In [1]:
```
#Experiment No.3
```

In [2]:
```
#Aim: To perform Outlier Detection and Removal Using IQR

#Name: Sakshi Padmakar Yeole
#Class: 3rd yr(B)
#Subject:ET-II
#Roll no.:69
```

In [6]:
```
import pandas as pd
import os
```

In [7]:
```
os.getcwd()
```

Out[7]:
```
'C:\\Users\\hp'
```

In [8]:
```
os.chdir("C:\\Users\\hp\\Downloads")
```

In [9]:
```
df = pd.read_csv("height.csv")
df
```

Out[9]:

	Gender	Height
0	Male	73.847017
1	Male	68.781904
2	Male	74.110105
3	Male	71.730978
4	Male	69.881796
...	...	...
9995	Female	66.172652
9996	Female	67.067155
9997	Female	63.867992
9998	Female	69.034243
9999	Female	61.944246

10000 rows × 2 columns

In [10]:
```
df.describe()
```

Out[10]:

	Height
count	10000.000000
mean	66.367560
std	3.847528
min	54.263133
25%	63.505620
50%	66.318070
75%	69.174262
max	78.998742

In [11]:
```
Q1 = df.Height.quantile(0.25)
Q3 = df.Height.quantile(0.75)
Q1, Q3
```

Out[11]:
```
(63.505620480000005, 69.17426172750001)
```

In [12]:
```
IQR = Q3 - Q1
```

```
IQR
```

Out[12]:
```
5.668641247500005
```

In [13]:
```
lower_limit = Q1 - 1.5*IQR
upper_limit = Q3 + 1.5*IQR
lower_limit, upper_limit
```

Out[13]:
```
(55.00265860875, 77.67722359875002)
```

In [14]:
```
df[(df.Height<lower_limit)|(df.Height>upper_limit)]
```

Out[14]:

	Gender	Height
994	Male	78.095867
1317	Male	78.462053
2014	Male	78.998742
3285	Male	78.528210
3757	Male	78.621374
6624	Female	54.616858
7294	Female	54.873728
9285	Female	54.263133

In [15]:
```
df_no_outlier = df[(df.Height>lower_limit)&(df.Height<upper_limit)]
df_no_outlier
```

Out[15]:

	Gender	Height
0	Male	73.847017
1	Male	68.781904
2	Male	74.110105
3	Male	71.730978
4	Male	69.881796
...	...	...
9995	Female	66.172652
9996	Female	67.067155
9997	Female	63.867992
9998	Female	69.034243
9999	Female	61.944246

9992 rows × 2 columns

In [16]:
```
import pandas as pd
import matplotlib
from matplotlib import pyplot as plt
%matplotlib inline
matplotlib.rcParams['figure.figsize'] = (12,8)
```

In [17]:
```
df = pd.read_csv("height.csv")
df.head()
```

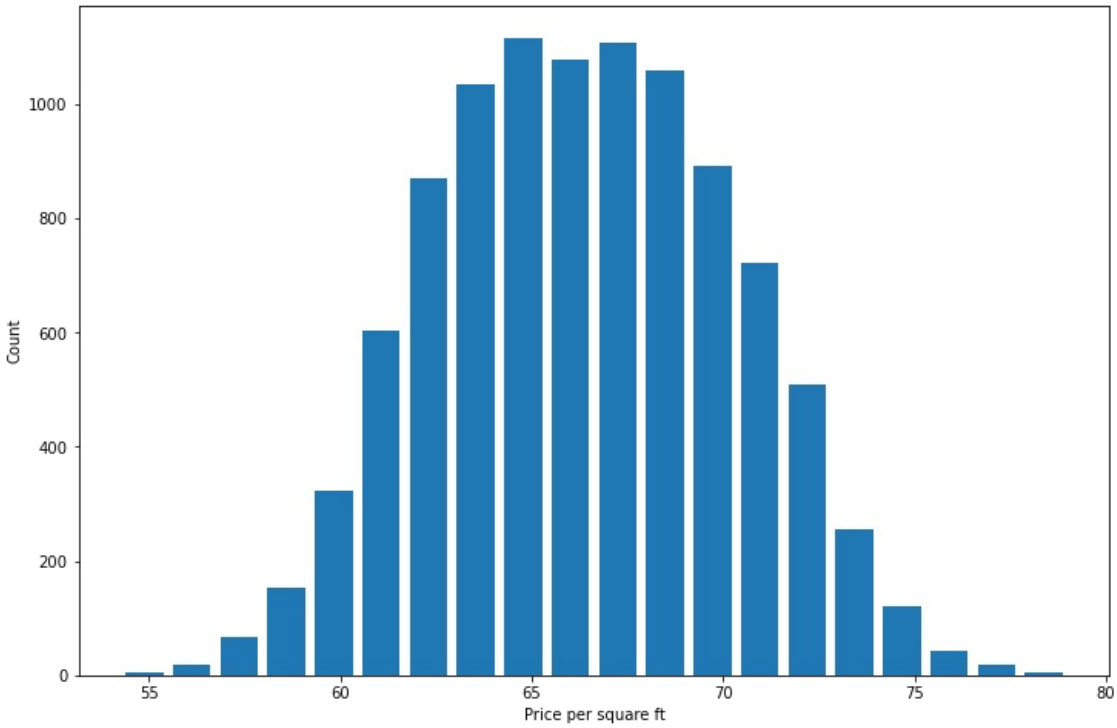
Out[17]:

	Gender	Height
0	Male	73.847017
1	Male	68.781904
2	Male	74.110105
3	Male	71.730978
4	Male	69.881796

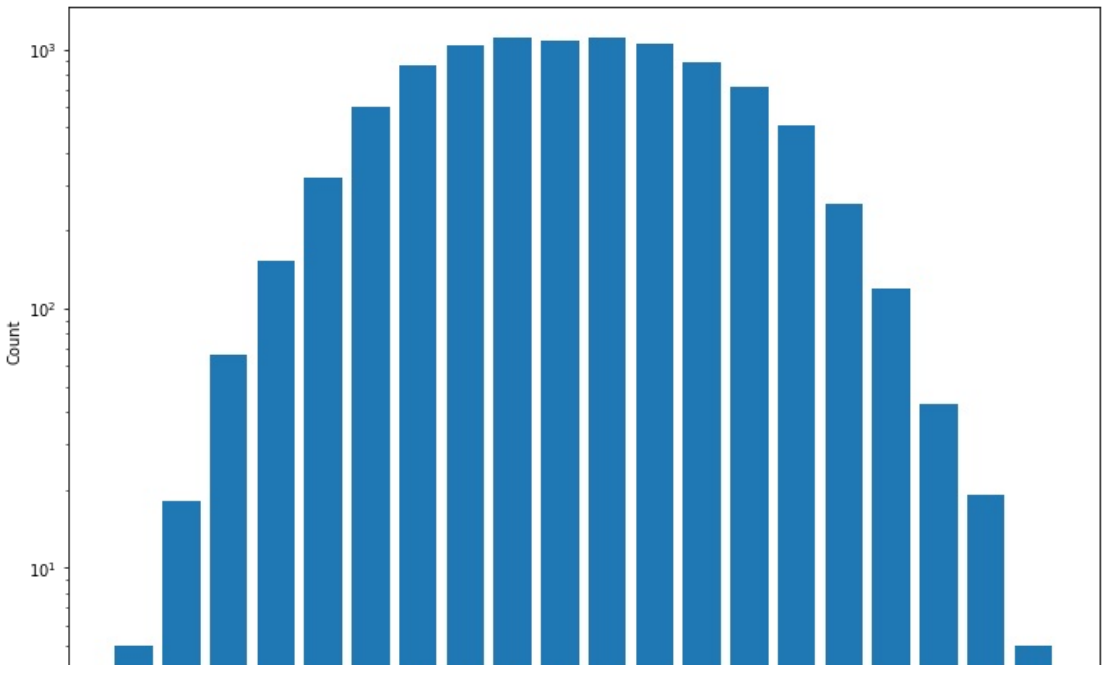
In [18]:
```
df.Height.describe()
```

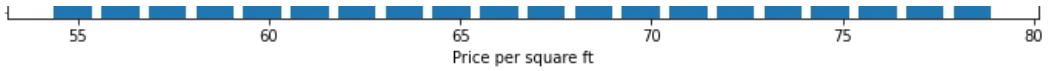
Out[18]:
```
count    10000.000000
mean        66.367560
std          3.847528
min         54.263133
25%         63.505620
50%         66.318070
75%         69.174262
max         78.998742
Name: Height, dtype: float64
```

In [19]:
```
plt.hist(df.Height, bins=20, rwidth=0.8)
plt.xlabel('Price per square ft')
plt.ylabel('Count')
plt.show()
```



