Assigned: 01/21/2021
Due: Sun. 01/31/2021, 11:59pm

Instructions: This project will cover some questions related to topics of data types, attribute types, exploratory data analysis, and data preprocessing. The project will also serve as a further introduction to using a high-level language for analysis (e.g., R, MATLAB, Python).

Submission Requirements: Your answers must be computer generated (including text and diagrams). Your final document submission should include text responses to questions, tables, R/Matlab/Python code used to calculate answers, and figures.

Formatting of submissions: The following methods are acceptable ways to submit your assignment:

- If you are using MATLAB consider:
  - .m file + markup, publishing matlab code → PDF
    Incorporate your answers directly into your MATLAB code (code, comments, results), publish the code creating an PDF file.
  - .mlx files + LiveScript Editor → PDF
    Answer your questions in the Matlab LiveScript editor, embedding code and results from matlab .m file
- If you are using R consider:
  - Rmd → PDF
    Use `knitr` or `rmarkdown` to collect all text responses, figures, tables, and code in the R Markdown file and process it to produce a PDF file.
  - Snw → PDF
    Use `R Sweave` to collect all text responses, figures, tables, and code in the Snw file and process it to produce a PDF.
- If you are using Python consider:
  - Jupyter notebook (.ipynb) → PDF
    Incorporate all text responses, figures, tables, and code in the jupyter notebook and process it to produce a PDF file.
  - Colab notebook (.ipynb) → PDF
    Incorporate all text responses, figures, tables, and code in the jupyter notebook and process it to produce a PDF file.

Any other packages or tools, outside those listed in the assignments, for reproducible research should be cleared by Dr. Brown before use in your submission.

Template files for this project are available in the three language choices. You may deviate from these files in terms of styling, but please keep to simple, clean, choices, e.g., *make it easy for Dr. Brown and the graders to find and grade your responses.*

Name your main submission files as *P1_LastName_FirstName*, create a zip-file called *P1_LastName_FirstName.zip* and submit on Canvas. For example, if I was using R, I would submit either:

- *P1_Brown_Laura.Rmd*, *P1_Brown_Laura.pdf*, or
- *P1_Brown_Laura.Snw*, *P1_Brown_Laura.pdf*

For Matlab, I would submit:

- *P1_Brown_Laura.m*, *P1_Brown_Laura.pdf*, or
- *P1_Brown_Laura.mlx*, *P1_Brown_Laura.pdf*

For Python, I would submit:

- *P1_Brown_Laura.ipynb*, *P1_Brown_Laura.pdf*

Questions:

1. Census Data

   Consider the Census Income data set available at the UCI ML archive. Specifically, you will be interested in the `adult.data` file which contains the data and `adult.names` files which contains documentation about the data.

   You should explore the files a bit in a text editor to understand the format. Then load the data for you analysis, the first samples of the data set should be:

   ```
   39, State-gov, 77516, Bachelors, 13, Never-married, Adm-clerical, Not-in-family,
   White, Male, 2174, 0, 40, United-States, <=50K
   ```

   The variables are made up of different types: numeric, nominal, etc. Answer the following questions:

   (a) (14 points) **Variable Definitions:** For each variable (column of the data set) excluding the final class/target variable (last column), write a clear 1-sentence description of what the variable is, i.e., what information does it describe and how is it defined collected.

   To answer this question, you may have to do a bit of reading and research into this data set. If you can not find a clear explanation of what a variable is and how it is defined say so.

   For example, the variable "age" could be described as, "Age is the age of an individual as reported by that person for the 1990 census; the value is reported in integer units of years."

   Refer to each variables by the hyphenated name given in the `adult.names` documentation.

   (b) **Missing Data:** There are missing data.

   i. (2 points) What is the symbol or symbols used to indicate missing values?

   ii. (7 points) For each variable, calculate and report the percentage of missing data for that variable (percentage of rows).

   Ignore missing values for the remainder of the question, e.g., if you are asked for the mean take the mean of all non-missing values ignoring the rows with missing values.

   (c) (7 points) **Variable Types:** Which of the variables are numerics and which are categorical? (Use column names)

   (d) **Numeric Data:** For the variables of `Age` and `Hours-per-week`, answer the following questions for each variable. Use visualization principles to consider how best to present the information.

   i. (4 points) Generate a histogram with an *appropriate* number of bins to visualize the distribution of the data.

   ii. (5 points) Create a figure with 2 histograms side-by-side or a single histogram with 2 distributions overlaid, where one is for the data samples with the *class/target variable* is "$> 50k$"; in the other, only consider data samples where the class variables is "$\leq 50k$".

   iii. (5 points) Generate a figure with 2 boxplots side-by-side, with the two boxplots corresponding to samples for the two classes: "$\leq 50k$" and "$> 50k$".

iv. (6 points) Describe what the plots have revealed about the data (2-4 sentences).

(e) **Categorical Data:** For the variables of `Education` and `Marital-status`, answer the following questions. Use visualization principles to consider how best to present the information.

i. (4 points) Generate a bar plot, where each bar corresponds to the number of unique values. Include, missing values as a possible value in the plot if appropriate.

ii. (5 points) Create a figure with either (a) 2 bar plots in a single figure (stacked one on top of the other), where the top bar plot is for the data with the class '$\leq 50k$" and the bottom plot is for data with the class "$> 50k$", or (b) a grouped bar chart with the two groups being the class variable.

iii. (6 points) Describe what the plots have revealed about the data (2-4 sentences).

2. Sports Data

The use of data analysis in sports is becoming increasing more common (and a high profit business). Interest in this analysis grew substantially with the publishing of the book *Moneyball* (and the subsequent movie). Statistical analysis has spread to many other sports including basketball, football (both American and soccer), tennis, and many others.

Consider the data set provided: `nfl-20-running-stats.csv` that consists of the 2020 NFL Running statistics available at: https://www.pro-football-reference.com/years/2020/rushing.htm.

First select the pool of players to focus on Pos, only those that are running backs or full backs that played in 6 games or more G. The following analyses will only consider this subset of players.

(a) (2 points) What is the size of the pool of players considered for further analysis?

(b) (6 points) Calculate the *mean*, *median*, and *mode* of TD and Fmb.

(c) (6 points) Calculate the first and third quartiles, Q_1 and Q_3 and $37th$ percentile of Yds and 1D

(d) (5 points) Present the *five-number summary* of Y/G and Lng as a table

(e) (6 points) Compare the distributions of the number of Yards of each runner based on whether they had less than 5 touchdowns TD or 5 or more touchdowns.

(f) (5 points) Draw the scatter plot of Yds vs. Fmb.

(g) (5 points) Draw a scatter plot of 1D vs. Y/A.