Machine Learning

Development Team Project

Airbnb Business Analysis Using a Data Science Approach

# By

# Gayathridevi Durairaj, Pitshou Bokoro, Sabanathan Rubanathan, Theodoros Theodorou

# (Group 4)

# Table of Contents

[**Table of Contents 1**](#_id0d0n7xmig8)

[**Introduction 2**](#_3hud0r7bsa3q)

[1. Context 2](#_le6un58yb8ay)

[2. Rationale 2](#_qhjyr958vapk)

[3. Business Question 2](#_jlin785n5mkm)

[4. Business Impact 2](#_e0k3tfc42jzu)

[**Methodology 3**](#_hwfoz8q0n67z)

[1. Data Cleaning 3](#_lp3ibc3iyekj)

[2. Feature Engineering 3](#_twgnfmneflf9)

[3. Exploratory Data Analysis (EDA) 3](#_7unw84hzxinn)

[4. Analytical Methods 5](#_hewapmayi30e)

[**Regression 6**](#_2zr1zkffa4aq)

[1. Results 6](#_5byawsfw5n66)

[**Clustering 8**](#_hdeh3dj0k3uf)

[1. Results 8](#_qxtxgqwpg835)

[**Conclusion 9**](#_blzmvn6tux96)

[**Bibliography 10**](#_rdxm6ctl0q4v)

[**Appendix 12**](#_2zbfwsiv5m48)

[1. Code Used in EDA 12](#_gcpwrwdflhyd)

[a. Importing Libraries 12](#_fndx4dypyar0)

[b. Loading and Inspecting Data 12](#_wqjknx66zmps)

[c. Data Cleaning 12](#_hgoee88mwk1b)

[d. Feature Engineering 13](#_abye0edmp0wb)

[e. Exploratory Data Analysis (EDA) 13](#_t44iy9yfvwez)

[2. Code Used in Regression Analysis 13](#_q72uufo5bv0)

[3. Code Used in Clustering Analysis 14](#_2fz963q2cj3w)

[a. Selecting the numerical features that are appropriate for clustering 14](#_kh5a6eqnbsyc)

[b. Standardising Data 15](#_c2sdmtr5ox41)

[c. Determine the optimal number of clusters (k) 15](#_jhfxnq2gsldm)

[d. Applying k-means Clustering 15](#_3mmzx1rt9tks)

[e. Analysing the Clusters 15](#_aqybug43j0a6)

[4. Supplementary Chart 16](#_j0apsyiirmy0)

# Introduction

## 1. Context

Airbnb has rapidly evolved into a transformative player within the global hospitality industry, redefining traditional accommodation models by offering a digital marketplace that seamlessly connects hosts with guests through short-term rentals. Its success is underpinned by the complex interplay of multiple factors, including pricing strategies, listing reviews, and booking patterns, all of which shape consumer decision-making and host performance (Ke, 2017; Alharbi, 2023). With millions of active listings distributed across diverse geographic contexts, Airbnb produces a vast and dynamic dataset. This dataset offers valuable opportunities for empirical research examining market dynamics, consumer preferences, and revenue generation patterns (Dogru et al., 2019).

## 2. Rationale

Understanding the determinants of revenue generation on Airbnb is of both academic and practical significance. Prior research highlights the role of pricing, location, and amenities in shaping consumer preferences, while customer reviews act as critical trust-building mechanisms (Gibbs et al., 2017; Lin and Yang, 2024). However, the combined influence of these factors on host performance and platform-level outcomes remains underexplored. Examining these relationships contributes to knowledge on digital platform economics, consumer behaviours, and data-driven decision-making in the sharing economy.

## 3. Business Question

How do host portfolio sizes (measured by calculated\_host\_listings\_count) influence listing reviews and Airbnb’s business performance and market dynamics?

## 4. Business Impact

Addressing this question provides theoretical contributions to platform-based market literature while delivering practical insights into refining pricing algorithms, enhancing host competitiveness, and strengthening trust through effective review systems. Furthermore, such findings can inform evidence-based strategies that support Airbnb’s long-term growth, resilience, and profitability, ensuring sustained relevance and competitiveness within an increasingly dynamic global hospitality marketplace.

