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:

```
library(tidyverse)  # I would like to use dplyr's glimpse function, as well as other things from tidyve
glimpse(inc)
```

```
## Rows: 5,001
## Columns: 8
## $ Rank        <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,~
## $ Name        <chr> "Fuhu", "FederalConference.com", "The HCI Group", "Bridger~
## $ Growth_Rate <dbl> 421.48, 248.31, 245.45, 233.08, 213.37, 179.38, 174.04, 17~
## $ Revenue     <dbl> 1.179e+08, 4.960e+07, 2.550e+07, 1.900e+09, 8.700e+07, 4.5~
## $ Industry    <chr> "Consumer Products & Services", "Government Services", "He~
## $ Employees   <int> 104, 51, 132, 50, 220, 63, 27, 75, 97, 15, 149, 165, 250, ~
## $ City        <chr> "El Segundo", "Dumfries", "Jacksonville", "Addison", "Bost~
## $ State       <chr> "CA", "VA", "FL", "TX", "MA", "TX", "TN", "CA", "UT", "RI"~
```

I see that we have more detailed knowledge of the types of numerical columns. For example, `Rank` and `Employees` are of the type integer. `Growth_Rate` and `Revenue` are of the type double.

I want to see the proportion of missing values in each column.

```
(colSums(is.na(inc)) / nrow(inc)) * 100
```

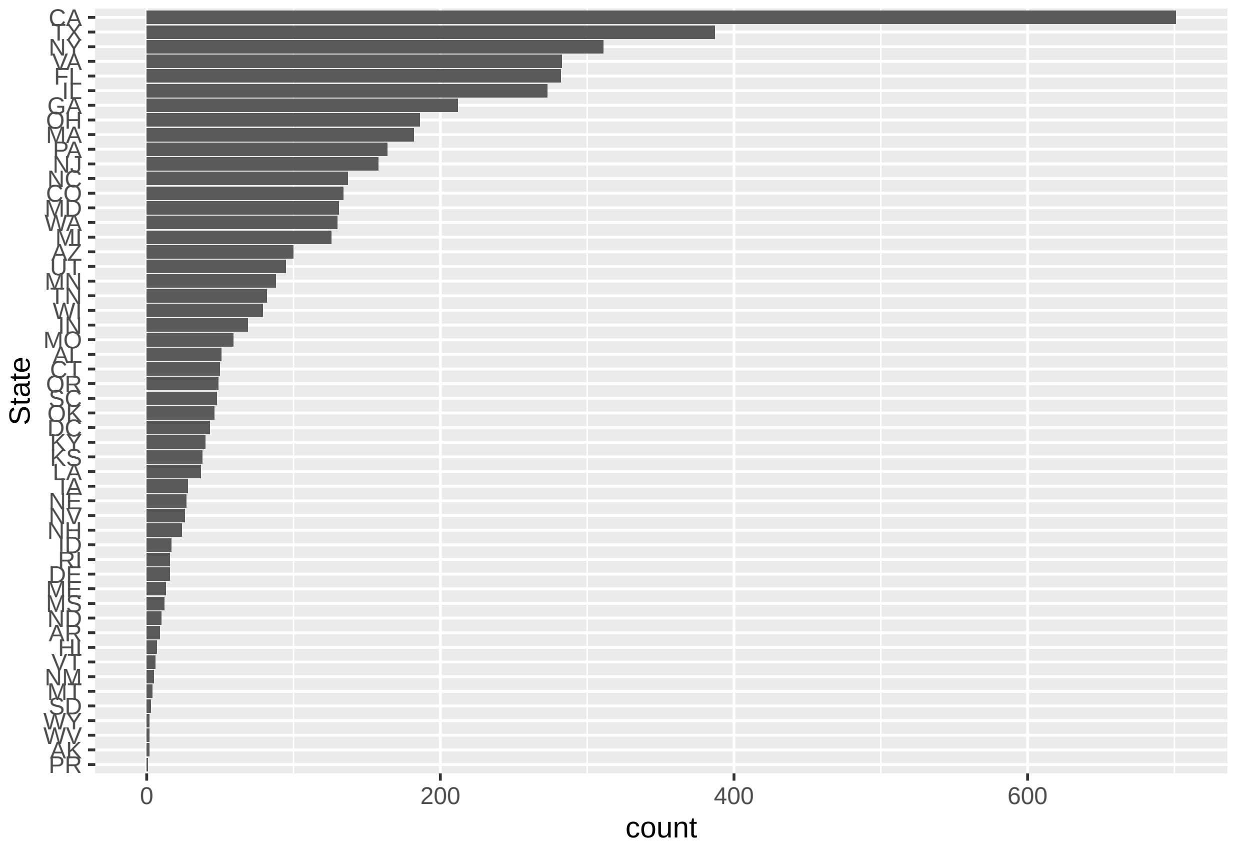
```
##        Rank        Name Growth_Rate     Revenue    Industry   Employees
##    0.000000    0.000000    0.000000    0.000000    0.000000    0.239952
##        City       State
##    0.000000    0.000000
```

