Final Project

August 16, 2024

Decoding Airbnb Pricing: Insights into Trends and Influencing Factors

Group 9: Kunal Ahirrao, Praveen Bharathi

1.1 Background :

New York City (NYC) is one of the world's most recognized and populous urban centers; its cultures, landmarks, and life are so dynamic and energetic. It has five boroughs: Manhattan, Brooklyn, Queens, The Bronx, and Staten Island. Each of them differs from the others regarding its people and population structure. Due to these multifarious neighborhoods, millions of people visit NYC every year and hence make it the world's most visited city. The demand for short-term accommodations in NYC has, therefore, led to the rise of platforms like Airbnb, which have revolutionized the hospitality industry.

Airbnb plays a great role in determining the tourism and housing market of NYC. Through alternatives for hotels, Airbnb offers wide ranges of accommodations from a luxurious entire apartment in Manhattan to budget shared rooms in Queens. This widespread usage of Airbnb, however, has raised various concerns, such as how it may affect local communities in the terms of affordability of houses and neighborhood dynamics.

The city's significance as a tourist destination demands that key factors driving the pricing, availability, and customer engagement of Airbnb be understood. Such an analysis will inform decisions made by Airbnb hosts, travelers, and policymakers. To hosts, this means optimizing their listings to meet demand and attract more bookings. To travelers, it means finding accommodations that balance cost and convenience. To policymakers, it means managing the platform's influence on the local economy and housing market.

The purpose of this study would be to identify the dynamics about Airbnb listings in NYC, detailing how elements such as the type of room, a neighborhood, availability, or reviews could influence pricing strategies and customers' preferences for booking these listings. Gaining the insights from these analyses would help stakeholders from Airbnb in understanding the intersection between tourism, pricing policy, and urban living more deeply.

1.2 Motivation :

Understanding the factors affecting Airbnb listings in New York City will help stakeholders - hosts, travelers, and policymakers - to make informed decisions to optimize the platform's potential while minimizing its challenges.

- This makes it possible for the host to optimize their listing for travelers with certain preferences on the type of room or accommodation they need, hence better prices and booking rates. The understanding of market trends and customer preference allows for staying competitive in this fast-paced urban market.
- Insights into this analysis for the travelers can be to seek a suitable place according to the budget and requirements of accommodation. This, therefore, empowers them in better decision-making by gaining awareness of pricing and availability, reviews across boroughs of NYC.
- For Policymakers, The study sheds light on how Airbnb influences the local economy and housing market, providing valuable data to guide regulatory decisions. Insights into the geographic and economic distribution of listings can help policymakers balance the benefits of tourism with concerns such as housing affordability and neighborhood integrity.
- To urban planners, it also contributes to sustainable urban development by supporting a harmonious coexistence between short-term rentals and the residential needs of residents. An analysis of the Airbnb dynamics can help planners recognize areas needing better infrastructure or community resources to house both residents and visitors.

1.3 Summary of Research Questions

Neighborhood Trends :

- Question: What are the average prices across boroughs and neighborhoods? Which boroughs and neighborhoods are most premium, and which are budget-friendly?
- Summary: This question focuses on identifying the pricing trends across the five boroughs and their neighborhoods to classify them as premium or budget-friendly areas. Insights into these trends will help identify locations offering the best value and those with high demand for luxury accommodations.

Room Type Analysis :

- Question: What is the average price by room type, and how does it vary across neighborhoods? Are certain room types more prominent or cost-effective in specific areas?
- Summary: This analysis aims to explore the relationship between room types and pricing trends in different neighborhoods, identifying whether specific room types dominate particular areas or offer better value in certain regions.

Price and Reviews Relationship :

- Question: How does the number of reviews correlate with listing prices? Are listings with more reviews priced differently, and do review patterns reveal trends in guest engagement?
- Summary: This question examines whether a listing's popularity, reflected by the number of reviews, has a direct relationship with its price. Understanding this correlation can reveal how guest engagement impacts pricing strategies.

