

# 1\_Measure of Central Tendency

```
In [2]: import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Mean, Media, Mode are often used in data cleaning

```
In [3]: dataset = pd.read_csv("titanic.csv")
```

```
In [4]: dataset.head(3)
```

```
Out[4]:
```

	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 [42]: dataset["Age"]
```

```
Out[42]: 0      22.0
1      38.0
2      26.0
3      35.0
4      35.0
...
882    27.0
883    19.0
884     7.0
885    26.0
886    32.0
Name: Age, Length: 887, dtype: float64
```

## Find Median

To remove null values in age column

```
In [5]: # In order to see how many null entries are present in all columns
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 [6]: # There are no null values above, in case there are null values we can remove them
dataset["Age"].fillna(dataset["Age"].mean(), inplace=True)
```

```
In [7]: np.median(dataset["Age"])
```

```
Out[7]: 28.0
```

## Find Mean

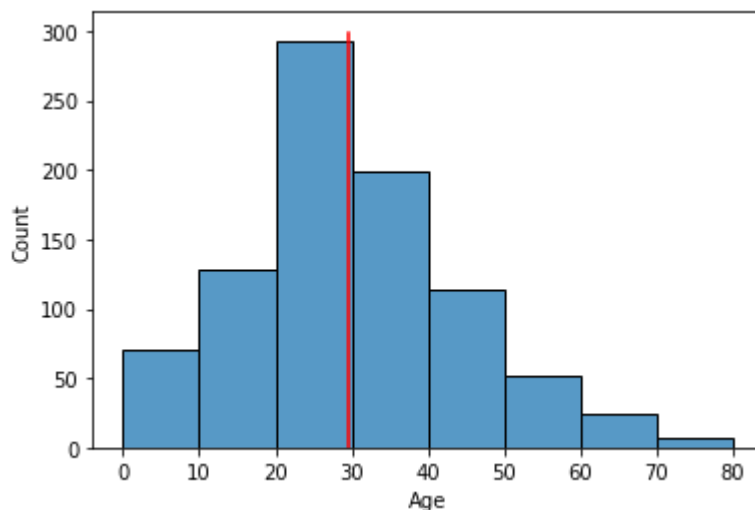
```
In [8]: dataset["Age"].mean()
```

```
Out[8]: 29.471443066516347
```

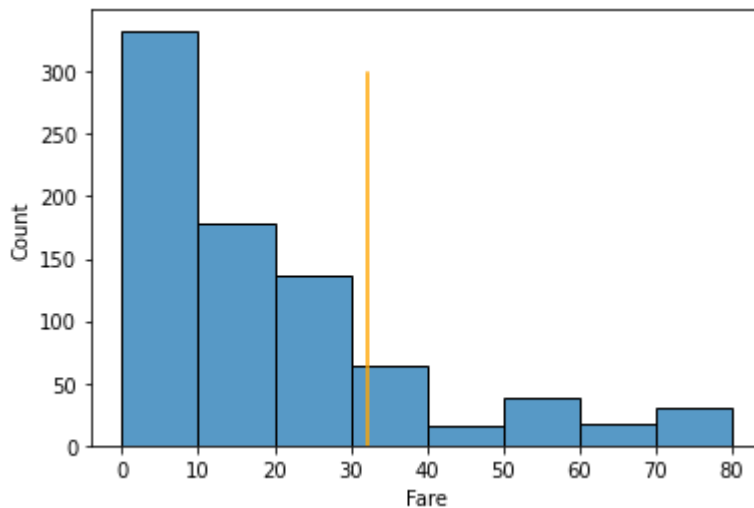
```
In [13]: mn = np.mean(dataset["Age"])
md = np.mean(dataset["Fare"])
md
```

```
Out[13]: 32.30542018038331
```

```
In [10]: sns.histplot(x="Age", data=dataset, bins= [i for i in range(0,81,10)])
plt.plot([mn for i in range(0,300)],[i for i in range(0,300)], c="red")
plt.show()
```



```
In [49]: sns.histplot(x="Fare", data=dataset, bins=[i for i in range(0,81,10)])
plt.plot([md for i in range(0,300)], [i for i in range(0,300)], c="orange")
plt.show()
```



## Finding Mode

```
In [21]: dataset["Fare"].mode()
```

```
Out[21]: 0    8.05
         Name: Fare, dtype: float64
```

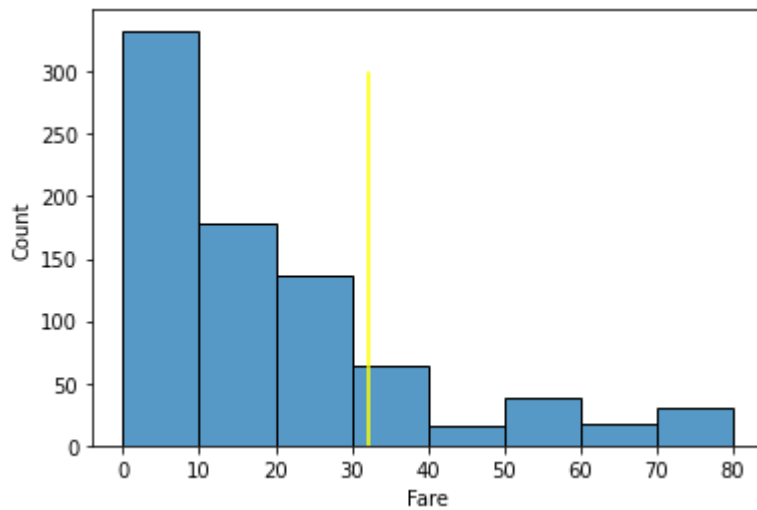
```
In [27]: mo = dataset["Fare"].mode()[0]
         mo
```

```
Out[27]: 8.05
```

```
In [28]: # To determine the frequency of fare
         dataset["Fare"].value_counts()
```

```
Out[28]: 8.0500    43
         13.0000   42
         7.8958   36
         7.7500   33
         26.0000  31
         ..
         35.0000   1
         28.5000   1
         6.2375   1
         14.0000   1
         10.5167   1
         Name: Fare, Length: 248, dtype: int64
```

```
In [48]: # to plot the mode of Fare ind dataset
         sns.histplot(x="Fare", data=dataset, bins=[i for i in range(0,81,10)])
         plt.plot([md for i in range(0,300)], [i for i in range(0,300)], c="yellow")
         plt.show()
```



To show all variables in one plot

```
In [56]: sns.histplot(x="Fare", data=dataset, bins=[i for i in range(0,81,10)])
plt.plot([mn for i in range(0,300)], [i for i in range(0,300)], c="red", label="Mean")
plt.plot([md for i in range(0,300)], [i for i in range(0,300)], c="green", label="Median")
plt.plot([mo for i in range(0,300)], [i for i in range(0,300)], c="blue", label="Mode")
plt.legend()
plt.show()
```

