Exploring Airbnb houses prices

Author: Pedro Londono Arbelaez

***Abstract*—This paper studies the analysis and predictions of Airbnb in city of New York. The research answers a number of questions about housing and predicts the price of houses.**

INTRODUCTION

Analysing prices in the city of new York has a meaningful impact in economics and in behaviour on how people decide to leave. The paper describes how the prices of houses behave for different thing such as the are where people is located, or the number of houses in each area group.

Airbnb is one of the largest companies nowadays, what it started as a small company now helps thousands of people is their lives, specially for tourism.

Airbnb business rents houses through their app and allow people all over the word to start for whatever time they wish in exchange of a price.

This dataset was chosen due to the many questions that can be answer and the passion for predicting housing prices.

I. DATA, MOTIVATION AND PLAN

1. Data source

This data source contains different columns about housing data in the city of New York:

- the neighbourhood
- neighbourhood area
- the names of the people who host on airbnb
- Altitude and longitude
- Type of room
- Number of days

The data source was downloaded from Kaggle. This topic was chosen due to how many questions you can find.

Although the dataset has relatively clean data, it was enough to show every phase of the data wrangling skills, from checking nulls to deleting finding outsiders and deleting outsiders. The research was conducted to provide the most accurate information.

II. RESEARCH QUESTIONS

The aim of this study is to answer questions about the population and area of the state of NY about their housing market.

The Research question that were chosen to answer are:

- What are the landlords with the highest number of houses?
- Do the neighbourhood and neighbourhood group have a correlation with the price of rent?
- How many houses are per neighbourhood and neighbourhood group areas?
- What are the types of rooms broken down by neighbourhood group?
- Can we predict which neighbourhood it lies in?
- Can we predict the price of a home based on the area is in?

Although there are many more questions that could be asked, and answer, such as what are the richest landlord on the city of NY, or who was the one who rented more houses in less time.

Those, however are not the topic of this paper.

III. RESEARCH STRATEGY

1. Find the dataset
2. Load the dataset.
3. Look for duplicate rows, null values in columns and outsiders.
4. Clean the dataset, either by filling or dropping the columns or rows that has null values. Also remove outsiders.
5. Search for relationships between data to answer the questions.
6. Plot the data in graphs to find relationships.
7. Find insights from the graphs.
8. Try different methods to find relationships from the data.
9. Apply featuring engineering to see which
10. Build models to predict the price of a home, and the area it lies in.

IV. DATA PREPARATION.

1. CORRELATION BETWEEN COLUMNS

The correlation between columns before analysing price prediction showed the following results.

- Price has a 15% correlation with the calculated_host_listings_count and a 10% correlation with availability.

- The longitud and latitude have a 87% correlation.

- Minimum night have a correlation of 26% with availability_365 and 32% with calculated_host_listings_count

- Number of reviews have a correlation of 16% with availability_365

- id have a correlation of 58% with host_id and 15% with calculated_host_listings_count

1. DATA CLEANING

The data was cleaned by removing outsiders and null values, after checking.

V. FINDINGS AND DISCUSSION

1. NUMBER OF PROPERTIES HOSTED BY LANDLORDS

The analysis was carried after cleaning the data and finding relationships between the different features.

The purpose of this question was to find who were the people who had the larger amount of hosted properties.

This does not mean that they own the properties, neither is meant to that they have the largest number of properties or bedrooms in New York, as some people may not use Airbnb to host the properties or bedrooms.

The study showed the following results:

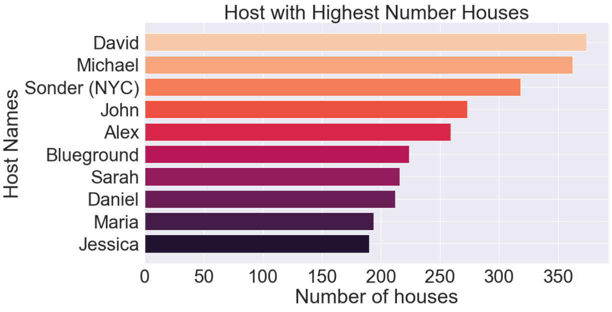


Figure 1: Number of properties listed by host

Figure 1 shows the ten largest number of people that have hosted on Airbnb.

