NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

CB0494 Introduction to Data Science and Artificial Intelligence

Semester 2 (2024/25 Academic Year)

Project Report: Milan Airbnb Prices Prediction

3 April 2025

Group Member	Name (as stated in the matriculation card)	Matriculation No.
#1	Dalia Betinjaneh	N2402643E
#2	Emma Anastassova	N2402084D

1. BACKGROUND AND OBJECTIVES

Airbnb prices in cities like Milan vary widely based on location, room type, host behavior, and reviews. Understanding and predicting these prices using data science techniques improves customer experience and supports better pricing strategies for hosts.

The objective of this project is to analyze and model the pricing behavior of Airbnb listings in Milan using a dataset with numerical and categorical attributes. We aim to 1. Perform thorough exploratory data analysis (EDA) to understand the dataset, detect anomalies, and extract insight. 2.Use clustering techniques to group similar Airbnb listings based on relevant attributes. 3. Implement supervised learning models (Linear Regression and Classification Trees) to predict price or price categories based on selected attributes.

2. METHODOLOGY

2.1 Data Preparation

After loading the dataset, we performed an initial inspection and identified 14 numerical and 5 categorical variables. The categorical variables include **name**, **host_name**, **neighbourhood**, **room_type**, and **last_review**. Since **last_review** was initially stored as a string, we transformed it into a numerical feature called **days_since_last_review** by computing the number of days between the last review date and the current date. This new variable helps quantify the recency of guest activity.

Summary statistics revealed that both **price** and **minimum_nights** are heavily right-skewed, with a small number of extreme outliers. Variables such as **reviews_per_month** and **availability_365** also showed high variability, which likely reflects differences in listing popularity and host behavior.

2.2 Handling Missing Values

name and **host_name** contain a small number of missing values (10 and 124 respectively), but since they are not used in modeling, no action was required. However, **last_review** and **reviews_per_month** have around 5,000 missing values - most likely for listings with no reviews. We explored two strategies: imputing missing values vs. dropping those rows. The final decision was made after further EDA and model evaluation.

3. EXPLORATORY DATA ANALYSIS (EDA)

We apply EDA to understand variable distributions, detect anomalies, and identify patterns. We analyzed both individual variables and joint relationships using univariate, bivariate, and multivariate visualizations.

3.1 Univariate Analysis

We examined the distribution of all relevant numerical variables using histograms, boxplots, and violin plots. We also explored descriptive statistics (mean, median, IQR, standard deviation) and visualized value ranges to detect skewness and variability. Key findings:

- **price**, **minimum_nights**, and **review-related features** (number of reviews, reviews per month) are highly skewed with many outliers.
- **latitude** and **longitude** are relatively symmetric, listings are evenly distributed around central Milan.
- **days_since_last_review** is moderately skewed - many listings haven't been reviewed in a long time.
- **availability_365** shows a bimodal distribution, with many listings either almost always available (close to 365) or rarely available (close to 0).

- **host_listing_count** reveals that while most hosts manage a single listing (occasional hosts), a few operate many - likely professional hosts or property management companies.
- **minimum_nights** is low for most listings, but a few require long stays, indicating targeting of different rental strategies.

3.2. Anomaly detection : Outliers analysis

To better understand the price distribution, we analyzed extreme values, as very high-priced listings can skew the data and distort model training. These luxury listings likely follow different pricing dynamics and would ideally require additional features—like apartment size or amenities—which are not available in our dataset. To reduce their impact, we applied **anomaly detection** using the **IQR method**, flagging listings with price ≥ **€200** as outliers. This identified 1,595 listings (8.71%) with an average price of €500.64 and a maximum of €11,999.

We then analyzed the characteristics of these outliers:

- **85.2%** of outliers were **Entire home/apt**, and most were located in **central upscale neighborhoods** such as Duomo, Brera, and Buenos Aires – Venezia, popular with high-end tourists.
- Outliers had slightly higher average availability, but the difference was marginal - suggesting availability alone does not predict price well.
- Some hosts (e.g., **host_id** 175128252) manage many high-priced listings, pointing to professional hosts or companies. Such hosts might charge higher prices than those managing only 1–2 listings.

These findings confirm that outliers behave differently and are likely influenced by unobserved luxury-related features. Removing them avoids distortion in modeling, while **host_id** may still serve as a valuable predictor among standard listings.

3.3 Joint Distributions

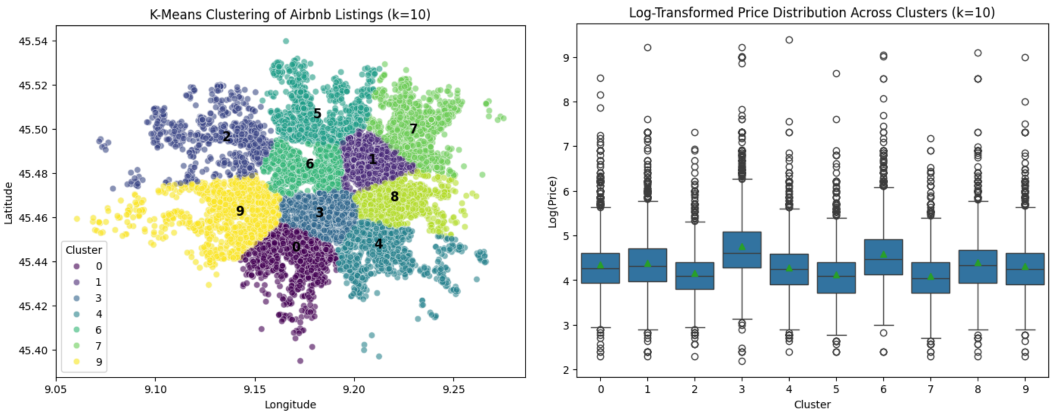
- **Scatter plots** showed no clear linear trends between numerical variables and price.
- **Boxplots** showed significant price differences across **room types** and **neighborhoods**. Upscale areas like **Duomo** and **Brera** had median prices around **€100**, while some outer neighborhoods were closer to **€50**. Room type and neighborhood clearly influence price, so we included them in our models. Initially, boxplots were compressed due to extreme values; we regenerated them on the outlier-filtered dataset (without log transform) to better capture actual price ranges across neighborhoods.
- The **correlation matrix** confirmed that most numeric variables have **weak correlation** with price ($r < 0.2$), indicating no single dominant predictor.

4. UNSUPERVISED LEARNING (CLUSTERING)

KMeans clustering was applied to **longitude** and **latitude** to identify geographic groupings:

- The **elbow method** suggested an optimal number of clusters at $k = 4$, but we also tested $k = 8$ and $k = 10$ to obtain more granular clusters - especially to separate central vs outer areas of Milan.
- In the $k = 10$ clustering, we used boxplots of log(price) to compare price distributions across clusters. Clusters 3 and 6, which correspond to the central zones of Milan, had higher median and mean prices and a greater share of outliers, confirming the strong impact of location.

- We then computed the correlation between each cluster assignment and price, and found that **cluster_8** had the strongest correlation, so it was selected as a feature in our predictive models.
- Clustering was also used to build separate models per cluster (region-based modeling), motivated by the idea that different areas of Milan might follow distinct pricing patterns. This approach allowed us to compare localized models with a global one.



5. FEATURE SELECTION

The goal of this step was to identify the most relevant and informative features for our supervised learning models. We combined insights from EDA, correlation analysis, and domain logic to guide this process.

5.1 Feature Engineering

We introduced three new binary features to capture hidden patterns not reflected in the raw variables:

- **is_professional_host**: Equals 1 if the host manages 10 or more listings, 0 otherwise. This captures differences in pricing strategies between professional and casual hosts.
- **is_high_availability**: Equals 1 if availability_365 ≥ 250, indicating full-time rental properties, as opposed to occasional listings.
- **is_long_min_stay**: Equals 1 if minimum_nights ≥ 7, which may correspond to listings targeting business travelers or long-term stays.

5.2 Variable Exclusion : we excluded variables with no predictive value, such as:

- **IDs and textual fields**: **id**, **name**, **host_name**, **last_review**
- **Target variable**: **price** (and its log version)

We retained **host_id** as some hosts consistently manage high-priced listings, as shown in the outlier analysis. We avoided including multiple location-based features (longitude, latitude, cluster_4, cluster_10) to reduce redundancy. We kept only **cluster_8**, which had the highest correlation with price and already captures geographic variation.

5.3 Final Feature List : selected based on correlation with price, interpretability, and EDA insights.

- **Numerical**: **minimum_nights**, **number_of_reviews**, **reviews_per_month**, **host_id**, **days_since_last_review**, **is_professional_host**, **is_long_min_stay**, **is_high_availability**, **cluster_8**
- **Categorical Encoded**: **room_type**, **neighbourhood**

6. DATA PREPROCESSING AND CLEANING

We prepared the dataset by handling missing values, encoding categorical features, and removing outliers. Missing values were either imputed (**reviews_per_month** = 0, **days_since_last_review** = max) or dropped entirely; both versions were kept for model comparison. Categorical variables (**room_type**, **neighbourhood**) were **ordinally encoded** based on comfort level and average price, respectively. Listings with price ≥ €200 were removed using the IQR method to reduce distortion from extreme values.

7. SUPERVISED LEARNING : MODELING RESULTS

7.1 Linear Regression

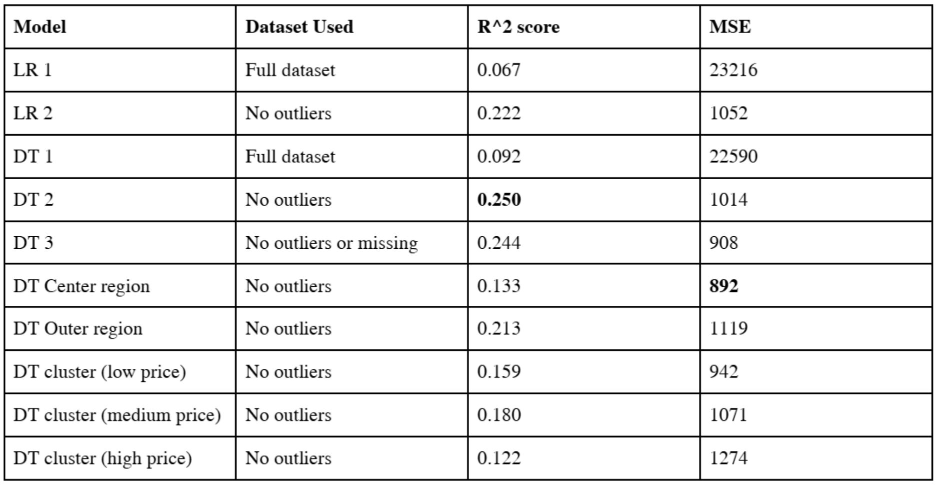
- **Model 1** : trained on the full dataset with outliers; poor performance due to extreme values.
- **Model 2** : trained on the dataset after removing outliers (using the IQR method); resulted in significant improvement in both R² and MSE.
- We also tried a separate linear regression for each cluster region, which we did not include, since it showed little explainability (R^2) and significantly worse MSE than the general model.

7.2 Decision Tree Regression

- **DT Model 1**: Trained on full dataset with outliers; poor performance, similar to Linear Regression 1.
- **DT Model 2**: Trained on the outlier-filtered dataset; outperformed both linear models.
- **DT Model 3**: Same as DT2, but rows with missing values were dropped instead of imputed; achieved slightly worse results than DT2, since less training data (dropped 5k rows).
- **Cluster-Based DT Models**: Trained separate models for low, medium, and high-priced cluster groups (based on average price in **cluster_10**). Aimed to capture different pricing patterns by price level, but results were mixed and did not outperform the global DT model.
- **Region-Based DT Models**: Trained separately on listings from central vs outer zones of Milan; outer model performed better (higher R²), but still worse than DT2.

7.3 Decision Tree Classification

- **Binary Classification Model**: Reframed the task as a binary classification (low vs high price, split at the median); trained on outlier-filtered dataset with missing values dropped. Achieved **70.3% accuracy**, outperforming all regression models (see figure below).



Model	Dataset Used	R^2 score	MSE
LR 1	Full dataset	0.067	23216
LR 2	No outliers	0.222	1052
DT 1	Full dataset	0.092	22590
DT 2	No outliers	**0.250**	1014
DT 3	No outliers or missing	0.244	908
DT Center region	No outliers	0.133	**892**
DT Outer region	No outliers	0.213	1119
DT cluster (low price)	No outliers	0.159	942
DT cluster (medium price)	No outliers	0.180	1071
DT cluster (high price)	No outliers	0.122	1274

8. CONCLUSION

We analyzed Airbnb prices in Milan using EDA, anomaly detection, clustering, and supervised learning (linear regression and decision trees). Our goal was to understand pricing patterns and build models to predict price or price categories.

Key Findings

- The dataset was noisy and skewed, with ~8.7% high-price outliers. Removing them improved model performance.
- Price prediction with regression was difficult due to noise, weak correlations, and non-linear patterns. Classification into two price groups (low vs high) was more feasible and reached 0.703 accuracy.
- Prices were mainly influenced by room type and location (neighbourhood and cluster) ; categorical variables had more predictive value than numerical ones; most other features had limited impact.
- Decision trees outperformed linear regression by capturing non-linear patterns in the data.
- Clustering captured spatial variation in prices; cluster-based models were tested but didn't outperform global ones.

Limitations :

- No features on information such as apartment size, year built, and amenities limited predictive power.
- Noise and high variability in the data made it hard to capture consistent pricing patterns.
- Weak linear correlations between variables and price reduced regression effectiveness.

Final Recommendation

- Use classification models for price segmentation rather than exact price prediction.
- Collect additional features (e.g., apartment size, year built, amenities) to improve model accuracy.
- Explore ensemble and non-linear models (e.g., Random Forest, XGBoost) for better performance.