Exploratory Data Analytics Project-2

Startups

C. Yashwanth Chowdary

Exploratory Data Analysis Project

Venture Capital Investments Analysis

1. Introduction

Venture capital (VC) is a crucial source of funding for startups and early-stage companies with high growth potential. Analyzing VC investment data can provide valuable insights into industry trends, funding patterns, and factors influencing startup success. This project focuses on performing Exploratory Data Analysis (EDA) on a venture capital investments dataset to uncover meaningful patterns and relationships.

2. Objective

The primary objectives of this project are

- To clean and preprocess the venture capital investment data
- To perform Exploratory Data Analysis (EDA) to understand the dataset
- To generate meaningful insights using visualisations
- To identify patterns such as industry trends, funding distributions, and factors influencing startup success
- To develop predictive models for startup success probability

3. Data Overview

The dataset used in this analysis consists of data about startups and their respective investment details. The columns in the dataset include:

- **name**: Name of the startup
- **category_list**: Industry category of the startup
- **city**: City where the startup is located
- **state_code**: State code
- **country_code**: Country code
- **funding_rounds**: Number of funding rounds
- **founded_year**: Year when the startup was founded
- **status**: Current status of the startup (operating, closed, etc.)
- **raised_amount_usd**: Total funding amount in USD
- And more…

4. Data Cleaning and Preprocessing

The following steps were performed to clean and prepare the data:

1. **Loading Required Libraries**: Essential R packages were loaded for data manipulation, visualisation, and modelling.
2. **Data Import**: The dataset was imported from the CSV file.
3. **Data Type Conversion**: Columns with numerical values (e.g., funding_total_usd, funding_rounds, founded_year) were converted to appropriate numeric types.
4. **Handling Special Characters**: Commas in the funding_total_usd column were removed to convert it to a numeric format.
5. **Column Selection**: Only relevant columns were kept for the analysis.
6. **Handling Missing Data**: Rows with missing values in important columns (funding_rounds, founded_year, status) were removed.
7. **Feature Engineering**: A new column, “startup_age," was created by calculating the difference between the current year (2025) and the founding year.

5. Exploratory Data Analysis (EDA)

5.1. Basic Summary

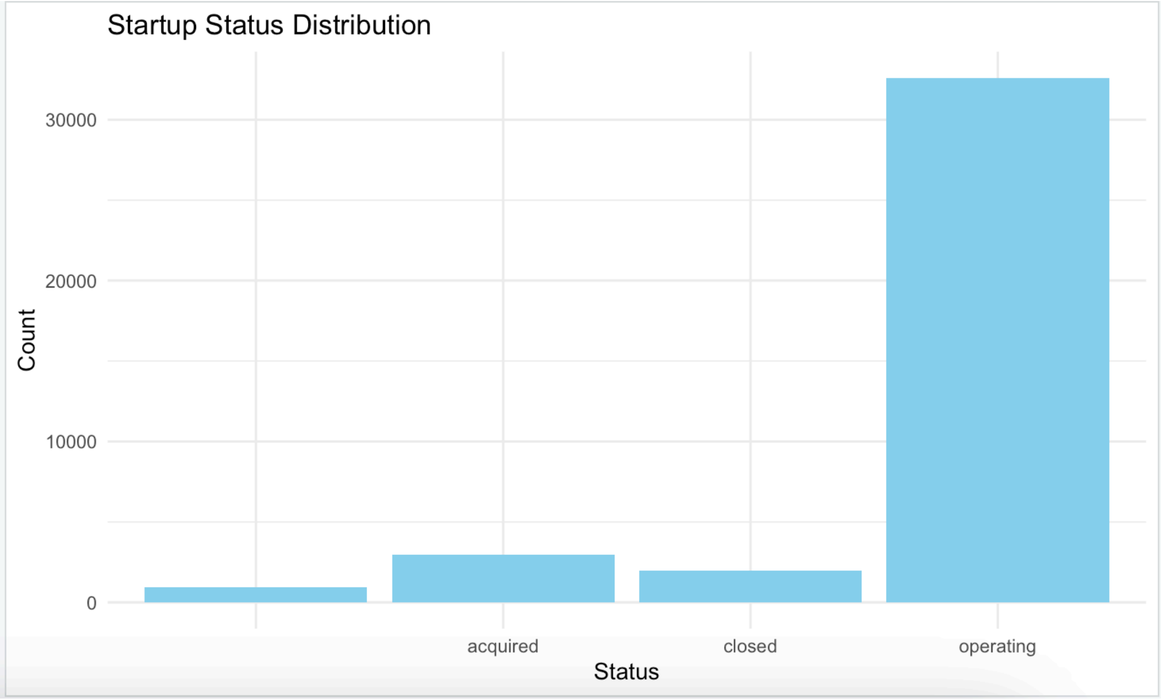
- A summary of the dataset was obtained using the `summary()` function, providing an overview of the data, including the mean, median, and range for numeric variables.
- The missing data was identified using the `colSums(is.na())` function to ensure no critical data points were left unaddressed.

