PHASE – 2

Data Preprocessing

AI Driven Exploration:

Name	Banu.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()

```
df.head()
```

	CORPORATE_IDENTIFICATION_NUMBER	COMPANY_NAME	COMPANY_STATUS	COMPANY_CLASS	COMPANY_CATEGORY	COMPANY_SUB_CATEGORY	DATE_OF_REGISTR
0	F00643	HOCHTIEF AG	NAEF	NaN	NaN	NaN	01-1
1	F00721	SUMITOMO CORPORATION (SUMITOMO SHOJI KAISHA LL	ACTV	NaN	NaN	NaN	
2	F00802	SRILANKAN AIRLINES LIMITED	ACTV	NaN	NaN	NaN	01-0
3	F01208	CALTEX INDIA LIMITED	NAEF	NaN	NaN	NaN	
4	F01218	GE HEALTHCARE BIO-SCIENCES LIMITED	ACTV	NaN	NaN	NaN	

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

Checking for Missing Values:

df.isnull().sum()

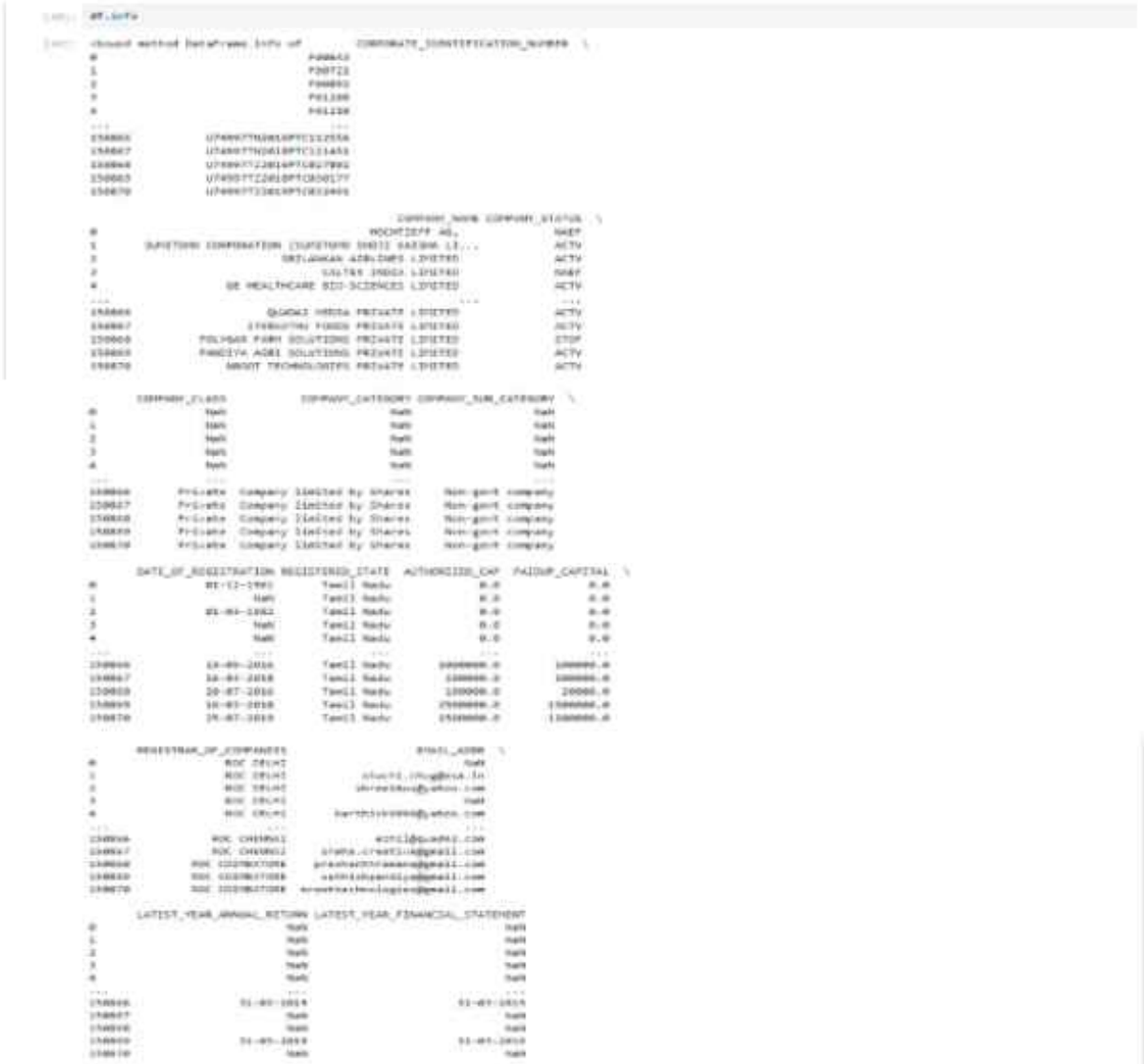
```
[19]: df.isnull().sum()

[19]: CORPORATE_IDENTIFICATION_NUMBER            0
      COMPANY_NAME                               0
      COMPANY_STATUS                             0
      COMPANY_CLASS                            334
      COMPANY_CATEGORY                         334
      COMPANY_SUB_CATEGORY                     334
      DATE_OF_REGISTRATION                      59
      REGISTERED_STATE                           0
      AUTHORIZED_CAP                             0
      PAIDUP_CAPITAL                             0
      INDUSTRIAL_CLASS                         310
      PRINCIPAL_BUSINESS_ACTIVITY_AS_PER_CIN     0
      REGISTERED_OFFICE_ADDRESS                 99
      REGISTRAR_OF_COMPANIES                   174
      EMAIL_ADDR                             38129
      LATEST_YEAR_ANNUAL_RETURN              75869
      LATEST_YEAR_FINANCIAL_STATEMENT        75782
      dtype: int64
```

- 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()

```
df.isnull()
```

	CORPORATE_IDENTIFICATION_NUMBER	COMPANY_NAME	COMPANY_STATUS	COMPANY_CLASS	COMPANY_CATEGORY	COMPANY_SUB_CATEGORY	DATE_OF_RE
0	False	False	False	True	True	True	
1	False	False	False	True	True	True	
2	False	False	False	True	True	True	
3	False	False	False	True	True	True	
4	False	False	False	True	True	True	
...	...	...	...	...	...	...	
150866	False	False	False	False	False	False	
150867	False	False	False	False	False	False	
150868	False	False	False	False	False	False	
150869	False	False	False	False	False	False	
150870	False	False	False	False	False	False	

150871 rows × 17 columns

- 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'})

```
df.fillna({'COMPANY_CLASS':'Private','COMPANY_CATEGORY':'Company limited by Shares','COMPANY_SUB_CATEGORY':'Non-govt company'})
```

	CORPORATE_IDENTIFICATION_NUMBER	COMPANY_NAME	COMPANY_STATUS	COMPANY_CLASS	COMPANY_CATEGORY	COMPANY_SUB_CATEGORY	DATE_OF_RE
0	[illegible]	HOCHTIEF AG	NAEF	Private	Company limited by Shares	Non-govt company	
1	[illegible]	SUMITOMO CORPORATION (SUMITOMO SHOJI KAISHA LI...	ACTV	Private	Company limited by Shares	Non-govt company	
2	[illegible]	SRILANKAN AIRLINES LIMITED	ACTV	Private	Company limited by Shares	Non-govt company	
3	[illegible]	CALTEX INDIA LIMITED	NAEF	Private	Company limited by Shares	Non-govt company	
4	[illegible]	GE HEALTHCARE BIO-SCIENCES LIMITED	ACTV	Private	Company limited by Shares	Non-govt company	
...	...	...	...	...	...	...	
150866	U74999TN2019PTC112559	QUAD42 MEDIA PRIVATE LIMITED	ACTV	Private	Company limited by Shares	Non-govt company	
150867	U74999TN2019PTC121461	VEMARTHU FOODS PRIVATE LIMITED	ACTV	Private	Company limited by Shares	Non-govt company	
150868	U74999TZ2019PTC027802	POLYGAR FARM SOLUTIONS PRIVATE LIMITED	STOF	Private	Company limited by Shares	Non-govt company	
150869	U74999TZ2019PTC030177	PANDIYA AGRI SOLUTIONS PRIVATE LIMITED	ACTV	Private	Company limited by Shares	Non-govt company	
150870	U74999TZ2019PTC032491	NROOT TECHNOLOGIES PRIVATE LIMITED	ACTV	Private	Company limited by Shares	Non-govt company	

150871 rows × 17 columns

- 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)

```
df.dropna(axis=0)
```

	CORPORATE_IDENTIFICATION_NUMBER	COMPANY_NAME	COMPANY_STATUS	COMPANY_CLASS	COMPANY_CATEGORY	COMPANY_SUB_CATEGORY	DATE_OF_RI
310	U01117TZ1943PLC000117	NELLAMALAI AGRO INDUSTRIES LIMITED	ACTV	Public	Company limited by Shares	non-govt company	
311	U01100TN1988PLC013473	ABAN OFFSHORE LIMITED	ACTV	Public	Company limited by Shares	non-govt company	
313	U21139TN1992PLC024278	SOFTECH INFINIUM SOLUTIONS LIMITED	ACTV	Public	Company limited by Shares	non-govt company	
315	U01122TZ1989PLC010782	RODHIRAJU INDUSTRIES LIMITED	ACTV	Public	Company limited by Shares	non-govt company	
318	U01132TZ1922PLC000234	THE UNITED NILGIRI TEA ESTATES COMPANY LIMITED	ACTV	Public	Company limited by Shares	non-govt company	
...	...	...	...	...	...	...	
150862	U74993TN2018PTC112109	MIXIT COMMUNICATIONS PRIVATE LIMITED	ACTV	Private	Company limited by Shares	non-govt company	
150864	U74993TN2018PTC112257	ETHNICINDIAN FASHION RETAIL PRIVATE LIMITED	ACTV	Private	Company limited by Shares	non-govt company	
150864	U74993TN2018PTC112257	ETHNICINDIAN FASHION RETAIL PRIVATE LIMITED	ACTV	Private	Company limited by Shares	non-govt company	
150865	U74993TN2018PTC112312	SAVIDYA EDUCATION PRIVATE LIMITED	ACTV	Private	Company limited by Shares	non-govt company	
150866	U74993TN2018PTC112556	QUADAS MEDIA PRIVATE LIMITED	ACTV	Private	Company limited by Shares	non-govt company	
150869	U74993TZ2018PTC030177	RASIYA AGRI SOLUTIONS PRIVATE LIMITED	ACTV	Private	Company limited by Shares	non-govt company	

72725 rows × 17 columns

- 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()

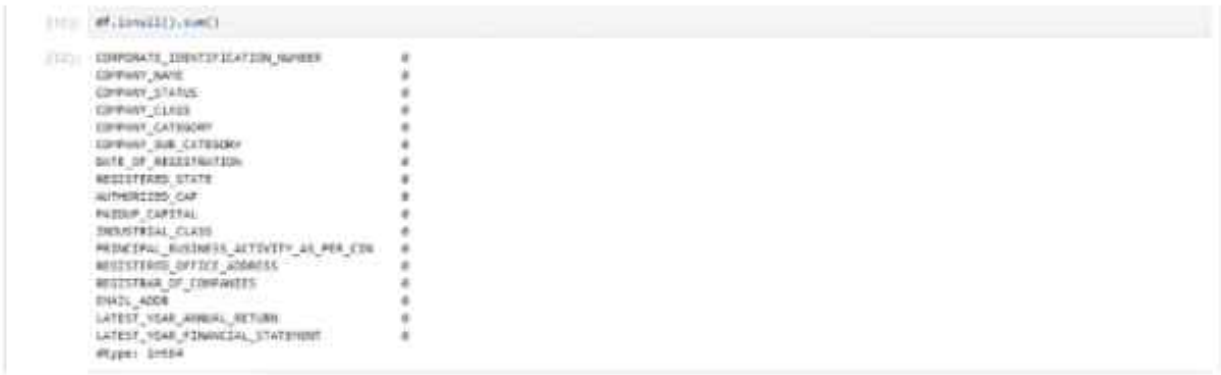
```
[11]: df.info()

<class 'pandas.core.frame.DataFrame'>
Index: 73739 entries, 318 to 258888
Data columns (total 17 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   CORPORATE_IDENTIFICATION_NUMBER         73734 non-null  object
 1   COMPANY_NAME                            73734 non-null  object
 2   COMPANY_STATUS                          73733 non-null  object
 3   COMPANY_CLASS                           73733 non-null  object
 4   COMPANY_CATEGORY                        73733 non-null  object
 5   COMPANY_SUB_CATEGORY                    73733 non-null  object
 6   DATE_OF_REGISTRATION                    73733 non-null  object
 7   REGISTERED_STATE                        73733 non-null  object
 8   AUTHORIZED_CAP                          73734 non-null  float64
 9   PAIDUP_CAPITAL                          73734 non-null  float64
 10  INDUSTRIAL_CLASS                        73733 non-null  object
 11  PRINCIPAL_BUSINESS_ACTIVITY_AS_PER_CIN  73733 non-null  object
 12  REGISTERED_OFFICE_ADDRESS               73733 non-null  object
 13  REGISTRAR_OF_COMPANIES                  73734 non-null  object
 14  EMAIL_ADDR                              73734 non-null  object
 15  LATEST_YEAR_ANNUAL_RETURN               73733 non-null  object
 16  LATEST_YEAR_FINANCIAL_STATEMENT         73733 non-null  object
dtypes: float64(2), object(15)
memory usage: 10.1+ MB
```

- 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.