LOVELY PROFESSIONAL UNIVERSITY

INT353

B-TECH CSE 3rd YEAR (DATA SCIENCE -AI & ML)

NAME: Susnata Das Mahapatra

REGISTRATION NO:12111532

SECTION: K21UP

ROLL NO: RK21UPA11

MY DATASET: Singapore Airbnb Dataset

SIGNATURE

1. INTRODUCTION OF THE DATABASE:

Singapore, a vibrant and dynamic city-state in Southeast Asia, is a global hub for business, tourism, and culture. As a popular travel destination, Singapore offers a diverse range of accommodations, including hotels, hostels, and the ever-expanding market of Airbnb rentals. The "Singapore Airbnb Database" is a comprehensive collection of data that provides valuable insights into the world of short-term vacation rentals in this thriving city.

This database captures a snapshot of the Singaporean Airbnb ecosystem, encompassing various aspects of Airbnb listings, host information, pricing trends, and guest reviews. It offers an opportunity to delve into the details of the local short-term rental market and conduct exploratory data analysis (EDA) to uncover patterns, trends, and interesting findings.

2. WHY I HAVE CHOSEN THIS DATABASE:

1. **Relevance to Your Interests:** I have a personal interest in travel, tourism, or the hospitality industry, and exploring the Singapore Airbnb database aligns with my passion and curiosity.

2. **Data Availability:** The availability of comprehensive and well-structured data is a crucial factor in any data analysis project. The Singapore Airbnb database offers a rich source of data with various variables, making it suitable for in-depth analysis.

3. **Data Science Skills Development:** Working with this dataset allows me to practice and enhance my data analysis and visualization skills, which are valuable in various professional contexts.

4. **Academic Research:** I am a student. So, the Singapore Airbnb dataset can be a valuable resource for academic studies related to tourism, economics, urban planning, or data analysis.

3.DOMAIN:

The domain of the "Singapore Airbnb Database" is primarily related to the hospitality and tourism industry within the context of Singapore

1. **Tourism Domain:** This dataset is directly related to the broader hospitality and tourism industry, which encompasses various aspects of travel, accommodations, and visitor experiences.

2. **Short-Term Vacation Rentals**: The dataset specifically focuses on short-term vacation rentals offered by hosts on the Airbnb platform. It includes details about properties, hosts, pricing, and guest reviews for such accommodations.

3. **Singapore Location Domain:** The dataset is geographically limited to Singapore, providing insights into the short-term rental market in this city-state. It may be used to understand tourism trends, accommodation choices, and pricing strategies in Singapore.

4. **Data Analysis and Data Science:** Beyond the industry-specific domains, the dataset falls within the domain of data analysis and data science, as it serves as a valuable resource for individuals looking to analyze and draw insights from structured data.

4.INFORMATION:

This dataset consists of 7907 rows and 16 columns. some columns contain missing or null values,

Column names are:

1. Id
2. Name
3. Host_id
4. Host_name
5. neighbourhood_group
6. neighbourhood
7. latitude
8. longitude
9. room_type
10. price
11. minimum_nights
12. number_of_reviews
13. last_review
14. reviews_per_month
15. calculated_host_listings_count
16. availability_365

QUENTIONS/PLANS:

1. What is the distribution of Airbnb listing prices in Singapore?
2. How does the price vary by room type (e.g., private room, entire home)?
3. What is the average price per neighbourhood in Singapore?
4. Is there a correlation between the number_of_reviews and the price of listings?
5. What is the distribution of the minimum nights required for bookings?
6. How does the availability of listings change over the year (seasonality)?
7. Are there any outliers in the price distribution?
8. What is the average rating score for Airbnb listings in Singapore?

9. Is there a relationship between the number of reviews and the rating score?

10. How many listings are offered by each host?

11. What is the distribution of the size (square footage) of listings?

12. How does the price vary by neighborhood and room type?

13. What is the distribution of the number of bathrooms in listings?

14. How does the price change based on the number of people a listing can accommodate?

15. What is the average price per neighborhood, considering only listings with a high rating (e.g., 4.5 stars and above)?

16. How does the minimum nights requirement vary by neighborhood?

17. What is the distribution of the host response time?

18. Is there a correlation between the price and the number of reviews?

19. How does the price change over the year, considering only listings with a high rating (e.g., 4.5 stars and above)?

20. What are the most common amenities offered in Airbnb listings?

LIBRARIES USED :

NumPy:

NumPy is a Python library for numerical computations.
It provides support for handling large arrays and matrices efficiently.
Offers a wide range of mathematical functions for array operations.
Fundamental for scientific computing and data analysis tasks.

Pandas:

Pandas is a Python library for data manipulation and analysis.
It introduces two primary data structures, DataFrames and Series.
Simplifies data cleaning, filtering, aggregation, and transformation.
Essential for data preprocessing and data exploration.

Matplotlib:

Matplotlib is a Python library for creating static, animated, or interactive visualizations.

It offers extensive options for generating various types of plots and charts.
Customizable to create publication-quality visualizations.
Widely used for data visualization and reporting.

Seaborn:

Seaborn is a data visualization library built on Matplotlib.
Specializes in creating informative and visually appealing statistical graphics.
Simplifies complex plotting tasks with built-in themes and color palettes.
Ideal for exploratory data analysis and presenting data insights.

APPROACHES:

Step 1: Define the Problem

- Clearly define the objectives of your EDA and the specific questions you want to answer.

Step 2: Data Collection

- Obtain the dataset and load it using Pandas (e.g., pd.read_csv()).
- Ensure data integrity and verify that it contains relevant information.

Step 3: Data Understanding

- Perform initial data exploration:
- Check the first few rows with df.head().
- Get the dataset's dimensions using df.shape.
- Inspect data types and missing values with df.info().
- Calculate summary statistics using df.describe().

Step 4: Data Cleaning

- Address missing values using NumPy and Pandas:
- Use np.nan or Pandas' fillna() to impute missing values.
- Remove rows or columns with missing values using df.dropna().
- Remove duplicate rows if they exist (df.drop_duplicates()).
- Handle outliers, if necessary, using NumPy's statistical functions.

Step 5: Feature Engineering

- Create new features or transform existing ones using Pandas:
- Use df.apply() or Pandas' built-in functions.
- Convert data types as needed (e.g., date/time conversion).

Step 6: Data Visualization

- Generate visualizations to gain insights using Matplotlib and Seaborn:
- Create histograms, box plots, and scatter plots to explore distributions and relationships.
- Use bar charts, count plots, and heatmaps for categorical data.
- Construct correlation matrices and use Seaborn's heatmap() for visualizing relationships.
- Create geographic visualizations if applicable.

Step 7: In-Depth Analysis

- Answer specific questions or explore patterns:
- Calculate aggregates (e.g., mean, median, counts) using Pandas' aggregation functions.
- Group and filter data as needed (df.groupby() and df.query()).
- Perform statistical tests using NumPy or other libraries (e.g., t-tests, ANOVA).

Step 8: Reporting and Visualization

- Summarize findings and insights in clear, concise language.
- Create informative visualizations and charts using Matplotlib and Seaborn.
- Provide actionable recommendations or insights based on your analysis.

Step 9: Documentation and Code Sharing

- Document your EDA process, including data cleaning steps, transformations, and analysis methods.
- Share your code and findings with team members or stakeholders.
- Consider using Jupyter Notebooks for creating reproducible reports.

Step 10: Iterate as Necessary

- Be prepared to iterate through the steps as new questions or insights arise during your analysis.

STEPS OF EDA:

Step 1: Importing Necessary Libraries You will typically need the following libraries for EDA in Python:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Step 2: Loading the Dataset Load your Singapore Airbnb dataset into a Pandas DataFrame:

```
df = pd.read_csv("singapore_airbnb_data.csv")
```

Step 3: Initial Data Exploration Start by exploring basic information about your dataset:

```
# Check the first few rows of the dataset
print(df.head())

# Get the shape of the dataset
print(df.shape)

# Check data types and missing values
print(df.info())

# Summary statistics
print(df.describe())
```

