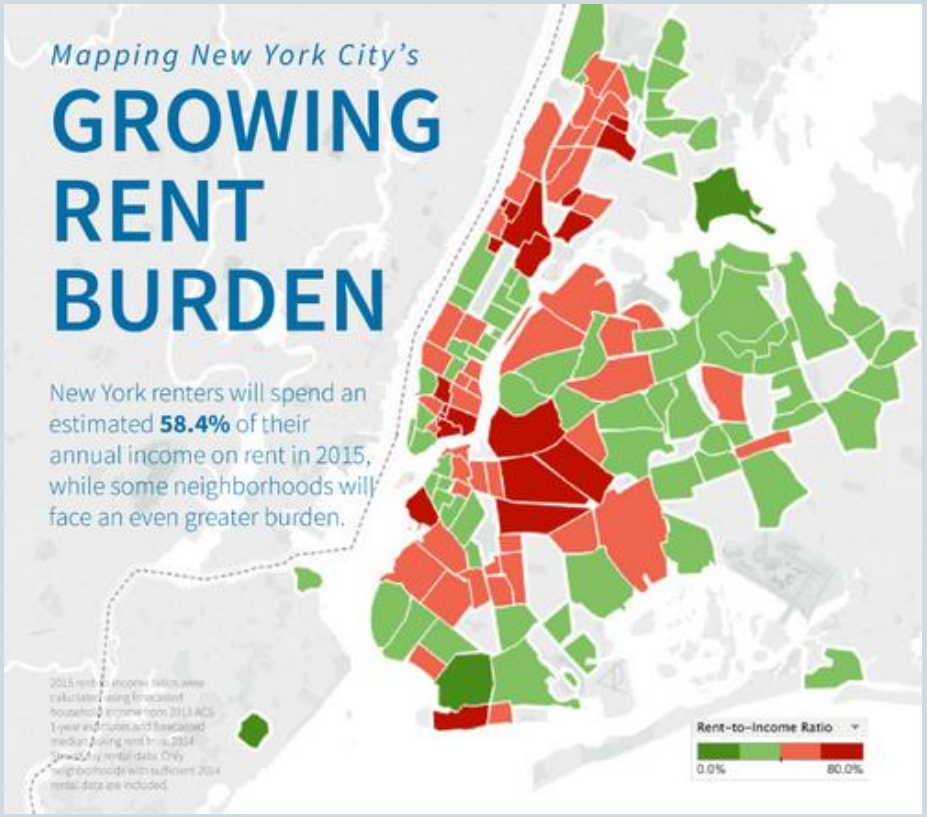
New York City Rental Price Prediction

NICOLAS JORQUERA

INTRODUCTION



Street Easy Prediction

Moving to a new city can be a daunting task, especially when you have no connections and know little of the city. Although rental agencies can assist in helping you find a location, they can be expensive.

naked apartments

Average Rent in NYC

Here's a snapshot of average rents in Brooklyn, the Bronx, Manhattan and Queens, broken down by apartment size.

3 bedrooms
2 bedrooms
1 bedroom
Studio

GOAL

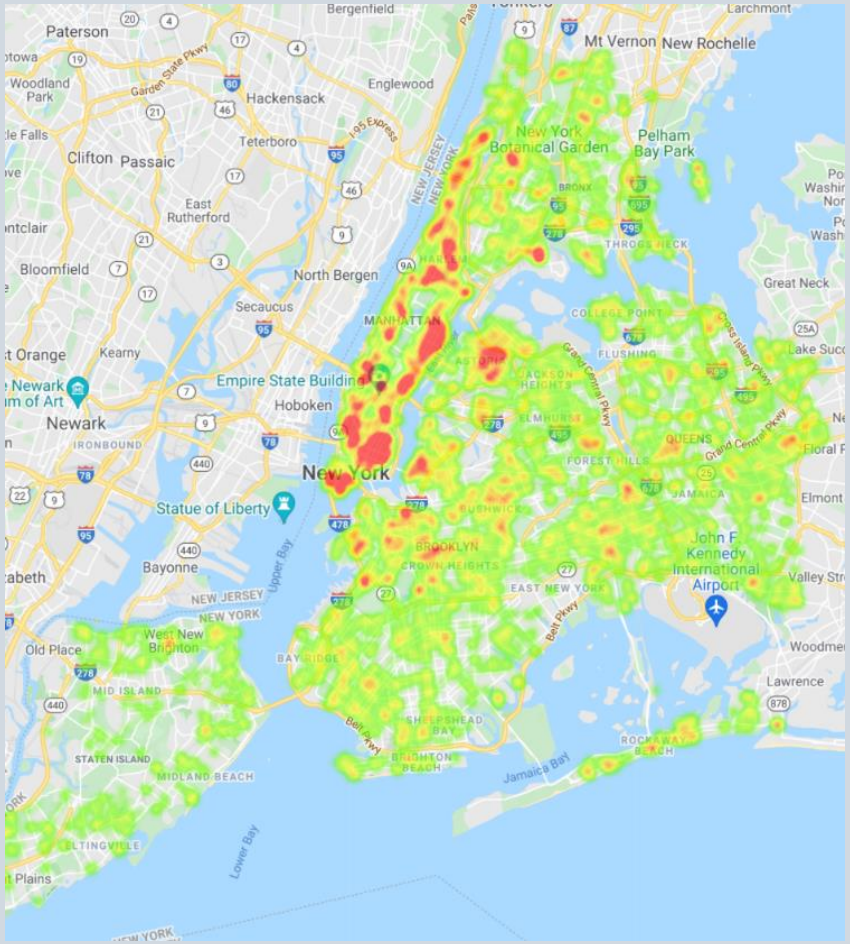
This project aims at helping individuals find a new home in New York City based on their own criteria! We will examine the extrinsic factors of each Neighborhood, analyze which apartments are over market values; and narrow down this choice to a few select listings so that apartment hunting becomes a less daunting task!

