DATA 608: HW 1

Andrew Carson

August 28, 2018

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:

Answer:

- Growth rate: ranges from 0.34 to 421 with a median of 1.4. Consequently, there will be lots of small values and very few large values, creating a skew right.
- Revenue: this is similar, in that the range from 2,000,000 to 10,000,000,000 has an average of 48,220,000, making it also skewed right.
- Industry: The top category of industry by count is IT, followed by Business and Marketing. I'd be curious to know how this relates to growth rate and revenue. Using code below, the average growth rate by industry does not match with count by industry. IT and Business are not in the top 10. At the top is energy and Consumer products. For average Revenue by industry, computer hardware and energy are on top, so this also does not match with the top counts by industry.
- Employees: most companies are small. The range from 1 to 66,803 has a 3rd quartile of 132. This doesn't surprise me as it is very easy for small companies to have large growth rates (e.g., moving from 1 to 2 employees doubles the company size), while it is much more difficult for large companies to have large growth rates. This can be confirmed by comparing Employee size to Growth Rate. By grouping the employee size into five buckets of equal counts of companies, we can see that the lowest employee size bucket (1, 25) has the highest average growth rate (4.9%), followed by the next bucket (25,53), the next bucket (53, 132), and the final bucket (132, 66,803), each in order of employee size descending.
- City and State: While the top city is New York, the top state is CA. This must mean that CA has more cities with growing businesses than NY, so the growth is less concentrated.

```r
library(tidyr)
library(dplyr)

# industry vs. growth rate
industryVsGrowth<-inc %>%
  group_by(Industry) %>%
  summarise(avg_Growth_Rate = mean(Growth_Rate, na.rm=TRUE)) %>%
  arrange(desc(avg_Growth_Rate))

head(industryVsGrowth, 10)
```

```
## # A tibble: 10 x 2
##                         Industry avg_Growth_Rate
##                           <fctr>           <dbl>
##  1                        Energy        9.603303
##  2 Consumer Products & Services        8.776108
##  3                   Real Estate        7.746667
##  4          Government Services        7.238168
##  5      Advertising & Marketing        6.225478
##  6                        Retail        6.184729
##  7           Financial Services        5.435308
##  8                      Software        5.020643
##  9                        Health        4.856394
## 10                         Media        4.374074
```

```r
# industry vs. revenue
industryVsRevenue<-inc %>%
  group_by(Industry) %>%
  summarise(avg_Revenue = mean(Revenue, na.rm=TRUE)) %>%
  arrange(desc(avg_Revenue))
```

```
head(industryVsRevenue, 10)
```

```
## # A tibble: 10 x 2
##                          Industry avg_Revenue
##                             <fctr>       <dbl>
##  1            Computer Hardware   270129545
##  2                       Energy   126344954
##  3              Food & Beverage    98559542
##  4   Logistics & Transportation    95745161
##  5 Consumer Products & Services    73676847
##  6                 Construction    70450802
##  7           Telecommunications    56855814
##  8 Business Products & Services    54705187
##  9                     Security    52230137
## 10       Environmental Services    51741176
```

```
# employees vs. growth rate
inc$Employees_Cut <- cut(inc$Employees,fivenum(inc$Employees))

employeesVsGrowth<-inc %>%
  group_by(Employees_Cut) %>%
  summarise(avg_Growth_Rate = mean(Growth_Rate, na.rm=TRUE)) %>%
  arrange(desc(avg_Growth_Rate))

head(employeesVsGrowth, 10)
```

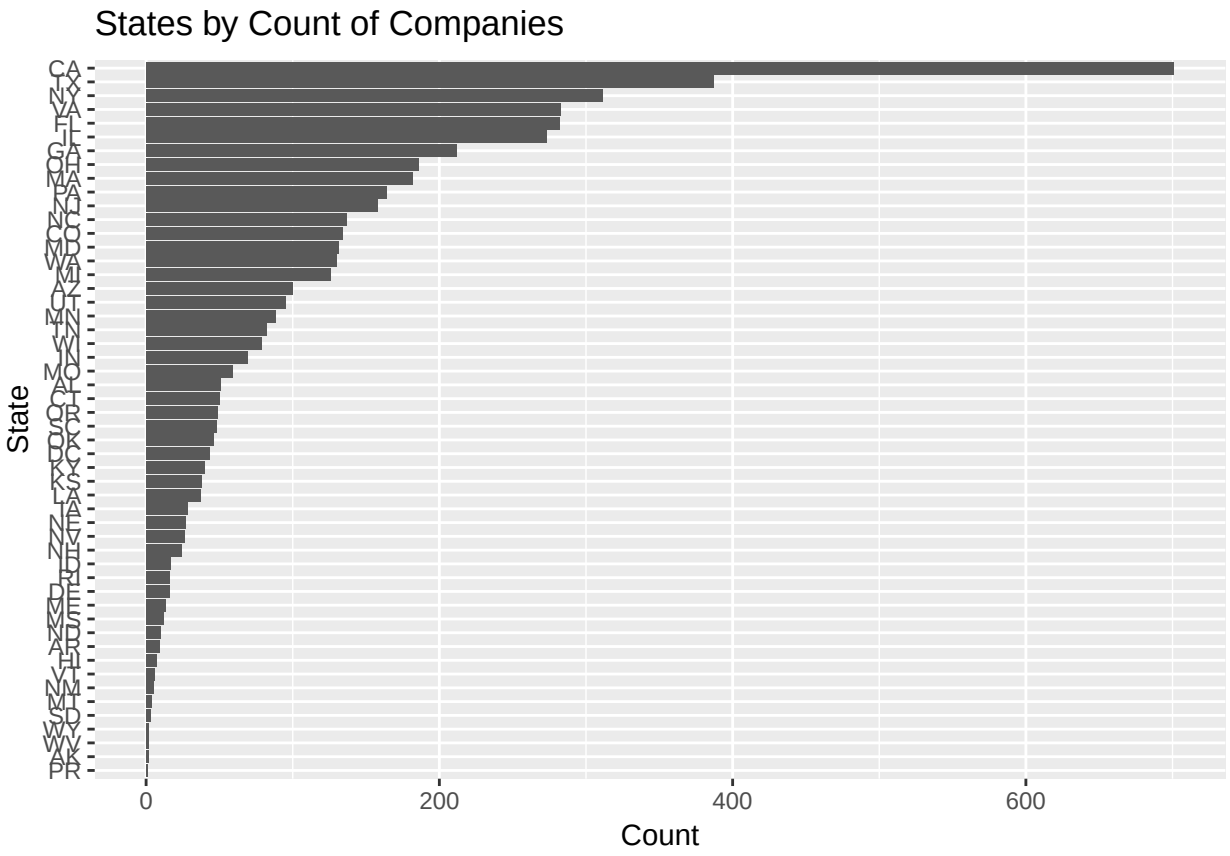
```
## # A tibble: 5 x 2
##    Employees_Cut avg_Growth_Rate
##           <fctr>           <dbl>
## 1         (1,25]        4.907083
## 2        (25,53]        4.843713
## 3       (53,132]        4.769115
## 4 (132,6.68e+04]        3.935764
## 5           <NA>        3.706111
```

Question 1

Create a graph that shows the distribution of companies in the dataset by State (i.e. 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)

#bar plot
qplot(data=inc, x=reorder(State, table(inc$State)[State]), ylab = "Count", xlab = "State", main = "Stat
```



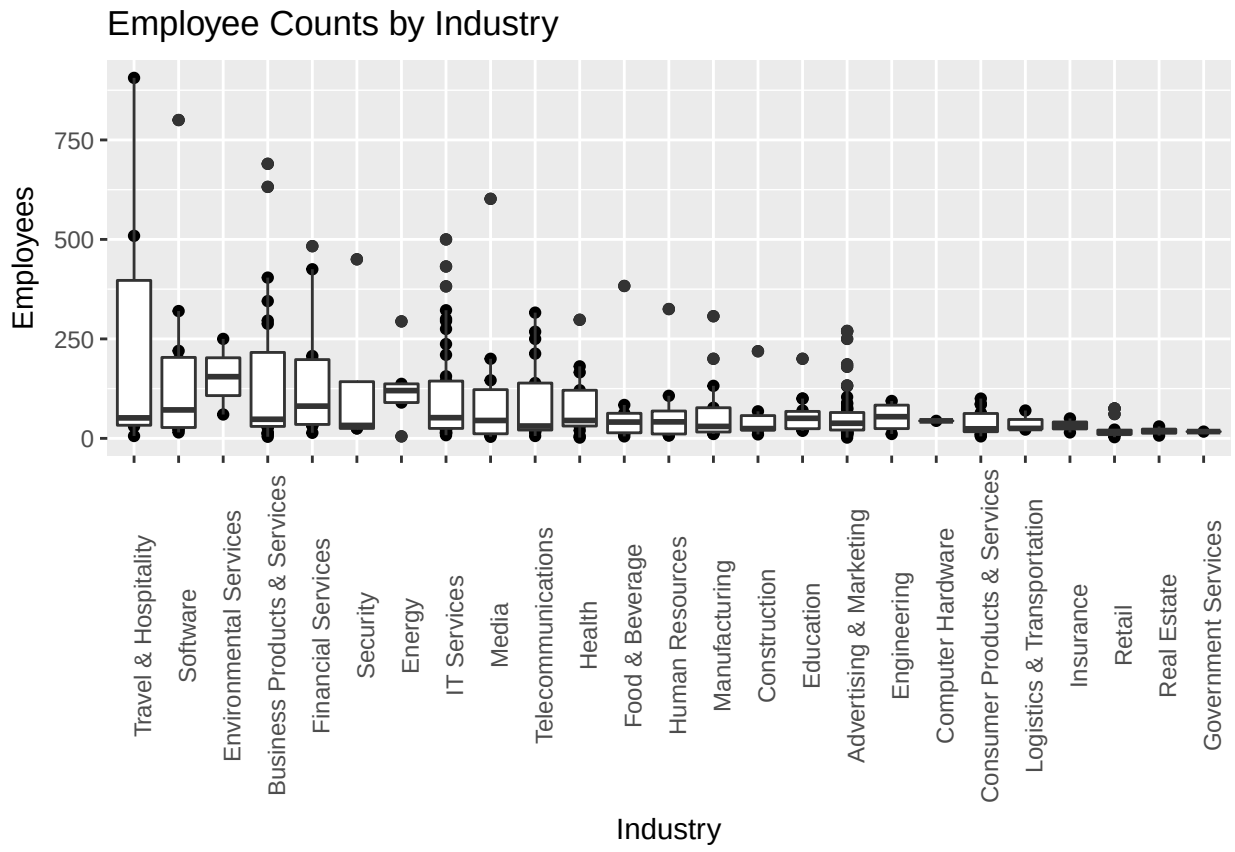
Question 2

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.

```
# state with 3rd most companies is: NY
NewYork <- inc %>%
  filter(State=="NY") %>%
  filter(Employees < 1000) #handle outliers based on visual inspection

#complete.cases
NewYork <- NewYork[complete.cases(NewYork),]

#boxplot: handles outliers, ranges, median
qplot(data=NewYork, x=reorder(Industry, -Employees), y=Employees, xlab="Industry", ylab="Employees", ma
```



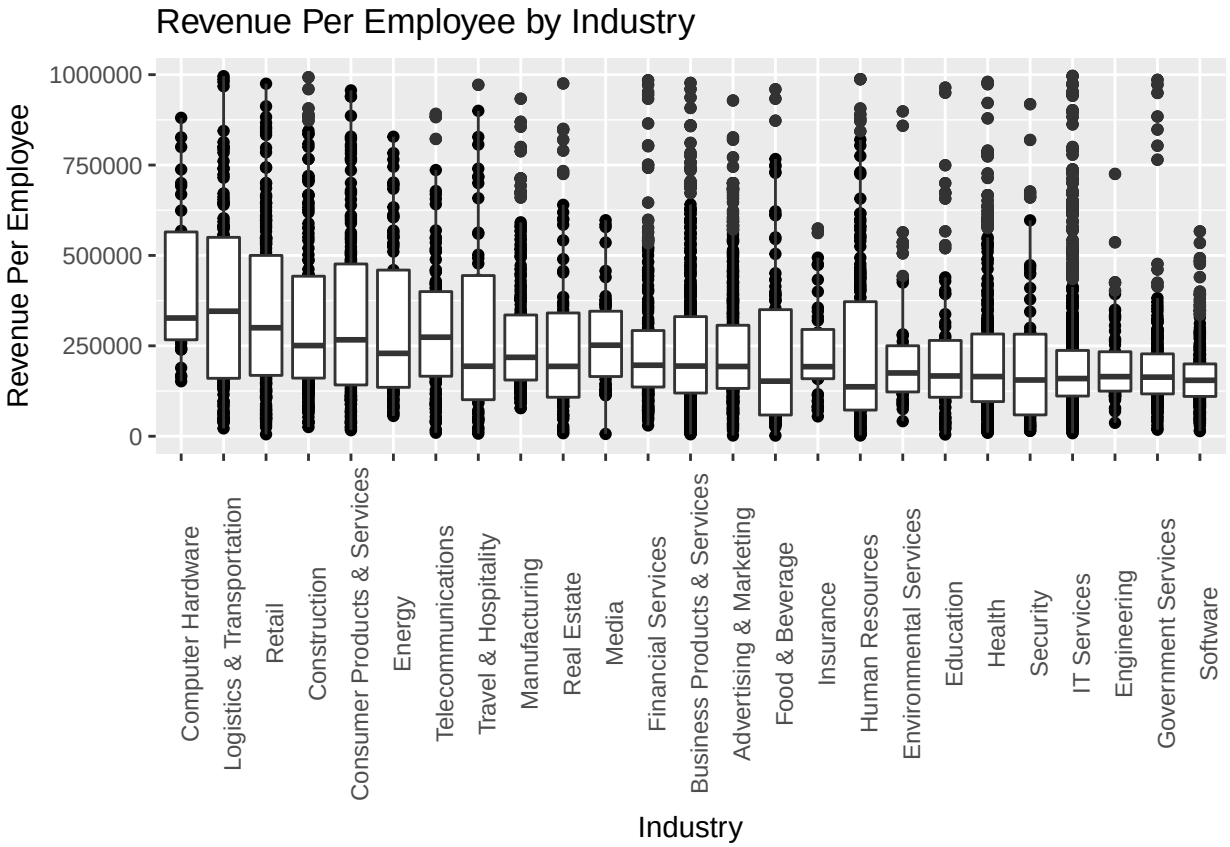
Question 3

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.

```
#revenue per employee
inc$RevPerEmp <- inc$Revenue / inc$Employees

#complete cases and remove outliers
inc_complete <-inc[complete.cases(inc),]
inc_complete <- inc %>% filter(RevPerEmp < 1000000)

#boxplot: handles outliers, ranges, median
qplot(data=inc_complete, x=reorder(Industry, -RevPerEmp), y=RevPerEmp, xlab="Industry", ylab="Revenue P
```


Revenue Per Employee by Industry
Revenue Per Employee
1000000
750000
500000
250000
0
Computer Hardware
Logistics & Transportation
Retail
Construction
Consumer Products & Services
Energy
Telecommunications
Travel & Hospitality
Manufacturing
Real Estate
Media
Financial Services
Business Products & Services
Advertising & Marketing
Food & Beverage
Insurance
Human Resources
Environmental Services
Education
Health
Security
IT Services
Engineering
Government Services
Software
Industry