# Exploratory Data Analysis and Data Visualization of Airbnb in New York City

Surya Teja Mothukuri
*under the guidance of Dr. Yong Yeol Ahn*
*Indiana University, Bloomington*
Indiana, USA
smothuk@iu.edu

***Abstract*—This project will use a thorough exploratory data analysis (EDA) and visualization to investigate the Airbnb Open Data for New York City. Through an analysis of multiple variables like pricing distribution, types of rooms, and geographic areas, this study seeks to identify patterns and trends that can be informative to hosts and hosts alike within the Airbnb ecosystem. The research will leverage a wide range of data visualization approaches to display the findings, including geographical mapping, correlation analysis, and temporal trend visualizations. The final objective is to produce a collection of engaging, interactive visualizations that offer a comprehensive and detailed understanding of the New York City Airbnb market.**

***Index Terms*—EDA - Exploratory Data Analysis, Correlation, Geospatial Mapping**

## I. Introduction

### A. Motivation

The richness and complexity of the NYC Airbnb dataset make it a special place for exploratory data analysis (EDA) and data visualization. New York City is one of Airbnb's biggest markets, and the company has completely changed the hospitality business globally. With more than 65 million tourists each year, New York City is one of the most visited cities in the world and a center for discovery of culture, commerce, and leisure (Statista, 2023). Since this surge has increased Airbnb's appeal, it is crucial to comprehend how host tactics and user behavior are influenced by variables like availability, pricing, and location.

Such assessments are important, and empirical data backs them up. For example, studies show that central boroughs like Manhattan and Brooklyn always have the most Airbnb listings and fetch higher rates because of their close proximity to famous sites like Times Square and Central Park. At the same time, vacationers on a tight budget are increasingly choosing less central boroughs, such as Staten Island and the Bronx, where average rates are much lower. These patterns highlight the economic and social effect of Airbnb in urban areas by revealing a substantial interaction between user preferences, pricing trends, and geographic location.

Additionally, research has shown how Airbnb affects local housing dynamics. The NYU Furman Center claims that the rise in short-term rentals has sparked concerns about gentrification and housing affordability in NYC communities. Policymakers may learn about the subtleties of the short-term rental market, hosts can modify their tactics, and the analysis helps optimize the user experience by offering data-driven visual storytelling. Because of this, the initiative has relevance not only for Airbnb stakeholders but also for the larger conversation about housing policy and urban development.

Fig. 1. The Might of New York City

### B. Background and Objectives

The NYC Airbnb dataset has been thoroughly examined in earlier studies, which gives your project important background information. The NYU Furman Center conducted a noteworthy study that looked at the regional trends and legal ramifications of Airbnb use. It demonstrated how the distribution of Airbnb listings in various areas of New York City correlates with variables such as rental profitability and median family income. The economic and regulatory effects of short-term rentals, including their profitability in comparison to long-term rentals in middle-class communities, were highlighted in this study.

The NYC Data Science Academy conducted another study that employed machine learning to forecast Airbnb rental pricing. It addressed issues including skewed pricing distributions and missing data while leveraging elements like neighborhood data, room kinds, and amenities. The study's visualizations, such log-transformed price distributions, provide important

information on how prices fluctuate among various listing categories.

Some criticisms of current visualizations include the need for greater integration of temporal components of listing activity and for more clear representation of neighborhood-level patterns. By include user reviews or occupancy rates, several studies might improve their spatial visualizations and present a more complete picture. Opportunities for your project to provide creative and powerful visuals are highlighted by these criticisms.

This project's main goal was to look at the main elements affecting the demand and pricing of Airbnb listings in New York City. Finding regional patterns in supply and demand, establishing links between revenue and host reputation (e.g., review ratings), and investigating seasonal changes in occupancy were the objectives of the study. It was predicted that homes with better ratings would get more reservations, while central locations and regions close to popular tourist destinations would have higher listing costs. The project focused on the practical uses of EDA and data visualization approaches by doing exploratory data analysis and producing meaningful visualizations that aimed to offer actionable insights for Airbnb hosts, guests, and policymakers.

Additionally, a socioeconomic analysis of Airbnb data looked at relationships between area demographics and listing attributes. It used heatmaps to show how numerical factors relate to one another, exposing trends in host behavior, availability, and cost. Although heatmaps offer a great summary, their excessive dependence on color scales may make them inaccessible to viewers with color vision impairments, indicating the necessity for other encodings such as annotated scatterplots.

Together, these studies show how Airbnb data can be effectively visualized to reveal insights that can be put to use. But they also point out areas that need work, such sophisticated clustering approaches, accessibility, and interaction. Your proposal fills these gaps by providing dynamic and user-friendly visuals that improve the NYC Airbnb dataset's interpretability and usability.

### C. Contribution

Finding trends at the borough and neighborhood levels using interactive spatial and statistical representations is a significant contribution. Heatmaps and cluster-based visualizations, for example, show how boroughs like Manhattan and Brooklyn dominate not only in terms of the number of listings but also in terms of price and popularity. This stands in contrast to more affordable offerings in places like Staten Island and the Bronx, which serve distinct tourist demographics and tastes. These findings are consistent with research conducted by the NYU Furman Center and others, but the project's dynamic methodology makes the findings more relevant by enabling stakeholders to interactively investigate trends.

Additionally, the project identifies relationships between listing attributes (e.g., room type, price, availability) and host behavior using correlation matrices and detailed feature-based analysis. By integrating data such as word usage in listing descriptions and host policies on minimum stays, the project offers new perspectives on how hosts optimize listings to cater to diverse audiences. These granular insights, often overlooked in broader studies, have practical implications for both hosts and Airbnb's platform strategies. This project addresses gaps in accessibility and usability found in prior work. By employing interactive charts and intuitive visual designs, it ensures that insights are accessible to a broader audience, including policy makers, urban planners, and everyday users.

## II. Process

### A. Data Analysis

A wide range of parameters are included in the NYC Airbnb dataset, including categorical attributes like room type, neighborhood, and host identification as well as numerical attributes like price, availability, and reviews. Cleaning and preparing the dataset were the main goals of the initial data analysis. Depending on their importance, problems like missing values in numerical fields (such price and reviews) were either removed or imputed. Boxplots were used to identify outliers, especially in pricing, and capping methods or log-transformations were used to correct skewed distributions.

### B. Data Visualization Methods

To properly examine and display the data, a number of visualization techniques were taken into consideration.

- Distributions of categorical data, like room kinds and borough-level numbers, were shown using bar and pie charts.
- Violin plots were used to display the distribution and density of prices in order to highlight how prices varied among boroughs and room kinds.

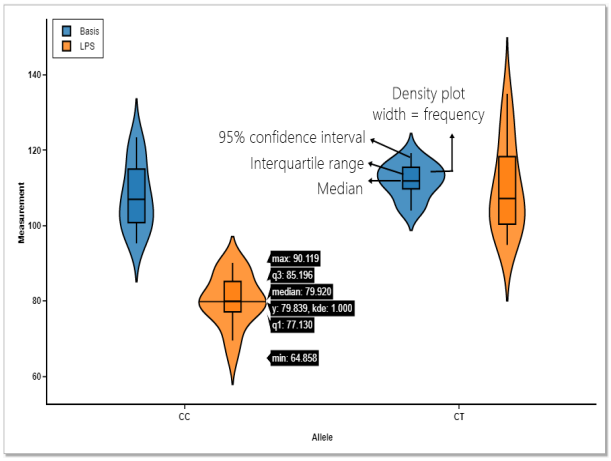


Fig. 2. Violin Plots Features

- Geographical Heat-maps were made to show how many listings there are and how close they are to important places in New York City.

- Interactive Scatter Plots and Cluster Maps techniques increased user interest by enabling dynamic investigation of location and pricing patterns.
- Frequently used phrases were highlighted in word clouds to assess host descriptions and provide information on listing attributes.

### C. Failed Experiments

- There were a lot of data points for price related trends, so having a statis scatter plot for those visualization didn't work. Replacing them with interactive tools improved usability.
- Having stacked bar charts for borough comparisons and their visualizations failed to capture the nuances of data distribution and were replaced with violin plots and density charts.
- When doing temporal analysis, the charts failed to create significant differences between borough-specific trends because of the overlapping lines and instead chose to replace them with separate plots for boroughs.

The effective visuals provided depth, clarity, and interest. Violin plots are perfect for evaluating pricing changes among boroughs because they efficiently integrate statistical summaries and data distributions. By enabling users to select data and zoom into certain regions, interactive scatter plots improved usability and reduced clutter while facilitating dynamic exploration. High-density regions like Manhattan and Brooklyn are prominently displayed in geographic heatmaps, which graphically depict spatial patterns. Host descriptions were condensed into word clouds, which gave the study a qualitative component. Strong numerical associations were rapidly found using correlation heatmaps, which provided information on the interactions between factors like availability and pricing.

On the other hand, several visualizations failed because of limitations in data or design. Particularly in locations with a high population density, static scatter plots become congested with overlapping spots, making them difficult to understand. Despite their simplicity, stacked bar charts were unable to convey the subtleties of multi-dimensional data, such as pricing differences between boroughs. Because they were not separated, overlapping line charts for temporal analysis made specific patterns difficult to see. Compared to bar charts, pie charts were less useful and had poor legibility for tiny amounts. Better palette selections were also required to increase inclusiveness because heatmaps' unoptimized color scales made them inaccessible to colorblind people.

The use of clustering algorithms and interactive representations made it possible to comprehend the data more thoroughly. For example, heatmaps effectively drew attention to densely populated places, such as Manhattan, while cluster maps dynamically displayed pricing patterns, allowing visitors to delve into particular districts. Exploratory plots and sophisticated visualizations, including violin plots, were combined to create a balance between analytical depth and simplicity.

The efficiency of the visualizations was increased by this iterative approach, which also made sure that the conclusions were understandable, useful, and available to a range of audiences.

## III. Results and Discussions

The NYC Airbnb dataset's visualization techniques were chosen after a thorough assessment of the dataset's features, the analysis's objectives, and the intended audience. Numerous potential approaches were taken into consideration, each with unique benefits and drawbacks.

The distribution of Airbnb listing prices among the five boroughs of New York City is depicted in Fig 3. Given its premium status and strong demand, Manhattan stands out for having the largest selection and densest density in the mid-to-high price categories ($100–$300). Brooklyn is a little more reasonably priced, with most of the costs falling between $50 and $200, although it still has a wide range. With the majority of items priced around $150, Queens, Staten Island, and the Bronx have more affordable price distributions. Of these three, Queens offers a balanced range, whereas the Bronx and Staten Island have the least amount of price concentration. All things considered, Manhattan and Brooklyn serve tourists looking for upscale accommodations, but Queens, Staten Island, and the Bronx are more appealing to tourists on a tight budget.

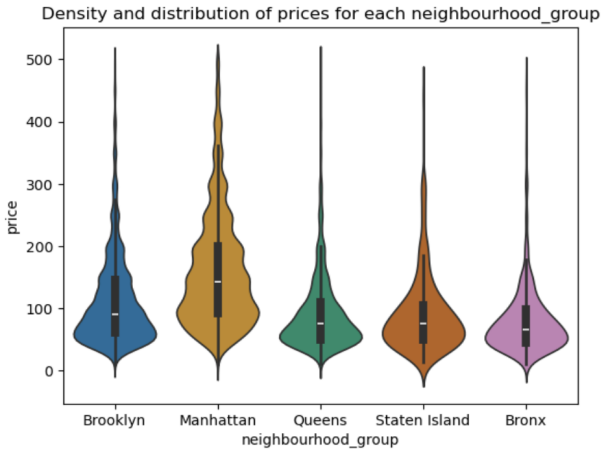


Fig. 3. Violin Plot of all boroughs with prices

In Fig 4, here we can visualize all the listings on Airbnb on a map of the city itself. These have been geographically mapped using the longitude and latitudes of each listings. We can observe that Brooklyn and Manhattan have a lot of Airbnb listings in each and every corner of the borough, where as in the other areas, the listings are sparsely divided in some of the areas.

In Fig 5, the interactive heatmap shows the cost of Airbnb rentals in the New York City region. Warmer tones (orange/red) imply higher rental rates, whereas cooler tones (blue/purple) suggest lower prices. The heatmap's color intensity shows the concentration of rental pricing. Manhattan has the biggest concentration of expensive rental properties,

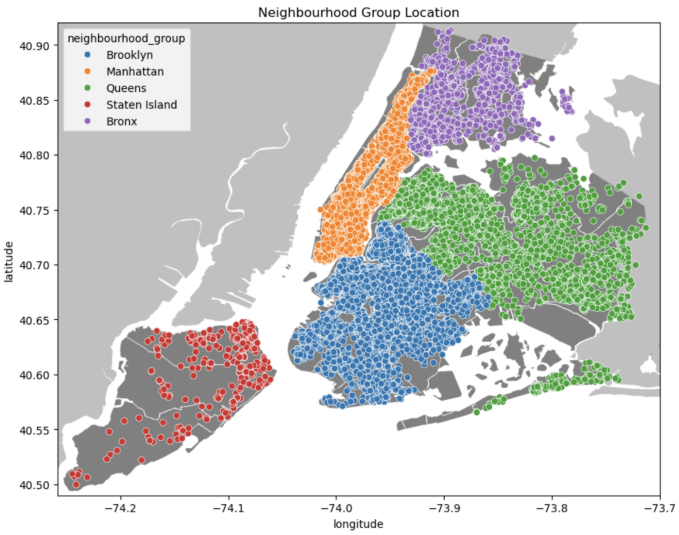


Fig. 4. Mapping of all the listings

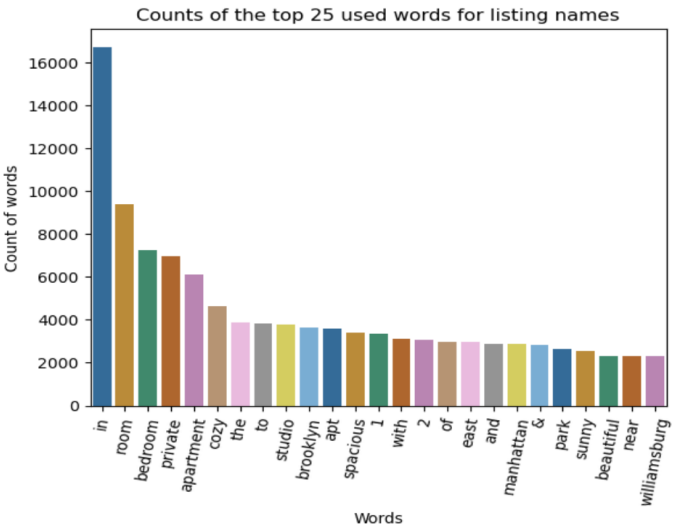


Fig. 6. Counts of Top 25 words used in the listings on Airbnb

especially in the Upper East and West Sides, Midtown, and Lower Manhattan. Prices in Brooklyn are somewhat high, particularly in areas near Manhattan like Williamsburg and Downtown Brooklyn. The cooler tones seen in places farther from the city center, such as sections of Queens, Staten Island, New Jersey, and Long Island, indicate more reasonably priced rental properties. Users may investigate certain neighborhoods and their pricing patterns in greater detail thanks to the map's interactive features.

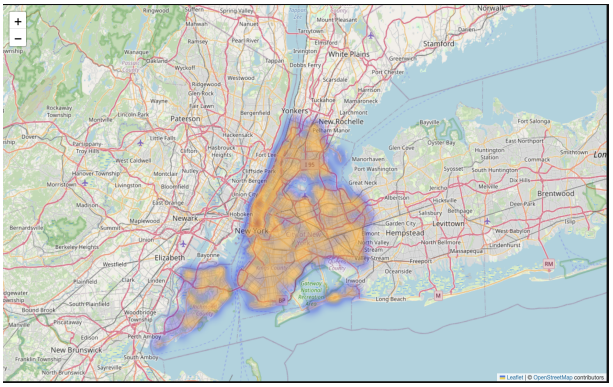

Fig. 5. Heatmap of all the listings wrt prices

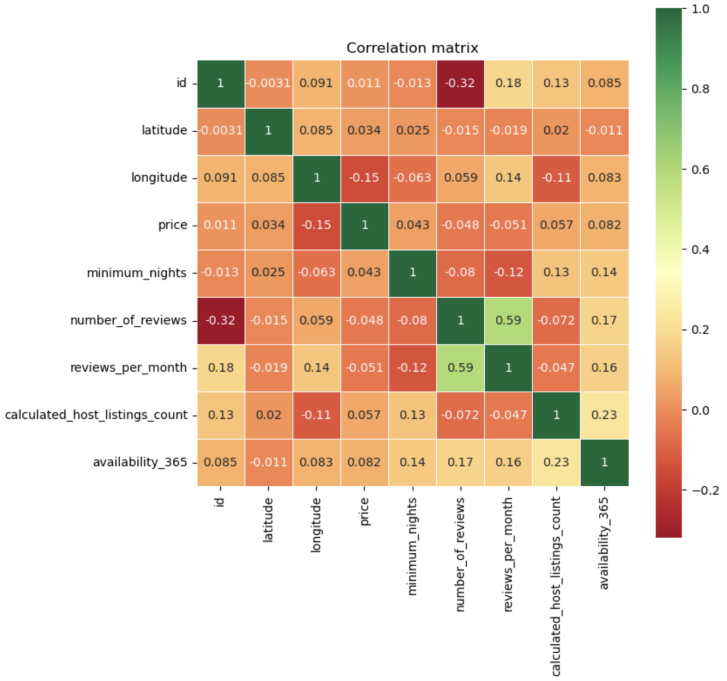


Fig. 7. Correlation Matrix of Numerical Features

Here as shown in Fig 6, we observe that all the hosts have been using (other than the preposition 'in') words that make utter sense to anyone who is reading it like 'room', 'bedroom', 'private', 'apartment', indirectly taking care of all multi-lingual guests as well and not making any use of fancy phrases or words.

Fig 8 gives us a interesting insight. It shows that some property managers in say, Bronx or Staten Island have flexible minimum nights almost close to a single digit but if you observe where as in Manhattan, it might be due to stricter policies or maybe some hosts have very few properties listed, so they want to maximize money by giving it for long duration stays.

In Fig 9, the boxplot shows the pricing distribution of Airbnb rentals in New York City by borough for the three distinct room kinds (private, complete home/apartment, and shared). Manhattan's premium market is reflected in its continually high median costs for all room categories, particularly for full homes or flats. With somewhat lower but still high pricing in comparison to other boroughs, Brooklyn comes next. There is less price variance in Queens and the Bronx, where prices are more reasonably priced and have smaller

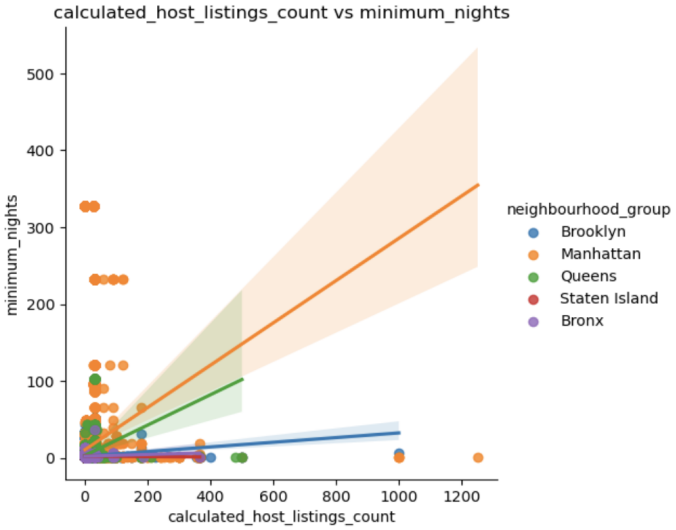


Fig. 8. Host Listings Count vs Minimum Nights wrt 5 Boroughs

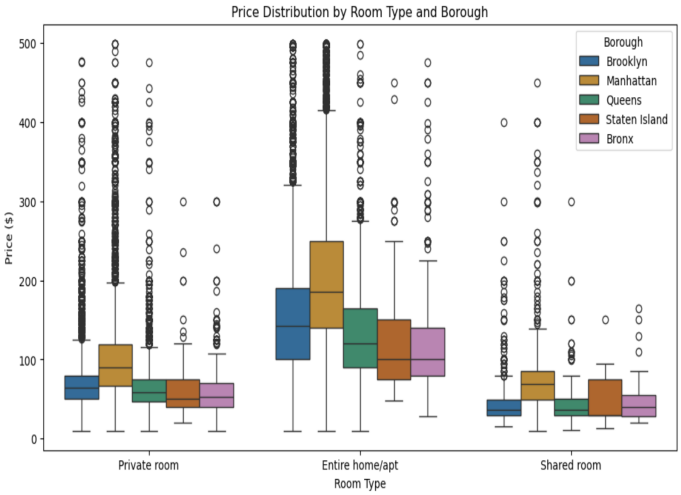


Fig. 9. Prices of different room types in each borough

interquartile ranges. With relatively fewer high outliers, Staten Island has the lowest median costs in every category. Although shared rooms continue to be the most affordable choice, with costs concentrating at the lower end across all boroughs, whole houses or flats have the largest price range.

