Analysis and Forecast of Start-Up Companies

Stat 405 Final Project

Aayush Raman
Yinsen Miao
Gabriel Rubio Breternitz

December 9, 2013

Contents

1 Introduction

Crunchbase [1] released a dataset detailing approximately 18,500 startups since 1920. The dataset contains a number of interesting features including startup locations, rounds of investments by type, early investors, monetary value of investments, company status and outcome. We recognize that identifying a promising startup can be a tricky problem, because the most innovative startups tend to disrupt existing markets. However, this dataset allows us to "follow the money" and gain insights into what a successful startup looks like, with a focus towards how an investor could exploit this knowledge for monetary gain. We follow the startups all the way to their IPO and beyond into the stock market to see which companies that came from humble beginnings transformed themselves into large and financially stable companies ruling the market today.

2 Dataset

The Crunchbase website is a part of Techcrunch that focuses on being the definitive database of start-ups, entrepreneurs, and investors and making information about the Startup world available to everyone. We used the dataset that was published on November 4th, 2013 and our central assumption made is that the companies' funding information is reliable and trustworthy. The dataset consists of 3 different csv files that contains information about Companies, Investments and Rounds of Funding. The data files have unique variables but contain common features among them. The description of the data files are as follows: "companies.csv" contains 18505 startup companies' profile and 18 features, which includes companies' status, total funding received, geographical locations and time events of various investments by investors; "rounds.csv" includes 33102 observations and 13 variables explaining funding date, amount of funding and fund type of each round; "investment.csv" has 55240 instances and 20 features which mainly deals with the information of the investors. After cleaning the duplicated instances, we analyzed the conjunction of those three datasets. We used the stock information from Yahoo Finance[2] to get the information about the public companies in our dataset. The description is given in the Post IPO section.

3 Trends in Investments

While working on the time series dataset, we chose to look at two important factors: total investment and number of startup companies in a given year. These two factors are important because they can be indicative of the general state of the economy in the United States as well as the success of the startup companies. Figure **??** provides detailed information about a pattern of investment since 1987. Before 1987, the dataset contained information about less than 50 startups companies and the dollar amount of investments was also minimal. Therefore, we concentrated on the companies that were founded post-1987. One interesting thing to note is that when the number of startups is low, the total dollar amount of investments is also low. This is rather intuitive, but it is still important to confirm. Also notice that many of the most heavily funded industries today are web-related, which shows us how startup funding exploded with the rise to prominence of software companies. Also note that total investment rose briefly around when Google started in 1998. However, shortly after, investment declined in reaction to the implosion of the market following the bursting of the dot-com bubble or internet bubble.

America experienced an overall increase in investment after 2004, which coincides with when Facebook was founded. One possible reason for the swift increase of investment could be the increased popularity of Internet usage for social networking and marketing through sites like Orkut, Linkedin

[1] http://www.crunchbase.com/
[2] http://finance.yahoo.com/

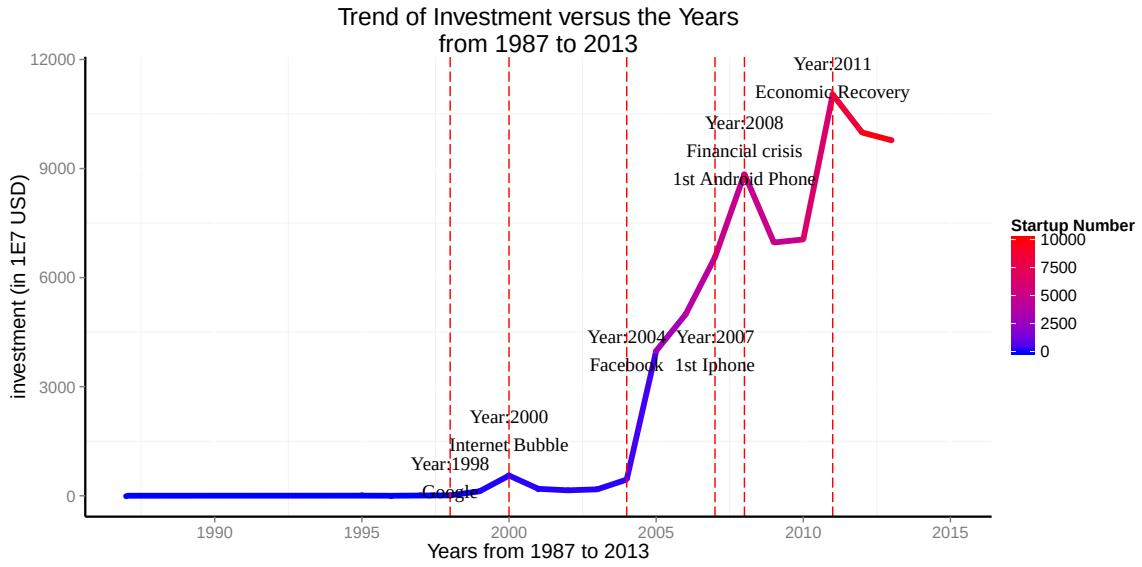

Figure 1: **The plot above shows the number of investment in United States from 1987 to 2013. The color filling indicates the number of Startup companies in that year. The six vertical red dotted lines represent important time events.**

and small online businesses. In the mid and late 90's, investors seemed to want to look for new and better search engines. We may suggest the investors were in search of the next Google in the area of social networks leading to an exponential increase in the investment. From year 2005, the slope of investment decreased since the U.S. housing bubble bursted and it peaked in 2006. This was the time when the loans and investment from financial institutions decreased. While the number of startup companies was booming till 2007, the interbank credit crisis led to a 23% decrease in total investment. The software companies seemed less affected by the financial recession however, since the number of software and mobile startup companies actually still increased. One explanation could be the ascent of smart phones like the iPhone and Android phones which created entirely new markets for app development and mobile applications. Total investment remained rather low from 2009 to 2010 and then it revived as the economic conditions improved. After investment reached its highest point in 2011, the graph again shows a decrease in seed investments and number of new startups founded. We may say that the investors are struggling to attract the best founders and make seed investments in promising companies. Also, we can say that most young accelerators and incubators seem destined to fail because of the overcrowded market for early stage funding. It appears that the market may have found a saturation point; in the past, investment tended to increase whenever a new market emerged, but there hasn't been an entirely new market created due to technological advancement since the mobile app market emerged. Should another new technology prove to be transformative such that it creates more room for startups, then we could assume another spike in investments and number of startups founded.

4 Emerging Industries

The breakdown of the companies in terms of status is described in Table **??**. We have very few data points for the companies that were acquired, IPOed (went public) or closed down. Most companies in the dataset are currently operating which means that they are private running companies based on the various rounds of funding and investments. The hottest categories are Software, Biotech, Web and Mobile. The heat map in Figure **??** shows that the startups in the area of clean technology, E-

Status of the Company	Number of Companies
OPERATING	15078
ACQUIRED	1879
CLOSED	1170
IPO	377

Table 1: **Status of Companies**

commerce and mobile sector has become increasingly popular within the last decade. The web, mobile and video game industries have increased thanks in large part to the growth of the software industry.

Company Type	Operating	IPO	Acquired	Closed
Software	2275	37	358	154
Biotech	1836	119	119	80
Web	1139	16	289	262
Mobile	945	15	145	73
Enterprise	933	13	147	41
E-Commerce	609	9	47	56
Advertising	581	12	107	44
Cleantech	544	22	44	41
Video-Games	523	5	94	86
Hardware	518	16	44	40

Table 2: **Status for various categories of companies**

However, Table **??** shows that the Web, E-commerce, Cleantech, Video games and Hardware industry have an equal chance of being acquired or getting shut down. (Note that "Cleantech" industry refers to companies that look to be innovative in the area of biofuels, electric vehicles, solar panels and advanced nuclear technologies). One of the reasons for this could be the presence of already big established companies especially in Web, E-commerce and Video games sector whereas, in the case of Hardware and Cleantech industry, the innovations take significant capital and require more time to develop and commercialize their products successfully. Since those types of innovation require more investments and funding by the investors. Therefore, there is a huge amount of sustainability issues.

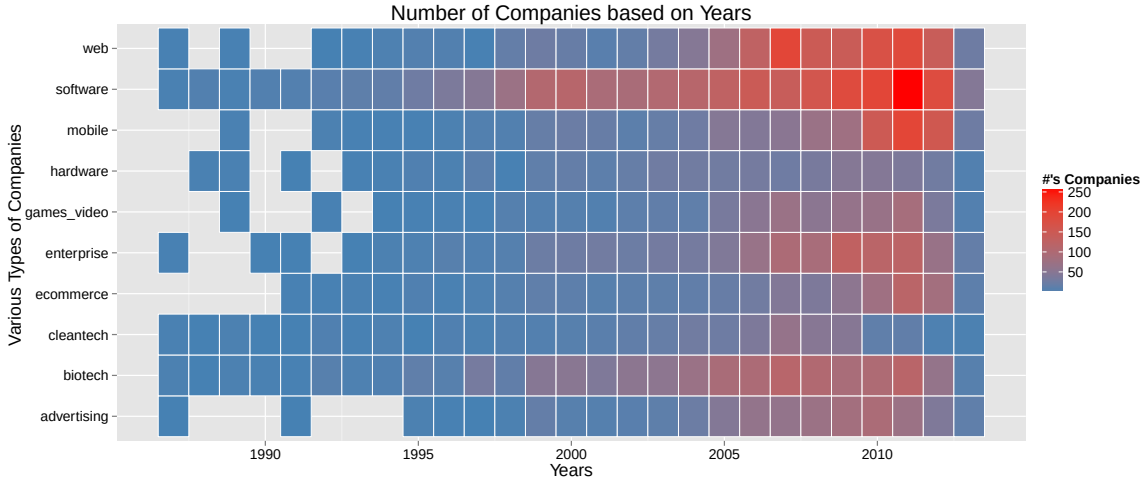

Figure 2: **The heat map shows the temporal trend of different types of industry**

5 Geographical Trends

The dataset contains 42 categories of companies, spread across 601 locations. The map in Figure **??** shows the state-by-state distribution of total investment dollars invested. Massachusetts, California, Texas and Washington are the top states for startups, since investors in those four states collectively raise the most money for new companies. Also note that in addition to a strong presence in the software industry, Massachusetts and SF Bay are inundated with Biotech companies as well. California has a large amount of small and diverse biotech companies whereas Massachusetts has a more focused segment of the Biotech industry, with most companies working in the field of drug discovery and development. Therefore, the investment is large as compared to other states. Texas and Washington also raise very high dollar amounts due to their advantageous-to-business tax environments - most notably, the lack of a state personal or corporate income tax.

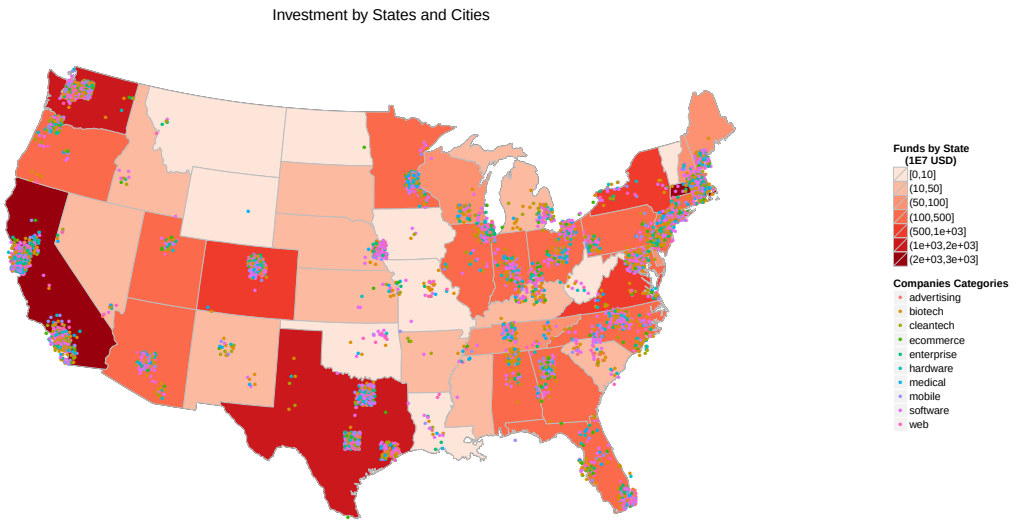

Figure 3: **The maps shows the geographic location of the investment and investment states received from 1987 to 2013. We can see that Startup companies grouped into regions and by sum of investment received by each state.**

As one might expect, most of the startups in our data set are clustered around major cities. However, we must be careful when making assumptions from this map, because our data set is biased towards companies that received some form of venture funding. Thus, there may be many startups in other places that did not receive funds, so they are not represented here.

The heat map in Figure **??** shows the top ten regions for startups and how they represent the top ten types of industries. It shows that San Francisco (SF) Bay area, Greater New York, Greater Los Angeles and Boston are the most popular hubs for startup companies of virtually every top industry. The software industry especially is omnipresent in all the ten region, with SF Bay area not surprisingly being the most popular one. The Biotech industry is dominant in SF Bay, Boston, Greater Los Angeles and San Diego, Washington DC and Seattle. We believe this could be due to the presence of elite universities and research institutes like Harvard and MIT in the Boston Area, Stanford and UC

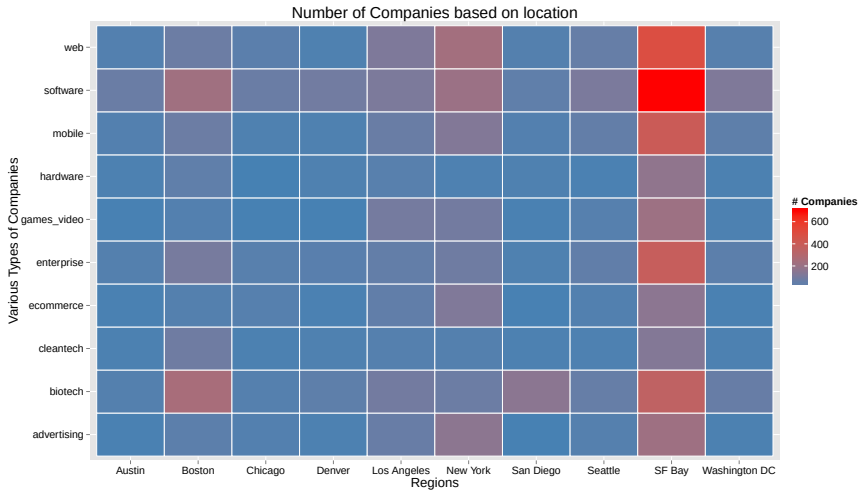

Figure 4: **The heat map shows the geographical trend of different types of industry**

Berkeley around the SF Bay area, the University of Washington in Seattle, and the National Institute of Health and John Hopkins University around the DC area.

Different areas seem to engender different types of startups. New York has a strong showing in the advertising, web and software industries, but seems bizarrely lacking in Biotech and Cleantech startups one might expect to find in such a large and investor-dense region. Boston also has a noteworthy amount of "enterprise" companies, which are geared more towards facilitating business functions rather than selling products or services to consumers. This could help explain the region's remarkably dominant software industry, which rely a lot on the business-to-business types of services.

6 Type of Investment

The type of funding is an important criteria for the entrepreneurs to support the early costs associated with starting a business. Series A refers to the first round of stock offered to investors during early-stage rounds. Typical Series A rounds fall in the range of $2-5M, with the offer options for 20-40% of the company, and are intended to support a company through the early stages of building a business, from product development to hiring to marketing. Series B refers to second-stage financing. Series C/C+ funding is used as companies grow, they might continue to seek additional funds to meet future milestones. Angel Investors and Venture Capitalists are either rich individuals or small company that provides small seed funding for the new start up industry. They may also buy the stock of the public company in exchange of convertible debt, for example Warren Buffet bought stocks in Goldman Sachs during the recession period. Equity funding or Crowdfunding is an umbrella term that refers to any means of financing your company in which you receive money in exchange for issuing shares of your stock.

The figure **??** shows an obvious trend of the Financial institutions investing more than the individuals. Individual Investors are more interested in providing funds to the software, mobile and web as they are interested in immediate gains as compared to financial institutions. It is easily noticeable that the distributions of total rounds of Series A/B/C+ or Venture funding is more in case of software than biotech, although the total investments in biotech industry is more than software. This is obvious because the number of years to develop a product takes a lot of time and money in case of biotech in-

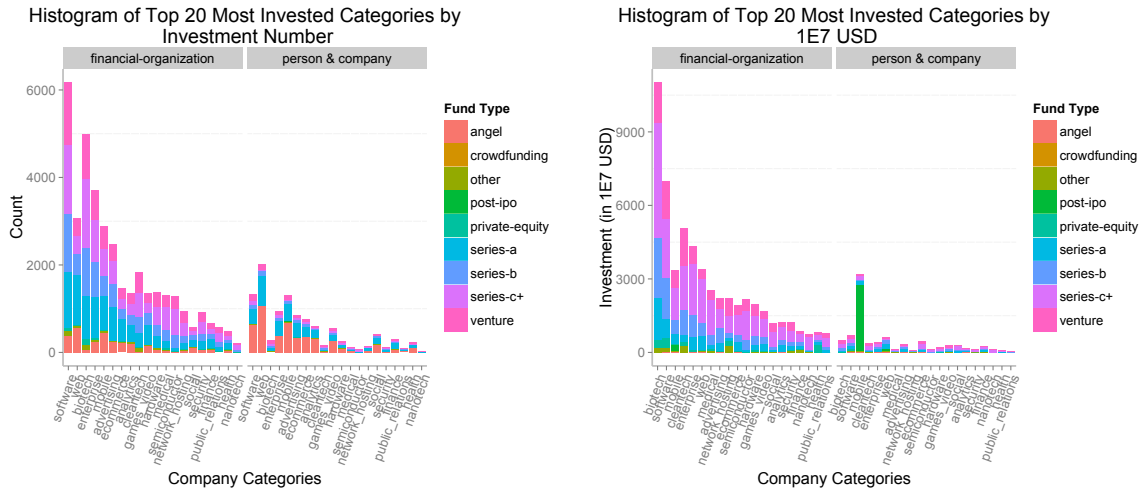

Figure 5: **Type, Rounds and Amount of Investment in various different types of Industry**

dustries compared to software products. This also suggests that – although the risk is high in biotech industry, the investors are interested in investing because the return of investment is very high as compared to software industry. For example: if a biotech company comes up with a drug that cures cancer or tuberculosis or a successful diagnostic product then percentage of profit is definitely higher than software industries.

7 Performance of Investors

We were curious about most important investors, and so we took a look at how successful different investors are at picking startups that will either be acquired or IPO or shut down. To do this, we looked at companies in the data set that were not currently operating and instead focused on those which had a definite end, be it a buyout, going public, or bankruptcy. Since we could not obtain data that could give us the insights on how much of a company was sold during a merger or acquisition or the amount of shares that a particular investor sold during the IPO, the link between fund outcomes and financial performance is tenuous. For example, in case of Google and Apple, an investor would have gained a lot if they would have kept its shares rather than selling them whereas, on the other hand in case of Zynga, a famous online video game company, the shares plummeted 4 times lower than the initial IPO stock price.

