An investigation into Air BnBs prices into New York

Rosany Antonyvincent

Graduate data scientist - technical assessment
ARM

October 2020

Table of contents

Chapter 1

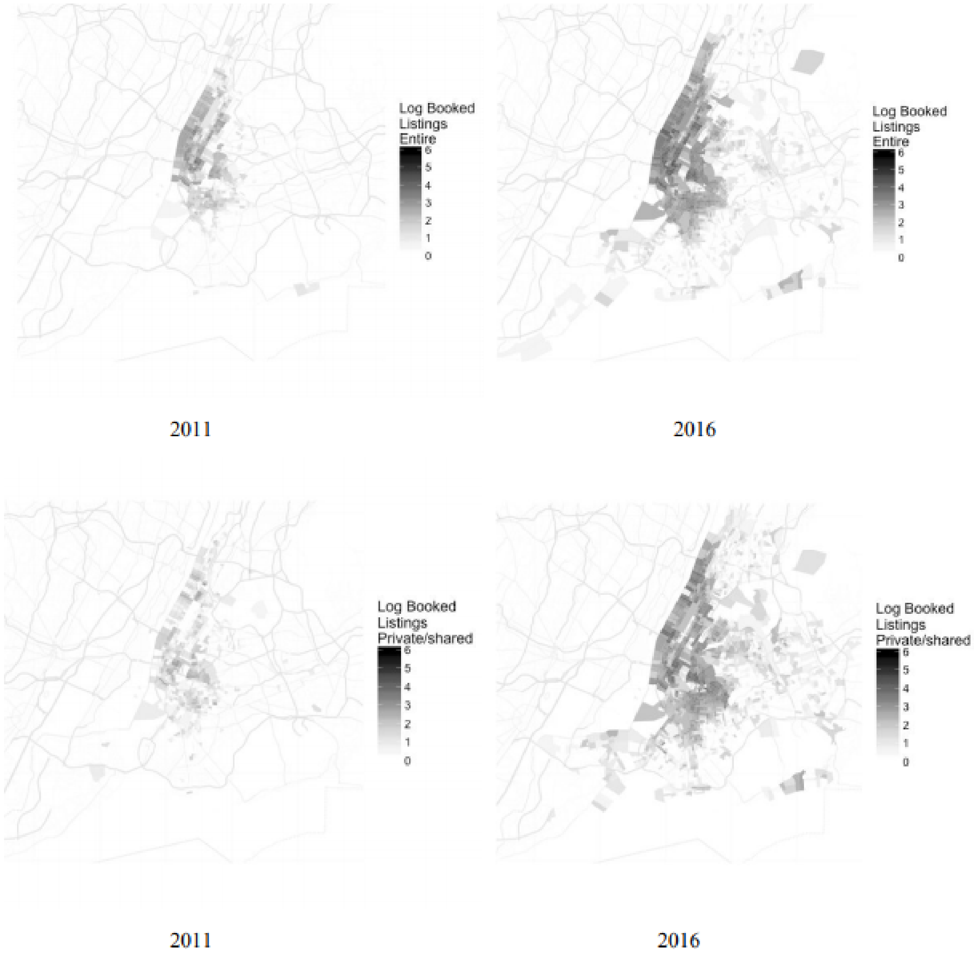
Introduction

1.1 Background

From 2008, Air BnBs opened a new avenue into the travel industry, offering a more personalised experience to exploring new locations. Usually Air BnBs work out cheaper than hotels, and therefore this has had a big impact on the hotel industry. Renting Air BnBs have now become the cheaper, more convenient option for visiting New York [1]. Air BnB has made the process even simpler. It offers over 50,000 listings for New York City alone [2].

The ease of both creating extra income for home owners and for cheaper accommodation for travellers has allowed for a lot of competition. One point to note is that Air BnBs generally have less flexibility than hotels, varying amenities and some accounts are potentially untrustworthy. Nonetheless, Air BnB has had a big impact on the travel industry [3][4].

Air BnBs have also come into the spotlight due to many illegal listings [5]. Most listings are in 3 high profile locations in Manhattan. This is disrupting the spread of wealth as a small handful of landlords are offering up entire homes or apartments. The density increase in listings can be seen in 1.1. BY 2016, AirBnB listings were considerably more geographically dispersed than hotels. An issue that has arisen from this increase in properties is tax evasion which has been under consistent investigation since 2013 [6].



2011

2016

2011

2016

Fig. 1.1 Figure showing density increase in listings in NYC

1.2 Problem

With Air BnBs on the rise, the competition has increased massively. The rising media attention these properties are receiving for illegal practise is also of concern and therefore this can be analysed within the data. To this end, it's necessary to analyse what kind of hostels/rooms are selling and what features are attributed to the price.

The aim of this project is to evaluate whether there is a correlation between different features on the price of the BnB. If there is a correlation, then the features can be evaluated individually and predictive modelling will be used to evaluate feature significance.

