IST 652

SCRIPTING FOR DATA ANALYSIS

PROJECT REPORT

Enhancing Host Strategy Using Data Analysis on Airbnb

"Optimizing Guest Experience and Revenue Through Insights and Analytics"

Presented By

Amith Joseph
Master of Science in Business Analytics
University of Louisville

This project presents an in-depth analysis of Airbnb hosting strategies, focusing on data-driven decision-making to enhance profitability and guest experience. Through statistical analysis, visualization techniques, and structured data preprocessing, key insights have been extracted to provide actionable recommendations for hosts and investors.

DATA DESCRIPTION:

We used publicly available datasets from **Inside Airbnb** (https://insideairbnb.com/get-the-data/), which provides detailed data about Airbnb properties for cities around the world. Our focus was on **San Diego, California**, and we used three main datasets: **Listings**, **Calendar**, and **Reviews**. Each dataset provided valuable information that helped us analyze Airbnb hosting strategies.

1. Listings Dataset

The Listings dataset contains detailed information about each Airbnb property. This includes:

- **Basic property details** like the number of bedrooms, bathrooms, and amenities (e.g., Wifi, Pool, Kitchen).
- **Host information** such as response rates, whether the host is a "superhost," and how long the host has been active.
- **Pricing details** showing the cost per night for each property.

This dataset helped us understand the features that properties offer and how these features might influence guest ratings, prices, and popularity.

2. Calendar Dataset

The Calendar dataset provides daily availability and pricing information for each listing over a period of time. It includes:

- **Daily prices** for each property.
- **Availability status** for each day (whether a property was available or booked).
- **Minimum and maximum stay requirements** set by the host.

This dataset allowed us to study how prices change over time, such as during peak seasons or special events, and how often properties are booked. It also helped analyze how availability patterns differ between neighborhoods or property types.

3. Reviews Dataset

The Reviews dataset includes feedback from guests who stayed at the properties. It contains:

- **Review dates** to track guest experiences over time.
- **Guest details** like reviewer names.
- **Guest comments** describing their experiences and opinions about the property.

This dataset helped us explore guest satisfaction by analyzing review scores and identifying common themes in feedback. It also provided insights into what guests value most during their stay.

Data Utilization:

By combining these datasets, we created a comprehensive view of Airbnb properties in San Diego. This allowed us to:

1. **Analyze amenities and their impact** on guest ratings and pricing. For example, we looked at whether properties with amenities like pools or Wi-Fi receive better ratings or higher prices.
2. **Study pricing trends** across different times of the year (e.g., seasonal changes) and neighbourhoods to identify factors influencing prices.
3. **Understand guest preferences** by analysing reviews to uncover what guests appreciate most and what hosts can improve.

Data Importance:

These datasets work together to provide a full picture of the Airbnb market in San Diego. The Listings dataset gives us property and host details, the Calendar dataset shows trends in availability and pricing, and the Reviews dataset captures guest feedback. This combination allows us to offer practical advice to Airbnb hosts, such as how to improve guest satisfaction, optimize prices, and attract more bookings. It also helps investors identify profitable areas and understand what drives demand in the Airbnb market.

DATA PREPROCESSING:

The following steps were taken for data preprocessing:

1. Simplifying the Dataset

To make the dataset easier to work with, we removed columns that weren't relevant to our analysis, such as URLs, metadata, and other unnecessary details. This reduced clutter and helped focus on the most important information. We also renamed columns by adding prefixes like listings_ or calendar_ to clearly distinguish between datasets and avoid confusion after merging.

2. Cleaning and Transforming Data

- **Date Handling**: Dates, such as the host's start date and review dates, were converted into a standard format. We used these to calculate useful fields, like how long a host has been active (in years) and the number of days since the last review.
- **Text Cleanup**: Columns with symbols like % or $ were cleaned to remove these characters and converted into numerical values for analysis.
- **Binary Conversion**: Columns with true or false values, such as whether a host is a superhost, were converted into 1s and 0s to make them easier to analyze.

3. Handling Missing Values

- Numerical columns, such as response rates and review scores, were filled with their average values to maintain data consistency.
- Fields like the number of beds or reviews per month were filled with 0, as missing data here likely indicated no activity rather than an error.
- For some fields, like review activity duration and days since the last review, missing values were filled with -1 to mark them as distinct from valid data.
- Text fields, such as descriptions or neighborhood overviews, were filled with 'Unknown' to ensure no empty values remained.

