# 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
##  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

# There are numerous R functions that provide descriptive & exploratory statistics, such as functions i

library(Hmisc)
```

```
## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

## 
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
## 
##     format.pval, units
```

```
describe(inc)
```

```
## inc 
## 
##  8  Variables      5001  Observations
## --------------------------------------------------------------------------------
## Rank 
##        n  missing distinct     Info     Mean      Gmd      .05      .10 
##     5001        0     4999        1     2502     1667      252      502 
##      .25      .50      .75      .90      .95 
##     1252     2502     3751     4501     4751 
## 
## lowest :    1    2    3    4    5, highest: 4996 4997 4998 4999 5000
## --------------------------------------------------------------------------------
## Name 
##        n  missing distinct 
##     5001        0     5001 
## 
```

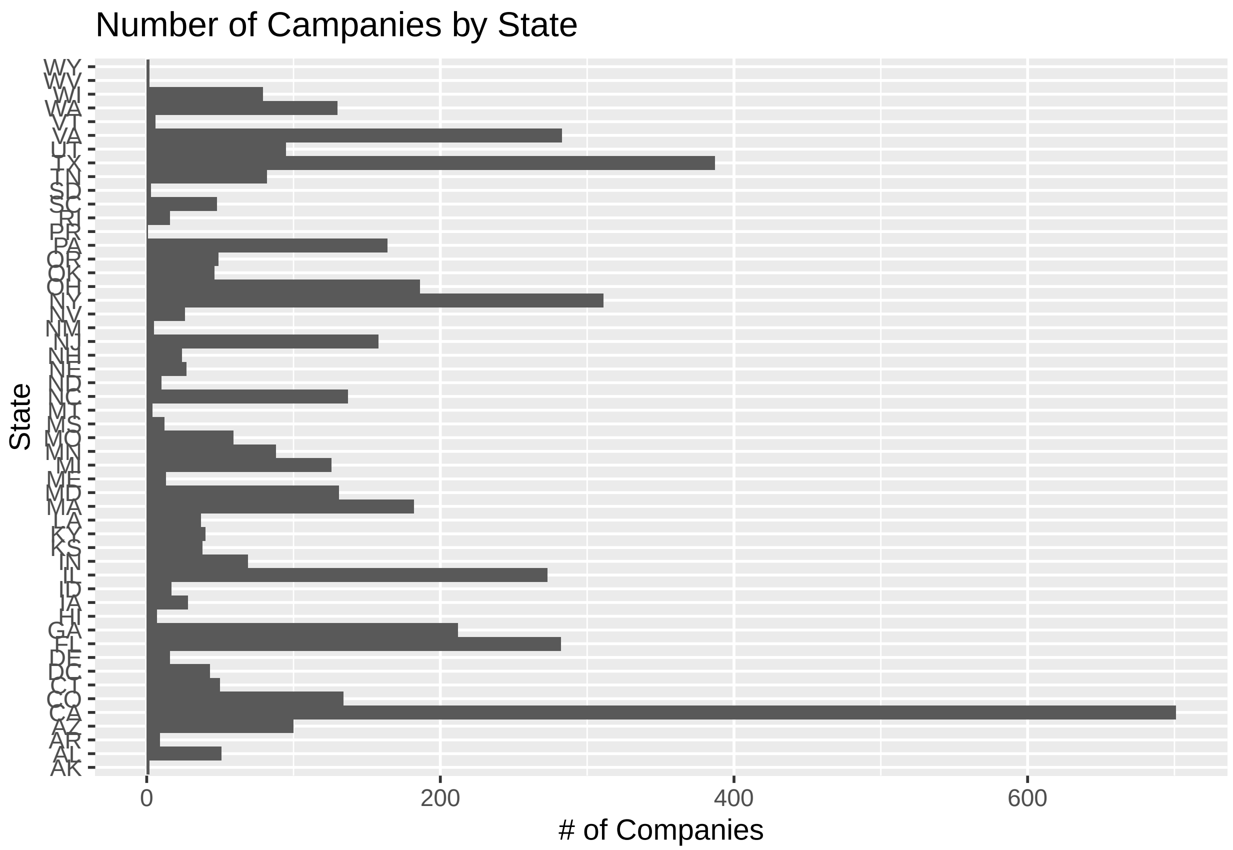
```
## lowest : (Add)ventures                          @Properties                            1-Stop Transla
## highest: Zoup!                                  ZT Wealth and Altus Group of Companies Zumasys
## --------------------------------------------------------------------------------
## Growth_Rate
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     5001        0     1147        1    4.612    6.493     0.43     0.50
