New York City Rental Price Prediction

1. Introduction

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. Therefore; 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!

2. Zillow – New York City Rent Prices for 2020 Dataset

Data Cleaning and Data Wrangling

We will first examine the Zillow Dataset, which has close to 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. However it's important to note that half of the Area feature was missing for listings. We will explore this feature later on.

Among the features includes were latitude / longitude which allowed us to map these listings utilizing Google Maps API. We then use a heat map [**Figure 1**], setting the weight to be Rent Cost so that we can easily visualize which neighborhoods tend to be more expensive.

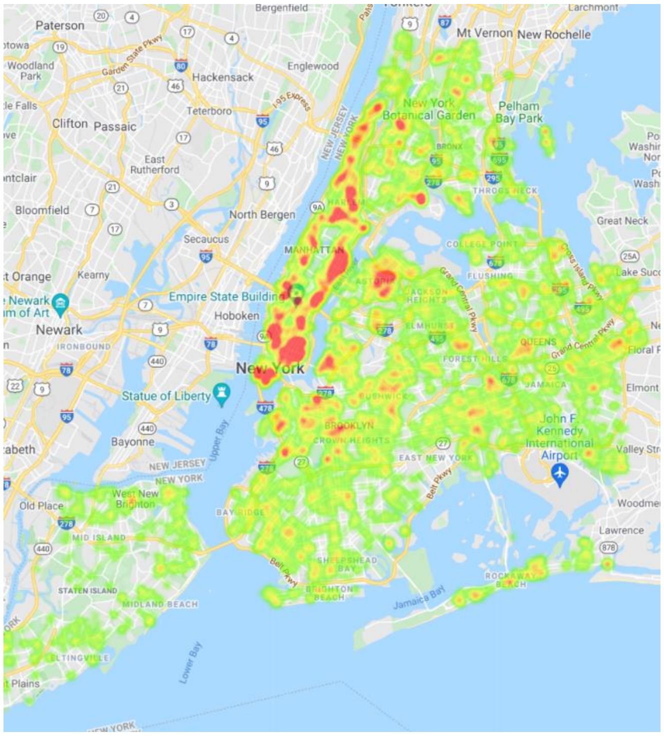

Figure 1 – Heat Map of Listings [Google API]

Exploratory Data Analysis

Comparing different Neighborhoods / Buroughs:

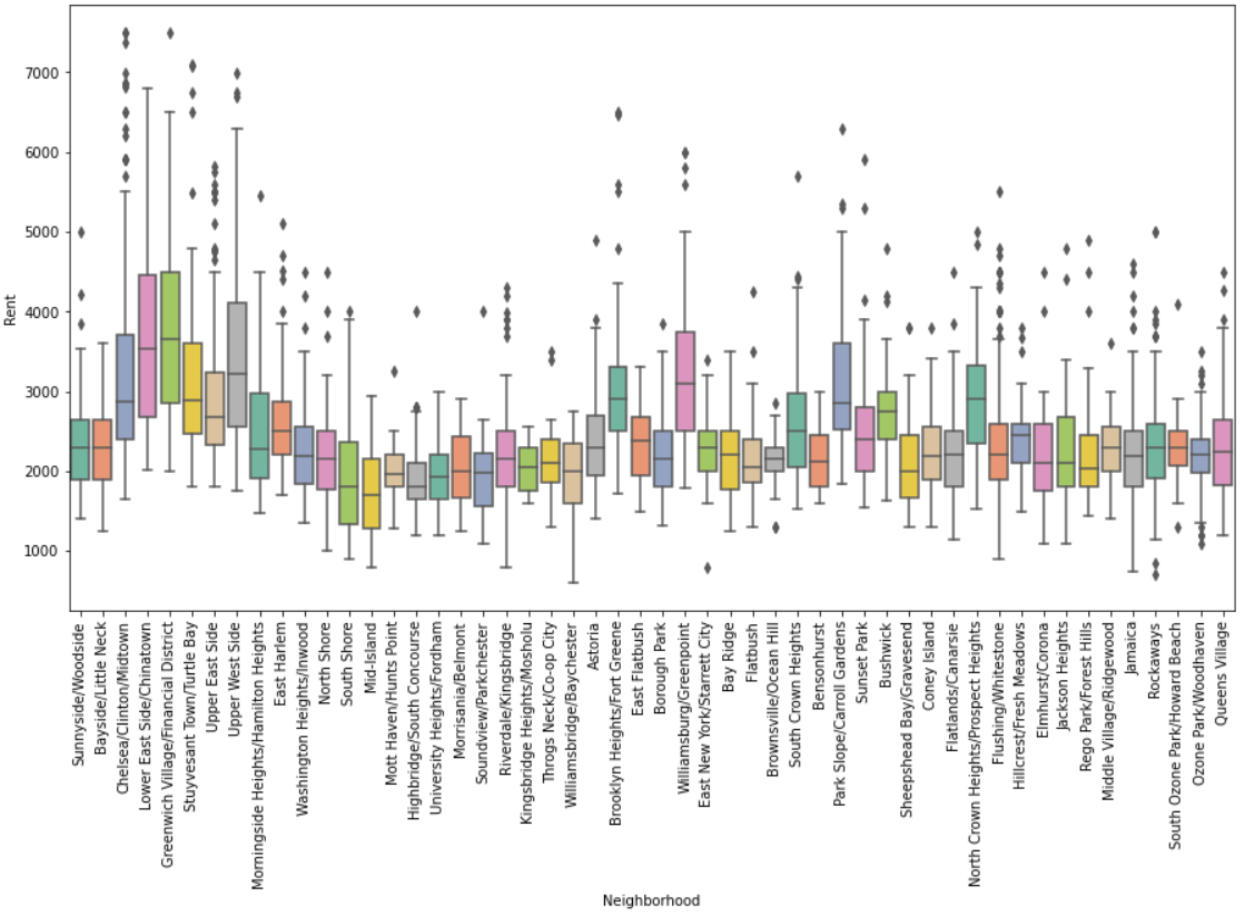

Figure 2 – Average Rent Price Fluctuation by Neighborhood

Figure 2 allows us compare the average rent per Neighborhoods. Neighborhoods in Manhattan have a higher Average Rent on average, while neighborhoods in Bronx and Staten Island are significantly less.

Figure 3 allows us to differentiate different types of listings by Borough:

- 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 the majority of units for rent in Staten Island are houses.

