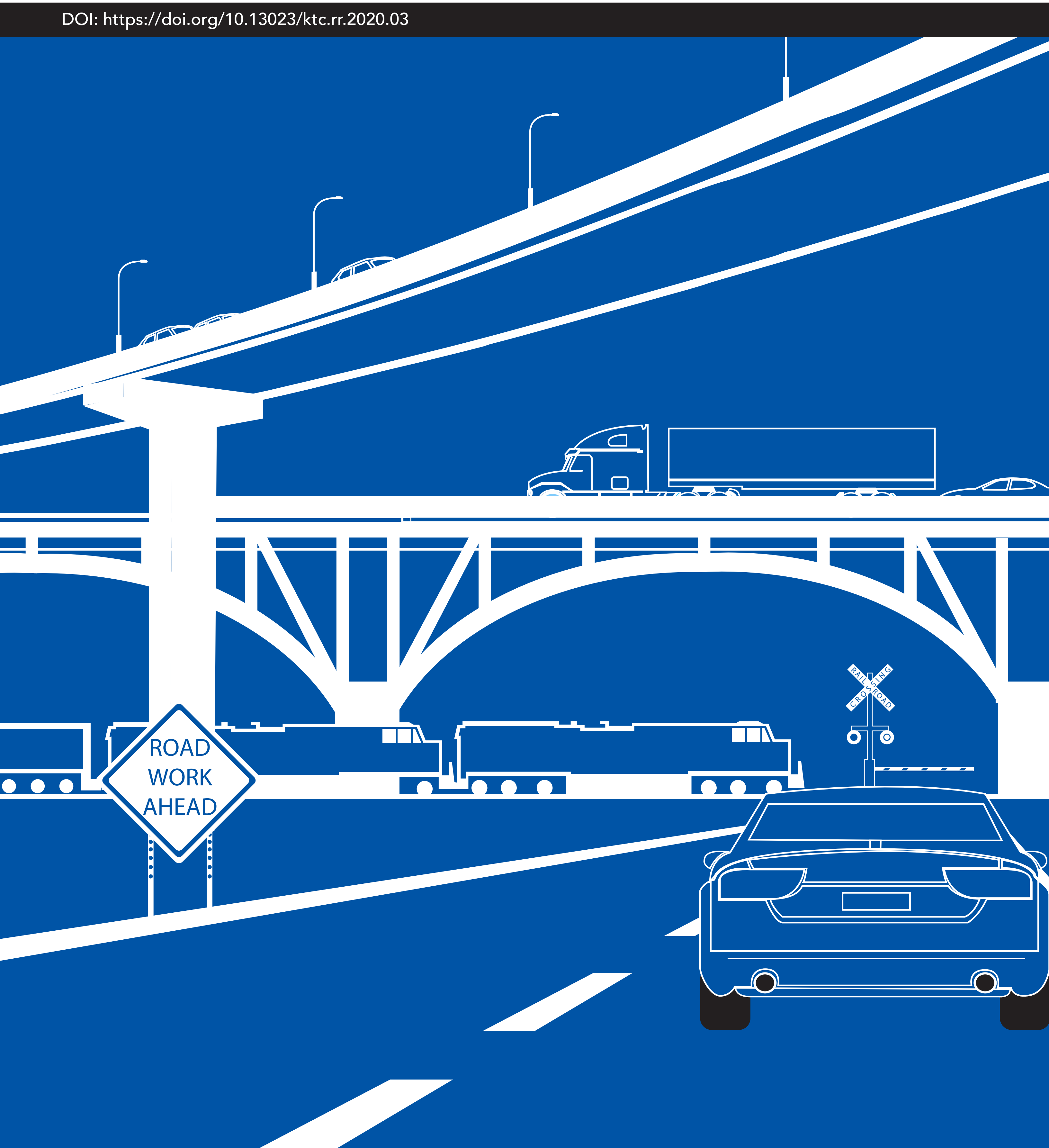
Effect of Socioeconomic Factors on Crash Occurrence

Report Number: KTC-20-03/SPR19-56-2-1F

DOI: https://doi.org/10.13023/ktc.rr.2020.03



Kentucky Transportation Center
College of Engineering, University of Kentucky, Lexington, Kentucky

in cooperation with
Kentucky Transportation Cabinet
Commonwealth of Kentucky

The Kentucky Transportation Center is committed to a policy of providing equal opportunities for al persons in recruitment, appointment, promotion, payment, training, and other employment and education practices without regard for economic, or social status and will not discriminate on the basis of race, color, ethnic origin, national origin, creed, religion, political belief, sex, sexual orientation, marital status or age.

Kentucky Transportation Center
College of Engineering, University of Kentucky, Lexington, Kentucky

in cooperation with
Kentucky Transportation Cabinet
Commonwealth of Kentucky

KTC KENTUCKY
Transporation Center

Kentucky Transportation Center • University of Kentucky
176 Raymond Building • Lexington, KY 40506 • 859.257.6898 • www.ktc.uky.edu

University of
Kentucky

Research Report
KTC-20-03/SPR19-56-2-1F

Effect of Socioeconomic Factors on Crash Occurrence

Nikiforos Stamatiadis, Ph.D.
Professor of Civil Engineering

Shraddha Sagar
Research Associate

Samantha Wright
Senior Lecturer in Civil Engineering

and

Aaron Cambron
Research Associate

Kentucky Transportation Center
College of Engineering
University of Kentucky
Lexington, Kentucky

In Cooperation With
Kentucky Transportation Cabinet
Commonwealth of Kentucky

The contents of this report reflect the views of the authors, who are responsible for the facts and accuracy of the data represented herein. The contents do not necessarily reflect the official views or policies of the University of Kentucky, the Kentucky Transportation Center, the Kentucky Transportation Cabinet, the United States Department of Transportation, or the Federal Highway Administration. This report does not constitute a standard, specification, or regulation. The inclusion of manufacturer names or trade names is for identification purposes and should not be considered an endorsement.

May 2020

1. Report No. KTC-20-03/19-56-2-1F	**2. Government Accession No.**	**3. Recipient's Catalog No**
4. Title and Subtitle Effect of Socioeconomic Factors on Crash Occurrence		**5. Report Date** May 2020
		6. Performing Organization Code
7. Author(s): Nikiforos Stamatiadis, Shraddha Sagar, Samantha Wright, Aaron Cambron		**8. Performing Organization Report No.** KTC-20-03/19-56-2-1F
9. Performing Organization Name and Address Kentucky Transportation Center College of Engineering University of Kentucky Lexington, KY 40506-0281		**10. Work Unit No. (TRAIS)**
		11. Contract or Grant No. 19-56-2
12. Sponsoring Agency Name and Address Kentucky Transportation Cabinet State Office Building Frankfort, KY 40622		**13. Type of Report and Period Covered** Final
		14. Sponsoring Agency Code
15. Supplementary Notes Prepared in cooperation with the Kentucky Transportation Cabinet		

16. Abstract

Road traffic crashes are a leading cause of death in the United States. In Kentucky, per capita crash rates and crash-related fatalities have outpaced the national average for over a decade. Wanting to explain why the U.S. Southeast sees higher crash rates than other regions, researchers have argued the region's unique socioeconomic conditions provide a compelling explanation. Taking this observation as a starting point, this study examined the relationship between highway safety and socioeconomic characteristics using an extensive crash dataset from Kentucky. This research sought to identify at-risk drivers based on the socioeconomic and demographic attributes of the zip codes in which they reside. Using the quasi-induced exposure approach, binary logistic regression was used to develop predictions of driver at-fault probability based on socioeconomic characteristics of their residence zip code. Statistical analysis found that variables such as income, education level, poverty level, employment, age, gender, rurality, and number of traffic-related convictions of a driver's zip code influence the likelihood of their being at fault in a crash. This finding can be used to identify groups of drivers most likely to be involved in crashes and develop targeted and efficient safety programs. Spatial analysis did not uncover robust correlations between county-level socioeconomic characteristics and at-fault driver involvement across the state. The results can be used to identify target groups for safety improvements and aid in the Kentucky Safety Circuit Rider Program activities.

17. Key Words highway safety; socioeconomic factors; social factors; quasi-induced exposure; traffic crashes		**18. Distribution Statement** Unlimited with approval of the Kentucky Transportation Cabinet	
19. Security Classification (report) Unclassified	**20. Security Classification (this page)** Unclassified	**21. No. of Pages** 83	**19. Security Classification (report)**

Table of Contents

List of Figures

List of Tables

1. Introduction

According to World Health Organization (WHO), every year 1.25 million people die in road traffic crashes, while at least 20 million suffer from non-fatal crashes [1]. In the United States (U.S.), road traffic crashes are a leading cause of death. The National Highway Traffic Safety Administration (NHTSA) estimated that all traffic crashes in 2010 cost the U.S. economy $836 billion [2]. Kentucky has a higher overall crash rate per population than the national average. In 2016, the NHTSA estimated 22.5 crashes per 1,000 persons for the country, while Kentucky's rate was 37.3. In 2018, deaths per 100 million vehicle miles traveled (VMT) in Kentucky was 1.46 while the national average was 1.13. According to an Insurance Institute of Highway Safety report from 2018, Kentucky ranked 5th in fatalities per 100 million VMT [3]. In 2019, 728 fatalities were reported in Kentucky while in 2018 there were 722 [4]. These trends underscore the importance of addressing the factors that could influence high collision rates and implementing effective policies to reduce them. Addressing the underlying issues that lead to safety problems will improve overall roadway safety.

If transportation agencies are to implement effective countermeasures, it is important for them to understand the underlying factors contributing to crashes. In several previous attempts, driver behavior, demographic factors, socioeconomic features, geometric design, and roadway characteristics have been identified as associated factors [5-12]. Past research efforts have demonstrated the significant influence of macro-level socioeconomic features on crash occurrence (e.g., poverty, income, employment and education) [5, 8, 10, 13, 14]. Many of these studies concentrated on the socioeconomic factors of the region where the crash occurred. Maciag [15] compiled fatal pedestrian crashes reported in Fatality Analysis Reporting System (FARS) for the 2008 to 2012 period to study the relationship between fatal crashes and economic conditions of the crash location. He found that fatalities are generally more common in poor socioeconomic areas. Also, historical crash data analysis by NHTSA indicated that crash rates are 2.5 times higher in rural areas than in urban areas [16]. These studies underscore the greater potential for crashes to occur in socially and economically disadvantaged areas. Though it is important to examine the socioeconomic characteristics of the region where a crash occurs, focusing on the residence characteristics associated with the origin of the drivers causing a crash may yield more information.

Prior research attempted to determine the association between socioeconomic factors related to driver residence and crash occurrence and estimate their role in crash occurrence [17-20]. A recent WHO study identified that people of lower socioeconomic background are more likely to be involved in crashes, with causation factors including human errors (e.g., speeding, lack of restraints, distracted driving, driving under the influence, inadequate roadway infrastructure, traffic law enforcement) [1]. Blatt and Furman also reached the same conclusion through an examination of the correlation between socioeconomic characteristics of the driver residence and crash occurrence [18]. They demonstrated that fatal crashes are more likely to occur on rural roads, while drivers who reside in rural areas or small towns have significantly higher involvement in such crashes. Several other studies have confirmed the high risk of crash involvement for drivers residing in a rural/poor neighborhood [5, 10, 17, 21].

Stamatiadis and Puccini [14] showed that the U.S. Southeast experiences consistently higher fatality rates compared to other regions. They noted that the distinct socioeconomic characteristics of the region are a significant reason that could explain the high fatality rate. They identified potential socioeconomic factors that could explain the high fatality rates in those regions, including the median household income, unemployment, educational attainment, and percentage of rural population. The study suggests the socioeconomic data of the zip code in which a driver resides could serve as a potential surrogate measure for explaining the high fatality rates.

A plausible explanation for Kentucky's higher crash rates may be the differences in a variety of socioeconomic characteristics of the state compared to other states. Based on statistics from the U.S. Census Bureau, Kentucky has lower percentages of high school completion and university attendance than the national average [22]. With respect to income characteristics, most of the counties have a median family income 19 percent lower than the national median income, are at the bottom of the national rankings with respect to both income and disposable income per capita, and have one of the largest percentages among the states of people below the poverty level. These socioeconomic characteristics could influence highway safety by affecting the age of vehicles owned (older, less safe vehicles), vehicle condition (not properly maintained), the attitudes of the drivers toward safety and risk-taking behaviors, and the level of driving education available (Stamatiadis and Puccini, 1999). Moreover, Kentucky is considered a rural state since more than fifty percent of its counties are classified as rural [23].

The number of traffic collisions is gradually increasing in Kentucky. Being in the Southeast, Kentucky's socioeconomic profile is suspected to be a significant reason for these recent increasing crash trends. It is apparent that there may be some connection between socioeconomic factors and crash occurrence, and therefore it is critical to examine their impact on each other. It is also important to determine how demographic data of driver residence influences crash involvement. Analyzing these factors might help identify the major causes for increasing crash trends, and in turn, highlight areas that may require more attention for improving overall roadway safety.

The primary goal of this research is to define at-risk groups of drivers based on the socioeconomic characteristics of the driver residence. Leveraging Kentucky historical crash data, this study used statistical and spatial analyses to investigate the socioeconomic and geodemographic factors of driver residence zip code. Spatial analysis sought to identify correlations between income patterns throughout the state and crash involvement by age and gender. Statistical analysis attempted to forge a predictive approach for estimating crash involvement probability based on socioeconomic and demographic characteristics of the driver's residence zip code. The main objective of the study is to identify factors that could potentially be indicators of crash occurrence. Study findings identify groups of drivers at a higher risk for being involved in crashes. It is important to determine whether these drivers, who may contribute to the future crash risk, belong to a particular group (e.g., age, gender) or region (e.g., rural/urban). This will provide better evidence for implementing efficient safety programs that target such groups.

2. Literature Review

Significant research has been undertaken globally to investigate whether socioeconomic and demographic factors impact crash occurrences. Some methods investigate demographics surrounding the crash location, while others use surrogate descriptors associated with the residence location of drivers involved in a crash. The following sections discuss past research focused on identifying socioeconomic and demographic factors that can explain crash involvement as well as methods used to investigate these relationships.

2.1 Socioeconomic and Demographic Variables

Various socioeconomic and demographic variables have been examined to identify their potential contribution to crash occurrence. Prior research shows some common threads among explanatory variables, which agree with a priori expectations: income, poverty, employment, education, rurality, and driver age all seem to have an impact [5, 8, 10, 14, 17, 24].

Rural areas are generally cited as having higher fatality crash rates than urban areas, and a large portion of previous research dealt with the levels of rural and urban percentages of a region. Muelleman and Mueller [25] investigated fatal commercial motor vehicle (CMV) crash characteristics as they relate to population density. Information on human variables (age, gender, restraint use, alcohol, ejection from vehicle, seating position, and driving record), vehicle variables (vehicle make, crash type, manner of leaving scene, and most harmful event), and crash variables (crash location, crash time, posted speed limit, first harmful event, surface type, and emergency medical system (EMS) times) were included in the analysis. Counties in the study regions were categorized as urban and rural; rural counties were subdivided into three groups based on population density. The major factors significantly related to the high fatality rates in low density areas were prevalence of alcohol use and higher levels of intoxication, delayed medical care, use of light and heavy trucks, frequent non-collisions (defined as a crash with no injuries or damages) on less travelled roads, and frequent crashes on gravel roads. Also, the study confirmed the previously known inverse relationship between population density and motor vehicle crash (MVC) fatality rates. They concluded that the fatality rate per 100 million VMT was 44 percent higher in rural than urban areas. They also noted that rural areas are not homogeneous, and comparisons based only on urban/rural groupings can obscure. However, variables like restraint use, crash severity, and older occupants showed no difference between the three rural regions, raising concerns regarding their contribution to explaining the relationship between fatality rate and population density. Though this research recognized many crash variables associated with population density, it did not determine the relative contribution of each factor to explaining the differences in fatality rates within rural areas. The authors recommended further research to determine how the fatality rate increases in areas with low population density are associated with pre-crash, crash, and post-crash variables. However, there has not been relevant research conducted on this.

Blatt and Furman [18] conducted a similar geodemographic analysis at the zip code level, with a focus on the residential location of the driver (characterized as rural and urban). Five levels of population density were identified for classifying driver residence location: rural, small town, second city, suburban, and urban. Other driver characteristics were divided into social clusters (age groups, gender, involvement in crash resulting in death of a child, and blood alcohol concentration level). Using geodemographic analysis, the percentage of drivers in fatal crashes in each social cluster was compared to the base population of that social cluster. Overall findings indicated that drivers from rural areas or small towns were more likely to be involved in fatal crashes and that those fatal crashes were more likely to take place on rural roads. The authors acknowledged that roadway features (e.g., two-lane highways, narrow shoulder, limited sight distance) may play a bigger role in rural crashes while economic and behavioral factors (e.g., use of seat belts, poor EMS response time, longer travel time to reach the nearest medical facility) could contribute to serious crash outcomes. Zwerling et al. [26] investigated the factors associated with increased fatal crash involvement rates in rural areas. They found that fatal crash incidence density was more than two times higher in rural than in urban areas. The major reason for this is the high rate of increased injury severity in rural crashes, which is three times higher in rural areas compared to that in urban areas.

Noland and Quddus [24] used negative binomial (NB) regression to explore the association between crash casualties and land use variables (proportion of urbanized area, population density, employment density), road characteristics (length of various road types, number of junctions and roundabouts) and area-wide demographics (age, level of social deprivation, percent of economically active population). NB models were developed for total fatalities, serious injuries, and slight injuries. The results indicated that densely populated urban areas had fewer traffic causalities, while areas with higher employment had more traffic causalities. Roadway characteristics did not exhibit any influence on traffic casualties, although the length of the road segments showed some effects on serious injuries. Social deprivation showed a positive relationship with traffic causalities but no significance for motorized (excluding bicyclists and pedestrians) casualties. The residual cause for high causality rate in areas with higher levels of social deprivation was not investigated. They offered as a possible explanation for their findings the possibility that lower income people tend to live in areas with low cost of living and cheap housing; such areas are likely to have unsafe roadway conditions. Further reviewing this dimension would be useful to identify target areas or populations that need more attention.

Hasselberg et al. [9] determined that drivers with a relatively low educational attainment level show an excess risk for overall crashes and crashes leading to fatality or serious injury. Their study also estimated that 33 percent of minor injuries and 53 percent of severe injuries would be avoided if all subjects had the same injury rate as subjects with a higher education. Similarly, Zephaniah et al. [10] revealed that driving under the influence (DUI) crash rates (normalized by population) are influenced by employment, income, education, and housing characteristics. Areas with high rental housing percentages exhibited lower DUI crash rates. The rate of DUI crashes was higher in rural areas, possibly indicating acceptance of drunk driving among communities living in those regions. Also, the overall percentage of residents with at least a high school education in a postal code reduced the frequency of DUI crashes. Their study also showed that DUI crashes were related to lower male employment and lower female education achievement, while Cook et al. [27] confirmed higher DUI crash involvement for male drivers. These studies used characteristics of the driver's residence and showed that a higher education has a positive impact (i.e., reduction) on vehicle crashes.

Several researchers have cited income and poverty as relevant predictors for crash-related analysis. Although income and poverty could be closely related, as poverty status is generally based on income below a certain level. Lee et al. [17] investigated the relationship between at-fault driver residence characteristics and all types of crashes using three years of data from Florida. They found that median family income had a negative relationship with the number of at-fault drivers, indicating that drivers from lower income communities were more likely to be at fault. Maciag [15] indicated that in metro areas, low-income tracts recorded pedestrian fatality rates were approximately twice that of more affluent neighborhoods; tracts with high poverty rates displayed a similar trend. Aguero-Valverde et al. [11] also concluded that the percentage of the population living under the poverty line had a highly significant and positive correlation with crash risk when using an NB prediction model.

Another factor that has been cited is employment, measured in terms of unemployment rates, portion of people working from home, or portion of unskilled workers. Factor et al. [8] used a sample of the Israeli population with detailed socioeconomic data and nine years of crash data for their analysis. They found that non-skilled workers were over-involved in fatal crashes relative to their percentage of the total population of all workers. Conversely, Lee et al. [17] found that a higher proportion of the population working from home resulted in a lower number of at-fault drivers, though it was proposed that this resulted from travel exposure. Later, Adanu et al. [7] found that unemployed drivers had a probability of 0.23 of being at-fault in a crash, while the probability of being at-fault in a serious injury crash was 0.57. They suggested that the odds of an unemployed driver being at-fault for a serious crash were 1.32 times higher than for a driver who was employed, self-employed, or retired. In addition to employment, they attempted to demonstrate that average credit scores (lower scores equal higher risk) and average commute times (longer times equal higher risk) are significant predictors for severe injury crash risk. At the driver level, the results showed higher proportions of serious injury crashes involved no seat belt usage, unemployed drivers, young drivers, distracted driving, and the driver's race. The model also showed a previously known inverse relationship between population density and crash severity, however, the authors made a counterintuitive argument. Based on their opinion, larger populations are more likely to live in urban areas having higher overall incomes and educational levels, which are factors that may influence crash occurrence and severity. Even though the influence of population density and vulnerability of rural areas to severe crashes had been established by previous studies, the authors suggested more detailed investigation of less populated regions is needed to better understand the relationship between driver characteristics and specific crash types.

Age is a key factor that contributes to a driver's involvement in a crash. Brown et al. [5] attempted to identify and analyze the socioeconomic and demographic factors related to the residential characteristics (at zip-code level) of drivers involved in crashes. Their study showed that drivers in the 15-19 age group have the highest odds of being at risk for an injury or fatal crash, followed by the 20-24 age group. The middle age group (45-54) drivers had the lowest odds of being at fault in a crash. Chen et al. [28], Factor et al. [8] and Hanna et al. [12] all indicated that undesirable crash results, such as more crashes or higher fatality rates, were present for young or new drivers, but the impact of elderly drivers varied. It might be that the young drivers have a tendency to speed more than older drivers [16]. Lee et al. [17] determined that a larger proportion of elderly population decreases the likelihood of drivers being at fault. Also, Adanu et al. [7] found that older drivers (above 65 years) have the least contribution to fatal crashes. This might be because the older drivers contribute less to the socioeconomic features (for example, median income) of a region, compared to the other age groups. Males [29] showed a joint effect of age and income-related factors on young driver fatalities. Using a multivariate regression analysis, he concluded that driver age is not a significant predictor of fatal crash risk when controlling for poverty-related factors (such as older vehicle age, lower state per capita income, and lower education levels). Aguero-Valverde and Jovanis [11] indicated that counties with a higher percentage of the population under the poverty line; higher percentage of their population in age groups 0–14, 15–24, and over 64; and those with increased road mileage and road density have a significantly increased crash risk. Several studies of older adult drivers discuss the risk factors they create for themselves and others. Lyman et al. [30] observed an increasing fatal crash rate for drivers over 70 years of age. The study found that drivers over 65 years of age will account for more than half of the total increase in fatal crashes by 2030. However, the contribution of different age groups to crash severity was unclear and requires further investigation.

In addition to age, crash occurrence is often associated with the driver's gender and marital status (separated or widowed) [8, 13]. Factor et al. [8] provided evidence that separated and widowed drivers are 50 percent more likely to be involved in a crash than married drivers. In terms of at-fault drivers, the proportion of males is higher than that of females. For the state of Kentucky, 55 percent of the drivers involved in collisions during 2016 (where the gender was listed) were male and 45 percent female. In fatal collisions, 74 percent of the drivers were male and 26 percent female. Zephaniah et al. [10] showed that DUI crashes are related to male employment and female educational attainment. Additionally, there might be a joint relationship between other socioeconomic factors (like income) and gender and age, which requires more investigation.

Another interesting factor contributing to crash occurrence is proximity to driver residence [5, 31]. A latent class analysis (a model-based clustering method), considered by Adanu et al. [31], indicated that more than 75 percent of young at-fault driver crashes occurred within 25 miles of the driver's residence. However, Brown [5] showed that approximately 35 percent of the crashes occur within 5 miles of the driver's residence. Additional investigation has been recommended to analyze how crash occurrence is influenced by proximity to driver residence for specific target groups (e.g., age, gender, educational attainment) or regions (e.g., rural/urban area).

Apart from socioeconomic and demographic characteristics of the driver's residence, driver history plays a major role in dictating future crash risk. Chandraratna [19] demonstrated that a driver with one previous at-fault crash is about 150 percent more likely to be involved in another crash than a driver with no previous at-fault crash involvements. His study also demonstrated that drivers whose driving records have citations, crashes, or both are high-risk drivers. Even though his research estimated the likelihood of a driver being involved in a future crash, estimates were limited to only at-fault drivers with previous crash records. Many past research investigations looked at the impact of crash-prone drivers on safety and developed models predicting how a driver's past crash history could affect their crash occurrence(s) in the upcoming year [32, 33].

2.2 Citation and Crash History

According to the NHTSA about 94 percent of serious crashes are due to dangerous choices or errors people make behind the wheel [34]. It is critical to identify high-risk drivers and their characteristics to reduce crashes through targeted efforts such as safety education and enforcement programs. Many researchers have demonstrated that one driver being involved in multiple crashes is more than a coincidence. Greenwood and Yule [35] first documented the existence of crash-prone drivers. Other research has investigated the impact of crash-prone drivers on safety and developed models predicting how a driver's crash history could affect their crash occurrence(s) in the upcoming year [40, 43].

Blasco et al. [33] investigated how the probability of a driver involved in a crash changes when they already have one previous crash involvement. They noted that the less the time elapsed between two crashes, the higher the probability of a driver being involved in another crash. Therefore, drivers with convictions and a crash history are considered to be a higher risk. In 2002, Daigneualt et al. [36] examined older drivers' previous conviction record and crash data and concluded that prior crashes are a better predictor of crash risk than prior convictions. Chen et al. [37] identified crash-prone drivers based on their at-fault crash involvement in prior records and found that a statistical model using prior at-fault crash data can recognize up to 23 percent more drivers who will have one or more at-fault crash involvements in the next two years than those using conviction information only.

Using Louisiana data, Sun et al. [32] investigated the impact of crash-prone drivers on safety to predict how a driver's past crash history affects their crash involvement in the upcoming year. Their findings showed that 5 percent of drivers were responsible for 35 percent of crashes during a seven-year time period. They concluded that the probability of a driver with crash history being involved in a future crash is more than seven times higher than the probability of drivers with zero crashes. Chandraratna [19] also demonstrated that a driver with one previous at-fault crash was about 150 percent more likely to be involved in a another crash within the next two years than a driver who had no previous at-fault crash involvements. His study also demonstrated that drivers whose records contain citations, crashes, or both are high-risk drivers. Even though his research estimated the likelihood of a driver being involved in a future crash, estimation was limited to only at-fault drivers having previous crash records.

2.3 Analysis Methods

The NB distribution is a discreet probability distribution often used when dealing with crash counts, and NB regressions are used to model crash counts for a roadway segment. Noland and Quddus [24] used NB count data models to analyze the associations between demographic factors (e.g., land use types, road characteristics and area-wide demographics, including level of social deprivation) with traffic fatalities and serious or slight injuries. Social deprivation is measured using an index developed in the United Kingdom that examines six socioeconomic factors: income, employment, health deprivation and disability, education skills and training, housing, and geographical access to services. They used census blocks in England as the spatial units of the crash location to connect these demographics with crash fatalities. More recently, the *Highway Safety Manual* (HSM) recommended developing Safety Performance Functions (SPFs) using NB regressions, which are primarily based on average annual daily traffic (AADT) for homogeneous roadway segments. However, Ivan et al. (2016) demonstrated an alternative for predicting crashes on local roads if traffic volumes are not available. The study estimated SPFs for local road intersections and segments at the Traffic Analysis Zone (TAZ) level using socio-demographic and network topological data. There are approximately 1,800 TAZs in Connecticut, which cluster into six analysis groups based on land use and population density. SPFs were developed using Poisson regression models, which predicted intersection and segment crashes within each TAZ using the number of intersections and the total local roadway length, respectively.

Other forms of regression modeling have been used in crash analysis. La Torre et al. [38] and Rivas-Ruiz et al. [39] used multiple linear regression in their analysis, while Chen et al. (2015) used a Bayesian random intercept regression model. La Torre et al. [38] investigated the association between regional differences in traffic crash mortality and crash rates with socio-demographic factors and variables describing road behavior, vehicles, infrastructure, and medical care in Italy. Rivas-Ruiz et al. [39] used simple and multiple linear regression with a backwards stepwise elimination approach to study the variability of Road Traffic Injury (RTI) mortality on Spanish roads, adjusted for vehicle kilometers traveled (VKT) in each Spanish province. Both studies found some significance in area-wide socioeconomic factors, such as employment rates, alcohol use, and education levels. Chen et al. [40] analyzed injury or fatal truck driver crashes. The study concluded that the presence of alcohol or drugs correlates positively with crash severity.

Some have found other regression models to be more useful, such as logistic and lognormal regressions. Logistic regression is the simplest form of regression that can be used when the dependent variable is binary. This technique fits the best when the effect of more than one independent variable (categorical, continuous, or both) is examined. Factor et al. [8] created a binary response variable to describe crash fatality level. The model used demographic factors to predict the probability of being involved in a fatal crash versus a non-fatal crash. The research linked nine years of injury and fatal road-crash records with census data and used several socioeconomic factors grouped into discrete categories, such as gender, education groups, and age groups. The binary dependent variable indicated whether the driver had been involved in a fatal or severe accident within the past nine years. They also used categorical independent variables such as gender, age groups, and marital status for analysis. Findings of the regression were then expressed as probabilities, which is one of the major contributions of logistic regression. Vachal [41] used logistic regression to study crash factors in relation to injury outcomes for single and multivehicle truck crashes. The research noted that while drugs and alcohol are potentially a contributing factor for truck drivers, substance use is more common and more dangerous for drivers of passenger vehicles.

Similarly, Hanna et al. [12] considered fatal crashes involving unlicensed young drivers (under age 19) in the U.S. using conditional and unconditional logistic modeling. This analysis was based on the urbanicity (which categorizes all U.S. counties as urban, suburban or rural based on population and proximity to metropolitan areas) and the Townsend Index of Relative Material Deprivation (which serves as a proxy measure for socioeconomic status based on access to local goods, services, resources, and amenities). To allow for the simultaneous study of driver characteristics and region information, Adanu et al. [7] used multilevel logistic modeling, which recognizes "the hierarchical structure in data and also provide[s] information to compute the amount of variability in the data attributable to each level of the hierarchy." They created a binary response variable which identifies crashes as fatal or non-fatal. They used a two-level hierarchical logit model with driver characteristics at level 1 and regional information at level 2. In sequential or hierarchical logistic regression models, explanatory variables can be added to the model step by step, which allows the examination of how the model changes with the addition of each set of variables. This approach would allow for the development of models at each level and understanding of the effects of these predictors on the response variable, at the driver level and regional level. Similarly, Chen et al. [42]

used multinomial logit models to examine the influence of drugs or alcohol in increasing the probability of injury or fatality for CMV drivers. Khorashadi et al. [43] used a multinomial logit model to examine the effect of alcohol or drug use on rural road truck crashes. They concluded that the probability of severe/fatal injury increased 246 percent compared to crashes not involving alcohol or drugs.

Das et al. [44] conducted an explanatory data analysis to develop a crash prediction model which estimated the likelihood that at-fault drivers will be involved in future crashes. They categorized drivers into four types: not-at-fault prone drivers (involved in multiple crash but not responsible for), at-fault prone drivers (responsible for multiple crashes), not-at-fault non-prone drivers (involved in only one crash but not responsible for), and at-fault non-prone drivers (responsible for only one crash). Extensive data analysis was conducted to determine the association of these four driver categories with variables such as human-related factors, crash-related variables, roadway-related variables, environmental factors, and vehicle-related variables. The results of data analysis emphasized the importance of understanding the behavior and other associated characteristics of drivers involved in multiple crashes (i.e., crash prone drivers). A logistic regression model was developed for crash-prone drivers, with the dependent variable being the driver's fault status. The idea of categorizing the at-fault and not-at-fault drivers based on crash risk was a creative idea, however the model did not include all of them. The final model predicting fault status was limited to crash-prone drivers (i.e., drivers involved in more than one crash). To address this issue, a multinomial logistic regression modeling technique can be used – an extension of binomial logistic regression, allowing for a dependent variable with more than two categories. In this case, the dependent variable can be split in four driver categories as defined by the researchers. Using multinomial logistic regression, the crash proneness (or any other categorical variables such as gender and educational attainment) can be added as a categorical explanatory variable. This will help to understand how the categorical explanatory variables vary within the binary dependent variable. For example, this will help to determine how much more likely a crash-prone driver is to be at-fault than a non-crash prone driver.

Chandraratna et al. [19] approached this scenario differently. They tried to predict the likelihood of a driver's involvement in a crash based on previous crash involvement. The dependent variable was whether the driver had a previous crash involvement during the study period. They used the fault status of the driver as one of the independent variables. The results demonstrated that drivers who were previously at fault in a crash were more likely to be involved in additional crashes than other drivers. However, in this case a driver with one previous crash was considered riskier than a driver with five (for instance) previous crashes.

Other methods such as spatial analysis have been used in crash analysis utilizing socioeconomic factors. Brown [5] considered the residential locations of at-risk drivers (drivers reported as contributing to fatal crashes) and the demographic characteristics associated with those residential locations at the Census Block Group level. Socioeconomic variables for higher-risk block groups (more than 8 at-risk drivers per 1,000 driving population) were compared to those of lower risk groups to determine trends. This study used a cluster analysis identifying hot spots of high-or low-risk areas that can be targeted for specific safety programs. Of note here, is the fact that this study examined demographic characteristics tied to the driver's home location instead of the commonly used method of socioeconomic characteristics tied to the crash location. Kocatepe et al. [13] used hotspots to investigate the exposure of different age groups to severe injury crashes in the Tampa, Florida area. Severity-weighted crash hot spots were identified using the Getis-Ord Gi method, weighted by the number of severely injured occupants involved in each crash. The study examined the proximity of residents in different age groups (17 and younger, 18 to 21, 22 to 64, and 65 and older) to severity-weighted crash hotspots. Age, ethnicity, education, poverty level, and vehicle ownership all had an effect on crash injury exposure.

A less defined but widely used method for this type of research involves separating crash or socioeconomic data into groups and comparing them with descriptive statistics. Abdalla et al. [45] studied the effect of driver social circumstance on crash occurrence and casualty by linking crash records and census data in

Scotland's Lothian Region. The research showed a correlation between fatal crashes and a driver's distance from home. Socioeconomic variables were bundled into a Deprivation Index and postal codes were separated into the most affluent and most deprived to compare traffic casualties normalized by population. Similarly, Blatt et al. [18] considered fatal crashes occurring in rural areas, with a focus on driver residential location. Five years of crash data from FARS were linked to driver home zip code and other factors, including driver age, gender, and blood alcohol concentration. Five levels of population density were identified for classifying each driver's residence location: rural, small town, second city, suburban, or urban. Other driver characteristics were divided into social clusters (e.g., age groups). Using geodemographic analysis, the percentage of drivers in fatal crashes in each social cluster was compared to the base population of that social cluster. In additional research involving traffic fatalities, Maciag [15] investigated the differences in demographics between census tracts in relation to pedestrian fatalities in that tract. Census tracts were broken into categories by income and poverty to facilitate a direct comparison of pedestrian fatalities.

2.4 Summary

The socioeconomic factors most relevant to crash occurrence investigation are income, education level, poverty percentage, employment, driver age, and the rurality of an area. Education and income are typically negatively correlated with crash response. Poverty is positively correlated, while employment varied across studies. Young drivers, and areas with a high proportion of young drivers, tend to have a higher proportion of crashes and fatalities. In general, crashes in more rural areas exhibit more fatalities.

Past research has shown a relationship between crash involvement and age. Most literature shows a positive association between young (under 25) and older (over 65) drivers and crashes or fatalities. Several studies of older drivers identified their increased crash involvement and demonstrated the risk factors they create for themselves and others. Studies have also noted that young and old drivers have a positive relationship with crash involvement, indicating their higher propensity to be at fault in a crash. This study further examines these trends to determine whether they hold for Kentucky drivers.

Gender and marital status (separated or widowed) of the driver have also been identified as good predictors of crash occurrence. In Kentucky, 55 percent of the drivers involved in collisions during 2016 (where the gender was listed) were male [46]. In fatal collisions, 74 percent of the drivers were male. Similar trends have been observed over the years, and there may be crucial relationships between gender and crash occurrence (or crash severity) as it would be influenced by socioeconomic factors in Kentucky. In Alabama, Zephaniah et al. [10] showed that DUI crashes are related to male employment and female educational attainment. The percentage of drivers divorced and separated was considered in the preliminary analysis but not included in the final model due to multicollinearity. The current study investigates these interactions to determine whether they influence crashes.

Apart from socioeconomic and demographic characteristics of the driver's residence, previous crash records and citations are good predictors of crash occurrence. Even though few researchers have attempted to include crash history/citation in their analysis, its relationship with crash occurrence, adjusting for a driver's socioeconomic attributes, has not been examined. Das et al. [44] investigated crash-prone drivers (with multiple crash records) to define their likelihood of being at-fault in the future, while Chandraratna [19] attempted to predict the likelihood of a driver's involvement in a crash based on previous crash involvement. The former did not consider drivers with single crash involvement, leaving room for future research. The latter used previous crash involvement as the dependent variable for predicting the likelihood of a driver with previous crash involvement being involved in another crash. However, in this case a driver with one previous crash was considered as risky as the driver with five (for instance) previous crashes. This study accounts for citation information to predict the probability of a driver causing a future crash when adjusted for socioeconomic characteristics.

To investigate the role of these factors on crashes, many different methods have been used, and while all are valid, a wide range of analytical practices for relating socioeconomic characteristics with crash data are available. Many regression techniques have been applied as well as spatial statistics, clustering, and comparative grouping. The main objective of the current research is to identify factors that may predict the fault status of a driver using the socioeconomic and demographic characteristics of their residence zip code. In other words, the response variable is the driver's fault status, which is categorical. Logistic regression is the most appropriate and widely used method to answer this question due to the categorical nature of the dependent variable. This modeling technique is beneficial for examining the effects of more than one explanatory variable. Binary logistic regression is used to estimate the probability of a driver's fault status based on multiple independent variables.

3. Research Methodology

This project's main objective was to establish the relationship between crash occurrence and socioeconomic factors associated with the residence zip codes of at-fault drivers. Analysis also considers associated factors: crash-related factors, driver characteristics, and the socioeconomic and demographic features of the driver residence zip code. The final model can help decision makers identify driver groups needing attention, with a goal of increasing the safety of those groups. The following section describes the data and methodology used for analysis.

3.1 Socioeconomic Descriptor Factors

The literature review identified several factors which can help explain crash occurrence. This section summarizes how this study uses data on socioeconomic factors gathered from the literature review.

The most widely used variables are income, education level, poverty percentage, employment level, driver age, and the rurality of an area. Preliminary analysis showed typical correlations of these variables with crash occurrence; however, analysis considered crash data only for at-fault drivers [47]. These variables were also evaluated to address crash exposure in a more systematic manner and to investigate how crash exposure could affect the association between these variables and crash occurrence.

Driver age has been shown to be a good predictor for determining driver at-fault status. Previous research has shown that both young (under 25) and old (over 65) drivers are more likely to be at fault in a crash than the not at fault. This study also investigates these age groups in light of the socioeconomic factors by grouping of drivers into age groups.

In addition to age, the literature review identified gender and marital status (separated or widowed) of the driver as good predictors of crash occurrence. Cambron et al. [47] considered the percentage of drivers divorced and separated in their preliminary analysis, however, this was not included in the final model due to multicollinearity. Multicollinearity is a statistical phenomenon in which multiple factors are related to each other. It can cause unstable estimates and inaccurate variances, which affect confidence intervals and hypothesis tests. The data used here test the possibility of multicollinearity by examining the correlation matrix formed between the predictor variables. However, examining the correlation matrix may be helpful but insufficient for detecting multicollinearity. Cambron et al. estimated the variance inflation factor (VIF), a measure of multicollinearity, which assesses how the variance of a regression coefficient increases if predictors are correlated. But VIF is limited to ordinary least squares regression analysis and therefore cannot be used in binary logistic regression. Thus, the current study uses a Feasible Solution Algorithm (FSA) to detect possible interactions between predictor variables. It investigates these interactions to determine whether they influence crashes, since the proposed approach considers crash exposure as well.

Previous research showed a well-defined relationship between level of education and crashes. The percentage of people with different education levels and their relationship linked with gender are also significant descriptors of crash propensity [10]. Further, race of the driver is also a factor associated with crash occurrence [7]. However, research on the association between race and crashes is sparse. The current study evaluates the influence of race on crash occurrence.

The negative correlation of income and poverty level with crashes has been established. These variables have an underlying relationship with rurality, education, and employment status. It is more likely that people with more education have better employment and higher income. These people tend to live in urban areas with better housing facilities. Therefore, it is expected that the housing characteristics of zip codes are also a significant predictor of crash involvement.

The association of crash occurrence with previous crash records and citations is widely established. This information can be utilized as a predictive variable. This analysis is deemed appropriate, since the current study evaluates driver history while considering the socioeconomic and demographic characteristics of driver residence zip codes.

3.2 Variable Selection Methods

Many socioeconomic variables need to be tested against driver at-fault status. It is tedious and time-consuming to test all possible variable combinations to develop the best model with the most appropriate variables. As a first step toward variable selection and to better understand how socioeconomic variables relate to driver at-fault status, two statistical analyses were conducted: correlation analysis and recursive partitioning analysis. A stepwise selection process was used, where variables were added and removed, back and forth, in the logistic regression model to find the best candidates for predicting the response variable. Possible interactions were tested to develop a statistically stronger and mathematically stable model.

3.2.1 Correlation Test

A correlation test investigates the relationship between two variables. Point-biserial correlation is the statistical test used to measure the strength of association between a continuous variable and a binary variable [48]. It is a special case of the Pearson's product-moment correlation coefficient (or Pearson correlation coefficient) which is applied when the correlation test is conducted for a binary variable. It measures the strength of association of two variables in a single measure, the correlation coefficient (r), which ranges from -1 to +1. A result with a coefficient value equal to -1 indicates a perfect negative association, a value of +1 indicates a perfect positive association, and a value of 0 indicates no association. A value greater than 0 indicates a positive association (i.e., as the value of one variable increases the value of the other variable also increases). A value less than 0 indicates a negative association (i.e., as the value of one variable increases the value of the other variable decreases). This test also calculates a p-value, which indicates whether the association between two variables is statistically significant.

3.2.2 Recursive Partitioning Analysis

Recursive partitioning analysis is a statistical algorithm used for predictive modeling in statistics and machine learning [49]. It attempts to correctly classify data along a decision tree by splitting them into subgroups based on the variables at hand. It is an iterative process that builds a decision tree by sorting the independent variables down the tree based on how accurately they predict the target variables. The process continues until no more useful splits can be found. This method examines all the variables in a dataset to find the one that gives the best classification or prediction by splitting the data into subgroups. It helps in understanding the importance of the variables that should be considered in the modeling. The objective of using this approach was to obtain a set of variables that can be used for logistic regression and modeling of at-fault probability.

3.2.3 Stepwise Selection

Stepwise regression is a modeling technique in which variables are added to or removed from a model to find the best candidates for predicting the response variable [50]. This technique was used here not as a modeling technique but as a variable selection process. Using binary logistic regression, all candidate socioeconomic and demographic variables were examined to evaluate whether their p-value fell below the specified level of statistical significance. Non-significant variables were removed from the model. Following this process, the best subset of variables that defined the response variable was selected. In spite of these advantages, stepwise regression has many drawbacks as well. It is a bad idea to just select variables in the final model based only on their p-value. The removal of less significant predictors tends to increase the significance of a model's remaining predictors. Also, in the process of adding or removing variables one at a time, it is possible to miss the optimal model. Therefore, this study uses the findings from the correlation analysis and the classification and regression tree (CART) model to make the right choices regarding which variables to use.

In this variable-selection method, the variables with the largest correlation coefficient that end up in the CART model are added to the model. One by one, the strongest variables identified using the variable-selection methods are added and the model is refitted to estimate the new model parameters. The variation in p-value and the parameter estimates are noted after the addition of every variable. At each step after adding a variable, variables that are not significant at that level are eliminated. This process continues until every remaining variable is significant.

Following these steps, several models were developed with the best pair of variables. The models were then evaluated using different evaluation criteria to develop the best possible model for predicting driver at-fault probability for crash involvement.

3.2.4 Identifying Interactions

Interactions offer a better understanding of the relationship between predictors in a model. The inclusion of interaction terms, in addition to the main effects, can improve a model's mathematical stability [51]. Two (or more) independent variables interact if the effect of one of the variables shows dependence on the other variable(s). As noted above, several potential interactions among the socioeconomic variables might influence crash occurrence. It is tedious and time consuming to test all the combinations of variables that can potentially form an interaction, and for this reason many previous analyses have not attempted to explore interactions. In some cases, interaction terms are identified based on prior knowledge and they are screened one by one. This research attempts to search for an optimal model containing interactions using an algorithm developed by the Department of Statistics at the University of Kentucky [52]. A tool called "Shiny" uses an FSA to detect interactions. The algorithm allows for fixed, specified explanatory variables in the model and the addition of a feasibly best interaction [52]. It allows one to formulate new or to improve upon existing models. Several criterion functions (such as R^2 and adjusted R^2, interaction p-values, Akaike's Information Criterion and the Bayesian Information Criterion) were evaluated to examine model quality. FSA allows higher order interactions; however, this study is limited to two-way interactions.

Based on the results from variable selection methods, several combinations of explanatory variables were tested in the Shiny application to find the best solution. The results of the test are presented in the next section.

3.3 Crash Exposure – Quasi-Induced Exposure Technique

It is important to examine crash exposure when considering crashes and attempting to identify what factors contribute to a crash. Crash databases do not contain information on driver exposure. Typically, VMT, number of licensed drivers, registered vehicles, and similar exogenous factors have been used to define exposure. With these conventional metrics, the exposure proportion of the driving population may vary depending on other factors such as time of day, driver gender or age, road type, and so on. This has raised questions about the reliability and applicability of these exposure metrics when examining safety issues as they pertain to more specific groups of drivers or conditions, since the denominator in the ratio of crash occurrence for such subgroups and conditions cannot be obtained. The quasi-induced exposure technique developed by Carr (1969) overcomes this problem. The approach assumes that not-at-fault drivers represent the total population in question, and the crash rate measure of exposure is developed in terms of the relative accident involvement ratio (RAIR), which is the ratio of the percentage of at-fault drivers to the percentage of not-at-fault drivers from the same subgroup.

This ratio is defined in Equation 1:

$$\text{RAIR} = \frac{\text{proportion of at}-\text{fault drivers}}{\text{proportion of not}-\text{at}-\text{fault drivers}} \quad (1)$$

Chandraratna and Stamatiadis [53] examined the validity of this assumption using two samples of not-at-fault driver data: one with not-at-fault drivers selected from the first two vehicles in a multi-vehicle crash and a second that included not-at-fault drivers (excluding the first two drivers) from multi-vehicle crashes with more than two vehicles involved. They concluded that the two samples were statistically identical and therefore that "estimating relative crash propensities for any given driver type by using the quasi-induced exposure approach will yield reasonable estimates of exposure."

3.4 Statistical Modeling

Logistic regression is a classification algorithm generally used to model the probability of a certain group. As discussed in the literature review, logistic regression is the most appropriate and widely used method when the dependent variable is categorical. This modeling technique is beneficial when effects of more than one explanatory variable influence an outcome [44]. Independent variables can be discrete and/or continuous. In linear regression, expected values of the response variable are modeled based on a combination of explanatory variables while logistic regression is a linear model for binary classification predictive modeling. Model coefficients in a logistic regression model are estimated using a probabilistic framework called maximum likelihood estimation.

Mathematically, a logistic regression estimates a multiple linear regression function defined as:

$$\text{y} = \text{a} + b_1X_1 + b_2X_2 + \ldots + b_nX_n \quad (2)$$

where *y* is the dependent variable, *X*'s are the explanatory variables, *a* is the intercept and *b*'s are the coefficients of the explanatory variables. In this case, the left-hand side of the equation could result in negative values or values greater than 1, while *y* (the dependent variable) is categorical in nature (i.e., *y* should be 0 or 1). This problem is solved by transforming *y* so that the regression process can be used. The logit transform of the response variable is called log-odds or logit.

Mathematically,

$$\text{log odds or logit (P)} = \text{a} + b_1X_1 + b_2X_2 + \ldots + b_nX_n \quad (3)$$

$$\forall \text{ log odds or logit (P)} = \ln\left(\frac{p}{1-p}\right) \quad (4)$$

$$= \ln\left(\frac{\text{probability of presence of characterestics}}{\text{probability of absence of characterestis}}\right)$$

Here, *p* is the probability an event will occur. In the context of the current study, *p* is the probability of a driver being at-fault when involved in a crash. The logit transform of the response variable is called log odds or logit. Therefore, the logistic regression defines the log odds for the response variable as a linear combination of explanatory variables.

Combining Equations 3 and 4,

$$\ln\left(\frac{p}{1-p}\right) = \text{a} + b_1X_1 + b_2X_2 + \ldots + b_nX_n \quad (5)$$

Here, the ratio of the probability of at-fault drivers to the probability of not-at-fault drivers is called the odds ratio. It is equivalent to the RAIR, which is the driver exposure measure in the quasi-induced exposure technique.

After taking the anti-logarithm of Equation 5 and replacing the regression equation with $f(X)$, the equation for the probability of the characteristics of interest is expressed as a function of the regression equation:

$$p = \frac{e^{f(X)}}{1 + e^{f(X)}} \tag{6}$$

On further mathematical manipulation, Equation 6 takes its final form,

$$p = \frac{1}{1 + e^{-f(X)}} \tag{7}$$
$$\forall \quad f(X) = \mathrm{a} + b_1X_1 + b_2X_2 + \ldots + b_nX_n$$

where $f(X)$ is the regression model, X_i is the ith explanatory variable, a is the intercept, and b_i is the ith coefficient estimated using the maximum likelihood method.

Logistic regression results can be displayed as odds ratios or probabilities. Odds ratios quantify the strength of association between two events. In simpler words, it is the ratio between the odds describing two events.

3.4.1 Relative Accident Involvement Ratio

Binary logistic regression was used in this research to develop a regression model to predict a driver's fault status based on different socioeconomic and demographic variables. Equation 7 can be used to estimate the likelihood of a driver belonging to a particular zip code (with specific socioeconomic and demographic factors) being the at-fault driver in a crash. Here, p is the probability of a driver being at fault, while considering as exposure drivers with the same characteristics not at-fault in a crash. Equation 7 is analogous to the RAIR used in the quasi-induced exposure methodology and is the measure of crash propensity (as discussed in the previous section).

Considering the probability of a driver being at fault calculated as p, the RAIR of a driver group is calculated using Equation 8.

$$\text{RAIR (at-fault)} = \frac{\mathrm{p}}{1 - \mathrm{p}} \tag{8}$$

The following example demonstrates the use and interpretation of RAIR. Stamatiadis and Puccini [14] indicated that in the Southeast male drivers cause 78 percent of single-vehicle fatal crashes and 70 percent of multivehicle crashes. This indirectly means that the female drivers are responsible for the remaining fatal crashes. Considering exposure data, males represent 73 percent of the driving population involved in multivehicle crashes. So, the RAIR for men causing a single-vehicle fatal crash is 78/73 = 1.06, while for female the ratio is (100-78)/(100-73) = 0.81. Similarly, the risk ratio for male and female drivers for multivehicle crashes can be calculated. When they analyzed the involvement ratios by gender, they concluded that even though males are more likely to cause single-vehicle crashes, females are more likely to cause multivehicle crashes (Figure 1). This may be explained by the different levels of risk that each gender is willing to take.

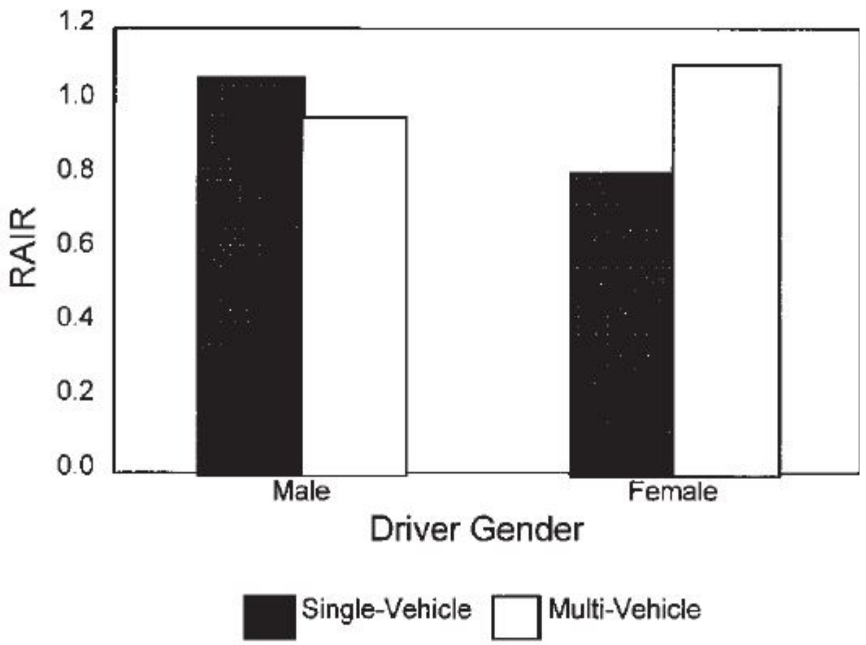


Figure 1 RAIR for Driver Gender

The quasi-induced exposure approach was used here to define the exposure of the driver by assuming that not-at-fault drivers represent the general population. The response variable is categorical (i.e., at-fault and not-at-fault driving status of the driver), and logistic regression is the most appropriate method to analyze this binary dependent variable.

