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:

To start, I imported ggplot2. This data visualization library is very comprehensive and useful for anybody who is plotting data in R.

```
# Insert your code here, create more chunks as necessary

library(ggplot2)
```

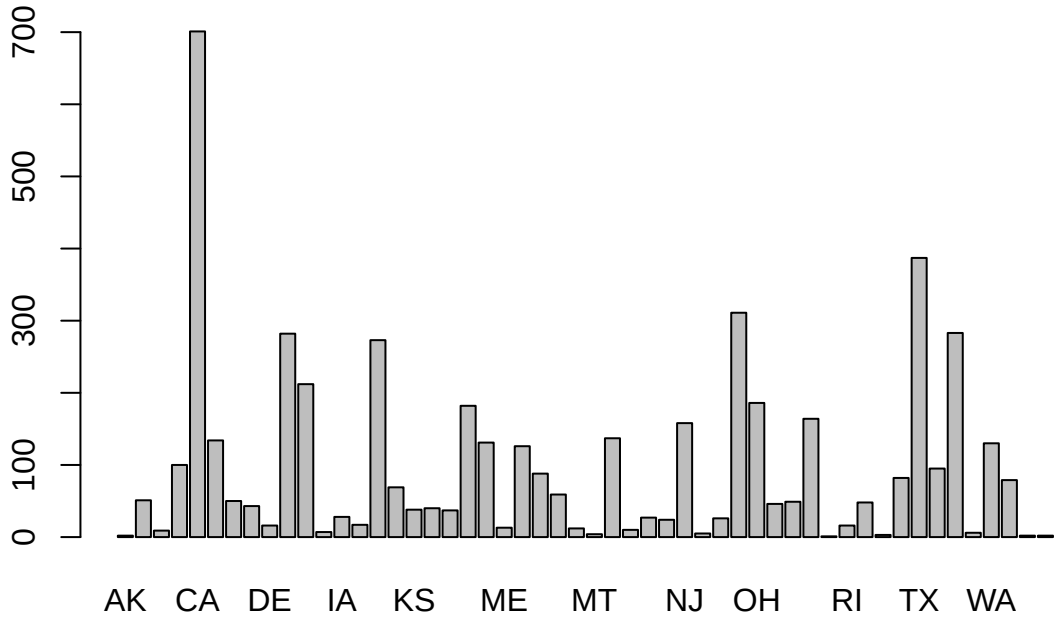
Question 1

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.

Visualization Attempts

For Question 1, I threw in a simple plot just to see what the data looks like. Already, I can see a lot of problems. Given that there are fifty states, they do not all show up on the x-axis. I can make the x-axis wider to accomodate for each state, but the portrait limitation would make it very difficult to see the small text labels. Another problem is that the distribution of the number of companies varies significantly. There are some states like Alaska and West Virginia with 2 companies and states like California with over 700 companies. Given the wide range of data, it will be difficult to visually distinguish the smaller counts from each other; a bar with 2 companies looks very similar to a bar with 3 companies.

```
plot(inc$State)
```



Initally, my first thought upon seeing this data was to use a map to display the data. Maps are very intuitive for people who are not familiar with data and are more visually interesting than a bar graph. However, a map of the United States would need to be displayed horizontally, which will not work for this particular problem. Also, small states like New Jersey, Rhode Island, and Maryland will be very hard to see on a small screen.

Given the accessibility problems with a map, there are a few things we can do to improve the original bar graph. To make the states more visible, we can flip the x and y axis so that states names will appear vertically. Given that we have more vertical space than horizontal space, it will be much easier to see the state names. The next problem is the distribution of the company counts by state. This problem was a little trickier to address and required a few additions to the graph. The first thing I did was sort the states by the number of companies instead of sorting by alphabetical order. Visually, it is much easier to distinguish the states with more companies from the states with fewer companies. I then added a color gradient that shifts from a darker blue to a lighter blue to further distinguish the states with differing numbers of companies.

