DATA608 - Module 1 R Notebook, Author - Peter Gatica

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.1     v dplyr   1.0.5
## 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(tidytext)
library(ggplot2)
library(gcookbook)
library(dplyr)
library(knitr)
```

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:

Taking a look at the summary of the data and I immediately notice the maximum number of employees of a company and the minmum number. This is a very large gap between the maximum and the minimum and I cannot help but wonder what affect that may have on the average number of employees per company.

Question 1

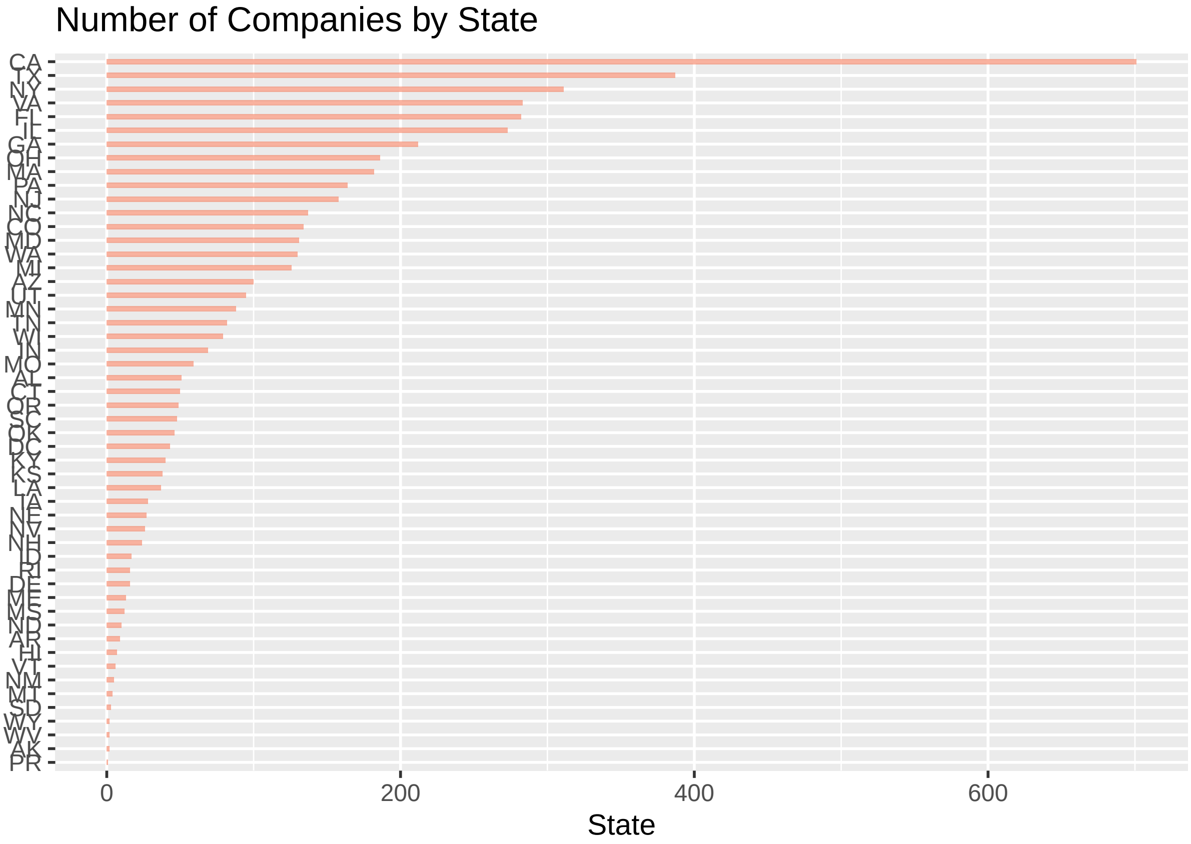
Create a graph that shows the distribution of companies in the dataset by State (ie how many are in each state). There are a lot of States, so consider which axis you should use. This visualization is ultimately going to be consumed on a 'portrait' oriented screen (ie taller than wide), which should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

```
# Count the number of companies by state
inc_agg_by_state <- aggregate(inc$Name, by=list(inc$State), FUN=length)
# rename columns
names(inc_agg_by_state) <- c("State", "Count")
# inc_agg_by_state
```

```
head(inc_agg_by_state[with(inc_agg_by_state, order(-Count)),],5) # List in order by the total number of
```

```
##    State Count
## 5     CA   701
## 45    TX   387
## 35    NY   311
## 47    VA   283
## 10    FL   282
```

```
inc_agg_by_state %>%
    ggplot(aes(fct_reorder(`State`,`Count`), `Count`))+
      geom_bar(stat="identity", fill="#f68060", alpha=.6, width=.4) +
      coord_flip() +
      xlab("") +
      ylab("State")+
      ggtitle("Number of Companies by State")
```

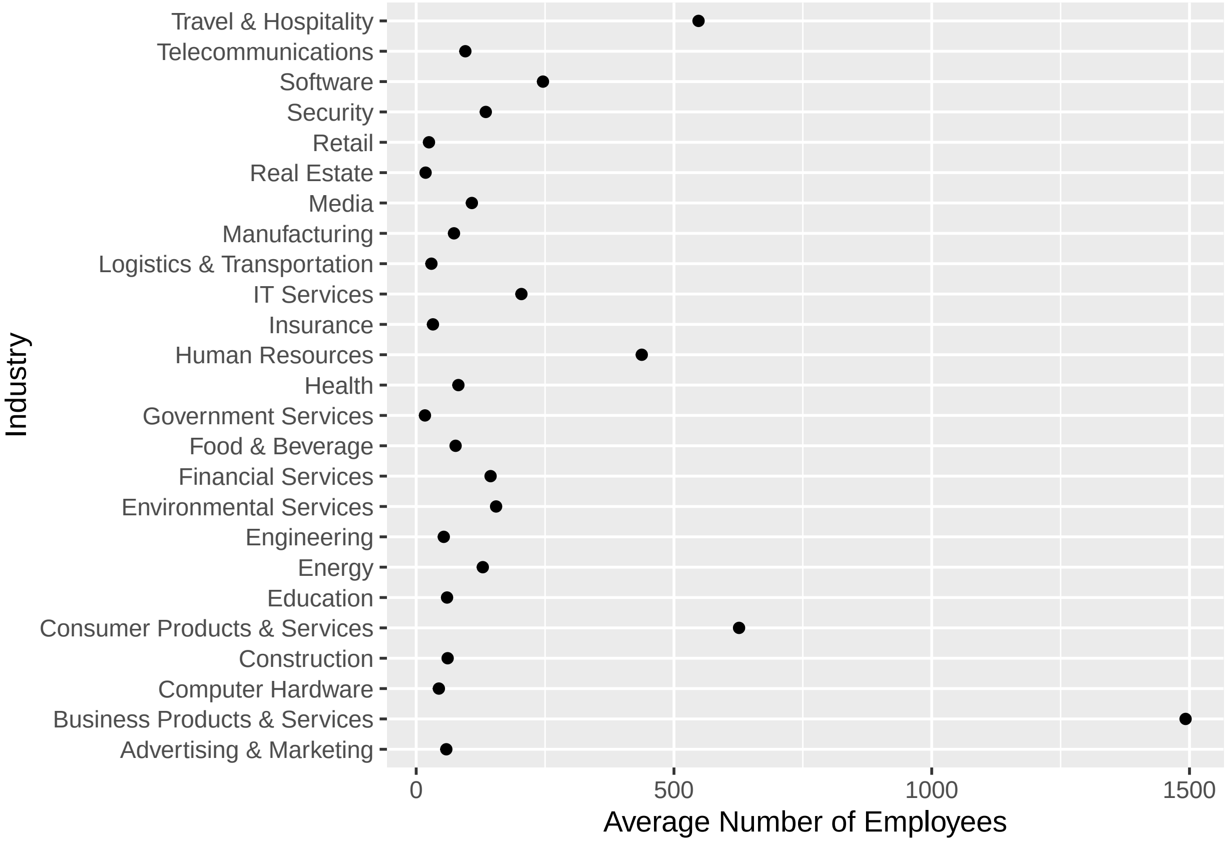


Quesiton 2

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.

```
# Answer Question 2 here
# Get the cases with full data in the state with the 3rd most companies
inc_comp_cases_ny <- inc[complete.cases(inc), ] %>%
  filter(State == 'NY')

# Get the average/mean of employment by Industry
inc_ny <- aggregate(Employees ~ Industry, data=inc_comp_cases_ny, FUN=mean)
names(inc_ny) <- c("Industry", "AvgNumEmps")
```

```
# Create a scatter plot to show the average and/or median employment by industry for companies in this s
ggplot(inc_ny, aes(x = AvgNumEmps, y = Industry)) +
  xlab("Average Number of Employees") +
  geom_point()
```



Notice the big difference of the average number of employees by industry for the state of New York. I might call this an outlier that may have an inpact on the overall average number of employees for the state of New York. Looking at the actual averages of each industry, one will notice the that the average is almost double between the first and the second highest average.

```
head(inc_ny[with(inc_ny, order(-AvgNumEmps)),],5) # List in order by the total number of companies in th
```

```
##                        Industry AvgNumEmps
## 2  Business Products & Services  1492.4615
## 5  Consumer Products & Services   626.2941
## 25         Travel & Hospitality   547.7143
## 14              Human Resources   437.5455
## 23                     Software   245.9231
```

Also notice that the number of employees for the largest company is over 3 times the number of the next larger company in the New York. This will definitely affect the average number of employees which can in turn can skew other averages. I will exclude the Sutherland Global Services company to see what kind of affect it will have and will consider it an outlier.

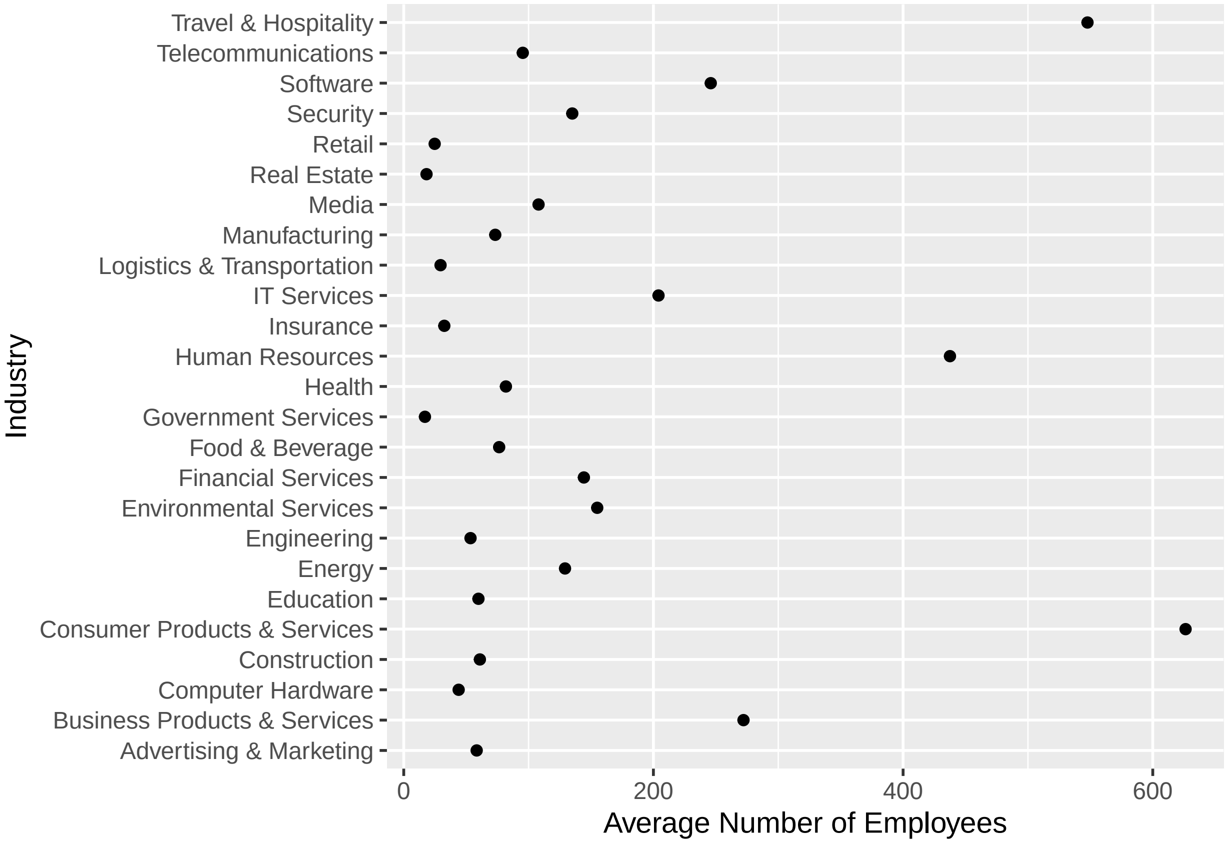
```
head(inc_comp_cases_ny[with(inc_comp_cases_ny, order(-Employees)),c("Rank","Name","Industry","Employees
```

```
##     Rank                       Name                     Industry Employees
## 274 4577 Sutherland Global Services Business Products & Services     32000
## 307 4936                       Coty Consumer Products & Services     10000
## 287 4716               Westcon Group                  IT Services      3000
## 228 3899  Denihan Hospitality Group         Travel & Hospitality      2280
## 254 4363                TransPerfect Business Products & Services      2218
```

```
# Filter the Sutherland company
inc_ny_emp_no_out <- inc_comp_cases_ny %>%
  filter(Employees < 10001)

# Recalculate the average/mean of employment by Industry
inc_ny_emp_no_out <- aggregate(Employees ~ Industry, data=inc_ny_emp_no_out, FUN=mean)
names(inc_ny_emp_no_out) <- c("Industry", "AvgNumEmps")
inc_ny_emp_no_out[with(inc_ny_emp_no_out, order(-AvgNumEmps)),] # List in order by the total number of
```

```
##                        Industry AvgNumEmps
## 5  Consumer Products & Services  626.29412
## 25         Travel & Hospitality  547.71429
## 14              Human Resources  437.54545
## 2  Business Products & Services  272.16000
## 23                     Software  245.92308
## 16                  IT Services  204.09302
## 9        Environmental Services  155.00000
## 10           Financial Services  144.30769
## 22                     Security  135.00000
## 7                        Energy  129.20000
## 19                        Media  108.00000
## 24           Telecommunications   95.35294
## 13                       Health   81.84615
## 11              Food & Beverage   76.44444
## 18                Manufacturing   73.30769
## 4                  Construction   61.00000
## 6                     Education   59.85714
## 1       Advertising & Marketing   58.43860
## 8                   Engineering   53.50000
## 3             Computer Hardware   44.00000
## 15                    Insurance   32.50000
## 17   Logistics & Transportation   29.50000
## 21                       Retail   24.78571
## 20                  Real Estate   18.25000
## 12          Government Services   17.00000
```

```
ggplot(inc_ny_emp_no_out, aes(x = AvgNumEmps, y = Industry)) +
    xlab("Average Number of Employees") +
  geom_point()
```



```
summary(inc_ny_emp_no_out)
```

```
##    Industry           AvgNumEmps
##  Length:25          Min.   : 17.00
##  Class :character   1st Qu.: 53.50
##  Mode  :character   Median : 81.85
##                     Mean   :149.24
##                     3rd Qu.:155.00
##                     Max.   :626.29
```

As you can see by the scatterplot and the summary that excluding the Sutherland company shows a more accurate measure of the average number of employees by company across all industires for the state of New York at 149.24 number of employees. I believe that considering Sutherland company as an outlier is the right thing to do.

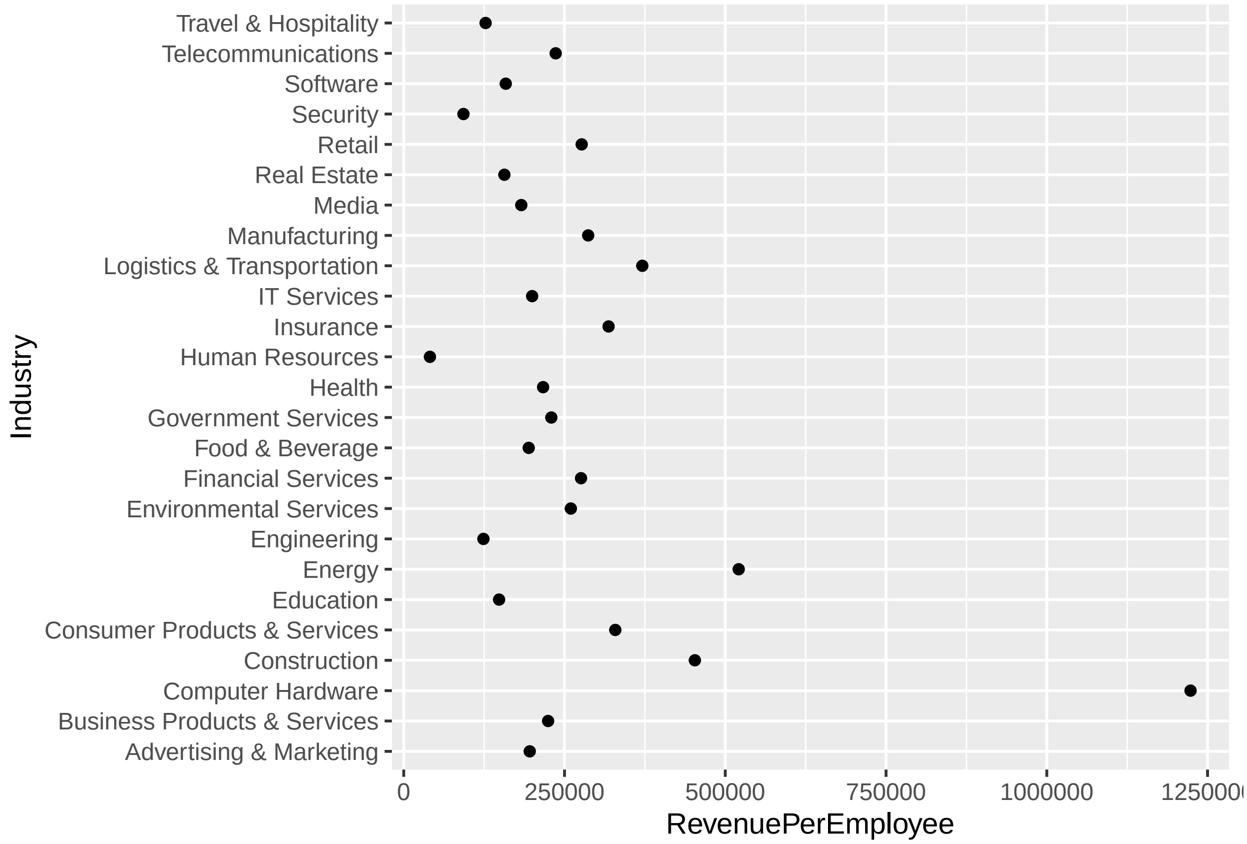
Question 3

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.

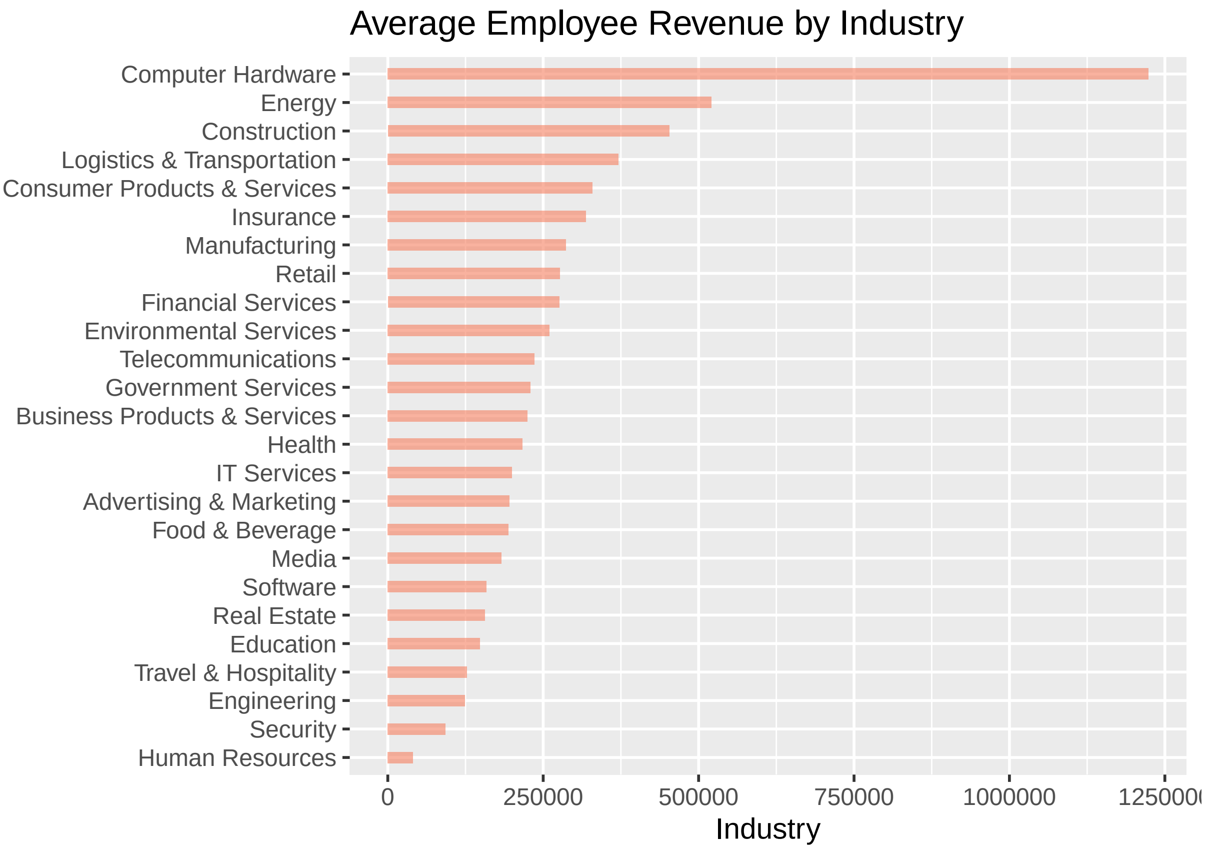
```
# Answer Question 3 here
# Get the average/mean of employment by Industry
inc_rev_emps <- aggregate(cbind(Revenue, Employees) ~ Industry, data=inc, FUN=sum)

inc_rev_by_emp <- group_by(inc_rev_emps, Industry, Revenue / Employees)

# rename columns
names(inc_rev_by_emp) <- c("Industry", "Revenue", "Employees" , "RevenuePerEmployee")
inc_rev_by_emp
```

```
## # A tibble: 25 x 4
## # Groups:   Industry, RevenuePerEmployee [25]
##    Industry                         Revenue Employees RevenuePerEmployee
##    <chr>                              <dbl>     <dbl>              <dbl>
##  1 Advertising & Marketing       7785000000     39731            195943.
##  2 Business Products & Services 26345900000    117357            224494.
##  3 Computer Hardware            11885700000      9714           1223564.
##  4 Construction                 13174300000     29099            452741.
##  5 Consumer Products & Services 14956400000     45464            328972.
##  6 Education                     1139300000      7685            148250.
##  7 Energy                       13771600000     26437            520921.
##  8 Engineering                   2532500000     20435            123930.
##  9 Environmental Services        2638800000     10155            259852.
## 10 Financial Services           13150900000     47693            275741.
## # ... with 15 more rows
```

```
ggplot(inc_rev_by_emp, aes(x = RevenuePerEmployee, y = Industry)) +
  geom_point()
```



```
inc_rev_by_emp %>%
    ggplot(aes(fct_reorder(`Industry`,`RevenuePerEmployee`), `RevenuePerEmployee`))+
      geom_bar(stat="identity", fill="#f68060", alpha=.6, width=.4) +
      coord_flip() +
      xlab("") +
      ylab("Industry")+
      ggtitle("Average Employee Revenue by Industry")
```



In conclusion, as an investor I would most likely invest my money in the Computer Hardware industry if I was making that decision based on the average revenue generated per employee.