OUTPUT:

```
      id                                name  host_id  host_name  \
0  49091    COZICOMFORT LONG TERM STAY ROOM 2   266763  Francesca
1  50646      Pleasant Room along Bukit Timah   227796    Sujatha
2  56334                        COZICOMFORT   266763  Francesca
3  71609  Ensuite Room (Room 1 & 2) near EXPO   367042    Belinda
4  71896      B&B  Room 1 near Airport & EXPO   367042    Belinda

  neighbourhood_group neighbourhood  latitude  longitude     room_type  price  \
0       North Region     Woodlands   1.44255  103.79580  Private room     83
1     Central Region   Bukit Timah   1.33235  103.78521  Private room     81
2       North Region     Woodlands   1.44246  103.79667  Private room     69
3        East Region      Tampines   1.34541  103.95712  Private room    206
4        East Region      Tampines   1.34567  103.95963  Private room     94

   minimum_nights  number_of_reviews last_review  reviews_per_month  \
0             180                  1  2013-10-21               0.01
1              90                 18  2014-12-26               0.28
2               6                 20  2015-10-01               0.20
3               1                 14  2019-08-11               0.15
4               1                 22  2019-07-28               0.22

   calculated_host_listings_count  availability_365
0                               2               365
1                               1               365
2                               2               365
3                               9               353
4                               9               355
(7907, 16)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7907 entries, 0 to 7906
```

```
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              7907 non-null   int64
 1   name                            7905 non-null   object
 2   host_id                         7907 non-null   int64
 3   host_name                       7907 non-null   object
 4   neighbourhood_group             7907 non-null   object
 5   neighbourhood                   7907 non-null   object
 6   latitude                        7907 non-null   float64
 7   longitude                       7907 non-null   float64
 8   room_type                       7907 non-null   object
 9   price                           7907 non-null   int64
 10  minimum_nights                  7907 non-null   int64
 11  number_of_reviews               7907 non-null   int64
 12  last_review                     5149 non-null   object
 13  reviews_per_month               5149 non-null   float64
 14  calculated_host_listings_count  7907 non-null   int64
 15  availability_365                7907 non-null   int64
dtypes: float64(3), int64(7), object(6)
memory usage: 988.5+ KB
None
                 id       host_id     latitude    longitude         price  \
count  7.907000e+03  7.907000e+03  7907.000000  7907.000000   7907.000000
mean   2.338862e+07  9.114481e+07     1.314192   103.848787    169.332996
std    1.016416e+07  8.190910e+07     0.030577     0.043675    340.187599
min    4.909100e+04  2.366600e+04     1.243870   103.646560      0.000000
25%    1.582180e+07  2.305808e+07     1.295795   103.835825     65.000000
50%    2.470627e+07  6.344891e+07     1.311030   103.849410    124.000000
75%    3.234850e+07  1.553811e+08     1.322110   103.872535    199.000000
max    3.811276e+07  2.885676e+08     1.454590   103.973420  10000.000000

       minimum_nights  number_of_reviews  reviews_per_month  \
count     7907.000000        7907.000000        5149.000000
mean        17.510054          12.807386           1.043669
std         42.094616          29.707746           1.285851
min          1.000000           0.000000           0.010000
25%          1.000000           0.000000           0.180000
50%          3.000000           2.000000           0.550000
75%         10.000000          10.000000           1.370000
max       1000.000000         323.000000          13.000000

       calculated_host_listings_count  availability_365
count                     7907.000000      7907.000000
mean                        40.607689       208.726318
std                         65.135253       146.120034
min                          1.000000         0.000000
25%                          2.000000        54.000000
50%                          9.000000       260.000000
75%                         48.000000       355.000000
max                        274.000000       365.000000
```

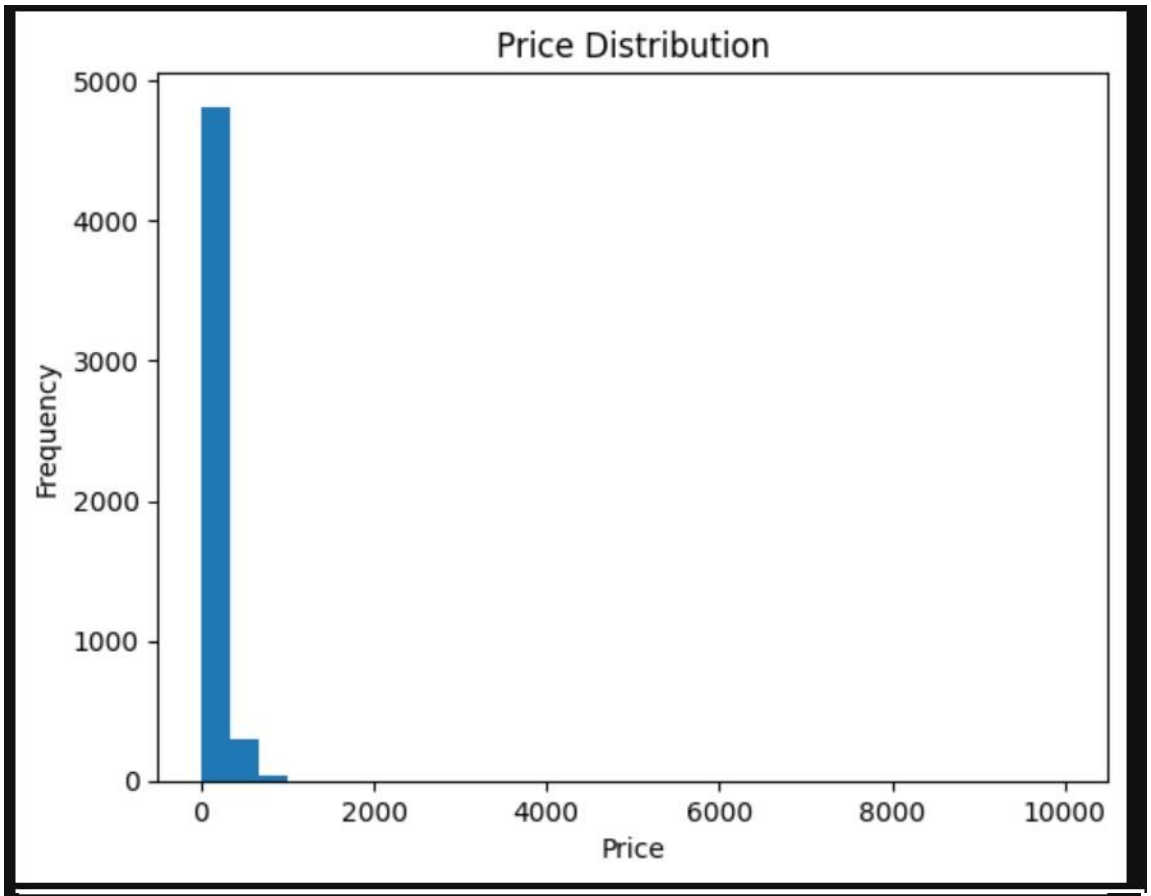
Step 4: Data Cleaning Address missing values, duplicate rows, and any obvious data quality issues:

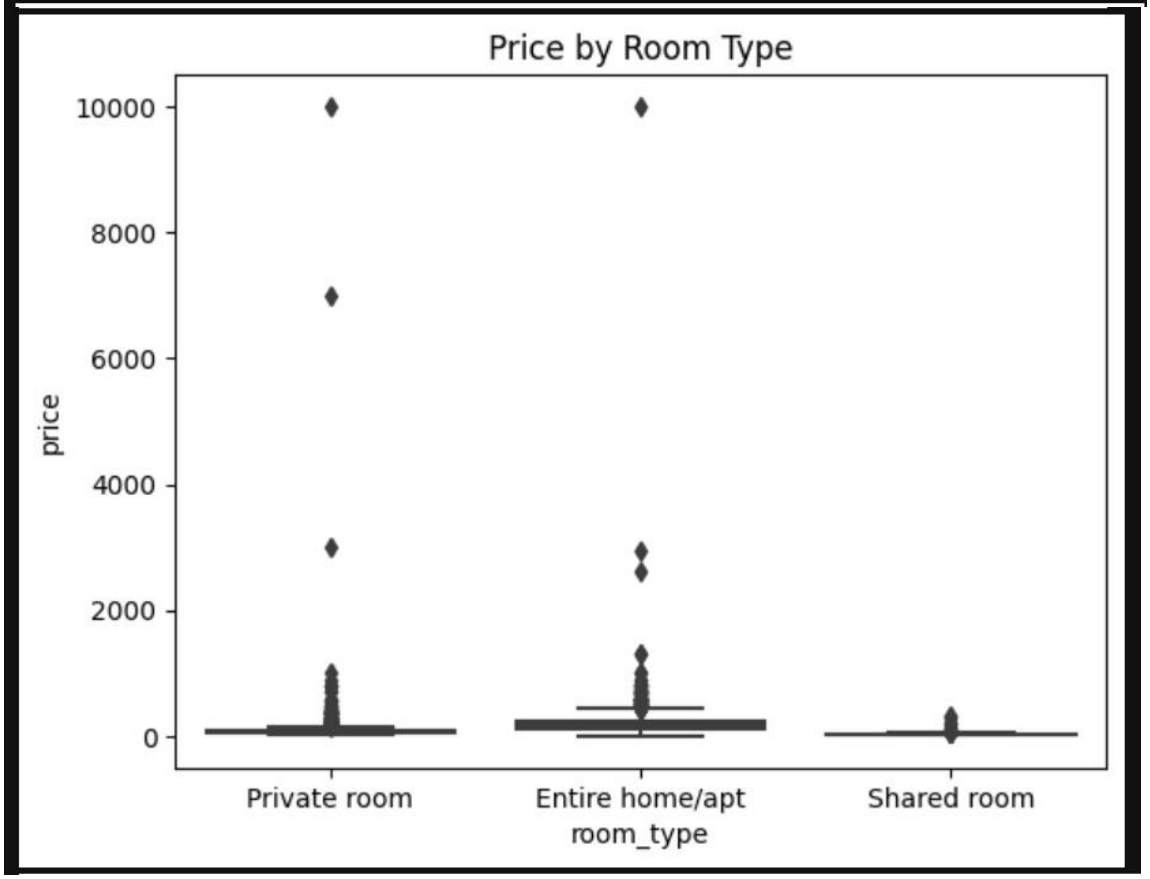
```
# Handle missing values
df.dropna(inplace=True)

# Remove duplicates
df.drop_duplicates(inplace=True)
```

