Importing the Libraries

```
import numpy as np
import sklearn
import pandas as pd
from sklearn.datasets import load_breast_cancer
import matplotlib.pyplot as plt
```

Load the breast cancer dataset

```
br=load_breast_cancer()
data=np.c_[br.data,br.target]
columns=np.append(br.feature_names, ["target"])
df=pandas.DataFrame(data, columns=columns)
df
```

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.30010	0.14710
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.08690	0.07017
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.19740	0.12790
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.24140	0.10520
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.19800	0.10430
...	...	...	...	...	...	...	...	...
564	21.56	22.39	142.00	1479.0	0.11100	0.11590	0.24390	0.13890
565	20.13	28.25	131.20	1261.0	0.09780	0.10340	0.14400	0.09791
566	16.60	28.08	108.30	858.1	0.08455	0.10230	0.09251	0.05302
567	20.60	29.33	140.10	1265.0	0.11780	0.27700	0.35140	0.15200
568	7.76	24.54	47.92	181.0	0.05263	0.04362	0.00000	0.00000

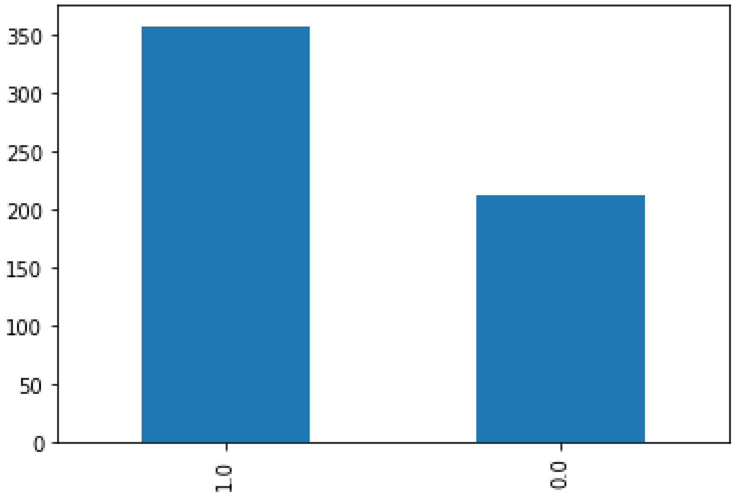
569 rows × 31 columns

Find the class distribution

```
df['target'].value_counts().plot(kind='bar',y=['benign','malignant'])
v=df['target'].value_counts().to_dict()
```

```
print("Benign tumour counts:"+str(v[1.0]))
print("Malignant tumour counts:"+str(v[0.0]))
```

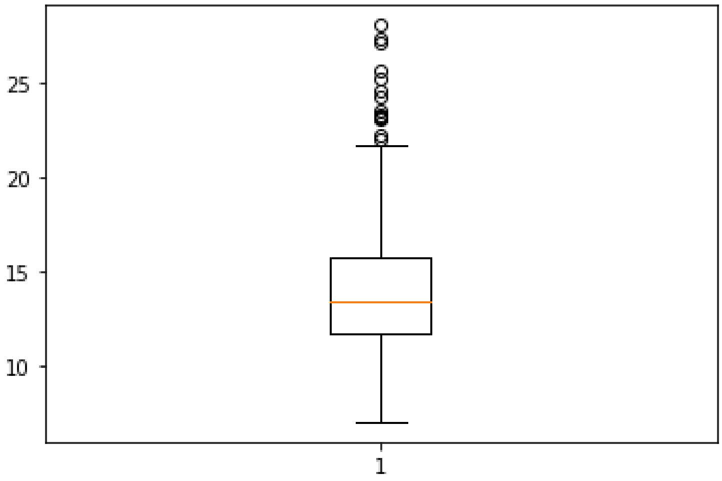
```
Benign tumour counts:357
Malignant tumour counts:212
```



Demonstrate five number summary and boxplot.

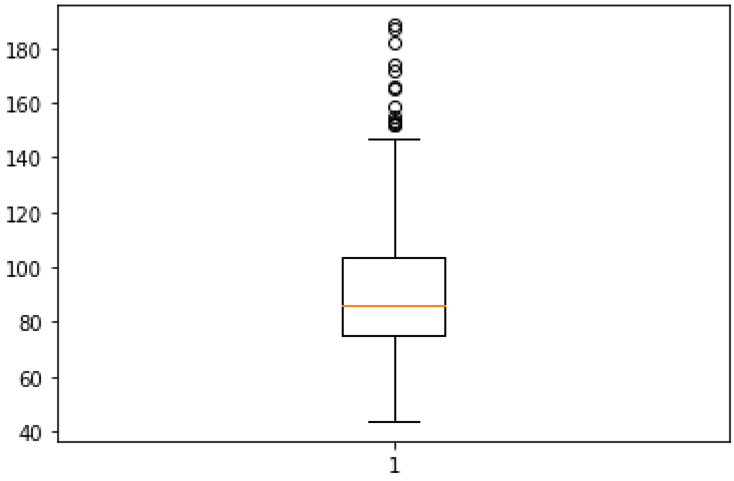
```
print("5 point summary and box plot for mean-radius")
print(df['mean radius'].describe(percentiles=[.25,.5,.75]))
p=plt.boxplot(df['mean radius'])
```

```
5 point summary and box plot for mean-radius
count    569.000000
mean      14.127292
std        3.524049
min        6.981000
25%       11.700000
50%       13.370000
75%       15.780000
max       28.110000
Name: mean radius, dtype: float64
```



