**Clearing the air: understanding the factors that influences the number of bookings an AirBnB receives – compiled from AirBnB listings in New York City**

**INTRODUCTION**

In the past decade, AirBnB has exploded onto the global market and dominated the new digital frontier as a major player. It’s now one of the staple businesses in the ‘sharing economy’, a marketplace where retail investors and mum and dads looking for a side income can enter a market that was traditionally reserved for those with large capital, and use assets they already have to create their own revenue.

AirBnB provides the means for a customer who owns property to lease out a room or entire hom to lodgers. An AirBnB customer, or host, can set the terms for their listing: the minimum number of nights a lodger has to stay, price per night, rooms in the property etc. providing greater flexibility and freedom in how they run their business.

The impact of AirBnb has been monumental. In 2015-16, AirBnB was believed to have contributed $1.6 billion to the Australian GDP alone. The increasing popularity of AirBnB has led to real estate in many cities being swooped up and used as AirBnBs, driving up both real estate prices and competition from other AirBnBs.

As such establishing an AirBnB and ensuring its success can be fraught with risks due to both the increased competition and the greater capital a person has to put down to purchase real estate. It is now more paramount than before to ensure every advantage is taken to safeguard the success of an AirBnB. Therefore it would be vital to know what elements contributes to an AirBnB’s success, and how an AirBnB host can control these elements when setting up an AirBnB such as property location, the minimum number of nights a lodger is required to stay, the price per night and more.

By understanding what impact these factors might have on the rate of bookings brought in, AirBnB hosts can be prepared in entering the market and protecting their investment. As such an analysis that looks at the factors that influences the number of bookings an AirBnB receives would be incredibly helpful for both interested and experienced AirBnB hosts.

**DATA**

The primary data used was lifted from a dataset of AirBnB listings in New York City during 2019, compiled by user Dgomovon on Kaggle.com. The data was pulled directly from listings on the AirBnB website that were listed for New York City and advertised during 2019. Therefore, the data consists of the features seen in an AirBnB listing such as the name of the listings, the listing ID, the name of the host, the host ID, the borough it was located in, the specific neighborhood it was in, it’s address, the latitude and longitude, room type, price per night, minimum nights required to stay, number of reviews it has received, date of the review and other miscellaneous details.

How many bookings a listing receives is not publicly available so instead the number of reviews a listing has will be used as a proxy for the number of bookings a listing received, given that to provide a review, a reviewer had to have first booked and stayed at the AirBnB. The features found in a listing will also be used to assess how they influence the number of reviews a listing receives.

To supplement that data, data will also be taken from the FourSquare API which provides geographical and location data based on a specific area. It would be interesting to see how the close proximity of amenities could influence the booking rate of an AirBnB listing. This could help determine where property should be purchased to set up an AirBnB.

New York City is globally renowned for its museums, with many famous cultural institutions often right next to a museum. As such an analysis would be conducted to see how the proximity of an AirBnB listing near a museum could influence its success. To accomplish this, the FourSquare API will be used to pull the location of museums throughout New York City and this will be analyzed in conjunction with the geographical locations of AirBnB listings.

**METHODOLOGY**

The analysis was done in Python, using modules such as numpy, pandas, SciKit Learn and Folium. The dataset was loaded in as a csv and saved onto a data frame. From there it was analysed and explored.

First the variables in the dataset were explored to see what they measured, what data type they possessed and to build a general understanding of the data. Descriptive statistics were then run to understand the distribution of the data.

Upon examining the descriptive statistics, it was determined some data cleaning should be conducted. Firstly, certain listings had a price of zero listed. These listings were removed from the dataset. Listings that also had a price greater than $500 a night was removed to provide a better fit for the statistical models and provide a stronger generalizability to any results were found. In addition, listings that required a minimum stay of 31 nights or more, which is lodgers were required to stay for a minimum of a month, were also removed to provide a stronger statistical fit and better generalizability.

