Thursday, 4 December 2025 6:00 PM

* Re-learning EDA

EDA

- Univariate
- Bi-variate
- but very few know the purpose behind them.
  - why do we perform univariate analysis & bi-variate analysis.
  - what exactly should we analyze under univariate & bivariate
  - what mistake we should avoid.

3. major points.

- What comes under univariate analysis.
- what comes under bi-variate analysis.
- Why each part is important from the modelling perspective.

Uni-variate

Uni → one
variate → variable.

Analyze one - single variable

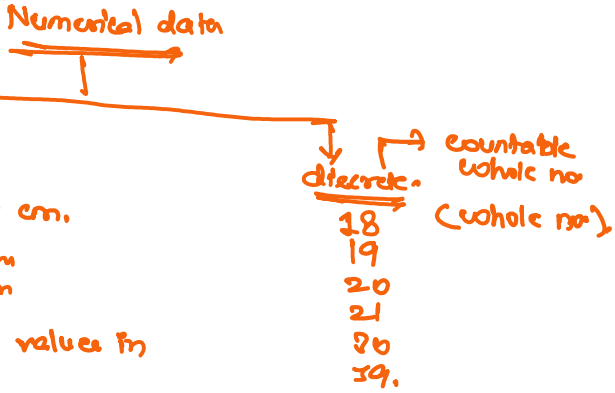
- Type of Analysis.
  - Visual
  - Non - visual analysis. (statistical)

Bi-variate

Variable

- **Categorical** (Qualitative.)
  - Gender
  - Grade
  - Occupation
  - Smoking status etc.
    - **Nominal data** (No order)
      - Gender
      - Color.
      - City.
      - Occupation.
    - **Ordinal data.** (orders).
      - Grades.
      - Education level.
      - Customer satisfaction Rating.
- **Numerical.** (Quantitative)
  - Age.
  - height
  - weight
  - salary
  - marks etc.

This distinction is very important because it will affect
- How we visualize.
- What Statistical measure we calculate
- How we encode these data (columns) before modelling.



Height, weight, temp → values in a range.

* Uni-variate analysis

Categorical column
- Non-visual analysis
- Visual analysis

* Non-visual analysis

✓ Count of values in the columns
Count() or df.shape()

✓ Most frequent value → mode

✓ Number of unique value → nunique()

Before doing any kind of distribution analysis
How many unique categories are present.
- nunique → no. of distinct categories.

Technical term for no. of unique categorical values.

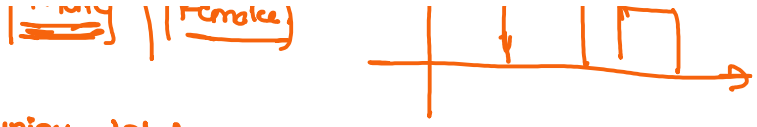
Cardinality.

High Cardinality → many unique values.
Low cardinality → few unique values.

✓ Distribution of categories → value_counts()

value_counts() is most important non-visual analysis for categorical value.

Titanic



✓ Actual unique labels — unique ().

Gender → [male and female]

Gender.nunique — 2.

Gender.unique — [male, female].

Gender.value_counts()

male

female

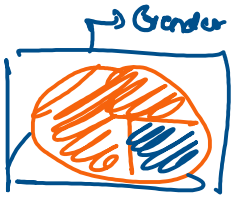
* Visual analysis

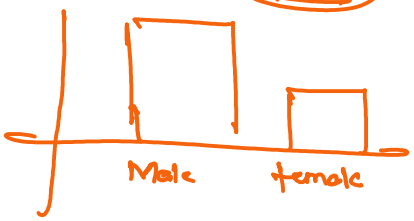
✓ count plot
✓ Bar plot } → frequency
✓ Pie chart → %

Univariate
- scatterplot
- 2 column.

value_count → count plot / bar plot → frequency.

↳ pie chart → %





- Numerical
  - Non-visual analysis

✓ count → how many value are there in the colum.

✓ Mean (Average) / Median. → average value

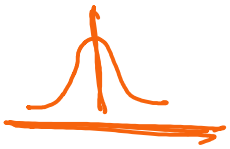
✓ min → minimum value

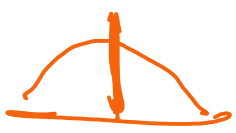
✓ max → maximum value

✓ standard deviation

$\sigma^2 = var = \frac{\sum (x-\mu)^2}{N}$ → square of unit

$std = \sqrt{\sigma^2} = \sigma$ → Better measure → same unit.





- A low standard deviation → values are close to/around the mean
- A high standard deviation → values are widely spread around the mean.

- We can also check

- Quantile / Percentile.
- Skewness
- Kurtosis.

95 percentile → means 95% of the total population who have attended the exam score lower than yours.

% → your score / total cs.

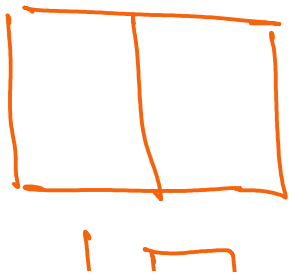
- IQR →
- MAD

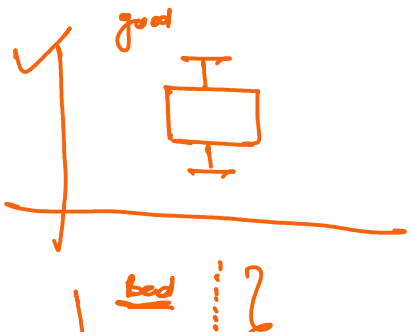
* Visualization for Numerical column

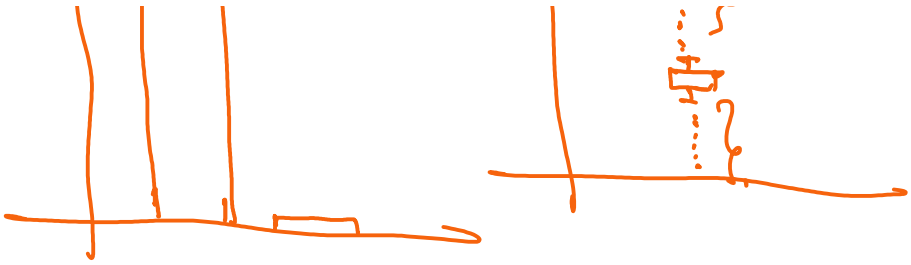
- Histogram. → frequency distribution
- KDE plot → smooth continuous probability distribution.
  - ↳ good for understanding skewness
- Box plot
  - ↳ Best for detecting.
    - Outliers
    - Median.
    - Quartiles
    - Spread (IQR).

WHY ?

- To understand the data. → help us to describe, summarize and understand the data.
- To learn about the distribution data.
  - Distribution tell us
    - Is the data is balanced or imbalanced?
    - Is it skewed?
    - Are there outliers.
    - Are the categories are evenly spread or highly dominated.
    - Is the data symmetric, bimodal, long tailed.
    - What is spread or range.
- Feature selection







- To detect outliers
  - Outliers needs to treated.
    - Mislead our mean.

$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3.$

$\frac{1+2+3+4+100}{5} = \frac{110}{5} = 22.$

- Affect ML model performance
- Create incorrect interpretation.

- To identify missing values

Reason to do Univariate analysis.

- Describe the data
- Learn data distribution
- Feature transformation
- Outlier detection
- Identify missing value
- Feature selection
- Business insight

Numerical → outliers. → boxplot

Categorical → outliers → countplot