CUNY DATA 608: HW #1

Zachary Herold

Sept. 15, 2019

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:

Loading the csv file and preserving only records with no missing values

```
inc <- read.csv("https://raw.githubusercontent.com/ZacharyHerold/CUNY-DATA-608/master/inc5000_dat
a.csv", header= TRUE)
inc <- inc[complete.cases(inc),]
```

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
##  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   :2501   110 Consulting        :   1   Mean   :  4.615
##  3rd Qu.:3750   11thStreetCoffee.com  :   1   3rd Qu.:  3.290
##  Max.   :5000   123 Exteriors         :   1   Max.   :421.480
##                 (Other)               :4983
##     Revenue                                    Industry      Employees
##  Min.   :2.000e+06   IT Services                 : 732   Min.   :    1.0
##  1st Qu.:5.100e+06   Business Products & Services: 480   1st Qu.:   25.0
##  Median :1.090e+07   Advertising & Marketing     : 471   Median :   53.0
##  Mean   :4.825e+07   Health                      : 354   Mean   :  232.7
##  3rd Qu.:2.860e+07   Software                    : 341   3rd Qu.:  132.0
##  Max.   :1.010e+10   Financial Services          : 260   Max.   :66803.0
##                      (Other)                     :2351
##             City          State
##  New York     : 160   CA     : 700
##  Chicago      :  90   TX     : 386
##  Austin       :  88   NY     : 311
##  Houston      :  76   VA     : 283
##  San Francisco:  74   FL     : 282
##  Atlanta      :  73   IL     : 272
##  (Other)      :4428   (Other):2755
```

Observations:

We observe the minimum qualifying revenue for inclusion is $2 million.

The quantitative data (Growth_rate, Revenue, Employees) have maximum values far beyond the 3rd quartile values, reflecting a heavy rightward skew due to outliers. Indeed, the mean is beyond the 3rd quartile values for each of the three variables. In the case of Growth_Rate, this is likely because a low starting base, while for Revenue and Employees, this is from the inclusion of companies of scale beyond the norm.

From the data it is unclear if the Growth_Rate is expressed as a percentage or decimal. We would assume that these are decimals with a minimum Growth_Rate of 34% since the dataset is looking at fast-growing companies. For growth rate to have meaning, there should be a minimum number of years for the company to be in operation, since year-on-year growth will be inflated when the base revenue of the year prior is very low, something expected in the initial years of operation.

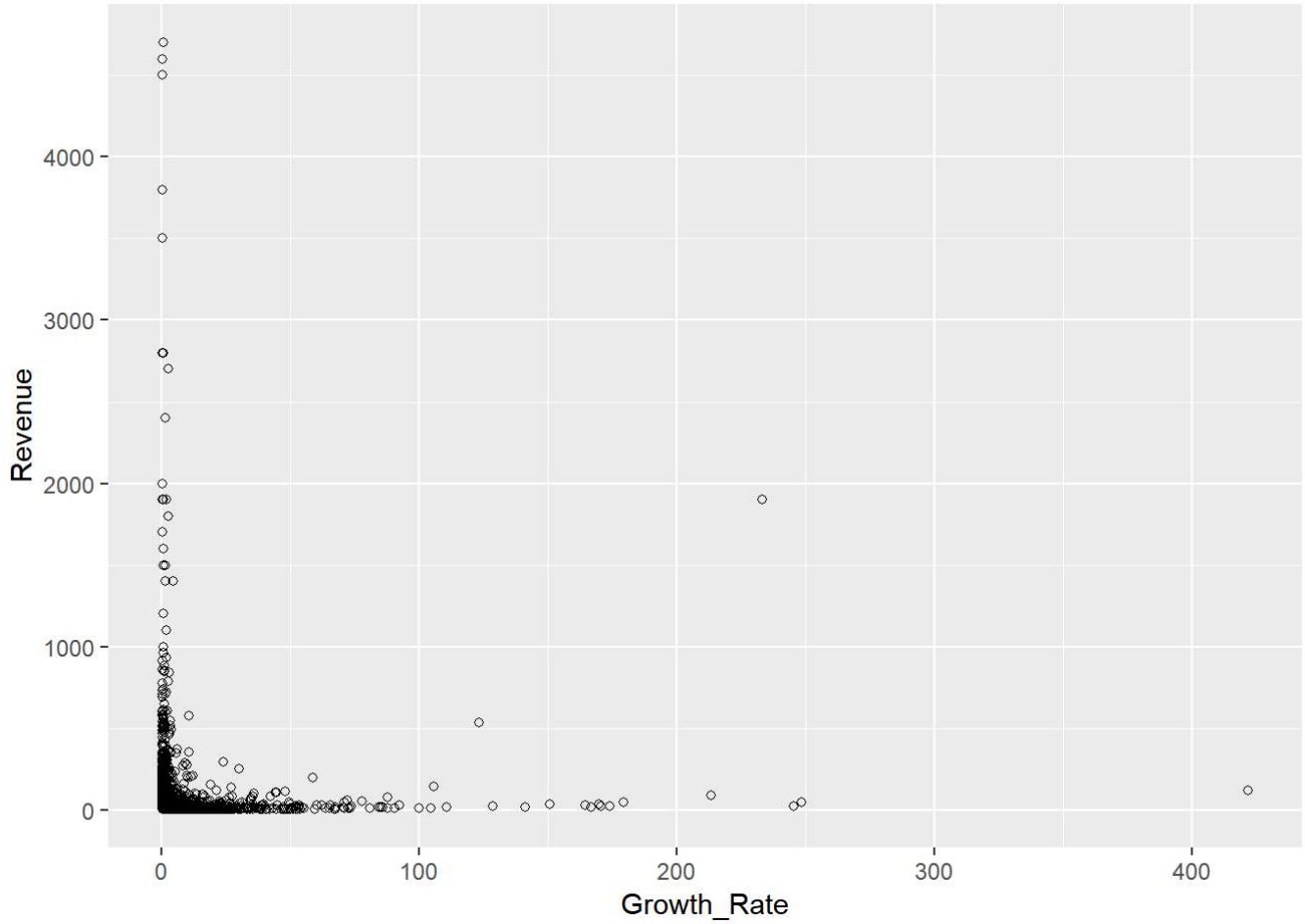
To better show the revenue amounts, we convert the values into million dollars.

```
inc$Revenue <- inc$Revenue  / 1000000
order.Rev <- order(inc$Revenue)
tail(inc[order.Rev, ],6)
```

```
##      Rank                         Name Growth_Rate Revenue
## 4246 4246 American Tire Distributors        0.59    3500
## 4716 4716              Westcon Group        0.44    3800
## 4997 4997                  Dot Foods        0.34    4500
## 4936 4936                       Coty        0.36    4600
## 3853 3853                 ABC Supply        0.73    4700
## 4788 4788                        CDW        0.41   10100
##                            Industry Employees         City State
## 4246 Consumer Products & Services      3341 Huntersville    NC
## 4716                  IT Services      3000    Tarrytown    NY
## 4997              Food & Beverage      3919 Mt. Sterling    IL
## 4936 Consumer Products & Services     10000     New York    NY
## 3853                 Construction      6549       Beloit    WI
## 4788            Computer Hardware      6800 Vernon Hills    IL
```

From the scatterplot below, we see two distinct company types clusted around the axes: large companies with relative low growth and small companies with extremely high growth.

```
inc <- subset(inc, Revenue < 5000) #removing CDW (Computer Hardware) due to its incomparable scale
ggplot(inc, aes(x = Growth_Rate, y = Revenue)) + geom_point(shape = 21)
```




```
order.Growth <- order(inc$Growth_Rate)
tail(inc[order.Growth, ],6)
```

```
##   Rank                        Name Growth_Rate Revenue
## 6    6 MileStone Community Builders      179.38    45.7
## 5    5                      DataXu      213.37    87.0
## 4    4                     Bridger      233.08  1900.0
## 3    3               The HCI Group      245.45    25.5
## 2    2       FederalConference.com      248.31    49.6
## 1    1                        Fuhu      421.48   117.9
##                       Industry Employees         City State
## 6                  Real Estate        63       Austin    TX
## 5     Advertising & Marketing       220       Boston    MA
## 4                      Energy        50      Addison    TX
## 3                      Health       132 Jacksonville    FL
## 2         Government Services        51     Dumfries    VA
## 1 Consumer Products & Services       104   El Segundo    CA
```

The energy company Bridger stands out above the rest as one that has extremely brisk growth along with $1.9 billion in revenue.

Despite the observation of the two classes of companies, there is not a negative correlation between Growth_Rate and Revenue as would be expected.

```
cor(inc$Growth_Rate, inc$Revenue)
```

```
## [1] 0.01093116
```

The lack of strong relationship between Revenue and Growth_Rate is further reflected in the high p-value of the linear regression.

```
mod <- lm(Growth_Rate ~ Revenue, inc)
summary(mod)
```

```
##
## Call:
## lm(formula = Growth_Rate ~ Revenue, data = inc)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7.88  -3.84  -3.19  -1.32 416.81
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.5787780  0.2057971  22.249   <2e-16 ***
## Revenue     0.0007956  0.0010306   0.772     0.44
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.14 on 4986 degrees of freedom
## Multiple R-squared:  0.0001195,  Adjusted R-squared:  -8.105e-05
## F-statistic: 0.5958 on 1 and 4986 DF,  p-value: 0.4402
```

Question 1

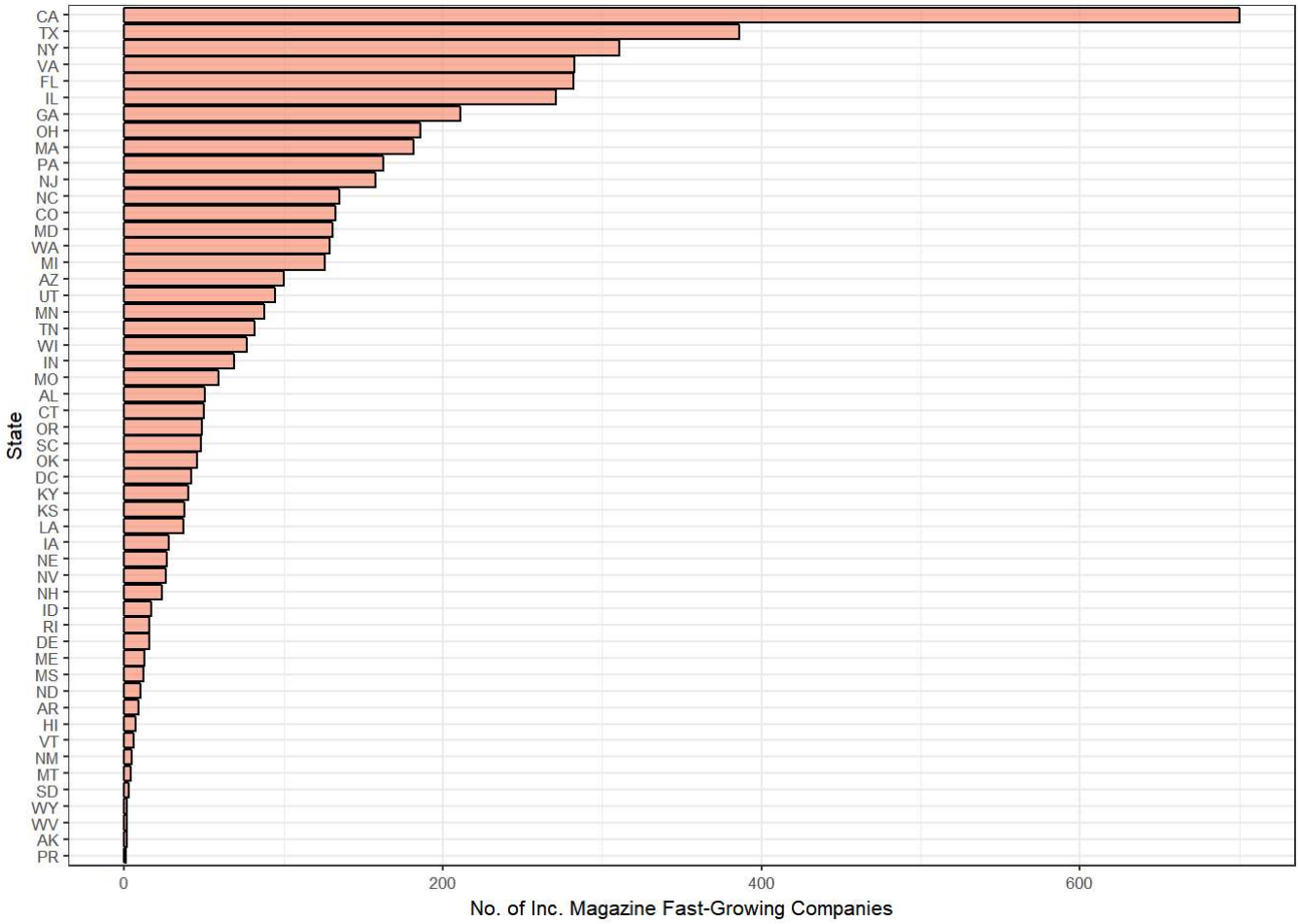
Create a graph that shows the distribution of companies in the dataset by State (ie how many are in each state). There are a lot of States, so consider which axis you should use. This visualization is ultimately going to be consumed on a 'portrait' oriented screen (ie taller than wide), which should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

```
count.State <- data.frame(table(inc$State))
head(count.State)
```

```
##   Var1 Freq
## 1   AK    2
## 2   AL   51
## 3   AR    9
## 4   AZ  100
## 5   CA  700
## 6   CO  133
```

Fast-growing Companies by State

```
ggplot(count.State, aes(x = reorder(Var1, Freq), y = Freq)) +
  geom_bar(stat = "identity", fill = "#f68060", alpha =.6, colour = "black") +
  coord_flip() +
  theme_bw(base_size = 8) +
  xlab("State") +
  ylab("No. of Inc. Magazine Fast-Growing Companies")
```



As expected, States with the largest populations have the most fast-growing companies. There are some discrepancies however. For example, comparing these figures to national population statistics, we note that Pennslyvania is ranked 5th in population, but only 10th here, while Massachusetts is 15th in population but 9th here. Surprisingly, Delaware is not ranked higher due to its favorable incorporation laws, suggesting that the state refers to location of headquarters, rather than place of incorporation.

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.

Here we have reduced the dataset to only New York-based companies.

```
order.Freq <- order(count.State$Freq, decreasing = TRUE)
sort.State <- count.State[order.Freq, ]
state3 <- as.character(sort.State[3,1])
inc.sub <- subset(inc, inc$State == state3)
inc.sub <- inc.sub[complete.cases(inc.sub),]
head(inc.sub)
```

```
##    Rank                     Name Growth_Rate Revenue
## 26   26              BeenVerified       84.43    13.7
## 30   30                  Sailthru       73.22     8.1
## 37   37              YellowHammer       67.40    18.0
## 38   38                 Conductor       67.02     7.1
## 48   48 Cinium Financial Services       53.65     5.9
## 70   70                  33Across       44.99    27.9
##                       Industry Employees      City State
## 26 Consumer Products & Services        17  New York    NY
## 30      Advertising & Marketing        79  New York    NY
## 37      Advertising & Marketing        27  New York    NY
## 38      Advertising & Marketing        89  New York    NY
## 48           Financial Services        32 Rock Hill    NY
## 70      Advertising & Marketing        75  New York    NY
```

There are 311 NY companies represented in the dataset.

```
nrow(inc.sub)
```

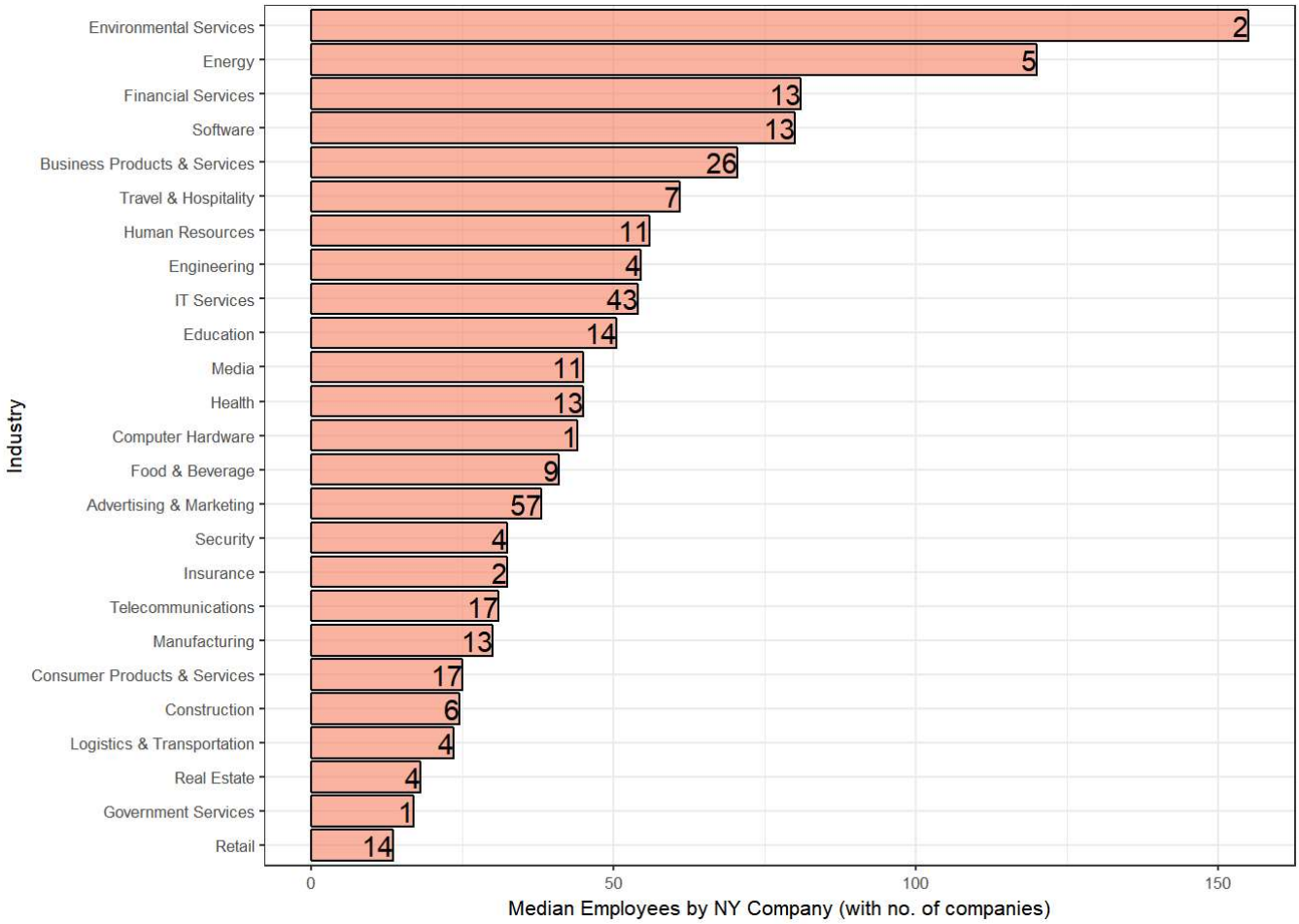
```
## [1] 311
```

```
state3.Industry <- data.frame(table(inc.sub$Industry)) # grouping NY companies by industry
state3.Employees <- aggregate(Employees ~ Industry, inc.sub, median) # taking the median number of
 employees by industry in NY
```

Median number of corporate employees by industry (in New York)

(The sample size is marked within the bar)

```
ggplot(state3.Employees, aes(x = reorder(Industry, Employees), y = Employees)) +
  geom_bar(stat = "identity", fill = "#f68060", alpha =.6, colour = "black") +
  geom_text(aes(label = state3.Industry$Freq), hjust = 1) +
  coord_flip() +
  theme_bw(base_size = 8) +
  xlab("Industry") +
  ylab("Median Employees by NY Company (with no. of companies)")
```



Plotting the range of employees by Industry in NY

We note that two companies have in excess of 10,000 employees, clearly outliers.

```
head(inc.sub[order(-inc.sub$Employees),], 6)
```

```
##      Rank                       Name Growth_Rate Revenue
## 4577 4577 Sutherland Global Services        0.48   597.6
## 4936 4936                       Coty        0.36  4600.0
## 4716 4716             Westcon Group         0.44  3800.0
## 3899 3899  Denihan Hospitality Group        0.71   280.8
## 4363 4363               TransPerfect        0.55   341.3
## 1498 1499       Sterling Infosystems        2.66   214.9
##                          Industry Employees      City State
## 4577 Business Products & Services     32000 Pittsford    NY
## 4936 Consumer Products & Services     10000  New York    NY
## 4716                  IT Services      3000 Tarrytown    NY
## 3899         Travel & Hospitality      2280  New York    NY
## 4363 Business Products & Services      2218  New York    NY
## 1498              Human Resources      2081  New York    NY
```

We first define outliers that those cases with values more than 1.5 times higher or lower than the interquartile range. When we do so, we are left with 253 companies, with 58 classified as “outliers” according to this definition.

```
outlier_limit <- median(inc.sub$Employees) + 1.5 * (fivenum(inc.sub$Employees)[4] - fivenum(inc.su
b$Employees)[2])
inc.sub2 <- subset(inc.sub, Employees < outlier_limit)
nrow(inc.sub2)
```

```
## [1] 253
```

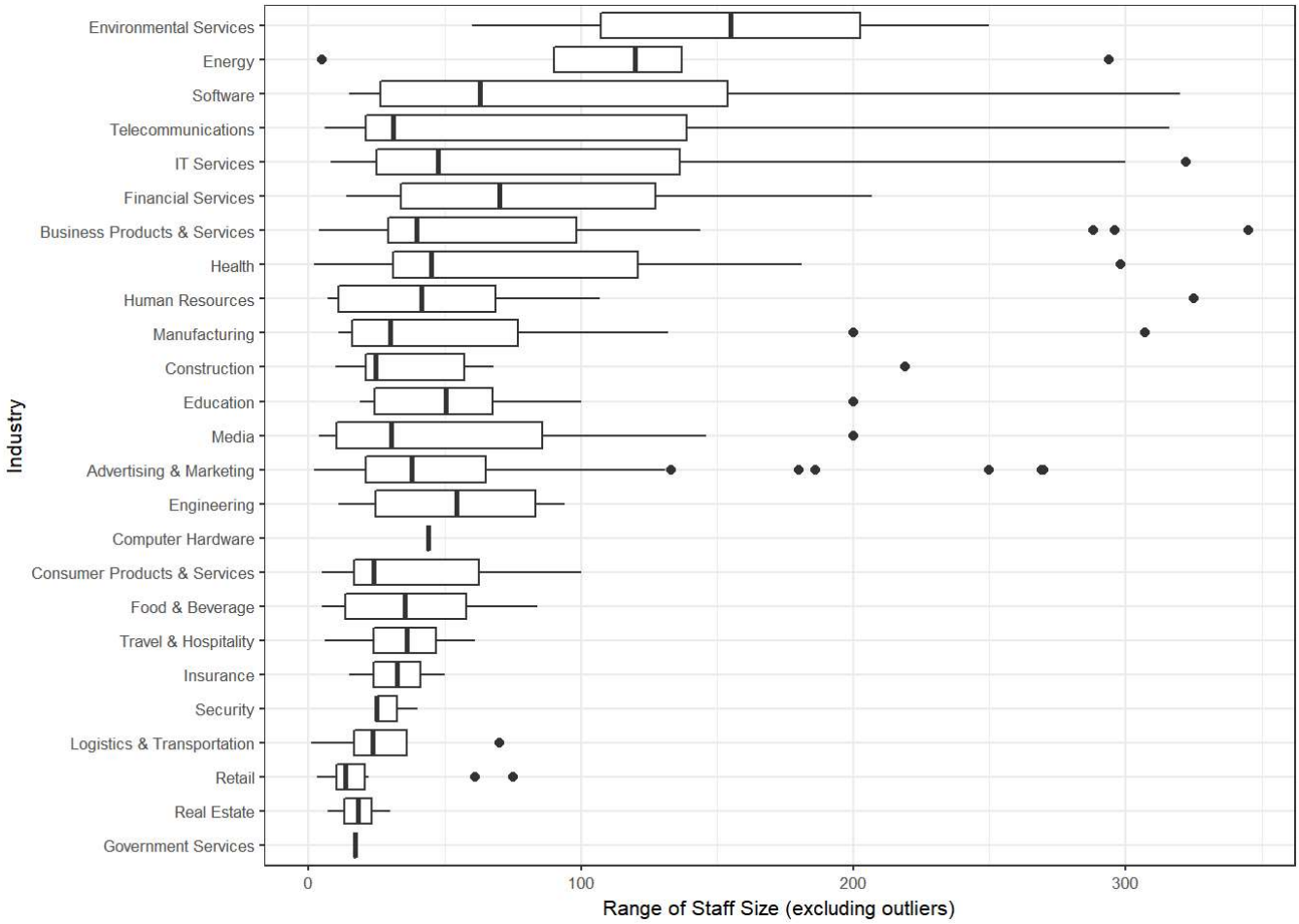
The IQR definition yields too many outliers, so we redefine it as companies with over 350 employees, thereby removing on 25 companies, or about 8%.

```
outlier_limit2 <- 350
inc.sub2 <- subset(inc.sub, Employees < outlier_limit2)
nrow(inc.sub2)
```

```
## [1] 286
```

Creating boxplots, we observe the ranges of employees by segment. IT services seem to have the widest spread.

```
ggplot(inc.sub2, aes(x = reorder(Industry, Employees), y = Employees)) +
  geom_boxplot() +
  coord_flip() +
  theme_bw(base_size = 8) +
  xlab("Industry") +
  ylab("Range of Staff Size (excluding outliers)")
```



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.

We note that Computer Hardware was the highest revenue per employee (even after CDW was excised from the dataset) with over $600,000 per worker. Energy ranks second and construction third. Somewhat ironically, Human Resources is the least value-generating for labor.

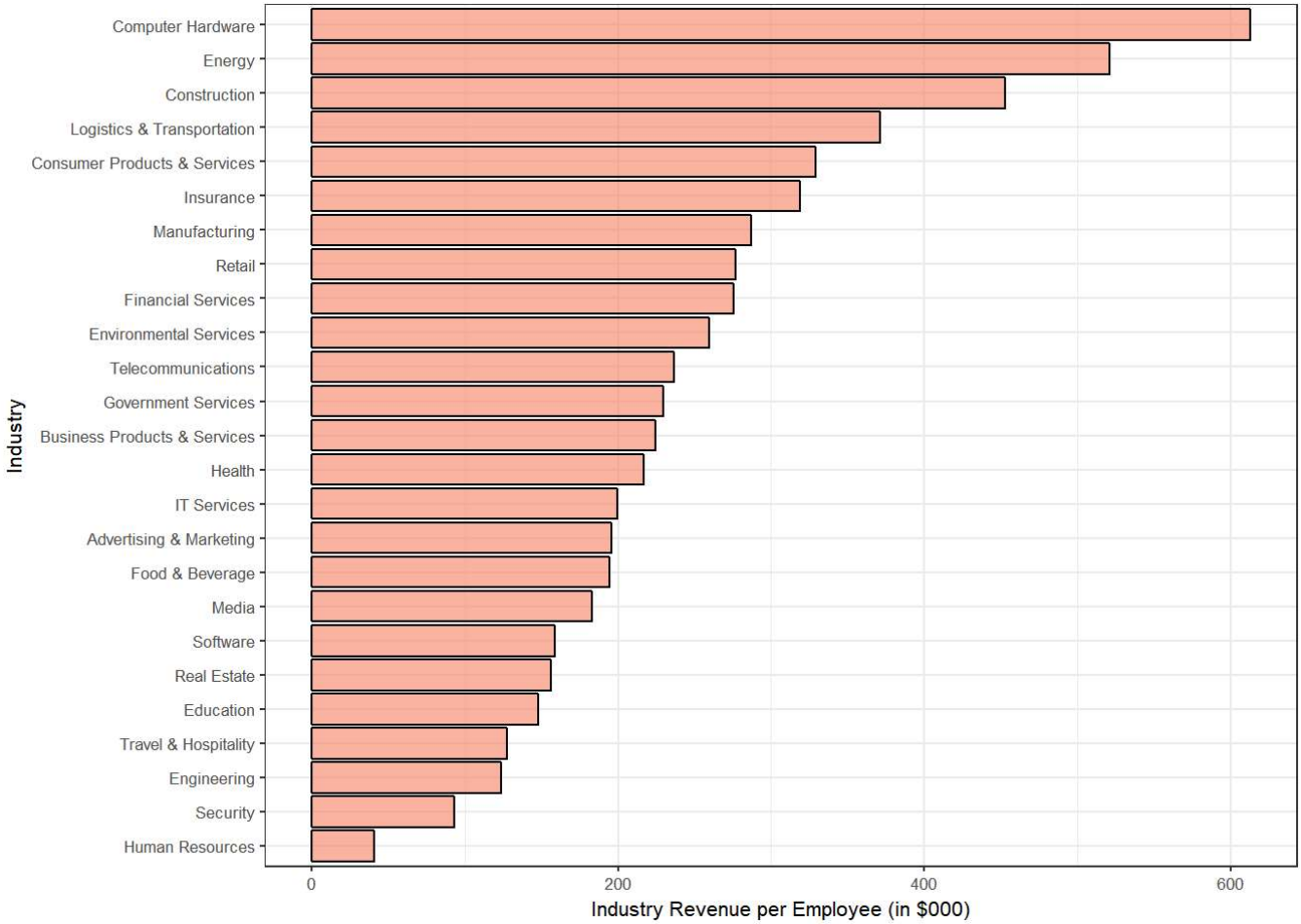
```
grp <- group_by(inc, Industry)
ind.rev <- summarise(grp, rev.per.employee=sum(Revenue)/sum(Employees))

order.Freq <- order(ind.rev$rev.per.employee, decreasing = TRUE)
sort.rev <- ind.rev[order.Freq, ]
head(sort.rev)
```

```
## # A tibble: 6 x 2
##   Industry                      rev.per.employee
##   <fct>                                    <dbl>
## 1 Computer Hardware                        0.613
## 2 Energy                                   0.521
## 3 Construction                             0.453
## 4 Logistics & Transportation               0.371
## 5 Consumer Products & Services             0.329
## 6 Insurance                                0.319
```

Revenue per Employee, Industry-wide, by Industry Categories

```
ggplot(sort.rev, aes(x = reorder(Industry, rev.per.employee), y = rev.per.employee * 1000)) +
  geom_bar(stat = "identity", fill = "#f68060", alpha =.6, colour = "black") +
  coord_flip() +
  theme_bw(base_size = 8) +
  xlab("Industry") +
  ylab("Industry Revenue per Employee (in $000)")
```



At the individual company level, we note the range of Revenue per Employee using a boxplot. This is after first removing outliers based on IQR (occuring especially at energy companies)

```
#Removing outliers based on IQR * 1.5
inc$rev.employee <- inc$Revenue / inc$Employees
outlier_limit2 <- median(inc$rev.employee) + 1.5 * (fivenum(inc$rev.employee)[4] - fivenum(inc$re
v.employee)[2])
inc.sub3 <- subset(inc, inc$rev.employee < outlier_limit2)
```

Range of Employee Revenue Generation (at Individual Companies)

```
ggplot(inc.sub3, aes(x = reorder(Industry, rev.employee), y = rev.employee * 1000)) +
  geom_boxplot() +
  coord_flip() +
  theme_bw(base_size = 8) +
  xlab("Industry") +
  ylab("Range of Revenue per Employee (in $000, excluding outliers)")
```