DSBA. Data Visualization course

Assignment 1.
Exploratory Data Analysis

CentraleSupélec

Olivia de la Chapelle
Théo Belen-Halimi
Alexandra Pricop
Nithin Alfred

3th of march of 2024

SUMMARY

1. Data Description

1.1. Topic of interest & dataset selection

After looking at a significant amount of datasets online, trying to find the perfect balance between all of the elements we wanted from a dataset we found "***Airbnb in NY 2019***". We all were interested in working on a topic related to the study of geographical distribution of a set of variables. We all agreed on the fact that we wanted to explore a dataset that none of us already went through and from which we could extract useful insights to fulfill our personal interests. As we all want to travel to NY in the future, it made a perfect match.

This dataset was perfect since it checked all of our requirements:

- Complete enough dataset which could allow us to make unique and complementary data visualizations
- Composed of latitude and longitude variables as some of us where particularly interested in using geographical visualization
- We were particularly interested in a dataset that would permit us to focus each analysis on completely different factors: localization, price, host.
- To perform data visualization we wanted to be able to try clustering techniques as well as NLP methods. This was possible thanks to one text's variable corresponding to a description.

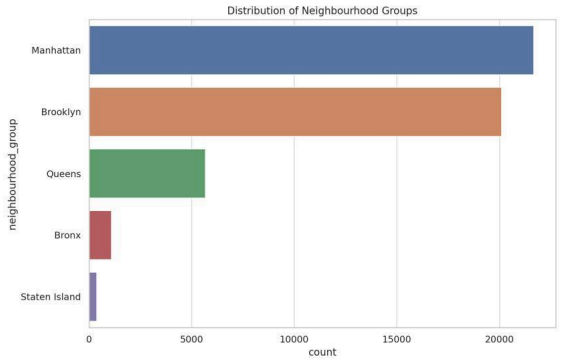
Finally, we believe that getting familiar with simple Kaggle datasets is very useful to get to know how standard dataframe are built from future projects. Starting from scratch, without any indication on what variables are actually correlated with, is perfect to test our creativity.

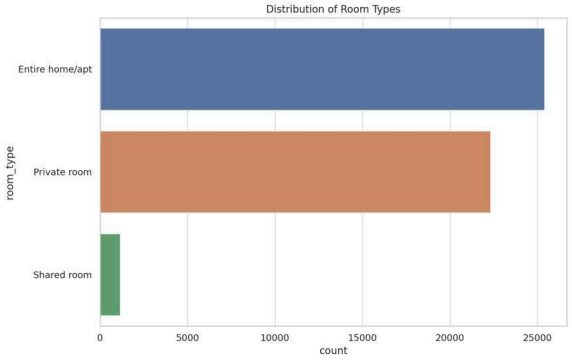
The data collection phase was very straightforward as we could directly download it from the Kaggle website in the csv format. We then started explorations on google colab.

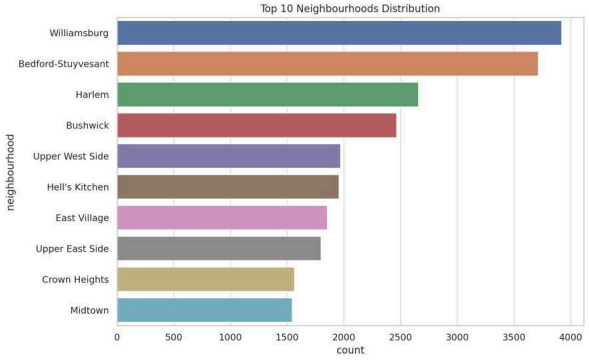
1.2. Insights of data repartition & quality

Column Name	***Description***	***Type***	***Missing Values***
ID	Accommodation ID	Number (int)	0
Name	Accommodation title on Airbnb	Text	15
Host_id	Host ID	Number (int)	0
Host_name	Name of the host	Text	21
Neighbourhood_group	Borough of NY	Text	0
Neighbourhood	Different neighborhood of NY	Text	0
Latitude	Accommodation Latitude	Number (float)	0