Chapter 2

Data acquisition and cleaning

2.1 Data source

The labelled data used is a public dataset part of Airbnb on Kaggle pulled from the original website. The dependent variable being explored is the price. Features explored will be:

- neighbourhoodgroup
- neighbourhood
- longitude
- latitude
- lastreview - date of last review left
- hostlistingscount - number of listings for host
- price - price of BnB per night

To this end, all other features can be dropped from the dataframe to improve runtime.

The packages used for this project were mostly pandas, numpy, sklearn and seaborn [7].

2.2 Cleaning for modelling

Missing rows are removed from each dataframe created for analysis. Equally, the year was extracted from lastreview using strftime.

Different cleaning was used depending on whether the feature was categorical or continuous. Continuous data only required for rows with missing values to be removed. One hot encoding was used for categorical features.

2.2.1 One Hot Encoding

One hot encoding is useful in machine learning by transforming categorical variables into Boolean dummy variables consisting of 0s and 1s. The features evaluated were neighbourhoodgroup and roomtype. These were transformed into Boolean columns to be used for predictive modelling [8]. The original features were then removed from the dataframe.

Chapter 3

Exploratory Data Analysis

The dataset allows for a lot of analysis. Firstly, the neighbourhood with the most listings was evaluated. This simply required using pandas groupby function then taking the neighbourhood with the highest value. This showed Williamsburg had the most listings.

Having evaluated the neighbourhood with the most listings, it is interesting to note what the average number of listings per host is across time. This required the time the listing was made available, and the number of listings per host (hostid). The year from the lastreview variable has been used for the year indicator. This however is unreliable given that the last review does not necessarily mean year of BnB use or when the BnB was made open to the public.

Equally, the dataset included rows where if the hostid was the same per id (unique), the number of listings were the same regardless of date of review. Therefore, the data was organised in date order, and duplicate host*id*s after appearing were removed. In this instance, the hostid count would not be included multiple times.

A quick analysis is taken of the dataframe to look at the mean, which came out as 5.16, rounded to 5 listings per host id. In order to find the listing count across time, a for loop is setup to find the average listing count per year. Importantly, the average of the previous years in included year on year, where the assumption that any properties that existed the year before, still exist in the current year.

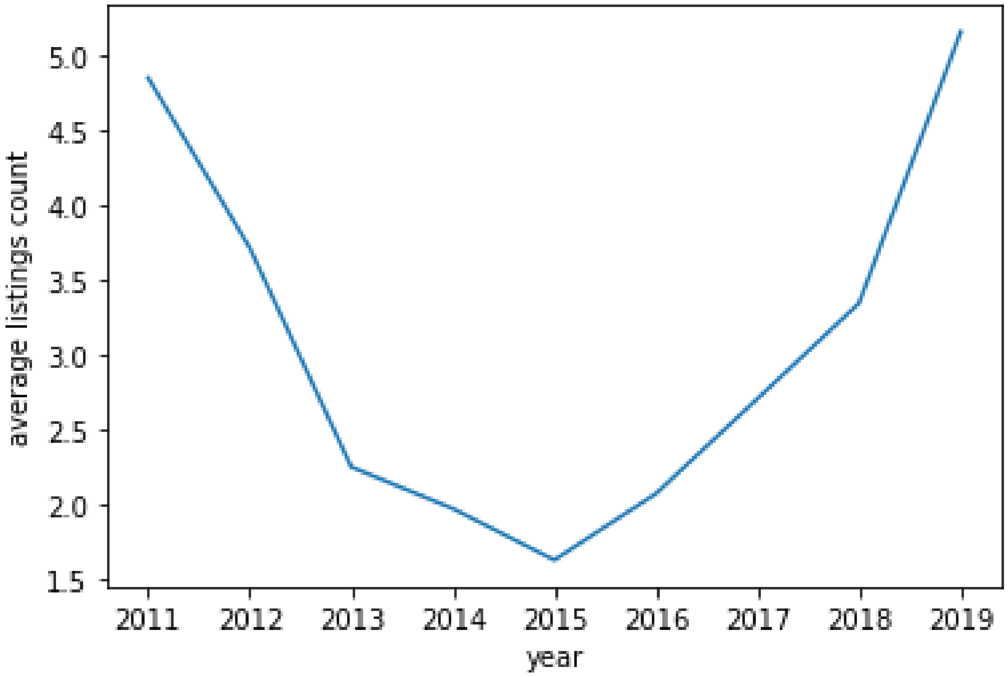


Fig. 3.1 Graph to show how the average listings count varies with time.

As can be seen in 3.1, the average listings count has roughly a quadratic relationship with time. The dip in 2015 could be due to bigger properties being let and therefore more rooms but fewer listings [9].

Having established how the listings vary across time, it's interesting to also look at how the price varies across the neighbourhood groups across time.

