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)

S/W Packages and Libraries used:

For the following assignment, the interpreter used was Google Collab and the Primary Library used was-

- Matplotlib: Matplotlib is a versatile plotting library in Python, offering a wide range of visualization capabilities, including histograms for illustrating feature distributions.

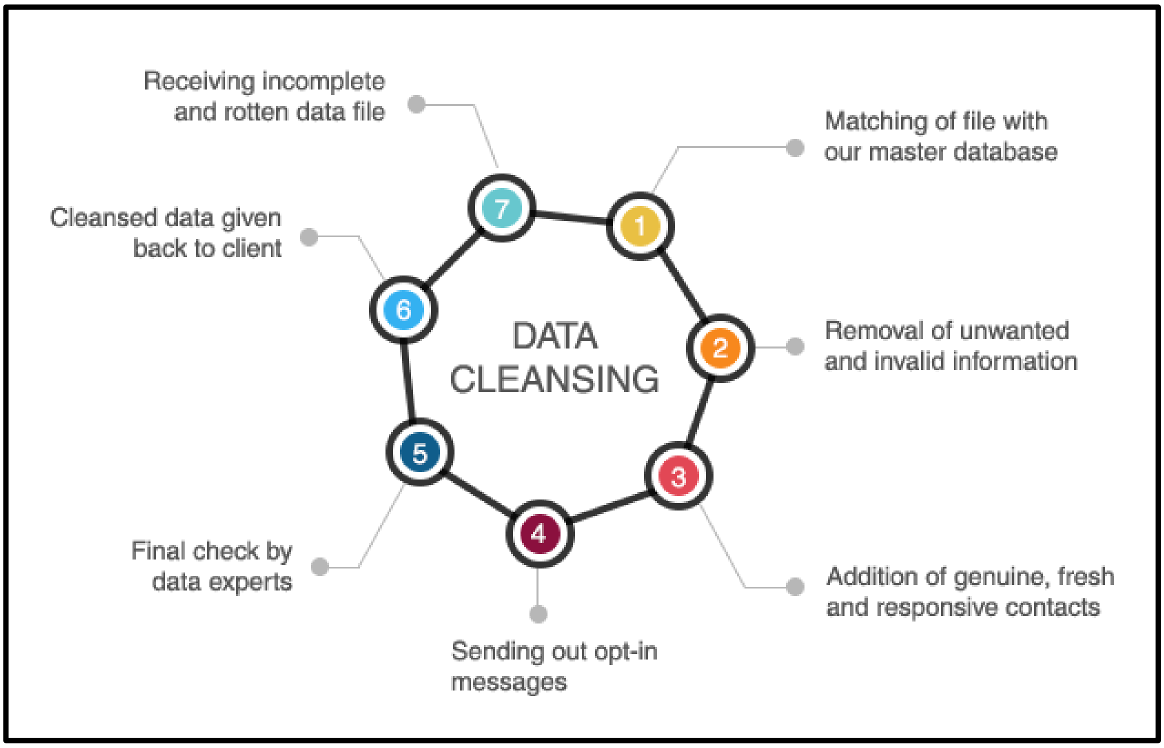
Theory-

- Data Visualization:
    - Data visualization is the graphical representation of information and data. It uses visual elements like charts, graphs, and maps to help viewers understand trends, patterns, and outliers in the data.

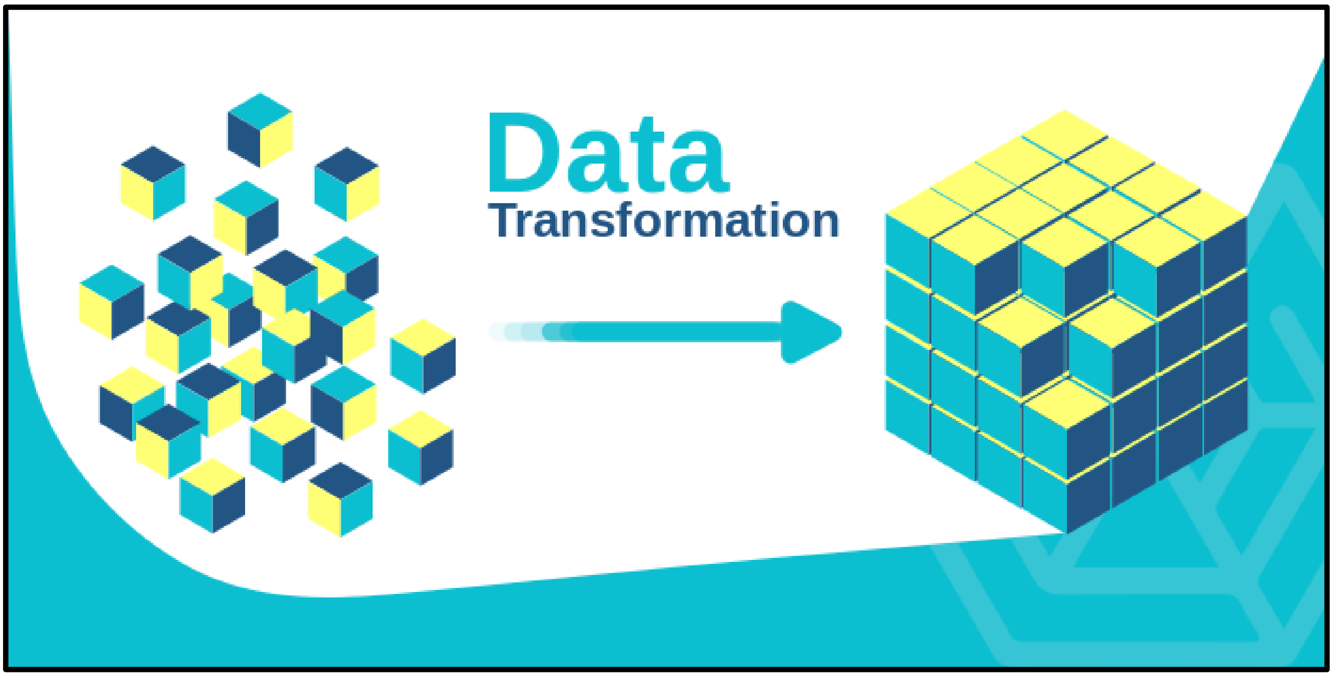
    - Choosing the appropriate visualization depends on the type of data and the insights you want to convey.

YoY Comparison
800
600
400
200
0
INFOGRARHIC
Sales Volume in First-half Year
500
400
300
200
100
0
Trending Products
500
400
300
200
100
0
Regional Completion Rate
40%
20%
60%
0%
80%
40%
20%
60%
0%
80%
Popular Products
500
400
300
200
100
0
Percentage of Products
24%
16%
60%
Regional Sales
TH
VIE
MAS
ID
KH
CN
SG
Regional Sales Completion Rate of Each Product
0%
25%
50%
75%
100%

- Data Cleaning:
    - Data cleaning, also known as data cleansing, is the process of identifying and correcting errors, inconsistencies, and inaccuracies in the dataset to improve its quality and reliability.


Receiving incomplete and rotten data file
Matching of file with our master database
Cleansed data given back to client
DATA CLEANSING
Removal of unwanted and invalid information
Final check by data experts
Addition of genuine, fresh and responsive contacts
Sending out opt-in messages
1
2
3
4
5
6
7

- Data Transformation:
    - Data transformation involves converting raw data into a suitable format for analysis or modeling.
    - It includes tasks such as normalization, aggregation, and feature engineering.



Methodology-

- For Computing Summary Statistics:
  - Utilize libraries like Pandas to load the dataset into a DataFrame.
  - Use Pandas' describe() method to generate summary statistics such as minimum, maximum, mean, standard deviation, variance, and percentiles for each feature.
- For Data Visualization - Histogram Creation:
  - Employ Matplotlib, a widely-used plotting library in Python, to create histograms for each feature.

    - Iterate through each feature in the dataset, plot its histogram using Matplotlib's hist() function, and customize the plot as necessary to illustrate feature distributions effectively.

- Data Cleaning:

    - Identify and handle missing values using techniques like imputation or removal.

    - Detect and handle outliers that may affect the integrity of the dataset.

- Data Integration:

    - Merge or concatenate multiple datasets if required for analysis.

    - Ensure consistency in data formats and representations across integrated datasets.

Applications:

- Descriptive Analysis: Summary statistics provide insights into the central tendency, variability, and distribution of features, aiding in descriptive analysis.

- Exploratory Data Analysis (EDA): Histograms visualize the distribution of individual features, revealing patterns, trends, and potential outliers.

Limitations:

- Data Quality: Incomplete or inaccurate data can lead to biased summary statistics and misleading visualizations.

- Model Performance: The success of classification models depends on factors like feature selection, model choice, and hyperparameter tuning, which may require iterative experimentation.

Conclusion:

By computing summary statistics, visualizing feature distributions, and executing data cleaning, integration, transformation, and model-building processes, it addresses essential aspects of exploratory data analysis and classification tasks. While offering valuable insights and enabling the construction of predictive models, it's important to acknowledge potential limitations such as data quality issues and computational constraints. Nevertheless, leveraging these libraries empowers practitioners to navigate datasets efficiently, derive meaningful insights, and develop robust classification models for informed decision-making.