DATA LOADING AND PRE-PROCESSING :

The data loading and preprocessing process involves several important steps to prepare the data for analysis or machine learning tasks. Here's a detailed description of the preprocessing process:

1. Data Loading:

The code begins by loading the dataset from a CSV file named 'Data_Gov_Tamil_Nadu.csv' into a pandas DataFrame. The 'encoding' parameter is set to 'ISO-8859-1' to handle any character encoding issues that may exist in the CSV file. This step is essential to make the data accessible and ready for further processing.

2. Checking for Missing Values:

After loading the data, the code checks for missing values in the dataset. Missing values can be problematic for data analysis and machine learning, so identifying and addressing them is crucial.

The code uses the `isnull().sum()` method to calculate the number of missing values for each column in the DataFrame. This information is stored in the 'missing_values' variable, and it is printed to the console to provide an overview of which columns have missing data.

3. Filling Missing Values (Imputation):

- To handle missing values, the code focuses on two specific columns: 'LATEST_YEAR_ANNUAL_RETURN' and 'LATEST_YEAR_FINANCIAL_STATEMENT.' Since these columns are time-related, it's essential to address missing values appropriately.
- The code employs a method called "linear interpolation" using the `interpolate()` method. This method estimates missing values based on the values of adjacent data points. In this case, missing values are replaced with interpolated values that follow a linear trend.
- The `inplace=True` argument is used, indicating that the changes should be applied directly to the DataFrame, ensuring that missing values are filled in these two columns.

4. Converting Categorical Features to Numerical:

- Many machine learning algorithms require input data to be in numerical format. Categorical features, which represent non-numeric data such as labels or categories, need to be converted into a numerical form.
- In this code, three categorical features, 'COMPANY_CLASS,' 'COMPANY_CATEGORY,' and 'COMPANY_SUB_CATEGORY,' are converted into numerical representations using Label Encoding.
- A `LabelEncoder` instance is created for each feature, and the `fit_transform()` method is applied to encode the categorical data into a numeric format. The resulting numeric values are stored in the same columns in the DataFrame, replacing the original categorical data.

5. Displaying Descriptive Statistics (After Label Encoding):

- After the data preprocessing steps are completed, the code displays descriptive statistics for the DataFrame. This summary includes statistics such as mean, standard deviation, minimum, maximum, and quartile values for each numerical feature in the dataset.
- This summary provides insights into the central tendencies and distributions of the data, both before and after the preprocessing steps.

Overall, the preprocessing process is a critical step in data analysis and machine learning projects. It involves handling missing data, converting categorical data to a numerical format, and ensuring the data is in a suitable state for further analysis or predictive modeling. These steps help improve the quality and readiness of the data for downstream tasks.