# Methodology

## 1. Data Cleaning

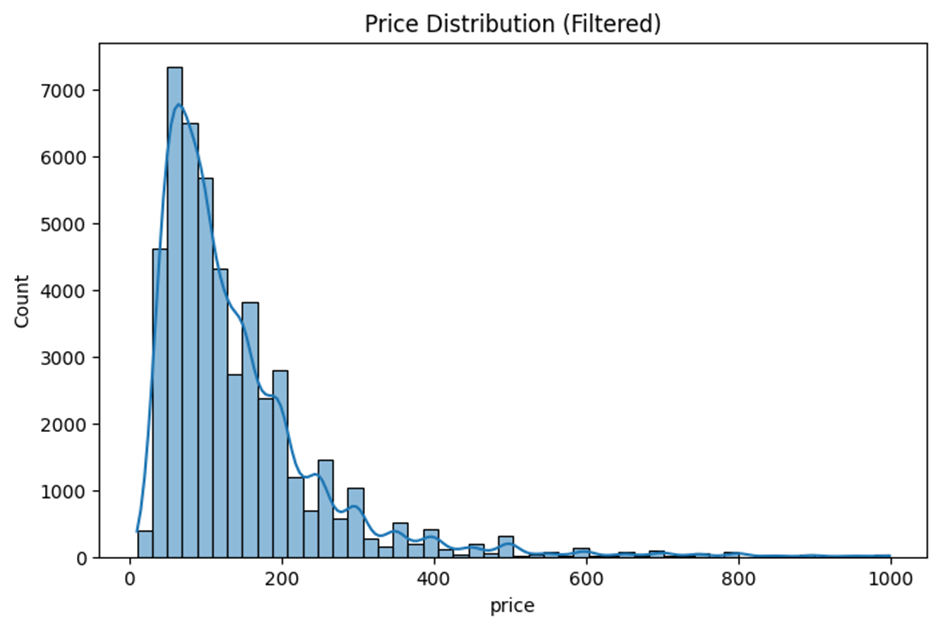
The first step in research analysis is conducting exploratory data analysis (EDA) as it provides critical insight into data structures (Behrens et al., 2012). To understand how a host’s portfolio size affects Airbnb pricing and demand in New York City, the dataset was first prepared and cleaned to ensure reliable insights (Cunningham and Muir, 2023). Columns like host\_name and listing\_ID were removed, as they do not influence market behaviour. Missing values in the reviews\_per\_month column were replaced with zeros, representing listings without reviews with no recent activity. Outliers in the price column were filtered to focus on realistic rates between $1 and $1,000, preventing a few extreme listings from distorting trends. These data cleaning steps lay the foundation for accurate exploratory data analysis (Sharifnia et al., 2025).

