**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.**

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?
   2. Report the column names of the data set.
   3. Which types of data are there in the dataset? Numeric, categorical, ordinal?
   4. Which columns contain missing values and how much (what percent) of those columns are missing?
   5. How do you think we should deal with missing values?
   6. With this data, would you fit a supervised or an unsupervised learning model? Why?
   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.
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”?*
   3. Plot a histogram of the variables in b above. Do the shapes of the histograms confirm the skewness you found in b?
   4. How would your convert the *“rep\_education”* column into numerical data? Name two ways.
3. **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.
  2. Plot a bar graph for the *“rep\_education"* column and describe it.
  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?