Predicting Airbnb Prices in New York City 2019

Aneesha Sreerama

Introduction

The goal of this analysis is to analyze the factors that determine the price of Airbnb listing in New York City. Using a variety of factors such as neighborhood, burrow, proximity to New York City attractions, number of bedrooms, and type of listing, the aim is to estimate the price of listings. With a model to estimate the average prices of Airbnbs in the area, Airbnb owners can better understand how to modify the prices off their listings to become competitive in the market. In addition, for those planning to set up an Airbnb, understanding which features are given the most importance can be crucial to the success of a listing.

Data Visualizations and Analysis

The following section shows visualizations/results and a brief description explaining the finding. This section explores the regarding:

- Location: Manhattan, Brooklyn, Queens, Bronx, Staten Island
- Neighborhood: the neighborhood the listing is in (ex: Harlem, Midtown)
- Latitude: Coordinate of the listing
- Longitude: Coordinate of the listing
- Listing Space Type: Private Home vs Entire home/apt
- Distances from each of the most popular tourist sites in NYC
- Cost: Price of the listing (in dollars)

The distribution of the NYC Airbnb prices has a strong positive skew. A large proportion of prices are within 40-100 dollars.

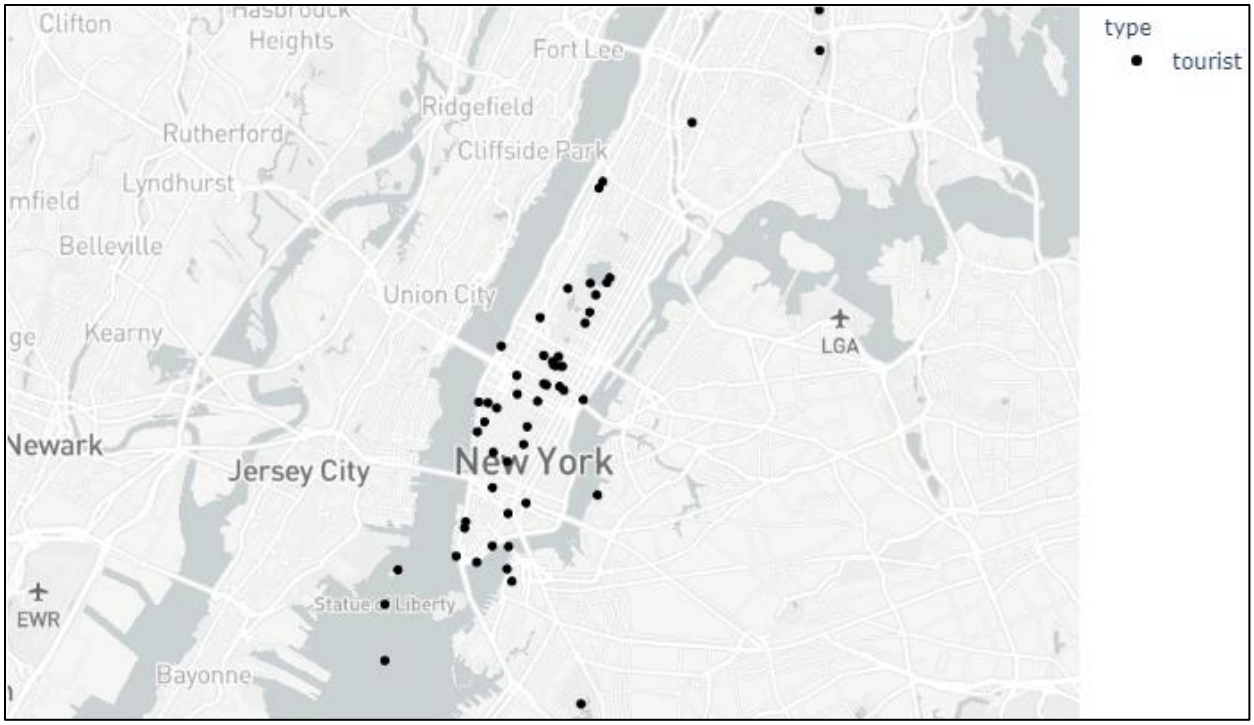
The distribution of the NYC Airbnb prices across the burrows is indicative of a few patterns. The range of prices in Brooklyn and Manhattan is significantly greater in comparison to Staten Island or Bronx. Across all the burrows, the median house prices are about the same across the board.

		count	mean	std	min	\
neighbourhood_group	room_type					
Bronx	Entire home/apt	377.0	179.076923	147.832215	27.0	
	Private room	649.0	169.211094	134.395756	25.0	
	Shared room	60.0	222.833333	162.078571	45.0	
Brooklyn	Entire home/apt	9514.0	130.199916	118.598127	24.0	
	Private room	10079.0	140.682508	129.637931	24.0	
	Shared room	410.0	190.134146	176.793427	35.0	
Manhattan	Entire home/apt	13153.0	145.628678	133.974287	24.0	
	Private room	7926.0	138.958491	117.966417	24.0	
	Shared room	476.0	167.060924	155.216692	25.0	
Queens	Entire home/apt	2089.0	165.479177	150.676201	25.0	
	Private room	3355.0	167.938301	148.893416	25.0	
	Shared room	198.0	199.939394	158.877363	35.0	
Staten Island	Entire home/apt	176.0	165.000000	127.223807	30.0	
	Private room	187.0	155.641711	138.057667	25.0	
	Shared room	9.0	273.000000	189.459098	135.0	

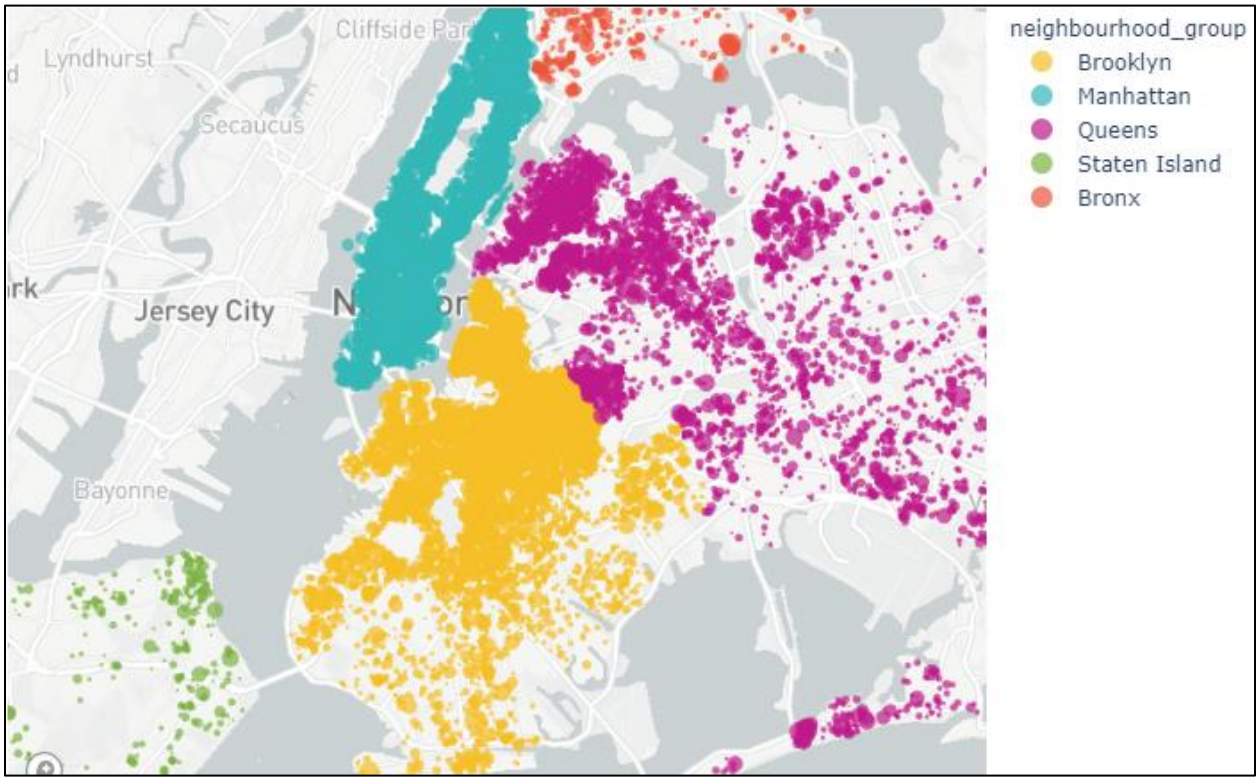
		25%	50%	75%	max
neighbourhood_group	room_type				
Bronx	Entire home/apt	90.0	140.0	210.00	1000.0
	Private room	95.0	135.0	200.00	1000.0
	Shared room	115.0	195.0	300.00	1100.0
Brooklyn	Entire home/apt	60.0	99.0	150.00	1700.0
	Private room	70.0	105.0	169.50	1763.0
	Shared room	95.0	150.0	213.75	1315.0
Manhattan	Entire home/apt	65.0	100.0	185.00	1750.0
	Private room	70.0	100.0	175.00	1600.0
	Shared room	75.0	129.0	200.00	1497.0
Queens	Entire home/apt	80.0	125.0	199.00	1600.0
	Private room	89.0	130.0	200.00	1700.0
	Shared room	109.0	165.0	250.00	1369.0
Staten Island	Entire home/apt	88.0	132.5	204.75	900.0
	Private room	93.0	125.0	170.00	1000.0
	Shared room	165.0	175.0	315.00	737.0

