# Pre-requisite

- Understanding of Python, Power BI or Tableau
- Understanding of Data Cleaning
- Understanding Data Visualization

## Data Analytics of Airbnb Data:

### Objective:

In this exericise, you will be performing Data Analytics on an Open Dataset dataset coming from Airbnb. Some of the tasks include

- Data Cleaning.
- Data Transformation
- Data Visualization.

### Overview of Airbnb Data:

People's main criteria when visiting new places are reasonable accommodation and food. Airbnb (Air-Bed-Breakfast) is an online marketplace created to meet this need of people by renting out their homes for a short term. They offer this facility at a relatively lower price than hotels. Further people worldwide prefer the homely and economical service offered by them. They offer services across various geographical locations

### Dataset Source

YOu can get the dataset for this assessment using the following link:
https://www.kaggle.com/datasets/arianazmoudeh/airbnbopendata

This dataset contains information such as the neighborhood offering these services, room type, price,avaliabilty, reviews, service fee, cancellation policy and rules to use the house. This analysis will help airbnb in improving its services.

So all the best for your Data Analytics Journey on Airbnb data!!!

# Task 1: Data Loading (Python)

1. Read the csv file and load it into a pandas dataframe.
2. Display the first five rows of your dataframe.
3. Display the data types of the columns.

```
## Read the csv fileimport pandas as pd
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Read the CSV file
```

```
#df = pd.read_csv(r'C:\Users\Vuxenutbildningen\Downloads\archive\
Airbnb_Open_Data.csv')
data_file = '~/Downloads/archive/Airbnb_Open_Data.csv'

# Display the first five rows of the dataframe
df.head(5)
```

```
        id                                                NAME      host
id  \
0  1001254                Clean & quiet apt home by the park
80014485718
1  1002102                               Skylit Midtown Castle
52335172823
2  1002403               THE VILLAGE OF HARLEM....NEW YORK !
78829239556
3  1002755                                                NaN
85098326012
4  1003689  Entire Apt: Spacious Studio/Loft by central park
92037596077

  host_identity_verified host name neighbourhood group
neighbourhood  \
0            unconfirmed  Madaline            Brooklyn    Kensington

1               verified     Jenna           Manhattan       Midtown

2                    NaN     Elise           Manhattan        Harlem

3            unconfirmed     Garry            Brooklyn  Clinton Hill

4               verified    Lyndon           Manhattan   East Harlem


        lat      long        country  ... service fee minimum
nights  \
0  40.64749 -73.97237  United States  ...       $193            10.0

1  40.75362 -73.98377  United States  ...        $28            30.0

2  40.80902 -73.94190  United States  ...       $124             3.0

3  40.68514 -73.95976  United States  ...        $74            30.0

4  40.79851 -73.94399  United States  ...        $41            10.0


  number of reviews last review  reviews per month review rate number
\
0               9.0  10/19/2021               0.21                4.0

1              45.0   5/21/2022               0.38                4.0
```

```
2                0.0         NaN                  NaN                   5.0

3              270.0    7/5/2019                 4.64                   4.0

4                9.0  11/19/2018                 0.10                   3.0


  calculated host listings count  availability 365  \
0                            6.0             286.0
1                            2.0             228.0
2                            1.0             352.0
3                            1.0             322.0
4                            1.0             289.0

                                          house_rules license
0  Clean up and treat the home the way you'd like...     NaN
1  Pet friendly but please confirm with me if the...     NaN
2  I encourage you to use my kitchen, cooking and...     NaN
3                                                NaN     NaN
4  Please no smoking in the house, porch or on th...     NaN

[5 rows x 26 columns]
```

```
# Display the data types of the columns
df.dtypes
```

```
id                                int64
NAME                             object
host id                           int64
host_identity_verified           object
host name                        object
neighbourhood group              object
neighbourhood                    object
lat                             float64
long                            float64
country                          object
country code                     object
instant_bookable                 object
cancellation_policy              object
room type                        object
Construction year               float64
price                            object
service fee                      object
minimum nights                  float64
number of reviews               float64
last review                      object
reviews per month               float64
review rate number              float64
calculated host listings count  float64
```

```
availability 365                  float64
house_rules                        object
license                            object
dtype: object
```

# Task 2a: Data Cleaning (Any Tool)

1. Drop some of the unwanted columns. These include `host id`, `id`, `country` and `country code` from the dataset.
2. State the reason for not including these columns for your Data Analytics.

If using Python for this exercise, please include the code in the cells below. If using any other tool, please include screenshoots before and after the elimination of the columns.

```
# Drop the unwanted columns

df = df.drop(['host id', 'id', 'country', 'country code'], axis=1)
# Display the updated dataframe
df.head()
```

```
                                                NAME
host_identity_verified  \
0                Clean & quiet apt home by the park
unconfirmed
1                            Skylit Midtown Castle
verified
2               THE VILLAGE OF HARLEM....NEW YORK !
NaN
3                                                NaN
unconfirmed
4  Entire Apt: Spacious Studio/Loft by central park
verified

  host name neighbourhood group neighbourhood       lat      long  \
0  Madaline            Brooklyn    Kensington  40.64749 -73.97237
1     Jenna           Manhattan       Midtown  40.75362 -73.98377
2     Elise           Manhattan        Harlem  40.80902 -73.94190
3     Garry            Brooklyn  Clinton Hill  40.68514 -73.95976
4    Lyndon           Manhattan   East Harlem  40.79851 -73.94399

  instant_bookable cancellation_policy        room type  ...  service
fee  \
0            False              strict     Private room  ...
$193
1            False            moderate  Entire home/apt  ...
$28
2             True            flexible     Private room  ...
$124
3             True            moderate  Entire home/apt  ...
$74
```

```
4            False            moderate  Entire home/apt  ...
$41

  minimum nights number of reviews  last review  reviews per month  \
0           10.0               9.0   10/19/2021               0.21
1           30.0              45.0    5/21/2022               0.38
2            3.0               0.0           NaN                NaN
3           30.0             270.0     7/5/2019               4.64
4           10.0               9.0   11/19/2018               0.10

  review rate number  calculated host listings count  availability 365
\
0                4.0                             6.0             286.0

1                4.0                             2.0             228.0

2                5.0                             1.0             352.0

3                4.0                             1.0             322.0

4                3.0                             1.0             289.0


                                          house_rules  license
0  Clean up and treat the home the way you'd like...      NaN
1  Pet friendly but please confirm with me if the...      NaN
2  I encourage you to use my kitchen, cooking and...      NaN
3                                                NaN      NaN
4  Please no smoking in the house, porch or on th...      NaN

[5 rows x 22 columns]
```

