📈 Analysis of Venture Capital Investments

A Comprehensive Examination of The Intricate Venture Capital Investments Landscape

Author: Bhalisa Sodo
Code: GitHub

Introduction

Venture capital investments play a pivotal role in fostering innovation and driving economic growth. This report delves into the intricate landscape of venture capital investments, with a focus on understanding the factors that contribute to the success of startups. By examining the intricate dynamics between founders, investors, and the broader ecosystem, we aim to unravel the complexities that shape the trajectory of these ventures.

The report commences by presenting a comprehensive visual analysis, offering insights into the characteristics and profiles of founders and investors. This visual exploration serves as a foundation for comprehending the multifaceted nature of the venture capital landscape, illuminating the intricate relationships and patterns that underlie this domain.

Subsequently, we embark on an in-depth investigation of the start-up and funding landscape, dissecting the various variables and elements that influence the performance and outcomes of these ventures. This analysis provides a holistic understanding of the factors that contribute to the success or failure of startups, enabling us to identify the critical determinants that drive their growth trajectories.

Leveraging the insights gained from this comprehensive analysis, our goal is to develop a robust predictive model that accurately forecasts the success metrics of startups. By harnessing the power of data-driven approaches and machine learning techniques, we strive to create a tool that can assist stakeholders in making informed decisions, optimizing resource allocation, and maximizing the potential for successful ventures.

Through this report, we aim to shed light on the intricate dynamics of the venture capital ecosystem, empowering founders, investors, and policymakers with valuable insights and actionable recommendations. By fostering a deeper understanding of the factors that shape start-up success, we contribute to the collective effort of nurturing innovation, driving economic growth, and unlocking the full potential of entrepreneurial endeavours.

Key Findings

1. Founder Profiles: The report visualizes founder profiles, highlighting the prevalence of for-profit companies and the dominance of Caucasian males in the technology venture capital landscape.

2. Investments & Portfolio Companies: Notable founders like Anne Wojcicki and Naval Ravikant lead in investments, with a new focus on AI start-ups and a trend towards more investments in this sector.

3. Investment Categories: Both male and female founders generally invest in similar types of start-ups, with slight variations such as females showing interest in Fashion and Education, while males focus on Analytics and Information Technology.

4. Lead Investments & Exits: Lead investors play a crucial role in providing initial funding to start-ups, while exits through acquisitions or IPOs are key milestones for investors to realize returns.

5. Founder Education: Stanford University and Harvard Business School are the most attended institutions by founders, with degrees predominantly in Computer Science and Economics.

6. Popularity: Founders like Mark Zuckerberg and Elon Musk have garnered significant media attention due to their companies' impact on society and the tech industry.

7. Start-up Landscape: California and New York host a significant number of start-ups, with California being a hub for technology start-ups due to its access to talent, venture capital, and a supportive ecosystem.

8. Start-up Success Factors: The age of a start-up at the time of its first funding round plays a crucial role in its success, with early funding correlating with higher chances of success and acquisitions typically occurring within the 0-8-year range.

9. Funding Rounds: The average amount raised by start-ups has shown a steady increase over the years, with different types of funding rounds contributing to the growth of these ventures.

10. Model Performance: Machine learning models like Decision Tree Classifier have demonstrated near-perfect performance in classifying start-up success, indicating the potential for accurate predictive models in this domain.

Data Visualisation

Founders

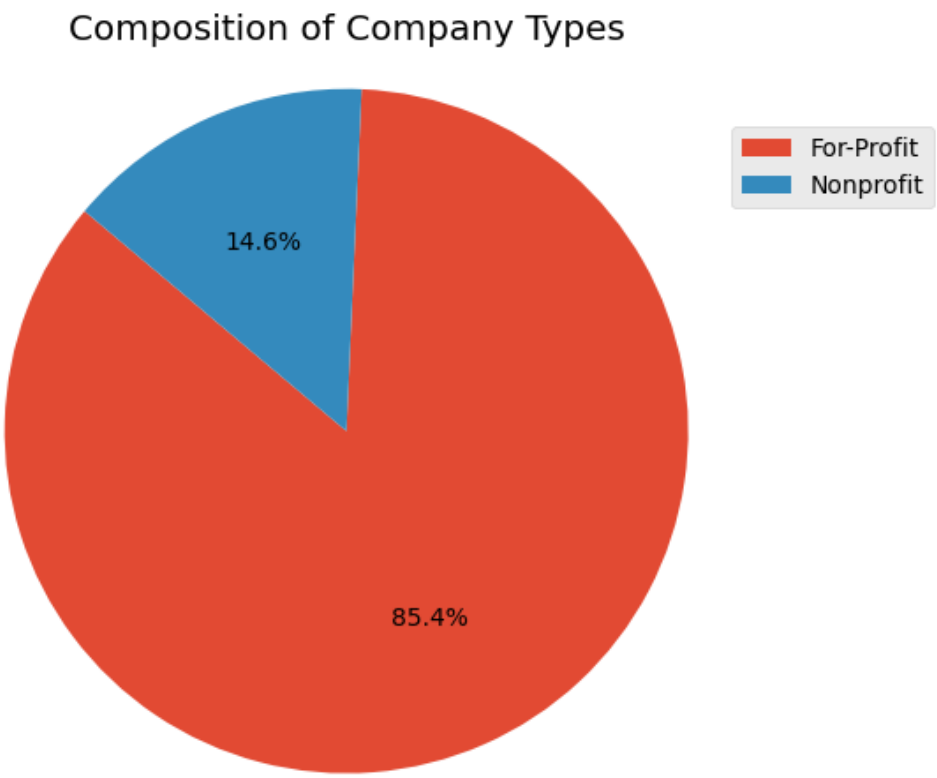
Founder data has a cut-off year of 2017. Meaning some facts may have changed in the 7 years that have passed. I have made the visualisations to get an idea of the founder profiles and to better understand our dataset. I will give updated numbers for some, but not all founders, as we delve into the report. For the convenience of this presentation, I will use the terms ‘founders’ and ‘investors’ interchangeably in this section.

Founder Biographies

Since it is practically impossible for me to read all the biographies of the many start-up founders in the dataset, I decided to create a word cloud to give us an idea of what our founders are made of, so to speak. Shown in the word cloud above are the most prevalent keywords in the bios of the founders in the dataset - the size of a word tells us how frequently a word appears - the bigger the word the more frequently it appears in the bios. For instance, Stanford University is the only alma mater shown in the word cloud, in this context it means the institution has a higher count of the founders that appear in the dataset than any other institution - as we'll later observe in our analysis.

Company Types

For-profit companies dominate the composition of the type of companies founded. This is not surprising as this is a venture capital investment dataset, where maximising shareholder value is of utmost importance.

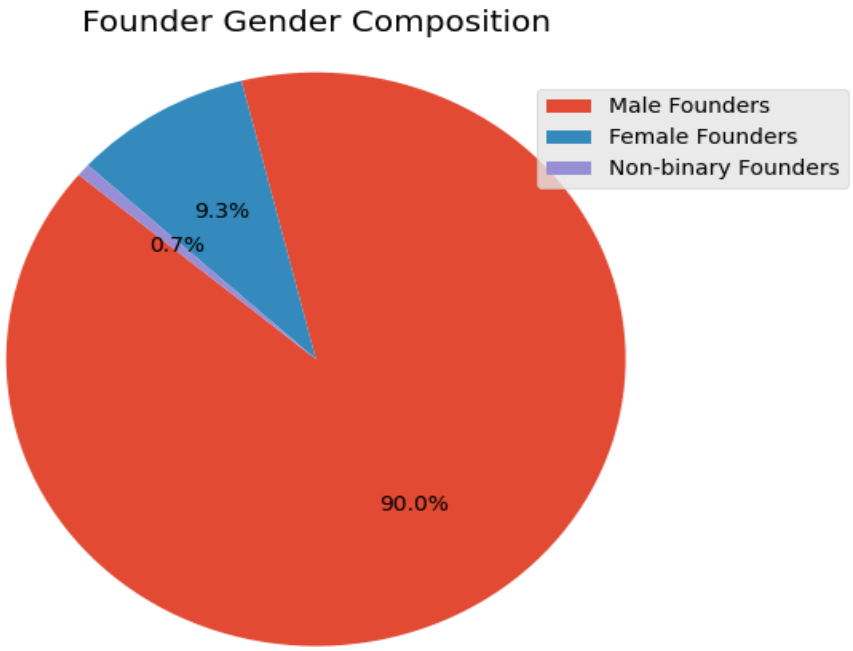


Gender-based Analysis

In my dataset there are three codes that identify a founder's gender; 0, 1 and 2. I assumed that these codes represent female, male and non-binary categories respectively. For the former two (0 and 1) I have cross-referenced the gender with the names, and they seem to check out, indeed these people identify as one of the two genders (female or male). I did further Crunchbase searches of the names I had identified to be non-binary, and they were classified as either female or male. This could mean two things; my assumptions are wrong, or I need to make further investigations. I did the latter. After a Google search & some reading; it seemed safe to assume that the discrepancy arises from the fact Crunchbase's gender classification system is still evolving. While they have made strides in recognising women-led companies, other gender identities (such as non-binary) may not yet be fully captured. So, I am sticking with my initial assumption. But for this portion I will stick to visualisations mostly pertaining to male and female, whom I am confident I have classified with a high degree of accuracy - also to mitigate the risk

of being downright wrong. But all other analyses are inclusive of all the dataset's genders unless stated otherwise.

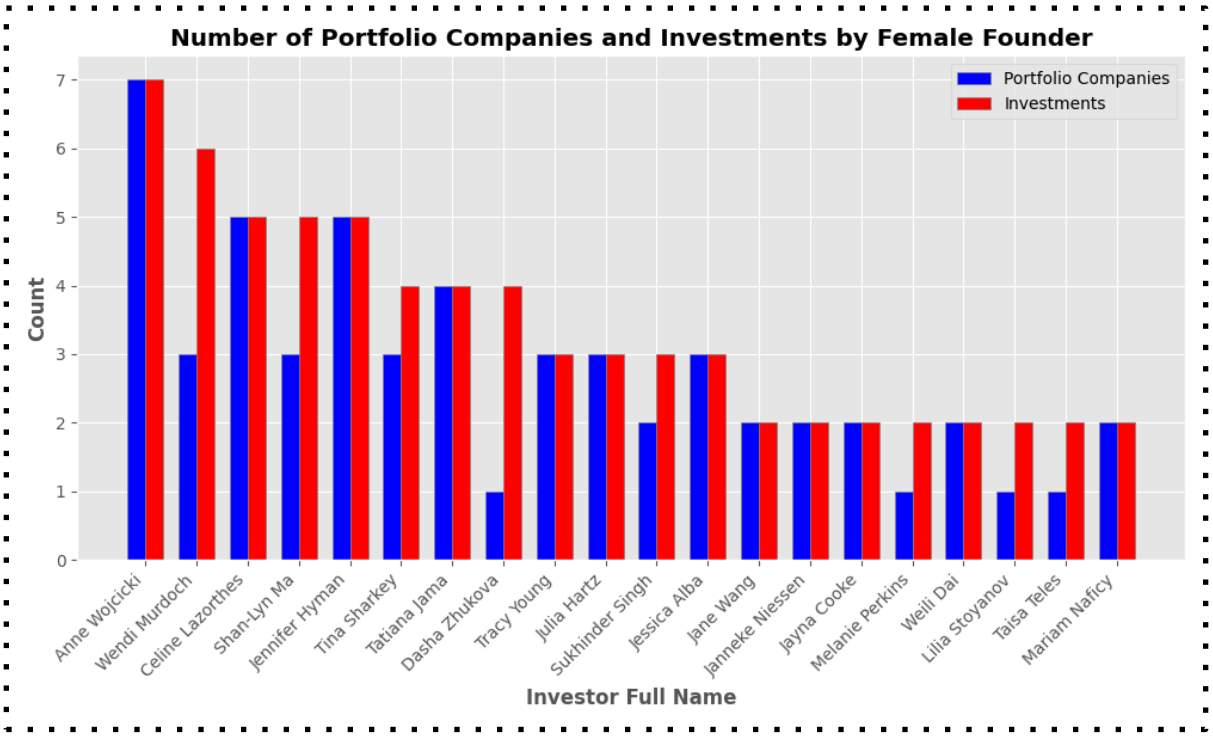
Let's look at the gender composition of the founders.