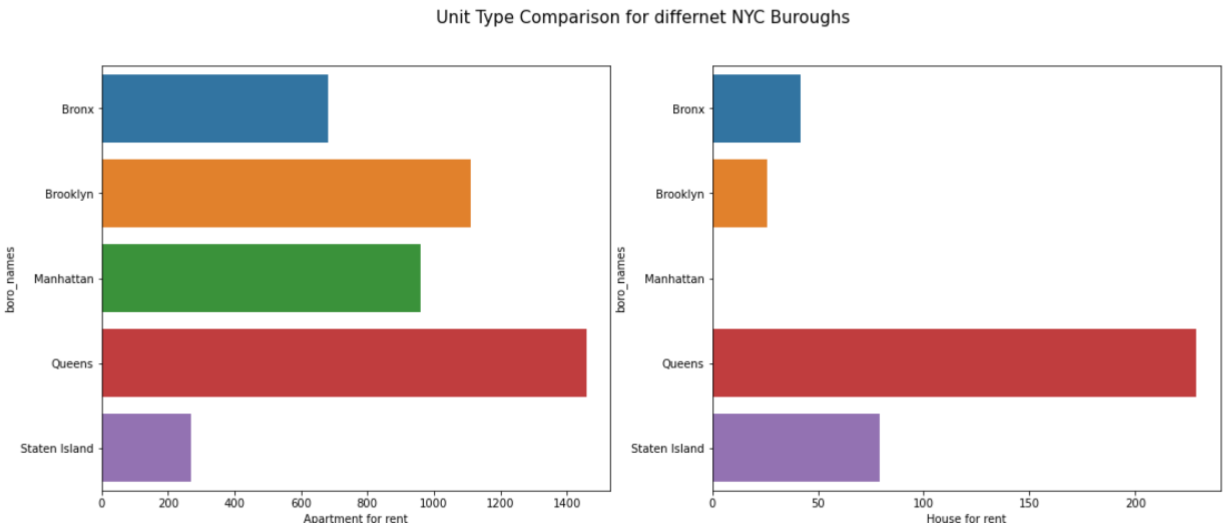

Figure 3 – Unit Type Comparison for different NYC Boroughs

3. Neighborhood Extrinsic Factors - Housing Authority Dataset

Data Cleaning and Data Wrangling

To accurately predict and model Rent Prices for listings in NYC it was very apparent we needed more information about the location of these listings. 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 and Diversity of Neighborhood; among others.[1]

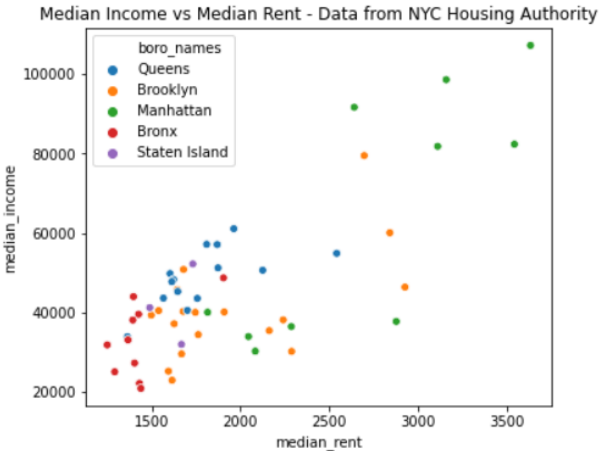

Figure 4 – Median Income vs Median Rent

Exploratory Data Analysis

We will begin by exploring different correlations that may exist between these features. All these scatter plots have positive correlations:

[1] A full listing of all column features and their definitions can be found in the project repository.

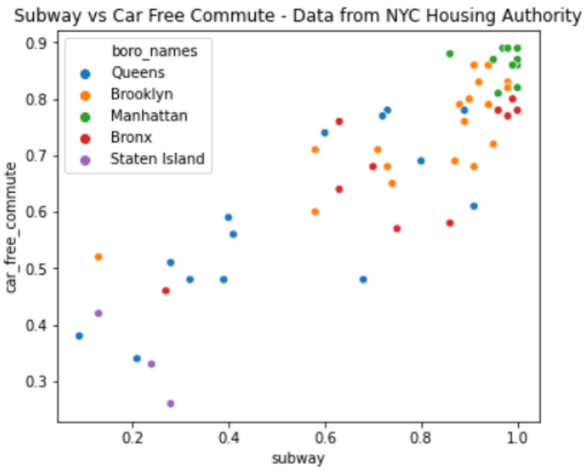

Figure 5 – Subway vs Car Free Commute

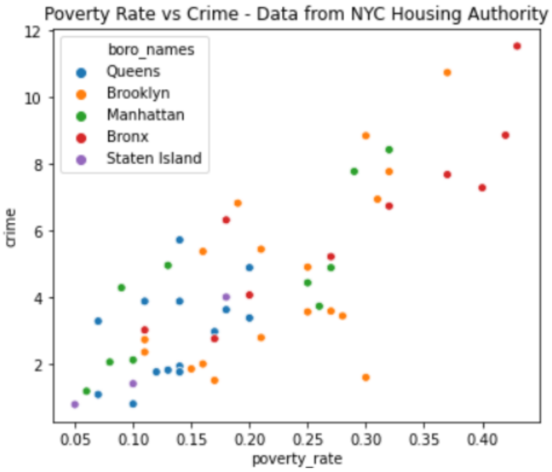

Figure 6 – Poverty Rate vs Crime

Observations:

1. **Figure 4 –** 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.
2. **Figure 5 -** 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.
3. **Figure 6 -** 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.

