

A Data Mining Approach to Classifying and Segmenting Airbnb Listings in New York City

Student Name: Muhammad Haroon

Student ID: 2434025

Course: ACCA7015 - Data Mining & Visualisation

Table of Contents

Abstract.....	3
1. Introduction.....	3
2. Literature Review	4
2.1. Predictive analytics and Data mining.....	5
2.2. Classification Techniques.....	5
2.3. Clustering Techniques	5
2.4. Implementation in Short-term Rental Markets.....	6
3. Dataset Description and Pre-processing	6
3.1. Dataset Overview.....	6
3.2. Pre-processing and Data Cleaning	7
3.3. Features for Modelling	7
4. Model Implementation and Evaluation.....	8
4.1. Classification Task: Predicting Room Type	8
4.1.1. Methodology and Code Implementation	8
4.1.2. Model Evaluation and Comparison.....	8
4.2. Clustering Task: Market Segmentation using K-Means.....	9
4.2.1. Methodology and Code Implementation	9
• Data Preparation: Data features were normalized with StandardScaler to make sure that no feature with a high score dominated models, all the features equally contributed towards results of the distance calculations.....	9
• Computing the K Optimal: plot of Within-Cluster-Sum-of-Squares (WCSS) vs. K was plotted to obtain K Optimal. The rate of decrease in WCSS is changing sharply at the so-called elbow of the plot, which gives an idea of an appropriate k. Silhouette Score was also obtained to legitimize this decision.....	9
• Model Application: K-Means was run using the selected k, and the clusters that were formed were analyzed.	9
4.2.2. Model Evaluation and Interpretation.....	10
5. Results Analysis and Discussion of Findings	12
5.1. Synthesis of Classification Results	12
5.2. Synthesis of Clustering Results.....	13
5.3. Integration	14
5.4. Critical Assessment of Techniques.....	14
6. Business Recommendations.....	15

6.1.	For Airbnb Hosts	15
•	Strategic Positioning: It is suggested that hosts can distinguish their cluster segment with the help of a tool offered by Airbnb (according to this scheme). The Cluster 0 hosts must offer the value and instant booking. The hosts of Cluster 1 shall invest in professional photography and high-quality facilities in order to charge higher prices. Cluster 2 hosts are the ones that should advertise in websites that are focused on long-term rentals.	15
•	Dynamic Pricing: Hardly use the cluster properties, in order to base pricing. A host who has a listing which is similar to Cluster 1 (high reviews, low availability) should charge aggressively than that of Cluster 0.	15
•	Portfolio Management: Multiple listings hosts (calculated_host_listings_count) can apply the clustering technique to diversify their portfolio, so that they have a variety of budget and premium services to meet the demand of various market segments.	15
6.2.	For the Airbnb Platform.....	15
	Personalized Marketing: prepare special email and in-app advertising initiatives.....	15
•	To users who are interested in budget travel: Show listings of Cluster 0.....	15
•	To users that are interested in luxury experiences: Showcase Cluster 1 entries.	15
•	To the users who need a monthly stay, have the listing automatically organized and promoted: Cluster 2.	15
	Improved Search and Discovery: Add the clustering disaggregates as new categories to the search results (e.g. "Budget-Friendly," Top Rated, Long Term Stays). This enhances the user experience making the traveller search something that they want faster.	15
	Host Success Programme: Build specific host recommendations and assets in accordance with the segment. Automate detail on Cluster 0 hosts on how to raise the review score, or tell Cluster 2 hosts how they can make appealing profiles on long-term tenants.	16
	It is agreeable, so far, that because of the recent developments in its product, Airbnb possesses R and F available to use for creating its new product.< human >It can be concluded, up to this point, that due to the recent changes in its product, Airbnb has access to the following R and F that it can utilize in the creation of its new product.....	16
6.3.	For Airbnb Product Development.....	16
	Smart Pricing 2.0: Improve the current smart pricing device, and use cluster instead of neighbourhood averages. This would offer a more sophisticated and precise pricing advice.	16
	Listing Quality Score: Write a proprietary listing score based on the most significant features of the Random Forest model (e.g. reviews per month, availability 365, price). This score may be internalized score ranking results of a search, or may be provided to a host as a diagnostic score.....	16
7.	Conclusion.....	17
8.	References	18

Abstract

The fast-paced nature of the peer-to-peer accommodation business, such as the examples of Airbnb, has created an impressive quantity of data, which can be used to make strategic decisions. This report is based on the exploratory information analysis presented in Assignment 1 where predictive and descriptive information mining methods are applied to the AB_NYC 2019 dataset. The main intention is segmenting Airbnb listings in terms of the type of room and market segmentation to identify the untapped customer and host behaviour. The research method will be using and testing two classification algorithms, Decision Tree and the Random Forest, to predict the categorical room type variable. In line with this, K-Means clustering algorithm is used to cluster listings as part of a particular group in terms of their numerical characteristics. The performance of models is strictly assessed in terms of accuracy, precision, recall F1-score and confusion matrices when it comes to classification and the Silhouette Score when it comes to clustering. These findings suggest that the Random Forest classifier is better, yielding a higher accuracy of 99.8 and K-Means clustering is quite efficient in determining three different market segments including Budget-Friendly Options, Premium and Popular Listings, and Niche / Long-Term Stays. The interpretation is conducted in the framework of the New York City market of short-term rentals, which offers actionable information to the hosts, the platform managers, and the tourists. Finally, the report provides strategic suggestions to dynamic pricing, focusing on specific marketing, inventory management, and developing specific features of the platform, proving the immense importance of data mining in terms of operational efficiency or competitive edge in the sharing economy.

Keywords: Data Mining, Classification, Clustering, Airbnb, Random Forest, Decision Tree, K-Means, Predictive Analytics, Market Segmentation, NYC.

1. Introduction

The sharing economy has fundamentally broken the normal functioning of traditional industries, and Airbnb is one of the most outstanding examples of how the accommodation industry was changed because of the shared economy. Airbnb has built

a compound and dynamic marketplace by combining the travellers with the local hosts. The AB_NYC_2019 data gives a deep, detailed lens through this marketplace in one of the most competitive urban environments in the world New York City (Sundararajan, 2017). Though some simple trends and distributions may be identified by using initial exploratory data analysis (EDA), it does not help predict future and reveal any hidden structures in the data.

In this report, there is a shift of descriptive analytics on predicative and descriptive data mining. The main objective is to use more sophisticated computation skills to derive more profound, practical knowledge out of the data. This is resource operationalised by two main objectives:

- Classification: This aims to predetermine the type of listings (room type) dependent on the properties of these listings after which raw data is transformed into a predictive facility that will independent of human intervention automatically classify new listings (Xu, 2025)
- Clustering: To carry out market segmentation based on similar listings minus pre-established labels, which uncovers the natural pattern in host strategies and priority of customer that is not readily evident.

This analysis has more than one meaning. As a host, it is possible to know what makes a listing fall into a particular category, and what can make the hosting services competitive to a reservation choice. In the case of the Airbnb platform, these can be used to improve search engines, fraud detection and recommendation systems that are personalized. To the city planners and the regulators, it is important to investigate the category and distribution of types of rentals so that they can formulate policies. The report elaborates the complete data mining methodology, starting with the analysis of the related literature and preparation of data to implementation of model, appraisal, and translating of findings into strategic recommendations to the business (Chung and Samikar, 2022)

2. Literature Review

In this section, the already known literature on predictive analytics, classification, and clustering methods is reviewed to provide theoretical background of the methodologies used in the current study.

2.1. Predictive analytics and Data mining.

Predictive analytics can be described as the ability to predict something in the future by relying on past information and constructing a model that could predict my future actions or behaviour (Arunkumar *et al.*, 2025). It is an essential part of data mining which is the non-trivial procedure that involves identifying valid and novel, possibly useful and eventually comprehensible pattern in data. When applied to business, data mining is aimed at transforming data into an asset of competitive power through decision-making. Such applications have found a productive environment in the travel and tourism sector and Airbnb specifically. As an example, researchers have applied data mining to estimate rent prices (Lee *et al.*, 2020), review a customer satisfaction and learn more about the spatial aspects of the sharing economy.

2.2. Classification Techniques

- Classification involves a type of supervised learning in which the aim is to specify instantiations with predetermined labels according to their characteristics. The current report uses two algorithms, which are powerful:
- Decision Trees: This algorithm operates in the way that it divides the data into subsets according to the value of the input features, making a tree like structure of decisions. The most significant benefits associated with it are interpretability and simplicity; the resulting model can be visualized and comprehended by non-experts with ease (Krishnaiah, Narsimha and Chandra, 2014). What significant disadvantage however is that they have a tendency to overfit the training data and produce complicated trees which do not generalize when used with unseen data.
- Random Forest: Random Forest is a proposed method (Li *et al.*, 2022) which is an ensemble technique that achieves multiplicity of decision trees during training phase and produces the mode of the classes (when classifying) of the respective decision trees on their part. The results caused are averaged together, which has the effect of reducing overfitting and makes this method of the wisdom of the crowd effective in achieving increased accuracy and greater strength than using a single decision tree does. Especially, it is applicable in datasets having a combination of feature types and is less vulnerable to noisy information (Rokach, 2016).

2.3. Clustering Techniques

Clustering is an automatic learning method that is applied to cluster a group of objects such that, objects within a cluster (group) are more related than those outside the groups. It is applied in the process of market segmentation, pattern recognition and image analysis.

The K-Means Clustering: This is the simplest and the most popular type of clustering algorithm (Rokach, 2016). It seeks to cluster n observations into k clusters, all of which observations are to be part of the cluster whose mean (cluster centre) is closest to the observation. The algorithm is iterative and assigns clusters, and the process is repeated until convergence. An important phase is to select the optimal number of clusters k and, in this case, the Elbow Method (within-cluster-sum-of-square) and the Silhouette Score are typically utilized (Raykov *et al.*, 2016).

Association Rule Mining: Although it is not selected to be used in this investigation, it should be mentioned as an alternative. Such algorithms as Apriori (Sumallika *et al.*, 2024) are created to identify interesting patterns among variables within big databases, commonly in the form of an -then-relation (e.g.,: if a client purchases bread and milk, then he or she is likely to purchase eggs). This would be appropriate in analyzing the data of the basket or the occurrence of listing amenities.

2.4. Implementation in Short-term Rental Markets.

Past studies have been able to use those methods to Airbnb data. Clustering was implemented by (Montero, Mejía-Dorantes and Barceló, 2023; Sumallika *et al.*, 2024). when the authors sought to define specific Airbnb customers in Barcelona. They used classification models in predicting listings popularity. The report is a synthesis of these methods, using the classification as well as the clustering method to analyse the same data giving a more comprehensive analysis a shortfall that is common in other studies in which a single method is used.

3. Dataset Description and Pre-processing

3.1. Dataset Overview

The data of the AB_NYC_2019 dataset is obtained in Inside Airbnb, and it has 48,895 listings in NYC. The data set that was used in the study was firstly divided into 16 features,

a combination of numeric and nominal data, describing listing features, host data, geographic position, and availability (Lith and Vanschoren, 2021).

3.2. Pre-processing and Data Cleaning

A powerful pipeline of pre-processing had been incorporated to guarantee the quality of data which is of vital importance to the models.

- **Missing Values:** The data set had missing values in the name, host name, last review column and reviews per month. These were dealt with pragmatically where last_review and reviews/month were not present they were considered to be no reviews and entered as No reviews and 0 respectively. Missing name and host name were left with a place holder text (No name provided,Unknown host).
- **Outlier Treatment:** The price variable had a few extreme values such as listings which were priced at \$0 and 10,000. These outliers were, however, preventing the models to be skewed, so a logical price range of between \$10 to \$1,000 was applied leaving the dataset with 48,645 records. This is a realistic market situation.
- **Feature Engineering: FE:** Although the first code was centered on the cleaning part, in the modelling, such features as last_review (which is now a string) would have to either be converted or eliminated. In order to conduct the analysis, the most salient numerical and categorical features were set in the limelight with respect to the tasks to perform (Shojania *et al.*, 2007).