Fig 10 shows a very interesting and interactive visualization where we can find prices of clusters of all the Airbnb's listed. This price will change interactively as we zoom in/out from the map as the clusters get assigned on the basis of our scrolling. The map shows us where the more costly (red) and less expensive (green) listings are situated. Generally speaking, Manhattan and other central locations, such as Brooklyn Heights or Midtown, have more red marks, which denote greater pricing. The Bronx and Queens, on the other hand, may have more green and orange indicators, which would indicate cheaper costs. More crowded areas, like Manhattan, will have more Airbnb properties because of the denser labeling. This suggests that travelers prefer these places since they are close to major attractions (such Times Square, Central Park, and museums).

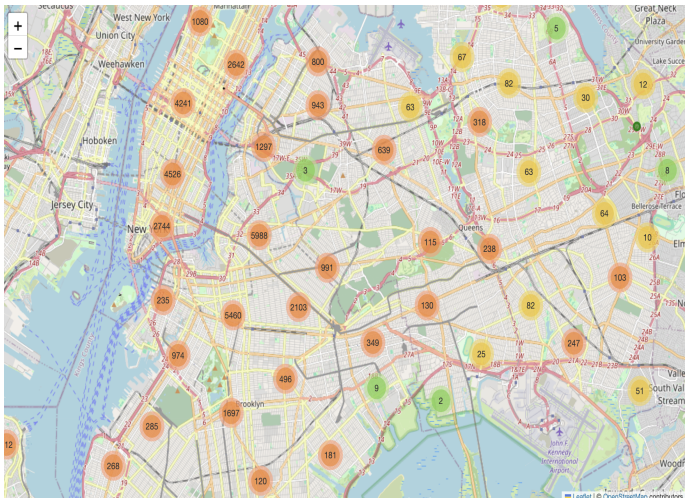


Fig. 10. Interactive Map of Prices of Listings (Clusters)

In Fig 11, we observe that most boroughs have more reviews for more expensive listings, indicating that buyers are more inclined to leave feedback for these houses. Probably, these are in popular tourist destinations like Manhattan or Brooklyn. Off-peak or less-demanded places could be more affordable and accessible all year round, since data may indicate that listings with lower prices often have greater availability. For example, posts in the Bronx and Staten Island may be less expensive due to lower demand. If money is tight, boroughs like Staten Island and The Bronx may provide more affordable and readily available housing alternatives, but you may have to forgo being near well-known tourist attractions.

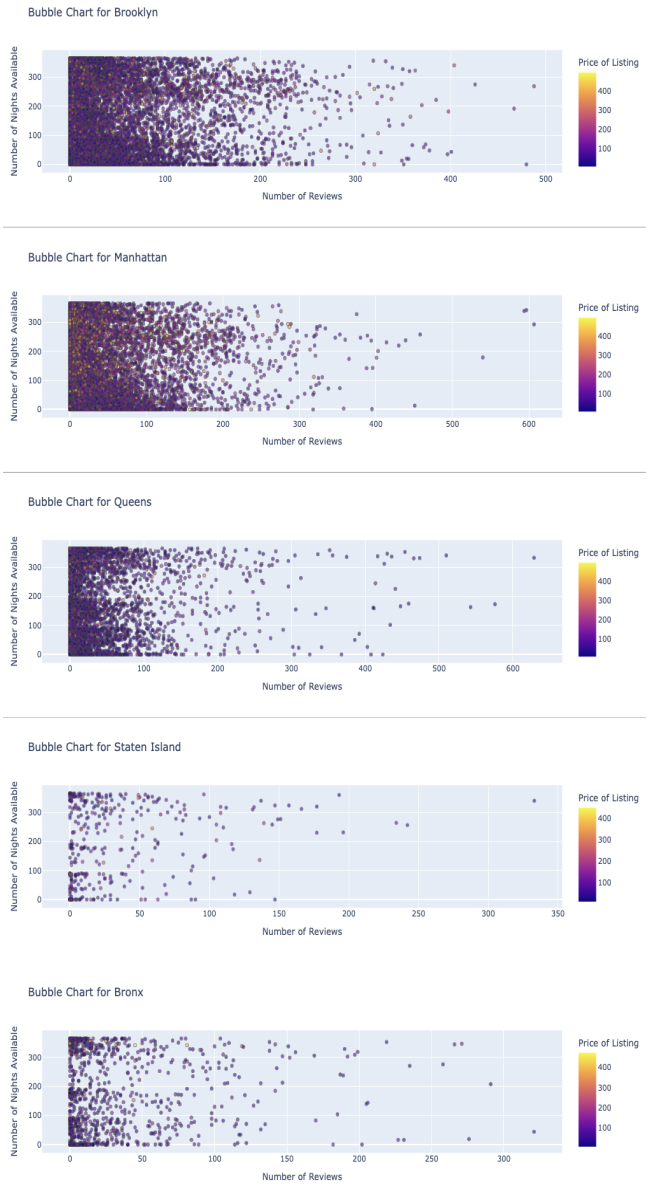


Fig. 11. Interactive Map of Prices of Listings (Clusters)

