Suyash pratap singh(181B226)

Aim:- Remove outliers using percentile based on price per night for a given apartment/home

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

In [4]:

```
df = pd.read_csv(r'C:\Users\Admin\Downloads\AB_NYC_2019.csv')
df.head(6)
```

Out[4]:

	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
5	5099	Large Cozy 1 BR Apartment In Midtown East	7322	Chris	Manhattan	Murray Hill	40.74767

In [5]:
```
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 [6]:
```
df.describe()
```

Out[6]:

	id	host_id	latitude	longitude	price	minimum_nights
count	4.889500e+04	4.889500e+04	48895.000000	48895.000000	48895.000000	48895.000000
mean	1.901714e+07	6.762001e+07	40.728949	-73.952170	152.720687	7.029962
std	1.098311e+07	7.861097e+07	0.054530	0.046157	240.154170	20.510550
min	2.539000e+03	2.438000e+03	40.499790	-74.244420	0.000000	1.000000
25%	9.471945e+06	7.822033e+06	40.690100	-73.983070	69.000000	1.000000
50%	1.967728e+07	3.079382e+07	40.723070	-73.955680	106.000000	3.000000
75%	2.915218e+07	1.074344e+08	40.763115	-73.936275	175.000000	5.000000
max	3.648724e+07	2.743213e+08	40.913060	-73.712990	10000.000000	1250.000000

In [7]:
```
df.shape
```

Out[7]:
```
(48895, 16)
```

Set a Threshold value for example with .95

```
In [8]: max_threshold = df['price'].quantile(.95)
        max_threshold
```

```
Out[8]: 355.0
```

Find out the value greater than threshold

In [9]:
```
df[df['price']>max_threshold]
```

Out[9]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	l
61	15396	Sunny & Spacious Chelsea Apartment	60278	Petra	Manhattan	Chelsea	4(
85	19601	perfect for a family or small group	74303	Maggie	Brooklyn	Brooklyn Heights	4(
103	23686	2000 SF 3br 2bath West Village private townhouse	93790	Ann	Manhattan	West Village	4(
121	27659	3 Story Town House in Park Slope	119588	Vero	Brooklyn	South Slope	4(
158	38663	Luxury Brownstone in Boerum Hill	165789	Sarah	Brooklyn	Boerum Hill	4(
...	...	...	...	...	...	...	
48748	36417250	US Open special 2-bed luxury condo	133288905	Cherie	Manhattan	Midtown	4(
48755	36419291	Wyndham Midtown 45 New York City 1 Bedroom Deluxe	273812306	Kelly	Manhattan	Midtown	4(
48757	36419574	Luxury & Spacious 1500 ft² MANHATTAN Townhouse	11454384	Ellen	Manhattan	Tribeca	4(
48833	36450896	Brand New 3-Bed Apt in the Best Location of FiDi	29741813	Yue	Manhattan	Financial District	4(
48839	36452721	Massage Spa. Stay overnight. Authors Artist dr...	274079964	Richard	Brooklyn	Sheepshead Bay	4(

2441 rows × 16 columns

Set a minimum threshold value with .05

In [10]:

```
min_threshold = df['price'].quantile(.05)
min_threshold
```

Out[10]:

```
40.0
```

Find out the value Smaller than threshold

In [11]:

```
df[df['price']<min_threshold]
```

Out[11]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhoo
36	11452	Clean and Quiet in Brooklyn	7355	Vt	Brooklyn	Bedforc Stuyvesar
249	62452	A SpeciaL!! Private Room in NY	303939	Lissette	Staten Island	Tompkinsvill
250	62461	B NYC Staten Alternative...	303939	Lissette	Staten Island	Tompkinsvill
251	62787	C Private Room By The Ferry	303939	Lissette	Staten Island	Tompkinsvill
256	63320	D Private Che@p Room 2 Explore NYC	303939	Lissette	Staten Island	Tompkinsvill
...	...	...	...	...	...	.
48851	36455649	#7 New Hotel-Like Private Room KING bed near JFK	263504959	David	Queens	Woodhave
48852	36455809	Cozy Private Room in Bushwick, Brooklyn	74162901	Christine	Brooklyn	Bushwic
48867	36473044	The place you were dreaming for. (only for guys)	261338177	Diana	Brooklyn	Gravesen
48868	36473253	Heaven for you(only for guy)	261338177	Diana	Brooklyn	Gravesen
48871	36475746	A LARGE ROOM - 1 MONTH MINIMUM - WASHER&DRYER	144008701	Ozzy Ciao	Manhattan	Harler

2042 rows × 16 columns

Filter dataset with min and max threshold contain no outliers

In [15]:
```
df1=df[(df['price']<max_threshold) & (df['price']>min_threshold)]
df1.head(6)
```

Out[15]:

	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
5	5099	Large Cozy 1 BR Apartment In Midtown East	7322	Chris	Manhattan	Murray Hill	40.74767

In [17]:
```
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 43631 entries, 0 to 48894
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   id                              43631 non-null  int64  
 1   name                            43617 non-null  object 
 2   host_id                         43631 non-null  int64  
 3   host_name                       43612 non-null  object 
 4   neighbourhood_group             43631 non-null  object 
 5   neighbourhood                   43631 non-null  object 
 6   latitude                        43631 non-null  float64
 7   longitude                       43631 non-null  float64
 8   room_type                       43631 non-null  object 
 9   price                           43631 non-null  int64  
 10  minimum_nights                  43631 non-null  int64  
 11  number_of_reviews               43631 non-null  int64  
 12  last_review                     35092 non-null  object 
 13  reviews_per_month               35092 non-null  float64
 14  calculated_host_listings_count  43631 non-null  int64  
 15  availability_365                43631 non-null  int64  
dtypes: float64(3), int64(7), object(6)
memory usage: 5.7+ MB
```

In [18]:
```
df1.shape
```

Out[18]:
```
(43631, 16)
```

In [19]:
```
df1.describe()
```

Out[19]:

	id	host_id	latitude	longitude	price	minimum_nights
count	4.363100e+04	4.363100e+04	43631.000000	43631.000000	43631.000000	43631.000000
mean	1.874542e+07	6.494019e+07	40.729397	-73.952633	128.201187	6.850359
std	1.097545e+07	7.726130e+07	0.054071	0.045249	70.515988	20.148614
min	2.539000e+03	2.438000e+03	40.499790	-74.244420	41.000000	1.000000
25%	9.218810e+06	7.362414e+06	40.690500	-73.982890	72.000000	1.000000
50%	1.932280e+07	2.869966e+07	40.723110	-73.955730	109.000000	2.000000
75%	2.878742e+07	1.020121e+08	40.763615	-73.937710	169.000000	5.000000
max	3.648724e+07	2.743213e+08	40.911690	-73.712990	353.000000	1250.000000

In [20]:
```
min_threshold , max_threshold = df1.host_id.quantile([.001,.999])
min_threshold , max_threshold
```

Out[20]:
```
(8543.95, 273373168.13000005)
```

```
In [22]: df1[df1.host_id<min_threshold]
```

Out[22]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem
5	5099	Large Cozy 1 BR Apartment In Midtown East	7322	Chris	Manhattan	Murray Hill
6	5121	BlissArtsSpace!	7356	Garon	Brooklyn	Bedford-Stuyvesant
8	5203	Cozy Clean Guest Room - Family Apt	7490	MaryEllen	Manhattan	Upper West Side
9	5238	Cute & Cozy Lower East Side 1 bdrm	7549	Ben	Manhattan	Chinatown
10	5295	Beautiful 1br on Upper West Side	7702	Lena	Manhattan	Upper West Side
11	5441	Central Manhattan/near Broadway	7989	Kate	Manhattan	Hell's Kitchen
69	16821	Large Room in Amazing East Village Apt	4396	Casey	Manhattan	East Village
184	46544	Park Slope haven 15 mins from Soho	8198	Monica	Brooklyn	Park Slope
272	64707	Amazing Sunny & Breezy Home In the Heart of NYC	7310	Tilly	Manhattan	Little Italy
650	246916	Quality Cozy Studio Next to Subway	3647	Rafael	Queens	Elmhurst
1378	609559	Queens Quality Convenient Apartment	3647	Rafael	Queens	Elmhurst
1418	636391	Charming Sunny W. Village Apt.	8425	Sharon	Manhattan	West Village

	id	name	host_id	host_name	neighbourhood_group	neighbourhood
2290	1101224	THE PUTNAM	2571	Teedo	Brooklyn	Bedford-Stuyvesant
3465	2075600	Delicious & Airy Apt in Landmark Brownstone	5089	Subhana	Brooklyn	Bedford-Stuyvesant
4446	3040654	Sunny Bedroom In Astoria!	6041	Miranda	Queens	Ditmars Steinway
4555	3172212	Spacious room in artist Loft	7351	Tanda	Brooklyn	South Slope
4767	3373030	Cute,Cozy Lower East Side 1bdrm	7549	Ben	Manhattan	Lower East Side
6073	4445185	2 Level Loft room in Artist Studio	7351	Tanda	Brooklyn	South Slope
6698	4815848	Elegant NYC, 10mins to Manhattan!	3211	Catherine	Queens	Long Island City
6699	4815886	Amazing Artist Loft	7351	Tanda	Brooklyn	South Slope
7875	6027345	Newly renovated historic brownstone	2881	Loli	Brooklyn	Bedford-Stuyvesant
8283	6373770	Huge private bedroom for quiet sleep in Manhat...	3867	Luke	Manhattan	Two Bridges
8302	6385039	Historic room in renovated brownstone	2881	Loli	Brooklyn	Bedford-Stuyvesant
8802	6751450	Next to Empire State building	7209	Liz	Manhattan	Midtown
9409	7208745	Beautiful Furnished Master Bedroom	3415	Nataraj	Queens	Fresh Meadows
9932	7645359	Old World Charm in Hip Brooklyn	7500	Russ	Brooklyn	Clinton Hill
10372	7937553	Riomaggiore Room. Queen Bedroom in Bklyn Townh...	2787	John	Brooklyn	Bensonhurst
13583	10160215	Torre del Lago Room.	2787	John	Brooklyn	Gravesend
13688	10267242	Cinque Terre Room. Clean and Quiet Queen Bedroom	2787	John	Brooklyn	Gravesend
13963	10593675	La Spezia room. Clean, quiet and comfortable bed	2787	John	Brooklyn	Bensonhurst

	id	name	host_id	host_name	neighbourhood_group	neighbourhood
14656	11574785	Queen bed & Air Conditioning, views of Chatham Sq	3867	Luke	Manhattan	Chinatown
16512	13234457	Cozy Clinton Hill Crib On Classon	2868	Letha M.	Brooklyn	Bedford-Stuyvesant
17631	13864551	Comfy Room in Amazing East Village Apt	4396	Casey	Manhattan	East Village
21556	17263207	Brooklyn home. Comfort and clean. Liguria room.	2787	John	Brooklyn	Bensonhurst
22728	18393354	Midtown Sanctuary	2845	Jennifer	Manhattan	Midtown
30604	23669201	Great Price: Williamsburg Brooklyn Loft off L ...	2438	Tasos	Brooklyn	Williamsburg
32078	25054120	DREAMY! Huge + sunny mid-century apt with balcony	8440	Michelle	Brooklyn	Flatbush
34627	27466647	Le Bain	6485	Saeko	Brooklyn	Bedford-Stuyvesant
42825	33245975	Crashpad in Clinton Hill	3151	Eric	Brooklyn	Clinton Hill

```
In [23]: df1.isnull()
```

Out[23]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitu
0	False	False	False	False	False	False	False	Fa
1	False	False	False	False	False	False	False	Fa
2	False	False	False	False	False	False	False	Fa
3	False	False	False	False	False	False	False	Fa
4	False	False	False	False	False	False	False	Fa
...	...	...	...	...	...	...	...	
48889	False	False	False	False	False	False	False	Fa
48890	False	False	False	False	False	False	False	Fa
48892	False	False	False	False	False	False	False	Fa
48893	False	False	False	False	False	False	False	Fa
48894	False	False	False	False	False	False	False	Fa

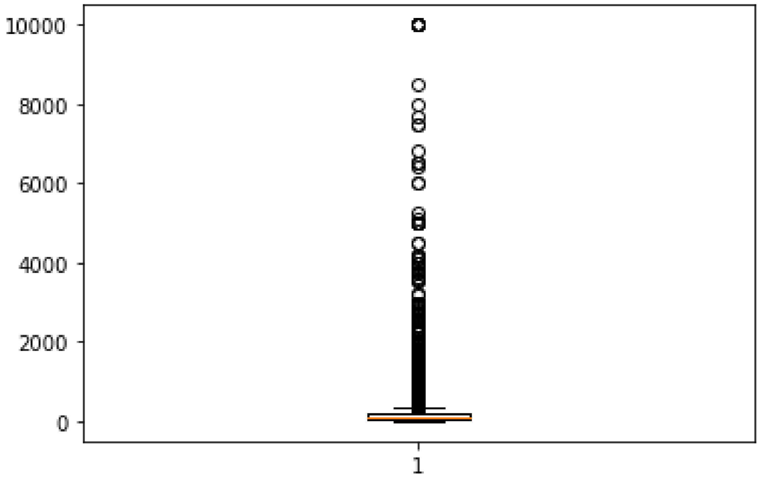
43631 rows × 16 columns

In [24]:
```
df1.isnull().sum()
```

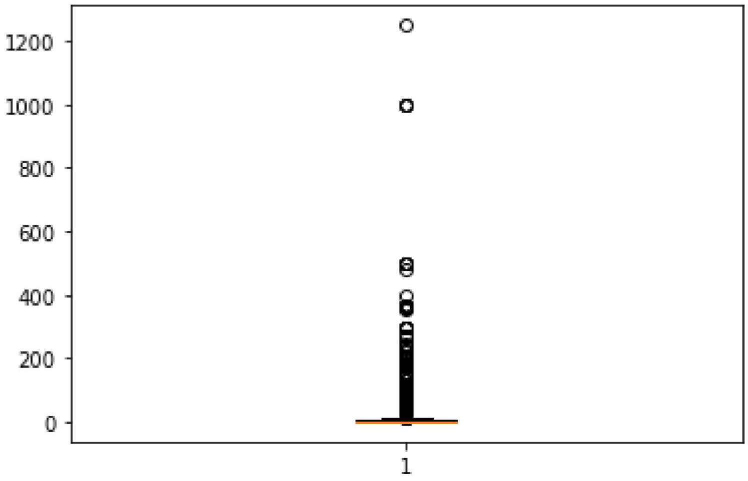
Out[24]:
```
id                                   0
name                                14
host_id                              0
host_name                           19
neighbourhood_group                  0
neighbourhood                        0
latitude                             0
longitude                            0
room_type                            0
price                                0
minimum_nights                       0
number_of_reviews                    0
last_review                       8539
reviews_per_month                 8539
calculated_host_listings_count       0
availability_365                     0
dtype: int64
```

PLOT BOXPLOT

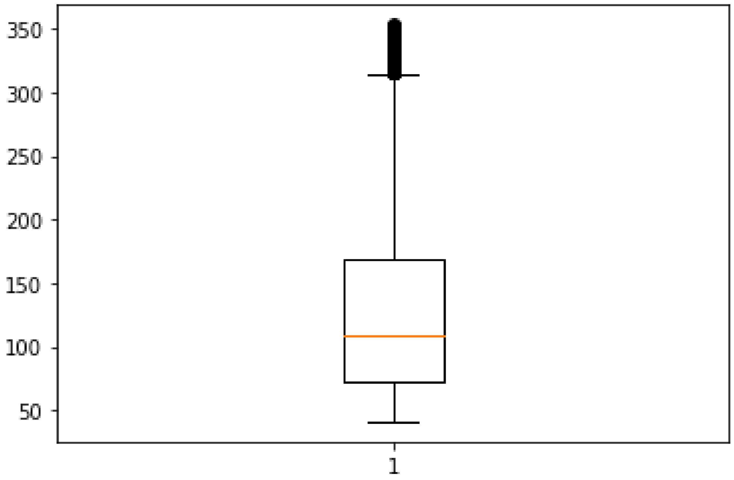
In [26]:
```
plt.boxplot(df['price'])
plt.show()
```



```
In [28]: plt.boxplot(df['minimum_nights'])
         plt.show()
```

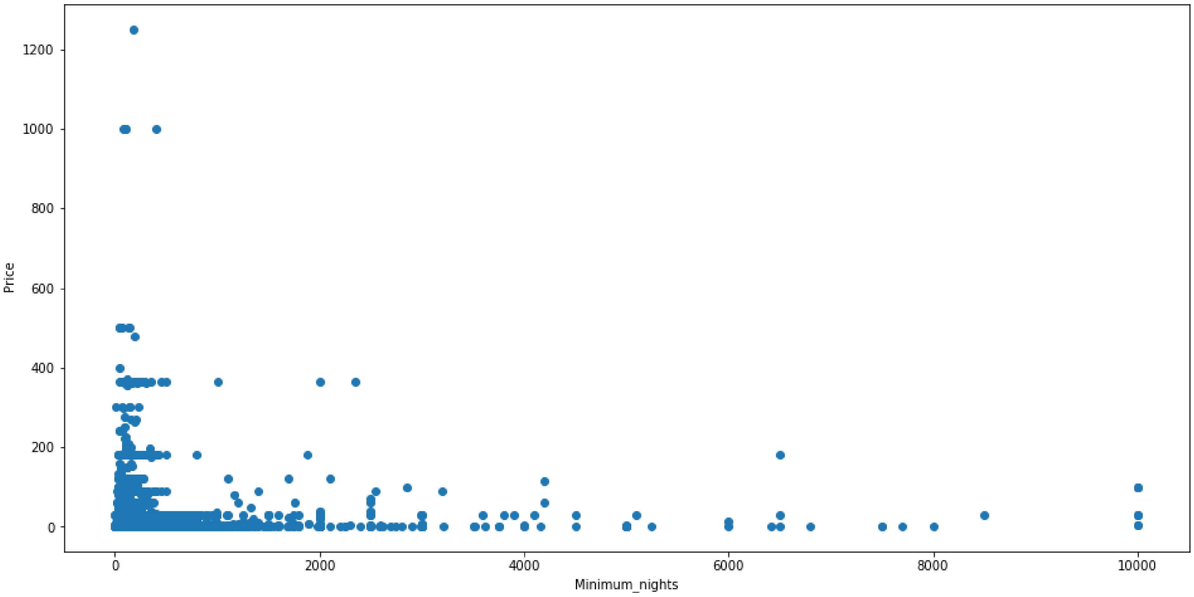


```
In [29]: plt.boxplot(df1['price'])
         plt.show()
```



Plot Scatter Plot

In [49]:
```
fig, ax = plt.subplots(figsize=(16,8))
ax.scatter(df['price'], df['minimum_nights'])
ax.set_xlabel('Minimum_nights')
ax.set_ylabel(' Price')
plt.show()
```



InterQuartile Range = Q3(.75)-Q1(.25)

In [37]:
```
Q1=df['price'].quantile(0.25)
Q3=df['price'].quantile(0.75)
IQR=Q3-Q1
IQR
```

Out[37]:
```
106.0
```

In [55]:

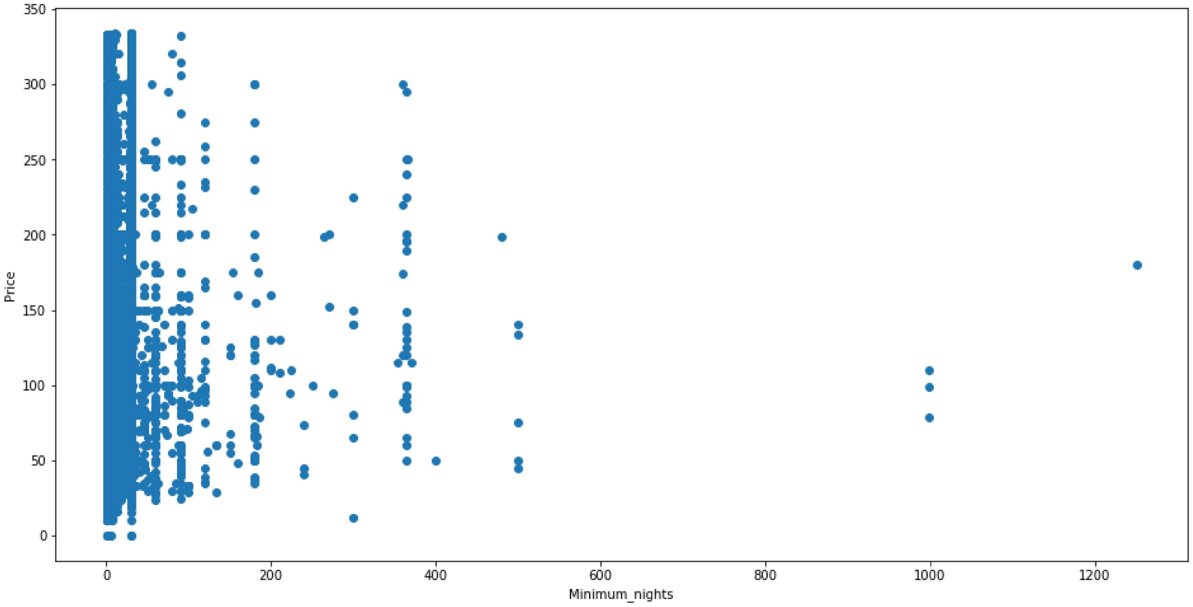
```
df2=df[~((df['price'] < (Q1 - 1.5*IQR)) | (df['price'] > (Q3+ 1.5*IQR)))]
df2.shape
print(df[((df['price'] < (Q1 - 1.5*IQR)) | (df['price'] > (Q3+ 1.5*IQR)))])
```

