Airbnb Exploratory Data Analysis

Monalisa Burma

Date: 19/02/2024

Project Overview

The purpose of this project was to conduct an exploratory data analysis (EDA) on the Airbnb dataset. The analysis aimed to gain insights into various aspects such as pricing trends, property types, room preferences, and geographic distribution of listings.

Data Sources

The project utilized three main datasets:

Dataset Link: https://www.kaggle.com/datasets/airbnb/seattle/data

Calendar Dataset:

Description: Contains information about the availability and pricing of Airbnb listings on specific dates.

Listings Dataset:

Description: Provides detailed information about individual Airbnb listings, including property details, host information, and reviews.

Reviews Dataset:

Description: Includes information about reviews submitted by guests, including reviewer details and comments.

Steps Taken

1. Data Loading and Exploration

Loaded the calendar, listings, and reviews datasets.

Explored basic information about each dataset using the info() function.

[3]:
```
print("Calendar dataset info:")
print(calendar.info())
```

```
Calendar dataset info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1393570 entries, 0 to 1393569
Data columns (total 4 columns):
 #   Column      Non-Null Count    Dtype
---  ------      --------------    -----
 0   listing_id  1393570 non-null  int64
 1   date        1393570 non-null  object
 2   available   1393570 non-null  object
 3   price       934542 non-null   object
dtypes: int64(1), object(3)
memory usage: 42.5+ MB
None
```

[4]:
```
print("\nListings dataset info:")
print(listings.info())
```

```
Listings dataset info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3818 entries, 0 to 3817
Data columns (total 92 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   id                                3818 non-null   int64
 1   listing_url                       3818 non-null   object
 2   scrape_id                         3818 non-null   int64
 3   last_scraped                      3818 non-null   object
 4   name                              3818 non-null   object
 5   summary                           3641 non-null   object
 6   space                             3249 non-null   object
 7   description                       3818 non-null   object
 8   experiences_offered               3818 non-null   object
 9   neighborhood_overview             2786 non-null   object
 10  notes                             2212 non-null   object
 11  transit                           2884 non-null   object
 12  thumbnail_url                     3498 non-null   object
 13  medium_url                        3498 non-null   object
 14  picture_url                       3818 non-null   object
 15  xl_picture_url                    3498 non-null   object
 16  host_id                           3818 non-null   int64
```

```
[5]: print("\nReviews dataset info:")
     print(reviews.info())
```

```
Reviews dataset info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 84849 entries, 0 to 84848
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype 
---  ------         --------------  ----- 
 0   listing_id     84849 non-null  int64 
 1   id             84849 non-null  int64 
 2   date           84849 non-null  object
 3   reviewer_id    84849 non-null  int64 
 4   reviewer_name  84849 non-null  object
 5   comments       84831 non-null  object
dtypes: int64(3), object(3)
memory usage: 3.9+ MB
None
```

2. Data Cleaning

Handled missing values in the calendar dataset by filling NaNs in the 'price' column with 0.

