

### 3\_Percentage, Percentile and Quartile

```
In [1]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
In [2]: dataset = pd.read_csv('titanic.CSV')
```

```
In [3]: dataset.head(3)
```

```
Out[3]:
```

	Survived	Pclass	Name	Sex	Age	Siblings/Spouses Aboard	Parents/Children Aboard	Fare
0	0	3	Mr. Owen Harris Braund	male	22.0	1	0	7.2500
1	1	1	Mrs. John Bradley (Florence Briggs Thayer) Cum...	female	38.0	1	0	71.2833
2	1	3	Miss. Laina Heikkinen	female	26.0	0	0	7.9250

```
In [5]: dataset.isnull().sum()
```

```
Out[5]: Survived          0
Pclass          0
Name            0
Sex             0
Age             0
Siblings/Spouses Aboard  0
Parents/Children Aboard  0
Fare            0
dtype: int64
```

```
In [ ]: # So no null value is present in above data
```

```
In [7]: np.percentile(dataset['Age'], 25), np.percentile(dataset['Age'], 75)
```

```
Out[7]: (20.25, 38.0)
```

```
In [13]: np.percentile(dataset['Age'], 0), np.percentile(dataset['Age'], 100), np.percentile
```

```
Out[13]: (0.42, 80.0, 28.0)
```

```
In [14]: dataset['Age'].min(), dataset['Age'].max(), dataset['Age'].median()
```

```
Out[14]: (0.42, 80.0, 28.0)
```

```
In [16]: # So in above 2 rows, min. age account for 0% percentile and max. age accounts for  
# and median age is 50% percentile of age
```

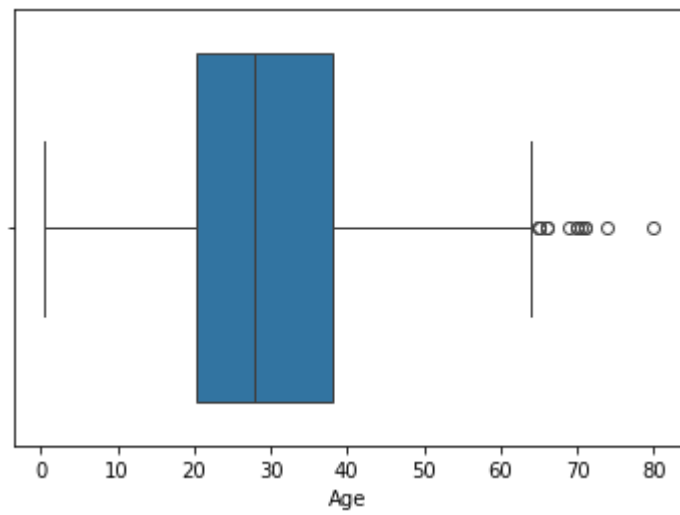
```
In [17]: dataset.describe()
```

```
Out[17]:
```

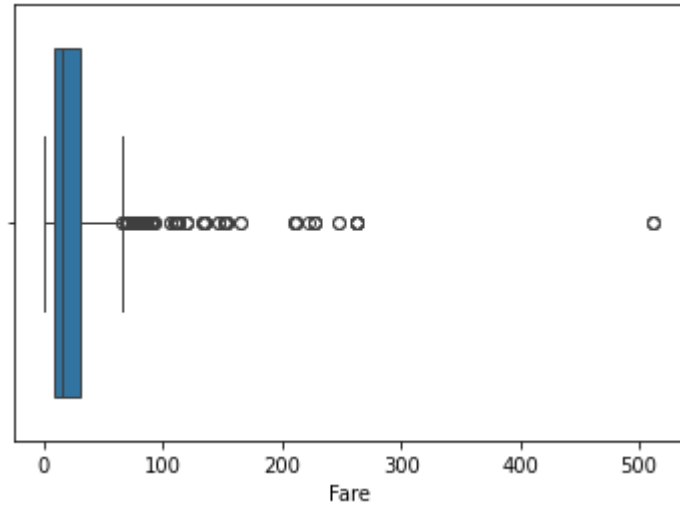
	Survived	Pclass	Age	Siblings/Spouses Aboard	Parents/Children Aboard	Fare
<b>count</b>	887.000000	887.000000	887.000000	887.000000	887.000000	887.000000
<b>mean</b>	0.385569	2.305524	29.471443	0.525366	0.383315	32.30542
<b>std</b>	0.487004	0.836662	14.121908	1.104669	0.807466	49.78204
<b>min</b>	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
<b>25%</b>	0.000000	2.000000	20.250000	0.000000	0.000000	7.92500
<b>50%</b>	0.000000	3.000000	28.000000	0.000000	0.000000	14.45420
<b>75%</b>	1.000000	3.000000	38.000000	1.000000	0.000000	31.13750
<b>max</b>	1.000000	3.000000	80.000000	8.000000	6.000000	512.32920

```
In [20]: #If you see closely on age you can see that  
# min(0%) : 0.42  
# Q1 : 25% : 20.25  
# Q2 : 50% : 28.00  
# Q3 : 75% : 38.00  
# Q4 : max(80%): 80.00  
# So you can see the huge difference between Q3 and Q4. So it is clear that outlier  
# Also difference between min (0%) and Q1 is significant larger, so there is also c  
# median (Q2) is 28, so it is evident that the median is inclined towards left side  
# So this whole analysis tell that there is definitely outlier present in this data
```

```
In [23]: # To show it in the boxplot  
sns.boxplot(x='Age', data=dataset)  
plt.show()
```



```
In [25]: # To show it in the boxplot
sns.boxplot(x='Fare', data=dataset)
plt.show()
```



In [ ]:

In [ ]: