Project Report: Data Cleaning and Analysis on Company Registration Dataset

****

**LOVELY PROFESSIONAL UNIVERSITY, PHAGWARA, PUNJAB**

**SUBMITTED BY -** Prashant Panwar  
**Registration Number:** 12301139  
**Section – K23FD**  
**Roll no - 55**

**DECLARATION :**

I, **Prashant Panwar**, hereby declare that the work done by me on “Company Registration Data Analysis” from March 2025 to April 2025, is a record of original work for the partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology in Computer Science - Data Science** , Lovely Professional University, Phagwara.

**Signature**                 **Signature**  
Name: Prashant Panwar          Ms. Baljinder Kaur  
Reg No: 12301139             UID: *27952*

**ACKNOWLEDGMENT**

First and foremost, I would like to express my deepest gratitude to my college for providing me with the opportunity and resources to undertake this project.  
I extend my sincere thanks to my mentor, **Ms. Baljinder Kaur**, for his invaluable guidance, constructive feedback, and constant encouragement throughout the project. His expertise and support were instrumental in achieving the objectives of this work.  
I would also like to acknowledge the support of my friends and family, whose encouragement and motivation kept me focused and dedicated.  
Finally, I am grateful to everyone whose efforts, directly or indirectly, contributed to the successful completion of this project.  
**Thank you all for your unwavering support.**  
**Prashant Panwar**

**Table of Contents**

1. Abstract
2. Introduction
3. Methodology
4. Results and Analysis
   * Key Findings
   * Visualizations and Tables
   * Interpretation of Results
5. Conclusion
6. References
   * Cited Papers, Articles, and Resources
7. GitHub Repository
   * **PROJECT LINK -** [**https://github.com/Prashant-Best/python-project-on-up-data**](https://github.com/Prashant-Best/python-project-on-up-data)

**Linkedin Post:-**[**https://www.linkedin.com/posts/prashant-panwar-948265259\_python-dataanalytics-visualization-activity-7316488962099793920-nC2g?utm\_source=share&utm\_medium=member\_desktop&rcm=ACoAAD-TdQgByYQ\_24KX3ml6gRx-cDifDM0DvxY**](https://www.linkedin.com/posts/prashant-panwar-948265259_python-dataanalytics-visualization-activity-7316488962099793920-nC2g?utm_source=share&utm_medium=member_desktop&rcm=ACoAAD-TdQgByYQ_24KX3ml6gRx-cDifDM0DvxY)

**Dataset:-**[**http://data.gov/**](http://data.gov/)

**Project Report: Data Cleaning and Analysis on Company Registration Dataset**

**1. Introduction**

Data plays a crucial role in decision-making, especially when it comes to understanding public participation in elections. This project focuses on analyzing a real-world dataset containing election company statistics across various States and Union Territories in India. The primary goal is to clean the dataset, explore the key features, derive insights, and visualize the patterns using Python.

By leveraging libraries such as **Pandas, Matplotlib, and Seaborn**, we perform Exploratory Data Analysis (EDA) and draw meaningful conclusions that could help in understanding registration trends and classifications.

**2. Source of Dataset**

The dataset used for this project is titled **"UP DATA.csv"**, and it contains demographic details of companies categorized into Authorized Capital, Paidup Capital, and Company Status, along with the total number of companies across each State/UT.

**Format:** CSV (Comma-Separated Values)  
**Columns:**

* **State/UT & Code**
* **Authorized Capital**
* **Paidup Capital**
* **Company Status**
* **Total**

**3. EDA (Exploratory Data Analysis) Process**

Steps followed:

* Loaded the dataset using pandas.read\_csv()
* Displayed basic information: shape, column names, data types
* Checked for missing values and removed rows with nulls
* Performed statistical analysis (mean, median, mode)
* Plotted graphs to visualize data

**4. Analysis on Dataset**

### **1. Distribution of Companies by Class**

**Plot:** Bar Plot (sns.countplot)  
**Column Used:** CompanyClass

**Detailed Explanation:**  
This analysis identifies how newly registered companies are classified in legal terms. Common classifications include:

* **Private Limited Companies**: Preferred by startups and small businesses.
* **Public Limited Companies**: Typically larger enterprises.
* **One Person Companies (OPC)**: Introduced to encourage solo entrepreneurship.

**Objective Insight:**  
Understanding the proportions of each class helps assess the entrepreneurial landscape. A dominance of Private Ltd. companies may suggest a startup-driven economy, while more Public Ltd. firms could indicate mature industrial participation.

### **2. Company Status Distribution**

**Plot:** Horizontal Bar Plot  
**Column Used:** CompanyStatus

**Detailed Explanation:**  
This section identifies how many companies are:

* **Active**: Currently operational
* **Strike Off**: Removed due to non-compliance or inactivity
* **Dissolved/Wound up**: Officially closed

**Objective Insight:**  
This metric is vital to understand **business survival rates**. A high ratio of inactive or dissolved companies can point to challenges in business sustainability, policy barriers, or economic instability.

### **3. Capital Structure Analysis**

**Plot:** Box Plot (log scale)  
**Columns Used:** AuthorizedCapital, PaidupCapital

**Detailed Explanation:**  
This evaluates the financial capital declared by companies:

* **Authorized Capital**: The maximum capital a company can raise.
* **Paid-up Capital**: The actual money received from shareholders.

Using a **logarithmic scale** helps visualize highly skewed data, where a few companies might have very high capital.

**Objective Insight:**  
This shows how well-funded the companies are, and whether most firms are **small-cap** startups or **well-capitalized** ventures

### **4. Top 10 Industrial Classifications**

**Plot:** Bar Plot  
**Column Used:** CompanyIndustrialClassification

**Detailed Explanation:**  
This determines which industry verticals are attracting the most company registrations, such as:

* Information Technology
* Retail
* Real Estate
* Manufacturing

**Objective Insight:**  
Helps in identifying **economic growth sectors**. For example, a rise in tech startups indicates digital transformation trends, while dominance in real estate could suggest infrastructure investment.

### **5. Company Registrations Over the Years**

**Plot:** Line Plot  
**Column Used:** RegistrationYear

**Detailed Explanation:**  
Analyzing company formation trends by year reveals:

* Booms (possibly after government startup schemes)
* Declines (due to economic slowdowns or global events like COVID-19)

**Objective Insight:**  
This objective maps **temporal trends** in entrepreneurship and helps correlate external events (like demonetization or GST) with changes in business registrations.

### **6. Distribution of Authorized Capital**

**Plot:** Histogram (log scale)  
**Column Used:** AuthorizedCapital

**Detailed Explanation:**  
This histogram shows how many companies fall into different capital ranges. It answers:

* Are most businesses small-scale (< ₹1 lakh)?
* How many companies are high-cap (₹10 crore+)?

**Objective Insight:**  
It provides a **financial depth analysis** of the entrepreneurial base. It can also help investors or policy-makers identify which segments to support.

### **7. Correlation Between Numerical Features**

**Plot:** Heatmap  
**Columns Used:** All numeric fields (e.g., capital values, year)

**Detailed Explanation:**  
A heatmap displays pairwise correlations between numerical columns:

* Are higher authorized capitals associated with more recent registrations?
* Is there a linear relationship between authorized and paid-up capital?

**Objective Insight:**  
Understanding these correlations helps discover **hidden patterns**, useful in predictive modeling or strategic planning.

**5. Conclusion**

This project successfully delivered a comprehensive exploratory data analysis of company registration records from Uttar Pradesh using Python and key data science libraries like Pandas, Seaborn, and Matplotlib.

Key takeaways include:

* **Private Limited Companies** emerged as the most common company type, indicating a healthy startup ecosystem and entrepreneurial inclination toward scalable business models.
* The majority of companies were classified as **Active**, but a noticeable portion were found in the **Strike Off** category, pointing toward potential challenges in long-term business viability.
* Analysis of **Authorized and Paid-up Capital** revealed a highly skewed distribution, with a small number of companies holding very high capital. This underlines the presence of a few dominant players amidst a majority of small-scale businesses.
* **Industry classification data** highlighted the leading economic sectors attracting company formation, enabling a better understanding of regional industrial focus.
* **Temporal analysis** of company registrations showed year-wise fluctuations, revealing the possible effects of external economic policies or conditions on entrepreneurship.
* The **correlation matrix** provided insight into relationships among capital metrics and registration timelines, serving as a basis for future predictive modeling.

Through this analysis, the project demonstrates how structured data analysis can reveal business patterns, support decision-making, and provide groundwork for deeper investigations such as forecasting or policy evaluation.