```
             id                                               name    host_id
\
61        15396                 Sunny & Spacious Chelsea Apartment      60278
85        19601                perfect for a family or small group      74303
103       23686  2000 SF 3br 2bath West Village private  townhouse      93790
114       26933  2 BR / 2 Bath Duplex Apt with patio! East Village      72062
121       27659                   3 Story Town House in Park Slope     119588
...         ...                                                ...        ...
48758  36420289    Rustic Garden House Apt, 2 stops from Manhattan   73211393
48833  36450896   Brand New 3-Bed Apt in the Best Location of FiDi   29741813
48839  36452721  Massage Spa. Stay overnight. Authors Artist dr...  274079964
48842  36453160  LUXURY MANHATTAN PENTHOUSE+HUDSON RIVER+EMPIRE...  224171371
48856  36457700  Large 3 bed, 2 bath , garden , bbq , all you need   66993395

                     host_name neighbourhood_group       neighbourhood  \
61                       Petra           Manhattan             Chelsea
85                      Maggie            Brooklyn    Brooklyn Heights
103                        Ann           Manhattan        West Village
114                      Bruce           Manhattan        East Village
121                       Vero            Brooklyn         South Slope
...                        ...                 ...                 ...
48758                LaGabrell              Queens    Long Island City
48833                      Yue           Manhattan  Financial District
48839                  Richard            Brooklyn      Sheepshead Bay
48842  LuxuryApartmentsByAmber           Manhattan             Chelsea
48856                   Thomas            Brooklyn  Bedford-Stuyvesant

       latitude  longitude        room_type  price  minimum_nights  \
61     40.74623  -73.99530  Entire home/apt    375             180
85     40.69723  -73.99268  Entire home/apt    800               1
103    40.73096  -74.00319  Entire home/apt    500               4
114    40.72540  -73.98157  Entire home/apt    350               2
121    40.66499  -73.97925  Entire home/apt    400               2
...         ...        ...              ...    ...             ...
48758  40.75508  -73.93258  Entire home/apt    350               2
48833  40.70605  -74.01042  Entire home/apt    475               2
48839  40.59866  -73.95661     Private room    800               1
48842  40.75204  -74.00292  Entire home/apt    350               1
48856  40.68886  -73.92879  Entire home/apt    345               4

       number_of_reviews last_review  reviews_per_month  \
61                     5  2018-11-03               0.12
85                    25  2016-08-04               0.24
103                   46  2019-05-18               0.55
114                    7  2017-08-09               0.06
121                   16  2018-12-30               0.24
...                  ...         ...                ...
48758                  0         NaN                NaN
48833                  0         NaN                NaN
48839                  0         NaN                NaN
48842                  0         NaN                NaN
48856                  0         NaN                NaN

       calculated_host_listings_count  availability_365
61                                  1               180
85                                  1                 7
103                                 2               243
```

```
114                                   4                 298
121                                   2                 216
...                                 ...                 ...
48758                                 1                 364
48833                                 1                  64
48839                                 1                  23
48842                                 1                   9
48856                                 3                 354

[2972 rows x 16 columns]
```

In [51]:

```
fig, ax = plt.subplots(figsize=(16,8))
ax.scatter(df2['minimum_nights'],df2['price'])
ax.set_xlabel('Minimum_nights')
ax.set_ylabel('Price')
plt.show()
```



In [ ]:

In [ ]: