R for Business Analytics Final Project Guidelines

Total Points: 200

Group Project (Groups of 3 or 4)

Project Objective: The purpose of the final project is to implement the techniques and information learned in the class to a real-world data set. The final project need not be an innovation of your own. But it should involve taking an untidy data set, tidying it, performing EDA, and running some statistical inferences.

Phase 1

Step 1: Identify a raw data set where you have some tidying to do and there are opportunities for performing data analysis. Tidying does not mean that you need to use all the tidying techniques that the course covers. Instead, the data set should involve usage of at least three tidying techniques. Data analysis includes exploratory data analysis (Week 7), creating visualizations (Week 9), and drawing statistical inferences (Week 10).

You can obtain the raw data of your choice – either from your professional setting (if it is not proprietary) or from some other source that you know. There are many websites that provide open-source data sets that you can use.

Step 2 *(25 points)*: Introduction to the data

1. Write an introduction to the data set.
2. State where the data was obtained from – a hyperlink to the source of the data set.
3. Why did you choose this data set?

Step 3 *(25 points)*: Data Preparation

1. Import the data into `R`. Is this data tidy? If not, which principles of tidy data does your data set violate? Please do not worry about cleaning the data yet. You will perform data clean up that in step 4.
2. List the variables and their data types.
3. Examine the structure of the data set.
4. Show the dimensions of the data set.
5. Show the first 6 rows and the last 6 rows of the data set.
6. Identify how many missing values are in your data set.
7. How many missing values are there in each column?
8. Summarize what cleaning/transforming needs to be done.

Phase 2

Step 4 *(50 points)*: Clean and Tidy the data Apply the principles of tidy data and clean the data. Note: This can entail some of the following (not everything listed below is required):

- Performing any coercions (such as changing the variable data type to integer/numeric/factor.)
- Using the `tidyr` package or the `dplyr` package in `R`.
- Using basic imputation techniques to deal with missing values.
- Apply other ways to tidy the data like splitting and uniting columns.
- Any other techniques covered or not covered in the course.

Step 5 *(30 points)*: Perform Exploratory Data Analysis (EDA)

1. Run the summary of the variables in the data set.
2. Learn about the data visually by plotting whichever plot from the list below applies:
   a) Bar plot
   b) Histogram
   c) Box plot
   d) Scatter plot
3. Summarize your findings from the various plots and tables.

4. Uncover new information in the data that is not self-evident (i.e. besides plotting the data as above, slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information).

Step 6 *(35 points)*: Inferences Apply some inferential statistics techniques covered in Week 10 and in your statistics courses.

Step 7 *(35 points)*: Predictive Analytics/Modeling Apply some predictive analytics/modeling techniques covered in Week 10 and in your statistics courses.