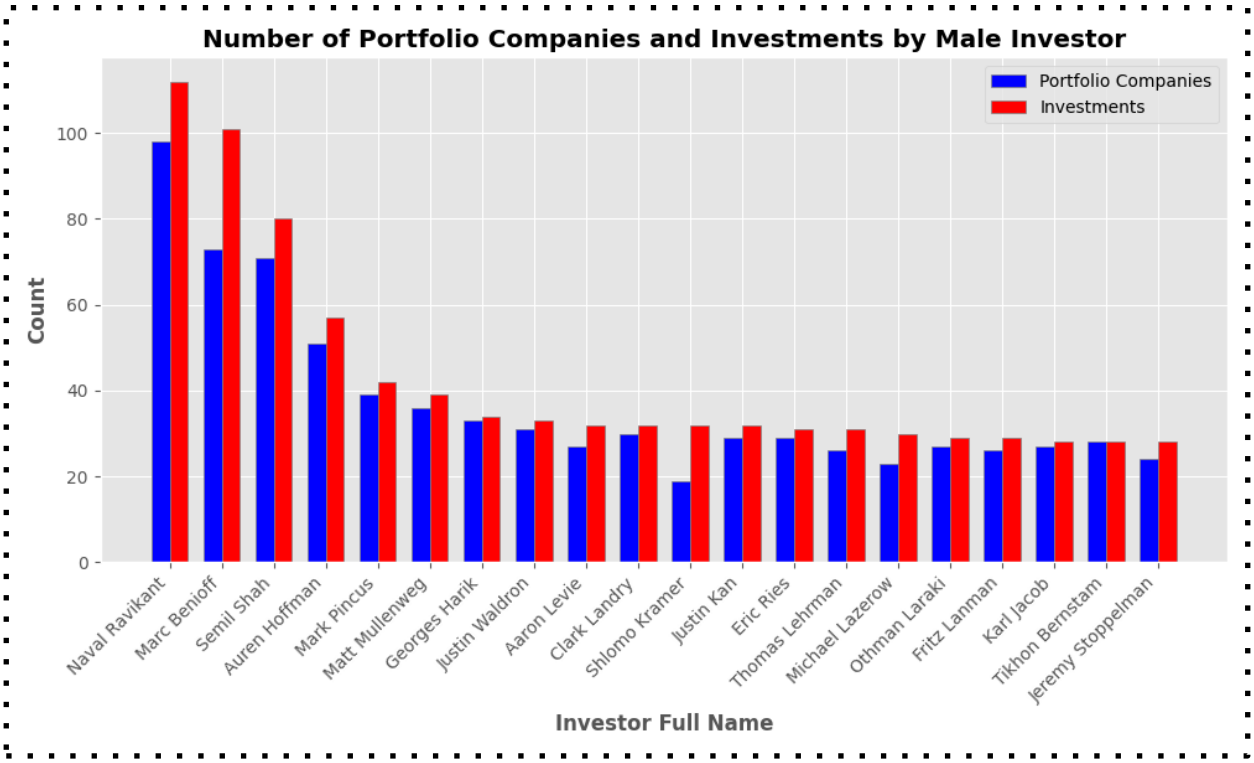
As expected in the technology venture capital investment landscape, males dominate the industry. More specifically Caucasian males. Not much can be said about this chart except the industry needs to become more gender & minority inclusive. I have read companies like Crunchbase have committed to improving representation over the years. They have added race and ethnicity data alongside gender tags, and they are actively exploring additional representation categories. By collecting and sharing this data, they hope to create more opportunities for historically underrepresented founders.

Start-up founders tend to invest in other start-ups, as we will see in this section. For the convenience of this report I will sometimes use the terms 'founder' and 'investor' interchangeably unless explicitly stated otherwise.

Investments & Portfolio Companies



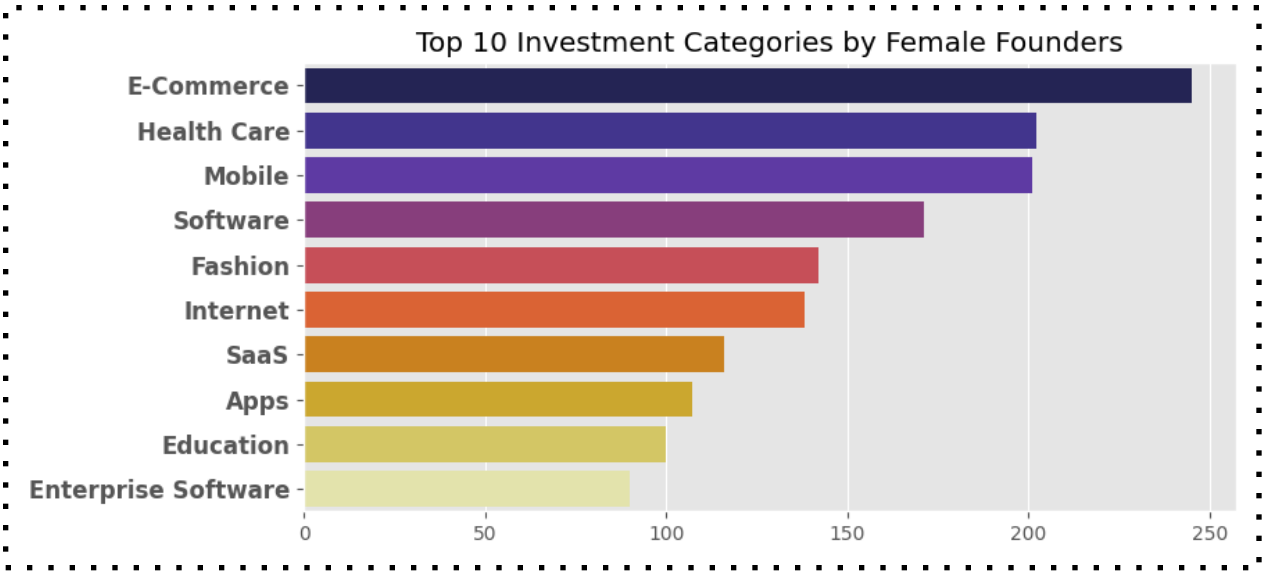
Above are the top 20 founders with primarily the most investments and then portfolio companies, as of 2017. My focus will be on the top 20 founders in each metric. Anne Wojcicki the co-founder and CEO of 23andMe is an angel investor and has since raised her investment count to 50, as of 20 March 2024. In second place we have Wendi Murdoch co-founder of Artsy, who has since raised her investment count to 14 as of 20 March 2024. Her last and most recent investment was in ByteDance, the parent company of TikTok, in 2018.

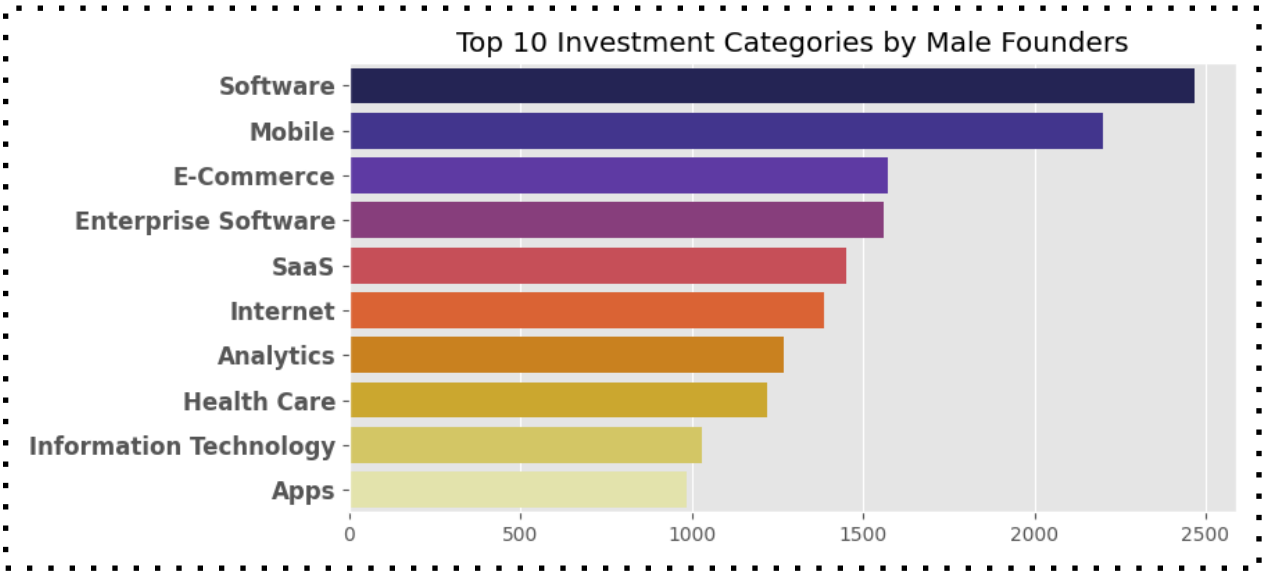


Naval Ravikant, known for his co-founder and former CEO of AngelList roles, is leading the charts, followed by Marc Benioff the CEO of Salesforce and Semil Shah, General Partner at Haystack - a seed investment company. Naval has increased his portfolio companies and investments to 222 and 264 respectively. His latest and most notable investment in 2024, thus far, is in Perplexity AI - an AI-chat-based conversational search engine. Marc increased his to 183 total investments by 2024. We may experience a proliferation of AI start-ups and see more investments directed to these AI start-ups going forward. Thanks to the introduction of and rapid advancements in Large Language Models (LLM), by companies like OpenAI (GPT-4) and open-source implementations like Llama-2 and Mistral, developed by Meta and MistralAI respectively.

Let’s investigate investment categories to uncover if gender plays a role in choosing which start-ups to invest in.

Investment Categories

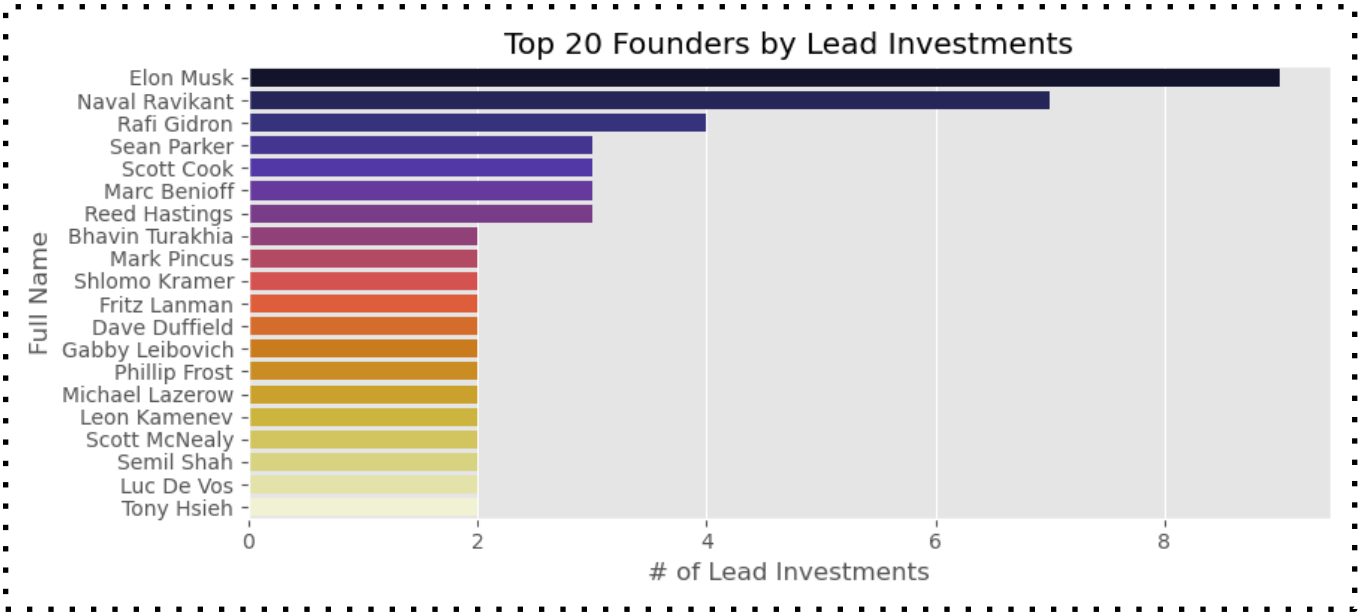




Both genders generally invest in the same kind of start-ups. I was not expecting much of a difference because tech trends are typically the same for everyone. However, it is interesting to note that Fashion and Education are in the top 10 for females, while males have Analytics and Information Technology in their top 10.

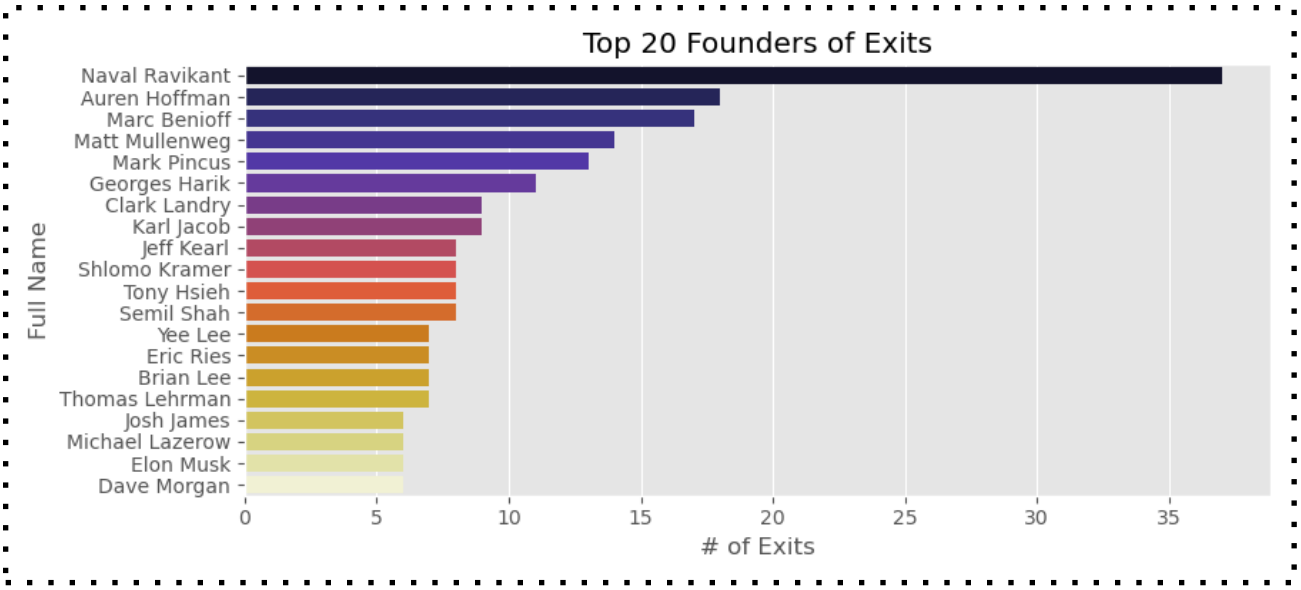
Lead Investments

According to Confluence VC, a lead investment refers to the initial investment made into a company, and the lead investor is the first entity to provide funding to that company. The lead investor serves as the initial endorsement, acting as the primary point of contact between the company and other potential investors during subsequent fundraising rounds. In exchange for their equity stake, lead investors contribute their time, expertise, and professional networks to support the companies they back. A company can have more than one lead investors.



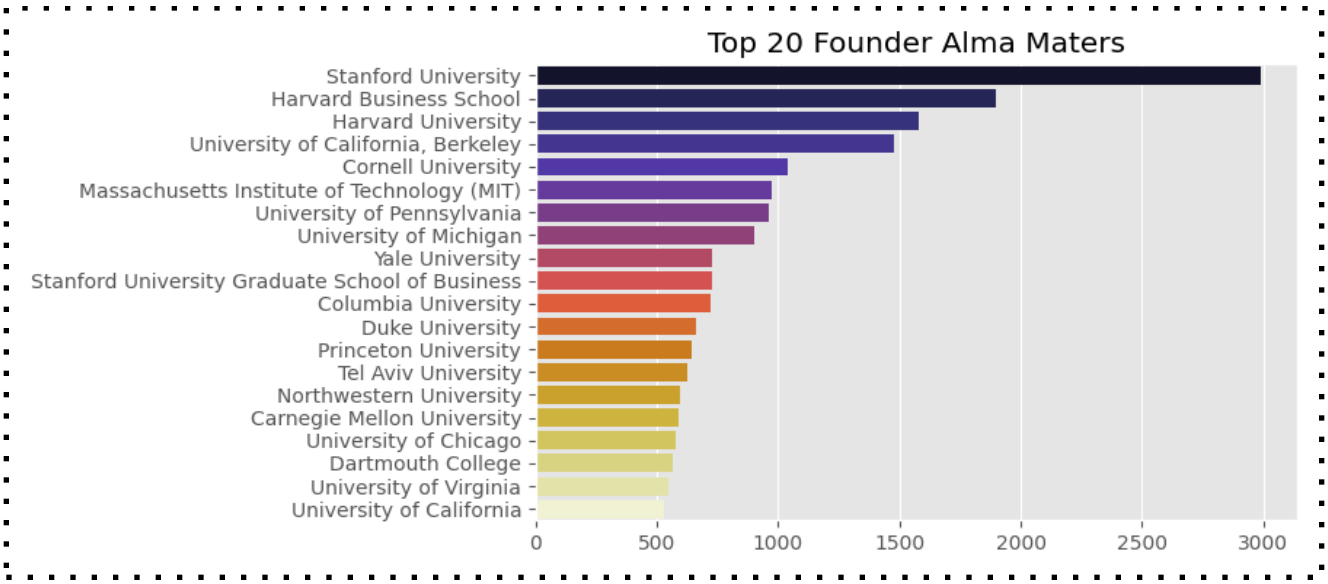
Investment Exits

Investors, often, invest their resources in start-ups in order to earn returns from their initial investment as a start-up scales and increases its value. Exits typically refer to successful liquidity events for investors, such as an acquisition or initial public offering (IPO).



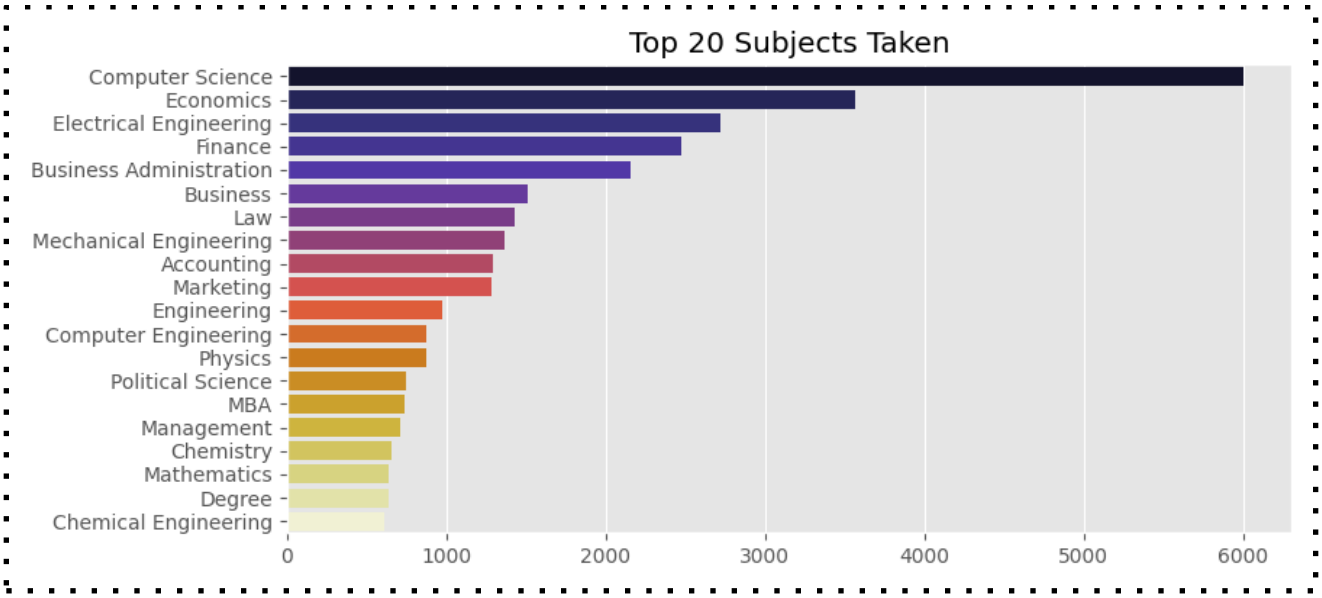
Auren Hoffman has since increased his no. of exits to 68, eclipsing Naval who has increased his to 65, as of March 2024. They are both followed by Marc Benioff with 47 exits.

Founder Education

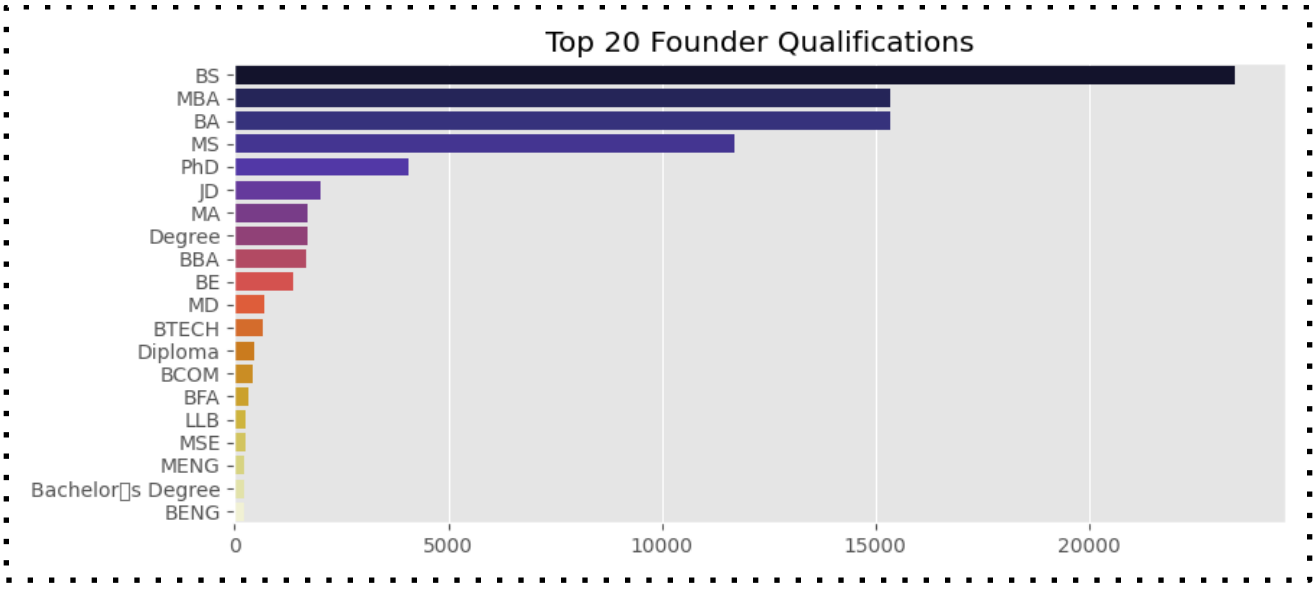


As we previously observed from the biography word cloud, Stanford is the single-most attended university by the founders, followed by Harvard Business School.

