AirBnb Data Analysis Using Hive

Samyuktha Muralidharan, Sanjana Boddireddy, Shilpa Konde Deshmukh, Verusca Aimie Capuno
Department of Information Systems, California State University
Los Angeles
Tel. 424-345-5537, Fax. 323-343--5209
e-mail : smurali2@calstatela.edu , sboddir@calstatela.edu, skonded@calstatela.edu, vcapuno@calstatela.edu

Abstract: This paper illustrates the analysis conducted on "Airbnb" Data that contains information about Airbnb Listings.Our analysis is focused on Airbnb Listings in New York City and Los Angeles County and the goal is to perform an in-depth analysis on these most densely populated cities in the USA.This analysis helps us understand the distribution of listings in New York city and Los Angeles county.It gives an insight on the seasonality in demand and we also focus on various factors such as revenue,renting prices and variation of prices with the property type.The analysis is done for the period 2013-March 2020 . This detailed analysis helps Airbnb to showcase the best listings and gives an idea about how they can help their hosts to improve bookings and in turn increase the value of Airbnb. The analysis also helps the property hosts to determine various aspects of renting and gives them insights about the factors to be considered if they are planning to invest in properties in the future. It also helps Airbnb to grow as success of Airbnb depends on the success of their hosts.This paper illustrates the usage of Hadoop, MapReduce, and Hive on big data for easy summarization by utilizing the knowledge gained in lab sessions.We use HiveQL in order to query about 3GB of data and visualize it on Tableau and Excel 3D Maps. We present here our exploratory data analysis, visualizations, animations and lots of other interesting insights into the Airbnb data.
URL:https://www.kaggle.com/samyukthamurali/airbnb-ratings-dataset
Dataset size: 3GB
Cluster version: Hadoop 2.7.1.2.4.2.0-258
No of nodes: 3
HDFS Capacity: 1003.6 GB
CPU Speed: 2.20GHz
Memory: 241.821 GB

1. Introduction

Airbnb, Inc. is an online marketplace that connects people who want to rent out their homes with people looking for accommodations in that locale. It was founded in 2008 and provides a platform for hosts to accommodate guests with short-term lodging and tourism-related activities. It covers more than 81,000 cities and 191 countries worldwide. Travelers can book an Airbnb property through the Airbnb website for less than the cost of a hotel room and also provided with more amenities. Airbnb generates revenue by deducting a commission from hosts for every booking done through the platform. Hosts also pay a certain amount of fee for the processing of the payments of the guests. In the first part we are determining the distribution of Airbnb listings worldwide and also in the States of USA. The second part of our analysis includes identifying the distribution and the number of listings focusing New York and Los Angeles areas, evaluating revenue of their neighborhoods and determining the months fetching the maximum revenue in both the cities. We also analyze the seasonality in demand for listings and various property types using tempo-spatial analysis, determine the average price of listings of several property types and neighborhoods and figuring out the supply and demand for various accommodation/guest group configurations. In the last part we focus on Airbnb Super hosts, analyzing what it takes to become a Super host. The above steps performed gives detailed insights about Airbnb data. It will help the hosts choose more profitable locations and the best neighborhoods to invest in. The amazing hosts and listings have allowed Airbnb to grow to what it is today, allowing travelers to have more interesting and personalized visits around the world. Therefore, it is in Airbnb's best interests to help their hosts succeed. We conducted our analysis by querying the data using HiveQL and visualized using Tableau, a software known for its interactive data visualization feature which can transform data from various sources (i.e. excel, access, csv) into insightful and actionable information. This will help simplify raw data into an intuitive format.

