Project 1

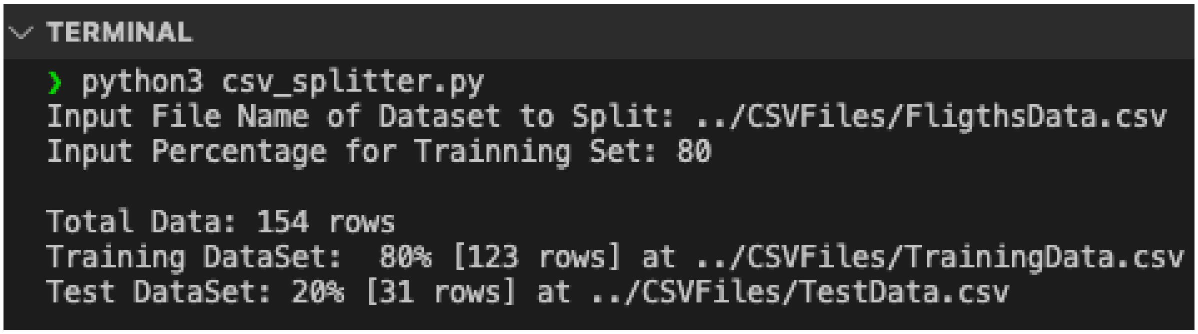
Data Collections

Data collected is from Flights from the month of March from Houston to Seattle. Data was scraped from the website Kayak using selenium and a chrome driver. The dependent variable collected was the flight **price** and the independent variables are the airline's name, departure time, departure meridiem, date and weekday. Data is collected into a pandas Data frame to write it to the csv file.

Organize Data

Slitting of the data is performed by accessing random data rows of the original csv and adding them to the training csv set first and the rest to the test set csv.

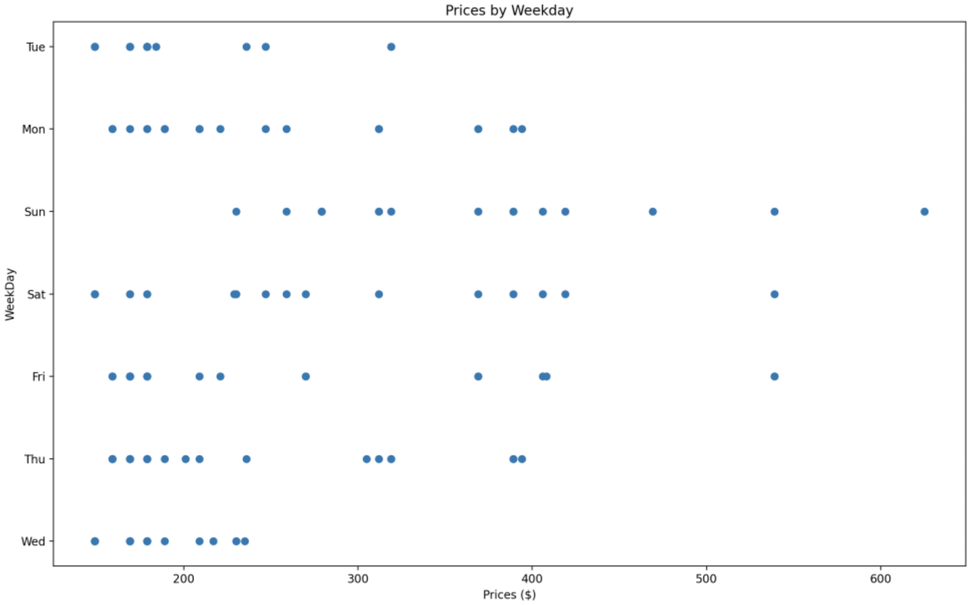
The command line should as follow:

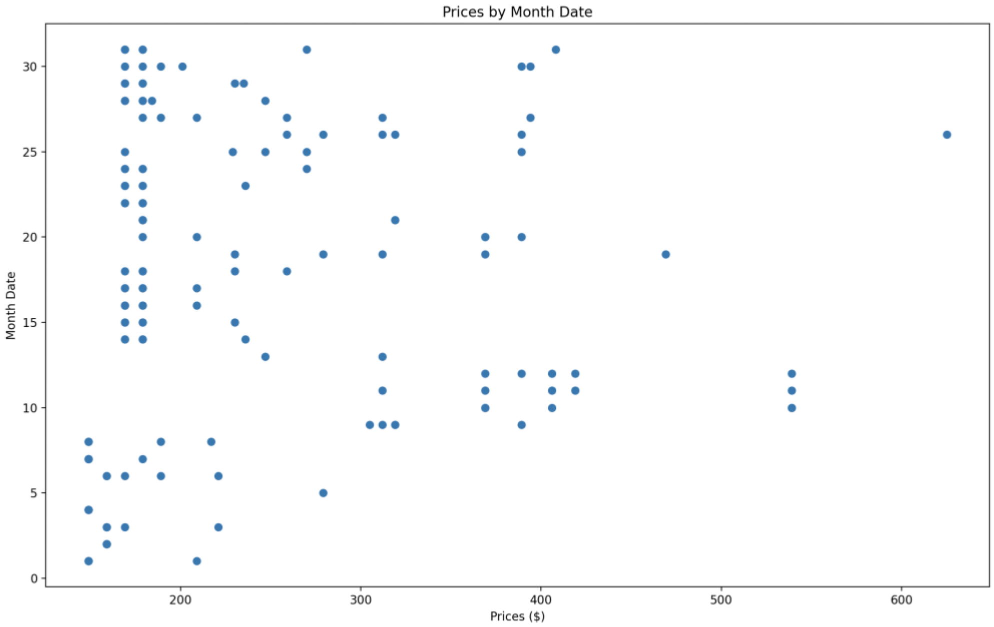


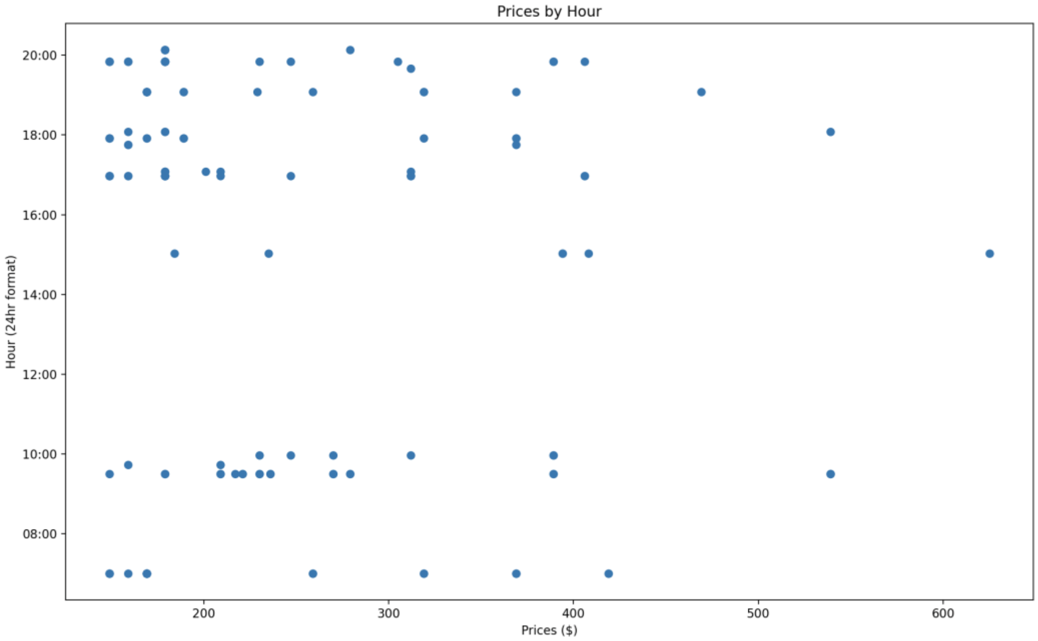
```
❯ python3 csv_splitter.py
Input File Name of Dataset to Split: ../CSVFiles/FligthsData.csv
Input Percentage for Trainning Set: 80

Total Data: 154 rows
Training DataSet:  80% [123 rows] at ../CSVFiles/TrainingData.csv
Test DataSet: 20% [31 rows] at ../CSVFiles/TestData.csv
```