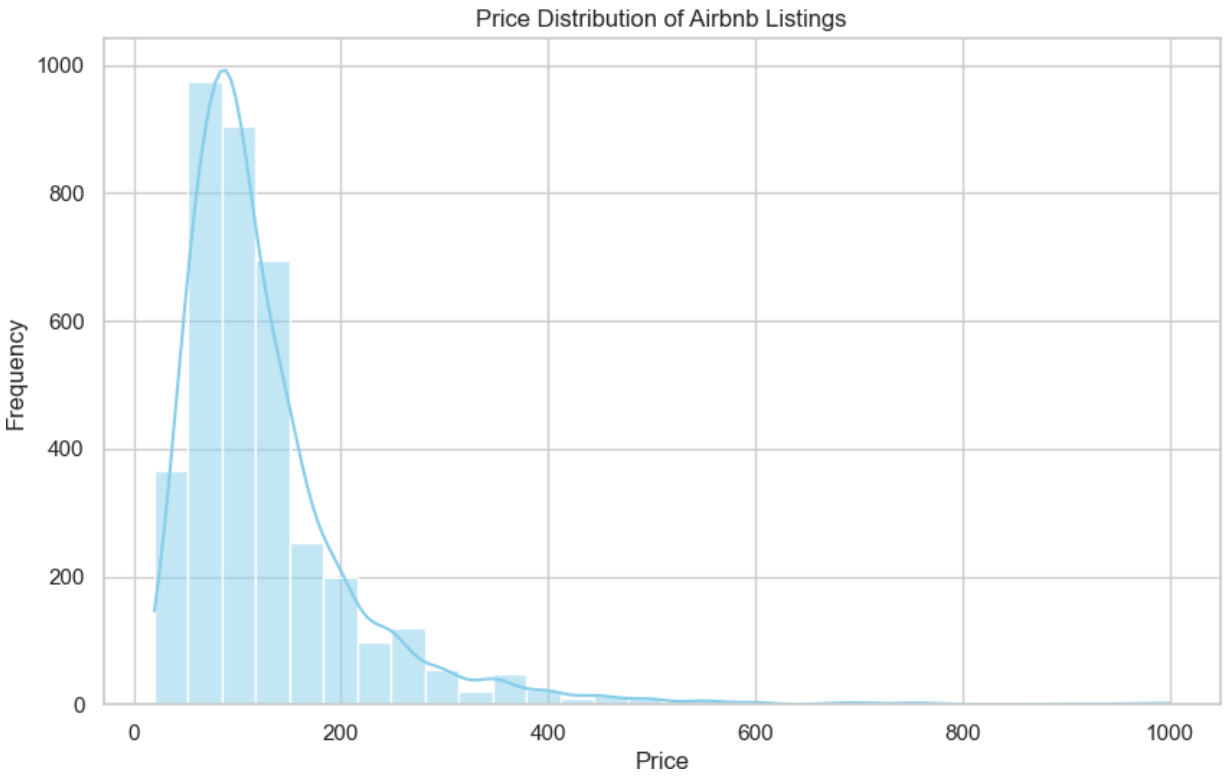
Removed duplicate rows in the listings dataset.

3. Average Daily Price Analysis

Calculated and visualized the average daily price trend using the calendar dataset and got approximately $92.50.

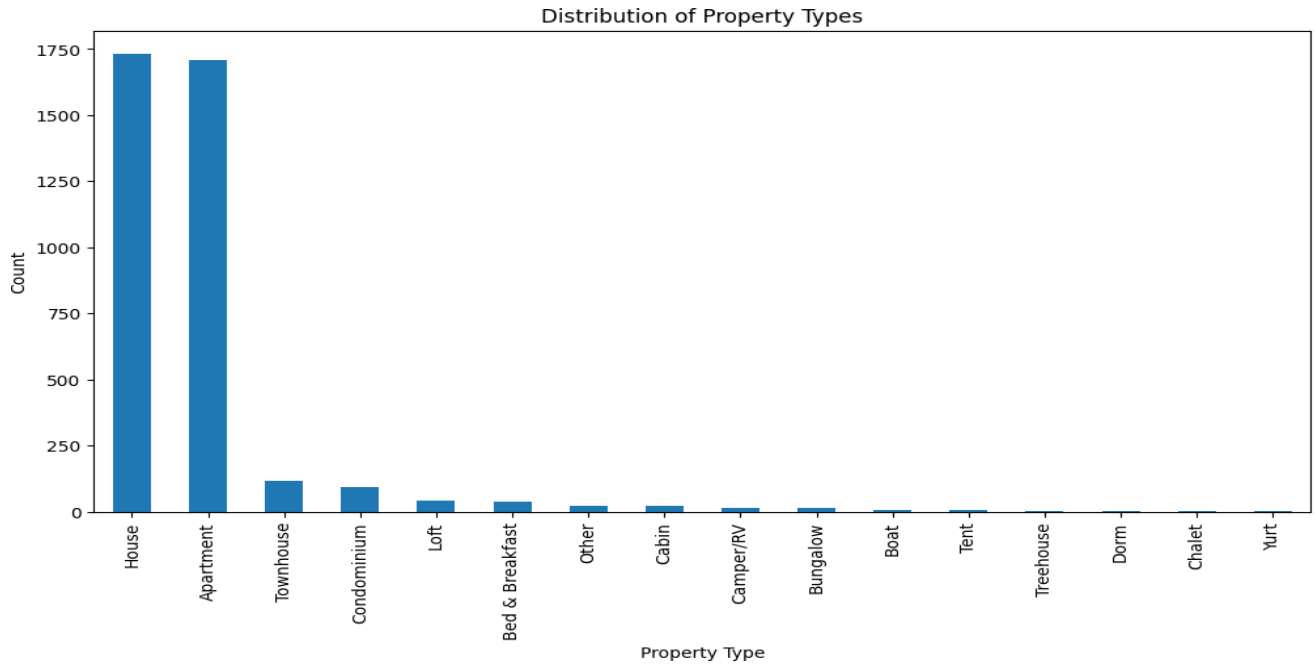
4. Some Visualization Graphs

Price Distribution Analysis



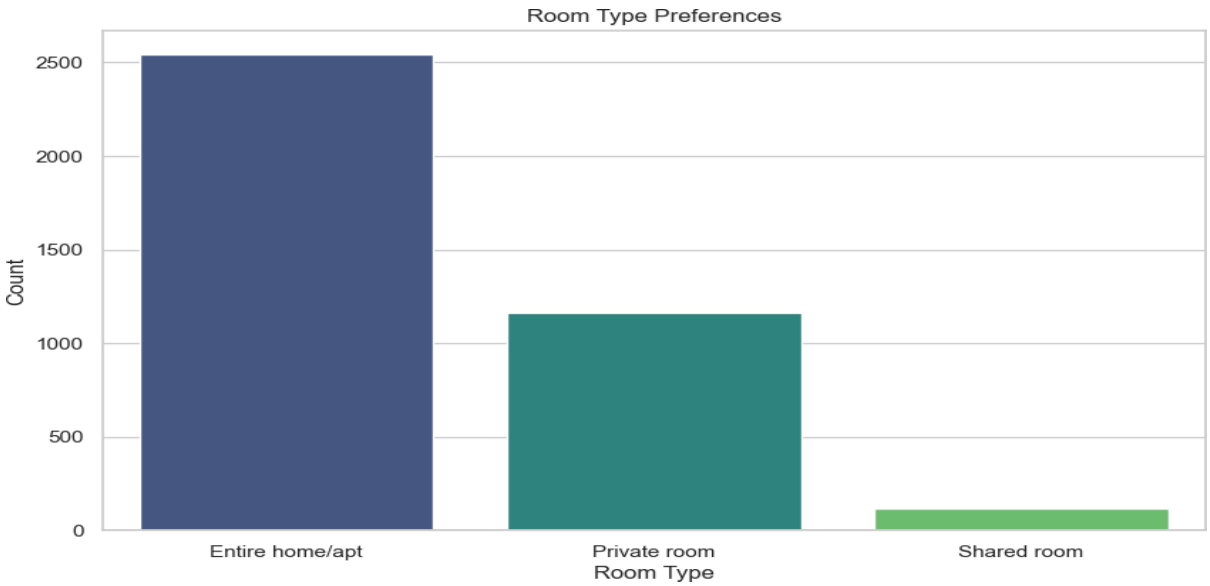
- From this graph we can observe that the highest frequency is occurring around the **$100** price range. This suggests that a significant number of listings are priced in this range.
- The majority of listings are clustered on the **lower-priced** side, indicating that more affordable options are prevalent.
- As prices increase beyond $200, the number of listings declines sharply.
- The smooth line (KDE) overlaid on the histogram represents the **probability density** of prices.
- It follows a similar pattern to the histogram but provides a continuous view of data density.

Property Type Distribution



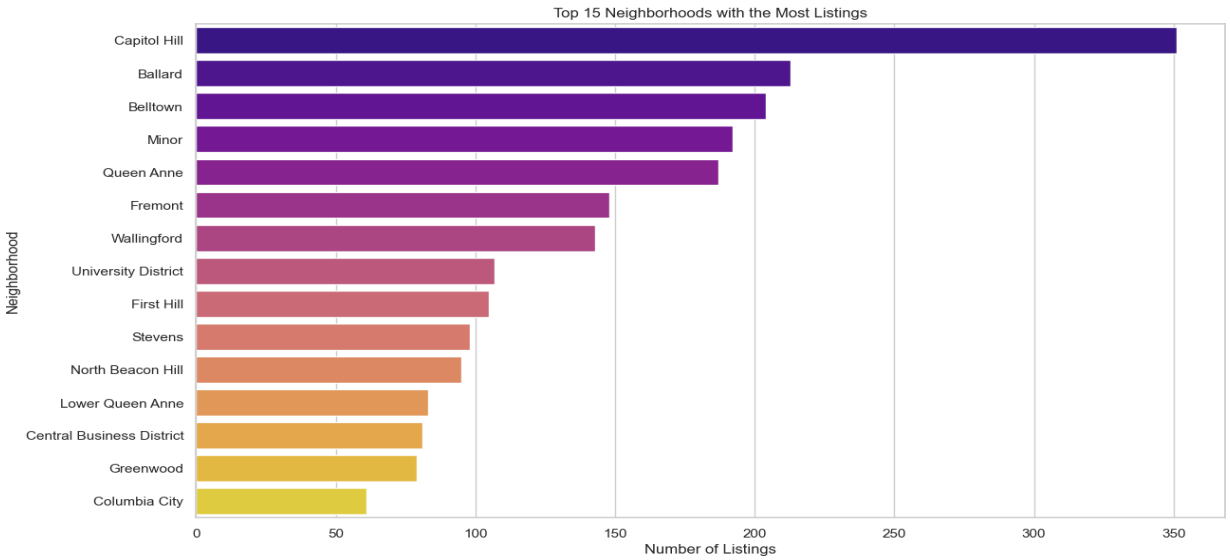
- **House** and **Apartment** are the most common property types, with significantly higher counts than other categories. There is a steep drop in count after House and Apartment, indicating their dominance.
- Property types like **Townhouse**, **Condominium**, and **Loft** have moderate representation.
- Types like **Bed & Breakfast**, **Other**, and **Cabin** are less common.
- The least common property types include **Tent**, **Treehouse**, **Dorm**, **Chalet**, and **Yurt**.