4. Feature Engineering

- **Bathroom Adjustments**: Bathroom counts were extracted from text descriptions, and adjustments were made based on whether the bathrooms were shared or private. For instance, shared bathrooms were counted as half, while private ones remained unchanged.
- **New Columns**: Additional columns were created, such as the total duration of review activity and the number of days since the last review, to provide insights into property activity over time.

5. Merging Datasets

We combined multiple datasets to create a unified view:

- The calendar dataset was merged with listings data to connect pricing and availability information with property details.
- The reviews dataset was joined with the listings data to link customer feedback to specific properties. After merging, we checked for and removed duplicates to ensure each record was unique and accurate.

ANALYSIS:

Questions Answered

1. How do specific amenities (e.g., Wifi, Kitchen) impact ratings?

- Statistical Method Used: T-Test analysis was conducted to compare the average ratings of listings with and without specific amenities to determine the significance of the impact.
- Fields Used:
  - listings_review_scores_rating: To measure guest satisfaction through ratings.
  - listings_amenities: To identify the presence or absence of specific amenities.

2. What are the pricing trends across neighborhoods?

- Statistical Method Used: Descriptive statistics were employed to calculate averages and distribution of prices, along with correlation analysis to explore relationships between pricing and neighborhood characteristics.
- Fields Used:
  - listings_price: To analyze the pricing for each listing.
  - listings_neighborhood_overview: To group listings based on neighborhood locations.

3. How does host behavior (e.g., response rate, acceptance rate) impact occupancy and revenue?

- Statistical Method Used: Correlation analysis was applied to assess the relationships between host attributes and guest engagement, occupancy rates, and revenue.
- Fields Used:
  - listings_host_response_rate: To evaluate how prompt host responses influence engagement.
  - listings_host_acceptance_rate: To assess the impact of host willingness to accept bookings.
  - listings_reviews_per_month: As a proxy for occupancy and guest activity.

4. How does the duration of review activity impact overall ratings and guest trust?

- Statistical Method Used: Descriptive statistics and correlation analysis were used to explore trends and relationships between review activity duration and guest ratings.
- Fields Used:
  - listings_review_activity_duration: To measure the time span of reviews for each listing.
  - listings_review_scores_rating: To evaluate the overall guest trust and satisfaction.

5. What are the most profitable and in-demand neighborhoods?

- Statistical Method Used: Descriptive statistics and interactive visualizations were created to analyze pricing and occupancy trends across neighborhoods.
- Fields Used:
  - listings_price: To identify high-revenue neighborhoods.
  - listings_reviews_per_month: To measure demand based on the number of reviews.
  - listings_neighborhood_overview: To group and analyze neighborhoods.

PROGRAM DESCRIPTION:

This program analyzes Airbnb property data to explore how different amenities affect customer satisfaction and ratings. It starts by loading and cleaning datasets, which include information about property listings, calendars, and customer reviews. Unnecessary columns are removed, missing values are filled using appropriate methods, and text data is cleaned to make everything consistent. We also created new features, like how long a host has been active, the duration of review activity, and adjusted bathroom counts based on whether they are shared or private.

After cleaning the data, the program combines the datasets into one, connecting property details, pricing, and customer reviews. Then, it analyzes how specific amenities, like Wifi, Pool, and Kitchen, impact customer ratings. The results are visualized using charts, maps, and word clouds to show patterns and trends.

The program uses several Python libraries:

- pandas and numpy for cleaning and analyzing data.
- datetime for working with dates, like calculating durations.
- matplotlib.pyplot and seaborn for creating charts to visualize the data.
- folium and MarkerCluster for building interactive maps to display property locations and patterns.
- wordcloud for generating word clouds from text data.
- scikit-learn's CountVectorizer for processing and analyzing text.

These libraries made it easy to clean, analyze, and visualize the data, helping us understand how amenities influence customer satisfaction in Airbnb properties.

OUTPUT DOCUMENTATION:

The program produced a variety of outputs that provided critical insights into the data, facilitated visualization, and supported statistical analyses

1. **Statistical Outputs**:

- **T-Test Results**: The results of T-tests for different amenities revealed whether the differences in ratings and prices for listings with and without specific amenities were statistically significant. For instance, amenities like Wifi and Kitchen showed marginal yet significant impacts on guest ratings, highlighting the importance of these features in influencing guest satisfaction.
- **Correlation Matrices**: Generated correlation matrices outlined relationships between various numerical fields, such as host response rates, review scores, and listing prices. This helped identify strong and weak dependencies among the variables.

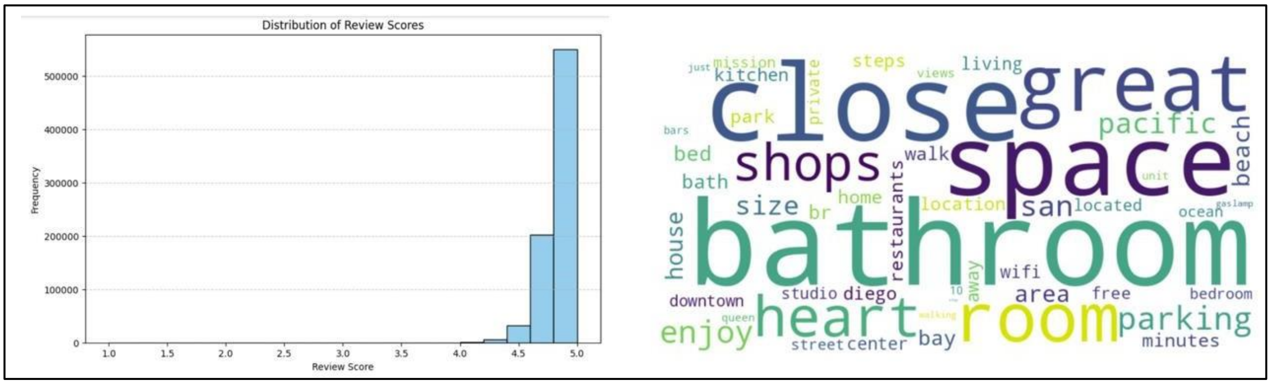
2. **Visualizations**:

- **Bar Charts**: Comparisons of average ratings and prices were visualized using bar charts for listings with and without specific amenities. These charts emphasized trends and helped identify factors contributing to higher ratings or revenue.
- **Scatterplots**: Scatterplots illustrated relationships between key variables, such as host response rate and review scores, enabling an intuitive understanding of correlations.
- **Heatmaps**: Heatmaps visualized correlations among multiple variables in a single chart, providing a comprehensive overview of interdependencies within the dataset.
- **Interactive Maps**: Created using folium, these maps displayed the spatial distribution of Airbnb listings. They highlighted high-demand neighborhoods and provided geographical context to the analysis.

CONCLUSION AND KEY FINDINGS:

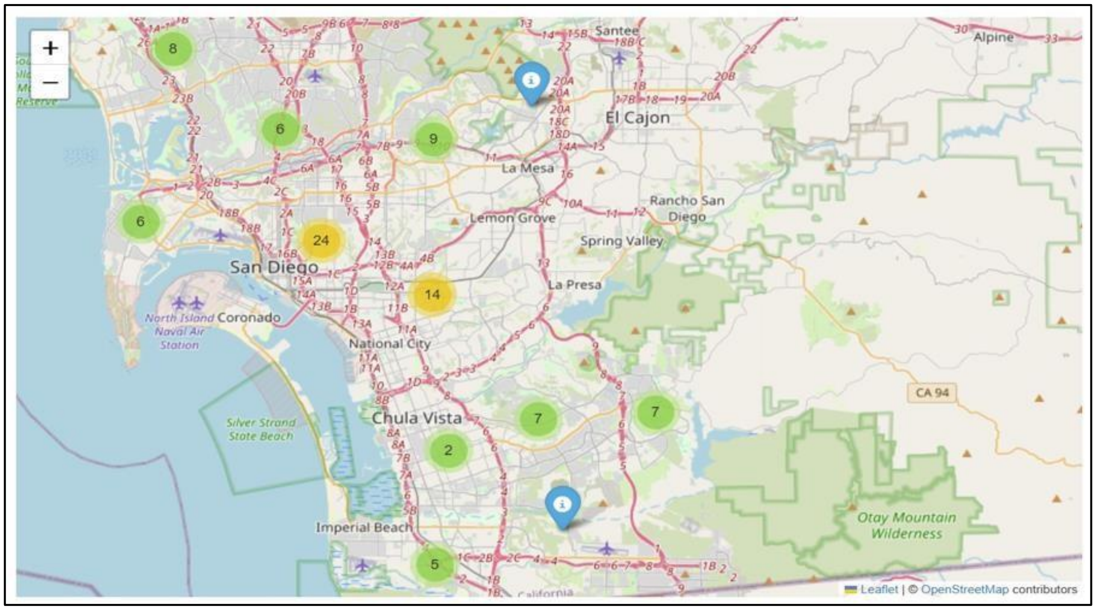
Key Findings:

1. Keyword Association with Review Scores and Popularity:



- The histogram shows that the majority of listings have high review scores, concentrated near 4.5 to 5.0, indicating overall guest satisfaction.
- The word cloud highlights keywords like "bathroom," "close," "shops," and "space" as commonly mentioned in reviews. These terms suggest that cleanliness, proximity to amenities, and spaciousness are critical factors influencing guest experience.

