**BUSAN 302**

**Group Project Report**

**Analysis of European AirBnB’s using   
Machine Learning**

**Dataset chosen: AirBnB**

**Fri 8-10 Group B**

| Name | UPI |
| --- | --- |
| Lance Nillo | lnil547 |
| Danny Weng | dwen293 |
| Nancy Wu | nwu928 |
| Daniel Chua | dchu804 |
| Winson Liang | wlia752 |

# Executive Summary

Airbnb has emerged as a global game-changer, redefining how people access accommodation and explore new destinations. This report analyzes European Airbnb listings, primarily focusing on constructing a predictive pricing model and uncovering the variables affecting price.

Our dataset consists of 41,714 records and 16 attributes. We identified 8 variables to focus on, however we took into account all 16 variables to ensure our analysis remained balanced and unbiased.

To predict pricing, we employed a Linear Regression model, implementing feature engineering techniques such as log-transformations. The final model achieved an R-squared value of 0.39 and pinpointed the following key explanatory variables with the most influence on Price:

* Person Capacity, with a coefficient of 1.04
* Bedrooms, with a coefficient of 1.21
* Normalized Attraction Index, with a coefficient of 1.51
* Entire home/apt, with a coefficient of 1.29

However, our analysis uncovered specific limitations:

* The dataset includes Airbnb listings from various European cities, making it challenging to predict prices universally across all cities
* Missing data resulted in the exclusion of approximately 8,000 records from model training, potentially valuable information
* Linear Regression assumes a linear relationship between input features and the continuous target variable, which may not always hold true

This report lays the foundation for understanding the determinants of Airbnb pricing in European cities. It empowers Airbnb hosts and property owners to optimize their pricing strategies and property offerings, enabling Airbnb to maintain its competitive edge and continue establishing itself as the premier platform for hospitality.

# Introduction

In today’s world of travel and accommodation, Airbnb has emerged as a prominent player, fundamentally altering the way individuals experience hospitality services. Airbnb is a globally renowned online marketplace that facilitates short-term lodging and travel experiences, connecting hosts with travelers to allow individuals to explore diverse destinations while also enjoying the comforts of a home that is away from home.

This report addresses the question: "How can we effectively assess the appropriate rental rates for AirBnB listings in European cities while considering the significance of various factors influencing pricing?”

We hypothesise that City Center, Bedrooms, and Attraction Index are significant determinants to European Airbnb pricing. This hypothesis is supported by our proposal's recommendation to employ a linear regression machine learning model as our primary predictive tool. This report leverages the Airbnb dataset with data analytic techniques to provide insights into the factors which determine pricing for Airbnb listings.

# Descriptive statistics

The dataset is an extraction of the Airbnbs in a few cities in Europe. It comprises 41,714 rows and 16 columns; each row details one record of an Airbnb and the columns, or attributes, are made up of nominal, ordinal, continuous, and discrete data which provide information about the Airbnb.



Its attributes can be grouped into 6 categories:

* **Price:** Comprises one floating point attribute, Price (continuous), that denotes the price of the AirBnB.
* **Features:** Variables that describe the type of AirBnB. This comprises five attributes including Person Capacity (discrete), Multiple Rooms (nominal), Business (nominal), Cleanliness Rating (ordinal) and Bedrooms (discrete)
* **Proximity:** Distance from a certain location. This comprises two attributes including City Center (km) (continuous) and Metro Distance (km) (continuous)
* **Indices:** Measures of value of the AirBnB. Comprises 4 attributes, including Attraction Index (continuous), Normalised Attraction Index (continuous), Restaurant Index (continuous), and Normalised Restaurant Index (continuous)
* **Room type**: Describes what type of room it is, using boolean values. There are three attributes including Entire home/apt (nominal), Private room (nominal), and Shared room (nominal)
* **Guest** **Rating**: Provides details about the guest stay experience. This includes one attribute: Guest Satisfaction (ordinal).

Based on these categories, we chose 8 variables to focus on, however our data analysis was conducted on all 16 variables to prevent bias and to fully understand the correlation between them before narrowing them down to our variables of focus. These variables are:

| Price   * Most data lies between the lower range of $0-$2,000 * Some extremely high price values around $13,000 * Potential to be skewed data | Bedrooms   * Most AirBnBs have 1 bedroom, with the next highest being 2, 0, and 3. * Not many records with 4+ bedrooms |
| --- | --- |
| City Center (km)   * Most of the data lies within 10km away from city center * Largest point is just above 25km away from city center | Metro Distance (km)   * Most of the data lies within 4km of the metro station * Largest point is just above 14km away from the metro station |
| Restaurant Index   * Most of the data lies between 0-3,000 * One extremely high value of around 6,600 | Person Capacity   * The mode is a person capacity of 2, with the next highest being 4 * No Airbnbs have a person capacity of 1, and the largest person capacity an Airbnb can hold is 6 |
| Cleanliness Rating   * The mode is 10 for cleanliness rating (the highest rating for an Airbnb) with the next highest being 9 * The data appears to be exponential * No Airbnbs have a cleanliness rating of 1 | Guest Satisfaction   * The mode is a rating of "Excellent" for Airbnbs * Very few records have a rating of "Bad" or "Very Bad" * Imbalance of records, potentially making it difficult to train a model to predict this variable as it will be biased to "Excellent" rated Airbnbs |

# 

# Data pre-processing

Before we can begin creating our model, we need to clean the data.

1. Drop Price null values

Price is our dependent variable for this project. There were 33,372 non-null values and 8,342 null values for this column and because this is our dependent variable, we need to handle the missing values.

When handling missing values, there are three main categories of methods that can be used: deletion, imputation, and modelling (Kang, 2013). Price is our target variable, so we chose not to impute the values to prevent bias in our trained model. With a large dataset (41,714 records), deletion will have a marginal impact training the model, so we opted to delete the records with null Price values.





1. Clear remaining null values using mean imputation

We can see in Figure 9 that every column now has 33,372 entries, except for two of our independent variables, City Center (km) and Metro Distance (km), which have missing values. We have lost a significant amount of data from Step 1 after deleting records of null Price values and a further deletion of records with null City Center (km) and Metro Distance (km) values could significantly compromise the strength of our model. Hence, we chose to use mean imputation, i.e. computing the mean value of each column and replacing the null values with the mean value.





This makes up for 677 City Center (km) values and 3,366 Metro Distance (km) values.

1. Use Label encoding for Guest Satisfaction

Guest Satisfaction comprises data categories of "Very Bad", "Bad", "Not bad", "Good", and "Excellent". To use this column as input for our machine learning model, we need to use label encoding to turn these into numeric values, i.e. allocating each category with a number and replacing each instance of that category with the number.



# Data exploration

Now that we have cleaned our data, we can do some initial data exploration to understand if there is any significant correlation between the variables. 

We initially hypothesized that closer proximity to the city center would be associated with higher rental prices, as the advantages and convenience of staying near the city center tend to increase the demand for AirBnb listings, and that there would be a negative correlation where a decrease in the metro station distance results in an increase in rental prices.

Figure 11 shows that AirBnBs positioned closer to the City Center exhibit relatively higher rental prices, compared to those with a more rural status. Similarly, Figure 12 illustrates the relationship between the proximity of a metro station and the rental prices. This relationship is also intuitive, as it reflects the convenience of being close to transportation for easy citywide travel. Both plots may not exhibit a particularly strong relationship as there are data points that could be potential outliers. However, the underlying trend for both relationships depict a right skewed distribution where an abundance of data points lie on the left side of the plot.

Looking at a full correlation table on all 16 variables, we can see from Figure 13 that there does not appear to be any significant correlation between Price and another variable.



This is surprising as we would expect that these variables do have an impact on how the prices of rentals are determined. Looking into this further, it appears that City Center (km) and Metro Distance (km) have no correlation at all, with values of -0.058 and -0.036 respectively, while the highest correlation is Normalised Attraction Index at 0.29 — still fairly low.

In the initial data exploration, we removed the non-correlated variables (defined to be correlation -0.2 to 0.2) and created a new correlation table with the updated features.





The updated correlation table in Figure 14 shows the relatively most correlated variables to be Bedrooms, Normalised Attraction Index, and Normalised Restaurant Index. These have a correlation to price of 0.22, 0.29, and 0.23 respectively.

