**Citi bike Analysis**

This is the analysis of the 2017 Citi bike data. The aim is to predict the time taken to complete a journey given the start and end station.

## Motivation

This project can help the riders get a rough estimate as to how much time would they take if they are going from station A to station B given their characteristics and weather characteristics. This helps them in planning their journey and also avoiding paying extra by missing the Initial free time. This also can help the Citi bike team to get an idea of how many bikes they have per station. They can give riders a rough estimate of the arrival time of a bike at a station if the station has no bike and also suggest riders to change their end destination if the station would be full by the time they would get there.

## Built With

* Jupyter- Interactive computing environment
* Anaconda Distribution- Open source distribution of the Python
* Citi bike - Used to generate Citi bike Data
* Weather Underground – Used to get Historical Weather data

## Features

This notebook is divided into 4 sections:

1. Imports :- Grouping all the imports together so that anyone who needs to run this code knows what all packages are required.
2. Run Time File Reading:- Created a way to read the file directly from the Citi bike server. This will help if the code needs to be automated and run every month when new files release. This saves a pickle file which is read back in section 3. This pickle file acts as a checkpoint as well so that not everyone needs to run section 2 if they have output of section 2.
3. Exploratory Analysis:- This is where the initial cleaning and some features engineering is done. Also, in this section the visualizations are created to answer the questions presented by the Mayor. At the end of this section a pickle file is saved which is read back in section 4. This pickle file acts as a checkpoint as well so that not everyone needs to run section 3 if they have output of section 3.
4. Modelling Phase:- In this section first the Citi bike data is merged with weather data and some feature are created for patterns observed while doing Exploratory Analysis in step 3. Now since the behavior of subscriber and customer as so different that I have created two different models, unique for each user type.

**Section 1: Imports**

All imports are combined in this section to help readers to understand the packages this code requires and so the user can download them beforehand.

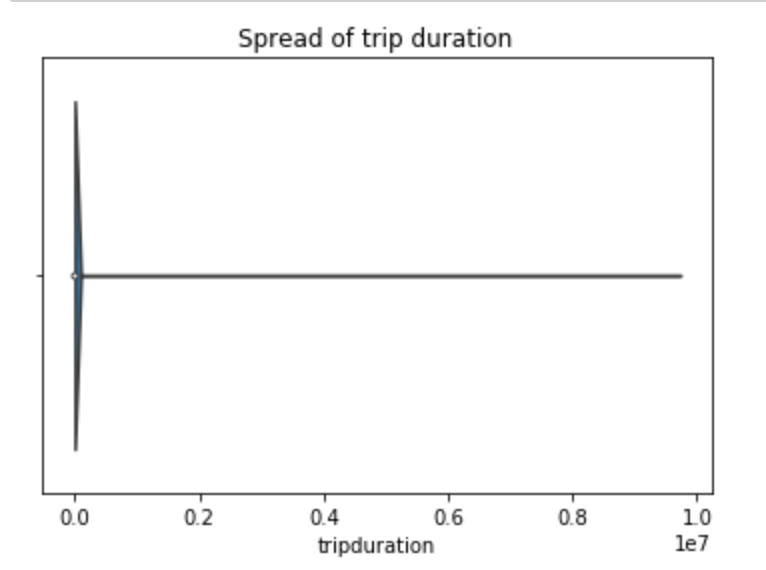
**Section 2: Run Time File Reading**

In this section I have made a way to download the files directly from the Citi bike website for the Citi bike data. This code handles inconsistencies in date format and works fine with 2017 data. This code handles inconsistencies in the naming convention in columns and works fine with 2017 data. The weather data currently is being taken from weather underground website as it is freely available but that can also be automated by using the weather.com api in Bluemix console (it was not done in this version as that is not freely available).

