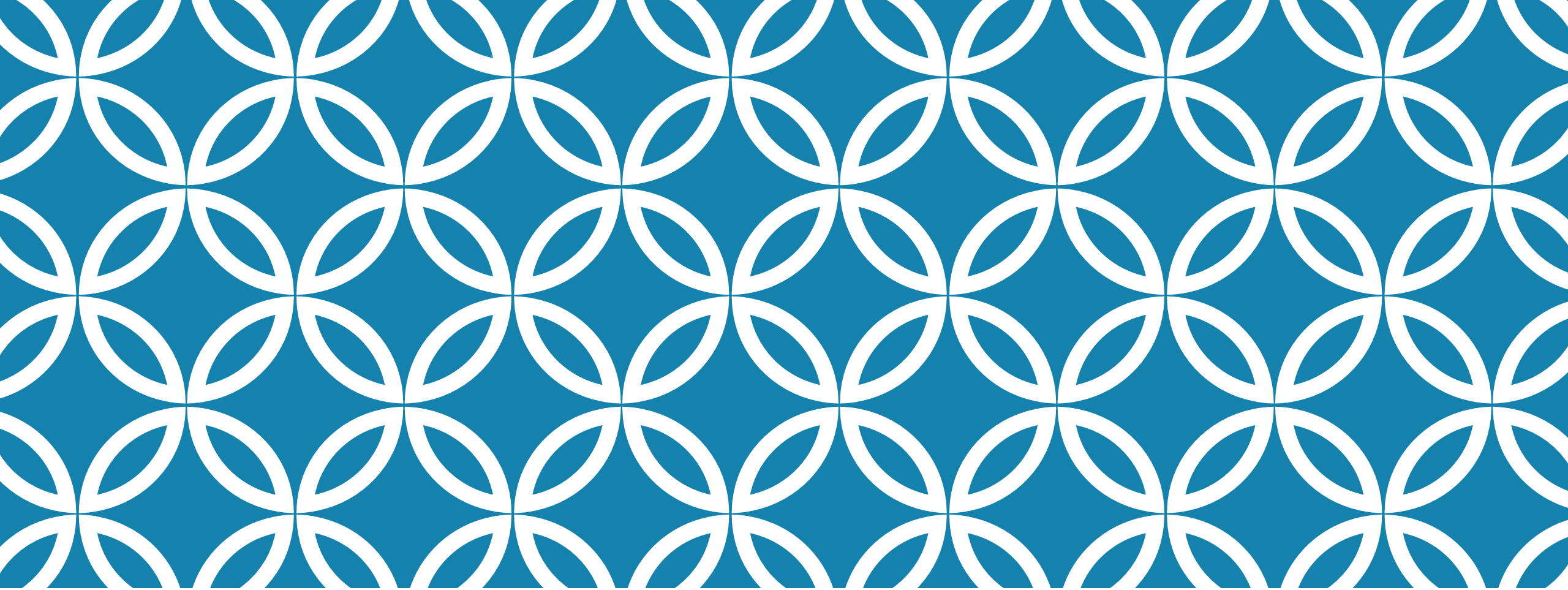


EDA OVERVIEW

Sharath Srivatsa

TYPES OF ANALYTICS

Descriptive Analytics

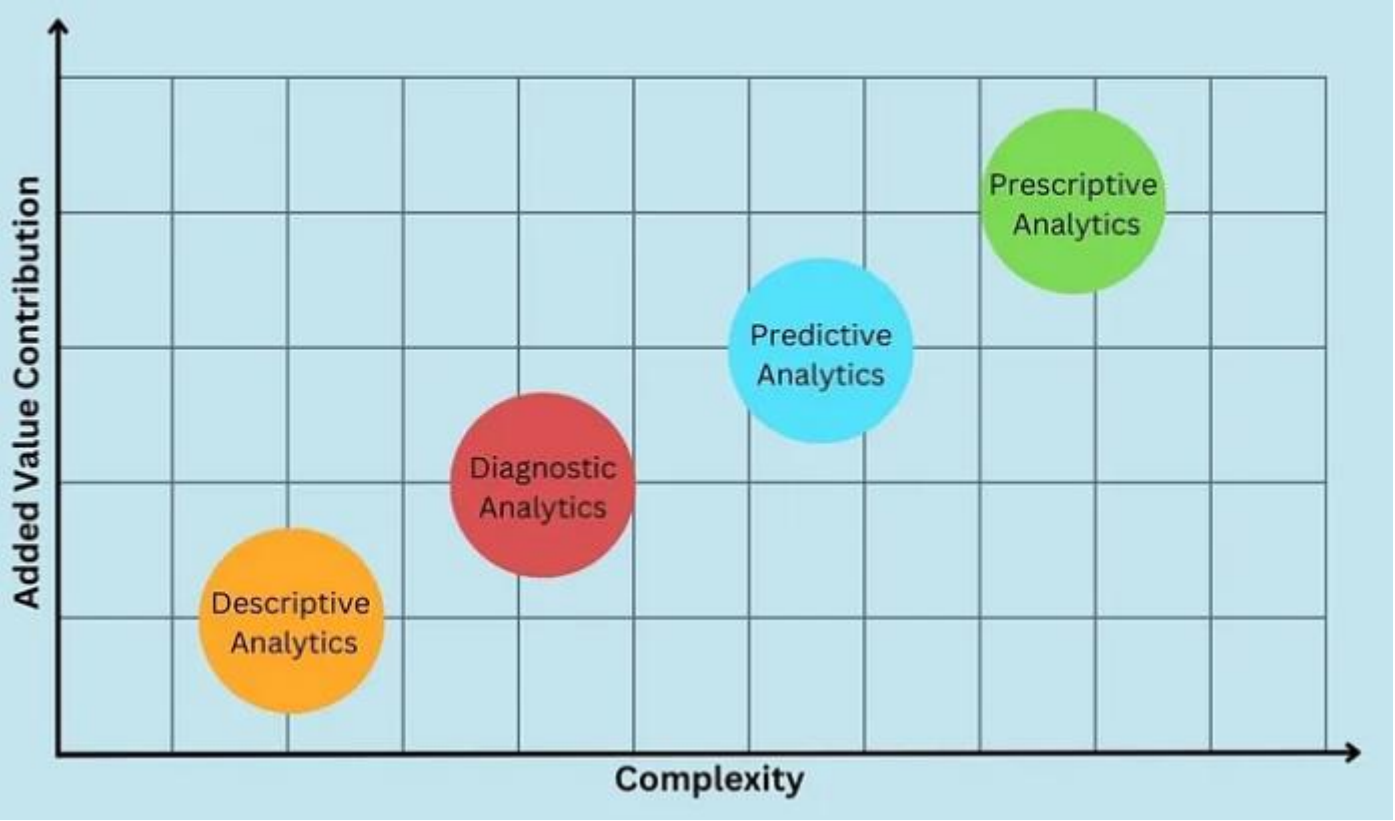
- We can consider this type of data analysis as the explainer of data

- 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

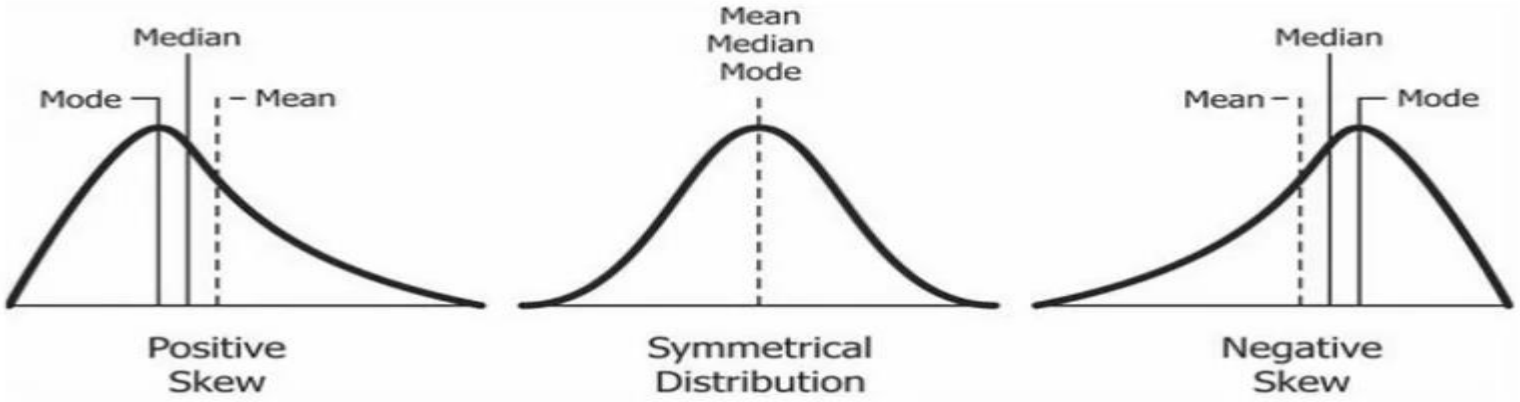
- Here, the final type of data analysis is prescriptive data analysis, which goes beyond descriptive and predictive analysis by recommending a course of action based on the analysis results

EDA – DESCRIPTIVE ANALYTICS

Agenda

- Understand Problem Statement
- Understand Data Definitions
- Comprehensive Data Review
- Univariate Analysis
- Multi-variate Analysis
- Basic Cleaning

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:

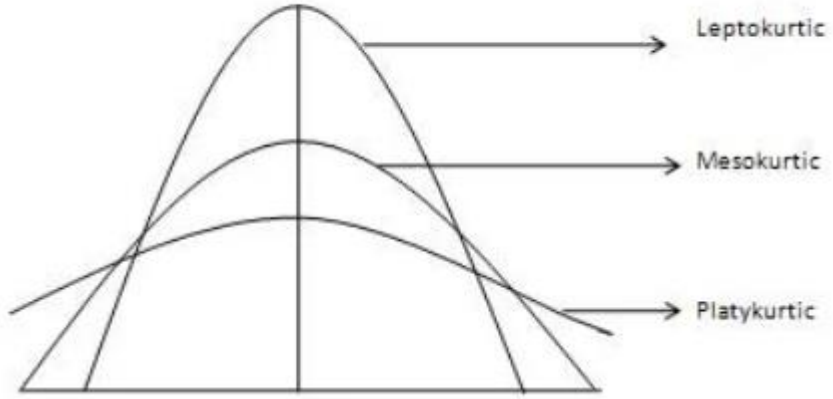
- Values between -1 and -0.5 or between 0.5 and 1

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



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

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