**6. Future Scope**

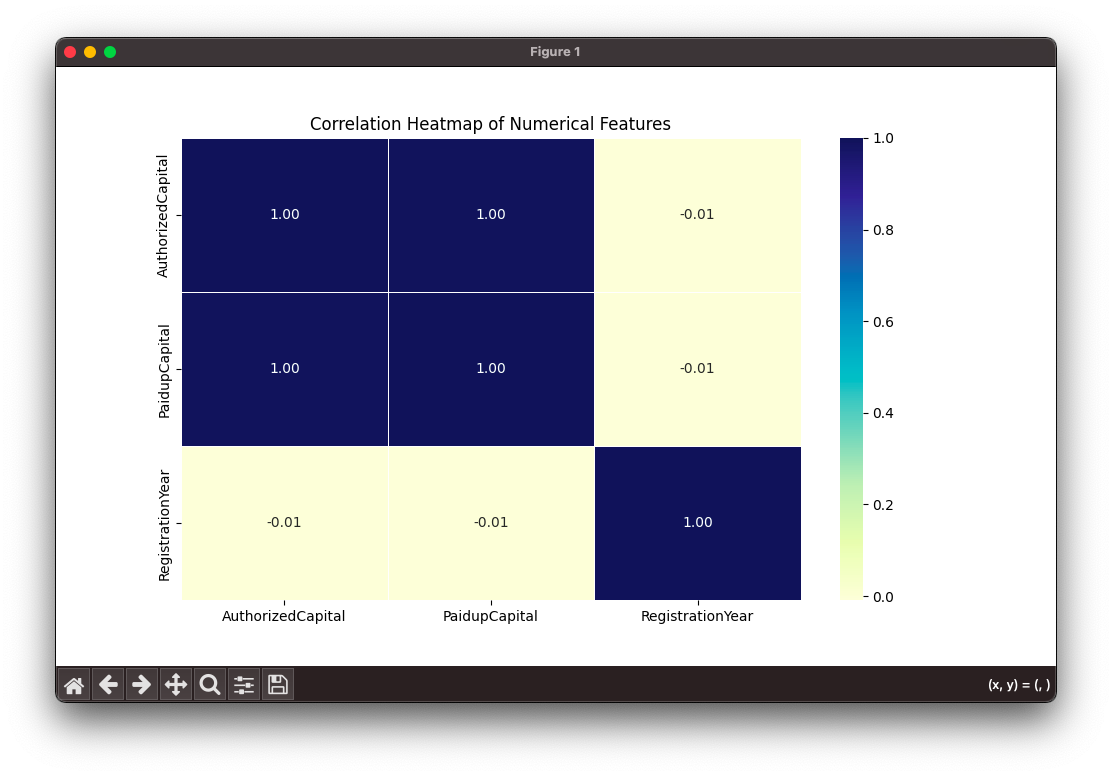
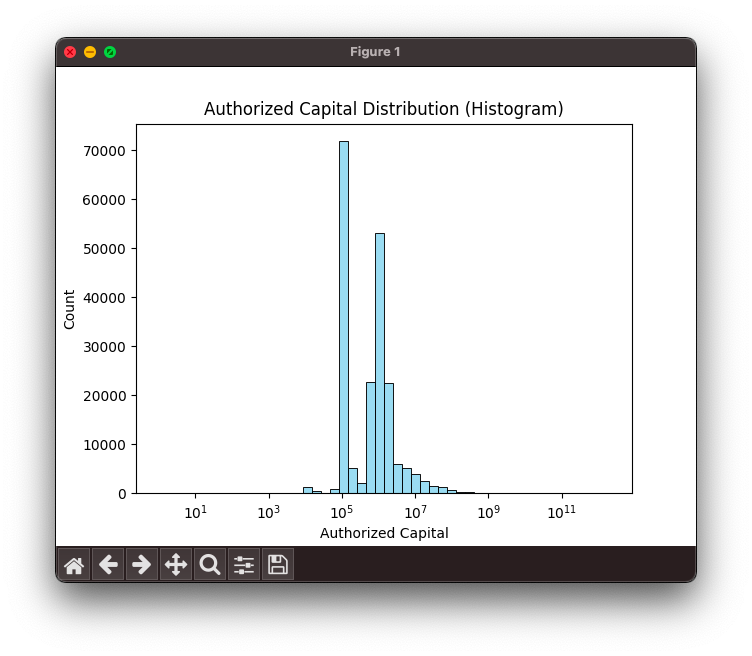
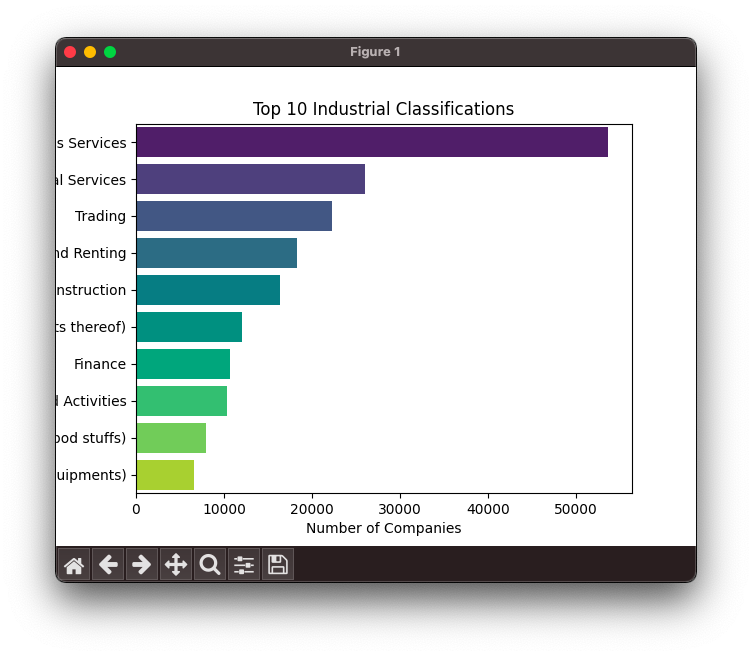
