Task Sheet 6.1

Ashwani Sherawat

**Data Source:**

The dataset, sourced from **insideairbnb.com**, captures the Airbnb landscape in Amsterdam as of **December 6th, 2018**, comprising approximately 20,000 listings, detailed variables, 365-day availability and pricing data, text-based reviews for analysis, and a neighborhood shapefile for spatial insights.

**Data Collection:**

The data was collected form <http://insideairbnb.com/amsterdam/> and uploaded by Erik Bruin on [Kaggle](https://www.kaggle.com/datasets/erikbruin/airbnb-amsterdam/data?select=reviews_details.csv). The data set is mainly talk about airbnb lisitngs and there customer review for all the properties in the Amsterdam.

The dataset consists of the following files:

* **calendar.csv**: Contains data on the availability of each listing, including dates and whether the listing is available or booked.
* **listings.csv**: Provides detailed information about each Airbnb listing, such as price, location, number of bedrooms, and host details.
* **listings\_details.csv**: Offers additional characteristics for each listing, including room type, amenities, and other attributes not found in the main **listings.csv** file.
* **neighbourhoods.csv**: Contains data about the various neighborhoods in Amsterdam where Airbnb listings are located, useful for location-based analysis.
* **neighbourhoods.geojson**: A geo-spatial file representing the geographical boundaries of Amsterdam’s neighborhoods, ideal for mapping or spatial analysis.
* **reviews.csv**: Includes review data for each listing, such as review text, review scores, and other related information.

**Data Overview:**

The dataset includes the following files:

* **calendar.csv:** The calendar has 365 records for each listing. It specifies the whether the listing is available on a particular day (365 days ahead), and the price on that day.
* **listings.csv:** A listing is basically an advertisement. This file holds the most useful variables that can be used visualizations.
* **listings\_details.csv:** This file holds the same variables as the listing file plus 80 additional variables.
* **neighbourhoods.csv:** Simple file with the Dutch names of the neighbouhoods
* **neighbourhoods.geojson:** This is the shape file that can be used in conjunction with interactive maps (such as Leaflet for R of the Python folium package).
* **reviews.csv:** This is a simple file that can be used to count the number of reviews by listing (for a specific period).
* **reviews\_details.csv:** This file holds the full details of all reviews, and can also be used for instance for text mining.

**Data Limitations:**

* The dataset focuses on data collected exclusively on **December 6th, 2018**, which does not account for changes to Airbnb listings, prices, or reviews that occurred after this date, particularly during the post-COVID era. Since the hospitality industry has undergone significant transformations due to the pandemic, this dataset may not accurately reflect current trends in Amsterdam's Airbnb market.

**Reason for Choosing This Data:**

As a frequent traveler, I’ve observed that making the right decisions when booking hotels and flights can significantly save both time and money. This observation inspired me to explore this dataset, as it provides valuable insights into the hospitality industry, helping identify trends and opportunities for cost and time optimization.

**Ethical Considerations:**

When working with the Airbnb dataset, it is crucial to address several ethical considerations. First, privacy and confidentiality should be prioritized, ensuring that any **personal information(PII data)** in customer reviews or listing details is anonymized. Since the dataset includes reviews, which can reflect personal experiences, it is essential to avoid any misuse of this data by respecting the individuals’ privacy. Bias and fairness should also be considered, as the dataset may reflect biases based on cultural, socio-economic, or geographic factors. Informed consent should be taken into account, ensuring that individuals whose data is included are aware of how it will be used. Additionally, transparency in data collection and analysis is essential, including acknowledging any limitations or assumptions made during the process. Lastly, the potential for data manipulation or misrepresentation should be minimized to prevent harm to hosts or guests.

**Questions to Explore:**

Here are the top 7-8 questions for hypothesis and analysis based on the dataset:

1. How does the price of listings vary across different neighborhoods in Amsterdam?
2. Is there a correlation between the type of room (e.g., entire home, private room) and the price of the listing?
3. What is the relationship between the number of reviews per month and the price of the listing?
4. How does the host's listing count affect the average reviews per month?
5. Does the availability of listings correlate with their price and review count?
6. Which neighborhoods have the highest-priced listings, and how does this compare with other variables like room type and review count?
7. Do hosts with multiple listings charge higher prices or receive more reviews compared to hosts with single listings?
8. How can sentiment analysis on customer reviews provide insights into guest experiences and their relationship with room types or pricing?

**Data Wrangling & Profiling:**

**Listings.csv**

* Found 38 **missing values** in the name column which are replaced with unkown.
* Found 2406 **missing values** in the reviews\_per\_month & last\_review cloumns which are replaced with unkown.
* No Inconsistencies found in the categorical columns.
* The 'minimum\_nights' column has been analyzed, and a **high-value outlier** of 1001 nights was detected. Since such an extreme value is unrealistic for minimum nights, it has been removed from the dataset to ensure more accurate analysis.
* No **Duplicate values** were found in the Dataset

|  |  |
| --- | --- |
| Listings |  |
| id | **int64** |
| name | **object** |
| host\_id | **int64** |
| host\_name | **object** |
| Neighbourhood\_Group | **Float64** |
| Neighbourhood | **Object** |
| Latitude | **Float64** |
| Longitude | **Float64** |
| Room\_Type | **Object** |
| Price | **Int64** |
| Minimum\_Nights | **Int64** |
| Number\_Of\_Reviews | **Int64** |
| Last\_Review | **Object** |
| Reviews\_Per\_Month | **Object** |
| Calculated\_Host\_Listings\_Count | **Int64** |
| Availability\_365 | **Int64** |

**Listings\_details.csv Column Overview :**

|  |  |  |
| --- | --- | --- |
| Column Name | Description | Variable Type |
| 1. id | Unique identifier for each listing | int64 |
| 1. listing\_url | URL of the listing | object |
| 1. scrape\_id | Identifier for the scrape job | int64 |
| 1. last\_scraped | Date the listing was last scraped | object |
| 1. name | Name of the listing | object |
| 1. summary | Short description of the listing | object |
| 1. space | Information about the space | object |
| 1. description | Detailed description of the listing | object |
| 1. experiences\_offered | Experiences available through listing | object |
| 1. neighborhood\_overview | Overview of the neighborhood | object |
| 1. notes | Additional notes about the listing | object |
| 1. transit | Information about local transit | object |
| 1. access | Access instructions for the listing | object |
| 1. interaction | Details on host-guest interaction | object |
| 1. house\_rules | Rules for guests | object |
| 1. thumbnail\_url | URL for the thumbnail image | float64 |
| 1. medium\_url | URL for medium-sized image | float64 |
| 1. picture\_url | URL for the picture of the listing | object |
| 1. xl\_picture\_url | URL for extra-large image | float64 |
| 1. host\_id | Unique identifier for the host | int64 |
| 1. host\_url | URL of the host's profile | object |
| 1. host\_name | Name of the host | object |
| 1. host\_since | Date when the host joined | object |
| 1. host\_location | Location of the host | object |
| 1. host\_about | Host's description or profile | object |
| 1. host\_response\_time | Response time to guest inquiries | object |
| 1. host\_response\_rate | Response rate of the host | object |
| 1. host\_acceptance\_rate | Host's acceptance rate | float64 |
| 1. host\_is\_superhost | Whether the host is a superhost | object |
| 1. host\_thumbnail\_url | URL for host's thumbnail | object |
| 1. host\_picture\_url | URL for host's profile picture | object |
| 1. host\_neighbourhood | Host's neighborhood | object |
| 1. host\_listings\_count | Number of listings the host has | float64 |
| 1. host\_total\_listings\_count | Total number of listings by the host | float64 |
| 1. host\_verifications | Verification details for the host | object |
| 1. host\_has\_profile\_pic | Whether the host has a profile picture | object |
| 1. host\_identity\_verified | Whether the host's identity is verified | object |
| 1. street | Street address of the listing | object |
| 1. neighbourhood | Neighborhood of the listing | object |
| 1. neighbourhood\_cleansed | Cleaned-up version of neighborhood | object |
| 1. neighbourhood\_group\_cleansed | Cleaned-up group for neighborhood | float64 |
| 1. city | City of the listing | object |
| 1. state | State of the listing | object |
| 1. zipcode | Zip code of the listing | object |
| 1. market | Market where the listing is located | object |
| 1. smart\_location | Smart location flag (e.g., geolocation) | object |
| 1. country\_code | Code of the country | object |
| 1. country | Country of the listing | object |
| 1. latitude | Latitude coordinate of the listing | float64 |
| 1. longitude | Longitude coordinate of the listing | float64 |
| 1. is\_location\_exact | Whether the location is exact | object |
| 1. property\_type | Type of the property (e.g., apartment) | object |
| 1. room\_type | Type of room (e.g., entire home) | object |
| 1. accommodates | Number of people the property can accommodate | int64 |
| 1. bathrooms | Number of bathrooms | float64 |
| 1. bedrooms | Number of bedrooms | float64 |
| 1. beds | Number of beds | float64 |
| 1. bed\_type | Type of bed (e.g., king, queen) | object |
| 1. amenities | List of amenities available | object |
| 1. square\_feet | Size of the property in square feet | float64 |
| 1. price | Price of the listing | object |
| 1. weekly\_price | Weekly price for the listing | object |
| 1. monthly\_price | Monthly price for the listing | object |
| 1. security\_deposit | Security deposit required | object |
| 1. cleaning\_fee | Cleaning fee charged | object |
| 1. guests\_included | Number of guests included | int64 |
| 1. extra\_people | Extra charge for additional people | object |
| 1. minimum\_nights | Minimum number of nights to stay | int64 |
| 1. maximum\_nights | Maximum number of nights to stay | int64 |
| 1. calendar\_updated | Date when calendar was last updated | object |
| 1. has\_availability | Whether the listing has availability | object |
| 1. availability\_30 | Availability in the next 30 days | int64 |
| 1. availability\_60 | Availability in the next 60 days | int64 |
| 1. availability\_90 | Availability in the next 90 days | int64 |
| 1. availability\_365 | Availability in the next 365 days | int64 |
| 1. calendar\_last\_scraped | Date when the calendar was last scraped | object |
| 1. number\_of\_reviews | Number of reviews | int64 |
| 1. first\_review | Date of the first review | object |
| 1. last\_review | Date of the last review | object |
| 1. review\_scores\_rating | Overall rating based on reviews | float64 |
| 1. review\_scores\_accuracy | Rating for accuracy of the listing | float64 |
| 1. review\_scores\_cleanliness | Rating for cleanliness of the listing | float64 |
| 1. review\_scores\_checkin | Rating for check-in experience | float64 |
| 1. review\_scores\_communication | Rating for communication | float64 |
| 1. review\_scores\_location | Rating for location | float64 |
| 1. review\_scores\_value | Rating for value of the listing | float64 |
| 1. requires\_license | Whether a license is required | object |
| 1. license | License information | object |
| 1. jurisdiction\_names | Names of legal jurisdictions | object |
| 1. instant\_bookable | Whether the listing is instant bookable | object |
| 1. is\_business\_travel\_ready | Whether the listing is business ready | object |
| 1. cancellation\_policy | Cancellation policy for the listing | object |
| 1. require\_guest\_profile\_picture | Whether guest profile picture is required | object |
| 1. require\_guest\_phone\_verification | Whether guest phone verification is required | object |
| 1. calculated\_host\_listings\_count | Total number of listings by the host | int64 |
| 1. reviews\_per\_month | Average number of reviews per month | float64 |

**Droped all Columns from List\_detail.csv file:**

Here is the list of columns :

* 'host\_id'
* 'host\_url'
* 'host\_name'
* 'host\_since'
* 'host\_location'
* 'host\_about'
* 'host\_response\_time'
* 'host\_response\_rate'
* 'host\_acceptance\_rate'
* 'host\_is\_superhost'
* 'scrape\_id'
* 'last\_scraped'
* 'notes'
* 'transit'
* 'interaction'
* 'thumbnail\_url'
* 'medium\_url'
* 'xl\_picture\_url'
* 'host\_picture\_url'
* 'first\_review'
* 'last\_review'
* 'jurisdiction\_names'
* 'instant\_bookable'
* 'is\_business\_travel\_ready'
* 'require\_guest\_profile\_picture'
* 'require\_guest\_phone\_verification'
* 'host\_thumbnail\_url'
* 'host\_neighbourhood'
* 'host\_listings\_count'
* 'host\_total\_listings\_count'
* 'host\_verifications'
* 'host\_has\_profile\_pic'
* 'host\_identity\_verified'
* 'summary'
* 'listing\_url'
* 'description'
* 'house\_rules'
* 'picture\_url'
* 'market'
* 'calendar\_last\_scraped'
* 'reviews\_per\_month'

"**A total of 41 columns were removed as they contained sensitive information primarily related to the host and internal use cases, which could potentially violate PII (Personally Identifiable Information) regulations.**"

"No changes were made to the **neighborhood, calendar,** and **review datasets** as the information was compliant with privacy regulations and can be merged."

| **Column** | **Non-Null Count** | **Data Type** | **Summary** |
| --- | --- | --- | --- |
| neighbourhood\_group | 0 | float64 | All values are missing (NaN). |
| neighbourhood | 22 | object | All values are non-null and valid. |

**Additional Insights:**

* **Total Rows:** 22
* **Total Columns:** 2
* **Memory Usage:** 484.0+ bytes

Here’s the data profiling information for the Review\_details dataset in tabular format:

| **Column** | **Non-Null Count** | **Data Type** | **Summary** |
| --- | --- | --- | --- |
| listing\_id | 431830 | int64 | No missing values, numerical identifier for listings. |
| id | 431830 | int64 | No missing values, numerical identifier. |
| date | 431830 | object | No missing values, date of the review. |
| reviewer\_id | 431830 | int64 | No missing values, numerical identifier for reviewers. |
| reviewer\_name | 431830 | object | No missing values, name of the reviewer. |
| comments | 431300 | object | 530 missing values, textual comments from reviewers. |

**Additional Insights:**

* **Total Rows:** 431,830
* **Total Columns:** 6
* **Missing Values:**
  + The comments column has 530 missing values.

"The following code changes the **530 missing values** in the 'comments' column of the df\_reviews\_details DataFrame to the placeholder string **'Unknown'**, ensuring there are no missing values in this column."