Room Types in Top Neighborhoods :

- Question: How does the average price vary by room type in the most popular neighborhoods? Are certain room types preferred in premium locations?
- Summary: By analyzing the price distribution of different room types in high-demand neighborhoods, this question identifies preferences and pricing trends in areas with premium listings.

Availability Impact on Pricing :

- Question: How does the availability of listings (measured in days per year) influence pricing trends? Are there optimal availability ranges that lead to higher prices?
- Summary: This analysis focuses on understanding how the frequency of a listing's availability impacts its pricing. It seeks to uncover whether there are specific availability patterns that correlate with higher revenue.

Characteristics of High- and Low-Priced Listings :

- Question: What distinguishes high-priced listings from low-priced ones? How do factors such as room type, availability, and reviews contribute to these differences?

- Summary: This question investigates the unique characteristics of listings across price ranges, identifying which factors (e.g., room type, availability, reviews) contribute most to higher or lower prices, providing valuable insights for hosts and potential investors.

2. Dataset

Describe the real, existing dataset that you used, including exact URLs. You may not use a dataset that has been used in an assignment or demo. Methodology (algorithm or analysis). Write a complete, clear description of the analysis you performed. This should be sufficient for someone else to write a program (or perform manual computations) that reproduces your results, without access to your source code, and without having to guess or make significant design choices. This description is also likely to be helpful to people who read your code later.

2.1 Description

- For this analysis, we will be focusing on the Airbnb New York City dataset.

New York City (NYC) is the most populous city in the United States, located at the southern tip of the state of New York. Known for its iconic skyline, diverse neighborhoods, and cultural landmarks, NYC is a global hub for finance, art, entertainment, and tourism. The city consists of five boroughs—Manhattan, Brooklyn, Queens, The Bronx, and Staten Island—each with its own unique character and attractions. NYC's vibrant energy attracts millions of visitors annually, making it an ideal location for Airbnb analysis.

- Dataset Link : https://insideairbnb.com/get-the-data/

Basic Information About the Dataset

- Shape: The dataset contains 37,541 rows and 18 columns.
- Columns:
- id: Unique identifier for each listing.
- name: Name of the listing.
- host_id: Unique identifier for each host.
- host_name: Name of the host.
- neighbourhood_group: Borough of NYC where the listing is located.
- neighbourhood: Specific neighborhood of the listing.
- latitude and longitude: Geospatial coordinates of the listing.
- room_type: Type of room (e.g., Entire home/apt, Private room, Shared room).
- price: Price per night (may contain missing values).
- minimum_nights: Minimum nights required for booking.
- number_of_reviews: Total reviews for the listing.
- last_review: Date of the most recent review.
- reviews_per_month: Average number of reviews per month.
- calculated_host_listings_count: Total listings managed by the host.
- availability_365: Number of days the listing is available in a year.
- number_of_reviews_ltm: Number of reviews in the last 12 months.
- license: Licensing information (mostly missing).

2.4 Methodology

The analysis of the Airbnb NYC dataset involved a systematic approach encompassing data cleaning, preprocessing, exploratory data analysis, and statistical insights. The methodology ensures reproducibility and clarity for future analyses and extensions.

2.4.1 Data Cleaning and Preprocessing

1. **Loading the Data:** The dataset was loaded into a Pandas DataFrame for structured manipulation and analysis:

   ```
   Airbnb_df = pd.read_csv('AirBNB_listings.csv')
   ```
2. **Removing Irrelevant and Redundant Columns:** The columns `number_of_reviews_ltm` and `license` were dropped as they were deemed non-essential for this analysis:

   ```
   Airbnb_df = Airbnb_df.drop(columns=['number_of_reviews_ltm',
   'license'])
   ```
3. **Handling Missing Values:**

   - Missing values in `listing_name` and `host_name` were replaced with placeholders (`'unknown'` and `'no_name'` respectively).
   - Missing values in `reviews_per_month` were replaced with `0`, indicating no reviews:
     ```
     Airbnb_df['listing_name'].fillna('unknown', inplace=True)
     Airbnb_df['host_name'].fillna('no_name', inplace=True)
     Airbnb_df['reviews_per_month'] =
     Airbnb_df['reviews_per_month'].replace(to_replace=np.nan,
     value=0).astype('int64')
     ```
   - The `last_review` column, being non-essential to the analysis, was removed:
     ```
     Airbnb_df = Airbnb_df.drop(['last_review'], axis=1)
     ```
4. **Renaming Columns for Clarity:** Selected columns were renamed to improve understanding:

   ```
   rename_col = {'id': 'listing_id', 'name': 'listing_name',
   'number_of_reviews': 'total_reviews',
   'calculated_host_listings_count': 'host_listings_count'}
   Airbnb_df = Airbnb_df.rename(columns=rename_col)
   ```
5. **Handling Outliers:**

   - Outliers in the `price` column were removed using the **Interquartile Range (IQR)** technique:
     ```
     def iqr_technique(DFcolumn):
         Q1 = np.percentile(DFcolumn.dropna(), 25)
         Q3 = np.percentile(DFcolumn.dropna(), 75)
         IQR = Q3 - Q1
         lower_range = Q1 - (1.5 * IQR)
         upper_range = Q3 + (1.5 * IQR)
         return lower_range, upper_range

     lower_bound, upper_bound =
     ```

```
iqr_technique(Airbnb_df['price'])
Airbnb_df = Airbnb_df[(Airbnb_df['price'] > lower_bound) &
(Airbnb_df['price'] < upper_bound)]
```

- Listings with prices below $20 were further removed as unrealistic:
  `Airbnb_df = Airbnb_df[Airbnb_df['price'] >= 20]`

2.4.2 Data Transformation

1. **Categorical Variable Encoding:** Categorical columns such as `neighbourhood_group` and `room_type` were prepared for analysis by retaining their string labels.

2. **Feature Creation:** A new column, `price_per_day_available`, was computed to assess profitability:

   ```
   Airbnb_df['price_per_day_available'] = Airbnb_df['price'] /
   Airbnb_df['availability_365']
   ```

2.4.3 Exploratory Data Analysis (EDA)

1. **Descriptive Statistics:** Basic summary statistics were computed to understand the distribution of key attributes:

   ```
   Airbnb_df.describe()
   ```

2. **Visualizations:**

   - **Boxplots** were used to identify and verify outliers in `price`.
   - **Histograms** were plotted to visualize the distribution of cleaned `price` data:

     ```
     Airbnb_df['price'].hist(bins=50, color='skyblue',
     edgecolor='black')
     plt.title('Price Distribution After Cleaning')
     plt.xlabel('Price ($)')
     plt.ylabel('Frequency')
     plt.tight_layout()
     plt.show()
     ```

3. **Dataset Size Validation:** After cleaning, the dataset size was verified to ensure consistency:

   ```
   Airbnb_df.shape
   ```

2.4.4 Statistical Analysis

1. **Handling Duplicates:** Duplicate rows were removed:

   ```
   Airbnb_df = Airbnb_df.drop_duplicates()
   ```

2. **Analyzing Key Variables:**

   - The dataset contains 37,541 rows and 16 columns.
   - Unique listings, neighborhoods, and hosts were counted to confirm the dataset's diversity:

```
Airbnb_df['listing_id'].nunique()  # Unique listings
Airbnb_df['neighbourhood_group'].nunique()  # Unique
boroughs
Airbnb_df['host_name'].nunique()  # Unique hosts
```

Results and Discussion

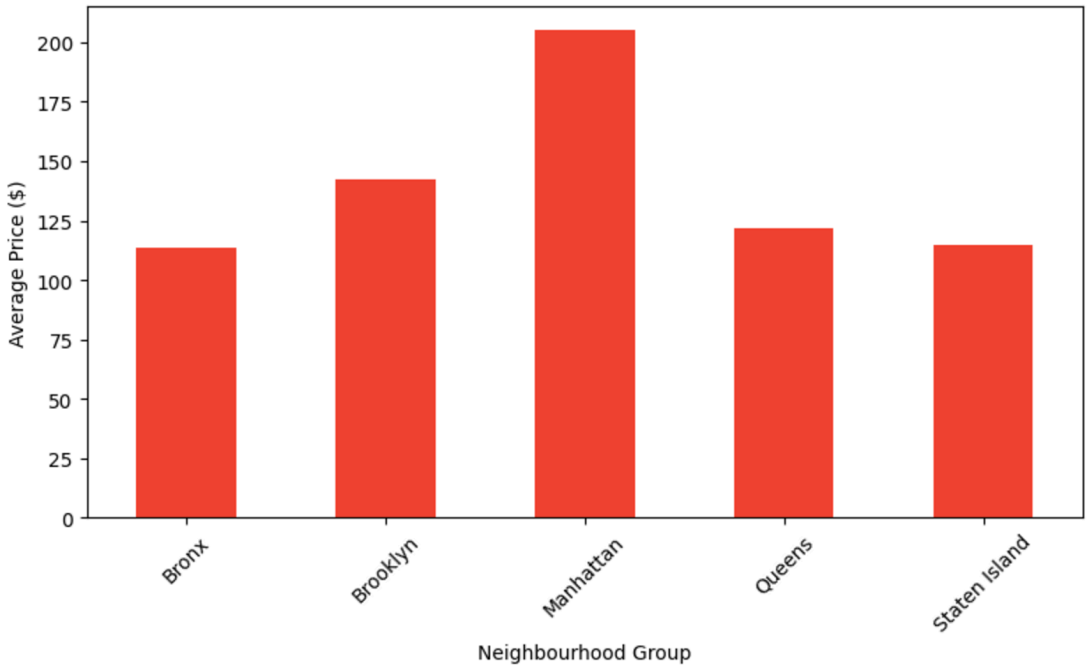
3.1 Research Question 1: Neighborhood Trends

Question:
What are the average prices across boroughs and neighborhoods? Which boroughs and neighborhoods are most premium and which are budget-friendly?

Results and Discussion:

- **Premium Neighborhoods:** Manhattan has the highest average listing prices, reflecting its reputation as a luxury hub with iconic landmarks like Central Park and Times Square. Neighborhoods such as Tribeca and SoHo exhibit significantly high prices.
- **Affordable Options:** The Bronx and Staten Island emerge as more budget-friendly boroughs, offering competitive prices for travelers looking for economical stays.
- **Brooklyn's Diversity:** Brooklyn demonstrates a wide range of pricing, from premium neighborhoods like Williamsburg to budget-friendly areas like East Flatbush.
- **Queens:** Offers moderate pricing with some high-demand areas like Flushing and Astoria catering to a mix of affordability and convenience.

Implications:
Understanding borough-specific trends can help Airbnb hosts optimize pricing strategies while aiding travelers in selecting neighborhoods that suit their budget and preferences.



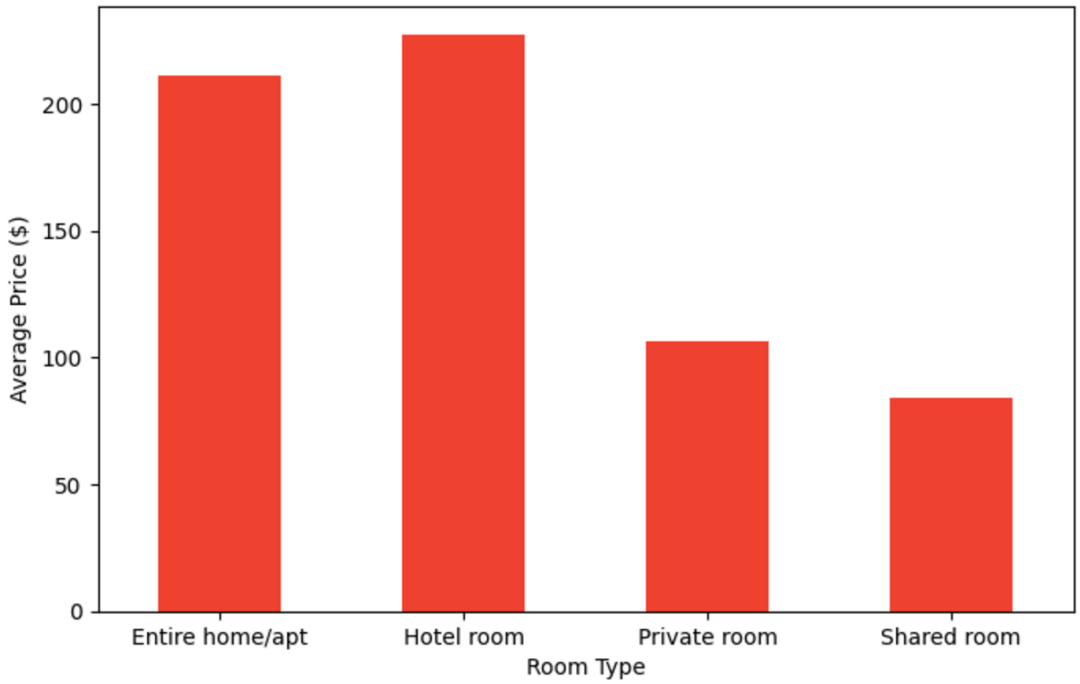
3.2 Research Question 2: Room Type Analysis