EXPLORING THE DATASETS

In this project we utilize two datasets:

1. **New York City Rent Prices for 2020 Dataset**

2. **Neighborhood Extrinsic Factors - Housing Authority Dataset**

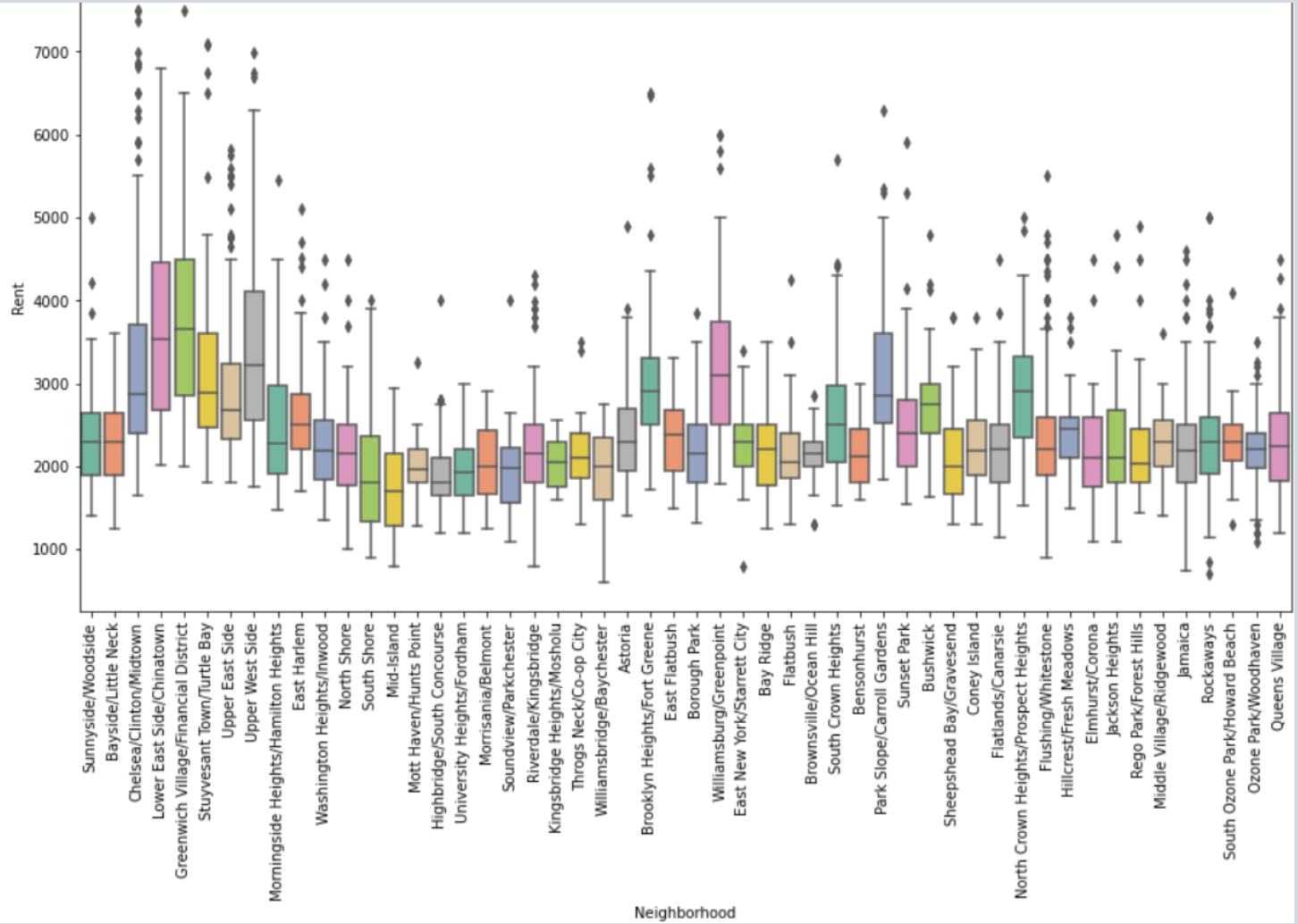


Heat Map of Listings [Google API]

Zillow – New York City Rent Prices for 2020 Dataset

Data Wrangling

- 7000 unique listings for apartments in New York City in 2020, with over 20 features
- However, to ensure that all this data is accurate, we had to filter the data, by removing listings not in NYC, or listings whose rent cost were astronomical. After filtering the data, we removed 20% of the listings.
- Utilized Google Maps API to view location of all filtered listings.



