**PHASE – 2**

**Data Preprocessing**

**AI Driven Exploration:**

|  |  |
| --- | --- |
| **Name** | **Jeevitha. J** |
| **Date** | **09/10/2023** |
| **Team ID** | **Proj-2121-Team(4)** |
| **Project Name** | **AI Driven Exploration** |

**Program with Explanation:**

**Importing Libraries:**

import pandas as pd import pandas as np

* Here, you are importing the pandas library with the alias "pd," which is a common practice. However, you also attempted to import pandas with the alias "np," which is usually used for NumPy, another popular Python library. It's better to use "pd" consistently for pandas.

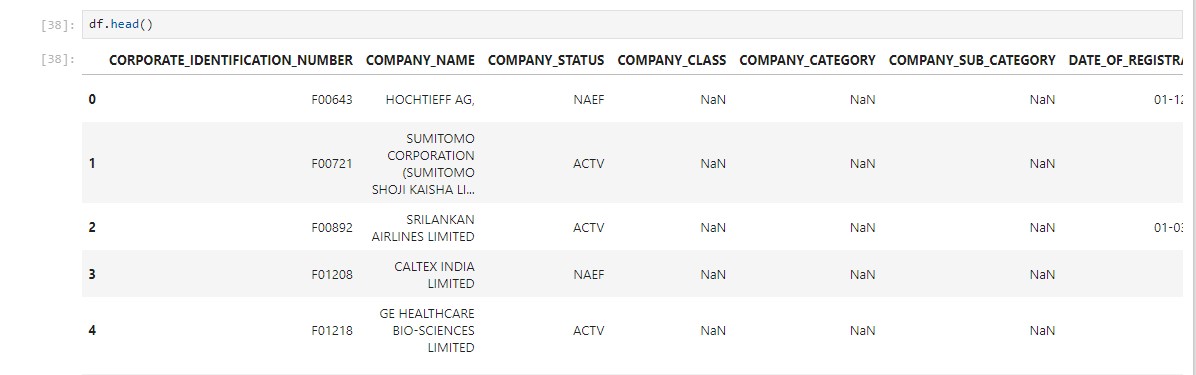
**Loading Data form CSV file:**

df = pd.read\_csv('C:\\Users\\win10\\Desktop\\Data\_Gov\_Tamil\_Nadu.csv', encoding='latin-1')

* This code reads data from a CSV file located at the specified path and stores it in a pandas DataFrame called df. The encoding='latin-1' parameter is used to specify the character encoding of the file.

**Displaying the First Rows of the DataFrame:**

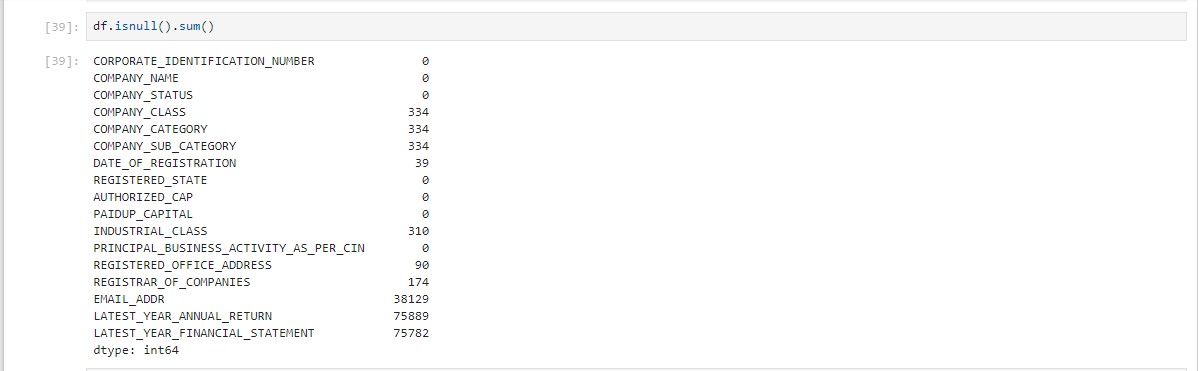
df.head()



* This line of code displays the first few rows of the DataFrame df to inspect its contents.

**Checking for Missing Values:**

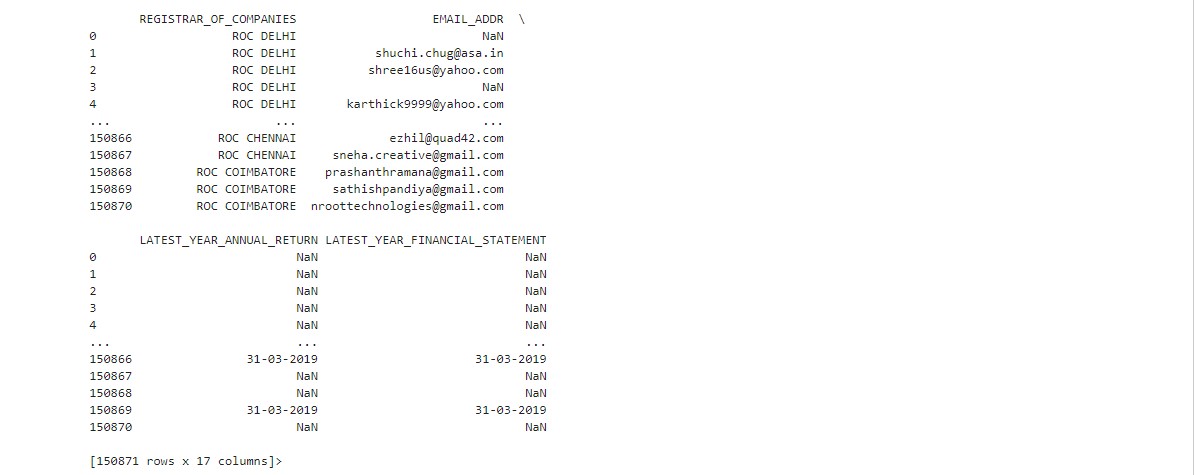
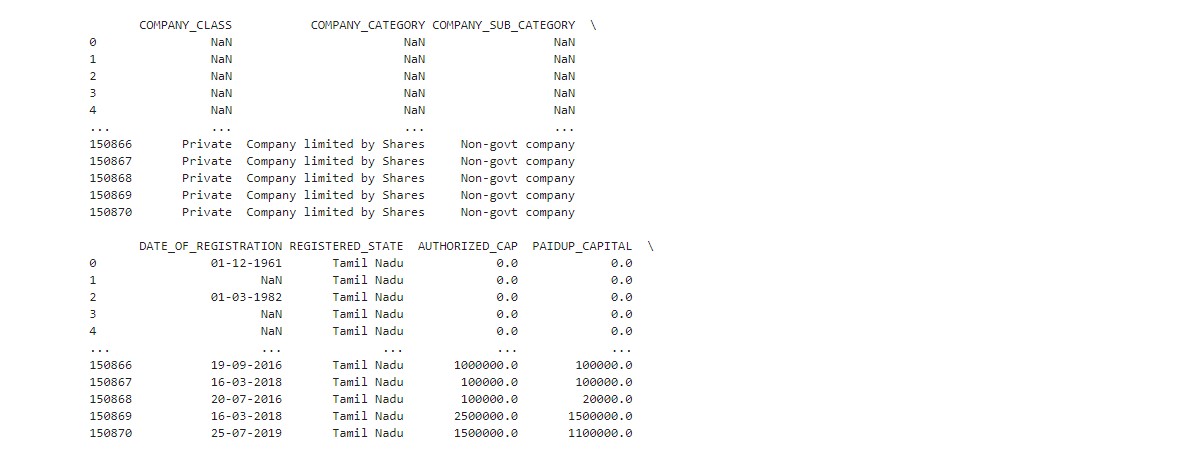
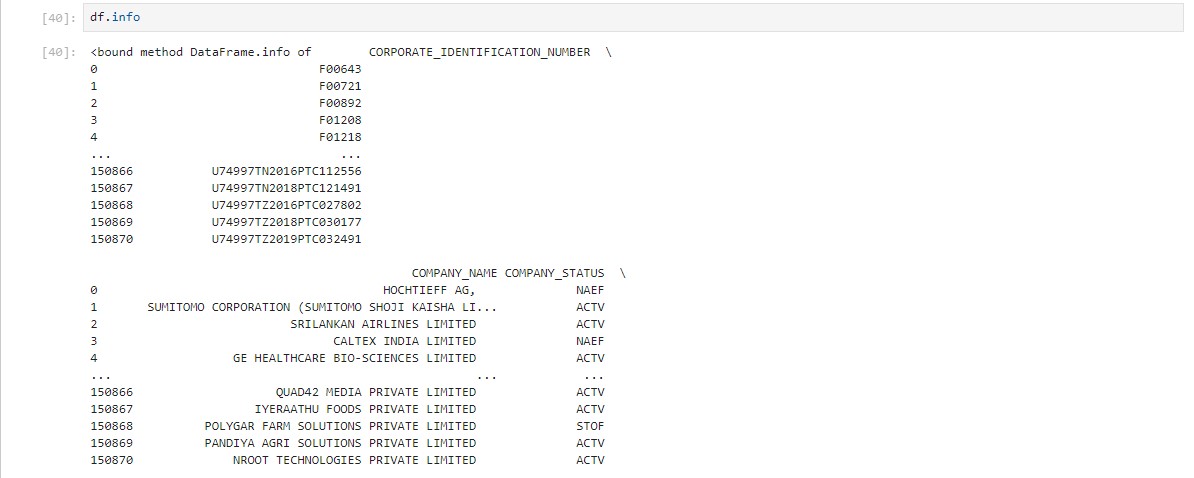
df.isnull().sum()



* Here, you are checking for missing values (NaN) in each column of the DataFrame df. The isnull().sum() function counts the number of missing values in each column.

**Displaying DataFrame Information:**

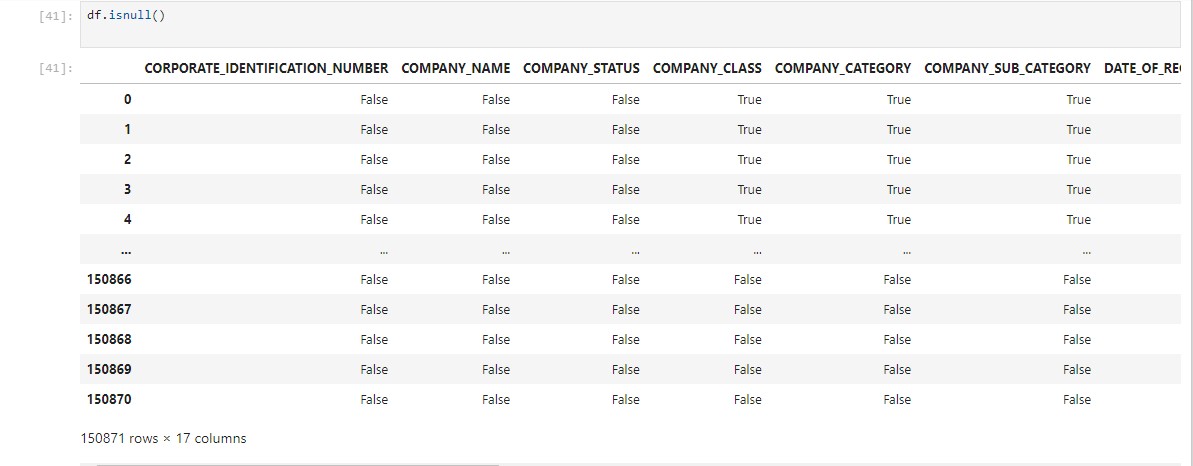
df.info



* This line of code attempts to display information about the DataFrame. However, it should be corrected to df.info() (with parentheses) to call the info() method.

**Checking for Missing Values (Again):**

df.isnull()

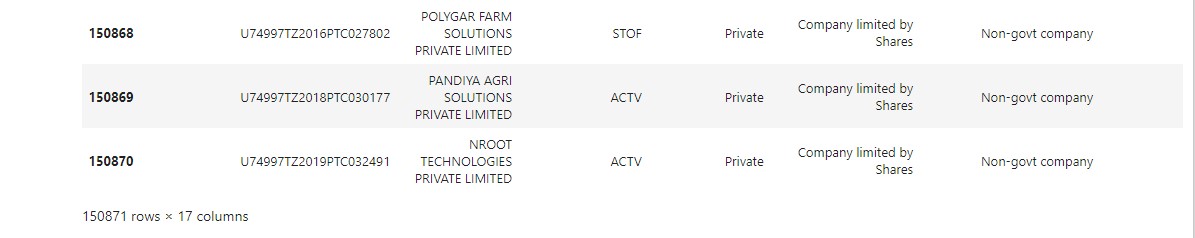


* Similar to the previous check, this code checks for missing values in the entire DataFrame. It returns a DataFrame of Boolean values indicating whether each element is missing or not.

**Filling Missing Values:**

df.fillna({'COMPANY\_CLASS': 'Private', 'COMPANY\_CATEGORY': 'Company limited by

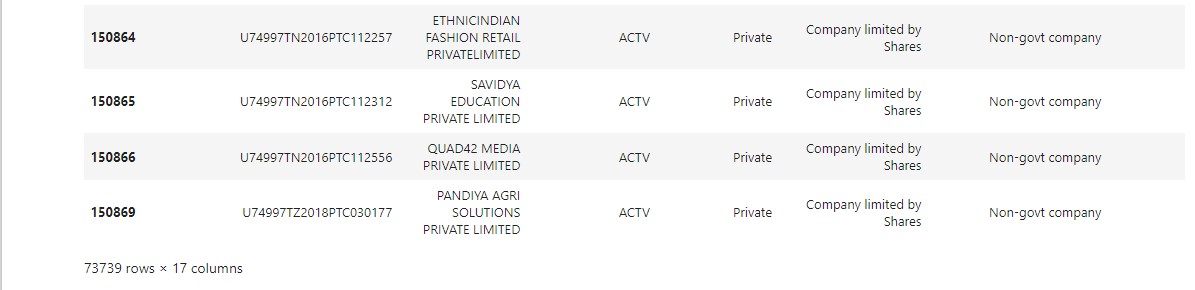
Shares', 'COMPANY\_SUB\_CATEGORY': 'Non-govt company'})



* This line attempts to fill missing values in specific columns ('COMPANY\_CLASS', 'COMPANY\_CATEGORY', 'COMPANY\_SUB\_CATEGORY') with predefined values. However, it doesn't modify the original DataFrame. You should assign the result back to df for the changes to take effect.

**Dropping Rows with Missing Values:**

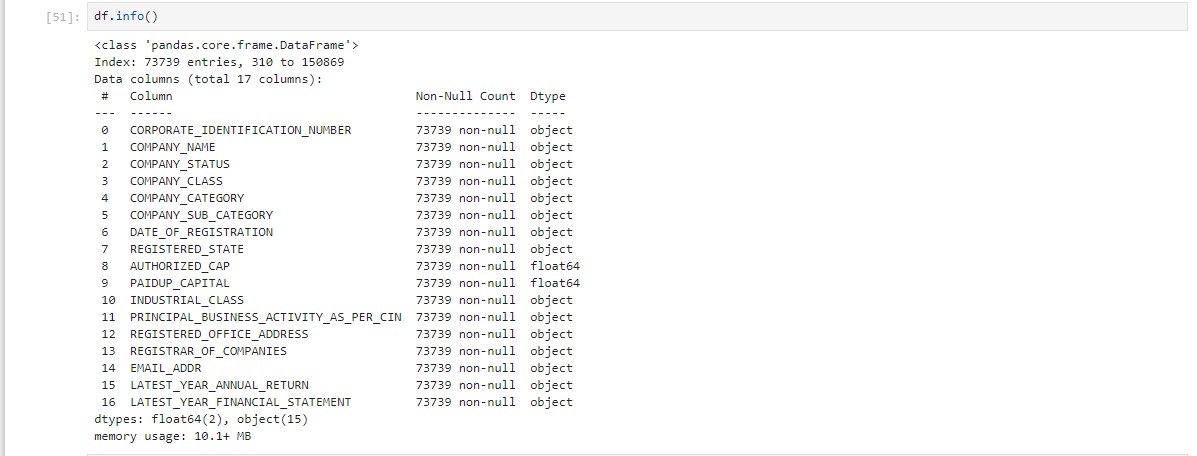
df.dropna(axis=0)



* This line attempts to drop rows with missing values from the DataFrame, but it doesn't modify the original DataFrame. You should assign the result back to df if you want to keep the changes.

**Displaying DataFrame Information (Again):**

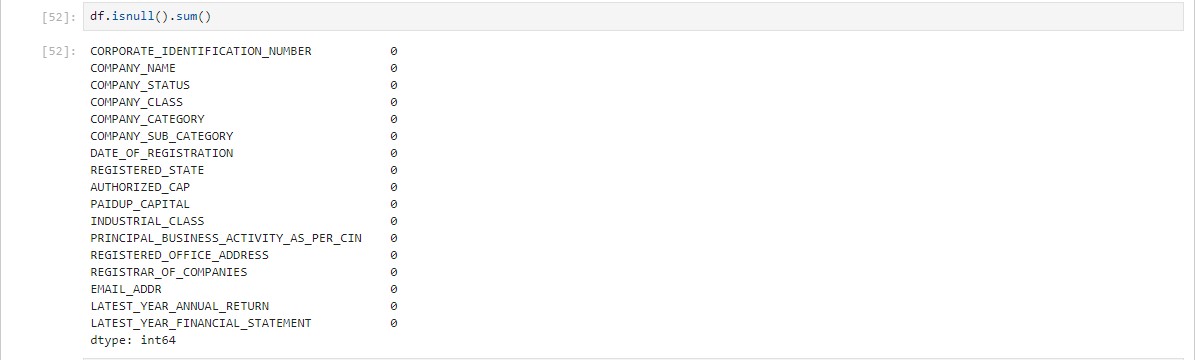
df.info()



* This line correctly displays information about the DataFrame, including data types and non-null counts.

**Checking for Missing Values (Once More):**

df.isnull().sum()



* This line checks for missing values again and displays the count of missing values in each column. However, this will still show the original DataFrame with missing values since steps 7 and 8 did not modify it.
* To summarize, you should make sure to assign the results of operations like filling missing values or dropping rows back to the DataFrame df if you want to apply those changes to the original data.

* Note that some of the operations like 'fillna' and 'dropna' don't modify the

DataFrame in place unless you reassign it as shown in the comments above.