Venture Capitals: Shedding Light on Hyperconnected Global Network of Investments and Startups.

A Submission to Ocean Data Challenge hosted through Desights AI

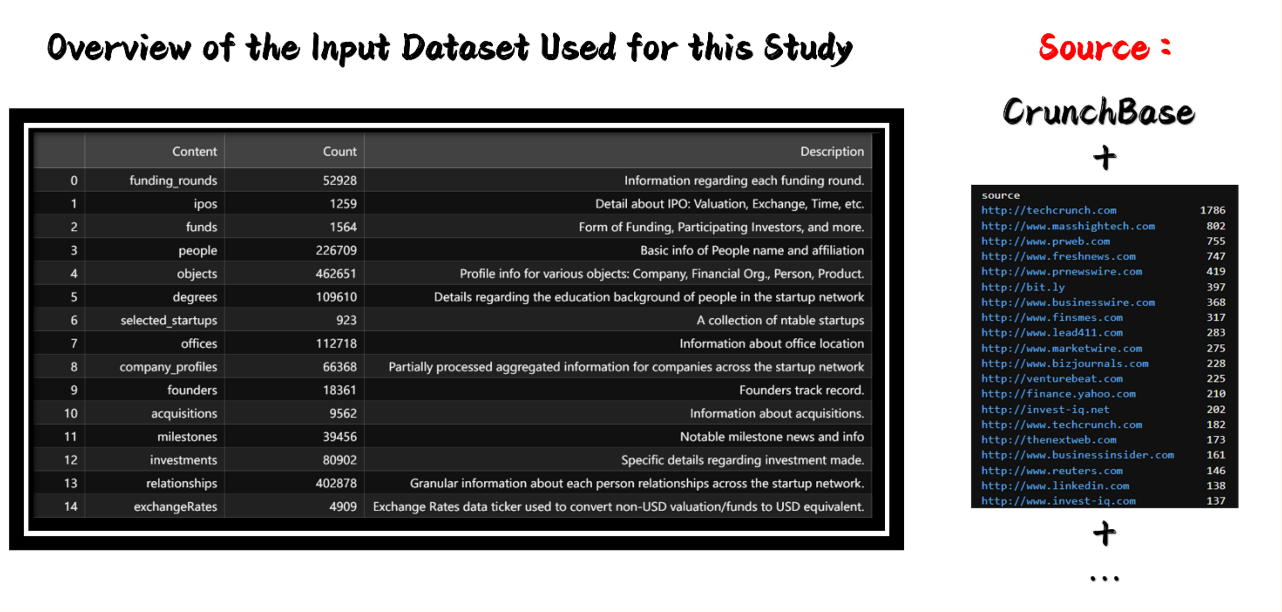
Dominikus Brian 钟鸿盛 | domi@dreambrook.tech
www.dreambrook.tech

Introduction

In this study we uncover the implicit and explicit relationship across the complex startup investment and venture capital (VC) landscape. We dive into the intricate role of various factors such as fundings, relationship, education background, relationship, location, and other factors might determine the likelihood for a startup's failure and success. We deliberately construct various metrics and schemes to help logically dissect possible chain of events that lead to success of startup. A classification model for distinguishing successful startup from those that fail were developed. In constructing the machine learning model, we put emphasize onto the innate structure of context and cause-effect relationship, allowing the developed classification model to have high degree of interpretability. Assessment of accuracy, recall, precision, and the so-called ROC curve were also performed.

Throughout the data analysis we also discover a number of interesting insights that highlight patterns of value related to the global hypernetwork of startup investment by venture capitals, big corporation, and financial organization. Our data exploration aims to answers 8 key questions posed for this dataset, before then expanding on other insights, like investment strategy and such. The 8 key questions are: (1) To uncover most interesting insights from overall analysis, (2) determine most important factors in a startup's success, (3) identify common characteristics of failed startup, (4) cluster VC funds by their existing investment, (5) exploring founder characteristics and their educational background, (6) cluster funder of investment based on typical characteristics, (7) map up network of startup ecosystem member, and (8) track also analyze interesting time-dependent investment trend.

We first begin with summarizing the raw dataset used for this work. The raw dataset for this study comprises of 4 sets of data, the first dataset provides detailed information that are useful in understanding startup ecosystem such as about their financing rounds, investors, acquisitions, and so on. The second focus on the founders for this startup. The third peer into the industry trends and other info for nearly 1000 curated startup list. The fourth provide the rest of the raw company information for 65,000 startups. To begin with, we first clean up and further curate the raw dataset into 14 individual datasets, appropriately treating missing values, and performed numerical encoding for categorical data when necessary. Overview of this processed input dataset is shown in **Fig.1**. This table will also serve as a handy reference when referring to a particular dataset. In the next section we will start off with presenting the key findings along with answering key questions before then proceeds with presenting developed model and explain on its interpretability.



Overview of the Input Dataset Used for this Study

	Content	Count	Description
0	funding_rounds	52928	Information regarding each funding round.
1	ipos	1259	Detail about IPO: Valuation, Exchange, Time, etc.
2	funds	1564	Form of Funding, Participating Investors, and more.
3	people	226709	Basic info of People name and affiliation
4	objects	462651	Profile info for various objects: Company, Financial Org., Person, Product.
5	degrees	109610	Details regarding the education background of people in the startup network
6	selected_startups	923	A collection of ntable startups
7	offices	112718	Information about office location
8	company_profiles	66368	Partially processed aggregated information for companies across the startup network
9	founders	18361	Founders track record.
10	acquisitions	9562	Information about acquisitions.
11	milestones	39456	Notable milestone news and info
12	investments	80902	Specific details regarding investment made.
13	relationships	402878	Granular information about each person relationships across the startup network.
14	exchangeRates	4909	Exchange Rates data ticker used to convert non-USD valuation/funds to USD equivalent.

Source :

CrunchBase

+

```
source
http://techcrunch.com              1786
http://www.masshightech.com         802
http://www.prweb.com                755
http://www.freshnews.com            747
http://www.prnewswire.com           419
http://bit.ly                       397
http://www.businesswire.com         368
http://www.finsmes.com              317
http://www.lead411.com              283
http://www.marketwire.com           275
http://www.bizjournals.com          228
http://venturebeat.com              225
http://finance.yahoo.com            210
http://invest-iq.net                202
http://www.techcrunch.com           182
http://thenextweb.com               173
http://www.businessinsider.com      161
http://www.reuters.com              146
http://www.linkedin.com             138
http://www.invest-iq.com            137
```

+

...

Figure 1. Overview of the data input processed for the analyses in this study. This naming of dataset will also serve as a useful reference table when mentioning which dataset were used for analysis.

Key Findings

(1) To uncover most interesting insights from overall analysis

The most interesting pattern we uncover in this data-intensive study can be crystallized into the following three golden nuggets of insights:

(a) Computer Science, Stanford University, and San Fransisco Bay region exhibit powerful domination across various spectrum of the datasets, be it number of companies, relationships, and number of funding rounds that happens. As a consequence, both community of founders, funders, and other factors are correlated higher with these factors.

(b) Along the years more and more funding rounds series before initial public offering (IPO), such as series D, E, F, G become more frequent after dot-com bubble about the year of 2000. Other forms of funding or equity staking such as convertible, post-IPO debt, post-IPO equity, and private equity also started to gain popularity few years before and after the 2008 financial crisis.

(c) The trend for number of funds raising is increasing by order of magnitudes almost every 5 years, starting from the year of 1995 with only less than 10 entries, up to more than 10,000 by the year of 2013. This trend coincides nicely with the rise of internet, software, and web-based startup, that tend to pursue the rapid growth scheme fueled by VC investments.

(2) Determine most important factors in a startup's success

Through performing correlation analysis and cause-effect tracing of these correlations, here we conclude three of the most important factors that increase the odds of startup to succeed are: (Current) Relationship, have association with San Fransisco Region or New York Region, and being in the right industry/product/service category and having a computer science (from Stanford) / economics (from Harvard). In order to explain and describe how we arrived to these interesting insights and also quickly go through other insights like it, here we first present the **Fig 2, Fig 3, and Fig 4.** The two figures provided a quick snapshot into the distribution of key data and metrics that we will discuss along the report, such as the funding count, product count, funding type, and categorization of the company within the dataset.

Here we highlight several interesting points of observations. Most funding rounds (almost 2/3) involved less then 10 investors or so, with some other having around 12-20 investors, and on some limited instance more than 25 investors. Most fundings are disclosed as seed round, Series funding round C, D, E, F have generally more mean number of investors, as it gets closer to IPO and have potential for bigger exit. Web, software, ecommerce, and education company, have disproportionate number of products compared to those from other categories. We also present our "Correlation to Success" result in **Fig. 5** based on our Pearson correlation analysis for the features in both selected startups and company profile dataset. Key takeaway from the correlation studies shown that there's inherent connection that be based in relationship, especially current relationship (of individual to company), which typically increase the likelihood for a company to attract more funding rounds of larger sum, and then to have more products and being reported to achieve milestones toward success in general. In this report we will take the company that continuously maintain operation, reached IPO, or being through merger & acquisition with bigger company as the basic definition of success, other metrics such as being a fortune 500 company or explicitly labelled as success will also be used when appropriate along the discussion going forward.

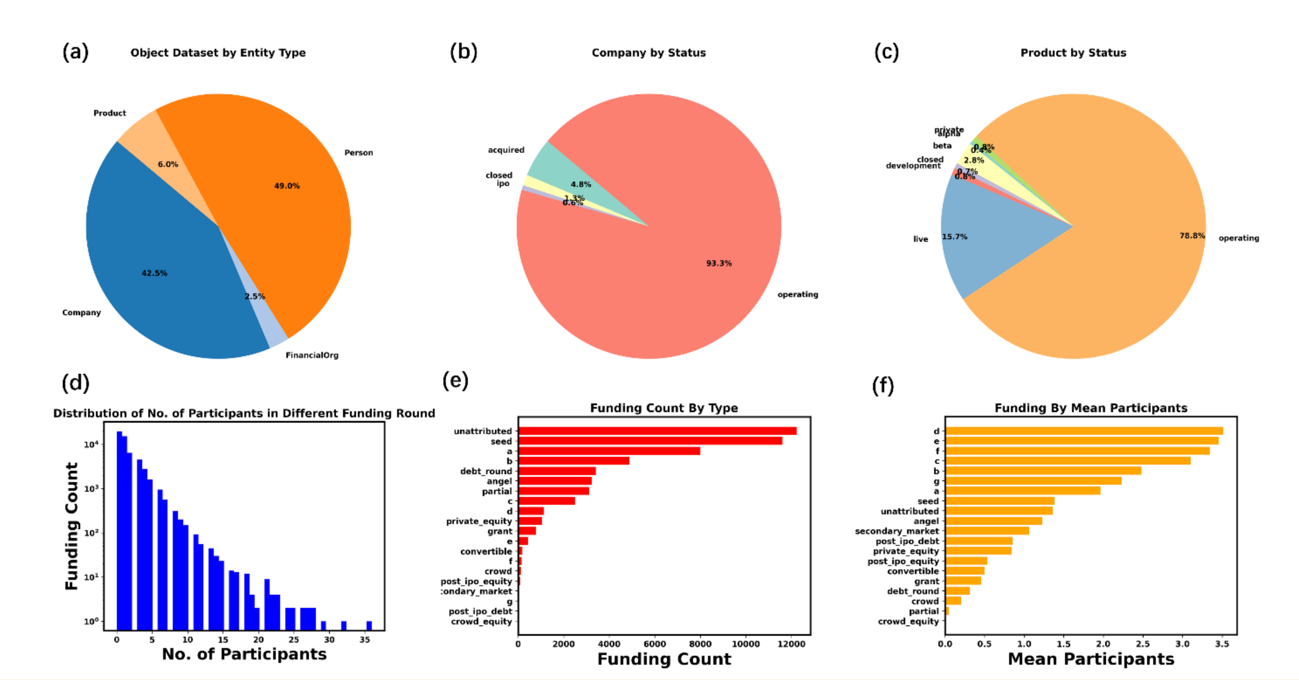


Figure 2. A snapshot into basic objects and funding related the key data contained in the datasets. (a) object dataset by entity type, (b) company distribution by status, (c) product by status, (d) distribution of no. of participants across startup funding rounds in semi-log plot, (e)funding count by type of investment, and (f) funding round by No. of Mean Participant.

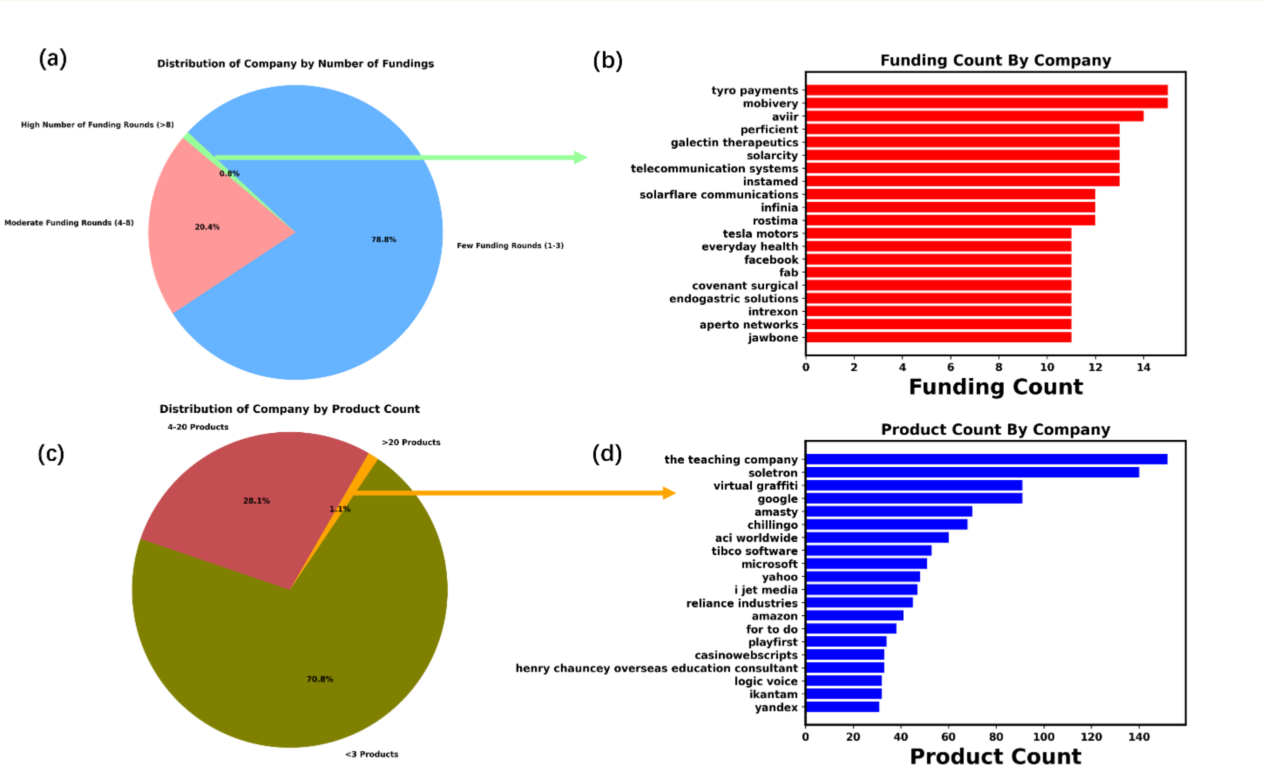


Figure 3. Company by Funding and Product Count. (a) categorization of company by number of fundings (1-3, 4-8, and >8), (b) top 20 companies with the most funding count, (c) categorization of company by product counts (<3, 4-20, and >20), (d) top 20 companies with the most product count.

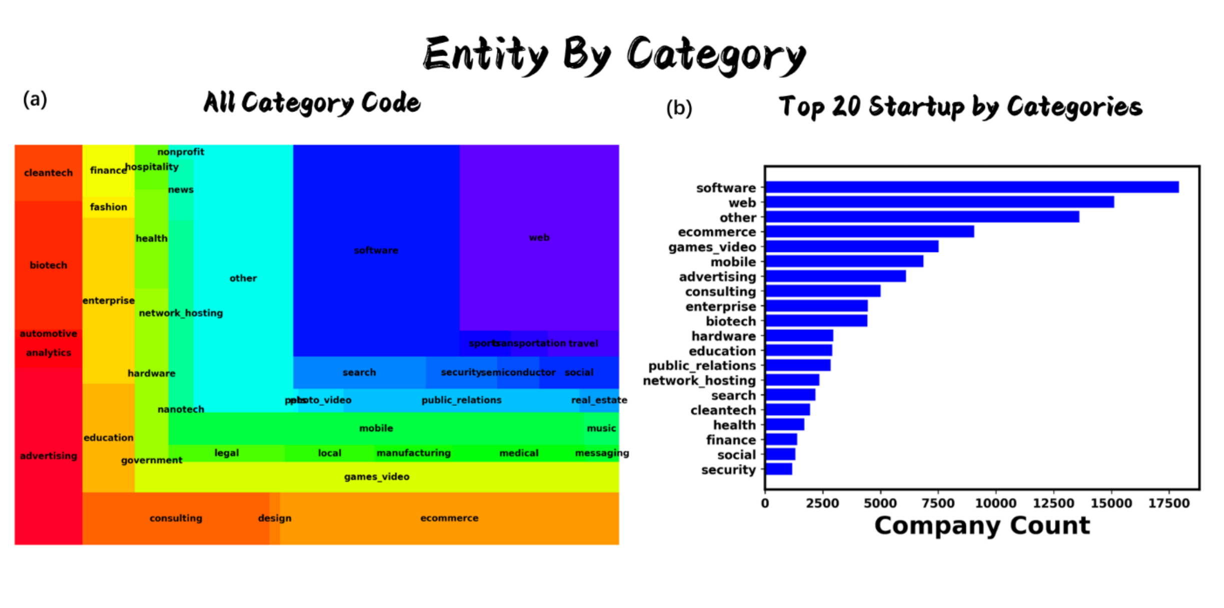


Figure 4. Entity by category, with Treemap for (a) all category code, and (b) company count for top 20 categories

(3) Identifying common characteristics of failed startup

As failure is to certain degree can be regarded as approximately the negative or reverse of success, lessons and factors we discuss in the previous part also applies in the reverse direction here. Moreover specifically, we want to highlight a common characteristic of failed startup, in which we here define as company with the status of closed. To this end, we show in **Fig. 6** a list of instances where failure happens, and emphasize the important difference across failure likelihood. Building a startup in the messaging, and semiconductor have a failure ratio of about 7% and 5%, respectively. These numbers are almost two-three folds higher compared to most other category such as software, ecommerce, and advertising that have only about 1-2% failure ratio.

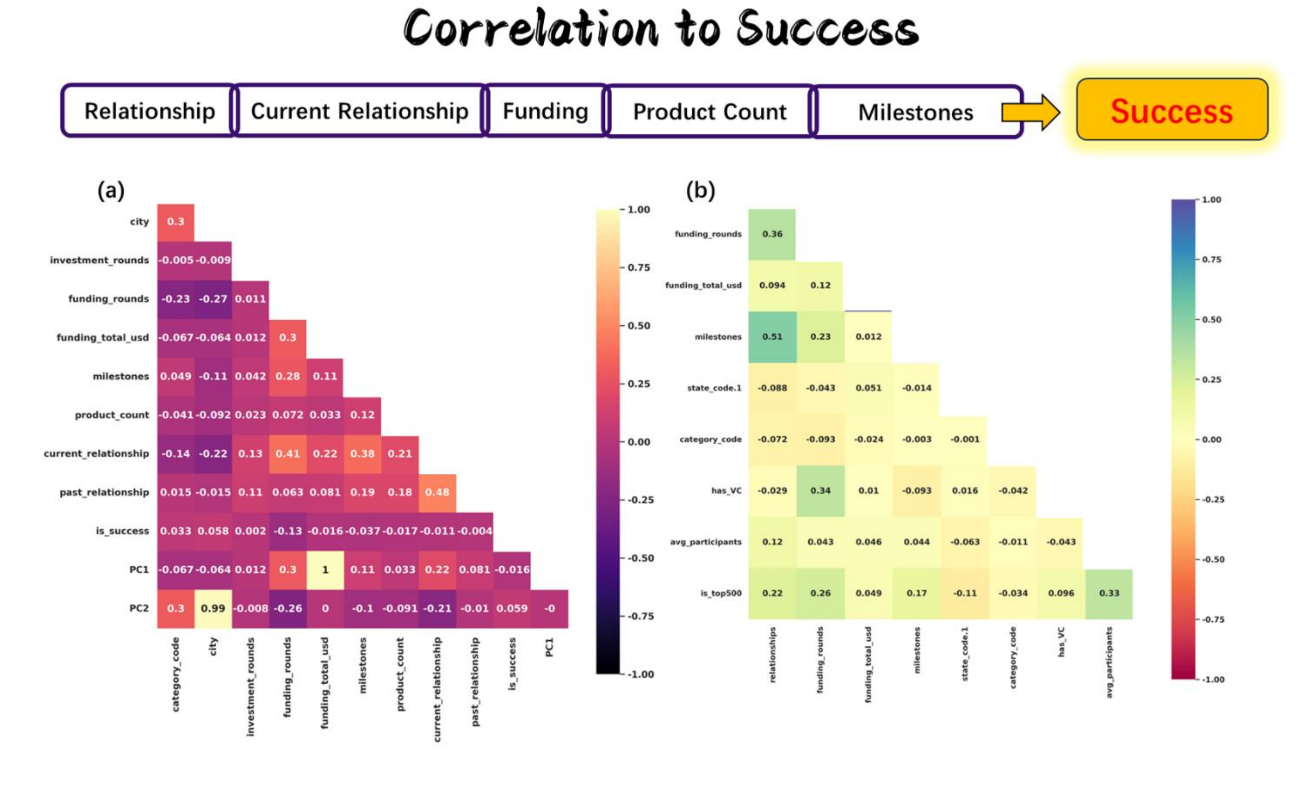


Figure 5. Chain of correlation of success along with Pearson correlation heatmap from the (a) curated company dataset, and (b) full company dataset. PC here stands for Principal Component, obtained from Principal Component analysis.

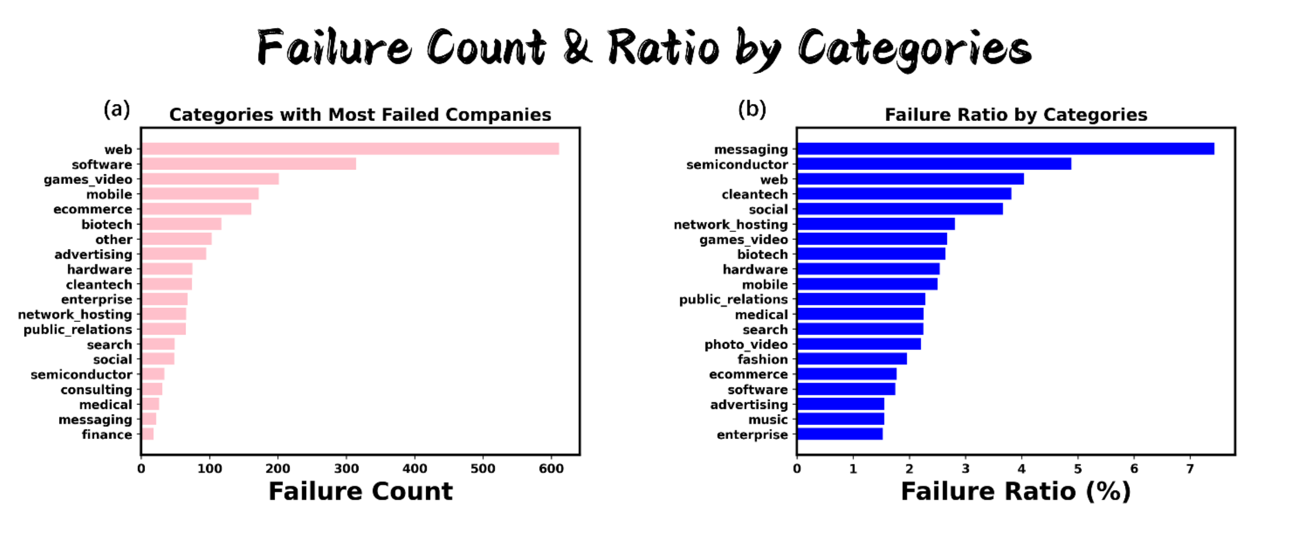


Figure 6. Failure count and ratio by categories with (a) showing for absolute failure count in the dataset

(4) VC funds clustered by their existing investment

In **Fig. 7** we present our clustering result for the Venture Capital data based on their existing investment. To this end, we utilize the principal component analysis (PCA) to first reduced the multidimensional data of the investment dataset into a reduced dimension data with only 2 most important factors. Here the principal component obtained are by definition orthogonal to each other (i.e., inherently have no correlation to each other) thereby is a great tool to have a distinguishable feature useful for clustering and other model construction purposes. In **Fig.5** in which similar PCA analysis were also performed on the startup dataset, we can observe that the PC1 (1^{st} component), and PC2 (2^{nd} component) each are connected to other meaningful metrics in a specific way. For instance, in **Fig.5** PC1 is physically mainly representing the funding total in usd, along with noticeable correlation to funding rounds, and current relationship. The PC2 however is completely unrelated to PC1 and is mainly represent city (and other location related features such as Zip code, state, country which is not shown for simplicity), and also a decent correlation with category code. Our clustering reveals an optimal total of 6 Clusters, each with its own characteristics. For instance, in the cluster 0 the apparent feature is investment provided by prominent VC like Greylock or Accel to individuals, in this case Kim and Robert Rodriquez and Evan Nisselson. Other cluster, such as cluster 5 exhibit pattern of investment made by different entity onto single person or corporation. The clusters were obtained by k-means clustering method, and the optimal number of clusters is determined by performing so-called "elbow scan" procedures on number of possible clusters. For technically interested readers, details are provided in methodology section.

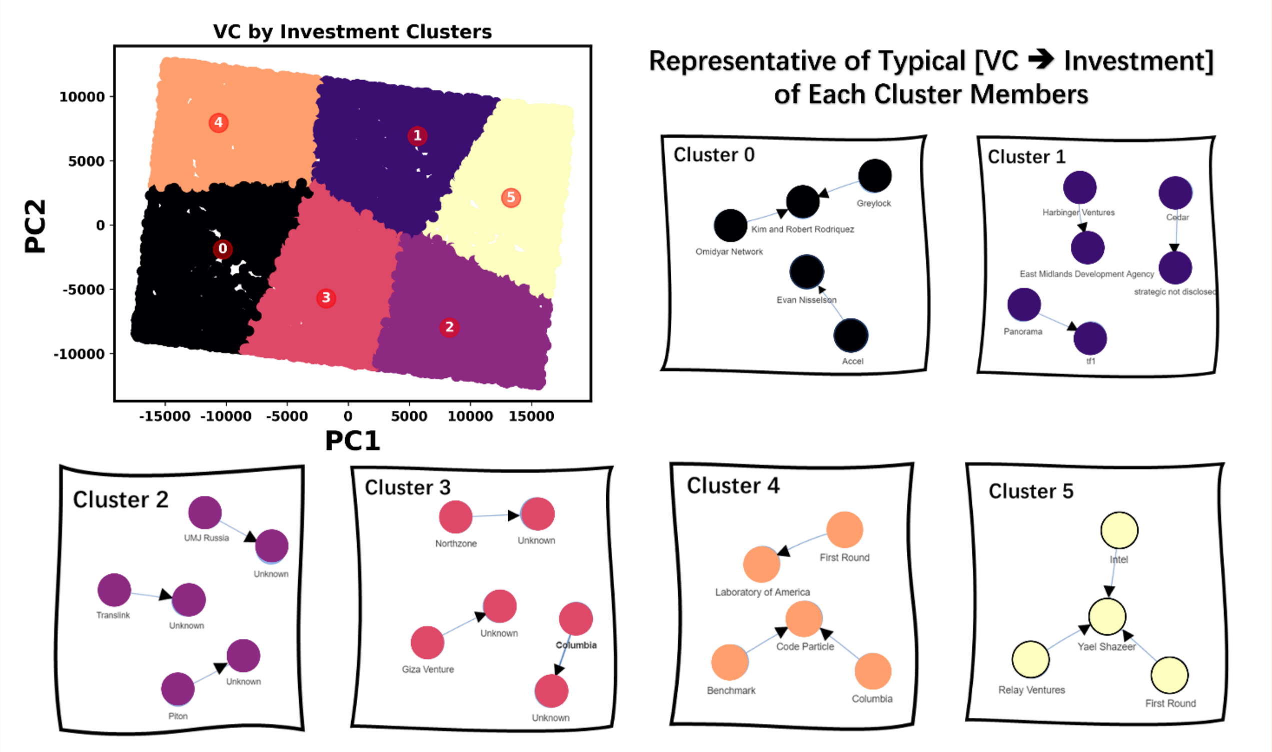
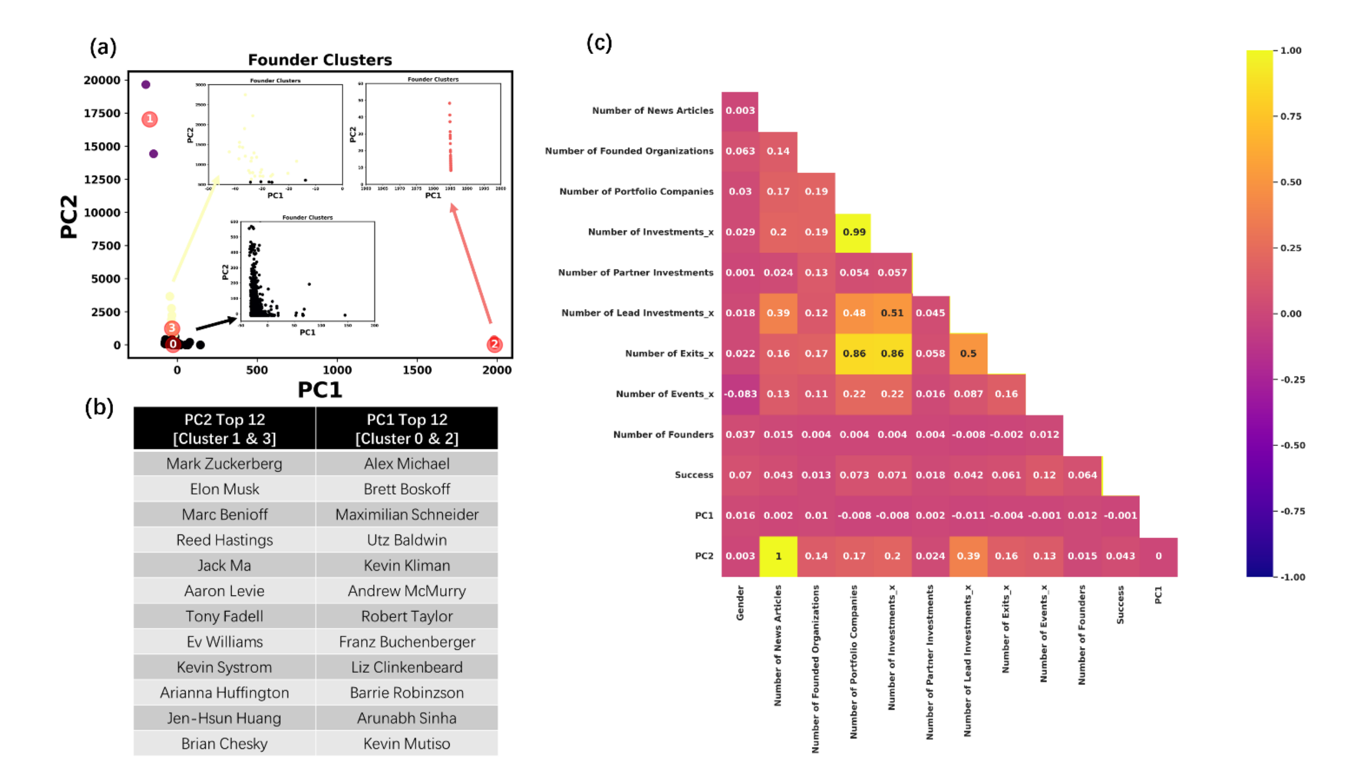


Figure 7. Venture Capital Cluster by existing Investment based on the reduced dimensions feature obtained through principal component analysis. Representative of typical VC-> Invested company/individual for each cluster were also shown.

(5) Exploring founder characteristics and their educational background

In this section we will peer through founder background and correlation between different features associated with him/her/they. We then proceed with looking into the trend of how academic background of these individuals led toward them becoming founder or member of startup ecosystem. In **Fig.8 (a)** we again utilize the PCA dimensional reduction technique for the purpose of clustering the founder data. The optimized number of clusters is given by 4 clusters, and essentially we observe that cluster 1 & 3 are broadly in the same spectrum of more well-known or successful founder, while cluster 0 & 2 are more toward a less known or considerably less successful founders. We then extract top 12 data of cluster member based on sorted PC2 or PC1 axis. In the PC2 representative data , we see household names of startup world such as Mark Zuckerberg, Elon Musk, Jack Ma, and Jensen Huang, while the other set based on PC1 data listed out names such as Utz Baldwin, Brett Boskoff, and so on. Notably here PC2 is highly correlated with number of news article which then naturally biased or inclined toward founder with reputation and fame (which often comes from being successful), and to number of lead investment (with funds which usually only made possible by previous successful ventures). PC1 on the other hand are less correlated with the other factors. Cross correlation between other factors can also be observed in **Fig.8(c)** such as number of exits by a founder is highly correlated with the number of portfolio companies and number of investment, and so on.

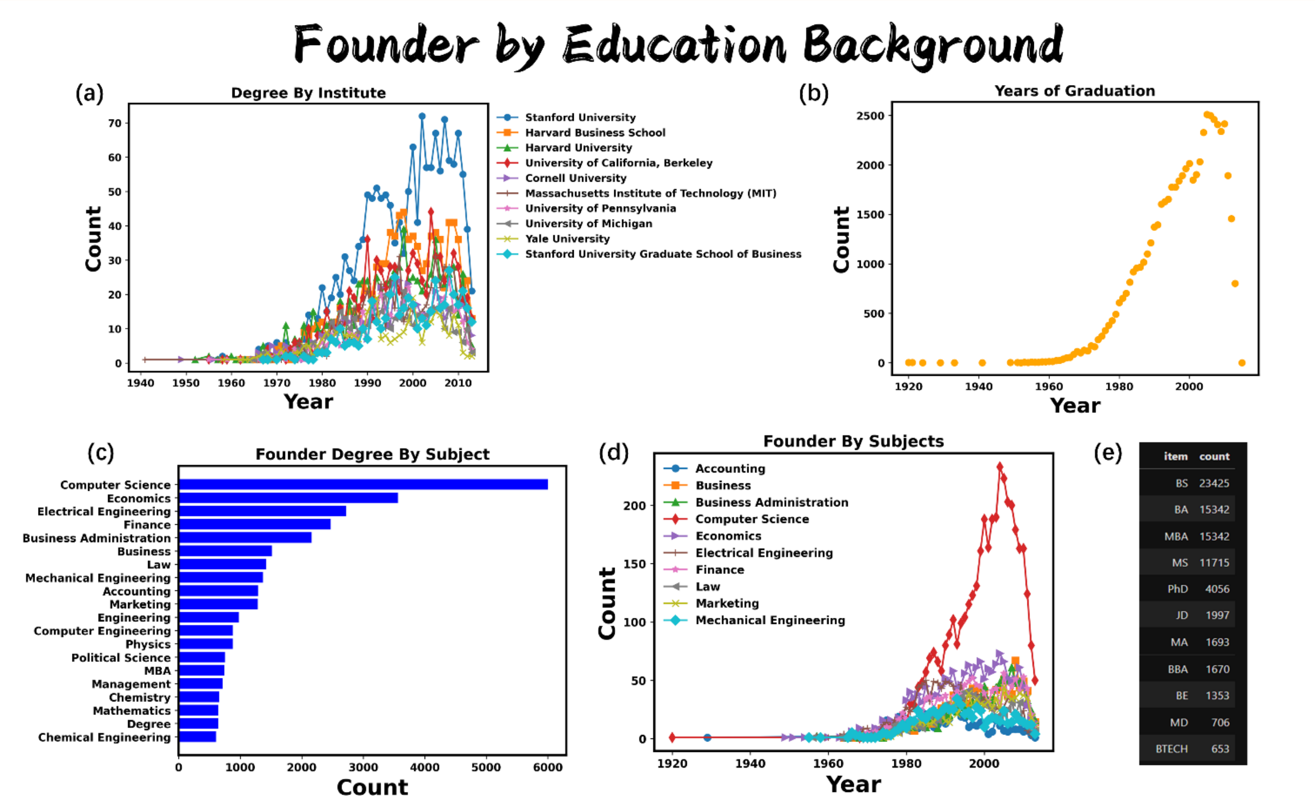


PC2 Top 12 [Cluster 1 & 3]	PC1 Top 12 [Cluster 0 & 2]
Mark Zuckerberg	Alex Michael
Elon Musk	Brett Boskoff
Marc Benioff	Maximilian Schneider
Reed Hastings	Utz Baldwin
Jack Ma	Kevin Kliman
Aaron Levie	Andrew McMurry
Tony Fadell	Robert Taylor
Ev Williams	Franz Buchenberger
Kevin Systrom	Liz Clinkenbeard
Arianna Huffington	Barrie Robinzson
Jen-Hsun Huang	Arunabh Sinha
Brian Chesky	Kevin Mutiso

	Gender	Number of News Articles	Number of Founded Organizations	Number of Portfolio Companies	Number of Investments_x	Number of Partner Investments	Number of Lead Investments_x	Number of Exits_x	Number of Events_x	Number of Founders	Success	PC1
Number of News Articles	0.003											
Number of Founded Organizations	0.063	0.14										
Number of Portfolio Companies	0.03	0.17	0.19									
Number of Investments_x	0.029	0.2	0.19	0.99								
Number of Partner Investments	0.001	0.024	0.13	0.054	0.057							
Number of Lead Investments_x	0.018	0.39	0.12	0.48	0.51	0.045						
Number of Exits_x	0.022	0.16	0.17	0.86	0.86	0.058	0.5					
Number of Events_x	-0.083	0.13	0.11	0.22	0.22	0.016	0.087	0.16				
Number of Founders	0.037	0.015	0.004	0.004	0.004	0.004	-0.008	-0.002	0.012			
Success	0.07	0.043	0.013	0.073	0.071	0.018	0.042	0.061	0.12	0.064		
PC1	0.016	0.002	0.01	-0.008	-0.008	0.002	-0.011	-0.004	-0.001	0.012	-0.001	
PC2	0.003	1	0.14	0.17	0.2	0.024	0.39	0.16	0.13	0.015	0.043	0

Figure 8. Founder's Characteristics, Correlation, and Clusters. (a) Founder data clustered by reduced dimension principal component analysis, (b) representative sample of the founder/startup individual cluster members obtained, (c) correlation between features used for dimension reduction.

Next, let us dive deeper toward understanding the influence or effect of founder academic background through visualization of data and analysis shown in **Fig. 9.**



item	count
BS	23425
BA	15342
MBA	15342
MS	11715
PhD	4056
JD	1997
MA	1693
BBA	1670
BE	1353
MD	706
BTECH	653

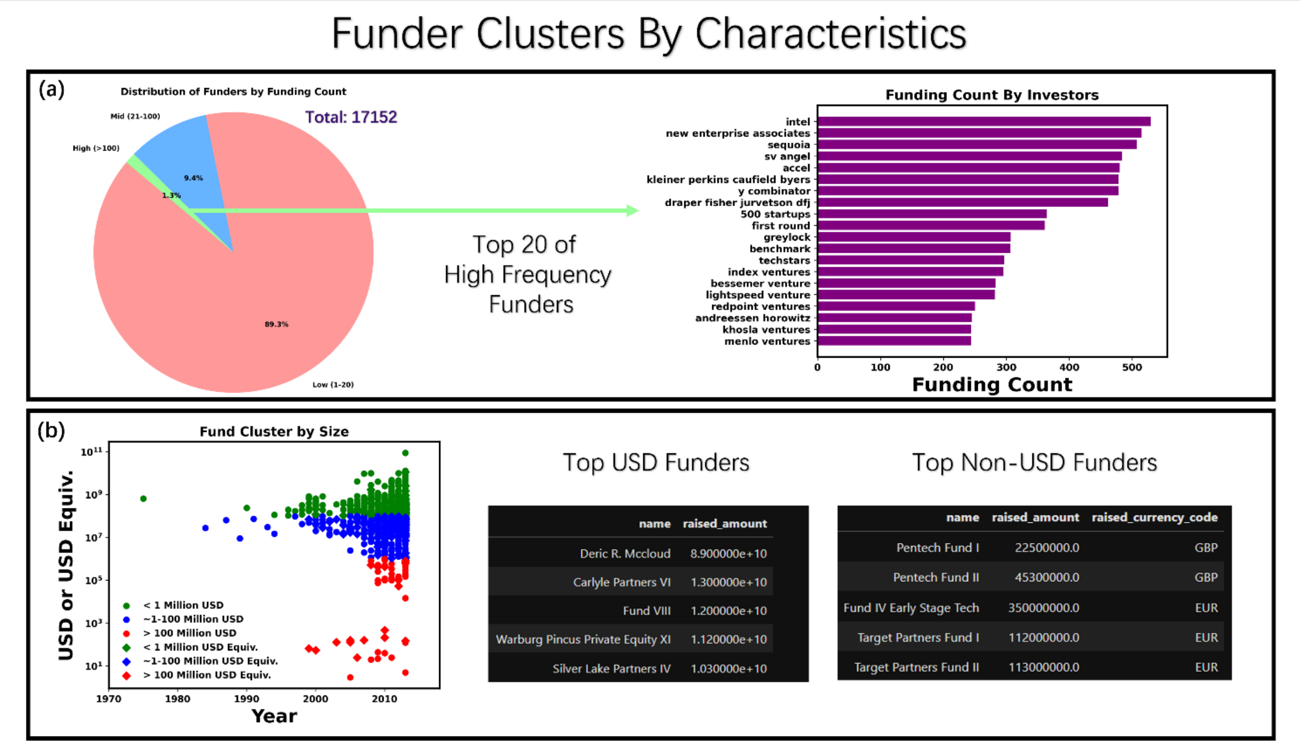
Figure 9. Founder by Education background analyzed based on (a) institution in which founder's degree was obtained, (b) years of graduation, (c) subject in which degree was obtained, (d) evolution of the number of graduates from specific major along the years, (e) type of degree obtained by startup founders.

Fig. 9 (a) shown that Ivy league and top universities in USA are the dominant force in the startup ecosystem world. Further, from the analysis result shown in **Fig. 9 (b)**, we notice that individuals that graduate between 1985-2005 are the dominant players across the startup ecosystems, emphasizing importance of role for those born in the 70s to late 80s. Moreover, in **Fig. 9 (c)** we observe that 5 majors in particulars: Computer Science, Economics, Electrical Engineering, Finance, and Business Administration made up almost more than half of all founder's academic background by degree subject. **Fig. 9 (a & d)** elucidate the fact that the increase of founders graduating in the span of late 80s to early 2000s coincide with spike of computer science major graduates, which also peaked out about the same time frame, notably the numbers of graduate from Stanford University for the same period also increases by similar multitude. In terms of degree obtained, most are of BS (Bachelor of Science) or BA (Bachelor of Art), making about ~48,000 in number, while advanced degree like MBA, MS, PhD, and JD, MD are also dominant, together amount to ~32,000 individuals or so.

(6) Funder clusters and typical characteristics

Taking a different approach in clustering the funders, we perform a categorization-based clustering using features that are straightforward and insightful. In **Fig. 10.** and **Fig. 11** We display the resulting clusters of funders by characteristics. Fig.10 (a) shows distribution of funders by funding count, dividing them into low (<1-20 fundings), mid (21-100 funding counts) and high (>100 fundings) clusters. We then listed out top 20 of the high frequency funders. On top of the list are Corporation and VC such as Intel, Sequoia, SV Angel, accel, Kleiner Perkins, Y Combinators, and so on, the very name we would expect for such list, with funding counts ranging from 300-more than 500 for the first top 10 of the list. In the lower panel of Fig 10. We also show for funder cluster by size of the investment made, we divide this into three for one with < 1 million USD (or USD equivalent), between 1-100 million USD, and above 100 Million USD. When it comes to the funders identity it is interesting to also see individual name on huge investment that are amounting hundreds of million dollars alongside common institutional names.

In **Fig.11(a-d)** we display the sorted leaderboard for the financial organization in the investment dataset. One of the most interesting findings here is that in terms of investment round and number of invested companies the leaderboard is completely dominated by funding institution located in SF Bay region, taking all 10 spots for all two categories. So, when people say “that’s where the money is” people are more likely than not does meant it whole heartedly as we can see through the data. When we look into relationship, we start to see importance of investment bankers such as Goldman Sachs, Morgan Stanley, Merrill Lynch, and so on. These institutions being in New York and Wallstreet are almost indispensable when it comes to late-stage startup investment and IPO. In terms of Milestone the data are more dispersed owing to the diversity and incompleteness of how milestone varies.



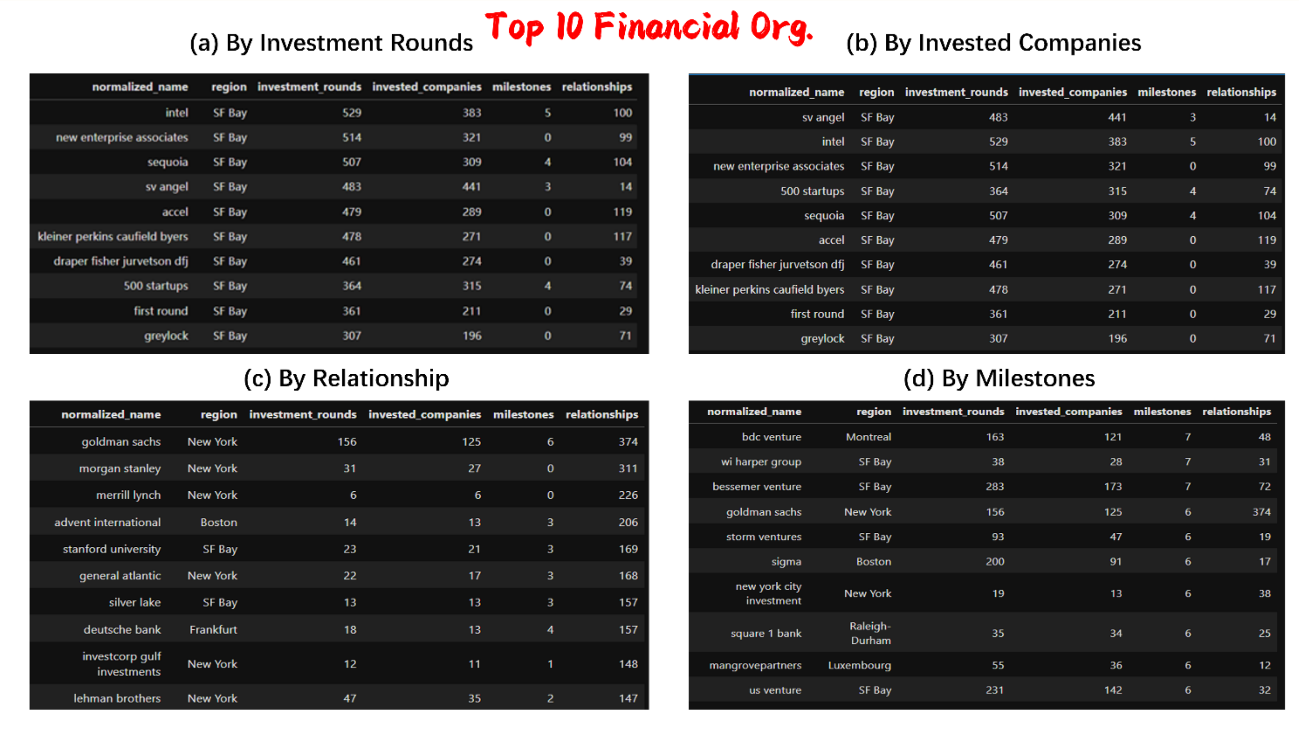
Top USD Funders

name	raised_amount
Deric R. Mccloud	8.900000e+10
Carlyle Partners VI	1.300000e+10
Fund VIII	1.200000e+10
Warburg Pincus Private Equity XI	1.120000e+10
Silver Lake Partners IV	1.030000e+10

Top Non-USD Funders

name	raised_amount	raised_currency_code
Pentech Fund I	22500000.0	GBP
Pentech Fund II	45300000.0	GBP
Fund IV Early Stage Tech	350000000.0	EUR
Target Partners Fund I	112000000.0	EUR
Target Partners Fund II	113000000.0	EUR

Figure 10. Funder Cluster by other characteristics, here we present two cluster categorization based on the number of funding an organization involved in and also the particular funding size



Top 10 Financial Org.

(a) By Investment Rounds

normalized_name	region	investment_rounds	invested_companies	milestones	relationships
intel	SF Bay	529	383	5	100
new enterprise associates	SF Bay	514	321	0	99
sequoia	SF Bay	507	309	4	104
sv angel	SF Bay	483	441	3	14
accel	SF Bay	479	289	0	119
kleiner perkins caufield byers	SF Bay	478	271	0	117
draper fisher jurvetson dfj	SF Bay	461	274	0	39
500 startups	SF Bay	364	315	4	74
first round	SF Bay	361	211	0	29
greylock	SF Bay	307	196	0	71

(b) By Invested Companies

normalized_name	region	investment_rounds	invested_companies	milestones	relationships
sv angel	SF Bay	483	441	3	14
intel	SF Bay	529	383	5	100
new enterprise associates	SF Bay	514	321	0	99
500 startups	SF Bay	364	315	4	74
sequoia	SF Bay	507	309	4	104
accel	SF Bay	479	289	0	119
draper fisher jurvetson dfj	SF Bay	461	274	0	39
kleiner perkins caufield byers	SF Bay	478	271	0	117
first round	SF Bay	361	211	0	29
greylock	SF Bay	307	196	0	71

(c) By Relationship

normalized_name	region	investment_rounds	invested_companies	milestones	relationships
goldman sachs	New York	156	125	6	374
morgan stanley	New York	31	27	0	311
merrill lynch	New York	6	6	0	226
advent international	Boston	14	13	3	206
stanford university	SF Bay	23	21	3	169
general atlantic	New York	22	17	3	168
silver lake	SF Bay	13	13	3	157
deutsche bank	Frankfurt	18	13	4	157
investcorp gulf investments	New York	12	11	1	148
lehman brothers	New York	47	35	2	147

(d) By Milestones

normalized_name	region	investment_rounds	invested_companies	milestones	relationships
bdc venture	Montreal	163	121	7	48
wi harper group	SF Bay	38	28	7	31
bessemer venture	SF Bay	283	173	7	72
goldman sachs	New York	156	125	6	374
storm ventures	SF Bay	93	47	6	19
sigma	Boston	200	91	6	17
new york city investment	New York	19	13	6	38
square 1 bank	Raleigh-Durham	35	34	6	25
mangrovepartners	Luxembourg	55	36	6	12
us venture	SF Bay	231	142	6	32

Figure. 11. Hall of Fame Leaderboard for top 10 financial organization in terms of (a) investment rounds, (b) number of invested companies, (c) relationship, and (d) milestones achieved.

(7) Landscape of startup ecosystem member/individual

In mapping the member/individual network, we will partly refer to the **Fig.7** and **Fig.8** that have defines how the individuals across the ecosystem is connected to institutions/companies (**Fig.7**) and to other startup founders, co-founders and executives (**Fig. 8**). To enrich the analysis, here we present additional mapping, by literal means, with justification that these individuals are working from offices and headquarter of their companies/startup. Therefore, by mapping the location of this offices, we also get a snapshot of how these startup ecosystem members are networked and distributed globally. The mapped result is shown in **Fig.12** below. The map displayed the startup Hot Spots across the globe as contained in the dataset.

(8) Tracking and analyzing time-dependent investment trend

We also performed a comprehensive analysis to time dependent investment trend which can uncover many timely insights and hindsight onto the underlying changes across the startup investments over the years, shown in **Fig 13(a-d).** The first insights from the investment trend are in the orders of magnitude increase of fund-raising investment count starting from 1995, with rate of about 1 order of magnitude every 5 years. The second observation is that most IPO are predominantly happened on NASDAQ and NYSE. Getting over the dot-com bubble there's a significant increase in series funding investment along with various creative or innovative funding scheme such as crowd equity, convertible, post-IPO debt and so on. The number of series round being raised are also notably increases over the years.

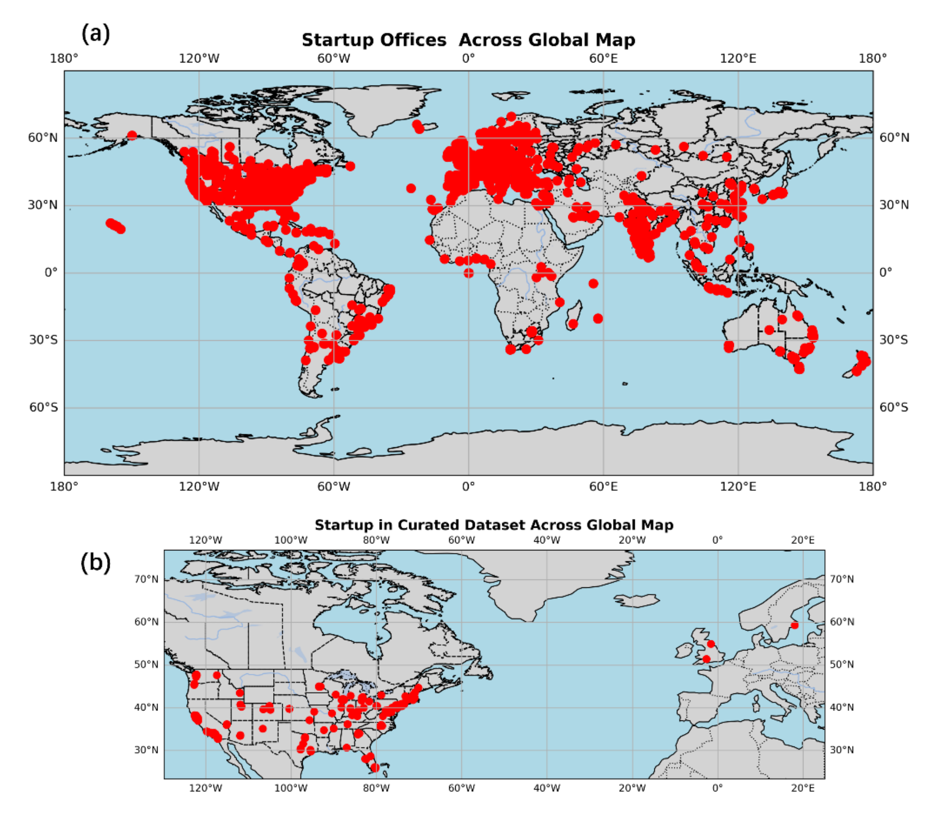


Figure 12. location of startups in the (a) company_profile dataset (112718) in total in terms of location and there are 923 startup in the (b) curated startup dataset.

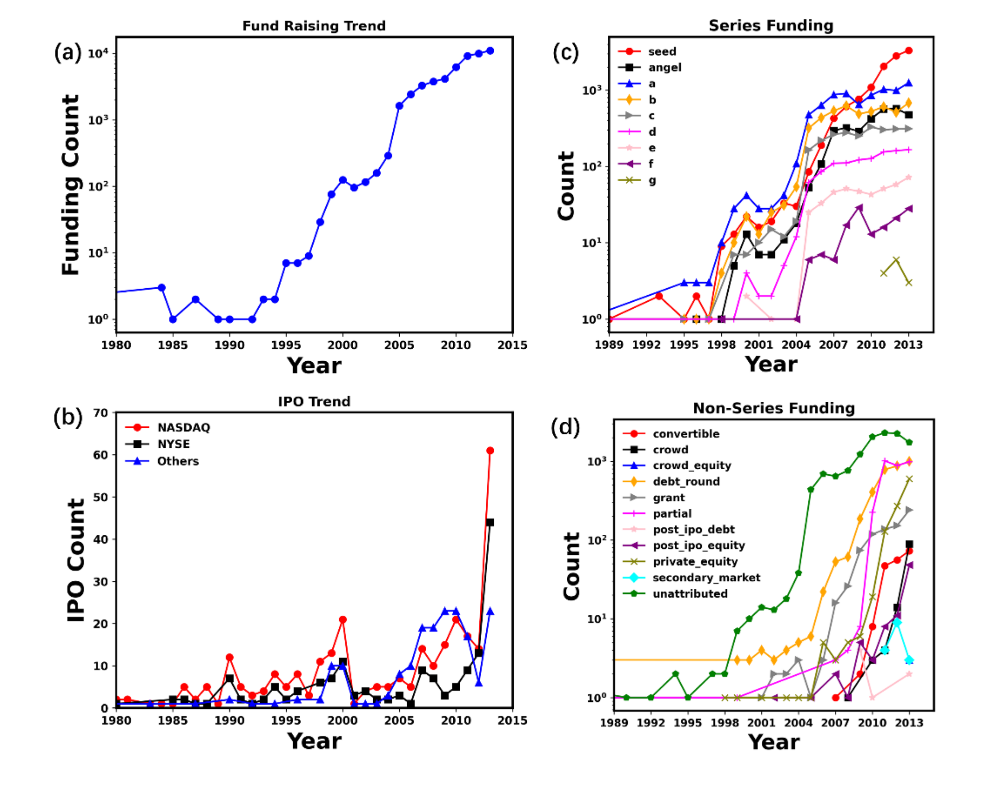


Figure 13. Funding and Investment trend as a function of years. (a) By funding count, (b) by IPO count in different stock exchange (c) categorized based on series funding investment, (d) investment type categorized by non-series form.

Success vs Failure: A Development of Classification Model for Classifying Startups

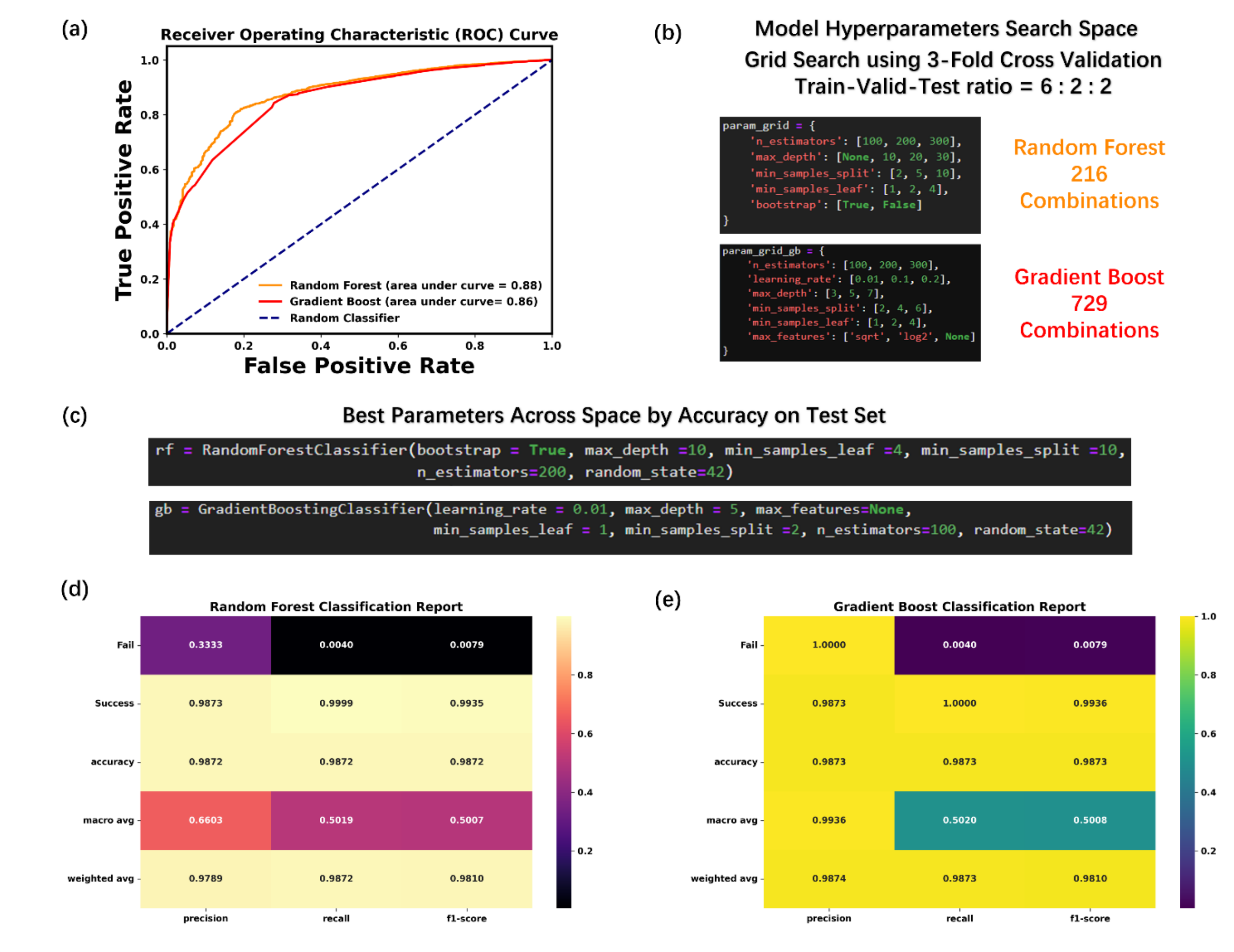
In this section we will proceed with developing and presenting a classification model that determine whether a startup will belong to the successful breed or fail group of startups. The models are then assessed by means of calculating the accuracy, precision, recall, as well as the so-called ROC (Receiver Operating Characteristic) Curve, along with the AUC (area under curve) of the plot. Main results of this model are shown in **Fig. 14.** In building the model, we first success prediction model, we use the object dataset that belong to category of company. A total of ~196,000 data entries were used as raw input. We performed data cleaning, and then encode all the categorical data into numerical values. The next step is then to perform dimension reduction to be used as a useful condensed feature for classification purpose. We opt to use the PCA as before and also tested the use of t-SNE for dimensional reduction, after several parameter test, we decide to stick with PCA and proceed with model development.

In developing the machine learning (ML) classifier model, we make use of the Random Forest（RF） and Gradient Boost (GB) classifier. The pipeline used are the commonly adopted train valid test, with split ratio of 6:2:2. The input dataset were first used to performed a through grid search for optimizing hyperparameters used for prediction. The hyperparameter space we define for this study is shown in **Fig.14 (b)** amounting to 216 possible combinations for the RF model and 729 possible combinations for the GB model. The grid search was performed based on the 3-fold cross validation technique. Each of the iteration takes about 15-30s each totaling to decent several hours calculation on 8 core CPUs. The obtained best model hyperparameters were shown in **Fig.14 (c)** which then used for deployment to obtained the results for ROC curve and classification reports in **Fig.14 (a)**, and **Fig.14 (d-e)**, respectively. In terms of ROC curve and area under curve (1 being perfect classifier), the RF model seemingly slightly outperform the GB model with AUC of 0.88 against 0.86 for GB model. However, looking closer to the classification report we notice that GB model is actually performed better.

The RF model for Class 'Fail' is only 0.3333, suggests that there were many false positives. The low recall value is also indication of high false negatives, consequently the F1-score (which a form of combination from both precision and recall) also is low. In predicting the Class ‘Success’ the model is performing rather well with precision and accuracy > 0.98. The recall score is also close to 1., the balance between precision and recall are also decent, justified by the high F1 score of 0.9935. Nevertheless, in general the RF model is not as good as the GB model.

The GB model have a high precision and accuracy score for both classes ‘Fail’ and ‘Success’, with 1.00, and 9.87, respectively. However, the recall score for class ’Fail’ is only 0.004. For class ‘Success’ the model outperforms the RF model, with higher precision and recall score. Like the RF model previously discussed, this Gradient Boost model is also displaying a bias towards class 'Success'. This notably is an inherent issue deriving from the imbalance in the dataset in which only a small fraction (~1%) of the data actually have an explicit ‘Fail’ label, while the rest are considered as ‘Success’. In the future, improvement can be done by redefining the success criteria and produce a more balanced labelled input data. Other classifier models might also be tried with more hyperparameters to attempts toward getting a more balanced classification model.

In conclusion, the two models developed along with the optimized hyperparameters indicated that with the current dataset (in which majority are considered as "Success") the model have high confidence in predicting startup that is successful or have better chance of success. However, predicting failure with certainty is a bit tougher for the model. Provided the same limitation on the labelled data quantity The optimized Gradient Boost algorithm fare better in predicting absolute case of failure and success compared to the Random Forest model.



```
param_grid = {
    'n_estimators': [100, 200, 300],
    'max_depth': [None, 10, 20, 30],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
    'bootstrap': [True, False]
}
```

```
param_grid_gb = {
    'n_estimators': [100, 200, 300],
    'learning_rate': [0.01, 0.1, 0.2],
    'max_depth': [3, 5, 7],
    'min_samples_split': [2, 4, 6],
    'min_samples_leaf': [1, 2, 4],
    'max_features': ['sqrt', 'log2', None]
}
```

```
rf = RandomForestClassifier(bootstrap = True, max_depth =10, min_samples_leaf =4, min_samples_split =10,
                            n_estimators=200, random_state=42)
```

```
gb = GradientBoostingClassifier(learning_rate = 0.01, max_depth = 5, max_features=None,
                                min_samples_leaf = 1, min_samples_split =2, n_estimators=100, random_state=42)
```

Random Forest Classification Report

	precision	recall	f1-score
Fail	0.3333	0.0040	0.0079
Success	0.9873	0.9999	0.9935
accuracy	0.9872	0.9872	0.9872
macro avg	0.6603	0.5019	0.5007
weighted avg	0.9789	0.9872	0.9810

Gradient Boost Classification Report

	precision	recall	f1-score
Fail	1.0000	0.0040	0.0079
Success	0.9873	1.0000	0.9936
accuracy	0.9873	0.9873	0.9873
macro avg	0.9936	0.5020	0.5008
weighted avg	0.9874	0.9873	0.9810

Figure. 14 Classification Models for Predicting Startup Success. (a) ROC Curve and the area under curve , (b) hyperparameter space defined and used for grid search optimization, (c) Obtained best parameters across the grid space, (d) classification report for the Random Forest Classification model developed, (e) classification report for the Gradient Boost Classification model developed

Discussions on Model Interpretability

To inherently include model interpretability, we purposefully build the model based on a condensed feature out of the many real value features that have straightforward interpretation. The Condensed feature can always be mapped back-and-forth with a certain degree of correlation confidence since its inception is fixed, which were illustrated in previous section and displayed in **Fig.15.**

In terms of clarity, we can consider that the originally diverse and straightforward features (in this case, city, funding rounds, milestones, product count, etc) has been aggregated and is 'merged' into a new feature PC1 and PC2, each of which still retain fixed relationship with each individual original feature. In this example PC1 is completely correlated with funding total USD value and are representative of also funding rounds and current relationship with a given correlation of 1, 0.3, and 0.22 respectively. This then suggest that when used for classification, the basis of this classification still retain a portion of the original multi-dimensional and complex data, but now in terms of single value of PC1 or PC2.

In terms of insightfulness in predicting startup success, a lot more factor can be taken into consideration. In our case, the currently presented of 10 features is actually also already a compromise from the originally more than 39 columns. Which most of it are correlated to each other so to certain degree are information wasted.

