AI - DRIVEN EXPLORATION AND PREDICTION OF COMPANY REGISTRATION TRENDS WITH REGISTRAR OF COMPANIES(ROC)

PHASE 5 SUBMISSION DOCUMENT

510521205033: Renuka H

Project Title:ROC Company Analysis

Phase 5:Project Documentation & Submission

Topic:In this section we will document the complete project and prepare it for submission.

ROC Company Analysis

INTRODUCTION:

Registrar of Company (ROC) Company Analysis is a detailed and in-depth examination of a specific business entity's operations, financial performance, regulatory compliance, and legal standing in accordance with the guidelines and regulations stipulated by the Registrar of Companies in a given jurisdiction. The ROC, functioning as a governmental regulatory authority, holds a central role in overseeing, monitoring, and administering corporate entities, ensuring adherence to statutory provisions, corporate governance standards, and financial transparency.

This comprehensive analysis is of paramount importance to a wide range of stakeholders, including investors, shareholders, regulators, and financial analysts, as it offers a holistic understanding of a company's adherence to legal obligations, its financial health, and its position within the regulatory framework. ROC Company Analysis is a meticulous investigation that encompasses various dimensions of the company's operations and governance.

Key components integral to this ROC Company Analysis encompass:

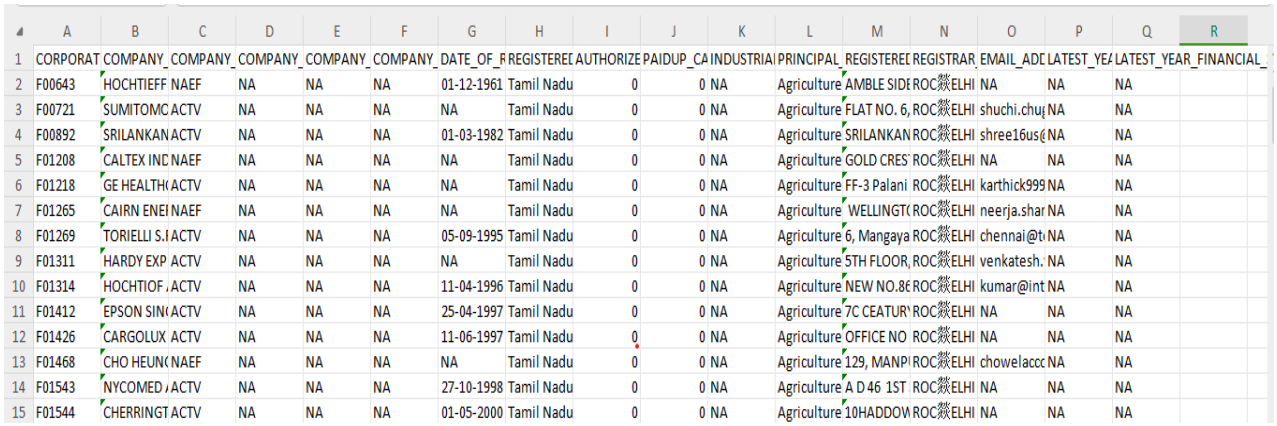
- Registrar of Companies (ROC) Overview: A comprehensive explanation of the Registrar of Companies and its function in regulating and supervising the activities of companies within a specified jurisdiction.
- Company Profile: A detailed introduction to the company under examination, featuring information such as its name, industry sector, geographical location, key business activities, and a brief history.
- Legal Compliance and Regulatory Adherence: A rigorous evaluation of the company's conformity with statutory requirements, including its registration status, filing of annual reports, compliance with tax regulations, and adherence to other pertinent legal and regulatory mandates.

- Financial Health and Performance: A meticulous assessment of the company's financial well-being, with a focus on its financial statements, profitability, solvency, liquidity, and overall financial sustainability.
- Corporate Governance Structure: An exploration of the company's governance framework, including the composition of its board of directors, executive leadership, corporate governance principles in place, and their practical implementation.
- Shareholder Landscape: An overview of the company's ownership structure, identification of major shareholders, institutional investors, public ownership, and any notable trends in shareholding.
- Operational Efficiency and Market Positioning: A critical examination of the company's operational efficiency, core business processes, competitive strengths, market positioning, and any notable advantages or challenges in its industry.
- Market Analysis and Competitive Landscape: An analysis of the company's position within its industry, market trends, competitive landscape, and key factors impacting its competitive standing.
- Legal and Regulatory Issues: Insights into any ongoing or past legal disputes, regulatory compliance concerns, or matters affecting the company's legal standing.
- Sustainability and Future Prospects: A discussion of the company's sustainability initiatives, commitment to ethical practices, and its potential for future growth and development.
- Investment Considerations and Strategic Implications: Guidance on how the findings of the ROC Company Analysis can influence investment decisions, risk assessment, and strategic planning for the company under scrutiny.

Throughout this analysis, practical examples and case studies may be provided to illustrate the real-world relevance of ROC Company Analysis in assessing a company's legal compliance, financial performance, and its overall position within the regulatory landscape. This comprehensive assessment serves as an indispensable resource for investors, regulators, and

corporate professionals, enabling them to make informed decisions about companies operating under the purview of the Registrar of Companies. Understanding a company's adherence to legal obligations and its regulatory compliance is foundational to ensuring transparency, accountability, and legal integrity in the corporate sphere.

GivenDataset:https://tn.data.gov.in/resource/company-master-data-tamil-nadu-upto-28th-february-2019



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	CORPORAT	COMPANY_	COMPANY_	COMPANY_	COMPANY_	COMPANY_	DATE_OF_R	REGISTERE	AUTHORIZE	PAIDUP_CA	INDUSTRIAI	PRINCIPAL_	REGISTERE	REGISTRAR	EMAIL_ADI	LATEST_YE	LATEST_YEAR_FINANCIAL_	
2	F00643	HOCHTIEFF	NAEF	NA	NA	NA	01-12-1961	Tamil Nadu	0	0	NA	Agriculture	AMBLE SIDE	ROC-ELHI	NA	NA	NA	
3	F00721	SUMITOMC	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture	FLAT NO. 6,	ROC-ELHI	shuchi.chu	NA	NA	
4	F00892	SRILANKAN	ACTV	NA	NA	NA	01-03-1982	Tamil Nadu	0	0	NA	Agriculture	SRILANKAN	ROC-ELHI	shree16us(	NA	NA	
5	F01208	CALTEX INC	NAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture	GOLD CRES	ROC-ELHI	NA	NA	NA	
6	F01218	GE HEALTH(	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture	FF-3 Palani	ROC-ELHI	karthick999	NA	NA	
7	F01265	CAIRN ENEI	NAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture	WELLINGT(	ROC-ELHI	neerja.shar	NA	NA	
8	F01269	TORIELLI S.I	ACTV	NA	NA	NA	05-09-1995	Tamil Nadu	0	0	NA	Agriculture	6, Mangaya	ROC-ELHI	chennai@t	NA	NA	
9	F01311	HARDY EXP	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture	5TH FLOOR,	ROC-ELHI	venkatesh.	NA	NA	
10	F01314	HOCHTIOF /	ACTV	NA	NA	NA	11-04-1996	Tamil Nadu	0	0	NA	Agriculture	NEW NO.8(	ROC-ELHI	kumar@int	NA	NA	
11	F01412	EPSON SIN(	ACTV	NA	NA	NA	25-04-1997	Tamil Nadu	0	0	NA	Agriculture	7C CEATUR	ROC-ELHI	NA	NA	NA	
12	F01426	CARGOLUX	ACTV	NA	NA	NA	11-06-1997	Tamil Nadu	0	0	NA	Agriculture	OFFICE NO	ROC-ELHI	NA	NA	NA	
13	F01468	CHO HEUN(	NAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture	129, MANP	ROC-ELHI	chowlaccc	NA	NA	
14	F01543	NYCOMED /	ACTV	NA	NA	NA	27-10-1998	Tamil Nadu	0	0	NA	Agriculture	A D 46 1ST	ROC-ELHI	NA	NA	NA	
15	F01544	CHERRINGT	ACTV	NA	NA	NA	01-05-2000	Tamil Nadu	0	0	NA	Agriculture	10HADDOV	ROC-ELHI	NA	NA	NA	

.

.

.

150860	U74997TN2	BRAINY BO	STOF	Private	Company li	Non-govt o	22-04-2015	Tamil Nadu	1500000	1000000	74997	Real estate	4/619, BAB	ROC-HEN	karthickalir	NA	NA	
150861	U74997TN2	AUROFLUX	ACTV	Private	Company li	Non-govt o	23-04-2015	Tamil Nadu	500000	500000	74997	Real estate	Plot No. 3,	ROC-HEN	mcmadhav	31-03-2019	31-03-2019	
150862	U74997TN2	CHENNAI R	ACTV	Private	Company li	Non-govt o	09-07-2016	Tamil Nadu	1000000	100000	74997	Real estate	NO. 189, 3R	ROC-HEN	shans@dci.	NA	NA	
150863	U74997TN2	FITFYT WEL	UPSO	Private	Company li	Non-govt o	19-07-2016	Tamil Nadu	100000	100000	74997	Real estate	NEW NO. 2(	ROC-HEN	caysilkarun	NA	NA	
150864	U74997TN2	MRKR COM	ACTV	Private	Company li	Non-govt o	22-08-2016	Tamil Nadu	100000	100000	74997	Real estate	11C,AASAR	ROC-HEN	RKRPRIYAN	31-03-2019	31-03-2019	
150865	U74997TN2	XYZT TECH	ACTV	Private	Company li	Non-govt o	30-08-2016	Tamil Nadu	1000000	1000000	74997	Real estate	Flat no. 120	ROC-HEN	deepak@b	NA	NA	
150866	U74997TN2	ETHNICIND	ACTV	Private	Company li	Non-govt o	30-08-2016	Tamil Nadu	1000000	1000000	74997	Real estate	VNR Milfor	ROC-HEN	ALEX.CHAN	31-03-2019	31-03-2019	
150867	U74997TN2	SAVIDYA EI	ACTV	Private	Company li	Non-govt o	01-09-2016	Tamil Nadu	1000000	100000	74997	Real estate	Mahalaksh	ROC-HEN	sridevis77(	31-03-2019	31-03-2019	
150868	U74997TN2	QUAD42 MI	ACTV	Private	Company li	Non-govt o	19-09-2016	Tamil Nadu	1000000	100000	74997	Real estate	Old No.37,	ROC-HEN	ezhil@qua(	31-03-2019	31-03-2019	
150869	U74997TN2	IYERAATHU	ACTV	Private	Company li	Non-govt o	16-03-2018	Tamil Nadu	100000	100000	74997	Real estate	G-3, G-Bloc	ROC-HEN	sneha.crea	NA	NA	
150870	U74997TZ2(	POLYGAR F,	STOF	Private	Company li	Non-govt o	20-07-2016	Tamil Nadu	100000	20000	74997	Real estate	31, LIC Colo	ROC-OIM	prashanthr	NA	NA	
150871	U74997TZ2(	PANDIYA A	ACTV	Private	Company li	Non-govt o	16-03-2018	Tamil Nadu	2500000	1500000	74997	Real estate	10/10 C3, V	ROC-OIM	sathishpan	31-03-2019	31-03-2019	
150872	U74997TZ2(	NROOT TEC	ACTV	Private	Company li	Non-govt o	25-07-2019	Tamil Nadu	1500000	1100000	74997	Real estate	139/1BPUD	ROC-OIM	nroottechn	NA	NA	
150873																		

150872 Rows x 17 Columns

Here's a listed of tools and software commonly used in the process:

Registrar of Company (ROC) Company Analysis requires a variety of tools and software to effectively gather, analyze, and present data related to a company's financial, legal, and operational aspects. Here are some of the commonly used tools and software for conducting a comprehensive ROC Company Analysis:

1.Financial Analysis and Modeling:

- Microsoft Excel: Excel is a versatile tool for financial modeling, data analysis, and creating financial statements. It's widely used for financial ratio analysis, trend analysis, and data organization.

2.Financial Data Sources:

- Bloomberg Terminal: Bloomberg provides real-time financial data, market news, and analytics, making it invaluable for financial research and analysis.
- Reuters Eikon: Similar to Bloomberg, Reuters Eikon offers financial data, news, and analytics for comprehensive financial research.

3.Regulatory Filings and Legal Research:

- EDGAR (Electronic Data Gathering, Analysis, and Retrieval): For companies registered with the U.S. Securities and Exchange Commission (SEC), EDGAR is the primary source for accessing and analyzing regulatory filings, including annual reports and financial statements.
- Company Registry Websites: Many countries have online company registries that offer information on a company's legal status, filings, and ownership details.

4.Corporate Governance and Board Information:

- BoardEx: BoardEx provides information on corporate boards, executive leadership, and governance structures of public companies.

5.Market and Industry Analysis:

- Statista: Statista offers industry reports, market data, and statistics, which are valuable for understanding a company's position within its industry.
- IBISWorld: A source for industry research reports and market data.
- MarketResearch.com: Provides a vast collection of market research reports.

6.Accounting Software:

- QuickBooks or Xero: These accounting software platforms can be used to assess a company's financial records and practices.

7.Data Visualization and Reporting Tools:

- Tableau or Power BI: These tools are helpful for creating interactive data visualizations, dashboards, and reports to present analysis findings effectively.

8.Statistical and Data Analysis Software:

- R or Python: These programming languages are commonly used for advanced statistical analysis and data manipulation, especially when dealing with large datasets.

9.Investment Analysis Software:

- Morningstar Direct: This platform is used by investment professionals for in-depth investment analysis, including equities and fixed-income instruments.
- FactSet: Provides financial and economic data and analytics for investment research.

10.Legal and Regulatory Compliance Tools:

- Compliance Management Software: These tools assist in assessing a company's compliance with legal obligations and regulatory requirements.

11.Screening and Due Diligence Software:

- Thomson Reuters World-Check: Used for screening entities against global sanctions lists, politically exposed persons (PEPs), and adverse media for due diligence and compliance purposes.

12.Communication and Project Management:

- Slack, Microsoft Teams, or similar platforms: These are essential for team communication, task management, and project coordination.

The specific tools and software used for ROC Company Analysis may vary depending on the scope of the analysis, the jurisdiction of the company, and the available resources. Analysts often employ a combination of these tools to ensure a thorough and accurate assessment of the company's financial, legal, and operational aspects.

1.PROBLEM STATEMENT AND DESIGN THINKING PROCESS:

Problem statement:

The Registrar of Companies (ROC) Company Analysis project seeks to address the challenge of effectively evaluating and comprehensively understanding the financial, regulatory, and operational aspects of a specific company under the purview of the ROC. The project recognizes the need for stakeholders, including investors, regulators, and financial analysts, to make informed decisions and assessments of a company's performance, regulatory compliance, and overall standing within the legal and financial landscape.

1.Lack of Comprehensive Information: There is a lack of a centralized, comprehensive source for accessing and analyzing all the relevant data needed for a thorough ROC Company Analysis. This includes financial statements, regulatory filings, legal compliance records, and operational data.

2.Data Accuracy and Reliability: The reliability and accuracy of data, especially when dealing with international or complex corporate structures, can be a significant challenge. Ensuring that the data used for analysis is up-to-date and trustworthy is crucial.

3.Complex Regulatory Frameworks: Companies operating in various jurisdictions need to comply with multiple regulatory frameworks. Understanding and assessing these diverse regulations can be complex and time-consuming.

4.Efficient Data Collection and Management: Efficiently gathering, organizing, and managing the vast amount of data required for ROC Company Analysis is often a logistical challenge.

5.Data Privacy and Security: Ensuring that sensitive corporate information and data are handled securely and in compliance with data protection regulations is a concern.

6.Interpreting Legal Compliance: Evaluating a company's adherence to legal and regulatory obligations can be intricate, as it requires a deep understanding of local and international legal frameworks.

7.Strategic Decision-Making: The project aims to enable stakeholders to make strategic decisions based on the analysis. Identifying key insights and translating them into actionable strategies is a critical challenge.

Design Thinking Process:

1. Empathize: Understand Stakeholder Needs

- Identify the stakeholders involved, such as investors, regulators, and financial analysts.
- Conduct interviews, surveys, and focus groups to understand their specific needs and pain points in ROC Company Analysis.
- Create personas to represent different stakeholder groups and their goals.

2. Define: Problem Statement and User Requirements

- Synthesize the information gathered and define a clear problem statement for the ROC Company Analysis project.
- Create a user journey map to visualize the steps stakeholders take in the analysis process.
- Develop a list of user requirements and project objectives based on the empathy stage findings.

3. Ideate: Generate Innovative Solutions

- Organize brainstorming sessions with a cross-functional team, including data analysts, legal experts, and software developers.
- Generate creative ideas for tools, software, and methodologies that can improve the ROC Company Analysis process.
- Encourage open, free-flowing idea generation without immediate judgment.

4. Prototype: Create Low-Fidelity Solutions

- Develop low-fidelity prototypes of potential tools or software solutions.
- These prototypes could be paper sketches, wireframes, or simplified software mockups.
- Keep the prototypes simple to focus on core functionalities and user experience.

5. Test: Gather Feedback and Iterate

- Engage with a select group of stakeholders to test the prototypes.
- Collect feedback on usability, functionality, and whether the proposed solutions address their needs.
- Iterate on the prototypes based on feedback, making necessary adjustments.

6. Develop: Build Functional Software Tools

- Based on the feedback and refined prototypes, move forward with the development of the selected tools or software solutions.
- Work closely with developers, data analysts, and domain experts to ensure the solutions meet the defined user requirements.
- Ensure that data security and privacy regulations are adhered to during development.

7. Implement: Pilot the Solutions

- Launch a pilot phase of the tools or software with a smaller group of users.
- Gather real-world feedback and performance data.
- Monitor for any technical issues, data accuracy, and user satisfaction.

8. Deliver: Full-Scale Implementation

- After successful testing and refinement in the pilot phase, roll out the tools or software for broader use.
- Provide training and support to users to ensure a smooth transition to the new solutions.

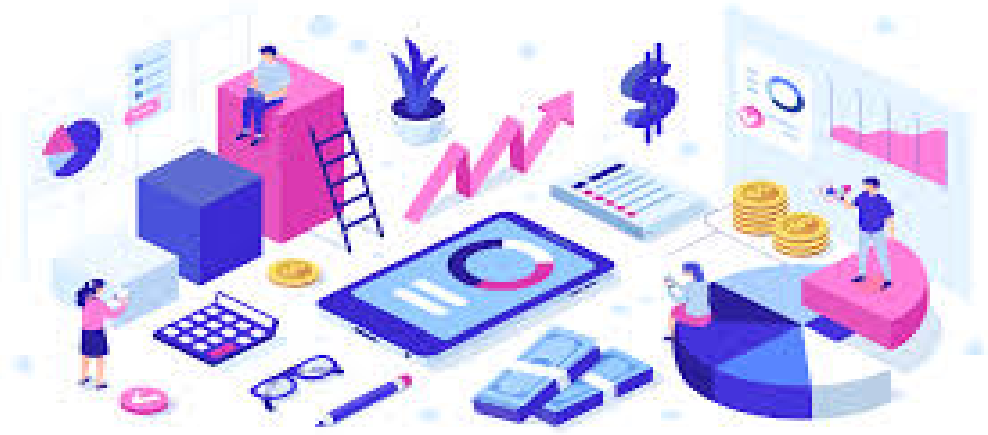
9. Measure: Evaluate Impact and Continuous Improvement

- Collect data on the impact of the new tools or software on the ROC Company Analysis process.
- Analyze key performance indicators, such as efficiency gains, data accuracy, and user satisfaction.

- Continuously gather feedback and make improvements based on user needs and changing regulatory requirements.

10. Scale and Sustain: Expand and Maintain the Solutions

- If the solutions prove successful, consider expanding their use to a broader audience.
- Develop a maintenance plan to ensure the ongoing functionality, security, and compliance of the tools and software.



2.DESIGN INTO INNOVATION:

Data source:

The data source for ROC (Registrar of Companies) company analysis typically includes datasets and information provided by the Registrar of Companies or government authorities responsible for corporate registrations.

GivenDataset:https://tn.data.gov.in/resource/company-master-data-tamil-nadu-upto-28th-february-2019

Data Collection:

- Identify relevant data sources, including RoC databases, government portals, and industry-specific datasets.
- Obtain necessary permissions and access rights to collect the data.
- Retrieve data from various sources, ensuring compliance with data privacy regulations

Data Preprocessing:

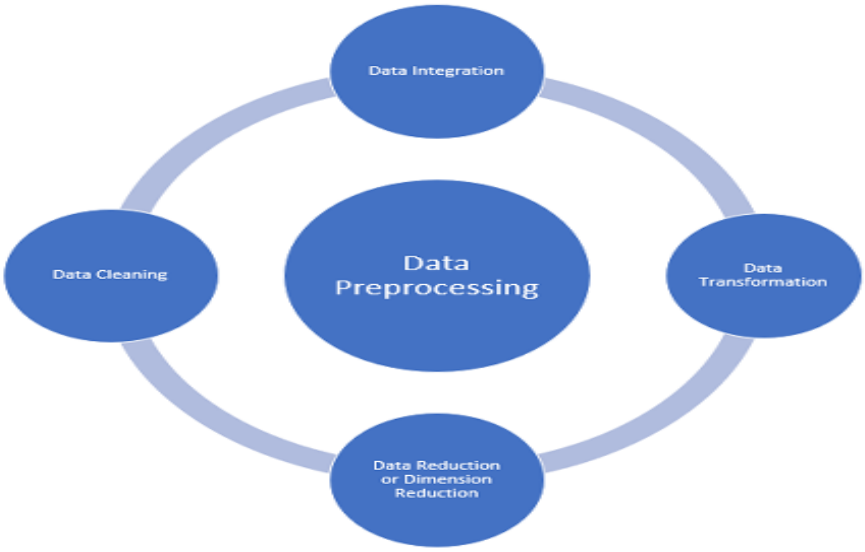
- Clean the data by handling missing values, outliers, and duplicates.
- Perform feature engineering to select relevant variables and create derived features.
- Normalize or scale numerical data and encode categorical variables.
- Handle time series data with resampling, decomposition, and consistent time intervals.
- Split the data into training, validation, and testing sets for modeling.
- Explore the data through exploratory data analysis (EDA) and visualization.
- Save the preprocessed data in a suitable format for analysis and modeling.

Data preprocessing Steps:

- Data Collection:
  - Gather the raw data from various sources, such as databases, files, APIs, or sensors.
- Data Cleaning:
  - Handle missing data: Impute missing values by filling them in with appropriate values (e.g., mean, median, or mode) or removing rows or columns with too many missing values.
  - Remove duplicates: Identify and remove duplicate entries in the dataset.

    - Outlier detection and treatment: Identify and handle outliers that may affect the quality of your analysis or model.
- Data Transformation:
    - Data encoding: Convert categorical data into numerical format using techniques like one-hot encoding or label encoding.
    - Feature scaling: Scale numerical features to have similar scales, often using methods like Min-Max scaling or standardization.
    - Feature engineering: Create new features based on domain knowledge or data analysis to enhance the model's performance.
    - Binning or discretization: Convert continuous numerical data into discrete bins.
    - Text preprocessing: Tokenize, remove stop words, and perform stemming or lemmatization for natural language processing tasks.
- Data Reduction:
    - Dimensionality reduction: Use techniques like Principal Component Analysis (PCA) or feature selection methods to reduce the number of features while preserving important information.
- Data Splitting:
    - Divide the dataset into training, validation, and test sets to assess model performance and prevent overfitting.
- Data Normalization:
    - Normalize data to ensure that all features have a similar scale, making it easier for some machine learning algorithms to converge.
- Handling Time Series Data:
    - Resampling: Adjust time intervals in time series data, such as aggregating or interpolating data points.
    - Lag features: Create lag features to capture temporal patterns.
- Handling Imbalanced Data (for classification problems):
    - Over-sampling: Increase the number of instances in the minority class.
    - Under-sampling: Decrease the number of instances in the majority class.

    - Synthetic data generation: Create synthetic examples for the minority class.
- Data Visualization:
    - Explore and visualize the data to gain insights and identify patterns or trends.
- Data Integration:
    - Merge or join multiple datasets if needed for your analysis or modeling.
- Data Quality Assurance:
    - Check for data consistency, accuracy, and validity.
    - Address any data quality issues that may arise during preprocessing.
- Documentation:
    - Document the preprocessing steps and any assumptions made to maintain transparency and facilitate reproducibility.



PYTHON PROGRAM:

#Import necessary libraries

```
import numpy as np

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt
```

#Loading dataset

In[1]:

```
dataset=pd.read_csv('Data_Gov_Tamil_Nadu.csv',encoding=("ISO-8859-1"),low_memory=False)
```

In[2]:

```
print(dataset.columns)
```

Out[ ]:

```
Index(['CORPORATE_IDENTIFICATION_NUMBER', 'COMPANY_NAME',
'COMPANY_STATUS',
    'COMPANY_CLASS', 'COMPANY_CATEGORY', 'COMPANY_SUB_CATEGORY',
    'DATE_OF_REGISTRATION', 'REGISTERED_STATE', 'AUTHORIZED_CAP',
```

'PAIDUP_CAPITAL', 'INDUSTRIAL_CLASS',

'PRINCIPAL_BUSINESS_ACTIVITY_AS_PER_CIN',
'REGISTERED_OFFICE_ADDRESS',

'REGISTRAR_OF_COMPANIES', 'EMAIL_ADDR',
'LATEST_YEAR_ANNUAL_RETURN',

'LATEST_YEAR_FINANCIAL_STATEMENT'],
dtype='object')

In[3]:

dataset.info()

Out[ ]:

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 150871 entries, 0 to 150870

Data columns (total 17 columns):

#	Column	Non-Null Count	Dtype
0	CORPORATE_IDENTIFICATION_NUMBER	150871 non-null	object
1	COMPANY_NAME	150871 non-null	object
2	COMPANY_STATUS	150871 non-null	object

```
3  COMPANY_CLASS                    150537 non-null  object
4  COMPANY_CATEGORY                 150537 non-null  object
5  COMPANY_SUB_CATEGORY             150537 non-null  object
6  DATE_OF_REGISTRATION             150832 non-null  object
7  REGISTERED_STATE                 150871 non-null  object
8  AUTHORIZED_CAP                   150871 non-null  float64
9  PAIDUP_CAPITAL                   150871 non-null  float64
10 INDUSTRIAL_CLASS                 150561 non-null  object
11 PRINCIPAL_BUSINESS_ACTIVITY_AS_PER_CIN 150871 non-null  object
12 REGISTERED_OFFICE_ADDRESS        150781 non-null  object
13 REGISTRAR_OF_COMPANIES           150697 non-null  object
14 EMAIL_ADDR                       112742 non-null  object
15 LATEST_YEAR_ANNUAL_RETURN        74982 non-null   object
16 LATEST_YEAR_FINANCIAL_STATEMENT  75089 non-null   object
dtypes: float64(2), object(15)
memory usage: 19.6+ MB
```

#Display the first few rows of the dataset to get an overview

In[5]:

```
print("Dataset Preview:")

print(dataset.head())
```

Out[ ]:

```
Dataset Preview:

 CORPORATE_IDENTIFICATION_NUMBER  \

0                  F00643

1                  F00721

2                  F00892

3                  F01208

4                  F01218



                          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

COMPANY_CLASS COMPANY_CATEGORY COMPANY_SUB_CATEGORY DATE_OF_REGISTRATION \

0 NaN NaN NaN 01-12-1961

1 NaN NaN NaN NaN

2 NaN NaN NaN 01-03-1982

3 NaN NaN NaN NaN

4 NaN NaN NaN NaN

REGISTERED_STATE AUTHORIZED_CAP PAIDUP_CAPITAL INDUSTRIAL_CLASS \

0 Tamil Nadu 0.0 0.0 NaN

1 Tamil Nadu 0.0 0.0 NaN

2 Tamil Nadu 0.0 0.0 NaN

3 Tamil Nadu 0.0 0.0 NaN

4 Tamil Nadu 0.0 0.0 NaN

PRINCIPAL_BUSINESS_ACTIVITY_AS_PER_CIN \

0 Agriculture & allied

1 Agriculture & allied

2 Agriculture & allied

3 Agriculture & allied

4 Agriculture & allied

REGISTERED_OFFICE_ADDRESS REGISTRAR_OF_COMPANIES \

0 AMBLE SIDE, NO.8(OLD NO.30),3RD FLOOR KHADER N... ROC DELHI

1 FLAT NO. 6, 1st FLOOR, 113/113ARAMA NAICKEN ST... ROC DELHI

2 SRILANKAN AIRLINES LIMITED, VIJAYA TOWERSNO-4,... ROC DELHI

3 GOLD CREST 24 55 NORTHUSMAN ROAD T NAGAR ROC DELHI

4 FF-3 Palani Centre32 Venkat Naryan Road Nagar ROC DELHI

EMAIL_ADDR LATEST_YEAR_ANNUAL_RETURN \

0 NaN NaN

1 shuchi.chug@asa.in NaN

2 shree16us@yahoo.com NaN

3 NaN NaN

4 karthick9999@yahoo.com NaN

LATEST_YEAR_FINANCIAL_STATEMENT

0 NaN

1 NaN

2 NaN

3 NaN

4 NaN

In[6]:

print(len(dataset))

dataset.head(1)

Out[]:

150871

DATA PREPROCESSING:

#Handling Missing values

In[7]:

```
print(f"Total Values : {len(dataset)}\n")

for x in dataset.columns:

    print(f'{len(dataset)-dataset[x].count()} values missing in {x}')
```

Out[ ]:

```
Total Values : 150871

0 values missing in CORPORATE_IDENTIFICATION_NUMBER

0 values missing in COMPANY_NAME

0 values missing in COMPANY_STATUS

334 values missing in COMPANY_CLASS

334 values missing in COMPANY_CATEGORY

334 values missing in COMPANY_SUB_CATEGORY

39 values missing in DATE_OF_REGISTRATION

0 values missing in REGISTERED_STATE

0 values missing in AUTHORIZED_CAP

0 values missing in PAIDUP_CAPITAL

310 values missing in INDUSTRIAL_CLASS

0 values missing in PRINCIPAL_BUSINESS_ACTIVITY_AS_PER_CIN
```

90 values missing in REGISTERED_OFFICE_ADDRESS

174 values missing in REGISTRAR_OF_COMPANIES

38129 values missing in EMAIL_ADDR

75889 values missing in LATEST_YEAR_ANNUAL_RETURN

75782 values missing in LATEST_YEAR_FINANCIAL_STATEMENT

In[8]:

Dataset.iloc[:,:-1].values

Out[ ]:

```
array([['F00643', 'HOCHTIEFF AG,   ', 'NAEF', ..., 'ROC\xa0DELHI', nan,
        nan],
       ['F00721',
        'SUMITOMO CORPORATION (SUMITOMO SHOJI KAISHA LIMITED)  ', 'ACTV',
        ..., 'ROC\xa0DELHI', 'shuchi.chug@asa.in', nan],
       ['F00892', 'SRILANKAN AIRLINES LIMITED   ', 'ACTV', ...,
        'ROC\xa0DELHI', 'shree16us@yahoo.com', nan],
       ...,
       ['U74997TZ2016PTC027802',
```

'POLYGAR FARM SOLUTIONS PRIVATE LIMITED ', 'STOF', ...,

'ROC\xa0COIMBATORE', 'prashanthramana@gmail.com', nan],

['U74997TZ2018PTC030177',

'PANDIYA AGRI SOLUTIONS PRIVATE LIMITED ', 'ACTV', ...,

'ROC\xa0COIMBATORE', 'sathishpandiya@gmail.com', '31-03-2019'],

['U74997TZ2019PTC032491', 'NROOT TECHNOLOGIES PRIVATE LIMITED ',

'ACTV', ..., 'ROC\xa0COIMBATORE', 'nroottechnologies@gmail.com',
nan]], dtype=object)

Exploratory Data Analysis (EDA):

- Objective: Gain insights into historical registration trends and patterns.
- Activities:
    - Create visualizations (time series plots, histograms, maps).
    - Conduct statistical analysis (seasonality, trends, anomalies).
    - Explore correlations and relationships.
    - Hypothesis testing (if relevant).

PYTHON PROGRAM:

In[9]:

```
number__of__companies=number__of__companies.groupby(by='DATE__OF__REGISTRATION').size().reset__index(name='No__of__companies')
```

```
number__of__companies
```

Out[ ]:

	DATE_OF_REGISTRATION	No_of_companies
0	01-01-1914	1
1	01-01-1930	1
2	01-01-1937	1
3	01-01-1942	1
4	01-01-1945	1
...	...	...
13535	31-12-2013	23
13536	31-12-2014	6

13537	31-12-2015	18
13538	31-12-2018	21
13539	31-12-2019	54

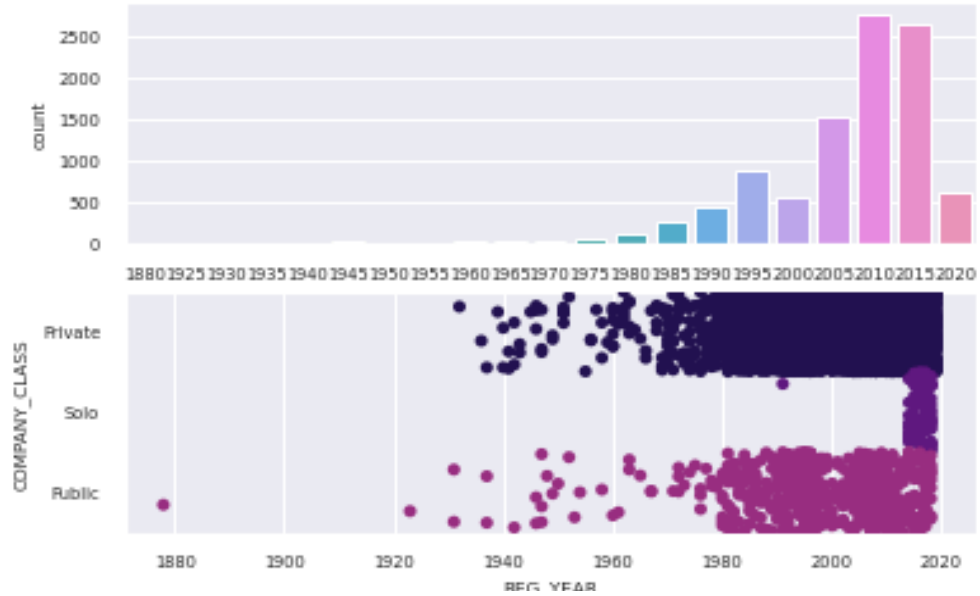
13540 rows × 2 columns

In[10]:

```
f, ax = plt.subplots(2)

#Counting all the number of companies by REG_YEAR

sns.countplot(x="REG_YEAR_5BIN",dataset=df, ax = ax[0])

#Year of registration by COMPANY_CLASS

sns.stripplot(x="REG_YEAR",  y="COMPANY_CLASS",   data=df, jitter=0.5,

 ax = ax[1])
```

Out[ ]:

```
<AxesSubplot:xlabel='REG_YEAR', ylabel='COMPANY_CLASS'>
```

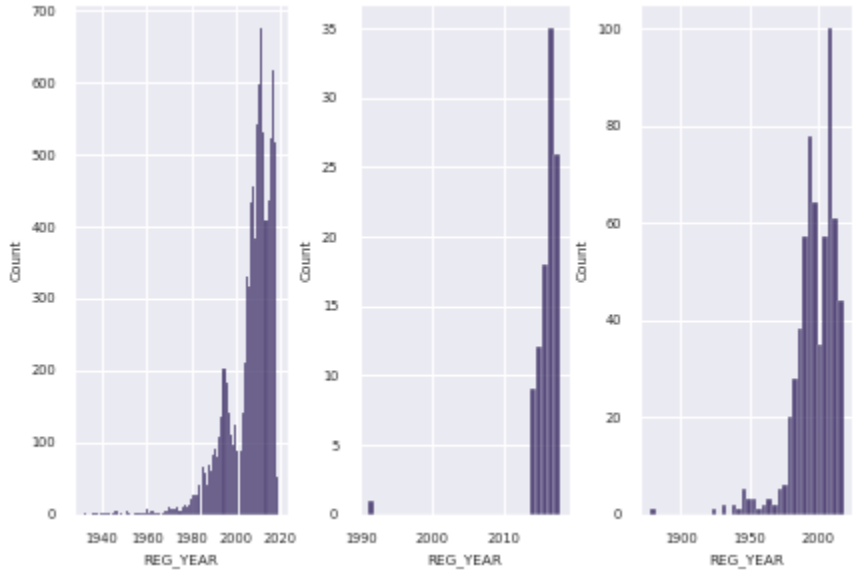


In[11]:

```
f, ax = plt.subplots(1, len(df["COMPANY_CLASS"].unique()))

f.tight_layout()

y=0

print(df["COMPANY_CLASS"].unique())

for x in df["COMPANY_CLASS"].unique():

  sns.histplot(x="REG_YEAR",

        data=df[df["COMPANY_CLASS"]==x],

        ax=ax[y])

  y+=1
```

Out[ ]:

```
['Private' 'Solo' 'Public']
```

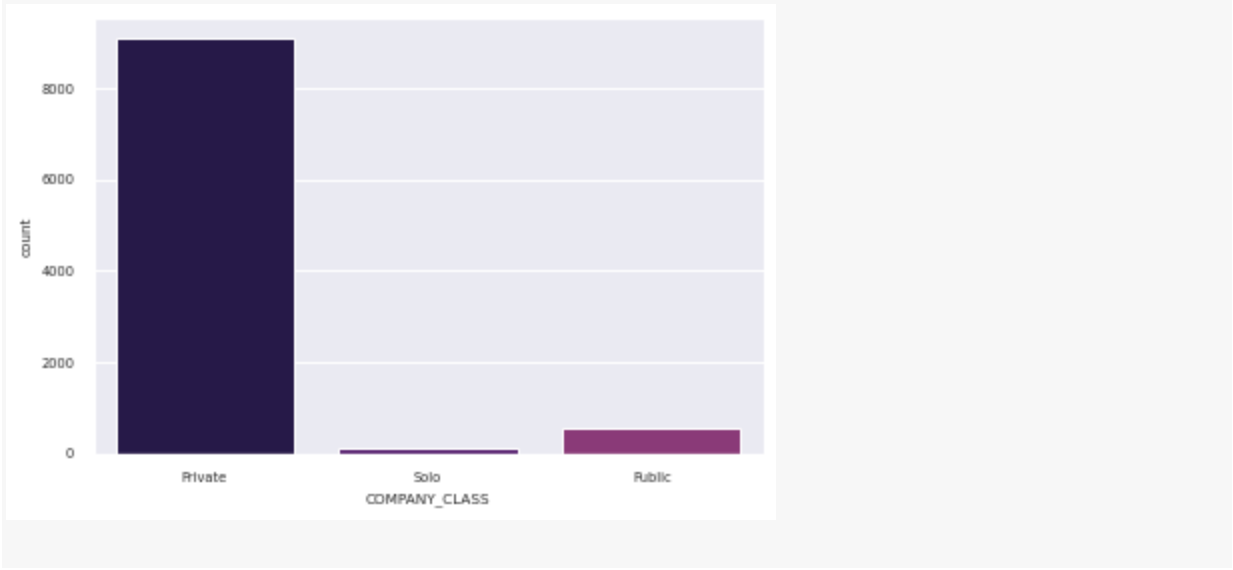


In[12]:

sns.countplot(x="COMPANY_CLASS"data=df[df["REG_YEAR"] >= 1982]) #Change year

Out[ ]:

```
<AxesSubplot:xlabel='COMPANY_CLASS', ylabel='count'>
```



Feature Engineering:

- Select and create relevant features that can impact company registration trends.
- Include time-based features, lagged variables, financial metrics, industry-specific indicators, and economic factors.
- Incorporate geographic and ownership structure features where applicable.

- Utilize natural language processing (NLP) for textual data analysis.
- Ensure appropriate scaling and normalization of features.

Ensemble methods :

Bagging (Bootstrap Aggregating):

- Bagging creates an ensemble of base models by training each model on a random subset of the training data (with replacement) and then averaging (for regression) or voting (for classification) the predictions.
- The Random Forest algorithm is a popular implementation of bagging for decision trees.

Boosting:

- Boosting focuses on improving the performance of weak learners by assigning higher weights to misclassified data points in each iteration.
- Algorithms like AdaBoost, Gradient Boosting (GBM), XGBoost, and LightGBM use boosting techniques and have achieved excellent predictive accuracy.

Stacking (Stacked Generalization):

- Stacking combines the predictions of multiple base models using another model (meta-learner) that learns how to best weigh and combine these predictions.
- Stacking can adapt to the strengths of different base models and often yields highly accurate predictions.

Voting Classifiers/Regression:

- Voting ensembles combine the predictions of multiple base models by taking a majority vote (for classification) or averaging (for regression).
- Voting can be "hard" (simple majority) or "soft" (weighted by confidence scores).

Gradient Boosting Variants:

- Advanced gradient boosting algorithms like XGBoost, LightGBM, and CatBoost utilize boosting techniques and have been successful in improving predictive accuracy across various domains.

Ensemble of Diverse Models:

- Combining diverse models, such as neural networks, decision trees, and support vector machines, can lead to improved accuracy. Each model may capture different aspects of the data.

Feature Engineering and Selection:

- Ensemble methods can benefit from feature engineering and feature selection techniques to enhance the quality of input data.

Hyperparameter Tuning:

- Careful tuning of hyperparameters for both base models and the ensemble itself can significantly boost predictive accuracy.

3.BUILD LOADING AND PREPROCESSING THE DATASET:

Loading dataset:

• Loading the dataset using machine learning is the process of bringing the data into the machine
learning environment so that it can be used to train and evaluate a model.
• The specific steps involved in loading the dataset will vary depending on the machine learning
library or framework that is being used.
However, there are some general steps that are common to most machine learning frameworks:

1..Identify the dataset:

The first step is to identify the dataset that you want to load. This dataset may be stored in a local file, in a
database, or in a cloud storage service.

2.Load the dataset:

Once you have identified the dataset, you need to load it into the machine learning environment. This may
involve using a built-in function in the machine learning library, or it may involve writing your own code.

3.Preprocess the dataset:

Once the dataset is loaded into the machine learning environment, you may need to preprocess it before you
can start training and evaluating your model. This may involve cleaning the data, transforming the data into
a suitable format, and splitting the data into training and test sets

PYTHON PROGRAM:

```
#load the dataset
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150871 entries, 0 to 150870
Data columns (total 17 columns):
#   Column                          Non-Null Count   Dtype
---  ------                          --------------  -----
 0   CORPORATE_IDENTIFICATION_NUMBER  150871 non-null  object
 1   COMPANY_NAME                     150871 non-null  object
 2   COMPANY_STATUS                   150871 non-null  object
 3   COMPANY_CLASS                    150537 non-null  object
 4   COMPANY_CATEGORY                 150537 non-null  object
 5   COMPANY_SUB_CATEGORY             150537 non-null  object
 6   DATE_OF_REGISTRATION             150832 non-null  object
 7   REGISTERED_STATE                 150871 non-null  object
 8   AUTHORIZED_CAP                   150871 non-null  float64
 9   PAIDUP_CAPITAL                   150871 non-null  float64
 10  INDUSTRIAL_CLASS                 150561 non-null  object
 11  PRINCIPAL_BUSINESS_ACTIVITY_AS_PER_CIN  150871 non-null  object
 12  REGISTERED_OFFICE_ADDRESS        150781 non-null  object
 13  REGISTRAR_OF_COMPANIES           150697 non-null  object
 14  EMAIL_ADDR                       112742 non-null  object
 15  LATEST_YEAR_ANNUAL_RETURN        74982 non-null   object
 16  LATEST_YEAR_FINANCIAL_STATEMENT  75089 non-null   object
dtypes: float64(2), object(15)
memory usage: 19.6+ MB
```

In[13]:

```
for x in dataset.columns:
  print(f'{x} :
{len(dataset[x].unique())}\n{dataset[x].unique()[:20]}\n')
```

Out[ ]:

```
CORPORATE_IDENTIFICATION_NUMBER : 150871
['F00643' 'F00721' 'F00892' 'F01208' 'F01218' 'F01265' 'F01269' 'F01311'
 'F01314' 'F01412' 'F01426' 'F01468' 'F01543' 'F01544' 'F01563' 'F01565'
 'F01566' 'F01589' 'F01593' 'F01618']
```

COMPANY_NAME : 150560
['HOCHTIEFF AG, '
 'SUMITOMO CORPORATION (SUMITOMO SHOJI KAISHA LIMITED) '
 'SRILANKAN AIRLINES LIMITED ' 'CALTEX INDIA LIMITED '
 'GE HEALTHCARE BIO-SCIENCES LIMITED '
 'CAIRN ENERGY INDIA PTY. LIMITED ' 'TORIELLI S.R.L '
 'HARDY EXPLORATION & PRODUCTION (INDIA) INC.. '
 'HOCHTIOF AKTIENGESELLSHARFF VORM GFBR HELFMANN '
 'EPSON SINGAPORE PVT LTD ' 'CARGOLUX AIRLINES INTERNATIONAL S A '
 'CHO HEUNG ELECTRIC INDUSTRIAL COMPANY LIMITED '
 'NYCOMED ASIA PACIFIC PTE LIMITED ' 'CHERRINGTON ASIA LTD '
 'SHIMADZU ASIA PACIFIC PTE LIMITED '
 'CORK INTERNATIONAL PTY LIMITED ' 'ERBIS ENGG COMPANY LIMITED '
 'RALF SCHNEIDER HOLDING GMBH ' 'MITRAJAYA TRADING PRIVATE LIMITED '
 'HEAT AND CONTROL PTY LIMITED ']

COMPANY_STATUS : 11
['NAEF' 'ACTV' 'ULQD' 'LIQD' 'AMAL' 'DISD' 'UPSO' 'STOF' 'D455' 'CLLP'
 'CLLD']

COMPANY_CLASS : 4
[nan 'Public' 'Private' 'Private(One Person Company)']

COMPANY_CATEGORY : 4
[nan 'Company limited by Shares' 'Company Limited by Guarantee'
 'Unlimited Company']

COMPANY_SUB_CATEGORY : 6
[nan 'Non-govt company' 'Union Govt company'
 'Subsidiary of Foreign Company' 'State Govt company'
 'Guarantee and Association comp']

DATE_OF_REGISTRATION : 13541
['01-12-1961' nan '01-03-1982' '05-09-1995' '11-04-1996' '25-04-1997'
 '11-06-1997' '27-10-1998' '01-05-2000' '13-07-1999' '02-11-1999'
 '14-06-2000' '17-07-2000' '24-01-2001' '08-03-2001' '22-03-2001'
 '16-08-2001' '21-11-2001' '24-12-2001' '23-09-1995']

REGISTERED_STATE : 1
['Tamil Nadu']

AUTHORIZED_CAP : 1623
[0.000e+00 1.250e+07 1.500e+10 1.500e+08 5.000e+05 2.500e+08 5.000e+06

5.000e+07 1.600e+07 5.500e+07 8.000e+06 3.000e+08 4.000e+08 4.000e+07
1.353e+08 1.600e+08 2.000e+08 1.000e+08 1.000e+05 7.000e+07]

PAIDUP_CAPITAL : 16294
[0.00000000e+00 6.27350000e+06 1.16730000e+08 3.83500000e+07
4.00000000e+07 1.89066750e+08 4.99656600e+07 1.04117300e+07
4.62420000e+07 3.00000000e+01 2.63300000e+08 2.04715000e+08
2.81676800e+08 3.45500000e+07 1.07000000e+08 1.14525000e+08
1.59827370e+08 4.93480000e+07 7.00000000e+03 2.16140336e+09]

INDUSTRIAL_CLASS : 1458
[nan '01117' '01119' '01122' '01132' '01133' '01211' '01222' '01409'
'01542' '02310' '02511' '03210' '05001' '08031' '11101' '13206' '14200'
'14299' '15100']

PRINCIPAL_BUSINESS_ACTIVITY_AS_PER_CIN : 17
['Agriculture & allied' 'Mining and quarrying' 'Manufacturing'
'Electricity gas and water supply' 'Construction'
'Wholesale and retail trade repair of motor vehicles motorcycles and personal and household goods'
'Hotels and restaurants' 'Transport storage and communications'
'Financial intermediation' 'Real estate renting and business activities'
'Health and social work'
'Other community social and personal service activities'
'Extraterritorial organizations and bodies' 'Unclassified'
'Activities of private households as employers and undifferentiated production activities of private households'
'Public administration and defence compulsory social security'
'Education']

REGISTERED_OFFICE_ADDRESS : 142911
['AMBLE SIDE, NO.8(OLD NO.30),3RD FLOOR KHADER NAWAZ KHAN ROAD,NUGAMBA '
'FLAT NO. 6, 1st FLOOR, 113/113ARAMA NAICKEN STREET, NUNGAMBAKKAM '
'SRILANKAN AIRLINES LIMITED, VIJAYA TOWERSNO-4, KODAMBAKKAM HIGH ROAD, NUNGAMBAKKA '
'GOLD CREST 24 55 NORTHUSMAN ROAD T NAGAR '
'FF-3 Palani Centre32 Venkat Naryan Road Nagar '
' WELLINGTON PLAZA 90,ANNA SALAI,CHENNAI-600002 '
'6, Mangayarkarsi Nagar, Lakshmi Nagar5th Street, Nanganallur '
'5TH FLOOR,WESTMINISTER BUILDING,108 DR.RADHAKRISHNAN SALAI '
'NEW NO.86, OLD NO.98, POLYHOSE TOWER 4TH FLOOR,MOUNT ROAD, GUINDY, '
'7C CEATURY PLAZA560 MOUNT ROAD ' 'OFFICE NO 91MEENAKBAKAM AIRPORT '
'129, MANPUR VILLAGE,SRIPERUMBUDUR, '
'A D 46 1ST STREETANNA NAGAR MADRAS ' '10HADDOWS ROAD '
'FIRST FLOOR, NO:1063, MUNUSAMY SALAI,K K NAGAR '

'ARJAY APEX CENTRE NO-24COLLEGE ROAD '
'39,2nd Main Road,Raja Annamalaipuram, '
"FLAT C, 'SAI VASANTHAM', 3rd FLOOR,NEW NO.41, 3rd MAIN ROAD,GANDHI NAGAR, A "
'OLD NO 148 NEW NO 52 ELDAMSROAD TEYNAMPET CHENNAI '
'A40 OLD NO 26 6TH STREETANNA NAGAR ']

REGISTRAR_OF_COMPANIES : 5
['ROC\xa0DELHI' 'ROC\xa0COIMBATORE' 'ROC\xa0CHENNAI' nan
'ROC\xa0HYDERABAD']

EMAIL_ADDR : 79941
[nan 'shuchi.chug@asa.in' 'shree16us@yahoo.com' 'karthick9999@yahoo.com'
'neerja.sharma@cairnindia.com' 'chennai@torielliindia.com'
'venkatesh.v@hardyoil.co.in' 'kumar@international.hochtief.de'
'chowelaccounts@gmail.com' 'kousik@vsnl.com' 'ncrajagopal@gmail.com'
'direx@vsnl.com' 'stsogawa@minebea.co.in' 'joe@obara.co.in'
'pverma@vkvermaco.com' 'jeeva@fecindia.com' 'chennaiadmin@ksaiyar.com'
'accounts.ho@sinarjernih.co.in' 'svrajacdm@yahoo.com' 'socamkr@vsnl.net']

LATEST_YEAR_ANNUAL_RETURN : 170
[nan '31-03-2019' '31-03-2018' '31-03-2012' '31-03-2016' '31-03-2017'
'31-03-2010' '31-03-2014' '31-03-2013' '30-06-2007' '31-03-2009'
'31-03-2015' '30-06-2013' '31-03-2008' '31-03-2007' '30-09-2019'
'31-03-2011' '30-09-2013' '31-12-2010' '31-03-2006']

LATEST_YEAR_FINANCIAL_STATEMENT : 139
[nan '31-03-2019' '31-03-2018' '31-03-2012' '31-03-2011' '31-03-2017'
'31-03-2016' '31-03-2010' '31-03-2014' '31-03-2013' '30-06-2007'
'31-03-2009' '31-03-2015' '31-12-2006' '31-03-2008' '31-03-2007'
'30-06-2011' '30-09-2013' '30-06-2012' '31-12-2010']

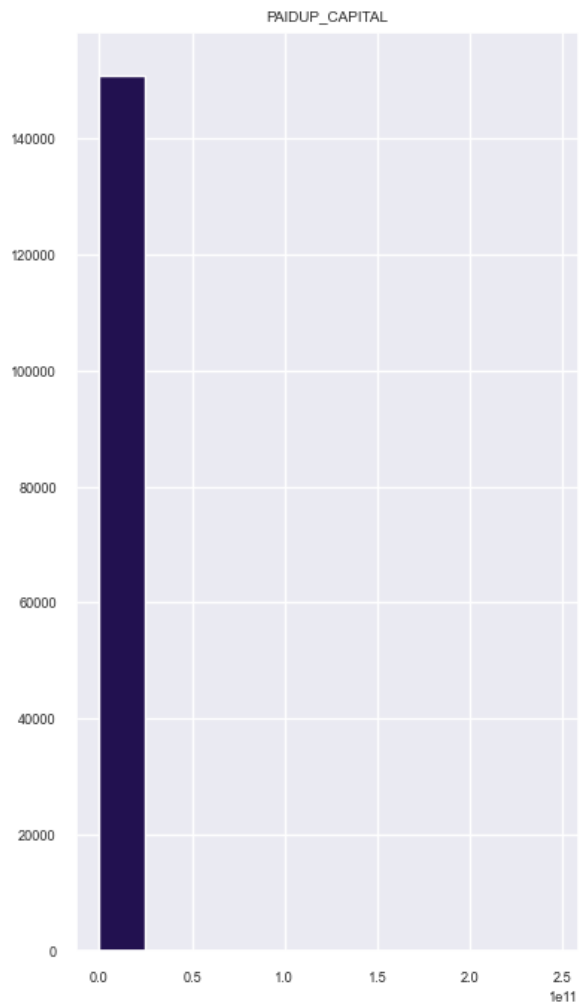
Visualization and preprocessing of dataset:

In[ ]:

```
dataset.hist(figsize=(10,8))
```

Out[ ]:

```
array([[<Axes: title={'center': 'AUTHORIZED_CAP'}>,
        <Axes: title={'center': 'PAIDUP_CAPITAL'}>]], dtype=object)
```



Importance of loading and preprocessing dataset:

1. Data Quality Assurance:

● Loading data allows you to inspect its quality. You can identify missing values, outliers, inconsistencies, and errors in the dataset. Addressing these issues is vital to ensure accurate and reliable results.

2. Data Exploration:

● Processing the dataset enables you to explore and understand its characteristics. You can calculate

basic statistics, visualize the data, and identify patterns or trends. This exploration informs subsequent analysis.

3. Data Preprocessing:

● Datasets are rarely in the perfect format for analysis. Preprocessing involves cleaning, transforming, and structuring the data for analysis. It may include handling missing values, encoding categorical variables, scaling, and feature engineering.

4. Feature Selection:

● Careful data processing allows you to select relevant features (variables) and discard irrelevant or
redundant ones. This simplifies models, improves interpretability, and reduces overfitting.

5. Model Performance:

● High-quality data and preprocessing contribute to better model performance. Clean and well-processed data improves the accuracy and generalization of machine learning models.

6. Data Security:

● Handling data includes measures for data security and privacy. Ensuring sensitive information is protection is crucial to meet legal and ethical requirements.

7. Efficiency:

● Processing data efficiently can save computational resources and time. Techniques like dimensionality reduction can make analysis faster and more manageable.

8. Interpretability:

● Well-processed data leads to more interpretable results. Understanding how the data was manipulated allows for a clearer interpretation of analysis outcomes.

9. Consistency:

● Consistent data processing practices across projects and datasets facilitate collaboration, documentation, and maintenance of data analysis workflows.

10. Data Reproducibility: - By documenting the data loading and processing steps, you enable others to
reproduce your analysis, which is crucial forvalidation and peer review.

4.PERFORMING EXPLORATORY DATA ANALYSIS,FEATURE ENGINEERING,AND PREDICTIVE MODELING:

1. Exploratory Data Analysis (EDA):

Exploratory Data Analysis is the first crucial step in our journey. EDA involves understanding the dataset's characteristics, uncovering patterns, and identifying relationships within the data. Key components of EDA include data cleaning, summary statistics, and data visualization. By examining the data in depth, we gain insights that pave the way for effective predictive modeling. EDA allows us to:

- Identify missing values and outliers, ensuring data quality.
- Summarize data through statistical measures.
- Visualize data using various charts and plots to understand its distribution and relationships.
- Recognize patterns and trends that can guide feature engineering and modeling.

In[15]:

```
print(f"Total Values : {len(dataset)}\n")
for x in dataset.columns:
  print(f'{len(dataset)-dataset[x].count()} values missing in {x}')
```

Out[ ]:

```
Total Values : 150871

0 values missing in CORPORATE_IDENTIFICATION_NUMBER
0 values missing in COMPANY_NAME
0 values missing in COMPANY_STATUS
334 values missing in COMPANY_CLASS
334 values missing in COMPANY_CATEGORY
334 values missing in COMPANY_SUB_CATEGORY
39 values missing in DATE_OF_REGISTRATION
```

0 values missing in REGISTERED_STATE
0 values missing in AUTHORIZED_CAP
0 values missing in PAIDUP_CAPITAL
310 values missing in INDUSTRIAL_CLASS
0 values missing in PRINCIPAL_BUSINESS_ACTIVITY_AS_PER_CIN
90 values missing in REGISTERED_OFFICE_ADDRESS
174 values missing in REGISTRAR_OF_COMPANIES
38129 values missing in EMAIL_ADDR
75889 values missing in LATEST_YEAR_ANNUAL_RETURN
75782 values missing in LATEST_YEAR_FINANCIAL_STATEMENT

In[16]:

```
sns.stripplot(x="REG_YEAR",
              y="PRINCIPAL_BUSINESS",
              hue="COMPANY_CLASS",
              data=df, jitter=0.3)
```

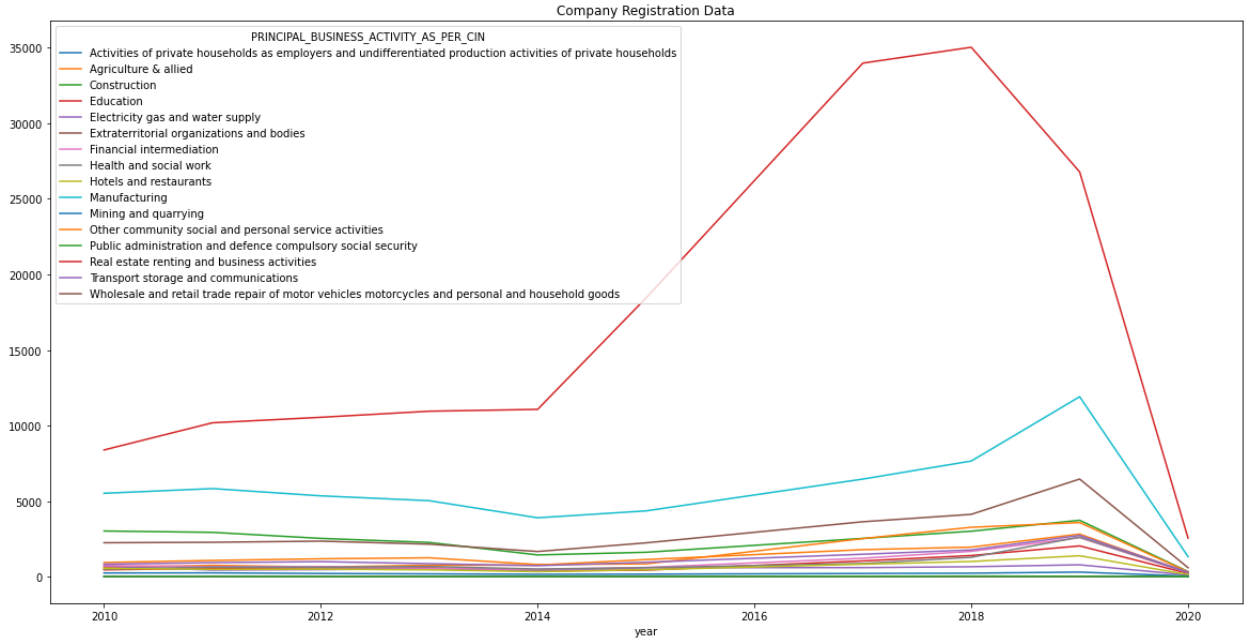
Out[ ]:

```
<AxesSubplot:xlabel='REG_YEAR', ylabel='PRINCIPAL_BUSINESS'>
```

In[ ]:
table=search.groupby('year')['PRINCIPAL_BUSINESS_ACTIVITY_AS_PER_CIN'].value_counts().unstack().fillna(0)

In[ ]:
plotting = table.plot.line(figsize=(20,10), fontsize=10, title="Company Registration Data", legend=True)

Out[ ]:



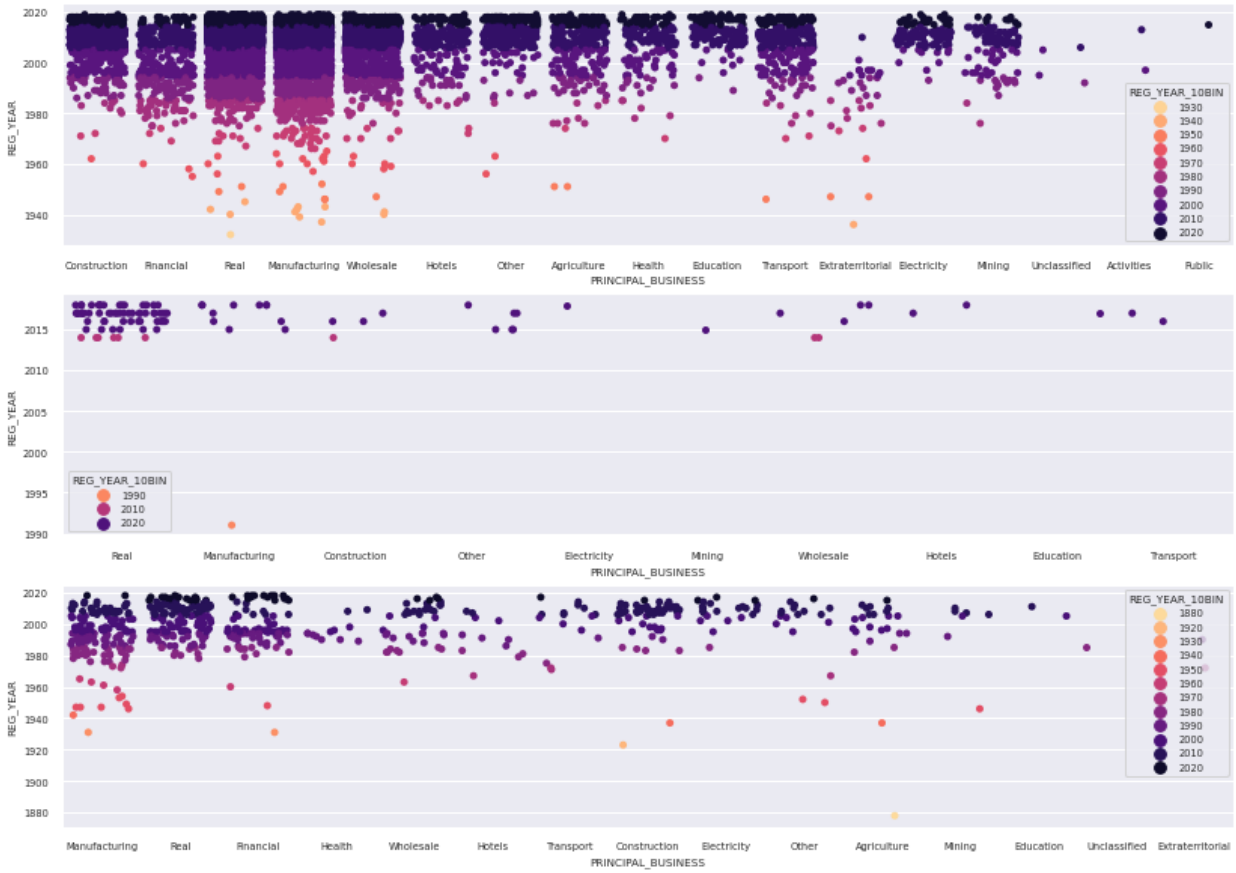
In[17]:

```
#Companies in respect to COMPANY_CLASS over time.
y=0
f, ax = plt.subplots(len(df["COMPANY_CLASS"].unique()),1)
f.subplots_adjust(top=1, bottom=-0.9, left=-0.9, hspace=0.2)
print(df["COMPANY_CLASS"].unique())
for x in df["COMPANY_CLASS"].unique():
    sns.stripplot(y="REG_YEAR",
                  x="PRINCIPAL_BUSINESS",
```

```
                  data=df[df["COMPANY_CLASS"]==x],
                  hue="REG_YEAR_10BIN",
                  palette="magma_r",
                  jitter=0.4,
                  ax=ax[y])
    y+=1
```

Out[]:

```
['Private' 'Solo' 'Public']
```



In[18]:

```
y=0; f, ax = plt.subplots(len(df["REGISTERED_STATE"].unique()),1)
f.subplots_adjust(top=10, bottom=-0.9, left=-0.5, hspace=0.2)
for x in df["REGISTERED_STATE"].unique():
    sns.histplot(y='REG_YEAR',
                 x='PRINCIPAL_BUSINESS',
                 #jitter=0.3,
                 hue="REGISTERED_STATE",
```

```
                    palette="magma_r",
                    data= df[df["REGISTERED_STATE"]==x],
                    ax=ax[y])
    y+=1
```

In[19]:
df["AUTHORIZED_CAP"].describe()

Out[ ]:
count 1.127180e+05
mean   4.574342e+07
std   1.628890e+09
min   0.000000e+00
25%   1.000000e+05
50%   1.000000e+06
75%   2.500000e+06
max   3.000000e+11
Name: AUTHORIZED_CAP, dtype: float64

In[ ]:
group =
search.groupby('PRINCIPAL_BUSINESS_ACTIVITY_AS_PER_CIN')
print(group.size())
group.size().plot.pie(y='PRINCIPAL_BUSINESS_ACTIVITY_AS_PER_CIN', figsize=(40,40), fontsize=40)

Out[ ]:

```
PRINCIPAL_BUSINESS_ACTIVITY_AS_PER_CIN

Activities of private households as employers and undifferentiated

production activities of private households        60

Agriculture & allied

11425
```

Construction

23378

Education

7636

Electricity gas and water supply

5221

Extraterritorial organizations and bodies

15

Financial intermediation

9413

Health and social work

8552

Hotels and restaurants

6087

Manufacturing

57281

Mining and quarrying

1958

Other community social and personal service activities

15744

Public administration and defence compulsory social security

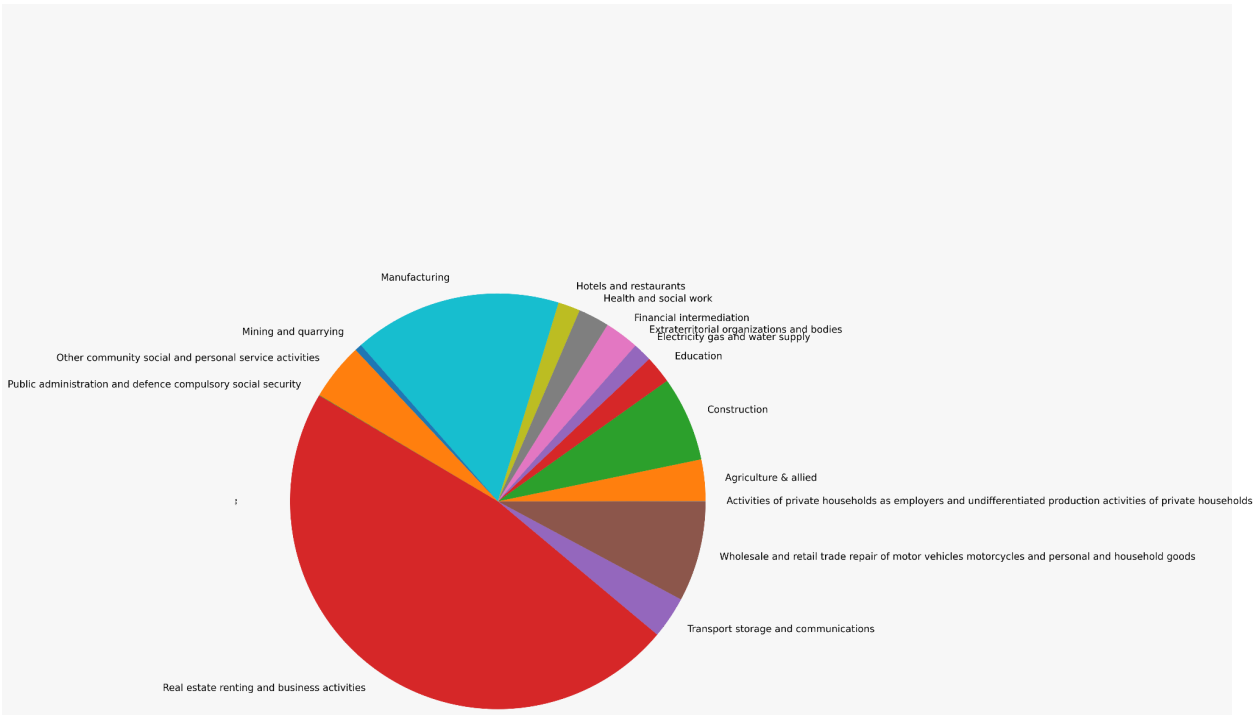
113

Real estate renting and business activities

167907

```
Transport storage and communications

11535

Wholesale and retail trade repair of motor vehicles motorcycles and

personal and household goods                    27700

dtype: int64

Out[ ]:
```

<AxesSubplot:ylabel='None'>



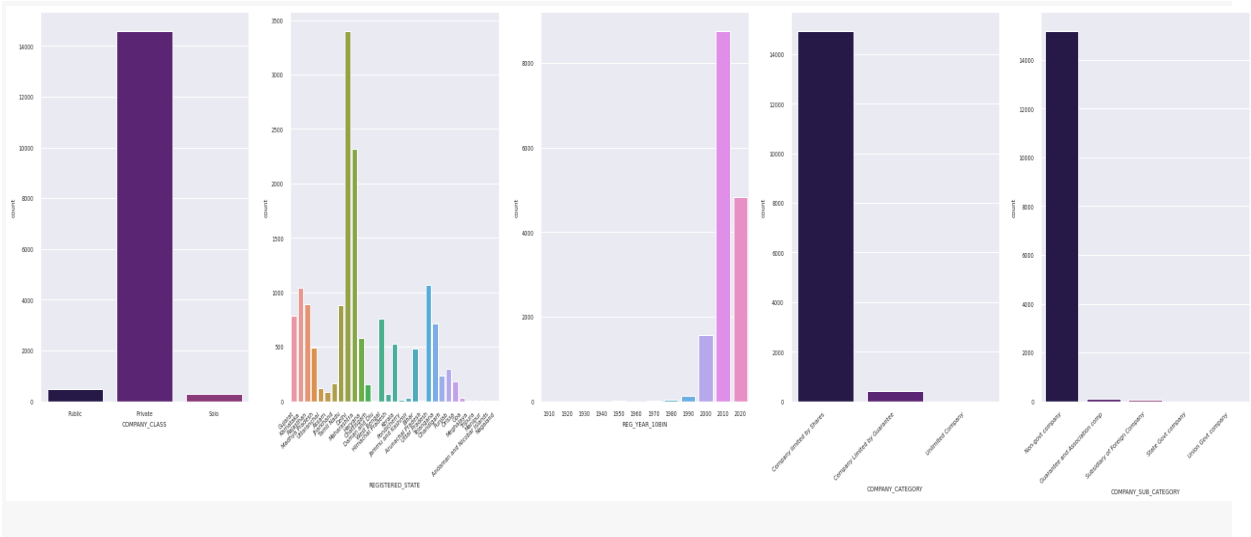
In[20]:

```
f, axes = plt.subplots(1,5)
f.subplots_adjust(top=0.5, bottom=-0.9, left=-3, hspace=0.2)
ax1 = sns.countplot(x="COMPANY_CLASS",
data=df2[df2["PRINCIPAL_BUSINESS"] == "Education"], ax=axes[0])
```

```
ax2 = sns.countplot(x="REGISTERED_STATE",
data=df2[df2["PRINCIPAL_BUSINESS"] == "Education"], ax=axes[1])
ax3 = sns.countplot(x="REG_YEAR_10BIN",
data=df2[df2["PRINCIPAL_BUSINESS"] == "Education"], ax=axes[2])
ax4 = sns.countplot(x="COMPANY_CATEGORY",
data=df2[df2["PRINCIPAL_BUSINESS"] == "Education"], ax=axes[3])
ax5 = sns.countplot(x="COMPANY_SUB_CATEGORY",
data=df2[df2["PRINCIPAL_BUSINESS"] == "Education"], ax=axes[4])
ax2.set_xticklabels(ax2.get_xticklabels(), rotation=40, ha="right")
ax4.set_xticklabels(ax4.get_xticklabels(), rotation=40, ha="right")
ax5.set_xticklabels(ax5.get_xticklabels(), rotation=40, ha="right")
```

Out[ ]:

```
[Text(0, 0, 'Non-govt company'),
 Text(1, 0, 'Guarantee and Association comp'),
 Text(2, 0, 'Subsidiary of Foreign Company'),
 Text(3, 0, 'State Govt company'),
 Text(4, 0, 'Union Govt company')]
```



Visualizing Correlation:

In[20]:
dataset.corr(numeric_only=True)

Out[ ]:

	AUTHORIZED_CAP	PAIDUP_CAPITAL
AUTHORIZED_CAP	1.000000	0.971255
PAIDUP_CAPITAL	0.971255	1.000000

In[21]:

```
plt.figure(figsize=(10,5))
sns.heatmap(dataset.corr(numeric_only = True), annot=True)
```

Out[ ]:
<Axes: >

Feature Engineering:

Feature engineering is the process of creating or transforming features (variables) in the dataset to enhance model performance.performance. Effective feature engineering can significantly impact the predictive power of our models. In this phase, we will:

- Select relevant features and discard irrelevant ones.
- Create new features that capture domain-specific information.
- Handle categorical variables through one-hot encoding or other suitable techniques.
- Normalize or scale numerical features for improved model convergence.
- Prepare the data for modeling by addressing any data-specific challenges.

3. Predictive Modeling:

With the data well-prepared, we will venture into predictive modeling. Predictive modeling is the core of our project, where we utilize various machine learning and AI algorithms to make informed predictions. This phase involves:

- Data splitting: Separating the dataset into training, validation, and test sets.
- Model selection: Choosing the most suitable algorithm for the specific prediction task.
- Model training: Training the selected model using the training data.
- Model evaluation: Assessing the model's performance using various metrics.
- Hyperparameter tuning: Optimizing the model's parameters to enhance performance.
- Model validation: Testing the model on unseen data to ensure generalizability.
- Interpretability: Employing techniques to understand how the model makes predictions,ensuring transparency.

★ Linear Regression
★ Decision Trees
★ Random Forest
★ Gradient Boosting (e.g., XGBoost or LightGBM)
★ Neural Networks (Deep Learning)

PYTHON PROGRAM:

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

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import r2_score,
mean_absolute_error,mean_squared_error
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import Lasso
from sklearn.ensemble import RandomForestRegressor
from sklearn.svm import SVR
import Xgboost as Xg
%matplotlib inline

import warnings
warnings.filterwarnings("ignore")
/opt/conda/lib/python3.10/site-packages/scipy/__init__.py:146:
UserWarning: A NumPy version >=1.16.5 and <1.23.0 is required
for this version of SciPy (detected version 1.23.5
warnings.warn(f"A NumPy version >={np_minversion} and
<{np_maxversion}"
```

In[21]:

```
dataset=pd.read_csv('Data_Gov_Tamil_Nadu.csv',encoding=("ISO-88
59-1"),low_memory=False)
```

Linear Regression:

In[22]:

```
model_lr=LinearRegression()
```

In[23]:

```
model_lr.fit(X_train_scal, Y_train)
```

In[24]:

```
lin_reg = LinearRegression()
lin_reg.fit(X_train, y_train)
predictions = lin_reg.predict(X_test)
```

```
mae, mse, rmse, r_squared = evaluation(y_test, predictions)
print("MAE:", mae)
print("MSE:", mse)
print("RMSE:", rmse)

print("R2 Score:", r_squared)
print("-"*30)rmse_cross_val = rmse_cv(lin_reg)
print("RMSE Cross-Validation:", rmse_cross_val)
new_row = {"Model": "LinearRegression","MAE": mae, "MSE": mse,
"RM
SE": rmse, "R2 Score": r_squared, "RMSE (Cross-Validation)":
rmse_cross
_val}models = models.append(new_row, ignore_index=True)
```

Out[ ]:

```
MAE: 23567.890565943395

MSE: 1414931404.6297863

RMSE: 37615.57396384889

R2 Score: 0.8155317822983865

------------------------------

RMSE Cross-Validation: 36326.451444669496
```

Ridge Regression:

In [25]:

```
ridge = Ridge()ridge.fit(X_train, y_train)predictions =
ridge.predict(X_test)
mae, mse, rmse, r_squared = evaluation(y_test, predictions)
```

```
print("MAE:", mae)
print("MSE:", mse)
print("RMSE:", rmse)
print("R2 Score:", r_squared)
print("-"*30)rmse_cross_val = rmse_cv(ridge)
print("RMSE Cross-Validation:", rmse_cross_val)
new_row = {"Model": "Ridge","MAE": mae, "MSE": mse, "RMSE":
rmse,
"R2 Score": r_squared, "RMSE (Cross-Validation)":
rmse_cross_val}model
s = models.append(new_row, ignore_index=True)
```

Out[ ]:

```
MAE: 23435.50371200822

MSE: 1404264216.8595588

RMSE: 37473.513537691644

R2 Score: 0.8169224907874508

------------------------------

RMSE Cross-Validation: 35887.852791598336
```

Lasso Regression:

In [26]:

```
lasso = Lasso()lasso.fit(X_train, y_train)predictions =
lasso.predict(X_test)
mae, mse, rmse, r_squared = evaluation(y_test, predictions)
print("MAE:", mae)
print("MSE:", mse)
print("RMSE:", rmse)
```

```
print("R2 Score:", r_squared)
print("-"*30)rmse_cross_val = rmse_cv(lasso)
print("RMSE Cross-Validation:", rmse_cross_val)
new_row = {"Model": "Lasso","MAE": mae, "MSE": mse, "RMSE":
rmse,
"R2 Score": r_squared, "RMSE (Cross-Validation)":
rmse_cross_val}model
s = models.append(new_row, ignore_index=True)
```

Out[ ]:

```
MAE: 23560.45808027236

MSE: 1414337628.502095

RMSE: 37607.680445649596

R2 Score: 0.815609194407292
------------------------------

RMSE Cross-Validation: 35922.76936876075
```

Elastic Net:

In [27]:

```
elastic_net = ElasticNet()elastic_net.fit(X_train, y_train)predictions =
elasti
c_net.predict(X_test)
mae, mse, rmse, r_squared = evaluation(y_test, predictions)
print("MAE:", mae)
print("MSE:", mse)
print("RMSE:", rmse)
print("R2 Score:", r_squared)
print("-"*30)rmse_cross_val = rmse_cv(elastic_net)
```

```
print("RMSE Cross-Validation:", rmse_cross_val)
new_row = {"Model": "ElasticNet","MAE": mae, "MSE": mse, "RMSE":
r

mse, "R2 Score": r_squared, "RMSE (Cross-Validation)":
rmse_cross_val}
models = models.append(new_row, ignore_index=True)
```

Out[ ]:

```
MAE: 23792.743784996732

MSE: 1718445790.1371393

RMSE: 41454.14080809225

R2 Score: 0.775961837382229

------------------------------

RMSE Cross-Validation: 38449.00864609558
```

Support Vector Machines:

In [28]:

```
svr = SVR(C=100000)svr.fit(X_train, y_train)predictions =
svr.predict(X_test)
mae, mse, rmse, r_squared = evaluation(y_test, predictions)
print("MAE:", mae)
print("MSE:", mse)
print("RMSE:", rmse)
print("R2 Score:", r_squared)
print("-"*30)rmse_cross_val = rmse_cv(svr)
```

print("RMSE Cross-Validation:", rmse_cross_val)
new_row = {"Model": "SVR","MAE": mae, "MSE": mse, "RMSE": rmse,
" R2 Score":
r_squared, "RMSE (Cross-Validation)": rmse_cross_val}models =
models.append(new_row, ignore_index=True)
Out[ ]:

MAE: 17843.16228084976

MSE: 1132136370.3413317

RMSE: 33647.234215330864

R2 Score: 0.852400492526574

RMSE Cross-Validation: 30745.475239075837

Random Forest Regressor:

In [29]:
random_forest=RandomForestRegressor(n_estimators=100)random_f
orest. fit(X_train,y_train)predictions = random_forest.predict(X_test)
mae, mse, rmse, r_squared = evaluation(y_test, predictions)
print("MAE:", mae)
print("MSE:", mse)
print("RMSE:", rmse)
print("R2 Score:", r_squared)
print("-"*30)rmse_cross_val = rmse_cv(random_forest)
print("RMSE Cross-Validation:", rmse_cross_val)
new_row = {"Model": "RandomForestRegressor","MAE": mae, "MSE":
mse, "RMSE": rmse, "R2 Score": r_squared, "RMSE

(Cross-Validation)": rmse_cross_val}models =
models.append(new_row, ignore_index=True)
Out[ ]:

MAE: 18115.11067351598

MSE: 1004422414.0219476

RMSE: 31692.623968708358

R2 Score: 0.869050886899595

RMSE Cross-Validation: 31138.863315259332

XGBoost Regressor:

In [30]:

```
xgb = XGBRegressor(n_estimators=1000,
learning_rate=0.01)xgb.fit(X_train, y_train)predictions =
xgb.predict(X_test)
mae, mse, rmse, r_squared = evaluation(y_test, predictions)
print("MAE:", mae)
print("MSE:", mse)
print("RMSE:", rmse)
print("R2 Score:", r_squared)
print("-"*30)rmse_cross_val = rmse_cv(xgb)
print("RMSE Cross-Validation:", rmse_cross_val)
new_row = {"Model": "XGBRegressor","MAE": mae, "MSE": mse,
"RMSE": rmse, "R2 Score": r_squared, "RMSE
(Cross-Validation)":rmse_cross_val}models =
models.append(new_row, ignore_index=True)
```

Out[ ]:

MAE: 17439.918396832192

MSE: 716579004.5214689

RMSE: 26768.993341578403

R2 Score: 0.9065777666861116

RMSE Cross-Validation: 29698.84961808251

Polynomial Regression (Degree=2)

In [31]:

```
poly_reg = PolynomialFeatures(degree=2)X_train_2d =
poly_reg.fit_transform(X_train)X_test_2d =
poly_reg.transform(X_test)lin_reg =
LinearRegression()lin_reg.fit(X_train_2d, y_train)predictions =
lin_reg.predict(X_test_2d)mae, mse, rmse, r_squared =
evaluation(y_test, predictions)
print("MAE:", mae)
print("MSE:", mse)
print("RMSE:", rmse)
print("R2 Score:", r_squared)
print("-"*30)rmse_cross_val = rmse_cv(lin_reg)
print("RMSE Cross-Validation:", rmse_cross_val)
new_row = {"Model": "Polynomial Regression (degree=2)","MAE":
mae, " MSE": mse,"RMSE": rmse, "R2 Score": r_squared, "RMSE
(Cross-Validation)": rmse_cross_val}models =
models.append(new_row, ignore_index=True)
```

Out[ ]:

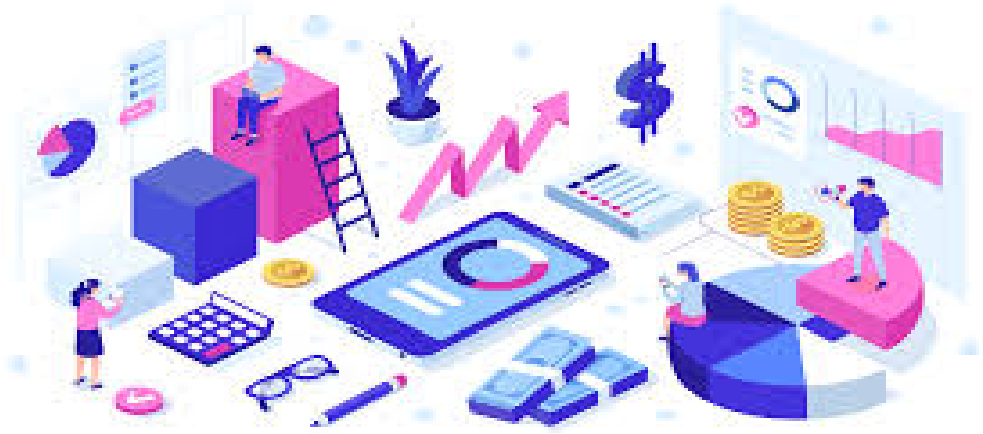
MAE: 2382228327828308.5

MSE: 1.51399115441823 42e+32

RMSE: 1.2304434787580 59e+16

R2 Score: -1.9738289005226644e+22

RMSE Cross-Validation: 36326.451444669496



ADVANTAGES:

1.Enhanced Transparency: Access to ROC data promotes transparency in the corporate sector. It allows stakeholders to access official and credible information about companies, including their structure, management, financial health, and legal compliance.

2.Legal Compliance Verification: ROC data enables verification of a company's compliance with legal and regulatory requirements. This includes the filing of financial statements, annual reports, and other necessary documents, ensuring that companies operate within the legal framework.

3.Informed Decision-Making: Investors, creditors, and business partners can make informed decisions based on ROC data. This data provides valuable insights into a company's history, financial stability, and governance, facilitating better decision-making.

4.Investor Confidence: ROC data enhances investor confidence by providing reliable and official information. This can encourage investment in the market, leading to economic growth and job creation.

5.Credit Risk Assessment: Financial institutions use ROC data to assess the creditworthiness of companies seeking loans or credit. This helps lenders manage credit risk and make sound lending decisions.

6.Market Research: Researchers and analysts can leverage ROC data for market research and industry analysis. It aids in understanding market trends, competitive landscapes, and emerging opportunities.

7.Mergers and Acquisitions (M&A) Due Diligence: Companies involved in M&A activities benefit from ROC data for due diligence, reducing risks associated with transactions and ensuring successful acquisitions.

8.Legal Proceedings and Disputes: Legal professionals use ROC data as evidence and supporting documentation in legal cases, ensuring a fair and transparent legal process.

9.Government Oversight and Regulation: Regulatory bodies and government agencies rely on ROC data to monitor and enforce corporate governance standards, taxation

compliance, and regulatory adherence, thereby maintaining the integrity of the business environment.

10.Public Awareness and Accountability: ROC data promotes public awareness of corporate activities, making companies more accountable for their actions and encouraging informed discussions among the public.

11.Policy Formulation and Reform: Policymakers use ROC data to inform the development and adjustment of policies related to corporate governance, taxation, and economic development, contributing to better business environments.

12.Efficient Resource Allocation: Companies can utilize ROC data to allocate resources more efficiently, identify growth opportunities, and mitigate risks.

13.Economic Development: A robust ROC company analysis promotes economic development by creating a fair and well-regulated business environment that attracts investment and supports business growth.

14.Strategic Decision-Making: Companies can use ROC data for strategic decision-making, such as evaluating potential business partners, assessing the financial stability of suppliers and customers, and identifying competitive advantages.

15.Anti-Corruption Efforts: Access to official company data can support anti-corruption efforts by exposing unethical practices and enhancing corporate governance.

16.Data-Driven Insights: ROC company analysis provides valuable data-driven insights into the corporate world, facilitating evidence-based decision-making in various sectors.

DISADVANTAGES:

1.Data Accuracy and Timeliness: ROC data may not always be up-to-date or entirely accurate. Companies might delay submitting required documents, resulting in outdated information. Timely updates can be crucial for decision-making.

2.Limited Financial Information: While ROC data provides financial statements, it may not offer a comprehensive view of a company's financial health. Companies can engage in creative accounting, making it difficult to assess their true financial status.

3.Privacy Concerns: Access to ROC data can raise privacy concerns for individuals associated with a company, such as directors and shareholders. Making such information publicly available can lead to potential privacy violations.

4.Complexity and Volume: Analyzing ROC data can be complex, especially in countries with a large number of registered companies. The sheer volume of data can make it challenging to extract meaningful insights efficiently.

5.Lack of Context: ROC data provides factual information but may not offer the context needed to fully understand a company's operations, market position, or strategies. Additional research may be required to fill in these gaps.

6.Limited Non-public Information: ROC data primarily consists of publicly available information. It may not include sensitive or confidential business information, making it challenging to gain a complete understanding of a company.

7.Regulatory Gaps: Depending on the jurisdiction, there may be regulatory gaps in the type and quality of information available in ROC data. Certain information may not be required to be disclosed, limiting the depth of analysis.

8.Potential for Manipulation: In some cases, companies may attempt to manipulate or misrepresent information in their ROC filings, either intentionally or due to errors, which can mislead those relying on the data.

9.Bureaucratic Delays: The process of updating and maintaining ROC data may be subject to bureaucratic delays, affecting the timely availability of information.

10.Compliance Costs: Companies often incur compliance costs in terms of time and resources to fulfill regulatory requirements for ROC filings. These costs may be burdensome, particularly for small businesses.

11.Limited Coverage of Business Types: ROC data primarily covers registered companies, which may not account for various types of business entities, such as sole proprietorships or partnerships. This can limit the scope of analysis.

12.Access Restrictions: In some jurisdictions, access to ROC data may be restricted or subject to fees, making it less accessible to the general public and researchers.

13.Data Security Concerns: Storing and managing large volumes of ROC data can pose data security risks, especially when it includes sensitive information about companies and individuals.

14.Inconsistencies in Data Format: Data format and terminology may vary between jurisdictions, making it challenging to conduct cross-border comparisons and analysis.

15.Over Reliance on Data: Over Reliance on ROC data without corroborating it with other sources of information can lead to incomplete or inaccurate assessments of companies.

BENEFITS:

1.Legal Compliance: ROC ensures that companies adhere to legal requirements. Analyzing ROC data allows you to verify whether a company is in compliance with statutory regulations, such as filing financial statements and other necessary documents.

2.Transparency and Accountability: ROC data promotes transparency in corporate operations. It allows stakeholders, including investors, creditors, and the general public, to access official and credible information about a company's structure, management, and financial health, enhancing corporate accountability.

3.Investor Confidence: Investors can use ROC data to conduct due diligence and assess the credibility and legitimacy of a company. This, in turn, can increase investor confidence in the market.

4.Credit Risk Assessment: Financial institutions and creditors use ROC data to evaluate the creditworthiness of a company when considering lending. It helps them assess the financial stability and credibility of potential borrowers.

5.Market Research: Researchers and analysts can use ROC data to study market trends, identify industry players, and gain insights into market dynamics. This information is valuable for making informed decisions and forecasts.

6.Mergers and Acquisitions (M&A) Due Diligence: In M&A transactions, ROC data is crucial for due diligence. It provides a comprehensive understanding of a target company's financials, structure, and regulatory compliance.

7.Legal Proceedings: Legal professionals can use ROC data in litigation and legal disputes. It serves as evidence and supporting documentation for legal claims, ensuring that the legal process is fair and transparent.

8.Government Oversight: Government bodies and regulators use ROC data to monitor and enforce corporate governance standards, taxation compliance, and regulatory adherence, ensuring the integrity of the business environment.

9.**Public Awareness and Accountability**:ROC data is a tool for increasing public awareness of corporate activities. It holds companies accountable for their actions and provides a mechanism for public engagement and scrutiny.

10.Policy Formulation: Policymakers can use ROC data to inform the creation and modification of policies related to corporate governance, taxation, and economic development.

11.Investment Decisions: ROC data assists investors in making informed investment decisions. It provides critical information about a company's history, financials, and leadership that is essential for assessing investment opportunities.

12.Business Partnerships and Alliances: Companies considering forming partnerships, joint ventures, or alliances can use ROC data to evaluate potential partners' credibility and financial stability.

13.Due Diligence for Suppliers and Customers: Businesses can use ROC data to assess the financial stability and legitimacy of suppliers and customers, minimizing the risk of working with unreliable entities.

CONCLUSION:

In conclusion, Registrar of Companies (ROC) company analysis is a crucial practice with a multitude of benefits and some limitations. This process offers transparency, verifies legal compliance, and fosters informed decision-making. It aids in credit risk assessment, supports market research, streamlines M&A due diligence, and enhances government oversight and regulation. Additionally, ROC data promotes public awareness, assists in policy formulation, and contributes to economic development and anti-corruption efforts.

However, there are potential drawbacks to consider. ROC data may suffer from inaccuracies and delays, limited financial information, and privacy concerns. The complexity of data and lack of

context can make it challenging to draw meaningful insights. Regulatory gaps, potential for data manipulation, and bureaucratic delays may hamper the effectiveness of ROC analysis.

To harness the advantages of ROC company analysis while mitigating its limitations, it is essential to approach the data with caution, corroborate findings with other sources, and acknowledge the inherent constraints. ROC data remains a valuable tool for investors, researchers, regulatory bodies, and the public in their pursuit of understanding and engaging with the corporate landscape, advancing transparency and accountability, and facilitating well-informed decision-making in the realm of business and finance.

PREPARED BY,

Renuka H