Real-World Applications for Analytics Teaching and Learning

Anita Lee-Post
University of Kentucky, Anita.Lee-Post@uky.edu

Right click to open a feedback form in a new tab to let us know how this document benefits you.
To my husband, C.A. Post,

for his example of

integrity, generosity, and resourcefulness
Preface

I wrote *Real-World Applications for Analytics Teaching and Learning* as a solution to the limited availability of open educational resources (OERs) on Analytics software skill development, in particular, skills in cleansing data, dealing with big data, building dynamic data visuals such as dashboards, and text mining. SAS JMP became our Analytics software of choice because its core capabilities align well with developing the aforementioned Analytics skillsets. However, the marketplace is dominated by textbooks using Excel and/or R but not SAS JMP. Therefore, my primary goal in creating this OER is to provide materials using SAS JMP for Analytics skill development.

Learning-by-doing has been demonstrated to be effective in teaching and learning software skills. As such, I incorporated a learning-by-doing pedagogy to elevate students’ learning experience from passive knowledge transfer to active knowledge exploration and discovery in this OER. Specifically, the OER is created to integrate Analytics concepts and techniques with real-world scenarios to illustrate how real data can be transformed into actionable insights to support decision making. Guiding questions are used to engage students to think critically, relate concepts to the given situation, reason through a problem and come up with their own answers/solutions. The COVID-19 pandemic provides a current and realistic real-world context for this OER. The URL references cited in the OER are accurate at the time of writing. In the future, updates to the URLs will be made as needed.

This OER is distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (https://creativecommons.org/licenses/by-nc-sa/4.0/), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided that the author and publication source are credited, that changes (if any) are clearly indicated, and that the derivative work is distributed under the same license.

I welcome instructors in Analytics to adopt the OER in your class. To access the OER, instructors can use this digital object identifier: https://doi.org/10.13023/msc.textbook.01. I also have a set of teaching notes that can be shared upon instructors’ requests. If you are an instructor and are interested in receiving the teaching notes and/or collaborating with me to further develop OERs on Analytics, please use this link to provide me with your contact information: https://uky.az1.qualtrics.com/jfe/form/SV_6RUsuyRHXck9zbo

Last but not least, I would like to thank the University of Kentucky Libraries for providing me funding support with the Alternative Textbook Grant.

Dr. Anita Lee-Post
Gatton College of Business and Economics
University of Kentucky
Introduction

Since the coronavirus was first confirmed in Wuhan, China in December of 2019, the disease was named COVID-19 and declared a pandemic in March of 2020 as it spread to every continent infecting 175 million and killing 4 million people worldwide as of June 2021. In the U.S., the number of confirmed cases grew from 1 to over 33 million with more than 600,000 deaths in an 18-month period beginning in January of 2020. To help understand and mitigate the spread of the disease, research communities (e.g., Johns Hopkins Coronavirus Resource Center, University of Maryland COVID-19 Impact Analysis Platform) and government agencies (e.g., Center for Disease Control and Prevention, World Health Organization) have collected and made available massive amounts of data for public access and sharing.

As a health services research analyst, you have the Analytics knowledge and skill to make sense of the data to generate insights and offer decision support for COVID-related issues. In particular, you are interested in the following two aspects of COVID-19 interventions to prevent and control the disease. First, for the interventions to be effective, the risk factors that associate strongly with COVID-19 mortality must be identified. Knowing the risk factors can help individuals understand their COVID-19 mortality risks so that they can behave in ways that lower their likelihood of being infected (e.g., practice social distancing, wear a mask, avoid crowds, get vaccinated). The risk factors can also inform health-care providers to make appropriate treatment and resource allocation decisions to save lives. Moreover, policymakers...
can perform risk assessment and prediction based on the risk factors to identify vulnerable population for targeted use of such interventions as stay-at-home, viral testing, or contact tracing to stem the infection/transmission rate. Furthermore, the ability to stratify COVID-19 risk can help government prioritize vaccination efforts to protect our most vulnerable population first.