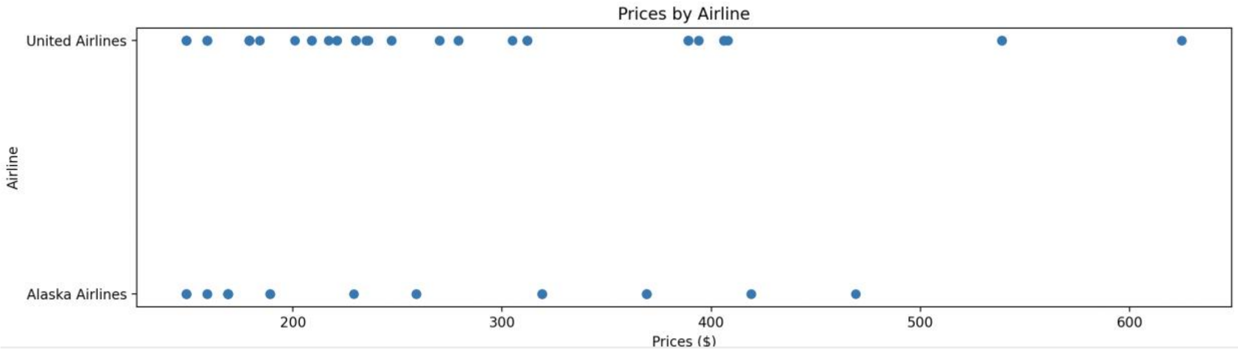
Data Plotting

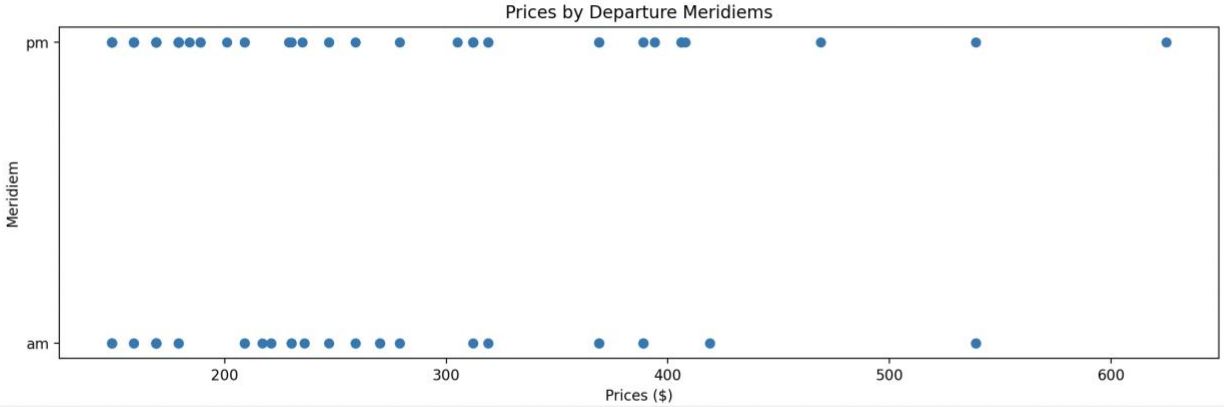
Plotting of each independent variable with the dependent variable (flight price):



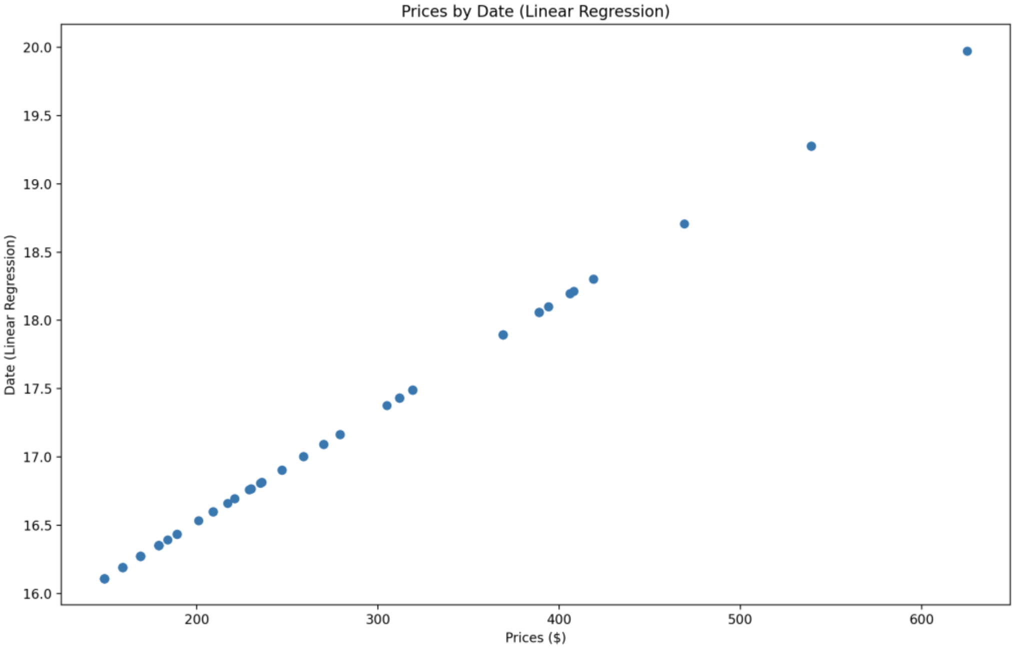

Prices by Month Date
Month Date
0
5
10
15
20
25
30
200
300
400
500
600
Prices ($)


Prices by Hour
Hour (24hr format)
08:00
10:00
12:00
14:00
16:00
18:00
20:00
200
300
400
500
600
Prices ($)





Linear Regression



Acknowledgments

- Selenium Locating Elements
- Append pandas DataFrame to existing csv file
- pandas.DataFrame.to_csv
- Linear Regression
- Scraping Prices from Kayak
- Plotting timestamps (hour/minute/seconds) with Matplotlib
- sklearn.linear_model.LinearRegression