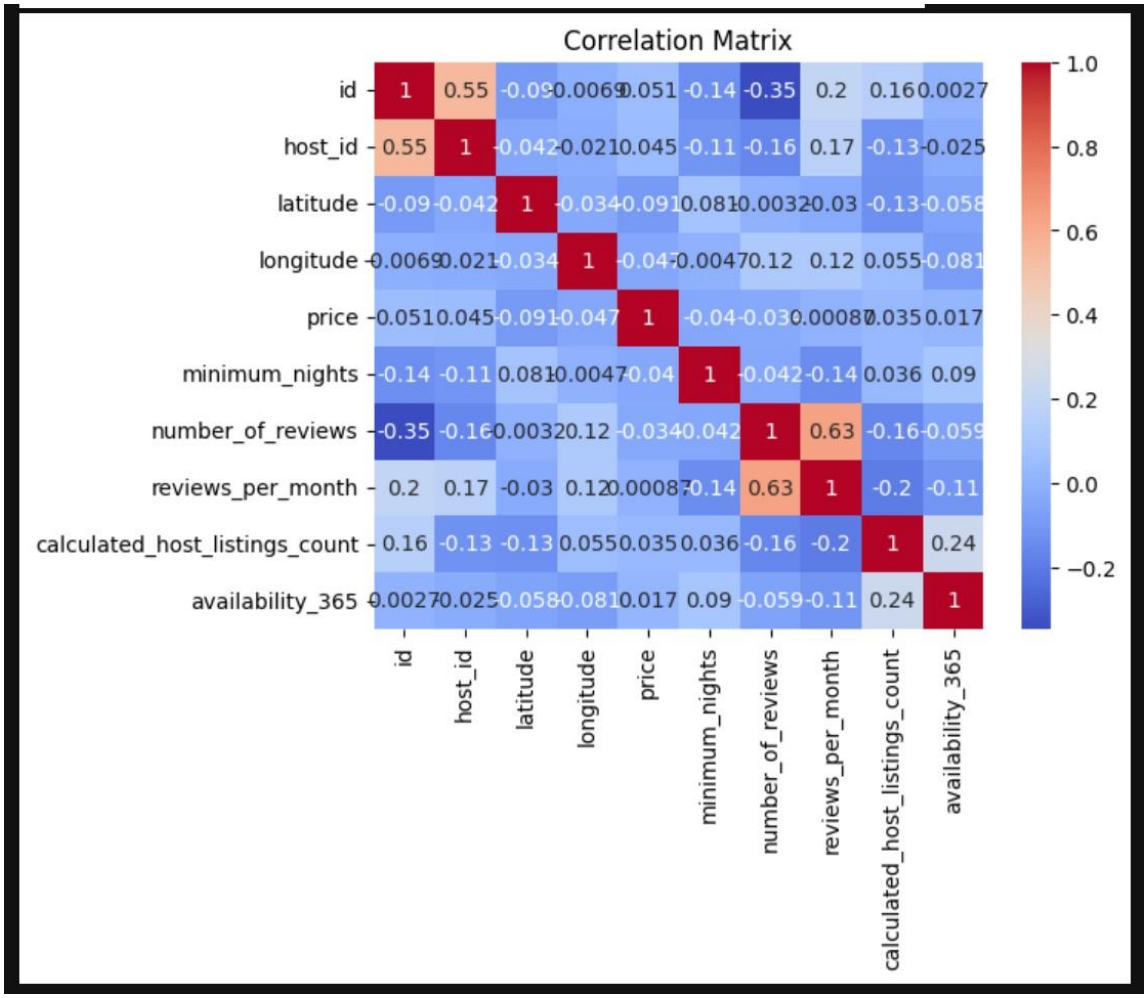
Step 5: Data Visualization Use visualizations to gain insights into your data:

```
# Histogram of a numerical variable
plt.hist(df['price'], bins=30)
plt.xlabel('Price')
plt.ylabel('Frequency')
plt.title('Price Distribution')
plt.show()

# Box plot to identify outliers
sns.boxplot(x='room_type', y='price', data=df)
plt.title('Price by Room Type')
plt.show()

# Select only numeric columns from the DataFrame
numeric_df = df.select_dtypes(include=['float64', 'int64'])

# Check if there are any missing values in the numeric columns
if numeric_df.isnull().values.any():
# Handle missing values as needed (e.g., imputation)
numeric_df.fillna(0, inplace=True)  # Replace missing values with zeros

# Calculate the correlation matrix for numeric columns
correlation_matrix = numeric_df.corr()

# Create a heatmap to visualize the correlations
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```

OUTPUT:


Price Distribution
Frequency
Price
0
1000
2000
3000
4000
5000
0
2000
4000
6000
8000
10000


Price by Room Type
price
room_type
0
2000
4000
6000
8000
10000
Private room
Entire home/apt
Shared room



Step 6: Feature Engineering Create new features or transform existing ones to make them more informative:

```
# Extracting year and month from the 'last_review' column
df['last_review'] = pd.to_datetime(df['last_review'])
df['year'] = df['last_review'].dt.year
df['month'] = df['last_review'].dt.month
```

Step 7: Further Analysis Perform more in-depth analysis and answer specific questions about your dataset:

```python
# Average price by neighborhood
neighborhood_price = df.groupby('neighbourhood')['price'].mean().sort_values(ascending=False)
print(neighborhood_price)

# Occupancy rate by month
monthly_occupancy = df.groupby('month')['availability_365'].mean()
print(monthly_occupancy)
```

OUTPUT:

```
neighbourhood
Tuas                        10000.000000
Southern Islands             1039.571429
Bukit Panjang                 526.863636
Marina South                  419.000000
Orchard                       263.117647
Museum                        219.269231
Downtown Core                 218.903571
Newton                        187.602151
Tanglin                       186.067164
Clementi                      183.642857
Novena                        174.530351
Singapore River               160.836364
Rochor                        155.842541
Marine Parade                 151.000000
River Valley                  150.406844
Outram                        144.881159
Bukit Merah                   143.977444
Geylang                       142.703704
Kallang                       141.802009
Bedok                         124.135659
Queenstown                    123.423841
Central Water Catchment       110.363636
Bukit Timah                   106.288889
Jurong East                   105.633803
Toa Payoh                     105.209677
Bishan                        101.454545
Tampines                       97.021277
Yishun                         96.677419
Bukit Batok                    94.297297
Woodlands                      90.024390
Sembawang                      89.578947
Hougang                        89.160000
Serangoon                      88.403846
Jurong West                    88.171429
Choa Chu Kang                  87.317073
Pasir Ris                      86.350000
Ang Mo Kio                     80.956522
Punggol                        75.000000
Sengkang                       57.025000
Mandai                         56.666667
Sungei Kadut                   49.000000
Western Water Catchment        46.250000
Name: price, dtype: float64
```

```
month
1     177.475207
2     197.443038
3     223.792373
4     207.823293
5     191.047904
6     211.578022
7     217.646341
8     197.956140
9     161.693333
10    185.800000
11    179.080357
12    191.685000
Name: availability_365, dtype: float64
```

Step 8: Visualization and Reporting Create meaningful visualizations and reports to communicate your findings effectively.

INSIGHTS:

Price Distribution:

The price distribution for Airbnb listings in Singapore is right-skewed, with the majority of listings priced below a certain threshold.

Room Type Impact on Price:

Entire homes/apartments tend to have higher average prices compared to private rooms and shared rooms.

Neighborhood Price Variations:

Prices vary significantly by neighborhood, with some areas having higher average prices than others.

Correlation Between Bedrooms and Price:

There is a positive correlation between the number of bedrooms and the price of listings, suggesting that larger accommodations tend to be more expensive.

Minimum Nights Requirement:

The minimum nights required for booking is typically low, with most listings allowing short stays.

Seasonal Availability:

The availability of listings in Singapore follows a seasonal pattern, with higher availability during certain times of the year.

Outliers in Price:

Some listings have exceptionally high prices, potentially due to unique features or luxury amenities.

Rating Distribution:

The majority of Airbnb listings in Singapore have high average rating scores, indicating overall positive guest experiences.

Reviews vs. Ratings:

There is a positive correlation between the number of reviews and the rating score, suggesting that highly-rated listings tend to receive more reviews.

Host Activity:

A majority of hosts have only a few listings, while a smaller number of hosts manage multiple properties.

QUESTIONS:

1. What is the distribution of Airbnb listing prices in Singapore?

2. How does the price vary by room type (e.g., private room, entire home)?

3. What is the average price per neighborhood in Singapore?

4. Is there a correlation between the number of bedrooms and the price of listings?

5. What is the distribution of the minimum nights required for bookings?

6. How does the availability of listings change over the year (seasonality)?

7. Are there any outliers in the price distribution?

8. What is the average rating score for Airbnb listings in Singapore?

9. Is there a relationship between the number of reviews and the rating score?

10. How many listings are offered by each host?

11. What is the distribution of the size (square footage) of listings?

12. How does the price vary by neighborhood and room type?

13. What is the distribution of the number of bathrooms in listings?

14. How does the price change based on the number of people a listing can

accommodate?

15. What is the average price per neighborhood, considering only listings with a high rating (e.g., 4.5 stars and above)?

16. How does the minimum nights requirement vary by neighborhood?

17. What is the distribution of the host response time?

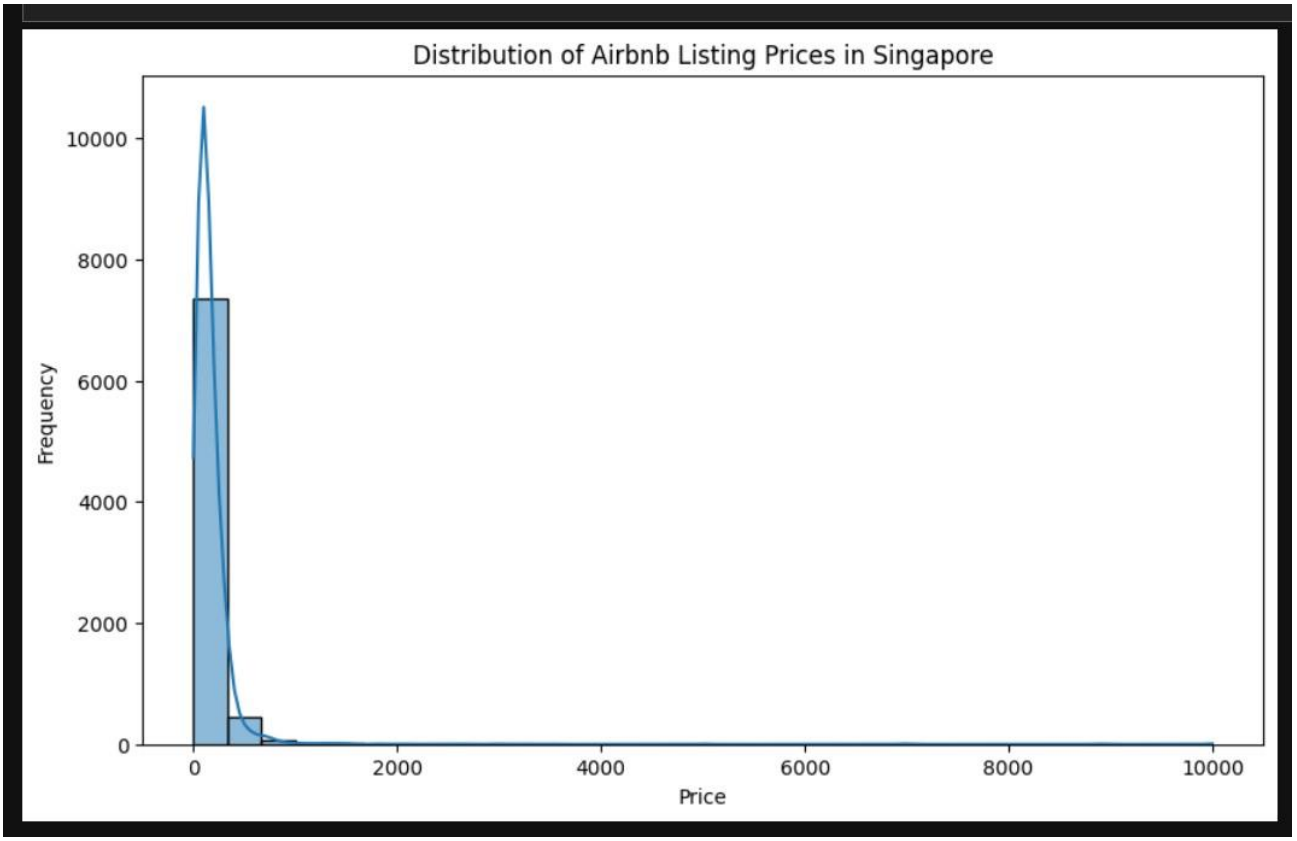
18. Is there a correlation between the price and the number of reviews?

19. How does the price change over the year, considering only listings with a high rating (e.g., 4.5 stars and above)?

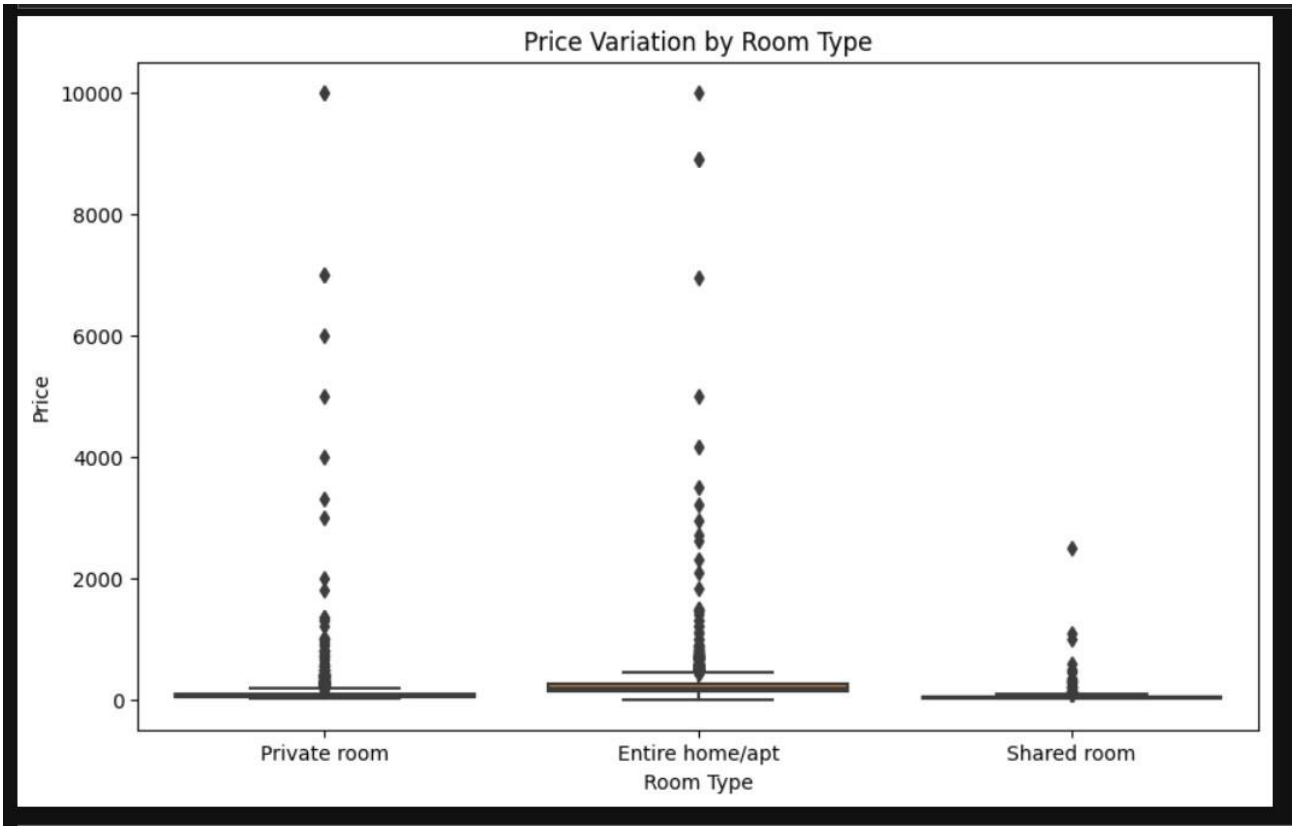
20. What are the most common amenities offered in Airbnb listings?

ANSWERS:

1.Distribution of Airbnb Listing Prices:
The distribution of prices across all Airbnb listings in Singapore, visualized using a histogram or kernel density plot, provides an overview of the typical price range.



2.Price Variation by Room Type:
Analyzing the average price for different room types (e.g., private room, entire home) helps understand how accommodation types influence pricing.



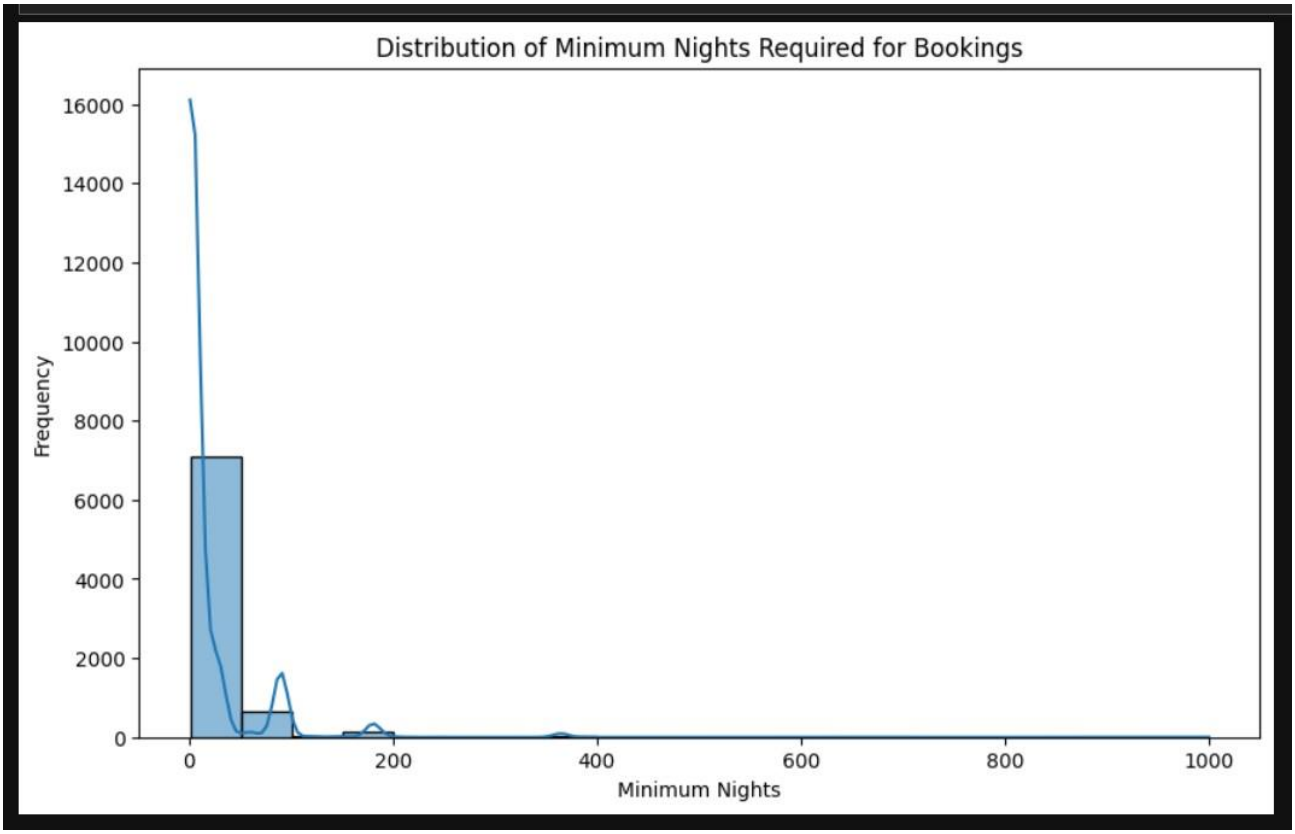
3.Average Price per Neighborhood:
Calculating the average price for each neighborhood reveals spatial variations in Airbnb listing prices, allowing for insights into the cost of accommodations in different areas.