## IV. Conclusion and Future Work

### A. Discussion and Conclusion

Several important findings were discovered using the NYC Airbnb dataset's exploratory data analysis and visualization. Significant borough-level differences were noted by the investigation, with Manhattan and Brooklyn leading in terms of listing density and cost. These results highlight the boroughs' cultural and economic significance to NYC's tourist sector and are in line with other studies. Boroughs like the Bronx and Staten Island, on the other hand, offered information on reasonably priced choices that catered to longer stays and people on a tight budget.

The visualization of correlations between numerical variables, including price and availability, showed several intriguing patterns, such as the how minimum stay requirements differ between boroughs and how room type affects cost. In particular, interactive visualizations like heatmaps and cluster maps were very good at revealing geographical trends and enabling stakeholders to interactively explore individual areas. The study of listing descriptions using word clouds also revealed frequently used words and phrases, offering qualitative insights into how hosts promote their homes.

The significance of interactive visualizations was one of the key lessons learned. More actionable insights were obtained from tools that let users delve into certain areas of the data than from static displays. Another lesson learned was the need of diversity and accessibility in visualization design, including the use of colorblind-friendly palettes.

### B. Future Work

More in-depth analysis of temporal trends, including seasonal and event-based variations in demand, availability, and price, may be possible in future research on the NYC Airbnb dataset. This would provide tourists useful information to plan economical stays and hosts useful information to improve listing tactics. Analyzing statistics from other international cities, such as Paris or Tokyo, might reveal commonalities and draw attention to NYC's distinct market dynamics. Machine learning-based predictive modeling may be able to pinpoint the main elements influencing listing prices, assisting stakeholders in forecasting market movements. Additionally, by identifying areas for host offers to enhance, sentiment analysis of user evaluations may give qualitative insights into visitor happiness.

Lastly, evaluating the socioeconomic effects of Airbnb, such as its impact on housing affordability and neighborhood demography, may help guide urban policy and promote fair urban growth. The project's scope and usefulness would be greatly increased by these extensions.

## References

[1] Ingrid Gould Ellen et al., "Airbnb Usage Across New York City Neighborhoods: Geographic Patterns and Regulatory Implications," NYU Furman Center, 2017. [Online]. Available: https://furmancenter.org

[2] NYC Data Science Academy, "Airbnb Data Analysis and Modeling of New York City," 2021. [Online]. Available: https://nycdatascience.com

[3] Inside Airbnb, "Inside Airbnb: Adding Data to the Debate," [Online]. Available: http://insideairbnb.com/

[4] Gábor Dudás, György Vida, Tamás Kovalcsik, and Lajos Boros, "A Socio-Economic Analysis of Airbnb in New York City," *Regional Statistics*, vol. 7, no. 1, pp. 135–151, 2017. [Online]. Available: https://www.researchgate.net

[5] Statista, "Tourism in New York City," 2023. [Online]. Available: https://www.statista.com

[6] A. Gupta et al., "Predictive Analysis of Airbnb Listings in New York City," in *International Conference on Data Science and Applications*, 2019.

[7] M. Bateman, "Visualizing Trends in NYC Airbnb Data," *Data Visualization Journal*, vol. 3, no. 2, 2020. [Online]. Available: https://datavizjournal.com

[8] Zillow Research, "Neighborhood Rent and Price Trends in New York City," 2022. [Online]. Available: https://www.zillow.com/research

[9] S. Chen and L. Zhang, "Dynamic Pricing Models for Short-term Rentals: Insights from New York City Airbnb Data," *Tourism Economics*, vol. 26, no. 5, pp. 1084–1103, 2020.

[10] D. Jones et al., "Interactive Visualization of Airbnb Data with Python," *Data Science for Travel Journal*, vol. 5, no. 1, 2021. [Online]. Available: https://traveldatascience.com