Data 608: Assignment 1

Bharani Nittala

2022-02-12

```
suppressMessages(suppressWarnings(library(tidyverse)))
suppressMessages(suppressWarnings(library(ggplot2)))
suppressMessages(suppressWarnings(library(RColorBrewer)))
```

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:

```
#Understand the structure of the dataframe apart from summary statistics
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" ...
```

```
#Understand mean, median and standard deviation of the dataframe:

mean(inc$Growth_Rate)
```

```
## [1] 4.611826
```

```
mean(inc$Revenue)
```

```
## [1] 48222535
```

```
mean(inc$Employees, na.rm = TRUE)#A few companies have missing employee counts
```

```
## [1] 232.718
```

```
median(inc$Growth_Rate)
```

```
## [1] 1.42
```

```
median(inc$Revenue)
```

```
## [1] 10900000
```

```
median(inc$Employees, na.rm = TRUE)#A few companies have missing employee counts
```

```
## [1] 53
```

```
sd(inc$Growth_Rate)
```

```
## [1] 14.12369
```

```
sd(inc$Revenue)
```

```
## [1] 240542281
```

```
sd(inc$Employees, na.rm = TRUE)#A few companies have missing employee counts
```

```
## [1] 1353.128
```

```
#We can also do IQR(Q3-Q1) in case the data is skewed

IQR(inc$Growth_Rate)
```

```
## [1] 2.52
```

```
IQR(inc$Revenue)
```

```
## [1] 23500000
```

```
IQR(inc$Employees, na.rm = TRUE)
```

```
## [1] 107
```

Question 1

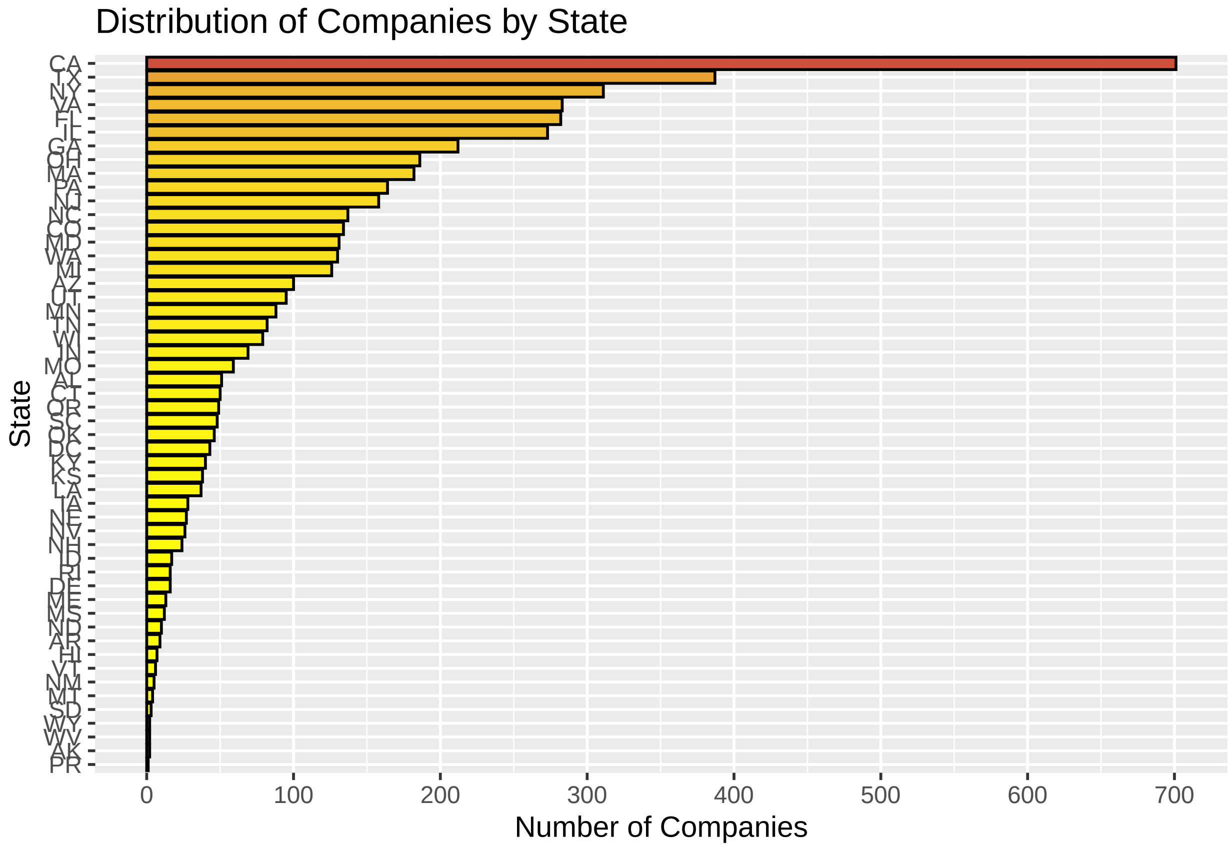
Create a graph that shows the distribution of companies in the dataset by State (ie how many are in each state). There are a lot of States, so consider which axis you should use. This visualization is ultimately going to be consumed on a 'portrait' oriented screen (ie taller than wide), which should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

