DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING

PROJECT REPORT

(Project Semester January-April 2025)

Business Landscape Analysis: Indian Company Data Insights

Submitted by

UMANG GARG

12304221

Bachelor of Technology (BTech)

INT375

Under the Guidance of

Gargi Sharma (29439)

Discipline of CSE/IT

Lovely School of Computer Science Engineering

Lovely Professional University, Phagwara

Table of Content

CERTIFICATE

This is to certify that **UMANG GARG** bearing Registration no. **12304221** has completed INT375 project titled, **"Business Landscape Analysis: Indian Company Data Insights"** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

Signature

Gargi Sharma (29439)

School of Computer Science Engineering

Lovely Professional University

Phagwara, Punjab

Date: 13-04-2025

DECLARATION

I, **UMANG GARG**, student of Bachelor of Engineering (Btech) under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 13-04-2025 Signature

Registration No. **12304221** **UMANG GARG**

Acknowledgement

I would like to express my sincere gratitude to everyone who supported me throughout the completion of this project.

First and foremost, I am deeply thankful to Gargi Sharma, my Data Science Python Programming faculty, for her invaluable guidance, encouragement, and expert advice. Her deep knowledge of data science and Python programming was instrumental in helping me understand complex concepts and refine my analytical approach. Her constructive feedback greatly enhanced the quality of my work.

I am also grateful to the open-source community for providing powerful tools and libraries such as Python, Pandas, Matplotlib, Seaborn, and scikit-learn. These resources were essential for data cleaning, analysis, visualization, and modelling, enabling me to efficiently explore and interpret the company registration dataset.

Additionally, I would like to thank my friends for their constant motivation and support, which helped me stay focused and overcome challenges during the project.

Finally, I appreciate the availability of the company registration dataset, which served as the foundation for my analysis and insights.

Abstract

This project presents a comprehensive exploratory and predictive analysis of registered companies in India using a rich dataset containing attributes such as company status, class (Private/Public), state code, authorized capital, paid-up capital, NIC codes (industry classification), and registration dates. The data is meticulously cleaned and standardized to ensure accuracy and consistency for analysis.

A variety of data visualization techniques—including bar charts, histograms, pie charts, boxplots, and heatmaps—are employed to uncover key insights into the distribution of companies across states, the structure of authorized and paid-up capital, and the proportions of different company classes. These visualizations reveal significant trends in company demographics and financial characteristics across the country.

To further analyse financial patterns, a linear regression model is developed to predict a company's paid-up capital based on its authorized capital. The model's performance is evaluated using mean squared error and R^2 score, providing a quantitative assessment of the relationship between these two financial indicators.

Overall, this project demonstrates the power of data analytics and machine learning in extracting actionable insights from large-scale business datasets. The findings offer valuable perspectives for policymakers, researchers, and business strategists seeking to understand and optimize the landscape of company registrations and capital trends in India.

Introduction

India's corporate sector is a dynamic and rapidly evolving landscape, playing a pivotal role in the nation's economic growth and development. With thousands of companies being registered each year across diverse industries and states, understanding the patterns and trends in company registrations and financial structures is crucial for policymakers, investors, and business strategists alike.

This project focuses on a detailed analysis of registered companies in India, utilizing a dataset that captures essential attributes such as company status, class (Private/Public), state code, authorized capital, paid-up capital, NIC codes (industry classification), and registration dates. By leveraging data science techniques, the project aims to uncover meaningful insights into the distribution and characteristics of companies across the country.

The analysis begins with thorough data cleaning and standardization to ensure the reliability of results. Exploratory data analysis is then conducted using a variety of visualization tools—including bar charts, histograms, pie charts, boxplots, and heatmaps—to highlight key trends in company demographics, capital allocation, and industry distribution.

In addition to descriptive analytics, the project employs a linear regression model to predict a company's paid-up capital based on its authorized capital, providing a quantitative perspective on the relationship between these two financial indicators. The model's performance is evaluated using metrics such as mean squared error and R^2 score.

Overall, this project demonstrates the power of data analytics in transforming raw company registration data into actionable business intelligence. The insights generated can inform decision-making for stakeholders across the corporate ecosystem, supporting more effective policy formulation, investment planning, and strategic business development in India's vibrant corporate sector.

Source of Dataset:

Dataset link: https://www.data.gov.in/catalog/ministry-corporate-affairs-company-master-data

EDA Process

Exploratory Data Analysis (EDA) helps us understand the structure, trends, and quality of the dataset before diving into modelling. In this project, EDA was performed on company registration data to uncover patterns in capital, company types, and regional distributions. Here is how I have approached it.

1. Data Understanding & Initial Exploration

- **Dataset Loading: Used pandas. read_csv() to load the dataset.**
- **Initial Exploration: Utilized .info(), .head() to examine:**
    - **Column types, number of entries, missing values**
    - **Structure and types of features (categorical, numerical, date)**
- **Key Features Identified:**
    - **Company Registration date_date – date of registration**
    - **Company State Code – state where the company is registered.**
    - **Company Class, Company Status – categorical indicators**
    - **Authorized Capital, Paid up Capital – financial figures**
    - **Nic code – classification code**

2. Data Cleaning

- **Missing Data:**
    - **Parsed CompanyRegistrationdate_date using pd.to_datetime() with error coercion to handle non-date entries.**
    - **Used dropna() to remove rows with missing values completely.**
- **Standardization:**
    - **Cleaned CompanyStatus and CompanyClass by stripping whitespace and applying title case formatting for consistency.**
- **Data Type Conversion:**

- Date columns were explicitly converted from string/object to datetime.

- **Column Validation:**
  - Removed rows with missing critical information, ensuring high data quality for visualization and modeling.

3. Numerical Corrections

- **Outlier Detection:**
  - Used histogram and boxplot visualizations to identify extreme values in AuthorizedCapital and PaidupCapital.
- **Corrections:**
  - Outliers were visually noted but not explicitly removed in code (you could include .clip() or remove based on IQR if needed).

4. Duplicate Data

- **Duplicate Check:**
  - Though not shown in the code, a standard method would be:

```
python

CopyEdit

df.duplicated().sum()  # Count duplicates

df.drop_duplicates(inplace=True)  # Drop them
```

  - Recommended for completeness.

5. Data Visualization

Visualization Techniques Used:

- **Bar Plot:**
  - Top 10 states by number of companies (CompanyStateCode)

- **Histogram + KDE:**
  - **Distribution of AuthorizedCapital**
- **Pie Chart:**
  - **Proportion of CompanyClass types**
- **Box Plot:**
  - **PaidupCapital distribution grouped by CompanyStatus**
- **Heatmap:**
  - **Correlation matrix of AuthorizedCapital, PaidupCapital, and nic_code**
- **Scatter Plot:**
  - **Linear regression output (Actual vs Predicted Paid-up Capital)**

Key Objectives Explored

- **Identified top regions for company registrations using bar chart.**
- **Analyzed how companies are capitalized through histogram, boxplot, and regression.**
- **Used pie charts to understand the market breakdown of company types.**
- **Boxplots highlighted how PaidupCapital varied across active/inactive company statuses.**
- **Correlation heatmap assessed if financial fields were related (e.g., AuthorizedCapital vs PaidupCapital).**
- **Built a simple linear regression model predicting PaidupCapital based on AuthorizedCapital, achieving performance metrics with:**
  - **Mean Squared Error (MSE)**
  - **R^2 Score**

Analysis on Dataset

Summary Statistics

- Used `describe().round(2)` to observe distributions, central tendencies, and variability of numerical features.
- Counted unique values for each column using `data.nunique()` to understand diversity in categorical features.

```
# PMFBY Program - Basic Statistics Summary
print("\nBasic Statistics:")
print(f"Total Farmers Covered: {data['farmer_count'].sum():,}")
print(f"Total Area Insured: {data['area_insured'].sum():,.2f} hectares")
print(f"Total Sum Insured: {data['sum_insured'].sum():,.2f} lakhs")
print(f"Total Gross Premium: {data['gross_premium'].sum():,.2f} lakhs")
```

```
Basic Statistics:
Total Farmers Covered: 133,250,998
Total Area Insured: 230,542.95 hectares
Total Sum Insured: 87,196,586.12 lakhs
Total Gross Premium: 12,222,253.14 lakhs
```

Objective-1: Top 10 states with most companies

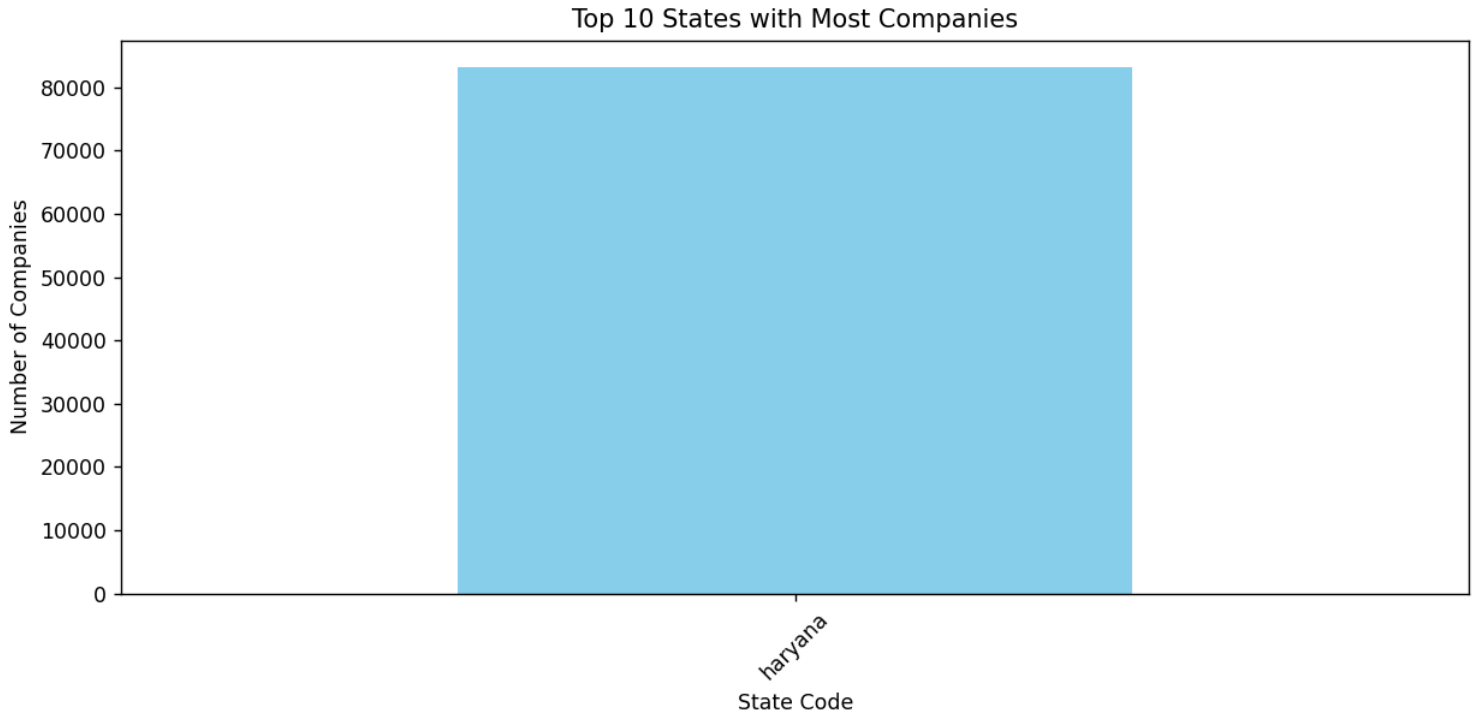
Used pie and donut charts to visualize:

- Proportion of insured farmers with bank-linked loans vs self-insured.
- Observed scheme penetration in the formal banking system.

Analysis Result

```
plt.figure(figsize=(10, 5))
df_cleaned['CompanyStateCode'].value_counts().head(10).plot(kind='bar', color='skyblue')
plt.title('Top 10 States with Most Companies')
plt.ylabel('Number of Companies')
plt.xlabel('State Code')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

Visualization



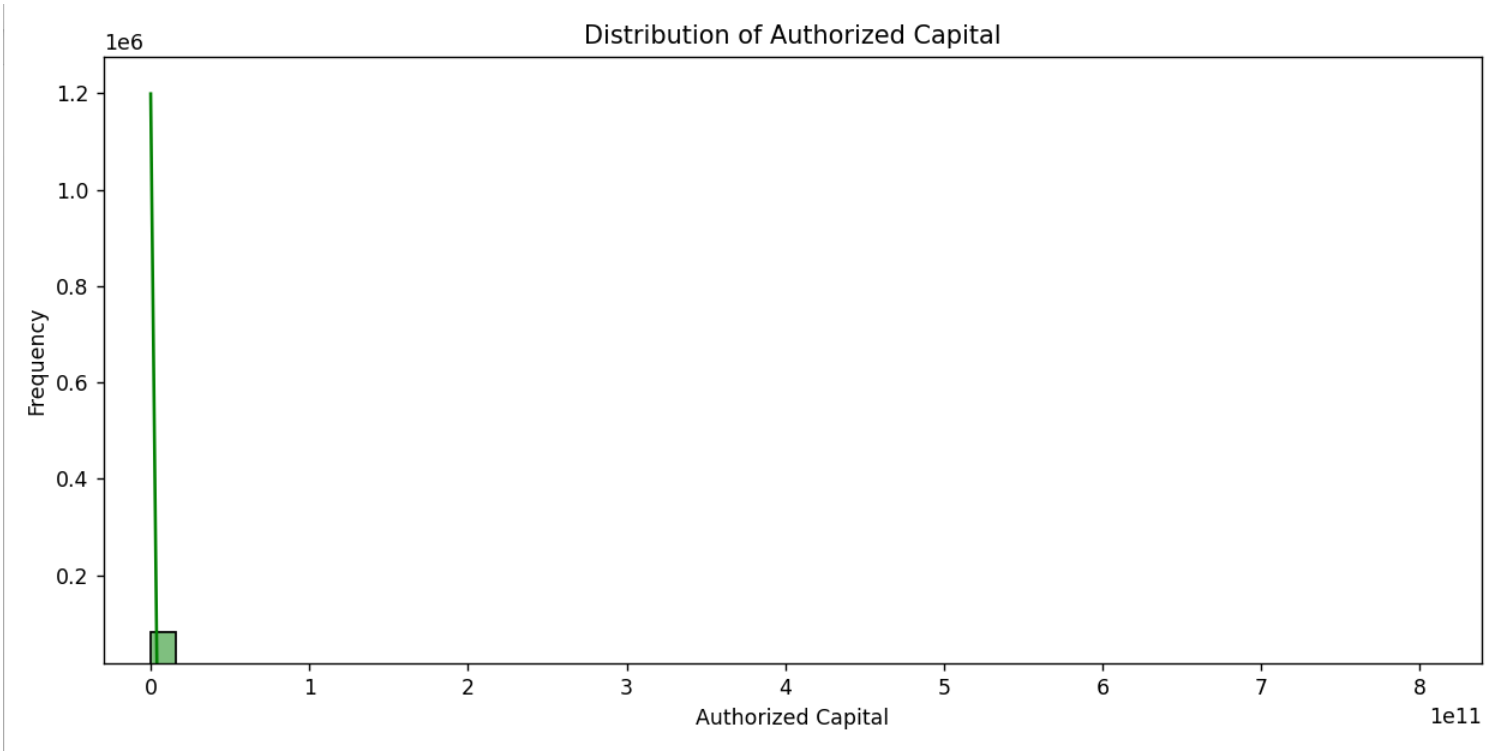
Objective-2 Distribution of Authorized Capital

- Aggregated state-wise claims, premiums, and farmers covered. ○ Sorted to find top-performing or underperforming states. ○ Detected policy reach imbalance between larger and smaller states.

Analysis Result

