Exploratory Data Analysis

K Iwasaki

September 15, 2017

Contents

Set-up

Most of columns are categorical variables.

```
drops = c("title", "link", "excerpt", "Company", "money_raised", "linkedin_link",
          "Company_at_Linkedin", "Specialties", "Website", "Location",
          "zip_code","State",
          "Description", "Also.viewed", "Industry")
data = data[, !(names(data) %in% drops) ]

summary(data[, 1:13 ])
```

```
##      published_at  funding_round money_raised_float
##   1/14/2016 :  3            : 12   Min.   : 10.00
##   7/29/2015 :  3   Series B: 48    1st Qu.: 15.00
##   11/15/2016:  2   series C:  1    Median : 25.00
##   11/3/2015 :  2   Series C:152    Mean   : 41.17
##   12/11/2013:  2   Series D: 14    3rd Qu.: 45.00
##   2/11/2008 :  2   Series E:  5    Max.   :793.50
##   (Other)   :218
##                                        CompanyName       CompanySize
##   2U                                          :  1   51-200    :101
##   3D Robotics                                 :  1   201-500   : 64
##   aCommerce - Ecommerce Solutions for Southeast Asia:  1   Nov-50    : 27
##   Affle                                       :  1   1001-5000: 21
##   App Annie                                   :  1   501-1000 : 14
##   Appear Here                                 :  1   10,001+  :  2
##   (Other)                                     :226   (Other)  :  3
##      Founded            City      address_check         Country
##   Min.   :1939   San Francisco:49   False: 59   United States :178
##   1st Qu.:2007                :46   True :173   United Kingdom: 17
##   Median :2010   New York     :27               Germany       :  8
##   Mean   :2009   Mountain View:11               Canada        :  4
##   3rd Qu.:2012   San Mateo    : 8               India         :  4
##   Max.   :2017   Boston       : 6               Singapore     :  4
##                  (Other)      :85               (Other)       : 17
```

```
##      latitude          longitude
##  Min.    :25.78    Min.    :-122.67
##  1st Qu.:37.44    1st Qu.:-122.39
##  Median :37.78    Median :-121.95
##  Mean    :38.47    Mean    :-103.87
##  3rd Qu.:40.74    3rd Qu.: -77.28
##  Max.    :47.62    Max.    : -71.04
##  NA's    :46       NA's    :46
##                                Industry_consolidated
##  Internet                                 :68
##  Computer Software                        :48
##  Information Technology and Services:25
##  Financial Services                       :15
##  Consumers Goods & Services               :14
##  Infrastructure                           :10
##  (Other)                                  :52
##  spc_Logistics.and.Supply.Chain
##  Min.    :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean    :0.1034
##  3rd Qu.:0.0000
##  Max.    :1.0000
##
```

```
# check columns 1:13. Columns 13: have same format.
str(data[, 1:13 ])
```

```
## 'data.frame':    232 obs. of  13 variables:
##  $ published_at              : Factor w/ 209 levels "1/11/2010","1/14/2016",..: 189 176 160 158 
##  $ funding_round             : Factor w/ 6 levels "","Series B",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ money_raised_float        : num  45 39 40 48 90 20.2 29 32 36 20 ...
##  $ CompanyName               : Factor w/ 232 levels "2U","3D Robotics",..: 25 29 185 126 39 127 
##  $ CompanySize               : Factor w/ 8 levels "10-Jan","10,001+",..: 4 7 7 7 4 4 7 7 7 7 ...
##  $ Founded                   : num  2013 2013 2015 2011 2013 ...
##  $ City                      : Factor w/ 68 levels "","Arlington",..: 1 1 36 1 55 1 52 36 52 23 
##  $ address_check             : Factor w/ 2 levels "False","True": 1 1 2 1 2 1 2 2 2 2 ...
##  $ Country                   : Factor w/ 23 levels "Belgium","Brazil",..: 8 22 23 22 23 7 23 23 
##  $ latitude                  : num  NA NA 40.7 NA 37.4 ...
##  $ longitude                 : num  NA NA -74 NA -122 ...
##  $ Industry_consolidated     : Factor w/ 16 levels "Computer & Network Security & Hardware",..: 
##  $ spc_Logistics.and.Supply.Chain: int  0 0 0 1 0 0 0 0 0 0 ...
```

```
# show columns with na
na = lapply(data, function(x) sum(ifelse(is.na(x) | x == "" | x == "not found", TRUE, FALSE)))
na[na > 0]
```

```
## $funding_round
## [1] 12
##
## $City
## [1] 46
##
## $latitude
## [1] 46
##
```

```
## $longitude
## [1] 46
```