While this layout was already significantly better than the original plot, I ran into another problem. Given that the graph is vertically long, it is not possible to see the x-axis markers and all the states without compressing the image. I decided to add the company count to the end of the bars to show exactly how many companies each state had, without having to look at the x-axis.

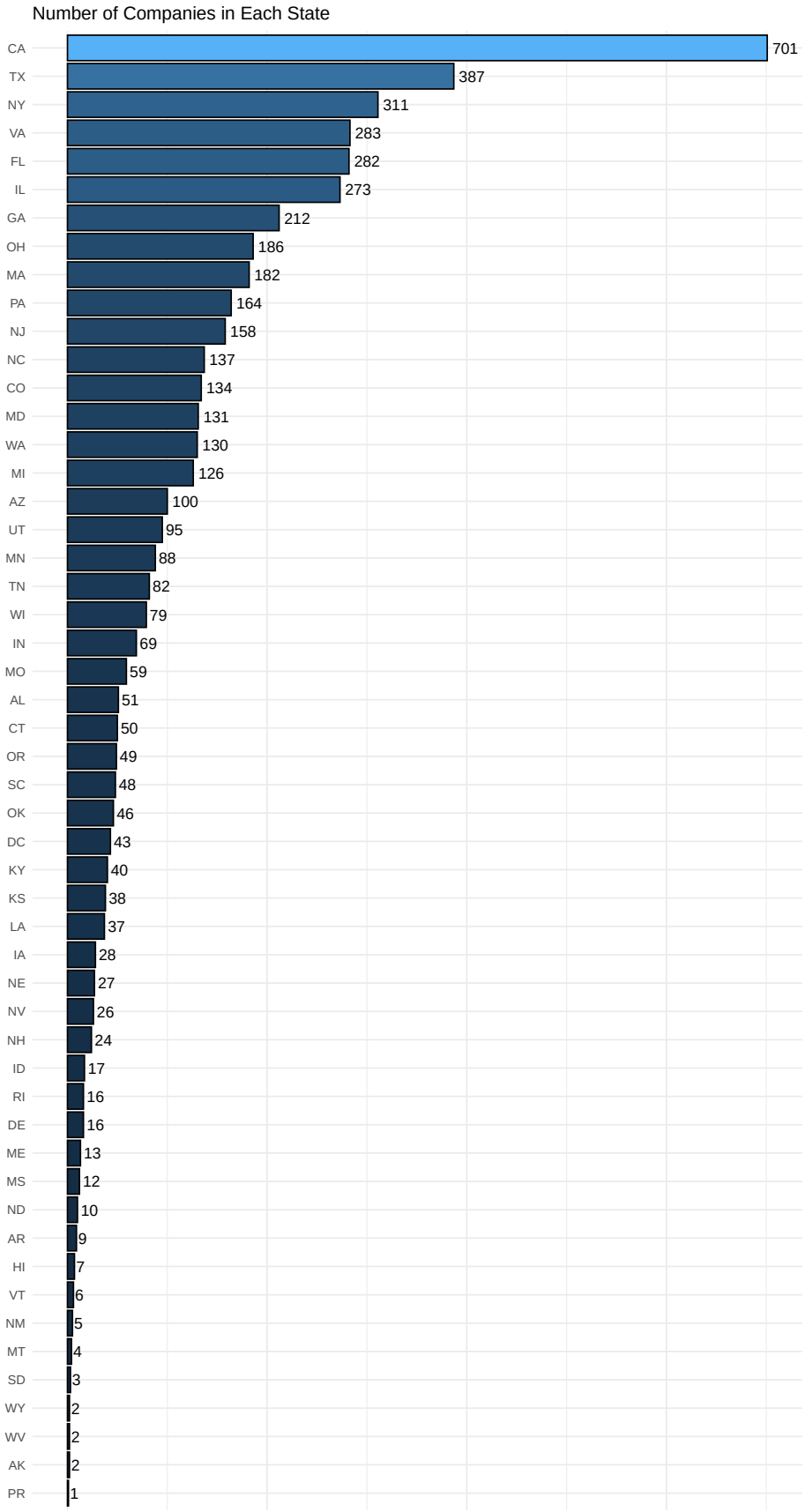
Question 1 Answer

The resulting plot is shown below:

```
# Answer Question 1 here

ggplot(inc, aes(x = factor(State, levels = names(sort(table(State),decreasing= F))))) + #sorts state co
```

```
  geom_bar(aes(fill = ..count..), colour = "black") + #creates bar graph colored by number of companies
  coord_flip() + #flips x and y coordinates to make graph longer
  geom_text(stat = 'count', aes(label = ..count..), hjust = -0.2) + #adds number label to each bar
