DATA 608 Module 1

Md Jalal Uddin

February 13, 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:

```
library(ggplot2)
library(tidyr)
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
```

```
inc <- read.csv("https://raw.githubusercontent.com/charleyferrari/CUNY_DATA_608/master/module1/Data/inc5
```

And lets preview this data:

```
head(inc)
```

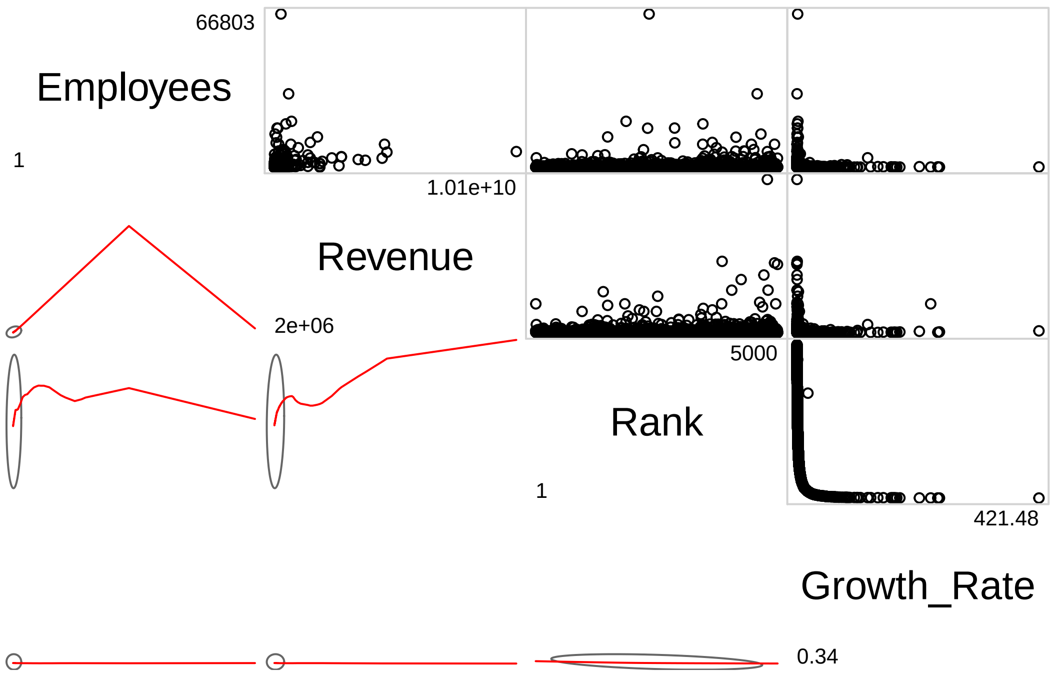
```
##   Rank                         Name Growth_Rate   Revenue
## 1    1                         Fuhu      421.48 1.179e+08
## 2    2        FederalConference.com      248.31 4.960e+07
## 3    3                The HCI Group      245.45 2.550e+07
## 4    4                      Bridger      233.08 1.900e+09
## 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:

```r
library(corrgram)
corrgram(inc, order=TRUE, lower.panel=panel.ellipse,
  upper.panel=panel.pts, text.panel=panel.txt,
  diag.panel=panel.minmax)
```



```
summary(lm(Employees ~ Revenue, data = inc))
```

```
##
## Call:
## lm(formula = Employees ~ Revenue, data = inc)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9130   -148   -128    -74  66211
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.574e+02  1.877e+01   8.384   <2e-16 ***
## Revenue     1.562e-06  7.643e-08  20.432   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1300 on 4987 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.07725,    Adjusted R-squared:  0.07706
## F-statistic: 417.5 on 1 and 4987 DF,  p-value: < 2.2e-16
```

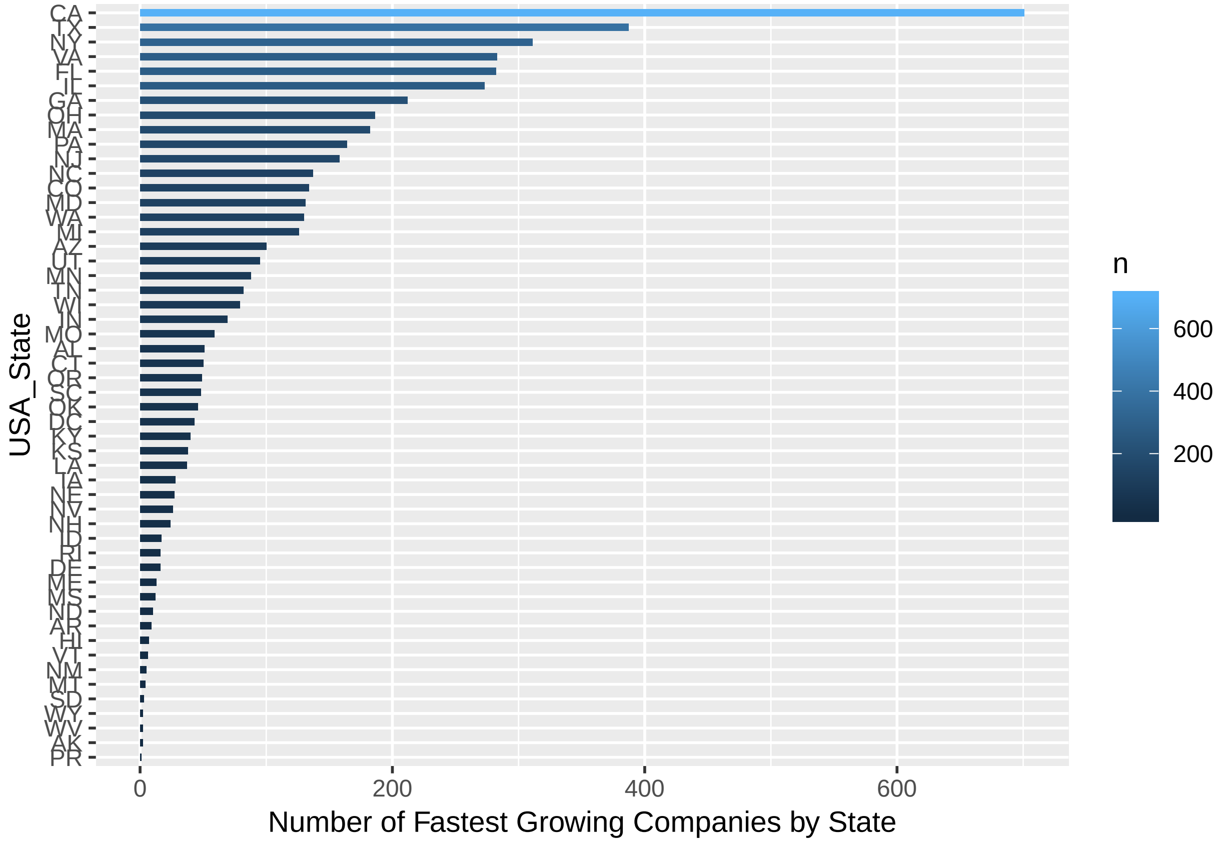
Question 1 Create a graph that shows the distribution of companies in the dataset by State (ie how many are in each state). There are a lot of States, so consider which axis you should use. This visualization is ultimately going to be consumed on a 'portrait' oriented screen (ie taller than wide), which should further guide your layout choices.

Answer:

```
state = inc %>%
  group_by(State) %>%
  count(State)%>%
  arrange(desc(n))
head(state)
```

```
## # A tibble: 6 x 2
## # Groups:   State [6]
##   State     n
##   <fct> <int>
## 1 CA      701
## 2 TX      387
## 3 NY      311
## 4 VA      283
## 5 FL      282
## 6 IL      273
```

```
f <- ggplot(state, aes(x=reorder(State, n), y=n, fill=n))
f + geom_bar(stat="identity", width=0.5, position = position_dodge(width=1.5)) + coord_flip() + labs(x
```

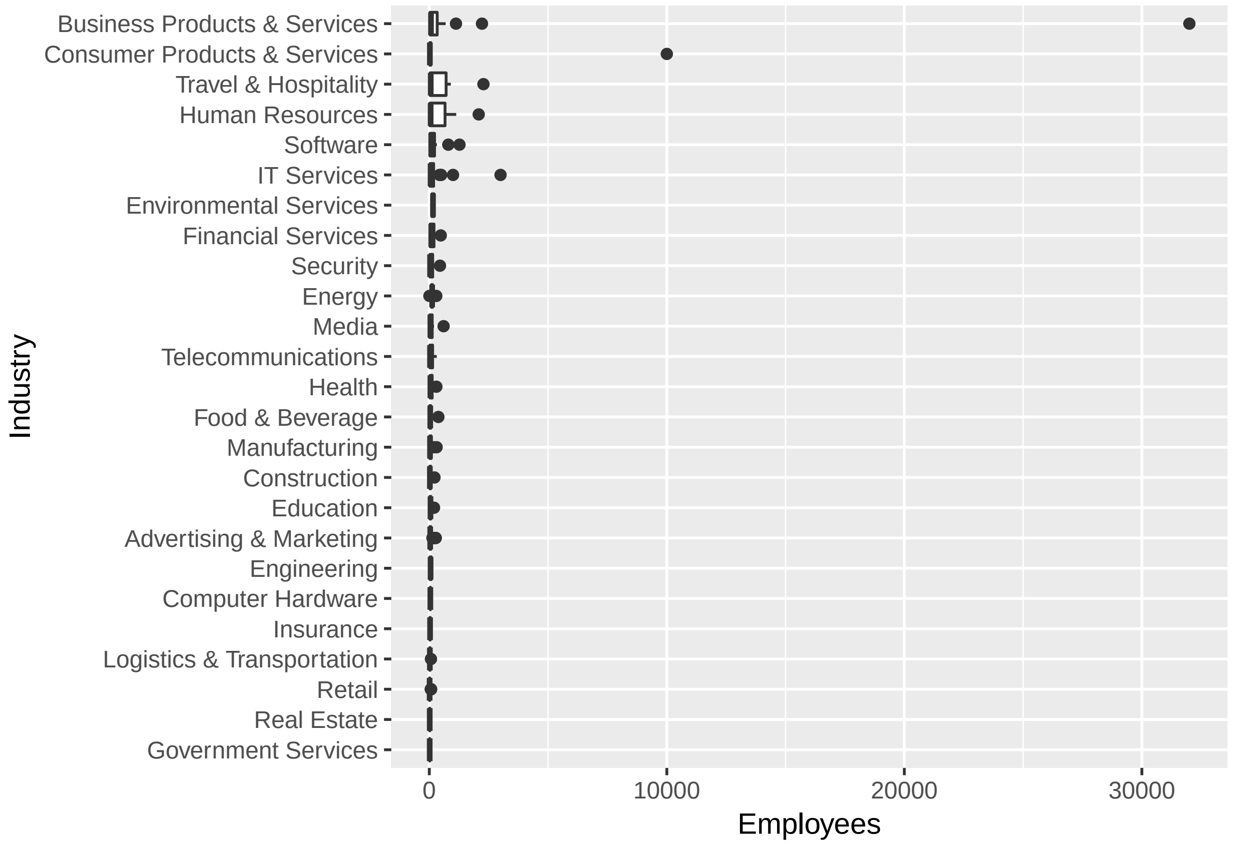


Question 2: Lets dig in on the state with the 3rd most companies in the data set. Imagine you work for the state and are interested in how many people are employed by companies in different industries. Create a plot that shows the average and/or median employment by industry for companies in this state (only use

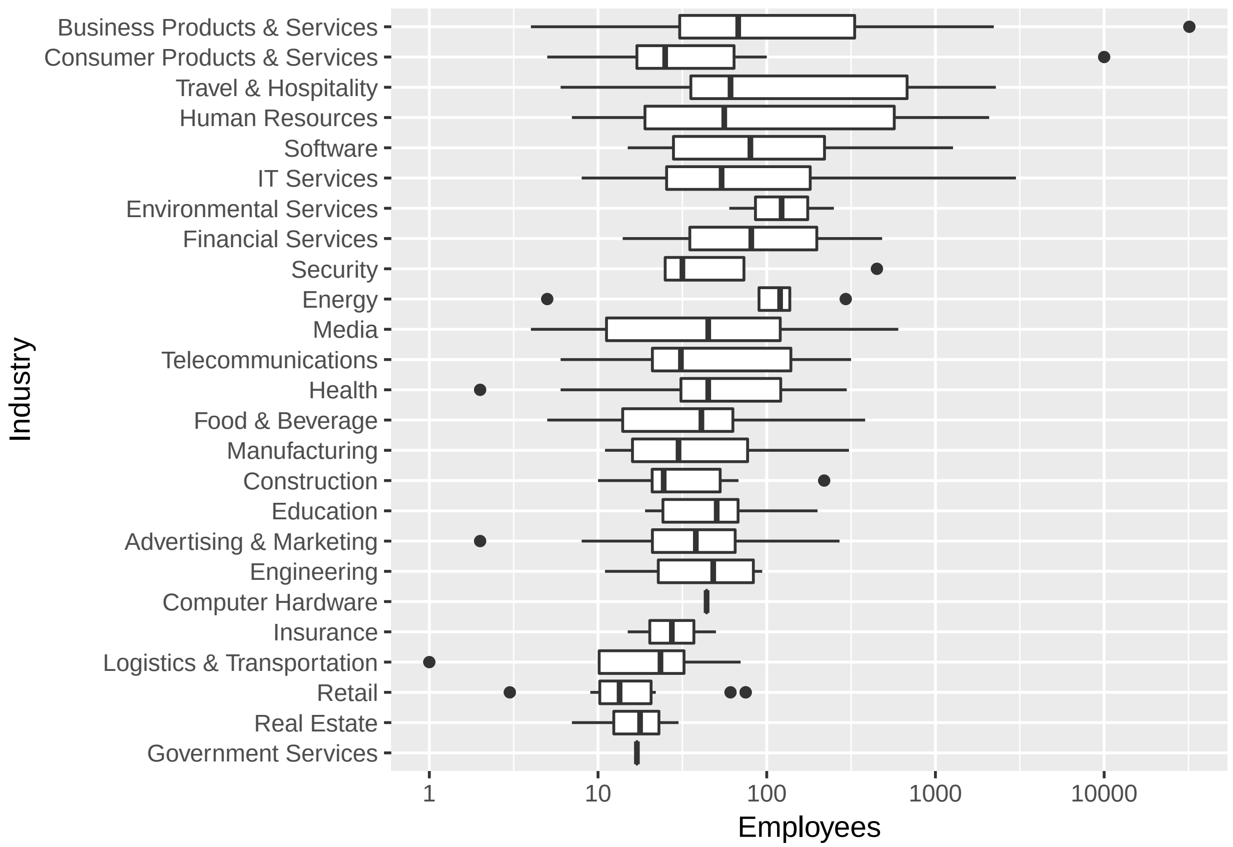
cases with full data, use R's complete.cases() function.) In addition to this, your graph should show how variable the ranges are, and you should deal with outliers.

Answer:

```
inc <- inc[complete.cases(inc),]
New_York = inc %>%
  filter(State == "NY")

Graph <- ggplot(New_York, aes(reorder(Industry,Employees,mean), Employees))
Graph <- Graph + geom_boxplot() + coord_flip() + labs(x = "Industry", y = "Employees")
Graph
```

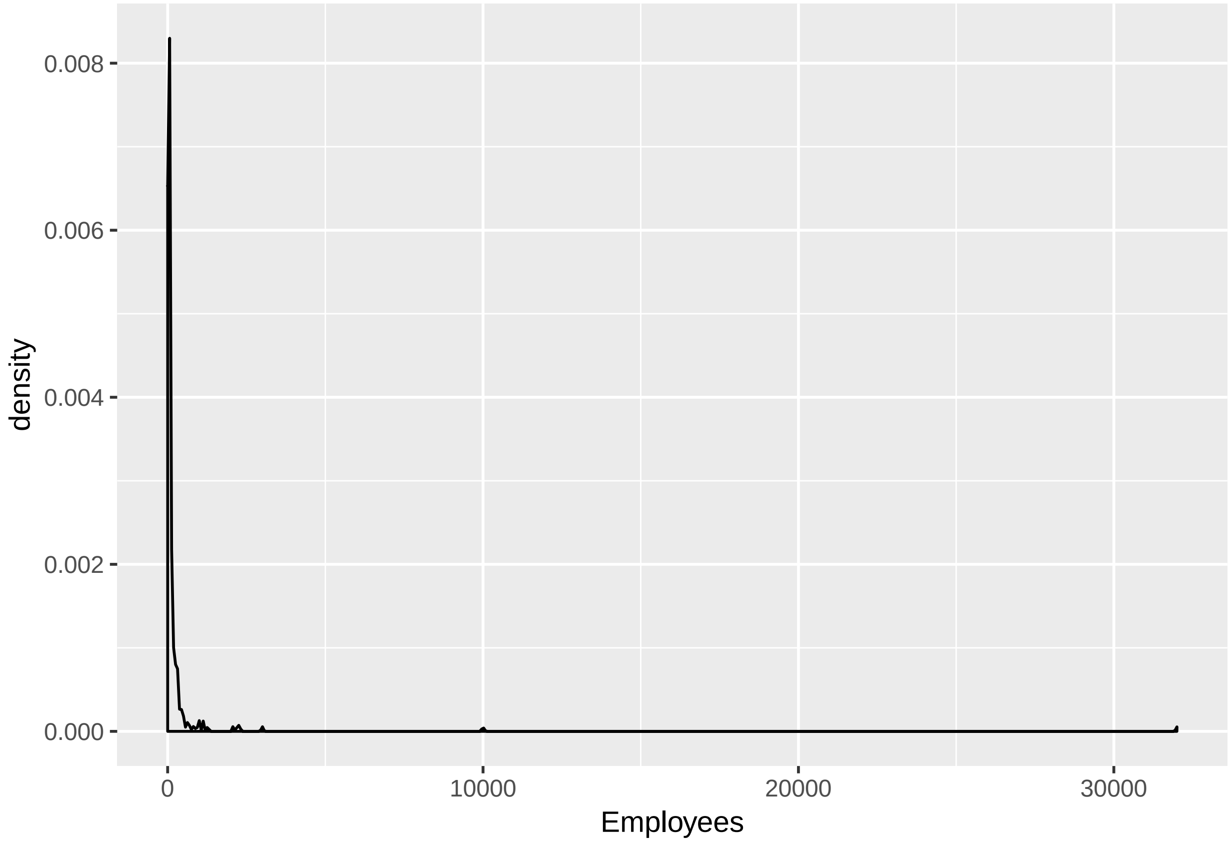


```
Graph + scale_y_log10()
```



Remove outlier:

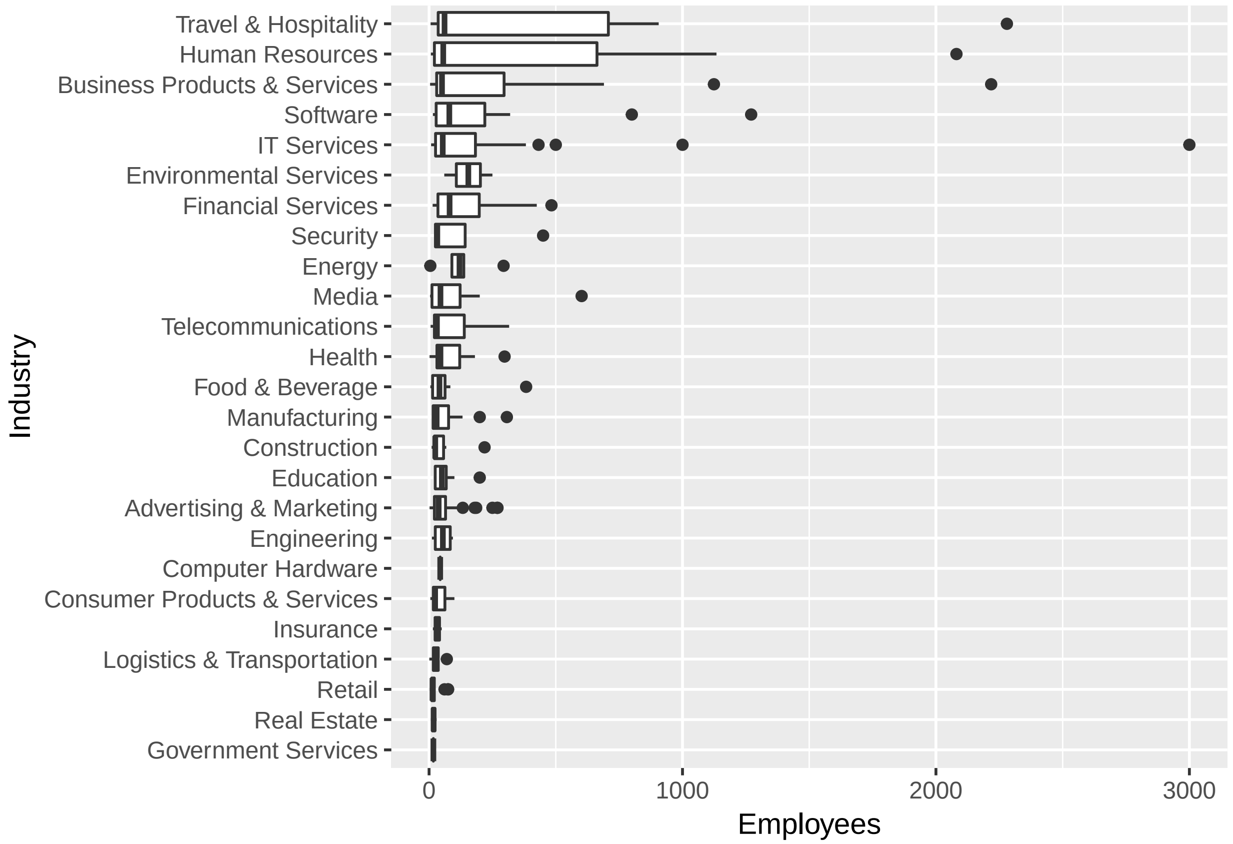
```
c <- ggplot(New_York, aes(Employees))
c + geom_density(kernel = "gaussian")
```



```
head(New_York %>% arrange(desc(Employees)))
```

```
##   Rank                        Name Growth_Rate   Revenue
## 1 4577 Sutherland Global Services        0.48 5.976e+08
## 2 4936                        Coty        0.36 4.600e+09
## 3 4716               Westcon Group        0.44 3.800e+09
## 4 3899  Denihan Hospitality Group        0.71 2.808e+08
## 5 4363                TransPerfect        0.55 3.413e+08
## 6 1499       Sterling Infosystems        2.66 2.149e+08
##                      Industry Employees      City State
## 1 Business Products & Services     32000 Pittsford    NY
## 2 Consumer Products & Services     10000  New York    NY
## 3                  IT Services      3000 Tarrytown    NY
## 4          Travel & Hospitality      2280  New York    NY
## 5 Business Products & Services      2218  New York    NY
## 6              Human Resources      2081  New York    NY
```

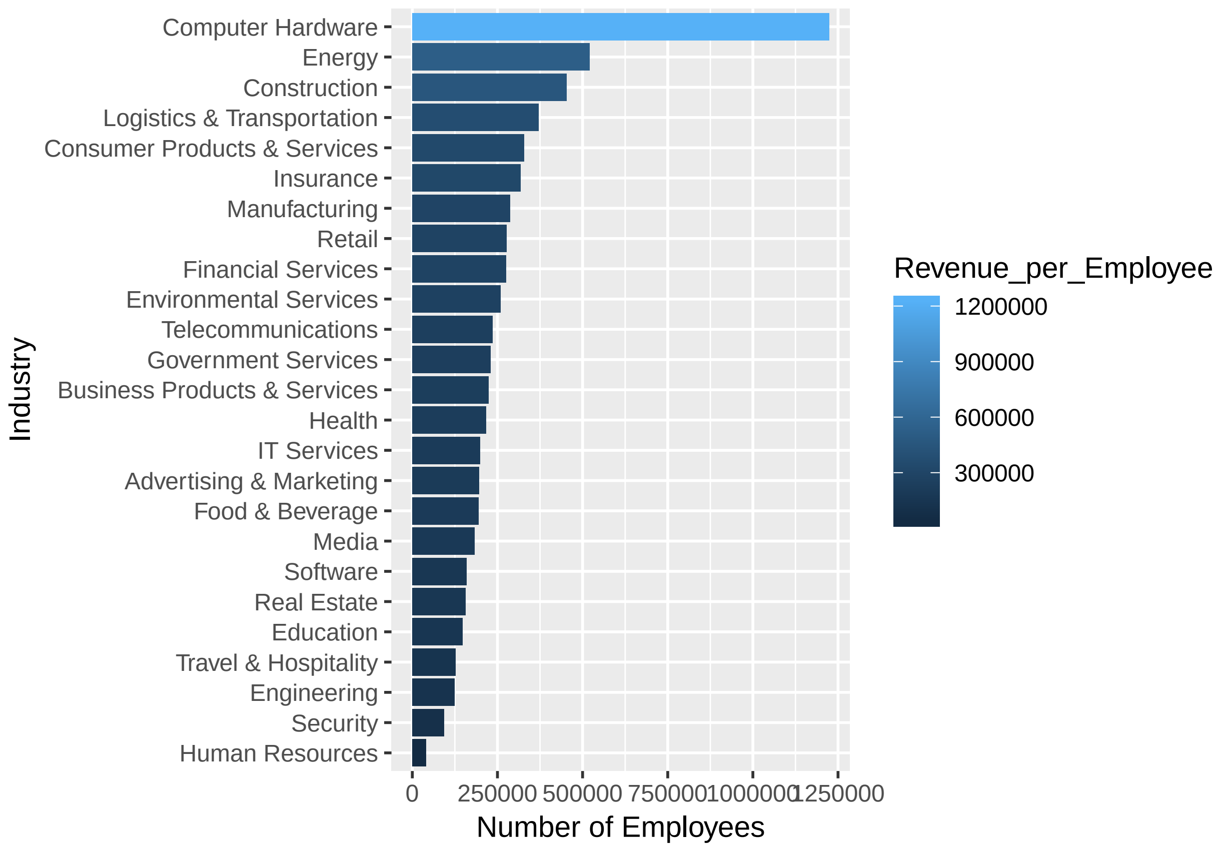
```
New_York_Without_outliers = New_York %>%
  filter(Employees <= 3000)

graph <- ggplot(New_York_Without_outliers, aes(reorder(Industry,Employees,mean), Employees))
graph <- graph + geom_boxplot() + coord_flip() + labs(x = "Industry", y = "Employees")
graph
```



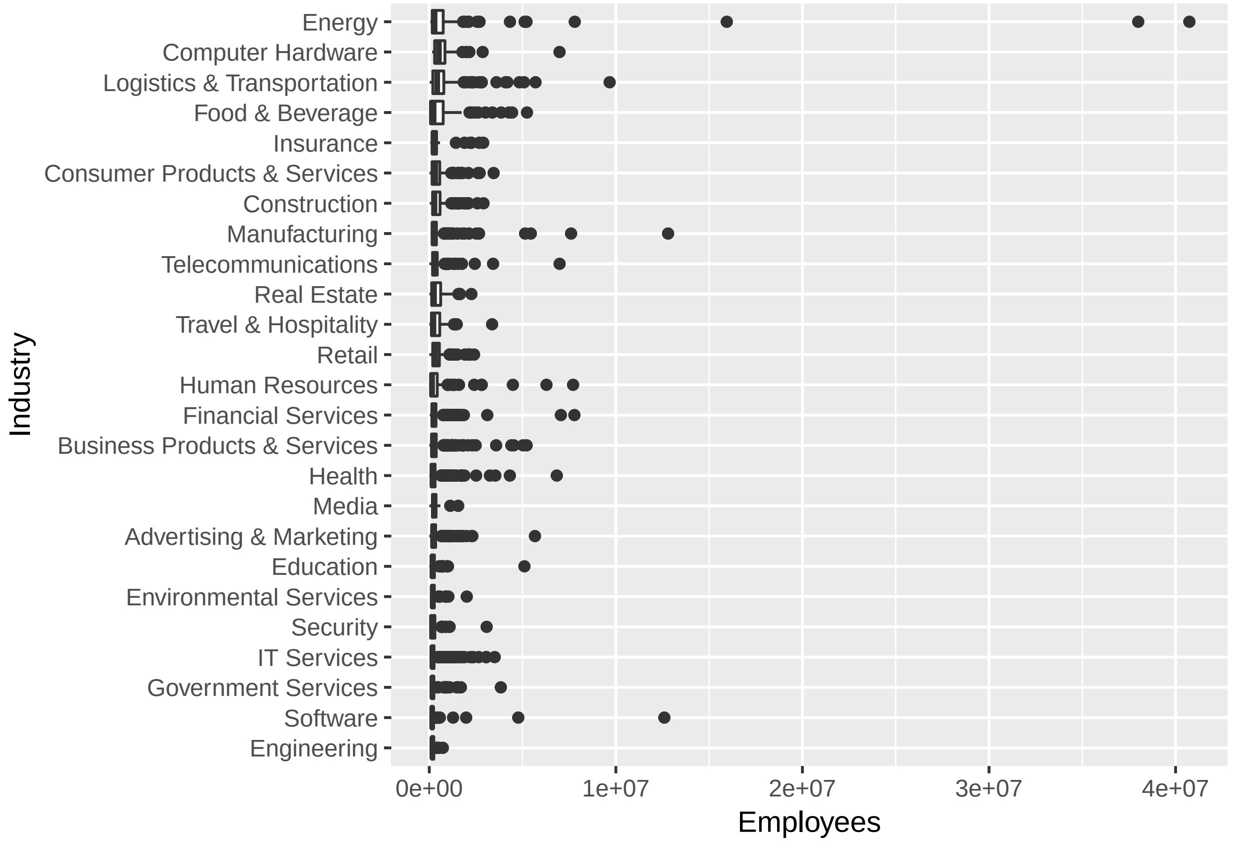
Question:3 Now imagine you work for an investor and want to see which industries generate the most revenue per employee. Create a chart that makes this information clear. Once again, the distribution per industry should be shown.

Answer:

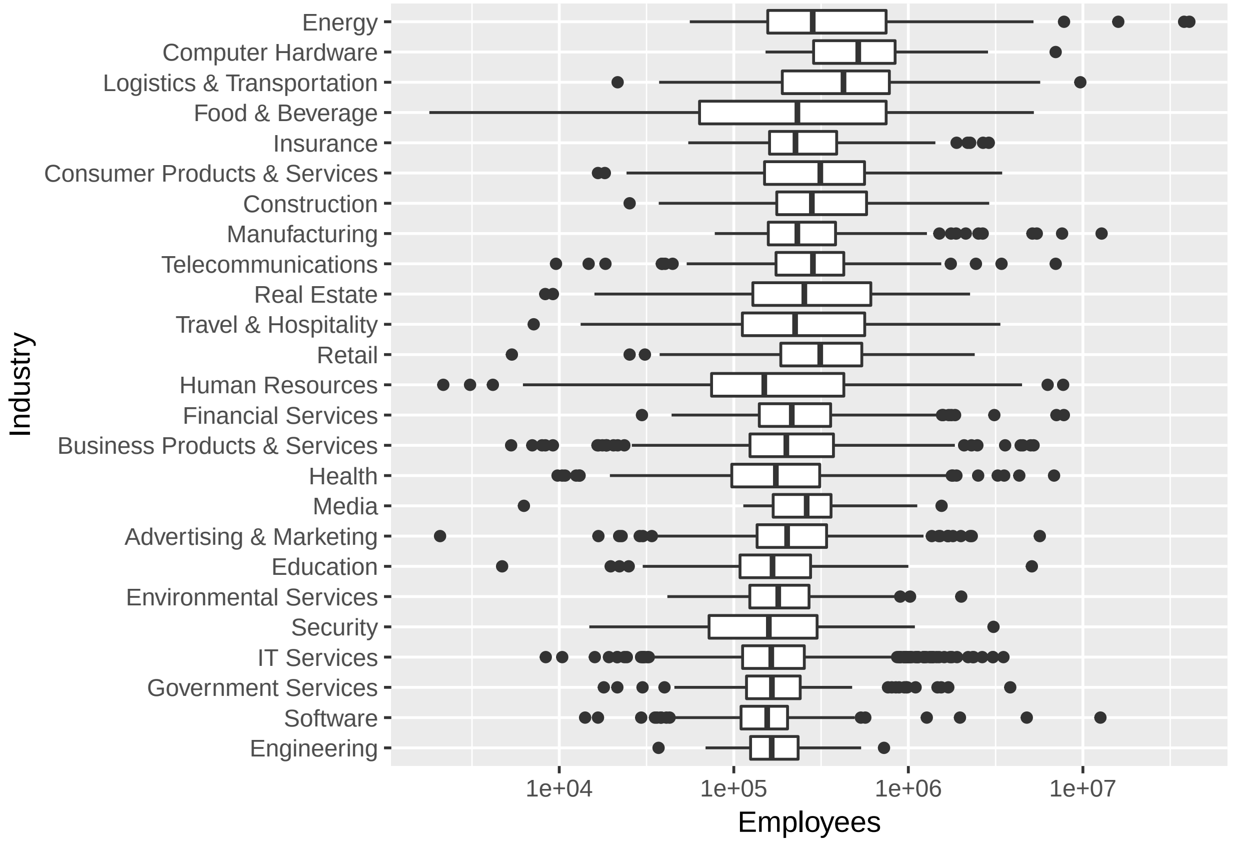
```
inc <- inc[complete.cases(inc),]

industry_emp = inc %>%
  group_by(Industry) %>%
  summarise(Revenue=sum(Revenue), Employees=sum(Employees)) %>%
  mutate(Revenue_per_Employee = Revenue/Employees)

Chart <- ggplot(industry_emp, aes(x=reorder(Industry, Revenue_per_Employee), y=Revenue_per_Employee, fil
Chart + geom_bar(stat="identity") + coord_flip() + labs(x = "Industry", y = "Number of Employees")
```



```
rev_emp = inc %>%
  mutate(Revenue_per_Employee = Revenue/Employees)

graph1 <- ggplot(rev_emp, aes(reorder(Industry,Revenue_per_Employee,mean), Revenue_per_Employee))
graph1<- graph1 + geom_boxplot() + coord_flip() + labs(x = "Industry", y = "Employees")
graph1
```



```
graph1 + scale_y_log10()
```


Energy
Computer Hardware
Logistics & Transportation
Food & Beverage
Insurance
Consumer Products & Services
Construction
Manufacturing
Telecommunications
Real Estate
Travel & Hospitality
Retail
Human Resources
Financial Services
Business Products & Services
Health
Media
Advertising & Marketing
Education
Environmental Services
Security
IT Services
Government Services
Software
Engineering
Industry
1e+04
1e+05
1e+06
1e+07
Employees