Team Project Report

Machine Learning August 2023

Ruth Allison, Nicholas Bandy, Tasweem Beelunkhan, Danilo De Sousa, Lojayne Diab

# Introduction

Airbnb operates as an online intermediary for property rentals, earning commissions from bookings (Airbnb, 2023). While property owners determine their own pricing, comprehending the local market dynamics is crucial for success. This report aims to identify the most successful neighbourhood groups based on property concentration and occupancy rates. Such insights will guide both current and potential hosts in making informed decisions for their ventures.

# Data

The dataset used was rentals in New York City during 2019. Each entry pinpoints a property's geographical location, type, price per night, and minimum stay requirements. These rentals span neighbourhoods, with each borough housing between 32 to 51 distinct areas. The dataset's depth extends to financial aspects and availability metrics, indicating potential rental opportunities throughout the year.

To sharpen the focus of the analysis, some variables were set aside. Descriptive elements such as property names, primarily used for advertising, were excluded. Information about hosts, including their ID, name, and listing count, was deemed peripheral given the property-centric nature of the study. Similarly, review-related metrics, which aren't mandatory for renters, were omitted for their potential unreliability in reflecting a property's success.

A crucial assumption shaped the understanding of the availability\_365 column. Instead of viewing it as the total nights a property was open for booking, it was interpreted as nights remaining post existing reservations. Thus, booked nights were calculated by deducting this value from 365. This dataset offers a glimpse into New York City's 2019 rental trends, presenting a vivid picture of available options and market dynamics.

# Methodology

The dataset consists of 10 numeric columns and 6 categoric columns, 4 of which contain null values as shown in Figure 1. Host and listing identification columns are not required, therefore Id, name, host\_id, and host\_name are dropped from the dataframe. Last\_review is dropped as discussed above. Null values in the reviews\_per\_month column is replaced with 0 as they indicate that there has not yet been a review.

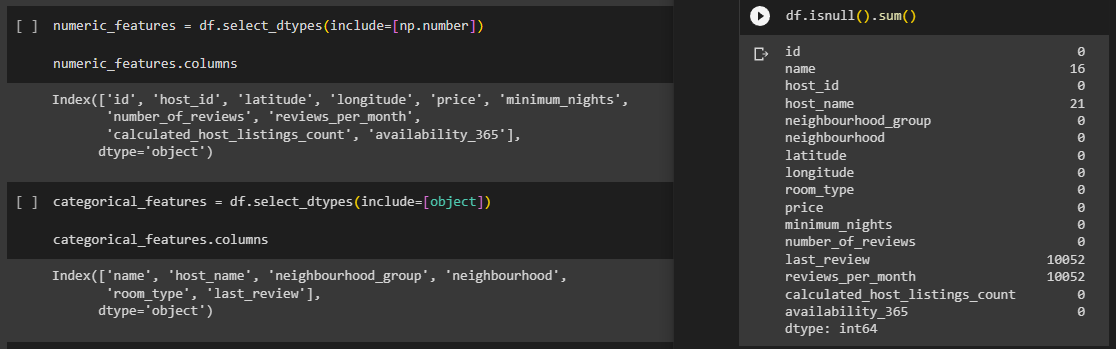
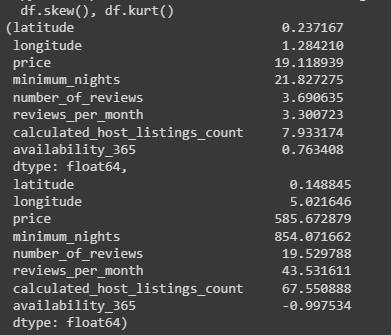
*Figure 1: Data types and Null Values*

Figure 2 shows the skewness and kurtosis on the remaining numeric columns. Price and minimum nights show the highest skew values, with number of reviews, reviews per month, and host listing count also having significant skew. This indicates that there is a lack of symmetry in these columns. Similarly, these columns also show the highest kurtosis indicating a heavy tail and peakedness relative to the normal distribution in these columns (DeCarlo, 1996). All the other columns appear symmetrical with weak tails.



*Figure 2: Skewness and Kurtosis results*

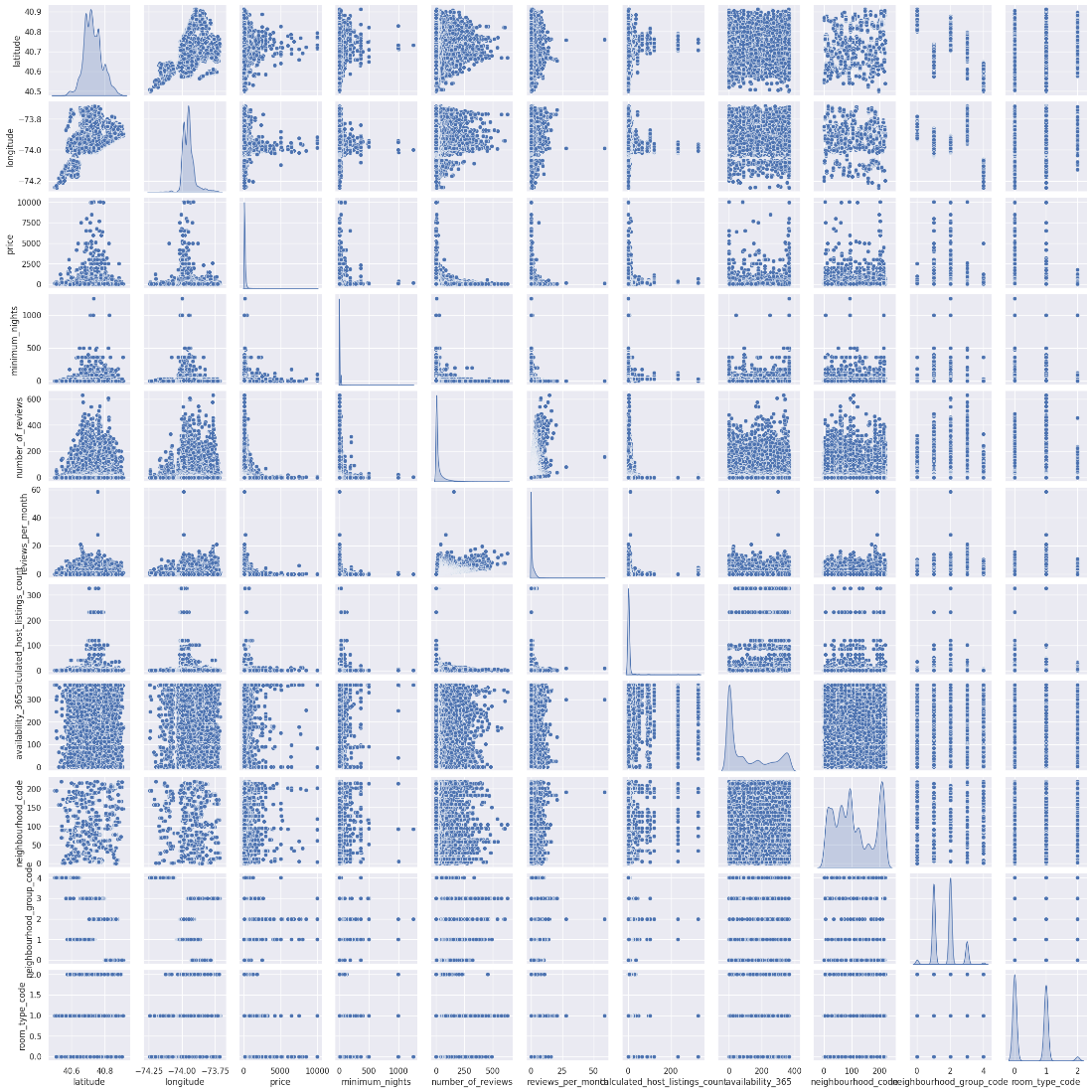
A screenshot of a computer screen

Description automatically generated

*Figure 3: Correlation matrix*

A correlation matrix (Figure 3) is used to investigate the dependence between multiple numerical variables at the same time. The heatmap's colour spectrum ranges from blue (negative correlation) to red (positive correlation). No strong correlation is seen between numerical variables however much of the data is categorical

Numeric codes are added for the remaining categoric columns to view their relationship with other columns within the pair plot shown in Figure 4. The pair plot shows that the price, number of reviews, reviews per month, and the number of listings a host has vary greatly between different neighbourhood groups and different room types. The number of reviews, reviews per month, and number of listings a host has all show a decreasing trend as the price increases.



