Exploring Airbnb Listings in Lisbon: A Data-driven Analysis

Baran Can Çelik (20232067@novaims.unl.pt, 20232067); Carlos Eduardo Nunes Lourenço (20232020@novaims.unl.pt, 20232020); Priyá Dessai (20232053@novaims.unl.pt, 20232053); Raysa Rocha (20232051@ novaims.unl.pt, 20232051).
Group. B7 **TP.** 2

Abstract. This project analyzes Airbnb listings in Lisbon to provide insights for hosts, investors, and guests. We imported, cleaned, and identified key numerical and categorical variables in the dataset. Outliers and missing values were addressed using statistical methods and the K-NN Imputer.
We normalized the data and explored K-Means, DBSCAN, and K-Means with PCA for clustering. K-Means, chosen for its balanced clusters, identified five distinct groups: Luxury Rentals, Standard Rentals, Budget-Friendly Rentals, Corporate Housing, and Extended Stay Rentals. These clusters were further analyzed with categorical variables.
The study revealed patterns in pricing, rental frequency, and property types across neighborhoods. Despite challenges with outliers and unclear listing dates, our findings highlight Lisbon's diverse Airbnb market. The results suggest targeted marketing and improved guest facilities, showcasing the benefits of data science in optimizing Airbnb listings.

Keywords. Clustering, Airbnb, Lisbon Market Analysis

Introduction

We chose to study Airbnb listings in Lisbon due to the market's potential for data-driven insights beneficial to hosts, investors and guests. Motivated by the growing popularity of Airbnb rentals, we seek to uncover insights into pricing trends, property characteristics, and geographic patterns of Airbnb listings in Lisbon. By leveraging data science techniques, particularly unsupervised machine learning algorithms such as clustering, we aim to assist potential guests in making informed decisions, uncover hidden patterns, and segment the market effectively.

Data and Methods

We began by importing the raw dataset "listings.csv" from insideairbnb.com/lisbon webpage and confirmed that there were no duplicates. By reviewing the DataFrame, we verified that it contained 22,929 rows and 18 columns, as well as 4 floats, 7 integers, and 7 objects. After that, we focused on the numerical variables: 'price', 'host_id', 'minimum_nights', 'reviews_per_month', and 'calculated_host_listings_count', as well as the categorical variables: 'host_name', 'neighbourhood', 'neighbourhood_group', and 'room_type'.

Using visual exploration techniques like boxplots, we identified outliers and later checked for missing values in 'price' and 'reviews_per_month' columns. In the preprocessing phase, we removed observations with prices above 2000 and minimum nights exceeding 365 as these values did not fit the overall data pattern and we retained the remaining outliers to maintain the integrity of the data. Additionally, the missing values in 'reviews_per_month' were filled with zeros and for the 'price' column, we utilized the K-NN Imputer, based on 'price', 'neighbourhood', and 'room_type' columns.

Lastly, we merged 'host_name' and 'host_id' into a new column named 'host_info' to accurately distinguish hosts.

Before applying the clustering algorithms, we standardized the data using StandardScaler, as it handles outliers more effectively than MinMaxScaler. In order to obtain well-defined clusters and meaningful insights from the data, we explored different algorithms such as K-Means, DBSCAN and K-Means combined with PCA. During K-Means application, the Elbow and Silhouette Methods were used to determine the optimum number of clusters. Throughout these stages, we utilized essential Python libraries like Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn and Plotly to conduct and visualize our analysis efficiently.

Results and Discussion

Despite DBSCAN achieving the highest Silhouette Score, K-Means was chosen due to its performance and more balanced cluster distribution, resulting the following five clusters:

- **Cluster 0**: This cluster consists of 692 data points and features high-end properties with an average price of 651.69. These listings cater to exclusive clientele, indicated by low rental frequency.
- **Cluster 1**: Comprising 15,746 data points, this cluster represents the most common rental type with a moderate average price of 106.25 and balanced rental frequency.
- **Cluster 2**: Containing 5,231 data points, this cluster is characterized by affordability with an average price of 95.84 and high rental turnover.
- **Cluster 3**: Including 1,191 data points, this cluster features listings with a moderate average price of 107.33 and high minimum night requirements, catering to long-term renters.
- **Cluster 4**: The smallest cluster with 46 data points, featuring niche properties for very long-term stays with an average price of 133.97.

Having defined our clusters using numerical variables, in the following section we aim to enhance our analysis by integrating categorical variables such as 'neighbourhood', 'neighbourhood_group', 'host_info', and 'room_type'.

Cluster 0, labeled as Luxury Rentals, includes high-end properties with significantly higher prices, averaging €652 per night. In this cluster, we predominantly find entire home/apartment as room types (87%), with moderate minimum night requirements averaging around 3 nights. The low number of rentals suggests that these hosts serve a selected group of people who appreciate luxury commodities. The reviews per month are quite low, averaging 0.50, but this does not seem to impact the customer choice. Leading hosts in this segment, such as 'RoomPicks' and 'OhMyGuest', are presumed to be enterprises or rental management agencies. Lastly, they are situated in exclusive neighborhoods like Cascais e Estoril, Santo António, Santa Maria Maior, and Colares.

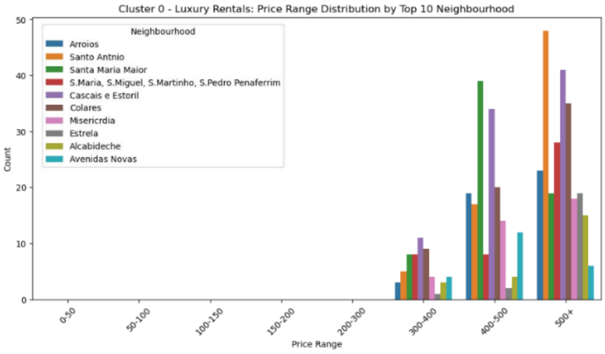

Figure 1: Cluster 0 - Luxury Rentals: Price Range Distribution by Top 10 Neighbourhood - click on the link to view it bigger

Cluster 1, labeled as Standard Rentals: This is the largest cluster consisting of typical rental properties with moderate pricing, averaging €106 per night, targeting primarily families and business travelers. These listings, which are mostly entire home/apartment (70%) and private rooms (27%), generally have slightly higher minimum night requirements, averaging 4 nights. With 0.63 reviews per month, this cluster is the second most reviewed. The properties are located in popular neighborhoods like Santa Maria Maior, Arroios, Misericórdia, and Cascais e Estoril. Top hosts managing these rentals include companies 'Travelservices', similar to Cluster 0, as well as individuals such as 'Luís' and 'Albertino'.

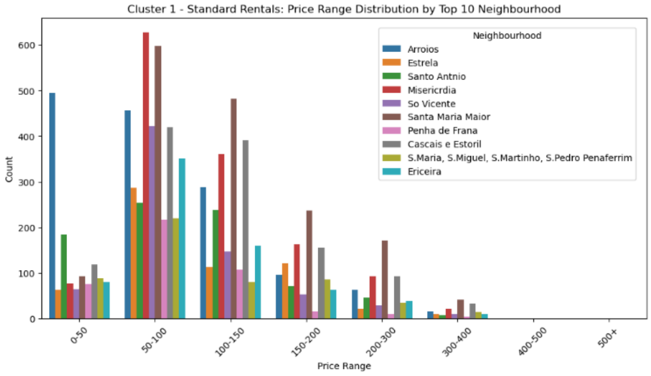

Figure 2: Cluster 1 - Standard Rentals: Price Range Distribution by Top 10 Neighbourhood - click here to view it bigger

Cluster 2, labeled as Budget-Friendly Rentals, is characterized by the most affordable prices, averaging €96 per night and high rental frequency, targeting mostly budget-conscious travelers, students, and frequent Airbnb renters. Properties in this cluster have the lowest minimum night requirements, averaging almost 2 nights, and are frequently booked, indicating high demand. Moreover, this cluster stands out with the highest average number of reviews per month, at 3.36. The cluster predominantly consists of entire home/apartment (82%) and private rooms (17%). Hosts like 'Alexandra Pedro And Team' and 'Sarah & Pedro' specialize in providing cost-effective accommodations in accessible neighborhoods such as Santa Maria Maior, Misericórdia, Arroios, and São Vicente.

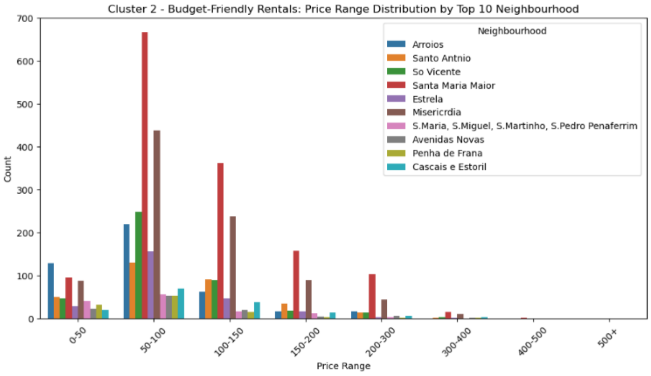

Figure 3: Cluster 2 – Budget-Friendly Rentals: Price Range Distribution by Top 10 Neighbourhood - click here to view it bigger

Cluster 3, labeled as Corporate Housing: This cluster consists of listings for long-term business stays, with moderate prices averaging €107 per night and higher minimum night requirements, averaging nearly 13 nights. Targeting primarily business travelers and expatriates, these properties are generally entire home/apartment (79%). Additionally, with moderate reviews per month, averaging 0.54, these listings are managed by property companies like 'Feels Like Home' and 'Blueground'. They are situated in central neighborhoods, including Santo António, Santa Maria Maior, Arroios, and Avenidas Novas.

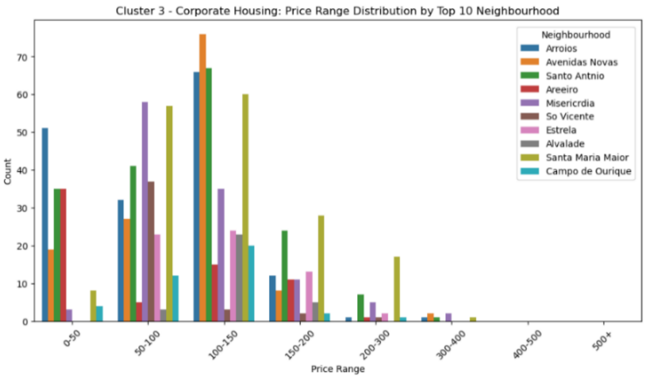

Figure 4: Cluster 3 - Corporate Housing: Price Range Distribution by Top 10 Neighbourhood - click here to view it bigger

Cluster 4, labeled as Extended Stay Rentals, is the smallest cluster, offering properties for very long-term stays with exceptionally high minimum night requirements, averaging nearly 269 nights. These listings are priced slightly higher compared to both Standard and Budget-Friendly Rentals, at €134 per night, and predominantly entire home/apartment (93%). Additionally, with moderate reviews per month, averaging 0.54, similar to Cluster 3. Targeting expatriates and long-term travelers, these rentals are managed by notable hosts including companies such as 'Martinhal Hotels' and individuals like 'Paulo Eduardo' and 'Tiago'. These properties are situated in neighborhoods including Misericórdia, Santo António, Arroios, and Parque das Nações.

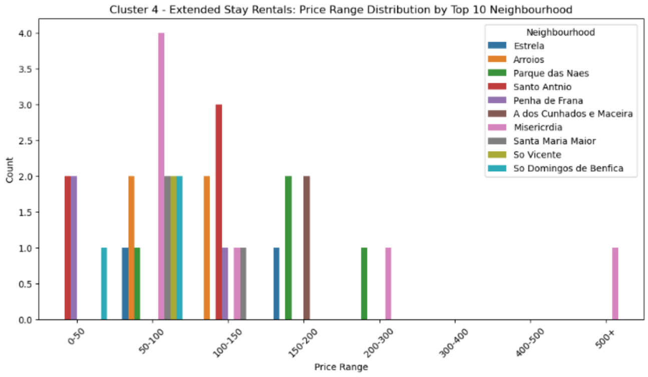

Figure 5: Cluster 4 - Extended Stay Rentals: Price Range Distribution by Top 10 Neighbourhood - click here to view it bigger

Conclusions

After analyzing Airbnb listings in Lisbon through the K-Means application, we identified five clusters: Cluster 0 (Luxury Rentals), Cluster 1 (Standard Rentals), Cluster 2 (Budget-Friendly Rentals), Cluster 3 (Corporate Housing), and Cluster 4 (Extended Stay Rentals). These rentals vary in price, target guests, and property types. Some receive more reviews, while others have loyal customers. Location also matters, as rentals in different neighborhoods attract different guests. Our study shows that Lisbon's Airbnb market is diverse, helping hosts, investors, and guests make better decisions.

During our analysis, we encountered several challenges. The wide range of outliers across all selected numerical variables, combined with unclear listing dates, caused challenges for the analysis and calculation of reviews. Varying minimum night requirements resulted in some places receiving fewer reviews. Additionally, the lack of demographic data prevented us from displaying the dominant gender, age range, or income in our clusters. The DBSCAN's imbalanced data distribution made it unsuitable for our analysis. Lastly, PCA was challenging due to the inefficacy of 4 components and the difficulty in interpreting 3-component results. Hence, PCA was excluded from clustering as its Silhouette Score was lower than K-Means.

Consequently, our analysis highlights Lisbon's complex Airbnb market, offering valuable insights. Based on these findings, we recommend implementing several actionable strategies:

- **Cluster 0 (Luxury Rentals)**: Partnering with high-end travel agencies to reach wealthy travelers seeking luxury accommodations.
- **Cluster 1 (Standard Rentals)**: Promoting listings in popular neighborhoods through local tourism websites and business travel platforms to increase visibility.
- **Cluster 2 (Budget-Friendly Rentals)**: Encouraging guests to leave reviews by offering small incentives, as high review counts can attract more bookings. Promoting listings on budget travel websites to reach the target audience.
- **Cluster 3 (Corporate Housings)**: Providing business travelers with essentials such as high-speed internet, workspaces, and early check-in/late check-out options. Partnering with local businesses and corporations to offer accommodation packages for relocating employees and business trips.
- **Cluster 4 (Extended Stays Rentals)**: Offering long-term stay discounts and flexible lease options to attract expatriates and long-term travelers.

In conclusion, our project showed how data science can improve Airbnb listings in Lisbon. By breaking down the market into different segments, property managers can refine their strategies, boost marketing efforts, make smarter investments, and enhance guest satisfaction, leading to better overall performance in the Airbnb rental market.

Statement of contribution & Acknowledgments

Contributions of each group member were divided as follows: Raysa and Priyá primarily worked on the Jupyter Notebook, while Carlos and Baran focused on the report, but all group members contributed to both parts and to the presentation. Additionally, Prof. Vitor provided valuable Data Preprocessing assistance.

References

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt & J. Millman (Eds.), Proceedings of the 9th Python in Science Conference (pp. 51-56).

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585(7825), 357-362.

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90-95.

Waskom, M. L. (2021). Seaborn: Statistical Data Visualization. Journal of Open Source Software, 6(60), 3021.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Inc., Plotly Technologies. (2015). Collaborative data science. Montréal, QC: Plotly Technologies Inc. Available at: https://plotly.com.

Grus, J. (2015). Data science from scratch: First principles with Python (First edition). O'Reilly.

VanderPlas, J. (2016). Python Data Science Handbook—Essential Tools for Working with Data (First Edition). O'Rilley.