Therefore to further explore these correlations we will create a correlation graph. The observations can be found below:

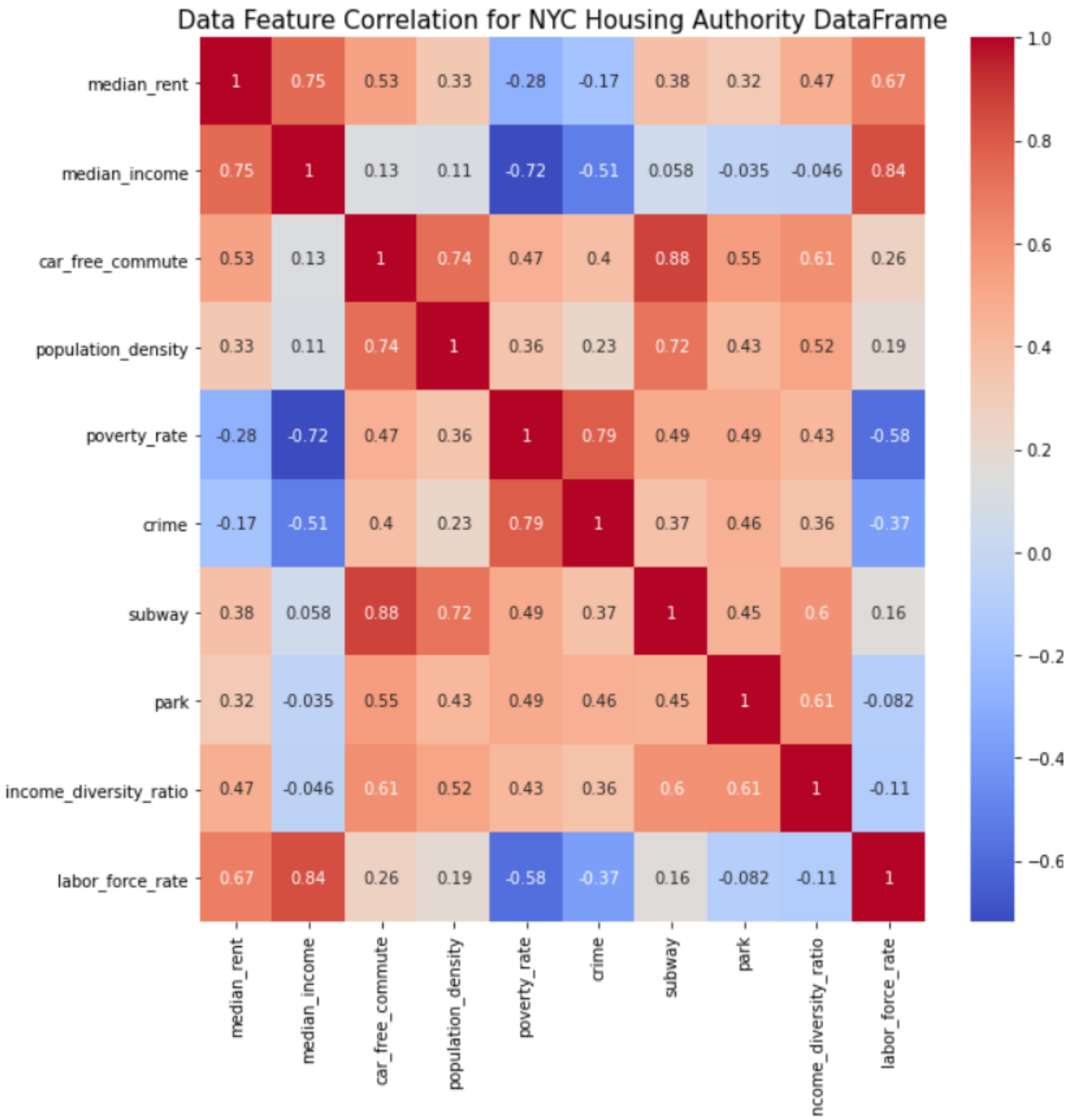

Figure 6 – Correlation Matrix for NYC Housing Dataset

Strong Positive Correlations	**Moderately Strong Negative Correlations**
• Car Free Commute / Subway (0.89) • Labor Force Rate / Median Income (0.85) • Median Income / Median Rent (0.8) • Population Density / Car Free Commute (0.73)	• Poverty Rate / Median Income (-0.67) • Labor Force Rate / Poverty Rate (0.85) • Median Income / Median Rent (0.8)

Combining both Datasets:

We can now merge both datasets on their Neighborhood, by first categorizing each listing into a neighborhood [*dependent on zip code*]. Once merged, we now have a complete dataset that can be inserted into a Pipeline.