# Task 2b: Data Cleaning (Python)

- Check for missing values in the dataframe and display the count in ascending order. **If the values are missing, impute the values as per the datatype of the columns.**
- Check whether there are any duplicate values in the dataframe and, if present, remove them.
- Display the total number of records in the dataframe before and after removing the duplicates.

```
## Check for missing values in the dataframe and display the count in
ascending order.

missing_values = df.isnull().sum().sort_values(ascending=True)
print(missing_values)
```

```
room type                                   0
lat                                         8
long                                        8
```

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

```
# Impute missing values based on column data types
for column in df.columns:
    if df[column].dtype == 'object':
        df[column] = df[column].fillna('Unknown')  # Replace missing
values with 'Unknown' for object columns
    elif df[column].dtype == 'float64':
        df[column] = df[column].fillna(df[column].mean())  # Replace
missing values with mean for float columns
    elif df[column].dtype == 'int64':
        df[column] = df[column].fillna(df[column].median())  # Replace
missing values with median for integer columns

df.head(30)
```

```
                                                 NAME
host_identity_verified  \
0                  Clean & quiet apt home by the park
unconfirmed
1                               Skylit Midtown Castle
verified
2                 THE VILLAGE OF HARLEM....NEW YORK !
Unknown
3                                             Unknown
unconfirmed
4    Entire Apt: Spacious Studio/Loft by central park
verified
5           Large Cozy 1 BR Apartment In Midtown East
verified
6                                     BlissArtsSpace!
```

```
Unknown
7                                      BlissArtsSpace!
unconfirmed
8                       Large Furnished Room Near B'way
verified
9                    Cozy Clean Guest Room - Family Apt
unconfirmed
10                   Cute & Cozy Lower East Side 1 bdrm
verified
11                     Beautiful 1br on Upper West Side
verified
12                      Central Manhattan/near Broadway
verified
13     Lovely Room 1, Garden, Best Area, Legal rental
verified
14   Wonderful Guest Bedroom in Manhattan for SINGLES
verified
15                        West Village Nest - Superhost
verified
16                     Only 2 stops to Manhattan studio
unconfirmed
17                    Perfect for Your Parents + Garden
verified
18                                     Chelsea Perfect
verified
19    Hip Historic Brownstone Apartment with Backyard
Unknown
20                    Huge 2 BR Upper East  Cental Park
verified
21                     Sweet and Spacious Brooklyn Loft
verified
22                       CBG CtyBGd HelpsHaiti rm#1:1-4
verified
23                              CBG Helps Haiti Room#2.5
Unknown
24                                 CBG Helps Haiti Rm #2
unconfirmed
25              MAISON DES SIRENES1,bohemian apartment
Unknown
26                   Sunny Bedroom Across Prospect Park
Unknown
27  Magnifique Suite au N de Manhattan - vue Cloitres
verified
28                                  Midtown Pied-a-terre
unconfirmed
29       SPACIOUS, LOVELY FURNISHED MANHATTAN BEDROOM
verified

    host name neighbourhood group       neighbourhood       lat
```

```
long  \
0    Madaline            Brooklyn          Kensington  40.64749 -
73.97237
1       Jenna           Manhattan             Midtown  40.75362 -
73.98377
2       Elise           Manhattan              Harlem  40.80902 -
73.94190
3       Garry            Brooklyn        Clinton Hill  40.68514 -
73.95976
4      Lyndon           Manhattan         East Harlem  40.79851 -
73.94399
5    Michelle           Manhattan         Murray Hill  40.74767 -
73.97500
6     Alberta            Brooklyn  Bedford-Stuyvesant  40.68688 -
73.95596
7        Emma            Brooklyn  Bedford-Stuyvesant  40.68688 -
73.95596
8      Evelyn           Manhattan      Hell's Kitchen  40.76489 -
73.98493
9        Carl           Manhattan     Upper West Side  40.80178 -
73.96723
10    Miranda           Manhattan           Chinatown  40.71344 -
73.99037
11       Alan           Manhattan     Upper West Side  40.80316 -
73.96545
12    Unknown           Manhattan      Hell's Kitchen  40.76076 -
73.98867
13      Darcy             brookln         South Slope  40.66829 -
73.98779
14   Leonardo           Manhattan     Upper West Side  40.79826 -
73.96113
15     Daniel           Manhattan        West Village  40.73530 -
74.00525
16    Heather            Brooklyn        Williamsburg  40.70837 -
73.95352
17       Ryan            Brooklyn         Fort Greene  40.69169 -
73.97185
18    Alberta            manhatan             Chelsea  40.74192 -
73.99501
19     Martin            Brooklyn       Crown Heights  40.67592 -
73.94694
20     Audrey           Manhattan         East Harlem  40.79685 -
73.94872
21     Alissa            Brooklyn        Williamsburg  40.71842 -
73.95718
22       Mary            Brooklyn          Park Slope  40.68069 -
73.97706
23    William            Brooklyn          Park Slope  40.67989 -
73.97798
```

```
24  Charlotte            Brooklyn          Park Slope  40.68001 -
73.97865
25    Miranda            Brooklyn  Bedford-Stuyvesant  40.68371 -
73.94028
26     Carlos            Brooklyn     Windsor Terrace  40.65599 -
73.97519
27   Adrianna           Manhattan              Inwood  40.86754 -
73.92639
28     Andrew           Manhattan      Hell's Kitchen  40.76715 -
73.98533
29      Daryl           Manhattan              Inwood  40.86482 -
73.92106

   instant_bookable cancellation_policy        room type  ...  service
fee  \
0             False              strict     Private room  ...
$193
1             False            moderate  Entire home/apt  ...
$28
2              True            flexible     Private room  ...
$124
3              True            moderate  Entire home/apt  ...
$74
4             False            moderate  Entire home/apt  ...
$41
5              True            flexible  Entire home/apt  ...
$115
6             False            moderate     Private room  ...
$14
7             False            moderate     Private room  ...
$212
8              True              strict     Private room  ...
$204
9             False              strict     Private room  ...
$58
10            False            flexible  Entire home/apt  ...
$64
11             True            flexible  Entire home/apt  ...
$121
12            False              strict     Private room  ...
$143
13             True            moderate     Private room  ...
$116
14            False            flexible     Private room  ...
$30
15             True            flexible  Entire home/apt  ...
Unknown
16          Unknown            moderate  Entire home/apt  ...
Unknown
```

```
17          Unknown             flexible  Entire home/apt  ...
Unknown
18          Unknown             moderate     Private room  ...
Unknown
19          Unknown             moderate  Entire home/apt  ...
Unknown
20          Unknown             moderate  Entire home/apt  ...
$56
21          Unknown             flexible  Entire home/apt  ...
$95
22          Unknown             moderate     Private room  ...
$27
23          Unknown             moderate     Private room  ...
$210
24          Unknown               strict     Private room  ...
$163
25          Unknown               strict  Entire home/apt  ...
$235
26          Unknown             moderate     Private room  ...
$106
27          Unknown               strict     Private room  ...
$55
28          Unknown             moderate  Entire home/apt  ...
$42
29          Unknown               strict     Private room  ...
$86

   minimum nights number of reviews  last review  reviews per month  \
0            10.0                9.0   10/19/2021           0.210000
1            30.0               45.0    5/21/2022           0.380000
2             3.0                0.0      Unknown           1.374022
3            30.0              270.0     7/5/2019           4.640000
4            10.0                9.0   11/19/2018           0.100000
5             3.0               74.0    6/22/2019           0.590000
6            45.0               49.0    10/5/2017           0.400000
7            45.0               49.0    10/5/2017           0.400000
8             2.0              430.0    6/24/2019           3.470000
9             2.0              118.0    7/21/2017           0.990000
10            1.0              160.0     6/9/2019           1.330000
11            5.0               53.0    6/22/2019           0.430000
12            2.0              188.0    6/23/2019           1.500000
13            4.0              167.0    6/24/2019           1.340000
14            2.0              113.0     7/5/2019           0.910000
15           90.0               27.0   10/31/2018           0.220000
16            2.0              148.0    6/29/2019           1.200000
17            2.0              198.0    6/28/2019           1.720000
18            1.0              260.0     7/1/2019           2.120000
19            3.0               53.0    6/22/2019           4.440000
20            7.0                0.0      Unknown           1.374022
```

```
21              3.0                 9.0   12/28/2021            0.070000
22              2.0               130.0     7/1/2019            1.090000
23              1.0                39.0     1/1/2019            0.370000
24              2.0                71.0     7/2/2019            0.610000
25              2.0                88.0    6/19/2019            0.730000
26              1.0                19.0    6/23/2019            1.370000
27              4.0                 0.0      Unknown            1.374022
28             10.0                58.0    8/13/2017            0.490000
29              3.0               108.0    6/15/2019            1.110000

   review rate number  calculated host listings count  availability
365  \
0              4.000000                                6.0
286.0
1              4.000000                                2.0
228.0
2              5.000000                                1.0
352.0
3              4.000000                                1.0
322.0
4              3.000000                                1.0
289.0
5              3.000000                                1.0
374.0
6              5.000000                                1.0
224.0
7              5.000000                                1.0
219.0
8              3.000000                                1.0
180.0
9              5.000000                                1.0
375.0
10             3.000000                                4.0
1.0
11             4.000000                                1.0
163.0
12             4.000000                                1.0
258.0
13             4.000000                                3.0
47.0
14             3.000000                                1.0
68.0
15             3.000000                                1.0
100.0
16             3.000000                                1.0
197.0
17             5.000000                                1.0
96.0
18             3.000000                                1.0
```

```
325.0
19           5.000000                                1.0
345.0
20           3.000000                                2.0
347.0
21           3.000000                                1.0
193.0
22           4.000000                                6.0
54.0
23           3.000000                                6.0
9.0
24           4.000000                                6.0
344.0
25           4.000000                                2.0
372.0
26           5.000000                                2.0
344.0
27           3.279106                                1.0
96.0
28           3.279106                                1.0
103.0
29           3.279106                                3.0
172.0
```