```
plt.figure(figsize=(10, 5))
sns.histplot(df_cleaned['AuthorizedCapital'], bins=50, kde=True, color='green')
plt.title('Distribution of Authorized Capital')
plt.xlabel('Authorized Capital')
plt.ylabel('Frequency')
plt.tight_layout()
plt.show()
```

Visualization

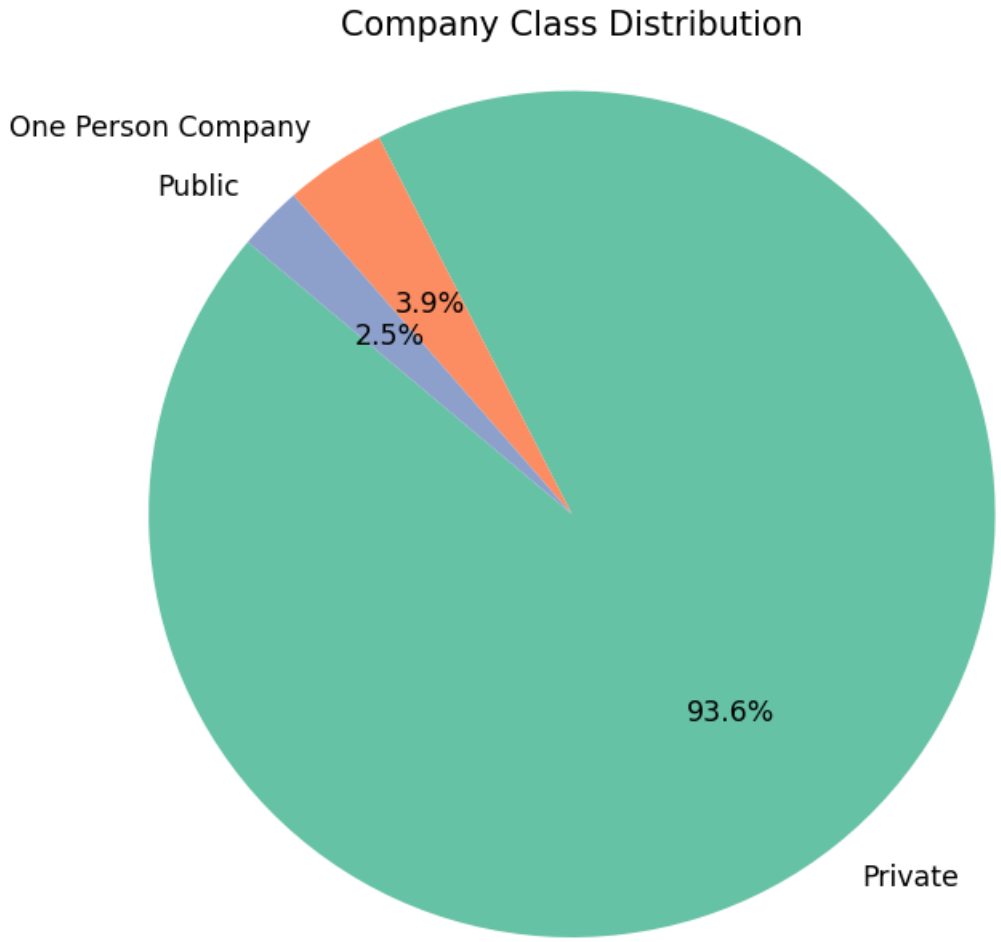


Objective-3: Company Class Distribution

Analysis Result

```
class_counts = df_cleaned['CompanyClass'].value_counts()
plt.figure(figsize=(6, 6))
plt.pie(class_counts, labels=class_counts.index, autopct='%1.1f%%', startangle=140, colors=sns.color_palette('Set2'))
plt.title('Company Class Distribution')
plt.axis('equal')
plt.show()
```

Visualization

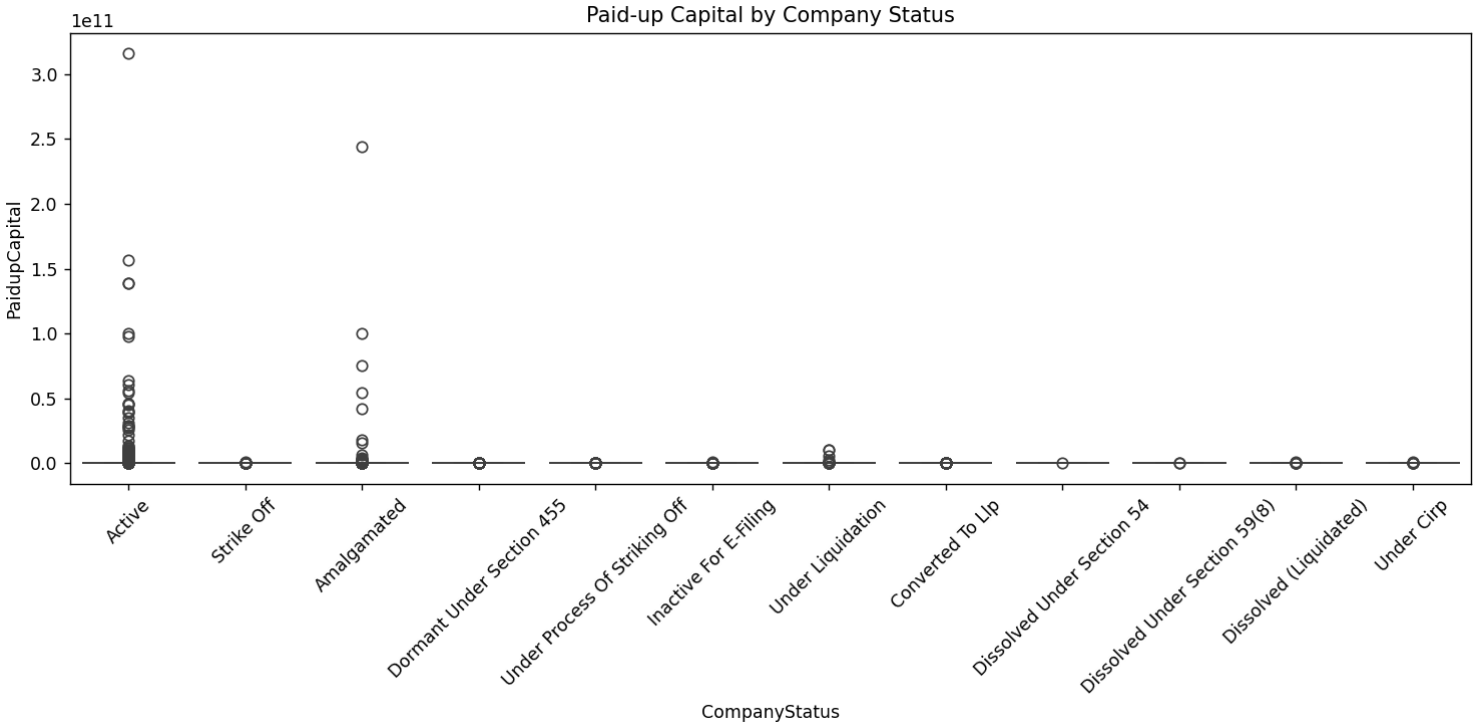


Objective-4: Paid-up Capital by Company Status

Analysis Result

```
plt.figure(figsize=(12, 6))
sns.boxplot(x='CompanyStatus', y='PaidupCapital', data=df_cleaned)
plt.title('Paid-up Capital by Company Status')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