Longitude	Accommodation Longitude	Number (float)	0
Room_Type	Room Type	Text	0
Price	Accommodation Price by Night	Number (int)	0
Minimum_nights	Minimum Night you have to stay in the accommodation	Number (int)	0
Number_of_reviews	Accommodation numbers of reviews	Number (int)	0
Last_review	Last review date	Date	10052
Review_per_month	Number of review per month	Number (int)	10052
Calculated_host_listings_count	Number of accommodations of the host		0
Avalability_365	Number of day available for an accommodation	Number (int)	0







Distribution of Neighbourhood Groups

- Manhattan has the highest count of listings, followed by Brooklyn, Queens, Bronx, and Staten Island, in descending order.
- “Brooklyn” and “Manhattan” airbnb represent more than 90 % of all airbnb listed on the platform

Distribution of Room Types

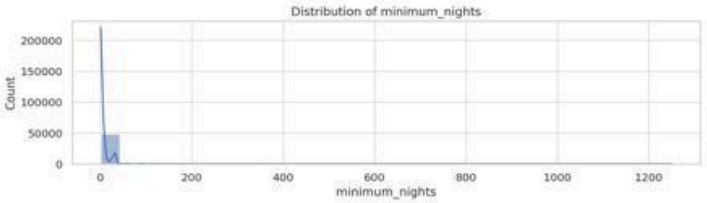
- The most common room type is "Entire home/apt," followed by "Private room," and "Shared room" is the least common.
- “Private room” and “Entire home/apt” represent more that 90% of all accommodations available on the platform

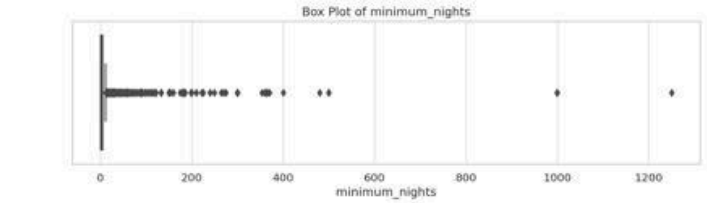
Top 10 Neighborhoods Distribution

- Williamsburg is the most popular neighborhood, followed by Bedford-Stuyvesant and Harlem. Other neighborhoods in descending order are Bushwick, Upper West Side, Hell's Kitchen, East Village, Upper East Side, Crown Heights, and Midtown.

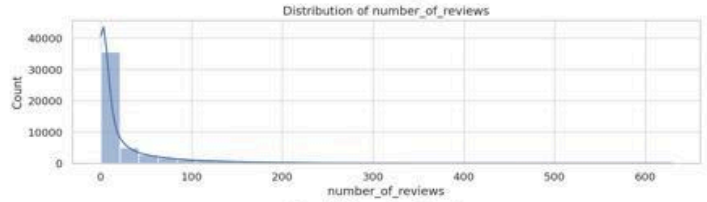
From these charts, one can infer that Manhattan and Brooklyn are the most popular areas for rentals, entire homes or apartments are the most listed room type, and Williamsburg is the most popular neighborhood among the top 10 listed. This information could be useful for understanding market demand, planning for property investments, or strategizing for hospitality services in these areas.

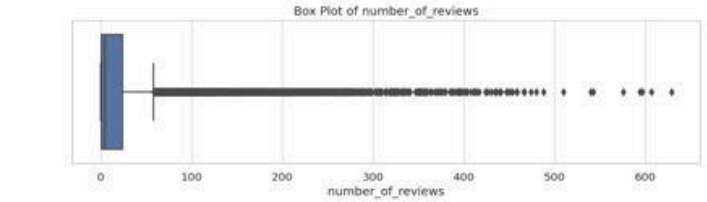
The top pair of plots shows the distribution and box plot for price. The histogram indicates that most of the prices are clustered at the lower end of the scale with a long tail towards the higher prices, suggesting some very high values (possible outliers). The box plot also shows a number of outliers beyond the upper whisker.





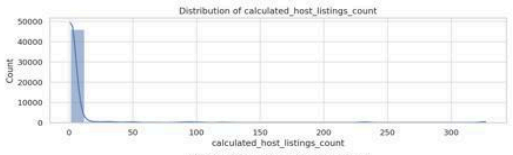
The second pair of plots represents the distribution and box plot for minimum_nights. Similar to the price, most of the data is concentrated at the lower end with a long tail indicating outliers.





