# Final Project 527”

Satoshi Ido, Shu-Ming Meng, Robert Jung

17 April 2023

## library

```
library("MASS")
library("lmtest")
library("stringr")
library("dplyr")
library("data.table")
library("stringi")
library("openxlsx")
library("ggplot2")
library("lattice")
library("epiDisplay")
library("car")
library("treemap")
library("d3treeR")
library("htmlwidgets")
```

## import the dataset

```
print(getwd())
```

```
## [1] "/Users/satoshiido/Documents/statistical-analysis/527"
```

```
df <- read.csv("/Users/satoshiido/Documents/statistical-analysis/527/unicorn_data.csv")
head(df, 3)
```

```
##        Company Valuation.Billion Date.Joined       Country          City
## 1       stripe               $95   1/23/2014 United States San Francisco
## 2    oyo rooms              $9.6   9/25/2018         India      Gurugram
## 3 servicetitan              $9.5  11/14/2018 United States      Glendale
##                       Industry
## 1                      Fintech
## 2                       Travel
## 3 Internet software & services
##                                                   Select.Investors Founded.Year
## 1                     Khosla Ventures, LowercaseCapital, capitalG         2010
## 2 SoftBank Group, Sequoia Capital India,Lightspeed India Partners         2012
## 3     Bessemer Venture Partners, ICONIQ Capital, Battery Ventures         2012
##   Total.Raised.Million Financial.Stage Investors.Count Deal.Terms
## 1              $2.901B           Asset              39         12
## 2              $3.114B            None              20         11
## 3              $1.099B            None              16          6
##                                                          Investors
## 1                     Khosla Ventures, LowercaseCapital, capitalG
## 2 SoftBank Group, Sequoia Capital India,Lightspeed India Partners
## 3     Bessemer Venture Partners, ICONIQ Capital, Battery Ventures
##   Portfolio.Exits Entrepreneur Final.Degree Final.School
## 1               1                        NA
## 2            None                        NA
## 3            None                        NA
```

## preprocess

clean the data

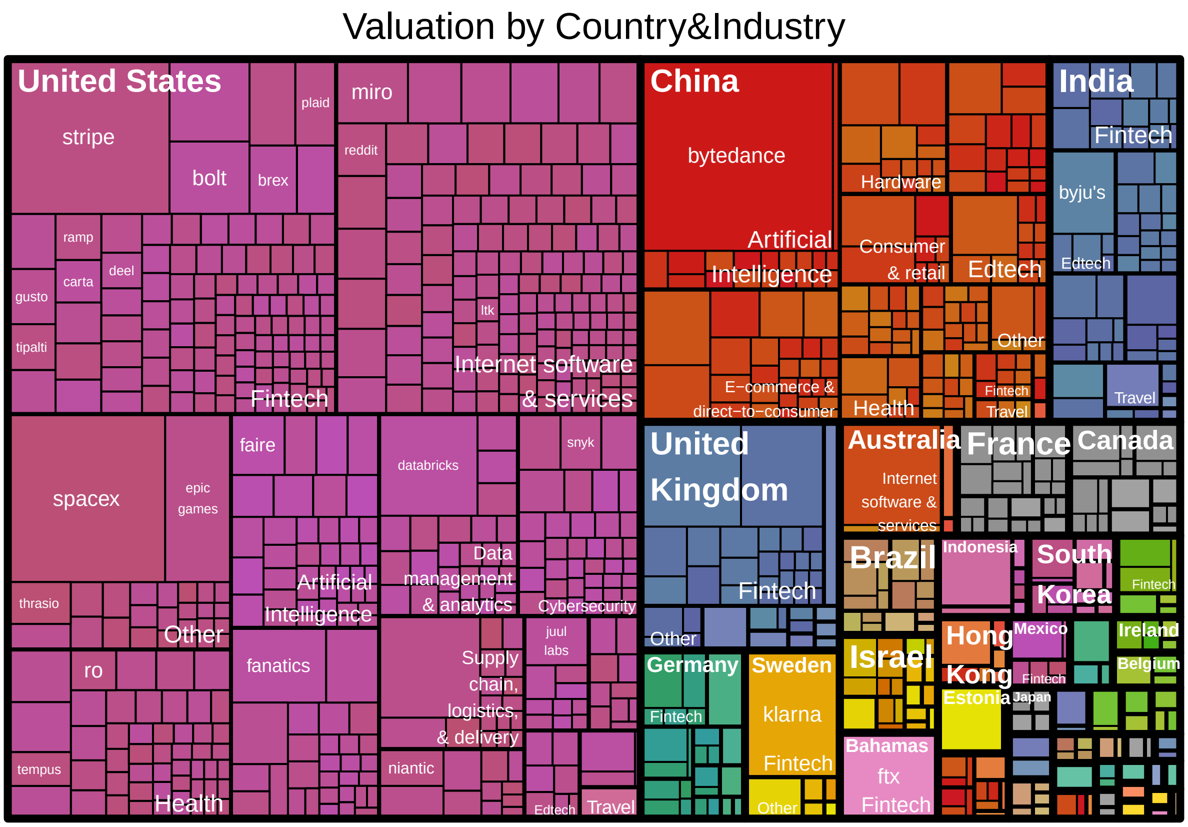
```
billion_func <- function(y){
    # if the value in the column include "B", extract them
    tmp <- y %>% filter(str_detect(y[, "Total.Raised.Million"], "B"))
    # remove the $ and B signs
    tmp$Total.Raised.Million <- stri_replace_all_regex(
            tmp$Total.Raised.Million, pattern = c("[$]", "[B]"),
            replacement = c("", ""),
            vectorize = FALSE
            )
    # change the data type to numeric
    tmp <- tmp %>% mutate_at("Total.Raised.Million", as.numeric)
    # unit convert from $B to $M
    tmp[, "Total.Raised.Million"] <- tmp[, "Total.Raised.Million"] * 1000

    return(tmp)
    }

million_func <- function(y){
    # if the value in the column include "M", extract them
    tmp <- y %>% filter(str_detect(y[, "Total.Raised.Million"], "M"))
    # remove the $ and M signs
    tmp$Total.Raised.Million <- stri_replace_all_regex(
            tmp$Total.Raised.Million, pattern = c("[$]", "[M]"),
            replacement = c("", ""),
            vectorize = FALSE
            )
    # change the data type to numeric
    tmp <- tmp %>% mutate_at("Total.Raised.Million", as.numeric)
    return(tmp)
    }
```

```r
cleaner <- function(x){
    # delete the unncessary columns
    x <- subset(
        x,
        select = -c(
            Investors, Portfolio.Exits, Entrepreneur,
            Final.Degree, Final.School)
        )
    # divide into two dataset and convert
    tmp0 <- billion_func(x)
    tmp1 <- million_func(x)
    x <- rbind(tmp0, tmp1)

    # remove the $ and convert character into numeric
    x$Valuation.Billion <- stri_replace_all_regex(
            x$Valuation.Billion, pattern = "[$]",
            replacement = "",
            vectorize = FALSE
            )
    x <- x %>% mutate_at("Valuation.Billion", as.numeric)

    # create the another column with unit of "M"
    x[, "Valuation.Million"] <- x[, "Valuation.Billion"] * 1000

    # change string to date format
    x$Date.Joined <- as.Date(x$Date.Joined, format = "%m/%d/%Y")
    # return the list of Investors given delimited strings
    x$Select.Investors <- strsplit(x$Select.Investors, split = ", ")

    # calculate how many months pasted since each company has joined the unicorn
    floor_dec <- function(y, level=1) {round(y - 5 * 10^(-level - 1), level)}
    YearMonth <-
        (20231200 - round(as.numeric(gsub("-", "", x$Date.Joined)), -2)) / 100
        for (i in 1 : length(x$Founded.Year)) {
            if ((YearMonth[i] %% 100 - 8) < 0) {
                x$Date.Joined_in_months[i] <- (
                    floor_dec(YearMonth[i]/100,0)-1)*12+abs(YearMonth[i]%%100-4
                )
            }
            else {
                x$Date.Joined_in_months[i] <- (
                    floor_dec(YearMonth[i]/100,0))*12+abs(YearMonth[i]%%100-4
                )
            }
        }
    return(x)
    }

df <- cleaner(df)
```

# Visualization

## Tree Map by Country and Industry

```
# basic treemap
treemap(df,
        index = c("Country", "Industry", "Company"),
        type = "index",
        vSize = "Valuation.Billion",
        fontcolor.labels = c("white", "white"),
        fontsize.title = 14,
        fontsize.labels = c(12, 9, 8),
        palette = "Set2",
        bg.labels = c("transparent"),
        align.labels = list(
          c("left", "top"),
          c("right", "bottom"),
          c("center", "center")
        ),
        border.col = c("black", "black", "black"),
        border.lwds = c(4, 2, 1),
        title = "Valuation by Country&Industry",
        overlap.labels = 0,
        inflate.labels = F
    )
```

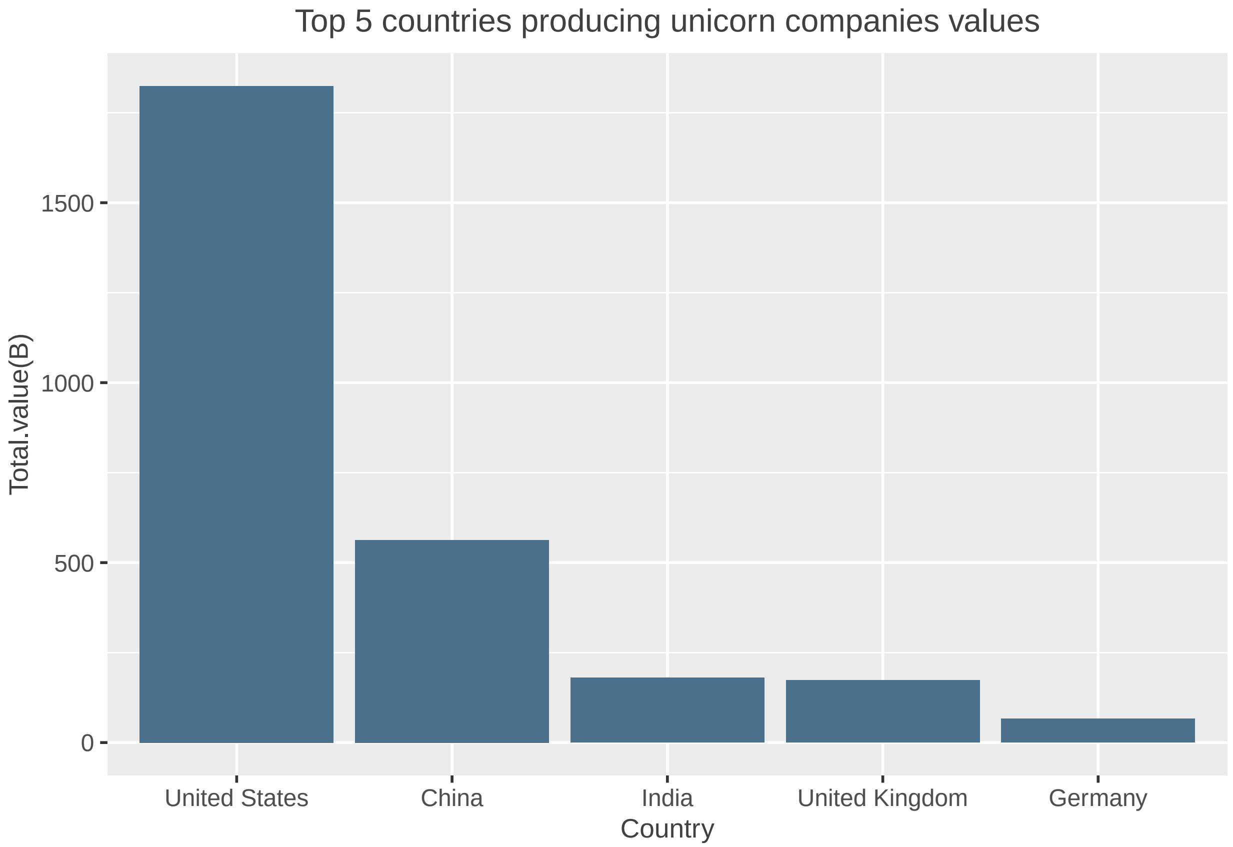


4

We can see some countries (US, China mainly) are the major countries productiong unicorn companies. It is noticeable that there are some popular industries (such as Fintech, IT, AI, E-commerce) and some countries have advantage in specific industries For example, UK has many unicorn companies in Fintech industry while AI and E-commerce are popular indsutries in China. From this plot, we decided to mainly look into Country, City, and Industry.

## Top 5 Countries

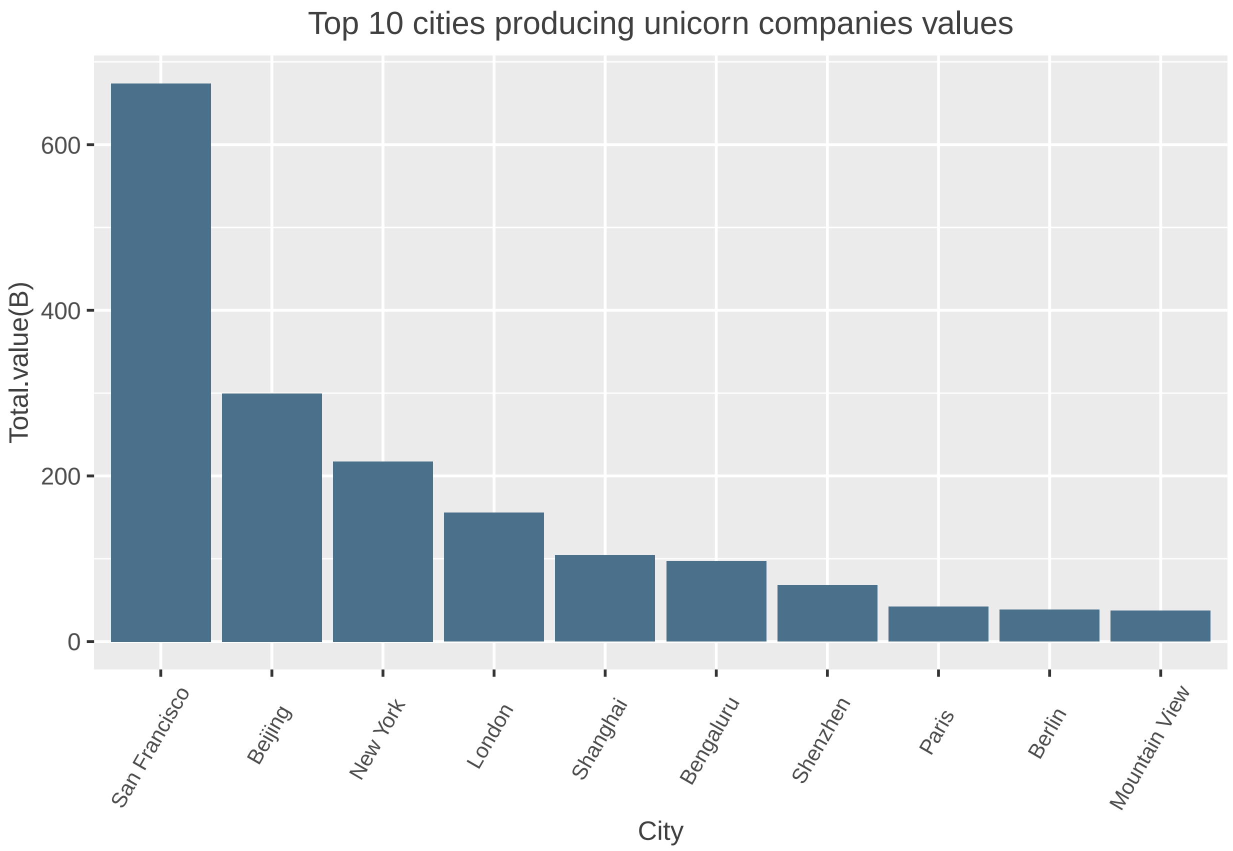
```r
# group by country and count the number of each country
tmp <- df %>%
    group_by(Country) %>%
    summarise(total.country.count = n(), .groups = "drop")
# pick the top 5 countries which produce the unicorn the most
top5countries <- unlist(tmp[order(tmp$total.country.count, decreasing = TRUE), ][0:5, 1])
tmp <- df[which(df$Country %in% top5countries), ] %>%
    group_by(Country) %>%
    summarise(total.value.Billion = sum(Valuation.Billion), .groups = "drop") %>%
    as.data.frame()

# plot
ggplot(
        tmp,
        aes(x = reorder(Country, -total.value.Billion),
        y = total.value.Billion)) +
    geom_bar(stat = "identity", fill = "skyblue4") +
    labs(
        title = "Top 5 countries producing unicorn companies values",
        x = "Country",
        y = "Total.value(B)") +
    theme(
        plot.title = element_text(size = 12, hjust = 0.5, color = "gray25"),
        axis.title.x = element_text(size = 10, hjust = 0.5, color = "gray25"),
        axis.title.y = element_text(size = 10, hjust = 0.5, color = "gray25")
        )
```



Top countries producing the unicorn companies most are US, China, India, UK, and Germany. US and China are leading unicorn booms.

### Top 10 Cities

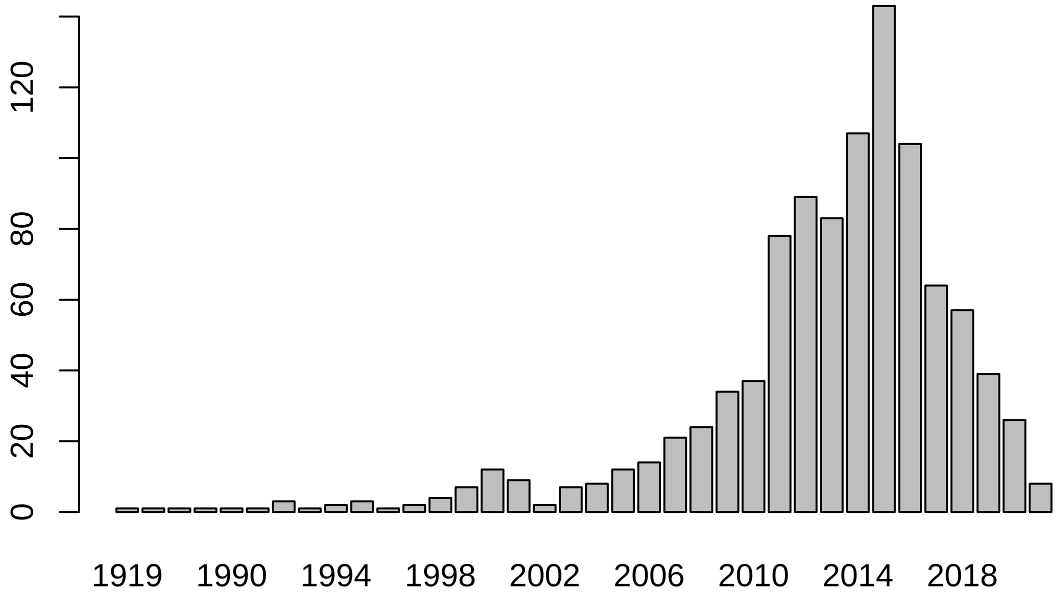
```
# group by city and count the number of each city
tmp <- df %>%
    group_by(City) %>%
    summarise(total.city.count = n(), .groups = "drop")
# pick the top 10 cities which produce the unicorn the most
top10cities <- unlist(tmp[order(tmp$total.city.count, decreasing = TRUE), ][0:10, 1])
tmp <- df[which(df$City %in% top10cities), ] %>%
    group_by(City) %>%
    summarise(total.value.Billion = sum(Valuation.Billion), .groups = "drop") %>%
    as.data.frame()

# plot
ggplot(
        tmp,
        aes(x = reorder(City, -total.value.Billion),
        y = total.value.Billion)) +
    geom_bar(stat = "identity", fill = "skyblue4") +
    labs(
        title = "Top 10 cities producing unicorn companies values",
        x = "City",
        y = "Total.value(B)") +
    theme(
        plot.title = element_text(size = 12, hjust = 0.5, color = "gray25"),
        axis.title.x = element_text(size = 10, hjust = 0.5, color = "gray25"),
        axis.title.y = element_text(size = 10, hjust = 0.5, color = "gray25"),
        axis.text.x = element_text(size = 8, angle = 60, vjust = 0.7, hjust = 0.7)
        )
```



Top cities producing many unicorn companies are SF, Beijing, NY, London, Shanghai, Bengaluru, Shenzhen, and so on. Mostly, the major metropolitan areas in US, China, and India

## Founded Year

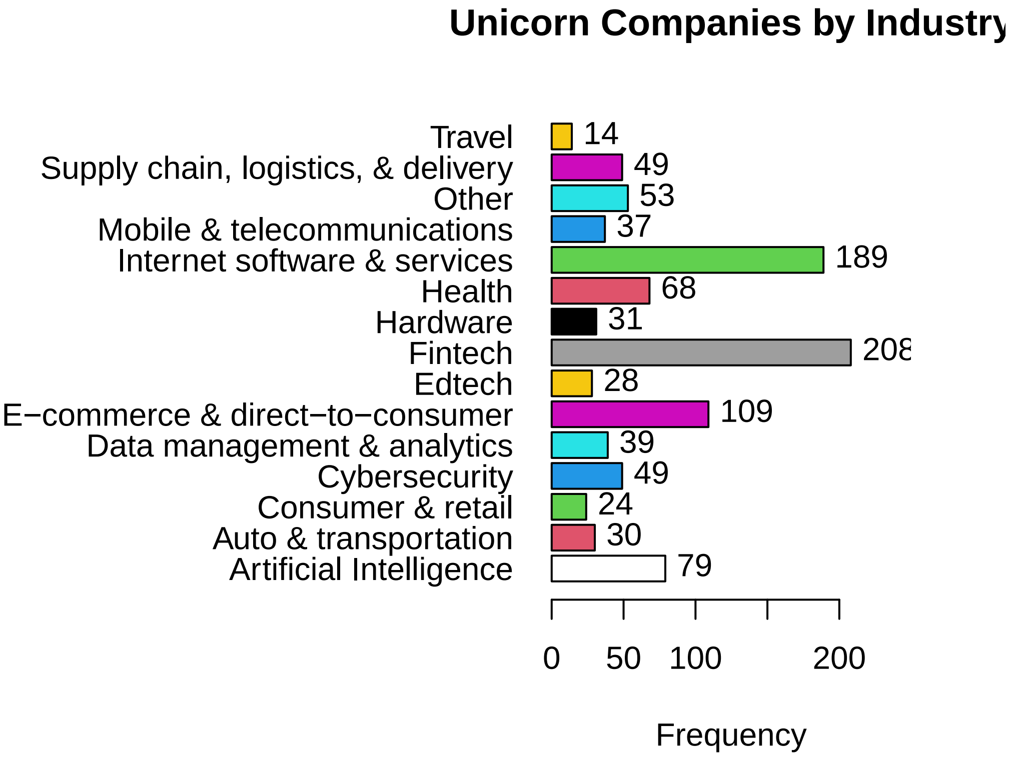
```
# barchart
barplot(table(df$Founded.Year))
```



As we expected, the majority of unicorn companies are founded in 2010's. It is obvious that it is a recent trend among startups for not doing IPO.

### Number of Unicorn Companies by Each Industry

```
# frequency barchart
tab1(df$Industry, main = "Unicorn Companies by Industry", cum.percent = TRUE)
```



```
## df$Industry :
##                                       Frequency Percent Cum. percent
## Artificial Intelligence                      79     7.8          7.8
## Auto & transportation                        30     3.0         10.8
## Consumer & retail                            24     2.4         13.2
## Cybersecurity                                49     4.9         18.1
## Data management & analytics                  39     3.9         21.9
## E-commerce & direct-to-consumer             109    10.8         32.8
## Edtech                                       28     2.8         35.6
## Fintech                                     208    20.7         56.2
## Hardware                                     31     3.1         59.3
## Health                                       68     6.8         66.0
## Internet software & services                189    18.8         84.8
## Mobile & telecommunications                  37     3.7         88.5
## Other                                        53     5.3         93.7
## Supply chain, logistics, & delivery          49     4.9         98.6
## Travel                                       14     1.4        100.0
##   Total                                    1007   100.0        100.0
```

We can clearly see the Fintech and Internet software & services are the two most dominating Industries in this case with 208 and 189 companies respectively.

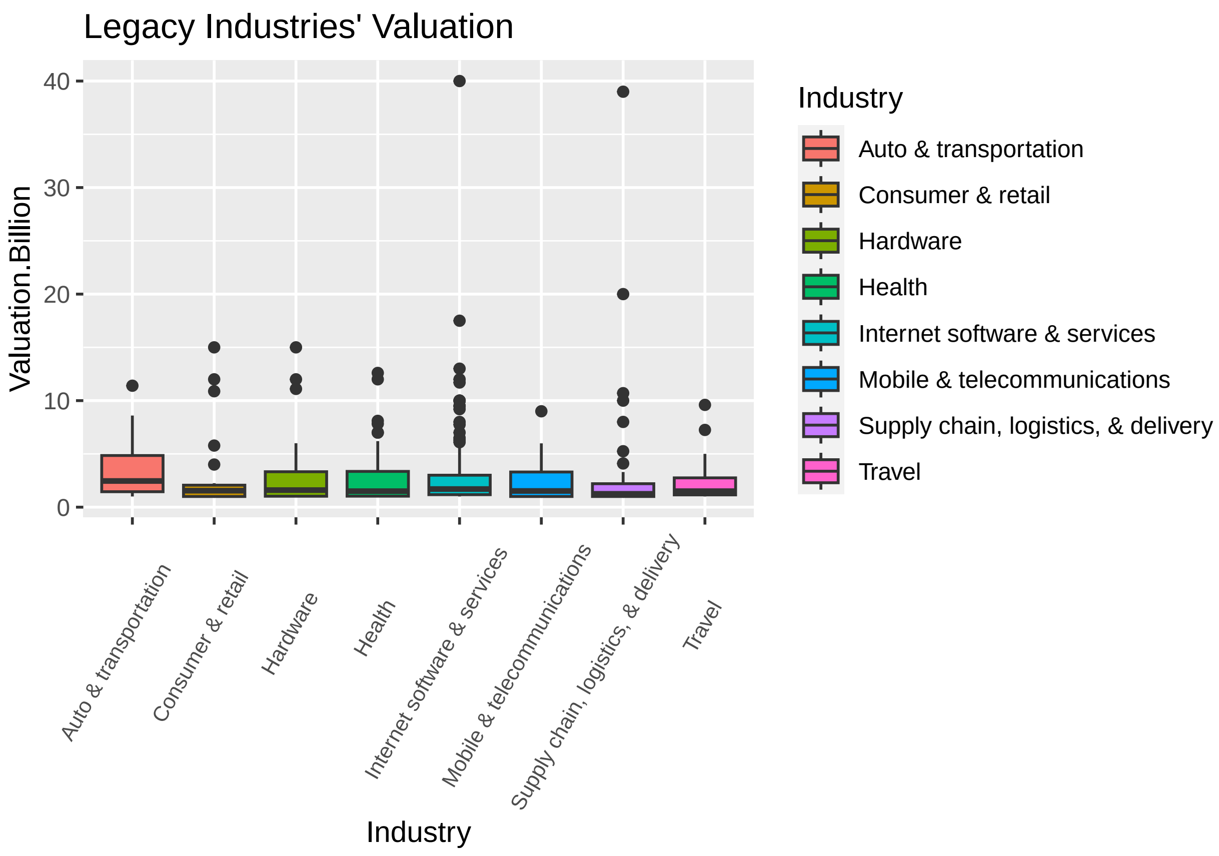
## Legacy or Growing Industries vs Companies Valuation

Preprocess

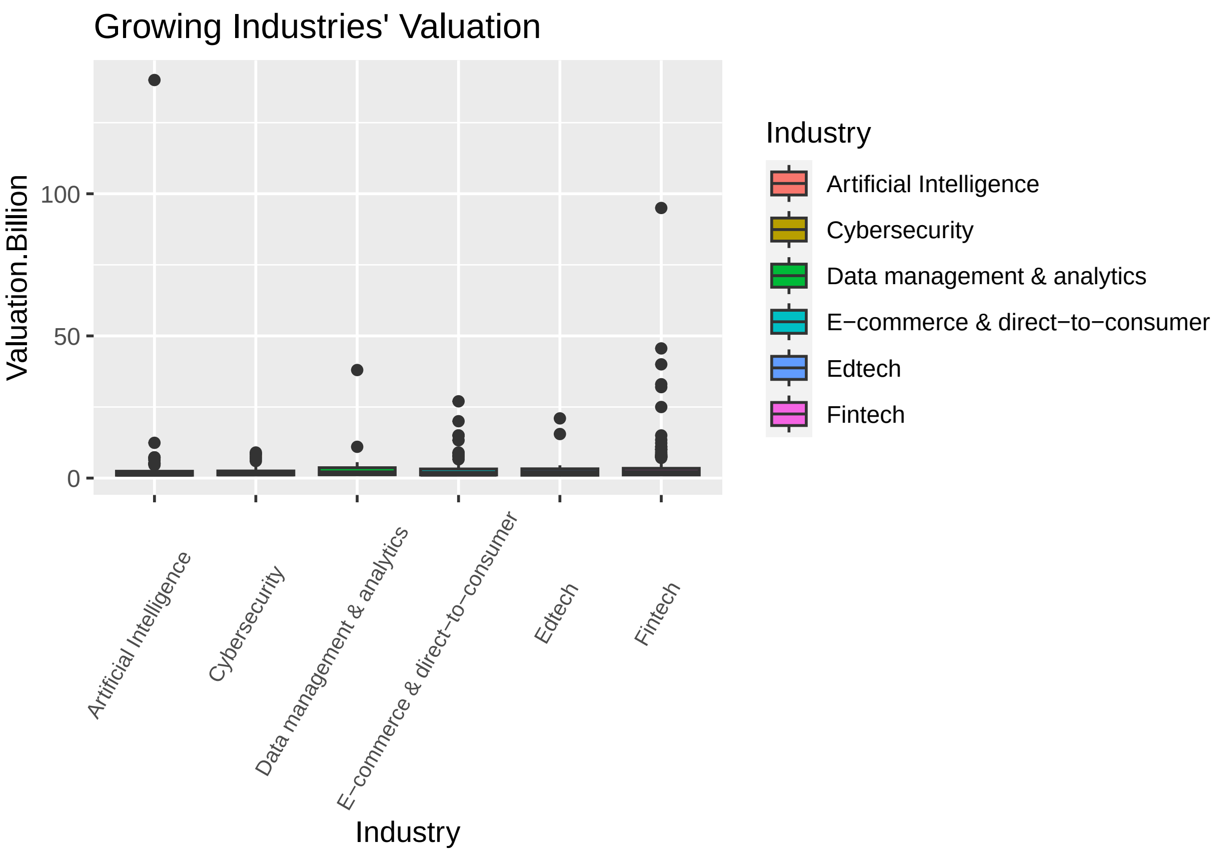
```r
# divide into subgroup
# Growing IND
growing_list <- c("Artificial Intelligence", "Cybersecurity",
        "Data management & analytics", "Edtech", "Fintech",
        "E-commerce & direct-to-consumer"
        )
df_GI <- df[which(df$Industry %in% growing_list), ]

# Legacy IND
legacy_list <- c("Auto & transportation","Consumer & retail","Hardware",
            "Health","Supply chain, logistics, & delivery","Travel",
            "Internet software & services","Mobile & telecommunications"
            )
df_LI <- df[which(df$Industry %in% legacy_list), ]
```

### Legacy or Growing Industries vs Companies Valuation

```r
# plotting
ggplot(df_LI, aes(x=Industry, y=Valuation.Billion, fill=Industry)) +
    geom_boxplot()+
    theme(axis.text.x = element_text(size = 8, angle = 60, vjust = 0.7, hjust = 0.7)) +
    labs(
        title = "Legacy Industries' Valuation",
        x = "Industry",
        y = "Valuation.Billion")
```



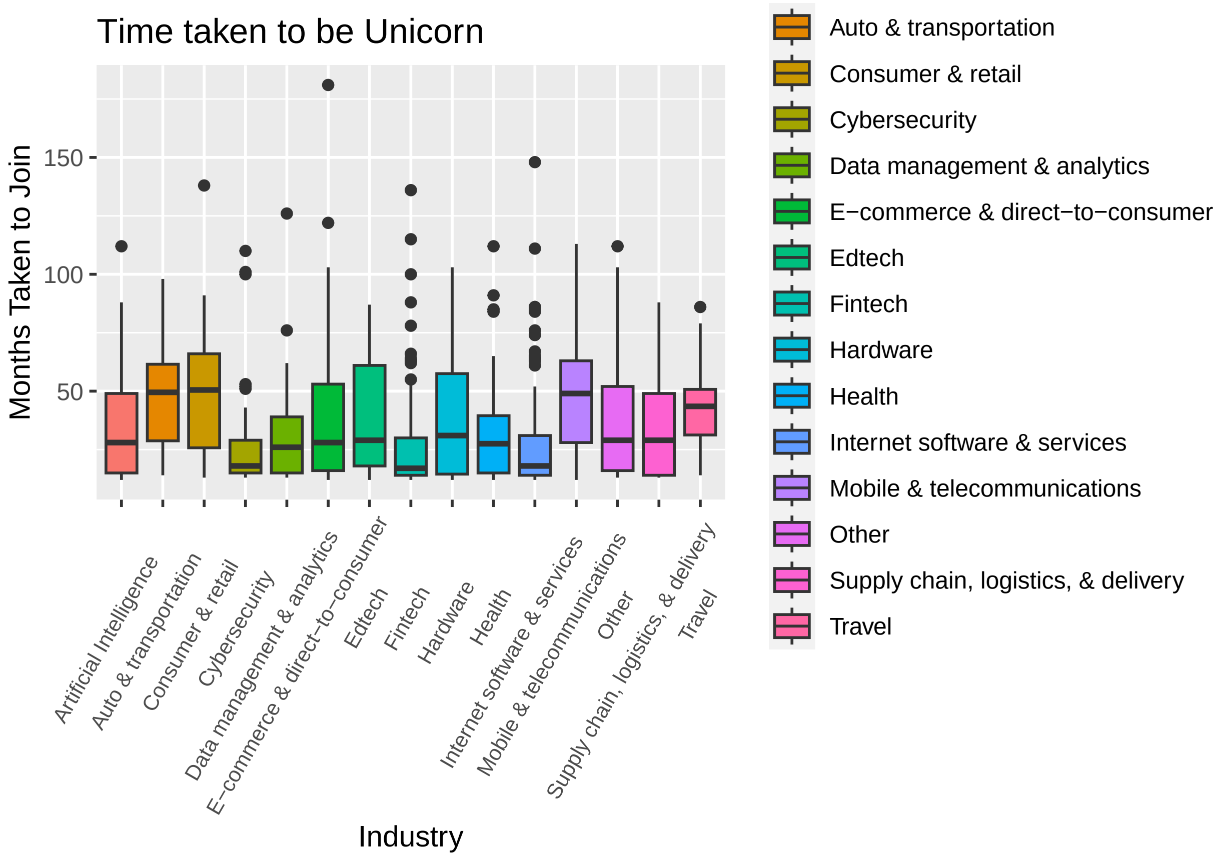
```
ggplot(df_GI, aes(x=Industry, y=Valuation.Billion, fill=Industry)) +
    geom_boxplot()+
    theme(axis.text.x = element_text(size = 8, angle = 60, vjust = 0.7, hjust = 0.7)) +
    labs(
        title = "Growing Industries' Valuation",
        x = "Industry",
        y = "Valuation.Billion")
```



There are apparently a lot of outliers in both legacy and growing industries, but the scale of growing industries is larger. That could mean even with similar median company valuation, more companies from growing industries have way higher company valuation.

### Time of Each Industry Taking to Become Unicorn Comapnies

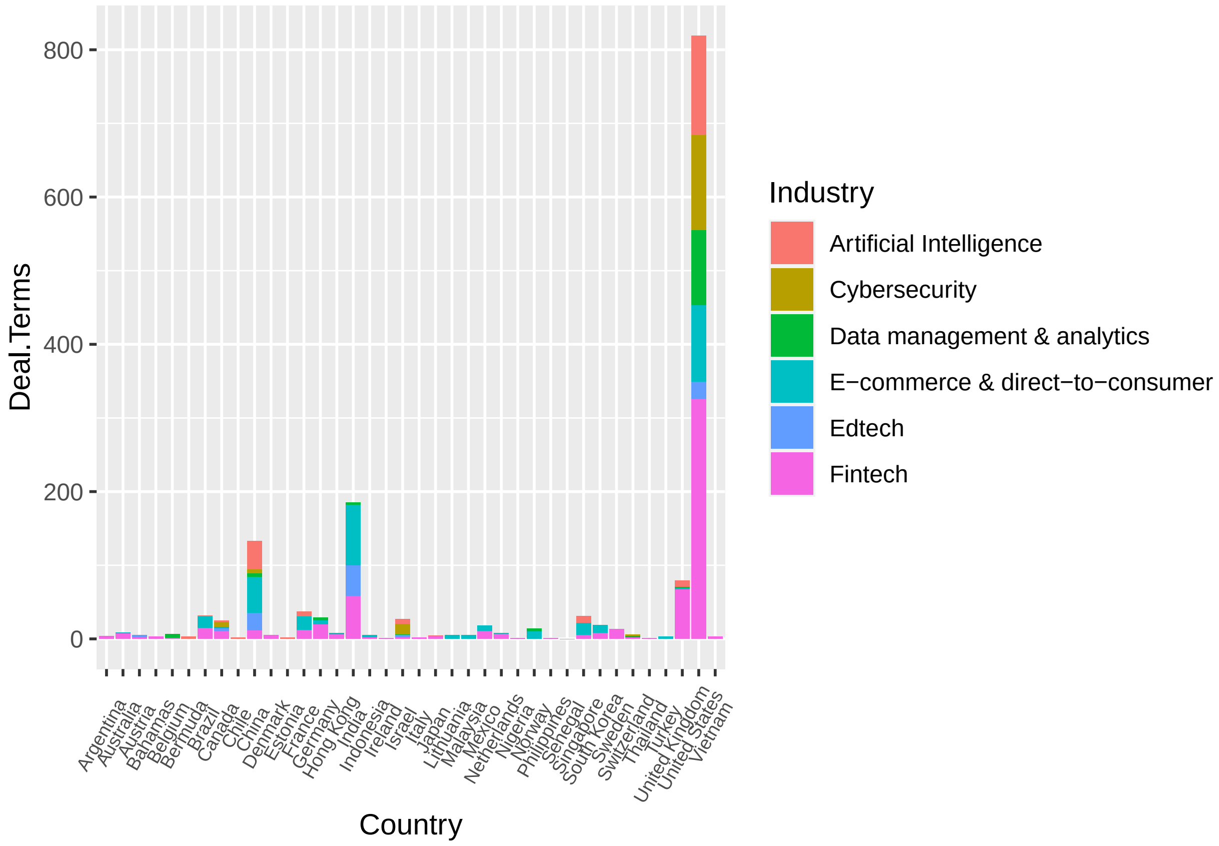
```
ggplot(df, aes(x=Industry, y=Date.Joined_in_months, fill=Industry)) +
    geom_boxplot()+
    theme(
        axis.text.x = element_text(size = 8, angle = 60, vjust = 0.7, hjust = 0.7)) +
    labs(
    title = "Time taken to be Unicorn",
    x = "Industry",
    y = "Months Taken to Join")
```



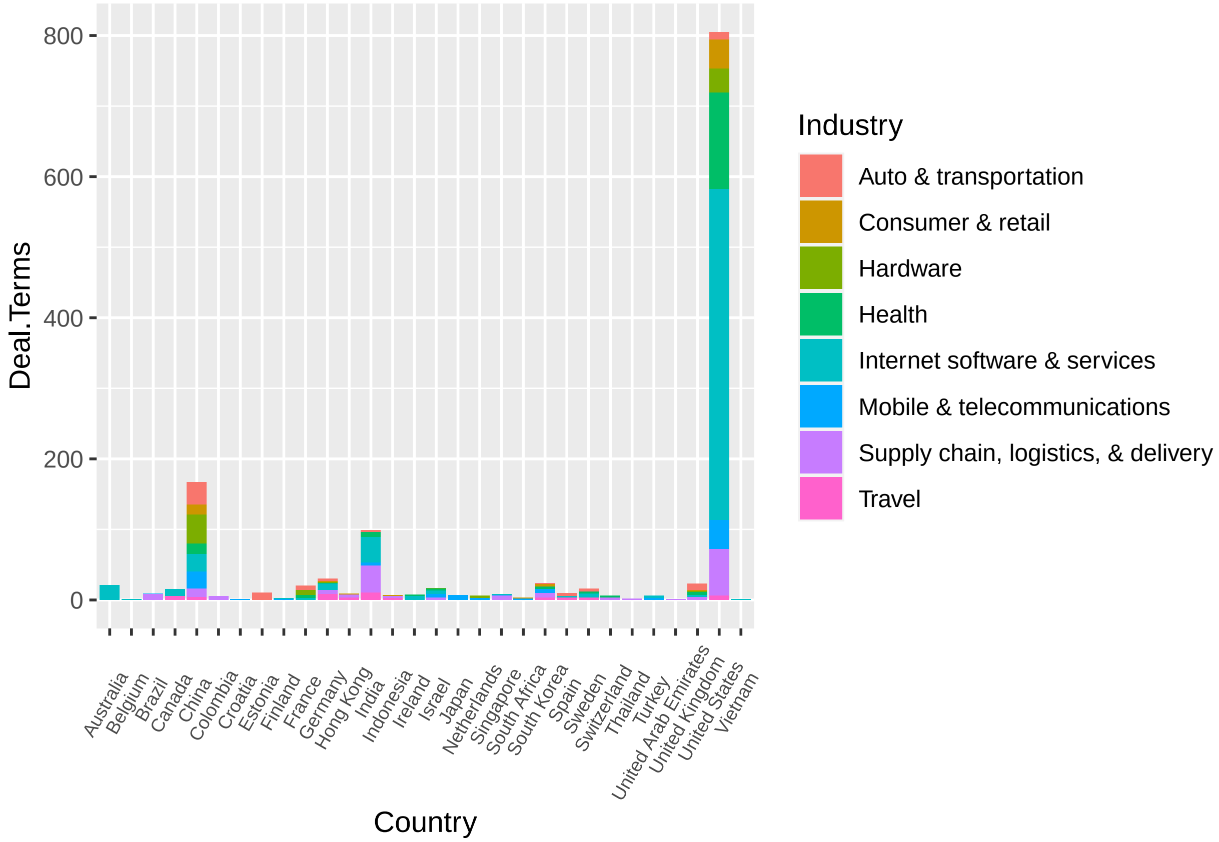
We roughly calculated how many months each company took to become unicorn company and group them by Industry. Clearly, there's difference between some these industries.

## Deal.Term vs Country by Industry

```
ggplot(
    data = df_GI,
    aes(x = Country, y = Deal.Terms, fill = Industry)) +
    geom_bar(position = "stack", stat="identity") +
    theme(axis.text.x = element_text(size = 7, angle = 60, vjust = 0.7, hjust = 0.7)
    )
```

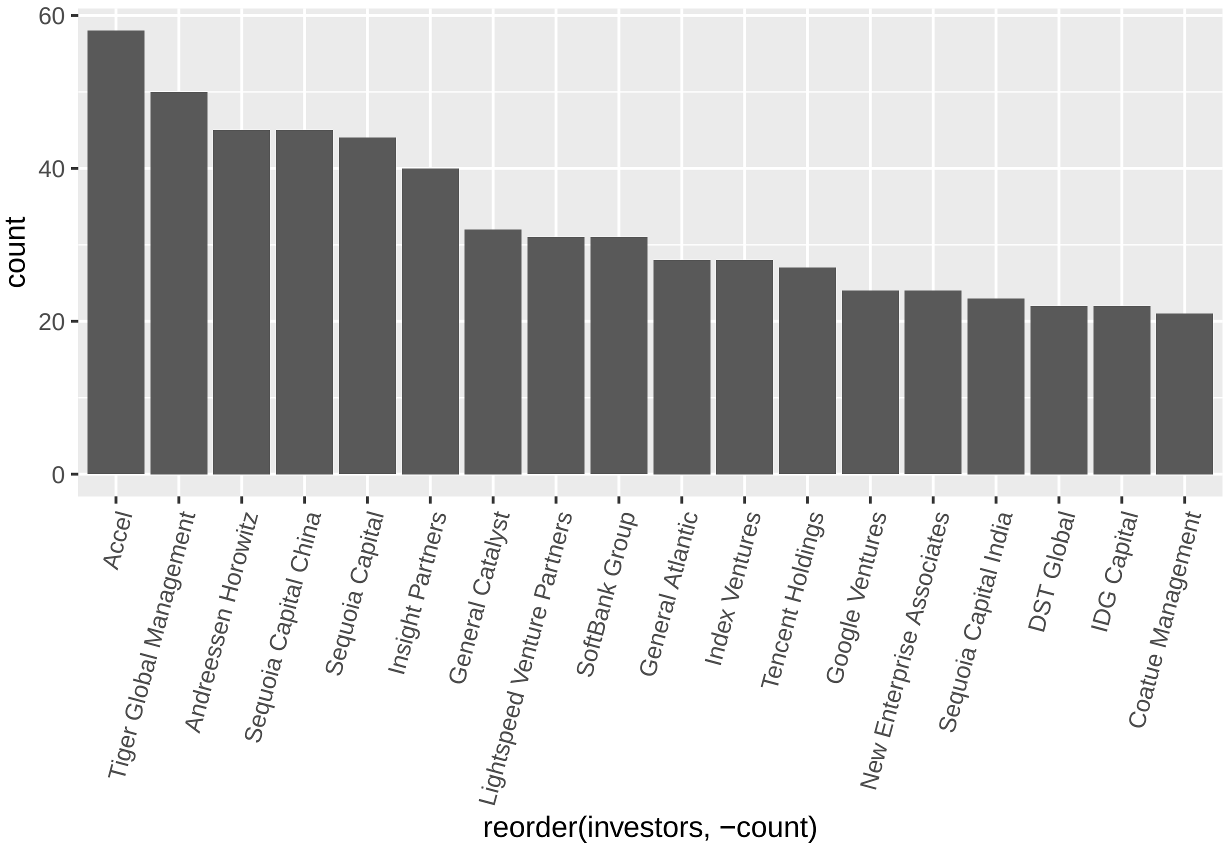


```
ggplot(
    data = df_LI,
    aes(x = Country, y = Deal.Terms, fill = Industry)) +
    geom_bar(position = "stack", stat = "identity") +
    theme(axis.text.x = element_text(size = 7, angle = 60, vjust = 0.7, hjust = 0.7)
    )
```



## Investors

```
# count the number of unicorn startups each investor has invested
# create the matrix of investors frequency count
words.freq <- table(unlist(df$Select.Investors))
investors <- cbind(names(words.freq), as.integer(words.freq))
investors <- as.data.frame(investors)
investors$V2 <- as.numeric(investors$V2)
colnames(investors) <- c("investors", "count")
investors <- investors[order(investors$count, decreasing = TRUE), ]
top_investors <- filter(investors, count > 20)


options(repr.plot.width = 15, repr.plot.height = 8)
top_investors <- filter(investors, count > 20)
ggplot(data = top_investors, aes(x = reorder(investors, -count), y = count)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 75, vjust = 1, hjust = 1))
```



As we rank the investors based on how many unicorn companies each of them has invested, we notice that there are some succesful investors who invested more than 20 startups which ended up becoming unicorn companies. We picked those successful investors and did plot them

## Modeling

### Data Preprocess

```
# Country
## group by country and count the number of each country
tmp <- df %>%
    group_by(Country) %>%
    summarise(total.country.count = n(), .groups = "drop")
topcountries.list <- unlist(tmp[tmp$total.country.count > 5, ][, 1])

# city
## group by city and count the number of each city
tmp <- df %>%
    group_by(City) %>%
    summarise(total.city.count = n(), .groups = "drop")
topcities.list <- unlist(tmp[tmp$total.city.count > 10, ][, 1])


# Country.label & City.label
```

```r
# Add a columns to Dataframe regarding to top countries and top cities
df2 <- df %>%
        mutate(Country.label = ifelse(
                Country %in% topcountries.list, 1, 0)) %>%
        mutate(City.label = ifelse(
                City %in% topcities.list, 1, 0))

# replace the county names if they produce less than or equal to 5 companies
df2$Country[df2$Country.label == 0] <- "others"
# replace the city names if they produce less than or equal to 10 companies
df2$City[df2$City.label == 0] <- "others"

# Investor.label
words.freq <- table(unlist(df2$Select.Investors))
investors <- cbind(names(words.freq), as.integer(words.freq))
investors <- as.data.frame(investors)
investors$V2 <- as.numeric(investors$V2)
colnames(investors) <- c("investors", "count")
investors <- investors[order(investors$count, decreasing = TRUE), ]
top_investors <- filter(investors, count > 20)
## x = df2
investor.fun <- function(x) {
    y <- x$Select.Investors
    for (i in 1:length(y)) {
        for (k in 1:length(y[[i]])) {
            # if "k"th investor in "i"th list in
            # dataframe is in top_investors$investors,
            ifelse(
                y[[i]][k] %in% top_investors$investors,
                # And if "i"th value in "Investor.label"
                # column is blank, simply insert "1",
                ifelse(
                    is.null(x$Investor.label[i]),
                    x$Investor.label[i] <- 1,
                    # if "i"th value in "Investor.label" column = "1",
                    # keep it, if = "0", replace it with "1"
                    ifelse(
                        x$Investor.label[i] == 1,
                        x$Investor.label[i],
                        x$Investor.label[i] <- 1)
                    # if "k"th investor in "i"th list in dataframe is
                    # not in top_investors$investors, insert "0"
                ), x$Investor.label[i] <- 0
            )
        }
    }
    return(x)
}
df2 <- investor.fun(df2)
df2$Investor.label <- as.character(df2$Investor.label)
```

# Modeling

```r
lm <- lm(data = df2,
        Valuation.Billion ~ Country + City + Industry
        + Founded.Year + Total.Raised.Million + Investors.Count
        + Deal.Terms + Date.Joined_in_months + Investor.label)
summary(lm)
```

```
##
## Call:
## lm(formula = Valuation.Billion ~ Country + City + Industry +
##     Founded.Year + Total.Raised.Million + Investors.Count + Deal.Terms +
##     Date.Joined_in_months + Investor.label, data = df2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.352  -1.393  -0.083   1.181  92.142
##
## Coefficients: (1 not defined because of singularities)
##                                                Estimate Std. Error t value
## (Intercept)                                  -7.458e+01  6.954e+01  -1.072
## CountryBrazil                                -5.438e+00  3.495e+00  -1.556
## CountryCanada                                -4.125e+00  2.722e+00  -1.516
## CountryChina                                 -6.072e+00  2.630e+00  -2.309
## CountryFrance                                -7.160e+00  3.532e+00  -2.027
## CountryGermany                               -4.561e+00  3.232e+00  -1.411
## CountryHong Kong                             -6.818e+00  3.353e+00  -2.033
## CountryIndia                                 -6.373e+00  2.561e+00  -2.488
## CountryIndonesia                             -6.399e+00  3.361e+00  -1.904
## CountryIsrael                                -4.696e+00  2.716e+00  -1.729
## CountryMexico                                -5.840e+00  3.325e+00  -1.756
## CountryNetherlands                           -4.404e+00  3.333e+00  -1.321
## Countryothers                                -5.112e+00  2.473e+00  -2.067
## CountrySingapore                             -7.764e+00  3.187e+00  -2.436
## CountrySouth Korea                           -5.320e+00  2.893e+00  -1.839
## CountrySweden                                -6.716e+00  3.349e+00  -2.005
## CountryUnited Kingdom                        -6.021e+00  3.056e+00  -1.970
## CountryUnited States                         -5.231e+00  2.386e+00  -2.193
## CityBengaluru                                -1.728e+00  2.029e+00  -0.851
## CityBerlin                                   -3.263e+00  2.953e+00  -1.105
## CityBoston                                   -1.604e+00  2.066e+00  -0.776
## CityChicago                                  -2.012e+00  2.056e+00  -0.979
## CityHangzhou                                 -1.770e+00  1.725e+00  -1.026
## CityLondon                                   -7.699e-01  2.590e+00  -0.297
## CityMountain View                            -2.087e+00  2.031e+00  -1.027
## CityNew York                                 -1.265e+00  1.549e+00  -0.817
## Cityothers                                   -1.165e+00  1.364e+00  -0.854
## CityPalo Alto                                -1.495e+00  2.014e+00  -0.742
## CityParis                                     4.394e-01  3.260e+00   0.135
## CityRedwood City                             -2.493e+00  2.238e+00  -1.114
## CitySan Francisco                             1.489e-01  1.499e+00   0.099
## CitySao Paulo                                -1.207e+00  3.378e+00  -0.357
## CityShanghai                                -3.451e-02  1.185e+00  -0.029
## CityShenzhen                                 9.743e-01  1.617e+00   0.603
## CitySingapore                                       NA         NA      NA
## IndustryAuto & transportation               -4.254e+00  1.306e+00  -3.257
## IndustryConsumer & retail                   -1.062e+00  1.367e+00  -0.777
## IndustryCybersecurity                       -1.114e+00  1.073e+00  -1.038
## IndustryData management & analytics         -1.181e-01  1.142e+00  -0.103
## IndustryE-commerce & direct-to-consumer     -1.623e+00  8.799e-01  -1.845
## IndustryEdtech                              -2.275e+00  1.294e+00  -1.758
## IndustryFintech                              4.757e-01  7.901e-01   0.602
## IndustryHardware                            -2.514e+00  1.245e+00  -2.020
## IndustryHealth                              -7.486e-01  9.704e-01  -0.771
## IndustryInternet software & services        -2.727e-01  7.948e-01  -0.343
## IndustryMobile & telecommunications         -1.437e+00  1.173e+00  -1.225
## IndustryOther                               -2.226e-01  1.048e+00  -0.212
## IndustrySupply chain, logistics, & delivery -1.960e+00  1.063e+00  -1.843
## IndustryTravel                              -3.492e+00  1.739e+00  -2.008
## Founded.Year                                 3.956e-02  3.446e-02   1.148
## Total.Raised.Million                         5.855e-03  2.918e-04  20.069
## Investors.Count                             -9.861e-03  2.144e-02  -0.460
## Deal.Terms                                   4.531e-01  1.110e-01   4.083
## Date.Joined_in_months                        2.634e-02  9.783e-03   2.692
## Investor.label1                             -2.784e-01  4.794e-01  -0.581
##                                             Pr(>|t|)
## (Intercept)                                  0.28380
## CountryBrazil                                0.12005
## CountryCanada                                0.12996
## CountryChina                                 0.02116 *
## CountryFrance                                0.04294 *
## CountryGermany                               0.15853
## CountryHong Kong                             0.04230 *
## CountryIndia                                 0.01301 *
## CountryIndonesia                             0.05724 .
## CountryIsrael                                0.08413 .
## CountryMexico                                0.07936 .
## CountryNetherlands                           0.18670
## Countryothers                                0.03903 *
## CountrySingapore                             0.01502 *
## CountrySouth Korea                           0.06627 .
## CountrySweden                                0.04522 *
## CountryUnited Kingdom                        0.04915 *
## CountryUnited States                         0.02857 *
## CityBengaluru                                0.39473
## CityBerlin                                   0.26940
## CityBoston                                   0.43780
## CityChicago                                  0.32805
## CityHangzhou                                 0.30510
## CityLondon                                   0.76632
## CityMountain View                            0.30459
## CityNew York                                 0.41439
## Cityothers                                   0.39337
## CityPalo Alto                                0.45804
## CityParis                                    0.89281
## CityRedwood City                             0.26556
## CitySan Francisco                            0.92087
## CitySao Paulo                                0.72092
## CityShanghai                                 0.97679
## CityShenzhen                                 0.54689
## CitySingapore                                     NA
## IndustryAuto & transportation                0.00117 **
## IndustryConsumer & retail                    0.43738
## IndustryCybersecurity                        0.29959
## IndustryData management & analytics          0.91761
## IndustryE-commerce & direct-to-consumer      0.06535 .
## IndustryEdtech                               0.07903 .
## IndustryFintech                              0.54728
## IndustryHardware                             0.04370 *
## IndustryHealth                               0.44063
## IndustryInternet software & services         0.73156
## IndustryMobile & telecommunications          0.22080
## IndustryOther                                0.83183
## IndustrySupply chain, logistics, & delivery  0.06562 .
## IndustryTravel                               0.04492 *
## Founded.Year                                 0.25127
## Total.Raised.Million                         < 2e-16 ***
## Investors.Count                              0.64571
## Deal.Terms                                  4.83e-05 ***
## Date.Joined_in_months                        0.00723 **
## Investor.label1                              0.56161
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 5.708 on 953 degrees of freedom
## Multiple R-squared:  0.4387, Adjusted R-squared:  0.4075
## F-statistic: 14.05 on 53 and 953 DF,  p-value: < 2.2e-16
```

```
summary(stepAIC(lm))
```

```
## Start:  AIC=3560.69
## Valuation.Billion ~ Country + City + Industry + Founded.Year +
##     Total.Raised.Million + Investors.Count + Deal.Terms + Date.Joined_in_months +
##     Investor.label
##
##                         Df Sum of Sq   RSS    AIC
## - Country               16     306.8 31359 3538.6
## - City                  16     372.9 31425 3540.7
## - Investors.Count        1       6.9 31059 3558.9
## - Investor.label         1      11.0 31063 3559.0
## - Founded.Year           1      42.9 31095 3560.1
## <none>                               31052 3560.7
## - Industry              14     992.2 32044 3564.4
## - Date.Joined_in_months  1     236.1 31288 3566.3
## - Deal.Terms             1     543.1 31595 3576.2
## - Total.Raised.Million   1   13123.8 44176 3913.7
##
## Step:  AIC=3538.59
## Valuation.Billion ~ City + Industry + Founded.Year + Total.Raised.Million +
##     Investors.Count + Deal.Terms + Date.Joined_in_months + Investor.label
```

```
##
##                        Df Sum of Sq   RSS    AIC
## - City                 17     464.3 31823 3519.4
## - Investors.Count       1       4.3 31363 3536.7
## - Investor.label        1       6.3 31365 3536.8
## - Founded.Year          1      37.7 31397 3537.8
## <none>                              31359 3538.6
## - Date.Joined_in_months  1     219.1 31578 3543.6
## - Industry             14    1128.8 32488 3546.2
## - Deal.Terms            1     610.6 31970 3556.0
## - Total.Raised.Million   1   13229.1 44588 3891.0
##
## Step:  AIC=3519.39
## Valuation.Billion ~ Industry + Founded.Year + Total.Raised.Million +
##     Investors.Count + Deal.Terms + Date.Joined_in_months + Investor.label
##
##                        Df Sum of Sq   RSS    AIC
## - Investors.Count       1       0.2 31823 3517.4
## - Investor.label        1       0.3 31824 3517.4
## - Founded.Year          1      51.4 31875 3519.0
## <none>                              31823 3519.4
## - Date.Joined_in_months  1     337.2 32160 3528.0
## - Industry             14    1388.3 33212 3534.4
## - Deal.Terms            1     586.7 32410 3535.8
## - Total.Raised.Million   1   13470.7 45294 3872.8
##
## Step:  AIC=3517.4
## Valuation.Billion ~ Industry + Founded.Year + Total.Raised.Million +
##     Deal.Terms + Date.Joined_in_months + Investor.label
##
##                        Df Sum of Sq   RSS    AIC
## - Investor.label        1       0.3 31824 3515.4
## - Founded.Year          1      51.7 31875 3517.0
## <none>                              31823 3517.4
## - Date.Joined_in_months  1     341.2 32165 3526.1
## - Industry             14    1388.1 33212 3532.4
## - Deal.Terms            1     678.3 32502 3536.6
## - Total.Raised.Million   1   13930.3 45754 3881.0
##
## Step:  AIC=3515.41
## Valuation.Billion ~ Industry + Founded.Year + Total.Raised.Million +
##     Deal.Terms + Date.Joined_in_months
##
##                        Df Sum of Sq   RSS    AIC
## - Founded.Year          1      51.5 31875 3515.0
## <none>                              31824 3515.4
## - Date.Joined_in_months  1     341.3 32165 3524.2
## - Industry             14    1388.2 33212 3530.4
## - Deal.Terms            1     678.1 32502 3534.6
## - Total.Raised.Million   1   13939.7 45763 3879.2
##
## Step:  AIC=3515.04
## Valuation.Billion ~ Industry + Total.Raised.Million + Deal.Terms +
##     Date.Joined_in_months
```

```
##
##                         Df Sum of Sq   RSS    AIC
## <none>                               31875 3515.0
## - Date.Joined_in_months  1     291.5 32167 3522.2
## - Industry              14    1361.5 33237 3529.2
## - Deal.Terms             1     664.9 32540 3533.8
## - Total.Raised.Million   1   13915.6 45791 3877.8
```

```
##
## Call:
## lm(formula = Valuation.Billion ~ Industry + Total.Raised.Million +
##     Deal.Terms + Date.Joined_in_months, data = df2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.384  -1.472  -0.035   1.085  93.101
##
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                -1.2536815  0.7095097  -1.767
## IndustryAuto & transportation              -4.5738283  1.2320611  -3.712
## IndustryConsumer & retail                  -1.2299663  1.3312241  -0.924
## IndustryCybersecurity                      -0.9193939  1.0352440  -0.888
## IndustryData management & analytics        -0.3179762  1.1131008  -0.286
## IndustryE-commerce & direct-to-consumer     -1.9614410  0.8407636  -2.333
## IndustryEdtech                             -2.4688360  1.2507056  -1.974
## IndustryFintech                             0.4073257  0.7539162   0.540
## IndustryHardware                           -2.4284497  1.2057689  -2.014
## IndustryHealth                             -0.8947252  0.9393314  -0.953
## IndustryInternet software & services       -0.0778442  0.7662880  -0.102
## IndustryMobile & telecommunications        -1.3868557  1.1420798  -1.214
## IndustryOther                              -0.4432245  1.0110104  -0.438
## IndustrySupply chain, logistics, & delivery -2.2320149  1.0349054  -2.157
## IndustryTravel                             -4.1059883  1.6503209  -2.488
## Total.Raised.Million                        0.0057679  0.0002776  20.779
## Deal.Terms                                  0.4167047  0.0917409   4.542
## Date.Joined_in_months                       0.0250729  0.0083369   3.007
##                                            Pr(>|t|)
## (Intercept)                                0.077542 .
## IndustryAuto & transportation              0.000217 ***
## IndustryConsumer & retail                  0.355745
## IndustryCybersecurity                      0.374706
## IndustryData management & analytics        0.775193
## IndustryE-commerce & direct-to-consumer     0.019852 *
## IndustryEdtech                             0.048665 *
## IndustryFintech                            0.589126
## IndustryHardware                           0.044277 *
## IndustryHealth                             0.341070
## IndustryInternet software & services       0.919106
## IndustryMobile & telecommunications        0.224914
## IndustryOther                              0.661194
## IndustrySupply chain, logistics, & delivery 0.031267 *
## IndustryTravel                             0.013010 *
## Total.Raised.Million                        < 2e-16 ***
## Deal.Terms                                  6.25e-06 ***
## Date.Joined_in_months                       0.002701 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 5.677 on 989 degrees of freedom
## Multiple R-squared:  0.4238, Adjusted R-squared:  0.4139
## F-statistic: 42.79 on 17 and 989 DF,  p-value: < 2.2e-16
```

Before we conclude the output, we have to be careful about the limitation of linear model. Applying a Linear model to these compled data is not the optimal method since it does not take in account its collinearity and the risk of incorporating categorical values. According to the output, Industry is actually significant while Country, City, and Investor are insignificant. Adjusted R-squared: 0.4139 and AIC = 3515.04 tells us the model fit is not good.