2. Manipulating Datasets

2.1 Tools and Data Processing

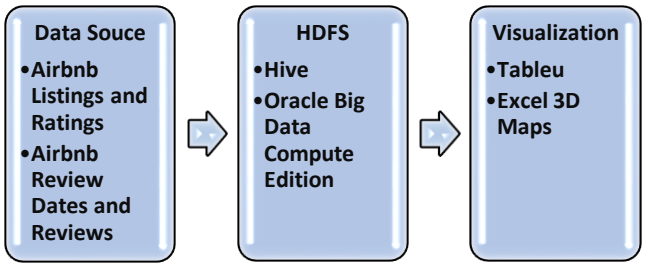

Figure1.Data processing

This analysis is based on dataset downloaded from Kaggle and Airbnb website which were loaded into Oracle Big Data Compute Edition Hadoop cluster. We used Win SCP to transfer the files downloaded from Kaggle to the cluster and used the fundamental command 'wget' to download the files from Airbnb website into the remote node of Oracle BDCE. We connected to the beeline CLI and created tables in Hive to analyze the data. We built intricate codes to query and

evaluate each dataset. The resultant data was then downloaded as a 'csv' file and was visualized in Tableau to derive the results.

3. Related Work

Identifying the patterns in the Airbnb data is very important as it gives insights about the rental market trend and the saturation levels of properties. It also helps identify investment opportunities and mitigate the risks faced. We have analyzed the Airbnb data on various factors that will help on the above and here is a list of papers our work is based upon,

1. Qing Ke (2017) measures the cumulative number of hosts and plots its geo locations. The paper also determines the room types and its distribution in various countries. It also analyses the review count and the distribution of the number of reviews based on various factors.
2. Gábor Dudás, György Vida (2017) analyses the listings in NYC,its price,availability,minimum nights,number of reviews for various property types using statistical methods. It also discusses the geographical distribution of listings. Regression analysis of Airbnb listings is also done.
3. J. Li, F. Biljecki(2019) This paper focuses on number of reviews over years, neighborhood price sensitivity and recommendations for regulating short-term rental businesses based on neighborhood's price sensitivity.

The above papers have used Machine Learning,Python(Numpy,Pandas),Statistical methods to analyze Airbnb data. Based on the above-mentioned analyses we have used HiveQL to query the data and visualize the same using Tableau , that gives an animated and interactive data visualization.

4. Analysis

4.1 Analysis on number of Listings worldwide and in the States of USA.

To start the analysis, we wanted to look at the data at a high level.Figure 2 below shows the number of listings around the globe and we identified that the USA has the highest number of Airbnb listings with a total of 143,954 followed by United Kingdom with 64,407 listings. The states New York and California have the maximum listings in the United States as shown in Figure 3.

Figure 2 . Worldwide Listings

Figure 3. Listings in United States

We are focusing on New York City and Los Angeles County for further analysis.

4.2 Seasonality in Demand

The next few visualizations are analyses of the demand for rentals in both NYC and LA over the years from 2017 to 2020.We used tempo-spatial analysis for the same. We see that the number of bookings for New York City has gradually increased from 2017 and it is the maximum in the year 2019 (Figure 4). There are 5 main boroughs in NYC : Manhattan,Brooklyn,Queens,Staten Island and Bronx. The different colors differentiate the boroughs .We see that neighborhoods in Brooklyn borough have got maximum bookings over years.Figure 5 shows the variation of bookings with months in NYC.We see that bookings are maximum in the months of September and October in 2018 and 2019 (Figure 5).This gives an insight about the trend and the seasonal pattern of bookings.Due to unavailability of booking dates we used 'Review Dates' as a proxy for booking dates.Airbnb claims that about 60% of the guests review their stay and the reviews are given within two weeks of their stay.Hence review dates is a good estimation to study the demand.

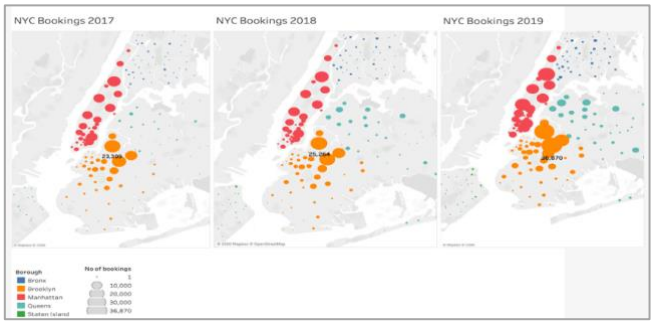

Figure 4. New York City Bookings

Animation link:
https://drive.google.com/file/d/1RnCDN_1Th4bZaHYwLL6KRy0grJtpZcW8/view

