T5 project - Week3

Course End Project : Data Analysis Module

Maram Alsalamah
Lamia alasmari

```
[ ] # import liberaries
    import numpy as np
    import pandas as pd
    import matplotlib.pyplot as plt
    import seaborn as sns
```

```
[ ] # ploting style
    plt.style.use('ggplot')
```

Data Loading and Initial Exploration :

```
[ ] # data paths for loading
    Folder_path = r'C:\Users\maram\Desktop\Airbnb Data'
    Listings_file = Folder_path+r'\Listings.csv'
    Reviews_file = Folder_path+r'\Reviews.csv'
```

Importing the 2 CSV files wiht ANSI format :

```
listining_data = pd.read_csv(Listings_file,encoding='ANSI')
Reviews_data = pd.read_csv(Reviews_file,encoding='ANSI')
```

Explore the structure of the dateset (number of rows and columns, data types, etc.):

```
print("Dataset shape:", listining_data.shape)
print("\nData Columns types:")
print(listining_data.dtypes)
```

```
Dataset shape: (279712, 33)

Data Columns types:
listing_id                        int64
name                             object
host_id                           int64
host_since                       object
host_location                    object
host_response_time               object
host_response_rate              float64
host_acceptance_rate            float64
host_is_superhost                object
host_total_listings_count       float64
host_has_profile_pic             object
host_identity_verified           object
neighbourhood                    object
district                         object
city                             object
latitude                        float64
longitude                       float64
property_type                    object
room_type                        object
accommodates                      int64
bedrooms                        float64
amenities                        object
price                             int64
minimum_nights                    int64
maximum_nights                    int64
review_scores_rating            float64
review_scores_accuracy          float64
review_scores_cleanliness       float64
review_scores_checkin           float64
review_scores_communication     float64
review_scores_location          float64
review_scores_value             float64
instant_bookable                 object
dtype: object
```

```python
print("Dataset shape:", Reviews_data.shape)
print("\nData Columns types:")
print(Reviews_data.dtypes)
```

```
Dataset shape: (5373143, 4)

Data Columns types:
listing_id      int64
review_id       int64
date           object
reviewer_id     int64
dtype: object
```

Examine the first few rows of the dataset to understand its contents:

```python
pd.set_option('display.max_columns', None)
listining_data.head()
```

	listing_id	name	host_id	host_since	host_location	host_response_time	host_response_rate	host_acceptance_rate	host_is_superhost	host_total_listings_count	host_has_profile_pic	host_identity_verified	neighbourhood	district	city	latitude	longitude	property_type	room_type	accommodates	bedrooms	amenities	pri
0	281420	Beautiful Flat in le Village Montmartre, Paris	1466919	2011-12-03	Paris, Ile-de-France, France	NaN	NaN	NaN	f	1.0	t	f	Buttes-Montmartre	NaN	Paris	48.88668	2.33343	Entire apartment	Entire place	2	1.0	["Heating", "Kitchen", "Washer", "Wifi", "Long...	
1	3705183	39 mÃ‚Â² Paris (Sacre CÃ... â€œur)	10328771	2013-11-29	Paris, Ile-de-France, France	NaN	NaN	NaN	f	1.0	t	t	Buttes-Montmartre	NaN	Paris	48.88617	2.34515	Entire apartment	Entire place	2	1.0	["Shampoo", "Heating", "Kitchen", "Essentials"...	1
2	4082273	Lovely apartment with Terrace, 60m2	19252768	2014-07-31	Paris, Ile-de-France, France	NaN	NaN	NaN	f	1.0	t	f	Elysee	NaN	Paris	48.88112	2.31712	Entire apartment	Entire place	2	1.0	["Heating", "TV", "Kitchen", "Washer", "Wifi"...	
3	4797344	Cosy studio (close to Eiffel tower)	10668311	2013-12-17	Paris, Ile-de-France, France	NaN	NaN	NaN	f	1.0	t	t	Vaugirard	NaN	Paris	48.84571	2.30584	Entire apartment	Entire place	2	1.0	["Heating", "TV", "Kitchen", "Wifi", "Long ter...	
4	4823489	Close to Eiffel Tower - Beautiful flat : 2 rooms	24837558	2014-12-14	Paris, Ile-de-France, France	NaN	NaN	NaN	f	1.0	t	f	Passy	NaN	Paris	48.85500	2.26979	Entire apartment	Entire place	2	1.0	["Heating", "TV", "Kitchen", "Essentials", "Ha...	

```
Reviews_data.head()
```

	listing_id	review_id	date	reviewer_id
0	11798	330265172	2018-09-30	11863072
1	15383	330103585	2018-09-30	39147453
2	16455	329985788	2018-09-30	1125378
3	17919	330016899	2018-09-30	172717984
4	26827	329995638	2018-09-30	17542859

Data Cleaning :

Handle missing values appropriately (e.g., imputation, deletion, etc.).

```
# first we need to see the nan values of out data to know how can we deal with it
listining_data.isna().sum()
```

```
listing_id                          0
name                              173
host_id                             0
host_since                        165
host_location                     840
host_response_time             128782
host_response_rate             128782
host_acceptance_rate           113087
host_is_superhost                 165
host_total_listings_count         165
host_has_profile_pic              165
host_identity_verified            165
neighbourhood                       0
district                       242700
city                                0
latitude                            0
longitude                           0
property_type                       0
room_type                           0
accommodates                        0
bedrooms                        29435
amenities                           0
price                               0
minimum_nights                      0
maximum_nights                      0
review_scores_rating            91405
review_scores_accuracy          91713
review_scores_cleanliness       91665
review_scores_checkin           91771
review_scores_communication     91687
review_scores_location          91775
review_scores_value             91785
instant_bookable                    0
dtype: int64
```

```
# first we need to see the nan values of out data to know how can we deal with it
Reviews_data.isna().sum()
```

```
listing_id     0
review_id      0
date           0
reviewer_id    0
dtype: int64
```

```
## more information about our data
listining_data.describe()
```

	listing_id	host_id	host_response_rate	host_acceptance_rate	host_total_listings_count	latitude	longitude	accommodates	bedrooms	price	minimum_nights	maximum_nights	review_scores_rating	review_scores_accuracy
count	2.797120e+05	2.797120e+05	150930.000000	166625.000000	279547.000000	279712.000000	279712.000000	279712.000000	250277.000000	279712.000000	279712.000000	2.797120e+05	188307.000000	187999.000000
mean	2.638196e+07	1.081658e+08	0.865939	0.827168	24.581612	18.761862	12.595075	3.288736	1.515509	608.792737	8.050967	2.755860e+04	93.405195	9.565476
std	1.442576e+07	1.108570e+08	0.283744	0.289202	284.041143	32.560343	73.081309	2.133379	1.153080	3441.826611	31.518946	7.282875e+06	10.070437	0.990878
min	2.577000e+03	1.822000e+03	0.000000	0.000000	0.000000	-34.264400	-99.339630	0.000000	1.000000	0.000000	1.000000	1.000000e+00	20.000000	2.000000
25%	1.384462e+07	1.720656e+07	0.900000	0.780000	1.000000	-22.964390	-43.198040	2.000000	1.000000	75.000000	1.000000	4.500000e+01	91.000000	9.000000
50%	2.767098e+07	5.826911e+07	1.000000	0.980000	1.000000	40.710785	2.382780	2.000000	1.000000	150.000000	2.000000	1.125000e+03	96.000000	10.000000
75%	3.978485e+07	1.832853e+08	1.000000	1.000000	4.000000	41.908610	28.986730	4.000000	2.000000	474.000000	5.000000	1.125000e+03	100.000000	10.000000
max	4.834353e+07	3.901874e+08	1.000000	1.000000	7235.000000	48.904910	151.339810	16.000000	50.000000	625216.000000	9999.000000	2.147484e+09	100.000000	10.000000

```python
listining_data.columns
```

```
Index(['listing_id', 'name', 'host_id', 'host_since', 'host_location',
       'host_response_time', 'host_response_rate', 'host_acceptance_rate',
       'host_is_superhost', 'host_total_listings_count',
       'host_has_profile_pic', 'host_identity_verified', 'neighbourhood',
       'district', 'city', 'latitude', 'longitude', 'property_type',
       'room_type', 'accommodates', 'bedrooms', 'amenities', 'price',
       'minimum_nights', 'maximum_nights', 'review_scores_rating',
       'review_scores_accuracy', 'review_scores_cleanliness',
       'review_scores_checkin', 'review_scores_communication',
       'review_scores_location', 'review_scores_value', 'instant_bookable'],
      dtype='object')
```

```python
## see the values count of each column
print ('Cities Values \n' , listining_data['city'].value_counts())
```

```
Cities Values
 Paris              64690
New York           37012
Sydney             33630
Rome               27647
Rio de Janeiro     26615
Istanbul           24519
Mexico City        20065
Bangkok            19361
Cape Town          19086
Hong Kong           7087
Name: city, dtype: int64
```

```python
# see the values count of each column
print ('room_type Values \n' , listining_data['room_type'].value_counts())
```

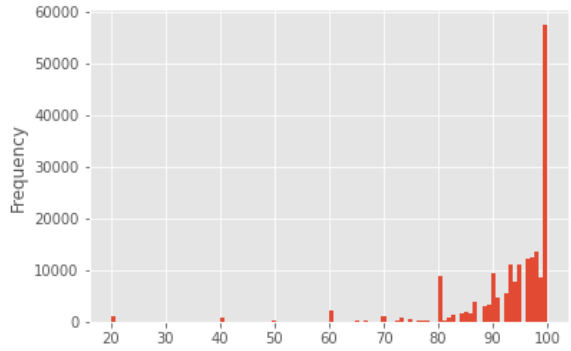
```
room_type Values
 Entire place    182005
Private room     86988
Hotel room        5857
Shared room       4862
Name: room_type, dtype: int64
```

```python
## see the values count of each column
print ('review_scores_rating Values \n' , listining_data['review_scores_rating'].value_counts())
```

```
review_scores_rating Values
 100.0    57458
98.0     13616
97.0     12425
96.0     12261
93.0     10995
         ...
31.0         1
36.0         1
27.0         1
61.0         1
44.0         1
Name: review_scores_rating, Length: 63, dtype: int64
```

```
[ ] ## plot of review_scores_rating
    listining_data['review_scores_rating'].plot(kind ='hist',bins=100)
```

```
<Axes: ylabel='Frequency'>
```




```
[ ] # Replace missing values in specific columns
    listining_data['name'].fillna("Unknown", inplace=True)
    listining_data['host_since'].fillna(listining_data['host_since'].mode()[0], inplace=True)
    listining_data['host_location'].fillna("Unknown", inplace=True)
    # Continue with similar operations for other columns...

    # For numeric columns : i can replaced missing values with median or mode
    numeric_cols = ['host_total_listings_count', 'bedrooms', 'review_scores_rating', 'review_scores_accuracy',
                    'review_scores_cleanliness', 'review_scores_checkin', 'review_scores_communication', 'review_scores_location', 'review_scores_value']
    for col in numeric_cols:
        listining_data[col].fillna(listining_data[col].median(), inplace=True)
```

```
# nan values of out data ( how can we deal with it ?)
listining_data.isna().sum()
```

```
listing_id                        0
name                              0
host_id                           0
host_since                        0
host_location                     0
host_response_time           128782
host_response_rate           128782
host_acceptance_rate         113087
host_is_superhost               165
host_total_listings_count         0
host_has_profile_pic            165
host_identity_verified          165
neighbourhood                     0
district                     242700
city                              0
latitude                          0
longitude                         0
property_type                     0
room_type                         0
accommodates                      0
bedrooms                          0
amenities                         0
price                             0
minimum_nights                    0
maximum_nights                    0
review_scores_rating              0
review_scores_accuracy            0
review_scores_cleanliness         0
review_scores_checkin             0
review_scores_communication       0
review_scores_location            0
review_scores_value               0
instant_bookable                  0
dtype: int64
```

```
[ ] # missing values for remaining columns
    listining_data['host_response_time'].fillna("Unknown", inplace=True)
    listining_data['host_response_rate'].fillna(listining_data['host_response_rate'].median(), inplace=True)
    listining_data['host_acceptance_rate'].fillna(listining_data['host_acceptance_rate'].median(), inplace=True)
    listining_data['district'].fillna("Unknown", inplace=True)
```

```
# nan values
listining_data.isna().sum()
```

```
listing_id                     0
name                           0
host_id                        0
host_since                     0
host_location                  0
host_response_time             0
host_response_rate             0
host_acceptance_rate           0
host_is_superhost              0
host_total_listings_count      0
host_has_profile_pic           0
host_identity_verified         0
neighbourhood                  0
district                       0
city                           0
latitude                       0
longitude                      0
property_type                  0
room_type                      0
accommodates                   0
bedrooms                       0
amenities                      0
price                          0
minimum_nights                 0
maximum_nights                 0
review_scores_rating           0
review_scores_accuracy         0
review_scores_cleanliness      0
review_scores_checkin          0
review_scores_communication    0
review_scores_location         0
review_scores_value            0
instant_bookable               0
dtype: int64
```

Check for any duplicate entries:

```
[ ] # Check for duplicate entries
    duplicate_rows = listining_data.duplicated().sum()
    print("\nNumber of duplicate rows are:", duplicate_rows)
```

```
Number of duplicate rows are: 0
```

Convert categorical variables into the appropriate data type if necessary :

```
[ ] # categorical columns
    categorical_cols = listining_data.select_dtypes(include=['object']).columns

    print("Categorical Columns:")
    print(categorical_cols)
```

```
Categorical Columns:
Index(['name', 'host_since', 'host_location', 'host_response_time',
       'host_is_superhost', 'host_has_profile_pic', 'host_identity_verified',
       'neighbourhood', 'district', 'city', 'property_type', 'room_type',
       'amenities', 'instant_bookable'],
      dtype='object')
```

```
listining_data[categorical_cols].head()
```

```
listining_data[categorical_cols].head()
```

	name	host_since	host_location	host_response_time	host_is_superhost	host_has_profile_pic	host_identity_verified	neighbourhood	district	city	property_type	room_type	amenities	instant_bookable
0	Beautiful Flat in le Village Montmartre, Paris	2011-12-03	Paris, Ile-de-France, France	NaN	f	t	f	Buttes-Montmartre	NaN	Paris	Entire apartment	Entire place	["Heating", "Kitchen", "Washer", "Wifi", "Long...	f
1	39 mÃ‚Â² Paris (Sacre CÃ…â€œeur)	2013-11-29	Paris, Ile-de-France, France	NaN	f	t	t	Buttes-Montmartre	NaN	Paris	Entire apartment	Entire place	["Shampoo", "Heating", "Kitchen", "Essentials"...	f
2	Lovely apartment with Terrace, 60m2	2014-07-31	Paris, Ile-de-France, France	NaN	f	t	f	Elysee	NaN	Paris	Entire apartment	Entire place	["Heating", "TV", "Kitchen", "Washer", "Wifi",...	f
3	Cosy studio (close to Eiffel tower)	2013-12-17	Paris, Ile-de-France, France	NaN	f	t	t	Vaugirard	NaN	Paris	Entire apartment	Entire place	["Heating", "TV", "Kitchen", "Wifi", "Long ter...	f
4	Close to Eiffel Tower - Beautiful flat : 2 rooms	2014-12-14	Paris, Ile-de-France, France	NaN	f	t	f	Passy	NaN	Paris	Entire apartment	Entire place	["Heating", "TV", "Kitchen", "Essentials", "Ha...	f

```
[ ] ## see the values count of each column
    print ('host_is_superhost Values \n' , listining_data['host_is_superhost'].value_counts())
```

```
host_is_superhost Values
 f    229294
t     50253
Name: host_is_superhost, dtype: int64
```

```
[ ] # Alternatively, using replace
    # host_is_superhost
    # host_has_profile_pic
    # host_identity_verified
    # instant_bookable
    convert_to_bool = ('host_is_superhost','host_has_profile_pic','host_identity_verified','instant_bookable')

    for Col in convert_to_bool :
        listining_data[Col] = listining_data[Col].replace({'f': False, 't': True})
        listining_data[Col] = listining_data[Col].astype(bool)
```

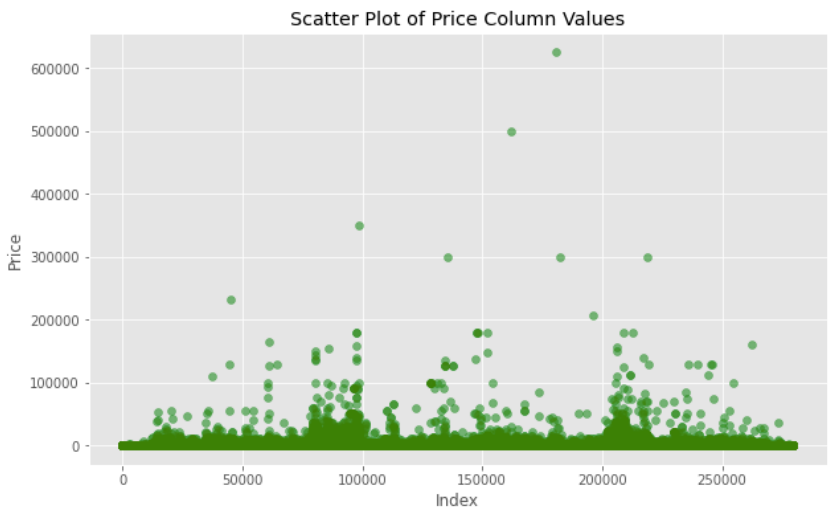
```
listining_data[categorical_cols].head()
```

	name	host_since	host_location	host_response_time	host_is_superhost	host_has_profile_pic	host_identity_verified	neighbourhood	district	city	property_type	room_type	amenities	instant_bookable
0	Beautiful Flat in le Village Montmartre, Paris	2011-12-03	Paris, Ile-de-France, France	NaN	False	True	False	Buttes-Montmartre	NaN	Paris	Entire apartment	Entire place	["Heating", "Kitchen", "Washer", "Wifi", "Long...	False
1	39 mÃ‚Â² Paris (Sacre CÃ…â€œur)	2013-11-29	Paris, Ile-de-France, France	NaN	False	True	True	Buttes-Montmartre	NaN	Paris	Entire apartment	Entire place	["Shampoo", "Heating", "Kitchen", "Essentials"...	False
2	Lovely apartment with Terrace, 60m2	2014-07-31	Paris, Ile-de-France, France	NaN	False	True	False	Elysee	NaN	Paris	Entire apartment	Entire place	["Heating", "TV", "Kitchen", "Washer", "Wifi",...	False
3	Cosy studio (close to Eiffel tower)	2013-12-17	Paris, Ile-de-France, France	NaN	False	True	True	Vaugirard	NaN	Paris	Entire apartment	Entire place	["Heating", "TV", "Kitchen", "Wifi", "Long ter...	False
4	Close to Eiffel Tower - Beautiful flat : 2 rooms	2014-12-14	Paris, Ile-de-France, France	NaN	False	True	False	Passy	NaN	Paris	Entire apartment	Entire place	["Heating", "TV", "Kitchen", "Essentials", "Ha...	False

