Report ETL Project

Presented by:

Juan Jose Rendon

Juan Camilo Burbano

Teacher:

Javier Alejandro Vergara Zorrila

Universidad Autónoma De Occidente

ELT
08/03/2024

Introduction

In this project, undertaken by a team of data engineers, we selected a dataset through the Kaggle platform for analysis and interpretation. The project was executed using Python and several of its libraries. Additionally, a MySQL database was created to manage all the gathered information. The visualization of our findings will be showcased through dashboards created in Power BI.

Our focus for this project is the real estate sector, specifically analyzing an Airbnb dataset available on Kaggle. This dataset was split into multiple CSV files, which we merged to create a comprehensive database.

This project aims to benefit individuals looking to list their properties on Airbnb within California.

Technologies Used

The project leveraged the following technologies:

Python: The core programming language of the project.

Visual Studio Code: The chosen code editor for project management and development.

MySQL: The database management system used for storing project data.

Power BI: Used for creating dashboards to visualize our findings.

Kaggle: This platform provided the initial data source and is valuable for data-related projects.

Project Architecture

Our project's architecture is organized into three simple and practical folders containing essential documents:

config: Contains the requirements file for library installation and credentials for database connection.

data: Houses the CSV files used to create our final dataset, which are then analyzed and visualized and the documentation files.

notebooks: Contains two Jupyter notebooks. One for merging and cleaning the dataset, and another for transferring the data to a database.

Implementation

The implementation process began with running the project_eda.ipynb notebook in the notebooks folder. This notebook handles the creation and cleaning of the final dataset and performs an Exploratory Data Analysis (EDA).

Data Selection

We encountered many problems when selecting the files to be used. After analyzing them one by one, we opted to perform our analysis by combining the market analysis and amenities files, which have a better structure compared to the other CSVs, had more than the requested number

We also realized that there is no data dictionary, so we decided to conduct research based on the context of our data and create it, which will be specified later.

In turn, we chose not to use other CSVs because their columns did not provide any relevant or necessary information. The data structure or the way they were stored was incorrect or nonexistent. Moreover, many CSVs (especially the property sales) only had a range of 10 to 80 records or rows, which would not help us compared to the 49,000 or 29,000 records of the already selected files. of rows, and provided us with more valuable information.

After the selection we started by importing essential libraries and then merged the selected CSV files related to amenities, and market analysis. This merge highlighted a challenge with null data due to varying data amounts across CSVs.

The cleaning process involved eliminating irrelevant columns, deletion and imputation of null data. A detailed cleaning was performed on the remaining columns in our final CSV.

The decision is made to use the following variables:

- ***hot_tub**, pool: Evaluate the impact of these amenities on the property's appeal, occupancy, and nightly rate.*
- ***city**: Analyze property performance based on geographical location and detect regional differences.*
- ***host_type**: Understand how the type of host affects revenue, occupancy, and guest satisfaction.*
- ***bedrooms, bathrooms, guests**: Determine the influence of property size and capacity on demand and price.*
- ***revenue**: Directly measure the financial success of properties and identify factors that maximize revenue.*
- ***openness:** (If referring to availability or promotion) Analyze how exposure affects occupancy and revenue.*
- ***occupancy**: Identify demand patterns and evaluate the effectiveness of pricing or promotional strategies.*
- ***nightly_rate**: Understand pricing strategy, its impact on occupancy, and guests' perception of value.*
- ***lead_time**: Explore guest booking behavior and its impact on property management.*
- ***length_stay**: Analyze guest preferences for stay duration and their influence on revenue and management.*

The rest will be eliminated as they are not considered relevant for the analysis. This process ensured data consistency across the board.

This is the final data dictionary for our dataset, providing a foundation for brainstorming potential analyses and insights:

1. ***hot_tub**: Indicates whether the property has a hot tub. A numeric value (usually 0 or 1, where 1 means the property includes a hot tub and 0 means it does not. This could influence guest preferences and potentially the nightly rate.*
2. ***pool**: Indicates whether the property has a pool. Similar to the hot tub, a value of 1 indicates the presence of a pool, and a 0 indicates its absence. Having a pool might be a deciding factor for some guests.*
3. ***city**: The city where the property is located. This categorical variable helps analyze demand and pricing across different locations.*
4. ***host_type**: The type of host offering the property, such as individual hosts or property management companies. This can influence the guest experience and property policies.*
5. ***bedrooms**: Number of bedrooms in the property. This is an important measure of the property's capacity and the type of groups it can accommodate.*
6. ***bathrooms**: Number of available bathrooms. Like bedrooms, the number of bathrooms is a critical factor in guest comfort and convenience.*
7. ***guests**: Maximum number of guests the property can accommodate. This is crucial for understanding the property's target market (couples, families, large groups, etc.).*

8. ***revenue****: Revenue generated by the property over a specific period. This variable is key to assessing the financial performance of the property investment.*
9. ***openness****: Likely refers to the property's booking availability or willingness to accept reservations. It might indicate how often the property is available for rent.*
10. ***occupancy****: The property's occupancy rate. It shows the percentage of time the property was rented compared to available time.*
11. ***nightly_rate****: Nightly price for the property. This is a critical factor affecting both the property's demand and the revenue it generates.*
12. ***lead_time****: The lead time for bookings. It can provide insights into travel planning and seasonal or specific event demand.*
13. ***length_stay****: The duration of guest stays. Important for understanding the type of travel (short-term, long vacations, etc.) and managing bookings and availability.*

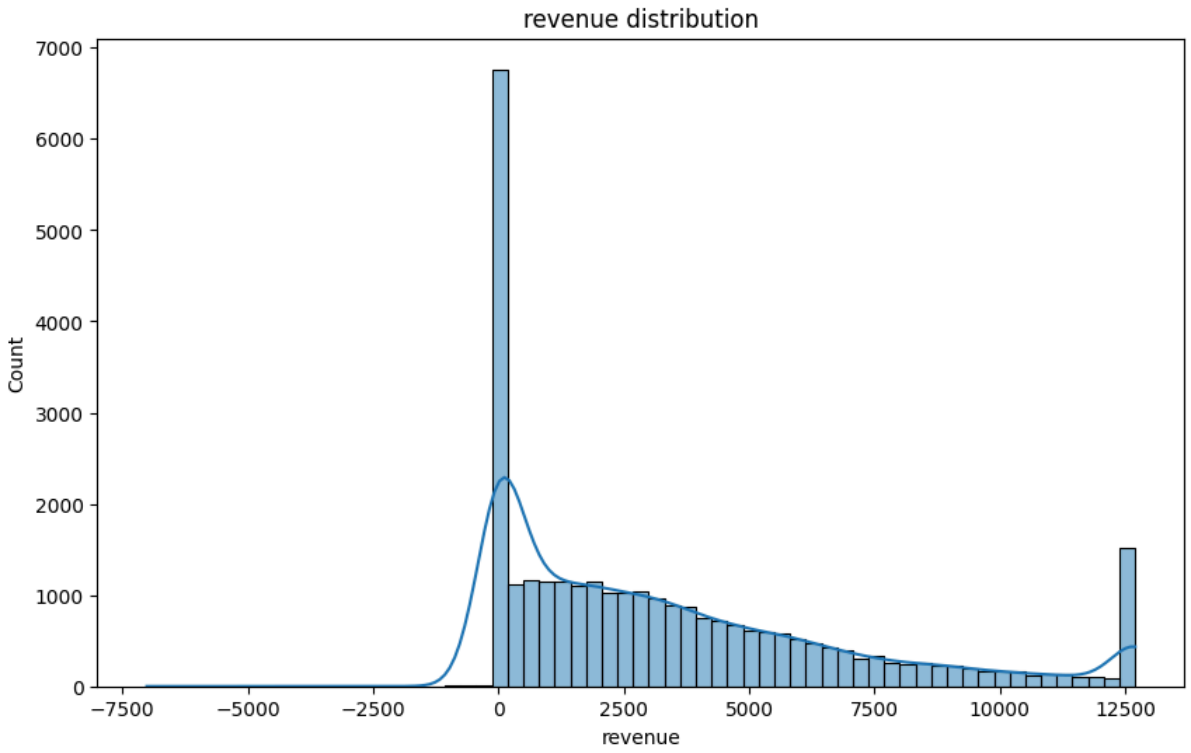
Following the data cleaning, we conducted an EDA to draw significant insights and conclusions.