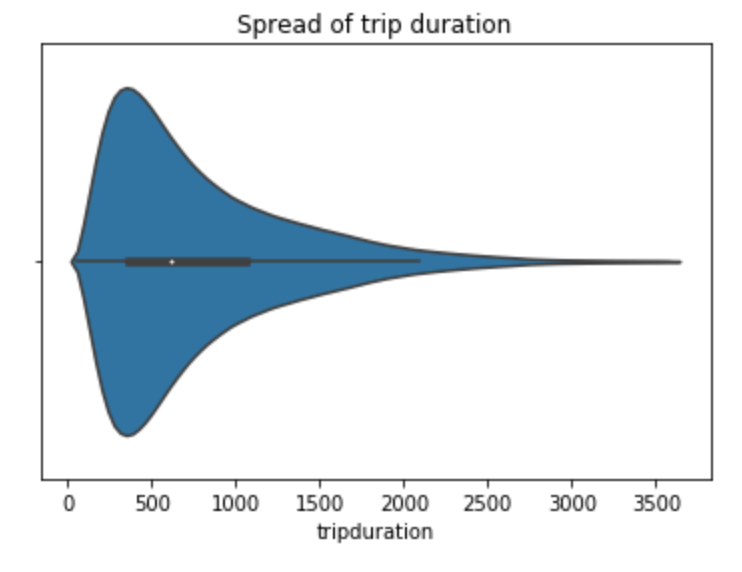
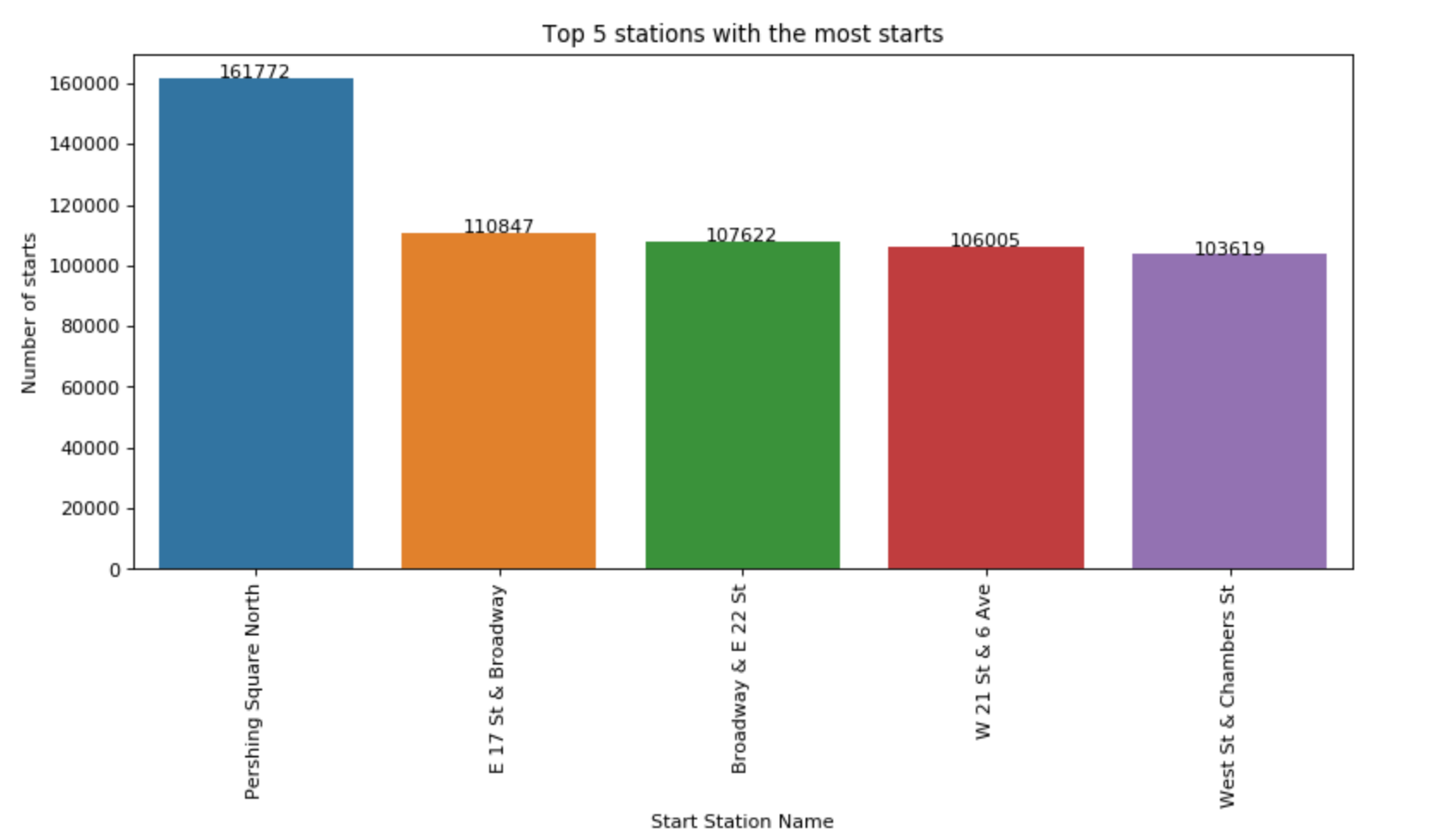
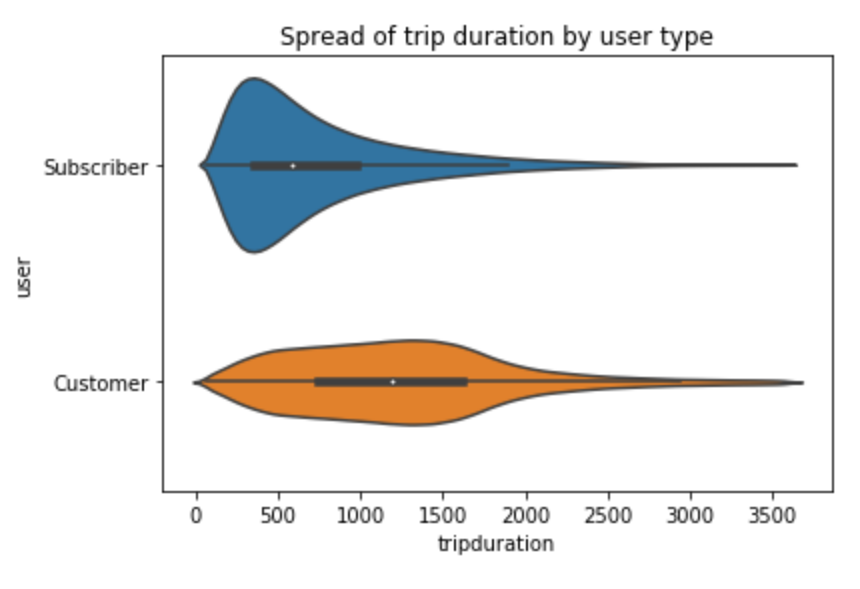
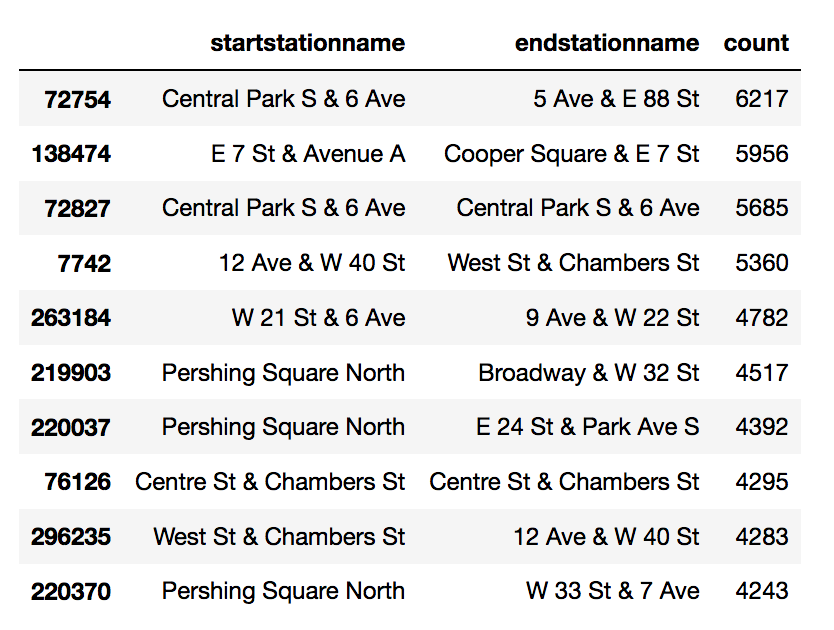
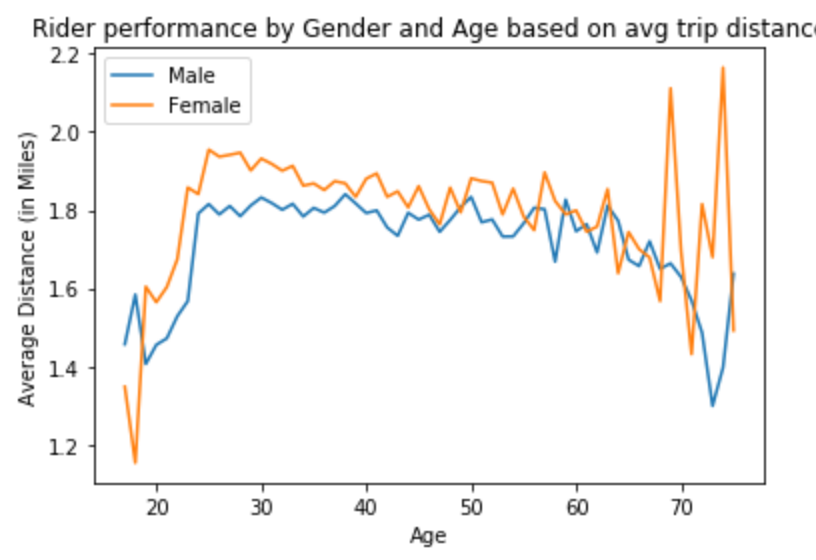
**Section 3: Exploratory Analysis**

