Assignment No:- 2

Problem Statement:

Perform the following operations using Python on the data sets:

a) Compute and display summary statistics for each feature available in the dataset. (e.g.minimum value, maximum value, mean, range, standard deviation, variance and percentiles)

b) Data Visualization-Create a histogram for each feature in the dataset to illustrate the feature distributions.

c) Data cleaning, Data integration, Data transformation, Data model building (e.g. Classification)

Software Library Package:

For this task, we'll use the following Python libraries:

- `pandas` for data manipulation and analysis.
- `matplotlib` for data visualization.
- `scikit-learn` for data cleaning, transformation, and model building.

Theory:

i) Methodology:

a) Computing Summary Statistics: Summary statistics provide a concise summary of the dataset's characteristics, including measures like mean, median, standard deviation, variance, and percentiles.

b) Data Visualization (Histograms): Histograms visualize the distribution of data values for each feature, aiding in understanding the data's distribution and identifying patterns.

c) Data Cleaning, Integration, Transformation, and Model Building: These steps are essential in the data analysis pipeline. Data cleaning involves handling missing values and outliers. Integration combines multiple datasets if applicable. Transformation involves scaling or encoding features. Model building refers to creating predictive models for tasks like classification.

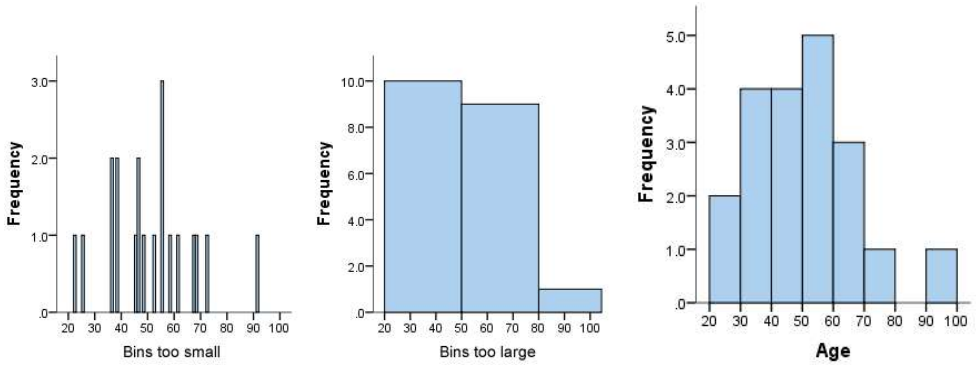

Fig 1: Sample Diagrams of Histograms

ii) Advantages and Applications and Limitations/Example:

- Advantages: Summary statistics and histograms provide quick insights into the dataset's characteristics and distribution. Data cleaning, integration, transformation, and model building are crucial for accurate analysis and prediction.
- Applications: These techniques are widely used in various fields such as finance, healthcare, marketing, etc., for tasks like risk assessment, customer segmentation, etc.
- Limitations/Example: Summary statistics may not capture the full complexity of data. Histograms may not reveal patterns in sparse data. Data cleaning and integration can be time-consuming. Incorrect transformation or model selection can lead to inaccurate results.

Working/Algorithm:

a) Computing Summary Statistics:

To compute summary statistics for each feature in the dataset:

- Calculate the minimum value, maximum value, mean, range, standard deviation, variance, and percentiles for each feature.
- Use statistical functions provided by libraries like `pandas` or `numpy` to compute these statistics.

b) Data Visualization - Creating Histograms:

To create histograms for each feature in the dataset:

- Use a plotting library such as `matplotlib` to generate histograms.
- Plot the frequency distribution of each feature by dividing the data into bins and counting the number of occurrences in each bin.
- Customize the histograms by adjusting bin width, color, labels, etc.

c) Data Cleaning, Integration, Transformation, and Model Building:

For data cleaning, integration, transformation, and model building:

- Data cleaning involves handling missing values, duplicates, and outliers.
- Data integration combines multiple datasets if applicable.
- Data transformation may include scaling numerical features, encoding categorical variables, or creating new features.
- Model building involves selecting an appropriate machine learning algorithm, training the model on the data, and evaluating its performance.

Diagram:

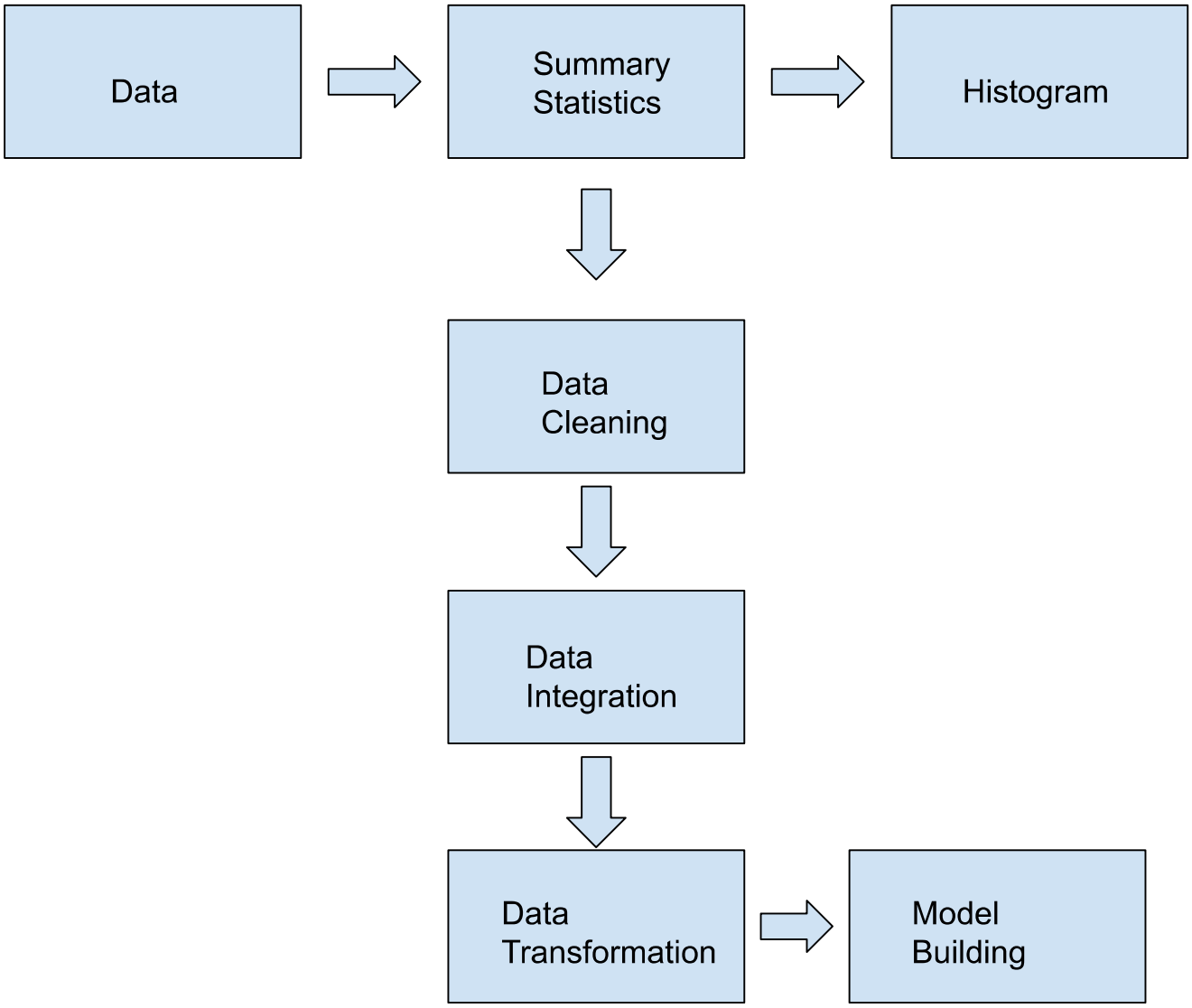

Fig 2: Workflow diagram

Conclusion:

Performing operations such as computing summary statistics, creating histograms for data visualization, and conducting data cleaning, integration, transformation, and model building are essential steps in the data analysis process. These steps help in gaining insights from data, identifying patterns, and building predictive models for various applications. Choosing appropriate methodologies and tools ensures efficient and accurate data analysis and model building.