STOR 538 Sports Analytics

Practice 1

Objectives:

- Data Cleaning: Demonstrate You Can Carefully Clean Data Based Off Instructions
- Data Visualization: Demonstrate You Can Produce Visuals of Data

Introduction:

I fully expect you to complete this entire assignment by yourself without the help of another person. The work you submit should be your own, but you can use the internet to help you find solutions. You will be required to either show all your code and/or provide step-by-step instructions on how you completed the various tasks. If there is evidence that you either submitted someone else's work or failed to submit work that illustrates how you achieved certain output, expect your instructor to pursue an Honor Code violation.

For this assignment, you will be cleaning and analyzing a dataset called **GameStats**. This dataset contains statistics from NCAA football games from 2015 to 2019. The first thing you need to notice about this dataset is that every football game has two rows. One row that contains stats for the home team and one row that contains stats for the away team. The @ symbol in the 4th column indicates if the school is the home team or away team. To understand this, examine the 48th and 49th row, where there was only one football game between Ohio State and Virginia Tech.

Now, read the next section to find out all of the tasks I want you to complete for this dataset.

Assignment:

- Data Cleaning: There are 7,361 rows in this dataset. I expect you to clean the data following the steps below using code in some programming language.
  a. You need to figure out how to convert this dataset where there are two rows for each game to one row for each game.
  b. I want you to remove all games that are played on a neutral field. This means there is no home team and no away team. You can identify these games based off the value "N" in the 4th column.
  c. In your cleaned dataset, the first three variables need to be *Date, Home, Away* containing the date of the game, name of the home team, and name of the away team for each individual game.

  d. In your cleaned dataset, the fourth variable should be called *HomeWins* and this variable should be binary with a 1 indicating the home team won and a 0 indicating the home team lost.

  e. In the original dataset, *PassCmp* to *TotalTO* are all team statistics. In the cleaned dataset, there should be two variables for each team statistic. I want to create a variable starting with "H" for the home team, and a variable starting with "A" for the away team. For example, in your cleaned dataset, I want the fifth variable to be *HPassCmp* (Passing completions for the home team) and the sixth variable to be *APassCmp* (Passing completions for the away team). Following this pattern, the next variable in your cleaned dataset should be *HPassAtt,* then *APassAtt*, then *HPassPct*, then *APassPct*¸ then … you should get the point.

You need to submit your cleaned dataset as a CSV file. The title of this submitted file should be "CleanedGameStats.csv".

If you plan on using R, Python, or another programming language, you also need to submit a PDF file containing all of your code with comments explaining what each line of code is doing. This code should start with importing the original dataset and should end with writing the cleaned dataset to your computer as a CSV. The title of this submitted PDF should be "DataCleaning.pdf"

If you plan on doing this without using R or Python, you should write enough paragraphs that explains every single step you took to clean the dataset according to my instructions. You should submit your explanation of all of your steps with the data in a PDF. This should be written up professionally. If you choose to not use a statistical programming language such as R or Python, I should be able to read everything you wrote and accomplish what was asked for. The title of this submitted PDF should be "DataCleaning.pdf"

- Data Visualization: After you clean the data so each game is one row, I want you to create three visuals that summarize information from this dataset. In each of the visuals, make sure you use titles and axis labels and make sure everything is readable for each picture.

  a. Visual #1: Pick one NCAA conference and show side by side boxplots of one variable for all the schools in the conference that you chose. You may need to use Google to find all of the schools in the conference you choose.

  b. Visual #2: Create TWO new variables using subtraction to compare the home team to the away team. For example, if we had home points and away points, I could subtract one from the other to find the difference in points. Once you create your two variables, create a scatterplot with one variable on the y-axis and the other on the x-axis. Also, fit a linear regression, and add the linear regression line to the plot. I should see the raw data and the fitted linear regression model.

c. Visual #3: Pick a different NCAA conference (not the one in Visual #1) and summarize one variable using a statistic for each school in that conference. Then, use a bar plot to plot the summary statistic for each school in the conference and order the bar plots either in ascending or descending order. For example, I could summarize the average number of home points for each school in the ACC, and then use a bar plot to plot the average number of points for each school, and then reorder the bars so the school that scores the fewest points on average when home is on the left and the school that scores the highest points on average when home is on the far right.

If you plan on using R, Python, or another programming language, you need to submit a PDF file containing all of your code and output of these three visuals. Make sure you follow all of my instructions or you could lose a significant number of points. Your PDF should contain exactly three visual and the code you submit should work. The title of this submitted PDF should be DataVisualization.pdf".

If you plan on doing this without using R or Python, you should still submit a PDF with the three visuals and write enough paragraphs that explains every single step you took to create the three visuals in the software you used. The title of this submitted PDF should be DataVisualization.pdf".

If you fail to clean the dataset appropriately, you will not be able to do this part, and you will lose a significant number of points if you submit a bunch of visuals when the clean data required had issues or significant problems.

Submission:

You will need to submit multiple things to get full credit on this assignment. If you submit any visuals, results, output, etc. without providing code or a detailed summary of how you achieved the visuals, results, output, etc., then you will be treated as if you didn't accomplish the task and possibly be targeted for an Honor Code Violation. Pay attention to the bullet points below. You are submitting exactly three things.

- You should submit the cleaned data as a CSV file named "CleanedGameStats.csv"
- You should submit a PDF file named "DataCleaning.pdf" that contains all of your commented code and/or an explanation of your work that explains all of the steps it took to clean the original dataset named **GameStats**
- You should submit a PDF file named "DataVisualization.pdf" that contains all of the code (or explanation) and output of the three visuals that you were required to create. I only want to see three visuals. If you submit more than three visuals, you will lose points.

Rubric:

The table below shows how many points are allocated to each item along with some explanation or reminder of key points. You will lose points for not following the detailed instructions given above.

Criteria	**Points**
Data Cleaning: Code or Detailed Explanation	8 Points (*Working code with a lot of comments or detailed explanation*)
Data Cleaning: Submitted CSV file	4 Points (*Data should be cleaned correctly*)
Data Cleaning: Submitted PDF file	2 Points (*Appropriately named)*
Data Visualization: Boxplots (1)	4 Points *(All Schools from One Conference and side-by-side boxplots, one boxplot per school in conference)*
Data Visualization: Scatterplot (2)	6 Points *(Create two "differenced" variables, scatterplot, and linear regression line)*
Data Visualization: Barplot (3)	8 Points *(Summarize one variable of your choice using a statistic of your choice like the mean, barplot where height of bar represents the summary statistic of chosen variable for each school in a different conference, ascending or descending order)*
Data Visualization: Submitted PDF file	2 Points (*Appropriately named)*