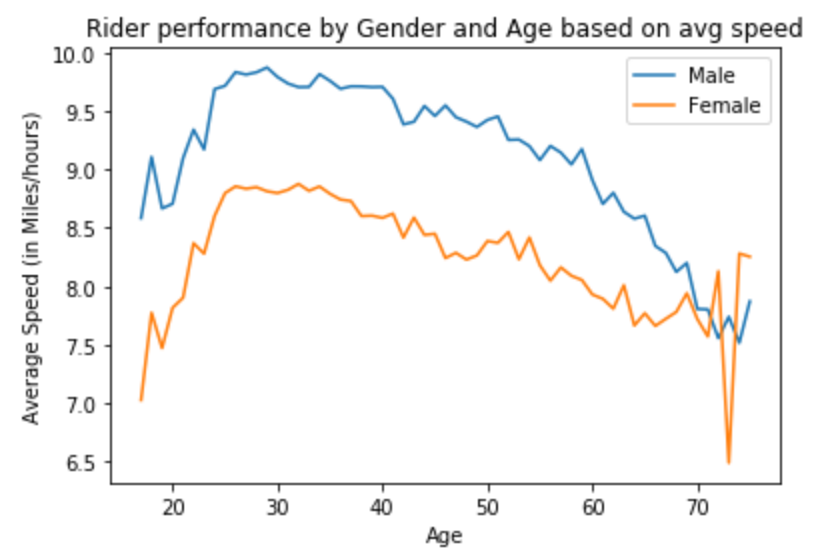
This section is further divided into 4 parts

1. In this part the file is read that was created in the previous section and any rows that have all empty values is dropped.
2. In this part, I start by calculating distance, speed and age. Then according to the data dictionary mentioned by Citi bike, Subscribers should have gender and birth year in the data and the Customer can’t have any. So, based on this preliminary cleaning is done. The spread of trip duration is plotted (using violin plot) to understand the spread of the data. Below is initial spread.

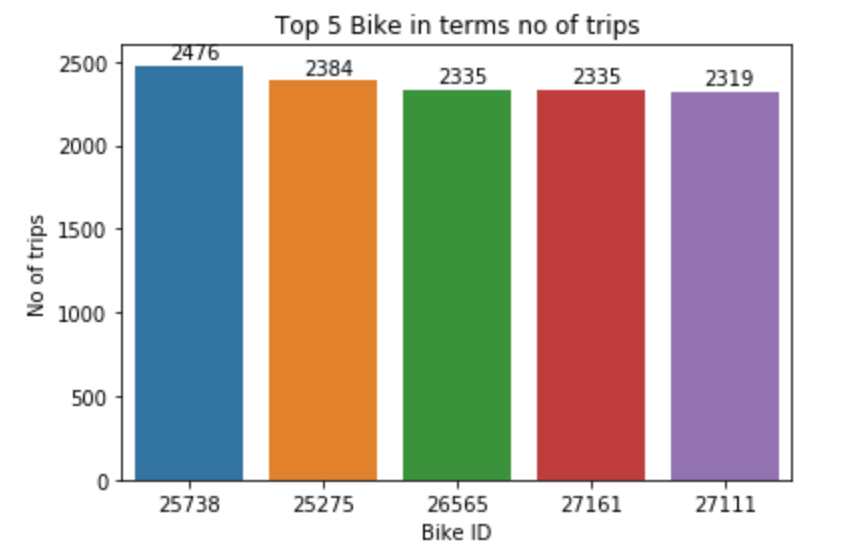


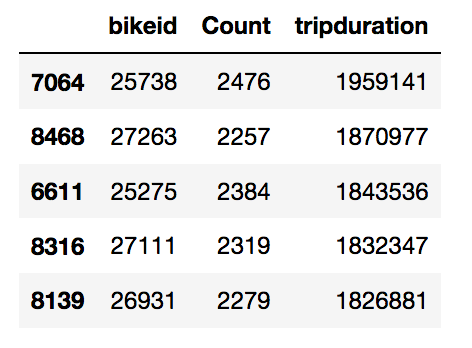
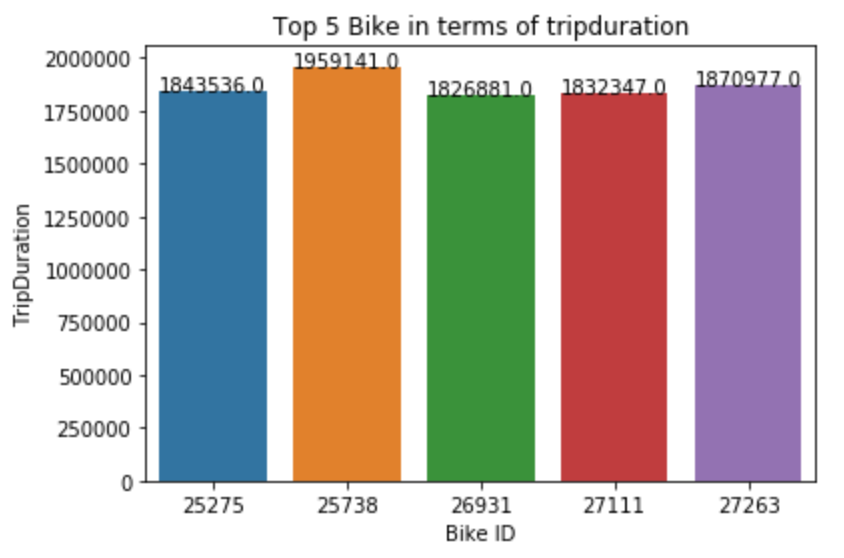
This clearly showed that duration had a lot of outliers. Also, the split for user type and gender was plotted.

1. In this part the cleaning operations were performed. First distance greater than 10 miles were removed as it is really hard to bike more than 10 Miles. Second, Trip duration greater than an hour was removed based on the assumption that people would change bikes if they were going for long distance instead of paying additional charges (since for charges are applicable for customers after first 30 mins and for subscribers after first 45 min in intervals of 15 minutes). Then I further cleaned by removing trips where the age of the rider was above 75 years as people older than that was highly unlikely, it seemed more of an outlier or error in capturing the data. After performing this cleaning, the trip duration was again plotted, and the result is below.
2. In this part, I create the visualization to answer the questions presented by the Mayor. Below are the questions and the visualizations created to answer that question.
3. Top 5 stations with the most starts (showing # of starts)
4. Trip duration by user type
5. Most popular trips based on start station and stop station)
6. Rider performance by Gender and Age based on avg trip distance (station to station), median speed (distance traveled / trip duration)



