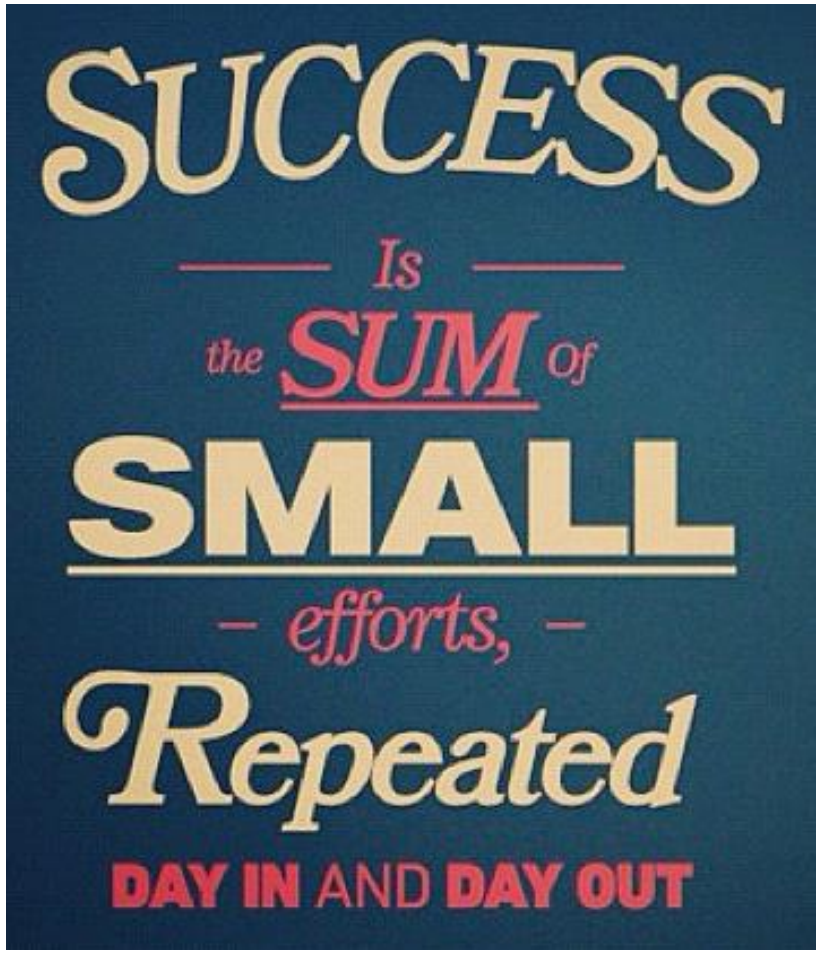
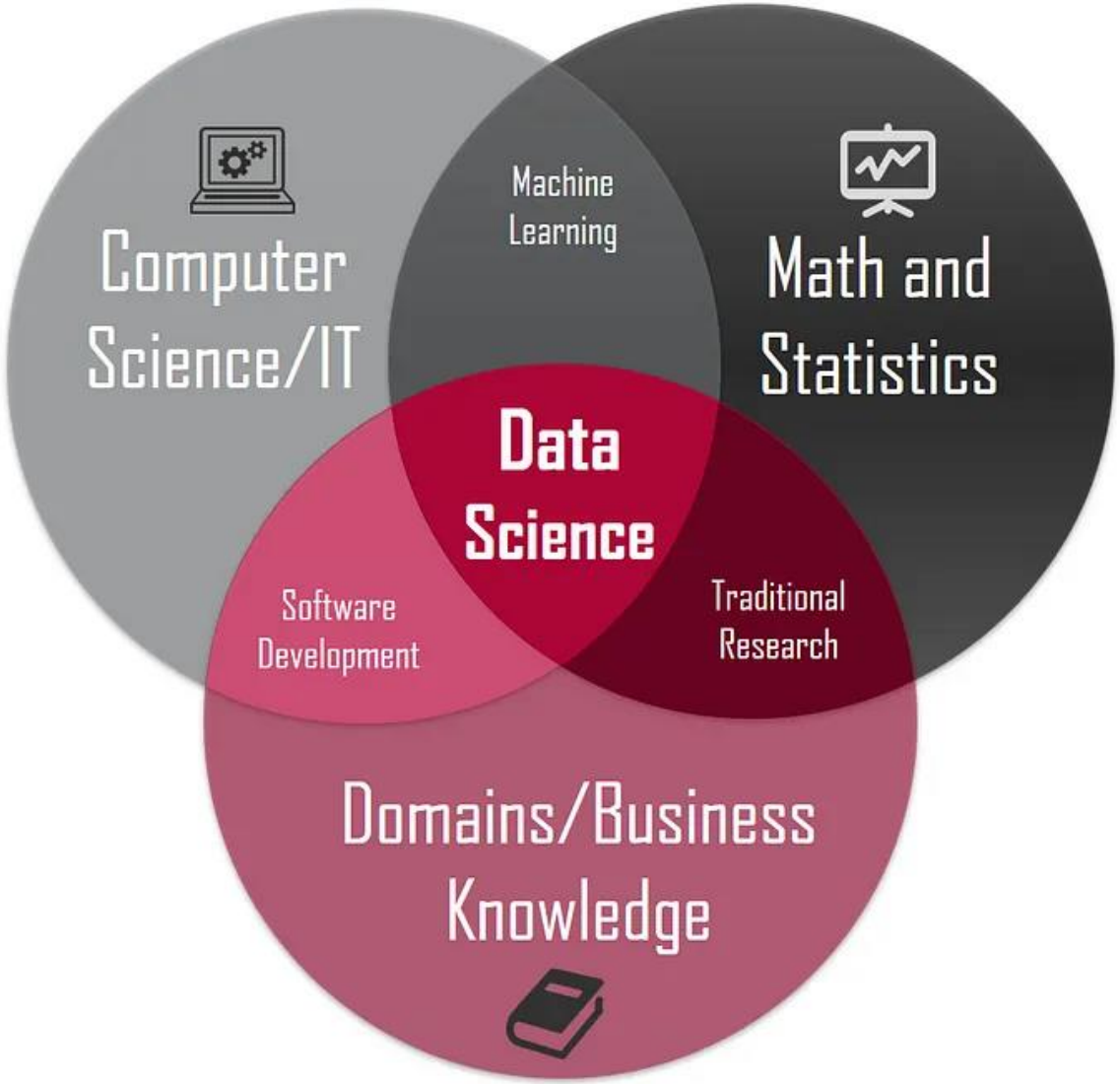

the MORE
YOU PRACTICE
THE BETTER
YOU GET


Practice
Habit
Practice
Practice


SUCCESS
Is
the SUM of
SMALL
efforts,
Repeated
DAY IN AND DAY OUT


Computer Science/IT
Machine Learning
Math and Statistics
Data Science
Software Development
Traditional Research
Domains/Business Knowledge

# 4 Pillars of Data Science

## COMPUTER SCIENCE

- Develops algorithms and data structures
- Works with databases and cloud computing
- Implements machine learning and AI models
- Optimizes code for efficiency and scalability

***Tools**: Python, SQL, Git, Power Apps, Power Automate*

## COMMUNICATION & VISUALIZATION

- Converts complex data into clear insights
- Builds reports, dashboards, and presentations
- Uses charts, graphs, and storytelling techniques
- Bridges the gap between data and business

***Tools**: Tableau, Power BI, Matplotlib*

## MATHEMATICS & STATISTICS

- Applies probability and statistical methods
- Uses linear algebra and calculus for ML
- Conducts hypothesis testing and A/B testing
- Ensures accurate data analysis and interpretation

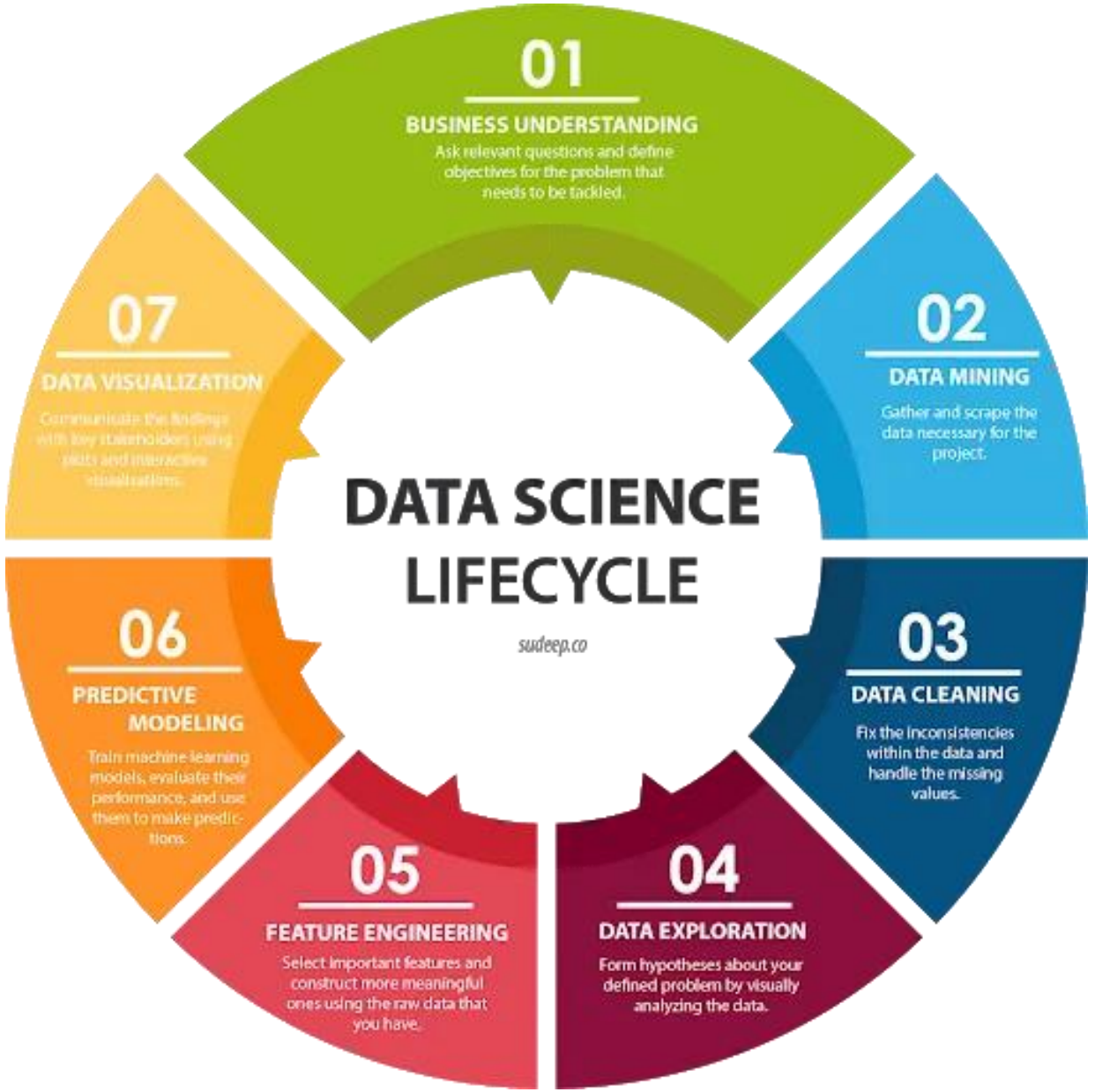
***Tools**: NumPy, SciPy, R*

## DOMAIN KNOWLEDGE

- Understands industry-specific problems
- Translates data insights into business impact
- Aligns models with real-world applications
- Supports strategic decision-making

***Tools**: SAP, Tally, Google Analytics, Salesforce*

1. Business Understanding
2. Data Mining
3. Data Cleaning
4. Data Exploration
5. Feature Engineering
6. Modeling
7. Visualization/ Presentation



# PYTHON FOR DATA SCIENCE
## MODULE 1 – EXPLORING DATA

PGD-ERP (DS)
SJCC

# PYTHON FOR DATA SCIENCE – MODULE 1
## EXPLORING DATA – 5 QUESTION METHODOLOGY

| DESCRIBE IT | QUANTIFY IT | DETAIL IT | PICTURE IT | ANALYZE IT |
|---|---|---|---|---|
| What data did I give you? | How much data did I give you ? | Tell me some specifics ? | What did you observe in the data ? | What can I do with this data ? |
| *Describe in a sentence Geography, Measure, Time, Product* | *Rows Columns File size Table size* | *Data types Missing values Value Counts* | *Top level observations (using charts, graphs)* | *Calculate Measures, Predictive analysis, Build dashboards* |

# PYTHON FOR DATA SCIENCE – MODULE 1
## EXPLORING DATA

Understanding Characteristics of the dataset –

- Size, Shape
- Value counts
- Data Types and change data types
- Treating missing values

# PYTHON FOR DATA SCIENCE

—

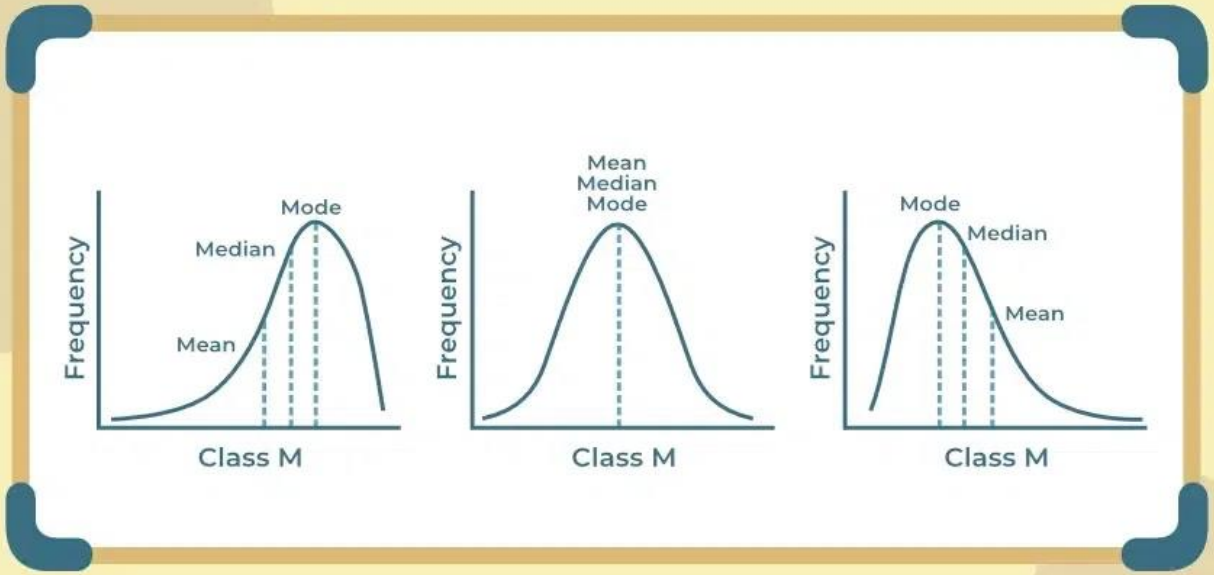
# MODULE 2 & 3

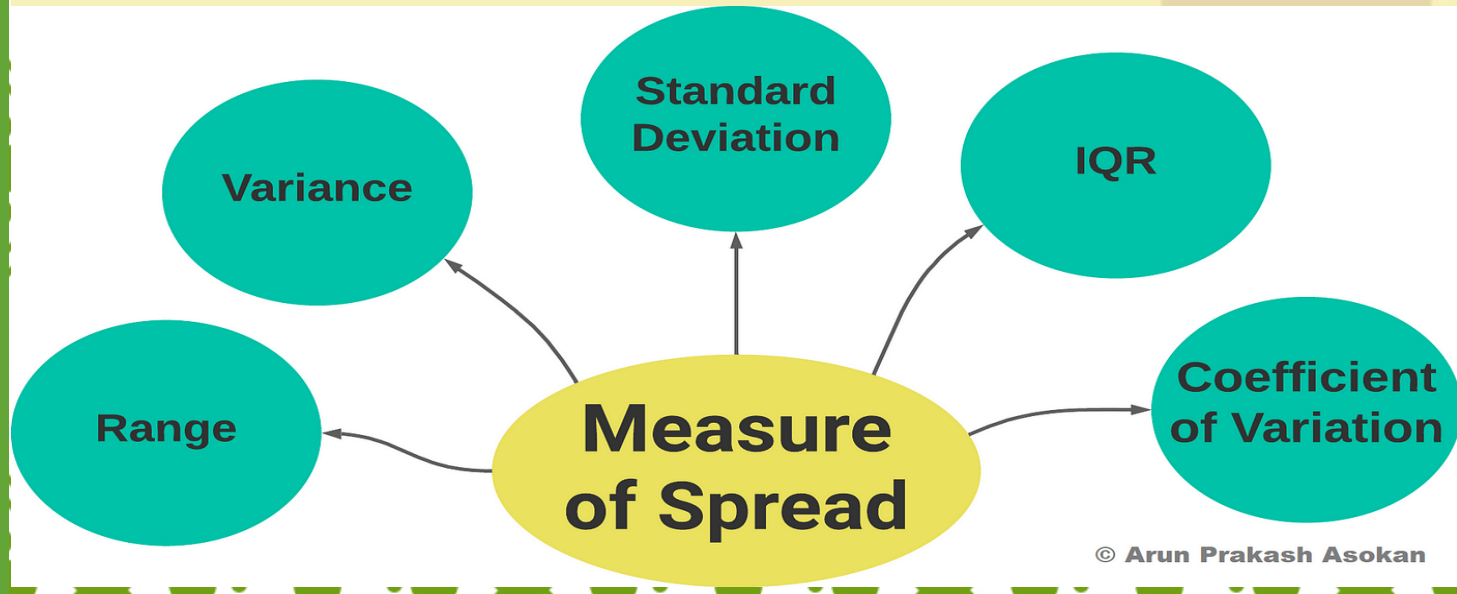
**Measures of Central Tendency**

Mean, Median, Mode

**Measures of Dispersion**

Range, Variance, Quartiles, IQR
Standard Deviation, Coefficient of Variation



# PYTHON FOR DATA SCIENCE – MODULE 2 & 3

## MEASURES OF CENTRAL TENDENCY – MEAN (AVERAGE)

- There are 3 means: **Arithmetic (balance point)**, Geometric, Harmonic

**DEFINITION: The mean $\bar{x}$**

To find the **mean** $\bar{x}$ (pronounced "x-bar") of a set of observations, add their values and divide by the number of observations. If the $n$ observations are $x_1, x_2, \ldots, x_n$, their mean is

$$\bar{x} = \frac{\text{sum of observations}}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

or, in more compact notation,

$$\bar{x} = \frac{\sum x_i}{n}$$

- The mean tells us how large each data value would be if the total were split equally among all the observations

