2024

Navigating AIRBNB Market

ANALYSING AIRBNB PRICES IN BUENOS AIRES

ANNA CINCOTTA, SAHANA BAPAT AND SAHIBNOOR CHAHAL

655: Data Mining in Business – Capstone Project

Introduction

The competitive and fluctuating market for Buenos Aires Airbnbs creates an array of challenges for hosts. They may struggle to maximize occupancy, optimally price their listings, and understand guest preferences. Traditional pricing and demand forecasting lacks the ability to account for the complex interaction of factors. Guest feedback provides valuable insights, however, manually parsing through wastes time and effort and results in an inadequate analysis. This project seeks to develop a machine learning model that tackles these challenges through three interconnected components: demand analysis, price prediction, and topic modeling.

Problem Statement

Airbnb hosts in Buenos Aires face challenges predicting demand for their listings, due to the city's fluctuating seasons, economic factors, and traveler's unique preferences. Upon thorough review of the data we had access to, we decided to develop a model which identifies trends over time and predicts the price of a particular listing according to basic amenities. This model will allow hosts to make data-backed pricing decisions, maximize occupancy and profitability, and better accommodate the city's seasonal demand cycles.

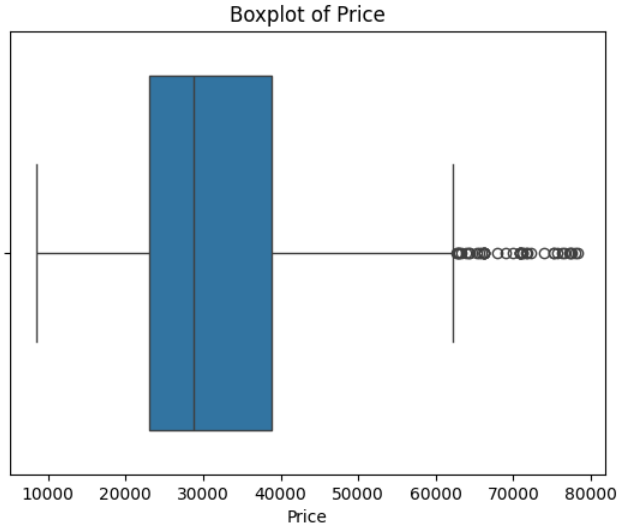
Data Preprocessing

Upon initial review of data for Airbnb listings of Buenos Aires, we noticed several columns that displayed null or irrelevant data regarding the problem statement we decided to address. We evaluated every field that was provided in the dataset and decided which ones would provide us with an understanding of the city's market for Airbnb listings. For this purpose, we used the following techniques to arrive at the most appropriate and cleaned data for analysis.

- **Feature importance:** *RandomForestClassifier* was used to perform feature importance to highlight what features need to be included for the most efficient price prediction model.
- **Relatability:** Columns like *latitude* and *longitude* don't make much sense when considering price prediction. Hence, these 2 columns were excluded.
- **Missing data handling and data processing**:
    - ***Price:*** Delete records since it is the variable we're predicting, and not having price will not make sense for us.
    - ***Property Type and Room Type*:** One-hot encoding to find a correlation between price and property type and room_type.
    - ***Bathrooms, Bedrooms and NumberOfBeds:*** The most appropriate technique would be to use mode (the highest occurring value) as mode gives a whole number that makes sense. Arriving at mean and median would alter the correctness.

- ***Has_Review_Scores:*** Missing values filled with 0, and one-Hot_encoding to convert 'Review_Scores_rating' to a boolean value.

- ***Outliers in price***: Prices beyond lower bound (less than 10000) and above upper bound (greater than 62000) were eliminated to reduce the effect of outliers on the model.



Plot 1*: Identifying outliers in price*

We also used the reviews data file to perform sentiment analysis.

- All the comments were translated from their original language to English.
- These translated comments were cleaned using lemmatization by removing all the stop words.
- Sentiment polarity was calculated for these cleaned comments, following which they were all categorized into 3 categories; **positive**: sentiment>0.8, **moderate:** sentiment >0, **negative:** rest of the values.
- This final dataframe was then merged with the rest of the data and one-hot-encoding was performed to convert all the data into numeric values.

