

Start-Up Analysis​

**1. Introduction and Research Questions**

Startups play a crucial role in driving innovation and economic growth and have the potential to create significant value for investors and society as a whole. However, launching and scaling a startup is a challenging and risky undertaking, and many startups fail in their early stages. Understanding the factors that contribute to startup success or failure is therefore of great interest to entrepreneurs, investors, and researchers.

The Big Startup Success/Fail Dataset is a comprehensive and informative dataset that provides a wealth of data on startups that have either succeeded or failed in the past. The dataset was obtained from Kaggle, a platform for sharing and discovering datasets and other resources related to data science and machine learning.

This project aims to analyze the Big Startup Success/Fail Dataset to gain insights into the factors that contribute to startup success or failure. Specifically, we aim to answer the following research questions:

1. What are the key factors that contribute to startup success or failure, such as funding rounds, founding year, industry, location, and other key metrics?
2. How does the impact of these factors vary across different industries and locations?

To answer these questions, we will use a variety of statistical and machine learning techniques, including data cleaning, exploratory data analysis, regression analysis, and classification algorithms.

The motivation for this project is to provide valuable insights for entrepreneurs, investors, and policymakers seeking to improve the success rate of startups. By analyzing this dataset, we can identify patterns and trends that can inform strategic decision-making, such as identifying promising industries or geographic regions for startups or understanding the importance of certain factors such as funding and founding year.

In conclusion, this project seeks to leverage the power of analytics to study a comprehensive and informative dataset on startup success and failure. By answering our research questions, we aim to provide valuable insights for entrepreneurs, investors, and policymakers seeking to improve the success rate of startups and drive innovation and economic growth.

**2. Data Collection**

The data was mined and engineered from various sources and file formats. They were then extracted and transformed to the format of our requirement. Transformation of data included Data Cleaning. Steps like Removing Outliers, Identifying Missing Values, dropping missing values, imputing missing values and Interpolation were followed to ensure Data Integrity. The data dictionary provided a comprehensive guide that outlined the metadata including data types, data constraints, relationships between data elements, and other relevant information. Our data dictionary contains information like Variable and Description which provides us with information about the predictors available and their nature.

Data Sample:

A screenshot of a computer

Description automatically generated with low confidence

Data Dictionary:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| permalink | Link to Organization |
| name | Name of the Startup |
| homepage\_url | Website URL of the Startup |
| category\_list | Field of the company |
| funding\_total\_usd | Total funding compensation received |
| status | Operational status |
| country\_code | The country of Origin |
| state\_code | The state at which the company is located |
| region | Region of company location |
| city | City of company location |
| funding\_rounds | The number of rounds company went for funding |
| founded\_at | When the company was established |
| first\_funding\_at | Date at which the company was first funded |
| last\_funding\_at | The last time the company was funded |

**3. Data Preprocessing**

The dataset, which was obtained from various sources, contains 66368 rows and 14 columns. Real-world datasets are often imperfect, containing missing values, incorrect data types, unreadable characters, and unexpected values. Poor data quality can manifest in various forms, such as empty cells, data in the wrong format, outliers, and duplicates. Empty cells refer to data points that are not present in the dataset, typically represented as NaN (Not a Number) in Pandas. These missing values can be handled by dropping them, filling them with a constant value, or interpolating them to estimate the missing value. All the NaN and NULL values have been removed.

Data Summary:

A picture containing shape

Description automatically generated

Graphical user interface, table

Description automatically generated

Above, we replaced all the place holders with NULL values.

Table

Description automatically generated

In the above screenshot, we replaced all the null values with zero or considered as other categories.

Text

Description automatically generated

Changing the data types accordingly.

Number of Rows after cleaning: **51119**

Number of columns after cleaning: **13**

**4. Data Analysis**

The final dataset has over 13 columns and 50 thousand + rows. The bar graph shows us that there are still 40,000 companies still under operation of all the companies listed. The boxplot gives us a representation of what the status of the startups are currently and we can see that most of them converted into IPOs. The correlation heatmap gives us the highest correlation between two variables. The tableau dashboards give us a clear representation that each region specialises in different domains of startups and the possibility of success depends majorly on the location. Hence, geospatial plotting would be ideal for the use case.

Chart, bar chart

Description automatically generated Chart, box and whisker chart

Description automatically generated

Chart, treemap chart

Description automatically generated

**5. Limitations and Future Work**

Limitations:

**Data availability**: One of the major limitations in analyzing startup companies over funding, category, and geography is the availability of accurate and up-to-date data. While there are various sources of information such as Crunchbase, AngelList, and Pitchbook, these databases can have incomplete, inconsistent, or outdated information.

**Data quality:** Even when data is available, the quality of the data can vary greatly, and it can be difficult to verify the accuracy of the information. This can lead to incorrect conclusions and inaccurate analysis.

**Limited scope**: The analysis of startup companies over funding, category, and geography is limited to the data that is available. It may not capture important factors such as the quality of the management team, the strength of the product, or the competitive landscape.

**Future Work:**

Text

Description automatically generated

Analyzing additional attributes like government support and geographical location can provide valuable insights for startups. Here are some potential areas of focus for future analysis:

**Government support**: Examining the level of support that startups receive from their government can provide insight into the overall business climate and regulatory environment. This might include looking at factors such as tax incentives, grants, and funding opportunities specifically targeted at startups.

**Geographical location**: The location of a startup can have a significant impact on its success. Factors such as access to talent, infrastructure, and markets can all be affected by the company's location. In addition, different regions may have different regulatory environments or cultural attitudes towards entrepreneurship that can affect a startup's chances of success.

**Industry-specific factors:** Depending on the industry in which a startup operates, there may be additional factors that are particularly relevant to its success. For example, in the tech industry, access to venture capital funding and the availability of technical talent may be particularly important.

**Founder experience**: The experience and background of the startup's founders can also be an important factor in its success. For example, founders with previous startup experience or experience in the industry in which they are launching their business may be better equipped to navigate the challenges of starting a new venture.

**Market size and competition**: Analyzing the size of the market and the level of competition can help to identify potential opportunities and challenges for startups. This might include examining factors such as market growth rates, barriers to entry, and the competitive landscape.

By considering these additional attributes, it may be possible to gain a more comprehensive understanding of the factors that contribute to startup success or failure. However, it's important to note that there is no one-size-fits-all approach to analyzing startups, and different startups may face unique challenges and opportunities based on their individual circumstances.