The third pair of plots illustrates the distribution and box plot for number_of_reviews. The data heavily skews towards the lower end, meaning most listings have a low number of reviews, with a few outliers with a very high number of reviews.

The fourth pair shows reviews_per_month, which again, has most of its data concentrated at the lower end and a few higher values stretching out to the right.



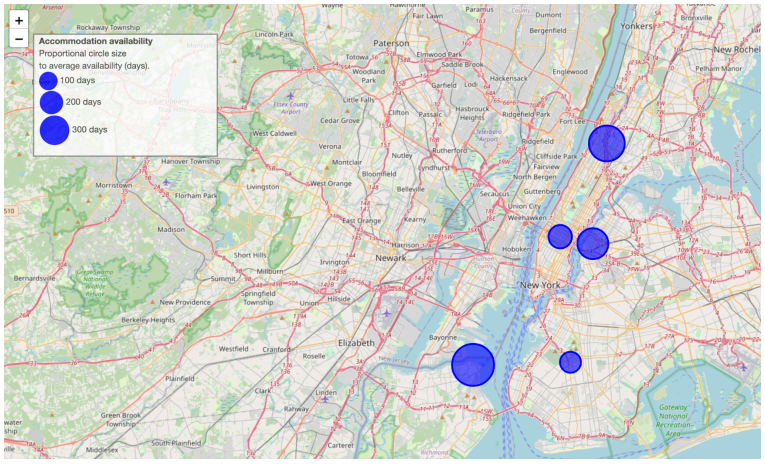
The fifth pair of plots
is for calculated_host_listings_count.
The histogram shows that most hosts have a few listings, with a tail suggesting that there are hosts with many listings. The box plot confirms this with several outliers indicated.

1.3. URL link

The dataset is accessible for free on the Kaggle website following this link.

2. Exploratory Analysis

2.1. Do boroughs in New York City have an impact on Airbnb availability, price and density ? (Théo Belen-Halimi)

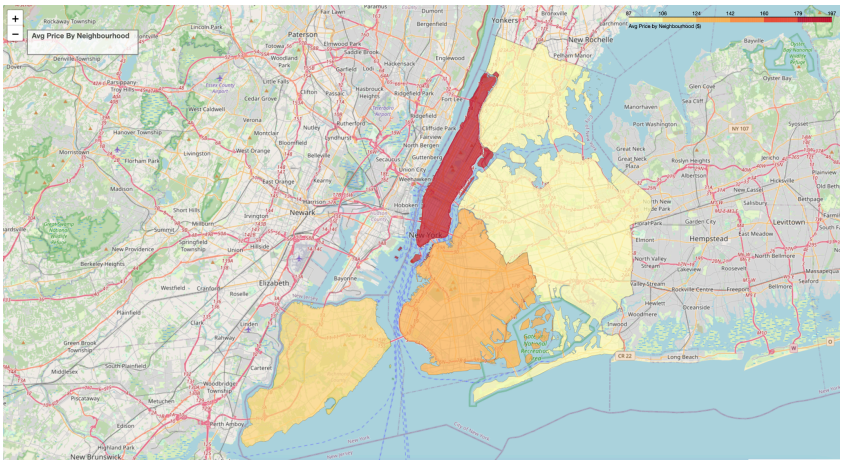


Description of Figure:
This visualization maps the average availability of Airbnb accommodations within New York City's boroughs. Using proportional circles to represent the availability, it reveals variations in how often listings are available for booking, ranging from 100 to 300 days annually. This geographical representation provides a clear, at-a-glance understanding of availability trends across the Bronx, Brooklyn, Manhattan, Queens, and Staten Island.

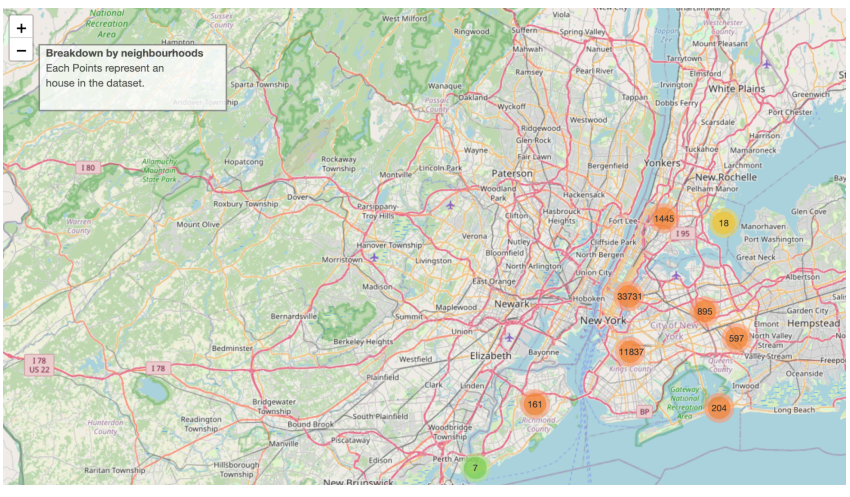
We can notice that Manhattan and Brooklyn are the boroughs the most booked. Two reasons can explain why.

- They both are very touristic boroughs and people usually rent an Airbnb in those areas close to the city center to visit NY.
- Maybe the density of Airbnb in those areas is really small, which explains the fact that they are less available due to high demand and low supply.

We need to explore more on the second hypothesis:



This second visualization highlights the average price per borough and explains the two hypotheses we just made. Tourist areas are usually more expensive and the law of supply and demand (high demand, low supply) makes the price rise. But we may wonder, what's the density of Airbnb in the different boroughs to confirm the low supply theory.



This visualization underlines the density of Airbnb in New York. This visualization allows us to have more information the affirm

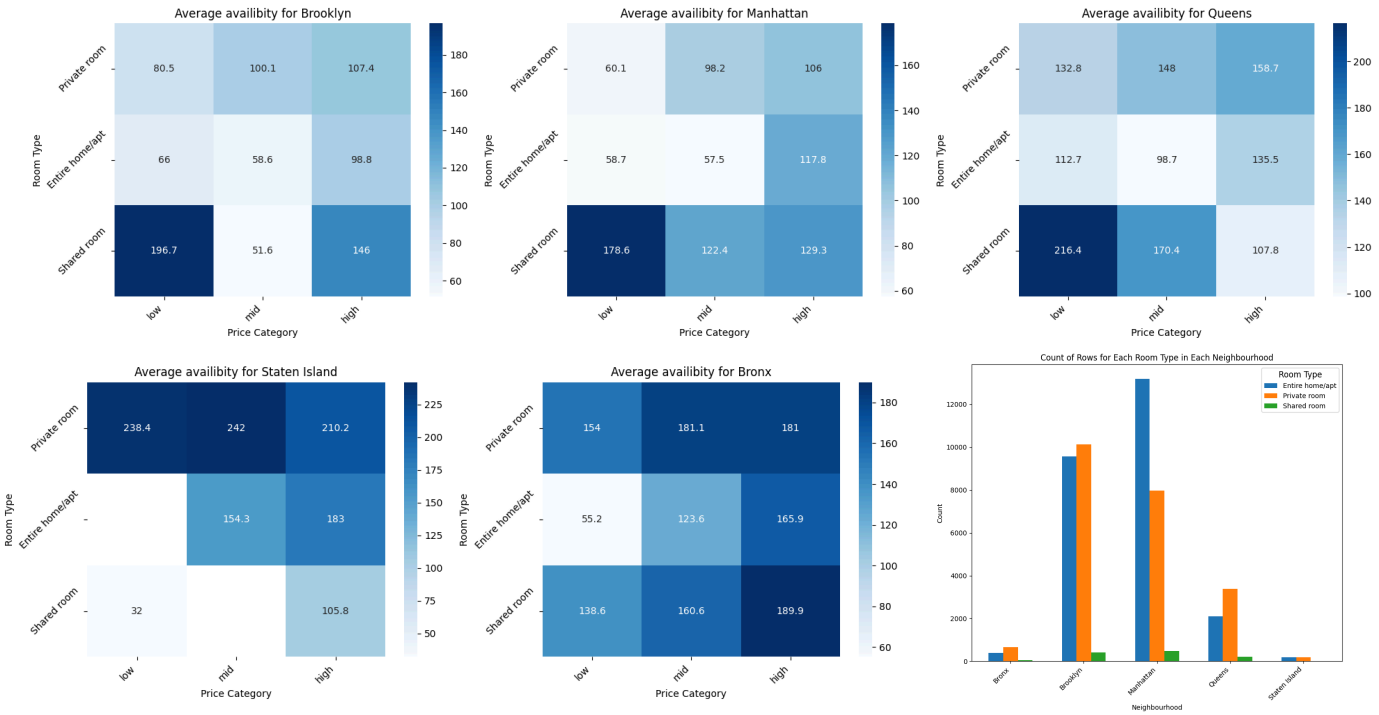
or reject the hypothesis by knowing the supply of Airbnb per borough.