For all three questions, we are asked to give distributions of categorical data: States or Industries. The simple Bar Chart seems to be the most intuitive way to present the information. I used horizontal bars to make the charts fit the portrait orientation better.

```
suppressMessages(suppressWarnings(library(tidyverse)))
suppressMessages(suppressWarnings(library(ggplot2)))
suppressMessages(suppressWarnings(library(RColorBrewer)))

state_level  <- inc %>% group_by(State) %>% summarise(total = n()) %>% arrange(desc(total))

q1 <- ggplot(data = state_level, aes(x=reorder(State, total) , y=total, fill=total)) +
      geom_bar(stat="identity", position=position_dodge(), colour="black", width = 0.9) +
      coord_flip() +   scale_fill_gradient(low="yellow", high="tomato3") + scale_y_continuous(breaks = s
      guides(fill=FALSE) +
      ggtitle("Distribution of Companies by State") +
      xlab("State") + ylab("Number of Companies")
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```

```
q1
```



I think it's key to understand the most important signal from the dataframe than having mixed signals with rainbow colors popping from the chart. Hence, I used color combination that makes the important signal pop in the chart.

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.

New York is in third place for most companies in the dataframe. To show variable ranges, I leveraged mean and median and filtered the outliers that are 1.5 times or higher than IQR.

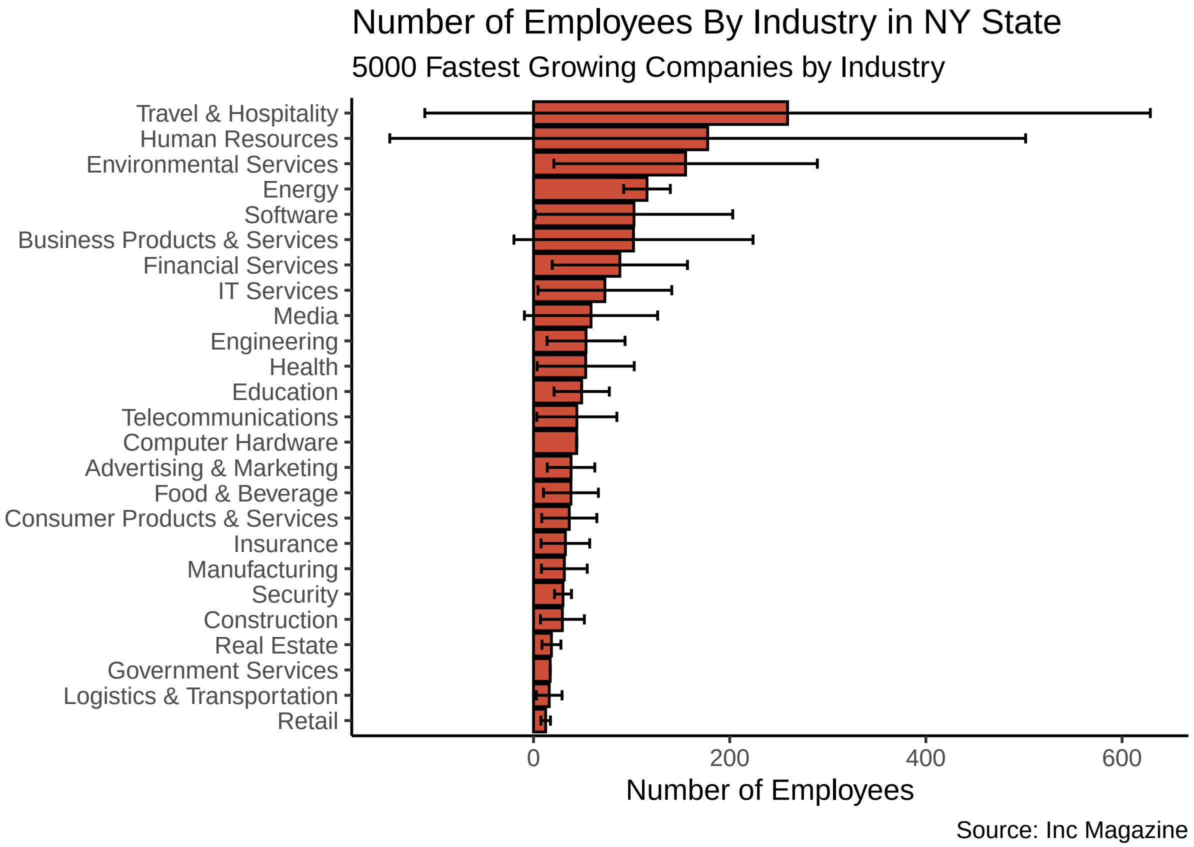
```r
library(ggplot2)

ny_state <- inc  %>% filter(State == 'NY', complete.cases(.)) %>% arrange(Industry) %>% select(Industry
ny_state <- ny_state %>% group_by(Industry) %>% filter(!(abs(Employees - median(Employees)) > 1.5*IQR(Em
ny_state
```

```
## # A tibble: 262 x 2
## # Groups:   Industry [25]
##    Industry               Employees
##    <chr>                      <int>
##  1 Advertising & Marketing       79
##  2 Advertising & Marketing       27
##  3 Advertising & Marketing       89
##  4 Advertising & Marketing       75
##  5 Advertising & Marketing       42
##  6 Advertising & Marketing       15
##  7 Advertising & Marketing       46
##  8 Advertising & Marketing       19
##  9 Advertising & Marketing       45
## 10 Advertising & Marketing       12
## # ... with 252 more rows
```

```r
industry_means <- ny_state %>% group_by(Industry) %>% summarise(mean_emp = mean(Employees), emp_sd = sd
industry_means$emp_sd[is.na(industry_means$emp_sd)] <- 0
industry_means
```

```
## # A tibble: 25 x 3
##    Industry                     mean_emp emp_sd
##    <chr>                           <dbl>  <dbl>
##  1 Advertising & Marketing          38.2   24.2
##  2 Business Products & Services    102.   122.
##  3 Computer Hardware                44      0
##  4 Construction                     29.4   22.4
##  5 Consumer Products & Services     36.5   28.1
##  6 Education                        49.1   28.2
##  7 Energy                          116.    23.8
##  8 Engineering                      53.5   39.8
##  9 Environmental Services          155    134.
## 10 Financial Services               88     68.9
## # ... with 15 more rows
```

```r
ggplot(industry_means, aes(x=reorder(Industry, mean_emp),y=mean_emp)) +
  geom_bar(stat='identity', color = 'black', fill="tomato3") +
  geom_errorbar(aes(ymin = mean_emp - emp_sd, ymax = mean_emp + emp_sd), width=0.4) +
  theme(legend.position="none") +
      labs(title="Number of Employees By Industry in NY State",
         subtitle="5000 Fastest Growing Companies by Industry",
         caption="Source: Inc Magazine",
         y="Number of Employees",
         x="") +
  coord_flip() +
  theme_classic()
```



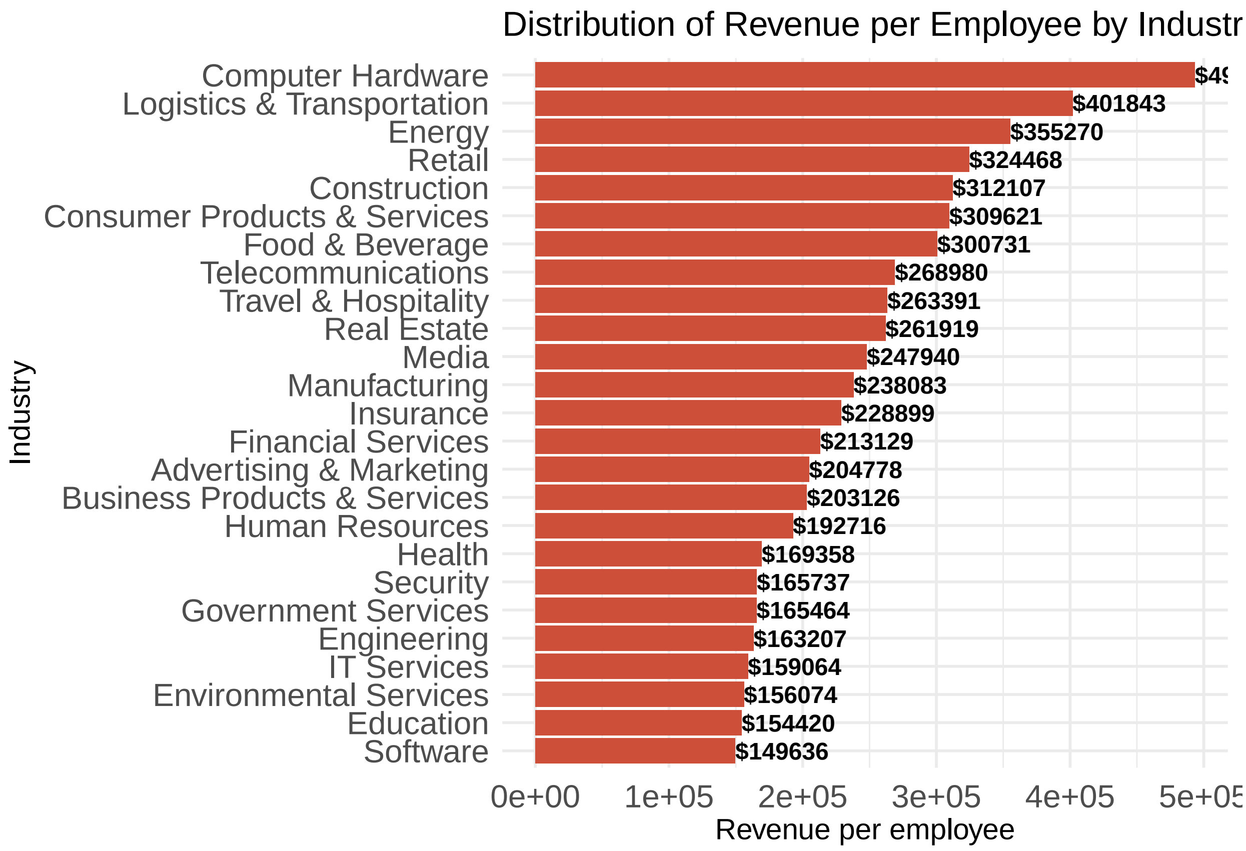
Question 3

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.

```
inc <- inc %>% mutate(rev_per_empl = Revenue/Employees)
rev_per_industry <- inc %>% filter(complete.cases(.)) %>% group_by(Industry) %>% filter(!(abs(rev_per_e
rev_per_industry
```

```
## # A tibble: 25 x 3
##    Industry                     revenue_per_employee  rev_sd
##    <chr>                                       <dbl>   <dbl>
##  1 Advertising & Marketing                   204778. 107797.
##  2 Business Products & Services              203126. 128333.
##  3 Computer Hardware                         493371. 286003.
##  4 Construction                              312107. 207266.
##  5 Consumer Products & Services              309621. 216944.
##  6 Education                                 154420.  91885.
##  7 Energy                                    355270. 288510.
##  8 Engineering                               163207.  58792.
##  9 Environmental Services                    156074.  57728.
## 10 Financial Services                        213129. 117896.
## # ... with 15 more rows
```

```
ggplot(data = rev_per_industry, aes(x=reorder(Industry,revenue_per_employee),y = revenue_per_employee))
  geom_bar(stat="identity",  fill="tomato3")+
  geom_text(aes(label=sprintf("$%0.0f",round(revenue_per_employee, digits=0))), fontface="bold",  vjust=
  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
Logistics & Transportation
Energy
Retail
Construction
Consumer Products & Services
Food & Beverage
Telecommunications
Travel & Hospitality
Real Estate
Media
Manufacturing
Insurance
Financial Services
Advertising & Marketing
Business Products & Services
Human Resources
Health
Security
Government Services
Engineering
IT Services
Environmental Services
Education
Software
$49
$401843
$355270
$324468
$312107
$309621
$300731
$268980
$263391
$261919
$247940
$238083
$228899
$213129
$204778
$203126
$192716
$169358
$165737
$165464
$163207
$159064
$156074
$154420
$149636
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
Revenue per employee