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
```

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:

Some additional interesting information we can get from this data set is the number of unique values in certain columns.

For example, the number of unique values in the "Industry" column is:

```
nrow(unique(inc["Industry"]))
```

```
## [1] 25
```

And the number of unique values in the "State" column is:

```
nrow(unique(inc["State"]))
```

```
## [1] 52
```

Interestingly there are 52 unique values for the State column which tells us that most likely every state is being respresented and that there are two additional values that likely represent Washington DC and Puerto Rico.

Another interesting stat we can get from this data non-visually is checking the number of NA values in each column. This can be done by using the `colSums(is.na(inc))` function.

```
colSums(is.na(inc))
```

```
##        Rank        Name Growth_Rate     Revenue    Industry   Employees
##           0           0           0           0           0          12
##        City       State
##           0           0
```

Here we see that only the "Employees" column has NA values. This tells us we likely will not need to do much data cleaning for this data set.

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.

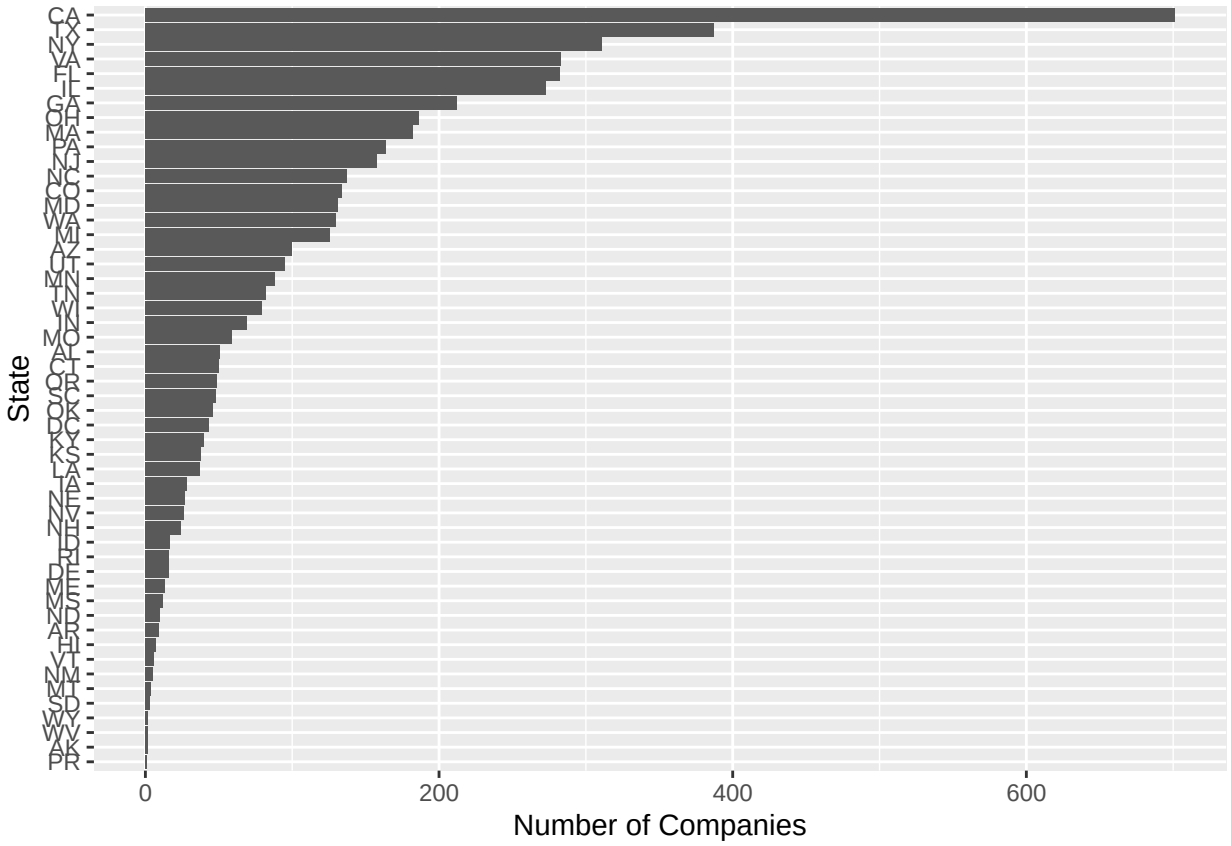
```
# Answer Question 1 here
library(dplyr)
```

```
##
## 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
```

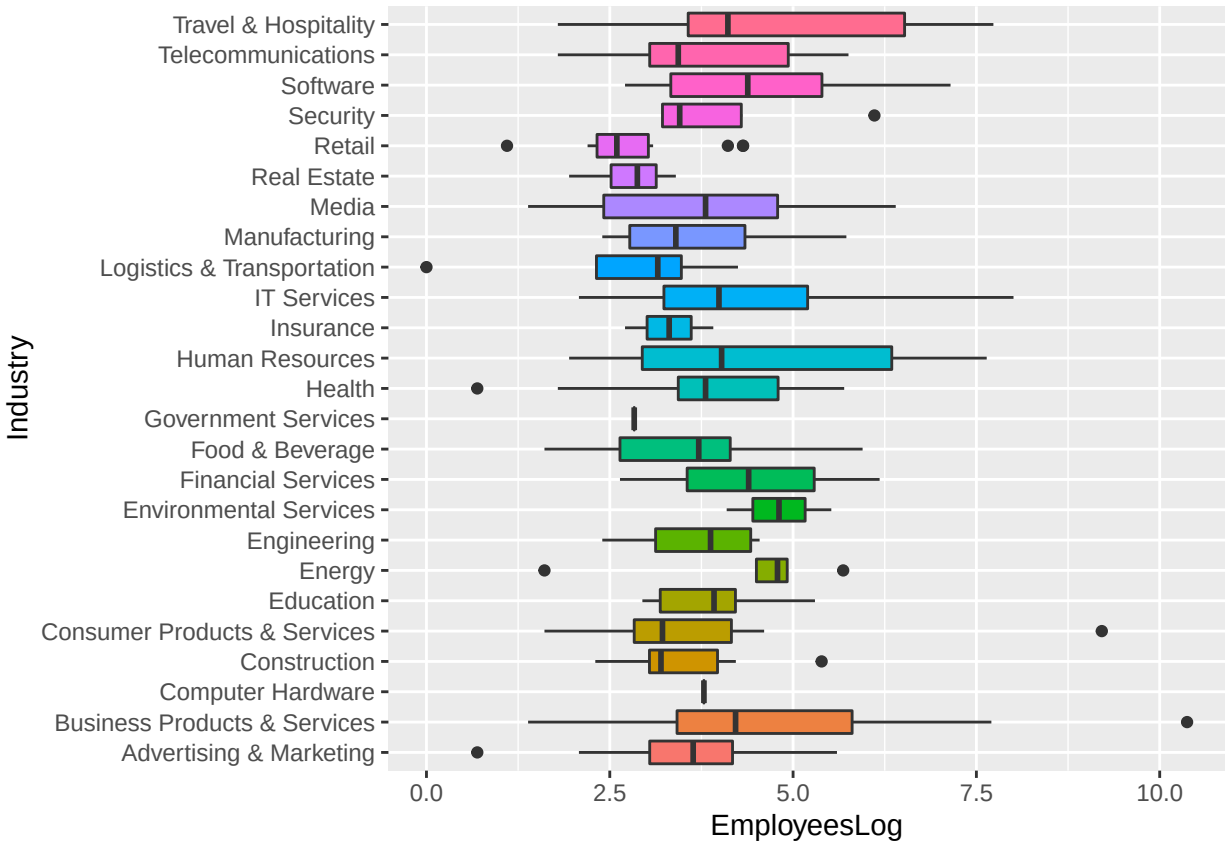
```
library(ggplot2)

# Answer Question 1 here
inc %>%
  count(State, n(), sort = T) %>%
  ggplot(aes(x = reorder(State, n), y = n)) +
  geom_col() +
  coord_flip() +
  labs(x = "State", y = "Number of Companies")
```



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.

```
incNy <- inc %>% filter(complete.cases(.) & State == 'NY')
incNy$EmployeesLog <- log(incNy$Employees)
# plot a violin plot
ggplot(incNy, aes(x = EmployeesLog, y = Industry, fill = Industry)) +
  geom_boxplot() +
  theme(legend.position = "none")
```

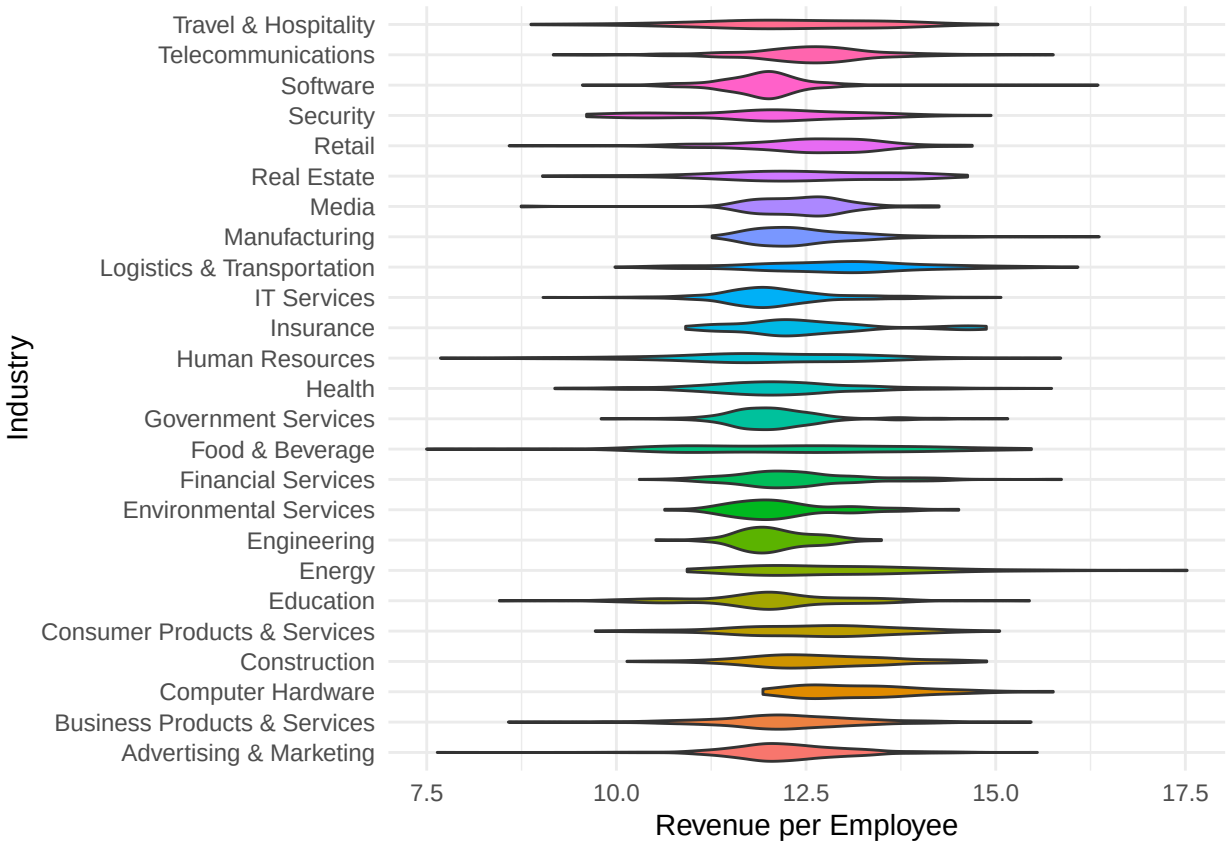


Question 3

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
ggplot(inc, aes(x = Industry, y = log(Revenue / Employees), fill = Industry)) +
  geom_violin() +
  scale_y_continuous(name = "Revenue per Employee") +
  theme_minimal() +
  coord_flip() +
  theme(legend.position = "none")
```

```
## Warning: Removed 12 rows containing non-finite values (stat_ydensity).
```


Industry
Travel & Hospitality
Telecommunications
Software
Security
Retail
Real Estate
Media
Manufacturing
Logistics & Transportation
IT Services
Insurance
Human Resources
Health
Government Services
Food & Beverage
Financial Services
Environmental Services
Engineering
Energy
Education
Consumer Products & Services
Construction
Computer Hardware
Business Products & Services
Advertising & Marketing
7.5
10.0
12.5
15.0
17.5
Revenue per Employee