As can be seen in 3.2, generally Manhattan is the most expensive to stay in on average, including the most recent 2019. Equally, the most expensive year and neighbourhood group was 2015 Staten Island. This was to be expecting as 2015 saw an increase in larger properties, as opposed to single or share rooms. The price per night seems to be increasing for Bronx and Staten Island on average across time whereas for Brooklyn, Manhattan and Queens, the price seems to be decreasing steadily.

The most common type of accommodation across these properties is an entire home/apartment, with 25409. This is followed by private rooms with a count of 22326. A shared room is the least common, with a count of 1160. This suggests that people who come to New York, generally come in bigger groups as opposed to on their own or as a pair.

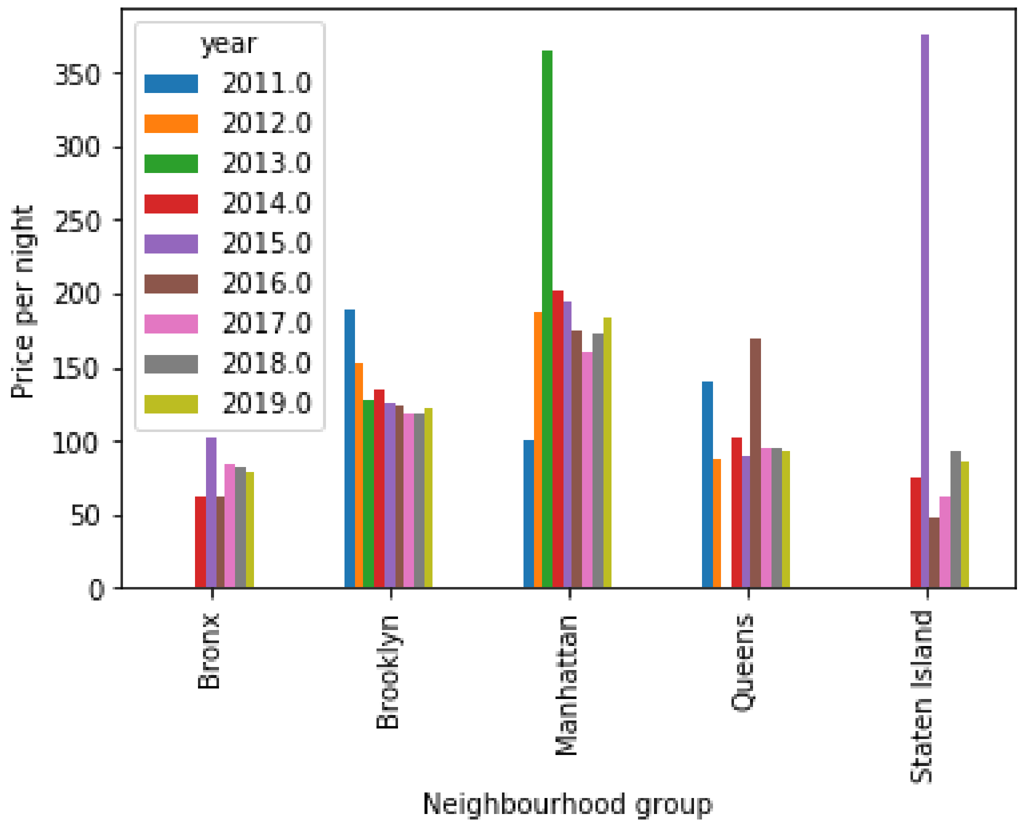


Fig. 3.2 Graph to show how the price of neighbourhoods varies across time.

Having analysed the price of neighbourhoods, it is useful to know the cheapest place to stay in terms of neighbourhood and type of accommodation. In 2019, the cheapest place was a shared room in Randall Manor, Staten Island.

3.1 Price prediction

The price of a night on average in NY can be evaluated using this data. By grouping by year and taking the mean of the prices, a plot can be produced of price per night across year.

As can be seen in 3.3, the price fluctuates across time. There is a noticeable peak in 2013. This could be due to the rise in additional admin fees across all AirBnB sites [10]. There are many factors that can influence the price per night on average across time such as the aviation industry, economy, weather and natural disasters. It is difficult therefore form a 2020 prediction as these factors are not accounted for in the data. At best, the last two points for 2018 and 2019 could be used to form a prediction but this would be a very weak hypothesis given the affect of coronavirus is unknown but only assumed to have caused a very big reduction in price per night on average.

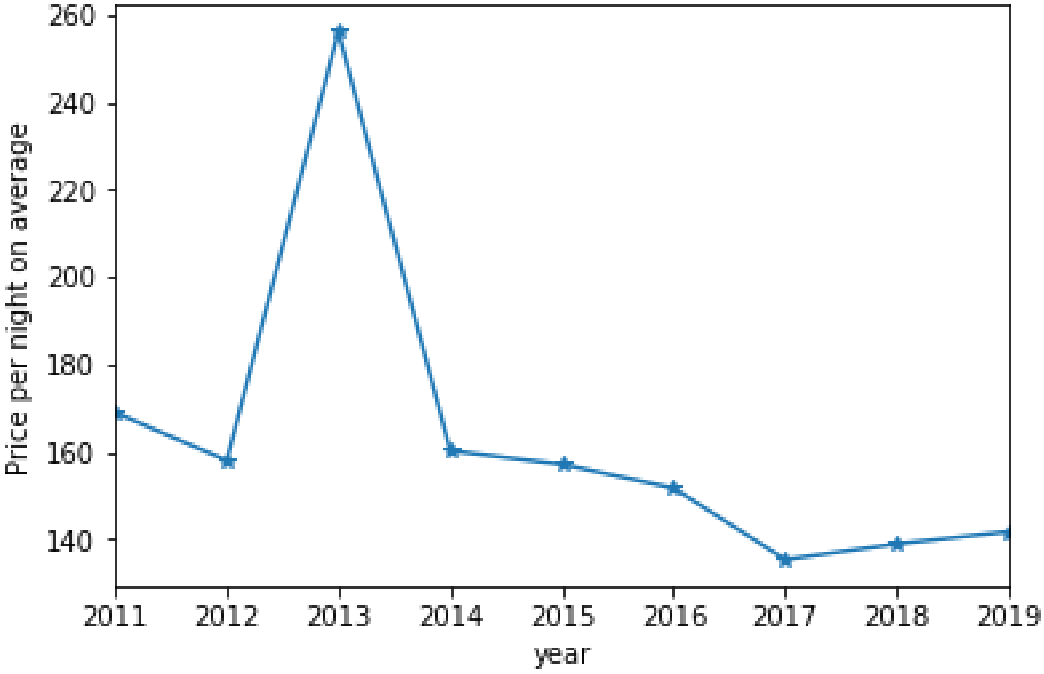


Fig. 3.3 Graph to show how the average price per night has varied across time.

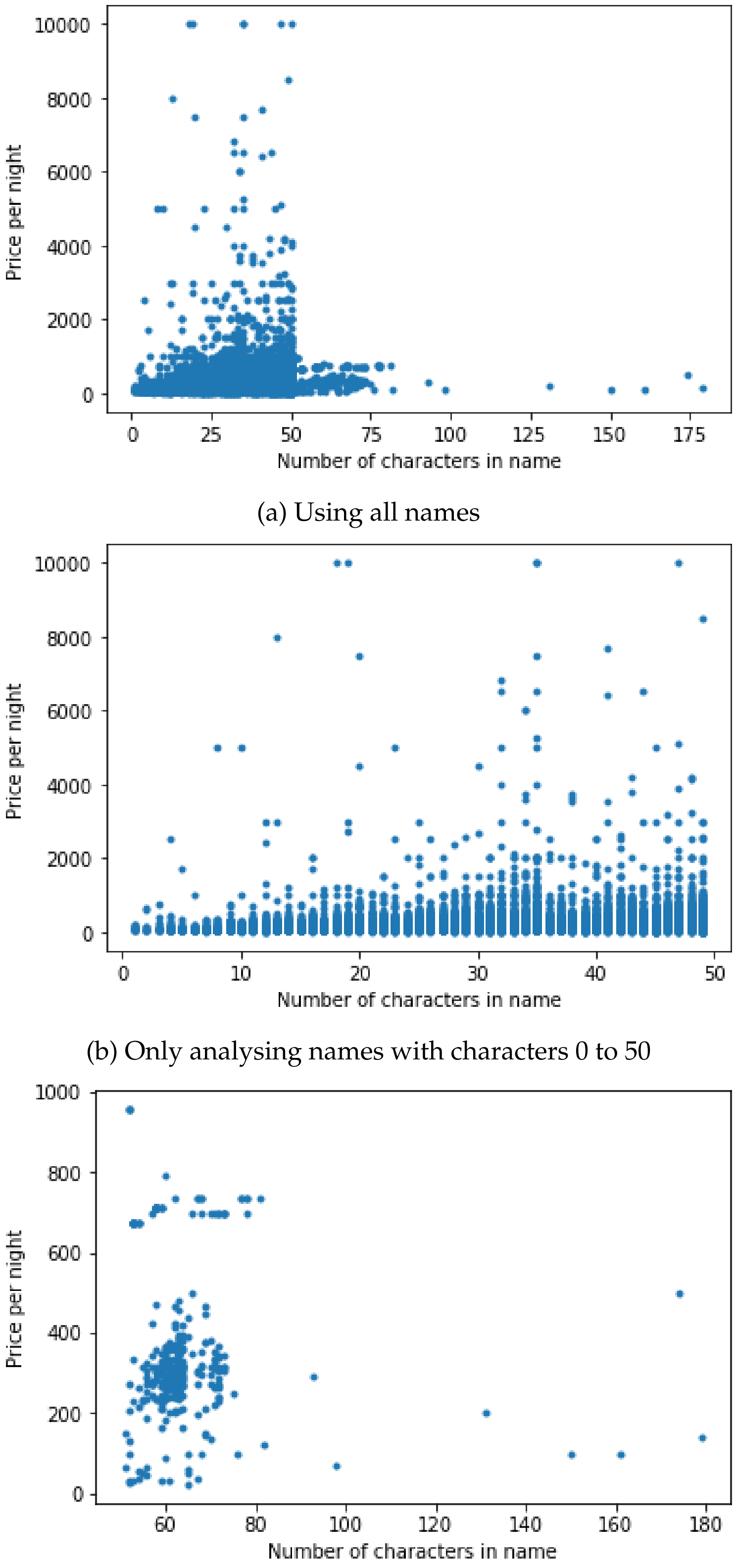
3.2 Characters in name analysis

Given that the dataset includes the names of each BnB listing, it is possible to analyse whether the number of characters in the name has any impact on the price. In 3.4a, the variance of price per night against characters in name can be seen. Noticeably, the price significantly drops at 50 characters. This has then been analysed into characters between 0 and 50, and 50 onwards. There is a large difference in price between names that are less than 50 characters and names that are more.

In order to evaluate whether there is any statistical significance between the two variables, Spearman correlation will be used as opposed to Pearson since figure 3.4a doesn't show a linear relationship but does seem to show a monotonic relationship. The Spearman correlation returned 0.04, 0.07 and -0.07 respectively. For a good indicator of statistical significance, the Spearman correlation should be 0.7 or greater. Therefore we can conclude that these two variables are not statistically significant.

3.3 Statue of Liberty visit analysis

Another property that can be evaluated is the nearest house by price and number of reviews. The assumption made is that a person can travel in straight lines. If for example, someone would like to visit the Statue of Liberty, where would the closest entire house be which is less than $300 but also has a high maximum number of reviews per month? This analysis requires the absolute difference between longitude and latitude coordinates and settings the conditions to "room type == "Entire home/apt" and "price < 300" then sorting the dataframe by minimum absolute distance and highest reviews per month. This results in *Spacious neo-country seaside loft* being the optimum BnB.



(a) Using all names

(b) Only analysing names with characters 0 to 50

(c) Analysing names with characters above 50

Fig. 3.4 Graphs to show how the price per night varies with number of characters in name.

Chapter 4

Predictive modelling

In order to apply predictive modelling, first one needs to identify which features need to be used. As stated in chapter 2, either continuous or categorical variables can be used.

The predictive modelling used will be:

- K Nearest Neighbours - K-Nearest Neighbours classifies a new data point, based on the majority vote of the classes of it's nearest neighbours. The K value describes how many nearest neighbours should be taken into account in order to classify a new data point. The training process involves simply finding what value of K allows for the highest classification score on the training data.
- Decision Tree - Within a decision tree, a series of Boolean questions are used to classify data points. The training process involves fine tuning these questions, in order to maximize the accuracy of the classification on the training data set. Finally, data that is previously unseen by the model is classified.
- Support Vector Machine - Support vector machine classifies cases by finding a decision boundary in feature space that splits the two classes. The training process involves finding the best location for this separator.

The accuracy will be measured using:

- Jaccuard index - Measures similarity of predicted values and true values
- f1-score - Measures accuracy based on true/false positives and true/false negatives

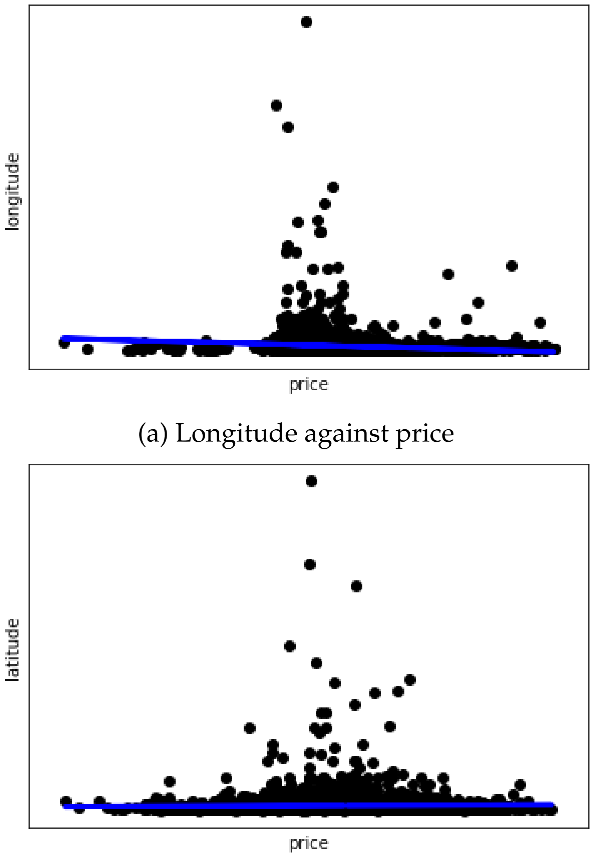
Jaccuard index and f1-score.

4.1 Continuous variables

The longitude and latitude variables are a good start to modelling since predictive modelling requires continuous variables. However, it is important to analyse these variables together since they don't give much information individually.

4.1.1 Linear regression

A very simple analysis to begin with would be linear regression. As an example, when performing linear regression on longitude and latitude against price individually, using a train/test ratio of 0.85/0.15.



(a) Longitude against price

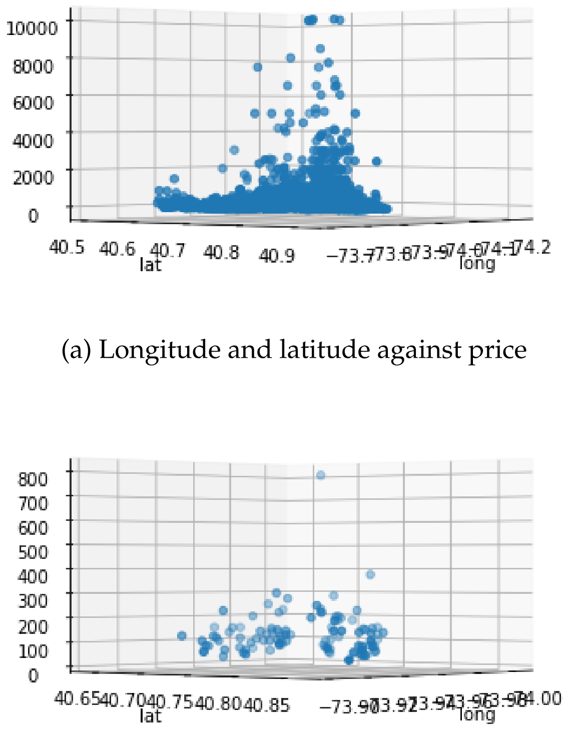
(b) Latitude against price

Fig. 4.1 Graphs to show a linear regression analysis over longitude and latitude features.

As can be seen in 4.2, individually, they do not hold a linear relationship with price.

4.1.2 Longitude-latitude combination

A scatterplot can be used first to view the price against longitude and latitude.



(a) Longitude and latitude against price

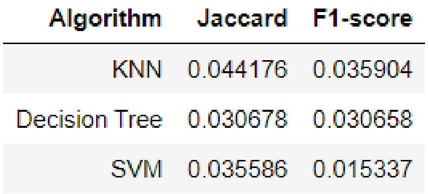
(b) First 100 longitude and latitude rows against price

Fig. 4.2 Graphs to show combined effect of longitude and latitude on price.

These scatterplots aren't indicative of a relationship between longitude, latitude and price for predictive modelling as most data is clumped and the dataset is very large. Even when taking only the first 100 rows as in 4.2b, the graph doesn't suggest a relationship and thus a multiple linear regression model is not necessary. In this instance, categorical variables may be more useful for predictive modelling.

4.1.3 Prediction results & discussion

Since longitude and latitude aren't good variables to act as predictors, the results of predictive modelling are very low, as expected 4.3. All Jaccard index values are less than 0.04, suggesting the test and train data are not similar at all. This is likely due to longitude and latitude being a very poor indicator of price as the price is affected by many different factors. In this case, the dataset hasn't provided enough information with respect to the location, people's preferences.



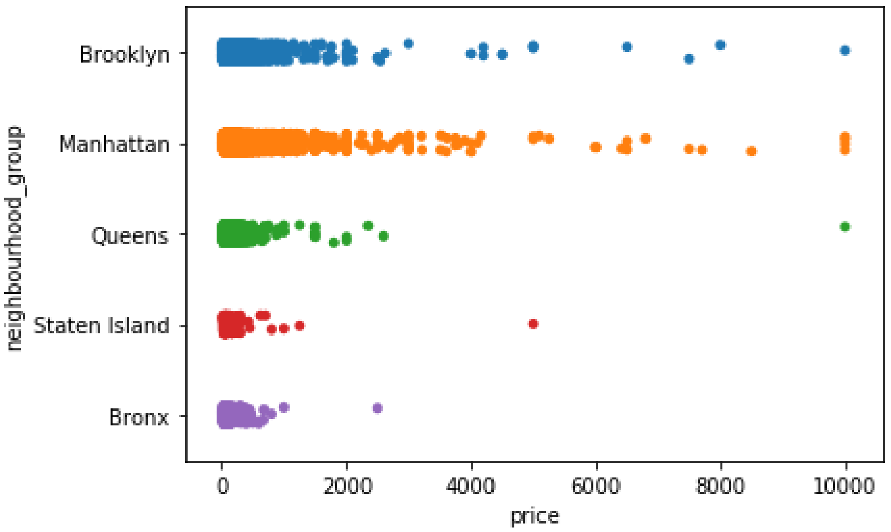
Algorithm	Jaccard	F1-score
KNN	0.044176	0.035904
Decision Tree	0.030678	0.030658
SVM	0.035586	0.015337

Fig. 4.3 Table showing the accuracy of different means of classification.