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



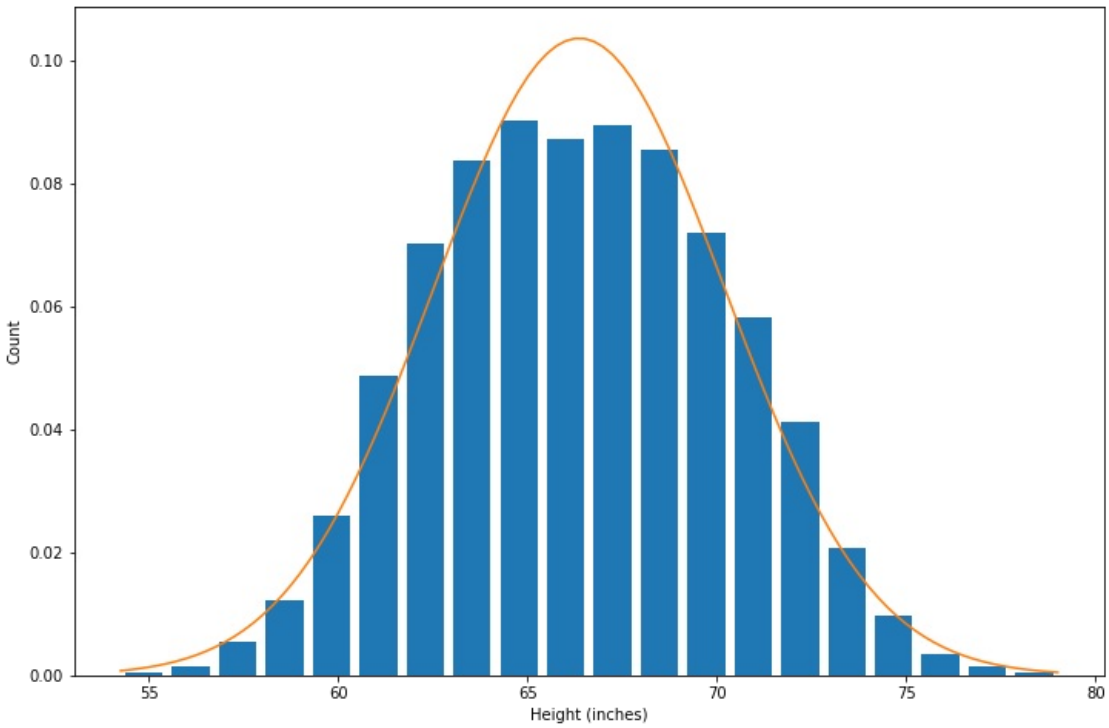


In [21]:
```
from scipy.stats import norm
import numpy as np

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

rng = np.linspace(54.263133, df.Height.max(), 100)
plt.plot(rng, norm.pdf(rng,df.Height.mean(),df.Height.std()))
```

Out[21]:
```
[<matplotlib.lines.Line2D at 0x1e2d9b2d160>]
```



In [22]:
```
lower_limit, upper_limit = df.Height.quantile([0.001, 0.999])
lower_limit, upper_limit
```

Out[22]:
```
(56.066548911530006, 77.06738853708)
```

In [23]:
```
outliers = df[(df.Height>upper_limit) | (df.Height<lower_limit)]
outliers.sample(10)
```

Out[23]:

	Gender	Height
5360	Female	55.668202
4297	Male	77.100872
7294	Female	54.873728
9825	Female	55.979198
2014	Male	78.998742
9285	Female	54.263133
994	Male	78.095867
7617	Female	55.148557
6624	Female	54.616858
5345	Female	55.336492

In [24]:
```
df2 = df[(df.Height<upper_limit) & (df.Height>lower_limit)]
df2.shape
```

Out[24]: (9980, 2)

In [25]:
```
df.shape
```

Out[25]: (10000, 2)

In [26]:
```
df.shape[0] - df2.shape[0]
```

Out[26]: 20

In [27]:
```
max_limit = df.Height.mean() + 4*df.Height.std()
min_limit = df.Height.mean() - 4*df.Height.std()
max_limit, min_limit
```

Out[27]: (81.75767223804789, 50.9774472716833)

In [28]:
```
max_limit = df2.Height.mean() + 3 * df2.Height.std()
min_limit = df2.Height.mean() - 3 * df2.Height.std()
```

In [29]:
```
df[(df.Height>max_limit) | (df.Height<min_limit)].sample(5)
```

Out[29]:

	Gender	Height
6624	Female	54.616858
3285	Male	78.528210
1317	Male	78.462053
9285	Female	54.263133
2014	Male	78.998742

In [30]:
```
df = df[(df.Height>min_limit) & (df.Height<max_limit)]
df.shape
```

Out[30]: (9992, 2)

In [31]:
```
df['zscore'] = (df.Height-df.Height.mean())/df.Height.std()
df.sample(10)
```

Out[31]:

	Gender	Height	zscore
9110	Female	62.995039	-0.878973
1388	Male	67.294018	0.242310
9481	Female	67.726674	0.355158
1812	Male	66.475562	0.028836
2446	Male	68.874876	0.654638
2845	Male	68.845117	0.646876
8466	Female	62.218599	-1.081489
1153	Male	69.013335	0.690752
4196	Male	70.187849	0.997095
3098	Male	70.839872	1.167160

In [32]:
```
outliers_z = df[(df.zscore < -4) | (df.zscore>4)]
outliers_z.shape
```

Out[32]: (0, 3)

In [33]:
```
outliers_z.sample(0)
```

Out[33]:

Gender	Height	zscore

In [ ]:
```

```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js