DATA608: Assignment 1

Eric Lehmphul

09/11/2022

```
library(tidyverse)
```

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:

The default summary function does not provide the standard deviation of the numeric variables. The standard deviation can be leveraged to understand the spread of the numeric data.

```
# Insert your code here, create more chunks as necessary

# Standard Deviations
inc %>% summarise(across(
    .cols = is.numeric,
    .fns = list(SD = sd), na.rm = TRUE,
    .names = "{col}_{fn}"
    ))
```

```
##    Rank_SD Growth_Rate_SD Revenue_SD Employees_SD
## 1 1443.506       14.12369  240542281     1353.128
```

The categorical variables of `Name`, `Industry`, `City`, `State` were stored as a character data type rather than a factor data type. The most frequent factor levels of each variable are displayed below.

```
inc$Name <- as.factor(inc$Name)
inc$Industry <- as.factor(inc$Industry)
inc$City <- as.factor(inc$City)
inc$State <- as.factor(inc$State)
```

```
get_most_least_freq <- function(variable){
  top <- inc %>%
    count({{variable}}) %>%
    arrange(desc(n)) %>%
    slice_head(n = 5)

  bottom <- inc %>%
    count({{variable}}) %>%
    arrange(desc(n)) %>%
    slice_tail(n = 5)

  return(rbind(top, bottom))
}
```

```
get_most_least_freq(Name)
```

```
##                                   Name n
## 1                      (Add)ventures 1
## 2                        @Properties 1
```

```
## 3                  1-Stop Translation USA 1
## 4                          110 Consulting 1
## 5                    11thStreetCoffee.com 1
## 6                                 Zoup! 1
## 7  ZT Wealth and Altus Group of Companies 1
## 8                                Zumasys 1
## 9                                 Zurple 1
## 10                             ZweigWhite 1
```

```
get_most_least_freq(Industry)
```

```
##                        Industry   n
## 1                   IT Services 733
## 2  Business Products & Services 482
## 3       Advertising & Marketing 471
## 4                        Health 355
## 5                      Software 342
## 6          Travel & Hospitality  62
## 7                         Media  54
## 8        Environmental Services  51
## 9                     Insurance  50
## 10            Computer Hardware  44
```

```
get_most_least_freq(City)
```

```
##              City   n
## 1        New York 160
## 2         Chicago  90
## 3          Austin  88
## 4         Houston  76
## 5   San Francisco  75
## 6  Woodland Hills   1
## 7       Woodville   1
## 8      Wyomissing   1
## 9         Yonkers   1
## 10       Zumbrota   1
```

```
get_most_least_freq(State)
```

```
##    State   n
## 1     CA 701
## 2     TX 387
## 3     NY 311
## 4     VA 283
## 5     FL 282
## 6     SD   3
## 7     AK   2
## 8     WV   2
## 9     WY   2
## 10    PR   1
```

After exploring the data summary information it is clear that the `Growth_Rate` column is heavily skewed as the mean is 4.612, median is 1.420, and the sd is 14.12369. There also appears to be outliers as the 3rd quartile value is 3.290 and the max is 421.480.

The variable `Employee` also has a large amount of variance. Most of the companies are relatively small in size 75% of the companies had less than 132 employees with some just having a singular worker. This data also contains much larger companies as the max value is 66,803 employees. It would be interesting to investigate whether employee size affects the growth rate of a company.

