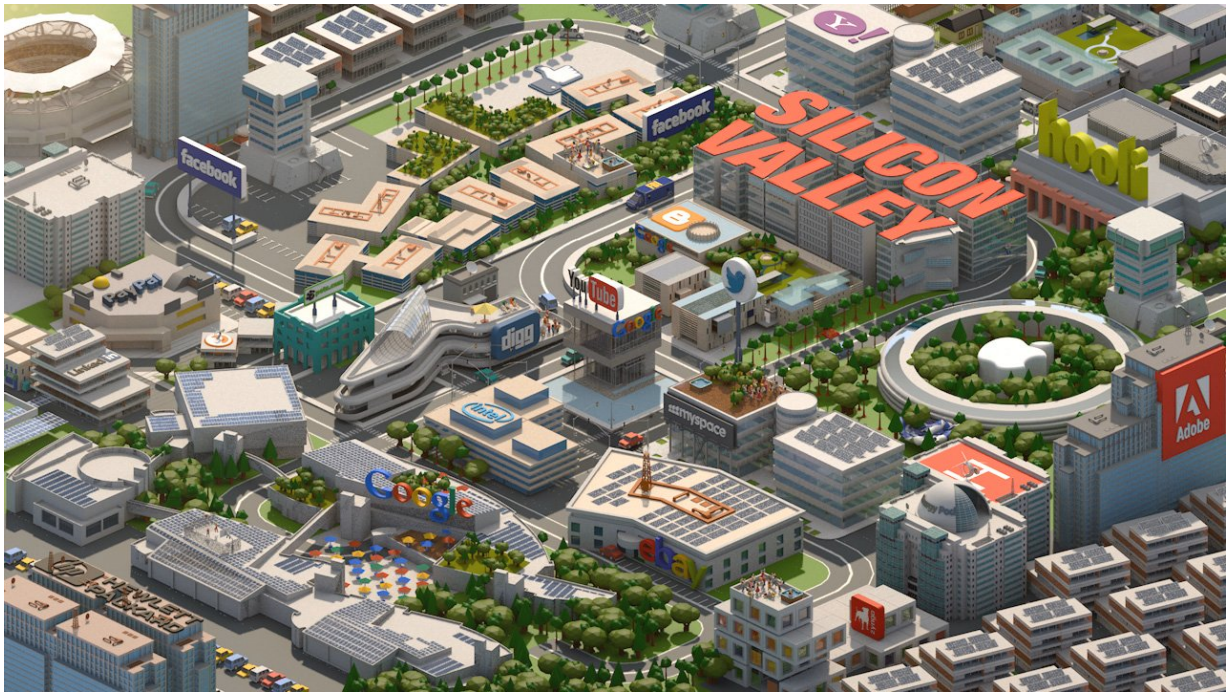
Surviving Silicon Valley's Series A Crunch

Can we predict success in crossing the chasm?



Hitomi Umeki, Paulynn Yu, Vaughn Baker

Stat 701: Modern Data Mining
Professor Linda Zhao
The Wharton School

April 2015

1. Background

1.1 Motivating Questions

A common problem in venture capital and the startup ecosystem for the last several years has been the "Series A Crunch." Thanks in part due to the advent of crowdfunding, an explosion in the number of startup accelerators, and more wealthy individuals looking to invest in private equity, more entrepreneurs raise seed funding than can possibly go on to raise a Series A.[1] As a result, many startups flounder in what has come to be called the "**Valley of Death**" as shown in Figure 1:

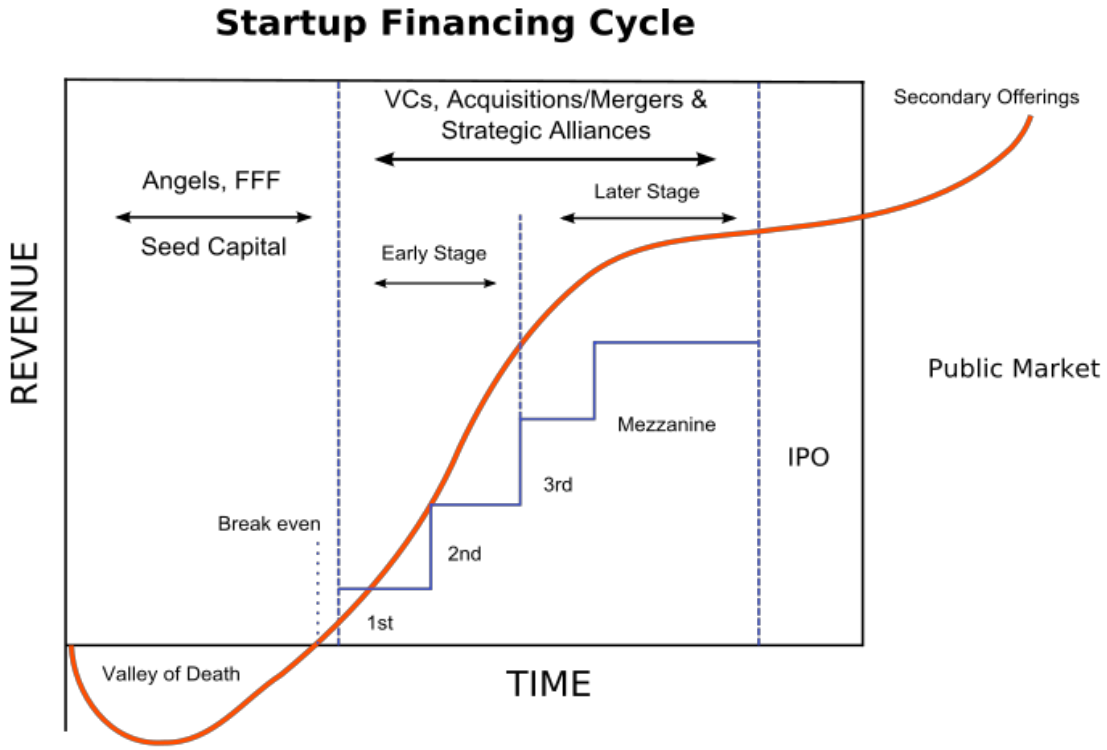


Figure 1: The Funding Lifecycle: many startups fail to ever climb out of the Valley of Death

Our key motivating question is: **can we identify characteristics that make a startup more likely to successfully raise funding?** For example, does the name of the startup matter? Would, as the adage goes, "a rose by any other name smell just as sweet?" Or can the name a company founder chooses impact its ultimate success or failure to raise funding? Can we identify factors that correlate with greater success in making the jump from seed to professional venture capital funding? If so, we might be able to find diamonds in the rough - companies not yet invested in by VCs but worthy of a second look.

These questions are not mere academic musings. Their answers could have real implications for both entrepreneurs and investors:

[1] While there is no formal boundary between Seed and Series A, a startup's first *priced round* of formal venture capital funding is commonly known as a Series A. This is the first round of investment by professional investors who negotiate a specific value for the company and take equity rather than convertible debt in exchange for cash investment.

Implications of our research for Entrepreneurs:

1. How should I name my company?
2. At what age should my company expect to raise seed funding?
3. At what age should we expect to raise our Series A?
4. By comparing 2) and 3): How long should we expect to spend in the Valley of Death?
5. When might it be time to throw in the towel if I haven't yet raised a Series A?

Implications of our research for Investors:

1. Of the hundreds (or thousands) of companies vying for my attention, how can I prioritize them?
2. What does the current universe of eligible Series A startups look like, and which ones should I consider investing in?

For the analysis, we accessed funding data from CrunchBase, the leading repository of startup funding activity, as of April 2015. Considerable effort was required to merge and clean up several separate databases to get the full dataset we needed to do the analysis. A caveat on the data: CrunchBase data is largely self-reported, which could result in inaccurate information (which we did our best to clean out) and reporting bias (successful companies tend to report more than failed ones).

1.2 Industry Landscape

Though venture capital is a relatively new vehicle - only having come about in the last 40 years - its importance in fueling technological innovation has made it a critical bellwether for the future of the American economy.

> "Venture capital in the United States began as a cottage industry, notable in the early years for investments in companies such as Intel, Microsoft, and Apple. In 1990, 100 VC firms were actively investing, with slightly less than $30 billion under management, according to the NVCA. During that era venture capital generated strong, above-market returns, and performance by any measure was good.
>
> What happened? During the peak of the internet boom, in 2000, the number of active firms grew to more than 1,000, and assets under management exceeded $220 billion. VC didn't scale well. As in most asset classes, when the money flooded in, returns fell, and venture capital has not yet recovered. The number of firms and the amount of capital have declined since the boom, though they are both still far above the levels of the early and middle 1990s."
> *- Diane Mulcahy, former venture capitalist, Six Myths About Venture Capitalists HBR May 2013*

As VCs vie for better returns in a more competitive environment, they must find better ways to allocate their time and capital than they have in the past, and that has many turning to data.[2] One of the best sources of data on the venture capital industry is quarterly MoneyTree report published by PriceWaterhouseCoopers in partnership with the National Venture Capital Association.[3] This report offers the best aggregate-level accounting of what's going on in venture capital, but unfortunately yields very few actionable insights at the company-level, or the 'why' behind what's happening. For example, one quick and obvious insight we can gain from these reports is that there may be a location bias for VC fundings as shown in Figure 2.

[2] http://techcrunch.com/2013/06/01/the-quantitative-vc/
[3] Available at https://www.pwcmoneytree.com/

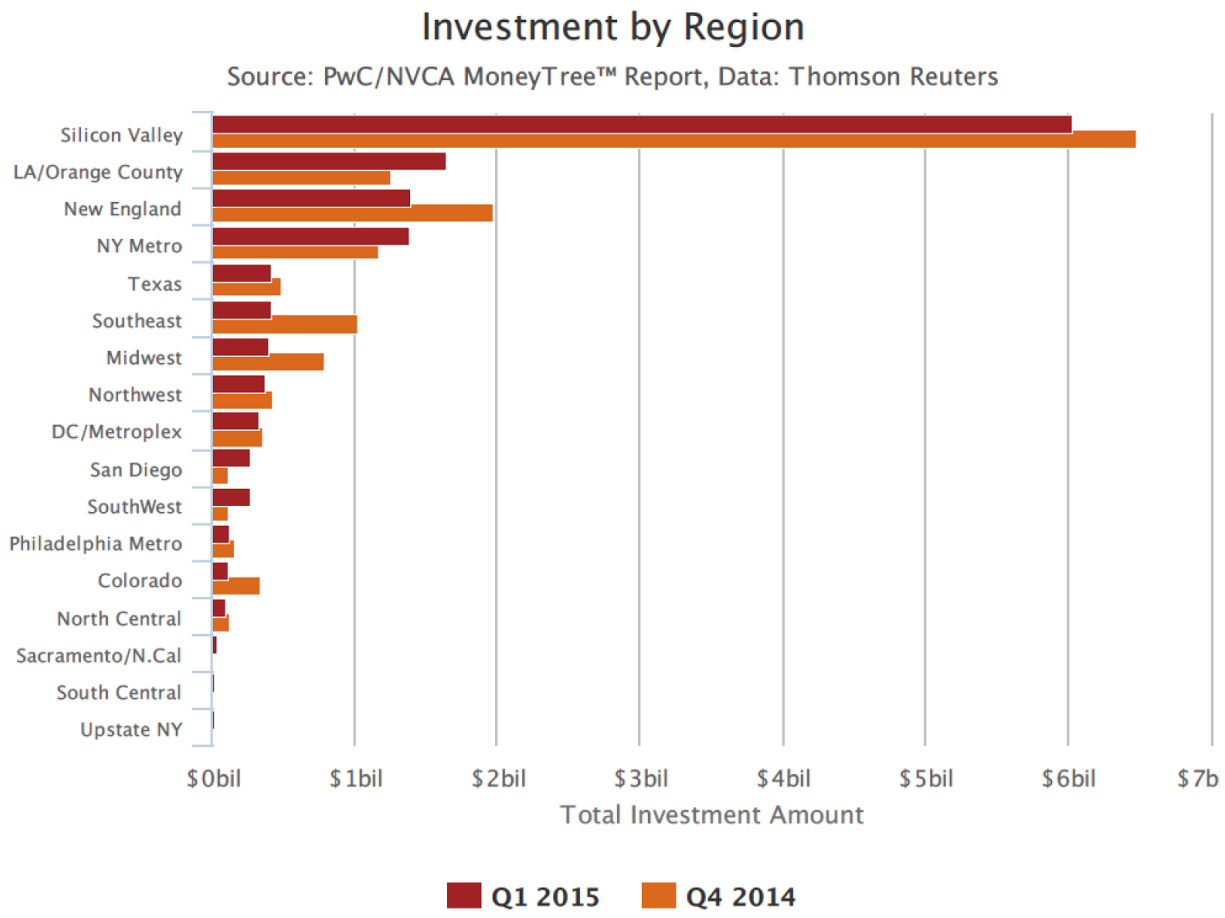


Figure 2: Investment by Region: *Silicon Valley & Boston historically claim >50% of all VC dollars*

New companies, such as CB Insights, are cropping up to fill the void in deeper analysis, but subscriptions cost anywhere from $1,200 to $4,400. Our goal is to see if we can generate some insights using free data sources.

2. Summary Of Data

2.1 Data Cleanup and Extraction

We retrieved data from Crunchbase 2.0[4], the largest and most well-respected public aggregator of information on major companies, investors and executives of the technology ecosystem. The data was received upon request, and in .xls format.

The raw data contained several separate spreadsheets as per the following:

- Investment information for 93,000 investment rounds in crunchbase dating back to the 1980's.
- Company text-based descriptions for 263,568 companies listed on Crunchbase as of April 2015.
- Company demographic information for 53,000 funded companies on Crunchbase.

These sheets were merged and organized into one master file using lookup algorithms in Excel. Their logic was as follows:

- Merged company ***descriptions*** from CrunchBase's Open Data Map (an export of all organizations and individuals listed on Crunchbase with selected description info) into company information tab
- Created a response variable, "***success***" using Excel's `vlookup()` function to indicate whether a given round for a company indicated successful crossing of the "Valley of Death" from pre-VC funding to formal VC funding. (marked as "***success***" = 1(Pass))
- Merged ***funded_date*** information for when companies successfully "Passed" using `vlookup()`
- Assigned a ***unique_id*** for each record by merging "***Company&Name_City***" because there can be different companies with the same name. While it is possible there could be two (or more) different companies with the same name in the same city, we think that would be extremely unlikely.
- Analyzed and grouped ***market*** information to business type (***b2b_b2c***) (sales-driven v.s. marketing-driven). This was done manually at the market level (not individual company level) for ~500 of the roughly 760 markets listed. Some misclassification is known to have occurred but assumed to be approximately uniformly distributed across markets.
- ***founded year***: contains 5,858 NAs and 191 observations with founding dates ranging all the way from the 1898 to the distant future: 8850 AD (talk about investing in the future of technology!). We omitted observations with founding dates prior to 1/1/1950 and more recent than 4/1/2015 as either obsolete or nonsensical.

The merged data set contained a total of 54,581 records. Further cleanup was conducted in R:

- ***country***: data contained records of 100+ countries with U.S. consisting of about ⅓ of the data. There are some 126 country codes listed, but we know that the US accounts for most of the world VC activity - nearly 70% according to Dow Jones VentureSource (Figure 3). International trends can also differ greatly from those at play in the US, so we can both simplify our analysis and make it clearer by focusing on the US firms in our database (n=32,278).

[4] https://www.crunchbase.com

VC investment by region 2013

Region	Invested capital (US$b)	Invested rounds	% change (amount invested)	% change (deals)	% of the global VC activity
United States	33.1	3,480	0.9%	-4.6%	68.2%
Europe	7.4	1,395	19.4%	5.7%	15.3%
Canada	1.0	176	14.4%	23.0%	2.1%
China	3.5	314	-30.0%	20.3%	7.2%
India	1.8	222	12.5%	-2.2%	3.7%
Israel**	1.7	166	54.5%	17.7%	3.5%
Total	48.5	5,753	1.9%	0.2%	100%

Source: Dow Jones VentureSource, 2014
**All-site Israeli companies

Figure 3: VC Investment by Region: US accounts for 68.2 % of VC activity

- ***region***: data contained 1,058 unique regions within U.S. We therefore identified the top 10 regions (SF Bay Area, NYC, Boston, Los Angeles, Seattle, Washington D.C., Chicago, San Diego, Denver and Austin), which consists of about ⅔ of the data, and classified the remaining as "others".
- **Creating new date variables**:
  - ***first_funding_at/last_funding_at*** -> since year is most important, we created a new variable *first_funding_year* and *last_funding_year*
- **Handling NAs**: Random forest does not deal with NAs
  - ***market***: 1199 -> eliminate because market information is essential in analysis
  - ***first_funding year***: 11
  - ***last_funding year***: 4 that overlaps with the NA records for *first_funding_year*

The final 'clean' data set (`crunchbase_final.csv`) is a sample of 26,003 observations.

2.2 Predictor Variables

We will be working with the following predictor variables post cleanup.

- ***unique_id***: (character, added variable) unique identifier of the records. Some companies may have the same names hence indistinguishable. We created a new variable that is a merge of "name" and "city".
- ***name***: (character) company name
- ***market***: (character) the one primary market a company operates in
- ***funding_total_usd***: (numerical) total funding raised to date (as of April 3, 2015) in $USD
- ***status***: (factor) status of the company defined by acquired, closed, ipo, operating
- ***country_code***: (categorical) three-letter country code. USA for all records.
- ***state_code***: (categorical) two letter state codes
- ***region***: (factor) indicates major metropolitan area. (SF Bay Area, NYC, Boston, Los Angeles, Seattle, Washington D.C., Chicago, San Diego, Denver, Austin and others)
- ***funding_rounds***: (numerical) funding rounds received
- ***founded_at***: (date) exact date of founding (YYYY-MM-DD)
- ***founded_year***: (numeric) year of founding (YYYY)
- ***first_funding_at***: (date) first reported round of funding (YYYY-MM-DD)
- ***last_funding_at***: (date) last reported round of investment (YYYY-MM-DD)
- ***first_funding_year***: (numeric, added variable) year of first reported round of funding (YYYY)
- ***first_funding_year***: (numeric, added variable) year of first reported round of funding (YYYY)

- ***description***: (character) free form text description of the company
- ***success***: (factor) binary response variable created to indicate whether or not that company has received venture funding or not
- ***funded_date***: (date) date of successful venture capital funding (YYYY-MM-DD)
- ***funded_amt***: (numeric) size of investment for first 'successful' venture capital round in $USD
- ***b2b_b2c****:* (factor, added variable) categorical variable that represent the business type.
- ***age***: (numeric, added variable) time in years between System Date (today) and ***founded_at*** date
- ***seedTime***: (numeric, added variable) time in years between company founding and first funding
- ***valleyTime***: (numeric, added variable) time in years between first funding and first VC round

2.3 Data Exploration

In order to gain a general understanding, we first explored the data through univariate and bivariate analyses.

Average Age at Funding: Seed

There is clearly still some noise in the data - negative age values make no sense here - but the median age gives us a good sense of the truth as it is less influenced by outliers: the typical company will have raised its first round of seed funding by around 3.6 years.

```
> summary(seedTime)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-48.180   0.489   1.574   3.621   4.261  64.120
```

Among companies that raise outside Seed funding, the 25th percentile are able to do it in just 6 months, while the bottom quartile don't raise till 4.3 years in. In answer to one of our motivating questions, if a company hasn't raised a seed round by year 5, it's probably time to throw in the towel or expect that it will be a self-funded business.

Average Time in the Valley of Death

For our roughly 26,000 observations, just under 2,500 of them - 9.5% - successfully make it to a Series A funding round. To put that in perspective, one has better odds of getting accepted to Harvard Business School (12% acceptance rate in 2011) than of successfully making it through the Valley of Death, and that's *after* having already raised at least one round of Seed funding!

```
> summary(valleyTime)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.000   0.766   1.297   1.661   2.067  29.310   23527
```

The mean time spent by a Seed-funded startup in the Valley of Death is 86 weeks - about 1.7 years. Any company that hasn't raised its Series A by 108 weeks(2.1 years) after their Seed round faces dwindling odds of ever doing so.

Status of Companies

Table 1: Proportion of pass and non-pass companies that are acquired, closed, ipo or operating

	acquired	closed	ipo	operating
pass	0.11933535	0.03889728	0.01510574	0.82666163
non-pass	0.09546793	0.05754091	0.03200520	0.81498597

Our hypothesis is that pass companies are more successful and is the 'goal'. From looking at the status proportion in Table 1, we learn that pass companies are more likely to be acquired and still in operation. They are also less likely to be closed. Interestingly however more non-pass companies IPO than pass companies. This may indicate that these companies that 'leapfrogged' venture funding were able to bootstrap their way to IPO, funding growth either from the founders' own well of financial resources or from internal company cash flows.

Regional Breakdown

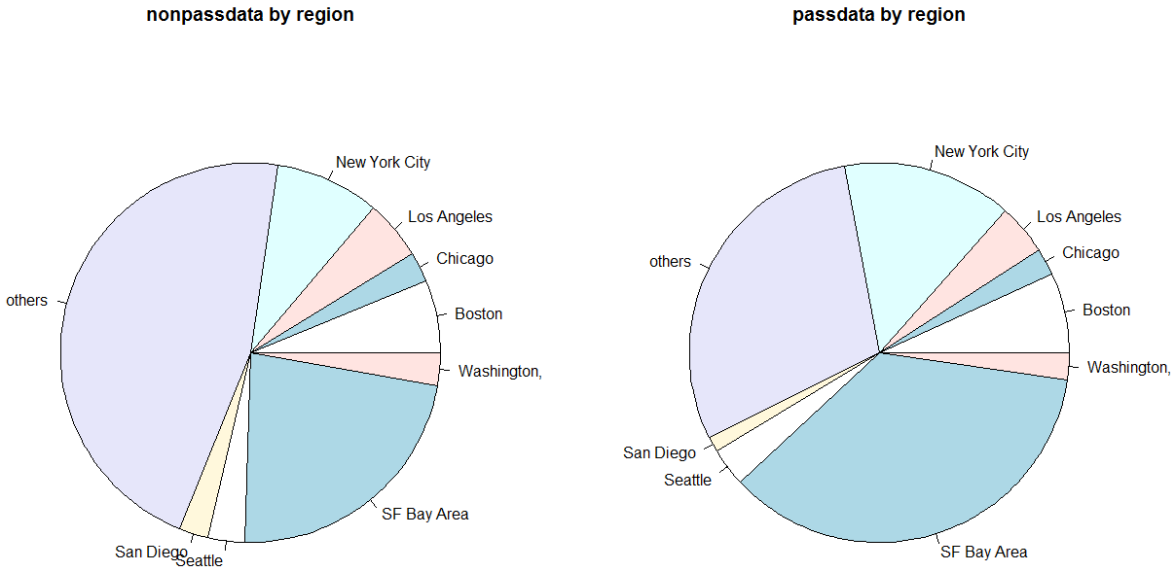


Figure 5: Regional Breakdown of pass and non-pass data

Looking at the companies that pass to receive VC funding, a higher proportion of them are from New York City and SF Bay Area. Non pass data seems pretty dispersed to other locations. This indicates that there are more opportunities for VC funding in these areas compared to other places.

3. Analysis

3.1 Text Classification: How Should I Name my Company?

3.1.1 Objective

Naming a startup is perhaps one of the most challenging initial tasks an entrepreneurs faces. Investors normally make up their minds in less than three minutes during a pitch and even before a pitch, so bad names have the possibility to distract important discussions. In today's search-engine economy, a company's name is also important for discovery and marketing purposes and can be a tremendous competitive advantage. There are various sites[5] that help entrepreneurs name a company as well as blogs[6] that lists recommendations in naming. Below are some general recommendations from various sites:

- use 2 syllables
- keep it to fewer than 10 letters
- more critical for B2C than B2B companies
- numbers make it more interesting
- trendy prefix: "get-" or "app-"
- trendy suffix: "-ly" or ".com"

As startups become more and more competitive, it is probably useful for both investors and entrepreneurs to gain additional insights into the association between company namings and success.

3.1.2 General Stats

We began by trying simply to understand the differences in the names between ***Pass*** and ***Non-Pass*** data. We used Perl to extract out the name, and calculate the average counts of the company name and descriptions.

As suggested from the above recommendations, more successful companies (in our context, success in raising VC funding) seem to have shorter names. We can see this from lower number of average words and letters for pass companies in the table below. Another interesting observation is lower average number of all caps for pass companies. It seems that to get investor attention, the shorter and simpler the company name is better. The more descriptive the number of words (up to 14) the more likely the company is to get funding. On the other hand, length bodes well for company descriptions.

Table 2: Comparison of Company names and descriptions between pass and non-pass

	company name				**company descriptions**	
	average # of words	**average # of letters**	**average # of numerical**	**average # of all caps**	**average # of words**	**average # of letters**
pass	1.410	10.143	0.021	0.015	13.967	84.00
non-pass	1.662	11.875	0.026	0.027	9.885	60.20

[5] http://www.namemesh.com/company-name-generator
[6] http://mashable.com/2012/10/04/startup-naming/

While we cannot use the statistical tools to precisely measure the number of syllables without getting into advanced natural language processing, we were able to use an algorithm to do a fuzzy syllable count, which we user in later analysis.

3.1.3 Wordcloud

To analyze frequent words in company names and description between pass and non-pass companies, we created a word cloud as illustrated in Figure 6. Generally speaking, the more frequent the word is used, the larger and bolder the words will be displayed. It is useful for visually identifying trends and patterns. We used the tm packages to create a corpus (a collection of text documents).

Some words like technologies, inc and systems are predominant in both pass and non-pass companies. However, there are some words that are more dominant in non-pass companies like solutions, medical and pharmaceuticals. While it may be tempting to want to avoid the names solutions and medical in your startup, this may be a case of more healthcare startups existing overall thus increasing its likelihood to appear in non-pass companies. The point is, it is difficult to draw strong conclusions from word cloud but they provide an idea of what are the common names.

Figure 6: Wordclouds of "Pass" data (left) and "Non-Pass" data (right) for company names.

Figure 7 below shows the common words appearing in description of pass and non-pass companies. This is a slightly more interesting visual as there are more underlying words in description than names. Pass companies seem to have fewer prominent words like platform, mobile, users, social, software and technology. Non-pass companies have similar 'big' words (platform, mobile etc.) but have a wider range of trendy words like management, services and solutions. This may indicate that there are some clear 'hype' words like platform but they exist in both categories. We would refrain to conclude that a company gets penalized for words like product and management.

Figure 7: Wordclouds of "Pass" data (left) and "Non-Pass" data (right) for company descriptions.

3.1.4 Random Forest

Because our purpose here is to identify factors associated with company name that correlate to successful funding, we decided to use random forest to gain further insights. Random forests are a popular method in text classification application, and studies have shown them to outperform other classifiers.[7] Random forest is an ensemble classifier which allows us to capture classification results through multiple models with each catching intricacies of a complex model.

We will run random forest with the following predictors, which were all additionally added as variables.

- b2b_b2c: should have interaction effects with names.
- total word count
- total letter count
- has a number
- ends with "-ly"
- ends with ".com"

Response variable is pass for companies who have successfully made it to venture round, and non-pass for those who have not.

Random Forest was run with 100 trees, which proved to be sufficient through the plot of trees to the errors. The sample size was selected to be about 2/3 of the smaller classification size, which was about 1800. We adjust the degree of randomness and the cost ratio by tuning the sample size, to a cost ratio of 10 to 1 (false negatives are 10 times more costly than false positives). In a perfect world, both the investors and entrepreneurs will be most concerned with avoiding false negatives because investors would not want to

[7] Segnini A. and Tayou J. Random Forest and Text Mining.

miss an opportunity for investment, and entrepreneurs would not want to give up an opportunity to make it big.

Output

Through the Random Forest algorithm, out-of-bag data is used to construct the confusion table as shown in Table 3, in which rows represent the truth, while columns represent the classifications.

Table 3: Confusion Matrix

```
Confusion matrix:
      0     1 class.error
0 14084 10323   0.4229524
1   939  1250   0.4289630
```

The model error is 39.8% for the *non-pass*, while 42.9% for the *pass* data. This means that within the startups that succeed and make it to a venture round, the algorithm correctly classifies 57.8% of the time, while for those that does not pass, the model correctly classifies 57.1% of the time. Use error is 6% for the *non-pass* and 89.1% for *pass*. This means that when startups are forecasted to not succeed, it is correct about 86% of the time, and when startups are forecasted to succeed, the forecast is incorrect 88% of the time. This high level of inaccuracy is because we are assigning asymmetric costs, which should be acceptable due to the difference in opportunity and lost costs.

Interpretation of variable importance plots

Figure 8 illustrates the variable importance plots for both categorical outcomes of *pass* and *non-pass*. Through these plots, we can gain insights on the significance of each variable in classifying outcomes. The plot is an average measurement of the decrease in accuracy when a certain predictor is shuffled and prohibited from making contribution to a forecast. The plot exclude the interaction effects, and therefore the sum of the declines does not equal the model accuracy.

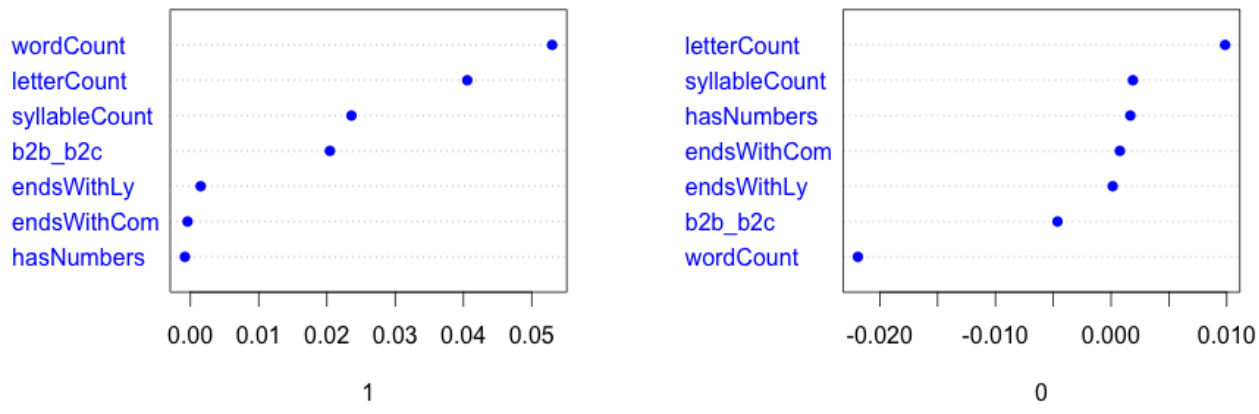


Figure 8: Variable Importance Plots

From Table3, we saw that *pass (success=1)* is correctly forecasted about 57.8% of the time. Combining that with the results from the variable importance plot, we see that the accuracy drops by 5.5% to 52.3% if

wordCount is shuffled, drops another 4% to 48.3% if *letterCount* is shuffled, and drops 2.5% to 45.8% if *syllableCount* is shuffled. The other three variables do not seem to have an affect on the model accuracy hence we can conclude that in this analysis, we were able to see how the content of the name

Partial Dependence Plots

Partial dependence plots allow us to understand the relationship between a predictor and the response given the other predictors are held constant. Figure 9 shows the partial dependence plots in a probabilities scale. Because this is a binary outcome, the results for the other class should be reciprocal.

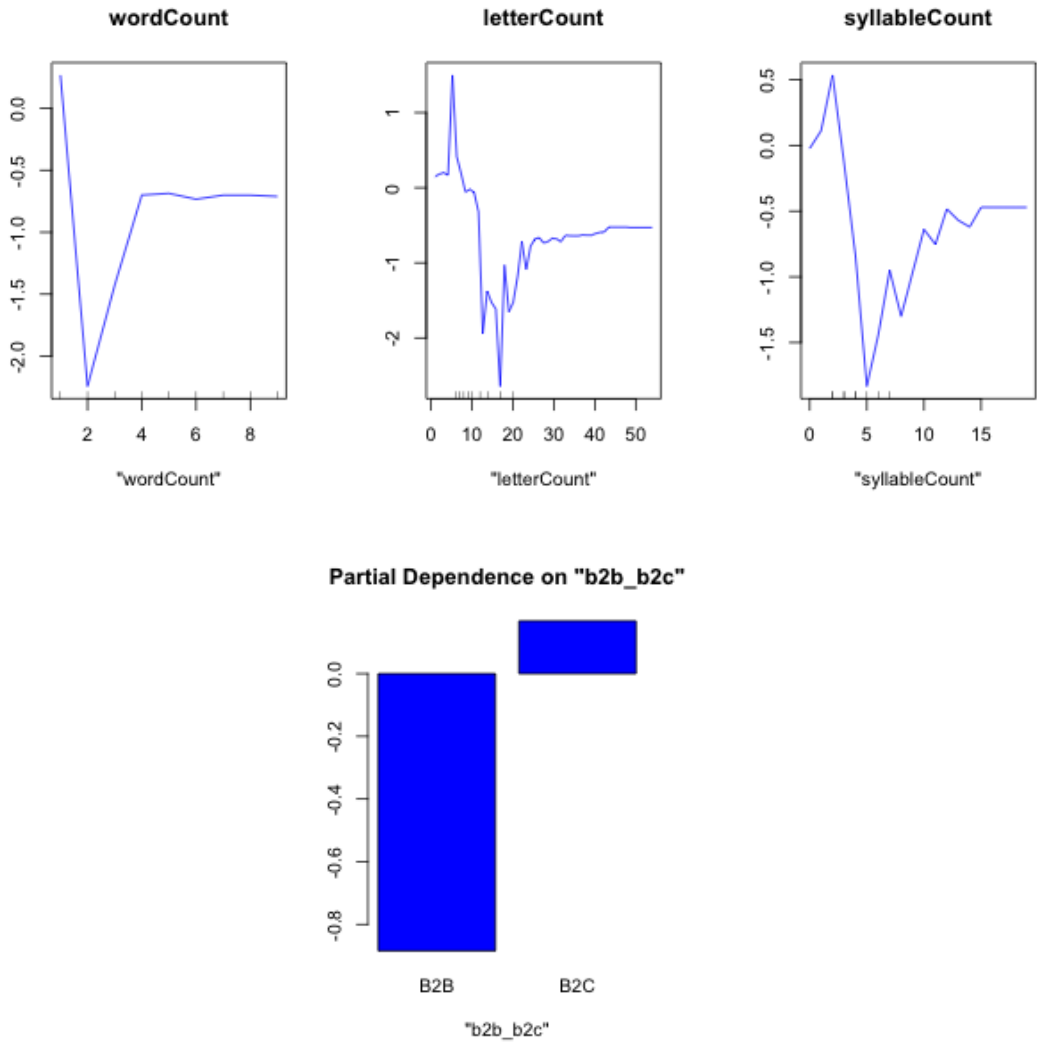


Figure 9: Partial Dependence Plots

We can understand the following for each variable:

- ***wordCount:*** most of the data is between 1-2, and anything beyond that is out of range. Odds of *pass* is high when *wordCount* is 1, and drops significantly when word count is 2.

- ***letterCount***: most of the data is between 5-15, and anything beyond that is out of range. Odds of *pass* is highest when *letterCount* is between 5-8, then drops significantly from 8 to 18.
- ***syllableCount***: most of the data is between 2-7, and anything beyond that is out of range. Odds of *pass* is highest when *syllableCount* is 2-3, then drops significantly.
- ***b2b_b2c***: b2c has more of an effect to company's success. This variable should more likely have an interaction effect with the other predictor variables as name matters more to B2C companies than B2B companies.

Conclusions

In understanding the associations between startups and their names, we looked at how different characteristics of the name is associated with the success of a company. As with all things, trends in startup names change over time. We would not argue that entrepreneurs choose their company's name based on data over long-term vision or alignment with brand. Through this analysis we find that structural features of a company's name - features such as the number of words, syllables, and letters - are more strongly correlated with company success in fundraising than content-based characteristics (e.g. ending in -ly, .com, or containing numbers). The findings for name structure match anecdotal suggestions made by several prominent blogs[8] (but not with supporting data). Some of these suggestions include limiting a company's name to less than ten letters and 2 syllables. We gained additional insights where we observed significant drops in the probabilities of a company's success when a company name has more than one word, more than 10 letters and more than two syllables.

3.2 Philly's Startup Scene: What Drives Investment and Can We Identify Good Investment Candidates?

3.2.1 Objective

Having looked at some national trends, we'll bring things home to Philly. As graduate students at Wharton we know that Philly's tech scene is not as robust as Silicon Valley or Boston. Nevertheless, the City of Philadelphia, local universities, and local venture funds have made a concerted effort to foster entrepreneurship in the area, and anecdotal evidence seems to indicate their efforts are working.

So, what trends are at play in Philadelphia's emerging tech scene? And, can we help predict which companies might or should receive venture funding based on local market factors? Our objective now is to answer these questions.

The premise of our analysis is admittedly a bit contrived: in order to demonstrate modern techniques such as lasso for handling high dimensionality in a data set, we needed to restrict the number of observations to be much smaller than the number of predictor variables.

3.2.2 LASSO

We chose LASSO (Least Absolute Shrinkage and Selection Operator) as our initial approach in understanding important variables that may help in the prediction of companies that successfully make it to

[8] http://mashable.com/2012/10/03/startup-naming-importance/

the venture round. There are 754 unique markets within the variable markets, which brings us to to a total number of variables of almost 760+ while we only have 574 records of data that represent Philadelphia. In class, we learned that LASSO is penalized regression method which is useful in performing variable selections when the sample size is small and the number of prediction variables are large. In contrast to the Ridge regression, it allows for some coefficients to have a coefficient value of zero. Upon learning that correlations have the tendency to affect Lasso prediction results[9], we first analyzed the correlation plots as shown in Figure 10.

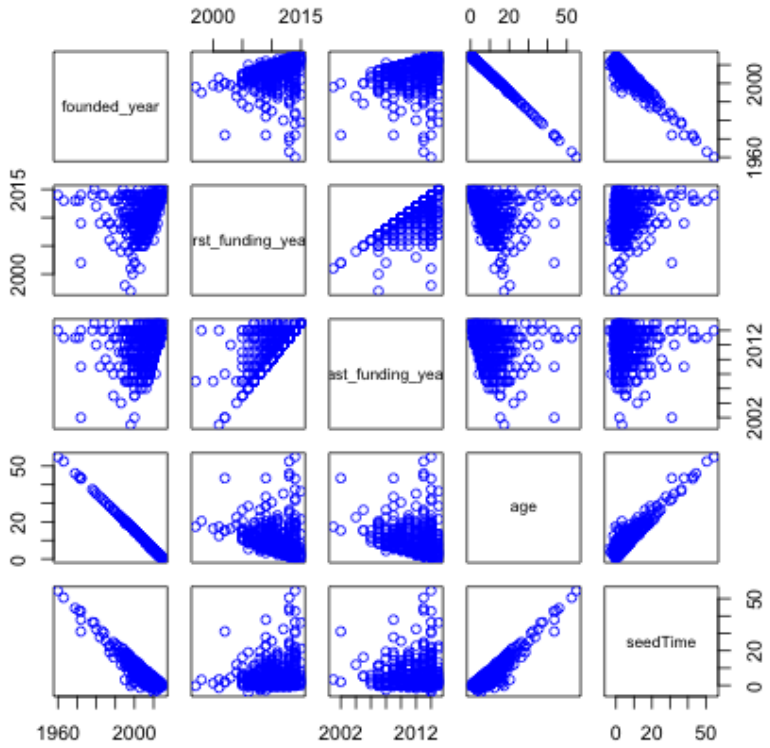


Figure 10: LASSO Analysis Output: Plot of Lambdas

We found strong correlation between time related variables such as age, seedTime and founded_year. We selected age out of the three because the variable was most interpretable. Therefore we will be running lasso with:

- **market -> 754 unique categories**
- funding_total_usd
- funding_rounds
- first_funding_year
- seedTime
- wordCount
- letterCount

[9] Hebiri, M. and Lederer J.C. (2012) How Correlation Influences Lasso Prediction.

In understanding which lambda to use, we selected lambda so that cross validation prediction error is minimized. Lambda is a tuning parameter which allows us to penalize the betas and is a biased trade off (when lambda is too big, we penalize beta too much). The plots of the lambdas are shown in Figure 10.

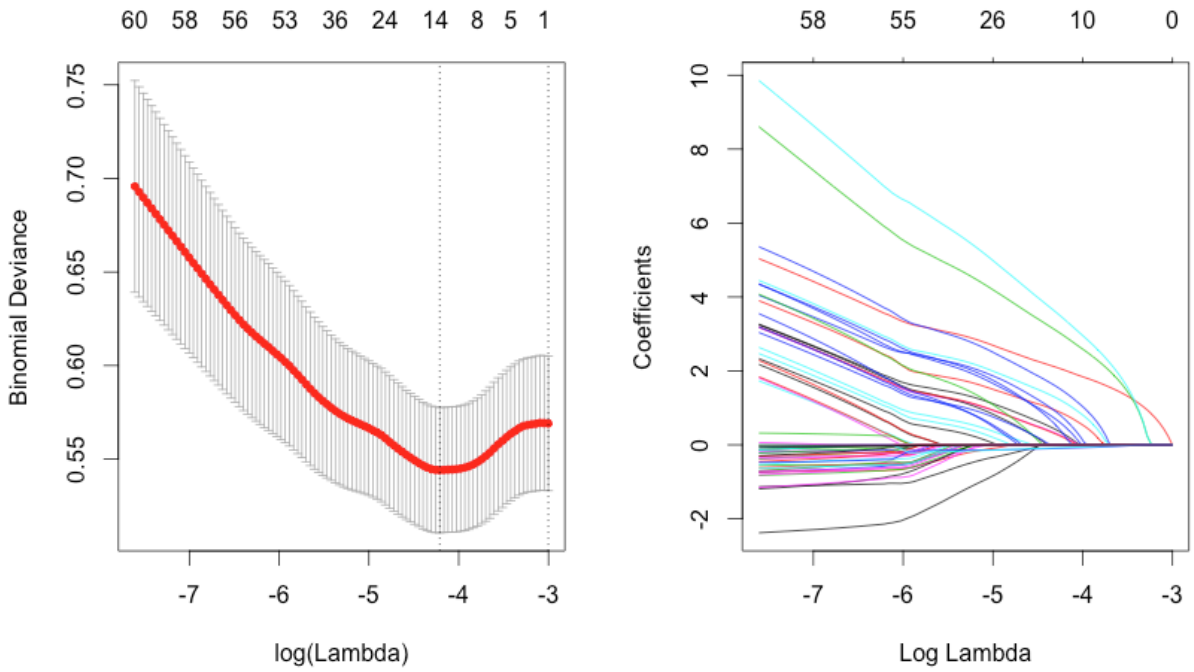


Figure 10: LASSO Analysis Output: Plot of Lambdas

We then ran LASSO with different tuning parameters and different set of matrices and concluded with the following lists of coefficients. We used the combination of these coefficients to run a linear model in an attempt to understand the relationship.

```
> coef.min
                (Intercept)          marketAdvertising           marketAnalytics
              210.476707553                0.022897349                1.885560772
              marketAndroid      marketCloud Computing             marketContent
                2.731023596                0.652735013                0.262960024
            marketInternet      marketMedical Devices      marketNanotechnology
                0.927132001                3.031068896                0.002942715
marketPredictive Analytics        first_funding_year                        age
                1.097400508               -0.105824107               -0.025856759

> coef.min
         (Intercept)           marketAdvertising              marketAnalytics
               193.20813319                0.45481154                2.11317490
             marketAndroid      marketCloud Computing             marketContent
                 3.18256110                1.01960054                0.91770889
          marketE-Commerce  marketEnterprise Software marketInformation Technology
                 0.33967898                0.28875518                0.56997207
            marketInternet      marketMedical Devices      marketNanotechnology
```

```
                1.34022522                  3.57941622                  0.64612049
 marketPredictive Analytics                founded_year          first_funding_year
                1.59133182                  0.05261532                 -0.14990658
```

3.2.3 Logistic Regression

We take the dramatically reduced set of explanatory variables (X_i) & categories the from LASSO analysis aboveand run them through the generalized linear model using logistic regression to yield directly interpretable (β_i) coefficients.

```
data was split into 1:5
Call:
glm(formula = success ~ marketAnalytics + marketAdvertising +
    marketAndroid + marketCloudComputing + marketContent + marketECommerce +
    marketEnterpriseSoftware + marketInformationTechnology +
    marketInternet + marketMedicalDevices + marketNanotechnology +
    marketPredictiveAnalytics + first_funding_year + age, family = "binomial",
    data = data.phl)
```

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	**-2.38926**	**0.30577**	**-7.814**	**5.54e-15**	*******
marketAnalytics	**2.81215**	**0.74134**	**3.793**	**0.000149**	*******
marketAdvertising	**1.47589**	**0.83254**	**1.773**	**0.076269**	**.**
marketAndroid	18.19046	1455.39755	0.012	0.990028	
marketCloudComputing	**3.00827**	**1.44534**	**2.081**	**0.037401**	*****
marketContent	1.93030	1.24535	1.550	0.121140	
marketECommerce	**1.42919**	**0.67433**	**2.119**	**0.034054**	*****
marketEnterpriseSoftware	**1.21570**	**0.58373**	**2.083**	**0.037285**	*****
marketInformationTechnology	1.50933	1.17838	1.281	0.200247	
marketInternet	NA	NA	NA	NA	
marketMedicalDevices	18.15108	1455.39755	0.012	0.990049	
marketNanotechnology	**2.25319**	**1.24988**	**1.803**	**0.071431**	**.**
marketPredictiveAnalytics	**2.87600**	**1.42858**	**2.013**	**0.044095**	*****
age	**-0.05850**	**0.03258**	**-1.796**	**0.072533**	**.**

```
---
```

However, kicking out variables one by one from the least significant, we were only left with a very simple model as per below. This is probably due to the fact that we have such limited data sets and even few data records were highly influential in driving regression results. We understand from the below output that highly significant coefficients were age and market where the value is “Analytics”.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.99783	0.26864	-7.437	1.03e-13	***
marketAnalytics	2.45065	0.73286	3.344	0.000826	***
age	-0.06267	0.03059	-2.049	0.040479	*

3.2.4 Conclusion

Our study suggests Philadelphia is a burgeoning hotbed for analytics and enterprise IT companies. Companies in these markets that have raised seed funding are most likely to successfully go on to raise a Series A, and among them analytics startups especially so.

But, we find that there is no grace in old age for local tech startups seeking VC funding: two of the strongest correlations in our logistic classification model are with ***age*** and ***first_funding_year***. The older a company is and the longer it has been since they raised their Seed funding, the less likely it becomes that they will survive the Valley of Death. Because we use a logistic model, the coefficients cannot be directly interpreted as % increase or decrease in probability, but the model indicates these are highly statistically significant factors.

So, our advice to local entrepreneurs wanting to stay in Philly and looking to raise venture funding:

1) we hope you're doing analytics
2) to quote the poet Dylan Thomas: "Do not go gentle into that good night but rage, rage against the dying of the light."

3.2.5 Possible Further Analysis

Investors may be interested in identifying which companies should have received venture funding but have not yet. These would be companies which have so far been passed over for investment show good promise of success in going on to raise a Series A round. To do this, we could look for 'false negatives' in a confusion matrix of predicted vs. actual where the response variable is ***success***. The method we followed in the above analysis of limiting our sample to only companies in Pennsylvania yielded a thin pool of companies that successfully raised venture funding which resulted in a high degree of bias: for any of the 750+ markets, there are only 1-2 success story companies. As a result, splitting the data into training and testing data sets resulted in dramatic changes to significance of factors - removing 1/10 to 1/5th of the observations would mean taking out the only success story in a given sector within Pennsylvania.

A follow-up analysis with a focus on identifying investment candidates could attempt to build a better predictive model (ours focused on inference) using the full national dataset (rather than restricting to Pennsylvania alone as we did to demonstrate lasso). Using the full national data would provide many more 'successful' observations, which would in turn reduce the problem of extreme bias.

The predictor variables for this analysis could include: region, market, age, whether the company is B2B or B2C, # of rounds raised to date, total funding raised to date, and several factors related to the company name: # of words, # of letters, # of syllables. Interesting response variables might be either a binary prediction for outcome: whether a company is likely to raise funding or not, or a linear prediction of how much funding the company has raised. Both of these could be compared against actual outcomes to identify candidates who have been overlooked for investment or underinvested in (possibly highly capital efficient teams).

We are very excited about this dataset and the potential to apply modern data mining techniques to it, and there is seemingly no end to interesting questions which might be answered in this space, so we have included a list of other potentially interesting managerial questions for future analysis in the Appendix.

Appendix

Output

```
> summary(working)
  unique_id            name                          market      funding_total_usd
 Length:31851       Length:31851       Software           : 3320   Min.   :    1
 Class :character   Class :character   Biotechnology      : 2899   1st Qu.: 1424
 Mode  :character   Mode  :character   Mobile             : 1237   Median : 5998
                                       Curated Web        : 1048   Mean   : 6344
                                       Enterprise Software :  975   3rd Qu.:10828
                                      (Other)             :21225   Max.   :16103
                                      NA's                : 1147

      status      country_code   state_code              region      funding_rounds
 acquired : 3104   USA:31851    CA     :11018   others       :14286   Min.   : 1.000
 closed   : 1784                NY     : 3334   SF Bay Area  : 7563   1st Qu.: 1.000
 ipo      :  975                MA     : 2118   New York City: 2968   Median : 1.000
 operating:25988                TX     : 1616   Boston       : 1979   Mean   : 1.881
                                FL     : 1065   Los Angeles  : 1595   3rd Qu.: 2.000
                                (Other):12664   Seattle      : 1008   Max.   :19.000
                                NA's   :   36   (Other)      : 2452

  founded_at     founded_year  first_funding_at     last_funding_at
 1/1/11 : 1472   Min.   :1902   Min.   :1973-04-15   Min.   :1973-04-15
 1/1/12 : 1391   1st Qu.:2005   1st Qu.:2009-09-09   1st Qu.:2011-01-07
 1/1/10 : 1323   Median :2009   Median :2011-12-21   Median :2013-04-01
 1/1/09 : 1173   Mean   :2007   Mean   :2011-04-24   Mean   :2012-05-13
 1/1/13 : 1114   3rd Qu.:2012   3rd Qu.:2013-10-02   3rd Qu.:2014-06-01
 (Other):19678   Max.   :2015   Max.   :2060-01-01   Max.   :2060-01-01
 NA's   : 5700   NA's   :5774

 description         success    funded_date    first_funding_year last_funding_year
 Length:31851       0:29204   1/1/12 :   12   Min.   :1973       Min.   :1973
 Class :character   1: 2647   4/1/13 :   12   1st Qu.:2009       1st Qu.:2011
 Mode  :character             1/1/08 :   11   Median :2011       Median :2013
                              7/1/07 :   10   Mean   :2011       Mean   :2012
                              7/1/14 :   10   3rd Qu.:2013       3rd Qu.:2014
                              (Other): 2592   Max.   :2060       Max.   :2060
                              NA's   :29204
```

Other managerial questions of interest for further analysis:

1. Does amount of funding raised before first venture round impact a company's ability to successfully raise VC funding (or amount)?
2. Which regions have the best "batting average" of companies succeeding in raising a Series A? (On a per capita basis. We suspect this is San Francisco)
3. What markets are "hot" at seed and venture levels right now? Are they they same or do we see different trends in each?
4. How long does it take to get to a series A on average? How about by market sector, geography?
5. Given that a company has raised Venture funding, what determines HOW MUCH funding they go on to raise?
6. Does the perception of herd mentality among investors bear out in the data? (i.e. do we see loads of funding pouring into different 'hot' sectors each year)