FINAL_REPORT

May 10, 2024

1 IST 652 - SCRIPTING FOR DATA ANALYSIS - PROJECT REPORT

AIR-BNB NYC 2019 DATASET
GROUP 1 - TANVI PRADHAN, PRIYA VORA & KESHAV CH

1.1 Motivation And Background

The Airbnb platform has disrupted the traditional hospitality industry by empowering individuals to offer short-term rental accommodations. New York City, being a global hub for tourism and business travel, has witnessed a significant rise in Airbnb listings. Understanding the dynamics of this sharing economy model is crucial for various stakeholders, including hosts, guests, policymakers, and the city's tourism industry.

The motivation behind this project is multifaceted. The research questions are worth computing because they provide valuable insights that can inform decision-making processes for various stakeholders involved in the Airbnb ecosystem and the broader hospitality industry.

Understanding pricing patterns, spatial distribution, and the impact of reviews can help hosts optimize their listings, set competitive prices, and improve their overall guest experience. This knowledge can lead to increased occupancy rates, higher revenue, and a better return on investment for hosts.

From a guest's perspective, knowing the answers to these questions can aid in making informed decisions when selecting accommodations. Guests can factor in pricing, neighborhood preferences, and the influence of reviews to find listings that best suit their needs and budget.

For policymakers and city officials, these insights are valuable for regulating the short-term rental market, addressing concerns related to housing affordability, and implementing measures that balance the interests of residents, hosts, and the tourism industry.

Knowing the answers to these research questions can make a significant difference in various aspects of the Airbnb ecosystem and the broader hospitality industry:

Pricing Strategies: Hosts can use the insights on pricing patterns across neighborhoods and room types to set competitive and attractive prices, potentially increasing their bookings and revenue.

Neighborhood Preferences: Understanding the spatial distribution of listings and popular neighborhoods can guide hosts in strategically acquiring or listing properties in high-demand areas, leading to better occupancy rates.

Guest Experience: By analyzing the impact of reviews on pricing and occupancy, hosts can prioritize providing exceptional guest experiences, which can lead to positive reviews and higher demand for their listings.

Market Trends: Insights into room type preferences can help hosts cater to the demand for specific accommodation types, potentially leading to better utilization of their properties.

Regulatory Framework: Policymakers can use the findings related to minimum nights, availability, and pricing strategies to develop regulations that strike a balance between promoting tourism and addressing concerns like housing affordability and neighborhood disruptions. Overall, having a comprehensive understanding of these research questions can empower stakeholders to make data-driven decisions, optimize their strategies, and contribute to the sustainable growth and regulation of the sharing economy in urban settings like New York City.

1.2 Summary of Research Questions and Results

Pricing Analysis: How does the average price vary across different neighborhoods and room types? The analysis revealed significant variations in average prices across different neighborhoods and room types in New York City. Entire home/apartment listings were generally more expensive than private and shared rooms.

Spatial Distribution: What are the patterns in the spatial distribution of Airbnb listings across NYC's neighborhoods? The spatial distribution analysis uncovered distinct clusters of Airbnb listings in certain neighborhoods, with higher concentrations observed in areas like Manhattan, Brooklyn, and Queens.

Review Impact: How do reviews and their frequency influence a listing's price and occupancy rates? The study found a positive correlation between the number of reviews, review frequency, and listing prices. Listings with higher review counts and more frequent reviews tended to command higher prices and have better occupancy rates.

Room Type Preference: What is the distribution and popularity of different room types among Airbnb listings? Entire home/apartment listings were the most prevalent room type, followed by private rooms and shared rooms. The distribution varied across different neighborhoods, reflecting diverse preferences and housing market dynamics.

Regulatory Insights: How do the number of minimum nights and listing availability correlate with pricing strategies? The analysis revealed that listings with a higher minimum number of nights tended to have higher prices, potentially to compensate for the longer commitment required from guests. Additionally, listings with higher availability (fewer blocked dates) generally had lower prices

1.3 Dataset

The dataset used for this project is the "NYC Airbnb Open Data" for the year 2019, which can be accessed through the Hugging Face datasets at the following URL: https://huggingface.co/datasets/gradio/NYC-Airbnb-Open-Data/tree/main

This dataset contains comprehensive information on Airbnb listings in New York City. It includes data on 48,895 unique Airbnb listings, with each listing characterized by 16 different attributes. Key attributes include:

id: Unique identifier for the listing

name: Name of the listing

host_id: Unique identifier for the host

host_name: Name of the host

neighbourhood_group: Borough the listing is located in

neighbourhood: Neighborhood of the listing

latitude: Latitude coordinate of the listing

longitude: Longitude coordinate of the listing

room_type: Type of room offered (Entire home/apt, Private room, Shared room)

price: Price per night for the listing

minimum_nights: Minimum number of nights required for a booking

number_of_reviews: Total number of reviews the listing has received

last_review: Date of the last review

reviews_per_month: Average number of reviews received per month

calculated_host_listings_count: Total number of listings the host manages

availability_365: Number of days the listing is available for booking in a year

1.4 Methodology

1.4.1 Data Preparation and Cleaning

Loading Data: We loaded the dataset into a Pandas DataFrame using pd.read_csv(‘AIRBNB_NYC_2019.csv’). This step converted the CSV file into a DataFrame, a structured format ideal for detailed data manipulation and analysis in Python.

Handling Missing Values We quantified missing values using df.isnull().sum(), which was crucial for assessing data quality and deciding on necessary preprocessing steps. To handle missing values effectively:

Textual data in columns like name and host_name with missing entries were replaced with ‘Unknown’, ensuring that subsequent analyses could proceed without dropping records. For the last_review, missing dates were set to ‘Unknown’, and for reviews_per_month, missing entries were filled with 0, preserving the integrity of date-related operations and review count analyses.

1.4.2 Data Transformation

Date Formatting: We converted last_review from string format to datetime using pd.to_datetime(), treating ‘Unknown’ as NaT (Not a Time). This enabled us to perform time-series analyses and make chronological comparisons. Price Normalization: We corrected unrealistic pricing data by replacing zero prices with the mean price of non-zero listings using df[‘price’].replace(0, df[df[‘price’] != 0][‘price’].mean()).

1.4.3 Feature Engineering

Categorical Variable Encoding One-Hot Encoding: We implemented One-Hot Encoding that converts categorical variables like room_type and neighbourhood_group into a series of binary columns. This process creates a new column for each category value, filled with zeros and ones (0s and 1s), to indicate the absence or presence of that category. This encoding is essential for including categorical data in machine learning models, which typically require numerical input.

Geospatial Mapping Interactive Mapping: Folium is a powerful library for creating interactive maps. We utilised it to visualize geographic data, such as the location of Airbnb listings, by plotting latitude and longitude points on a map. This helped us reveal geographic patterns such as clustering of listings in tourist-heavy neighborhoods or the distribution of room types across the city. These maps were interactive, allowing users to zoom in/out and click on markers to get more information, enhancing the usability of spatial analyses.

1.4.4 In-depth Analysis

Pricing Strategy Exploration Group Analysis: After grouping the data by attributes like neighbourhood and room_type, we calculated the average prices to understand the pricing trends across different areas and types of accommodations. This analysis helped in identifying which neighborhoods are more expensive and which types of rooms (like entire homes vs. shared rooms) command higher prices.

Correlation and Causality Correlation Matrix: We developed a correlation matrix using df.corr() to explore the relationships between continuous variables, such as the number of reviews, pricing, availability, and review frequencies. This matrix helps in identifying potential factors that may influence the pricing or popularity of listings. High correlation coefficients can indicate a strong relationship, whereas low coefficients suggest little to no linear relationship.