Only the `Employees` column contains missing values, ~24%.

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
# I am using ggplot2 from tidyverse to make a bar chart.
df <- inc %>% group_by(State) %>%   mutate(count_name_occurr = n())
ggplot(data=df, aes(x=reorder(State,count_name_occurr))) +
  geom_bar(stat="count") +
  xlab('State') +
  coord_flip()
```

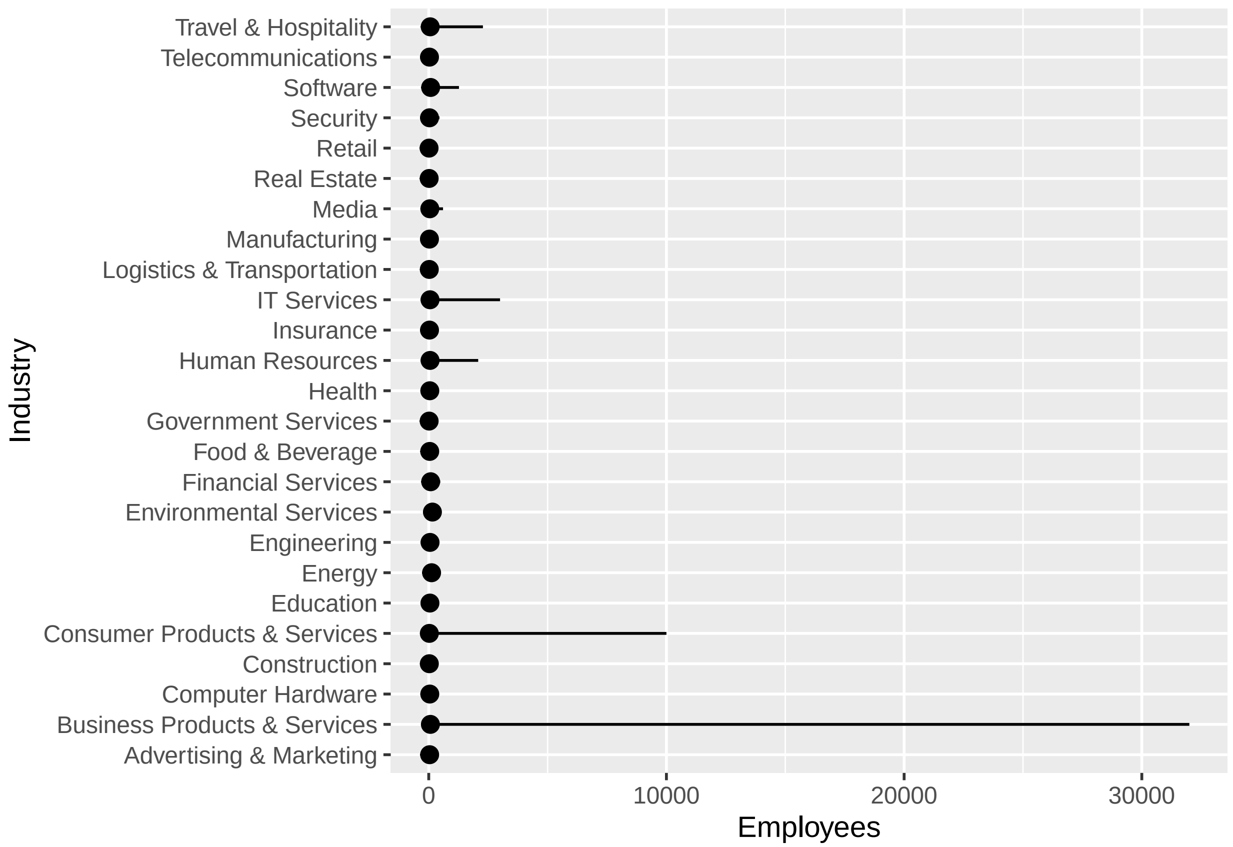


Quesiton 2

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.

```r
# Answer Question 2 here
df1 <- df %>% filter(State == 'NY')
df1 <- df1[complete.cases(df1),]
```

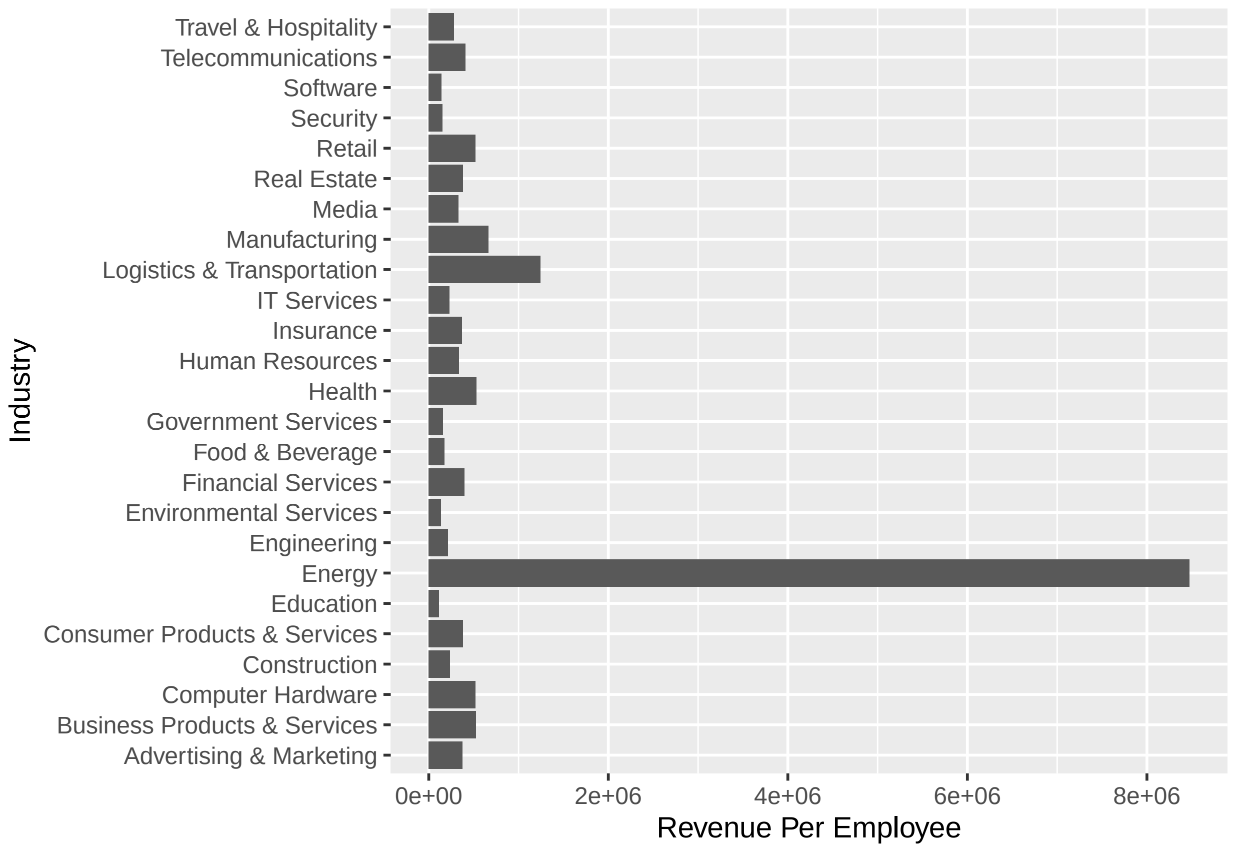
```r
ggplot(df1, aes(Industry, Employees)) +
    stat_summary(
    mapping = aes(x = Industry, y = Employees),
    fun.min = min,
    fun.max = max,
    fun = median
  ) +
  coord_flip()
```



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(df1, aes(Industry, Revenue/Employees)) +
  geom_bar(stat = "summary_bin", fun = mean) +
  ylab('Revenue Per Employee') +
  coord_flip()
```


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
Industry
0e+00
2e+06
4e+06
6e+06
8e+06
Revenue Per Employee