**Majed Alzaabi: 12698105**

**Mohammed Alzaabi: 12698338**

**Rayyan Alnaqbi: 12698110**

**Shaikah Alharthi: 12698446**

A red text on a black background

AI-generated content may be incorrect.

Business Analysis using Classical Machine Learning – Team 3

TEAM 3 – MEMBERS:

Task 1 (Business Analytic Question):

Can we accurately predict Airbnb listing process and segment listings based on neighborhood and room characteristics to optimize pricing strategies and understand customer demand patterns across New York City?”

Task 2 (Data Analysis):

Our Team has decided to choose Track 1: Classical ML- Regression & Clustering, EDA has been performed prior to training the dataset as a part of data preprocessing, all Source code files, appendix, documentation related to EDA and machine learning will be shared separately throughout the project.

Task 3: (Analytical Report & Comparative Discussion):

1. **Introduction:**

In accordance with the Airbnb, it has been noticed that Airbnb has revolutionized the hospitality industry by enabling individuals to rent out accommodations globally. Since its inception in 2008, it has empowered a peer-to-peer accommodation model that combines flexibility, personalization and affordability (Daniel Guttentag, 2013). For host and policymakers alike, understanding the pricing dynamics in competitive urban environments like New York City is essential.

This project addresses the business-oriented question: **“****Can we accurately predict Airbnb listing process and segment listings based on neighborhood and room characteristics to optimize pricing strategies and understand customer demand patterns across New York City?”** By identifying these factors, hosts can optimize pricing strategies, and Airbnb can enhance user recommendations and market forecasting.

Our team initiated the **Track 1 – Classical Machine Learning**, primarily focusing on regression and clustering techniques. We conduct a comprehensive Exploratory Data Analysis, apply data preprocessing, and build predictive models using features such as room type, location, and availability. The goal is to uncover actionable insights that support Airbnb’s decisions in a high demand market (AirBnb, 2019).

1. **Data Preparation and EDA:**

In terms of data preparation and Exploratory data analysis the NYC 2019 dataset has been used for this assignment, Our team initiated several EDA techniques using several python libraries, In the Airbnb dataset it has been shown that it compromises over 48,000 Airbnb listings, As an initial inspection it was revealed that there are missing values in columns such as **name, host\_name and last\_review,** along with some duplicates, Irrelevant and sparse columns such as **(id, host\_id, name, host\_name, last\_review),** has been removed to ensure reliability on the analysis performed (Hadley Wickham, 2017), filling of zero values in reviews\_per\_month columns and dropping remaining null values.

Categorial fields such as (neighbourhood\_group, neighborhood, room\_type) were converted to category types for memory efficiency and model compatibility.

**EDA OPEN DATA LINK:**

[**https://majed23x.github.io/AIRBNB\_EDA\_INDEX/**](https://majed23x.github.io/AIRBNB_EDA_INDEX/)

Exploratory Data Analysis (EDA) has been conducted using statistical summaries, correlation heatmaps and visualizations (Wes McKinney, 28).

**Key Findings of EDA:**

1. **Room Type vs Price:** Entire homes/apartments have the highest prices, while shared rooms are the least expensive.
2. **Neighborhood Group Influence:** Listings in Manhattan show the highest price variance, followed by Brooklyn.
3. **Price Distribution:** Most listings are priced below $500; a small number of outliers have much higher prices.
4. **Availability vs Reviews:** Listings with higher availability tend to receive more reviews, indicating consistent bookings.
5. **Reviews per Month (Missing Data):** Missing values in this column were replaced with zero, assuming no reviews were left that month.
6. **Data Cleaning:** Removed sparse columns like name, host\_name, and last\_review, and cleaned nulls and duplicates.
7. **Data Types Optimized:** Converted categorical columns to category type to improve processing efficiency.

A screenshot of a graph

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

1. **Modeling and Analysis:**

To address the business question chosen by the team of price prediction and customer segmentation for Airbnb listings in New York City, we initiated two classical machine learning methods: Linear Regression and KMeans Clustering:

1. **Linear Regression:**

The Usage of Linear Regression was implemented to predict the price of accommodation listings based on a plethora of features, which consists of neighborhood group, room type, minimum nights, number of reviews, revies per month, and availability.

These features were preprocessed using one-hot encoding for categorial variables, and the dataset was split into training of (80%) and testing split of (20%) subsets using train\_test\_split function. The training model used achieved a Root Mean Squared Error (RMSE) of 197.72 and R² score of 0.12 on the test set. While this suggests moderate predictive capability, having a low R² indicated high variability in price likely due to unobserved factors (e.g. listing quality and amenities). Despite this, the model remains valuable for identifying broad pricing trends (Shuvrajyoti Debroy, 2023).