As we can observe from the graph all of them have more than 200 houses, being the largest David listing more than 350 properties or bedrooms.

We can observe further than the range is quite large given the number of people, increasing exponentially.

They have on average 264 combining properties and rooms. This analysis answers our first question on the largest landlords of properties in Airbnb.

2. NUMBER OF PROPERTIES IN NEIGHBOURHOOD AREA AND NEIGHBOURHOOD

The study on data showed that there were 5 neighbourhood areas. Brooklyn, Manhattan, Queens, State Island and Bronx.

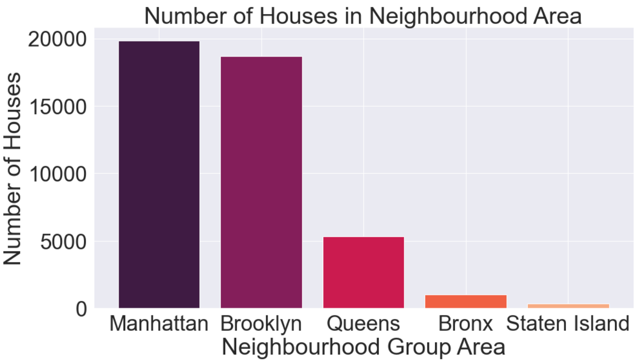


Figure 2: Number of properties listed by neighbourhood area.

Table 1	Number of properties listed by area	
	Neighbourhood Area	***Number***
	Manhattan	19831
	Brooklyn	18693
	Queens	5315
	Bronx	1021
	Staten Island	344

Figure 2 and Table 1 shows the results obtain containing the number of properties listed by area. We can observe that Manhattan is the neighbourhood area having largest number of listings in Airbnb.

This by no means shows that Manhattan has more houses than other areas in total, followed by Brooklyn and Queens, and lastly Bronx and Staten Island. This figure only shows that Manhattan has the largest from Airbnb which is a sample from the population, i.e. Brooklyn or any other district could have a larger proportion of houses than Manhattan, however there are less listed on Airbnb.

The Analysis showed the following results studying the number of properties in neighbourhood.

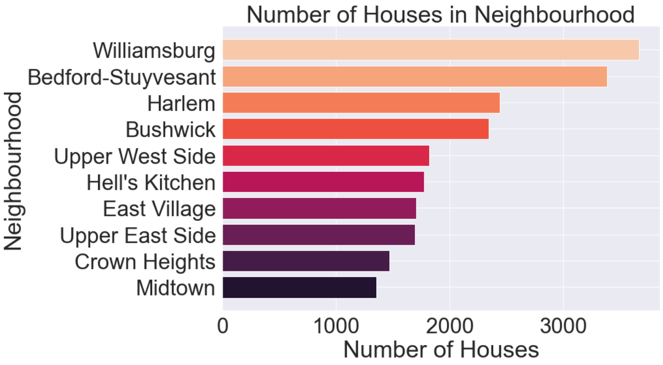


Figure 3: Number of properties listed by neighbourhood.

Figure 3 shows the ten neighbourhoods having the largest number of listings in Airbnb.

We can observe that Williamsburg is the neighbourhood having largest number of listings in Airbnb. Followed by Bedford, Harlem, Bushwick and the rest. We find similar that there is a range that triples the properties listed in the Midtown are. Furthermore, 5 out of the ten neighbourhoods belong to the neighbourhood area of Manhattan.

This answers our second question about number of houses in each area neighbourhood and neighbourhood. We can find that the results shown in Figure 3 compose the results of Figure 2.

3. WHAT IS THE TYPES OF ROOMS BROKEN DOWN BY NEIGHBOURHOOD GROUP

The study showed that there are less people renting shared rooms, we can observe that in Manhattan there are more expensive prices than in other areas in all room types. We can observe that there are shared rooms located in Manhattan that are more expensive than entire homes and private rooms in other areas. We started with the initial hypothesis that entire homes should be more expensive than and private rooms should be more expensive than share rooms.

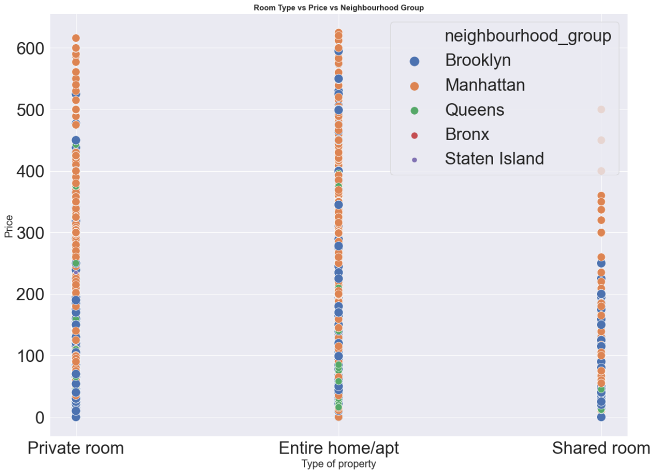


Figure 4: Room Type vs Price vs Neighbourhood

Figure 3 shows that we were correct. However the area where it is located has a big effect on the price. There is a higher supply of entire homes in Airbnb compare to private rooms and shared rooms.

4. CORRELATION BETWEEN NEIGHBOURHOODS AND NEIGHBOURHOOD AREA AND PRICE

As we can observe from Figure 3, this gives us a good visualisation of how the types of property rented differs between areas and between price, we can find the following conclusions:

1. The price generally depends of the type of property being rented.

We can observe that share rooms are less expensive than private rooms and entire home/apt. Another interesting thing is that an entire home/apt could be as expensive as a private room, as the number of points reaching 600 is slightly higher in entire home/apt than private room.

2. The price generally depends of the area of the property is located in.

We can observe that in areas such as Manhattan or Brooklyn, the price is much higher, even a share room could be more expensive than entire home/apt on other areas such as Queens or Staten Island.

We can also see that in the case of Manhattan there are more houses in a price higher than in other areas therefore the average price is higher. From the research we can conclude there is a correlation between the price of rent and the area where the property is located in.

Further research was done to see a different picture of data, The following figure compares how the number of reviews change over the neighbourhood group and the price of bedrooms.

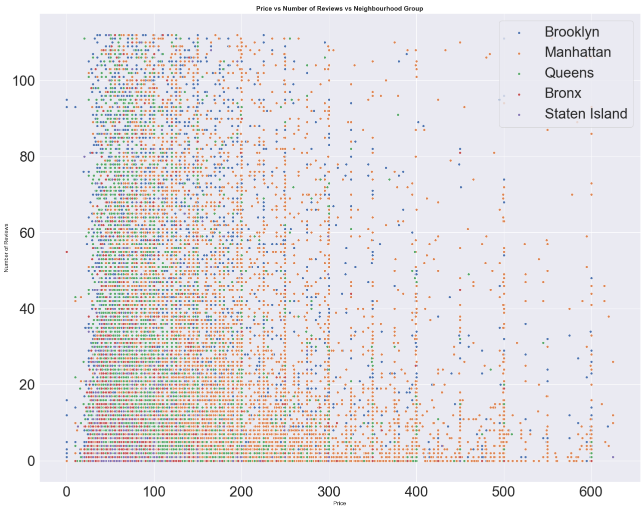


Figure 5:Price vs Number of Reviews vs Neighbourhood

As we can observe, we can see similar behaviour found in the price vs type of property vs Neighbourhood Area.

We can found the following that

- As the price increases number of reviews decrease, therefore less people give reviews, also, the people who give the most reviews after 200 are generally from Manhattan. There are lower reviews given by people living in Manhattan when the price is less than 150.

- Queens, Staten Island and Bronx reviews are mostly given when the price is less than 200, and there are a higher number of reviews given and from more diverse background. Queens has a higher number of reviews in that price range than other areas.

- Brooklyn State shows that the reviews of that state are less dependent on price, We can see that on that people can find expensive locations as well as cheaper locations.

VI. PRICE PREDICTION

All the research done previously has been used to determine which columns and which data would be used to predict the price, and get the different distributions of the features. The model used was linear regression which was not good enough results with the given data. To find the best parameters we used grid search and we used K-Fold cross-validation. This was accurate though as before the analysis the data showed little correlation between numerical features. The data was normalised

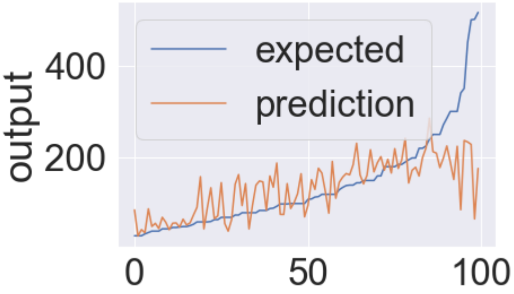


Figure 6: Room Type vs Price vs Neighbourhood

As we can observe, the model starts fitting the data, but it does not behave well to predict accurately. Using other model such as polynomial regression would be more appropriate. The results are shown by table 2:

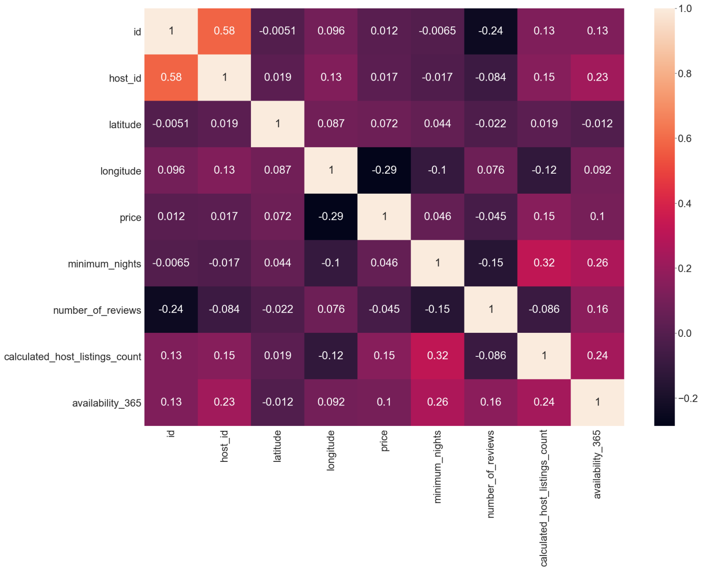
Table 2

Metrix	***Number***
MAE	51.52
MSE	6078.302
RMSE	77.96
R2	0.482

Polinomial regression was attempted, however it was taking to long to make the calculations.

Conclusion

To conclude, we can confirm that the price depends of the area of residence, although the price is dependent on the type of room rented. The price had no correlation at the beginning when we plotted a correlation matrix of all the columns.



As we can observe from the correlation matrix, there is not much correlation between numerical columns. However the area is not considered a numerical column, and it has been excluded, in the prediction was encoded using one-hot encoding.

References

[1] Data Wrangling in Python by Wes McKinnie

[2] Python Documentation, Pandas, Numpy and matplotlib

[3] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html

[4] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[5] https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.concat.html

[6] https://seaborn.pydata.org

Identify applicable funding agency here. If none, delete this text box.