DefineAndSolveMLProblem

July 26, 2025

1 Lab 8: Define and Solve an ML Problem of Your Choosing

[3]:
```
import pandas as pd
import numpy as np
import os
import matplotlib.pyplot as plt
import seaborn as sns
```

In this lab assignment, you will follow the machine learning life cycle and implement a model to solve a machine learning problem of your choosing. You will select a data set and choose a predictive problem that the data set supports. You will then inspect the data with your problem in mind and begin to formulate a project plan. You will then implement the machine learning project plan.

You will complete the following tasks:

1. Build Your DataFrame
2. Define Your ML Problem
3. Perform exploratory data analysis to understand your data.
4. Define Your Project Plan
5. Implement Your Project Plan:
    - Prepare your data for your model.
    - Fit your model to the training data and evaluate your model.
    - Improve your model's performance.

1.1 Part 1: Build Your DataFrame

You will have the option to choose one of four data sets that you have worked with in this program:

- The "census" data set that contains Census information from 1994: `censusData.csv`
- Airbnb NYC "listings" data set: `airbnbListingsData.csv`
- World Happiness Report (WHR) data set: `WHR2018Chapter2OnlineData.csv`
- Book Review data set: `bookReviewsData.csv`

Note that these are variations of the data sets that you have worked with in this program. For example, some do not include some of the preprocessing necessary for specific models.

Load a Data Set and Save it as a Pandas DataFrame The code cell below contains filenames (path + filename) for each of the four data sets available to you.

Task: In the code cell below, use the same method you have been using to load the data using `pd.read_csv()` and save it to DataFrame `df`.

You can load each file as a new DataFrame to inspect the data before choosing your data set.

[4]:
```
# File names of the four data sets
adultDataSet_filename = os.path.join(os.getcwd(), "data", "censusData.csv")
airbnbDataSet_filename = os.path.join(os.getcwd(), "data", "airbnbListingsData.
 ↪csv")
WHRDataSet_filename = os.path.join(os.getcwd(), "data",␣
 ↪"WHR2018Chapter2OnlineData.csv")
bookReviewDataSet_filename = os.path.join(os.getcwd(), "data", "bookReviewsData.
 ↪csv")


df = pd.read_csv(airbnbDataSet_filename)

df.head()
```

[4]:
```
                                                name  \
0                                Skylit Midtown Castle
1  Whole flr w/private bdrm, bath & kitchen(pls r…
2            Spacious Brooklyn Duplex, Patio + Garden
3                    Large Furnished Room Near B'way
4                  Cozy Clean Guest Room - Family Apt

                                            description  \
0  Beautiful, spacious skylit studio in the heart…
1  Enjoy 500 s.f. top floor in 1899 brownstone, w…
2  We welcome you to stay in our lovely 2 br dupl…
3  Please don't expect the luxury here just a bas…
4  Our best guests are seeking a safe, clean, spa…

                                neighborhood_overview    host_name  \
0  Centrally located in the heart of Manhattan ju…     Jennifer
1  Just the right mix of urban center and local n…  LisaRoxanne
2                                                NaN      Rebecca
3    Theater district, many restaurants around here.     Shunichi
4  Our neighborhood is full of restaurants and ca…    MaryEllen

                       host_location  \
0  New York, New York, United States
1  New York, New York, United States
2  Brooklyn, New York, United States
3  New York, New York, United States
4  New York, New York, United States

                                       host_about  host_response_rate  \
```

```
0  A New Yorker since 2000! My passion is creatin…                0.80
1  Laid-back Native New Yorker (formerly bi-coast…                0.09
2  Rebecca is an artist/designer, and Henoch is i…                1.00
3  I used to work for a financial industry but no…                1.00
4  Welcome to family life with my oldest two away…                 NaN

   host_acceptance_rate  host_is_superhost  host_listings_count  …  \
0                 0.17              True                  8.0  …
1                 0.69              True                  1.0  …
2                 0.25              True                  1.0  …
3                 1.00              True                  1.0  …
4                  NaN              True                  1.0  …

   review_scores_communication  review_scores_location  review_scores_value  \
0                         4.79                    4.86                 4.41
1                         4.80                    4.71                 4.64
2                         5.00                    4.50                 5.00
3                         4.42                    4.87                 4.36
4                         4.95                    4.94                 4.92

  instant_bookable calculated_host_listings_count  \
0            False                               3
1            False                               1
2            False                               1
3            False                               1
4            False                               1

   calculated_host_listings_count_entire_homes  \
0                                            3
1                                            1
2                                            1
3                                            0
4                                            0

   calculated_host_listings_count_private_rooms  \
0                                             0
1                                             0
2                                             0
3                                             1
4                                             1

   calculated_host_listings_count_shared_rooms  reviews_per_month  \
0                                            0               0.33
1                                            0               4.86
2                                            0               0.02
3                                            0               3.68
4                                            0               0.87
```

```
  n_host_verifications
0                    9
1                    6
2                    3
3                    4
4                    7

[5 rows x 50 columns]
```

1.2 Part 2: Define Your ML Problem

Next you will formulate your ML Problem. In the markdown cell below, answer the following questions:

1. List the data set you have chosen.
2. What will you be predicting? What is the label?
3. Is this a supervised or unsupervised learning problem? Is this a clustering, classification or regression problem? Is it a binary classificaiton or multi-class classifiction problem?
4. What are your features? (note: this list may change after your explore your data)
5. Explain why this is an important problem. In other words, how would a company create value with a model that predicts this label?

1. I have chosen the Airbnb NYC "listings" data set: `airbnbListingsData.csv`.
2. I will be predicting the `price` of a listing. The label is the `price` column.
3. This is a supervised learning problem because we are using labeled data (`price`) to train the model. It is a regression problem since the output variable is continuous.
4. `room_type` `neighbourhood_group` `latitude`, `longitude` `minimum_nights` `number_of_reviews` `reviews_per_month` `availability_365` `calculated_host_listings_count` (This list may be revised after data cleaning and exploration.)
5. Predicting listing prices is valuable for platforms like Airbnb to help hosts set competitive pricing, provide price recommendations, detect outlier or fraudulent listings, and optimize revenue. It can also help customers filter and compare listings more effectively based on predicted value for money.

1.3 Part 3: Understand Your Data

The next step is to perform exploratory data analysis. Inspect and analyze your data set with your machine learning problem in mind. Consider the following as you inspect your data:

1. What data preparation techniques would you like to use? These data preparation techniques may include:
    - addressing missingness, such as replacing missing values with means
    - finding and replacing outliers

- renaming features and labels
- finding and replacing outliers
- performing feature engineering techniques such as one-hot encoding on categorical features
- selecting appropriate features and removing irrelevant features
- performing specific data cleaning and preprocessing techniques for an NLP problem
- addressing class imbalance in your data sample to promote fair AI

2. What machine learning model (or models) you would like to use that is suitable for your predictive problem and data?

    - Are there other data preparation techniques that you will need to apply to build a balanced modeling data set for your problem and model? For example, will you need to scale your data?

3. How will you evaluate and improve the model's performance?

    - Are there specific evaluation metrics and methods that are appropriate for your model?

Think of the different techniques you have used to inspect and analyze your data in this course. These include using Pandas to apply data filters, using the Pandas `describe()` method to get insight into key statistics for each column, using the Pandas `dtypes` property to inspect the data type of each column, and using Matplotlib and Seaborn to detect outliers and visualize relationships between features and labels. If you are working on a classification problem, use techniques you have learned to determine if there is class imbalance.

Task: Use the techniques you have learned in this course to inspect and analyze your data. You can import additional packages that you have used in this course that you will need to perform this task.

Note: You can add code cells if needed by going to the Insert menu and clicking on Insert Cell Below in the drop-drown menu.

1.4 Part 3: Understand Your Data

The next step is to perform exploratory data analysis. Inspect and analyze your data set with your machine learning problem in mind. Consider the following as you inspect your data:

1. What data preparation techniques would you like to use? These data preparation techniques may include:

    - addressing missingness, such as replacing missing values with means
    - finding and replacing outliers
    - renaming features and labels
    - finding and replacing outliers
    - performing feature engineering techniques such as one-hot encoding on categorical features
    - selecting appropriate features and removing irrelevant features
    - performing specific data cleaning and preprocessing techniques for an NLP problem
    - addressing class imbalance in your data sample to promote fair AI

2. What machine learning model (or models) you would like to use that is suitable for your predictive problem and data?
    - Are there other data preparation techniques that you will need to apply to build a balanced modeling data set for your problem and model? For example, will you need to scale your data?
3. How will you evaluate and improve the model's performance?
    - Are there specific evaluation metrics and methods that are appropriate for your model?

Think of the different techniques you have used to inspect and analyze your data in this course. These include using Pandas to apply data filters, using the Pandas `describe()` method to get insight into key statistics for each column, using the Pandas `dtypes` property to inspect the data type of each column, and using Matplotlib and Seaborn to detect outliers and visualize relationships between features and labels. If you are working on a classification problem, use techniques you have learned to determine if there is class imbalance.

Task: Use the techniques you have learned in this course to inspect and analyze your data. You can import additional packages that you have used in this course that you will need to perform this task.

Note: You can add code cells if needed by going to the Insert menu and clicking on Insert Cell Below in the drop-drown menu.

[5]:
```
# Load the dataset
df = pd.read_csv(airbnbDataSet_filename)

# Show the first few rows
print("Preview of dataset:")
display(df.head())

# Overview of data types and non-null counts
print("\nDataFrame info:")
df.info()

# Summary statistics
print("\nSummary statistics:")
display(df.describe())

# Check for missing values
print("\nMissing values per column:")
display(df.isnull().sum())

# Check data types
print("\nData types:")
print(df.dtypes)

# Boxplot for price to check outliers
plt.figure(figsize=(10, 4))
```

```
sns.boxplot(x=df['price'])
plt.title("Price Distribution with Outliers")
plt.show()

# Distribution plot of price
plt.figure(figsize=(10, 4))
sns.histplot(df['price'], bins=50, kde=True)
plt.title("Histogram of Price")
plt.xlabel("Price")
plt.ylabel("Frequency")
plt.show()

# Correlation heatmap for numerical features
plt.figure(figsize=(10, 8))
sns.heatmap(df.select_dtypes(include=['number']).corr(), annot=True,␣
 ↪cmap="coolwarm")
plt.title("Correlation Heatmap")
plt.show()
```

```
Preview of dataset:

                                                name  \
0                                Skylit Midtown Castle
1  Whole flr w/private bdrm, bath & kitchen(pls r...
2            Spacious Brooklyn Duplex, Patio + Garden
3                    Large Furnished Room Near B'way
4                  Cozy Clean Guest Room - Family Apt

                                          description  \
0  Beautiful, spacious skylit studio in the heart...
1  Enjoy 500 s.f. top floor in 1899 brownstone, w...
2  We welcome you to stay in our lovely 2 br dupl...
3  Please don't expect the luxury here just a bas...
4  Our best guests are seeking a safe, clean, spa...

                               neighborhood_overview    host_name  \
0  Centrally located in the heart of Manhattan ju...     Jennifer
1  Just the right mix of urban center and local n...  LisaRoxanne
2                                                NaN      Rebecca
3    Theater district, many restaurants around here.     Shunichi
4  Our neighborhood is full of restaurants and ca...    MaryEllen

                       host_location  \
0  New York, New York, United States
1  New York, New York, United States
2  Brooklyn, New York, United States
3  New York, New York, United States
4  New York, New York, United States
```

```
                                          host_about  host_response_rate  \
0  A New Yorker since 2000! My passion is creatin...                0.80
1  Laid-back Native New Yorker (formerly bi-coast...                0.09
2  Rebecca is an artist/designer, and Henoch is i...                1.00
3  I used to work for a financial industry but no...                1.00
4  Welcome to family life with my oldest two away...                 NaN

   host_acceptance_rate  host_is_superhost  host_listings_count  ...  \
0                  0.17               True                  8.0  ...
1                  0.69               True                  1.0  ...
2                  0.25               True                  1.0  ...
3                  1.00               True                  1.0  ...
4                   NaN               True                  1.0  ...

   review_scores_communication  review_scores_location  review_scores_value  \
0                         4.79                    4.86                 4.41
1                         4.80                    4.71                 4.64
2                         5.00                    4.50                 5.00
3                         4.42                    4.87                 4.36
4                         4.95                    4.94                 4.92

  instant_bookable calculated_host_listings_count  \
0            False                               3
1            False                               1
2            False                               1
3            False                               1
4            False                               1

   calculated_host_listings_count_entire_homes  \
0                                            3
1                                            1
2                                            1
3                                            0
4                                            0

   calculated_host_listings_count_private_rooms  \
0                                             0
1                                             0
2                                             0
3                                             1
4                                             1

   calculated_host_listings_count_shared_rooms  reviews_per_month  \
0                                            0               0.33
1                                            0               4.86
2                                            0               0.02
3                                            0               3.68
```

```
4                                                  0               0.87

  n_host_verifications
0                    9
1                    6
2                    3
3                    4
4                    7

[5 rows x 50 columns]


DataFrame info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 28022 entries, 0 to 28021
Data columns (total 50 columns):
 #   Column                                        Non-Null Count  Dtype
---  ------                                        --------------  -----
 0   name                                          28017 non-null  object
 1   description                                   27452 non-null  object
 2   neighborhood_overview                         18206 non-null  object
 3   host_name                                     28022 non-null  object
 4   host_location                                 27962 non-null  object
 5   host_about                                    17077 non-null  object
 6   host_response_rate                            16179 non-null  float64
 7   host_acceptance_rate                          16909 non-null  float64
 8   host_is_superhost                             28022 non-null  bool
 9   host_listings_count                           28022 non-null  float64
 10  host_total_listings_count                     28022 non-null  float64
 11  host_has_profile_pic                          28022 non-null  bool
 12  host_identity_verified                        28022 non-null  bool
 13  neighbourhood_group_cleansed                  28022 non-null  object
 14  room_type                                     28022 non-null  object
 15  accommodates                                  28022 non-null  int64
 16  bathrooms                                     28022 non-null  float64
 17  bedrooms                                      25104 non-null  float64
 18  beds                                          26668 non-null  float64
 19  amenities                                     28022 non-null  object
 20  price                                         28022 non-null  float64
 21  minimum_nights                                28022 non-null  int64
 22  maximum_nights                                28022 non-null  int64
 23  minimum_minimum_nights                        28022 non-null  float64
 24  maximum_minimum_nights                        28022 non-null  float64
 25  minimum_maximum_nights                        28022 non-null  float64
 26  maximum_maximum_nights                        28022 non-null  float64
 27  minimum_nights_avg_ntm                        28022 non-null  float64
 28  maximum_nights_avg_ntm                        28022 non-null  float64
```

```
 29  has_availability                                28022 non-null  bool
 30  availability_30                                 28022 non-null  int64
 31  availability_60                                 28022 non-null  int64
 32  availability_90                                 28022 non-null  int64
 33  availability_365                                28022 non-null  int64
 34  number_of_reviews                               28022 non-null  int64
 35  number_of_reviews_ltm                           28022 non-null  int64
 36  number_of_reviews_l30d                          28022 non-null  int64
 37  review_scores_rating                            28022 non-null  float64
 38  review_scores_cleanliness                       28022 non-null  float64
 39  review_scores_checkin                           28022 non-null  float64
 40  review_scores_communication                     28022 non-null  float64
 41  review_scores_location                          28022 non-null  float64
 42  review_scores_value                             28022 non-null  float64
 43  instant_bookable                                28022 non-null  bool
 44  calculated_host_listings_count                  28022 non-null  int64
 45  calculated_host_listings_count_entire_homes     28022 non-null  int64
 46  calculated_host_listings_count_private_rooms    28022 non-null  int64
 47  calculated_host_listings_count_shared_rooms     28022 non-null  int64
 48  reviews_per_month                               28022 non-null  float64
 49  n_host_verifications                            28022 non-null  int64
dtypes: bool(5), float64(21), int64(15), object(9)
memory usage: 9.8+ MB

Summary statistics:

       host_response_rate  host_acceptance_rate  host_listings_count  \
count        16179.000000          16909.000000         28022.000000
mean             0.906901              0.791953            14.554778
std              0.227282              0.276732           120.721287
min              0.000000              0.000000             0.000000
25%              0.940000              0.680000             1.000000
50%              1.000000              0.910000             1.000000
75%              1.000000              1.000000             3.000000
max              1.000000              1.000000          3387.000000

       host_total_listings_count  accommodates     bathrooms      bedrooms  \
count               28022.000000  28022.000000  28022.000000  25104.000000
mean                   14.554778      2.874491      1.142174      1.329708
std                   120.721287      1.860251      0.421132      0.700726
min                     0.000000      1.000000      0.000000      1.000000
25%                     1.000000      2.000000      1.000000      1.000000
50%                     1.000000      2.000000      1.000000      1.000000
75%                     3.000000      4.000000      1.000000      1.000000
max                  3387.000000     16.000000      8.000000     12.000000

              beds         price  minimum_nights  ...  review_scores_checkin  \
count  26668.000000  28022.000000    28022.000000  ...           28022.000000
```

mean	1.629556	154.228749	18.689387	...	4.814300
std	1.097104	140.816605	25.569151	...	0.438603
min	1.000000	29.000000	1.000000	...	0.000000
25%	1.000000	70.000000	2.000000	...	4.810000
50%	1.000000	115.000000	30.000000	...	4.960000
75%	2.000000	180.000000	30.000000	...	5.000000
max	21.000000	1000.000000	1250.000000	...	5.000000

	review_scores_communication	review_scores_location
count	28022.000000	28022.000000
mean	4.808041	4.750393
std	0.464585	0.415717
min	0.000000	0.000000
25%	4.810000	4.670000
50%	4.970000	4.880000
75%	5.000000	5.000000
max	5.000000	5.000000

	review_scores_value	calculated_host_listings_count
count	28022.000000	28022.000000
mean	4.647670	9.581900
std	0.518023	32.227523
min	0.000000	1.000000
25%	4.550000	1.000000
50%	4.780000	1.000000
75%	5.000000	3.000000
max	5.000000	421.000000

	calculated_host_listings_count_entire_homes
count	28022.000000
mean	5.562986
std	26.121426
min	0.000000
25%	0.000000
50%	1.000000
75%	1.000000
max	308.000000

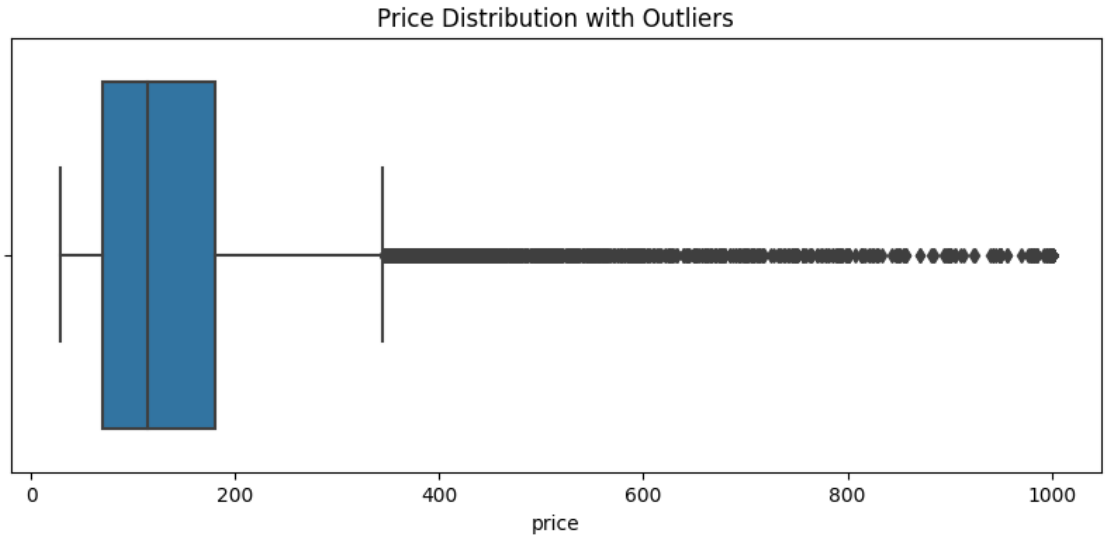
	calculated_host_listings_count_private_rooms
count	28022.000000
mean	3.902077
std	17.972386
min	0.000000
25%	0.000000
50%	0.000000
75%	1.000000
max	359.000000

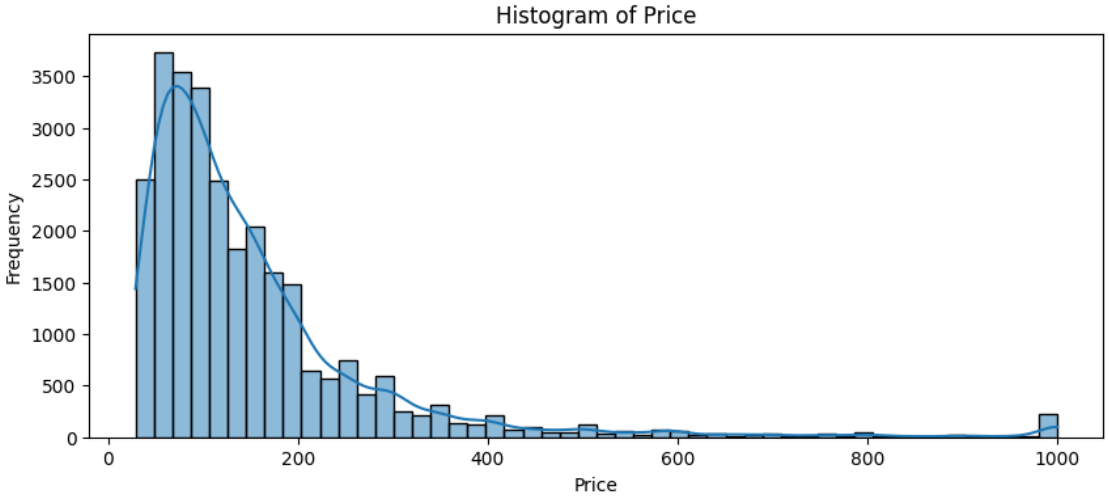
```
       calculated_host_listings_count_shared_rooms  reviews_per_month  \
count                                  28022.000000       28022.000000
mean                                       0.048283           1.758325
std                                        0.442459           4.446143
min                                        0.000000           0.010000
25%                                        0.000000           0.130000
50%                                        0.000000           0.510000
75%                                        0.000000           1.830000
max                                        8.000000         141.000000

       n_host_verifications
count          28022.000000
mean               5.169510
std                2.028497
min                1.000000
25%                4.000000
50%                5.000000
75%                7.000000
max               13.000000

[8 rows x 36 columns]


Missing values per column:

name                                    5
description                           570
neighborhood_overview                9816
host_name                               0
host_location                          60
host_about                          10945
host_response_rate                  11843
host_acceptance_rate                11113
host_is_superhost                       0
host_listings_count                     0
host_total_listings_count               0
host_has_profile_pic                    0
host_identity_verified                  0
neighbourhood_group_cleansed            0
room_type                               0
accommodates                            0
bathrooms                               0
bedrooms                             2918
beds                                 1354
amenities                               0
price                                   0
minimum_nights                          0
maximum_nights                          0
```

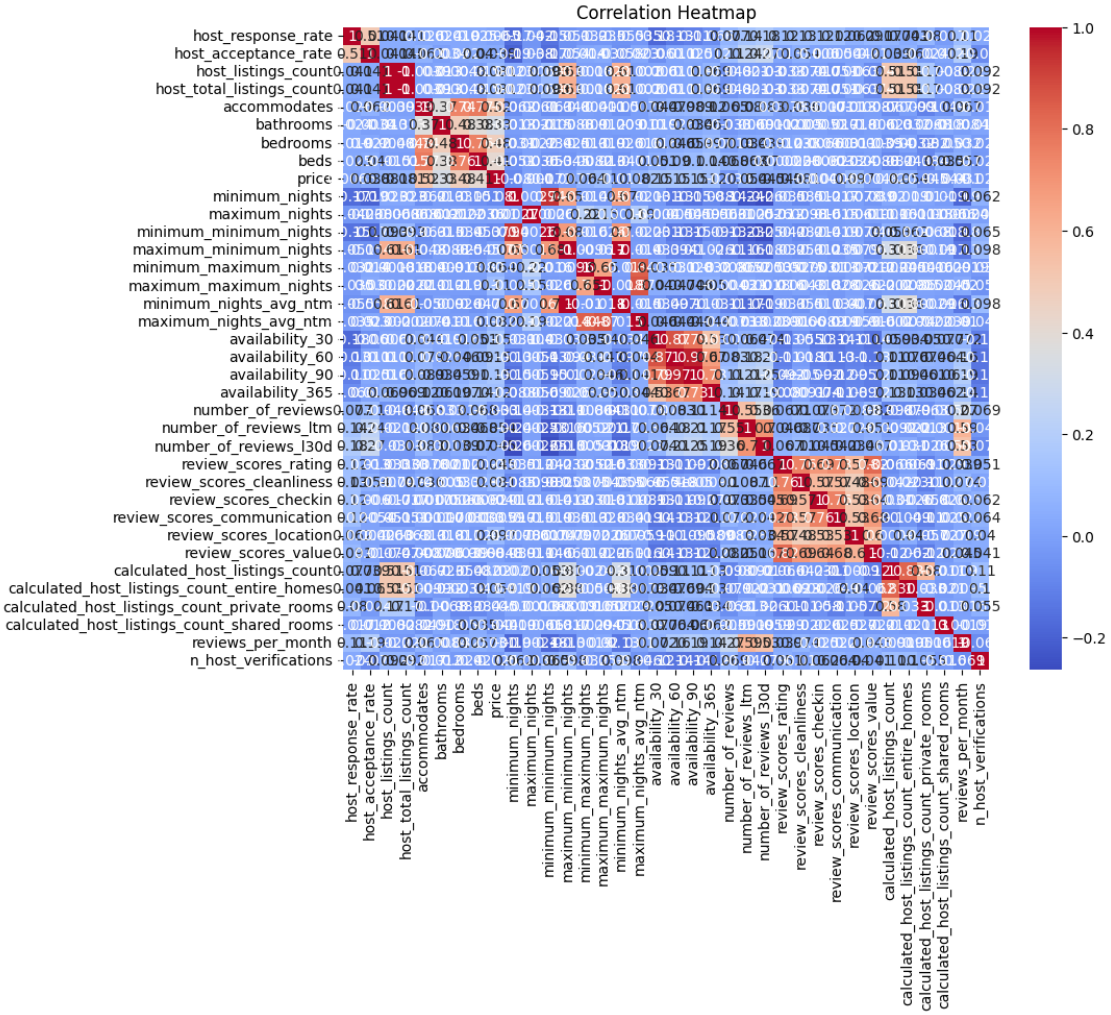
```
minimum_minimum_nights                          0
maximum_minimum_nights                          0
minimum_maximum_nights                          0
maximum_maximum_nights                          0
minimum_nights_avg_ntm                          0
maximum_nights_avg_ntm                          0
has_availability                                0
availability_30                                 0
availability_60                                 0
availability_90                                 0
availability_365                                0
number_of_reviews                               0
number_of_reviews_ltm                           0
number_of_reviews_l30d                          0
review_scores_rating                            0
review_scores_cleanliness                       0
review_scores_checkin                           0
review_scores_communication                     0
review_scores_location                          0
review_scores_value                             0
instant_bookable                                0
calculated_host_listings_count                  0
calculated_host_listings_count_entire_homes     0
calculated_host_listings_count_private_rooms    0
calculated_host_listings_count_shared_rooms     0
reviews_per_month                               0
n_host_verifications                            0
dtype: int64
```

```
Data types:
name                                    object
description                             object
neighborhood_overview                   object
host_name                               object
host_location                           object
host_about                              object
host_response_rate                     float64
host_acceptance_rate                   float64
host_is_superhost                         bool
host_listings_count                    float64
host_total_listings_count              float64
host_has_profile_pic                      bool
host_identity_verified                    bool
neighbourhood_group_cleansed            object
room_type                               object
accommodates                             int64
```

```
bathrooms                                       float64
bedrooms                                        float64
beds                                            float64
amenities                                        object
price                                           float64
minimum_nights                                    int64
maximum_nights                                    int64
minimum_minimum_nights                          float64
maximum_minimum_nights                          float64
minimum_maximum_nights                          float64
maximum_maximum_nights                          float64
minimum_nights_avg_ntm                          float64
maximum_nights_avg_ntm                          float64
has_availability                                   bool
availability_30                                   int64
availability_60                                   int64
availability_90                                   int64
availability_365                                  int64
number_of_reviews                                 int64
number_of_reviews_ltm                             int64
number_of_reviews_l30d                            int64
review_scores_rating                            float64
review_scores_cleanliness                       float64
review_scores_checkin                           float64
review_scores_communication                     float64
review_scores_location                          float64
review_scores_value                             float64
instant_bookable                                   bool
calculated_host_listings_count                    int64
calculated_host_listings_count_entire_homes       int64
calculated_host_listings_count_private_rooms      int64
calculated_host_listings_count_shared_rooms       int64
reviews_per_month                               float64
n_host_verifications                              int64
dtype: object
```


Price Distribution with Outliers
0
200
400
600
800
1000
price


Histogram of Price
Frequency
3500
3000
2500
2000
1500
1000
500
0
0
200
400
600
800
1000
Price



[5]:
```
# Load the dataset
df = pd.read_csv(airbnbDataSet_filename)

# Show the first few rows
print("Preview of dataset:")
display(df.head())

# Overview of data types and non-null counts
print("\nDataFrame info:")
df.info()

# Summary statistics
print("\nSummary statistics:")
display(df.describe())

# Check for missing values
```

```
print("\nMissing values per column:")
display(df.isnull().sum())

# Check data types
print("\nData types:")
print(df.dtypes)

# Boxplot for price to check outliers
plt.figure(figsize=(10, 4))
sns.boxplot(x=df['price'])
plt.title("Price Distribution with Outliers")
plt.show()

# Distribution plot of price
plt.figure(figsize=(10, 4))
sns.histplot(df['price'], bins=50, kde=True)
plt.title("Histogram of Price")
plt.xlabel("Price")
plt.ylabel("Frequency")
plt.show()

# Correlation heatmap for numerical features
plt.figure(figsize=(10, 8))
sns.heatmap(df.select_dtypes(include=['number']).corr(), annot=True,␣
↪cmap="coolwarm")
plt.title("Correlation Heatmap")
plt.show()
```

```
Preview of dataset:

                                                name  \
0                                Skylit Midtown Castle
1  Whole flr w/private bdrm, bath & kitchen(pls r...
2            Spacious Brooklyn Duplex, Patio + Garden
3                    Large Furnished Room Near B'way
4                  Cozy Clean Guest Room - Family Apt

                                           description  \
0  Beautiful, spacious skylit studio in the heart...
1  Enjoy 500 s.f. top floor in 1899 brownstone, w...
2  We welcome you to stay in our lovely 2 br dupl...
3  Please don't expect the luxury here just a bas...
4  Our best guests are seeking a safe, clean, spa...

                                neighborhood_overview    host_name  \
0  Centrally located in the heart of Manhattan ju...     Jennifer
1  Just the right mix of urban center and local n...  LisaRoxanne
2                                                NaN      Rebecca
```

```
3    Theater district, many restaurants around here.     Shunichi
4  Our neighborhood is full of restaurants and ca...    MaryEllen

                     host_location  \
0  New York, New York, United States
1  New York, New York, United States
2  Brooklyn, New York, United States
3  New York, New York, United States
4  New York, New York, United States

                                        host_about  host_response_rate  \
0  A New Yorker since 2000! My passion is creatin...                0.80
1  Laid-back Native New Yorker (formerly bi-coast...                0.09
2  Rebecca is an artist/designer, and Henoch is i...                1.00
3  I used to work for a financial industry but no...                1.00
4  Welcome to family life with my oldest two away...                 NaN

   host_acceptance_rate  host_is_superhost  host_listings_count  ...  \
0                  0.17               True                  8.0  ...
1                  0.69               True                  1.0  ...
2                  0.25               True                  1.0  ...
3                  1.00               True                  1.0  ...
4                   NaN               True                  1.0  ...

   review_scores_communication  review_scores_location  review_scores_value  \
0                         4.79                    4.86                 4.41
1                         4.80                    4.71                 4.64
2                         5.00                    4.50                 5.00
3                         4.42                    4.87                 4.36
4                         4.95                    4.94                 4.92

  instant_bookable calculated_host_listings_count  \
0            False                               3
1            False                               1
2            False                               1
3            False                               1
4            False                               1

   calculated_host_listings_count_entire_homes  \
0                                            3
1                                            1
2                                            1
3                                            0
4                                            0

   calculated_host_listings_count_private_rooms  \
0                                             0
1                                             0
```

```
2                                                 0
3                                                 1
4                                                 1

   calculated_host_listings_count_shared_rooms  reviews_per_month  \
0                                            0               0.33
1                                            0               4.86
2                                            0               0.02
3                                            0               3.68
4                                            0               0.87

  n_host_verifications
0                    9
1                    6
2                    3
3                    4
4                    7

[5 rows x 50 columns]


DataFrame info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 28022 entries, 0 to 28021
Data columns (total 50 columns):
 #   Column                                        Non-Null Count  Dtype
---  ------                                        --------------  -----
 0   name                                          28017 non-null  object
 1   description                                   27452 non-null  object
 2   neighborhood_overview                         18206 non-null  object
 3   host_name                                     28022 non-null  object
 4   host_location                                 27962 non-null  object
 5   host_about                                    17077 non-null  object
 6   host_response_rate                            16179 non-null  float64
 7   host_acceptance_rate                          16909 non-null  float64
 8   host_is_superhost                             28022 non-null  bool
 9   host_listings_count                           28022 non-null  float64
 10  host_total_listings_count                     28022 non-null  float64
 11  host_has_profile_pic                          28022 non-null  bool
 12  host_identity_verified                        28022 non-null  bool
 13  neighbourhood_group_cleansed                  28022 non-null  object
 14  room_type                                     28022 non-null  object
 15  accommodates                                  28022 non-null  int64
 16  bathrooms                                     28022 non-null  float64
 17  bedrooms                                      25104 non-null  float64
 18  beds                                          26668 non-null  float64
 19  amenities                                     28022 non-null  object
```

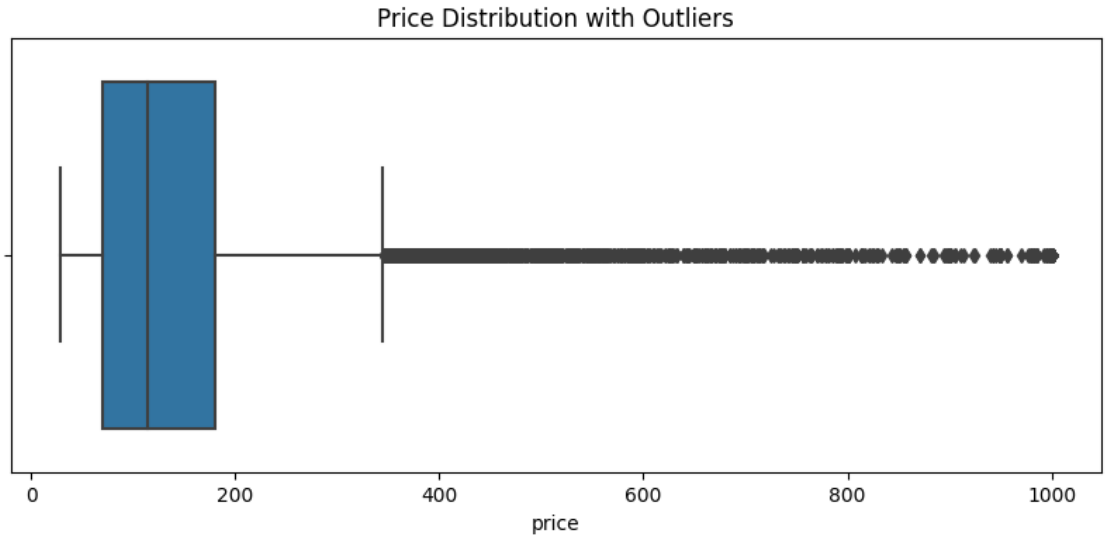
```
 20  price                                          28022 non-null  float64
 21  minimum_nights                                 28022 non-null  int64
 22  maximum_nights                                 28022 non-null  int64
 23  minimum_minimum_nights                         28022 non-null  float64
 24  maximum_minimum_nights                         28022 non-null  float64
 25  minimum_maximum_nights                         28022 non-null  float64
 26  maximum_maximum_nights                         28022 non-null  float64
 27  minimum_nights_avg_ntm                         28022 non-null  float64
 28  maximum_nights_avg_ntm                         28022 non-null  float64
 29  has_availability                               28022 non-null  bool
 30  availability_30                                28022 non-null  int64
 31  availability_60                                28022 non-null  int64
 32  availability_90                                28022 non-null  int64
 33  availability_365                               28022 non-null  int64
 34  number_of_reviews                              28022 non-null  int64
 35  number_of_reviews_ltm                          28022 non-null  int64
 36  number_of_reviews_l30d                         28022 non-null  int64
 37  review_scores_rating                           28022 non-null  float64
 38  review_scores_cleanliness                      28022 non-null  float64
 39  review_scores_checkin                          28022 non-null  float64
 40  review_scores_communication                    28022 non-null  float64
 41  review_scores_location                         28022 non-null  float64
 42  review_scores_value                            28022 non-null  float64
 43  instant_bookable                               28022 non-null  bool
 44  calculated_host_listings_count                 28022 non-null  int64
 45  calculated_host_listings_count_entire_homes    28022 non-null  int64
 46  calculated_host_listings_count_private_rooms   28022 non-null  int64
 47  calculated_host_listings_count_shared_rooms    28022 non-null  int64
 48  reviews_per_month                              28022 non-null  float64
 49  n_host_verifications                           28022 non-null  int64
dtypes: bool(5), float64(21), int64(15), object(9)
memory usage: 9.8+ MB

Summary statistics:

       host_response_rate  host_acceptance_rate  host_listings_count  \
count        16179.000000          16909.000000         28022.000000
mean             0.906901              0.791953            14.554778
std              0.227282              0.276732           120.721287
min              0.000000              0.000000             0.000000
25%              0.940000              0.680000             1.000000
50%              1.000000              0.910000             1.000000
75%              1.000000              1.000000             3.000000
max              1.000000              1.000000          3387.000000

       host_total_listings_count  accommodates     bathrooms      bedrooms  \
count               28022.000000  28022.000000  28022.000000  25104.000000
mean                   14.554778      2.874491      1.142174      1.329708
```

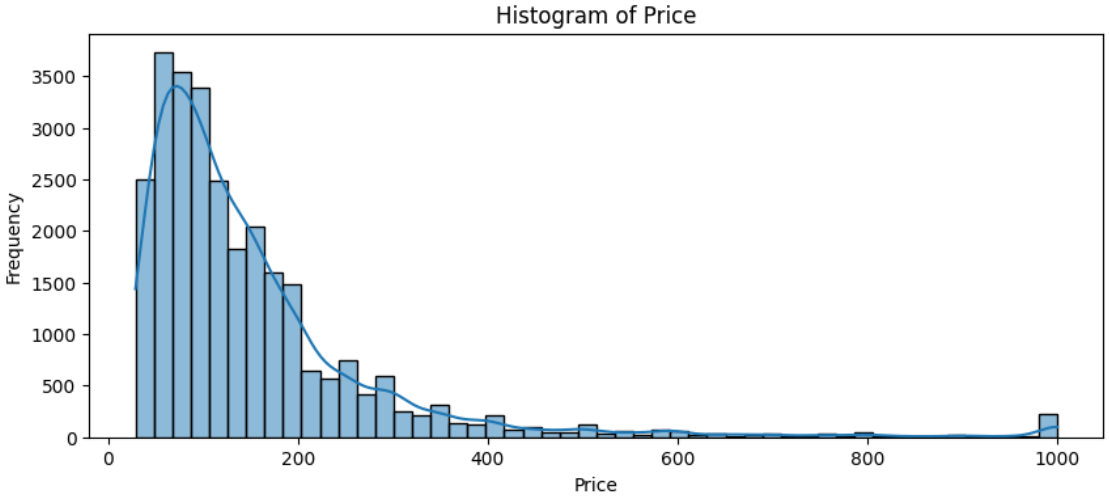
```
std                  120.721287      1.860251      0.421132      0.700726
min                    0.000000      1.000000      0.000000      1.000000
25%                    1.000000      2.000000      1.000000      1.000000
50%                    1.000000      2.000000      1.000000      1.000000
75%                    3.000000      4.000000      1.000000      1.000000
max                 3387.000000     16.000000      8.000000     12.000000

                beds         price  minimum_nights  ...  review_scores_checkin  \
count  26668.000000  28022.000000    28022.000000  ...           28022.000000
mean       1.629556    154.228749       18.689387  ...               4.814300
std        1.097104    140.816605       25.569151  ...               0.438603
min        1.000000     29.000000        1.000000  ...               0.000000
25%        1.000000     70.000000        2.000000  ...               4.810000
50%        1.000000    115.000000       30.000000  ...               4.960000
75%        2.000000    180.000000       30.000000  ...               5.000000
max       21.000000   1000.000000     1250.000000  ...               5.000000

       review_scores_communication  review_scores_location  \
count                 28022.000000            28022.000000
mean                      4.808041                4.750393
std                       0.464585                0.415717
min                       0.000000                0.000000
25%                       4.810000                4.670000
50%                       4.970000                4.880000
75%                       5.000000                5.000000
max                       5.000000                5.000000

       review_scores_value  calculated_host_listings_count  \
count         28022.000000                    28022.000000
mean              4.647670                        9.581900
std               0.518023                       32.227523
min               0.000000                        1.000000
25%               4.550000                        1.000000
50%               4.780000                        1.000000
75%               5.000000                        3.000000
max               5.000000                      421.000000

       calculated_host_listings_count_entire_homes  \
count                                 28022.000000
mean                                      5.562986
std                                      26.121426
min                                       0.000000
25%                                       0.000000
50%                                       1.000000
75%                                       1.000000
max                                     308.000000

       calculated_host_listings_count_private_rooms  \
```

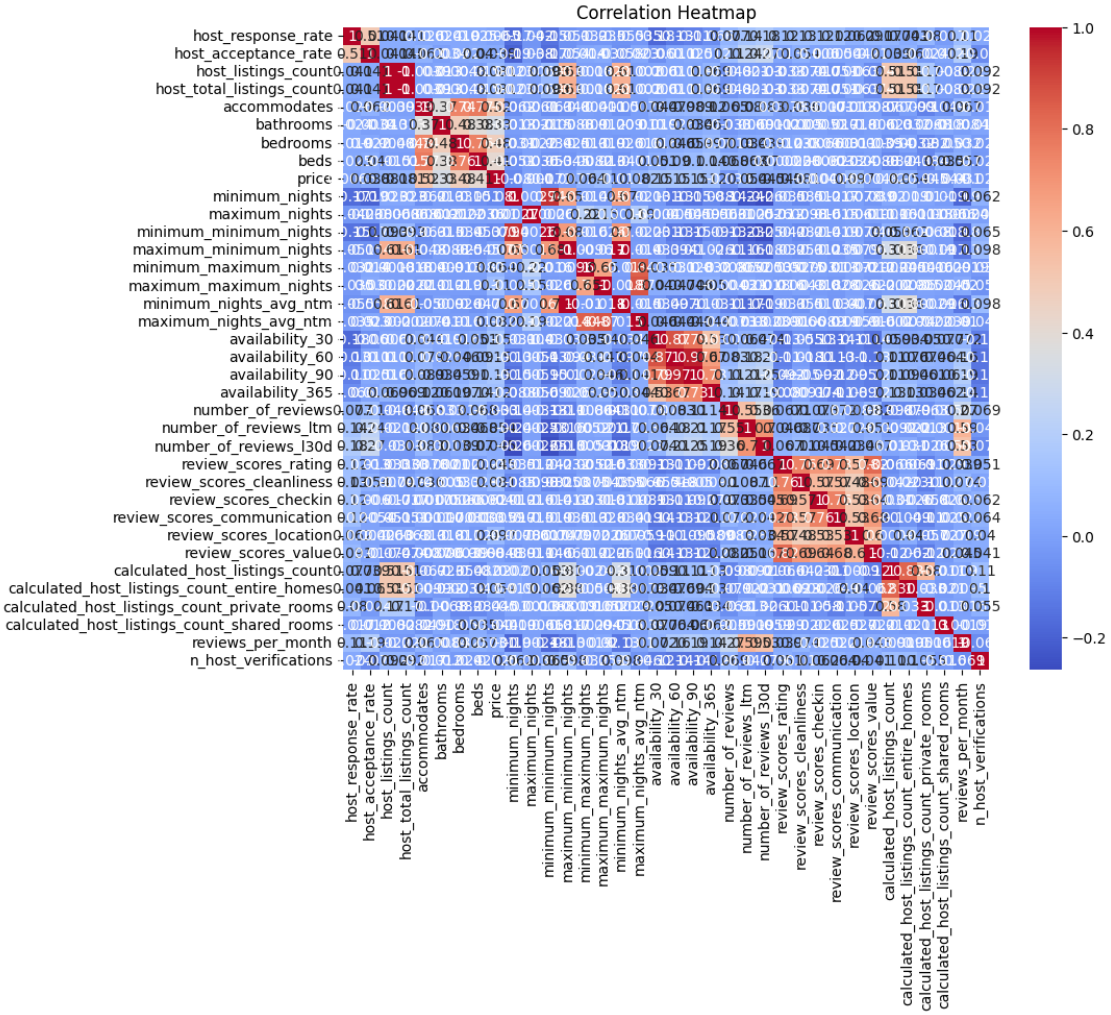
```
count                                   28022.000000
mean                                        3.902077
std                                        17.972386
min                                         0.000000
25%                                         0.000000
50%                                         0.000000
75%                                         1.000000
max                                       359.000000

       calculated_host_listings_count_shared_rooms  reviews_per_month  \
count                                  28022.000000       28022.000000
mean                                       0.048283           1.758325
std                                        0.442459           4.446143
min                                        0.000000           0.010000
25%                                        0.000000           0.130000
50%                                        0.000000           0.510000
75%                                        0.000000           1.830000
max                                        8.000000         141.000000

       n_host_verifications
count          28022.000000
mean               5.169510
std                2.028497
min                1.000000
25%                4.000000
50%                5.000000
75%                7.000000
max               13.000000

[8 rows x 36 columns]


Missing values per column:

name                                        5
description                               570
neighborhood_overview                    9816
host_name                                   0
host_location                              60
host_about                              10945
host_response_rate                      11843
host_acceptance_rate                    11113
host_is_superhost                           0
host_listings_count                         0
host_total_listings_count                   0
host_has_profile_pic                        0
host_identity_verified                      0
neighbourhood_group_cleansed                0
```

```
room_type                                        0
accommodates                                     0
bathrooms                                        0
bedrooms                                      2918
beds                                          1354
amenities                                        0
price                                            0
minimum_nights                                   0
maximum_nights                                   0
minimum_minimum_nights                           0
maximum_minimum_nights                           0
minimum_maximum_nights                           0
maximum_maximum_nights                           0
minimum_nights_avg_ntm                           0
maximum_nights_avg_ntm                           0
has_availability                                 0
availability_30                                  0
availability_60                                  0
availability_90                                  0
availability_365                                 0
number_of_reviews                                0
number_of_reviews_ltm                            0
number_of_reviews_l30d                           0
review_scores_rating                             0
review_scores_cleanliness                        0
review_scores_checkin                            0
review_scores_communication                      0
review_scores_location                           0
review_scores_value                              0
instant_bookable                                 0
calculated_host_listings_count                   0
calculated_host_listings_count_entire_homes      0
calculated_host_listings_count_private_rooms     0
calculated_host_listings_count_shared_rooms      0
reviews_per_month                                0
n_host_verifications                             0
dtype: int64


Data types:
name                                          object
description                                   object
neighborhood_overview                         object
host_name                                     object
host_location                                 object
host_about                                    object
host_response_rate                           float64
```

```
host_acceptance_rate                            float64
host_is_superhost                                  bool
host_listings_count                             float64
host_total_listings_count                       float64
host_has_profile_pic                               bool
host_identity_verified                             bool
neighbourhood_group_cleansed                     object
room_type                                        object
accommodates                                      int64
bathrooms                                       float64
bedrooms                                        float64
beds                                            float64
amenities                                        object
price                                           float64
minimum_nights                                    int64
maximum_nights                                    int64
minimum_minimum_nights                          float64
maximum_minimum_nights                          float64
minimum_maximum_nights                          float64
maximum_maximum_nights                          float64
minimum_nights_avg_ntm                          float64
maximum_nights_avg_ntm                          float64
has_availability                                   bool
availability_30                                   int64
availability_60                                   int64
availability_90                                   int64
availability_365                                  int64
number_of_reviews                                 int64
number_of_reviews_ltm                             int64
number_of_reviews_l30d                            int64
review_scores_rating                            float64
review_scores_cleanliness                       float64
review_scores_checkin                           float64
review_scores_communication                     float64
review_scores_location                          float64
review_scores_value                             float64
instant_bookable                                   bool
calculated_host_listings_count                    int64
calculated_host_listings_count_entire_homes       int64
calculated_host_listings_count_private_rooms      int64
calculated_host_listings_count_shared_rooms       int64
reviews_per_month                               float64
n_host_verifications                              int64
dtype: object
```


