Project Report: AIRBNB Market Analysis

Mukand Krishna **6/25/23**

1. Abstract:

This report presents a comprehensive analysis of the Airbnb market using Python. The analysis includes data cleaning, exploratory data analysis (EDA), and prediction on pricing feature using linear regression models. Additionally, a Tableau dashboard was created to provide visual insights into the data. The purpose of this report is to showcase the findings and insights obtained from the analysis.

2. Introduction:

The Airbnb market has grown significantly in recent years, providing an alternative accommodation option for travelers. This analysis aims to gain insights into the Airbnb market by examining various aspects such as pricing, reviews, and host behavior.

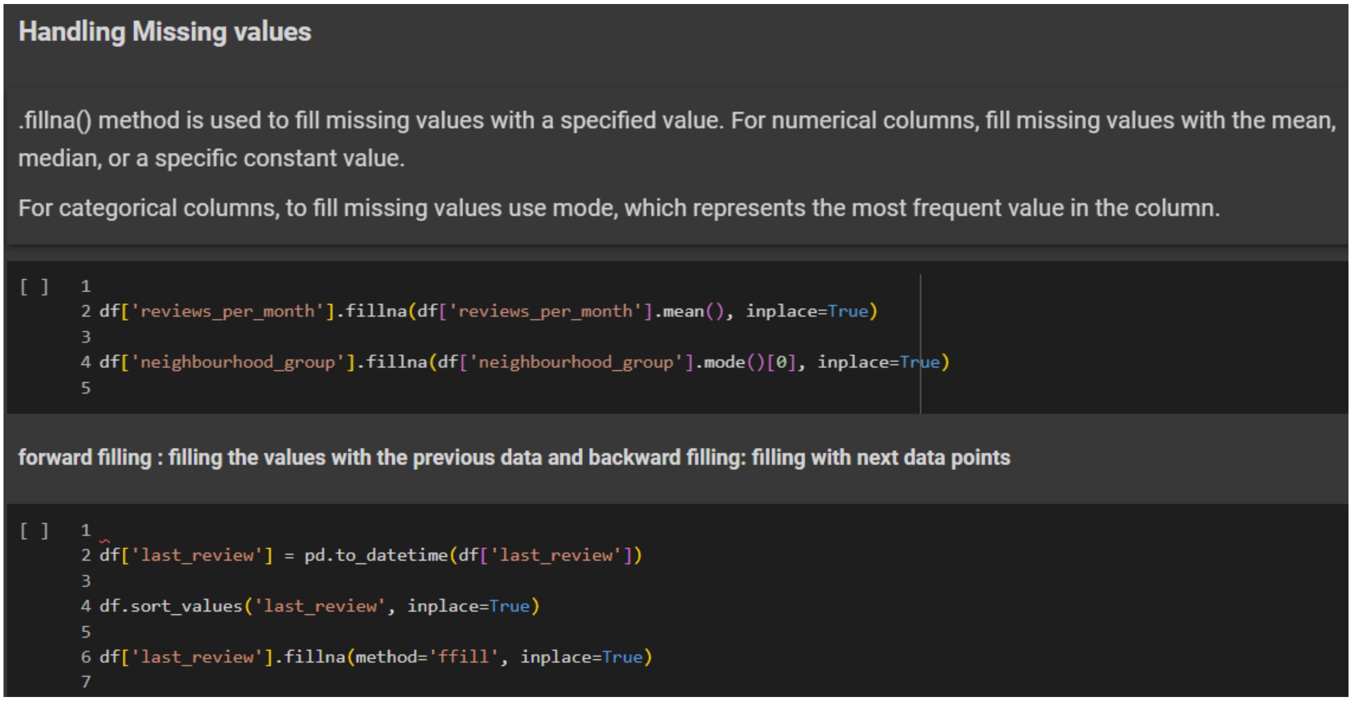
2.1 Objectives:

The main objectives of the project were as follows:

- Analyze the CSV file of AIRBNB 2023.
- Clean and preprocess the data to ensure data quality.
- Visualize the data using charts and graphs for analysis.
- Prediction on price features.

3. Data Cleaning:

The collected data underwent a thorough cleaning process, including handling missing values, and removing outliers. This ensured the accuracy and reliability of the analysis.



Handling Missing values

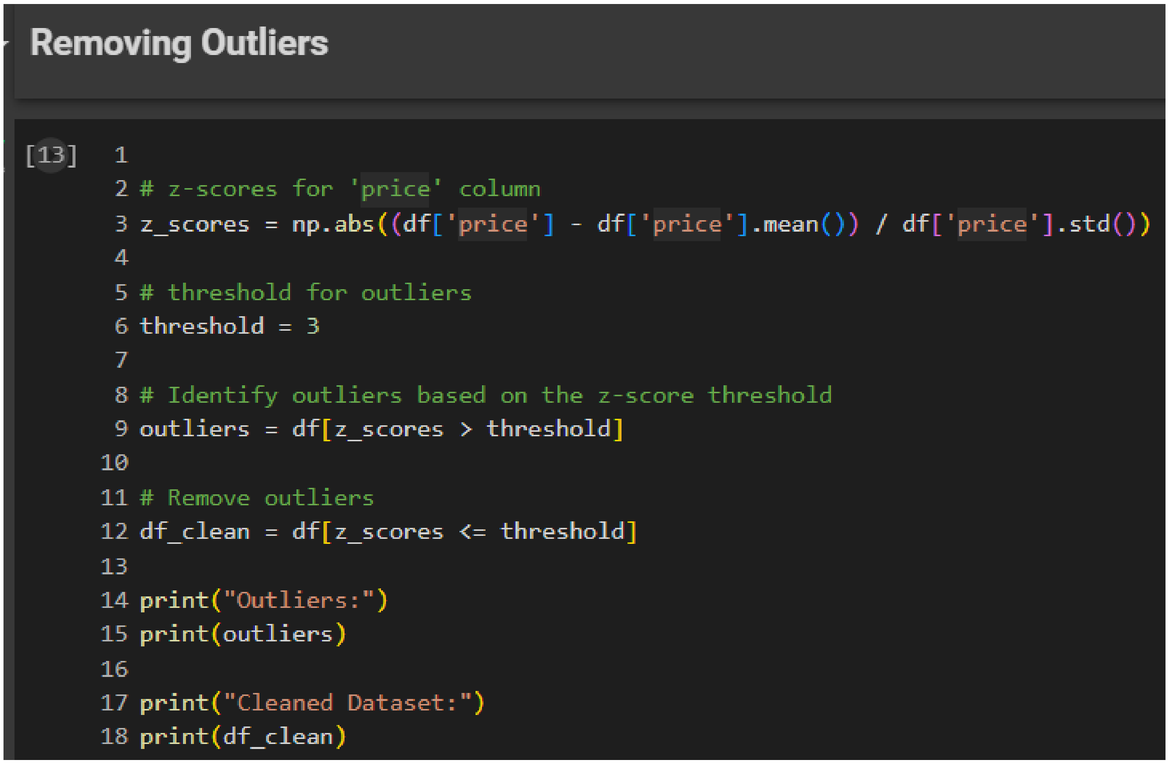
.fillna() method is used to fill missing values with a specified value. For numerical columns, fill missing values with the mean, median, or a specific constant value.

For categorical columns, to fill missing values use mode, which represents the most frequent value in the column.

```

df['reviews_per_month'].fillna(df['reviews_per_month'].mean(), inplace=True)

df['neighbourhood_group'].fillna(df['neighbourhood_group'].mode()[0], inplace=True)

```

forward filling : filling the values with the previous data and backward filling: filling with next data points

```

df['last_review'] = pd.to_datetime(df['last_review'])

df.sort_values('last_review', inplace=True)

df['last_review'].fillna(method='ffill', inplace=True)

```



Removing Outliers

```python

# z-scores for 'price' column
z_scores = np.abs((df['price'] - df['price'].mean()) / df['price'].std())

# threshold for outliers
threshold = 3

# Identify outliers based on the z-score threshold
outliers = df[z_scores > threshold]

# Remove outliers
df_clean = df[z_scores <= threshold]

print("Outliers:")
print(outliers)

print("Cleaned Dataset:")
print(df_clean)
```

4. Exploratory Data Analysis(EDA):

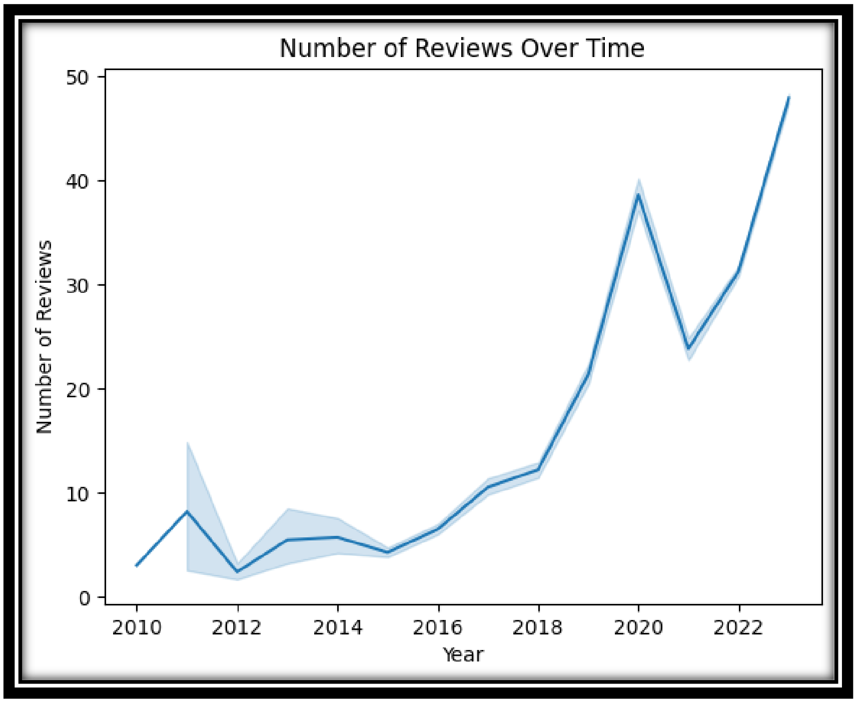
During the EDA phase, several key variables were analyzed to understand their distributions and relationships. The following analyses were conducted:

- Distribution Analysis: Examining the distribution of prices, ratings, and number of reviews.
- Temporal Analysis: Investigating temporal patterns and trends in bookings and availability.
- Property Types Analysis: Analyzing the distribution and popularity of different property types.
- Heatmap to visualize the relationships and correlations between variables.

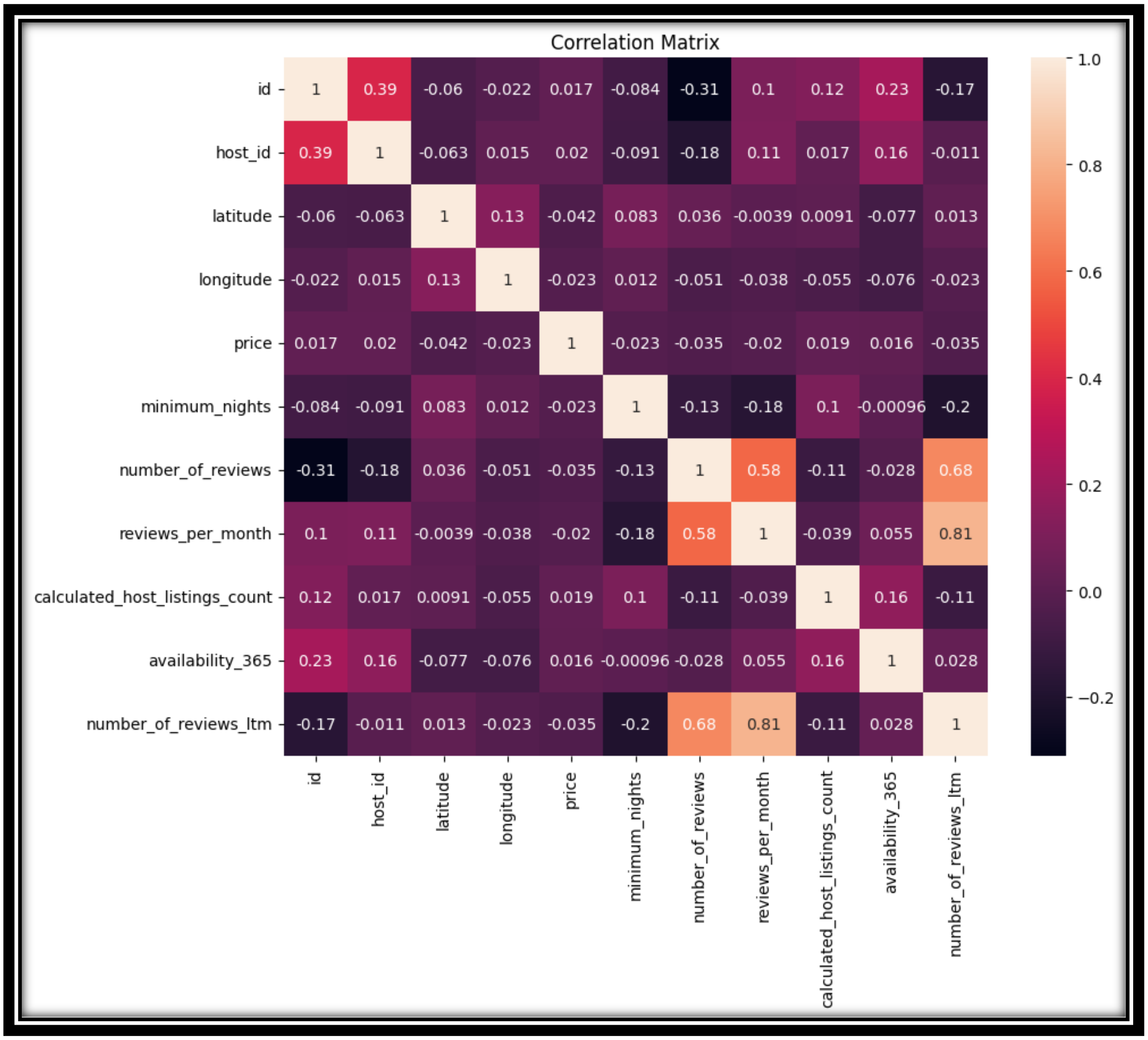
5. Graphical Representation of Data:

Few graphs are shown below:

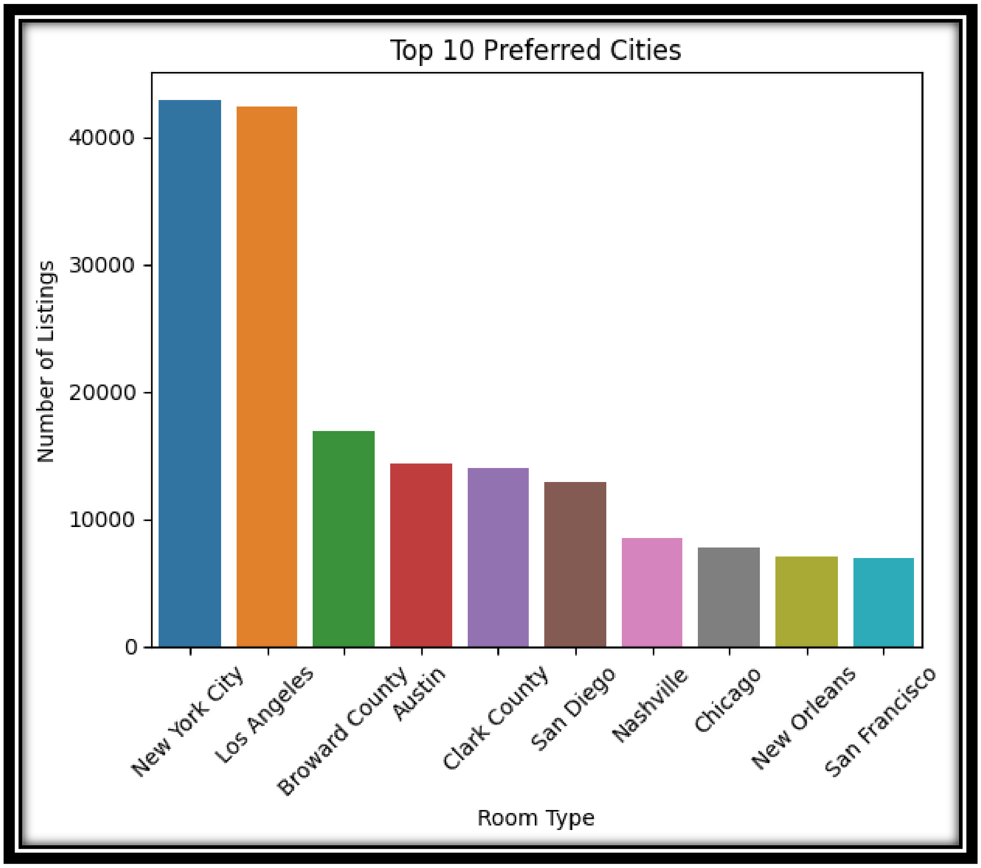
Line Graph b/w reviews and year



Relationship b/w variables



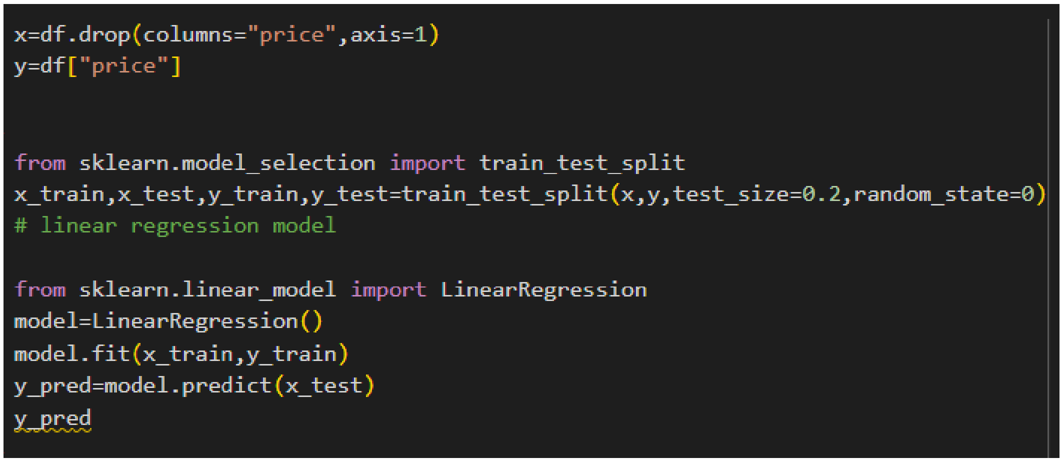
Bar chart of cities



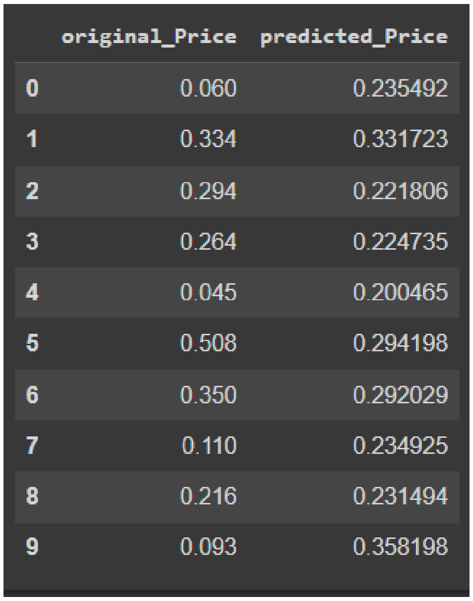
6. Linear Regression model:

Linear regression models were developed to predict price feature of the Airbnb market.

80% of data was used for training and 20% was used as testing, then the difference b/w the original and predicted price was printed.

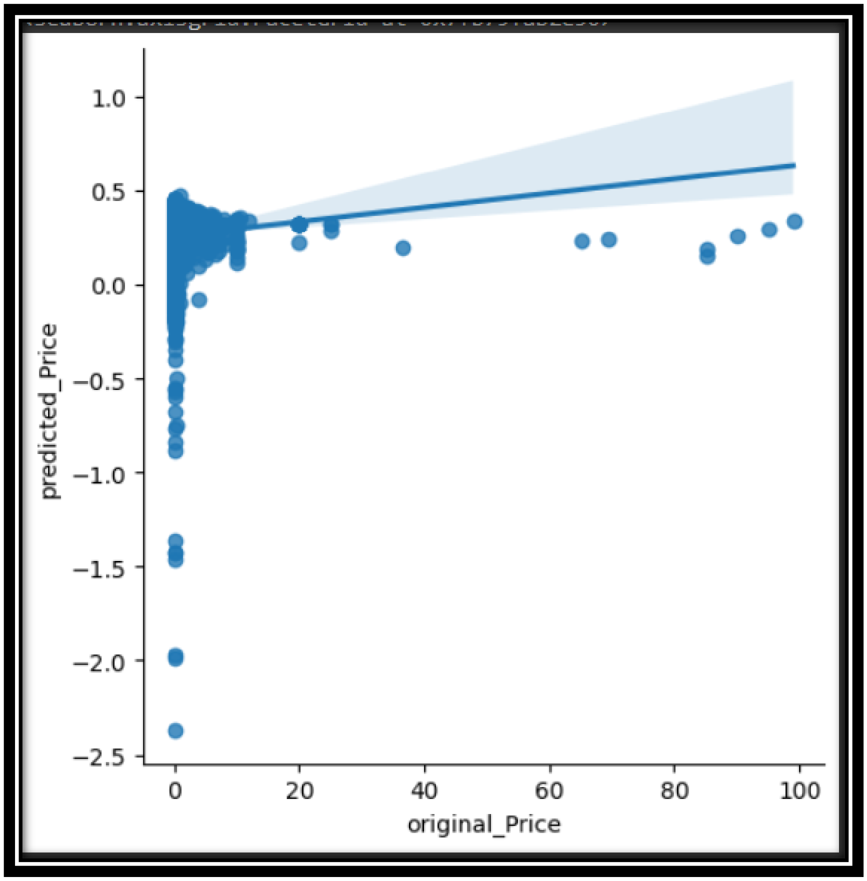


```
x=df.drop(columns="price",axis=1)
y=df["price"]


from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=0)
# linear regression model

from sklearn.linear_model import LinearRegression
model=LinearRegression()
model.fit(x_train,y_train)
y_pred=model.predict(x_test)
y_pred
```

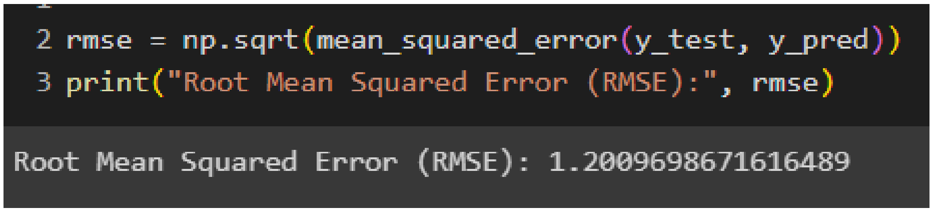


	original_Price	predicted_Price
0	0.060	0.235492
1	0.334	0.331723
2	0.294	0.221806
3	0.264	0.224735
4	0.045	0.200465
5	0.508	0.294198
6	0.350	0.292029
7	0.110	0.234925
8	0.216	0.231494
9	0.093	0.358198

Prediction of price is performed, and scatter graph is plotted b/w actual and predicted price.



Root Mean Squared Error is used to measure the accuracy of a predictive model, in regression analysis. It represents the square root of the average squared difference between the predicted values and the actual values. It quantifies the average deviation between the predicted values and the observed values. It tells how well the **model fits the data**, and lower RMSE indicates better predictive performance.

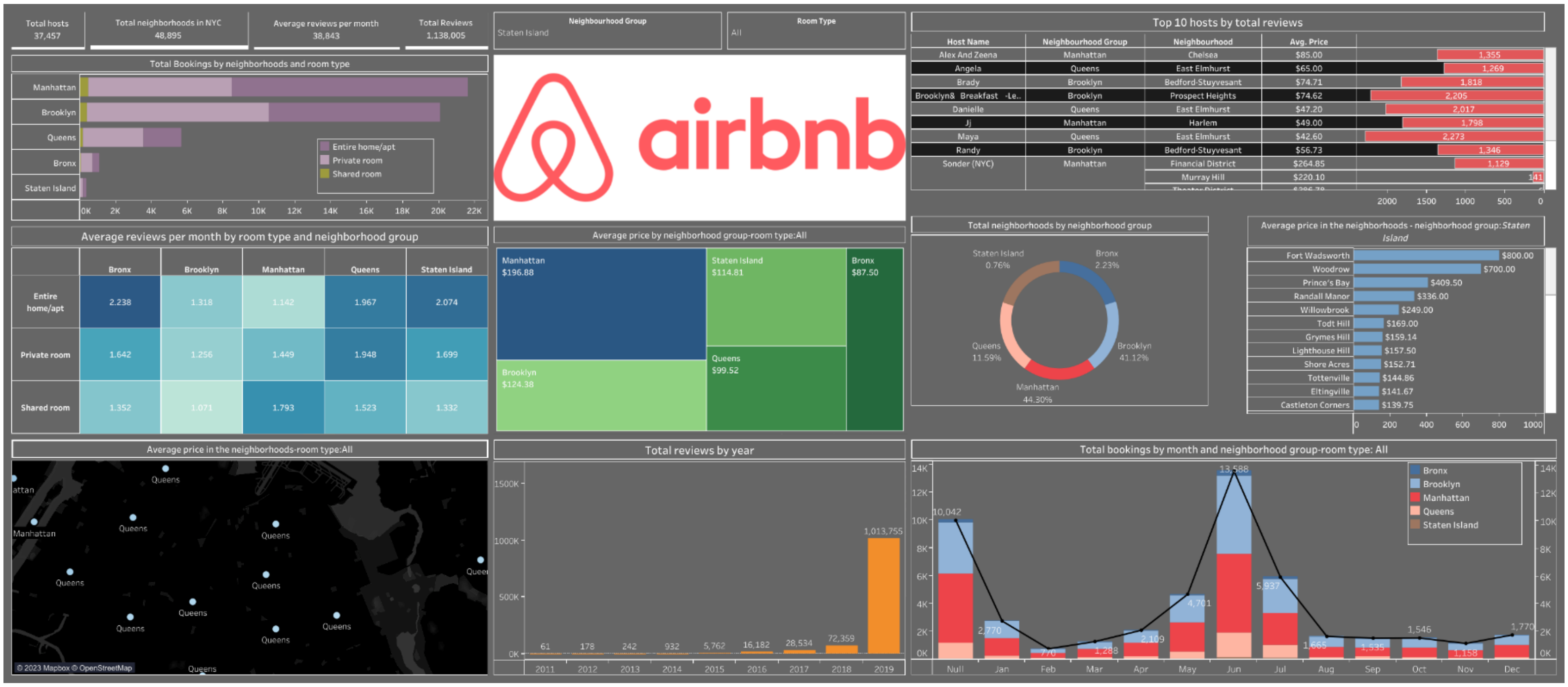


```
rmse = np.sqrt(mean_squared_error(y_test, y_pred))
print("Root Mean Squared Error (RMSE):", rmse)
```

```
Root Mean Squared Error (RMSE): 1.2009698671616489
```

7. Tableau Dashboard:

The dashboard includes interactive visualizations that allow users to explore the data and gain insights.



Total hosts 37,457

Total neighborhoods in NYC 48,895

Average reviews per month 38,843

Total Reviews 1,138,005

Neighbourhood Group: Staten Island

Room Type: All

Top 10 hosts by total reviews

Host Name	Neighbourhood Group	Neighbourhood	Avg. Price	
Alex And Zeena	Manhattan	Chelsea	$85.00	1,355
Angela	Queens	East Elmhurst	$65.00	1,269
Brady	Brooklyn	Bedford-Stuyvesant	$74.71	1,818
Brooklyn& Breakfast -Le..	Brooklyn	Prospect Heights	$74.62	2,205
Danielle	Queens	East Elmhurst	$47.20	2,017
Jj	Manhattan	Harlem	$49.00	1,798
Maya	Queens	East Elmhurst	$42.60	2,273
Randy	Brooklyn	Bedford-Stuyvesant	$56.73	1,346
Sonder (NYC)	Manhattan	Financial District	$264.85	1,129
		Murray Hill	$220.10	141

Average reviews per month by room type and neighborhood group

	Bronx	Brooklyn	Manhattan	Queens	Staten Island
Entire home/apt	2.238	1.318	1.142	1.967	2.074
Private room	1.642	1.256	1.449	1.948	1.699
Shared room	1.352	1.071	1.793	1.523	1.332

Average price in the neighborhoods - neighborhood group: *Staten Island*

Fort Wadsworth	$800.00
Woodrow	$700.00
Prince's Bay	$409.50
Randall Manor	$336.00
Willowbrook	$249.00
Todt Hill	$169.00
Grymes Hill	$159.14
Lighthouse Hill	$157.50
Shore Acres	$152.71
Tottenville	$144.86
Eltingville	$141.67
Castleton Corners	$139.75

8. Conclusion:

In summary, by using the information obtained from analyzing the Airbnb 2019 data, hosts can make smarter choices about how much to charge for their listings, what type of rooms to offer, how to keep guests happy, and how to expand their business strategically. These recommendations are intended to improve the overall experience for both hosts and guests, which should result in more bookings, happier guests, and ultimately, more money earned.