**Resubmission**

**Machine Learning Module: Development Team Project Report**

**Introduction**

Airbnb (Air Bed and Breakfast) is a short term rental property, and it is one of the most relevant modern accommodation services; with around 5 million listings across more than 81000 cities as of 2008 (Perez-Sanchez, 2018).Understanding the factors which contribute to the pricing of listings helps to optimize pricing strategies. A study by Pittala (2004) shows that location, property characteristics and host attributes are key determinants of pricing. By analyzing these factors, stakeholders such as guests and policymakers can make informed decisions that enhance the fairness and efficiency of the Airbnb market. More importantly, hosts can use the knowledge to maximize their revenue while remaining competitive. In this report, the Airbnb New York 2019 dataset will be analysed in order to answer the Business Analytic Question: “*Which factors influence Airbnb listing prices in the different boroughs of New York City, and how can hosts optimize their pricing strategy?*”

**The Dataset**

The dataset contains data such as host details, boroughs (Bronx, Brooklyn, Manhattan, Queens, Staten Island), location coordinates, **room type** (entire home/apt, private room, shared room), **price** (nightly rate), minimum stay duration, **number of reviews**, date of the latest review, number of **reviews per month**, number of listings per host, and the number of available days per year. Python programming language will be used to carry out Data Pre-processing, Exploratory Data Analysis (EDA) as well as the K-Means Clustering.

**Data Analysis**

The first step is to get an overview of the dataset, that is, which columns are in the dataset and what their data types are. According to the table below there are 15 columns in total. However columns like host\_id, host\_name, longitude and latitude are not imperative to this analysis.

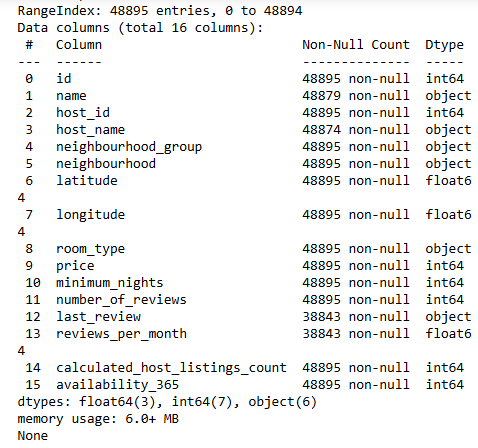
****

Table 1. Info about AB\_NYC\_2019.csv dataset

**Data Pre-processing**

Data pre-processing involves checking for missing values, and other inconsistencies to ensure data quality before analysis (Singh, 2021). The dataset was checked for missing values, and the outliers in the price column were removed using the Interquartile Range method.

**EDA**

Next, EDA was carried out in order to discover the trends that exist in relation to price (Jebb, 2017). Here, a Correlation Matrix, bar plots and scatter plots are some of the visualisations which were created. Last but not least the K-Means Clustering ML algorithm was used to analyse and visualize the effect of the availability and minimum nights on the pricing.

**Correlation Matrix**

Below is a Correlation Matrix which shows how each attribute relates to the others. We can conclude from the matrix that the price has a negative correlation to the number of reviews per month, and a positive correlation to the number of listings as well as the minimum nights required for booking. Price also has a very strong correlation to the availability (number of days in a year when listing is available for booking).

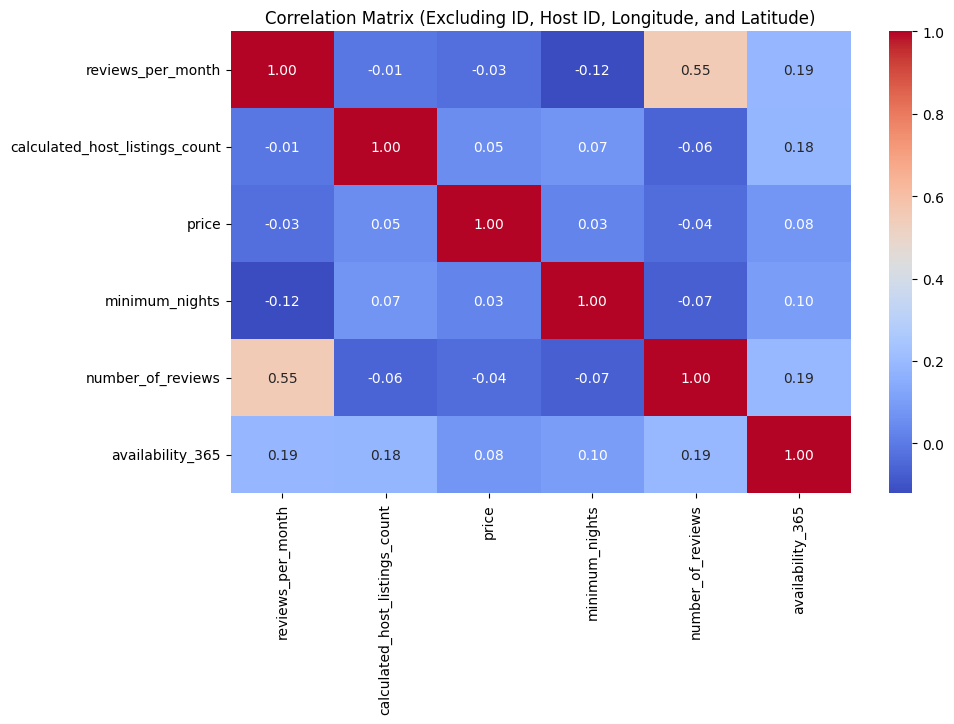
****

Figure 1 Correlation Matrix

**Effect of Boroughs**

When comparing the average prices by borough, listings in Manhattan are priced the highest (more than $175), followed by those in Brooklyn. Those in Bronx are priced the lowest (just above $75). The listings in Staten Island and Queens are priced between $80 and $100 on average. This pattern may be due to the fact that the five boroughs in New York have varying socio-economic profiles. According to the 2024 article by the New York Post, the average rents varied significantly across the boroughs; with the rent in Manhattan at $3,790, Brooklyn at $3,790, Queens at $3,427 and the Bronx at $3,163.

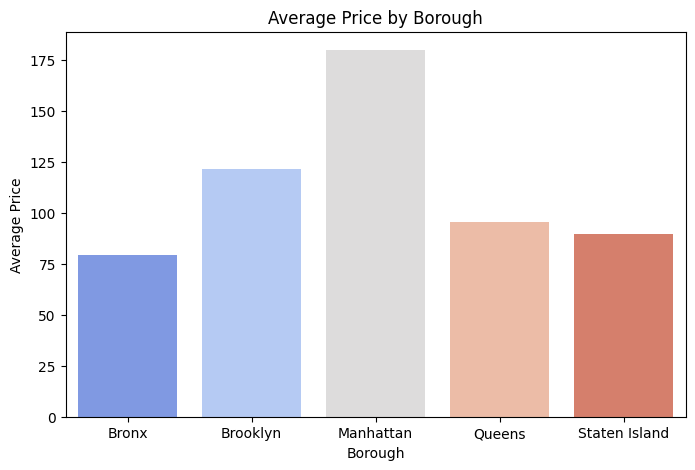
****

Figure 2 Bar plot of Average Price by Borough

**Effect of Neighbourhood Group (Borough) and Room Type**

We can see from the figure below that in general, entire homes/ apartments have significantly higher prices, followed by private rooms, and then shared rooms in all the boroughs except for Staten Island and Bronx. In these two boroughs, private rooms and shared rooms are priced at just above $50 on average, while entire homes or apartments are priced at around $125.

