Analysis of factors that affect startup success

Introduction to Data Analysis, UCSC Extension

ChiaYu Kuo / Patricia Huang - March 14, 2020

Introduction

In the article "Why some startups succeed (and why most fail)", the author analyzed the results of several studies to come up with an overall list that help ensure the survival of a startup business. He came up with the following conclusions: 1.) Have a plan 2.) Believe in what you are doing; be persistent and have perseverance 3.) Willingness to be flexible to change direction 4.) Have a leadership team with good business knowledge and experience that complement the technical side of the company 5.) Cultivate good mentors. The writer also added that "…the pace at which the United States produces $100-million companies has been stable over the last 20 years despite changes in the economy." We wondered about other factors besides those mentioned in the article.

Problem Definition

We found a dataset in Kaggle "Startup Investments (Crunchbase)" that we could use to answer our question: What are the important factors that influence the success of a startup? Could the location, founding year, market, etc. be relevant also?

Data Description

The original dataset has 49438 observations of 39 variables. We cleaned the dataset and removed the null values resulting in a dataset of 21840 observations. We further cleaned the dataset to focus our efforts on specific features that matched our problem definition resulting in 19 variables:

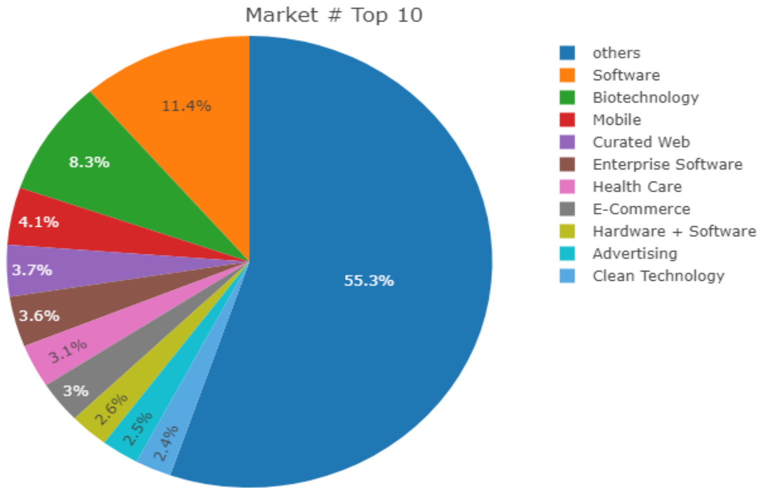
```
$ market : chr  "News" "Software" "Health and
           Wellness" "Health and Wellness" ...
$ funding_total_usd : chr  "17,50,000" "-" "-"
           "17,50,000" ...
$ status : chr  "acquired" "operating" "operating"
           "operating"
$ country_code : chr  "USA" "USA" "USA" "USA" ...
$ state_code : chr  "NY" "IL" "CA" "NJ" ...
$ region : chr  "New York City" "Springfield, Illinois"
           "Los Angeles" "Newark" ...
$ city : chr  "New York" "Champaign" "Los Angeles"
           "Iselin" .
$ founded_quarter: chr  "2012-Q2" "2010-Q1"
           "1986-Q1" "1984-Q1" ...
```

```
$ founded_year : num  2012 2010 1986 1984
           2001 ...
$ seed : num  1750000 0 0 0 0 40000 15000 0
           420000 750000 ...
$ venture : num  0 0 0 0 0 ...
$ 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 1750000 0 0 0 0 0….
$ debt_financing  : num  0 0 0 0 2050000 ...
$ 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 ...
$ product_crowdfunding: num  0 0 0 0 0 0 0 0 0 0 …
```

Descriptive/Exploratory Statistics

- Market -

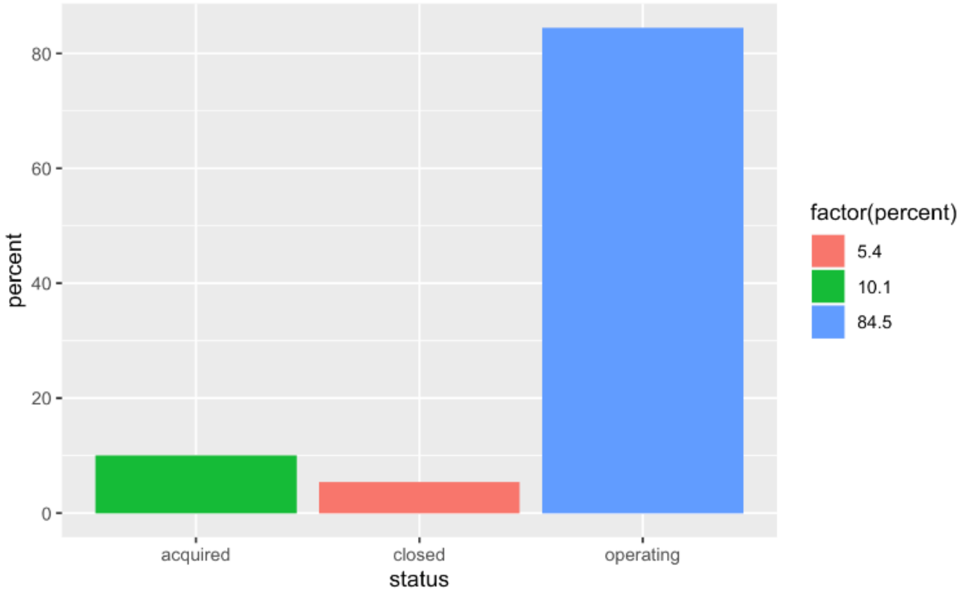
In this graph we have the percentages of the top markets. We have combined the smaller markets into ‘Others’. The top 3 markets are : Software, Biotechnology and Mobile.



```
perCent <- mydata$market
marKet <- cbind(freq = table(perCent), percentage = prop.table(table(perCent))*100)
m1 <- group_by(mydata, market)
m2 <- summarise(m1, firms = n())
m3 <- select(arrange(m2,firms), market, firms)
m4 <- tail(m3, 10)
total_firms <- 21240
m5 <- mutate(m4, percent = (firms/total_firms)*100)
m6 <- data.frame(group=c("Clean Technology", "Advertising", "Hardware + Software", "E-Commerce", "Health Care", "Enterprise Software",
"Curated Web", "Mobile", "Biotechnology", "Software", "others"), perCent = c(2.4, 2.5, 2.6, 3.0, 3.1, 3.6, 3.7, 4.1, 8.3, 11.4, 55.3))
m7 <- plot_ly(m6, labels = ~group, values = ~perCent, type = 'pie') %>%
layout(title = "Market # Top 10",  xaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE), yaxis = list(showgrid = FALSE,
zeroline = FALSE, showticklabels = FALSE))
```

- Status -

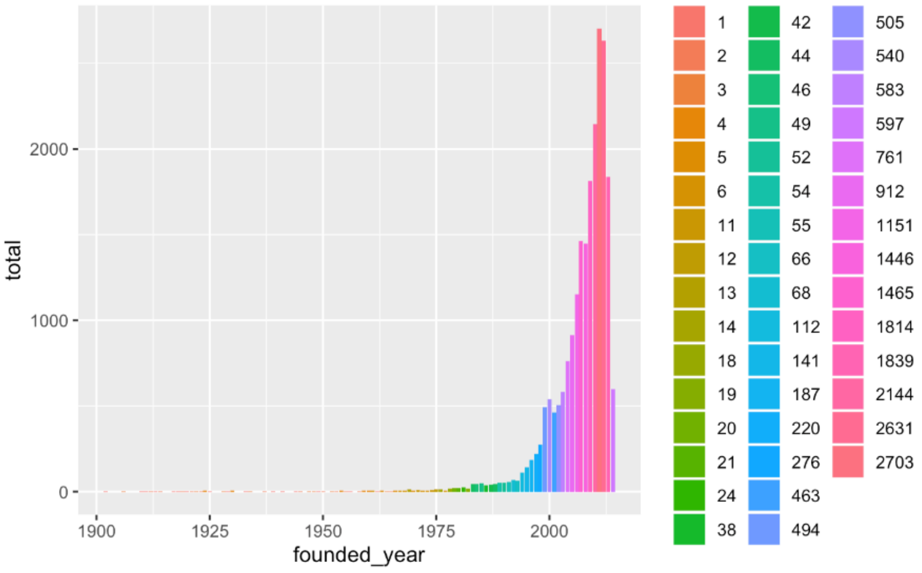
Here we want to focus on the successful startups with 84.5% having status=operating.



```
s1 <- group_by(mydata, status)
s2 <- summarise(s1, total = n())
s3 <- select(arrange(s2,total),status, total)
s4 <- mutate(s3, percent = round(total / sum(total)*100, digits = 1))
ggplot(data = s4, aes(x=status, y=percent, fill=factor(percent)))+geom_bar(stat="identity")
```