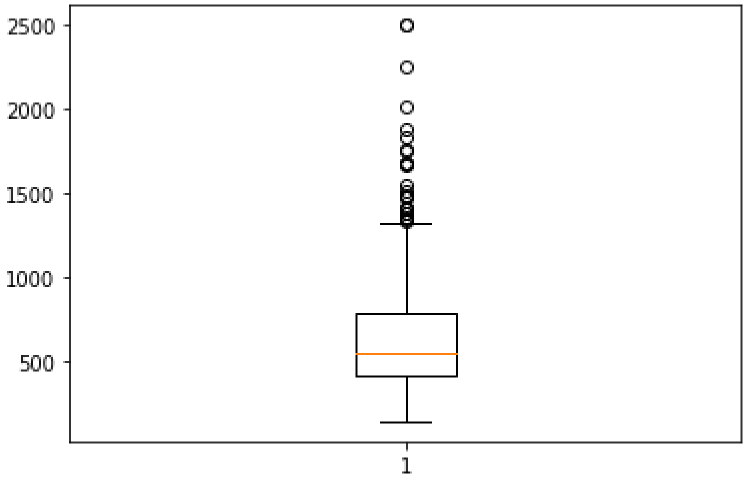
```
print("5 point summary and box plot for mean-perimeter")
print(df['mean perimeter'].describe(percentiles=[.25,.5,.75]))
p=plt.boxplot(df['mean perimeter'])
```

```
5 point summary and box plot for mean-perimeter
count    569.000000
mean      91.969033
std       24.298981
min       43.790000
25%       75.170000
50%       86.240000
75%      104.100000
max      188.500000
Name: mean perimeter, dtype: float64
```



```
print("5 point summary and box plot for mean-area")
print(df['mean area'].describe(percentiles=[.25,.5,.75]))
p=plt.boxplot(df['mean area'])
```

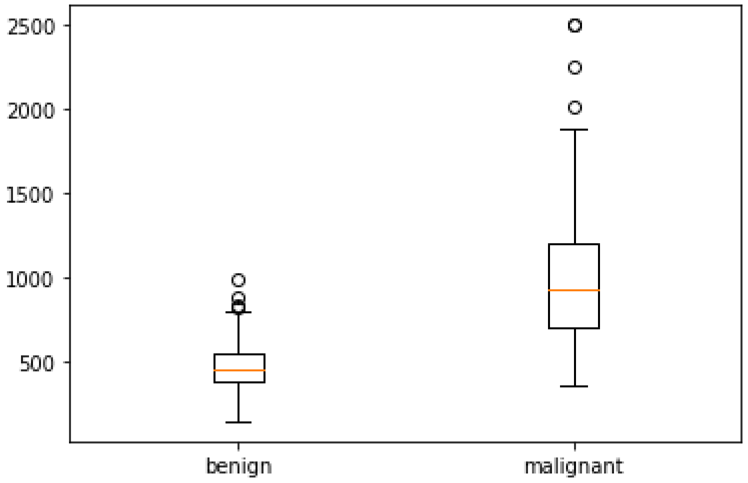
```
5 point summary and box plot for mean-area
count     569.000000
mean      654.889104
std       351.914129
min       143.500000
25%       420.300000
50%       551.100000
75%       782.700000
max      2501.000000
Name: mean area, dtype: float64
```



Compare attributes with respect to classes using boxplot

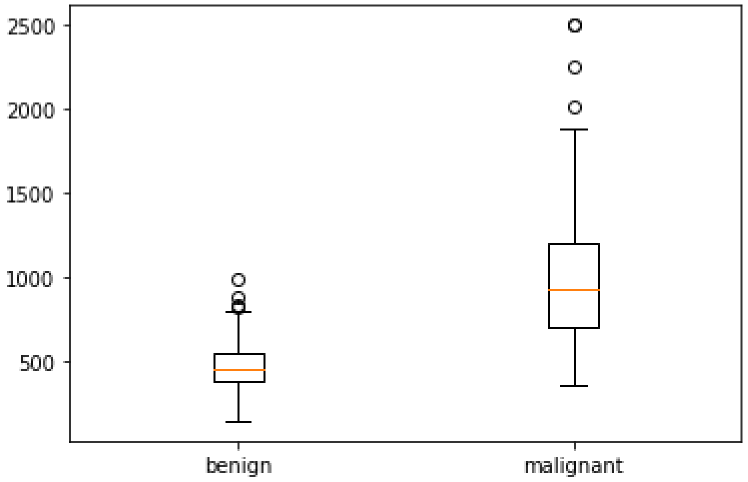
```
benign_radius=df[df['target']==1]['mean radius']
malig_radius=df[df['target']==0]['mean radius']
print("For mean radius:")
p=plt.boxplot([benign_area,malig_area],labels=['benign','malignant'])
```

```
For mean radius:
```



```
benign_radius=df[df['target']==1]['mean perimeter']
malig_radius=df[df['target']==0]['mean perimeter']
print("For mean perimeter:")
p=plt.boxplot([benign_area,malig_area],labels=['benign','malignant'])
```

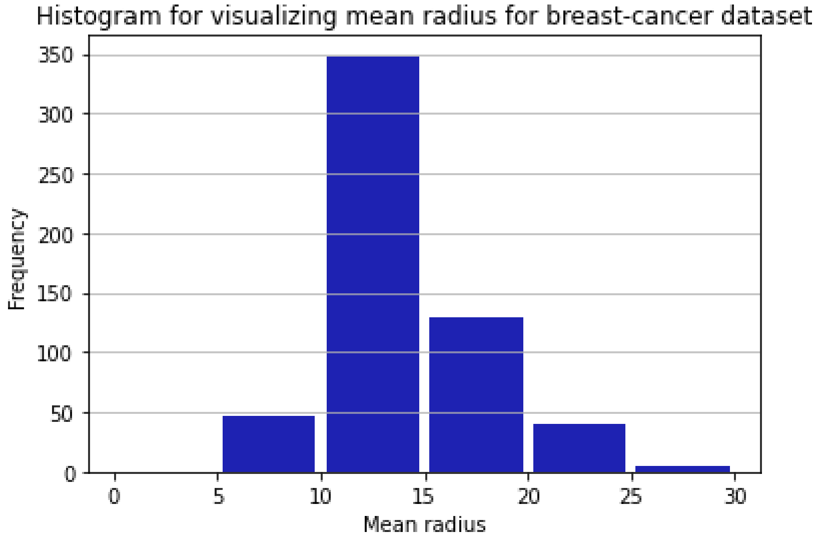
```
For mean perimeter:
```



Demonstrate histogram on numerical attributes

```
gaps=int(df['mean radius'].max()/5)
n, bins, patches = plt.hist(x=df['mean radius'], bins=range(0,int(df['mean radius'].max
                            alpha=0.9, rwidth=0.9)
plt.grid(axis='y', alpha=0.9)
plt.xlabel('Mean radius')
plt.ylabel('Frequency')
plt.title('Histogram for visualizing mean radius for breast-cancer dataset')
```

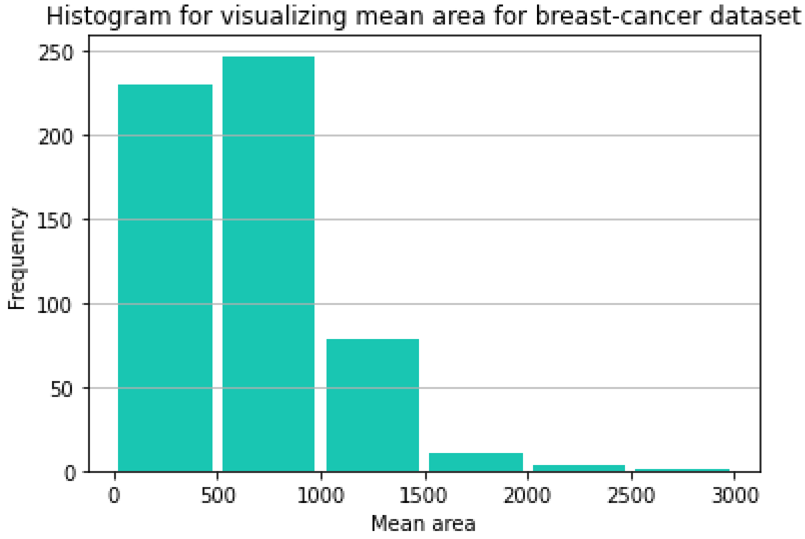
```
Text(0.5, 1.0, 'Histogram for visualizing mean radius for breast-cancer dataset')
```



- Most of the mean radius is between 10-15
- More than 70% of mean radius is less than 20

```
gaps=int(df['mean area'].max()/5)
n, bins, patches = plt.hist(x=df['mean area'], bins=range(0,int(df['mean area'].max())+
                            alpha=0.9, rwidth=0.9)
plt.grid(axis='y', alpha=0.9)
plt.xlabel('Mean area')
plt.ylabel('Frequency')
plt.title('Histogram for visualizing mean area for breast-cancer dataset')
```

```
Text(0.5, 1.0, 'Histogram for visualizing mean area for breast-cancer dataset')
```



- The mean area is max between 500-1000
- More than 80% of mean area is less than 1500