1. Data Preprocessing

I. Missing data

The original data set contains 726676 observations and 15 variables. For Bike ID and Birth Year, there are some missing values:

Variable	Number of Missing Values	Percentage in the whole data set
Bike ID	3193	0.44%
Birth Year	29076	4%

Because the numbers of missing values are not large, so my solution is to drop these missing values.

II. Abnormal data and outliers

(1) For Start Station Latitude, Start Station Longitude, End Station Latitude, End Station Longitude, there are some observations equals to 0. After I calculated the distance by using these four variables, I dropped those observations with the distance of 0 because there might be a data error, or a round trip, for which we can not predict the drip duration.

(2) I deleted the observations with the gender variable equals to 0 (unknown). Since we want to know how gender affects the trip duration, unknown observations cannot help us to build a model.

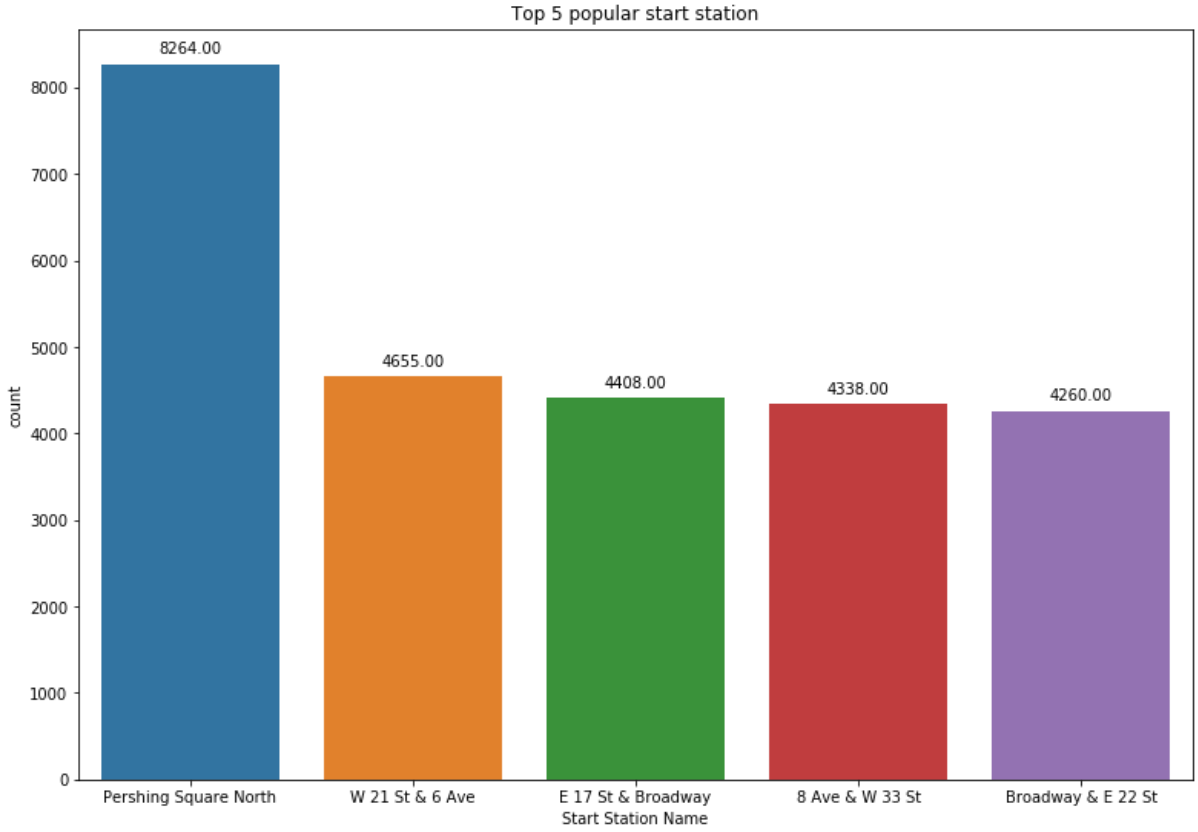
(3) After plotting trip duration, I observed that there are some abnormal values, such as 61 seconds (1 minute) and 5325688 seconds (1480 hours). I assumed that the time for Citi Bike trip durations should be greater than 4 minutes and within 5 hours, so I deleted the observations with trip duration less than 240 seconds and greater than 18000 seconds.

(4) After calculating age using the birth year, there are some age greater than 80 years old. I considered them as data error so I deleted the observations with age greater than 80.

