PRICE PREDICTION OF AIRBNB LISTINGS

SUPERVISED BY: DR. YUAN TIAN

GROUP NO.2

SALMA MOHAMED MAHER

YASMIN GAMAL MOHAMED

AGENDA

- Introduction
- Motivation
- Dataset Statistics
- Methodology

Introduction:

Airbnb is an online community marketplace that connects people looking to rent their homes with people who are looking for accommodations.

This dataset focuses on Airbnb listings and activities in New York City, a prime destination for both tourists and business travelers.

Motivation:

The primary goal of analyzing this dataset is to gain insights related to pricing and guest satisfaction. This can help hosts optimize their listings, guests find suitable accommodations, and Airbnb as a platform to enhance its services.

Dataset Statistics:

The dataset is a collection of Airbnb listings. It contains various attributes related to Airbnb properties, hosts, and user interactions.

The Airbnb dataset consists of 102,599 entries and 28 features.

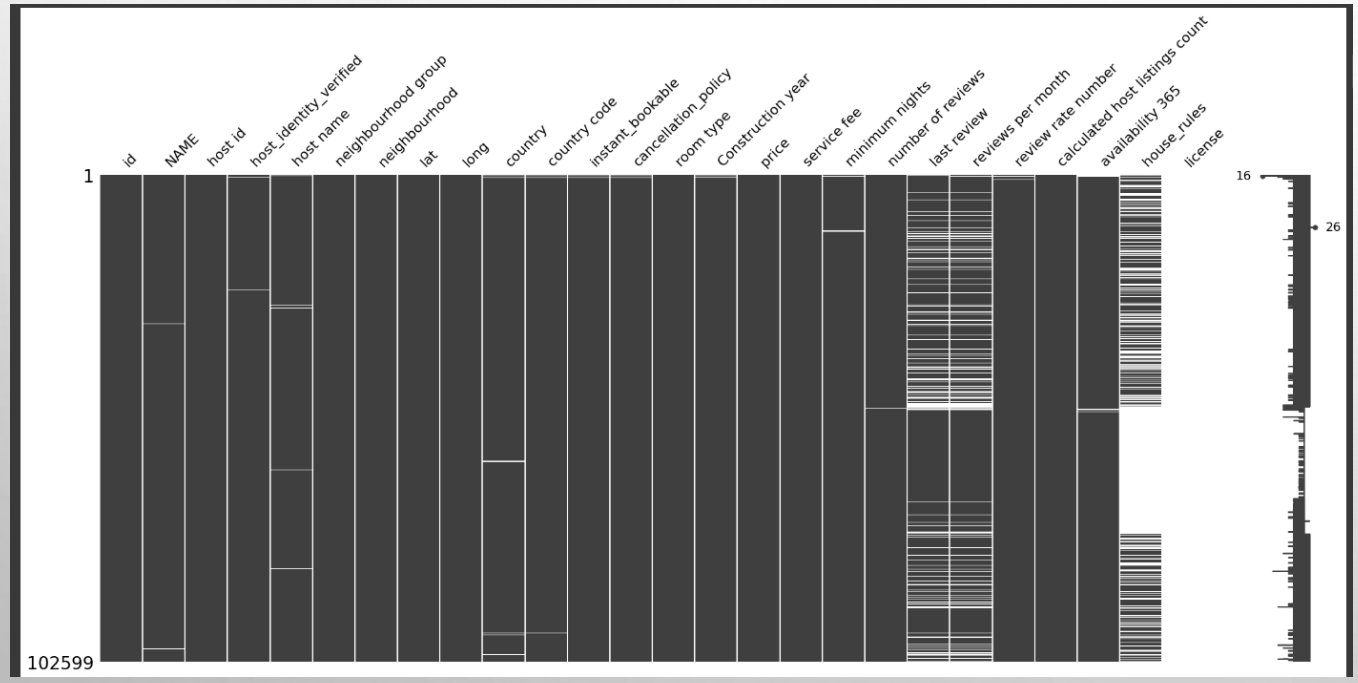
Key Features of the Dataset:

1. Listings details
2. Geographical Information
3. Pricing Information
4. Booking and Availability
5. Review Information

Methodology:

1. Data Cleaning and Preparation.

- Handling Missing Values:




```
license                           102597
house_rules                        52131
last review                        15893
reviews per month                  15879
country                              532
availability 365                     448
minimum nights                       409
host name                            406
review rate number                   326
calculated host listings count       319
host_identity_verified               289
service fee                          273
NAME                                 250
price                                247
Construction year                    214
number of reviews                    183
country code                         131
instant_bookable                     105
cancellation_policy                   76
neighbourhood group                   29
neighbourhood                         16
long                                   8
lat                                    8
id                                     0
host id                                0
room type                              0
dtype: int64
```

1-Mean/Median Imputation

- Price
- Service Fee Column
- Number of reviews

2-Mode Imputation

- Review rate number

3-Deletion

- License (it has only 2 values and all other values are missing)
- Country-Country Code (they have only one unique value)
- Id, host-id, host name

4-Imputation Based on Existing Relationships

- Calculated host listings count Column
- Neighborhood group Column

```
# Filter rows with missing 'neighbourhood group'
missing_neighbourhood_group = airbnb_raw_data[airbnb_raw_data['neighbourhood group'].isnull()]

# Display latitude and longitude,neighbourhood for rows with missing 'neighbourhood group'
print(missing_neighbourhood_group[['id', 'lat', 'long','neighbourhood']])

# Create the mapping between neighbourhood group and neighbourhood
neighbourhood_and_neighbourhood_group = airbnb_raw_data.dropna(subset=['neighbourhood group']).set_index('neighbourhood')['neighbourhood group'].to_dict()
neighbourhood_and_neighbourhood_group
```

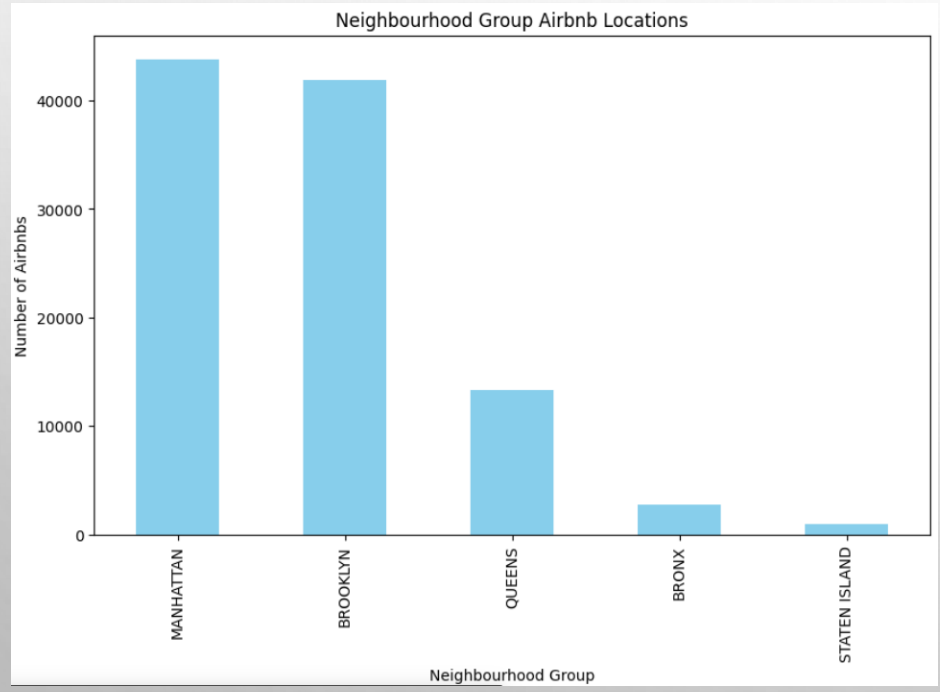
```
# Function to fill missing 'neighbourhood group'
def fill_neighbourhood_group(row):
    if pd.isna(row['neighbourhood group']):
        return neighbourhood_and_neighbourhood_group.get(row['neighbourhood'], None)
    return row['neighbourhood group']

# Apply the function to fill missing values
airbnb_raw_data['neighbourhood group'] = airbnb_raw_data.apply(fill_neighbourhood_group, axis=1)
```

```
#check again for missing data in neighbourhood group
airbnb_raw_data['neighbourhood group'].isnull().sum()
```

```
0
```

2-DATA VISUALIZATION

1- WHAT IS THE DISTRIBUTION OF LISTINGS ACROSS DIFFERENT NEIGHBOURHOOD GROUPS?



2- WHAT IS THE DISTRIBUTION OF ROOM TYPES?