**DATA 180: INTRODUCTION TO DATA SCIENCE**

**FALL 2023**

**Final (40 points)**

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

**Denys Sharpylo**

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

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?
      * 20000 rows, 21 columns
   2. Report the column names of the data set.

A screenshot of a computer code

Description automatically generated

* 1. Which types of data are there in the dataset? Numeric, categorical, ordinal?
     + There is numeric (such as avg\_bal\_cards), categorical (such as Def\_ind), and ordinal data (such as rep\_education)
  2. Which columns contain missing values and how much (what percent) of those columns are missing?

A screenshot of a computer code

Description automatically generated

* 1. How do you think we should deal with missing values?
     + We can either remove/drop rows that contain these missing values or substitute them with a “filler value” that would have little to no impact on the rest of the dataset (such as the mean). However, the first option seems to be better, and would allow us to interact with a dataset full of existing values.
  2. With this data, would you fit a supervised or an unsupervised learning model? Why?
     + Supervised, because we have predictor variables that are obtained from the application in order to obtain a response variable of whether to approve or deny the credit request through a model that a financial institution would use.
  3. 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.

A screenshot of a computer code

Description automatically generated

1. **Data summary statistics:**
   1. Find the summary statistics of the data set. You can use the summary function from dplyr.
      * [Answered in Rstudio through the summary function. Available in the knitted Markdown file.]
   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”?* 
      * Num\_card\_inq\_24\_month – mean > median, so the variable is skewed right.
      * Tot\_amount\_currently\_past\_due – mean > median, so the variable is skewed right.
      * Credit\_age – mean < median, even though the values are close, it is shaped like a bell curve with a very slight skew left.
   3. Plot a histogram of the variables in b above. Do the shapes of the histograms confirm the skewness you found in b?

A graph of a person with a card

Description automatically generated with medium confidenceA graph of a graph with numbers and a number

Description automatically generated with medium confidenceA graph of credit age

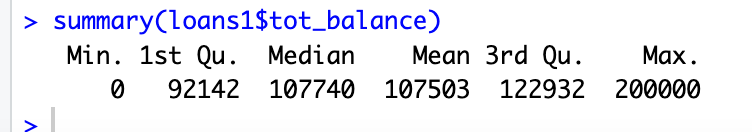
Description automatically generated  
- The graphs seem to be agreeing with the conclusions in part b.

* 1. How would your convert the *“rep\_education”* column into numerical data? Name two ways.
     + Use the different education levels as factors, and convert them using as.numeric() function.
     + Count number of appearances of each type of education, and then use the numeric values for comparison.

1. **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.
     + We can see from the bar graph for a Default Indicator variable that there are a lot more accounts that did not default than those that defaulted in the past 18 months after it was approved and opened.
  2. Plot a bar graph for the *“rep\_education"* column and describe it.
     + From the Reported Education bar graph, we see that there is an evident predominance of college graduates. High school graduates (or below) are the second biggest group, followed by graduates from graduate school and those reported as “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?



* + - According to the box plot, we can observe that there are many outlier values below the lower boundary (45,957) and above the upper boundary (169,117) – they are colored red for visual recognition in the box plot.