

Stat 133 - Post 01. Using VC / Startup data to further apply concepts from class including: Data Types, Vectors, Data Tables, and Frames

Mudit Goyal

10/23/2017

Introduction

Nowadays, startup investing is at an all time high. Check out this [article \(Ref 1\)](#) for further proof. Hundreds of companies **everyday** are being invested in according to data from popular sources such as **Pitchbook** and **Crunchbase**. However, is there a way to predict the success of these companies? What even defines success of these companies? There is almost unlimited data about the backgrounds of both successful and unsuccessful company founders. Many VC firms have tried (none have succeeded) to increase their batting averages in picking successful founders. I think there may be a way to create a model to predict success. For this case, we'll just look at the output data (including valuation at Seed Stage, Series A, and Series B) to illustrate a few of the concepts in class (Data Types, Vectors, Tabular Data, and exploring Tables). This is the background for this post whose intention is to show how techniques of computational data analysis can be applied.

In this assignment we're illustrating how you can work with an actual real world data frame and get meaningful insights even after some simple manipulation. We will read the data, understand how data frames work, use both bracket and the dollar operator, and learn a little bit about some of the startup ecosystem!

The Data

Motivation

Imagine being a venture capitalist - wouldn't it be useful to have a tool which can help you do a good amount of the due diligence on a company you'd normally have to do manually? This is something highly sought after from VC's. I've actually reached out to quite a few of them, and they've all said wonderful things, including a few that are willing to put in some capital in my project. We're illustrating some aspects of computational data analysis, mainly on **data frames**.

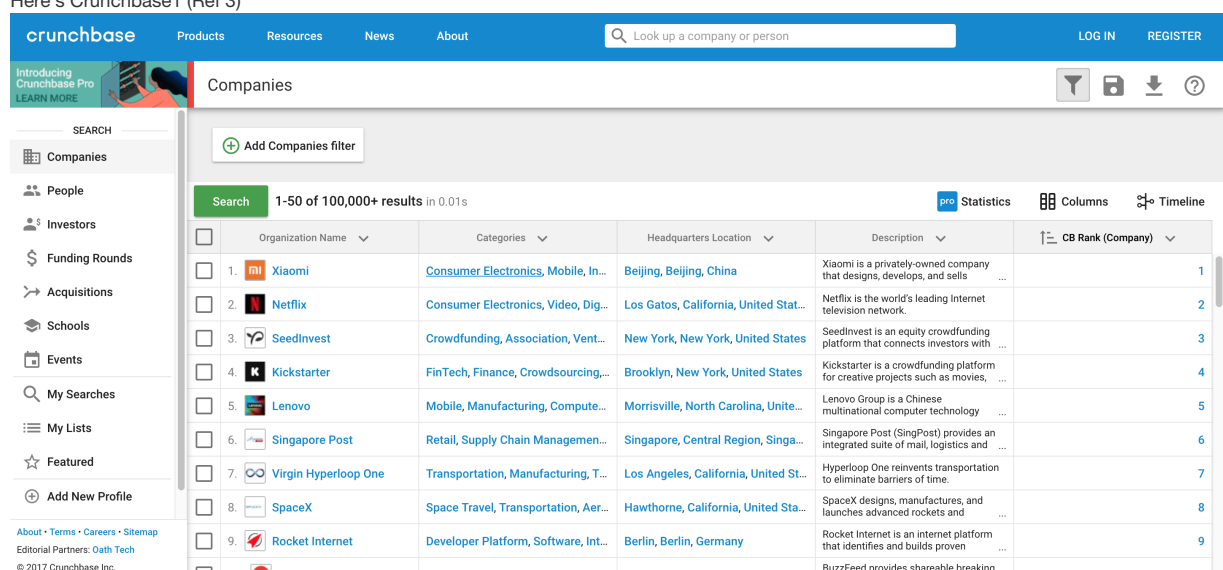
Background

For the purpose of this post, we're just working with the output data (CEO Name, Company, Seed Valuation, Series A Valuation, and Series B Valuation). Normally, we'd work with a huge amount of features which would predict these valuations. These include (but most definitely not limited to!): Education, Major in College, Hometown, Average Income, of hometown, Siblings...etc. As you can see, some data points are a lot more obscure than others. However, it's sometimes the more obscure data points which give you the most information. For example, please read [this \(Ref 2\)](#) article. The key takeaway that Paul Arnold, Venture Capitalist, had was that the most successful founders usually came from McKinsey. There is most definitely a lot of data and a lot of good output that can come from these data. In addition, some of the material used came from "Tidy Data", by Hadley Wickham (Ref 6), and also the **Introduction to dplyr** source. (Ref 7).

