Analytical Report

**Insights from Airbnb Rentals in New York City**

# Introduction

This report aims to explore the results of an analysis conducted on New York City Airbnb rental data from 2019, in which three key business questions were answered through in-depth exploratory data analysis (EDA) and the use of machine learning (ML). The insights gained will also be discussed in this report, as well as recommendations. For reproducibility, a simplified version of the CRISP-DM process was used in this analysis.

# Data understanding

The “New York City Airbnb Open Data” dataset was used in this analysis, which offers a snapshot of Airbnb listings and metrics in New York City in 2019. It is available at https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data.

This dataset contains 16 columns, with 10 of them being numerical (notably, 2 of them are ID-based) and 6 being categorical. This dataset covers diverse aspects of each listing, such as host name, latitude and longitude, reviews per month, room type, neighbourhood, rental availability in the next year and more, providing a solid basis for understanding the city’s Airbnb market.

# Data preparation

An initial analysis of this dataset revealed a total of 10,074 null rows (20.6% of all records) that were concentrated in the columns that represented the rental name, host name, last review and reviews per month. As removing these rows from the dataset would have resulted in data loss, it was decided to fill those null values with appropriate replacements, such as “Unknown” and “Never” for null names and dates.

The next step involved the transformation of the categorical columns. Conducting numerical operations on text data is not very efficient, which led to the creation of a mapping structure to assign numerical values to categorical data.

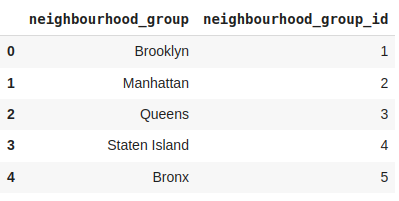


Figure 1: The mapping process applied to the neighbourhood group column

# Exploratory data analysis

Considering the scope of the data, uncovering rental patterns would be of interest to Airbnb. As such, the following three guiding questions were posed:

* How does location and room type affect the cost of the rental?
* How does availability over the next year correlate to room type and location?
* What is the distribution of listings between hosts and property types?

The first step to determine underlying relationships in the data was to calculate the dataset’s correlation. Before doing so, it was important to determine whether the data followed a normal distribution, given correlation methods depend on this property.

