R Notebook

Principles of Data Visualization and Introduction to ggplot2

```
library(ggplot2)
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
```

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
##  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
# Insert your code here, create more chunks as necessary
less_100_emp = subset(inc, Employees < 100)
btwn_100_1000_emp = subset(inc,  (Employees <= 1000) & (Employees >= 100))
grtr_1000_emp = subset(inc,  Employees > 1000)
na_emp = subset(inc, is.na(inc$Employees) )

summary(less_100_emp)
```

```
##       Rank                                     Name       Growth_Rate
##  Min.   :   2   (Add)ventures                    :   1   Min.   :  0.340
##  1st Qu.:1134   @Properties                      :   1   1st Qu.:  0.810
##  Median :2384   1-Stop Translation USA           :   1   Median :  1.515
##  Mean   :2415   11thStreetCoffee.com             :   1   Mean   :  4.707
##  3rd Qu.:3674   1st American Systems and Services:   1   3rd Qu.:  3.700
##  Max.   :5000   1st Equity                       :   1   Max.   :248.310
##                 (Other)                          :3418
##     Revenue                                 Industry      Employees
##  Min.   :2.00e+06   IT Services                 : 493   Min.   : 1.00
##  1st Qu.:4.10e+06   Advertising & Marketing     : 396   1st Qu.:19.00
##  Median :7.20e+06   Business Products & Services: 334   Median :32.00
##  Mean   :1.43e+07   Software                    : 216   Mean   :39.14
##  3rd Qu.:1.38e+07   Health                      : 210   3rd Qu.:55.00
##  Max.   :1.90e+09   Manufacturing               : 175   Max.   :99.00
##                     (Other)                     :1600
##             City          State
##  New York     : 112   CA     : 507
##  Austin       :  64   TX     : 240
##  Chicago      :  63   NY     : 227
##  San Francisco:  54   FL     : 195
##  Atlanta      :  53   VA     : 187
##  San Diego    :  53   IL     : 182
##  (Other)      :3025   (Other):1886
```

```
summary(btwn_100_1000_emp)
```

```
##       Rank                             Name       Growth_Rate
##  Min.   :   1   110 Consulting           :   1   Min.   :  0.350
##  1st Qu.:1491   123 Exteriors            :   1   1st Qu.:  0.730
##  Median :2704   2020 Exhibits            :   1   Median :  1.300
##  Mean   :2646   21c Museum Hotels        :   1   Mean   :  4.523
##  3rd Qu.:3855   22nd Century Technologies:   1   3rd Qu.:  2.680
##  Max.   :4990   29 Prime                 :   1   Max.   :421.480
##                 (Other)                  :1391
##     Revenue                                 Industry     Employees
##  Min.   :2.10e+06   IT Services                 :227   Min.   : 100.0
##  1st Qu.:1.72e+07   Business Products & Services:131   1st Qu.: 138.0
##  Median :3.14e+07   Health                      :128   Median : 202.0
##  Mean   :6.99e+07   Software                    :118   Mean   : 275.8
##  3rd Qu.:6.62e+07   Financial Services          : 87   3rd Qu.: 345.0
##  Max.   :2.70e+09   Manufacturing               : 74   Max.   :1000.0
##                     (Other)                     :632
##             City          State
##  New York     :  42   CA     :176
##  Houston      :  30   TX     :129
##  Austin       :  24   VA     : 92
##  Chicago      :  20   FL     : 76
##  San Francisco:  20   NY     : 75
##  Atlanta      :  18   IL     : 71
##  (Other)      :1243   (Other):778
```

```
summary(grtr_1000_emp)
```

```
##       Rank                   Name      Growth_Rate        Revenue
##  Min.   :  15   ABC Supply        :  1   Min.   :  0.340   Min.   :2.900e+06
##  1st Qu.:2090   Acadian Companies :  1   1st Qu.:  0.630   1st Qu.:8.358e+07
##  Median :3404   Accurate Home Care:  1   Median :  0.915   Median :2.265e+08
##  Mean   :3050   Acro Service      :  1   Mean   :  3.488   Mean   :5.603e+08
##  3rd Qu.:4128   Addison Group     :  1   3rd Qu.:  1.785   3rd Qu.:5.091e+08
##  Max.   :4997   Advanced Disposal :  1   Max.   :123.330   Max.   :1.010e+10
##                 (Other)           :162
##                          Industry   Employees            City
##  Human Resources             :21   Min.   : 1001   Chicago   :  7
##  Health                      :16   1st Qu.: 1325   New York  :  6
##  Business Products & Services:15   Median : 1948   Dallas    :  5
##  Food & Beverage             :15   Mean   : 3820   Houston   :  5
##  Financial Services          :12   3rd Qu.: 3899   Charlotte :  4
##  IT Services                 :12   Max.   :66803   Cincinnati:  3
##  (Other)                     :77                   (Other)   :138
##      State
##  IL     :19
##  CA     :17
##  TX     :17
##  FL     :11
##  NY     : 9
##  MI     : 7
##  (Other):88
```

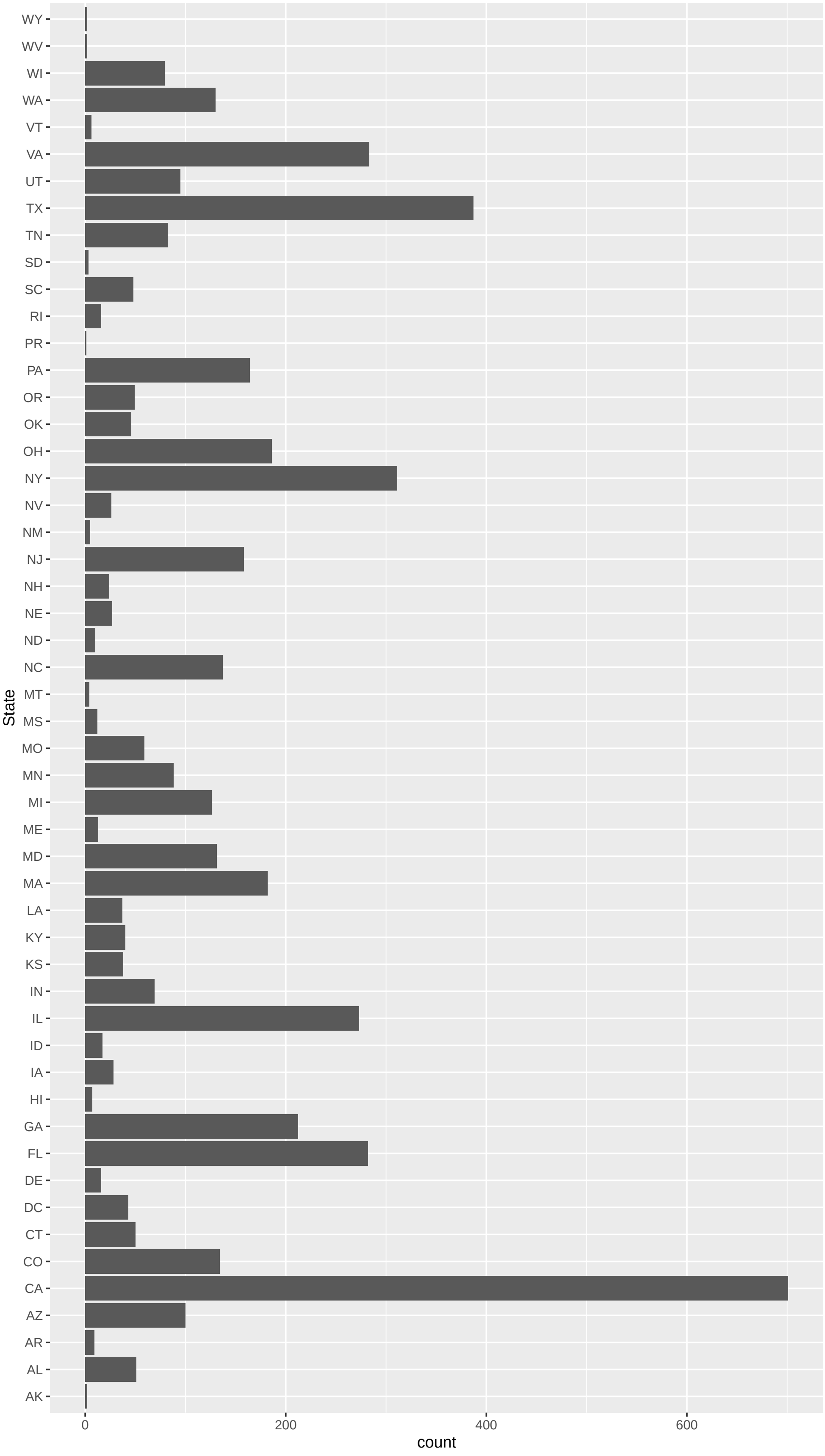
```
summary(na_emp)
```

```
##       Rank                                  Name     Growth_Rate
##  Min.   : 183   Carolinas Home Medical Equipment:1   Min.   : 0.350
##  1st Qu.:1521   Excalibur Exhibits              :1   1st Qu.: 0.670
##  Median :2470   First Flight Solutions          :1   Median : 1.475
##  Mean   :2606   Global Communications Group     :1   Mean   : 3.408
##  3rd Qu.:4012   Heartland Business Systems      :1   3rd Qu.: 2.700
##  Max.   :4968   Higher Logic                    :1   Max.   :22.320
##                 (Other)                         :6
##     Revenue                                 Industry   Employees
##  Min.   :  2700000   Business Products & Services:2   Min.   : NA
##  1st Qu.:  5025000   Food & Beverage             :2   1st Qu.: NA
##  Median :  9400000   Telecommunications          :2   Median : NA
##  Mean   : 35408333   Health                      :1   Mean   :NaN
##  3rd Qu.: 52275000   IT Services                 :1   3rd Qu.: NA
##  Max.   :156300000   Logistics & Transportation  :1   Max.   : NA
##                      (Other)                     :3   NA's   :12
##            City       State
##  Atlanta     :1   NC     :2
##  Bellevue    :1   WI     :2
##  Emerald Isle:1   CA     :1
##  Englewood   :1   CO     :1
##  Horsham     :1   DC     :1
##  houston     :1   GA     :1
##  (Other)     :6   (Other):4
```

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.

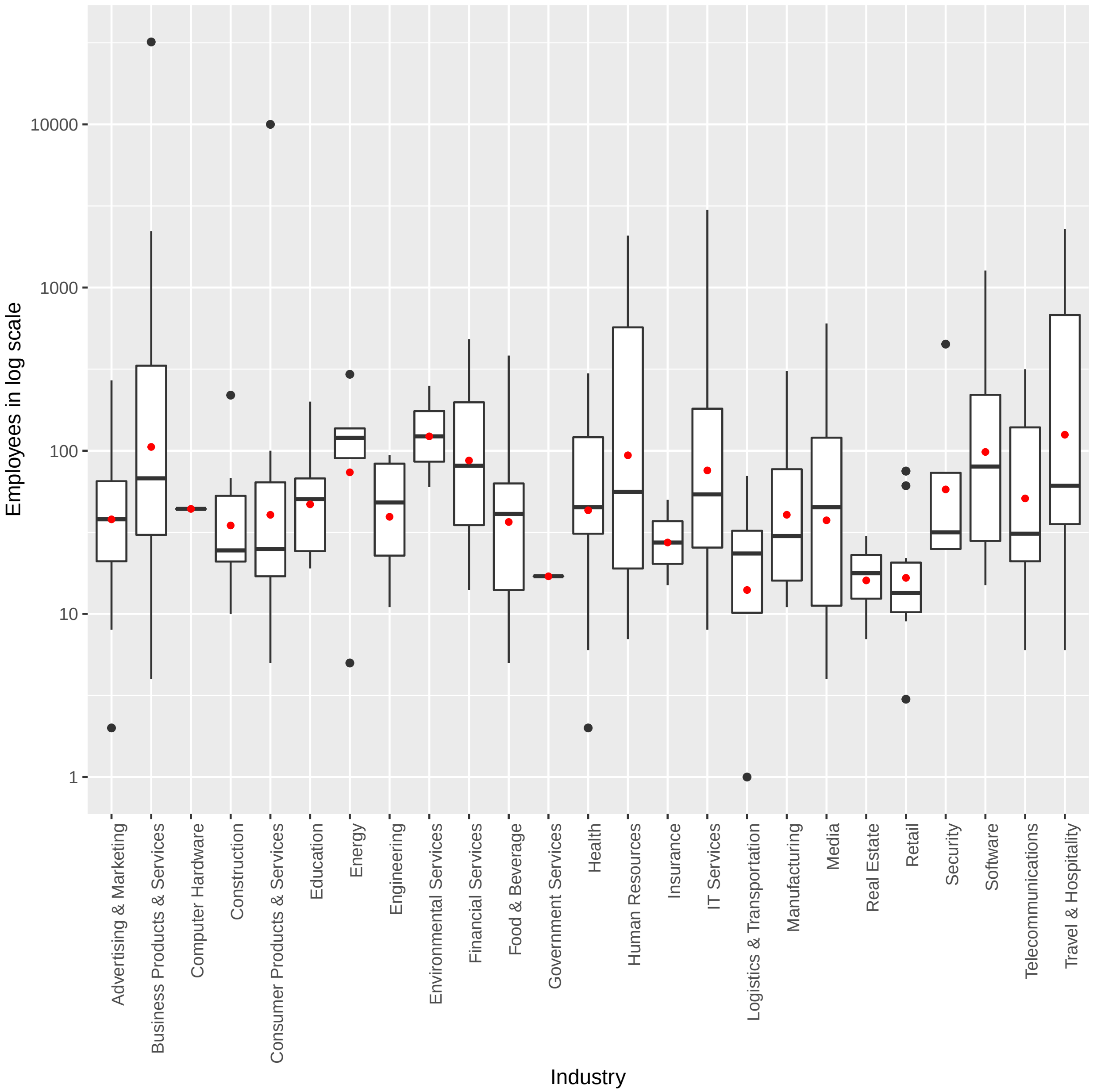
```
# Answer Question 1 here
ggplot(inc, aes(x=State)) + geom_bar() + coord_flip()
```


WY
WV
WI
WA
VT
VA
UT
TX
TN
SD
SC
RI
PR
PA
OR
OK
OH
NY
NV
NM
NJ
NH
NE
ND
NC
MT
MS
MO
MN
MI
ME
MD
MA
LA
KY
KS
IN
IL
ID
IA
HI
GA
FL
DE
DC
CT
CO
CA
AZ
AR
AL
AK
State
0
200
400
600
count

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.

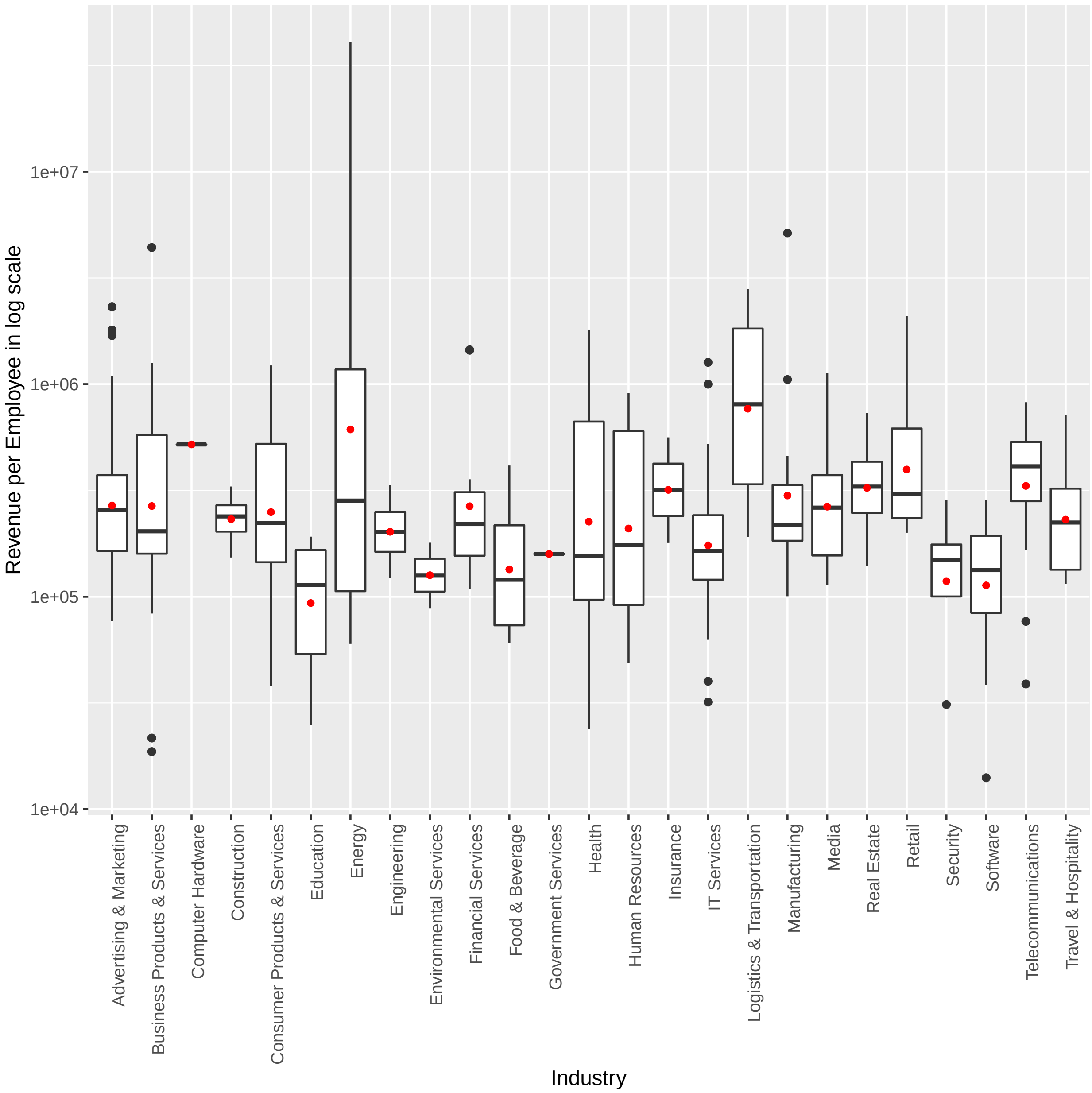
```r
# Answer Question 2 here
counts = count(inc, vars=State, sort = TRUE)
ordered_states = counts %>% pull(vars)
third_most_state = ordered_states[3]
NY_state = subset(inc, State == third_most_state)
subset = NY_state[complete.cases(NY_state),]
ggplot(subset, aes(y = Employees, x = Industry)) + geom_boxplot() + scale_y_continuous(trans='log10') +
```



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
# Assuming still for NY state
NY_state$Rev_per_emp = NY_state$Revenue/NY_state$Employees
subset = NY_state[complete.cases(NY_state),]
ggplot(subset, aes(y = Rev_per_emp, x = Industry)) + geom_boxplot() + scale_y_continuous(trans='log10')
```


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