##      .25      .50      .75      .90      .95
##     0.77     1.42     3.29     9.12    17.16
##
## lowest :   0.34   0.35   0.36   0.37   0.38, highest: 213.37 233.08 245.45 248.31 421.48
## --------------------------------------------------------------------------------
## Revenue
##          n   missing  distinct      Info      Mean       Gmd       .05       .10
##       5001         0      1069         1  48222535  75111227   2400000   3000000
##        .25       .50       .75       .90       .95
##    5100000  10900000  28600000  76900000 155600000
##
## lowest : 2.00e+06 2.10e+06 2.20e+06 2.30e+06 2.40e+06
## highest: 3.80e+09 4.50e+09 4.60e+09 4.70e+09 1.01e+10
## --------------------------------------------------------------------------------
## Industry
##        n  missing distinct
##     5001        0       25
##
## lowest : Advertising & Marketing      Business Products & Services Computer Hardware            Cons
## highest: Retail                       Security                     Software                     Tele
## --------------------------------------------------------------------------------
## Employees
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     4989       12      691        1    232.7    365.6     10.0     14.0
##      .25      .50      .75      .90      .95
##     25.0     53.0    132.0    351.2    688.0
##
## lowest :     1     2     3     4     5, highest: 17057 18887 20000 32000 66803
## --------------------------------------------------------------------------------
## City
##        n  missing distinct
##     5001        0     1519
##
## lowest : Acton        Addison      Adrian       Agoura Hills Aiea
## highest: Worthington  Wyomissing   Yonkers      Youngsville  Zumbrota
## --------------------------------------------------------------------------------
## State
##        n  missing distinct
##     5001        0       52
##
## lowest : AK AL AR AZ CA, highest: VT WA WI WV WY
## --------------------------------------------------------------------------------
```

### 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
# To represent the distribution of companies by state, I used histogram to represent the number of coun
library(ggplot2)
ggplot(inc, aes(x=State)) + geom_histogram(stat = "count", width = 1) +
    ggtitle("Number of Campanies by State") +
    xlab("State") + ylab("# of Companies") +
    coord_flip()
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



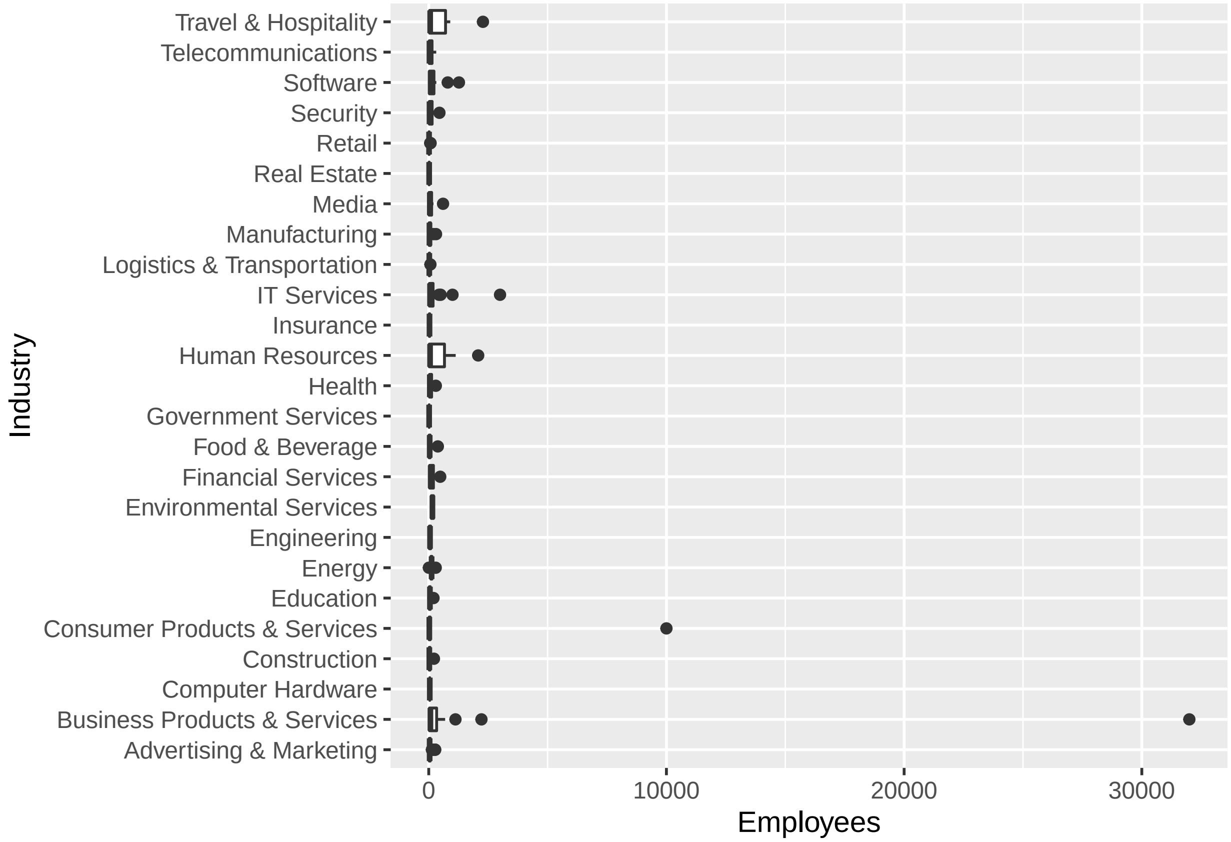
## Quesiton 2

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.

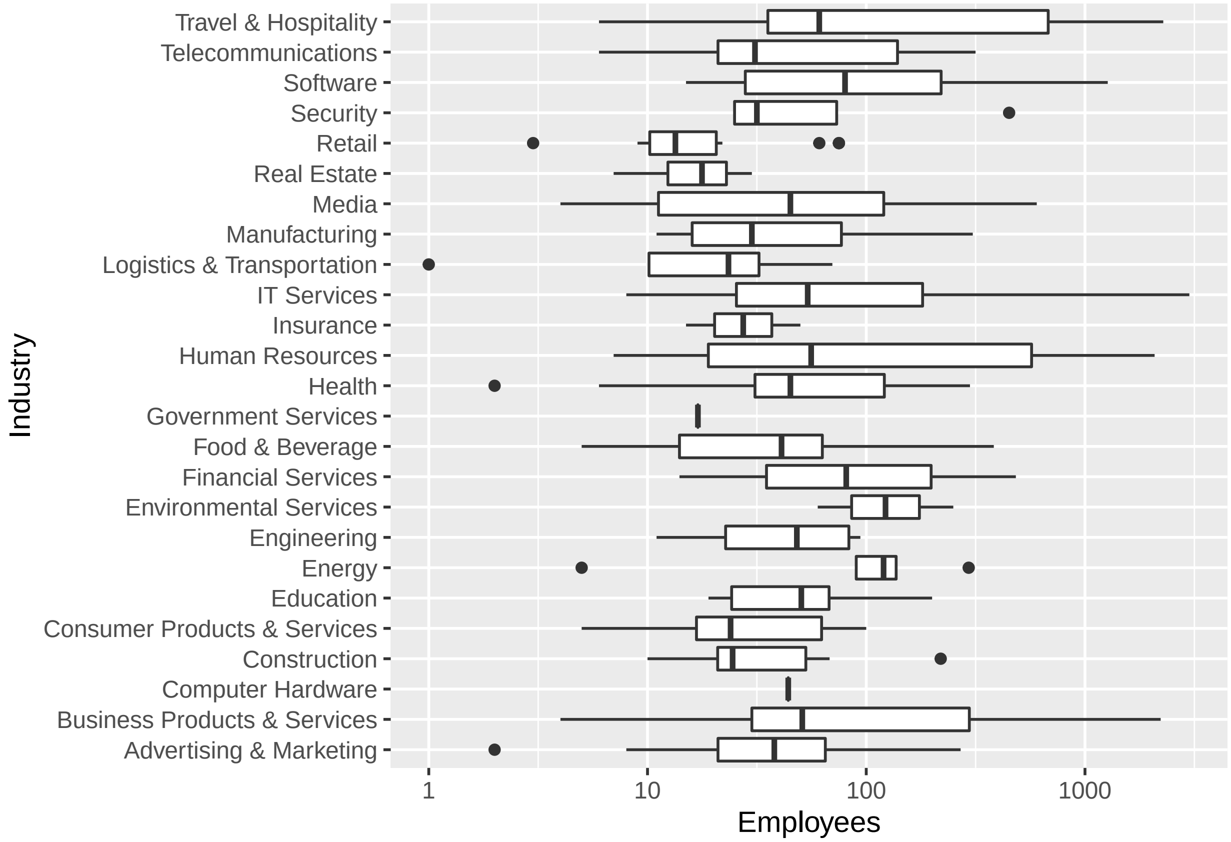
```
# Answer Question 2 here

