Exploratory Data Analysis

EMPOWERING High Performance Technology Teams

OVERVIEW

Overview

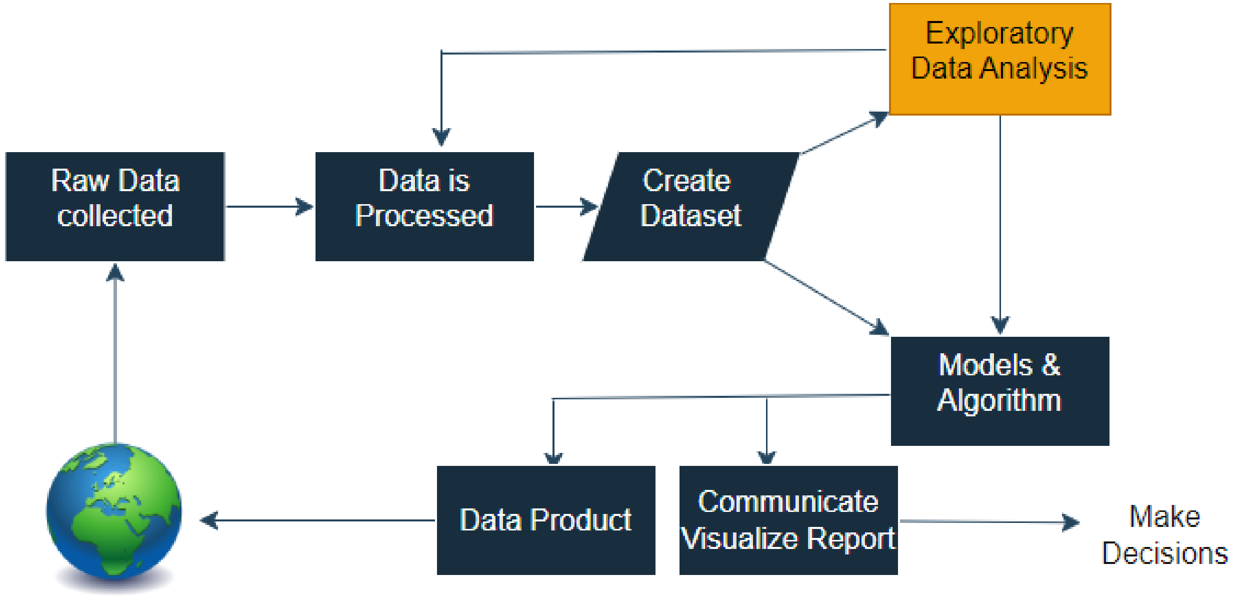
- Total Sessions (2 hours)
- Focus majorly on Exploratory Data Analysis
- Covering EDA steps, Types of EDA, Univariate analysis and Multivariate analysis

Agenda

- Data Science process
- Exploratory Data Analysis & Steps Involved
- Types of EDA
- Univariate analysis and its types
- Multivariate analysis and its types
- Pros & Cons of Uni and Multivariate Analysis

Data Science Process

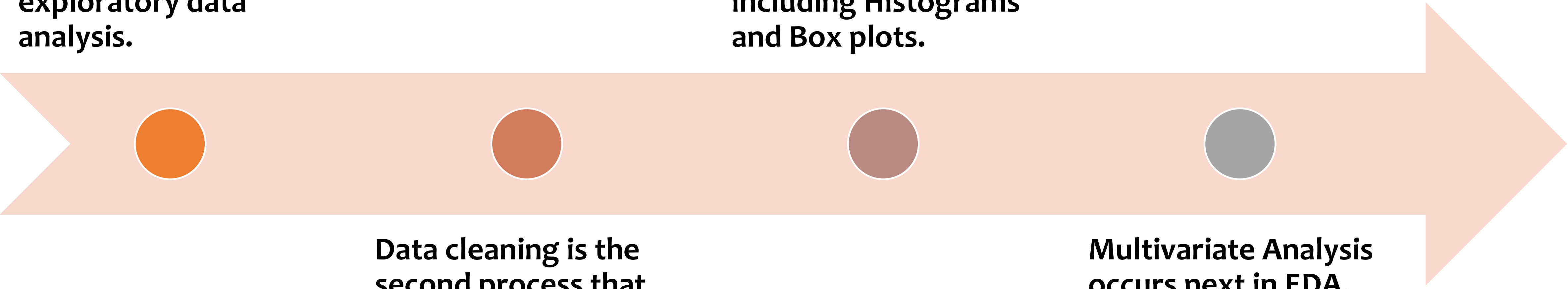
- The flow chart depicts the steps that occur during the Data Science process where the EDA process is performed after creating the dataset
- EDA is an iterative process



What is Exploratory Data Analysis?

- Exploratory Data Analysis was developed by American mathematician “John Tukey” in 1970s and are used mostly in the data discovery process
- EDA is used by data scientists and analysts to analyze and investigate data sets and summarizes their main characteristics
- It is beneficial in determining how best to manipulate data sources to achieve best answer
- It provides a clear vision for data scientists to discover patterns, test a hypothesis, spot anomalies, and check assumptions
- It also helps in determining whether statistical techniques considered are appropriate or not

Steps Involved in EDA



Data collection is the process of finding and loading data into a system and is considered as an essential part of exploratory data analysis.

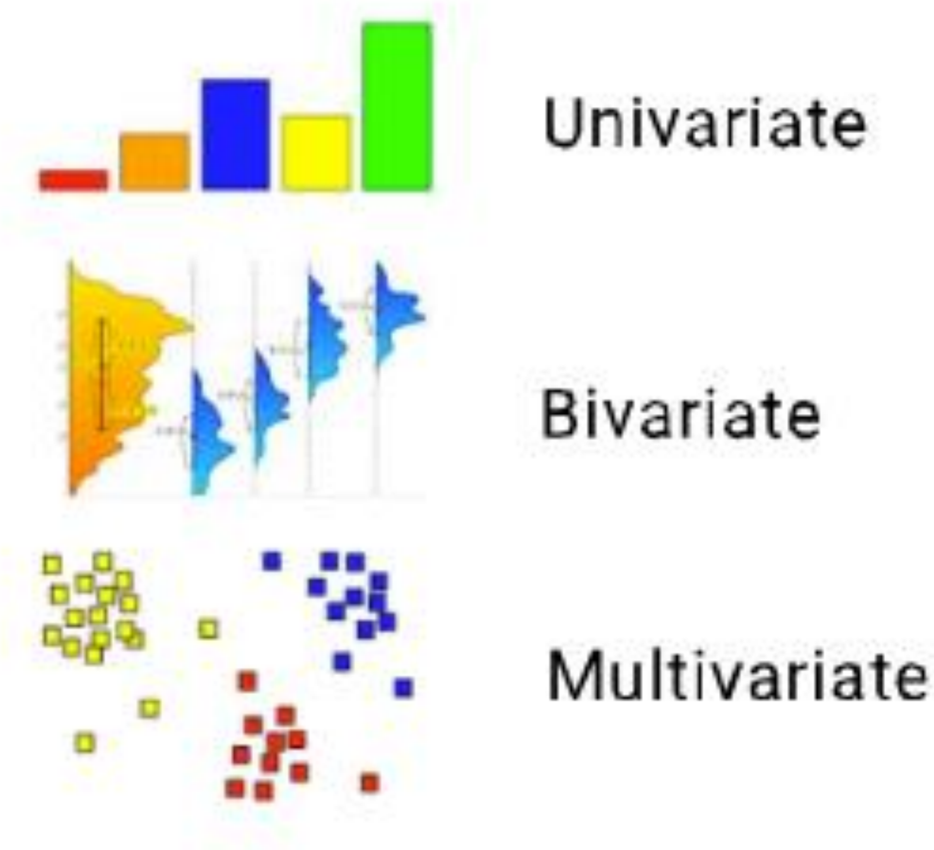
Data cleaning is the second process that involves the removal of unwanted variables and values from the data set.

The next step is Univariate analysis where the visual methods are used including Histograms and Box plots.

Multivariate Analysis occurs next in EDA.

Types of EDA

- Univariate non-graphical
- Univariate graphical
- Multivariate non-graphical
- Multivariate graphical

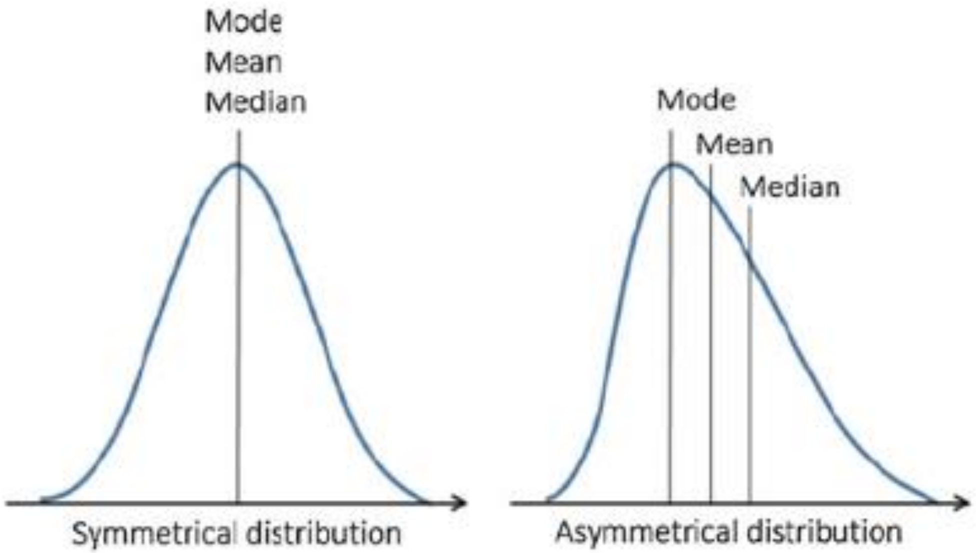


What is Univariate Analysis?

- Univariate analysis is the simplest form of analyzing data where “Uni” means “one”
- In other words, only a single variable is being explored separately at a time
- It looks at the range of values and describes the pattern of response to the variable
- It does not deal with causes or relationships between variables

Univariate Non-Graphical EDA

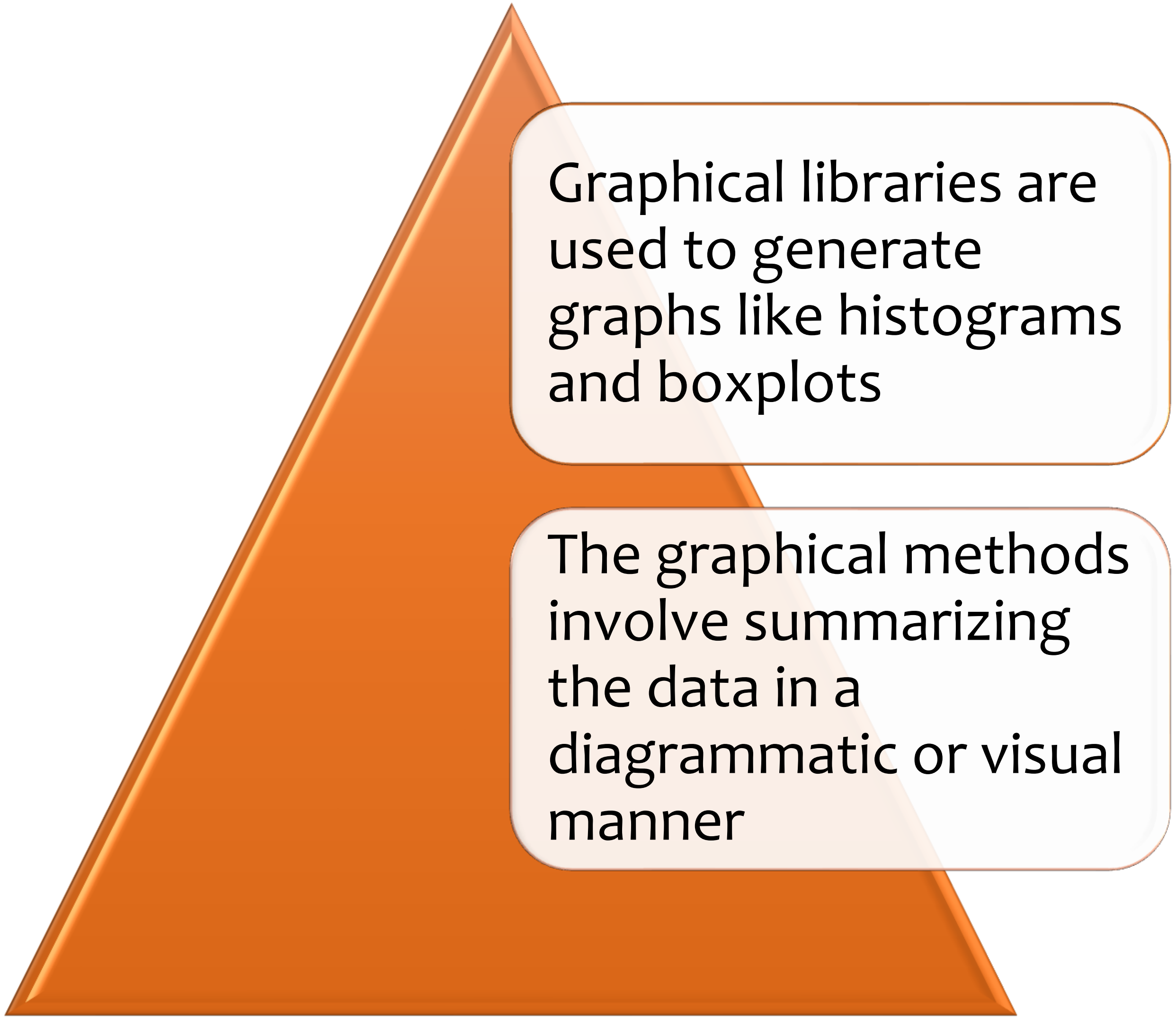
- Univariate non-graphical EDA is the analysis of a single variable in a dataset without using visual representations such as plots or graphs
- It examines the distribution and summary statistics of the variable to gain insights into its characteristics and understand its behavior
- The main goal is to describe and summarize the variables central tendency, dispersion, and shape of the distribution



Univariate Non-Graphical EDA

- Some common statistics considered in univariate non-graphical EDA are mean, median, mode, standard deviation, variance, range, and quartiles
- Univariate non-graphical EDA is useful when large datasets are used where graphical representation focuses on understanding the variables numerical properties
- Non-graphical EDA alone may not provide a comprehensive understanding of the relationships between variables
- It is combined with other forms of exploratory data analysis, such as bivariate or multivariate analysis, creating better understanding of the dataset

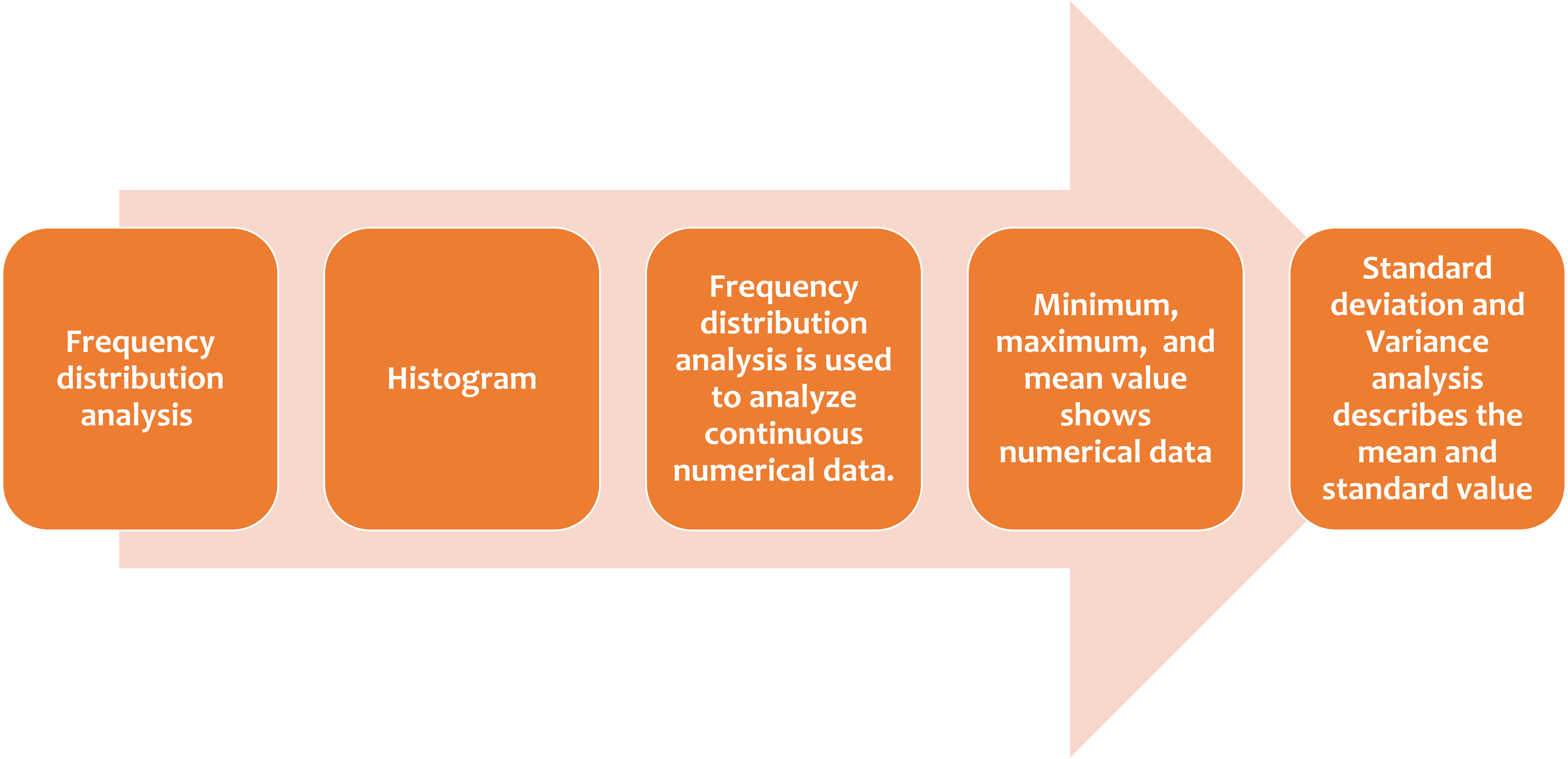
Univariate Graphical EDA



Graphical libraries are used to generate graphs like histograms and boxplots

The graphical methods involve summarizing the data in a diagrammatic or visual manner

Types of Univariate Analysis



Frequency distribution analysis

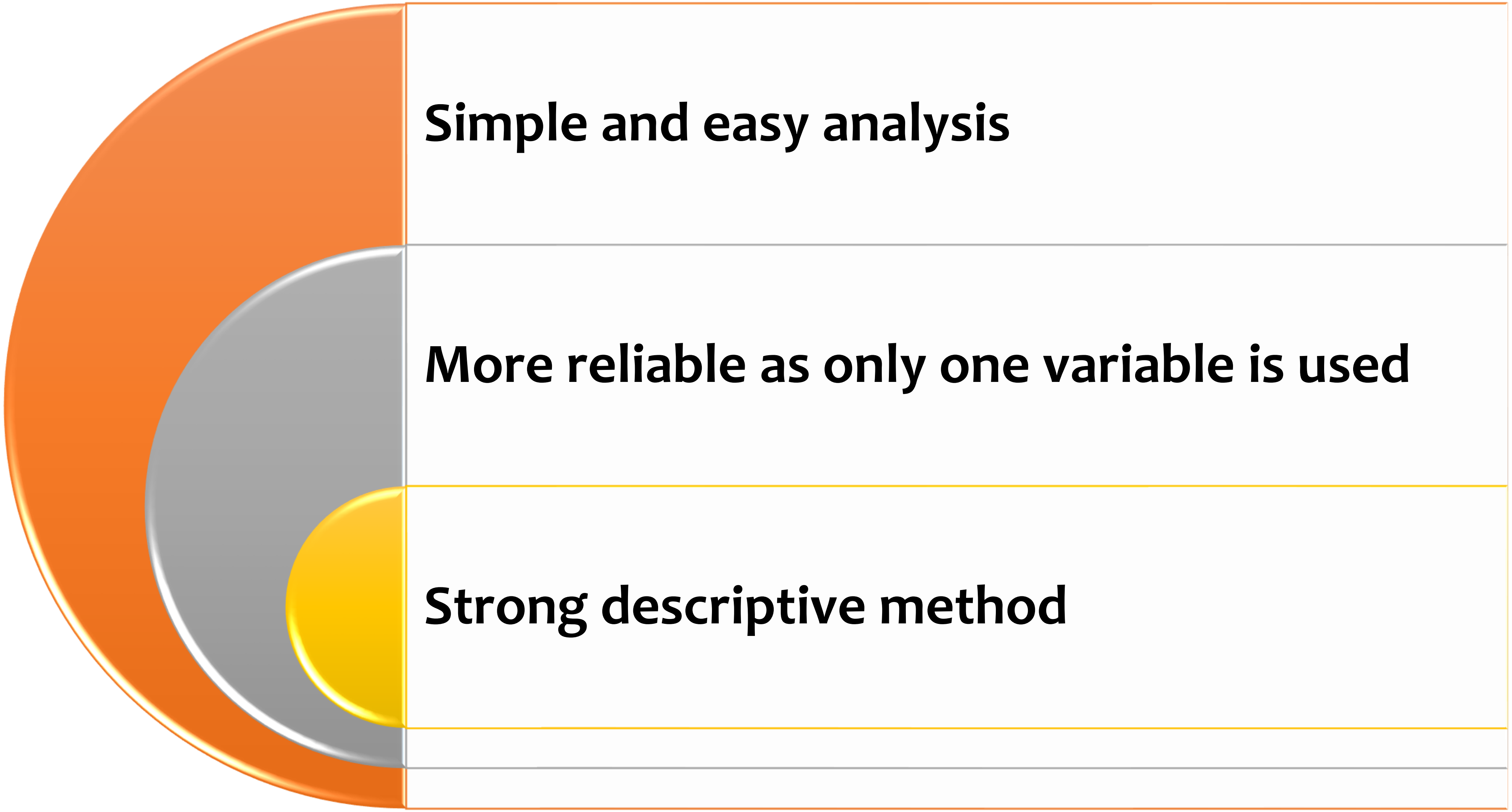
Histogram

Frequency distribution analysis is used to analyze continuous numerical data.

Minimum, maximum, and mean value shows numerical data

Standard deviation and Variance analysis describes the mean and standard value

Advantages of Univariate Analysis

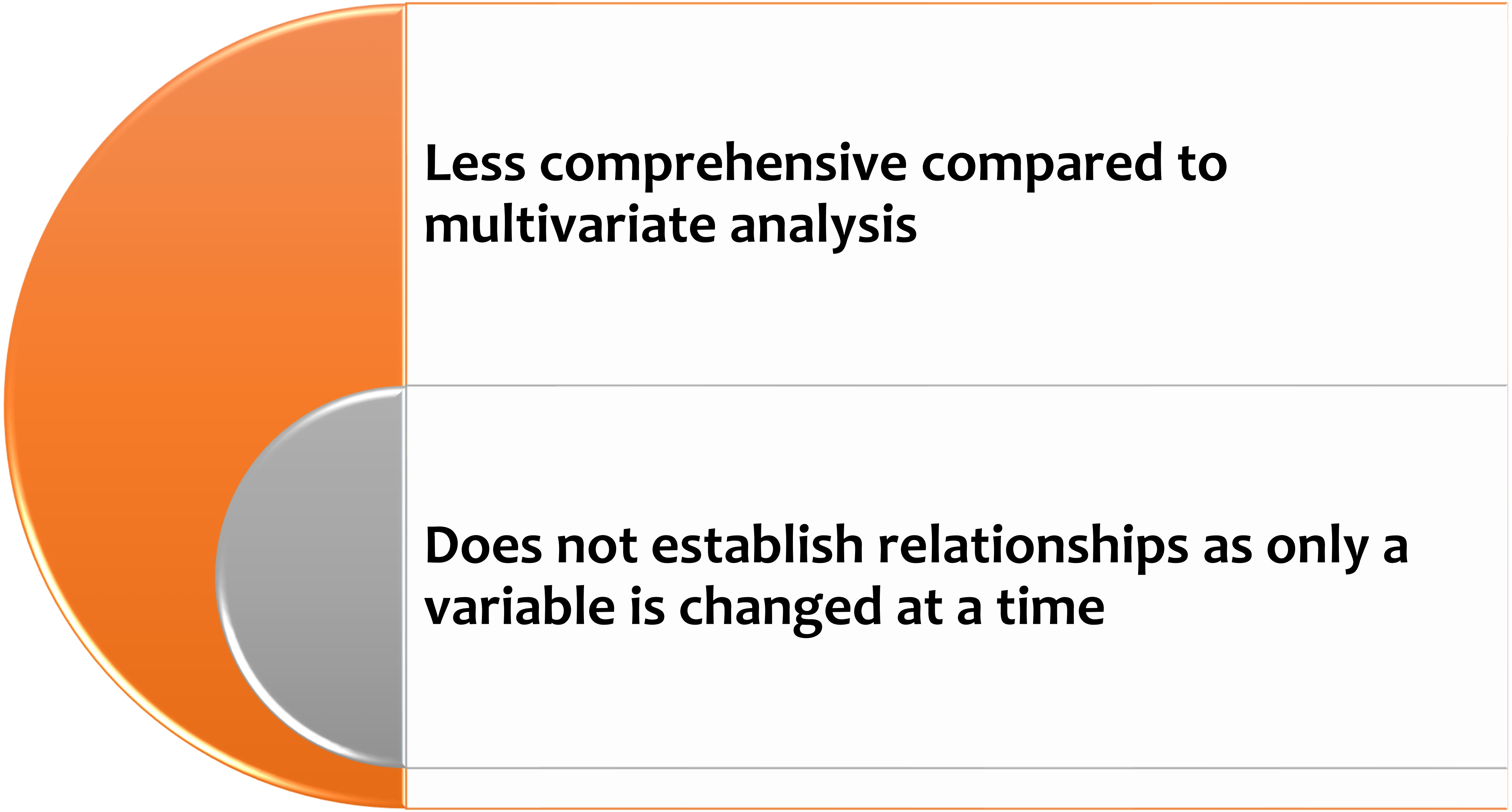


Simple and easy analysis

More reliable as only one variable is used

Strong descriptive method

Disadvantages of Univariate Analysis



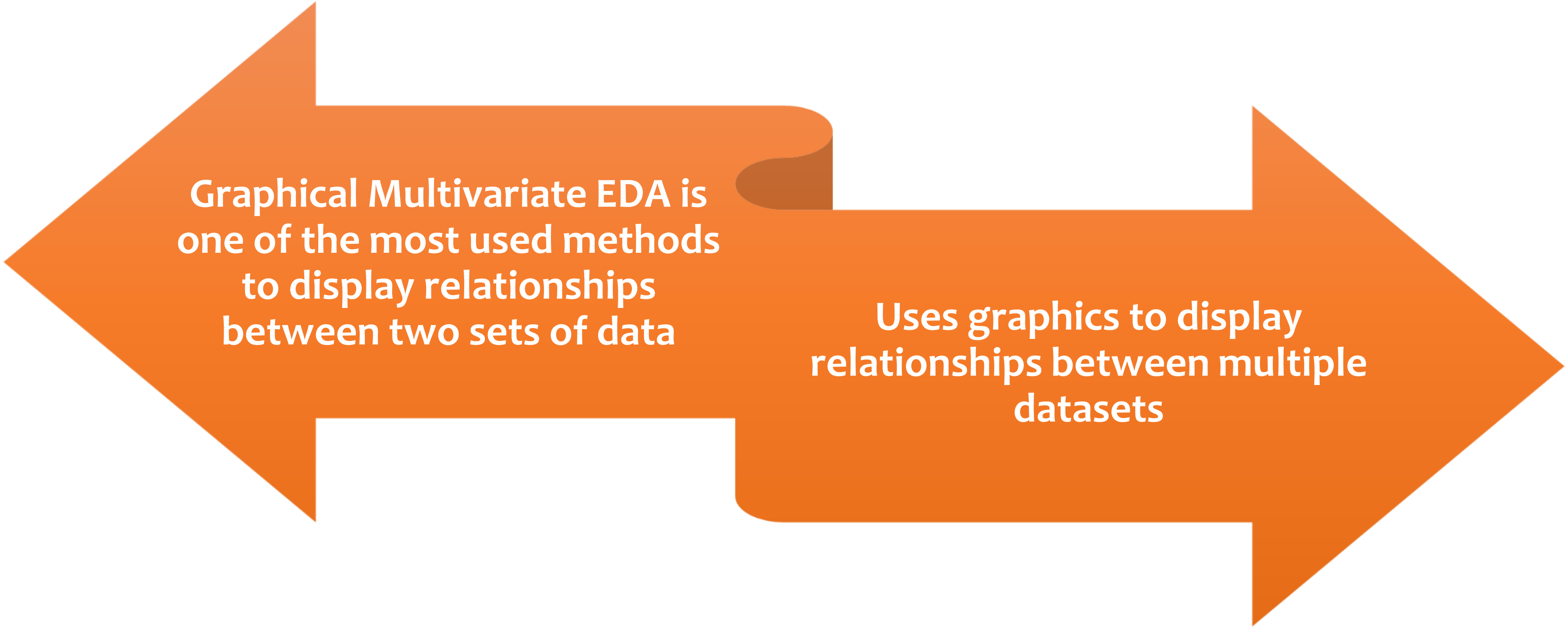
Less comprehensive compared to multivariate analysis

Does not establish relationships as only a variable is changed at a time

What is Multivariate Analysis?

- The statistical study of data where measurements are made on each experimental unit
- Multivariate analysis is a statistical technique that examines the relationships between multiple variables
- It can help companies to gain a more complete understanding of their data which can be used to improve efficiency, make better decisions, and identify new opportunities

Multivariate Graphical EDA



Graphical Multivariate EDA is one of the most used methods to display relationships between two sets of data

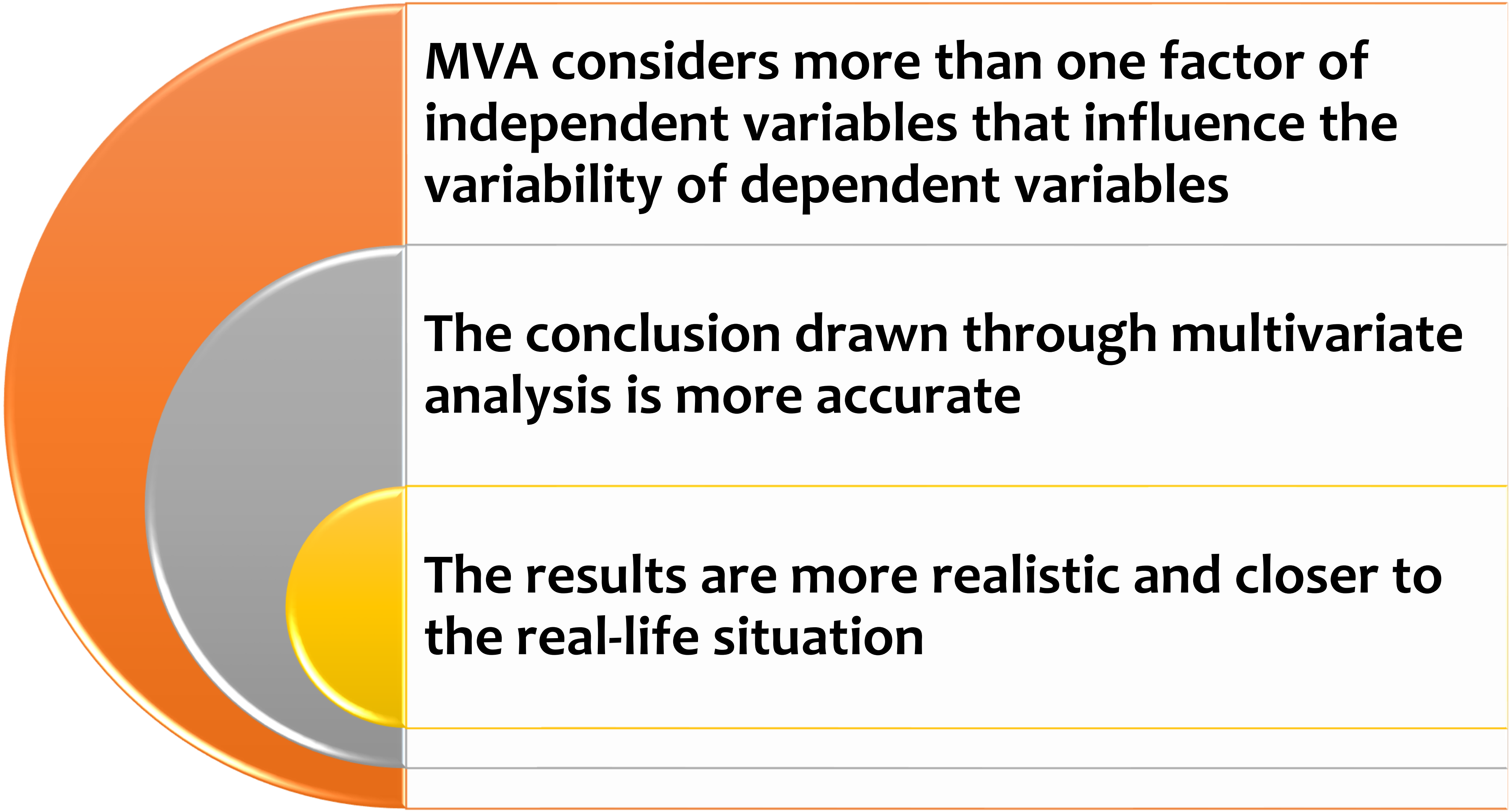
Uses graphics to display relationships between multiple datasets

Multivariate Non-Graphical EDA

A multivariate non-graphical EDA technique shows relationships between two or more variables of data through cross-tabulation or statistics

Tabulations are used in non-graphical methods to analyze and visualize tests for association or independence of categorical variables

Advantages of Multivariate Analysis

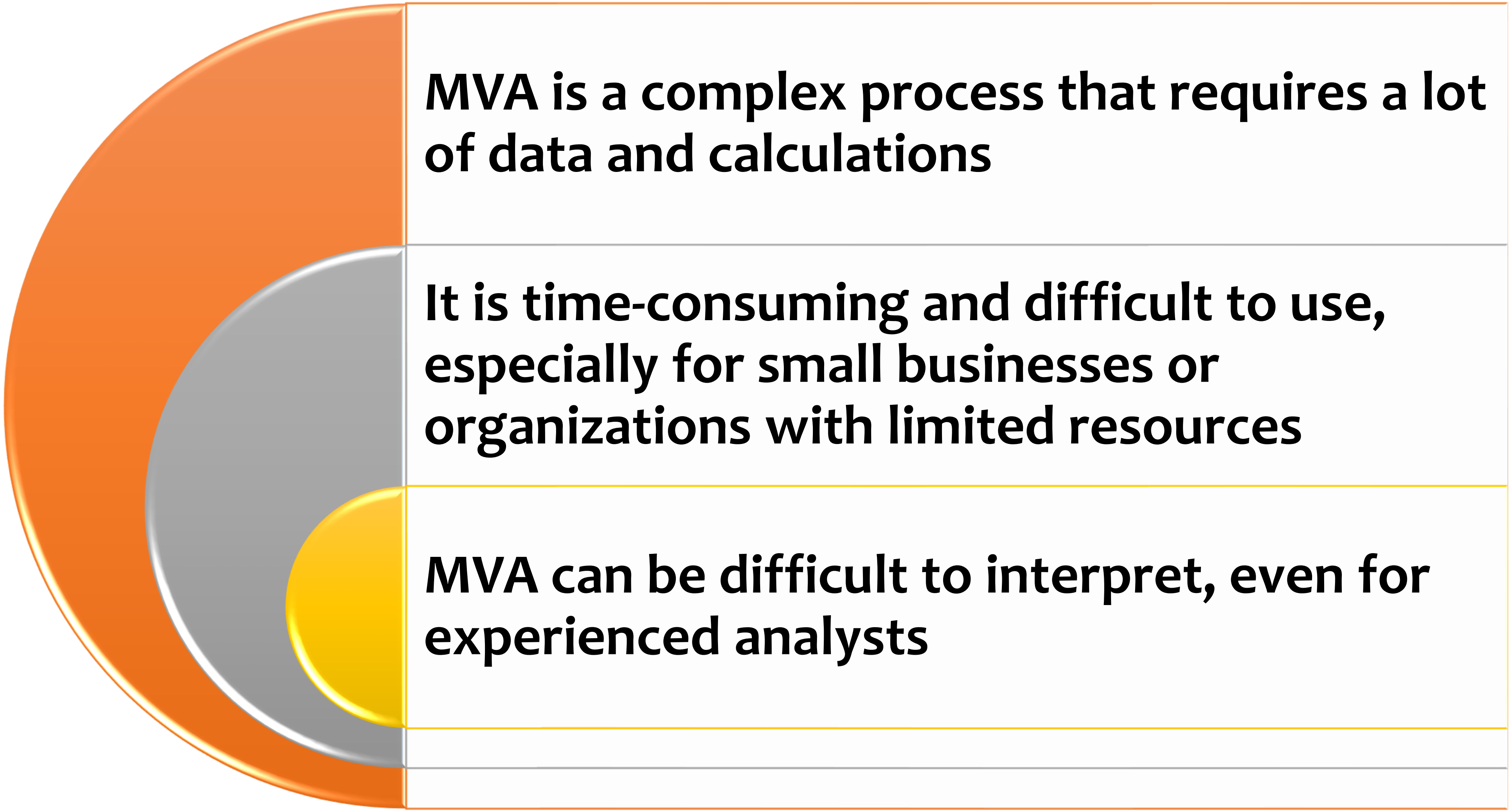


MVA considers more than one factor of independent variables that influence the variability of dependent variables

The conclusion drawn through multivariate analysis is more accurate

The results are more realistic and closer to the real-life situation

Disadvantages of Multivariate Analysis



MVA is a complex process that requires a lot of data and calculations

It is time-consuming and difficult to use, especially for small businesses or organizations with limited resources

MVA can be difficult to interpret, even for experienced analysts

Demo

SUMMARY

- EDA is an iterative process to analyze and investigate data sets and summarizes their main characteristics, and is beneficial in determining how best to manipulate data sources to achieve best answer
- Two major types of EDA are Univariate and Multivariate analysis which are further segregated into graphical and non-graphical kinds
- Univariate models analyze only one variable and does not reveal any relationship(s) between two or more factors
- Multivariate analysis unearths relationships between multiple variables

THANK YOU !