Computer Science is the most pursued subject followed by Economics as shown below. This checks out, as most of the top subjects are either technology or commerce related.

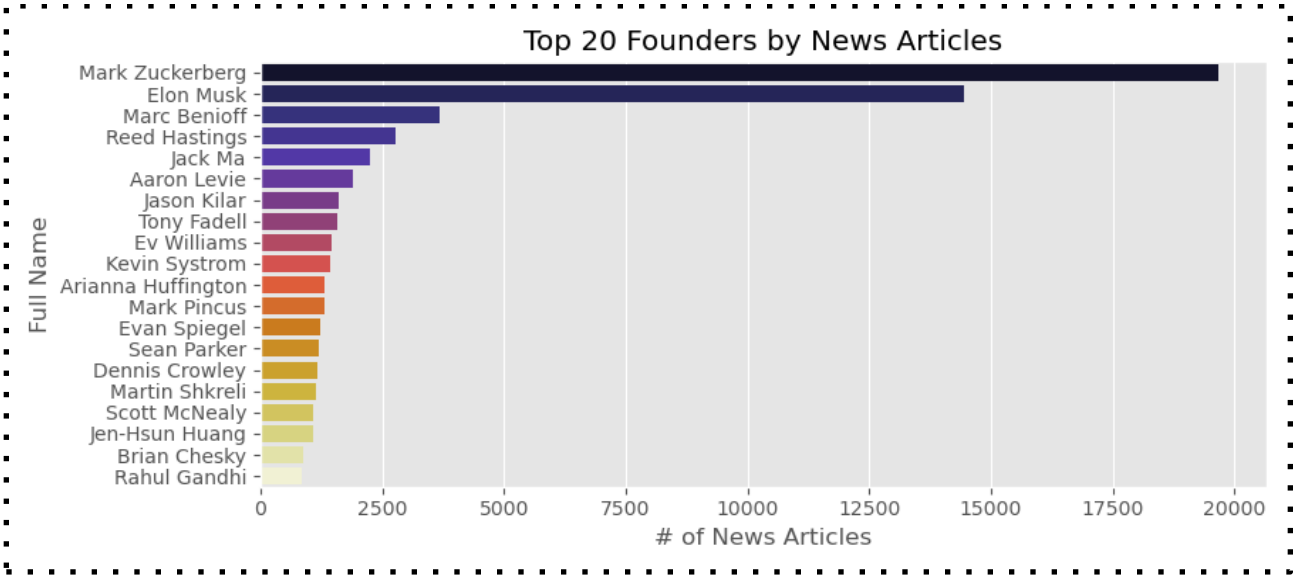


Below we observe degrees are overwhelmingly dominated by Bachelor of Science (BS), followed by Master of Business Administration (MBA) and Bachelor of Arts (BA). In line with our previous observations in alma maters and subjects taken.



Popularity

I have used the no. of article appearances as an indication of a founder's popularity. I thought this would be interesting.



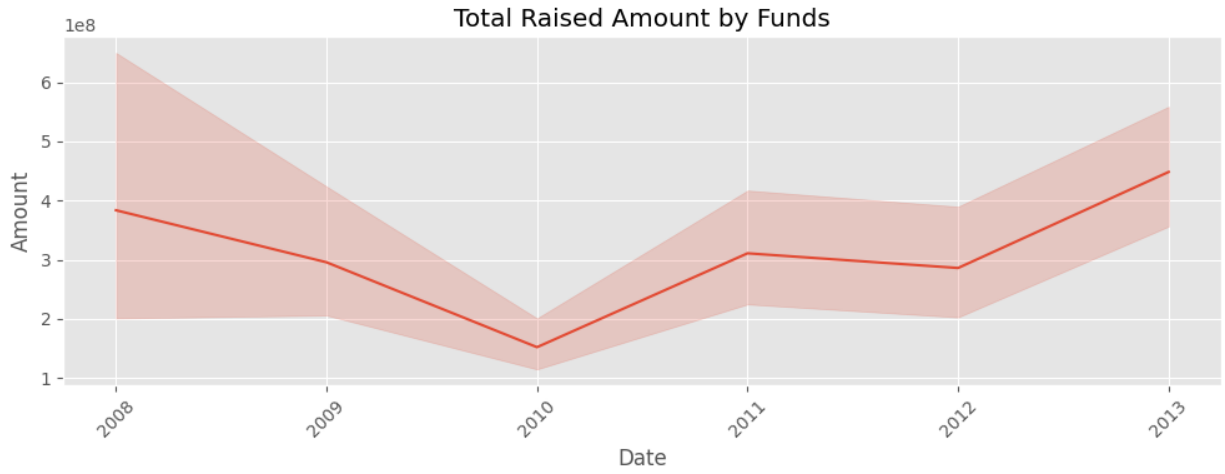
Mark Zuckerberg garnered significant media attention during the period assessed, primarily due to the user privacy concerns surrounding Facebook and the company's high-profile acquisitions of WhatsApp and Instagram. The privacy issues peaked in 2017 and continued to persist in subsequent years, keeping Zuckerberg in the media spotlight.

Elon Musk has arguably become the most widely covered tech founder in recent times, thanks to his entrepreneurial endeavours at SpaceX, Tesla, and Neuralink. Musk's acquisition of Twitter (now known as X) and his connections with OpenAI have further amplified his media presence, making him a frequent subject of daily news coverage.

While the assessment accurately captures the reasons behind Zuckerberg's and Musk's significant media coverage, it's important to note that the popularity of founders can fluctuate over time, subject to changing news cycles, emerging

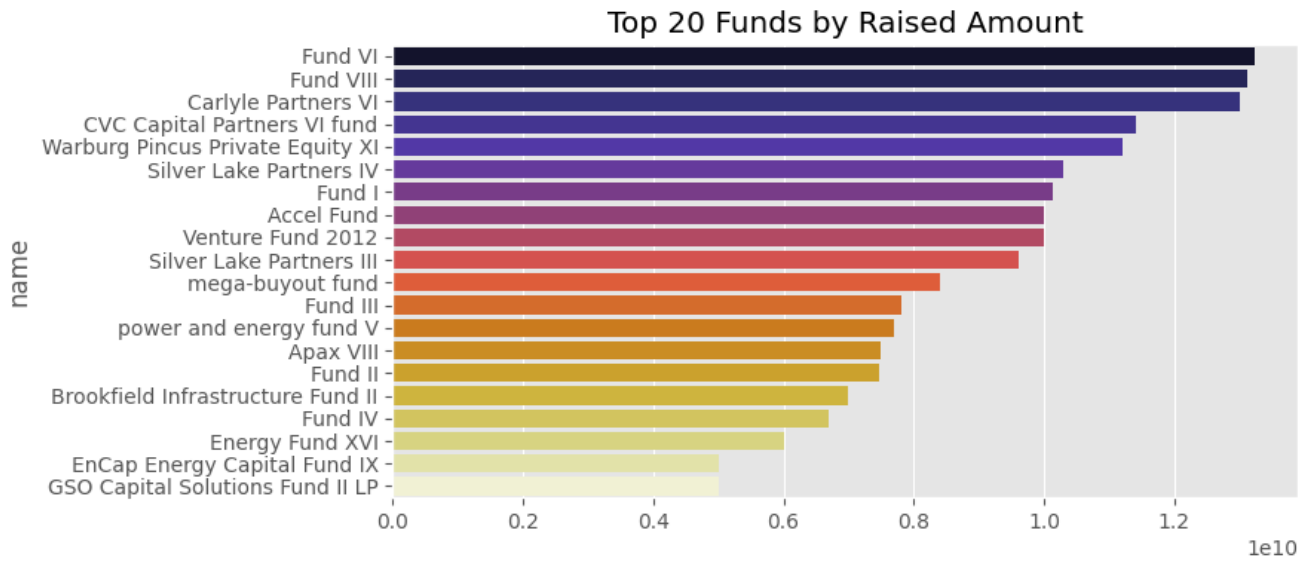
controversies, and the overall impact of their companies on society and the tech industry.

Investment Funds



The graph depicts the total funds raised by investment funds from 2008 to 2013. Initially, there was a sharp decline in funds raised from 2008 to 2010. However, starting from 2010, there was a steady upward trend, indicating a recovery and growth in the raised amounts.

I had to remove a fund name / person called Deric R. Mccloud from the data who had raised what seems to be an anomalous value of $89 billion. I could not find backing for this anywhere on the internet, so I decided to discard the record. Below are the top 20 funds by how much they raised.

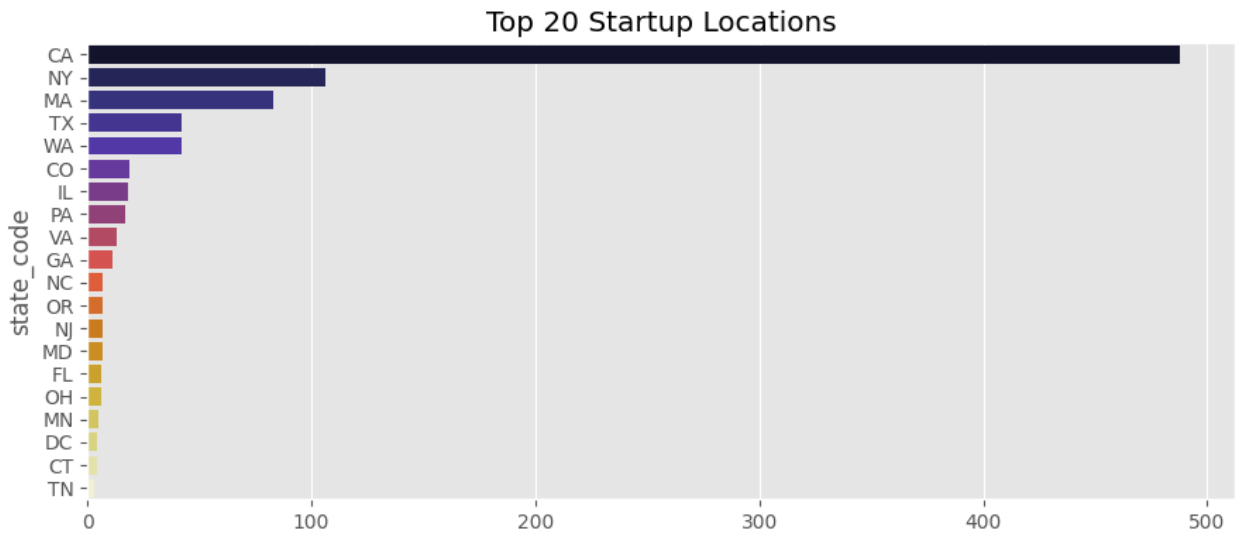


Start-ups

This section explores various characteristics of the ever evolving and complex start-up landscape. From where they are located to their funding and operating status. Which I should say, about the latter, will be used as a start-up success metric later in our analysis.

Start-up Locations

California has the highest number of startups compared to any other state, accounting for 20% of the startups in our dataset. It is followed by New York, which hosts 6% of the total number of startups. These numbers can be attributed to several factors:



1. Presence of major tech hubs: California is home to Silicon Valley, widely regarded as the global epicentre of the technology industry. The San Francisco Bay Area, which encompasses Silicon Valley, provides a conducive ecosystem for start-ups, with access to venture capital, talent, and a well-established entrepreneurial culture. Similarly, New York City is a major hub for various industries, including finance, media, and technology, making it an attractive destination for start-ups.