## 2. Feature Engineering

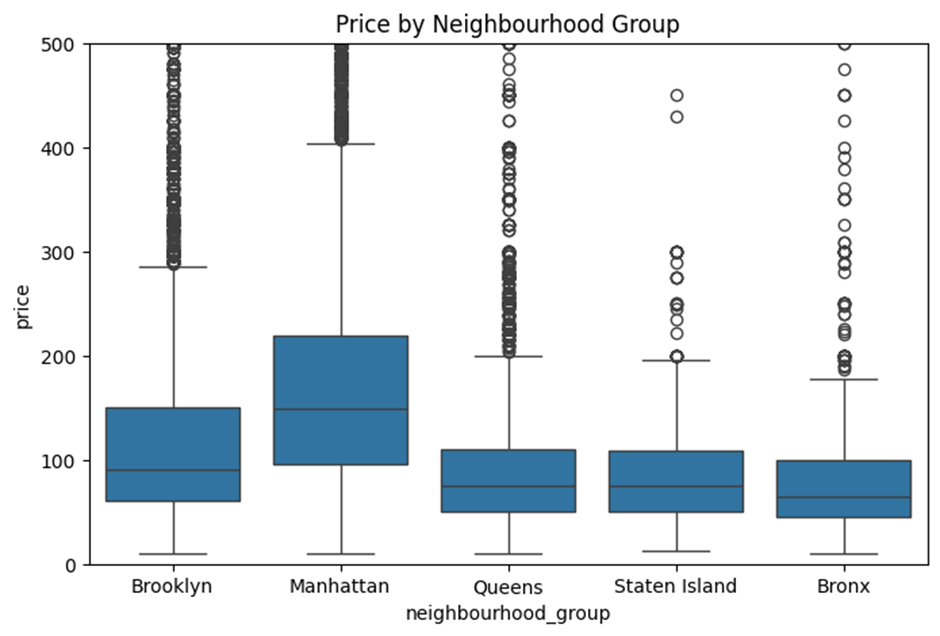
Feature engineering added meaningful business metrics. Host portfolio size (calculated\_host\_listings\_count field), represents how many listings each host operates; so to measure demand, a demand\_score was created by multiplying total reviews by reviews per month, capturing both the scale and recent activity of each listing. Neighbourhood group categories (Manhattan, Brooklyn, etc.) were retained to evaluate how host scale and demand vary across distinct market segments.

## 3. Exploratory Data Analysis (EDA)

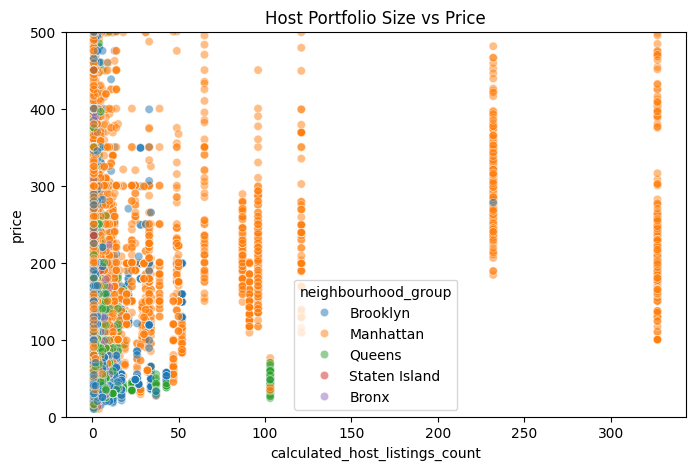
Visualisations help translate these insights for business audiences. A price distribution histogram (Figure 1) shows the overall range of nightly rates and highlights where most listings are priced, helping identify typical market levels. A boxplot of price by neighbourhood group (Figure 2) makes it clear which boroughs tend to command higher or lower rates, highlighting regional pricing patterns that may influence host strategy. Finally, a scatterplot of host portfolio size versus price (Figure 3), coloured by neighbourhood, illustrates how professional hosts with multiple listings position themselves across different market segments. These visuals assist with portfolio analysis, as they provide intuitive summaries of complex data before machine learning methods are applied.



*Figure 1. A histogram representing average prices for an amount of listings*

**

*Figure 2. A boxplot showing the price of listings within each neighbourhood group*

**

*Figure 3. A scatterplot to show the variation of price with the host’s listing count for each neighbourhood group*

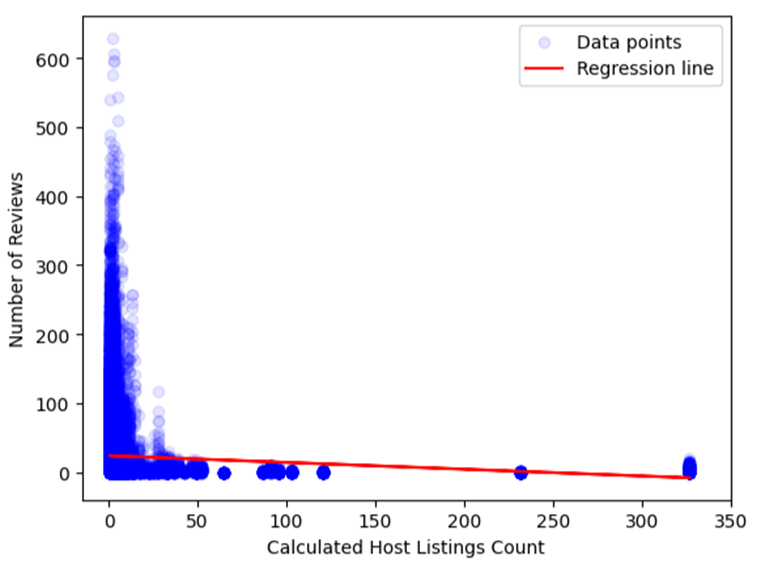
## 4. Analytical Methods

To investigate how host portfolio size influences pricing and demand, this study applies two complementary machine learning methods. Regression analysis will be used to estimate the impact of portfolio size, availability, and neighbourhood on guest demand and pricing (Magno, Cassia and Ugolini, 2018). In business terms, this helps identify whether larger hosts charge higher rates or whether availability and location drive more reviews. In parallel, clustering analysis will segment listings into groups with similar characteristics (such as small or large listings) providing insight into market structure, competitive positioning, and targeted pricing strategies (Dolgui and Proth, 2010).

# Regression

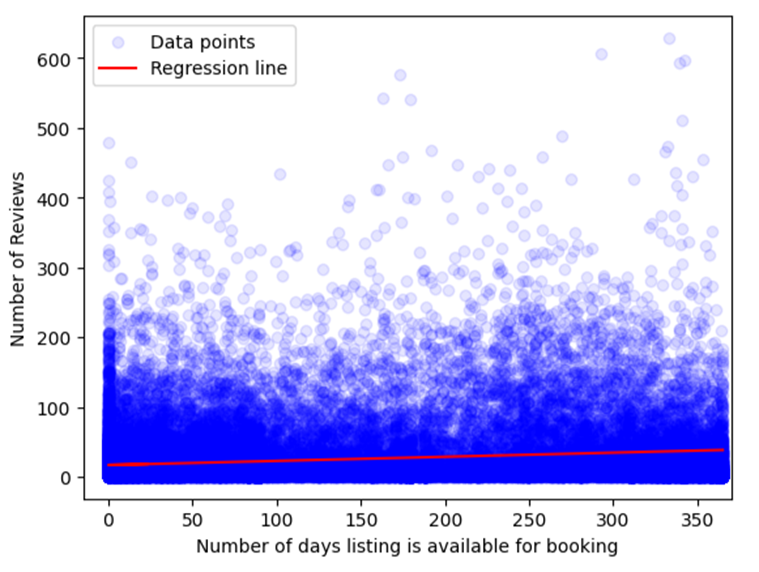
## 1. Results

Our goal was to identify any patterns related to the host portfolios (Calculated\_Host\_Listings\_Count) and the number of reviews for each listing and how it can affect Airbnb's business. Furthermore, to identify any other factors that can affect the number of reviews for the Airbnb business. According to the regression line between calculated\_host\_listings\_count and number of reviews (James et al., 2023), we can clearly see that high listing counts can harm the business regarding the number of reviews. In Figure 4 we can see that the higher number of reviews is situated from 3 to 15 calculated\_host\_listing\_counts, which indicates that this is an ideal host listings to maximise the number of reviews for Airbnb's business. Regarding the business aspect, these results show that hosts regardless of, if they are running a business or not, which are more likely to be aiming for higher calculated\_host\_listing\_counts, should be aiming to have a number of listings count of around 3 to 15.



*Figure 4. Regression Line for calculated host listings count and number of reviews*

In addition, a note worth taking for Airbnb's business is that as shown in Figure 5 there is a positive regression line between number of days listing is available for booking (availability\_365) and number of reviews. This analysis was performed to identify if a higher number of availability for the listings can impact the number of reviews for Airbnb's business. The results show that hosts can benefit from having their Airbnb available throughout the year if that is possible for them.