A boxplot was created exploring the distribution of price of listings and how they varied between different boroughs. Another boxplot was also created examining the distribution of number of reviews between different boroughs.

A scatterplot was created exploring the relationship between price per night of a listing and the minimum number of nights required.

A map of New York City was then created using Folium and the geopy library with the geographical data of the AirBnBs within the dataset, to help visualize the spread of AirBnBs in New York City. Due to the computational load required, only 100 AirBnBs were visualized in the map.

The FourSquare API was used to find museums within New York City, with this data then published into a data frame. The museum data was then appended to the New York map, to provide a visualization of both AirBnBs and museums within New York.

Given the three most popular museums in New York, the Museum of Modern Art, Museum of Natural History and the Guggenheim were all located in Central Park, it was determined that the distance of Central Park from an AirBnb should be considered as a factor that could influence the number of reviews and ergo the number of bookings a listing receives.

The distance of each AirBnB from the geographical coordinates of Central Park was calculated and inputted into the data frame.

A boxplot was constructed to measure the distribution of distances from Central Park for the AirBnBs. Two scatterplots were constructed, exploring the relationships between the price of a listing and distance from Central Park, and the number of reviews from a listing and Central Park.

Two additional scatterplots were also constructed examining the relationship between the price of a listing and the number of reviews, and the minimum number of nights required to stay and the number of reviews received.

A linear regression model was built examining the relationship between the number of reviews an AirBnB had received and its distance from Central Park. The data was split into two subsets of training and testing data, with 20% of the data allocated for testing. The model was trained and tested. Actual and predicted values from the model was used for evaluation. This was also reinforced with the metrics of Mean Absolute Error (MAE), Mean Squared Error (MSE) and the Root Mean Squared Error (RMSE) for evaluation.

To determine if the relationship between the number of reviews received and the distance from Central Park could be polynomial, a polynomial regression model was also constructed, iterating through different degrees until the best fit was found. Evaluation metrics of RMSE and the Pearson correlation was also used. A scatterplot was constructed to visualize the line of best fit for the model.

A multiple linear regression model was built examining the effect that the price of a listing and the minimum number of nights is required to stay could have no the number of reviews a listing receives. Due to the results found for distance from Central Park, it was decided they would be omitted from the multiple linear regression model.

The data was split, trained and tested with a split of 20% testing data. It was evaluated with MAE, MSE and RMSE and comparisons of actual against predicted values. An iterative polynomial model was also applied to the multiple linear regression to see if a polynomial relationship existed.

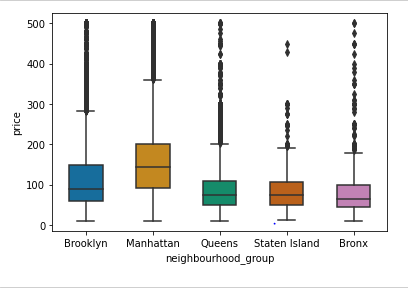
From the map of museums and AirBnBs in New York constructed earlier, museums were concentrated in the southern end of Manhattan, with a large cluster in Soho. It was decided that museums from a 2 kilometer radius of Soho would be pulled using the FourSquare API to best express the effect that distance from a museum might have on an AirBnB listing. The average latitude and longitude of the museums pulled was calculated and used as the distancing point.

A dataframe was created, holding only AirBnB listings that were based in Manhattan. The distance of each AirBnB listing in the set from the average co-ordinates of museums in Soho was calculated.

A boxplot was constructed looking at the distribution of distances from the average museum in SoHo for each listing.