Relevant insights from this information include

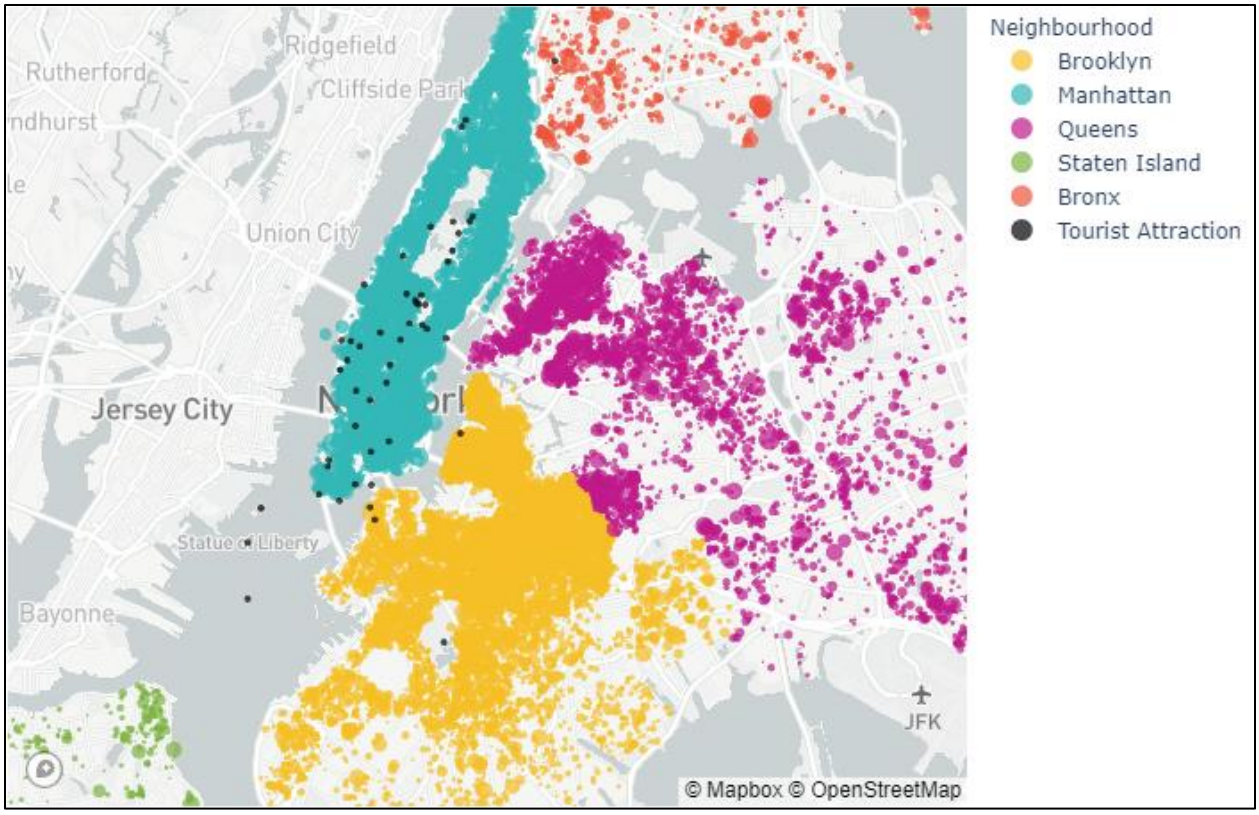
1. Private rooms tended to cost more than an entire home/apt.
2. There were a lot of outliers on the higher end of the price range (represents the proportion of Airbnb owners who rent out their listing for extended periods of time)
3. Brooklyn and Manhattan prices tended to significantly higher than in any other location
4. The average price of a listing was the highest in Staten Island, as opposed to Brooklyn or Manhattan (due to the lower number of listing here)



Most tourist locations were located in the Manhatten burrow in New york City.



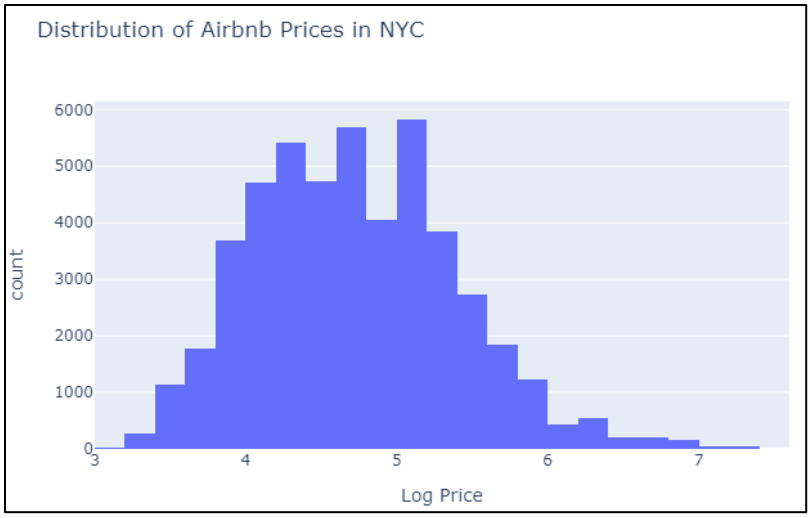
The density of Airbnb locations in Manhattan and Brooklyn tend to be a lot higher than in any other part of New York City.



This is a visual representing the Tourist attractions and the Airbnbs located in New York City and this image led to the question; Does proximity to tourist sites influence the prices of Airbnbs?

```
LINEAR REGRESSION
------------------------

        R-squared value for training set:  0.11282686997584668
        R-squared value for testing set:  0.09149973269559786

        Average Deviation from Expected Value:  126.82718112353577

RIDGE REGRESSION
------------------------

        R-squared value for training set:  0.10033135987708608
        R-squared value for testing set:  0.09450210881398835

        Average Deviation from Expected Value:  126.6174409432067

LASSO REGRESSION
------------------------

C:\Users\Aneesha\anaconda3\lib\site-packages\sklearn\linear_model\_coordinate_descent.py:529: ConvergenceWarning:

Objective did not converge. You might want to increase the number of iterations. Duality gap: 137170338.7073434, tolerance: 621
52.17085359385


        R-squared value for training set:  0.11131008100929207
        R-squared value for testing set:  0.0922844943014014

        Average Deviation from Expected Value:  126.77239269447692

K-NEIGHBORS REGRESSION
------------------------

        R-squared value for training set:  0.48789226908138517
        R-squared value for testing set:  0.34015335928071977

        Average Deviation from Expected Value:  108.08646545873214

SVR REGRESSION
------------------------

        R-squared value for training set:  0.021636022731539195
        R-squared value for testing set:  0.01952027571774173

        Average Deviation from Expected Value:  131.7556108228378

MLP REGRESSION
------------------------

        R-squared value for training set:  0.47241905918842775
        R-squared value for testing set:  0.4218124262946913

        Average Deviation from Expected Value:  101.17755656416473
```

This result represents the r-squared value for the training and testing sets using all 290 features to predict the price of an Airbnb. The Linear, Ridge, and Lasso regressors poorly fit the dataset as these regressors scale poorly when using multiple dimensions. K-Nearest Neighbors Regressor and the MLP Regressor performed significantly better with this dataset.



In order to improve the performace of the algorithm I noticed that the prices data had a very skewed distribution as seen in the first histogram. I took the log of each of the prices in order to normalize the dataset, in hopes that a niormally distributed dataset would perform better with kNN algorithm.

```
LINEAR REGRESSION
------------------------

        R-squared value for training set:  0.1770967970708398
        R-squared value for testing set:  0.16587991150773362

        Average Deviation from Expected Value:  0.6195582787150735

RIDGE REGRESSION
------------------------

        R-squared value for training set:  0.1664101504169586
        R-squared value for testing set:  0.16781291147460953

        Average Deviation from Expected Value:  0.6188399764042829

LASSO REGRESSION
------------------------

        R-squared value for training set:  0.15706576045403853
        R-squared value for testing set:  0.15999363758059892

        Average Deviation from Expected Value:  0.6217405055258263

K-NEIGHBORS REGRESSION
------------------------

        R-squared value for training set:  0.5068850233616201
        R-squared value for testing set:  0.3634705185856607

        Average Deviation from Expected Value:  0.5412241110588201

SVR REGRESSION
------------------------

        R-squared value for training set:  0.16283663648908964
        R-squared value for testing set:  0.1633737605951836

        Average Deviation from Expected Value:  0.6204883258520628

MLP REGRESSION
------------------------

        R-squared value for training set:  0.43127313689864133
        R-squared value for testing set:  0.4017052005162177

        Average Deviation from Expected Value:  0.5247174268102723
```

This change resulted in the Linear, Ridge, and Lasso regressors to perform significantly better. The k-NN regressor slightly overfit the training dataset but still had an improved performance when used on the testing dataset. The MLP regressor saw a slight decrease in terms of performance.

```
LINEAR REGRESSION
------------------------

        R-squared value for training set:  0.09455321010558759
        R-squared value for testing set:  0.0916929835523711

        Average Deviation from Expected Value:  126.81369143849533

RIDGE REGRESSION
------------------------

        R-squared value for training set:  0.08831050304176213
        R-squared value for testing set:  0.08800712196206772

        Average Deviation from Expected Value:  127.07073254715483

LASSO REGRESSION
------------------------

C:\Users\Aneesha\anaconda3\lib\site-packages\sklearn\linear_model\_coordinate_descent.py:529: ConvergenceWarning:

Objective did not converge. You might want to increase the number of iterations. Duality gap: 236711876.742755, tolerance: 6215
2.17085359385

        R-squared value for training set:  0.09351615437146343
        R-squared value for testing set:  0.09207755850353261

        Average Deviation from Expected Value:  126.78684229409308

K-NEIGHBORS REGRESSION
------------------------

        R-squared value for training set:  0.4888859946790073
        R-squared value for testing set:  0.34552288646188767

        Average Deviation from Expected Value:  107.64578825499038

SVR REGRESSION
------------------------

        R-squared value for training set:  0.01820565234593685
        R-squared value for testing set:  0.01628552734500599

        Average Deviation from Expected Value:  131.97277253186337

MLP REGRESSION
------------------------

        R-squared value for training set:  0.40661440570855334
        R-squared value for testing set:  0.41134186811668527

        Average Deviation from Expected Value:  102.08957223740389
```

This result was generated after the use of Recursive Feature Elimination. The goal was to remove features that had a negligible impact on the result of the dataset. Scaling down the dataset failed to result in any significant improvements.

Conclusions

In order to analyze our data, I began by visualizing the dataset using boxplots, histograms, and a series of maps that that plotted the individual tourist sites, listings, and both together on a map on New York City. Then we proceeded to use feature engineering to exclude a few categorical variables and split variables such as Location and Neighborhood into groups. I proceeded to test a few classifiers such as Linear, Lasso, Ridge, kNN, SVR, and the MLP regressors. To improve the performance, I ran the regressors once again while using the natural log of the prices in order to have a normal distribution. This greatly improved the accuracy of the Linear, Ridge, SVR, and Lasso regressors. However, the R-squared values were quite low. Based on these results, I used Recursive Feature Elimination to reduce the number of variables in the dataset in hopes of improving the accuracy of the kNN regressor. As a result, this modification slightly improved the performance of the kNN regressor. However, the MLP regressor had the better performance when using all 290 features.

I compared Linear, Ridge, Lasso, kNN, SVR, and the MLP regressors. MLP Regression had an R-squared value of 0.4724 on the training set and 0.4218 on the testing set. This indicates that the model fit the dataset quite well and the testing set showed that this model has potential as it has ~65% accuracy. kNN Regression had an R-squared value of 0.4889 on the training set and 0.3455 on the testing set when using Recursive Feature Elimination. This indicates that the model slightly overfit the dataset and the testing set showed that this model was an average fit for the dataset but did not perform as well as the MLP regressor for the given data.

The MLP regressor had the best fit for this dataset and utilized a variety of variables to output better predictions. This regressors also embodies the fact that many features are important when determining where to open an Airbnb and at what price to do so. This model has a ~65% accuracy and that is a reasonable accuracy for a versatile question.

References

1. 55 Best Things to Do in New York City (New York). (2020, January 26). Retrieved January 02, 2021, from https://www.thecrazytourist.com/top-25-things-to-do-in-new-york-city/