In [ ]:
```
import pandas as pd
csv_file = './AB_NYC_2019.csv'
airbnb = pd.read_csv(csv_file)
df = pd.DataFrame(airbnb)
```

In [ ]:
```
df.shape
```

Out[ ]:
```
(48895, 16)
```

In [ ]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48895 entries, 0 to 48894
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   id                              48895 non-null  int64  
 1   name                            48879 non-null  object 
 2   host_id                         48895 non-null  int64  
 3   host_name                       48874 non-null  object 
 4   neighbourhood_group             48895 non-null  object 
 5   neighbourhood                   48895 non-null  object 
 6   latitude                        48895 non-null  float64
 7   longitude                       48895 non-null  float64
 8   room_type                       48895 non-null  object 
 9   price                           48895 non-null  int64  
 10  minimum_nights                  48895 non-null  int64  
 11  number_of_reviews               48895 non-null  int64  
 12  last_review                     38843 non-null  object 
 13  reviews_per_month               38843 non-null  float64
 14  calculated_host_listings_count  48895 non-null  int64  
 15  availability_365                48895 non-null  int64  
dtypes: float64(3), int64(7), object(6)
memory usage: 6.0+ MB
```

In [ ]:
```
df.head()
```

Out[ ]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851

In [ ]:
```
df.isnull().sum()
```

Out[ ]:
```
id                                0
name                             16
host_id                           0
host_name                        21
neighbourhood_group               0
neighbourhood                     0
latitude                          0
longitude                         0
room_type                         0
price                             0
minimum_nights                    0
number_of_reviews                 0
last_review                   10052
reviews_per_month             10052
calculated_host_listings_count        0
availability_365                  0
dtype: int64
```

In [ ]:
```
colsToDrop = ['id','host_name','last_review']

df.drop(colsToDrop, axis=1, inplace=True)
df.isnull().sum()
```

Out[ ]:
```
name                             16
host_id                           0
neighbourhood_group               0
neighbourhood                     0
latitude                          0
longitude                         0
room_type                         0
price                             0
minimum_nights                    0
number_of_reviews                 0
reviews_per_month             10052
calculated_host_listings_count        0
availability_365                  0
dtype: int64
```

In [ ]:
```
df.dropna(inplace=True)
```

In [ ]:
```
df.shape
```

Out[ ]:
```
(38837, 13)
```

In [ ]:
```
df.dtypes
```

Out[ ]:
```
name                             object
host_id                           int64
neighbourhood_group              object
neighbourhood                    object
latitude                        float64
longitude                       float64
room_type                        object
price                             int64
minimum_nights                    int64
number_of_reviews                 int64
reviews_per_month               float64
calculated_host_listings_count      int64
availability_365                  int64
dtype: object
```

In [ ]:
```
df.describe()
```

Out[ ]:

	host_id	latitude	longitude	price	minimum_nights	number_of_re
count	3.883700e+04	38837.000000	38837.000000	38837.000000	38837.000000	38837.0
mean	6.424425e+07	40.728135	-73.951145	142.314442	5.868450	29.3
std	7.589301e+07	0.054993	0.046696	196.959053	17.386079	48.1
min	2.438000e+03	40.506410	-74.244420	0.000000	1.000000	1.0
25%	7.033514e+06	40.688640	-73.982460	69.000000	1.000000	3.0
50%	2.837193e+07	40.721710	-73.954800	101.000000	2.000000	9.0
75%	1.018905e+08	40.762990	-73.935020	170.000000	4.000000	33.0
max	2.738417e+08	40.913060	-73.712990	10000.000000	1250.000000	629.0

In [ ]:
```
# convert room_type to quantitative values using pandas get_dummies
df_dummies = pd.get_dummies(df["room_type"], prefix="room_type")
df = df.join(df_dummies)
df.head()
```

Out[ ]:

	name	host_id	neighbourhood_group	neighbourhood	latitude	longitude	room_type
0	Clean & quiet apt home by the park	2787	Brooklyn	Kensington	40.64749	-73.97237	Private room
1	Skylit Midtown Castle	2845	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt
3	Cozy Entire Floor of Brownstone	4869	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt
4	Entire Apt: Spacious Studio/Loft by central park	7192	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt
5	Large Cozy 1 BR Apartment In Midtown East	7322	Manhattan	Murray Hill	40.74767	-73.97500	Entire home/apt

In [ ]:
```
df.drop(["room_type"], axis=1, inplace=True)
```

In [ ]:
```
df.dtypes
```

Out[ ]:

```
name                                object
host_id                              int64
neighbourhood_group                 object
neighbourhood                       object
latitude                           float64
longitude                          float64
price                                int64
minimum_nights                       int64
number_of_reviews                    int64
reviews_per_month                  float64
calculated_host_listings_count       int64
availability_365                     int64
room_type_Entire home/apt            uint8
room_type_Private room               uint8
room_type_Shared room                uint8
dtype: object
```