1. What is the busiest bike in NYC in 2017? How many times was it used? How many minutes was it in use?

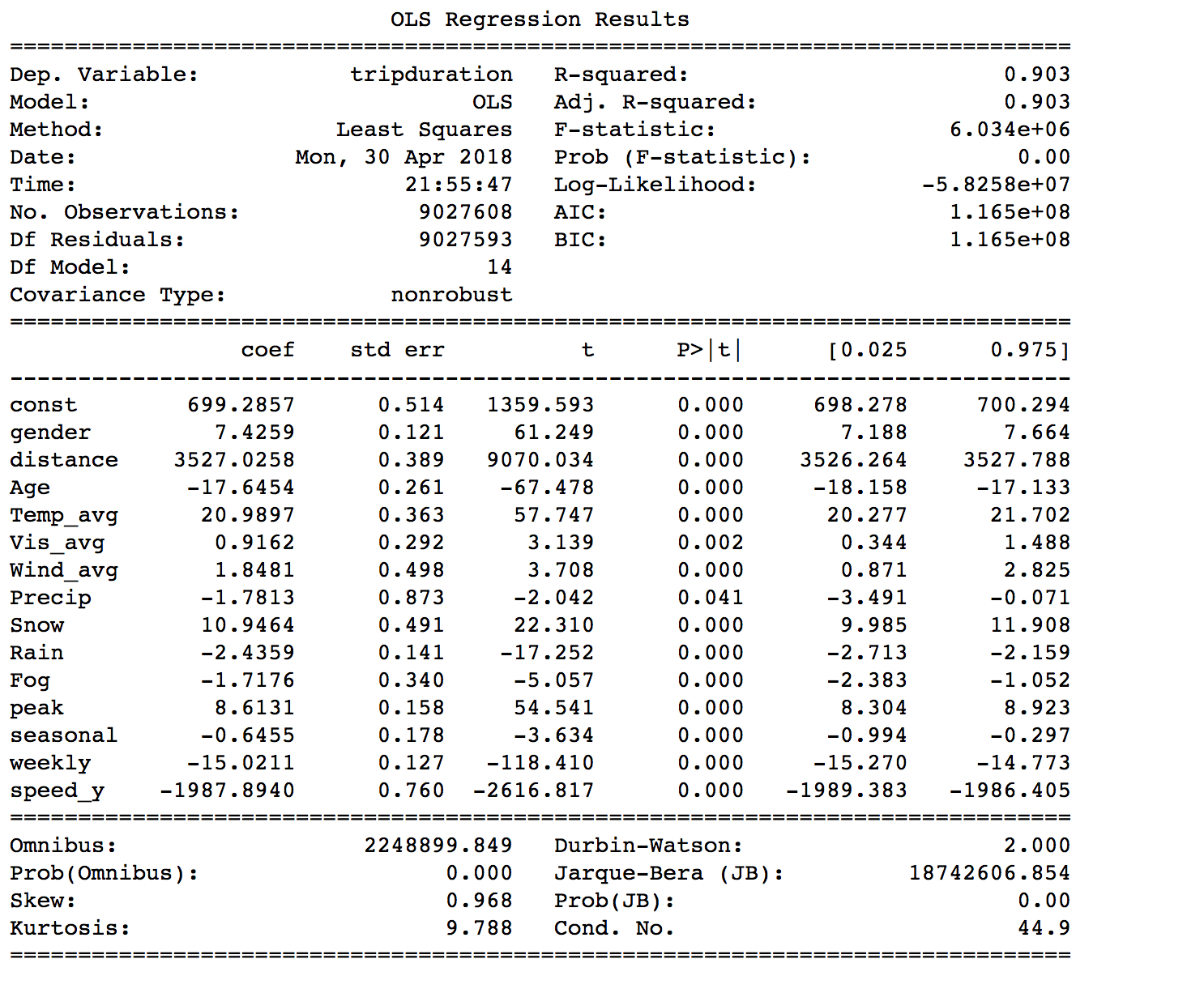




**Section 4: Modelling Phase**

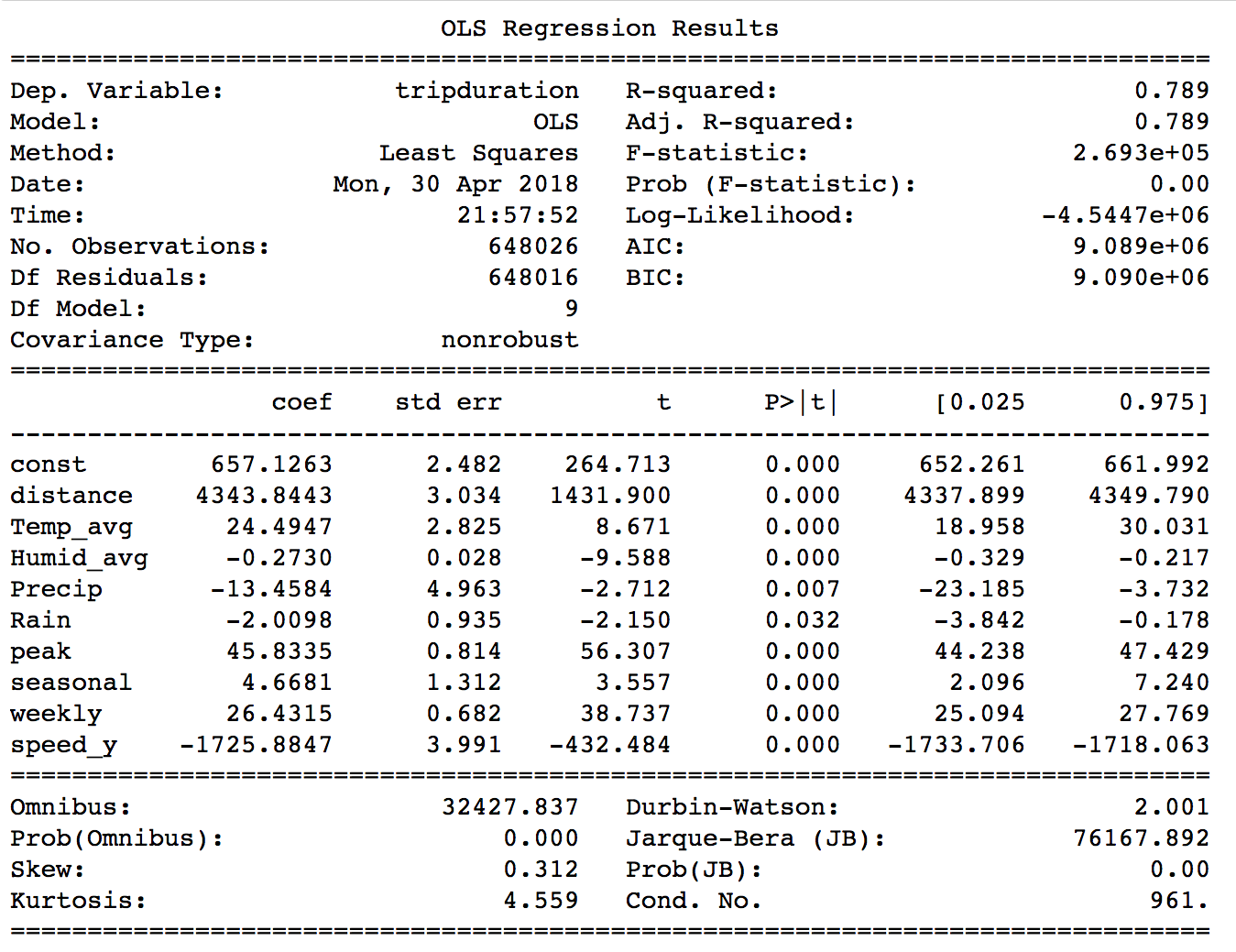
This section is further divided into 4 parts