  ggtitle("Number of Companies in Each State") + #adds title
  guides(fill = F) + #removes legend
  theme_minimal() + #minimal theme for less intrusive visual elements
  theme(axis.title.x = element_blank(), #remove x-axis label
        axis.text.x = element_blank(), #remove x-axis text
        axis.ticks.x = element_blank(), #remove x-axis tick marks
        axis.title.y = element_blank()) #remove y-axis label
```


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

Question 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.

Data Prep

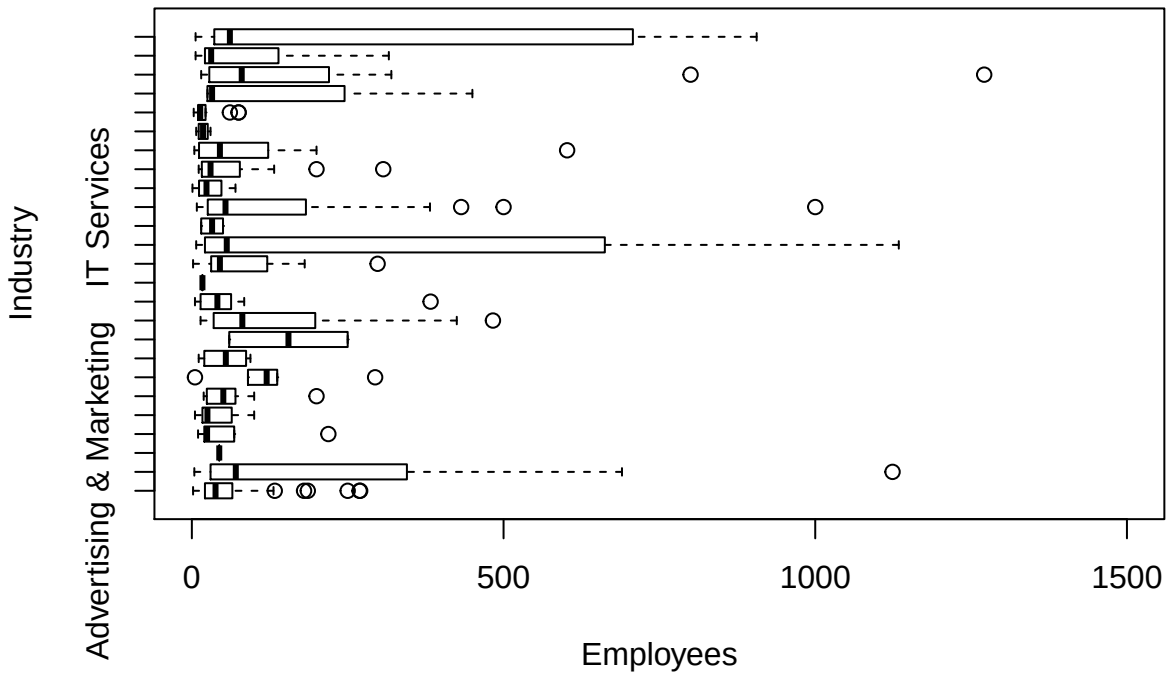
According to the graph from question 1, the state with the 3rd most companies is New York. In the code below, I used the `complete.cases` function to select the rows with complete cases. Afterwards, I selected the rows from New York as my new dataset.

```
inc_complete = inc[complete.cases(inc),]
new_york = inc_complete[inc_complete$State == "NY",]
```

Visualization Attempts

The first type of visualization I tried was the boxplot. The advantage of a boxplot is that it clearly shows the median as well as the spread of the data. Unfortunately, for this particular set of data, the ranges vary significantly by industry and it is hard to see each individual median. Also, the axes are not very informative.

```
boxplot(Employees~Industry, data = new_york, horizontal = T, ylim = c(0,1500))
```



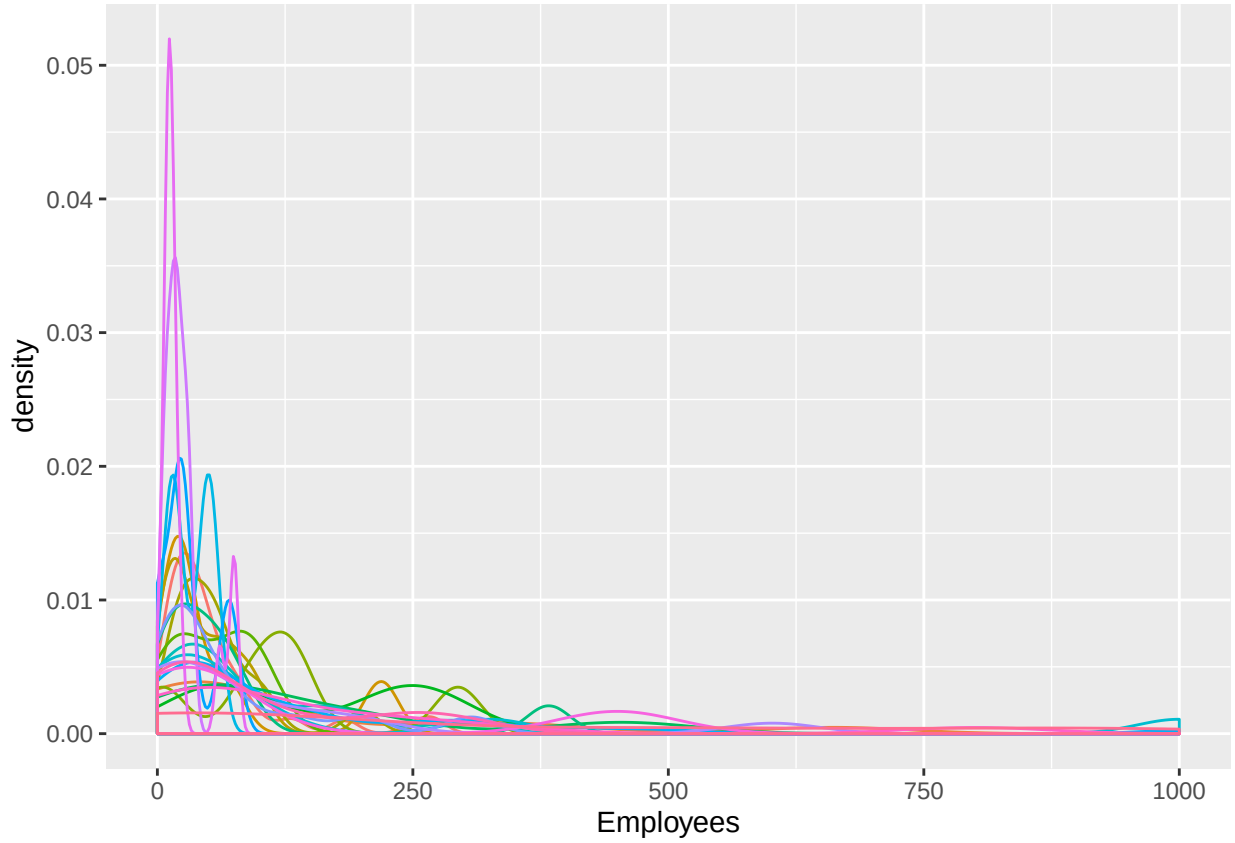
My second attempt was to make a stacked density plot. Stacked density plots are nice because it allows the user to compare multiple distributions directly. However, when there are more than 3 distribution, stacked density plots become more difficult to distinguish from each other. Finding the median for each individual industry is even more difficult with this graph. It is possible to improve this graph by adding some interactable elements that lets the user select specific distributions to display, however, as a static image, the stacked density plot is not a good way to visualize this data.

```
ggplot(new_york, aes(x = Employees, color = Industry)) +
  geom_density() +
  #facet_wrap(~Industry) +
  theme(legend.position = "none") +
  xlim(0,1000)
```

```
## Warning: Removed 9 rows containing non-finite values (stat_density).
```

```
## Warning: Groups with fewer than two data points have been dropped.
```

```
## Warning: Groups with fewer than two data points have been dropped.
```



