In [5]:
```
import pandas as pd
import matplotlib
from matplotlib import pyplot as plt
%matplotlib inline
matplotlib.rcParams['figure.figsize'] = (12,8)
```

In [2]:
```
df = pd.read_csv("bhp.csv")
```

In [3]:
```
df
```

Out[3]:

	location	size	total_sqft	bath	price	bhk	price_per_sqft
0	Electronic City Phase II	2 BHK	1056.0	2.0	39.07	2	3699
1	Chikka Tirupathi	4 Bedroom	2600.0	5.0	120.00	4	4615
2	Uttarahalli	3 BHK	1440.0	2.0	62.00	3	4305
3	Lingadheeranahalli	3 BHK	1521.0	3.0	95.00	3	6245
4	Kothanur	2 BHK	1200.0	2.0	51.00	2	4250
...	...	...	...	...	...	...	...
13195	Whitefield	5 Bedroom	3453.0	4.0	231.00	5	6689
13196	other	4 BHK	3600.0	5.0	400.00	4	11111
13197	Raja Rajeshwari Nagar	2 BHK	1141.0	2.0	60.00	2	5258
13198	Padmanabhanagar	4 BHK	4689.0	4.0	488.00	4	10407
13199	Doddathoguru	1 BHK	550.0	1.0	17.00	1	3090

13200 rows × 7 columns

In [14]:
```
lower = df.price.quantile(0.1)
lower
```

Out[14]:
```
38.0
```

In [16]:
```
upper = df.price.quantile(0.999)
upper
```

Out[16]:
```
2000.0
```

In [32]:
```
df1_outliers= df[(df['price'] < lower) | (df['price'] > upper)]
df1_outliers
```

Out[32]:

	location	size	total_sqft	bath	price	bhk	price_per_sqft
20	Kengeri	1 BHK	600.0	1.0	15.00	1	2500
24	Thanisandra	1 RK	510.0	1.0	25.25	1	4950
26	Electronic City	2 BHK	660.0	1.0	23.10	2	3500
31	Bisuvanahalli	3 BHK	1075.0	2.0	35.00	3	3255
35	Kanakpura Road	2 BHK	700.0	2.0	36.00	2	5142
...	...	...	...	...	...	...	...
13105	Chandapura	1 BHK	520.0	1.0	14.04	1	2700
13124	Kereguddadahalli	2 BHK	1015.0	2.0	35.00	2	3448
13153	Raja Rajeshwari Nagar	1 BHK	510.0	1.0	22.00	1	4313
13171	other	1 Bedroom	812.0	1.0	26.00	1	3201
13199	Doddathoguru	1 BHK	550.0	1.0	17.00	1	3090

1309 rows × 7 columns

In [19]:
```
df.total_sqft.std(0.0)
```

Out[19]:
```
1237.3234454015123
```

In [21]:
```
df.head(n = 10)
```

Out[21]:

	location	size	total_sqft	bath	price	bhk	price_per_sqft
0	Electronic City Phase II	2 BHK	1056.0	2.0	39.07	2	3699
1	Chikka Tirupathi	4 Bedroom	2600.0	5.0	120.00	4	4615
2	Uttarahalli	3 BHK	1440.0	2.0	62.00	3	4305
3	Lingadheeranahalli	3 BHK	1521.0	3.0	95.00	3	6245
4	Kothanur	2 BHK	1200.0	2.0	51.00	2	4250
5	Whitefield	2 BHK	1170.0	2.0	38.00	2	3247
6	Old Airport Road	4 BHK	2732.0	4.0	204.00	4	7467
7	Rajaji Nagar	4 BHK	3300.0	4.0	600.00	4	18181
8	Marathahalli	3 BHK	1310.0	3.0	63.25	3	4828
9	other	6 Bedroom	1020.0	6.0	370.00	6	36274