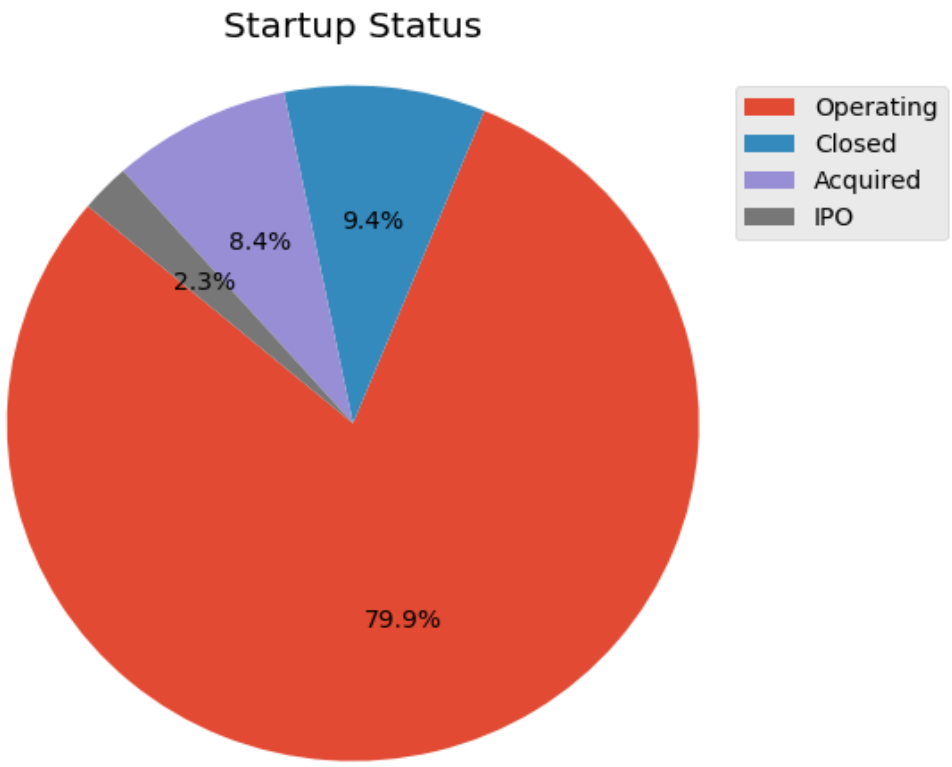
2. Availability of resources and talent: Both California and New York boast a large pool of highly skilled professionals, top-tier universities, and research institutions, which provide a steady supply of talent and innovative ideas for start-ups. Additionally, these states offer robust infrastructure, extensive transportation networks, and a diverse range of support services essential for businesses.

3. Investor confidence and capital access: The concentration of venture capital firms and angel investors in California and New York plays a crucial role in attracting and funding start-ups. These states have a long-standing reputation for successful start-up exits, which further reinforces investor confidence and attracts more capital inflow.

While the exact percentages may vary over time, the combination of these factors has contributed to California and New York consistently ranking among the top states for start-up activity in the United States.

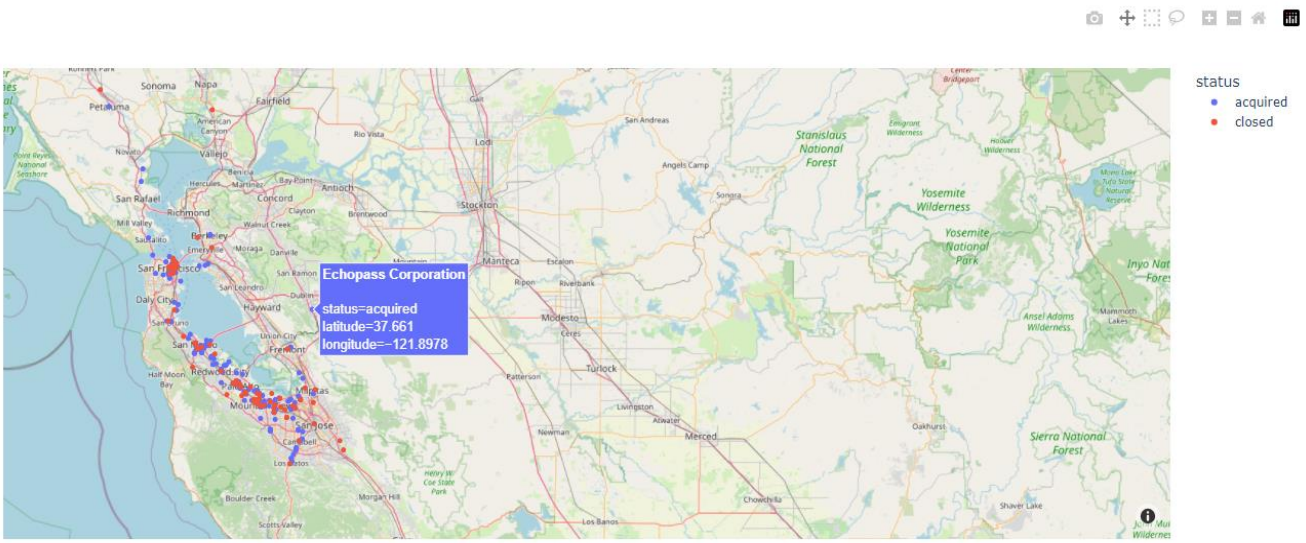
Start-up Status

The below chart shows the status of over 65,000 startups.

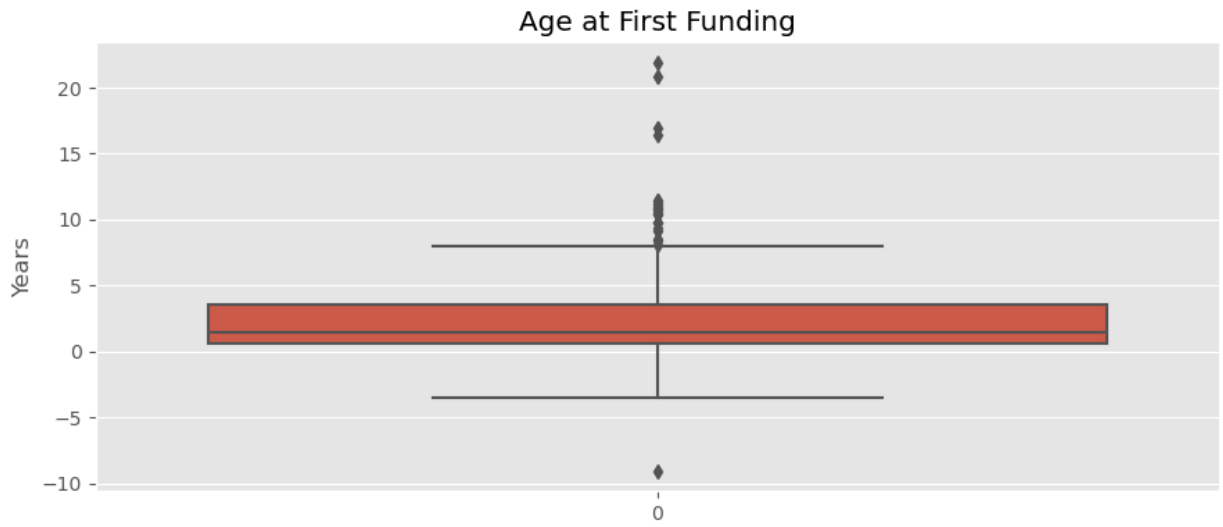


Majority of the start-ups are still operating, meaning they have not yet closed, been acquired or publicly listed (IPO). Almost 10% of the recorded start-ups have closed. This means just over 10% of the start-ups succeeded or got acquired. I will consider an acquisition or IPO as a win for a start-up. Whilst start-ups typically get acquired because of their successes – sometimes it could be because of bankruptcy or some failure. The latter acquisition scenario happens much less often in the start-up ecosystem.

When I make the code available you will be able to interact with the map that shows under 1,000 US-based startups and whether they had been acquired or have since closed. I may deploy it before then and make a link available in a LinkedIn post of the report.

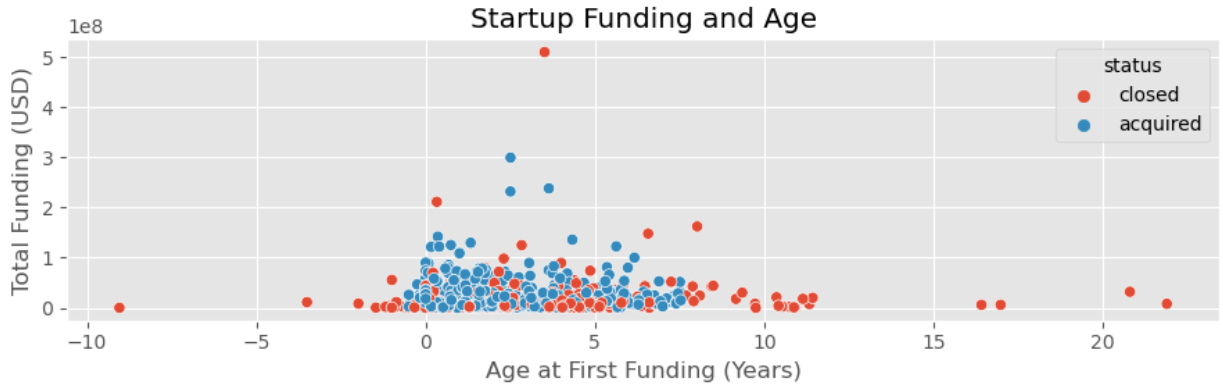


First Funding



The plot shows a distribution of the ages of the start-ups (in years) when they received their first round of funding. On average, the start-ups received their first funding after 2 years and 3 months since they had been operating. There are startups who received their first funding after 10 years. These are rare outlier cases. There are many reasons for this, one of them is bootstrapping. In this case the founders fund the start-up with personal income, loans? Or tap into the so-called 'rich uncle' reservoir.

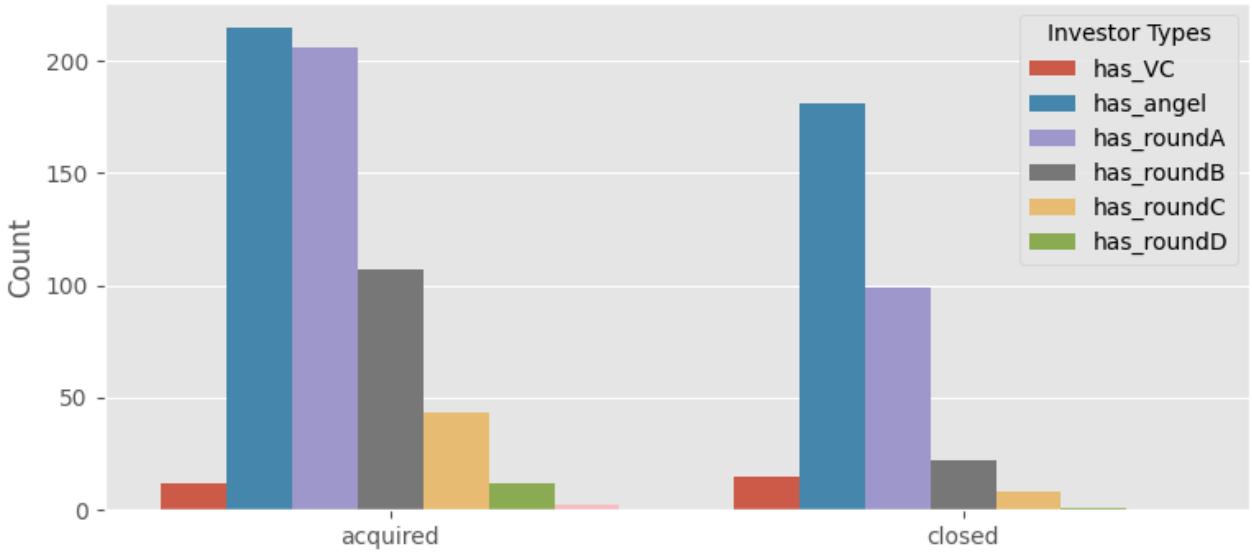
Let's investigate the effect of first-funding age on a start-up's success.



