EDA OVERVIEW

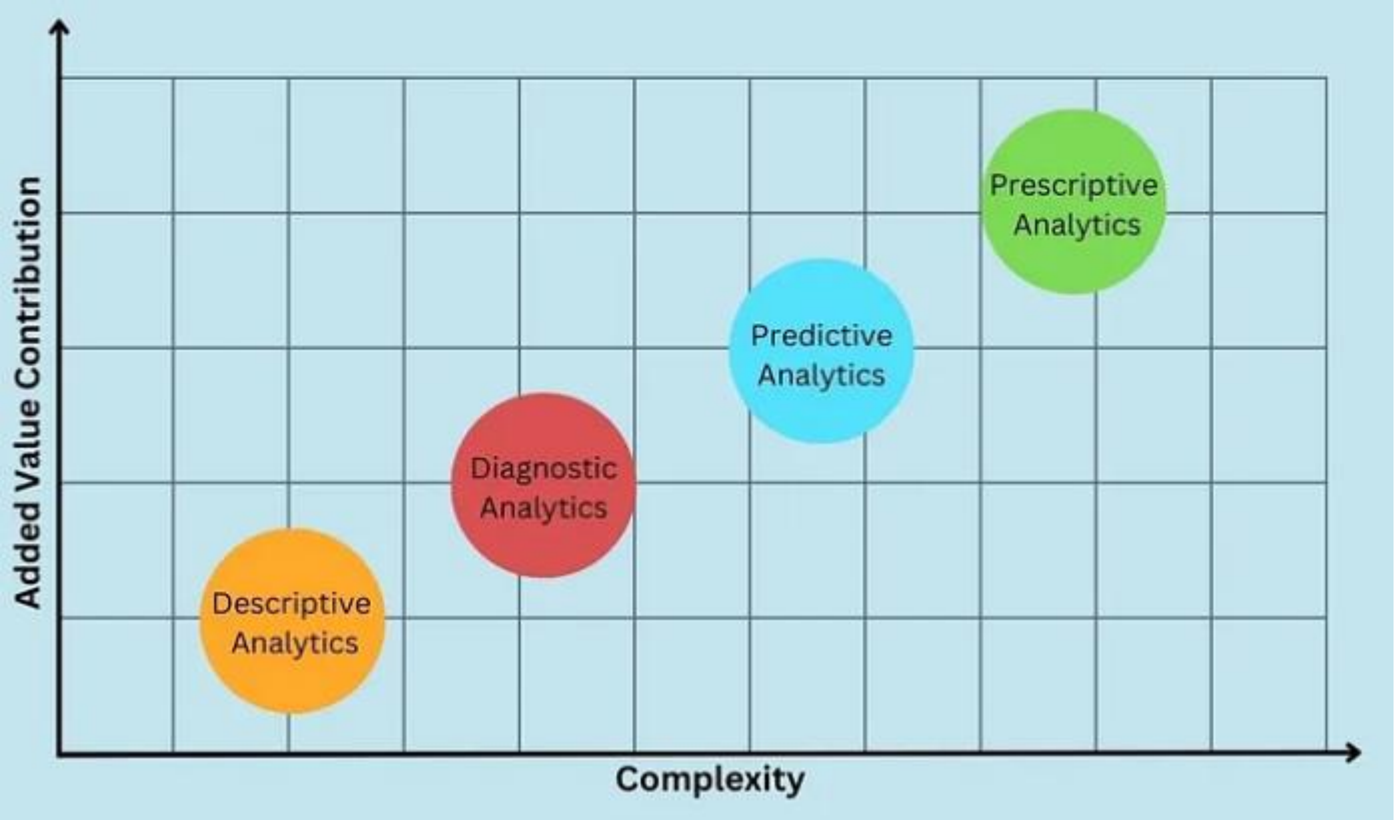
Sharath Srivatsa

TYPES OF ANALYTICS

Descriptive Analytics

- We can consider this type of data analysis as the explainer of data

- Talking in deep descriptive data analysis tells about what happened in the past and usually combines dashboards and graphs with it



Diagnostic Analytics

- After performing the descriptive analysis and getting to know what happened, one must look forward to seeking why it happened

- We perform diagnostic data analysis to find the cause of output from the descriptive analysis

Predictive Analytics

- After knowing and understanding what happened and the root cause of what happened, one needs to answer the question of what is likely to happen.

- Predictive analysis is used to predict future outcomes using the previous data

Prescriptive Analytics

- Here, the final type of data analysis is prescriptive data analysis, which goes beyond descriptive and predictive analysis by recommending a course of action based on the analysis results

EXAMPLES

Descriptive Analytics **Examples:**

○ A company tracks its website traffic over time to identify peak visiting hours and popular pages.

○ A sales report shows the total revenue, number of units sold, and best-selling products for the last quarter.

○ A marketing team analyzes social media engagement metrics (likes, shares, comments) to see which posts resonated most with the audience.

Diagnostic Analytics **Examples:**

○ A company investigates a sudden drop in sales by examining factors like changes in marketing campaigns, competitor activity, or customer reviews.

○ A manufacturer analyzes production data to identify the root cause of a recent increase in product defects.

○ A hospital analyzes patient data to determine factors contributing to higher readmission rates.

EXAMPLES

Predictive Analytics Examples:

○ A retailer forecasts future demand for a product based on past sales trends, seasonality, and economic indicators.

○ A bank uses credit scoring models to predict the likelihood of a customer defaulting on a loan.

○ A streaming service recommends movies or shows to users based on their viewing history and preferences.

Prescriptive Analytics Examples:

○ A transportation company uses route optimization software to determine the most efficient delivery routes, minimizing fuel consumption and delivery time.

○ A marketing team uses A/B testing to identify the most effective email subject line for a marketing campaign.

○ A doctor uses a clinical decision support system to determine the best treatment plan for a patient based on their specific condition and medical history.

PRIVATE AND PUBLIC DATA

Feature	Private Data	Public Data
Identifiability	Relates to an identified or identifiable individual	Does not identify individuals
Sensitivity	Often highly sensitive	Usually non-sensitive or anonymized
Access	Restricted	Open and freely available
Regulations	Strict data protection laws apply	Fewer restrictions
Examples	Medical records, financial data	Census data, weather data

Law	Region	Law	Region
GDPR	***European Union***	***PDPA***	***Singapore***
CPRA	***California, USA***	***PIPA***	***South Korea***
PIPEDA	***Canada***	***LGPD***	***Brazil***
PIPL	***China***	***Privacy Act 1988***	***Australia***
		POPIA	***South Africa***

DATA CLEANING

Simple Imputation Methods

- Mean/Median/Mode Imputation
- Constant Value Imputation
- Last Observation Carried Forward (LOCF) / Next Observation Carried Backward (NOCB)

Model-Based Imputation Methods

- ***Regression Imputation:*** assumes a linear relationship between variables.
- ***K-Nearest Neighbors (KNN) Imputation:*** Effective for complex relationships but computationally expensive for large datasets.

Advanced Imputation Methods

Deep Learning Imputation:

- Utilizing deep learning models like autoencoders or generative adversarial networks (GANs) to learn complex patterns in the data and generate imputations.
- Can be highly accurate but requires large datasets and computational resources.

MICE (Multivariate Imputation by Chained Equations): Suitable for datasets with multiple missing variables and complex relationships.

OUTLIER DETECTION

Z-score

Measures how many standard deviations a data point is from the mean. A Z-score with an absolute value greater than 3 is often considered an outlier.

Pros: Simple to calculate, widely applicable.

Cons: Sensitive to extreme values, assumes a normal distribution.

Interquartile Range (IQR)

Calculates the range between the first quartile (Q1) and third quartile (Q3) of the data. Outliers are defined as values falling below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR.

Pros: Robust to extreme values, doesn't assume a normal distribution.

Cons: May not be suitable for small datasets.

CENTRAL TENDENCY AND VARIABILITY MEASURES

Measures of Central Tendency: These tell you where the "center" of your data lies.

- Mean: The average value.
- Median: The middle value when the data is sorted.
- Mode: The most frequent value.

Measures of Variability (Spread): These describe how spread out the data is.

- Range: The difference between the maximum and minimum values.
- Variance: The average of the squared differences from the mean.
- Standard Deviation: The square root of the variance.

Interquartile Range (IQR): The range between the 25th and 75th percentiles.

- Frequency Distribution: Shows how often different values occur in the data.
- Histograms: A graphical representation of the frequency distribution.

Shape of the Distribution:

- Skewness: Measures the asymmetry of the data distribution.
- Kurtosis: Measures the "peakedness" of the data distribution.

EXPLORING CATEGORICAL VARIABLES

Univariate Analysis

- **Frequency Distribution:** Calculate the count or percentage of observations for each category.
- **Bar Charts:** Visualize the frequency distribution using rectangular bars.
- **Pie Charts:** Represent proportions of categories within a whole (best for a small number of categories).
- **Mode:** Identify the most frequent category.

Bivariate Analysis

- **Crosstabulation:** Analyse the relationship between two categorical variables by creating a contingency table.
- **Stacked Bar Charts:** Compare the distribution of one categorical variable across different categories of another.
- **Grouped Bar Charts:** Visualize the counts or proportions of each category for different groups.
- **Chi-Square Test:** Determine if there's a significant association between two categorical variables

Multivariate Analysis

- **Cluster Analysis:** Group similar categories based on their characteristics.
- **Correspondence Analysis:** Explore relationships between categorical variables in a multidimensional space similar to PCA for continuous variables

Additional Considerations

- **Data Cleaning:** Handle missing values and inconsistencies in categorical data.
- **Data Transformation:** Consider creating new categorical variables or combining existing ones.
- **Visualization Tools:** Utilize libraries like Matplotlib, Seaborn, or Plotly for effective visualizations.

EXPLORING CONTINUOUS VARIABLES

Univariate Analysis

- **Summary Statistics:** Calculate measures like mean, median, mode, standard deviation, quartiles, minimum, and maximum to understand the data's central tendency, dispersion, and shape.
- **Histograms:** Visualize the distribution of the data using bars to represent frequency within specified intervals.
- **Box Plots:** Display the distribution's median, quartiles, and outliers graphically.
- **Density Plots:** Smooth out the histogram to reveal the underlying probability density function.

Bivariate Analysis

- **Scatter Plots:** Explore the relationship between two continuous variables by plotting their values as points on a graph.
- **Correlation Coefficient:** Quantify the strength and direction of the linear relationship between two variables.
- **Line Plots:** Visualize the trend of one continuous variable over another (e.g., time series data).

Multivariate Analysis

- **Correlation Matrix:** Analyse the relationships between multiple continuous variables simultaneously.
- **Principal Component Analysis (PCA):** Reduce the dimensionality of your data while preserving most of the information.
- **Cluster Analysis:** Group similar data points based on their values.

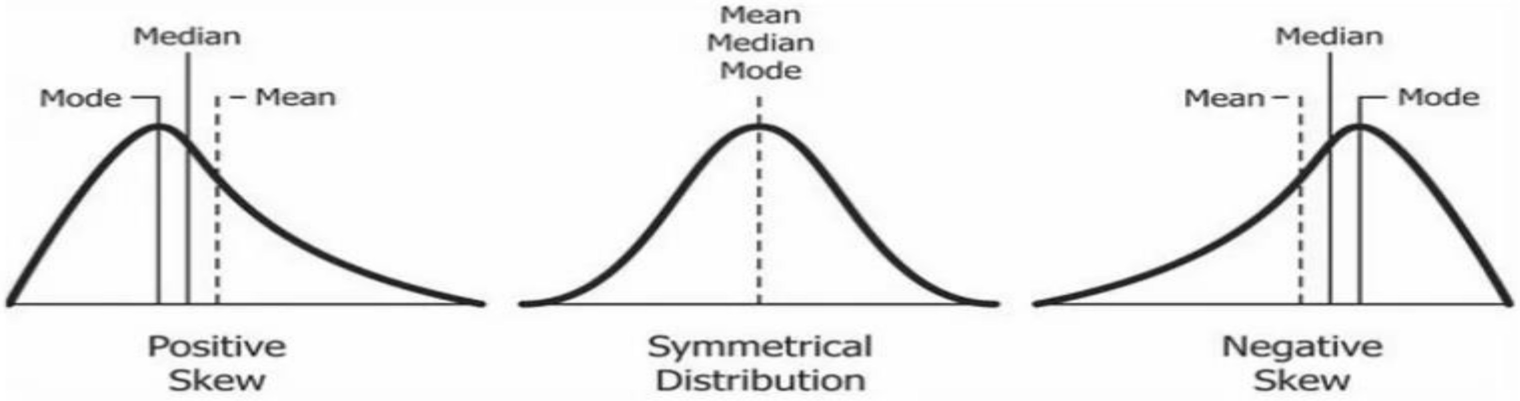
Additional Considerations

- **Data Cleaning:** Handle missing values, outliers, and inconsistencies in the data.
- **Data Transformation:** Consider applying transformations like log, square root, or normalization to improve data distribution or model performance.
- **Visualization Tools:** Utilize libraries like Matplotlib, Seaborn, or Plotly for effective visualizations.

STRUCTURE OF EDA

- Make plots of categorical and numerical data
- Interpret each plot to know the insights given in the data and document
- EDA is making as many plots and followed by interpretation for each plot
- After detailed EDA, summarize key observations through EDA which explains the current state of the data or current state of Affairs/Business the data represents

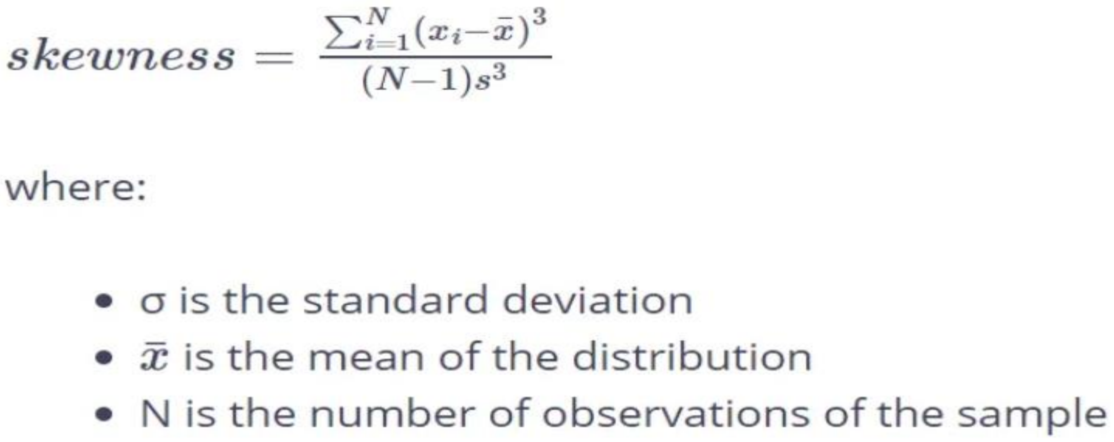
SKEWNESS



Symmetrical: When the skewness is close to 0 and the mean is almost the same as the median

Negative skew: When the left tail of the histogram of the distribution is longer and the majority of the observations are concentrated on the right tail. In this case, we can use also the term “left-skewed” or “left-tailed”. and the median is greater than the mean.

Positive skew: When the right tail of the histogram of the distribution is longer and the majority of the observations are concentrated on the left tail. In this case, we can use also the term “right-skewed” or “right-tailed”. and the median is less than the mean.



$$skewness = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^3}{(N-1)s^3}$$

where:

- σ is the standard deviation
- $\bar{x}$ is the mean of the distribution
- N is the number of observations of the sample

Symmetric:

- Values between -0.5 to 0.5

Moderated Skewed data:

- Values between -1 and -0.5 or between 0.5 and 1

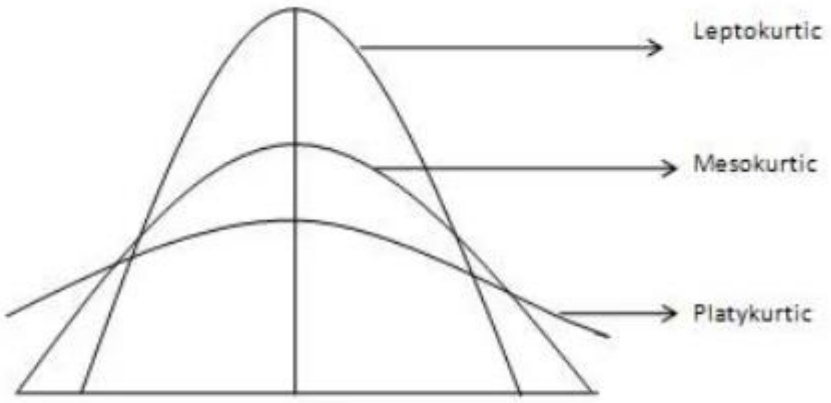
Highly Skewed data:

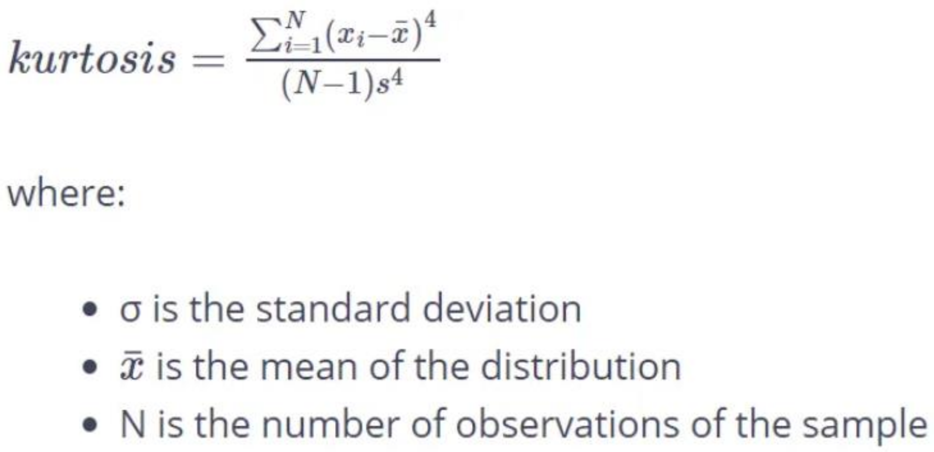
- Values less than -1 or greater than 1

KURTOSIS

1. In statistics, we use the kurtosis measure to describe the “tailedness” of the distribution as it describes the shape of it.
2. It is also a measure of the “peakedness” of the distribution
3. Three types of kurtosis.
   1. Mesokurtic: This is the normal distribution
   2. Leptokurtic: This distribution has fatter tails and a sharper peak. The kurtosis is “positive” with a value greater than 3
   3. Platykurtic: The distribution has a lower and wider peak and thinner tails. The kurtosis is “negative” with a value less than 3

Feature	Skewness	Kurtosis
Measures	Asymmetry	Tailedness (outliers)
Focus	Tail length	Peak and tails
Direction	Can be positive, negative, or zero	Can be leptokurtic, mesokurtic, or platykurtic





$$kurtosis = \frac{\sum_{i=1}^{N}(x_i-\bar{x})^4}{(N-1)s^4}$$

where:

- σ is the standard deviation
- $\bar{x}$ is the mean of the distribution
- N is the number of observations of the sample

** Skewness measures the degree of asymmetry of the distribution, while Kurtosis measures the degree of peakedness and flatness of a distribution

*** The value of both Skewness and Kurtosis ranges from -infinity to +infinity

WHEN SKEWNESS AND KURTOSIS

When you need to understand the direction of asymmetry:

Skewness tells you if your data is concentrated more on one side of the mean than the other

When you need to understand the probability of extreme values:

Kurtosis tells you how often outliers occur in your data.