In [22]:
```
df
```

Out[22]:

	location	size	total_sqft	bath	price	bhk	price_per_sqft
0	Electronic City Phase II	2 BHK	1056.0	2.0	39.07	2	3699
1	Chikka Tirupathi	4 Bedroom	2600.0	5.0	120.00	4	4615
2	Uttarahalli	3 BHK	1440.0	2.0	62.00	3	4305
3	Lingadheeranahalli	3 BHK	1521.0	3.0	95.00	3	6245
4	Kothanur	2 BHK	1200.0	2.0	51.00	2	4250
...	...	...	...	...	...	...	...
13195	Whitefield	5 Bedroom	3453.0	4.0	231.00	5	6689
13196	other	4 BHK	3600.0	5.0	400.00	4	11111
13197	Raja Rajeshwari Nagar	2 BHK	1141.0	2.0	60.00	2	5258
13198	Padmanabhanagar	4 BHK	4689.0	4.0	488.00	4	10407
13199	Doddathoguru	1 BHK	550.0	1.0	17.00	1	3090

13200 rows × 7 columns

In [33]:
```
df1_outliers
```

Out[33]:

	location	size	total_sqft	bath	price	bhk	price_per_sqft
20	Kengeri	1 BHK	600.0	1.0	15.00	1	2500
24	Thanisandra	1 RK	510.0	1.0	25.25	1	4950
26	Electronic City	2 BHK	660.0	1.0	23.10	2	3500
31	Bisuvanahalli	3 BHK	1075.0	2.0	35.00	3	3255
35	Kanakpura Road	2 BHK	700.0	2.0	36.00	2	5142
...	...	...	...	...	...	...	...
13105	Chandapura	1 BHK	520.0	1.0	14.04	1	2700
13124	Kereguddadahalli	2 BHK	1015.0	2.0	35.00	2	3448
13153	Raja Rajeshwari Nagar	1 BHK	510.0	1.0	22.00	1	4313
13171	other	1 Bedroom	812.0	1.0	26.00	1	3201
13199	Doddathoguru	1 BHK	550.0	1.0	17.00	1	3090

1309 rows × 7 columns

In [2]:
```
df = pd.read_csv("bhp.csv")
df.head()
```

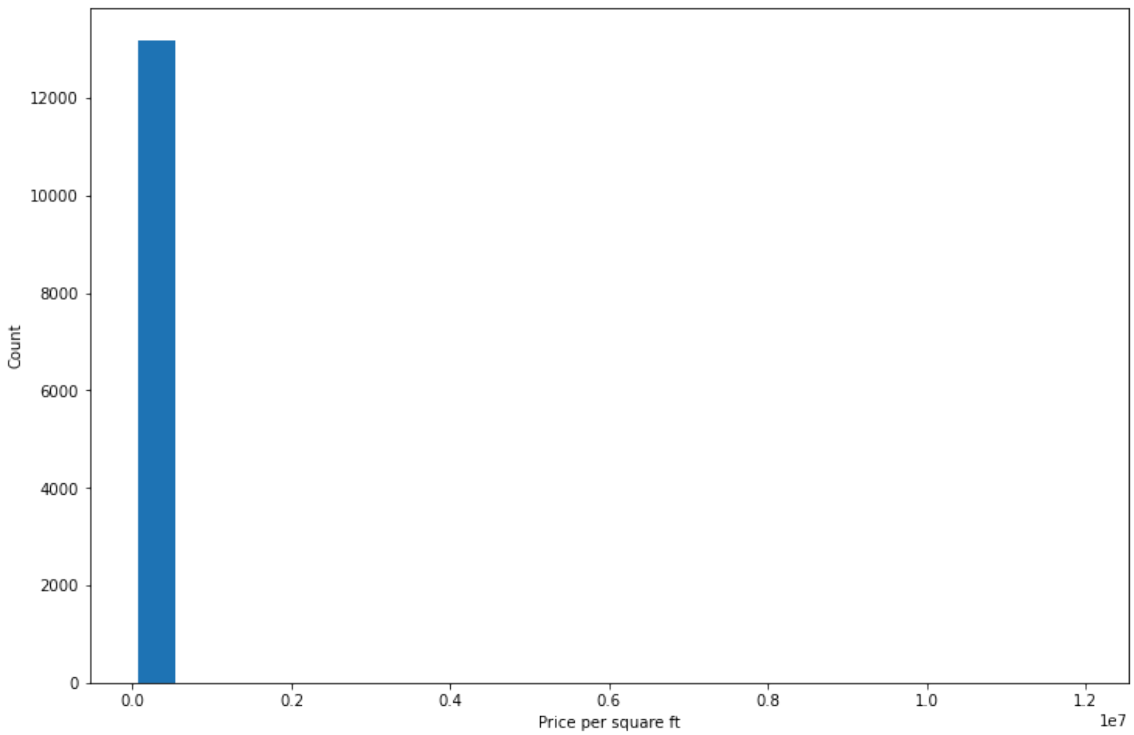
Out[2]:

	location	size	total_sqft	bath	price	bhk	price_per_sqft
0	Electronic City Phase II	2 BHK	1056.0	2.0	39.07	2	3699
1	Chikka Tirupathi	4 Bedroom	2600.0	5.0	120.00	4	4615
2	Uttarahalli	3 BHK	1440.0	2.0	62.00	3	4305
3	Lingadheeranahalli	3 BHK	1521.0	3.0	95.00	3	6245
4	Kothanur	2 BHK	1200.0	2.0	51.00	2	4250

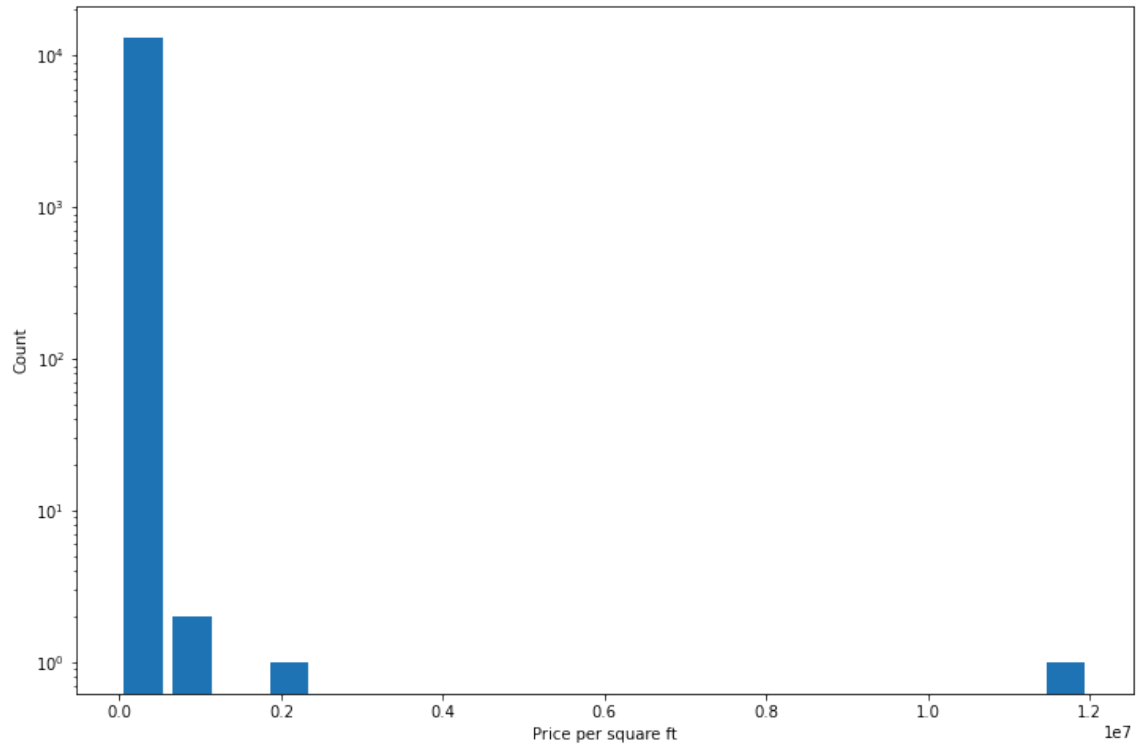
In [3]:
```
df.price_per_sqft.describe()
```

Out[3]:
```
count    1.320000e+04
mean     7.920337e+03
std      1.067272e+05
min      2.670000e+02
25%      4.267000e+03
50%      5.438000e+03
75%      7.317000e+03
max      1.200000e+07
Name: price_per_sqft, dtype: float64
```

In [6]:
```
plt.hist(df.price_per_sqft, bins=20, rwidth=0.8)
plt.xlabel('Price per square ft')
plt.ylabel('Count')
plt.show()
```



In [7]:
```
plt.hist(df.price_per_sqft, bins=20, rwidth=0.8)
plt.xlabel('Price per square ft')
plt.ylabel('Count')
plt.yscale('log')
plt.show()
```



In [8]:
```
lower_limit, upper_limit = df.price_per_sqft.quantile([0.001, 0.999])
lower_limit, upper_limit
```

Out[8]: (1366.184, 50959.36200000099)

In [9]:
```
lower_limit, upper_limit = df.price_per_sqft.quantile([0.001, 0.999])
lower_limit, upper_limit
```

Out[9]: (1366.184, 50959.36200000099)

In [10]:
```
df2 = df[(df.price_per_sqft<upper_limit) & (df.price_per_sqft>lower_limit)]
df2.shape
```

Out[10]: (13172, 7)

In [11]:
```
df.shape
```

Out[11]: (13200, 7)

In [12]:
```
df.shape[0] - df2.shape[0]
```

Out[12]: 28

In [13]:
```
max_limit = df2.price_per_sqft.mean() + 4*df2.price_per_sqft.std()
min_limit = df2.price_per_sqft.mean() - 4*df2.price_per_sqft.std()
max_limit, min_limit
```

Out[13]: (23227.73653589429, -9900.429065502549)

In [14]:
```
df2[(df2.price_per_sqft>max_limit) | (df2.price_per_sqft<min_limit)].sample
```

Out[14]:

	location	size	total_sqft	bath	price	bhk	price_per_sqft
3420	other	3 Bedroom	1350.0	3.0	380.0	3	28148
3816	Domlur	6 BHK	2400.0	4.0	600.0	6	25000
9711	Rajaji Nagar	2 Bedroom	1056.0	1.0	250.0	2	23674
3752	other	4 Bedroom	1200.0	4.0	300.0	4	25000
1721	other	5 Bedroom	2400.0	5.0	625.0	5	26041
3144	other	5 BHK	8321.0	5.0	2700.0	5	32448
12631	Rajaji Nagar	5 Bedroom	2500.0	4.0	650.0	5	26000
3488	Banashankari Stage III	8 Bedroom	1200.0	7.0	350.0	8	29166
13078	other	4 Bedroom	9200.0	4.0	2600.0	4	28260
11919	other	3 Bedroom	1524.0	4.0	400.0	3	26246

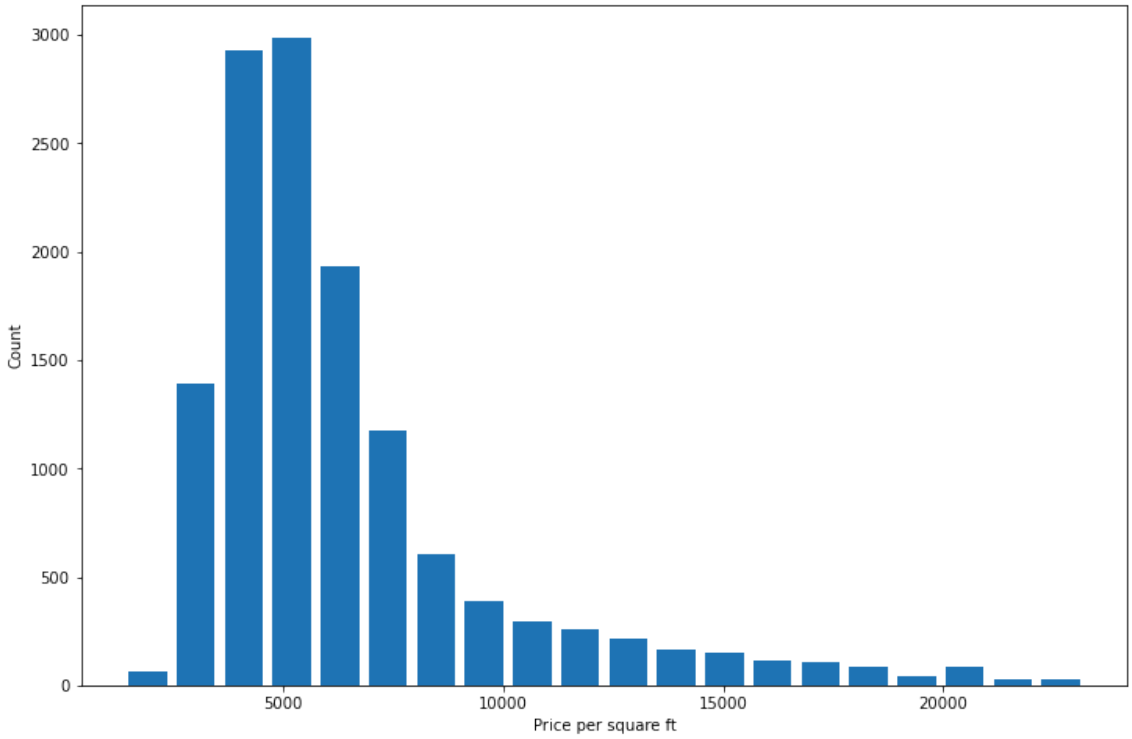
In [15]:
```
df3 = df2[(df2.price_per_sqft>min_limit) & (df2.price_per_sqft<max_limit)]
df3.shape
```

Out[15]: (13047, 7)

In [16]:
```
df2.shape[0]-df3.shape[0]
```

Out[16]: 125

In [17]:
```
plt.hist(df3.price_per_sqft, bins=20, rwidth=0.8)
plt.xlabel('Price per square ft')
plt.ylabel('Count')
plt.show()
```

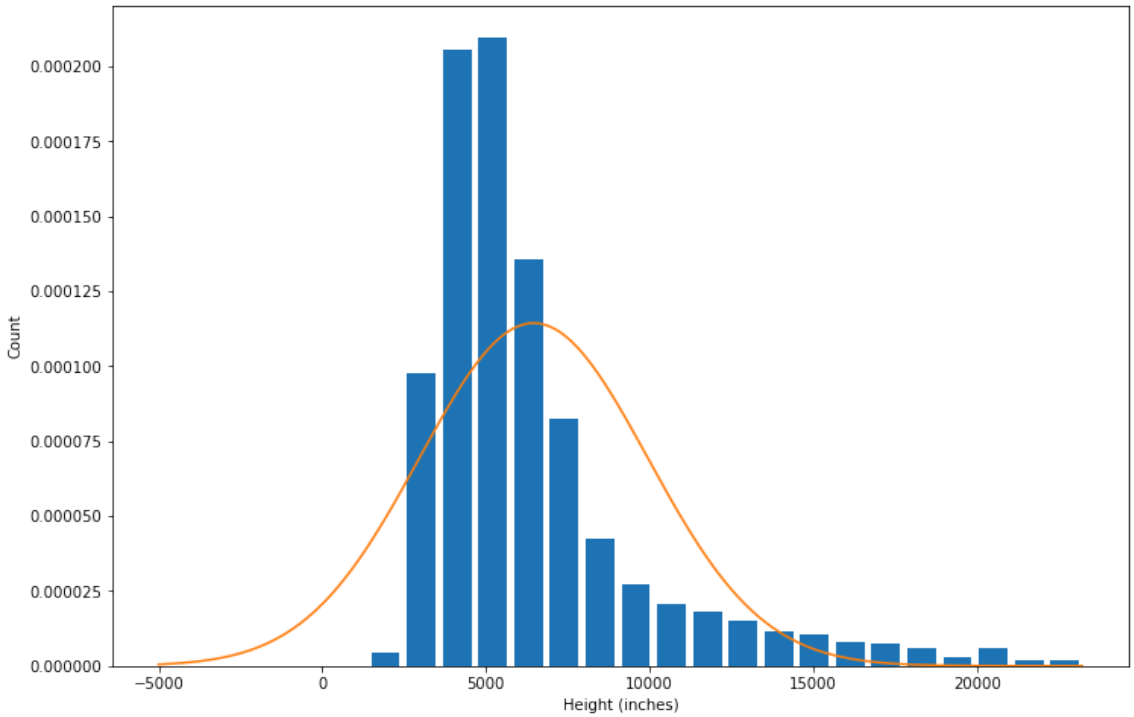


In [18]:
```
from scipy.stats import norm
import numpy as np

plt.hist(df3.price_per_sqft, bins=20, rwidth=0.8, density=True)
plt.xlabel('Height (inches)')
plt.ylabel('Count')

rng = np.arange(-5000, df3.price_per_sqft.max(), 100)
plt.plot(rng, norm.pdf(rng,df3.price_per_sqft.mean(),df3.price_per_sqft.std
```

Out[18]:
```
[<matplotlib.lines.Line2D at 0x1c6a784feb0>]
```



In [19]:
```
df2['zscore'] = (df2.price_per_sqft-df2.price_per_sqft.mean())/df2.price_pe
df2.sample(10)
```

```
<ipython-input-19-70754eb1335e>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  df2['zscore'] = (df2.price_per_sqft-df2.price_per_sqft.mean())/df2.price
_per_sqft.std()
```

Out[19]:

	location	size	total_sqft	bath	price	bhk	price_per_sqft	zscore
4992	Green Glen Layout	3 BHK	1530.0	3.0	105.0	3	6862	0.047898
7639	Indira Nagar	2 BHK	1475.0	2.0	171.0	2	11593	1.190370
1423	other	3 BHK	2197.0	4.0	280.0	3	12744	1.468321
677	Whitefield	2 BHK	1140.0	2.0	56.0	2	4912	-0.423000
5128	Sahakara Nagar	3 BHK	1200.0	2.0	75.0	3	6250	-0.099892
963	Ramamurthy Nagar	5 Bedroom	1640.0	4.0	240.0	5	14634	1.924730
4801	Kengeri Satellite Town	2 BHK	1030.0	2.0	50.0	2	4854	-0.437007
7508	Hormavu	4 Bedroom	3500.0	4.0	289.0	4	8257	0.384771
1152	Yelahanka	2 BHK	1170.0	2.0	62.5	2	5341	-0.319403
1226	Uttarahalli	2 BHK	1150.0	2.0	49.0	2	4260	-0.580450

In [20]:
```
outliers_z = df2[(df2.zscore < -4) | (df2.zscore>4)]
outliers_z.shape
```

Out[20]: (125, 8)

In [21]:
```
outliers_z.sample(5)
```

Out[21]:

	location	size	total_sqft	bath	price	bhk	price_per_sqft	zscore
978	Rajaji Nagar	4 Bedroom	315.0	4.0	90.0	4	28571	5.290325
13185	Hulimavu	1 BHK	500.0	1.0	220.0	1	44000	9.016218
4119	other	4 Bedroom	7000.0	5.0	2050.0	4	29285	5.462746
3752	other	4 Bedroom	1200.0	4.0	300.0	4	25000	4.427977
13094	other	4 Bedroom	1200.0	5.0	325.0	4	27083	4.930994

In [22]:
```
df4 = df2[(df2.zscore>-4)&(df2.zscore<4)]
df4.shape
```

Out[22]: (13047, 8)

In [23]:
```
df2.shape[0] - df4.shape[0]
```

Out[23]:
```
125
```

In [ ]:
```

```