MSDS 593 EDA and Visualization

Instructor: Shan Wang

Summer 2021

Homework 1 Due Friday July 16 11:59pm

Introduction: submit **a .pdf file** containing all you answers/python codes for the homework and submit **a separate notebook for any python code** you have run. The purpose of this is the grader can view your pdf file directly when grading on canvas. When necessary, they might need to go back to your notebook to run certain part. Failing to submit either file would automatically reduce your grade. Mark the questions numbers clearly and organize your answer nicely so the graders won't miss some points. **Avoid copy and paste** and try to type all the codes yourself. A lot of questions have a lot of flexibility: it's hard to have two people having the same codes, it's easy to detect and please follow academic honesty.

I. This part of questions will help you practice the basic numpy and pandas.

1. Create a 1D ndarray of numbers from 100 to 112 (step=1) inclusively and use python to
   (a) Print the shape of the array
   (b) print the data type of the array
   (c) Create a new array that is a slice of the original array with index $[5 : 10]$ inclusively, and assign all values of the the new array to be 0.
   (d) Create a boolean vector from the array to indicate if any element is greater than 105, and less than or equal to 110.
   (e) Replace all the elements in the array that are greater than 105 and less than or equal to 110 with 0.
2. Make a *Series* object with year values: 1991,1992,1993,1994,1995,1996,1997,1998,1999,2000.
   (a) Print out how many total values there are using code not manual counting
   (b) Make another series with rainfall values 12.09,12.35,12.51,10.25,10.18,10.59,10.26,10.48,8.67,10.23
   (c) Imagine the order of rainfall values follow the order the years. Print out the years for which rainfall was less than 11.
   (d) Normalize that rainfall series by $\frac{x-mean}{std}$ and print it out.
   (e) Set the year starting 1996 and forward as np.nan. Count the number of missing values.
   (f) Fill all the NaNs with 0.

3. For the *cars.csv* data set:

   (a) Use numpy's function $np.corrcoef(x, y)$ to compute the correlation between a car's weight and the miles per gallon; that function returns a matrix of x with x, x with y, etc... so the diagonal will always be correlation 1.0. What is the correlation between a car's weight and the MPG? What does the correlation tell us about their relationship?

   (b) Display the records for all 8 cylinder cars

   (c) Create a new column called 'ENG2WGT' that has the engine to weight ratio

4. For the $kaggle - uber - other - federal.csv$ data set:

   (a) Create a new data frame containing 'Time', 'Status', and 'PU_Address' columns

   (b) Create a new column 'Hour' extracting hour information from 'Time'. Hint: make sure 'Time' is the correct data type and keep the date added.

   (c) Set the index of the data frame to 'Time'.

   (d) Display records at positions between 10 and 15 inclusively

   (e) Display the 'PU_Address' for records whose index is 'Today's date 20:00:00': hint: today's date depends on the day you are working on the question.

   (f) Reset the data frame so that 'Time' is a column again

II. This part of questions will help you practice the basic EDA and matplotlib.

5. You work for a large school district as a data analyst. Your boss wants to purchase a large amount of cereal for school breakfasts. He needs to choose a manufacturer and product. He wants you to prepare a presentation for the executive team. You are provided with some data in *cereal.csv*:

   Dataset: Columns in the dataset:

   - Name: Name of cereal
   - mfr: Manufacturer of cereal
     - A = American Home Food Products;
     - G = General Mills
     - K = Kelloggs
     - N = Nabisco
     - P = Post
     - Q = Quaker Oats
     - R = Ralston Purina
   - type: cold/hot
   - calories: calories per serving
   - protein: grams of protein
   - fat: grams of fat
   - sodium: milligrams of sodium

- fiber: grams of dietary fiber
- carbo: grams of complex carbohydrates
- sugars: grams of sugars
- potass: milligrams of potassium
- vitamins: vitamins and minerals - 0, 25, or 100, indicating the typical percentage of FDA recommended
- shelf: display shelf (1, 2, or 3, counting from the floor)
- weight: weight in ounces of one serving
- cups: number of cups in one serving
- rating: a rating of the cereals

(a) Initially explore the dataset by looking at the number of records, column names, column types and few record values. Through this initial look, combined with your boss's goal above, list five analytics queries or questions that you would have about this dataset in your exploratory process.

(b) To answer the questions you have listed above, what columns from this data you would need? Print some basic statistical summaries and plots for **at least 5 columns** you would need to check missing values, extreme values and value distributions. List three findings you have at this point.

(c) To answer the questions you have listed above, describe 5 visualizations that can help you explore this data. (Example: line graph between variable x and y.).

(d) Plot at three DIFFERENT graphs you have mentioned above. Summarise some findings from each plot. (For example, line chart of x and y and line chart of a and b doesn't count as different; line chart and box plot does.)

(e) Besides the information provided in the data, what else information might be helpful to achieve your boss' goal?

- Guideline:
  - Questions in (a) should be related to boss's goal.
  - At step (b), your findings don't have to answer the goal directly yet, it can be just simple exploration about each column. Findings can be about missing values, extreme values, value distributions or between variable relationships.
  - Visualizations in (c) should follow the correct data type and relate to your questions from (a).
  - You don't have to plot some fancy designs in (d) yet. Focus on the findings from observing the plots.
  - Extra information in (e) should NOT be in the current data, but be helpful to your boss's goal.