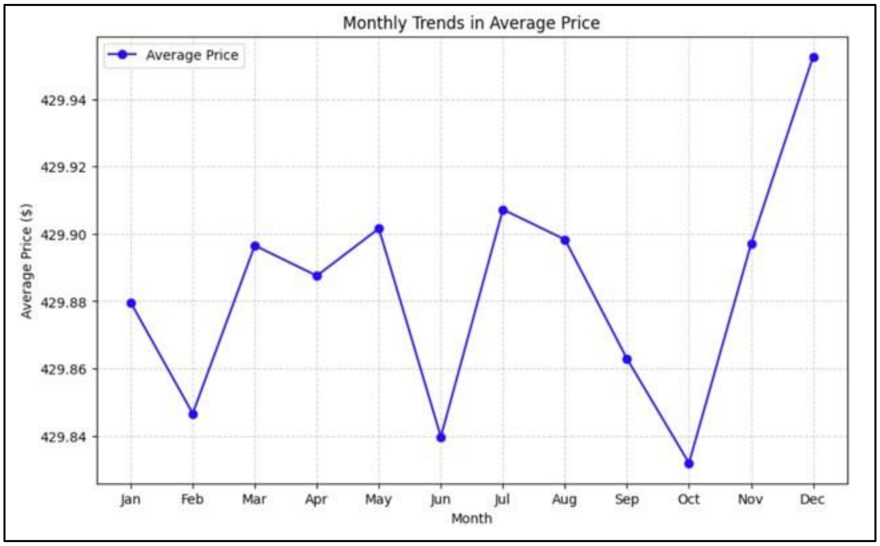
2. Neighborhood Influence on Prices and Availability:



- The interactive map highlights that neighborhoods closer to central San Diego have higher availability and demand, indicated by denser cluster sizes.
- Popular areas like downtown San Diego show a concentration of listings, suggesting that location significantly impacts booking rates and revenue.

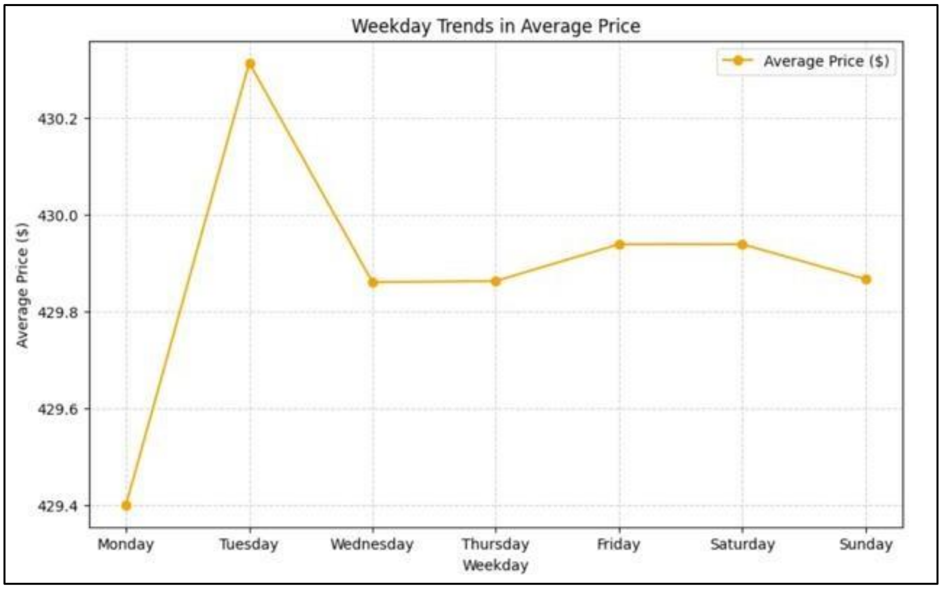
3. Seasonal Fluctuations in Listing Prices:

- **Monthly Trends**:



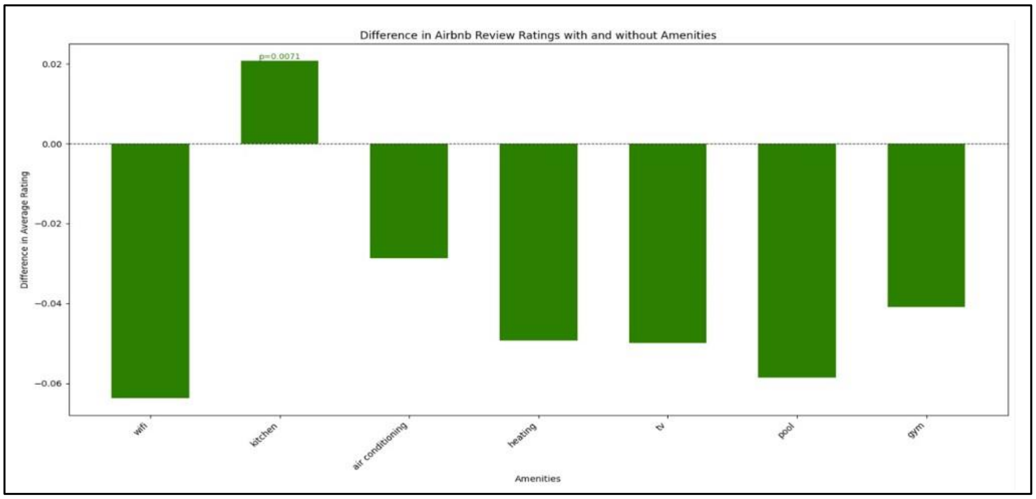
  - There is a noticeable fluctuation in average prices across months, with peaks in certain periods (e.g., December) likely due to seasonal demand.
  - A decline in mid-year months indicates reduced demand during these times.

- **Weekly Trends**:



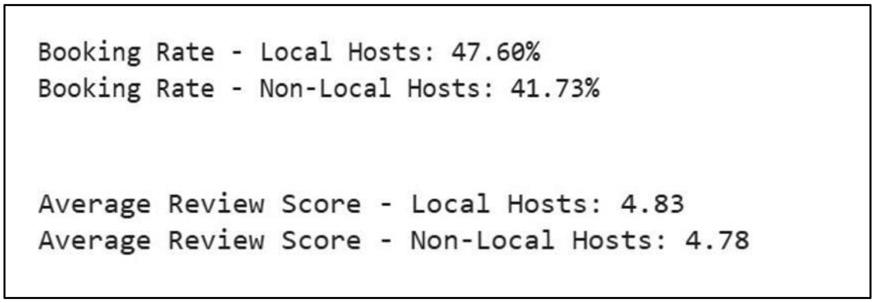
  - Tuesdays and weekends exhibit higher average prices compared to other weekdays, suggesting pricing strategies capitalize on mid-week and weekend travel patterns.

4. **Effect of Amenities on Average Ratings:**



- The bar chart shows a slight positive impact on ratings for listings with a kitchen, while amenities like Wifi, air conditioning, and pools exhibit marginal or negative effects.
- Statistical significance (e.g., p=0.0071 for the kitchen) confirms that some amenities are more influential than others in enhancing guest satisfaction.

5. **Impact of Local Hosts on Booking Rates and Review Scores:**



```
Booking Rate - Local Hosts: 47.60%
Booking Rate - Non-Local Hosts: 41.73%


Average Review Score - Local Hosts: 4.83
Average Review Score - Non-Local Hosts: 4.78
```

- Local hosts outperform non-local hosts in terms of booking rates (47.60% vs. 41.73%), indicating better guest engagement and familiarity with the area.
- Average review scores are slightly higher for local hosts (4.83 vs. 4.78), reflecting better service quality and guest satisfaction.

Conclusion:

1. **Guest Preferences**:
    - Cleanliness, proximity to attractions, and spaciousness are pivotal factors influencing guest satisfaction and review scores.
    - Amenities such as kitchens significantly enhance guest experiences, while Wifi and air conditioning show limited influence.
2. **Location Matters**:
    - Neighborhoods closer to city centers or popular attractions drive higher demand and occupancy rates.
    - Location plays a critical role in optimizing revenue for hosts.
3. **Dynamic Pricing Strategies**:
    - Seasonal and weekly pricing trends indicate opportunities for hosts to adjust prices based on demand, maximizing profitability during peak times.
    - During the low-demand months of February, June, and October, property owners can lower prices to attract bookings while using this period to carry out repairs or renovations with minimal disruption to occupancy.
4. **Local Host Advantage**:
    - Local hosts demonstrate higher booking rates and guest satisfaction, likely due to better area knowledge and personalized services.

5. **Actionable Recommendations**:
    - Hosts should prioritize maintaining cleanliness, improving amenities like kitchens, and targeting high-demand neighborhoods.
    - Dynamic pricing models aligned with seasonal and weekly trends can help maximize revenue.