Romesh Lulla D17C 31 ADS Exp 2

Aim: Apply data cleaning techniques (e.g. Data Imputation)

Theory:

Data Cleaning:

Data cleaning, also known as data cleansing or data preprocessing, is a crucial step in the data science pipeline that involves identifying and connecting or removing errors, inconsistencies, and inaccuracies in the data to improve its quality and usability.

Dealing with missing data is a prevalent and inherent challenge in data collection, particularly when dealing with extensive datasets. Numerous factors contribute to missing data, including participants providing incomplete information, non-responses from individuals who decide not to share data, poorly designed survey instruments, or the necessity to exclude data due to confidentiality concerns.

Importance of Data Cleaning

Data cleaning is essential because raw data is often noisy, incomplete, and inconsistent, which can negatively impact the accuracy and reliability of the insights derived from it. Inaccuracies, outliers, missing values, and inconsistencies can compromise the validity of analytical results. Moreover, clean data facilitates more effective modelling and pattern recognition, as algorithms perform optimally when fed high-quality, error-free input.

Steps to Perform Data Cleaning

Performing data cleaning involves a systematic process to identify and rectify errors, inconsistencies, and inaccuracies in a dataset. The following are essential steps to perform data cleaning:

1. Remove irrelevant data: First, figure out what analysis you'll be running and what your downstream needs are. What questions do you want to answer or problems do you want to solve? Remove unwanted observations from your dataset, including duplicate observations or irrelevant observations.

2. Deduplicate your data: Duplicate observations will happen most often during data collection. When you combine data sets from multiple places, scrape data, or receive data from clients or multiple departments, there are opportunities to create duplicate data.

3. Fix structural errors: Structural errors are when you measure or transfer data and notice strange naming conventions, typos, or incorrect capitalization.

4. Deal with missing data: This is where data imputation comes into play.

5. Filter out data outliers: Outliers can skew your analysis and lead to incorrect conclusions.

6. Validate your data: Ensure that your data cleaning has been successful and that your data is now ready for analysis.

Data Imputation

Data imputation is a method for retaining the majority of the dataset's data and information by substituting missing data with a different value4. These methods are employed because it would be impractical to remove data from a dataset each time5.

Importance of Data Imputation

Imputation is important because missing data can cause issues such as incompatibility with most of the Python libraries

Various Methods

1. Drop Missing Data:

    - Dropping Columns with NaN Values: This method involves removing entire columns that contain missing values. It is suitable when the missing values are concentrated in specific variables.

    - Dropping Rows with NaN Values: In this approach, rows with missing values are dropped, creating a dataset without any missing entries. This method is effective when the missing values are sporadic.

2. Simple Imputation Methods: Simple imputation methods replace missing values with summary statistics.

    - Mean Imputation: Replacing missing values with the mean of the respective column assumes a normal distribution and is effective when the missing values are missing completely at random (MCAR).

    - Median Imputation: Similar to mean imputation, this method replaces missing values with
the median of the column. It is more robust in the presence of outliers.

Conclusion:

In this experiment, we analyse the results by applying different methods on how to deal with missing data. In the first stage, we identified missing data and analysed the density in the data set visually and proportionally. We continued by coding how to deal with the detected missing data.

we analysed the results using dropping and simple imputation methods. Mean and median imputation for a variable with a high proportion of missing data gives biased results. Or it distorts the correlation with other variables.