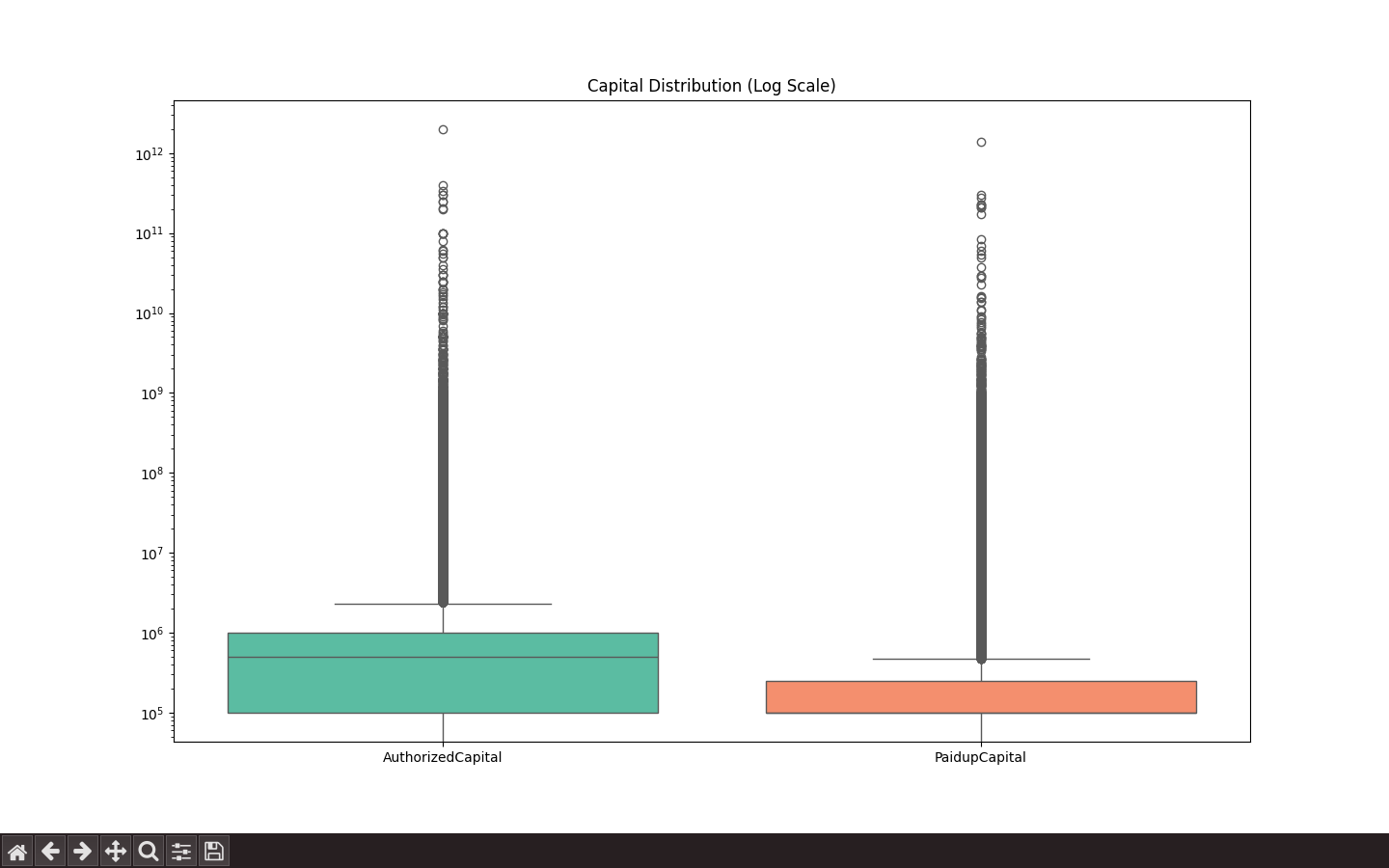
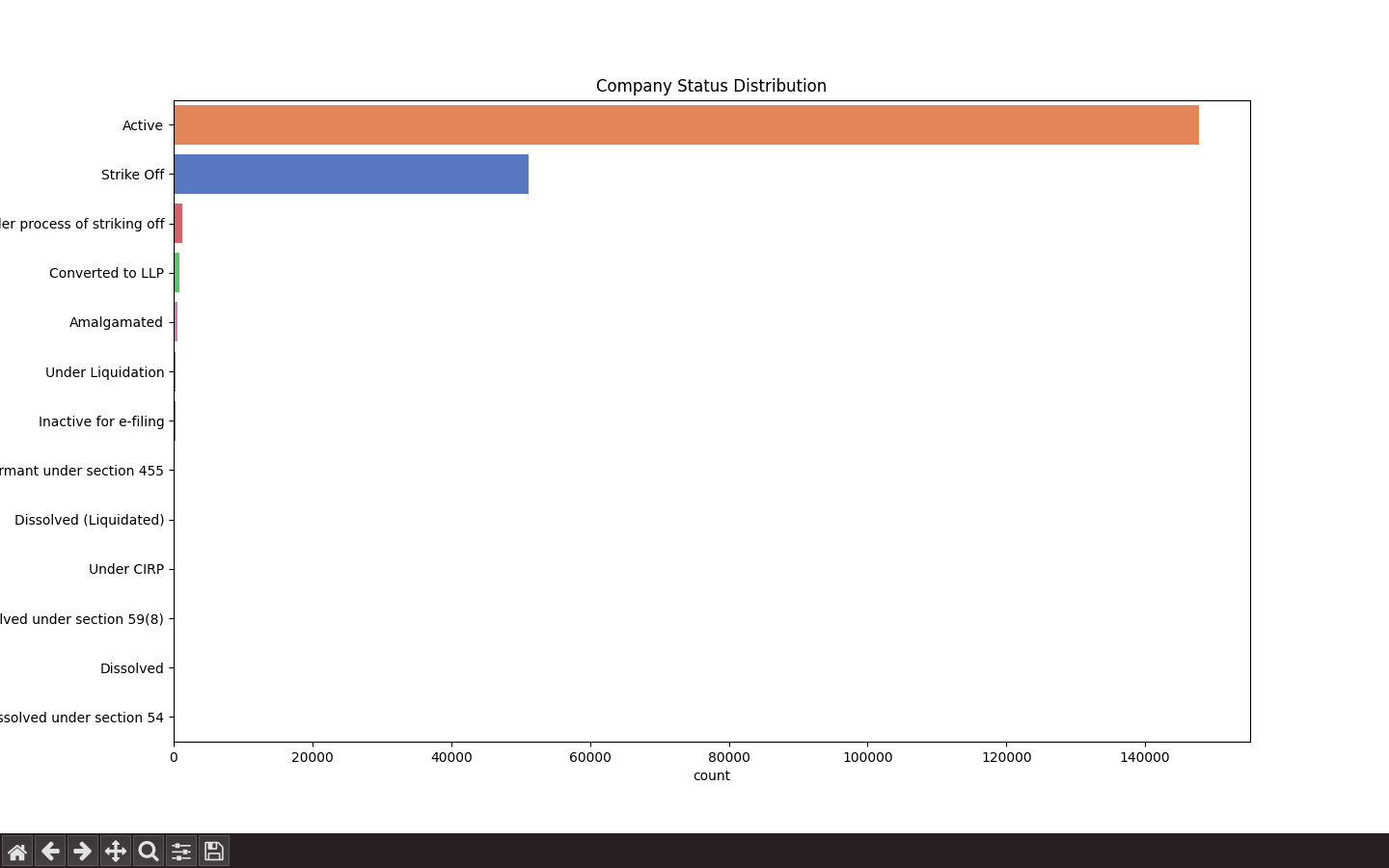
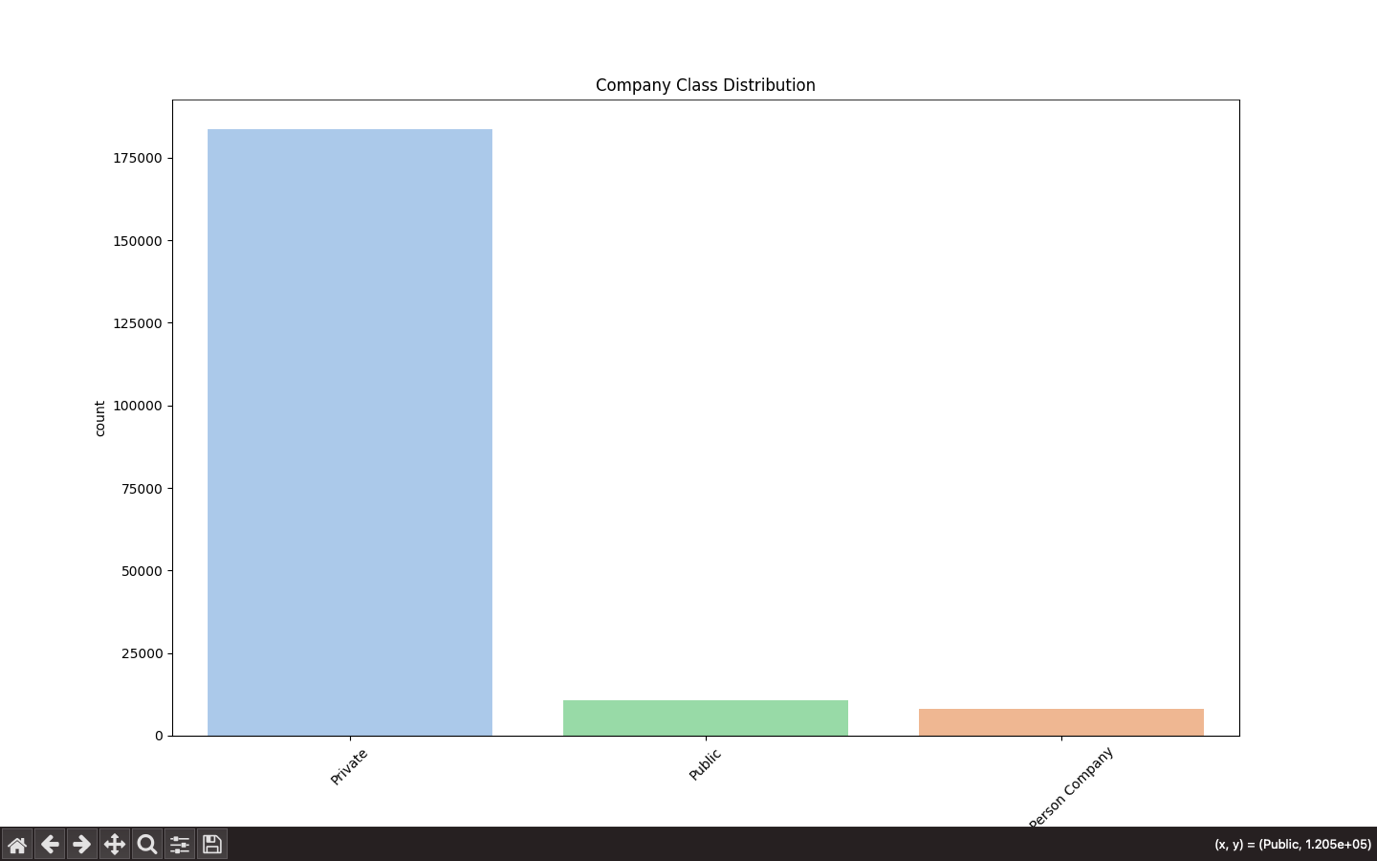
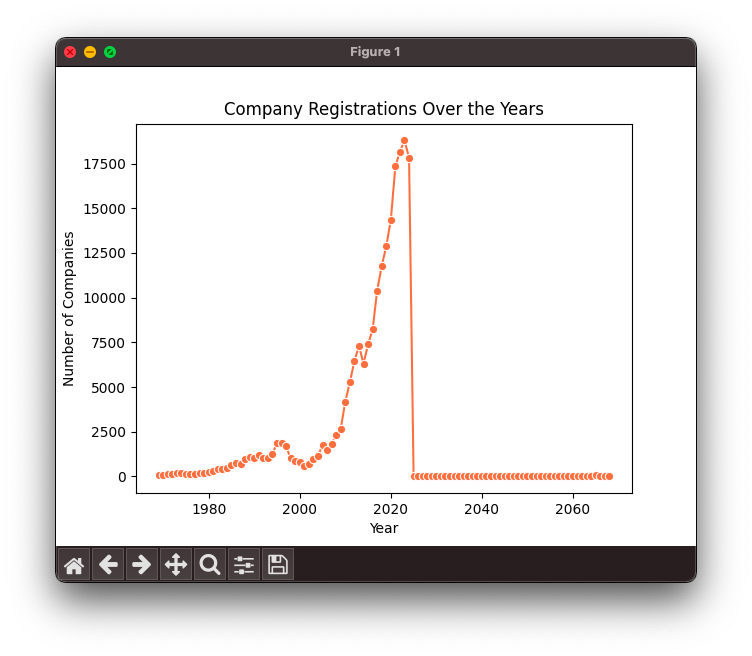
The current analysis lays the foundation for more advanced and targeted explorations. Here are several directions for future work:

1. **📊 Interactive Dashboards:**
   * Use libraries like **Plotly**, **Dash**, or **Streamlit** to build user-friendly dashboards for real-time exploration of company data by stakeholders and policymakers.
2. **📍 Geospatial Analysis:**
   * Incorporate geographic data (e.g., city/district/state) to perform **regional mapping** of company trends using tools like **GeoPandas** or **Folium**.
   * This can help identify business hotspots and underdeveloped zones.
3. **📅 Time-Series Forecasting:**
   * Apply models such as **ARIMA, Prophet**, or **LSTM** to forecast future company registrations based on past trends.
   * Useful for economic planning and startup ecosystem monitoring.
4. **📂 Industry-Specific Deep Dives:**
   * Perform sectoral analysis on specific industries (e.g., healthcare, tech, agriculture) to understand unique characteristics, growth rates, and capital patterns.
5. **🤖 Predictive Modeling:**
   * Train machine learning models to predict:
     + Company survival chances based on initial registration data.
     + High-growth sectors for new business opportunities.
6. **📄 Policy Impact Analysis:**
   * Study the impact of government schemes (like Startup India or GST reforms) on company registration rates and capital structure changes.
7. **💼 Integration with Financial & Tax Data:**
   * Merging with external datasets (e.g., tax filings, funding records) can allow more holistic financial health analysis of registered companies.

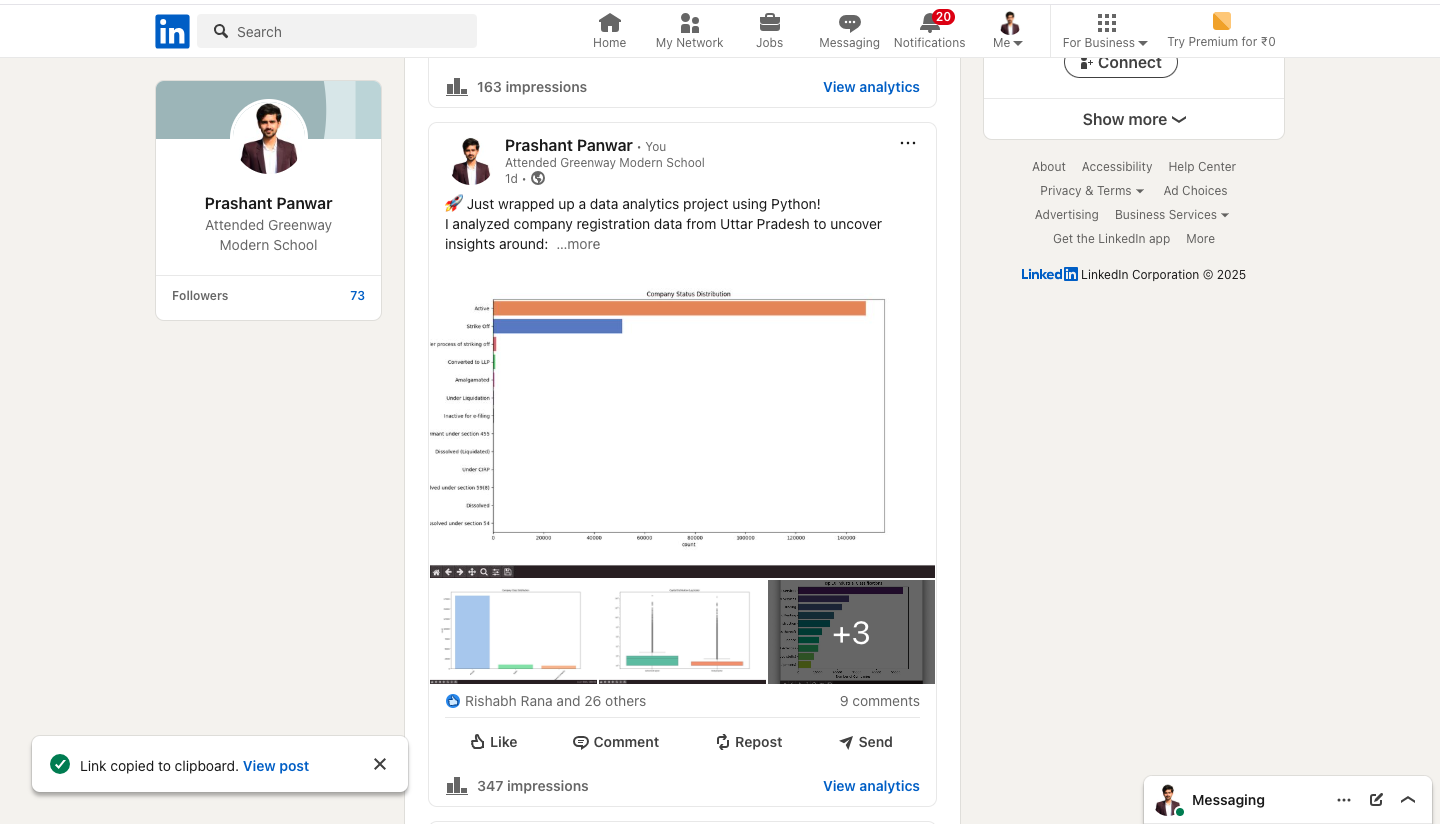
**7. References**

* [Python Documentation](https://docs.python.org)
* [Pandas Library](https://pandas.pydata.org)
* [Seaborn Visualization](https://seaborn.pydata.org)
* [Matplotlib](https://matplotlib.org)
* Dataset: *Local "UP DATA.csv" file provided*

*Screenshots:-*



Linkedin Post:-



Code:-

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

# Load data

raw = pd.read\_csv('UP DATA.csv', low\_memory=False)

# Data overview

print("=== HEAD ===")

print(raw.head(), "\n")

print("=== TAIL ===")

print(raw.tail(), "\n")

print("=== INFO ===")

print(raw.info(), "\n")

print("=== DESCRIPTION ===")

print(raw.describe(include='all'), "\n")

print("=== NULL VALUES PER COLUMN BEFORE ANY CLEANING ===")

print(raw.isnull().sum(), "\n")

# Changing null values to its mean

raw["AuthorizedCapital"] = raw["AuthorizedCapital"].fillna(raw["AuthorizedCapital"].mean())

raw["PaidupCapital"] = raw["PaidupCapital"].fillna(raw["PaidupCapital"].mean())

print("=== NULL VALUES PER COLUMN AFTER CLEAING 2 COLUMS CLEANING ===")

print(raw.isnull().sum(), "\n")

# Drop null values and work on a clean copy

df = raw.dropna().copy()

print("=== NULL VALUES PER COLUMN AFTER CLEAING CLEANING ===")

print(df.isnull().sum(), "\n")

# Convert registration date to datetime format (dd/mm/yy)

df["CompanyRegistrationdate\_date"] = pd.to\_datetime(

df["CompanyRegistrationdate\_date"],

format="%d/%m/%y",

errors="coerce"

)

# Drop rows with invalid dates

df = df[df["CompanyRegistrationdate\_date"].notnull()].copy()

# Extract registration year

df["RegistrationYear"] = df["CompanyRegistrationdate\_date"].dt.year

# Plot 1: Company Class Distribution

sns.countplot(

data=df,

x="CompanyClass",

order=df["CompanyClass"].value\_counts().index,

hue="CompanyClass",

palette="pastel",

legend=False

)

plt.title("Company Class Distribution")

plt.xticks(rotation=45)

plt.show()

# Plot 2: Company Status Distribution

sns.countplot(

data=df,

y="CompanyStatus",

order=df["CompanyStatus"].value\_counts().index,

hue="CompanyStatus",

palette="muted",

legend=False

)

plt.title("Company Status Distribution")

plt.show()

# Plot 3: Capital Distribution (log scale)

sns.boxplot(data=df[["AuthorizedCapital", "PaidupCapital"]], palette="Set2")

plt.yscale("log")

plt.title("Capital Distribution (Log Scale)")

plt.show()

# Plot 3.1: Capital Distribution (After outliner removal)

Q1\_AC = df["AuthorizedCapital"].quantile(0.25)

Q3\_AC = df["AuthorizedCapital"].quantile(0.75)

IQR\_AC = Q3\_AC - Q1\_AC

df = df[

(df["AuthorizedCapital"] >= (Q1\_AC - 1.5 \* IQR\_AC)) &

(df["AuthorizedCapital"] <= (Q3\_AC + 1.5 \* IQR\_AC))

]

# Remove outliers for PaidupCapital

Q1\_PC = df["PaidupCapital"].quantile(0.25)

Q3\_PC = df["PaidupCapital"].quantile(0.75)

IQR\_PC = Q3\_PC - Q1\_PC

df = df[

(df["PaidupCapital"] >= (Q1\_PC - 1.5 \* IQR\_PC)) &

(df["PaidupCapital"] <= (Q3\_PC + 1.5 \* IQR\_PC))

]

sns.boxplot(data=df[["AuthorizedCapital", "PaidupCapital"]], palette="Set2")

plt.yscale("log")

plt.title("Capital Distribution (Log Scale) After Removing Outliers")

plt.show()

# Plot 4: Top 10 Industrial Classifications

top\_industries = df["CompanyIndustrialClassification"].value\_counts().nlargest(10)

sns.barplot(

y=top\_industries.index,

x=top\_industries.values,

hue=top\_industries.index,

palette="viridis",

legend=False

)

plt.title("Top 10 Industrial Classifications")

plt.xlabel("Number of Companies")

plt.show()

# Plot 5: Company Registrations Over the Years

reg\_years = df["RegistrationYear"].value\_counts().sort\_index()

sns.lineplot(x=reg\_years.index, y=reg\_years.values, marker="o", color="coral")

plt.title("Company Registrations Over the Years")

plt.xlabel("Year")

plt.ylabel("Number of Companies")

plt.show()

# Plot 6: Distribution of Authorized Capital (Histogram)

sns.histplot(df["AuthorizedCapital"], bins=50, color='skyblue', log\_scale=True)

plt.title("Authorized Capital Distribution (Histogram)")

plt.xlabel("Authorized Capital")

plt.show()

# Plot 7: Full Correlation Heatmap of Numerical Features

plt.figure(figsize=(10, 6))

corr = df.select\_dtypes(include=["number"]).corr()

sns.heatmap(corr, annot=True, fmt=".2f", cmap="YlGnBu", linewidths=0.5)

plt.title("Correlation Heatmap of Numerical Features")

plt.show()