CSC 535/635 – HW 1  
Due Date: Friday, September 25, 2020 by 11:59 PM

# **Instructions**

This homework assignments consists of three parts. In part one, you will perform data analysis tasks on a provided dataset. In part 2, you will use a dataset of your choice and perform at least one data analysis task of your choice. In part 3, you will write a short report about the analysis you did in pars 1 and 2. Format and guideline for the report are in the file Report\_Template.docx

You have the option to either work alone or with a partner. You cannot work with more than one partner. Undergraduate and graduate students may work together. If you work with a partner, please make only one submission and include both names at the top of your notebook and report. To find a partner, you can make a post on the discussion board “Homework Partners”. Notice that: students are not allowed to work with the same partner for more than one homework assignment. This rule does not apply to the course project.

# **Part 1** [total 70 points]

Tips received by waitresses are essential portion of the salary for many waitresses. Fair treatment of waitresses includes giving them equal opportunity to receive equal tips. In this homework assignment, you will use IPython tools to analyze a tips dataset, which is provided in the file tips.csv. The dataset contains information about tips given by customers of a restaurant over a period of time in 1990. The attributes of the dataset are as follows:

* total\_bill: cost of the meal including tax, in US dollars
* tip: gratuity in US dollars
* sex: sex of person paying for the meal
* smoker: Is a smoker in the party? With possible values No and Yes
* day: day of the meal with possible values Thur, Fri, Sat, and Sun
* time: time of meal with possible values Dinner and Lunch
* size: number of people in the party

As part of the analysis, you will investigate the following questions:

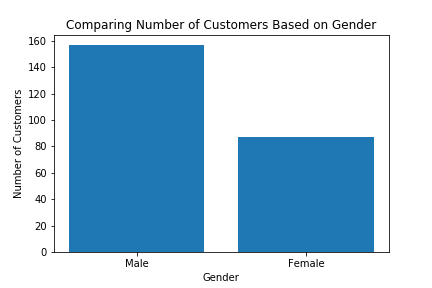
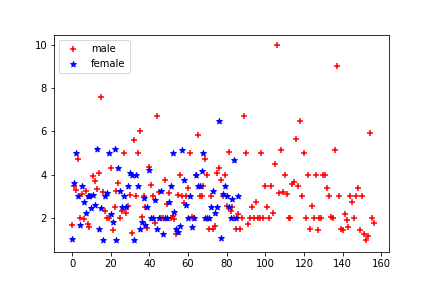
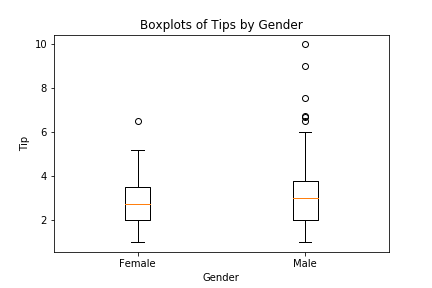
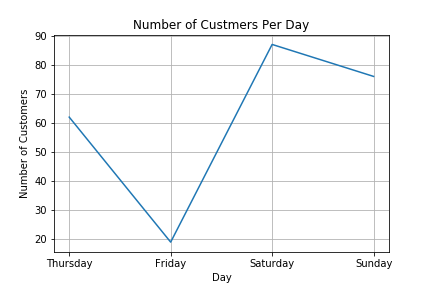
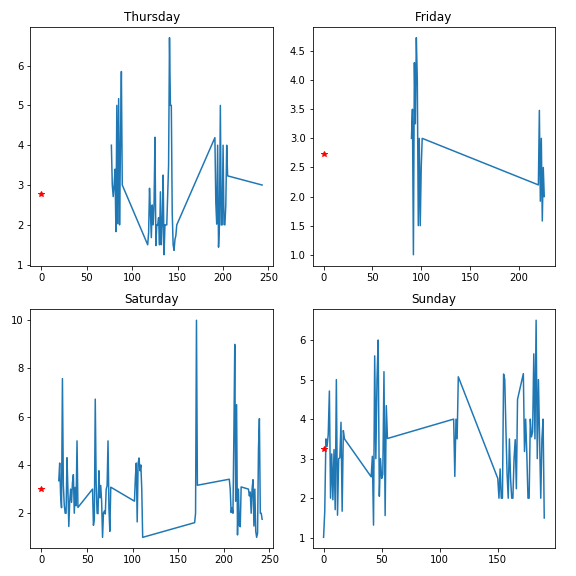
1. How much tip do customers usually give on average?
2. Do male customers usually give more tips than female customers?
3. What day(s) of the week is the restaurant the busiest?
4. Do customers tend to give more tips on certain days?

To help you perform the analysis, you are asked to perform the steps below.

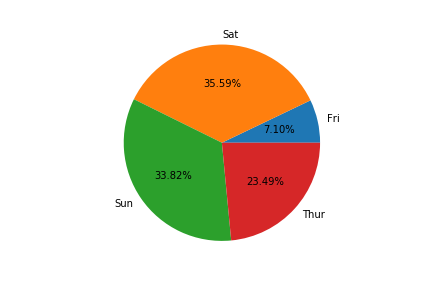
Write IPython code to do the following. Save your code in a file with the name hw1.ipynb. [2 points each item, except where specified differently]

## **Getting to know the data:**

1. Load the data in a DataFrame (DF) named tips.
2. Find the names of the columns.
3. Rename the columns “time” as “meal” and “size” as “party size”.
4. Use the head function to view parts of the data.
5. Retrieve the information in the second and third rows using two different methods (you can use .loc, .iloc, or the slicing notation, :).
6. Retrieve the information in the first and second columns (i.e., under total\_bill and tip). Use the head function to display the first three rows under these two columns.
7. Get the information stored in the cell at the intersection of the second row and third column. I.e., what is the sex of the customer (that paid for the meal) that is stored in the second row of the DF?
8. Find the number of observations in the dataset.
9. Use the describe function to get basic statistics about the numerical attributes.
10. Use the describe function to get basic statistics about all attributes. It is okay to have the value NaN shown in some places.   
      
    **How much tip do customers usually give on average?**
11. Find the average tip amount.
12. Get basic statistics about the tip column using the describe function. This helps give an idea about the min, max, median values of tip in addition to other basic statistics.
13. Draw a boxplot of the tip column. This gives a visual representation of the tips data and shows outliers.
14. From the boxplot, it looks that the outliers are for tips >= 6. Retrieve the rows where the tip amount is >= 6.   
      
    **Do male customers usually give more tip than female customers?**

1. [3 points] Use the groupby function to find the average tip amount per sex. i.e., find the average tip by male customers and the average tip by female customers. It is okay if your command gives the averages for all numeric attributes.
2. Create a DF named males that contains the male samples in the tips DF. Then, call males.sample(10) to draw a random sample of 10 male observations.
3. Create a DF named females that contains the female samples in the tips DF. Then, call females.sample(frac=0.1, replace = True) to generate and display a random sample with replacement that consists of 10% of the size of the females DF.
4. Write code to find out the number of male customers and the number of female customers.
5. [4 points] Create a bar chart comparing the number of males and number of females in the dataset. Give labels to the x-axis, y-axis, and a title to the plot. Sample display is as follows:  
     
   
6. [4 points] Draw a scatter plot of the tips given by male customers and a scatter plot of the tips given by female customers. Notice, for easier comparison, you need to show the two scatter plots in the same subplot area. Make sure to label the data and show a legend. Sample display is as follows:  
     
   
7. [5 points] Draw two box plots to show tips per sex. One box plot to show the tips by females while the other to shows tips by females. Show both plots in the same area (i.e., do not create separate subplot areas for each boxplot because that is visually easier to compare.) Label the diagram as show in the following sample display:  
     
     
   **What day(s) of the week is the restaurant the busiest?**
8. Apply the describe function on the “day” column of the tips DF and observe the result.
9. Use the groupby function to form groups based on the days of the week and find the number of customers that were served every data?
10. [5 points] Draw a line plot that shows number of customers served every day. Notice: you do not need to worry about the individual customers as shown under the size attribute. We are treating each observation, regardless of party size, as one customer.   
      
    Use the following line to specify tick locations and labels of the x-axis:  
      
     plt.xticks([0.0,1.0,2.0,3.0],["Thursday",'Friday',"Saturday","Sunday"])  
      
    Use the following line to show a grid:  
      
     plt.grid(True)  
      
    Sample diagram is as follows. Notice the order of the days on the x-axis. You need to follow the same order in your line plot.   
      
     **Do customers tend to give more tips on certain days?**
11. Use the groupby function to find the average tip per day? Store these averages in variables. These variables will be used next to display the averages in plots.
12. [5 points] Draw a figure with 2 by 2 subplots. Each subplot shows a line plot of the tips given on a certain day. Show the average tip in each subplot. For example, these are shown as red asterisks in the diagram below. Give titles to the different subplots as well as a title to the whole figure. The line plt.figure(figsize=(8, 8)) can be used to specify figure dimensions in inches. Sample output is as follows:  
      
      
    

Tips Given on Different Days

1. [4 points] Draw a pie chart to compare the **total** tip amounts collected on the different days. Make sure the pie chart shows as a perfect square. Show percentages (including the percent sign, %,) and labels as shown in the following figure:  
     
   

# **Part 2 [15 points]**

Find a dataset of your choice. Propose at least one “non-trivial” data analysis question on that dataset. Then, write IPython code to answer the proposed question or questions. Make sure to cite the source of your dataset in the report and as a comment in the notebook.

Possible places to look for datasets include:

<https://www.kaggle.com/datasets>

<https://www.kdnuggets.com/datasets/index.html>

<https://www.kdnuggets.com/2011/02/free-public-datasets.html>

<http://www.inf.ed.ac.uk/teaching/courses/dme/html/datasets0405.html>

<https://aws.amazon.com/opendata/open-data-sponsorship-program/>

<https://www.statcrunch.com/>

<http://archive.ics.uci.edu/ml/index.php>  
<https://datacatalog.worldbank.org/>

<http://shawndra.pbworks.com/w/page/16303773/datasets>  
<http://swiki.cc.gatech.edu/compfreak/5>

<https://think.cs.vt.edu/corgis/csv/>

<https://www.ll.mit.edu/r-d/datasets/1998-darpa-intrusion-detection-evaluation-dataset>

<http://www.cs.cornell.edu/~shuochen/lme/data_page.html>

<https://www.quora.com/Where-can-I-find-large-datasets-open-to-the-public>

<https://stackoverflow.com/questions/2674421/free-large-datasets-to-experiment-with-hadoop>

Several cities are making data available:  
<https://data.mo.gov/>

<https://data.kcmo.org/>

<https://www.kcdigitaldrive.org/article/kcmo-best-open-data/>   
<https://data.cityofchicago.org/>   
[https://data.sfgov.org](https://data.sfgov.org/)   
<https://www.opendataphilly.org/>

**Part 3 [15 Points]**

Write a report about the analysis you did in parts 1 and 2. Follow the format and guidelines as provided in the file Report\_Template.docx.

**What to turn in?**

Upload your code and report to Blackboard. Name the code file as hw1.ipynb and the report as hw1\_report.docx. Make sure to include your name(s) at the beginning of your report and notebook.