| | house_rules | license |
|---|---|---|
| 0 | Clean up and treat the home the way you'd like... | Unknown |
| 1 | Pet friendly but please confirm with me if the... | Unknown |
| 2 | I encourage you to use my kitchen, cooking and... | Unknown |
| 3 | Unknown | Unknown |
| 4 | Please no smoking in the house, porch or on th... | Unknown |
| 5 | No smoking, please, and no drugs. | Unknown |
| 6 | Please no shoes in the house so bring slippers... | Unknown |
| 7 | House Guidelines for our BnB We are delighted ... | Unknown |
| 8 | - Please clean up after yourself when using th... | Unknown |
| 9 | NO SMOKING OR PETS ANYWHERE ON THE PROPERTY 1.... | Unknown |
| 10 | Unknown | Unknown |
| 11 | My ideal guests would be warm, friendly, and r... | Unknown |
| 12 | - One of the bedroom closets is not accessible... | Unknown |
| 13 | Unknown | Unknown |
| 14 | Unknown | Unknown |
| 15 | Arrival time can be no later than 9:00PM unles... | Unknown |
| 16 | Absolutely no smoking in the building, handlin... | Unknown |
| 17 | - Please be mindful of the neighbors, quiet ti... | Unknown |
| 18 | Unknown | Unknown |
| 19 | LAUNDRY - Laundry can be done by the visitor b... | Unknown |
| 20 | No smoking, No pets.  No shoes in the house. V... | Unknown |
| 21 | - No smoking or open flames on the property - ... | Unknown |
| 22 | Arrival time can be no later than 10:00PM. No ... | Unknown |
| 23 | Unknown | Unknown |

```
24  We take great care of our home and expect you ...  Unknown
25                                            Unknown  Unknown
26  Quiet neighborhood, middle apartment of big ho...  Unknown
27  To treat our home with respect.  No smoking in...  Unknown
28  Please no pets or smoking in the house, though...  Unknown
29  My ideal guests would be warm, friendly, and r...  Unknown

[30 rows x 22 columns]
```

```
# Verify if missing values have been imputed
missing_values_after_imputation =
df.isnull().sum().sort_values(ascending=True)
print(missing_values_after_imputation)
```

```
NAME                              0
availability 365                  0
calculated host listings count    0
review rate number                0
reviews per month                 0
last review                       0
number of reviews                 0
minimum nights                    0
service fee                       0
price                             0
Construction year                 0
room type                         0
cancellation_policy               0
instant_bookable                  0
long                              0
lat                               0
neighbourhood                     0
neighbourhood group               0
host name                         0
host_identity_verified            0
house_rules                       0
license                           0
dtype: int64
```

```
## Check whether there are any duplicate values in the dataframe
duplicate_rows = df.duplicated()
print("Number of duplicate rows:", duplicate_rows.sum())

## Display the total number of records in the dataframe befores
removing the duplicates
total_records_before_droping = df.shape[0]
print("Total number of records  before removing duplicates:",
total_records_before_droping)

# If present remove duplicate values
df = df.drop_duplicates()
```

```
Number of duplicate rows: 3436
Total number of records  before removing duplicates: 102599

## Display the total number of records in the dataframe after removing
the duplicates
total_records_after_droping = df.shape[0]
print("Total number of records after removing duplicates:",
total_records_after_droping)

# Check for duplicate values
duplicate_rows = df.duplicated()
print("Number of duplicate rows:", duplicate_rows.sum())

Total number of records after removing duplicates: 99163
Number of duplicate rows: 0
```

# Task 3: Data Transformation (Any Tool)

- Rename the column `availability 365` to `days_booked`
- Convert all column names to lowercase and replace the spaces in the column names with an underscore "_".
- Remove the dollar sign and comma from the columns `price` and `service_fee`. If necessary, convert these two columns to the appropriate data type.

If using Python for this exercise, please include the code in the cells below. If using any other tool, please include screenshoots of your work.

```
## Rename the column availability 365 to days_booked
df = df.rename(columns={'availability 365': 'days_booked'})

# Display the updated dataframe
df.head()

                                                  NAME
host_identity_verified  \
0                Clean & quiet apt home by the park
unconfirmed
1                             Skylit Midtown Castle
verified
2               THE VILLAGE OF HARLEM....NEW YORK !
Unknown
3                                           Unknown
unconfirmed
4  Entire Apt: Spacious Studio/Loft by central park
verified

  host name neighbourhood group neighbourhood       lat      long  \
0  Madaline            Brooklyn    Kensington  40.64749 -73.97237
1     Jenna           Manhattan       Midtown  40.75362 -73.98377
2     Elise           Manhattan        Harlem  40.80902 -73.94190
```

```
3     Garry            Brooklyn  Clinton Hill  40.68514 -73.95976
4    Lyndon           Manhattan   East Harlem  40.79851 -73.94399

  instant_bookable cancellation_policy        room type  ...  service
fee  \
0            False              strict     Private room  ...
$193
1            False            moderate  Entire home/apt  ...
$28
2             True            flexible     Private room  ...
$124
3             True            moderate  Entire home/apt  ...
$74
4            False            moderate  Entire home/apt  ...
$41

  minimum nights number of reviews  last review  reviews per month  \
0           10.0                9.0   10/19/2021           0.210000
1           30.0               45.0    5/21/2022           0.380000
2            3.0                0.0      Unknown           1.374022
3           30.0              270.0     7/5/2019           4.640000
4           10.0                9.0   11/19/2018           0.100000

  review rate number  calculated host listings count  days_booked  \
0                4.0                             6.0        286.0
1                4.0                             2.0        228.0
2                5.0                             1.0        352.0
3                4.0                             1.0        322.0
4                3.0                             1.0        289.0

                                          house_rules  license
0  Clean up and treat the home the way you'd like...  Unknown
1  Pet friendly but please confirm with me if the...  Unknown
2  I encourage you to use my kitchen, cooking and...  Unknown
3                                            Unknown  Unknown
4  Please no smoking in the house, porch or on th...  Unknown

[5 rows x 22 columns]
```

```
## Convert all column names to lowercase and replace the spaces with
an underscore "_"
df.columns = df.columns.str.lower().str.replace(' ', '_')

# Display the updated dataframe with modified column names
df.head()
```

```
                                             name
host_identity_verified  \
0                Clean & quiet apt home by the park
unconfirmed
```

```
1                              Skylit Midtown Castle
verified
2               THE VILLAGE OF HARLEM....NEW YORK !
Unknown
3                                           Unknown
unconfirmed
4  Entire Apt: Spacious Studio/Loft by central park
verified

  host_name neighbourhood_group neighbourhood       lat      long  \
0  Madaline            Brooklyn    Kensington  40.64749 -73.97237
1     Jenna           Manhattan       Midtown  40.75362 -73.98377
2     Elise           Manhattan        Harlem  40.80902 -73.94190
3     Garry            Brooklyn  Clinton Hill  40.68514 -73.95976
4    Lyndon           Manhattan   East Harlem  40.79851 -73.94399

  instant_bookable cancellation_policy        room_type  ...
service_fee  \
0            False              strict     Private room  ...
$193
1            False            moderate  Entire home/apt  ...
$28
2             True            flexible     Private room  ...
$124
3             True            moderate  Entire home/apt  ...
$74
4            False            moderate  Entire home/apt  ...
$41

  minimum_nights number_of_reviews  last_review  reviews_per_month  \
0           10.0               9.0   10/19/2021           0.210000
1           30.0              45.0    5/21/2022           0.380000
2            3.0               0.0      Unknown           1.374022
3           30.0             270.0     7/5/2019           4.640000
4           10.0               9.0   11/19/2018           0.100000

  review_rate_number  calculated_host_listings_count  days_booked  \
0                4.0                             6.0        286.0
1                4.0                             2.0        228.0
2                5.0                             1.0        352.0
3                4.0                             1.0        322.0
4                3.0                             1.0        289.0

                                        house_rules  license
0  Clean up and treat the home the way you'd like...  Unknown
1  Pet friendly but please confirm with me if the...  Unknown
2  I encourage you to use my kitchen, cooking and...  Unknown
3                                          Unknown  Unknown
4  Please no smoking in the house, porch or on th...  Unknown
```

```
[5 rows x 22 columns]
```

```
## Remove the dollar sign and comma from the columns. If necessary,
convert these two columns to the appropriate data type.
df['price'] = df['price'].replace({'\$': '', ',': '', 'Unknown':
np.nan}, regex=True).astype(float)
df['service_fee'] = df['service_fee'].replace({'\$': '', ',': '',
'Unknown': np.nan}, regex=True).astype(float)

# Display the updated dataframe
df.head()
```

```
                                                 name
host_identity_verified  \
0                Clean & quiet apt home by the park
unconfirmed
1                             Skylit Midtown Castle
verified
2               THE VILLAGE OF HARLEM....NEW YORK !
Unknown
3                                           Unknown
unconfirmed
4  Entire Apt: Spacious Studio/Loft by central park
verified

  host_name neighbourhood_group neighbourhood       lat      long  \
0  Madaline            Brooklyn    Kensington  40.64749 -73.97237
1     Jenna           Manhattan       Midtown  40.75362 -73.98377
2     Elise           Manhattan        Harlem  40.80902 -73.94190
3     Garry            Brooklyn  Clinton Hill  40.68514 -73.95976
4    Lyndon           Manhattan   East Harlem  40.79851 -73.94399

  instant_bookable cancellation_policy        room_type  ...
service_fee  \
0            False              strict     Private room  ...
193.0
1            False            moderate  Entire home/apt  ...
28.0
2             True            flexible     Private room  ...
124.0
3             True            moderate  Entire home/apt  ...
74.0
4            False            moderate  Entire home/apt  ...
41.0

   minimum_nights  number_of_reviews  last_review
reviews_per_month  \
0            10.0                9.0   10/19/2021           0.210000
```

```
1             30.0                  45.0    5/21/2022            0.380000

2              3.0                   0.0       Unknown           1.374022

3             30.0                 270.0     7/5/2019            4.640000

4             10.0                   9.0   11/19/2018            0.100000


  review_rate_number  calculated_host_listings_count  days_booked  \
0                4.0                             6.0        286.0
1                4.0                             2.0        228.0
2                5.0                             1.0        352.0
3                4.0                             1.0        322.0
4                3.0                             1.0        289.0

                                          house_rules  license
0  Clean up and treat the home the way you'd like...  Unknown
1  Pet friendly but please confirm with me if the...  Unknown
2  I encourage you to use my kitchen, cooking and...  Unknown
3                                            Unknown  Unknown
4  Please no smoking in the house, porch or on th...  Unknown

[5 rows x 22 columns]
```

## Task 4: Exploratory Data Analysis (Any Tool)

- List the count of various room types avaliable in the dataset.
- Which room type has the most strict cancellation policy?
- List the average price per neighborhood group, and highlight the most expensive neighborhood to rent from.

If using Python for this exercise, please include the code in the cells below. If using any other tool, please include screenshoots of your work.

```
## List the count of various room types avaliable with Airbnb

room_type_counts = df['room_type'].value_counts()

# Display the count of room types
print(room_type_counts)
```

```
Entire home/apt    52003
Private room       44895
Shared room         2150
Hotel room           115
Name: room_type, dtype: int64
```

```
## Which room type adheres to more strict cancellation policy
# Group by room type and count the occurrence of each cancellation
policy
```

```
room_type_cancellation_counts = df.groupby(['room_type',
'cancellation_policy']).size()
print(room_type_cancellation_counts)
```

```
room_type        cancellation_policy
Entire home/apt  Unknown                50
                 flexible            17368
                 moderate            17344
                 strict              17241
Hotel room       flexible               44
                 moderate               37
                 strict                 34
Private room     Unknown                23
                 flexible            14834
                 moderate            15101
                 strict              14937
Shared room      Unknown                 3
                 flexible              714
                 moderate              715
                 strict                718
dtype: int64
```

```
# Get the cancellation policy with the highest count for each room
type
most_strict_cancellation =
room_type_cancellation_counts.groupby(level='room_type').idxmax()
print(most_strict_cancellation)
```

```
room_type
Entire home/apt    (Entire home/apt, flexible)
Hotel room              (Hotel room, flexible)
Private room          (Private room, moderate)
Shared room              (Shared room, strict)
dtype: object
```

```
# Filter the room types with the most strict cancellation policy
strict_room_types =
most_strict_cancellation[most_strict_cancellation.apply(lambda x: x[1]
== 'strict')]

# Display the room types with the most strict cancellation policy
print("Room Types with the Most Strict Cancellation Policy:")
print(strict_room_types)
```

```
Room Types with the Most Strict Cancellation Policy:
room_type
Shared room    (Shared room, strict)
dtype: object
```

```
## List the prices by neighborhood group and also mention which is the
most expensive neighborhood group for rentals
```

```
# Calculate the average price per neighborhood group
average_price_neighborhood = df.groupby('neighbourhood_group')
['price'].mean().sort_values(ascending=False)
print("Average Price per Neighborhood Group:")
print(average_price_neighborhood)
```

```
Average Price per Neighborhood Group:
neighbourhood_group
Unknown          658.357143
Queens           629.712735
Bronx            626.614412
Staten Island    626.431843
Brooklyn         626.428192
Manhattan        622.683781
brookln          580.000000
manhatan         460.000000
Name: price, dtype: float64
```

```
# Highlight the most expensive neighborhood
most_expensive_neighborhood = average_price_neighborhood.idxmax()

print("The most expensive neighborhood:")
print(most_expensive_neighborhood)
```

```
The most expensive neighborhood:
Unknown
```

```
# Display the most expensive neighborhood to rent from
print("Most Expensive Neighborhood to Rent from:",
most_expensive_neighborhood)
print(f"{most_expensive_neighborhood}: ${most_expensive_price:.5f}")
```

```
Most Expensive Neighborhood to Rent from: Unknown
Unknown: $658.35714
```

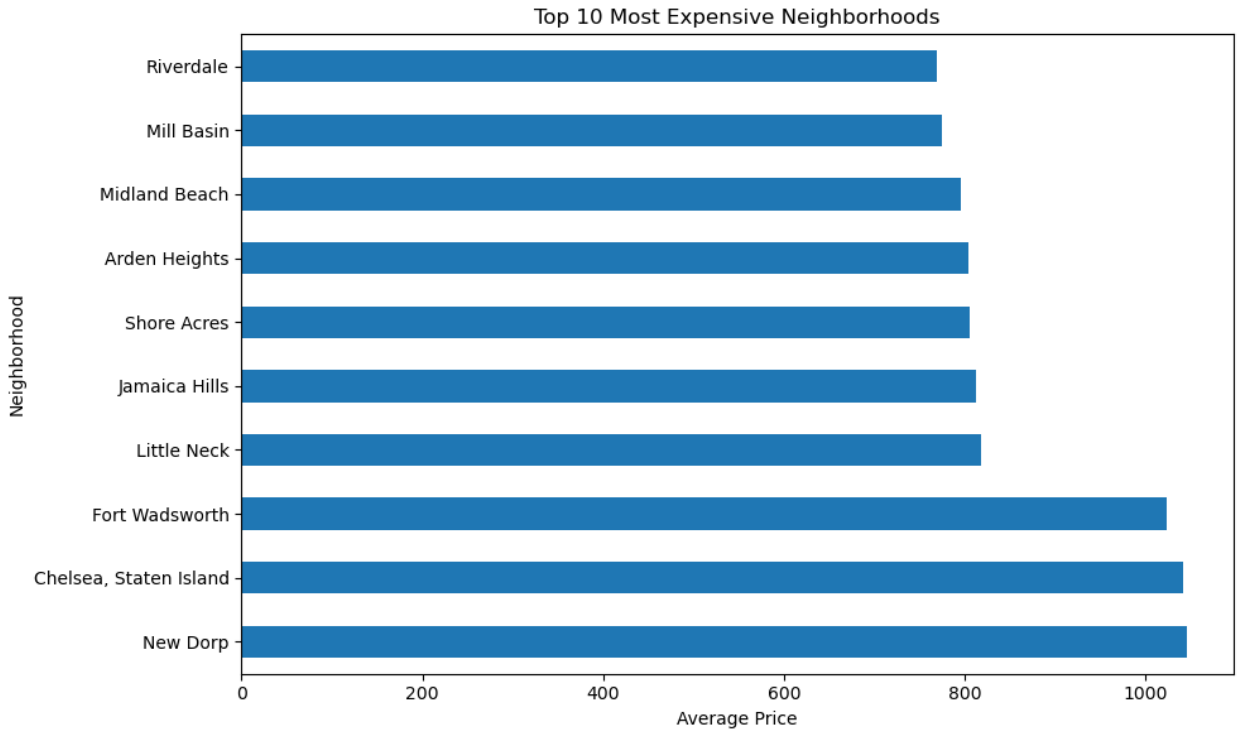
# Task 5a: Data Visualization (Any Tool)

- Create a horizontal bar chart to display the top 10 most expensive neighborhoods in the dataset.
  - Create another chart with the 10 cheapest neighborhoods in the dataset.
- Create a box and whisker chart that showcases the price distribution of all listings split by room type.

If using Python for this exercise, please include the code in the cells below. If using any other tool, please include screenshoots of your work.

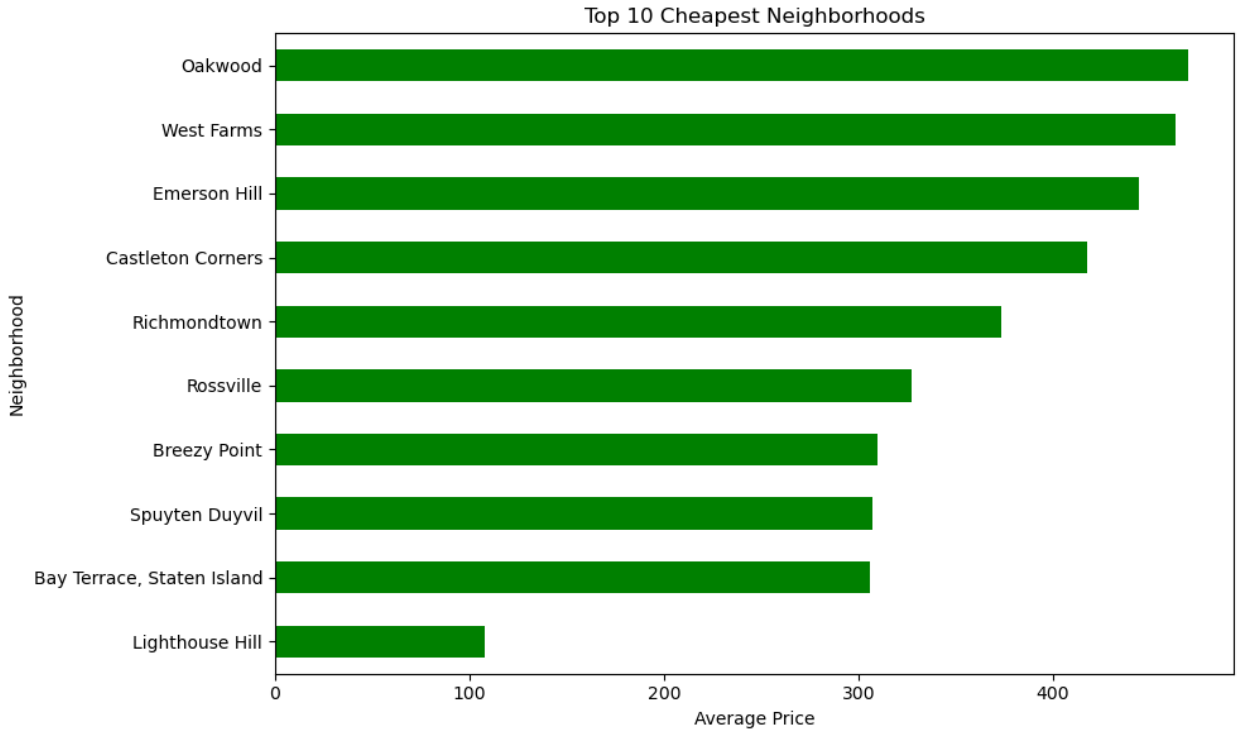
```
#Create a horizontal bar chart to display the top 10 most expensive
neighborhoods in the dataset.
# Convert 'price' column to numeric
df['price'] = pd.to_numeric(df['price'], errors='coerce')
```

```
# Calculate the average price per neighborhood
average_price_neighborhood = df.groupby('neighbourhood')
['price'].mean().sort_values(ascending=False)

# Select the top 10 most expensive neighborhoods
top_10_expensive = average_price_neighborhood.head(10)

# Create the horizontal bar chart
plt.figure(figsize=(10, 6))
top_10_expensive.plot(kind='barh')
plt.xlabel('Average Price')
plt.ylabel('Neighborhood')
plt.title('Top 10 Most Expensive Neighborhoods')
plt.tight_layout()
plt.show()
```



