Startup Investments

Technical Report – Part 1

Group 6

Vineeth Penugonda, Sai Rachana Bandi, Chandra Vardhan

04 May 2020

# All Libraries  
library(lubridate) # Date  
library(binaryLogic) # Binary Encoding  
  
# Visualizations  
library(dplyr)   
library(ggplot2)   
library("tm")  
library("SnowballC")  
library("wordcloud")  
library("RColorBrewer")  
library(stringr)  
  
# Correlation Analysis  
library("PerformanceAnalytics")  
library(Hmisc)

# Data Preprocessing

The dataset was downloaded from the following kaggle link: <https://www.kaggle.com/arindam235/startup-investments-crunchbase>

The dataset related to startup investments in 116 countries has been collected from the CrunchBase database. We considered only the startup companies based in USA. The factors we considered - the type of market, the state, the region the startup belongs to, total funding (in USD), different types of funding – seed, angel, crowdfunding etc., founded date, last funding date, operating status.

### Key Information In The Dataset

**name**: Name of the startup company

**category\_list**: Category the startup belongs to

**market**: Type of market the startup belongs to

**funding\_total\_usd**: Total funding the startup received

**status**: Status of the startup – acquired, operational, closed

**state\_code**: State in which the startup was founded in

**region**: Region in which the startup was founded in

**city**: City in which the startup was founded in

**funding\_rounds**: Number of funding rounds the startup went through

**founded\_at**: Day on which the startup was founded

**founded\_month**: Month in which the startup was founded

**founded\_quarter**: Quarter in which the startup was founded – Q1, Q2, Q3, Q4

**founded\_year**: Year in which the startup was founded

**first\_funding\_day**: Day on which the startup started receiving funding

**first\_funding\_month**: Month in which the startup started receiving funding

**first\_funding\_year**: Year in which the startup started receiving funding

**last\_funding\_day**: Day on which the startup stopped receiving funding

**last\_funding\_month**: Month in which the startup stopped receiving funding

**last\_funding\_year**: Year in which the startup stopped receiving funding

**seed**: Seed rounds are among the first rounds of funding a company will receive, generally while the company is young and working to gain traction. A seed round typically comes after an angel round (if applicable) and before a company’s Series A round.

**venture**: Venture funding refers to an investment that comes from a venture capital firm and describes Series A, Series B, and later rounds. This funding type is used for any funding round that is clearly a venture round but where the series has not been specified.

**equity\_crowdfunding**: Equity crowdfunding platforms allow individual users to invest in companies in exchange for equity. Typically, on these platforms the investors invest small amounts of money, though syndicates are formed to allow an individual to take a lead on evaluating an investment and pooling funding from a group of individual investors.

**undisclosed**: Undisclosed amount on the last funding date.

**convertible\_note**: A convertible note is an ‘in-between’ round funding to help companies hold over until they want to raise their next round of funding. When they raise the next round, this note ‘converts’ with a discount at the price of the new round. You will typically see convertible notes after a company raises, for example, a Series A round but does not yet want to raise a Series B round.

**debt\_financing**: In a debt round, an investor lends money to a company, and the company promises to repay the debt with added interest.

**angel**: An angel round is typically a small round designed to get a new company off the ground. Investors in an angel round include individual angel investors, angel investor groups, friends, and family.

**grant**: A grant is when a company, investor, or government agency provides capital to a company without taking an equity stake in the company.  
**private\_equity**: A private equity round is led by a private equity firm or a hedge fund and is a late stage round. It is a less risky investment because the company is more firmly established, and the rounds are typically upwards of $50M.

**post\_ipo\_equity**: A post-IPO equity round takes place when firms invest in a company after the company has already gone public.

**post\_ipo\_debt**: A post-IPO debt round takes place when firms loan a company money after the company has already gone public. Similar to debt financing, a company will promise to repay the principal as well as added interest on the debt.

**secondary\_market**: A secondary market transaction is a fundraising event in which one investor purchases shares of stock in a company from other, existing shareholders rather than from the company directly. These transactions often occur when a private company becomes highly valuable and early stage investors or employees want to earn a profit on their investment, and these transactions are rarely announced or publicized.

**product\_crowdfunding**: In a product crowdfunding round, a company will provide its product, which is often still in development, in exchange for capital. This kind of round is also typically completed on a funding platform.

**round\_A**: Round A is a funding round for earlier stage companies and range on average between $1M–$30M.

**round\_B**: Round B is a funding round for earlier stage companies and range on average between $1M–$30M.

**round\_C**: Round C is a funding round for later stage and more established companies. These rounds are usually $10M+ and are often much larger.

**round\_D**: Round D is a funding round for later stage and more established companies. These rounds are usually $10M+ and are often much larger.  
**round\_E**: Round E is a funding round for later stage and more established companies. These rounds are usually $10M+ and are often much larger.  
**round\_F**: Round F is a funding round for later stage and more established companies. These rounds are usually $10M+ and are often much larger.  
**round\_G**: Round G is a funding round for later stage and more established companies. These rounds are usually $10M+ and are often much larger.  
**round\_H**: Round H is a funding round for later stage and more established companies. These rounds are usually $10M+ and are often much larger.  
**post\_success**: Indicates if the startup was successful or not successful based on the rules mentioned above.

dat <- read.csv("../Dataset/US\_StartUps\_Investments\_Data.csv", na.strings = c("", "NA"), stringsAsFactors = FALSE)

On initial observation, the dataset downloaded from Kaggle had a lot of missing data. So, we removed rows which had missing data in one or more columns.

# Replacing commas in the column - funding\_total\_usd  
dat$funding\_total\_usd <- as.numeric(gsub(",","",dat$funding\_total\_usd))

The values in the column “funding\_total\_usd” had commas. We removed commas for the values.

# Remove rows with NAs  
dat <- dat[complete.cases(dat),]

# Remove trailing spaces in column - market  
dat$market <- trimws(dat$market, which = c("both"))

The **market** column had left and right trailing spaces for each keyword. So, we trimmed it using trimws() function.

# Remove country\_code, homepage\_URL, X and permalink  
dat$homepage\_url <- NULL  
dat$X <- NULL  
dat$permalink <- NULL  
dat$country\_code <- NULL

We didn’t require **homepage URL of the startup company**, **X which is the ID**, **permalink - permanent link in the crunchbase**, **country code**. We are considering only startups in United States of America (USA).

# Strip values in columns  
# Month  
dat$founded\_month = substring(dat$founded\_month,6)  
dat$founded\_month = as.numeric(dat$founded\_month)  
  
# Quarter  
dat$founded\_quarter = substring(dat$founded\_quarter,6)  
  
# YY/MM/DD. We only need the day.  
dat$founded\_at = substring(dat$founded\_at,9)  
dat$founded\_at = as.numeric(dat$founded\_at)

We observed that the values in the **founded\_month** column were in **Year-Month**. Since we needed only the month we removed the year from those values. It was a similar case for **founded\_quarter** column.

**founded\_at** is the day in the YY/MM/DD format. So we removed year and month information from the values to just get the day(dd) information.

# Date DataType  
dat$first\_funding\_at <- as.Date(dat$first\_funding\_at)  
dat$last\_funding\_at <- as.Date(dat$last\_funding\_at)

# Separate date into three columns  
dat$first\_funding\_day <- day(ymd(dat$first\_funding\_at))  
dat$first\_funding\_month <- month(ymd(dat$first\_funding\_at))   
dat$first\_funding\_year <- year(ymd(dat$first\_funding\_at))  
  
dat$last\_funding\_day <- day(ymd(dat$last\_funding\_at))  
dat$last\_funding\_month <- month(ymd(dat$last\_funding\_at))   
dat$last\_funding\_year <- year(ymd(dat$last\_funding\_at))  
  
# Remove Funding Dates  
dat$first\_funding\_at <- NULL  
dat$last\_funding\_at <- NULL

**first\_funding\_at** and **last\_funding\_at** are in YY/MM/DD format. We split the dates into three columns - **first\_funding\_day** (DD), **first\_funding\_month** (MM),**first\_funding\_year** (YY).

# Create new column: Success or not Success  
# Acquired, post\_ipo\_equity or post\_ipo\_equity and Acquired  
  
dat$post\_success <- ifelse(dat$status=="acquired", 1,  
 ifelse(dat$post\_ipo\_equity > 0, 1,  
 ifelse((dat$status=="acquired") && (dat$post\_ipo\_equity > 0), 1,  
 0 ))) # all other values map to 0  
  
dat$post\_success <- factor(dat$post\_success)

Both an IPO (Initial Public Offering) and a process of M&A (Mergers & Acquisitions) are the critical events that classify a start-up as successful.

We created the “post\_success” column which indicates if the startup was successful or not successful based on these rules – if the status of the company is “acquired” or if the post IPO equity is greater than 0 or if the status of the company is “acquired” and the post IPO equity is greater than 0.

# Saving file  
write.csv(dat,'Data\_Cleansed.csv')  
saveRDS(dat, file = "Data\_Cleansed.rds")

### Binary Encoding

There are more than 32 markets, regions and states. We binary encoded the values in *market*, *region*, *state\_code*. Then, digits from that binary string are split into separate columns. This is useful for machine learning. Better encoding of categorical data can mean better model performance.

# Function for Binary Encoding  
  
encode\_binary <- function(x, order = unique(x), name = "v\_") {  
 x <- as.numeric(factor(x, levels = order, exclude = NULL))  
 x2 <- as.binary(x)  
 maxlen <- max(sapply(x2, length))  
 x2 <- lapply(x2, function(y) {  
 l <- length(y)  
 if (l < maxlen) {  
 y <- c(rep(0, (maxlen - l)), y)  
 }  
 y  
 })  
 d <- as.data.frame(t(as.data.frame(x2)))  
 rownames(d) <- NULL  
 colnames(d) <- paste0(name, 1:maxlen)  
 d  
}

# dat -> Visualizations  
# filter\_data -> ML  
  
new\_df <- cbind(dat, encode\_binary(dat[["market"]], name = "market\_"))  
  
new\_df <- cbind(new\_df, encode\_binary(dat[["region"]], name = "region\_"))  
  
new\_df <- cbind(new\_df, encode\_binary(dat[["state\_code"]], name = "state\_code\_"))  
  
saveRDS(new\_df, file = "Data\_Cleansed\_Encoded.rds")

### Filtering and Factorizing

We selected data from year 1990 until 2014. We converted all the categorical variables to factors. We also converted the data types from “num” to “logical” (boolean) for the binary encoded columns.

# Filter the data and Factorize it!  
  
filter\_data <- new\_df[ which(new\_df$founded\_year >= 1990), ]  
  
# Datatypes  
# Categorical Variables: name, category\_list, market, status, state\_code, region, city, founded\_at, founded\_month, founded\_quarter,  
# founded\_year, first\_funding\_day, first\_funding\_month, first\_funding\_year, last\_funding\_day, last\_funding\_month,  
# last\_funding\_year  
  
filter\_data$name <- NULL  
filter\_data$category\_list<- NULL  
filter\_data$market<- NULL  
  
filter\_data$status<-factor(filter\_data$status)  
  
filter\_data$state\_code<- NULL  
filter\_data$region<- NULL  
filter\_data$city<- NULL  
  
filter\_data$founded\_at<-factor(filter\_data$founded\_at)  
filter\_data$founded\_month<-factor(filter\_data$founded\_month)  
filter\_data$founded\_quarter<-factor(filter\_data$founded\_quarter)  
filter\_data$founded\_year<-factor(filter\_data$founded\_year)  
filter\_data$first\_funding\_day<-factor(filter\_data$first\_funding\_day)  
filter\_data$first\_funding\_month<-factor(filter\_data$first\_funding\_month)  
filter\_data$first\_funding\_year<-factor(filter\_data$first\_funding\_year)  
filter\_data$last\_funding\_day<-factor(filter\_data$last\_funding\_day)  
filter\_data$last\_funding\_month<-factor(filter\_data$last\_funding\_month)  
filter\_data$last\_funding\_year<-factor(filter\_data$last\_funding\_year)  
  
filter\_data$market\_1 <- as.logical(filter\_data$market\_1)  
filter\_data$market\_2 <- as.logical(filter\_data$market\_2)  
filter\_data$market\_3 <- as.logical(filter\_data$market\_3)  
filter\_data$market\_4 <- as.logical(filter\_data$market\_4)  
filter\_data$market\_5 <- as.logical(filter\_data$market\_5)  
filter\_data$market\_6 <- as.logical(filter\_data$market\_6)  
filter\_data$market\_7 <- as.logical(filter\_data$market\_7)  
filter\_data$market\_8 <- as.logical(filter\_data$market\_8)  
filter\_data$market\_9 <- as.logical(filter\_data$market\_9)  
filter\_data$market\_10 <- as.logical(filter\_data$market\_10)  
  
filter\_data$region\_1 <- as.logical(filter\_data$region\_1)  
filter\_data$region\_2 <- as.logical(filter\_data$region\_2)  
filter\_data$region\_3 <- as.logical(filter\_data$region\_3)  
filter\_data$region\_4 <- as.logical(filter\_data$region\_4)  
filter\_data$region\_5 <- as.logical(filter\_data$region\_5)  
filter\_data$region\_6 <- as.logical(filter\_data$region\_6)  
filter\_data$region\_7 <- as.logical(filter\_data$region\_7)  
filter\_data$region\_8 <- as.logical(filter\_data$region\_8)  
  
filter\_data$state\_code\_1 <- as.logical(filter\_data$state\_code\_1)  
filter\_data$state\_code\_2 <- as.logical(filter\_data$state\_code\_2)  
filter\_data$state\_code\_3 <- as.logical(filter\_data$state\_code\_3)  
filter\_data$state\_code\_4 <- as.logical(filter\_data$state\_code\_4)  
filter\_data$state\_code\_5 <- as.logical(filter\_data$state\_code\_5)  
filter\_data$state\_code\_6 <- as.logical(filter\_data$state\_code\_6)  
  
saveRDS(filter\_data, file="Data\_CE\_Filtered.rds")

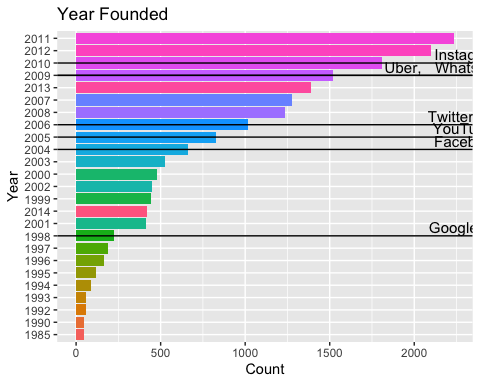
str(filter\_data)

