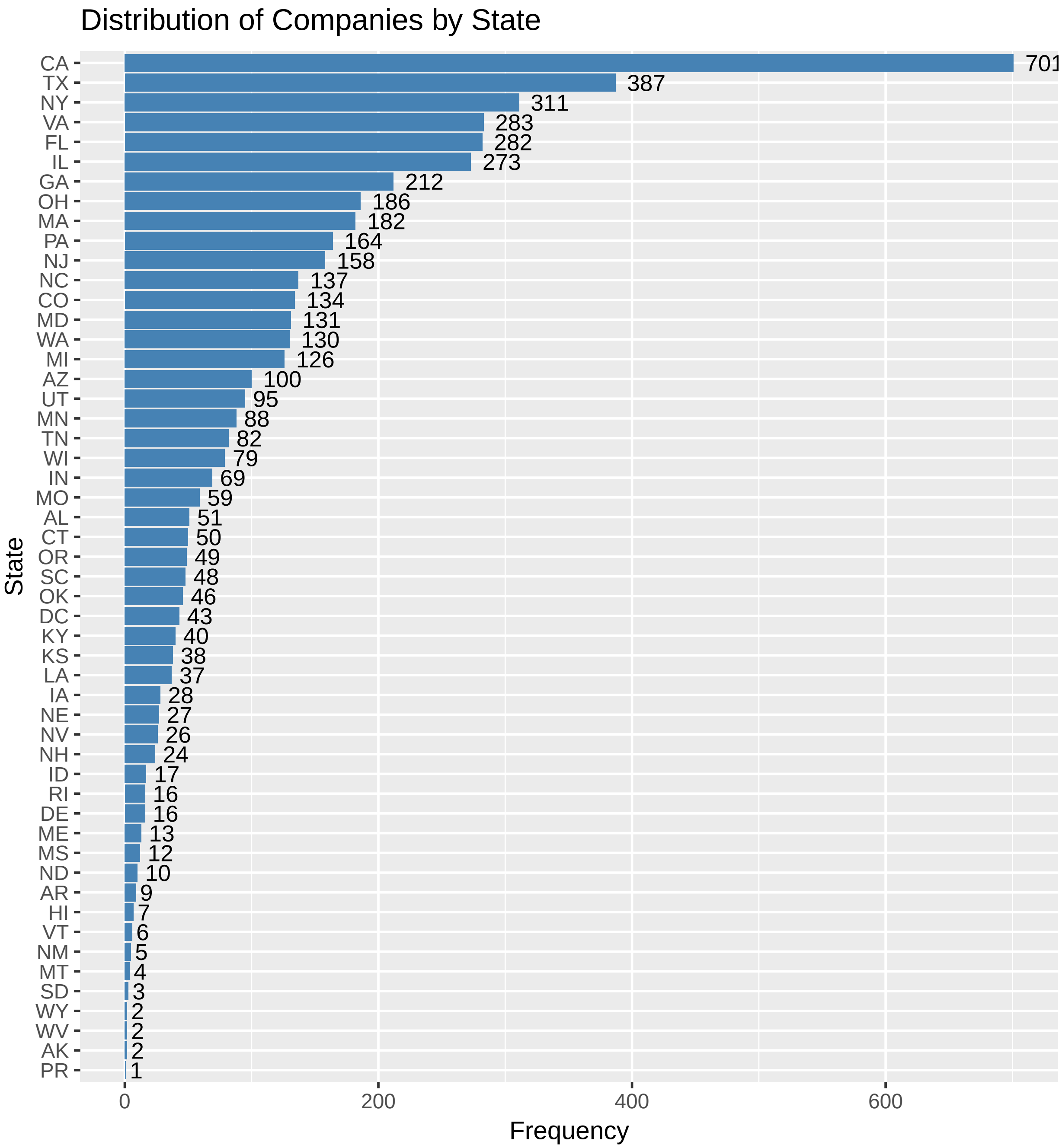
Data 608 Homework 1

Aadi Kalloo

2/3/2017

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 assuming I am using a 'portrait' oriented screen (ie taller than wide).