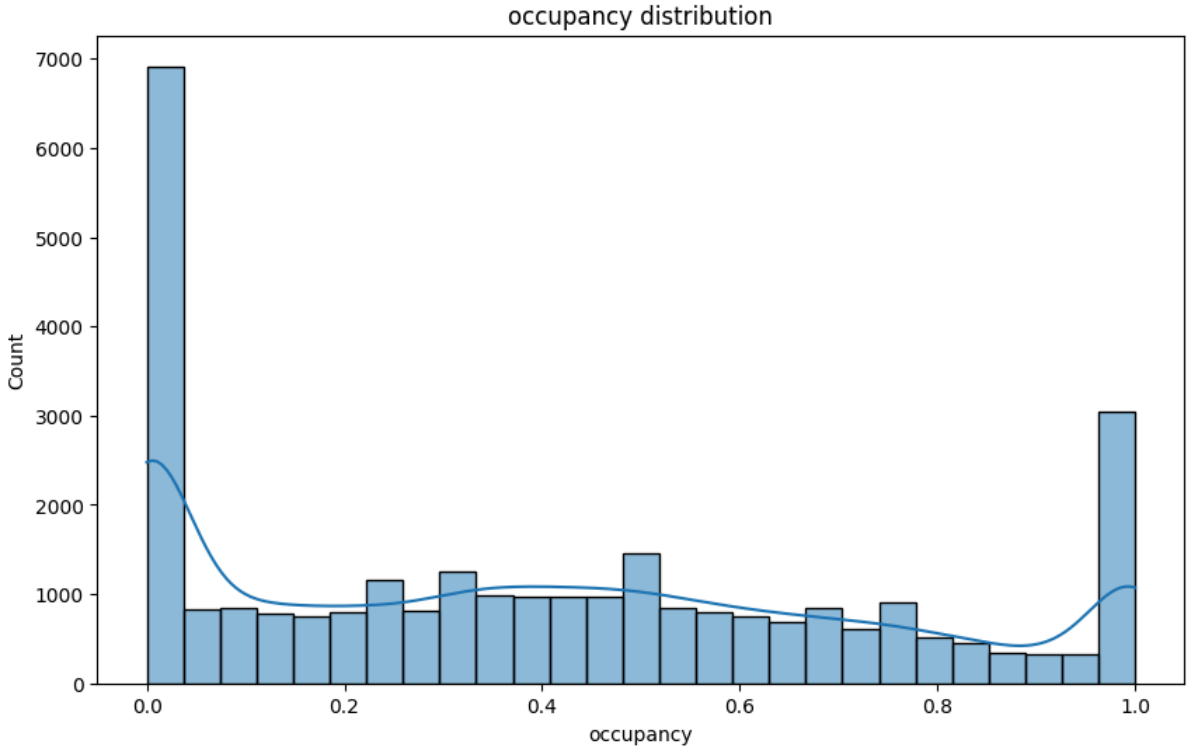
After completing the EDA, we established a connection to our MySQL database to upload the cleaned dataset.

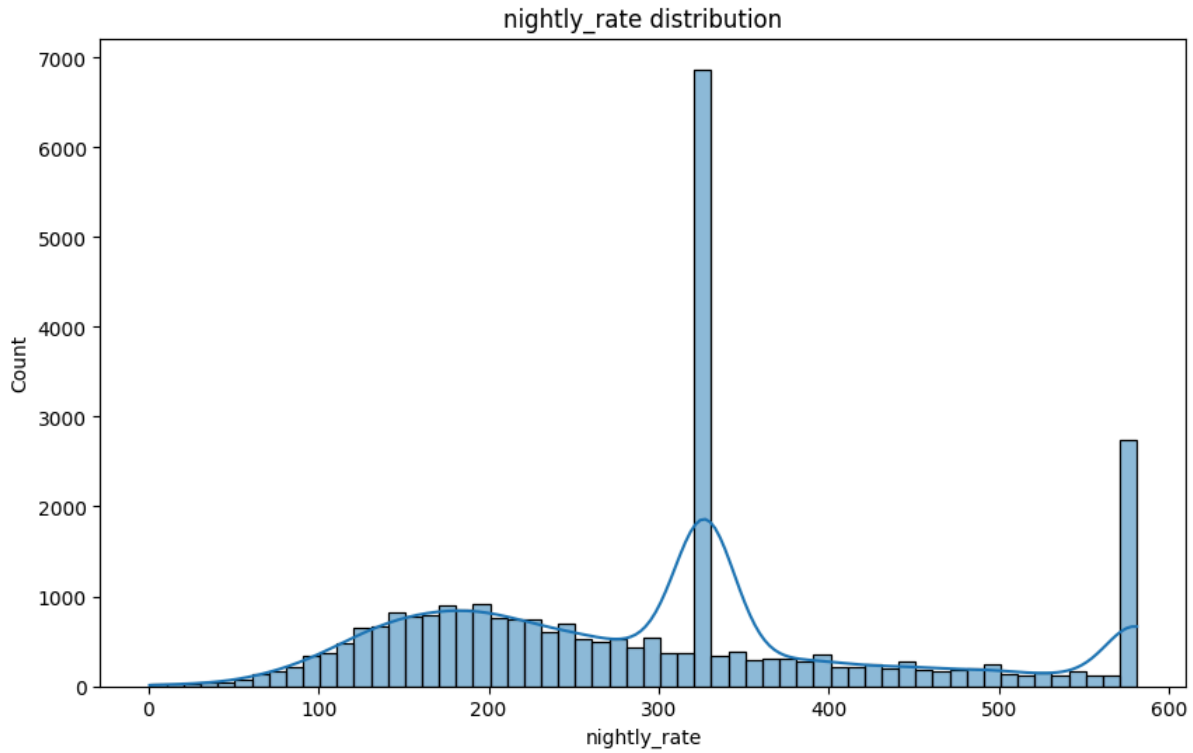
Next, we ran the connection.ipynb notebook. This involved installing necessary libraries, connecting to the database, creating a database named etl, and inserting the cleaned data for dashboard creation.

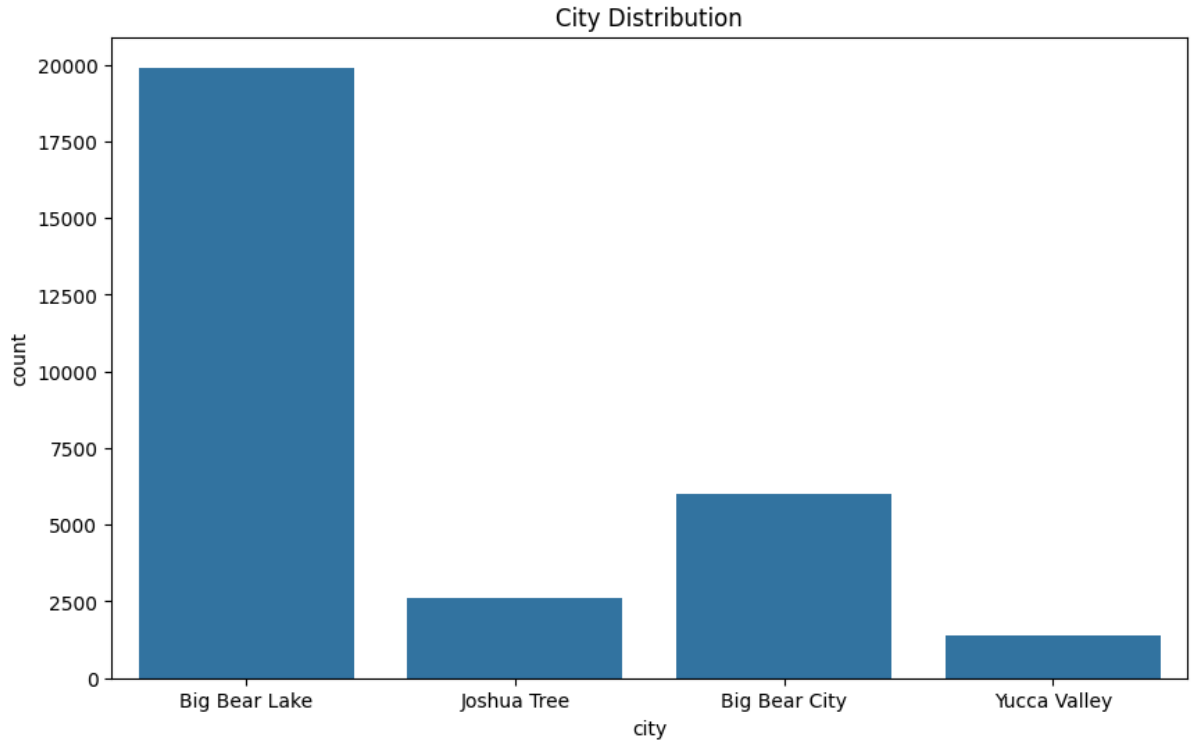
Exploratory Data Analysis (EDA)

Based on our previous selection, analysis, and data handling, we proceeded to conduct an exploratory study which provided us with valuable information. This allowed us to make informed decisions and better illustrate these in a dashboard for an enhanced visualization.


revenue distribution
Count
revenue
7000
6000
5000
4000
3000
2000
1000
0
−7500
−5000
−2500
0
2500
5000
7500
10000
12500


occupancy distribution
Count
occupancy
7000
6000
5000
4000
3000
2000
1000
0
0.0
0.2
0.4
0.6
0.8
1.0





Based on these graphs, the following interpretations can be made:

- Revenue:

Distribution: There is a prominent vertical bar at the value 0 on the X-axis, indicating a high number of transactions that were either canceled or stored.

Positive Revenue: To the right of value 0, there are smaller bars showing a positive distribution of revenue, representing sales that were completed.

Negative Revenue: To the left of value 0, the values represent sales that were stored and then canceled.

- Occupancy:

  Distribution: There are two prominent peaks in the graph at the occupancy values of 0.0 and 1.0, indicating that most observations fall at these two values.

  Peaks: There is a significant peak in the count when the occupancy is 0.0, with more than 6000 counts. Another notable peak occurs when the occupancy is 1.0, though this peak is much smaller than the first.

  Intermediate Values: Between these two peaks, the count significantly decreases and fluctuates between approximately 100 and just over 1000 counts for intermediate occupancy values.

- Nightly Rate:

  Distribution: The bars represent the frequency of each nightly rate. The bars are more prominent in specific ranges of nightly rate, especially around 300 and 500. This indicates that most nightly rates fall within these ranges.

  Peaks: The peaks in the graph represent the most common values of nightly rates. A higher peak means this nightly rate value is more common in the data.

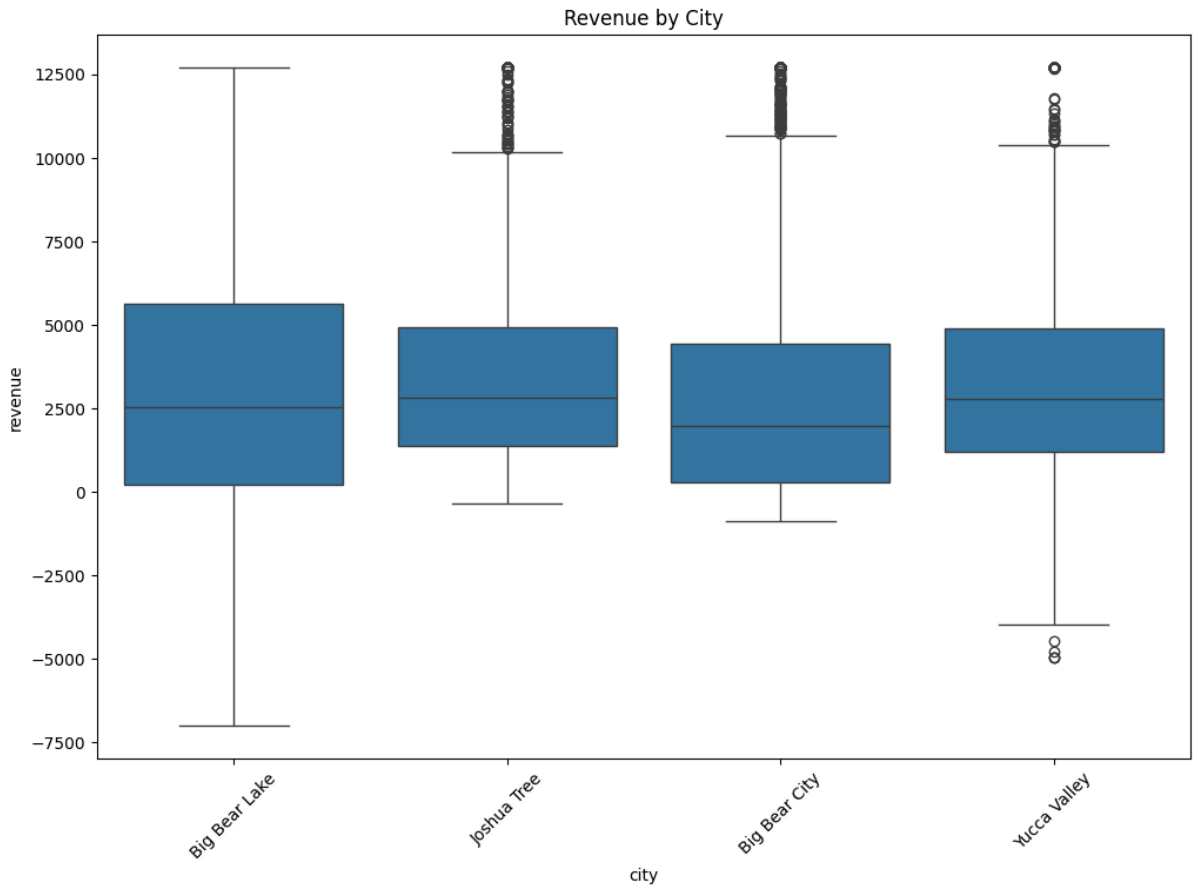
- City:

  Distribution: The bars represent the frequency of each city.

  Big Bear Lake has the highest frequency, reaching close to 20,000.

  Joshua Tree and Yucca Valley have the lowest frequencies, just above and below 2,500 respectively.

  Big Bear City has an intermediate frequency around 7,500.

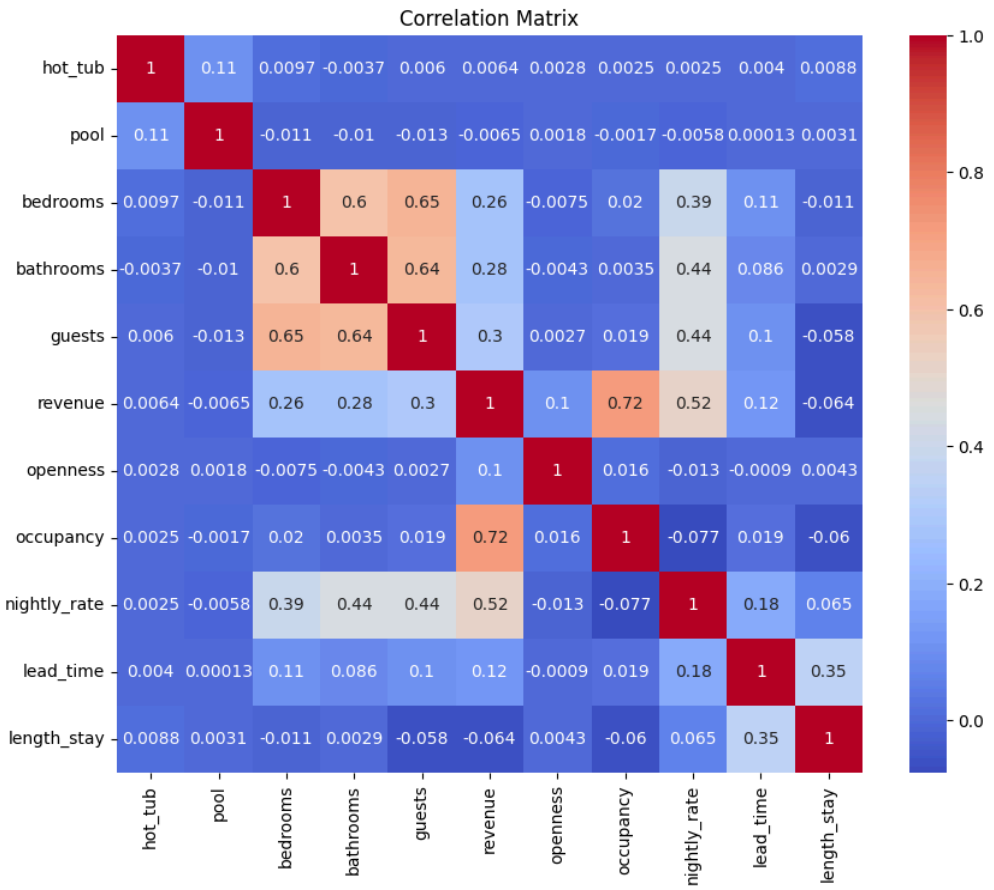


- Boxes: Each box in the graph represents the interquartile range of revenue for each city, which is the range within which the central half of the data lies. The line within each box indicates the median revenue, which is the middle value of the data.

- Whiskers: The whiskers extending from each box show the range of revenue within 1.5 times the interquartile range. This provides an indication of the variability of revenue in each city.

- Circles: The small circles above and below the whiskers represent outlier values in revenue for each city. These are values that fall outside the typical range of revenue.

- General Interpretation: The median revenue for Big Bear Lake is higher than the other cities, while Lucerne Valley has a wider interquartile range, indicating greater variability in revenue. There are also several outlier values in the data for each city.



- Correlation Coefficients: The numbers in the matrix represent the correlation coefficients between the variables. A correlation coefficient ranges between -1 and 1. A value close to 1 indicates a strong positive correlation, while a value close to -1 indicates a strong negative correlation.

- The variable "bedrooms" and "bathrooms" seem to have a strong positive correlation, meaning that as the number of bedrooms increases, the number of bathrooms also tends to increase.

- Guests and Bedrooms: These two variables also seem to have a positive correlation. This could indicate that properties with more rooms tend to accommodate more guests.

- Revenue and Bedrooms: These variables seem to have a positive correlation. This could suggest that properties with more rooms generate more income.

Key Findings

Impact of Amenities on Occupancy Rate and Financial Performance: Properties with certain amenities, such as hot tubs and pools, tend to show a higher occupancy rate and can command higher nightly rates. This suggests that investing in property improvements could be an effective strategy to enhance financial performance.

Influence of Location on Demand and Revenue: The property's location (city) significantly impacts its performance. Properties in areas like Big Bear Lake and Palm Springs exhibit higher demand and revenue potential, highlighting the importance of location when considering investments in vacation rental properties.

Relationship Between Property Size and Performance: There is a positive correlation between the property size (number of bedrooms and bathrooms) and financial performance, indicating that larger properties can attract broader groups and generate higher revenues.

Stay Preferences and Booking Behavior: Variables such as booking lead time and stay duration offer insights into guests' booking behaviors. Properties that accommodate various stay duration preferences can attract a wider range of guests.

Effectiveness of Pricing Strategies: Analyzing the nightly rate alongside occupancy rate provides valuable insights into the effectiveness of pricing strategies. Adjusting prices based on demand and season can optimize revenue and occupancy rates.

Strategic Recommendations

Invest in Amenities: Consider investing in significant improvements and amenities (like pools and hot tubs) that increase property appeal.

Focus on Key Locations: Assess the location of current properties and future investments, prioritizing areas with high demand and financial performance.

Optimize Property Size: Tailor the offering to market segments looking for larger properties to accommodate families or groups.

Dynamic Pricing Strategies: Implement dynamic pricing strategies based on demand analysis, seasonality, and local events to maximize revenue.

Continuous Analysis: Continue with regular market analysis and adjust strategies based on emerging trends and performance data.

General conclusion

This project on the analysis of Airbnb properties in California represents an exhaustive effort to understand the determining factors behind the success of listings in the vacation rental space. By overcoming obstacles related to data integrity and strategic file selection for analysis, we have managed to identify essential patterns and correlations that influence the profitability and demand for these properties. Through a careful examination of key variables such as revenue, occupancy, and the influence of amenities and location, this study has revealed important insights for optimizing listing strategies on the Airbnb platform.