DATA CLEANING AND PREPROCESSING USING PYTHON

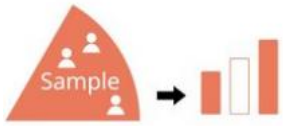
PRESENTED BY – REMYA R S

Analyzing Data for a Business

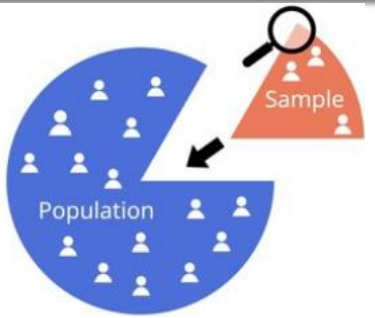
Understanding Problem Statement

Data Collection

Data Cleaning



Descriptive Analysis



Inferential Analysis

Data Visualization

Predictive Analysis

Data Cleaning

Steps

- Data Type Conversion
- Removing Duplicates
- Handling Outliers
- Handling Missing Data
  - For numerical column :
    - Using mean – For columns having no extreme outliers.
    - Using median – For columns having extreme outliers.
  - For categorical column :
    - Using mode

Descriptive Analysis

10,20,20,30,40,50,50,50,60

Count =9

Mean (Average):

Mean= (10+20+20+30+40+50+50+50+60) / 9= 36.67

Median, middle value of the sorted list = 40

Mode, value that appears most frequently =50

Maximum=60

Minimum=10

Descriptive Analysis

10,20,20,30,40,50,50,50,60

$$\text{Standard Deviation} = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

$$\mu = \frac{10 + 20 + 20 + 30 + 40 + 50 + 50 + 50 + 60}{9} = \frac{330}{9} \approx 36.67$$

N = 9

$$\sum(x_i - \mu)^2 = (10 - 36.67)^2 + (20 - 36.67)^2 + (20 - 36.67)^2 + (30 - 36.67)^2 + (4$$

$$= (-26.67)^2 + (-16.67)^2 + (-16.67)^2 + (-6.67)^2 + (3.33)^2 + (13.33)^2 + (13.33)^2 \cdot$$

$$= 711.11 + 277.78 + 277.78 + 44.45 + 11.09 + 177.69 + 177.69 + 177.69 + 544.45$$

$$= 2399.73$$

Standard deviation = 16.10

Descriptive Analysis

10,20,20,30,40,50,50,50,60

A box plot is a standardized way of displaying the distribution of data

Q2 → rep. middle value and 50 % of data lies below it

Q1 → 25 % of data lies below it

Q3 → 75 % of data lies below it


Thank
You