Missing Data:

As mentioned previously more than half the listings had the 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.

- Manhattan Apartments on average have less size, which correlates since the population density in these neighborhoods is so high. On the other hand; boroughs that are more suburban tend to offer more space.

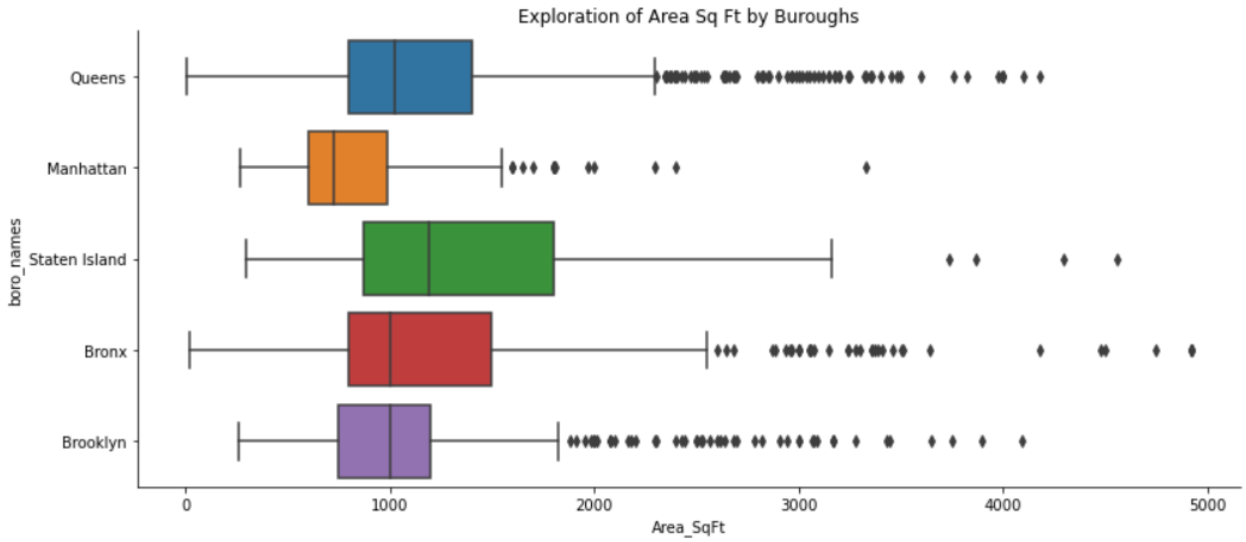

Figure 7 – Exploration of Area [Sq Ft] by Borough

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, as **Figure 8** showed this relationship in these two boroughs had the largest juxtaposition. Here we clearly see a linear relationship; that varies depending on Borough.

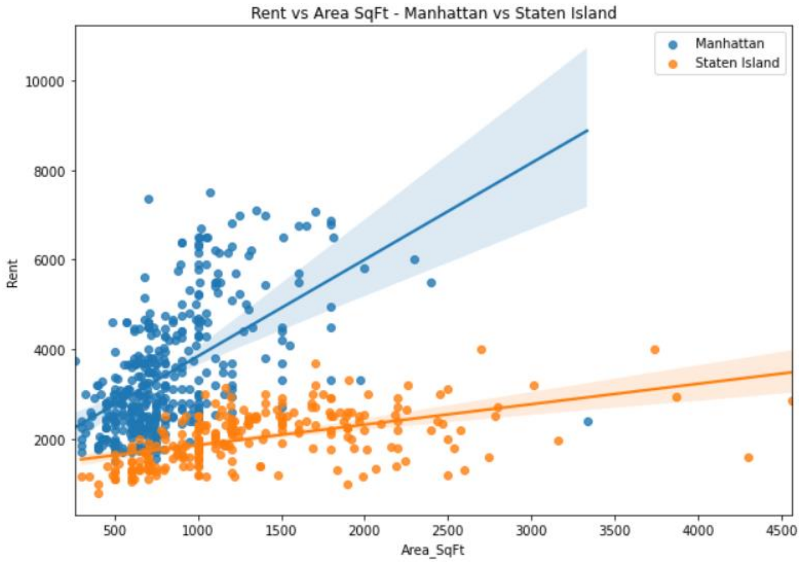

Figure 8 – Area [Sq Ft] vs Rent – Manhattan vs Staten Island

4. Principal Component Analysis

Because 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

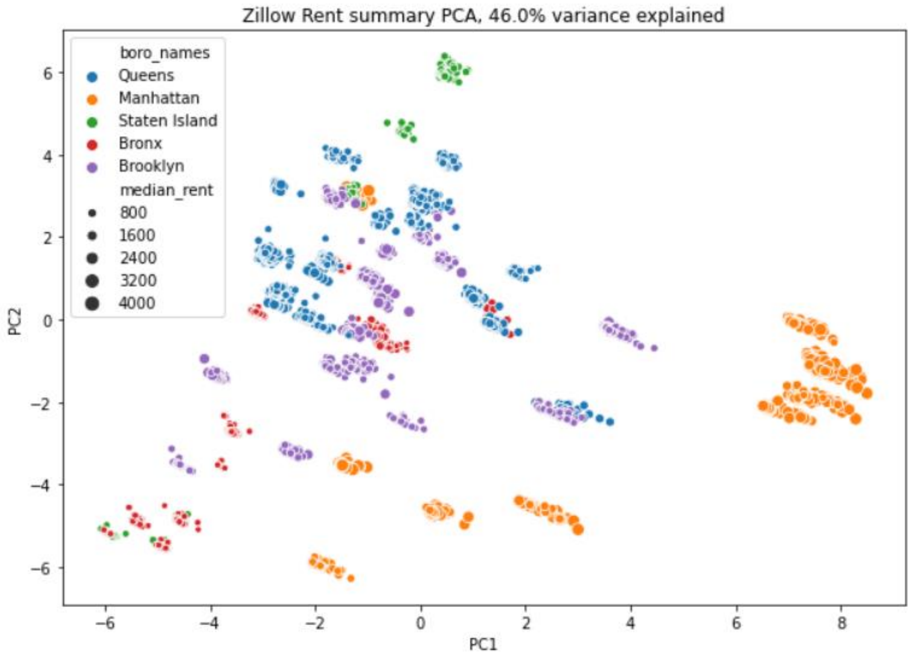

Figure 9 – PCA Analysis with Two Dimensions

Observations

1.2 components only explains 46% of the variance.

2.Manhattan listings clearly cluster by borough. However other boroughs intersect one another[2].

[2] After performing PCA Analysis on specific Burroughs, we note that the majority of the data clusters by Neighborhoods. Jupyter Notebook shows how neighborhoods cluster in Manhattan and Brooklyn

5. Modeling

Preprocessing

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.

Hyperparameter Tuning

A few important things to note is that 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[3].

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 Selection

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.

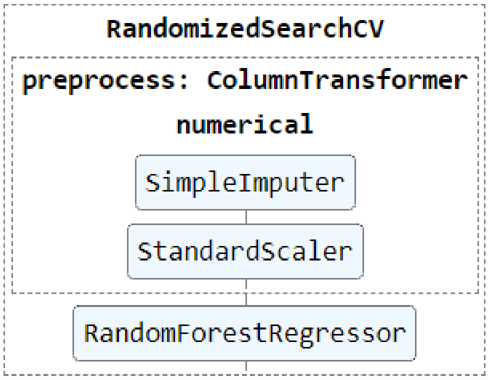

Figure 10 – Random Forest Model

[3] These Hyperparameters can be found on the Project Repository.

6. 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 exploring how different extrinsic factors vary neighborhood to neighborhood. 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[4].

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.

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.

[4] This process can be found on the Project Repository.