Phase 3 submission document

AI-Driven Exploration and Prediction of Company Registration Trends with Registrar of Companies(ROC)

510521205032-Rasmi S
Bharathidasan Engineering College
Project title-ROC Company Analysis
Phase 3-Development Part 1

Development part 1:Start building the AI-driven exploration and prediction project by loading and preprocessing the dataset.

INTRODUCTION:

★ In the ever-evolving landscape of commerce and industry, staying ahead of registration trends and understanding the dynamics of new business formations is crucial.

The "AI-Driven Exploration and Prediction of Company Registration Trends with Registrar of Companies (RoC)" project seeks to address this need by leveraging the power of artificial intelligence and data analysis to provide valuable insights into the world of company registrations and regulatory compliance.

- Comprehensive Exploration: The primary aim of this project is to comprehensively explore the historical and emerging trends in company registrations with the RoC. By scrutinizing data patterns, regional variations, and industry-specific behaviors, we intend to unearth insights that offer value to stakeholders across diverse sectors.
- Predictive Modeling: This project's core ambition is to employ advanced machine learning and AI techniques to develop precise predictive models for company registration trends. By understanding the historical underpinnings, we aim to forecast the future, offering businesses, investors, and policymakers an informed crystal ball.
- Regulatory Compliance Assessment: A vital aspect of this project is the assessment of regulatory compliance among registered companies. By analyzing financial reports, legal compliance, and corporate governance practices, we aspire to promote transparency and responsible corporate conduct.
- Risk Evaluation: This project includes a thorough examination of the financial health and risk profiles of registered companies. Evaluating solvency, debt exposure, and profitability, our goal is to provide valuable insights empowering stakeholders to mitigate risks effectively.
- Industry Benchmarking: A key element of this project is benchmarking companies within their respective industries. This comparative analysis offers a means to evaluate performance in relation to peers and competitors.
- Policy Insights: The insights derived from this project have the potential to shape policy development. By contributing to the formulation of effective regulations and fostering a business-friendly environment, we aim to stimulate economic growth.

★ This project relies on cutting-edge data analytics, artificial intelligence, and machine learning techniques to analyze extensive datasets sourced from the RoC and other relevant sources. Data is meticulously cleaned, preprocessed, and analyzed to extract actionable insights, with a central focus on predictive modeling for future trends.The success of the "AI-Driven Exploration and Prediction of Company Registration Trends with RoC" project is rooted in collaboration. Regulatory authorities,

industry associations, and businesses are pivotal partners on this journey. By working together, we can continually refine our goals and objectives in alignment with the evolving corporate landscape.

- ★ This project is a dynamic and evolving initiative. In the future, we intend to explore advanced AI algorithms, including deep learning and ensemble methods, to enhance predictive accuracy. Real-time data integration and user-friendly interfaces are on the horizon, enabling up-to-the-minute insights and accessibility for a broader audience.

As we embark on this journey of exploration and analysis, the "AI-Driven Exploration and Prediction of Company Registration Trends with RoC" project holds the promise of delivering valuable insights and opportunities for growth and innovation across the corporate world. It signifies a commitment to data-driven decision-making, regulatory compliance, and responsible corporate governance.

GIVEN DATA SET:

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

Data_Gov_Tamil_Nadu...

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	CORPORATE	COMPANY_N	COMPANY_S	COMPANY_C	COMPANY_C	COMPANY_S	DATE_OF_RE	REGISTERED	AUTHORIZED	PAIDUP_CAP	INDUSTRIAL_	PRINCIPAL_BI	REGISTERED	REGISTRAR_	EMAIL_ADDR	LATEST_YEA	LATEST_YEAR_
2	F00643	HOCHTIEFF A	NNAEF	NA	NA	NA	1/12/1961	Tamil Nadu	0	0	NA	Agriculture &	AMBLE SIDE	R O C CHEN	NA	NA	NA
3	F00721	SUMITOMO C	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	FLAT NO. 6 1	SR O C CHEN	shuchi.chug@	NA	NA
4	F00892	SRILANKAN A	IACTV	NA	NA	NA	1/3/1982	Tamil Nadu	0	0	NA	Agriculture &	SRILANKAN AI	R O C CHEN	shreel0us@y	NA	NA
5	F01208	CALTEX INDIA	NAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	GOLD CREST	R O C CHEN	NA	NA	NA
6	F01218	GE HEALTHC	IACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	FF- 3 Palani	R O C CHEN	karthick9999	NA	NA
7	F01265	CAIRN ENERG	NAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	WELLINGTON	R O C CHEN	neerja.sharr	NA	NA
8	F01269	TORELLI S.R.L	ACTV	NA	NA	NA	5/9/1995	Tamil Nadu	0	0	NA	Agriculture &	6, Mangayar	R O C CHEN	chennai@tor	NA	NA
9	F01311	HARDY EXPL	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	5TH FLOOR,	R O C CHEN	venkatesh.v@	NA	NA
10	F01314	HOCHTIEF AK	ACTV	NA	NA	NA	11/4/1996	Tamil Nadu	0	0	NA	Agriculture &	NEW NO.86,	R O C CHEN	kumar@inter	NA	NA
11	F01412	EPSON SINGA	ACTV	NA	NA	NA	25-04-1997	Tamil Nadu	0	0	NA	Agriculture &	7C CEATUR	R O C CHEN	NA	NA	NA
12	F01426	CARGILL(UK	ACTV	NA	NA	NA	11/6/1997	Tamil Nadu	0	0	NA	Agriculture &	OFFICE NO 9	R O C CHEN	NA	NA	NA
13	F01468	CHO HEUNG E	NAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	29, MANPUR	R O C CHEN	chowlacoun	NA	NA
14	F01543	NYCOMED AS	ACTV	NA	NA	NA	27-10-1998	Tamil Nadu	0	0	NA	Agriculture &	A D -46 1ST S	R O C CHEN	NA	NA	NA
15	F01544	CHERRINGTO	ACTV	NA	NA	NA	1/5/2000	Tamil Nadu	0	0	NA	Agriculture &	OHADDOWS	R O C CHEN	NA	NA	NA
16	F01563	SHIMADZU AS	NAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	FIRST FLOOR	R O C CHEN	kousik@vsnl	NA	NA
17	F01565	CORK INTER	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	ARAY APEX	R O C CHEN	NA	NA	NA
18	F01566	ERBS ENGG	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	39 2nd Main	R O C CHEN	NA	NA	NA
19	F01589	RALF SCHNEI	NAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	FLAT C, SAI V	R O C CHEN	NA	NA	NA
20	F01593	MITRAYA TA	CTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	OLD NO 148	NR O C CHEN	NA	NA	NA
21	F01618	HEAT AND C	ACTV	NA	NA	NA	13-07-1999	Tamil Nadu	0	0	NA	Agriculture &	A40 OLD NO	R O C CHEN	ncrajagopal	NA	NA
22	F01628	DREX SYSTE	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	F-1, FIRST FL	R O C CHEN	drex@vsnl.n	NA	NA
23	F01641	NMB-MINEBE	NAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	Level - 2 Re	R O C CHEN	stsogawa@n	NA	NA
24	F01643	ARROW INTE	ACTV	NA	NA	NA	2/11/1999	Tamil Nadu	0	0	NA	Agriculture &	BLUE HAVEN	R O C CHEN	NA	NA	NA
25	F01694	GAMBRO CHE	ACTV	NA	NA	NA	14-06-2000	Tamil Nadu	0	0	NA	Agriculture &	5 1ST FLOOR	R O C CHEN	NA	NA	NA
26	F01703	OBARA CORP	NAEF	NA	NA	NA	17-07-2000	Tamil Nadu	0	0	NA	Agriculture &	INDIA BRANCH	R O C CHEN	joe@obara.c	NA	NA
27	F01752	GUPTA WAW	ACTV	NA	NA	NA	24-01-2001	Tamil Nadu	0	0	NA	Agriculture &	4 AVVAI SHA	R O C CHEN	NA	NA	NA
28	F01753	AUCHAN INTE	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	RK Tower, N	R O C CHEN	pverma@vk	NA	NA
29	F01767	TOSHIBA PLA	NAEF	NA	NA	NA	8/3/2001	Tamil Nadu	0	0	NA	Agriculture &	HOTEL AMBA	R O C CHEN	NA	NA	NA
30	F01768	YAMAZEN CO	NAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	PLOT 69, SIVA	R O C CHEN	NA	NA	NA
31	F01770	OWL INTERN	ACTV	NA	NA	NA		Tamil Nadu	0	0	NA	Agriculture &	NO 1 SAPTHA	R O C CHEN	NA	NA	NA
32	F01826	LEXMARK IN	ACTV	NA	NA	NA	16-08-2001	Tamil Nadu	0	0	NA	Agriculture &	APEEJAY BUS	R O C CHEN	NA	NA	NA
33	F01830	FLUID ENER	GACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	FLUID ENER	GR O C CHEN	jeeva@fleci	NA	NA
34	F01881	WATCH GUA	RACTV	NA	NA	NA	21-11-2001	Tamil Nadu	0	0	NA	Agriculture &	54/2, padwe	R O C CHEN	chennaiadmi	NA	NA
35	F01878	SINAR JERUH	ACTV	NA	NA	NA	24-12-2001	Tamil Nadu	0	0	NA	Agriculture &	57/4 SEVEN	R O C CHEN	accountsho@	NA	NA
36	F01968	SIPLEC INTER	ACTV	NA	NA	NA	23-09-1995	Tamil Nadu	0	0	NA	Agriculture &	II FLOOR, PA	RR O C CHEN	svrajacdm@	NA	NA
37	F01935	INTELSAT GL	OACTV	NA	NA	NA	20-05-2005	Tamil Nadu	0	0	NA	Agriculture &	TPL HOUSE,	2R O C CHEN	NA	NA	NA
38	F01940	PGS GEOPHY	SACTV	NA	NA	NA	27-05-2002	Tamil Nadu	0	0	NA	Agriculture &	R.00M3054	&R O C CHEN	NA	NA	NA
39	F01987	SEVERN GLO	ACTV	NA	NA	NA	29-08-2002	Tamil Nadu	0	0	NA	Agriculture &	8B SRV AVEN	R O C CHEN	NA	NA	NA
40	F02028	LAGERWEY W	ACTV	NA	NA	NA	24-10-2002	Tamil Nadu	0	0	NA	Agriculture &	SUJATHA CEN	R O C CHEN	NA	NA	NA
41	F02001	SOCAM MAN	NAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	0 No.II N No.9	R O C CHEN	socamkr@vs	NA	NA
42	F02098	VAN DE NUL	NACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	ENMORE CO	MR O C CHEN	NA	NA	NA
43	F02104	BUCKMAN LA	WACTV	NA	NA	NA	5/2/2003	Tamil Nadu	0	0	NA	Agriculture &	50 ANNA SAI	R O C CHEN	vassociates2	NA	NA
44	F02110	ZWICK ASIA P	ACTV	NA	NA	NA	13-02-2002	Tamil Nadu	0	0	NA	Agriculture &	3D SAI KRAN	R O C CHEN	NA	NA	NA
45	F02122	INVE THALA	NNAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	PLOT NO 34	ER O C CHEN	chandra68@	NA	NA
46	F02126	SUNLEY FASH	ACTV	NA	NA	NA	12/3/2003	Tamil Nadu	0	0	NA	Agriculture &	100B, TV RO	AR O C CHEN	NA	NA	NA
47	F02143	ROTHE ERDE	NAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	6E GEE GEE	ER O C CHEN	NA	NA	NA
48	F02157	RANGASWAM	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	SITE NO.25/2	R O C CHEN	apsanandh@	NA	NA
49	F02189	EASTMAN FIL	ACTV	NA	NA	NA	18-08-2003	Tamil Nadu	0	0	NA	Agriculture &	28 ARUNACH	AR O C CHEN	admin_hyd@	NA	NA
50	F02222	XAMBALA IN	NNAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	No 63, Valluv	aR O C CHEN	design@xa	NA	NA
51	F02235	DANTEE LIMI	ACTV	NA	NA	NA	5/11/2003	Tamil Nadu	0	0	NA	Agriculture &	NO 21 GROUN	DR O C CHEN	NA	NA	NA
52	F02253	COLUMBIA SP	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	MF- 7 CIPET	IR O C CHEN	VSree@colu	NA	NA
53	F02261	KISTLER INST	NAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	2B CENTUR	YR O C CHEN	manojmalik@	NA	NA
54	F02262	AJINOMOTO	CNAEF	NA	NA	NA	21-01-2004	Tamil Nadu	0	0	NA	Agriculture &	23/1, POONA	MR O C CHEN	info@ajinom	NA	NA
55	F02297	DANKOTUWA	ACTV	NA	NA	NA	15-04-2004	Tamil Nadu	0	0	NA	Agriculture &	Old No 15 &	16R O C CHEN	maheerama	NA	NA
56	F02337	PUNCAK NAG	ACTV	NA	NA	NA	26-07-2004	Tamil Nadu	0	0	NA	Agriculture &	5TH FLOOR,	TR O C CHEN	NA	NA	NA
57	F02339	SIGMA CORP	NAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	NO5, EMERA	LR O C CHEN	jth@sigma.c	NA	NA
58	F02372	CARGO COMM	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	S NO 105 DES	LR O C CHEN	gerard_fm@	NA	NA
59	F02378	HETTIGODA D	IACTV	NA	NA	NA	17-09-2004	Tamil Nadu	0	0	NA	Agriculture &	STALLNO.4	SR O C CHEN	NA	NA	NA
60	F02394	PROPOLIS SY	SACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	PLOT 45, KAL	R O C CHEN	NA	NA	NA
61	F02448	DEUTSCHE W	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	54 K P N COL	RO C CHEN	NA	NA	NA
62	F02443	NORDROTEX	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	31 VENKATNA	R O C CHEN	NA	NA	NA
63	F02446	PANASIA RES	NAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	C/ 0 PADMAV	AR O C CHEN	NA	NA	NA
64	F02466	SAIPEM PORT	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	NO 4, FOOR T	HR O C CHEN	kpramaswar	NA	NA
65	F02478	KDDS INFOTE	ACTV	NA	NA	NA	31-03-2005	Tamil Nadu	0	0	NA	Agriculture &	IITH FLOOR,	ER O C CHEN	NA	NA	NA
66	F02492	SEBCORP ENA	CTV	NA	NA	NA	27-04-2005	Tamil Nadu	0	0	NA	Agriculture &	8 TH FLOOR,	JR O C CHEN	NA	NA	NA
67	F02507	ETS MONTEL	ACTV	NA	NA	NA	20-05-2005	Tamil Nadu	0	0	NA	Agriculture &	Old no.8, New	R O C CHEN	sriranjani.ra	NA	NA
68	F02522	ITAC OK LTD	NAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	ANMOL PALA	ZR O C CHEN	NA	NA	NA
69	F02537	EOS GMBH EL	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	Ground Floor,	R O C CHEN	pcchnl@gm	NA	NA
70	F02594	CHANDRA IM	ACTV	NA	NA	NA	8/9/2005	Tamil Nadu	0	0	NA	Agriculture &	NO 46 RK MI	ZR O C CHEN	NA	NA	NA
71	F02608	SCG TRADING	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	FLAT NO 4-B	R O C CHEN	yongyosp@s	NA	NA
72	F02630	GLOBAL BRA	NACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	SUITE NO 40	8R O C CHEN	NA	NA	NA
73	F02673	EXPONENTIA	NAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	NEW NO - 28	R O C CHEN	sri@percipi	NA	NA
74	F02694	ATHEROS IN	DNAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	4, TVH AGNI	R O C CHEN	pjagadee@a	NA	NA
75	F02695	DEICHMANN	EACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &	PRINCE TOW	ER O C CHEN	nizam_syed@	NA	NA
76	F02734	NSK LIMITED	NAEF	NA	NA	NA	29-05-2006	Tamil Nadu	0	0	NA	Agriculture &	Old No.5 / N	eR O C CHEN	jojo@nsk.co	NA	NA