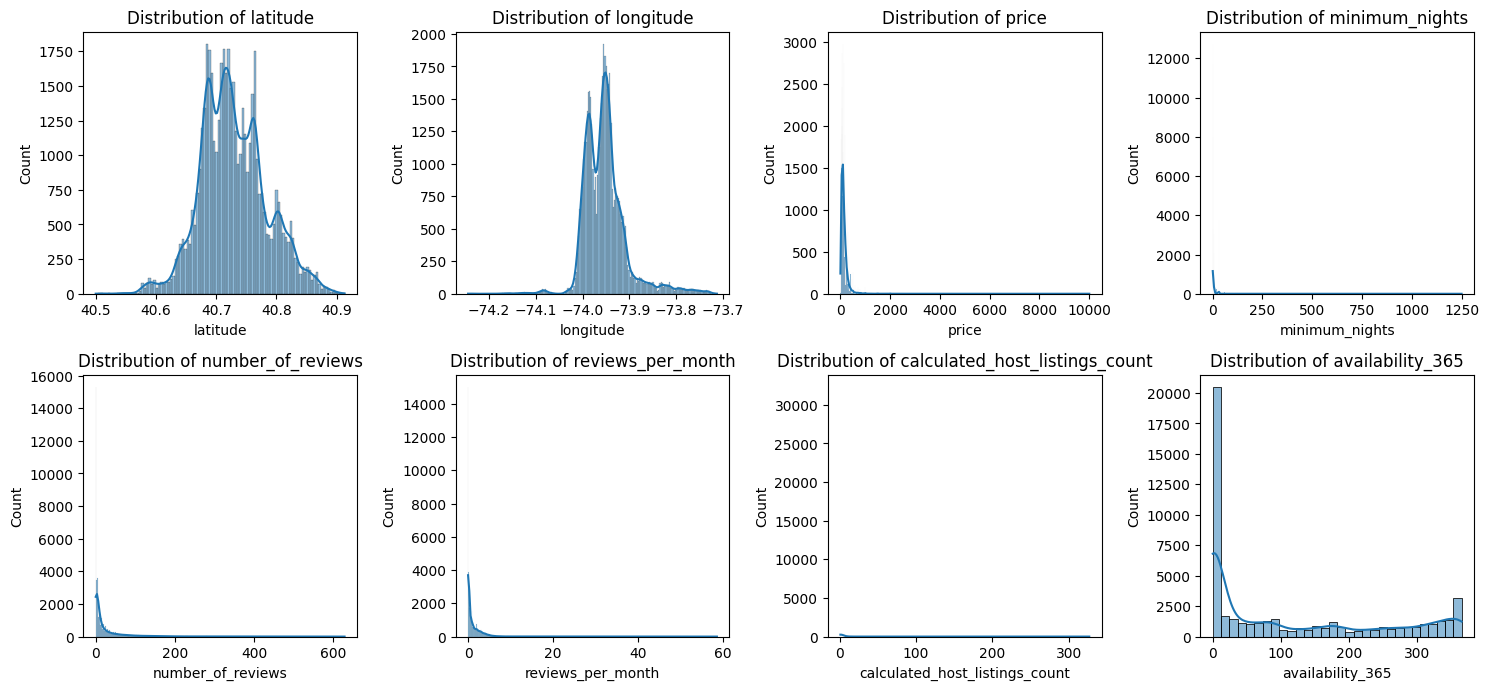
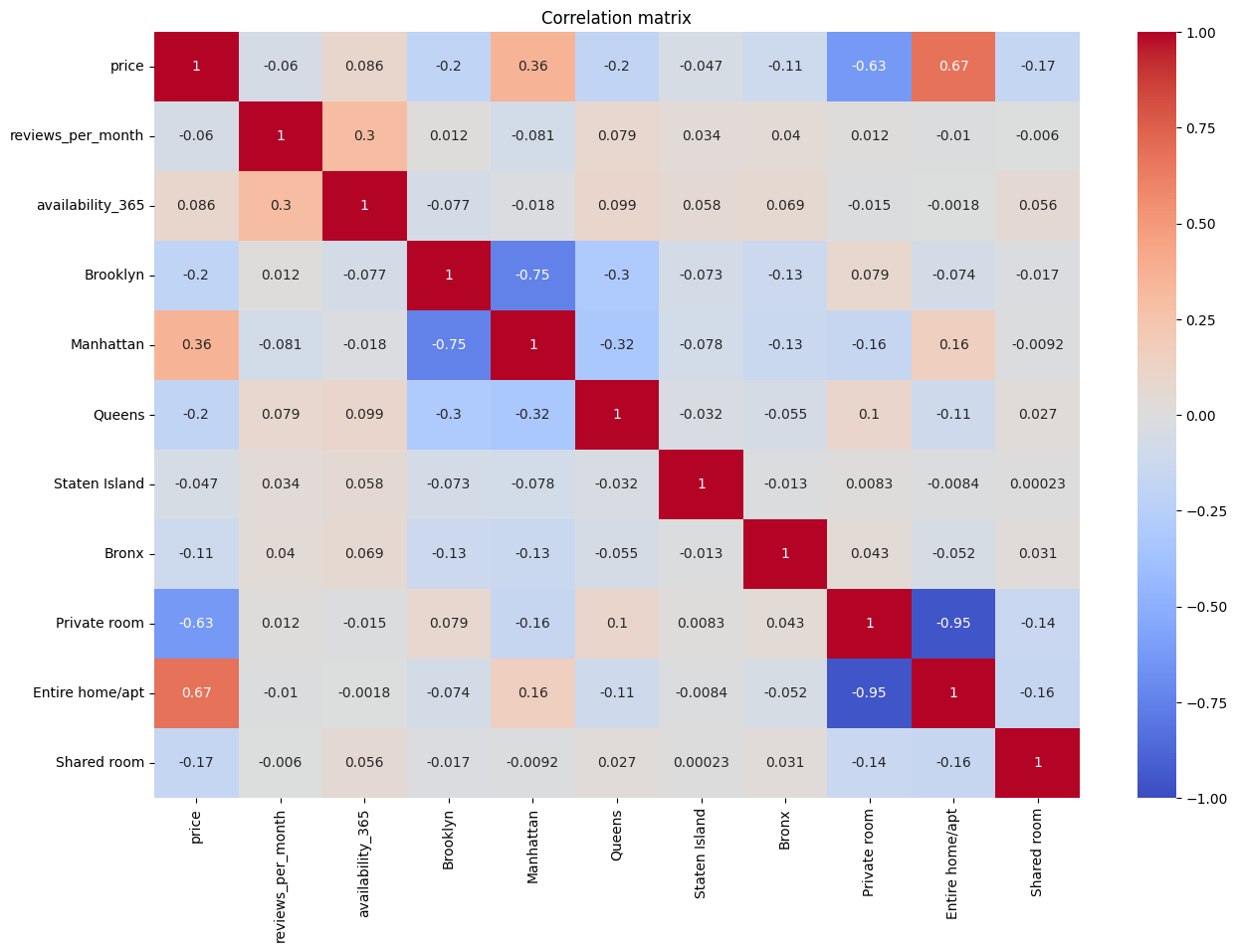


Figure 2: Numerical data distribution (excluding ID data)

As seen in the figure, the data did not follow a normal distribution, which meant that correlation methods like Pearson’s correlation were unsuited to this data (Schober et al., 2018), thus necessitating a different approach.

## Correlation analysis

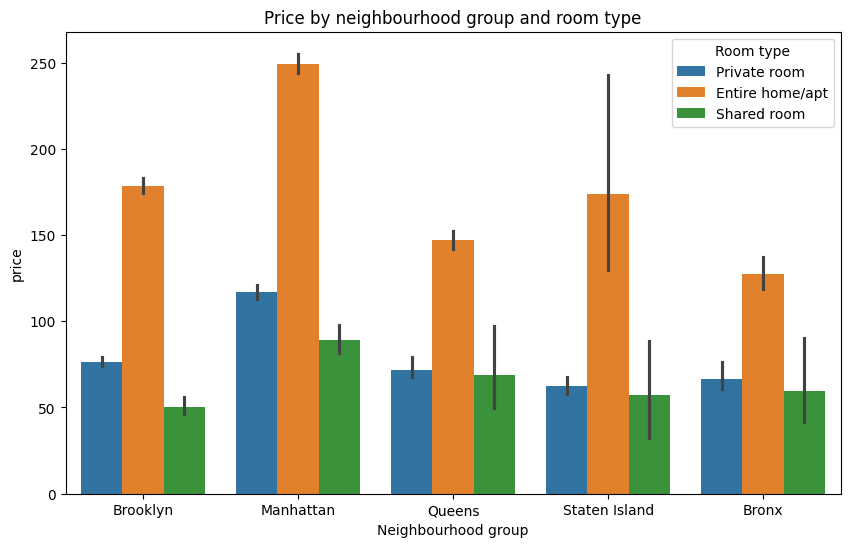
Point-biserial correlation was used, which is the value of Pearson’s product correlation when one of the variables is dichotomous (i.e., a categorical variable with two possible values) and the other variable is metric (Kornbrot, 2005). Before applying it, the categorical columns were first encoded via one-hot encoding, which converts them into binary features.

