In [55]:

```python
# importing necessary libraries

import numpy as np # linear algebra
import pandas as pd # data processing
import matplotlib.pyplot as plt
import seaborn as sns
import os
for dirname, _, filenames in os.walk('AB_NYC_2019.csv'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
from sklearn.preprocessing import StandardScaler
```

In [3]:

```python
# to read .csv file of dataset

df=pd.read_csv("AB_NYC_2019.csv")
df.head(10)
```

Out[3]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calc
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	2018-10-19	0.21	
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	2019-05-21	0.38	
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	0	NaN	NaN	
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	2019-07-05	4.64	
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	2018-11-19	0.10	
5	5099	Large Cozy 1 BR Apartment In Midtown East	7322	Chris	Manhattan	Murray Hill	40.74767	-73.97500	Entire home/apt	200	3	74	2019-06-22	0.59	
6	5121	BlissArtsSpace!	7356	Garon	Brooklyn	Bedford-Stuyvesant	40.68688	-73.95596	Private room	60	45	49	2017-10-05	0.40	
7	5178	Large Furnished Room Near B'way	8967	Shunichi	Manhattan	Hell's Kitchen	40.76489	-73.98493	Private room	79	2	430	2019-06-24	3.47	
8	5203	Cozy Clean Guest Room - Family Apt	7490	MaryEllen	Manhattan	Upper West Side	40.80178	-73.96723	Private room	79	2	118	2017-07-21	0.99	
9	5238	Cute & Cozy Lower East Side 1 bdrm	7549	Ben	Manhattan	Chinatown	40.71344	-73.99037	Entire home/apt	150	1	160	2019-06-09	1.33	

In [4]:

```python
# Exploratory Data Analysis (EDA) to identify number of null

#Basic EDA
df.isnull().sum()
```

Out[4]:

```
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 [5]:

```python
# descriptive analysis to summarize data

df.describe()
```

Out[5]:

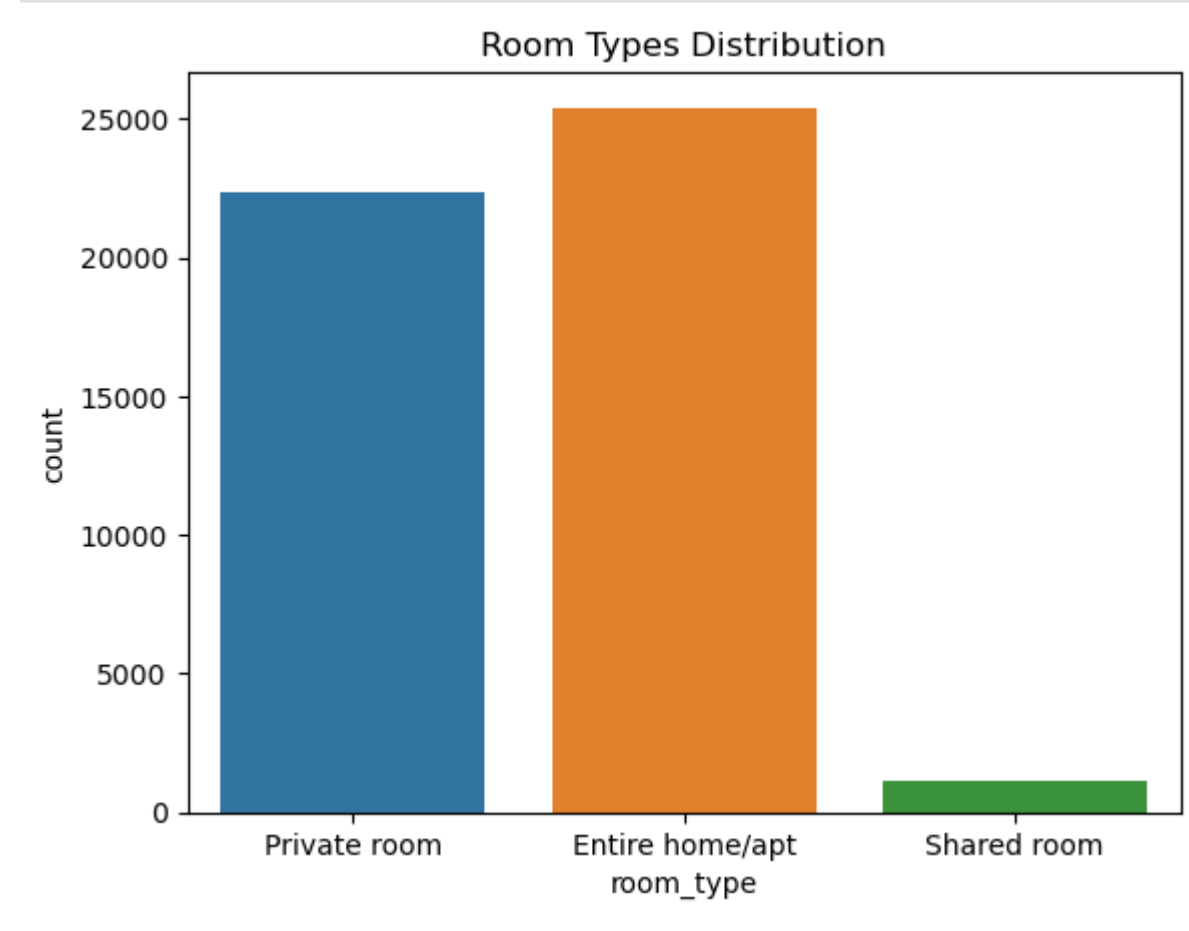
	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
count	4.889500e+04	4.889500e+04	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	38843.000000	48895.000000	48895.000000
mean	1.901714e+07	6.762001e+07	40.728949	-73.952170	152.720687	7.029962	23.274466	1.373221	7.143982	112.781327
std	1.098311e+07	7.861097e+07	0.054530	0.046157	240.154170	20.510550	44.550582	1.680442	32.952519	131.622289
min	2.539000e+03	2.438000e+03	40.499790	-74.244420	0.000000	1.000000	0.000000	0.010000	1.000000	0.000000
25%	9.471945e+06	7.822033e+06	40.690100	-73.983070	69.000000	1.000000	1.000000	0.190000	1.000000	0.000000
50%	1.967728e+07	3.079382e+07	40.723070	-73.955680	106.000000	3.000000	5.000000	0.720000	1.000000	45.000000
75%	2.915218e+07	1.074344e+08	40.763115	-73.936275	175.000000	5.000000	24.000000	2.020000	2.000000	227.000000
max	3.648724e+07	2.743213e+08	40.913060	-73.712990	10000.000000	1250.000000	629.000000	58.500000	327.000000	365.000000

In [6]:

```python
# Data Cleaning : by removing / dropping non-numeric columns

non_numeric_columns=df.select_dtypes(exclude=[np.number]).columns
df_numeric=df.drop(columns=non_numeric_columns,axis=1)
```

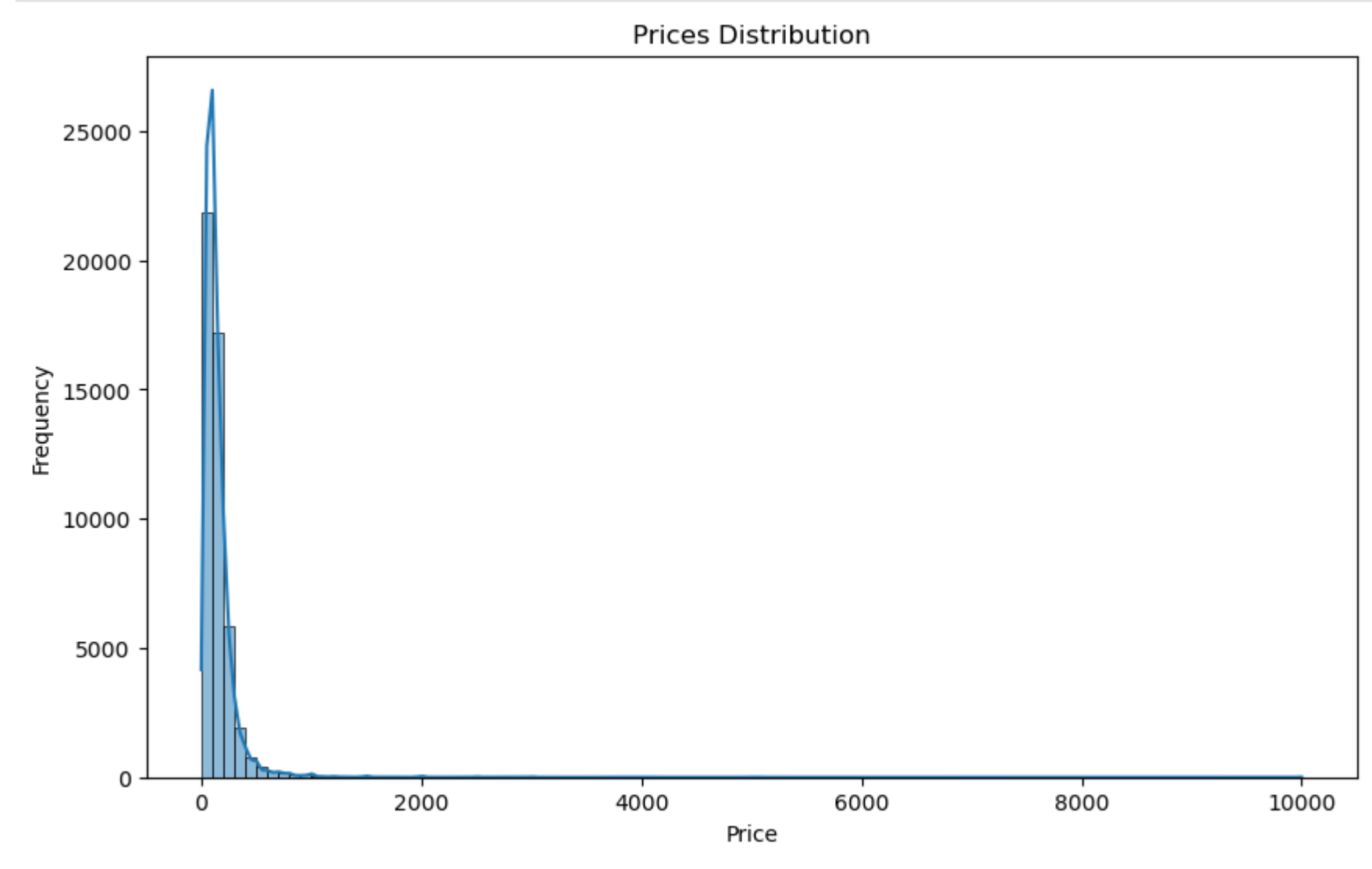
In [9]:

```python
# Distribution of different room types

sns.countplot(x='room_type',data=df)
