Module1

Zachary Palmore

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)
library(latex2exp)
library(ggpubr)
library(psych)
theme_set(theme_minimal()) # Set plot theme
sum(is.na(inc)) # 12 missing values
```

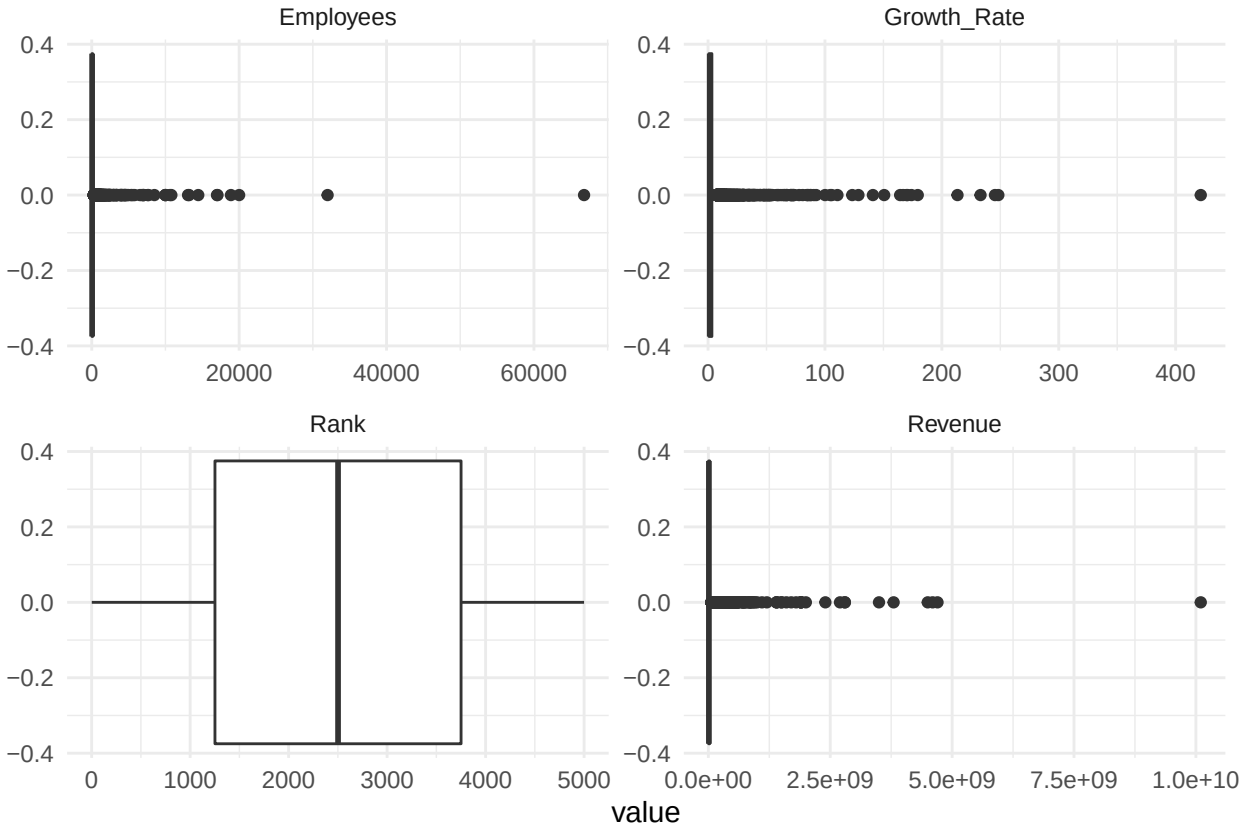
```
## [1] 12
```

```
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.omitted <- na.omit(inc) # remove missing values; equivalent to complete.cases() but worked up front
inc.ccs <- inc[complete.cases(inc),] # To show equivalency
```

```
# 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 at NY
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
```

```r
# Look at how industries are represented
inc %>%
  dplyr::count(Industry, sort=T)
```

```
##                        Industry   n
## 1                   IT Services 733
## 2  Business Products & Services 482
## 3       Advertising & Marketing 471
## 4                        Health 355
## 5                      Software 342
## 6             Financial Services 260
## 7                 Manufacturing 256
## 8  Consumer Products & Services 203
## 9                        Retail 203
## 10          Government Services 202
## 11              Human Resources 196
## 12                 Construction 187
## 13   Logistics & Transportation 155
## 14              Food & Beverage 131
## 15          Telecommunications 129
## 16                      Energy 109
## 17                 Real Estate  96
## 18                   Education  83
## 19                 Engineering  74
## 20                    Security  73
## 21         Travel & Hospitality  62
## 22                       Media  54
## 23       Environmental Services  51
## 24                   Insurance  50
## 25            Computer Hardware  44
```

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

```
##             vars    n        mean           sd    median     trimmed
## Rank           1 5001     2501.64      1443.51 2.502e+03     2501.73
## Name*          2 5001     2501.00      1443.81 2.501e+03     2501.00
## Growth_Rate    3 5001        4.61        14.12 1.420e+00        2.14
## Revenue        4 5001 48222535.49 240542281.14 1.090e+07 17334966.26
## Industry*      5 5001       12.10         7.33 1.300e+01       12.05
## Employees      6 4989      232.72      1353.13 5.300e+01       81.78
## City*          7 5001      732.00       441.12 7.610e+02      731.74
## State*         8 5001       24.80        15.64 2.300e+01       24.44
##                    mad     min        max      range  skew kurtosis         se
## Rank            1853.25 1.0e+00 5.0000e+03 4.9990e+03  0.00    -1.20      20.41
## Name*           1853.25 1.0e+00 5.0010e+03 5.0000e+03  0.00    -1.20      20.42
## Growth_Rate        1.22 3.4e-01 4.2148e+02 4.2114e+02 12.55   242.34       0.20
## Revenue     10674720.00 2.0e+06 1.0100e+10 1.0098e+10 22.17   722.66 3401441.44
## 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*            604.90 1.0e+00 1.5190e+03 1.5180e+03 -0.04    -1.26       6.24
## 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
```

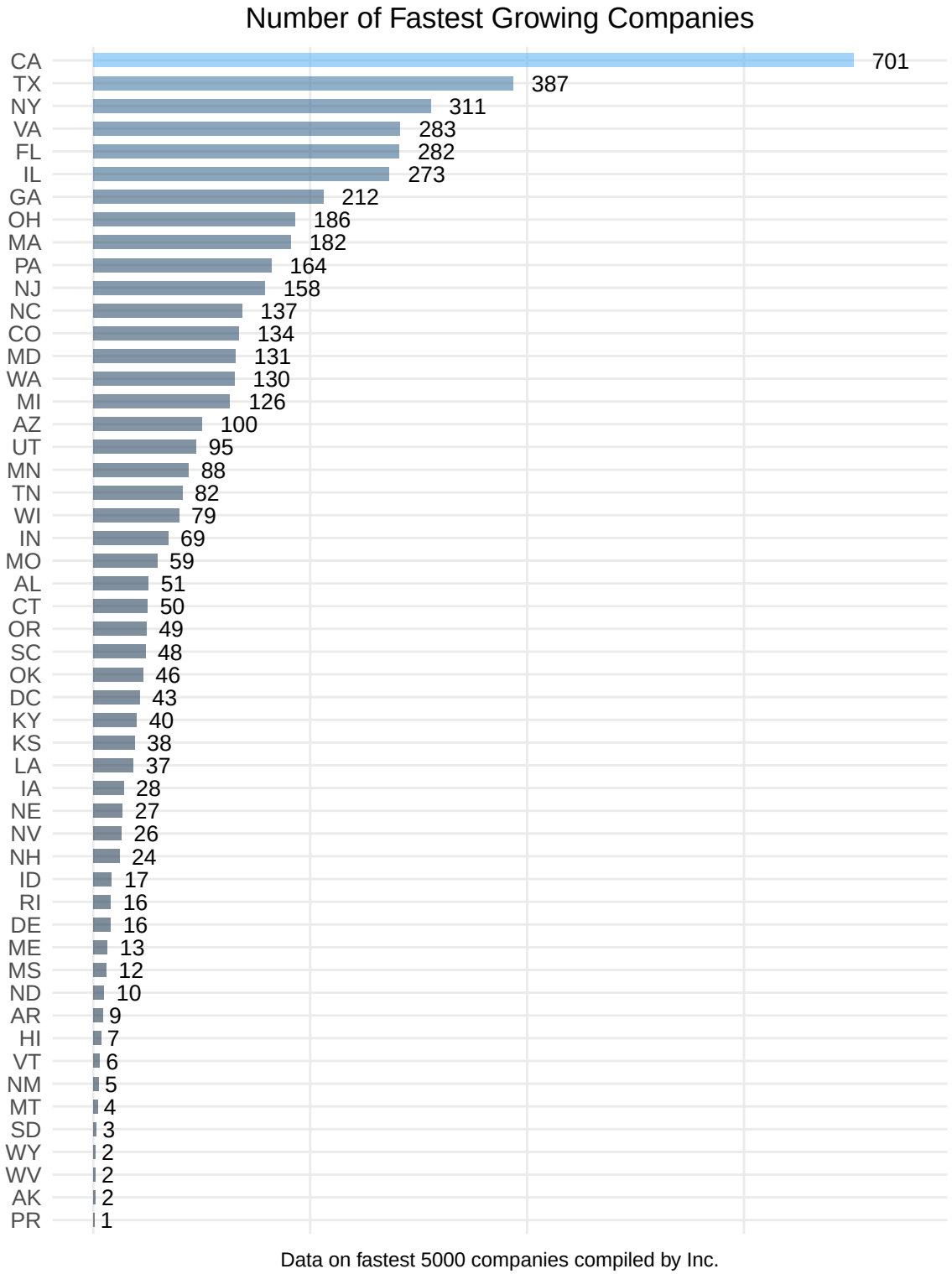
```
# Out of curiosity
# Which states did best in their ranking?
inc %>%
  arrange(desc(Rank)) %>%
  group_by(State) %>%
  summarise(StateRank = (sum(Rank)/nrow(inc))) %>%
  ggplot(aes(reorder(State, StateRank), StateRank)) +
  geom_col(aes(fill = StateRank, alpha = .80)) + coord_flip() +
  labs(y = "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(y = reorder(Var1, Freq), x = Freq)) +
  geom_col(aes(fill = Freq, alpha = .80, width = .6)) +
  labs(x = "State", y = "Count of \"Fastest\" Companies",
    title = "Number of Fastest Growing Companies",
    caption = "Data on fastest 5000 companies compiled by Inc.") +
  geom_text(aes(label = round(Freq, 0)), size = 3.6, hjust = -.5) +
  xlim(c(0,750)) +
  theme(legend.position = "none",
        panel.grid.minor.x = element_blank(),
        panel.grid.minor.y = element_line(color = "lightgrey",
                                          linetype = "dotted", size = 3),
        plot.title = element_text(hjust = 0.5),
        plot.caption = element_text(hjust = 0.5),
        axis.ticks.x = element_blank(),
        axis.ticks.y = element_blank(),
        axis.text.y = element_text(size = 10),
        axis.text.x = element_blank(),
        axis.title.x = element_blank(),
        axis.title.y = element_blank(),)
```



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
# Method of removing outliers? None specified.
# We view several cases to see the full picture to decide
library(ggpubr)
# Let's first view it pre-outlier removal

# Find state to filter by
data.frame(table(inc$State)) %>%
  arrange(desc(Freq))
# Based on this table of frequencies; NY is 3rd

# Create boxplot with complete cases and all outliers shown
box.inc.ccs <- inc.ccs %>%
  filter(State == "NY") %>%
  ggplot(aes(x = reorder(Industry, Employees, median), Employees)) +
  geom_boxplot(aes(fill = median(Employees), alpha = .75)) + coord_flip() +
  stat_summary(fun.y=median, geom="point", shape=21,
               size=2, color="lightblue", alpha = 0.9, fill = "white") +
  labs(y = "Employees", x = "Industry",
       title = "NY Employment by Industry",
       subtitle = "Median shown as light blue dot for each industry",
       caption = "From 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),
        axis.ticks.x = element_blank(),
        axis.text.y = element_text(face = "bold", size = 12),
        axis.text.x = element_text(size = 12),
        axis.title.x = element_text(face = "bold", size = 12),
        axis.title.y = element_blank())  +
  scale_x_discrete(limits = rev(levels(inc$Var1)), expand = c(-0.8, -4))  +
  annotate("text", x = 3, y = 2000, label = "No Outliers Removed")

# Now create boxplot with two clearest outliers not shown
box.inc.tworemoved <- inc.ccs %>%
  filter(State == "NY") %>%
  ggplot(aes(x = reorder(Industry, Employees, median), Employees)) +
  geom_boxplot(aes(fill = median(Employees), alpha = .75)) + coord_flip() +
  stat_summary(fun.y=median, geom="point", shape=21,
               size=2, color="lightblue", alpha = 0.9, fill = "white") +
  labs(y = "Employees", x = "Industry",
       title = "NY Employment by Industry",
       subtitle = "Median shown as light blue dot for each industry",
       caption = "From data on fastest 5000 companies as compiled by Inc. magazine") +
  theme(legend.position = "none",
```

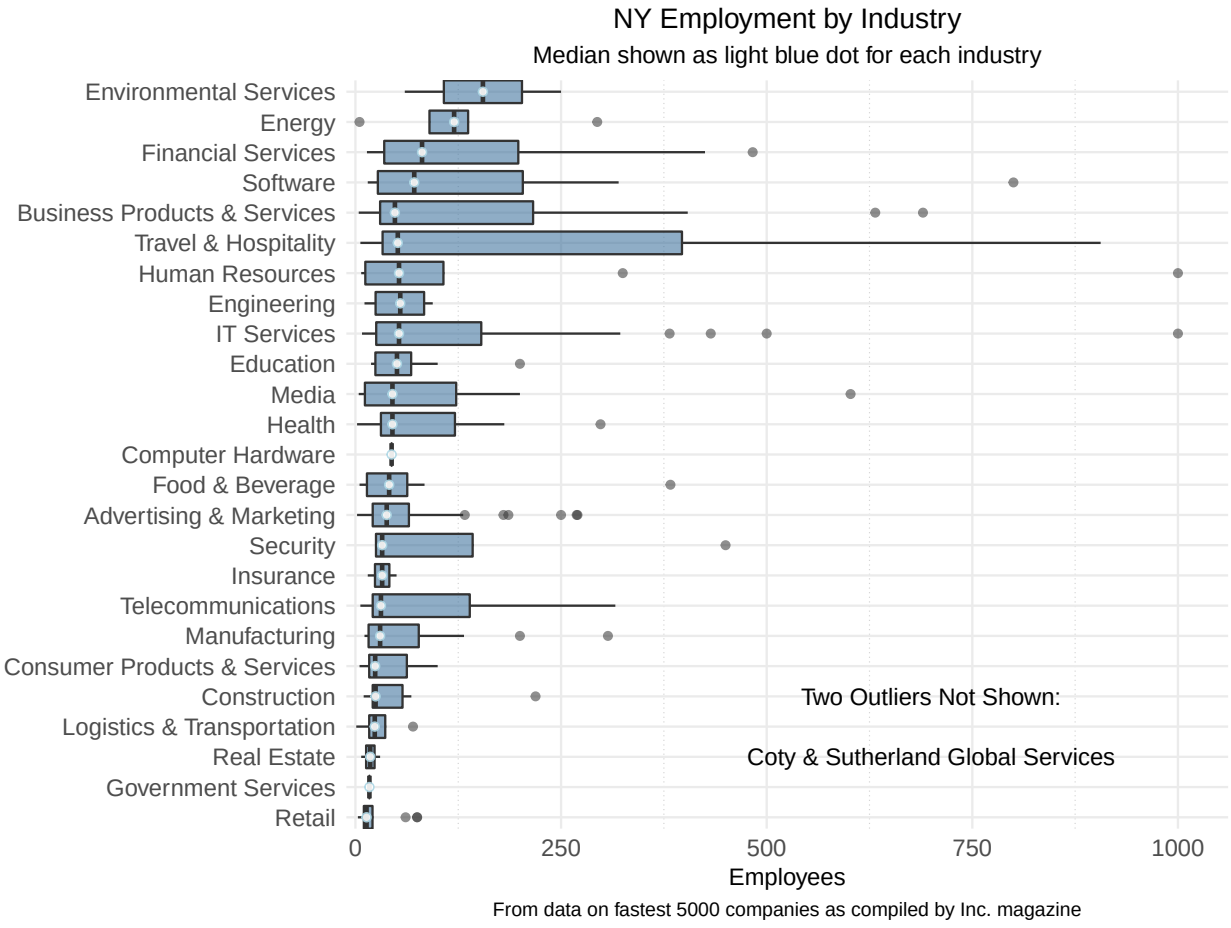
```
        panel.grid.minor.x = element_line(color = "lightgrey",
                                          linetype = "dotted"),
        panel.grid.minor.y = element_line(color = "lightgrey",
                                          linetype = "dotted"),
        plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        plot.caption = element_text(hjust = 0.5),
        axis.ticks.x = element_blank(),
        axis.text.y = element_text(face = "bold", size = 12),
        axis.text.x = element_text(size = 12),
        axis.title.x = element_text(face = "bold", size = 12),
        axis.title.y = element_blank())  +
  scale_x_discrete(limits = rev(levels(inc$Var1)), expand = c(-0.8, -4))  +
  annotate("text", x = 3, y = 600, label = "Two Outliers Removed") +
  scale_y_continuous(limits = c(0, 1000), breaks = seq(0, 1000, 500), expand = c(0.01, 0.5))

# Alternative options for removal and clarity
# Remove outliers based on IQR of R boxplot
outs <- boxplot(inc.ccs$Employees, plot=F)$out
outs.num <- outs %>% as.numeric() %>% data.frame()
outs.num %>%
  count() # 610 Outliers Identified via Boxplot
# Down select to remove them
inc.boxremoval <- inc.ccs[-which(inc.ccs$Employees %in% outs),]

# Applying the above method on average employment per industry we have this
# Boxplot with outliers determined by 1.5 times each industy's IQR removed
box.inc.boxremoval <- inc.boxremoval %>%
  filter(State == "NY") %>%
  ggplot(aes(x = reorder(Industry, Employees, median), Employees)) +
  geom_boxplot(aes(fill = median(Employees), alpha = .75)) + coord_flip() +
  stat_summary(fun.y=median, geom="point", shape=21,
               size=2, color="lightblue", alpha = 0.9, fill = "white") +
  labs(y = "Employees", x = "Industry",
       title = "NY Employment by Industry",
       subtitle = "Median shown as light blue dot for each industry",
       caption = "From 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),
        axis.ticks.x = element_blank(),
        axis.text.y = element_text(face = "bold", size = 12),
        axis.text.x = element_text(size = 12),
        axis.title.x = element_text(face = "bold", size = 12),
        axis.title.y = element_blank())  +
  scale_x_discrete(limits = rev(levels(inc$Var1)), expand = c(-0.8, -4)) +
  annotate("text", x = 2, y = 164, label = "Outliers Removed beyond 1.5xIQR")
#  scale_y_continuous(limits = c(0, 200), breaks = seq(0, 200, 50), expand = c(0.01, 0.5)) +
```

```
inc.boxremoval %>%
  filter(State == "NY") %>%
  filter(Industry == "Government Services") # Cipher Tech Solutions - only one in NY for "Government Se
inc.boxremoval %>%
  filter(State == "NY") %>%
  filter(Industry == "Environmental Services") # Creative Environment Solutions and one other environme
```

```
ggarrange(box.inc.ccs, box.inc.tworemoved, box.inc.boxremoval)
```

```
inc.ccs %>%
  filter(State == "NY") %>%
  filter(Industry == "Consumer Products & Services") %>%
  arrange(desc(Employees))
inc.ccs %>%
  filter(State == "NY") %>%
  filter(Industry == "Business Products & Services") %>%
  arrange(desc(Employees))
```

```
# Answer Question 2 here
inc.ccs %>%
  filter(State == "NY") %>%
  ggplot(aes(x = reorder(Industry, Employees, median), Employees)) +
  geom_boxplot(aes(fill = median(Employees), alpha = .75)) + coord_flip() +
  stat_summary(fun.y=median, geom="point", shape=21,
               size=1.6, color="lightblue", alpha = 0.9, fill = "white") +
  labs(y = "Employees", x = "Industry",
       title = "NY Employment by Industry",
       subtitle = "Median shown as light blue dot for each industry",
       caption = "From 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),
        axis.ticks.x = element_blank(),
        axis.text.y = element_text(size = 11),
        axis.text.x = element_text(size = 11),
        axis.title.x = element_text(size = 11),
        axis.title.y = element_blank())  +
  scale_x_discrete(limits = rev(levels(inc$Var1)), expand = c(-0.8, -4))  +
  annotate("text", x = 5, y = 700, label = "Two Outliers Not Shown:") +
  annotate("text", x = 3, y = 700, label = "Coty & Sutherland Global Services") +
  scale_y_continuous(limits = c(0, 1050), breaks = seq(0, 1000, 250), expand = c(0.01, 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.

```
# Testing Question 3 here
# There is more than one way to calculate "which industries generate the most revenue per employee"
# We take a look at NY out of curiosity for comparison
inc.ccs %>%
  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 = "Dollars (Millions)", x = "Industry",
                      title = "Revenue Per Employee (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))
```

```
# Then without NY filter (for entire US)
# Example 1: calculates the revenue per employee as a per capita dollar value in each company first
# then sums the dollars values within each industry
inc.ccs %>%
  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 = "Dollars (Millions)", x = "Industry",
                      title = "U.S. Revenue Per Employee",
                      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),
        axis.ticks.x = element_blank(),
        axis.text.y = element_text(face = "bold", size = 12),
        axis.text.x = element_text(size = 12),
        axis.title.x = element_text(face = "bold", size = 12),
        axis.title.y = element_blank())
```

```
# Scaled features
inc.ccs %>%
  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 = "Dollars (Millions)", x = "Industry",
                      title = "U.S. Revenue Per Employee",
                      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),
```

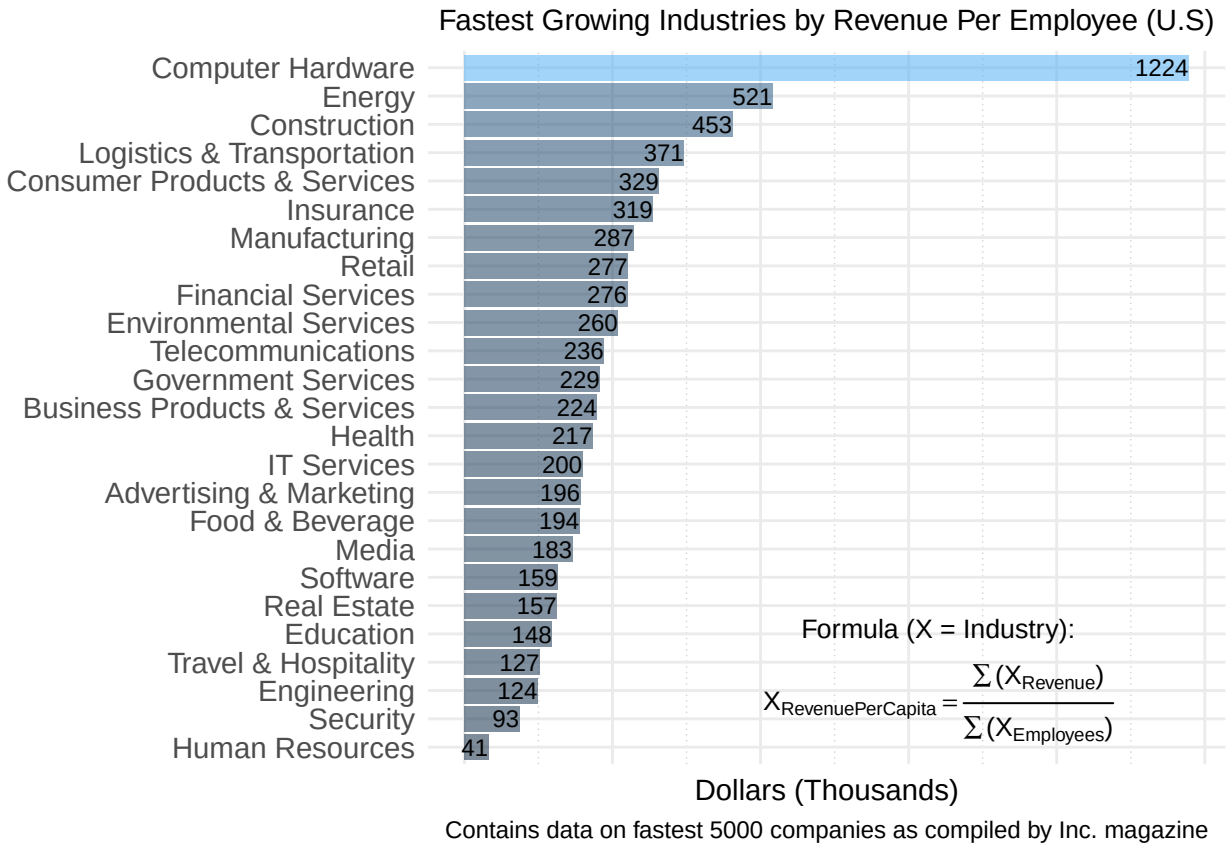
```
        plot.subtitle = element_text(hjust = 0.5),
        plot.caption = element_text(hjust = 0.5),
        axis.ticks.x = element_blank(),
        axis.text.y = element_text(face = "bold", size = 12),
        axis.text.x = element_text(size = 12),
        axis.title.x = element_text(face = "bold", size = 12),
        axis.title.y = element_blank()) +
  scale_x_discrete(limits = rev(levels(inc$Var1)), expand = c(-0.8, -4))  +
  scale_y_continuous(limits = c(0, 200), breaks = seq(0, 200, 50), expand = c(0.01, 0.5))
```

```
# Example 2: calculates the total revenue as a dollar amount for each industry and
# calculates separately the total number of employees in each industry
# then creates a ratio of revenue dollars to the number of employees in each industry to find
# the industry revenue per employee
inc.ccs %>%
  dplyr::select(Industry, Revenue, Employees) %>%
  group_by(Industry) %>%
  summarise(TotalRev = sum(Revenue),
            TotalEmp = sum(Employees),
            TotRevPerEmp = (TotalRev / TotalEmp)/1000) %>%
  ggplot(aes(reorder(Industry, TotRevPerEmp), TotRevPerEmp)) +
  geom_col(aes(y = TotRevPerEmp,
           fill = TotRevPerEmp, alpha = .80)) + coord_flip() +
  labs(y = "Dollars (Thousands)", x = "Industry",
                      title = "U.S. Revenue Per Employee",
                      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),
        axis.ticks.x = element_blank(),
        axis.text.y = element_text(face = "bold", size = 12),
        axis.text.x = element_text(size = 12),
        axis.title.x = element_text(face = "bold", size = 12),
        axis.title.y = element_blank())
```

```
# Example 3: calculates the revenue per employee as a per capita dollar value in each company
# then finds the mean of those values for each industry
inc.ccs %>%
  dplyr::select(Industry, Revenue, Employees) %>%
  mutate(RevPerEmp = Revenue / Employees) %>%
  group_by(Industry) %>%
  summarise(TotRevPerEmp = mean(RevPerEmp)/1000) %>%
  ggplot(aes(reorder(Industry, TotRevPerEmp), TotRevPerEmp)) +
  geom_col(aes(y = TotRevPerEmp,
           fill = TotRevPerEmp, alpha = .80)) + coord_flip() +
  labs(y = "Dollars (Thousands)", x = "Industry",
                      title = "Revenue Per Employee (U.S)",
                      caption = "Contains data on fastest 5000 companies as compiled by Inc. magazine")
```

```
  geom_text(aes(label = round(TotRevPerEmp, 0)), size = 3, hjust = 1.19) +
  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),
        axis.ticks.x = element_blank(),
        axis.text.y = element_text(face = "bold", size = 12),
        axis.text.x = element_blank(),
        axis.title.x = element_text(face = "bold", size = 12),
        axis.title.y = element_blank())
```

```
# Answer Question 3 here
inc.ccs %>%
  dplyr::select(Industry, Revenue, Employees) %>%
  group_by(Industry) %>%
  summarise(TotalRev = sum(Revenue),
            TotalEmp = sum(Employees),
            TotRevPerEmp = (TotalRev / TotalEmp)/1000) %>%
  ggplot(aes(reorder(Industry, TotRevPerEmp), TotRevPerEmp)) +
  geom_col(aes(y = TotRevPerEmp,
           fill = TotRevPerEmp, alpha = .80)) + coord_flip() +
  labs(y = "Dollars (Thousands)", x = "Industry",
                      subtitle = "Fastest Growing Industries by Revenue Per Employee (U.S)",
                      caption = "Contains data on fastest 5000 companies as compiled by Inc. magazine")
  geom_text(aes(label = round(TotRevPerEmp, 0)), size = 3, hjust = 1.01) +
  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),
        axis.ticks.x = element_blank(),
        axis.text.y = element_text(size = 11),
        axis.text.x = element_blank(),
        axis.title.x = element_text(size = 11),
        axis.title.y = element_blank())   +
  annotate("text", x = 5.2, y = 800, label = "Formula (X = Industry):", size = 3.5) +
  annotate("text", x = 2.6, y = 800,
           label = latex2exp::TeX("$X_{RevenuePerCapita} = \\frac{\\sum(X_{Revenue})}{\\sum(X_{Employees
           size = 3.5, output='character', parse=T)
```


Fastest Growing Industries by Revenue Per Employee (U.S)
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
1224
521
453
371
329
319
287
277
276
260
236
229
224
217
200
196
194
183
159
157
148
127
124
93
41
Formula (X = Industry):
$X_{RevenuePerCapita} = \frac{\sum (X_{Revenue})}{\sum (X_{Employees})}$
Dollars (Thousands)
Contains data on fastest 5000 companies as compiled by Inc. magazine