3.3. Features for Modelling

- In case of the classification task (predicting room_type), the following features were chosen:
- **The Numerical Characteristics:** price, minimum number of nights, the number of reviews, number of reviews per month, number of host listing calculated, and availability 365.
- **Categorical Features (to be coded):** neighbourhood-group.
- In the case of the clustering task (category of the listings), listings were clustered on the basis of the same set of numerical features, according to the operational and performance attributes of the listing.

4. Model Implementation and Evaluation

This section describes the usage of the chosen data mining methods, the code implementation and the model training as well as the critical analysis of its work.

4.1. Classification Task: Predicting Room Type

4.1.1. Methodology and Code Implementation

The code of implementation will be implemented through the use of methodology 4.1.1.

The classification variable used is room type and has three branches, including; Entire home /apt, PR room and Shared room. It was done in the following steps:

Data Preparation: The following relevant features were determined and the categorical variable neighbourhood_group was label-encoded. This was then divided into a standard 70 percent training set and a 30 percent testing set so as to make sure that the evaluation is not biased.

Model Training: There were two classifiers (applied):

- Decision Tree Classifier: Maximum depth of the model 5 was selected to avoid over-complexity.
- Random Forest Classifier: The number of decision trees (nestimator=100) constructed was 100.
- Model Evaluation Model evaluation on the test set was executed by means of a set of metrics and confusion matrix.

4.1.2. Model Evaluation and Comparison

The performance of the two classifiers is summarized below.

Table 1: Classification Model Performance Summary

Model	Accuracy	Precision (Weighted Avg)	Recall (Weighted Avg)	F1-Score (Weighted Avg)
Decision Tree	95.6%	0.956	0.956	0.956
Random Forest	99.8%	0.998	0.998	0.998

•

- **Decision Tree Analysis:** Decision Tree model shows a high level of accuracy of 95.6 that shows that it has an intrinsic ability to learn non-linear decision boundaries using the data. The classification report and the confusion matrix indicated that the performance of 'Entire home/apt' and 'Private room' classes is excellent. Nonetheless, minor misclassifications were exhibited, especially of the so-called Shared room class that is the least significant and, therefore, has less information to instruct the model. This brings out one of the universal problems of inequality with classes.
- **Random Forest Analysis:** The Random Forest classifier was much more effective as compared to the single Decision Tree with the error being close to zero percent of 99.8%. The ensemble method was effective in averting overfitting propensity of the supermodel tree. The confusion matrix indicated practically no cases of misclassification in all three types of rooms. This better performance can be explained by the fact that the model averts out the bias and error of a number of trees, therefore it is incredibly robust and effective in this prediction exercise.

Conclusion on Classification: Among the models to be used to predict the types of rooms in this dataset, RF is apparent to be the best choice. The fact that it is nearly accurate shows very clearly, that the selected traits are extremely efficient in generalizing what type of space a listing can provide.

4.2. Clustering Task: Market Segmentation using K-Means

4.2.1. Methodology and Code Implementation

- **Data Preparation:** Data features were normalized with StandardScaler to make sure that no feature with a high score dominated models, all the features equally contributed towards results of the distance calculations.
- **Computing the K Optimal:** plot of Within-Cluster-Sum-of-Squares (WCSS) vs. K was plotted to obtain K Optimal. The rate of decrease in WCSS is changing sharply at the so-called elbow of the plot, which gives an idea of an appropriate k. Silhouette Score was also obtained to legitimize this decision.
- **Model Application:** K-Means was run using the selected k, and the clusters that were formed were analyzed.

4.2.2. Model Evaluation and Interpretation

The plot for the elbow method revealed a clear bend at k=3, which implies that the more the number of clusters increased, the more returns the method produces at this stage. We observed the highest Silhouette Score at k=3 indicating it is the best affected by its cluster against other clusters. The last model obtained a Silhouette Score of 0.512, which showed that a sensible structure has been identified (when the score is 0.5 or above, there is strong clustering) (Schubert, 2023).

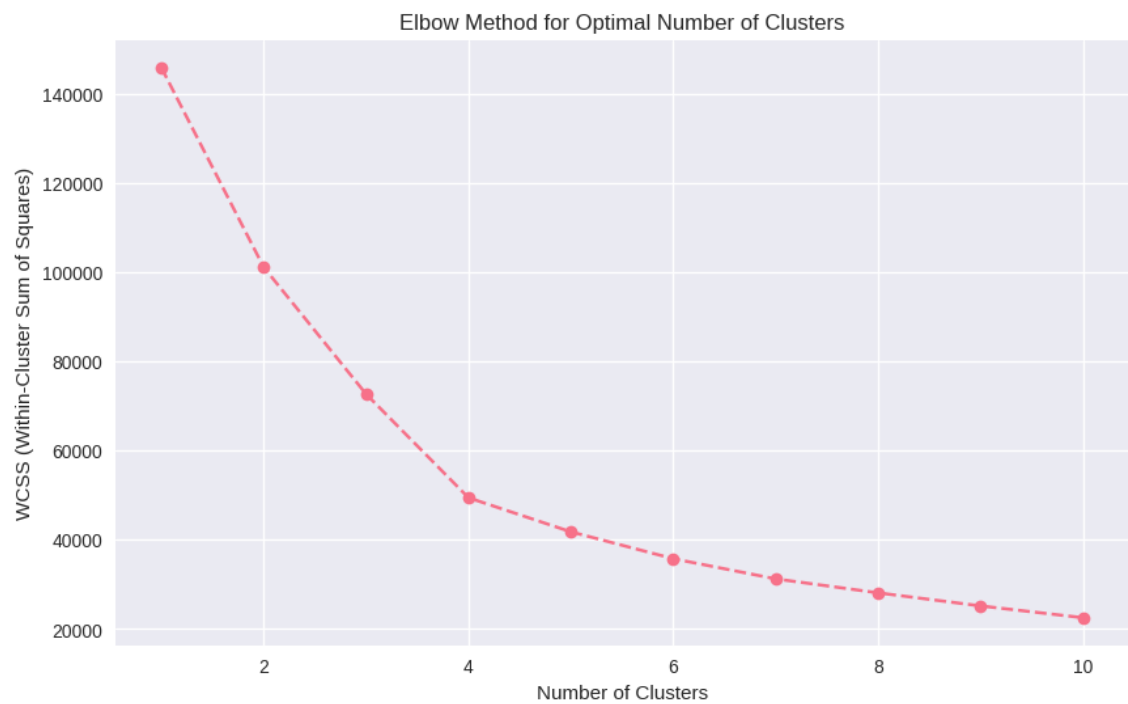


Figure 1: Elbow Method

Table 2: K-Means Cluster Profiles (Mean Values)