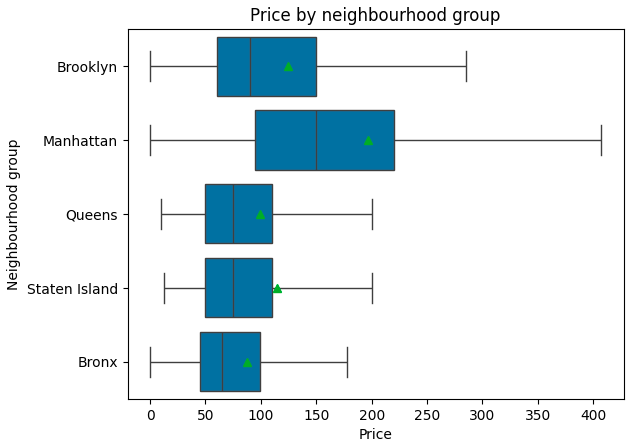
Figure 3: Correlation matrix, showing the various relationships between the dataset’s neighbourhood groups, price and room types

The most important records for this analysis were focused upon. Price is seen to correlate with all neighbourhood groups and has a particularly strong positive relationship with Manhattan. It also has a correlation with all rental types, thus hinting that these variables share a mutual relationship.

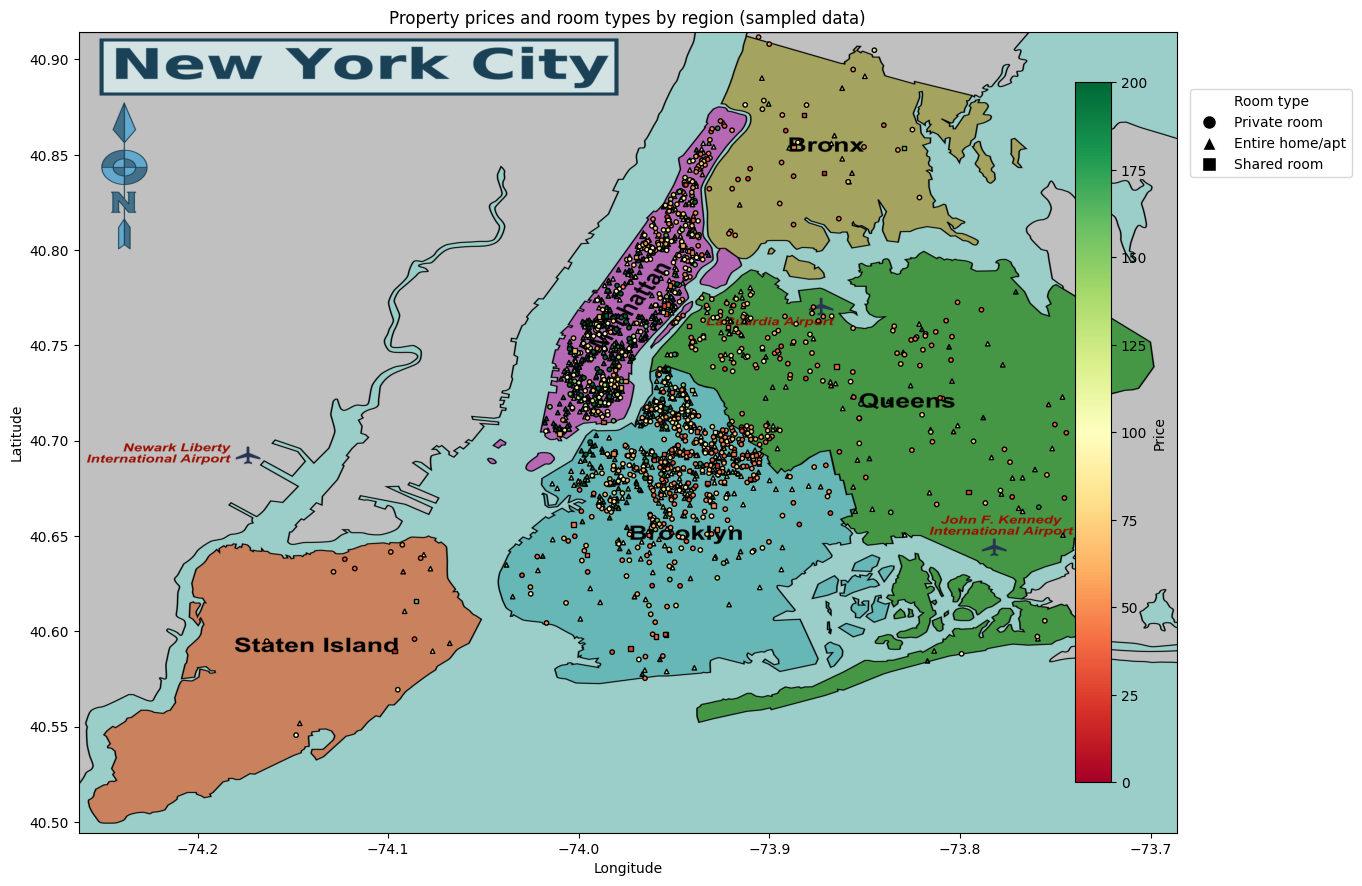
## Rental patterns

The first insights were gained on rental prices and its influencing variables. Using a visual approach was instrumental in evidencing the relationships shown in the correlation analysis.

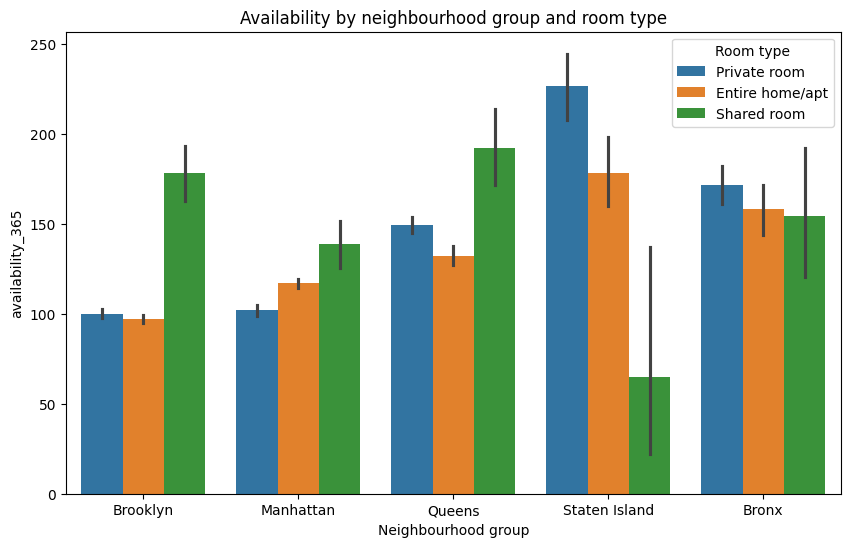
Figure 4: Rental prices by neighbourhood group and room type

Figure 5: A box plot showing rental prices by neighbourhood group and without outliers

The figures above evidence the role that neighbourhood group and rental type have in influencing rental prices, with entire home/apartment having the highest prices across all types and Manhattan being the most expensive location out of all neighbourhood groups.

Figure 6: A map of New York City showing a sample of two thousand rentals. The high concentration of rentals in Manhattan (top) and Brooklyn (centre) is apparent, as is the presence of expensive options (shaded in green) (adapted from NYC Map 360º, N.D.)

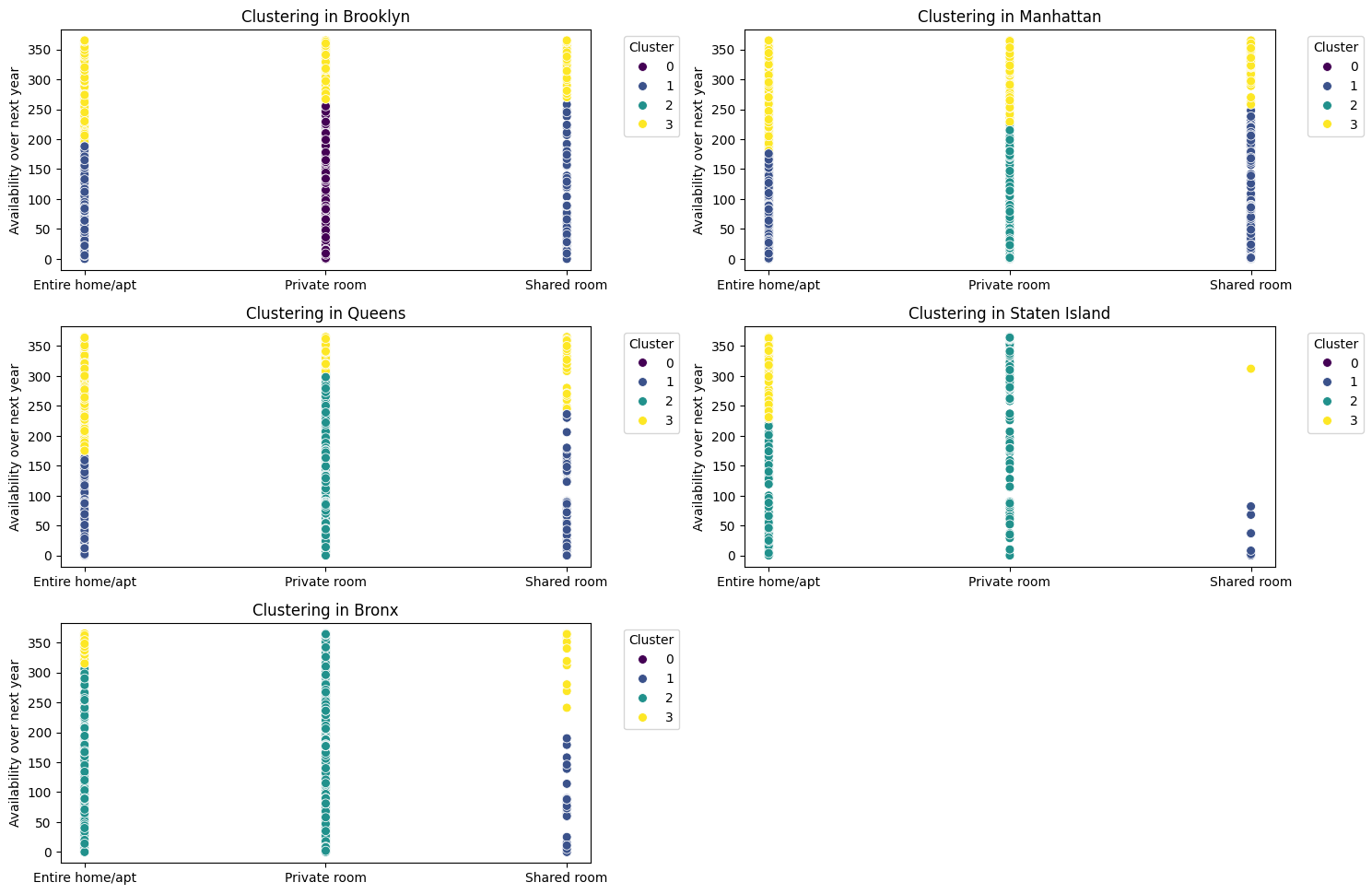
Shifting the focus to availability over the next year (the “availability\_365” column), the correlation analysis suggested that room type and neighbourhood group play a role in influencing this variable. Once more, a visual approach evidenced these relationships.