plt.title("Room Types Distribution")
plt.show()
```



In [12]:

```python
# Distributiion of prices

plt.figure(figsize=(10,6))
sns.histplot(df['price'],bins=100,kde=True)
plt.title("Prices Distribution")
plt.xlabel("Price")
plt.ylabel("Frequency")
plt.show()
```



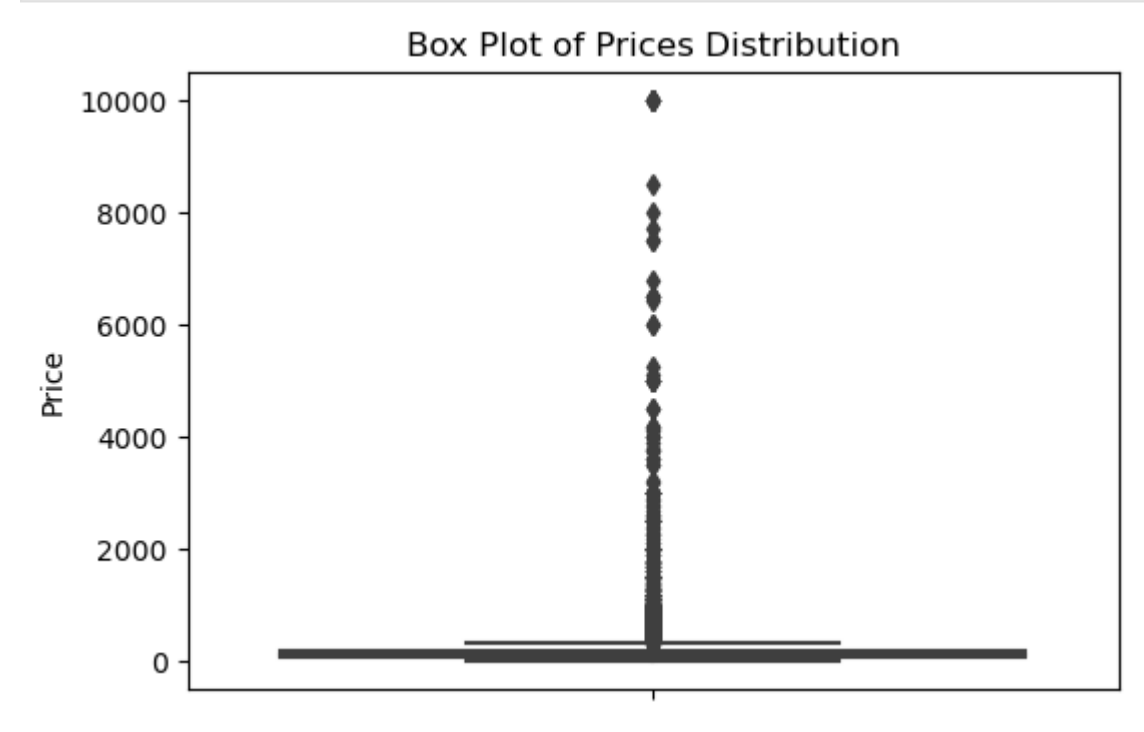
In [32]:

```python
print("Skewness: %f" % df['price'].skew())
print("Kurtosis: %f" % df['price'].kurt())

# as per skewness and kurtosis, we can say that the distribution is not normal or it doesn't provide equal distribution
```

```
Skewness: 19.118939
Kurtosis: 585.672879
```

In [41]:

```python
# Outliers detectios

plt.figure(figsize=(6,4))
sns.boxplot(data=df, y='price')
plt.title("Box Plot of Prices Distribution")
plt.ylabel("Price")
plt.show()
```



In [45]:

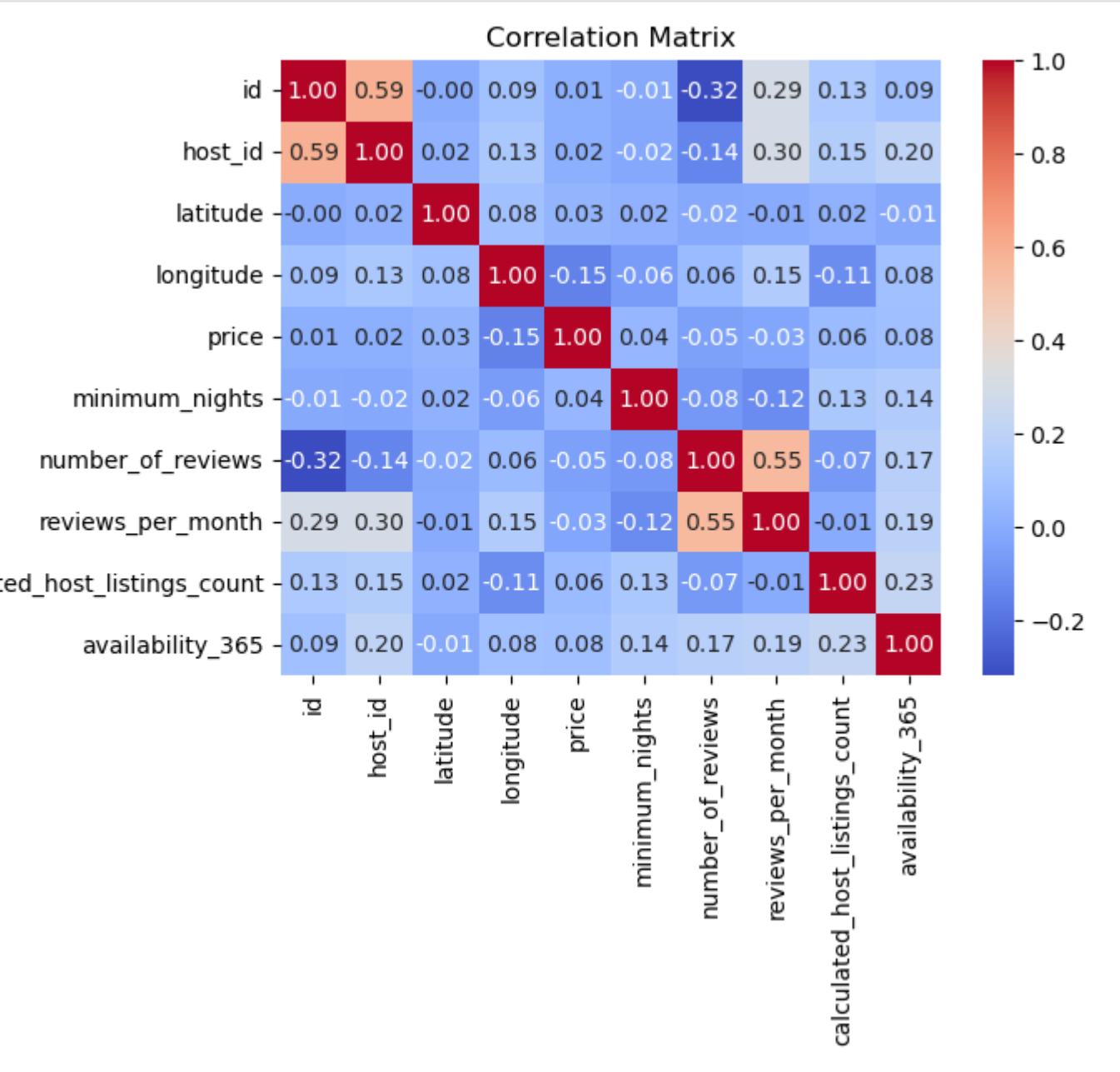
```python
Q1 = df['price'].quantile(0.25)
Q3 = df['price'].quantile(0.75)
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
outliers = df[(df['price'] < lower_bound) | (df['price'] > upper_bound)]
outliers.head()
```

Out[45]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculat
61	15396	Sunny & Spacious Chelsea Apartment	60278	Petra	Manhattan	Chelsea	40.74623	-73.99530	Entire home/apt	375	180	5	2018-11-03	0.12	
85	19601	perfect for a family or small group	74303	Maggie	Brooklyn	Brooklyn Heights	40.69723	-73.99268	Entire home/apt	800	1	25	2016-08-04	0.24	
103	23686	2000 SF 3br 2bath West Village private townhouse	93790	Ann	Manhattan	West Village	40.73096	-74.00319	Entire home/apt	500	4	46	2019-05-18	0.55	
114	26933	2 BR / 2 Bath Duplex Apt with patio! East Village	72062	Bruce	Manhattan	East Village	40.72540	-73.98157	Entire home/apt	350	2	7	2017-08-09	0.06	
121	27659	3 Story Town House in Park Slope	119588	Vero	Brooklyn	South Slope	40.66499	-73.97925	Entire home/apt	400	2	16	2018-12-30	0.24	

In [46]:

```python
#Correlation matrix
correlation_matrix=df_numeric.corr()
sns.heatmap(correlation_matrix,annot=True,cmap="coolwarm",fmt='.2f')
plt.title("Correlation Matrix")
plt.show()
```



Number of Reviews with Review per Month shows positive correlation with the specific value of 0.55

Feature Engineering

In [47]:

```python
# desc analysis of summarised information of Price
df.price.describe()
```

Out[47]:

```
count    48895.000000
mean       152.720687
std        240.154170
min          0.000000
25%         69.000000
50%        106.000000
75%        175.000000
max      10000.000000
Name: price, dtype: float64
```

In [48]:

```python
#Setting the min and max thresold for cleaning the data
minThresold,maxThresold=df.price.quantile([0.01,0.999])
minThresold,maxThresold
```

Out[48]:

```
(30.0, 3000.0)
```

In [49]:

```python
#Datapoints where price is less than minThresold (1 percentile) and greater than the maxThresold (99 percentile)

df1=df[(df.price<minThresold)|(df.price>maxThresold)]
df1.shape
```

Out[49]:

```
(449, 16)
```

In [50]:

```python
df1.head(10)
```

Out[50]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	ca
957	375249	Enjoy Staten Island Hospitality	1887999	Rimma & Jim	Staten Island	Graniteville	40.62109	-74.16534	Private room	20	3	80	2019-05-26	0.92	
1862	826690	Sunny, Family-Friendly 2 Bedroom	4289240	Lucy	Brooklyn	Prospect Heights	40.67919	-73.97191	Entire home/apt	4000	4	0	NaN	NaN	
2675	1428154	Central, Peaceful Semi-Private Room	5912572	Tangier	Brooklyn	Flatbush	40.63899	-73.95177	Shared room	29	2	5	2014-10-20	0.07	
2698	1448703	Beautiful 1 Bedroom in Nolita/Soho	213266	Jessica	Manhattan	Nolita	40.72193	-73.99379	Entire home/apt	5000	1	2	2013-09-28	0.03	
2860	1620248	Large furnished 2 bedrooms- - 30 days Minimum	2196224	Sally	Manhattan	East Village	40.73051	-73.98140	Entire home/apt	10	30	0	NaN	NaN	
3020	1767037	Small Cozy Room Wifi & AC near JFK	9284163	Antonio	Queens	Woodhaven	40.68968	-73.85219	Private room	29	2	386	2019-06-19	5.53	
3537	2110145	UWS 1BR w/backyard + block from CP	2151325	Jay And Liz	Manhattan	Upper West Side	40.77782	-73.97848	Entire home/apt	6000	14	17	2015-02-17	0.27	
3695	2224896	NYC SuperBowl Wk 5 Bdrs River View	11353904	Todd	Manhattan	Upper West Side	40.79476	-73.97299	Entire home/apt	4000	1	0	NaN	NaN	
3720	2243699	SuperBowl Penthouse Loft 3,000 sqft	1483320	Omri	Manhattan	Little Italy	40.71895	-73.99793	Entire home/apt	5250	1	0	NaN	NaN	
3774	2271504	SUPER BOWL Brooklyn Duplex Apt!!	11598359	Jonathan	Brooklyn	Clinton Hill	40.68766	-73.96439	Entire home/apt	6500	1	0	NaN	NaN	

In [51]:

```python
# Price distribution within min- and max- Thresold

df2=df[(df.price>minThresold)&(df.price<maxThresold)]
df2.shape #Datapoints where prices is in between the min and max thresold  (data points where price is lie between 1 percentile to 99 percentile)
```

Out[51]:

```
(48183, 16)
```

In [52]:

```python
df2.head(10)
```

Out[52]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calc
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	2018-10-19	0.21	
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	2019-05-21	0.38	
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	0	NaN	NaN	
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	2019-07-05	4.64	
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	2018-11-19	0.10	
5	5099	Large Cozy 1 BR Apartment In Midtown East	7322	Chris	Manhattan	Murray Hill	40.74767	-73.97500	Entire home/apt	200	3	74	2019-06-22	0.59	
6	5121	BlissArtsSpace!	7356	Garon	Brooklyn	Bedford-Stuyvesant	40.68688	-73.95596	Private room	60	45	49	2017-10-05	0.40	
7	5178	Large Furnished Room Near B'way	8967	Shunichi	Manhattan	Hell's Kitchen	40.76489	-73.98493	Private room	79	2	430	2019-06-24	3.47	
8	5203	Cozy Clean Guest Room - Family Apt	7490	MaryEllen	Manhattan	Upper West Side	40.80178	-73.96723	Private room	79	2	118	2017-07-21	0.99	
9	5238	Cute & Cozy Lower East Side 1 bdrm	7549	Ben	Manhattan	Chinatown	40.71344	-73.99037	Entire home/apt	150	1	160	2019-06-09	1.33	

In [53]:

```python
# descriptive analysis to summarize data of Price

df2.price.describe()
```

Out[53]:

```
count    48183.000000
mean       148.772036
std        153.594795
min         31.000000
25%         70.000000
50%        110.000000
75%        179.000000
max       2999.000000
Name: price, dtype: float64
```

In [54]:

```python
# Outliers removal

data = df[(df['price'] > lower_bound) & (df['price'] < upper_bound)]
data.describe()
```

Out[54]:

	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
count	4.591800e+04	4.591800e+04	45918.000000	45918.000000	45918.000000	45918.000000	45918.000000	36909.000000	45918.000000	45918.000000
mean	1.889785e+07	6.632478e+07	40.728487	-73.950728	119.947014	6.935973	23.944945	1.378257	6.620193	109.359358
std	1.091889e+07	7.756044e+07	0.055334	0.046471	68.117249	19.857728	45.317122	1.692033	30.938400	130.272996
min	2.539000e+03	2.438000e+03	40.499790	-74.244420	0.000000	1.000000	0.000000	0.010000	1.000000	0.000000
25%	9.436114e+06	7.722615e+06	40.689230	-73.981920	65.000000	1.000000	1.000000	0.190000	1.000000	0.000000
50%	1.952542e+07	3.028359e+07	40.721770	-73.954360	100.000000	2.000000	5.000000	0.710000	1.000000	39.000000
75%	2.891184e+07	1.054798e+08	40.763390	-73.934310	159.000000	5.000000	24.000000	2.020000	2.000000	216.000000
max	3.648724e+07	2.743213e+08	40.913060	-73.712990	333.000000	1250.000000	629.000000	58.500000	327.000000	365.000000

In [ ]: