Springboard Capstone 3

Carlee Price

July 8, 2017

A STUDY ON THE ANGEL INVESTING LANDSCAPE AND OUTCOMES FOR INVESTORS

WHAT WE HAVE:

A Crunchbase data set that reports investments in seed-stage companies. This data is a sample, reflecting some portion of the total amount of activity in this space. It was published in 2014 and reflected activity in the space up to that point. We're going to select from the global, long-horizon data set those companies that received first funding in the US after 2007.

The goal here is to characterise for potential investors in this space, what the range of potential outcomes (returns on their invested capital) might be, and what proper portfolio construction looks like and means for those same returns.

PART I: WORKING WITH THE CRUNCHBASE DATA

a) LOAD & TRANSFORM

```
#read in the file that includes all the company information for this database
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
```

```
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 how may dollars are involved, it becomes difficult to draw conclusions. So we'll remove any companies for which the funding total shows N/A.

This is a sufficiently robust set to draw conclusions about the space.

Exploration #1): How often do investors see their money returned (company acquired) compared to the frequency with which they lose all their money (company closed)?

Transform the funding column into useable form (by stripping punctuation, converting to integer) and changing blanks in status column to read "unknown". Here is where we remove companies that lack information on funding totals. Let's see how these companies performed over the study horizon, how many remained operational.

```
companies$funding_total_usd <- as.numeric(gsub("[[:punct:]]", "", companies$funding_total_usd))
companies <- subset(companies, funding_total_usd > 1)
nrow(companies)
```

```
## [1] 21892
```

```
levels(companies$status)[1] <- "unknown"
table(companies$status)
```

```
##
##   unknown  acquired    closed operating
##       451      1632      1051     18758
```

This is our first look at the win/loss ratio in this space. Counting exits as a win (7.5%) and closeds as a loss (4.8%) 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] 1276
```

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

```
##
##   unknown  acquired    closed operating
##         7       326       168       775
```

Here we see the wins (25.5%) and losses (13.2%) are much higher; an even better ratio. It seems that spending more time in the market leads to more resolutions (closure or acquisition). Remember this data was captured in 2014.

#2. What about 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.

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.

```
acquisitions$price_amount[5344] <- 400000000
```

We're not able to screen these new tables for "first funding" date, as we did with our companies table. So instead we'll use a left join, which allows us to start with the screened companies table and add information from the new data sets only where it matches. lib

```
companies2 <- subset(companies, select = c(name, funding_total_usd, status, state_code, region, city, fu
#remove Aptalis Pharma
companies2 <- companies2[-c(1584), ]
acquisitions2 <- subset(acquisitions, select = c(company_name, company_state_code, company_region, compa
#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
joinedset <- merge(x = companies2, 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.

```
library(dplyr)
```

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

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

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

```
nrow(joinedset)
```

```
## [1] 21944
```

```
#subset(joinedset, name == "Catch.com")
#this one and 11 others can be eliminated with:
joinedset <- distinct(joinedset, name, first_funding_at, acquired_at, .keep_all = TRUE)
nrow(joinedset)
```

```
## [1] 21925
```

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).

```
nrow(joinedset)
```

```
## [1] 21925
```

```
#subset for where acquisition price information is disclosed
test <- subset(joinedset, price_amount > 1, funding_total_usd >1)
nrow(test)
```

```
## [1] 510
```

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

```
## [1] 8.399708
```

Only 510 of the 1632 companies described as "acquired" in our joined set include pricing information. Still a reasonable sample size, and the numbers are encouraging (8.40X return on total amount raised) but we wander into speculative territory if we fail to understand the biases in the sample (is it fair to say that acquisitions with price disclosed are generally more or less generous to investors?) extrapolate to conclusions about the population.

If we consider just those companies with disclosed acquisition prices, compared to the total amount raised across the data set, what does that tell us? So if all of the companies that were not acquired, or were but did not disclose the acquisition price, were ultimately worth $0 what would returns on this money be (annualized IRR)?