1. In this part I start by loading the dataset created in the previous section and the weather data and did some analysis to see if there is any trend for user type based on hour, month and day of the week. Also, the Citi bike and weather data are merged in this part
2. In this part some additional features are created based on the analysis done in the previous step. These features reflected the historical trend of most usage for each user type based on day of week, hours and month. It was interesting to find out that customers tend use the bike more on weekend while subscriber use it more on weekday. These also gave a sense of the traffic patterns and seasonality which were not there in the data directly. Also, some records were removed as they were round trips, so they added no information in the model and some had irrational speed. The threshold for speed was assumed to be 5 mph to 20 mph. NYC follows a policy of 25 mph for streets and assuming bikes drive slower than car hence 20 also average human walking speed is 3.5 mph, so a rider would pedal at least 5 mph as it is hard to pedal at such low speeds. Here also the average speed for subscribers and customers are created since actual speed can’t be used. The average speed for subscribers is calculated based on their age, gender and route. The average speed for customers is calculated based on their route.
3. In this part the model for subscriber is created. I first merge this Dataframe with Average speed df to bring in average speed for subscribers based on their route, age and gender. Then, I drop first the columns based on their correlation with the dependent variable and with other independent variable. After the initial model result, some variables were dropped being insignificant based on P – Value. This model is evaluated based on R2(Variance Score) which is 0.90. The result for this model is below.



After getting this result I checked Ridge, SGD Regressor and Lasso regression models to make sure that I wasn’t overfitting and do some dimensionality reduction. Lasso helped in reducing some variables and then the model was cross validated to make sure over fitting was not happening. The cross validated r2 is 0.903.

1. In this part the model for customers is created. I first merge this Dataframe with Average speed df to bring in average speed for customers based on their route. Then, I drop first the columns based on their correlation with the dependent variable and with other independent variable. After the initial model result, some variables were dropped being insignificant based on P – Value. This model is evaluated based on R2(Variance Score) which is 0.79. The result for this model is below.