1.4.5 Reporting and Insights

Synthesis of Findings Insight Extraction: We synthesized insights from both the exploratory and targeted analyses to provide stakeholders with actionable information on pricing strategies, market positioning, and potential investment opportunities.

Advanced Visualization Detailed Reporting: We employed advanced visualization tools like Seabolin for statistical plots and Folium for geographic data visualization. These tools helped us effectively highlight key trends and patterns, presenting our findings in an accessible and visually appealing format.

1.5 Results

1. Pricing Analysis: Variation Across Neighborhoods and Room Types Our analysis revealed significant price variations across different neighborhoods and room types in New York City. The highest average prices were noted in Manhattan, particularly for entire homes/apartments, aligning with its status as a premium location. In contrast, Brooklyn and Queens offered more economical options, especially for private rooms. This disparity highlights the influence of location and room type on pricing strategies within the city's Airbnb market.

Implications:
These findings suggest potential areas of investment for new Airbnb hosts and could influence pricing adjustments for existing listings to maximize occupancy and revenue. Understanding these dynamics is crucial for hosts aiming to compete effectively in the saturated NYC market.

2. Spatial Distribution of Airbnb Listings The distribution of Airbnb listings was notably dense in central Manhattan and parts of Brooklyn, reflecting higher tourist footfall and the appeal of proximity to major attractions. The outer boroughs showed a sparser distribution, which could be attributed to fewer tourist attractions or possibly zoning regulations affecting short-term rentals.

Implications:
The concentrated nature of listings in specific areas could lead to regulatory scrutiny, as seen in other cities. It might also influence urban planning and local housing policies, considering the impact on rental prices and availability for local residents.

3. Impact of Reviews on Pricing and Occupancy Our findings indicate a strong positive correlation between the number of reviews, review frequency, and both pricing and occupancy rates. Listings with higher review volumes and frequent recent reviews commanded higher prices and greater occupancy, suggesting that prospective guests trust well-reviewed listings more.

Implications:
This underscores the importance of customer satisfaction and engagement for hosts. Encouraging guests to leave reviews could significantly enhance listing attractiveness and profitability.

4. Room Type Preferences and Popularity Entire homes/apartments were the most popular, particularly in Manhattan, followed by private rooms, mainly in budget-conscious boroughs like Queens and Brooklyn. The preference for entire homes/apartments could be linked to the privacy they offer, appealing especially to tourists and small groups.

Implications: This trend might influence Airbnb's marketing strategies and host offerings, potentially spurring more hosts to convert their properties into spaces that can be listed as entire homes/apartments.

5. Minimum Nights and Listing Availability Correlation with Pricing There was a notable correlation between longer minimum stays and higher pricing, particularly in sought-after neighborhoods like Manhattan. This might be a strategy by hosts to reduce turnover costs and stabilize occupancy rates.

Implications:
Policy makers could look into this aspect when designing regulations to ensure a fair balance between tourist accommodation and housing availability for residents.

1.6 Reflection

Learning Outcomes: - Gained a deep understanding of data manipulation and analysis techniques, particularly in Python, which was instrumental in processing and interpreting the extensive Airbnb dataset. - Learned the importance of visual and statistical analysis in revealing underlying patterns that are not immediately obvious.

Challenges Faced: - Initially underestimated the time and effort required for data cleaning and preparation, which are crucial for accurate analysis. - Encountered challenges in applying geospatial analysis due to the complexity of the data and the nuances of spatial data interpretation.

Future Improvements: - Would invest more time in preliminary data exploration to better understand its structure and quality before diving into complex analyses. - Intend to explore more advanced statistical techniques and machine learning models to predict trends and possibly automate some of the analysis for real-time insights.

This project not only enhanced our technical skills but also provided valuable insights into the operational dynamics of Airbnb's market in NYC, offering a blend of theoretical knowledge and practical application.

[ ]:

```

```