Bach Quang Minh
MS Candidate in Computer Science – Oriented Data
John Von Neumann Institute
Mail: minh.bach2017@ict.jvn.edu.vn

JVN Institute
Vietnam National University HCM

AIRBNB REPORT

Data description

The data contains 30478 records of hosts in New York City for travelers. There are 13 attributes and here are some attributes of interests:

Neighbourhood: Where the hosts locate.

Property Type: whether the place is an apartment, house, loft or other types of properties.

Room Type: The type of room to be offered.

Price: The cost for renting

Review Scores Rating: Reviews of users for each host.

22155 records have been reviewed, while the rest have not.

There are a significant number of missing values. These will be preprocessed before the analysis.

Purpose

The purpose of this analysis is to answer these following questions:

Question 1: How does the price distribute by Neighbourhood, Property Type and Room Type? This means we are interested in knowing some of the more specific questions such as which neighbourhood has the highest/lowest price? Or which property type and room type usually cost more,…

Question 2: Is there a relationship between the factors Neighbourhood, Property Type, Room Type and the price? If so, which factor has the most impact on the price?

Question 3: Which sector is offering good service?, which mean earning high review scores. The sectors can be viewed in terms of Neighbourhood, Property Type and Room Type.

Question 4: An attempt to build a predictive model for review score using Neighbourhood, Property Type, Room Type and Price.

Data exploration

This section, we start building some charts to get the overall structure of the data. We will also see the focus of the market and its size in different aspects. The figure 1 shows a dashboard created by Power BI and the interactive version can be found here.

Starting at the upper left panel, listings are highly concentrated with more than 80% in Manhattan and Brooklyn while the rest of NYC makes up for a humble share of less than 20%. Among the two neighbourhoods of highest number of listings, higher supply is seen in Manhattan than Brooklyn (52.61% versus 38.31%). Hence most of the market is dominated by Manhattan and Brooklyn. On the other hand, Bronx and Statenisland are the two least favorable places to stay in NYC, accounting for only 1.13% and 0.48% respectively.

On the upper right panel, Apartment stands out to be the most usual property type although there are a handful of different property types. Apartment might be the best-fit type in terms of supply and demand.

In the third panel, most records contain room types as Entire home (17k) and Private room (12.6k). Shared room only is the least favorable room type, only around 800 listings.

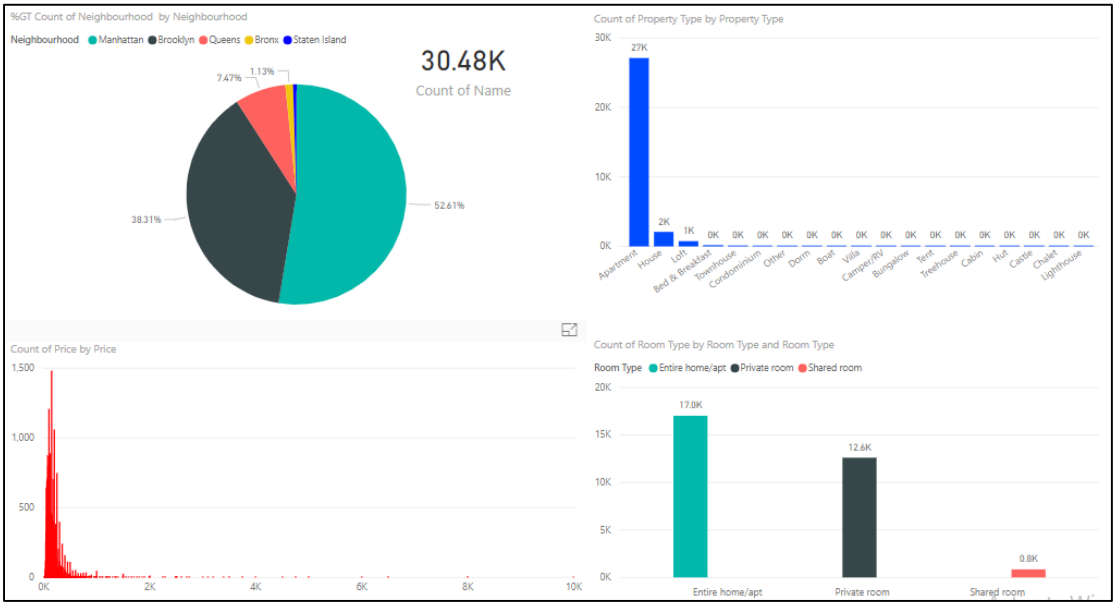

Figure 1.1: The dashboard shows the distribution of the records by Neighbourhood, Property Type and Room Type.

Moreover, we can see more out of this dashboard. For instance, we would like to see where the House listings mostly locates. By clicking the House column in the upper right panel, we can see that most houses locate in Brooklyn (57.51%) and other boroughs (34.45%), as opposed to the overall distribution, which centralizes in Manhattan.