library(dplyr)
```

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

## The following objects are masked from 'package:Hmisc':
##
##     src, summarize

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

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

```
# From the summary() function in the previous section, we can see that the state with the thrid most co
# Apply complete.cases() function.

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

# Prepare dataset.
NY <- filter(inc_complete, State == "NY")
#head(NY)

# Create box plot WITH outliers for initial data exploration.
chart_initial <- ggplot(NY, aes(Industry, Employees)) + geom_boxplot() + coord_flip()
chart_initial
```



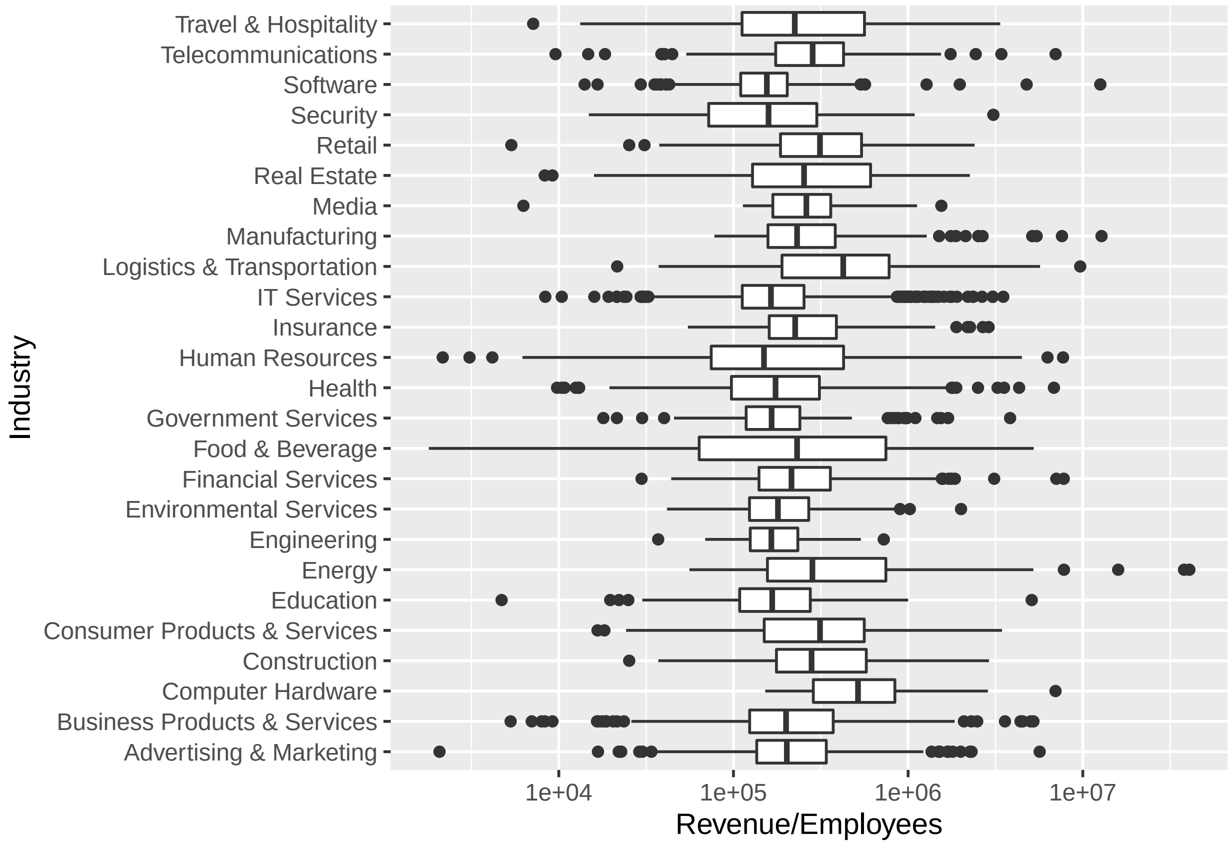
```
# Find and remove outliers - remove the two biggest outliers based on the original plot.

new_NY <- subset(NY, Employees < 10000)