The final dataframe had the following variables:

- *price*
- *latitude*
- *property_type*
- *room_type*
- *accommodates*
- *bathrooms*
- *bedrooms*
- *beds*
- *minimum_nights*
- *maximum_nights*

- *has_availability*
- *number_of_reviews*
- *instant_bookable*

Exploratory Data Analysis

Descriptive Analysis

1. **Latitude**: Minimal variation (std = 0.016), indicating all listings are geographically close within Buenos Aires.
2. **Bedrooms**: Average listings have about 0.97 bedrooms, with a maximum of 4. Most listings (75th percentile = 1) cater to smaller accommodations, as the majority have one bedroom.
3. **Minimum Nights**: Mean ~ 2.9 nights, with a highly skewed distribution (max = 100, 75th percentile = 3). Most listings have short minimum stay requirements, but a few have significantly higher values.
4. **Maximum Nights**: Large variability (std = 449.77), with a mean of ~485.36 nights. Most listings allow long stays, as seen from the 75th percentile (1,125 nights).
5. **Sentiment**: Average sentiment is neutral to positive (mean = 0.45). The range spans from -0.7 to 1.0, indicating variability in guest feedback.

The data shows that most Airbnb listings in Buenos Aires are highly rated, cater to smaller groups (1-bedroom), and allow for flexible stays (short minimum nights, long maximum nights). There is minimal geographic variation, and sentiment analysis shows mixed but slightly positive feedback overall.

	count	mean	std	min	25%	50%	75%	max
latitude	868.0	-34.590670	0.015965	-34.65539	-34.599972	-34.589384	-34.581313	-34.53726
bedrooms	868.0	0.972350	0.656207	0.00000	1.000000	1.000000	1.000000	4.00000
minimum_nights	868.0	2.993088	5.552098	1.00000	1.000000	2.000000	3.000000	100.00000
maximum_nights	868.0	485.365207	449.777206	2.00000	65.000000	365.000000	1125.000000	1125.00000
sentiment	868.0	0.456320	0.252222	-0.70000	0.324716	0.450000	0.600000	1.00000

Table 1 *Descriptive Analysis of numerical data*

Geo-Map of listings in Buenos Aires *(GeoMap)*

By visualizing a sample of 1000 records for the density of listings across different areas of Buenos Aires, we can identify that the eastern parts of Buenos Aires are the most popular neighborhoods or areas with a high concentration of Airbnb properties. The reason being that the most popular tourist sites are found in the historic core of the city, in the Montserrat and San Telmo neighborhoods and have a higher proximity to the airports and beaches. This

information can be useful for understanding market dynamics and potentially identifying areas with higher demand.

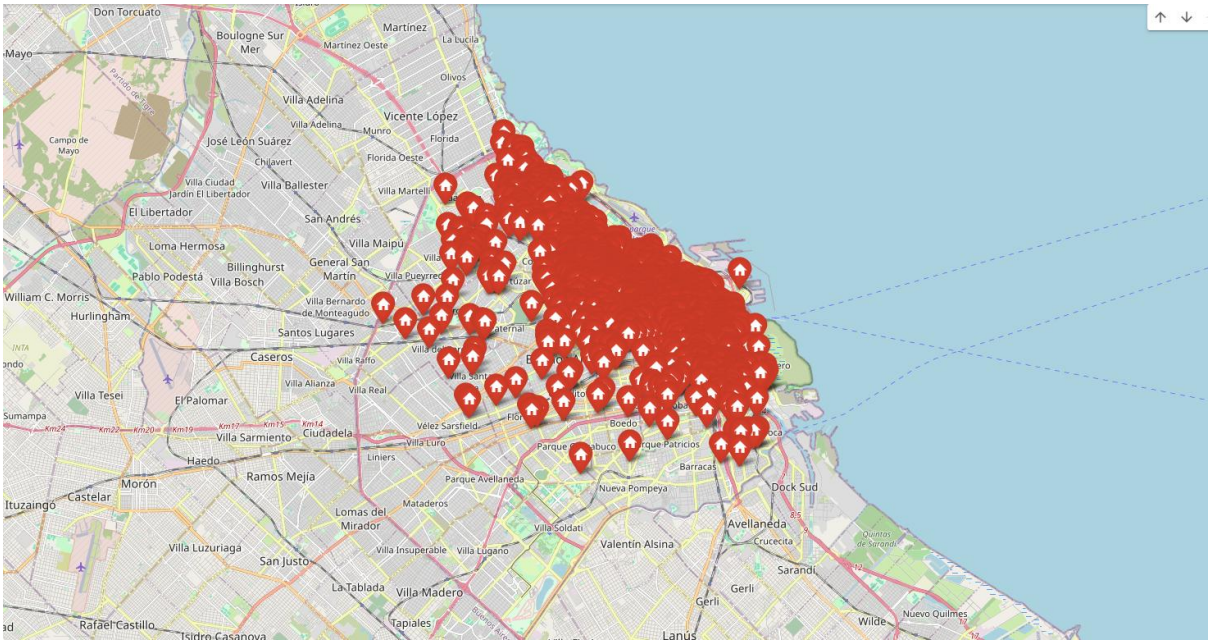
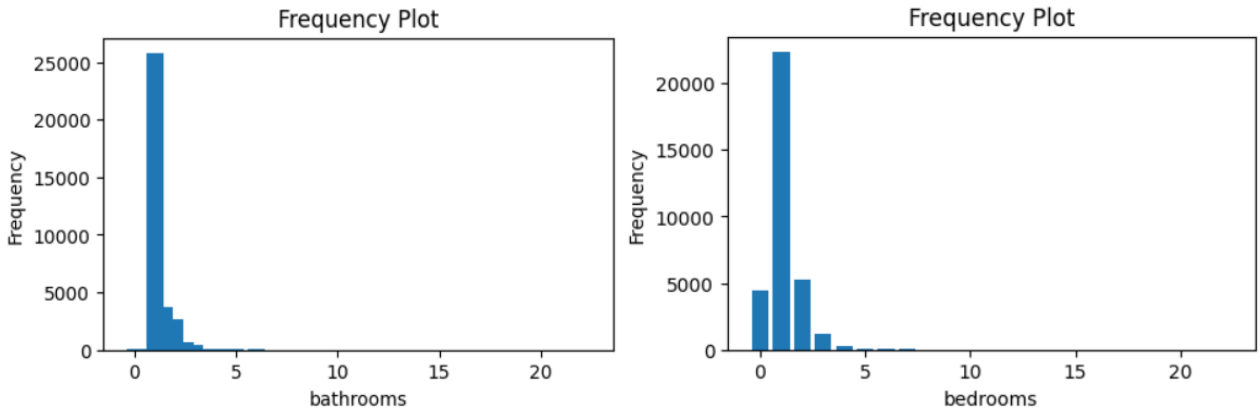


Figure: 1 *Geo Map of listings in Buenos Aires*

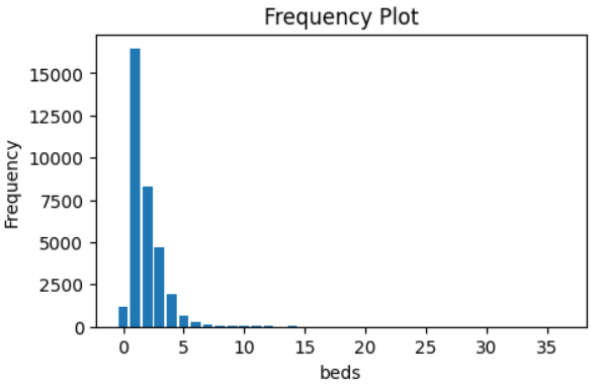
Histograms

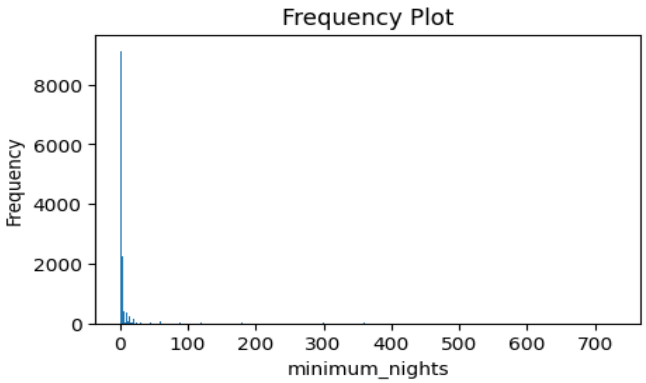
Categorical Data:

1. ***Bathrooms:*** The distribution shows a concentration of listings with 1 to 4 bathrooms, with fewer properties having more. This could indicate that most Airbnb listings cater to small to medium-sized groups, aligning with typical guest needs.
2. ***Bedrooms:*** The distribution is right-skewed, with a majority of listings having 2-3 bedrooms. This reinforces the idea that the market primarily serves smaller groups or families, as larger homes are less common.

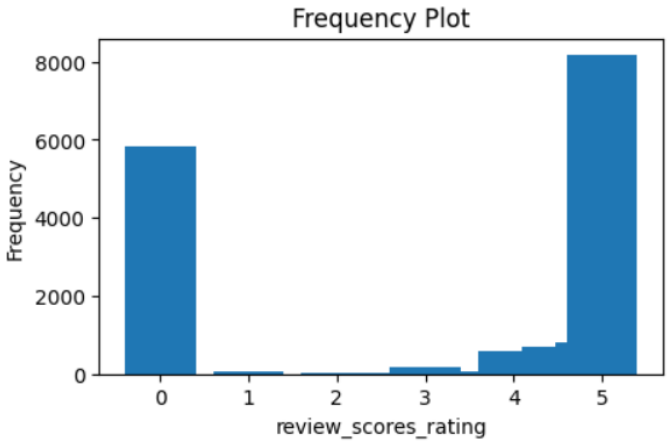


3. ***Beds:*** The distribution, similar to bedrooms and bathrooms, is right skewed indicating that airbnb's with 1-4 beds on an average are preferred. This shows a higher tendency for customers to choose listings that are quiet and less crowded, making privacy a higher priority.
4. ***Minimum_nights:*** This histogram may show peaks at certain values, indicating common minimum stay requirements of 0-98 nights. The distribution shows that the hosts are accommodating and accept a few hours or even a lengthy duration of stay.



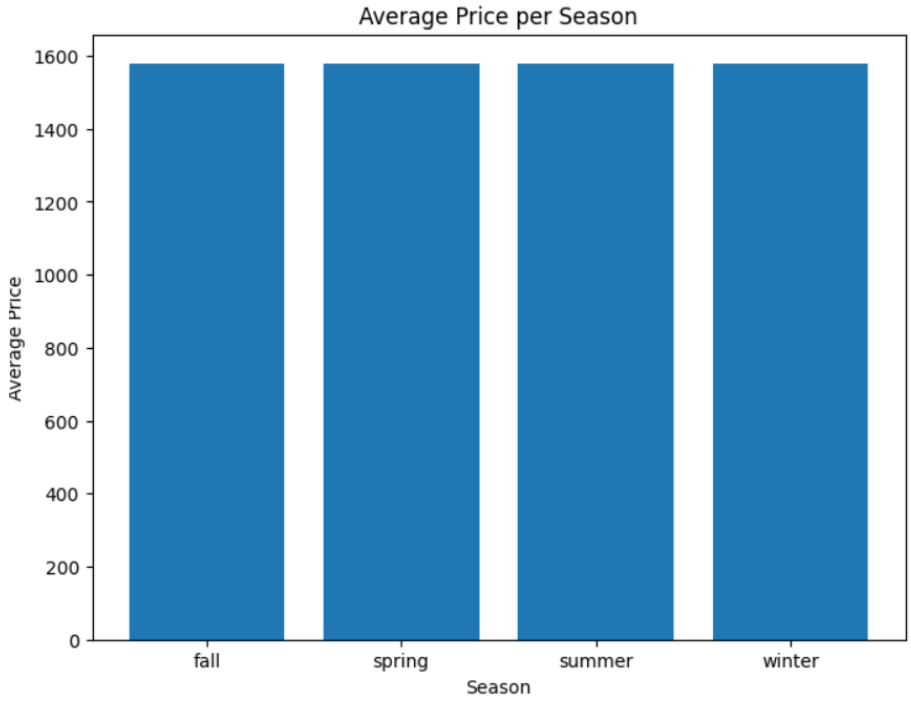


5. ***5. Review_scores_rating:*** *most listings have a review score of 0 or 5, with very few ratings between 1 and 4, indicating a dominance of unrated or highly-rated listings*.



Seasonal Demand Variation

- After importing the data and running analysis, it was clear that this new data file was not good enough to provide any seasonal data. As seen from the chart below, the average price per season remained the same. Upon further study, average price remained the same across months as well.



Looking at the table below, the price range is huge ranging from 0 to 10,000,000. Also, this data does not match with the listings data, which is our main data file.

	price
count	1.343457e+07
mean	1.579721e+03
std	6.094136e+04
min	0.000000e+00
25%	2.900000e+01
50%	4.000000e+01
75%	6.000000e+01
max	1.007510e+07

Topic Modelling for Reviews

LDA Topics:

- **Topic 1:** location, apartment, great, stay, place, clean, recommend, would, host, nice

  This topic revolves around the overall positive experience of the stay, highlighting location, apartment quality, cleanliness, and host hospitality.

- **Topic 2:** palermo, area, walk, restaurants, safe, close, everything, nice, barrio, apartment

  Possible Interpretation: Focuses on the neighborhood of Palermo, known for its safety, walkability, access to restaurants, and overall pleasant atmosphere.
- **Topic 3:** recommend, stay, would, host, definitely, apartment, great, place, highly, back

  Possible Interpretation: Emphasizes strong recommendations and intentions to return, suggesting high guest satisfaction with the host and accommodation.
- **Topic 4:** check, time, easy, communication, quick, host, great, apartment, stay, helpful.

  Possible Interpretation: Relates to the ease of communication, check-in process, and helpfulness of the host, contributing to a smooth and convenient experience.
- **Topic 5:** beautiful, comfortable, clean, spacious, well, stay, really, apartment, everything

  Possible Interpretation: Highlights the physical attributes of the apartment, such as cleanliness, comfort, spaciousness, and having everything guests need.

Correlation Matrix

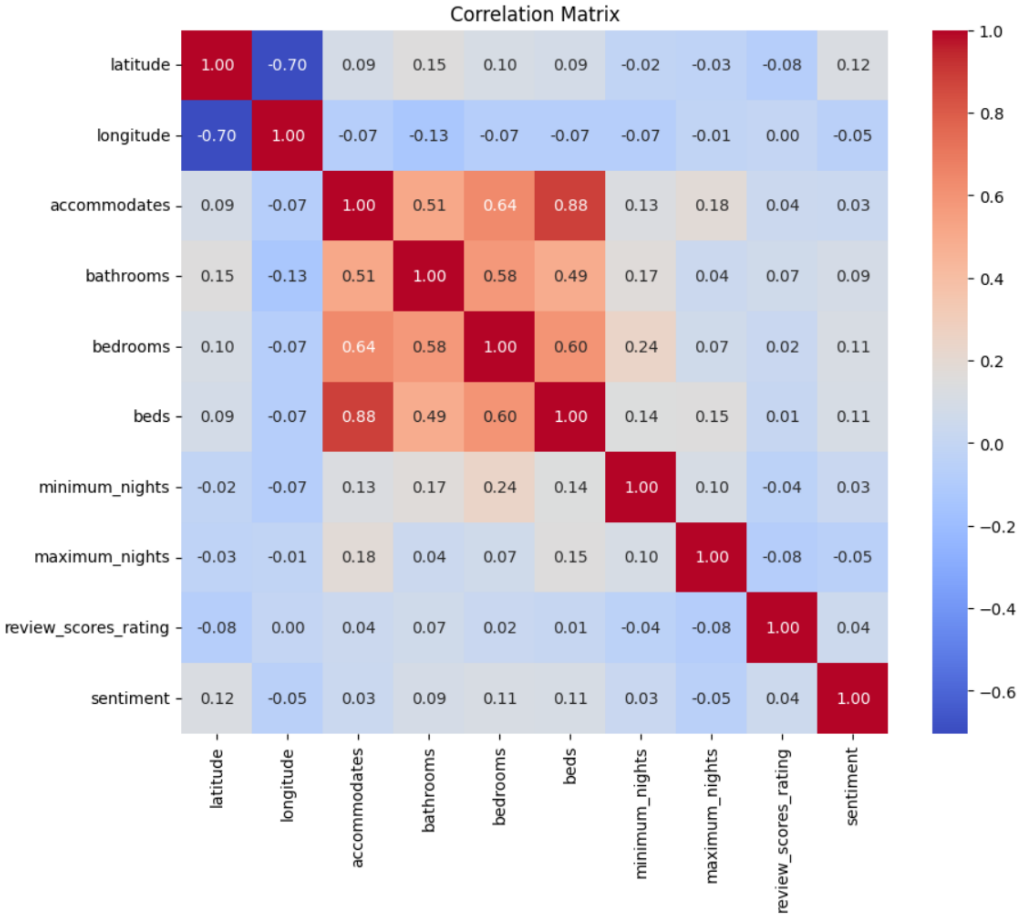


***Figure 2**: Correlation matrix before deleting columns*

- **Latitude and Longitude:**
    - Strong negative correlation (-0.69), indicating a geographic relationship (likely location-based clustering or trends).
    - *We will delete one column to deal with this (longitude).*

- **Accommodation-related Features:**
  accommodates has strong positive correlations with:
    - beds (0.87): More beds generally imply higher accommodation capacity.
    - bedrooms (0.57): Larger accommodations tend to have more bedrooms.
    - bathrooms (0.43): Accommodation capacity relates to the number of bathrooms.
    - We will delete *accommodates* to tackle this high correlation issue.
- **beds and bedrooms:**
    - Strong positive correlation (0.52): More bedrooms generally include more beds but not proportionally.
    - Since *bedrooms* had a higher value when we performed feature importance, we will keep it and delete *beds*.
- **minimum_nights and maximum_nights:**
    - Weak correlations with most features: Suggests these variables are not strongly tied to physical accommodation attributes.
- **Review Scores and Sentiment:**
    - Weak positive correlation between review_scores_rating and sentiment (0.06): Indicates a slight relationship where better sentiment might lead to higher ratings, though the effect is minimal.
- **sentiment with Location:**
    - Weak positive correlation with latitude (0.13): Sentiment may slightly vary by geographic location.

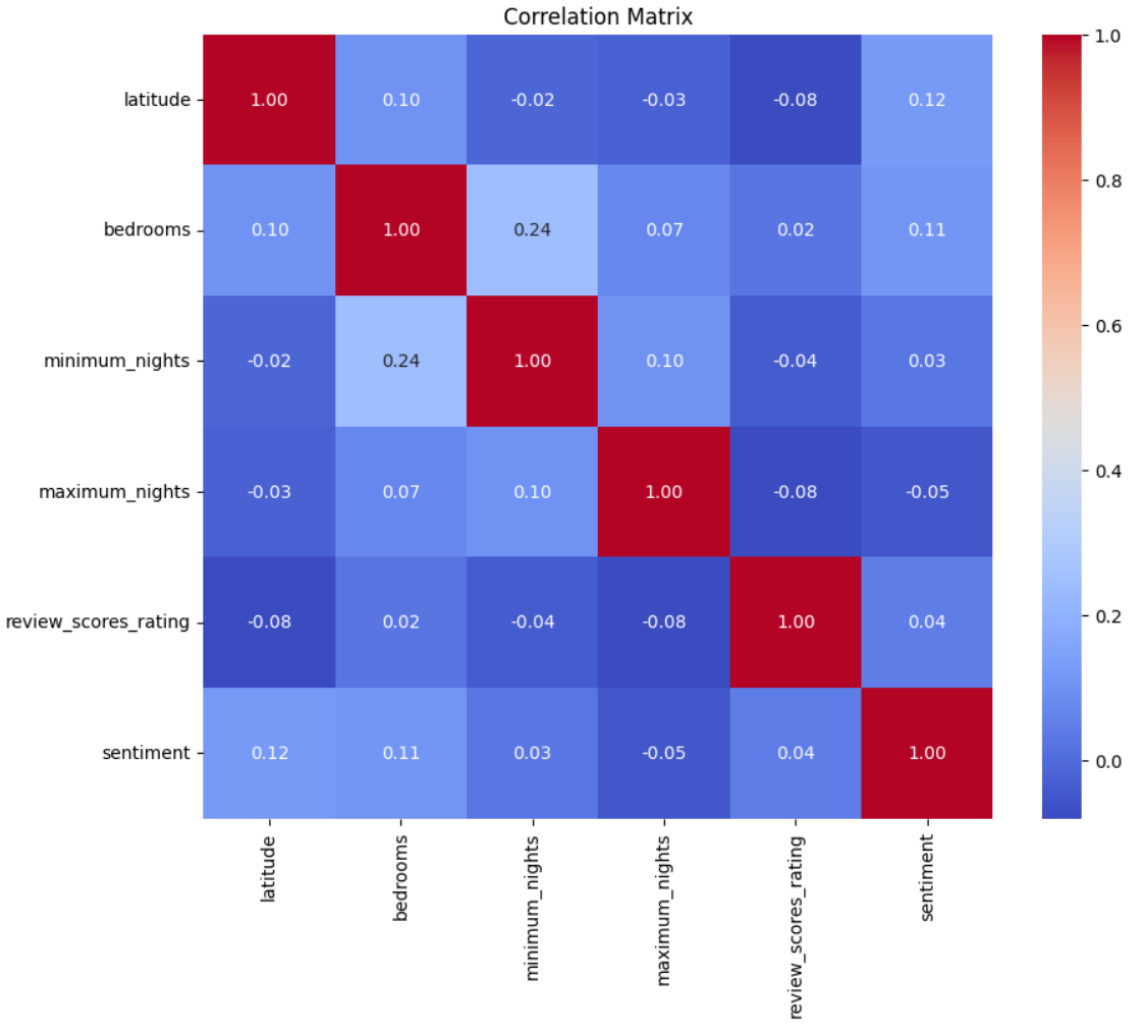


Figure 3: *Final correlation matrix*

Takeaways:

- Physical attributes of listings (e.g., beds, bedrooms, bathrooms) are strongly interrelated.
- Geographic location (latitude and longitude) significantly affects clustering but doesn't correlate strongly with other features.
- Sentiment and review scores have minimal impact on physical characteristics, suggesting they may need additional analysis for predictive modeling.

Model Analysis

Several machine learning models were evaluated to predict prices, with their performance summarized in terms of **Root Mean Squared Error (RMSE)** and **R-squared**. The results are as follows:

1. **Linear Regression**:
    a. **RMSE**: 2.4073×10^{16}, **R-squared**: -3.0591×10^{24}

    b. This model failed to capture the underlying patterns in the data, likely due to non-linearity or the presence of extreme outliers, making it unsuitable for this task.
2. **Decision Tree**:
    a. **RMSE**: 11,734.34, **R-squared**: 0.273
    b. With a shallow depth (max_depth: 5), the model underfits the data, explaining only 27.31% of the variance.
3. **Random Forest**:
    a. **RMSE**: 10,583.71, **R-squared**: 0.409
    b. This model improves over the Decision Tree, capturing more complex relationships with deeper trees (max_depth: 10) and 700 estimators. It explains 40.9% of the variance.
4. **Gradient Boosting**:
    a. **RMSE**: 10,246.25, **R-squared**: 0.446
    b. Gradient Boosting performed the best among the models, achieving the lowest error and explaining 44.6% of the variance. It effectively captures non-linear relationships in the data.
5. **XGBoost**:
    a. **RMSE**: 10,582.85, **R-squared**: 0.409
    b. XGBoost delivered performance comparable to Random Forest, with similar error metrics and variance explained. Further hyperparameter tuning may improve results.

Key Takeaways

- **Gradient Boosting** emerged as the best-performing model, with the lowest RMSE and highest R-squared, indicating its ability to effectively model non-linear patterns in the data.
- **Linear Regression** was the least effective, highlighting the need for models that can handle non-linear relationships and outliers..
- Both **Random Forest** and **XGBoost** provided reasonable performance but fell slightly short of Gradient Boosting.

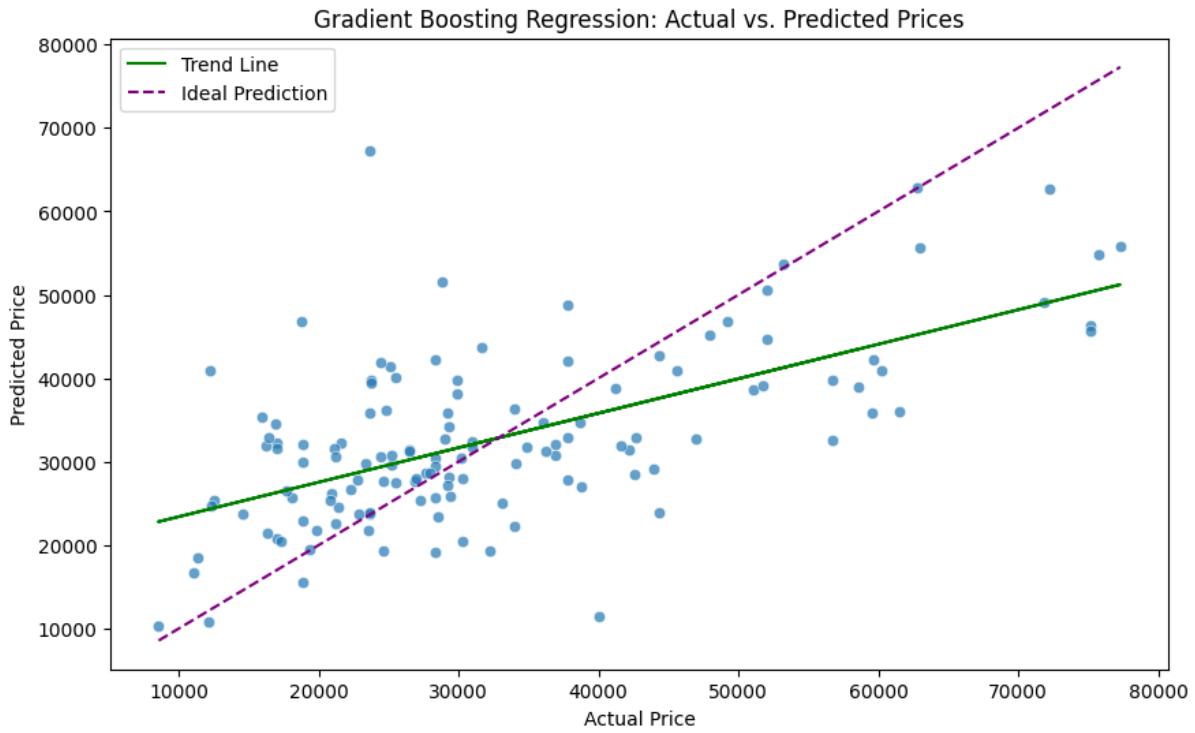


Figure 4: *Scatter plot showing how Gradient Boost Regression*

Recommendations for model training:

- **Deeper exploration:** Investigate why features like property_type and location do not contribute significantly. Consider whether the dataset lacks granularity in these features.
- **Alternative feature engineering:** Generate derived features like proximity to landmarks, density of nearby properties, or categorization of luxury vs economy properties.
- **Computing power:** The dataset is too large to be executed on google colab. This makes it difficult to work with the entire dataset. We sampled about 1000 entries and that itself took around 7-8 minutes.
- **Loss of data:** Since the data had to be sampled, we potentially lost a lot of data that could otherwise have been crucial in predicting the value of price. Sampling also introduces randomness and can provide different results every time it is run. Here are 2 images to show the differences:

	Model	RMSE	R-squared	Best Parameters
0	Linear Regression	2.407337e+16	-3.059137e+24	{}
1	Decision Tree	1.173434e+04	2.731539e-01	{'max_depth': 5}
2	Random Forest	1.058371e+04	4.087095e-01	{'max_depth': 10, 'n_estimators': 700}
3	Gradient Boosting	1.024625e+04	4.458153e-01	{'learning_rate': 0.05, 'n_estimators': 100}
4	XGBoost	1.058285e+04	4.088055e-01	{'learning_rate': 0.05, 'n_estimators': 100}

Table 2: *Model results with first random selection of data*

	Model	RMSE	R-squared	Best Parameters
0	Linear Regression	3.955562e+15	-1.214620e+23	{}
1	Decision Tree	1.135229e+04	-4.388360e-04	{'max_depth': 5}
2	Random Forest	9.555252e+03	2.912253e-01	{'max_depth': 7, 'n_estimators': 300}
3	Gradient Boosting	9.799304e+03	2.545570e-01	{'learning_rate': 0.1, 'n_estimators': 100}
4	XGBoost	9.783267e+03	2.569951e-01	{'learning_rate': 0.05, 'n_estimators': 100}

Table 3: *Model results with second random selection of data*

➔ The 2 tables show a drastic difference in the metrics values.

Conclusion

The project successfully developed a Gradient Boosting model to predict Airbnb listing prices in Buenos Aires. The model achieved an RMSE of approximately 10,246.25 and an R-squared value of 0.45, indicating moderate predictive performance. While the model effectively captures general pricing trends, there is room for improvement in capturing finer price variability. Future enhancements could include incorporating additional features such as seasonal trends, amenities, proximity to landmarks, and a few more derived features to improve accuracy. These insights will help hosts make data-driven pricing decisions, maximize occupancy, and adapt to market fluctuations more effectively.

Implications and Recommendations to the Stakeholders:

Airbnb

- **Implications:**
  - **Market Dynamics:** The model highlights the complex interplay of various factors influencing listing prices, including property features, location, and guest sentiment.
  - **Pricing Strategies:** Airbnb can leverage the model to develop more sophisticated pricing tools, enabling hosts to optimize their listings for revenue generation while maintaining competitiveness.
  - **Guest Experience**: Sentiment analysis integrated into the model offers valuable insights into guest satisfaction, allowing Airbnb to address potential issues and enhance overall platform experience.
- **Recommendations:**
  - **Develop dynamic pricing algorithms**: Implement algorithms that adjust listing prices in real-time based on market demand, competitor analysis, and seasonal trends.

- **Provide personalized host recommendations:** Offer tailored suggestions to hosts on optimizing listing features, pricing strategies, and guest communication based on the model's predictions.
- **Invest in sentiment analysis dashboards**: Create comprehensive dashboards that visualize guest feedback and highlight areas for improvement across the platform, such as property amenities, customer service, and booking processes.

Hosts

- **Implications:**
  - **Pricing Optimization:** Hosts can utilize the model's insights to set competitive and attractive prices for their listings, maximizing revenue and occupancy rates.
  - **Listing Enhancement:** By understanding the factors that significantly influence pricing and guest satisfaction, hosts can prioritize improvements to their listings, such as amenities, cleanliness, and communication responsiveness.
  - **Reputation Management:** The model emphasizes the importance of guest feedback, encouraging hosts to actively manage reviews and address any concerns promptly to maintain a positive reputation.
- **Recommendations:**
  - **Adopt data-driven pricing strategies**: Adjust listing prices based on the model's predictions, considering factors like property size, location, seasonality, and competitor analysis.
  - **Highlight key listing features**: Emphasize features that significantly impact pricing and guest satisfaction in listing descriptions and photographs, such as comfortable beds, well-equipped kitchens, and convenient locations.
  - **Engage with guest feedback:** Respond to reviews, address any issues raised by guests, and demonstrate proactive communication to build trust and enhance reputation.

Customers

- **Implications:**
  - **Price Transparency**: The model's predictions provide customers with greater transparency into pricing dynamics, enabling them to make informed booking decisions based on property value and market conditions.
  - **Personalized Recommendations:** By leveraging the model's insights, Airbnb can offer more relevant and personalized recommendations to customers, enhancing their search experience and helping them find ideal accommodations.

        - **Trust and Reliability**: The model's integration of sentiment analysis improves the reliability of guest reviews and ratings, allowing customers to make more confident booking choices.
- **Recommendations:**
    - **Utilize pricing filters and tools:** Explore the platform's features for filtering listings by price range, understanding price drivers, and comparing options effectively.
    - **Provide detailed preferences:** Share specific preferences and needs with Airbnb to receive more personalized recommendations and tailored search results.
    - **Rely on verified reviews and ratings:** Make booking decisions based on authentic guest feedback and ratings to ensure a positive and reliable experience.

By leveraging data insights, all stakeholders can maximize satisfaction, profitability, and operational efficiency.

References:

- Gemini for help with coding.
- Stackoverflow with code debugging along the way.
- We also referred to the code provided in our course to build the basic framework.