Design Rationale:
The decision to use a map overlaid was driven by the need to convey geographical data intuitively. The size of each circle directly correlates with the average days of availability, offering a straightforward visual comparison across regions. This method is particularly effective for geographic datasets, as it allows the viewer to quickly grasp spatial patterns and variances in availability. By employing this visual encoding, the visualization communicates the dispersion and density of available Airbnb listings in an easily interpretable manner, without necessitating additional explanation.

Storytelling
In this visualization, the focus is on elucidating the patterns of Airbnb availability across New York City's boroughs, aiming to unveil how these patterns might influence visitor accommodation choices. By highlighting the differential availability—from highly accessible to more limited—this graphic tells a story of urban dynamics, visitor demand, and the spatial distribution of Airbnb listings. It seeks to inform viewers about potential areas of high or low competition for bookings, and how these dynamics could reflect underlying economic, social, or regulatory factors affecting short-term rental markets in an urban context.

2.2. What are the correlations between price, availability and room types per neighborhood in the NYC airbnb market ?(Olivia de la Chapelle)

According to me, answering the question above helps clients and hosts understand better the dynamics behind the current market to make wiser decisions.

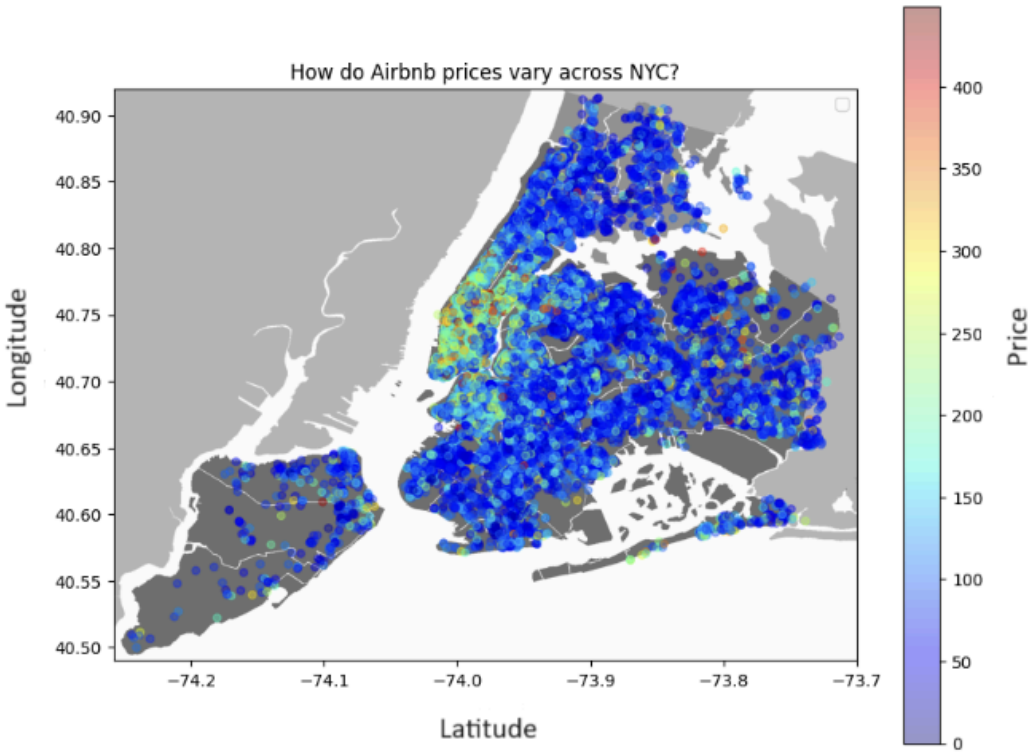


Average availability in days of airbnb accommodations in NYC

Description and storytelling:
My objective was to form a very easy to read and straightforward visualization while being able to verify as many hypotheses as possible from the four more important variables. From these dense heatmaps, it is possible to study relationships between 4 dimensions.
Let's consider the context of top NYC neighborhoods (Manhattan, Brooklyn, Queens) being popular and expensive. Accommodations are concentrated in these areas, as they meet airbnb' target preferences being tourist-friendly locations. Low-priced shared rooms are particularly available which indicates a low demand for this category. For other room types though, patterns emerge in top neighborhoods. As city living costs are high, low-priced private rooms are particularly not available. Conversely, mid-range priced entire homes are the least available. Cheaper ones might be too tiny, suggesting that larger groups seek a minimum of comfort either by paying more or going for more affordable housing in other neighborhoods such as the Bronx and Staten Island. These show high availability, attracting less clients in general on the platform.

Design Rationale:
When looking at accommodation availability, I categorized the data into three categories to simplify future analysis as a big part of accommodations are almost never available and the rest is homogeneously distributed. Similarly, I defined price categories (low, mid, and high) to identify clusters in the heatmaps, considering a price distribution centered around $89 per day but could go up to $10,000. Even though geographical maps are more appealing, I decided to use heat maps as it helped me analyze more data at the same time while using real values to compare them.

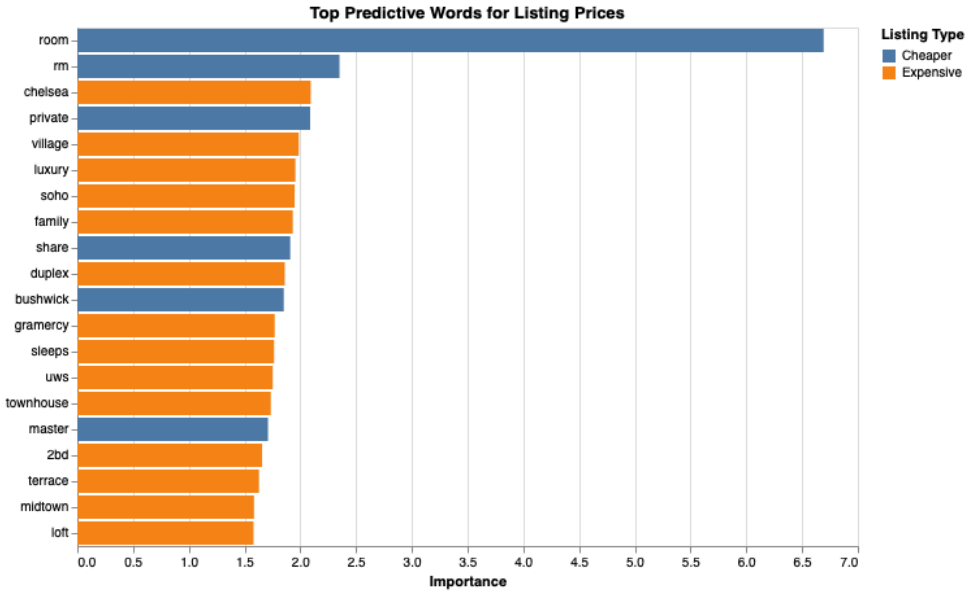
2.3. How do AirBnB prices vary across NYC? (Nithin Alfred)



Description and storytelling:
The figure is a vivid depiction of the Airbnb pricing landscape across New York City. In this visual narrative, I'm keen to illustrate the diversity and range of accommodation prices available to travelers. The varying hues from deep blue to intense red across the map signify a price escalation, with cooler tones representing more economical options and warmer tones signaling premium lodging. This scatter plot, superimposed on a grayscale map, allows the viewer to spot at a glance the hotspots of high-priced rentals, predominantly clustered in Manhattan, as well as the more affordable stays, which are widely dispersed. It's a story of economic geography, where the vibrancy of urban life intersects with the affordability of staying in one of the world's most bustling metropolises.
(In an effort to avoid outliers and make the graph more visually appealing, I subsetted the dataset to remove above the 98th percentile of prices)

Design Rationale:
From a design standpoint, my visualization leverages color as a primary encoding mechanism, capitalizing on its intuitive association with temperature to denote price intensity. The choice of a jet colormap, transitioning smoothly from cool to warm shades, is deliberate; it taps into a universally understood color spectrum to convey increasing value. The underlying map's muted tones ensure that the colored data points stand out, drawing immediate attention. The decision to use a scatter plot aligns with the geographical nature of the data, mapping the exact locations of Airbnb listings onto the city's layout. This design choice facilitates an effective communication of spatial patterns and price distributions, enabling viewers to draw insights into the economic geography of Airbnb listings in New York City.

2.4. What is the impact of descriptive keywords on the price ranges of NYC Airbnb listings? (Alexandra Pricop)



Description and Storytelling:

These visualizations explore the influence of language on the perceived value of Airbnb listings in New York City. The word clouds highlight the most frequent terms found in the names of Airbnb listings, distinguishing between those associated with higher-priced and lower-priced options. The bar chart complements this by quantifying the importance of these terms in predicting the price category. Collectively, they tell a story of how hosts use specific descriptive terms to position their listings within the market: terms like "luxury," "Chelsea," and "duplex" are prevalent in higher-priced listings, whereas "comfort," "central," and "midtown" appear to denote more affordable options. This suggests that hosts strategically employ language that resonates with the expectations of their target audience, potentially influencing the listing's appeal and profitability.

Design Rationale:

The word clouds use size and color to indicate the frequency and category (expensive or cheaper) of words, providing an immediate visual representation of their significance. This allows for easy comparison of which features are emphasized in different price ranges. The bar chart employs horizontal bars to quantify this relationship, with color coding for quick categorization and length representing the relative importance of each term. The decision to use horizontal bars enables the inclusion of longer descriptive words without compromising readability. Both visualizations employ a color scheme that differentiates the two price categories while maintaining aesthetic harmony across the visuals. By aligning the word clouds with the bar chart, viewers can cross-reference the qualitative impact of words with their quantitative contribution to the pricing model, thus enhancing comprehension and reinforcing the narrative.

3. Findings

The different visualizations and analysis made from each member of the group helped identify different but complementary valuable insights across various dimensions about data repartition and the dynamics between accommodation offers on Airbnb in NYC.

While analyzing the variation of prices across NYC, Nithin highlighted dense clusters of high-priced Airbnbs in Manhattan, especially around tourist hotspots and commercial centers. The outer boroughs and certain areas of Brooklyn featured more economical options, with prices often below $100 per night. The color distribution on his visualization reflects the economic landscape of NYC neighborhoods in the context of short-term rentals.

Théo explored the scarcity of Airbnb options in Manhattan and Brooklyn, hypothesizing that high demand in these touristic areas leads to reduced availability. Contrary to the initial belief that a limited number of listings might be the cause, further investigation revealed that Manhattan and Brooklyn host 90% of New York City's Airbnb listings, dismissing the limited supply theory. With the highest average nightly rates of $197 and $142, respectively, the

scarcity and elevated prices in these boroughs are primarily due to the high demand for lodging in these sought-after locations, rather than a shortage of available properties.

Olivia's examination of the correlation between price, availability, and room types per neighborhood provided additional insights. Apart from the specificity of “shared rooms” that do show demand from the client side, in top and more expensive neighborhoods (Manhattan, Brooklyn, and Queens) low-priced private rooms and middle-range entire homes were the most researched and least available ones. Although most accommodations are not from the Bronx and Staten Island, they gather the higher availability, meaning that both neighborhoods are not that attractive to the Airbnb public. The target clients must have a smaller paycheck to offer and are redirected to this area of NYC. Consequently, high price accommodations are particularly available in these areas.

Leveraging text, Alexandra studied the impact of keywords on listing prices, using word cloud visualization and regression analysis. Luxury and exclusive amenities featured prominently in higher-priced listings, aligning with the geographical analysis that highlighted prime locations and their correlation with higher prices. This analysis reinforced our initial hypothesis about critical factors influencing Airbnb pricing in NYC.

In summary, we were able to understand the general intricate dynamics of Airbnb listings in NYC. From the influence of demand and supply on availability to the impact of keywords on pricing and the variation of prices across neighborhoods, the findings provided a comprehensive understanding of the Airbnb market in New York City. These insights directly relate to the initial topic of exploring and understanding the factors influencing Airbnb availability, pricing, and patterns across different neighborhoods in NYC.