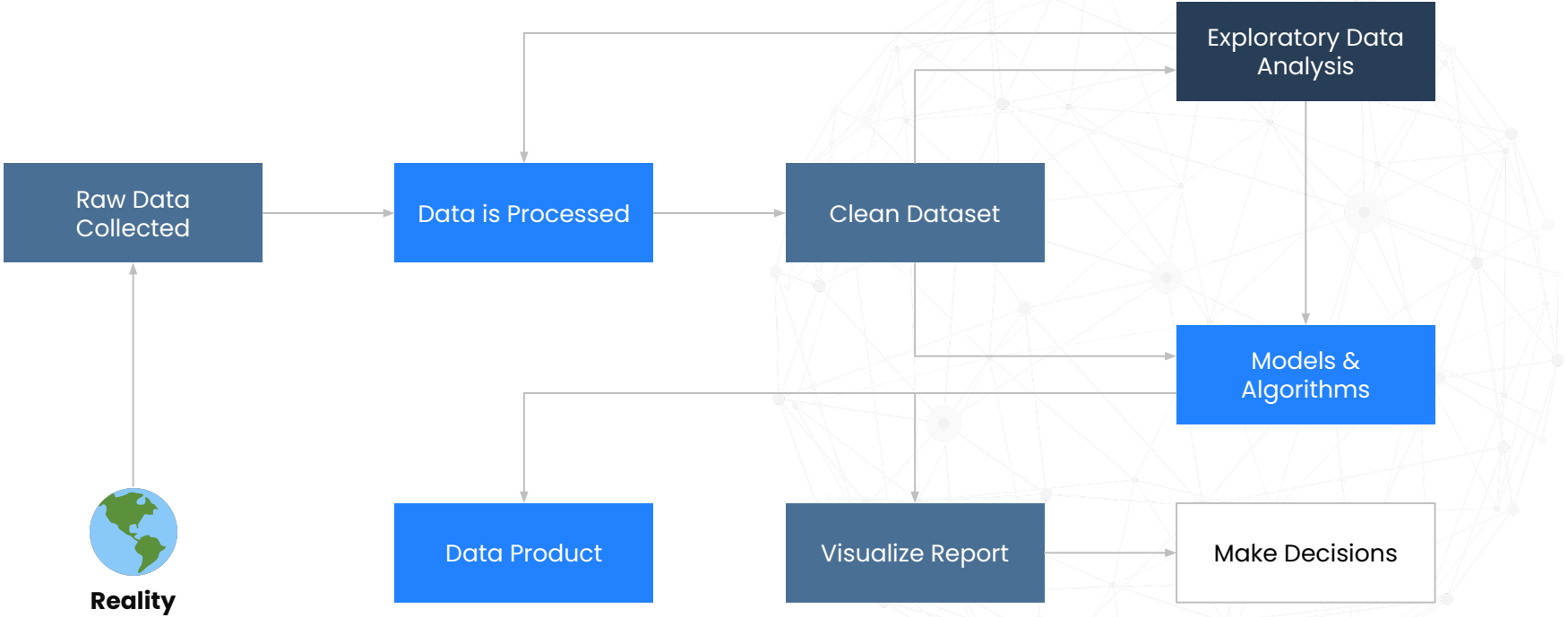
Exploratory Data Analysis

Agenda

1. What is Exploratory Data analysis?
2. Why EDA is important?
3. Visualization
   - Important charts for visualization.
4. Steps involved in EDA:-
   - Data Sourcing
   - Data Cleaning
   - Univariate analysis with visualization
   - Bivariate analysis with visualization
   - Derived Metrics
5. Use Cases

Data Analytics/Science Process



What is Exploratory Data Analysis

- Exploratory Data Analysis is an approach to analyze the datasets to summarize their main characteristics in form of visual methods.
- EDA is nothing but a data exploration technique to understand various aspects of the data.
- The main aim of EDA is to obtain confidence in a data to an extent where we are ready to engage a machine learning model.
- EDA is important to analyze the data it's a first steps in data analysis process.
- EDA give a basic idea to understand the data and make sense of the data to figure out the question you need to ask and find out the best way to manipulate the dataset to get the answer to your question.
- Exploratory data analysis help us to finding the errors, discovering data, mapping out data structure, finding out anomalies.
- Exploratory data analysis is important for business process because we are preparing dataset for deep thorough analysis that will detect you business problem.
- EDA help to build a quick and dirty model, or a baseline model, which can serve as a comparison against later models that you will build.

Visualization

Visualization is the presentation of the data in the graphical or visual form to understand the data more clearly. Visualization is easy to understand the data

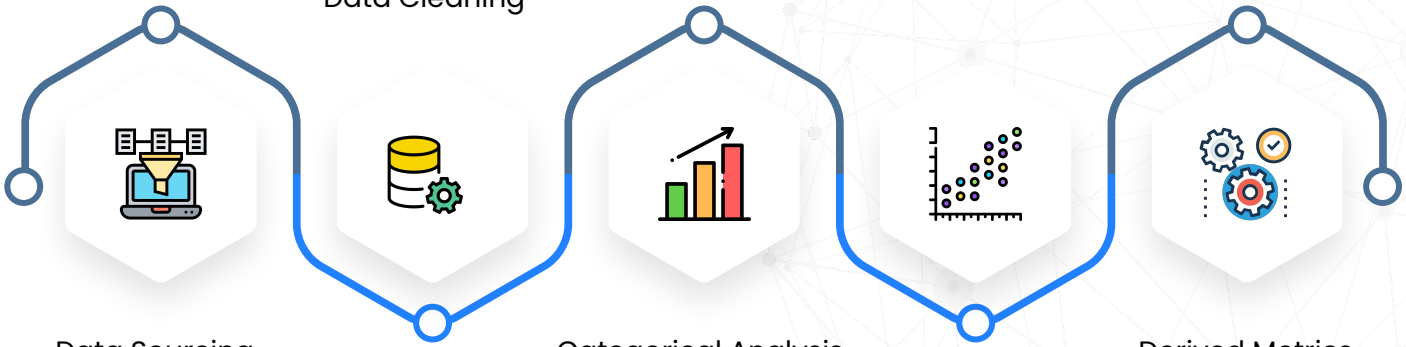
Easily understand the features of the data

Easily analyze the data and summarize it.

Help to get meaningful insights from the data.

Help to find the trend or pattern of the data.

Steps involved in EDA



Data Sourcing

- Data Sourcing is the process of gathering data from multiple sources as external or internal data collection.
- There are two major kind of data which can be classified according to the source:
  1. Public data
  2. Private data

Public Data

The data which is easy to access without taking any permission from the agencies is called public data. The agencies made the data public for the purpose of the research,

- **Example:** government and other public sector or ecommerce sites made the data public.

Private Data

Private Data:- The data which is not available on public platform and to access the data we have to take the permission of organisation is called private data.

- **Example:** Banking ,telecom ,retail sector are there which not made their data publicly available.

After collecting the data , the next step is data cleaning. Data cleaning means that you get rid of any information that doesn't need to be there and clean up by mistake.

Data Cleaning is the process of clean the data to improve the quality of the data for further data analysis and building a machine learning model.

The benefit of data cleaning is that all the incorrect and irrelevant data is gone, and we get the good quality of data which will help in improving the accuracy of our machine learning model.

The following are some steps involve in Data Cleaning

Handle Missing Values

Standardization of the data

Outlier Treatment

Handle Invalid values

Some Questions you need to ask yourself about the data before cleaning the data

Handle Missing Values

Delete Rows/Columns

This method we commonly used to handle missing values. Rows can be deleted if it has insignificant number of missing value Columns can be delete if it has more than 75% of missing value

Replacing with mean/ median/mode

This method can be used on independent variable when it has numerical variables. On categorical feature we apply mode method to fill the missing value.

Algorithm Imputation

Some machine learning algorithm supports to handle missing value in the datasets. Like KNN, Naïve Bayes, Random forest.

Predicting the missing values

Prediction model is one of the advanced method to handle missing values. In this method dataset with no missing value become training set and dataset with missing value become the test set and the missing values is treated as target variable.

Example

For Numerical Data

Airlines	Ticket Price
Indigo	3887
Air Asia	7662
Jet Airways	-
Air India	5221
SpiceJet	4321

Suppose we have Missing values in the categorical data:

Then we take the mode of the dataset a to fill the missing values: Here :

Mode = Indigo

We substitute the Indigo in place of missing value in Airline column

The following are some steps involve in Data Cleaning

Suppose we have Airlines ticket price data in which there is missing value.

Steps to fill the numeric missing value:-

1. Compute the mean/median of the data
   (3887+7662+5221+4321)/4 = **5272.75**
2. Substitute the Mean of the value in missing place.

For categorical Data

Airlines	Ticket Price
Indigo	3887
Indigo	7675
Air Asia	4236
-	6524
Jet Airways	4321

Standardization/Feature Scaling

Feature scaling is the method to rescale the values present in the features. In feature scaling we convert the scale of different measurement into a single scale. It standardize the whole dataset in one range.

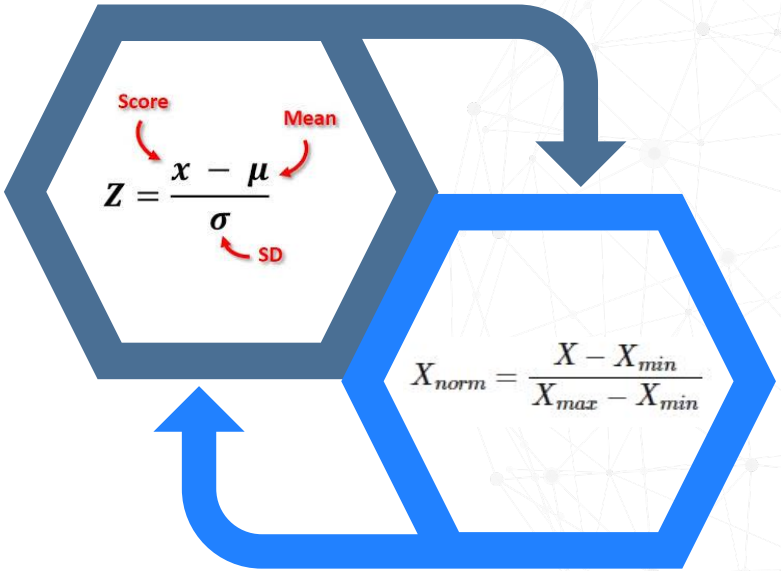
Importance of Feature Scaling

When we are dealing with independent variable or features that differ from each other in terms of range of values or units of the features, then we have to normalize/standardize the data so that the difference in range of values doesn't affect the outcome of the data.

Feature Scaling Method.

Standard Scaler

Standard scaler ensures that for each feature, the mean is zero and the standard deviation is 1, bringing all feature to the same magnitude. In simple words Standardization helps you to scale down your feature based on the standard normal distribution



Min-Max Scaler

Normalization helps you to scale down your features between a range 0 to 1

Example

Normalization

Age	Income (£)	New value
24	15000	(15000 – 12000)/18000 = 0.16667
30	12000	(12000 – 12000)/18000 =0
28	30000	(30000 – 12000)/18000 =1

Income Minimum = 12000
Income Maximum = 30000
(Max – min) = (30000 – 12000) = 18000

Hence, we have converted the income values between 0 and 1

Please note, the new values have
Minimum = 0
Maximum = 1

Example

Standardization

Age	Income (£)	New value
24	15000	(15000 - 19000)/9643.65 = -0.4147
30	12000	(12000 - 19000)/9643.65 = -0.7258
28	30000	(30000-19000)/9643.65 = 1.1406

Average = (15000 + 12000 + 30000)/3 = 19000
Standard deviation = 9643.65

Hence, we have converted the income values to lower values using the z-score method.

x = c(-0.4147, -0.7258, 1.1406)
mean(x) = -0.000003 ~ 0
var(x) = 0.999 ~1

Outlier Treatment

Outliers are the most extremes values in the data. It is an abnormal observations that deviate from the norm. Outliers do not fit in the normal behavior of the data.

Detect Outliers using following methods

1. Boxplot
2. Histogram
3. Scatter plot
4. Z-score
5. Inter quartile range(values out of 1.5 time of IQR)

Handle Outlier using following methods

1. Remove the outliers.
2. Replace outlier with suitable values by using following methods:-
   - Quantile method
   - Inter quartile range
3. Use that ML model which are not sensitive to outliers
4. Like:-KNN, Decision Tree, SVM, Naïve Bayes, Ensemble methods

Handle Invalid Value

Encode Unicode properly

In case the data is being read as junk characters, try to change encoding, E.g. CP1252 instead of UTF-8.

Convert incorrect data types

Correct the incorrect data types to the correct data types for ease of analysis. E.g. if numeric values are stored as strings, it would not be possible to calculate metrics such as mean, median, etc.

Some of the common data type corrections are — string to number: "12,300" to “12300”; string to date: "2013-Aug" to “2013/08”; number to string: “PIN Code 110001” to "110001"; etc.

Correct values that go beyond range

If some of the values are beyond logical range, e.g. temperature less than -273° C (0° K), you would need to correct them as required.

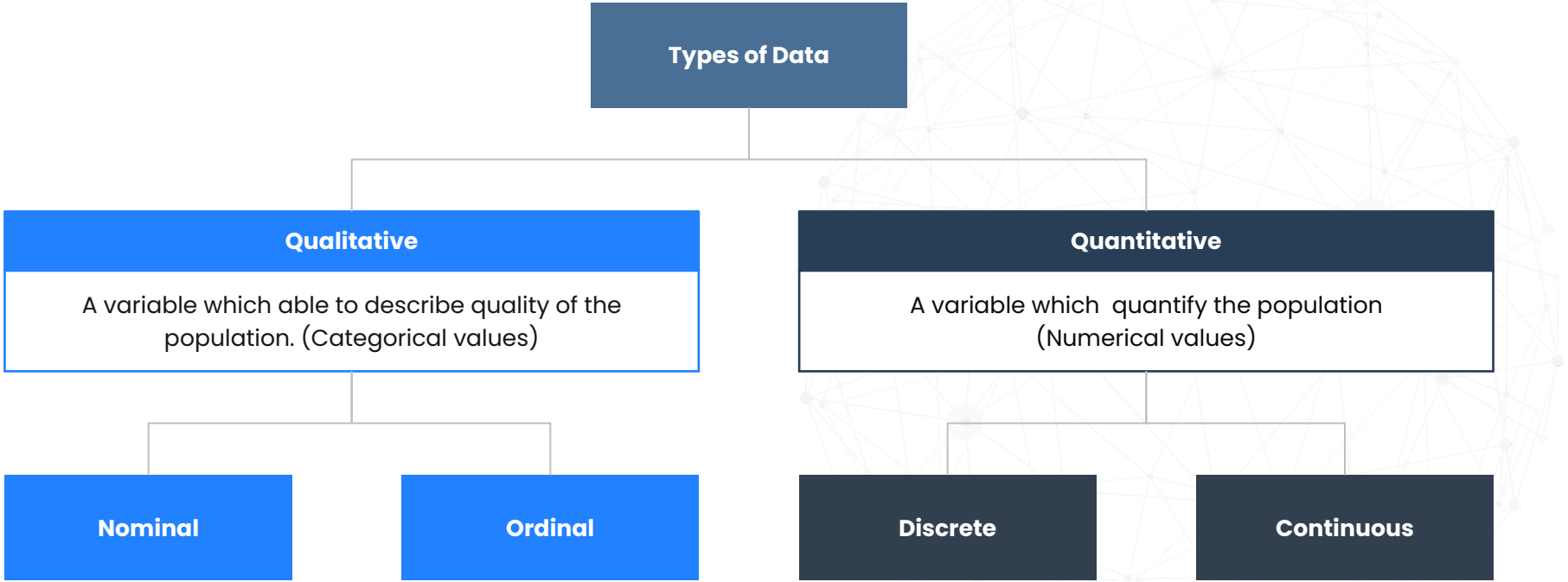
A close look would help you check if there is scope for correction, or if the value needs to be removed.

Correct wrong structure

Values that don’t follow a defined structure can be removed.

E.g. In a data set containing pin codes of Indian cities, a pin code of 12 digits would be an invalid value and needs to be removed. Similarly, a phone number of 12 digits would be an invalid value

Types of Data



Types of Data

Discrete

It has a discrete value that means it take only counted value not a decimal values. Like count of student in class

Nominal

It represent qualitative information without order. Value represent a discrete units.

Like Gender: Male/Female ,Eye colour.

Continuous

A number within a range of a value is usually measured, such as height

Ordinal

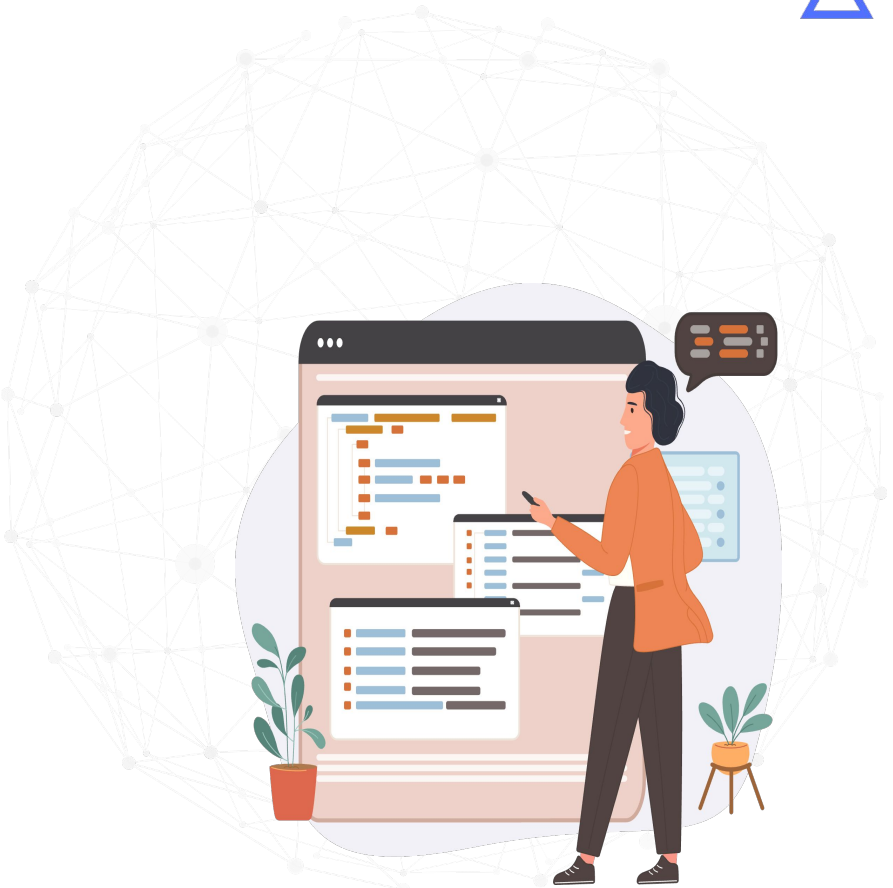
It represent qualitative information with order. It indicate the measurement classification are different and can be ranked. Lets say Economic status: high/ medium /low which can ordered as low, medium, high.

Types of Analysis

Univariate Analysis

Univariate analysis is the simplest form of analyzing data.

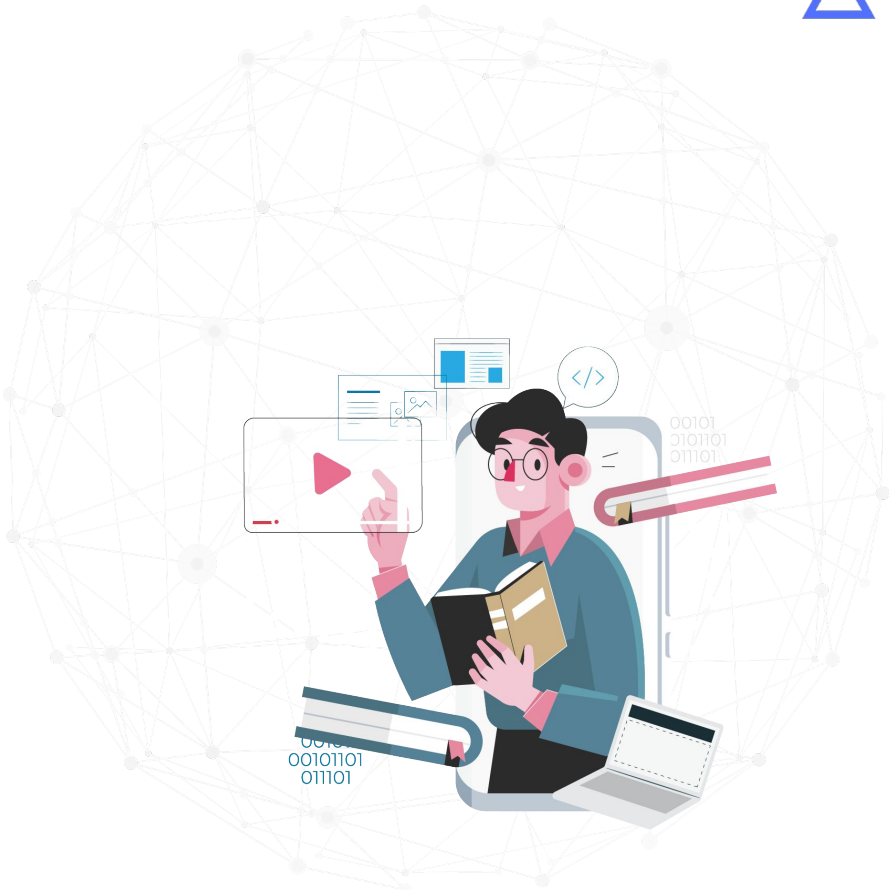
“Uni” means “one”, so in other words your data has only one variable.



Bivariate Analysis

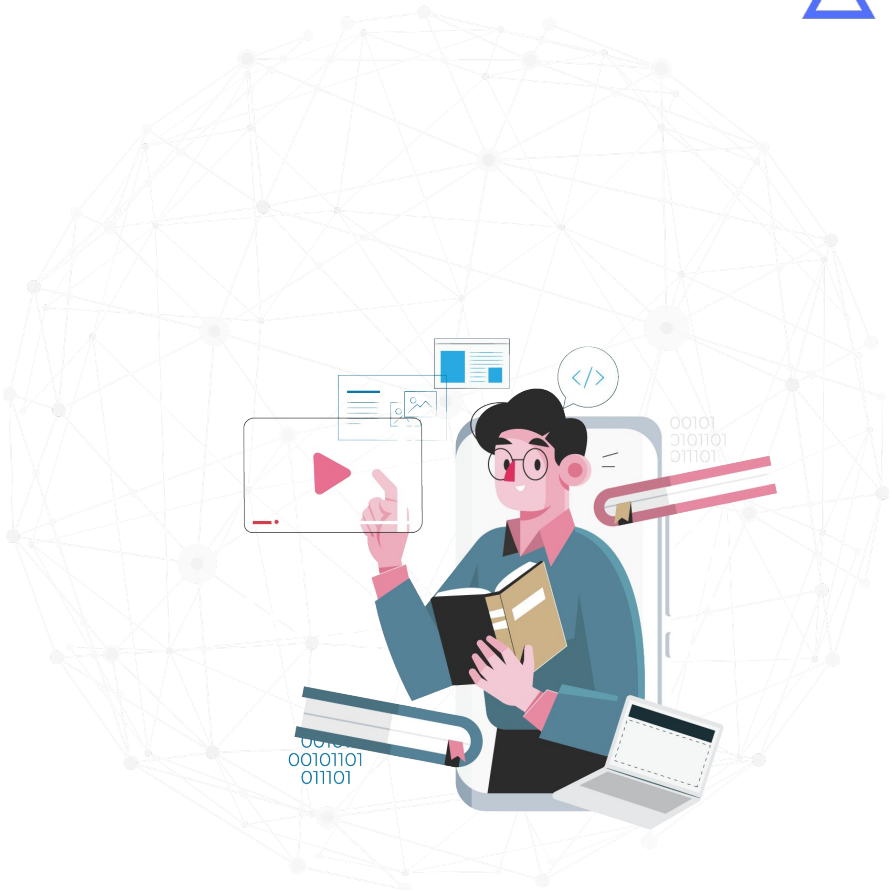
Data which has two variables, you often want to measure the relationship that exists between these two categorical variables.

Bivariate Analysis can also be performed with numerical values, or a combination of numerical & categorical values



Multivariate Analysis

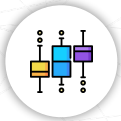
Data which has more than two variables, you often want to measure the relationship that exists between these features



Numerical Analysis

We also perform various analysis over Numerical data.

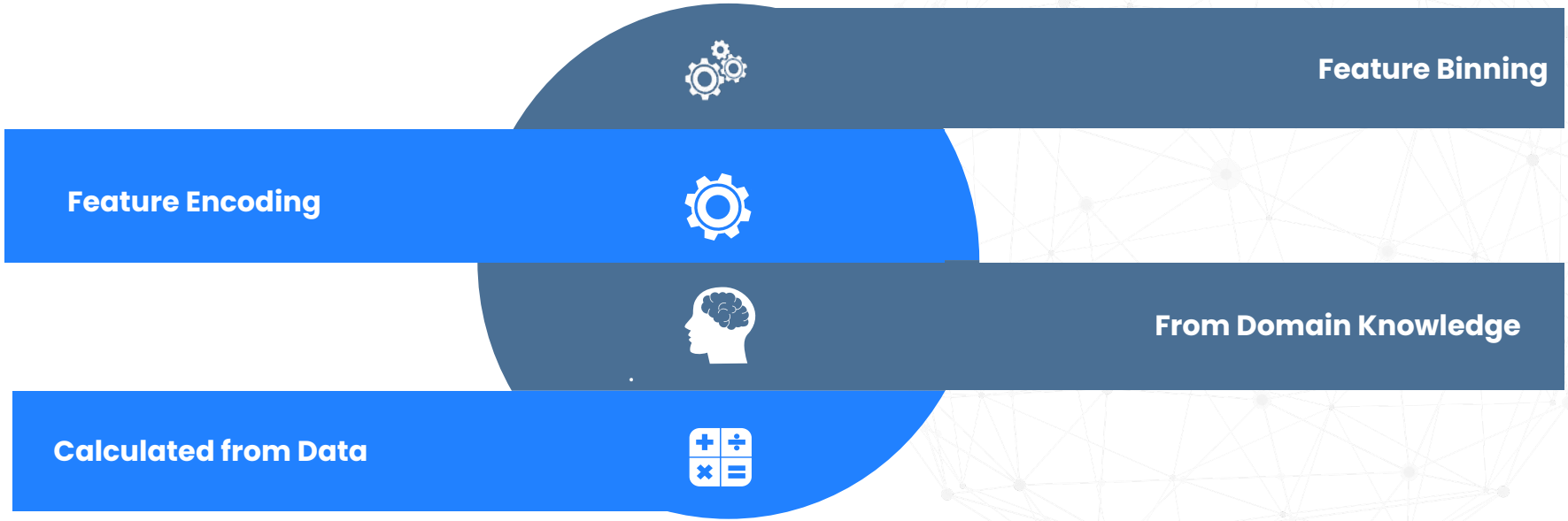
For example, dealing with a single numerical variable, we might be interested in knowing their statistical information such as mean, median, 25th Percentile, 75th Percentile, min, max etc.



Similarly, while analyzing multiple features, we might be interested in knowing their correlation with each other.

Derived Metrics

Derived metrics create a new variable from the existing variable to get a insightful information from the data by analyzing the data.



Feature Binning

Feature binning converts or transform continuous/numeric variable to categorical variable.
It can also be used to identify missing values or outliers.

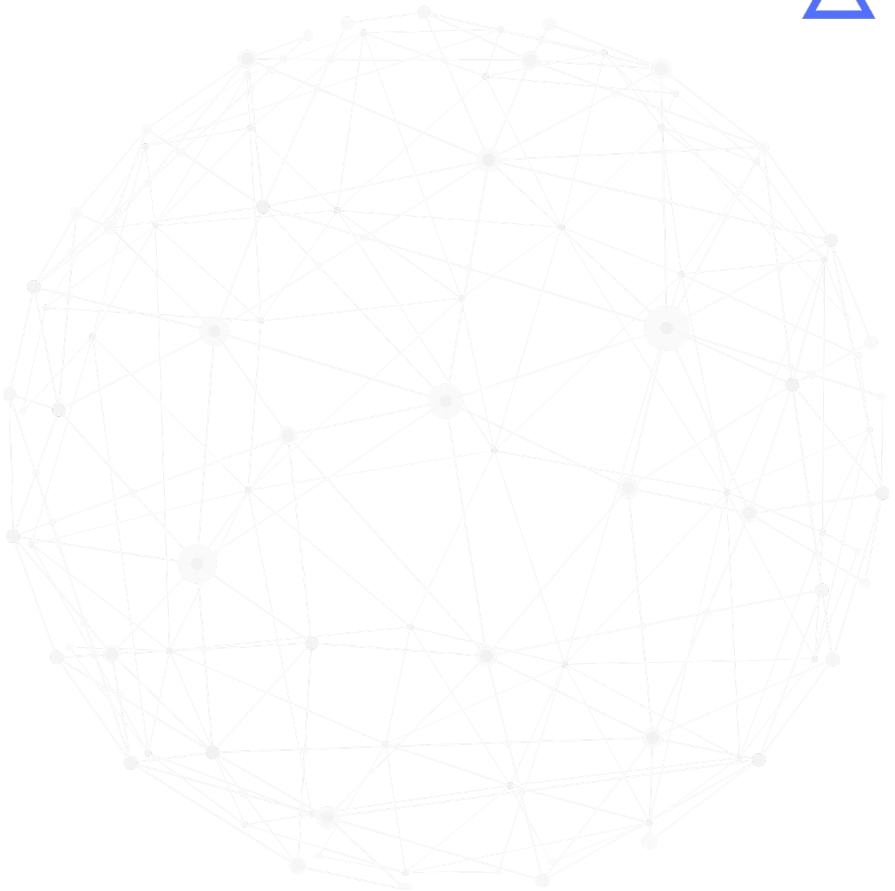
Type of Binning

a. Equal width binning
b. Equal frequency binning

Feature Binning

It transform continuous or numeric variable into categorical value without taking dependent variable into consideration

Equal Width	Equal width separate the continuous variable to several categories having same range of width.
Equal Frequency	Equal frequency separate the continuous variable into several categories having approximately same number of values.



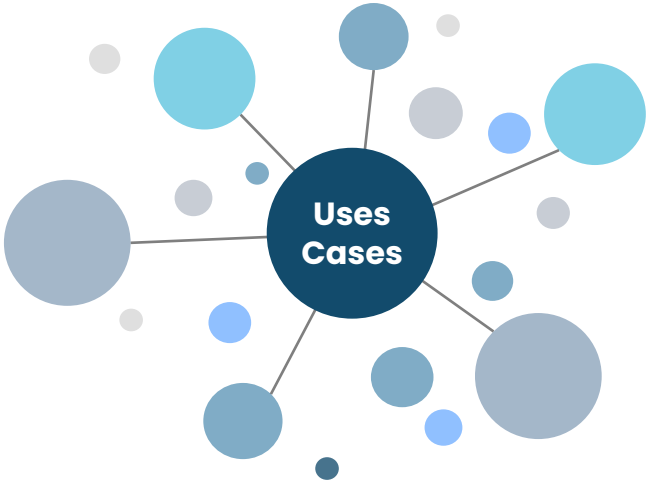
Feature Encoding

Feature encoding help us to transform categorical data into numeric data.

Label encoding	Label encoding is technique to transform categorical variables into numerical variables by assigning a numerical value to each of the categories.
One-Hot encoding	This technique is used when independent variables are nominal. It creates k different columns each for a category and replaces one column with 1 rest of the columns is 0. Here, 0 represents the absence, and 1 represents the presence of that category.

Use cases

Basically EDA is important in every business problem, it's a first crucial step in data analysis process.



Some of the use cases where we use EDA is:-

Cancer Data Analysis

In this data set we have to predict who are suffering from cancer and who's not.

Fraud Data Analysis in E-commerce Transactions

In this dataset we have to detect the fraud in a E-commerce transaction.

Thank you