# Apply log transformation to further normalize the data.
chart_log_transformed <- ggplot(new_NY, aes(Industry, Employees)) +
    geom_boxplot() +
    scale_y_log10() +
    coord_flip()

chart_log_transformed
```



### Question 3

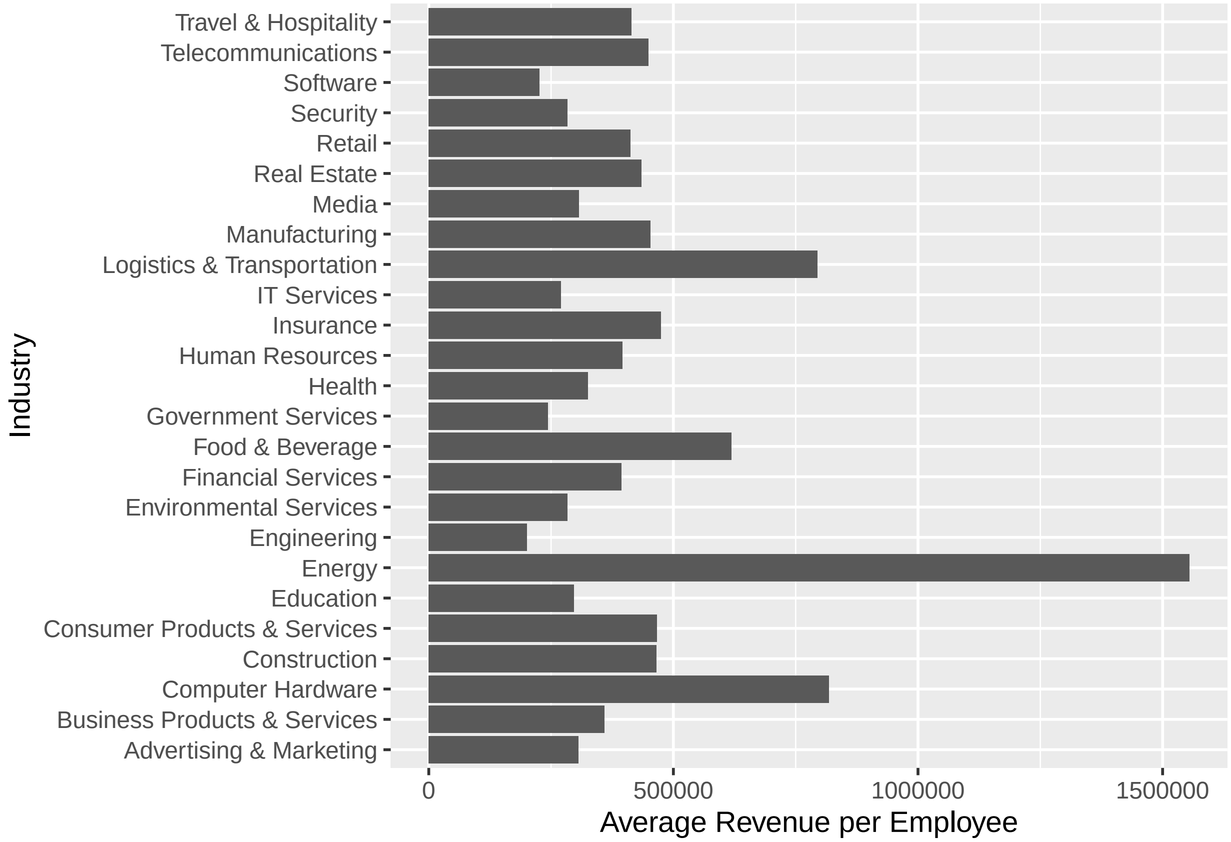
Now imagine you work for an investor and want to see which industries generate the most revenue per employee. Create a chart that makes this information clear. Once again, the distribution per industry should be shown.

```
# Answer Question 3 here
# Plot boxplot.
ggplot(inc_complete, aes(Industry, Revenue/Employees)) +
    geom_boxplot() +
    scale_y_log10() +
    coord_flip()
```



```
# The boxplot looks chaotic and does not show any trend without extensive data cleaning. I decided to u

# Add new column to dataframe.
RevenuePerEmployee = inc_complete$Revenue / inc_complete$Employees
inc_complete <- cbind(inc_complete, RevenuePerEmployee)

# Create bar charts for revenue per employee by industry.
ggplot(inc_complete, aes(x=Industry, y=RevenuePerEmployee)) +
    stat_summary(fun.y="mean", geom="bar")+
    xlab("Industry") + ylab("Average Revenue per Employee") + coord_flip()
```

```
## Warning: `fun.y` is deprecated. Use `fun` instead.
```



```
# The bar chart shows a much clearer trend, with the Energy industry generating the most revenue per em
```

```
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
```