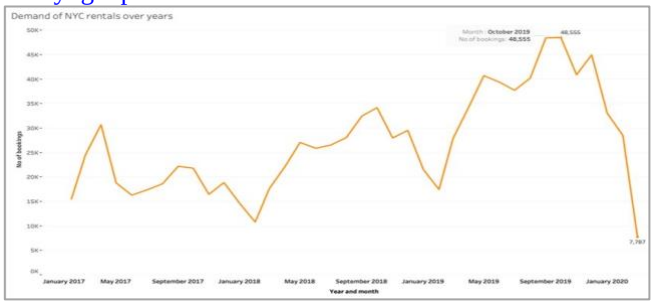

Figure 5. Demand for NYC rentals in each month

Next we analyze the Seasonality in Demand in LA County.The bookings have gradually increased from 2017 and it is the maximum in 2019 (Figure 6).Also Hollywood and Venice neighborhoods have the maximum bookings that shows that travelers wish to stay near the major tourist spots.

A further analysis on the monthly rental demand (Figure 7) shows that July,August and September months have the maximum bookings in 2018 and 2019 that can be related to the summer vacation time in USA and LA sees a lot of visitors during that period.This helps Airbnb and in turn the hosts to determine the trend of bookings for the rental properties.

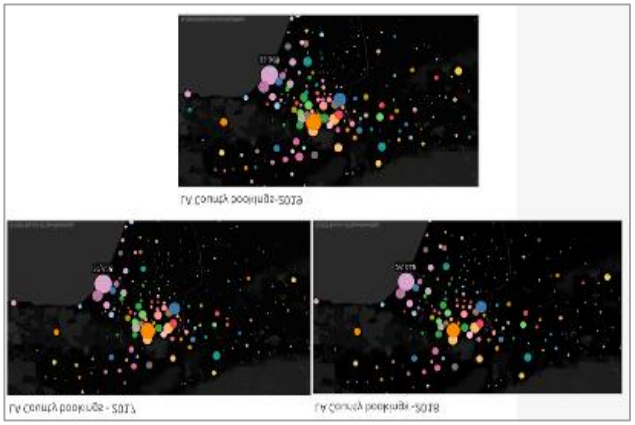

Figure 6. Los Angeles County Bookings
Animation link:
https://drive.google.com/open?id=1fYHQPU-ofW9cfv0slyhra17tA5MbMSMH

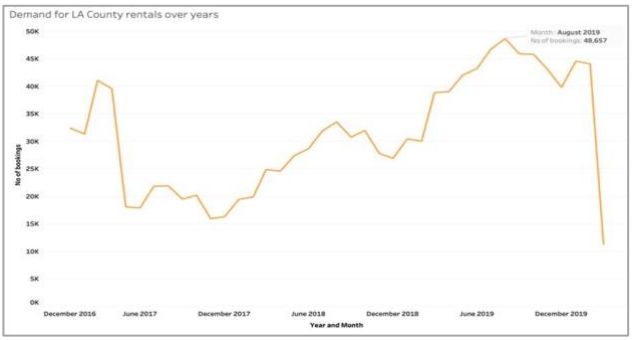

Figure 7. Demand for LA county rentals each month

4.3 Seasonality in Demand for Property Types

There are different kinds of property Listings in Airbnb.We have Apartments, townhouses, villa,hotels, guesthouses,condos etc. In this part we are analyzing the seasonality in demand for various property types in NYC and LA County using tempo-spatial analysis.We wanted to find out if there's a property type that is highly marketable and high in demand for bookings.We chose Apartment property type for NYC as there are many Listings of that type in NYC. Figure 8 shows the bookings for Apartments in NYC over years. We see that the bookings are maximum in the year 2017 and they have reduced in the subsequent years. Although the bookings have reduced they are still in demand in NYC when compared to other properties (Figure 9)and are highly sought after for bookings in NYC.

We chose to analyze the demand for House Property in LA as there are more number of Listings that are of House type in LA. We see that 2017 has he maximum number of bookings with a drop in 2018 and it has slightly increased in 2019 (Figure 10).

