Friday, 5 December 2025 5:56 PM

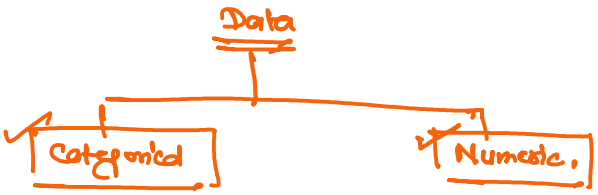
- What is Bi-variate analysis?

Bi-variate
two → columns / features.

- So, here we analyze two variable together.

- Why two variable?
  - How one variable affects another
  - What kind of relationship or pattern exist b/w the two.
  - Whether one variable depends on another.

3 possible combination in Bi-variate analysis.



- Possible scenario.
  - Numerical – Numerical. →
    - Age v/s salary.
    - Height v/s weight
    - Price v/s Distance.

    → Both the column are numerical
    - continuous / discrete.
  - Numerical v/s categorical
    - Salary v/s Gender
    - Marks v/s department
    - Age v/s city.

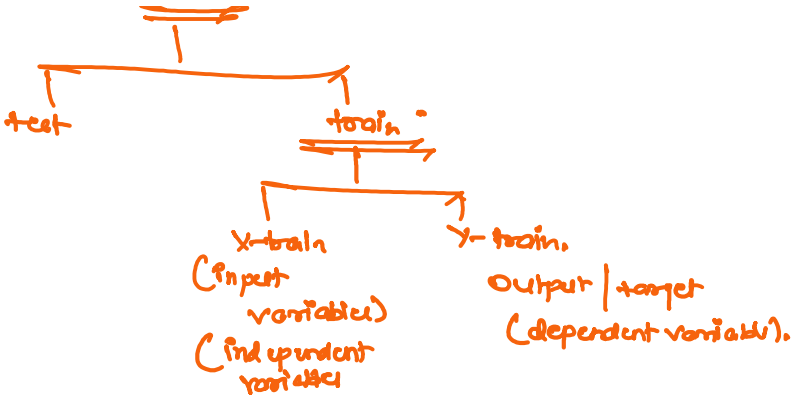
    → one column → numerical
    other col → categorical.
  - Categorical v/s categorical
    - Gender v/s Purchase
      City v/s Product Type
    - Vehicle Type v/s Fuel Type

- Goal
  - Univariate analysis
    - Describe one variable → statistical summary.
    - Identify the distribution, range, outliers, missing value etc.
  - Bi-variate analysis
    - To identify the relationship / pattern b/w two variables.

- Why do we need it (Importance)

In ML

Data



- What does the algorithm do?
  - It learns the relationship b/w X and y.
  - It builds a model that expresses how x influence y.

Model (data + algo) → trained algorithm, on the data.

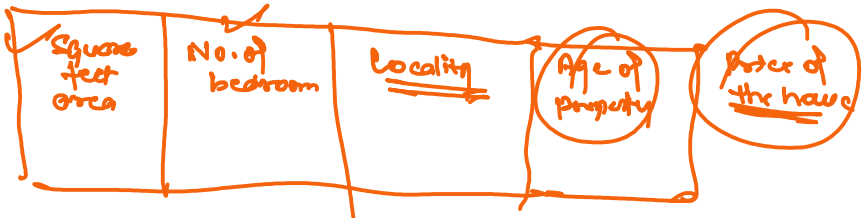
- A model is nothing but a mathematical representation of relationship.
  - ↳ we must study the relationship manually using bi-variate analysis.

- Type of analysis under Bi-variate variate
  - Visual. → graphs or plots.
  - Non-visual. → statistical and numerical measure

But which one to use & when to use depend upon.
- Whether the variables → numerical.
  - → categorical
  - → mixed.

- Why study relationship?

Suppose In a dataset



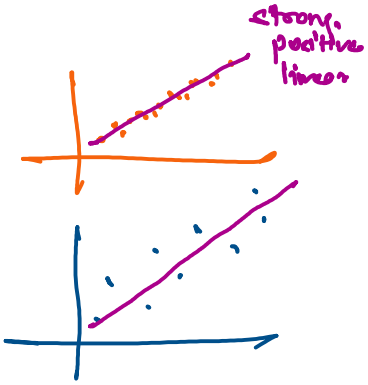
price = X_1 (sqft) → X_2 (bedroom) → X_3 (locality) → X_4 (Age of

- Numerical - Numerical
  - Non-visual method → Correlation.
    - Pearson Correlation Coefficient
      - Measure linear relationship
      - Works well data is normally distributed.
      - Value range −1 to 1.
    - Spearman Rank Correlation coefficient.
      - Measure monotonic relationship. (non-linear)
      - Works well data → not normally distributed

Both these method help us to see.

When x ↑ , does Y ↑
Y ↓
Y stay unrelated.

- Visual approach → Scatter plot.



Why?

Trends.
- Direction. (positive / -ve).
- Strength.
- Clusters.
- Outliers.
- Linear / non-linear patterns

Case-2 Categorical v/s Categorical.

- Non-visual approach → crosstab.
- It helps us to understand,
  - which category combination are common
  - which combination rarely happens.
  - how to categorical feature interacts with each other.

Crosstab

Gender v/s Purchase → this directly shows if male or females purchase more.

2. Visual method (will discuss soon)
   - Clustered bar chart
   - Stacked bar chat
   - Heatmap of crosstabs. → b/w two variables
   - Mosaic plot

Case-3 Numerical v/s Categorical.

- Non-visual approach
  - Groupby + descriptive statistics (aggregate functions).
    - min
    - max
    - median
    - standard deviation
    - count
    - percentile
    - value_counts.
  - Pivot table.

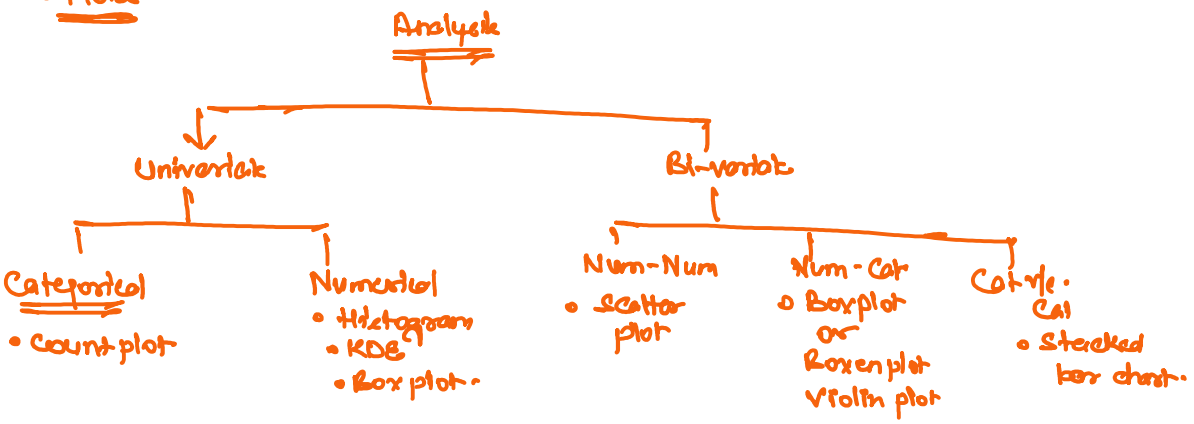
Gender v/s salary.

- It tells us.
  - average salary of male
  - average salary of female

- Visual-approach
  - Box plot (Best choice)

- Boxen plot - (Alternative).

- Plotly



Total plots we covered.

Univariate

- Count plot
- Histogram
- KDE.
- Box plot

Bivariate

- Scatterplot
- Box plot (Boxen plot) / violin.
- Stacked bar chart
- Violin plot (Optional).

7-8 → extremely powerful plots.