Data 608 - Module 1

Gabriel Santos

2023-02-08

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
```

```
library(tidyverse)
library(openintro)
library(ggplot2)
```

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:

```
dim(inc)
```

```
## [1] 5001    8
```

```
names(inc)
```

```
## [1] "Rank"        "Name"        "Growth_Rate" "Revenue"     "Industry"   
## [6] "Employees"   "City"        "State"
```

```
Indus <- inc %>%
  dplyr::select(Industry)
table(Indus)
```

```
## Industry
##      Advertising & Marketing Business Products & Services 
##                          471                          482 
##            Computer Hardware                 Construction 
##                           44                          187 
## Consumer Products & Services                    Education 
##                          203                           83 
##                       Energy                  Engineering 
##                          109                           74 
##       Environmental Services           Financial Services 
##                           51                          260 
##              Food & Beverage          Government Services 
##                          131                          202 
##                       Health              Human Resources 
##                          355                          196 
##                    Insurance                  IT Services 
##                           50                          733 
##   Logistics & Transportation                Manufacturing 
##                          155                          256 
##                        Media                  Real Estate 
##                           54                           96 
##                       Retail                     Security 
##                          203                           73 
##                     Software           Telecommunications 
##                          342                          129 
##         Travel & Hospitality 
##                           62
```

Question 1

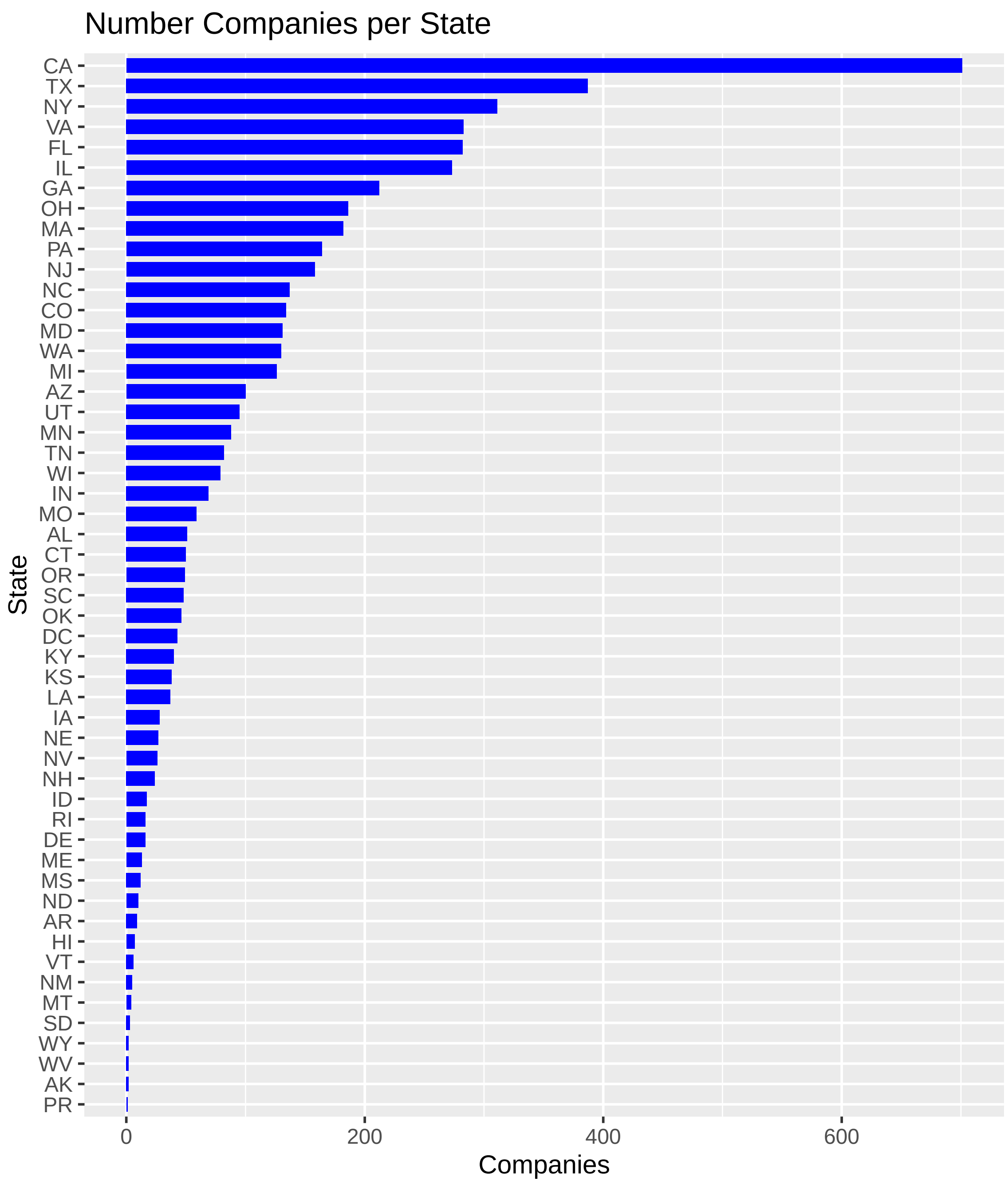
Create a graph that shows the distribution of companies in the dataset by State (ie how many are in each state). There are a lot of States, so consider which axis you should use. This visualization is ultimately going to be consumed on a 'portrait' oriented screen (ie taller than wide), which should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

```
distribcompanies <- inc %>%
  dplyr::select(State)
table(distribcompanies)
```

```
## State
##  AK  AL  AR  AZ  CA  CO  CT  DC  DE  FL  GA  HI  IA  ID  IL  IN  KS  KY  LA  MA
##   2  51   9 100 701 134  50  43  16 282 212   7  28  17 273  69  38  40  37 182
##  MD  ME  MI  MN  MO  MS  MT  NC  ND  NE  NH  NJ  NM  NV  NY  OH  OK  OR  PA  PR
## 131  13 126  88  59  12   4 137  10  27  24 158   5  26 311 186  46  49 164   1
##  RI  SC  SD  TN  TX  UT  VA  VT  WA  WI  WV  WY
##  16  48   3  82 387  95 283   6 130  79   2   2
```

```
distribcompplot <- inc %>%
  group_by(State) %>%
  count(State) %>%
  arrange(desc(n)) %>%
  as_tibble(distribcompplot)

ggplot(distribcompplot, aes(x=reorder(State,n), y=n)) +
    geom_bar(stat="identity", fill="blue", width=0.7) +
    coord_flip() +
    xlab("State") + ylab("Companies") +
    ggtitle("Number Companies per State")
```



Question 2

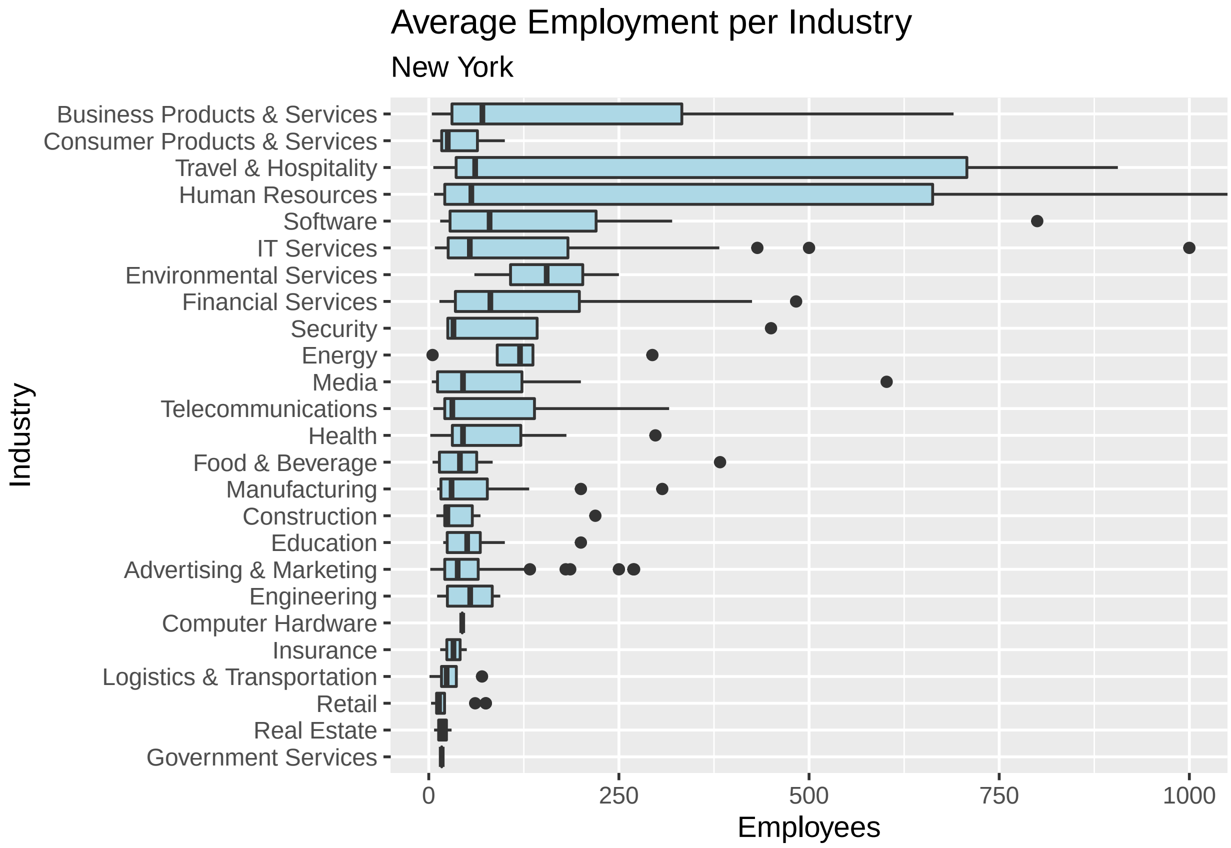
Lets dig in on the state with the 3rd most companies in the data set. Imagine you work for the state and are interested in how many people are employed by companies in different industries. Create a plot that shows the average and/or median employment by industry for companies in this state (only use 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.

```
ny_total <- filter(inc, (State=="NY"))
summary(ny_total)
```

```
##       Rank          Name            Growth_Rate        Revenue
##  Min.   :  26   Length:311         Min.   : 0.350   Min.   :2.000e+06
##  1st Qu.:1186   Class :character   1st Qu.: 0.670   1st Qu.:4.300e+06
##  Median :2702   Mode  :character   Median : 1.310   Median :8.800e+06
##  Mean   :2612                      Mean   : 4.371   Mean   :5.872e+07
##  3rd Qu.:4005                      3rd Qu.: 3.580   3rd Qu.:2.570e+07
##  Max.   :4981                      Max.   :84.430   Max.   :4.600e+09
##    Industry           Employees          City              State
##  Length:311         Min.   :    1.0   Length:311         Length:311
##  Class :character   1st Qu.:   21.0   Class :character   Class :character
##  Mode  :character   Median :   45.0   Mode  :character   Mode  :character
##                     Mean   :  271.3
##                     3rd Qu.:  105.5
##                     Max.   :32000.0
```

```
ny_industry <- ny_total %>%
  filter(complete.cases(.)) %>%
  group_by(Industry) %>%
  dplyr::select(Industry, Employees)

ggplot(ny_industry, aes(x=reorder(Industry,Employees), y=Employees)) +
    geom_boxplot(fill="lightblue") + xlab("Employees") +
    theme(legend.position="none") +
    xlab("Industry") + ylab("Employees") +
    coord_flip(ylim= c(0, 1000)) +
    ggtitle("Average Employment per Industry", subtitle = "New York")
```

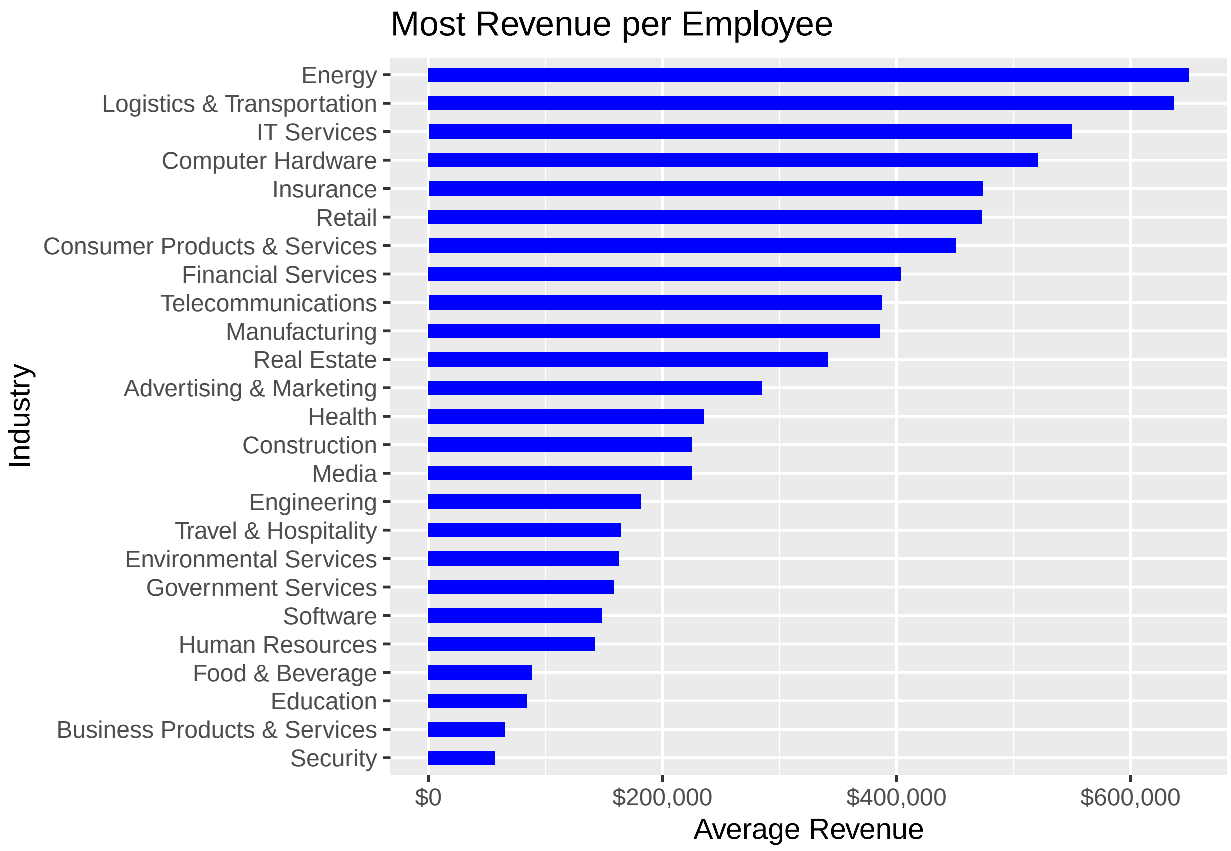


Question 3

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_revenue <- ny_total %>%
  group_by(Industry) %>%
  summarize(total_revenue = sum(Revenue), total_employee = sum(Employees), avg_revenue = total_revenue/
  arrange(desc(avg_revenue)) %>%
  na.omit()


ggplot(ny_revenue, aes(x=reorder(Industry,avg_revenue), y=avg_revenue)) +
    geom_bar(stat="identity", fill="blue", width=0.5) +
    coord_flip() +
    xlab("Industry") + ylab("Average Revenue") +
    ggtitle("Most Revenue per Employee") +
    scale_y_continuous(labels = scales::label_dollar())
```



Conclusion

According to the graphs we can see that the state with the largest number of companies is California, followed by Texas and New York in third place.In New York State, the industries that employ the most staff are travel *and hospitality and then Human Resources. In the same state of New York, the industry that generates the highest amount of income per employee is Energy, followed by Logistics & Transportation.*