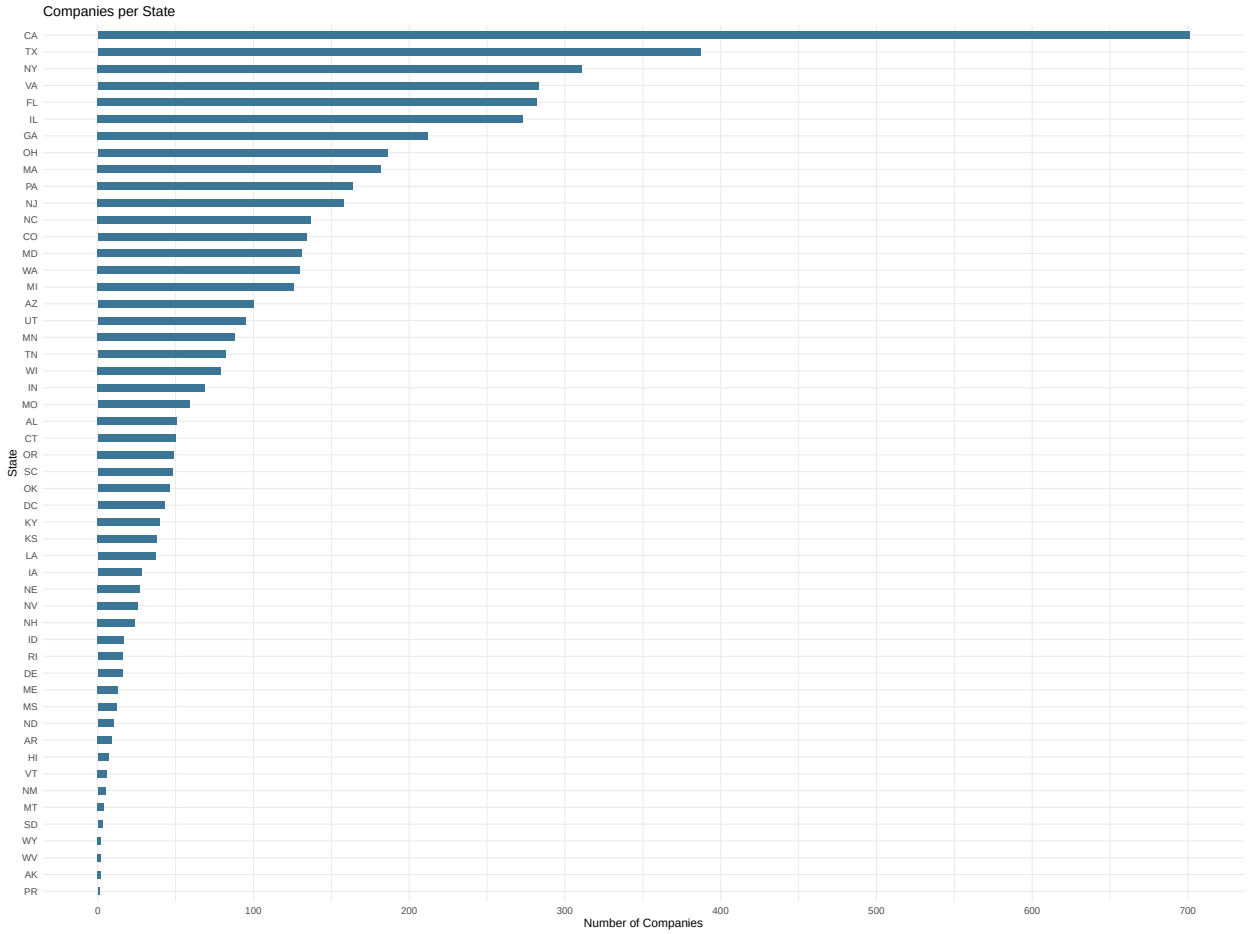
A large portion of the companies reside in large commercial cities and states. IT Services is the most popular industry by far in this dataset.

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.

```
# Answer Question 1 here

# Order the states in descending order
ordered.states <- inc %>% count(State)

# Create plot
ggplot(ordered.states, aes(x = reorder(State, n), y = n)) +
  geom_bar(stat = "identity", width = 0.475, position = "dodge", fill = "#3B7696") +
  ylim(0, 725) +
  scale_y_continuous(breaks = (seq(0, 700, by = 100))) +
  coord_flip() +
  ylab("Number of Companies") +
  xlab("State") +
  ggtitle("Companies per State") +
  theme_minimal()
```

```
## Scale for 'y' is already present. Adding another scale for 'y', which will
## replace the existing scale.
```



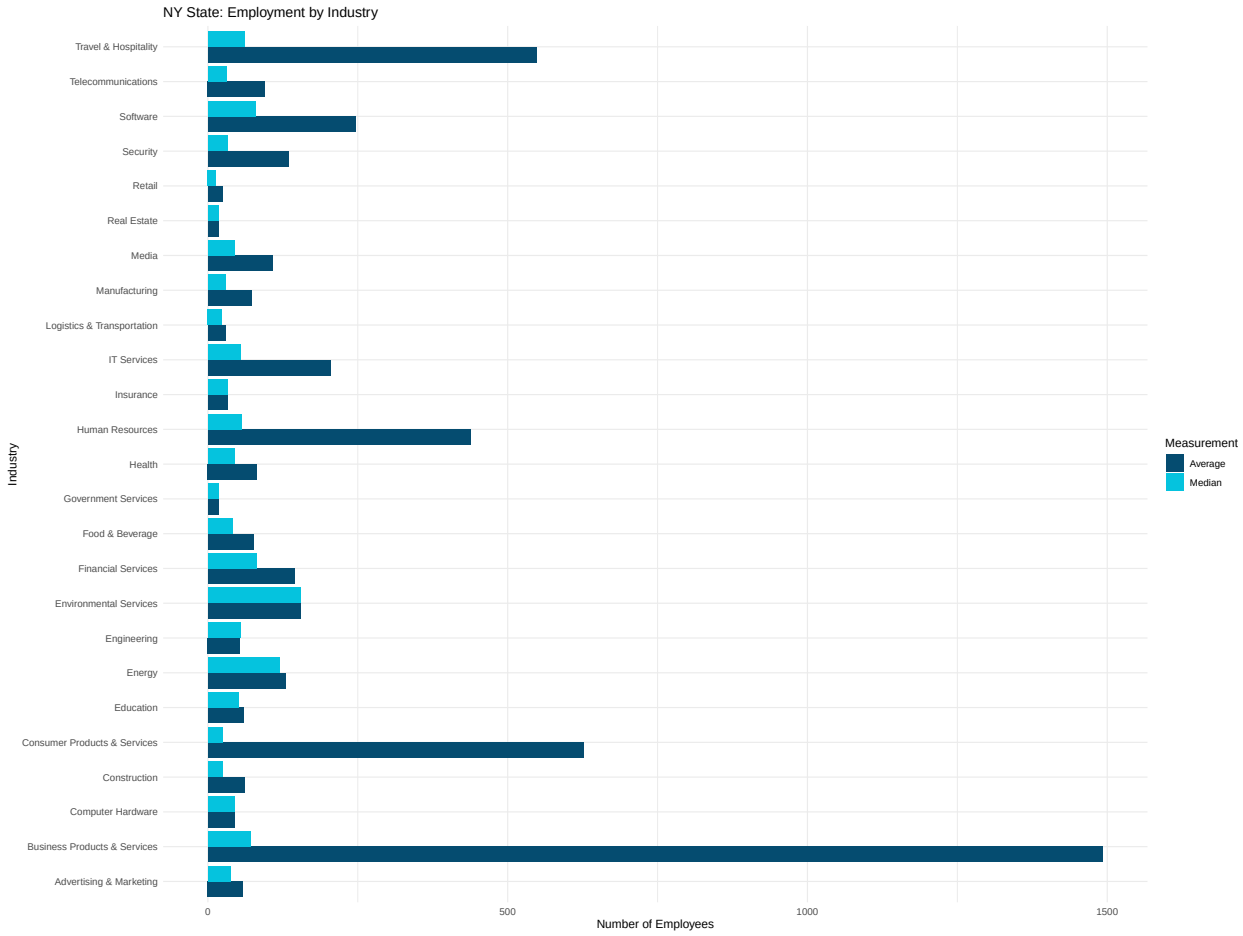
Quesiton 2

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.

Using Barplots

```
# Answer Question 2 here

# Get data for the state with the 3r most companies (NY)
q2.data.barplot <- inc %>%
  filter(State == "NY") %>%
  filter(complete.cases(.)) %>%
  group_by(Industry) %>%
  summarise(Average = mean(Employees), Median = median(Employees)) %>%
  gather("Measurement", "value", 2:3)
```

```r
# Create Plot
q2.data.barplot %>% ggplot(aes(x = Industry, y = value)) +
  geom_bar(stat = "identity", position = position_dodge(), aes(fill=Measurement)) +
  coord_flip() +
  xlab("Industry") +
  ylab("Number of Employees") +
  ggtitle("NY State: Employment by Industry") +
  scale_fill_manual(values = c("#054C70","#05C3DE")) +
  theme_minimal()
```

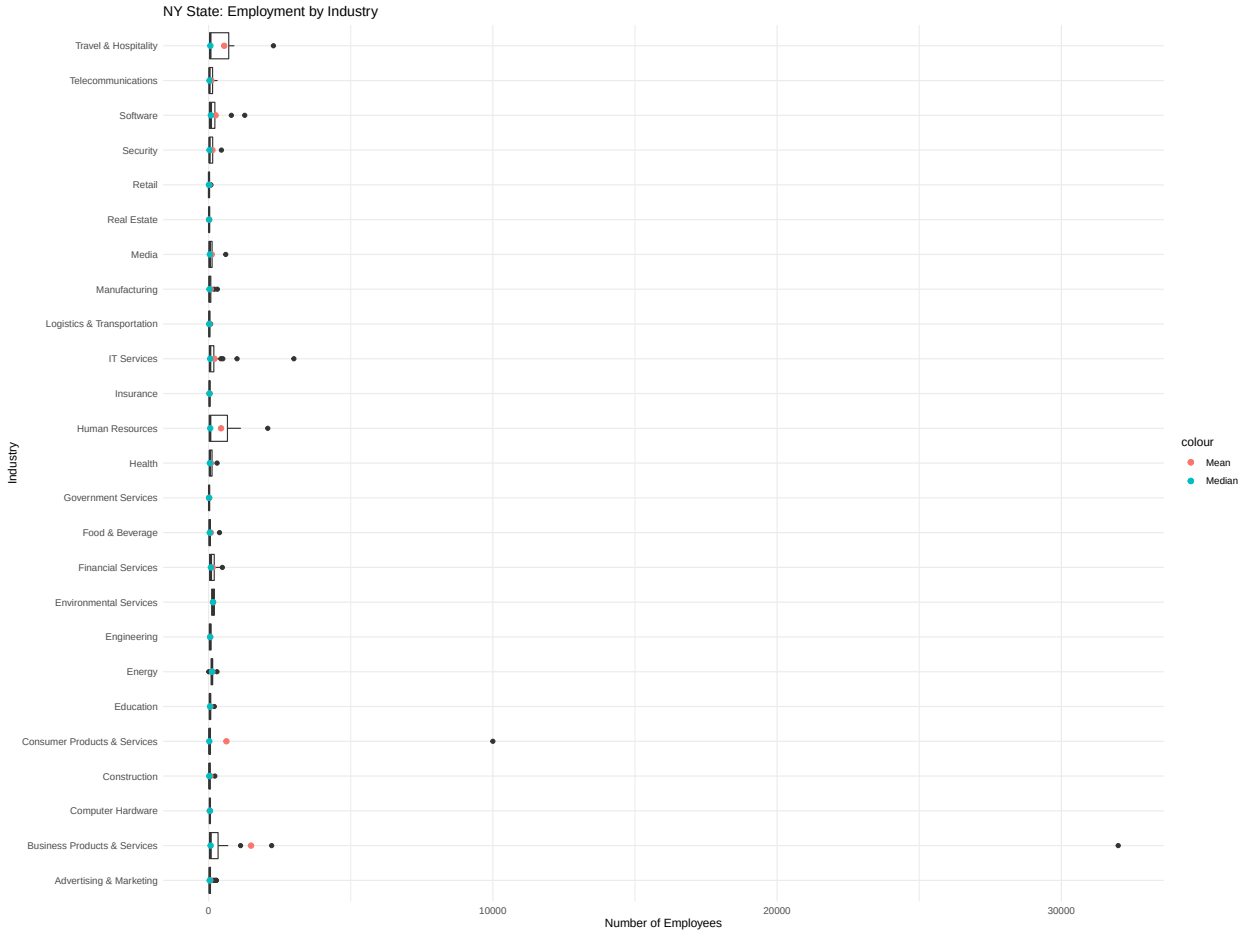


Using Boxplots

With all Outliers

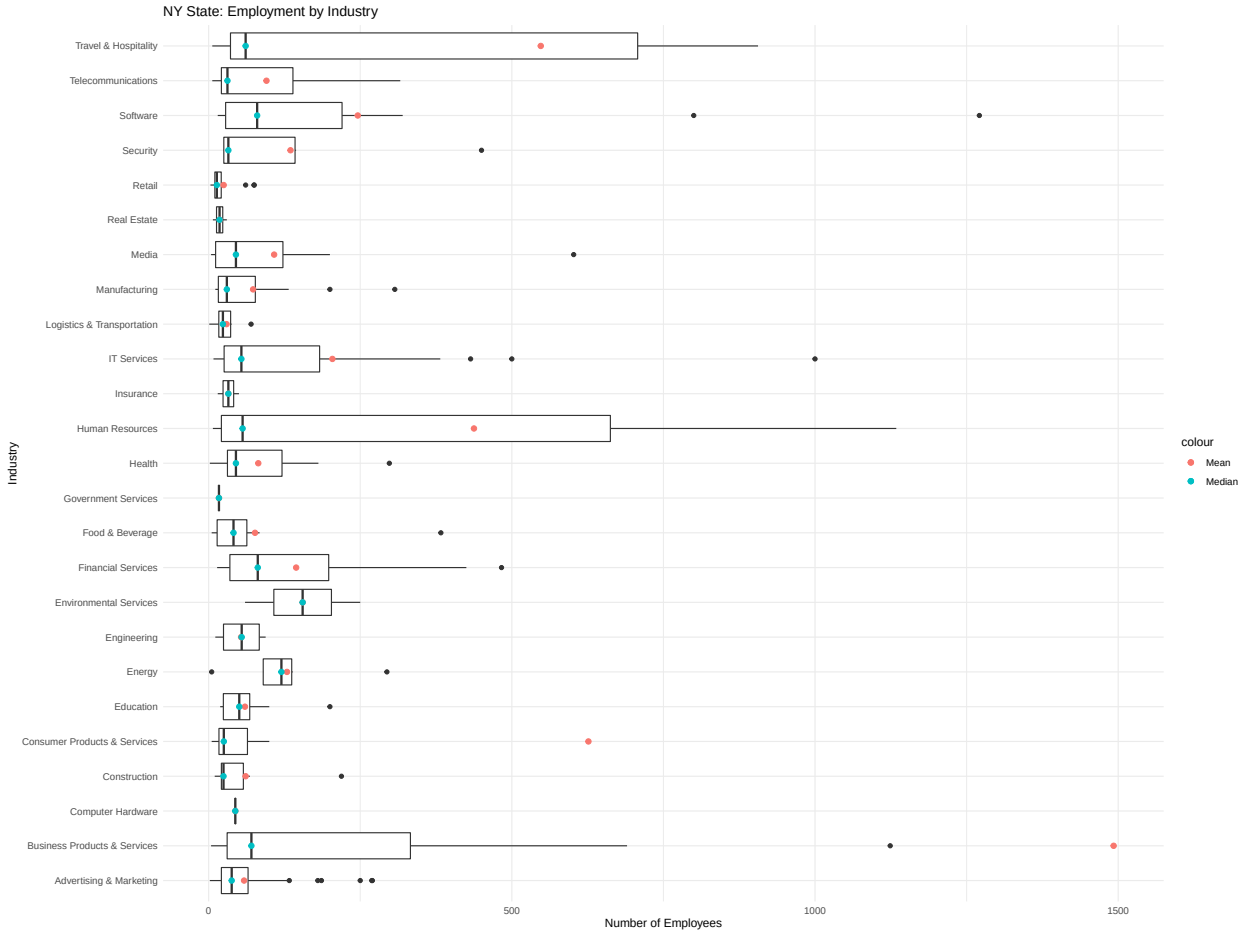
```r
q2.data <- inc %>%
  filter(State == "NY") %>%
  filter(complete.cases(.))

# Create plot
q2.data %>% ggplot(aes(x = Employees, y = Industry)) +
  geom_boxplot() +
```

```
  stat_summary(fun = "mean", size = 2, geom = "point", aes(color = "Mean")) +
  stat_summary(fun = "median", size = 2, geom = "point", aes(color = "Median")) +
  ggtitle("NY State: Employment by Industry") +
  xlab("Number of Employees") +
  theme_minimal()
```



Excluded Extreme Outliers

```
q2.data %>% ggplot(aes(x = Industry, y = Employees)) +
  geom_boxplot() +
  coord_flip(ylim = c(0, 1500)) +
  stat_summary(fun = "mean", size = 2, geom = "point", aes(color = "Mean")) +
  stat_summary(fun = "median", size = 2, geom = "point", aes(color = "Median")) +
  ggtitle("NY State: Employment by Industry") +
  ylab("Number of Employees") +
  theme_minimal()
```



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.

```
# Answer Question 3 here

# remove scientific notation
options(scipen = 5)

# Generate the revenue per employee
revenue.data <- inc %>%
  filter(complete.cases(.)) %>%
  group_by(Industry) %>%
  summarise(total_revenue = sum(Revenue), total_employees = sum(Employees), revenue_per_employee = (tota

# Create plot

revenue.data %>% ggplot(aes(x = revenue_per_employee, y = reorder(Industry, revenue_per_employee))) +
  geom_bar(stat = "identity", width = 0.475, position = "dodge", fill = "#04354F") +
  xlab("Revenue per Employee (in $)") +
```

```
  ylab("Industry") +
  xlim(0, 1250000) +
  ggtitle("Revenue per Employee by Industry") +
  theme_minimal()
```