Second, none of the interventions can succeed without the full cooperation of people behaving responsibly, i.e., in compliance with the protection and preventive practices as stipulated in the intervention. Our country-wide lockdown will have no effect on slowing the spread of COVID-19 if people defy the stay-at-home directive. Similarly, the reopening plans will have to be paused if people ignore effective and low-cost practices in public such as social distancing and mask wearing, causing a resurgence of the virus. Likewise, the effort to return closer to pre-pandemic normal (e.g., lifting restrictions on businesses, large gathering, and travel) will be hampered because of vaccine hesitancy. It is, therefore, essential that we take responsibility for our behavior to protect ourselves from being infected and at the same time reduce the risk of infecting others. In short, an understanding of how to influence individuals to behave responsibly is critical to the success of any intervention.

You have planned five studies focusing on the above two areas (i.e., identify risk factors, and influence individuals’ behaviors) to impact the success of interventions to prevent and control the disease. These five studies are: (1) characterize COVID-19 mortality demographic risk factors, (2) visualize COVID-19 mortality demographics, (3) conduct COVID-19 mortality time series forecasting, (4) predict COVID-19 mortality, and (5) analyze COVID-19 vaccine acceptance, uptake, and experiences.

The first four studies are concerned with COVID-19 mortality risk factor analyses with the use of individual cases of COVID-19 data. Results from the analyses will help define the demographic characteristics of the population most vulnerable to dying from the disease. They also help track changes to COVID-19 deaths over time, compare COVID-19 deaths by the demographic characteristics, examine the relationships between COVID-19 deaths and the risk factors, and make predictions of COVID-19 deaths. More importantly, the analyses will facilitate individuals, healthcare providers, and policymakers making evidence-based decisions to best prevent and control the disease.

The last study is about acquiring, cleansing, modifying, and analyzing COVID-19 vaccination-related data to generate insights about vaccine acceptance. Such insights will further our understanding of individuals’ attitudes towards the vaccine to help address vaccine hesitancy concerns, improve vaccine distribution coordination, and expedite vaccination efforts to reach population-level immunity to COVID-19.

You have structured each study with a specific goal, related tasks, and expected results. Guiding questions are used to help critical thinking and reasoning through the Analytics process from start to finish to accomplish the goals. The details of the five studies are described below.
Study 1: Characterize COVID-19 Mortality Demographic Risk Factors

According to the Centers for Disease Control and Prevention (CDC) and research studies in Nature and Clinical Infectious Disease, demographic risk factors for COVID-19 deaths include:

1. Age
2. Race/ethnicity
3. Gender

Your goal is to characterize COVID-19 mortality demographic risk factors to facilitate subsequent analyses (e.g., visualize the relationship between COVID-19 deaths and the risk factors, make predictions of COVID-19 deaths). Specifically, you are to perform the following data preparation and preliminary analysis tasks:

1. Acquire relevant data on COVID-19 deaths and the demographic risk factors mentioned above.
2. Cleanse the data from duplications, errors, missing values, and outliers.
3. Modify the data so that they are suitable for achieving the goal.
4. Prepare a summary of the demographic risk factors of COVID-19 mortality after a preliminary analysis of the modified data.

Results from the study can provide evidence to support the association between an individual’s demographic characteristics and their COVID-19 mortality risk (e.g., a 60-year-old Hispanic female has a higher risk of dying from the disease than a 40-year-old non-Hispanic male).
Task 1: Data Acquisition
Answer the following questions to acquire the right data for COVID-19 mortality demographic risk factor characterization:
   1. What type of data source is appropriate?
   2. Where is the data source located?
   3. What data is available from the source identified?
   4. How do you acquire the data from the source identified?
   5. What data is relevant from the source identified?

Task 2: Data Cleansing
Answer the following questions to cleanse the data acquired from Task 1 from relevant anomalies:
   1. What anomalies are detected and corrected during data cleansing?
   2. Should all the anomalies identified above be cleansed in this case? Explain.
   3. How do you perform the data cleansing that is needed in this case?

Task 3: Data Modification
Answer the following questions to prepare the data cleansed in Task 2 for COVID-19 mortality demographic risk factor characterization:
   1. What kind of data modification is needed to ensure the data cleansed is suitable for achieving the goal?
   2. How do you perform the data modification that is needed in this case?

Task 4: Preliminary Analysis
Answer the following questions about conducting a preliminary analysis of the data modified in Task 3:
   1. What are the demographic risk factors of COVID-19 mortality?
   2. What kind of preliminary analysis is needed in this case?
   3. How do you perform the needed preliminary analysis in JMP?
   4. What conclusions can be drawn from the preliminary analysis?
Study 2: Visualize COVID-19 Mortality Demographics

Building on the results from data preparation and preliminary analysis in Study 1, your goal is to continue analyzing COVID-19 mortality data with the use of data visualization techniques. Specifically, you are to perform the following data visualization tasks:

1. Visualize the distribution of COVID-19 deaths by gender, age, and ethnicity individually.
2. Visualize the comparisons of COVID-19 deaths by gender, age, and ethnicity.