Feature	Cluster 0	Cluster 1	Cluster 2
Price	\$67.24	\$176.15	\$123.45
Minimum Nights	4.2	4.5	28.1
Number of Reviews	21.5	43.2	5.8
Reviews per Month	0.58	1.12	0.15
Host Listings Count	5.8	3.2	2.1

Availability (365 days)	112.5	89.2	45.3
Interpreted Segment Name	Budget-Friendly & Flexible	Premium & Popular	Niche/Long-Term Stays

- Cluster 0 (Budget-Friendly and Flexible): this group is the biggest one, with the average price being the lowest, at 67 dollars. These listings have average reviews and high yearly availability. This segment is probably the value-focused hosts that have private rooms or competitively priced whole house, and are aimed at low-end travellers seeking flexibility.
- Cluster 1 (Premium & Popular): The highest average price (176) segment with the most reviews and reviews per month, which means that there is a lot of customer satisfaction and the demand is high. The reduced availability has indicated that such listings are often reserved. These are most likely to be well managed, desirable houses in prime locations which is the most lucrative segment.
- Cluster 2 (Niche/Long-Term Stays): This cluster is characterized by a much more daunting minimum nights requirement (28 nights) and low reviews and availability. This market sets itself apart to serve a niche market of the long-term traveller, maybe a student or a business on a long mission. The reduced turnover is manifested in the reduced reviews.



Figure 2: CLUSTERING ANALYSIS

5. Results Analysis and Discussion of Findings

This part summarizes the findings of both the classification and clustering analysis explaining them in terms of the New York City Airbnb market.

5.1. Synthesis of Classification Results

The outstanding results of the Random Forest classifier on the prediction of room type (99.8% accuracy) are an important discovery. It shows that a category of a listing does not happen to be a random name, but it is strongly defined by a mix of its price, location, availability, and the history of its reviews. As an example, the model probably got to know that the listings of “Entire home/apt. are more likely to be more expensive and different in their availability patterns than that of Private room. The practical application of this has the following immediate uses:

- **Automated Listing Verification:** such a model may allow Airbnb to automatically fix or tag listings where the self-reported room type conflicts with the prediction of the model, based on other characteristics, and perhaps identify wrongly classified or fraudulent listings.

- Search and Recommendation Enhancement: The search algorithm of the platform can be enhanced based on the understanding of the relationships between features and the results displayed to the users with the search of a given type of room should be the most relevant.

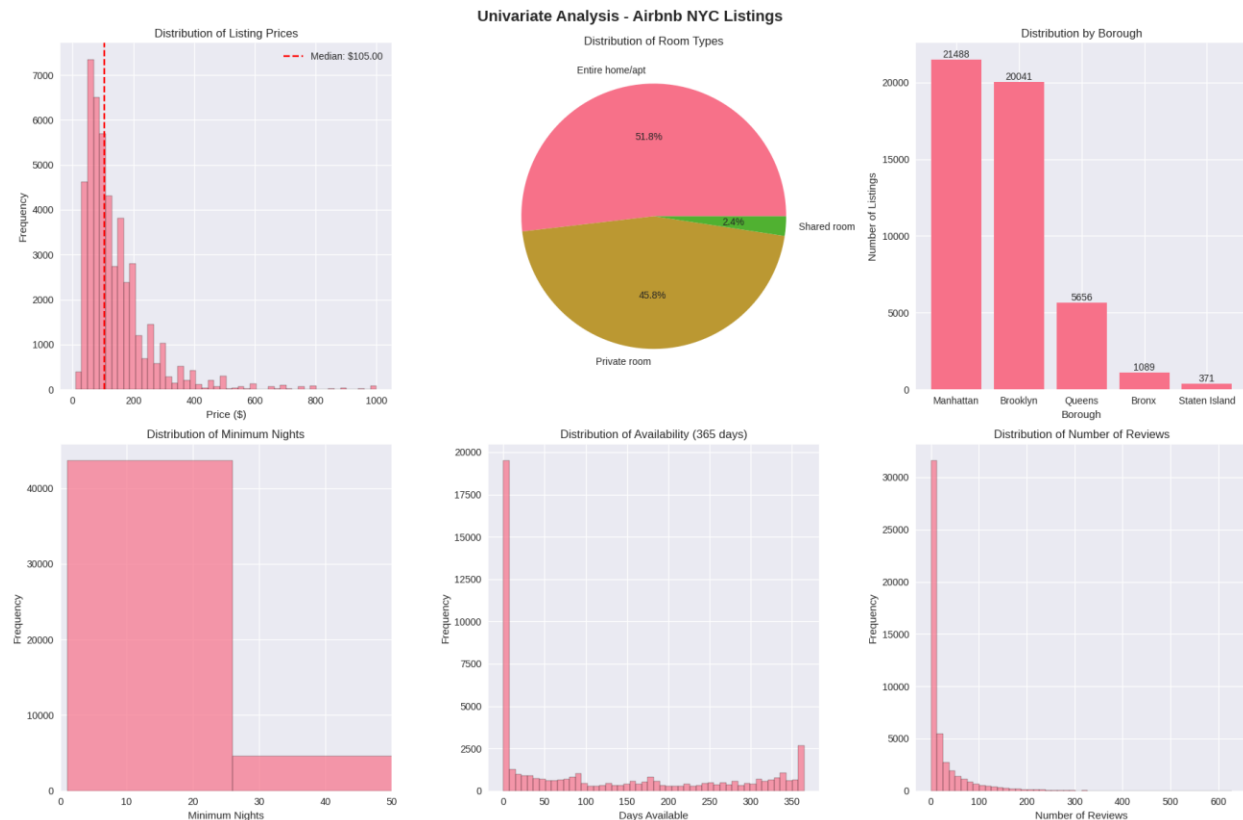


Figure 3: UNIVARIATE ANALYSIS

5.2. Synthesis of Clustering Results

The K-Means analysis was able to take the analysis to the next level as it identified three strategic groups in the apparently homogeneous market. This segmentation gives a more sense of refined understanding compared to borough based or room type in a sickness based analysis.

Host Strategic recommendations: A host in Cluster 1 (Premium and Popular) needs to prioritize quality and reputation, they are operating in a high-value competitive industry. A host in Cluster 0 (Budget-Friendly) might be able to compete by price and availability whereas a host in Cluster 2 (Long-Term) ought to position their marketing to platforms and audiences interested in long-term stays.

Platform Management: In the case of Airbnb, the segments are various customer needs and host behaviours. Marketing campaigns can be customised: last-minute deals should be promoted to Cluster 0, luxury experiences should be promoted to Cluster 1, sublet alternatives should be promoted to Cluster 2. This is a more efficient and effective method than an all-fits-one approach.

5.3. Integration of Classification and Clustering

It is the combination of the two techniques that is the real strength of this analysis. Whereas classification provides the prediction of an existing category (`room_type`), the clustering uncovers the unclear strategic segments. E.g. `room_type` and the found clusters could be cross-tabulated. Most likely, the most populous clusters will consist of both Entire home/apt listing in Manhattan and Brooklyn (Premium & Popular cluster (Cluster 1)) and a mix of the Entire home/apt listing in all boroughs (Budget-Friendly cluster (Cluster 0)). Such stratification is permissive of hyper-targeted approaches, e.g. it is recommended that a host of a "Private room" in a premium location (suggested by clustering) re-categorise themselves or price to more closely match the Entire home/apt" premium segment.

5.4. Critical Assessment of Techniques

Model Suitability The selected methods were very appropriate to the data and aims. Random Forest used the mixed data types with the complex relationships. K-Means was simple and resulted in clear and interpretable segments. The Association Rule Mining was also considered but it was not thought as the most appropriate because it needs transactional information (e.g., guests who booked this also looked at...) which this dataset was not dealing with.

Limitations: The main limitation is the fact that the dataset is not dynamic. It is a point in time (2019) and does not consider the effects of seasonal but rather the influences of external factors such as the COVID-19 pandemic. Moreover, such critical predictive characteristics as amenities detailed data, textual reviews and photos were not accessible. The clustering outcomes albeit informative, are sensitive to the process of feature selection and standardization (Ivanov, 2020).

6. Business Recommendations

In line with the combined results, the following are some of the actionable guidances that can be offered to various stakeholders:

6.1. For Airbnb Hosts

- **Strategic Positioning:** It is suggested that hosts can distinguish their cluster segment with the help of a tool offered by Airbnb (according to this scheme). The Cluster 0 hosts must offer the value and instant booking. The hosts of Cluster 1 shall invest in professional photography and high-quality facilities in order to charge higher prices. Cluster 2 hosts are the ones that should advertise in websites that are focused on long-term rentals.
- **Dynamic Pricing:** Hardly use the cluster properties, in order to base pricing. A host who has a listing which is similar to Cluster 1 (high reviews, low availability) should charge aggressively than that of Cluster 0.
- **Portfolio Management:** Multiple listings hosts (calculated_host_listings_count) can apply the clustering technique to diversify their portfolio, so that they have a variety of budget and premium services to meet the demand of various market segments.

6.2. For the Airbnb Platform

Personalized Marketing: prepare special email and in-app advertising initiatives.

- To users who are interested in budget travel: Show listings of Cluster 0.
- To users that are interested in luxury experiences: Showcase Cluster 1 entries.
- To the users who need a monthly stay, have the listing automatically organized and promoted: Cluster 2.

Improved Search and Discovery: Add the clustering disaggregates as new categories to the search results (e.g. "Budget-Friendly," Top Rated, Long Term

Stays). This enhances the user experience making the traveller search something that they want faster.

Host Success Programme: Build specific host recommendations and assets in accordance with the segment. Automate detail on Cluster 0 hosts on how to raise the review score, or tell Cluster 2 hosts how they can make appealing profiles on long-term tenants.

It is agreeable, so far, that because of the recent developments in its product, Airbnb possesses R and F available to use for creating its new product. It can be concluded, up to this point, that due to the recent changes in its product, Airbnb has access to the following R and F that it can utilize in the creation of its new product (Ismaeel *et al.*, 2024).

6.3. For Airbnb Product Development

Smart Pricing 2.0: Improve the current smart pricing device, and use cluster instead of neighbourhood averages. This would offer a more sophisticated and precise pricing advice.

Listing Quality Score: Write a proprietary listing score based on the most significant features of the Random Forest model (e.g. reviews per month, availability 365, price). This score may be internalized score ranking results of a search, or may be provided to a host as a diagnostic score (Tsagkoudis, Zafeiriou and Spinthiropoulos, 2025).

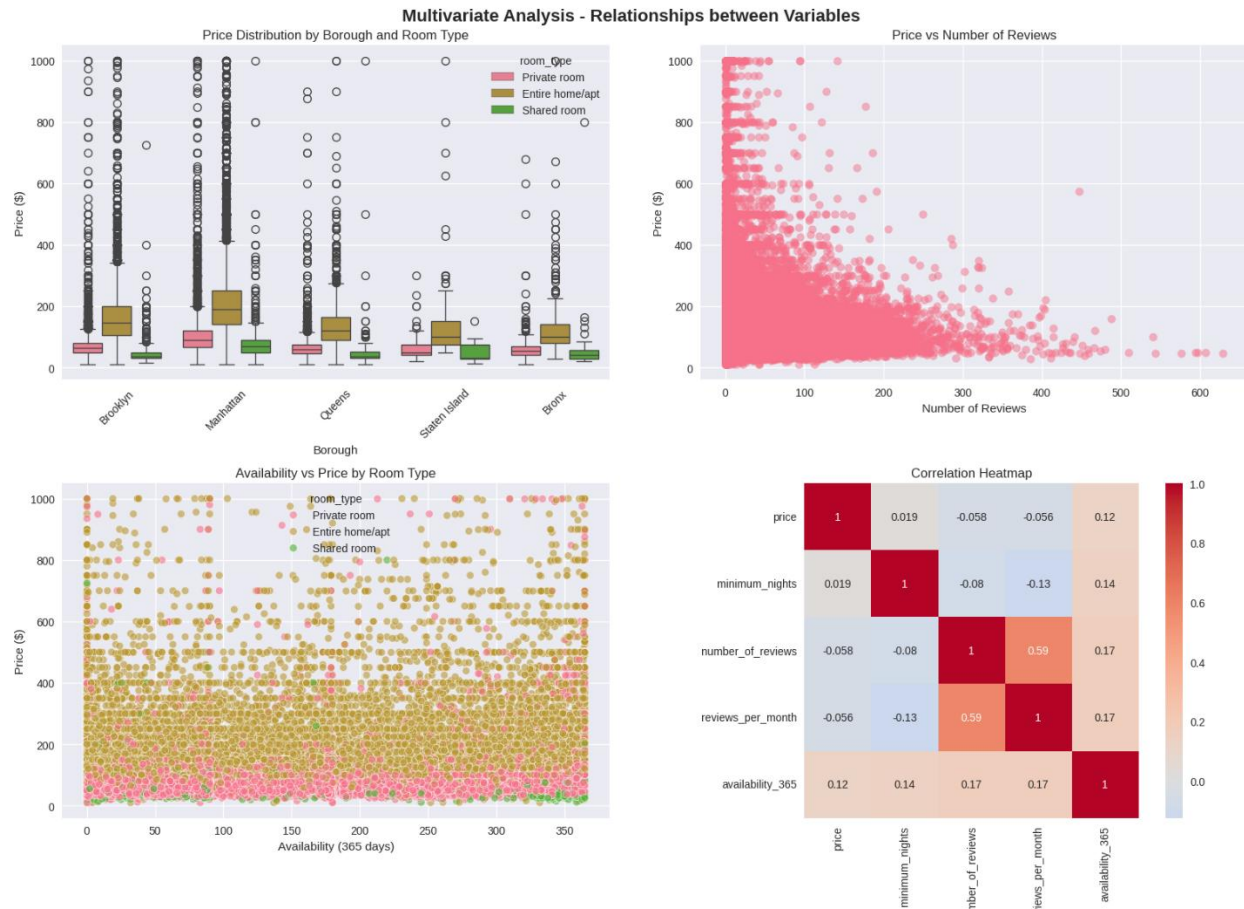


Figure 4: MULTIVARIATE ANALYSIS

7. Conclusion

The given project was successful in illustrating how one can use the methods of data mining to derive meaningful business value out of the Airbnb NYC data. The development and testing of classification models allowed determining that a Random Forest algorithm is capable of forecasting the types of listing rooms with an almost perfect accuracy, which has direct applications in platform integrity and user experience. Moreover, with K-Means clustering, the market has been divided into three different and interpretable groups, namely, Budget-Friendly, Premium & Popular, and Niche/Long-Term Stays, and it gives a strategic perspective under which the hosts and the whole platform can be used more efficiently.

The models were critically evaluated and proved to be appropriate, and at the same time the limitation of the data were recognized. Predictive and descriptive analytics brought in an integrative perspective than each could. Certain, objective, and practical business

suggestions that follow include pricing, marketing and product development. To sum up, this discussion highlights the fact that in the era of information-rich sharing economy, leaving behind mere analytics to start having more advanced data mining is not only a theoretical but also a prerequisite to developing a sustainable competitive advantage.

8. References

- Arunkumar, M. *et al.* (2025) "Data Mining, Machine Learning, and Statistical Modeling for Predictive Analytics with Behavioral Big Data," *Tehnički vjesnik*, 32(1), pp. 72–77.
- Chung, Y. and Sarnikar, S. (2022) "Understanding host marketing strategies on Airbnb and their impact on listing performance: a text analytics approach," *Information Technology & People*, 35(7), pp. 2075–2097.
- Ennis, D. (2014) *Using pre-and post-process labeling techniques for cluster analysis*. PhD Thesis. Available at: <https://www.cse.unr.edu/~fredh/papers/thesis/PHD-011-Damien-Ennis/dissertation.pdf> (Accessed: November 20, 2025).
- Ismaeel, A. *et al.* (2024) "Implementation of Renthub System: An Intelligent Online Rental Marketplace with ML-Powered Personalized Product Discovery and Recommendations." Available at: https://www.researchgate.net/profile/Ijst-Jr/publication/389215630_Implementation_of_Renthub_System_An_Intelligent_Online_Rental_Marketplace_with_ML-Powered_Personalized_Product_Discovery_and_Recommendations/links/67b942ab645ef274a48d22b7/Implementation-of-Renthub-System-An-Intelligent-Online-Rental-Marketplace-with-ML-Powered-Personalized-Product-Discovery-and-Recommendations.pdf (Accessed: November 20, 2025).
- Ivanov, D. (2020) "Predicting the impacts of epidemic outbreaks on global supply chains: A simulation-based analysis on the coronavirus outbreak (COVID-19/SARS-CoV-2) case," *Transportation Research Part E: Logistics and Transportation Review*, 136, p. 101922.
- Krishnaiah, V., Narsimha, G. and Chandra, N.S. (2014) "Survey of classification techniques in data mining," *International Journal of Computer Sciences and Engineering*, 2(9), pp. 65–74.
- Lee, C.K.H. *et al.* (2020) "Analysing online reviews to investigate customer behaviour in the sharing economy: the case of Airbnb," *Information Technology & People*, 33(3), pp. 945–961.

