608-HW1

Jose Mawyin

2/8/2020

Principles of Data Visualization and Introduction to ggplot2

I have provided you with data about the 5,000 fastest growing companies in the US, as compiled by Inc. magazine. lets read this in:

```
inc <- read.csv("https://raw.githubusercontent.com/charleyferrari/CUNY_DATA_608/master/module1/Data/inc5
library(ggplot2)
library(magrittr)
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
```

And lets preview this data:

```
head(inc)
```

```
##   Rank                          Name Growth_Rate   Revenue
## 1    1                          Fuhu      421.48 1.179e+08
## 2    2       FederalConference.com      248.31 4.960e+07
## 3    3               The HCI Group      245.45 2.550e+07
## 4    4                     Bridger      233.08 1.900e+09
## 5    5                      DataXu      213.37 8.700e+07
## 6    6 MileStone Community Builders      179.38 4.570e+07
##                       Industry Employees         City State
## 1 Consumer Products & Services       104   El Segundo    CA
## 2          Government Services        51     Dumfries    VA
## 3                       Health       132 Jacksonville    FL
## 4                       Energy        50      Addison    TX
## 5      Advertising & Marketing       220       Boston    MA
## 6                  Real Estate        63       Austin    TX
```

```
summary(inc)
```

```
##       Rank                          Name       Growth_Rate
##  Min.   :   1   (Add)ventures         :   1   Min.   :  0.340
##  1st Qu.:1252   @Properties           :   1   1st Qu.:  0.770
```

```
##  Median :2502   1-Stop Translation USA:   1   Median :  1.420
##  Mean   :2502   110 Consulting        :   1   Mean   :  4.612
##  3rd Qu.:3751   11thStreetCoffee.com  :   1   3rd Qu.:  3.290
##  Max.   :5000   123 Exteriors         :   1   Max.   :421.480
##                 (Other)               :4995
##     Revenue                                  Industry      Employees
##  Min.   :2.000e+06   IT Services                 : 733   Min.   :    1.0
##  1st Qu.:5.100e+06   Business Products & Services: 482   1st Qu.:   25.0
##  Median :1.090e+07   Advertising & Marketing     : 471   Median :   53.0
##  Mean   :4.822e+07   Health                      : 355   Mean   :  232.7
##  3rd Qu.:2.860e+07   Software                    : 342   3rd Qu.:  132.0
##  Max.   :1.010e+10   Financial Services          : 260   Max.   :66803.0
##                      (Other)                     :2358   NA's   :12
##             City          State
##  New York     : 160   CA     : 701
##  Chicago      :  90   TX     : 387
##  Austin       :  88   NY     : 311
##  Houston      :  76   VA     : 283
##  San Francisco:  75   FL     : 282
##  Atlanta      :  74   IL     : 273
##  (Other)      :4438   (Other):2764
```

Think a bit on what these summaries mean. Use the space below to add some more relevant non-visual exploratory information you think helps you understand this data:

```
#How many companies are there per state?
state.count <- inc %>% count(State)
state.count <- state.count[with(state.count, order(-n)), ]
head(state.count,10)
```

```
## # A tibble: 10 x 2
##    State     n
##    <fct> <int>
##  1 CA      701
##  2 TX      387
##  3 NY      311
##  4 VA      283
##  5 FL      282
##  6 IL      273
##  7 GA      212
##  8 OH      186
##  9 MA      182
## 10 PA      164
```

```
#What is the total revenue generated by this fast growing companies per state.
state.income <- aggregate(inc$Revenue, by=list(Category=inc$State), FUN=sum)
state.income <- state.income[with(state.income, order(-x)), ]
head(state.income,10)
```

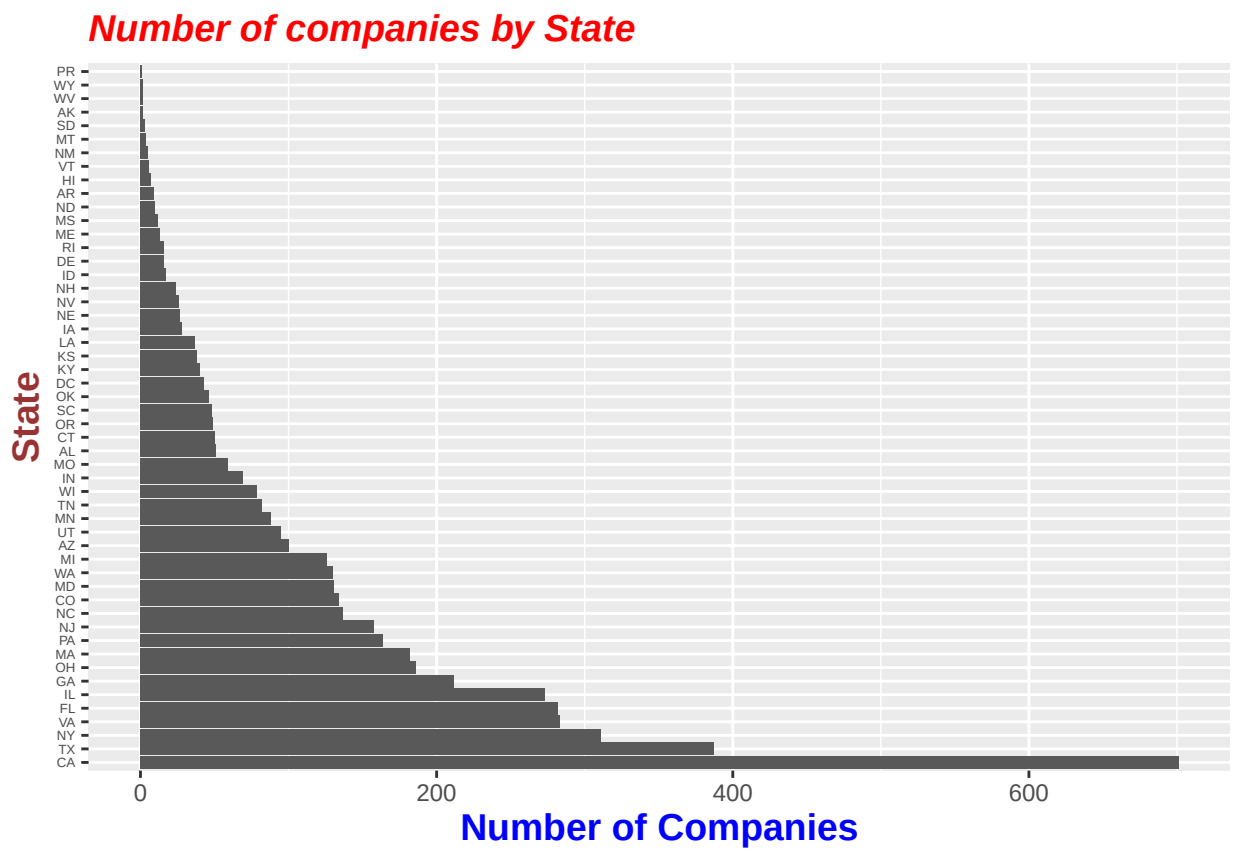
```
##    Category           x
## 15       IL 33244300000
## 5        CA 23457900000
## 45       TX 22164200000
```

```
## 35       NY 18260400000
## 36       OH 12786600000
## 10       FL 10610300000
## 28       NC  9258500000
## 47       VA  8667700000
## 23       MI  7805800000
## 50       WI  7296600000
```

Question 1

Create a graph that shows the distribution of companies in the dataset by State (ie how many are in each state). There are a lot of States, so consider which axis you should use. This visualization is ultimately going to be consumed on a 'portrait' oriented screen (ie taller than wide), which should further guide your layout choices.

```
state.count$State <- factor(state.count$State, levels = state.count$State[order(-state.count$n)])
state.count <- arrange(state.count, n)
g <- ggplot(data=state.count, aes(x=State, y=n))
g + geom_bar(stat="identity") + coord_flip() + ggtitle("Number of companies by State") +
  ylab("Number of Companies") + xlab("State") + theme(
plot.title = element_text(color="red", size=14, face="bold.italic"),
axis.title.x = element_text(color="blue", size=14, face="bold"),
axis.title.y = element_text(color="#993333", size=14, face="bold"),
axis.text.y = element_text(size = 5))
```



Question 2

Lets dig in on the state with the 3rd most companies in the data set. Imagine you work for the state and are interested in how many people are employed by companies in different industries. Create a plot that shows the average and/or median employment by industry for companies in this state (only use cases with full data, use R's `complete.cases()` function.) In addition to this, your graph should show how variable the ranges are, and you should deal with outliers.

From the ordered graph above, we see that the state with the 3rd most companies is New York.

```
NY.companies <- filter(inc, State == "NY")
nrow(NY.companies)
```

```
## [1] 311
```

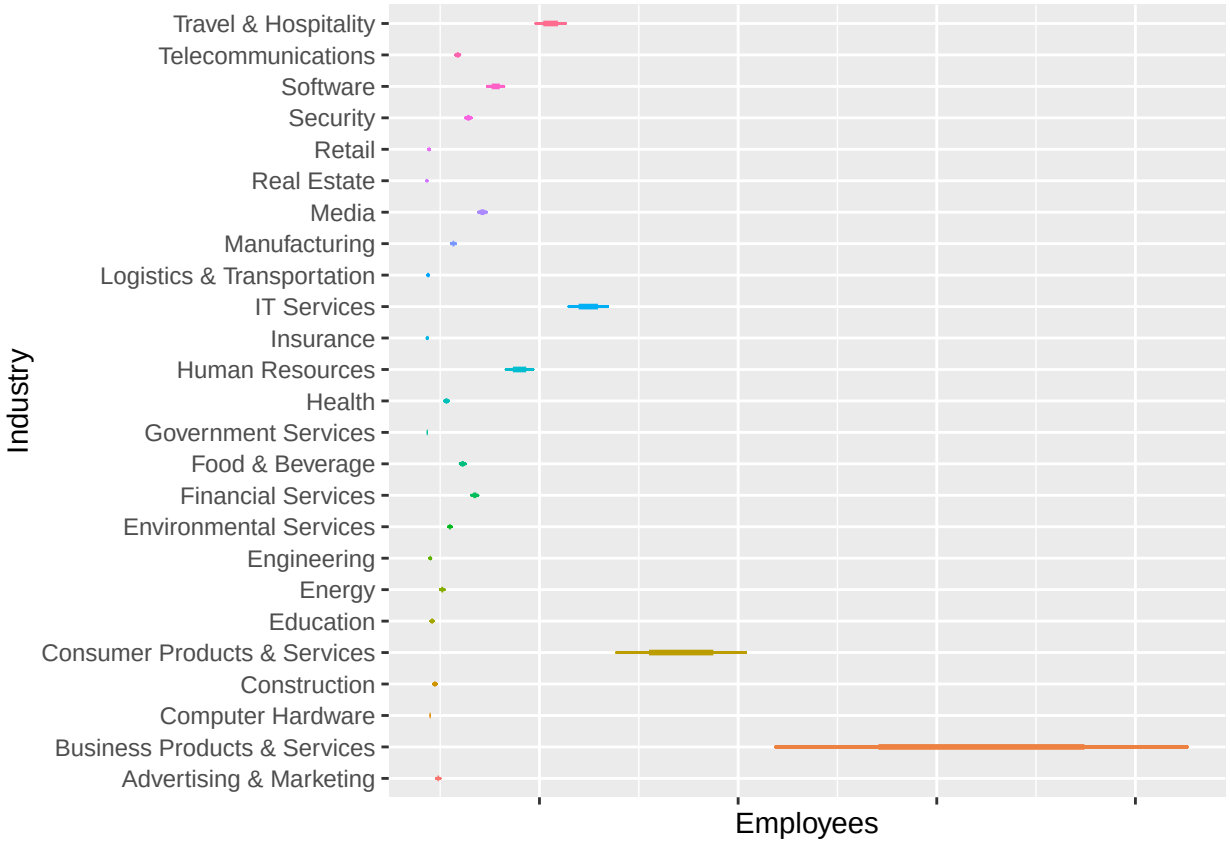
```
NY.companies <- NY.companies[complete.cases(NY.companies[1:8]),]
NY.companies$Employees <- as.double(NY.companies$Employees)
head(NY.companies)
```

```
##   Rank                     Name Growth_Rate  Revenue
## 1   26             BeenVerified       84.43 13700000
## 2   30                 Sailthru       73.22  8100000
## 3   37             YellowHammer       67.40 18000000
## 4   38                Conductor       67.02  7100000
## 5   48 Cinium Financial Services       53.65  5900000
## 6   70                 33Across       44.99 27900000
##                       Industry Employees      City State
## 1 Consumer Products & Services        17  New York    NY
## 2      Advertising & Marketing        79  New York    NY
## 3      Advertising & Marketing        27  New York    NY
## 4      Advertising & Marketing        89  New York    NY
## 5           Financial Services        32 Rock Hill    NY
## 6      Advertising & Marketing        75  New York    NY
```

```
NY.industry.count <- NY.companies %>% count(Industry)

d <- ggplot(NY.companies, aes(x = Employees,  y = Industry, color=Industry))

d + geom_boxplot(outlier.shape=2, aes(group=Industry), alpha = 0.3,outlier.size=16, notch=TRUE)  + theme
```



Using a boxplot on the New York State per industry we can show the median number of employees per industry and the range of employnmeent in these industries. We see how the fastest growing companies with the most employees are in the Business Products and Services as compared to smaller employers such as Retail, Real State, etc.

Question 3

Now imagine you work for an investor and want to see which industries generate the most revenue per employee. Create a chart that makes this information clear. Once again, the distribution per industry should be shown.

```
NY.companies$rev.per.empl <- cbind(NY.companies$Revenue/NY.companies$Employees)
head(NY.companies)
```

```
##   Rank                     Name Growth_Rate  Revenue
## 1   26              BeenVerified       84.43 13700000
## 2   30                  Sailthru       73.22  8100000
## 3   37              YellowHammer       67.40 18000000
## 4   38                 Conductor       67.02  7100000
## 5   48 Cinium Financial Services       53.65  5900000
## 6   70                  33Across       44.99 27900000
##                       Industry Employees      City State rev.per.empl
## 1 Consumer Products & Services        17  New York    NY    805882.35
## 2      Advertising & Marketing        79  New York    NY    102531.65
## 3      Advertising & Marketing        27  New York    NY    666666.67
```

```
## 4      Advertising & Marketing        89  New York    NY     79775.28
## 5           Financial Services        32 Rock Hill    NY    184375.00
## 6      Advertising & Marketing        75  New York    NY    372000.00
```

```r
# Answer Question 3 here
ggplot(NY.companies, aes(x=factor(Industry), y=rev.per.empl, color=Industry)) + stat_summary(fun.y="medi
  xlab("Industry") + ylab("Median Revenue per Employee")
```