2. Data Visualizations

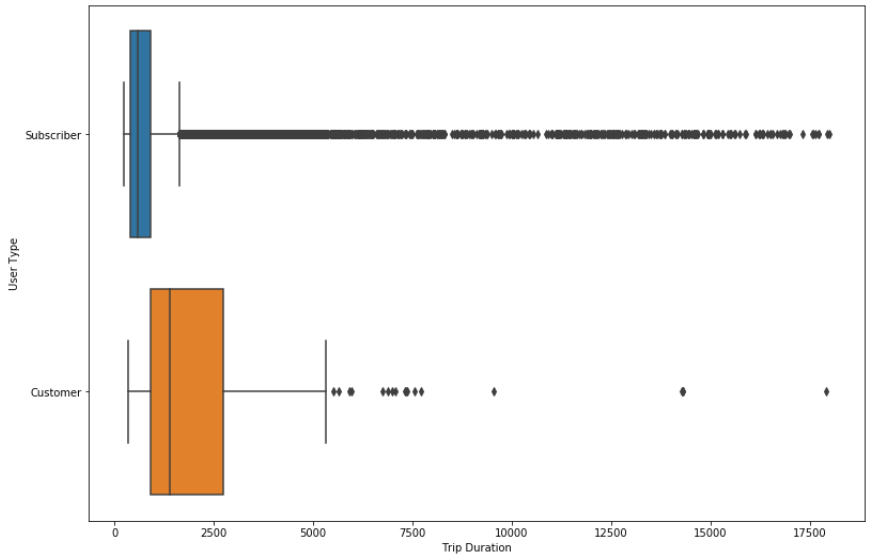
From the data set, I found some interesting patterns:

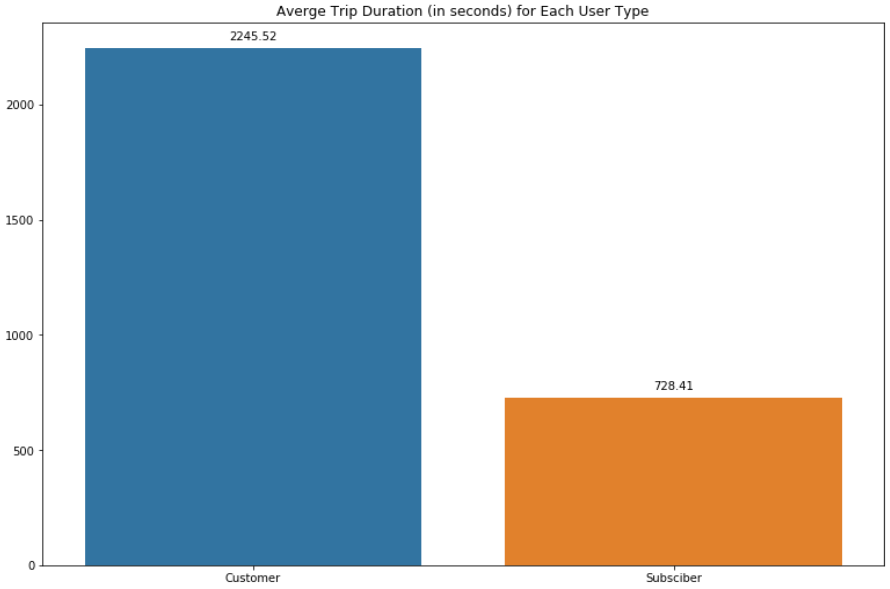
(1) Top 5 stations with the most starts:



From the plot, we can tell the most popular start station is Pershing Square North, with 8264 records during Jan 2017. This analysis can help Citi bike to allocate their bikes.

