R Notebook

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:

```
library(psych)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'

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

```
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
##  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:

```
# Insert your code here, create more chunks as necessary

#Check structure of data
str(inc)
```

```
## 'data.frame':    5001 obs. of  8 variables:
##  $ Rank       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Name       : Factor w/ 5001 levels "(Add)ventures",..: 1770 1633 4423 690 1198 2839 4733 1468 186
##  $ Growth_Rate: num  421 248 245 233 213 ...
##  $ Revenue    : num  1.18e+08 4.96e+07 2.55e+07 1.90e+09 8.70e+07 ...
##  $ Industry   : Factor w/ 25 levels "Advertising & Marketing",..: 5 12 13 7 1 20 10 1 5 21 ...
##  $ Employees  : int  104 51 132 50 220 63 27 75 97 15 ...
##  $ City       : Factor w/ 1519 levels "Acton","Addison",..: 391 365 635 2 139 66 912 1179 131 1418 .
##  $ State      : Factor w/ 52 levels "AK","AL","AR",..: 5 47 10 45 20 45 44 5 46 41 ...
```

```
#look at the bottom of the data
tail(inc)
```

```
##      Rank               Name Growth_Rate  Revenue                      Industry
## 4996 4996             cSubs        0.34 1.34e+07 Business Products & Services
## 4997 4997         Dot Foods        0.34 4.50e+09              Food & Beverage
## 4998 4998 Lethal Performance        0.34 6.80e+06                      Retail
## 4999 4999   ArcaTech Systems        0.34 3.26e+07           Financial Services
## 5000 5000               INE        0.34 6.80e+06                IT Services
## 5001 5000              ALL4        0.34 4.70e+06       Environmental Services
##      Employees         City State
## 4996        19     Montvale    NJ
## 4997      3919 Mt. Sterling    IL
## 4998         8   Wellington    FL
## 4999        63       Mebane    NC
## 5000        35     Bellevue    WA
## 5001        34    Kimberton    PA
```

```
#Check for missing variables across all columns
colSums(is.na(inc))
```

```
##        Rank        Name Growth_Rate     Revenue    Industry   Employees
##           0           0           0           0           0          12
##        City       State
##           0           0
```

```
#Employees column has missing values
```

```
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
```

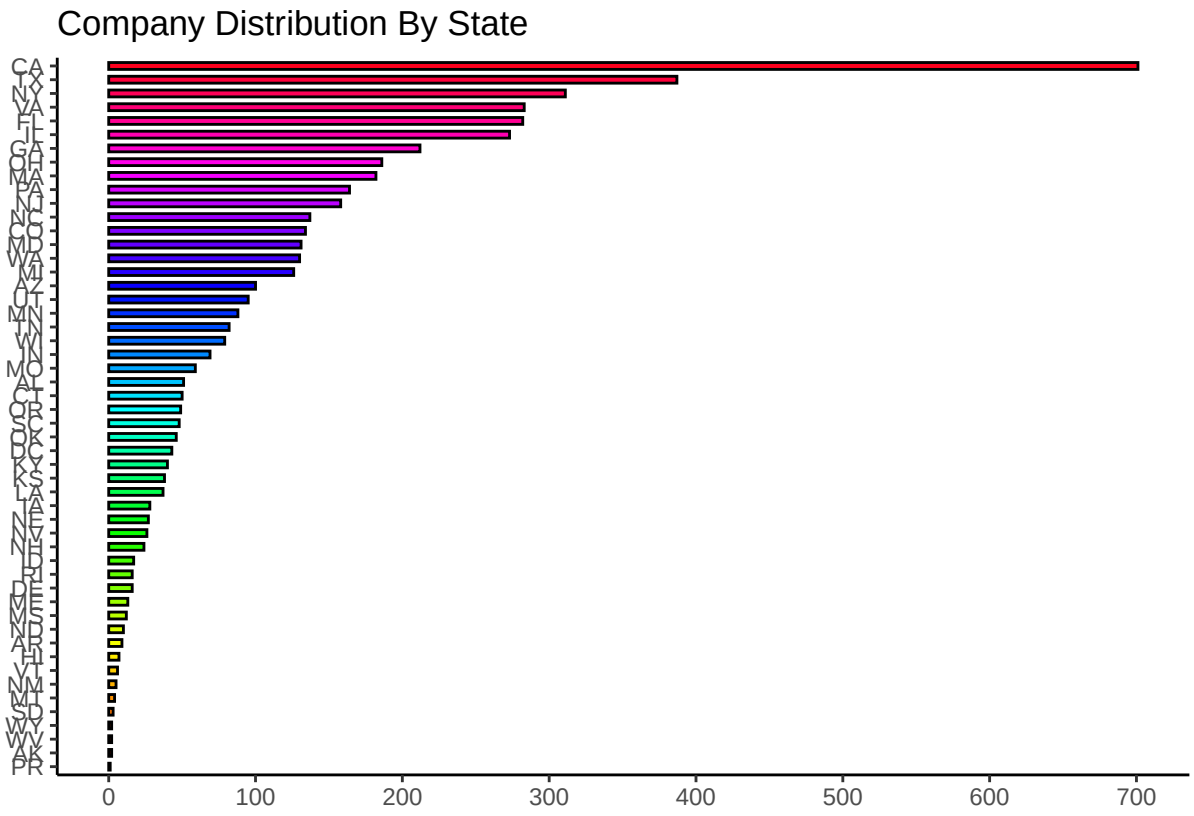
```
#count for distinct values of state
#Top 36 states have 100 or more companies
```

```
count_state <- dplyr::count(inc,State)

#count for distinct values of City
#count_city <- dplyr::count(inc, City)
#Decided against using distinct count for cities as 1519 rows were calcultated, not a useful summarizat

#count for distinct values of industry
count_industry <- dplyr::count(inc, Industry)
```

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

desc_cs <- count_state %>% arrange(desc(n))


#the multiple colors helps distinguish the many states presented in the graph
ggplot(desc_cs, aes(x=reorder(State, n),y=n, color=State)) +
  geom_bar(stat='identity', width = 0.5, color = 'black', fill=rainbow(52)) +
  coord_flip() +
  labs(title = 'Company Distribution By State', x='', y='')+
  scale_y_continuous(breaks = seq(0, 700, 100))+
  theme_classic()
```




```
#source: https://stackoverflow.com/questions/29587881/increase-plot-size-width-in-ggplot2
ggsave(file="Distribution By State.png", width=10, height=5, dpi=300)
```

Quesiton 2

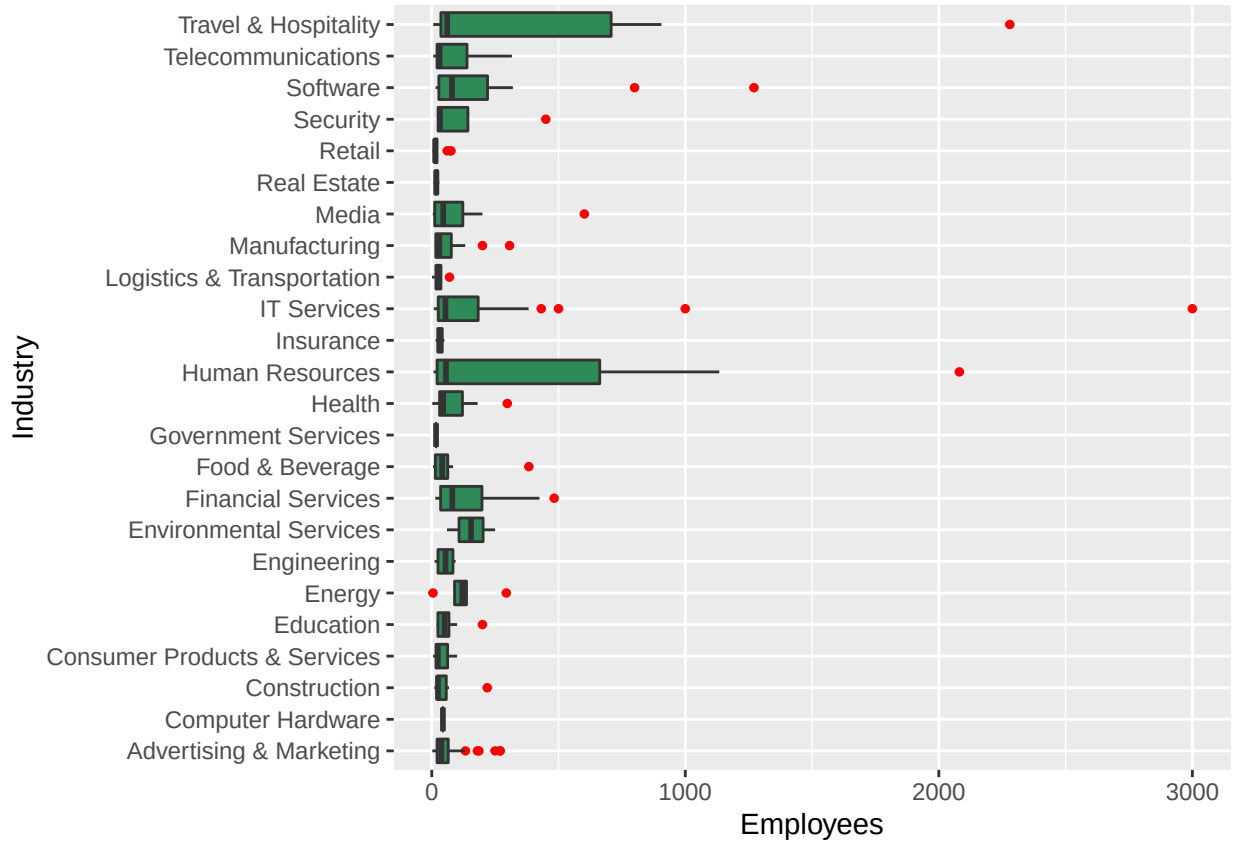
Lets dig in on the state with the 3rd most companies in the data set. Imagine you work for the state and are interested in how many people are employed by companies in different industries. Create a plot that shows the average and/or median employment by industry for companies in this state (only use 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.

Based on the graphic and data from above, the state with the 3rd most companies is NY. So we will be digging into the employment of different industries within the state of NY.

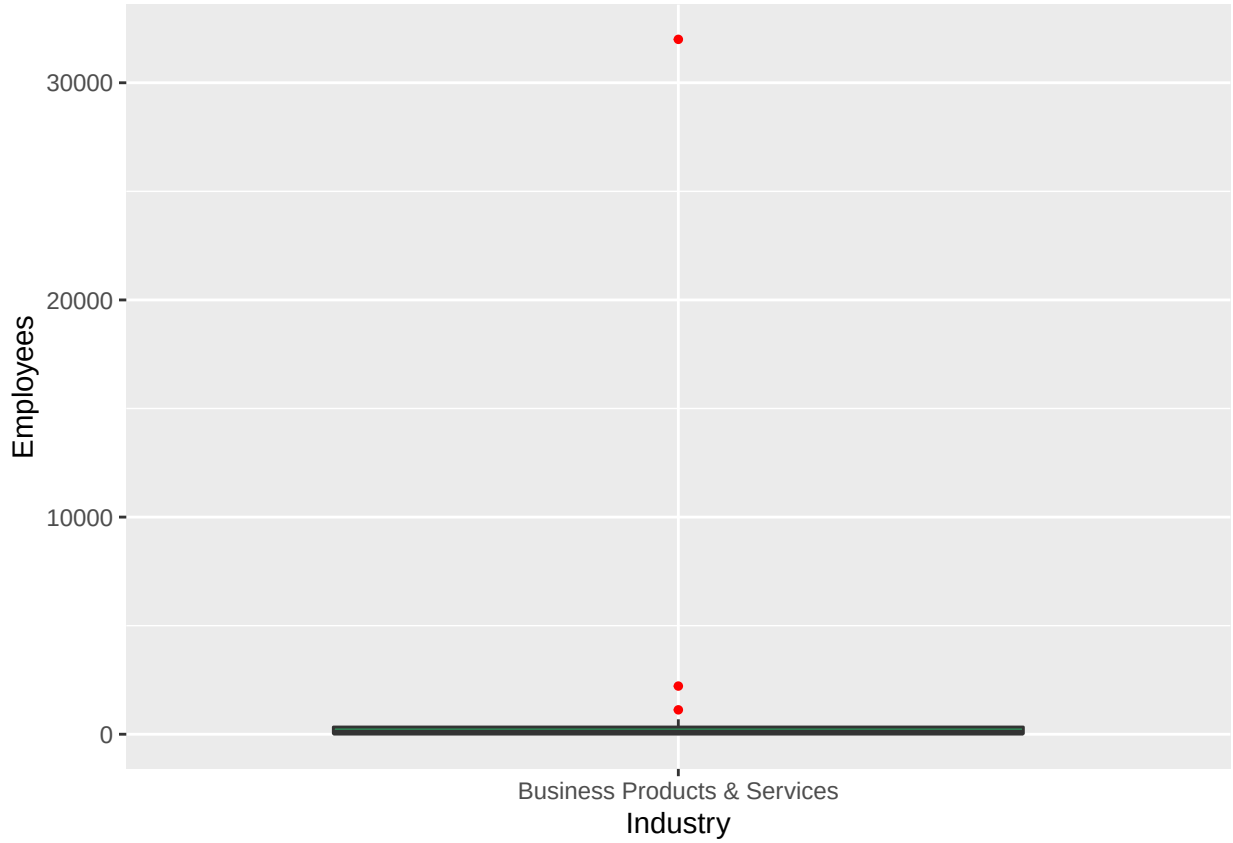
```
# Answer Question 2 here

inc_complete <- inc[complete.cases(inc),]

ny_industry <- inc_complete %>%filter(State == 'NY')

#Seperated Business products and services, they had an outsized number that distorted the rest of the v

ny_industry_business <- ny_industry %>% filter(Industry == 'Business Products & Services')

nyi_no_business <- ny_industry %>% filter(Industry != 'Business Products & Services')
```

```r
#Going to utilize boxplots to illustrate the range/average/median employment by industry
# source: https://www.quora.com/What-is-the-best-graph-to-illustrate-ranges-in-a-data-series?share=1

ggplot(nyi_no_business, aes(x = Industry, y=Employees)) +
    coord_flip() +
    geom_boxplot(fill="seagreen", outlier.color = "red", outlier.size = 1) +
    ylim(0,3000)
```

```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```



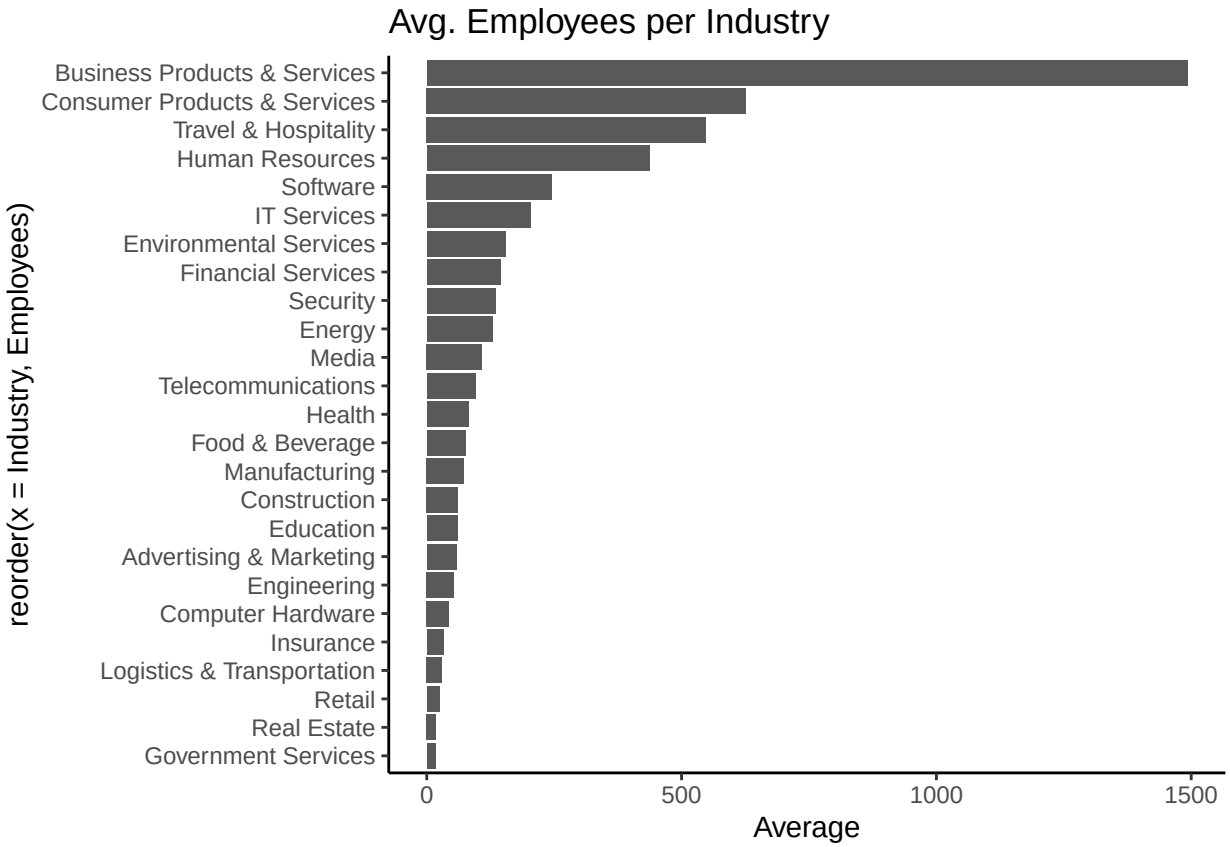

```r
#The outlier in for the Business Products and Services created a very flat boxplot, I played with minim
ggplot(ny_industry_business, aes(x = Industry, y=Employees)) +
    geom_boxplot(fill="seagreen", outlier.color = "red", outlier.size = 1)
```




```
ggplot(ny_industry, aes(reorder(x=Industry, Employees), y = Employees)) +
    stat_summary(fun = "mean", geom = "bar") +
    coord_flip() +
    labs(title = "Avg. Employees per Industry", y = "Average")+
    theme_classic()
```

```
## Warning: Ignoring unknown parameters: fun
```

```
## No summary function supplied, defaulting to `mean_se()
```

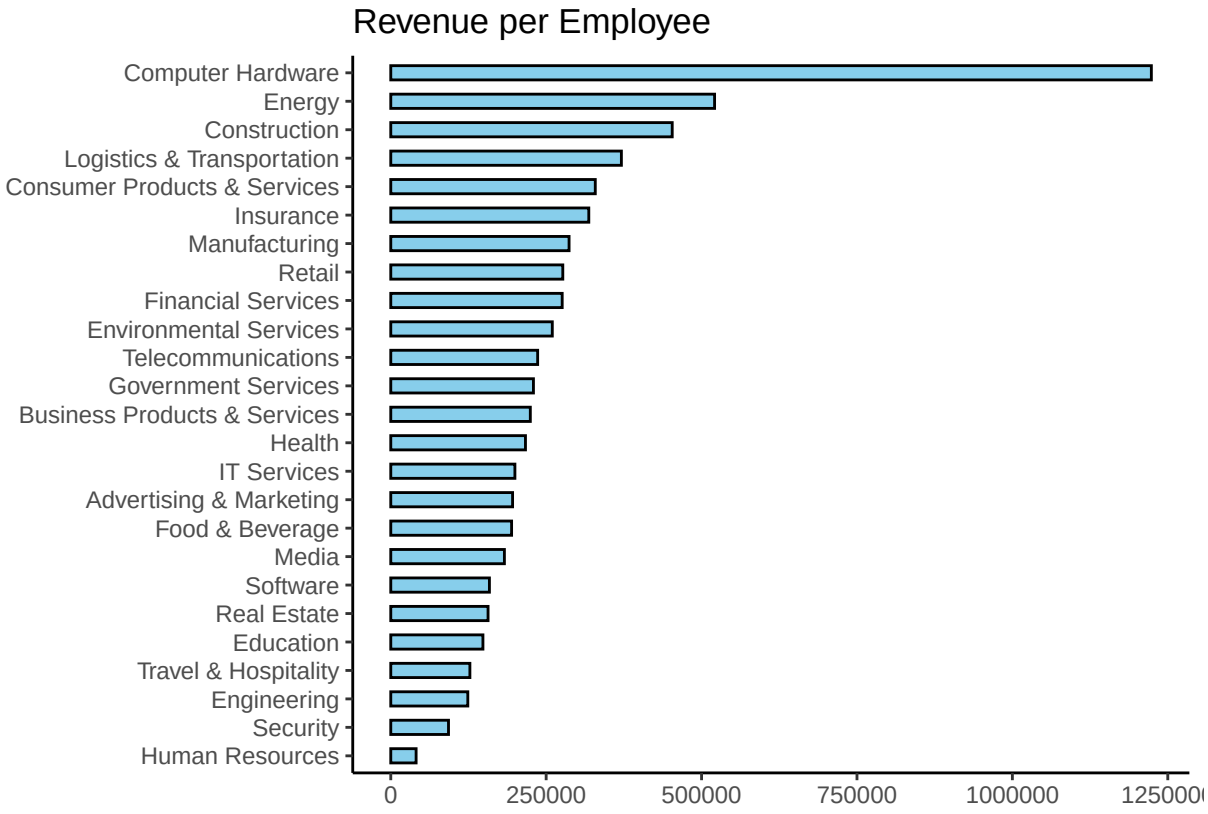


Question 3

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.

The revenue per employee here is shown for the national dataset.

```
# Answer Question 3 here


#Let's calculate a new field, revenue per employee

rev_per_employee <- inc_complete %>% group_by(Industry) %>% summarise(revenue=sum(Revenue), employees=su


ggplot(rev_per_employee, aes(x=reorder(Industry, revenue_per_employee),y=revenue_per_employee)) +
  geom_bar(stat='identity', width = 0.5, color = 'black', fill='skyblue') +
  coord_flip() +
  labs(title = 'Revenue per Employee', x='', y='')+
  theme_classic()
```



Sources: https://www.tutorialgateway.org/r-ggplot2-boxplot/

https://www.quora.com/What-is-the-best-graph-to-illustrate-ranges-in-a-data-series?share=1

https://stackoverflow.com/questions/29587881/increase-plot-size-width-in-ggplot2

https://stackoverflow.com/questions/11857935/plotting-the-average-values-for-each-level-in-ggplot2#11858054