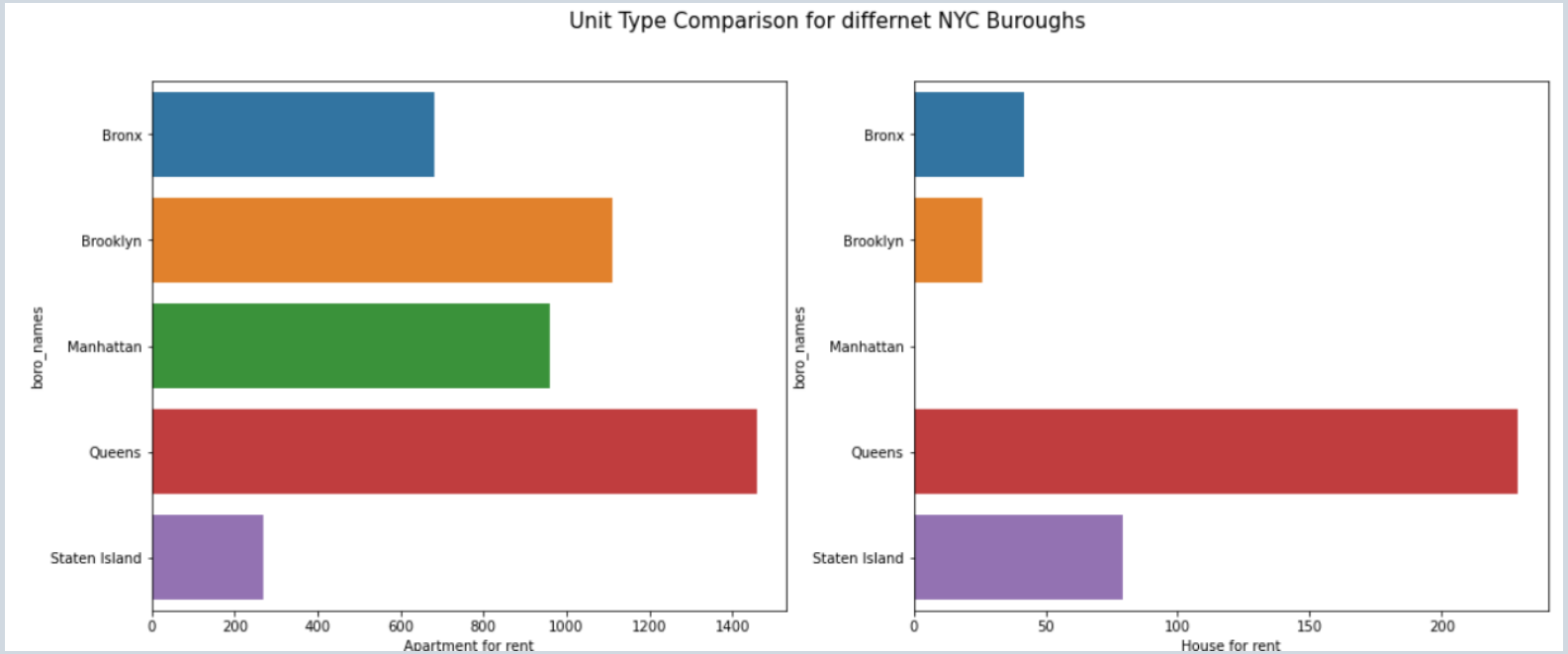
Average Rent Price Fluctuation by Neighborhood

EXPLORATORY DATA ANALYSIS

Comparing different Neighborhoods / Burroughs:

- Neighborhoods in Manhattan have a higher Average Rent on average, while neighborhoods in Bronx and Staten Island are significantly less.

EXPLORATORY DATA ANALYSIS



Unit Type Comparison for Different NYC Burroughs

Comparing different Neighborhoods / Burroughs:

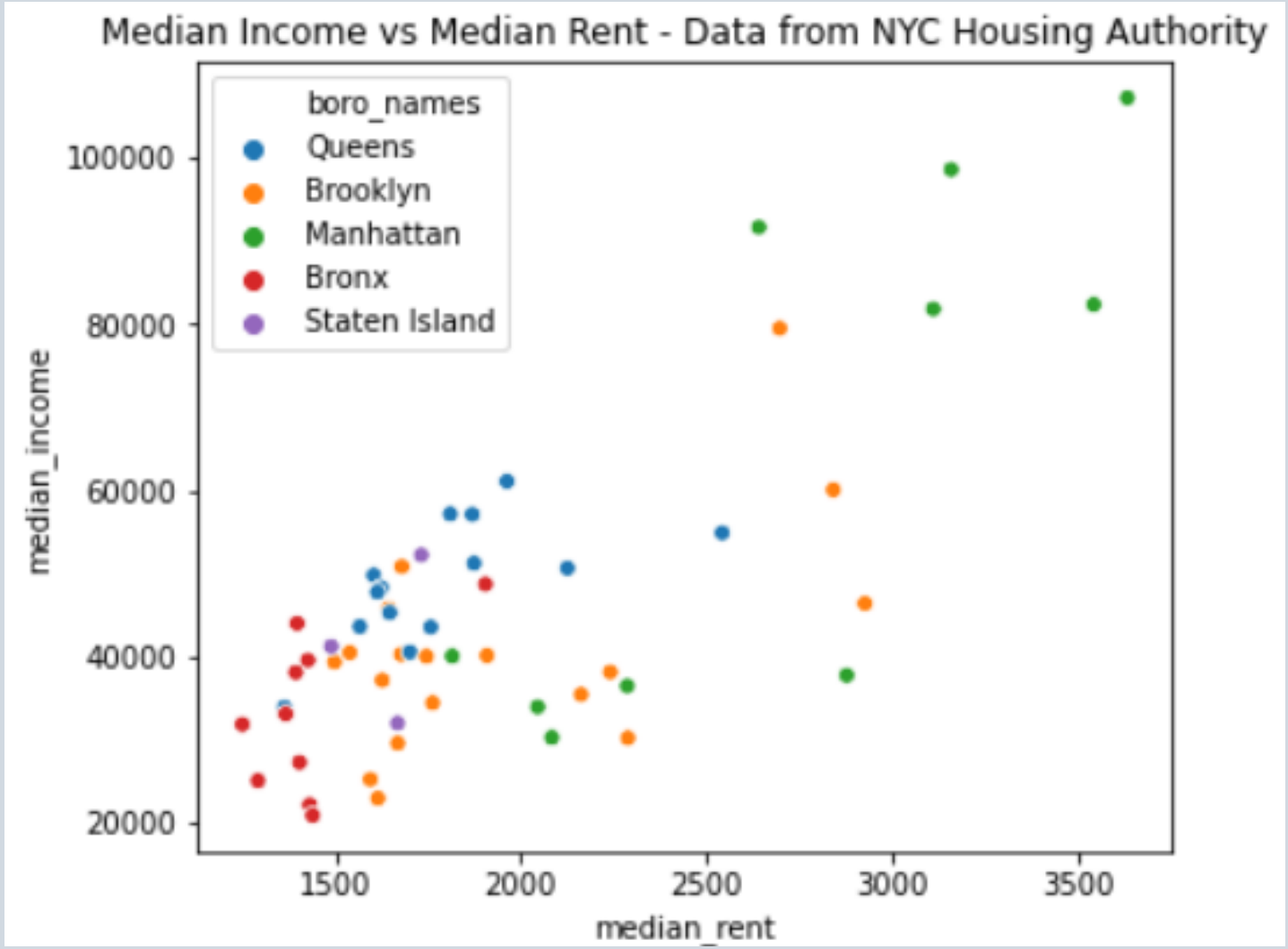
- Queens has the most units for rent. However, Manhattan has no houses for rent; which makes sense as it's not as suburban as the other boroughs. On the other hand, most units for rent in Staten Island are houses.

Neighborhood Extrinsic Factors – NYC Housing Authority

- This dataset includes the 55 neighborhoods in NYC, and 33 features. These features are useful for comparing extrinsic factors, including Number of Housing Units, Average Price, Locations of Public Transportation

Observations

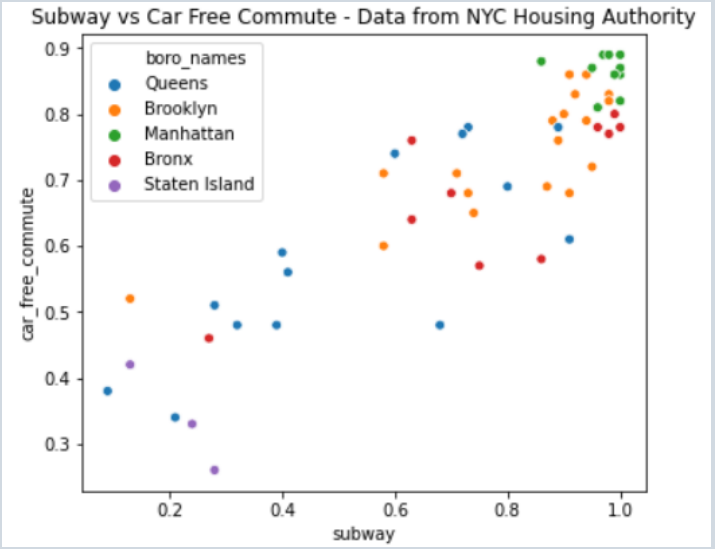
- Manhattan has the highest median rent and median income; which correlates as these Neighborhoods are significantly more expensive than other boroughs. Then comes Brooklyn, with Bronx significantly less expensive than these boroughs.

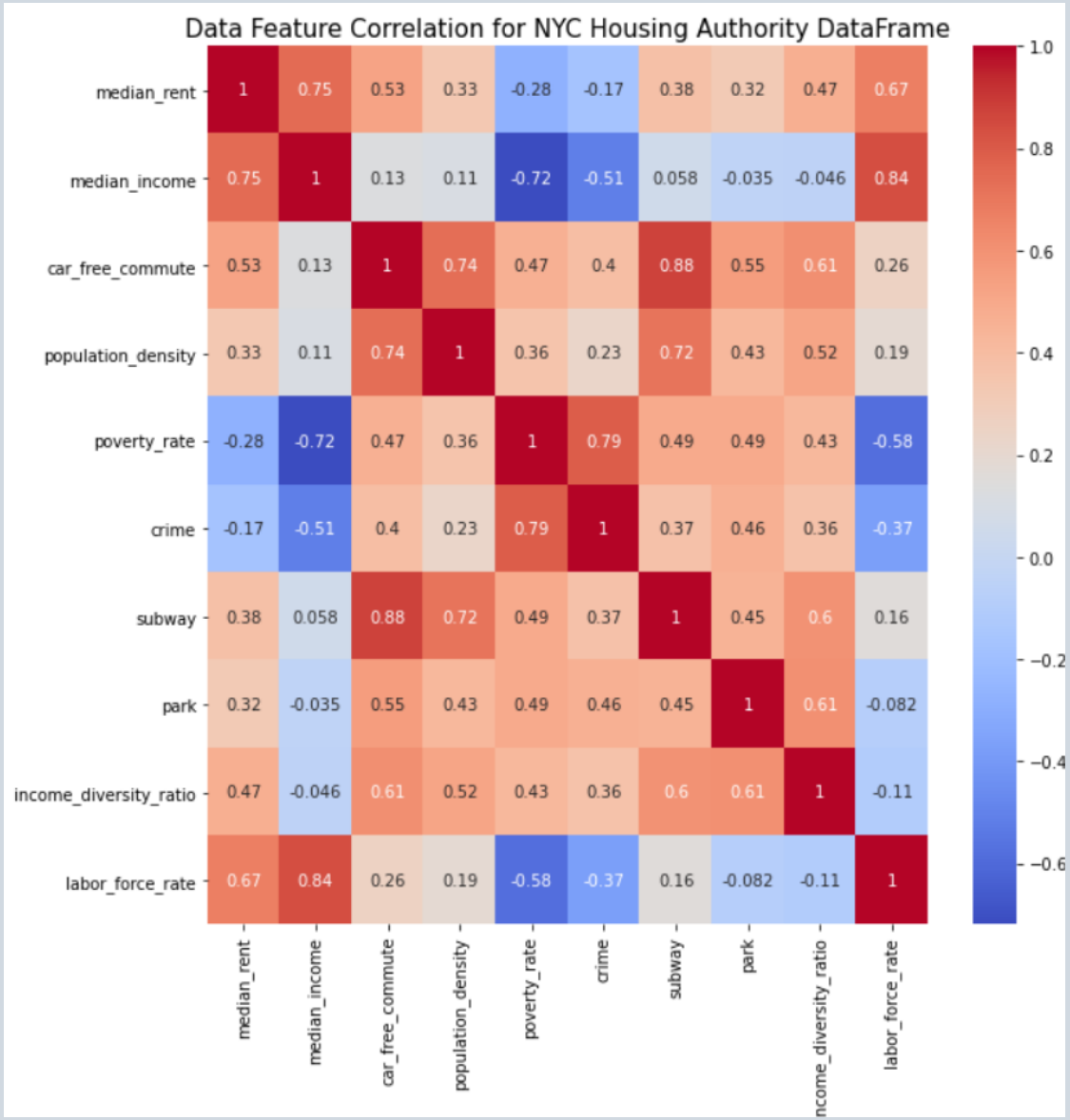


EXPLORATORY DATA ANALYSIS

Observations – *cont'd*

- Manhattan has the highest subway ratio, which signifies the percentage of residential units that are within a ½ mile walk of a station entrance for the NYC Subway. On the other hand, Staten Island has the lowest.
- Bronx has the highest Poverty Rates and Crime Rates;. On the other hand Queens is closest to the origin in this graph, showing lowest crime rates. Manhattan and the other boroughs seem to be scattered.