The plot indicates that start-ups who were funded earlier in their start-up's lifecycle, or less than 0 years of operating, have mostly failed and those who

received their first funding too late down the line, ~ after 9 years or later tend to have closed. The period that has the highest number of acquisitions is 0 – 8 years, where we also observe the highest number of acquisitions. But as the years go by lesser-and-lesser startups get funded.

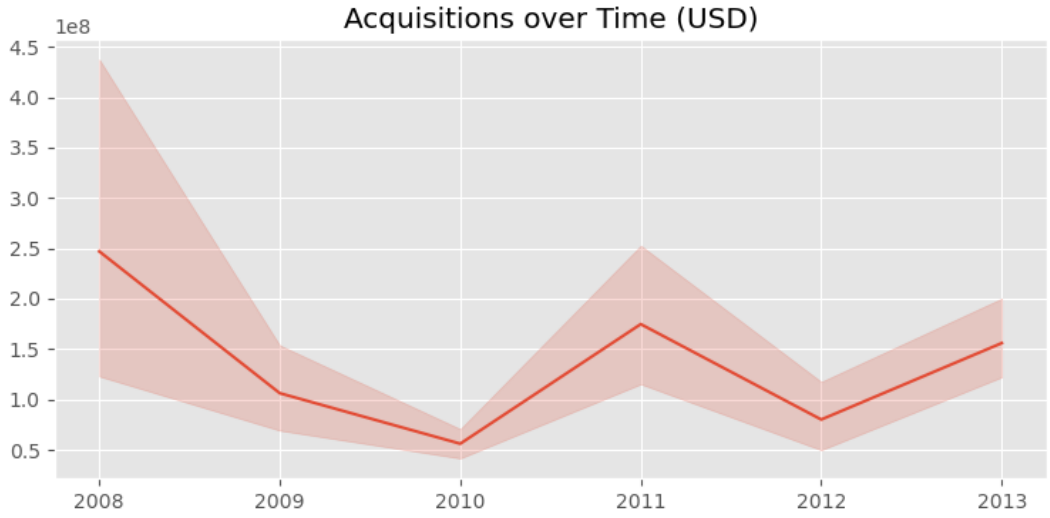
Start-up Status by Investment Type



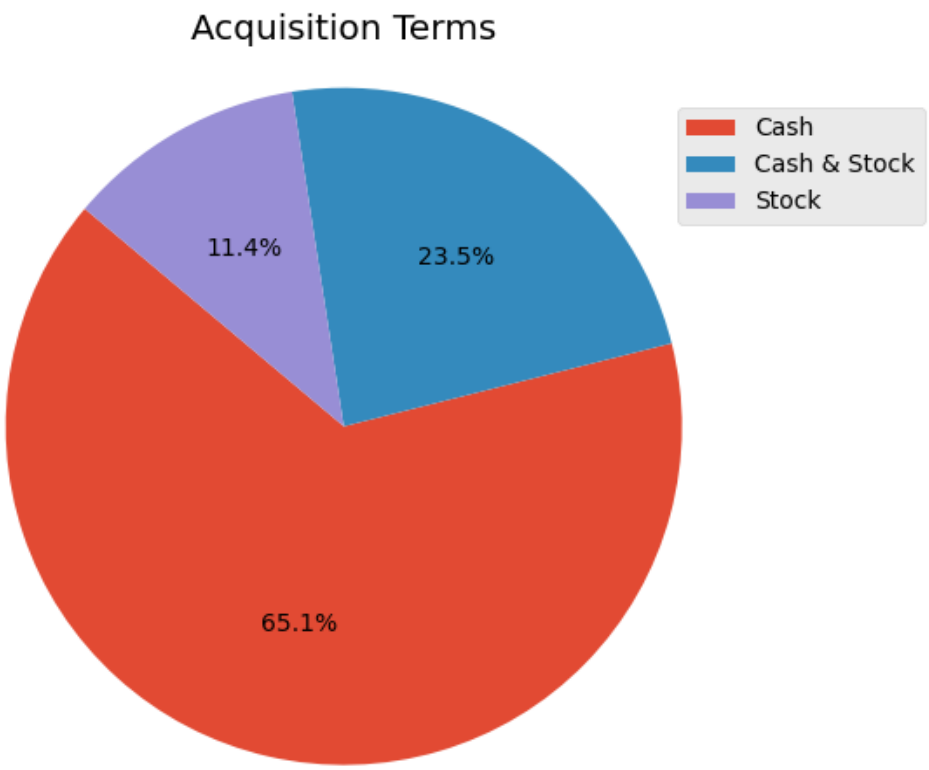
The composition of the kind of investments start-ups receive are generally the same in both closed and acquired start-ups. Angel investment is the most popular investment type. However, we observed that most closed start-ups did not make it to funding round D. But this is just a subset of under a thousand older start-ups. Which also most likely had small seed funding needs. As a start-up matures in its lifecycle and needs much higher levels of funding, venture capital becomes the best option.

Acquisitions

The average acquisition is $122 million. However, we do have rare cases (2.4%) when a start-up was acquired with a billion dollars or higher. These start-ups are called “unicorns.”



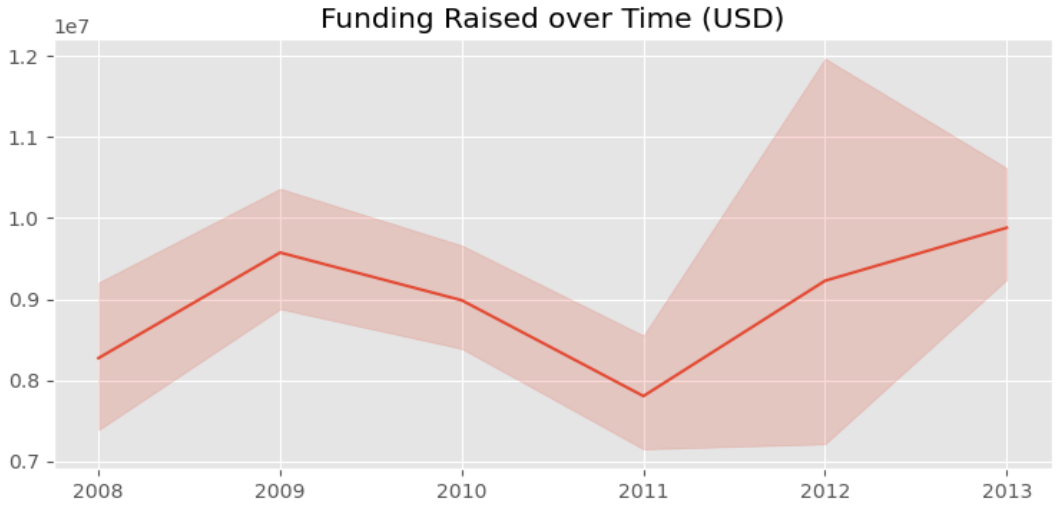
The highest acquisition recorded in the original dataset is of EDS by Hewlett-Packard, in 2008. The acquisition is recorded to be valued at $2.6 trillion. However, after some fact-checking, I came back with the value of $13. 9 billion, which I used to replace the anomalous value.



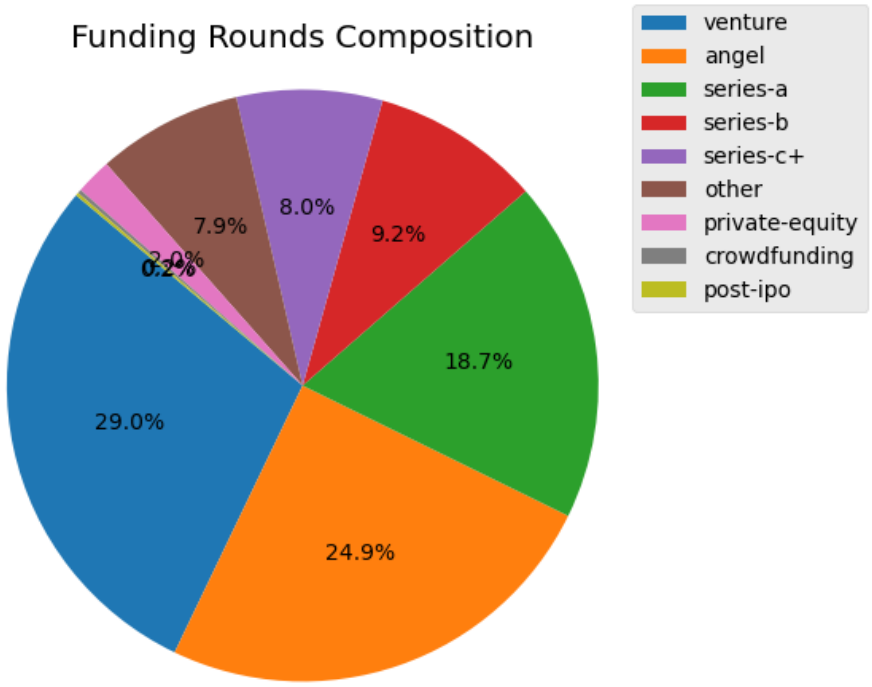
An overwhelming majority of acquisition terms are cash-based followed by a combination of cash & stock, and lastly purely stock.

Funding Rounds

The average amount raised by a start-up in the period (2009 – 2014) recorded was $9 million dollars. We observed a steady rise in funding raised over the period 2011 – 2013.



The composition of the funding round types are as follows;



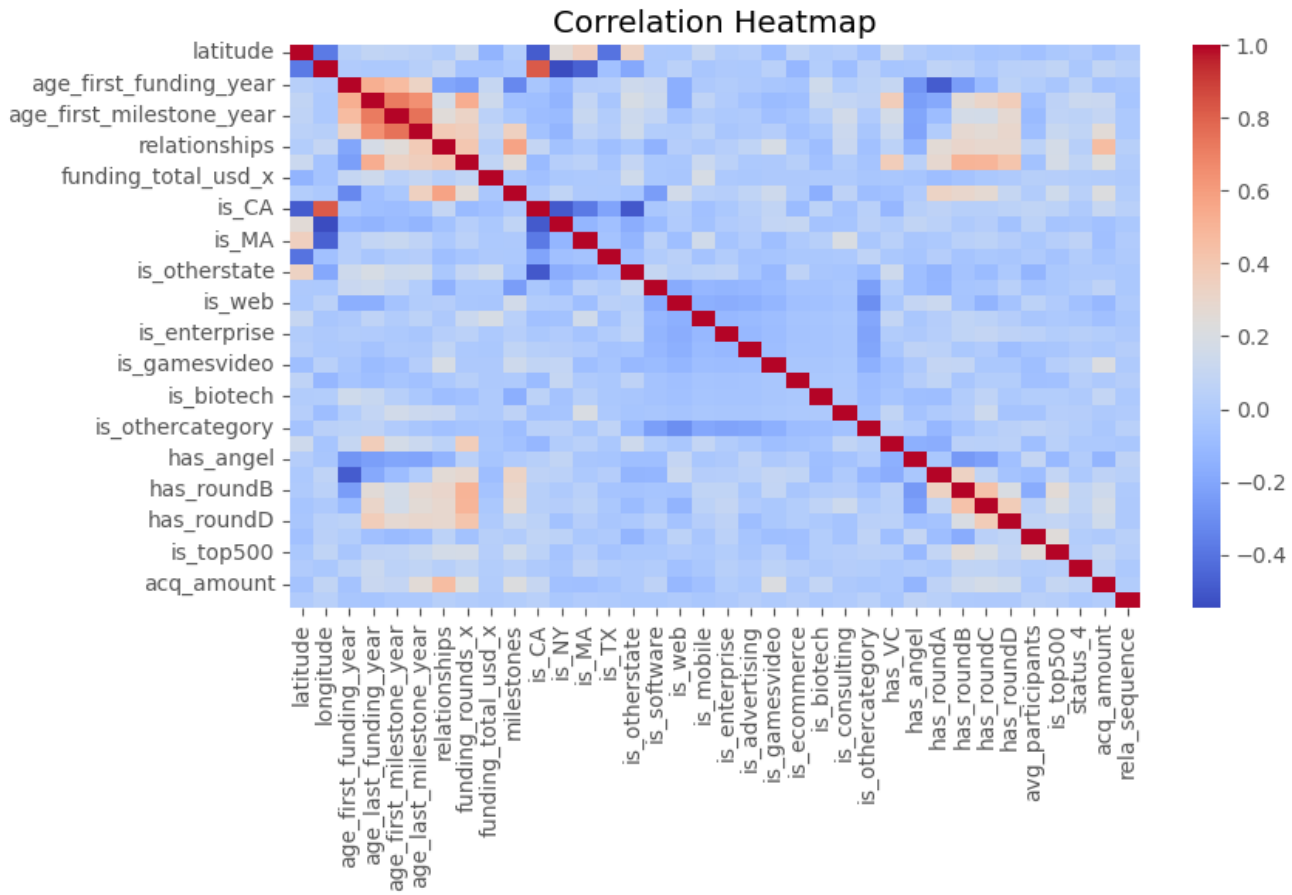
The two most prevalent funding vehicles are angel and venture investment.

In-depth Analysis

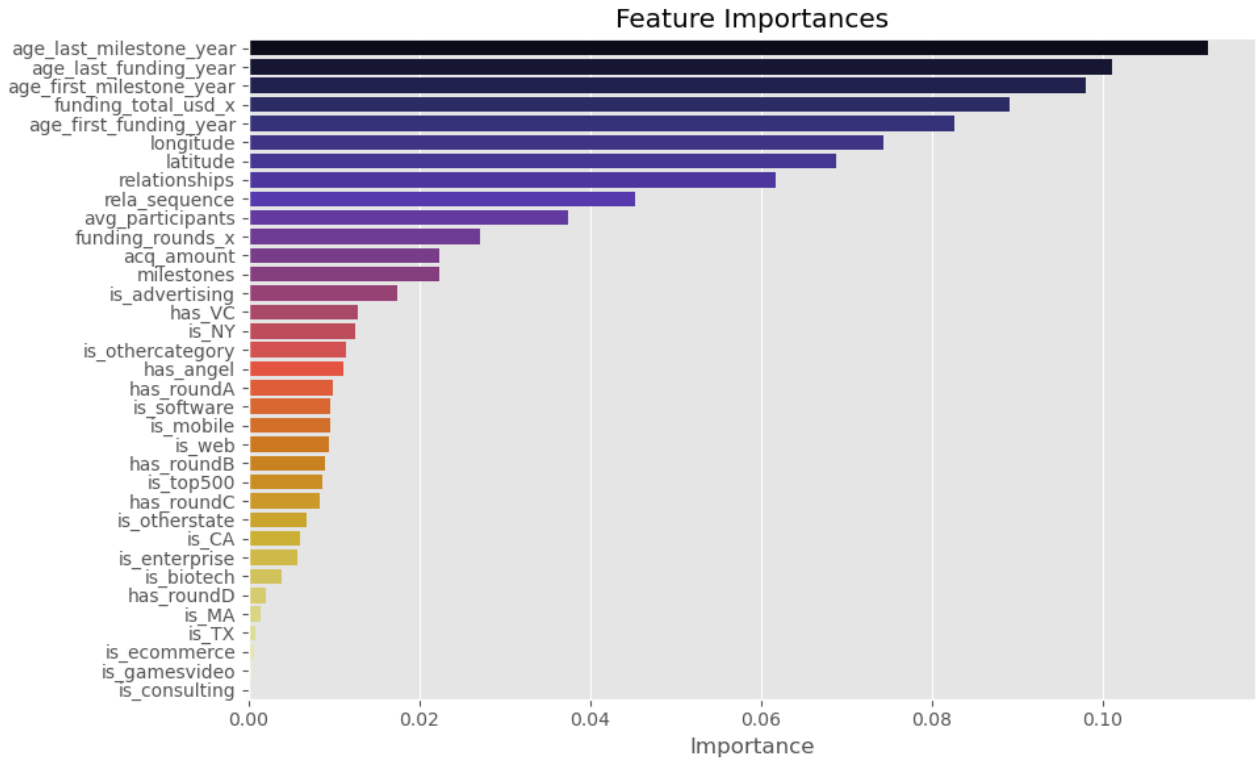
Correlations

I have selected the status of a start-up to determine their success; closed (0), operating (1) or acquired (2). I will use a column ‘status_4’ as my response variable to benchmark success.

Below we observe weak positive correlations with our response. And cluster of relatively high correlations amongst age-related features. But let’s dig deeper into the importances of our features in determining the outcome of our response.



Feature Importance



1**. Age-related features**: The top four most important features are 'age_last_milestone_year', 'age_last_funding_year', 'age_first_milestone_year', and 'age_first_funding_year'. This suggests that the age of the company, as measured by the timing of its milestones and funding rounds, is a crucial factor in determining its 'status_4' value.

2. Funding-related features: The features 'funding_total_usd_x' (total funding amount) and 'funding_rounds_x' (number of funding rounds) have relatively high importance, indicating that the amount of funding and the number of funding rounds are also significant predictors of 'status_4'.

3. Geographic features: The geographic features 'longitude' and 'latitude' have moderate importance, suggesting that the location of the company may play a role in determining its 'status_4'.

4. Company characteristics: Features like 'relationships' (number of relationships), 'rela_sequence' (sequence of relationships), and 'avg_participants' (average number of participants in funding rounds) have moderate to low importance, indicating that these company characteristics have some influence on 'status_4', but not as much as the age and funding-related features.

5. Industry categories: The industry category features (e.g., 'is_advertising', 'is_software', 'is_mobile') generally have low importance, suggesting that the industry category may not be a strong predictor of 'status_4' compared to other features.

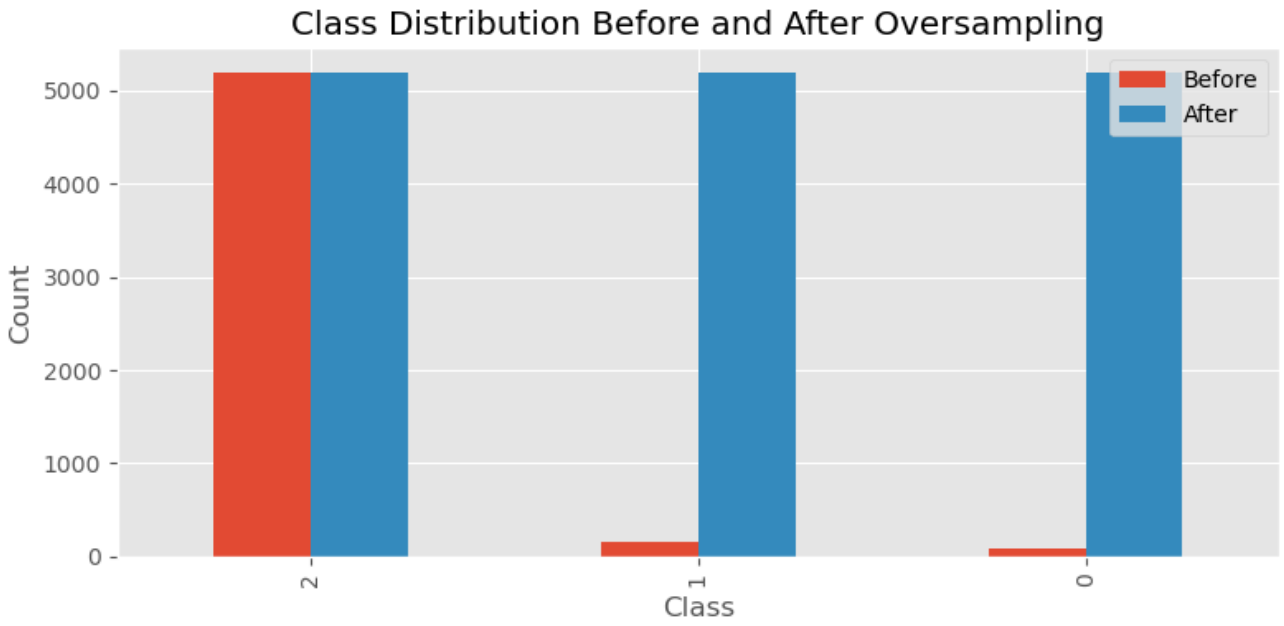
6. Funding sources: Features related to funding sources, such as 'has_VC' (venture capital), 'has_angel' (angel investors), and 'has_roundA/B/C/D' (different funding rounds) have relatively low importance, indicating that the source of funding may not be as influential as other factors in determining 'status_4'.

7. State/location features: The state-related features (e.g., 'is_CA', 'is_NY', 'is_MA', 'is_TX') generally have low importance, suggesting that the specific state where the company is located may not be a strong predictor of 'status_4'.

Overall, the feature importances highlight the significance of age-related and funding-related features, while other factors like industry categories, funding sources, and specific state locations seem to have a lesser impact on the target variable 'status_4'.

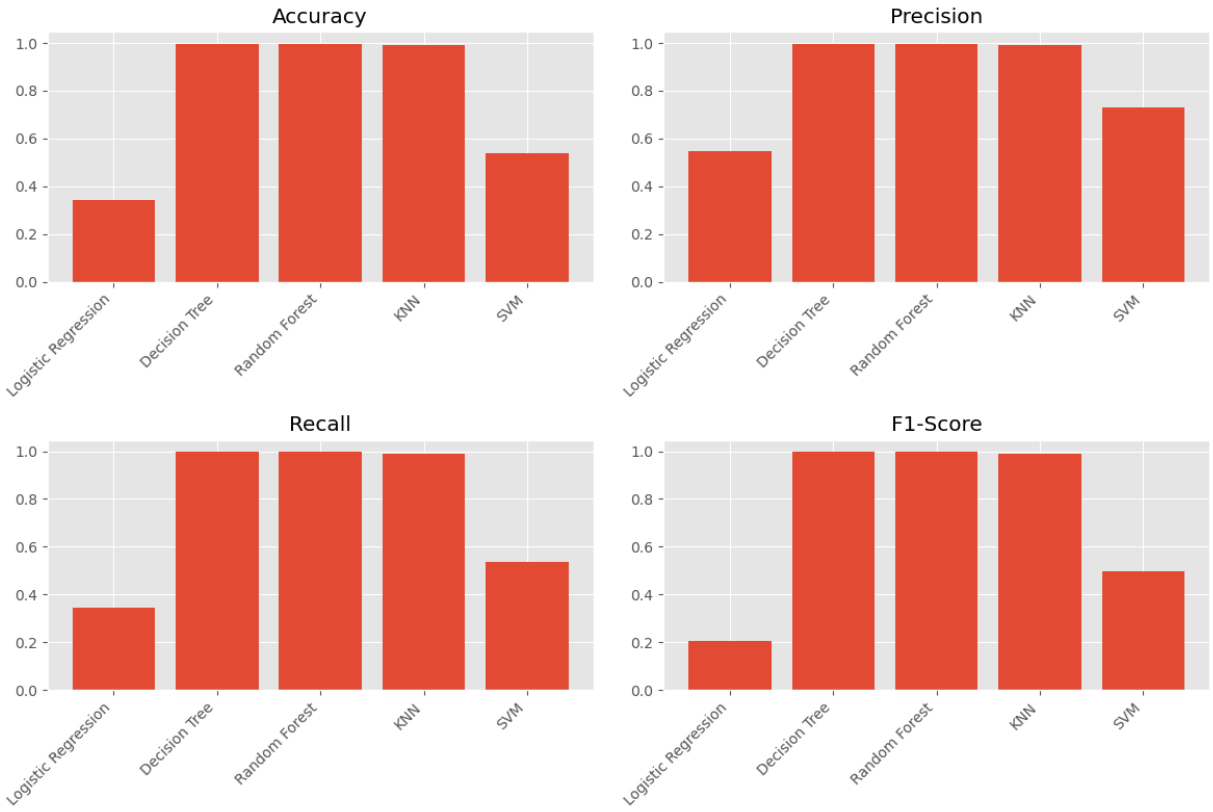
Addressing Class Imbalance

To address the class imbalance, I went with oversampling the minority classes: oversampling involves replicating instances from the minority class to balance the class distribution. I have used SMOTE (Synthetic Minority Over-sampling Technique) for this purpose.



