**Project Document: Seattle Airbnb Listings Analysis**

**I. Summary of the Data Source (listings.csv)**

**1. Data Source:**

- Name: listings.csv

- Format: Comma-separated values (CSV)

- Source: Airbnb dataset

- Date of Acquisition: Sep. 1, 2023

**2. Data Collection:**

- Collection Method: The data was obtained from Airbnb's platform or an authorized data provider.

**3. Data Contents:**

The listings.csv dataset contains information related to Airbnb listings. It includes a wide range of attributes for each listing. Some key columns include:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Description** | **Time variant/**  **Time invariant** | **Structured/**  **Unstructured** | **Quantitative/Qualitative** | **Nominal/**  **Ordinal**  **/Discrete/**  **Continuous** |
| id | A unique identifier for each listing. | Time invariant | Structured | Quantitative | Discrete |
| room\_type | Indicates the type of room available | Time invariant | Structured | Qualitative | Nominal |
| accommodates | The number of guests the listing can accommodate. | Time invariant | Structured | Quantitative | Discrete |
| bedrooms | the number of bedrooms in the listing. | Time invariant | Structured | Quantitative | Discrete |
| bathrooms | the number of bathrooms in the listing. | Time invariant | Structured | Quantitative | Discrete |
| amenities |  | Time variant | Unstructured | Qualitative | Nominal |
| price | the price of the listing | Time variant | Structured | Quantitative | Continuous |
| minimum\_nights | the minimum number of nights a guest must stay. | Time variant | Structured | Quantitative | Discrete |
| maximum\_nights | the maximum number of nights a guest can stay. | Time variant | Structured | Quantitative | Discrete |
| number\_of\_reviews | the total number of reviews for the listing. | Time variant | Structured | Quantitative | Discrete |
| review\_scores\_rating | the overall rating score for the listing. | Time variant | Structured | Quantitative | Continuous |
| review\_scores\_accuracy | review scores related to accuracy | Time variant | Structured | Quantitative | Continuous |
| review\_scores\_cleanliness | review scores related to cleanliness | Time variant | Structured | Quantitative | Continuous |
| review\_scores\_checkin | review scores related to  check-in | Time variant | Structured | Quantitative | Continuous |
| review\_scores\_communication | review scores related to communication | Time variant | Structured | Quantitative | Continuous |
| review\_scores\_location | review scores related to location | Time variant | Structured | Quantitative | Continuous |
| review\_scores\_value | review scores related to value | Time variant | Structured | Quantitative | Continuous |
| latitude | The latitude coordinate of the listing's location | Time invariant | Structured | Quantitative | Continuous |
| longitude | The longitude coordinate of the listing's location. | Time invariant | Structured | Quantitative | Continuous |

**4. Data Cleaning**

**4.1. Wrangling**

|  |  |  |  |
| --- | --- | --- | --- |
| **Columns Dropped** | **Columns renamed** | **Columns Type Changed** | **Comment/Reason** |
| 56 irrelevant columns |  |  |  |
|  | 'bathrooms\_text' 🡪'bathrooms' |  | for a more accurate description. |
|  |  | bathrooms:  object 🡪 float64 | To enhance the data's usability |
|  |  | price:  object 🡪 float64 | To enhance the data's usability |
|  |  | amenities:  object 🡪 DataFrame  of binary columns | The 'amenities' column includes a textual list of provided amenities for each listing. To enhance the data's usability for analysis, the analyst has decided to split the text based on the delimiter (comma) and create new binary columns for each amenity. |

**4.2. Consistency Checks**

|  |  |  |  |
| --- | --- | --- | --- |
| **Missing values** | **Missing Value Treatment** | **Duplicates** | **Comment/Reason** |
|  |  | none |  |
| bedrooms: 1576 | Imputed value 1 |  | Upon examining the dataset, it appears that most entries with missing values correspond to private rooms. |
| bathrooms: 14 | Dropped |  | The number of missing values is less than 5%. |
| review\_scores\_rating: 1074  review\_scores\_accuracy: 1077  review\_scores\_cleanliness: 1077  review\_scores\_checkin: 1077  review\_scores\_communication: 1077  review\_scores\_location: 1077  review\_scores\_value: 1077 | Dropped |  | Most missing values are primarily associated with listings with 0 availability and the analyst cannot reasonably impute these missing values. Hence, the analyst decides to exclude them to improve the quality of your analysis. |

**II. Data Relevance**

The 'listings.csv' dataset is highly relevant for investigating key questions related to the Seattle Airbnb market. It provides essential information on property types, neighborhoods, pricing, guest preferences, and seasonal trends. By analyzing columns such as 'property\_type' and 'neighborhood,' we can identify prevalent property types in different Seattle neighborhoods. The 'bedrooms' and 'price' columns enable the exploration of correlations between bedroom counts and rental prices across neighborhoods. Additionally, the 'neighborhood' and 'reviews' data offer insights into guest preferences and the popularity of neighborhoods. When combined with 'price' and 'availability' columns, the dataset provides a means to understand how neighborhood characteristics influence rental prices and occupancy rates. Finally, by examining 'availability' columns and the 'price' column, seasonal trends in rental pricing and occupancy rates can be uncovered. Moreover, the dataset allows for an assessment of the correlation between review scores and listing prices, shedding light on the relationship between property quality and pricing. In summary, the 'listings.csv' dataset serves as a comprehensive resource for exploring various facets of the Seattle Airbnb market, offering valuable insights for property management and guest experiences.

Bottom of Form

**III. Data Limitations and Ethics:**

Data Limitations:

- Missing Data: The dataset may have missing values, particularly in columns related to reviews and ratings.

- Data Quality: Data quality issues, such as outliers or inaccuracies, may be present and require cleaning.

- Representativeness: The dataset may not represent the entire Airbnb market in the specified location, as it only includes properties listed on the platform.

Ethics:

- Privacy: Care must be taken to handle any potentially sensitive information, ensuring that no personally identifiable information (PII) is disclosed.

- Bias: There may be bias in the dataset due to selection bias (only properties listed on Airbnb) and potential bias in reviews.

**IV. Questions to Explore:**

Based on the available data, the following questions will be explored:

1. What are the most common property types within different Seattle neighborhoods?
2. Is there a correlation between the number of bedrooms and rental prices across neighborhoods?
3. Which neighborhoods are the most popular among Airbnb guests in Seattle?
4. How do neighborhood characteristics influence rental prices and occupancy rates?
5. Are there seasonal trends in rental prices and occupancy rates?
6. Do the review scores correlate with listing prices?