Results from the study can highlight trends of COVID-19 deaths (e.g., the number of COVID-19 deaths increases exponentially from March to July 2020), occurrences of COVID-19 deaths within a demographic risk factor (e.g., Asians have a higher risk of dying from the disease than non-Asians), and occurrences of COVID-19 deaths across gender, age, and ethnicity (e.g., more Black males over 60 years old have died from COVID-19 than non-Black females under 60 years old in the past six months).

Task 1: COVID-19 Mortality Demographics Distributions

Answer the following questions using the data modified in Task 3 of Study 1 to visualize the distributions of COVID-19 deaths by the three COVID-19 mortality demographics (i.e., gender, age, and ethnicity) one at a time:

1. What is the date type for each COVID-19 mortality demographic?
2. What are the methods you would use to visualize each demographic individually?
3. Do you need to modify the data prior to performing this visualization task?
4. How do you perform this visualization task in JMP?
5. What can be concluded from this visualization task?
Task 2: COVID-19 Mortality Demographics Comparison
Answer the following questions using the data modified in Task 3 of Study 1 to compare COVID-19 mortality by demographics:

1. What are the data visualization methods you would use to compare COVID-19 deaths by two or more COVID-19 mortality demographics?
2. Do you need to modify the data for this visualization task?
3. How do you perform this visualization task in JMP?
4. What conclusions can be drawn from the resulting visualization?

Task 3: COVID-19 Mortality & Demographics Over Time
Answer the following questions using the data modified in Task 3 of Study 1 to visualize COVID-19 mortality and demographics over time:

1. What are the methods you would use to visualize COVID-19 mortality & demographics over time?
2. Do you need to modify the data for the above visualization task?
3. How do you perform the above visualization task in JMP?
4. What conclusions can be drawn from the resulting visualization?
Study 3: Predict COVID-19 Mortality Using Time Series Forecasting


Building on the results from data preparation and preliminary analysis in Study 1, as well as data visualization of COVID-19 deaths over time in Study 2, your goal is to predict COVID-19 mortality by projecting the underlying time series data pattern into the future. The ability to predict COVID-19 deaths based on time is instrumental for individuals, healthcare providers, and policymakers to make temporal-based decisions to best prevent and control the disease (e.g., when to lift COVID-19 restrictions). Specifically, you are to perform the following time series forecasting tasks:

1. Modify the data acquired and cleansed in Studies 1 and 2 for time series forecasting.
2. Perform time series forecasting for data with a level or sudden shifting pattern.
3. Perform time series forecasting for data with a trend and/or seasonal pattern.

Results of the study can confirm the best time series forecasting method for COVID-19 mortality predictions over time.
Task 1: Data Modification
Answer the following questions to prepare the data modified in Task 3 of Study 2 for time series forecasting:

1. What are the underlying data patterns that can be discovered with time series forecasting?
2. What are the data requirements for time series forecasting?
3. Do you need to modify the data prior to making time series forecasts on COVID-19 mortality?
4. How do you perform the needed data modification in JMP?

Task 2: Time Series Forecasting for Data with a Level/Sudden Shifting Pattern
Various smoothing methods are useful to make forecasts for time series data with a level/sudden shifting pattern. In particular, (1) Naïve, (2) Simple Average, (3) Moving Average of small k, and (4) Exponential Smoothing with a large alpha are appropriate for tracking recent data movements. Note that Moving Average with k=1 is equivalent to Simple Average, and Exponential Smoothing with alpha=1 is equivalent to Naïve. For the sake of simplicity, you will focus on (1) Naïve and (2) Simple Average as the competing methods in this case. To compare the performance of Naïve and Simple Average, you will use (1) Cumulative Forecast Error and (2) Mean Forecast Error. To compare the accuracy of the two smoothing methods, you will use (1) Mean Absolute Error, (2) Sum Squared Error, and (3) Mean Absolute Percentage Error. To select the best smoothing method for time series forecasting, you use forecast error-based metrics to evaluate competing methods’ performance (e.g., cumulative forecast error) and/or accuracy (e.g., mean absolute error). The best method is the one with the least biased and/or smallest value of the metrics. The best method can then be used to make COVID-19 mortality predictions. As such, Task 2 can be further broken down into the following:

2. Generate simple average forecasts on COVID-19 mortality.
3. Select the best method using the appropriate forecast error metrics.
4. Use the best method to make COVID-19 mortality predictions.

