Experiment 09: Implement different techniques and functions to clean unprocessed data including removing missing values, transforming incorrect data types

Learning Objective: Analyze different methods to preprocess data such as removing missing values or transforming incorrect data types.

Theory:

Removing Missing Values:

dropna(): This function is used to drop rows or columns with missing values from a DataFrame.
fillna(): It allows filling missing values with specified values such as mean, median, mode, or a custom value.
isnull(): This function returns a boolean DataFrame indicating where missing values are present, which can be used for further analysis or filtering.
notnull(): Similar to isnull(), but returns the opposite boolean DataFrame indicating non-missing values.

Transforming Incorrect Data Types:

astype(): This function is used to convert the data type of a column to a specified data type. For example, converting a column from object (string) to integer.
to_numeric(): Converts the values of a column to numeric data type. It has options to handle errors, such as coercing invalid values to NaN or raising an error.
to_datetime(): Converts a column to datetime data type, useful for handling date or timestamp data.
apply(): This function can apply a custom function to each element of a DataFrame or Series, enabling custom data type conversions or transformations.
Handling Duplicate Values:

duplicated(): This function identifies duplicate rows in a DataFrame based on specified columns.
drop_duplicates(): Removes duplicate rows from a DataFrame, optionally based on specified columns.

Data Normalization and Scaling:

StandardScaler: Scales features to have a mean of 0 and a standard deviation of 1.

MinMaxScaler: Scales features to a specified range, usually between 0 and 1.
RobustScaler: Scales features using robust statistics to handle outliers.
Data Imputation:

SimpleImputer: Provides simple strategies for imputing missing values, such as mean, median, most frequent, or a constant value.

Text Preprocessing:

Text cleaning: Removing special characters, punctuation, and unnecessary whitespace.
Tokenization: Splitting text into individual words or tokens.
Stopword removal: Removing common words that do not carry significant meaning.
Stemming or Lemmatization: Converting words to their base or root form.
Outlier Detection and Treatment:

Z-score: Identifying outliers based on their distance from the mean in terms of standard deviations.
Interquartile Range (IQR): Identifying outliers based on the difference between the third quartile (Q3) and the first quartile (Q1).
Winsorization: Capping or flooring extreme values to a specified percentile to mitigate the impact of outliers.

Implementation code:

Output:

Result and discussion:

Learning Outcomes: Students should have the ability to

LO 9.1: Implementing data cleaning techniques using libraries like pandas in Python improves your programming skills, particularly in data manipulation and data preprocessing tasks.

LO 9.2: Ability to clean unprocessed data requires critical thinking skills to evaluate the impact of different cleaning methods on the dataset and choose the most appropriate approach.

Course Outcomes: Understand and apply Data Importing and Metadata Management concepts.

Conclusion:

Viva Questions:

Q. 1 What are some common data quality issues that necessitate data cleaning?

Q.2 How can you identify incorrect data types in a dataset, and what methods can be used to transform them into the correct types?

Q.3 Can you discuss the role of data cleaning in ensuring data integrity and reliability?

-

For Faculty Use

Correction Parameters	**Formative Assessment [40%]**	**Timely completion of Practical [40%]**	**Attendance / Learning Attitude [20%]**	
Marks Obtained				