*Figure 4: Pairplot for bivariate analysis*

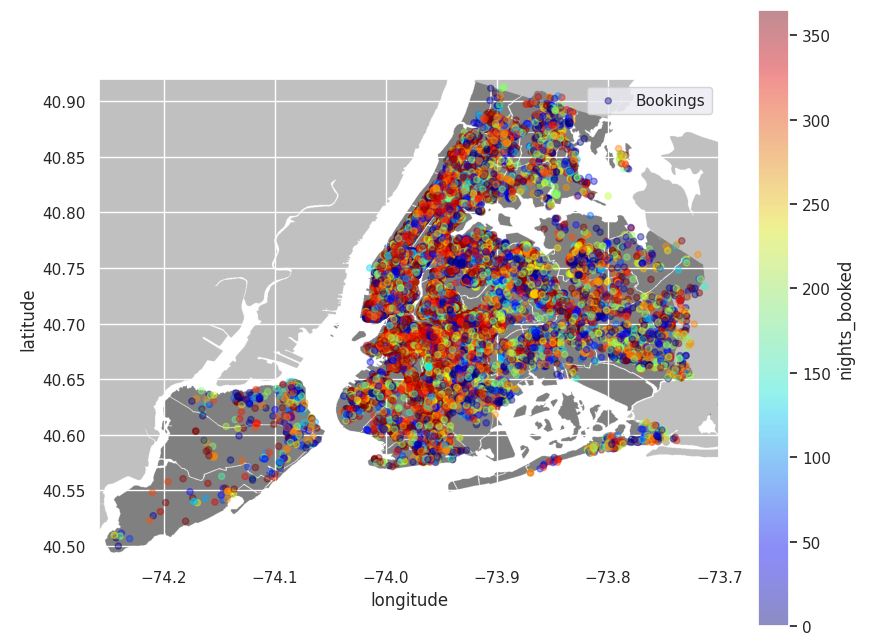
Table 1 shows that Manhattan has the highest average price at $196.88 per night, indicating it is the most expensive neighbourhood group. Brooklyn follows closely with an average price of $124.38 per night, making it the second most expensive.

Brooklyn has the highest average number of nights booked throughout the year at 265, suggesting that listings in this area are very popular. Staten Island has an average number of nights booked of 165, indicating that the area is less popular, possibly because it is further from the main tourist areas.

*Table 1*

|  |  |  |
| --- | --- | --- |
| **Neighbourhood group** | **Mean price ($)** | **Mean number of nights booked** |
| Bronx | 87.50 | 199 |
| Brooklyn | 124.38 | 265 |
| Manhattan | 196.88 | 253 |
| Queens | 99.52 | 221 |
| Staten Island | 114.81 | 165 |

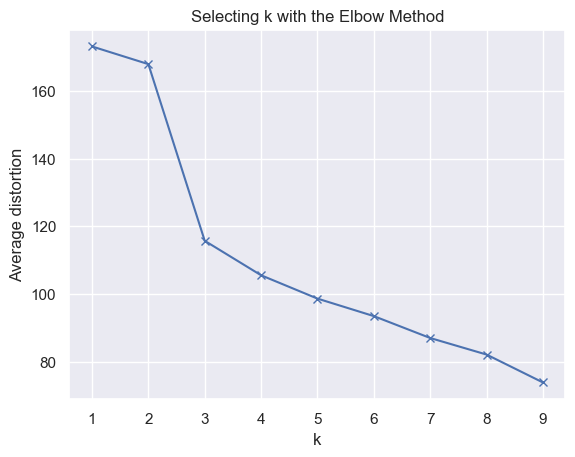
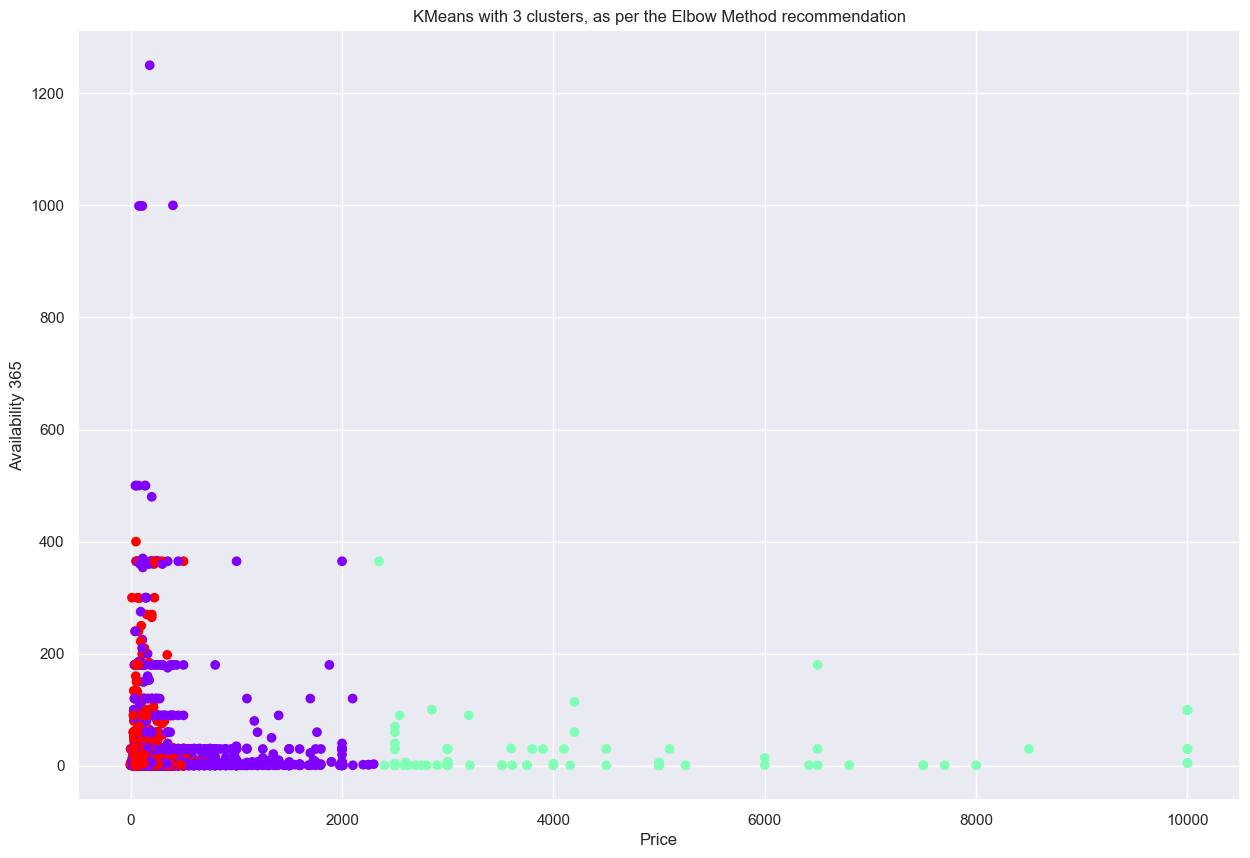
Figure 5 shows the bookings made plotted onto a map of New York City alongside the five boroughs as context for determining which neighbourhoods were most successful.

 A map of the united states

Description automatically generated

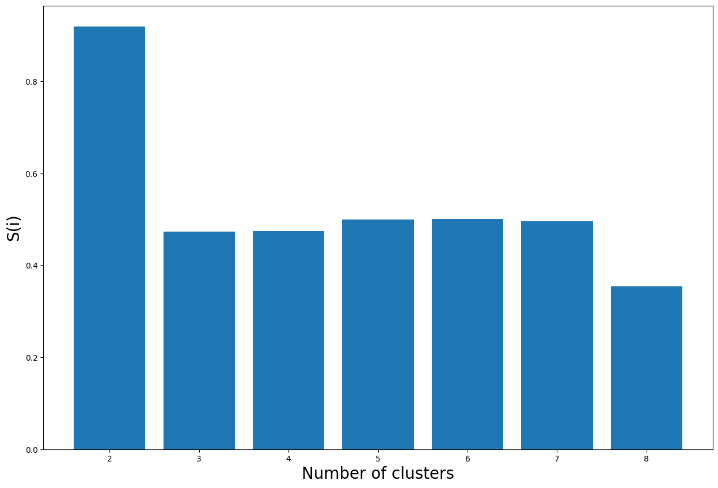
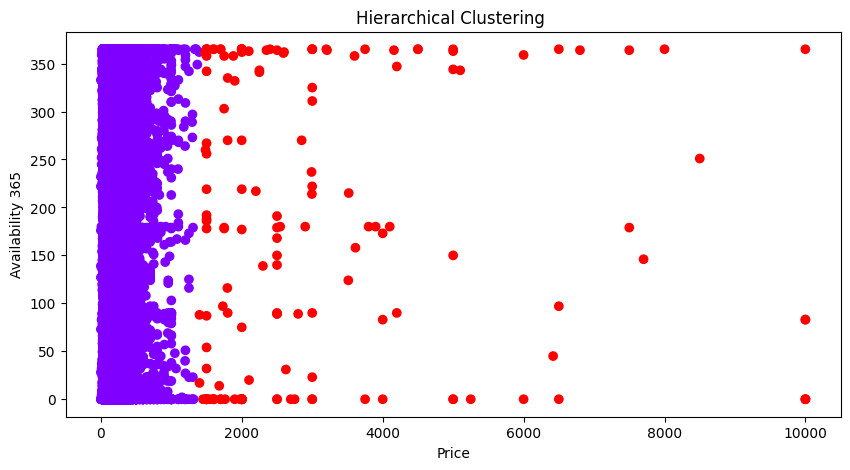
*Figure 5: Plot of bookings by location*

An initial attempt to determine the optimum number of clusters was carried out by plotting SSE against number of clusters which gave a k value of 3 (Figure 6) but when these were plotted a large amount of overlap was seen (Figure 7).

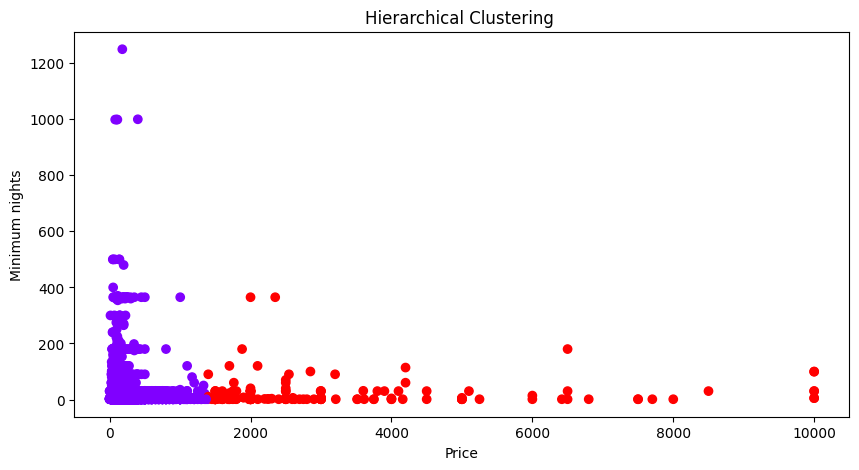
*Figure 6: Elbow method to determine k Figure 7: Plot showing four clusters*

To address the challenge of data sparsity when segmenting New York City listings required a nuanced approach. We chose the Agglomerative Clustering algorithm, a hierarchical clustering method known for its efficiency in handling sparse data (Vijaya et al., 2019). Furthermore, we applied a conversion-based method to transform categorical data into numerical representations before clustering (Bai & Liang, 2022). For the critical task of identifying the optimal number of clusters, we relied on the Silhouette Coefficient metric, a reliable measure of cluster cohesion and separation, as shown in Figure 8 (scikit-learn, 2023).

*Figure 8: Silhouette coefficient values*  *Figure 9: Plot of hierarchical clustering: price and availability*

Utilizing the Silhouette Coefficient as our metric, we determined that the ideal number of clusters for our dataset is two, showcasing a clear separation of the segments as per Figures 9 and 10.



*Figure 10: Plot of hierarchical clustering: price and minimum nights*

Thus, the chosen methods and tools effectively segmented the listings, providing clear insights into the structure and distribution of New York City rental data.

# Conclusions

In conclusion, our research highlighted that in 2019, Manhattan, with its attractions like Broadway and Central Park, was a prime choice for renters. Brooklyn, especially areas adjacent to Manhattan, also garnered significant attention. The combination of iconic landmarks in Manhattan and the competitive pricing of neighbouring Brooklyn made these regions the most successful rental destinations in New York City.

# References

Airbnb. (2023) How much does Airbnb charge hosts? Available from: <https://www.airbnb.co.uk/resources/hosting-homes/a/how-much-does-airbnb-charge-hosts-288> [Accessed 8 September 2023]

Bai, L. & Liang,J. (2022) A categorical data clustering framework on graph representation. *Pattern Recognition* 128(1):108694. DOI: <https://doi.org/10.1016/j.patcog.2022.108694>

DeCarlo, L. T. (1997) On the Meaning and Use of Kurtosis. Psychological methods. [Online] 2 (3), 292–307. Available From <https://web-s-ebscohost-com.uniessexlib.idm.oclc.org/ehost/pdfviewer/pdfviewer?vid=0&sid=51b16d11-8f19-40e6-a38e-993423b297d6%40redis>

Scikit-learn. (2023) Selecting the number of clusters with silhouette analysis on KMeans clustering. Available from: <https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html>

Vijaya, Sharma, S. & Batra,N. (2019) Comparative Study of Single Linkage, Complete Linkage, and Ward Method of Agglomerative Clustering. *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)* 1(1): 568-573. DOI: <https://doi.org/10.1109/COMITCon.2019.8862232>