Figure 7: Rental availability by neighbourhood group and room type

General availability in Manhattan and Brooklyn over the next year is lower than in other locations. Interestingly, some outliers seem to indicate the role that other circumstances, such as price, play in availability.

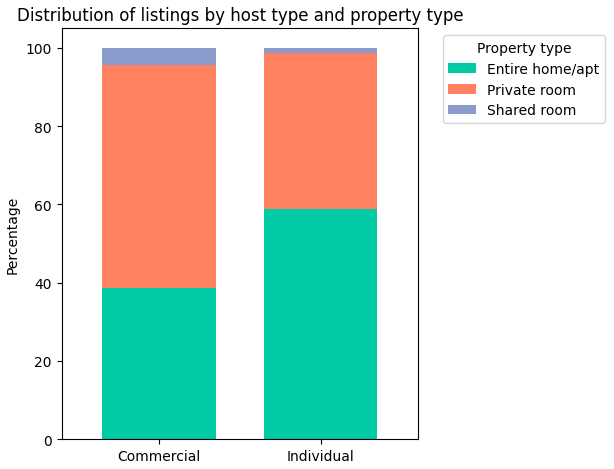
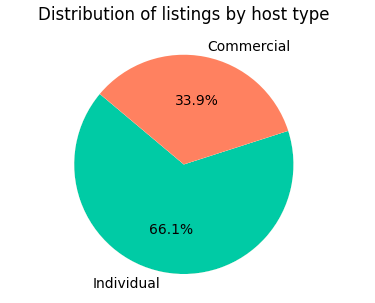
Using a clustering approach, it’s possible to group the data based on availability, room type and neighbourhood group. The following clusters have been defined for this task:

* **Cluster 0 (purple):** listings which are likely fully booked or have limited availability, possibly being the most popular;
* **Cluster 1 (blue):** listings with moderate availability, suggesting these are booked regularly but still have open days**;**
* **Cluster 2 (green):** listings with generally higher availability, such as those with low demand or used seasonally;
* **Cluster 3 (yellow):** listings with highest availability, representing those that are rarely booked or are consistently available, possibly due to premium prices.

Figure 8: Clustering in various neighbourhood groups by availability over the next year and room type

This analysis suggests that Brooklyn and Queens tend to have the highest demand for shared rooms and private rooms, while Staten Island and the Bronx generally have lower demand, particularly for entire homes. Manhattan shows a mixed pattern, with higher availability for entire homes, possibly due to higher prices.

When evaluating the distribution of listings between hosts and property types, visualisation can once more be resorted to as a reliable and effective method to generate insights.

Figure 9: Distribution of listings by host and property types

As shown in the figure, individual hosts manage most listings (66.1%), while commercial hosts account for about one-third. Among property types, entire homes/apartments and private rooms dominate, with shared rooms being rare.

# Conclusion and recommendations

The analysis shows location and room type drive rental prices, with Manhattan commanding the highest rates and Brooklyn and Queens sustaining demand for private and shared rooms. Listings in Manhattan could benefit from competitive pricing to improve occupancy rates. Budget-focused marketing in Brooklyn and Queens and dynamic pricing across neighbourhoods can optimize performance in New York City’s rental market.

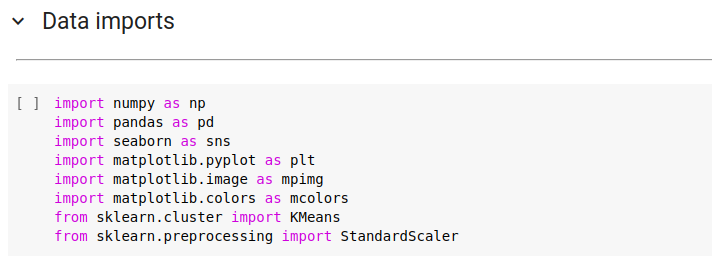
# References

Schober, P, Boer, C. & Schwarte, L.A. (2018) Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia* 126(5): 1763-1768. DOI: https://doi.org/10.1213/ANE.0000000000002864.

Kornbrot, D. (2014) ‘Point Biserial Correlation’, in: N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri & J.L. Teugels (eds). *Wiley StatsRef: Statistics Reference Online.* Hoboken, New Jersey: John Wiley & Sons, Inc. 1-2.

NYC Map 360º. (N.D.) New York City Boroughs & Neighborhoods Map. Available from: https://nycmap360.com/nyc-boroughs-map (Accessed 27 November 2024).

# Appendix

 Figure 10: Importing necessary packages

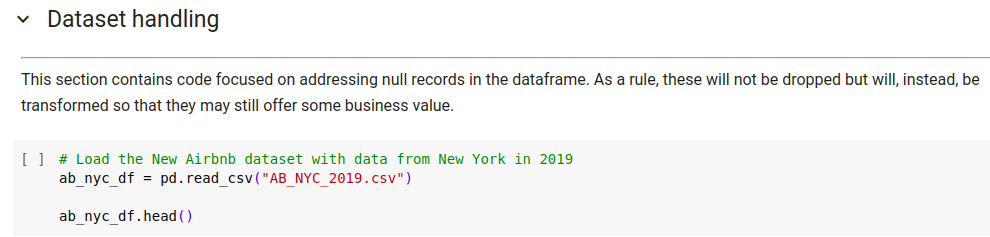
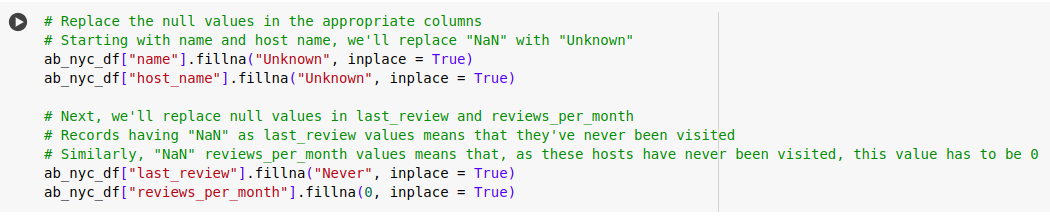
Figure 11: Reading the 2019 New York City Airbnb rentals dataset

Figure 12: Checking the number and proportion of null records in the dataset

Figure 13: Replacing null data with appropriate values

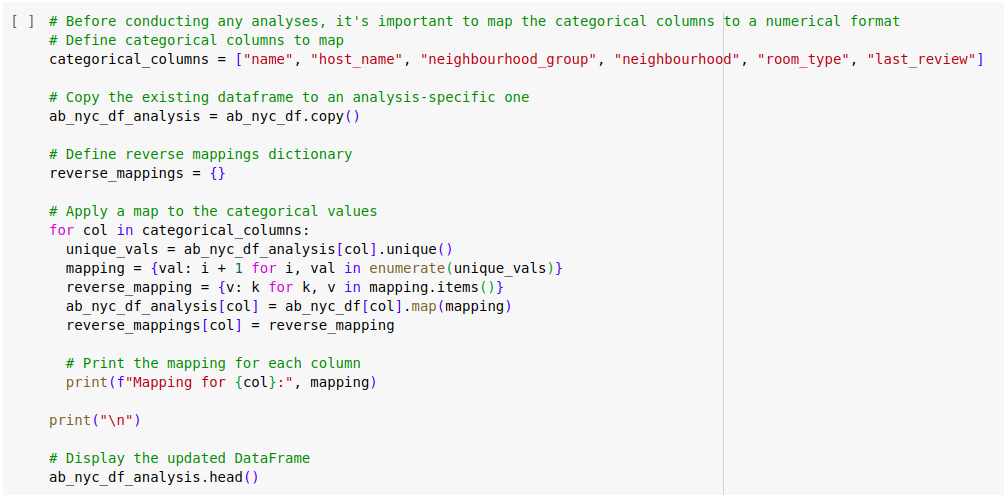
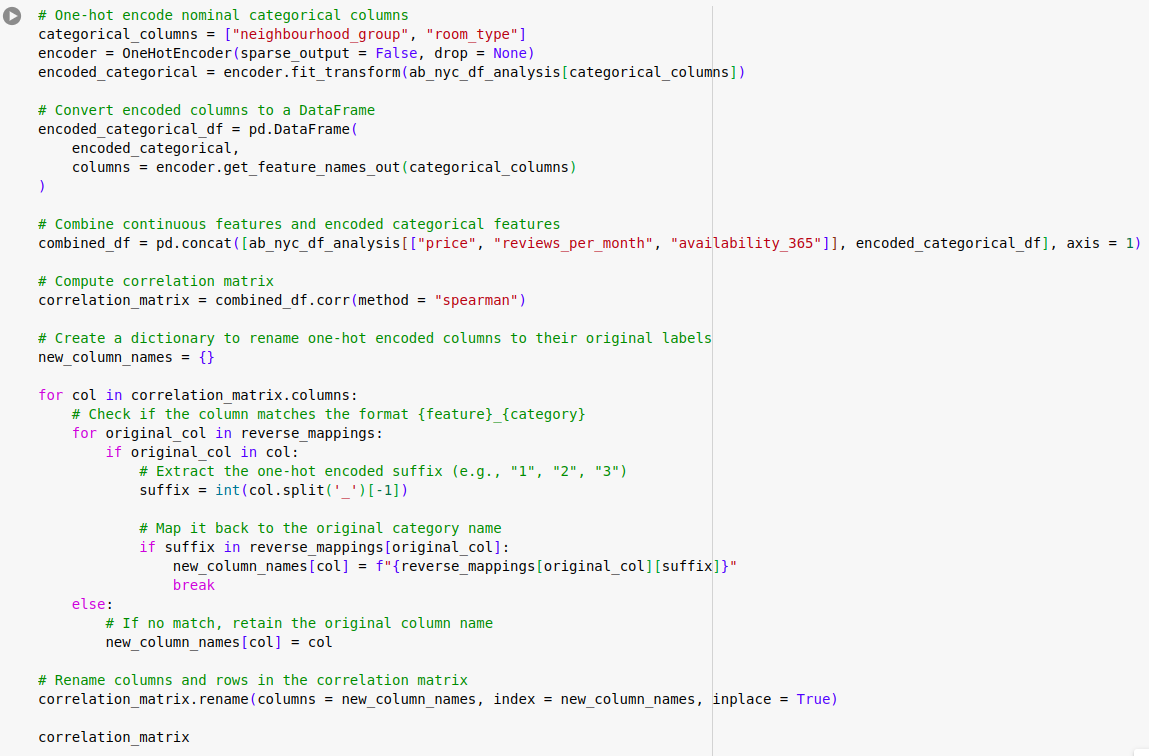
Figure 14: Mapping categorical columns to a numerical format