- Year Founded -

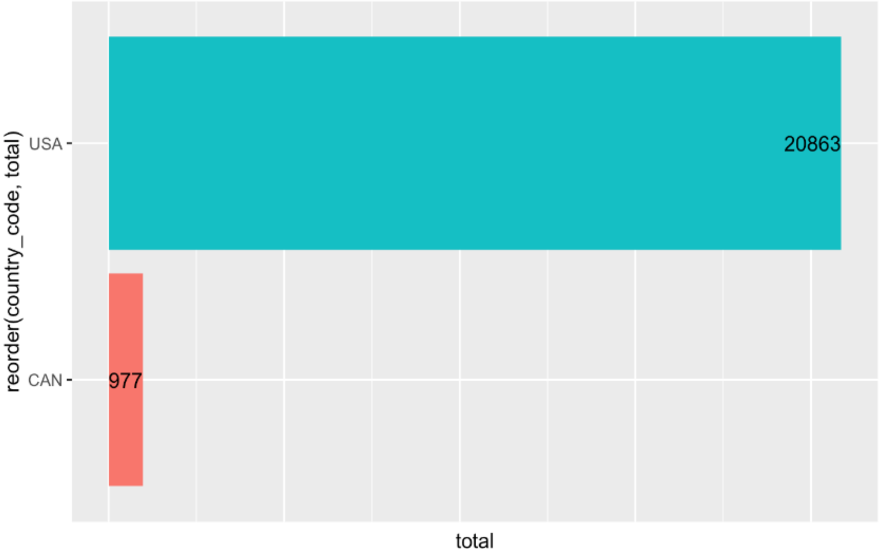
We have the peak time period for startups during the time period 2011-2012.



```
# founded year
f1 <- group_by(mydata, founded_year)
f2 <- summarise(f1, total = n())
ggplot(data = f2, aes(x=founded_year, y=total, fill=factor(total)))+geom_bar(stat="identity")
```

- Country -

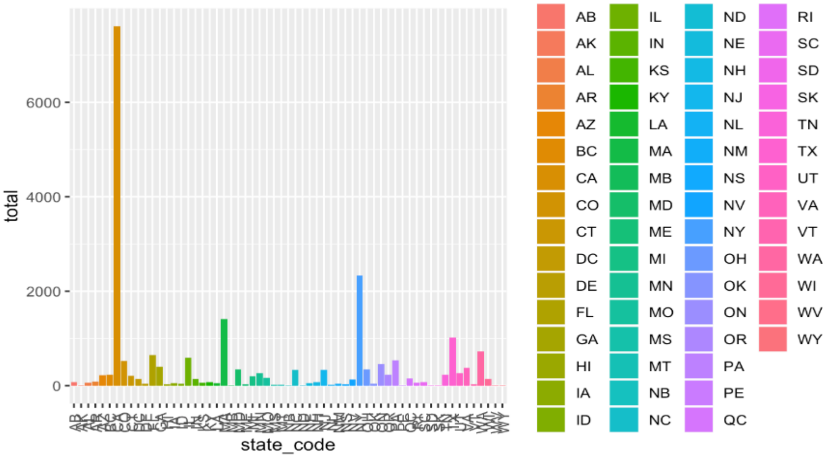
The USA dominates in the data with 20863 startups.



```
#country
c1 <- group_by(mydata, country_code)
c2 <- summarise(c1, total = n())
c3 <- select(arrange(c2,total), country_code, total)
ggplot(c3, aes(x = reorder(country_code, total), y = total, fill = country_code))+
  geom_bar(stat = "identity") + geom_text(aes(label = total), hjust = 1)+ coord_flip()+ theme(axis.text.x = element_blank(),
        axis.ticks.x = element_blank(), legend.position = "none")
```

- State -

The top 3 states with the most startups are California, New York and Massachusetts.



```
# states
st1 <- group_by(mydata, state_code)
st2 <- summarise(st1, total = n())
st3 <- select(arrange(st2,total), state_code, total)
ggplot(data = st3, aes(x=state_code, y=total, fill=factor(state_code)))+geom_bar(stat="identity")+ theme(axis.text.x =
element_text(angle=90, vjust=0.6))
```

Diagnostic Statistics

Based on our exploratory graphs, we will focus on the following for our analysis:

- Status: Operating
- Year founded: 1995-2014
- State: Top 10
- Markets: Top 10

We performed the Chi Square Test on these categorical variables to discover their dependency on status = “operating”. Here are the results:

- State p value<0

```
Number of cases in table: 17611

Number of factors: 2

Test for independence of all factors:

 Chisq = 9.804e-30, df = 0, p-value = 0

 # Chi Square - state code: summary(table(mydata1$status, mydata1$state_code))
```

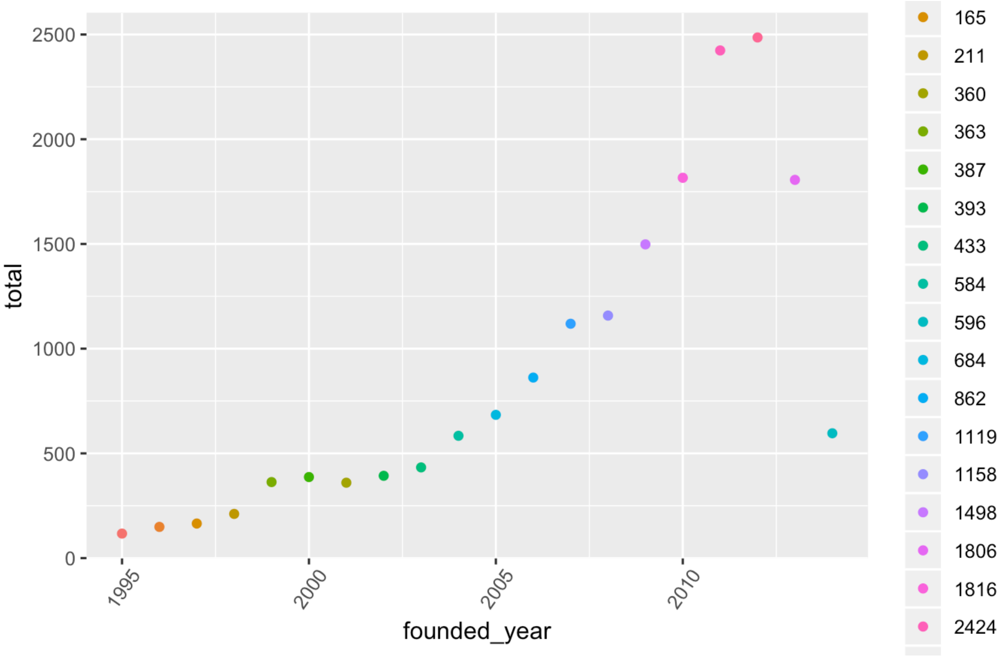
- Market p value<0

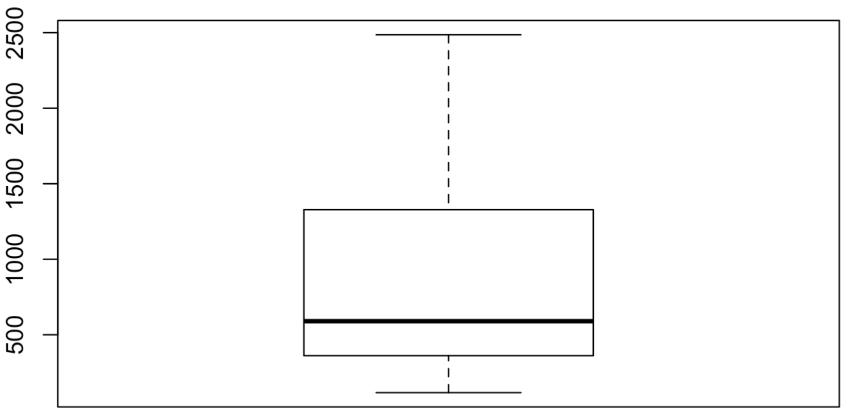
```
Number of cases in table: 17611

Number of factors: 2

Test for independence of all factors:

 Chisq = 1.5234e-29, df = 0, p-value = 0

 Chi-squared approximation may be incorrect

 # Chi Square - market: summary(table(mydata1$status, mydata1$market))
```

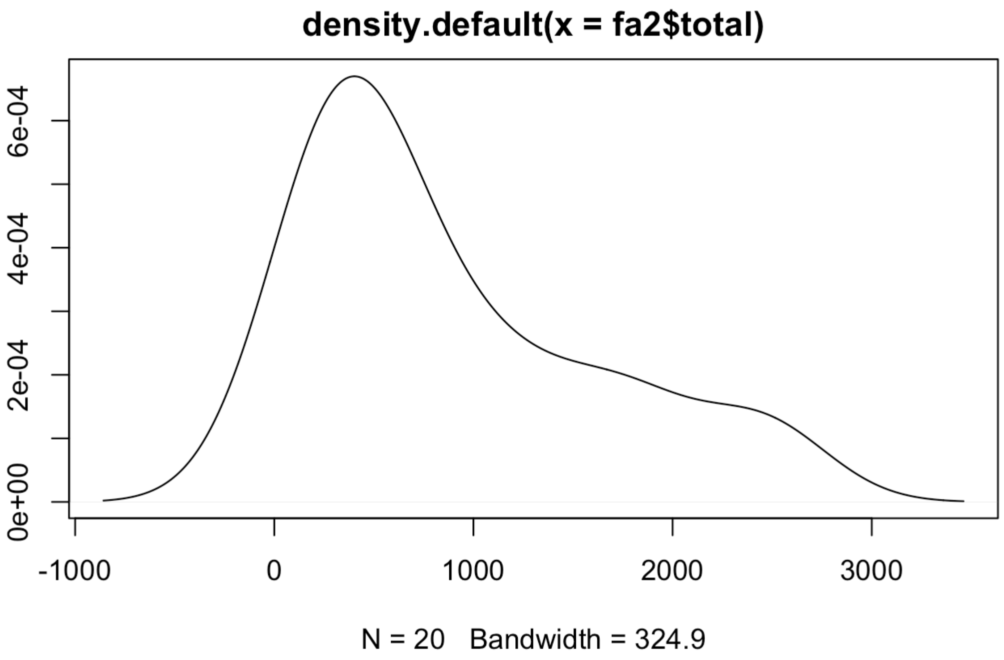
- Founded year p value<0

```
Number of cases in table: 17611

Number of factors: 2

Test for independence of all factors:

 Chisq = 8.531e-29, df = 0, p-value = 0

# Chi Square - founder_year: summary(table(mydata1$status, mydata1$founded_year))
```

For additional analysis on the year founded, we have the summary statistics, box plot and density graph:

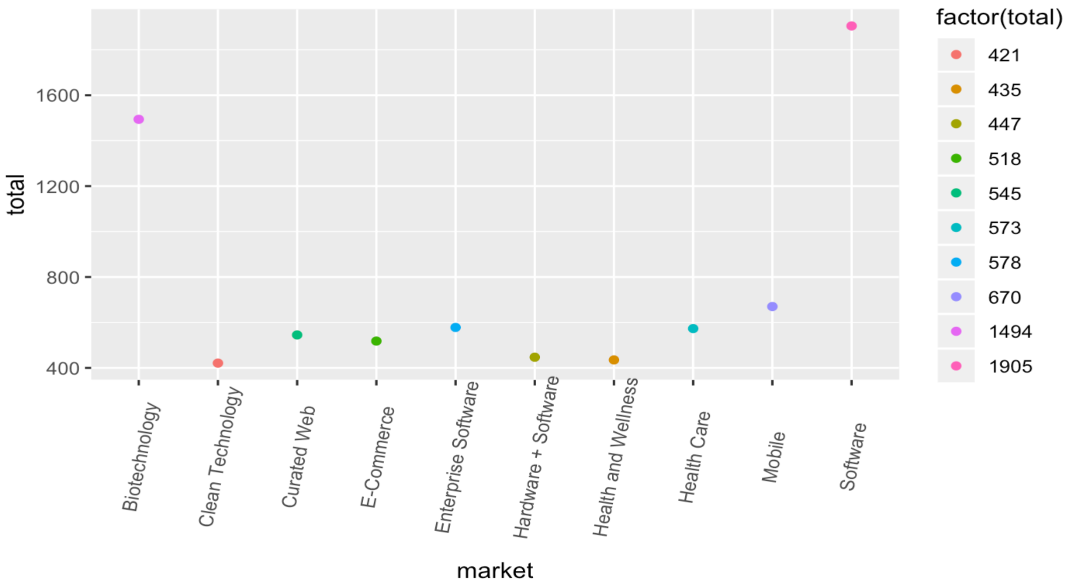


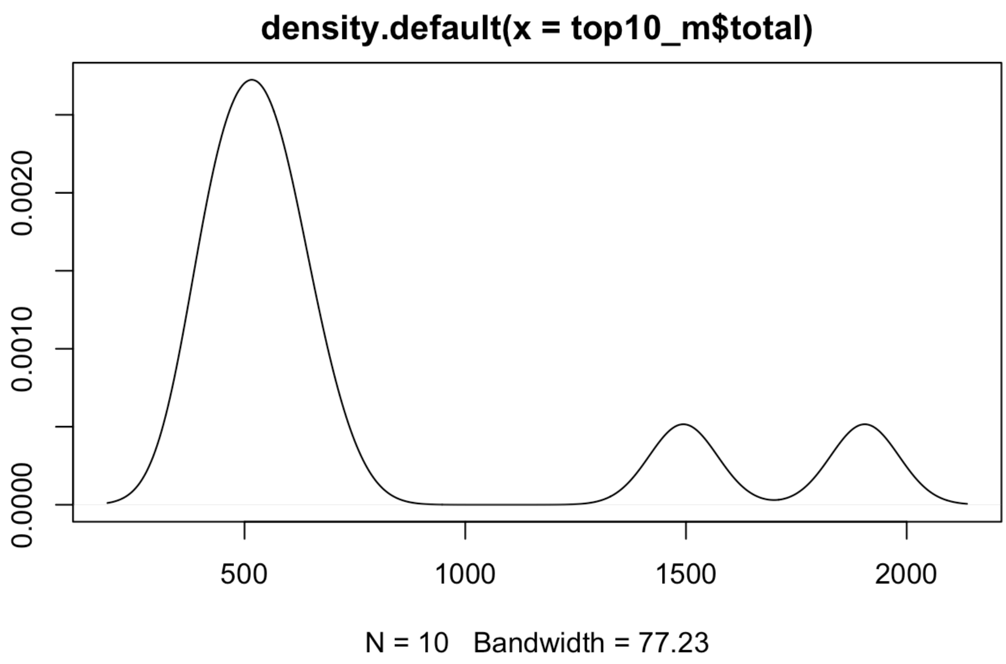




```
# # Density and Exploratory Plot - founded year > 1995
fa1 <- group_by(mydata1, founded_year)
fa2 <- summarise(fa1, total = n())
fa3 <- select(arrange(fa2,total), founded_year, total)
fa4 <- ggplot(fa3, aes(x = founded_year, y = total, color = factor(total))) + geom_point() + theme(axis.text.x =
element_text(angle=55, vjust=0.6))
print(fa4)
boxplot(fa2$total)
fa5 <- density(fa2$total)
```

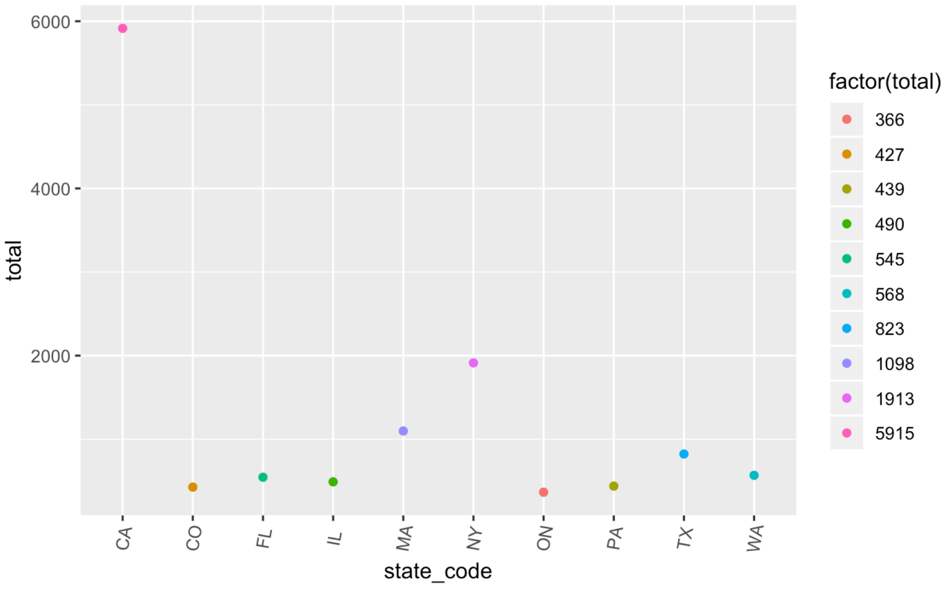
Here are the summary statistics and graphs for the Top 10 Markets:

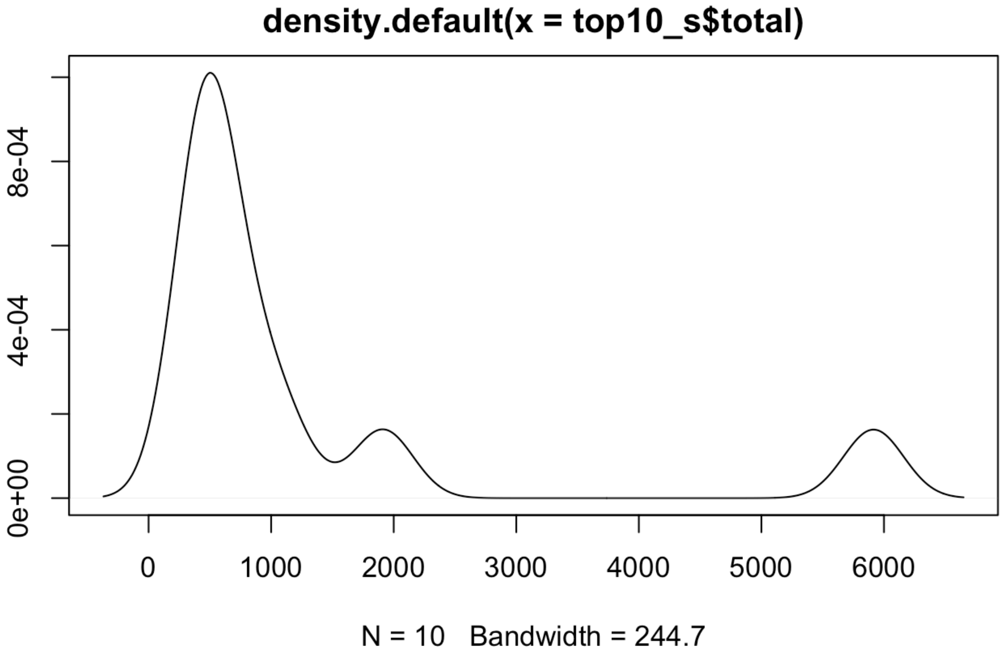




```
# Density and Exploratory Plot – market (top 10)
ma1 <- group_by(mydata1, market)
ma2 <- summarise(ma1, total = n())
ma3 <- select(arrange(ma2,total), market, total)
top10_m <- tail(ma3, 10)
top10_m
m <- ggplot(top10_m, aes(x = market, y = total, color = factor(total))) + geom_point() + theme(axis.text.x =
element_text(angle=80, vjust=0.6))
print(m)
top10_m <- density(top10_s$total)
```

Here are the summary statistics and graphs for the Top 10 States:





```
# Density and Exploratory Plot – state (Top 10)
sa1 <- group_by(mydata1, state_code)
sa2 <- summarise(sa1, total = n())
sa3 <- select(arrange(sa2,total), state_code, total)
top10_s <- tail(sa3, 10)
top10_s
s <- ggplot(top10_s, aes(x = state_code, y = total, color = factor(total))) + geom_point() + theme(axis.text.x =
element_text(angle=80, vjust=0.6))
print(s)
d_top10 <- density(top10_s$total)
```

Findings

We reject the null hypotheses and conclude that market, region, state have a dependency with the success of a startup status = “operating”. We tried to do further analysis on the available categorical and continuous variables using ggpairs, however, we could find anything relevant to our main problem. We could not do extra correlation tests to support our p values above because our dataset, with its large number of observations, requires extensive transformation to get more continuous data.

Recommendations

In Kaggle the dataset we used is for a classification problem. In order to do regression analysis, this dataset needs to be extended with more continuous data to boost support of the Chi Square test results. We attempted to compute the summary numbers (min/max, mean, median, etc.) for the founding years, top 10 markets and states to find some meaningful results. We were able to find that years 1999-2008 gave a normal distribution, however, for the top 10 markets and states we had doubts about the resulting numbers. We discovered after researching and attempting different creative ways of analysis on this dataset that we need more continuous data to support our Chi Square results.