Data 608 - Module 1 Assignment

Zhouxin Shi

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
```

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         Revenue
##  Min.   :   1   Length:5001        Min.   :  0.340   Min.   :2.000e+06
##  1st Qu.:1252   Class :character   1st Qu.:  0.770   1st Qu.:5.100e+06
##  Median :2502   Mode  :character   Median :  1.420   Median :1.090e+07
##  Mean   :2502                      Mean   :  4.612   Mean   :4.822e+07
##  3rd Qu.:3751                      3rd Qu.:  3.290   3rd Qu.:2.860e+07
##  Max.   :5000                      Max.   :421.480   Max.   :1.010e+10
##
##    Industry           Employees          City              State
##  Length:5001        Min.   :    1.0   Length:5001        Length:5001
##  Class :character   1st Qu.:   25.0   Class :character   Class :character
##  Mode  :character   Median :   53.0   Mode  :character   Mode  :character
##                     Mean   :  232.7
##                     3rd Qu.:  132.0
##                     Max.   :66803.0
##                     NA's   :12
```

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:

```
str(inc)
```

```
## 'data.frame':    5001 obs. of  8 variables:
##  $ Rank       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Name       : chr  "Fuhu" "FederalConference.com" "The HCI Group" "Bridger" ...
##  $ Growth_Rate: num  421 248 245 233 213 ...
##  $ Revenue    : num  1.18e+08 4.96e+07 2.55e+07 1.90e+09 8.70e+07 ...
##  $ Industry   : chr  "Consumer Products & Services" "Government Services" "Health" "Energy" ...
##  $ Employees  : int  104 51 132 50 220 63 27 75 97 15 ...
##  $ City       : chr  "El Segundo" "Dumfries" "Jacksonville" "Addison" ...
##  $ State      : chr  "CA" "VA" "FL" "TX" ...
```

```
library(sqldf)
```

```
## Loading required package: gsubfn
```

```
## Loading required package: proto
```

```
## Loading required package: RSQLite
```