If we look at the exact number of bookings for all the property types in LA(Figure 11) we see that the bookings are maximum for House type followed by Apartments.Tourists might have preferred other property types in 2018 and 2019 that resulted in the distribution of bookings.This gives Airbnb insights about the recent trend of property types and hosts get an idea if their property is saturated in the market.

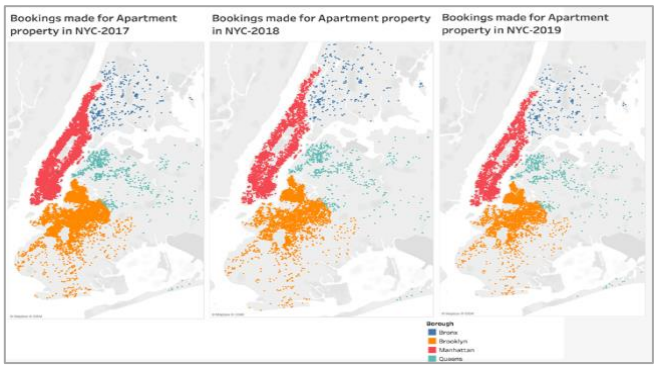

Figure 8.Bookings for Apartments in NYC
Animation link:
https://drive.google.com/open?id=1D3hlrDOvhyDcu4KLB0emLMc0sAxZuoBW

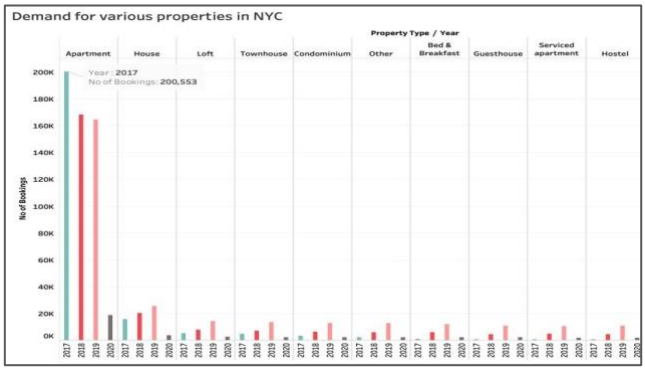

Figure 9. Bookings for various properties in NYC

Figure 10.Bookings for House in LA
Animation link:
https://drive.google.com/open?id=1ZCUwqSegVzsfpflRn_3DuyCa3tYKSgx3

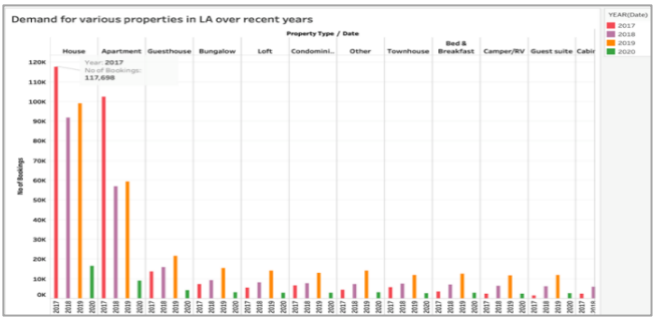

Figure 11. Bookings for various properties in LA County

4.4 Distribution and Number of Listings in NYC and LA

We wanted to dig deeper and determine the neighborhoods that have the highest number of listings to get a better idea of which areas are highly competitive.In NYC , we see that Bedford-Stuyvesant and Williamsburg neighborhoods of Brooklyn borough have a greater number of listings followed by Harlem of Manhattan (Figure 12).These are the major boroughs of NYC that accommodate many travelers.

We analysed on the number of Listings in various neighborhoods of LA County (Figure 13).We see that Hollywood and Venice have the maximum listings followed by Long Beach,Downtown and Santa Monica.These neighborhoods constitute the top attractions in LA and hence there are more listings.

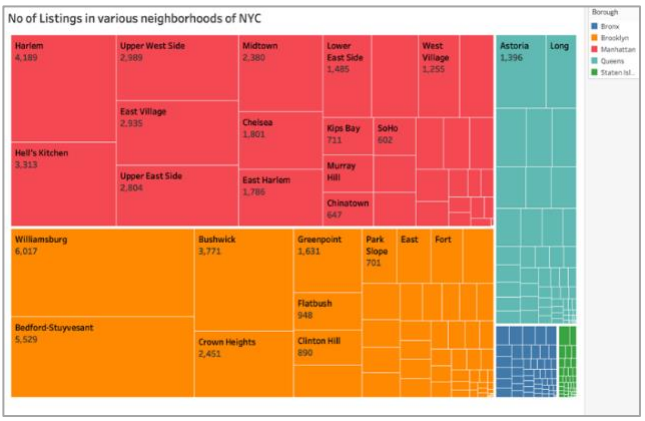

Figure 12. Distribution of Listings in NYC

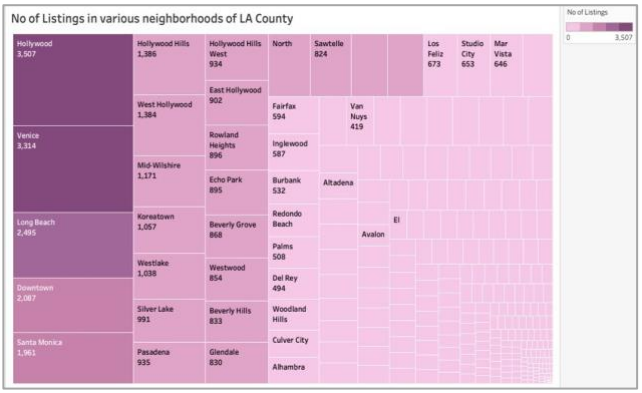

Figure 13. Distribution of Listings in LA County

4.5 Revenue of Neighborhoods in NYC and LA

This analysis is to identify which neighborhood generates the most revenue from Airbnb Listings and to determine which months of the year fetch more revenue in both the cities.We see that Williamsburg neighborhood in Brooklyn borough fetches the highest revenue followed by Bedford Stuyvesant (Figure 14).In general,Brooklyn borough fetches maximum revenue for Airbnb properties followed by Manhattan. So this gives Airbnb and the hosts an idea about which neighborhood they can invest in to gain more revenue.

We further look into the months that provide more revenue in Williamsburg neighborhood of NYC (Figure 15).We see that the months of May,August,September and October generate more revenue.So hosts can utilise this period to rent out their properties and they can utilise the months before May for maintenance of their properties.

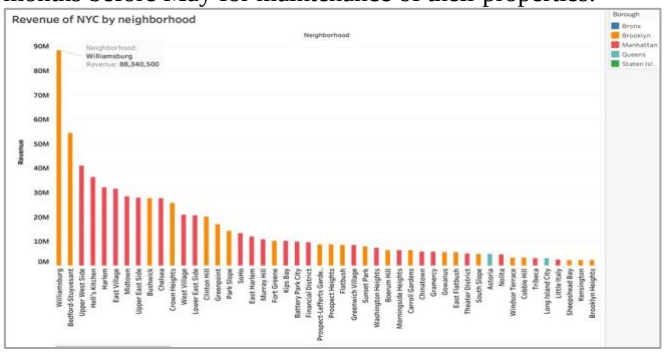

Figure 14. Revenue of various neighborhoods in NYC

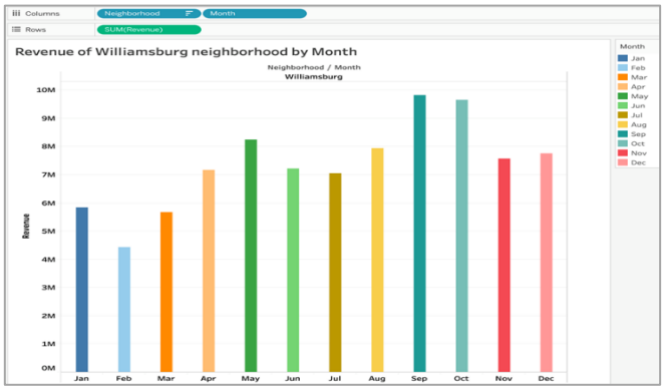

Figure 15. Monthly Revenue for Williamsburg

Animation link:
https://drive.google.com/open?id=13Dogb1SeK4XVZtBz-7bWtEIoP1qKm08Y