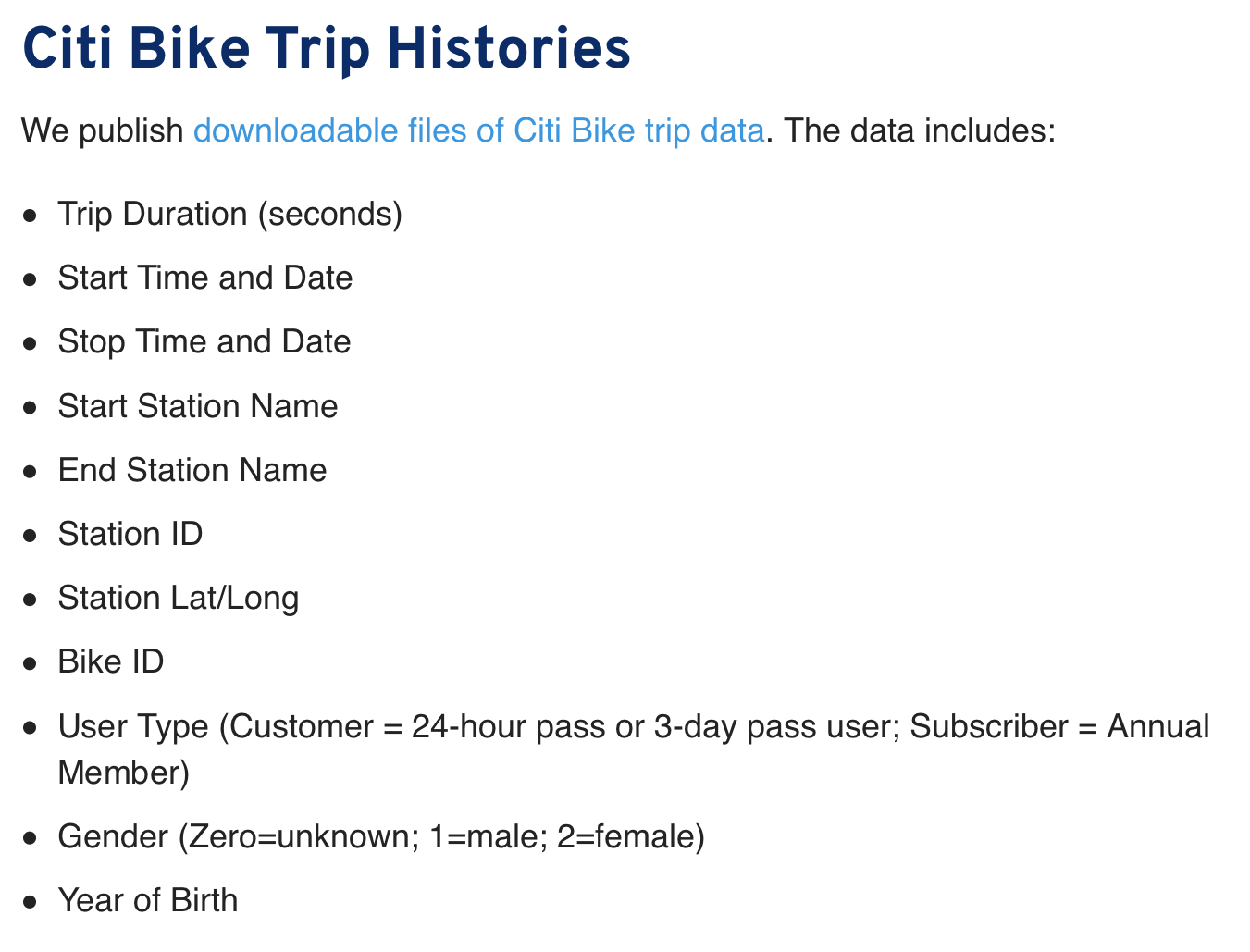


After getting this result I checked Lasso regression models to make sure that I wasn’t overfitting and do some dimensionality reduction. Lasso helped in reducing some variables and then the model was cross validated to make sure over fitting was not happening. The cross validated r2 is 0.787.

## Citi bike Data Dictionary

The data dictionary for the Citibike data is mentioned on this [link](https://www.citibikenyc.com/system-data).

Below you can find the screen shot for the same.



## How to use?

To use this code, you will need Python 3 along with the Jupyter notebooks. All the packages that are used in this code need to be installed, I prefer installing the Anaconda Distribution of python as it would take care of all this. To download Anaconda [click here](https://anaconda.org/). All the work done for this analysis is present in this [link](https://ibm.box.com/s/xn1icip0qn1biib09y67tnibq0j1q216).

**Versioning**

This is the first version of this work.

**Authors**

* **Aishwarya Ajay Das**- *Initial work*

**Acknowledgments**

* <http://toddwschneider.com/posts/a-tale-of-twenty-two-million-citi-bikes-analyzing-the-nyc-bike-share-system/>
* Predicting Bike Usage for New York City’s Bike Sharing System - Divya Singhvi, Somya Singhvi, Peter I. Frazier, Shane G. Henderson, Eoin O’ Mahony, David B. Shmoys, Dawn B. Woodard
* Demand and Trip Prediction In Bike Share Systems - Zhaonan Qu SU
* Bicycle-Sharing System Analysis and Trip Prediction - Jiawei Zhang, Xiao Pan, Moyin Li, Philip S. Yu
* Efficient Journey Planning and Congestion Prediction Through Deep Learning - Muhammad Shalihin Bin Othman, Sye Loong Keoh, Gary Tan

## Contribute

Please read CONTRIBUTING.md for details on our code of conduct, and the process for submitting pull requests to me.

## License

This project is licensed under the MIT License - see the LICENSE.md file for details.