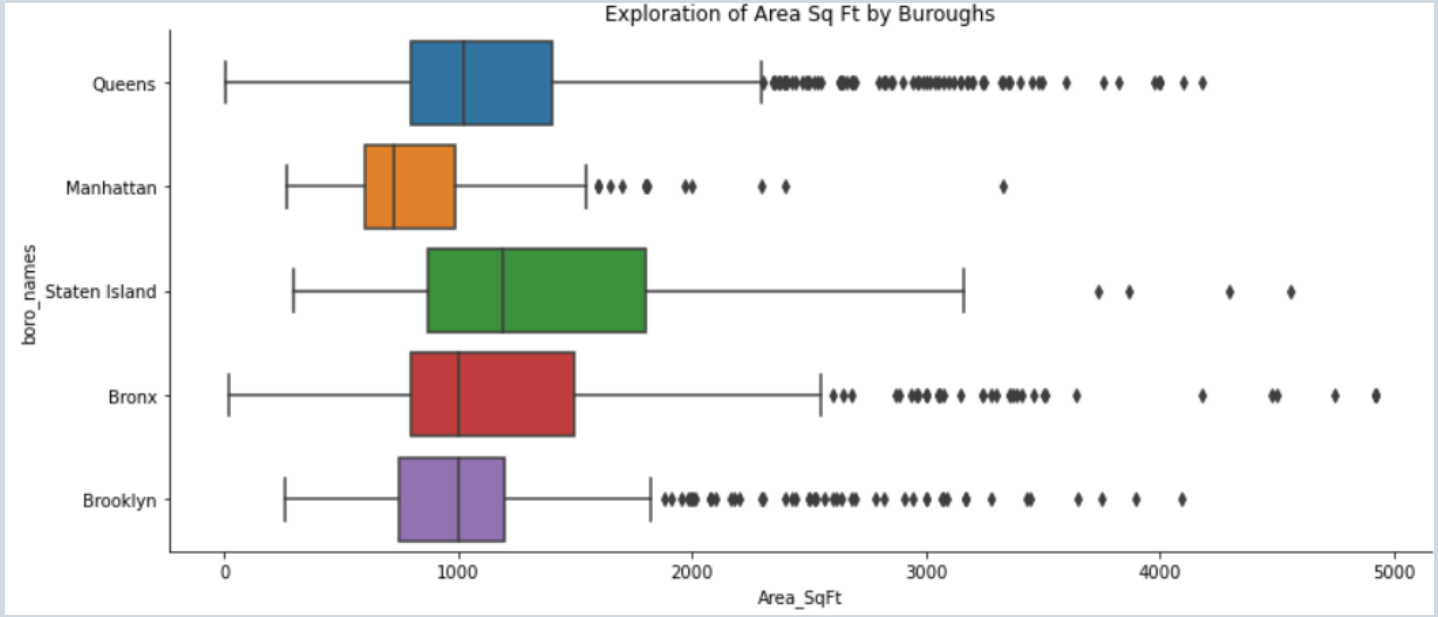




Correlation Matrix

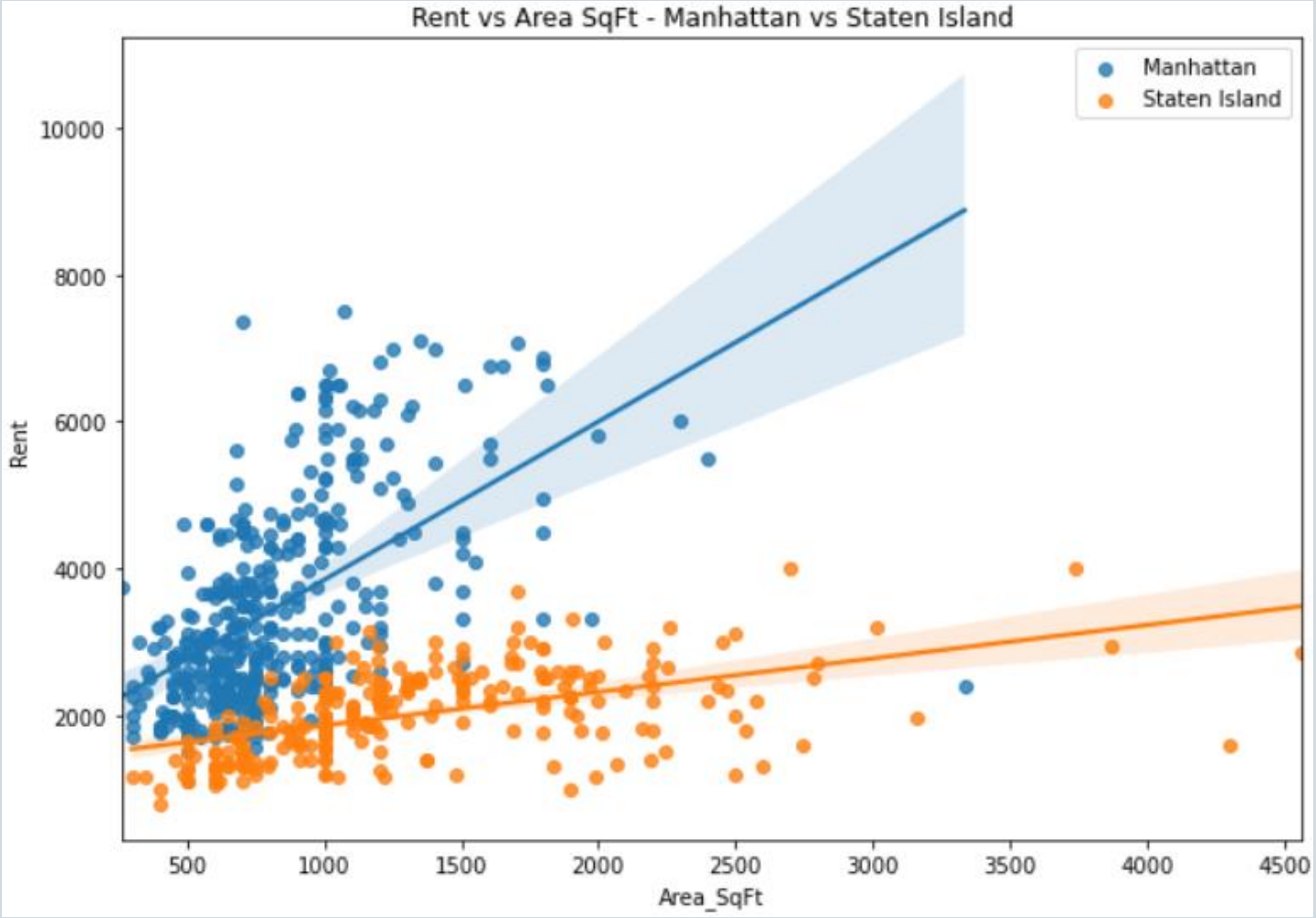
DEALING WITH MISSING DATA

More than half the listings have area missing. Instead of removing it, we examine how linear this relationship is. By isolating the listings which have area specified, we were able to visualize how this feature fluctuation between different boroughs.



EDA – RENT AND AREA

After plotting a scatterplot between these two we found no clear linear relationship, therefore we imposed the borough name and re-examined this relationship between Manhattan and Staten Island. Here we clearly see a linear relationship; that varies depending on Borough.

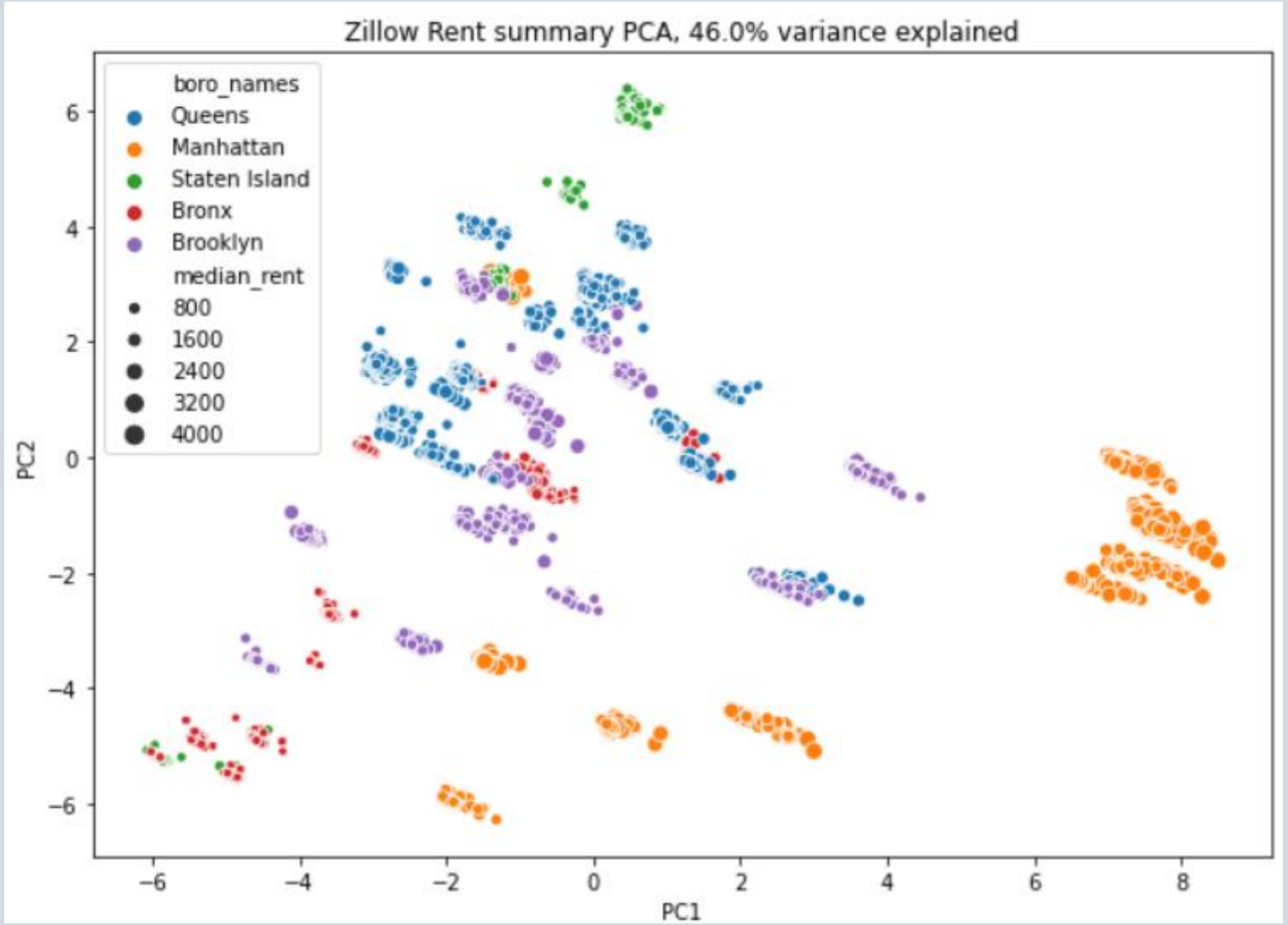


PRINCIPAL COMPONENT ANALYSIS

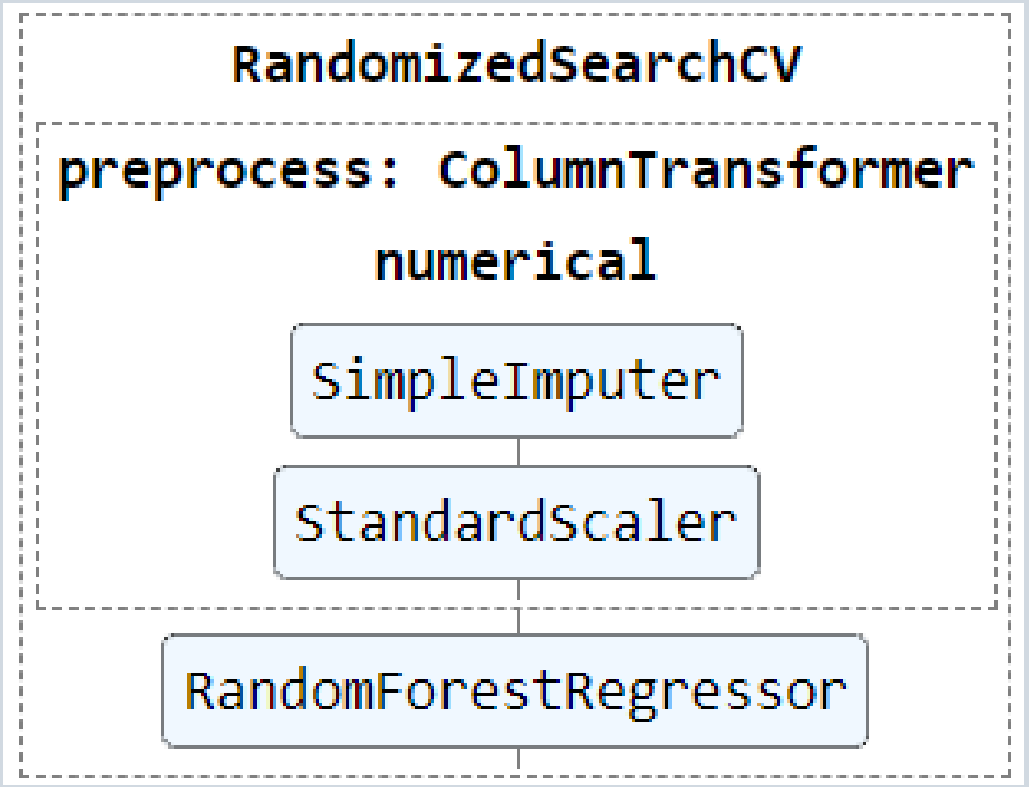
Our Dataset has more than 40 features, we will apply PCA to visualize this high dimensional data. By imposing the different Boroughs and Rent Prices on our PCA visualization, we can see note how they tend to cluster.

