NEXT PROFITABLE SHARE STATION-

ANALYZING THE CITI BIKE SHARE USAGE IN NYC

INTRODUCTION

- GOAL: Using New York City 2015 (5-year estimate) census data to analyze the next most profitable location to add a bike share station.
- APPROACH: Merge the 3 datasets together, determine the distance within the boroughs and the stations. Using the stations closest to the boroughs to create a models.
- RESULT: Using the Gradient Regression Boost, Queens has the highest profitable possible stations.

BACKGROUND ON CITI BIKE SHARE

CITI BIKE SHARE IS THE NATIONS LARGEST BIKE SHARE PROGRAM

- 12,000 bikes and 750 stations within Manhattan, Brooklyn, Queens and Jersey City.
- Designed for quick, convenient , and affordable trips
- Have Annual Member or Day Passes
- One year of bike share is cheaper than two monthly subway passes
- Can be more convenient than owning your own bike (no locks, no storage needed)

SCOPE

- Analyze bike usage in the five boroughs based off of census data
- Using NY Census Data predict the next most profitable location to add a bike share station.
- Linear Regression, Random Forest Regression, Gradient Boosting Regressor, PLSRegression

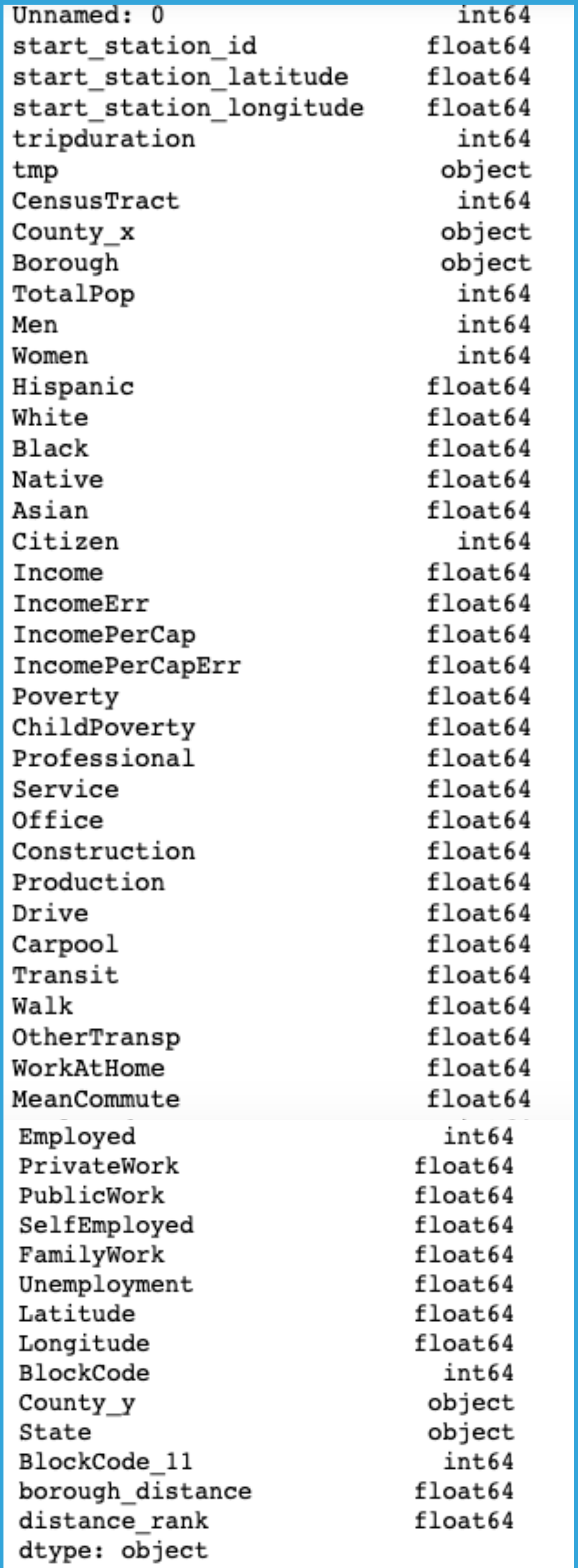
DATA:

- Data Set 1: from Kaggle-https://www.kaggle.com/muonneutrino/new-york-city-census-data

- Data Set 2:from Citi Bike Share- https://www.citibikenyc.com/system-data

Two Datasets

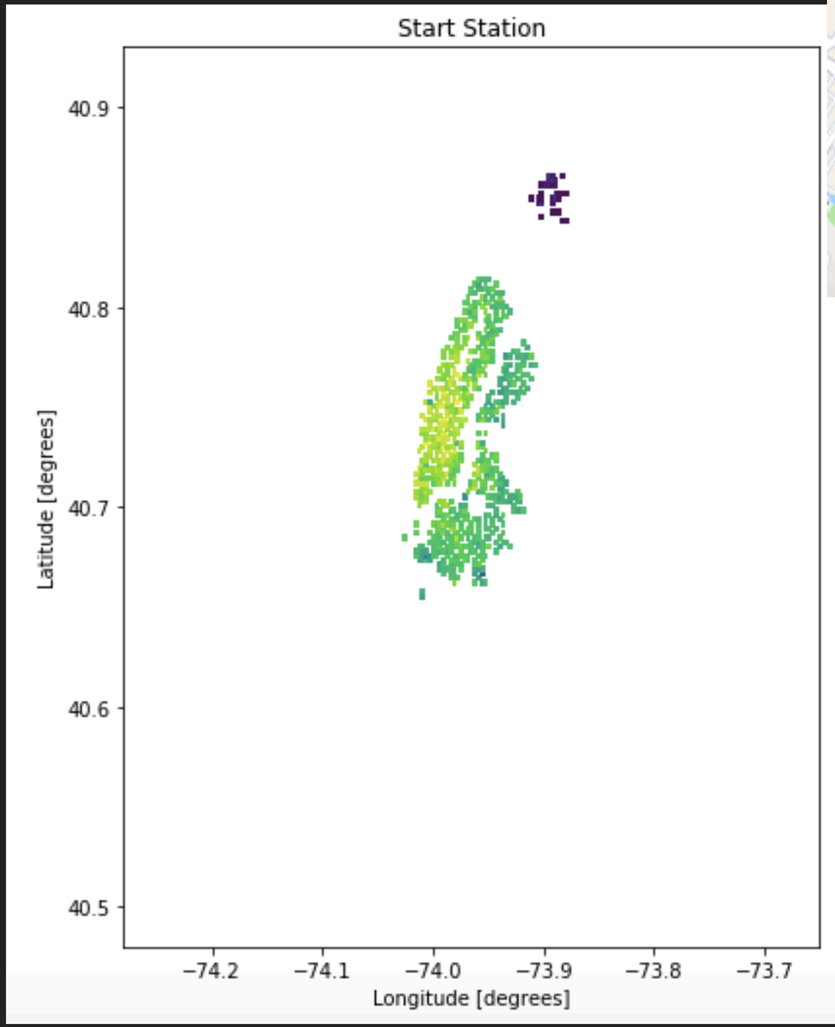
Merged:

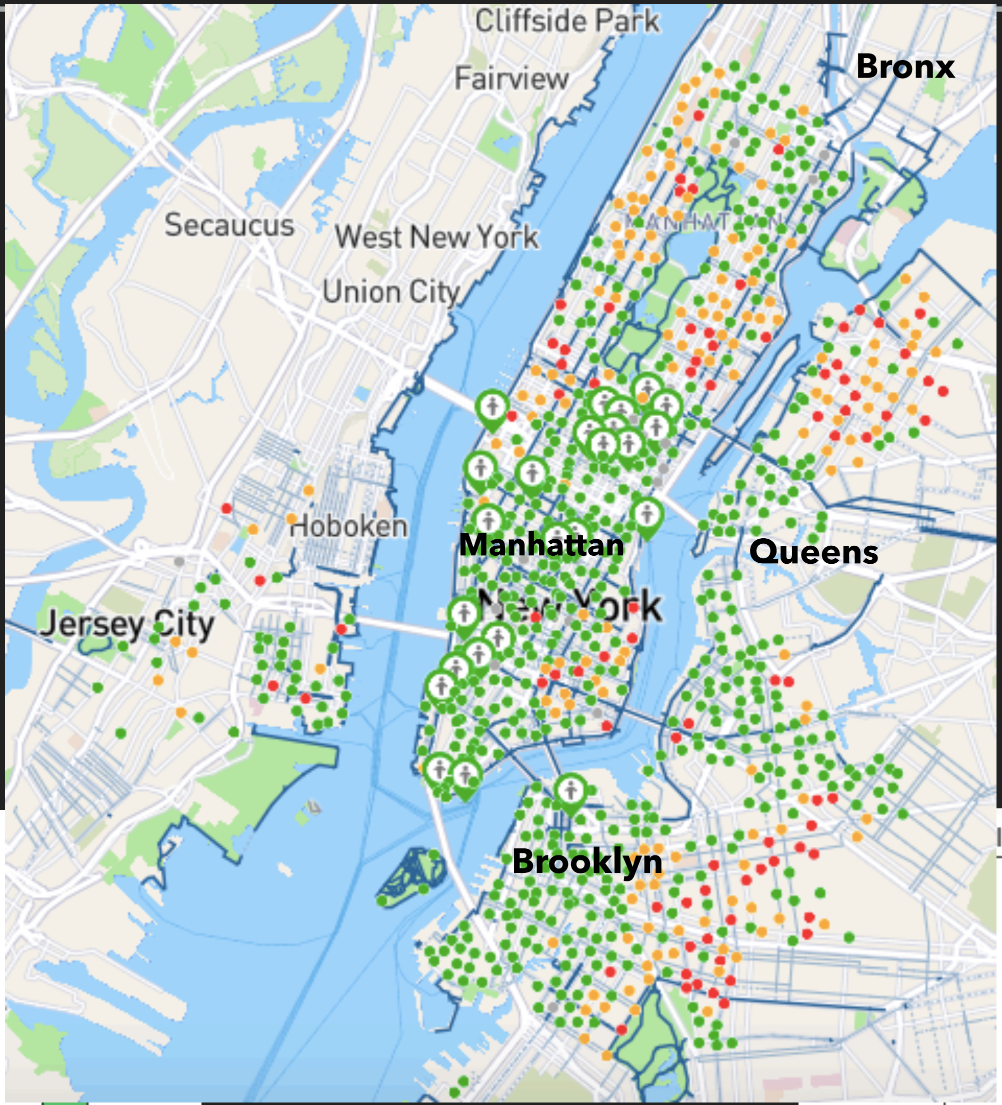


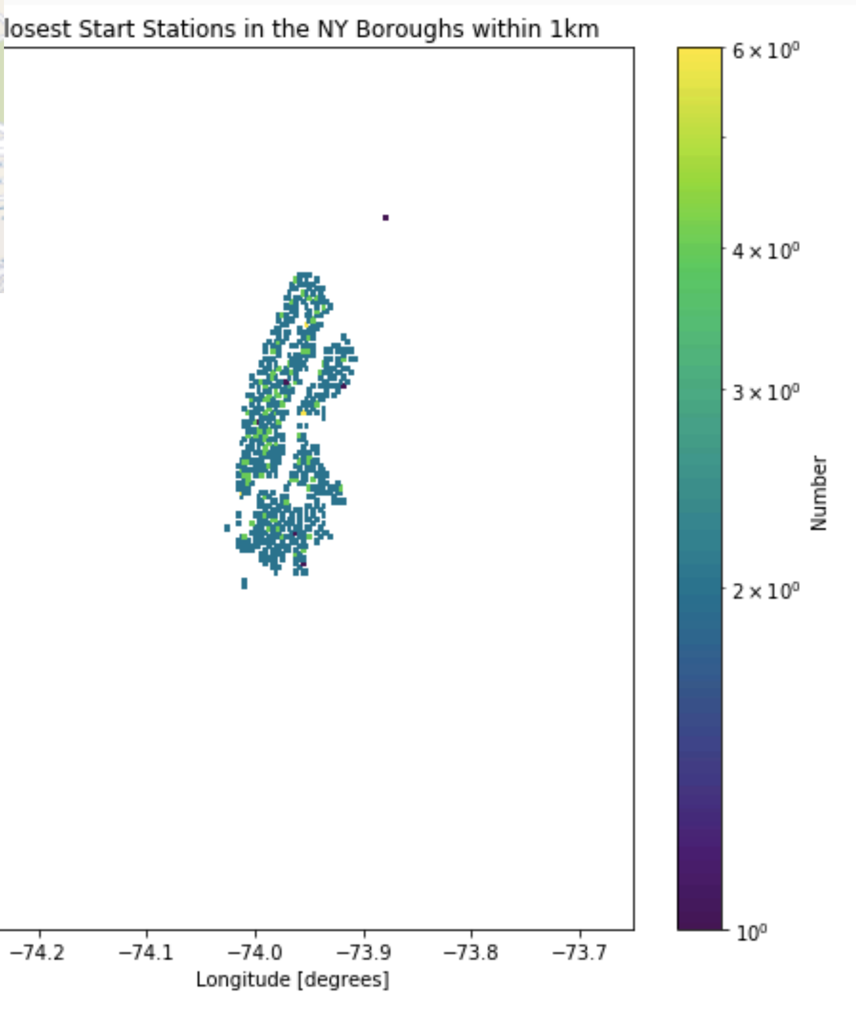
```
Unnamed: 0                    int64
start_station_id            float64
start_station_latitude      float64
start_station_longitude     float64
tripduration                  int64
tmp                          object
CensusTract                   int64
County_x                     object
Borough                      object
TotalPop                      int64
Men                           int64
Women                         int64
Hispanic                    float64
White                       float64
Black                       float64
Native                      float64
Asian                       float64
Citizen                       int64
Income                      float64
IncomeErr                   float64
IncomePerCap                float64
IncomePerCapErr             float64
Poverty                     float64
ChildPoverty                float64
Professional                float64
Service                     float64
Office                      float64
Construction                float64
Production                  float64
Drive                       float64
Carpool                     float64
Transit                     float64
Walk                        float64
OtherTransp                 float64
WorkAtHome                  float64
MeanCommute                 float64
Employed                      int64
PrivateWork                 float64
PublicWork                  float64
SelfEmployed                float64
FamilyWork                  float64
Unemployment                float64
Latitude                    float64
Longitude                   float64
BlockCode                     int64
County_y                     object
State                        object
BlockCode_11                  int64
borough_distance            float64
distance_rank               float64
dtype: object
```