# 

# Modelling

In our pursuit of predicting rental prices for Airbnb listings, we have opted for the Linear Regression method. This type of machine learning model will allow us to recognise which features or explanatory variables are the most influential to predict rental prices, enabling us to optimise rental predictability. After conducting preprocessing alterations to our dataset, we fitted an initial model as illustrated in Figure 15.





On average, the model’s prediction exhibits a deviation of approximately 106.13 Euros, as reflected by the Mean Absolute Error (MAE) of 106.13. Given that the bulk of rental prices in our dataset fall within a range of 0 to 2000 Euros, a deviation of 106.13 Euros raises a significant concern. This level of discrepancy could potentially impede the model’s ability to predict Airbnb rental prices. The Root Mean Squared Error (RSME) shows us a deviation of 201.66 Euros, but this may be due to the outliers in the dataset. Furthermore, an R-Squared value of 0.21 signifies that just 21% of the variability in rental prices are accounted for by the factors included in this initial model. This emphasises the model’s limitations and its inability to extensively explain the variations in rental prices.

Additionally, we took a more comprehensive evaluation of the model’s performance, particularly focusing on the residuals. This is a crucial part of verifying the underlying assumptions within linear regression modelling. With close examination of the residuals plot (Figure 16), we can see that as predicted prices increase, there is an increase in variability of residuals. This represents a violation of the constant-scatter criterion, a critical assumption in linear regression.



We then examined the distribution of each continuous explanatory variable, and observed that all of them exhibit skewness. Skewness in explanatory variables can compromise regression assumptions and diminish model accuracy. To address this, we applied a log transformation to the explanatory variables, ensuring they follow a more normal distribution. For instance, in Figure 17, it illustrates the distribution of Airbnbs and their respective distances from the City Center. The distribution appears to be right skewed, deviating from a normal distribution.







As part of the log transformation process, we excluded the “Normalised Restaurant Index” from the columns we transform. This was mainly due to the lack of improvement regarding skewness as observed in Figure 18. The explanatory variables that we will log transform are: City Center, Metro Distance, Attraction Index, Normalised Attraction Index, and Restaurant Index. 

Moreover, we found it necessary to apply log transformation to our response variable “Price”, due to its strong skewness.

A correlation analysis of the log-transformed dataframe can be found in Appendix A.

Logged Model Code:





The final model, as shown in Figure 22, better aligns data points with the ideal 1:1 trend line. However, significant variability persists, impacting the R-Squared value and the model's validity. Acknowledging and addressing these dataset limitations is essential.

Interpretation



* Y-intercept: With no explanatory variable, Price is estimated to be 50.09 Euros.
* A one unit increase in Person Capacity, resulted in an average increase of about 1.04 times in Price.
* An extra Bedroom is estimated to result in an increase of 1.21 times in Price.
* Notably, a one unit increase in the Normalised Attraction Index was estimated to have an increase by about 1.51 times in Price.
* Entire home/apt categorised as ‘Home’ saw an estimated increase in Price by about 1.29 times.

Dataset Limitations

According to the dataset’s metadata, each city represents a unique cultural and lifestyle tapestry. This variability posed challenges capturing niche city variations using a single linear regression model. Future improvements could include modelling linear regression for each specified city. Additionally, we excluded 8,000 null target values and opted for mean imputation for other missing features, potentially reducing variability and introducing bias to the dataset.

# Discussion/Conclusion

Our analysis of European Airbnb listings aimed to uncover the dynamics influencing pricing in the hospitality industry. As we delved into the dataset, we explored the significance of factors like location, reputation, amenities, and various indices. Notably, the data revealed that properties nearer to city centers commanded higher rental prices, reflecting the appeal of convenience and accessibility. Additionally, proximity to metro stations correlated inversely with rental prices, as expected. However, the complexities of these relationships hinted at potential outliers. Surprisingly, certain variables, such as City Center (km) and Metro Distance (km), displayed minimal correlation with pricing, suggesting the existence of subtler, multifaceted determinants in Airbnb pricing.

One key outcome involves the development of educational resources for hosts. By emphasizing factors such as location, reputation, and amenities, hosts can gain a deeper understanding of the elements that influence rental prices. Educated hosts are better equipped to create listings that align with guest expectations, potentially leading to greater guest satisfaction, improved reviews, and, ultimately, more bookings. This, in turn, can elevate Airbnb's reputation and establish it as the premier platform for hospitality. Another critical business implication pertains to enhancing transparency concerning pricing determinants. By furnishing more detailed information on pricing, Airbnb can cultivate trust among hosts and guests. Transparent pricing allows users to make informed choices, reducing uncertainty. This elevated trust and satisfaction can enhance the reputation of individual hosts and the Airbnb platform, potentially resulting in increased bookings and greater user loyalty.

Our analysis not only delves into the intricacies of Airbnb's pricing dynamics but also underscores opportunities for Airbnb to bolster host support and improve user experiences through education and transparency. By implementing these business strategies, Airbnb can cement its position as a frontrunner in the hospitality industry while fostering mutually beneficial relationships between hosts and guests. While our linear regression model holds promise, the lack of strong correlations between particular variables and pricing underscores the intricate nature of Airbnb pricing dynamics. Our hypothesis suggested that factors like location, reputation, amenities, and various indices substantially impact Airbnb's pricing structure. The data revealed several noteworthy trends, such as the positive correlation between proximity to city centers and higher rental prices, indicating the significance of location.

# 

# References

Georgevici, A., & Terblanche, M. (2019). Neural networks and deep learning: a brief introduction. *Intensive Care Medicine*, *45*(5), 712–714. https://doi.org/10.1007/s00134-019-05537-w

Kang, H. (2013). The prevention and handling of the missing data. Korean Journal of Anesthesiology, 64(5), 402–406. https://doi.org/10.4097/kjae.2013.64.5.402

Verma, D. P. S., & Gupta, S. S. (2004). Does higher price signal better quality? *Vikalpa*, *29*(2), 67–78. https://doi.org/10.1177/0256090920040206

Zhang, Z., Chen, R. J. C., Han, L. D., & Yang, L. (2017). Key Factors Affecting the Price of Airbnb Listings: A Geographically Weighted Approach. *Sustainability, 9*(9), 1635. https://doi.org/10.3390/su9091635

# Appendices

Appendix A

A correlation analysis of the log transformed dataframe in Figure A.1 shows a much better correlation of variables than our first table from Figure 13, albeit still not overly high. We can see the potential for multicollinearity between a multitude of variables, in which we decided to eliminate the variable with the weakest correlation with Price:

Attraction Index (0.36) and Restaurant Index (0.3) >> Collinearity value of 0.94

* **Drop Restaurant Index**

Attraction Index (0.36) and Normalised Attraction Index (0.48) >> Collinearity value of 0.78

* **Drop Attraction Index**

Normalised Attraction Index (0.48) and Normalised Restaurant Index (0.39) >> 0.74

* **Drop Normalized Restaurant Index**

Room\_Type\_Enture home/apt (0.29) and Room\_Type\_Private room (-0.27) >> -0.98

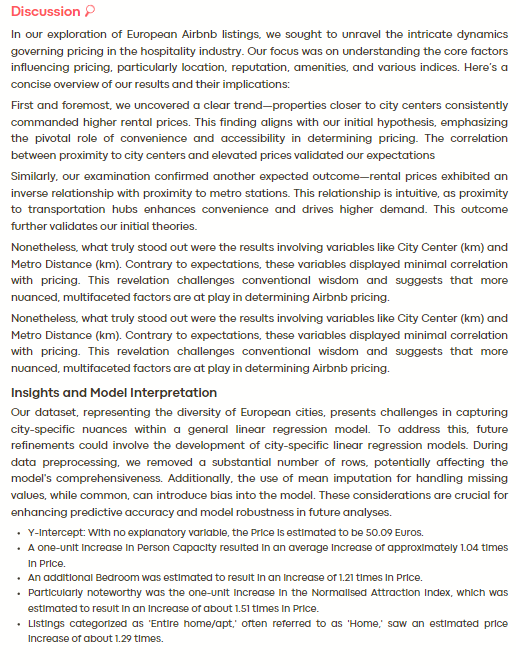
* **Drop Room\_Type\_Private room**

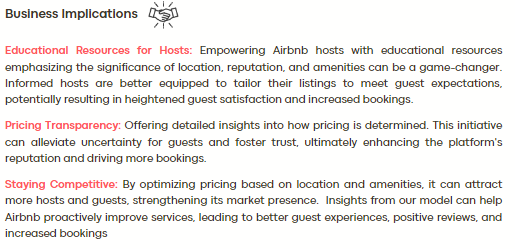


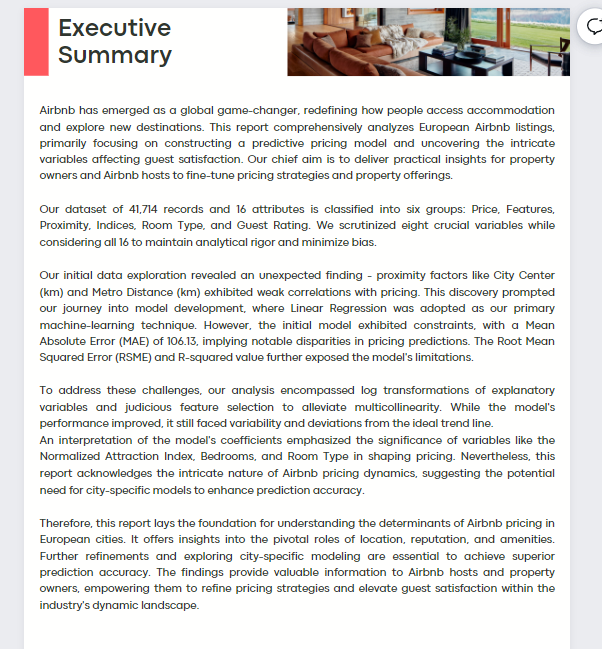


After addressing multicollinearity, we arrived at a final set of explanatory variables showing enhanced correlations compared to the initial dataset. This improvement is evident in our final model's Actual vs Predicted plot.

# 







Airbnb has emerged as a global game-changer, redefining how people access accommodation and explore new destinations. This report comprehensively analyzes European Airbnb listings, primarily focusing on constructing a predictive pricing model and uncovering the intricate variables affecting guest satisfaction. Our goal is to provide actionable insights for property owners and Airbnb hosts to fine-tune their pricing strategies and property offerings.

Our dataset consists of 41,714 records and 16 attributes, categorized into Price, Features, Proximity, Indices, Room Type, and Guest Rating. We identified 8 variables as critical out of a total of 16. Despite focusing on these 8, we took into account all 16 variables to ensure our analysis remained balanced and unbiased.

Our initial data exploration revealed an unexpected finding - proximity factors like City Center (km) and Metro Distance (km) exhibited weak correlations with pricing. This discovery prompted our journey into model development, where Linear Regression was adopted as our primary machine-learning technique. However, the initial model exhibited constraints, with a Mean Absolute Error (MAE) of 106.13, implying notable disparities in pricing predictions. The Root Mean Squared Error (RSME) and R-squared value further exposed the model's limitations.

To predict pricing, we employed a Linear Regression model, implementing feature engineering techniques such as log-transformations. The final model achieved an R-squared value of 0.39 and pinpointed the following key explanatory variables with the most influence on Price:

* Person Capacity, with a coefficient of 1.04.
* Bedrooms, with a coefficient of 1.21.
* Normalized Attraction Index, with a coefficient of 1.51.
* Entire home/apt, with a coefficient of 1.29.

However, our analysis uncovered specific limitations::

* The dataset includes Airbnb listings from various European cities, making it challenging to predict prices universally across all cities.
* Missing data resulted in the exclusion of approximately 8,000 records from model training, potentially valuable information.
* Linear Regression assumes a linear relationship between input features and the continuous target variable, which may not always hold true.

This report lays the foundation for understanding the determinants of Airbnb pricing in European cities. It offers insights into the pivotal roles of location, reputation, and amenities. It empowers Airbnb hosts and property owners to optimize their pricing strategies and property offerings, enabling Airbnb to maintain its competitive edge and continue establishing itself as the premier platform for hospitality.