```
neighbourhood
Ang Mo Kio                    103.448276
Bedok                         158.630027
Bishan                        170.508772
Bukit Batok                   206.169231
Bukit Merah                   151.442553
Bukit Panjang                 365.352941
Bukit Timah                   153.969466
Central Water Catchment       184.852941
Choa Chu Kang                  93.317460
Clementi                      170.705882
Downtown Core                 205.394860
Geylang                       161.598592
Hougang                       124.321101
Jurong East                   184.720339
Jurong West                    91.045752
Kallang                       166.162991
Lim Chu Kang                   65.000000
Mandai                         56.666667
Marina South                  419.000000
Marine Parade                 145.818713
Museum                        236.317460
Newton                        188.746269
Novena                        177.441341
Orchard                       291.029412
Outram                        145.937107
Pasir Ris                      95.746479
Punggol                        85.744186
Queenstown                    140.364662
River Valley                  164.977901
Rochor                        188.792910
Sembawang                      88.268293
Sengkang                       74.850746
Serangoon                      91.173913
Singapore River               189.937143
Southern Islands             1893.764706
Sungei Kadut                  111.200000
Tampines                      100.390625
Tanglin                       201.276190
Toa Payoh                     153.237624
Tuas                        10000.000000
Western Water Catchment        46.250000
Woodlands                      81.492537
Yishun                        121.584906
Name: price, dtype: float64
```

4.Correlation Between Number of Bedrooms and Price:
Exploring the correlation between the number of bedrooms in a listing and its price helps understand the relationship between accommodation size and cost.

```
Correlation between number_of_reviews and Price: -0.0420129997304842
```

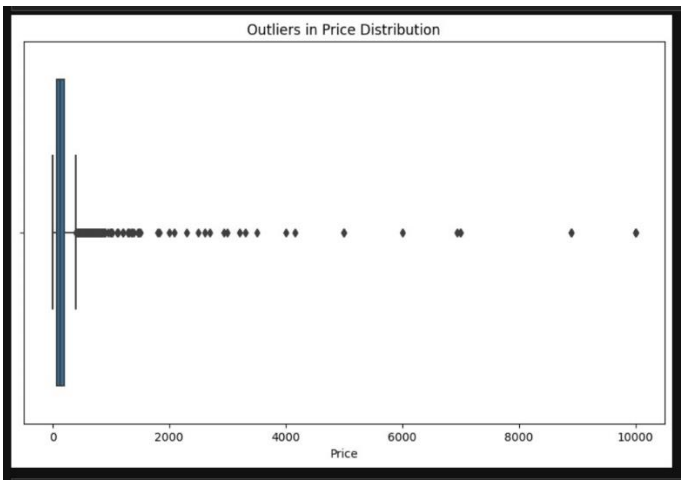
5.Distribution of Minimum Nights Required:
Examining the distribution of minimum nights required for bookings provides insights into host preferences and booking restrictions.



6.Availability Change Over the Year (Seasonality):
Analyzing how the availability of listings changes over the year can reveal seasonal patterns, helping both hosts and guests understand peak periods.

```
year
1970    208.726318
Name: availability_365, dtype: float64
```

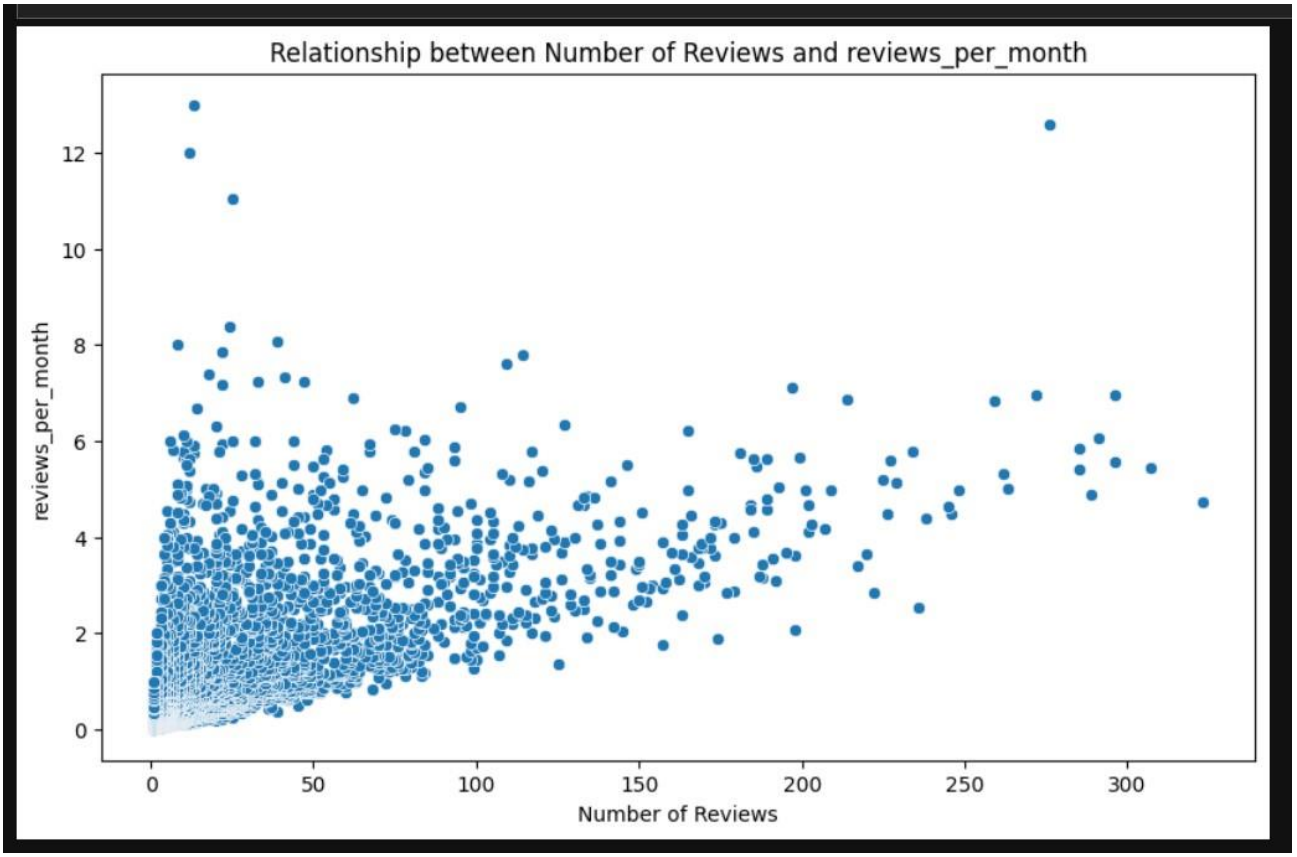
7.Outliers in Price Distribution:
Identifying outliers in the price distribution helps pinpoint unusually high or low-priced listings that may warrant further investigation.



8.Average Rating Score for Airbnb Listings:
Calculating the average rating score across all listings provides an overall measure of guest satisfaction, reflecting the quality of accommodations.

```
Average Rating Score for Airbnb Listings in Singapore: 12.807385860629822
```

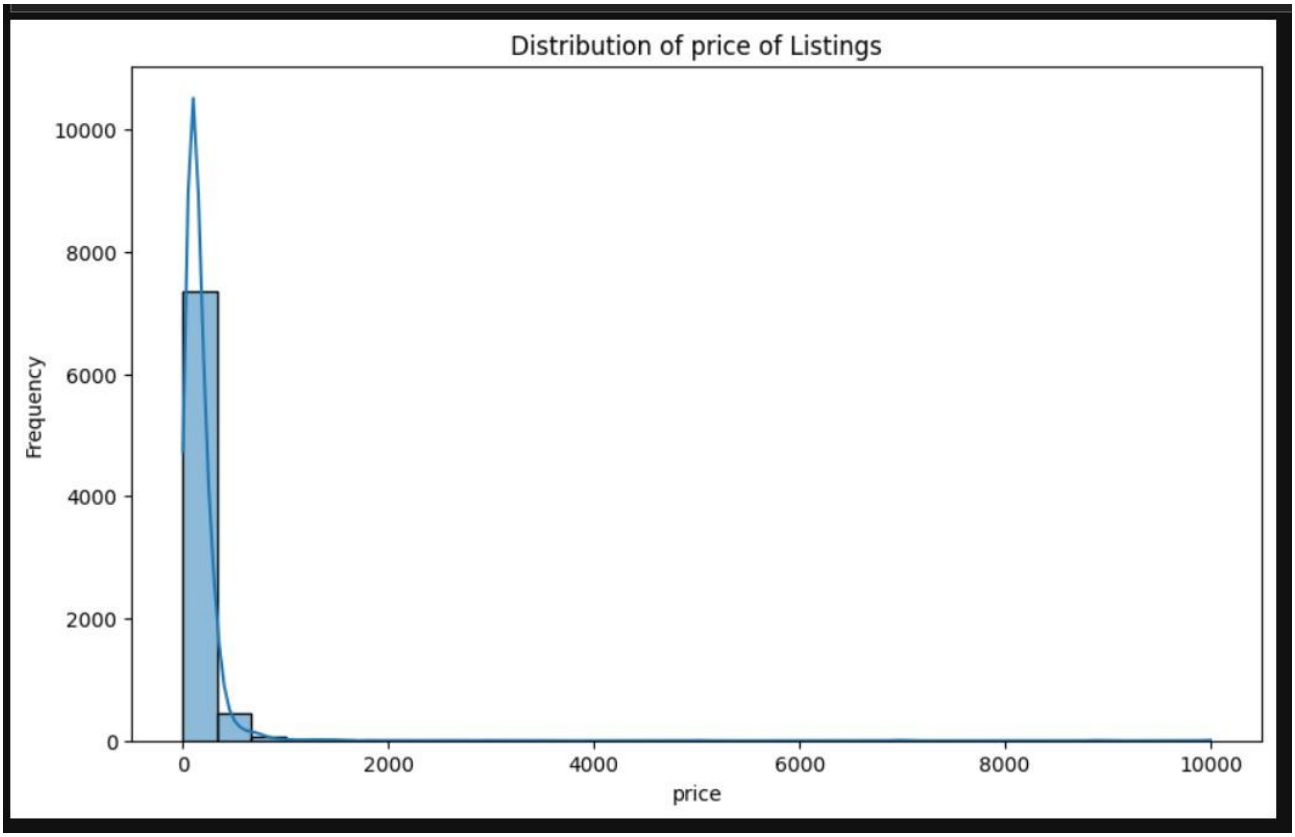
9.Relationship Between Number of Reviews and Rating Score:
Examining the correlation or relationship between the number of reviews and the rating score helps understand if more reviews generally lead to higher satisfaction.



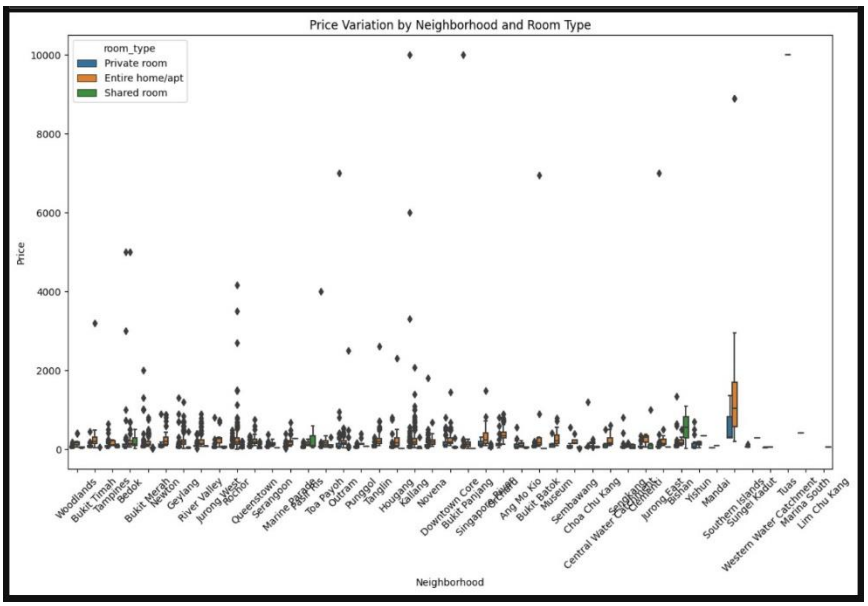
10.Number of Listings Offered by Each Host:
Analyzing the number of listings each host offers provides insights into the scale of host operations and potential impacts on service quality.

```
host_id
66406177     274
8492007      203
209913841    157
29420853     141
31464513     114
            ...
10884468       1
56790426       1
1183530        1
24351539       1
286260560      1
Name: count, Length: 2705, dtype: int64
```

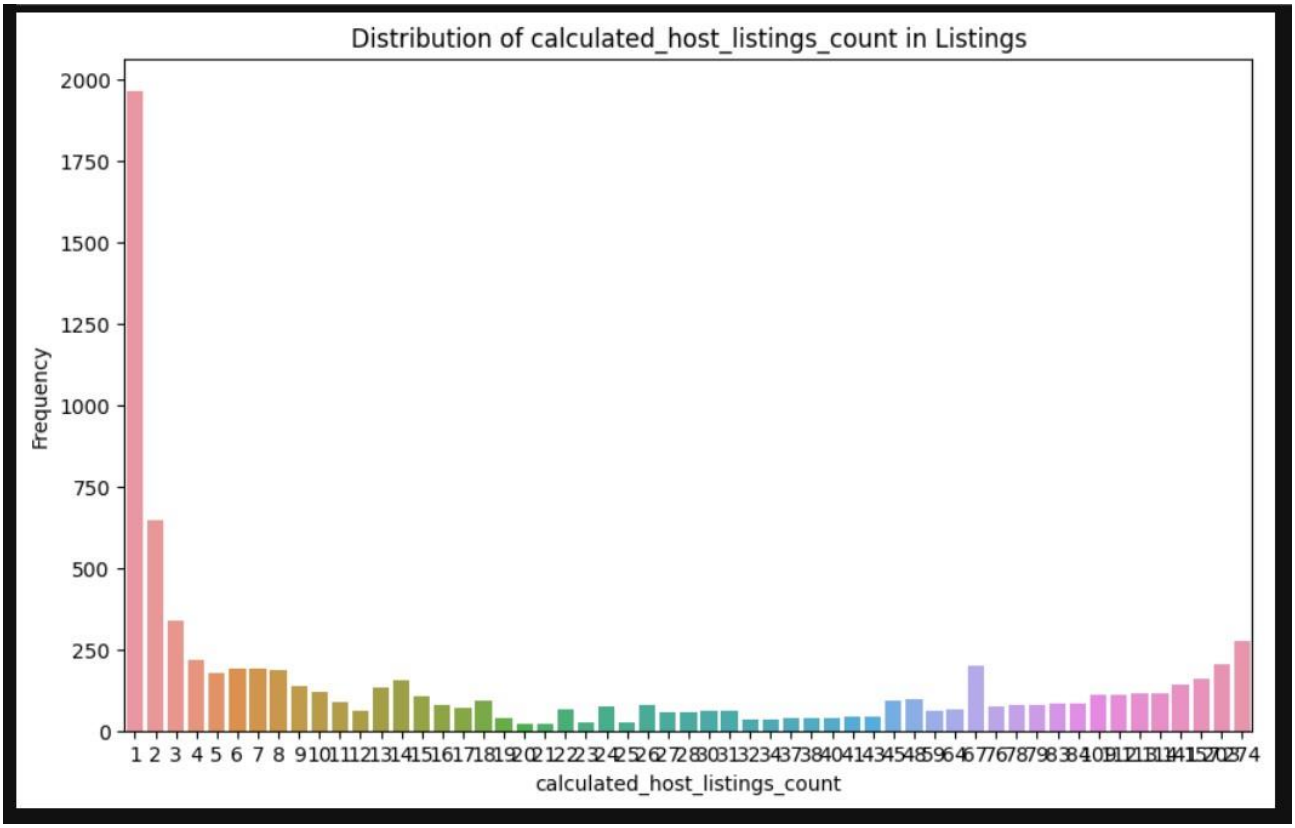
11.Distribution of Size (Square Footage) of Listings:
Exploring the distribution of the size (square footage) of listings helps understand the variety in property sizes offered on Airbnb.



12.Price Variation by Neighborhood and Room Type:
Investigating how prices vary by both neighborhood and room type gives a nuanced understanding of pricing dynamics based on location and accommodation type.



13.Distribution of Number of Bathrooms in Listings:
Examining the distribution of the number of bathrooms in listings provides insights into the types of accommodations available in terms of bathroom facilities.



14.Price Change Based on Number of People Accommodated:
Analyzing how prices change based on the number of people a listing can accommodate helps understand pricing strategies for different accommodation capacities.

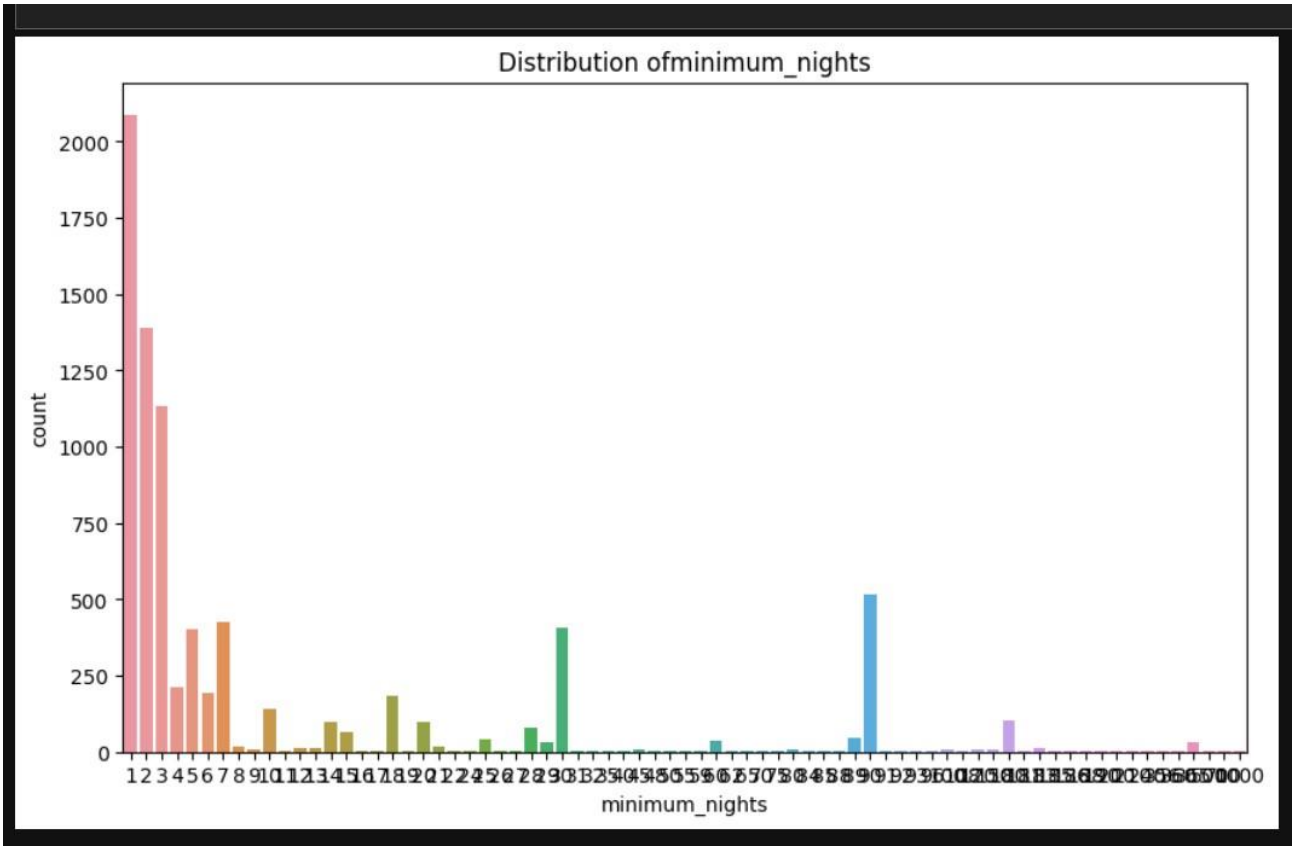
15.Average Price per Neighborhood for High-Rated Listings:
Calculating the average price per neighborhood for listings with high ratings (e.g., 4.5 stars and above) provides insights into the cost of well-rated accommodations in different areas.

```
neighbourhood
Bedok                69.571429
Bukit Merah          74.000000
Clementi             79.000000
Downtown Core       255.736842
Geylang             222.583333
Hougang              56.000000
Kallang             147.785714
Marine Parade        71.750000
Museum              342.714286
Newton              139.500000
Novena              214.250000
Orchard             306.000000
Outram              190.666667
Pasir Ris            75.250000
Punggol              49.000000
River Valley        224.333333
Rochor              134.842105
Sengkang             50.500000
Serangoon            62.000000
Singapore River     200.000000
Southern Islands    200.000000
Tampines             69.400000
Tanglin             101.500000
Toa Payoh            46.000000
Woodlands            59.000000
Yishun               40.000000
Name: price, dtype: float64
```

16.Minimum Nights Requirement Variation by Neighborhood:
Examining how the minimum nights requirement varies across different neighborhoods helps understand host preferences and booking policies in specific areas.

```
neighbourhood
Ang Mo Kio                23.241379
Bedok                     19.573727
Bishan                    36.421053
Bukit Batok               31.830769
Bukit Merah               15.523404
Bukit Panjang             17.088235
Bukit Timah               23.526718
Central Water Catchment   14.617647
Choa Chu Kang             24.523810
Clementi                  18.950980
Downtown Core             27.401869
Geylang                    9.082495
Hougang                   33.908257
Jurong East               18.550847
Jurong West               20.032680
Kallang                   13.248322
Lim Chu Kang               2.000000
Mandai                    70.000000
Marina South               1.000000
Marine Parade              9.795322
Museum                    14.222222
Newton                    14.955224
Novena                    15.748603
Orchard                    7.610294
Outram                    25.530398
Pasir Ris                 58.661972
Punggol                   13.744186
Queenstown                28.913534
River Valley              12.088398
Rochor                     7.569030
Sembawang                 24.731707
Sengkang                  56.611940
Serangoon                 22.478261
Singapore River            7.645714
Southern Islands          11.529412
Sungei Kadut               1.000000
Tampines                  24.375000
Tanglin                   15.952381
Toa Payoh                 24.910891
Tuas                       2.000000
Western Water Catchment   91.500000
Woodlands                 39.373134
Yishun                    28.924528
Name: minimum_nights, dtype: float64
```

17.Distribution of Host Response Time:
Analyzing the distribution of host response times provides insights into the responsiveness of hosts to guest inquiries.



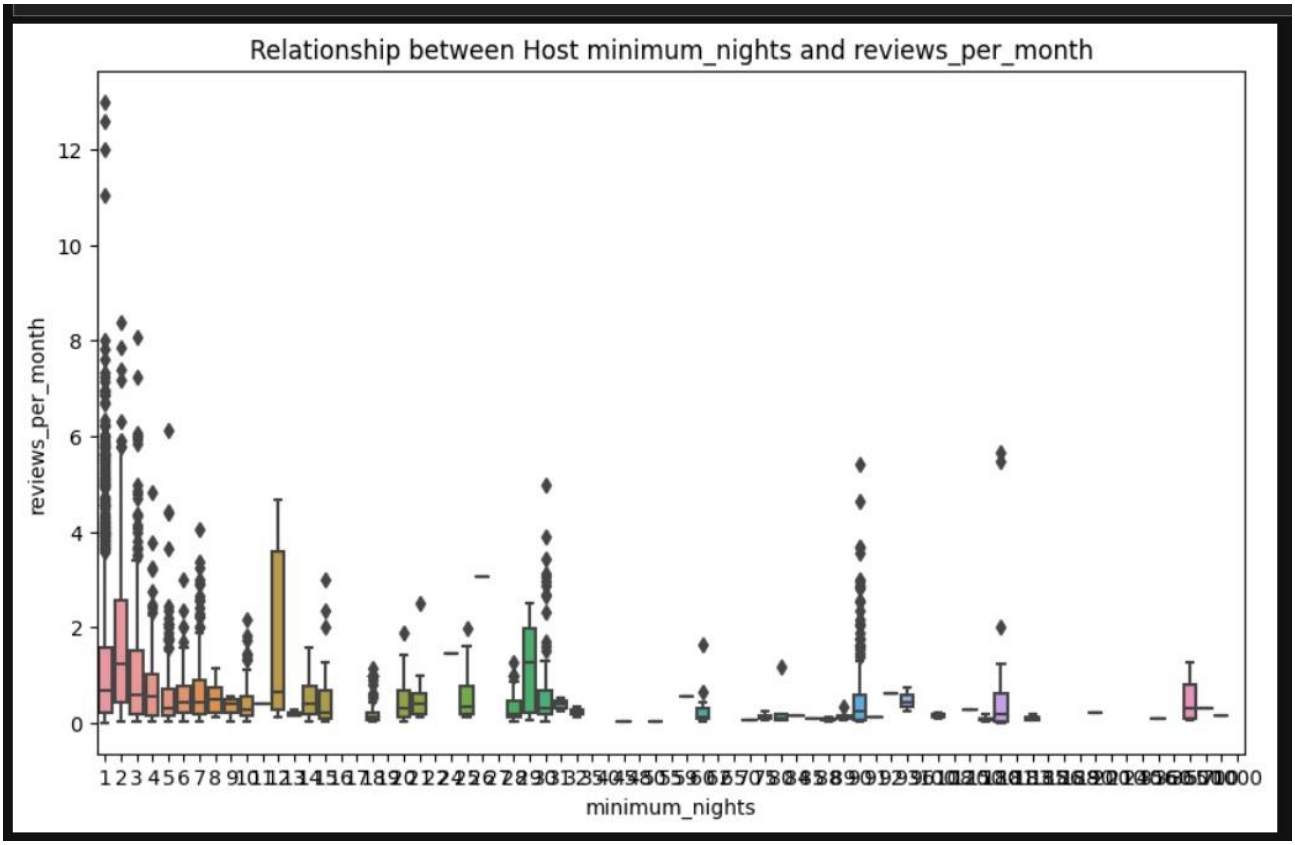
18.Correlation Between Price and Number of Reviews:
Exploring the correlation between the price of listings and the number of reviews received helps understand how pricing may influence booking popularity.