Models & Model Performance

I have trained and compared five different machine learning classification models; LogisticRegression, Decision Tree, RandomForestClassifier, K-Nearest Neighbours (KNN) Classifier and a Support Vector Machine (SVM). And here is how they all performed relative to each other.



Here's a summary of the performance of the different classification models:

1. **Logistic Regression:**
   - Accuracy: 0.341 (low)
   - Precision: 0.547 (moderate)
   - Recall: 0.347 (low)
   - F1-Score: 0.205 (very low)

Overall, Logistic Regression performed poorly on this dataset, with low accuracy, recall, and F1-score, indicating that it struggled to classify the instances correctly.

2. **Decision Tree Classifier:**
   - Accuracy: 0.996 (very high)
   - Precision: 0.996 (very high)
   - Recall: 0.996 (very high)
   - F1-Score: 0.996 (very high)

The Decision Tree Classifier performed exceptionally well, achieving very high scores across all metrics, suggesting that it can classify the instances with high accuracy and balance between precision and recall.

3. **Random Forest Classifier:**
   - Accuracy: 0.997 (very high)
   - Precision: 0.997 (very high)
   - Recall: 0.997 (very high)
   - F1-Score: 0.997 (very high)

Similar to the Decision Tree Classifier, the Random Forest Classifier also performed remarkably well, with very high scores across all metrics, indicating its effectiveness in classifying the instances accurately.

4. **K-Nearest Neighbors (KNN) Classifier:**
   - Accuracy: 0.991 (very high)
   - Precision: 0.991 (very high)
   - Recall: 0.991 (very high)
   - F1-Score: 0.991 (very high)

The KNN Classifier also performed exceptionally well, achieving very high scores across all metrics, comparable to the Decision Tree and Random Forest classifiers.

5. **Support Vector Machine (SVM):**
   - Accuracy: 0.539 (moderate)
   - Precision: 0.730 (high)
   - Recall: 0.535 (moderate)
   - F1-Score: 0.498 (moderate)

The SVM model performed moderately, with a moderate accuracy, high precision, and moderate recall and F1-score, suggesting that it may have struggled with some instances but maintained a good balance between precision and recall.

The tree-based models (Decision Tree and Random Forest) and the KNN model outperformed the other models by a significant margin, achieving very high scores

across all metrics. These models appear to be well-suited for this particular dataset and classification task.
It's important to note that these results may vary depending on the specific data and problem at hand, as well as the hyperparameter tuning and preprocessing steps applied to the models.

While zooming into one of the most performant models in our analysis, the RandomForectClassifier, we get the following report:

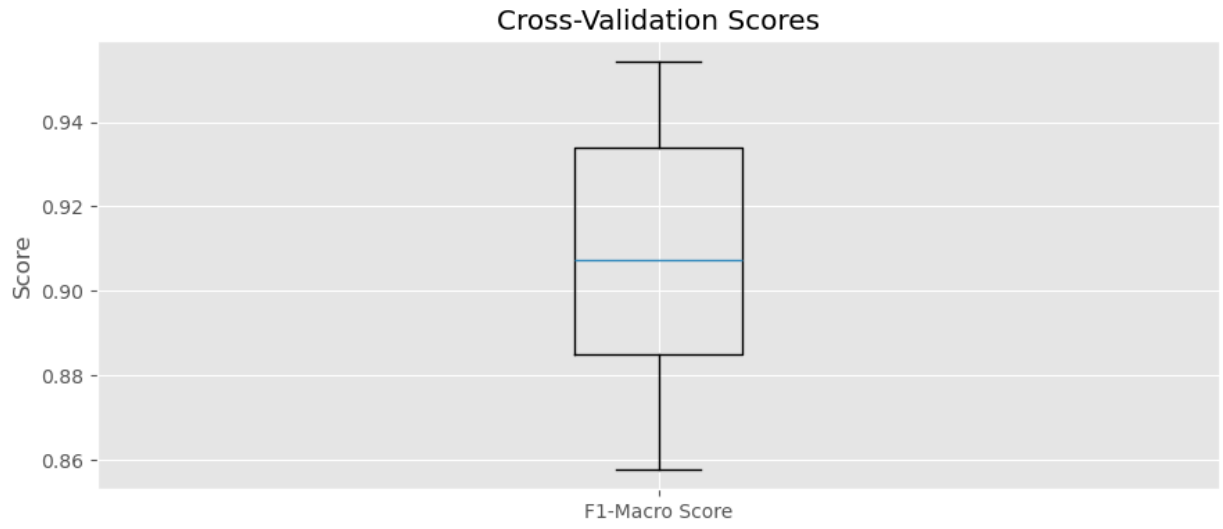
	precision	recall	f1-score	support
0	1.00	1.00	1.00	1057
1	0.99	1.00	1.00	1039
2	1.00	0.99	1.00	1021
accuracy			1.00	3117
macro avg	1.00	1.00	1.00	3117
weighted avg	1.00	1.00	1.00	3117

The classification report suggests that the model has achieved near-perfect performance on this dataset, with no significant issues in terms of precision, recall, or F1-score for any of the classes. The balanced class distribution and the model's ability to correctly classify instances across all classes contribute to this outstanding performance.

However, it's important to note that these results may be specific to the test set used for evaluation. It's always recommended to validate the model's performance on additional data or through cross-validation techniques to ensure its robustness and generalisation capabilities. Let’s perform some cross-validation.

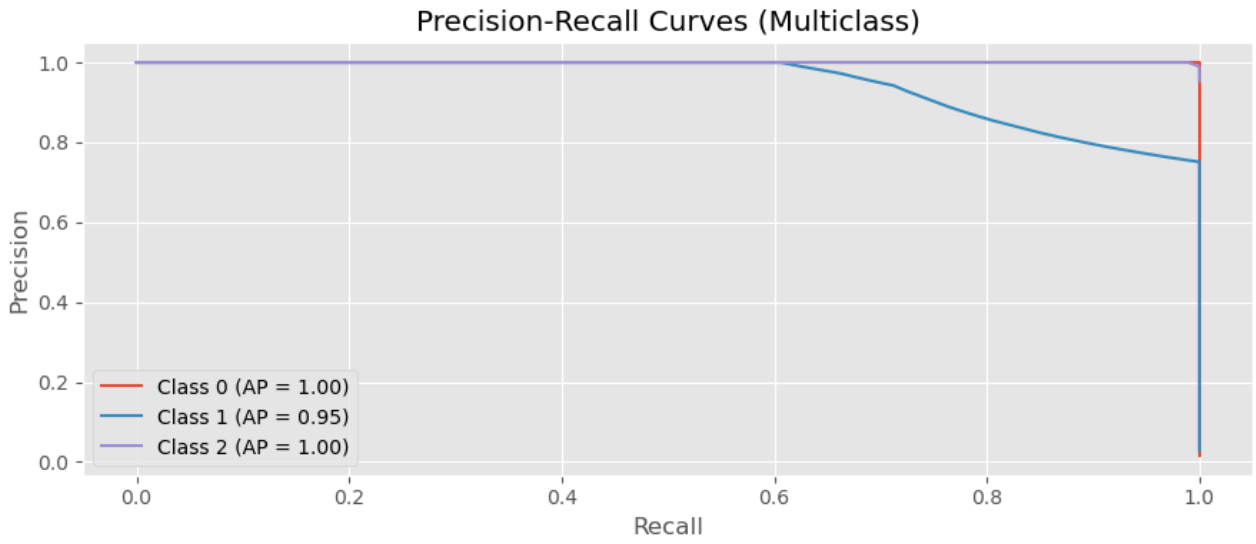
I perform stratified 10-fold cross-validation using *StratifiedKFold* from *sklearn.model_selection.* The *cross_val_score* function is used to compute the F1-

macro score for each fold. The cross-validation scores are visualised using a boxplot, which shows the distribution of the scores across the folds



The boxplot demonstrates that the model's performance, as measured by the F1-Macro score, is stable and consistent across the cross-validation folds. The tight distribution of scores around a high median value (0.91) indicates that the model is performing well and generalising effectively to unseen data.

The following chart shows the precision-recall curves for our multiclass classification problem with three classes (Class 0, Class 1, and Class 2). Remember 0, 1 & 2 indicate whether a start-up has since closed, still operating or has been acquired. The curves plot the trade-off between precision (y-axis) and recall (x-axis) for different probability thresholds.



In an ideal scenario, the precision-recall curve would reach the top-right corner of the plot, representing a precision and recall of 1.0 for all classes. The curves for Class 0 and Class 2 are very close to this ideal case, while the curve for Class 1 is slightly lower but still performs well.

The precision-recall curves suggest that the model performs exceptionally well in identifying instances of all three classes, with Class 0 and Class 2 achieving near-perfect precision and recall, and Class 1 having a slightly lower but still very good performance. The high average precision scores for all classes further reinforce the model's strong performance on this multiclass classification problem.

Conclusion

The realm of venture capital investments is a multifaceted and ever-evolving landscape, where complexities abound, and the pursuit of success is a constant endeavour. As we have witnessed throughout this report, numerous factors intertwine, shaping the trajectories of startups and the decisions of investors. The dynamic nature of this ecosystem, coupled with the rapid pace of technological advancements, often presents challenges in determining the precise ingredients that propel a start-up towards success – the very essence that fuels the vitality of the venture capital domain.

In our quest to unravel the intricate tapestry of start-up success, we have adopted acquisitions as a key indicator of promise and potential. The acquisition of a start-up signifies not only its ability to attract interest from established entities but also a recognition of the value it has created. Moreover, as startups navigate the passage of time, their ability to secure continued funding and sustain growth becomes a testament to the viability of their endeavours. Age, therefore, emerges as a crucial factor, serving as a beacon that illuminates the path towards realising a start-up's true potential. If a start-up can consistently capture the attention of investors and demonstrate its capacity for innovation, it increases the likelihood of achieving the milestone of an acquisition.

While initial public offerings (IPOs) represent the pinnacle of success for startups, their rarity and the extended journey required to reach this stage have led us to focus our attention on acquisitions as a more accessible and pragmatic measure of achievement. The ultimate stage of an IPO, though highly coveted, remains an elusive goal for the vast majority of startups, underscoring the importance of recognising and celebrating intermediate milestones.

As we navigate the intricate tapestry of the venture capital landscape, this report serves as a guiding light, illuminating the factors that shape start-up success and equipping stakeholders with the insights necessary to make informed decisions. By fostering a deeper understanding of the dynamics that govern this ever-evolving ecosystem, we contribute to the collective effort of nurturing innovation, driving economic growth, and unlocking the full potential of entrepreneurial endeavours.