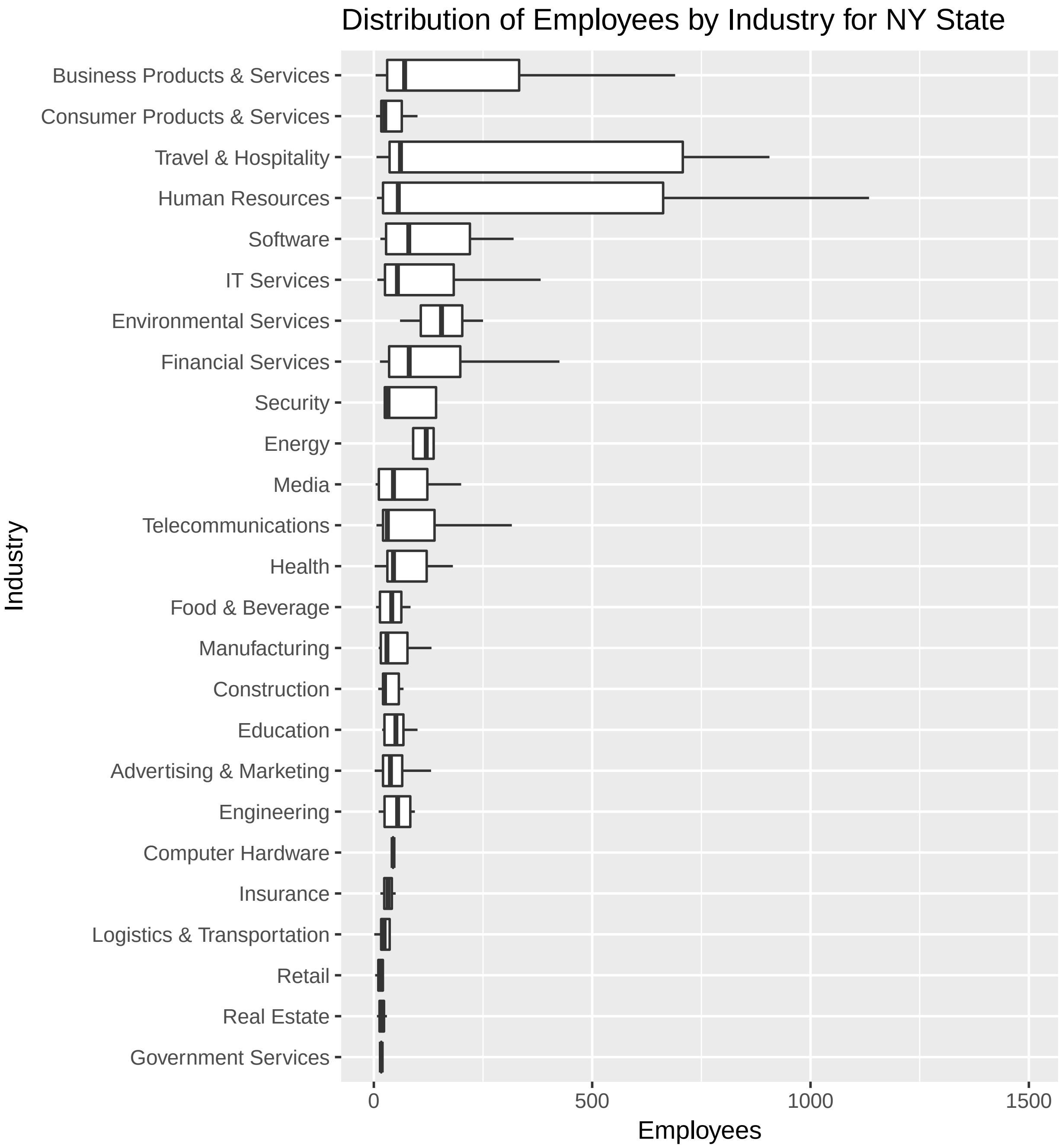
```
data = read.csv(data_url, stringsAsFactors = FALSE)
state_count = plyr::count(data, vars = "State")
state_count = state_count[order(-state_count$freq), ]
row.names(state_count) = NULL

ggplot(data = state_count, aes(x = reorder(State, freq), y = freq)) +
    coord_flip() + geom_bar(stat = "identity", fill = "steelblue") +
    geom_text(aes(label = freq), hjust = -0.3, color = "black",
        size = 3.5) + labs(title = "Distribution of Companies by State",
    x = "State", y = "Frequency")
```



2. Let's 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 employ. Create a plot of average employment by industry for companies in this state (only use cases with full data (user R's complete.cases() function). Your graph should show how variable the ranges are, and exclude outliers.

```
state_3 = state_count$State[3]
state_3_df = data[data$State == state_3, ]
state_3_df = state_3_df[complete.cases(state_3_df), ]
state_3_df_group_summary = ddply(state_3_df, .(Industry), summarize,
    Average = mean(Employees))
ylim1 = c(0, boxplot.stats(state_3_df_group_summary$Average)$out[1])
```

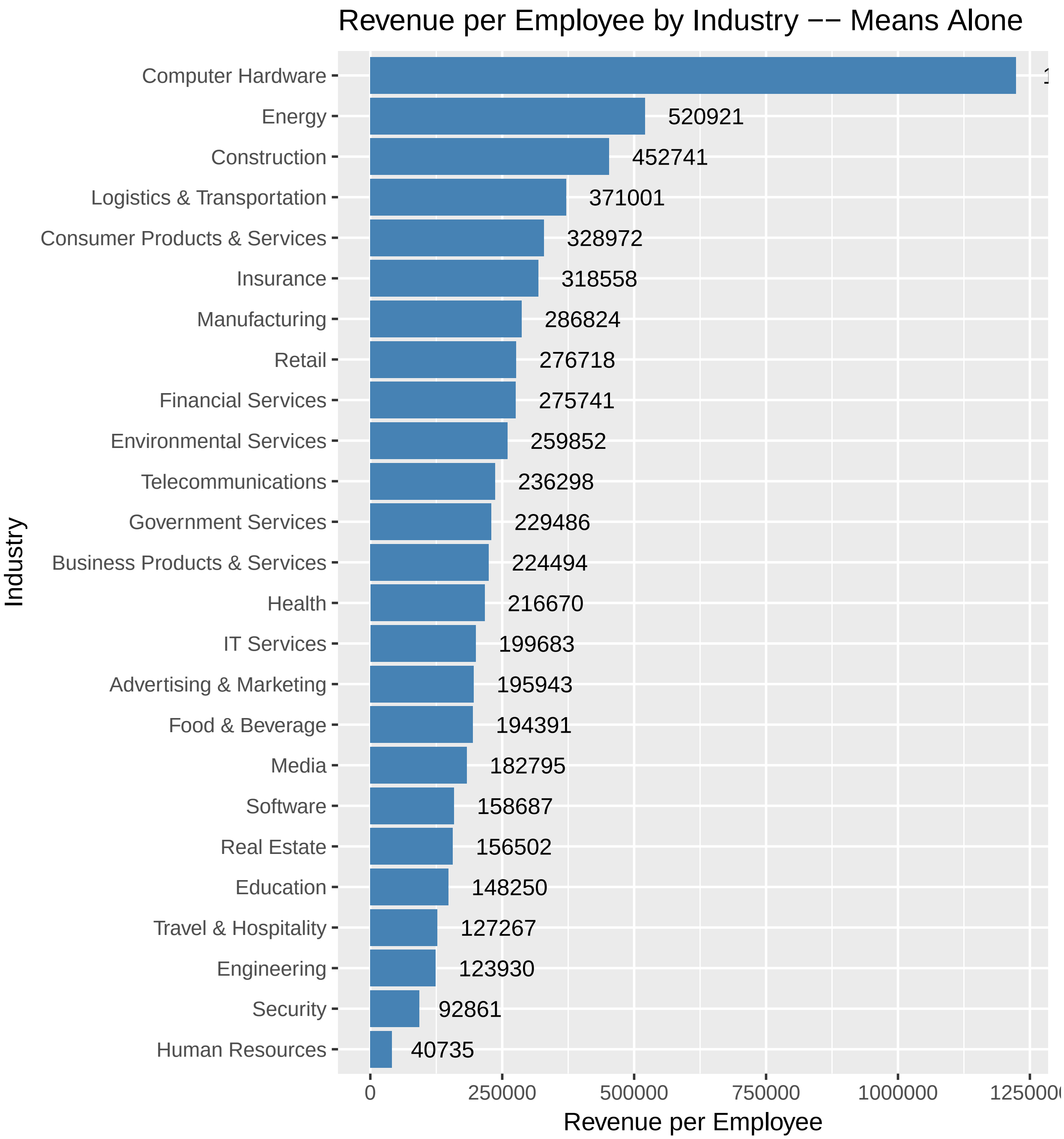
```
ggplot(data = state_3_df, aes(x = reorder(Industry, Employees),
    y = Employees)) + geom_boxplot(outlier.size = NA) + coord_flip(ylim = ylim1) +
    labs(title = paste0("Distribution of Employees by Industry for ",
        state_3, " State"), x = "Industry", y = "Employees")
```



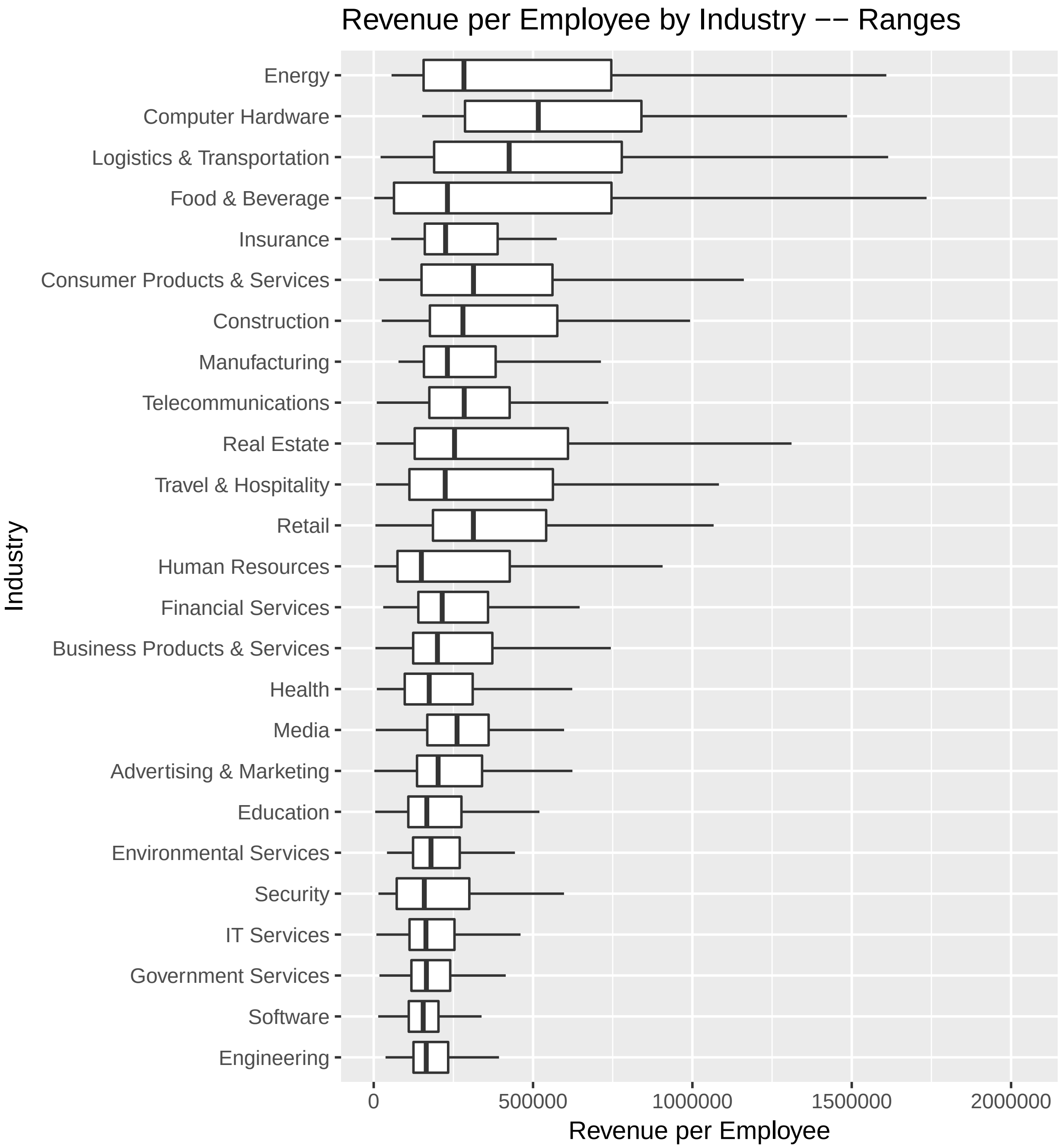
3. Now imagine you work for an investor and want to see which industries generate the most revenue per employee. Create a chart makes this information clear.

```
c_data = data[complete.cases(data), ]
c_data_gs = ddply(c_data, .(Industry), summarize, Average_Income = mean(Revenue/Employees))
c_data_gs1 = ddply(c_data, .(Industry), summarize,
    Average_Income = sum(Revenue)/sum(Employees))
```

```r
ylim2 = c(0, boxplot.stats(c_data_gs$Average_Income)$out[1])


ggplot(data = c_data_gs1, aes(x = reorder(Industry,
    Average_Income), y = round(Average_Income))) +
    coord_flip() + geom_bar(stat = "identity", fill = "steelblue") +
    geom_text(aes(label = round(Average_Income)), hjust = -0.3,
        color = "black", size = 3.5) + labs(title = "Revenue per Employee by Industry -- Means Alone",
    x = "Industry", y = "Revenue per Employee")
```



```r
ggplot(data = c_data, aes(x = reorder(Industry, Revenue/Employees),
    y = Revenue/Employees)) + geom_boxplot(outlier.size = NA) +
    coord_flip(ylim = ylim2 * 2.5) + labs(title = paste0("Revenue per Employee by Industry -- Ranges"),
    x = "Industry", y = "Revenue per Employee")
```


Revenue per Employee by Industry -- Ranges
Industry
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
Revenue per Employee