Project 1

CUNY MSDS DATA 608

Duubar Villalobos Jimenez

February 10, 2018

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:

```
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
```

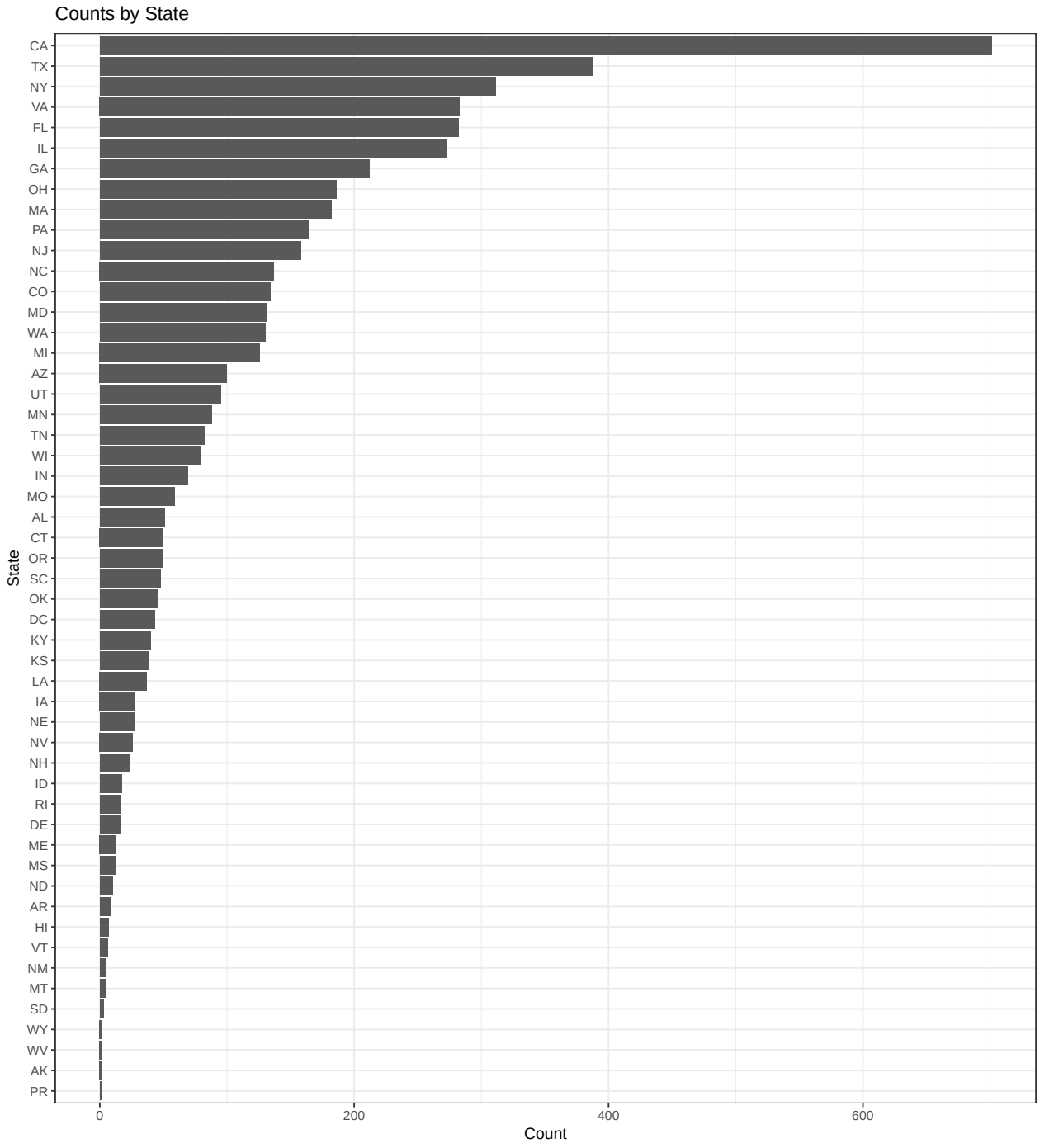
Think a bit on what these summaries mean. Use the space below to add some more relevant non-visual exploratory information you think helps you understand this data:

```
# Need to create individual summaries
Growth_Rate <- summary(inc$Growth_Rate)

Revenue <- summary(inc$Revenue)

Industry <- summary(inc$Industry)

Employees <- summary(inc$Employees)

City <- summary(inc$City)

State <- summary(inc$State)
```

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
my.data <- inc %>%
            group_by(State) %>%
            summarise('Count' = n()) %>%
            arrange(desc(`Count`))

# Basic plot
p <- ggplot(my.data, aes(x = reorder(State, Count), y = Count)) +
      geom_bar(stat='identity') +
      coord_flip() +
      xlab("State") +
      ylab("Count") +
      ggtitle("Counts by State") +
      theme(plot.title = element_text(hjust = 0.5)) +
      scale_y_continuous() +
      theme_bw()
```



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.

State with the 3rd most companies

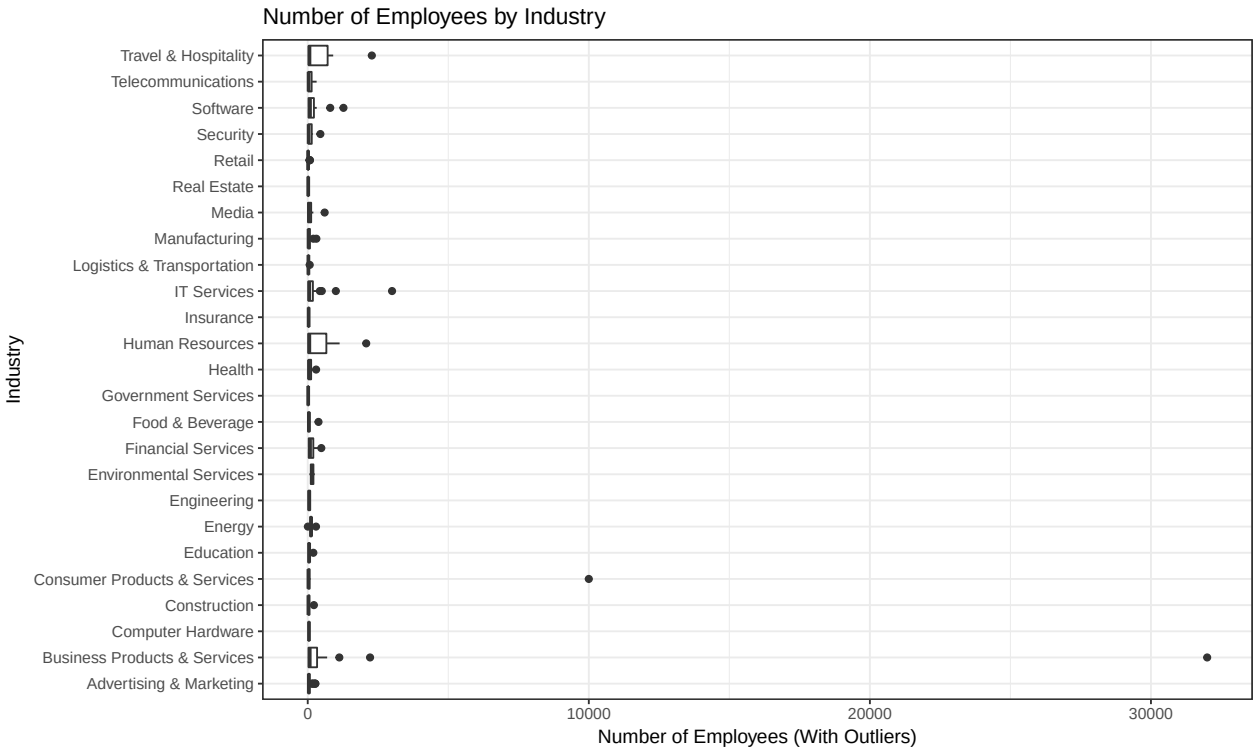
In order to get the State with the 3rd most companies, we can select as follows:

```
# Answer Question 2 here
x = arrange(my.data,desc(Count))
x1 <- x[3:3,]  # Return the State with the 3rd most companies
```

From the above code we see that the State with the 3rd most companies is as follows:

```
##   State Count
## 1    NY   311
```

```
# Subsetting data for the selected State
ind_by_state <- subset(inc, State == as.character(x1$State[1]))

# Complete Cases
ind_by_state <- ind_by_state %>%
                filter(complete.cases(Employees))

# Identifying outliers
ind_outliers <- ggplot(ind_by_state, aes(Industry, Employees)) +
                geom_boxplot() +
                coord_flip() +
                xlab("Industry") +
                ylab("Number of Employees (With Outliers)") +
                ggtitle("Number of Employees by Industry") +
                theme(plot.title = element_text(hjust = 0.5)) +
                scale_y_continuous() +
                theme_bw()
```

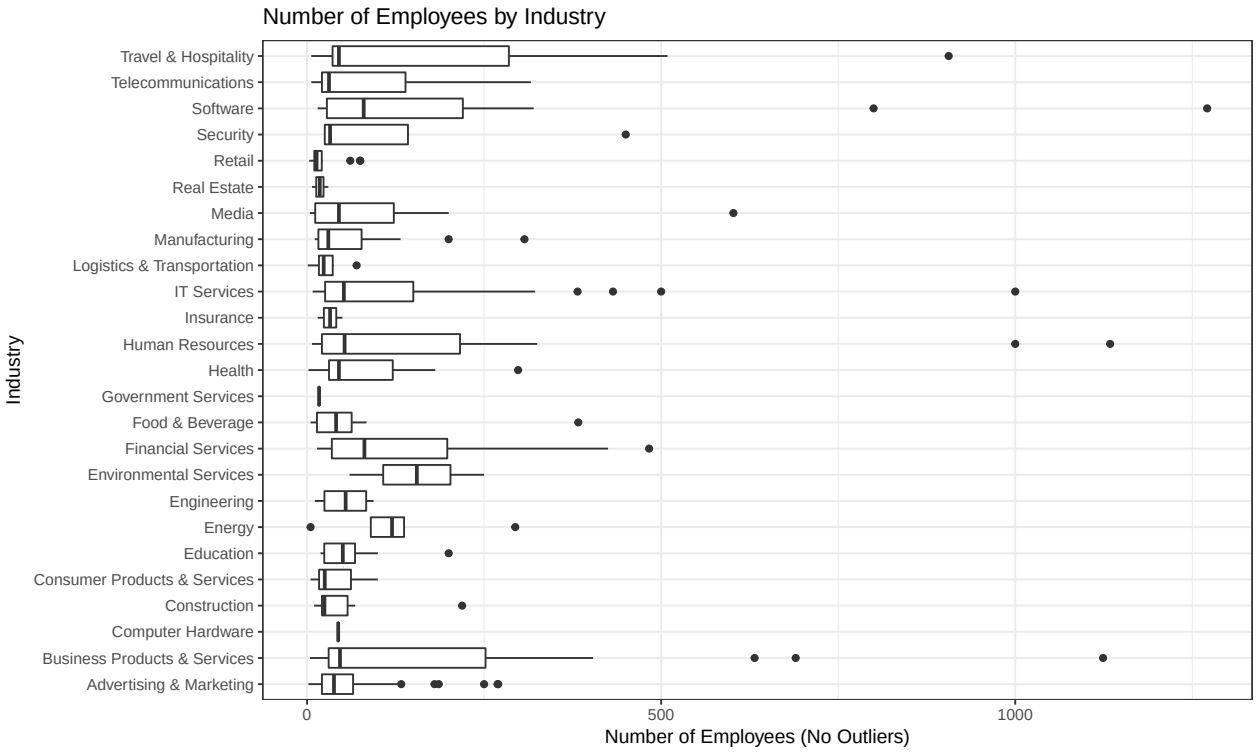


Discarding Outliers (We will repeat the process 6 times in this case in order to identify if an outlier is detected and confirmed by statistical tests, this function can remove it or replace by sample mean or median.)

```r
# https://cran.r-project.org/web/packages/outliers/outliers.pdf

ind_no_outliers <- rm.outlier(ind_by_state$Employees, fill = TRUE, median = TRUE, opposite = FALSE)
ind_by_state$Emp_No_Outliers <- ind_no_outliers

ind_no_outliers <- rm.outlier(ind_by_state$Emp_No_Outliers, fill = TRUE, median = TRUE, opposite = FALS
ind_by_state$Emp_No_Outliers <- ind_no_outliers

ind_no_outliers <- rm.outlier(ind_by_state$Emp_No_Outliers, fill = TRUE, median = TRUE, opposite = FALS
ind_by_state$Emp_No_Outliers <- ind_no_outliers

ind_no_outliers <- rm.outlier(ind_by_state$Emp_No_Outliers, fill = TRUE, median = TRUE, opposite = FALS
ind_by_state$Emp_No_Outliers <- ind_no_outliers

ind_no_outliers <- rm.outlier(ind_by_state$Emp_No_Outliers, fill = TRUE, median = TRUE, opposite = FALS
ind_by_state$Emp_No_Outliers <- ind_no_outliers

ind_no_outliers <- rm.outlier(ind_by_state$Emp_No_Outliers, fill = TRUE, median = TRUE, opposite = FALS
ind_by_state$Emp_No_Outliers <- ind_no_outliers
```

```r
# Identifying outliers
ind_outliers <- ggplot(ind_by_state, aes(Industry, Emp_No_Outliers)) +
                geom_boxplot() +
                coord_flip() +
                xlab("Industry") +
                ylab("Number of Employees (No Outliers)") +
                ggtitle("Number of Employees by Industry") +
                theme(plot.title = element_text(hjust = 0.5)) +
                scale_y_continuous() +
                theme_bw()
```

Result after 6 iterations; Outliers have been replaced by the Median value.



Obtaining data with outliers

```
# Obatining  summary data for the state
my.data <- ind_by_state %>%
            group_by(Industry) %>%
            summarise('Count' = n(),
                      'N_Employees' = sum(Employees),
                      'Average' = round(mean(Employees),0),
                      'Median' = round(median(Employees),0)) %>%
            arrange(desc(`Count`))

my.data <- data.frame(my.data)

# Basic plot Company Counts by Industry
p1 <- ggplot(my.data, aes(x = reorder(Industry, Count), y = Count)) +
      geom_bar(stat='identity') +
      coord_flip() +
      xlab("Industry") +
      ylab("Count") +
      ggtitle("Company Counts by Industry") +
      theme(plot.title = element_text(hjust = 0.5)) +
      scale_y_continuous() +
      theme_bw()

# Basic plot Number of Employees by Industry
p2 <- ggplot(my.data, aes(x = reorder(Industry, N_Employees), y = N_Employees)) +
      geom_bar(stat='identity') +
      coord_flip() +
      xlab("Industry") +
      ylab("Count") +
```

```
      ggtitle("Number of Employees by Industry") +
      theme(plot.title = element_text(hjust = 0.5)) +
      scale_y_continuous() +
      theme_bw()

# Basic plot Average Number of Employees by Industry
p3 <- ggplot(my.data, aes(x = reorder(Industry, Average), y = Average)) +
      geom_bar(stat='identity') +
      coord_flip() +
      xlab("Industry") +
      ylab("Count") +
      ggtitle("Average Number of Employees by Industry") +
      theme(plot.title = element_text(hjust = 0.5)) +
      scale_y_continuous() +
      theme_bw()

# Basic plot Median Number of Employees by Industry
p4 <- ggplot(my.data, aes(x = reorder(Industry, Median), y = Median)) +
      geom_bar(stat='identity') +
      coord_flip() +
      xlab("Industry") +
      ylab("Count") +
      ggtitle("Median Number of Employees by Industry") +
      theme(plot.title = element_text(hjust = 0.5)) +
      scale_y_continuous() +
      theme_bw()
```

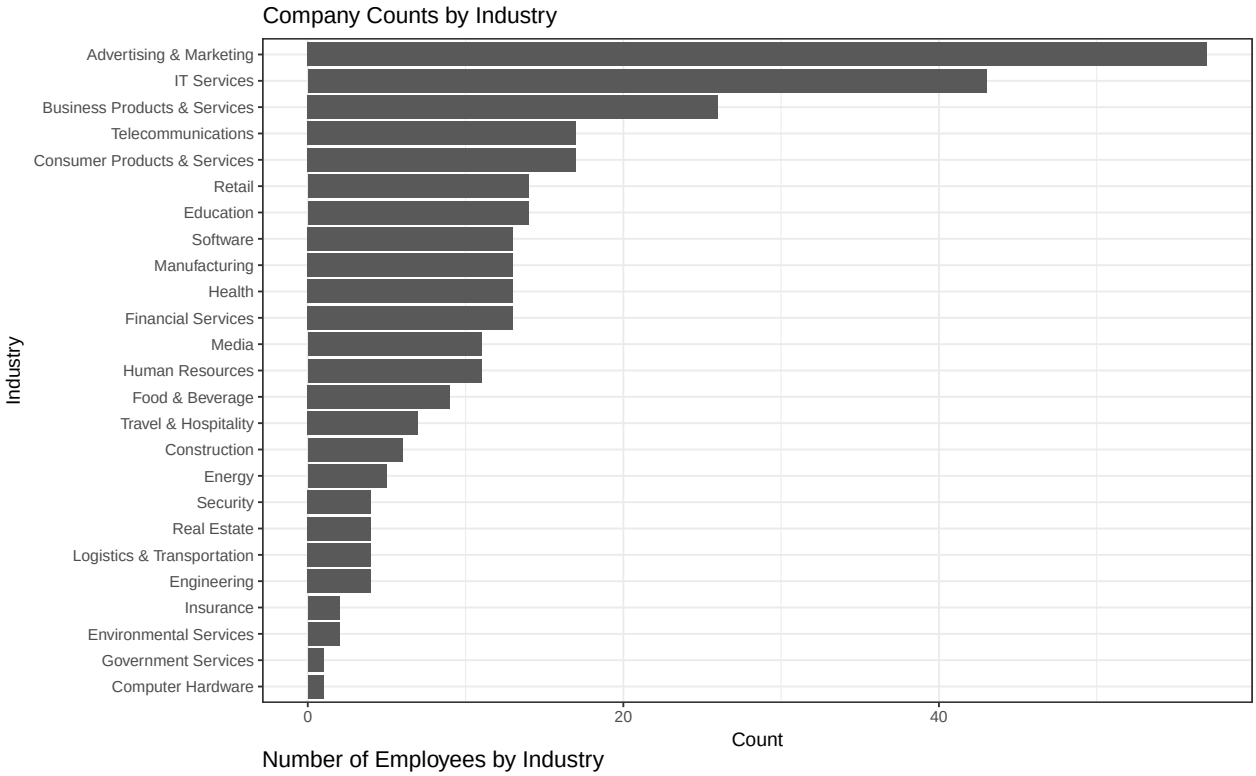
Let's have a visual of the first few rows of the data:

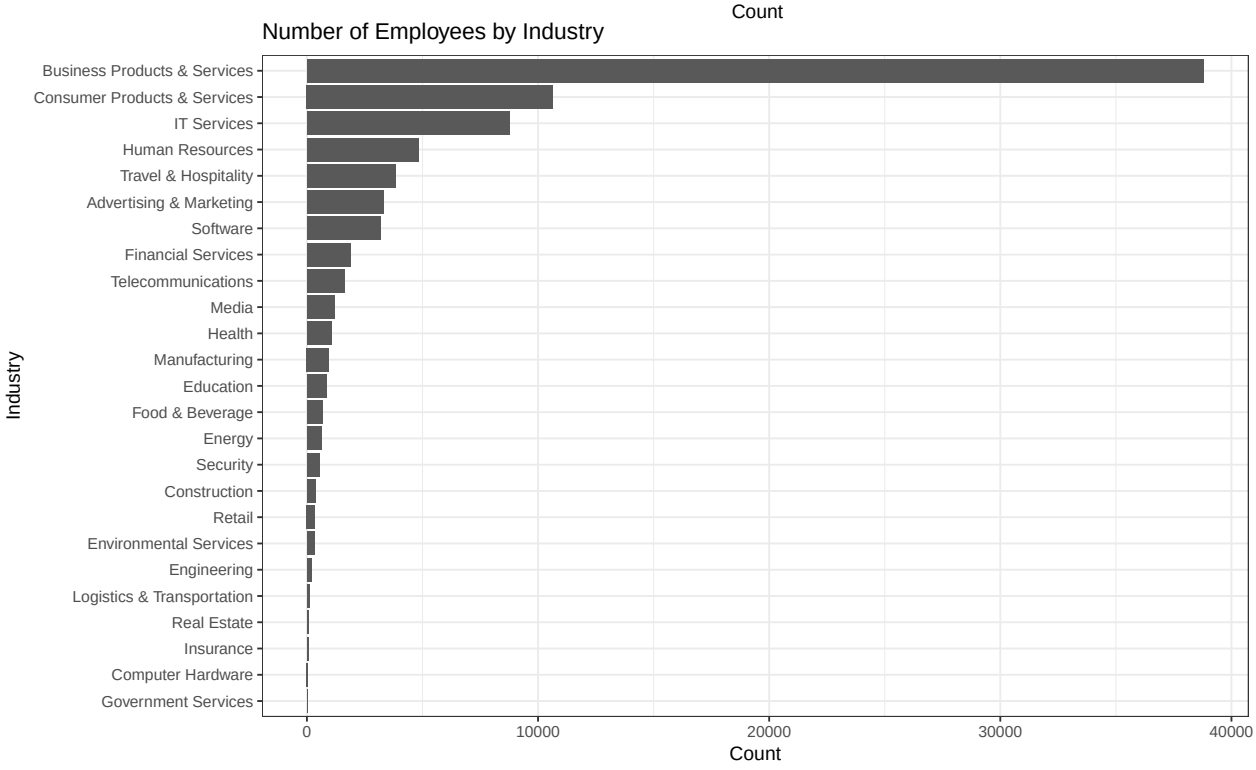
```
my.data
```

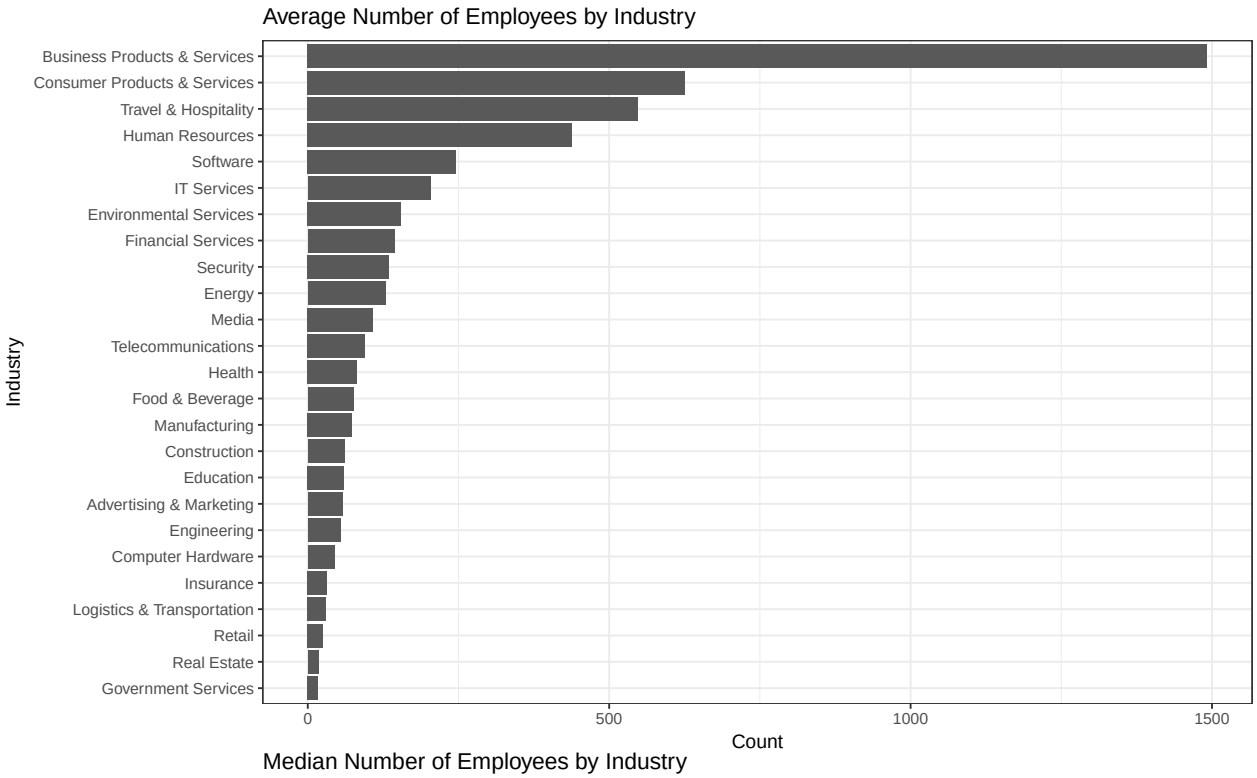
```
##                        Industry Count N_Employees Average Median
## 1       Advertising & Marketing    57        3331      58     38
## 2                   IT Services    43        8776     204     54
## 3  Business Products & Services    26       38804    1492     70
## 4  Consumer Products & Services    17       10647     626     25
## 5            Telecommunications    17        1621      95     31
## 6                     Education    14         838      60     50
## 7                        Retail    14         347      25     14
## 8            Financial Services    13        1876     144     81
## 9                        Health    13        1064      82     45
## 10               Manufacturing    13         953      73     30
## 11                    Software    13        3197     246     80
## 12              Human Resources    11        4813     438     56
## 13                       Media    11        1188     108     45
## 14             Food & Beverage     9         688      76     41
## 15         Travel & Hospitality     7        3834     548     61
## 16                Construction     6         366      61     24
## 17                      Energy     5         646     129    120
## 18                  Engineering     4         214      54     54
## 19   Logistics & Transportation     4         118      30     24
## 20                  Real Estate     4          73      18     18
## 21                     Security     4         540     135     32
```

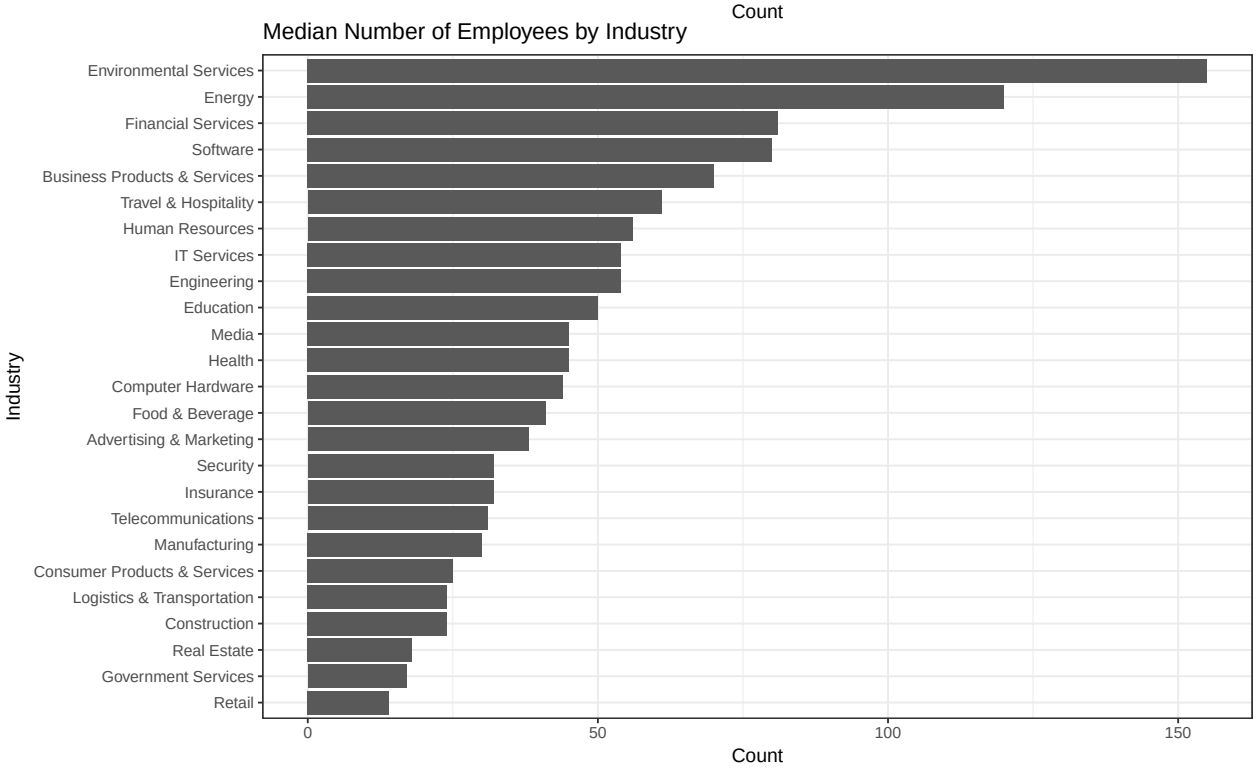
```
## 22       Environmental Services     2         310      155     155
## 23                    Insurance     2          65       32      32
## 24             Computer Hardware     1          44       44      44
## 25           Government Services     1          17       17      17
```

Graphical representation iof the data.









Obtaining data with NO outliers

