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,

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

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

MEAN

Sum of observation/n

MEDIAN

- The median is the mid-point of a distribution,

1. Arrange all observations in order of size, from smallest to largest if n= odd, the median is the center observation in the ordered list. In n= even, the median is the average of the two center observations in the ordered list

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 that is the median
- College fees, home prices, and salaries are all skewed, so here it is better to use Median

MODE

- Most occurring observation
- shirt sizes, footwear sizes etc.

RANGE

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

QUARTILE

- 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

SD,VAR, COEFF

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

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**

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 peaked ness**

CORRELATION

- ▪ Correlation describes the strength of an association between two variables
- ▪ It is completely symmetrical, correlation between A & B is same as correlation between B & A
- ▪ Two variables are linearly related (meaning they change together at a constant rate)
- ▪ The sample correlation coefficient, r, quantifies the strength of the relationship. Correlations are also tested for statistical significance
- ▪ Correlation can't look at the presence or effect of other variables outside of the two being explored
- ▪ Importantly, correlation doesn't tell us about cause and effect
- ▪ Correlation also cannot accurately describe curvilinear relationships
- ▪ It is a unit-free measure called the correlation coefficient which ranges from -1 to +1 and is denoted by r
- ▪ The closer r is to zero, the weaker the linear relationship
- ▪ Positive r values indicate a positive correlation, where the values of both variables tend to increase together
- ▪ Negative r values indicate a negative correlation, where the values of one variable tend to increase when the values of the other variable decrease

SQC

- ▪ Statistical Quality Control the predecessor of Total Quality Management still continues to exert its influence in the quality management of corporations
- ▪ SQC is about employing inspection methodologies derived from statistical sampling theory to ensure conformance to requirements (Nicholas, 1998)
- ▪ Statistical Quality Control (SQC) is a methodology used to monitor and control the quality of products or services
- ▪ Statistical Quality Control aims to identify and eliminate defects or variations in production processes, improving product or service quality and reducing waste
- ▪ SQC can help identify patterns and trends that can be used to make data-driven decisions that improve overall quality and productivity
- ▪ SQC is widely used across many industries, including manufacturing, healthcare, and service industries

- ▪ Processes have some degree of inherent variability
- ▪ Process variability can be classified:
  - ▪ that caused by common sources (also referred to as chance causes)
  - ▪ that caused by special sources (also known as assignable causes)
- ▪ The common faults are those caused by problems with the processing system itself
- ▪ The special factors are usually unpredictable and are disturbances to 'routine' operation

SQC Tools and Techniques

- ▪ Process flowcharts
- ▪ Check sheets
- ▪ Pareto diagrams
- ▪ Histograms
- ▪ Cause-and-Effect diagrams
- ▪ Scatter diagrams
- ▪ Control charts

➢ In statistics, Control charts are tools to determine whether a process is in a controlled statistical state. They are also known as Shewhart charts or process-behavior charts.

➢ The data is plotted in a timely order

➢ It is bound to have a central line of average, an upper line of upper control limit and a lower line of lower control limit.