5.2. Visualizations

The following visualizations were created to provide deeper insights:

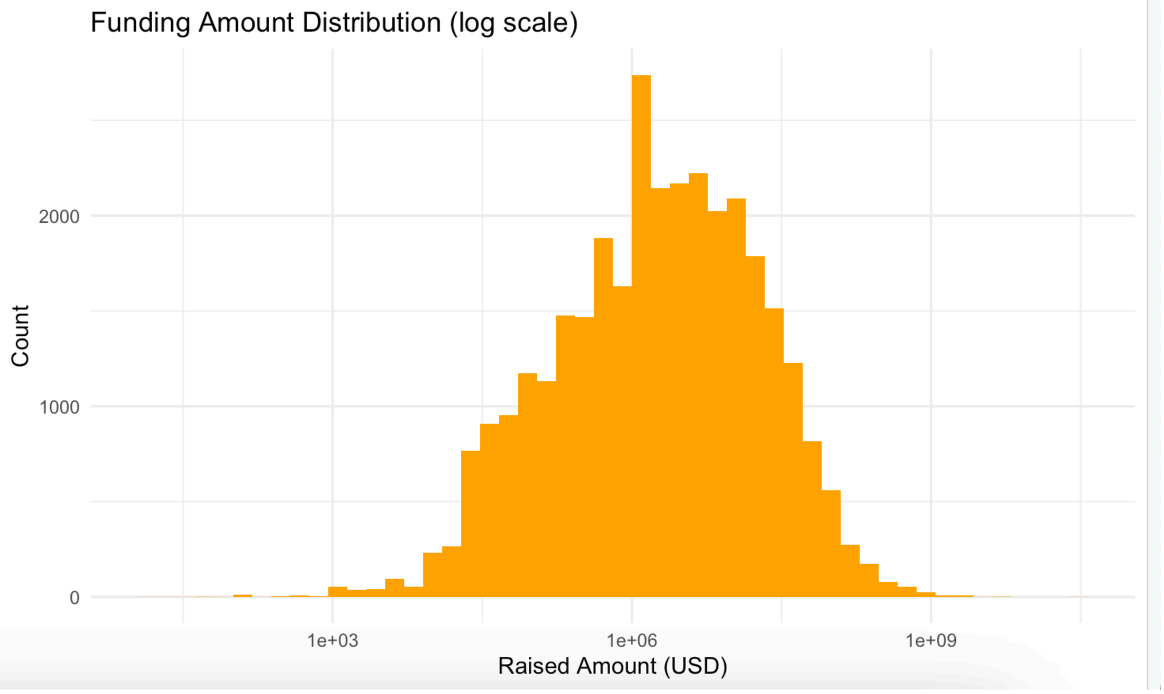
1. Startup Status Distribution

A bar chart was plotted to visualize the distribution of startup status (operating, closed, etc.), showing the success rate of ventures in the dataset.