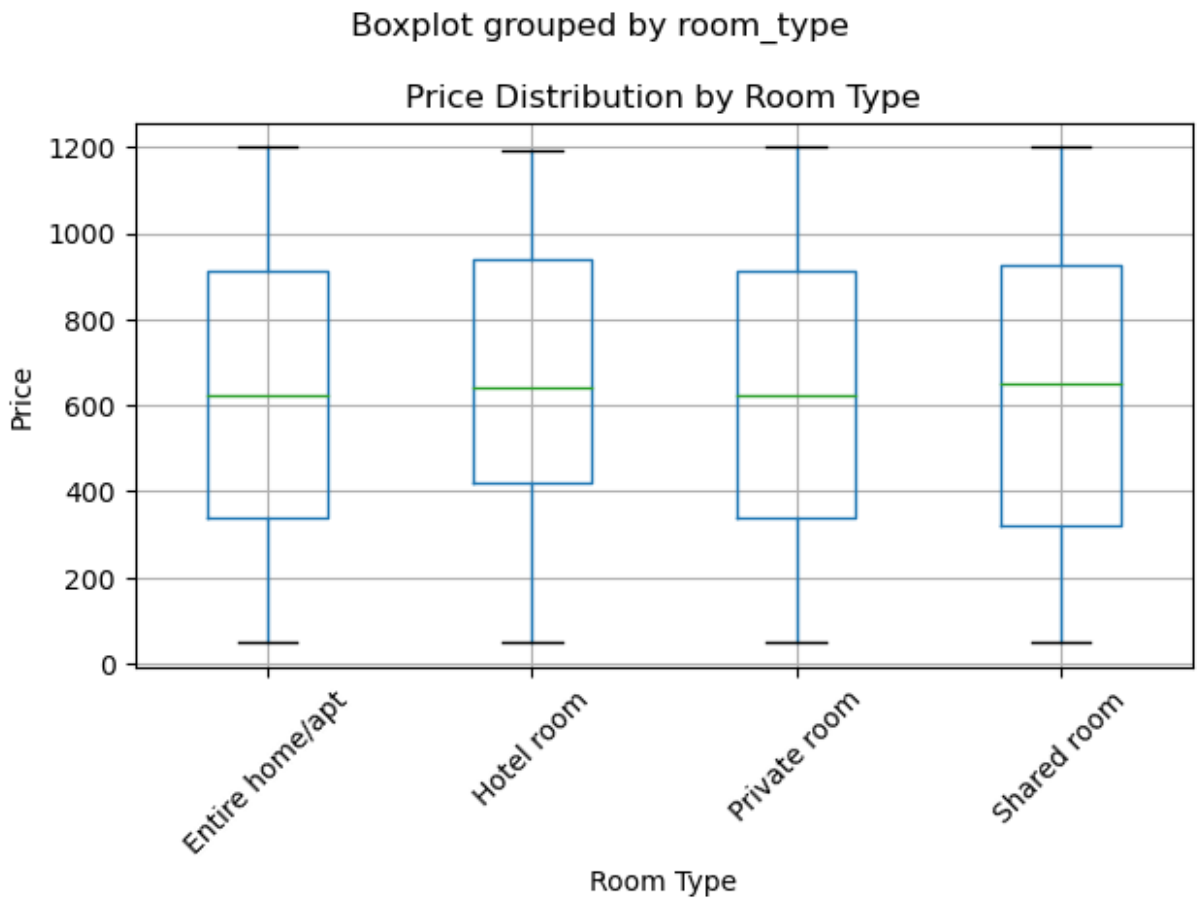

```
#Create another chart with the 10 cheapest neighborhoods in the
dataset.
# Convert 'price' column to numeric
df['price'] = pd.to_numeric(df['price'], errors='coerce')

# Calculate the average price per neighborhood
average_price_neighborhood = df.groupby('neighbourhood')
['price'].mean().sort_values()

# Select the 10 cheapest neighborhoods
```

```
top_10_cheapest = average_price_neighborhood.head(10)

# Create the horizontal bar chart
plt.figure(figsize=(10, 6))
top_10_cheapest.plot(kind='barh', color='green')
plt.xlabel('Average Price')
plt.ylabel('Neighborhood')
plt.title('Top 10 Cheapest Neighborhoods')
plt.tight_layout()
plt.show()
```




```
#Create a box and whisker chart that showcases the price distribution
of all listings split by room type.
# Convert 'price' column to numeric
df['price'] = pd.to_numeric(df['price'], errors='coerce')

# Create the box and whisker chart
plt.figure(figsize=(10, 6))
df.boxplot(column='price', by='room_type')
plt.xlabel('Room Type')
plt.ylabel('Price')
plt.title('Price Distribution by Room Type')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

```
<Figure size 1000x600 with 0 Axes>
```



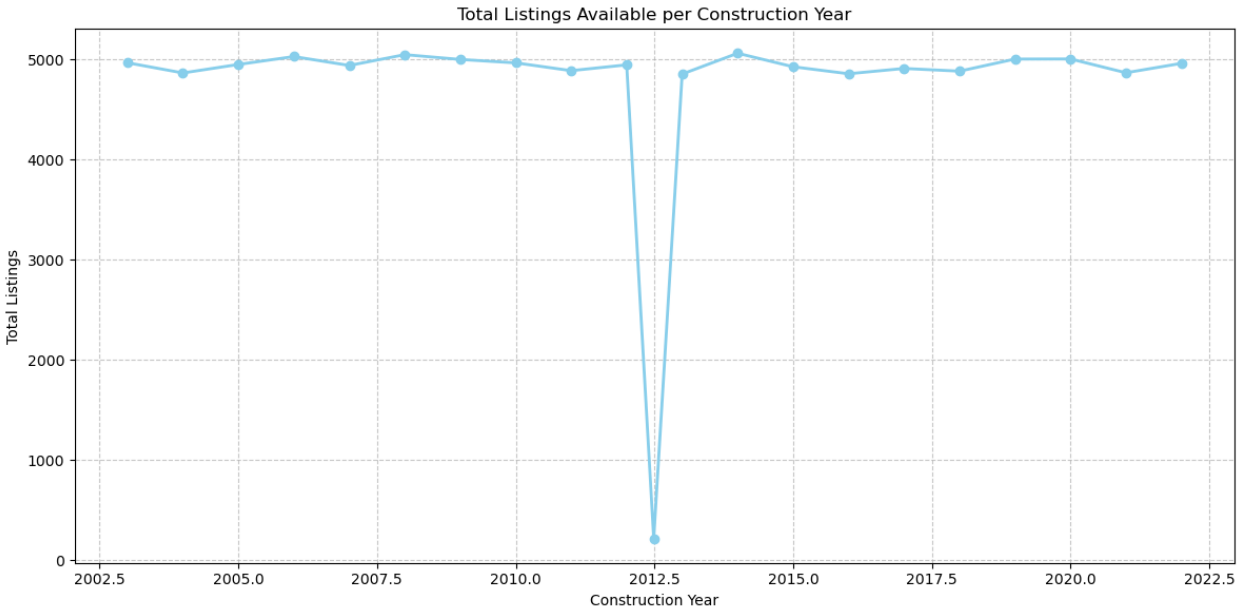

# Task 5b: Data Visualization (Any Tool)

- Create a scatter plot to illustrate the relationshi between the cleaning fee and the room price and write down the kind of correlation, if any, that you see.
- Create a line chart to showcase the total amount of listings available per year.

If using Python for this exercise, please include the code in the cells below. If using any other tool, please include screenshoots of your work.

