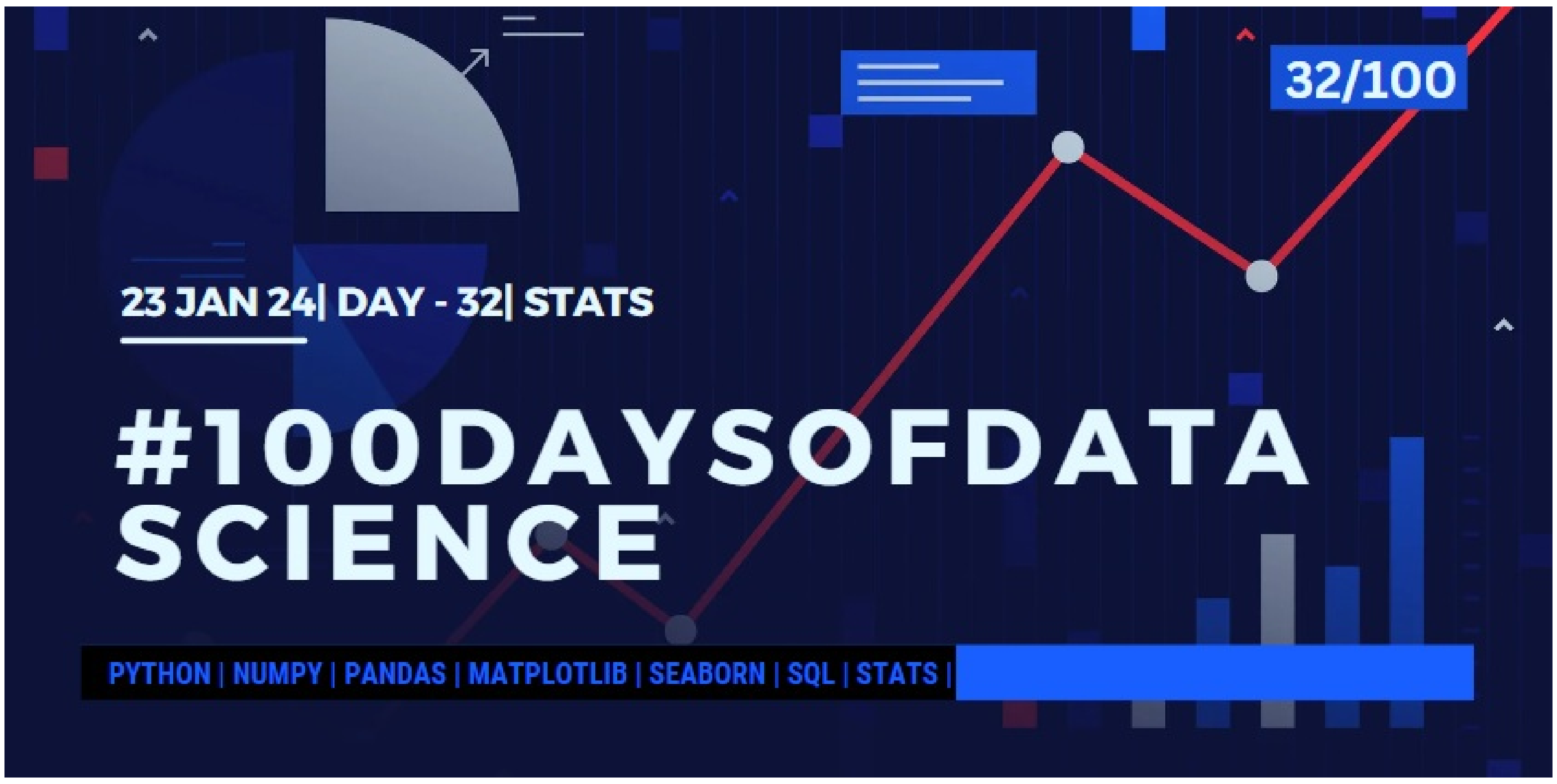
32/100
23 JAN 24| DAY - 32| STATS
#100DAYSOFDATA SCIENCE
PYTHON | NUMPY | PANDAS | MATPLOTLIB | SEABORN | SQL | STATS |

Statistics - 1 ▯

- Statistics is a branch of mathematics that deals with collecting, analyzing, interpreting, and visualizing empirical data.
- Statistics are of two types
  1. Descriptive Statistics
  2. Inferential Statistics
- We have 2 types of variable
  1. Quantitative/Numerical
  2. Qualitative/Categorical

1. Descriptive Statistics:

- Descriptive statistics is understanding, analyzing, summarizing the data in form of numbers and graphs. We analyze the data using different plots and charts on different kinds of data(numerical and categorical) like bar plot, pie chart, scatter plot, Histogram, etc.
- It is a method use to summarize & describe the data such that we can represent the entire population and sample by single value
- It invlove measure of central tendency(mean, median, mode), it invlove measure of dispersion(variablity)

2. Inferential Statistics

- We are extracting some sample of data from population data, and from that sample of data, we are inferencing something for population data.
- It is a process that allows us to not only test the hypothesis but also estimates value in population in short
- It is used to draw conclusion about the distribution
- It means we perform some tests on sample data and make a conclusion specific to that population. we use various techniques to drive conclusions including data visualization, manipulation, etc.

Types of Variable:

- There are 2 types of data we get as Numerical and Categorical which we need to handle and analyze.
- In Short if the data is in numerical form it will go under Quantitative variable and if it is in characters/obeject it wil go under categorical variable
- Quantitaive variable have two sub types:

  1. Discrete -
     - All the numerical parameter which can be counted eg. Population, no. of vehicle, bank balance, etc it is only integer
  2. Continous -
     - All the numerical parameter which can be measured eg. Height, Weight, Volume, Mass, etc it can be float

- Quatitaive variable have three sub types:
  1. Binomial -
     - It will have only two outcomes eg. head&tail, True&False, Win&Loss
  2. Nominal -
     - If we change the order of this outocme and its meaning doesn't changes it will be under nomial
  3. Odinal -
     - If we change the order of this outcome and its meaning changes it will go under odinal(That means order need to be maintained)

Population and Sample

- Population:
  - In statistics, the population comprises all observations (data points) about the subject under study.
  - An example of a population is studying the voters in an election. In the 2019 Lok Sabha elections, nearly 900 million voters were eligible to vote in 543 constituencies.
- Sample:

- In statistics, a sample is a subset of the population. It is a small portion of the total observed population.
- An example of a sample is analyzing the first-time voters for an opinion poll.

In [2]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import skew
from scipy.stats import kurtosis
```

In [3]:

```
Call_duration = pd.Series([23, 3, 13, 4, 45, 35, 48, 98, 65, 45, 75, 24, 15, 25, 34, 17, 16, 17, 19, 37, 46])
Call_duration
```

Out[3]:

```
0     23
1      3
2     13
3      4
4     45
5     35
6     48
7     98
8     65
9     45
10    75
11    24
12    15
13    25
14    34
15    17
16    16
17    17
18    19
19    37
20    46
dtype: int64
```

Measure of Central Tendency

- **Measures of central tendency are the measures that are used to describe the distribution of data using a single value. Mean, Median and Mode are the three measures of central tendency.**
- **Types:**
  1. **Mean**
  2. **Median**
  3. **Mode**

1. Mean:

- **Mean is only the average of all numbers in a particular numeric variable.**
- **When data contains outliers then finding mean and using it in any kind of manipulation is not suggested because a single outlier affects mean badly.**
- **So its solution is median.**

In [6]:

```
print("Mean of Call Duration: ",Call_duration.mean())
```

```
Mean of Call Duration:  33.523809523809526
```

2. Median:

- **The median is a centre value after sorting all the numbers.**
- **If the total number is even then it is the average of centre 2 values.**
- **It does not depend on or affected outliers till half of the data does not become outliers.**

In [12]:

```
print("Median of Call Duration: ",Call_duration.median())
```

```
Median of Call Duration:  25.0
```

3. Mode:

- **The mode is the observation (value) that occurs most frequently in the data set.**
- **There can be over one mode in a dataset.**

In [8]:

```
print("Mode of Call Duration: ",Call_duration.mode())
```

```
Mode of Call Duration:  0    17
1    45
dtype: int64
```

In [10]:

```
print("Getting first mode of Call Duration: ",Call_duration.mode()[0])
print("Getting second mode of Call Duration: ",Call_duration.mode()[1])
```

```
Getting first mode of Call Duration:  17
Getting second mode of Call Duration:  45
```

Measures of Spread

- **Measures of spread help to understand spreads of data means where your data is more spread (positive, negative, center)**
- **Types:**
  1. **Range**
  2. **Percentile**
  3. **Quartiles**
  4. **Interquartile Range(IQR)**
  5. **Mean Absolute Deviation**
  6. **Variance**
  7. **Standard Deviation**

1. Range:

- **The range describes the difference between the largest and smallest point in your data (max-min).**

In [11]:

```
maximum = Call_duration.max()
minimum = Call_duration.min()
ranges = maximum - minimum
print("Range of Call Duration: ", ranges)
```

```
Range of Call Duration:  95
```

2. Percentile:

- **A percentile is a measure used in statistics indicating the value below which a given percentage of observation in a group of observations falls.**

In [21]:

```
print('15% Data: ', Call_duration.quantile(0.15))
print("-" * 25)
print('37% Data: ', Call_duration.quantile(0.37))
print("-" * 25)
print('59% Data: ', Call_duration.quantile(0.59))
print("-" * 25)
print('73% Data: ', Call_duration.quantile(0.73))
print("-" * 25)
```

```
15% Data:  15.0
-------------------------
37% Data:  20.6
-------------------------
59% Data:  34.8
-------------------------
73% Data:  45.0
-------------------------
```

3. Quartiles:

- **Quartiles are the values that divide a list of numbers into quarters. the steps to find the quartile is.**
  1. **Put the list of numbers in order**
     A. **Then cut the list into 4 equal parts**
     B. **The quartiles are at the cuts**
- **Q2 is also known as the median and we can find the 4 quartiles by depicting the percentile value at 25, 50, 75, and 100.**

In [23]:

```
print('25% Data: ', Call_duration.quantile(0.25))
print("-" * 25)
print('Median/50% Data: ', Call_duration.median())
print("-" * 25)
print('75% Data: ', Call_duration.quantile(0.75))
print("-" * 25)
print('100% Data: ', Call_duration.quantile(1))
print("-" * 25)
```

```
25% Data:  17.0
-------------------------
Median/50% Data:  25.0
-------------------------
75% Data:  45.0
-------------------------
100% Data:  98.0
-------------------------
```

4. Interquartile Range(IQR):

- **It is a measure of dispersion between upper(75th) and lower(25th) Quartiles.**
- **It is a very important term in statistics that is used in most calculations and data preprocessing like dealing with outliers.**

In [26]:

```
Q1 =  Call_duration.quantile(0.25)
```

```
Q1 = Call_duration.quantile(0.25)
Q3 = Call_duration.quantile(0.75)
IQR = Q3 - Q1
print("IQR of Call Duration: ",IQR)
```

```
IQR of Call Duration:  28.0
```

5. Mean Absolute Deviation:

- **The absolute deviation from the mean, also called Mean absolute deviation(MAD), describes the variation in the data set.**
- **In simple words, it tells the average absolute distance of each point in the set.**

In [33]:

```
from scipy.stats import median_abs_deviation
median_absolute_deviation = median_abs_deviation(Call_duration)
print("Mean Absolute Deviation of Call Duration: ",median_absolute_deviation)
```

```
Mean Absolute Deviation of Call Duration:  12.0
```

6. Variance

- **Variance measure how far is data point is from the mean, only the difference from MAD and variance is we take square here.**
- **The variance is computed by finding the difference between each data point and mean, squaring them, summing them up, and take the average of all those numbers.**
- **The numpy has a direct function to calculate variance.**

In [34]:

```
variance = np.var(Call_duration)
print("Variance of Call Duration: ", variance)
```

```
Variance of Call Duration:  543.6780045351474
```

7. Standard Deviation:

- **Standard deviation in statistics is the square root of the variance.**
- **Variance and standard deviation represent the measures of fit, meaning how well the mean represents the data.**

In [36]:

```
std = np.std(Call_duration)
print("Standard Deviation of Call Duration: ", std)
```

```
Standard Deviation of Call Duration:  23.316903836812198
```

In [ ]:

```

```