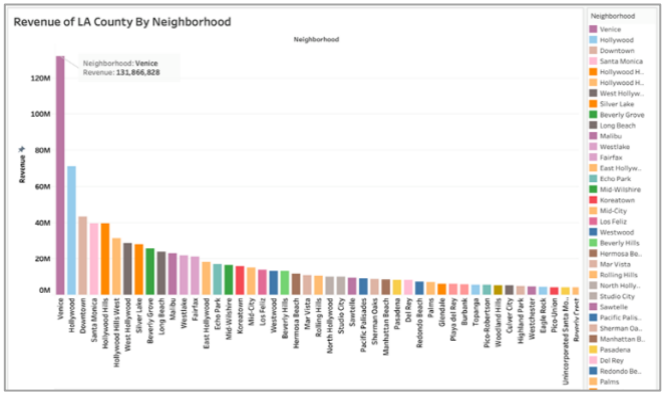

Figure 16. Revenue of various neighborhoods in LA County

We further look into the revenue of various neighborhoods in LA County. We see that Venice fetches maximum revenue in LA followed by Hollywood (Figure 16).If we further drill down we see that July, August, September and October months generate more revenue in Venice neighborhood (Figure 17).Hosts can utilize this period to rent out and they can maintain their properties in the remaining months.

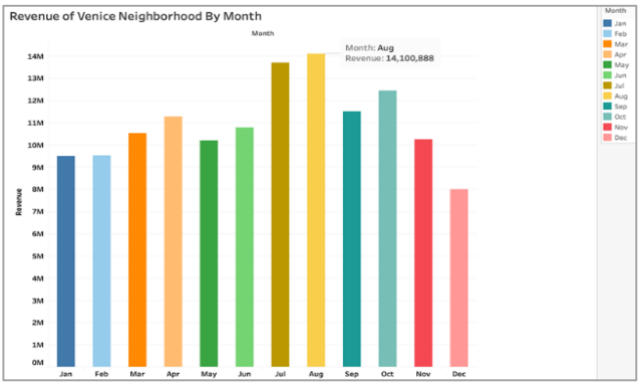

Figure 17. Monthly Revenue for Venice

Animation link:
https://drive.google.com/open?id=1nFKsV9WFvQtb0NUvAnPTYnthLmXaMco0

4.6 Average Price of Listings in various neighborhoods of NYC

Effective pricing is essential to help the hosts know what price to offer for their listings, while maintaining sufficient profit margin and keep up with the competition. We analyzed the data to determine the average price of listings in NYC. We see that the average price is higher in Upper West Side, Theater District neighborhoods (Manhattan Borough), Manhattan Beach (Brooklyn borough) (Figure 18). These can be considered the most expensive neighborhoods to stay in New York City.

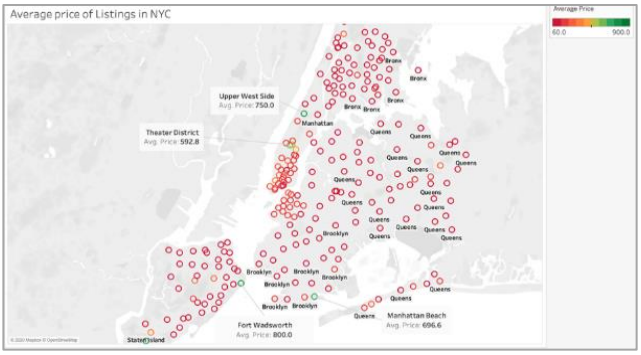

Figure 18. Average price in various neighborhoods

Number of Listings for various prices in NYC: Here we determine the number of listings for various 'prices per night' for the properties. We see that there are around 3000 listings each with price/night of $100 and $150 (Figure 19). The Listings are concentrated in the price range of $50 to $200 and the number of listings is decreasing as the price/night increases. So the number of expensive listings is lesser in NYC.

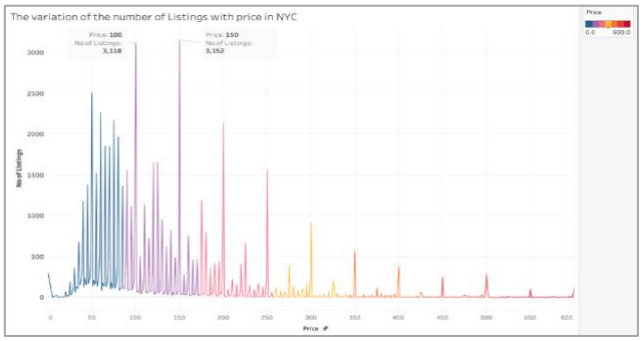

Figure 19. Number of Listings for each Price/Night in NYC

Average price of various Property Types in NYC: We drill down to determine the average price of each property type in various neighborhoods of NYC. We chose five popular neighborhoods for this analysis. Apartments have the highest average price in Upper West side neighborhood of Manhattan. Bungalow has the highest average price in Tribeca neighborhood. Condominium is expensive in Battery park City (Figure 20). This helps us determine the expensive property in each neighborhood and the variation of price with property types.

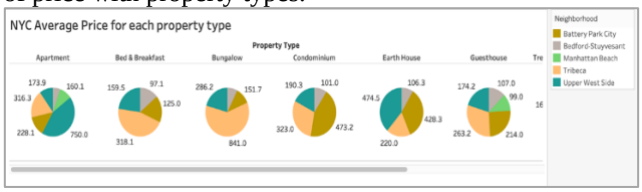

Figure 20. Average price for various property types in NYC

4.7 Demand/Supply for Bedroom configurations in LA

We have calculated the ratio of No.Of.Bookings to No.Of.Listings of each bedroom configuration for properties in LA. We see that properties with one and two bedrooms have the highest ratio signifying the highest demand in LA (Figure 21). This helps Airbnb to understand which configuration is well sought after and helps the host to determine if their property is over saturated in the market.

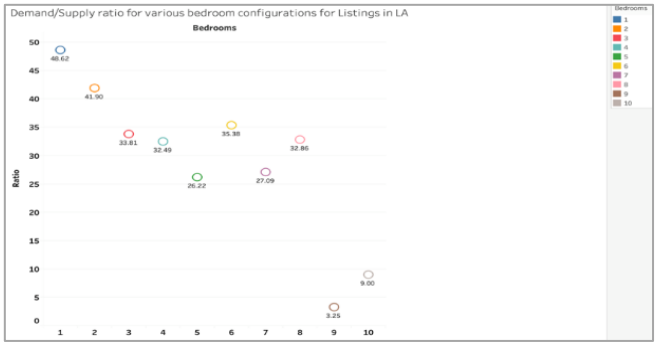

Figure 21. Demand/Supply – Bedroom configuration

Demand/Supply for Accommodation configurations in LA: We calculated the ratio of No.Of.Bookings to the No.Of.Listings of various accommodation configurations in LA. We see that properties that accommodate 2-3 people are well sought after (Figure 22) and it gives the host pretty good regular rentals. This helps the Airbnb host to determine the

common group size of visitors and if their property is over saturated in the market.

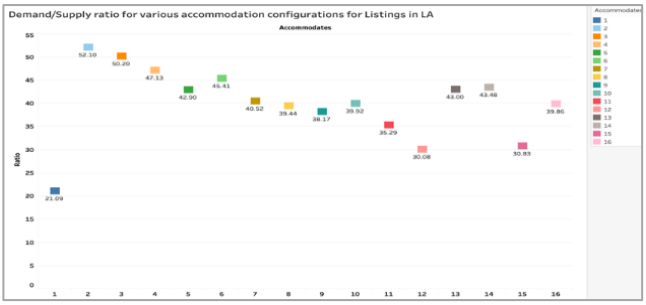

Figure 22.Demand/Supply – Accommodation configuration

4.8 Analyzing what it takes to become a Superhost

Airbnb awards the title of "Superhost" to a small fraction of its dependable hosts. This is an incentive program that is advantageous to the hosts, Airbnb, and their customers. The superhost gets more business in the form of higher bookings, and Airbnb gets happy satisfied customers.

We first determine the percentage of superhosts in LA and NYC. We find LA has around 24% of superhosts and NYC has 15% of superhosts (Figures 23 and 24) out of the total listings in LA and NYC respectively.