```
# Create a scatter plot to illustrate the relationshi between the
cleaning fee and the room price
plt.figure(figsize=(11, 6))
sns.scatterplot(data=df, x='service_fee', y='price')
plt.xlabel('Cleaning Fee')
plt.ylabel('Room Price')
plt.title('Relationship between Cleaning Fee and Room Price')
plt.show()
```


```
# Group the data by construction_year and calculate the total count of
listings for each year
listings_per_year = df.groupby('construction_year').size()
##Create a line chart to showcase the total amount of listings
available per year
plt.figure(figsize=(12, 6))
listings_per_year.plot(kind='line', marker='o', color='skyblue',
linewidth=2)
plt.xlabel('Construction Year')
plt.ylabel('Total Listings')
plt.title('Total Listings Available per Construction Year')
plt.grid(True, linestyle='--', alpha=0.7)
plt.tight_layout()

# Display the plot
plt.show()
```




# Task 5c: Data Visualization (Any Tool)

- Create a data visualization of your choosing using one of the review columns in isolation or in combination with another column.
- Create a visualization to compare at least two different variables between super hosts and regular hosts.

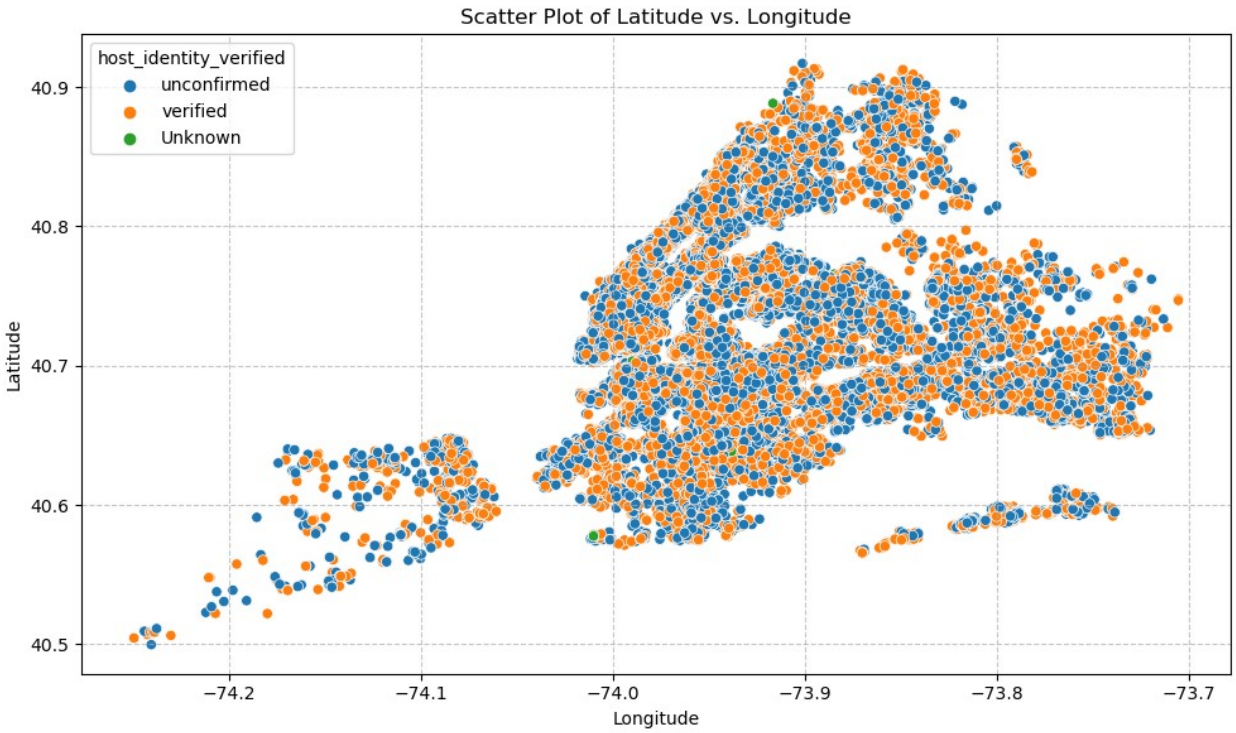
If using Python for this exercise, please include the code in the cells below. If using any other tool, please include screenshoots of your work.

```
#Create a data visualization of your choosing using one of the review columns
plt.figure(figsize=(10, 6))
# Adjust column names here
sns.scatterplot(data=df, x='reviews_per_month', y='price', alpha=0.7)
plt.xlabel('Review Scores Rating')
plt.ylabel('Price ($)')
plt.title('Relationship between Review Scores Rating and Price')
plt.grid(True, linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()  # Display the plot
```


Create a visualization to compare at least two different variables between super hosts and regular hosts >>> change to latitude and longitude for both "unconfirmed" and "verified" because there is no dataset super hosts and regular hosts .

```
# Compute the mean latitude and longitude for both "unconfirmed" and
"verified"
#host_identity_verified categories
mean_lat = df.groupby("host_identity_verified").lat.mean().round(2)
mean_long = df.groupby("host_identity_verified").long.mean().round(2)

print("Mean Latitude:")
print(mean_lat)

print("\nMean Longitude:")
print(mean_long)

# Create the scatter plot
plt.figure(figsize=(10, 6))
sns.scatterplot(x='long', y='lat', hue='host_identity_verified',
data=df)
plt.xlabel('Longitude')
plt.ylabel('Latitude')
plt.title('Scatter Plot of Latitude vs. Longitude')
plt.grid(True, linestyle='--', alpha=0.7)
plt.tight_layout()
```

```
# Display the plot
plt.show()

Mean Latitude:
host_identity_verified
Unknown        40.73
unconfirmed    40.73
verified       40.73
Name: lat, dtype: float64

Mean Longitude:
host_identity_verified
Unknown       -73.96
unconfirmed   -73.95
verified      -73.95
Name: long, dtype: float64
```




---

END----------