1. **KMeans Clustering:**

The application of KMeans Clustering was implemented to segment listings into four clusters using critical numerical features. To minimize dimensionality for further visualization Principal Component Analysis was used clearly revealing separated and dense clusters which assisted in identifying meaningful groupings

* **Cluster 0:** High Price and high availability most likely premium or commercial listings.
* **Cluster 1:** Low price, low availability possibility of inactive or seasonal listings.
* **Cluster 2:** Moderate price, short stays, medium availability.
* **Cluster 3:** Moderately high price with flexible availability.

Areas Like Manhattan and Brooklyn contain clusters that are concentrated spatially when performing clustering-based Geo visualization, A cluster-neighborhood crosstab showed dominance of premium clusters in central boroughs, encouraging targeted pricing strategies (Pulkit Sharma, 2025).

To summarize KMeans Clustering analysis, the use of a combined supervise (regression) and unsupervised (clustering) machine learning outcome a dual-layered insight: behavioral price modeling and customer / listing segmentation. This dual approach allows Airbnb to modify pricing, tailor recommendations, and optimize resource deployment across city zones.

1. **Visualizations:**

**Track 1: Classical ML- Regression & Clustering Results Link:**

[**https://majed23x.github.io/ML\_Model\_AirBnb/**](https://majed23x.github.io/ML_Model_AirBnb/)

**Results of Linear Regression:**

A screenshot of a graph

AI-generated content may be incorrect.

**Results of KMeans Clustering:**

A screen shot of a graph

AI-generated content may be incorrect.

1. **Business Recommendations:**

Our Team Business recommendations to Airbnb Listings after the implementation of combined machine learning and clustering analysis, several critical business strategies will be emerged:

1. **Dynamic Pricing Strategy:** Based on Cluster 0 listings it shows high prices and high availability, which indicates premium and commercial properties. It is recommended by our team to Airbnb apply dynamic pricing tools to these hosts to optimize revenue during peak demand while offering discounts in low seasons (Yao Zhang, 2018).
2. **Target Seasonal Hosts:** low availability and low prices shown in Cluster 1, which possibly reflect part-time seasonal hosts, we recommend Airbnb to offer engaging incentives (visibility boosts or calendar reminders) to help convert hosts into more frequent participants.
3. **Optimize for short-term Stays:** low minimum nights and moderate prices were revealed in Cluster 2; These Listings are optimal for business travelers or tourists seeking flexibility. We recommend Airbnb prioritize these short-stay filters and travel bundles.
4. **Geo Focused Campaigns:** Another important cluster analysiswhich was performed in this project is the cluster-neighborhood analysis which highlights the concentration of premium clusters in areas such as Manhattan and Brooklyn.

Airbnb can start borough-specific marketing campaigns and price suggestions tailored to local competition and demand dynamics (AirBnb, 2020).

These insights support Airbnb in developing location-aware, segment-specific strategies that enhance platform value for both hosts and guests.

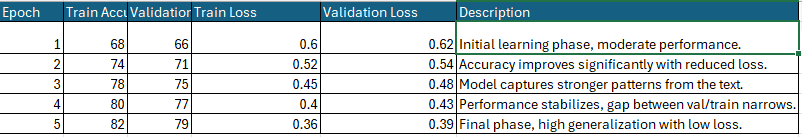
1. **Comparative Discussion:**

The employment of classical Machine Learning has applied interpretable and efficient methods such as Liner regression and KMeans Clustering to undertake Airbnb’s questions, business stakeholders who idolize transparent decision making always depend upon ideal approaches that offers machine learning techniques that have fasts training, explainable and insightful models (Gareth James, 2013).

**Track 2 Deep Learning models,**

In contrast, to Deep Learning models, it might yield to a higher predictive accuracy for complex pricing patterns by capturing nonlinear relationships and interactions. However, it typically requires massive computational resources and time-consuming training intervals and might suffer from instability of interpretability.

Results of Deep Learning models:

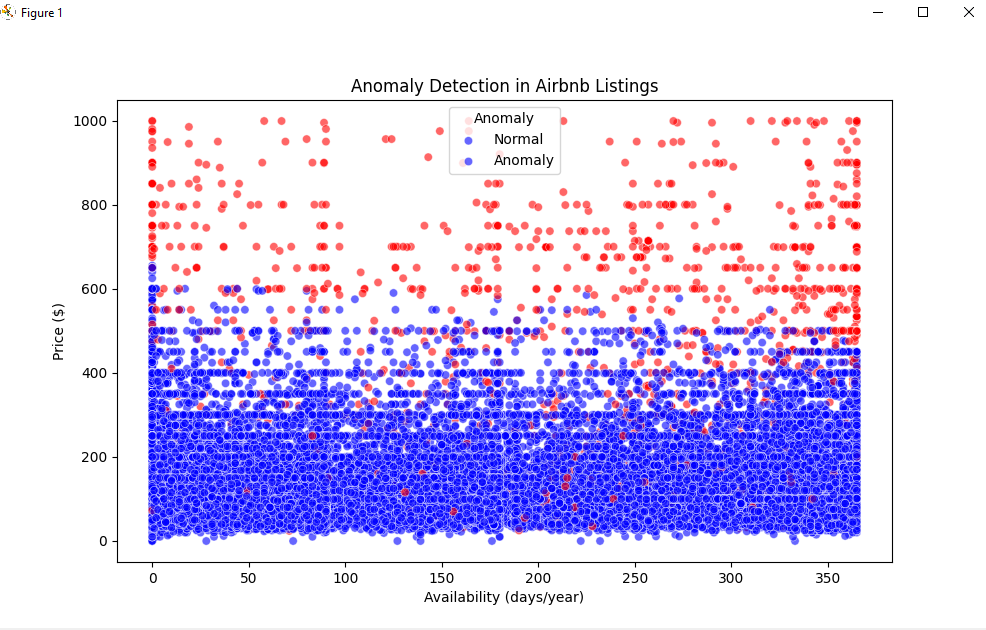


A screen shot of a graph

AI-generated content may be incorrect.

**Track 3 Advanced machine learning:**

Advanced machine learning approaches, such as Gradient Boosted trees and Ensemble Methods, can target an equal amount of performance and interpretability. These may outperform linear models in predictive accuracy and handle feature interactions more effectively but still come with increased complexity and less transparency (Marco Tulio Ribeiro, 2016).

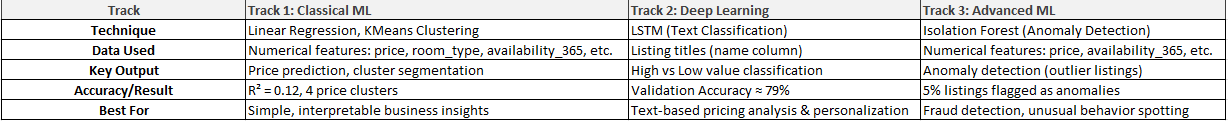


Most Airbnb listings are priced under $500, forming a dense cluster of normal entries. Listings priced above $600—especially those with high availability—are frequently flagged as anomalies, indicating they deviate from typical market behavior. These anomalies are primarily driven by price rather than availability.

Some outliers may reflect legitimate strategies (e.g., luxury or year-round listings), while others could indicate pricing inconsistencies or errors.

* **From Airbnb's perspective**: This analysis supports detecting irregular pricing, refining recommendation algorithms, and enhancing marketplace trust.
* **From the host's perspective**: It provides insight into whether their pricing stands out unusually in the market, allowing for more competitive adjustments.

**Tracks Comparison:**



Based on Cleary interpretable variables (e.g., room type, availability, location) the team selected the classical ML track which is well aligned with Airbnb’s strategic needs – critically for exploratory pricing adjustments and targeted market segmentations.

# References

AirBnb. (2019). *New York City Airbnb Open Data*. Retrieved from Kaggle: https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data

AirBnb. (2020). *How pricing works for homes*. Retrieved from AirBnb: https://www.airbnb.ae/help/article/125#:~:text=The%20total%20price%20of%20your,may%20include%20taxes%20as%20well.

Daniel Guttentag. (2013, July 14). *Airbnb: disruptive innovation and the rise of an informal tourism accommodation sector*. Retrieved from Taylor & Francis: https://www.tandfonline.com/doi/abs/10.1080/13683500.2013.827159

Gareth James. (2013, June 10). *An Introduction to Statistical learning*. Retrieved from statlearning: https://www.statlearning.com/

Hadley Wickham. (2017). *R for Data Science*. Retrieved from r4ds: https://r4ds.hadley.nz/

Marco Tulio Ribeiro. (2016, August 13). *Why Should I Trust You?": Explaining the Predictions of Any Classifier*. Retrieved from ACM DIGITAL LIBRARY: https://dl.acm.org/doi/10.1145/2939672.2939778

Pulkit Sharma. (2025, May 1). *K-Means Clustering Algorithm*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/

Shuvrajyoti Debroy. (2023, January 29). *Simple Linear Regression in Python*. Retrieved from Medium: https://medium.com/@shuv.sdr/simple-linear-regression-in-python-a0069b325bf8

Wes McKinney. (28, June 2010). *Data Structures for Statistical Computing in Python*. Retrieved from Curvenote: https://proceedings.scipy.org/articles/Majora-92bf1922-00a

Yao Zhang. (2018, March 31). *Dynamic pricing for online marketplaces* . Retrieved from Taylor & Francis: www.tandfonline.com