

## **PHASE – 2**

### **Data Preprocessing**

#### **AI Driven Exploration:**

<b>Name</b>	<b>Banu.J</b>
<b>Date</b>	<b>09/10/2023</b>
<b>Team ID</b>	<b>Proj-2121-Team(4)</b>
<b>Project Name</b>	<b>AI Driven Exploration</b>

## **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()`

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

	CORPORATE_IDENTIFICATION_NUMBER	COMPANY_NAME	COMPANY_STATUS	COMPANY_CLASS	COMPANY_CATEGORY	COMPANY_SUB_CATEGORY	DATE_OF_REGISTERATION
0	F00643	HOCHTIEF AG	NATP	nan	nan	nan	01-0
1	F08721	ELI LILLY CORPORATION ELI LILLY SHOEI KASHI K.K.	ACTV	nan	nan	nan	
2	F08852	SKYLINE AIRLINES LIMITED	ACTV	nan	nan	nan	01-0
3	F01028	CATERPILLAR LIMITED	NATP	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()`

```
[10]: df.isnull().sum()
[10]: CORPORATE_IDENTIFICATION_NUMBER      0
      COMPANY_NAME                          0
      COMPANY_STATUS                        0
      COMPANY_CLASS                         234
      COMPANY_CATEGORY                      234
      COMPANY_SUB_CATEGORY                 234
      DATE_OF_REGISTRATION                 39
      REGISTERED_STATE                     0
      AUTHORIZED_CAP                       0
      PAIDUP_CAPITAL                       0
      INDUSTRIAL_CLASS                     310
      PRINCIPAL_BUSINESS_ACTIVITY_AS_FOR_CS 0
      REGISTERED_OFFICE_ADDRESS            99
      REGISTERED_OFF_COMPLIANCE           174
      EMAIL_ADDRESS                       38129
      LATEST_YEAR_ANNUAL_RETURN            75009
      LATEST_YEAR_FINANCIAL_STATEMENT      75781
      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.

## df.info()

07. info	
0	cloud method [name] info of COMPANY_NOTIFICATION_NUMBER %
1	000000
2	000000
3	000000
4	000000
5	000000
6	000000
7	000000
8	000000
9	000000
10	000000
11	000000
12	000000
13	000000
14	000000
15	000000
16	000000
17	000000
18	000000
19	000000
20	000000
21	000000
22	000000
23	000000
24	000000
25	000000
26	000000
27	000000
28	000000
29	000000
30	000000
31	000000
32	000000
33	000000
34	000000
35	000000
36	000000
37	000000
38	000000
39	000000
40	000000
41	000000
42	000000
43	000000
44	000000
45	000000
46	000000
47	000000
48	000000
49	000000
50	000000
51	000000
52	000000
53	000000
54	000000
55	000000
56	000000
57	000000
58	000000
59	000000
60	000000
61	000000
62	000000
63	000000
64	000000
65	000000
66	000000
67	000000
68	000000
69	000000
70	000000
71	000000
72	000000
73	000000
74	000000
75	000000
76	000000
77	000000
78	000000
79	000000
80	000000
81	000000
82	000000
83	000000
84	000000
85	000000
86	000000
87	000000
88	000000
89	000000
90	000000
91	000000
92	000000
93	000000
94	000000
95	000000
96	000000
97	000000
98	000000
99	000000
100	000000
101	000000
102	000000
103	000000
104	000000
105	000000
106	000000
107	000000
108	000000
109	000000
110	000000
111	000000
112	000000
113	000000
114	000000
115	000000
116	000000
117	000000
118	000000
119	000000
120	000000
121	000000
122	000000
123	000000
124	000000
125	000000
126	000000
127	000000
128	000000
129	000000
130	000000
131	000000
132	000000
133	000000
134	000000
135	000000
136	000000
137	000000
138	000000
139	000000
140	000000
141	000000
142	000000
143	000000
144	000000
145	000000
146	000000
147	000000
148	000000</

- 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()`

```
[101] df.isnull()
```

```
[102]
```

	CORPORATE_IDENTIFICATION_NUMBER	COMPANY_NAME	COMPANY_STATUS	COMPANY_CLASS	COMPANY_CATEGORY	COMPANY_SUB_CATEGORY	DATE_OF_REGISTRATION
0	False	False	False	True	True	True	True
1	False	False	False	True	True	True	True
2	False	False	False	True	True	True	True
3	False	False	False	True	True	True	True
4	False	False	False	True	True	True	True
...	...	...	...	...	...	...	...
150666	False	False	False	False	False	False	False
150667	False	False	False	False	False	False	False
150668	False	False	False	False	False	False	False
150669	False	False	False	False	False	False	False
150670	False	False	False	False	False	False	False

150671 rows x 8 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'})`

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

```
[74]:
```

	CORPORATE_IDENTIFICATION_NUMBER	COMPANY_NAME	COMPANY_STATUS	COMPANY_CLASS	COMPANY_CATEGORY	COMPANY_SUB_CATEGORY	DATE_OF_REGISTRATION
0	R00643	HOOCHTOPP AG	NAEP	Private	Company limited by Shares	Non-govt company	2015-01-01
1	R00721	SUNTORNO CORPORATION (SUNTORNO SINGAPORE PTE LTD)	ACTV	Private	Company limited by Shares	Non-govt company	2015-01-01
2	R00862	SILANGAIR AIRLINES LIMITED	ACTV	Private	Company limited by Shares	Non-govt company	2015-01-01
3	R01208	CALFEI ROKA LIMITED	NAEP	Private	Company limited by Shares	Non-govt company	2015-01-01
4	R01218	GE HEALTHCARE BIO-SCIENCES LIMITED	ACTV	Private	Company limited by Shares	Non-govt company	2015-01-01
...	...	...	...	...	...	...	...
150866	U148677N2016PTC12355	QUAD4 MEDIA PRIVATE LIMITED	ACTV	Private	Company limited by Shares	Non-govt company	2016-01-01
150867	U148677N2016PTC123481	VIRMAATHU FOODS PRIVATE LIMITED	ACTV	Private	Company limited by Shares	Non-govt company	2016-01-01
...	...	...	...	...	...	...	...
150868	U148677Z2016PTC037802	POUSAA FARM SOLUTIONS PRIVATE LIMITED	STOP	Private	Company limited by Shares	Non-govt company	2016-01-01
150869	U148677Z2016PTC030177	PAUCHA AGRI SOLUTIONS PRIVATE LIMITED	ACTV	Private	Company limited by Shares	Non-govt company	2016-01-01
...	...	...	...	...	...	...	...
150870	U148677Z2016PTC033481	WROOF TECHNOLOGIES PRIVATE LIMITED	ACTV	Private	Company limited by Shares	Non-govt company	2016-01-01

150871 rows x 8 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)`

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

	CORPORATE_IDENTIFICATION_NUMBER	COMPANY_NAME	COMPANY_STATUS	COMPANY_CLASS	COMPANY_CATEGORY	COMPANY_SUB_CATEGORY	DATE_OF_R
310	UB1117023443PJC00117	NEELAMALAI AGRO INDUSTRIES LIMITED	ACTV	Public	Company limited by Shares	Non-govt company	
311	UB11181N1966PJC013473	ABAN OFFSHORE LIMITED	ACTV	Public	Company limited by Shares	Non-govt company	
313	UB11197H2950PJC024078	SOFTECH INFRASTRUCTURE SOLUTIONS LIMITED	ACTV	Public	Company limited by Shares	Non-govt company	
315	UB1122723469PJC010763	ROCKRAJ INDUSTRIES LIMITED	ACTV	Public	Company limited by Shares	Non-govt company	
316	UB1123727820PJC000234	THE UNITED NUGRI TEA ESTATES COMPANY LIMITED	ACTV	Public	Company limited by Shares	Non-govt company	
---	---	---	---	---	---	---	---
150862	U14977N2018PJC112108	MIOR COMMUNICATIONS PRIVATE LIMITED	ACTV	Private	Company limited by Shares	Non-govt company	
150864	U14977N2018PJC112257	ETHNICROWAN FASHION RETAIL PRIVATE LIMITED	ACTV	Private	Company limited by Shares	Non-govt company	
150864	U14977N2018PJC112257	ETHNICROWAN FASHION RETAIL PRIVATE LIMITED	ACTV	Private	Company limited by Shares	Non-govt company	
150865	U14977N2018PJC112312	SARVHA EDUCATION PRIVATE LIMITED	ACTV	Private	Company limited by Shares	Non-govt company	
150866	U14977N2018PJC112356	CLUSTAL MEDIA PRIVATE LIMITED	ACTV	Private	Company limited by Shares	Non-govt company	
150869	U1497722018PJC001127	INDIANA AGRI SOLUTIONS PRIVATE LIMITED	ACTV	Private	Company limited by Shares	Non-govt company	

15739 rows x 8 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'>
Int64: 73730 entries, 0 to 258805
Data columns (total 17 columns):
 #   Column                                  Non-Null Count  Dtype
---  -
 0   CORPORATE_IDENTIFICATION_NUMBER        73730 non-null  object
 1   COMPANY_NAME                           73730 non-null  object
 2   COMPANY_STATUS                         73730 non-null  object
 3   COMPANY_CLASS                          73730 non-null  object
 4   COMPANY_CATEGORY                      73730 non-null  object
 5   COMPANY_SUB_CATEGORY                  73730 non-null  object
 6   DATE_OF_REGISTRATION                  73730 non-null  object
 7   REGISTERED_STATE                     73730 non-null  object
 8   AUTHORIZED_CAP                        73730 non-null  float64
 9   RAISED_CAPITAL                       73730 non-null  float64
10  INDUSTRIAL_CLASS                      73730 non-null  object
11  PRINCIPAL_BUSINESS_ACTIVITY_AS_PER_CIN 73730 non-null  object
12  REGISTERED_OFFICE_ADDRESS             73730 non-null  object
13  REGISTERED_OFF_COMPANIES              73730 non-null  object
14  EMAIL_ADDRESS                         73730 non-null  object
15  LATEST_YEAR_ANNUAL_RETURN              73730 non-null  object
16  LATEST_YEAR_FINANCIAL_STATEMENT        73730 non-null  object
dtypes: float64(2), object(15)
memory usage: 18.1+ 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()`

```
[110]: df.isnull().sum()
[111]: CORPORATE_IDENTIFICATION_NUMBER      0
CORPORATE_NAME                             0
CORPORATE_STATUS                           0
CORPORATE_CLASS                            0
CORPORATE_CATEGORY                         0
CORPORATE_SUB_CATEGORY                    0
SITE_OF_REGISTRATION                      0
REGISTERED_STATE                          0
AUTHORIZED_CAP                            0
PAIDUP_CAPITAL                            0
INDUSTRIAL_CLASS                          0
PRINCIPAL_BUSINESS_ACTIVITY_AS_PER_CDB   0
REGISTERED_OFFICE_ADDRESS                 0
REGISTRAR_OF_COMPANIES                    0
DINCL_ADDR                                0
LATEST_YEAR_ANNUAL_RETURN                 0
LATEST_YEAR_FINANCIAL_STATEMENT           0
#Type: 0m54
```

- 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.