Module1

Group 3

9/2/2021

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:

```
# Insert your code here, create more chunks as necessary
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.3     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(psych)
```

```
##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

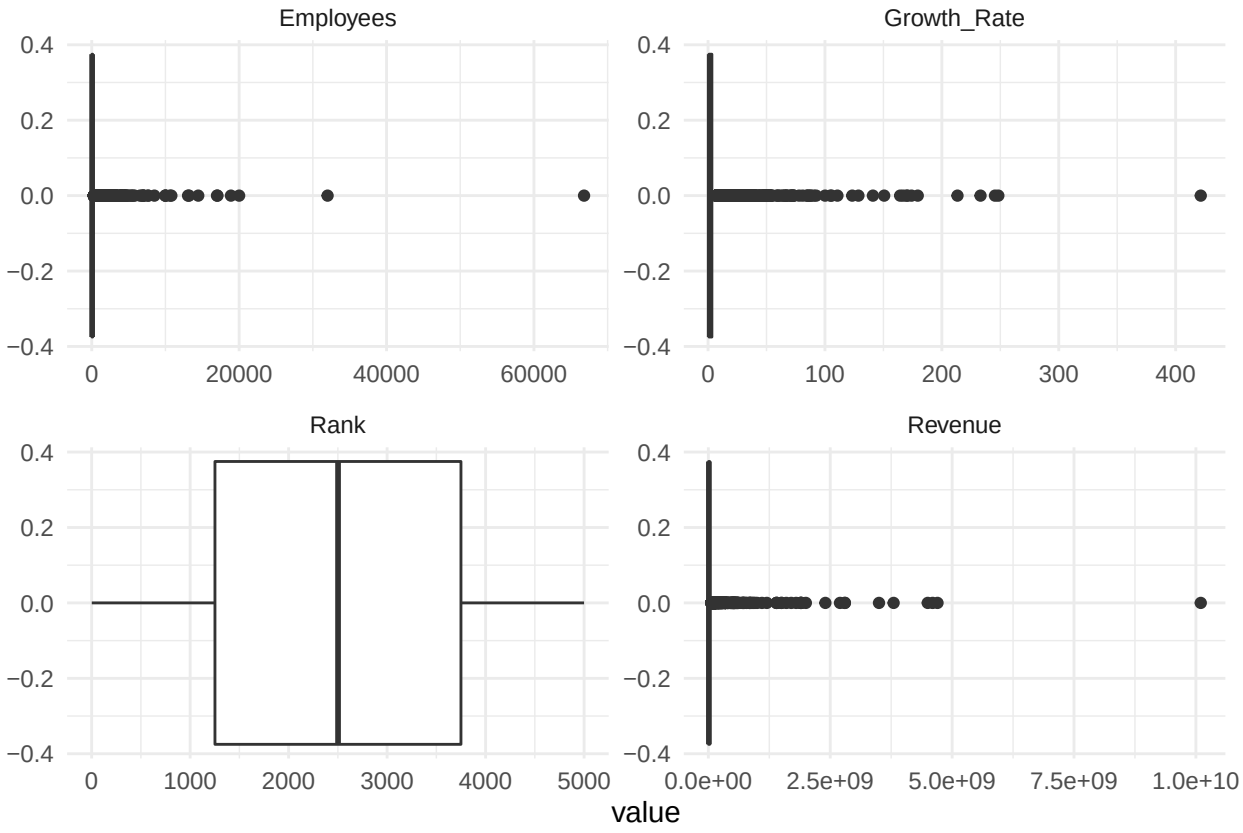
```
theme_set(theme_minimal()) # Set plot theme
sum(is.na(inc)) # 12 missing values
```

```
## [1] 12
```

```
inc[which(is.na(inc)),] # Print all missing value indecies
```

```
##       Rank Name Growth_Rate Revenue Industry Employees City State
## NA      NA <NA>          NA      NA     <NA>        NA <NA>  <NA>
## NA.1    NA <NA>          NA      NA     <NA>        NA <NA>  <NA>
## NA.2    NA <NA>          NA      NA     <NA>        NA <NA>  <NA>
## NA.3    NA <NA>          NA      NA     <NA>        NA <NA>  <NA>
## NA.4    NA <NA>          NA      NA     <NA>        NA <NA>  <NA>
## NA.5    NA <NA>          NA      NA     <NA>        NA <NA>  <NA>
## NA.6    NA <NA>          NA      NA     <NA>        NA <NA>  <NA>
## NA.7    NA <NA>          NA      NA     <NA>        NA <NA>  <NA>
## NA.8    NA <NA>          NA      NA     <NA>        NA <NA>  <NA>
## NA.9    NA <NA>          NA      NA     <NA>        NA <NA>  <NA>
## NA.10   NA <NA>          NA      NA     <NA>        NA <NA>  <NA>
## NA.11   NA <NA>          NA      NA     <NA>        NA <NA>  <NA>
```

```
# They hold no significance and can be removed completely
inc <- na.omit(inc) # remove missing values

# Only 8 variables; we can plot all
inc %>%
  dplyr::select(-Name, -City, -State, -Industry) %>% # 4 characters variables are useless in plot
  gather(key, value) %>% # gather into key value pairs
  ggplot(aes(value)) + # create ggplot
  geom_boxplot() + # as a geometric boxplot
  facet_wrap(~key, scales ="free") + # by each key
  theme(axis.ticks.x = element_blank()) # hide x axis tick marks
```



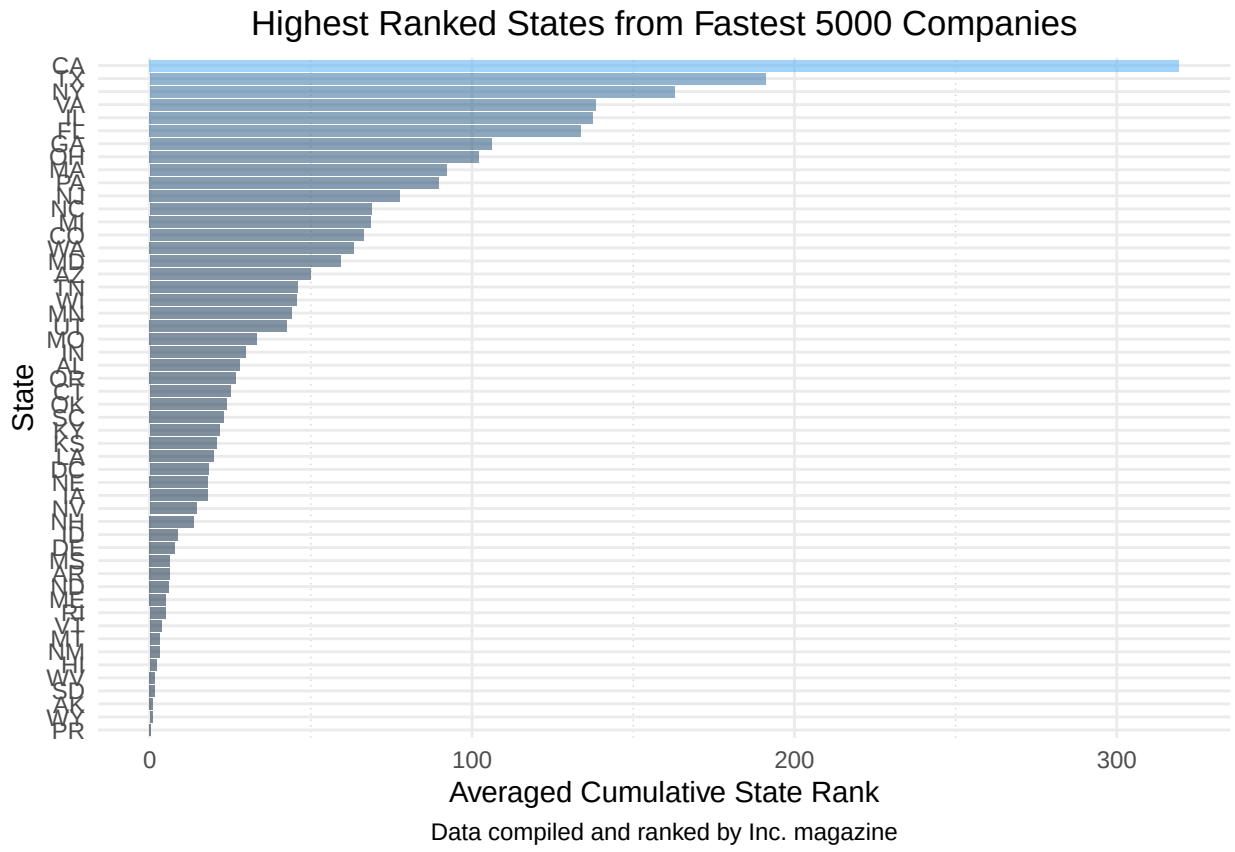

```
# Results:
# There are a lot of outliers
# Comparing other variables to rank which has no outliers
# The remaining variables are dominated by outliers

# Look closer at statistics
describe(inc)
```

```
##             vars    n        mean           sd    median     trimmed
## Rank           1 4989     2501.39      1443.42 2.502e+03     2501.47
## Name*          2 4989     2495.00      1440.34 2.495e+03     2495.00
## Growth_Rate    3 4989        4.61        14.14 1.420e+00        2.14
## Revenue        4 4989 48253357.39 240819468.86 1.090e+07 17328099.17
```

```
## Industry*      5 4989       12.09         7.33 1.300e+01       12.05
## Employees      6 4989      232.72      1353.13 5.300e+01       81.78
## City*          7 4989      730.98       440.33 7.600e+02      730.70
## State*         8 4989       24.80        15.63 2.300e+01       24.44
##                     mad     min        max      range  skew kurtosis         se
## Rank            1851.77 1.0e+00 5.0000e+03 4.9990e+03  0.00    -1.20      20.44
## Name*           1848.80 1.0e+00 4.9890e+03 4.9880e+03  0.00    -1.20      20.39
## Growth_Rate        1.22 3.4e-01 4.2148e+02 4.2114e+02 12.54   241.94       0.20
## Revenue     10674720.00 2.0e+06 1.0100e+10 1.0098e+10 22.15   721.05 3409454.05
## Industry*          8.90 1.0e+00 2.5000e+01 2.4000e+01 -0.10    -1.18       0.10
## Employees         53.37 1.0e+00 6.6803e+04 6.6802e+04 29.81  1268.67      19.16
## City*            603.42 1.0e+00 1.5170e+03 1.5160e+03 -0.04    -1.26       6.23
## State*            19.27 1.0e+00 5.2000e+01 5.1000e+01  0.12    -1.46       0.22
```

```
# Growth Rate, Revenue, and the number of Employees have high skew values
# Those same values are also curtailed sharply mid distribution
# These are to be expected in the fastest growing 5000 companies
```

```
# Out of curiosity
# Which states did best in their ranking?
inc %>%
  arrange(desc(Rank)) %>%
  group_by(State) %>%
  summarise(StateRank = (sum(Rank)/nrow(inc))) %>%
  ggplot(aes(reorder(State, StateRank), StateRank)) +
  geom_col(aes(fill = StateRank, alpha = .80)) + coord_flip() +
  labs(y = "Averaged Cumulative State Rank", x = "State",
       title = "Highest Ranked States from Fastest 5000 Companies", caption = "Data compiled and ranked
  theme(legend.position = "none",
        panel.grid.minor.x = element_line(color = "lightgrey",
                                          linetype = "dotted"),
        panel.grid.minor.y = element_line(color = "lightgrey",
                                          linetype = "dotted"),
        plot.title = element_text(hjust = 0.5),
        plot.caption = element_text(hjust = 0.5))
```

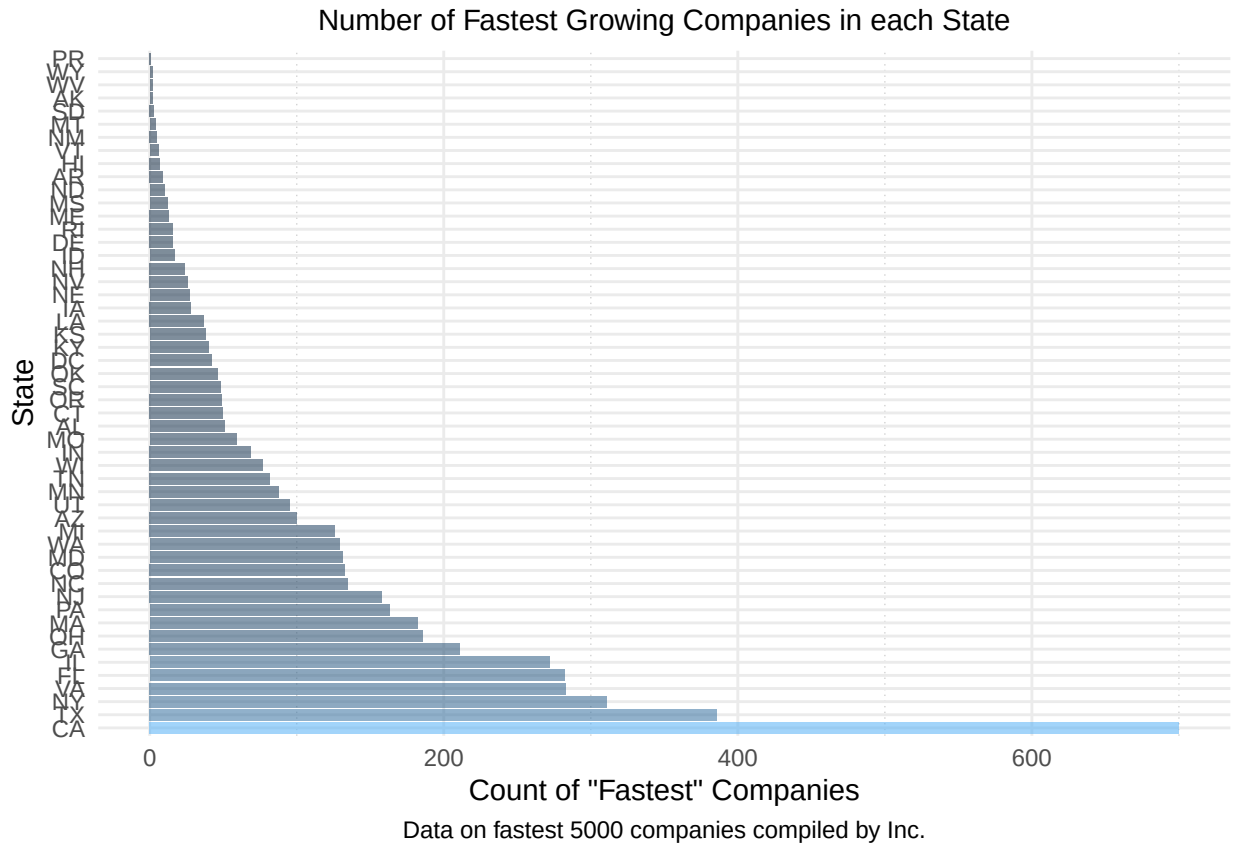



```
# Could have also used a histogram with stat = "count"
```

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
data.frame(table(inc$State)) %>%
  ggplot(aes(reorder(Var1, -Freq), Freq)) +
  geom_col(aes(fill = Freq, alpha = .80)) +
  coord_flip() + labs(x = "State", y = "Count of \"Fastest\" Companies",
    subtitle = "Number of Fastest Growing Companies in each State",
    caption = "Data on fastest 5000 companies compiled by Inc.") +
  theme(legend.position = "none",
        panel.grid.minor.x = element_line(color = "lightgrey",
                                          linetype = "dotted"),
        panel.grid.minor.y = element_line(color = "lightgrey",
                                          linetype = "dotted"),
        plot.subtitle = element_text(hjust = 0.5),
        plot.caption = element_text(hjust = 0.5))
```



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.

```
# This sounds like a boxplot/violin plot given an average/median employment by industry with a variable
# this would also show outliers unless removed
# However, the data is not averagable for the industry in the state
# Unless you did something like this
inc %>%
  filter(State == "NY") %>%
  group_by(Industry) %>%
  summarise(IndMed = median(Employees)) # Calc median employment by industry in this state
```

```
## # A tibble: 25 x 2
##    Industry                      IndMed
##    <chr>                          <dbl>
##  1 Advertising & Marketing         38
##  2 Business Products & Services    70.5
##  3 Computer Hardware               44
##  4 Construction                    24.5
```

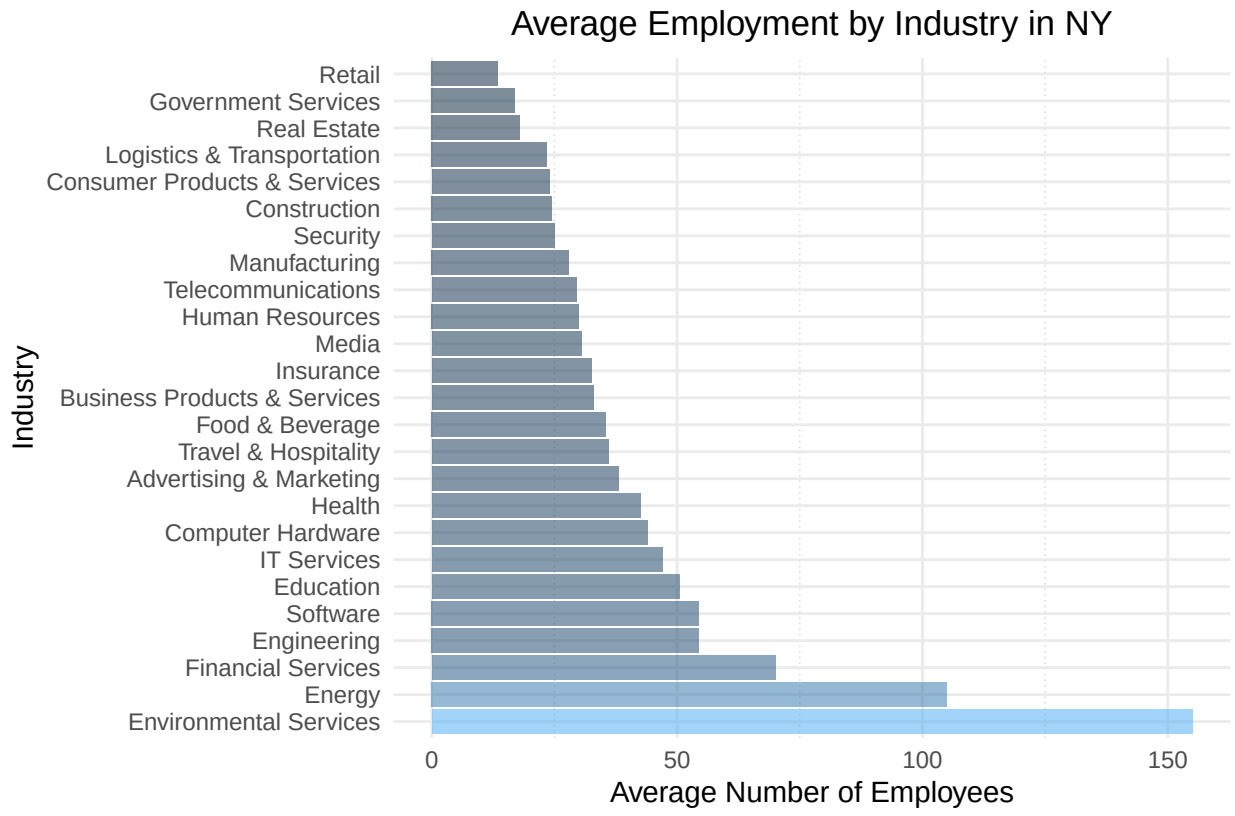
```
##  5 Consumer Products & Services   25
##  6 Education                      50.5
##  7 Energy                        120
##  8 Engineering                    54.5
##  9 Environmental Services        155
## 10 Financial Services             81
## # ... with 15 more rows
```

```
# Answer Question 2 here
data.frame(table(inc$State)) %>%
  arrange(desc(Freq))
```

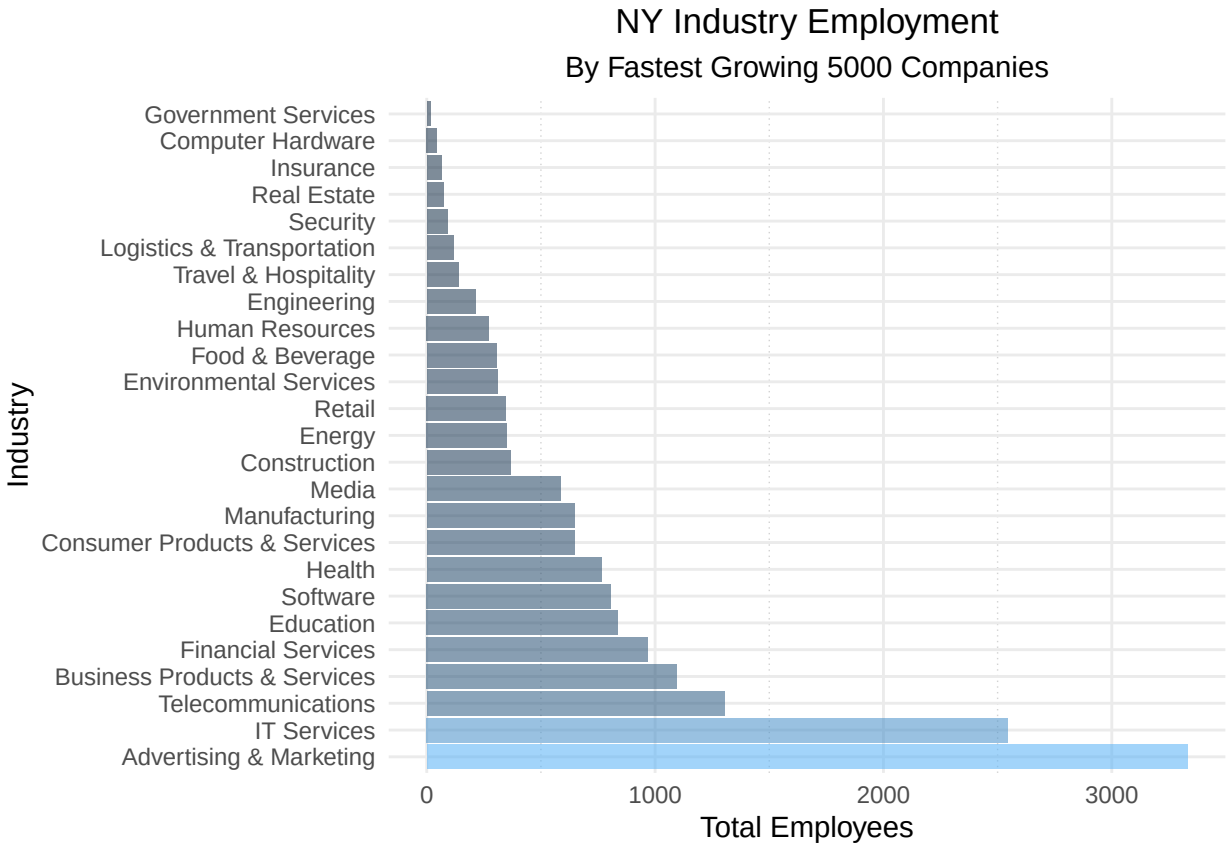
```
##    Var1 Freq
## 1    CA  700
## 2    TX  386
## 3    NY  311
## 4    VA  283
## 5    FL  282
## 6    IL  272
## 7    GA  211
## 8    OH  186
## 9    MA  182
## 10   PA  163
## 11   NJ  158
## 12   NC  135
## 13   CO  133
## 14   MD  131
## 15   WA  129
## 16   MI  126
## 17   AZ  100
## 18   UT   95
## 19   MN   88
## 20   TN   82
## 21   WI   77
## 22   IN   69
## 23   MO   59
## 24   AL   51
## 25   CT   50
## 26   OR   49
## 27   SC   48
## 28   OK   46
## 29   DC   42
## 30   KY   40
## 31   KS   38
## 32   LA   37
## 33   IA   28
## 34   NE   27
## 35   NV   26
## 36   NH   24
## 37   ID   17
## 38   DE   16
## 39   RI   16
## 40   ME   13
## 41   MS   12
```

```
## 42   ND   10
## 43   AR    9
## 44   HI    7
## 45   VT    6
## 46   NM    5
## 47   MT    4
## 48   SD    3
## 49   AK    2
## 50   WV    2
## 51   WY    2
## 52   PR    1
```

```r
# Based on this table of frequencies; NY is 3rd

# Remove outliers based on IQR of R boxplot
outs <- boxplot(inc$Employees, plot=F)$out
inc <- inc[-which(inc$Employees %in% outs),]

# Applying the above method on average employment per industry we have
inc %>%
  filter(State == "NY") %>%
  group_by(Industry) %>%
  summarise(IndMed = median(Employees)) %>%
  ggplot(aes(reorder(Industry, -IndMed)), IndMed) +
  geom_col(aes(y = IndMed, fill = IndMed, alpha = .80)) +
  coord_flip() + labs(y = "Average Number of Employees", x = "Industry",
                      title = "Average Employment by Industry in NY",
                      caption = "Contains data on fastest 5000 companies as compiled by Inc. magazine")
  theme(legend.position = "none",
        panel.grid.minor.x = element_line(color = "lightgrey",
                                          linetype = "dotted"),
        panel.grid.minor.y = element_line(color = "lightgrey",
                                          linetype = "dotted"),
        plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        plot.caption = element_text(hjust = 0.5))
```




```
# Total employees per industry in the state of NY
inc %>%
  filter(State == "NY") %>%
  group_by(Industry) %>%
  summarise(TotalEmployees = sum(Employees)) %>%
  ggplot(aes(reorder(Industry, -TotalEmployees)), TotalEmployees) +
  geom_col(aes(y = TotalEmployees, fill = TotalEmployees, alpha = .80)) +
  coord_flip() + labs(y = "Total Employees", x = "Industry",
                      title = "NY Industry Employment",
                      subtitle = "By Fastest Growing 5000 Companies") +
  theme(legend.position = "none",
        panel.grid.minor.x = element_line(color = "lightgrey",
                                          linetype = "dotted"),
        panel.grid.minor.y = element_line(color = "lightgrey",
                                          linetype = "dotted"),
        plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        plot.caption = element_text(hjust = 0.5))
```



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 %>%
  filter(State == "NY") %>%
  dplyr::select(Industry, Revenue, Employees) %>%
  mutate(RevPerEmp = Revenue / Employees) %>%
  group_by(Industry) %>%
  summarise(TotRevPerEmp = (sum(RevPerEmp)/1000000)) %>%
  ggplot(aes(reorder(Industry, -TotRevPerEmp), TotRevPerEmp)) +
  geom_col(aes(y = TotRevPerEmp,
           fill = TotRevPerEmp, alpha = .80)) + coord_flip() +
  labs(y = "Revenue Per Employee (Millions of Dollars)", x = "Industry",
                      title = "Total Revenue per Employee in NY",
                      caption = "Contains data on fastest 5000 companies as compiled by Inc. magazine")
  theme(legend.position = "none",
        panel.grid.minor.x = element_line(color = "lightgrey",
                                          linetype = "dotted"),
        panel.grid.minor.y = element_line(color = "lightgrey",
                                          linetype = "dotted"),
        plot.title = element_text(hjust = 0.5),
```

```
        plot.subtitle = element_text(hjust = 0.5),
        plot.caption = element_text(hjust = 0.5))
```