Exploratory Data Analysis - Report and Demonstration

Student Name: Mohammad Lakhi

Student ID: 2512430

Subject Code: ACCA7015

Subject Name: Data Mining & Visualization

University of Wales Trinity Saint David

Date: 21st November 2025

Abstract

The research uses data mining techniques to predict Airbnb listing prices and identify market segments in the New York City short-term rental market. Using the AB_NYC_2019 dataset, which contains 48,895 listings, the research employs supervised classification via Random Forest and Logistic Regression algorithms, as well as unsupervised clustering via K-Means, to discover operational patterns among hosts. Data cleaning and exploratory analysis revealed that room type and neighbourhood location are the most important price determinants, with entire apartments commanding higher prices than private rooms, and Manhattan and Brooklyn locations commanding premium prices. Random Forest had 68.99% classification accuracy compared to 64.75% for Logistic Regression, but both algorithms struggled with the premium price category due to class imbalance. K-Means clustering showed three distinct market segments: traditional sharing homes operating private rooms in Brooklyn, semi-professional hosts managing entire apartments in Queens with high availability, and commercial multi-listing operators in Manhattan managing large portfolios.The findings show that the Airbnb market is diverse, with various host types pursuing different strategies, and that combining supervised and unsupervised learning yields complementary insights into market structure. These findings have practical implications for hosts looking to position their listings more competitively, as well as policymakers developing regulations that distinguish between occasional home-sharing and commercial operations.

TABLE OF CONTENTS

1 Introduction **1**

1.1 Context and Background 1

1.2 Research Gap and Approach 1

1.3 Research and Aim Objectives 2

1.4 Significance of the Research 2

2 Literature Review **4**

3 Methodology **9**

3.1 Data Collection 9

3.2 Data Cleaning 9

3.3 Exploratory Data Analysis 10

3.4 Supervised Classification 11

3.5 Unsupervised Classification 11

3.6 Result Analysis 12

4 Results **13**

4.1 Classification Performance Comparison 13

4.2 K-Means Clustering Analysis 15

5 Discussion **17**

5.1 Classification Performance Comparison 17

5.2 Market Segmentation 18

5.3 Determinants of Price Classification 18

5.4 Limitations and research extensions 19

6 Conclusion and Future Scope **20**

6.1 Conclusion 20

6.2 Future Scope 21

REFERENCES **22**

Chapter 1

Introduction

1.1 Context and Background

The rapid growth of peer-to-peer accommodation platforms has presented hosts with complex pricing challenges, as they must independently determine competitive rates in the absence of common guidance systems. Recent studies on Airbnb markets have identified that listing prices are influenced by multiple interacting factors such as property location, room type, host characteristics, and temporal availability patterns. These relationships vary significantly across different urban contexts. The AB_NYC_2019 dataset is one of the most comprehensive publicly available snapshots of short-term rental markets, containing over 48,000 listings across New York City's five boroughs with detailed attributes such as geographical coordinates, pricing structures, review metrics, and host information. Understanding how to effectively classify these listings and identify underlying markets has become critical for both individual hosts looking for pricing strategies and platforms looking to keep their marketplaces competitive[1],[4],[15].

1.2 Research Gap and Approach

Previous machine learning applications on Airbnb data have primarily focused on price prediction using regression techniques, with models ranging from linear regression to ensemble methods like Random Forest and Gradient Boosting, achieving varying degrees of accuracy depending on feature engineering and dataset characteristics . However, divide listings into distinct price categories and using unsupervised clustering to reveal natural market segments represents an unexplored but potentially more actionable approach to strategic decision-making. The AB_NYC_2019 dataset is analyzed

using a dual-method framework that includes supervised classification using Random Forest and Logistic Regression models, as well as unsupervised learning using K-Means clustering [11],[16].

1.3 Research and Aim Objectives

The goal of this study is to use data mining techniques on the AB_NYC_2019 dataset to create classification models that predict Airbnb listing price categories and identify natural market segments using unsupervised clustering. The research objectives are threefold: First, to implement and evaluate supervised classification models including Random Forest and Logistic Regression for categorizing listings into distinct price bands based on property, location, and host attributes, and to assess their performance using standard metrics such as accuracy, precision, recall, and F1-score. Second, to apply K-Means clustering to uncover underlying groupings within the New York City Airbnb market, and examine.Third, to interpret the results of both classification and clustering in a business context, providing actionable insights for hosts seeking to optimize pricing strategies and platform operators aiming to understand market structure. [7][9][12].

1.4 Significance of the Research

This study adds to the growing literature on data-driven decision-making in peer-to-peer accommodation markets. It shows how classification and clustering techniques can be combined to provide predictive and descriptive insights into Airbnb pricing and market structure. This study's classification models provide hosts with a systematic framework to determine their listing's price category based on observable characteristics, allowing for more informed pricing decisions without platforms offering guidance tools. [11][15][16].

The clustering analysis reveals natural market segments within New York City's short-term rental landscape, highlighting how different types of properties and host strategies coexist within the same platform. This information can inform individual host position and platform policy design. The research demonstrates the value of integrating super-

vised and unsupervised learning approaches, and applies established best practices for data cleaning, exploratory analysis, and algorithm selection to a real-world dataset. This demonstrates a practical workflow for similar analyses in other urban contexts or on other sharing economy platforms [1][3][10].

Chapter 2

Literature Review

Research on Airbnb pricing has consistently found that listing prices vary by a variety of factors, including property characteristics, locational attributes, host involvement levels, reputation signals, and rental policies, with the relative importance of these determinants varying by geographical market and city type. Empirical research using hedonic regression approaches has shown that variables such as room type, property size, proximity to city centers, neighborhood income levels, and host response rates all exert significant effects on nightly rates. Models that account for spatial heterogeneity perform better than those assuming uniform price structures across regions [4][15].

Studies on predicting Airbnb prices in various urban contexts revealed distinct patterns across cities. According to research conducted in New York City, borough location, property type, and economic variables are strong predictors of listing prices, with Manhattan commanding significantly higher rates than the outer boroughs. Research on the London and Amsterdam markets has highlighted the importance of seasonal variations, holidays, and local demand fluctuations in determining optimal pricing strategies. Some studies have found that hosts frequently lack access to free and accurate pricing tools, despite the complexity of determining competitive rates [14][15][16].

The methodological foundations for supervised classification in this study are based on studies of Random Forest algorithms and their application to prediction tasks. Random Forest is an ensemble learning method that uses training data to build multiple decision trees and then uses a majority vote across trees to determine the output. The method provides internal validation via out-of-bag error estimation, which uses replacement samples from the training set to test each tree on samples not used in its construction, eliminating the need for a separate validation set. Random Forest can handle non-linear relationships, high-dimensional feature spaces, and predict outliers

or minority classes when properly configured [7].

However, in imbalanced datasets, traditional Random Forest methods may exhibit bias towards majority classes, resulting in poor recall for minority categories. This limitation has prompted the development of improved variants such as HCRF, which combines two key mechanisms to improve classification performance on imbalanced data: random hybrid-resampling and a cost function that favours minority classes. The hybrid resampling approach adjusts training distributions by oversampling minority classes and decreasing majority classes, while the cost function punishes misclassification of minority instances. Experiments on imbalanced benchmark datasets showed that HCRF achieved better F1-scores and minority class recall compared to standard Random Forest. This demonstrates that targeted modifications can address the inherent bias of ensemble methods towards dominant classes [8].

K-Means clustering is still a popular method for detecting natural groupings within data in unsupervised learning, but standard implementations have well-known limitations such as sensitivity to initial centroid placement and difficulty determining the optimal number of clusters. Research on improved K-Means algorithms has addressed these issues in two ways: using the elbow rule to identify appropriate k values based on within-cluster sum of squares, and employing distance-based initialization strategies that select initial centroids far apart to avoid local optima. The improved method calculates the distance between each data point and previously selected centroids, then selects new centroids from points that are the furthest away from existing ones. In comparison to standard random initialization, this improved K-Means achieved better convergence stability and a lower final within-cluster sum of squares, according to experimental results on both synthetic and real datasets.This makes it more appropriate for exploratory clustering where the true number of groups is unknown [9].

Building on these determinant studies, several researchers created machine learning models to predict Airbnb prices based on structured listing features, using algorithms ranging from linear regression and structured regression variants to ensemble methods like Random Forest, Gradient Boosting, and XGBoost. After applying predictive models, these studies usually begin with extensive data cleaning and exploratory analysis

to handle missing values, remove outliers, and engineer relevant features. One analysis of Airbnb pricing using machine learning techniques performed descriptive, prescriptive, and exploratory analysis before applying linear regression, logistic regression, and Random Forest models. This highlights the importance of careful preprocessing for achieving robust predictions [11].

instead of predicting continuous price values, complementary research has used machine learning to classify Airbnb listings into meaningful segments based on neighbourhood attributes and listing features. One study created a machine-learning-based categorization system for New York City listings that classified properties by neighbourhood group, room type, minimum nights, availability, and review metrics. This approach demonstrated that supervised classification techniques can reveal structure within large platform datasets, helping understand how different listing types and host strategies coexist within the same marketplace, providing insights that go beyond simple price prediction to examine market composition and competitive dynamics [12].

Another line of research has looked into listing survival during external shocks, using survival analysis to determine which Airbnb properties in New York City remained active during the Covid-19 pandemic. The study discovered that characteristics such as review counts, host responsiveness, room type, and neighbourhood location were all significant predictors of continued operation, with entire homes in high-demand areas outperforming shared rooms in peripheral neighbourhoods. This survival-oriented perspective that listing characteristics influence not only pricing and demand, but also the fundamental ability of properties to remain competitive during severe market downturns [13].

A study compared k-Nearest Neighbours, multiple linear regression, LASSO, Ridge, Random Forest, Gradient Boosting, and XGBoost to Amsterdam Airbnb data. XGBoost showed the best performance with a R^2 score of 0.6321 and RMSE of 59.97 after systematic hyperparameter tuning using cross-validation and random search. The study compared regularized regression methods, which help prevent overfitting by penalizing large coefficients, with combinations of techniques, which combine multiple weak learners to produce stronger predictions. XGBoost outperformed regularized re-

gression and alternative ensemble approaches, demonstrating that advanced boosting techniques can capture complex non-linear relationships between listing attributes and prices more effectively than simpler models when properly tuned [16].

Beyond price prediction, studies have looked into how Airbnb listings are distributed in urban areas and how host participation patterns change over time in response to sociol-economic and regulatory factors. A spatiotemporal analysis of Airbnb activity in San Francisco examined how listings emerged and spread across neighbourhoods between 2009 and 2015, revealing that host participation was strongly correlated with median household income, home-ownership rates, and proximity to tourist attractions, and that low-income neighbourhoods saw faster growth in commercial multi-listing hosts than wealthier areas where single-property hosts predominated. This study found that Airbnb's market structure reflects sociol-economic divides, with different types of hosts and listing strategies clustering in different areas of the city [5].

Efficient EDA frameworks may effectively manage relatively big datasets by converting dataframe operations into optimized database queries, as shown by research on exploratory data analysis utilizing database-backed dataframe libraries. By using PolyFrame to perform iterative cleaning, transformation, and aggregation without creating complicated distributed system infrastructure, data scientists were able to achieve performance on datasets with hundreds of thousands of rows that was either on par with or better than that of well-known dataframe libraries like Pandas and Dask, according to a case study that applied PolyFrame to Airbnb data. According to the study, database-backed approaches provide a useful compromise between single-machine tools and full distributed computing frameworks for moderately scaled data science tasks . This is because end-to-end exploratory workflows involving data loading, filtering, grouping, aggregation, and visualization preparation benefit from query optimization techniques traditionally used in database systems [1].

Effective application of machine learning models to real-world datasets requires careful data preparation, such as handling missing values, detecting and treating outliers, normalizing feature scales, and conducting exploratory data analysis to understand distributional properties and variables. Data cleaning research in quantitative studies

emphasizes the need to handle missing values systematically using methods such as deletion, mean or median calculation, or model-based imputation, depending on the missingness mechanism. Outliers should be identified using statistical criteria such as z-scores or median ranges, and then removed, transformed, or retained depending on whether they represent measurement errors or true extreme values. Data cleaning is an important requirement for trustworthy quantitative analysis, as the study points out that improper management of missing data and outliers can result in incorrect parameter estimations, decreased statistical power, and incorrect results [3].

Chapter 3

Methodology

This study uses a structured data mining workflow for the AB_NYC_2019 dataset, including data collection, cleaning, exploratory analysis, supervised classification, and unsupervised clustering. The methodology uses established quantitative data handling practices and machine learning algorithms to gain predictive and descriptive insights into Airbnb pricing patterns and market structure in New York City [1][3].

