ASSIGNMENT NO. 1

Title: Perform the following operations using R/Python on suitable data sets:

a) read data from different formats (like csv, xls)

b) indexing and selecting data, sort data,

c) describe attributes of data, checking data types of each column,

d) counting unique values of data, format of each column, converting variable data type

(e.g. from long to short, vice versa),

e) identifying missing values and fill in the missing values

S/W Packages and H/W apparatus used: Linux OS: Ubuntu/Windows , Jupyter notebook.

Theory:

Data Preparation

Data preparation also referred as “data preprocessing” is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step prior to processing and often involves reformatting data, making corrections to data and the combining of data sets to enrich data.

Importance of Data preparation

● Because most machine learning algorithms require data to be structured in a specific way, datasets must be prepared before they can offer useful insights. a number of databases having missing, invalid, or otherwise difficult to process values for an algorithm. If you're looking for information, The algorithm will be unable to use it if it is missing. If the data is incorrect, the algorithm will produce inaccurate or even incorrect results. the outcomes are deceiving

● Some datasets are relatively clean but need to be shaped (e.g., aggregated or pivoted) and many datasets are just lacking useful

business context (e.g., poorly defined ID values), hence the need for feature enrichment. Good data preparation produces clean and well curated data which leads to more practical, accurate model outcomes.

● Before entering the data into the machine learning model, this is the most important step. The reason for this is that the data set must be unique and specific to the model, thus we must identify the data's required characteristics. The data preparation process provides a mechanism for preparing data for project definition as well as project evaluation of machine learning algorithms.

● There are a variety of predicting machine learning models available, each with its own method. However, some processes are common to all models, and they allow us to identify the underlying business problem and its solutions. The following are some of the data preparation procedures:

● 1. Determine the problems

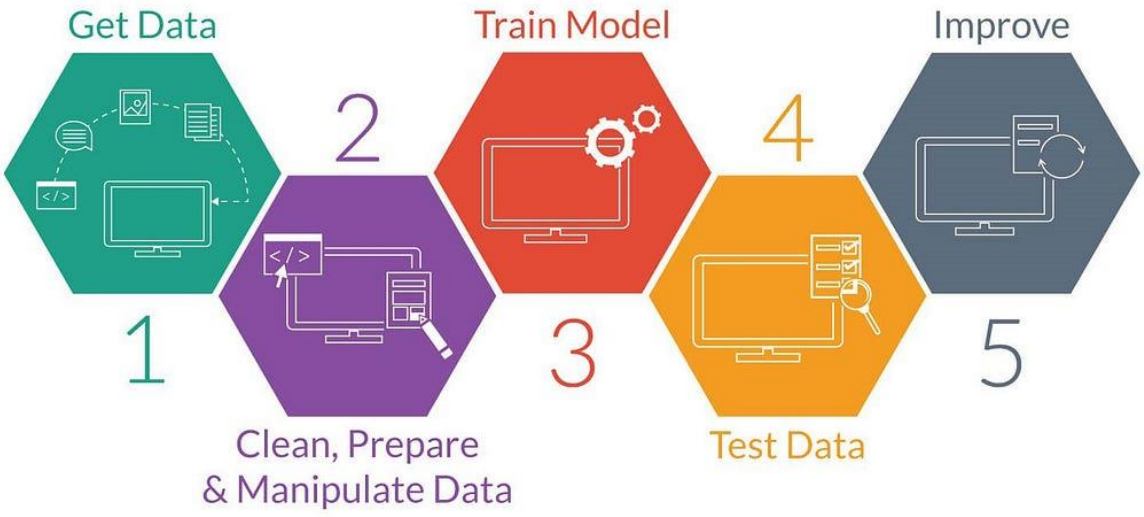
● 2. Data cleaning

● 3. Feature selection

● 4. Data transformation

● 5. feature engineering

● 6. Dimensionality reduction



a) Reading Data from Different Formats:

- CSV: Using `pd.read_csv()`
- Excel: Using `pd.read_excel()`

b) Indexing and Selecting Data, Sorting Data:

- Indexing: Using `.loc[]` and `.iloc[]`
- Selection: Using column names or numeric indices
- Sorting: Using `.sort_values()`

c) Describing Attributes of Data, Checking Data Types:

- Describing: Using `.describe()`
- Data Types: Using `.dtypes`

d) Counting Unique Values, Converting Variable Data Types:

- Unique Values: Using `.value_counts()`
- Converting Data Types: Using `.astype()`

e) Identifying and Handling Missing Values:

- Identifying: Using `.isna()` or `.isnull()`
- Handling: Using `.fillna()` or `.dropna()`

Working:

1. Read data from various formats using appropriate Pandas functions.
2. Perform indexing, selection, and sorting operations to extract relevant subsets of data.
3. Use descriptive statistics methods to understand the attributes of the data.
4. Count unique values and convert data types as needed.
5. Identify missing values and handle them by filling or removing them.
6. Utilize Pandas' functionalities to perform any additional data manipulation tasks required.

Conclusion:

Data preparation is recognized for helping businesses and analytics to get ready and prepare the data for operations. Pandas is a versatile tool for data manipulation and exploration in Python, offering numerous functionalities to handle various data preprocessing tasks efficiently. By leveraging its capabilities, users can streamline the process of preparing data for analysis and gain valuable insights from their datasets.