Analysis of Airbnb Listings in Cape Town: Pricing, Availability, and Review Insights

Name: ABC

Student ID: 001

Date: 04-Dec-2024

Abstract

This report presents an analysis of the Cape Town Airbnb dataset, focusing on key factors that influence the pricing, availability, and performance of Airbnb listings. The dataset includes detailed information on over 23,000 listings, such as property types, room types, pricing, host information, review scores, and availability. The analysis aims to identify trends, patterns, and anomalies in the data through exploratory data analysis (EDA), including data cleaning, visualization, and statistical techniques. Key findings reveal correlations between room type and price, the positive impact of high review scores, and the influence of the number of bedrooms on accommodation capacity. The report also proposes several research questions for further exploration, including the effect of property type on pricing and the relationship between review scores and listing success. This analysis provides valuable insights that can inform pricing strategies and enhance the guest experience on the Airbnb platform in Cape Town.

Table of Contents

1. Introduction

The goal of this assignment is to analyze a dataset of **Airbnb listings in Cape Town** to uncover insights related to the pricing, availability, and performance of these listings, as well as the factors influencing their success. The dataset contains detailed information on Airbnb listings, including listing details, pricing, availability, and host information, as well as review data, including review scores and the number of reviews.

The objectives of this analysis are:

1. **Explore and clean the dataset**: This involves loading the dataset, identifying and handling missing values, removing duplicates, and ensuring that the data is in the correct format for analysis.
2. **Conduct exploratory data analysis (EDA)**: We will analyze key variables such as price, number of reviews, review scores, room type, and location. Visualizations like histograms, scatter plots, and correlation heatmaps will help us understand the distribution of these variables and their relationships.
3. **Identify trends and patterns**: We will investigate patterns, trends, and anomalies within the data, such as the correlation between pricing and room type or the relationship between review scores and the number of reviews.
4. **Propose research questions**: Based on the data exploration, we will propose several research questions that can be explored further to gain deeper insights into the Airbnb listings and their pricing dynamics in Cape Town.

In this report, the analysis will focus on summarizing key findings and providing actionable insights for potential areas of improvement for Airbnb hosts, pricing strategies, and customer experience. The approach will be methodical and iterative, with a focus on data cleaning, visualization, and statistical analysis to uncover meaningful patterns and trends.

2. Dataset Overview

The dataset used in this analysis is sourced from **Airbnb**, focusing on listings in **Cape Town**. The dataset consists of two main components:

- **Listing Information**: Includes details such as listing name, description, host information, property type, room type, pricing, minimum/maximum nights, availability, amenities, and location (latitude and longitude).
- **Review Information**: Contains review-related details such as the number of reviews, review scores, review ratings for aspects like cleanliness, check-in, communication, and location, along with the review dates.

Together, these two datasets provide a complete picture of Airbnb's listing ecosystem in Cape Town, offering insights into property details, pricing patterns, and customer feedback.

3. Methodology

3.1 Data Loading and Initial Exploration

- **Data Loading**: The datasets (`df1` for listing details and `df2` for review information) were successfully loaded using **pandas**' `read_csv` function.
    - `df1`: Contains 62 columns and 23,564 rows.
    - `df2`: Contains 21 columns and 23,564 rows.
- **Initial Exploration**: After loading the data, an initial exploration was performed by inspecting the first few rows of each dataset to understand the structure, column names, and types of data. This helped identify that both datasets share a common column, `id`, which was used for merging them. Key observations included:
    - Dataset 1 (`df1`) contains detailed information about listings, while Dataset 2 (`df2`) contains review-related data.
    - Both datasets have a significant number of columns, which were explored to assess relevance and data quality.

3.2 Data Cleaning and Wrangling

The data cleaning process focused on ensuring that the data was ready for analysis by handling missing values, removing duplicates, and correcting data types.

- **Variable Selection**: A subset of relevant variables was selected to focus on key features, such as:
    - Listing Details: `name`, `description`, `price`, `neighborhood_overview`, `property_type`, `room_type`, etc.
    - Host Information: `host_name`, `host_since`, `host_location`, `host_is_superhost`, etc.
    - Review Information: `review_scores_rating`, `number_of_reviews`, etc.
- Columns with excessive missing data or that were not useful for the analysis were excluded, such as `host_neighbourhood` and `calendar_updated`.

- **Handling Missing Data**:
    - Columns with more than 50% missing data were dropped to maintain a clean dataset.
    - Remaining missing values were handled by dropping rows with missing values in the selected variables.
    - For the `price` column, the data was cleaned by removing non-numeric characters and converting the values to a numeric format.
- **Duplicates and Data Types**:

  - Duplicates were removed to avoid redundant data.
  - The data type for `price` was corrected to a numeric format, and `host_since` was converted to `datetime`.
- **Text Standardization and Feature Creation**:
  - Standardized text columns, such as converting `neighborhood_overview` to lowercase, to ensure consistency across text fields.
  - No new features were created during the cleaning process.

3.3 Exploratory Data Analysis (EDA)

In this phase, key variables were analyzed to understand their distribution, relationships, and to identify patterns within the dataset.

- **Summary Statistics**: Key numeric columns were analyzed, including `price`, `review_scores_rating`, `accommodates`, `bedrooms`, and `number_of_reviews`. The summary statistics revealed:
  - The average `price` was relatively high, with some outliers in the data.
  - `review_scores_rating` showed a high concentration of listings with positive reviews.
- **Distribution Analysis**: The distribution of several variables was visualized:
  - **Price Distribution**: The price distribution showed a skew, with some listings having very high prices, which may be outliers. These high-priced listings require further investigation to confirm if they are genuine or erroneous data points.
  - **Review Scores Distribution**: The majority of listings had positive review scores, with few extreme low ratings.
  - **Number of Reviews**: Listings had a wide range of reviews, with most having fewer than 100 reviews, and some having over 200.
- **Relationships Between Variables**: Key relationships between variables were analyzed:
  - **Price vs. Room Type**: A box plot showed that `entire homes` typically had higher prices than `private rooms` and `shared rooms`.
  - **Accommodates vs. Bedrooms**: A scatter plot revealed a positive correlation between the number of bedrooms and the accommodation capacity, indicating that larger properties tend to accommodate more guests.
  - **Beds vs. Price**: A scatter plot showed that more beds tended to correlate with higher prices.
- **Correlation Analysis**:

  - A correlation matrix heatmap revealed that variables like `price`, `bedrooms`, `accommodates`, and `beds` were positively correlated. However, variables like `latitude` and `longitude` showed minimal correlation with other features.

4. Proposed Research Questions

Based on the insights from the dataset, the following research questions can be explored further:

1. How does the property type affect the price of a listing?

- **Variables**: `price`, `property_type`
- **Analysis Type**: Group comparison using box plots and statistical tests (e.g., ANOVA or Kruskal-Wallis) to compare the average prices across different property types.

2. What is the relationship between the number of reviews and the review scores for listings?

- **Variables**: `number_of_reviews`, `review_scores_rating`
- **Analysis Type**: Correlation analysis and scatter plots to examine the relationship between review quantity and overall ratings.

3. Can we predict the price of a listing based on the number of bedrooms, bathrooms, and its location?

- **Variables**: `price`, `bedrooms`, `bathrooms`, `latitude`, `longitude`
- **Analysis Type**: Predictive modeling using linear regression or machine learning algorithms (e.g., Random Forest or Gradient Boosting) to predict price based on these features.

5. Findings and Conclusions

5.1 Main Findings

1. **Price Distribution**:

    - The price distribution was highly skewed, with a significant number of listings priced moderately. However, there were several listings with extremely high prices, which could be outliers or unique premium listings (e.g., luxury properties). These outliers might skew any analysis that involves pricing, and they warrant further investigation to ensure they represent legitimate data points.

2. **Review Scores**:

    - Most listings had positive review scores, indicating that the majority of Airbnb hosts in Cape Town provide a satisfactory experience for guests. However, a few listings had notably low ratings. This trend highlights the importance of ensuring quality control, as reviews and ratings can have a direct impact on future bookings and overall success on the platform.

3. **Room Type and Price**:
    - A clear pattern was observed in the relationship between **room type** and **price**. Listings for entire homes had significantly higher prices compared to private or shared rooms. This is expected, as entire homes offer more privacy and space, which typically translates to higher costs. This insight can help inform pricing strategies for hosts depending on their offering type.
4. **Accommodates vs. Bedrooms**:
    - The correlation between the number of **bedrooms** and **accommodates** was positive, which is logical as larger properties tend to accommodate more guests. This insight aligns with general expectations about the relationship between the number of bedrooms and the accommodation capacity. It suggests that hosts can potentially charge higher prices based on the number of guests they can accommodate.
5. **Beds vs. Price**:
    - A positive relationship between **beds** and **price** was observed, indicating that properties with more beds tend to have higher prices. This finding reinforces the notion that larger properties are generally priced higher, likely because they can accommodate more guests and offer more amenities.
6. **Correlation Between Location and Price**:
    - The **latitude** and **longitude** of the listings showed minimal correlation with other features like price or review scores, which suggests that pricing in Cape Town may not be strongly driven by geographic location alone. However, this might be worth exploring further by segmenting the data into different neighborhoods or areas.
7. **Missing Data**:
    - Some variables, particularly those related to host details (e.g., `host_location` and `host_about`), had significant amounts of missing data. This could affect analyses that depend on these features. The missing data for certain columns, especially the `license` and `host_is_superhost` columns, was mostly handled by dropping rows with missing values. This cleaning process ensured that the dataset remained useful for analysis without unnecessary noise.

5.2 Trends and Patterns

- **Luxury Listings**: A few listings with exceptionally high prices likely represent premium properties. These may have unique features such as luxury amenities, large spaces, or desirable locations (e.g., beachfront properties). Further segmentation of these listings could help understand the factors contributing to their higher prices.

- **Host Experience**: Listings by hosts with **Superhost** status tend to have better ratings and higher prices. This highlights the role of host reliability and reputation in shaping the Airbnb market, where highly-rated hosts may attract more bookings and justify higher prices.

- **Review Scores**: The high concentration of positive review scores suggests that most guests are satisfied with their stays. However, the existence of low review scores may point to inconsistencies in quality across properties. A deeper dive into these properties might reveal common issues such as cleanliness, accuracy of listings, or customer service.

5.3 Suggestions for Further Analysis

1. **Outlier Investigation**:

    - A deeper exploration of the high-priced listings could help identify whether these outliers are legitimate (e.g., luxury properties or unique offerings) or if there are data entry errors. Identifying the true nature of these high-priced listings could significantly improve the analysis of pricing trends.

2. **Geographic Analysis**:

    - Since **latitude** and **longitude** showed minimal correlation with other variables, a more detailed geographic analysis could help understand how location affects pricing. It might be useful to map listings by neighborhood or proximity to key landmarks (e.g., beaches, parks, tourist attractions) to see if location plays a stronger role than initially observed.

3. **Review Sentiment Analysis**:

    - Given the importance of guest reviews, sentiment analysis could be applied to the **description** and **neighborhood_overview** columns to understand the key factors driving positive or negative feedback. This analysis would provide insights into what guests value the most in their stays.

4. **Predictive Modeling**:

    - Although no formal predictive modeling was conducted in this exploration, using machine learning models (e.g., linear regression, random forests) to predict **price** based on variables like **accommodates**, **bedrooms**, **room_type**, and **review_scores_rating** could help identify key price drivers and develop a pricing model.

5. **Feature Engineering**:

- Further feature engineering could be valuable, especially in terms of creating new features from existing ones (e.g., calculating the average number of reviews per month or creating a "luxury" binary variable based on price). This would enrich the dataset and could help improve modeling accuracy in future analyses.

In conclusion, the analysis provided valuable insights into the characteristics and pricing dynamics of Airbnb listings in Cape Town. The findings revealed clear trends regarding room type, accommodation size, pricing, and review scores, while also highlighting some anomalies, such as high-priced listings that may need further investigation. Further analysis could explore the impact of location, host experience, and guest reviews in more depth, and predictive models could be employed to better understand the factors influencing pricing and review scores.

6. References

1. **Inside Airbnb Project**:
   The dataset used in this analysis is sourced from the **Inside Airbnb** project, which provides detailed and publicly accessible data on Airbnb listings across various cities worldwide. For more information on the project, visit the Inside Airbnb website.

2. **Python Libraries**:

   - **pandas**: A powerful library for data manipulation and analysis, used for loading, cleaning, and exploring the datasets.
     Documentation: https://pandas.pydata.org/

   - **NumPy**: A fundamental package for numerical computing in Python, used for handling arrays and numerical operations.
     Documentation: https://numpy.org/

   - **Matplotlib**: A plotting library for creating static, animated, and interactive visualizations in Python. Used for creating plots to visualize distributions and relationships.
     Documentation: https://matplotlib.org/

   - **Seaborn**: A Python visualization library based on Matplotlib that provides a high-level interface for drawing attractive and informative statistical graphics. Used for creating more advanced plots, such as histograms, box plots, and scatter plots.
     Documentation: https://seaborn.pydata.org/

   - **SciPy**: A library for scientific and technical computing, often used for statistical analysis and hypothesis testing.
     Documentation: https://scipy.org/

3. **pandas_profiling**: A tool used for quick exploratory data analysis and generating summary reports of datasets (if applicable).
   Documentation: https://pandas-profiling.github.io/pandas-profiling/

These sources and libraries played a crucial role in the data cleaning, wrangling, and exploratory analysis process, providing the necessary tools to manipulate and visualize the data effectively.