Task 2_1: Naïve Forecasting Method
Answer the following questions to perform naïve forecasting on COVID-19 mortality using the appropriate modified data from Task 1:

1. How do you generate naïve forecasts in JMP?
2. How do you compute naïve forecast errors in JMP?
3. How do you generate naïve forecast metrics in JMP?

Task 2_2: Simple Average Forecasting Method
Answer the following questions to perform simple average forecasting on COVID-19 mortality using the appropriate modified data from Task 1:

1. How do you generate simple average forecasts in JMP?
2. How do you compute simple average forecast errors in JMP?
3. How do you generate simple average forecast metrics in JMP?
Task 2_3: Time Series Forecasting Method Selection
Answer the following questions to select the best time series forecasting method for COVID-19 mortality forecasting with the use of the appropriate metrics from Tasks 2_1 and 2_2:
1. Based on the performance metrics, which is the best time series forecasting method?
2. Based on the accuracy metrics, which is the best time series forecasting method?

Task 2_4: Time Series Forecasting Prediction
Answer the following questions to predict COVID-19 mortality using the best time series forecasting method identified in Task 2_3:
1. According to performance, which time series forecasting method is the best?
2. Based on the best performing method, what is the predicted value of the COVID-19 mortality for the next period?
3. According to accuracy, which time series forecasting method is the best?
4. Based on the most accurate method, what is the predicted value of the COVID-19 mortality for the next period?

Task 3: Time Series Forecasting for Trend and/or Seasonality
Answer the following questions to perform time series forecasting for trend and/or seasonality on COVID-19 mortality using the appropriate data modified in Task 1:
1. How do you generate time series forecasts for seasonality in JMP?
2. How do you generate time series forecasts for trend in JMP?
3. How do you generate time series forecasts for trend and seasonality in JMP?
4. How do you select the best in this case?
5. How do you use the best method to predict COVID-19 mortality in the next period?
Study 4: Predict COVID-19 Mortality Using Linear Regression

Building on the results from data preparation and preliminary analysis in Study 1, and data visualization of COVID-19 mortality demographics in Study 2, your goal is to predict COVID-19 mortality by quantifying the relationship between COVID-19 mortality and one or more risk factors with the use of Linear Regression. The ability to predict COVID-19 deaths based on risk factors is valuable for individuals, healthcare providers, and policymakers to make demographic-related decisions to best prevent and control the disease (e.g., prioritize vaccination efforts to...
protect our most vulnerable population first). Specifically, you are to perform the following Linear Regression tasks:

1. Modify the data so that they are suitable for regression modeling.
2. Construct various Linear Regression models using a combination of the risk factors as independent variables to explain COVID-19 mortality.
3. Estimate the Linear Regression equation for each model.
4. Select the best fit model using the appropriate “goodness-of-fit” measure.
5. Use the best fit model to make COVID-19 mortality predictions.

Results of the study can determine the best Linear Regression model to predict the number of COVID-19 deaths based on demographic risk factors.

**Task 1: Data Modification**

Answer the following questions to prepare the data modified in Task 3 of Study 2 for Linear Regression analysis:

1. What are the data requirements for Linear Regression?
2. What kind of data preparation is needed to ensure the data cleansed is suitable for Linear Regression analysis?
3. How do you use JMP to perform the data modification that is needed in this case?

**Task 2: Linear Regression Modeling**

Answer the following questions regarding Linear Regression modeling using the data modified in Task 1:

1. What is the dependent variable?
2. What are the independent variables?
3. How many Linear Regression models can be formulated?
4. What are the possible Linear Regression models relating the dependent and their respective independent variable(s)?

**Task 3: Linear Regression Equation Estimating**

Answer the following questions regarding Linear Regression equation estimating using the data modified in Task 1:

1. Which of the models identified in Task 2 is a simple regression model?
2. Which of the models identified in Task 2 is a multiple regression model?
3. How do you use JMP to estimate the Linear Regression equation for each of the models identified in Task 2?

