Connected to .DATA-PROJ-VENV (Python 3.12.4)

In [ ]:
```python
# Import necessary libraries
import os
import numpy as np
import pandas as pd
```

In [ ]:
```python
# Read file from root .\data\raw\
script_dir = os.path.dirname(__file__)
listings_file_path = os.path.join(script_dir, "../../data/raw/listings.csv")
```

In [ ]:
```python
# Read the CSV file into a DataFrame
df_listings = pd.read_csv(listings_file_path)
```

In [ ]:
```python
# Preview Dataframes
df_listings.head()
```

Out[ ]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availabi
0	197677	Oshiage Holiday Apartment ...	964081	Yoshimi & Marek ...		Sumida Ku	35.717070	139.826080	Entire home/apt	12000	3	176	2024-05-15	1.13	1	
1	776070	Kero-kero house room 1 ...	801494	Kei ...		Kita Ku	35.738440	139.769170	Private room	9652	3	256	2024-06-03	1.81	1	
2	905944	4F - Near Shinjuku & Shibuya ...	4847803	Best Stay In Tokyo! ...		Shibuya Ku	35.678780	139.678470	Entire home/apt	25738	3	219	2024-06-19	1.60	6	
3	1016831	5 mins Shibuya Cat modern sunny Shimokita ...	5596383	Wakana ...		Setagaya Ku	35.658000	139.671340	Private room	23286	10	268	2024-06-26	1.96	2	
4	1196177	Stay with host Cozy private room Senju area ...	5686404	Yukiko ...		Adachi Ku	35.744731	139.797384	Private room	7500	2	120	2024-06-15	0.91	1	

In [ ]:
```python
# Overview of df_listings
df_listings.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16518 entries, 0 to 16517
Data columns (total 18 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   id                               16518 non-null  int64
 1   name                             16518 non-null  object
 2   host_id                          16518 non-null  int64
 3   host_name                        16518 non-null  object
 4   neighbourhood_group              16518 non-null  object
 5   neighbourhood                    16518 non-null  object
 6   latitude                         16518 non-null  float64
 7   longitude                        16518 non-null  float64
 8   room_type                        16518 non-null  object
 9   price                            16518 non-null  object
 10  minimum_nights                   16518 non-null  int64
 11  number_of_reviews                16518 non-null  int64
 12  last_review                      16518 non-null  object
 13  reviews_per_month                16518 non-null  object
 14  calculated_host_listings_count   16518 non-null  int64
 15  availability_365                 16518 non-null  int64
 16  number_of_reviews_ltm            16518 non-null  int64
 17  license                          16508 non-null  object
dtypes: float64(2), int64(7), object(9)
memory usage: 2.3+ MB
```

In [ ]:
```python
df_listings.describe()
```

Out[ ]:

	id	host_id	latitude	longitude	minimum_nights	number_of_reviews	calculated_host_listings_count	availability_365	number_of_reviews_ltm
count	1.651800e+04	1.651800e+04	16518.000000	16518.000000	16518.000000	16518.000000	16518.000000	16518.000000	16518.000000
mean	5.816491e+17	2.929799e+08	35.698748	139.736294	3.648868	37.783691	16.109335	154.140756	15.071013
std	4.970662e+17	1.846269e+08	0.041095	0.072995	8.920520	64.014216	19.674860	99.525334	21.505364
min	1.976770e+05	3.222340e+05	35.520940	139.081322	1.000000	0.000000	1.000000	0.000000	0.000000
25%	4.144522e+07	1.290963e+08	35.688305	139.699183	1.000000	4.000000	3.000000	75.000000	2.000000
50%	8.301174e+17	2.721294e+08	35.703990	139.728136	2.000000	17.000000	9.000000	144.000000	10.000000
75%	1.040939e+18	4.923137e+08	35.722587	139.790456	2.000000	45.000000	19.000000	234.000000	22.000000
max	1.189054e+18	5.859819e+08	35.840764	139.914020	365.000000	2660.000000	106.000000	365.000000	815.000000

In [ ]:
```python
df_listings.columns
```

Out[ ]:
```
Index(['id                  ',
       'name                                                                                         ',
       'host_id   ', 'host_name                                    ',
       'neighbourhood_group ', 'neighbourhood       ', 'latitude           ',
       'longitude          ', 'room_type       ', 'price  ', 'minimum_nights ',
       'number_of_reviews ', 'last_review ', 'reviews_per_month ',
       'calculated_host_listings_count ', 'availability_365 ',
       'number_of_reviews_ltm ', 'license'],
      dtype='object')
```

In [ ]:
```python
# Cleaning df_listings
listings_clean = (
    df_listings.copy()
)  # Creating copy of the df_listings before making changes
```

In [ ]:
```python
# Cleaning column names since they contain white spaces
listings_clean.columns = listings_clean.columns.str.strip().str.lower()
listings_clean = (
    listings_clean.drop(
        columns=[
            "neighbourhood_group",
            "minimum_nights",
            "number_of_reviews",
            "last_review",
            "reviews_per_month",
            "calculated_host_listings_count",
            "availability_365",
            "number_of_reviews_ltm",
            "license",
        ]
    ).drop_duplicates()  # Dropping duplicate data
)
```

In [ ]:
```python
# Quick check to see if changes were made
listings_clean.head()
```

Out[ ]:

	id	name	host_id	host_name	neighbourhood	latitude	longitude	room_type	price
0	197677	Oshiage Holiday Apartment ...	964081	Yoshimi & Marek ...	Sumida Ku	35.717070	139.826080	Entire home/apt	12000
1	776070	Kero-kero house room 1 ...	801494	Kei ...	Kita Ku	35.738440	139.769170	Private room	9652
2	905944	4F - Near Shinjuku & Shibuya ...	4847803	Best Stay In Tokyo! ...	Shibuya Ku	35.678780	139.678470	Entire home/apt	25738
3	1016831	5 mins Shibuya Cat modern sunny Shimokita ...	5596383	Wakana ...	Setagaya Ku	35.658000	139.671340	Private room	23286
4	1196177	Stay with host Cozy private room Senju area ...	5686404	Yukiko ...	Adachi Ku	35.744731	139.797384	Private room	7500

In [ ]:
```python
listings_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16518 entries, 0 to 16517
Data columns (total 9 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   id             16518 non-null  int64
 1   name           16518 non-null  object
 2   host_id        16518 non-null  int64
 3   host_name      16518 non-null  object
 4   neighbourhood  16518 non-null  object
 5   latitude       16518 non-null  float64
 6   longitude      16518 non-null  float64
 7   room_type      16518 non-null  object
 8   price          16518 non-null  object
dtypes: float64(2), int64(2), object(5)
memory usage: 1.1+ MB
```

In [ ]:
```python
# Replacing non-ASCII characters with blank spaces.
listings_clean["name"] = listings_clean["name"].apply(
    lambda x: "" if any(ord(char) > 127 for char in x) else x
)
listings_clean["host_name"] = listings_clean["host_name"].apply(
    lambda x: "" if any(ord(char) > 127 for char in x) else x
)
```

In [ ]:
```python
# Replace empty strings in the 'price' column with NaN.
listings_clean["price"] = pd.to_numeric(
    listings_clean["price"].replace("", np.nan), errors="coerce"
)
```

In [ ]:
```python
# Drop rows with NaN values in 'price'
listings_clean = listings_clean.dropna(subset=["price"])
```

In [ ]:
```python
# Convert to int64 (this removes decimal places)
listings_clean["price"] = listings_clean["price"].astype(int)
```

In [ ]:
```python
# Quick check to see if changes were made
listings_clean.head()
```

Out[ ]:

	id	name	host_id	host_name	neighbourhood	latitude	longitude	room_type	price
0	197677	Oshiage Holiday Apartment ...	964081	Yoshimi & Marek ...	Sumida Ku	35.717070	139.826080	Entire home/apt	12000
1	776070	Kero-kero house room 1 ...	801494	Kei ...	Kita Ku	35.738440	139.769170	Private room	9652
2	905944	4F - Near Shinjuku & Shibuya ...	4847803	Best Stay In Tokyo! ...	Shibuya Ku	35.678780	139.678470	Entire home/apt	25738
3	1016831	5 mins Shibuya Cat modern sunny Shimokita ...	5596383	Wakana ...	Setagaya Ku	35.658000	139.671340	Private room	23286
4	1196177	Stay with host Cozy private room Senju area ...	5686404	Yukiko ...	Adachi Ku	35.744731	139.797384	Private room	7500

In [ ]:
```python
listings_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 14805 entries, 0 to 16517
Data columns (total 9 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   id             14805 non-null  int64
 1   name           14805 non-null  object
 2   host_id        14805 non-null  int64
 3   host_name      14805 non-null  object
 4   neighbourhood  14805 non-null  object
 5   latitude       14805 non-null  float64
 6   longitude      14805 non-null  float64
 7   room_type      14805 non-null  object
 8   price          14805 non-null  int64
dtypes: float64(2), int64(3), object(4)
memory usage: 1.1+ MB
```

In [ ]:
```python
# Create a copy of the cleaned listings DataFrame
df_listings_cleaned = listings_clean.copy()
```

In [ ]:
```python
df_listings_cleaned.head()
```

Out[ ]:

	id	name	host_id	host_name	neighbourhood	latitude	longitude	room_type	price
0	197677	Oshiage Holiday Apartment ...	964081	Yoshimi & Marek ...	Sumida Ku	35.717070	139.826080	Entire home/apt	12000
1	776070	Kero-kero house room 1 ...	801494	Kei ...	Kita Ku	35.738440	139.769170	Private room	9652
2	905944	4F - Near Shinjuku & Shibuya ...	4847803	Best Stay In Tokyo! ...	Shibuya Ku	35.678780	139.678470	Entire home/apt	25738
3	1016831	5 mins Shibuya Cat modern sunny Shimokita ...	5596383	Wakana ...	Setagaya Ku	35.658000	139.671340	Private room	23286
4	1196177	Stay with host Cozy private room Senju area ...	5686404	Yukiko ...	Adachi Ku	35.744731	139.797384	Private room	7500

In [ ]:
```python
# Creating directory path for export of cleaned data.
clean_data_dir = os.path.join("..", "..", "data", "clean")
cleaned_listings_export_path = os.path.abspath(
    os.path.join(script_dir, clean_data_dir, "cleaned_listings.csv")
)
```

In [ ]:
```python
# Exporting cleaned data to directory.
df_listings_cleaned.to_csv(cleaned_listings_export_path, index=False)
print("Data Cleaning Completed!")
```

```
Data Cleaning Completed!
```