Based on the probabilities developed using logistic regression, target groups/target areas with high crash propensity can be identified for more detailed examination. This will help policymakers focus their efforts to improve safety using targeted efforts and specific road safety campaigns.

3.4.2 Evaluation Criteria

Several models were developed for Kentucky based on qualitative and quantitative variable selection. These models have undergone several model evaluations to produce the best possible model. The model evaluation criteria are explained below.

Likelihood Functions

The two likelihood functions used for model evaluation were Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). They are estimators of the relative quality of statistical models for a given dataset and criteria for model selection among a finite set of models. Models with the least likelihood function are preferred. One of the main drawbacks of these criteria is the possibility of an increase in likelihood when more parameters are added, which may result in overfitting.

Receiver Operating Characteristic Curve

The receiver operating characteristic (ROC) curve is a graphical plot that illustrates the performance measurements of a model. It is a probability curve plotted between the true positive rate (or sensitivity) and false positive rate (or 1-specificity) that represents the model's capability in distinguishing between the two classes (i.e., driver at-fault status). The area under the curve (AUC) represents the degree or measure of separability between the two classes. An excellent model has an AUC near to 1, which means it has good measure of separability. A poor model has an AUC closer to 0, which means it is reciprocating the result (i.e., predicting 0's as 1's and 1's as 0's).

Training and Validation Method

In this method, the dataset is randomly divided into two parts – a training set and a validation set. The model is developed using the training set and the fitted models are used to predict the responses for the validation set. The percentage correctly predicted is calculated to evaluate the model's capability to represent the data. In general, the training set is larger than the validation set to ensure the training set is a good representation of the overall dataset. Here, 80 percent of the cases were placed in the training dataset and 20 percent of cases were put into the validation dataset.

Probability Residual

In regression analysis, residuals play an important role in model validation. By definition, the residual is the difference between the observed value and the predicted value of the dependent variable. These residuals are an estimate of the model error and are used to validate the model. The smaller the residual, the better the model. This study used logistic regression model to predict the probability of a driver being at fault in a crash based on age, gender, and socioeconomic characteristics of the driver's residence zip code. Here,

$$\text{Residual} = \text{observed probability} - \text{predicted probability} \qquad (9)$$

The first step is to calculate the observed probability from raw data. The observed value is the actual probability of a driver being at fault. Seven age groups and two gender groups were formed. Hence there were 14 possible categories for age-gender combination. The probability of drivers in a particular age-gender group being at-fault was estimated using data for each zip code.

3.4 Model Development Approach

Many socioeconomic variables required testing against the driver at-fault status. To simplify the tedious and time-consuming process of testing all the possible variable combinations, two statistical analyses were conducted: correlation and recursive partitioning analysis. These processes reduced the number of factors or predictors that need to be considered in a model, and their results were used as a starting point for developing a logistic regression model. Correlation analysis investigated the relationship between the dependent variable and the socioeconomic variables. It calculated a p-value that represented the significance of the association between the variables. Statistically significant explanatory variables were narrowed down for a starting point in variable selection.

Since the dependent variable in this study was categorical, the Pearson coefficient may not be an appropriate measure to explain the relation between crash occurrence and socioeconomic variables. Instead the recursive partitioning analysis could be more appropriate, which is another statistical technique used to understand the association between the potential predictor and dependent variables. It helps in developing a tree-like model that aids in variable selection when the dependent variable is categorical. This approach was used to identify variables that can be used in the logistic regression model for predicting at-fault driver status. This method examined all variables in the dataset to find the one that gives the best prediction by splitting the data into subgroups. This approach estimated the relative importance of the variables being considered and indicated those variables that should be given priority for inclusion in the logistic regression modeling.

Results from the two techniques were used for the statistical modeling. In addition to the variables identified through these analyses, other variables were considered and tested to finalize the model with the most appropriate set of predictors. For example, if the education variable 'percent below high school graduate' was a descriptor of note in the recursive partitioning analysis, it was considered first in the modeling. However, other education variables (e.g., 'percent with high school graduate' and 'percent with bachelor's degree), which were significantly related to the dependent variables, based on the correlation analysis, were also tested. Each variable from the socioeconomic categories was tested to identify the best representation of that category in predicting at-fault driver crash involvement. Multiple variables from the same category were not used in the same model to avoid multicollinearity. Several models were developed for single- and two-unit crashes using this approach and their parameters were evaluated using the above explained criteria to select the final model.

4. Data Collection and Preparation

4.1 Crash Data

Kentucky crash data, aggregated at the zip code level, were used to examine the characteristics of drivers involved in crashes. Crash data for 2013 to 2016 – collected from the Kentucky State Police (KSP) records – with the 5-digit zip code of the driver residence, were used for the study. About 77 percent of the crashes that occurred during the four-year time period were two-unit crashes, 13.7 were single-unit crashes and the remainder involved three or more vehicles. This research primarily focused on single- and two-unit crashes, which limited the number of drivers involved to a maximum of two. The variables listed in Table 1 were extracted from the KSP database.

Typical information collected by KSP includes crash data (e.g., location, time of crash, environmental conditions), vehicles involved, driver and the occupants involved, and roadway characteristics. Information on crash severity, manner of collision, roadway characteristics, vehicle type, weather condition and lighting conditions were used in this analysis.

Table 1 List of Crash Record Variables

Variable Type	Variable
Crash	Master file number
	Year of collision
	Severity of crash (KABCO)
	Number of people injured
	Number of people killed
	Collision date & time
	Collision day week code
	Intersection crash indicator
	Number of units involved
	County code
	Crash location in lat\long
Vehicle	Unit number
	Unit type code
	Vehicle year
Roadway condition	Total number of lanes
	Roadway character code
	Roadway surface code
	Roadway condition code
	Weather code
	Light condition code
	Land use code
	Function class code
Person	Person number
	Person type code
	Zip code of driver residence
	Age at collision time
	Gender
	Human factors detected

This study did not consider information on passengers and pedestrians because driver fault status was central to the methodology.

Human factors coded for each driver were used to determine their fault status. For each crash, the driver with a human factor code recoded by the police officer was treated as the at-fault driver [53]. In the crash database, multiple human factors are recorded (if any) for drivers. For example, if three human factors are recorded for a driver involved in a two-vehicle crash, there are three entries for that particular Master File Number (MFN, a unique number identifying each crash). After using Python for data processing, human factors recorded to the same driver were aligned to convert the multiple entries to a single entry. Age and gender of the driver were used as the factors to correlate the entries belonging to the same driver. The first human factor recorded was used to determine fault status. For each MFN, the driver with the first human factor coded as "non-detected" was considered to be not at fault, while the driver with a human factor detected was treated as the at-fault driver. Crashes in which a human factor code was recorded for both or neither drivers were eliminated from analysis. These selection criteria avoided identifying multiple at-fault drivers for the same crash in two-unit crashes [53]. In single-unit crashes, only drivers with a human factor coded were included in the dataset, and these drivers were coded as at fault. As single-unit crashes involve just one vehicle, no not-at-fault driver group were identified for these crashes. Therefore, the not-at-fault driver group from the two-unit crashes were included in this dataset to facilitate the quasi-induced exposure technique.

4.1.1 Description on Crash Data

KSP's crash database recorded 725,935 drivers involved in two-unit crashes. Of the 128,422 single-unit crashes, only 80,340 have a human factor recorded. However, there are illogical entries in the information for some drivers, and they were removed. After the data processing and management (see Chapter 3), the final crash database used for analysis had 241,750 two-unit crashes (with 2×241750= 483,500 drivers involved) and 74,641 single-unit crashes.

In single-unit crashes, only drivers with a human factor recorded were included. It was assumed these drivers were at fault. The not-at-fault group from the two-unit crashes were included in the quasi-induced exposure analysis of single-unit crashes to account for driver exposure. The sample size of the not-at-fault group of drivers in the two-unit crashes was almost 3.2 times larger than the at-fault group of single-unit crashes. To avoid disparities in sample size of the not-at-fault group, a random sample equivalent to 75,000 was drawn from the original not-at-fault group. This sample was used as the not-at-fault group of drivers in the single-unit crash data.

Data were processed using the human factor process described here to develop the final dataset for single-unit and two-unit crashes. The final dataset included drivers with ages between 15 and 90 years. To analyze the RAIR of drivers in different age groups, ages were categorized into seven groups: < 20, 20-24, 25-39, 40-64, 65-74, 75-84 and > 85. Table 2 shows the distribution of age groups in the dataset prepared after data processing.

Table 2 Driver Age Distribution, 2013-2016

Two unit								
Fault Status	Age Group							
	<20	20-24	25-39	40-64	65-75	75-84	>84	Total
Not-at-fault	14,801	24,985	72,739	1,03,180	18,885	6,240	920	2,41,750
At-fault	30,582	36,579	68,634	75,568	18,168	9,916	2,303	2,41,750
Single unit								
Fault Status	Age Group							
	<20	20-24	25-39	40-64	65-75	75-84	>84	Total
Not-at-fault	4,600	7,778	22,464	31,740	5,840	1,948	271	74,641
At-fault	11,792	13,219	22,754	21,433	3,453	1,640	350	74,641

4.2 Socioeconomic Data

Socioeconomic and demographic variables were collected from the U.S. Census Bureau [22]. This database has two sets of information significant for this research: People and Housing. The information under the People category includes general information on the population (e.g., total population, race, marital status, age, gender, education, income, employment, poverty status) while the Housing category includes information on households (e.g., home value, number of housing units, household size, household type), in a particular geographical area. The choice of variables was made in response to the findings and suggestions of previous literature and the initial analysis conducted as part of this effort. Table 3 lists the socioeconomic variables chosen for the analysis; they were divided into 6 major categories: Race, Housing, Marital Status, Education, Income, and Other.

Table 3 List of Socioeconomic Variables

Category	Variable
Race	Percent white
	Percent black
	Percent American Indian
	Percent Asian
	Percent other races
Housing	Household units
	Household ownership total
	Owner occupied housing units
	Renter occupied housing units
	Median housing value
Marital Status	Percent now married
	Percent widowed
	Percent divorced
	Percent separated
	Percent never married
Education	Percent less than high school graduate
	Percent high school graduate
	Percent some college or associate degree
	Percent bachelor’s degree or higher
	Percent graduate or professional degree
Income	Median individual income
	Mean individual income
	Household mean income
	Household median income
Other	Employment population ratio
	Percentage rural
	Unemployment rate
	Percent below poverty level
	Total population

The 2016 U.S. Census Bureau population estimate indicated 85 percent of Kentucky’s population was white, and 8.3 percent was black [54]. Adanu et al. [7] found that race is a factor associated with crash occurrence. However, research on the relationship between race and crashes is sparse. This study tested the relationship between race and crash occurrence. To do so, the percent distribution of major races (White, Black, Indian, Asian, and Others) were extracted from population estimates. Other races included the sum of proportion of population belonging to races such as Native Hawaiian and other Pacific Islander alone, Two or More Races, and Hispanic or Latino. Information on all the races were included in this dataset for further investigation.

Housing is a well-established predictor of crash involvement. Housing density is most frequently considered as a surrogate for level of rurality. Noland and Quddus [24] and Hasselberg et al. [9] explained the relationship between housing and unsafe traffic conditions. Lower income people tend to live in rural areas where cost of living and housing are cheaper. These places are less likely to have adequate infrastructure and safe traffic conditions. Therefore, the number of household units and median housing value were considered and included in the analysis as they are viewed as surrogate indicators of rurality. Areas with high rental housing percentages exhibit lower DUI crash rates [10]. It is important to examine the potential effect of different housing ownership levels on crash occurrence. Therefore, data on housing characteristics (rental/owned) were also included in this analysis.

Marital status is expected to have a significant relationship with crash occurrence; however, its association has not been established adequately. Factor et al. [8] provided evidence that separated and widowed drivers are 50 percent more likely to be involved in a crash than married drivers. Stressful life events may inhibit safe decision making, resulting in an increased risk of causing a crash. Information on the proportion of population of now married, previously married (widowed, separated, and divorced) and never married were included in the dataset for further investigation.

Several researchers have investigated the correlation between the educational level of drivers and their involvement in crashes to uncover patterns that can prevent or decrease crashes. People with the lowest levels of educational attainment have the highest mortality rate [9, 10]. Cook et al. [27] discussed a positive relationship between female education attainment and crash involvement. This joint relationship between gender and educational attainment was further tested in this study to examine any possible relationships for Kentucky.

Income is another relevant predictor for crash-related analysis. Personal and household income have been cited as significant explanatory variables; however, personal income is more widely used to represent income [10, 14, 17, 24]. This research considered household and personal incomes to identify which is most representative for Kentucky drivers in relation to crash prediction. Therefore, different mathematical representations (mean and median) of both individual and household income were extracted from the U.S. Census Bureau database.

Other well-established predictors of crashes include employment rate, poverty level, and rurality. These variables are correlated with income, housing, and education. Their interdependency was also explored in this analysis.

The information on the above discussed variables were obtained from a five-year estimate of 2016 U.S. Census Bureau data at the zip code level. The data for the demographic and socioeconomic descriptors were joined at the zip code level and then merged with the data of crash-related variables matching the residence zip code of the driver. Python tools were used to prepare the final dataset. The variables in the final dataset were tested with the dependent variable (at-fault status) to understand their relationships with each other. Variables correlated with the dependent variable in the initial correlation analysis were retained for the final regression modeling. Multiple variables from same socioeconomic category were not used in the same model to avoid multicollinearity. For example, percent white and percent non-white are complementary, as it would produce ambiguous results if included in the same regression model.

4.3 Conviction Data

The literature review concluded that drivers who have driving records with convictions, crashes, or both are high-risk drivers. With driver crash history being unavailable, this research could not delve further into the effect of previous crash involvement on the fault status in a future crash. Instead, convictions — another representation of a historical driver performance — were considered. The initial idea was to combine crash data and conviction data at the driver level. However, this could not be achieved because there was no common element connecting the two databases. The crash database lacked driver license number, and it is the only factor that could be used to merge the two datasets at the driver level. Therefore, convictions were used at the zip code level in the form of average yearly convictions.

Conviction data for 2012 to 2018 were obtained from the Kentucky Driver License database. There were 1,196,762 conviction recorded for 612,295 drivers during the seven-year period. Each driver license number, license type, date and year of conviction, conviction type, zip code of the driver's residence, date of birth, and gender of the driver were extracted from the database. Multiple convictions were recorded for many drivers, and the maximum number of convictions entered for the same driver between 2012-2018 was 37. There were 113 different conviction types, which are related to DUI, speeding, reckless driving, ignorance of law, and failure to obey a court summon. For this analysis, convictions were categorized into six groups:

1. DUI: Drunk driving is generally charged as DUI. However, driving under the influence of an illicit substance or certain prescription medicines is also considered DUI. Over the seven-year period, the state recorded an average of 15,172 DUIs each year.
2. Speeding: One of the most common moving violations, this category includes all conviction types recorded for aggravated speeding. On average, 35,417 speeding convictions were recorded annually in Kentucky.
3. Driver behavior: This category includes moving violations related to driver behavior. Improper driving, driving on the wrong side of road, and texting while driving are some of the convictions under this category. About 16,401 such convictions each year were observed during the study period.
4. Negligence to law: This category includes other moving violations such as vehicle not under control, driving with a suspended license, and failure to dim headlights. Kentucky averaged 9,020 of these convictions each year during the study period.
5. Legal: Charges related to the violation of court or other legal proceedings. Examples are failure to answer court summons, license misrepresentation, and ignition interlock violation. Over the study period, the annual average was 72,879 legal charges.
6. Other: This category includes all other non-moving charges such as the refusal of chemical test, gasoline theft, and theft of motor vehicle/parts. Across the study period, 22,075 of these charges were recorded annually.

Non-moving convictions (legal and other) are the most frequent violations other than speeding. However, they are not considered to be closely associated with traffic safety and hence, not included in the analysis for the current study. The average convictions per year was calculated for every zip code; this number was normalized to 1,000 drivers for each zip code.

4.4 Data Processing

Python is widely used in data processing and management. The whole data manipulation procedure was carried out in Jupyter notebook which is an open source web application which allows the creation of live Python codes. The step-by-step procedure followed to manipulate the crash and census data is explained below.

First, crash data were converted into a useful format. KSP crash data contained 932,535 driver records. This dataset had 572,152 MFNs, and each MFN represents a unique crash. The initial step was to clean up the data by removing invalid entries which were probably due to human errors made while recording the information. For example, "Gender Code", which defines the driver's gender, had several invalid entries such as "+". This symbol has no definition in the crash data dictionary and therefore these were eliminated from further processing. Also, there were entries where the gender was unknown or missing. These cases were also eliminated from the database. In total, 1,389 crashes were removed. Age was another attribute with similar anomalies. There were 3,435 drivers whose age was under 16 years or over 90 years. These entries were assumed to be in error and thus not considered in the next step. Also, several driver residence zip codes were wrongly entered. The zip codes in Kentucky were obtained in the form of a shapefile from the Kentucky Geological Survey, maintained by the University of Kentucky [55]. There are 746 zip codes in Kentucky according to the shapefile. The crash database had 57,620 entries with zip codes not listed in the KGS database; they were eliminated from further processing.

Driver fault status was determined next. The fault status was decided based on the human factor code. For each crash, the driver with the first human factor coded as "non-detected" was considered to be not at fault, while the driver with a human factor detected was treated as the at-fault driver. Crashes in which a human factor code was recorded for both or neither drivers were eliminated from analysis. These selection criteria avoid multiple at-fault drivers for the same crash in two-unit crashes. In single-unit crashes, there is only one driver involved and hence that driver is supposed to be the one causing a crash.

Next, single-unit and two-unit crashes were extracted into two different files. There were 119,517 single-unit crashes recorded from 2013 to 2016. Single-unit crashes occur when a driver collides the vehicle with a non-moving object or an animal. There were only 74,691 crashes with a human factor recorded for the involved driver. Other uncontrollable factors, such as unfavorable weather conditions or an animal, could be the reason no human factor was recorded. Such crash entries were eliminated from next steps. About 24 MFNs were repeated in the dataset, probably due to double entries. They were also removed to avoid duplication.

The number of drivers involved in two-unit crashes during the study period was 679,106. Not every MFN had an at-fault and not-at-fault driver pair. Only 241,881 two-unit crashes had both at-fault and not-at-fault drivers. Therefore, only these MFNs were included for further processing.

Socioeconomic and demographic variables were obtained from different files and the first step was to combine them into a useful format. There were several attributes in each file not relevant here. For example, the data table on "Household Ownership" contained a column for margin of error estimate on the total households in each zip code. Such attributes were removed from each data table to ease the process of joining files. Using the "merge" command in Python, each file containing demographic and socioeconomic descriptors was joined to one another at the zip code level. According to the U.S. Census Bureau, 746 zip codes in Kentucky are fully or partially within the state boundary. However, Census data lack all of the socioeconomic variables for every Kentucky zip code. Exempted zip codes seemed to be tiny areas with probably very few or no people living in them. On combining each file at the zip code level, one final file was created that contained entries representing each zip code. Each column in the file represented the socioeconomic and demographic factors chosen for the analysis here.

After the data preparation, the next step was to combine crash data, census data, and conviction data. Using the "merge" command in Python, the files were joined by matching the zip code, which is a common field in both datasets. Finally, two files were prepared, one for single-unit crashes and the other for two-unit crashes. The final dataset had 74,641 single-unit and 241,750 two-unit crashes.

Variables in the final dataset were evaluated in relation to the dependent variable (at-fault status of the driver) to determine how strongly they correlated with one another. Variables that indicated correlation with the dependent variable in the initial correlation analysis were then tested using recursive partitioning analysis, followed by selection method. Interactions were also identified, and the logistic regression method used to develop the final models for single- and two-unit crash occurrence.

5. Spatial Analysis

Past research has uncovered strong relationships between crash occurrence and factors like driver age and gender. Younger (under 25) and older (over 65) drivers are more likely to be at fault in crashes than middle-aged drivers, while several studies have demonstrated that male drivers have a higher propensity for crash involvement than female drivers. Socioeconomic factors associated with crash risk include income, education level, poverty percentage, employment, and rurality of an area. Education and income tend to be negatively correlated with crash occurrence, while a positive correlation between poverty and rurality and crash risk is common. Kentucky is a predominantly rural state, with roughly 50 percent of its counties located in the federally designated Appalachian Region. Poverty, rurality, and unemployment rates are higher in Appalachia than the remainder of the state. Limited research has focused on disparities in motor vehicle crashes within the Appalachian region. The chapter presents the results of spatial analysis that examined whether trends related to age, gender, and income and at-fault status found in previous research hold for Kentucky. Along with investigating trends across the state, analysis also looked at whether at-risk propensity was higher or lower for drivers in Appalachia.

A quasi-induced exposure technique was used to assess the relative risk of drivers being at fault in a crash. RAIRs were calculated based on age, gender, and residence zip code. County-level heat maps were then generated using these ratios to visually represent the spatial variability in key trends. The following section elaborates on the methods used to generate heat maps followed by the spatial analysis results.

5.1. RAIR Calculation

Using a series of Python scripts, RAIRs for each age and gender category were calculated for zip codes (see Section 3.4.1 for a methodological explanation). Drivers were grouped into 7 age categories (< 20, 20-24, 25-39, 40-64, 65-74, 75-84 and > 85) for detailed statistical assessments. Conventional categories of young (< 25 years), middle-aged (25-64 years), and older drivers (> 64 years) were then used. To illustrate this approach, Table 4 shows the distribution of at-fault and not-at-fault drivers in the three age categories for ZIP Code 40003. Out of 100 not-at-fault drivers, 15 are in the < 25 age category. Therefore, the probability of a not-at-fault driver being in the young age group is: 15/100 = 0.15. The probability of an at-fault driver being young is 36/108 = 0.333. Therefore, the RAIR of young drivers in zip code 40003 is: 0.333/0.15 = 2.222.

Table 4 Distribution of Number of Drivers in Zip Code 40003

Status	Age Group			
	<25	25-64	>64	Total
Not At-Fault	15	74	11	100
At-Fault	36	58	14	108
RAIR	2.222	0.726	1.178	

RAIRs were horizontally arranged to the zip code level and saved as a csv file for use in the next step.

5.2. Aggregating RAIRs at County Level

Ratios were aggregated at the county level to produce heat maps. Two issues had to be addressed to generate the maps. First, the absence of age and gender distributions for people and drivers at the zip code level were required because RAIRs are calculated for each combination of variables. To develop county-level estimates, RAIRs must be weighted based on observed population distributions. Population data for different age categories are available at the county level in the U.S. Census Bureau's American Census Survey (ACS) database [22]. However, driver population is not available in a direct format. As such, it was assumed that all people older than 16 have a driver's license. The population of residents over 16 years of age was summed for each county to estimate driver numbers. A second assumption was that the population of each age and gender group follows a similar distribution throughout the county. To estimate the number of drivers in each group at the zip codes level, a distribution based on the area of the zip codes within the county was used. Because several zip codes are split between counties, ratios could not be directly aggregated at the county level. To estimate RAIRs for each category at the county level, a series of geospatial processes were carried out in ArcMap. The process used for proportionally allocating the RAIR among neighboring counties is thoroughly explained below.

The second issue that was addressed is that there are zip codes with similar RAIRs but different proportions of residents in each category. For example, zip codes 41301 and 40356 have similar RAIRs for all age and gender categories. However, their total driver populations differ quite significantly (Table 5). Driver population density in zip code 40356 is 9.6 times higher than in 41301. When aggregating ratios, it is important to also consider this factor. This was addressed through weighing the RAIRs by zip code population.

Table 5 RAIR of Example ZIP Codes

Zip code	RAIRs					Total Driver population	Area (sq mi)
	<25	25-64	>64	Male	Female		
41301	1.612	0.910	0.967	1.018	0.978	4,753	196.15
40356	1.618	0.847	1.083	1.049	0.949	33,359	142.48

5.2.1 Intersect Area of Zip Codes

First, the areas of zip codes split between counties was computed. Shapefiles of Kentucky at the zip code and county levels were obtained from web resources provided by the University of Kentucky and the U.S. Census Bureau, respectively [55, 56]. The csv file (see Section 5.1**Error! Reference source not found.**) was then joined to the shapefile of the zip codes. ArcMap's Intersect tool — which computes the geometric intersection between the input polygon features — was then used to intersect the county shapefile and new zip code shapefile. Next, the area of a zip code that coincides with the counties was extracted. The tool also calculates the proportion of area split between counties. As an example, Fayette County has 20 zip codes partially or completely coinciding with its border (Figure 2).

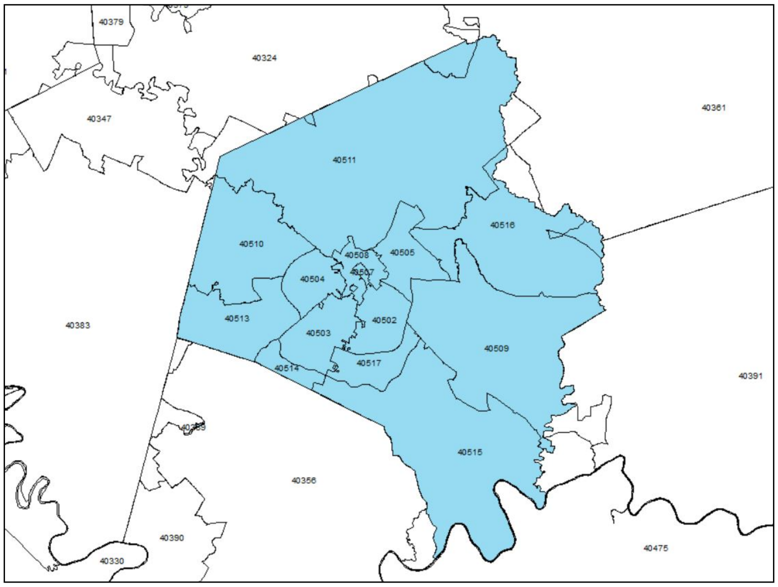


Figure 2 Intersection of Zip Codes in Fayette County

A portion of zip code 40324 is located in Fayette County, but it also stretches across Woodford, Harrison, Scott and Bourbon Counties. Only about 1.80 percent of zip code 40324 lies within Fayette County. Table 6 lists Fayette County zip codes and what percentage of their areas falls in the county.

Table 6 Zip Codes in Fayette County

County	Area of County (sq mi)	Zip Code	Total Area of Zip Code (sq mi)	Area in County (sq mi)	Percent of Area in County
Fayette	285.149	40324	158.268	2.850	1.80
		40347	35.825	0.006	0.02
		40356	142.483	0.011	0.01
		40361	266.262	1.739	0.65
		40383	155.501	0.013	0.01
		40391	240.249	0.036	0.01
		40502	7.427	7.427	100.00
		40503	9.047	9.047	100.00
		40504	6.256	6.256	100.00
		40505	7.837	7.837	100.00
		40507	0.407	0.407	100.00
		40508	4.004	4.004	100.00
		40509	46.300	42.650	92.12
		40510	21.619	21.610	99.96
		40511	87.739	80.791	92.08
		40513	14.421	14.404	99.88
		40514	3.000	2.989	99.62
		40515	56.072	47.811	85.27
		40516	32.362	29.097	89.91
		40517	6.165	6.165	100.00

5.2.2 Population at Zip Code Level

Next, the proportion of people by age group and gender was calculated for each zip code. First, the population of each county in available age and gender groups was collected from the ACS. Then, these population estimates were divided among zip codes by weighting the population based on county area (see Section 5.2.1). For example, Fayette County has 110,593 drivers in the < 25 age group. The total driver population was weighted by county area to estimate the driver population in each zip code for each age and gender category. Table 7 is an example calculation for Fayette County. For example, the < 25 driver population in zip code 40324 is: $110{,}593 \times \frac{2.850}{285.149} = 1{,}105$.

Table 7 Calculated Population of < 25 Drivers in Fayette County

County	Total <25 Population	Zip Code	Area in County (sq mi)	<25 Population
Fayette	110,593	40324	2.850	1,105
		40347	0.006	2
		40356	0.011	4
		40361	1.739	675
		40383	0.013	5
		40391	0.036	14
		40502	7.427	2,880
		40503	9.047	3,509
		40504	6.256	2,426
		40505	7.837	3,039
		40507	0.407	158
		40508	4.004	1,553
		40509	42.650	16,541
		40510	21.610	8,381
		40511	80.791	31,334
		40513	14.404	5,586
		40514	2.989	1,159
		40515	47.811	18,543
		40516	29.097	11,285
		40517	6.165	2,391

5.2.3 Weighted RAIR

Next, the RAIR for drivers in each category was calculated for every county adopting a weighted RAIR approach. To calculate a county's RAIR, the ratios of all zip codes in the county were weighted to the driver population in that zip code (Equation 10):

$$\text{Weighted RAIR of X county} = \frac{\sum_1^n RAIR_i P_i}{P_t} \qquad (10)$$

Where n = the number of zip codes in the county, $RAIR_i$ = RAIR of any category at the zip code i, P_i = population of the category in zip code i in the county, and P_t = total population of the county = $\sum Pi$. Table 8 illustrates the calculation of RAIR of < 25 drivers in Fayette County.

Table 8 Weighed Probability of Fayette County

County	Zip code	Population (P_i)	$RAIR_i$	$P_i \times RAIR_i$	Weighted RAIR
Fayette	40324	1,105	1.638	1,809.58	1.807
	40347	2	3.020	6.04	
	40356	4	1.618	6.47	
	40361	675	1.491	1,006.53	
	40383	5	1.747	8.73	
	40391	14	1.563	21.88	
	40502	2,880	1.555	4,477.95	
	40503	3,509	1.612	5,655.38	
	40504	2,426	1.393	3,380.40	
	40505	3,039	1.791	5,443.95	
	40507	158	1.517	239.73	
	40508	1,553	1.423	2,209.28	
	40509	16,541	1.715	28,363.12	
	40510	8,381	2.000	16,762.00	
	40511	31,334	1.826	57,214.03	
	40513	5,586	2.034	11,359.99	
	40514	1,159	1.736	2,012.02	
	40515	18,543	1.880	34,853.07	
	40516	11,285	1.866	21,052.69	
	40517	2,391	1.634	3,906.40	
	Total	110,593		199,789.24	

5.3 Heat Maps

Weighted RAIRs were used to generate heat maps for each county. In the maps which follow, counties are shaded to represent the crash involvement risk of drivers in various groups. As it is important to identify if drivers reside in areas where poverty or income are issues, the maps also display household income and indicate counties in Appalachia. The latter are denoted with hatching and using bold shading on the county borders. Median household income is shown on the maps as it is a socioeconomic factor widely recognized as influencing crash occurrences. It is correlated with other socioeconomic variables such as poverty and employment rate. Prior research has shown that household income is a better predictor of income [14, 17] than other factors because it better determines a family's overall economic status. Maps were developed for both single- and two-unit crashes.

5.3.1 Two-Unit Crashes

Figures 3–5 are heat maps for each age group, and Figures 6 and 7 are heat maps for each gender. Dark shading indicates high risk and light low risk. Section 3.3 describes how to interpret RAIRs.

For drivers under 25 (Figure 3), relative risk varies from 0.676 to 2.328. Statewide, RAIRs among young drivers are higher, which means that when a young driver is involved in a crash, they are more likely to be the at-fault driver than the not-at-fault one. Collectively, these findings speak to how the characteristics of young drivers — inexperience, lack of skill, and risk-taking behaviors — place them at greater risk. No strong trends were observed among drivers in Appalachia. This exemplifies the risk-taking behavior of the young drivers regardless of socioeconomic conditions.

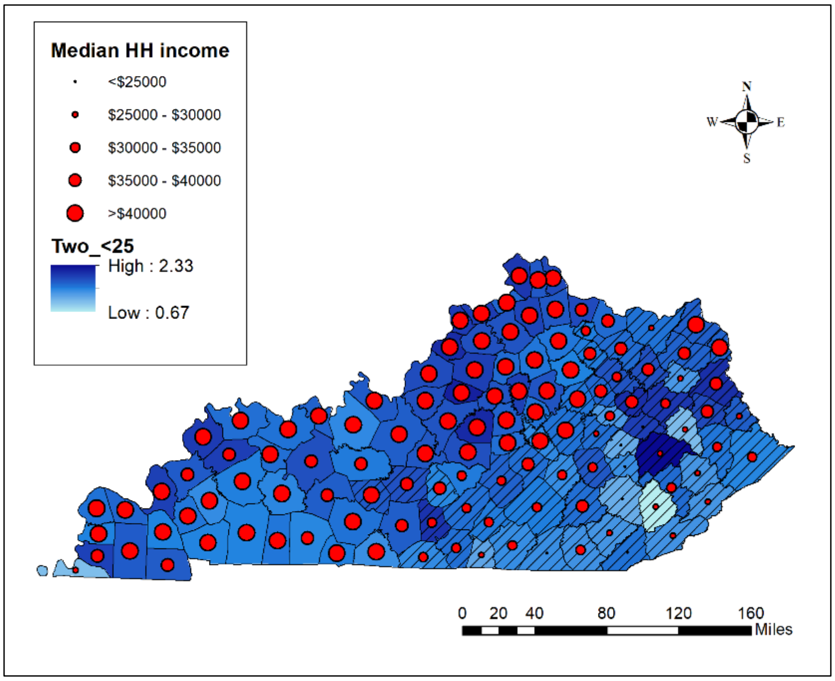


Figure 3 Heat Map for Young Drivers (< 25 years), Two-Unit Crashes

For drivers between 25 and 65 (Figure 4), RAIRs are between 0.689 and 1.034 — lower than the range for young drivers. Many high-income counties exhibit lower risk rates for middle-aged drivers than other age groups. But there is no evident regional pattern. Counties with higher RAIRs are mostly low-income areas in Appalachia. These ratios indicate the at-fault and not-at-fault probability of drivers are almost equal, demonstrating higher risk compared to counties elsewhere in the state. Overall, drivers in this age group are less likely to cause a crash than young drivers. This could be attributed to their better judgment and decision making, which are gained through experience.

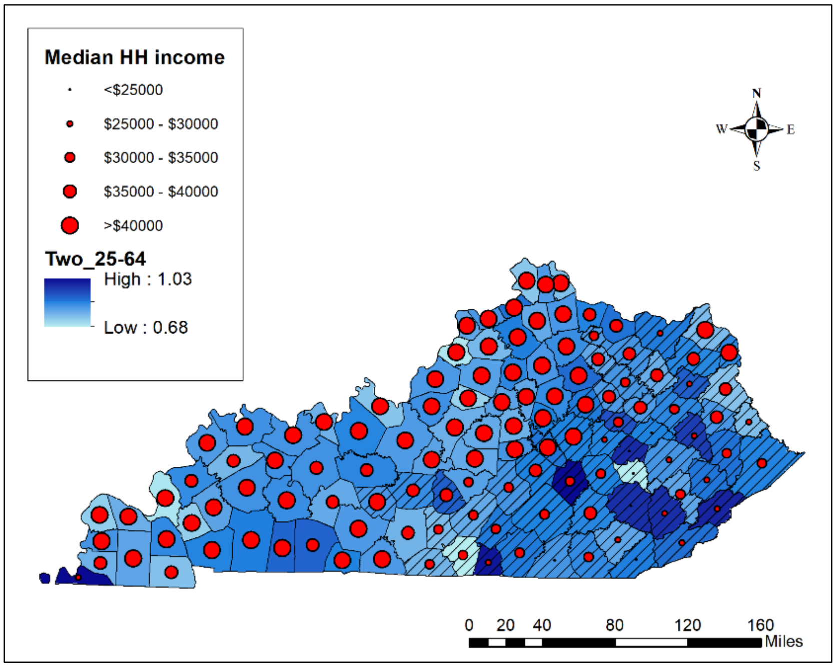


Figure 4 Heat Map for Middle Aged Drivers (25-64 years), Two-Unit Crashes

RAIRs for drivers over 64 range between 0.376 and 2.052. Across Kentucky, older drivers are more likely than young or middle-aged drivers to be at fault than not at fault when involved in a crash. This high risk could result from these drivers suffering from a loss of vision and/or cognitive ability [57, 58]. There are fewer old drivers in the dataset, which may impact exposure to crash occurrence, thus contributing to their higher risk ratio.

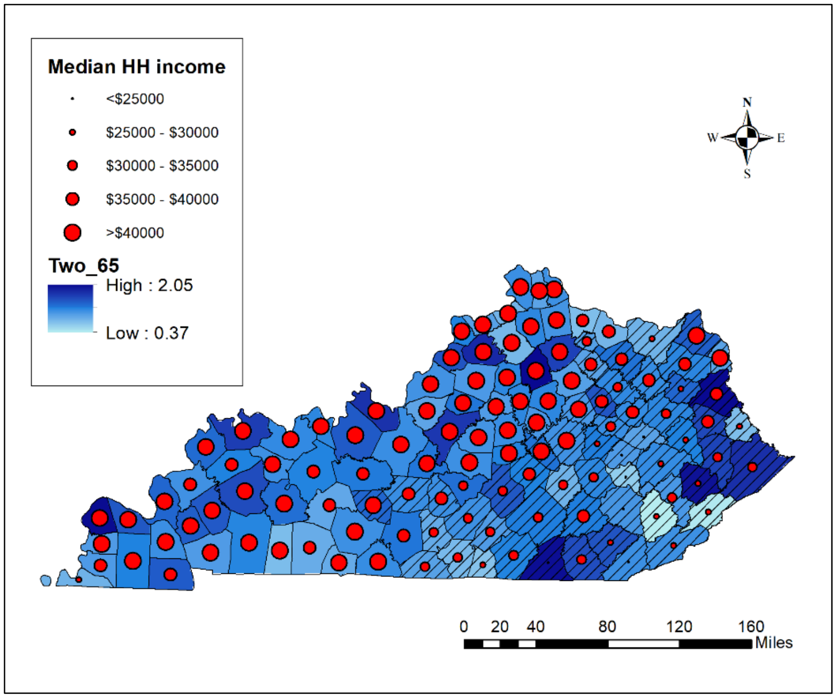


Figure 5 Heat Map for Old Drivers (>64 years), Two-Unit Crashes

For male drivers, RAIRs range from 0.89 to 1.30 (Figure 6). In most counties, RAIRs are close to 1.0, indicating high risk. While counties with the highest RAIRs are found in Appalachia, overall there are no strong regional trends. RAIRs for female drives are comparatively lower, with values ranging from 0.66 to 1.22 (Figure 7). Lower risk rates among females is likely due to male drivers exhibiting more aggressive driving behaviors and their willingness to take more risks [59].

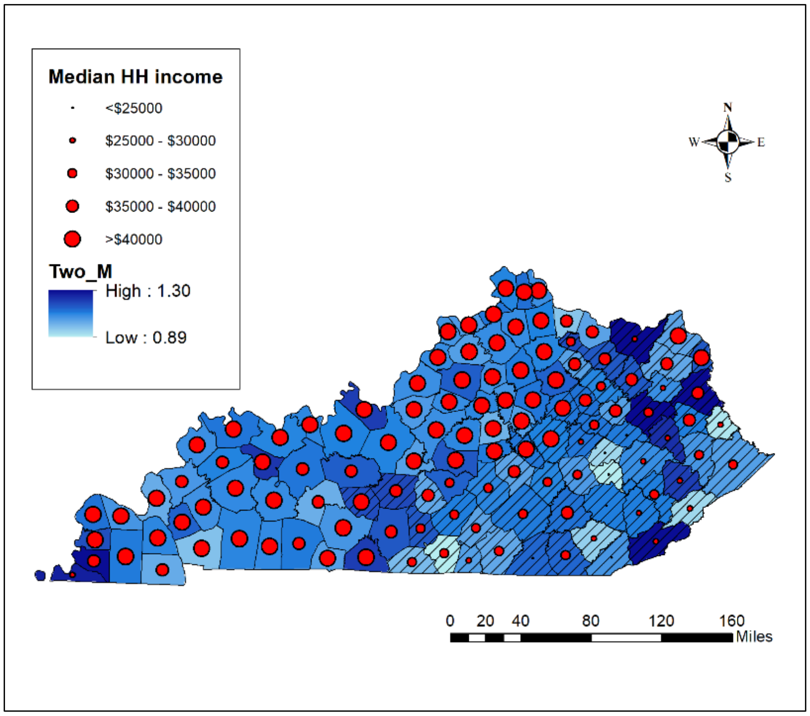


Figure 6 Heat Map for Male Drivers, Two-Unit Crashes

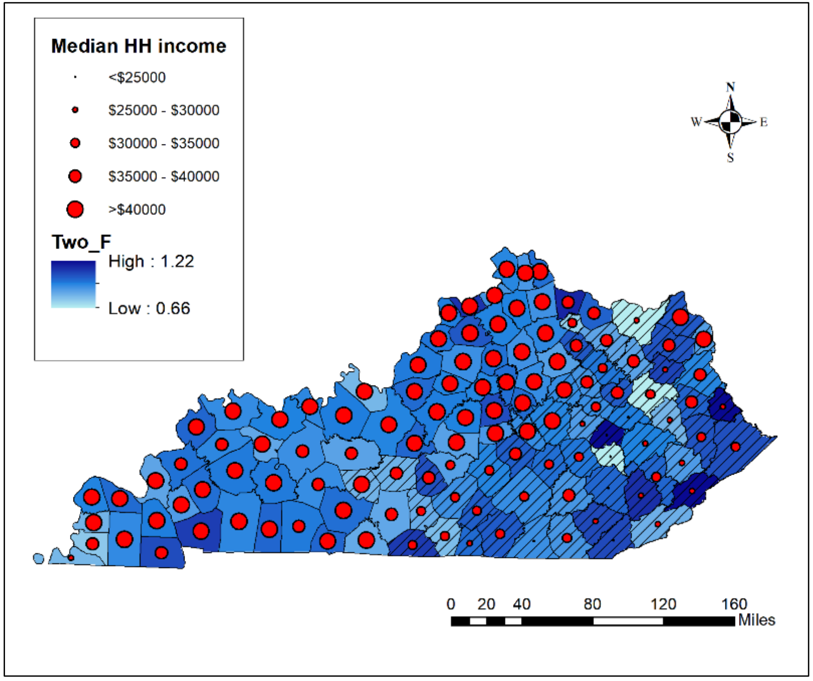


Figure 7 Heat Map for Female Drivers, Two-Unit Crashes

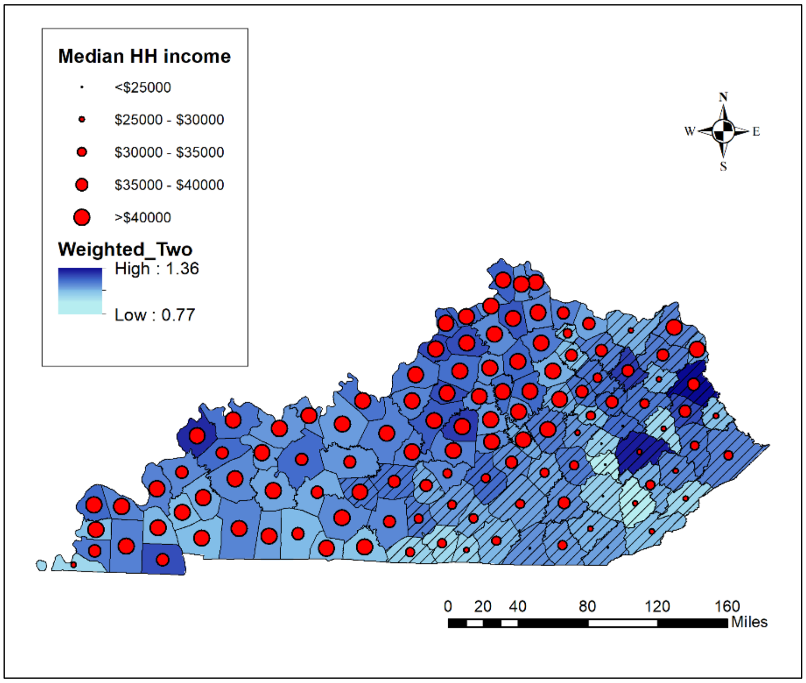


Figure 8 Weighted RAIR, Two-Unit Crashes

5.3.2 Single-Unit Crashes

Heat maps were also developed for single-unit crashes. Among young drivers, RAIRs range between 0.73 and 5.33, a wider spread than observed for two-unit crashes. In most counties, values are higher than two-unit crashes, although some areas of Appalachia have lower ratios. One explanation for this trend is that the datasets have a relatively small number of young drivers in these counties. Nonetheless, young drivers have a greater propensity to be at fault in single-unit crashes than two-unit crashes. These drivers' inexperience and lack of judgment may explain this phenomenon.

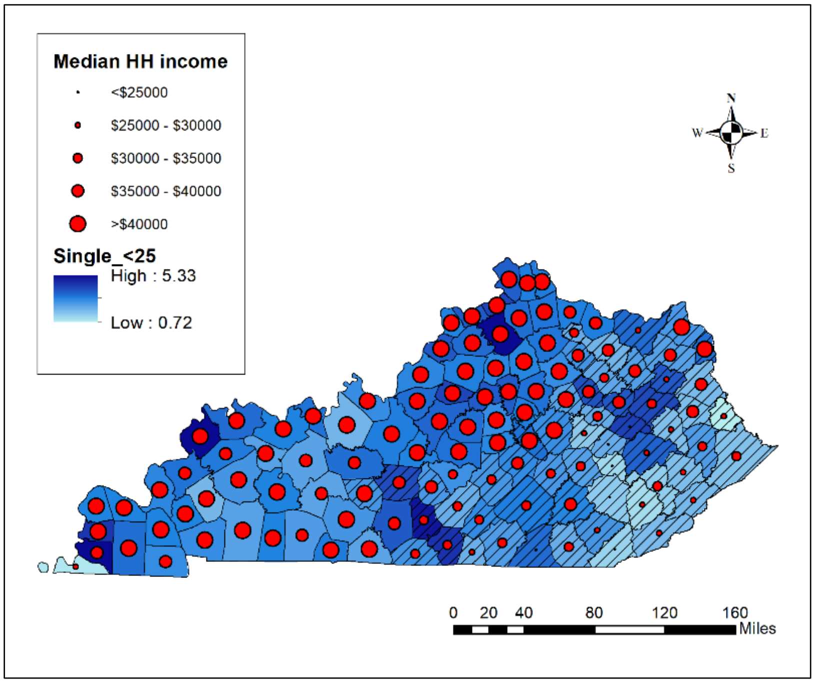


Figure 9 Heat Map for Young Drivers (<25 years), Single-Unit Crashes

Among middle-aged drivers, RAIR values for single-unit crashes are similar to those for two-unit crashes as they range between 0.657 and 1.07 (Figure 10). In Appalachia, risk levels are generally slightly higher than the rest of the state. Many counties with high household median income have lower risk for young drivers compared to middle-aged drivers.

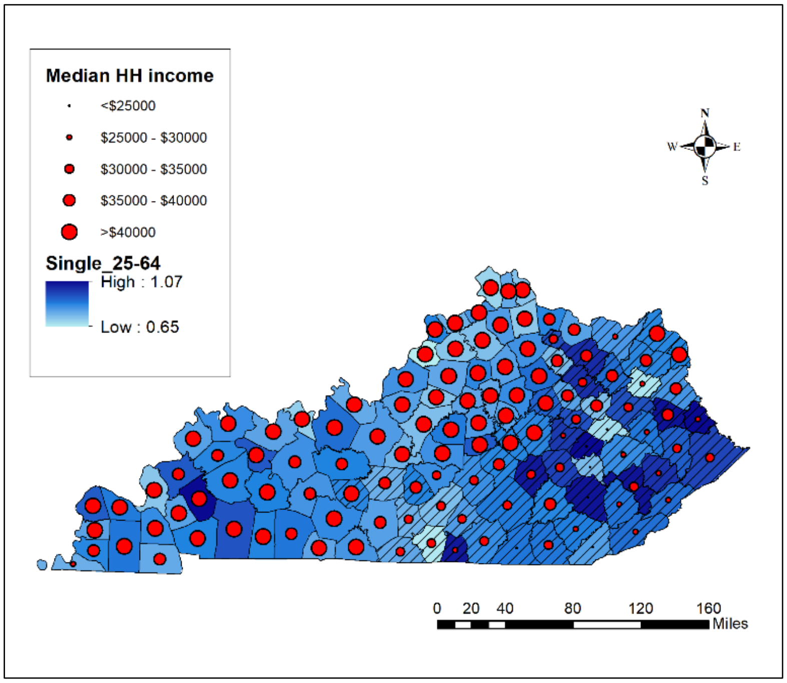


Figure 10 Heat Map for Middle Aged Drivers (25-64 years), Single-Unit Crashes

Among older drivers, trends for single-unit crashes diverge from those for two-unit crashes. RAIR values are from 0.124 to 1.43, with most counties falling into the lower range. While older drivers are less likely to cause a single-unit crash than drivers in the other age groups, they have a higher risk overall. This finding may result from the dataset having a relatively small number of older drivers. The contribution of older drivers to single-unit crash occurrence is discussed later in this report. No regional or socioeconomic trends are observed for this category.

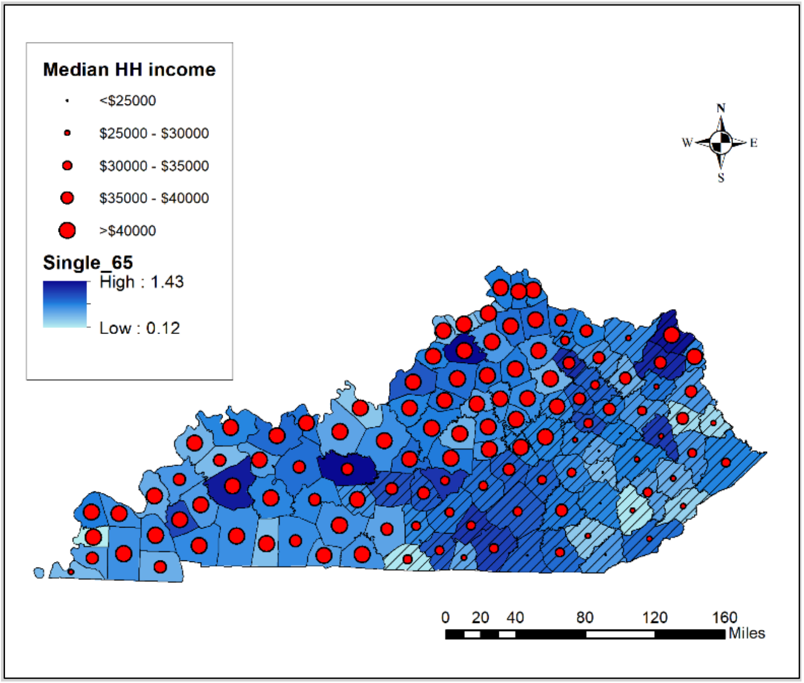


Figure 11 Heat Map for Old Drivers (>64 years), Single-Unit Crashes

Figures 12 and 13 shift the focus to the role of gender in single-unit crash risk. RAIRs for male drivers range from 0.76 and 1.68 (Figure 11) while for females the range is between 0.43 and 1.04 (Figure 12). These maps demonstrate that the likelihood of male drivers causing a single-unit crash is much higher than their propensity to be at fault in a two-unit crash (where the range was 0.89 to 1.30). Female drivers once again have lower RAIRs than males, demonstrating they are less likely to be the cause of a single-unit crash.

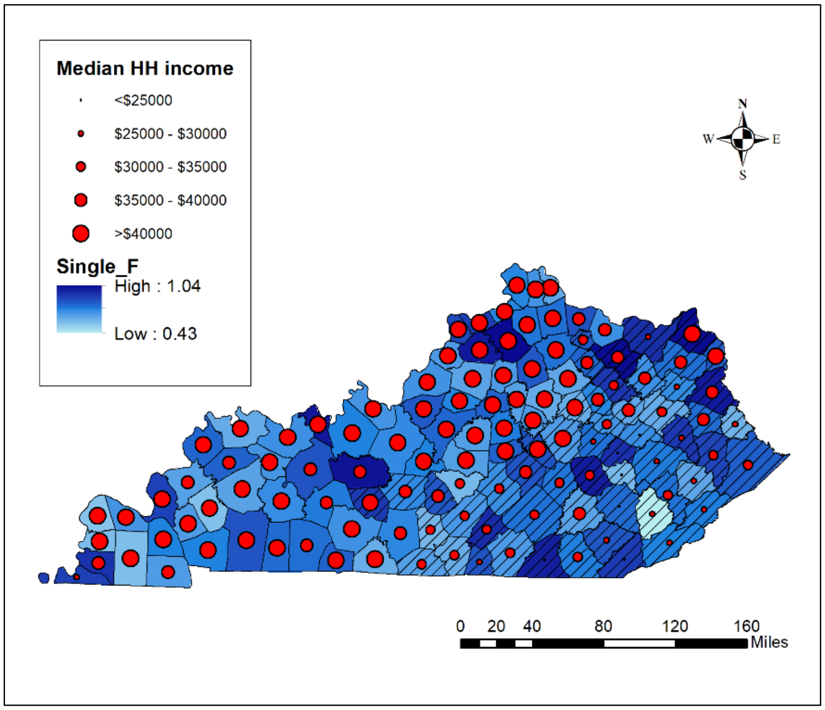


Figure 12 Heat Map for Male Drivers, Single-Unit Crashes

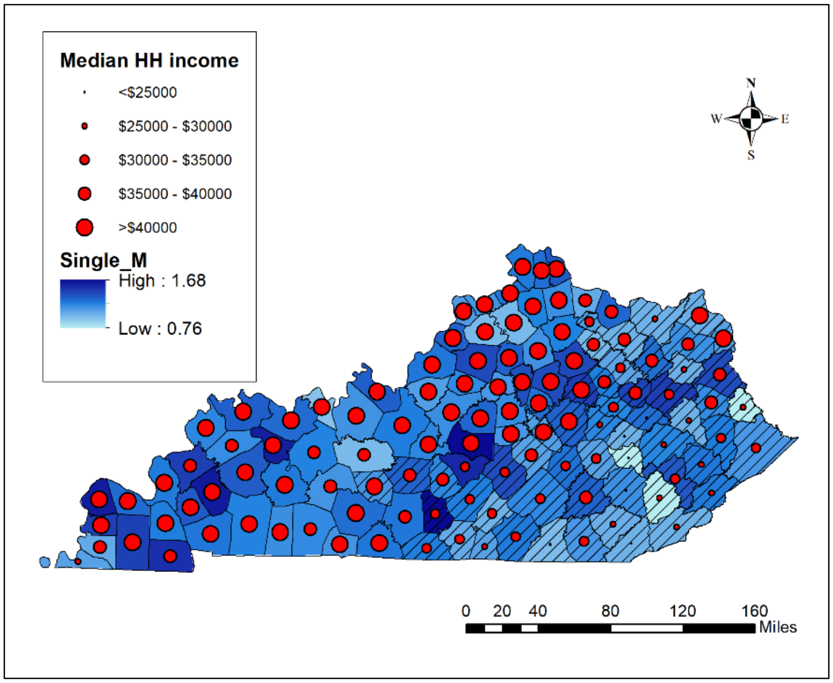


Figure 13 Heat Map for Female Drivers, Single-Unit Crashes

Weighted RAIRs are calculated for single-unit crashes as well and the heat map developed is shown in Figure 14. Few high-income counties seem to have higher risk rate while few low-income Appalachian counties are observed to have lower risk ratio in the state. Overall, no evident regional pattern is observed.

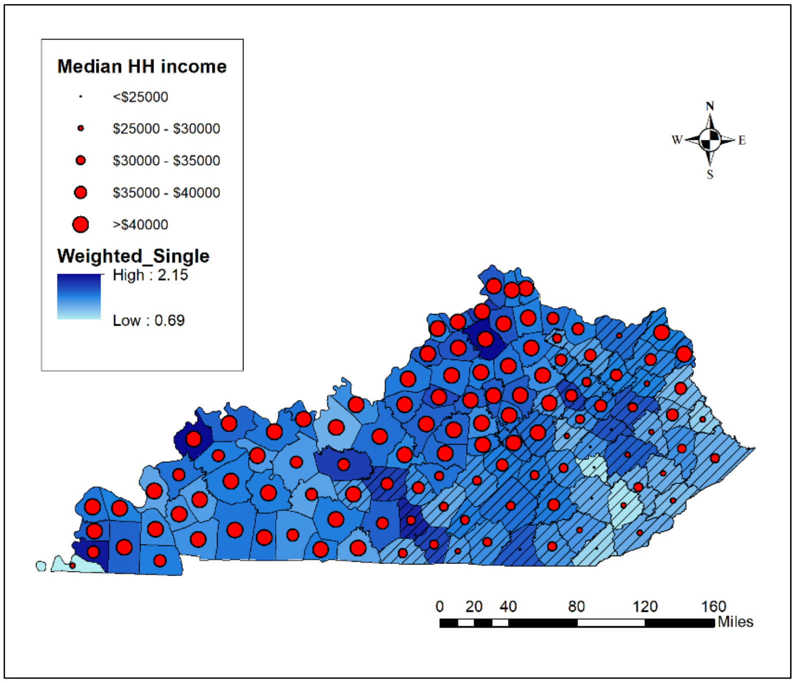


Figure 14 Weighted RAIR, Single-Unit Crashes

5.4 Application

The findings of the spatial analysis can be used to identify high risk counties that can be targeted for safety programs. The Safety Circuit Rider (SCR) program of the Federal Highway Administration (FHWA) is a safety program that provides safety-related support to agencies responsible for local road safety with a goal of reducing the frequency and severity of roadway crashes [60]. Kentucky implements the SCR program through the identification of six high risk counties annually and completion of a detailed crash data analysis and road safety audits on the county public roadways [61]. The goal is to develop a set of countermeasures to reduce crashes at the identified high-risk areas.

Using the weighted RAIR, the top six high risk counties are identified (Table 9). Programs that could address driver performance and crash involvement for drivers in these counties can be developed through the Kentucky Circuit Rider program or other efforts. The table displays high risk counties for both two-unit and single-unit crashes, identified based on its driver's propensity to cause a crash. Drivers in Union county are at high risk in causing both two-unit and single-unit crashes.

Table 9 Top 10 High Risk Counties

	Two-unit Crashes		Single-unit Crashes	
Rank	Name	Weighted RAIR	Name	Weighted RAIR
1	Lawrence	1.36	Owen	2.15
2	Breathitt	1.32	Union	1.77
3	Union	1.31	Hickman	1.56
4	Rowan	1.28	Metcalfe	1.52
5	Washington	1.28	Gallatin	1.44
6	Oldham	1.27	Grayson	1.42

5.5 Conclusions

Spatial analysis failed to uncover strong regional patterns in RAIR values. This finding is consistent with previous research on the relationships between driving behavior and factors such as age and gender (see Chapter 2). It is probable that socioeconomic trends were not detected by spatial analysis due to variables excluded from consideration, such as education and rurality, as well as interactions between them. The next chapter presents the results of regression analysis, which enabled a more robust statistical evaluation of these factors.

The spatial analysis indicated that median household income does not play a predominant role in single- and two-unit crash occurrence. The data did not show any correlation between income and the RAIRs for any of the age groups and genders examined. Young drivers in two-unit crashes are more prevalent in populated areas, while middle-aged drivers causing two-unit crashes are more prevalent in lower income counties in the Appalachian region. However, older drivers causing two-unit crashes have higher crash involvement statewide. With respect to gender, female drivers are less likely to be involved in a crash statewide than males; this was true for both single-unit and two-unit crashes. This trend is also consistent with prior research findings. Young drivers causing single-unit crashes are more prevalent in higher income counties, while middle-aged drivers causing single-unit crashes are more prevalent in lower income counties. Yet, older drivers are less likely to be involved in single-unit crashes statewide. The latter findings are in agreement with prior research and support the notion that older drivers are more likely to be involved in two-unit crashes than in single-unit crashes.

The weighted RAIR developed in this part of the analysis could be useful to the Kentucky Safety Circuit Rider Program and aid in the identification of target counties for establishing potential countermeasures for safety improvements. The RAIRs provide the opportunity to address driver-related issues and it could be an additional element that the Safety Circuit Rider Program could consider when identifying and selecting candidate counties for studying safety conditions and determining improvements. It is therefore conceivable that in addition to roadway improvement, driver training programs for specific target groups could be developed to complement the Safety Circuit Rider Program for a more complete approach to highway safety.

6. Statistical Modeling

This research primarily focused on two-unit and single-unit crashes. The objective of this effort was to identify the socioeconomic variables most useful for predicting a driver's fault status when involved in a crash. In addition to the descriptors (income, education, employment) most often discussed by previous research, many other variables (e.g., race, housing characteristics, marital status) were included in analysis. Several socioeconomic variables discussed in the literature review were collected from the U.S. Census Bureau, and the final datasets were prepared, as explained in Chapter 4. Several variable selection tests were conducted to make the appropriate choice of variables for the modeling.

As noted in the modeling section, correlation analysis was first used to examine the significance of each socioeconomic variable in predicting the dependent variable. Statistically significant explanatory variables were narrowed down to establish a starting set of variables for further selection. Recursive partitioning was used then to clarify the association between the potential predictors and the dependent variable. This step helped illustrate the importance of the variables that should be considered in the modeling. Next, the strongest variables identified in the previous tests were added or removed one by one and the model with the best estimates chosen. Possible interaction terms were also tested in this step. Based on these results, binary logistic regression models predicting crash occurrence were developed for single-unit and two-unit datasets. These steps are explained below in detail.

6.1 Two-Unit Crashes

6.1.1 Variable Selection

Correlation Test

Correlation matrices were developed to identify variables associated with at-fault status. Point biserial correlation coefficients represented each predictor's association with the dependent variable. Variables statistically significant at 95 percent are marked in green in Table 8.

Correlation analysis identified statistically significant variables. p-values less than 0.05 were considered to be significantly correlated with the at-fault status at the 95 percent level in a two-tailed test. The arithmetic sign of a Pearson coefficient indicates the nature of relationship between a socioeconomic variable and the indicator of crash occurrence. For example, income variables were positively correlated with at-fault status, meaning that as driver income increases, the likelihood of being at fault in a crash increase. An explanation of the results shown in Table 10 is provided below.

Table 10 Correlation Test Results for Two-Unit Crashes

Category	Variable	Correlation Coefficient	p-value
Race	Percent white (WH)	0.001	0.663
	Percent black (BL)	-0.001	0.298
	Percent American Indian (AI)	0.008	0.000
	Percent Asian (AS)	-0.004	0.004
	Percent other races (OR)	0.007	0.000
Housing	Household units (HH)	-0.002	0.285
	Household ownership total (HHO)	-0.002	0.178
	Owner occupied housing units (OHU)	-0.006	0.000
	Renter occupied housing units (RHU)	0.004	0.013
	Median housing value (HVL)	-0.010	0.000
Marital Status	Percent now married (MRD)	-0.007	0.000
	Percent widowed (WID)	0.007	0.000
	Percent divorced (DIV)	0.007	0.000
	Percent separated (SEP)	0.003	0.071
	Percent never married (NMD)	0.004	0.005
Education	Percent less than high school graduate (LHS)	0.008	0.000
	Percent high school graduate (HS)	0.005	0.001
	Percent some college/associate degree (COL)	0.000	0.933
	Percent bachelor’s degree or higher (BS)	-0.007	0.000
	Percent graduate or professional degree (GD)	-0.007	0.000
Income	Median individual income (MDIINC)	-0.011	0.000
	Household median income (MDHINC)	-0.012	0.000
	Household mean income (MIINC)	-0.011	0.000
	Mean individual income (MHINC)	-0.009	0.000
Other	Employment population ratio (EMP)	-0.006	0.000
	Percentage rural (RUR)	0.003	0.016
	Unemployment rate (UEMP)	0.004	0.014
	Percent below poverty level (POV)	0.011	0.000
	Total population (POP)	-0.003	0.043
	Driver Population (DOP)	-0.003	0.069
	Average Convictions per 1000 driver population (CON)	0.005	0.001
	Area per sq mi (A)	0.001	0.622
	Driver Population per sq mi (DOPSQM)	0.004	0.002
	Total Population per sq mi (POPSQM)	0.004	0.003
	Gender (G)	-0.038	0.000
	Age Group (AGE)	-0.095	0.000

Among the five categories of race, the proportion of Indians, Asians and others is significantly correlated with two-unit crashes. No relationship is observed between the predominant races (white and black) and the fault status. Though the other categories of races have a p-value < 0.05, the correlation coefficient is weak. These categories are generally minorities in Kentucky, and their p-values are significant probably due to their being a smaller proportion of the overall population. Race is thus not expected to serve as a predictor of crash occurrence for two-unit crashes. However, these variables were considered in statistical models in an attempt to examine whether they show any significance when considered along with other variables.

Housing density is unrelated to two-unit crashes; however, other housing variables are significant. As discussed previously, housing value is also another factor related to rurality. This could be related to household income as families with high income tend to live in areas with high housing value. Housing ownership characteristics (rental/owned) are also correlated with two-unit crashes, while rented house density is not related to their occurrence. These relationships are further investigated below.

Marital status has significant effects on two-unit crashes. All variables in this category, except for percent separated, have a p-value < 0.05. Percent separated is significant at 90 percent significance level. Therefore, a detailed investigation on the effect of marital status on the occurrence of two-unit crashes was conducted in later analysis. Furthermore, education is a potential descriptor of two-unit crashes and required more investigation.

Individual and household income show significant relationships with the at-fault status of the driver involved in two-unit crashes. Prior research has demonstrated household income is a better predictor of crash occurrence [14, 17]. Further analysis of two-unit crashes examined the various income categories to determine the most appropriate one for inclusion in the final model. As expected, convictions have a significant positive relationship with crash occurrence. Other variables such as rurality, poverty level, employment, and population, all of which have well-established relationships with crash occurrences, may also be correlated with income and educational level and their interaction was examined.

Recursive Partitioning Analysis

The classification tree confirms that age and gender are the most important factors influencing crash occurrence. The other variables that added value to the prediction of crash occurrence are average convictions, percent below poverty level, percent rural, and percent never married. These variables were further tested, along with the other potential variables identified from the literature review, and correlation tests.

Stepwise Selection

Inputs from the CART model and the correlation analysis were used as a starting point for this step. Along with the variables identified by the CART model, other variables identified as potential predictors were tested to develop the most suitable model representing two-unit crash occurrence. Note that multiple variables from the same category were not used in the same model to avoid complementary and dependency effects. First, predictors in the CART model were examined in a logistic regression model to evaluate whether their p-values fell below the specified level of statistical significance. Insignificant variables were removed from the model one by one. The variation in model parameters were recorded after adding each variable. Similarly, all socioeconomic and demographic variables were tested to select the best subset of predictor variables. During the process, driver population density was identified as another important predictor in the model that improves the predictability of crash occurrence. After testing several combinations of variables, the model including age, gender, convictions, rurality, poverty and driver population density has better parameter estimates than those tested in the process. It has an AIC and BIC of 33,332.7 and 33,465.7, respectively, and an improved percentage correctly classified of 61 percent. The ROC is 0.595. This model was finalized and further tested for interactions in the next step.

Interactions

The FSA was used to find interactions. The algorithm allows testing for interactions on a model with specified explanatory variables. Thus, it allows one to formulate new or to improve upon existing models by adding interactions. Two-way interactions were tested on the chosen models and several criterion functions (such as R^2 and adjusted R^2, interaction p-values, AIC and BIC) were evaluated to examine model quality.

The model finalized in the previous step was tested using the algorithm to identify potential interactions. The tool identified two interactions: between age and gender and between average convictions and driver population density. Among them age-gender was the strongest interaction which repeated the greatest number of times in the iterations. Models with the identified interactions were also developed and evaluated against the simpler model finalized in the previous step. Estimates of these three models and their evaluations were described in the previous section.

6.1.2 Regression Models

This section evaluates the three models for two-unit crashes. Likelihood function, ROC, and probability residuals of the models are compared. Training and validation datasets are also discussed to compare model accuracy.

Model 1

Table 11 shows Model 1, which is the simplest model developed for estimating at-fault driver propensity based on socioeconomic factors for two-unit crashes. This model defines probability of fault as a function of age group, gender, average convictions, driver population density, poverty level, and rurality. All model variables are significant at the 95 percent confidence level.

Table 11 Model 1 for Two-Unit Crashes

Variable	Variable name	Estimate (B)	Std. Error	95% Wald Confidence Interval		Hypothesis Test	p-value	Odds ratio (Exp(B))
				Lower	Upper	Wald Chi-Square		
(Intercept)		0.730	0.0138	0.703	0.757	2794.177	0.000	2.075
[Age Group=0]	<20	0.000						1.000
[Age Group=1]	20-24	-0.352	0.0130	-0.377	-0.326	734.958	0.000	0.703
[Age Group=2]	25-39	-0.792	0.0114	-0.815	-0.77	4842.875	0.000	0.453
[Age Group=3]	40-64	-1.047	0.0111	-1.069	-1.025	8852.281	0.000	0.351
[Age Group=4]	65-75	-0.773	0.0145	-0.801	-0.745	2861.289	0.000	0.462
[Age Group=5]	75-84	-0.270	0.0190	-0.308	-0.233	201.81	0.000	0.763
[Age Group=6]	>84	0.180	0.0403	0.101	0.259	20.013	0.000	1.198
[Gender=0]	Male	0.000						1.000
[Gender=1]	Female	-0.160	0.0058	-0.172	-0.149	752.709	0.000	0.852
Average Convictions/1000 driver population	CON	0.001	0.0004	0	0.002	5.305	0.021	1.001
Percentage rural	RUR	0.000	0.0001	6.91E-05	0.001	6.639	0.010	1.000
Percent below poverty level	POV	0.002	0.0004	0.001	0.003	31.095	0.000	1.002
Driver Population per sqmi	DOPSQM	1.24E-05	2.86E-06	6.81E-06	1.80E-05	18.827	0.000	1.000

The model evaluation parameters are given in Table 12. These values are used as a baseline to evaluate the quality of the two-unit crash models. The likelihood functions are estimators of the relative quality of statistical models for a given dataset. The AIC and BIC of the model are 33,332.7 and 33,465.7, respectively. The AUC of the model is 0.595, and 61 percent of the validation dataset was classified correctly. The probability residual, which is the difference between observed and predicted probability values, was also calculated to validate the model. The residual of 522 zip codes in Kentucky is less than or equal to 0.1. In other words, the differences between the actual probability and the model-predicted probability in these zip codes are less than or equal to 10 percent. This encompasses about 90 percent of Kentucky's overall area. These numbers are used as a baseline comparison when evaluating models with the other variables.

Table 12 Parameters of Model 1 for Two-Unit Crashes

Likelihood Functions			
Log likelihood	-16654.352		
AIC	33332.7		
BIC	33465.7		
ROC			
AUC	0.595		
Validation			
Percent correctly classified	61		
Probability residual			
Residual	Zip code	Area in sq mi	Percentage of area
≤0.10	522	35736.51	90.90
0.10 - 0.20	126	2689.31	6.84
0.20 - 0.30	51	701.04	1.78
>0.30	22	186.4	0.47

As described previously, interactions were tested on this model using the FSA and two two-way interactions identified – one, between average convictions and driver population density and the second between age and gender. These models were then evaluated and compared with Model 1 to identify the best option.

Model 2

Model 2 incorporates the interaction between average convictions and driver population density (Table 13). Along with the interactions and their main effect, the model includes age group, gender, indicator of poverty, and rurality.

Table 13 Model 2 for Two-Unit Crashes

Variable	Variable name	Estimate (B)	Std. Error	95% Wald Confidence Interval		Hypothesis Test	p-value	Odds ratio (Exp(B))
				Lower	Upper	Wald Chi-Square		
(Intercept)		0.681	0.017	0.648	0.714	1611.429	0.000	1.976
[Age Group=0]	<20	0.000						1.000
[Age Group=1]	20-24	-0.352	0.013	-0.378	-0.327	736.101	0.000	0.703
[Age Group=2]	25-39	-0.793	0.0114	-0.815	-0.77	4847.059	0.000	0.453
[Age Group=3]	40-64	-1.047	0.0111	-1.069	-1.025	8857.324	0.000	0.351
[Age Group=4]	65-75	-0.774	0.0145	-0.802	-0.746	2867.153	0.000	0.461
[Age Group=5]	75-84	-0.271	0.019	-0.309	-0.234	203.374	0.000	0.762
[Age Group=6]	>84	0.178	0.0403	0.099	0.257	19.591	0.000	1.195
[Gender=0]	Male	0.000						1.000
[Gender=1]	Female	-0.161	0.0058	-0.172	-0.149	753.193	0.000	0.852
Average Convictions/1000 driver population	CON	0.002	0.0005	0.001	0.003	24.163	0.000	1.002

Percentage rural	RUR	0.000	0.0001	5.66E-05	0	6.071	0.014	1.000
Percent below poverty level	POV	0.003	0.0004	0.002	0.004	48.441	0.000	1.003
Driver Population per sqmi	DOPSQM	4.38E-05	6.81E-06	3.04E-05	5.71E-05	41.293	0.000	1.000
Average Convictions/1000 driver population * Driver Population per sqmi	CON*DOP SQM	-1.11E-06	2.19E-07	-1.54E-06	-6.82E-07	25.774	0.000	1.000

Table 14 displays the evaluation parameters for Model 2. Model 2's AIC and BIC are 33,308.6 and 33,452.8, respectively, indicating better predictive power than Model 1. The AUC and percent correctly classified improved slightly to 0.597 and 61 percent, respectively. The probability residual is also better. The residual is less than or equal to 0.1 for 526 zip codes, which accounts for about 91.10 percent of Kentucky.

Table 14 Parameters of Model 2 for Two-Unit Crashes

<table>
<tr><td colspan="4"><u>Likelihood Functions</u></td></tr>
<tr><td>Log likelihood</td><td colspan="3">-16641.3</td></tr>
<tr><td>AIC</td><td colspan="3">33308.6</td></tr>
<tr><td>BIC</td><td colspan="3">33452.8</td></tr>
<tr><td colspan="4"><u>ROC</u></td></tr>
<tr><td>AUC</td><td colspan="3">0.597</td></tr>
<tr><td colspan="4"><u>Validation</u></td></tr>
<tr><td>Percent correctly classified</td><td colspan="3">61.2</td></tr>
<tr><td colspan="4"><u>Probability residual</u></td></tr>
<tr><td><u>Residual</u></td><td>Zip code</td><td>Area in sq mi</td><td>Percentage of area</td></tr>
<tr><td>≤0.10</td><td>526</td><td>35816.71</td><td>91.11</td></tr>
<tr><td>0.10 - 0.20</td><td>128</td><td>2649.13</td><td>6.74</td></tr>
<tr><td>0.20 - 0.30</td><td>46</td><td>661.07</td><td>1.684</td></tr>
<tr><td>>0.30</td><td>21</td><td>186.35</td><td>0.47</td></tr>
</table>

Table 15 shows the third model developed for predicting two-unit crashes. The predictor variables included in this model are rurality, poverty level, average convictions, driver population density, age groups, gender, and interaction terms between age and gender. The test for interaction using the FSA confirmed the strong correlation between age and gender in crash occurrence, concurring with the findings of previous researchers. Hence, Model 3 is expected to offer better performance than Models 1 and 2.

Table 15 Model 3 for Two-Unit Crashes

Variable	Variable name	Estimate (B)	Std. Error	95% Wald Confidence Interval		Hypothesis Test	p-value	Odds ratio (Exp(B))
				Lower	Upper	Wald Chi-Square		
(Intercept)	Intercept	0.771	0.0169	0.738	0.804	2071.575	0.000	2.162

[Age Group=0]	<20	0.000						1.000
[Age Group=1]	20-24	-0.369	0.0185	-0.405	-0.332	397.135	0.000	0.692
[Age Group=2]	25-39	-0.807	0.0162	-0.839	-0.775	2480.328	0.000	0.446
[Age Group=3]	40-64	-1.093	0.0158	-1.124	-1.062	4783.197	0.000	0.335
[Age Group=4]	65-75	-0.939	0.0202	-0.979	-0.899	2152.687	0.000	0.391
[Age Group=5]	75-84	-0.394	0.0263	-0.446	-0.343	224.289	0.000	0.674
[Age Group=6]	>84	0.068	0.0539	-0.038	0.174	1.595	0.207	1.07
[Gender=0]	Male	0.000						1.000
[Gender=1]	Female	-0.243	0.0201	-0.283	-0.204	147.205	0.000	0.784
Average Convictions per 1000 driver population	CON	0.001	0.0004	0	0.002	5.453	0.020	1.001
Percentage rural	RUR	0.0003	0.0001	7.74E-05	0.001	7.024	0.008	1.000
Percent below poverty level	POV	0.002	0.0004	0.001	0.003	32.288	0.000	1.002
Driver Population per sqmi	DOPSQM	1.22E-05	2.86E-06	6.54E-06	1.78E-05	18.023	0.000	1.000
[Age Group=0] * [Gender=0]	<20 Male	0.000						1.000
[Age Group=0] * [Gender=1]	<20 Female	0.000						1.000
[Age Group=1] * [Gender=0]	20-24 Male	0.000						1.000
[Age Group=1] * [Gender=1]	20-24 Female	0.032	0.0259	-0.019	0.083	1.537	0.215	1.033
[Age Group=2] * [Gender=0]	25-39 Male	0.000						1.000
[Age Group=2] * [Gender=1]	25-39 Female	0.027	0.0227	-0.017	0.072	1.442	0.230	1.028
[Age Group=3] * [Gender=0]	40-64 Male	0.000						1.000
[Age Group=3] * [Gender=1]	40-64 Female	0.091	0.0222	0.047	0.134	16.677	0.000	1.095
[Age Group=4] * [Gender=0]	65-75 Male	0.000						1.000
[Age Group=4] * [Gender=1]	65-75 Female	0.348	0.0289	0.291	0.404	144.322	0.000	1.416
[Age Group=5] * [Gender=0]	75-84 Male	0.000						1.000
[Age Group=5] * [Gender=1]	75-84 Female	0.257	0.0381	0.183	0.332	45.553	0.000	1.293
[Age Group=6] * [Gender=0]	>84 Male	0.000						1.000
[Age Group=6] * [Gender=1]	>84 Female	0.241	0.0812	0.082	0.4	8.822	0.003	1.273

Table 16 displays the evaluation parameters for Model 3. As expected, the model is improved. The AIC and BIC are 33,095.8 and 33,295.4, respectively, while the AUC and classification percentage are bother higher, at 0.612 and 62.9 percent, respectively. The residual did not significantly improve, yet Model 3 predicts 91.12 percent of Kentucky's area within 10 percent error.

Table 16 Parameters of Model 3 for Two-Unit Crashes

Likelihood Functions			
Log likelihood	-16529.9		
AIC	33095.8		
BIC	33295.4		
ROC			
AUC	0.612		
Validation			
Percent correctly classified	62.9		
Probability residual			
Residual	Zip code	Area in sq mi	Percentage of area
<=0.10	523	35818.87	91.12
0.10-=0.20	126	2607.35	6.63
0.20-=0.30	51	700.69	1.78
>0.30	21	186.35	0.47

Comparing the evaluation matrices of all three models, it is apparent that Model 3 has the most robust predictive abilities and capacity to represent two-unit crash occurrences.

6.1.3 Interpretation of Final Model

Model 3 demonstrates the best performance. The final model is a function of rurality, poverty, convictions, driver population density, age, gender, and age-gender interaction. Table 15 shows the coefficients or estimates of the model's variables.

The age group and gender coefficients behave as expected and agree with the findings of prior research. The value of the coefficient for the age group is higher for young and old drivers, which indicates their higher propensity to cause a crash. The negative coefficient for female drivers exhibits their lower susceptibility to be at fault compared to male drivers. The Wald score is the highest for age groups and gender, indicating their strong association with at-fault probability for a crash involvement. The coefficient for age and gender and their relationship with each other are explained later in the section.

Poverty, rurality, average convictions, and driver population density are other predictors of two-unit crash occurrence. The estimates of these variables is positive, concurring with the findings of correlation analysis. The probability of being at-fault increases when a driver resides in area with higher rates of poverty, rurality, population density, and convictions. Among these variables, percent below the poverty line is an important variable with a comparatively high Wald score. It seems to be a strong indicator of at-fault probability, and it agrees with the results of the recursive partitioning analysis.

Age and gender are categorical variables with age classified into seven groups and gender into two. The age groups were numbered from 0 to 6, where 0 is the youngest driver group (< 20 years old) and 6 is the oldest drivers (> 84 years old). For the gender category, 0 represents male drivers and 1 represents female drivers. Logistic regression defines the effect of categories with respect to a reference group. Here, < 20 is the reference group for age and male for gender category. Therefore, the coefficient and the odds ratio of the categories are defined in relation to the reference groups.

The final model takes the form,

$$y = 0.771 + B_1 \cdot age + B_2 \cdot gender + B_3 \cdot age \times gender + 0.0003 \cdot RUR + 0.002 \cdot POV + 0.001 \cdot CON + 1.22 \times 10^{-5} \cdot DOPSQM \quad (8)$$

where B_1, B_2 and B_3 are coefficients of age, gender, and age-gender interaction, respectively. The coefficient varies depending on the category of age and gender being considered.

Age and Gender

Regression model coefficients are interpreted assuming values for the predictor variables. Generally, the coefficients (or odds ratios) of categorical variables in logistic regression models are interpreted, assuming that the other variables take a value equal to zero. This is generally called the base condition. In the current study, the typical approach of assuming the continuous variable as zero did not make sense. Here, the logistic model deals with a zip code's socioeconomic and demographic factors. One of the predictor variables in the model is driver population per square mile; this value cannot be equal to zero for any zip code. To get a general idea about how the propensity to cause a crash varies in each category, odds ratios were calculated.

The RAIR is the ratio of the probability of being at fault to the probability of not being at fault. The RAIR of the quasi-induced exposure is analogous to the odds in logistic regression. In the current context, the odds ratio of being at-fault for each category is represented in terms of a reference group. The reference group for age and gender are < 20 years old and male, respectively.

Table 17 represents the odds ratio of female drivers in each age group with respect to the corresponding male group. Figure 15 represents the ratio in graphical format.

Table 17 Odds Ratio of Female Drivers with Respect to Male Drivers, Two-unit Crashes

Age-group	Odds ratio of female drivers with respect to male
<20	0.784
20-24	0.810
25-39	0.806
40-64	0.859
65-75	1.051
75-84	1.005
>84	0.999

Male drivers are more likely to cause a crash at younger ages. Among older population groups, male and female drivers become roughly equally likely to be at fault when involved in a crash.

The odds ratios exhibit that in young ages and until the age group of 65 – 75, male drivers are responsible for a greater proportion of two-unit crashes. The crash propensity is highest for < 20 male drivers while it gets better with age, probably due to judgment and decision improving with experience. Another reason for the higher involvement of young male drivers could be that they drive more miles than young females. This could increase their exposure and hence their likelihood of causing a crash. Also, young men are more susceptible to aggressive behavior and risk taking while driving, which may also explain their higher odds. The finding on age and gender agree with the findings of the previous research [14, 62].

The vulnerability of female drivers increases with age. Above 65 years of age, male and female drivers contribute almost equally to crash occurrences. This could be attributed to aging-related changes that affect their driving performance [63].

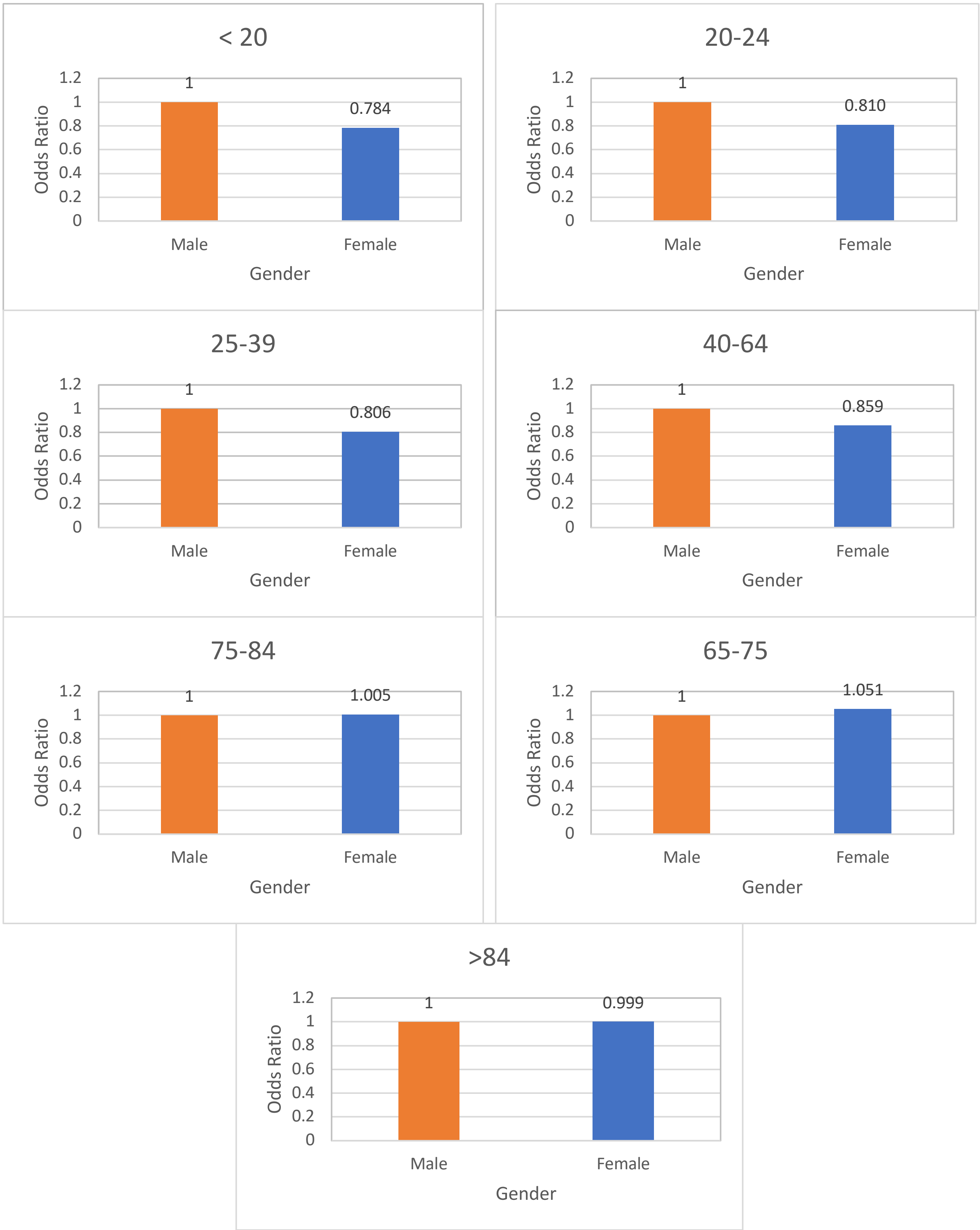


Figure 15 Odds Ratio of Male and Female Drivers by Age Group, Two-Unit Crashes

The model also allows for the comparison of the performance of male and female drivers in each age group. The odds ratios were calculated according to the previous description. Table 18 shows the odds ratios of each age group for males and females. Here the reference group is < 20. The odds ratio represents the propensity of a driver belonging to a particular age group to be at-fault, with respect to the reference group (i.e., < 20 group).

Table 18 Odds Ratio of Male and Female Drivers, Two-unit Crashes

Age Group	Odds Ratio of Male	Odds Ratio of Female
<20 (reference)	1.000	1.000
20-24	0.691	0.714
25-39	0.446	0.458
40-64	0.335	0.367
65-75	0.391	0.554
75-84	0.674	0.872
>84	1.070	1.362

Figure 16 and 17 graphically represent the odds ratios of male and female drivers. Among both male and female drivers, older drivers (> 84) have the highest odds ratio compared to young drivers. The odds ratios for the age groups follow the typical U-shape curve of crash involvement, with higher probabilities for younger and older drivers. For both males and females, younger and older drivers are more likely to be the driver at fault than middle-aged drivers. This concurs the findings from the literature review [14, 62].

From the table, it is evident that the odds ratios of male drivers are slightly higher for young drivers than young female drivers, while the odds ratios increase for female drivers as they grow older. This means that the female drivers are more likely to be at fault when older. This concurs with the results shown in Figure 2.

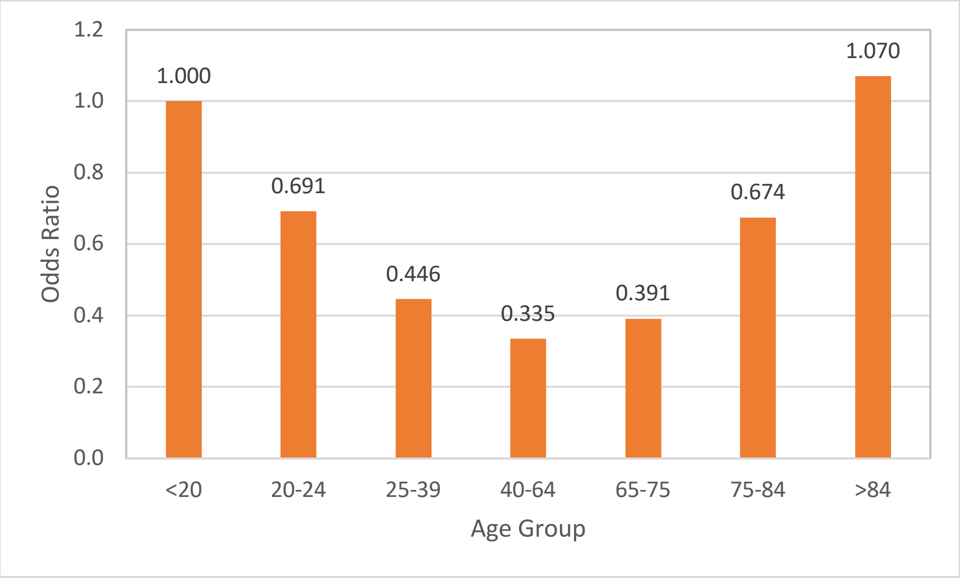


Figure 16 Odds Ratio of Male Drivers, Two-Unit Crashes

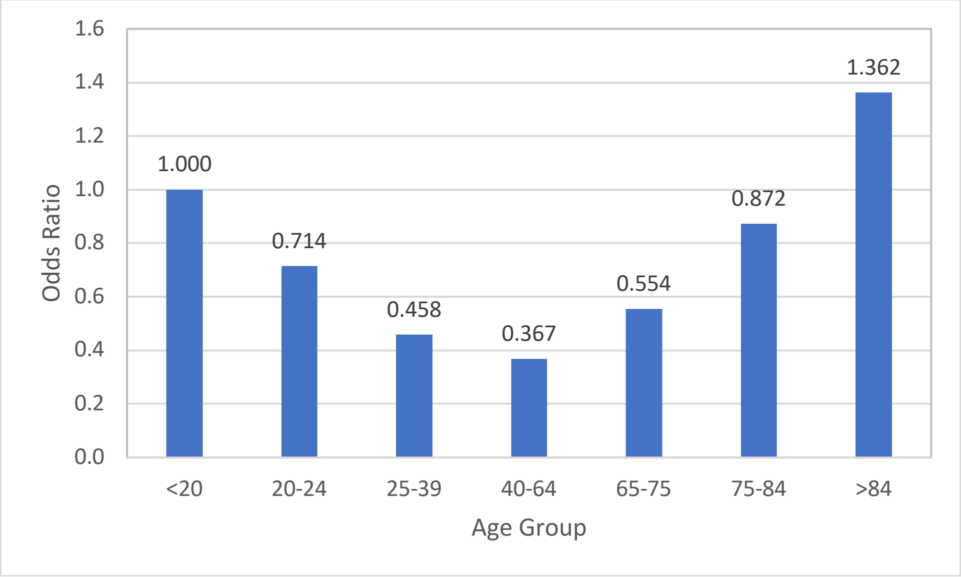


Figure 17 Odds Ratio of Female Drivers, Two-Unit Crashes

Socioeconomic Factors

Poverty and rurality are the two socioeconomic variables in the final model. The relationship of these variables can be interpreted in terms of odds, however, representing the relationship in a graphical format is easier to understand. The graphs show the predicted probability of each age-gender category in the y-axis while x-axis represents the socioeconomic variable of a zip code. The graphs demonstrate how the predicted probability for each category vary with change in their socioeconomic characteristics.

The coefficient of rurality in the final regression model (Table 15) is 0.0003. It represents the difference in log odds when the percent rural is increased by a unit. i.e., when percentage rural increases by 1, the log odds increase by 0.0003. In other words, the odds of being at fault = Exp (0.0003) = 1.0003 which implies that for one-unit increase in percent rural increases the odds of being at-fault by 0.03 percent. Therefore, for 33.33 unit increase in percent rural, one could expect a 1 percent increase in the odds of being at fault.

Figure 18 and Figure 19 show the relationship of rurality with the age-gender categories in the model. Throughout the analysis, rurality is observed to have a strong positive correlation with crash occurrence. However, rurality does not show any evident relationship with the at-fault probability of male and female drivers when age and other socioeconomic characteristics are considered. The effect of rurality is diminished probably due to its potential interaction with other socioeconomic variables in the model.

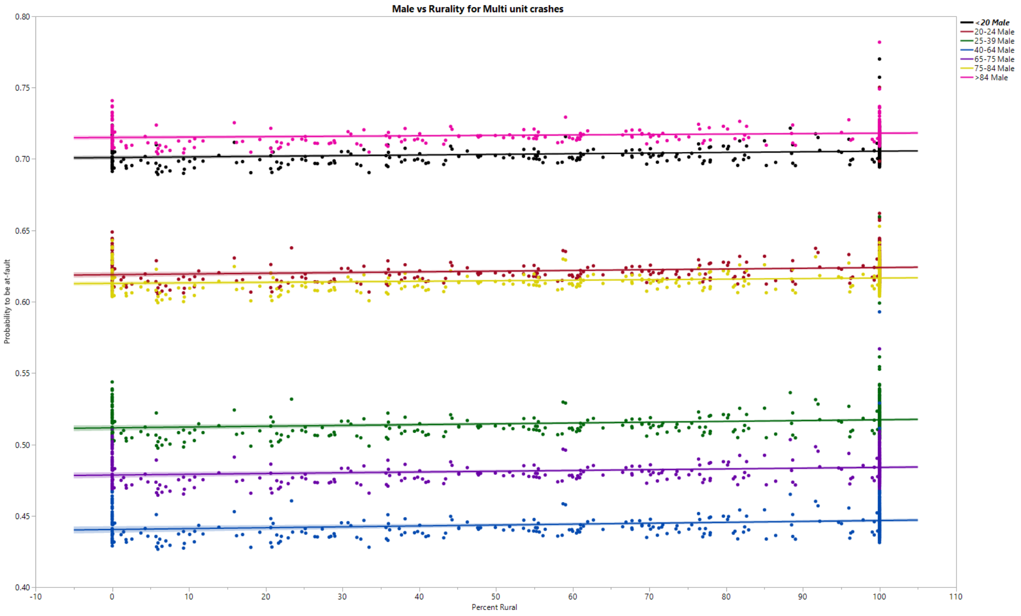


Figure 18 At-fault Probability of Male Drivers with Rurality, Two-Unit Crashes

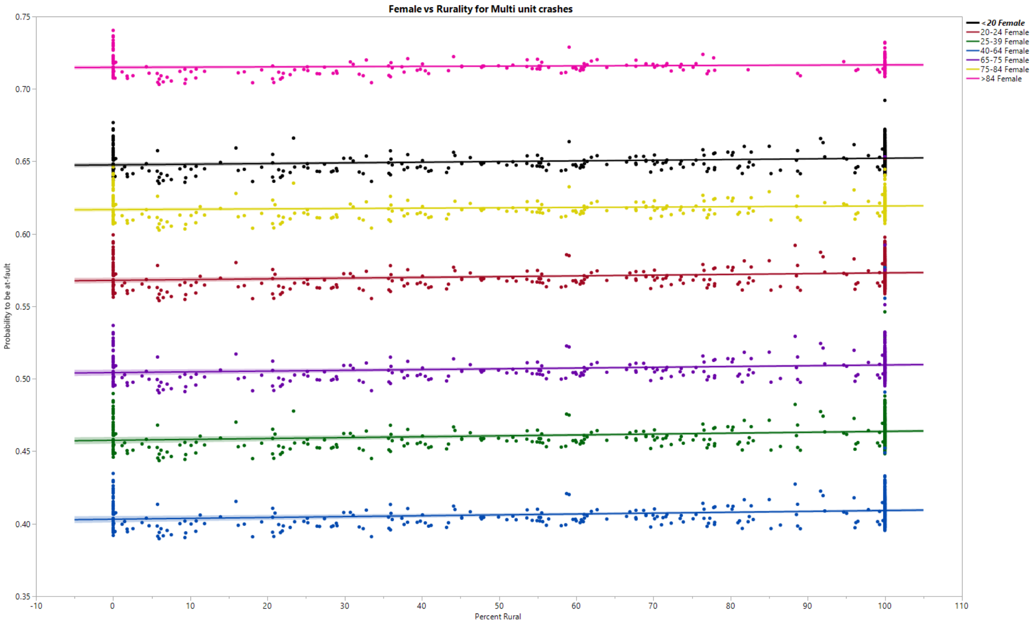


Figure 19 At-Fault Probability of Female Drivers with Rurality, Two-Unit Crashes

Poverty is the other socioeconomic predictor of two-unit crash occurrence. The estimate of the variable in the regression model (Table 13) is 0.002 which indicate that when poverty level increases by 1 unit, the log odds of being at-fault increased by 0.002. In other words, for every one-unit increase in poverty level, 0.2

percent increase in the odds is expected. Therefore, for 5 unit increase in poverty level, the odds of being at fault increases by 1 percent.

In the graphical format, poverty level has a positive relationship with predicted at-fault probability of male and female drivers in all age-groups (Figure 20 and Figure 21).

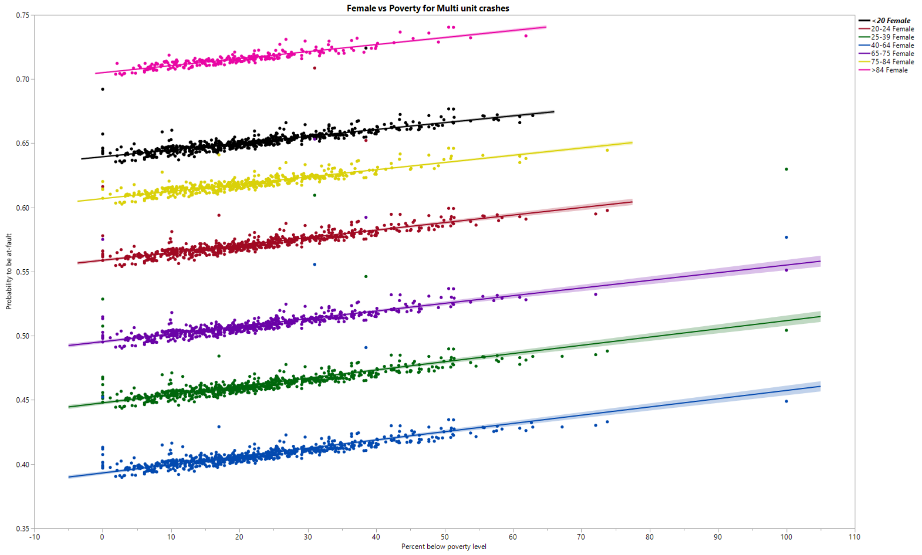


Figure 20 At-Fault Probability of Male Drivers with Poverty, Two-Unit Crashes

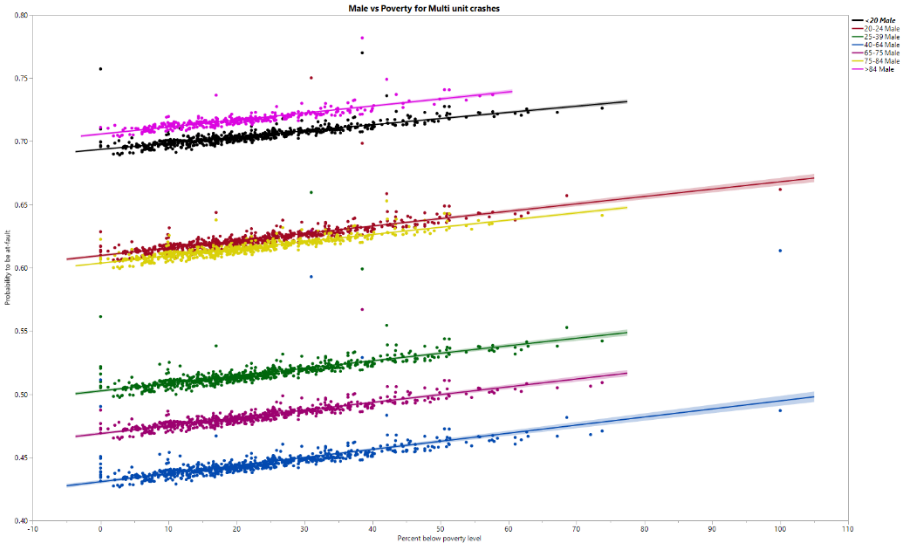


Figure 21 At-Fault Probability of Female Drivers with Poverty, Two-Unit Crashes

6.2 Single-Unit Crashes

To analyze the relationship between single-unit crashes and the socioeconomic characteristics of driver's zip code, initial data analysis was conducted following the same steps explained above. The results of the tests are discussed in this section. Based on the results from these tests, several models were tested, and their model parameters compared to recommend the most appropriate model for determining the probability of a driver being at fault.

6.2.1 Variable Selection

Correlation Test

A correlation matrix for single crashes was developed to identify variables associated with at-fault status. Correlation coefficients were calculated for each predictor variable, which represent its association with the dependent variable. Most of the variables are statistically significant at the 95 percent confidence level (shaded in green), and Table 19 shows their correlation with the dependent variable.

Table 19 Correlation Test Results for Single-Unit Crashes

Category	Variable	Correlation Coefficient	p-value
Race	Percent white (WH)	0.115	0.000
	Percent black (BL)	-0.097	0.000
	Percent American Indian (AI)	0.002	0.401
	Percent Asian (AS)	-0.130	0.000
	Percent other races (OR)	-0.070	0.000
Housing	Household units (HH)	-0.131	0.000
	Household ownership total (HHO)	-0.134	0.000
	Owner occupied housing units (OHU)	-0.132	0.000
	Renter occupied housing units (RHU)	-0.115	0.000
	Median housing value (HVL)	-0.121	0.000
Marital Status	Percent now married (MRD)	0.073	0.000
	Percent widowed (WID)	0.067	0.000
	Percent divorced (DIV)	-0.008	0.004
	Percent separated (SEP)	0.029	0.000
	Percent never married (NMD)	-0.102	0.000
Education	Percent less than high school graduate (LHS)	0.126	0.000
	Percent high school graduate (HS)	0.139	0.000
	Percent some college/associate degree (COL)	-0.074	0.000
	Percent bachelor's degree or higher (BS)	-0.141	0.000
	Percent graduate or professional degree (GD)	-0.123	0.000
Income	Median individual income (MDIINC)	-0.115	0.000
	Household median income (MDHINC)	-0.090	0.000
	Household mean income (MIINC)	-0.097	0.000
	Mean individual income (MHINC)	-0.114	0.000
Other	Employment population ratio (EMP)	-0.136	0.000
	Percentage rural (RUR)	0.189	0.000
	Unemployment rate (UEMP)	0.042	0.000
	Percent below poverty level (POV)	0.066	0.000
	Total population (POP)	-0.127	0.000
	Driver Population (DOP)	-0.127	0.000
	Average Convictions per 1000 driver population (CON)	-0.016	0.000
	Area per sq mi (A)	0.093	0.000
	Driver Population per sq mi (DOPSQM)	-0.128	0.000
	Total Population per sq mi (POPSQM)	-0.126	0.000
	Gender (G)	-0.121	0.000
	Age Group (AGE)	-0.199	0.000

Among the five races, the predominant categories, (i.e., the proportion of white and black) is significantly correlated with the occurrence of single-unit crashes. Percent white is positively correlated with single-unit crashes, which means that drivers from zip codes with more white population are likely to cause more single-unit crashes. At the same time, a negative correlation is observed with percent black. In Kentucky, the proportion of people belonging to other-than-white races is significantly smaller. Therefore, the other categories of race might not to be an important descriptor of a driver's at-fault status. However, a significant association between race and single-unit crashes is apparent. Therefore, these variables were considered in the statistical modeling to examine whether they remain significant when considered along with other variables.

Similarly, all of the housing variables are related to single-unit crashes, and they are negatively correlated with crash occurrence. Thus, the crash propensity of the drivers living in areas with high housing density or housing value (most likely urban areas) is low. Housing density and housing value are evidently related to rurality, and there could be a statistically important interaction among them when tested in a model. Housing value could be related to household income, as families with high income tend to live in areas with high housing value. Housing ownership characteristics (rental/owned) are also correlated with crash occurrence. These relationships were further investigated in the next step.

Marital status shows results that agree with prior research. Drivers previously married (widowed, separated and divorced) are correlated with the at-fault status and their crash involvement has been considered as a result of stressful life events. This was further investigated in the next level of analysis. Furthermore, education is also in agreement with prior research: less educated people are more likely to be the at-fault driver in a crash. In Table 14, as educational attainment increases, the sign of the correlation coefficient turns negative, which indicates lower crash involvement as an at-fault driver.

All types of income show a significant relationship with driver at-fault status according to previous research, but household median income is expected to be a better predictor of crash occurrence [14, 17]. These variables indicate a negative relationship with crash occurrence, agreeing with the findings of previous research. The analysis of this research examined the various income categories and determine the most appropriate one for inclusion in the final model predicting crash occurrence. Other variables such as rurality, poverty level, unemployment rate and population density that have well-established relationships with crash occurrence may be also correlated with income and educational level and their interaction was examined.

Recursive Partitioning Analysis

Recursive partitioning analysis was performed on the single-unit crash dataset and a CART model was developed to assist in variable selection. The classification tree finds that age, rurality, and gender are the most important factors influencing crash occurrence. However, education and unemployment rate also added value to the model. These variables were further tested along with other potential variables identified from the literature review and correlation analysis.

Stepwise Selection

As the initial step, the predictors in the CART model were examined in a logistic regression model to evaluate whether their p-values fell below the specified level of statistical significance. Percent never married and percent rural have p-values > 0.05 in this model. This is probably due to some interaction with the other predictor variables. Next, the income variables were tested. Household median income has a higher Wald score in the model compared to models tested with the other income variables. However, adding household median income influences the significance of rurality in the model. It is obvious that this is the product of the interrelation of these variables with income. Other socioeconomic variables such as poverty, marital status, and race were also tested, but their addition does not improve model parameters substantially. Conviction was also tested in the model to analyze its contribution to improve predictability.

It is one of the major descriptors of the two-unit crashes; hence it is important to check the contribution of convictions in a single-unit crash occurrence – it appears to be significant in the model. The final model includes age, gender, percent rural, percent with bachelor's degree, driver population density, and convictions. The AIC and BIC values are 261,96.99 and 26,315.95, respectively, and they are better than others tested in the process, The ROC is 0.6792, while percent correctly classified is 68.9 percent. Since this model represents the occurrence of single-unit crashes better than other models tested, it was finalized and the test for interactions proceeded.

Interactions

A process similar to two-unit crashes was conducted using the FSA to test for interactions. Two-way interactions were tested on the chosen models and several criterion functions evaluated to examine model quality. The algorithm identified two interactions on the model finalized in the previous step. The first one is between age and gender. This is similar to the findings of the test on two-unit crashes and prior research. The second one is between average conviction and percent with bachelor's degree. The term positively correlates with single-unit crash occurrence, and it needs further investigation. Among the two, the first exhibited a stronger existence in the iterations. However, the predictive power of both the models was tested along with the simpler one finalized in the stepwise selection process.

6.2.2 Regression Models

This section discusses the three models finalized for the single-unit crashes. The likelihood function, ROC, and probability residuals of the models are compared, followed by training and validation.

Model 1

Model 1 is the simplest one developed for single-unit crashes to estimate at-fault driver propensity based on socioeconomic factors of the driver's residence zip code. This model defines probability of fault as a function of age group, rurality, educational attainment, average convictions and driver population density. All of the variables in the model are significant at the 95 percent confidence level. Table 20 shows the model's estimates.

Table 20 Model 1 for Single-Unit Crashes

Variable	Variable name	Estimate (B)	Std. Error	95% Wald Confidence Interval		Hypothesis Test	p-value	Odds ratio (Exp(B))
				Lower	Upper	Wald Chi-Square		
(Intercept)		1.013	0.036	0.942	1.083	787.637	0.000	2.753
[Age Group=0]	<20	0.000						1.000
[Age Group=1]	20-24	-0.361	0.023	-0.407	-0.316	245.435	0.000	0.697
[Age Group=2]	25-39	-0.874	0.020	-0.914	-0.834	1857.861	0.000	0.417
[Age Group=3]	40-64	-1.325	0.02	-1.364	-1.286	4396.223	0.000	0.266
[Age Group=4]	65-75	-1.476	0.028	-1.531	-1.42	2712.757	0.000	0.229
[Age Group=5]	75-84	-1.105	0.038	-1.181	-1.029	812.851	0.000	0.331
[Age Group=6]	>84	-0.591	0.085	-0.758	-0.424	48.168	0.000	0.554
[Gender=0]	Male	0.000						1.000
[Gender=1]	Female	-0.505	0.011	-0.527	-0.484	2115.489	0.000	0.603
Percentage rural	RUR	0.008	0.000	0.007	0.008	1269.454	0.000	1.008
Percent bachelor's degree or higher	BS	-0.014	0.001	-0.016	-0.012	197.623	0.000	0.986
Driver Population per sqmi	DOPSQM	-6.06E-05	5.43E-06	-7.12E-05	-4.99E-05	124.58	0.000	1.000
Average Convictions/1000 driver population	CON	0.002	0.000	0	0.003	4.669	0.031	1.002

Table 21 shows the results of the model evaluation parameters. These values served as a baseline for evaluating the quality of the models developed for single-unit crashes. The AIC and BIC of the model are 26,196.9 and 26,315.9, respectively, the AUC is 0.679, and 63.1 percent of the validation dataset was classified correctly. The probability residual, which is the difference between observed and predicted probability values is less than or equal to 10 percent for 404 zip codes, which accounts for approximately 77.3 percent of Kentucky's overall area.

Table 21 Parameters of Model 1 for Single-Unit Crashes

Likelihood Functions			
Log likelihood	-13086.5		
AICc	26196.9		
BIC	26315.9		
ROC			
AUC	0.67925		
Validation			
Percent correctly classified	63.1		
Probability residual			
Residual	Zip code	Area in sq mi	Percentage of area
<=0.10	404	30331.92	77.344
0.10-=0.20	185	6943.38	17.705
0.20-=0.30	83	1438.91	3.669
>0.30	40	502.83	1.282

The two interactions identified through the FSA are between age and gender and average convictions and percent with bachelor's degree. These models are evaluated in the following section and compared with Model 1 to identify the best option.

Model 2

Model 2 incorporates the interaction age and gender (Table 22). Along with the interaction and its main effect, the model includes the other variables in Model 1 – rurality, education, driver population density, and average convictions.

Table 22 Model 2 for Single-Unit Crashes

Variable	Variable name	Estimate (B)	Std. Error	95% Wald Confidence Interval		Hypothesis Test	p-value	Odds ratio (Exp(B))
				Lower	Upper	Wald Chi-Square		
(Intercept)		1.008	0.0397	0.93	1.086	645.668	0.000	2.739
[Age Group=0]	<20	0.000						1.000
[Age Group=1]	20-24	-0.293	0.031	-0.355	-0.23	84.252	0.000	0.746
[Age Group=2]	25-39	-0.832	0.028	-0.887	-0.777	885.234	0.000	0.435
[Age Group=3]	40-64	-1.326	0.027	-1.38	-1.272	2329.867	0.000	0.265
[Age Group=4]	65-75	-1.635	0.038	-1.71	-1.56	1835.027	0.000	0.195
[Age Group=5]	75-84	-1.316	0.052	-1.418	-1.214	634.162	0.000	0.268
[Age Group=6]	>84	-0.911	0.115	-1.137	-0.685	62.403	0.000	0.402
[Gender=0]	Male	0.000						1.000
[Gender=1]	Female	-0.495	0.035	-0.565	-0.425	193.876	0.000	0.610
Percentage rural	RUR	0.008	0.0002	0.007	0.008	1273.488	0.000	1.008
Percent bachelor’s degree or higher	BS	-0.014	0.001	-0.016	-0.012	199.294	0.000	0.986
Driver Population per sqmi	DOPSQM	-6.07E-05	5.43E-06	-7.14E-05	-5.01E-05	124.945	0.000	1.000
Average Convictions/1000 driver population	CON	0.002	0.0007	0	0.003	4.679	0.031	1.002
[Age Group=0] * [Gender=0]	<20Male	0.000						1.000
[Age Group=0] * [Gender=1]	<20 Female	0.000						1.000
[Age Group=1] * [Gender=0]	20-24 Male	0.000						1.000
[Age Group=1] * [Gender=1]	20-24 Female	-0.148	0.046	-0.239	-0.058	10.318	0.001	0.862
[Age Group=2] * [Gender=0]	25-39 Male	0.000						1.000
[Age Group=2] * [Gender=1]	25-39 Female	-0.093	0.040	-0.173	-0.014	5.283	0.022	0.911
[Age Group=3] * [Gender=0]	40-64 Male	0.000						1.000
[Age Group=3] * [Gender=1]	40-64 Female	0.004	0.04	-0.074	0.082	0.01	0.920	1.004
[Age Group=4] * [Gender=0]	65-75 Male	0.000						1.000
[Age Group=4] * [Gender=1]	65-75 Female	0.380	0.056	0.269	0.491	44.872	0.000	1.462
[Age Group=5] * [Gender=0]	75-84 Male	0.000						1.000
[Age Group=5] * [Gender=1]	75-84 Female	0.476	0.077	0.324	0.627	37.798	0.000	1.609
[Age Group=6] * [Gender=0]	>84 Male	0.000						1.000
[Age Group=6] * [Gender=1]	>84 Female	0.678	0.169	0.346	1.009	16.019	0.000	1.969

The goodness of fit parameters for the model (AIC and BIC) are 26,023.9 and 26,202.3, respectively, indicating better predictive power than Model 1. The AUC is slightly improved while the percent correctly classified in the training and validation datasets remained the same. A greater number of zip codes (408) are predicted under the 10 percent error.

Table 23 Parameters of Model 2 for Single-Unit crashes

Likelihood Functions			
Log likelihood	-12994		
AICc	26023.9		
BIC	26202.3		
ROC			
AUC	0.68028		
Validation			
Percent correctly classified	63.1		
Probability residual			
Residual	Zip code	Area in sq mi	Percentage of area
<=0.10	408	30491.76	77.751
0.10-=0.20	182	6808.59	17.361
0.20-=0.30	84	1443.91	3.682
>0.30	38	472.78	1.206

Model 3

Table 24 shows the third model developed for predicting single-unit crashes. This model includes the interaction identified between percent with bachelor's degree and convictions. The other predictor variables, along with the main effects of the interaction terms, are percent rural, age, gender, and driver population per square mile.

Table 24 Model 3 for Single-Unit Crashes

Variable	Variable name	Estimate (B)	Std. Error	95% Wald Confidence Interval		Hypothesis Test	p-value	Odds ratio, Exp(B)
				Lower	Upper	Wald Chi-Square		
(Intercept)		1.100	0.0377	1.026	1.174	850.82	0.000	3.004
[Age Group=0]	<20	0.000						1
[Age Group=1]	20-24	-0.364	0.0231	-0.409	-0.319	248.45	0.000	0.695
[Age Group=2]	25-39	-0.879	0.0203	-0.919	-0.84	1876.7	0.000	0.415
[Age Group=3]	40-64	-1.327	0.02	-1.367	-1.288	4403.1	0.000	0.265
[Age Group=4]	65-75	-1.477	0.0284	-1.532	-1.421	2713.3	0.000	0.228
[Age Group=5]	75-84	-1.105	0.0388	-1.181	-1.028	810.23	0.000	0.331
[Age Group=6]	>84	-0.588	0.0852	-0.755	-0.421	47.526	0.000	0.556
[Gender=0]	Male	0.000						1
[Gender=1]	Female	-0.505	0.011	-0.527	-0.484	2111.1	0.000	0.603
Percentage rural	RUR	0.008	0.0002	0.008	0.009	1354.8	0.000	1.008
Percent bachelor's degree or higher	BS	-0.032	0.002	-0.036	-0.028	263.94	0.000	0.968
Driver Population per sqmi	DOPSQM	-6.18E-05	5.44E-06	-7.24E-05	-5.11E-05	128.78	0.000	1
Average Convictions/1000 driver population	CON	-0.005	0.0009	-0.006	-0.003	23.464	0.000	0.995
Percent bachelor's degree or higher * Average Convictions/1000 driver population	BS*CON	0.001	8.71E-05	0.001	0.001	116.1	0.000	1.001

The evaluation parameters of Model 3 are displayed in Table 25. Though percent correctly predicted, and the probability residual remain the same, the other model's parameters are worse than Model 2. The AIC and BIC increases to 26076.7 and 26205.5, respectively, while the AUC falls to 0.679.

Table 25 Parameters of Model 3 for Single-Unit Crashes

Likelihood Functions			
Log likelihood	-13025.4		
AIC	26076.7		
BIC	26205.5		
ROC			
AUC	0.67973		
Validation			
Percent correctly classified	63.1		
Probability residual			
Residual	Zip code	Area in sq mi	Percentage of area
<=0.10	408	30451.68	77.649
0.10-=0.20	176	6707.41	17.103
0.20-=0.30	81	1539.53	3.926
>0.30	47	518.41	1.322

Comparing the evaluation matrices of all three models, it is apparent that Model 2 has the most robust predictive power and offers the best representation of single-unit crash occurrence.

6.2.3 Interpretation of final model

Model 2 is best of the models examined. The final model is a function of rurality, education, convictions, driver population density, age, gender, and age-gender interactions.

Similar to two-unit crashes, the coefficients for age group and gender behave as expected and agree with the findings of prior research. The age group coefficient reflects the higher likelihood of young and older drivers being at fault in a crash. Female drivers are less likely to cause single-unit crashes than their male counterparts. The Wald score for age groups and gender are high, indicating their strong association with at-fault status. The interaction between age and gender are explained in detail later in the section.

Rurality is another predictor variable in the model, and it is one of the variables with the highest Wald score. This indicates the strong association between the rurality of the driver's residence zip code and the driver's likelihood of causing a single-unit crash. This agrees with the results of the recursive partitioning analysis. The other predictor variables in Model 2 are average convictions, percent with bachelor's degree, and driver population density. The coefficients for percent rural and average convictions have a positive relationship with fault status, concurring with the findings of previous research. Percent with bachelor's degree was included in the model, and it has a negative association with the dependent variable. This indicates that people with higher educational attainment have lower chance of causing single-unit crashes.

Driver population density displays an interesting relationship with single-unit crash occurrence. The variable has a negative estimate in the logistic regression model, meaning that drivers residing in less dense areas cause more singe-unit crashes. This can be explained by the positive coefficient of rurality in the model. It is highly likely that rural areas are less populated and thus there may be some interaction here that was not easily detected.

Age and gender are categorized and numbered similar to the two-unit crashes. Age is grouped into seven categories while there are two gender categories. Again, the coefficient and the odds ratio of the categories are defined in terms of the reference groups, which is < 20 years old for age and male for gender.

The final model takes the form,

$$y = 1.008 + B_1\cdot age + B_2\cdot gender + B_3\cdot age\times gender + 0.008\cdot RUR - 0.014\cdot BS + 0.002\cdot CON - 6.07\times10^{-5}\cdot DOPSQM \quad (10)$$

where B_1, B_2 and B_3 are coefficients of age, gender, and their interaction, respectively. The coefficients are given in Table 21 and they vary depending on what category of age and gender is under consideration.

Age and Gender

Through the process explained previously, the effects of these variables can be accounted for using the values of their estimates. Again, the coefficients of the categorical variables cannot be interpreted following the general process of assuming the value of continuous variables as zero. The single-unit model also has driver population per square mile as a predictor variable, a value which cannot equal zero.

Similar to the two-unit crashes, the odds ratios of being at-fault for each category were calculated. The ratios were represented in terms of a reference group. Table 26 represents the odds ratios for female drivers in each age group compared to their respective male group. Figure 22 represents the ratios graphically.

Table 26 Odds Ratio of Female Drivers with Respect to Male Drivers, Single-Unit Crashes

Age-group	Odds ratio of female drivers with respect to male drivers
<20	0.610
20-24	0.526
25-39	0.555
40-64	0.526
65-75	0.916
75-84	1.461
>84	1.110

It is evident from the figures that male drivers are more likely to cause a single-unit crash that those of younger ages. In older ages, both male and female drivers are equally likely to be at-fault when involved in a crash.

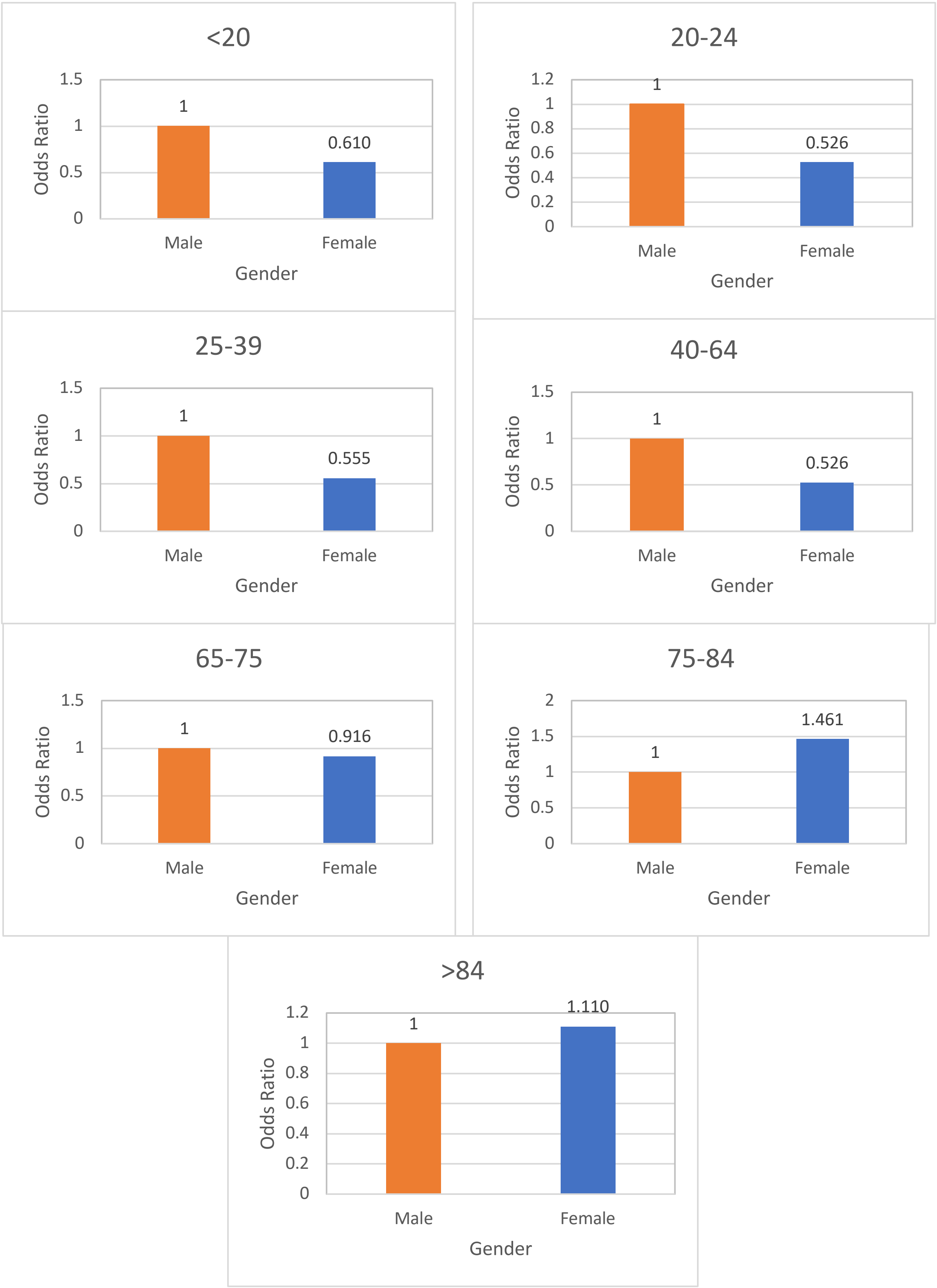


Figure 22 Odds Ratio of Male and Female Drivers by Age Group, Single-Unit Crashes

Male drivers are highly at-risk of causing single-unit crashes until middle age. The reason for high risk rate could be the aggressive and risk-taking behavior of male drivers [14]. The likelihood of causing a single-unit crash falls in older age groups, probably because of the greater experience in handling situations that may lead to a single-unit crash. Female drivers are better when they are young; their performance diminishes as they turn older, probably due to aging-related factors [63].

Male and female drivers in each age group can be compared to better understand their performance. Table 27 shows the odds ratios for male and female drivers. Here, the < 20 group is again chosen as the reference category.

Table 26 Odds Ratio of Male and Female Drivers, Single-Unit Crashes

Age Group	Odds ratio of male	Odds ratio of female
<20 (reference)	1.000	1.000
20-24	0.746	0.643
25-39	0.435	0.397
40-64	0.266	0.229
65-75	0.195	0.285
75-84	0.158	0.432
>84	0.402	0.792

The odds ratios listed in Table 27 are reproduced in figures below. Both figures show that these ratios follow the typical U-shaped curve for crash involvement, with higher probabilities for younger and older drivers. For both males and females, younger and older drivers are more likely to be at fault than the middle-aged drivers. This finding agrees with prior research [62].

The graphs also demonstrate the same findings as those discussed above. Male drivers have the highest odds ratios until middle age. At the same time, crash involvement rate of females is higher for younger and older drivers.

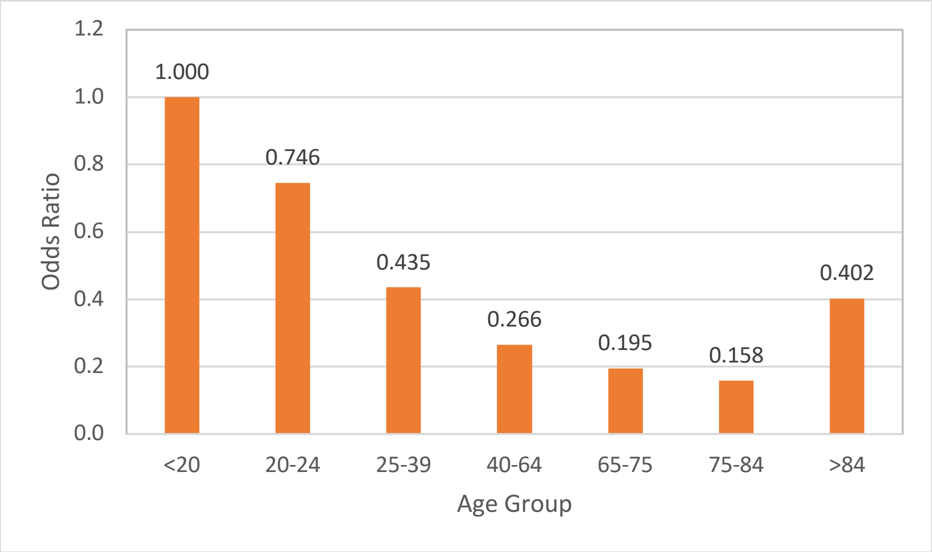


Figure 23 Odds Ratio for Male, Single-Unit Crashes

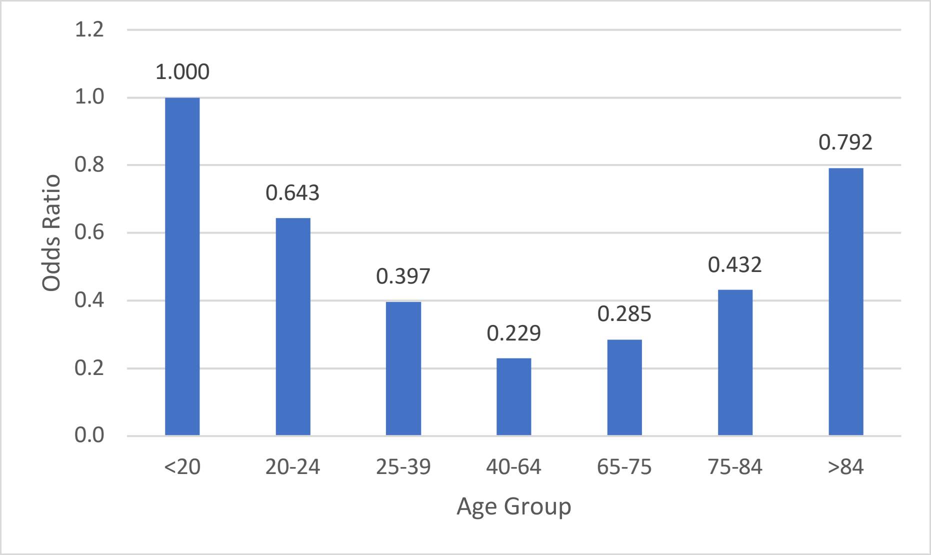


Figure 24 Odds Ratio for Female, Single-Unit Crashes

Socioeconomic Factors

The regression model predicting the occurrence of single-unit crashes includes two socioeconomic variables – rurality and percent with bachelor's degree (Table 22). Their estimates can be interpreted in terms of log-odds or odds of being at-fault.

The coefficient of rurality in the final regression model is 0.008, which is the difference in log odds when percent rural is increased by a unit. For every 1 percentage increase in rurality, the odds of being at fault in a crash increases by 0.8 percent. In other words, a 12.5 unit increase in percent rural increases the odds of being at fault by 1 percent.

The graphs depicting the influence of rurality on single-unit crash occurrence indicate strong positive association and concur with the findings of two-unit crashes (Figure 23 and Figure 26). For every age-gender category, their at-fault probability increases with rurality of their residence zip code.

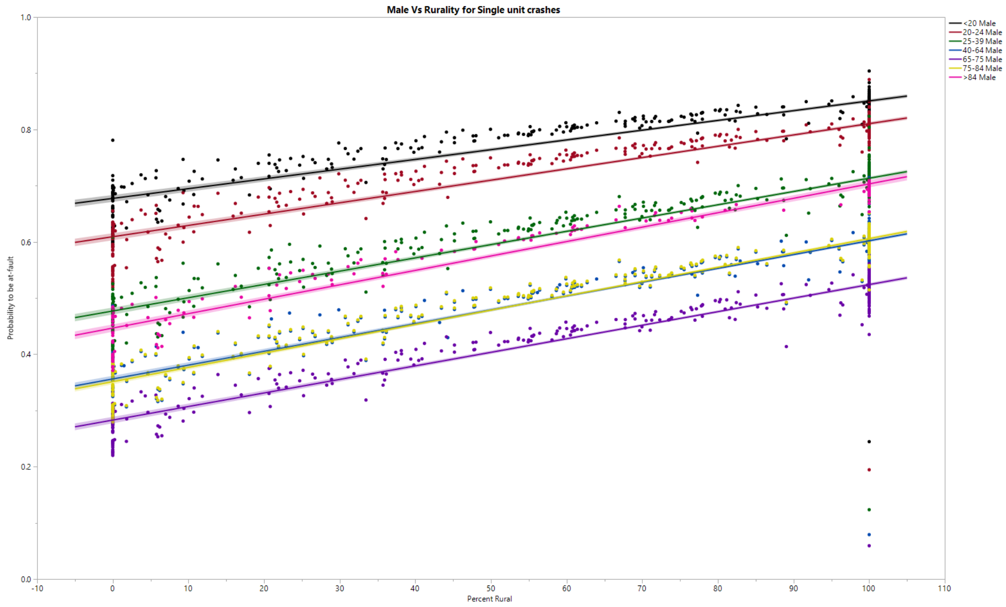


Figure 25 At-Fault Probability of Male Drivers with Rurality, Single-Unit Crashes

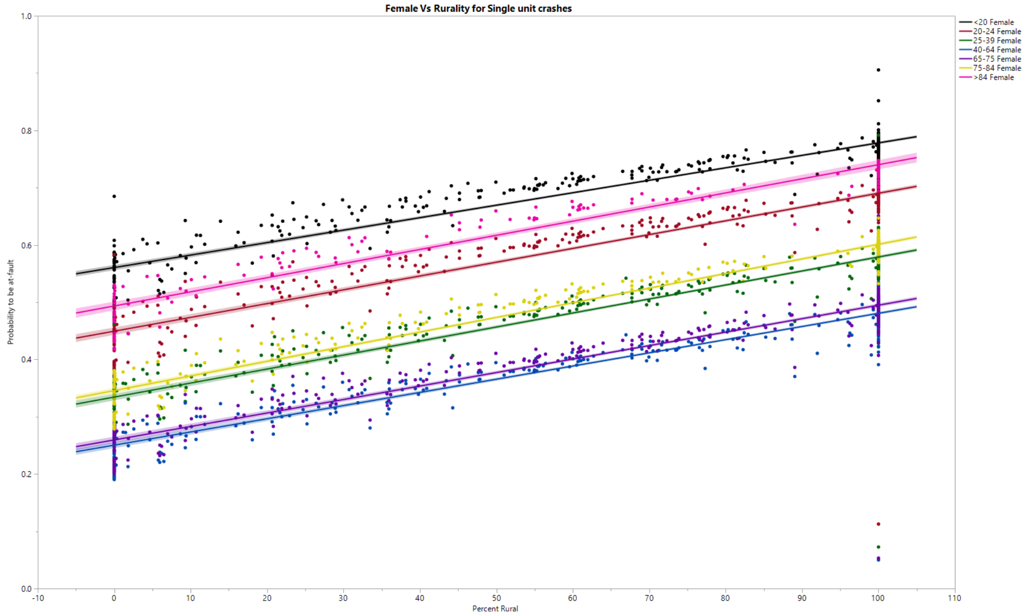


Figure 26 At-Fault Probability of Female Drivers with Rurality, Single-Unit Crashes

Similarly, the impact of educational attainment (percent with bachelor's degree) on crash occurrence can also be interpreted. The estimate of the educational descriptor in the model is -0.014. The negative sign indicates an inversely proportional relationship between the independent and dependent variable. The odds of the variable (i.e., Exp (-0.014)) is equal to 0.9861 which indicates that for every 1 unit of increase in educational attainment, 0.014 percent decrease is observed in the odds. Therefore, for 71.94 unit increase in percent with bachelor's degree, 1 percent increase in the odds of being at fault is observed.