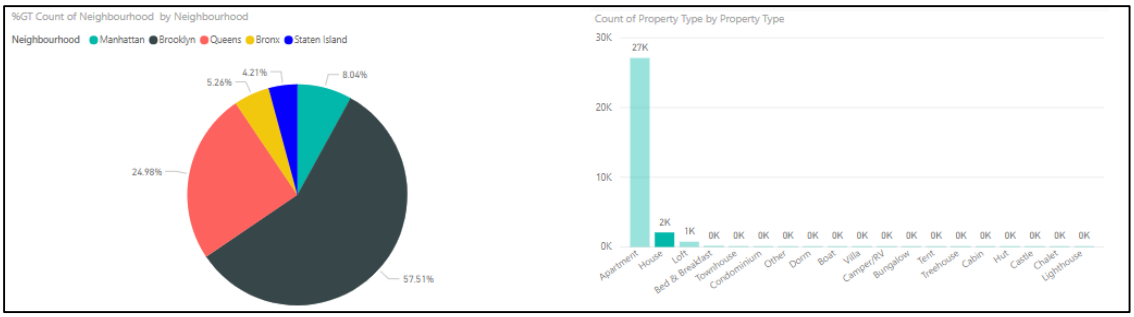

Figure 1.2: House listings locations

We might also interest in what property types are available in each neighbourhoods. As can be seen in figure 1.3, although Manhattan contain most listings, Brooklyn seems to offer more variety of property types. Queen contains quite even proportions of different property types.

Figure 1.3: Property types by the three most crowded neighbourhoods.

The analysis

For the analysis, Python is used as tools for making charts and building models.

Question 1: How does the price distribute by Neighbourhood, Property Type and Room Type?

Neighbourhood

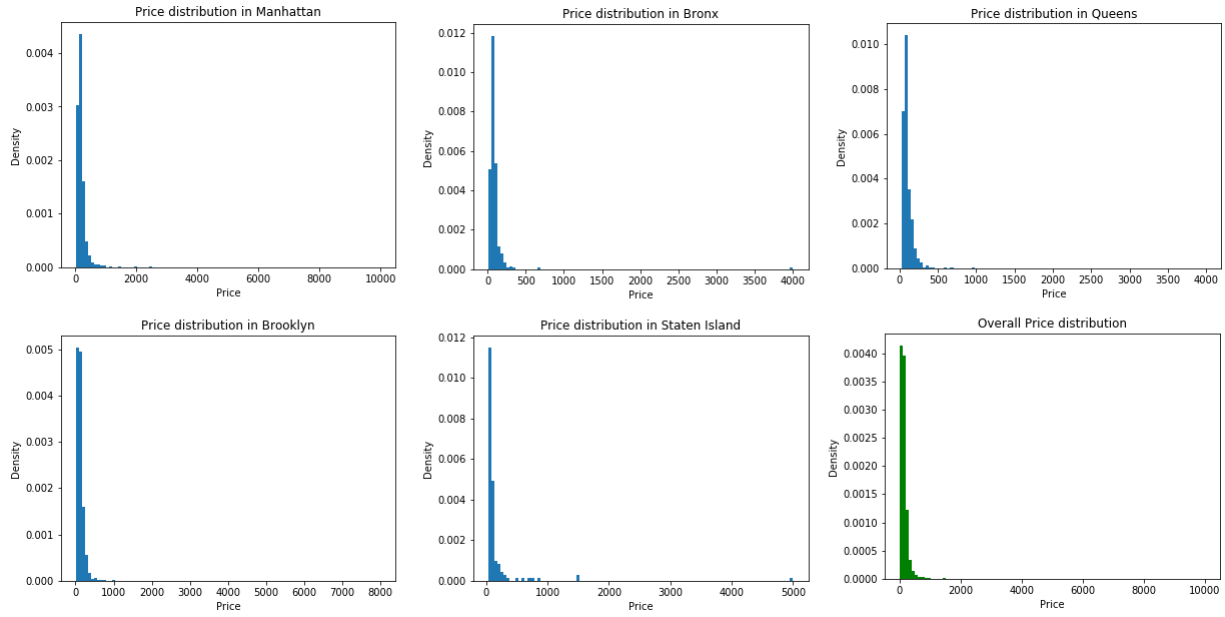

Figure 2: The panels show the price distributions by Neighbourhood and the overall distribution of the price.

First look at the overall price distribution, the market price centralized mostly under $500. The x-axis spreads up to $10000, indicating some places achieve very high price.

The price distributions by Neighbourhood and the overall price distribution have similar shape. They are skew distributions with high variance, therefore, the median should be chosen as the average to represent the distribution.

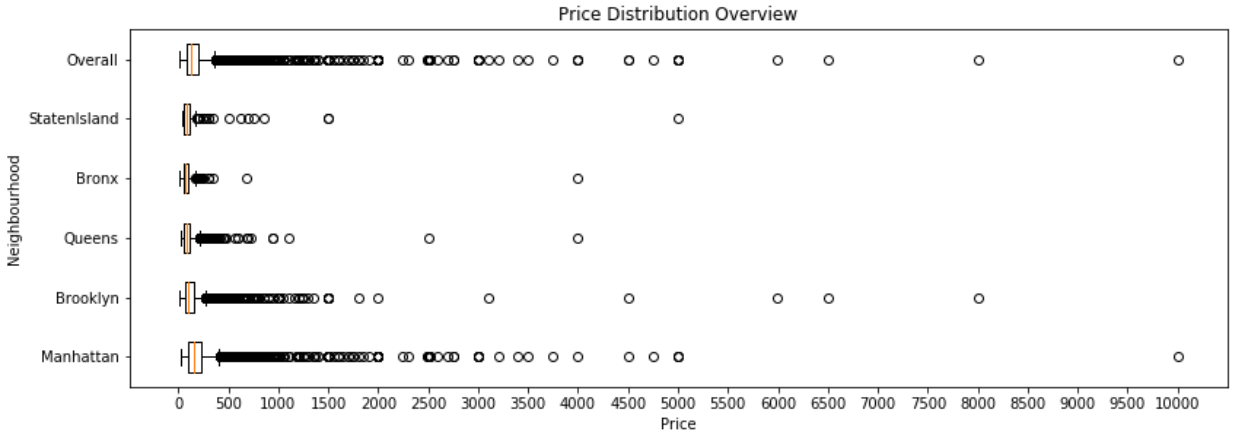

Figure 3: The box-whisker plot shows the overview of the price distribution by Neighbourhood.

Generally, the prices centralize mostly under $500. Manhattan sees quite a handful of high-priced records, over $500 to $2000. There is an unusually high-priced place here, toping at around $10000. Bronx and Queens have moderate price around $500.

The summary statistics of the price is computed in the following table.

Table 1: Summary statistics of the price by Neighbourhood.

Index	Brooklyn	Bronx	Manhattan	Queens	StatenIsland	Overall
count	11675	345	16033	2278	147	30478
mean	129.5	94.6609	198.475	103.222	163.463	163.59
std	155.387	218.421	221.815	119.56	450.109	197.785
min	10	10	20	25	35	10
25%	70	50	100	60	59	80
50%	100	69	155	80	79	125
75%	150	99	225	120	109	195
max	8000	4000	10000	4000	5000	10000

The 50th percentile is the median price, the average price, in each neighbourhood. The lowest average renting price is in Bronx and the highest average renting price, as expected, is in Manhattan.

To sum up, each neigbourhood has various prices catering different budgets of tourists. Manhattan and Brooklyn are tourists' favourite neigbourhoods to stay and there are more premium places than other neighbourhoods. Queens, StatenIsland and Bronx are more affordable, however, are not top tourist choices to stay.

Property Type

From the figure 1, the three significant Property Types are Apartment, House and Loft. The rest of them will be grouped into one group named Others for analysis.

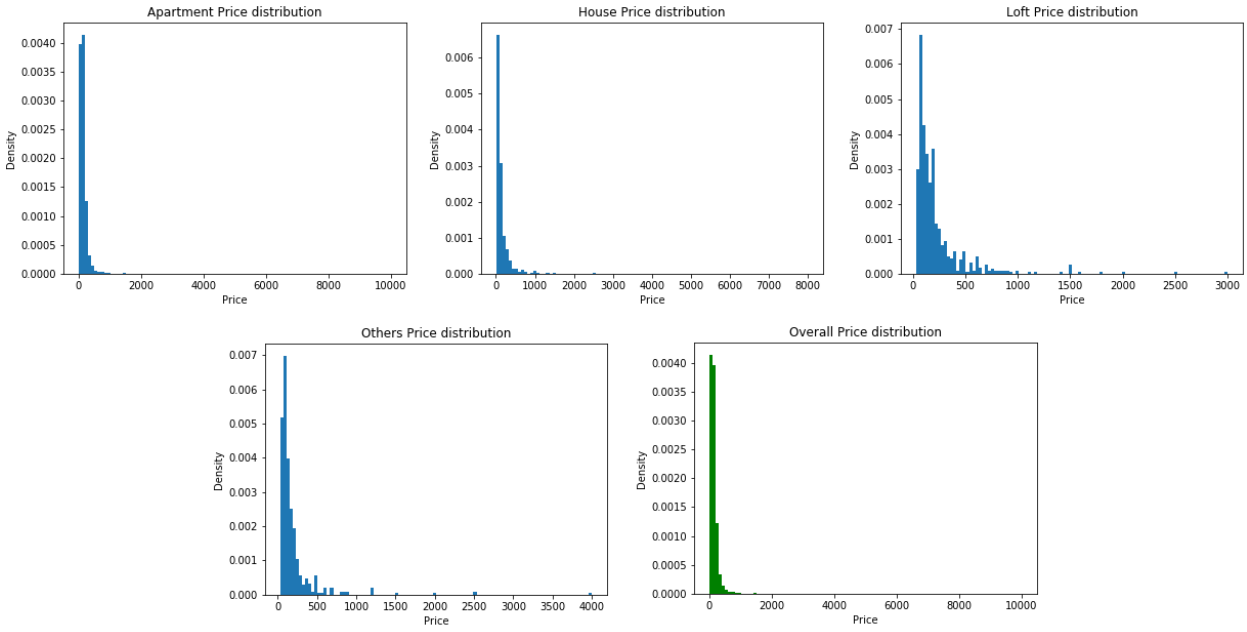

Figure 4: The panels show the price distributions by Property Type and the overall distribution of the price.

Apartment price distribution resembles the overall distribution. This fits the fact that a very large proportion of records are apartments, other types although available, however, only accounts for small amounts, leading different shapes distributions.

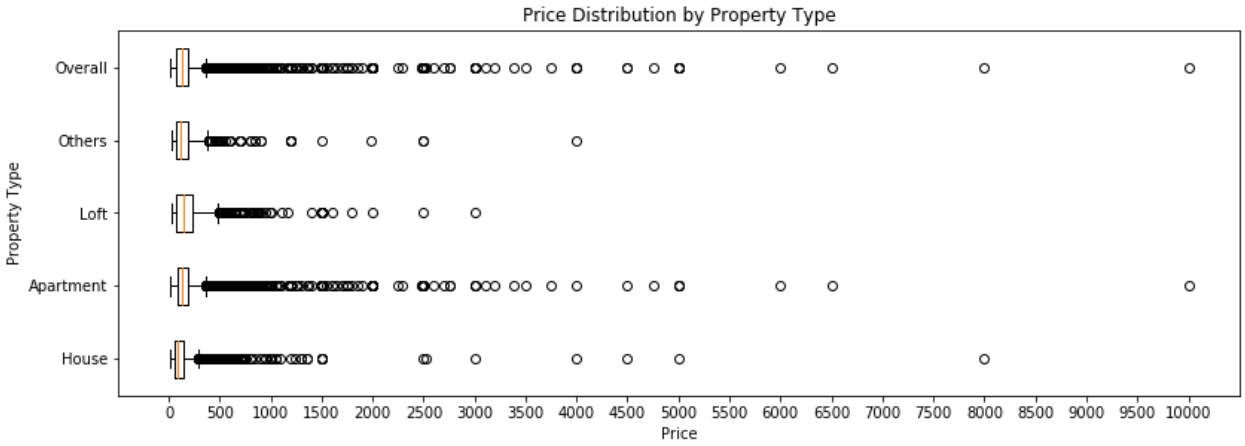

Figure 5: The box-whisker plot shows the overview of the price distribution by Property Type.

Apartment shows to offer variety of budgets. Prices are mostly around 85 to 195, although higher price places are also available. House seems to offer more choices than loft. With a price less than $150, one can find 1567 House listings (75%), meanwhile, only about 500 records of Loft type.

Table 2: Summary statistics of the price by Property Type.

Index	Apartment	House	Loft	Others	Overall
count	27105	2090	753	530	30478
mean	162.044	156.723	221.96	186.802	163.59
std	181.367	307.117	271.958	291.673	197.785
min	10	10	30	30	10
25%	85	60	80	75	80
50%	130	85	140	115	125
75%	195	150	240	195	195
max	10000	8000	3000	4000	10000

Room Type

In a similar way, we would like to see how the price distributes among the room types.

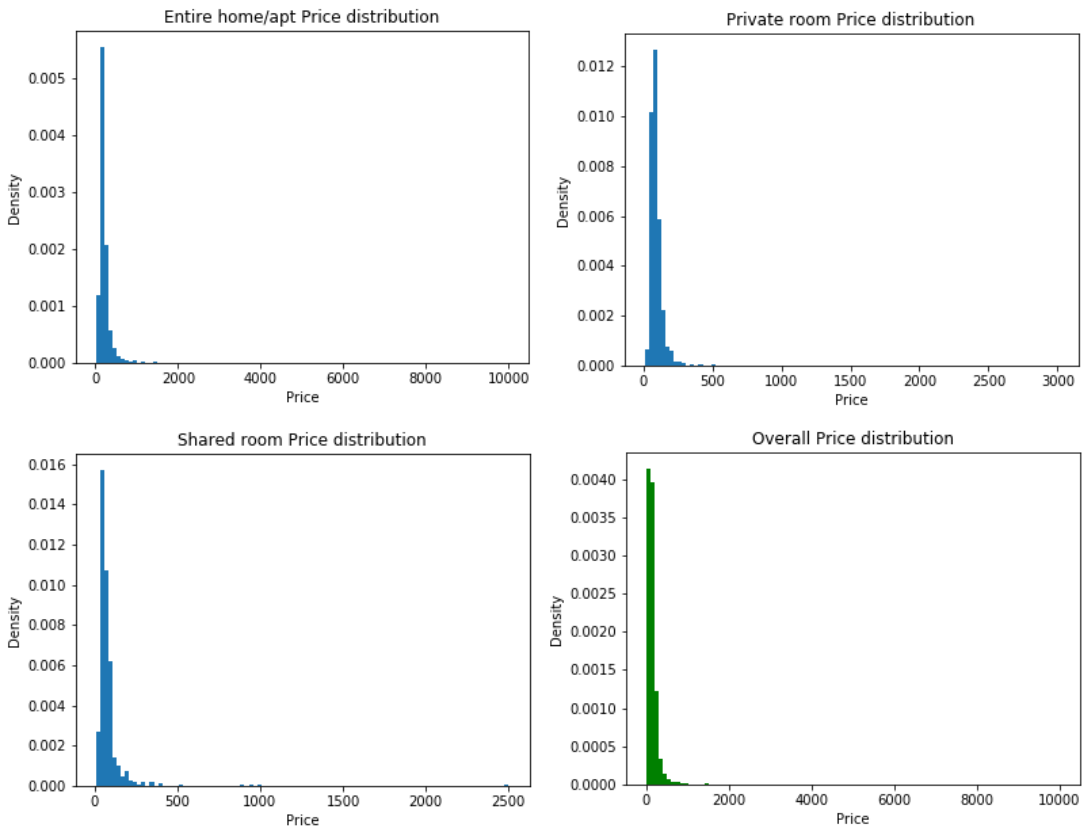

Figure 6: The panels show the price distributions by Room Type and the overall distribution of the price.

One may argue that from previous experience about these distributions we would also expect the distributions this time to be similar. In fact, this reasoning is not certain, there are no guarantee that the conditional distributions of the price given different contexts will be resemble the overall distribution of the price. Therefore, it is careful that we must plot these distributions. These are again skew distributions which should be presented using medians.

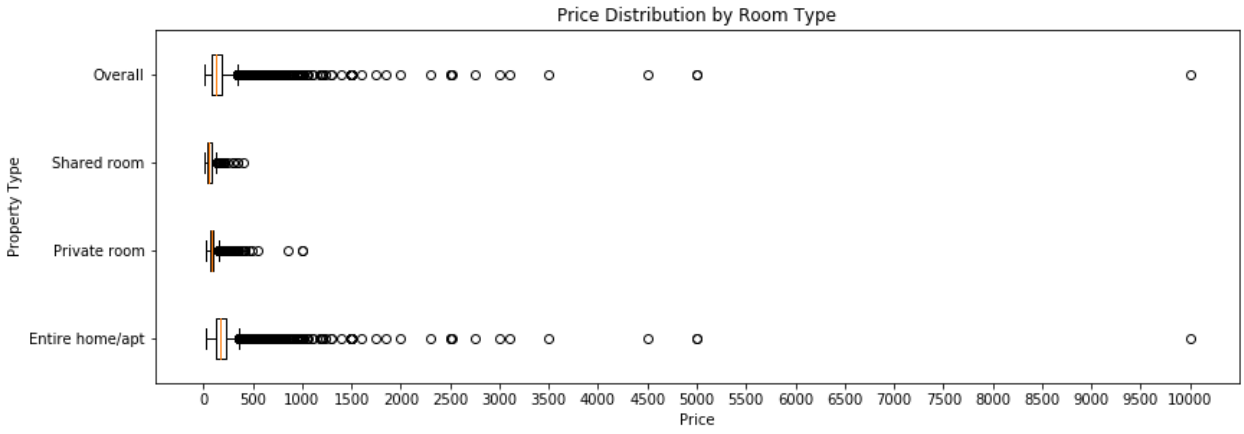

Figure 7: Price distribution by Room Type.

Shared-room and Private-room instances have their price strongly centralized about the median, wondering around $250. Entire home/apt captures most of the high-priced instances.

Table 3: Summary of the price by Room Type.

Index	Entire home/apt	Private room	Shared room	Overall
count	17024	12609	845	30478
mean	220.796	91.9197	80.5148	163.59
std	242.222	67.1197	110.756	197.785
min	10	10	10	10
25%	135	65	45	80
50%	175	80	60	125
75%	240	100	85	195
max	10000	3000	2500	10000

As can be seen from the table, the average prices of Entire, Private and Shared room are $175, $80 and $60 respectively.

Question 2: Is there a relationship between the factors Neighbourhood, Property Type, Room Type and the price? If so, which factor has the most impact on the price?

From the analysis above, we can see that the price of an instance changes with respect to its Neighbourhood, Property Type and Room Type. 'But **which of the features has the strongest impact on the price?** In this section, we attempt to build a regression model using the price as the dependent variable and the other features as independent ones.

Since Neighbourhood, Property Type and Room Type are all categorical variables and our goal is the fit a regression model for the price, therefore, we encode the values of these features to transform them into numerical data.

We will first try a simple linear regression to build the model. We first check for correlations among the features and between features and the price.

Table 4: The correlation values of features of interest.

Index	Neighbourhood	Property Type	Room Type	Price
Neighbourhood	1	-0.0495202	-0.0490931	0.078922
Property Type	-0.0495202	1	0.0942585	0.0337413
Room Type	-0.0490931	0.0942585	1	-0.313884
Price	0.078922	0.0337413	-0.313884	1

Most features have low correlation to each other, which indicates that they are quite independent variables. The red frame shows the correlation of each variables with the price. Generally they have quite low correlation values. However, we can see that Neighbourhood and Room Type have higher correlation with the price than the Property Type. This fits the intuition that most records are Apartment, therefore, the feature could not contribute much to the analysis. To that extent, we decide to fit a linear regression model with quadratic terms of Neighbourhood and Room Type.

OLS Regression Results

Dep. Variable:	Price	R-squared:	0.115
Model:	OLS	Adj. R-squared:	0.115
Method:	Least Squares	F-statistic:	633.0
Date:	Mon, 23 Apr 2018	Prob (F-statistic):	0.00
Time:	11:44:48	Log-Likelihood:	-1.6297e+05
No. Observations:	24382	AIC:	3.260e+05
Df Residuals:	24376	BIC:	3.260e+05
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	92.6919	7.040	13.167	0.000	78.893	106.490
Neighbourhood	135.2829	7.745	17.466	0.000	120.101	150.464
Property Type	30.2814	2.337	12.957	0.000	25.701	34.862
Room Type	-179.5426	5.946	-30.196	0.000	-191.197	-167.888
Neighbourhood squared	-32.4818	2.110	-15.396	0.000	-36.617	-28.347
Room Type squared	54.8411	4.281	12.812	0.000	46.451	63.231

Omnibus:	52020.307	Durbin-Watson:	2.005
Prob(Omnibus):	0.000	Jarque-Bera (JB):	365135210.617
Skew:	18.706	Prob(JB):	0.00
Kurtosis:	601.343	Cond. No.	38.4

Figure 8: Results of the regression model

R-squared shows the goodness of fit of the model. In this case, we don't seem to have a good R-squared value, which should be closed to 1. This is expected since this relationship between price and the features could be very complex. This model is just an initial rough estimation to the data.

The F-statistic gives an idea of whether the there is a relationship between the features and the response (price). The farther away from 1 the F-statistic is, the better the confirmation of the relationship. Therefore, in this case we can say that the selected features have a relationship with the price.

In the next red-framed zone, we can read the coefficient of each feature and its corresponding p-value. Here, all the p-values are very small, indicating all the features are statistically significant. The coefficients also show that the Neigbourhood (135.2829) and the Room Type (-179.5426) are weighted higher than the Property Type (30.2814) in terms of absolute weight, indicating higher effects of these features on the price.

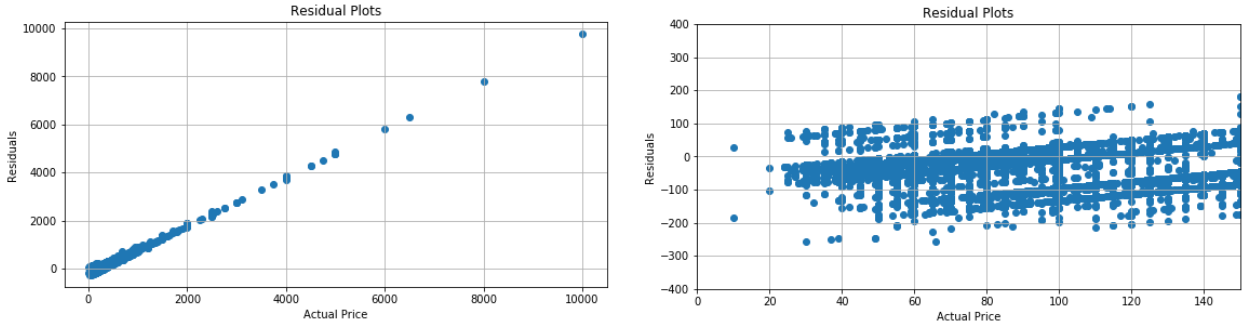

Figure 9: Residual plots in two different ranges of prices.

The residuals are distances between predicted values and true values. It could be a good indicator of how well the model perform. As can be seen from the left panel, the performance gets worse as the price increases, which is expected. If we zoom in the lower price segments, better performance is observed.

To measure the accuracy of the model, we use MAPE (Mean Absolute Percentage Errors) as the metrics. The training and test MAPE values are computed as 0.4547 and 0.4564 respectively.

In conclusion, this model is just a rough estimate. However, it offers us more confident in the features. The Neighbourhood and Room Type have higher effects on the price compared to Property Type.

Question 3: Which sector is offering good service?

This sector we will look at the review scores in different contexts. The data for this section are 22155 hosts which was reviewed. First, we want to see the overall distribution of the review score.

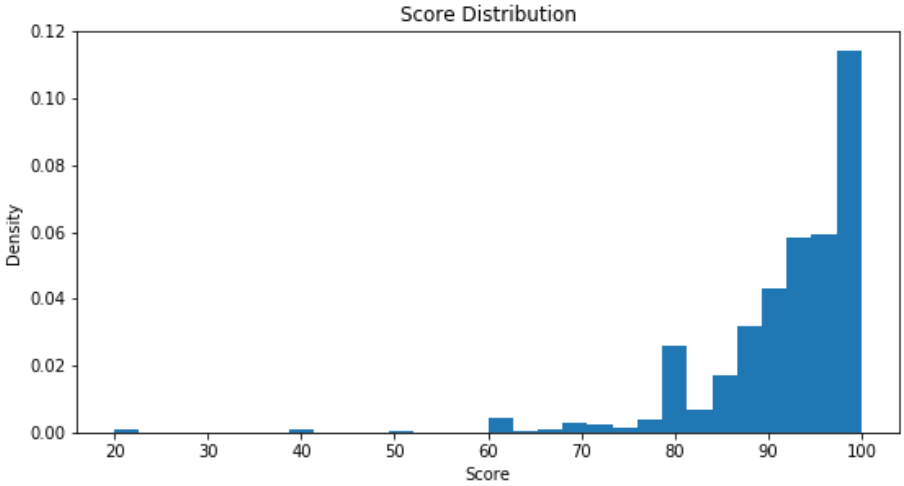

Figure 10: The overall score distribution

As the histogram suggested, most of the hosts are offering satisfying services. The scores centralize above 85 points.

Now, we dig deeper to see how the Neighbourhoods are doing.

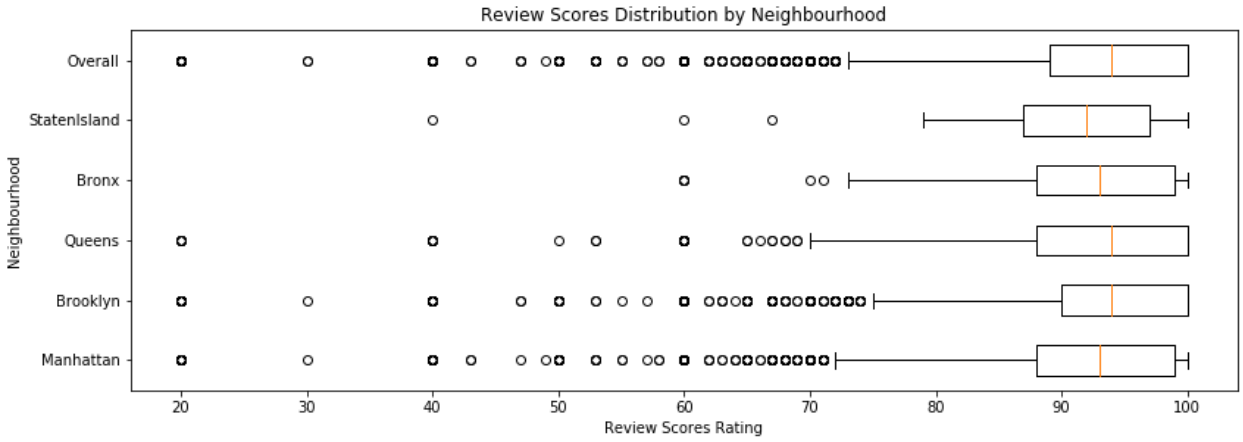

Figure 11: Review scores by Neighbourhood

It looks like that Manhattan and Brooklyn have a wide range of services in terms of quality. Most of the hosts locate in these boroughs have high score, however, a handful of places receive scores below 70. In contrast, Bronx hosts though don't have high scores in general, they have more consistent quality of service, the lowest host receive a score of 60.

Table 5: Summary of review score by Neighbourhoods

Index	Brooklyn	Bronx	Manhattan	Queens	StatenIsland	Overall
count	8487	217	11765	1590	96	22155
mean	92.3635	91.6544	91.8018	91.5491	90.8438	91.9932
std	8.65656	8.16924	8.84036	9.89586	9.13792	8.85037
min	20	60	20	20	40	20
25%	90	88	88	88	87	89
50%	94	93	93	94	92	94
75%	100	99	99	100	97	100
max	100	100	100	100	100	100

Overall, it can be said that the Neighbourhoods are doing quite evenly well.

Next step, we view the scores by Property Type.

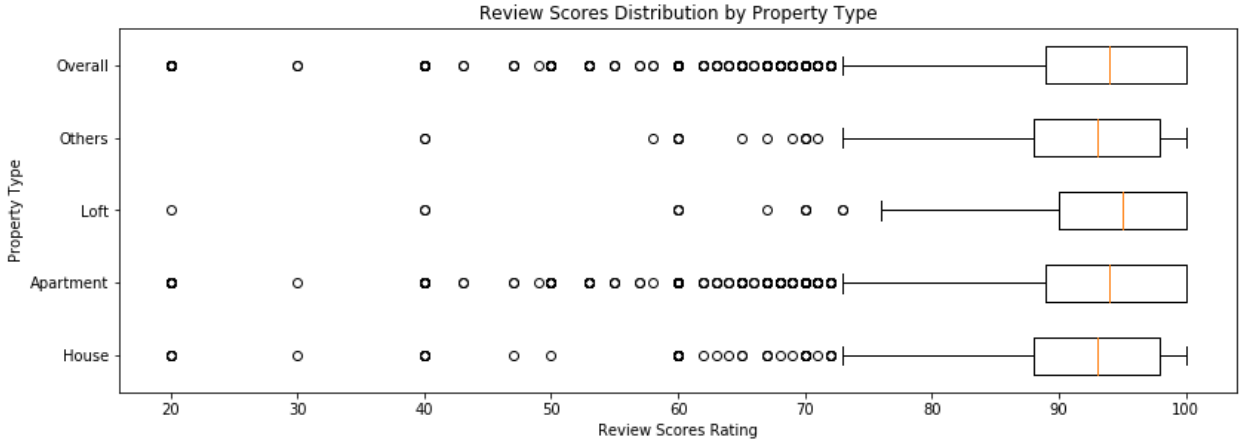

Figure 12: Review scores by Property Type

In the figure 12, Loft hosts seem to have best review scores. Only a few hosts of this type receive scores under 70 and the 3rd quartile is the same as the maximum value, which indicates that 25% of the hosts receive the maximum score. This can also be seen in Apartment, however, there are more Apartment hosts receiving scores under 70. Apartment and House see more scores below 70 compared to other types.

Table 6: Summary of Review scores by Property Type

Index	Apartment	House	Loft	Others	Overall
count	19658	1559	591	347	22155
mean	92.0506	90.8833	93.5076	91.1527	91.9932
std	8.75407	10.1077	7.89032	9.22188	8.85037
min	20	20	20	40	20
25%	89	88	90	88	89
50%	94	93	95	93	94
75%	100	98	100	98	100
max	100	100	100	100	100

The table shows that the scores are similar among the Property Type.

In terms of Room Type, the following figure shows how the scores distribute.

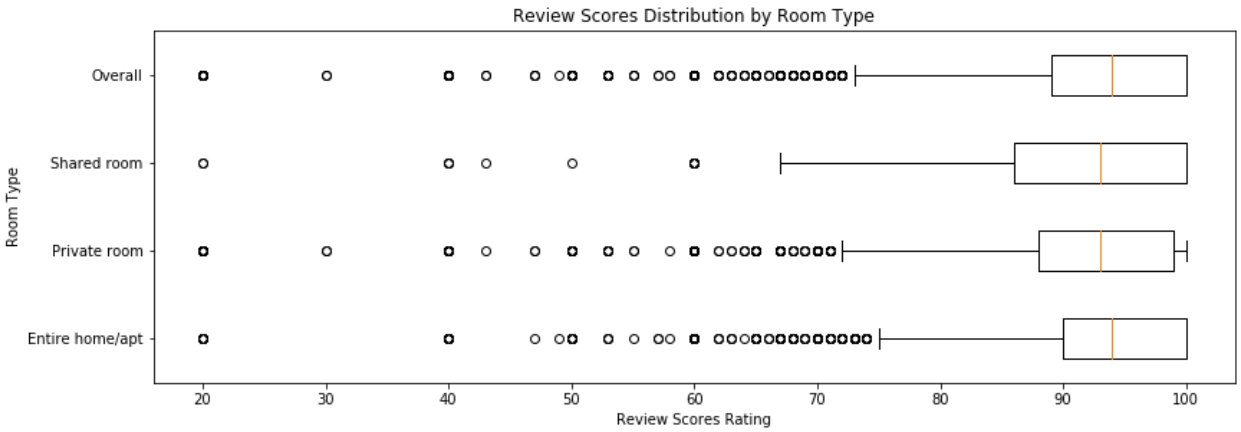

Figure 13: Review scores by Room Type

As can be seen, the Entire home/apt sector receives the most reviews under 70, the Private room sector is next and the Shared room sector receives the least reviews under 70. However if we look at the high score tail, we can see that the Entire home/apt and the Private room have higher concentration around 90 to 100, meanwhile, the scores for Shared room are more spread out.

Table 7: Summary of Review scores by Room Type

Index	Entire home/apt	Private room	Shared room	Overall
count	12884	8729	542	22155
mean	92.4987	91.3326	90.6181	91.9932
std	8.03478	9.78136	10.5687	8.85037
min	20	20	20	20
25%	90	88	86	89
50%	94	93	93	94
75%	100	99	100	100
max	100	100	100	100

Overall, there is not much different in quality service among these sectors.

Question 4: An attempt to build a predictive model for review score using Neighbourhood, Property Type, Room Type and Price.

As mentioned above, there are 8323 unreviewed hosts. This section, we attempt to build a predictive model for review scores given Neighbourhood, Property Type, Room Type and Price.

The motivation of predicting the review scores is that if we were able to predict the review score accurately, we could use the information for recommendation, business consultancy,…

Predicting the exact review score for a host is a challenging problem since rating depends on different factors of the service. In this context, we only have four basic features that could have influence on the

review score. To make it feasible, we decide to split the review scores into two group: Group A – review scores >= 80 and Group B – review score < 80. Group A contains 21027 records (95%) and Group B (5%) contains 1128 records. This indicates imbalanced class issue, one group takes only a small portion of the data. Predicting the review score is now narrowed down to an imbalanced classification problem.

We try various methods and see that Random Forests yield the highest results. To deal with imbalanced, we add class weights {A: 0.048, B: 1} to the model. Other configuration of the model is available in Python code.

Procedure for model evaluation: Train -> Evaluate the goodness of fit -> Test errors

On the training test (80% of the data), the model reports 70.33% of accuracy, which is not so bad. However, we would also like to see how accurate the model performs on each group.

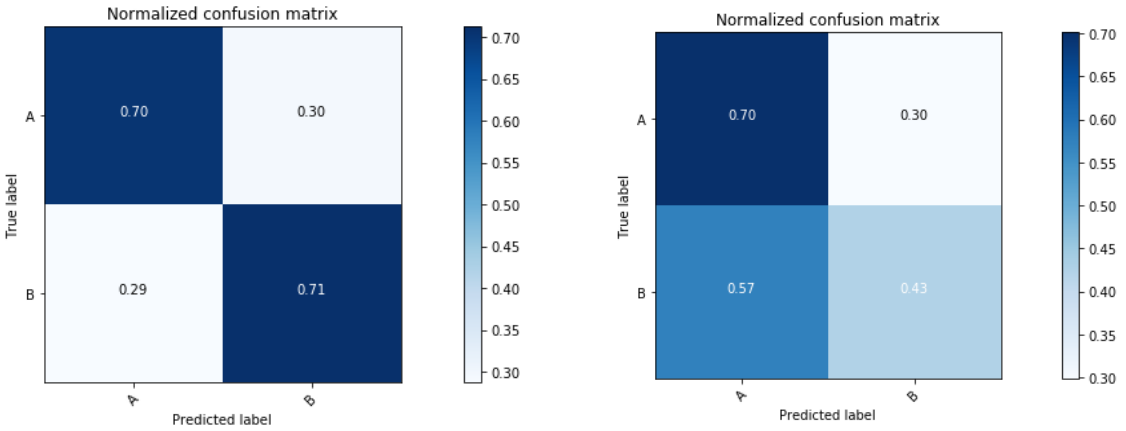

Figure 14: The left panel shows the confusion matrix for the training process and the right panel shows the confusion matrix for the test process.

Notice that the model classifies equally well on the two classes. However, we are really interested in the test process results. The test accuracy is 68.70%. As seen, the true positive rate for Group A remains the same, but reduces significantly for Group B.

To a certain level of acceptance we use this model to predict review group for 8323 unreviewed.

Conclusion

Question 1: How does the price distribute by Neighbourhood, Property Type and Room Type? This means we are interested in knowing some of the more specific questions such as which neighbourhood has the highest/lowest price? Or which property type and room type usually cost more,…

Most price is under 500. However, a handful of hosts have high cost.

Manhattan and Brooklyn have a lot of high-priced hosts. Bronx has the least number of high-priced host.

Most hosts are Apartment type and some have high costs. House-type hosts have the lowest price.

Most room types are of Entire house/apt or Private room. Entire house/apt costs more than Private room. Shared room has the lowest price.

Question 2: Is there a relationship between the factors Neighbourhood, Property Type, Room Type and the price? If so, which factor has the most impact on the price?

There is a significant relationship between the mentioned factors and the price. The model performs better with low-priced hosts.

Neighbourhood and Room type have more impact on the price than Property Type does.

Question 3: Which sector is offering good service?, which mean earning high review scores. The sectors can be viewed in terms of Neighbourhood, Property Type and Room Type.

Overall the sectors are offering good service. Review scores are mostly above 85.

Queens and Brooklyn have the highest review scores. However, Bronx has the highest consistent quality service.

Apartment and House types have high scores, but also contains a lot of low-scored hosts.

Room type receive similar review scores.

Question 4: An attempt to build a predictive model for review score using Neighbourhood, Property Type, Room Type and Price.

Predicting the exact score of the host is narrowed down to an imbalanced classification problem using Random Forests. Results look promising, however, more data and features should be collected to improve the model's performance.