Startup Investments Dataset Analysis

Thiago Lyra Ganem, Asha Anthony Ollukaran, Thalvaner Henrique Kuntze  
  
CSIS Department  
Douglas College  
New Westminster, Canada

# Introduction

The Startup Investment dataset provides a diverse range of information about Startups around the world. It contains financial information such as funding history, as well as general information such as their geographical location, industry and its success status. Therefore, analyzing this dataset and understanding its trends could provide valuable information about the Startup sector.

# Dataset Definition

## General Information **Dataset Name:**StartUp Investments **Link:** <https://www.kaggle.com/datasets/arindam235/startup-investments-crunchbase> **Number of Features:**40 **Dataset Header: Figure1 – Dataset Head showing first 10 rows** A screenshot of a computer Description automatically generated **Important Feature Descriptions:**

1. *Category List: A list of categories or industries that the startup is associated with.*
2. *Market: The market or industry segment that the startup operates in.*
3. *Funding Total (USD): The total amount of funding received by the startup in US dollars.*
4. *Status: The current status or stage of the startup (e.g., operational, closed, acquired, etc.).*
5. *Region: The broader geographical region where the startup is located.*
6. *City: The city where the startup is headquartered.*
7. *Funding Rounds: The number of funding rounds the startup has gone through.*
8. *Founded At: The date when the startup was founded.*
9. *Founded Month: The month in which the startup was founded.*
10. *Founded Quarter: The quarter in which the startup was founded.*
11. *Founded Year: The year when the startup was founded.*
12. *First Funding At: The date of the first funding round for the startup.*
13. *Last Funding At: The date of the most recent funding round for the startup.*
14. *Seed, Venture, Equity Crowdfunding, Undisclosed, Convertible Note, Debt Financing, Angel, Grant, Private Equity, Post-IPO Equity, Post-IPO Debt, Secondary Market, Product Crowdfunding, Round A, Round B, Round C, Round D, Round E, Round F, Round G, Round H: These columns seem to represent different types of funding rounds and the corresponding funding amounts for each round.*
15. *Success: A binary indicator (e.g., 0 or 1) representing whether the startup was successful or not.*

# Data Mining

## Initial Analysis

The dataset under examination offers comprehensive information about startups established between 1897 and 2014 accoss different countries, and sectors. By analyzing the first 10 rows of the dataset, important features such as Total Funding, Category, Country Code, Funding History and Success stand among others.

Therefore, we could try to understand important questions:

### Which countries are the best ones to host a Startup?

### Which sector are the ones receiving more funds?

### How long does it take to receive the first funding after the company is founded?

### What is the current trend on the investment?

### How many startups succeed and how many close?

### Due to the size of the dataset and the range variety of countries and sector, we will focus the analysis only on the top 10 countries and the top 10 sectors of the dataset.

## Treating the Dataframe

By checking the data frame an initial treatment must be made as the feature “funding\_total\_usd” is being set as a String. Therefore, it was cast as numeric, trimmed and also had both commas (“,”) and dollar sign (“$”) removed.

A white rectangular object with red text

Description automatically generated



Then, to check the top 10 markets and top 10 Countries two additional data frames were created from the original one.

A white background with blue text

Description automatically generated

A white background with black text

Description automatically generated

Afterwards, we filter the data frame that will be used in this analysis only contains the top 10 market and top 10 countries.

A screen shot of a computer code

Description automatically generated

## Understing the Top 10 Markets and Countries

To begin our analysis, a heatmap plot using Seaborn was produced for a better visualization showing the quantities of Startup per country and region.



A white background with black and red text

Description automatically generated

**Figure 2 – Heatmap Chart Number of Startups per Market and Country**

A screenshot of a graph

Description automatically generated

The heatmap chart offers a representation of the top 10 global startup market, and key factors are evident:

* The US is the leader in the startup market with the highest concentration of startups. Besides   
  Easter European countries, India, Russia, Israel and China also display significative startup activity. However, none of them comes close to the density observed in the USA.
* The sectors of Software, Biotechnology and mobile, dominate the market globally across all the countries analyzed.

Therefore, we can assume that the US is the country that has the most favorable conditions to create and manage a Startup.

## Analyzing Investments made by sector

To understand which markets are more propense to receive funding, a box plot showing the median and their variations is the most appropriate chart to answer this question.

A screen shot of a computer code

Description automatically generated

**Figure 3 – Boxplot of top 10 Markets vs Total Funding in USD original Scale**

A graph of different colored rectangular bars

Description automatically generated

Since the range of Fundings in USD were extensive, a box plot with original scale wasn’t going to allow a visualization. Thus, a transformation was necessary so a more comprehensible view of the funding distribution could be analyzed.

A screenshot of a computer program

Description automatically generated

**Figure 4 – Boxplot of top 10 Markets vs Total Funding in USD Log Scale**

A graph with different colored squares

Description automatically generated

By analyzing both charts, it is possible to visualize that the sectors of Clean Technology, Health Care and Biotechnology have a wide range of variation in their median. On the other hand, Software and Biotechnology sectors exhibited a higher number of outliers indicating that several startups in these areas secured an exceptionally high funding compared to the other sectors.

## Analyzing the time to receive first fuding.

In order to check how long does it usually takes to receive the first funding after the company was founded, a line chart with both features would be appropriate, this way we can verify the gap between the years.

A screenshot of a computer program

Description automatically generated

**Figure 5 – Lineplot of Startup Founded and First Funding per Year**

A graph with a line graph and numbers

Description automatically generated

By analyzing this chart, we can see that the number of startups and the number of Startups that were founded has increased in a linear way from 1995 until 2012, but had a strong decrease afterwards. On the other hand, the boom in investment started in 2004 and even though it increased in a linear way, it was faster than and never declined.

Also, it is possible to visualize that the difference between these two lines shows that there was a major delay since the startup creation and when they received their funding.

Besides, even though the number of new startups is declining, the number of fundings is increasing. Therefore, we can assume that the market is more propense to provide funding to existing companies rather than investing in new ones.

## Understading Funding Trend

To understand the trend in funding the Histogram is a good choice as we could easily represent the total amount of funding from 1995 until 2014, considering that the funding was all invested in the first round of funding made by each startup.

To create the histogram, we had to isolate a new data frame grouping the total amount invested by the date of the first funding.

A screenshot of a computer code

Description automatically generated

**Figure 6 – Histogram of Total Funded per Year**

A graph of blue bars

Description automatically generated

It is possible to notice that there was a big leap in investments from 2004 and 2005, and that afterwards a trend in investment was consistent until 2014 except for the years of 2009 and 2010.

This spike was however caused by investments made in big companies such as Verizon Communication, Ubber, AirBnB Pinterest, Lyft ,among others significant companies that thrived.

## Analyzing the success of Startups

The success of a Startup in this dataset is measured if the company worked or if it ceased its operations. In order to analyze it, a scatter plot showing how many company that was founded in a given year were closed or are still doing business can provide the trend among the market over the years.

To Create the scatter plot, first we had to create a new dataframe based with year and the count of success and closed companies. For that we used the following code.

A screenshot of a computer code

Description automatically generated

A computer screen shot of a computer code

Description automatically generated

**Figure 7 – Scatter Plot Displaying the number of startups that succeed vs the ones that shut down**

A graph with blue and red dots

Description automatically generated

The chart provides insights about a comparison on how many companies that were founded in a given year were still operating or were shut down. It is possible to evaluate that most companies thrived while a few ceased operations. However, a drastic reduction on new startups after 2012 happened, which may have happened due to uncertain to the market in the upcoming years.