Visualization

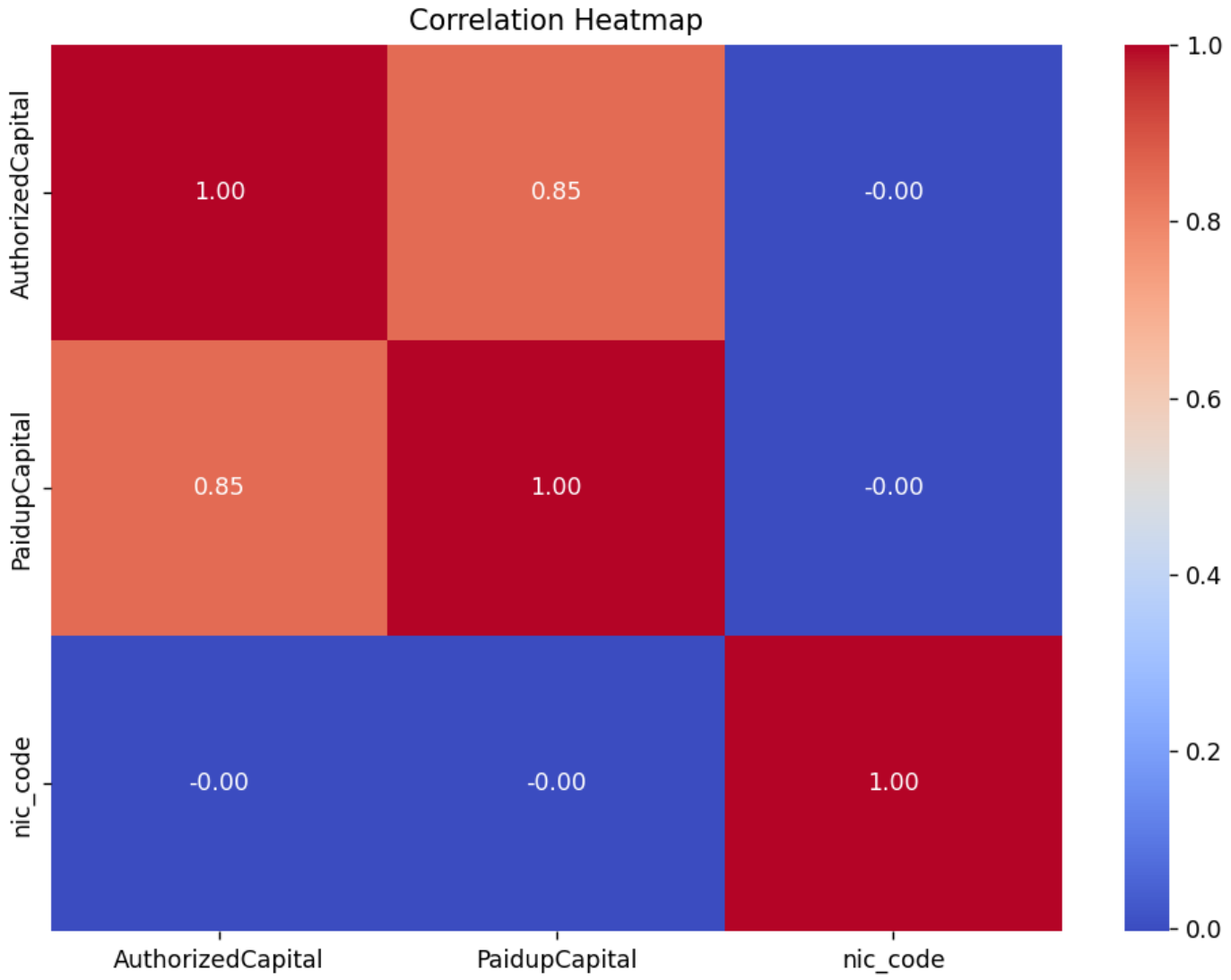


Objective-5: Correlation Heatmap

- **Seasonal Comparison**: ○ Compared Kharif vs Rabi using bar plots. ○ Explored if claims rise more in a particular season due to weather events. **Analysis Result**

```
plt.figure(figsize=(8, 6))
corr = df_cleaned[['AuthorizedCapital', 'PaidupCapital', 'nic_code']].corr()
sns.heatmap(corr, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Heatmap')
plt.tight_layout()
plt.show()
```

Visualization

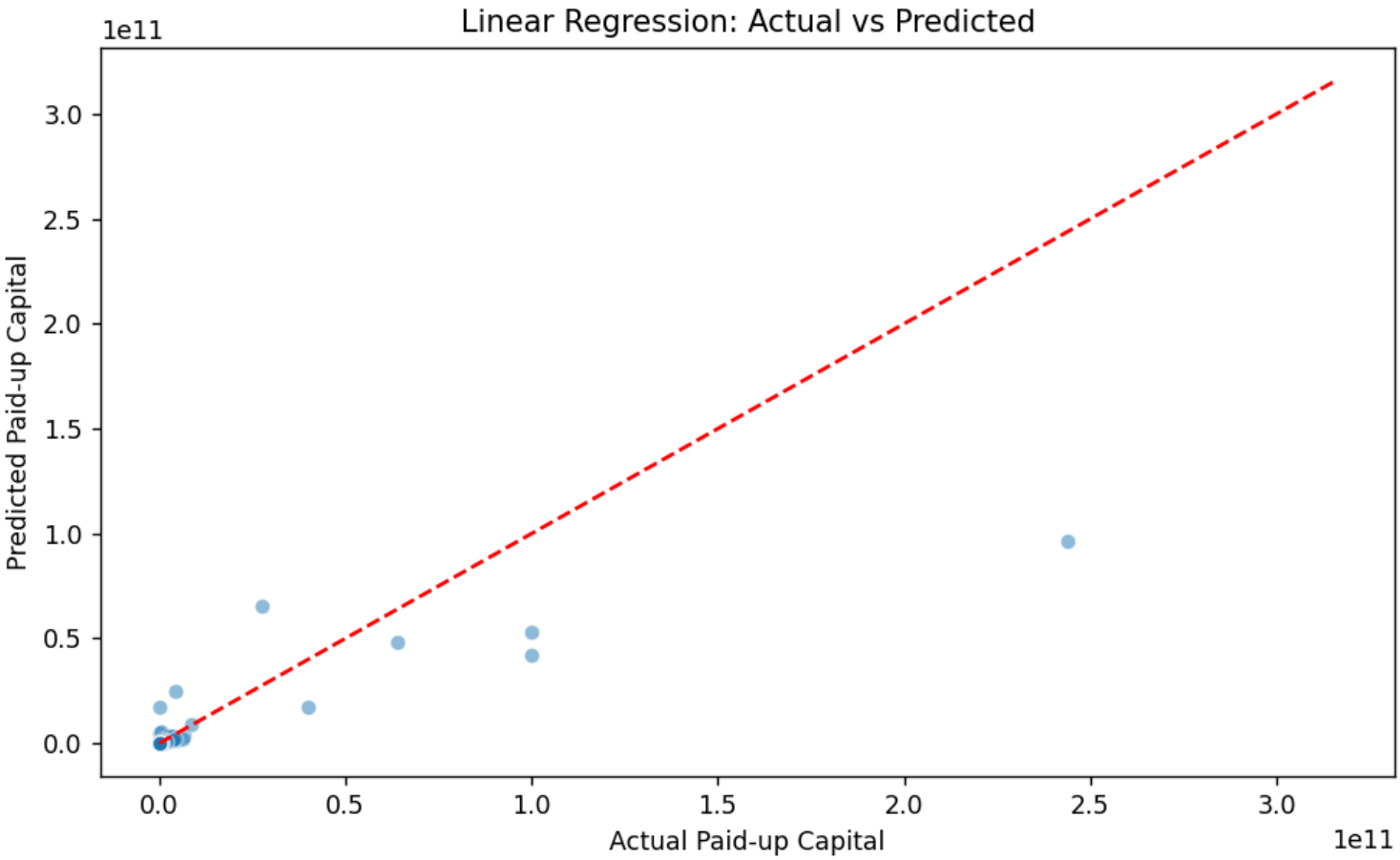


Objective -6: Linear Regression: Actual vs Predicted

Analysis Result

```
plt.figure(figsize=(8, 5))
sns.scatterplot(x=y_test, y=y_pred, alpha=0.5)
plt.xlabel('Actual Paid-up Capital')
plt.ylabel('Predicted Paid-up Capital')
plt.title('Linear Regression: Actual vs Predicted')
plt.plot([y.min(), y.max()], [y.min(), y.max()], 'r--')
plt.tight_layout()
plt.show()
```

Visualization



Coding Part

```python
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score
import numpy as np
import pandas as pd


df = pd.read_csv("4dbe5667-7b6b-41d7-82af
    -211562424d9a_a0d71eebebaa5ba00b5d1af1dd96a3dd.csv")

```

```
df.info(), df.head()

df_cleaned = df.copy()

df_cleaned['CompanyRegistrationdate_date'] = pd.to_datetime
    (df_cleaned['CompanyRegistrationdate_date'], errors='coerce')

df_cleaned.dropna(inplace=True)

df_cleaned['CompanyStatus'] = df_cleaned['CompanyStatus'].str.strip().str.title
    ()
df_cleaned['CompanyClass'] = df_cleaned['CompanyClass'].str.strip().str.title()


plt.figure(figsize=(10, 5))
df_cleaned['CompanyStateCode'].value_counts().head(10).plot(kind='bar', color
    ='skyblue')
plt.title('Top 10 States with Most Companies')
plt.ylabel('Number of Companies')
plt.xlabel('State Code')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

```
plt.figure(figsize=(10, 5))
sns.histplot(df_cleaned['AuthorizedCapital'], bins=50, kde=True, color='green')
plt.title('Distribution of Authorized Capital')
plt.xlabel('Authorized Capital')
plt.ylabel('Frequency')
plt.tight_layout()
plt.show()

class_counts = df_cleaned['CompanyClass'].value_counts()
plt.figure(figsize=(6, 6))
plt.pie(class_counts, labels=class_counts.index, autopct='%1.1f%%', startangle
    =140, colors=sns.color_palette('Set2'))
plt.title('Company Class Distribution')
plt.axis('equal')
plt.show()

plt.figure(figsize=(12, 6))
sns.boxplot(x='CompanyStatus', y='PaidupCapital', data=df_cleaned)
plt.title('Paid-up Capital by Company Status')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

```
plt.figure(figsize=(8, 6))
corr = df_cleaned[['AuthorizedCapital', 'PaidupCapital', 'nic_code']].corr()
sns.heatmap(corr, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Heatmap')
plt.tight_layout()
plt.show()

X = df_cleaned[['AuthorizedCapital']]
y = df_cleaned['PaidupCapital']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
    random_state=42)

model = LinearRegression()
model.fit(X_train, y_train)

y_pred = model.predict(X_test)

mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

plt.figure(figsize=(8, 5))
sns.scatterplot(x=y_test, y=y_pred, alpha=0.5)
```

```
plt.xlabel('Actual Paid-up Capital')
plt.ylabel('Predicted Paid-up Capital')
plt.title('Linear Regression: Actual vs Predicted')
plt.plot([y.min(), y.max()], [y.min(), y.max()], 'r--')
plt.tight_layout()
plt.show()

mse, r2
```

Visualization Graphs And Output

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 91268 entries, 0 to 91267
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   CIN                             91268 non-null  object
 1   CompanyName                     91268 non-null  object
 2   CompanyROCcode                  91268 non-null  object
 3   CompanyCategory                 91268 non-null  object
 4   CompanySubCategory              91268 non-null  object
 5   CompanyClass                    91268 non-null  object
 6   AuthorizedCapital               91268 non-null  float64
 7   PaidupCapital                   91268 non-null  float64
 8   CompanyRegistrationdate_date    91268 non-null  object
 9   Registered_Office_Address       91265 non-null  object
 10  Listingstatus                   91268 non-null  object
 11  CompanyStatus                   91268 non-null  object
 12  CompanyStateCode                91268 non-null  object
 13  CompanyIndian/Foreign Company   91268 non-null  object
 14  nic_code                        91268 non-null  int64
 15  CompanyIndustrialClassification 91268 non-null  object
dtypes: float64(2), int64(1), object(13)
memory usage: 11.1+ MB
```

Conclusion

This project delivers a comprehensive exploratory data analysis of company registration records in India, with a strong focus on identifying structural trends, financial distributions, and regional dynamics. Beginning with critical data preprocessing steps—such as converting date formats, removing missing and duplicate values, and standardizing categorical data—we ensured the dataset was reliable and clean. This step was crucial for eliminating inconsistencies and preparing the data for meaningful group-wise comparisons. By organizing and refining variables like CompanyStatus, CompanyClass, AuthorizedCapital, and PaidupCapital, we were able to surface patterns that would otherwise remain hidden. The dataset, once preprocessed, became a robust foundation for analysis that reflects the diverse and growing nature of India's corporate sector.

The use of insightful visualizations such as bar plots, pie charts, histograms, box plots, and heatmaps allowed for a deeper understanding of the data. For example, bar charts revealed the top states by company registrations, suggesting where economic activity is most concentrated, while box plots uncovered how capital distribution varies by company status. A pie chart showcased the proportional split among different company classes, giving us a quick overview of organizational structures. Additionally, the correlation heatmap and linear regression model highlighted the financial relationship between authorized and paid-up capital, providing a quantitative lens into capital management practices. Altogether, this EDA not only helped uncover valuable business insights but also laid the groundwork for more advanced modeling and strategic business intelligence, empowering data-driven decisions in future analyses.

Future Scope

This exploratory analysis opens the door to several promising avenues for future development in both academic and practical domains. One key direction is the creation of predictive models using machine learning techniques to estimate business success probability, capital growth trends, or company survival rates based on initial registration details. Historical company data can be leveraged to build classification or regression models that aid investors, policy-makers, and entrepreneurs in identifying high-potential ventures, forecasting risks, and optimizing regional economic planning.

Another important opportunity lies in integrating geospatial data, demographic indicators, and regional economic statistics with the existing dataset. This multi-layered approach can help identify clusters of high business activity, underserved regions, or policy-sensitive zones, enabling more targeted economic development strategies. Visualizing this through interactive dashboards or GIS-based tools can make insights more actionable for government agencies and investment boards. Additionally, further enrichment of the dataset with real-time updates (such as compliance status, annual returns, or default flags) can empower regulatory bodies and financial institutions to monitor corporate behavior and ensure transparency. As a long-term vision, partnerships with fintech platforms, startup incubators, and academic institutions could foster data-driven innovation in supporting India’s rapidly evolving business ecosystem.

Reference

[1] The Pandas Development Team, “Pandas Documentation.” [Online]. Available: https://pandas.pydata.org/docs
(Used for data manipulation, cleaning, and initial exploration)

[2] J. D. Hunter et al., “Matplotlib Documentation.” [Online]. Available: https://matplotlib.org/stable/contents.html
(Used for data visualization through plots, histograms, and scatter charts)

[3] M. Waskom et al., “Seaborn Documentation.” [Online]. Available: https://seaborn.pydata.org
(Used for creating advanced statistical visualizations like heatmaps and boxplots)

[4] Ministry of Corporate Affairs, Government of India, “MCA Services and Company Registration Information.” [Online]. Available: https://www.mca.gov.in
(Primary source for understanding company classification, registration norms, and capital structure)

[5] S. Sharma and A. Gupta, “Analyzing Trends in Company Formation and Capital Allocation in India,” Indian Journal of Economics & Development, 2022.
(Reference for context on economic trends and business registration analysis)

[6] M. Singh, “Exploratory Data Analysis in Business Data Using Python,” Data Science Journal, 2021.
(Referenced for EDA methodology and visualization best practices)

Linkedin post

Linkedin link: https://www.linkedin.com/posts/umanggarg0210_python-exceldashboard-datascience-activity-7317070285009010689-whEQ?utm_source=share&utm_medium=member_desktop&rcm=ACoAAEX_db0BwehLmAy1kRo0x2F2MbORWNW9SGo**Linkedin Post:**

Umang Garg • You
B.tech CSE @LPU || BS in Data Science and Applications @IIT Madras || Fronte...
1d •

🚀 Just wrapped up a data-driven project that combines the power of Python with the clarity of an Excel dashboard! 📊

I took a deep dive into a dataset of company registrations and used Python to extract actionable insights. Here's what I did:

- Cleaned & prepared the data using pandas
- Visualized trends like:
- Top states by company registrations
- Distribution of Authorized vs Paid-up Capital
- Company Class & Status breakdowns
- Built a Linear Regression model to predict Paid-up Capital based on Authorized Capital using `scikit-learn`
- Created an interactive Excel dashboard to present all key insights in a clean, business-friendly format

This project was a great way to blend data wrangling, statistical analysis, and visualization — all tied together with storytelling through a dashboard.

Skills sharpened: Python (pandas, seaborn, matplotlib, scikit-learn), Excel, Data Cleaning, EDA, Regression Modeling, and Dashboard Design.

Super proud of the final output — if you're curious to check it out or discuss similar ideas, let's connect!

#Python #ExcelDashboard #DataScience #DataAnalytics #MachineLearning #Pandas #Matplotlib #Seaborn #RegressionModel #BusinessIntelligence #DataVisualization #ProjectShowcase