SHARE STATIONS- MAP OF NYC

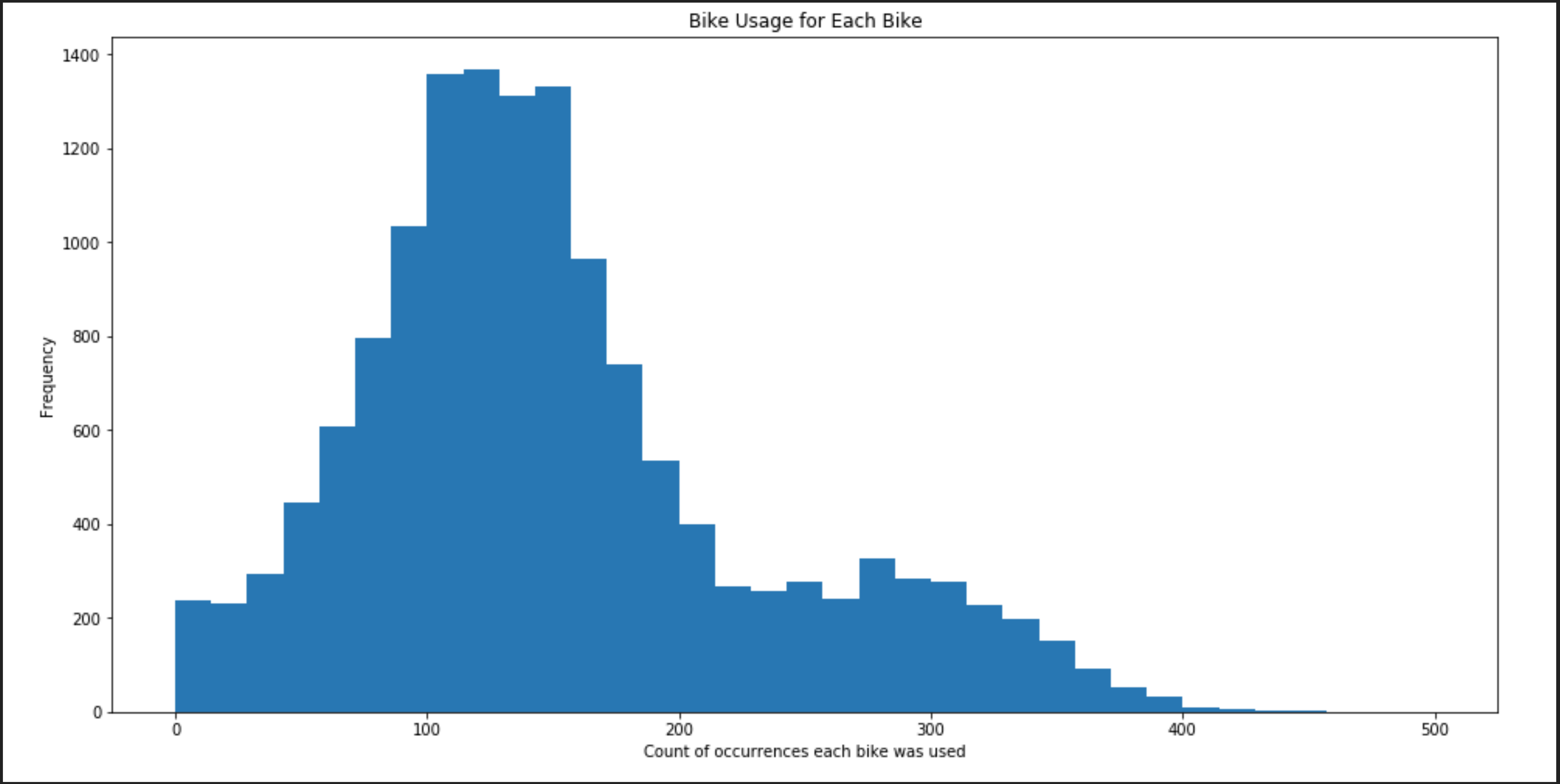
There are no bike share stations in the Bronx or Staten Island



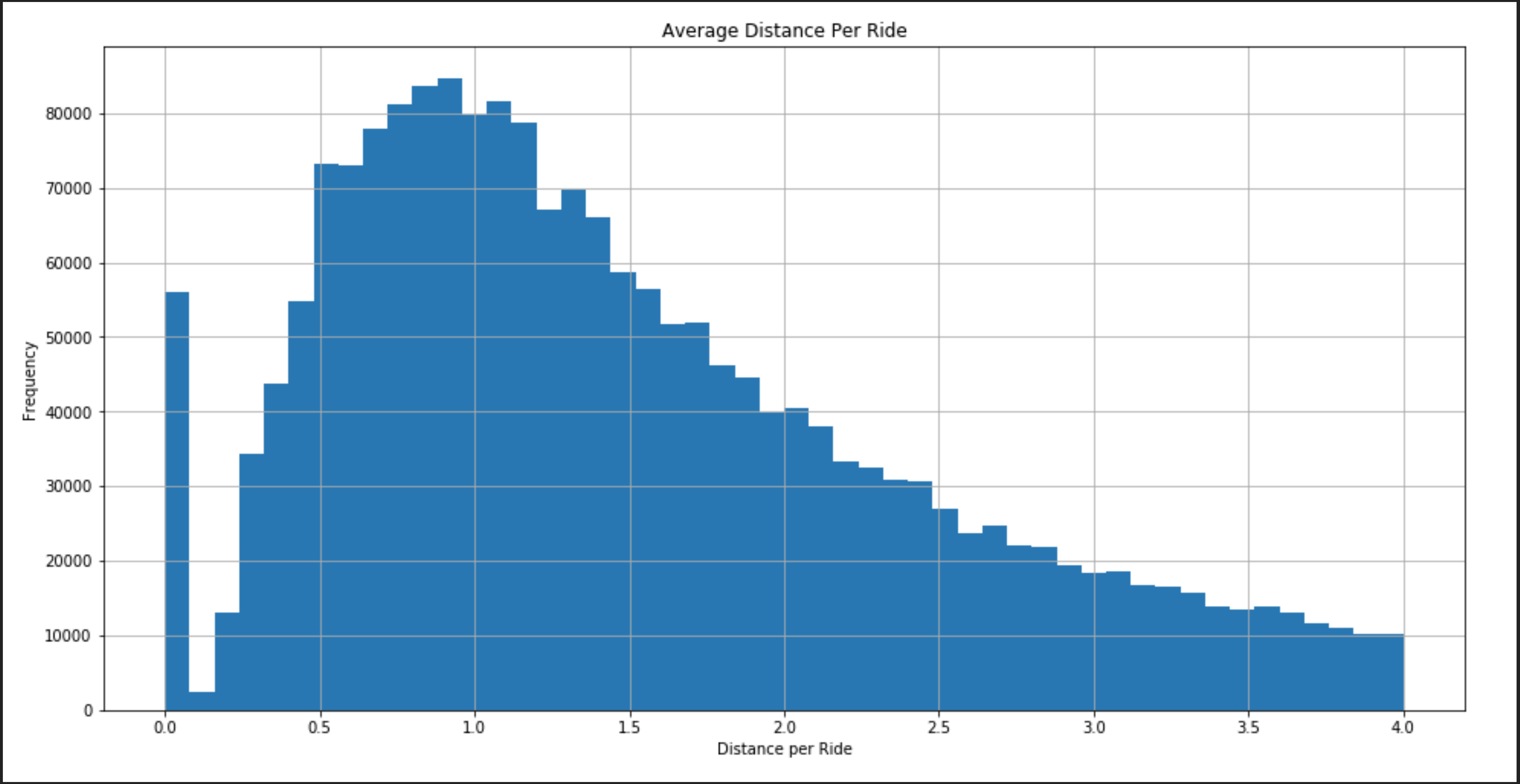




There are 14356 bikes and each has been used on average 152 times.

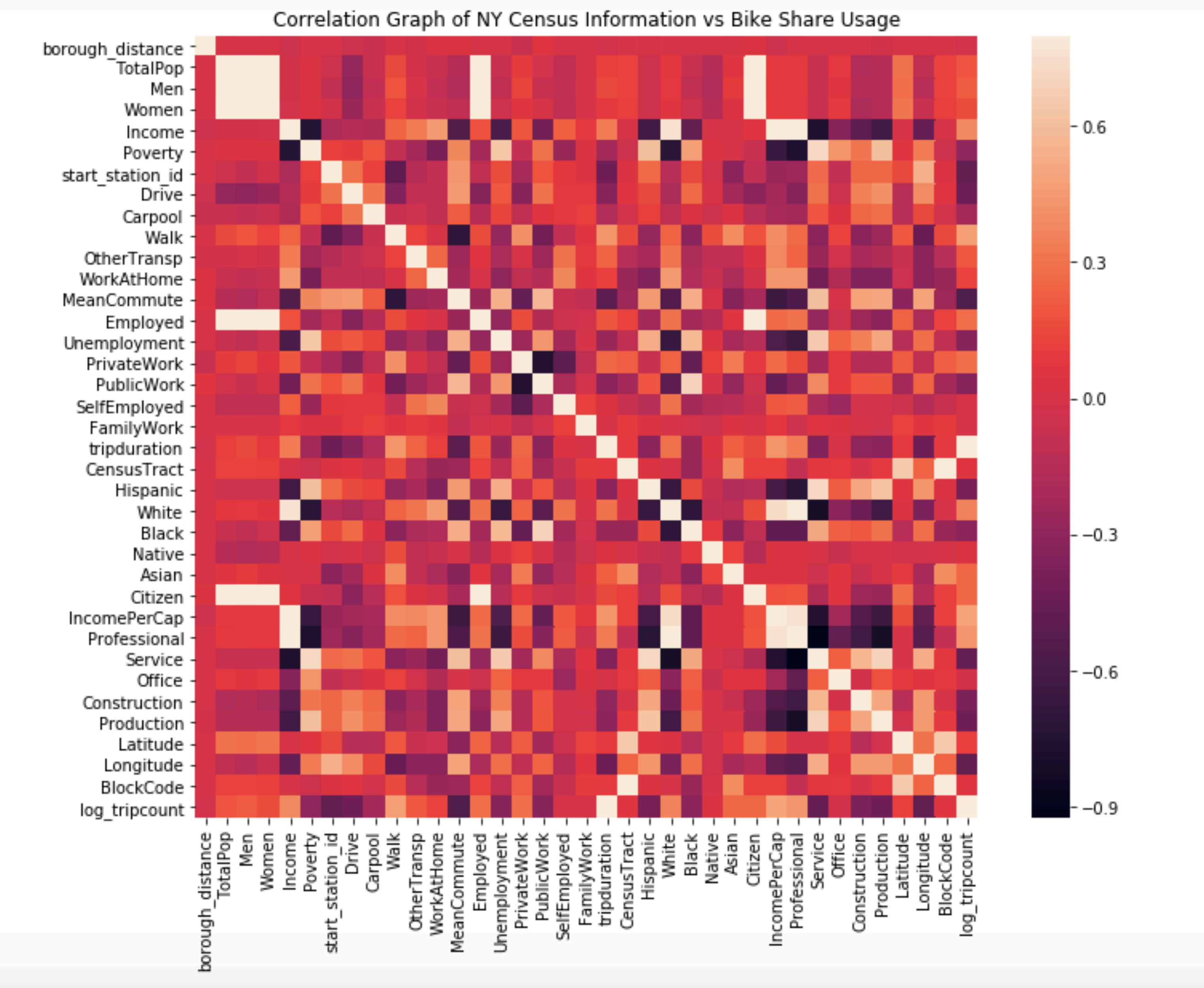


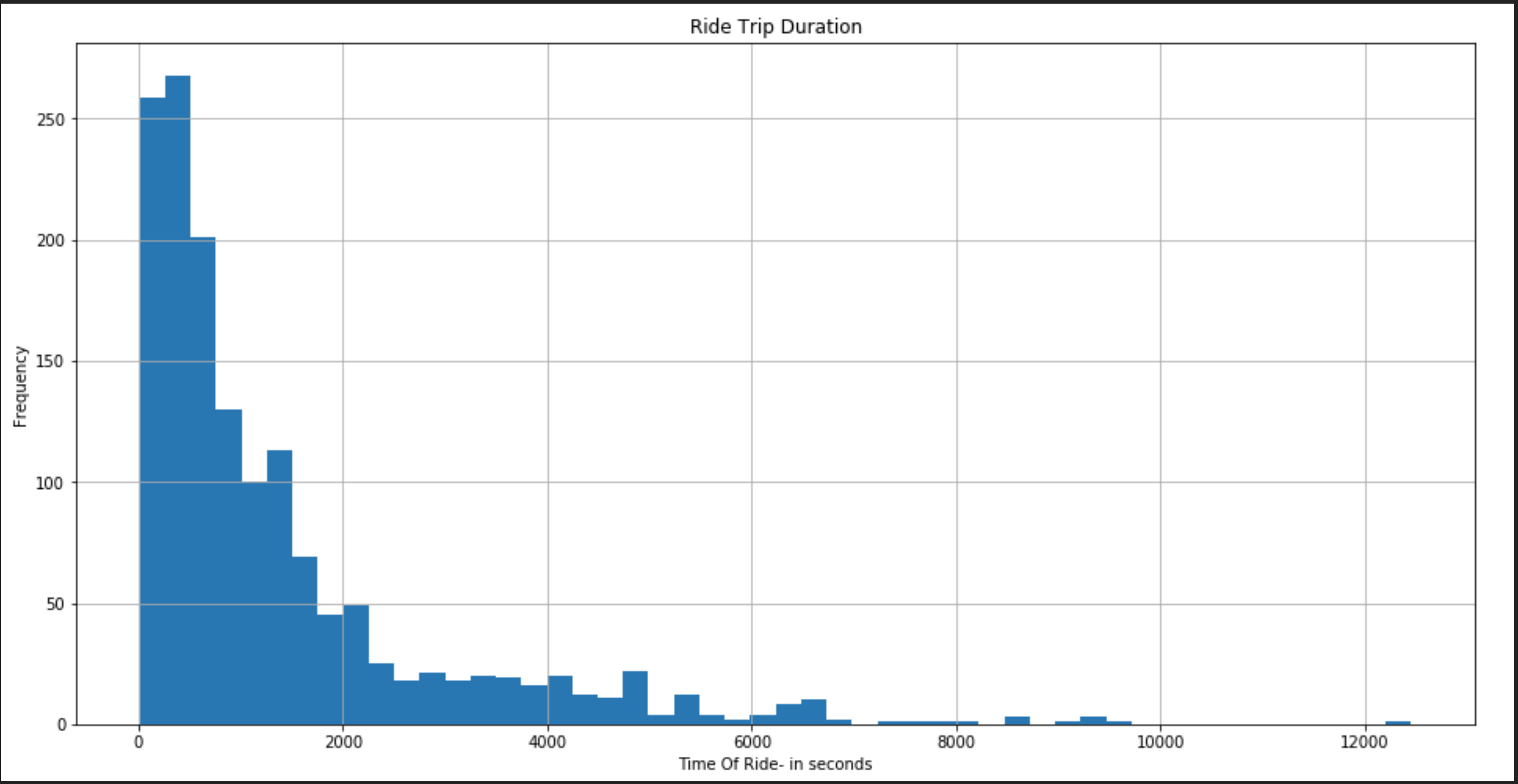
The average ride is slightly under 2km, likely riding from one station to another station



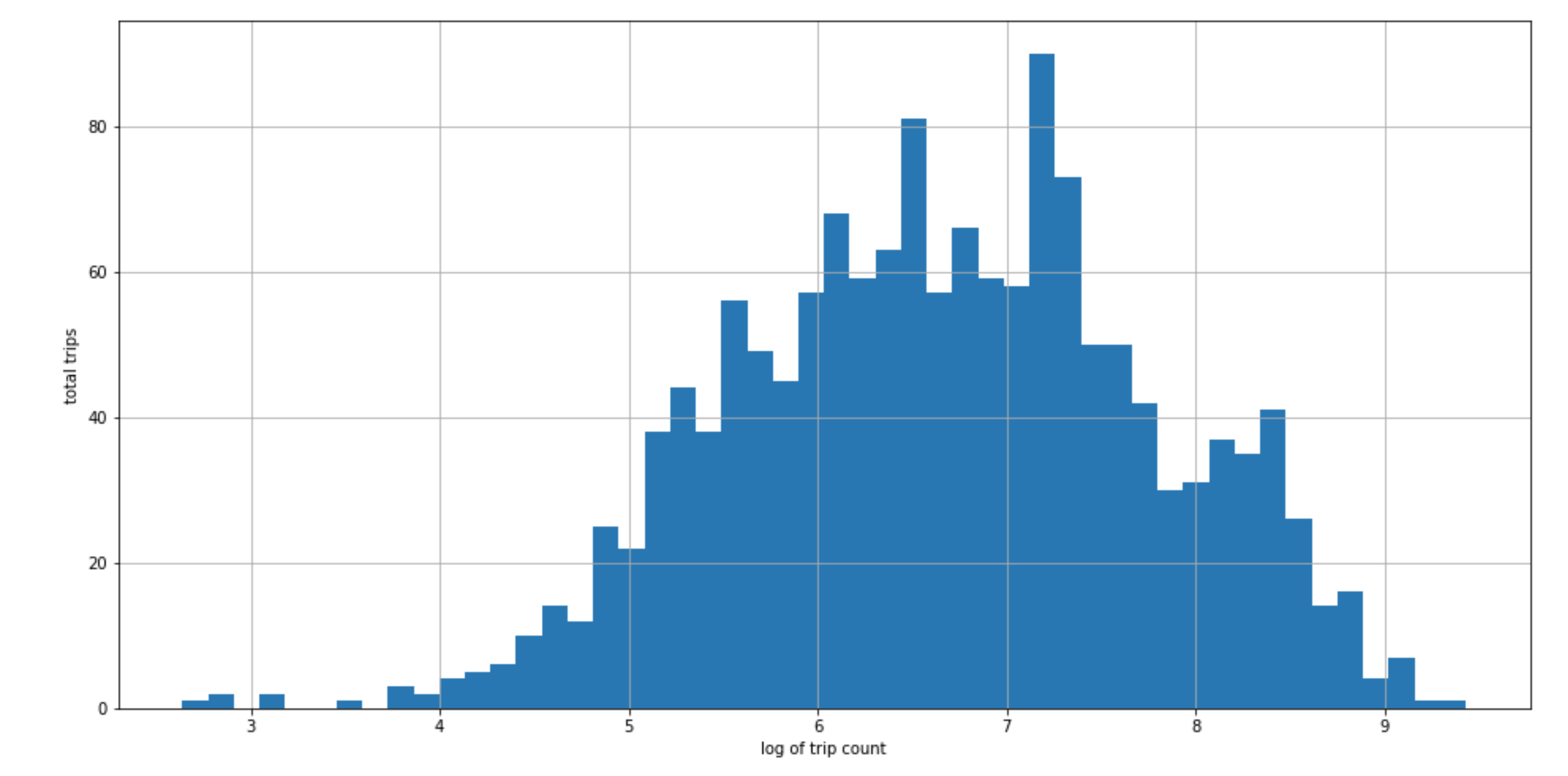
HEAT CORRELATION MAP

Showing high correlation with Income, Walking, Employed, Professional, Total Population, and Men

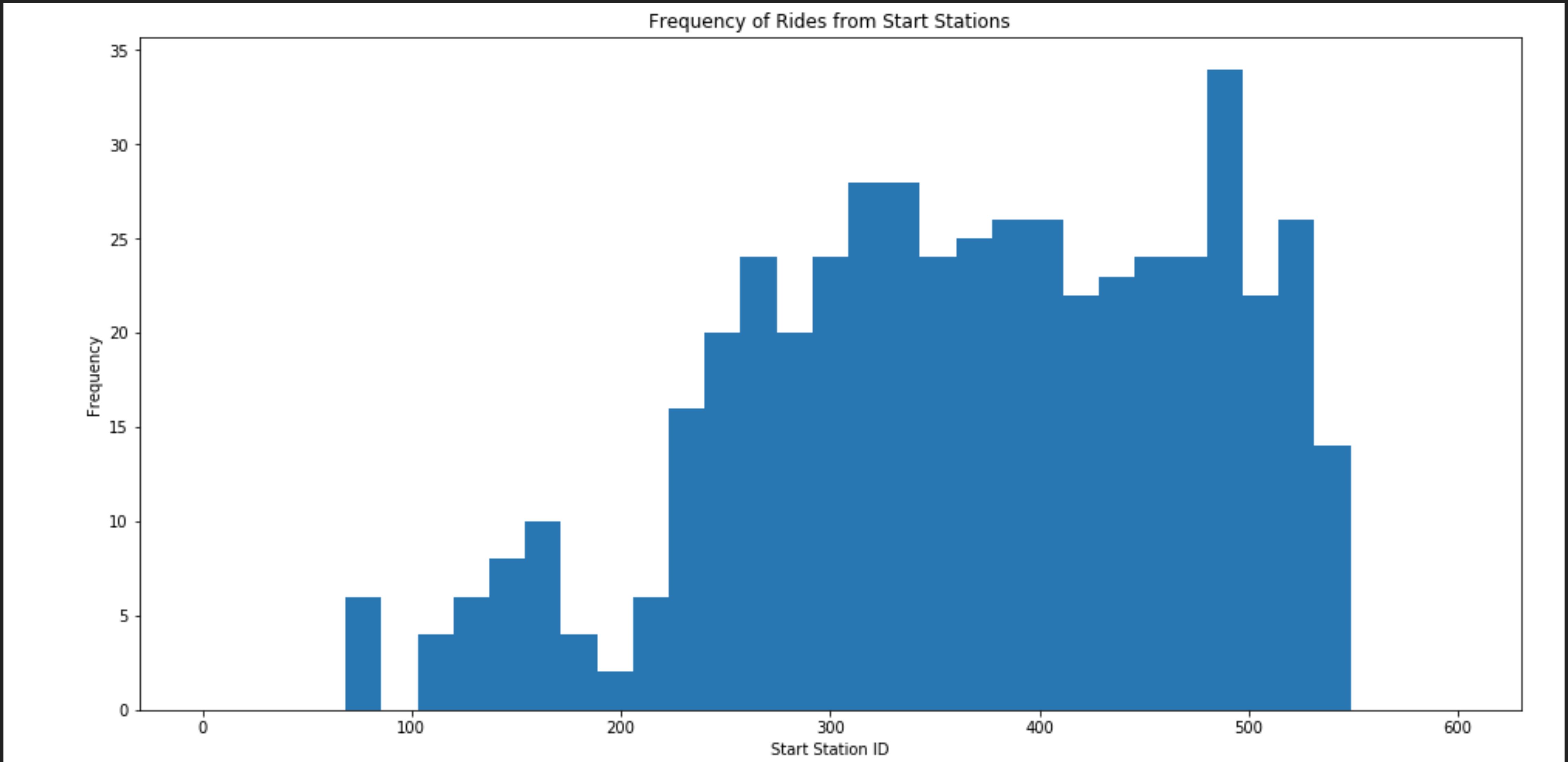




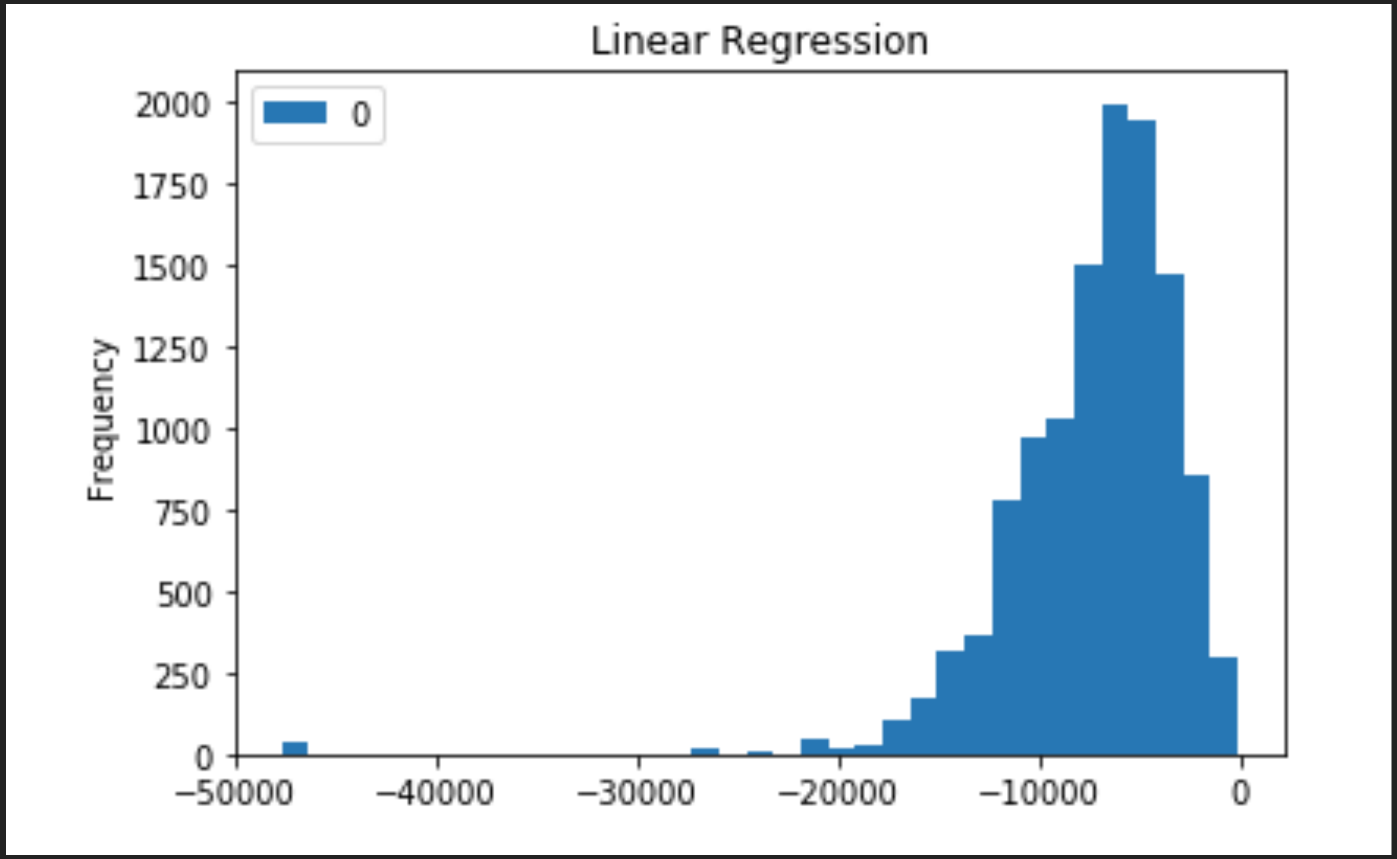
Uneven data shown here so I chose to use the log function to normalize the data for use within my models.

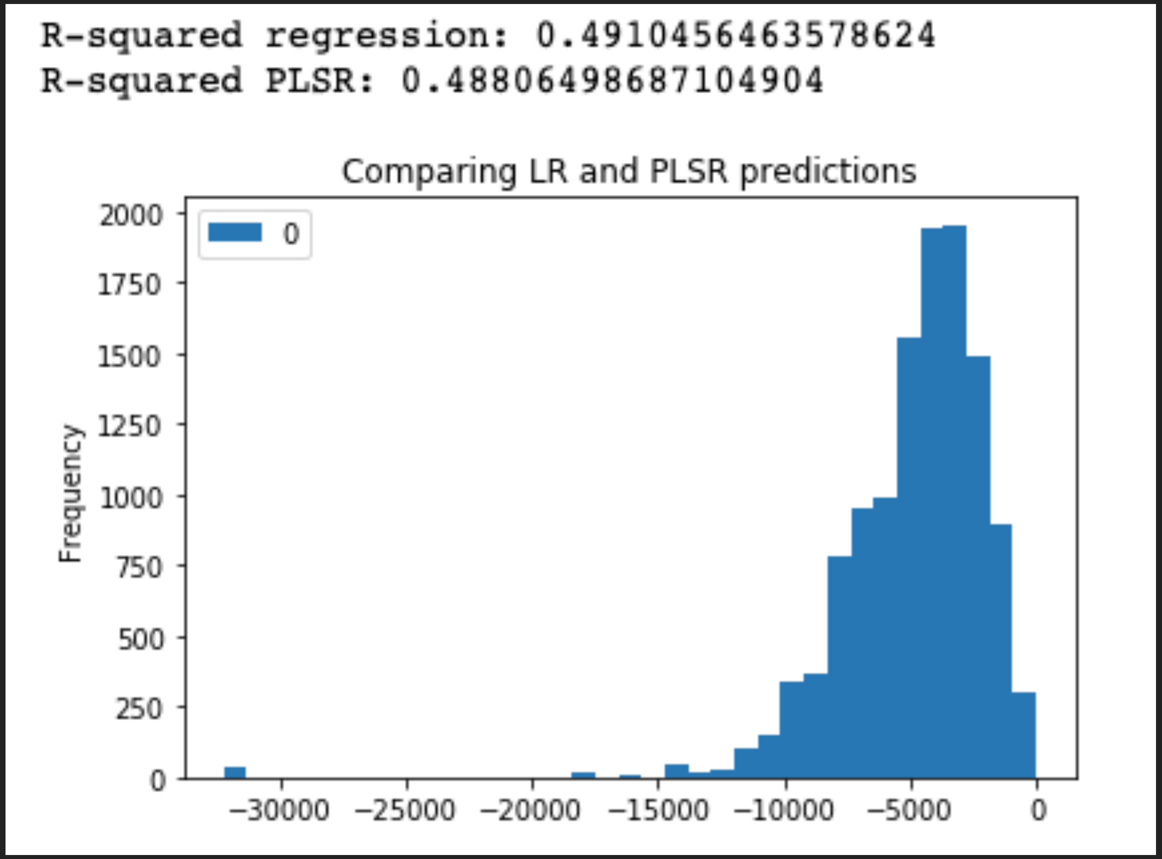


The number of rides that occur from each of the bike stations, showing the most use out of the Station ID's 300 to 500, which are mostly all in Manhattan.



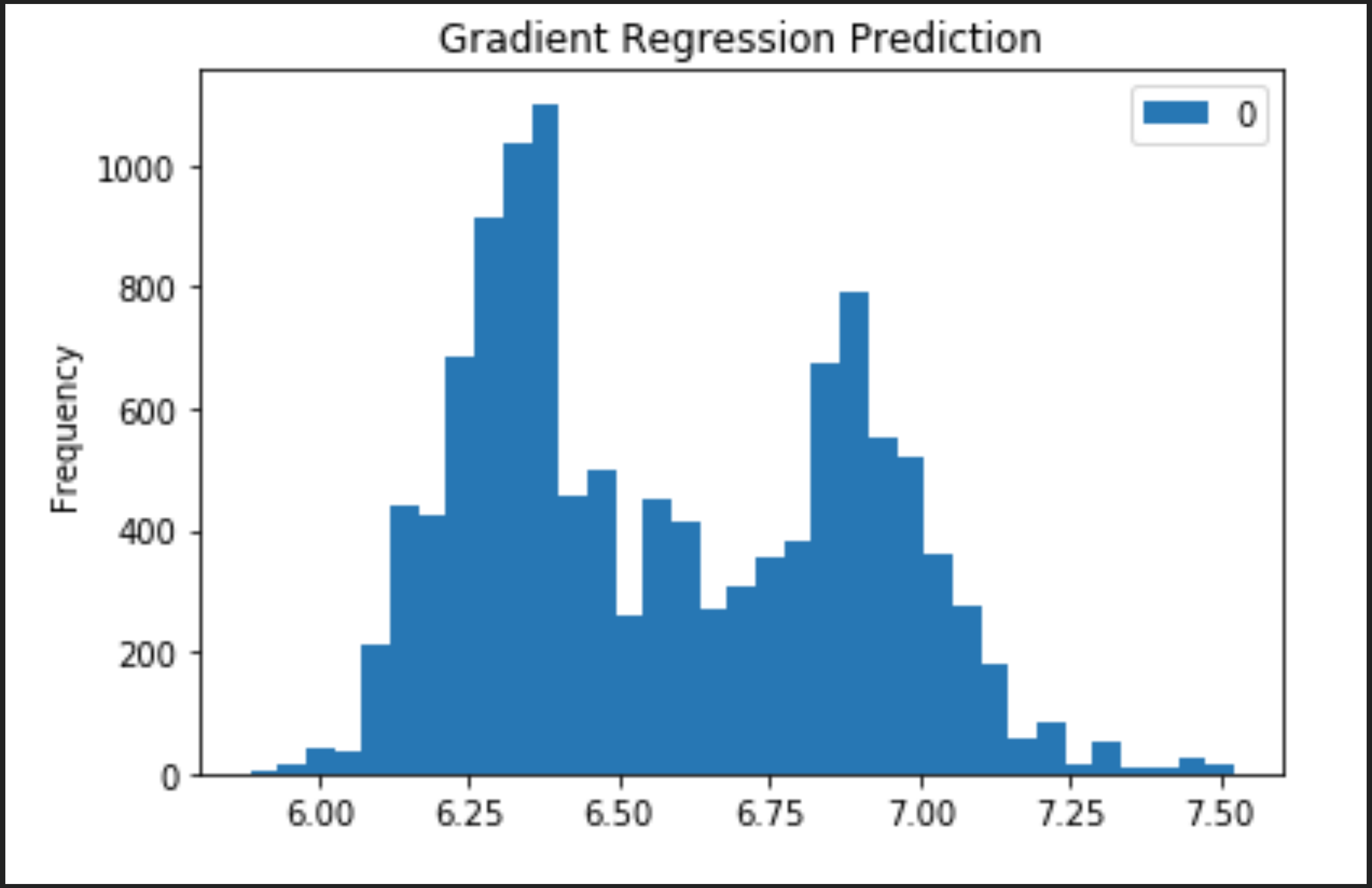
Used both Linear Regression and R-squared Linear Regression/R-square PLSR models to try and predict the most beneficial location for additional stations, however for both charts it came up negative. This could be because of using the log for the trip durations, however it did showed a 49% accuracy.



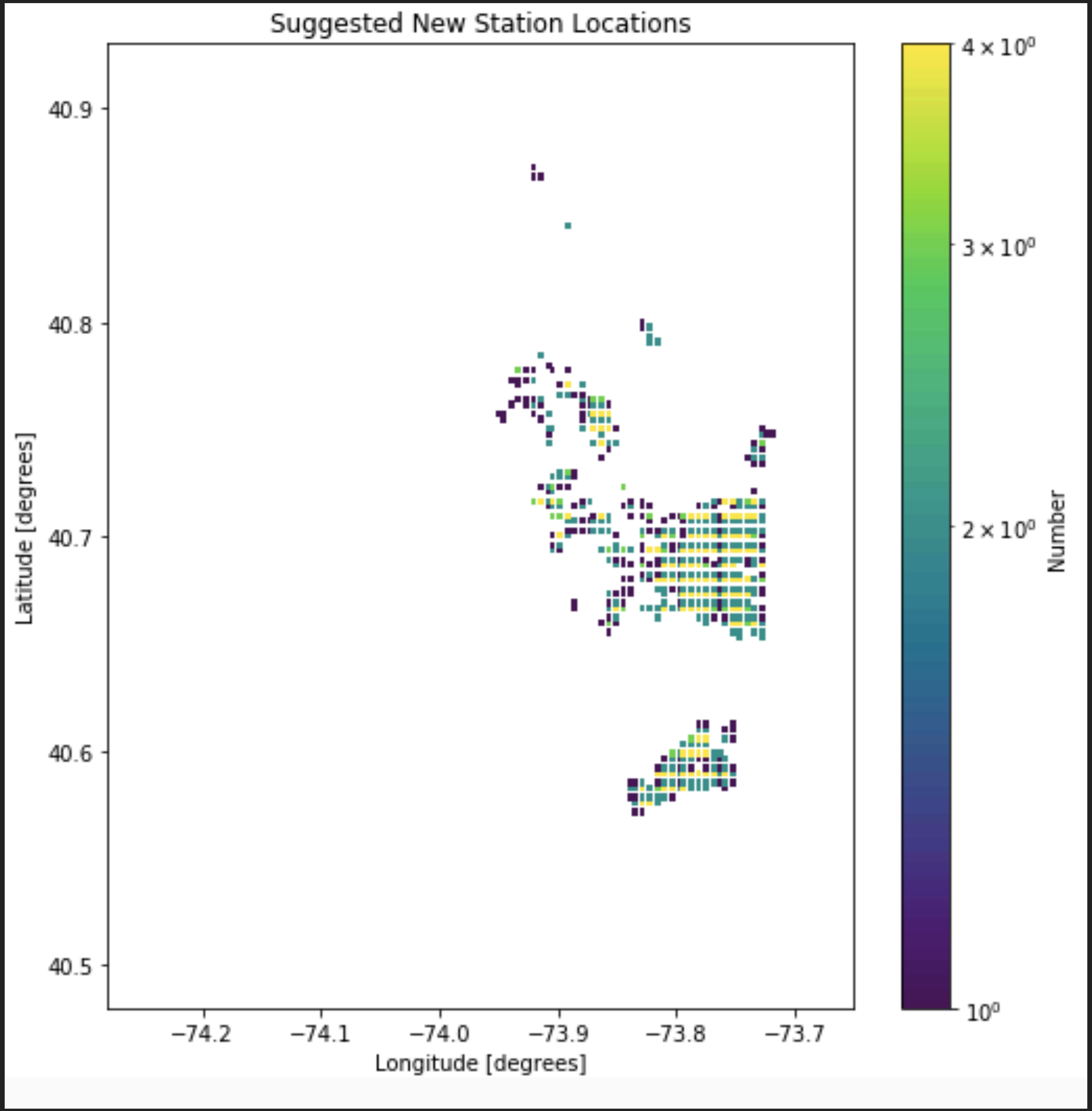


```
R-squared regression: 0.4910456463578624
R-squared PLSR: 0.48806498687104904
```

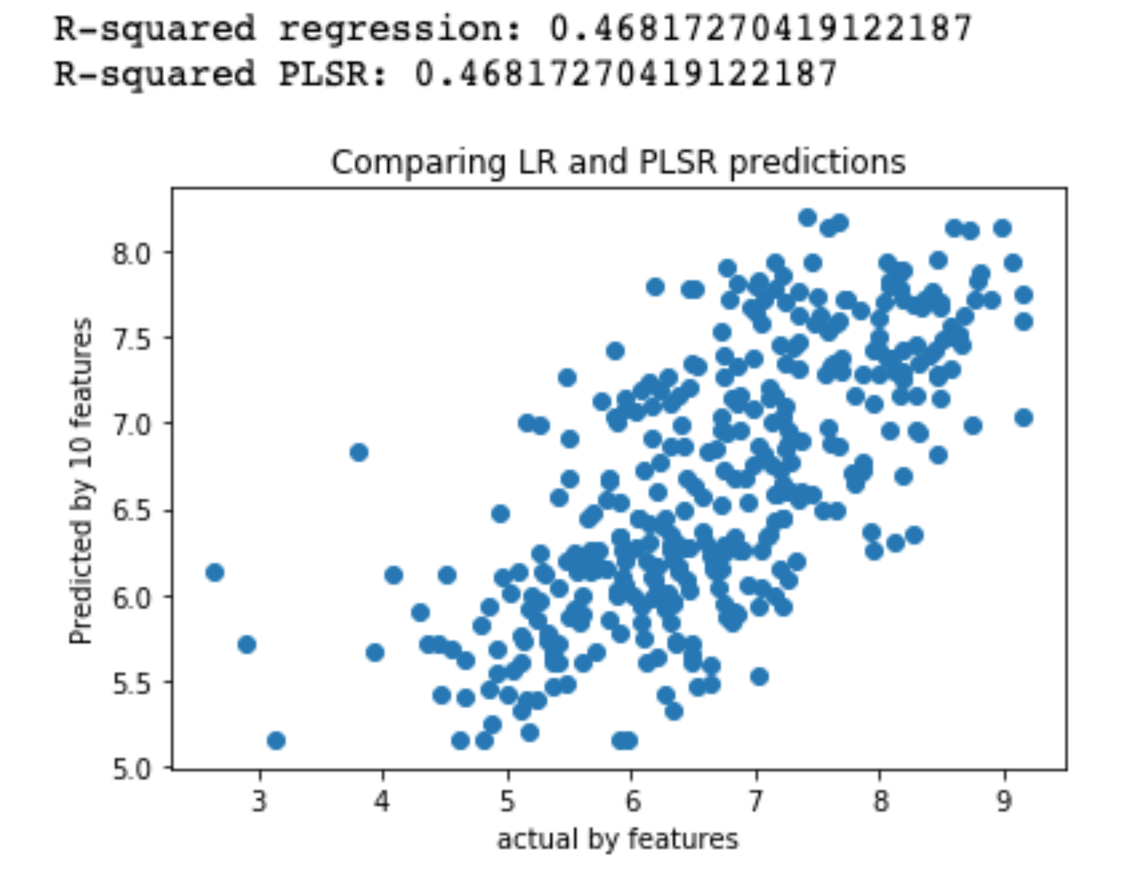
Gradient Regression proved to be the most effective prediction model for this data. With this histogram it shows the ride frequency, remembering that it is raised to the 6th from the log function. The right side of the tail is showing that there is room for growth in those boroughs, looking at about 7.25 for growth options.



Looking at the NY Census data (including income, gender, transportation preference, work type, etc) using the location based off of latitude and longitude with a gradient regression 7.25 or higher. This map shows that all the stations will fall into the Queens Borough. The yellow dots are the locations to start with out of the 128 new suggested locations within the borough that would be the most profitable stations to start with and then add onto.

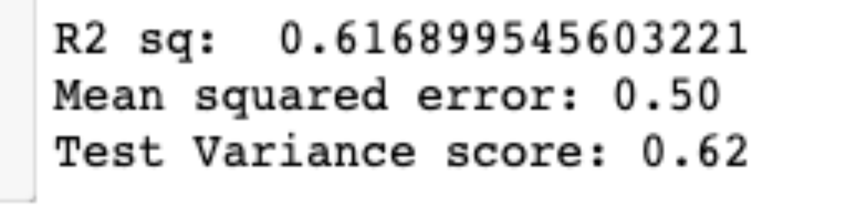


LINEAR REGRESSION & PARTIAL LEAST SQUARE REGRESSION

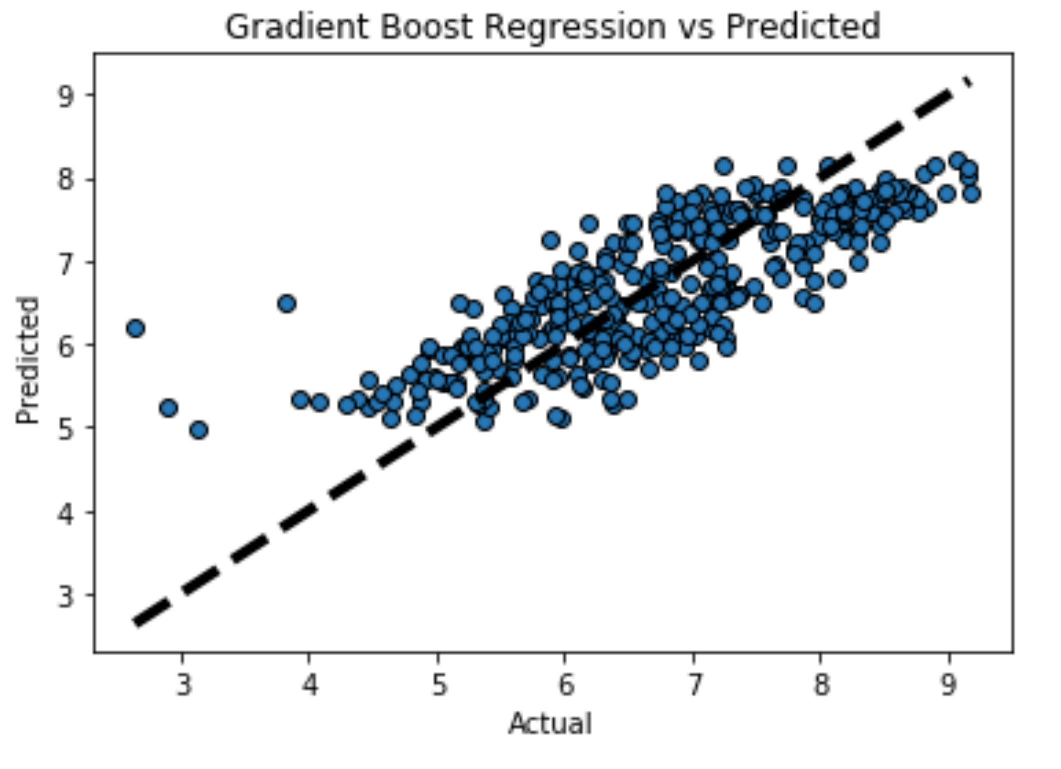


```
R-squared regression: 0.46817270419122187
R-squared PLSR: 0.46817270419122187
```

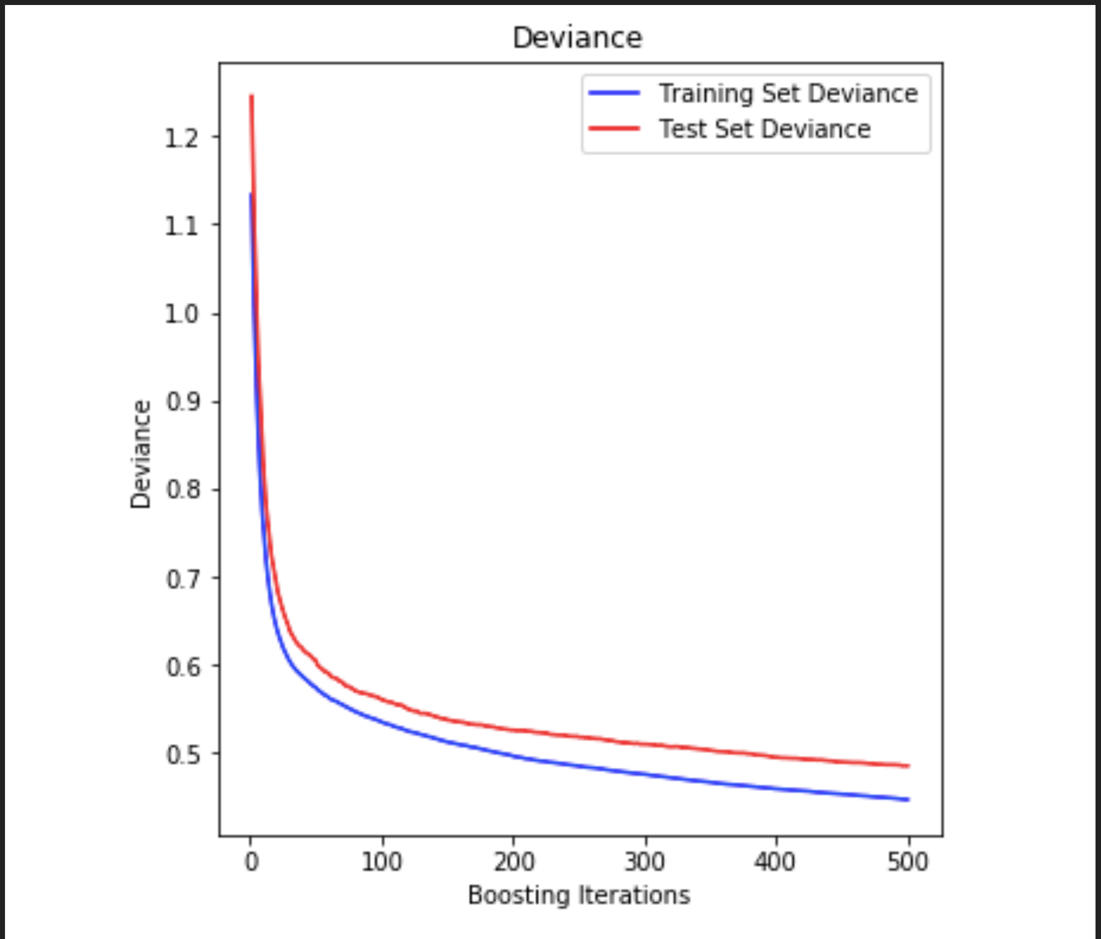
GRADIENT BOOST REGRESSION



```
R2 sq:  0.616899545603221
Mean squared error: 0.50
Test Variance score: 0.62
```



Gradient Boost Regression vs Predicted



Deviance

RANDOM FOREST