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:

```
# I used the describe feautre from Hmisc to get another look at the data.
library(Hmisc)
```

```
## Warning: package 'Hmisc' was built under R version 4.0.3
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## Warning: package 'Formula' was built under R version 4.0.3
```

```
## 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.

```
library(ggplot2)
library(ggthemes)
```

```
## Warning: package 'ggthemes' was built under R version 4.0.3
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------- tidyverse 1.3.0 --
```

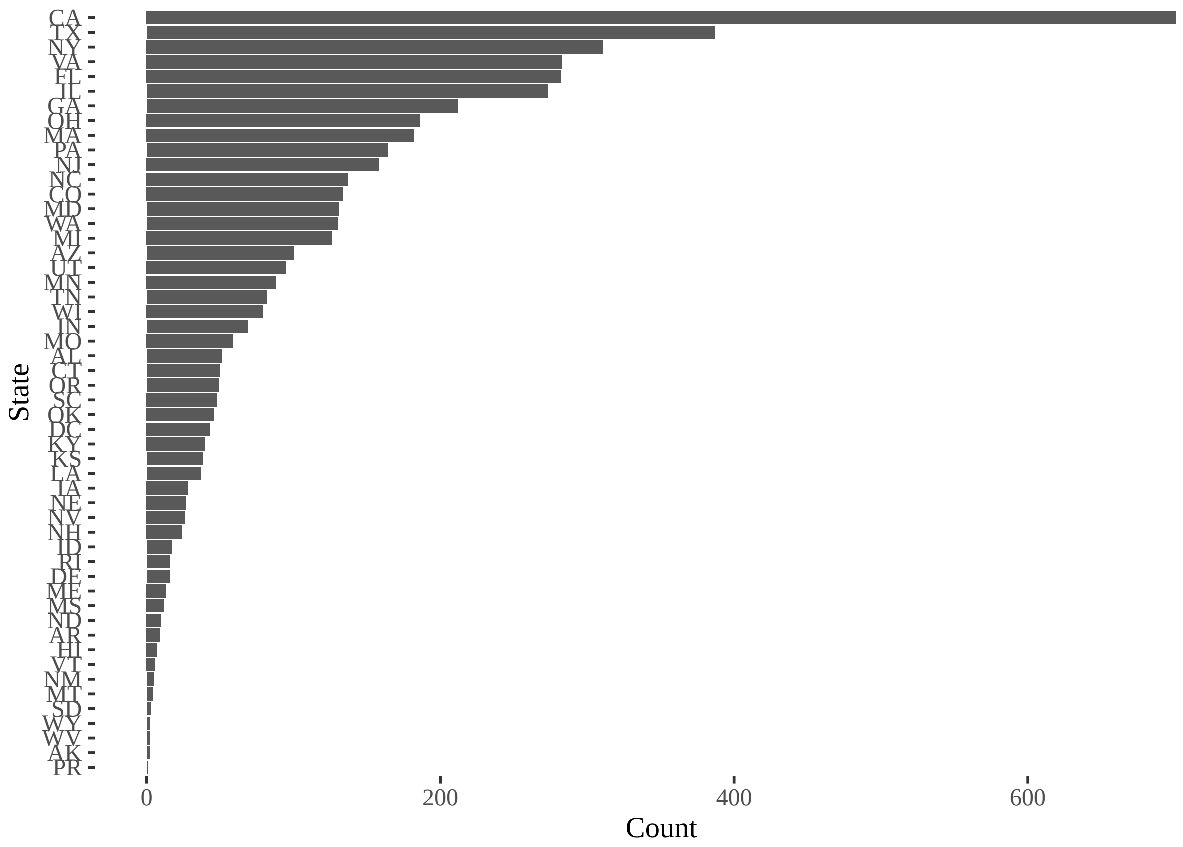
```
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
## v purrr   0.3.4
```

```
## -- Conflicts -------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter()    masks stats::filter()
## x dplyr::lag()       masks stats::lag()
## x dplyr::src()       masks Hmisc::src()
## x dplyr::summarize() masks Hmisc::summarize()
```

```
by_state <- inc %>%
  group_by(State) %>%
  summarise(count = n())
```

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

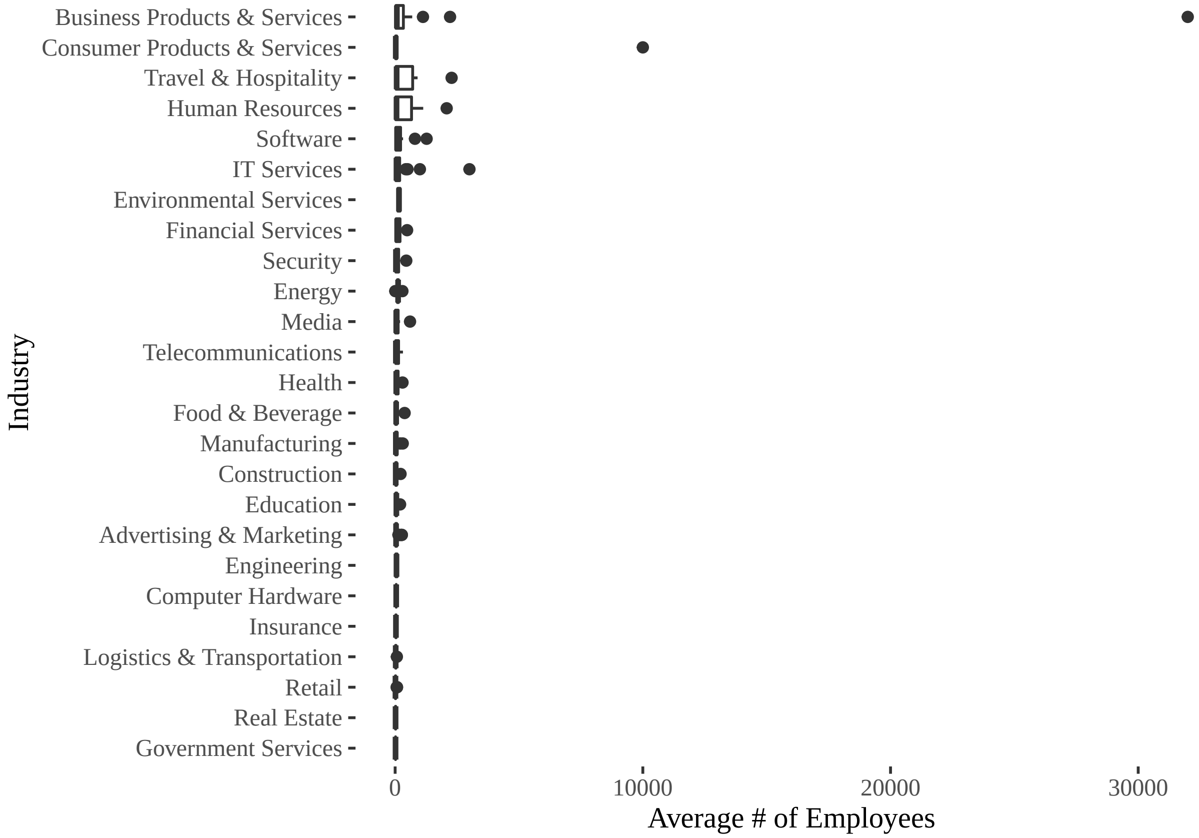
```
ggplot(by_state, aes(x = reorder(State, count), y = count)) +
  geom_col() +
  labs(x = "State",
       y = "Count") +
  coord_flip() +
  theme_tufte()
```



Quesiton 2

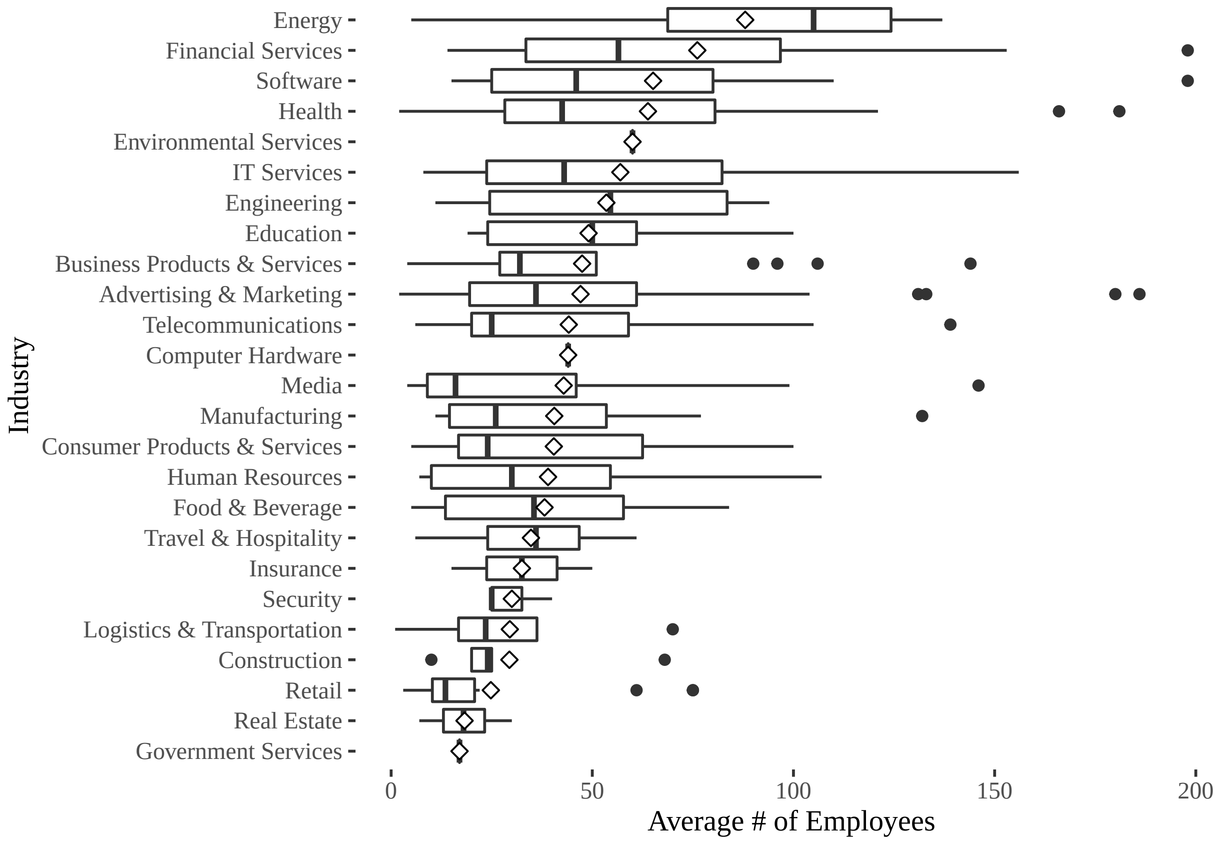
Lets dig in on the state with the 3rd most companies in the data set. Imagine you work for the state and are interested in how many people are employed by companies in different industries. Create a plot that shows the average and/or median employment by industry for companies in this state (only use 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.

```
# First I filtered down to just complete cases in NY.
ny <- filter(inc, State == "NY")
ny_complete <- ny[complete.cases(ny), ]

# Next, I looked at the data raw. I grouped the data by industry and created a boxplot to check for dis
ny_by_ave_empl <- ny_complete %>%
  group_by(Industry) %>%
  mutate(ave_empl = mean(Employees))

ggplot(ny_by_ave_empl, aes(x = reorder(Industry,ave_empl), y = Employees)) +
  geom_boxplot() +
  labs(x = "Industry",
       y = "Average # of Employees") +
  coord_flip() +
  theme_tufte()
```



```r
# Most companies had employees numbering in the hundreds, so the thousands were clear outliers. I remov
ny_by_ave_empl <- ny_complete %>%
  filter(Employees < 200) %>%
  group_by(Industry) %>%
  mutate(ave_empl = mean(Employees))

ggplot(ny_by_ave_empl, aes(x = reorder(Industry,ave_empl), y = Employees)) +
  geom_boxplot() +
  labs(x = "Industry",
       y = "Average # of Employees") +
  coord_flip() +
  stat_summary(fun.y = "mean", geom = "point", shape = 23, size = 2, fill = "white") +
  theme_tufte()
```

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



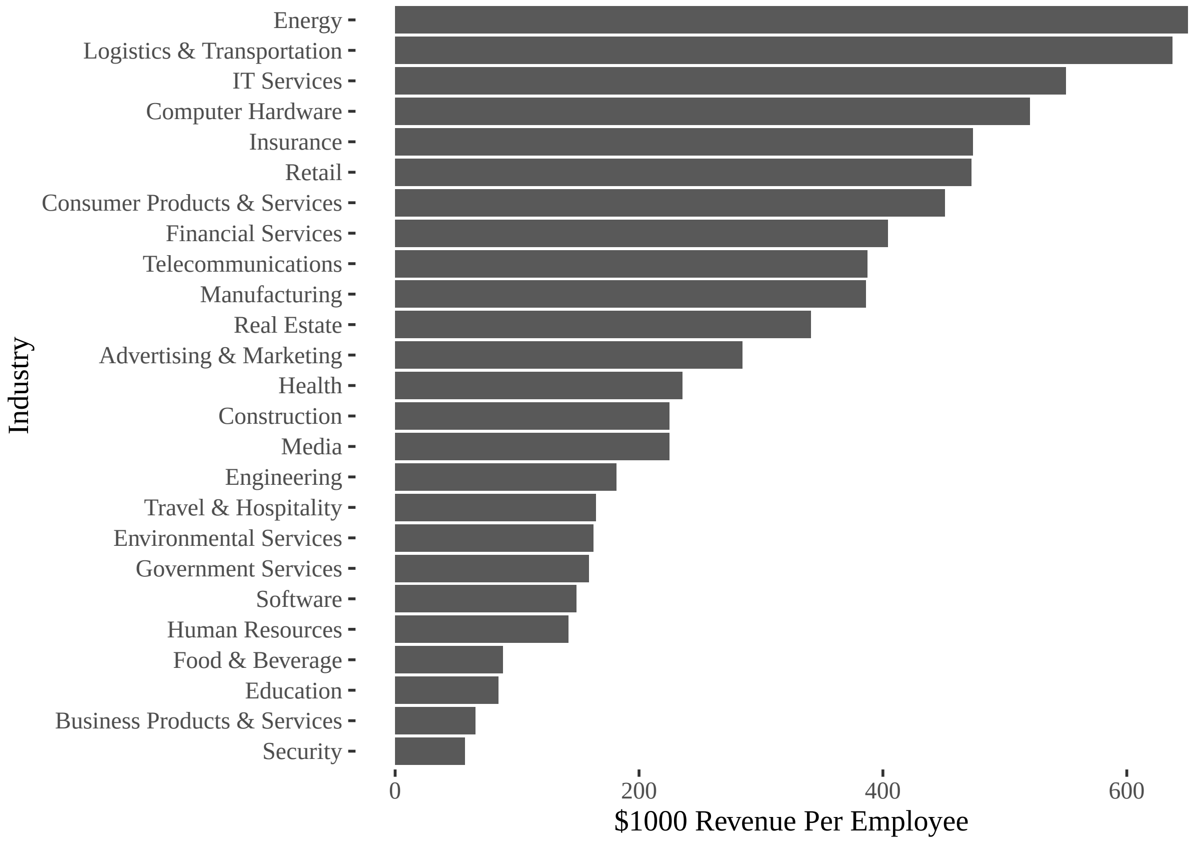
Question 3

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.

```r
ny_by_rev_per_empl <- ny_complete %>%
  group_by(Industry) %>%
  mutate(per_empl = (mean(Revenue)/1000)/mean(Employees)) %>%
  summarise(rev_per_empl = mean(per_empl))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```r
ggplot(ny_by_rev_per_empl, aes(x = reorder(Industry, rev_per_empl), y = rev_per_empl)) +
  geom_col() +
  labs(x = "Industry",
       y = "$1000 Revenue Per Employee") +
  coord_flip() +
  theme_tufte()
```


Energy
Logistics & Transportation
IT Services
Computer Hardware
Insurance
Retail
Consumer Products & Services
Financial Services
Telecommunications
Manufacturing
Real Estate
Advertising & Marketing
Health
Construction
Media
Engineering
Travel & Hospitality
Environmental Services
Government Services
Software
Human Resources
Food & Beverage
Education
Business Products & Services
Security
Industry
0
200
400
600
$1000 Revenue Per Employee