homework 1 608

Sam Reeves

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(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
```

```
library(cowplot)
library(stats)

library(reticulate)
use_python('/usr/bin/python3')

file <- 'inc5000_data.csv'
inc <- tibble(read.csv(file, header= TRUE))

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

head(inc)
```

```
## # A tibble: 6 x 8
##    Rank Name                  Growth_Rate Revenue Industry Employees City  State
##   <int> <fct>                       <dbl>   <dbl> <fct>        <int> <fct> <fct>
## 1     1 Fuhu                         421.  1.18e8 Consume~       104 El S~ CA
## 2     2 FederalConference.com        248.  4.96e7 Governm~        51 Dumf~ VA
## 3     3 The HCI Group                245.  2.55e7 Health         132 Jack~ FL
## 4     4 Bridger                      233.  1.9 e9 Energy          50 Addi~ TX
## 5     5 DataXu                       213.  8.7 e7 Adverti~       220 Bost~ MA
## 6     6 MileStone Community ~        179.  4.57e7 Real Es~        63 Aust~ 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   :2501   110 Consulting        :   1   Mean   :  4.615
##  3rd Qu.:3750   11thStreetCoffee.com  :   1   3rd Qu.:  3.290
##  Max.   :5000   123 Exteriors         :   1   Max.   :421.480
##                 (Other)               :4983
##     Revenue                                  Industry      Employees
##  Min.   :2.000e+06   IT Services                  : 732   Min.   :    1.0
##  1st Qu.:5.100e+06   Business Products & Services: 480   1st Qu.:   25.0
##  Median :1.090e+07   Advertising & Marketing     : 471   Median :   53.0
##  Mean   :4.825e+07   Health                       : 354   Mean   :  232.7
##  3rd Qu.:2.860e+07   Software                     : 341   3rd Qu.:  132.0
##  Max.   :1.010e+10   Financial Services           : 260   Max.   :66803.0
##                      (Other)                      :2351
##             City          State
##  New York     : 160   CA     : 700
##  Chicago      :  90   TX     : 386
##  Austin       :  88   NY     : 311
##  Houston      :  76   VA     : 283
##  San Francisco:  74   FL     : 282
##  Atlanta      :  73   IL     : 272
##  (Other)      :4428   (Other):2755
```

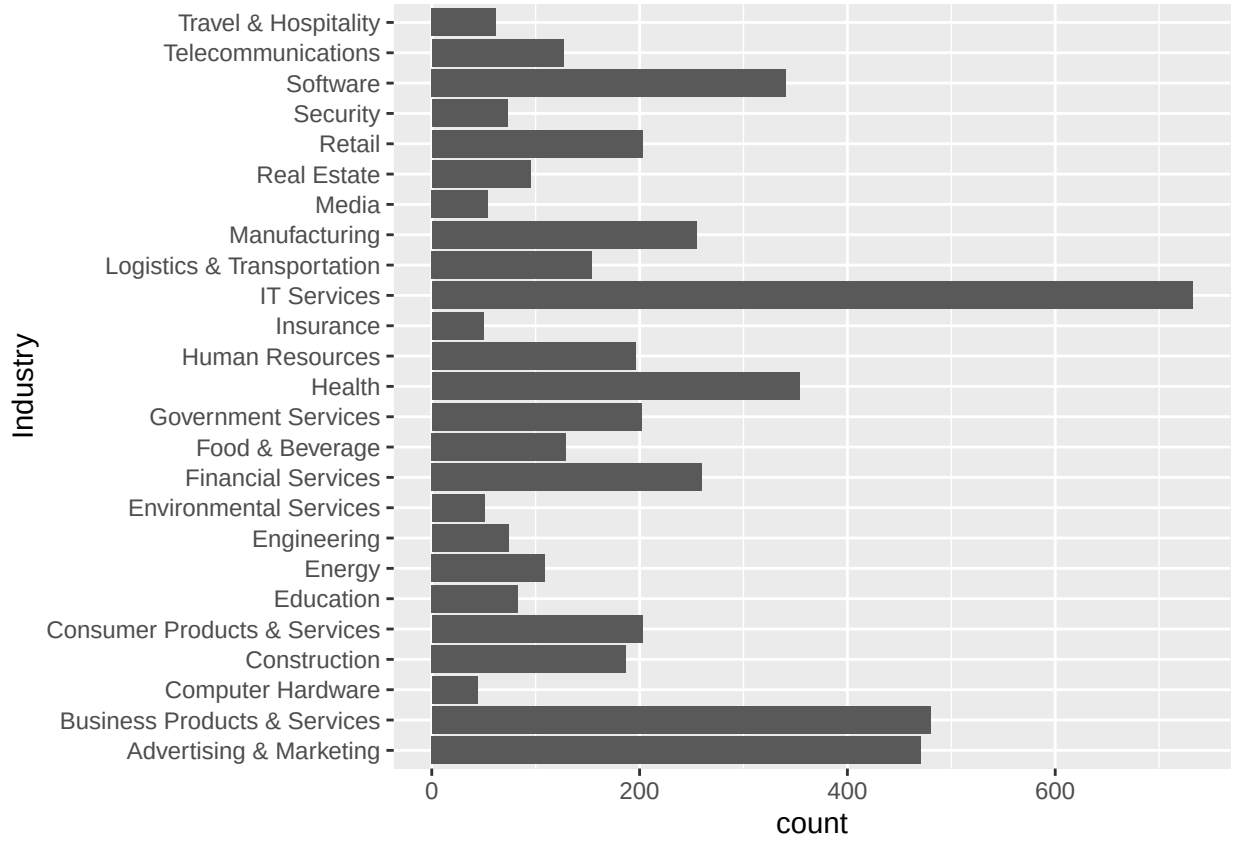
Think a bit on what these summaries mean. Use the space below to add some more relevant non-visual exploratory information you think helps you understand this data:

The mean of the Growth_Rate is higher than the 3rd quartile. There are some extreme growth outliers, and probably a few of them. Revenue goes from $2M to $10B. It's a very wide range with the median about 1/3 of the mean. The outliers in this group are probably the same ones from the Growth_Rate.

By far the highest growth industry is IT services, followed by Business Products & Services and Advertising and Marketing. These two combined are slightly higher than IT Services, and they are roughly equal.

Although CA has the highest number of growth companies from the list (strangely similar to the number of IT Services company), four non-California cities rank in the top places. These spots are held by New York, Chicago, Austin, and Houston. Austin and Houston account for nearly half the companies in Texas, the second highest-growth state.

```
ggplot(data = inc) +
  geom_bar(mapping = aes(y = Industry))
```

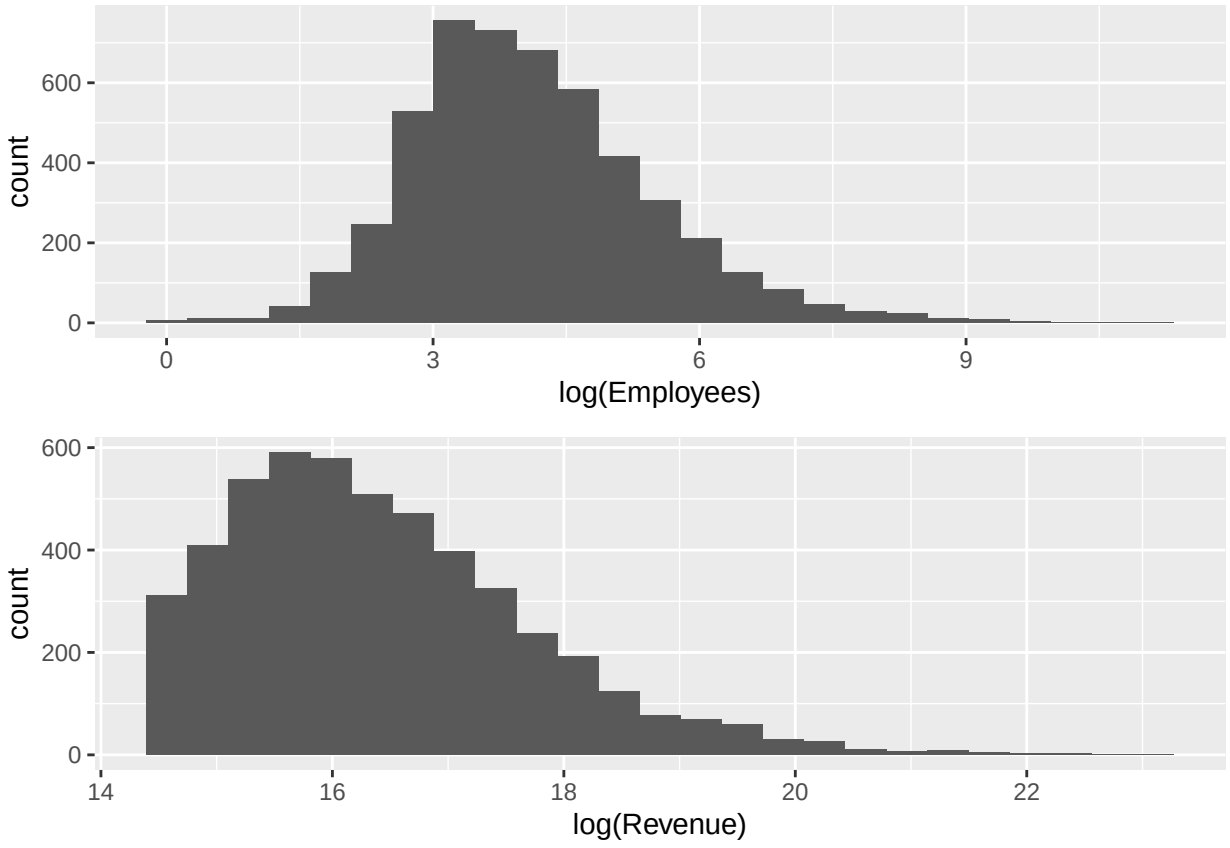


IT Services, Business Products, Advertising, Software and Health are leading the pack.

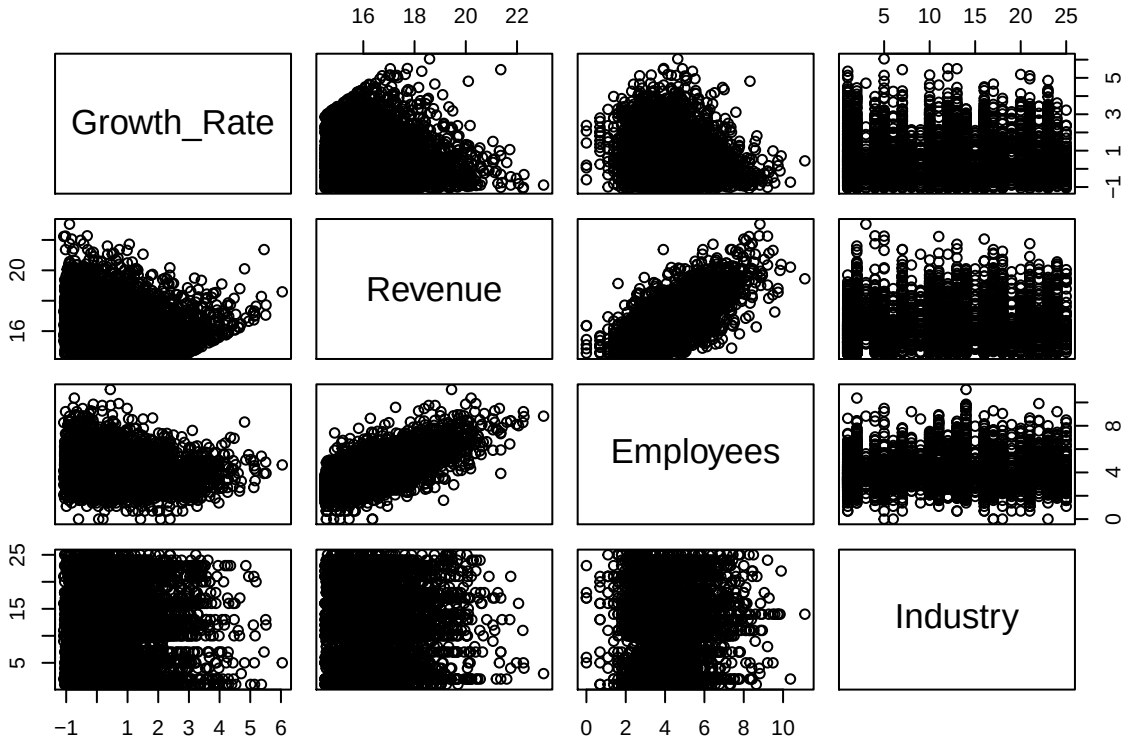
```
p1 <- ggplot(data = inc) +
  geom_histogram(mapping = aes(x = log(Employees)), bins = 25)

p2 <- ggplot(data = inc) +
  geom_histogram(mapping = aes(x = log(Revenue)), bins = 25)

plot_grid(p1, p2, ncol = 1)
```



```
p3 <- sapply(inc[,c(3,4,6)], log) %>% cbind(inc[,5]) %>%
  pairs()
```



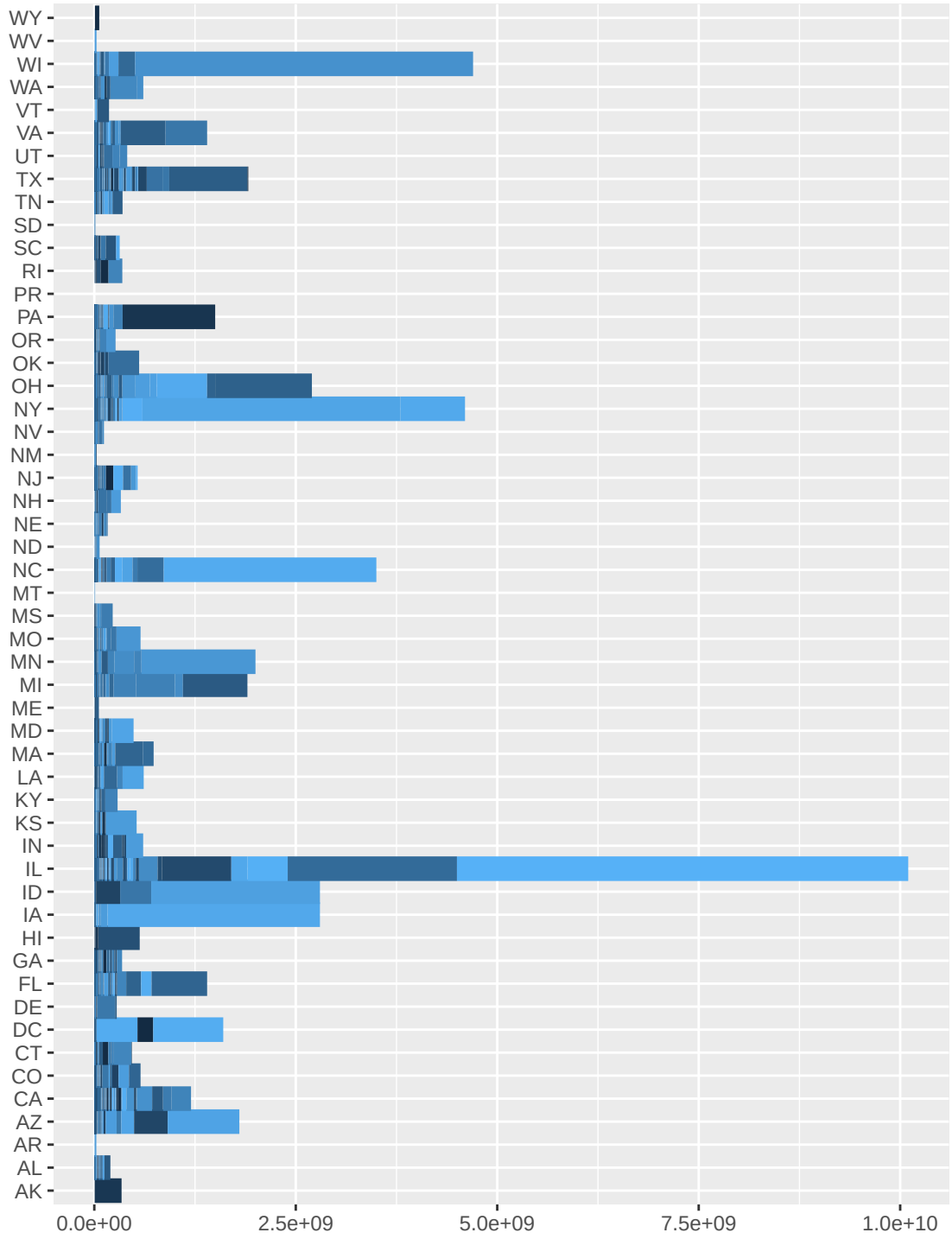
Revenue and number of employees seem strongly correlated, and there is a visible linear limit for revenue against growth rate.

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.

```
inc %>%
  ggplot(aes(x = Revenue, y = State, col = Rank)) +
    geom_line(size = 5) +
    theme(legend.position = 'none') +
    labs(x = '', y = '') +
    ggtitle("Each bar represents a company, revenue by color")
```

Each bar represents a company, revenue by color



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.

```
summary(inc$State) %>% sort() %>% tail()
```

```
##  IL  FL  VA  NY  TX  CA
## 272 282 283 311 386 700
```

```
inc <- inc %>%
  mutate(loc = paste0(City, ', ', State))

ny <- inc %>% filter(State == 'NY')
```

The third greatest number of companies on the growth list is 311 in New York. This includes Washington DC and Puerto Rico.

```
import pandas
import geopandas
import geopy

def locate(place) :
    locator = geopy.Nominatim(user_agent = 'robocop')
    location = locator.geocode([place])
    if location == None:
        coord = (0,0)
    else:
        coord = (location.latitude, location.longitude)
    return coord

def locate_all(data):
    la = list()
    lo = list()
    for i in range(len(data)):
        city = data['loc'][i]
        if city == 'New York, NY':
            coord = (40.7127281, -74.0060152)
        else:
            coord = locate(city)
        la.append(coord[0])
        lo.append(coord[1])
    return(la, lo)
```

```
la, lo = locate_all(r.ny)
```

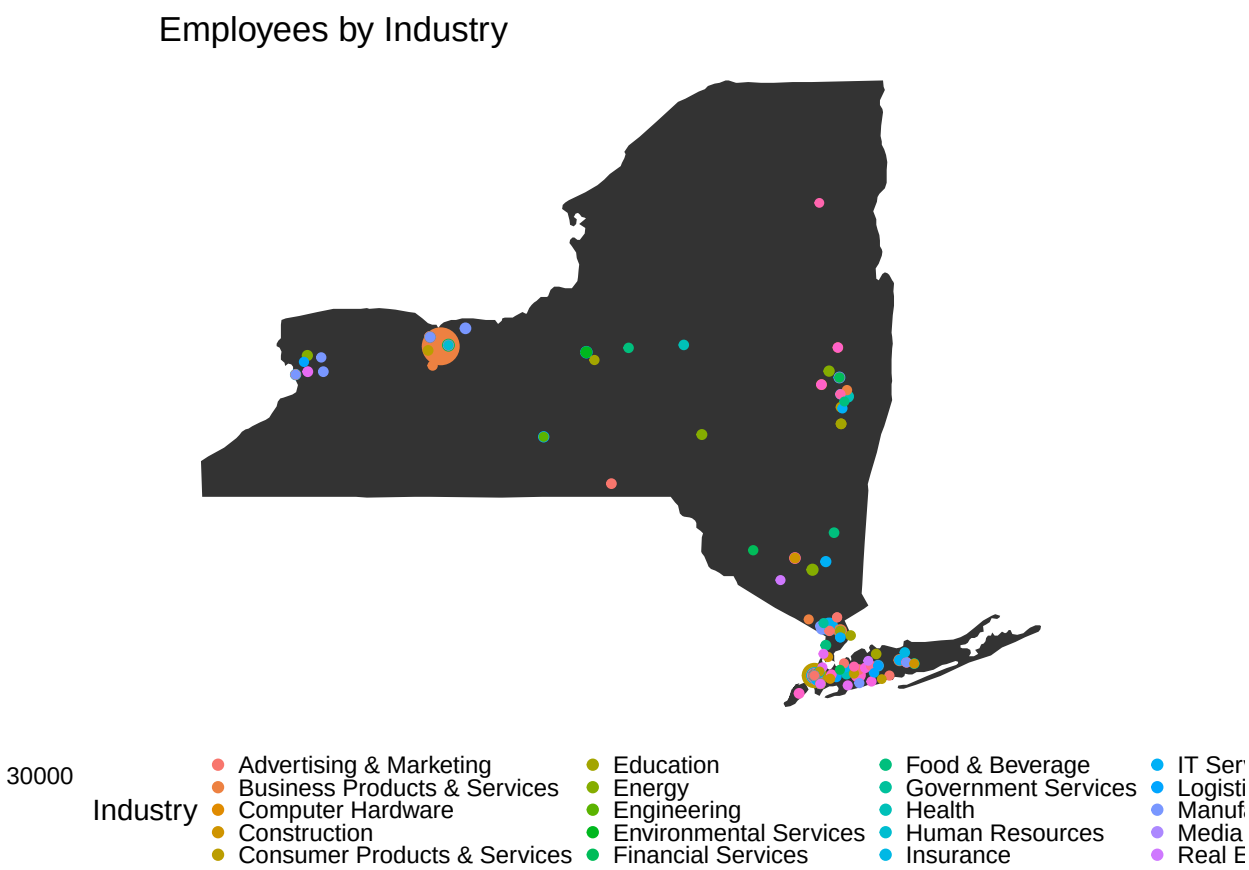
```
map <- map_data("state", region = 'New York')
ny <- cbind(ny, la = py$la, lo = py$lo)
```

```
summary(ny)
```

```
##       Rank                           Name      Growth_Rate        Revenue
##  Min.   :  26   1st Equity          :  1   Min.   : 0.350   Min.   :2.000e+06
##  1st Qu.:1186   33Across            :  1   1st Qu.: 0.670   1st Qu.:4.300e+06
##  Median :2702   5Linx Enterprises   :  1   Median : 1.310   Median :8.800e+06
##  Mean   :2612   Access Display Group:  1   Mean   : 4.371   Mean   :5.872e+07
##  3rd Qu.:4005   Adafruit            :  1   3rd Qu.: 3.580   3rd Qu.:2.570e+07
##  Max.   :4981   AdCorp Media Group  :  1   Max.   :84.430   Max.   :4.600e+09
##                 (Other)             :305
##                            Industry     Employees             City
##  Advertising & Marketing     : 57   Min.   :    1.0   New York :160
##  IT Services                 : 43   1st Qu.:   21.0   Brooklyn : 15
##  Business Products & Services: 26   Median :   45.0   Rochester:  9
##  Consumer Products & Services: 17   Mean   :  271.3   Buffalo  :  5
##  Telecommunications          : 17   3rd Qu.:  105.5   Fairport :  5
##  Education                   : 14   Max.   :32000.0   new york :  5
##  (Other)                     :137                     (Other)  :112
##      State         loc                 la              lo
##  NY     :311   Length:311         Min.   :40.58   Min.   :-78.878
##  AK     :  0   Class :character   1st Qu.:40.71   1st Qu.:-74.006
##  AL     :  0   Mode  :character   Median :40.71   Median :-74.006
##  AR     :  0                      Mean   :41.23   Mean   :-74.110
##  AZ     :  0                      3rd Qu.:41.03   3rd Qu.:-73.950
##  CA     :  0                      Max.   :60.38   Max.   :  5.334
##  (Other):  0
```

```
ny <- subset(ny, lo < 0)
ny <- ny[order(-ny$Employees), ]
```

```
ggplot() +
  geom_polygon(data = map,
               aes(x=long, y=lat, group = group)) +
  coord_fixed(1.3) +
  geom_point(data = ny,
             aes(x = lo, y = la,
                 color = Industry,
                 size = Employees)) +
  theme_void() +
  theme(legend.position = 'bottom', legend.key.size = unit(0.001, 'cm')) +
  ggtitle('Employees by Industry')
```


Employees by Industry
30000
Industry
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 Ser
Logisti
Manuf
Media
Real E

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.

```
inc <- inc %>%
  mutate(Efficiency = Revenue / Employees)

ggplot(inc) +
  geom_line(aes(y = Industry, x = log(Efficiency),
                col = log(Efficiency), size = 5)) +
  theme(legend.position = 'none') +
  labs(x = '', y = '') +
  ggtitle('Employee efficiency per Industry')
```