Springboard Capstone 3: A Study on Angel Investing

Carlee Price

September 4, 2017

A STUDY ON THE ANGEL INVESTING LANDSCAPE AND OUTCOMES FOR INVESTORS

BACKGROUND & OBJECTIVES

Angel Investing is a form of equity investing. It is similar to the stock market in the sense that it allows individuals to own a piece of a company in which they have no operating role and to benefit financially from that company's success. Angel Investing is different from stock market investing in important ways, most notably: liquidity (ability to turn the investment back into cash at will), visibility (how much is the investment worth at a given point in time) and history (how much do we know about the behavior of these type of investments over time).

We'll be using some industry vernacular in the report:

Equity: the type of ownership interest an investor in this space holds, appreciates most readily when the value of the underlying asset (the company) increases.

Rounds: a set of investors and a fixed amount of capital raised over a set period of time. Investors in each round invest at the same valuation, the rounds will become larger and the company more valuable as time passes and its operations grow.

Returns: the growth in value of the money invested from the time that the investment was made to the time it was closed out (money returned to investor in the form of cash). Returns can be reported as absolute (simple end value/start value), or can be annualized (standardized to a yearly return figure; this also accounts for compounding).

Private Market: Unlike stocks that are listed on an exchange, there exists no trading venue for Angel Investments. This means that once the investment is made, the investor is unable to sell their interest (get money back) until the entire company is liquidated (sold). Positioning the company for sale and executing a sale on favourable terms is therefore one of the major objectives of the company's management.

Exit: The sale of private angel-backed company. Angel Investors will only see the return of their investment on exit. This is the preferred outcome.

Acquisition: Can be either price-disclosed (typically through a joint press release) or undisclosed. Companies might choose to keep their transaction prices confidential for a variety of reasons. In these cases, a return will be generated for investors, but the amount must be modeled. This will be one of the primary goals of this project.

Closure: Another way in which the investment can be resolved (closed out). The company in which Angel invested ceases operations. Final terminal value of the investment is zero, and returns - 100%.

Portfolio Theory: Considers how the combination of individual securities/positions into a collection/portfolio can reduce the risk of the asset class. The mechanism is correlation; the underlying theory is that a portfolio of securities with imperfect correlation to each other will outperform on a risk-adjusted basis individual positions.

OBJECTIVES

Return exectations are critical for an investor in any asset class. At what type of annual rate should they (investor) expect to see their money grow, and what is the typical deviation around generalized, long-run returns; what is the risk? There is a dearth of information in this space as compared to public equities. In drawing broad conclusions about return and risk, the bulk of the work will be in populating missing data. We will build theories and models around how the missing data should best be populated, and visualize the characteristics of the resulting data set.

THE DATASET

Crunchbase is an aggregator of information (primarily qualitiative) around Angel-backed companies. Their data is private (and expensive) although in 2014 they published publicly a sample of seven trailing years of their data. This data forms the basis for our study. There are three distinct tables we will use.

Companies: largely descriptive, includes date of founding, location, name, etc. Rounds: also descriptive, but includes the amount of financing secured, dates, etc. Acquisitions: describes the exits in the space, including acquirers, date, values

The data has many shortcomings: for seed and angel rounds of investment, the data is largely self-reported by investors themselves rather than being an objective sample. There are some data integrity and cleaning issues that must be addressed. The dataset captures companies that are Angel financed (our area of focus) but also those with Venture backing and even debt. While the set is large (54k rows) our study necessarily focuses on a small subset (a few thousand). The dataset is complete in that the categories of data captured are meaningful. So the analytic steps established here can most certainly form the basis for processing data that may be more complete/of higher qualtiy, in the future.

PART I: WORKING WITH THE CRUNCHBASE DATA

LOAD & TRANSFORM

We start with the Companies data, keeping only the US based companies that were first funded in 2007 or later. Remember the most recent data in this set is from end-2014.

```
#read in the file that includes all the company information for this database
library(plyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.3
```

```
set.seed(261)
allcompanies = read.csv("companies.csv")
```

```
#keep only the companies that are US-based
companies <- subset(allcompanies, country_code == "USA")
#we also want to keep only the companies with first_funded_at dates of 2007 or later
#first cast this column as a date
companies$first_funding_at <- format(as.Date(companies$first_funding_at), "%Y/%m/%d")
#then screen for chosen date
companies <- subset(companies, first_funding_at >= '2007/01/01')
nrow(companies)
```

```
## [1] 25837
```

```
#check for completeness in the resulting set; list of rows that have missing values
nrow(companies[!complete.cases(companies),])
```

```
## [1] 4688
```

```
#what information has been captured for these companies?
colnames(companies)
```

```
##  [1] "ï..permalink"      "name"              "homepage_url"
##  [4] "category_list"     "market"            "funding_total_usd"
##  [7] "status"            "country_code"      "state_code"
## [10] "region"            "city"              "funding_rounds"
## [13] "founded_at"        "founded_month"     "founded_quarter"
## [16] "founded_year"      "first_funding_at"  "last_funding_at"
```

After subsetting the data, we have 25,837 rows of which 4,688 are incomplete. Some of this missing data won't bother us, but in the case of funding_total_usd, it is important. Our study focuses on companies that raise money from outsiders. Our conclusions will largely depend on what happens to the dollars these companies raise. If we don't know in a case (row) how may dollars are involved, it becomes difficult to draw conclusions. Below, we will remove any companies for which the funding total shows N/A.

DATA EXPLORATION

WIN/LOSS RATIO

Starting with the most simple outcome evaluation, let's consider how frequently companies are closed (return falls to zero, all money lost) versus acquired (return is cemented at some non-zero amount). What is the ratio of each within the total population of companies?

```
#transform funding column by stripping punctuation & converting to integer
companies$funding_total_usd <- as.numeric(gsub("[[:punct:]]", "", companies$funding_total_usd))
#strip out companies funded with too small a pool to be relevant for our audience
companies <- subset(companies, funding_total_usd > 100000)
nrow(companies)
```

```
## [1] 19401
```

```
#replace missing status fields
levels(companies$status)[1] <- "unknown"
table(companies$status)
```

```
##
##   unknown  acquired    closed operating
##       418      1592       859     16532
```

```
#find the ratios
table(companies$status)[2]/nrow(companies)
```

```
##   acquired
## 0.08205763
```

```
table(companies$status)[3]/nrow(companies)
```

```
##     closed
## 0.04427607
```

Counting exits (acquired) as a win (8.2%) and closeds as a loss (4.4%) looks encouraging; the ratio is > 1. Will the numbers be different if we look at just one vintage of companies? Let's look just at companies that were funded in 2007, not after. There are 1276 rows in this subset.

```
companies07 <- subset(companies, first_funding_at <= '2007/12/31')
nrow(companies07)
```

```
## [1] 1238
```

```
table(companies07$status)
```

```
##
##   unknown  acquired    closed operating
##         7       322       156       753
```

```
table(companies07$status)[2]/nrow(companies07)
```

```
##  acquired
## 0.2600969
```

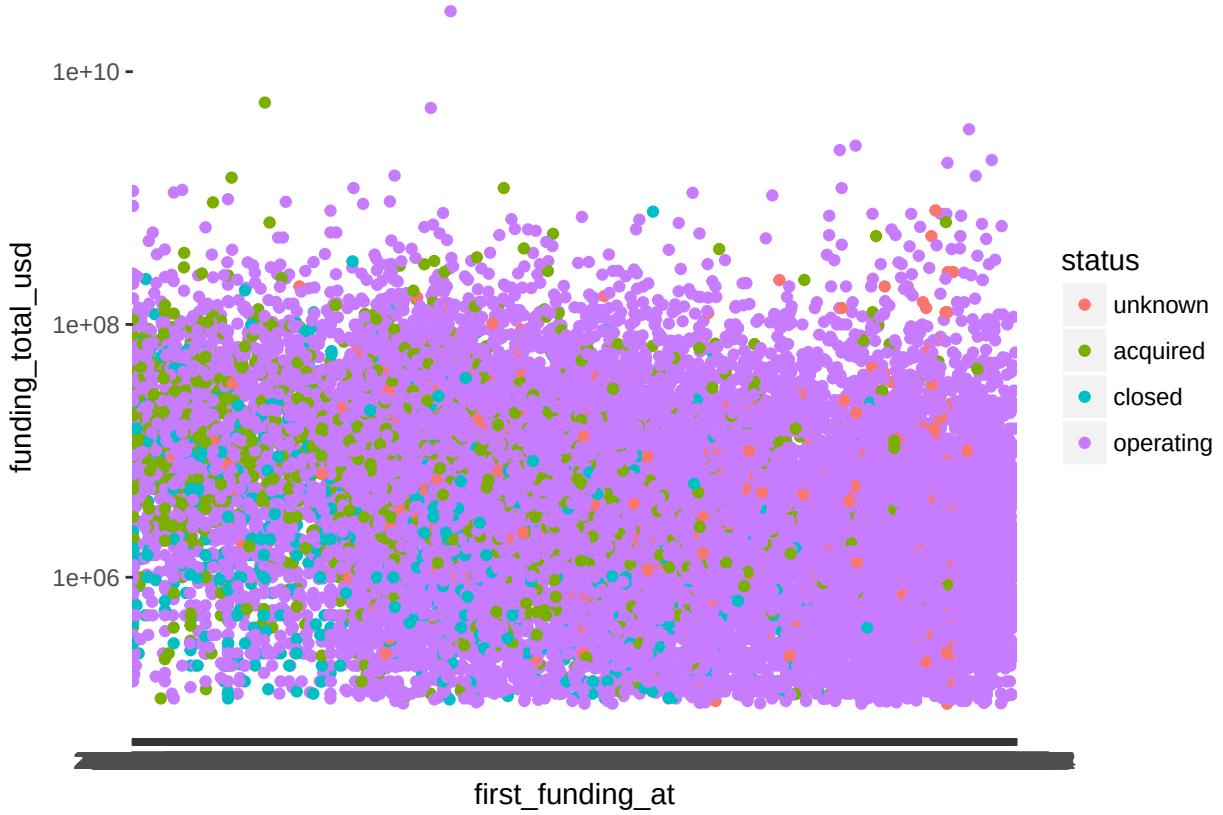
```
table(companies07$status)[3]/nrow(companies07)
```

```
##    closed
## 0.1260097
```

Here we see the wins (26.0%) and losses (12.6%) are much higher; an even better ratio. It seems that spending more time in the market leads to more resolutions (closure or acquisition).

We can also visualise how younger cohorts of companies skew towards "Operating", while older companies are more likely to be listed as "closed" or "acquired"

```
#first remove the rows with mising funding totals
companies3 <- subset(companies, !is.na(companies$funding_total_usd))
#then plot.  this could look better, specifically the x-axis.
ggplot(companies3, aes(x = first_funding_at, y = funding_total_usd, col = status)) + geom_jitter() + sca
```



DOLLAR-ON-DOLLAR RETURNS

Let's get more granular, and look not just at the outcome but at the degree of success. Closure of a business can only result in 100% loss but an exit can clearly generate > 100% returns. We're working towards a picture of total portfolio returns.

Add two new data sets from the same source: funding rounds (amounts raised) and acquisitions (exit values). Both will need to be tidied in a similar fashion to our companies table.

```
allrounds = read.csv("rounds.csv")
rounds <- subset(allrounds, company_country_code == "USA")
rounds$raised_amount_usd <- as.numeric(gsub("[[:punct:]]", "", rounds$raised_amount_usd))
allacquisitions = read.csv("acquisitions.csv")
acquisitions <- subset(allacquisitions, company_country_code == "USA")
acquisitions$price_amount <- as.numeric(gsub("[[:punct:]]", "", acquisitions$price_amount))
```

In the process of this exploration, some data-quality issues emerge. We address two of these here.

First, in acquisitions set: Riot Games was acquired by Tenecet for 400Mm, not the USD 4,000 noted.

```
#calling this by index rather than name is dangerous - consider changing
acquisitions$price_amount[5344] <- 400000000
```

Second, in companies set: Aptalis Pharma is listed as having raised a single round (173k seed capital) and then proceeding to exit for $2.9Bn. That would create an impressive return indeed for funders. The reality is that the company actually was spun out from the merger of two established pharmaceutical giants. At the date of formation had seven branded products in market, and a robust development pipeline. There was far

more than USD 173k in value within the company. Because this is incorrect and because the actual nature of the enterprise disqualifies it from this (startup) data set, it comes out as well.

We want to add this information to our set of companies, created above, for which funding data exists.

Each of these tables has information useful to our returns analysis. We need to create a unique key for each, in this case namecity (since many of these companies share names with other, unrelated companies created in different places). https://smbrate.com/We can also be selective about bringing only useful rows into the joined table.

```
nrow(rounds)
```

```
## [1] 54313
```

```
#select just the columns we want to see, and just the rows meeting our "seed or angel funding" criteria
rounds3 <- subset(rounds, rounds$funding_round_type %in% c("seed", "angel") & raised_amount_usd >= 10000
#create namecity field for proper joining
rounds3$namecity <- paste(rounds3$company_name, rounds3$company_city)
#eliminate duplicates
rounds2 <- distinct(rounds3, company_name, company_city, .keep_all = TRUE)
```

Repeat formatting steps for companies and acquisitions tables.

```
#select just the columns from companies that we want to see in the new dataset
companies2 <- subset(companies, select = c(name, funding_total_usd, status, state_code, region, city, fu
#remove Aptalis Pharma
companies2 <- companies2[!companies2$name == "Aptalis Pharma",]
acquisitions2 <- subset(acquisitions, select = c(company_name, company_city, acquirer_name, acquired_at
#we can't simply join on name, since there are a number of unique companies that share a name.
#create a new column namecity that includes both the name of the company & the city of its founding
companies2$namecity <- paste(companies2$name, companies2$city)
acquisitions2$namecity <- paste(acquisitions2$company_name, acquisitions2$company_city)
#then join them
fullset <- merge(x = rounds2, y = companies2, by = "namecity", all.x = TRUE)
fullset <- merge(x = fullset, y = acquisitions2, by = "namecity", all.x = TRUE)
```

There are duplicate rows in this merged dataframe. If we screen using unique, we remove 12 rows that are perfectly identical. There are also companies in here which have been acquired twice (see: Forrst). Stakeholders may in this case be getting paid twice, but not necessarily. Further, the amount of the gain in the second case would be the differential between that and the first bid, rather than the entire amount. This gets tricky, but can be properly addressed only where both prices (acquisition 1 and acquisition 2) are disclosed.

Uniqueness for our case will include name + first funding + acquired date. Eliminates 474 rows.

```
nrow(fullset)
```

```
## [1] 6640
```

```
fullset <- distinct(fullset, name, first_funding_at, acquired_at, .keep_all = TRUE)
nrow(fullset)
```

```
## [1] 6166
```

There are also some rows that have been incorrectly classified as “closed” or “unknown” for which an acquirer is listed. Let’s reclassify these.

```
fullset$status[which(fullset$status %in% c("closed", "unknown") & !is.na(fullset$acquirer_name))] <- "a
```

As a result of the join, we have several repeating columns. Let’s take them out.

```
drops <- c("company_name.x", "company_state_code", "company_region", "company_city.x", "company_city.y"
fullset <- fullset[ , !(names(fullset) %in% drops)]
```

And we have to filter out for repeats again. We'll also use this opportunity to correct several other faulty datapoints. And to refactor the status variable into a logical order, and add a field (metro) which indicates whether the company is located in one of the two geographic areas associated with greater starup success (as nearly half of these companies are). We're left with 6043 companies in the study set.

```
fullset$acquired_at[fullset$name == "DataPad"] <- "2014-09-30"
fullset$acquired_at[fullset$name == "Modern Feed"] <- "2009-06-01"
fullset <- fullset[!fullset$name == "Buccaneer",]
fullset <- fullset[!(fullset$name == "Roost" & fullset$founded_year == 2013),]
fullset <- subset(fullset, !is.na(fullset$name))
#refactor the levels of "status" to correspond with low value = less favourable outcome, high = better
fullset$status <- ordered(fullset$status, levels = c("closed", "unknown", "operating", "acquired"))
#assign closed companies a price amount value of 0
fullset$price_amount[which(fullset$status == "closed")] <- 0
#create new dummy variable to separate companies in areas of high return
fullset$metro = ifelse((fullset$region == "SF Bay Area" | fullset$region == "New York City"), 1, 0)
table(fullset$metro)
```

```
##
##    0    1
## 3373 2712
```

```
nrow(fullset)
```

```
## [1] 6085
```

Now we're ready to look at returns. Angels want to see exits, as that's the only way they're getting the money back, and the source of their returns. Our "status" field tells us which companies have been acquired and from that we got % exits (successes). We also want dollar-on-dollar % return, which requires acquisition price information (numerator) in addition to total funding information (denominator). Here we see another opportunity to remove corrupt data, specifically OhmData which is purported to have been funded for 185k and acquired for 3Mm just a month later.

We've also made the judgement to remove WhatsApp from the group. This transaction was an outlier to such an extent that it may unfairly impact our analysis & conclusions.

```
#subset for where acquisition price information is disclosed
acquired <- subset(fullset, price_amount > 1, funding_total_usd >1)
acquired <- acquired[!acquired$name == "OhmData",]
acquired <- acquired[!acquired$name == "WhatsApp",]  #is this the right thing to do?
acquired <- acquired[!acquired$name == "Medafor",]
nrow(acquired)
```

```
## [1] 90
```

```
#create a new column that shows total $ returned to investors against total $ raised
acquired$return <- acquired$price_amount/acquired$funding_total_usd
#use sum here instead of averaging the acquired$return column as it's more reflective
sum(acquired$price_amount)/sum(acquired$funding_total_usd, na.rm = TRUE)
```

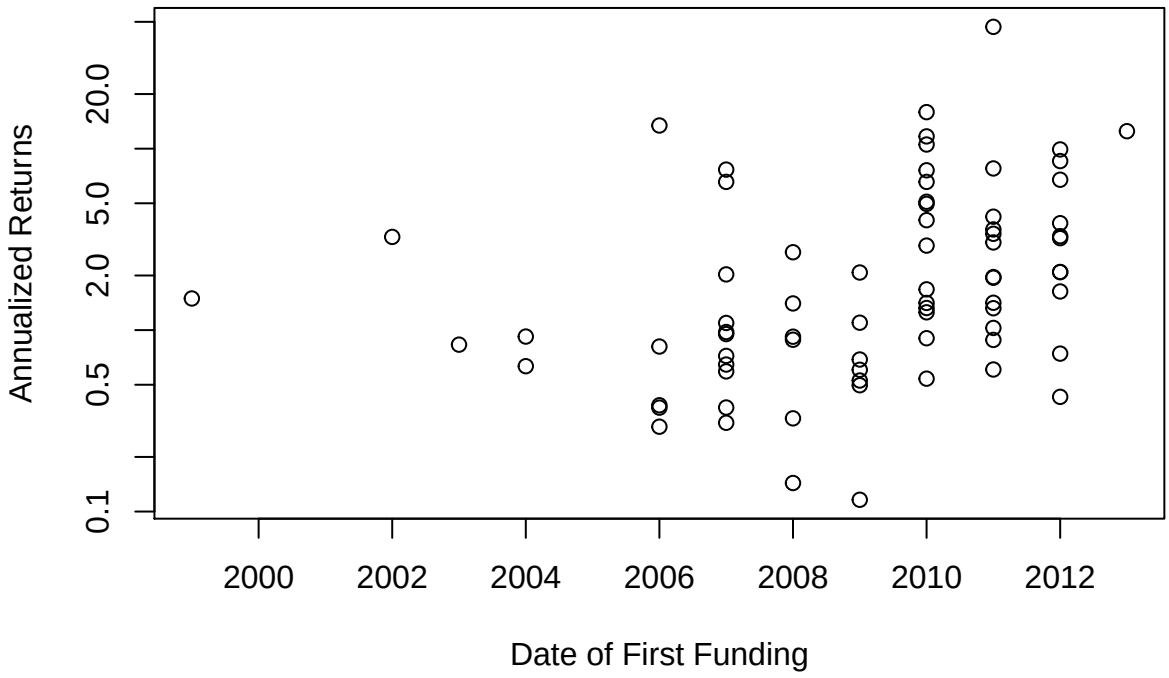
```
## [1] 8.245818
```

Only 90 of the 6085 companies described as "acquired" in our joined set include pricing information. Still a reasonable sample size, and the numbers are encouraging (8.24X return on total amount raised) but let's look closer. First, to annualize results.

Let's create a field that shows annualized returns by company, and then plot these against time to see if there are any interesting trends.

```
#first convert acquired_at to proper date format
acquired$acquired_at <- format(as.Date(acquired$acquired_at), "%Y/%m/%d")
#then to calculate how many days a company spent between funding and acquisition
acquired$age_at_acquisition <- (as.integer(as.Date(acquired$acquired_at) - as.Date(acquired$first_fundi
acquired <- subset(acquired, age_at_acquisition >= 1)
#whiche we then use to calculate annualized returns
acquired$annual_return <- (acquired$return ^ (365.25/as.integer(acquired$age_at_acquisition)) - 1)
plot(acquired$founded_year, acquired$annual_return, log = "y", ylab = "Annualized Returns", xlab = "Dat
```

```
## Warning in xy.coords(x, y, xlabel, ylabel, log): 8 y values <= 0 omitted
## from logarithmic plot
```



For companies that were acquired, those founded more recently have generated greater annual returns.

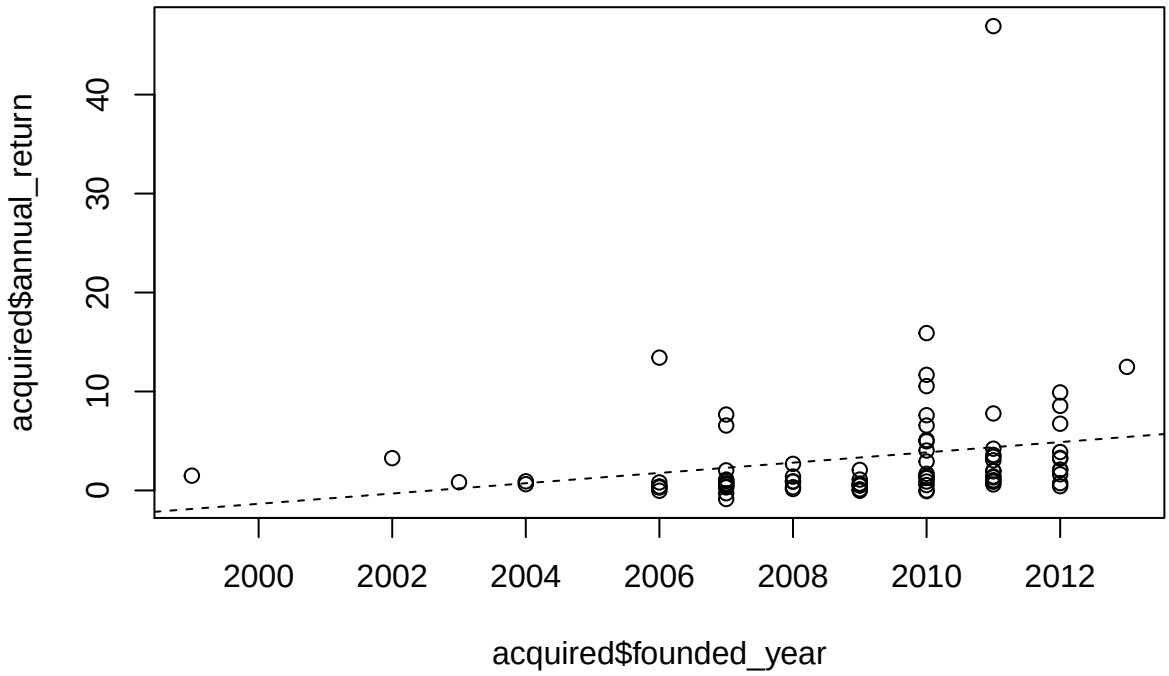
```
model1 <- lm(annual_return ~ founded_year, data = acquired)
summary(model1)
```

```
##
## Call:
## lm(formula = annual_return ~ founded_year, data = acquired)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.457 -2.678 -1.580  0.190 42.552
##
```

```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1040.6410   516.8066  -2.014   0.0475 *
## founded_year     0.5196     0.2573   2.020   0.0468 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.927 on 78 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.04971,    Adjusted R-squared:  0.03753
## F-statistic:  4.08 on 1 and 78 DF,  p-value: 0.04682
```

The relationship is positive (founded later = higher return) and significant at the 5% threshold. We can visualise the relationship.

```
plot(acquired$founded_year, acquired$annual_return)
abline(lm(annual_return ~ founded_year, data = acquired), lty = 2)
```

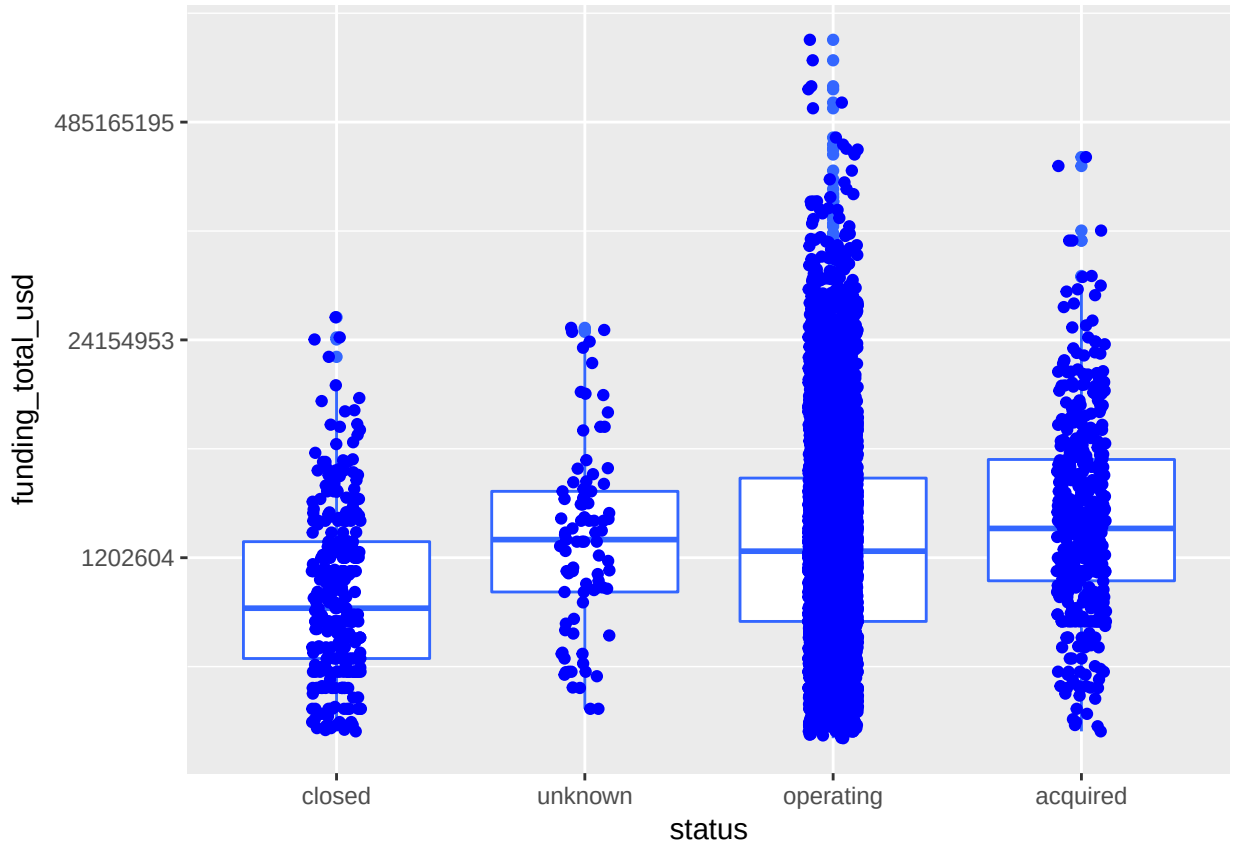


DEEP DIVE: AMOUNTS RAISED AND EFFECT ON STATUS & RETURNS

We can also examine the relationship between funding totals and stauts. Does total funding amount reliably influence the staus (operating, closed, acquired) for the company? Remember closed is bad and acquired is good. Operating can go either way, we address this later in the report. First we reorder our status factor levels from bad to good.

```
fullset$status <- ordered(fullset$status, levels = c("closed", "unknown", "operating", "acquired"))
#ggplot(fullset, aes(x=status, y=funding_total_usd)) + geom_boxplot() + scale_y_continuous(trans = "log
```

```
ggplot(fullset, aes(status, funding_total_usd)) + geom_boxplot(colour = "#3366FF") +
  geom_jitter(width = 0.1, colour = "blue") +
  scale_y_continuous(trans = "log")
```



Then we create a model on this factor.

```
model2 <- lm(as.integer(status) ~ funding_total_usd, data = fullset)
summary(model2)
```

```
##
## Call:
## lm(formula = as.integer(status) ~ funding_total_usd, data = fullset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.96743  0.04506  0.04648  0.04686  1.04700
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.953e+00  7.020e-03  420.66   <2e-16 ***
## funding_total_usd 4.387e-10  2.003e-10    2.19   0.0285 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5392 on 6083 degrees of freedom
## Multiple R-squared:  0.0007881,  Adjusted R-squared:  0.0006239
## F-statistic: 4.798 on 1 and 6083 DF,  p-value: 0.02853
```

Here again we see a relationship significant at the 5% threshold. More funding leads to better outcomes (remember our factor variable status has been ordered). The magic number here is 4 (our "acquired" level: we want these companies to be acquired).

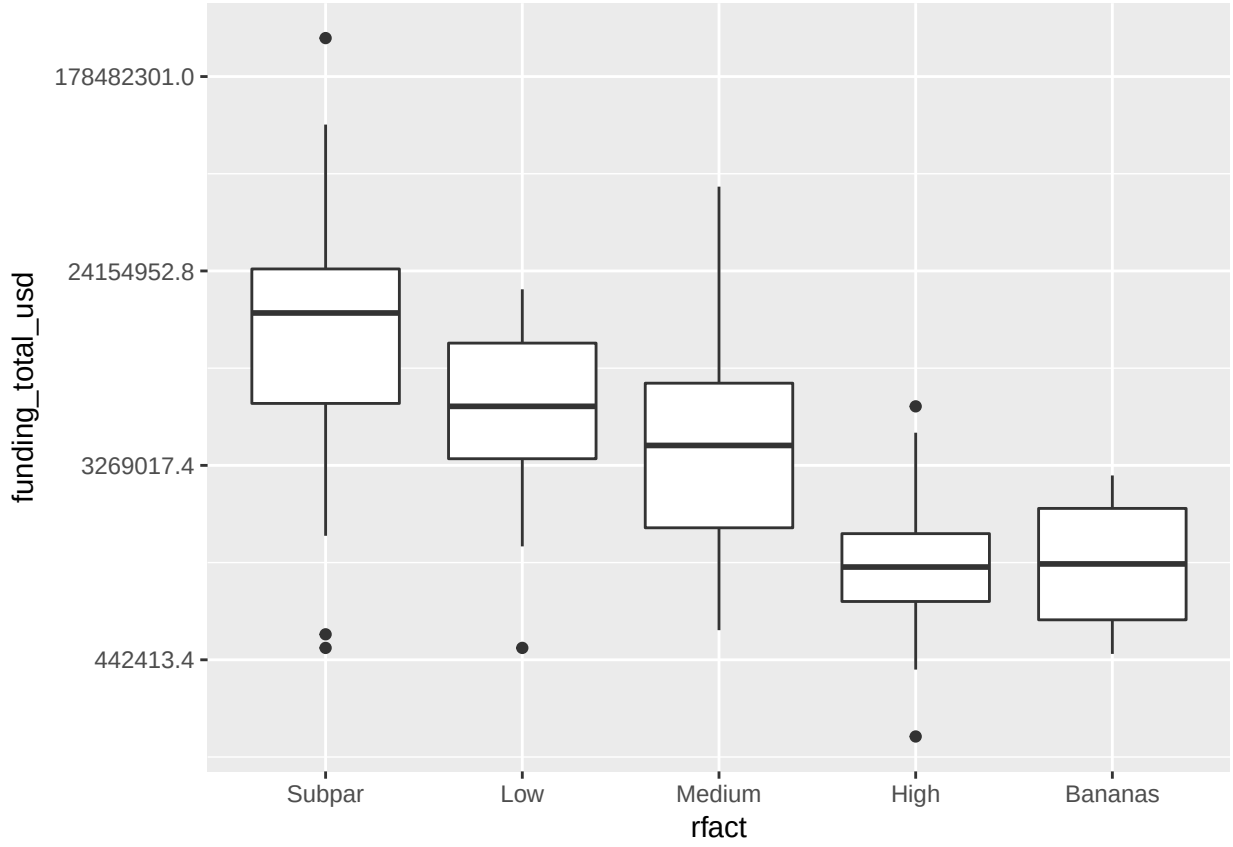
```
(4 - model2$coefficients[1])/model2$coefficients[2]
```

```
## (Intercept)
##  2386604633
```

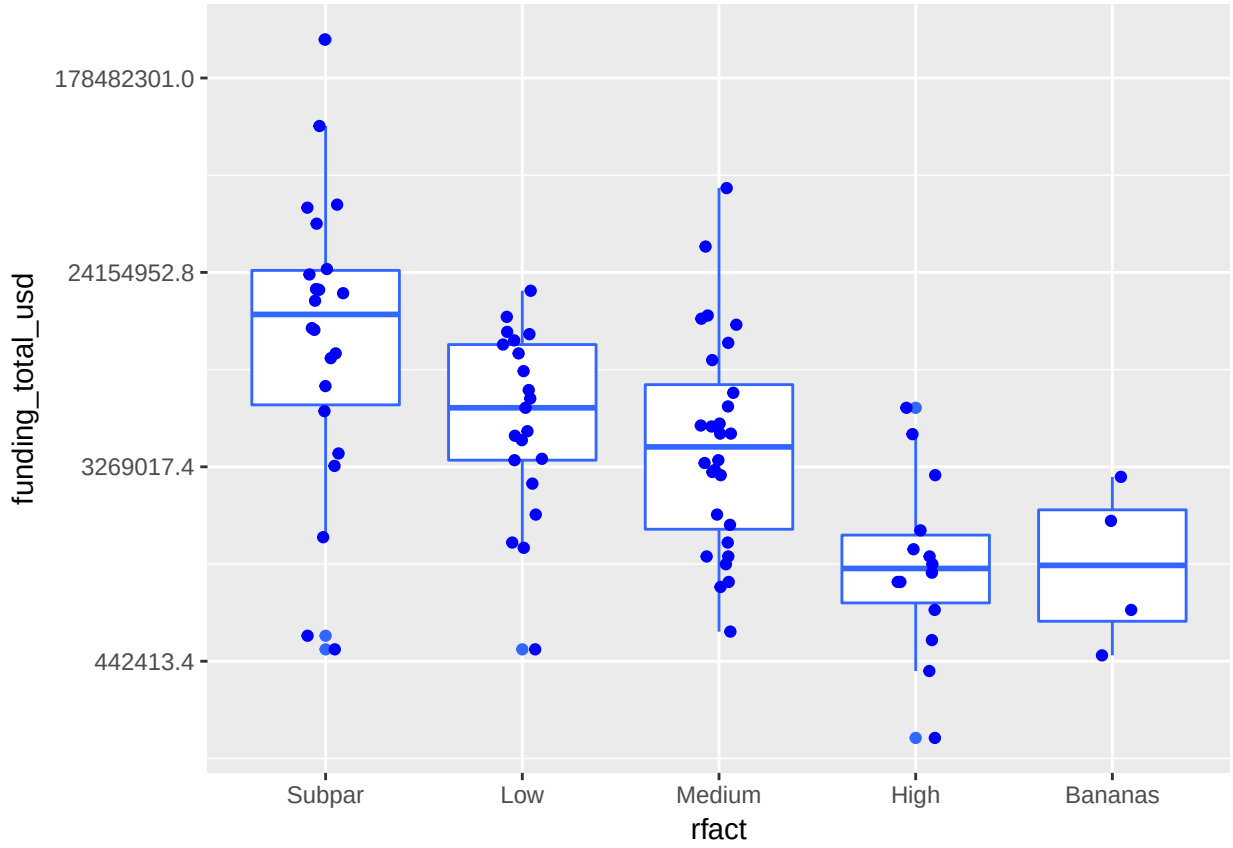
This model tells us that companies improve their probability of being acquired when they raise $2.43Bn or more.

And then WITHIN outcomes, so for companies specifically that were acquired, were those that received more funding acquired at higher prices?

```
#set breaks for each funding bucket, levels came about through trial & error
grp <- c(-1, .5575, 1.25, 5, 12, 60)
#set factors & assign names
acquired$rfact = cut(acquired$annual_return, breaks = grp,labels=c('Subpar','Low', 'Medium','High', 'Ba
#create a boxplot
#plot(x = acquired$rfact, y = acquired$funding_total_usd, log = 'y', ylab = "funds raised", xlab = "qua
ggplot(acquired, aes(x=rfact, y=funding_total_usd)) + geom_boxplot() + scale_y_continuous(trans = "log"
```




```
ggplot(acquired, aes(rfact, funding_total_usd)) + geom_boxplot(colour = "#3366FF") +
  geom_jitter(width = 0.1, colour = "blue") +
  scale_y_continuous(trans = "log")
```



Here again we can see that companies are most likely to appear generate "bananas" returns if they've raised less money. Likewise, those companies that have raised the most are in the lower-tier returns buckets. Interesting!

Is it a statistically significant relationship? Let's get the z score.

```
#there must be an easier way to do this
ftusd <- sd(acquired$funding_total_usd, na.rm = TRUE)
ftumn <- mean(acquired$funding_total_usd, na.rm = TRUE)
a <- subset(acquired, acquired$rfact == "Bananas")
amean <- mean(a$funding_total_usd)
(amean - ftumn)/(ftusd/(sqrt(nrow(a))))
```

```
## [1] -0.7047281
```

```
b <- subset(acquired, acquired$rfact == "Subpar")
bmean <- mean(b$funding_total_usd)
(bmean - ftumn)/(ftusd/(sqrt(nrow(b))))
```

```
## [1] 2.964359
```

Returns in the Subpar category are much more of an outlier than those in the Bananas category. Raising large amounts of money is coincident with lower returns.

Your company will very likely be acquired if you've raised gobs of money, but the resulting returns are likely to be low.

Finally, we'll eliminate the rfact column as it is no longer needed.

```
acquired$rfact <- NULL
```

OTHER EXPLORATIONS

We also know that companies in major metropolitan areas (Bay Area and New York specifically) tend to attract more attention and therefore improved exit opportunities.

```
model3 <- lm(as.numeric(status) ~ metro, data = fullset)
summary(model3)
```

```
##
## Call:
## lm(formula = as.numeric(status) ~ metro, data = fullset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.99299  0.00701  0.07441  0.07441  1.07441
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.92559    0.00927 315.595  < 2e-16 ***
## metro        0.06741    0.01389   4.855 1.24e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5384 on 6083 degrees of freedom
## Multiple R-squared:  0.003859,   Adjusted R-squared:  0.003695
## F-statistic: 23.57 on 1 and 6083 DF,  p-value: 1.237e-06
```

This is a very significant relationship. Companies that are located in NY & SF are much more likely to see a positive outcome (more acquireds, fewer closeds) than companies in other regions of the country.

In summary, we know that greater amounts of funding and a favourable geographic location influence positively the likelihood of a positive outcome for these companies, as do later founding years.

POPULATING MISSING ACQUISITION PRICES

We know what returns on our fully-documented transactions are, but in order to get a full picture of returns, we need to take into account the other transactions as well. For this we'll go back to our fullset and work to understand the metrics here. We can step closer to the truth on returns by keeping sum of disclosed prices in the numerator, but changing the denominator to all the funds put to work during this same period.

```
sum(acquired$price_amount)/sum(acquired$funding_total_usd)
```

```
## [1] 8.114933
```

```
sum(acquired$price_amount)/sum(fullset$funding_total_usd)
```

```
## [1] 0.2423171
```

This gives us a sense of the magnitude of effect including our incomplete cases in the analysis will have. We can't simply assume that the start by differentiating priced/acquired from unpriced/acquired

We'll need an age column for the remaining companies also. For those that haven't been acquired, and remain ongoing, this will be the end of the study period. For those that are closed, we assume a final operating date that is a full year after their last fundraise. For those that continue operating, and for the unknowns, this will be the end of the study period. This column will be used to calculate annual (rather than absolute) returns.

```
fullset$last_funding_at <- format(as.Date(fullset$last_funding_at), "%Y/%m/%d")
end <- as.Date("12/31/14", "%m/%d/%y")
fullset$age <- ifelse(fullset$status == "acquired", as.integer(as.Date(fullset$acquired_at) - as.Date(fu
```

And some cleaning here.

```
#remove the rows that show negative ages for companies
fullset <- subset(fullset, age >= 1)
#what do ages look like for companies in different "status" buckets?
fullset %>%  group_by(status) %>%
          summarise(raised = sum(funding_total_usd), rounds = mean(funding_rounds), age = mean(age))
```

```
## # A tibble: 4 × 4
##      status      raised   rounds       age
##       <ord>       <dbl>    <dbl>     <dbl>
## 1    closed   503098575 1.559105  532.0863
## 2   unknown   322472681 2.045977  978.9655
## 3 operating 33473608764 2.120755 1010.0462
## 4  acquired  2857931052 2.181406  902.2971
```

Now we'll separate out those companies that are acquired at undisclosed prices, as we did above with those that were acquired and disclosed. We'll call this new table "guess". There are 350 companies in this group.

```
table(fullset$status)
```

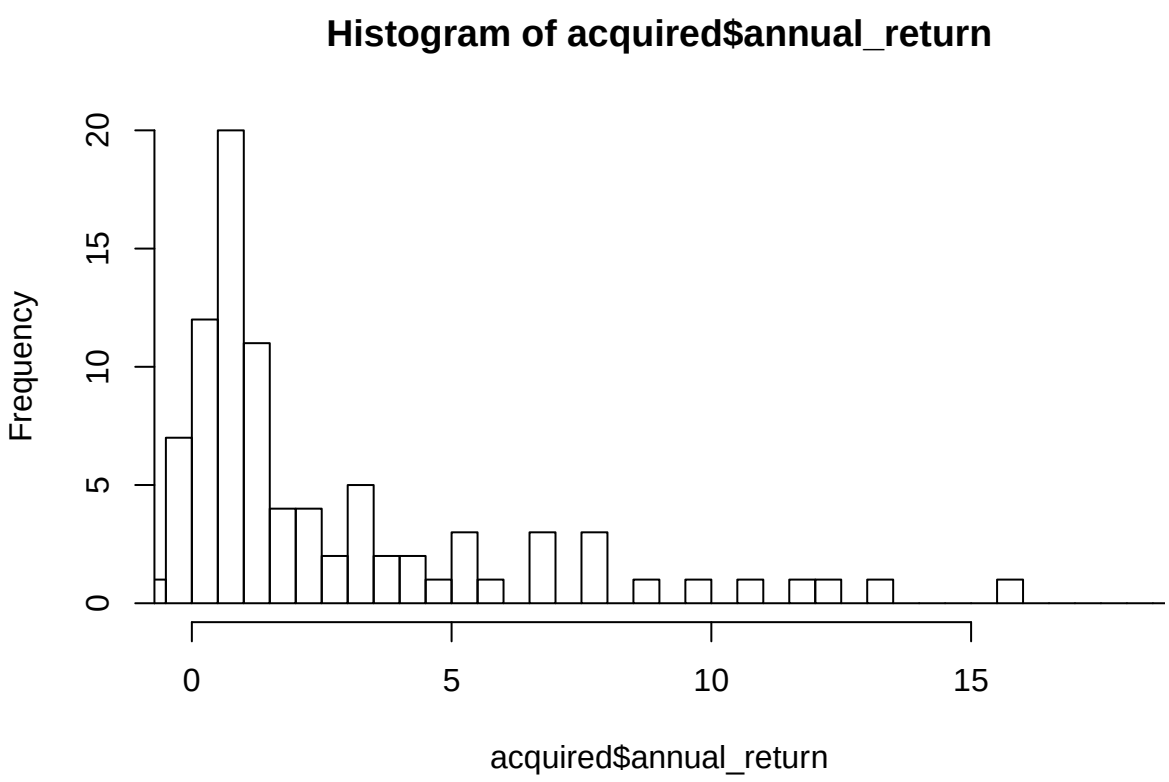
```
##
##    closed   unknown operating  acquired
##       313        87      5242       441
```

```
guess <- subset(fullset, fullset$status == "acquired" & is.na(fullset$price_amount), drop = TRUE)
nrow(guess)
```

```
## [1] 350
```

Let's have a crack at populating the "price_amount" column for these companies. The distribution should reasonably resemble that of our "priced acquired" group. Which is long-tailed with a single outlier: nPulse Technology, which was acquired in less than a year for 23X the invested amount, for annual return of 470%. Wow.

```
hist(acquired$annual_return, breaks = 80, xlim = c(0,18))
```




```
acquired[54,c(5, 6, 15, 16, 20)]
```

```
##                     name funding_total_usd acquirer_name acquired_at
## 3624 nPulse Technologies           2947189       FireEye  2014/05/06
##      age_at_acquisition
## 3624                299
```

Much of the modelling work done above described influences on "status", which is already known for this group. Let's see if some other numeric fields in our table will tell us about price

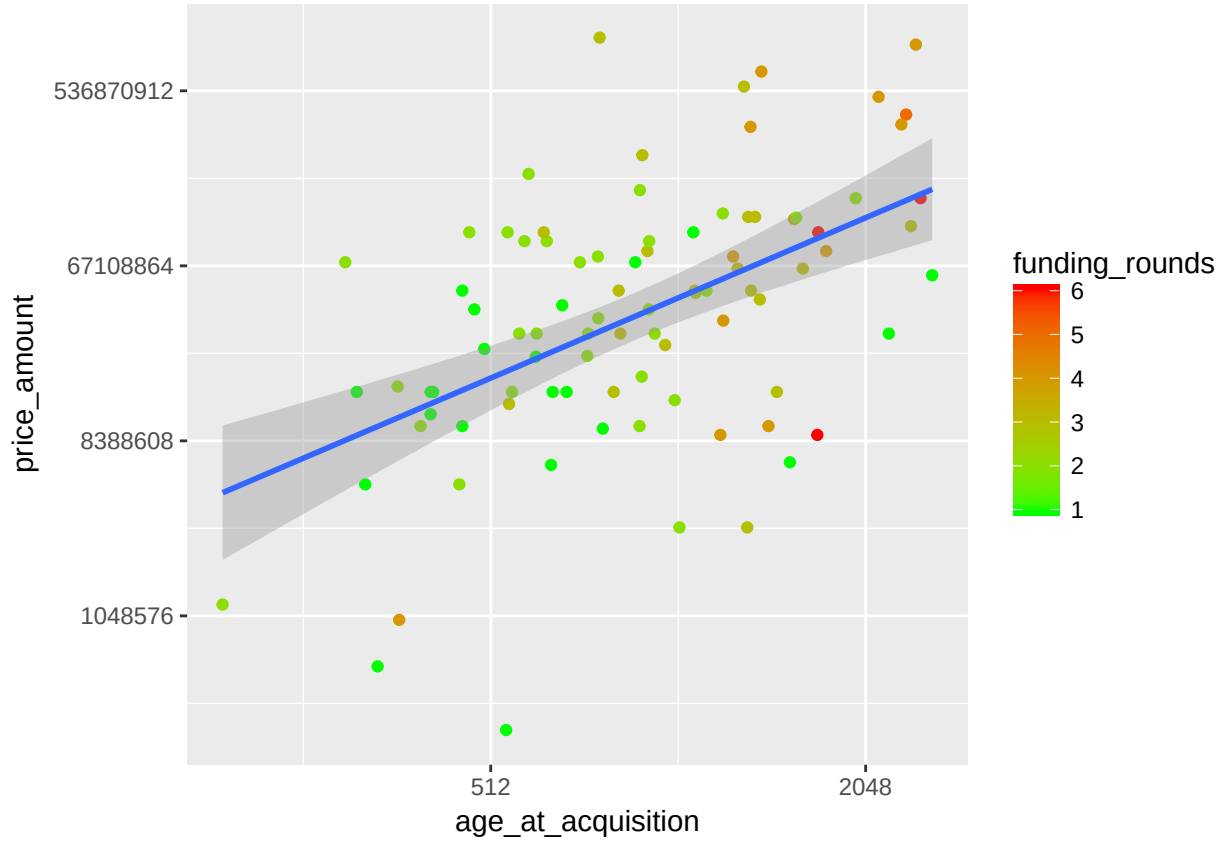
```
model4 <- lm(price_amount ~ age_at_acquisition, data = acquired)
summary(model4)
```

```
##
## Call:
## lm(formula = price_amount ~ age_at_acquisition, data = acquired)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -227217378  -64065482  -29150305    1270972  939721681
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -18688196   35628620  -0.525 0.601246
## age_at_acquisition    116758      30136   3.874 0.000207 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 166700000 on 87 degrees of freedom
## Multiple R-squared:  0.1471, Adjusted R-squared:  0.1373
## F-statistic: 15.01 on 1 and 87 DF,  p-value: 0.0002067
```

And a visualisation:

```
ggplot(acquired, aes(age_at_acquisition, price_amount)) +
 geom_point(aes(color = funding_rounds)) +
 scale_color_gradient(low = "green", high = "red") +
 geom_smooth(method ="lm") +
 scale_y_continuous(trans = "log2") +
  scale_x_continuous(trans = "log2")
```



So age influences price in a meaningful way. Older (which not surprisingly means more total funding) companies recieve higher prices on acquisitions.

Now to populate our missing price fields, using the information from our model: that older companies are worth more.

There's no reason to believe that the returns generated on exit by this group are distributed differently than for our "price-disclosed" group. So we'll assign returns to the "price undisclosed" group that are sampled from the same distribution.

We create a checkpoint here where we can experiment with modelling and come back to our unadjusted sets to reset if necessary

```
acquired2 <- acquired
guess2 <- guess
```

```
#here we can reset
acquired <- acquired2
guess <- guess2
```

If we use returns from ourprice disclosed set to populate the missing fields within our price undisclosed set, the "acquisition price" numbers that result can in some instances be far too high. This happens when returns from a short-lived company (which may be high) are randomly assigned to a long-lived recipient (power of compounding)

In order toadjust for this, we need to break the returns into two subgroups: young companies (which are distributed toward higher returns) and old companies (lower returns).

```
#start by capturing the 87 documented returns from price populated set into two different lists
guessreturnold <- acquired$annual_return[which(acquired$age_at_acquisition >= mean(acquired$age_at_acqui
length(guessreturnold)
```

```
## [1] 36
```

```
guessreturnyoung <- acquired$annual_return[which(acquired$age_at_acquisition < mean(acquired$age_at_acqu
length(guessreturnyoung)
```

```
## [1] 53
```

```
#then we draw randomly (with replacement) from that vector to populate the 381 "acquired price unknown"
#this ensures that the distribution of returns between our "disclosed" and "undisclosed" groups are the
guess$inferred_return[which(guess$age >= mean(guess$age))] <- sample(guessreturnold, size = length(whic
guess$inferred_return[which(guess$age < mean(guess$age))] <- sample(guessreturnyoung, size = length(whi
#we then work backward from the assigned annual return number to calculate the implied price amount
guess$inferred_price <- guess$funding_total_usd * (exp((log(guess$inferred_return + 1)) * (as.integer(gu
```

Let's have a quick look to see if our resulting inferred prices look correct-ish. We want to see if the inferred_prices we've generated are reasonable relative to those that are known.

```
summary(acquired$price_amount)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 2.700e+05 1.500e+07 4.200e+07 1.012e+08 1.000e+08 1.010e+09
```

```
summary(guess$inferred_price)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 3.001e+04 2.628e+06 8.246e+06 2.225e+08 3.933e+07 3.245e+10
```

```
summary(acquired$annual_return)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.8669  0.5922  1.3210  3.1790  3.5880 46.9200
```

```
summary(guess$inferred_return)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.8669  0.5400  1.1760  3.3180  3.2860 46.9200
```

There are a few billion-dollar acquisitions in there, which seems improbable on the basis that a transaction of this size could simply not be done without disclosure. We'll scale all companies that are in the top quantile on valuation down by a factor (itself drawn randomly from a range of 5 - 10). Somewhat arbitrary but functionally fair considering the realities of the industry. We'll have to recalculate returns for those companies we've adjusted. Note that this means our distribution for guess no longer matches the distribution of returns for acquired but having started from the right place and made reasonable assumptions, this should be ok.

```
#first create a cut-point for the companies we're going to scale down
cutoff1 = quantile(guess$inferred_price)[3]
guess$inferred_price[which(guess$inferred_price > cutoff1)] <- guess$inferred_price[which(guess$inferre
#finally a check if there are any companies still above the $Bn threshold
guess$inferred_price[which(guess$inferred_price > 1000000000)] <- guess$inferred_price[which(guess$infe
guess$newreturn = guess$inferred_price/guess$funding_total_usd
guess$newreturnann = (guess$newreturn ^ (365.25/guess$age) - 1)
summary(acquired$price_amount)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 2.700e+05 1.500e+07 4.200e+07 1.012e+08 1.000e+08 1.010e+09
```

```
summary(guess$inferred_price)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##     30010   1437000   2868000  13040000   6058000 430300000
```

And let's also make sure that our returns (before, after) have not become too distorted.

```
summary(acquired$annual_return)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.8669  0.5922  1.3210  3.1790  3.5880 46.9200
```

```
summary(guess$inferred_return)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.8669  0.5400  1.1760  3.3180  3.2860 46.9200
```

Looks alright, so let's make these new returns official, and eliminate the columns we no longer need.

```
guess$inferred_return <- guess$newreturnann
guess$price_amount <- guess$inferred_price
guess$newreturn = NULL
guess$newreturnann = NULL
guess$inferred_price = NULL
```

REXAMINING RETURNS FOR THE GROUP

Let's recombine our acquired and guess sets into a single table of "exits". This will give us a better sense for overall portfolio returns.

```
#change status to differentiate where we've made assumptions
guess$status <- "undisclosed acquired"
acquired$status <- "disclosed acquired"
#delete some of the rows in acquired that we no longer need
acquired$return <- NULL
#rename two other columns to match
colnames(acquired)[19] <- "age"
colnames(guess)[20] <- "annual_return"
colnames(guess) == colnames(acquired)
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [15] TRUE TRUE TRUE TRUE TRUE TRUE
```

And now that the tables match, we merge them together. Since we've taken pains to ensure that the columns all match, we can use an rbind.

```
exits <- rbind(acquired, guess)
```

Now we can have a more realistic view of annual returns for the full set:

```
sum(exits$price_amount)/sum(fullset$funding_total_usd)
```

```
## [1] 0.3651263
```

So these companies as a whole, over the seven years noted, returned 31.7% of the capital that was invested in them. Keeping in mind that those that were funded later may not yet be sufficiently mature to have generated any return, we can look year-by-year.

For 2007, 2008 & 2009:

```
exits07 <- subset(exits, first_funding_at <= '2007/12/31')
funding07 <- subset(fullset, first_funding_at <= '2007/12/31')
sum(exits07$price_amount)/sum(funding07$funding_total_usd)
```

```
## [1] 0.5501693
```

```
exits08 <- subset(exits, first_funding_at <= '2008/12/31')
funding08 <- subset(fullset, first_funding_at <= '2008/12/31')
sum(exits08$price_amount)/sum(funding08$funding_total_usd)
```

```
## [1] 0.6561342
```

```
exits09 <- subset(exits, first_funding_at <= '2009/12/31')
funding09 <- subset(fullset, first_funding_at <= '2009/12/31')
sum(exits09$price_amount)/sum(funding09$funding_total_usd)
```

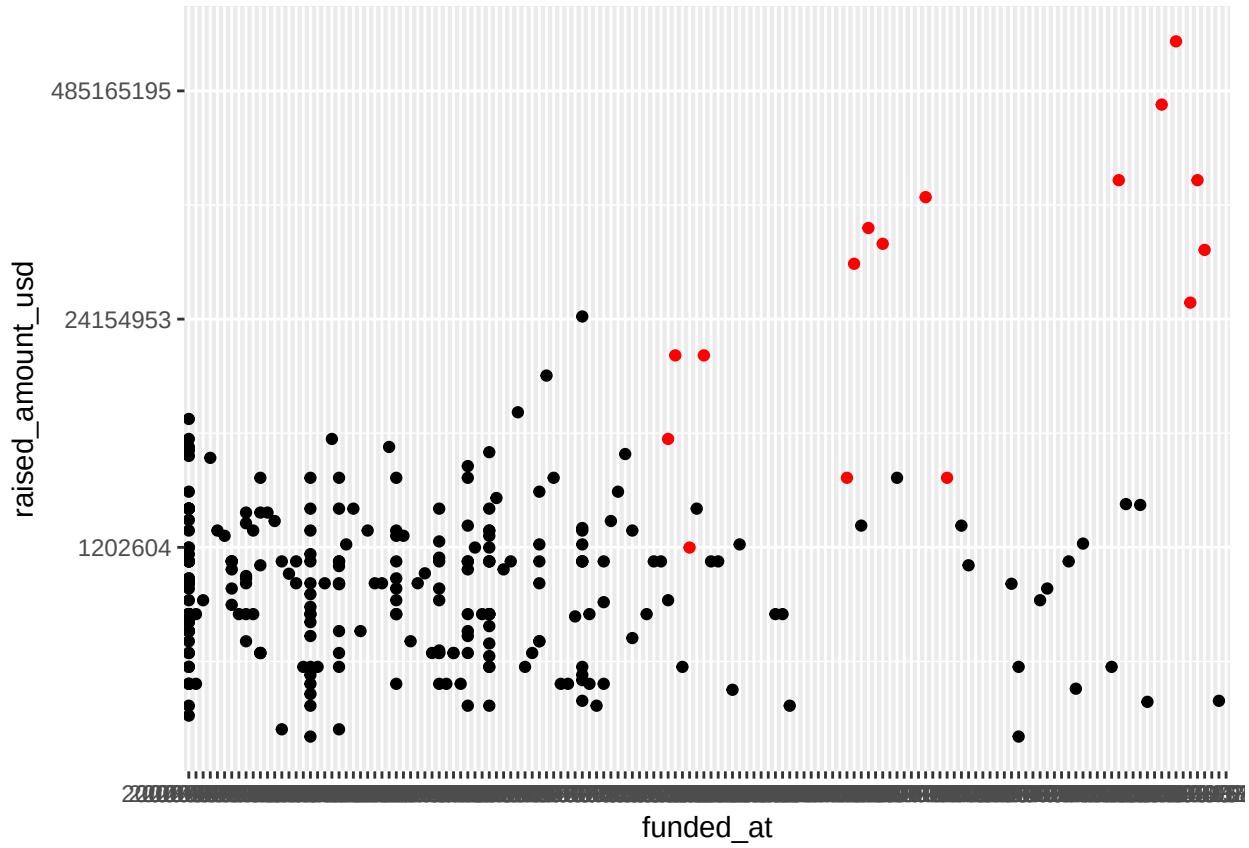
```
## [1] 0.4597542
```

This means that of the total amound invested in companies founded in 2007, 57% came back to investors via exits by 2014. Let's see if we can visualise this. On the x axis are our dates, from 2007 - 2014. The y axis is log(dollars). Black points show raises (investors put money into the company) and the red points are exits (investors recieve money back).

```
ggplot(funding07, aes(x=funded_at, y=raised_amount_usd)) + geom_point() + geom_point(aes (x = acquired_
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Removed 245 rows containing missing values (geom_point).
```



We can also create a dataset that considers each round of funding for our 2007 cohort of companies. We won't use this today except to create it from our list of those companies receiving initial funding in 2007/

```
names07 <- as.vector(funding07$name)
rounds07 <- subset(rounds3, company_name %in% names07)
```

The 30% figure (dollars returned/dollars invested) is concerning, as it represents an overall negative return on the capital dedicated to this space. Given the risk profile, investors would fairly expect significantly higher returns. Indeed, the most common figure for annual return expectation offered within the industry (Wiltbank & others) is 20% or more. Now, were that actually the case, the USD 4.9Bn invested in those companies first funded in 2007 would have grown to USD 17.5Bn realised returns by 2014, not the $2.8Bn actually reported. This is a significant gap, and suggests that those within industry who parrot the 20% number are way off.

```
sum(funding07$funding_total_usd)
```

```
## [1] 4903607980
```

```
sum(exits07$price_amount)
```

```
## [1] 2697814625
```

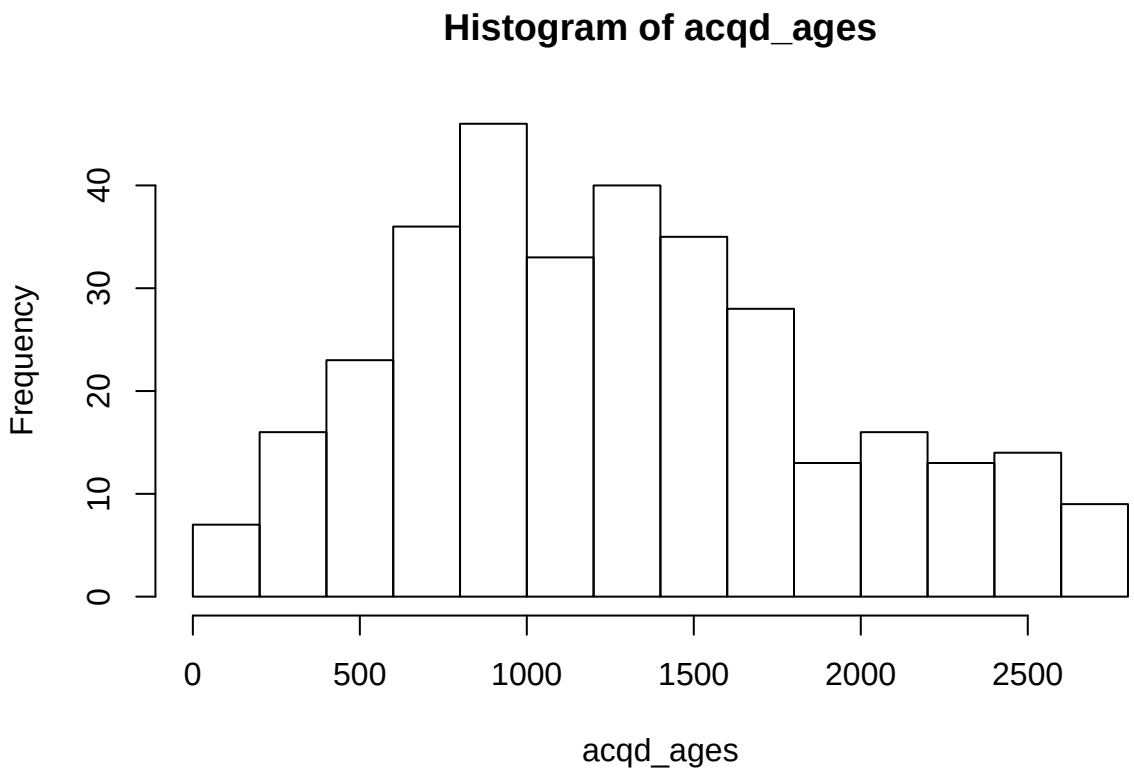
```
sum(funding07$funding_total_usd)*(1.2^7) - sum(exits07$price_amount)
```

```
## [1] 14872699340
```

ASSIGNING VALUE TO OPERATING COMPANIES

We can also look at the distribution of ages at which these companies were acquired:

```
acqd_ages <- c(as.vector(t(exits07$age)), as.vector(t(exits08$age)), as.vector(t(exits09$age)))
hist(acqd_ages)
```

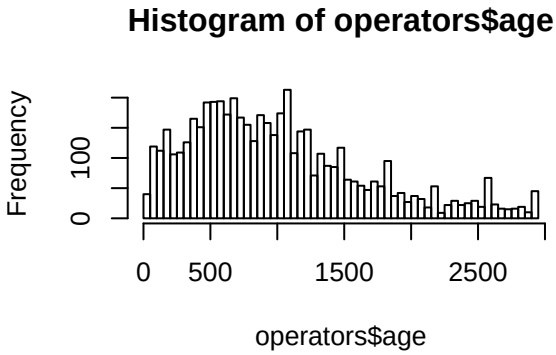


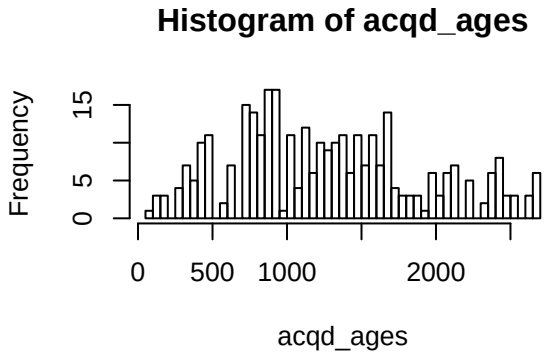

```
mean(acqd_ages)
```

```
## [1] 1269.69
```

And compare this set to those of our "still-operating companies". The idea here being that some of the companies that were listed as "operating" (and therefore without value in our returns analysis) may have gone on to be acquired (monetised) beyond our study horizon.

```
operators <- subset(fullset, status == "operating")
par(mfrow=c(2,2))
hist(operators$age, breaks = 50)
hist(acqd_ages, breaks = 50)
par(mfrow=c(1,1))
```





K-MEANS CLUSTERING

We can also use clustering to determine if there are any relationships in our data that we may have missed with our linear regressions and exploratory data analysis. We'll study exits specifically. First we'll size the dataset down to just the columns that are numeric. Then we'll use the NbClust package to plot a chart an see if there are any cluster counts that make more sense than the others.

```
library(NbClust)
```