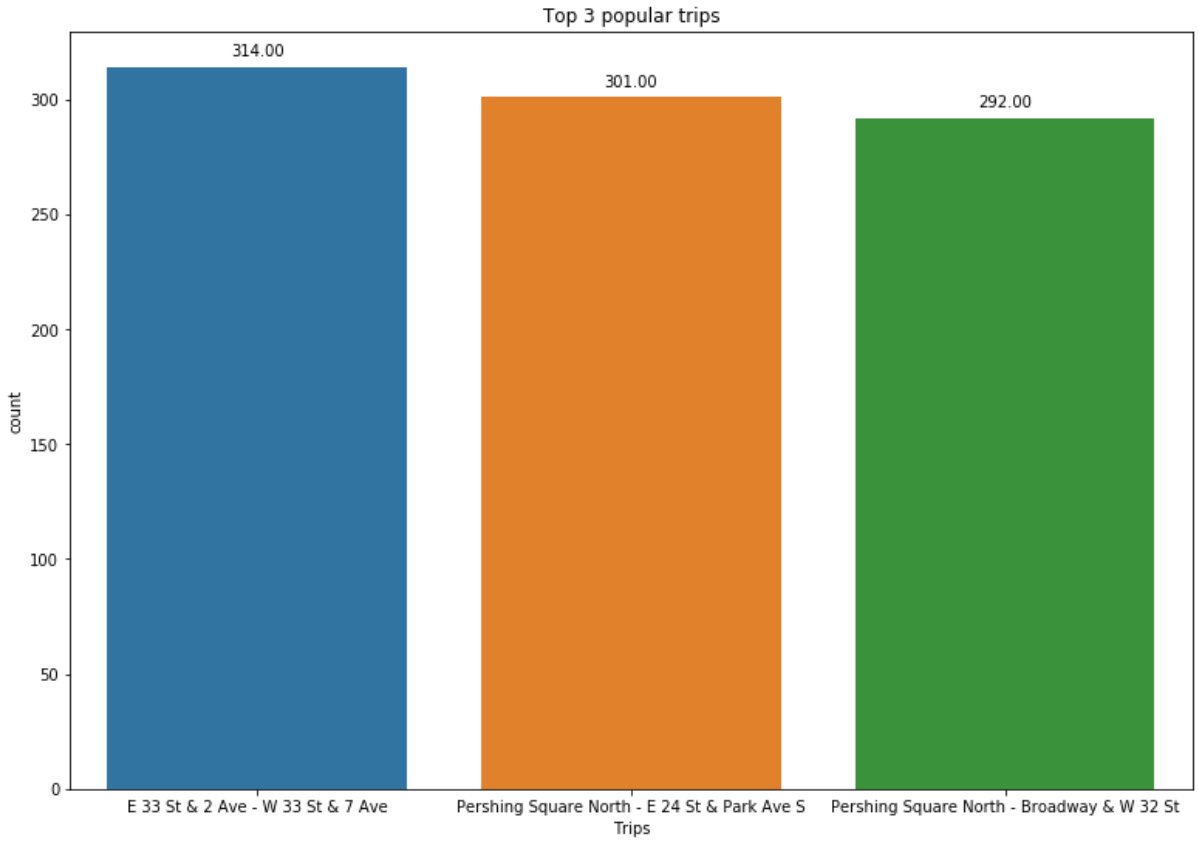
(2) Trip duration by user type





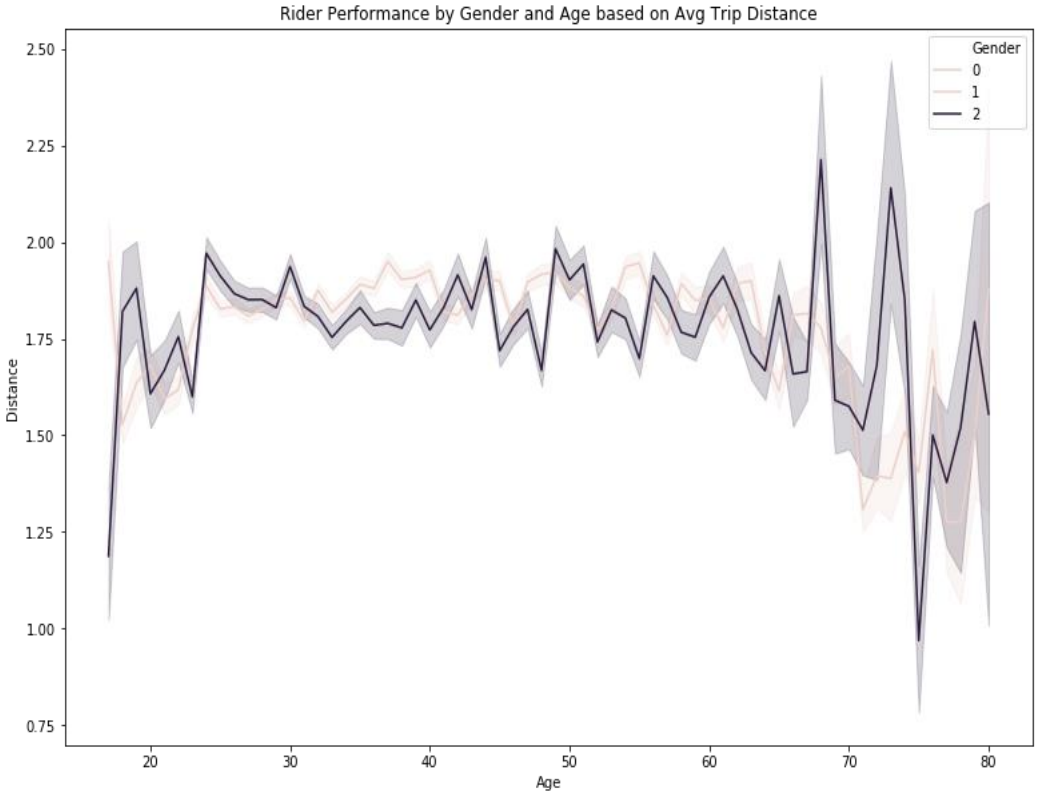
From the plot, we can tell customers have longer average trip durations than subscribers.

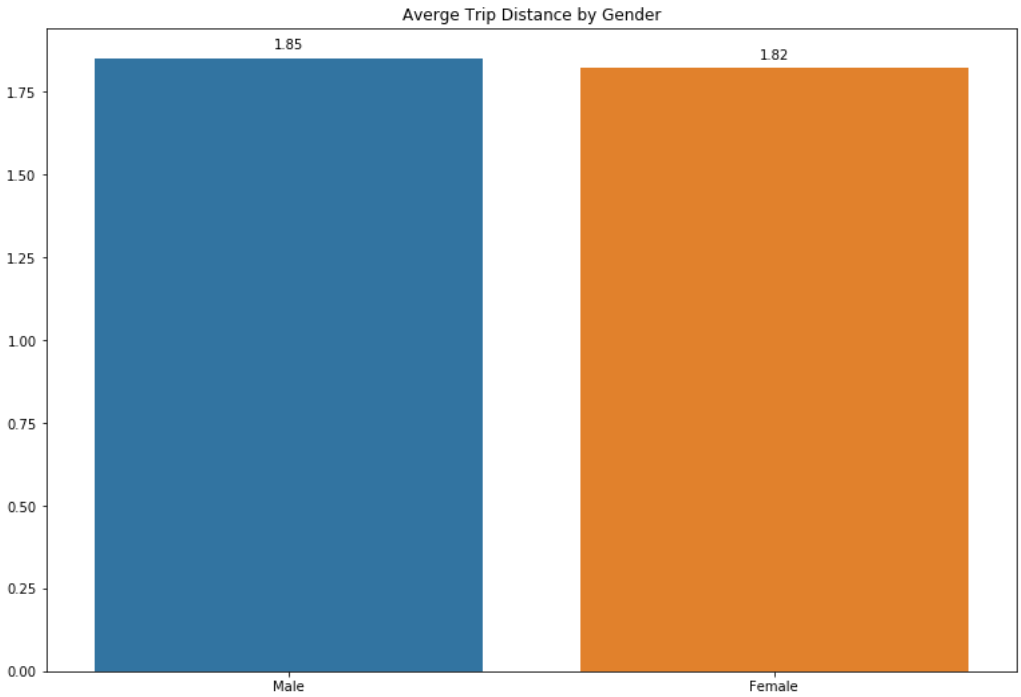
(3) Most popular trips based on start station and stop station (after delete round trip):



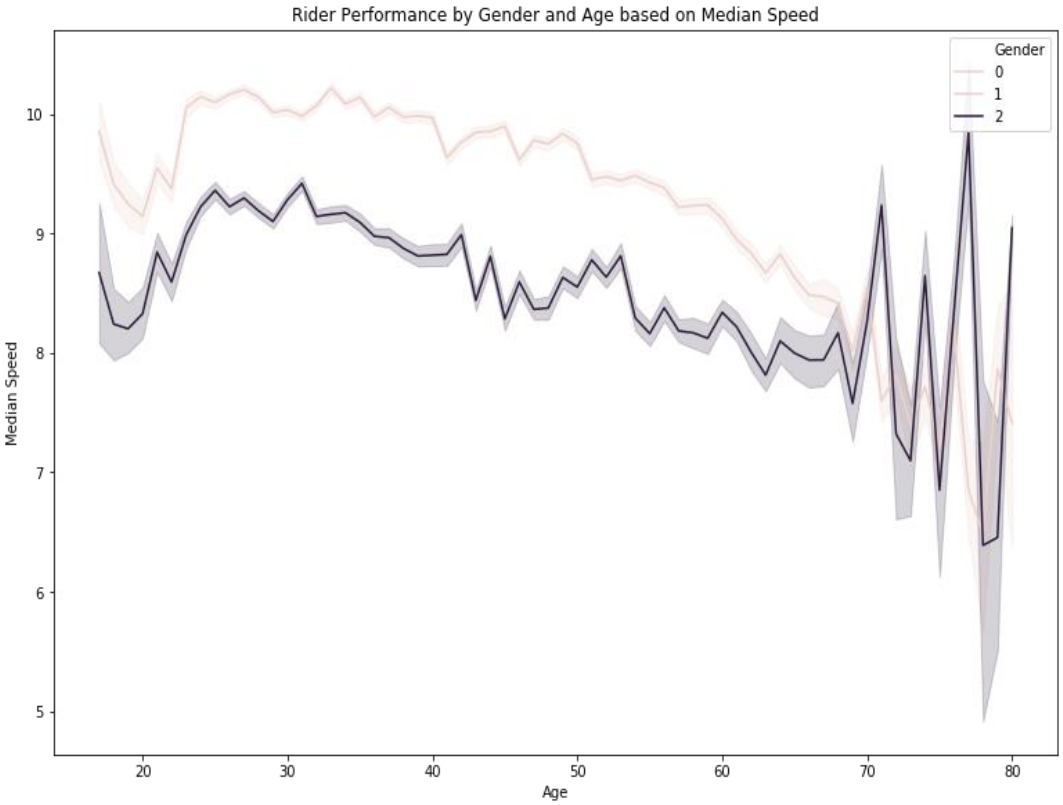
The most popular trip is from E33 St & 2 Ave to W 3 St & 7 Ave.

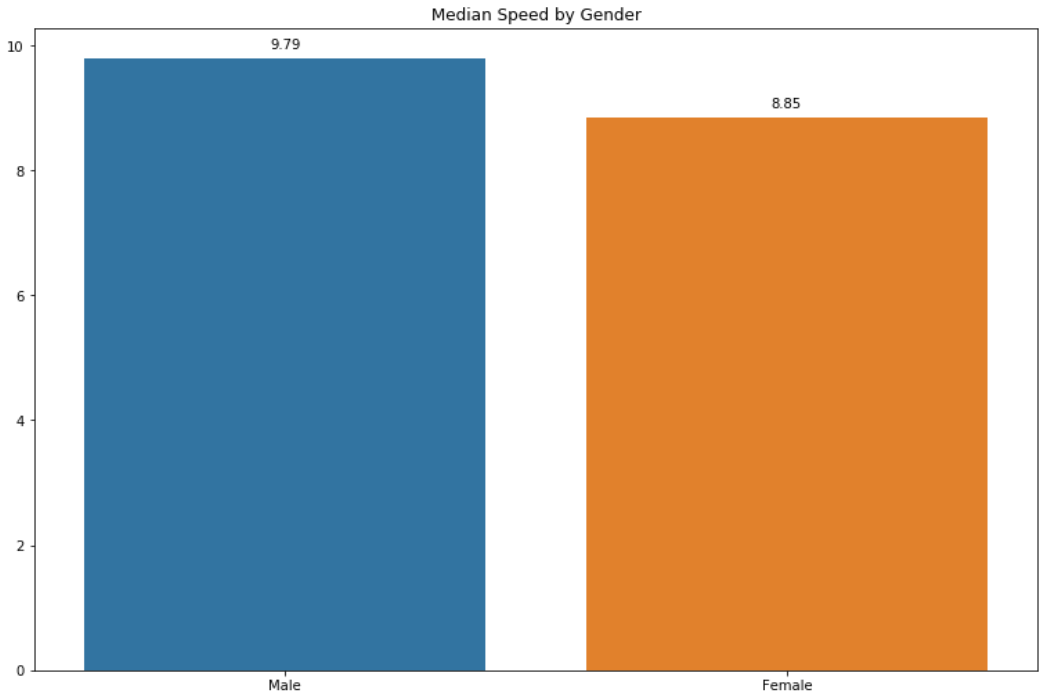
(4) Rider performance by Gender and Age based on average trip distance (station to station), median speed (distance traveled / trip duration)