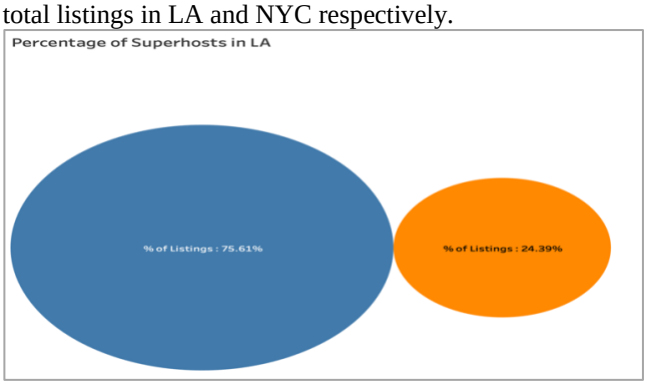

Figure 23. Percentage of Superhosts in LA

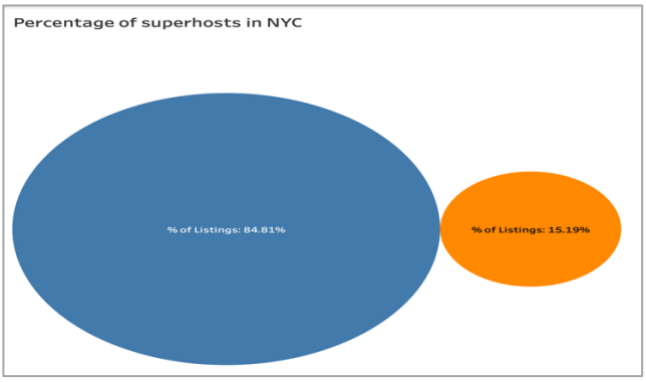

Figure 24. Percentage of Superhosts in NYC

What does it take to become a superhost?: Airbnb's site has certain criteria that must be fulfilled in order to become a Super host. Maintaining a review rate above 50% and a response rate above 90% are some of them.We plot a chart for LA Listings with these two dimensions (Figure 25). Our findings, that are mostly along Airbnb's guidelines, also show some interesting outliers. While most super-hosts are in the high-rating:high-response-rate region, we can also see a few Super hosts with response rates less than 75% (which violates the above 90% criteria set by Airbnb). This is a very small fraction of the hosts. In terms of Ratings, almost all hosts are rated 80% and above.

So most Airbnb hosts lie in the high-rating:high-response region, but only a small fraction get to be super hosts. Thus, becoming a Super host takes a lot more than high ratings & response rates. Some of the other factors being good communication, nice location, monthly availability, cancellation policy, cleanliness, genuineness of the property etc.

Figure 25. Average Ratings vs Host response Rate

5.Conclusion

While exploring Airbnb dataset we successfully used Hadoop, HiveQL,3-D maps and Tableau to store and manipulate data and get several interesting insights about the Airbnb rental market. We analyzed the dataset on various factors that will help Airbnb increase its value and also make successful business decisions. We made use of tempo-spatial analysis to determine the seasonality in demand for the listings and different property types. We also saw the variation of renting prices, revenue and the demand for various configurations of properties. The analyses made helps the property hosts determine the saturation level of their properties in the market and help them increase their bookings and hence the revenue.

References

[1] Qing Ke, "Sharing Means Renting?: An Entire-marketplace Analysis of Airbnb", *Indiana University, Bloomington*(2017)

[2] Gábor Dudás, György Vida, "A socio-economic analysis of Airbnb in New York City", *Regional Statistics,Vol.7.No.1*(2017).

[3] J. Li, F. Biljecki. "The implementation of big data analysis in regulating online short-term rental business". *4th International Conference on Smart Data and Smart Cities* (2019).

[4]http://www.columbia.edu/~sg3637/airbnb_final_analysis.html

[5] https://towardsdatascience.com/airbnb-rental-listings-dataset-mining-f972ed08ddec

[6] https://nycdatascience.com/blog/student-works/analysis-and-machine-learning-modeling-of-new-york-city-airbnb-data/

[7] https://github.com/Samyuktha-M/5200-project