The figure **??** suggests the top investors are Sequoia Capital, SV Angel and New Enterprise Accessories as they have the highest success rate. It is interesting to note that for most companies the top investors choose to fund, they end up getting acquired, and have a rather small chance of achieving an initial public offering. This seems to reinforce the trend that many startups prefer leveraging their success into a buyout and avoid the uncertainty of going public. Furthermore, due to the rather unpredictable nature of venture capital, it is likely that investors prefer a "sure thing" like an acquisition to a company that IPO's. It's important to note that likely the "Closed" category is heavily underrepresented in the data set which hints towards the case of "survivorship bias" in the data set. Survivorship bias is the logical error of concentrating on the companies that "survived" and inadvertently overlooking those that did not leading to lack of visibility.Since, the companies and individuals don't like to put up their failed attempts on display it is easy to neglect the failures. Indeed, finding data on failed companies proved to be a monumentally difficult task when cleaning the data set. Another interesting

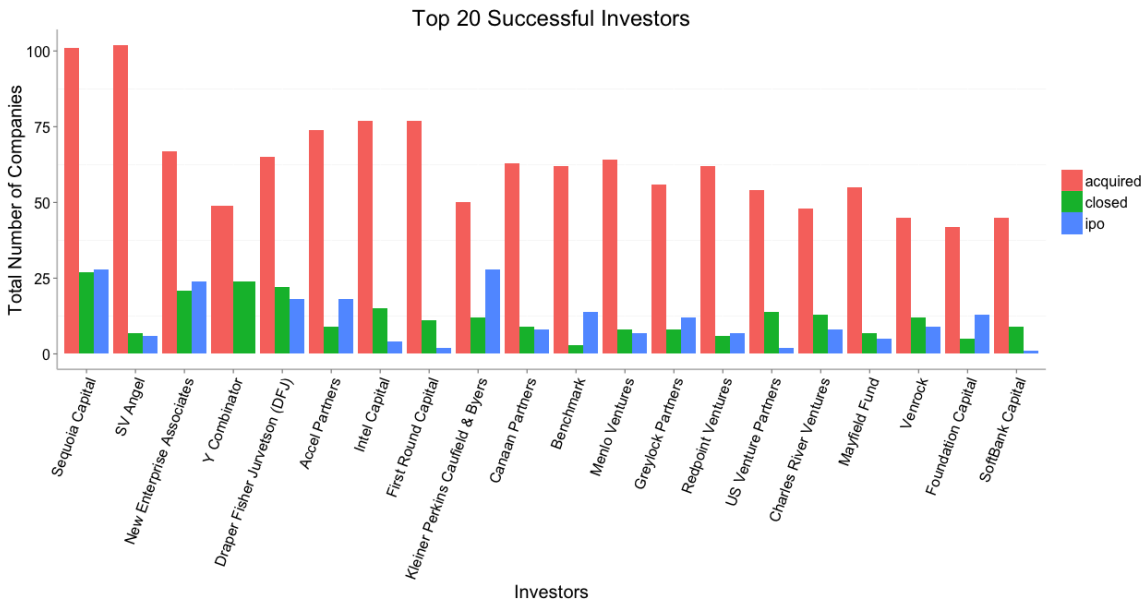

Figure 6: **The histogram shows the success of top 20 investors.**

thing to point out is KPCB's remarkably high IPO rate; indeed you were more likely to go public if KPCB invested than any other investor.

8 Predicting a Start-up's Success

We explored building a generalized linear model to see if we could predict whether a start-up will succeed (IPO/get acquired) or fail (close or shut down). We fit a regularized logistic regression model to the data that utilized predictive features such as Company category, geographical region, rounds of funding, total funding and the year in which the start up was founded. Using this model, we can try to make predictions about which companies in the dataset that are still operational are likely to succeed and which ones we think may shut down. Note that this may not be a fair measure of prediction because younger companies may not have enough time to build up their feature vectors. Therefore, we did analysis using the companies founded between 1997 and 2011. We have a mix of continuous numerical and categorical features. The company category, region, and the founding year were categorical variables. Categorical features are converted to numerical using the standard trick or expanding the features set by transforming a categorical feature with K possible values, into K features each with a binary value.

The most predictive features were total funding, category of the industry esp. Biotech, Cleantech, Software, Video Games and Hardware and region. The results in Table **??** predicts the following company has a great chance of being acquired or getting IPOed whereas, Table **??** suggests the chance of the company getting shutdown. The column Success or Failure suggests the probability of success or failure (1-Success) respectively. An interesting result was the prediction of Twitter's IPO which occurred on the 6th November (and the data set was published on 4th Nov. 2013). We were pleasantly surprised by the accuracy of our results, but the model is still not perfect. Another company it predicts as being highly likely to have an IPO is Solyndra, which famously went bankrupt amid scandal concerning accusations of fraud and misrepresenting corporate finances to obtain government funding. It is an important caution to not take the predictions made by the linear model by face value alone. Still,

Company Name	Type	Region	Found	Success
SurveyMonkey	software	SF Bay	1999	0.99
Sigmacare	health	New York	2005	0.99
ZeniMax	games_video	Washington DC	1999	0.98
Twitter	social	SF Bay	2006	0.97
Bloom Energy	cleantech	SF Bay	2002	0.97
Fisker Automotive	automotive	Los Angeles	2008	0.97
Pinterest	social	SF Bay	2009	0.96
Solyndra	manufacturing	SF Bay	2005	0.96
LivingSocial	ecommerce	Washington DC	2007	0.95
GreatPoint Energy	cleantech	Boston	2004	0.93

Table 3: **Predicting the Successful Companies**

Company Name	Type	Region	Found	Failure
JumpTheClub	mobile	Hartford	2010	0.9999999992
PureHistory	search	Somerset	2011	0.9999999990
Striped Sail	hardware	Champaign	2010	0.9999999987
Apps Genius	games_video	Red Bank	2009	0.9999999983
EnergyWeb Solutions	ecommerce	Allentown	2004	0.9999999980
GlucoSentient	biotech	Champaign	2011	0.9999999978
ANDalyze	other	Champaign	2005	0.9999999972
Forcura	biotech	Jacksonville	2010	0.9999999972
Southtree	web	Chattanooga	2009	0.9999999967
PrintEco	software	Champaign	2010	0.9999999966

Table 4: **Predicting the Shutdown Companies**

we believe that even this failure is positive, because our model is mostly predicated upon amount of funds raised; Solyndra had a management dispute that couldn't be predicted by any of the important features in the model. An argument could be made the the case of Solyndra was anomalous. Another thing to note is the model was trained upon a much smaller training set than its test predictions set, so a reasonable amount of over fitting may have taken place. Furthermore, because the model was trained to predict success, it does not do quite as good of a job predicting failures, since the features that predict success are not necessarily the same ones that predict failure. The model is likely unfit for any true prediction, but nevertheless performed admirably enough to make a case for which features are indeed the most predictive when looking at early investment in start-up companies.

9 Post-IPO

We wanted to look at the companies in our dataset that had an IPO to see if we could identify any notable trends. To do this, it first was necessary to discover the companies' ticker symbols. We built a function that called Yahoo! Finance's search function using the company name as an input. The function returned the first search result. However, due to the fact that the dataset has only information about companies in their startup phase, we found that the search function didn't operate properly for many companies that had gone through significant changes like mergers, post-IPO bankruptcies, buy-outs, name changes, and ticker symbol changes. Experimenting with other websites' search functions had the same issue. To overcome this, we fed the function a dummy value for those companies for which the search function failed to return the correct symbol and found the missing symbol manually. Because running the find_symbol function queries a website, it runs very slowly; to make our code run easier we saved the symbols in a file.

In the base data set, 377 companies are listed as having an IPO. After running the find_symbol func-

tion, we found symbols for 332 active companies. We chose to exclude companies that were bought out, had a merger, or went bankrupt because we wanted to focus on how the companies were performing currently and finding historical stock data for companies that are no longer in existence or underwent some significant change proved to be very difficult.

We used these symbols to query Yahoo! Finance using a function called getKeyStats_xpath, which scrapes a company profile page for any relevant valuation statistics. This function returned many, many interesting features like Market Cap, Revenue, EBITDA, etc. We restricted our analyses to the following features because we find them to be more reliable and had a minimal number of NA values: Market Cap, Enterprise Value, Enterprise Value/Revenue, Enterprise Value/EBITDA, Revenue, Revenue Per Share, Gross Profit, EBITDA , Total Cash , Total Cash Per Share, 52-Week Change , Shares Outstanding, Float, %Held by Insiders , PEG Ratio.

Company	Market Cap	Enterp. Value	Revenue	EBITDA*	Total Cash
Verizon	141	186	120	34	57
Google	354	297	57	18	55
Raytheon	27	29	24	3	4
Lockheed Martin	43	49	46	5	3
Texas Instruments	46	48	12	4	4
Honeywell	68	72	38	6	6
Xerox	14	20	22	3	1
News Corp	10	7	9	1	3
General Electric	270	652	145	29	10
Quintiles	5	7	4	1	1

Table 5: **The Top Ten IPO Companies**

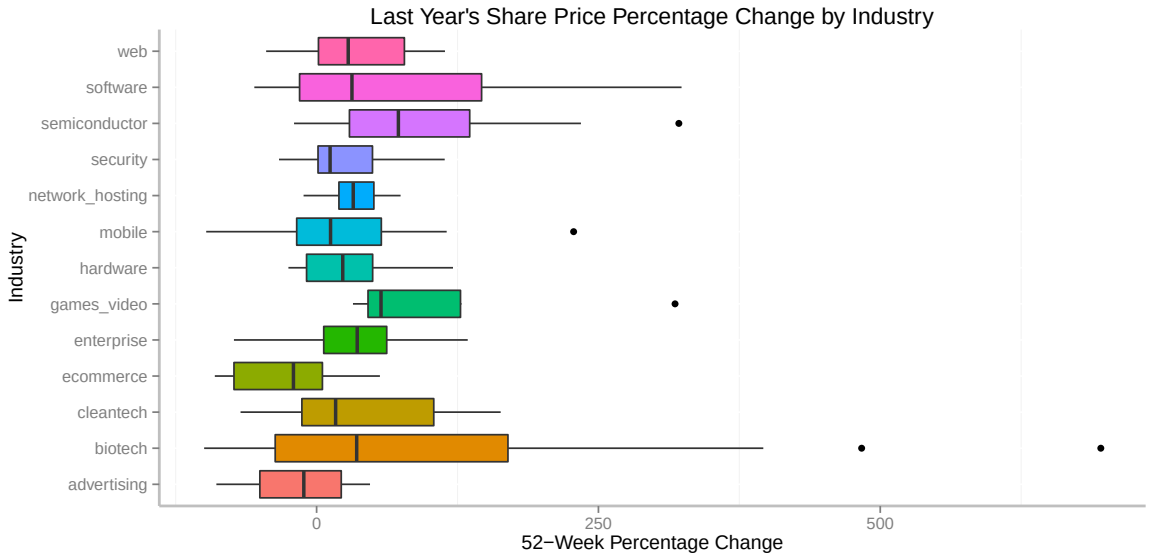

Figure 7: **The histogram shows the success of top 20 investors.**

Breaking down the percentage change within the last 52 weeks by industry, we can get a rough picture of which industries are booming, busting, predictable, or volatile. The figure **??** are the top industries identified earlier plus any industry with more than 15 companies represented in the data set. We can see that biotech, software, and cleantech have higher variances, making investing in those three industries high-risk, high-reward. Surprisingly, it appears that the safest industry for company growth currently

is video-games, with an average share price increase of over fifty percent. The worst performing industries in terms of growth were advertising and e-commerce. Note that this graphic must be interpreted with skepticism, because a number of factors can affect share price growth. For example, the youngest companies tend to have much higher absolute values of percentage growth because they're smaller, so industries with more young companies may have inflated figures. Furthermore, because this data excludes companies that failed post-IPO, had a merger, or were bought out, one cannot make inferences about these possibilities, which is likely what an investor looking at percentage growth may want to do.