Price Distribution with Outliers
0
200
400
600
800
1000
price


Histogram of Price
Frequency
3500
3000
2500
2000
1500
1000
500
0
0
200
400
600
800
1000
Price



[ ]:
```
1.
   - I handled missing values using `.isnull().sum()` to identify them.
   - I will impute missing numerical values with the mean and remove columns␣
↪with excessive missing data.
   - I used `boxplot` to detect outliers in `price`, and plan to remove␣
↪listings with prices > $1000.
   - I plan to one-hot encode categorical features like `room_type` and␣
↪`neighbourhood_group`.
   - I will scale features if needed for models that are sensitive to feature␣
↪range.

2. **Model Selection:**
   - I plan to use Linear Regression as a baseline model because it's easy to␣
↪interpret.
```

```
   - I may also try Random Forest or Gradient Boosting to capture nonlinear␣
↪patterns.
   - I may need to normalize numerical features if using certain models like␣
↪Ridge or KNN.

3. **Model Evaluation and Improvement:**
   - I will use Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) as␣
↪evaluation metrics.
   - I will perform a train/test split or use cross-validation.
   - I will improve the model by:
     - Removing outliers,
     - Tuning hyperparameters,
     - Trying ensemble models for better performance.
```

1.5 Part 4: Define Your Project Plan

Now that you understand your data, in the markdown cell below, define your plan to implement the remaining phases of the machine learning life cycle (data preparation, modeling, evaluation) to solve your ML problem. Answer the following questions:

- Do you have a new feature list? If so, what are the features that you chose to keep and remove after inspecting the data?
- Explain different data preparation techniques that you will use to prepare your data for modeling.
- What is your model (or models)?
- Describe your plan to train your model, analyze its performance and then improve the model. That is, describe your model building, validation and selection plan to produce a model that generalizes well to new data.

1.6 Part 5: Implement Your Project Plan

Task: In the code cell below, import additional packages that you have used in this course that you will need to implement your project plan.

[6]:
```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
```

[7]:
```
# 1. Load the dataset
df = pd.read_csv(airbnbDataSet_filename)

# 2. Drop irrelevant columns (if they exist)
columns_to_drop = ['id', 'name', 'host_name', 'last_review']
df = df.drop(columns=[col for col in columns_to_drop if col in df.columns])
```

```
# 3. Fill missing values for reviews_per_month
if 'reviews_per_month' in df.columns:
    df['reviews_per_month'] = df['reviews_per_month'].fillna(0)

# 4. One-hot encode categorical variables
categorical_cols = ['room_type', 'neighbourhood_group']
df = pd.get_dummies(df, columns=[col for col in categorical_cols if col in df.
 ↪columns], drop_first=True)

# 5. Define features (X) and target (y)
X = df.drop(columns=['price'])
y = df['price']

# Fill all numeric NaN in X with column mean
numeric_cols = X.select_dtypes(include=['number']).columns
X[numeric_cols] = X[numeric_cols].fillna(X[numeric_cols].mean())

# Keep only numeric columns (exclude object/text)
X = X.select_dtypes(include=['number'])

# 6. Feature scaling
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# 7. Split data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2,␣
 ↪random_state=42)

# 8. Train Linear Regression model
lr_model = LinearRegression()
lr_model.fit(X_train, y_train)
y_pred_lr = lr_model.predict(X_test)

# 9. Train Random Forest Regressor
rf_model = RandomForestRegressor(random_state=42)
rf_model.fit(X_train, y_train)
y_pred_rf = rf_model.predict(X_test)

# 10. Evaluate models
print("=== Linear Regression ===")
print("MAE:", mean_absolute_error(y_test, y_pred_lr))
print("MSE:", mean_squared_error(y_test, y_pred_lr))
print("R²:", r2_score(y_test, y_pred_lr))

print("\n=== Random Forest Regressor ===")
print("MAE:", mean_absolute_error(y_test, y_pred_rf))
```

```
print("MSE:", mean_squared_error(y_test, y_pred_rf))
print("R²:", r2_score(y_test, y_pred_rf))
```

```
=== Linear Regression ===
MAE: 64.20642427396994
MSE: 10983.69294838066
R²: 0.41852552915875285

=== Random Forest Regressor ===
MAE: 53.310656388452976
MSE: 8762.09622937331
R²: 0.53613640763819
```

[ ]: