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:

```
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
```

```
options(scipen=100)

library(dplyr)
library(ggplot2)
```

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

# Company Count by Industry
sort(table(inc$Industry), decreasing = T)
```

```
##
##                  IT Services Business Products & Services
##                          733                          482
##      Advertising & Marketing                       Health
##                          471                          355
##                     Software           Financial Services
##                          342                          260
##                Manufacturing Consumer Products & Services
##                          256                          203
##                       Retail          Government Services
##                          203                          202
##              Human Resources                 Construction
##                          196                          187
##   Logistics & Transportation              Food & Beverage
##                          155                          131
##           Telecommunications                       Energy
##                          129                          109
##                  Real Estate                    Education
##                           96                           83
##                  Engineering                     Security
##                           74                           73
##         Travel & Hospitality                        Media
##                           62                           54
##       Environmental Services                    Insurance
##                           51                           50
##            Computer Hardware
##                           44
```

```
# % of Companies w/ <= 100 Employees
sum(inc$Employees <= 100, na.rm = T) / nrow(inc)
```

```
## [1] 0.6910618
```

```r
# Company Count by State
sort(table(inc$State), decreasing = T)
```

```
##
##  CA  TX  NY  VA  FL  IL  GA  OH  MA  PA  NJ  NC  CO  MD  WA  MI  AZ  UT  MN  TN
## 701 387 311 283 282 273 212 186 182 164 158 137 134 131 130 126 100  95  88  82
##  WI  IN  MO  AL  CT  OR  SC  OK  DC  KY  KS  LA  IA  NE  NV  NH  ID  DE  RI  ME
##  79  69  59  51  50  49  48  46  43  40  38  37  28  27  26  24  17  16  16  13
##  MS  ND  AR  HI  VT  NM  MT  SD  AK  WV  WY  PR
##  12  10   9   7   6   5   4   3   2   2   2   1
```

```r
# Top 10 Cities by Company Count
group_by(inc, City) %>%
  summarize(count = n(), .groups = 'drop') %>%
  arrange(desc(count)) %>%
  top_n(10)
```

```
## Selecting by count
```

```
## # A tibble: 11 x 2
##    City          count
##    <chr>         <int>
##  1 New York        160
##  2 Chicago          90
##  3 Austin           88
##  4 Houston          76
##  5 San Francisco    75
##  6 Atlanta          74
##  7 San Diego        67
##  8 Seattle          52
##  9 Boston           43
## 10 Dallas           42
## 11 Denver           42
```

```r
# Top 10 Companies by Revenue
mutate(inc, Revenue_in_Billions = Revenue/1000000000) %>%
  select(Name, Industry, Revenue_in_Billions) %>%
  arrange(desc(Revenue_in_Billions)) %>%
  top_n(10)
```

```
## Selecting by Revenue_in_Billions
```

```
##                          Name                     Industry Revenue_in_Billions
## 1                         CDW            Computer Hardware                10.1
## 2                 ABC Supply                 Construction                 4.7
## 3                       Coty Consumer Products & Services                 4.6
## 4                  Dot Foods              Food & Beverage                 4.5
## 5              Westcon Group                  IT Services                 3.8
## 6  American Tire Distributors Consumer Products & Services                 3.5
## 7                    Kum & Go                      Retail                 2.8
## 8              Boise Cascade                 Construction                 2.8
## 9          EnvisionRxOptions                      Health                 2.7
## 10                 DLA Piper Business Products & Services                 2.4
```

```r
# Top 10 Fastest Growing Companies
select(inc, Name, Industry, Growth_Rate) %>%
  arrange(desc(Growth_Rate)) %>%
  top_n(10)
```

```
## Selecting by Growth_Rate
```

```
##                            Name                       Industry Growth_Rate
## 1                    Fuhu Consumer Products & Services      421.48
## 2         FederalConference.com          Government Services      248.31
## 3                 The HCI Group                       Health      245.45
## 4                       Bridger                       Energy      233.08
## 5                        DataXu      Advertising & Marketing      213.37
## 6  MileStone Community Builders                   Real Estate      179.38
## 7         Value Payment Systems           Financial Services      174.04
## 8          Emerge Digital Group      Advertising & Marketing      170.64
## 9                   Goal Zero Consumer Products & Services      169.81
## 10                    Yagoozon                       Retail      166.89
```

```r
# Top 10 Industries - Average Growth
group_by(inc, Industry) %>%
  summarize(avg_growth_rate = mean(Growth_Rate, na.rm = T), .groups = 'drop') %>%
  arrange(desc(avg_growth_rate)) %>%
  top_n(10)
```

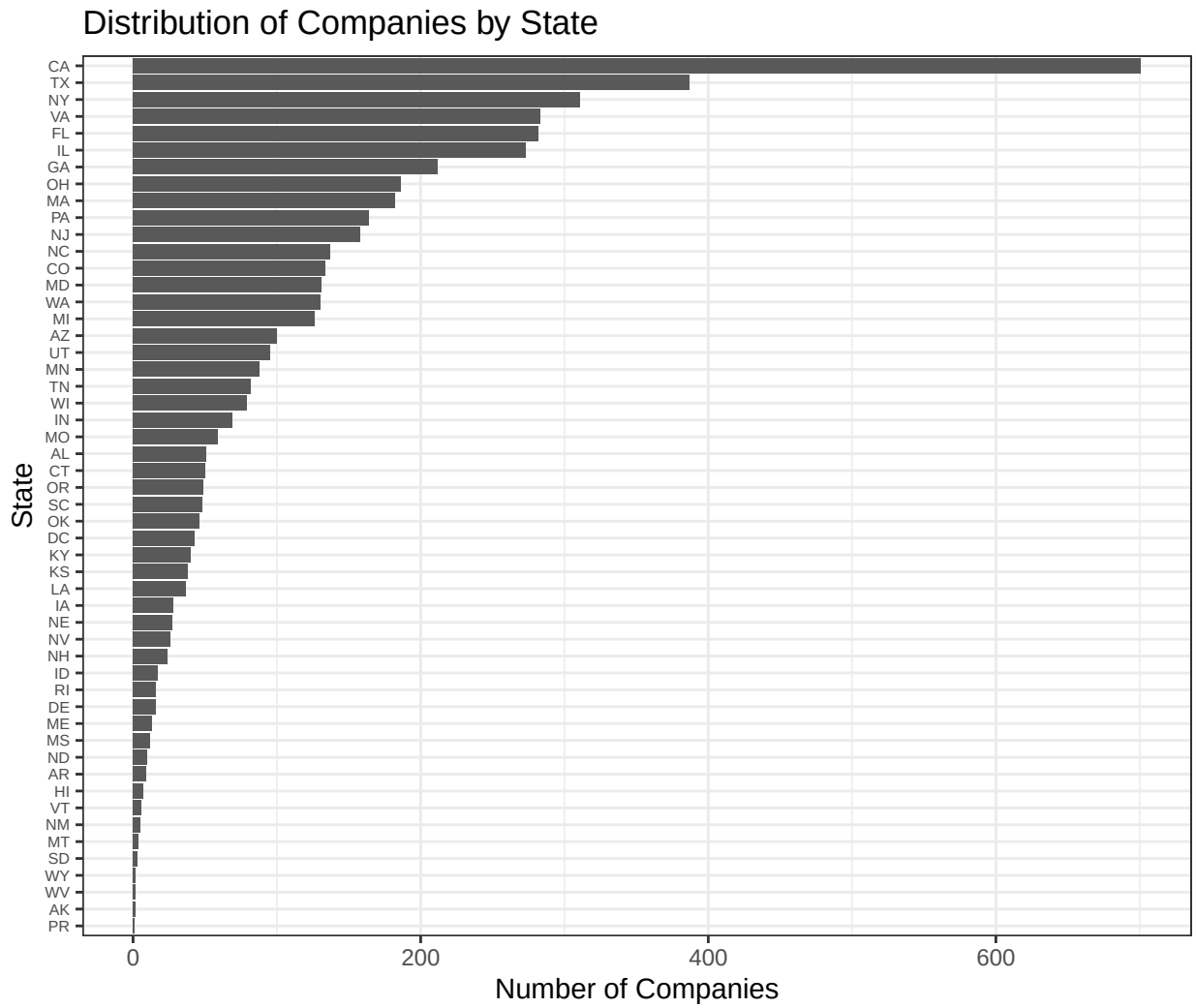
```
## Selecting by avg_growth_rate
```

```
## # A tibble: 10 x 2
##    Industry                     avg_growth_rate
##    <chr>                                  <dbl>
##  1 Energy                                  9.60
##  2 Consumer Products & Services            8.78
##  3 Real Estate                             7.75
##  4 Government Services                     7.24
##  5 Advertising & Marketing                 6.23
##  6 Retail                                  6.18
##  7 Financial Services                      5.44
##  8 Software                                5.02
##  9 Health                                  4.86
## 10 Media                                   4.37
```

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.

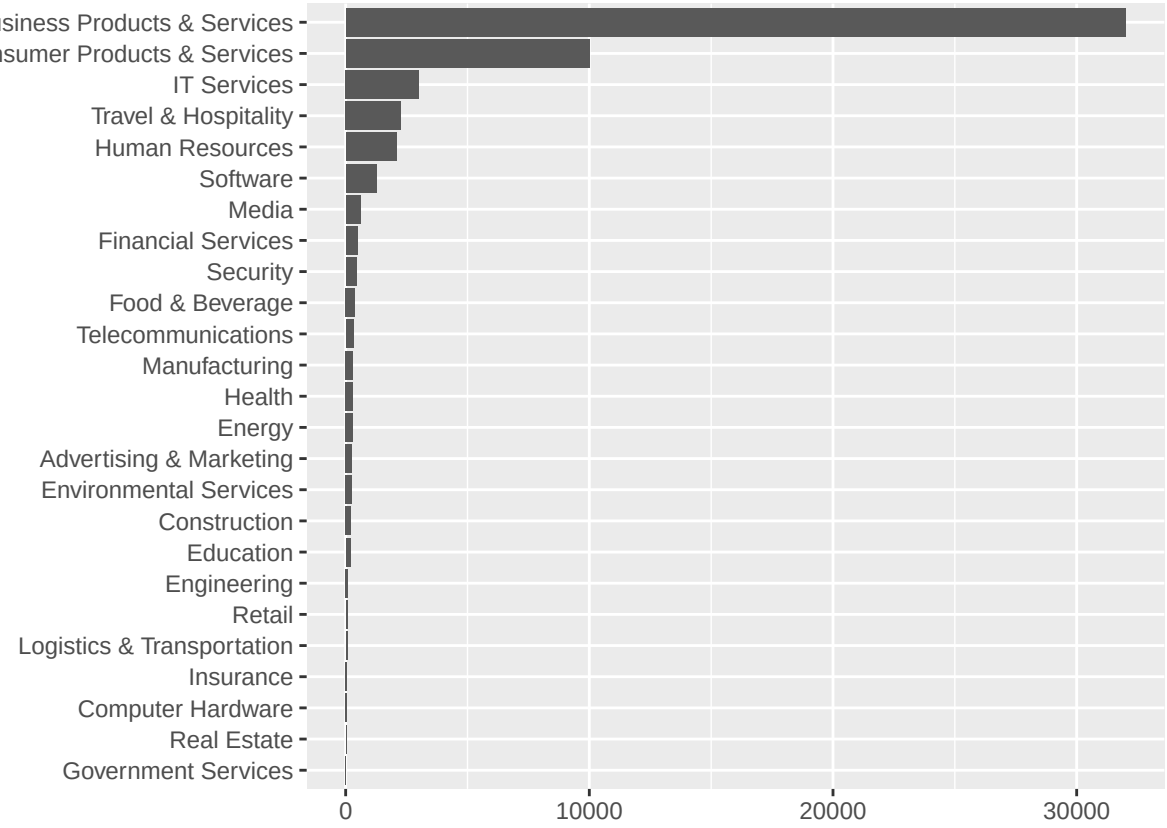
```
# Answer Question 1 here
# Companies by State
group_by(inc, State) %>%
  summarize(count = n(), .groups = 'drop') %>%
  ggplot(aes(y = reorder(State, count), x = count)) +
  geom_col() +
  theme_bw() +
  labs(x = 'Number of Companies',
       y = 'State',
       title = 'Distribution of Companies by State') +
  theme(axis.text.y = element_text(size = 6))
```



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
# ny <- filter(inc, State == 'NY' & !is.na(Industry) & !is.na(Employees))
ny <- filter(inc, State == 'NY' & complete.cases(inc) == T)

# Industry Summary
industry_scale <- group_by(ny, Industry) %>%
  summarize(max_e = max(Employees),
            avg_e = mean(Employees),
            med_e = median(Employees),
            .groups = 'drop')

# quick peek at max number of employees by industry
ggplot(industry_scale, aes(x = reorder(Industry, max_e), y = max_e)) +
  geom_col() +
  coord_flip() +
  labs(x = element_blank(),
       y = element_blank())
```

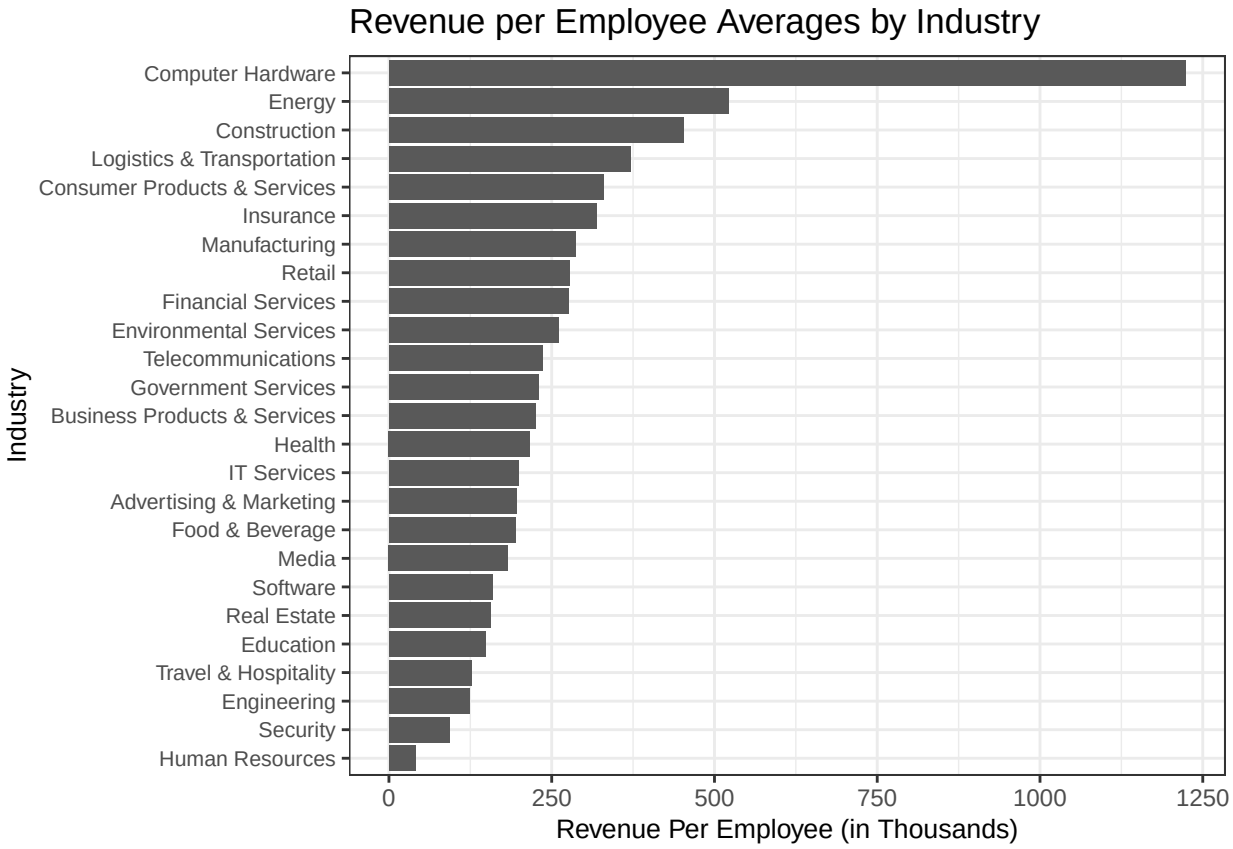


```r
# based on the max employee chart, lets split the industries into small, medium, and large
industry_scale$size <- with(industry_scale, ifelse(max_e >= 10000, 'Large',
                                                   ifelse(max_e >= 100, 'Medium', 'Small')))

industry_scale$size <- factor(industry_scale$size, levels = c('Small', 'Medium', 'Large'))

# Relabel Industries
industry_scale$label <- with(industry_scale,
                             paste0(Industry, ' (', round(avg_e, 0), ', ', round(med_e, 0), ')'))

# join industry size and labels
ny <- inner_join(ny, industry_scale, by = c('Industry' = 'Industry'))

# Distribution of Employees by Industry in NY State
ggplot(ny, aes(x = Employees, y = reorder(label, med_e))) +
  geom_boxplot() +
  facet_wrap(~size, scale = 'free_x') +
  labs(y = 'Industry (Average Employees, Median Employees)',
       title = 'Employees Distributions by Industry in NY',
       subtitle = 'Size Based on Largest Company in Each Industry') +
  theme(axis.title.y = element_text(size = 10),
        axis.text.y = element_text(size = 8),
        axis.title.x = element_text(size = 10),
        axis.text.x = element_text(size = 8))
```

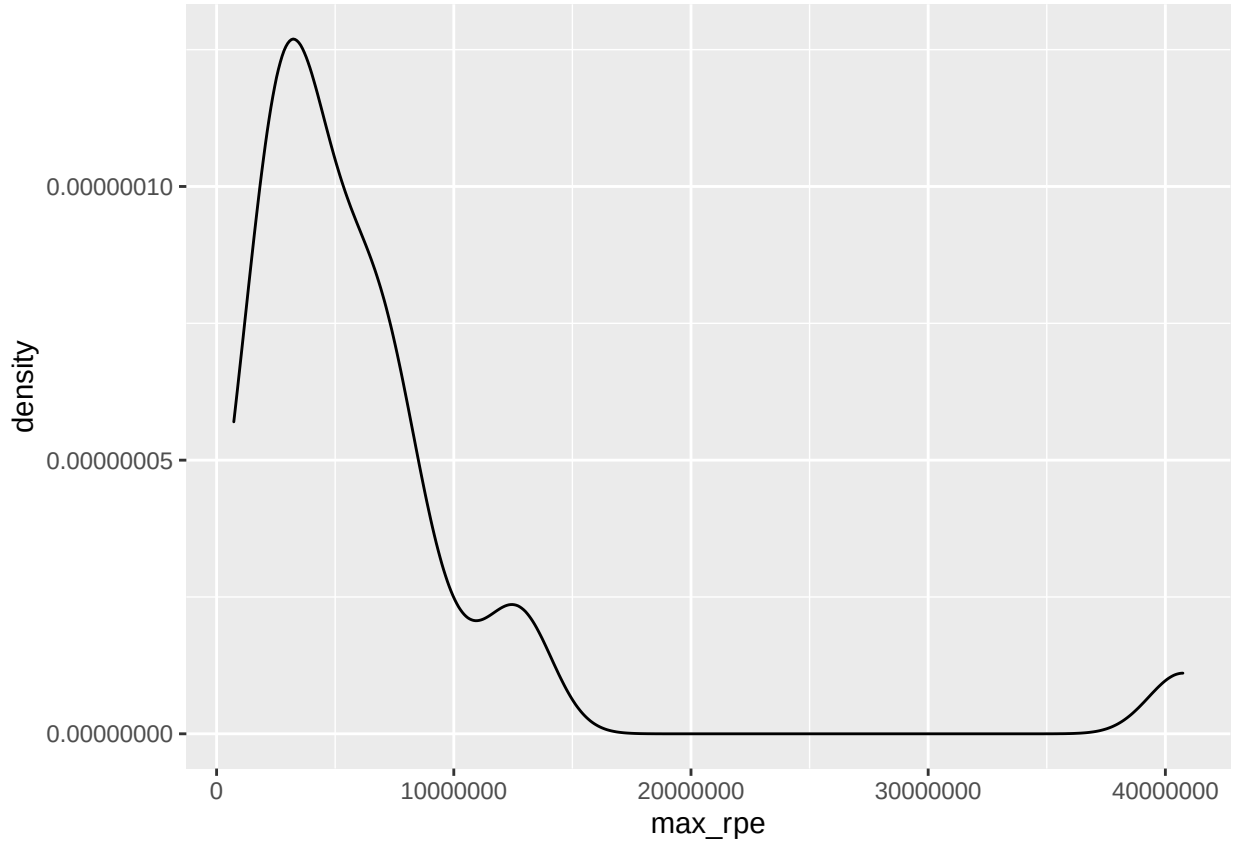
Question 3

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
filter(inc, !is.na(Revenue) & !is.na(Employees) & !is.na(Industry)) %>%
  group_by(Industry) %>%
  summarize(rpe = sum(Revenue) / sum(Employees), .groups = 'drop') %>%
  ggplot(aes(y = reorder(Industry, rpe), x = rpe/1000)) +
  geom_col() +
  theme_bw() +
  labs(x = 'Revenue Per Employee (in Thousands)',
       y = 'Industry',
       title = 'Revenue per Employee Averages by Industry') +
  theme(axis.text.y = element_text(size = 8),
        axis.title.y = element_text(size = 10),
        axis.title.x = element_text(size = 10))
```

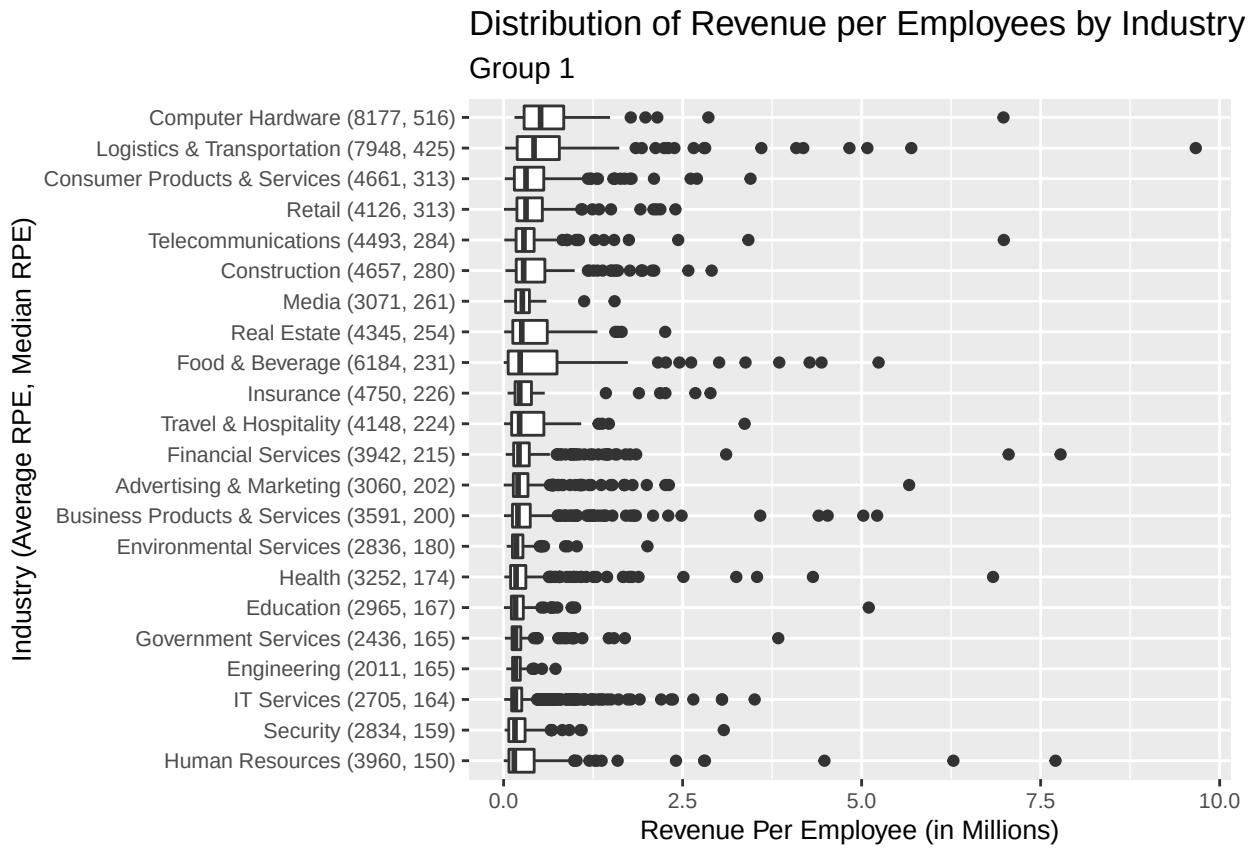


```r
# add calculated field
inc$rpe <- with(inc, Revenue / Employees)

# calculate summary stats per industry
med_rpe <- group_by(inc, Industry) %>%
  summarize(med_rpe = median(rpe, na.rm = T),
            avg_rpe = mean(rpe, na.rm = T),
            max_rpe = max(rpe, na.rm = T),
            .groups = 'drop')

# check distribution of max rpe to determine where to split chart
ggplot(med_rpe, aes(x = max_rpe)) +
  geom_density()
```




```r
# Looks like we can try to split around 10M
```

```r
# create label
med_rpe$label <- with(med_rpe,
                      paste0(Industry,
                             ' (', round(avg_rpe/100, 0), ', ', round(med_rpe/1000, 0), ')'))

# plot < 10M
inner_join(inc, med_rpe, by = c('Industry' = 'Industry')) %>%
filter(!is.na(rpe) & max_rpe < 10000000) %>%
  ggplot(aes(y = reorder(label, med_rpe), x = rpe/1000000)) +
  geom_boxplot() +
  theme(axis.title.y = element_text(size = 10),
        axis.text.y = element_text(size = 8),
        axis.title.x = element_text(size = 10),
        axis.text.x = element_text(size = 8)) +
  labs(x = 'Revenue Per Employee (in Millions)',
       y = 'Industry (Average RPE, Median RPE)',
       title = 'Distribution of Revenue per Employees by Industry',
       subtitle = 'Group 1')
```



```r
# plot >= 10M
inner_join(inc, med_rpe, by = c('Industry' = 'Industry')) %>%
filter(!is.na(rpe) & max_rpe >= 10000000) %>%
  ggplot(aes(y = reorder(label, med_rpe), x = rpe/1000000)) +
  geom_boxplot() +
  theme(axis.title.y = element_text(size = 10),
        axis.text.y = element_text(size = 8),
        axis.title.x = element_text(size = 10),
        axis.text.x = element_text(size = 8)) +
  labs(x = 'Revenue Per Employee (in Millions)',
       y = 'Industry (Average RPE, Median RPE)',
       title = 'Distribution of Revenue per Employees by Industry',
       subtitle = 'Group 2')
```