Homework 2

Robert Clements

Homework 2

- Submit two files: submit either a .html or .pdf file containing all your answers and code for the homework questions and submit a separate notebook for any python code you have run so that the grader can run it if necessary. To convert to html you can use the nbconvert library, or depending on the IDE you are using you may be able to export it using the IDE's features. To convert to .pdf you will need a version of TeX installed on your laptop. If you don't know anything about TeX, just do .html.

- Mark the question numbers clearly and organize your answers nicely so that the grader can easily find your answers. You can follow the style from the lecture notebooks to separate answers and code.

- Avoid copying and pasting and auto-complete and try to type all of the code yourself.

- Do not share code with each other.

I. These questions will help you practice basic numpy and pandas.

1. Create a 1D ndarray, data1, as numbers from 100 to 112 (step=1) inclusively (i.e. 100, 101, ..., 111, 112) and use python to:

(a) Print the shape of data1

(b) Print the data type of data1

(c) Create a boolean vector data1_boolean to indicate which elements of data1 are greater than 105 and less than or equal to 110.

(d) Replace all the elements in data1 that are greater than 105 and less than or equal to 110 with a 0.

(e) Create a new array data2 that is a copy of a slice of data1 with numeric array index 5 to 10 inclusively, and then assign all values of the new array to be 0.

2. Create a 4 x 3 matrix-like 2-D array with random numbers generated automatically by Python functions (you can do it any way you like). Find the max value of each row and store them in a 1-D array.

3. Make a Series object named 'year' with these year values: 1991,1992,1993,1994,1995,1996,1997,1998,1999,2000.

(a) Print out how many total values are in the Series, using code.

(b) Make another Series called 'rainfall' with these rainfall values 12.09, 12.35, 12.51, 10.25, 10.18, 10.59, 10.26, 10.48, 8.67, 10.23.

(c) Normalize the rainfall Series from (b) using the formula: $\frac{x-mean}{std}$, and save it to a new Series called 'rainfall_normalized'.

(d) Imagine the order of rainfall values follows the order of the years. Print out the years for which rainfall was less than 11.

(e) Set the rainfall values from 1996 to 2000 as `np.nan`. Count the number of missing values using code.

(f) Following (e), fill all the NaNs with 0.

4. For the cars.csv dataset:

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

(b) Create a table to show the unique values from 'CYL' and the counts of each value.

(c) Add a new column called 'ENG2WGT' to the dataset that has the engine-to-weight ratio.

5. For the kaggle—uber—other—federal.csv dataset, create a new DataFrame containing 'Date','Time', 'Status', and 'PU_Address' columns.

(a) Add a new column 'Datetime' to the DataFrame, which combines the values from columns 'Date' and 'Time'.

(b) Check the datatype of each column, and change them to the "correct" dtype based on your judgement. Print the dtypes of all columns after making all the appropriate changes.

(c) Create a new column 'Hour' extracting hour information from either 'Datetime' or 'Time'.

(d) Set the index of the DataFrame to 'Date'.

(e) Find the number of records whose index is '07/02/2014'.

(f) Reset the DataFrame index so that 'Date' is a column again.

II. These questions will help you practice basic EDA and get ready for matplotlib.

Guidelines:

– Questions in part (a) should be related to the boss's goal.
– In part (b), your findings don't have to answer the goal directly yet, it can be just simple exploration of 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 anything yet at this time.
– Extra information in (d) should NOT be in the current data but be helpful to your boss's goal.

Now the questions are as below:

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

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 a few record values. Through this initial look, combined with your boss's goal above, list four analytic 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 would you need to explore further? Print some basic statistical summaries (required) and plots (if you would like) for at least three columns. You might need to check missing values, extreme values and value distributions. List at least three findings about those columns you have at this point.

(c) To answer the questions you have created in (a), describe four visualizations that can help you explore this data (e.g. scatterplot between variable x and y to show the potential relationship). You only need to describe what you would do, no coding needed. You **DO NOT** need to plot anything yet.

(d) Besides the information provided in the data, what other information or data might be helpful to achieve your boss's goal? Again, just describe what you would need, no coding needed.