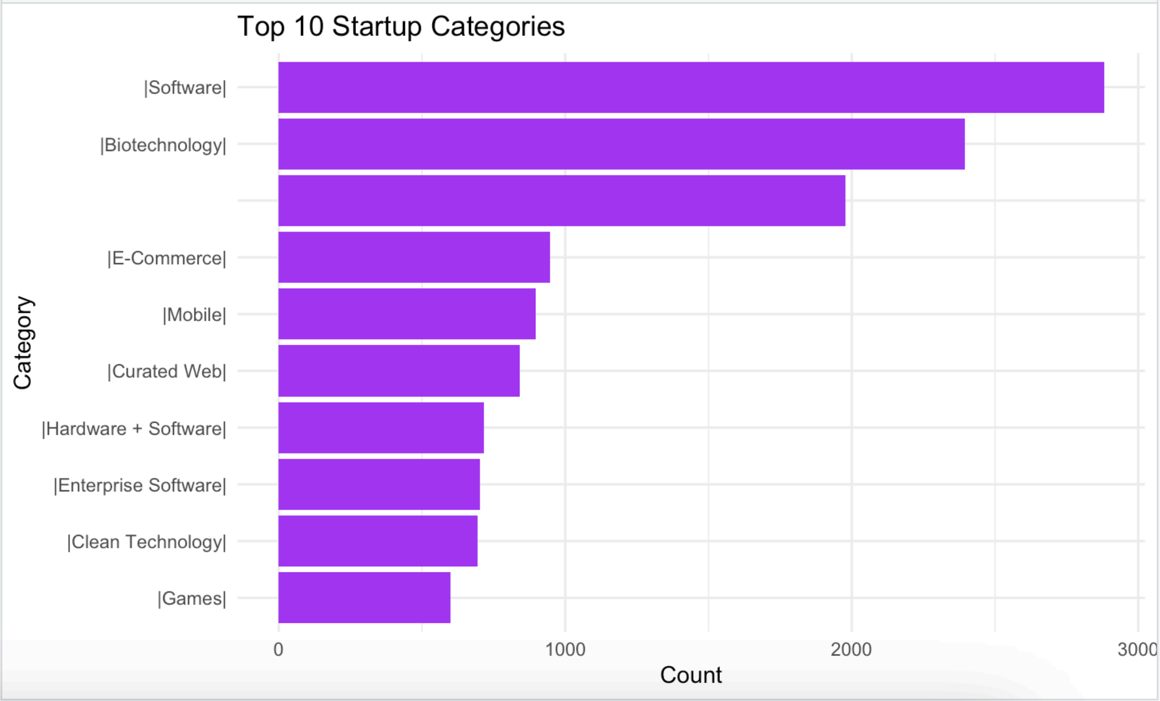
2. Funding Amount Distribution

A histogram displayed the distribution of funding amounts on a logarithmic scale, revealing the skewed nature of venture funding.



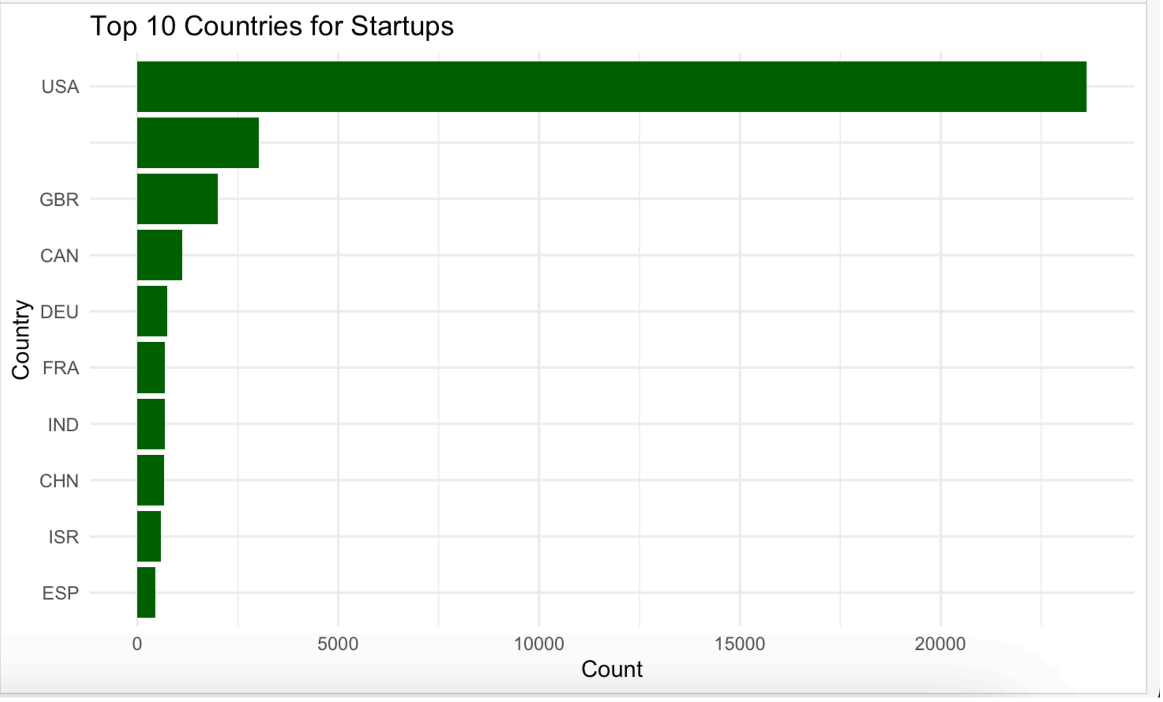
3. Top 10 Startup Categories

A bar chart highlighted the most common startup categories/industries, showing which sectors attract the most entrepreneurial activity.



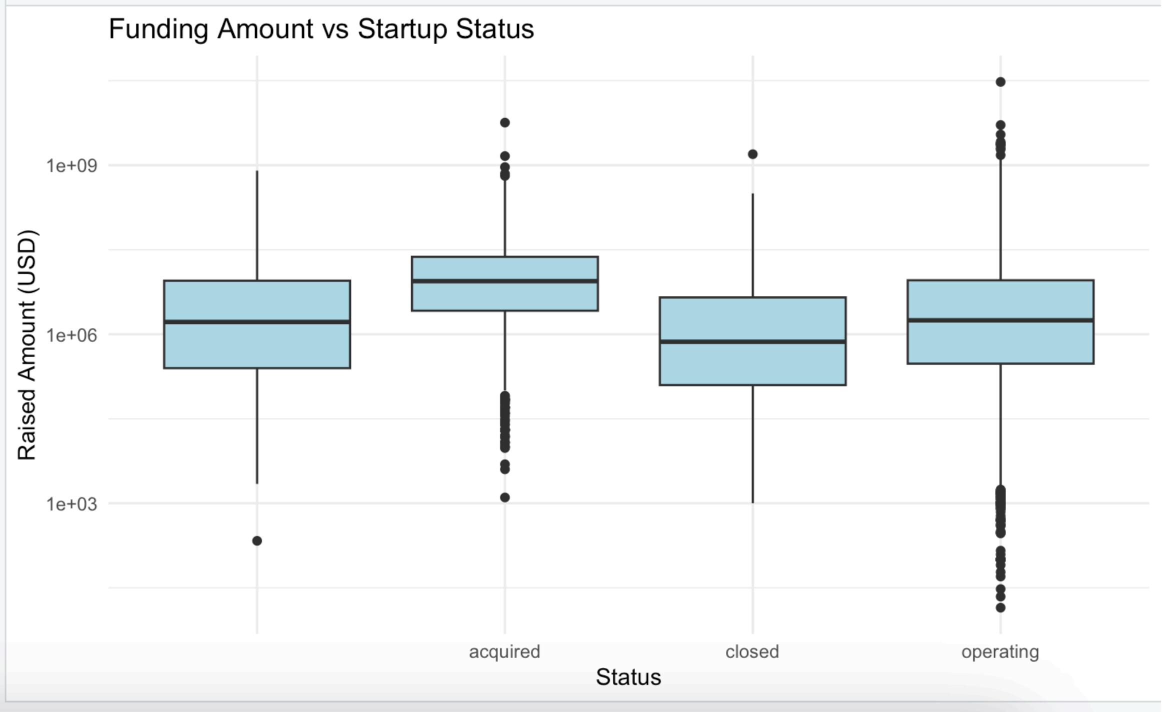
4. Top 10 Countries for Startups

A bar chart displayed the countries with the highest number of startups, revealing geographic hotspots for entrepreneurship.



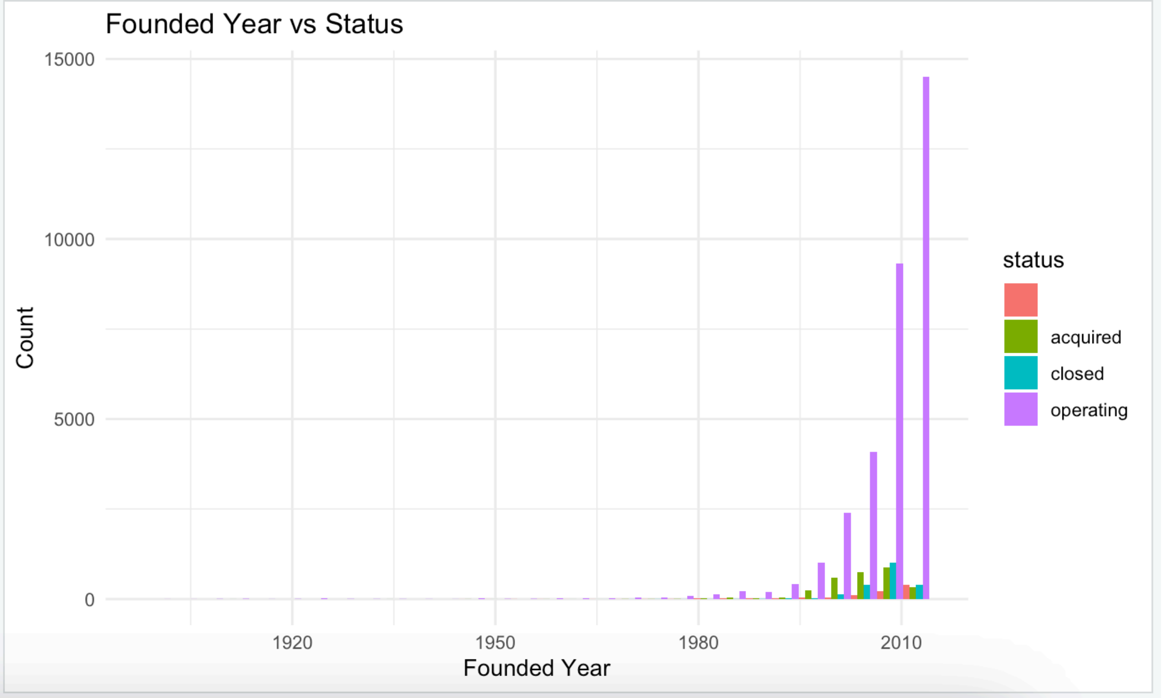
5. Funding Amount vs. Startup Status

A boxplot showed the relationship between funding amount and startup status, revealing whether higher funding correlates with operational status.



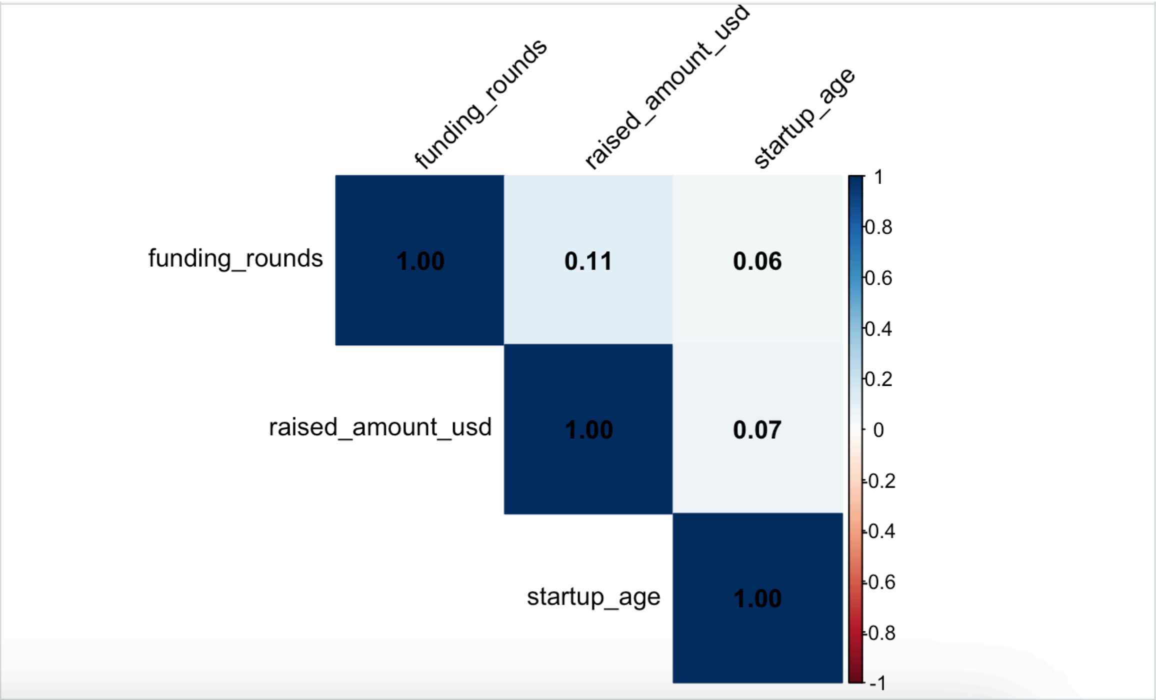
6. Founded Year vs. Status

A histogram displayed the distribution of founding years colored by status, showing if more recent startups have different success rates.



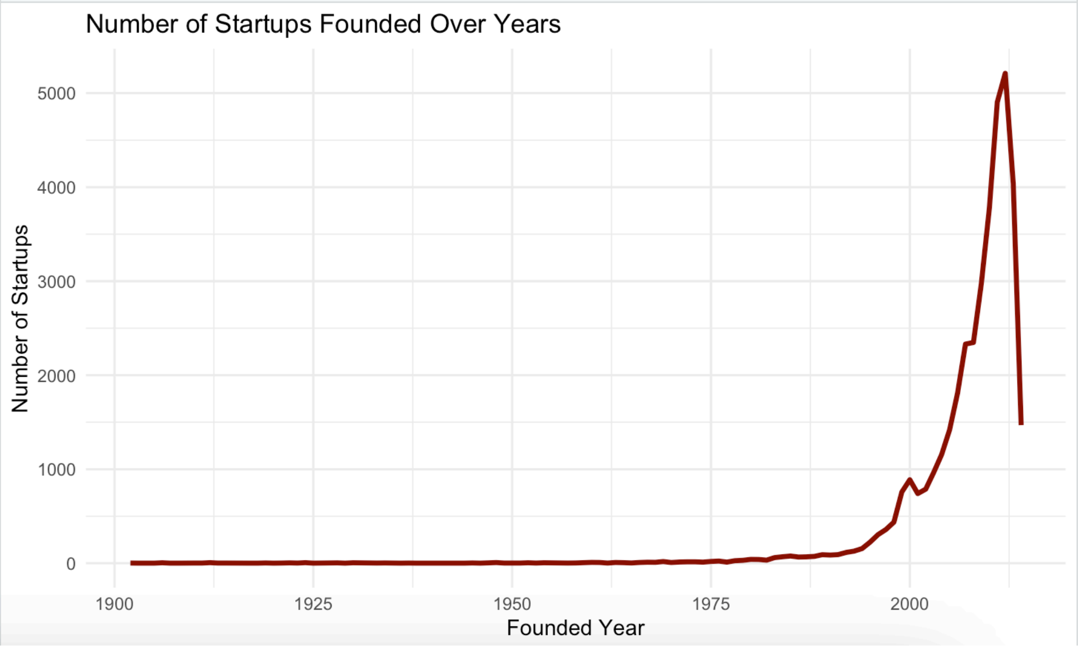
7. Correlation Matrix

A correlation heatmap was created to analyze the relationships between various numeric features in the dataset, such as funding rounds, raised amount, and startup age.



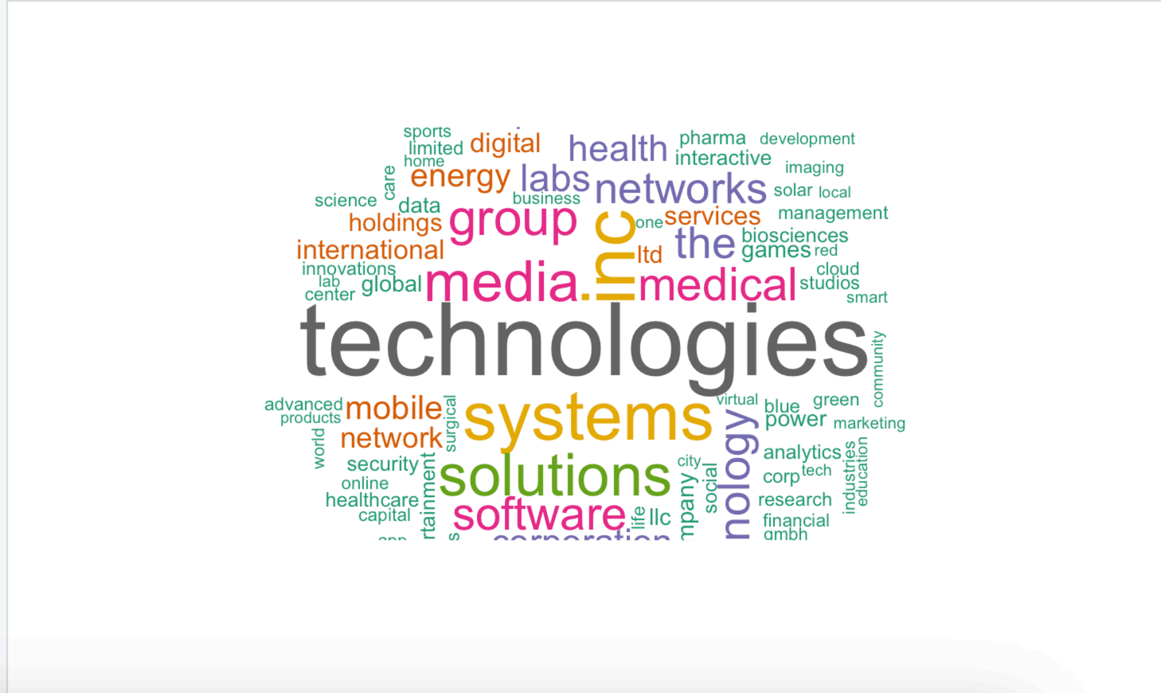
8. Startup Formation Trend

A line plot showed the number of startups founded over the years, revealing industry growth trends.



9. Startup Name WordCloud

A word cloud visualization of startup names was generated to identify common naming patterns and trends.



5.3. Insights from EDA

1. **Status Distribution**: The majority of startups in the dataset are still operational, with a significant portion having closed.

2. **Funding Distribution**: Venture funding follows a highly skewed distribution with most startups receiving modest amounts and a few securing very large investments.

3. **Industry Analysis**: Certain categories like software, e-commerce, and biotech dominate the startup landscape.

4. **Geographic Distribution**: The US, followed by countries like India, UK, and China, hosts the highest number of startups.

5. **Funding Impact**: Higher funding generally correlates with operational status, suggesting adequate capital is important for survival.

6. **Temporal Trends**: There has been a steady increase in startup formation over the years, with a potential slowdown in very recent years.

7. **Correlation Insights: The** number of funding rounds and the total amount raised show a strong positive correlation.

6. Predictive Analytics

Two machine learning models were developed to predict startup success:

Logistic Regression:

- Predicts the probability of a startup being operational based on funding rounds, raised amount, and age.
- The model's performance was evaluated using a confusion matrix.

Support Vector Machine (SVM):

- A more complex model for startup success prediction.
- Slightly outperformed logistic regression in classification accuracy.

7. Clustering Analysis

K-means clustering was performed to segment startups into distinct groups:

Optimal Cluster Determination:

- The elbow method was used to determine the optimal number of clusters (3).

Cluster Profiling:

- Three distinct startup profiles were identified based on funding, age, and success metrics.
- These clusters represent different startup growth trajectories and risk profiles.

8. Conclusion

The EDA on the venture capital investment data reveals several trends and patterns:

- **Funding and Success**: Startups with higher funding and moderate age have a higher survival chance.
- **Industry Impact**: Certain industries consistently attract more investment and show higher success rates.
- **Geographic Factors**: Startup ecosystems are concentrated in specific global locations.
- **Predictive Insights**: Both funding metrics and startup age are strong predictors of success.
- **Distinct Profiles**: The clustering analysis revealed different funding-age-success profiles among startups.

These insights can help investors make more informed decisions during the funding process, especially when considering industry trends, startup age, and funding history. Entrepreneurs can also benefit by understanding the factors that contribute to startup longevity and success.

9. GitHub Repository

https://github.com/mintu3770/startups