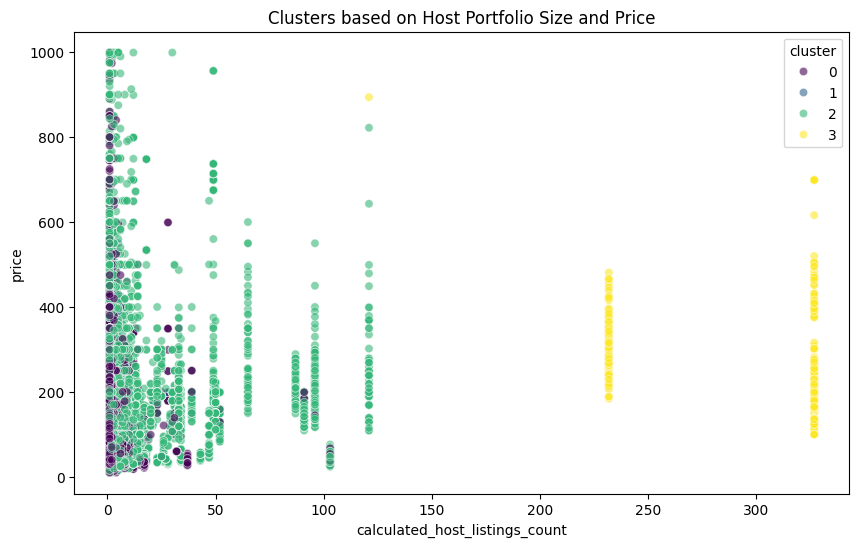


*Figure 5. Number of Reviews in regards to number of days listing is available for booking*

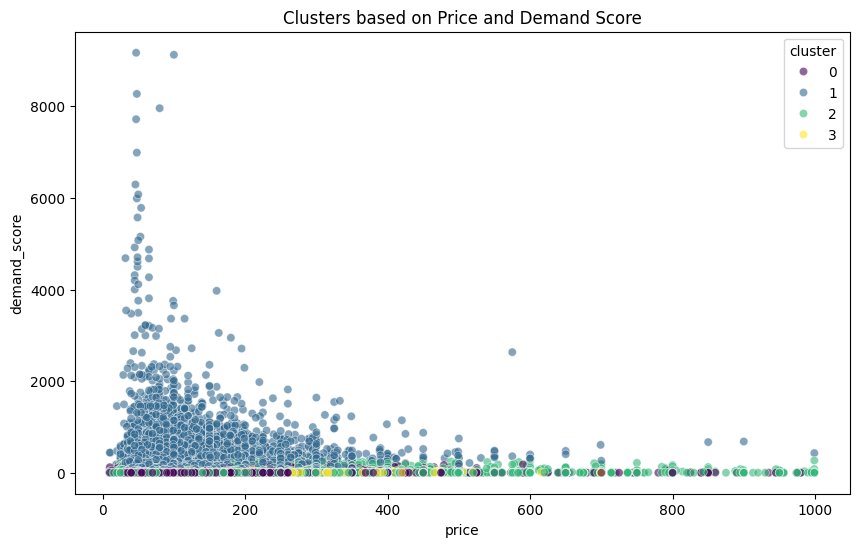
# Clustering

## 1. Results

Using unsupervised learning techniques, in this study we applied k-means clustering to segment Airbnb listings based on their quantitative attributes (Özçini, Yılmaz & Kaya, 2023). The elbow method facilitated the determination of an appropriate number of clusters (k=4). This technique was selected for its computational efficiency, interpretability, and scalability.

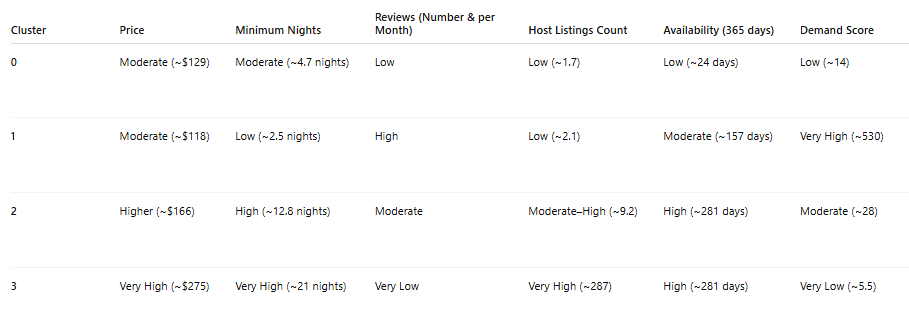
****

*Figure 6. Clusters based on Host Portfolio Size and Price*

****

*Figure 7. Cluster based on Price and Demand Score*

*Table 1: Cluster based on Price and Demand Score & Host Portfolio Size and Price Summary*

**

The cluster analysis indicates that larger portfolios (LPs) tend to attract higher prices and longer minimum stays. However, LPs receive fewer monthly reviews, which is consistent with the regression line findings.

# Conclusion

Regression and cluster analyses suggest that LPs attract fewer reviews than smaller or moderate ones. Assuming reviews reflect guest numbers, smaller portfolios may host guests more frequently. Therefore, hosts should be aiming to have listings count at no more than 20 to 25 times as we can see that it can be harmful to their number of reviews. LPs, however, optimises pricing strategies through higher prices and longer minimum stays, which reduce guest turnover but limit downtime.

Without financial data such as revenue or profit, firm business recommendations are not possible. More bookings do not necessarily mean higher profitability, as shown in a Turin study linking lower prices with higher review counts (Toppani, 2023). Nonetheless, if better pricing drives profitability, Airbnb could prioritise favourable terms for professional hosts managing LPs.

# Bibliography

Alharbi, Z. H. (2023). A Sustainable Price Prediction Model for Airbnb Listings Using Machine Learning and Sentiment Analysis. Sustainability, 15(17), 13159.<https://doi.org/10.3390/su151713159>.

Behrens, J.T., DiCerbo, K.E., Yel, N. and Levy, R. (2012). Exploratory Data Analysis. Handbook of Psychology, Second Edition, 2. doi: <https://doi.org/10.1002/9781118133880.hop202002>

Cunningham, S.A. and Muir, J.A. (2023). Data Cleaning. [online] Cambridge University Press. Available at: <https://www.cambridge.org/core/books/abs/cambridge-handbook-of-research-methods-and-statistics-for-the-social-and-behavioral-sciences/data-cleaning/44E664FD2372D182EE74BE39E8DAFD21>

Dogru, T., Mody, M., & Suess, C. (2019). Adding evidence to the debate: Quantifying Airbnb's disruptive impact on ten key hotel markets, Tourism Management, Vol. 72, pp. 27-38, ISSN 0261-5177,<https://doi.org/10.1016/j.tourman.2018.11.008>.

Dolgui, A. and Proth, J.-M. (2010). Pricing strategies and models. Annual Reviews in Control, 34(1), pp.101–110. doi: <https://doi.org/10.1016/j.arcontrol.2010.02.005>

Gibbs, C., Guttentag, D., Gretzel, U., Morton, J., & Goodwill, A. (2017). Pricing in the sharing economy: a hedonic pricing model applied to Airbnb listings. Journal of Travel & Tourism Marketing, 35(1), 46–56.<https://doi.org/10.1080/10548408.2017.1308292>.

James, G., Witten, D., Hastie, T., Tibshirani, R. and Taylor, J. (2023). Linear regression. In *An introduction to statistical learning: With applications in python* (pp. 69-134). Cham: Springer international publishing.

Ke, Q. (2017). Sharing Means Renting? An Entire-Marketplace Analysis of Airbnb. In Proceedings of the 2017 ACM on Web Science Conference. 131-139, Available at SSRN: <https://ssrn.com/abstract=2902840> or [http://dx.doi.org/10.2139/ssrn.2902840](https://dx.doi.org/10.2139/ssrn.2902840)

Lin, W. & Yang, F. (2024). The price of short-term housing: A study of Airbnb on 26 regions in the United States, Journal of Housing Economics, Elsevier, volume 65, issue C,<https://doi.org/10.1016/j.jhe.2024.102005>.

Magno, F., Cassia, F. and Ugolini, M.M. (2018). Accommodation prices on Airbnb: effects of host experience and market demand. The TQM Journal, 30(5), pp.608–620. doi: <https://doi.org/10.1108/tqm-12-2017-0164>

Özçini, M.A., Yılmaz, T. & Kaya, T. (2023). ‘Clustering AirBnB Listings in Istanbul Using Unsupervised Machine Learning Techniques’, in C. Kahraman, I.U. Sari, B. Oztaysi, S. Cevik Onar, S. Cebi & A. Ç. Tolga (eds), Intelligent and Fuzzy Systems – Intelligence and Sustainable Future: Proceedings of the INFUS 2023 Conference (Lecture Notes in Networks and Systems, vol. 759), Springer Science and Business Media Deutschland GmbH, pp. 197–205. doi: 10.1007/978-3-031-39777-6\_24. [online] doi: <https://link.springer.com/chapter/10.1007/978-3-031-39777-6_24>

Sharifnia, A.M., Kpormegbey, D.E., Thapa, D.K. and Cleary, M. (2025). A Primer of Data Cleaning in Quantitative Research: Handling Missing Values and Outliers. PubMed. [online] doi: <https://doi.org/10.1111/jan.16908>

Toppani, M. (2023). *The impact of professional hosts on Airbnb performance: Neighborhood dynamics, market sensitivity, and seasonal trends*. Master’s thesis. Politecnico di Torino. [online] doi: <https://webthesis.biblio.polito.it/35719/>

# Appendix

## 1. Code Used in EDA

### a. Importing Libraries

#Libraries for EDA

from google.colab import drive

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

#Libraries for Regression

from sklearn.linear\_model import LinearRegression

from sklearn.preprocessing import PolynomialFeatures

from scipy.optimize import curve\_fit

#Libraries for Clustering

from sklearn.preprocessing import StandardScaler

from sklearn.cluster import KMeans

from sklearn.metrics import silhouette\_score

### b. Loading and Inspecting Data

from google.colab import drive

drive.mount('/content/drive')

df = pd.read\_csv("/content/drive/MyDrive/Airbnb ML project/AB\_NYC\_2019.csv")

print(df.head())

print(df.info())

### c. Data Cleaning

# Drop irrelevant columns

cols\_to\_drop = ["id", "name", "host\_name", "last\_review"]

df = df.drop(columns=cols\_to\_drop)

# Handle missing values

df["reviews\_per\_month"] = df["reviews\_per\_month"].fillna(0)

# Remove extreme outliers in price

df = df[(df["price"] > 0) & (df["price"] < 1000)]

### d. Feature Engineering

# Demand proxy: number\_of\_reviews + reviews\_per\_month

df["demand\_score"] = df["number\_of\_reviews"] \* df["reviews\_per\_month"]

# Portfolio scale already exists: calculated\_host\_listings\_count

### e. Exploratory Data Analysis (EDA)

plt.figure(figsize=(8,5))

sns.histplot(df["price"], bins=50, kde=True)

plt.title("Price Distribution (Filtered)")

plt.show()

plt.figure(figsize=(8,5))

sns.boxplot(x="neighbourhood\_group", y="price", data=df)

plt.ylim(0, 500)

plt.title("Price by Neighbourhood Group")

plt.show()

plt.figure(figsize=(8,5))

sns.scatterplot(x="calculated\_host\_listings\_count", y="price", hue="neighbourhood\_group", data=df, alpha=0.5)

plt.ylim(0, 500)

plt.title("Host Portfolio Size vs Price")

plt.show()