```
(sum(joinedset$funding_total_usd, na.rm = TRUE)/sum(joinedset$price_amount, na.rm = TRUE))^(1/7)-1
```

```
## [1] 0.1120767
```

And the same just for our 2007 group:

```
joinedset07 <- subset(joinedset, first_funding_at <= '2007/12/31')
(sum(joinedset07$funding_total_usd, na.rm = TRUE)/sum(joinedset07$price_amount, na.rm = TRUE))^(1/7)-1
```

```
## [1] 0.1037692
```

10 - 11% is not a very good return, considering the amount of risk investors in this space take on. But again, this is just the reported cash in hand return for companies and to assume that all others in the data set were worthless is not realistic.

In order to capture a true picture of what total returns are, we need to add several pieces to this:

1) companies that were acquired for an undicslosed amount
2) companies that were acquired outside of the study timeframe

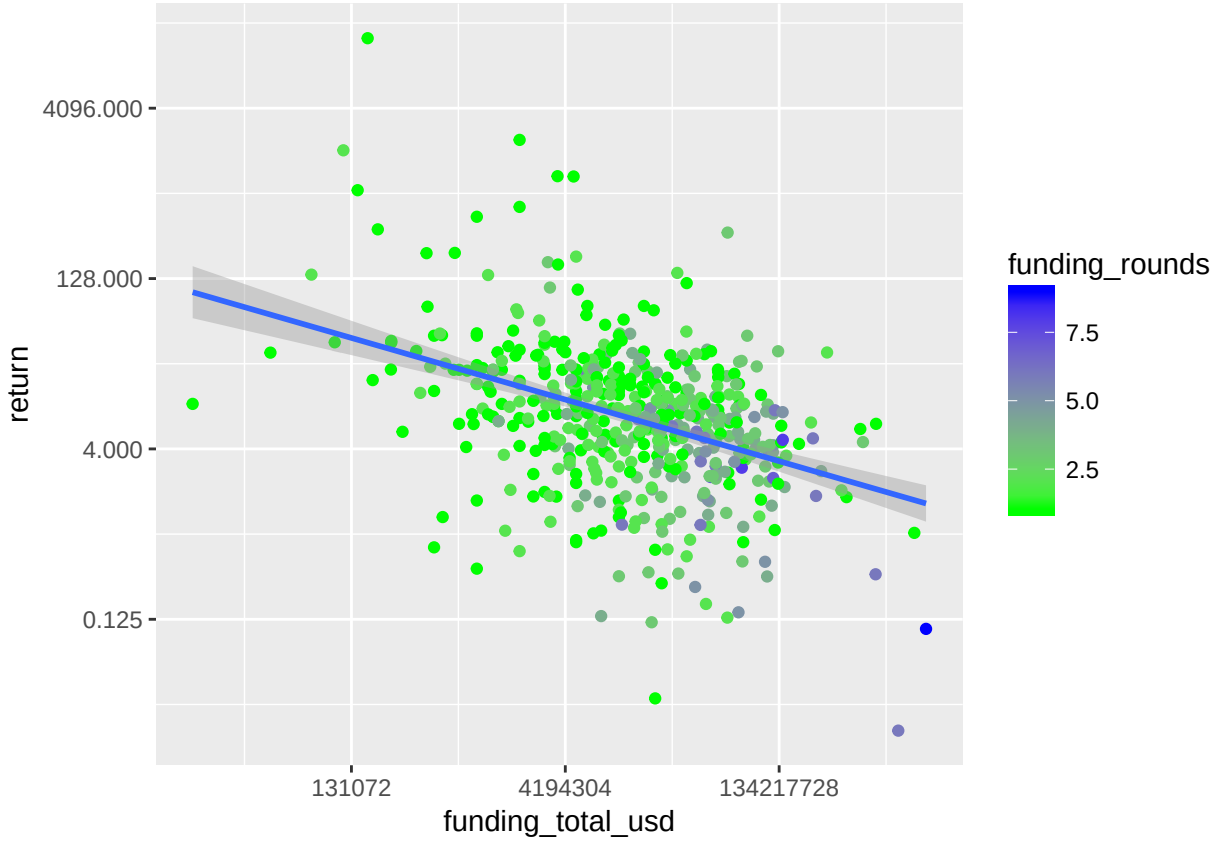
#3. Do companies that raise more money generate better returns for their stakeholders?

We can also use our returns column to draw some conclusions about returns relative to amounts raised. The companies we hear about in the press tend to be frequently in front of investors; they're cash-consumptive and generally percieved as fancy. But do these companies outperform their less capital-intensive peers? A regression analysis should give a clue.

```
library(ggplot2)
```

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

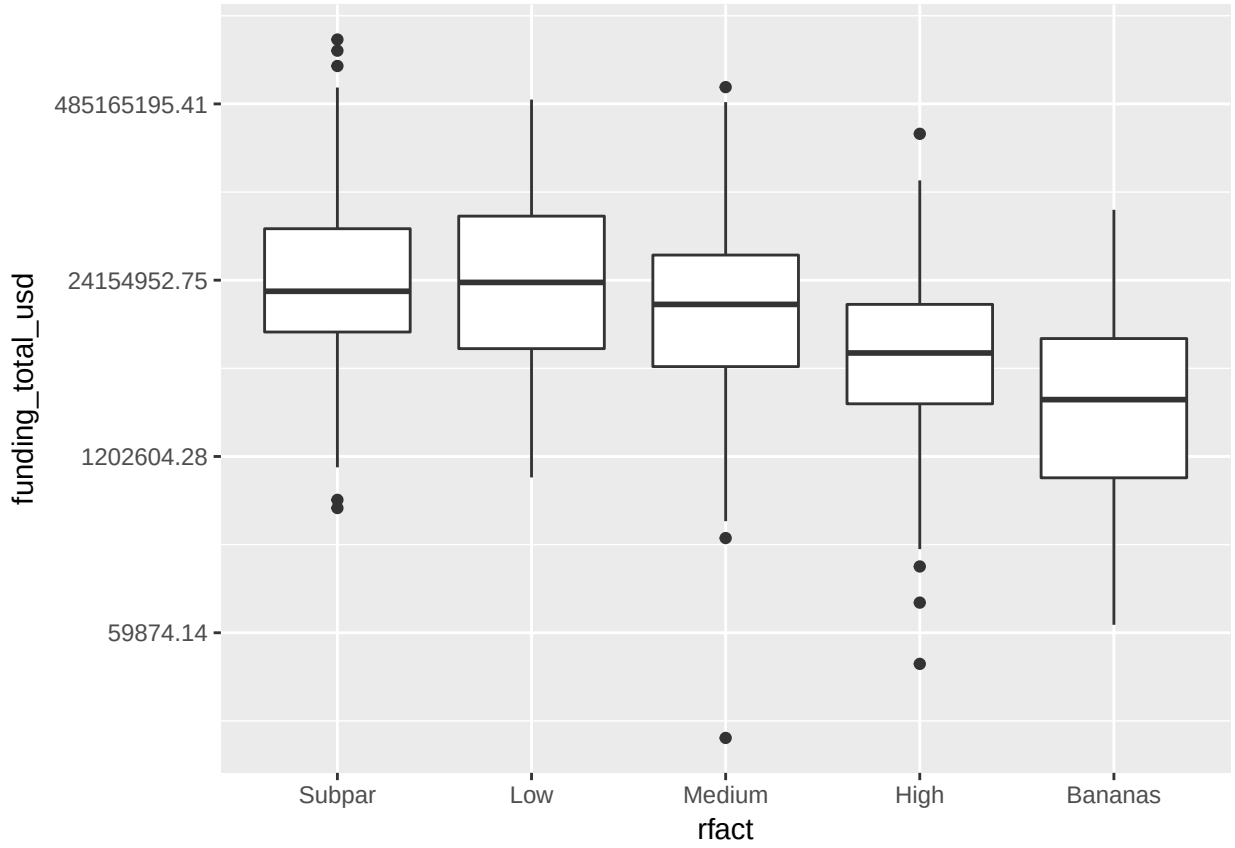
```
ggplot(test, aes(funding_total_usd, return)) +
 geom_point(aes(color = funding_rounds)) +
 scale_color_gradient(low = "green", high = "blue") +
 geom_smooth(method ="lm") +
 scale_y_continuous(trans = "log2") +
  scale_x_continuous(trans = "log2")
```



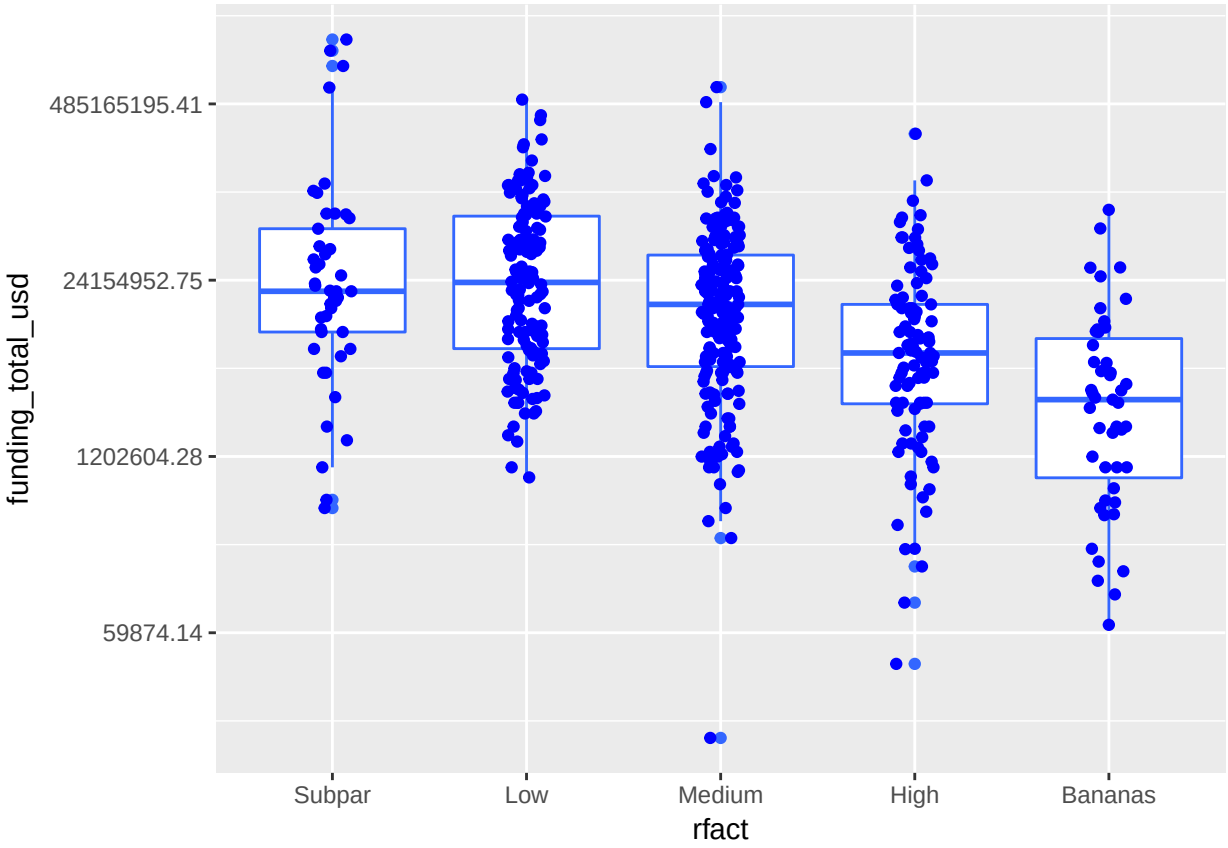
The slope here is negative, which tells us that companies which raise more tend to return less dollar-for-dollar. This may suprise some.

Next we factor these companies by returns, and look at the funding characteristics (amounts raised) by bucket.

```
#set breaks for each funding bucket, levels came about through trial & error
grp <- c(0, 1, 5, 15, 35, 20000)
#set factors & assign names
test$rfact = cut(test$return, breaks = grp,labels=c('Subpar','Low', 'Medium','High', 'Bananas'), ordere
#create a boxplot
#plot(x = test$rfact, y = test$funding_total_usd, log = 'y', ylab = "funds raised", xlab = "quality of
ggplot(test, aes(x=rfact, y=funding_total_usd)) + geom_boxplot() + scale_y_continuous(trans = "log")
```




```
ggplot(test, 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(test$funding_total_usd, na.rm = TRUE)
ftumn <- mean(test$funding_total_usd, na.rm = TRUE)
a <- subset(test, test$rfact == "Bananas")
amean <- mean(a$funding_total_usd)
(amean - ftumn)/(ftusd/(sqrt(nrow(a))))
```

```
## [1] -1.987669
```

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

```
## [1] 4.802753
```

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

We can get into SQL-like queries by using the dplyr pipe operator. Let's summarise some statistics by factor.

```
test %>%  group_by(rfact) %>%
          summarise(raised = mean(funding_total_usd), rounds = mean(funding_rounds), return = mean(retu
```

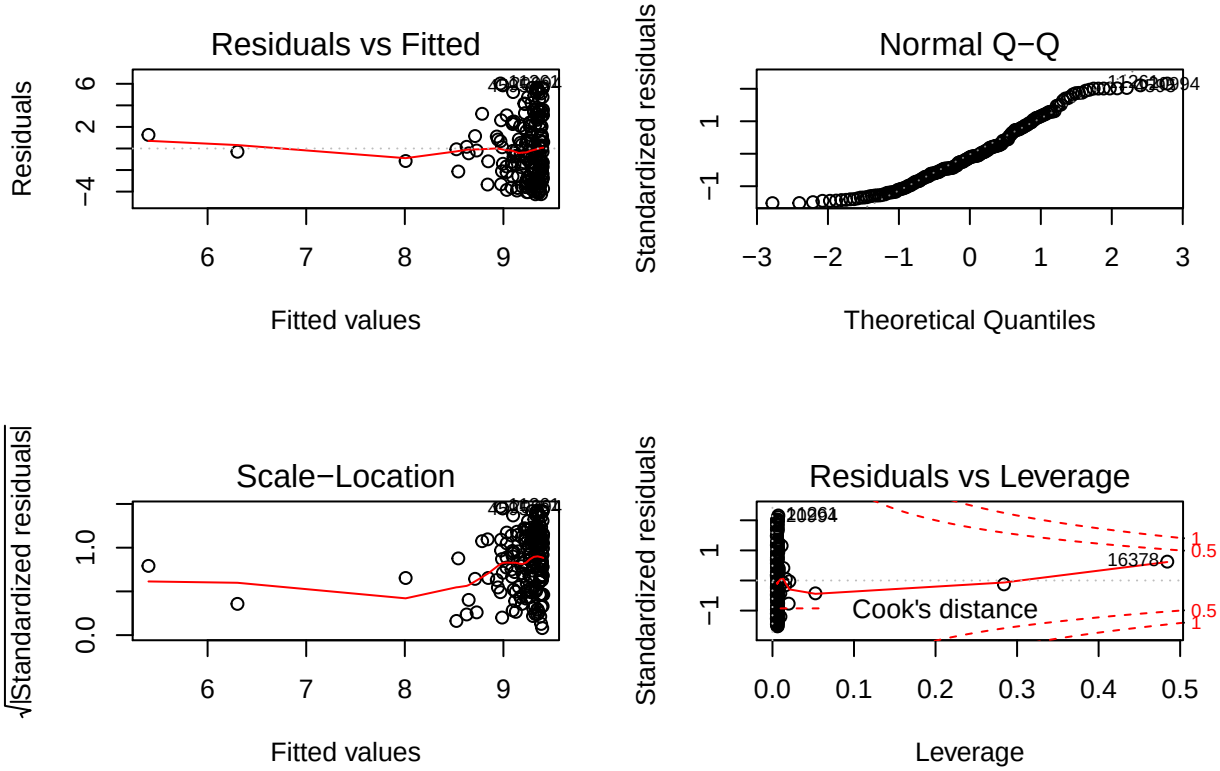
```
## # A tibble: 5 × 4
```

```
##     rfact    raised   rounds      return
##     <ord>     <dbl>    <dbl>       <dbl>
## 1  Subpar 120900430 2.711111   0.5396441
## 2     Low  53010228 2.698529   3.0496591
## 3  Medium  32406177 2.145946   9.2033200
## 4    High  18033085 1.762887  22.7315413
## 5 Bananas   8297109 1.510638 604.2179350
```

We can also have a look at how well these characteristics interact with each other.

Let's start by building a predictor model from the set of medium returns (5 - 15X):

```
medset <- subset(test, rfact == "Medium")
fit = lm(return ~ funding_total_usd, medset) # Run a regression analysis
par(mfrow=c(2,2)) # Change the panel layout to 2 x 2
plot(fit)
```




```
par(mfrow=c(1,1))
```