→ Exploratory Data Analysis Explore the data means analyze and Investigate the data. Summarized their main characteristic. Employing visualization method.

1. Univariate Non graphical.

- Univariate means Single feature or column.
- Description or Summary of this feature.

→ Mean
→ Median
→ Mode
→ Count
→ Quantile
→ Percentile
→ Min
→ Max
→ Standard deviation
→ Skewness
→ Outlier detection
→ Variance.

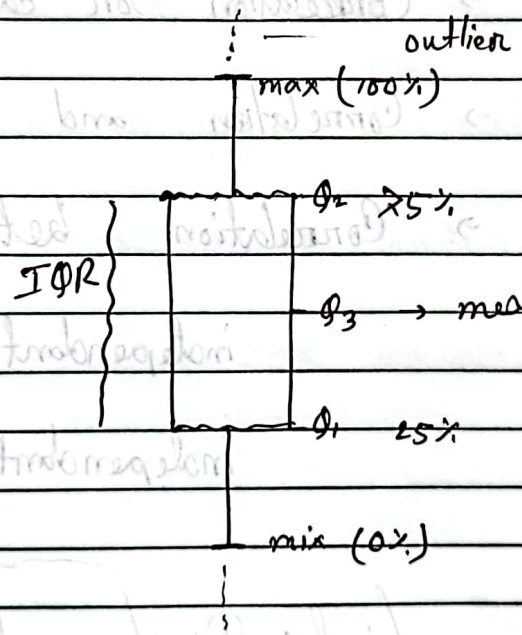
2. Univariate Graphical → Non Graphical can't provide full picture of data.

→ Summarize or Describe data in a Graphical way.

→ Histogram.

→ Steam and Leaf plot

→ Boxplot



five number Summary.

→ Bar - plot.

→ Pie - chart.

3. Bivariate means two feature.

Multivariate means multiple feature.

→ Multivariate non-graphical.

→ Describe the data non graphical way.

→ Relationship between two or more variables. (tabular formate)

→ Correlation for categorical variable

→ Correlation and Covariance.

→ Correlation between
independant vs independant
independant vs dependent.

→ Multivariate graphical: Display the Relationship between two or more variables.

→ Describe the information in a graphical way.

→ Scatter plot.

→ Pair plot

→ Heatmap (Correlation)

→ Facetgrid

→ Clustermap

→ Side by side box-plot.

→ Contingency table is one of the techniques for exploring two or more variables in a tabular or matrix formate.

[Multivariate non graphical]

Preprocessing

→ Data cleaning.
→ Data integration
→ Data Reduction.
→ Data Transformation.

Data Quality -

→ Accuracy.
→ Completeness.
→ Timeliness.
→ Consistancy.
→ Beleivability.
→ Interpretibility.

Accuracy
Completeness
Consistancy

Data cleaning,

→ Remove outlier, handle missing value.
→ smothout - Noisy data
→ Redundancy
→ Resolve - inconsistancy

Data Reduction. Reduce the data size or dimension.

- Dimensionality Reduction
- Numerosity Reduction.
- Compressed.

Data Transformation :- Discretization,
Normalization
Standardization

→ Raw data values for attributes are replace by Range or higher Conceptual levels.

Age → youths, adult, Senior.

Handle missing value,

→ Mean, Median, Mode,
→ Constant,
→ Most frequent,
→ Remove or skip or ignore
→ Manually.
→ Build a model for predict missing value - (est)

Noisy : Noise is a Random error or variance in a measured variable.

→ Binning
→ Regression analysis
→ Outlier Analysis

Data cleaning Process.

① → Discrepancy detection,

- human error data entry
- poorly design data " forms
- inconsistency data representation
- System Error.

→ Data Reduction

→ Data Reduction strategy