Observations

1. 2 components only explains 46% of the variance.
2. Manhattan listings clearly cluster by borough. However other boroughs intersect one another



MODELING


RandomizedSearchCV
preprocess: ColumnTransformer
numerical
SimpleImputer
StandardScaler
RandomForestRegressor

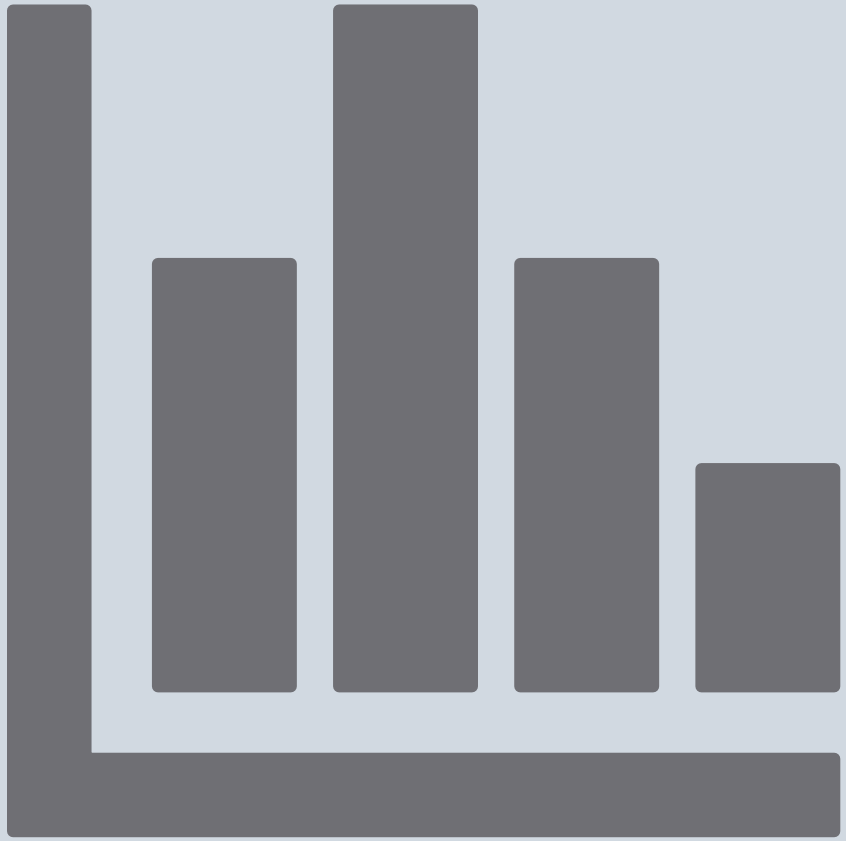
PREPROCESSING

MODEL SELECTION

MODEL PERFORMANCE

PREPROCESSIING

Used an Imputer and a Standard Scaler to perform preprocess the numerical features, in order build our Machine Learning Pipeline. For our categorical data we used a OneHotEncoder.



MODEL SELECTION

HYPERPARAMER TUNING

For Linear Regression and Ridge Regression model, we used Grid Search CV to determine which hyperparameters and would lead to more accurate model.

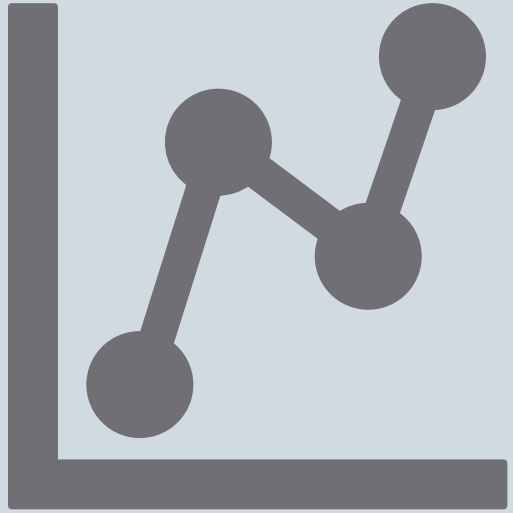
For the Random Forest Regressor we used Randomized Search CV to decrease computational complexity and asses the multiple hyperparameters associated with Random Forest

Machine Learning Models	Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)	r2 score
Linear Regression	266.62	413.13	0.78
Ridge Regression	266.32	412.32	0.79
Random Forest Regressor	119.37	261.33	0.91

MODEL PERFORMANCE

MODEL SELCTION

After creating these three models we deduced that a Random Forest Regressor outperformed the other models by 55% in MAE, improved the variance explained by 17%, and RMSE improved by 37%. Therefore, in order to predict rent prices for our listings we will apply the Random Forest Regressor model.



APPLICATION

Now we begin to dissect our data to determine which apartments are not priced over market and have the specifications our client desires. These specifications can change and be accommodated for different individuals. The first task is to remove Apartments who are overpriced, while keeping in mind that the MAE was 119.37.

Afterwards we can impose certain parameters based on client specifications, such as bedroom number, bath number, and ideal rent fluctuations. In a more sophisticated model, we can impose neighborhood constraints too

By determining which of these extrinsic neighborhood parameters [neighborhood features] are more important to individuals; we can reduce the listing selections to a few. Furthermore, we can even highlight which listings are below the predicted rent price; and separate these listings as ideal listings

FURTHER INVESTIGATION

In order to improve model performance, we could utilize Yelp API to determine other factors in Neighborhoods, such as average quality of life. Although there are no numerical interpretation of this, by examining average yelp reviews for restaurants and businesses in the area, we could factor these reviews in our model.

FINAL THOUGHTS

It is important to note that although this model can determine which neighborhood selection would be ideal for individuals, it comes down to the individual to explore the area and determine what's best for them. However, by utilizing our model it would hopefully significantly reduce the choices and can highlight which Neighborhoods are best suited for them based on the features they hope to find in that Neighborhood.