```
# Obatining  summary data for the state
my.data <- ind_by_state %>%
            group_by(Industry) %>%
            summarise('Count' = n(),
                      'N_Employees' = sum(Emp_No_Outliers),
```

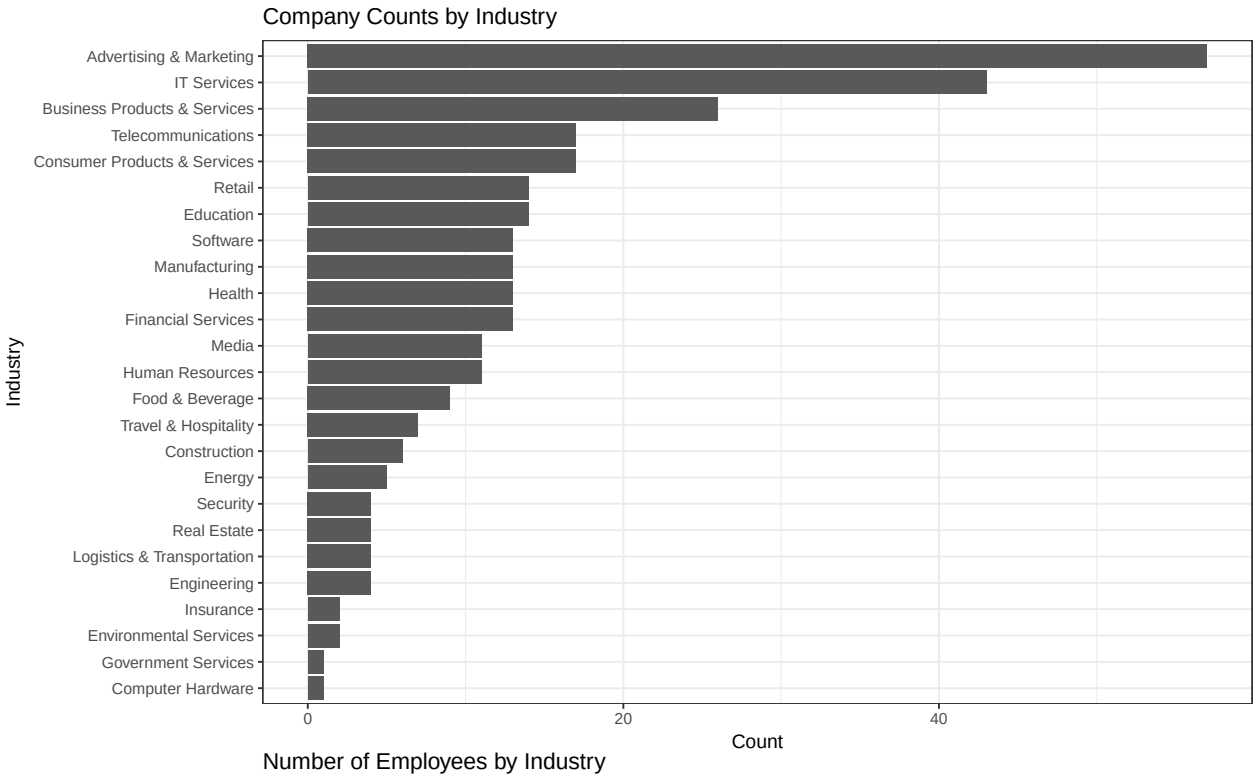
```
                      'Average' = round(mean(Emp_No_Outliers),0),
                      'Median' = round(median(Emp_No_Outliers),0)) %>%
            arrange(desc(`Count`))

my.data <- data.frame(my.data)

# Basic plot Company Counts by Industry
p1 <- ggplot(my.data, aes(x = reorder(Industry, Count), y = Count)) +
      geom_bar(stat='identity') +
      coord_flip() +
      xlab("Industry") +
      ylab("Count") +
      ggtitle("Company Counts by Industry") +
      theme(plot.title = element_text(hjust = 0.5)) +
      scale_y_continuous() +
      theme_bw()

# Basic plot Number of Employees by Industry
p2 <- ggplot(my.data, aes(x = reorder(Industry, N_Employees), y = N_Employees)) +
      geom_bar(stat='identity') +
      coord_flip() +
      xlab("Industry") +
      ylab("Count") +
      ggtitle("Number of Employees by Industry") +
      theme(plot.title = element_text(hjust = 0.5)) +
      scale_y_continuous() +
      theme_bw()

# Basic plot Average Number of Employees by Industry
p3 <- ggplot(my.data, aes(x = reorder(Industry, Average), y = Average)) +
      geom_bar(stat='identity') +
      coord_flip() +
      xlab("Industry") +
      ylab("Count") +
      ggtitle("Average Number of Employees by Industry") +
      theme(plot.title = element_text(hjust = 0.5)) +
      scale_y_continuous() +
      theme_bw()

# Basic plot Median Number of Employees by Industry
p4 <- ggplot(my.data, aes(x = reorder(Industry, Median), y = Median)) +
      geom_bar(stat='identity') +
      coord_flip() +
      xlab("Industry") +
      ylab("Count") +
      ggtitle("Median Number of Employees by Industry") +
      theme(plot.title = element_text(hjust = 0.5)) +
      scale_y_continuous() +
      theme_bw()
```

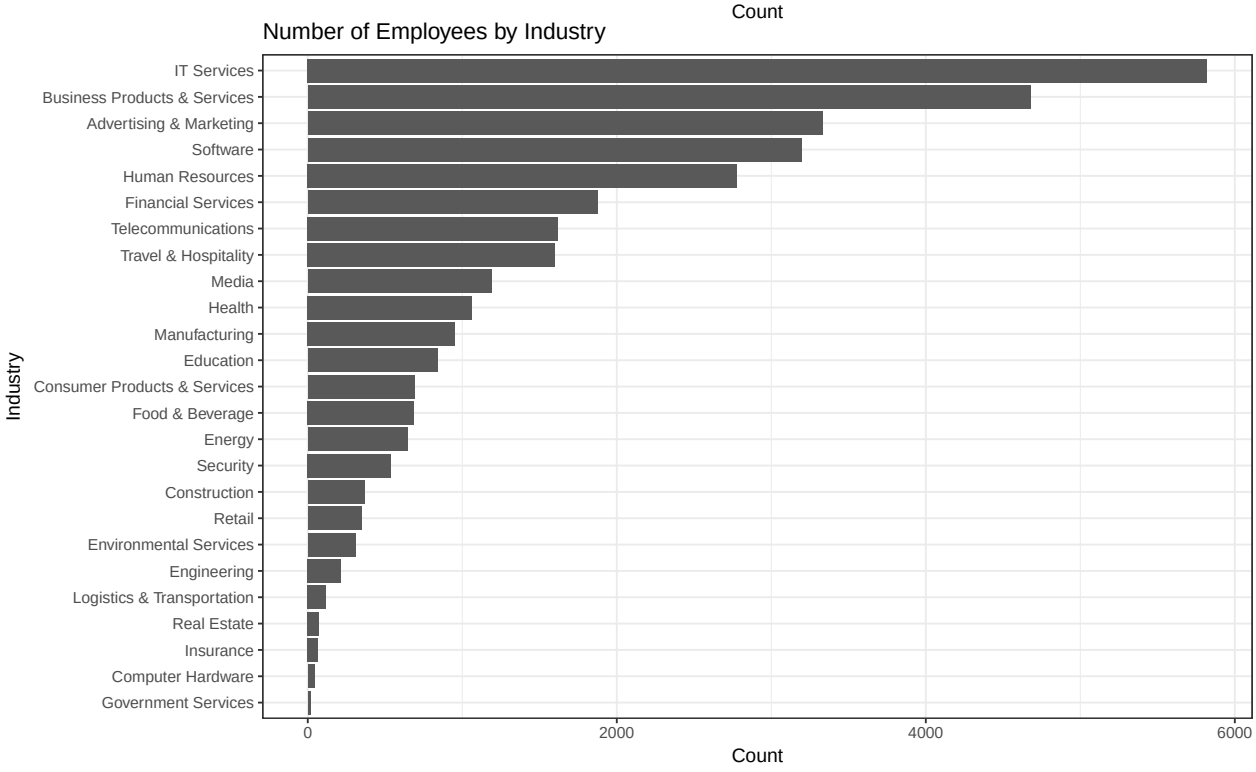
Let's have a visual of the first few rows of the data:

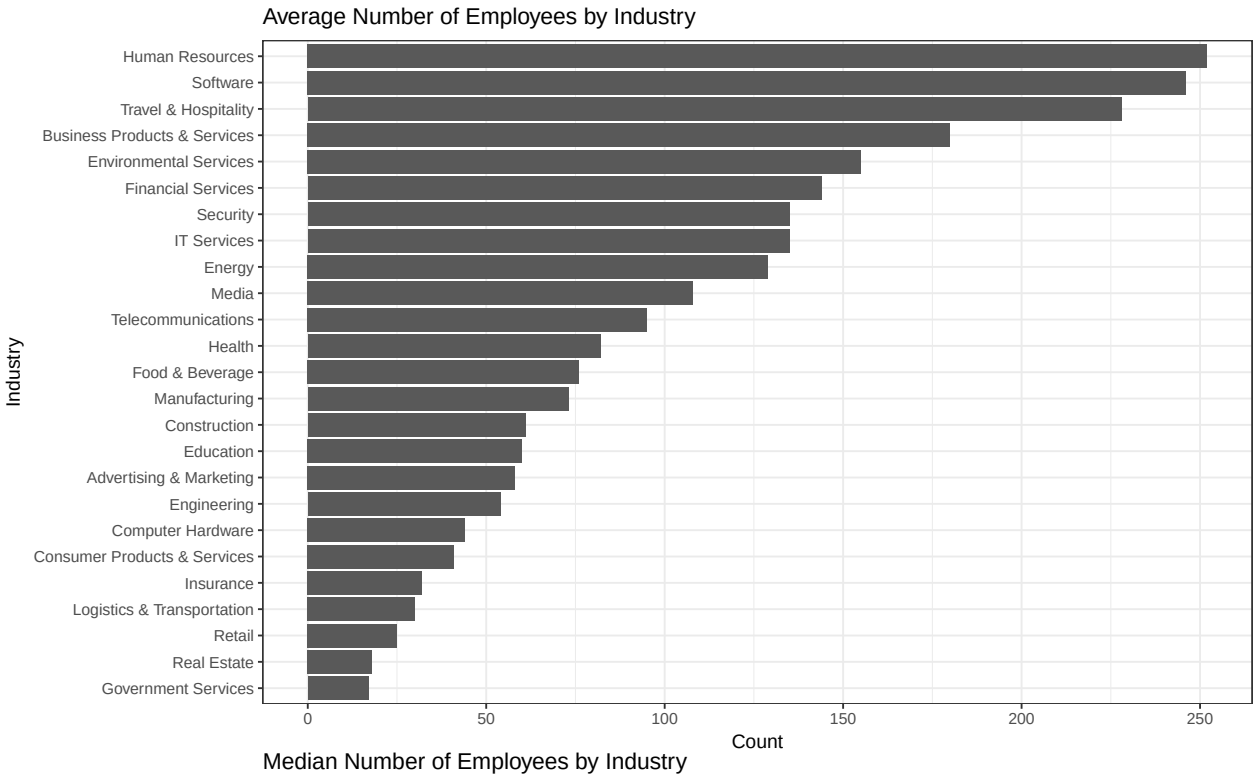
```
my.data
```

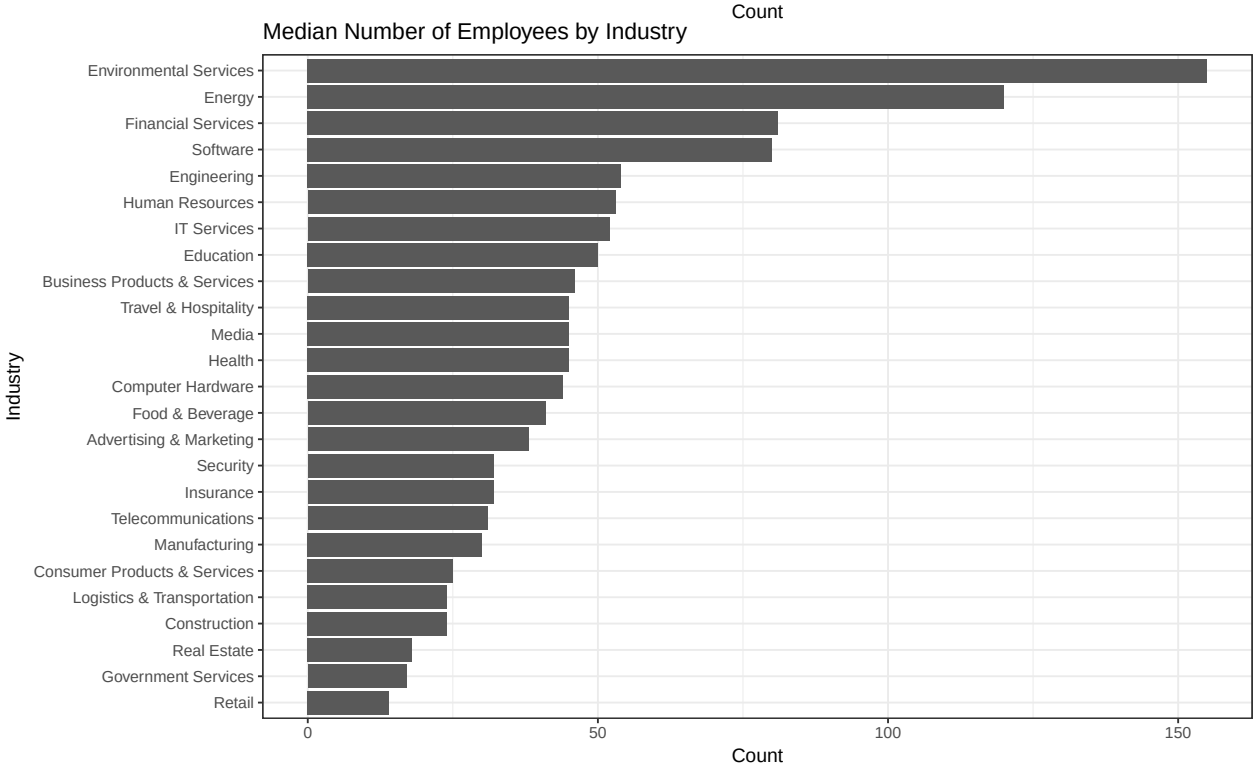
```
##                        Industry Count N_Employees Average Median
## 1       Advertising & Marketing    57        3331      58     38
## 2                   IT Services    43        5821     135     52
## 3  Business Products & Services    26        4676     180     46
## 4  Consumer Products & Services    17         692      41     25
## 5            Telecommunications    17        1621      95     31
## 6                     Education    14         838      60     50
## 7                        Retail    14         347      25     14
## 8             Financial Services    13        1876     144     81
## 9                        Health    13        1064      82     45
## 10               Manufacturing    13         953      73     30
## 11                    Software    13        3197     246     80
## 12              Human Resources    11        2777     252     53
## 13                        Media    11        1188     108     45
## 14              Food & Beverage     9         688      76     41
## 15         Travel & Hospitality     7        1599     228     45
## 16                Construction     6         366      61     24
## 17                      Energy     5         646     129    120
## 18                 Engineering     4         214      54     54
## 19   Logistics & Transportation     4         118      30     24
## 20                 Real Estate     4          73      18     18
## 21                    Security     4         540     135     32
## 22       Environmental Services     2         310     155    155
## 23                   Insurance     2          65      32     32
## 24            Computer Hardware     1          44      44     44
## 25          Government Services     1          17      17     17
```

Graphical representation iof the data (Since the Median and Mean are the same, I will present only one graphic).


Company Counts by Industry
Advertising & Marketing
IT Services
Business Products & Services
Telecommunications
Consumer Products & Services
Retail
Education
Software
Manufacturing
Health
Financial Services
Media
Human Resources
Food & Beverage
Travel & Hospitality
Construction
Energy
Security
Real Estate
Logistics & Transportation
Engineering
Insurance
Environmental Services
Government Services
Computer Hardware
Industry
0
20
40
Count


Number of Employees by Industry
IT Services
Business Products & Services
Advertising & Marketing
Software
Human Resources
Financial Services
Telecommunications
Travel & Hospitality
Media
Health
Manufacturing
Education
Consumer Products & Services
Food & Beverage
Energy
Security
Construction
Retail
Environmental Services
Engineering
Logistics & Transportation
Real Estate
Insurance
Computer Hardware
Government Services
Industry
0
2000
4000
6000
Count





Question 3

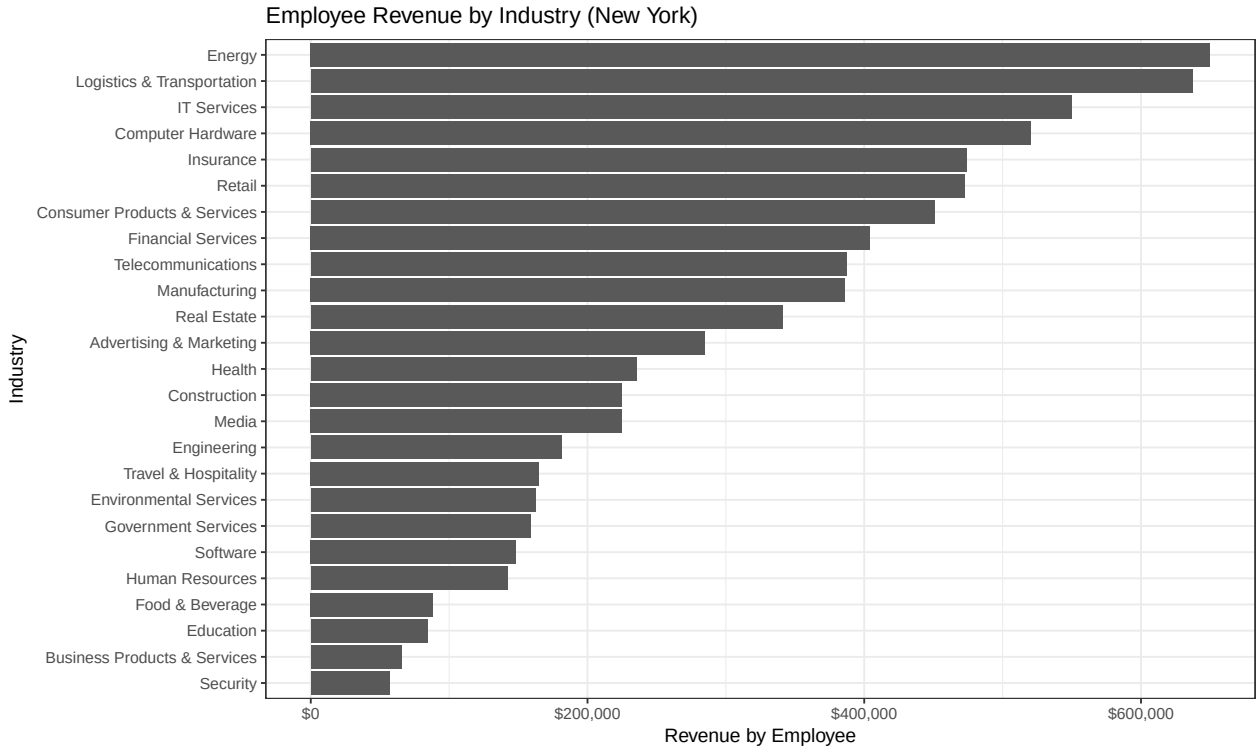
Now imagine you work for an investor and want to see which industries generate the most revenue per employee. Create a chart that makes this information clear. Once again, the distribution per industry should be shown.

```
# Answer Question 3 here
inc$Rev_by_Emp <- inc$Revenue / inc$Employees

# Selecting the #3 State  (I am assuming, we are still working with the third state with most companies
ind_by_state <- subset(inc, State == as.character(x1$State[1]))

# Complete Cases
ind_by_state <- ind_by_state %>%
                filter(complete.cases(Revenue, Employees))

# Obatining  summary data for the state
my.data <- ind_by_state %>%
            group_by(Industry) %>%
            summarise('Count' = n(),
                      'N_Employees' = sum(Employees),
                      'Tot_Revenue' = sum(Revenue)) %>%
            arrange(desc(`Tot_Revenue`))

my.data$Emp_Rev <- my.data$Tot_Revenue / my.data$N_Employees

my.data = arrange(my.data,desc(Emp_Rev))

my.data <- data.frame(my.data)
```

```
my.data
```

```
##                         Industry Count N_Employees Tot_Revenue   Emp_Rev
## 1                         Energy     5         646   419900000 650000.00
## 2    Logistics & Transportation     4         118    75200000 637288.14
## 3                   IT Services    43        8776  4826200000 549931.63
## 4             Computer Hardware     1          44    22900000 520454.55
## 5                     Insurance     2          65    30800000 473846.15
## 6                        Retail    14         347   164000000 472622.48
## 7  Consumer Products & Services    17       10647  4799300000 450765.47
## 8            Financial Services    13        1876   758100000 404104.48
## 9            Telecommunications    17        1621   627500000 387106.72
## 10                Manufacturing    13         953   368000000 386149.00
## 11                  Real Estate     4          73    24900000 341095.89
## 12      Advertising & Marketing    57        3331   949000000 284899.43
## 13                       Health    13        1064   250600000 235526.32
## 14                 Construction     6         366    82300000 224863.39
## 15                        Media    11        1188   267100000 224831.65
## 16                  Engineering     4         214    38800000 181308.41
## 17         Travel & Hospitality     7        3834   631800000 164788.73
## 18       Environmental Services     2         310    50400000 162580.65
## 19          Government Services     1          17     2700000 158823.53
## 20                      Software    13        3197   474600000 148451.67
## 21             Human Resources    11        4813   684100000 142135.88
## 22             Food & Beverage     9         688    60700000  88226.74
## 23                   Education    14         838    70800000  84486.87
## 24 Business Products & Services    26       38804  2549900000  65712.30
## 25                     Security     4         540    30800000  57037.04
```

```r
# Basic plot Number of Revenue per Employee by Industry
p <- ggplot(my.data, aes(x = reorder(Industry, Emp_Rev), y = Emp_Rev)) +
      geom_bar(stat='identity') +
      coord_flip() +
      xlab("Industry") +
      ylab("Revenue by Employee") +
      ggtitle("Employee Revenue by Industry (New York)") +
      theme(plot.title = element_text(hjust = 0.5)) +
      scale_y_continuous(labels = scales::dollar) +
      theme_bw()
```



For the whole Country

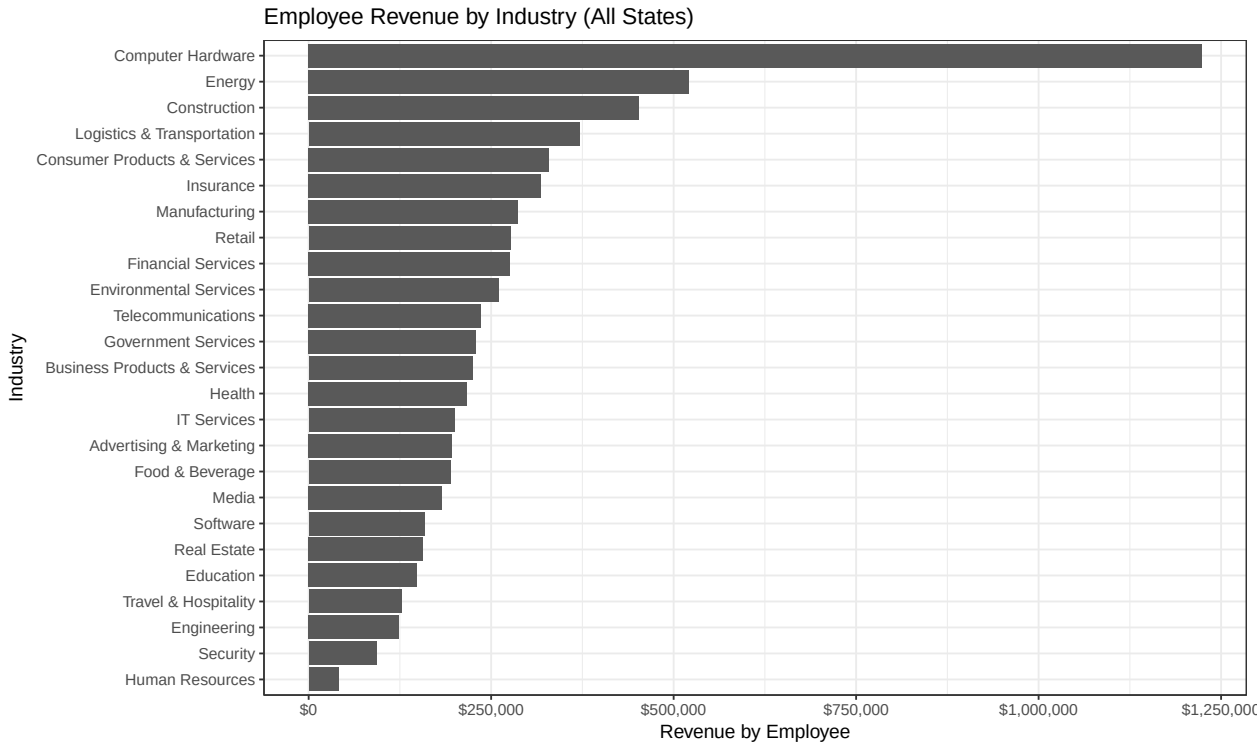
```r
# Answer Question 3 here
inc$Rev_by_Emp <- inc$Revenue / inc$Employees

# Selecting the #3 State  (I am assuming, we are still working with all of the states)
ind_by_state <- inc

# Complete Cases
ind_by_state <- ind_by_state %>%
                filter(complete.cases(Revenue, Employees))

# Obatining  summary data for the state
my.data <- ind_by_state %>%
            group_by(Industry) %>%
            summarise('Count' = n(),
                      'N_Employees' = sum(Employees),
                      'Tot_Revenue' = sum(Revenue)) %>%
            arrange(desc(`Tot_Revenue`))
```

```r
my.data$Emp_Rev <- my.data$Tot_Revenue / my.data$N_Employees


my.data <- data.frame(my.data)
```

```r
my.data
```

```
##                        Industry Count N_Employees Tot_Revenue    Emp_Rev
## 1  Business Products & Services   480      117357 26345900000  224493.64
## 2                   IT Services   732      102788 20525000000  199682.84
## 3                        Health   354       82430 17860100000  216669.90
## 4  Consumer Products & Services   203       45464 14956400000  328972.37
## 5    Logistics & Transportation   154       39994 14837800000  371000.65
## 6                        Energy   109       26437 13771600000  520921.44
## 7                  Construction   187       29099 13174300000  452740.64
## 8            Financial Services   260       47693 13150900000  275740.67
## 9               Food & Beverage   129       65911 12812500000  194390.92
## 10                Manufacturing   255       43942 12603600000  286823.54
## 11            Computer Hardware    44        9714 11885700000 1223563.93
## 12                       Retail   203       37068 10257400000  276718.46
## 13              Human Resources   196      226980  9246100000   40735.31
## 14                     Software   341       51262  8134600000  158686.75
## 15      Advertising & Marketing   471       39731  7785000000  195942.71
## 16           Telecommunications   127       30842  7287900000  236297.91
## 17          Government Services   202       26185  6009100000  229486.35
## 18                     Security    73       41059  3812800000   92861.49
## 19                  Real Estate    95       18893  2956800000  156502.41
## 20         Travel & Hospitality    62       23035  2931600000  127267.20
## 21       Environmental Services    51       10155  2638800000  259852.29
## 22                  Engineering    74       20435  2532500000  123929.53
## 23                    Insurance    50        7339  2337900000  318558.39
## 24                        Media    54        9532  1742400000  182794.80
## 25                    Education    83        7685  1139300000  148249.84
```

```r
# Basic plot Number of Revenue per Employee by Industry
p <- ggplot(my.data, aes(x = reorder(Industry, Emp_Rev), y = Emp_Rev)) +
      geom_bar(stat='identity') +
      coord_flip() +
      xlab("Industry") +
      ylab("Revenue by Employee") +
      ggtitle("Employee Revenue by Industry (All States)") +
      theme(plot.title = element_text(hjust = 0.5)) +
      scale_y_continuous(labels = scales::dollar) +
      theme_bw()
```


Employee Revenue by Industry (All States)
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
$0
$250,000
$500,000
$750,000
$1,000,000
$1,250,000
Revenue by Employee