**DATA 180: INTRODUCTION TO DATA SCIENCE**

**FALL 2023**

**Final (40 points)**

**Due December 15th at 11:59 pm EST.**

**This is a programming only assignment. Answer questions in both code and writing.**

**Workflow for final**

* **Create a new branch for the final project**
  + **Final\_Exam\_William\_Ferris**
* **Create r script ( .r file ) or R markdown file (.rmd) and store it in your local github repository**
* **Work on final**
* **Save file**
* **Commit to branch**
* **Create pull request**

For this exam we will be working with a real-world data set for data visualization, data wrangling, and creating summary statistics. You are expected to use base R or any modules we have covered in class to answer the questions in both code and writing. You can use either R script or a Markdown file to complete the assignment but make sure your code runs before you submit your assignment.

Here is a description of the data: Financial institutions that lend to consumers rely on models to help decide on who to approve or decline for credit (for lending products such as credit cards, automobile loans, or home loans). In this task, you are to use the skills we have learnt in class to understand this data. You are given historical data containing one response (binary) and 20 predictor variables from credit card accounts for a hypothetical bank XYZ.

Use the data set found [here](https://github.com/KennedyOdongo/DATA-180-Introduction-to-Data-Science--Section-2/blob/main/data/loan_default_data_set.csv) on the course website to answer the following questions. You can find a code book of the data [here](https://github.com/KennedyOdongo/DATA-180-Introduction-to-Data-Science--Section-2/blob/main/data/Appendix%20and%20data%20description.pdf) .

1. **Data wrangling:**
   1. What is the dimension (shape) of the dataset? How many rows and columns does the data set have?
      1. 20,000 rows and 21 columns
   2. Report the column names of the data set.
      1. tot\_balance, avg\_bal\_cards, credit\_age, credit\_age\_good\_account, credit\_card\_age, num\_acc\_30d\_past\_due\_12\_months, num\_acc\_30d\_past\_due\_6\_months, num\_mortgage\_currently\_past\_due, tot\_amount\_currently\_past\_due, num\_inq\_12\_month, num\_card\_inq\_24\_month, num\_card\_12\_month, num\_auto\_.36\_month, uti\_open\_card, pct\_over\_50\_uti, uti\_max\_credit\_line, pct\_card\_over\_50\_uti, ind\_XYZ, rep\_income, rep\_education, Def\_ind
   3. Which types of data are there in the dataset? Numeric, categorical, ordinal?
      1. When using the str() function, it tells me that there is mainly numeric data like ‘credit\_age’ and ‘tot\_balance’, and just one of the columns is categorical, ‘rep\_education.’
   4. Which columns contain missing values and how much (what percent) of those columns are missing?
      1. The columns missing values are: ‘pct\_card\_over\_50\_uti,’ and ‘rep\_income.’
      2. 9.79 percent of ‘pct\_card\_over\_50\_uti’ is missing and 7.795 percent of ‘rep\_income’ is missing
   5. How do you think we should deal with missing values?
      1. You can drop all the rows with missing values.
      2. You can add in predicted values into the missing spots.
      3. You can note where there is missing values in the data to account for the blank space. In a way, work around them.
      4. You can put in information from similar records/sources
   6. With this data, would you fit a supervised or an unsupervised learning model? Why?
      1. Supervised learning because we have a label on all the data and a goal to determine whether someone is worthy of approving or declining of credit.
   7. For part 2 and 3 drop all rows of the data that contain missing values. Print the dimensions of the resulting data set that has no missing values.
      1. 16653 rows and 21 columns
2. **Data summary statistics:**
   1. Find the summary statistics of the data set. You can use the summary function from dplyr.
   2. Based on the mean, mode, and median, is *“num\_card\_inq\_24\_month”* bell shaped, left, right skewed? How about “*tot\_amount\_currently\_past\_due”?* *“credit\_age”?* 
      1. For *num\_card\_inq\_24\_month:* 
         1. The mode is 0, the median is 0 and the mean is 1.053.
         2. Based on this info, since the mean is larger than the mode and median, I can assume that this graph will be right skewed.
      2. For *tot\_amount\_currently\_past\_due:*
         1. The mode is 0, the median is 0, and the mean is 352.5.
         2. Based on this info, since the mean is larger than the mode and the median, I can assume this is right skewed.
      3. For *credit\_age:*
         1. The mode is 295, the median is 280, and the mean is 280.7.
         2. Based on this info, because the mean, median, and mode are so close, I can assume credit\_age is bell shaped.
   3. Plot a histogram of the variables in b above. Do the shapes of the histograms confirm the skewness you found in b?
      1. Yes, the shapes of the histogram match the skewness I found from part b.
   4. How would your convert the *“rep\_education”* column into numerical data? Name two ways.
      1. You could assign each variable to a number. I believe there are 3 variables: high\_school, graduate, and college. You could make high\_school = 1, graduate = 2, and college = 3. By using the unclass() function, you could do this.
      2. You could make different columns for each option, so for this case: high\_school, college, and graduate. And for each one, if the row indicates one of these it would be filled in with a ‘1’ and the other two would be ‘0.’

**Data Visualization:**

For every graph in this section, remember to label your axes and to include a title. Feel free to play around with graphics and parameters. Have fun and explore!

* 1. Plot a bar graph for the *“Def\_Ind”* column and describe it.
     1. The def ind column of this data set is the indicator of default which means whether someone has no payments within the last three months. It is a binary column where if the value is 1, the account defaulted after an account was approved and opened with the bank in the past 18 months. If the value is 0, then there is not default. So in the graph, there are two bars where one represents when the value is defaulted at 1 and the other is at 0 for active accounts.
  2. Plot a bar graph for the *“rep\_education"* column and describe it.
     1. The rep education column has four options: high school, college, graduate, and other. The represents the unverified self-reported education level of the applicant. When looking at the graph, many are under college, the second highest is high school, then graduate, then other.
  3. Plot a histogram of the *“rep\_income”* variable.
  4. Plot a boxplot of the *“tot\_balance”* variable. Using the box plot report the five number summary of the variable? Are there any outliers for this variable?
     1. Yes, there are many outliers. When you run the ‘outliers’ variable in the bottom of my script, you can see all the outliers. Some clear outliers include: 200,000, 194485.99, 0, and 37431.07.