3.1 Data Collection

The AB_NYC_2019 dataset was obtained from Kaggle and contains 48,895 Airbnb listings across New York City's five boroughs. It includes 16 variables such as listing identifiers, host information, geographical coordinates, neighbourhood classifications, room type, price, minimum nights, review metrics, availability, and host listing counts. This dataset provides a comprehensive snapshot of New York City's short-term rental market, serving as the foundation for both supervised and unsupervised learning tasks [1].

3.2 Data Cleaning

Data cleaning was done in a systematic manner to address missing values and remove unimportant or low-quality records. Missing value analysis with Python's pandas library revealed that the name column had 16 missing entries, host_name had 21 missing entries, and last_review and reviews_per_month had 10,052 missing entries. Columns that do not contribute to analytical objectives, such as id, name, host_name, and last_review, were removed because they either serve as identifiers or contain a large amount of missing data, compromising analysis quality. Missing values in the reviews_per_month variable were imputed with zero, assuming that listings with no

recorded reviews have no review activity per month. Duplicate records and invalid entries were identified and removed to ensure data accuracy. After cleaning, the final dataset had 48,885 rows.Duplicate records and invalid entries were identified and removed to ensure data accuracy. After cleaning, the final dataset had 48,885 rows and 12 columns, with a loss of only 10 records from the original dataset[3].

3.3 Exploratory Data Analysis

Exploratory data analysis was used to better understand the distribution of key variables, identify relationships between features, and visualize patterns in the dataset. Univariate analysis looked at single variables to understand their distributions. A bar chart was used to visualize the distribution of room types, revealing that private rooms and entire home or apartment listings are the most common, with shared rooms accounting for the smallest segment of the market. A histogram was used to examine the distribution of listing prices, and it revealed that the majority of listings are priced below $200 per night, indicating that most travelers can afford it. A bar chart was used to analyze the distribution of listings by neighbourhood group, revealing that Manhattan and Brooklyn have the highest concentration of listings, indicating their popularity with tourists[1].

Multivariate analysis looked at the relationships between multiple variables to see what factors impact price and market structure. Box plots were used to examine price distribution by room type, revealing that entire home or apartment listings command the highest average prices, while private rooms and shared rooms are less expensive options, confirming that room type is a major determinant of listing price. Price variation across neighbourhood groups was examined using box plots, which revealed that Manhattan and Brooklyn have significantly higher median prices than Queens, the Bronx, and Staten Island, reflecting differences in demand and tourism concentration across boroughs. A scatter plot examining the relationship between price and number of reviews revealed that listings with more reviews tend to have lower prices, meaning that low-cost options receive more bookings, whereas high-cost listings accumulate

fewer reviews over time. The correlation analysis between numerical variables was performed and visualised using a heatmap, revealing that minimum nights required has the strongest relationship with pricing among continuous variables, while availability shows the lowest correlation with other features [1][11].

3.4 Supervised Classification

The continuous price variable was transformed into independent price categories for the supervised learning component, which framed the problem as a classification task. The prepared dataset was imported into Weka, a popular machine learning software platform that includes implementations of various classification algorithms as well as tools for model evaluation and comparison. Weka used multiple classification algorithms to predict price categories based on listing attributes like neighbourhood group, room type, geographical coordinates, minimum nights, review metrics, host listing counts, and availability. The tested classifiers include decision tree and Logistic regression, allowing for systematic comparison of algorithm performance on this dataset [7][10].

3.5 Unsupervised Classification

For the unsupervised learning component, K-Means clustering was used in Weka to find natural groupings within the dataset without using price labels as supervisory information. Clustering features include neighbourhood group, room type, geographical coordinates, minimum nights, review metrics, host listing counts, and availability, which enable the algorithm to identify market segments based on the intrinsic similarity of listings across these attributes. Prior to clustering, numerical features were normalized to ensure that variables measured on different scales contribute equally to the Euclidean distance calculations used by K-Means to assign clusters. The optimal number of clusters was determined by examining clustering quality metrics such as within-cluster sum of squares across different k values and using the elbow rule to identify the point at which additional clusters provide diminishing improvements in compactness. Weka's K-Means implementation was optimized with proper initialization and iteration settings to enhance convergence stability [9].

3.6 Result Analysis

After clusters were identified, they were characterized by examining the distribution of features within each cluster and comparing these distributions across clusters to determine whether natural market segments correspond to price levels, geographical areas, property types, or combinations of these factors. The performance of several categorization algorithms was carefully evaluated using accuracy, precision, recall, and F1-score criteria. The results were analyzed to identify which features are more predictive of price categories, which algorithms achieve the best balance of overall performance and minority class recognition, and what patterns develop from unsupervised clustering that shed light on the structure of the New York City short-term rental market [11][12].

Chapter 4

Results

4.1 Classification Performance Comparison

Two supervised classification algorithms, Random Forest and Logistic Regression, were used to predict Airbnb listing price categories on the AB_NYC_2019 dataset, with a 70/30 train-test split. The price categories included low, medium, high, and premium price bands, reflecting different market segments within New York City's short-term rental marketplace [7][10].

As shown in table 4.1, Random Forest achieved an overall classification accuracy of 68.99% on the test set of 14,668 instances, correctly classifying 10,120 of them. The Kappa statistic of 0.5053 indicates moderate agreement between predicted and actual categories that exceeds chance. The weighted average performance metrics show a precision of 0.685, recall of 0.690, and F-measure of 0.682, with a ROC area of 0.859 indicating strong discriminative ability across classes. When performance is examined by price category, Random Forest scored best on low-priced items (recall of 0.840 and F-measure of 0.827), followed by medium-priced listings (recall of 0.662 and F-measure of 0.620). The performance on high-priced listings was moderate with recall of 0.432 and F-measure of 0.492, while premium-priced listings proved most challenging with recall of only 0.094 and F-measure of 0.158, indicating the difficulty of predicting the smallest and most exclusive market segment [7][8].

On the same test set, Logistic Regression correctly classified 9,500 listings, resulting in an overall classification accuracy of 64.75%. The Kappa statistic of 0.4261 indicates moderate agreement, albeit slightly lower than Random Forest. The weighted average performance metrics have a precision of 0.644, recall of 0.648, and F-measure of 0.644, with a ROC area of 0.836. Logistic Regression performed best on low-priced listings

(recall = 0.818, F-measure = 0.800), followed by medium-priced listings (recall = 0.608, F-measure = 0.583). High-priced listings performed poorly, with a recall of 0.355 and an F-measure of 0.427, while premium listings proved the most difficult, with a recall of 0.065 and an F-measure of 0.119. Random Forest outperformed Logistic Regression in all price categories and accuracy measures, with considerable increases in high and premium category recall. However, both models struggled with the imbalanced premium class [8][10].

Table 4.1: Classification Performance Comparison: Random Forest vs Logistic Regression

Metric	**Random Forest**	**Logistic Regression**
Accuracy	68.99	64.75
Kappa Statistic	0.5053	0.4261
Precision	0.685	0.644
Recall	0.690	0.648
F-Measure	0.682	0.644
ROC Area	0.859	0.836
Correctly Classified	10,120	9,500
Incorrectly Classified	4,548	5,168
Total Instances	14,668	14,668

As shown in the Table 4.2, that both classifiers make similar errors, frequently identifying medium and high-priced listings as low-priced and failing to correctly identify premium listings due to their low representation in the dataset. According to Random Forest's confusion matrix, 3,366 of 5,087 medium-priced listings were correctly classified, while 1,094 were misclassified as low price. For high-priced listings, 1,126 of 2,607 were correct, with 1,310 misclassified as medium and 147 as low. Only 29 of 308 premium listings were correctly classified, with the vast majority misclassified as high or medium. These patterns confirm that class imbalance, especially the scarcity of premium listings, poses a significant challenge for both algorithms, consistent with research on imbalanced classification problems [8].

Table 4.2: Confusion Matrix

Actual / Predicted	**Medium**	**High**	**Low**	**Premium**	**Total**
Medium	3,366	622	1,094	5	5,087
High	1,310	1,126	147	24	2,607
Low	1,000	65	5,599	2	6,666
Premium	94	157	28	29	308

4.2 K-Means Clustering Analysis

K-Means clustering with k=3 was used to identify natural market segments in the dataset without using price labels as supervisory information. The algorithm converged after 8 iterations, with a within-cluster sum of squared errors of 72,006.66, indicating that the three identified clusters were reasonably compact. The clustering distributed listings into three segments: Cluster 0 with 23,888 listings (49 %), Cluster 1 with 6,812 listings (14%), and Cluster 2 with 18,195 listings (37%) [9].

Table 4.3: K-Means Clustering Results:

Characteristic	**Cluster 0**	**Cluster 1**	**Cluster 2**
Size (listings)	23,888 (49%)	6,812 (14%)	18,195 (37%)
Location	Brooklyn	Queens	Manhattan
Neighbourhood	Williamsburg	Astoria	Harlem
Room Type	Private room	Entire home/apt	Entire home/apt
Min. Nights	5.3	6.6	9.4
Reviews	20.3	38.0	21.7
Reviews/Month	0.98	1.68	1.02
Host Listings	2.2	3.7	15.0
Availability	67.7	216.9	133.0

Cluster centroids analysis reveals distinct market segments based on geographical location, room type, and host behaviour. Cluster 0, the largest segment, is located in Brooklyn's Williamsburg neighbourhood and is primarily made up of private rooms with relatively short minimum stay requirements (5.3 nights on average), moderate review counts (20.3 reviews), low host listing counts (2.2 listings per host), and limited availability (67.7 days per year). This profile suggests individual hosts offering spare rooms in their primary residences with selective availability, reflecting the classic peer-to-peer sharing model that Airbnb was originally designed to facilitate [5].

Cluster 1, the smallest segment, is located in Queens' Astoria neighbourhood and is primarily made up of entire home or apartment rentals with moderate minimum nights (6.6), the highest review activity (38.0 reviews and 1.68 reviews per month), moderate host listing counts (3.7) and the highest availability (216.9 days). This cluster represents mid-market entire-unit rentals operated by hosts who maintain a few properties with high availability and strong review performance, indicating professional or semi-professional hosts who prioritize guest satisfaction and consistent occupancy [12].

Cluster 2, the second-largest segment, is centered on Manhattan's Harlem neighbourhood and consists of entire home or apartment rentals with longer minimum stay requirements (9.4 nights), moderate review counts (21.7), the highest host listing counts (15.0) and moderate availability (133.0 days). This profile suggests commercial multi-listing hosts who manage portfolios of properties in Manhattan with longer stay requirements, potentially attracting to extended business travelers or tourists looking for weekly rentals rather than nightly bookings. The high number of host listings indicates that this cluster captures the commercial short-term rental sector, which has generated regulatory concern in many cities [5][13].

Comparing the clustering results to the classification problem demonstrates that K-Means' natural market groups do not precisely correlate with pricing categories, but rather reflect a combination of geographical location, property type, host strategy, and operational characteristics. This research shows that Airbnb market structure is complex, with pricing being one factor among others that define competitive positioning and market segmentation [12].

Chapter 5

Discussion

5.1 Classification Performance Comparison

Random Forest outperformed Logistic Regression in all metrics, with 68.99% accuracy compared to 64.75%. This performance advantage is due to Random Forest's ability to capture non-linear relationships via its ensemble of decision trees, whereas Logistic Regression assumes linear separation between classes. Previous studies have found that ensemble methods achieve superior accuracy on complex classification tasks, particularly when feature interactions influence outcomes [10].

The moderate Kappa statistics for both models (0.51 for Random Forest, 0.43 for Logistic Regression) show a significant improvement over random classification, but there is still room for improvement. Both algorithms struggled in the premium price category, with recall rates below 0.10 despite acceptable performance in other classes. This pattern is typical of imbalanced classification problems, where standard algorithms prioritize overall accuracy by insisting on majority classes at the expense of minority class recognition [8].

The confusion matrix reveal calculated misclassification patterns, with medium and high-priced listings frequently misidentified as lower-priced categories. This suggests that there is significant overlap in the feature space, with listings in adjacent price bands sharing similar observable characteristics. The overlap is most likely explained by the dataset's limited feature set, which lacks attributes such as interior quality, specific amenities, and host reputation that distinguish similarly located properties. According to a study, structured listing metadata alone cannot accurately predict price categories [11].

5.2 Market Segmentation

K-Means clustering revealed three distinct market segments that reflect various operational models. Cluster 0, which contains 49% of listings, represents traditional home-sharing with private rooms in Brooklyn, operated by hosts with small portfolios and selective availability, in line with the peer-to-peer vision of residents monetizing spare capacity in their primary residences. Cluster 1, which accounts for 14%, consists of entire Queens apartments with high review frequency and year-round availability, indicating semi-professional hosts managing small portfolios with an emphasis on occupancy and guest satisfaction. Cluster 2 accounts for 37% of the commercial sector in Manhattan, where hosts operate portfolios averaging 15 properties with longer minimum stays, indicating professional management aimed at extended-stay guests. This pattern reflects concerns about large-scale commercial operations and reduced long-term rental availability [5][12][13].

These segments demonstrate that market structure is multidimensional, including location, property type, host strategy, availability patterns, and pricing. For hosts, competitive positioning requires alignment across multiple operational dimensions rather than price optimization alone. For policymakers, the distinct typologies suggest that undifferentiated regulatory approaches may treat dissimilar operations equally [12].

5.3 Determinants of Price Classification

Exploratory analysis confirmed that room type and neighbourhood location are the primary price determinants, in line with previous unconventional pricing research. Entire apartments are significantly more expensive than private rooms, and Manhattan and Brooklyn have higher prices than the outer boroughs. The inverse relationship between review counts and prices implies that affordability drives booking frequency, creating a feedback mechanism where lower-priced listings accumulate social proof more quickly [4][15].

However, moderate classification accuracy indicates that observable features account for only a portion of the variance in price categories. Unobserved factors such as

amenities, interior condition, precise location relative to transportation and attractions, and host responsiveness likely contribute significantly to price differentiation, suggesting that incorporating richer feature sets could meaningfully improve predictive performance [11][16].

5.4 Limitations and research extensions

Several limitations limit these findings. The dataset is a static snapshot from 2019, so it cannot be used to analyze temporal dynamics such as seasonal pricing variation or host strategy evolution. Longitudinal analysis of individual listings would reveal adaptive pricing behaviours and market responses to external shocks [13].

Severe class imbalance, particularly in the premium category, significantly reduced classifier performance in rare classes. Addressing this issue with resampling techniques, cost-sensitive learning, or specialized algorithms may improve minority class recognition. Research shows that modified Random Forest implementations improve minority class recall while maintaining overall accuracy [8].

The analysis excluded unstructured data from listing descriptions and photographs. Natural language processing and computer vision analysis can extract features related to property quality and style that numerical attributes do not capture, but at the expense of model interpret-ability [16].

Finally, the elbow method was used to determine the clustering solution for k=3, but other cluster numbers or algorithms may reveal additional market structure. Hierarchical clustering or density-based methods may reveal finer-grained segmentation patterns [9][12].

Chapter 6

Conclusion and Future Scope

6.1 Conclusion

This study used data mining techniques on the AB_NYC_2019 dataset to classify Airbnb listing price categories and identify market segments using clustering. Random Forest outperformed Logistic Regression with 68.99% accuracy versus 64.75%, demonstrating that ensemble methods capture pricing relationships better than linear models. However, both algorithms struggled with the premium price category due to class imbalance, with recall below 0.10 for the rarest listings [7][10].

K-Means clustering revealed three distinct market segments that reflect various operational models. The largest segment consists of private rooms in Brooklyn operated by individual hosts, which represent traditional home-sharing. The smallest segment includes entire Queens apartments with high availability, indicating semi-professional operations.The third segment includes commercial multi-listing hosts in Manhattan that manage large portfolios, indicating professional property management. These findings show that the Airbnb market is heterogeneous, with various host types pursuing distinct strategies [5][12].

The analysis confirms that room type and neighbourhood location are the most important price determinants, with entire apartments commanding higher prices than private rooms and Manhattan and Brooklyn locations associated with premium pricing. The combination of supervised classification and unsupervised clustering provides complementary insights; classification reveals which attributes predict price categories, while clustering reveals operational strategies defining different host types[4][11][15].

6.2 Future Scope

Future research should use longitudinal analysis to investigate seasonal pricing variations and how hosts strategies change over time. Using techniques designed for imbalanced classification, like hybrid resampling or cost-sensitive learning, can significantly improve minority class recognition for premium listings [8][13].

Integrating unstructured data from listing descriptions and photographs using natural language processing and computer vision could reveal features about property quality and amenities that structured attributes do not. This would likely improve predictive performance by incorporating missing information from the dataset [16].

Alternative clustering approaches, such as hierarchical clustering or density-based methods, could reveal more detailed market segmentation patterns. Examining the relationship between identified market segments and pricing outcomes would clarify whether operational clusters correspond to distinct price ranges or if other factors drive price differentiation within segments [9][12].

References

[1] PhanwadeeSinthongand M. P. Carey, "Exploratory Data Analysis with Database-backed Dataframes: A Case Study on Airbnb Data," Dec. 2021, doi: https://doi.org/10.1109/bigdata52589.2021.9671603.

[2] V. D. Poianet al., "Exploratory data analysis (EDA) machine learning approaches for ocean world analog mass spectrometry," vol. 10, May 2023, doi: https://doi.org/10.3389/fspas.2023.1134141.

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

[4] V. Toader, A. L. Negruşa, O. R. Bode, and R. V. Rus, "Analysis of price determinants in the case of Airbnb listings,"Economic Research-EkonomskaIstraživanja, vol. 35, no. 1, pp. 1–17, Aug. 2021, doi: https://doi.org/10.1080/1331677x.2021.1962380.

[5] A. Sarkar, J. B. Pick, and S. Jabeen, "Spatiotemporal Patterns and Socioeconomic Influences on Host Participation in Short-Term Rental Markets: Airbnb in San Francisco,"ISPRS international journal of geo-information, vol. 13, no. 1, pp. 4–4, Dec. 2023, doi: https://doi.org/10.3390/ijgi13010004.

[6] S. L. Franconeri, L. M. Padilla, P. Shah, J. M. Zacks, and J. Hullman, "The science of visual data communication: What works," Psychological Science in the Public Interest, vol. 22, no. 3, pp. 110–161, Dec. 2021, doi: https://doi.org/10.1177/15291006211051956.

[7] Divya Pramasani Mohandoss, Y. Shi, and K. Suo, "Outlier Prediction Using Random Forest Classifier," Jan. 2021, doi: https://doi.org/10.1109/ccwc51732.2021.9376077.

[8] W. Lin, J. Gao, B. Wang, and Q. Hong, "An Improved Random Forest Classifier for Imbalanced Learning," 2021 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), pp. 703–707, Jun. 2021, doi: https://doi.org/10.1109/ICAICA52286.2021.9497933.

[9] H. Zhao, "Design and Implementation of an Improved K-Means Clustering Algorithm," Mobile Information Systems, vol. 2022, pp. 1–10, Sep. 2022, doi: https://doi.org/10.1155/2022/6041484.

[10] X.-X. Dong et al., "Comparison of Logistic Regression and Machine Learning Approaches in Predicting Depressive Symptoms: A National-Based Study," Psychiatry Investigation, vol. 22, no. 3, pp. 267–278, Mar. 2025, doi: https://doi.org/10.30773/pi.2024.0156.

[11] J. Dhillon et al., "Analysis of Airbnb Prices using Machine Learning Techniques," IEEE Xplore, Jan. 01, 2021. https://ieeexplore.ieee.org/document/9376144

[12] Kadri, "Machine Learning-based Categorization of Airbnb Listings in NYC," Journal of Soft Computing Paradigm, vol. 6, no. 3, pp. 299–313, Sep. 2024, Available: https://irojournals.com/jscp/article/view/6/3/6

[13] M. Hu, L. Yang, J. Park, and M. Liu, "Survival determinants and prediction for Airbnb listings," International Journal of Hospitality Management, vol. 128, p. 104132, Jul. 2025, doi: https://doi.org/10.1016/j.ijhm.2025.104132.

[14] Y. Gu, "Predictive Model on Airbnb Price in New York City." Available: https://www.clausiuspress.com/conferences/LNEMSS/ICEMGD%202020/442.pdf

[15] M. Mahyoub, A. A. Ataby, Y. Upadhyay, and J. Mustafina, "AIRBNB Price Prediction Using Machine Learning," IEEE Xplore, Jan. 01, 2023. https://ieeexplore.ieee.org/abstract/document/10099909/

[16] Y. Liu, "Airbnb Pricing Based on Statistical Machine Learning Models," 2021 International Conference on Signal Processing and Machine Learning (CONF-SPML), Nov. 2021, doi: https://doi.org/10.1109/conf-spml54095.2021.00042.