# PYTHON FOR DATA SCIENCE – MODULE 2 & 3

## MEASURES OF CENTRAL TENDENCY – MEDIAN

- The **median is the mid-point** of a distribution, the number such that about half the observations are smaller and about half are larger
- To find the median of a distribution:
  1. Arrange all observations in order of size, from smallest to largest.
  2. If the number of observations n is odd, the median is the center observation in the ordered list.
  3. If the number of observations n is even, the median is the average of the two center observations in the ordered list.

# PYTHON FOR DATA SCIENCE – MODULE 2 & 3

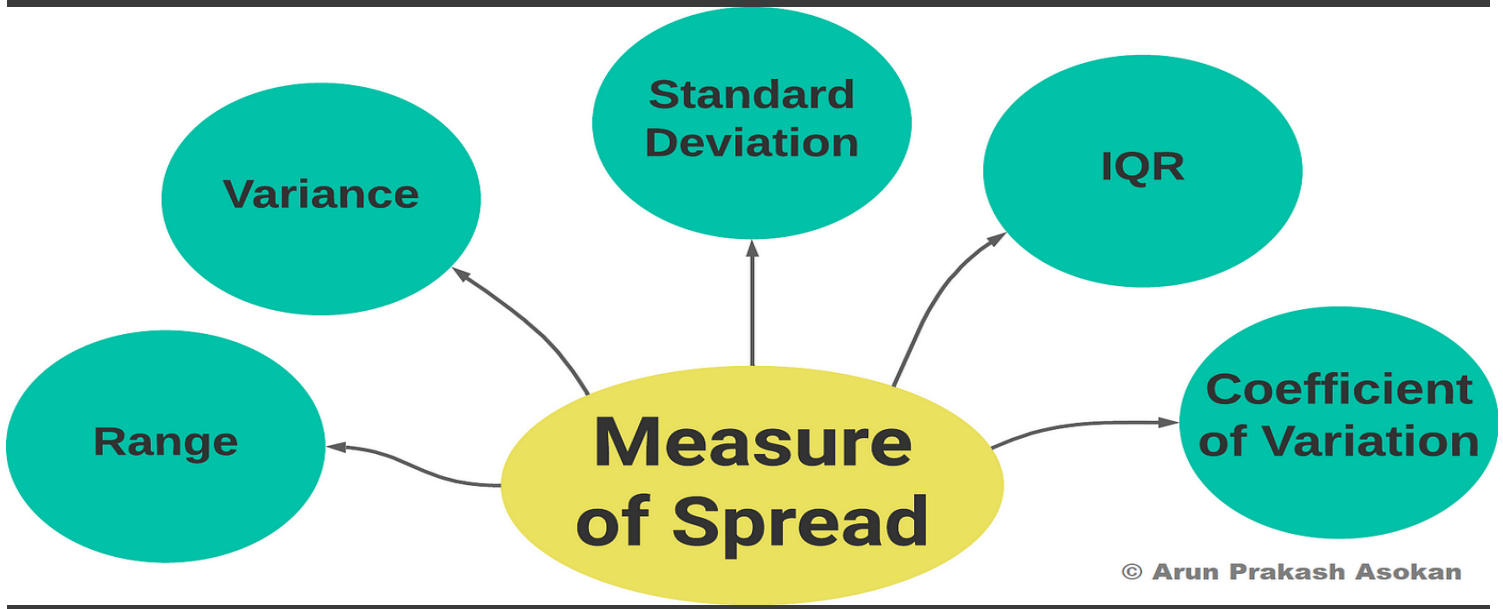
## MEASURES OF CENTRAL TENDENCE – MEAN vs MEDIAN

- The Mean is **affected** by extreme values, while Median is **resistant**
- The mean and median of a roughly symmetric distribution are close together
- If the distribution is exactly symmetric, the mean and median are exactly the same
- In a skewed distribution, the mean is usually farther out in the long tail than is the median
- *College fees, home prices, and salaries are all skewed, so here it is better to use Median*

# PYTHON FOR DATA SCIENCE – MODULE 2 & 3

## MEASURES OF CENTRAL TENDENCE – MODE

- Most occurring observation
- *Used when we have shirt sizes, footwear sizes etc.,*




# MEASURES OF DISPERSION

# PYTHON FOR DATA SCIENCE – MODULE 2 & 3

## MEASURES OF DISPERSION – RANGE

- the smallest observation (Min)
- the largest observation (Max)
- Range = Max – Min

# PYTHON FOR DATA SCIENCE – MODULE 2 & 3

## MEASURES OF DISPERSION – QUARTILES

- The first quartile **Q1** lies one-quarter of the way up the list
- The second quartile is the **median**, which is halfway up the list
- The third quartile **Q3** lies three-quarters of the way up the list
- The interquartile range (IQR) measures the range of the middle 50% of the data
- The interquartile range **(IQR) is defined as IQR = Q3 – Q1**
- *Be careful in locating the quartiles when several observations take the same numerical value. Write down all the observations, arrange them in order*
- **Outlier or Special Case** – Call an observation an outlier if it falls more than 1.5 × IQR above the third quartile or below the first quartile
  - **Q1 – 1.5 × IQR**
  - **Q3 + 1.5 × IQR**

# PYTHON FOR DATA SCIENCE – MODULE 2 & 3

## MEASURES OF DISPERSION – SD, VARIANCE, CO OF VAR

- The **standard deviation** sx measures the typical distance of the values in a distribution from the mean
- sx is always greater than or equal to 0. sx = 0 only when there is no variability
- This happens only when all observations have same value. Otherwise, sx > 0
- *As the observations become more spread out about their mean, sx gets larger*
- This average squared deviation is called the **variance**
- The **coefficient of variation (CV)** is the ratio of the standard deviation to the mean. The higher the coefficient of variation, the greater the level of dispersion around the mean. It is generally expressed as a percentage

# PYTHON FOR DATA SCIENCE – MODULE 2 & 3

## MEASURES OF DISPERSION – SD, VARIANCE, CO OF VAR

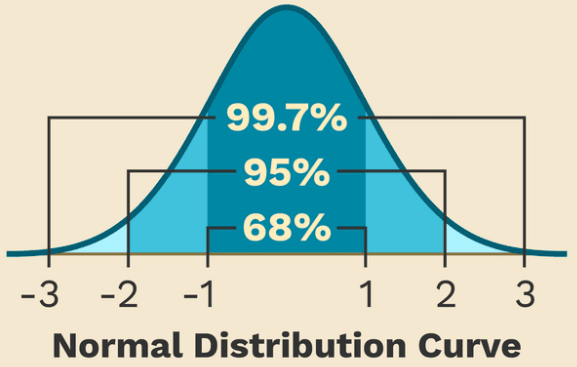
- The **standard deviation** sx measures the typical distance of the values in a distribution from the mean

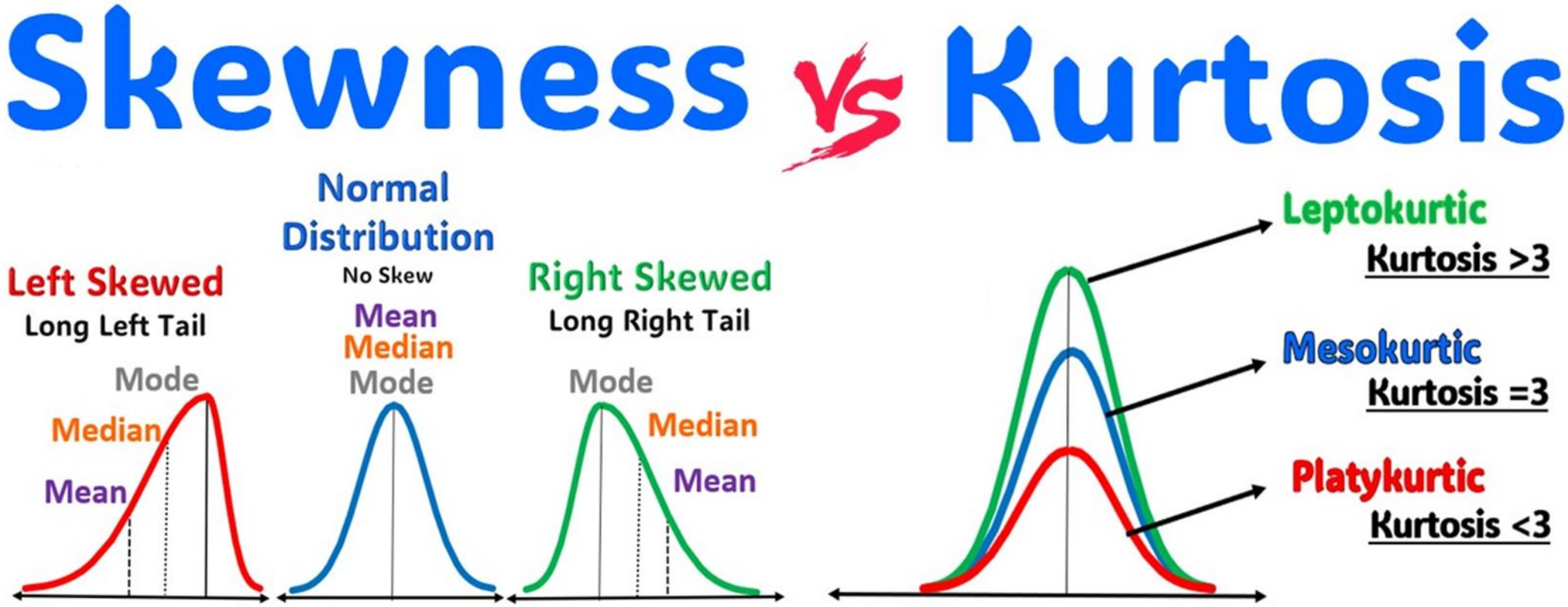
$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}}$$