## 2. Code Used in Regression Analysis

#Select the data used

x = df["calculated\_host\_listings\_count"].values.reshape(-1, 1)

y = df["number\_of\_reviews"].values

#Remove NaN values

mask = ~np.isnan(y)

x = x[mask]

y = y[mask]

#Fit regression model

model = LinearRegression()

model.fit(x, y)

y\_pred = model.predict(x)

plt.scatter(x, y, color="blue", alpha=0.1, label="Data points")

plt.plot(x, y\_pred, color="red", label="Regression line")

plt.xlim(-15, 350)

plt.xlabel("Calculated Host Listings Count")

plt.ylabel("Number of Reviews")

plt.legend()

plt.show()

#Select the data used

x = df["availability\_365"].values.reshape(-1, 1)

y = df["number\_of\_reviews"].values

#Remove NaN values

mask = ~np.isnan(y)

x = x[mask]

y = y[mask]

#Fit regression model

model = LinearRegression()

model.fit(x, y)

y\_pred = model.predict(x)

plt.scatter(x, y, color="blue", alpha=0.1, label="Data points")

plt.plot(x, y\_pred, color="red", label="Regression line")

plt.xlim(-15, 370)

plt.xlabel("Number of days listing is available for booking")

plt.ylabel("Number of Reviews")

plt.legend()

plt.show()

## 3. Code Used in Clustering Analysis

### a. Selecting the numerical features that are appropriate for clustering

numerical\_features = ['price', 'minimum\_nights', 'number\_of\_reviews', 'reviews\_per\_month', 'calculated\_host\_listings\_count', 'availability\_365', 'demand\_score']

df\_numerical = df[numerical\_features]

display(df\_numerical.head())

### b. Standardising Data

scaler = StandardScaler()

df\_scaled = scaler.fit\_transform(df\_numerical)

### c. Determine the optimal number of clusters (k)

inertia = []

# silhouette\_scores = [] # Remove silhouette\_scores list

k\_range = range(2, 11)

for k in k\_range:

kmeans = KMeans(n\_clusters=k, random\_state=42, n\_init=10)

kmeans.fit(df\_scaled)

inertia.append(kmeans.inertia\_)

# if k > 1: # Silhouette score is not defined for k=1 # Remove silhouette score calculation

# silhouette\_scores.append(silhouette\_score(df\_scaled, kmeans.labels\_))

plt.figure(figsize=(8, 5)) # Adjust figure size as only one plot

plt.plot(k\_range, inertia, marker='o') # Plot inertia against k\_range

plt.title('Elbow Method')

plt.xlabel('Number of clusters (k)')

plt.ylabel('Inertia')

plt.xticks(k\_range)

plt.tight\_layout()

plt.show()

### d. Applying k-means Clustering

kmeans = KMeans(n\_clusters=4, random\_state=42, n\_init=10)

df["cluster"] = kmeans.fit\_predict(df\_scaled)

display(df.head())

### e. Analysing the Clusters

cluster\_centers = df.groupby("cluster")[numerical\_features].mean()

display(cluster\_centers)

# Visualise clusters (example: price vs demand\_score)

plt.figure(figsize=(10, 6))

sns.scatterplot(x="price", y="demand\_score", hue="cluster", data=df, alpha=0.6, palette="viridis")

plt.title("Clusters based on Price and Demand Score")

plt.show()

# Visualise clusters (example: calculated\_host\_listings\_count vs price)

plt.figure(figsize=(10, 6))

sns.scatterplot(x="calculated\_host\_listings\_count", y="price", hue="cluster", data=df, alpha=0.6, palette="viridis")

plt.title("Clusters based on Host Portfolio Size and Price")

plt.show()

## 4. Supplementary Chart

The following supplementary chart was generated during the ML analysis stage but is not included in the main body of the report. It provides additional context for clustering.

*Figure 4a: finding the optimal number of clusters for k-means clustering using the elbow method and silhouette scores, iterating through different values of k, fitting the KMeans model, and calculating inertia and silhouette scores.*

