Project Report: Exploratory Data Analysis and Visualization of NYC Airbnb Dataset

Shilpa Singh
Indiana University Bloomington
School of Informatics and Computing
shilsing@iu.edu

Abstract

In this project I am trying to get preliminary insights about Airbnb data of NYC city through exploratory data analysis and visualization.The aim is to answer some interesting questions through a variety of different visualizations.

Introduction

Motivation

Airbnb has become increasingly popular since its launch in 2008 with the number of rentals listed on its website as 6 million in 100,000 cities and growing exponentially each year.(Hartmans Hartmans2017)It has become the primary accommodation provider not just for travellers who want to save money but also for business travellers who are also preferring this because they offer more business-friendly places to stay.New York City has been one of the hottest markets for Airbnb, with over 50,000 listings as of September 2019. This means there are over 40 homes being rented out per square km in NYC on Airbnb! (insideairbnb insideairbnb2019) One can perhaps attribute the success of Airbnb in NYC to the high rates charged by the hotels, which are primarily driven by the exorbitant rental prices in the city. I present here my exploratory data analysis and visualizations throwing some key insights into the Airbnb data of NYC about its property types,prices and demand.I focus on New York Citys data because I wish to perform an in-depth analysis of the real estate picture on one of the most densely populated cities in the world.

Related work

There are many related work in this domain which have tried to explore the listings of Airbnb in NYC area geographically.One such interesting work is presented in the link https://nycdatascience.com/blog/student-works/how-airbnb-is-in-nyc-interactive-data-visualization-in-r/

Here there is visualization with the name Airbnb Listings in NYC as shown on the other side.

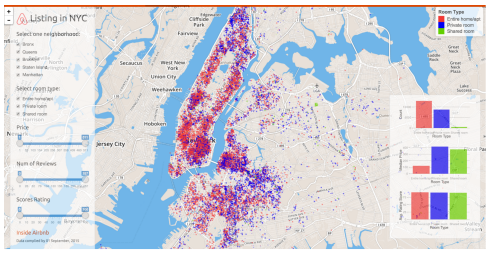

Figure 1: Airbnb NYC listings Dashboard

This is like an interactive dashboard where the user will enter the location on the left and the corresponding listing will be displayed on the map for 3 different categories shown with 3 different colors.Every listing on the map is shown by a circle.On the right side of the visualization there are bar charts showing the number of listings,their price and rating for each property type.This visualization technique using a dashboard is effective if we want to build a search mechanism for the listings present in a specific location in NYC.Also, the different categories are distinguished through different colours making it easy for the viewer to identify the type of property in a particular location.

There is another visualization on the same web page which shows details about the Airbnb property types in Manhattan and its neighbourhood.Overall,this visualization is good enough for displaying the listings types against their price in NYC and its neighbourhood.It also provides a scroll bar on the left to search for the listings within a certain price range.This is effective if a user is interested in the listings falling within a certain price range and can get an idea about the type of property whether it is a private room or hotel which is available to him in that price range.

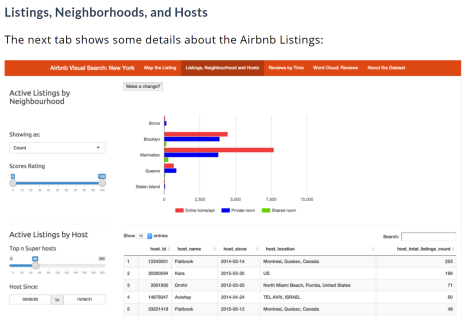

Figure 2: Airbnb NYC listings types

There is another cool visualization in the same link which analyzes the review text using Wordcloud.This idea seems really exciting to me and the visualization looks very convincing.It highlights the keywords in the review comments and throw some insight into the sentiment of the people who have provided their review for these listings.