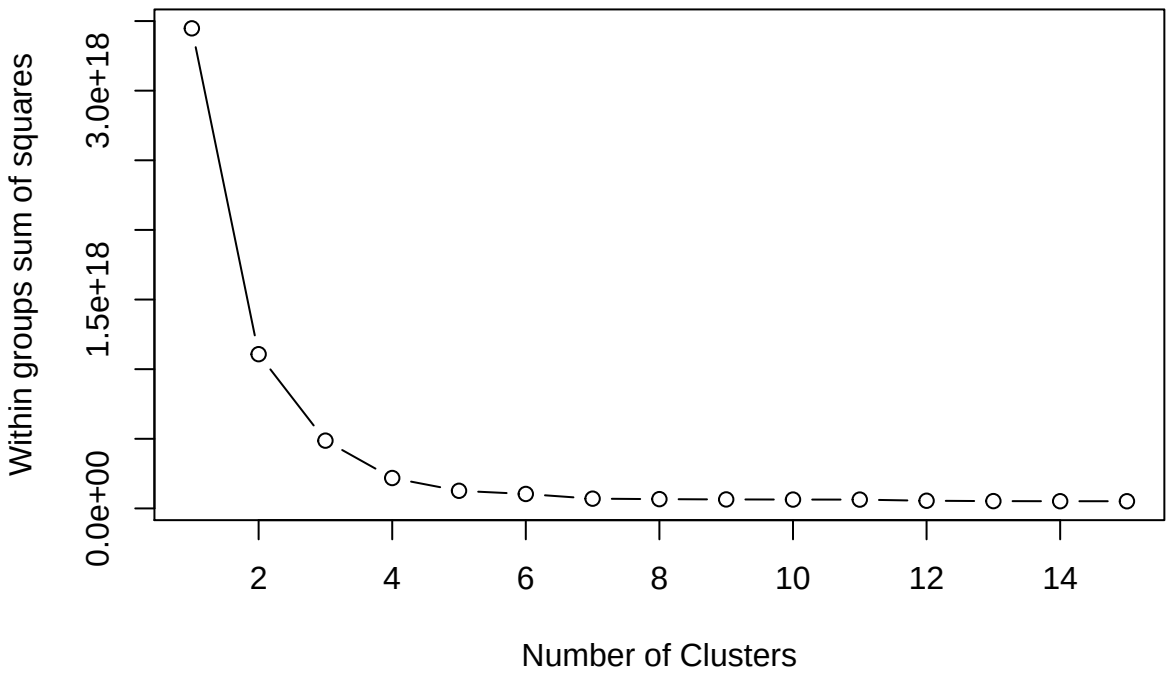
```
## Warning: package 'NbClust' was built under R version 3.3.2
```

```
smallexits <- subset(exits, !is.na(founded_year) & annual_return > 0, select = c(funding_total_usd, fou

wssplot <- function(data, nc=15, seed=1234){
                wss <- (nrow(data)-1)*sum(apply(data,2,var))
               for (i in 2:nc){
                set.seed(seed)
                wss[i] <- sum(kmeans(data, centers=i)$withinss)}

                  plot(1:nc, wss, type="b", xlab="Number of Clusters",
                       ylab="Within groups sum of squares")
       }

wssplot(smallexits)
```



They plot is pretty smooth so we'll choose 5 for our number of clusters.

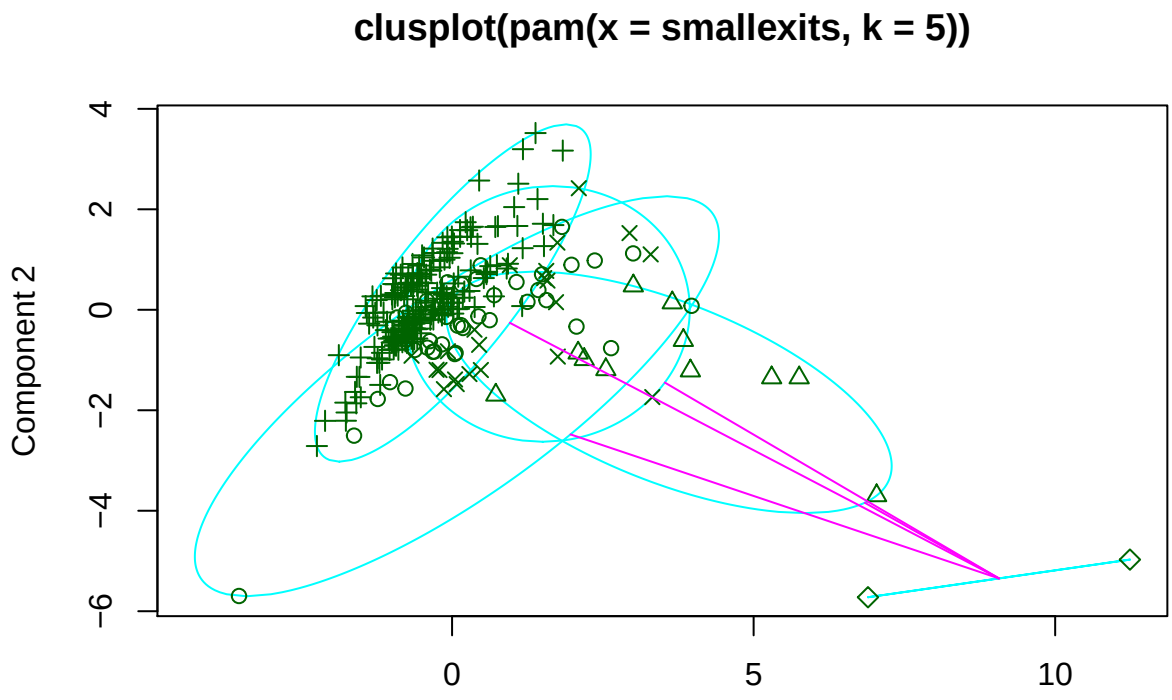
```
smallset <- subset(fullset, !is.na(founded_year), select = c(funding_total_usd, status, founded_year, m
savethis <- as.numeric(smallset$status)
smallset$status = NULL
invisible(scale(smallset))
smallfit <- kmeans(smallset, centers = 5 )
table(smallfit$cluster, savethis)
```

```
##    savethis
##        1    2    3    4
##   1    0    0    4    0
##   2    0    0    7    1
##   3  281   78 4411  372
##   4    0    0   33    3
##   5    3    3  278   21
```

```
library(cluster)
```

```
## Warning: package 'cluster' was built under R version 3.3.3
```

```
clusplot(pam(smallexits, 5))
```



Here you can see the effect of outliers on our clustering. Reading the columns left to right, nearly all of our companies listed as closed or unknown are grouped in Cluster 5. Clusters 1 and 4 are particularly small, containing just 4 and 8 observations respectively. I'd like to have a look at which rows in the set these outliers are, but don't quite know how to do this.

We can also see from this clustering that our "closed" companies appear in two clusters: 3 and 5. Also in Cluster 5 are 278 of our companies listed as "operating". Are these companies at greater risk for closure? If we had access to subsequent years data (say 2015 - 17) we might be able to test this theory (see: FUTURE WORK, below).

Let's use another approach for clustering. We'll use five groups again.

```
k = 5
KMC = kmeans(smallexits, centers = k, iter.max = 1000)
str(KMC)
```

```
## List of 9
##  $ cluster     : Named int [1:288] 5 5 3 2 5 2 2 4 2 5 ...
##   ..- attr(*, "names")= chr [1:288] "30" "45" "87" "104" ...
##  $ centers     : num [1:5, 1:6] 83200000 2095997 34628449 16715847 10741652 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:5] "1" "2" "3" "4" ...
##   .. ..$ : chr [1:6] "funding_total_usd" "founded_year" "price_amount" "metro" ...
##  $ totss       : num 3.45e+18
##  $ withinss    : num [1:5] 4.56e+15 2.36e+16 5.16e+16 2.39e+16 2.29e+16
##  $ tot.withinss: num 1.26e+17
##  $ betweenss   : num 3.32e+18
##  $ size        : int [1:5] 2 231 7 11 37
```

```
##  $ iter        : int 3
##  $ ifault      : int 0
##  - attr(*, "class")= chr "kmeans"
```

```
exitclusters = KMC$cluster
KMC$centers
```

```
##   funding_total_usd founded_year price_amount     metro       age
## 1          83200000     2008.000    970235000 1.0000000 1616.0000
## 2           2095997     2009.403      8207197 0.6147186  793.9264
## 3          34628449     2006.714    419120115 0.8571429 1912.4286
## 4          16715847     2009.000    196002338 0.4545455 1222.6364
## 5          10741652     2008.865     80478229 0.6216216 1180.6757
##   annual_return
## 1      1.648295
## 2      1.972675
## 3      1.036373
## 4      2.679870
## 5      3.669784
```

Cluster 2 contains the greatest number of observations at 203. This is the group that raised a very small amount of money relative to the others, but exited quickly and generated reasonable returns.

RESULTS

Basd on the information in this dataset, and using assumptions for our missing values that feel quite reasonable, we can fairly conclude that returns in this space are quite low. Rather than the 20% annual return espoused by industry organisations and some leaders, it appears that returns are significantly negative.

The data does suggest there are factors that are influential in more positive outcomes (thus increasing returns above the noted averages): companies in NY & the Bay Area are more likely to achieve exits. Companies that secure through their lifetimes billions of dollars in funding are also more likely to achieve an exit. Although moving into this domain of extreme levels of funding also acts as a decelerator on returns.

Our very best observed returns are generated by companies with only one or two rounds of funding for a total of a million or so dollars, are in the market for a year or two, and subsequently sold for USD10 to 70Mm.

Companies that register as closed generally did a single raise of just less than a million before going out of business. Sound familiar? Closed and acquireds have characteristics that read as similar in this dataset. Is the difference between them a matter of luck or just beyond the scope of the data we've gathered here?

FUTURE WORK

This area warrants further study. Additional data fields, particularly those that are numeric in nature, would allow us to test different relationships between outcomes and company characteristics. The industry is also notoriously opaque, and such data where it exists, is extremely expensive (Pitchbook charges USD20,000 for an annual subscription). Until data democratization arrives in this field, these may be the best conclusions we can draw.

Should Crunchbase release additional years of this limited data to the public, I'd love to test some of the hypotheses presented. Given additional time, I'd love also to populate valuation estimates around the operating companies, which is a good portion of the dataset.

For now, the main conclusion/takeaway is that investing in this space is a fool's errand. Clearly I need a new profession.