Room Type Preferences



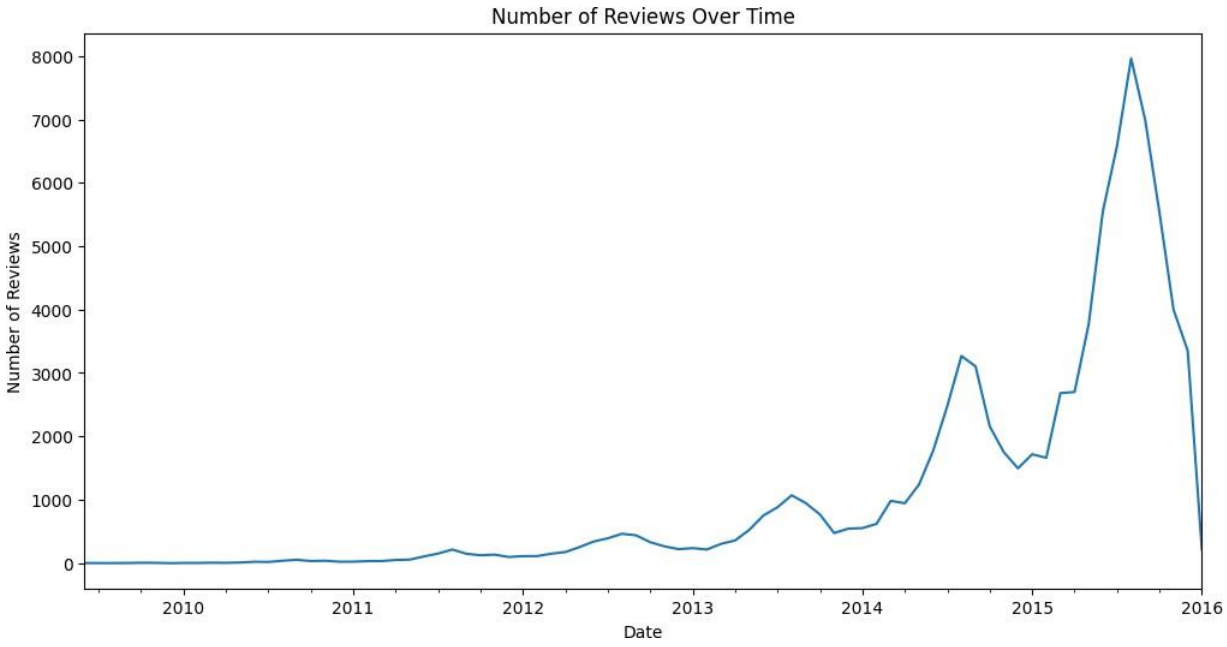
- **Entire home/apt** room type is the most preferred among guests.
- The second most preferred room is the **private room**, where gusts have private room within shared property.
- And the least preferred room type is **shared room**, and this is selected by budget-conscious travelers or those seeking a communal experience.

Number of Listings by Neighborhood



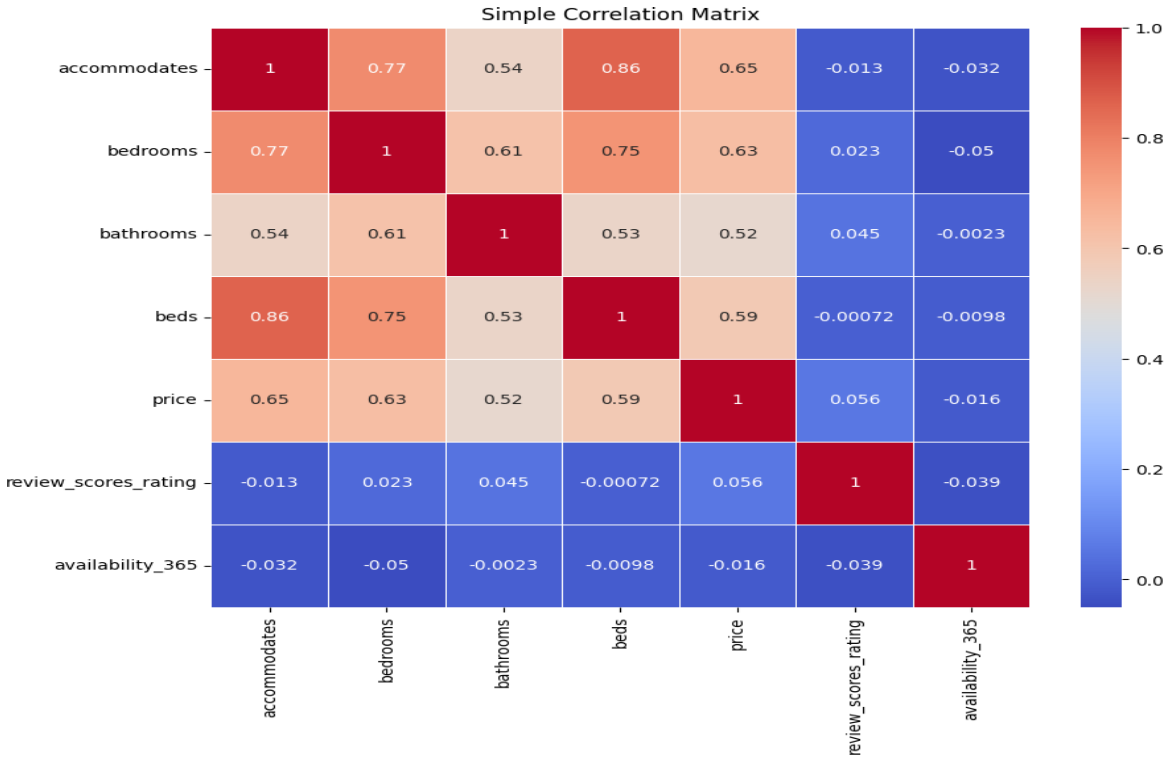
- **Capitol Hill** dominates with the highest number of listings, followed by **Ballard** and **Belltown**.
- The distribution of listings varies across neighborhoods, reflecting their popularity and demand.
- Hosts and travelers can use this information to make informed decisions about where to stay or list their properties.

Time Series Analysis of Reviews



- **Steady Low Volume (2010–2015)**:
  - From 2010 until mid-2015, there is a relatively **flat and low volume** of reviews.
  - The number of reviews remains consistent during this period.
- **Sudden Surge (Mid-2015)**:
  - Around mid-2015, there is a **significant increase** in the number of reviews.
  - The trend sharply rises, indicating a surge in guest feedback.
- **Peak (Early 2016)**:
  - The peak occurs in early 2016, with over **7000 reviews**.
  - This suggests a period of high activity and engagement on the platform.
- **Abrupt Decline (Post-Peak)**:
  - After reaching the peak, there is an **abrupt decline** in the number of reviews.
  - The downward trend continues beyond early 2016.

Correlation Analysis



- The number of **bedrooms** and **accommodates** has a strong positive correlation. **Price** is positively correlated with **accommodates**, **bedrooms**, **bathrooms**, and **beds**.
- There is a weak negative correlation between **availability_365** and **price**.
- **Review scores** are not strongly influenced by other variables in this subset.

Conclusion

In conclusion, the Airbnb Exploratory Data Analysis (EDA) has provided insightful findings into various aspects of the dataset. The price distribution analysis revealed a concentrated pricing landscape, essential for both hosts and guests. Property and room type distributions shed light on preferences and popular listing types. Geographic and neighborhood analyses assisted in understanding where listings are concentrated, aiding hosts in optimizing their offerings. The word cloud and time series analysis of reviews unveiled sentiments and temporal patterns in guest feedback. The correlation matrix highlighted connections between features, such as price and review scores, offering valuable insights for both hosts and potential guests. Overall, this EDA equips stakeholders with a deeper understanding of the Airbnb market dynamics, facilitating data-driven decision-making for hosts and enhancing the booking experience for guests.

Github Link: https://github.com/monalisaburma/Coding_Samurai/tree/main/Airbnb_EDA