15000 Rows &12 columns

NECESSARY STEP TO FOLLOW:

1.Import Libraries:

Start by importing the necessary libraries:

Program:

```
from sklearn import datasets from sklearn.tree import
DecisionTreeClassifier from sklearn.linear_model import
LogisticRegression from sklearn.metrics import roc_curve,
roc_auc_score from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
```

2.setup the dataset:

Here we have used datasets to load the inbuilt wine dataset and we have created objects X and y to store the data and the target value respectively.

```
dataset = datasets.load_wine()

 X = dataset.data

 y = dataset.target
```

3.Spliting the data and Training the model:

The module train_test_split is used to split the data into two parts, one is train which is used to train the model and the other is test which is used to check how our model is working on unseen

data. Here we are passing 0.3 as a parameter in the train_test_split which will split the data such that 30% of data will be in test part and rest 70% will be in the train part.

program:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)
```

Now we are creating objects for classifier and training the classifier with the train split of the dataset i.e x_train and y_train.);

```
clf_tree = DecisionTreeClassifier();

 clf_reg = LogisticRegression();

 clf_tree.fit(X_train, y_train);

 clf_reg.fit(X_train, y_train);
```

4.Using the models on test dataset

After traing the classifier on test dataset, we are using the model to predict the target values for test dataset. We are storing the predicted class by both of the models and we will use it to get the ROC AUC score

program:

```
y_score1 = clf_tree.predict_proba(X_test)[:,1]

 y_score2 = clf_reg.predict_proba(X_test)[:,1]
```

5.Creating False and True Positive Rates and printing Scores

We have to get False Positive Rates and True Postive rates for the Classifiers because these will be used to plot the ROC Curve. This can be done by roc_curve module by passing the test dataset and the predicted data through it. Here we are doing this for both the classifier.

program:

```
false_positive_rate1, true_positive_rate1, threshold1 = roc_curve(y_test,
y_score1) false_positive_rate2, true_positive_rate2, threshold2 =
roc_curve(y_test, y_score2)
```

Now, For getting ROC_AUC score we can simply pass the test data and the predected data into the function ruc_auc_score. We are printing it with print statements for better understanding.

```
print('roc_auc_score for DecisionTree: ', roc_auc_score(y_test, y_score1))
print('roc_auc_score for Logistic Regression: ', roc_auc_score(y_test,

y_score2))
```

6.Ploting ROC Curves

We are ploting two ROC Curve as subplots one for DecisionTreeClassifier and another for LogisticRegression. Both have their respective False Positive Rate on X-axis and True Positive Rate on Y-axis.

```
plt.subplots(1, figsize=(10,10)) plt.title('Receiver Operating Characteristic
- DecisionTree') plt.plot(false_positive_rate1, true_positive_rate1)
```

```
plt.plot([0, 1], ls="--") plt.plot([0, 0], [1, 0] , c=".7"), plt.plot([1, 1] , c=".7") plt.ylabel('True Positive Rate') plt.xlabel('False Positive Rate') plt.show() plt.subplots(1, figsize=(10,10)) plt.title('Receiver Operating Characteristic - Logistic regression') plt.plot(false_positive_rate2, true_positive_rate2) plt.plot([0, 1], ls="--") plt.plot([0, 0], [1, 0] , c=".7"), plt.plot([1, 1] , c=".7") plt.ylabel('True Positive Rate') plt.xlabel('False Positive Rate') plt.show()
```

As an output we get:

roc_auc_score for DecisionTree: 0.9539141414141414

roc_auc_score for Logistic Regression: 0.9875140291806959

Importance of loading and processing 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 protected 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 for validation and peer review.

Challenges involoved in loading and preprocessing a ROC company analysis dataset

- Data Volume: RoC datasets can be extensive, containing records of thousands or millions of companies. Handling large volumes of data may require specialized tools, infrastructure, and optimized algorithms.
- Data Quality: Ensuring data quality is paramount. Incomplete, inconsistent, or inaccurate data can lead to biased analysis and inaccurate predictions. Dealing with missing values, outliers, and errors is a significant challenge.
- Data Integration: If data is collected from multiple sources or in various formats, integrating it into a coherent dataset can be challenging. Data from different regions or branches may have variations in naming conventions or data formats.

- Data Security: RoC data often contains sensitive information about companies. Ensuring data security and complying with data protection regulations is essential.
- Data Imbalance: Imbalanced datasets, where certain outcomes (e.g., bankruptcies) are rare compared to others, can lead to model bias. Addressing this imbalance can be complex.
- Categorical Data: RoC data may include categorical variables, such as industry codes or company types. Encoding and handling these variables can be challenging, as different approaches may be required based on the analysis.
- Feature Engineering: Determining which features or variables are relevant for analysis can be intricate. Extracting meaningful information from textual data, such as company descriptions or reports, may require natural language processing (NLP) techniques.
- Time-Series Data: If the dataset includes historical registration data, time-series analysis is needed. Handling time-dependent features, seasonality, and trends can be complex.
- Dimensionality Reduction: Large datasets may lead to a high number of features. Dimensionality reduction techniques may be necessary to improve model efficiency and prevent overfitting.
- Regulatory Changes: RoC regulations and reporting standards may change over time. Ensuring that the dataset is up to date and consistent with current regulations can be challenging.
- Computational Resources: Processing large datasets may require substantial computational resources, including memory and processing power.
- Data Exploration Tools: Visualizing and exploring the data effectively is essential. Selecting the right tools and techniques for data exploration can be challenging.

- Documentation and Reproducibility: Documenting data preprocessing steps is crucial for reproducibility and transparency. Maintaining clear records of these steps can be challenging

How to overcome those challenges involved in roc company analysis dataset

Overcoming the challenges involved in working with a RoC (Registrar of Companies) dataset for company analysis requires a combination of effective strategies and best practices. Here's how to address these challenges:

1. Data Quality Assurance:

Data Cleaning: Implement data cleaning techniques to handle missing values, outliers, and inconsistencies. This may involve imputation, data validation, and outlier detection algorithms.

Data Validation: Cross-verify data with external sources or through expert review to validate the accuracy and completeness of the dataset.

2. Data Integration:

Standardize Data: Ensure that data from various sources or branches is standardized to a common format and naming convention.

ETL Processes: Implement Extract, Transform, Load (ETL) processes to integrate and consolidate data efficiently.

3. Data Security:

Data Encryption:Encrypt sensitive data to protect it from unauthorized access.

Access Control: Implement access control and authentication mechanisms to restrict access to authorized personnel only.

4. Data Imbalance:

Resampling Techniques: Apply resampling techniques like oversampling or undersampling to address class imbalance issues.

Advanced Algorithms: Consider using algorithms designed for imbalanced datasets, such as SMOTE (Synthetic Minority Over-sampling Technique).

5. Categorical Data:

-One-Hot Encoding: Convert categorical variables into a numerical format using one-hot encoding or other appropriate encoding methods.

Feature Selection:Use feature selection techniques to identify the most relevant variables for analysis.

6. Feature Engineering:

NLP Techniques:Utilize natural language processing (NLP) techniques to extract meaningful information from textual data.

Domain Expertise:Collaborate with domain experts to identify relevant features and create new features that capture valuable insights.

7. Time-Series Data:

Time-Series Analysis:Apply time-series analysis techniques to handle temporal data, including forecasting and trend detection.

Lag Variables:Create lag variables to account for time dependencies.

8. Dimensionality Reduction:

PCA (Principal Component Analysis): Use PCA or other dimensionality reduction techniques to reduce the number of features while preserving information.

Feature Importance: Assess feature importance through feature ranking methods to prioritize the most relevant features.

9. Regulatory Changes:

Continuous Monitoring:Stay informed about regulatory changes and ensure that the dataset is updated accordingly.

Regular Auditing: Conduct regular audits to verify data compliance with current regulations.

10. Computational Resources:

Cloud Computing:Consider using cloud computing platforms that offer scalable resources for processing and analyzing large datasets.

Parallel Processing: Utilize parallel processing to distribute the computational load efficiently.

11. Data Exploration Tools:

-Data Visualization Libraries: Utilize data visualization libraries (e.g., Matplotlib, Seaborn) to create informative visualizations for data exploration.

Interactive Dashboards: Develop interactive dashboards for exploring the data, making it more accessible and user-friendly.

12. Documentation and Reproducibility:

Version Control:Use version control systems to track changes in data preprocessing and analysis scripts.

Comprehensive Documentation: Maintain clear and comprehensive documentation of data preprocessing steps, code, and analysis methodology.

By addressing these challenges through a combination of data preprocessing techniques, collaboration with experts, and the adoption of relevant technologies, you can enhance the quality and reliability of your RoC company analysis dataset, making it a valuable resource for insightful analysis and decision-making.

1.LOADING THE 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:

a.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.

b.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.

c.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.

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}"
```

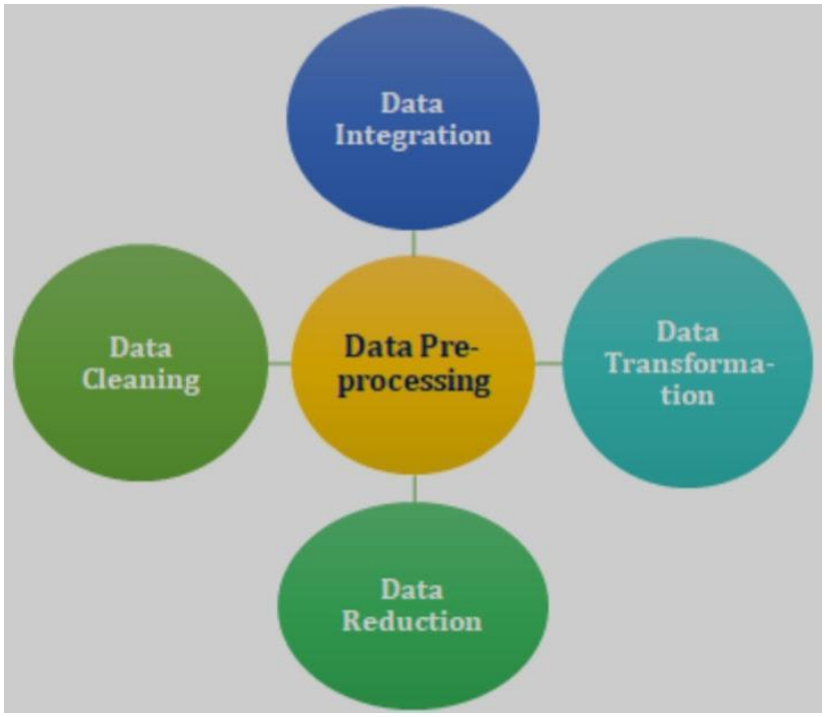
Loading Dataset:

```
dataset=pd.read_csv('D:data_Gov_Tamil_Nadu.csv')
```

2.PREPROCESSING THE DATASET:

- ➢ Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.
- ➢ When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task.
    - Data cleaning,
    - Data transformation,
    - Data reduction, and
    - Data integration

Those are the major steps in data preprocessing.



DATA INTEGRATION:

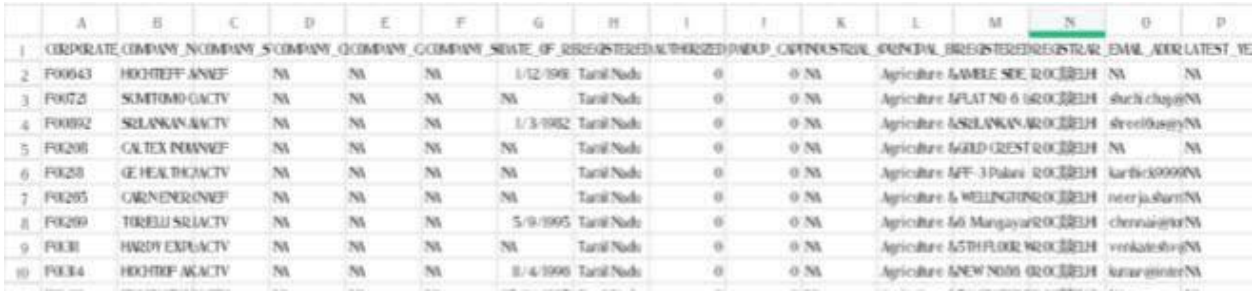
```
dataset1 = "https://tn.data.gov.in/resource/company-master-data-tamil-nadu-upto-28th-february-2019"
dataset2 = "https://tn.data.gov.in/resource/company-master-data-tamil-nadu-upto-28th-february-2019"
df1 = pd.read_csv(dataset1, header = 0)
df2 = pd.read_csv(dataset2, header = 0)
```

```
df1.head()
```

```
df2.head()
```

```
df = pd.merge(df1, df2, on = 'company_id')
```

df.head(10)



DATA CLEANING:

Data cleaning is the process of identifying and correcting errors, inconsistencies, and inaccuracies in a dataset. It involves tasks like handling missing data, removing duplicates, addressing outliers, standardizing data, and transforming it to ensure that the dataset is accurate, reliable, and ready for analysis or modeling.

DATA TRANSFORMATION:

Data transformation is the process of converting and modifying data to make it suitable for analysis, modeling, or other specific purposes. It may involve tasks like scaling, encoding categorical variables, aggregating data, or creating new features to extract meaningful information from the original dataset. Data transformation is a crucial step in data preprocessing to enhance the quality and utility of the data for various applications.

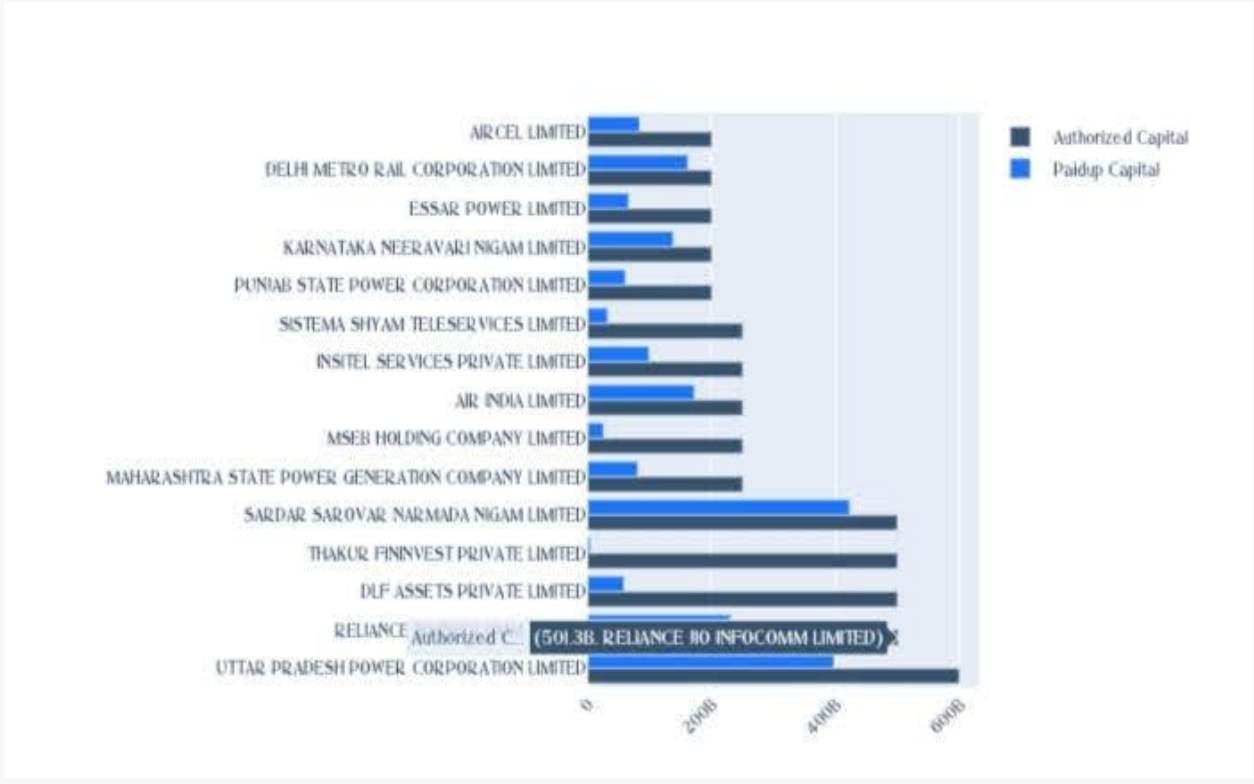
DATA REDUCTION:

Data reduction is the process of reducing the volume but producing the same or similar analytical results. It involves techniques like aggregation, sampling, and dimensionality reduction to make large

datasets more manageable while preserving essential information for analysis or modeling. Data reduction can improve the efficiency of data processing and reduce computational requirements.

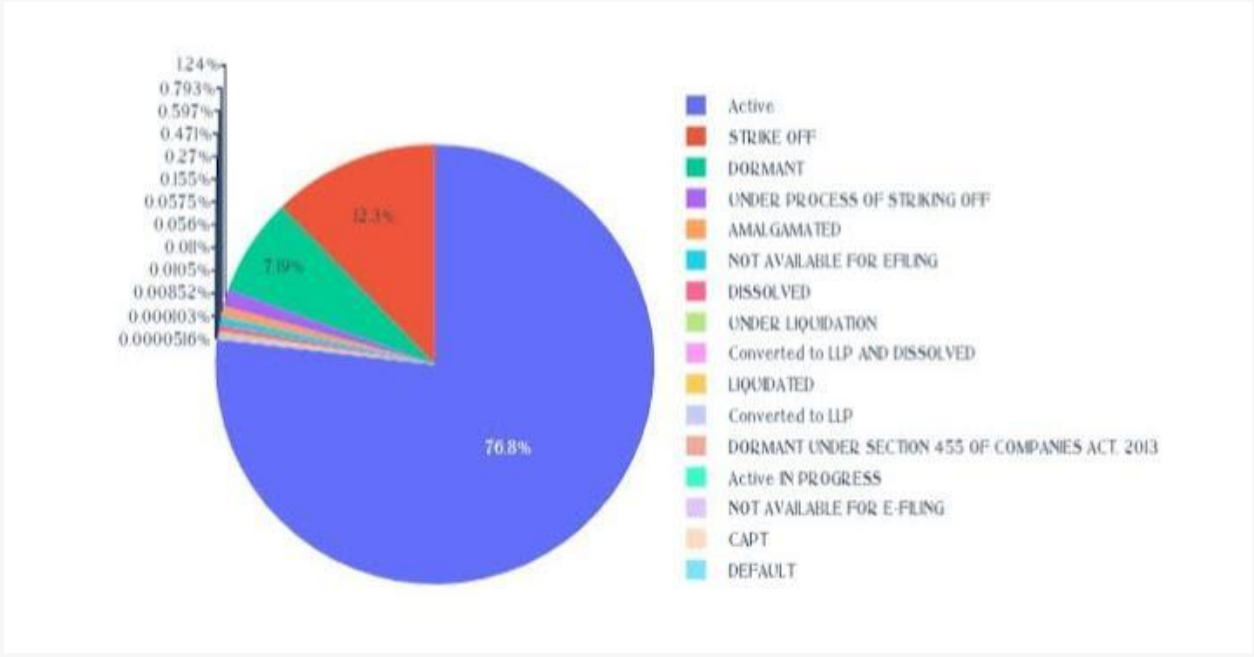
PROGRAM:

```
auth_capital_list = m.sort_values(by='AUTHORIZED_CAPITAL',
ascending=False)[0:20]
fig = go.Figure()
fig.add_trace(go.Bar(
    x=auth_capital_list['AUTHORIZED_CAPITAL'][0:15],
    y=auth_capital_list['COMPANY_NAME'][0:15] ,
    name='Authorized Capital',
    marker_color='rgb(55, 83, 109)',
    orientation='h'
))
fig.add_trace(go.Bar(
    x=auth_capital_list['PAIDUP_CAPITAL'][0:15] ,
    y=auth_capital_list['COMPANY_NAME'][0:15] ,
    name='Paidup Capital',
    marker_color='rgb(26, 118, 255)',
    orientation='h'
))
# Here we modify the tickangle of the xaxis, resulting in rotated labels.
fig.update_layout(barmode='group', xaxis_tickangle=-45)
fig.show()
```



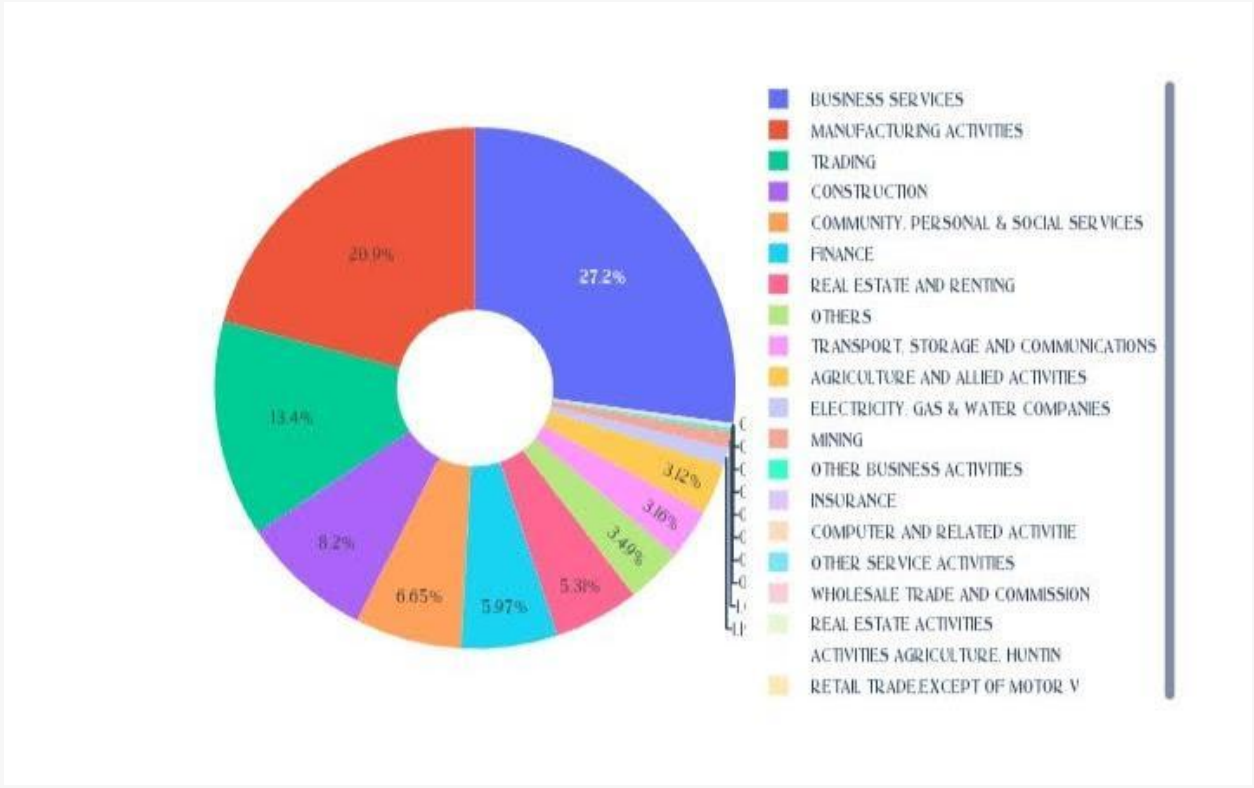

```
correction = {"ACTIVE": "Active",

          "ACTV": "Active",

          "Available": "available",

          "e-filing": "E-filing",

        "efiling": "E-filing",

          "CONVERTED TO LLP" : 'Converted to LLP'    }

m['COMPANY_STATUS'].replace(correction,regex = True,inplace = True)

fig3 = px.pie(m, values='COUNT', names='COMPANY_STATUS')
```

```
fig3.show()
```




```
n=m.groupby(['PRINCIPAL_BUSINESS_ACTIVITY'],as_index=False)['COUNT'].cout(
).sort_values(by='COUNT', ascending=False)[0:20]

fig4 = px.pie(n, values='COUNT',
names='PRINCIPAL_BUSINESS_ACTIVITY',hole=.3)

#fig4.update_layout(margin = dict(t=0, l=0, r=0, b=0))
```

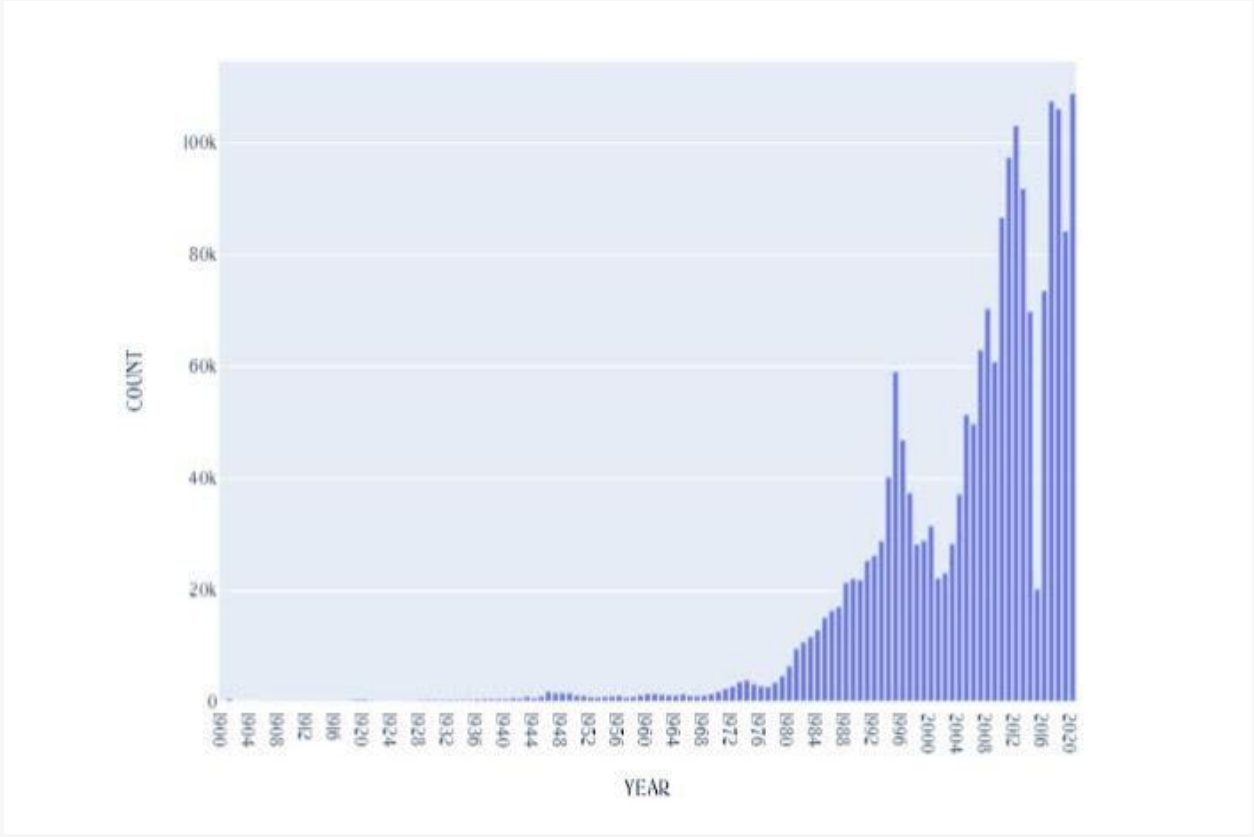
```
fig4.show()
```




```
df2 = m.groupby(['YEAR'],as_index=False)['COUNT'].count()
```

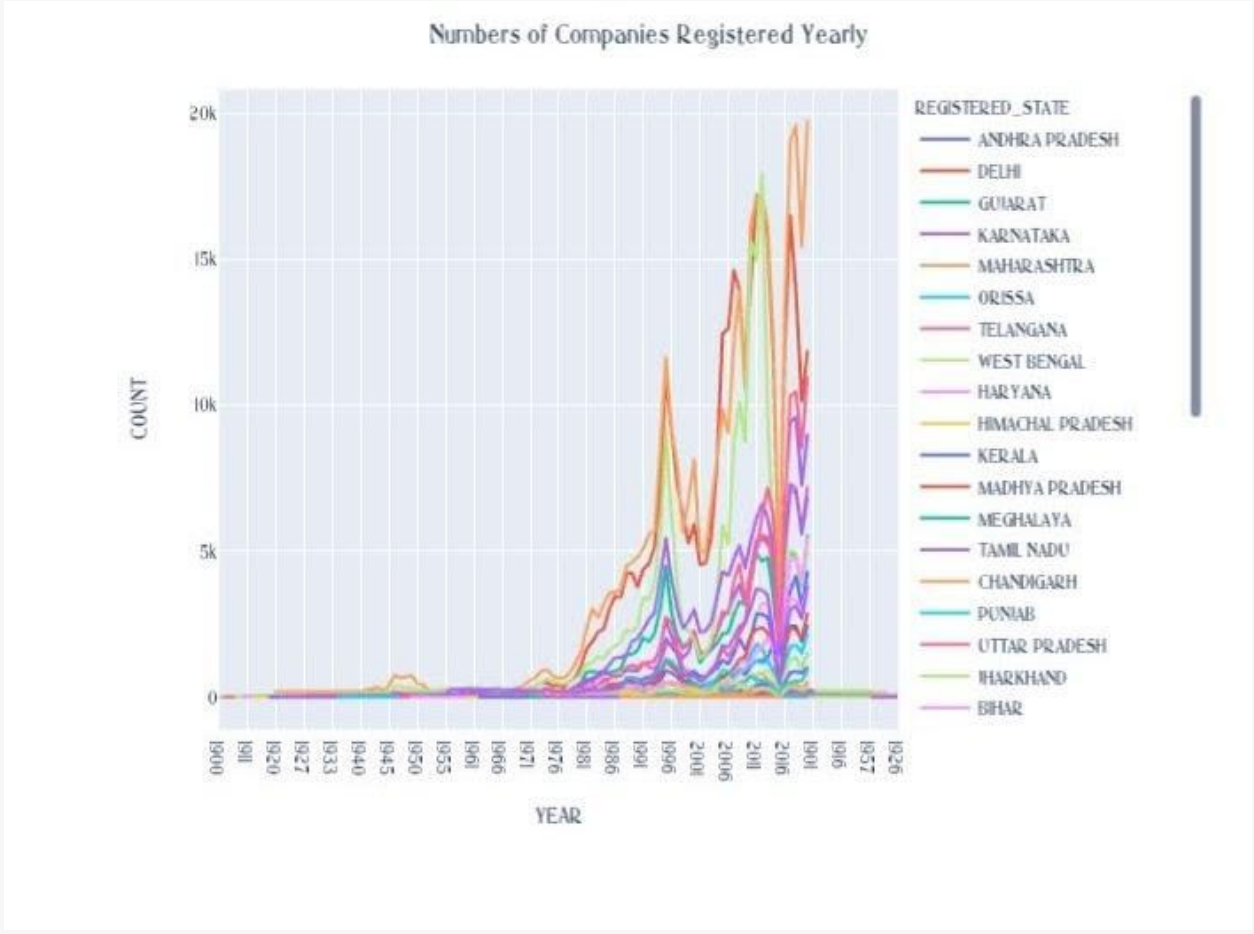
```
fig = px.bar(df2, x="YEAR", y="COUNT")
```

```
fig.show()
```




```
df =m.groupby(['YEAR','REGISTERED_STATE'],as_index=False)['COUNT'].count()
fig6 = px.line(df, x="YEAR", y="COUNT",color='REGISTERED_STATE')
fig6.update_layout( title={
'text': 'Numbers of Companies Registered Yearly',
        'y':0.96,
        'x':0.5,
        'xanchor': 'center',
        'yanchor': 'top'})
```

```
fig6.show()
```



CONCLUSION:

In summary, the analysis of the Registrar of Companies (ROC) reveals its pivotal role in regulating and maintaining corporate records. It ensures legal compliance, fosters transparency, and provides valuable data for various stakeholders. As businesses increasingly embrace digital processes, the ROC's role in streamlining and securing registration procedures is becoming ever more important for a modern and efficient business environment.