Figure 3: Wordcloud review text

There is another visualization which shows the number of reviews have increased over the time which is an indication of the increase in the popularity of Airbnb.The visualization technique is that of a time-series plot as the number of reviews is being plotted against the time dimension year.Looking at the trend of this time-series plot,we can easily conclude that the popularity of Airbnb has grown over the years.

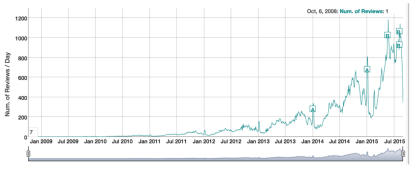

Figure 4: Number of Reviews over time

There is yet another visualization presented in the link https://towardsdatascience.com/airbnb-rental-listings-dataset-mining-f972ed08ddec which shows some interesting graphs highlighting the demand and price of Airbnb listings over the years.This is a scatter plot of number of reviews against the years fitted with a regression line and also shows how the trend changes over the months in 2017 and 2018.This is a typical time-series based analysis over the different time dimensions like years and months and help us analyze the trend in the growth of popularity.

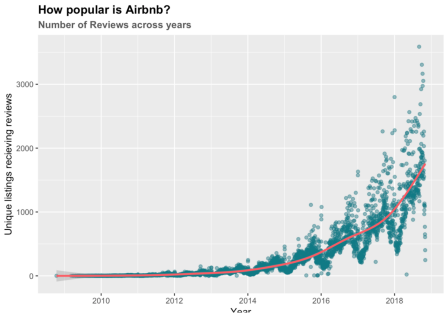

Figure 5: Seasonality in Demand

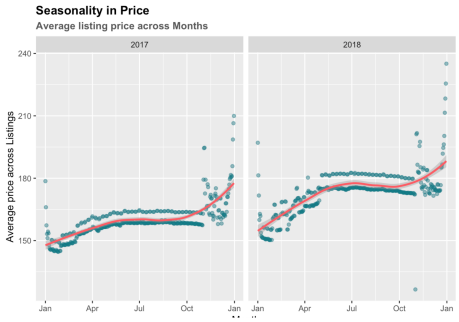

Figure 6: Seasonality in Price

Contribution

In this section, I would like to highlight my contributions to the existing work and explain how it is effective in this analysis.The objective of this work it to answer the questions through visualization which can gain deeper insights into the Airbnb dataset.We typically look for the following new insights.

1. What are the different listings of Airbnb in NYC and how they are located in the map.
2. What are the different types of properties rented in NYC? Do they vary by neighborhood?
3. How to visualize the price and popularity of listings across different regions in NYC and its neighbourhood on the map
4. What localities in NYC are most popular?
5. What locations are the most expensive?

6. In theory, we should see an increase in price the closer we get to the center of Manhattan. It will be very interesting to see if the popularity for the given accommodations also increase the closer we get to the center of Manhattan.
7. How the price of listings have changed over the time.Can we show this trend over the months in the last 2 years or specifically can we show some seasonality in price in the last few years?
8. How the popularity of listings have changed over the time.Can we show this trend over the months in the last 2 years or specifically can we show some seasonality in popularity in the last few years?
9. Are the demand and prices of the rentals correlated?
10. Does the weekend price varies from the weekday prices?
11. How the availability of the listings change over the weekdays and weekends for 2019.
12. Can we analyze the user review text using word cloud to get the sentiment from the review text.

Though the existing visualizations present a good dashboard to enable the visualization of listings based on the location and price,they do not offer any visualization which could show how the price and popularity of listings look like on the map.This will be a choropleth showing the popularity and price in different neighbourhoods of NYC.This is something which has been added through my contribution.Also,there is no visualization showing the analysis between the popularity and price in a time-series fashion extensively.Based on the kind of analysis, I have divided my work into 3 sections.

1. Spatial Analysis: This section will focus on the spatial analysis of the listings in different neighbourhoods by the property type.It will display the listings on the map of NYC neighbourhood and highlight each listing type with a different color.Also, it will show the choropleths showing the price and popularity on the map of NYC and its neighbourhood.It will also show the price variation according to the property type and the neighbourhood of NYC.Lastly,it will show the comparison in the number of property type in different neighbourhood and neighbourhood groups.
2. Price and Popularity Analysis: This section will highlight the trends in price and popularity over the number of years and the seasonality in price and popularity over the months for the last 2 years.It also shows the price trends in different weekdays and weekends for the last 2 years and the availability of the listings across the year.
3. Analysis of the review text in WordCloud to find the sentiments of the reviews.

Data and Methods

The dataset comprises of three main sources:

- listings - Detailed listings data showing 13 attributes for each of the listings. Some of the attributes used in the analysis are price (continuous), longitude (continuous), latitude (continuous), listing type (categorical), neighbourhood (categorical), number of reviews (continuous) among others.
- reviews - Detailed reviews given by the guests with 6 attributes. Key attributes include date (date time), listing id (discrete), reviewer id (discrete) and comment (textual).
- neighbourhood json - This is the geojson with the coordinates of the polygons to be plotted on the map and different attributes of these coordinates are mentioned in the property section of the json object.
- The data is downloaded from the link http://insideairbnb.com/get-the-data.html

In this section, I would like to discuss the various ideas,sketches and prototypes for the kind of insight I want to get.

- Spatial Data Analysis: In this section, I wanted to display the listings on the map and explore data-points like price,popularity and type of listing using map visualizations to answer questions related to differences in prices and popularity across different locations in NYC. A choropleth is a good visualization technique to see any quantitative parameter like price and popularity score across different regions on map. There are various libraries like geoplotlib,datashader and altair for choropleths.I planned to use altair for choropleth as it is one of my favourite libraries since it has many useful features.A typical choropleth will look like something below.

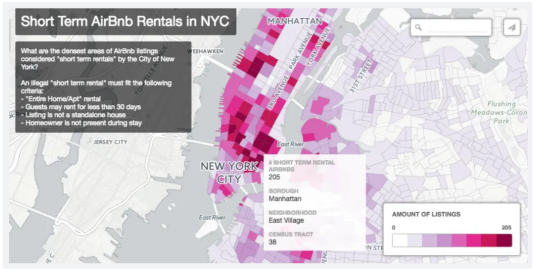

Figure 7: Choropleth of number of listings in NYC region

Also, as part of Spatial Data Analysis, I wanted to plot the different listing types on the map.An example of such a visualization would be like the below one.

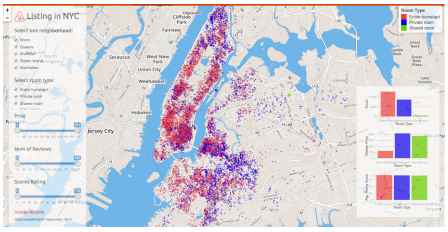

Figure 8: Plot of listings on the map

I have used the altair for the plot of price,popularity score and listings but it has not been layered on a mapbox item like the one in these visualizations shown above.I have plotted the map using the mark geoshape function of altair which takes input as the geographical coordinates which was there in the neighbourhoods geojson.

Another way of doing spatial data analysis in this dataset is to see the number of different category of listings like private room,entire home or hotel in the different neighbourhoods and neighbourhood groups.Since,this falls under categorical analysis,any kind of categorical plot would be suitable for this.A stacked bar chart or a grouped bar chart would be apt for this as we have count of listings which is a quantitative measure on the y-axis.Also, I have used grouped bar chart because it is easier to interpret it and compare the values on the y-axis.The categories used here are the listing types and the number of listings is the measure on the y-axis.Some of the visualizations which could be useful for this scenario are:

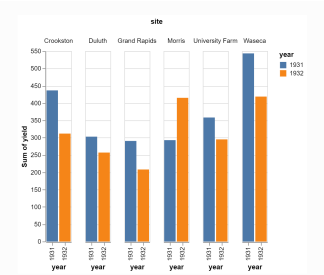

Figure 9: Grouped bar chart for categorical analysis

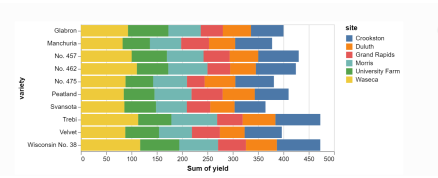

Figure 10: Stacked bar chart for categorical analysis

- Demand and Price Analysis: Based on the dataset,this is a time-series analysis as we have the price and number of reviews(proportional to popularity score) for different dates available from 2016 to 2019.Time-series analysis can be done using line charts or scatter plots effectively.Here, also I have preferred to use altair as it is convenient to use it with time-series data frames indexed with date-time columns in pandas. Also, I have used both scatter plots and line charts in order to visualize the density of the plot effectively and also tried to plot a rolling mean of the values.The line chart helped to analyze the seasonality in price and demand effectively across months and years.A typical time-series plot will look like the ones below:

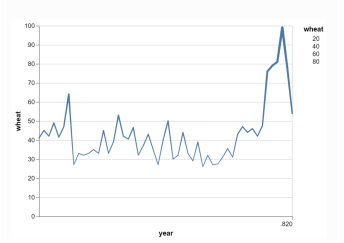

Figure 11: Line chart for time-series analysis

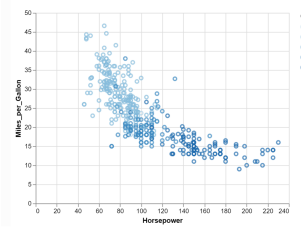

Figure 12: Scatter plot for time-series analysis

- User Review (Textual Data) Mining: Here I have used Wordcloud to analyze the review text and present the important words from the review text.Since, WordCloud is the most popular method,I have used this in my visualization.A typical word cloud will look like:

Figure 13: WordCloud for text analysis

Results

Here, I would like to present the visualizations in the context of the overall objective and the design goals.As before,I will divide the results into 3 categories based on our analysis.

Spatial Data Analysis

In this section,I have created the following visualizations:

- Plot of different category of listings on the NYC neighbourhood map.Color is used to distinguish between the different categories. I have used altair for this plot.The number of listings are too many and they look cluttered in some regions.It is a geotype plot in altair where we have to specify latitude and longitude information which was loaded from the neighbourhoods geojson.It clearly shows the type of listings on various regions of the map.

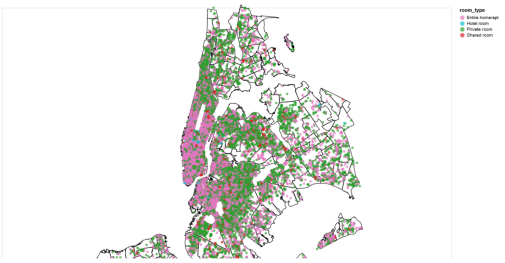

Figure 14: Listings on map in Altair

- Plot of different listing types in the NYC neighbourhood group.The groups are at the highest level of hierarchy.Through this visualization, we wish to study the relationship between property type and neighbourhood group. The primary question we aim to answer is whether different boroughs constitute of different rental types.There are majorly 4 types of property in this subset of data which are Entire Home,Hotel Room,Private Room and Shared Room.We can see from the visualization that Manhatten has highest number of Entire home/apt type property.

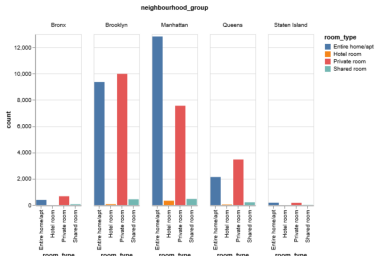

Figure 15: Listings in neighbourhood group

- Plot of different listing types in the NYC neighbourhood : There are many neighbourhood belonging to a group and we wish to see the listing type in the top 20 neighbourhoods which contain the maximum number of listings.

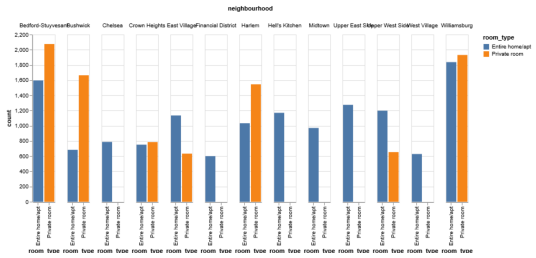

Figure 16: Listings in neighbourhood

- Categorical plot of price against listing type: In this visualization, we aim to plot the price against different categories of property type to visualize the comparison of prices across them.The categorical plot by seaborn is effective for this.

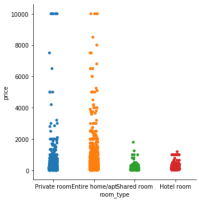

Figure 17: Price by Listings type

- Box plot of availability of listings against neighbourhood: In this visualization, we aim to plot the availability of listings throughout the year against different neighbourhoods and compare them.The boxplot by seaborn is effective for this.

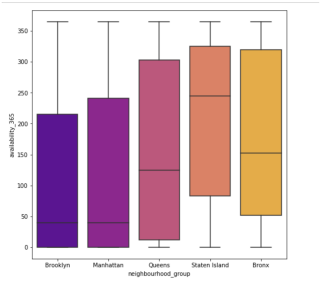

Figure 18: Year round Availability by neighbourhood

- Choropleth map to show the variation in the price of listings in the NYC region: The most important thing in a real estate property is its location.In this visualization, we aim to see the prices in different regions on the map.A choropleth is used for this with goldgreen shade and the most expensive areas are the darkest in color.This is done using altair and the tooltip in altair shows the neighbourhood and neighbourhood group on the map.We can see that the areas around Manhatten

are the most expensive.

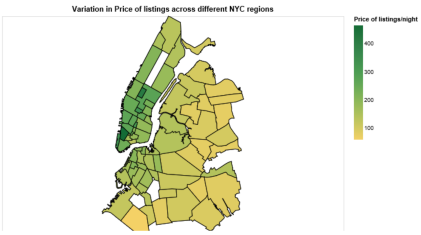

Figure 19: Variation in Price of listings across different NYC regions

- Choropleth map to show the variation in the popularity of listings in the NYC region: In this visualization, we aim to see the popularity in different regions on the map.A choropleth is used for this with goldgreen shade and the most popular areas are the darkest in color.This is done using altair and the tooltip in altair shows the neighbourhood and neighbourhood group on the map.The listing costs are largely in line with the location scores.However, there are few exceptions.We can see that the areas around Manhatten are not the most popular ones though they are the most expensive,in fact the areas near Brooklyn are the ones which have the highest popularity where the prices are fairly low.

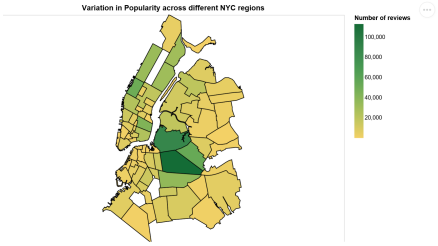

Figure 20: Variation in Popularity across different NYC regions

Price and Demand Analysis

In this section, we will analyse the demand for Airbnb listings in NYC. We will look at demand over the last few years and across months of the year to understand seasonality. We also wish to establish a relation between price and demand. The question we aspire to answer is whether prices of listings fluctuate with demand. We will also conduct a more granular analysis to understand how prices vary by days of the week.To study the demand, since we did not have data on the bookings made over the past year, we will use number of reviews variable as the indicator for demand.The visualizations in this section are:

- Price of listings in the last few years:We have used scatter plot to see the trend in the price of listings in the last few years.We can see that the price has increased in the current years.Also, the data has some outliers.

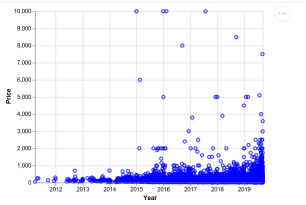

Figure 21: Scatter Plot showing the price of listings across the years

- Demand of listings in the last few years:We have used scatter plot to see the trend in the demand of listings in the last few years.We can see that the demand of listings has increased over the years. We can see an almost exponential increase in the demand of listings.

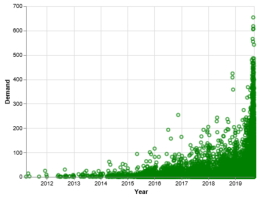

Figure 22: Scatter Plot showing the demand of listings across the years

- Seasonality in price in the last few years:We have used line plot as in case of time-series data to see the trend in the price of listings in the last 3 years.The average prices across listings tends to increase as one progresses along the year and spikes in December and August in 2017,December and September in 2018 and September in 2019.This could be attributed to the holiday seasons near Christmas where there are more demand of housing and in the fall season when the academic years starts and students look for new accommodation.

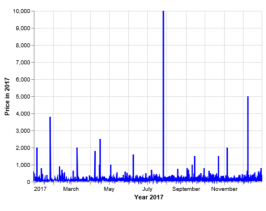

Figure 23: Scatter Plot showing the price of listings in 2017

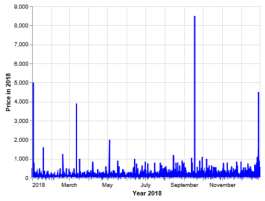

Figure 24: Scatter Plot showing the price of listings in 2018

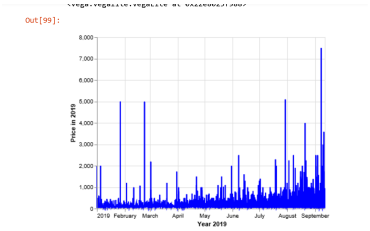

Figure 25: Scatter Plot showing the price of listings in 2019

- Seasonality in demand in the last few years:We have used line plot as in case of time-series data to see the trend in the demand of listings in the last 3 years.The average demand/number of reviews across listings tends to increase as one progresses along the year and there are many spikes throughout the year.This seems counter-intuitive as one would expect the demand to decrease with a decrease in price. This could possibly due to the assumption that we made that number of reviews is a reflection of the demand, which might not always be the case.

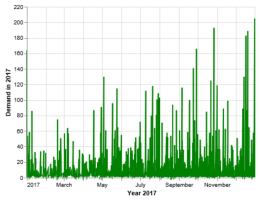

Figure 26: Scatter Plot showing the demand of listings in 2017

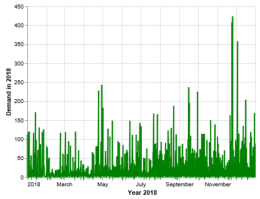

Figure 27: Scatter Plot showing the demand of listings in 2018

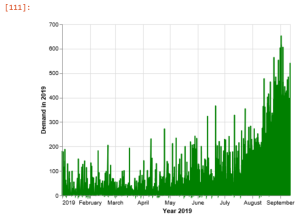

Figure 28: Scatter Plot showing the demand of listings in 2019

- Price Analysis on weekdays and weekends through boxplot for the year 2018 and 2019:We have used categorical plot and box plot to compare the prices in the weekdays.The price near weekend are more expensive compared to the other days of the weeks, perhaps due to higher demand for lodging.

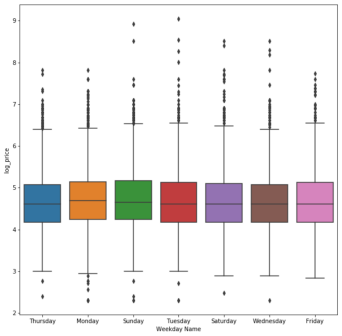

Figure 29: Boxplot of the prices across the weekdays

User Review (Textual Data) Analysis

Review text can tell us a lot about the customer mindset, their expectations and how well those were met. For the final result to make sense, the review text data requires a lot of cleaning like the words need to be stemmed, punctuation need to be removed along with stop words.I have not processed the text completely but after removing stopwords and plotting the WordCloud, I was able to see the dominant words in the review.

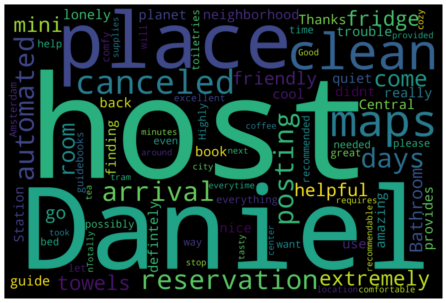

Figure 30: WordCloud of Review text

Discussion and Conclusion

Through this exploratory data analysis and visualization project, we gained several interesting insights into the Airbnb rental market. Below we will summarise the answers to the questions that we wished to answer at the beginning of the project:

- How do prices of listings vary by location? What localities in NYC are rated highly by guests?

  Manhattan has the most expensive rentals compared to the other boroughs. Prices are higher for rentals closer to city hotspots.However,the popularity is not necessarily greatest for the most expensive areas like we can seen on the map Brooklyn where the popularity is highest but the prices are fairly low.This could be because we are using number of reviews for a listing to relate its popularity and demand,which might only be partially correct.Also, areas where the prices are low could have more visitors to give more reviews.

- How does the demand for Airbnb rentals fluctuate across the year and over years?

  The demand (assuming that it can be inferred from the number of reviews) shows a seasonal pattern - demand increases from January to December with several spikes in between In general, the demand for Airbnb listings has been steadily increasing over the years.

- Are the demand and prices of the rentals correlated?

  Average prices of the rentals increase across the year, which correlates with demand. However, the prices show a spike in December and around Fall season as opposed to demand in this month, which is counter-intuitive. Prices are higher on average on weekends, compared to the other days of the week.

- What are the different types of properties in NYC? Do they vary by neighborhood?

  There are more than 20 different types of listings in NYC. The ratio of the type of listings to total numbers varies by borough. Manhatten and Brooklyn tend to have property types that are larger and can accommodate more number of people.

- Are there any common themes that can be identified from the free-text section of the reviews?

  The words are common like host,place,clean and nothing much can be inferred.

Limitations

We only had subset of data for past years and so the analysis was not very comprehensive. There was an assumption made, particularly in the demand and supply section of the report to understand the booking trends where the number of reviews is directly co-related to demand and popularity in all the analysis.

References

[Hartmans Hartmans2017] Avery Hartmans. 2017. Airbnb now has more listings worldwide than the top five hotel brands combined. (August 2017). https://www.businessinsider.com/airbnb-total-worldwide-listings-2017-8

[insideairbnb insideairbnb2019] insideairbnb. 2019. Adding data to the debate. (2019). http://insideairbnb.com/new-york-city/

- https://towardsdatascience.com/airbnb-rental-listings-dataset-mining-f972ed08ddec
- http://insideairbnb.com/get-the-data.html
- https://www.kaggle.com/sjrodgers1005/exploring-airbnb-nyc-homes-through-visuals
- https://nycdatascience.com/blog/student-works/how-airbnb-is-in-nyc-interactive-data-visualization-in-r/
- https://www.businessinsider.com/airbnb-total-worldwide-listings-2017-8
- https://observablehq.com/@chodimella/final-project-airbnb-data-visualisation