Sources

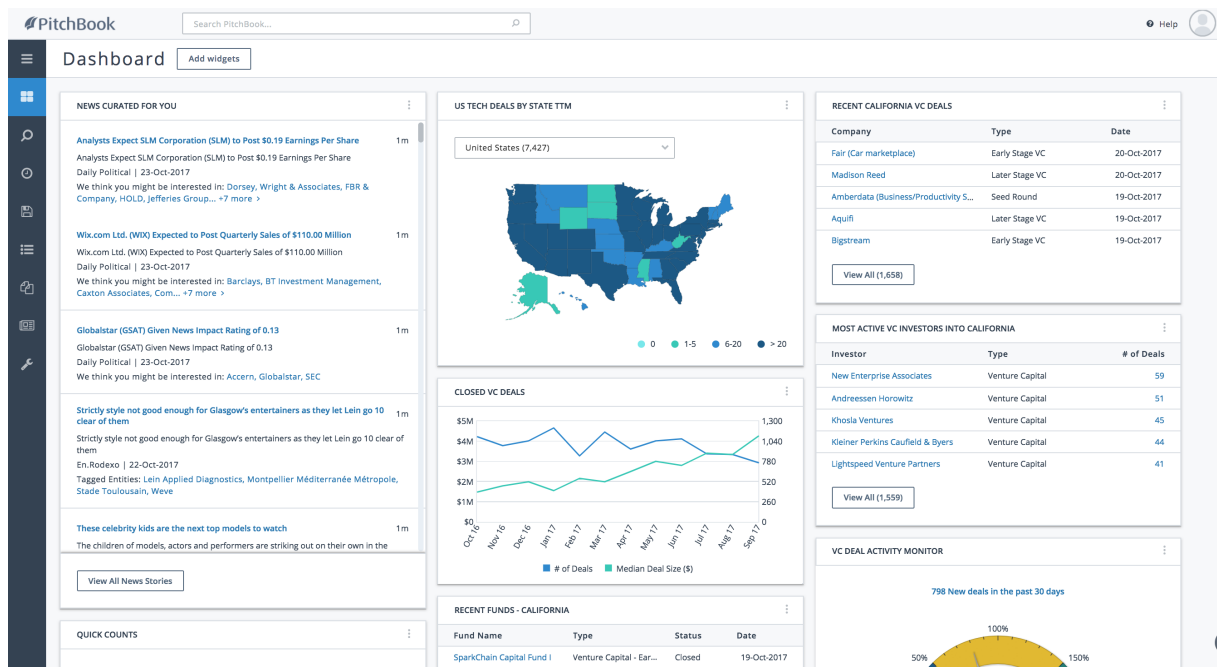
My data was taken primarily from [Crunchbase](#) and [Pitchbook](#). Unfortunately, you will not be able to access these without a paid membership, so I have attached screenshots here of what their database looks like.

Here's Crunchbase1 (Ref 3)



Organization Name	Categories	Headquarters Location	Description	CB Rank (Company)
1. Xiaomi	Consumer Electronics, Mobile, In...	Beijing, Beijing, China	Xiaomi is a privately-owned company that designs, develops, and sells ...	1
2. Netflix	Consumer Electronics, Video, Dig...	Los Gatos, California, United Stat...	Netflix is the world's leading Internet television network.	2
3. SeedInvest	Crowdfunding, Association, Vent...	New York, New York, United States	Seedinvest is an equity crowdfunding platform that connects investors with ...	3
4. Kickstarter	FinTech, Finance, Crowdsourcing...	Brooklyn, New York, United States	Kickstarter is a crowdfunding platform for creative projects such as movies, ...	4
5. Lenovo	Mobile, Manufacturing, Compute...	Morrisville, North Carolina, Unite...	Lenovo Group is a Chinese multinational computer technology ...	5
6. Singapore Post	Retail, Supply Chain Managemen...	Singapore, Central Region, Singa...	Singapore Post (SingPost) provides an integrated suite of mail, logistics and ...	6
7. Virgin Hyperloop One	Transportation, Manufacturing, T...	Los Angeles, California, United St...	Hyperloop One reinvents transportation to eliminate barriers of time.	7
8. SpaceX	Space Travel, Transportation, Aer...	Hawthorne, California, United Sta...	SpaceX designs, manufactures, and launches advanced rockets and ...	8
9. Rocket Internet	Developer Platform, Software, Int...	Berlin, Berlin, Germany	Rocket Internet is an internet platform that identifies and builds proven ...	9
10. BuzzFeed	News, Video, Social, Digital, Ent...	New York, New York, United States	BuzzFeed provides shareable breaking ...	10

Here's Pitchbook! (Ref 4)



As you can see, there's a ton of good data available. Let's start with my data set.

```
# Loading the required libraries here. In order to use the 'read_csv' method, you MUST
# load 'readr' otherwise it will not work. By default, R will load each column with type
# 'col_character()', however, we don't want that in this case. Since we have three
# columns with numerical values, we want to load the 'Seed Valuation', 'A Valuation',
# and 'B Valuation' as type 'col_number()'. This opens up more capabilities for us to do some more analysis.
```

```
library(readr)
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
```

```
# This is Ref 5.
vc = read_csv("/Users/Mudit/Desktop/stat-133/stat133-hws-fall17/post01/post01/data/output_data.csv",
              col_types = cols(
                'CEO Name' = col_character(),
                'Primary Company' = col_character(),
                'Seed Valuation' = col_number(),
                'A Valuation' = col_number(),
                'B Valuation' = col_number())
# Displays the data frame.
vc
```

```
## # A tibble: 170 x 5
##   `CEO Name` `Primary Company` `Seed Valuation` `A Valuation`
##   <chr>      <chr>              <dbl>         <dbl>
## 1 Tony Fadell Nest Labs           2700000       49210000
## 2 Kevin Systrom Instagram          2600000       32000000
## 3 Brian Wong Kiip              590000        11170000
## 4 Payal Kadakia ClassPass          15580000      80000000
## 5 Tom Chavez Krux              9590000       28920000
## 6 Tina Sharkey Brandless          14900000      35500000
## 7 Emmett Shear Twitch             1050000       24190000
## 8 Don Ressler Fabletics          68000000     210000000
## 9 Pau Sabria Olapic             5480000       22290000
## 10 Jack Conte Patreon             7980000       65500000
## # ... with 160 more rows, and 1 more variables: `B Valuation` <dbl>
```

As you can see, we have 5 columns, and 170 rows. Each column vector is **atomic** meaning that they are of the same type. The three right most rows are all of type double. We can check this by using the **typeof()** function. In addition, we can also find the length using the **length()** function.

In R, numeric vectors are of type **double** by default. Doubles are approximations.

The first two columns are of type **character** the most complex type of atomic vector.

```
# We use dollar sign notation here in order to access the column in the table 'vc'.
typeof(vc$`Seed Valuation`)
```

```
## [1] "double"
```

```
typeof(vc$`A Valuation`)
```

```
## [1] "double"
```

```
typeof(vc$`B Valuation`)
```

```
## [1] "double"
```

```
# Using the length function
length(vc)
```

```
## [1] 5
```

Now that we have our table and we've checked that the types of each column check out, we can begin our actual analysis.

Let's first pre-process the data a bit. Cleaning the data (as shown in Ref 6) is probably the most important part of actual data science, and can actually be the most time consuming as well.

Luckily our data is quite clean; however, the numbers are a bit gaudy to look at. Let's convert them into millions of dollars. Also, let's make sure that there are no empty cells anywhere in the data (there is one "NA" in the Seed Valuation Column) You can do that by using the is.na function.

```
vc$`Seed Valuation` = vc$`Seed Valuation`/1000000
vc$`A Valuation` = vc$`A Valuation`/1000000
vc$`B Valuation` = vc$`B Valuation`/1000000
# Removing the NA and replacing with zero.
vc[is.na(vc)] <- 0
vc
```

```
## # A tibble: 170 x 5
##   `CEO Name` `Primary Company` `Seed Valuation` `A Valuation`
##   <chr>      <chr>              <dbl>          <dbl>
## 1 Tony Fadell Nest Labs             2.70           49.21
## 2 Kevin Systrom Instagram             2.60           32.00
## 3 Brian Wong   Kiip                 0.59           11.17
## 4 Payal Kadakia ClassPass            15.58           80.00
## 5 Tom Chavez   Krux                 9.59           28.92
## 6 Tina Sharkey Brandless            14.90           35.50
## 7 Emmett Shear Twitch              1.05           24.19
## 8 Don Ressler  Fabletics            68.00          210.00
## 9 Pau Sabria   Olapic               5.48           22.29
## 10 Jack Conte  Patreon              7.98           65.50
## # ... with 160 more rows, and 1 more variables: `B Valuation` <dbl>
```

Let's now calculate the differences in valuation between Seed and A, and between A and B. We can do this using the dollar sign notation to refer to the columns. Note that positive numbers indicate that their valuation grew, whereas negative means that their company dropped.

```
vc$`Seed and A Difference` = vc$`A Valuation` - vc$`Seed Valuation`
vc
```

```
## # A tibble: 170 x 6
##   `CEO Name` `Primary Company` `Seed Valuation` `A Valuation`
##   <chr>      <chr>              <dbl>          <dbl>
## 1 Tony Fadell Nest Labs             2.70           49.21
## 2 Kevin Systrom Instagram             2.60           32.00
## 3 Brian Wong   Kiip                 0.59           11.17
## 4 Payal Kadakia ClassPass            15.58           80.00
## 5 Tom Chavez   Krux                 9.59           28.92
## 6 Tina Sharkey Brandless            14.90           35.50
## 7 Emmett Shear Twitch              1.05           24.19
## 8 Don Ressler  Fabletics            68.00          210.00
## 9 Pau Sabria   Olapic               5.48           22.29
## 10 Jack Conte  Patreon              7.98           65.50
## # ... with 160 more rows, and 2 more variables: `B Valuation` <dbl>, `Seed
## #   and A Difference` <dbl>
```

Let's sort them. We will work with the seed and a difference first.

```
arrange(vc, vc$`Seed and A Difference`)
```

```
## # A tibble: 170 x 6
##   `CEO Name` `Primary Company` `Seed Valuation` `A Valuation`
##   <chr>      <chr>             <dbl>         <dbl>
## 1 Brett Galloway      Xova Labs             8.00          0.00
## 2 John Payne          CircleUp              7.32          0.00
## 3 Gaurav Munjal       Flat.to              2.50          0.00
## 4 Brandon Rodman      Weave                0.63          0.00
## 5 Rob Salvatore       Tongal               2.22          2.47
## 6 Andy Ory            Acme Packet          9.83         12.00
## 7 Mike Farley         Tile                51.31         54.31
## 8 Oliver Roup         VigLink              2.00          5.00
## 9 Gene Wang           People Power         4.42          8.00
## 10 David Vivero        Amino               3.50          7.38
## # ... with 160 more rows, and 2 more variables: `B Valuation` <dbl>, `Seed`
## #   and A Difference` <dbl>
```

To interpret, we can see that the first few data points are negative, meaning that those companies were not succesful. Further research shows that we are correct in concluding that.

Let's now add the difference between A and B

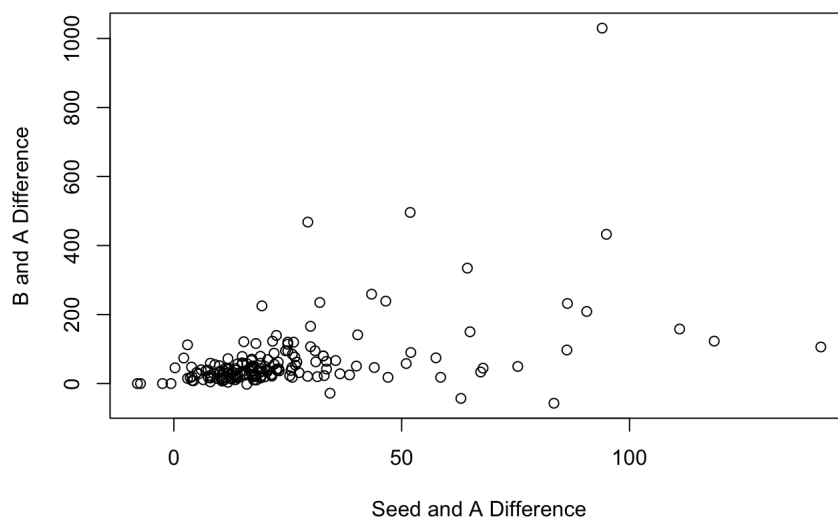
```
vc$`B and A Difference` = vc$`B Valuation` - vc$`A Valuation`
arrange(vc, desc(vc$`B and A Difference`))
```

```
## # A tibble: 170 x 7
##   `CEO Name` `Primary Company` `Seed Valuation` `A Valuation`
##   <chr>      <chr>             <dbl>         <dbl>
## 1 Louay Eldada  Quanergy Systems     16.00        110.00
## 2 Philip Krim   Casper Sleep         7.22         59.09
## 3 Kevin Systrom Instagram           2.60         32.00
## 4 Joshua Reeves Gusto              32.59        127.53
## 5 Payal Kadakia ClassPass           15.58         80.00
## 6 Michael Buckwald Leap Motion          7.60         51.00
## 7 Tony Fadell   Nest Labs            2.70         49.21
## 8 Kevin Gibbon  Shyp                 7.97         40.00
## 9 Max Ventilla  AltSchool            6.60         92.98
## 10 Tom Chavez   Krux                 9.59         28.92
## # ... with 160 more rows, and 3 more variables: `B Valuation` <dbl>, `Seed`
## #   and A Difference` <dbl>, `B and A Difference` <dbl>
```

Great. Now that we have a difference, let's see if the differences between Seed and Series A and the differences between Series A and B are correlated.

```
plot(vc$`Seed and A Difference`, vc$`B and A Difference`, xlab = "Seed and A Difference",
     ylab = "B and A Difference", main = "B and A Difference v. Seed and A Difference")
```

B and A Difference v. Seed and A Difference



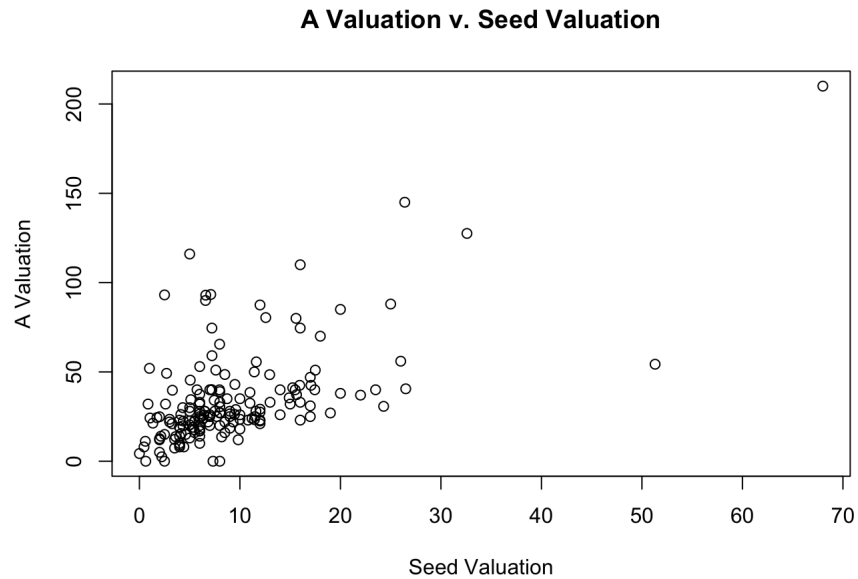
```
cor(vc$`Seed and A Difference`, vc$`B and A Difference`, use = "complete.obs")
```

```
## [1] 0.4560717
```

There is a correlation between the two of them, though it's not strong, with an R value of .45.

Now we do Seed Valuation and the A Valuation. We will also use the cor() function to find the R value.

```
plot(vc$`Seed Valuation`, vc$`A Valuation`, xlab = "Seed Valuation", ylab = "A Valuation",
     main = "A Valuation v. Seed Valuation")
```



```
cor(vc$`Seed Valuation`, vc$`A Valuation`, use = "complete.obs")
```

```
## [1] 0.610612
```

You can tell that there indeed is a moderately strong, linear correlation between the Seed Valuation and the Valuation at Series A concluded because we have an R value of about .6.

Finally we want to calculate the total difference from Series B and Seed to draw conclusions.

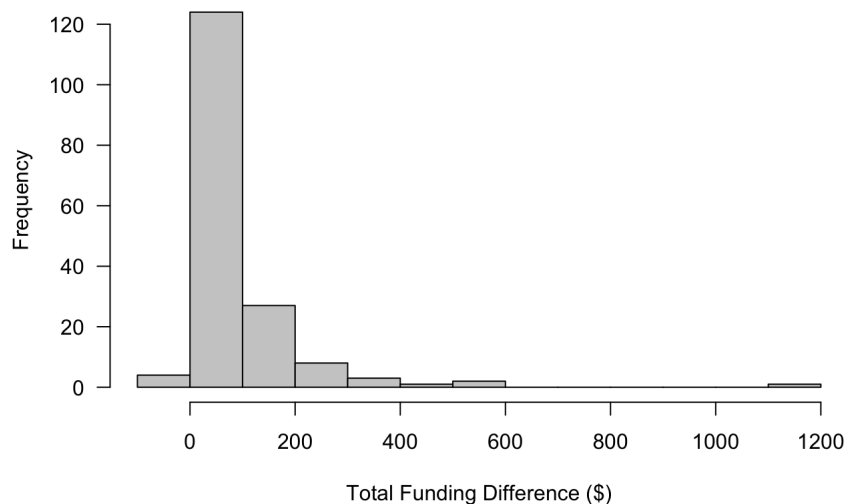
```
vc$`Total` <- vc$`B Valuation`-vc$`Seed Valuation`
arrange(vc, desc(vc$`Total`))
```

```
## # A tibble: 170 x 8
##       `CEO Name` `Primary Company` `Seed Valuation` `A Valuation`
##       <chr>      <chr>              <dbl>         <dbl>
## 1 Louay Eldada  Quanergy Systems      16.00         110.00
## 2 Philip Krim   Casper Sleep           7.22          59.09
## 3 Joshua Reeves Gusto                32.59         127.53
## 4 Kevin Systrom Instagram             2.60          32.00
## 5 Payal Kadakia ClassPass             15.58          80.00
## 6 Max Ventilla  AltSchool             6.60          92.98
## 7 Michael Buckwald Leap Motion           7.60          51.00
## 8 Amir Husain  SparkCognition         2.50          93.12
## 9 Tony Fadell   Nest Labs              2.70          49.21
## 10 Maxim Lobovsky Formlabs              5.00         116.00
## # ... with 160 more rows, and 4 more variables: `B Valuation` <dbl>, `Seed
## #   and A Difference` <dbl>, `B and A Difference` <dbl>, Total <dbl>
```

We're going to work primarily with the total column now.

```
hist(vc$Total, las = 1, col = 'gray80', xlab = 'Total Funding Difference ($)', main = 'Histogram of Total Funding
Differences')
```

Histogram of Total Funding Differences



This is very interesting - it's cool to

see that most of the differences are right in between 0 - 200 Million. Let's see the top 10 companies based on the funding differences. We can do it like displayed below.

```
top10 <- c('Primary Company', 'Total')
vc[order(vc$`Total`, decreasing = TRUE)[1:10], top10]
```

```
## # A tibble: 10 x 2
##   `Primary Company`   Total
##   <chr>             <dbl>
## 1 Quanergy Systems 1124.00
## 2 Casper Sleep      547.78
## 3 Gusto             527.41
## 4 Instagram         497.40
## 5 ClassPass         398.92
## 6 AltSchool         318.40
## 7 Leap Motion       302.40
## 8 SparkCognition    299.55
## 9 Nest Labs         285.42
## 10 Formlabs         269.00
```

Let's now see if there are any companies which have a negative amount for total

```
vc$`Primary Company`[vc$Total < 0]
```

```
## [1] "CircleUp" "Flat.to" "Weave" "Xova Labs"
```

We can conclude that these four companies were complete failures. Let's now compute the correlation coefficients between Total and all the variables used in the formula to compute it (B Valuation and Seed)

```
tot <- c('A Valuation', 'B Valuation', 'Seed Valuation')
tot_corrs <- cor(vc[,tot], vc$Total)
tot_corrs <- tot_corrs[order(tot_corrs, decreasing = TRUE), ]
tot_corrs
```

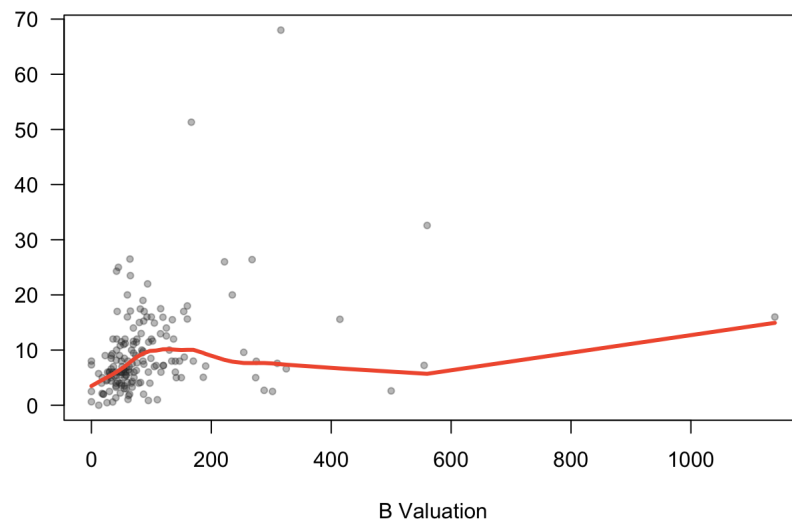
```
##   B Valuation   A Valuation Seed Valuation
##   0.9978714    0.5800205    0.2188862
```

It correlates with the B Valuation the most, but that's because that's the largest value compared to the total.

Let's now create a scatter plot between B and Seed Valuation, also including a lowess line.

```
plot(vc$`B Valuation`, vc$`Seed Valuation`, pch = 20, col = "#33333355", las = 1,
     xlab = 'B Valuation', ylab = '',
     main = 'B and Seed Valuations (with lowess line)')
lines(lowess(vc$`B Valuation`, vc$`Seed Valuation`), lwd = 3, col = "#F15C3C")
```

B and Seed Valuations (with lowess line)



correlation coefficient between them.

```
cor(vc$`B Valuation`, vc$`Seed Valuation`)
```

```
## [1] 0.282051
```

An R value of .28 is telling us that there is a very weak, positive, linear correlation between the B Valuation and the valuation and valuation at the Seed Stage. This proves to us that if we were to run a linear regression, we would not be able to accurately predict the potential valuation at Series B.

Concluding Remarks

This simple random sample of startup valuation can tell us a lot. After speaking to VC's for the past few years, it's valid to say that the valuation differences are a valid metric to look at when evaluating if a company is successful or not. I unfortunately had to take a simple random sample as not everyone is authorized to view the data that I have, so I took a simple random sample.

It is very interesting to see that even if you were to raise a large seed round or even a large round of funding for your Series A, that doesn't mean that your startup will be successful at the end. The four startups that we found here who completely failed all did raise large rounds of funding, which is a common theme, sadly, with other failed companies.

The purpose of this assignment was to illustrate how you can apply some of the stuff we have learned in Stats 133 to actual real life data. In this analysis, we worked with the 'vc' dataframe, and worked with manipulating that data in R such as (reading the table, understanding data frames, using the bracket notation, dollar operator, and more.)