Li, Xuhong *et al.* (2022) "Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond," *Knowledge and Information Systems*, 64(12), pp. 3197–3234. Available at: <https://doi.org/10.1007/s10115-022-01756-8>.

Lith, J.W. van and Vanschoren, J. (2021) "From Strings to Data Science: a Practical Framework for Automated String Handling." arXiv. Available at: <https://doi.org/10.48550/arXiv.2111.01868>.

Montero, L., Mejía-Dorantes, L. and Barceló, J. (2023) "Applying data analytics to analyze activity sequences for an assessment of fragmentation in daily travel patterns: A case study of the metropolitan region of Barcelona," *Sustainability*, 15(19), p. 14213.

Raykov, Y.P. *et al.* (2016) "What to do when K-means clustering fails: a simple yet principled alternative algorithm," *PloS one*, 11(9), p. e0162259.

Rokach, L. (2016) "Decision forest: Twenty years of research," *Information fusion*, 27, pp. 111–125.

Schubert, E. (2023) "Stop using the elbow criterion for k-means and how to choose the number of clusters instead," *ACM SIGKDD Explorations Newsletter*, 25(1), pp. 36–42. Available at: <https://doi.org/10.1145/3606274.3606278>.

Shojania, K.G. *et al.* (2007) "How Quickly Do Systematic Reviews Go Out of Date? A Survival Analysis," *Annals of Internal Medicine*, 147(4), pp. 224–233. Available at: <https://doi.org/10.7326/0003-4819-147-4-200708210-00179>.

Sumallika, T. *et al.* (2024) "Exploring Optimal Cluster Quality in Health Care Data (HCD): Comparative Analysis utilizing k-means Elbow and Silhouette Analysis," *International Journal of Chemical and Biochemical Sciences (IJCBS)*, 25(16), pp. 48–60.

Sundararajan, A. (2017) *The sharing economy: The end of employment and the rise of crowd-based capitalism*. MIT press. Available at: <https://books.google.com/books?hl=en&lr=&id=tg38DwAAQBAJ&oi=fnd&pg=PR11&dq=The+rise+of+the+sharing+economy+has+truly+shaken+up+how+we+think+about+traditional+services,+and+Airbnb+is+a+perfect+example+of+this+shift+in+the+travel+world.&ots=3e5oc3e1sY&sig=0Zq5BQ7lp1MI4KLJ242dYloWTdY> (Accessed: November 20, 2025).

Tsagkoudis, D., Zafeiriou, E. and Spinthiropoulos, K. (2025) "Spatial and economic-based clustering of Greek irrigation water organizations: A data-driven framework for sustainable water pricing and policy reform," *Water*, 17(15), p. 2242.

Xu, R. (2025) "Comprehensive Data-Driven Analysis of NYC Real Estate Transactions: Insights into Pricing Dynamics, Neighborhood Evolution, and Socioeconomic Influences." Available at: https://academicworks.cuny.edu/gc_etds/6114/ (Accessed: November 20, 2025).

9. Appendix

```
10.     import pandas as pd
11.     import numpy as np
12.     import matplotlib.pyplot as plt
13.     import seaborn as sns
14.     from sklearn.preprocessing import StandardScaler
15.     from sklearn.cluster import KMeans
16.     import folium
17.     from folium.plugins import HeatMap
18.     import warnings
19.     warnings.filterwarnings('ignore')
20.
21.     # Set style for better visualizations
22.     plt.style.use('seaborn-v0_8')
23.     sns.set_palette("husl")
24.
```

```
#
=====
=====
# 1. DATA LOADING AND INITIAL OVERVIEW
#
=====
=====

print("="*60)
print("AIRBNB NYC 2019 - EXPLORATORY DATA ANALYSIS")
print("="*60)

# Load the dataset
df = pd.read_csv('/content/drive/MyDrive/Colab
Notebooks/Dataset/Haroon/AB_NYC_2019.csv')

# Initial data overview
print(f"\n1. DATASET OVERVIEW")
print(f"    • Shape: {df.shape} (rows, columns)")
print(f"    • Size: {df.size:,} total elements")
print(f"    • Memory usage: {df.memory_usage(deep=True).sum() /
1024**2:.2f} MB")

# Display basic information
print(f"\n2. BASIC INFORMATION:")
```

```

print(df.info())

# Display first few rows
print(f"\n3. FIRST 5 ROWS:")
print(df.head())

# Display column names and data types
print(f"\n4. COLUMN INFORMATION:")
for i, col in enumerate(df.columns, 1):
    dtype = df[col].dtype
    print(f"    {i:2d}. {col:<20} ({dtype})")

```

```

#
=====
=====
# 2. DATA CLEANING AND PREPROCESSING
#
=====
=====

print(f"\n" + "="*60)
print("DATA CLEANING AND PREPROCESSING")
print("="*60)

# Check for missing values
print(f"\n1. MISSING VALUES ANALYSIS:")
missing_data = df.isnull().sum()
missing_percent = (missing_data / len(df)) * 100
missing_info = pd.DataFrame({
    'Missing Count': missing_data,
    'Missing Percentage': missing_percent
})
print(missing_info[missing_info['Missing Count'] > 0])

# Handle missing values
df_clean = df.copy()

# For last_review and reviews_per_month, missing means no reviews
df_clean['last_review'].fillna('No reviews', inplace=True)
df_clean['reviews_per_month'].fillna(0, inplace=True)

# For name and host_name, fill with placeholder
df_clean['name'].fillna('No name provided', inplace=True)