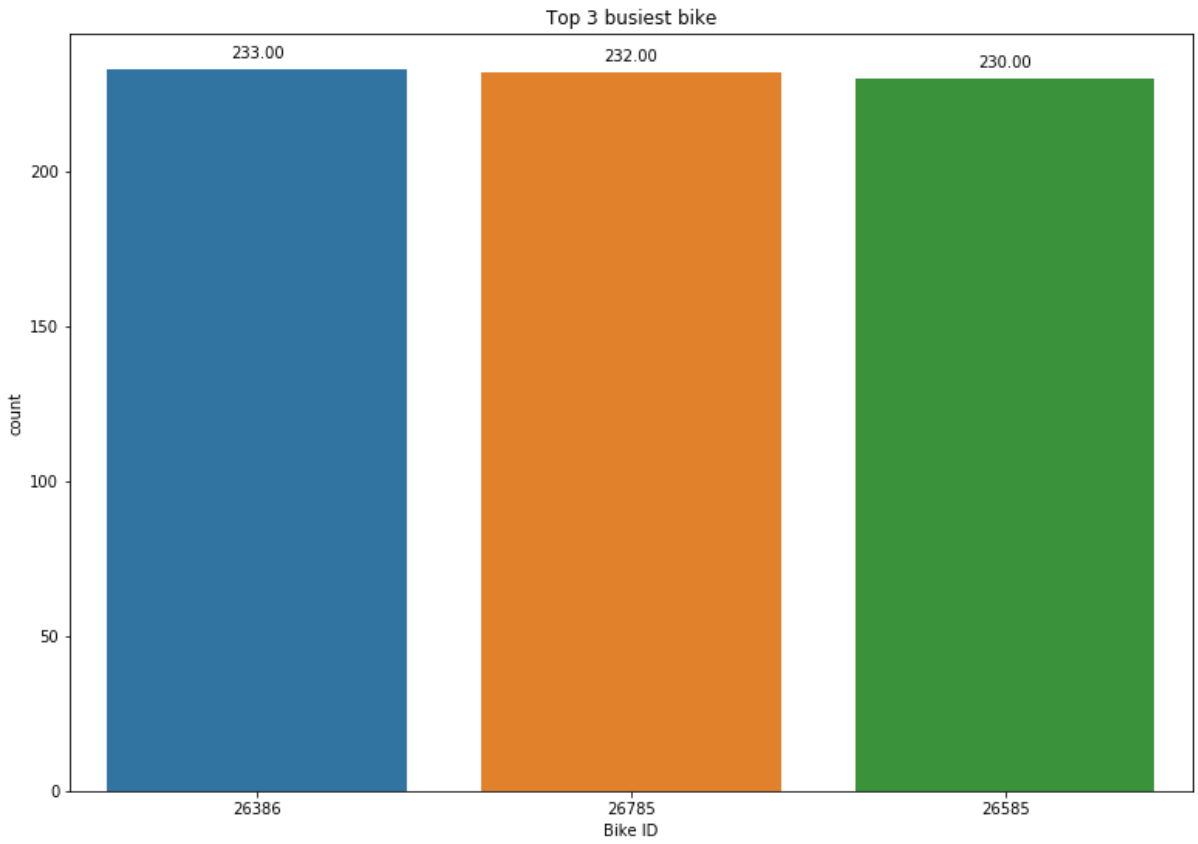
We can tell that for average trip distance, gender 1 (Male) is slightly greater than gender 2 (Female). Age does not affect the average trip distance too much for both males and females between 20 and 70 years old.





For median speed, Male is slightly faster than Female. Also, Age does not affect the median speed too much for both males and females between 20 and 70 years old.

(5) What is the busiest bike in NYC in 2017? How many times was it used? How many minutes was it in use?



The busiest bike in NYC on Jan 2017 is Bike ID 26386. It was used 233 times with 2882 minutes in total.

3.Prediction Model

I want to build a model that can predict how long a trip will take given a starting point and destination. The model I used is KNN Regressor. In KNN Regressor, the target is predicted by local interpolation of the targets associated with the nearest neighbours in the training set. The predictors I chose are Distance, User Type, Gender and Age, and the target variable is Trip Duration. The KNNRegressorModel

function fits the model on the training data set, and the KNNRegressor function predicts the target variable (Trip Duration) based on the test data set and returns the predictions. The Mean Square Error (MSE) for the model is 0.00053, which means the model has high accuracy.

When a Citi Bike users start to use the bike, we can calculate users' trip distance based on the start station and the end station/destination. We also can access users' gender, type and age. Then we fit these pieces of information into the KNN regressor model, the model can give the users the estimated trip duration.

4. Further business thoughts:

There is a further analysis I am also interested in:

Since the traffic and weather also can affect the trip duration, we can also collect and fit the traffic and weather condition in the model to predict the trip duration.