Figure 15: Creating a plot for data distribution analysis

Figure 16: Creating a correlation matrix

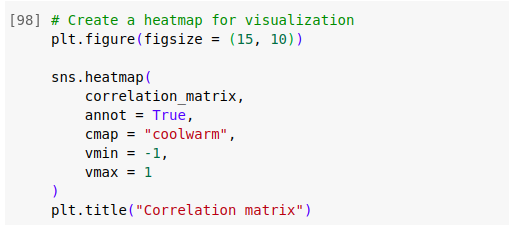


Figure 17: Plotting the correlation matrix

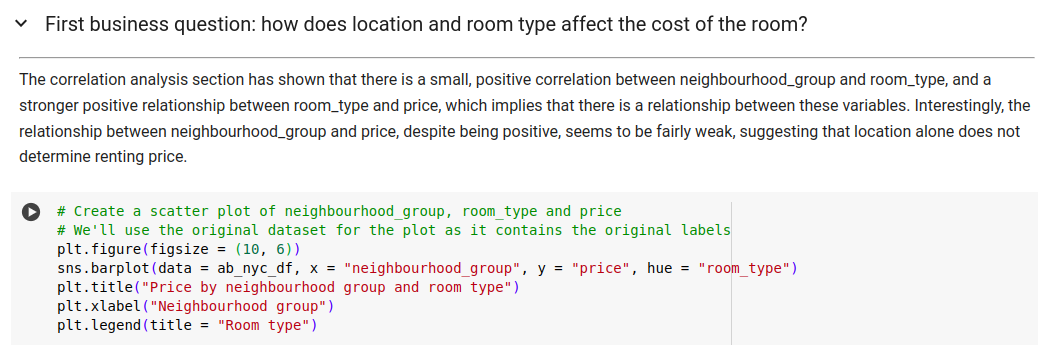
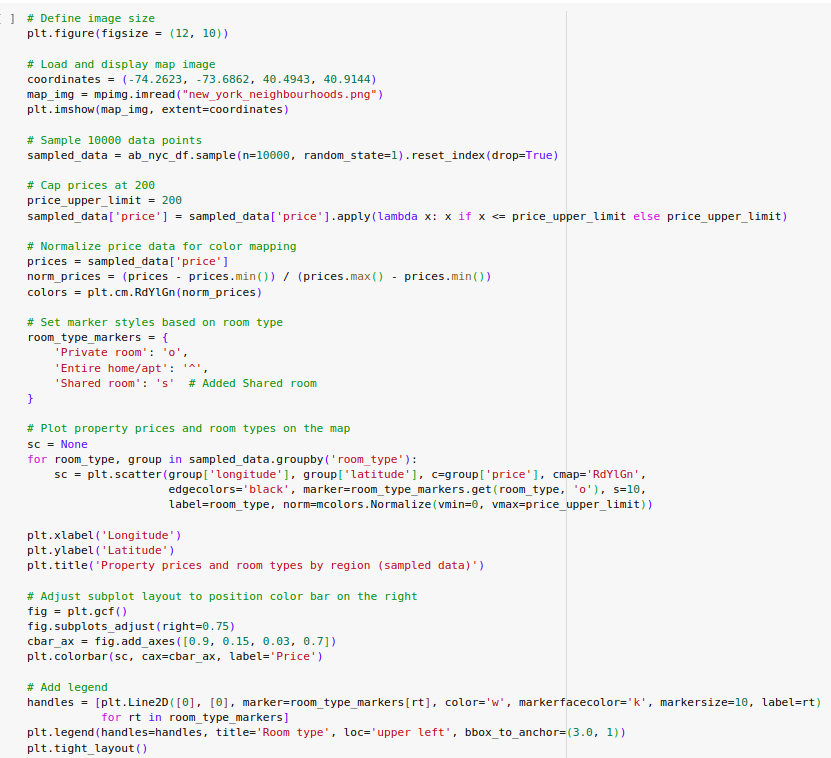
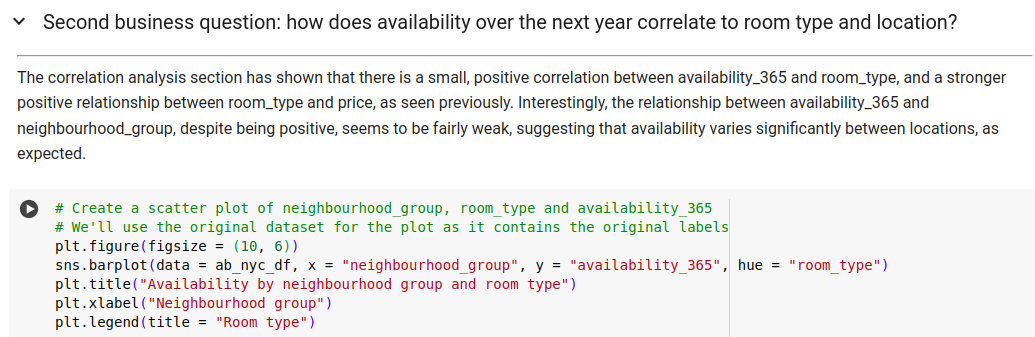
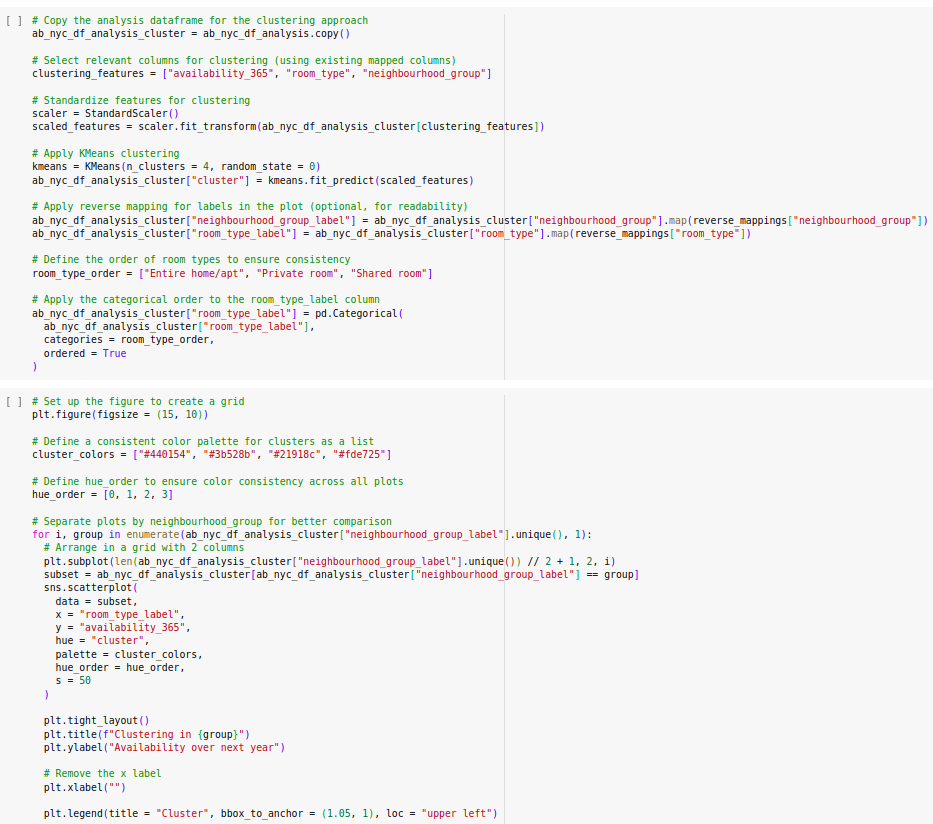
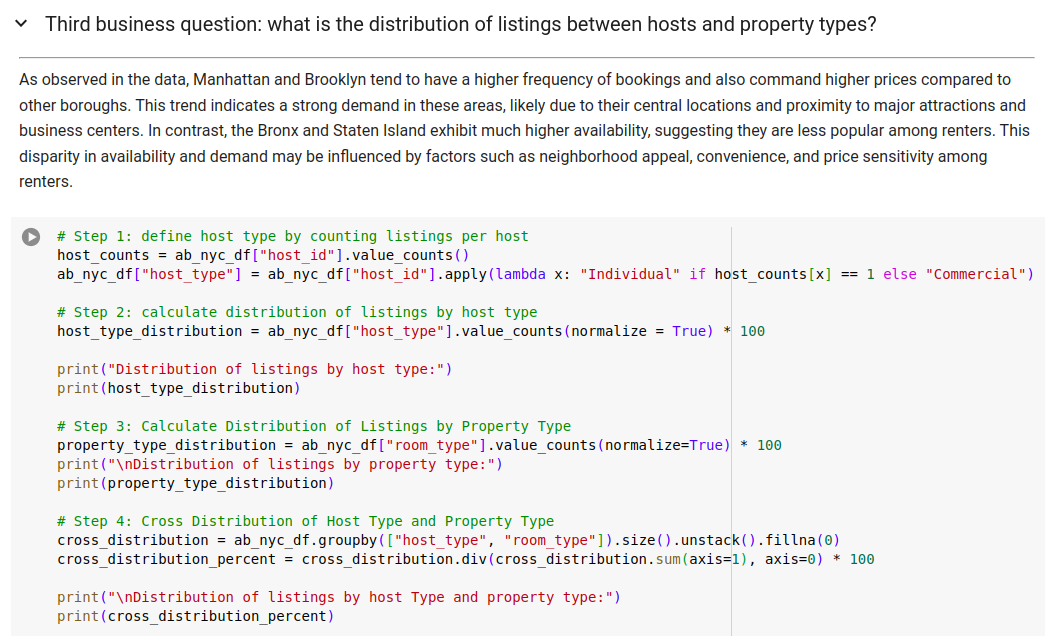
Figure 18: Creating a bar plot for the first question

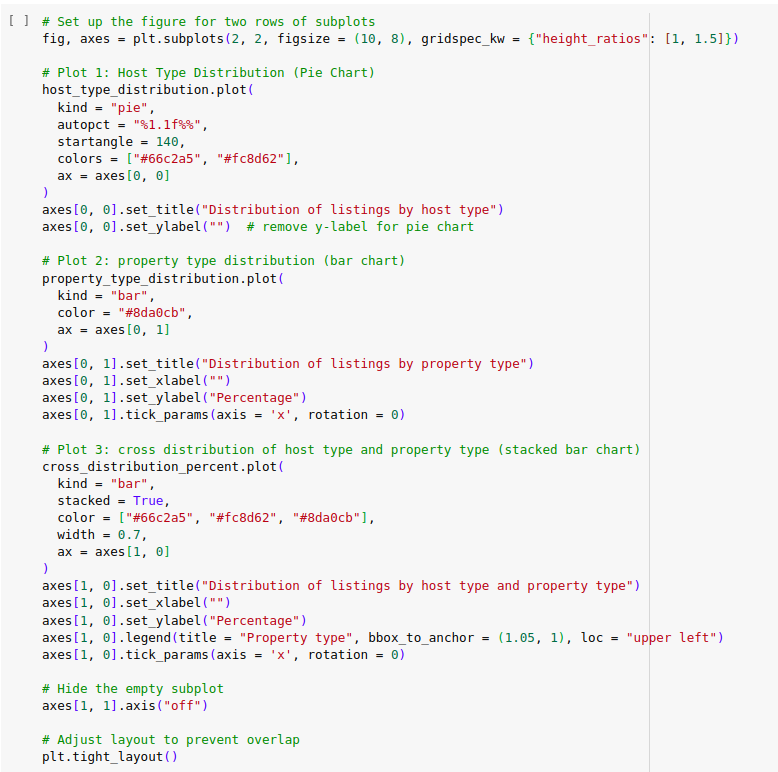
Figure 19: Creating a box plot for the first question

Figure 20: Creating a scatter plot for the New York City neighbourhoods' rental samples

Figure 21: Creating a bar plot for the second question

Figure 22: Clustering and plotting the data by availability over the next year, neighbourhood group and room type

Figure 23: Separating individual from commercial hosts

Figure 24: Creating several plots to showcase the distribution of rental types per host type