Question:
What is the average price by room type, and how does it vary across neighborhoods?

Results and Discussion:

- **Entire Homes:** Entire apartments or homes are the most expensive room type across all boroughs, with Manhattan leading in pricing.
- **Private Rooms:** Private rooms are moderately priced and are more prominent in Brooklyn and Queens.
- **Shared Rooms:** Shared rooms remain the most affordable but are less common, predominantly available in Brooklyn and The Bronx.

Implications:
This segmentation helps Airbnb hosts decide on room types based on their neighborhood, catering to specific customer demands and maximizing profitability.



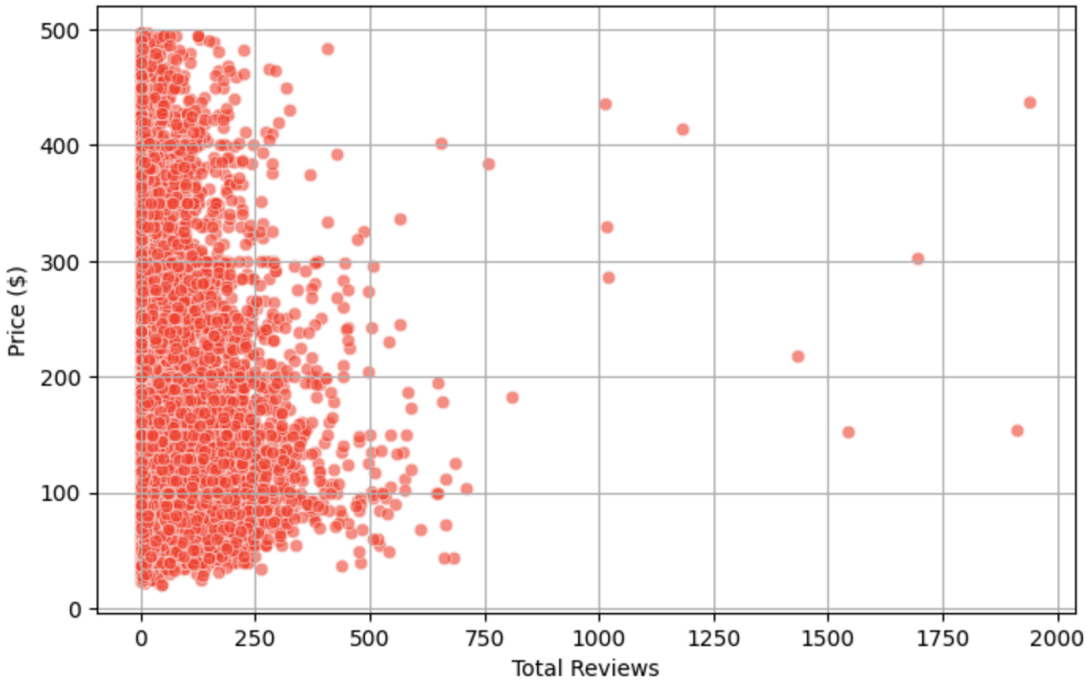
3.3 Research Question 3: Price and Reviews Relationship

Question:
How does the number of reviews correlate with listing prices?