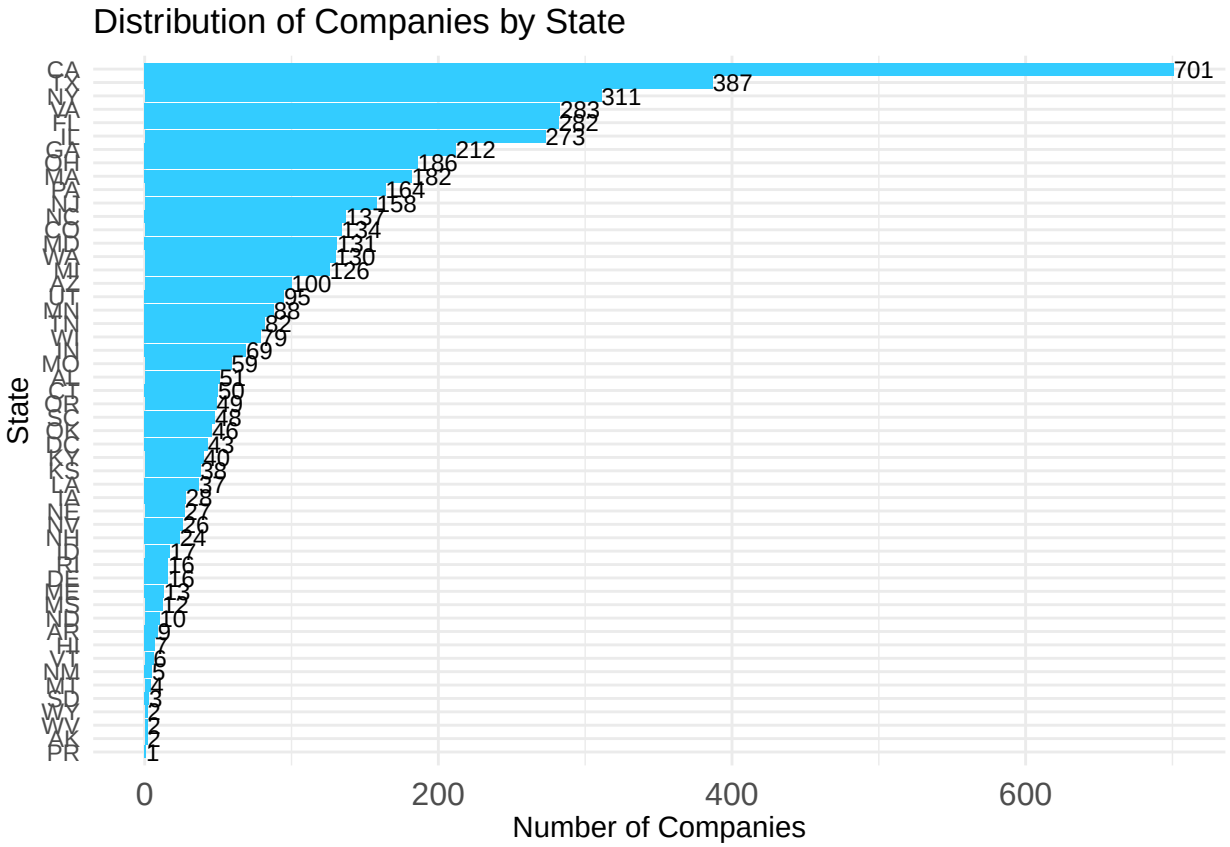
```
library(ggplot2)
```

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.

```
q1<-sqldf("select State
                ,count(distinct Name) as cnt_names
          from inc
          group by State")
ggplot(data = q1, aes(x=reorder(State,cnt_names),y = cnt_names))+
  geom_bar(stat="identity", fill="#33CCFF")+
  geom_text(aes(label=cnt_names), vjust=0.5, size=3, position=position_dodge(width = 2), hjust=0)+
  theme_minimal()+
  theme(axis.text.y=element_text(size=9, vjust=0.5))+
  theme(axis.text.x=element_text(size=12, vjust=0.5))+
  labs( x="State", y="Number of Companies")+
  coord_flip()+
  ggtitle("Distribution of Companies by State")
```

```
## Warning: position_dodge requires non-overlapping x intervals
```



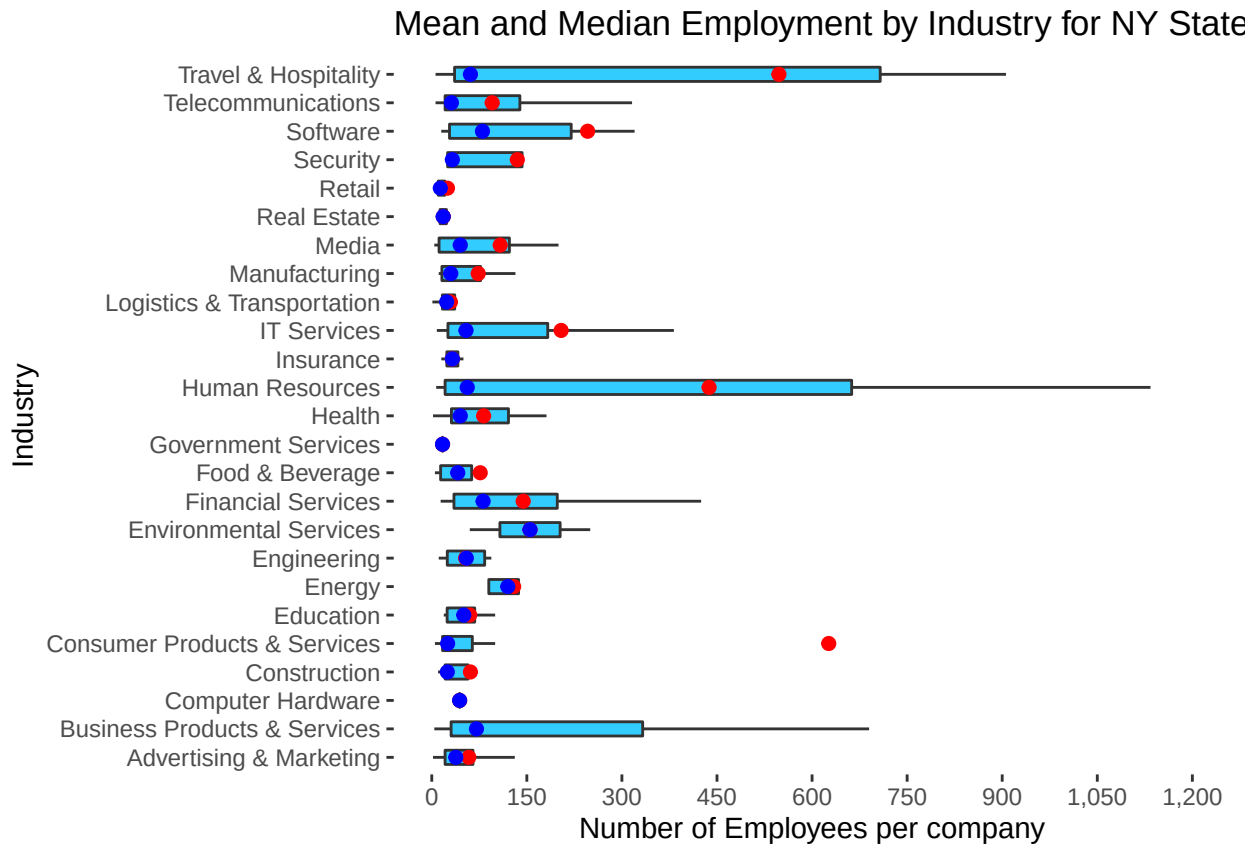
Quesiton 2

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.

```
inc_comp<- inc[complete.cases(inc), ]
q2 <- sqldf("select *from inc_comp where State = 'NY'")
ggplot(q2, aes(x=Industry, y=Employees)) +
    geom_boxplot(width=.5, fill="#33CCFF", outlier.colour=NA) +
    stat_summary(aes(colour = "mean"), fun.y = mean, geom="point", fill="red",
                 colour="red", shape=21, size=2, show.legend=TRUE) +
    stat_summary(aes(colour = "median"), fun.y = median, geom="point", fill="blue",
                 colour="blue", shape=21, size=2, show.legend=TRUE) +
    coord_flip(ylim = c(0, 1200), expand = TRUE) +
    scale_y_continuous(labels = scales::comma,
                       breaks = seq(0, 1500, by = 150)) +
    xlab("Industry") +
    ylab("Number of Employees per company") +
    ggtitle("Mean and Median Employment by Industry for NY State") +
    theme(panel.background = element_blank(), legend.position = "top")
```

```
## Warning: `fun.y` is deprecated. Use `fun` instead.
```

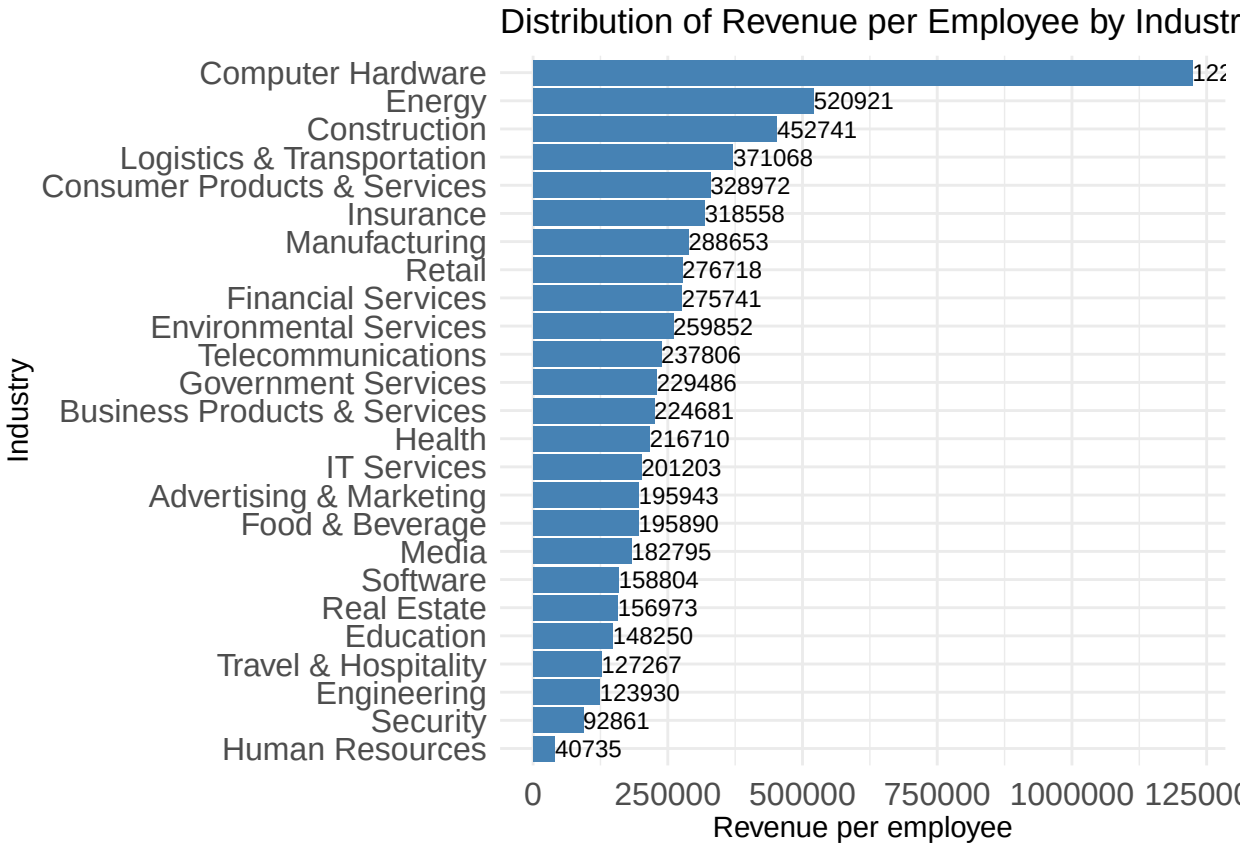
```
## Warning: `fun.y` is deprecated. Use `fun` instead.
```



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.

```
q3 <- sqldf("select Industry
, round(sum(Revenue)/sum(Employees),0) as revenue_per_employee
              from inc
group by Industry
order by sum(Revenue)/sum(Employees) desc")
ggplot(data = q3, aes(x=reorder(Industry,revenue_per_employee),y = revenue_per_employee))+
  geom_bar(stat="identity",  fill="steelblue")+
  geom_text(aes(label=round(revenue_per_employee, digits=0)), vjust=0.5, size=3, position=position_dodg
  theme_minimal()+
  theme(axis.text.y=element_text(size=12, vjust=0.5))+
  theme(axis.text.x=element_text(size=12, vjust=0.5))+
  labs( x="Industry", y="Revenue per employee")+
  coord_flip()+
  ggtitle("Distribution of Revenue per Employee by Industry")
```


Distribution of Revenue per Employee by Industr
Industry
Computer Hardware
Energy
Construction
Logistics & Transportation
Consumer Products & Services
Insurance
Manufacturing
Retail
Financial Services
Environmental Services
Telecommunications
Government Services
Business Products & Services
Health
IT Services
Advertising & Marketing
Food & Beverage
Media
Software
Real Estate
Education
Travel & Hospitality
Engineering
Security
Human Resources
122
520921
452741
371068
328972
318558
288653
276718
275741
259852
237806
229486
224681
216710
201203
195943
195890
182795
158804
156973
148250
127267
123930
92861
40735
0
250000
500000
750000
1000000
125000
Revenue per employee