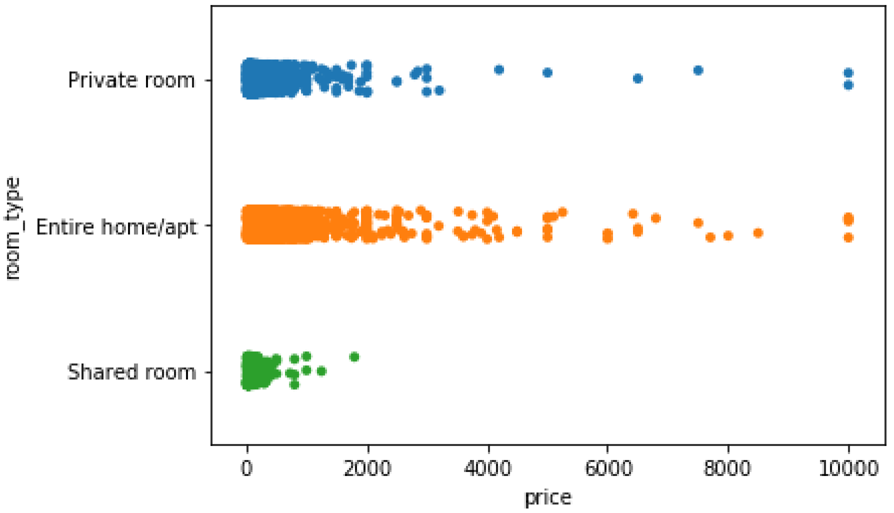
4.2 Categorical variables

4.2.1

The variables being analysed will be the neighbourhood group and room type. A stripplot is a convenient way to analyse if the input variables are going to make good predictors. What is expected is for data points to be grouped against different groups within the variable.



(a) Neighbourhood groups against price

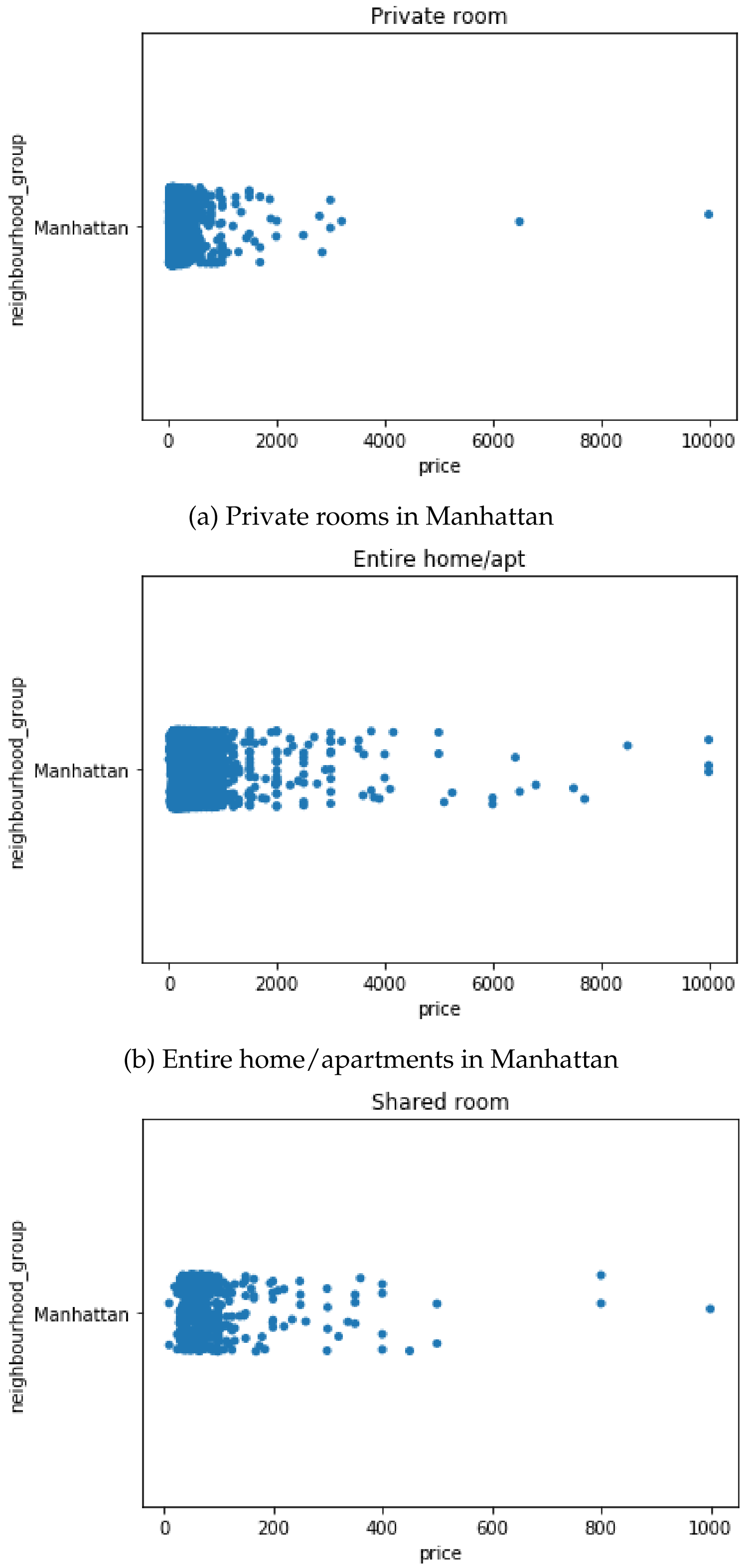


(b) Room type against price

Fig. 4.4 Graphs to show a stripplot of groups within variables.

As can be seen in 4.4, the neighbourhood group and room type groups will not make good predictors since the price stays consistently lumped for all groups. A good predictor would have clumped prices for each group, however what is evident is that the price does not vary depending on neighbourhood group and room type.

A deeper analysis into Manhattan and room type can be made. As can be seen in 4.5, the results are similar to that of 4.4 in that the groupings are not indicative of belonging to certain price ranges.



(a) Private rooms in Manhattan

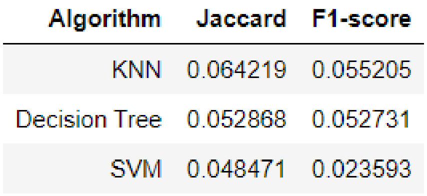
(b) Entire home/apartments in Manhattan

(c) Shared rooms in Manhattan

Fig. 4.5 Graphs to show a stripplot of room types in Manhattan.

4.2.2 Prediction results & discussion

As expected, the Jaccard index and f1-score are very low suggesting very little similarity between the test and train data. Although these values have improved from the continuous variables data, that is expected due to increased variables.



Algorithm	Jaccard	F1-score
KNN	0.064219	0.055205
Decision Tree	0.052868	0.052731
SVM	0.048471	0.023593

Fig. 4.6 Table showing the accuracy of different means of classification using neighbourhood group, room type, minimum night, longitude and latitude features.

Chapter 5

Conclusion

This report has explored the price of Air BnBs in New York from 2011 to 2019 under different conditions. What is noticeable is the increase entire homes being provided to consumers, with price per night on average staying fairly consistent (excluding 2013). The predictive modelling for both continuous and categorical variables have massively overfit the test data. The predictive modelling however has not been successful in this instance. With increased time, the price can be put into bands to reduce the dataframe size and thus memory. Having done this, the decision tree would make a good predictive model using the neighbourhood group and room type. Given the large range of price, the test and train datasets are skewed. In this instance, more data is important. Another method for data cleaning would be to replace missing values with the mean value in the same neighbourhood. This should result in more consistent data and could potentially provide a linear relationship with a combined longitude and latitude.

5.1 Future improvements

A new potential dataframe to analyse price, could be the reviews left on each property. This would utilise word count, type of words, positive or negative content, and amount of reviews left. In this instance, clustering would be a better method of predictive modelling.

References

[1] Hospitality research: Airbnb's impact on hotels. *EHL Insights*, Jan 2018.

[2] A federal judge blocked new york's latest attempt to crack down on airbnb. *Shared Economy Tax*, Jan 2019.

[3] How to navigate the nyc airbnb law. *Shared Economy Tax*, Dec 2019.

[4] Hotels vs. airbnb for new york city visitors. *Investopedia*, Jan 2020.

[5] David Streitfeld. Airbnb listings mostly illegal, new york state contends. *New York Times*, 15, 2014.

[6] Roberta A Kaplan and Michael L Nadler. Airbnb: A case study in occupancy regulation and taxation. *U. Chi. L. Rev. Dialogue*, 82:103, 2015.

[7] Jake VanderPlas. *Python data science handbook: Essential tools for working with data*. " O'Reilly Media, Inc.", 2016.

[8] Cedric Seger. An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing, 2018.

[9] Jun. New york city council to consider regulation of airbnb sharing economy. *The Guardian*, 2015.

[10] Airbnb: disruptive innovation and the rise of an informal tourism accommodation sector. *Current Issues in Tourism*, May 2015.