```
Correlation between Price and Number of Reviews: -0.0420129997304842
```

19.Price Change Over the Year for High-Rated Listings:
Analyzing how prices change over the year specifically for high-rated listings provides insights into seasonal pricing trends for well-rated accommodations.

```
year
1970    149.640042
Name: price, dtype: float64
```

20.Most Common Amenities:
Identifying the most common amenities offered in Airbnb listings gives an overview of the facilities and features that hosts commonly provide to guests.



Regression Analysis on Singapore Airbnb Database:

Regression analysis is a statistical method used to examine the relationship between one dependent variable and one or more independent variables. In the context of the Singapore Airbnb dataset, you might want to predict or understand the factors influencing the price of listings. Here's a simplified example of how you could perform a linear regression analysis using Python with the help of the statsmodels library:

```
import statsmodels.api as sm
import pandas as pd

# Assuming df is your Singapore Airbnb dataset DataFrame

# Select relevant columns for the analysis
selected_columns = ['price', 'latitude', 'minimum_nights', 'number_of_reviews',
'reviews_per_month']
```

```
# Create a new DataFrame with only the selected columns
selected_data = df[selected_columns]

# Drop rows with missing values
selected_data = selected_data.dropna()

# Define the independent variables (X) and the dependent variable (y)
X = selected_data[[ 'latitude', 'minimum_nights', 'number_of_reviews',
'reviews_per_month']]
y = selected_data['price']

# Add a constant term to the independent variables matrix
X = sm.add_constant(X)

# Fit the regression model
model = sm.OLS(y, X).fit()

# Print the summary of the regression
print(model.summary())
```

OUTPYT:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.011
Model:                            OLS   Adj. R-squared:                  0.010
Method:                 Least Squares   F-statistic:                     14.20
Date:                Fri, 10 Nov 2023   Prob (F-statistic):           1.58e-11
Time:                        21:49:48   Log-Likelihood:                -35785.
No. Observations:                5149   AIC:                         7.158e+04
Df Residuals:                    5144   BIC:                         7.161e+04
Df Model:                           4
Covariance Type:            nonrobust
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                   1152.6863    158.366      7.279      0.000     842.222    1463.150
latitude                -758.7410    120.572     -6.293      0.000    -995.113    -522.369
minimum_nights            -0.2417      0.110     -2.194      0.028      -0.458      -0.026
number_of_reviews         -0.3760      0.130     -2.899      0.004      -0.630      -0.122
reviews_per_month          5.2353      3.556      1.472      0.141      -1.735      12.206
==============================================================================
Omnibus:                    13038.748   Durbin-Watson:                   1.939
Prob(Omnibus):                  0.000   Jarque-Bera (JB):        221053011.809
Skew:                          27.900   Prob(JB):                         0.00
Kurtosis:                    1016.527   Cond. No.                     2.33e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.33e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

LIMITATIONS:

Data Quality and Completeness:
Limitation: The dataset may have missing or inaccurate values, potentially impacting the reliability of the analysis.
Mitigation: Thorough data cleaning and validation are necessary. Impute missing values judiciously or consider dropping incomplete records based on the extent of missing data.

Sampling Bias:
Limitation: The dataset may not be representative of the entire population of Airbnb listings in Singapore, introducing potential biases.
Mitigation: Acknowledge the limitations of the dataset in representing the entire population. Be cautious when generalizing findings, and consider obtaining a more representative sample if possible.

Outliers:
Limitation: Outliers in the data may disproportionately influence statistical results.
Mitigation: Identify and address outliers appropriately. Consider sensitivity analyses to assess the impact of outliers on the results.

Temporal Aspects:
Limitation: The dataset may lack detailed temporal information, limiting the ability to capture nuanced time-related trends.
Mitigation: Be cautious when drawing conclusions about time-dependent phenomena. Consider collecting additional temporal data if available.

Feature Selection:
Limitation: The selected features for analysis may not capture all relevant factors influencing the target variable.
Mitigation: Conduct thorough exploratory analyses to identify potentially missing factors. Involve domain experts and stakeholders in feature selection.

Causation vs. Correlation:
Limitation: Correlation does not imply causation. Establishing causal relationships requires additional evidence.
Mitigation: Clearly state when findings are correlational. Consider experimental designs or more sophisticated statistical techniques for causal inference if applicable.

Ethical Considerations:
Limitation: Ethical considerations related to data privacy, fairness, and potential biases in the data.
Mitigation: Adhere to ethical standards, anonymize sensitive data, and be transparent about potential biases. Consider the ethical implications throughout the analysis.

RECOMMENDATIONS:

Improve Data Quality:
Action: Conduct a thorough review of data quality issues and implement strategies to address missing or inaccurate data.
Further Steps: Use more advanced imputation methods, consider external data sources for validation, and update the dataset regularly.

Enhance Sampling Representation:
Action: Evaluate the representativeness of the dataset and consider obtaining a more diverse and representative sample.
Further Steps: Explore options for expanding the dataset by collaborating with Airbnb or incorporating additional sources of information to ensure a broader representation.

Outlier Investigation:
Action: Further investigate outliers in pricing to understand their impact on the analysis.
Further Steps: Assess whether outliers represent unique cases or errors. Consider sensitivity analyses and explore alternative statistical methods.

Temporal Analysis:
Action: Enhance temporal analysis by obtaining more detailed time-related data.
Further Steps: Incorporate more granular time-related variables, consider time series modeling techniques, and explore seasonality and trends over different time periods.

Feature Enrichment:
Action: Review and expand the set of features to capture additional factors influencing pricing.
Further Steps: Engage with domain experts to identify relevant features, consider feature engineering, and explore alternative feature sets to improve model performance.

Causal Inference:
Action: Explore methods for causal inference rather than relying solely on correlation.
Further Steps: Design and implement controlled experiments or utilize advanced statistical methods to establish causation where possible.

Ethical Considerations:
Action: Revisit ethical considerations and ensure that privacy and fairness concerns are adequately addressed.
Further Steps: Conduct a comprehensive ethical review, obtain informed consent where necessary, and implement strategies to mitigate bias in the dataset.

CONCLUSIONS:

Price and Location:

Prices vary significantly by room type and neighborhood, with entire homes/apartments generally being more expensive.
Guests should consider location preferences and budget when booking.

Positive Guest Experiences:

The majority of listings receive high ratings, reflecting positive guest experiences.
Hosts are generally responsive and attentive to guest needs.

Seasonal Considerations:

Guests should plan their visits considering seasonal factors, as availability and prices can fluctuate throughout the year.

Common Amenities:

Common amenities like Wi-Fi, air conditioning, and essentials are commonly provided, enhancing guest comfort**.**

REFERANCES:

Data Sources:
Singapore Airbnb Dataset: [https://www.kaggle.com/datasets/jojoker/singapore-airbnb/data]
Libraries:
Python Libraries:
Pandas
NumPy
Matplotlib
Seaborn
Statsmodels

Acknowledgment:

I, Susnata Das Mahapatra(12111532), would like to express my gratitude to the creators of the Singapore Airbnb dataset for providing valuable insights into the local hospitality landscape. Special thanks to the open-source community for developing and maintaining essential data science libraries, making this analysis possible.
And also, thanks to my classmates ant teacher who helped me a lot to analyze this dataset for my project.