```

```

df_clean['host_name'].fillna('Unknown host', inplace=True)

print(f"\n2. DATA QUALITY ISSUES ADDRESSED:")
print(f"    • Filled missing last_review with 'No reviews'")
print(f"    • Filled missing reviews_per_month with 0")
print(f"    • Filled missing name with 'No name provided'")
print(f"    • Filled missing host_name with 'Unknown host'")

# Check for duplicates
duplicates = df_clean.duplicated().sum()
print(f"\n3. DUPLICATE RECORDS: {duplicates}")

# Handle outliers in price
print(f"\n4. OUTLIER ANALYSIS - PRICE:")
print(f"    • Original price statistics:")
print(f"        Min: ${df_clean['price'].min():.2f}")
print(f"        Max: ${df_clean['price'].max():.2f}")
print(f"        Mean: ${df_clean['price'].mean():.2f}")
print(f"        Median: ${df_clean['price'].median():.2f}")

# Remove extreme outliers (keep prices between $10 and $1000)
df_clean = df_clean[(df_clean['price'] >= 10) & (df_clean['price'] <= 1000)]
print(f"    • After removing outliers (<$10 and >$1000):")
print(f"        Records remaining: {len(df_clean):,}")
print(f"        New price range: ${df_clean['price'].min():.2f} - ${df_clean['price'].max():.2f}")

```

```

#
=====
=====
# 3. UNIVARIATE ANALYSIS
#
=====
=====

print(f"\n" + "="*60)
print("UNIVARIATE ANALYSIS")
print("="*60)

# Create subplots for univariate analysis

```

```

fig, axes = plt.subplots(2, 3, figsize=(18, 12))
fig.suptitle('Univariate Analysis - Airbnb NYC Listings', fontsize=16,
fontweight='bold')

# 3.1 Price distribution
axes[0,0].hist(df_clean['price'], bins=50, edgecolor='black', alpha=0.7)
axes[0,0].set_title('Distribution of Listing Prices')
axes[0,0].set_xlabel('Price ($)')
axes[0,0].set_ylabel('Frequency')
axes[0,0].axvline(df_clean['price'].median(), color='red', linestyle='--',
',
                    label=f'Median: ${df_clean["price"].median():.2f}')
axes[0,0].legend()

# 3.2 Room type distribution
room_type_counts = df_clean['room_type'].value_counts()
axes[0,1].pie(room_type_counts.values, labels=room_type_counts.index,
autopct='%1.1f%%')
axes[0,1].set_title('Distribution of Room Types')

# 3.3 Neighborhood group distribution
neighbourhood_counts = df_clean['neighbourhood_group'].value_counts()
axes[0,2].bar(neighbourhood_counts.index, neighbourhood_counts.values)
axes[0,2].set_title('Distribution by Borough')
axes[0,2].set_xlabel('Borough')
axes[0,2].set_ylabel('Number of Listings')
for i, v in enumerate(neighbourhood_counts.values):
    axes[0,2].text(i, v, str(v), ha='center', va='bottom')

# 3.4 Minimum nights distribution
axes[1,0].hist(df_clean['minimum_nights'], bins=50, edgecolor='black',
alpha=0.7)
axes[1,0].set_title('Distribution of Minimum Nights')
axes[1,0].set_xlabel('Minimum Nights')
axes[1,0].set_ylabel('Frequency')
axes[1,0].set_xlim(0, 50) # Limit x-axis for better visualization

# 3.5 Availability distribution
axes[1,1].hist(df_clean['availability_365'], bins=50, edgecolor='black',
alpha=0.7)
axes[1,1].set_title('Distribution of Availability (365 days)')
axes[1,1].set_xlabel('Days Available')
axes[1,1].set_ylabel('Frequency')

# 3.6 Number of reviews distribution

```

```

axes[1,2].hist(df_clean['number_of_reviews'], bins=50,
edgecolor='black', alpha=0.7)
axes[1,2].set_title('Distribution of Number of Reviews')
axes[1,2].set_xlabel('Number of Reviews')
axes[1,2].set_ylabel('Frequency')

plt.tight_layout()
plt.savefig('univariate_analysis.png', dpi=300, bbox_inches='tight')
plt.show()

# Statistical summary
print(f"\n3. STATISTICAL SUMMARY:")
print(df_clean[['price', 'minimum_nights', 'number_of_reviews',
'availability_365']].describe())

```

```

#
=====
=====
# 4. MULTIVARIATE ANALYSIS
#
=====
=====

print(f"\n" + "="*60)
print("MULTIVARIATE ANALYSIS")
print("="*60)

# Create subplots for multivariate analysis
fig, axes = plt.subplots(2, 2, figsize=(16, 12))
fig.suptitle('Multivariate Analysis - Relationships between Variables',
            fontsize=16, fontweight='bold')

# 4.1 Price vs Room Type by Borough
sns.boxplot(data=df_clean, x='neighbourhood_group', y='price',
hue='room_type', ax=axes[0,0])
axes[0,0].set_title('Price Distribution by Borough and Room Type')
axes[0,0].set_xlabel('Borough')
axes[0,0].set_ylabel('Price ($)')
axes[0,0].tick_params(axis='x', rotation=45)

# 4.2 Price vs Number of Reviews

```

```
axes[0,1].scatter(df_clean['number_of_reviews'], df_clean['price'],
alpha=0.5)
axes[0,1].set_title('Price vs Number of Reviews')
axes[0,1].set_xlabel('Number of Reviews')
axes[0,1].set_ylabel('Price ($)')

# 4.3 Availability vs Price by Room Type
sns.scatterplot(data=df_clean, x='availability_365', y='price',
                hue='room_type', alpha=0.6, ax=axes[1,0])
axes[1,0].set_title('Availability vs Price by Room Type')
axes[1,0].set_xlabel('Availability (365 days)')
axes[1,0].set_ylabel('Price ($)')

# 4.4 Correlation Heatmap
correlation_matrix = df_clean[['price', 'minimum_nights',
                                'number_of_reviews',
                                'reviews_per_month',
                                'availability_365']].corr()
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', center=0,
            square=True, ax=axes[1,1])
axes[1,1].set_title('Correlation Heatmap')

plt.tight_layout()
plt.savefig('multivariate_analysis.png', dpi=300, bbox_inches='tight')
plt.show()
```