Univariate Analysis

- Funding round

As intended, most companies are in Series B and Series C. Need to merge "series C" and "Series C".

```
# counts = table(data$funding_round)
# counts

data$funding_round[data$funding_round == "series C"] = "Series C"

# remove the level does not occur ("series C")
data$funding_round = factor(data$funding_round)

counts = table(data$funding_round)
counts
```
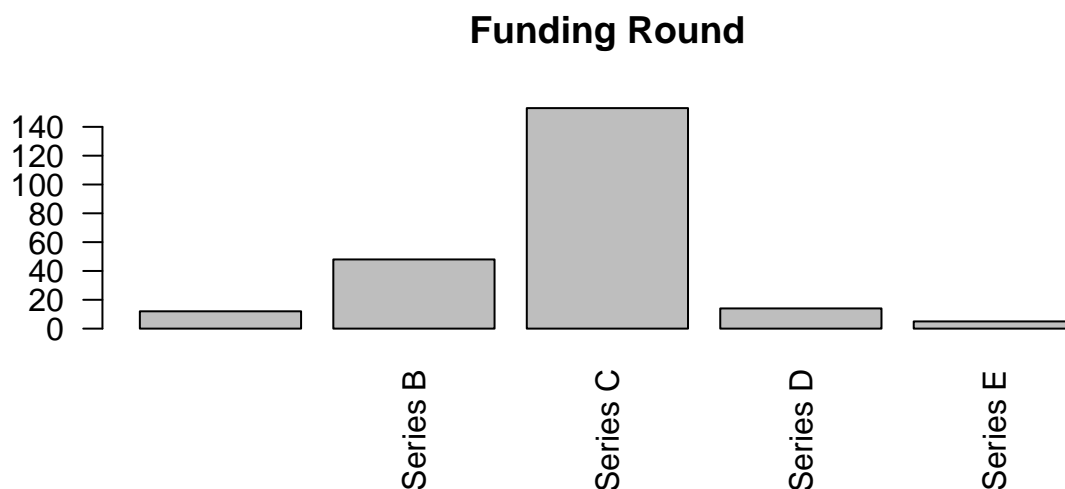
```
##
##        Series B Series C Series D Series E
##      12      48     153      14       5
```

```
prop.table(table(data$funding_round))
```

```
##
##         Series B   Series C   Series D   Series E
## 0.05172414 0.20689655 0.65948276 0.06034483 0.02155172
```

```
barplot(counts, main = "Funding Round", las = 2)
```
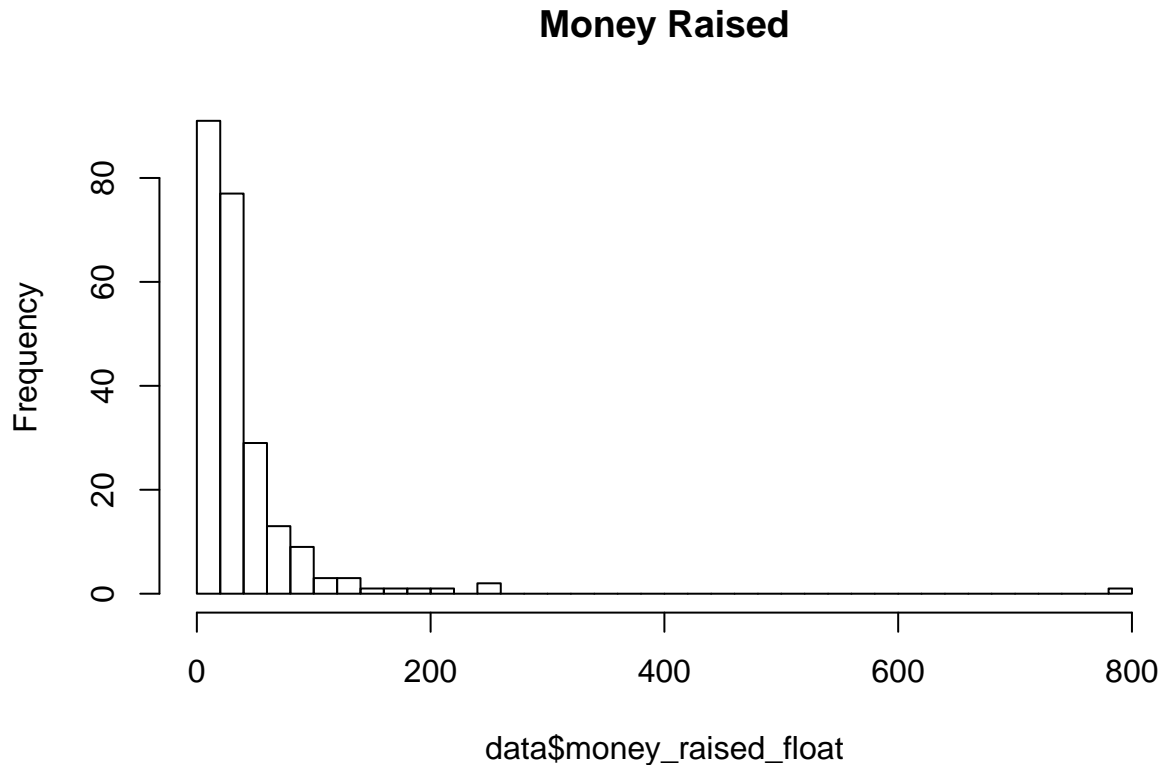
Funding Round

- Money raised

Most of companies raised money under $100M. We observe some outliers: Magic Leap, Pivotal, GitHub, and Opendoor.com.

```r
summary(data$money_raised_float)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.00   15.00   25.00   41.17   45.00  793.50
```

```r
hist(data$money_raised_float, breaks = 40, main = "Money Raised")
```

Money Raised


```r
# check the outliers
data[data$money_raised_float > 200, c("CompanyName", "funding_round", "CompanySize","money_raised_float
```

```
##         CompanyName funding_round CompanySize money_raised_float
## 28       Magic Leap      Series C   1001-5000              793.5
## 157         Pivotal      Series C   1001-5000              253.0
## 173          GitHub      Series B    501-1000              250.0
## 230    Opendoor.com      Series D     201-500              210.0
```

- Number of Employees

```r
# counts = table(data$CompanySize)
# counts

# str(data$CompanySize)

# clean up - factors
data$CompanySize = revalue(data$CompanySize, c("Nov-50"="11-50", "10-Jan"="1-10"))

# clean up - the level orders
data$CompanySize = factor(data$CompanySize, levels = c("1-10", "11-50","51-200","201-500", "501-1000",
```

```
                                                        "1001-5000","5001-10,000","10,001+"))
```

```
counts = table(data$CompanySize)
counts
```

```
##
##        1-10        11-50       51-200      201-500     501-1000    1001-5000
##           1           27          101           64           14           21
## 5001-10,000     10,001+
##           2           2
```

```
barplot(counts, main = "Company size", las=2)
```

Company size

- When Companies Are Founded

There are some companies founded before 2000. I suspect Hillshire Brands, founded in 1939, is a startup.

```
counts = table(data$Founded)
counts
```

```
##
## 1939 1976 1986 1989 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009
##    1    1    1    1    2    2    3    2    5    4   12   18   20   16   16
## 2010 2011 2012 2013 2014 2015 2016 2017
##   21   33   31   22   14    5    1    1
```

```
data[data$Founded < 2000, c("Founded", "CompanyName", "funding_round",
                            "CompanySize","money_raised_float")]
```

```
##     Founded                                  CompanyName
## 120    1999                                     Snagajob
## 142    1999 Trion Group, a Marsh & McLennan Agency, LLC Company
## 143    1986                                     Like.com
## 146    1976                                   Ticketmaster
## 152    1989                    La Jolla Pharmaceutical Company
```

```
## 214    1939                                 Hillshire Brands
##     funding_round CompanySize money_raised_float
## 120      Series C     201-500                   27
## 142      Series C     201-500                   70
## 143      Series C      11-50                    32
## 146      Series C 5001-10,000                   25
## 152      Series C      51-200                   12
## 214      Series B 5001-10,000                   24
```

- Country Companies Are Based In

Since I collected startups from TechCrunch, the US based news outlet, it turns out 77% startup in the dataset are based in the US. This might also be because the US produces the largest number of startups.

```
counts = table(data$Country)
counts
```

```
##
##        Belgium          Brazil          Canada          China          Denmark
##              1               1               4               1               1
##         France         Germany           India           iran           Israel
##              1               8               4               1               1
##          Italy           Japan           Korea     New Zealand          Norway
##              1               1               1               1               1
##         Poland          Russia       Singapore          Sweden         Thailand
##              1               1               4               1               1
##         Turkey  United Kingdom   United States
##              1              17             178
```

```
prop.table(counts)
```

```
##
##        Belgium          Brazil          Canada          China          Denmark
##    0.004310345     0.004310345     0.017241379     0.004310345     0.004310345
##         France         Germany           India           iran           Israel
##    0.004310345     0.034482759     0.017241379     0.004310345     0.004310345
##          Italy           Japan           Korea     New Zealand          Norway
##    0.004310345     0.004310345     0.004310345     0.004310345     0.004310345
##         Poland          Russia       Singapore          Sweden         Thailand
##    0.004310345     0.004310345     0.017241379     0.004310345     0.004310345
##         Turkey  United Kingdom   United States
##    0.004310345     0.073275862     0.767241379
```

- City

```
detach(package:plyr)
data %>%
  group_by(City) %>%
  summarize(n = n()) %>%
  arrange(desc(n)) %>%
  filter(City != "") %>%
  slice(1:10) %>%
  ggplot(., aes(x = reorder(City, -n), y = n)) +
  geom_bar(stat = "identity") +
  ggtitle("10 Top Cities Startups are based in") +
  theme(plot.title = element_text(hjust = 0.5, size=14)) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
```
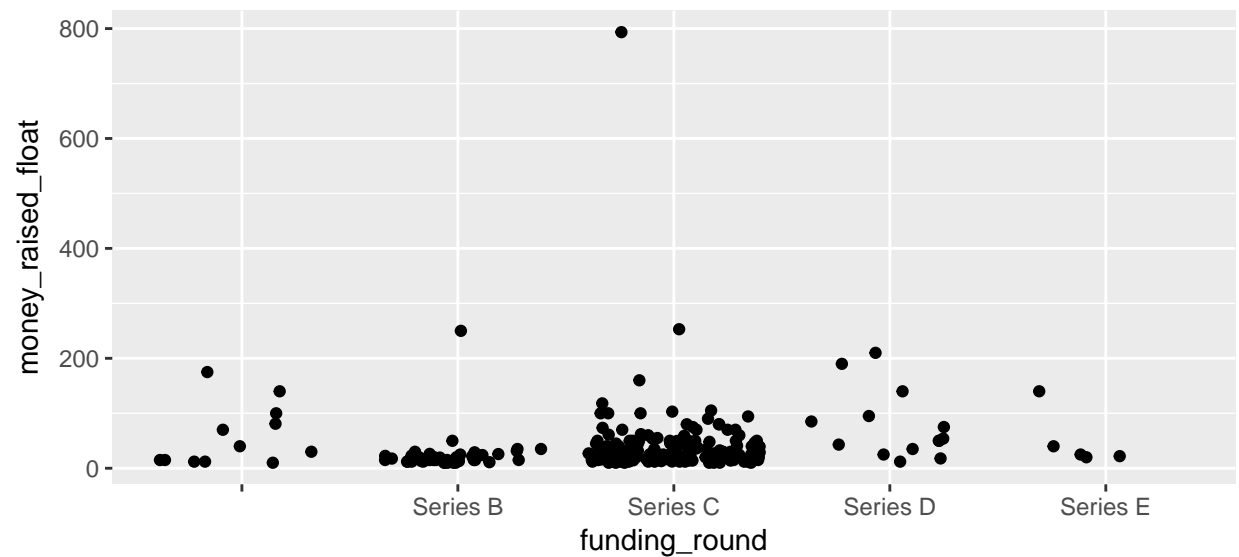
```
    theme(axis.title.x=element_blank()) +
    ylab("Number of Companies")
```

10 Top Cities Startups are based in
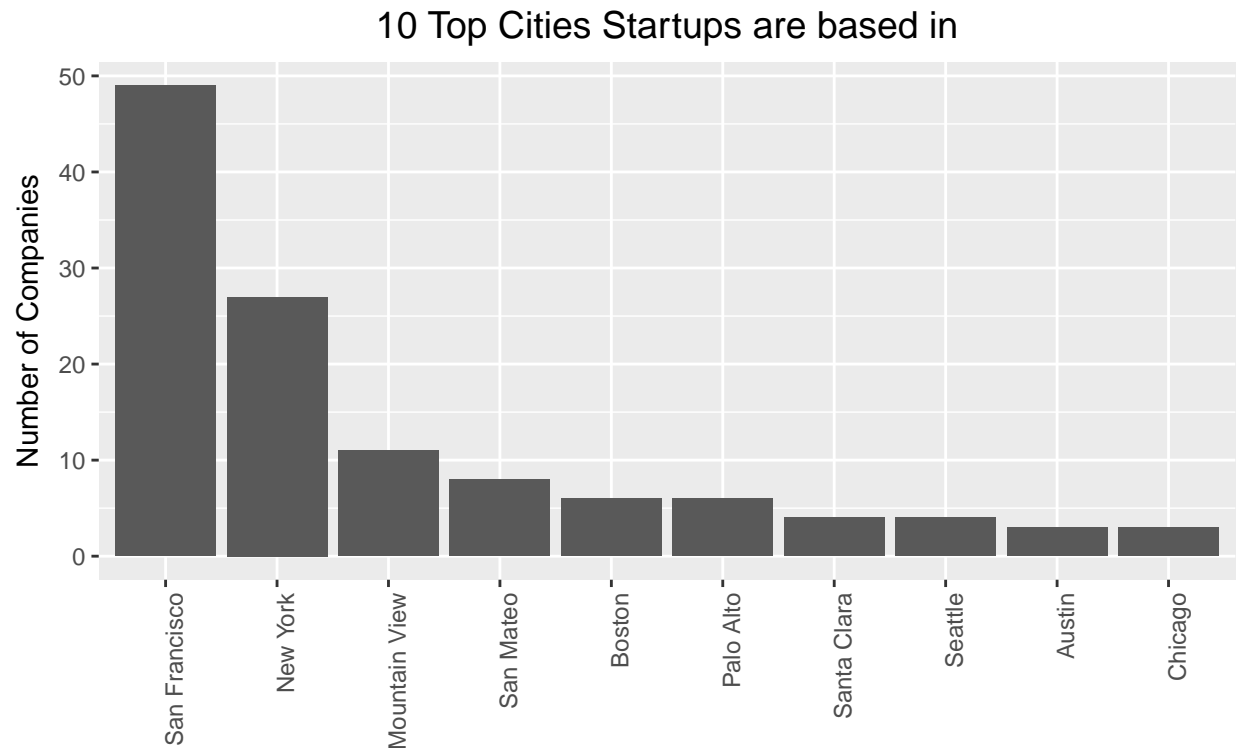
Bivariate Analysis

- Funding_round x money_raised

```
ggplot(data = data, aes(x= funding_round, y = money_raised_float)) +
  geom_jitter()
```

The table below shows mean and standard deviaiton of money raised for companies in each funding round. It makes sense that the mean increases as funding round progresses. Series E has lower mean than Series D. This might be because Series E is more of extension of Series D to sustain funding and not a funding round to drive a company to next level. Also note that stand devidations are quite larger for each round.
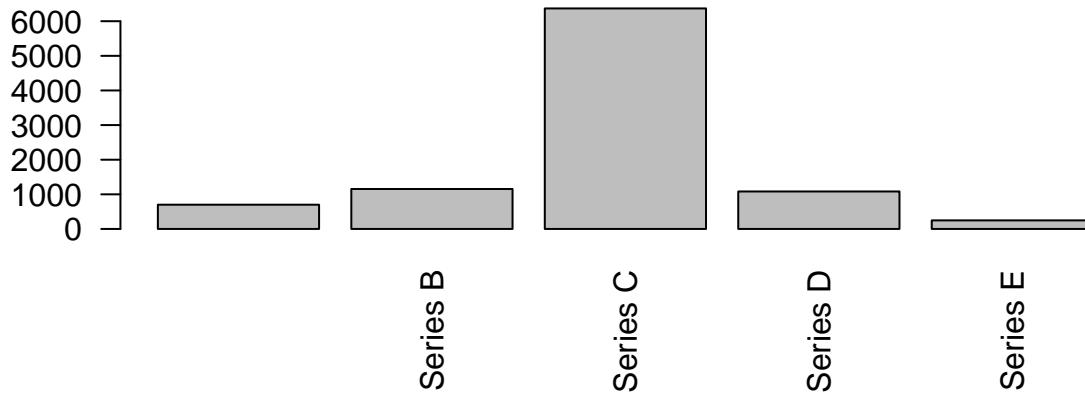
```
par(mfrow=c(2,1))

money = data %>%
  group_by(funding_round) %>%
  summarize(sum = sum(money_raised_float), mean = mean(money_raised_float), sd = sd(money_raised_float))
money
```

```
## # A tibble: 5 × 4
##   funding_round    sum      mean        sd
##          <fctr> <dbl>     <dbl>     <dbl>
## 1                700.0 58.33333 55.75161
## 2      Series B 1153.8 24.03750 34.28641
## 3      Series C 6369.2 41.62876 68.57037
## 4      Series D 1081.8 77.27143 62.11565
## 5      Series E  247.0 49.40000 51.25232
```
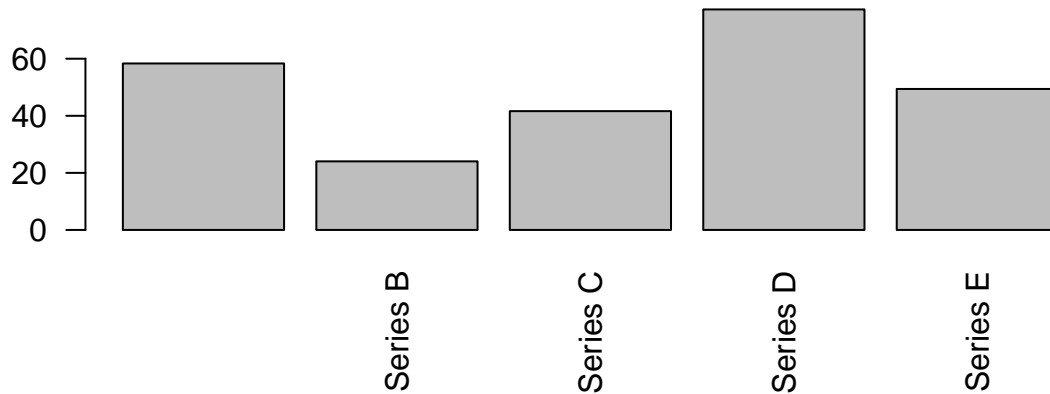
```
counts = money$sum
names(counts) = money$funding_round
barplot(counts, las = 2, main = "Total Money Raised by Funding Round")

counts = money$mean
names(counts) = money$funding_round
barplot(counts, las = 2, main = "Average Money Raised by Funding Round")
```

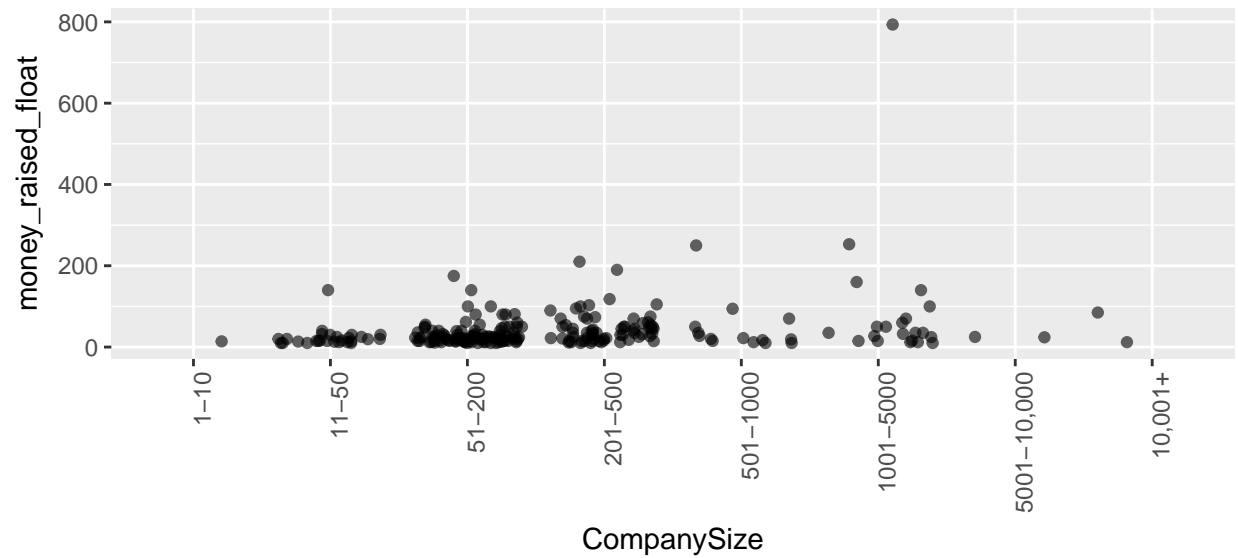**Total Money Raised by Funding Round**

Average Money Raised by Funding Round


- Company size x money_raised

```
ggplot(data = data, aes(x= CompanySize, y = money_raised_float)) +
  geom_jitter(alpha = 0.6) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

```
money = data %>%
  group_by(CompanySize) %>%
  summarize(mean = mean(money_raised_float), sd = sd(money_raised_float))

counts = money$mean
names(counts) = money$CompanySize
barplot(counts, las = 2, main = "Average Money Raised by Company Size")
```

Average Money Raised by Company Size
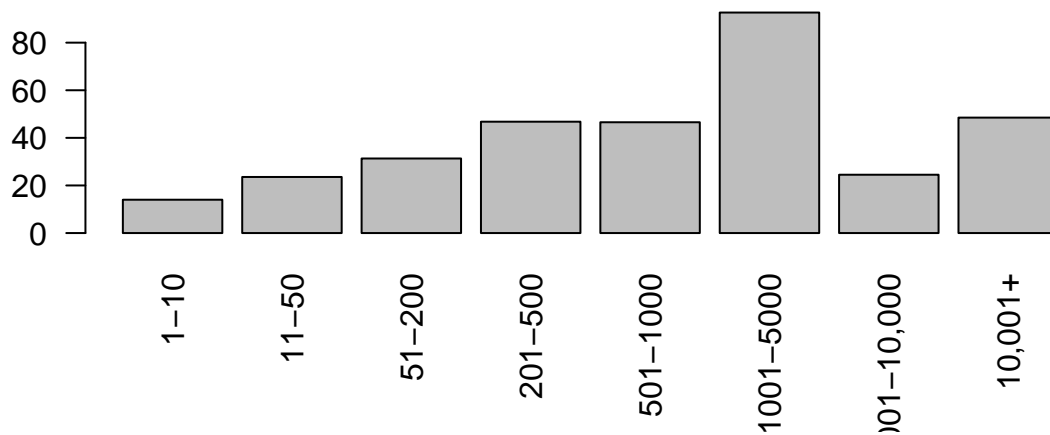
Location

```
# Set the center of map

all_states <- map_data("state")
```

```
ggplot() + geom_polygon(data = all_states, aes(x=long, y = lat, group = group),
                        colour="white", fill="grey10") +
  coord_fixed(1.3) +
  geom_jitter(data = data, mapping = aes(
    x = longitude, y = latitude, color = "Orangered" ,size = money_raised_float),
            alpha = 0.4, width = 0.7, height = 0.7) +
  scale_size(range = c(2, 12)) +
  ggtitle("Company Location with Money Raised") +
  theme(plot.title = element_text(hjust = 0.5, size=18)) +
  labs(x=NULL, y=NULL) +
  theme(panel.border = element_blank())
```

Company Location with Money Raised