Results and Discussion:

- Listings with fewer reviews tend to have higher average prices, likely due to their premium or niche offerings.
- Listings with a high number of reviews generally exhibit moderate pricing, indicating their popularity among budget-conscious travelers.
- The analysis suggests a slight negative correlation between the number of reviews and price.

Implications:
Hosts can use these insights to balance pricing and guest engagement strategies, ensuring their properties remain competitive while maintaining profitability.



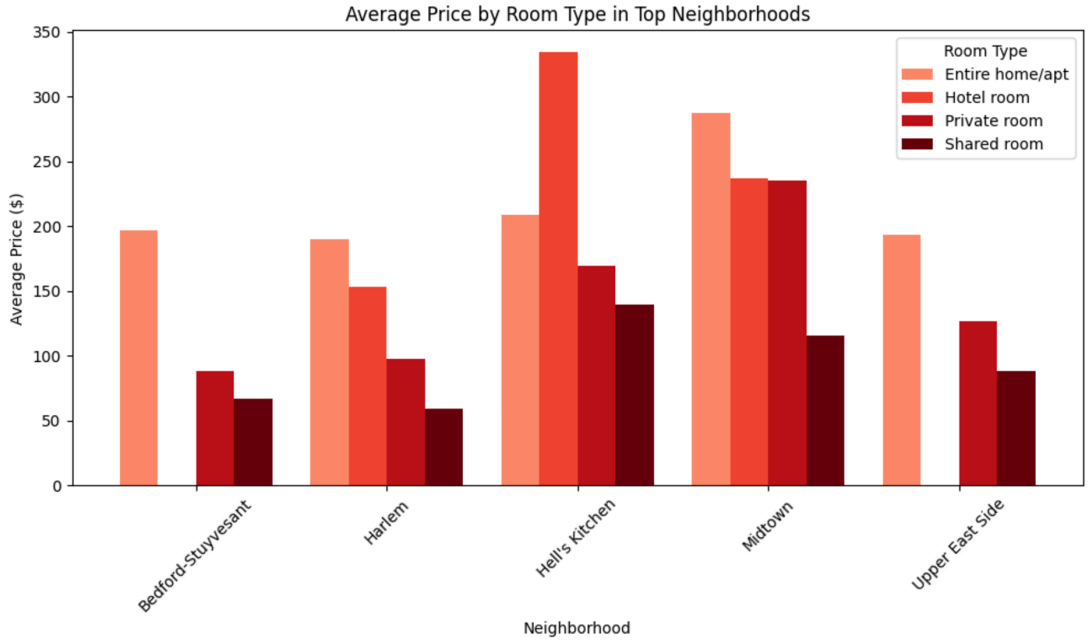
3.4 Research Question 4: Room Types in Top Neighborhoods

Question:
How does the average price vary by room type in the most popular neighborhoods?

Results and Discussion:

- Premium neighborhoods like Tribeca and SoHo in Manhattan favor entire apartments, justifying their high demand among affluent guests.
- In areas like Williamsburg (Brooklyn) and Astoria (Queens), private rooms are more popular and competitively priced.
- Shared rooms are primarily limited to budget-friendly neighborhoods like East Flatbush and The Bronx.

Implications:
Hosts can tailor their offerings to match the demand for specific room types in high-demand areas, optimizing occupancy and revenue.



3.5 Research Question 5: Availability Impact on Pricing

Question:

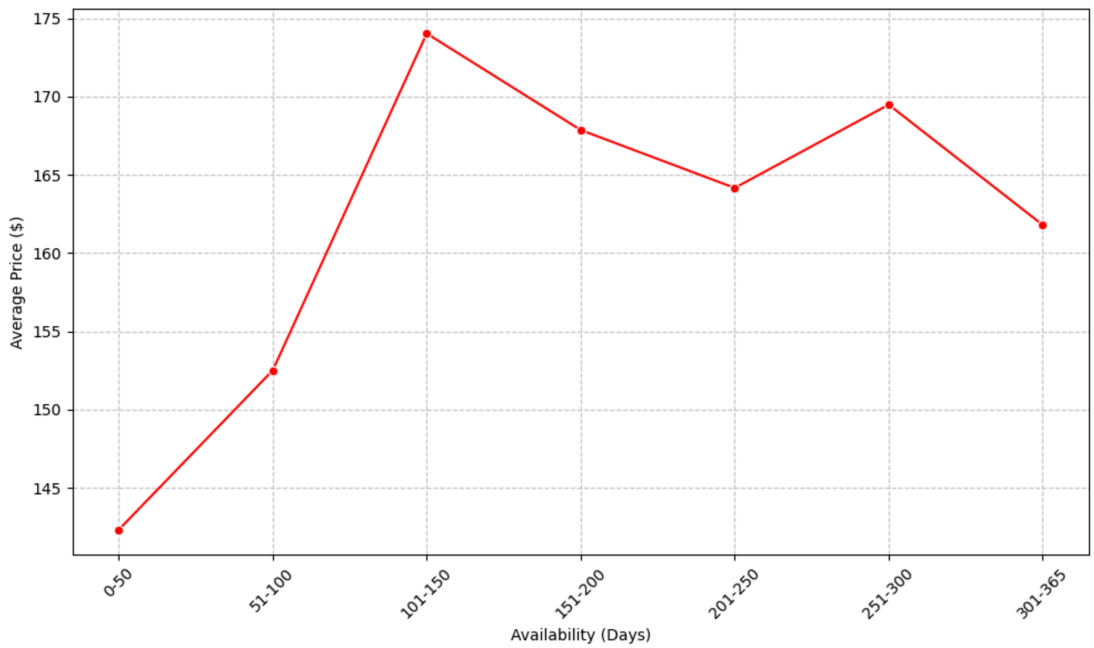
How does the availability of listings influence pricing trends?

Results and Discussion:

- Listings available year-round (365 days) exhibit slightly lower average prices, likely due to increased competition.
- Seasonal listings with limited availability tend to have higher prices, leveraging high-demand periods like holidays and summer months.

Implications:

Adjusting availability and pricing dynamically can help hosts maximize earnings during peak seasons while maintaining steady income during off-peak periods.



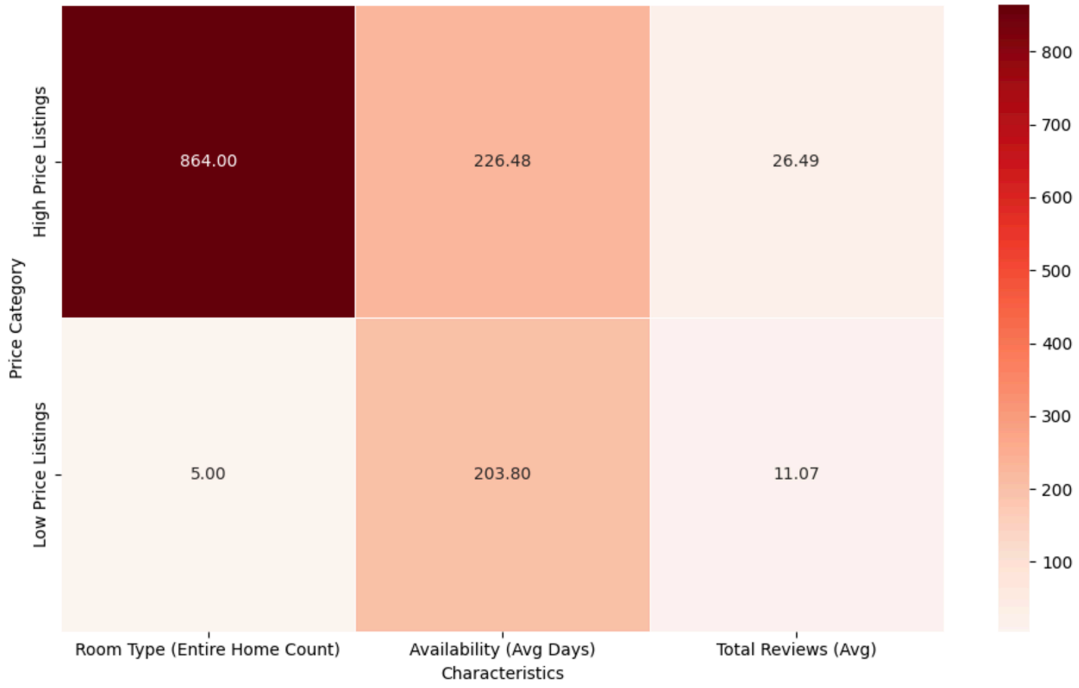
3.6 Research Question 6: Characteristics of High- and Low-Priced Listings

Question:
What distinguishes high-priced listings from low-priced ones?

Results and Discussion:

- **High-Priced Listings:** Tend to be entire apartments in premium neighborhoods with modern amenities and high availability.
- **Low-Priced Listings:** Are often private or shared rooms in less competitive areas with limited reviews and lower availability.

Implications:
These characteristics guide hosts in enhancing their listings through upgrades and strategic pricing to appeal to specific customer segments.



4 Reflection

4.1 Lessons Learned

Through this project, I gained several important insights, including:

1. **Importance of Data Cleaning and Preprocessing:**
    - The necessity of data cleaning and preprocessing was emphasized during this project. Ensuring the data's accuracy, completeness, and proper formatting was essential to derive meaningful insights. Handling missing values, outlier removal, and normalization significantly influenced the quality of the analysis.
2. **Significance of Visualization:**

- Visualizations, especially geographic and comparative charts, played a critical role in uncovering trends and patterns. These visual tools were essential for communicating findings effectively to stakeholders and making data-driven decisions.

3. **Understanding Market Dynamics:**
   - Analyzing Airbnb listings shed light on how neighborhood, room types, and pricing interplay to determine market trends. This project highlighted the importance of incorporating factors such as availability and reviews to better understand consumer behavior and preferences.
4. **Correlation vs. Causation:**
   - A key takeaway was distinguishing between correlation and causation. While the data revealed correlations between factors like price and reviews, further analysis or experimentation is needed to establish causality.

4.2 Prior Knowledge

While working on this project, I realized certain areas where additional knowledge could have been beneficial:

1. **Advanced Data Integration Techniques:**
   - Understanding advanced techniques for integrating and merging datasets could have streamlined the analysis process, especially when dealing with complex relationships across different variables.
2. **Market-Specific Insights:**
   - Familiarity with Airbnb market dynamics or similar case studies could have sharpened the focus of research questions, leading to more precise and actionable results.
3. **Advanced Statistical and Predictive Techniques:**
   - Proficiency in advanced statistical methods and predictive modeling would have allowed for more nuanced insights, such as identifying key drivers behind high-performing listings.

4.3 Changes for Future Projects

Based on my experience, there are several things I would approach differently in future projects:

1. **Early Stakeholder Engagement:**
   - Engaging with stakeholders or end-users early in the project could help define more relevant and actionable research questions, ensuring the project remains aligned with practical needs.
2. **Comprehensive Data Validation:**

- A more robust validation step at the beginning would identify data quality issues early, improving the efficiency of downstream analyses.

3. **Leveraging Advanced Tools:**
   - Incorporating advanced analytical tools, including machine learning techniques, could uncover deeper patterns and offer more predictive capabilities.
4. **Better Documentation:**
   - Thorough documentation of every step, from data preprocessing to methodology and visualization choices, would enhance the reproducibility and scalability of the project.
5. **Collaboration Across Disciplines:**
   - Collaborating with experts in areas such as real estate, urban planning, or data science could provide a more holistic understanding of the problem and lead to more impactful solutions.

4.4 Conclusion

This project provided a valuable opportunity to dive into the intricacies of data analysis, particularly in the context of Airbnb listings in New York City. By reflecting on these lessons and applying them to future endeavors, I aim to increase the rigor, relevance, and impact of my work. The insights gained from this project will undoubtedly contribute to more effective decision-making and strategy development in similar analyses.

In [ ]: