# ITP 449, Fall 2020

Homework 5
30 points

For each one of the following questions, write Python code in PyCharm.

- For each question, create a *new* Python file. Name each *hw5_q1_lastname_firstname.py* etc.
- Create a header in each file using *comments* to display your name and HW information. After that write your Python code.
  #Tommy Trojan
  #ITP 449 Fall 2020
  #HW5
  # Q1

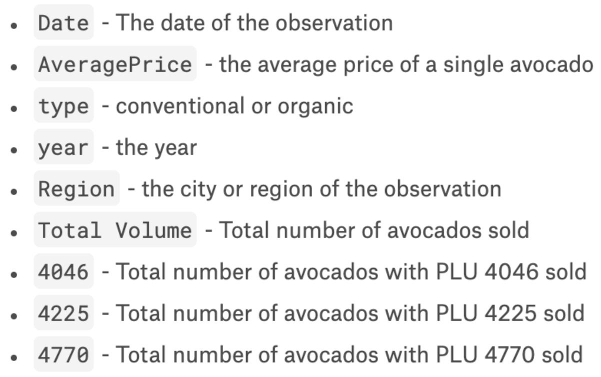
- Apart from the above comments, include single line comments describing the core logic of your algorithm / code.
  As an example,

  #Obtaining Tuples of Car Attributes (Car Name, mpg, cyl, disp, hp, gear) and generating pandas series.
  #Creating a DataFrame using the mtcars dataset.

Problem 1

The *avocado.csv* dataset contains price and quantity of avocados sold over time in various regions.

Additionally –



- `Date` - The date of the observation
- `AveragePrice` - the average price of a single avocado
- `type` - conventional or organic
- `year` - the year
- `Region` - the city or region of the observation
- `Total Volume` - Total number of avocados sold
- `4046` - Total number of avocados with PLU 4046 sold
- `4225` - Total number of avocados with PLU 4225 sold
- `4770` - Total number of avocados with PLU 4770 sold

1. Preparation
    a. Download the attached CSV file.
    b. Read the dataset into Python using Pandas.
    c. Include only these columns: *Date, AveragePrice, Total Volume*
    d. Store the data in a DataFrame *avocado*
    e. Convert Date column to a timestamp using *datetime*.
    f. Print the dataframe

2. Plotting
    a. Create a figure with 4 subplots
    b. Sort *avocado* by Date inplace in ascending order.
    c. Plot the average price of avocados over time in subplot 1. Use *scatter*.
    d. Plot the total volume of avocados sold over time in subplot 2. Use *scatter*.

You notice that the plots are cluttered. The reason is that there are many dates in the dataframe and there are several transactions on the same date!

To address this, we will aggregate the volume and price by date.
Create a new dataframe *avocado1* which sums the Total Volume for each date. Here are the steps

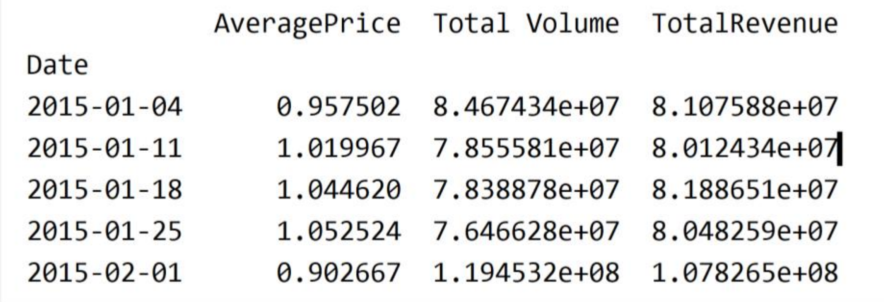
- Create a new column in *avocado* called *TotalRevenue* which is the product of *average price* and *total volume*
- Then create a new dataframe called *avocado1* which groups together the dataframe over the date

```
        avocado1 = avocado.groupby('Date').sum()
```

- Print avocado1. You will notice that the AveragePrice also got aggregated. This is not correct.
- Recalculate the average price using this