Figure 27 and Figure 28 shows this in a graphical format. For both male and female drivers in all age groups, the probability of being at-fault decreases if they reside in a zip code with higher educational attainment.

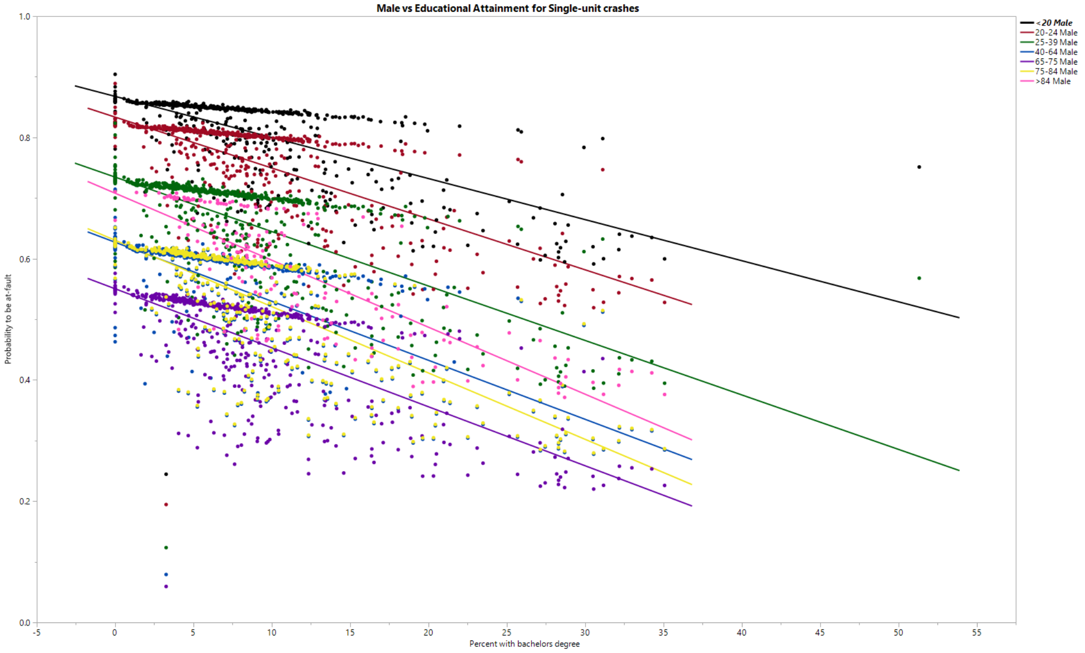


Figure 27 At-Fault Probability of Male Drivers with Educational Attainment, Single-Unit Crashes

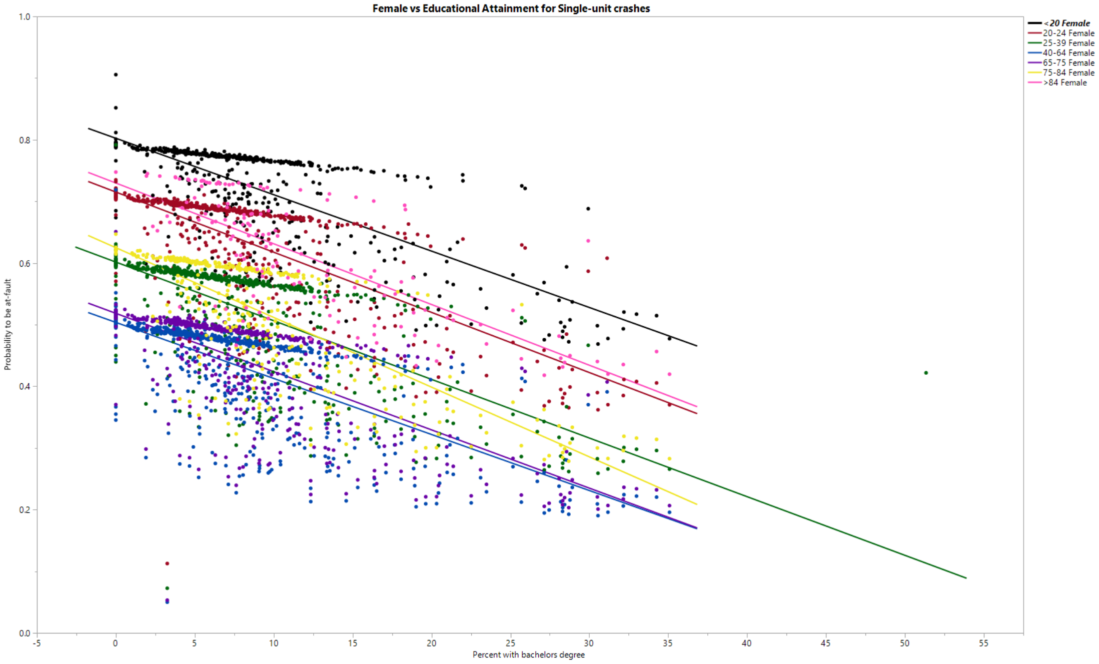


Figure 28 At-Fault Probability of Female Drivers with Educational Attainment, Single-Unit Crashes

7. Conclusions

This research examined the relationship between crash occurrence and socioeconomic factors associated with at-fault driver residence using data from the U.S. Census Bureau. To further investigate this, single- and two-unit crashes that occurred in Kentucky were analyzed separately. Mathematical models were developed that identified the socioeconomic and demographic characteristics of a driver's home zip code and how those factors make a driver more likely to cause crashes. Key socioeconomic factors considered include rurality, educational attainment, poverty percentage, population density, and convictions. Driver age and gender have a well-established relationship with probability of crash occurrence; hence they were also included in the models. Several other factors, including income, employment, marital status, and race, were also tested.

In this type of research, it is important to consider crash exposure when attempting to identify contributing factors to a crash. Crash databases lack information on driver exposure. The quasi-induced exposure technique was used here, which assumes that the not-at-fault drivers represent the total population in question; the crash rate measure of exposure was developed in terms of the RAIR. RAIR is the ratio of the percentage of at-fault drivers to the percentage of not-at-fault drivers in the same subgroup. Hence, the dependent variable used here was the fault status of a driver involved in a crash, which is binary.

Spatial analysis was used to investigate crash involvement trends and determine whether differences exist based on age group and gender as well as between Appalachia and the rest of the state. Heat maps developed using county-level RAIR values visualized key findings. For two-unit crashes, young and older drivers are more at risk of being the at-fault driver in a crash than middle-aged drivers. In case of single-unit crashes, older drivers exhibit lower risk than young and middle-aged drivers. Overall, female drivers have lower at-fault risk behavior than male drivers. No evident regional disparities were apparent with respect to Appalachia or economic status. A weighted average RAIR was calculated for both single- and two-unit crashes. The heat maps developed using the weighted RAIR can be used to identify the top at-risk counties in the state that then can be targeted for safety programs such as the Kentucky Safety Circuit Rider Program [61]. The analysis developed here could aid the Program in identifying driver-related issues in addition to the roadway elements considered, thus developing a more robust approach to improving overall safety for the targeted counties.

To further investigate the association between crash occurrence and socioeconomic characteristics of driver's residence zip code, logistic regression was used. This modeling technique is beneficial when effects of more than one explanatory variable influence an outcome. The independent variables can be discrete and/or continuous, and the response variable is the probability of the outcome, which is modeled based on a combination of the predictor values. Using this technique and series of variable selection methods, several regression models for two- and single-unit crashes were developed as a function of several socioeconomic and demographic variables. The models in each category were then evaluated to finalize the ones with the best predictive power. The predictors for the final model were selected through a series of steps, including correlation analysis, recursive partitioning analysis, and stepwise selection. The model finalized through the process was then tested for interactions using the FSA tool. Three models were developed for single-unit and two-unit crashes. Each underwent several evaluation processes to identify the best model.

Model results for the single-unit and two-unit crashes were quite similar. For two-unit crashes, fault status was found to be a function of age group, gender, rurality, poverty level, average convictions, and driver population density. For single-unit crashes, all of these variables had a significant effect. However, poverty level was dropped from the model when educational attainment (percent with bachelor's degree or higher) was added. All the predictors in the final models were significant at the 95 percent confidence level.

The odds ratios for younger and older drivers showed they are more likely than drivers in other age groups to cause two-unit and single-unit crashes, thus following the typical U-shape curve of crash involvement. This is consistent with past research, which has shown a relationship between crash involvement and age. Aguero-Valverde et al. [11] concluded that age groups under 25 and over 65 have a positive association with crash risk, and most of the previous literature has found a positive association between young drivers and crashes or fatalities. Several studies on older drivers identified their increased crash involvement and demonstrated the risk factors they create for themselves and other drivers. Other studies have also noted that young and old drivers have a positive relationship with crash involvement, indicating their higher propensity to be the at-fault driver in a crash. These are consistent with the findings of this study.

Male drivers have higher at-risk probability when younger but become better drivers with experience. The reason for the high-risk rate could be the aggressive and risk-taking behavior of young male drivers. The exposure of male drivers is higher as they most likely drive more miles than females; this could be another reason for the higher involvement of young males. Female drivers are better drivers when young, while their performance changes as they age.

The following lists provide a quick summary of the key findings of the research and they can be used to develop targeted efforts (as suggested below) to address them. The findings are separated into two lists and are based on the analysis that they were derived from. The first list discusses the findings of the spatial analysis, while the second presents the findings based on the statistical analysis completed and models developed.

Spatial analysis

- Young drivers in two-unit crashes are more prevalent in populated areas and median household income does not play a role.
- Middle-aged drivers in two-unit crashes are more prevalent in lower income counties in the Appalachian region.
- Older drivers in two-unit crashes have higher crash involvement statewide.
- Female drivers are less likely to be involved in a two-unit crash statewide than males.
- Young drivers in single-unit crashes are more prevalent in higher income counties.
- Middle-aged drivers in single-unit crashes are more prevalent in lower income counties.
- Older drivers are less likely to be involved in single-unit crashes statewide.

Statistical analysis

- Marital status has significant effects on two-unit crashes with percent divorced/widowed/separated being negatively correlated to at-fault status.
- Individual and household income are negatively correlated to the at-fault status of the driver involved in single- as well as two-unit crashes.
- The probability of being at-fault in a two-unit crash increases when a driver resides in area with higher rates of poverty, rurality, population density, and number of convictions/1,000 drivers.
- For both male and female drivers, the at-fault probability in two-unit crashes is higher for young (<25 years) and older (>75 years) drivers
- The crash propensity for two-unit crashes is highest for < 20 males and it reduces with age. On the contrary, the propensity increases with age for female drivers. Above 65 years of age, male and female drivers contribute almost equally to crash occurrences.
- The probability of being at-fault in single-unit crashes increase when a driver resides in area with lower educational attainment and higher rates of rurality and population density.
- Female drivers are less likely to cause single-unit crashes than their male counterparts.

The logistic regression models developed here accomplish this. It is critical to determine whether the most at-risk drivers have particular demographic characteristics (e.g., age, gender) and where they reside (e.g., zip codes, urban or rural setting). The findings of this study will help practitioners identify groups of drivers with a high crash-involvement risk factor. Based on this knowledge, safety programs can be designed to more efficiently target the most at-risk groups. The target demographic can be given compulsory safety awareness classes for driver's license renewal. As a control measure to prevent crashes, drivers in at-risk groups can be issued severe penalties (such as license suspension or revocation) if found guilty of a traffic violation or being at fault in a crash.

The findings of this study are limited to two-unit and single-unit crashes. Even though this was a limitation, since it did not allow for a complete investigation of the entire crash database, the study nonetheless uncovered meaningful trends regarding the propensity of driver groups to cause a future crash. Relying on police-reported crashes, as well as crashes that go unreported, could lead to a bias in any safety study or analysis; however, this is unavoidable. Census data lack information on the population of drivers in each age-gender category and prevented the study from taking into account the exposure of drivers in each category. Also, this study was limited to the socioeconomic and demographic factors of the driver's residence zip code. Hence, the primary cause of crashes (e.g., geometric and environmental conditions) at crash locations were not considered. Also, crash severity was excluded from the study. Consideration of crash severity as a dependent variable could give some more insights into how the socioeconomic variables influence crash severity. This may be an objective for a future study.

References

1. World Health Organization. *Road Traffic Injuries*. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries. Accessed 11/18/2018.
2. NHTSA. *The Economic and Societal Impact of Motor Vehicle Crashes 2010*. National Highway Traffic Safety Administration. US Department of Transportation. Washington DC. 2015.
3. *Fatality Facts 2018*. Insurance Institute for Highway Safety, Highway Loss Data Institute, 2019. https://www.iihs.org/topics/fatality-statistics/detail/state-by-state.
4. Kentucky Office of Highway Safety,, Kentucky Transportation Cabinet, Frankfort, KY. *Daily Fatality Statistics Update*. https://transportation.ky.gov/HighwaySafety/Pages/default.aspx. Accessed 1/14/2020.
5. Brown, K. T. A Safety Analysis of Spatial Pehonomena About the Residences of Drivers Involved in Crashes. *Dissertation Presented to the Graduate School of Clemson University,* 2016.
6. Noland, R. B., and Oh, L. The Effect of Infrastructure and Demographic Change on Traffic-Related Fatalities and Crashes: A Case Study of Illinois County-Level Data. *Accident Analysis and Prevention,* 2004. 36(4), 525-532.
7. Adanu, E. K., Smith, R., Powell, L., and Jones, S. Multilevel Analysis of the Role of Human Factors in Regional Disparities in Crash Outcomes. *Accident Analysis and Prevention,* 2017. 109(10-17).
8. Factor, R., Mahalel, D., and Yair, G. Inter-Group Differences in Road-Traffic Crash Involvement. *Accident Analysis and Prevention,* 2008. 40, 2000-2007.
9. Hasselberg, M., Vaeza, M., and Laflamme, L. Socioeconomic Aspects of the Circumstances and Consequences of Car Crashes among Young Adults. *Social Science & Medicine,* 2005. 60(2), 287-295.
10. Zephaniah, S., Jr., S. J., Smith, R., and Weber, J. Spatial Dependence among Socioeconomic Attributes in the Analysis of Crashes Attributable to Human Factors. *Analytic Methods in Accident Research (under review),* 2018.
11. Aguero-Valverde, J., and Jovanis, P. P. Spatial Analysis of Fatal and Injury Crashes in Pennsylvania. *Accident Analysis and Prevention,* 2006. 38(3), 618-625.
12. Hanna, C. L., Laflamme, L., and Bingham, C. R. Fatal Crash Involvement of Unlicensed Young Drivers: County Level Differences According to Material Deprivation and Urbanicity in the United States. *Accident Analysis and Prevention,* 2012. 45, 291-295.
13. Kocatepe, A., Ulak, M. B., Ozguven, E. E., Horner, M. W., and Arghandeh, R. Socioeconomic Characteristics and Crash Injury Exposure: A Case Study in Florida Using Two-Step Floating Catchment Area Method. *Applied Geography,* 2017. 87, 207-221.
14. Stamatiadis, N., and Puccini, G. Fatal Crash Rates in the Southeastern United States: Why Are They Higher? *Transportation Research Board,* 1999. 1665), 118-124.
15. Maciag, M. *America's Poor Neighbohoods Plagued by Pedestrian Deaths*. Governing Magazine, State and Local Government News for America's Leaders. 2014. www.governing.com/gov-data/pedestrian-deaths-poor-neighborhoods-report.html.
16. NHTSA. *National Survey of Speeding Attitudes and Behaviors 2011*. National Highway Traffic Safety Administration. U.S. Department of Transportation. Washington DC. 2013.
17. Lee, J., Abdel-Aty, M., and Choi, K. Analysis of Residence Characteristics of at-Fault Drivers in Traffic Crashes. *Safety Science,* 2014. 68(0), 6-13.
18. Blatt, J., and Furman, S. M. Residence Location of Drivers Involved in Fatal Crashes. *Accident Analysis and Prevention,* 1998. 30(6), 705-711.
19. Chandraratna, S., Stamatiadis, N., and Stromberg, A. Potential Crash Involvement of Young Novice Drivers with Previous Crash and Citation Records. *Human Performance; Simulation And Visualization,* 2005. 1937), 1-6.
20. Chandraratna, S. K. Crash Involvement Potential for Drivers with Multiple Crashes.In *Civil Engineering, No. Doctoral*, University of Kentucky, 2004.
21. Ivan, J., Burnicki, A., Wang, K., and Mamun, S. *Improvemnts to Road Safety Improvement Selection Procedures for Connecticut*. 2016.

22. United Census Bureau. *American Census Survey*. https://www.census.gov/programs-surveys/acs/. Accessed 12/10/2017.
23. Harrah, J. *Kentucky Metropolitan Areas out-Perform Rural and Small Urban Areas*. http://crcblog.typepad.com/crcblog/kentucky-metropolitan-areas-out-perform-rural-and-small-urban-areas.html.
24. Noland, R. B., and Quddus, M. A. A Spatially Disaggregate Analysis of Road Casualties in England. *Accident Analysis and Prevention,* 2004. 973-984.
25. Muellerman, R. L., and Mueller, K. Fatal Motor Vehicle Crashes: Variations of Crash Characteristics within Rural Regions of Different Population Densities. *The Journal of Trauma: Injury, Infection, and Critical Care,* 1996. 41(2), 315-320.
26. Zwerling, C., Peek-Asa, C., Whitten, P. S., Choi, S.-W., Sprince, N. L., and Jones, M. P. Fatal Motor Vehicle Crashes in Rural and Urban Areas: Decomposing Rates into Contributing Factors. *Injury Prevention,* 2005. 11(1), 24-28.
27. Cook, L. J., Knight, S., and Olson, L. M. A Comparison of Aggressive and Dui Crashes. *Journal of Safety Research,* 2005. 36(5), 491-493.
28. Chen, H. Y., Ivers, R. Q., Martiniuk, A. L. C., Boufous, S., Senserrick, Woodward, M., Stevenson, M., and Norto, R. Socioeconomic Status and Risk of Car Crash Injury, Independent of Place of Residence and Driving Exposure: Results from the Drive Study. *Journal of Epidemiology and Community Health,* 2010. 64(11).
29. Males, M. A. Poverty as a Determinant of Young Drivers' Fatal Crash Risks. *Safety Research,* 2009. 40(6), 443-448.
30. Lyman, S., Ferguson, S. A., Braver, E. R., and Williams, A. F. Older Driver Involvements in Police Reported Crashes and Fatal Crashes: Trends and Projections *Injury Prevention,* 2002. 8(2), 116-120.
31. Adanu, E. K., Penmetsa, P., Jones, S., and Smith, R. Gendered Analysis of Fatal Crashes among Young Drivers in Alabama. *Safety,* 2018. 4(3), 29.
32. Sun, X., Das, S., and He, Y. Analyzing Crash-Prone Drivers in Multiple Crashes for Better Safety Educational and Enforcement Strategies. *Journal of Transportation Technologies,* 2014. 4(1).
33. Blasco, R. D., Prieto, J. M., and Cornejo, J. M. Accident Probability after Accident Occurrence. *Safety Science,* 2003. 41(6), 481-501.
34. National Highway Traffic Safety Administration. *Automated Vehicles for Safety*. https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety. Accessed 1/14/2020.
35. Greenwood, M., and Yule, G. U. An Inquiry into the Nature of Frequency Distributions Representative of Multiple Happenings with Particular Reference to the Occurrence of Multiple Attacks of Disease or of Repeated Accidents. *Journal of the Royal Statistical Society,* 1920. 83(2), 255-279.
36. Daigneault, G., Joly, P., and Frigon, J.-Y. Previous Convictions or Accidents and the Risk of Subsequent Accidents of Older Drivers. *Accident Analysis and Prevention,* 2002. 34(2), 257-261.
37. Chen, W., Cooper, P., and Pinili, M. Driver Accident Risk in Relation to the Penalty Point System in British Columbia. *Accident Analysis and Prevention,* 1995. 26(1), 9-18.
38. Torre, G. L., Beeck, E. V., Quaranta, G., Mannocci, A., and Ricciardi, W. Determinants of within-Country Variation in Traffic Accident Mortality in Italy: A Geographical Analysis. *International Journal Of Health Geographics,* 2007. 6(1), 49.
39. Rivas-Ruiz, F., Perea-Milla, E., and Jimenez-Puente, A. Geographic Variability of Fatal Road Traffic Injuries in Spain During the Period 2002–2004: An Ecological Study. *BMC Public Health,* 2007. 7(1), 266.
40. Chen, C., Zhang, G., Tian, Z., Bogus, S. M., and Yang, Y. Hierarchical Bayesian Random Intercept Model-Based Cross-Levelinteraction Decomposition for Truck Driver Injury Severity Investigations. *Accident Analysis and Prevention,* 2015. 85, 186-198.
41. Vachal, K. *Analysis of Risk Factors in Severity of Rural Truck Crashes*. Upper Great Plains Transportation Institute, North Dakota State University, Fargo. 2016.
42. Chen, F., and Chen, S. Injury Severities of Truck Drivers in Single- and Multi-Vehicle Accidents on Rural Highways. *Accident Analysis and Prevention,* 2011. 43, 1677-1688.

43. Khorashadi, A., Niemeier, D., Shankar, V., and Mannering, F. Differences in Rural and Urban Driver-Injury Severities in Accidents Involving Large-Trucks: An Exploratory Analysis. *Accident Analysis and Prevention,* 2005. 37(5).
44. Das, S., Sun, X., Wang, F., and Leboeuf, C. Estimating Likelihood of Future Crashes for Crash-Prone Drivers. *Journal of Traffic and Tranportation Engineering (English Edition),* 2015. 2(3), 145-157.
45. Abdalla, I. M., Raeside, R., Barker, D., and David R.D, M. An Investigation into the Relationships between Area Social Characteristics and Road Accident Casualties. *Accident Analysis and Prevention,* 1997. 29(5), 583-593.
46. *Kentucky's Traffic Collision Facts*. Kentucky State Police. 2016.
47. Cambron, A., Stamatiadis, N., Wright, S., and Sagar, S. *Mri 1: Effect of Socioeconomic and Demographic Factors on Kentucky Crashes*. Southeastern Transportation Center, UT Center for Transportation Research, 309 Conference Center Building, Knoxville TN 37996-4133. 2018.
48. Laerd Statistics. *Point-Biserial Correlation Using Spss Statistics*. https://statistics.laerd.com/spss-tutorials/point-biserial-correlation-using-spss-statistics.php. Accessed 5/15/19.
49. PennState Eberly College of Science. Recursive Partitioning. https://newonlinecourses.science.psu.edu/stat555/node/100/. Accessed 5/10/19.
50. NCSS Statistical Software. Stepwise Regression. Accessed.
51.Frost, J. Understanding Interaction Effects in Statistics. https://statisticsbyjim.com/regression/interaction-effects/. Accessed 6/10/19.
52. Lambert, J., Gong, L., Elliott, C. F., Thompson, K., and Stromberg, A. Rfsa: An R Package for Finding Best Subsets and Interaction. *The R Journal,* 2018. 10, 295-308.
53. Chandraratna, S., and Stamatiadis, N. Quasi-Induced Exposure Method: Evaluation of Not-at-Fault Assumption. *Accident Analysis and Prevention,* 2009. 2009(41), 308-313.
54. Kentucky Public Health. *2017 Kentucky Racial and Ethnic Distribution* https://chfs.ky.gov/agencies/dph/Documents/2017KYEthnic_Distribution.pdf. Accessed.
55. University of Kentucky. *Kentucky Geological Survey*. http://www.uky.edu/KGS/gis/bounds.htm. Accessed 2/12/2019.
56. United States Census Bureau. *Mapping Files*. https://www.census.gov/geographies/mapping-files.html. Accessed 2/23/2020.
57. Sagar, S., Stamatiadis, N., Wright, S., and Green, E. Use of Codes Data to Improve Estimates of at-Fault Risk of Elderly Drivers. Submitted to Accident Analysis and Prevention*,*
58. Baldock, M., Mathias, J., McLean, A. J., and Berndt, A. Self-Regulation of Driving and Its Relationship to Driving Ability among Older Adults. *Accident Analysis and Prevention,* 2006. 38(1036-1045).
59. *Fatality Facts 2018 - Gender*. Insurance Institute for Highway Safety, Highway Loss Data Institute. 2018. https://www.iihs.org/topics/fatality-statistics/detail/gender.
60. Frank Gross, Nabors, D., Eck, R., and Hood, M. *Safety Ciruit Rider Programs Best Practices Guide*. Federal Highway Administration Office of Safety. 2009. https://safety.fhwa.dot.gov/local_rural/training/fhwasa09019/fhwasa09019.pdf.
61. Safety Circuit Rider Program. University of Kentucky, College of Engineering*,*
62. Stamatiadis, N., and Deacon, J. A. Quasi-Induced Exposure: Methodology and Insight. *Accident Analysis & Prevention,* 1997. 29(1), 37-52.
63. Staplin, L., Lococo, K. H., Stewart, J., and Decina, L. E. *Safety Mobility for Older Drivers Handbook*. National Highway Traffic Safety Administration, Washington, D.C. 1999.