Exploratory Data Analysis :

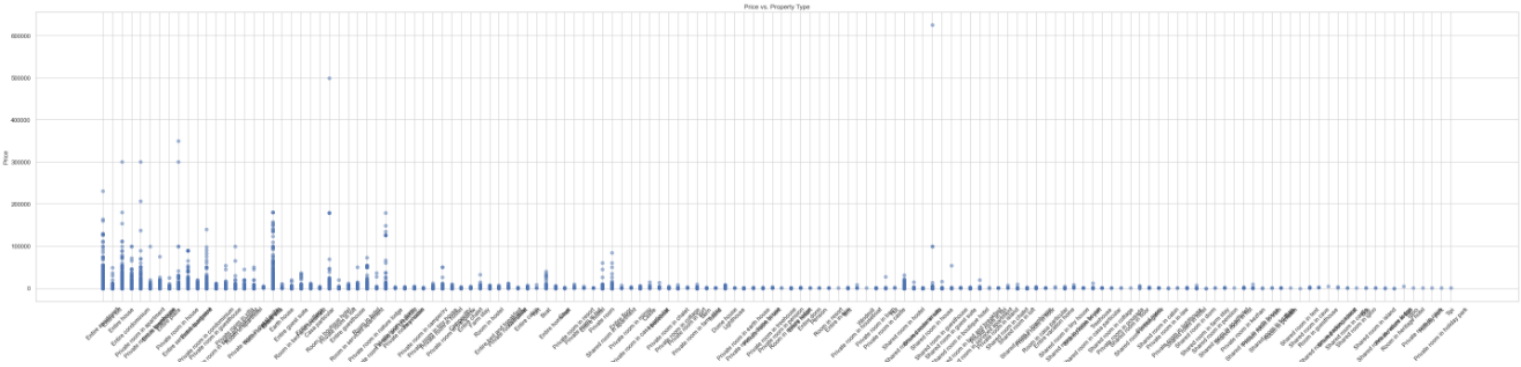
```python
plt.figure(figsize=(10, 6))
sns.histplot(listining_data['price'], bins=30, kde=True, color='skyblue', stat='density')
plt.title('Distribution of Listing Prices')
plt.xlabel('Price')
plt.ylabel('Density')
plt.xlim(listining_data['price'].min(), listining_data['price'].max())  # Set x-axis range
plt.show()
```


```python
# Scatter plot of 'price' column values
plt.figure(figsize=(10, 6))
plt.scatter(range(len(listining_data['price'])), listining_data['price'], color='green', alpha=0.5)
plt.title('Scatter Plot of Price Column Values')
plt.xlabel('Index')
plt.ylabel('Price')
plt.show()
```



```
[ ] plt.figure(figsize=(50, 10))
    plt.scatter(listining_data['property_type'], listining_data['price'], alpha=0.5)
    plt.title('Price vs. Property Type')
    plt.xlabel('Property Type')
    plt.ylabel('Price')
    plt.xticks(rotation=45)
    plt.show()

    # Scatter plot: Price vs. Neighborhood
    plt.figure(figsize=(50, 6))
    plt.scatter(listining_data['neighbourhood'], listining_data['price'], alpha=0.5)
    plt.title('Price vs. Neighborhood')
    plt.xlabel('Neighborhood')
    plt.ylabel('Price')
    plt.xticks(rotation=45)
    plt.show()
```




```
summary_statistics = listining_data[['price', 'review_scores_rating', 'review_scores_accuracy', 'review_scores_cleanliness',
                                     'review_scores_checkin', 'review_scores_communication', 'review_scores_location',
                                     'review_scores_value']].describe()

print(summary_statistics)
```

```
               price  review_scores_rating  review_scores_accuracy  \
count  279712.000000         279712.000000           279712.000000
mean      608.792737             94.253132                9.707950
std      3441.826611              8.351922                0.837567
min         0.000000             20.000000                2.000000
25%        75.000000             94.000000               10.000000
50%       150.000000             96.000000               10.000000
75%       474.000000             98.000000               10.000000
max    625216.000000            100.000000               10.000000

       review_scores_cleanliness  review_scores_checkin  \
count              279712.000000          279712.000000
mean                   9.538050               9.799458
std                    0.993508               0.724713
min                    2.000000               2.000000
25%                    9.000000              10.000000
50%                   10.000000              10.000000
75%                   10.000000              10.000000
max                   10.000000              10.000000

       review_scores_communication  review_scores_location  \
count                 279712.000000           279712.000000
mean                      9.797392                9.754083
std                       0.740778                0.704282
min                       2.000000                2.000000
25%                      10.000000               10.000000
50%                      10.000000               10.000000
75%                      10.000000               10.000000
max                      10.000000               10.000000

       review_scores_value
count        279712.000000
mean              9.553459
std               0.909803
min               2.000000
25%               9.000000
50%              10.000000
75%              10.000000
max              10.000000
```