$n$ = The number of data points

$x_i$ = Each of the values of the data

$\bar{x}$ = The mean of $x_i$



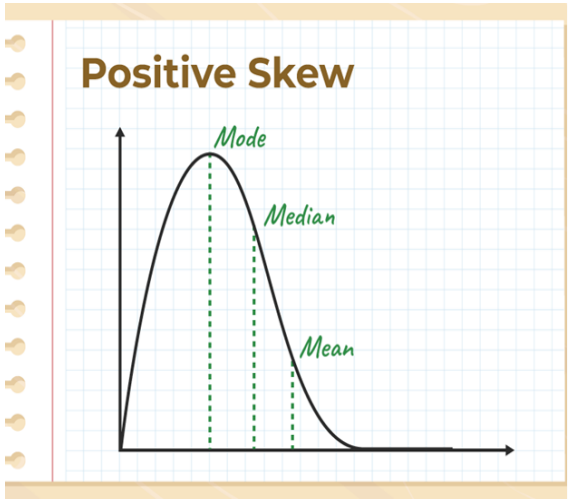


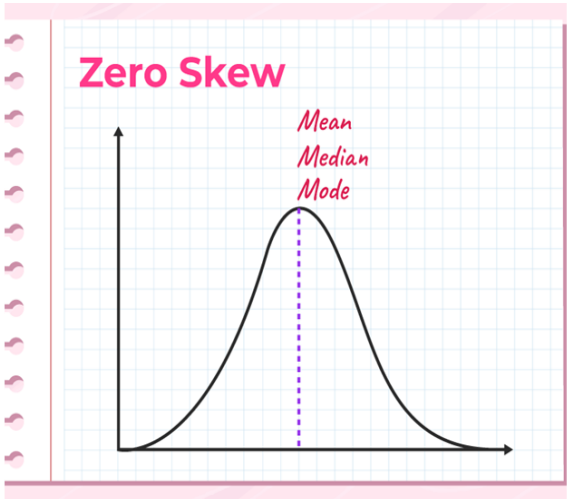

# MEASURES OF SHAPE

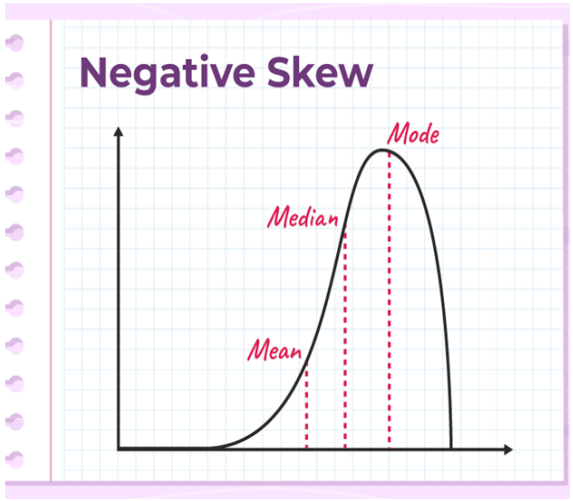
# PYTHON FOR DATA SCIENCE – MODULE 2 & 3

## MEASURES OF SHAPE – SKEWNESS

- The **Skewness** is the degree of asymmetry observed in a distribution on a bell curve to the left and right sides of the median
- Distributions can be positive and right-skewed, or negative and left-skewed. A normal distribution exhibits zero skewness
- *Skewness = 3(Mean – Median)/S.D*




· *If the skewness is between -0.5 and 0.5, the data are fairly symmetrical*

· *If the skewness is between -1 and -0.5 or between 0.5 and 1, the data are moderately skewed*

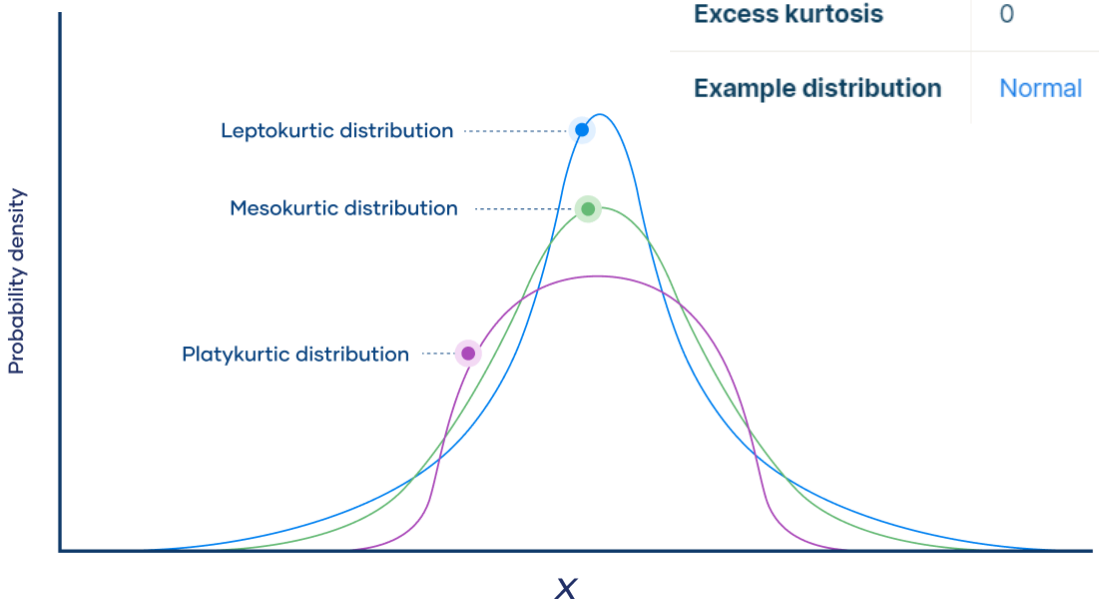
· *If the skewness is less than -1 or greater than 1, the data are highly skewed*

# PYTHON FOR DATA SCIENCE – MODULE 2 & 3

## MEASURES OF DISPERSION – SKEWNESS & KURTOSIS

- Statistical measure **Kurtosis** used to describe characteristic of a dataset.
- When normally distributed data is plotted on a graph, the plotted data that are farthest from the mean of the data usually form the tails on each side of the curve
- *Kurtosis indicates how much data resides in the tails or its peakedness*

| | Category | | |
|---|---|---|---|
| | **Mesokurtic** | **Platykurtic** | **Leptokurtic** |
| **Tailedness** | Medium-tailed | Thin-tailed | Fat-tailed |
| **Outlier frequency** | Medium | Low | High |
| **Kurtosis** | Moderate (3) | Low (< 3) | High (> 3) |
| **Excess kurtosis** | 0 | Negative | Positive |
| **Example distribution** | Normal | Uniform | Laplace |



# DATA SENT – CARS

DO NOT USE .DESCRIBE


***Load data, perform shape, dtypes etc.,***

1. Insert a column & calculate Length to Wheelbase ratio
2. Insert a column & calculate difference in milage between city & highway
3. Create 2 dataframes like this:

| MAKE | Avg of MPG_Highway | Avg of MPG_City | Avg of Milage diff |
|---|---|---|---|
| make1 | X | X | X |
| Make2 | X | X | X |
| Make3 | X | X | X |
| . | X | X | X |
| . | X | X | X |
| . | X | X | X |

| Type | Avg of MPG_Highway | Avg of MPG_City | Avg of Milage diff |
|---|---|---|---|
| Type1 | X | X | X |
| Type2 | X | X | X |
| Type3 | X | X | X |
| . | X | X | X |
| . | X | X | X |
| . | X | X | X |

***Using the Function Method***

1. For MSRP & Weight
   a. Calculate all Measures of Spread
   b. Calculate all measures of Central Tendency
   c. Calculate all Measures of Shape
2. Make 5 observations for each MSRP & Weight + from 1, 3 above

# NEW DATA SENT – TRAVEL TIME

## *Using the Function Method*

1. Calculate all the measures of Central Tendency
2. Calculate all the Measures of Spread
3. For Travel Time which is a better measure to use – Mean, Median or Mode…. Why ? Give reasons
4. After you have all the calculations, make 5 actionable observations

# NEW DATA SENT – TIPS

***Load data, perform shape, dtypes etc.,***

1. Insert a column & calculate tips as a percentage of Bill
2. Who Tips better Male or Female ?
3. What is the Average Bill per person ?

***Using the Function Method***

1. For Total Bill by Day & Tips by Day
   a. Calculate all the measures of Central Tendency
   b. Calculate all the Measures of Spread
2. Make 5 observations for each Total Bill and Tips