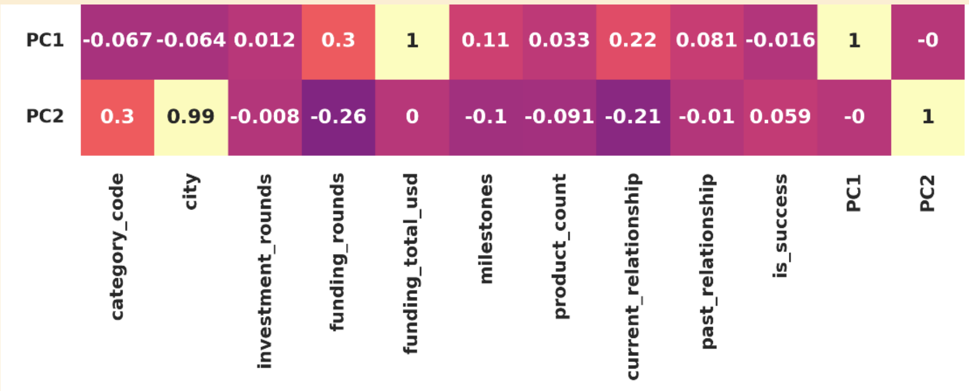


Figure 15. Illustration for the Model Interpretability based on reduced dimension feature.

Methodology

Here we provide a brief additional info regarding the principal component analysis methods utilized in the data analysis and model development.

Principal Component Analysis

Our goal in using the PCA method is to transform the original variables into a new set of uncorrelated variables, called principal components, which are linear combinations of the original variables. In our case we use only the first two principal components. Thus, the first principal component (PC1) captures the maximum variance in the data, the second (PC2) captures the next highest variance. Mathematically, if X is our original data matrix, the principal component Y can be obtained by: $Y = XW$ Here, W is the matrix of eigenvectors of the covariance matrix $X\text{^}TX$ (or correlation matrix). The eigenvectors are ordered by their corresponding eigenvalues in decreasing order. Each column of W represents a principal component. After obtaining the PC1 and PC2 we then append back the newly generated reduced dimension features into the original dataset.

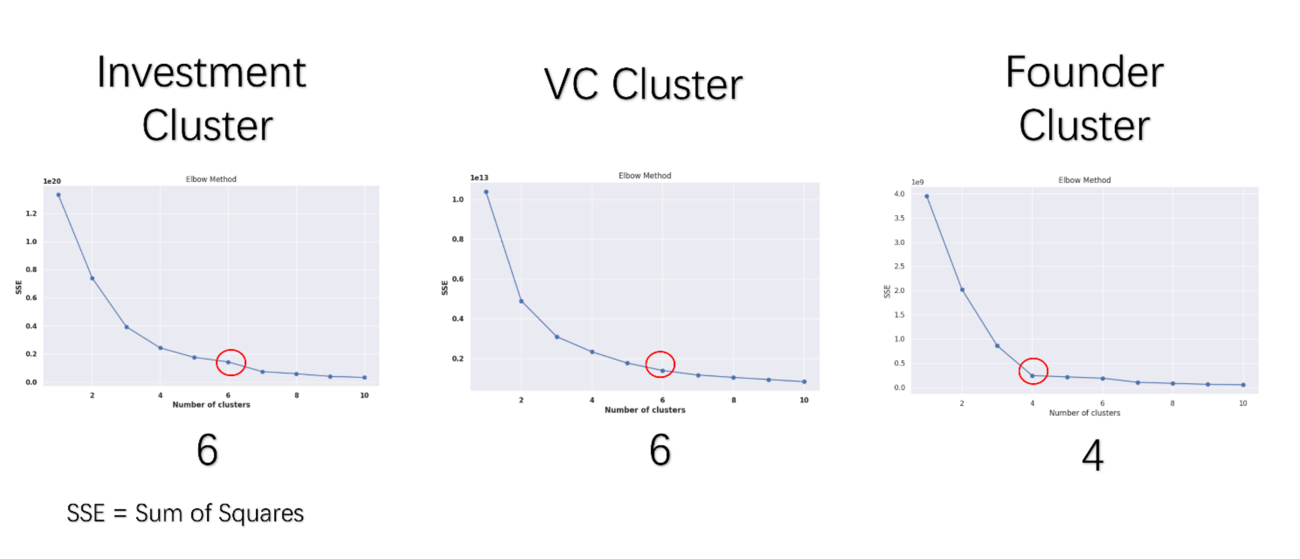
K-means Clustering

In this study, for our clustering approaches, K-means clustering has been used. K-means is a method used to divide a set of data points into k distinct groups or clusters. The goal is to find a balanced partition the data such that the sum of squared (SSE) distances between each point and the centroid of its respective cluster is minimized. The process goes on by first initializing a randomly placed number of centroids, we then calculate the Euclidian distance of data to the existing nearest centroid, this process then done for all data points and is iterated to find a balanced positioning of cluster centroids in which the SSE then minimized. We also present our result for determining the SSE for different number of clusters in the **Appendix A1.**

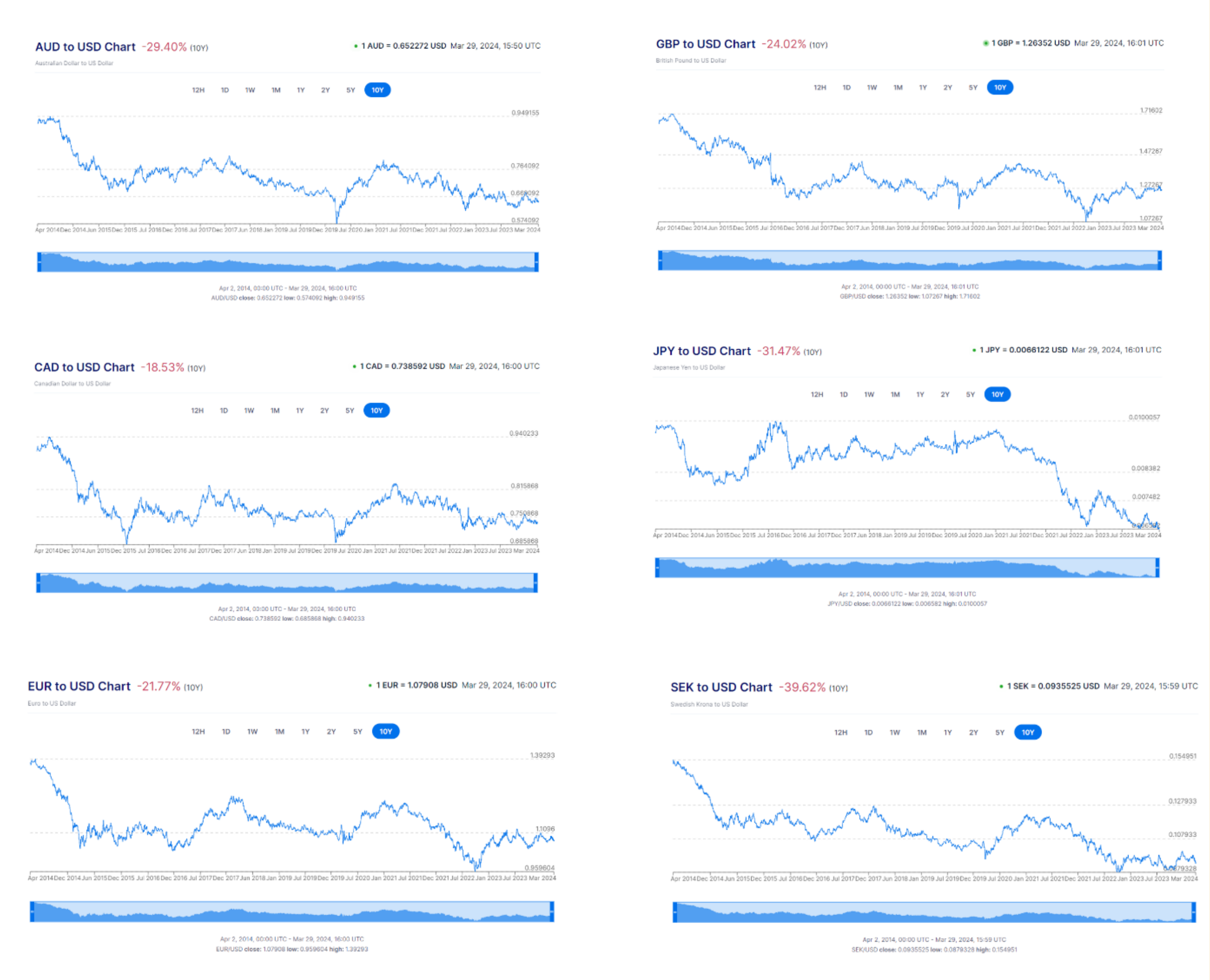
Conclusion

In this work we have deliberately explore the complex data sets with more than million rows entry and tenth of millions of data points. We then carefully performed data cleaning, curation and analysis in order to answer the key questions posed for this dataset. We also proceed with building and deploying clustering model, visualization, and classification model to uncover various intricate details and pattern across the data. One of the key insights in this work highlighted the rate of which fund raising is become increasingly frequent and with fund size that growing ever bigger, along with various innovation of how funds is disbursed appearing every other year. We also discover that among many factors that contribute and correlate to startup success, Computer Science, Stanford University, and San Fransisco Bay region exhibit powerful domination across various spectrum of the datasets. We also introduced the chain of correlation in understanding startup success supported along with a classification model based on Random Forest classifier and Gradient Boost classifier that can be used to predict whether a particular startup belong to a group of possibly success startup or not. All in all, the exploration of the dataset has provided various insights that is useful in understanding the dynamics of the global hyperconnected startup and VC investment ecosystem.

Appendices



Appendix A1. The scan of number of clusters used in developing the k-means clustering.



Appendix A2. Reference Conversion rate from non-USD currencies used in the dataset into USD.