Start Up Funding

Hiroki Hirayama

03/05/2020

'Startup Ecosystem - Funding Patterns'

Summary :

The aim of this project is to build a model was to predict whether a company will survive, get acquired or is no longer operating based on funding data. Startup trends were analysed through drawing up graphs to investigate trends.

Upon further examination, the data is imbalanced. Regardless, dimensionality reduction algorithm (PCA) and random tree classification was conducted. The results indicated that the nature of the data led to the inability to run a principle component analysis nor random tree classification algorithm - the classification algorithm implemented indicated that it performed no better than just precting a fixed outcome of the majority factor.

Whilst this is not the outcome hoped for, lessons were drawn from this project - both on how to pick datasets and about the start-up ecosystem.

Initial Setup

The following are the packages used : Tidyverse,dplyr,ggplot2,stats,ggthemes,stringr,gridExtra,caret, lubridate,nnet,psych,randomForest

Data : The dataset has been obtained from Kaggle @ https://www.kaggle.com/arindam235/startup-investments-crunchbase Thank you to Crunchbase (www.crunchbase.com) and Kaggle user 'Andy_M' for kindly providing and making the data set freely available.

Data Pre-processing and Feature Engineering

```r
head(datasource)
```

```
##                          permalink                name
## 1           /organization/waywire            #waywire
## 2 /organization/tv-communications &TV Communications
## 3   /organization/rock-your-paper  'Rock' Your Paper
## 4  /organization/in-touch-network  (In)Touch Network
## 5  /organization/r-ranch-and-mine -R- Ranch and Mine
## 6      /organization/club-domains      .Club Domains
##                   homepage_url
## 1        http://www.waywire.com
## 2          http://enjoyandtv.com
## 3  http://www.rockyourpaper.org
## 4 http://www.InTouchNetwork.com
## 5
## 6              http://nic.club/
##                                                           category_list
```

```
## 1                                  |Entertainment|Politics|Social Media|News|
## 2                                                                     |Games|
## 3                                                    |Publishing|Education|
## 4 |Electronics|Guides|Coffee|Restaurants|Music|iPhone|Apps|Mobile|iOS|E-Commerce|
## 5                                              |Tourism|Entertainment|Games|
## 6                                                                  |Software|
##          market funding_total_usd    status country_code state_code
## 1          News        17,50,000   acquired          USA         NY
## 2         Games        40,00,000  operating          USA         CA
## 3    Publishing           40,000  operating          EST
## 4   Electronics        15,00,000  operating          GBR
## 5       Tourism           60,000  operating          USA         TX
## 6      Software        70,00,000                     USA         FL
##           region         city funding_rounds founded_at founded_month
## 1  New York City     New York              1 2012-06-01       2012-06
## 2    Los Angeles  Los Angeles              2
## 3        Tallinn      Tallinn              1 2012-10-26       2012-10
## 4         London       London              1 2011-04-01       2011-04
## 5         Dallas   Fort Worth              2 2014-01-01       2014-01
## 6 Ft. Lauderdale Oakland Park              1 2011-10-10       2011-10
##   founded_quarter founded_year first_funding_at last_funding_at    seed venture
## 1         2012-Q2         2012       2012-06-30      2012-06-30 1750000   0e+00
## 2                           NA       2010-06-04      2010-09-23       0   4e+06
## 3         2012-Q4         2012       2012-08-09      2012-08-09   40000   0e+00
## 4         2011-Q2         2011       2011-04-01      2011-04-01 1500000   0e+00
## 5         2014-Q1         2014       2014-08-17      2014-09-26       0   0e+00
## 6         2011-Q4         2011       2013-05-31      2013-05-31       0   7e+06
##   equity_crowdfunding undisclosed convertible_note debt_financing angel grant
## 1                   0           0                0              0     0     0
## 2                   0           0                0              0     0     0
## 3                   0           0                0              0     0     0
## 4                   0           0                0              0     0     0
## 5               60000           0                0              0     0     0
## 6                   0           0                0              0     0     0
##   private_equity post_ipo_equity post_ipo_debt secondary_market
## 1              0               0             0                0
## 2              0               0             0                0
## 3              0               0             0                0
## 4              0               0             0                0
## 5              0               0             0                0
## 6              0               0             0                0
##   product_crowdfunding round_A round_B round_C round_D round_E round_F round_G
## 1                    0       0       0       0       0       0       0       0
## 2                    0       0       0       0       0       0       0       0
## 3                    0       0       0       0       0       0       0       0
## 4                    0       0       0       0       0       0       0       0
## 5                    0       0       0       0       0       0       0       0
## 6                    0       0 7000000       0       0       0       0       0
##   round_H
## 1       0
## 2       0
## 3       0
## 4       0
## 5       0
```

```
## 6       0
```

```r
str(datasource)
```

```
## 'data.frame':    54294 obs. of  39 variables:
##  $ permalink           : Factor w/ 49437 levels "","/organization/-qounter",..: 47060 44589 36160 202
##  $ name                : Factor w/ 49352 levels "","-R- Ranch and Mine",..: 18 17 12 13 2 10 11 21 22
##  $ homepage_url        : Factor w/ 45851 levels "","http://??????????-??????.??",..: 43585 4171 3703
##  $ category_list       : Factor w/ 16676 levels "","|3D Printing|",..: 5173 6143 11656 4665 15392 141
##  $ market              : Factor w/ 754 levels ""," 3D "," 3D Printing ",..: 471 279 550 213 689 643 9
##  $ funding_total_usd   : Factor w/ 14618 levels ""," -   "," 1 ",..: 3980 9443 9448 3484 11818 12802
##  $ status              : Factor w/ 4 levels "","acquired",..: 2 4 4 4 4 1 3 4 4 4 ...
##  $ country_code        : Factor w/ 116 levels "","ALB","ARE",..: 112 112 37 40 112 112 4 1 112 45 ..
##  $ state_code          : Factor w/ 62 levels "","AB","AK","AL",..: 42 8 1 1 55 13 1 1 18 1 ...
##  $ region              : Factor w/ 1090 levels "","\xc7an","\xc9vry",..: 693 572 952 570 251 347 147
##  $ city                : Factor w/ 4189 levels "","'s-hertogenbosch",..: 2560 2112 3657 2099 1245 267
##  $ funding_rounds      : int  1 2 1 1 2 1 1 1 1 1 ...
##  $ founded_at          : Factor w/ 3370 levels "","1636-09-08",..: 2576 1 2715 2175 3112 2348 1 1055
##  $ founded_month       : Factor w/ 421 levels "","1902-01","1903-01",..: 391 1 395 377 410 383 1 326
##  $ founded_quarter     : Factor w/ 219 levels "","1902-Q1","1903-Q1",..: 209 1 211 205 216 207 1 188
##  $ founded_year        : int  2012 NA 2012 2011 2014 2011 NA 2007 2010 NA ...
##  $ first_funding_at    : Factor w/ 3915 levels "","0001-05-14",..: 3035 2309 3075 2595 3807 3366 1200
##  $ last_funding_at     : Factor w/ 3658 levels "","0001-05-14",..: 2773 2157 2813 2338 3587 3104 987
##  $ seed                : int  1750000 0 40000 1500000 0 0 0 0 0 41250 ...
##  $ venture             : num  0e+00 4e+06 0e+00 0e+00 0e+00 7e+06 0e+00 2e+06 0e+00 0e+00 ...
##  $ equity_crowdfunding : int  0 0 0 0 60000 0 0 0 0 0 ...
##  $ undisclosed         : int  0 0 0 0 0 0 4912393 0 0 0 ...
##  $ convertible_note    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ debt_financing      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ angel               : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ grant               : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ private_equity      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ post_ipo_equity     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ post_ipo_debt       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ secondary_market    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ product_crowdfunding: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ round_A             : int  0 0 0 0 0 0 0 2000000 0 0 ...
##  $ round_B             : int  0 0 0 0 0 7000000 0 0 0 0 ...
##  $ round_C             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ round_D             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ round_E             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ round_F             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ round_G             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ round_H             : int  0 0 0 0 0 0 0 0 0 0 ...
```

```r
sum(complete.cases(datasource))/nrow(datasource)
```

```
## [1] 0.7087708
```

```r
sum(complete.cases(datasource))
```

```
## [1] 38482
```

```r
describe(datasource)
```

```
##                         vars     n       mean          sd  median    trimmed
## permalink*                 1 54294   22509.23    15336.63 22292.5   22292.27
## name*                      2 54294   22475.71    15310.63 22261.5   22260.47
## homepage_url*              3 54294   19419.69    14712.43 18785.5   18881.48
## category_list*             4 54294    6469.54     5268.49  5836.5    6187.46
## market*                    5 54294     293.16      239.14   257.0     281.64
## funding_total_usd*         6 54294    5183.12     4682.51  4439.0    4813.71
## status*                    7 54294       3.48        1.03     4.0       3.72
## country_code*              8 54294      71.72       46.96   112.0      75.52
## state_code*                9 54294      14.65       18.37     8.0      11.40
## region*                   10 54294     486.89      361.97   570.0     486.90
## city*                     11 54294    1736.37     1372.54  1984.0    1699.09
## funding_rounds            12 49438       1.70        1.29     1.0       1.40
## founded_at*               13 54294    1175.99     1064.97  1028.0    1100.41
## founded_month*            14 54294     238.98      163.28   324.0     248.06
## founded_quarter*          15 54294     135.49       88.90   187.0     142.50
## founded_year              16 38482    2007.36        7.58  2010.0    2008.63
## first_funding_at*         17 54294    2400.38     1165.08  2699.0    2524.70
## last_funding_at*          18 54294    2433.28     1105.47  2805.0    2591.33
## seed                      19 49438  217321.50  1056984.82     0.0   40974.00
## venture                   20 49438 7501050.54 28471124.16     0.0 2402967.81
## equity_crowdfunding       21 49438    6163.32   199904.82     0.0       0.00
## undisclosed               22 49438  130221.28  2981403.65     0.0       0.00
## convertible_note          23 49438   23364.10  1432045.73     0.0       0.00
## debt_financing            24 49438 1888156.89 138204566.17    0.0       0.00
## angel                     25 49438   65418.98   658290.79     0.0       0.00
## grant                     26 49438  162845.28  5612088.00     0.0       0.00
## private_equity            27 49438 2074285.75 31672313.29     0.0       0.00
## post_ipo_equity           28 49438  608873.65 26783480.52     0.0       0.00
## post_ipo_debt             29 49438  443435.97 34281689.53     0.0       0.00
## secondary_market          30 49438   38455.92  3864460.62     0.0       0.00
## product_crowdfunding      31 49438    7074.23   428216.59     0.0       0.00
## round_A                   32 49438 1243955.02  5531974.03     0.0  195745.59
## round_B                   33 49438 1492891.15  7472704.45     0.0   16875.24
## round_C                   34 49438 1205355.80  7993591.74     0.0       0.00
## round_D                   35 49438  737526.06  9815218.19     0.0       0.00
## round_E                   36 49438  342468.20  5406914.57     0.0       0.00
## round_F                   37 49438  169769.19  6277905.45     0.0       0.00
## round_G                   38 49438   57670.67  5252311.93     0.0       0.00
## round_H                   39 49438   14231.97  2716865.29     0.0       0.00
##                              mad min        max       range  skew kurtosis
## permalink*              20123.33   1      49437       49436  0.06    -1.26
## name*                   20093.68   1      49352       49351  0.06    -1.26
## homepage_url*           20059.58   1      45851       45850  0.15    -1.30
## category_list*           6375.92   1      16676       16675  0.33    -1.22
## market*                   292.07   1        754         753  0.31    -1.27
## funding_total_usd*       6578.30   1      14618       14617  0.38    -1.23
## status*                     0.00   1          4           3 -1.66     1.08
## country_code*               0.00   1        116         115 -0.47    -1.56
## state_code*                10.38   1         62          61  1.20    -0.02
## region*                   459.61   1       1090        1089 -0.13    -1.56
```

```
## city*                  1915.52    1       4189        4188   0.02    -1.51
## funding_rounds            0.00    1         18          17   2.93    12.34
## founded_at*            1522.63    1       3370        3369   0.33    -1.34
## founded_month*          103.78    1        421         420  -0.61    -1.40
## founded_quarter*         34.10    1        219         218  -0.77    -1.28
## founded_year              4.45 1902       2014         112  -4.67    39.92
## first_funding_at*      1141.60    1       3915        3914  -0.74    -0.55
## last_funding_at*        904.39    1       3658        3657  -1.00    -0.14
## seed                      0.00    0  130000000   130000000  61.54  6505.69
## venture                   0.00    0 2351000000  2351000000  24.67  1338.20
## equity_crowdfunding       0.00    0   25000000    25000000  73.80  7197.23
## undisclosed               0.00    0  292432833   292432833  57.59  4473.13
## convertible_note          0.00    0  300000000   300000000 188.85 39042.07
## debt_financing            0.00    0 30079503000 30079503000 209.05 45382.48
## angel                     0.00    0   63590263    63590263  42.17  2860.18
## grant                     0.00    0  750500000   750500000  83.31  8856.79
## private_equity            0.00    0 3500000000  3500000000  51.55  4357.09
## post_ipo_equity           0.00    0 4700000000  4700000000 122.82 19767.04
## post_ipo_debt             0.00    0 5800000000  5800000000 128.65 19229.35
## secondary_market          0.00    0  680611554   680611554 140.01 22125.99
## product_crowdfunding      0.00    0   72000000    72000000 135.18 20629.00
## round_A                   0.00    0  319000000   319000000  19.78   752.71
## round_B                   0.00    0  542000000   542000000  20.44   927.85
## round_C                   0.00    0  490000000   490000000  19.01   691.29
## round_D                   0.00    0 1200000000  1200000000  64.29  6549.53
## round_E                   0.00    0  400000000   400000000  32.90  1595.20
## round_F                   0.00    0 1060000000  1060000000 109.22 16825.68
## round_G                   0.00    0 1000000000  1000000000 155.72 27679.62
## round_H                   0.00    0  600000000   600000000 218.09 48110.92
##                          se
## permalink*            65.82
## name*                 65.71
## homepage_url*         63.14
## category_list*        22.61
## market*                1.03
## funding_total_usd*    20.10
## status*                0.00
## country_code*          0.20
## state_code*            0.08
## region*                1.55
## city*                  5.89
## funding_rounds         0.01
## founded_at*            4.57
## founded_month*         0.70
## founded_quarter*       0.38
## founded_year           0.04
## first_funding_at*      5.00
## last_funding_at*       4.74
## seed                4753.77
## venture           128048.40
## equity_crowdfunding  899.07
## undisclosed        13408.81
## convertible_note    6440.60
## debt_financing    621572.72
```

```
## angel                   2960.65
## grant                  25240.27
## private_equity        142445.70
## post_ipo_equity       120458.25
## post_ipo_debt         154181.32
## secondary_market       17380.35
## product_crowdfunding    1925.90
## round_A                24879.96
## round_B                33608.36
## round_C                35951.04
## round_D                44143.78
## round_E                24317.51
## round_F                28234.77
## round_G                23622.18
## round_H                12219.06
```

The data imported has a sample size of 54294 with 39 variables.Note that upon initial examination 70.87708% of the data is complete - indicating that there are missing data points.

Note that many duplicated features were found and removed. After removing for this duplicate and non-information rich features for model training, let us recheck the number of complete rows again.

After removing all NA data, the number of complete cases fell to 32822.

Feature Engineering

Upon examination of data, note that the data has to be further proprocessed before ready for fitting. This includes not only converting the features to the required data types, but also transforming the funding features to the natural log.

```
vcdata$market <- as.factor(vcdata$market) #Converting to factors.
#Status, Country Code, State Code, City are all already in factors.
#Further investigating "city" feature as it has 4189 levels.
vcdata %>% group_by(city) %>% summarise(n=n()) %>% arrange(desc(n))
```

```
## # A tibble: 3,435 x 2
##    city              n
##    <fct>         <int>
##  1 San Francisco  2221
##  2 New York       1953
##  3 London         1022
##  4 Palo Alto       484
##  5 Austin          462
##  6 Seattle         448
##  7 Cambridge       423
##  8 Chicago         423
##  9 Mountain View   414
## 10 Los Angeles     400
## # ... with 3,425 more rows
```

```
#Found features that contain \x in the data.
vcdata$city <- factor(str_replace_all(vcdata$city, "[^[A-Za-z]]", " "))
#Removed all \x in data. Reduced to 4127 levels after removing \x in data.
#Converting to date format.
vcdata$founded_at <- as.Date(vcdata$founded_at, "%Y-%m-%d")
vcdata$first_funding_at <- as.Date(vcdata$first_funding_at, "%Y-%m-%d")
vcdata$last_funding_at <- as.Date(vcdata$last_funding_at, "%Y-%m-%d")

vcdata %>% select(seed,venture,equity_crowdfunding,undisclosed,convertible_note,
  debt_financing,angel,grant,private_equity,post_ipo_equity,post_ipo_debt,
  secondary_market,product_crowdfunding) %>% range()
```

```
## [1]           0 30079503000
```

```
# Range is 0 to 30,079,503,000. This will be expensive to compute - hence, the
# data will be transformmed with log transformations.

vcdata_funding <- vcdata[,8:20]
#Coding 0s with NAs first as log 0 is infinity.
vcdata_funding[vcdata_funding == 0] <- NA
vcdata_funding <- log(vcdata_funding)
#Recoding NAs with 0.
vcdata_funding[is.na(vcdata_funding)] <- 0
#Checking if transformation was done correctly.
range(vcdata_funding)
```

```
## [1]  0.00000 24.12711
```

```r
#Repackaging dataset.
vcdata[,8:20] <- vcdata_funding
#Removing extra data sets to clean up environment.
rm(vcdata_funding)
```

Further engineering a new feature that computes the number of days between the first and last day of funding.

```r
vcdata <- vcdata%>%mutate(funding_days_gap=last_funding_at-first_funding_at)
#Checking if feature has been implemented correctly.
range(as.numeric(vcdata$funding_days_gap))
```

```
## [1]      0 733177
```

```r
#Finding for rows where date gaps are above 5000 days.
sum(vcdata$funding_days_gap>5000)
```

```
## [1] 20
```

```r
vcdata_outliers <- vcdata[vcdata$funding_days_gap>5000,]
#Extracting Outlier rows to be further investigated.
vcdata <- vcdata[!(vcdata$funding_days_gap>5000),]
#Removing Outlier rows for east of rebuilding dataset after.
vcdata_outliers$first_funding_at[c(2,5,15,17)] <- as.Date(c("2016-06-01","2012-08-01","2013-07-05","201
vcdata_outliers$last_funding_at[c(2,5,15,17)] <- as.Date(c("2016-07-08","2014-11-19","2019-07-26","2014
#Corrected outliers for the actual dates found on CrunchBase.
vcdata_outliers <- vcdata_outliers%>%mutate(funding_days_gap=last_funding_at-first_funding_at)
#Reconstructing Dataset
vcdata <- rbind(vcdata,vcdata_outliers)
#Checking if implemented correctly.
range(vcdata$funding_days_gap)
```

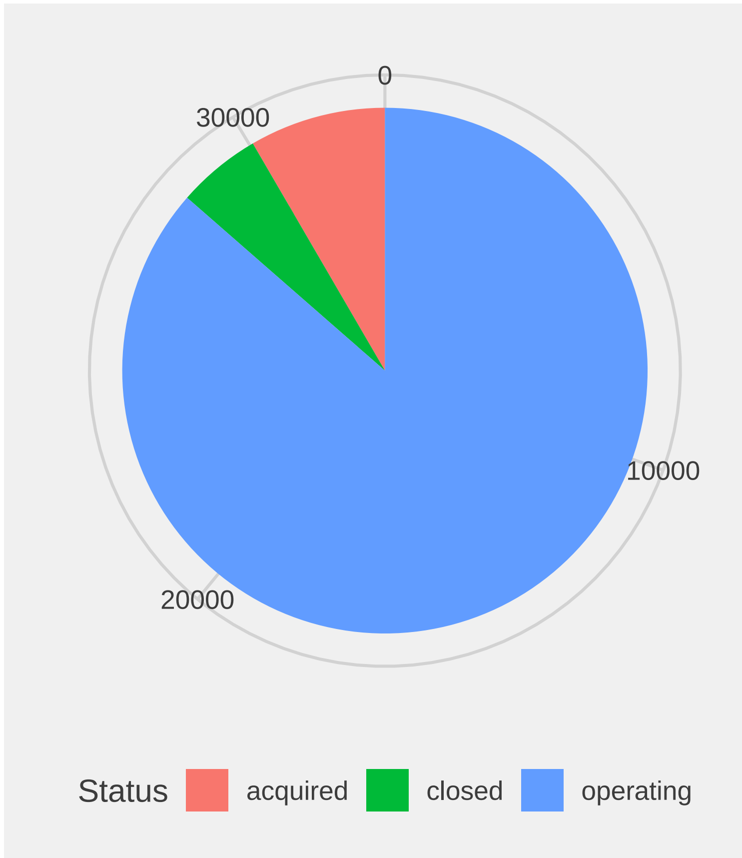
```
## Time differences in days
## [1]     0 10052
```

```r
#Removing redundant status factor level.
vcdata$status <- factor(as.character(vcdata$status))
```

All datasets have been converted to relevant data types. A new feature "funding_days_gap" which is the days in between the first funding date and last funding date has been computed as well.

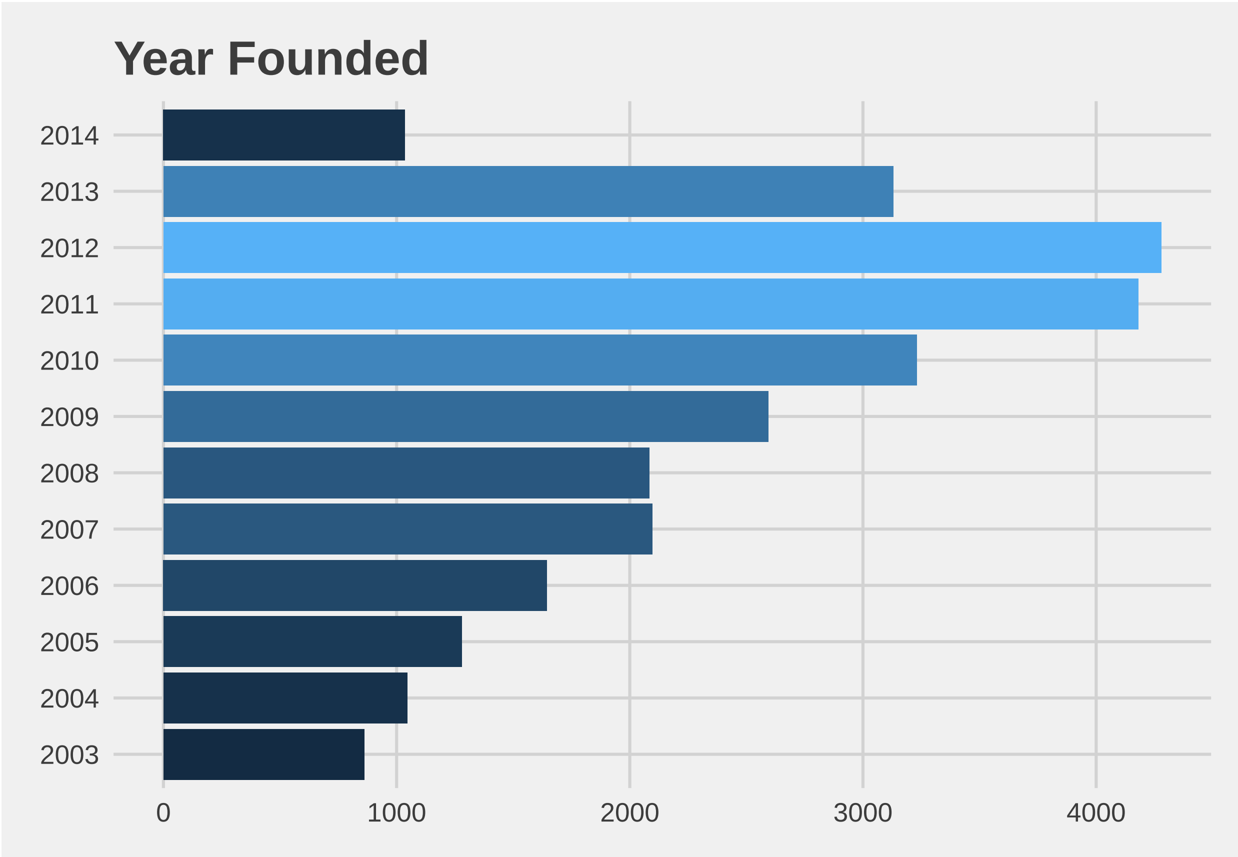
Analysis and Plots

Understanding the data through visualisations.



The data is imbalanced. This indicates that there might be some issues anticipated when running dimensionality reduction and prediction.

Start-up Formation Trends Over Time



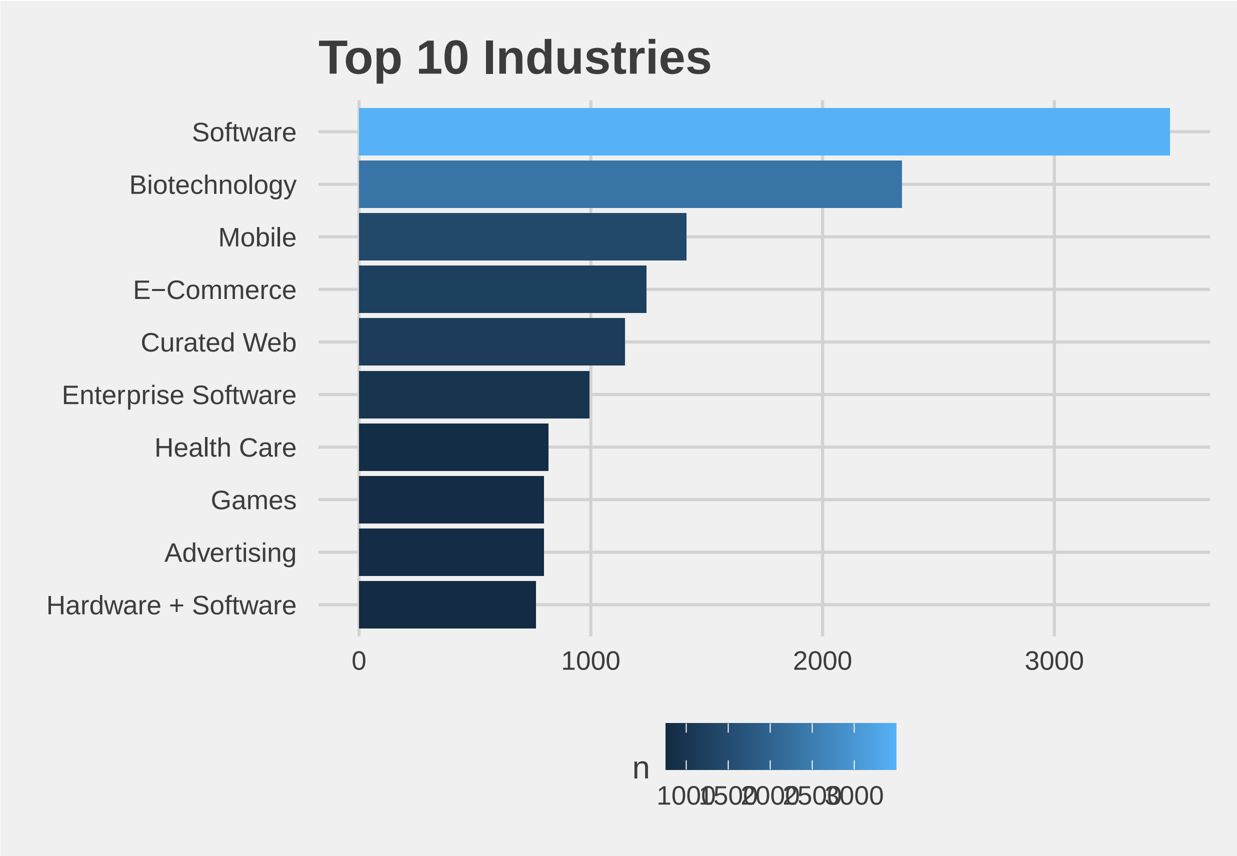
The number of start-ups that have been growing steadily with a huge leap in 2006 - 2007. Note that oddly, there was only a small dip in 2008; indicating that the collapse of the main wall street institutions due to the subprime mortgage crisis did not really impact entrepreneur's mindset that quickly.'

In fact, there was a quick recovery in 2009; indicating that perhaps that after losing their jobs, people are more willing to take risks. And indeed, after a quick search, University of Missouri did find such patterns as well : https://munewsarchives.missouri.edu/news-releases/2012/0731-economic-recession-leads-to-increased-entrepreneurship-mu-study-finds/

Recessions drive people to take risks! But obviously, there are confounding factors to take into account here: - Amazon Web Services was founded in 2016 (allowing for more access to compute power without high initial capital expenditure). - Launch of iPhone in 2007 - which gave rise to a large rise in the app industry. - Rise of social media and online entertainment (Facebook in 2004, Youtube in 2005, Twitter in 2006)

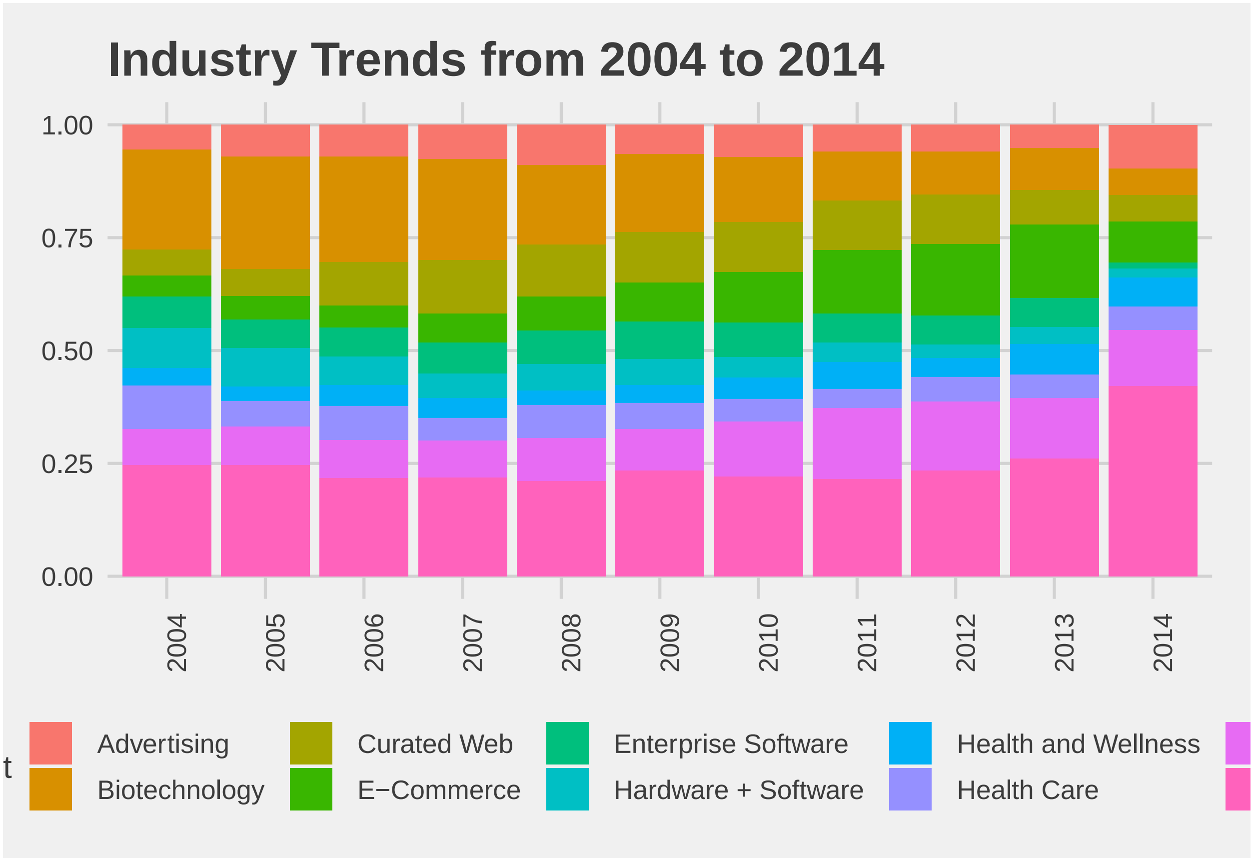
Regardless, interesting find!

Start-up Patterns by Industry



As expected, the software industry has the highest number of start-ups; with biotechnology right behind it.There seem to be a large gap in between biotechnology and mobile; the number of mobile is roughly half of biotechnology.

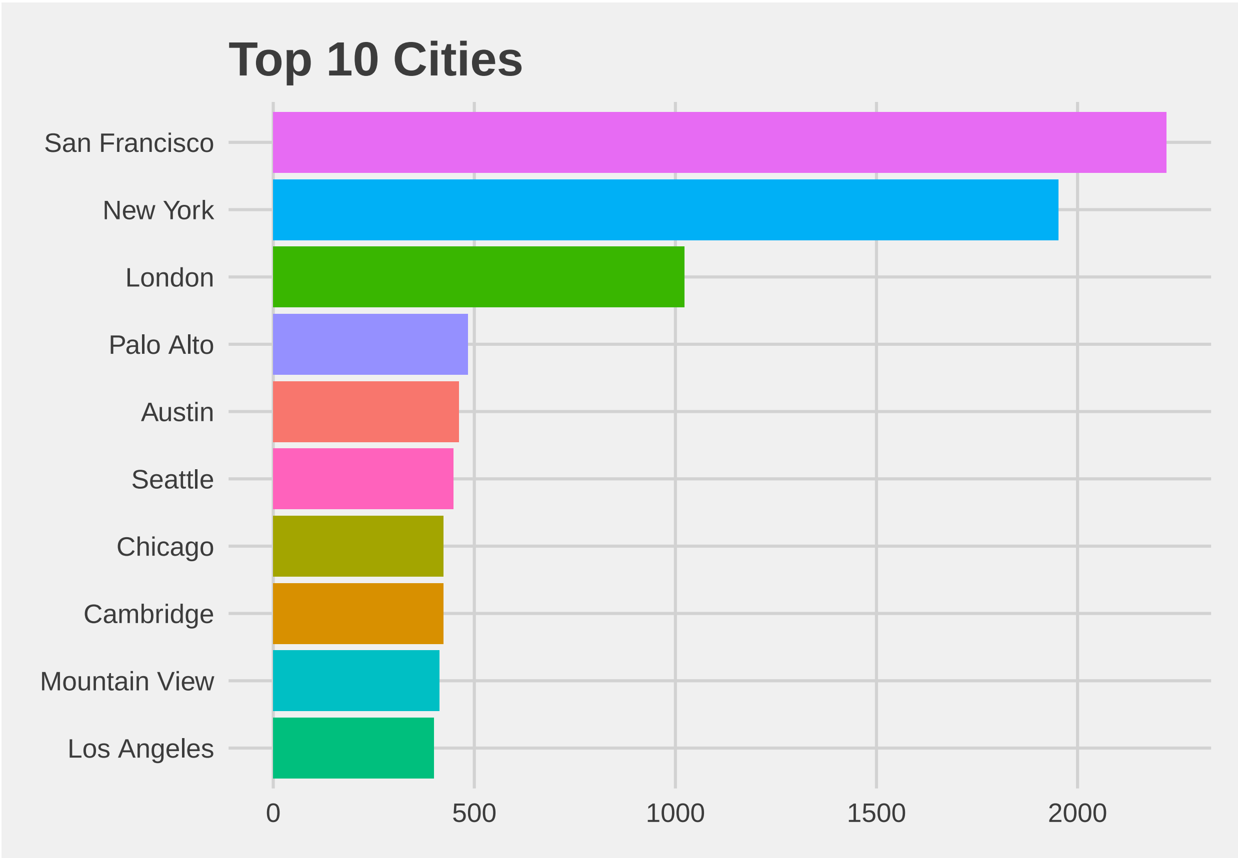
Start-up Patterns by Industry



The trends show that the patterns over the past 10 years tend to be relatively stable - with the exception of :

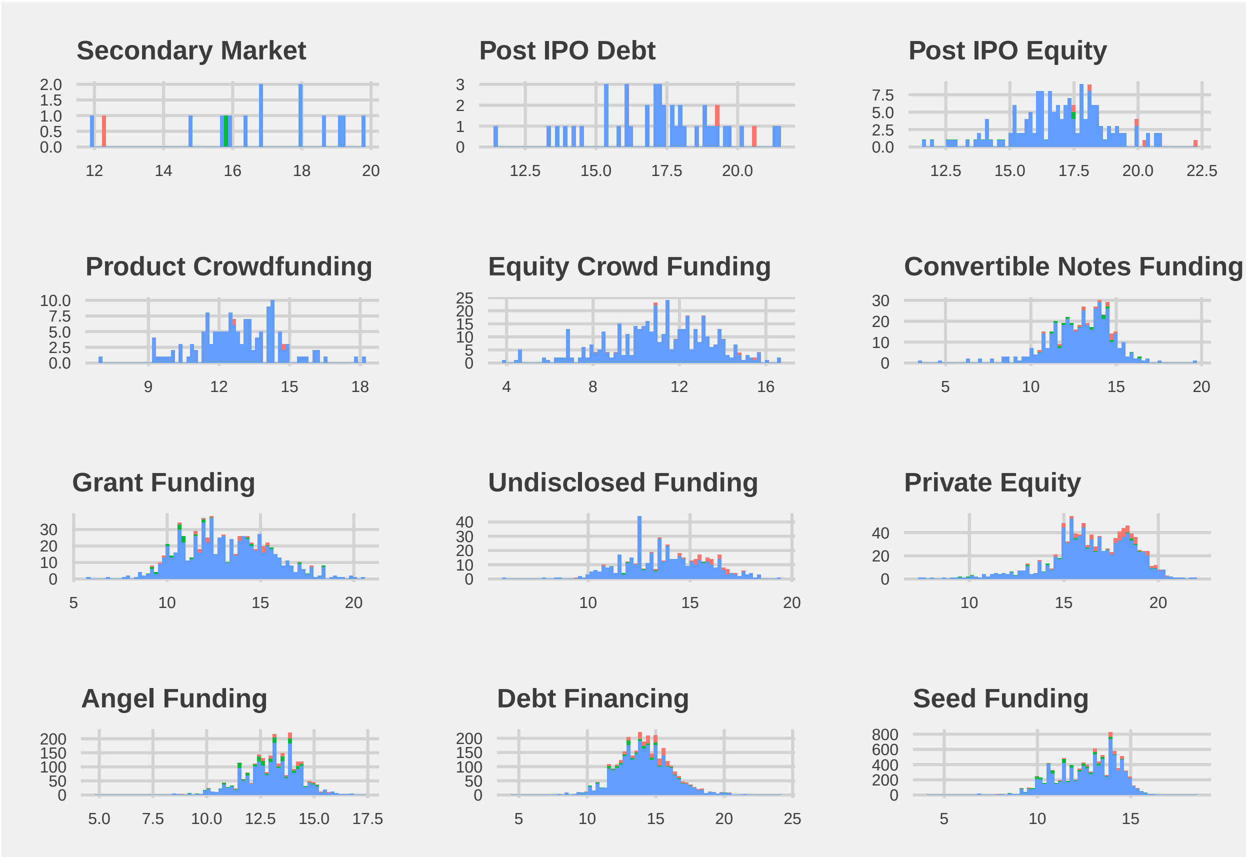
- Increment of software startups in 2014. This is due to post 2008 crisis funding increment (particularly in venture capital).
- Decrement of share of biotechnology startups starting from 2004 until 2014. Whilst it is not clear why, personal research indicate that hurdles required to bypass in order to make a successful marketable product is difficult - and the initial investment required in biotechnology companies are high.

Location



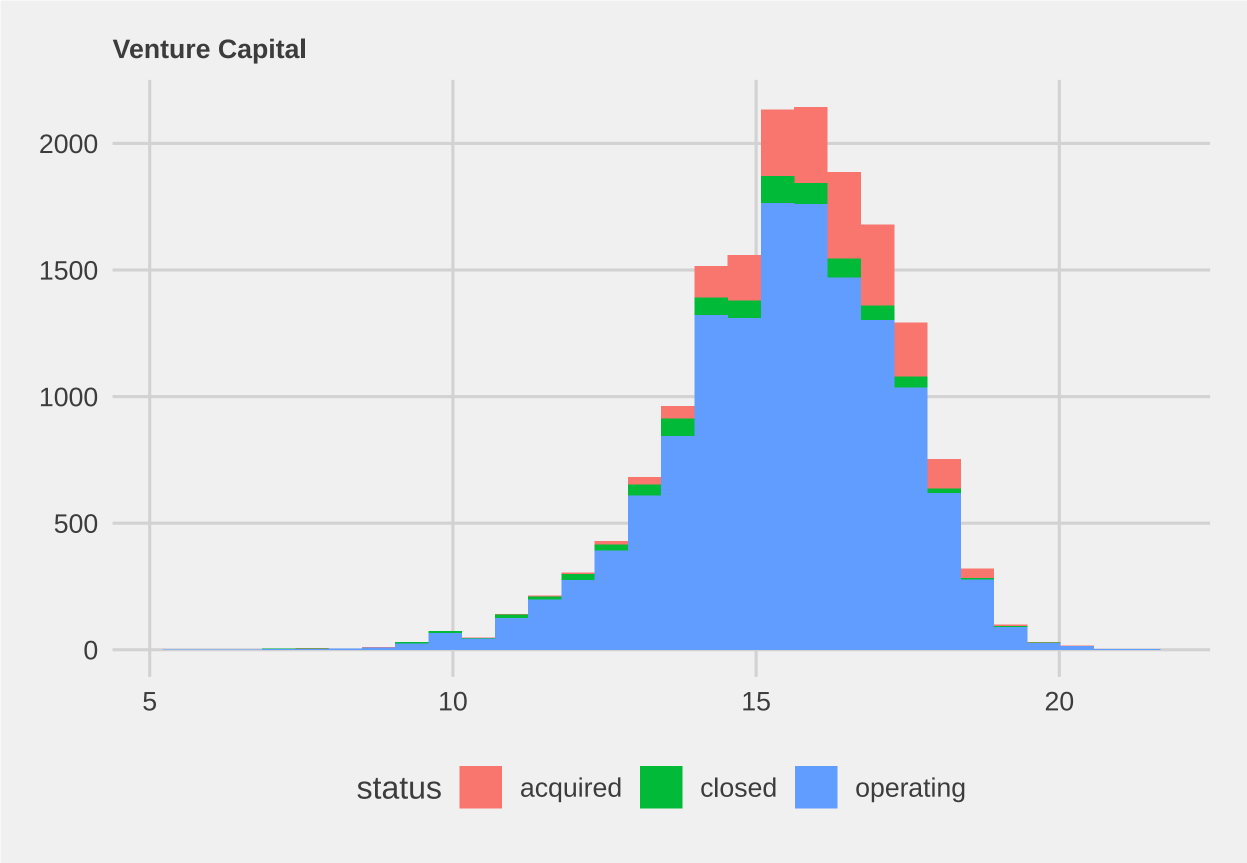
As expected, the highest number of startups in the world are all based in the United States. Let's find out which cities are reputable for start-ups outside of the United States.

Distribution of Funding



Note that the X axis has been log transformed.

As seen from the graph above, main sources are relatively normally distributed. However, the funding that is low in frequency (i.e., secondary market, post IPO debt) are not - this is likely due to the low numbers of funding given out in each of these funding categories.



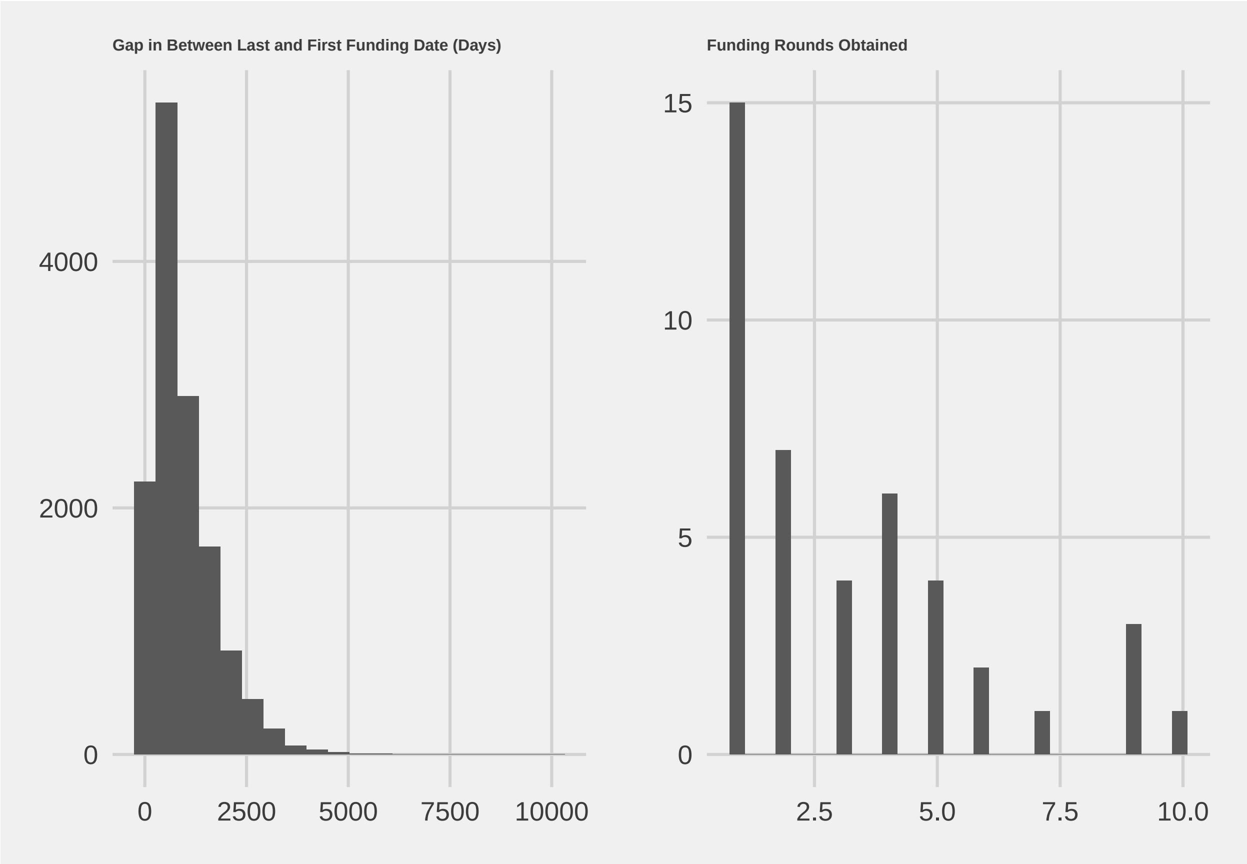
Let's analyse this further:

- Seed funding - This distribution displays a negative skew; indicating that most seed funding tend to be on the lower end. This is to be expected as seed funding is given for companies that are.
- Venture Capital - The distribution is slightly negative skewed. The mean is higher than seed funding's mean as well - this is to be expected as venture capital funding is normally for companies that are in the later development stages of a startup.

Note that the frequencies of both seed funding and venture capital is high - relative to other sources of funding.

The distribution of undisclosed funding is relatively normal - indicating that this category most likely contains data from all forms of funding aggregated together.

Funding Patterns



The mode of funding distributions is 1 funding round - indicating that most start-ups only go through one funding round.

Funding after the first round decreases significantly - most startups don't get through to the 2nd round of funding.

Sub-Conclusion

A lot can be learnt about the start-up funding scene just from these graphs. And, as observed before, the data is imbalanced, hence, it is very likely that algorithms implemented will be ineffective.

Regardless, let's get on with the algorithms.

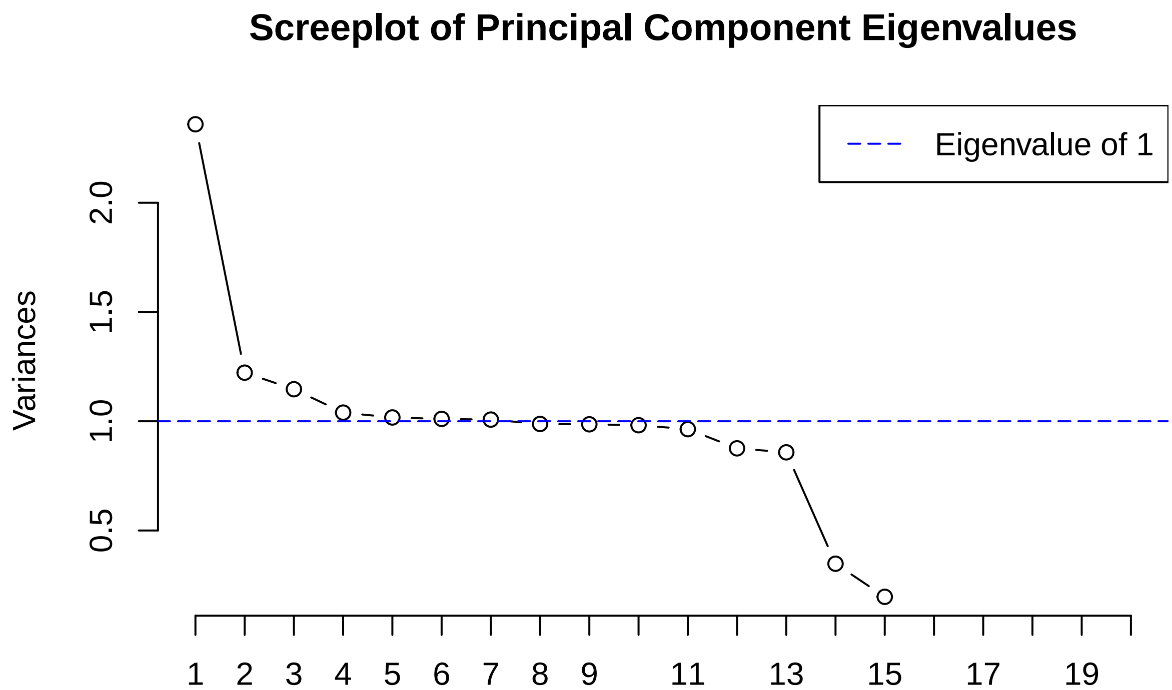
Algorithm Implementation 1 - Dimensionality Reduction : PCA

Principal Component Analysis was fitted onto the data. Before doing that, all categorical data was converted into dummy coding.

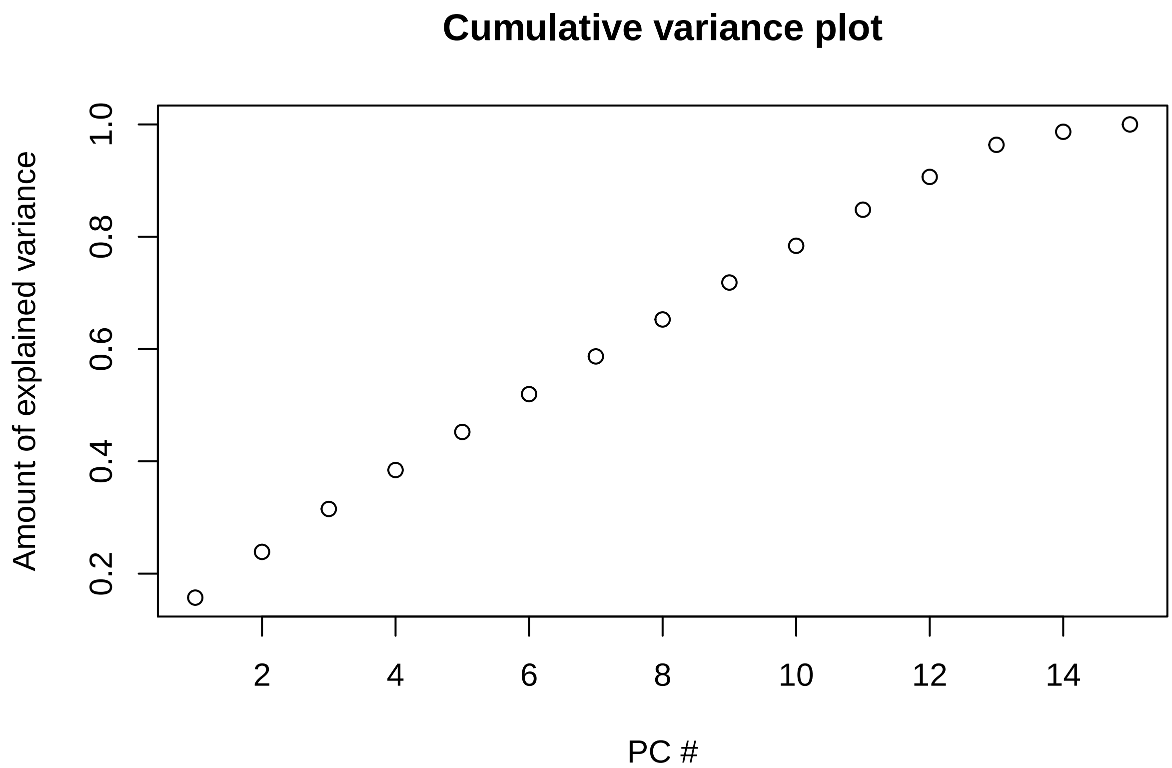
```
#Removing non.numeric vectors
vcdataforpc <- vcdata[,-c(1,2,3,5,6,7)]
vcdataforpc$funding_days_gap <- as.numeric(vcdataforpc$funding_days_gap)
vcdata.pr <- prcomp(vcdataforpc,center=TRUE,scale=TRUE)
summary(vcdata.pr)
```

```
## Importance of components:
##                           PC1     PC2     PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.5359 1.10557 1.07076 1.0195 1.00863 1.00533 1.00376
## Proportion of Variance 0.1573 0.08149 0.07643 0.0693 0.06782 0.06738 0.06717
## Cumulative Proportion  0.1573 0.23876 0.31519 0.3845 0.45232 0.51969 0.58686
##                            PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.99374 0.99297 0.99087 0.98172 0.93588 0.92611 0.58985
## Proportion of Variance 0.06584 0.06573 0.06546 0.06425 0.05839 0.05718 0.02319
## Cumulative Proportion  0.65270 0.71843 0.78389 0.84814 0.90653 0.96371 0.98690
##                          PC15
## Standard deviation     0.4432
## Proportion of Variance 0.0131
## Cumulative Proportion  1.0000
```

```
#Cut off point decided for Eigenvalue = 1
screeplot(vcdata.pr, type = "l", npcs = 20,
          main = "Screeplot of Principal Component Eigenvalues")
abline(h = 1, col="blue", lty=5)
legend("topright", legend=c("Eigenvalue of 1"),
       col=c("blue"), lty=5, cex=1.00)
```




```r
#Further checking if data is suitable for pca.
cumpro <- cumsum(vcdata.pr$sdev^2 / sum(vcdata.pr$sdev^2))
plot(cumpro[0:15], xlab = "PC #", ylab = "Amount of explained variance",
     main = "Cumulative variance plot")
```




```
#Not suitable for running PCA.
```

The results indicated that the data is not suitable for principle component analysis. This can be seen from how none of the factors captured more variance, in comparison to the others - leading to an almost linear graph in the "Cumulative variance plot".

This is to be expected as the data is structured in a manner where most features are either encoded with the relevant funding amount or coded as '0' for any specific funding category. For instance, a respective startup could have a large amount in VC funding; but nil for everything else.

Algorithm Implementation 2 : Random Tree

Let us try implementing classification algorithms to see how prediction performance would be like in this dataset.

Before that, let us split the data set to train and test set.

```
#Removing categorical predictors > 53 categories as Rtree cannot handle more than 53 categories.
train_forest <- train[,-c(1,3)]
train_rf <- randomForest(status ~ ., data=train_forest)
train_rf
```

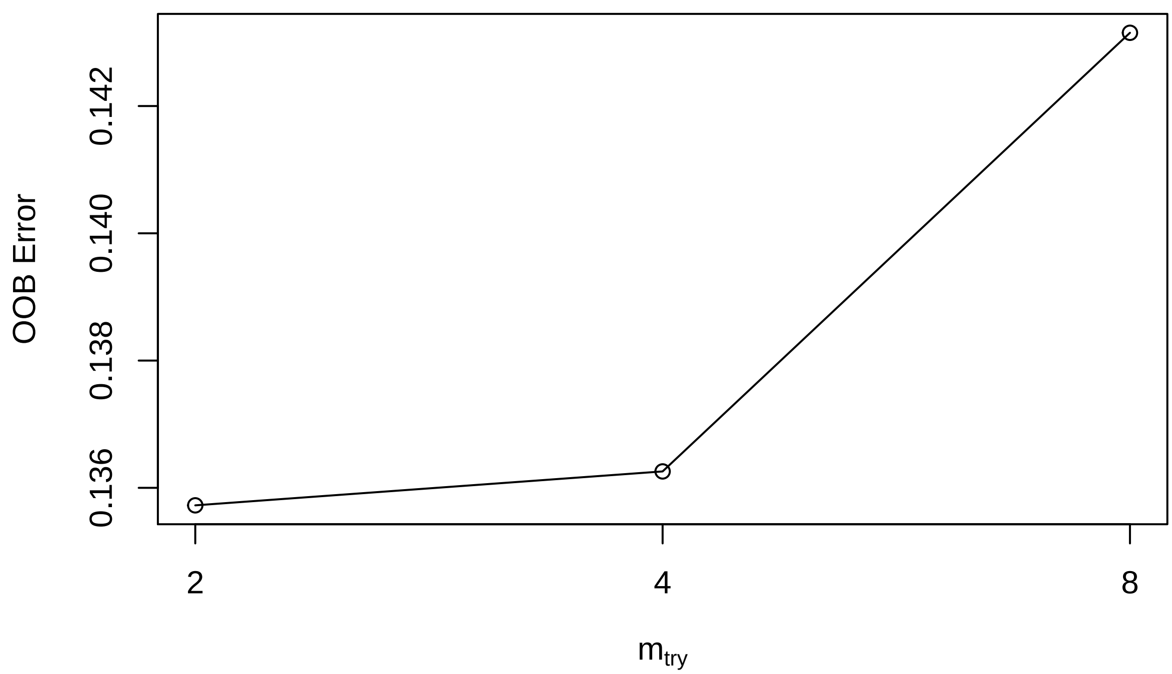
```
##
## Call:
##  randomForest(formula = status ~ ., data = train_forest)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 13.59%
## Confusion matrix:
##           acquired closed operating class.error
## acquired       158      1      2044 0.928279619
## closed          27      0      1335 1.000000000
## operating      161      1     22532 0.007138451
```

The initial random tree forest ran indicate that the algorithm cannot accurately predict whether a start up will be closed or acquired; this is expected as the data is imbalanced - there are more operating start-ups than ones that are acquired vs closed.

Let's tune the parameters just to see if a better outcome could be achieved.

```
#Finding best mtry tuning.
res <- tuneRF(x = subset(train_forest, select=-status),
              y = train_forest$status,
              ntreeTry = 500)
```

```
## mtry = 4  OOB error = 13.63%
## Searching left ...
## mtry = 2     OOB error = 13.57%
## 0.0039128 0.05
## Searching right ...
## mtry = 8     OOB error = 14.32%
## -0.05058692 0.05
```




```
#Lowest mtry = 2
#Finding best node size tuning.
nodesize <- seq(1, 10)
oob_err <- c()
for (i in 1:length(nodesize)) {
    model <- randomForest(formula = status ~ .,
                          data = train_forest,
                          nodesize = nodesize[i],mtry=2)
    oob_err[i] <- model$err.rate[nrow(model$err.rate), "OOB"]
}
which.min(oob_err)
```

```
## [1] 4
```

```r
#Optimal Random Forest Model : Nodesize 4, mtry = 2

finalrtree <- randomForest(status ~ ., data=train_forest,mtry=2,nodesize=4)
```

Hence, the final random forest model trained has paramaters mtry=2 and nodesize = 1. Let's further evaluate the performance of this algorithm on the train set.

```r
test_forest<- test[,-c(1,3)]
#Prediction by assuming that all outcomes are "operating"
guess <- mean(test_forest$status=="operating")
#Accuracy of just predicing that all startups are operating is 86.4%

#Random Forest Tree
random <- confusionMatrix(predict(finalrtree, test_forest), test_forest$status)$overall["Accuracy"]
#Rtree's final accuracy is 86.5%.

print(random-guess)
```

```
##     Accuracy
## 0.0001523693
```

The results indicate that the accuracy of the randomForest model is almost alike predicing all companies as still operating. The accuracy of prediction of randomForest performs marginally better by 00.015% - which is approximately equivalent to a NULL improvement in accuracy.

This indicates that there is no difference in performance here. This is to be expected due to the nature of the data.

Conclusion : Results and Limitations

The PCA and Random Tree Algorithm ran indicate that the data is not suitable for dimensionality reduction, nor for prediction. This is most likely due to the imbalanced nature of the data.

However, from the plots, a few conclusions about the start-up industry can be made :

- Most start-ups are still operating up until 2013; in fact, merely predicting that all start-ups are operating allows gives us an accuracy that is equivalent to a random tree algorithm prediction ran.
- VC funding and Seed funding are the most widespread - as seen from the plots done. Not many companies obtain post IPO financing.
- Most start-ups are based in the United States; this is unsurprising given the notion of "Silicon Valley"
- Economic recessions lead to a spike in entrepreneurship and the number of biotechnology startups have been decreasin - perhaps we will see a spike again post COVID-19 era?

Here are the lessons drawn from this project :

- Always check for data balance before even thinking of cleaning up/analysing the data.
- If data is imbalance, the algorithm will most likely performing as well as an educated guess of predicting that the test outcome is the majority group in the unbalanced data.

Thank you for reading through this. I hope you enjoyed the graphs!

End