## 'data.frame': 17784 obs. of 59 variables:  
## $ funding\_total\_usd : num 1750000 2050000 40000 2535000 4962651 ...  
## $ status : Factor w/ 3 levels "acquired","closed",..: 1 3 3 1 3 3 1 3 2 3 ...  
## $ funding\_rounds : num 1 4 1 2 6 1 2 1 1 1 ...  
## $ founded\_at : Factor w/ 31 levels "1","2","3","4",..: 1 1 1 1 1 4 16 1 1 1 ...  
## $ founded\_month : Factor w/ 12 levels "1","2","3","4",..: 6 1 1 7 1 7 9 1 4 1 ...  
## $ founded\_quarter : Factor w/ 4 levels "Q1","Q2","Q3",..: 2 1 1 3 1 3 3 1 2 1 ...  
## $ founded\_year : Factor w/ 25 levels "1990","1991",..: 23 12 19 21 19 25 22 11 20 23 ...  
## $ seed : num 1750000 0 40000 15000 0 420000 750000 0 50000 0 ...  
## $ venture : num 0 0 0 2520000 3814772 ...  
## $ equity\_crowdfunding : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ undisclosed : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ convertible\_note : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ debt\_financing : num 0 2050000 0 0 1147879 ...  
## $ angel : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ grant : num 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 : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ product\_crowdfunding: num 0 0 0 0 0 0 0 0 0 0 ...  
## $ round\_A : num 0 0 0 2520000 0 0 0 0 0 0 ...  
## $ round\_B : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ round\_C : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ round\_D : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ round\_E : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ round\_F : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ round\_G : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ round\_H : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ first\_funding\_day : Factor w/ 31 levels "1","2","3","4",..: 30 18 23 1 14 26 2 8 1 7 ...  
## $ first\_funding\_month : Factor w/ 12 levels "1","2","3","4",..: 6 6 8 1 10 11 11 3 4 11 ...  
## $ first\_funding\_year : Factor w/ 24 levels "11","1990","1993",..: 22 19 21 20 19 24 21 20 19 22 ...  
## $ last\_funding\_day : Factor w/ 31 levels "1","2","3","4",..: 30 28 23 16 19 26 30 8 1 7 ...  
## $ last\_funding\_month : Factor w/ 12 levels "1","2","3","4",..: 6 12 8 2 9 11 11 3 4 11 ...  
## $ last\_funding\_year : Factor w/ 22 levels "1990","1993",..: 20 19 19 19 22 22 19 18 17 20 ...  
## $ post\_success : Factor w/ 2 levels "0","1": 2 1 1 2 1 1 2 1 1 1 ...  
## $ market\_1 : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ market\_2 : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ market\_3 : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ market\_4 : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ market\_5 : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ market\_6 : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ market\_7 : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ market\_8 : logi FALSE FALSE TRUE TRUE TRUE TRUE ...  
## $ market\_9 : logi FALSE TRUE FALSE FALSE FALSE TRUE ...  
## $ market\_10 : logi TRUE TRUE FALSE TRUE TRUE FALSE ...  
## $ region\_1 : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ region\_2 : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ region\_3 : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ region\_4 : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ region\_5 : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ region\_6 : logi FALSE FALSE TRUE TRUE TRUE TRUE ...  
## $ region\_7 : logi FALSE TRUE FALSE FALSE TRUE TRUE ...  
## $ region\_8 : logi TRUE TRUE FALSE TRUE FALSE TRUE ...  
## $ state\_code\_1 : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ state\_code\_2 : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ state\_code\_3 : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ state\_code\_4 : logi FALSE FALSE TRUE TRUE TRUE TRUE ...  
## $ state\_code\_5 : logi FALSE TRUE FALSE FALSE TRUE TRUE ...  
## $ state\_code\_6 : logi TRUE TRUE FALSE TRUE FALSE TRUE ...

Finally, we have no null values in the dataset! We changed the data type of each column for further analysis.

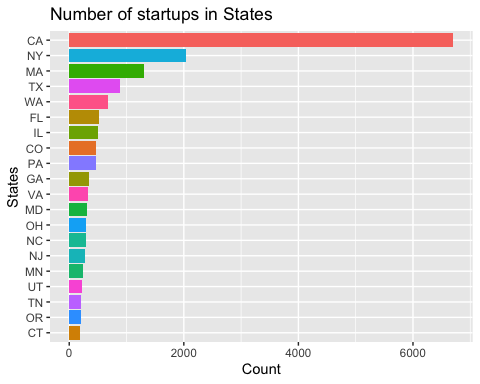
# Visualizations

# Founding year  
viz\_data <- dat  
viz\_data$founded\_year <- factor(viz\_data$founded\_year)  
  
viz\_data %>% count(founded\_year)%>% arrange(-n) %>% head(25) %>%  
 ggplot(aes(reorder(founded\_year, n), n)) + geom\_col(aes(fill=founded\_year)) + coord\_flip() +  
 theme(legend.position="none") +  
 ggtitle("Year Founded") + xlab("Year") + ylab("Count") +  
 geom\_vline(aes(xintercept = founded\_year), data = viz\_data %>% filter(name == "Facebook")) +  
 geom\_text(aes(x = founded\_year, y=2000, label = "Facebook"), data = viz\_data %>% filter(name == "Facebook"), size=4, vjust=-0.3, hjust=-0.3) +  
 geom\_vline(aes(xintercept = founded\_year), data = viz\_data %>% filter(name == "Twitter")) +  
 geom\_text(aes(x = founded\_year, y=2000, label = "Twitter"), data = viz\_data %>% filter(name == "Twitter"), size=4, vjust=-0.3, hjust=-0.3) +  
 geom\_vline(aes(xintercept = founded\_year), data = viz\_data %>% filter(name == "Google")) +  
 geom\_text(aes(x = founded\_year, y=2000, label = "Google"), data = viz\_data %>% filter(name == "Google"), size=4, vjust=-0.3, hjust=-0.3) +  
 geom\_vline(aes(xintercept = founded\_year), data = viz\_data %>% filter(name == "YouTube")) +  
 geom\_text(aes(x = founded\_year, y=2000, label = "YouTube"), data = viz\_data %>% filter(name == "YouTube"), size=4, vjust=-0.3, hjust=-0.3) +  
 geom\_vline(aes(xintercept = founded\_year), data = viz\_data %>% filter(name == "Instagram")) +  
 geom\_text(aes(x = founded\_year, y=2000, label = "Instagram"), data = viz\_data %>% filter(name == "Instagram"), size=4, vjust=-0.3, hjust=-0.3) +  
 geom\_vline(aes(xintercept = founded\_year), data = viz\_data %>% filter(name == "Uber")) +  
 geom\_text(aes(x = founded\_year, y=2000, label = "Uber,"), data = viz\_data %>% filter(name == "Uber"), size=4, vjust=-0.3, hjust=0.8) +  
 geom\_vline(aes(xintercept = founded\_year), data = viz\_data %>% filter(name == "WhatsApp")) +  
 geom\_text(aes(x = founded\_year, y=2000, label = "WhatsApp"), data = viz\_data %>% filter(name == "WhatsApp"), size=4, vjust=-0.3, hjust=-0.3)



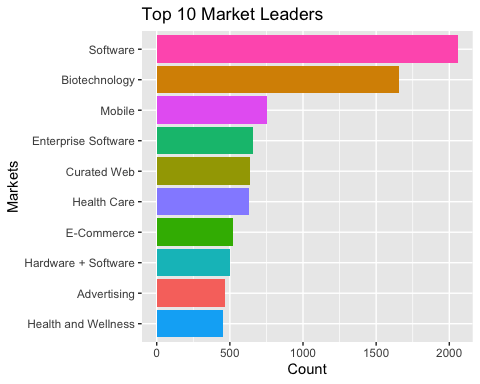
From the graph, we can observe that many successful companies like Facebook, Twitter, Uber, YouTube, WhatsApp intially started as startups after the year 2000. Also, the number of startup companies started in a year increased significantly from the year 2000.

# State Code  
dat %>% count(state\_code)%>% arrange(-n) %>% head(20) %>%  
 ggplot(aes(reorder(state\_code, n), n)) + geom\_col(aes(fill=state\_code)) + coord\_flip() +  
 theme(legend.position="none") +  
 ggtitle("Number of startups in States") + xlab("States") + ylab("Count")



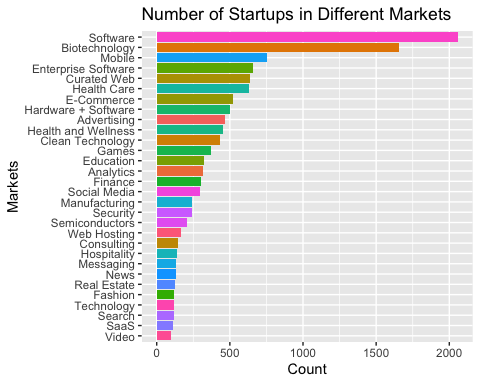
From the graph, we can observe that California (CA) state has the highest number of startups and the New York (NY) state is the second. Also, states like Massachusetts (MA), Texas (TX) and Washington (WA) have more number of startups.

# Top 10 market leaders  
dat %>% filter(market!="") %>% count(market)%>% arrange(-n) %>% head(10) %>%  
 ggplot(aes(reorder(market, n), n)) + geom\_col(aes(fill=market)) + coord\_flip() +  
 theme(legend.position="none") +  
 ggtitle("Top 10 Market Leaders") + xlab("Markets") + ylab("Count")



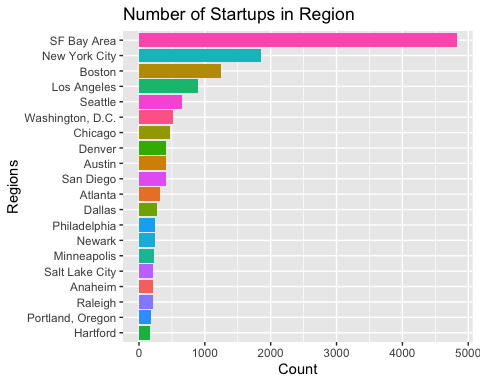
We can observe that many startups are in the field of software, Biotechnology, Health Care, E-Commerce.

# Markets  
dat %>% count(market)%>% arrange(-n) %>% head(30) %>%  
 ggplot(aes(reorder(market, n), n)) + geom\_col(aes(fill=market)) + coord\_flip() +  
 theme(legend.position="none") +  
 ggtitle("Number of Startups in Different Markets") + xlab("Markets") + ylab("Count")



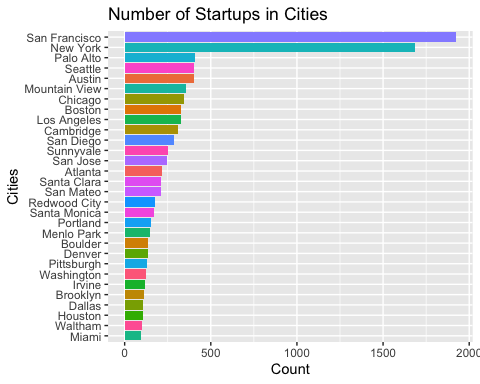
We can observe that many startups are in the field of software, Biotechnology, Health Care, E-Commerce.

# Regions  
dat %>% count(region)%>% arrange(-n) %>% head(20) %>%  
 ggplot(aes(reorder(region, n), n)) + geom\_col(aes(fill=region)) + coord\_flip() +  
 theme(legend.position="none") +  
 ggtitle("Number of Startups in Region") + xlab("Regions") + ylab("Count")



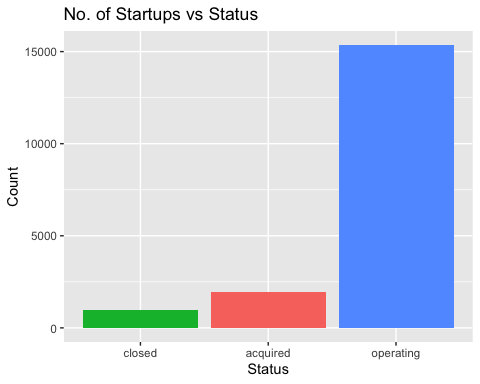
Regions like San Francisco Bay Area, New York City, Boston, Los Angeles and Washington DC have many number of startups.

# Cities  
dat %>% count(city)%>% arrange(-n) %>% head(30) %>%  
 ggplot(aes(reorder(city, n), n)) + geom\_col(aes(fill=city)) + coord\_flip() +  
 theme(legend.position="none") +  
 ggtitle("Number of Startups in Cities") + xlab("Cities") + ylab("Count")



San Francisco and New York have many more number of startups compared to other cities.

# No. of Startups vs Status  
dat %>% count(status)%>% arrange(-n) %>%  
 ggplot(aes(reorder(status, n), n)) + geom\_col(aes(fill=status)) +  
 theme(legend.position="none") +  
 ggtitle("No. of Startups vs Status") + xlab("Status") + ylab("Count")



We can observe that many startups established are still operating as of 2014. Also, many have been acquired and only few are closed which might operate in future.

docs <- apply(dat['category\_list'], 2, function(y) str\_sub(y, 2, -2))  
docs <- apply(docs, 2, function(y) unlist(strsplit(y, "\\|")))  
docs <- Corpus(VectorSource(docs))

# Convert the text to lower case  
docs <- tm\_map(docs, content\_transformer(tolower))

## Warning in tm\_map.SimpleCorpus(docs, content\_transformer(tolower)):  
## transformation drops documents

# Remove numbers  
docs <- tm\_map(docs, removeNumbers)

## Warning in tm\_map.SimpleCorpus(docs, removeNumbers): transformation drops  
## documents

# Remove english common stopwords  
docs <- tm\_map(docs, removeWords, stopwords("english"))

## Warning in tm\_map.SimpleCorpus(docs, removeWords, stopwords("english")):  
## transformation drops documents

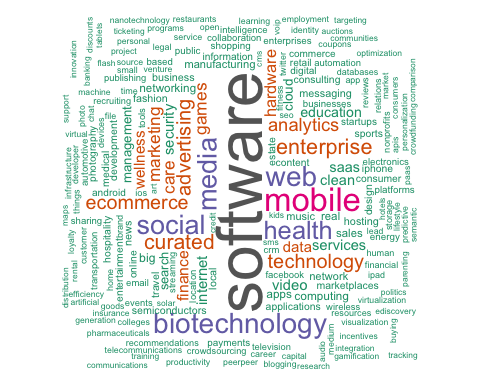
# Remove your own stop word  
# specify your stopwords as a character vector  
#docs <- tm\_map(docs, removeWords, c("blabla1", "blabla2"))   
# Remove punctuations  
docs <- tm\_map(docs, removePunctuation)

## Warning in tm\_map.SimpleCorpus(docs, removePunctuation): transformation drops  
## documents

# Eliminate extra white spaces  
#docs <- tm\_map(docs, stripWhitespace)  
# Text stemming  
# docs <- tm\_map(docs, stemDocument)

dtm <- TermDocumentMatrix(docs)  
m <- as.matrix(dtm)  
v <- sort(rowSums(m),decreasing=TRUE)  
d <- data.frame(word = names(v),freq=v)

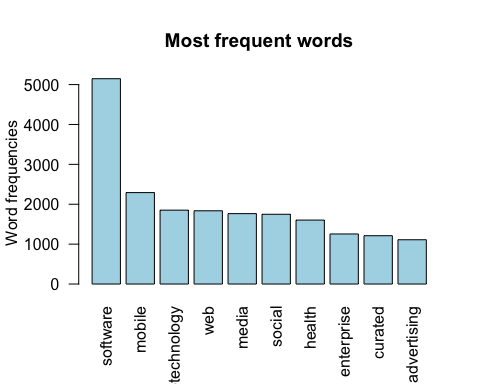
set.seed(1234)  
wordcloud(words = d$word, freq = d$freq, min.freq = 1,  
 max.words=200, random.order=FALSE, rot.per=0.35,   
 colors=brewer.pal(8, "Dark2"))



findAssocs(dtm, terms = "health", corlimit = 0.3)

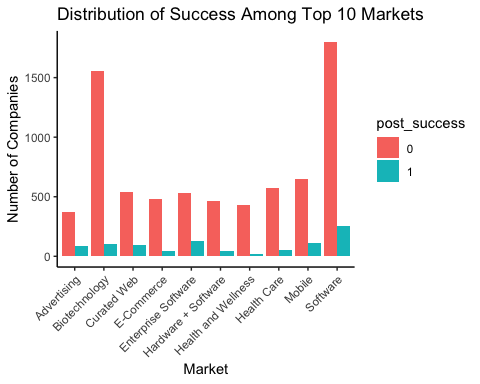
## $health  
## care wellness   
## 0.72 0.63

#Plot word frequencies  
barplot(d[1:10,]$freq, las = 2, names.arg = d[1:10,]$word,  
 col ="lightblue", main ="Most frequent words",  
 ylab = "Word frequencies")



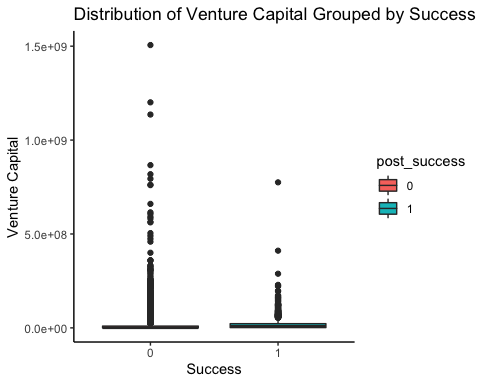
d = data.frame(dat['market'],dat['post\_success'],dat['status'],dat['founded\_year'])  
d = subset(d, dat$market == "Software"|dat$market == "Biotechnology"|dat$market == "Mobile"|dat$market == "Enterprise Software"|dat$market == "Curated Web"|dat$market == "Health Care"|dat$market == "E-Commerce"|dat$market == "Hardware + Software"|dat$market == "Advertising"|dat$market == "Health and Wellness")

ggplot(data = d, aes(x=market, fill = post\_success)) +   
 theme\_classic() +  
 ggtitle("Distribution of Success Among Top 10 Markets") +   
 xlab("Market") +  
 ylab("Number of Companies") +   
 geom\_bar(position="dodge") +   
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



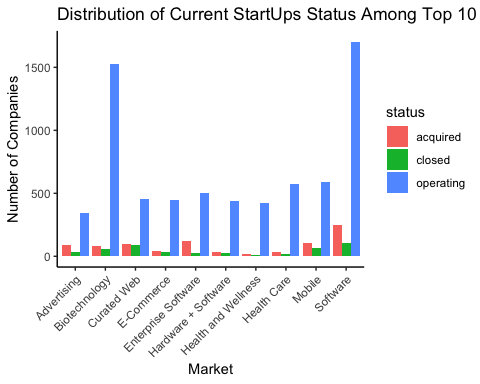
We can observe that the startups in the field of software, advertising and biotechnology have more success rate when compared to others.

ggplot(data=dat, aes(x=post\_success, y=venture, fill = post\_success)) +   
 theme\_classic() +  
 geom\_boxplot() +   
 ggtitle("Distribution of Venture Capital Grouped by Success") +   
 xlab("Success") +  
 ylab("Venture Capital")



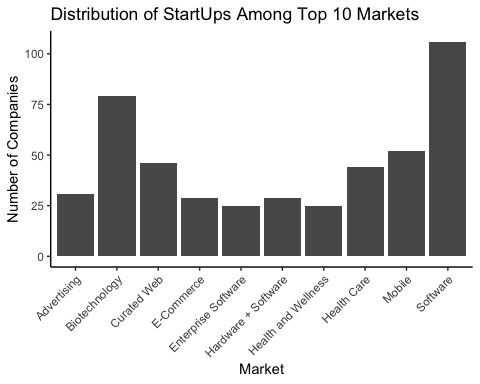
We can observe that successful companies are backed by Venture capital. Even top companies like facebook, youtube, google are backed by venture capital in the initial stages.

ggplot(data = d, aes(x=market, fill = status)) +   
 theme\_classic() +  
 ggtitle("Distribution of Current StartUps Status Among Top 10 Markets") +   
 xlab("Market") +  
 ylab("Number of Companies") +   
 geom\_bar(position="dodge") +   
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))

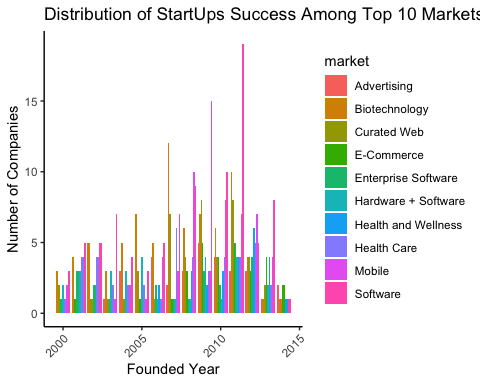


d1 = subset(d, d$founded\_year == 2000:2014)

ggplot(data = d1, aes(x=market, fill = founded\_year)) +   
 theme\_classic() +  
 ggtitle("Distribution of StartUps Among Top 10 Markets") +   
 xlab("Market") +  
 ylab("Number of Companies") +   
 geom\_bar(position="dodge") +   
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))

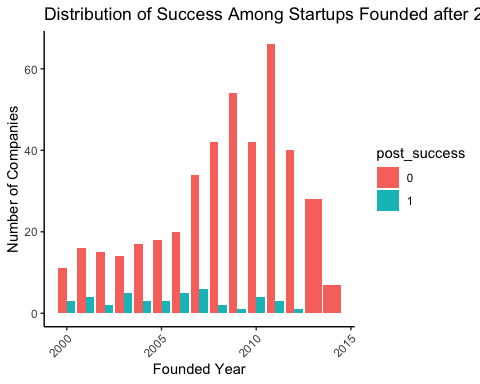


ggplot(data = d1, aes(x=founded\_year, fill = market)) +   
 theme\_classic() +  
 ggtitle("Distribution of StartUps Success Among Top 10 Markets") +   
 xlab("Founded Year") +  
 ylab("Number of Companies") +   
 geom\_bar(position="dodge") +   
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



We can observe that more number of successful startups are in the field of software and biotechnology

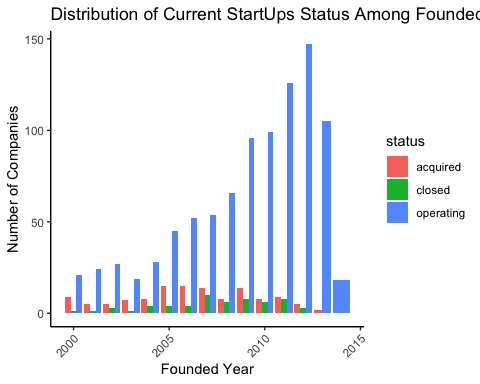
ggplot(data = d1, aes(x=founded\_year, fill = post\_success)) +   
 theme\_classic() +  
 ggtitle("Distribution of Success Among Startups Founded after 2000") +   
 xlab("Founded Year") +  
 ylab("Number of Companies") +   
 geom\_bar(position="dodge") +   
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



We can observe that the statups which have started before 2007 have good success rate. Between 2008-2010, the success rate of startups has reduced. After 2010, the success rate of startups increased.

d2 = subset(dat, dat$founded\_year == 2000:2014)

ggplot(data = d2, aes(x=founded\_year, fill = status)) +   
 theme\_classic() +  
 ggtitle("Distribution of Current StartUps Status Among Founded Year") +   
 xlab("Founded Year") +  
 ylab("Number of Companies") +   
 geom\_bar(position="dodge") +   
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



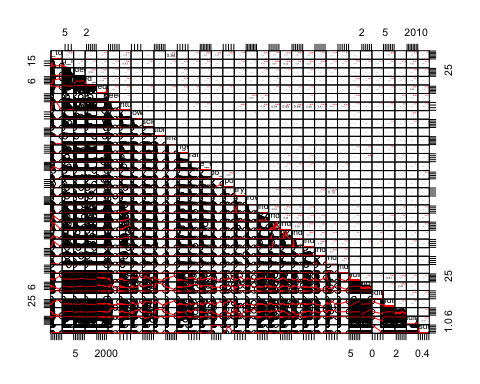
We can observe that every year few successful companies are acquired, whereas as few are closed and the remaining stay operated.

# Correlation Analysis

datCorr <- dat  
datCorr$name <- NULL  
datCorr$category\_list <- NULL  
datCorr$market <- NULL  
datCorr$status <- NULL  
datCorr$state\_code <- NULL  
datCorr$region <- NULL  
datCorr$city <- NULL  
datCorr$founded\_quarter <- NULL  
datCorr$post\_success <- as.numeric(as.character(datCorr$post\_success))

We removed columns which are of type “character”.

chart.Correlation(datCorr,   
 histogram=TRUE,   
 method = 'pearson',  
 pch=19)



rcorr(as.matrix(datCorr))  
## post\_success  
## funding\_total\_usd 0.0230   
## funding\_rounds 0.0000   
## founded\_at 0.0002   
## founded\_month 0.6359   
## founded\_year 0.0000   
## seed 0.0000   
## venture 0.0000   
## equity\_crowdfunding 0.5413   
## undisclosed 0.2209   
## convertible\_note 0.4312   
## debt\_financing 0.7988   
## angel 0.9547   
## grant 0.2661   
## private\_equity 0.5921   
## post\_ipo\_equity 0.0000   
## post\_ipo\_debt 0.5306   
## secondary\_market 0.5257   
## product\_crowdfunding 0.4366   
## round\_A 0.0000   
## round\_B 0.0000   
## round\_C 0.0000   
## round\_D 0.0363   
## round\_E 0.5459   
## round\_F 0.7626   
## round\_G 0.4709   
## round\_H 0.7219   
## first\_funding\_day 0.0000   
## first\_funding\_month 0.0310   
## first\_funding\_year 0.0000   
## last\_funding\_day 0.7852   
## last\_funding\_month 0.0003   
## last\_funding\_year 0.0000   
## post\_success

From the last matrix, we can see that **funding\_rounds**, **founded\_at**, **founded\_year, seed, venture, post\_ipo\_equity, round\_A, round\_B, round\_C, first\_funding\_day, first\_funding\_year, last\_funding\_month, last\_funding\_year are all significantly associated with outcome variable** post\_success (p-values < 0.01).

# Results and Conclusion

* California (CA) state has the highest number of startups and the New York (NY) state is in the second place.
* Most of the startups are in the field of software, Biotechnology, Health Care, E-Commerce.
* Regions like San Francisco Bay Area, New York City, Boston, Los Angeles and Washington DC have higher number of startups compared to other regions.
* Startups in the field of software, advertising and biotechnology have more success rate when compared to others.
* Successful companies are backed by Venture capital. Even top companies like facebook, youtube, google are backed by venture capital in the beginning.
* Every year few successful companies are acquired, whereas as few are closed and the remaining stay operated.
* More number of funding rounds, seed amount, venture capital, post IPO equity contributes to the success of a startup.
* More funding in the initial rounds A,B,C can contribute to the success of a stratup.
* We used 3 algorithms, namely the KNN, logistic regression and the gradient boosting model (GBM) to train 3 different machine learning models. Based, on AUC score we concluded that GBM is the best model.