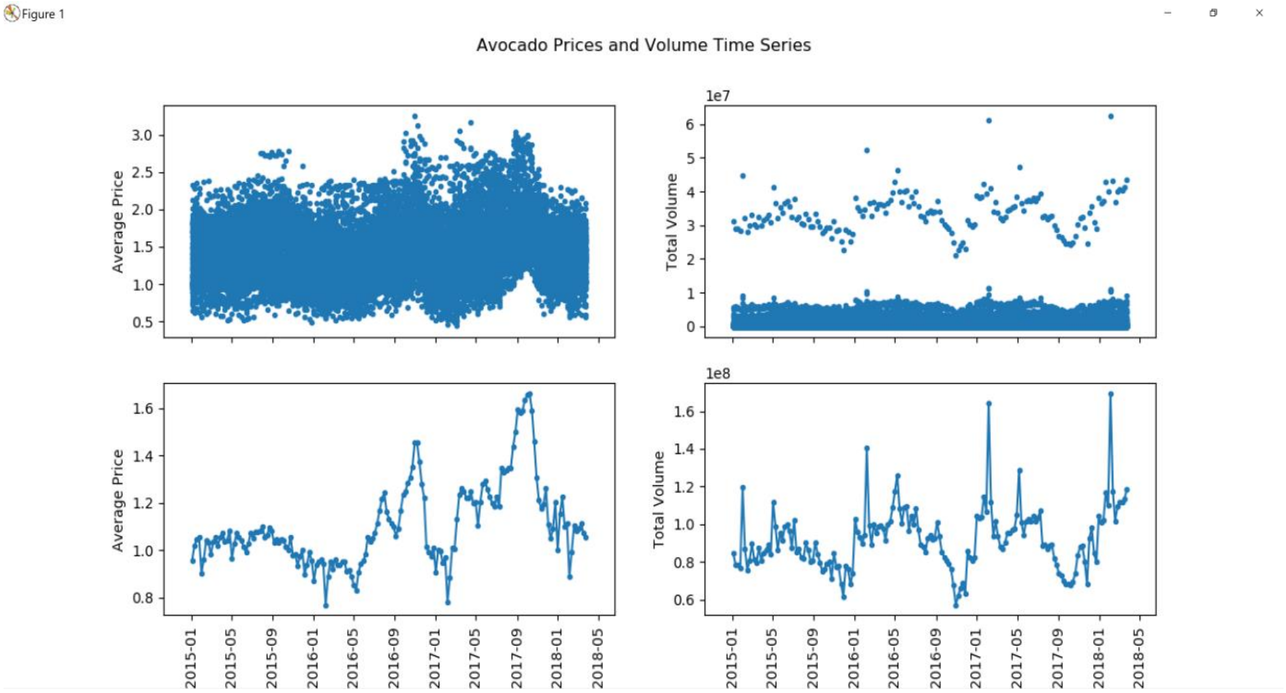
```
avocado1['AveragePrice'] =
avocado1['TotalRevenue']/avocado1['Total Volume']
```

- You should now have the following dataframe. Print the dataframe

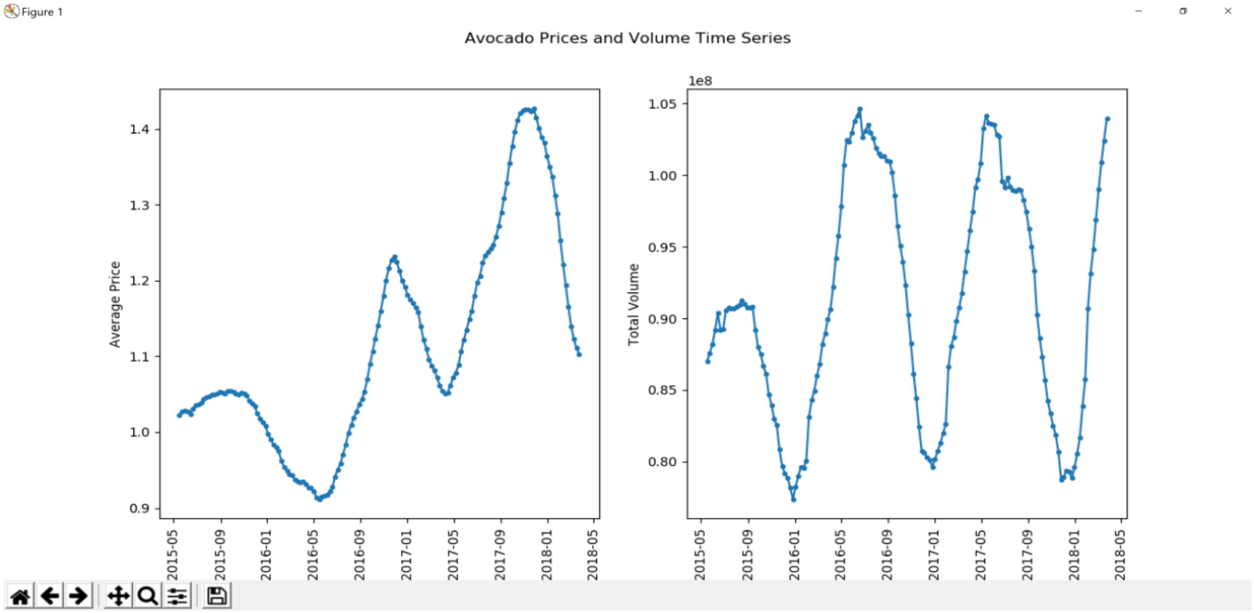


```
            AveragePrice  Total Volume  TotalRevenue
Date
2015-01-04      0.957502  8.467434e+07  8.107588e+07
2015-01-11      1.019967  7.855581e+07  8.012434e+07
2015-01-18      1.044620  7.838878e+07  8.188651e+07
2015-01-25      1.052524  7.646628e+07  8.048259e+07
2015-02-01      0.902667  1.194532e+08  1.078265e+08
```

    e. Plot the average price of *avocado1* over time in subplot 3. Use Plot.
    f. Plot the total volume of *avocado1* sold over time in subplot 4. Use Plot.



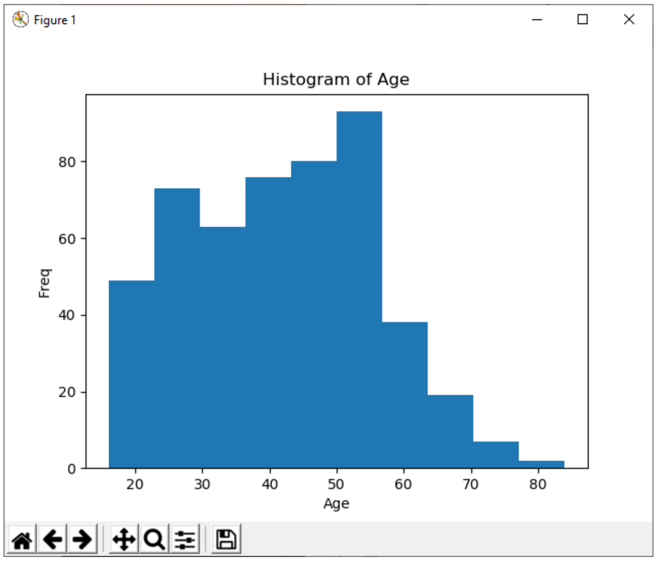
3. Plotting
    a. Create a figure with 2 subplots
    b. Use the code of examples in Lecture 5 to smooth out the last two plots from question 2. Plot the smoothed curves in subplots 1 and 2. You could use smoothing over 20 days
    c.



Problem 2

1. Create a statistical summary of the data in the file "*CommuteStLouis.csv*". Plot a histogram of *age* for the *CommuteStLouis* data.

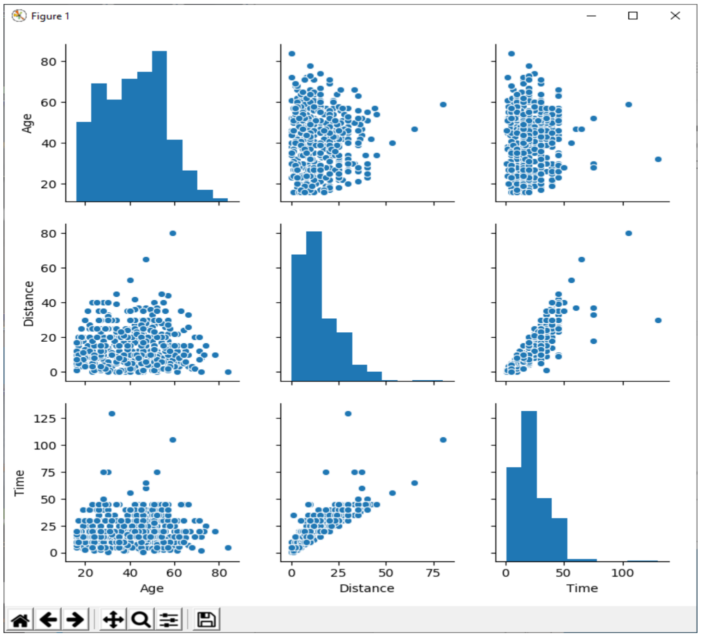
```
            Age   Distance      Time
count  500.00000  500.000000  500.000000
mean    41.38800   14.156000   21.970000
std     13.79994   10.748895   14.232436
min     16.00000    0.000000    1.000000
25%     30.00000    6.000000   11.500000
50%     42.00000   11.000000   20.000000
75%     52.00000   20.000000   30.000000
max     84.00000   80.000000  130.000000
```



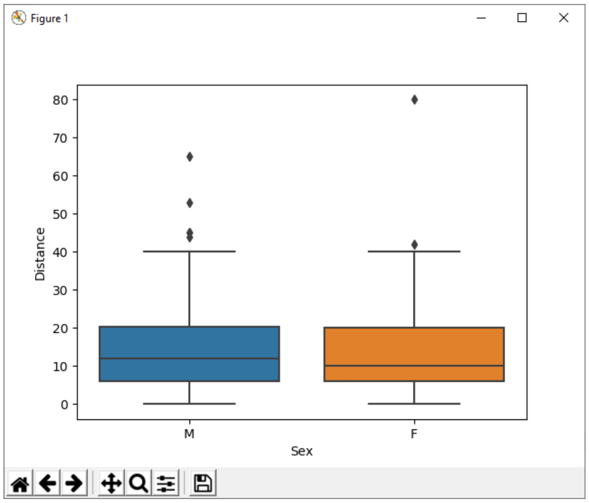
2. For the data *CommuteStLouis*:
   a. Produce a correlation matrix of age, distance and time. Which two numeric variables are most highly correlated? What is the correlation coefficient for the above pair?

```
               Age  Distance      Time
Age       1.000000 -0.000774  0.030292
Distance -0.000774  1.000000  0.830241
Time      0.030292  0.830241  1.000000
```

   b. Create a scatterplot matrix of the numeric variables in the data. What do the figures in the diagonal going from the top left to the bottom right show? What can you say about the skewness of the various attributes?

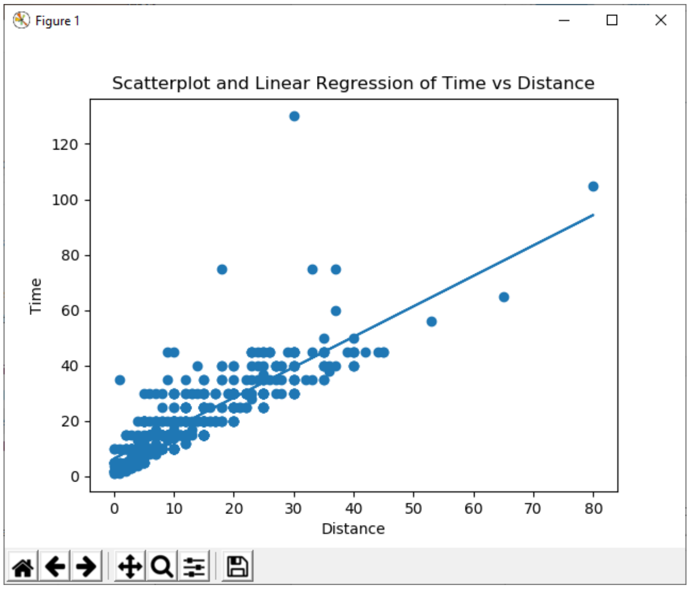


c. Produce a side-by-side boxplot of distance travelled by gender. Do the data in the file indicate that women tend to commute shorter distances?



Options: You can do Questions 3 and 4 as one figure, two subplots. Or two separate figures. Your choice.

3. For the pair in Question 2.a the scatter plot.
    a. Also superimpose a linear regression line on plot 1.



4. Show the distribution of residuals of the data from Question 3.

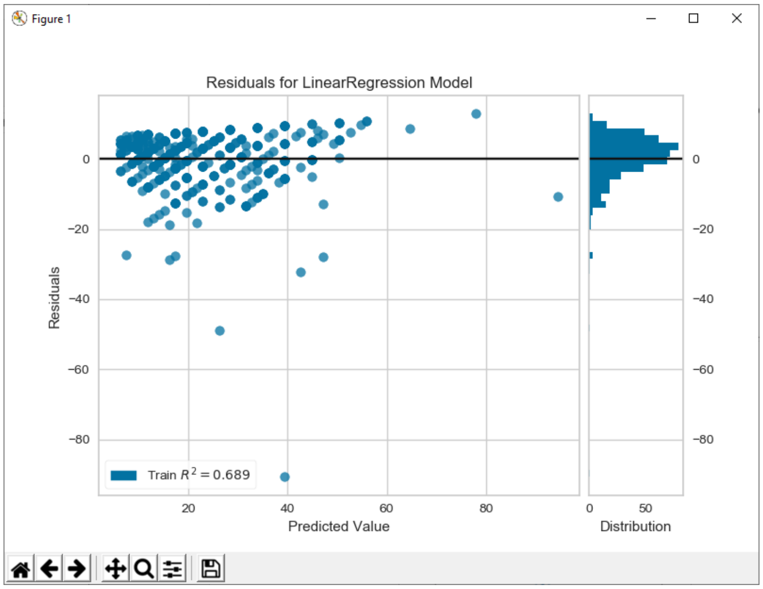


Figure 1
Residuals for LinearRegression Model
Residuals
Train $R^2 = 0.689$
Predicted Value
Distribution