AI_Phase3

Serial Number	Team Member Name	Registration Number
1	K.Navinraj	310821104064
2	M.P.Praveen Raja	310821104070
3	Sachin A	310821104081
4	K.P.Tharun	310821104100

Objective:

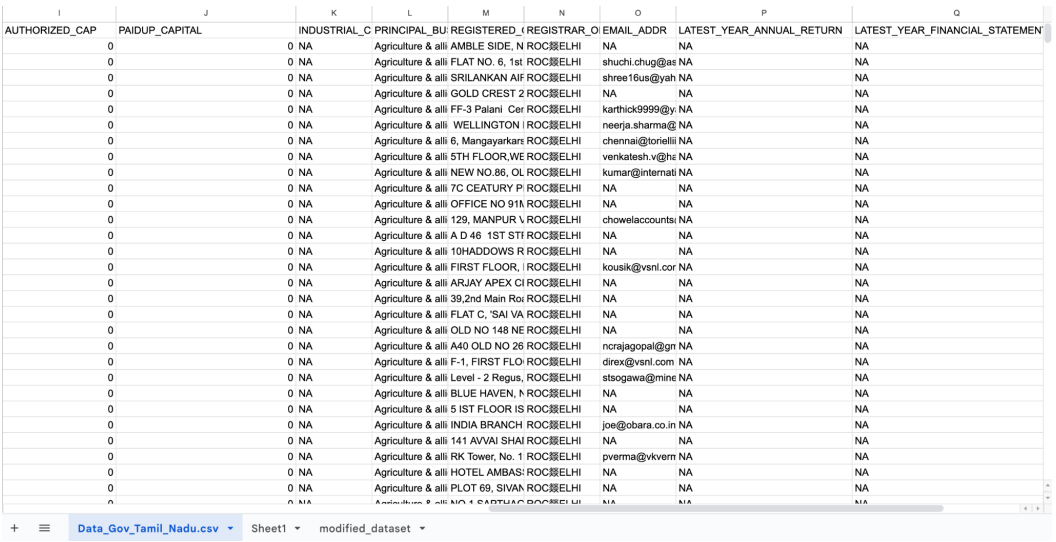
The technology stack described in the plan encompasses a range of tools and frameworks that are commonly used in data integration, preprocessing, modelling, real-time data processing, visualisation, deployment, monitoring, documentation, training, compliance, evaluation, and scaling. Let's elaborate on some of the key technologies and their purposes:

1.Dataset:

DatasetLink: [Company Master Data of Tamil Nadu upto 28th February 2019 | Open Government Data (OGD) Platform India]