Another interesting thing to look at was the comparison between Enterprise Value (EV) /Revenue (R) and Enterprise Value/EBITDA. These two ratios are popular for evaluating the value of the company. EV/R is a measure of the company's ability to generate funds, while EV/EBITDA is a more direct representation of the company's profits. We expected to see these numbers highly correlated, but found instead that they do not to a very high degree. A good example for how this can be is the medical industry; they have an average EV/R multiple of around 10, which seems high, however, the EV/EBITDA was around -2.7; this suggests that medical companies generate high revenue but do not enjoy very large profit margins. However, one must remember that there are many possible scenarios for any one ratio value, so not much weight can be given to these stats alone. Furthermore, it is generally considered meaningless to compare the investment multiples across industries, instead it is better to use them to compare similar companies. Finally, one must remember that the number of companies being compared per industry is very low, so a very large or a very small company could skew the industry average by a significant amount.

We also wanted to identify the "best" companies that are currently being publicly traded in our data set. To do this, we ranked every company in a number of key metrics, and then calculated an aggregate rank score from those rankings. We ranked the companies in the following metrics: Market Cap, Enterprise Value, Revenue, Revenue per share, Gross Profit, EBITDA, Total Cash, and PEG ratio.

This algorithm identified the top ten public companies in our data set as shown in Table **??**. Not surprisingly, nearly all of these company names are well known, and they are all very large and stable companies. These companies would be a good "safe bet" investment. However, investment in these companies is not guaranteed to return at very high levels, because the algorithm's calculations are based mostly on metrics that indicate a company's total size, not its growth.

10 Conclusions

Startups are essential for ensuring a competitive marketplace, and they are the front lines of innovation in both well-established industries and emerging new markets. The USA's ability to foster the growth of fledgling startups and support the smaller, exciting companies with promise is a significant strength of its economy. The explosion of the software industry and the subsequent related markets like web, mobile, social networking, etc. was made possible by the growing number of investors trying to get in on the next big company from an early phase and the ever-enlarging pool of investment dollars available for early startup funding.

A number of factors influence startup funding, and we found that startup funding fluctuates with the economy to a large degree. Furthermore, startups with a connection to San Francisco, New York, Los Angeles, Boston, or any other major investment center were much more likely to succeed. Indeed, we confirmed that most startup funding occurs around major metropolitan areas, as one might expect.

Certain industries have become much more prominent for startup investment over the years. For the last decade, software has been the undisputed giant, with more software companies being founded

yearly than any other industry. Biotech also has seen steady growth, and more recently, web, mobile apps, and enterprise new companies have become more common. Some industries sectors like Software are certainly more "safer" than others like Biotech, Cleantech and therefore, the chance of success can indeed be greatly dependent upon industry.

Our model was solely based to give an insight to an investor and its future investments. We found that the most predictive features of a startup's success are the amount of money received, region, industry, and year of founding. Using simply these features, one can predict with reasonable certainty which companies will go public or be acquired - one can assume investor confidence in those companies that receive more funding. However, no model is perfect, and predicting a company's future based solely off of external factors like funding received, location, etc. cannot be all-encompassing since it fails to take into account the strength of a companies internal qualities like employee ability, quality of product or service, etc.

Still, investors do a remarkable job of not wasting their money. The top 20 investors had very good success rates across the board, but it was interesting to note that most of them preferred an acquisition to an IPO. Very few companies manage to reach the IPO stage, and to get there one must reject many opportunities to cash in, which is not always in an investor's best interest. But, the low shut down rates definitely makes us skeptic about the survivorship bias.

Looking post-IPO, we tried to identify some factors that indicate a stable company, and look at which companies in our data set that began as startups managed to reach success. Big names like Verizon and Google dominate the list as one might expect, but perhaps a more interesting analysis was to look at industries and their growth; we found that the emerging industries with more startups were also those with much more unpredictable spread of a 52-week percent change in share price. Certain industries, like video gaming, appeared to offer more sure investment while other like advertisement seemed to only plummet post-IPO. This is interesting when analyzed in conjunction with the remarkably high rate of acquisition by advertising companies - perhaps advertising startups know that success in the market for their industry is less likely, so are more willing to cash out.

We believe we have painted an accurate and interesting picture of the startup landscape in the USA for the past 20 years. Startups drive the economy, and there seems to be an explosion in their prevalence after every new market emerges. The future holds promise and uncertainty for startups in the USA, but by analyzing their funding, we can make smart educated guesses on their future success and look to capitalize on smart investments.

11 Bibliography

a. Crunchbase Database: http://info.crunchbase.com/about/crunchbase-data-exports/

b. Yahoo! Finance: http://finance.yahoo.com/

c. Google Finance: http://www.google.com/finance

d. R Core Team (2013). R langauge : http://www.R-project.org/

e. Wikipedia: for the informations about the companies and its status.

12 Appendix

```
##############################################################################
# @Authors: Aayush Raman, Yansin Miao and Gabe Breternitz
# Date: 1st Dec. 2013
#
# Program is used for:
# Stat 405 Final Project-2 on Start-Up Companies
##############################################################################

##set the working directory and load data
setwd("../project2/complete_code_data")

## Remove all the data in the current workspace
rm(list = ls())

## Libraries
library(plyr)
library(ggplot2)
library(xtable)
library(gridExtra)
library(maps)
library(mapdata)
library(mapproj)
library(maptools)
library(ggmap)
library(sp)
library(stringr)
library(ggsubplot)
library(reshape2)
library(taRifx)
library(XML)
library(quantmod)
library(PerformanceAnalytics)
library(reshape2)

## Reading the  csv files from crunch base
investments = read.csv("investments.csv",
                       header=TRUE, stringsAsFactors = F)
companies = read.csv("companies.csv", header=TRUE,
                     stringsAsFactors = F)
ind.na = which(companies[, "category_code"] == "")
companies$category_code[ind.na] = "Others"
rounds = read.csv("rounds.csv", header=TRUE, stringsAsFactors = F)

##############################################################################
#                              Plot 1 line
#              Line Plot: Investment Versus Time
##############################################################################
#ind interest things about investor
#remove duplicate results in investments
```

```
investments <- unique(investments)

#split the investor_permalink and add a new variable called
#investor_type (company,financial-organization, person)
investments$investor_type =
  sapply(strsplit(investments$investor_permalink, "/"), "[", 2)
unique(investments$investor_type)
invest.cate <- subset(investments,investor_category_code!="" &
                        company_name != "Army & Air Force Exchange Service")
#summary by investor_type
invest.status<-ddply(investments,.(investor_type),
                     summarize,sum_funds=sum(raised_amount_usd,na.rm=T),
                     num_invest=length(investor_type))
invest.status <- arrange(invest.status,desc(sum_funds))
xtable(invest.status) #1 table

#summary by funded_year
invest.status<-ddply(investments,.(funded_year),
                     summarize,
                     sum_funds=sum(raised_amount_usd,na.rm=T),
                     Startup.num=length(funded_year))

invest.status$sum_funds = invest.status$sum_funds / 1E7
tem = read.csv("label.csv", header=TRUE, stringsAsFactors = F)

#a plot of total investment (funds and number) of by year
ggplot(invest.status,aes(x = funded_year,
                         y = sum_funds,
                         color = Startup.num, 1))+
  xlim(1987,2015)+
  geom_line(size = 2) +
  geom_point() +
  geom_vline(aes(xintercept = 1998), linetype = "longdash",color="red")+
  geom_vline(aes(xintercept = 2000), linetype = "longdash",color="red")+
  geom_vline(aes(xintercept = 2004), linetype = "longdash",color="red")+
  geom_vline(aes(xintercept = 2007), linetype = "longdash",color="red")+
  geom_vline(aes(xintercept = 2008), linetype = "longdash",color="red")+
  geom_vline(aes(xintercept = 2011), linetype = "longdash",color="red")+
  annotate("text", x=1998, y=800, label="Year:1998\nGoogle", size = 5,
           family="Times", colour="black") +
  annotate("text", x=2000, y=1800, label=
             "Year:2000\nInternet Bubble", size = 5,
           family="Times", colour="black") +
  annotate("text", x=2004, y=4000, label="Year:2004\nFacebook", size = 5,
           family="Times", colour="black") +
  annotate("text", x=2007, y=4000, label="Year:2007\n1st Iphone", size = 5,
           family="Times", colour="black") +
  annotate("text", x=2008, y=9500,
           label="Year:2008\nFinancial crisis\n1st Android Phone",
           size = 5,family="Times", colour="black") +
  annotate("text", x=2011, y=11500, label="Year:2011\nEconomic Recovery", size = 5,
```

```
          family="Times", colour="black") +
  theme(panel.background = element_blank(), text = element_text(size=15),
        axis.line = element_line(size = 0.8, colour = "grey75"))+
  scale_color_gradient(limits=c(0, 10000),high = "red", low = "blue",
                       name = "Startup Number")+
  labs(x="Years from 1987 to 2013",y="investment (in 1E7 USD)")+
  ggtitle("Trend of Investment versus the Years\nfrom 1987 to 2013")

##############################################################################
#                                 Table 1
#                          Status of the companies
##############################################################################
# status totalstatus
# 1 OPERATING       15078
# 2  ACQUIRED        1879
# 3    CLOSED        1170
# 4       IPO         377
## Status of the companies
comp.status = ddply(companies, .(status), summarize, totalstatus = length(status))
comp.status$status = toupper(comp.status[,1])
comp.status = comp.status[order(comp.status[,2] , decreasing = TRUE),]
row.names(comp.status) = NULL

# print table
print(xtable(comp.status, digits=0,
             caption="\\textbf{Status of Companies}", label="ByStatus_Table"),
      size="footnotesize", #Change size; useful for bigger tables
      include.rownames=FALSE, #Don't print rownames
      include.colnames=FALSE, #We create them ourselves
      hline.after=NULL, #We don't need hline; we use booktabs
      add.to.row = list(pos = list(-1,
                   nrow(comp.status)),
      command = c(paste("\\toprule \n",
                        "Status of the Company & Number of Companies  \\\\\n",
                        "\\midrule \n"),
                  "\\bottomrule \n")
      )
)

##############################################################################
#                                 Table 2
#              Top Companies types and its status -- overall
##############################################################################
# category_code NumofOper NumofIpo NumofAcq NumofClosed
# 1      software      2275       37      358         154
# 2       biotech      1836      119      119          80
# 3           web      1139       16      289         262
# 4        mobile       945       15      145          73
# 5    enterprise       933       13      147          41
# 6     ecommerce       609        9       47          56
# 7   advertising       581       12      107          44
```

```
# 8       cleantech        544        22        44          41
# 9     games_video        523         5        94          86
# 10       hardware        518        16        44          40
## Top Companies types and its status -- overall
comp.status = ddply(companies, .(category_code), summarize,
                    NumofOper = length(which(status == "operating")),
                    NumofIpo = length(which(status == "ipo")),
                    NumofAcq = length(which(status == "acquired")),
                    NumofClosed = length(which(status == "closed")))
comp.status = comp.status[order(comp.status[,2] , decreasing = TRUE),]
row.names(comp.status) = NULL

# printing table
print(xtable(comp.status[1:10, ], digits= 2,
             caption="\\textbf{Status for  various categories of companies}",
             label="ByCompType_Table"),
      size="footnotesize", #Change size; useful for bigger tables
      include.rownames=FALSE, #Don't print rownames
      include.colnames=FALSE, #We create them ourselves
      hline.after=NULL, #We don't need hline; we use booktabs
      add.to.row = list(pos = list(-1,
               nrow(comp.status[1:10,])),
      command = c(paste("\\toprule \n",
                        "Company Type & Operating & IPO & Acquired & Closed\\\\\n",
                        "\\midrule \n"),
                  "\\bottomrule \n")
      )
)

#############################################################################
#                              Plot 2 Heatmap 1
#             Heatmap: Hottest region and types of companies
#############################################################################
## Hottest region and types of companies
ind.geo = which(companies$category_code %in%
                  comp.status$category_code[1:10] ==  TRUE)
comp.region = companies[ind.geo, ]
comp.status = ddply(comp.region, .(region), summarize, Total = length(region))
comp.status = comp.status[order(comp.status[,2] , decreasing = TRUE),]
row.names(comp.status) = NULL
ind.geo = which(comp.region$region %in% comp.status$region[1:10] == TRUE)
comp.status = ddply(comp.region[ind.geo, ],
                    .(region, category_code), summarize,
                    Num.of.Comp = length(which(status == "operating")) +
                      length(which(status == "ipo")) +
                      length(which(status == "acquired")))

ggplot(comp.status, aes(y = category_code, x = region, fill = Num.of.Comp)) +
  geom_tile(colour = "white") +
  scale_fill_gradient(low = "steelblue", high = "red", name = "# Comapnies") +
  ylab("Various Types of Companies") +
```

```
  xlab("Regions") +
  ggtitle(label = "Number of Companies based on location") +
  theme(text = element_text(size=20),
        axis.text.x = element_text(vjust=1, color = "black"),
        axis.text.y = element_text(color = "black"))

###############################################################################
#                              Plot 3 Heatmap 2
#          Heatmap: Types of companies in terms of years
###############################################################################
## In terms of years
ind.geo = which(comp.region$founded_year > 1986)
comp.status = comp.region[ind.geo, ]
comp.status = ddply(comp.region[ind.geo, ],
                    .(founded_year, category_code),
                    summarize, Num.of.Comp = length(founded_year))
ggplot(comp.status,
       aes(y = category_code, x = founded_year, fill = Num.of.Comp)) +
  geom_tile(colour = "white") +
  scale_fill_gradient(low = "steelblue", high = "red", name = "#'s Comapnies") +
  xlab("Years") + ylab("Various Types of Companies ") +
  ggtitle(label = "Number of Companies based on Years") +
  theme(text = element_text(size=20),
        axis.text.x = element_text(vjust=1, color= "black"),
        axis.text.y = element_text(color = "black"))

###############################################################################
#                           Plot 4  Histogram 1
# Left: Number of funds categories receive descend by number of investment
# Right: Sum of funds categories receive descend by sum of investment
# Two Histograms are both color by funds type and faceted by investor type
###############################################################################
#the top investors' categories which receive most funds
#Prepare the data
investor.category <- ddply(investments,.(investor_name,investor_region,
                                         investor_country_code,
                                         investor_type), summarize,
                           sum_funds = sum(raised_amount_usd, na.rm = T),
                           num_invest = length(investor_type))
investor.category <- arrange(investor.category, desc(sum_funds))

#the top 20 investors' categories
investor.top20 <- head(investor.category, 20)
#the top invested companies' category
invested.category <- ddply(investments,.(company_category_code),
                           summarize,
                           sum_funds = sum(raised_amount_usd,na.rm=T),
                           num_invest = length(company_category_code))
invested.category <- arrange(invested.category, desc(sum_funds))
#the top 20 most invested companies' category
top20.index <- head(invested.category,20)$company_category_code
```

```
#subset the investment data with top 20
invested.top20<-subset(investments,company_category_code %in% top20.index)

invested.top20[invested.top20$investor_type
               == "person",]$investor_type = "person & company"
invested.top20[invested.top20$investor_type
               == "company",]$investor_type = "person & company"

temp1 <- ddply(invested.top20,.(company_category_code,
                                funding_round_type,investor_type),
               summarise,num_funds = length(company_category_code))

#Histogram Plot
p1<-ggplot(temp1, aes(reorder(company_category_code, -num_funds),
                      num_funds, fill=funding_round_type))+
  geom_histogram() +
  #scale_fill_brewer(palette = "Spectral") +
  theme(panel.background = element_blank(),
        axis.line = element_line(size = 0.8, colour = "grey75"),
        text = element_text(size=15),
        axis.text.x = element_text(angle = 70, hjust =1))+
  labs(x="Company Categories",y="Count") +
  scale_fill_discrete(name = "Fund Type") +
  facet_wrap(~investor_type)

temp <- ddply(invested.top20,.(company_category_code,
                               funding_round_type,investor_type),
              summarise,sum_funds=sum(raised_amount_usd,na.rm=T))
#combine investor_type person and compant into person+company as their
#investment are relatedly small compared to those of financial organizion
#temp[temp$investor_type == "person",]$investor_type = "person & company"
#temp[temp$investor_type == "company",]$investor_type = "person & company"

temp$sum_funds = temp$sum_funds / 1E7
#Plot the investment (in USD) of the top 20
#invested companies' categories facet by investor_type

p2<-ggplot(temp,aes(reorder(company_category_code, -sum_funds), sum_funds))+
  geom_bar(aes(fill=funding_round_type), stat="identity")+
  theme(panel.background = element_blank(), text = element_text(size=15),
        axis.line = element_line(size = 0.8, colour = "grey75"),
        axis.text.x = element_text(angle = 70, hjust =1)) +
  labs(x="Company Categories",y="Investment (in 1E7 USD)") +
  scale_fill_discrete(name = "Fund Type") +
  facet_wrap(~investor_type)
pushViewport(viewport(layout = grid.layout(1, 2)))
print(p1 + ggtitle("Histogram of Top 20 Most Invested Categories by
                   \n Investment Number"),
      vp = viewport(layout.pos.row = 1, layout.pos.col = 1))
print(p2 + ggtitle("Histogram of Top 20 Most Invested Categories by
```

```
                    \n 1E7 USD"),
      vp = viewport(layout.pos.row = 1, layout.pos.col = 2))

###############################################################################
#                          Plot 5  Histogram 2
#   Number of rounds companies receive Versus Companies' names
#   Companies who receive top 20 rounds(Left)
#   Companies who receive top 20 funds(Right)
#   Two Histograms are both decreased by num of rounds
###############################################################################
##Prepare the data
#romove the multiple instances in rounds
rounds <- unique(rounds)
#remove Army & Air Force Exchange Service
rounds <- subset(rounds, company_name != "Army & Air Force Exchange Service")

#the top investors' categories which receive most funds
rounds.status <- ddply(rounds,.(company_name), summarize,
                       sum_funds = sum(raised_amount_usd, na.rm = T),
                       num_invest = length(company_name))
top20.index1 <- head(arrange(rounds.status,desc(num_invest)),20)$company_name
rounds.top20<-subset(rounds, company_name %in% top20.index1)
rounds.temp1 <- ddply(rounds.top20,.(company_name, funding_round_type), summarize,
                      num_invest = length(company_name))

top20.index2 <- head(arrange(rounds.status,desc(sum_funds)),20)$company_name
rounds2.top20<-subset(rounds,company_name %in% top20.index2)
rounds.temp2 <- ddply(rounds2.top20,.(company_name, funding_round_type), summarize,
                      num_invest = length(company_name),
                      sum_invest = sum(raised_amount_usd, na.rm = T))

#Histogram Plot
p1 <- ggplot(rounds.temp1, aes(reorder(company_name, -num_invest),num_invest))+
  geom_bar(aes(fill = funding_round_type), stat = "identity") +
  scale_fill_discrete(name = "Fund Type") +
  theme(panel.background = element_blank(),
        axis.line = element_line(size = 0.8, colour = "grey75"),
        text = element_text(size=15),
        axis.text.x = element_text(angle = 70, hjust =1))+
  labs(x="Companies",y="Rounds")

p2 <- ggplot(rounds.temp2, aes(reorder(company_name, -num_invest), num_invest)) +
  geom_bar(aes(fill = funding_round_type), stat = "identity") +
  scale_fill_discrete(name = "Fund Type") +
  theme(panel.background = element_blank(),
        axis.line = element_line(size = 0.8, colour = "grey75"),
        text = element_text(size=15),
        axis.text.x = element_text(angle = 70, hjust =1))+
  labs(x="Companies",y="Rounds")

pushViewport(viewport(layout = grid.layout(1, 2)))
```

```
print(p1 + ggtitle("Histogram of Top 20 invested companies
                    \nwhich receive most rounds"),
      vp = viewport(layout.pos.row = 1, layout.pos.col = 1))
print(p2 + ggtitle("Histogram of Top 20 invested companies
                    \nwhich receive largest funds"),
      vp = viewport(layout.pos.row = 1, layout.pos.col = 2))


#############################################################################
#                          Data Preparation
#              Geom Plot Prepare map data and geom labels
#############################################################################
data(us.cities)
us.cities$name = str_replace_all(us.cities$name, " (.*)", "")

#merge the companies data with the geom position in us.cities data
companies1 <- merge(companies, us.cities, by.x="city", by.y="name")
companies1$funding_total_usd <- as.numeric(companies1$funding_total_usd,na.rm=T)
companies1[companies1$state_code == "",]$state_code = "SD"
#remove Army & Air Force Exchange Service since it is supported by government
companies1 <- subset(companies1, name != "Army & Air Force Exchange Service")

state.fund <- ddply(companies1,.(state_code),
                    summarize,
                    num_fund=sum(funding_total_usd,na.rm=T),
                    num_length=length(city))
state.fund<-arrange(state.fund,desc(num_fund))

#prepare the maping data
map <- map_data("state")
state = read.csv("state.csv", header=TRUE, stringsAsFactors = F)
center <- read.csv("center.csv", header=TRUE, stringsAsFactors = F)

# find center of every state via geocode and save the data
# state$lon = geocode(state$state_code)$lon
# state$lat = geocode(state$state_code)$lat
# write.csv(state, file = "center.csv")

state$region <- tolower(state$region)
state.fund1 <- merge(state.fund, state, by = "state_code")
df <- join(state.fund1, map, type = "full", by = "region")
df$num_fund <- df$num_fund / 1E7

#prepare the map
q <- ggplot(aes(long, lat), data = map_data('state')) +
  geom_polygon(aes(group = group), color = "black", fill = I('grey100'))
q <- q + coord_map("bonne",40)
q <- q +  theme(panel.grid = element_blank(),
                panel.background = element_rect(fill = "white"),
                axis.text = element_blank(), axis.ticks = element_blank())
q <- q + labs(x="",y = "")
```

```
#############################################################################
#                        Plot 6  Geomploy 1 (A heat map)
#            The color shows investment collected by states
#############################################################################
#map the state.fund2 onto the map
df$bin <-cut(df$num_fund,
             breaks = c(0,10,50,100,500,1000,2000,3000), include.lowest=T)

q + geom_polygon(data = df, aes(group = group, fill = df$bin)) +
  scale_fill_brewer("Funds (1E7 USD)\n Received by State", palette = "Reds")+
  theme(text = element_text(size = 20))+
  ggtitle("Funds received by states in 1E7 USD") +
  geom_text(aes(lon,lat,label = state_code), data = center) +
  coord_map('bonne', 31)

#############################################################################
#                    Plot 7 Geomploy 2 (A heatscatter map)
#            The color shows investment collected by states
#      The scatter points indicate companies categories in that city
#############################################################################
geo.comp <- ddply(companies1,.(city, state_code, status, lat, long, pop),
                  summarize,
                  num_fund=sum(funding_total_usd,na.rm=T),
                  num_length=length(city))
geo.comp1 <- subset(geo.comp,
       geo.comp$state_code != "AK" & geo.comp$state_code != "HI")

q + geom_polygon(data = df, aes(group = group, fill = df$bin)) +
  scale_fill_brewer("Funds (1E7 USD)\n Received by State",
                    palette = "Reds")+
  theme(text = element_text(size = 24))+
  ggtitle("Funds received by states in 1E7 USD
          and the most invested cities") +
  geom_jitter(data = geo.comp1,
              position = position_jitter(width = 0.8, height = 0.8),
              aes(x = long, y = lat, color = status)) +
  scale_area(range=c(2,6)) +
  geom_text(aes(lon,lat,label = state_code),
            data = center, color="white", size = 5)

#############################################################################
#                            Plot 8  Histogram 3
#                          Investors Performances
#############################################################################
## Investors Performances
colnames(companies)[2] = "company_name"
comp.invest = merge(companies[,c("company_name", "category_code",
                                 "status", "funding_rounds")],
                    investments[,c("company_name", "investor_name",
                                   "company_category_code",
```

```
                                        "funding_round_type")],
                      by="company_name", all=TRUE)
comp.status = ddply(comp.invest, .(investor_name, status), summarize,
                      Total = length(status))
comp.status = arrange(comp.status, desc(Total))
ind.na = which(is.na(comp.status$investor_name) == TRUE)
comp.status = comp.status[-ind.na, ]
top.investors = comp.status[1:20,1]
ind.invest = which(comp.status$investor_name %in% top.investors == TRUE)
comp.status = comp.status[ind.invest, ]

## Plot of VS Progress
ggplot(comp.status, aes(x = reorder(investor_name, -Total), y = Total))+
  geom_bar(aes(fill = status, order = desc(status)), stat = "identity") +
  theme(panel.background = element_blank(),
        legend.title=element_blank(),
        axis.text.x = element_text(angle = 70, hjust =1)) +
  labs(x="Investors",y="Total Number of Companies") +
