Assigned: 01/18/2017
Due: Mon 01/30/2017, midnight

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 an 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 and description of your efforts, tables, R/Matlab/Python code used to calculate answers, and figures. As well as the code to carry out the work.

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

- {Word + code}, {Open Office + code} → PDF
  This option may require taking screenshots or printing figures created in R/MATLAB/Python and importing them into the word processing software. Additional code and results should also be inserted into the word documents.
- If you are using R consider:
  - Rmd → PDF, Rmd → HTML
    Use `knitr` or `rmarkdown` to collect all text responses, figures, tables, and code in the Rmarkdown file and process it to produce a PDF or HTML file.
  - Snw → PDF, Stex → PDF
    Use `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 MATLAB consider:
  - .m file + markup, publishing matlab code → HTML
    Incorporate your answers directly into your MATLAB code (code, comments, results), publish the code creating an HTML file.
  - .m files + *Your favorite document editor*
    Answer your questions in your text editor, embedding code and results from matlab .m file
- If you are using Python consider:
  - iPython → HTML
  - LaTeX + Sphinx → PDF / HTML

I highly recommend following the ideas of reproducible research and embed the code, images, and results directly into the text using packages like `knitr/rmarkdown`, `Sweave`, `publish`, or `iPython` (other packages follow this practice using Latex and reStructureText as well and are open for you to use).

If you want to follow the style of the R introduction documents on Canvas (e.g., introA.html, introB.html, etc.), please use the provided CSS style file `min.css`), and follow instructions provided in R Studio documentation and the `rmarkdown` package. There are also a number of style and code highlighting styles available using `Bootstrap` themes.

Name your main submission files as *P1_LastName_FirstName*, create a zip-file called *Project1_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.Rmd*, *P1_Brown_Laura.html*, or
- *P1_Brown_Laura.Snw*, *P1_Brown_Laura.pdf*

along with any other supplemental `.R` files I created in *Project1_Brown_Laura.zip*.

Questions:

1. (4 points) From your reading of Ch. 1 of the text book, in your own words, what is the difference between classification and regression? How are they similar?

2. 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 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."

   (b) **Missing Data:** The symbol "?" indicates missing values in the data.

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

   Ignore missing values for the remainder of the question.

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

   (d) **Numeric Data:** Select two of the numeric variables, answer the following questions.

   i. (4 points) Explore the number of unique values the variable has.

- For a variable with less than 50 values, generate a histogram where each bin corresponds to one of the variable's values.
- For a variable with 50 or more values, generate a histogram using **50 bins** (the bin placement will be done automatically, via the program).

ii. (4 points) For each of the two variables, create 2 histograms as part of the same figure (stacked side-by-side or one on top of another). In one histogram, only consider data samples with the *class variable* is "$> 50k$"; in the other, only consider data samples where the class variables is "$\leq 50k$".

iii. (4 points) For each of the two variables, 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 two of the categorical variables (not including the class variable, last column), answer the following questions.

i. (4 points) For each of the two variables, generate a bar plot, where each bar corresponds to the number of unique values. Include, missing values as a possible value in the plot.

ii. (4 points) For each of the two variables, create 2 bar plots in a single figure (stacked one on top of the other). The top bar plot is for the data with the class '$\leq 50k$" and the bottom plot is for data with the class "$> 50k$".

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

(f) **Pairwise Analysis:**

i. (5 points) Pick a categorical variable, then explore how the variable changes with "Age". Explore whether the variable you select appears to depend on age, or whether it is independent of age. Clearly indicate what evidence there is to support your conclusion.

ii. (5 points) Pick any two numeric variables, and explore whether or not they depend on each other, i.e., are they independent or not? Clearly indicate what evidence there is to support your conclusion.

3. Automotive Data

Consider the data set provided: `Auto.csv` that describes several aspects of many cars.[1] Answer the following questions:

(a) (3 points) Which of the variables are quantitative and which are qualitative? (Use column names)

(b) (10 points) For each of the quantitative variables, calculate and report in a table the *mean*, *median*, *mode*, and *range*. For *mode* you will need to write your own function to perform the calculation.

For parts (c)-(e), only consider the variables displacement-DS and horsepower-HP.

(c) (3 points) Calculate the first quartile, Q_1, $37th$, and $89th$ percentile of DS and HP.

(d) (4 points) Present the *five-number summary* of DS and HP as a table.

(e) (4 points) Draw the box plots for DS and HP

[1] The `Auto` data set is part of the `ISLR` package, but is available as a csv on Canvas.

(f) (6 points) Investigate the relationship among different variables using scatterplots and other graphics utilities. Describe any interesting relationships found.

(g) (4 points) Let's say we want to predict MPG using the other variables. What other variables would be most useful in this prediction task. Why?

4. 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.

(a) **Tennis Data:**

For example, see the following articles on data analysis in tennis:

- *James Murphy Shares Remixes Made With Tennis Data Album*, Pitchfork
- *Despite Advanced Stats, Tennis Has a Data Problem*, Wall Street Journal
- *Why Tennis So Far Behind Other Sports in Data Analytics?*, Forbes
- *US Open 2015: Advanced Analytics in Tennis Takes a Minor Step Forward with IBM's SlamTracker*, IBTimes

Recenting a crowd-sourcing solution has been used to chart tennis match statistics. This project Match Charting Project has seen growth in data collected to over 2500 matches. The data is housed at Github tennis MatchChartingProject. For this analysis, you will focus on some basic statistics that can be calculated from this data set. You will only need the files: `charting-m-matches.csv`, `charting-m-stats-Overview.csv`, `charting-w-matches.csv`, and `charting-w-stats-Overview.csv`

Answer the following questions:

i. (16 points) For both the Men's and Women's tours, consider only matches at the four Grand Slams:
   - Australian Open
   - French Open
   - Wimbledon
   - US Open

   from 2011 - present. List the top 5 men and women with the most Aces/match along with this value.

ii. (16 points) For both the Men's and Women's tours in 2015, determine the top 5 men and women who have at least 5 matches charted with the highest break points saved percentage: Num. Break Points Saved / Num. of Break Points = bk_ptsbp_saved

Other data sets exist charting similar information: Github tennis slam pointbypoint.

Tennis enthusiatics have even created challenges on data analysis and visualization The Tennis Notebook - Storytelling Challenge

(b) (5 points (bonus)) Recreate one of the visualizations at Tennis Visuals using the data sets available above or create your own unique visualization of that tells a story about a match, a tournament, a career, a player, etc.