AI534 — Implementation Assignment 0 — Due 11:59PM Oct 1st, 2021

General instructions.

1. Please use Python 3 (preferably version 3.6+). You may use packages: Numpy, Pandas, and matplotlib, along with any from the standard library (such as 'math', 'os', or 'random' - for example).

2. You should complete this assignment alone. Please do not share code with other students, or copy program files/structure from any outside sources like Github. Your work should be your own.

3. Your source code and report will be submitted through Canvas.

4. You need to follow the submission instructions for file organization (located at the end of the report).

5. Please run your code before submission on the EECS servers (i.e. babylon01.eecs.oregonstate.edu). You can make your own virtual environment with the packages we've listed in either your user directory or on the scratch directory. If you're unfamiliar with any of this process, or have limited access, please contact one of the TA's.

6. Be sure to answer all the questions in your report. You will be graded based on your code as well as the report. The report should be clear and concise, with figures and tables clearly labeled with necessary legend and captions. The report should be a PDF document.

7. In your report, the **results should always be accompanied by discussions** of the results. Do the results follow your expectation? Any surprises? What kind of explanation can you provide?

Part I (5pts) The goal of this part is to get you familiarized with setting up environment such that your code will run smoothly on the server. What you need to submit for this part is mainly the readme file for your code.

To ensure that your submission runs and works as intended, we will test it on the server (babylon01) under Python virtual environment. Implemented on your local machine, your code is specific to your local Python environment (packages you use, revisions, etc.). It is your responsibility to test your code on the server prior to submission to ensure that it can be interpreted and run on babylon01. If the submission fails to run on the server, and/or does not have instructions on how to run it, it might get a zero grade.

Setting up the virtual environment on babylon01. The server might not have all the packages you chose to use. To be able to use packages of your choice, please refer to the following source (this should be done in babylon01 server):

`https://it.engineering.oregonstate.edu/setting-virtual-environments-python`

Note:

- Please feel free to ask TA's for help in case of issues with setting up your virtual environment on "babylon01" server.
- Some packages might be too large to fit in your directory. This is unlikely but might happen if you are running out of quota disk space on your account. You are welcome to use the "/scratch" space in your home directory, which should allow you to use extra disk space for class projects.

Providing instructions for TA's of how to test your code. Apart from your report, please provide a brief but detailed "readme.txt" file with specific steps of how to run your code. Here is an example what the readme file might look like:

> 1. Activate python virtual environment in the same directory where the code is.
> 2. To be able to run this code on the "babylon01" server under the virtual environment, please have these packages installed:
>   a. Numpy
>   b. Pandas
>   c. Etc...
> 3. Next, to run the program, please use the following command:
>   a. python test.py

Part II: Let's give it a try and play with some data (15 pts)

Note 1. When processing data, for common operations like loading csv file, deleting, inserting or splitting a column of data, etc, you can use existing APIs from numpy (np) or pandas (pd). For example, for loading csv file, you may need `open()`, `np.loadtxt()` or `pd.read_csv()`; For deleting a column, you may need `np.delete()` or `pd.drop()`; For inserting or splitting, you may need `np.insert()` or `(df.column.)str.split()` functions.

Note 2. Try to use full matrix operation whenever possible rather than using for loops. It will be drastically more efficient. You will make extensive use of matrix computing in engineering practices and scientific computations in the future, it's the time to get familiar with this.

Data. This data consisted of historic data on houses sold between May 2014 to May 2015 in csv format. You are also provided with a description of the features. The last 'feature' actually is the target y value for prediction. The intended use for this data is to train a predictive model for the sale price of a house based on these features, which we will explore more for the next implementation assignment. For this assignment, you will simply do some light exploration/manipulation of the data to warm up. Your code should implement the following steps. Each step may have some specific question(s) which you should answer in your report.

(a) One of the features is the ID of the entry. Please remove this feature from the data. This will reduce the feature by one.
`Question: Is it a good idea to use this feature in predicting the price of the house? why?`

(b) One of the features the date. Please split the date feature into three separate numerical features: *month*, *day* , and *year*. This will increase the dimension of the data by two. `Question: do you think the date feature is useful for this problem? Can you think of better ways of using this date feature than splitting them into three numerical features?`

(c) Several features are listed as numerical but have only a small number of discrete values. This includes "bedrooms", "bathrooms" and "floors". For each of these three features, identify the unique values that are observed (you can use the `unique` function from Pandas). For each of the 3 features, generate a boxplot (you can use the boxplot function from Pandas) of the price. Specifically, you should generate one plot for each feature grouped by the feature value. For example, for a feature with 3 distinct values (A, B and C), the box plot for this feature might look like the following:

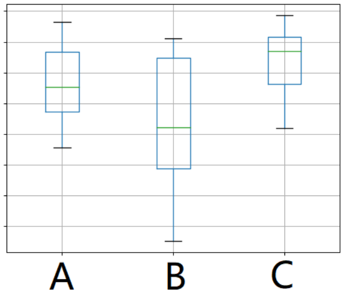

Figure 1: Sample boxplot for a feature with three values (A, B, C). Each column shows the median, Q1, Q3, min and max of the price for houses with a particular feature value.

(d) Consider the following numerical features: `sqrt_living`, `sqrt_lot`, `sqrt_living15`, `sqrt_lot15`. Calculate the co-variance matrix of these four features. Generate a scatter plot of the data using using sqrt_living and `sqrt_living15`, and another scatter plot using `sqrt_lot` with `sqrt_lot15`.

`Question: what do you observe from the scatter plot? Are these features redundant?`

Submission. For this assignment, your submission should be a single zip file that contains the following:
1) Your code
2) a readme file on how to run your code on the server
3) Your report. For this assignment, your report should be fairly short containing required figures with discussion of results and answers to the specific questions. The report should be typed (not handwritten), in PDF format.