  ggtitle("Top 20 investor") +
  theme(text = element_text(size=20),
        axis.text.x = element_text(vjust=1, color= "black"),
        axis.text.y = element_text(color = "black"))

###############################################################################
#                          Plot 9  Histogram 4
#                  Based on IPO, Closed and Acquired
###############################################################################
# Based on IPO, Closed and Acquired
ind.oper = which(comp.invest$status == "operating")
comp.status = comp.invest[-ind.oper, ]
comp.status = ddply(comp.status,
                    .(investor_name, status), summarize,
                    Total = length(which(status == "acquired"))
                    + length(which(status == "ipo")) +
                      length(which(status == "closed")))
comp.status = arrange(comp.status,desc(Total))
ind.na = which(is.na(comp.status$investor_name) == TRUE)
comp.status = comp.status[-ind.na, ]
top.investors = comp.status[1:20,1]
ind.invest = which(comp.status$investor_name %in% top.investors == TRUE)
comp.status.1 = comp.status[ind.invest, ]

#Plot
ggplot(comp.status.1, aes(x = reorder(investor_name, -Total), y = Total)) +
  geom_bar(aes(fill = status, order = desc(status)),
           stat = "identity", position = "dodge") +
  theme(panel.background = element_blank(),
        legend.title=element_blank(),
        axis.text.x = element_text(angle = 70, hjust =1)) +
  labs(x="Investors",y="Total Number of Companies") +
  ggtitle("Top 20 Successful Investors") +
```

```
  theme(text = element_text(size=20),
        axis.text.x = element_text(vjust=1, color= "black"),
        axis.text.y = element_text(color = "black"))

#############################################################################
#                    Model1  Logistics Regression (GLM)
#############################################################################
## factoring the variables for GLM
companies$category_code = as.factor(companies$category_code)
companies$region = as.factor(companies$region)
companies$funding_rounds = as.numeric(companies$funding_rounds)
companies$funding_total_usd = destring(companies$funding_total_usd)
ind.year = which(is.na(companies$funding_total_usd) == TRUE)
companies = companies[-ind.year, ]
ind.year = which(is.na(companies$founded_year) == TRUE | companies$founded_year == "")
companies = companies[-ind.year, ]
ind.year = which(companies$founded_year < 2012 & companies$founded_year > 1997)
companies = companies[ind.year, ]
companies$founded_year = as.factor(companies$founded_year)

## L-1 Logistic Regression Model for prediction
# whether start-up company will work or not
# Variables -- Region (Col# 4) + Category Type (Col# 5)
# + Total funding(#9) + Number of funding rounds(#11)
# Year Founded(#15)
train.data = companies[-which(companies$status == "operating"),
                       c(2,4,5,9,11,15,6)]
test.data = companies[which(companies$status == "operating"),
                      c(2,4,5,9,11,15,6)]
train.data$rank[which(train.data$status ==
                        "acquired"|train.data$status == "ipo")] = 1
train.data$rank[which(train.data$status == "closed")] = 0

## GLM
model.logit = glm(rank ~ category_code + region +
                    funding_rounds + log(funding_total_usd) + founded_year,
                  data = train.data, family = "binomial")
ind = which(test.data$category_code %in% train.data$category_code == TRUE)
test.data = test.data[ind, ]
ind = which(test.data$region %in% train.data$region == TRUE)
test.data = test.data[ind, ]
test.data$rankpred = predict(model.logit, newdata = test.data, type = "response")
test.data = test.data[order(test.data$funding_total_usd, decreasing = TRUE),]
test.data.1 = test.data[1:10, c(1,2,4,6,8)]
test.data.1 = test.data.1[order(test.data.1$rankpred, decreasing = TRUE),]

#############################################################################
#                  Table 3-4  Logistics Regression Result
#############################################################################
print(xtable(test.data.1, digits= 2, caption="\\textbf
            {Predicting the Successful Companies}",
```

```
            label="ByCompType_Table"),
      size="footnotesize", #Change size; useful for bigger tables
      include.rownames=FALSE, #Don't print rownames
      include.colnames=FALSE, #We create them ourselves
      hline.after=NULL, #We don't need hline; we use booktabs
      add.to.row = list(pos = list(-1,
                                   nrow(test.data.1)),
                        command = c(paste("\\toprule \n",
                                          "Company Name & Type
                                          & Region & Found & Prob.\\\\\n",
                                          "\\midrule \n"),
                                    "\\bottomrule \n")
      )
)
test.data.1 = test.data[order(test.data$rankpred),]
test.data.1 = test.data.1[1:10, c(1,2,4,6,8)]
test.data.1$rankpred = 1 - test.data.1$rankpred
print(xtable(test.data.1, digits= 10, caption="\\textbf
            {Predicting the Shutdown Companies}",
            label="ByCompType_Table"),
      size="footnotesize", #Change size; useful for bigger tables
      include.rownames=FALSE, #Don't print rownames
      include.colnames=FALSE, #We create them ourselves
      hline.after=NULL, #We don't need hline; we use booktabs
      add.to.row = list(pos = list(-1,
                                   nrow(test.data.1)),
                        command = c(paste("\\toprule \n",
                                          "Company Name & Type
                                          & Region & Found & 1 - Prob. of Success\\\\\n"
                                          "\\midrule \n"),
                                    "\\bottomrule \n")
      )
)

##############################################################################
#                    Query the Financial Data from Yahoo
##############################################################################
companies <- read.csv("companies.csv",stringsAsFactors = F)

#check if there are redundancies
length(unique(companies$name))
length(companies$name)
#they are the same, so data isn't redundant

#identify companies that had an IPO
ipo <- companies[which(companies$status == 'ipo'),]
ipo.bu <- companies[which(companies$status == 'ipo'),]

length(ipo$name) #377 companies went public

##############################################################################
```

```
# The following code queries Yahoo Finance for symbols, so sometimes
# it can take a very long time to run. I include it as a reference, but
# we have chosen not to have it run in this final version for ease of use.
# Instead, the data we collected using it is read from a .csv file.
##############################################################################
# #Function that finds symbols from company names
# find_symbol = function(company){
#
#   # construct url
#   root1 = 'http://finance.yahoo.com/lookup?s='
#   u1 = paste0(root1, company)
#
#   # encode url
#   url  = URLencode(u1)
#
#   # extract data from the right table
#   data=matrix(1,1)
#   data = readHTMLTable(url)[[2]]
#   symbol = as.character(data[1,1])
#   symbol
# }
#
# #update companies that changed name or have naming issues
# #used FB as a dummy to run the loop and then fixed the names from a backup later
# ipo$name[3] <- "AONEQ" #bankrupt
# ipo$name[6] <- 'FB' # 'ACCI' #not yet being traded, very new IPO
# ipo$name[26] <- 'FB' #'AMRS'
# ipo$name[27] <- 'FB' #'ANAC'
# ipo$name[29] <- 'FB' #'ACOM' #Doesnt return right stock
# ipo$name[30] <- 'FB' #'ANGI' #returns wrong stock
# ipo$name[50] <- 'FB' #'BCONQ'
# ipo$name[51] <- 'FB' #dummy stock, real stock is 'BYOC'
# ipo$name[56] <- 'FB' #dummy stock, real stock is 'BMODQ'
# ipo$name[57] <- 'FB' #'ANIP' #BioSante merged with ANIP, throw it out
# ipo$name[59] <- 'FB' #dummy, real is 'BFLY'
# ipo$name[62] <- 'FB' # #cancelled IPO
# ipo$name[65] <- 'FB' #'PRSS'
# ipo$name[67] <- 'FB' #'CALD'
# ipo$name[70] <- 'FB' #'CPRX'
# ipo$name[75] <- 'FB' #'CNC'
# ipo$name[93] <- 'FB' #dummy, real is 'CNVO'
# ipo$name[104] <- 'FB' # #acquired by Thoma Bravo for $1.1 billion
# ipo$name[116] <- 'FB' # #acquired by RR Donnely for $70 million
# ipo$name[123] <- 'FB' # #bankrupt september 2011
# ipo$name[137] <- 'FB' #'FMS'
# ipo$name[143] <- 'FB' #'GTIV'
# ipo$name[145] <- 'FB' #dumy, real is 'GCMI'
# ipo$name[149] <- 'FB' # #filed chapter 11 in August 2013
# ipo$name[151] <- 'FB' # #postponed IPO
# ipo$name[152] <- 'FB' # #Can't find - must have folded
# ipo$name[159] <- 'FB' #'GTAT'
```

```
# ipo$name[161] <- 'FB' #dummy, real is 'HS'
# ipo$name[173] <- 'FB' # #Not actually an IPO
# ipo$name[181] <- 'FB' # 'ALR'
# ipo$name[200] <- 'FB' # 'LOCM'
# ipo$name[204] <- 'FB' #dummy, real is 'LOOP'
# ipo$name[213] <- 'FB' # 'SUNE' #namechange
# ipo$name[219] <- 'FB' # 'BOTA' #merger
# ipo$name[221] <- 'FB' # 'NSTG'
# ipo$name[231] <- 'FB' # 'NGSX'
# ipo$name[233] <- 'FB' # 'NWSA'
# ipo$name[238] <- 'FB' # 'OLBK'
# ipo$name[257] <- 'FB' # 'PBPB'
# ipo$name[263] <- 'FB' # 'QLIK'
# ipo$name[265] <- 'FB' # 'MEET' #name change to MeetMe
# ipo$name[271] <- 'FB' # 'RZ' #going bankrupt
# ipo$name[275] <- 'FB' # 'RGLS'
# ipo$name[293] <- 'FB' # #Throw this one out - multiple mergers and name changes
# ipo$name[300] <- 'FB' # 'SRLS'
# ipo$name[303] <- 'FB' # 'QPOND'
# ipo$name[322] <- 'FB' # 'PULS' #name change to pulse electronics
# ipo$name[348] <- 'FB' # 'VRAZ'
# ipo$name[358] <- 'FB' # 'VRNG'
# ipo$name[363] <- 'FB' # 'WBSN'
# ipo$name[365] <- 'FB' # 'XRSC'
# ipo$name[10] <- 'FB' # 'ACTV'
# ipo$name[22] <- 'FB' # bought out - 'FTER'

# symbol <- vector()
# #find symbols
# for(i in 3:377){
#   symbol[i] <- find_symbol(ipo$name[i])
# }
# symbol.bu <- symbol

#correct symbol errors
# symbol[3] <- 'exclude' # symbol "AONEQ" , bankrupt
# symbol[6] <-   'exclude' # ACCI' #not yet being traded, very new IPO
# symbol[23] <- 'bought' #bought out
# symbol[26] <-  'AMRS'
# symbol[27] <-  'ANAC'
# symbol[29] <-  'bought' # 'ACOM' bought out
# symbol[30] <-  'ANGI'
# symbol[50] <-  'exclude' # symbol 'BCONQ' , bankrupt
# symbol[51] <-  'BYOC'
# symbol[56] <-  'exclude' # symbol 'BMODQ' , bankrupt
# symbol[57] <-  'merger' #BioSante merged with ANIP, throw it out
# symbol[59] <-  'bought' # 'BFLY' bought out
# symbol[62] <-  'exclude' #cancelled IPO
# symbol[65] <-  'PRSS'
# symbol[67] <-  'CALD'
# symbol[70] <-  'CPRX'
```

```
# symbol[72] <- 'exclude' #bankrupt
# symbol[75] <-  'CNC'
# symbol[93] <-  'bought' #CNVO'
# symbol[104] <-   'bought' #acquired by Thoma Bravo for $1.1 billion
# symbol[116] <-   'bought' #acquired by RR Donnely for $70 million
# symbol[122] <- 'exclude' #bankrupt
# symbol[123] <-   'exclude' #bankrupt september 2011
# symbol[137] <-  'FMS'
# symbol[143] <-  'GTIV'
# symbol[145] <-  'GCMI'
# symbol[149] <-  'exclude' #filed chapter 11 in August 2013
# symbol[151] <-  'exclude' #postponed IPO
# symbol[152] <-   'exclude' #Can't find - must have folded
# symbol[159] <-  'GTAT'
# symbol[161] <-  'bought' # 'HS' bought by Cigna for $3.8 billion
# symbol[173] <-   'exclude' #Not actually an IPO
# symbol[181] <-   'ALR'
# symbol[200] <-   'LOCM'
# symbol[204] <-  'bought' # 'LOOP' bought for $860 million
# symbol[206] <- 'bought' #taken over
# symbol[213] <- 'SUNE' #name change
# symbol[219] <- 'merger' #merger
# symbol[221] <- 'NSTG'
# symbol[231] <- 'NGSX'
# symbol[233] <- 'NWSA'
# symbol[238] <- 'OLBK'
# symbol[257] <- 'PBPB'
# symbol[263] <- 'QLIK'
# symbol[265] <- 'MEET' #name change to MeetMe
# symbol[271] <- 'exclude' # 'RZ' #going bankrupt
# symbol[275] <- 'RGLS'
# symbol[282] <- 'QUMU' #renamed
# symbol[293] <- 'exclude' #multiple mergers and name changes
# symbol[300] <- 'bought' # 'SRLS' bought for $80.8 million
# symbol[303] <- 'exclude' #SimplePons went private again maybe?
# symbol[322] <- 'PULS' #name change to pulse electronics
# symbol[348] <- 'VRAZ'
# symbol[358] <- 'VRNG'
# symbol[363] <- 'WBSN'
# symbol[365] <- 'XRSC'
# symbol[10] <- 'merger' # bought for $14.50 a share,
#                          56.7 mill shares sold 'ACTV'
# symbol[22] <- 'exclude' # 'FTER' #bankrupted
# symbol[339] <- 'exclude' #based in Hong Kong
# symbol[218] <- 'MSLP'
# symbol[348] <- 'merger' #with DIAG
# symbol[351] <- 'VMEM'
# symbol[336] <- 'exclude' #bankrupt
# symbol[363] <- 'bought' # for $907 million

#Save symbol names
```

```
#write.csv(symbol,"symbolnames.csv")

#load symbols
symbol <- read.csv("symbolnames.csv")
symbol <- as.vector(symbol[,2])

#restore names
ipo$name <- ipo.bu$name

#detects symbol
re.sym <- "([A-Z]){1,5}"

#remove excludable stocks
ind1 <- str_detect(symbol,re.sym)
ipo.c <- ipo[ind1,]
symbol.c <- symbol[ind1]

################################################################################
# The following code queries Yahoo Finance for valuation data, so sometimes
# it can take a very long time to run. I include it as a reference, but
# we have chosen not to have it run in this final version for ease of use.
# Instead, the data we collected using it is read from a .csv file.
################################################################################
#get key stats function (CITE THIS)
#citation:
#http://allthingsr.blogspot.com/2012/10/pull-yahoo-finance-key-statistics.html
# getKeyStats_xpath <- function(symbol) {
#   yahoo.URL <- "http://finance.yahoo.com/q/ks?s="
#   html_text <- htmlParse(paste(yahoo.URL, symbol, sep = ""), encoding="UTF-8")
#
#   #search for <td> nodes anywhere that have class 'yfnc_tablehead1'
#   nodes <- getNodeSet(html_text, "/*//td[@class='yfnc_tablehead1']")
#
#   if(length(nodes) > 0 ) {
#     measures <- sapply(nodes, xmlValue)
#
#     #Clean up the column name
#     measures <- gsub(" *[0-9]*:", "", gsub(" \\(.*?\\)[0-9]*:","", measures))
#
#     #Remove dups
#     dups <- which(duplicated(measures))
#     #print(dups)
#     for(i in 1:length(dups))
#       measures[dups[i]] = paste(measures[dups[i]], i, sep=" ")
#
#     #use siblings function to get value
#     values <- sapply(nodes, function(x)  xmlValue(getSibling(x)))
#
#     df <- data.frame(t(values))
#     colnames(df) <- measures
#     return(df)
```

```
#   } else {
#     break
#   }
# }
#
# #get company valuation data
# stats <- ldply(symbol.c,getKeyStats_xpath)
# rownames(stats) <- ipo.c$name
# write.csv(stats,"valuationdata.csv")

#load valuation data
stats <- read.csv("valuationdata.csv")
colnames.vdata <- as.vector(read.csv("colnames_vdata.csv")[,2])
rownames(stats) <- ipo.c$name

cols <- c("Market.Cap","Enterprise.Value",
          "Enterprise.Value.Revenue",
          'Enterprise.Value.EBITDA', 'Revenue','Revenue.Per.Share',
          'Gross.Profit','EBITDA','Total.Cash','Total.Cash.Per.Share',
          'X52.Week.Change','Shares.Outstanding', 'Float',
          'X..Held.by.Insiders','PEG.Ratio')
stats.c <- stats[,cols]
colnames(stats.c) <- colnames.vdata #restore colname formatting
stats.c.bu <- stats[,cols]

#replace spaces and dashes in colnames
colnames(stats.c) <- str_replace_all(colnames(stats.c), " ", "_")
colnames(stats.c) <- str_replace_all(colnames(stats.c), "52-", "FiftyTwo_")
colnames(stats.c) <- str_replace_all(colnames(stats.c), "/", "_by_")

#convert strings to numbers
mil <- "M"
thou <- "K"
bil <- "B"
for(i in 1:14){
  stats.indm <- str_detect(stats.c[,i],mil)
  stats.indt <- str_detect(stats.c[,i],thou)
  stats.indb <- str_detect(stats.c[,i],bil)
  stats.c[,i] <- str_replace(stats.c[,i],mil, "")
  stats.c[,i] <- str_replace(stats.c[,i],thou, "")
  stats.c[,i] <- str_replace(stats.c[,i],bil, "")
  stats.c[stats.indm,i] <- as.numeric(stats.c[stats.indm, i]) * 1000000
  stats.c[stats.indt,i] <- as.numeric(stats.c[stats.indt, i]) * 1000
  stats.c[stats.indb,i] <- as.numeric(stats.c[stats.indb, i]) * 1000000000
}

#combine data frames
comb <- as.data.frame(cbind(ipo.c,stats.c))
comb.bu <- comb
```

```
#look at count of companies by industry
code.count <- ddply(comb,.(category_code),
                    summarize, Total = length(category_code))
code.count2 <- code.count[which(code.count$Total > 5),]

##############################################################################
#                          Plot 10  Histogram 4
#                look at count of companies by industry
##############################################################################
g <- ggplot(code.count2, aes(x = category_code, y = Total))
g <- g + geom_bar(aes(fill = category_code))
g

#look at market cap avgs for industries
mcap.avgs <- ddply(comb, .(category_code), summarize,
                   avgMC = mean(as.numeric(Market_Cap), na.rm =T))
#remove others
mcap.avgs <- mcap.avgs[-1,]
mcap.avgs[,2] <- mcap.avgs[,2]

#Look at yearly percent change
comb$FiftyTwo_Week_Change <-
  as.numeric(str_replace(comb$FiftyTwo_Week_Change,"%",""))

comb.ten <- comb[which(comb$category_code == 'biotech' |
                         comb$category_code == 'software' |
                         comb$category_code == 'web' |
                         comb$category_code == 'cleantech' |
                         comb$category_code == 'mobile' |
                         comb$category_code == 'enterprise' |
                         comb$category_code == 'ecommerce' |
                         comb$category_code == 'advertising' |
                         comb$category_code == 'games_video' |
                         comb$category_code == 'hardware' |
                         comb$category_code == 'network_hosting' |
                         comb$category_code == 'security' |
                         comb$category_code == 'semiconductor'),]

##############################################################################
#                            Plot 11  Boxplot 1
#             percent change plot by industry categories
##############################################################################
#percent change plot by industry
pct.plot <- ggplot(comb.ten, aes(x = category_code,
                                 y = as.numeric(FiftyTwo_Week_Change),
                                 fill=category_code))
pct.plot <- pct.plot +
  geom_boxplot() + ylab("52-Week Percentage Change")
pct.plot <- pct.plot +
  xlab("Industry") + theme(legend.position="none") + coord_flip()
pct.plot <- pct.plot +
```

```
  theme(panel.background = element_blank(),
        axis.line = element_line(size = .8, color = "grey75"))
pct.plot<- pct.plot +
  ggtitle("Last Year's Share Price Percentage Change by Industry")
pct.plot

#Compare Enterprise Value/Revenue within industries and EV/EBITDA
#Note that high growth industries have high multiples

multiples <- ddply(comb,.(category_code),summarize,
     avg_EVbR = median(as.numeric(Enterprise_Value_by_Revenue),na.rm=T),
     ave_EVbEBIDTA = median(as.numeric(Enterprise_Value_by_EBITDA),na.rm=T))

#rank companies by key metrics
rvec <- colnames(comb)
rvec <- rvec[-c(1:18,21:22,28:32)]

#Tried to do with a loop but R didn't recognize get function for data frame
# names of the form paste0('comb$',name)
ranking <- matrix(nrow = 344, ncol = 8)
ranking[,1] <- rank(as.numeric(comb$Market_Cap))
ranking[,2] <- rank(as.numeric(comb$Enterprise_Value))
ranking[,3] <- rank(as.numeric(comb$Revenue))
ranking[,4] <- rank(as.numeric(comb$Revenue_Per_Share))
ranking[,5] <- rank(as.numeric(comb$Gross_Profit))
ranking[,6] <- rank(as.numeric(comb$EBITDA))
ranking[,7] <- rank(as.numeric(comb$Total_Cash))
ranking[,8] <- rank(as.numeric(comb$PEG_Ratio))

colnames(ranking) <- rvec
rownames(ranking) <- comb$name

agrank <- apply(ranking,1,sum)
comp.rank <- rank(agrank)
names(comp.rank) = NULL
comby <- as.data.frame(cbind(comb$name,comp.rank, symbol.c, ranking))
top <- comby[which(as.numeric(as.character(comby$comp.rank)) >= 329),]
write.csv(top, "top.csv")

list.com <- sort(rank(agrank),decreasing=T)
top.ten <- names(list.com[1:10])

#build table of top ten companies
top.ten.stat <- comb[c(317,135,246,180,300,145,333,208,125,242),]
ten.table <- top.ten.stat[,c(2,19,20,23,26,27)]
rownames(ten.table) <-NULL
#write.csv(ten.table,"ten.table.csv")

##############################################################################
# The following function calls quantmod and collects historical stock price
# data from Yahoo Finance. Because it queries the internet, the code can
```

```
# take a long time to run, so we have commented it out here. I include it as
# a reference, but read the data we collected using it from a csv file for
# ease of use.
##############################################################################
# ##Get daily stocks
# #take out problem symbols
# symbol.s <-symbol.c[-c(110,127,176,205,275,310)]
#
# #initialize comp.stock2
# comp.stock2 <- array()
#
# #break into 2 loops b/c yahoo only allows 200 queries at a time
# for(i in 1:200){
#   stocks <- getSymbols(symbol.s[i],from = "1960/1/1" ,auto.assign=F)
#   comp.stock <- stocks[,6]
#   comp.stock2 <- cbind(comp.stock2,comp.stock)
#   Sys.sleep(1)
# }
# Sys.sleep(20)
# for(i in 201:length(symbol.s)){
#   stocks <- getSymbols(symbol.s[i],from = "1960/1/1" ,auto.assign=F)
#   comp.stock <- stocks[,6]
#   comp.stock2 <- cbind(comp.stock2,comp.stock)
#   Sys.sleep(1)
# }
# #remove null set and fix colnames
# comp.stock2 <- comp.stock2[,-1]
# colnames(comp.stock2) <- gsub(".Adjusted","",colnames(comp.stock2))
# comp.stock2.bu <-comp.stock2
# #write.csv(ten.table.csv,"stocks_comp.csv")
#
# #get dates as a vector
#date.intv <- time(comp.stock2)
#write.csv(as.character(date.intv),"stocks_comp_dates.csv")
stock.tab = read.csv(file="ten.table.csv", header =TRUE)
stock.tab = stock.tab[,-1]
stock.tab[,2:6] = stock.tab[,2:6]/10^9
print(xtable(stock.tab, digits = 0, caption="\\textbf{The Top Ten IPO Companies}",
             label="ByCompType_Table"),
      size="footnotesize", #Change size; useful for bigger tables
      include.rownames=FALSE, #Don't print rownames
      include.colnames=FALSE, #We create them ourselves
      hline.after=NULL, #We don't need hline; we use booktabs
      add.to.row = list(pos = list(-1,
                                   nrow(stock.tab)),
                        command = c(paste("\\toprule \n",
                                          "Company & Market Cap & Enterp. Value
                                          & Revenue & EBITDA* & Total Cash\\\\\n",
                                          "\\midrule \n"),
                                    "\\bottomrule \n")
      )
```

)