	A	B	C	D	E	F	G	
1	CORPORATE_IDENTIFICATION	COMPANY_NAME	COMPANY_STATUS	COMPANY_CLASS	COMPANY_CATEGORY	COMPANY_SUB_CATEGO	DATE_OF_REG	RE
2	F00643	HOCHTIEFF AG,	NAEF	NA	NA	NA	01/12/1961	Ta
3	F00721	SUMITOMO CORPORATION (SUMITOMO SHOJI KAISHA LIMITED)	ACTV	NA	NA	NA	NA	Ta
4	F00892	SRILANKAN AIRLINES LIMITED	ACTV	NA	NA	NA	01/03/1982	Ta
5	F01208	CALTEX INDIA LIMITED	NAEF	NA	NA	NA	NA	Ta
6	F01218	GE HEALTHCARE BIO-SCIENCES LIMITED	ACTV	NA	NA	NA	NA	Ta
7	F01265	CAIRN ENERGY INDIA PTY. LIMITED	NAEF	NA	NA	NA	NA	Ta
8	F01269	TORIELLI S.R.L	ACTV	NA	NA	NA	05/09/1995	Ta
9	F01311	HARDY EXPLORATION & PRODUCTION (INDIA) INC..	ACTV	NA	NA	NA	NA	Ta
10	F01314	HOCHTIOF AKTIENGESELLSHARFF VORM GFBR HELFMANN	ACTV	NA	NA	NA	11/04/1996	Ta
11	F01412	EPSON SINGAPORE PVT LTD	ACTV	NA	NA	NA	25/04/1997	Ta
12	F01426	CARGOLUX AIRLINES INTERNATIONAL S A	ACTV	NA	NA	NA	11/06/1997	Ta
13	F01468	CHO HEUNG ELECTRIC INDUSTRIAL COMPANY LIMITED	NAEF	NA	NA	NA	NA	Ta
14	F01543	NYCOMED ASIA PACIFIC PTE LIMITED	ACTV	NA	NA	NA	27/10/1998	Ta
15	F01544	CHERRINGTON ASIA LTD	ACTV	NA	NA	NA	01/05/2000	Ta
16	F01563	SHIMADZU ASIA PACIFIC PTE LIMITED	NAEF	NA	NA	NA	NA	Ta
17	F01565	CORK INTERNATIONAL PTY LIMITED	ACTV	NA	NA	NA	NA	Ta
18	F01566	ERBIS ENGG COMPANY LIMITED	ACTV	NA	NA	NA	NA	Ta
19	F01589	RALF SCHNEIDER HOLDING GMBH	NAEF	NA	NA	NA	NA	Ta
20	F01593	MITRAJAYA TRADING PRIVATE LIMITED	ACTV	NA	NA	NA	NA	Ta
21	F01618	HEAT AND CONTROL PTY LIMITED	ACTV	NA	NA	NA	13/07/1999	Ta
22	F01628	DIREX SYSTEMS LIMITED	ACTV	NA	NA	NA	NA	Ta
23	F01641	NMB-MINEBEA THAI LIMITED	NAEF	NA	NA	NA	NA	Ta
24	F01643	ARROW INTERNATIONAL INC	ACTV	NA	NA	NA	02/11/1999	Ta
25	F01694	GAMBRO CHINA LTD	ACTV	NA	NA	NA	14/06/2000	Ta
26	F01703	OBARA CORPORATION	NAEF	NA	NA	NA	17/07/2000	Ta
27	F01752	CIPTA WAWASON MAJU ENGINEERING SDM BHD	ACTV	NA	NA	NA	24/01/2001	Ta
28	F01753	AUCHAN INTERNATIONAL S.A.	ACTV	NA	NA	NA	NA	Ta
29	F01767	TOSHIBA PLANT SYSTEMS AND SERVICES CORPORATION	NAEF	NA	NA	NA	08/03/2001	Ta
30	F01768	YAMAZEN CORPORATION	NAEF	NA	NA	NA	NA	Ta

Data_Gov_Tamil_Nadu.csv Sheet1 modified_dataset



AUTHORIZED_CAP	PAIDUP_CAPITAL	INDUSTRIAL_C	PRINCIPAL_BU	REGISTERED_	REGISTRAR_O	EMAIL_ADDR	LATEST_YEAR_ANNUAL_RETURN	LATEST_YEAR_FINANCIAL_STATEMEN
0	0	NA	Agriculture & alli	AMBLE SIDE, N	ROC簸ELHI	NA	NA	NA
0	0	NA	Agriculture & alli	FLAT NO. 6, 1st	ROC簸ELHI	shuchi.chug@as	NA	NA
0	0	NA	Agriculture & alli	SRILANKAN AIF	ROC簸ELHI	shree16us@yah	NA	NA
0	0	NA	Agriculture & alli	GOLD CREST 2	ROC簸ELHI	NA	NA	NA
0	0	NA	Agriculture & alli	FF-3 Palani Cer	ROC簸ELHI	karthick9999@y	NA	NA
0	0	NA	Agriculture & alli	WELLINGTON I	ROC簸ELHI	neerja.sharma@	NA	NA
0	0	NA	Agriculture & alli	6, Mangayarkars	ROC簸ELHI	chennai@toriellii	NA	NA
0	0	NA	Agriculture & alli	5TH FLOOR,WE	ROC簸ELHI	venkatesh.v@hs	NA	NA
0	0	NA	Agriculture & alli	NEW NO.86, OL	ROC簸ELHI	kumar@internati	NA	NA
0	0	NA	Agriculture & alli	7C CEATURY P	ROC簸ELHI	NA	NA	NA
0	0	NA	Agriculture & alli	OFFICE NO 91A	ROC簸ELHI	NA	NA	NA
0	0	NA	Agriculture & alli	129, MANPUR \	ROC簸ELHI	chowelaccountsi	NA	NA
0	0	NA	Agriculture & alli	A D 46 1ST STI	ROC簸ELHI	NA	NA	NA
0	0	NA	Agriculture & alli	10HADDOWS R	ROC簸ELHI	NA	NA	NA
0	0	NA	Agriculture & alli	FIRST FLOOR,	ROC簸ELHI	kousik@vsnl.cor	NA	NA
0	0	NA	Agriculture & alli	ARJAY APEX CI	ROC簸ELHI	NA	NA	NA
0	0	NA	Agriculture & alli	39,2nd Main Ro	ROC簸ELHI	NA	NA	NA
0	0	NA	Agriculture & alli	FLAT C, 'SAI VA	ROC簸ELHI	NA	NA	NA
0	0	NA	Agriculture & alli	OLD NO 148 NE	ROC簸ELHI	NA	NA	NA
0	0	NA	Agriculture & alli	A40 OLD NO 26	ROC簸ELHI	ncrajagopal@gn	NA	NA
0	0	NA	Agriculture & alli	F-1, FIRST FLO	ROC簸ELHI	direx@vsnl.com	NA	NA
0	0	NA	Agriculture & alli	Level - 2 Regus,	ROC簸ELHI	stsogawa@mine	NA	NA
0	0	NA	Agriculture & alli	BLUE HAVEN, N	ROC簸ELHI	NA	NA	NA
0	0	NA	Agriculture & alli	5 IST FLOOR IS	ROC簸ELHI	NA	NA	NA
0	0	NA	Agriculture & alli	INDIA BRANCH	ROC簸ELHI	joe@obara.co.in	NA	NA
0	0	NA	Agriculture & alli	141 AVVAI SHAI	ROC簸ELHI	NA	NA	NA
0	0	NA	Agriculture & alli	RK Tower, No. 1	ROC簸ELHI	pverma@vkvern	NA	NA
0	0	NA	Agriculture & alli	HOTEL AMBAS	ROC簸ELHI	NA	NA	NA
0	0	NA	Agriculture & alli	PLOT 69, SIVAN	ROC簸ELHI	NA	NA	NA

Data_Gov_Tamil_Nadu.csv | Sheet1 | modified_dataset

2. Required libraries

Importing required libraries

In [52]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

3. Loading the Dataset

We will load the dataset named “dialogs.txt” inside our python notebook using the pd.read_csv() method. To load the dataset for AI-driven exploration and prediction of company registration trends with the Registrar of Companies (RoC), you'll need to follow a series of steps. In this content guide, I'll outline the process for loading the dataset and preparing it for analysis using artificial intelligence (AI) techniques.

Loading the dataset.

In [65]:
```
#Reading the dataset from CSV format.
df = pd.read_csv('/Users/sachinanandharaj/Downloads/Data_Gov_Tamil_Nadu.csv', low_memory=False)
```

In [66]:
```
#Printing the given Dataset
df
```

Out[66]:

	CORPORATE_IDENTIFICATION_NUMBER	COMPANY_NAME	COMPANY_STATUS	COMPANY_CLASS	COMPANY_CATEGORY	COMPANY_SUB_CATEGORY
0	F00643	HOCHTIEFF AG,	NAEF	NaN	NaN	NaN
1	F00721	SUMITOMO CORPORATION (SUMITOMO SHOJI KAISHA LI...	ACTV	NaN	NaN	NaN
2	F00892	SRILANKAN AIRLINES LIMITED	ACTV	NaN	NaN	NaN
3	F01208	CALTEX INDIA LIMITED	NAEF	NaN	NaN	NaN
4	F01218	GE HEALTHCARE BIO-SCIENCES LIMITED	ACTV	NaN	NaN	NaN
...	...	...	...	...	...	...
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

4. Pre-processing

Preprocessing is a crucial step in AI-driven exploration and prediction of company registration trends with the Registrar of Companies (RoC). Proper preprocessing ensures that your dataset is clean, well-structured, and suitable for machine learning algorithms.

a often contains noise, irrelevant information, and inconsistencies that can interfere with the accuracy of NLP models.

Preprocessing

```
In [60]: #Cleaning the dataset by removing the NA value rows from the dataset
         df = df.dropna()

In [61]: #Printing the dataset
         df
```

5. Data Type Correction

Out[61]:

	CORPORATE_IDENTIFICATION_NUMBER	COMPANY_NAME	COMPANY_STATUS	COMPANY_CLASS	COMPANY_CATEGORY	COMPANY_SUB_CATEGORY
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
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

		LIMITED			Shares	
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
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

Data Type Correction

```
In [57]: df['INDUSTRIAL_CLASS'].astype('int32')

Out[57]: 310        1117
         311        1119
         313        1119
         315        1122
         318        1132
                    ...
         150862    74997
         150864    74997
         150865    74997
         150866    74997
         150869    74997
         Name: INDUSTRIAL_CLASS, Length: 73739, dtype: int32

In [58]: df['LATEST_YEAR_ANNUAL_RETURN'] = pd.to_datetime(df['LATEST_YEAR_ANNUAL_RETURN'])
```

CIN	REGISTERED_OFFICE_ADDRESS	REGISTRAR_OF_COMPANIES	EMAIL_ADDR	LATEST_YEAR_ANNUAL_RETURN	LATEST_YEAR_FINANCIAL_STATEMENT
llied	KATARY ESTATEKATARY POSTCOONOOR	ROC澳OIMBATORE	secneelamalai@avtplantations.co.in	2019-03-31	2019-03-31
llied	'JANPRIYA CREST'96, PANTHEON ROAD,EGMORE	ROC澳HENNAI	secretarial@aban.com	2019-03-31	2019-03-31
llied	29, PRECISION PLAZA, NEW 397, ANNA SALAITEYNAM...	ROC澳HENNAI	complianceofficer@softechinfinium.com	2018-03-31	2018-03-31
llied	1/102 SATHYAMANGALAM VILLAGEHOSUR TALUK	ROC澳OIMBATORE	mmreddyandco@gmail.com	2019-03-31	2019-03-31
llied	3 SAVITHRI SHANMUGHAM ROADRACE COURSE	ROC澳OIMBATORE	headoffice@chamrajtea.com	2019-03-31	2019-03-31

6. Exploratory Data Analysis

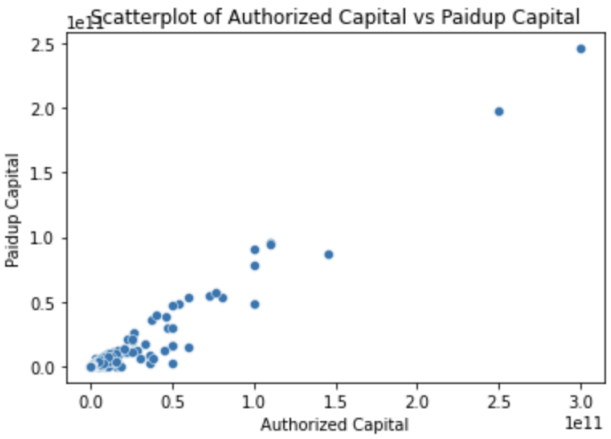
Exploratory data analysis (EDA) is used to analyses and investigate data sets and summarise their main characteristics, often employing data visualisation methods

AI-driven exploration and prediction of company registration trends with the Registrar of Companies (RoC) involves a series of steps, including data collection, cleaning, transformation, loading, EDA, modelling, and deployment. Careful attention to data quality and preprocessing is essential for building effective AI models in this context.

Exploratory Data Analysis

In [40]:
```
sns.scatterplot(x='AUTHORIZED_CAP', y='PAIDUP_CAPITAL', data= ds)
plt.title('Scatterplot of Authorized Capital vs Paidup Capital')
plt.xlabel('Authorized Capital')
plt.ylabel('Paidup Capital')
```

Out[40]:
```
Text(0, 0.5, 'Paidup Capital')
```

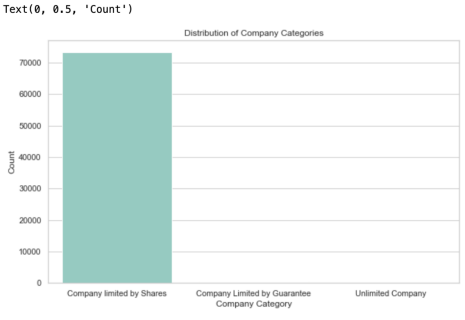


In [41]:
```
sns.set(style='whitegrid')  # Customize the plot style
plt.figure(figsize=(10, 6))  # Set the figure size

# Create the count plot for the COMPANY_CATEGORY column
sns.countplot(x='COMPANY_CATEGORY', data=ds, palette='Set3')  # Adjust the color palette if needed

# Customize the plot (optional)
plt.title('Distribution of Company Categories')
plt.xlabel('Company Category')
plt.ylabel('Count')
```

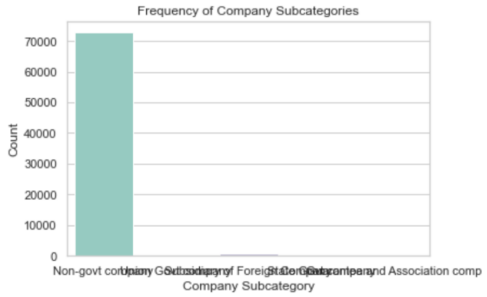
Out[41]:
```
Text(0, 0.5, 'Count')
```



In [43]:
```
### Create the count plot for the COMPANY_SUB_CATEGORY column
sns.countplot(x='COMPANY_SUB_CATEGORY', data=ds, palette='Set3')  # Adjust the color palette if needed

### Customize the plot
plt.title('Frequency of Company Subcategories')
plt.xlabel('Company Subcategory')
plt.ylabel('Count')
```

Out[43]: Text(0, 0.5, 'Count')

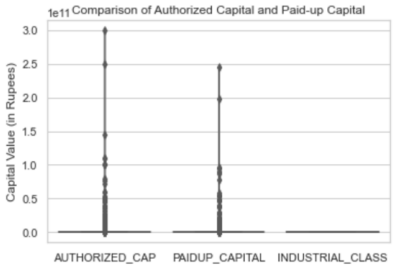


In [44]:
```
sns.boxplot(data=ds, palette='Set2')

# Add a violin plot on top of the box plot for better visualization
sns.violinplot(data=ds, palette='Set3', inner=None)

# Customize the plot (optional)
plt.title('Comparison of Authorized Capital and Paid-up Capital')
plt.ylabel('Capital Value (in Rupees)')
```

Out[44]: Text(0, 0.5, 'Capital Value (in Rupees)')



6. Extracting preprocessed csv file dataset

After performing data cleaning, the refined dataset can be saved and downloaded as a distinct CSV file.

Saving the Preprocessed dataset

In [16]:
```
df.to_csv('modified_dataset.csv', index=False)
```

In [18]:
```
ds = pd.read_csv('modified_dataset.csv')
```

7. Preprocessing Code

Importing required libraries

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

```
# Loading the dataset.
df = pd.read_csv('/Users/sachinanandharaj/Downloads/Data_Gov_Tamil_Nadu.csv',
low_memory=False)
print(df)
# Preprocessing - Removing NA values and changing the data type
df = df.dropna()
print(df)
# Data Type Correction
df['INDUSTRIAL_CLASS'].astype('int32')
df['LATEST_YEAR_ANNUAL_RETURN'] =
pd.to_datetime(df['LATEST_YEAR_ANNUAL_RETURN'])
df['LATEST_YEAR_FINANCIAL_STATEMENT'] =
pd.to_datetime(df['LATEST_YEAR_FINANCIAL_STATEMENT'])
df['INDUSTRIAL_CLASS'] = df['INDUSTRIAL_CLASS'].astype('int32')

# Exploratory Data Analysis
sns.scatterplot(x='AUTHORIZED_CAP', y='PAIDUP_CAPITAL', data= ds)
plt.title('Scatterplot of Authorized Capital vs Paidup Capital')
plt.xlabel('Authorized Capital')
plt.ylabel('Paidup Capital')
sns.set(style='whitegrid')  # Customize the plot style
plt.figure(figsize=(10, 6))  # Set the figure size

# Create the count plot for the COMPANY_CATEGORY column
sns.countplot(x='COMPANY_CATEGORY', data=ds, palette='Set3')  # Adjust the color
palette if needed

# Customise the plot
plt.title('Distribution of Company Categories')
plt.xlabel('Company Category')
plt.ylabel('Count')

# Create the count plot for the COMPANY_SUB_CATEGORY column
sns.countplot(x='COMPANY_SUB_CATEGORY', data=ds, palette='Set3')
plt.title('Frequency of Company Subcategories')
plt.xlabel('Company Subcategory')
plt.ylabel('Count')
sns.boxplot(data=ds, palette='Set2')

# Add a violin plot on top of the box plot for better visualization
sns.violinplot(data=ds, palette='Set3', inner=None)
plt.title('Comparison of Authorized Capital and Paid-up Capital')
plt.ylabel('Capital Value (in Rupees)')

# Saving the Preprocessed dataset
df.to_csv('modified_dataset.csv', index=False)
ds = pd.read_csv('modified_dataset.csv')
```