Feature Engineering :

```
[ ] listining_data['avg_rating'] = listining_data[['review_scores_rating', 'review_scores_accuracy', 'review_scores_cleanliness',
                                                   'review_scores_checkin', 'review_scores_communication', 'review_scores_location', 'review_scores_value']].mean(axis=1)
```

```
[ ] merged_data = pd.merge(listining_data, Reviews_data, on='listing_id', how='inner')
```

```
[ ] merged_data['occupancy_rate'] =  merged_data.groupby('listing_id')['review_id'].transform('count') / merged_data['maximum_nights']
```

```
[ ] merged_data.head()
```

	listing_id	name	host_id	host_since	host_location	host_response_time	host_response_rate	host_acceptance_rate	host_is_superhost	host_total_listings_count	host_has_profile_pic	host_identity_verified	neighbourhood	district	city	latitude	longitude	property_type	room_type	accommodates	bedrooms	amenities
0	281420	Beautiful Flat in le Village Montmartre, Paris	1466919	2011-12-03	Paris, Ile-de-France, France	Unknown	1.0	0.98	False	1.0	True	False	Buttes-Montmartre	Unknown	Paris	48.88668	2.33343	Entire apartment	Entire place	2	1.0	["Heating", "Kitchen", "Washer", "Wifi", "Long...
1	281420	Beautiful Flat in le Village Montmartre, Paris	1466919	2011-12-03	Paris, Ile-de-France, France	Unknown	1.0	0.98	False	1.0	True	False	Buttes-Montmartre	Unknown	Paris	48.88668	2.33343	Entire apartment	Entire place	2	1.0	["Heating", "Kitchen", "Washer", "Wifi", "Long...
2	3705183	39 mÃ‚Â² Paris (Sacre CÃ... â€œur)	10328771	2013-11-29	Paris, Ile-de-France, France	Unknown	1.0	0.98	False	1.0	True	True	Buttes-Montmartre	Unknown	Paris	48.88617	2.34515	Entire apartment	Entire place	2	1.0	["Shampoo", "Heating", "Kitchen", "Essentials"...
3	3705183	39 mÃ‚Â² Paris (Sacre CÃ... â€œur)	10328771	2013-11-29	Paris, Ile-de-France, France	Unknown	1.0	0.98	False	1.0	True	True	Buttes-Montmartre	Unknown	Paris	48.88617	2.34515	Entire apartment	Entire place	2	1.0	["Shampoo", "Heating", "Kitchen", "Essentials"...
4	3705183	39 mÃ‚Â² Paris (Sacre CÃ... â€œur)	10328771	2013-11-29	Paris, Ile-de-France, France	Unknown	1.0	0.98	False	1.0	True	True	Buttes-Montmartre	Unknown	Paris	48.88617	2.34515	Entire apartment	Entire place	2	1.0	["Shampoo", "Heating", "Kitchen", "Essentials"...