**Task 4: Best-fit Model Selection**

Answer the following questions to select the best-fit model based on the models generated in Task 3:

1. What is the goodness-of-fit measure that you should use to select the best-fit model?
2. Which is the best-fit model? Explain.
3. How do you select the best-fit model in JMP?
4. What conclusions can be drawn about the best-fit model?
Task 5: COVID-19 Mortality Prediction

Answer the following questions to predict COVID-19 mortality using the best-fit model identified in Task 4:

1. According to the best-fit model, what is the dependent variable?
2. According to the best-fit model, what are the independent variables?
3. What is the regression equation of the best-fit model?
4. What are the forecasting methods you can use to specify the values of the independent variables in the next month?
5. Use the simple average method to determine the values of the independent variables in the next month.
6. Based on the values of the independent variables determined with the simple average method, what is the predicted value of the dependent variable?
Study 5: Analyze COVID-19 Vaccine Acceptance, Uptake, and Experiences

Total Doses Administered Reported to the CDC by State/Territory and for Select Federal Entities per 100,000 of the Total Population

Understanding people’s attitude towards COVID-19 vaccine acceptance is central to achieving the goal of stopping the pandemic by inoculating enough people to reach herd immunity to COVID-19. To help gain insights into the implications of COVID-19 vaccine acceptance in the
U.S., you are charged with analyzing COVID-19 vaccination data with the use of (1) cluster analysis to discover geographic grouping of vaccine acceptance, (2) association rule analysis to express the likelihood of co-occurrences of vaccine acceptance and uptake, and (3) text mining to summarize Americans’ vaccine experiences. Specifically, you are to perform the following tasks:

1. Acquire relevant data on COVID-19 vaccine acceptance, uptake, and experiences.
2. Cleanse the data for cluster analysis, association rule analysis, and text mining.
4. Express the likelihood of co-occurrences of vaccine acceptance and uptake.
5. Summarize Americans’ vaccine experiences.

Results from Cluster Analysis can help answer questions such as whether there are geographical differences in attitudes towards vaccination and which locations need to increase efforts to overcome vaccine hesitation. Results from Association Rule Analysis help reveal what level of attitude towards vaccination associates most strongly with vaccine uptake across locations and whether there are temporal differences in the associations. Results from Text Mining help identify words that are frequently used to describe the vaccination experience and help distinguish between positive and negative perceptions of the experience.

**Task 1: Data Acquisition**

Answer the following questions to acquire the relevant data to analyze COVID-19 acceptance, uptake, and experiences:

1. What type of data source is appropriate?
2. Where is the data source located?
3. What data is available from the source identified?
4. How do you acquire the data from the source identified?
5. What data is relevant from the source identified?

**Task 2: Data Cleansing**

Answer the following questions to cleanse the data acquired in Task 1 from relevant anomalies:

1. What kind of data cleansing is needed for the vaccine acceptance data?
2. How do you perform the data cleansing that is needed for the vaccine acceptance data?
3. What kind of data cleansing is needed for the vaccine uptake data?
4. How do you perform the data cleansing that is needed for the vaccine uptake data?
5. What kind of data cleansing is needed for the vaccine experience data?
6. How do you perform the data cleansing that is needed for the vaccine experience data?
Task 3: Cluster Analysis
Answer the following questions using the appropriate cleansed data from Task 2 to discover geographic groupings of COVID-19 vaccine acceptance:

1. What is the method you would use to discover geographic groupings of COVID-19 vaccine acceptance?
2. Do you need to modify the data for the above task?
3. How do you perform the above task in JMP?
4. What conclusions can be drawn from the resulting analysis?

Task 4: Association Rule Analysis
Answer the following questions using the appropriate cleansed data from Task 2 to discover rules that associate vaccine acceptance and uptake:

1. What is the method you would use to discover rules that associate vaccine acceptance and uptake?
2. Do you need to modify the data for the above task?
3. How do you perform the above task in JMP?
4. What conclusions can be drawn from the resulting analysis?

Task 5: Text Mining
Answer the following questions using the appropriate cleansed data from Task 2 to discover useful information from articles about vaccine experiences:

1. What is the method you would use to discover useful information from articles describing Americans’ vaccine experiences?
2. Do you need to modify the data for the above task?
3. How do you perform the above task in JMP?
4. What conclusions can be drawn from the resulting analysis?