Question 2 Answer

The final plot is a violin plot displaying the median explicitly for each industry. I lengthened the graph vertically to accomodate for the industry text labels and the violin plot width. Compared to the boxplot, violin plots show the distribution of the data more clearly and can accomodate for larger variations in the

data. Though, the one downside of the violin plot is that outliers can stretch the data into a long thin strip, which can make it difficult to see the distribution. However, this is much better than boxplots, which display outliers as individual points. When there are thousands of outliers,

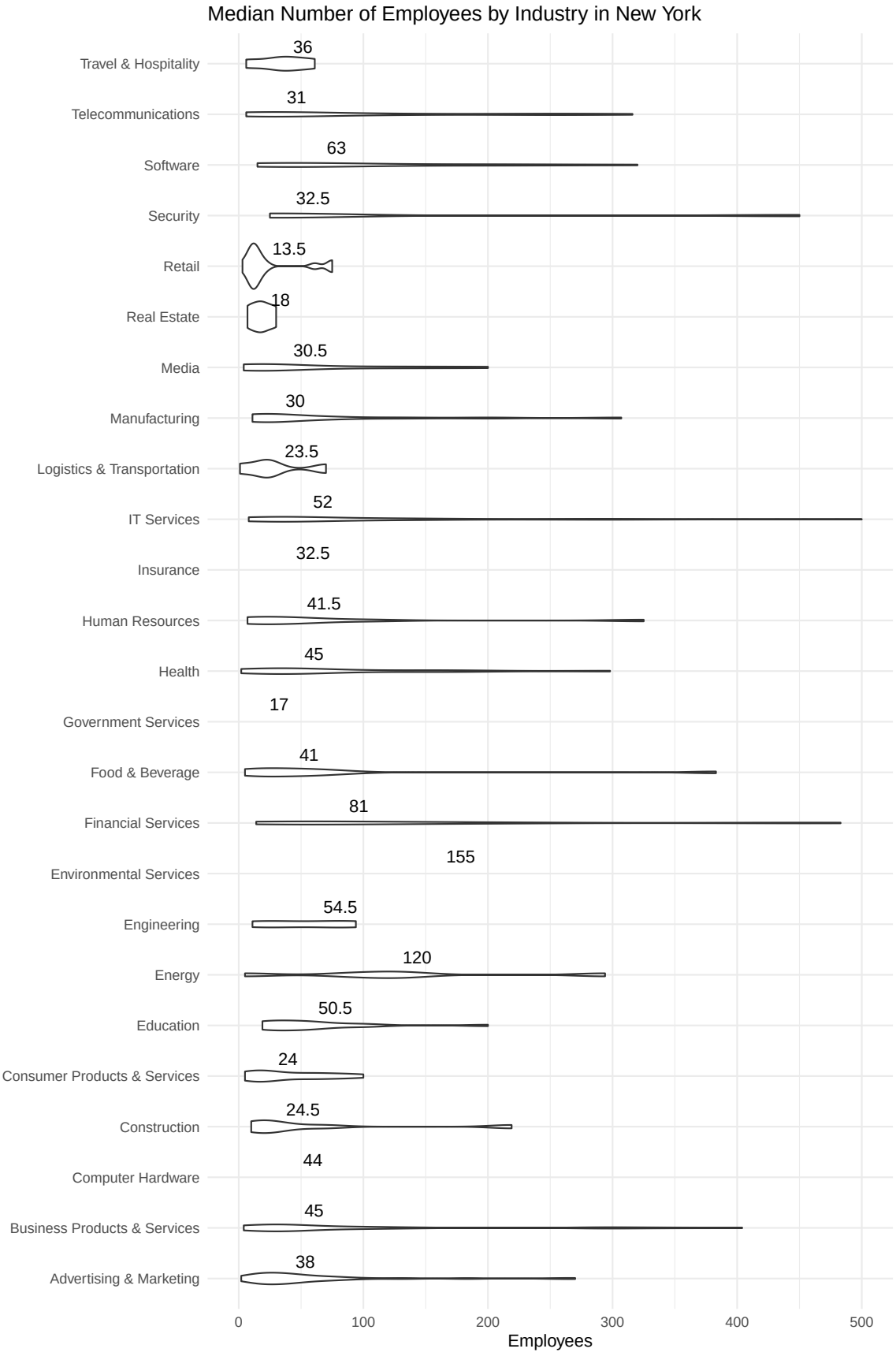
```
# Answer Question 2 here

fun_median = function(x){ #function to show the median for each industry
  return(data.frame(y=median(x),label=median(x,na.rm=T)))
}


ggplot(new_york, aes(x = Industry, y = Employees)) + #adds data to violin plot
  geom_violin(trim = T) + #creates violin plot
  coord_flip() + #flips x and y axes
  ylim(0,500) + #limits x-axis to 500
  stat_summary(fun.data = fun_median, geom="text", vjust = -0.9, hjust = -0.5) + #adds median labels to
  theme_minimal() + #minimal theme to remove extraneous visual elements
  ggtitle("Median Number of Employees by Industry in New York") + #adds title
  theme(axis.title.y = element_blank()) #remove y-axis label
```

```
## Warning: Removed 17 rows containing non-finite values (stat_ydensity).
```

```
## Warning: Removed 17 rows containing non-finite values (stat_summary).
```


Median Number of Employees by Industry in New York
Travel & Hospitality
36
Telecommunications
31
Software
63
Security
32.5
Retail
13.5
Real Estate
18
Media
30.5
Manufacturing
30
Logistics & Transportation
23.5
IT Services
52
Insurance
32.5
Human Resources
41.5
Health
45
Government Services
17
Food & Beverage
41
Financial Services
81
Environmental Services
155
Engineering
54.5
Energy
120
Education
50.5
Consumer Products & Services
24
Construction
24.5
Computer Hardware
44
Business Products & Services
45
Advertising & Marketing
38
0
100
200
300
400
500
Employees

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.

Data Prep

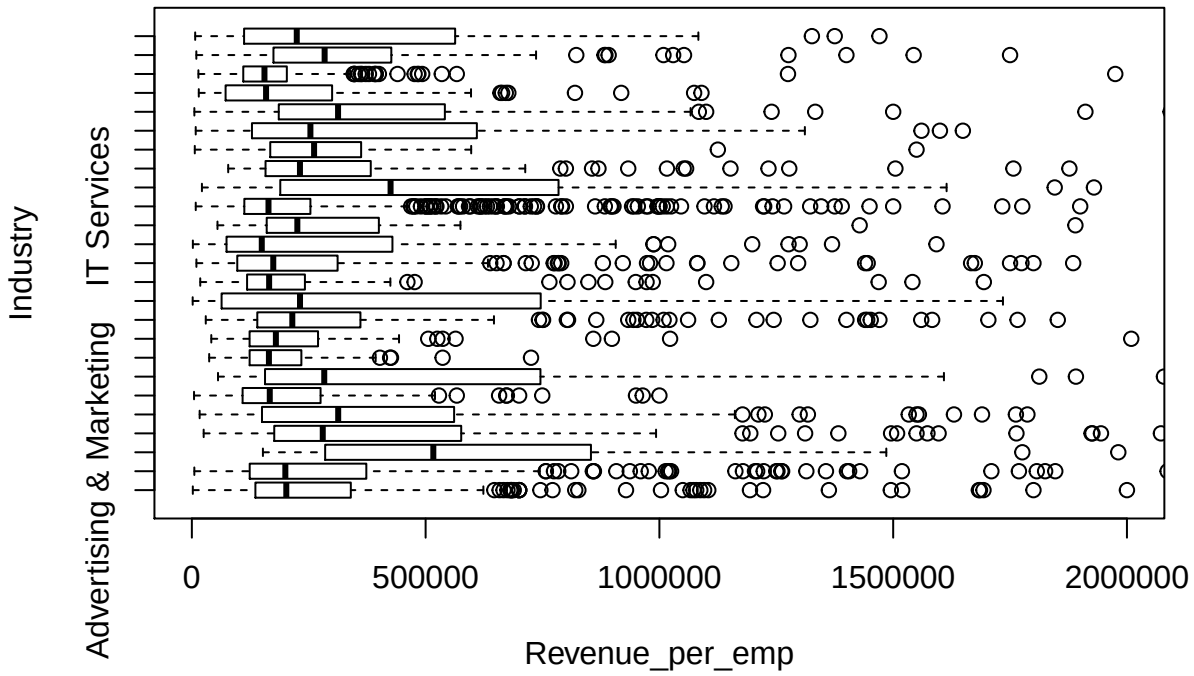
I added a variable to the complete data that takes the revenue of each company and divides it by the number of employees.

```
inc_complete$Revenue_per_emp = inc_complete$Revenue/inc_complete$Employees
```

Visualization Attempts

Like the previous problem, I wanted to see this data as a boxplot first. There were immediately problems. In addition to the problems specified in the previous question, this data has the issue of too many outliers. This boxplot is very cluttered and it is nearly impossible to tell which industry has the most revenue per employee.

```
boxplot(Revenue_per_emp~Industry, data = inc_complete, horizontal = T, ylim = c(0,2000000))
```



Question 3 Solution

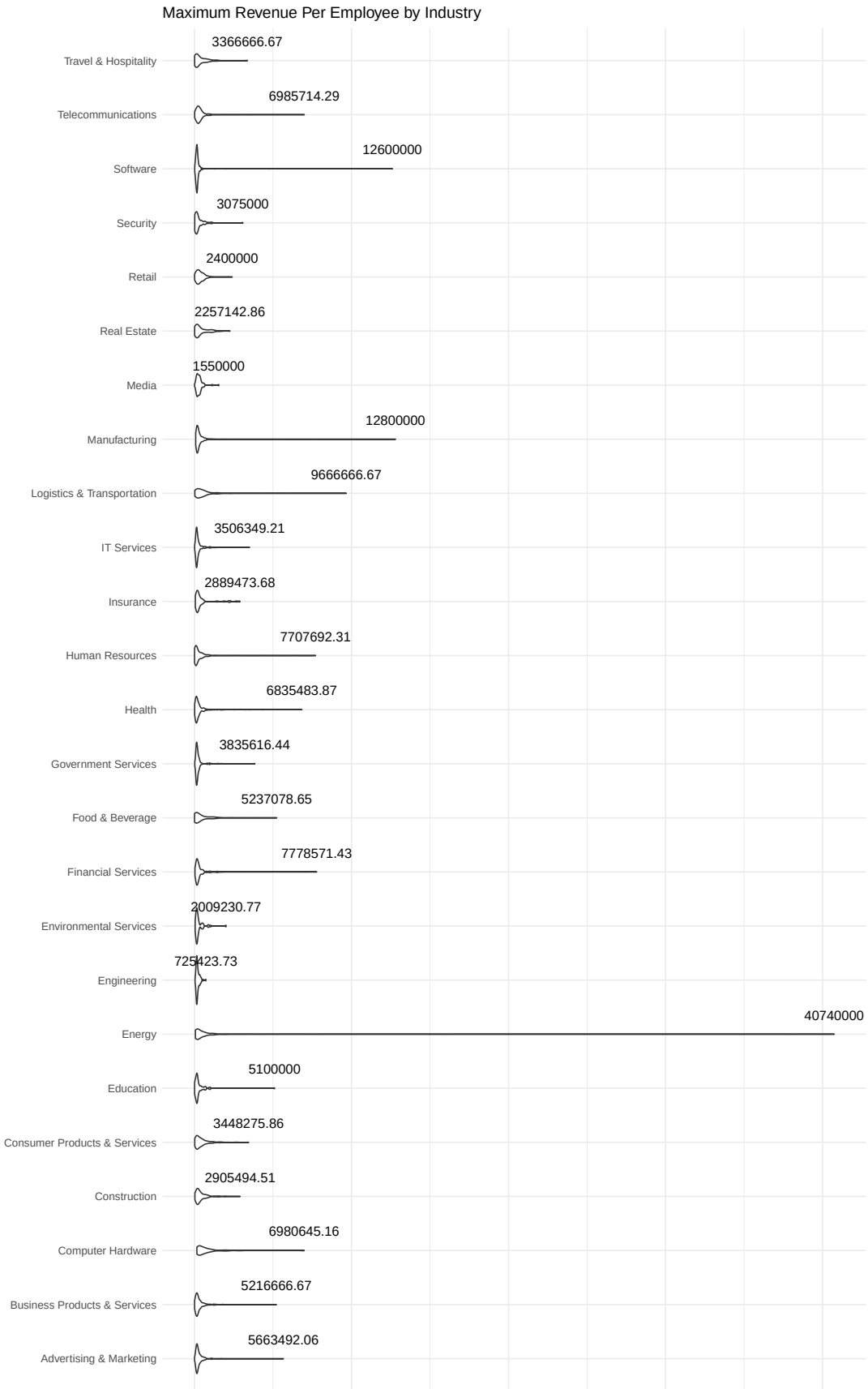
Violin plot again; hear me out. For the same reasons mentioned before, violin plots show the distribution much more clearly comapred to boxplots and are much more resistant to outliers. In the violin plot below, we can see the distribution of revenue per employee for each industry along with the maximum revenue per employee in each industry. In this case, the violin plot combines both the boxplot with the bar chart into a plot that shows the max values and distribution at the same time.

```
# Answer Question 3 here

fun_max = function(x){
  return(data.frame(y=round(max(x),2),label=round(max(x,na.rm=T),2))) #function to display max value by
}

options(scipen = 1000) #removes scientific notation

ggplot(inc_complete, aes(x = Industry, y = Revenue_per_emp)) + #adds data
  geom_violin(trim = T) + #creates violin plot
  coord_flip() + #flips x and y axes
  theme_minimal() + #minimal theme removes extraneous visual elements
  ggtitle("Maximum Revenue Per Employee by Industry") + #add title
  theme(axis.title.y = element_blank(),
        axis.title.x = element_blank(),
        axis.ticks.x = element_blank(),
        axis.text.x = element_blank()) + #remove axis labels
  stat_summary(fun.data = fun_max, geom="text", vjust = -1.5, hjust = 0.5) #adds max value label
```


Maximum Revenue Per Employee by Industry
Travel & Hospitality
3366666.67
Telecommunications
6985714.29
Software
12600000
Security
3075000
Retail
2400000
Real Estate
2257142.86
Media
1550000
Manufacturing
12800000
Logistics & Transportation
9666666.67
IT Services
3506349.21
Insurance
2889473.68
Human Resources
7707692.31
Health
6835483.87
Government Services
3835616.44
Food & Beverage
5237078.65
Financial Services
7778571.43
Environmental Services
2009230.77
Engineering
725423.73
Energy
40740000
Education
5100000
Consumer Products & Services
3448275.86
Construction
2905494.51
Computer Hardware
6980645.16
Business Products & Services
5216666.67
Advertising & Marketing
5663492.06