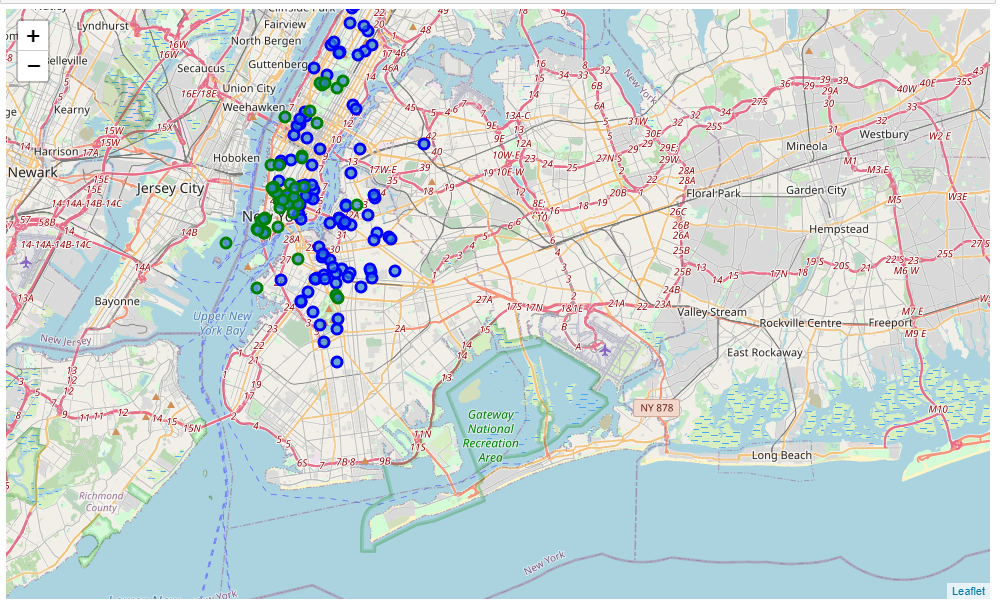
A linear regression model was constructed examining the relationship between distance from the average museum in SoHo and the number of reviews received for an AirBnB listing based in Manhattan. The data was split, trained and tested with a 20% testing split. It was evaluated using metrics of MAE, MSE, RMSE and comparison of actual values against predicted values by the model.

**RESULTS**

There were 48,895 AirBnB listings in the initial dataset. The largest minimum number of nights a lodger was required to stay was 1250 and the maximum price per night was $10,000. The average listing received 23 reviews, required 7 nights stay and cost $150 per night.

Listings in Manhattan had the highest mean prices, followed by Brooklyn. The Bronx had the lowest mean price and was close to price for listings in Queens and Staten Island. Conversely, Staten Island had the highest number of reviews with Manhattan having the least.

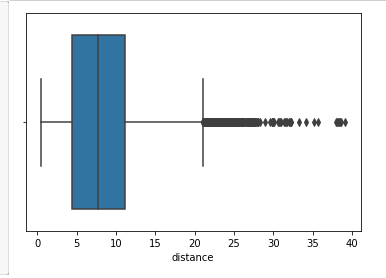
No clear relationship was found between price for a night and the minimum number of nights required to stay per listing. There appeared to be parabolic relationship with a dip in price near the 15-25 minimum night mark but it is not a clear trend.

The visualization of AirBnBs in New York, as represented by blue markers overlaid on the map of New York, shows that AirBnBs were concentrated in Manhattan and distributed axially. However it should be noted that due to computational load, only 100 AirBnBs were listed.

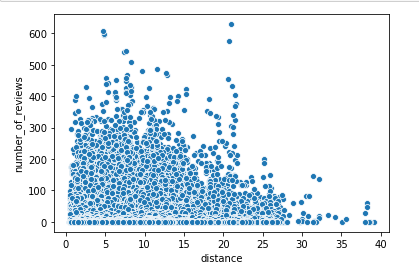
The FourSquare API pulled 50 museums in New York with the search query “museum” used as a parameter.

The data from the FourSquare API was loaded into a data frame and appended onto the map of New York, represented as green markers. As can be seen, the museums are primarily clustered in the south of Manhattan, with a strong concentration in the SoHo neighbourhood.

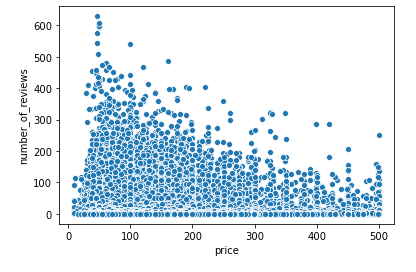
The distance of each AirBnB from Central Park where the three most popular museums in New York are located, as seen in the map at Central Park, was calculated for each AirBnB and loaded onto a dataframe.

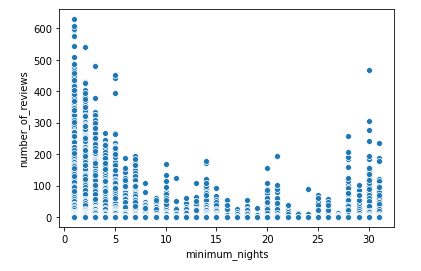
From the boxplot of distance of each AirBnB from Central Park, 50% of AirBnBs were 5 – 12 kms away from Central Park with the furthest AirBnB 40 kms away.

No clear relationship was found between the price of a listing and the distance of a listing from Central Park however they appeared to be a drop in price as the distance from Central Park increased, but this trend could not be substantiated.

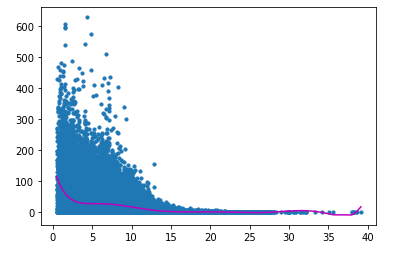
There appeared to be a relationship between the number of reviews a listing received and it’s distance from Central Park. As the distance increased, the number of reviews decreased with a sharp decrease after 20 kms away from Central Park. From 0 – 12 kms distance, the number of reviews stayed constant accounting for 75% of listings in the dataset.

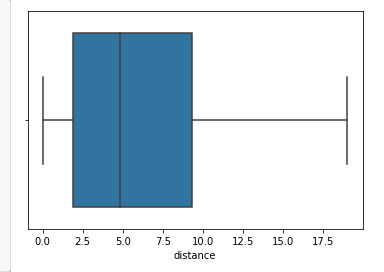
There appeared to be a downward trend for the relationship between price of a listing and the number of reviews it received. As the price of a listing increased, the number of reviews it received decreased. The highest number of reviews occurred around the $50 per night ma rk.

There was a parabolic relationship between the number of reviews a listing received and the number of minimum nights required for a stay, with the highest number of reviews occuring around the 1 night mark before slowly decreasing to a low at the 15 night mark and then increasing again to the 30 night mark. It appears that while shorter stays are better for more reviews coming in, one month stays also hold merit in getting bookings.

A linear regression model was built to assess the relationship between distance from Central Park and the number of reviews a listing recieves. An intercept of 20.17 a nd a co-efficient of 0.49 was found, indicating that for each increase in one unit of distance, the number of reviews increased by 0.49 of a unit. This is a positive relationship suggesting that distance away from Central Park gets more reviews. However the MAE and RMSE at 27.83 and 45.07 showed that this model was inaccurate and that these results should be treated with caution.

The polynomial model was evaluated to be inaccurate. The metrics found an r2 and RMSE of 0.12 and 41.92 respectively, showing that a polynominal relationship between number of reviews and distance from Central Park was both weak and inaccurate.

In addition, the model was prone to both underfitting in the general sense and overfitting at the tail extremes of the data, emphasizing its weak a nd inaccurate. It can concluded that a polynomial relationship does not exist for distance from Central Park and number of reviews.

The multiple regression model between the number of reviews and price of a listing and minimum nights of a listing found co-efficients of -0.02 and -0.77 respectively. This is a negative relationship, suggesting that increases in price and minimum nights decreased the number of reviews. The relationship between price and number of reviews was tenuous but the number of minimum nights had a strong impact on the number of reviews received. But the MAE and RMSE of 27.26 and 44.54 showed that this model is inaccurate, and should be treated with caution. The polynomial relationship was also found to be invalid, with poor metrics of 44.16 and 0.03 for the RMSE and r2 respectively.

The distance from the average museum in SoHo was f ound for AirBnB listings in Mahattan and loaded into a data frame. From there the boxplot showing the distribution of distances found that 50% of AirBnBs in Manhattan were 2.5 – 9.5 kms away fro m the average museum. This is a large spread showing AirBnBs are widely distributed around Manhattan.

The linear regression model examining the relationship between number of reviews and the distance from the average museum in SoHo showed a co-efficient of 0.42 indicating a positive relationship that as the distance increases, the number of reviews increased. This is an inaccurate model as shown by the metrics of 25.95 and 42.60 for the MAE and RMSE respectively.

**DISCUSSION**

The findings have elucidated many factors on the number of reviews a listing can receive. Firstly, the neighbourhoods shows that while Manhattan has highest average price for listings, Staten Island listings actually recieves the most reviews. This is corroborated by the relationship between price and number of reviews where increases in price decreases the number of reviews coming in. This implies that lower priced listings are more likely to get reviewed, which means they were more likely to be booked.

Likewise, the smaller the number of nights required for a stay, the more likely a customer would book that listing. However there was a resurgance in reviews at the 30 day mark, indicating that hosting an AirBnB which has month long stays is a viable option. The multiple regression model showed that the number of minimum nights had a stronger impact on a listing being booked while the effect of price was neglible in that spread, showing that more attention should be paid to the number of nights required. This could also explain the large number of reviews at the $50-$100 mark as a large portion of people were only staying for one night.

Distance from amneties such as Central Park and museums did not play a strong role in customers booking listings. If anything it was found that as distance increased, the number of reviews increased. Central Park and SoHo are both expensive neighbourhoods so any AirBnB in these areas are more likely to have higher prices and thus less likely to bring in bookings. It could that be Manhattan is small enough that customers are happy to travel around the island using the subway and stay somewhere further away from the amneties but cheaper.

It is important to note the assumption that a large portion of bookings resulted in reviews and this booking to review conversion rate was evenly distributed across price range and neighbourhoods. It could be that customers who are more likely to stay in Manhattan are also less likely to place a review. As such given that we are proxying the number of bookings on the number of reviews left, there will always be an information gap that should be minded.

**RECOMMENDATIONS**

If you are thinking of setting up an AirBnB in New York, Staten Island might be a good place for revenue. It has the highest average of reviews left, indicating the highest rate of bookings. Queens and Bronx are also good areas to consider setting up an AirBnB. Manhattan is the priciest so you can set up a high price rate but it also has the least bookings.

If you’re setting the minimum nights required to stay, you should preferably keep it short at 1-3 nights stay. You can take it up to 7 days stay however after that point the likelihood of bookings decrease substantially. Alternatively you could consider setting up a 30 day minimum stay as well as it brings in a fair amount of bookings.

Like all businesses, the cheaper it is the more customers you’re likely to attract. And AirBnB is no exception. A price range of $50-100 per night brings in a lot of bookings however the minimum number of nights is more important than the price.

And if you’re willing to set up an AirBnB in Manhattan alone, don’t feel pressured to set up shop near Central Park or the Soho neighbourhood. Most customers are happy to stay further away and travel around the island. Amneities don’t have as strong a pull compared to a low number of minimum nights and a low price.

**CONCLUSION**

The analysis shows that there are factors which can influence the number of reviews left, and thus the number of times a listing was booked. Understanding the strategy and locale behind the AirBnB is the best route to ensuring that the AirBnB continues to be successful and remains an important asset.