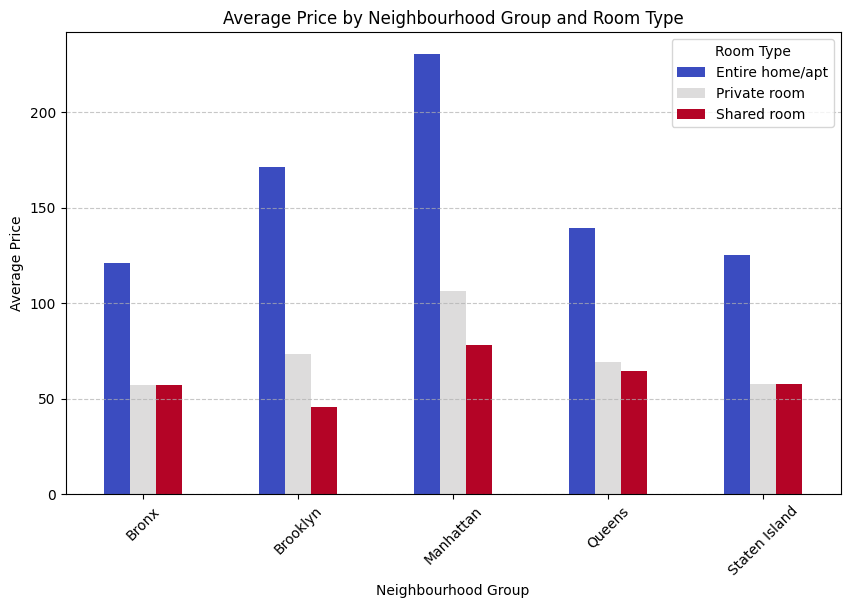
****

Figure 3 Bar Chart of Average Price by Borough and Room Type

**Effect of Number of Reviews**

The analysis shows that the lower the price, the higher the number of reviews is; below is a visualisation of this. The pricing is, however, not greatly influenced by the number of reviews as illustrated by the Bar plot of Price vs. Review Count in Figure 5. Lawani (2019) writes in his paper that Airbnb listing prices are influenced by the review score among other factors. However, the dataset that was analysed in this report only contains details about the number of reviews and not their content or rating (score), therefore it is lacking in this regard.

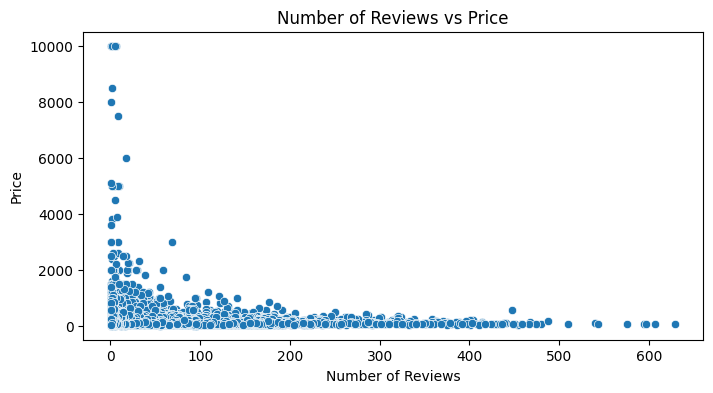
****

Figure 4 Scatter Plot of Number of Reviews vs Price

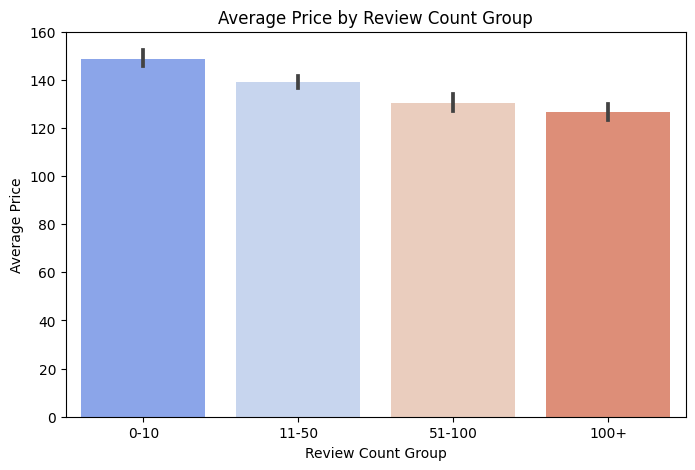
****

Figure 5

**K-Means Clustering**

K-Means Clustering visualizes the effect of not only the minimum nights, but also the availability on the pricing. The distinct clusters show that minimum nights do not have a significant effect on the pricing. However; the higher the availability, the higher the price is.

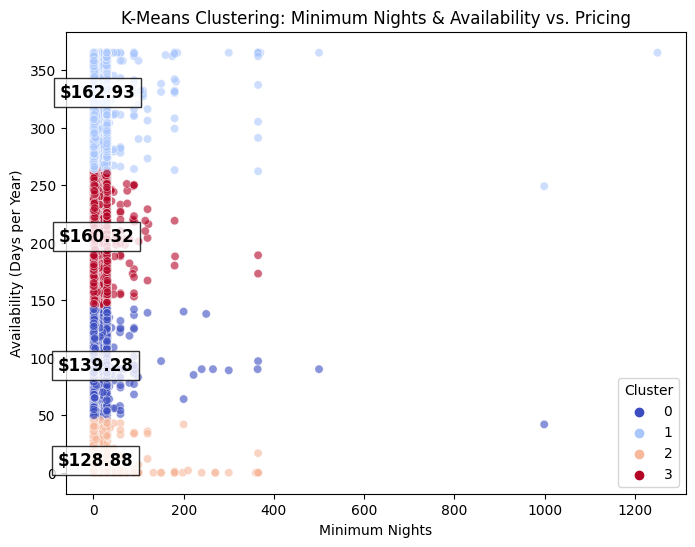
**\***

Figure 6 K-Means Clustering

**Conclusion**

From the analysis we can conclude that hosts can maximize revenue by adjusting their prices according to the location of their listings. Listings in Manhattan are the most highly priced, while those in Staten Island are more affordable. Moreover, hosts can optimize pricing based on accommodation type - entire homes have the highest prices, while shared rooms and private rooms are more budget-friendly.

The availability of a listing also has a strong influence on the pricing - listings that have high availability (likely long-term stays) tend to have higher pricing, while listings with low availability have more moderate prices. Lastly, we can conclude that a higher number of reviews correlate with lower pricing, and hosts who have a few reviews may need to adjust their prices in order to attract more guests.

**References:**

Gerstein, J. (2024). NYC rent continues to rise: Here’s how much you need to earn to live in each borough. Available at: <https://nypost.com/2024/09/26/us-news/nyc-rent-continues-to-rise-heres-how-much-you-need-to-earn-to-leave-in-each-borough/?utm_source=chatgpt.com> [Accessed: 27 March 2025]

Jebb, A. et al. (2017). Exploratory data analysis as a foundation of inductive research, Human Resource Management Review. 27(2),pg 265-276 DOI: <https://doi.org/10.1016/j.hrmr.2016.08.003>[Get rights and content](https://s100.copyright.com/AppDispatchServlet?publisherName=ELS&contentID=S1053482216300353&orderBeanReset=true)

Lawani, A. et al. (2019). Reviews and price on online platforms: Evidence from sentiment analysis of Airbnb reviews in Boston, Regional Science and Urban Economics. 75, pg 22-34. DOI: <https://doi.org/10.1016/j.regsciurbeco.2018.11.003>

Perez-Sanchez, V. et al. (2018).The What, Where, and Why of Airbnb Price Determinants, Sustainability. 10(12), 4596. DOI: <https://doi.org/10.3390/su10124596>

Singh, P. et al. (2021). Chapter 5 – Diagnosing of disease using machine learning, Machine Learning and the Internet of Medical Things in Healthcare*.*pg 89-111*.* DOI: <https://doi.org/10.1016/B978-0-12-821229-5.00003-3>