PHASE – 2

Data Preprocessing

AI Driven Exploration:

Name	Deva Dharshini.A
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.

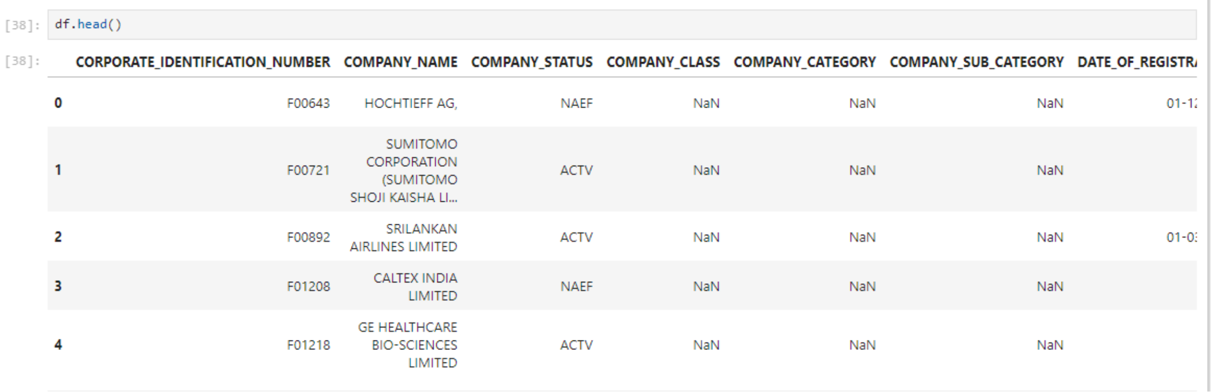
Loading Data form CSV file:

```
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()



```
[38]: df.head()
```

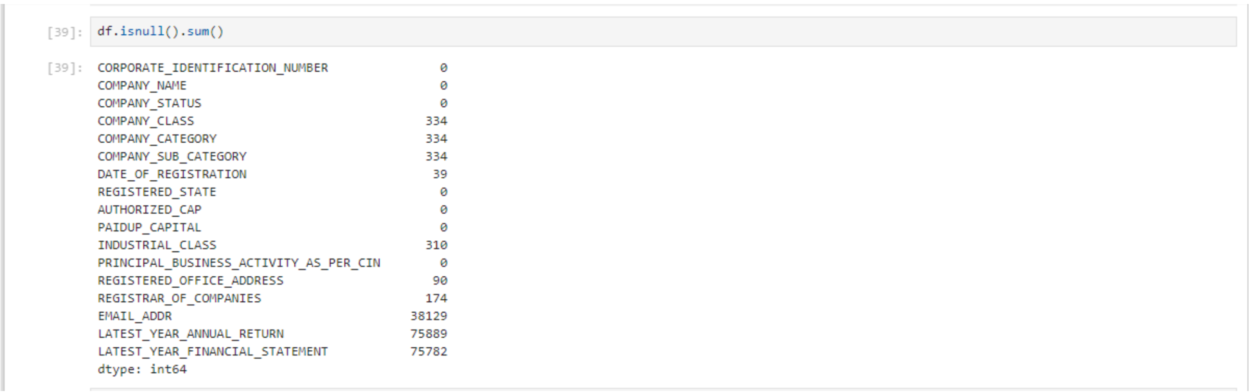
[38]:

	CORPORATE_IDENTIFICATION_NUMBER	COMPANY_NAME	COMPANY_STATUS	COMPANY_CLASS	COMPANY_CATEGORY	COMPANY_SUB_CATEGORY	DATE_OF_REGISTRA
0	F00643	HOCHTIEFF AG,	NAEF	NaN	NaN	NaN	01-1:
1	F00721	SUMITOMO CORPORATION (SUMITOMO SHOJI KAISHA LI...	ACTV	NaN	NaN	NaN	
2	F00892	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()

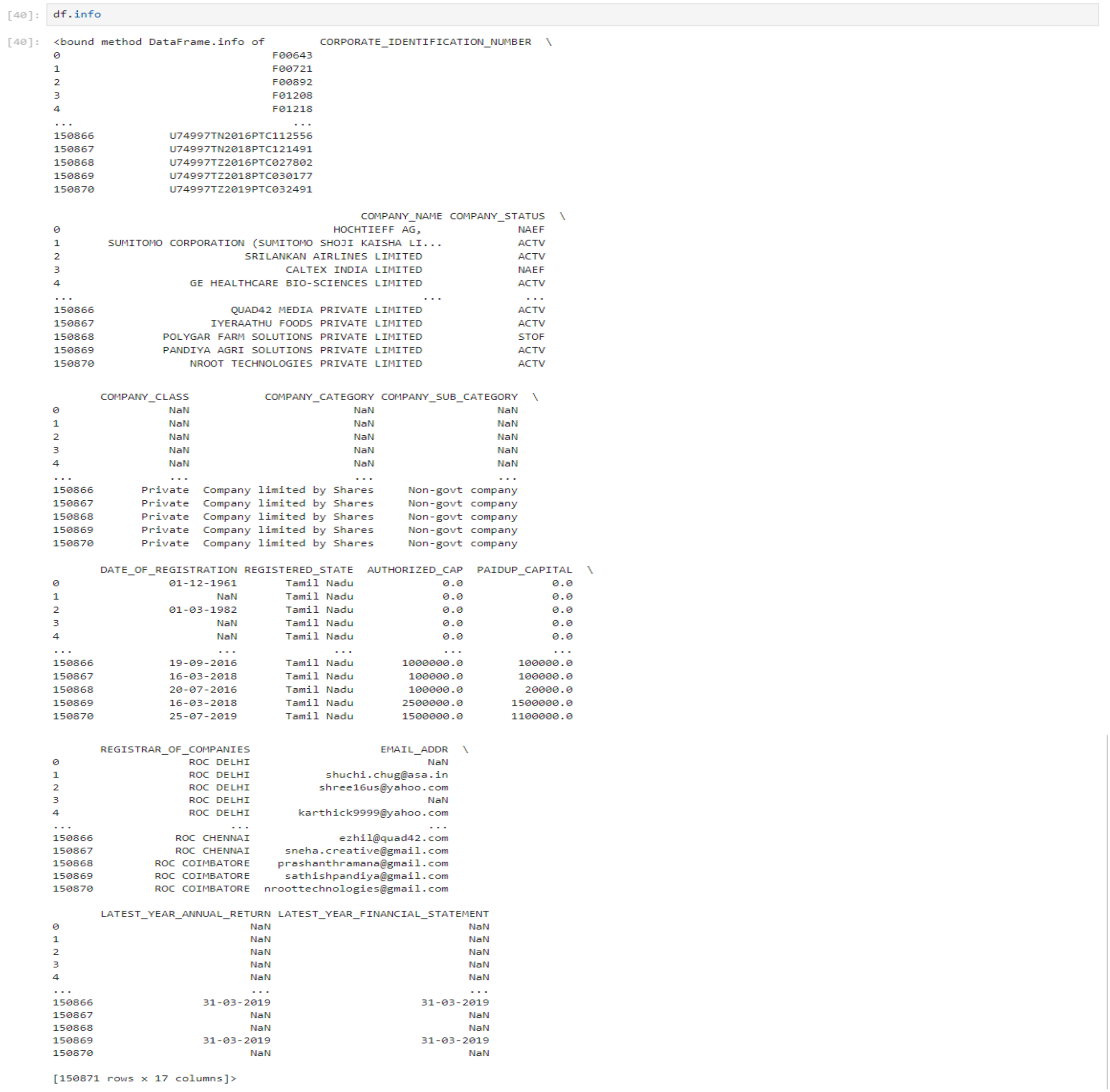


```
[39]: df.isnull().sum()

[39]: CORPORATE_IDENTIFICATION_NUMBER              0
      COMPANY_NAME                                 0
      COMPANY_STATUS                               0
      COMPANY_CLASS                              334
      COMPANY_CATEGORY                           334
      COMPANY_SUB_CATEGORY                       334
      DATE_OF_REGISTRATION                        39
      REGISTERED_STATE                             0
      AUTHORIZED_CAP                               0
      PAIDUP_CAPITAL                               0
      INDUSTRIAL_CLASS                           310
      PRINCIPAL_BUSINESS_ACTIVITY_AS_PER_CIN       0
      REGISTERED_OFFICE_ADDRESS                   90
      REGISTRAR_OF_COMPANIES                     174
      EMAIL_ADDR                               38129
      LATEST_YEAR_ANNUAL_RETURN                75889
      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



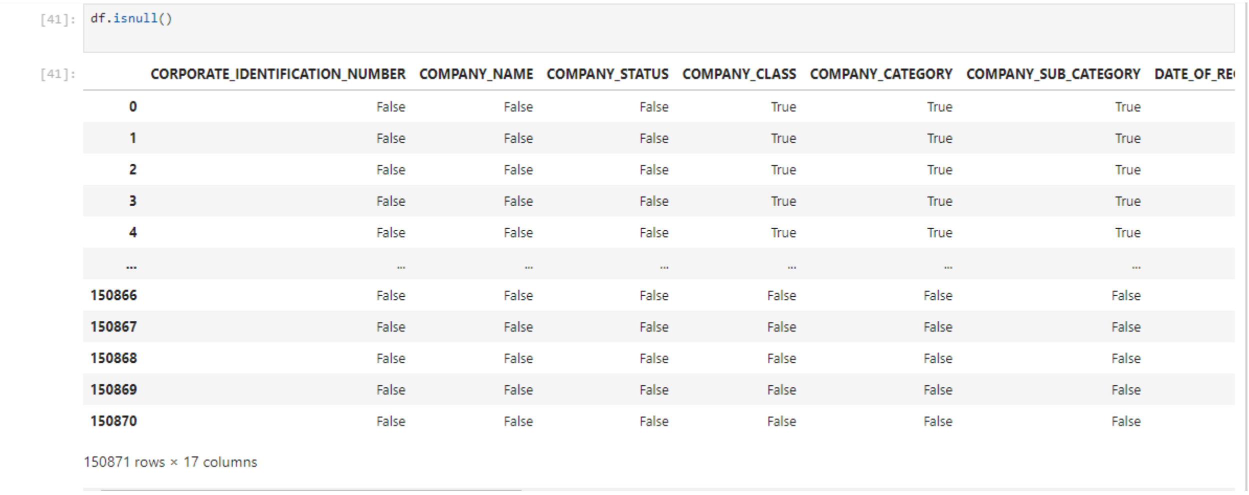
```
[40]: df.info
```

```
[40]: <bound method DataFrame.info of        CORPORATE_IDENTIFICATION_NUMBER  \
      0                               F00643
      1                               F00721
      2                               F00892
      3                               F01208
      4                               F01218
      ...                                ...
      150866           U74997TN2016PTC112556
      150867           U74997TN2018PTC121491
      150868           U74997TZ2016PTC027802
      150869           U74997TZ2018PTC030177
      150870           U74997TZ2019PTC032491

                                                   COMPANY_NAME COMPANY_STATUS  \
      0                                        HOCHTIEFF AG,              NAEF
      1       SUMITOMO CORPORATION (SUMITOMO SHOJI KAISHA LI...           ACTV
      2                           SRILANKAN AIRLINES LIMITED              ACTV
      3                                 CALTEX INDIA LIMITED              NAEF
      4                   GE HEALTHCARE BIO-SCIENCES LIMITED              ACTV
      ...                                                   ...            ...
      150866                    QUAD42 MEDIA PRIVATE LIMITED              ACTV
      150867                 IYERAATHU FOODS PRIVATE LIMITED              ACTV
      150868          POLYGAR FARM SOLUTIONS PRIVATE LIMITED              STOF
      150869          PANDIYA AGRI SOLUTIONS PRIVATE LIMITED              ACTV
      150870              NROOT TECHNOLOGIES PRIVATE LIMITED              ACTV

             COMPANY_CLASS           COMPANY_CATEGORY COMPANY_SUB_CATEGORY  \
      0                NaN                        NaN                  NaN
      1                NaN                        NaN                  NaN
      2                NaN                        NaN                  NaN
      3                NaN                        NaN                  NaN
      4                NaN                        NaN                  NaN
      ...              ...                        ...                  ...
      150866       Private  Company limited by Shares     Non-govt company
      150867       Private  Company limited by Shares     Non-govt company
      150868       Private  Company limited by Shares     Non-govt company
      150869       Private  Company limited by Shares     Non-govt company
      150870       Private  Company limited by Shares     Non-govt company

             DATE_OF_REGISTRATION REGISTERED_STATE  AUTHORIZED_CAP  PAIDUP_CAPITAL  \
      0                01-12-1961       Tamil Nadu             0.0             0.0
      1                       NaN       Tamil Nadu             0.0             0.0
      2                01-03-1982       Tamil Nadu             0.0             0.0
      3                       NaN       Tamil Nadu             0.0             0.0
      4                       NaN       Tamil Nadu             0.0             0.0
      ...                     ...              ...             ...             ...
      150866           19-09-2016       Tamil Nadu       1000000.0        100000.0
      150867           16-03-2018       Tamil Nadu        100000.0        100000.0
      150868           20-07-2016       Tamil Nadu        100000.0         20000.0
      150869           16-03-2018       Tamil Nadu       2500000.0       1500000.0
      150870           25-07-2019       Tamil Nadu       1500000.0       1100000.0

             REGISTRAR_OF_COMPANIES                   EMAIL_ADDR  \
      0                   ROC DELHI                          NaN
      1                   ROC DELHI           shuchi.chug@asa.in
      2                   ROC DELHI          shree16us@yahoo.com
      3                   ROC DELHI                          NaN
      4                   ROC DELHI       karthick9999@yahoo.com
      ...                       ...                          ...
      150866            ROC CHENNAI             ezhil@quad42.com
      150867            ROC CHENNAI     sneha.creative@gmail.com
      150868         ROC COIMBATORE    prashanthramana@gmail.com
      150869         ROC COIMBATORE     sathishpandiya@gmail.com
      150870         ROC COIMBATORE  nroottechnologies@gmail.com

             LATEST_YEAR_ANNUAL_RETURN LATEST_YEAR_FINANCIAL_STATEMENT
      0                            NaN                             NaN
      1                            NaN                             NaN
      2                            NaN                             NaN
      3                            NaN                             NaN
      4                            NaN                             NaN
      ...                          ...                             ...
      150866                31-03-2019                      31-03-2019
      150867                       NaN                             NaN
      150868                       NaN                             NaN
      150869                31-03-2019                      31-03-2019
      150870                       NaN                             NaN

      [150871 rows x 17 columns]>
```

- 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()



```
[41]: df.isnull()
```

[41]:

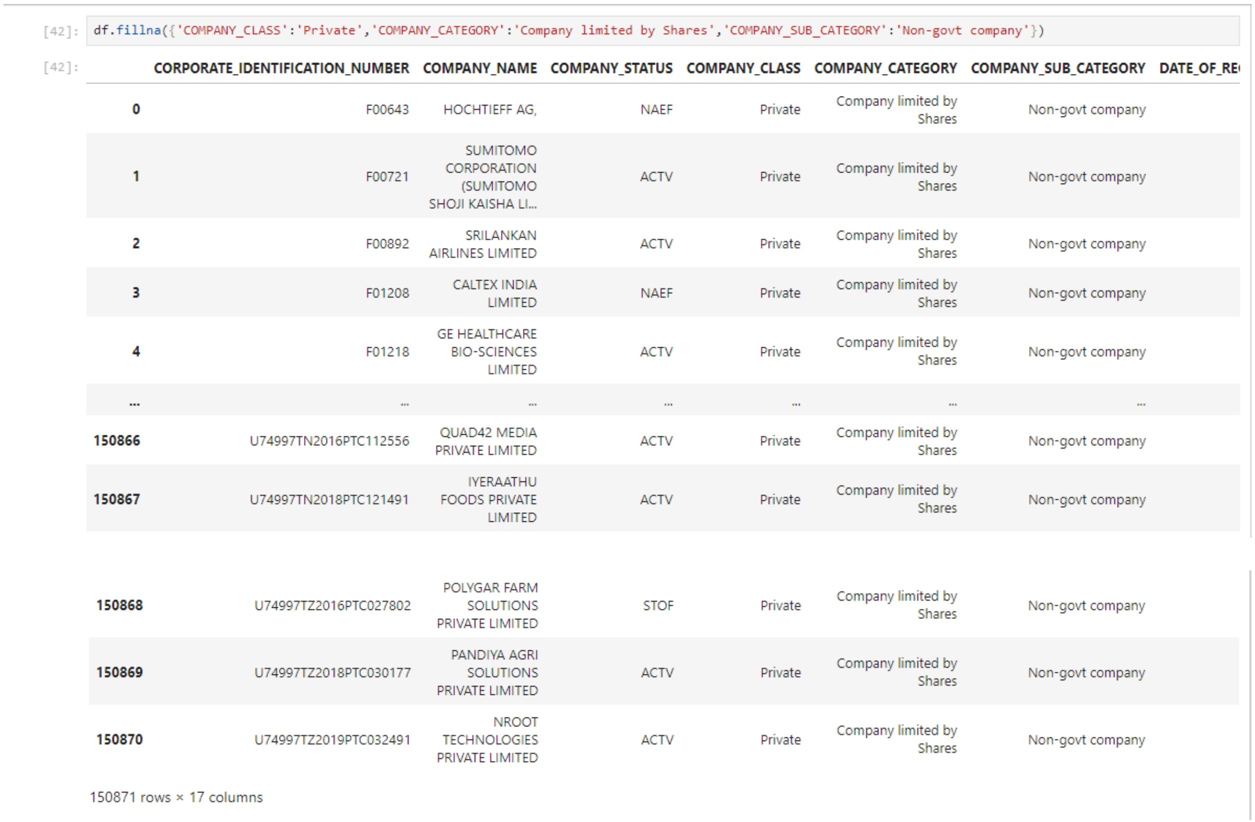
	CORPORATE_IDENTIFICATION_NUMBER	COMPANY_NAME	COMPANY_STATUS	COMPANY_CLASS	COMPANY_CATEGORY	COMPANY_SUB_CATEGORY	DATE_OF_RE
0	False	False	False	True	True	True	
1	False	False	False	True	True	True	
2	False	False	False	True	True	True	
3	False	False	False	True	True	True	
4	False	False	False	True	True	True	
...	...	...	...	...	...	...	
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'})



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

[42]:

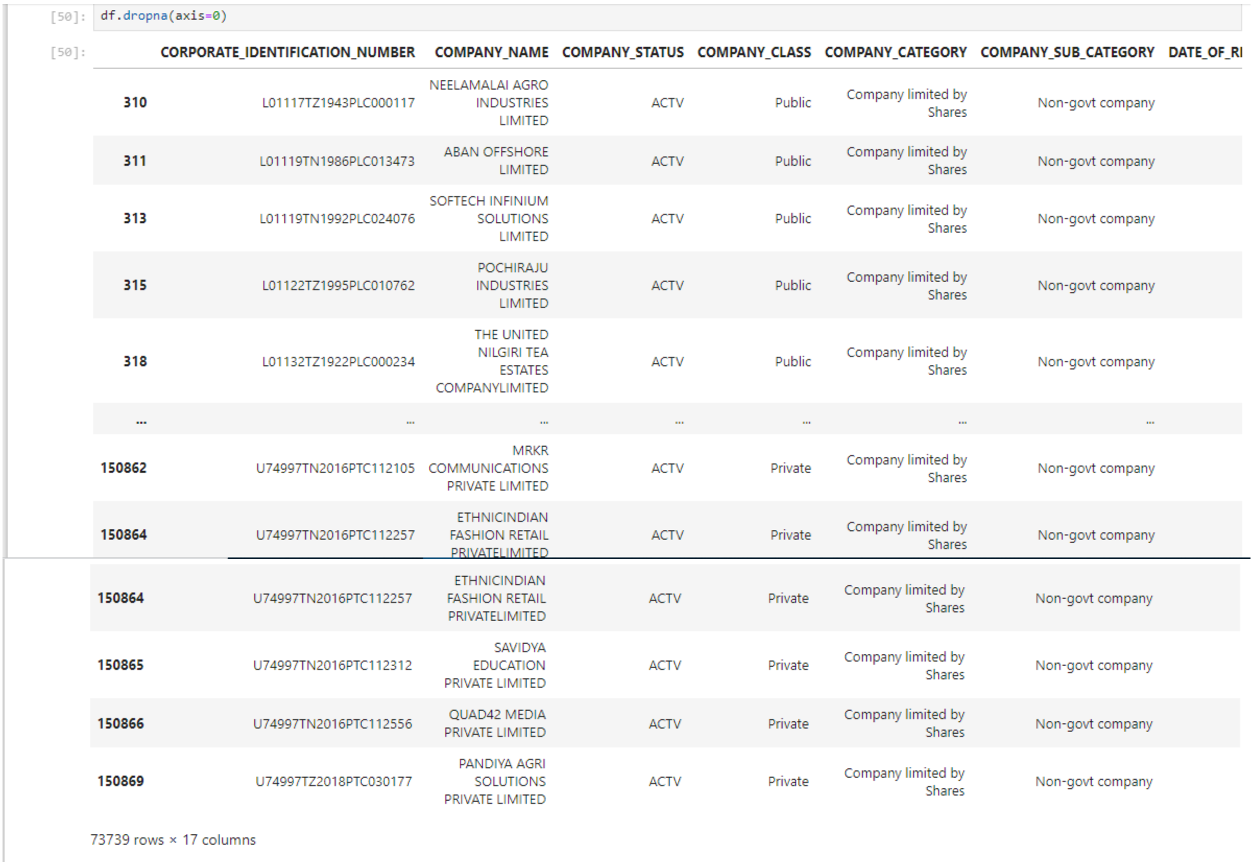
	CORPORATE_IDENTIFICATION_NUMBER	COMPANY_NAME	COMPANY_STATUS	COMPANY_CLASS	COMPANY_CATEGORY	COMPANY_SUB_CATEGORY	DATE_OF_RE
0	F00643	HOCHTIEFF AG,	NAEF	Private	Company limited by Shares	Non-govt company	
1	F00721	SUMITOMO CORPORATION (SUMITOMO SHOJI KAISHA LI...	ACTV	Private	Company limited by Shares	Non-govt company	
2	F00892	SRILANKAN AIRLINES LIMITED	ACTV	Private	Company limited by Shares	Non-govt company	
3	F01208	CALTEX INDIA LIMITED	NAEF	Private	Company limited by Shares	Non-govt company	
4	F01218	GE HEALTHCARE BIO-SCIENCES LIMITED	ACTV	Private	Company limited by Shares	Non-govt company	
...	...	...	...	...	...	...	
150866	U74997TN2016PTC112556	QUAD42 MEDIA PRIVATE LIMITED	ACTV	Private	Company limited by Shares	Non-govt company	
150867	U74997TN2018PTC121491	IYERAATHU FOODS PRIVATE LIMITED	ACTV	Private	Company limited by Shares	Non-govt company	
150868	U74997TZ2016PTC027802	POLYGAR FARM SOLUTIONS PRIVATE LIMITED	STOF	Private	Company limited by Shares	Non-govt company	
150869	U74997TZ2018PTC030177	PANDIYA AGRI SOLUTIONS PRIVATE LIMITED	ACTV	Private	Company limited by Shares	Non-govt company	
150870	U74997TZ2019PTC032491	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)



```
[50]: df.dropna(axis=0)
```

[50]:

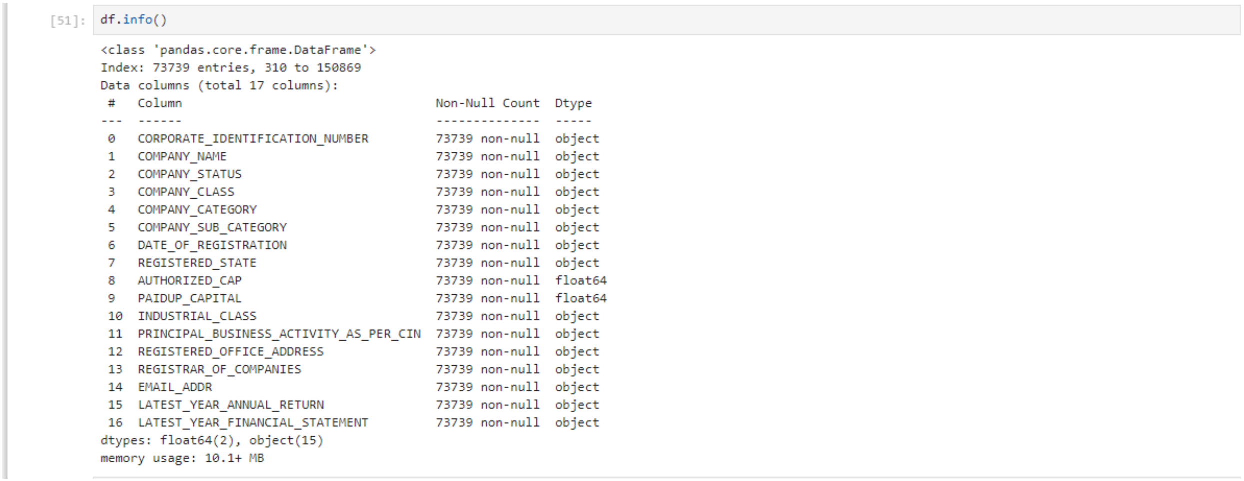
	CORPORATE_IDENTIFICATION_NUMBER	COMPANY_NAME	COMPANY_STATUS	COMPANY_CLASS	COMPANY_CATEGORY	COMPANY_SUB_CATEGORY	DATE_OF_RI
310	L01117TZ1943PLC000117	NEELAMALAI AGRO INDUSTRIES LIMITED	ACTV	Public	Company limited by Shares	Non-govt company	
311	L01119TN1986PLC013473	ABAN OFFSHORE LIMITED	ACTV	Public	Company limited by Shares	Non-govt company	
313	L01119TN1992PLC024076	SOFTECH INFINIUM SOLUTIONS LIMITED	ACTV	Public	Company limited by Shares	Non-govt company	
315	L01122TZ1995PLC010762	POCHIRAJU INDUSTRIES LIMITED	ACTV	Public	Company limited by Shares	Non-govt company	
318	L01132TZ1922PLC000234	THE UNITED NILGIRI TEA ESTATES COMPANYLIMITED	ACTV	Public	Company limited by Shares	Non-govt company	
...	...	...	...	...	...	...	
150862	U74997TN2016PTC112105	MRKR COMMUNICATIONS PRIVATE LIMITED	ACTV	Private	Company limited by Shares	Non-govt company	
150864	U74997TN2016PTC112257	ETHNICINDIAN FASHION RETAIL PRIVATELIMITED	ACTV	Private	Company limited by Shares	Non-govt company	
150864	U74997TN2016PTC112257	ETHNICINDIAN FASHION RETAIL PRIVATELIMITED	ACTV	Private	Company limited by Shares	Non-govt company	
150865	U74997TN2016PTC112312	SAVIDYA EDUCATION PRIVATE LIMITED	ACTV	Private	Company limited by Shares	Non-govt company	
150866	U74997TN2016PTC112556	QUAD42 MEDIA PRIVATE LIMITED	ACTV	Private	Company limited by Shares	Non-govt company	
150869	U74997TZ2018PTC030177	PANDIYA AGRI SOLUTIONS PRIVATE LIMITED	ACTV	Private	Company limited by Shares	Non-govt company	

73739 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()

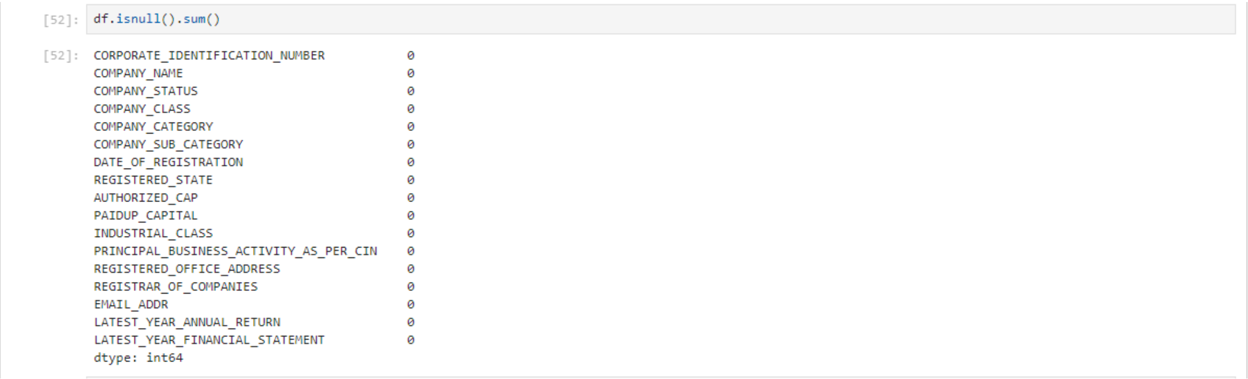


```
[51]: df.info()

<class 'pandas.core.frame.DataFrame'>
Index: 73739 entries, 310 to 150869
Data columns (total 17 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   CORPORATE_IDENTIFICATION_NUMBER         73739 non-null  object
 1   COMPANY_NAME                            73739 non-null  object
 2   COMPANY_STATUS                          73739 non-null  object
 3   COMPANY_CLASS                           73739 non-null  object
 4   COMPANY_CATEGORY                        73739 non-null  object
 5   COMPANY_SUB_CATEGORY                    73739 non-null  object
 6   DATE_OF_REGISTRATION                    73739 non-null  object
 7   REGISTERED_STATE                        73739 non-null  object
 8   AUTHORIZED_CAP                          73739 non-null  float64
 9   PAIDUP_CAPITAL                          73739 non-null  float64
 10  INDUSTRIAL_CLASS                        73739 non-null  object
 11  PRINCIPAL_BUSINESS_ACTIVITY_AS_PER_CIN  73739 non-null  object
 12  REGISTERED_OFFICE_ADDRESS               73739 non-null  object
 13  REGISTRAR_OF_COMPANIES                  73739 non-null  object
 14  EMAIL_ADDR                              73739 non-null  object
 15  LATEST_YEAR_ANNUAL_RETURN               73739 non-null  object
 16  LATEST_YEAR_FINANCIAL_STATEMENT         73739 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()



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

[52]: CORPORATE_IDENTIFICATION_NUMBER          0
      COMPANY_NAME                             0
      COMPANY_STATUS                           0
      COMPANY_CLASS                            0
      COMPANY_CATEGORY                         0
      COMPANY_SUB_CATEGORY                     0
      DATE_OF_REGISTRATION                     0
      REGISTERED_STATE                         0
      AUTHORIZED_CAP                           0
      PAIDUP_CAPITAL                           0
      INDUSTRIAL_CLASS                         0
      PRINCIPAL_BUSINESS_ACTIVITY_AS_PER_CIN   0
      REGISTERED_OFFICE_ADDRESS                0
      REGISTRAR_OF_COMPANIES                   0
      EMAIL_ADDR                               0
      LATEST_YEAR_ANNUAL_RETURN                0
      LATEST_YEAR_FINANCIAL_STATEMENT          0
      dtype: int64
```

- 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.