608-Module 1

Jagruti

9/3/2018

Reading Data

```
library(plyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
inc <- read.csv("https://raw.githubusercontent.com/charleyferrari/CUNY_DATA_608/master/module1/Data/inc5
```

```
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:

```
inc1 <- inc %>% group_by(State) %>% count(Name)
inc2 <- inc1 %>% group_by(State) %>% count(n)
inc2 <- subset(inc2,select = c("State","nn"))
inc2$State <- factor(inc2$State)
#inc2 <- as.data.frame(inc2)
inc2
```

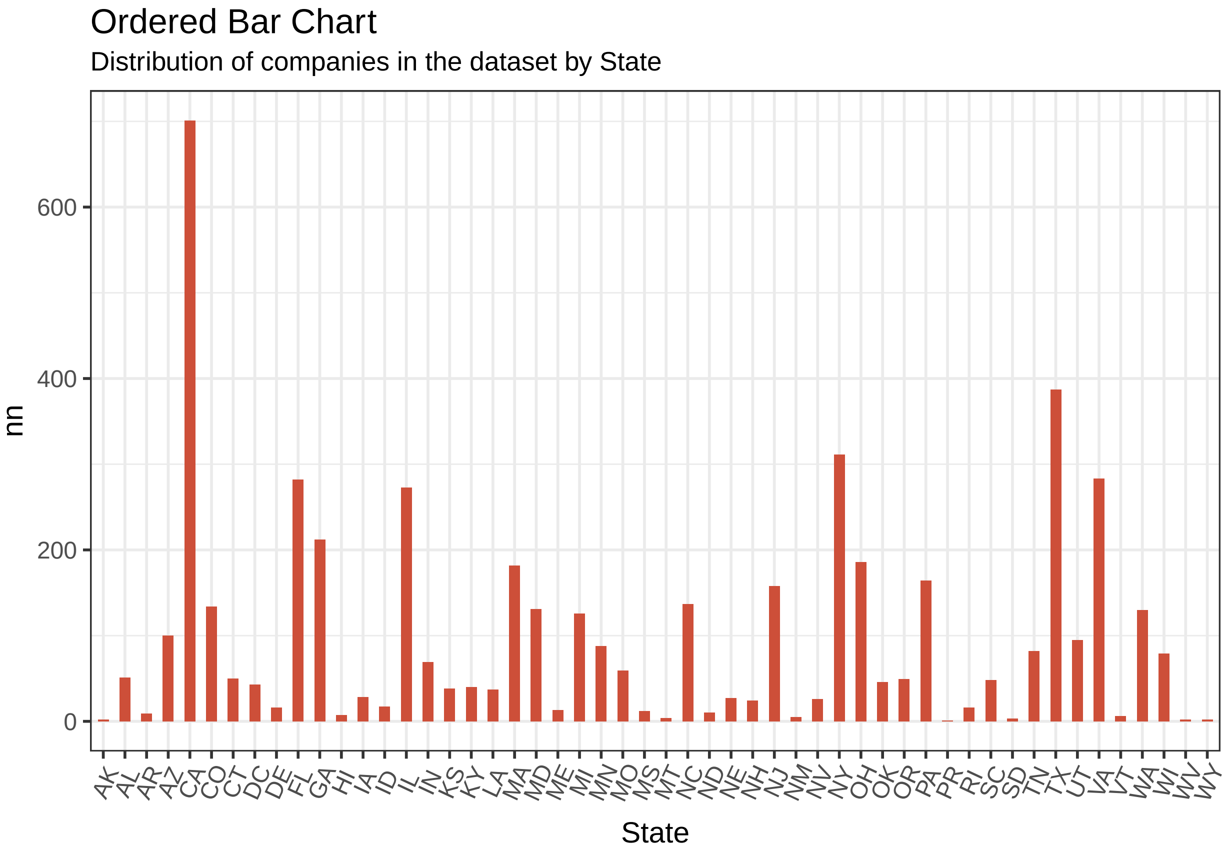
```
## # A tibble: 52 x 2
## # Groups:   State [52]
##     State    nn
##    <fctr> <int>
##  1     AK     2
##  2     AL    51
##  3     AR     9
##  4     AZ   100
##  5     CA   701
##  6     CO   134
##  7     CT    50
##  8     DC    43
##  9     DE    16
## 10     FL   282
## # ... with 42 more rows
```

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.

```
theme_set(theme_bw())
# Draw plot
```

```
ggplot(inc2, aes(x=State, y=nn)) +
  geom_bar(stat="identity", width=.5, fill="tomato3") +
  labs(title="Ordered Bar Chart",
       subtitle="Distribution of companies in the dataset by State") +
  theme(axis.text.x = element_text(angle=65, vjust=0.6))
```



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.

```
arrange(inc2,desc(nn))
```

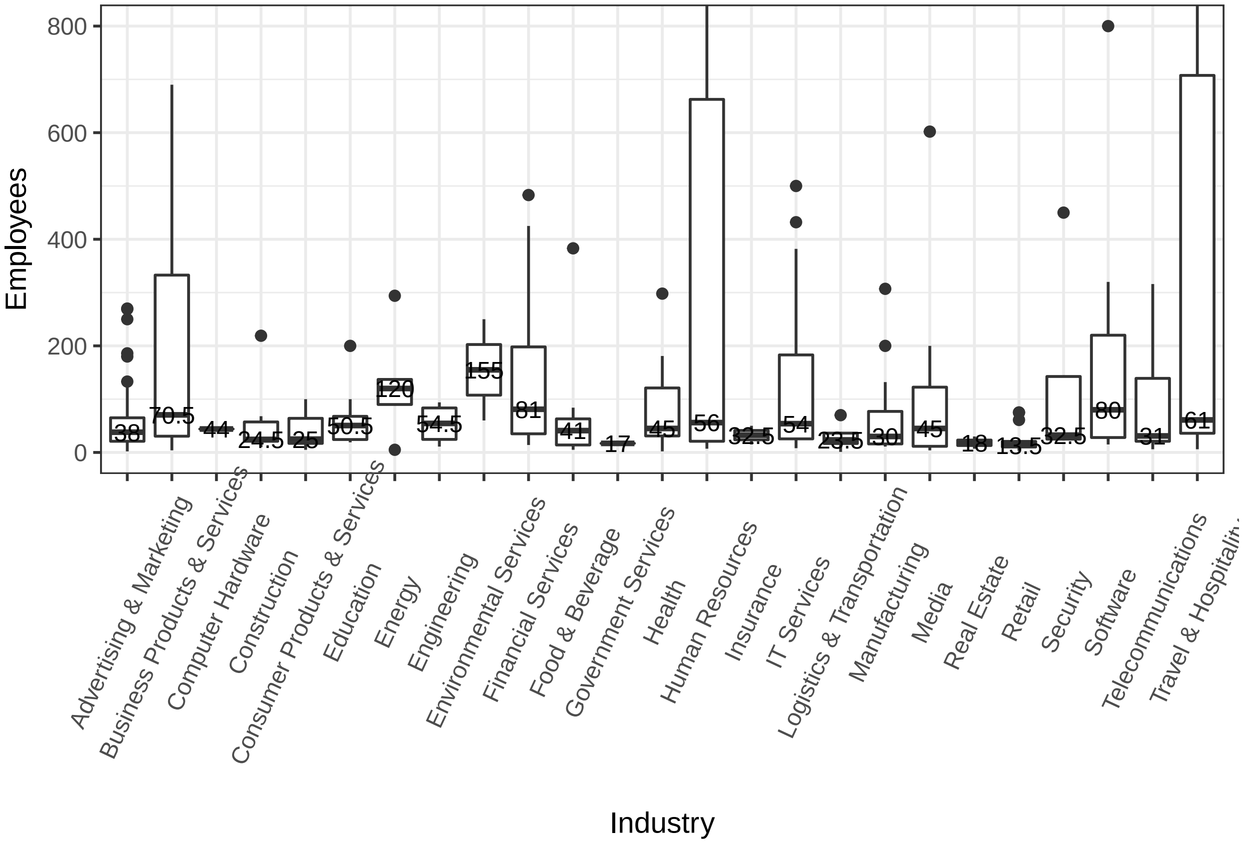
```
## # A tibble: 52 x 2
## # Groups:   State [52]
##     State    nn
##    <fctr> <int>
##  1     CA   701
##  2     TX   387
##  3     NY   311
##  4     VA   283
##  5     FL   282
##  6     IL   273
##  7     GA   212
##  8     OH   186
##  9     MA   182
## 10     PA   164
## # ... with 42 more rows
```

NY is state with 3rd most companies in the data set.

```
inc3 = inc[which(inc$State == "NY"),]
inc3 = inc3 %>% filter(complete.cases(.))
inc3 <- subset(inc3,select = c("Industry","Employees"))
head(inc3)
```

```
##                        Industry Employees
## 1 Consumer Products & Services        17
## 2      Advertising & Marketing        79
## 3      Advertising & Marketing        27
## 4      Advertising & Marketing        89
## 5           Financial Services        32
## 6      Advertising & Marketing        75
```

```
inc4 <- inc3 %>% group_by(Industry)
library(ggplot2)
p_meds <- ddply(inc3, .(Industry), summarise, med = median(Employees))
ggplot(inc3,aes(x = Industry, y = Employees)) +
    geom_boxplot() +
    geom_text(data = p_meds, aes(x = Industry, y = med, label = med),
              size = 3) + theme(axis.text.x = element_text(angle=65, vjust=0.6)) + coord_cartesian(ylim
```



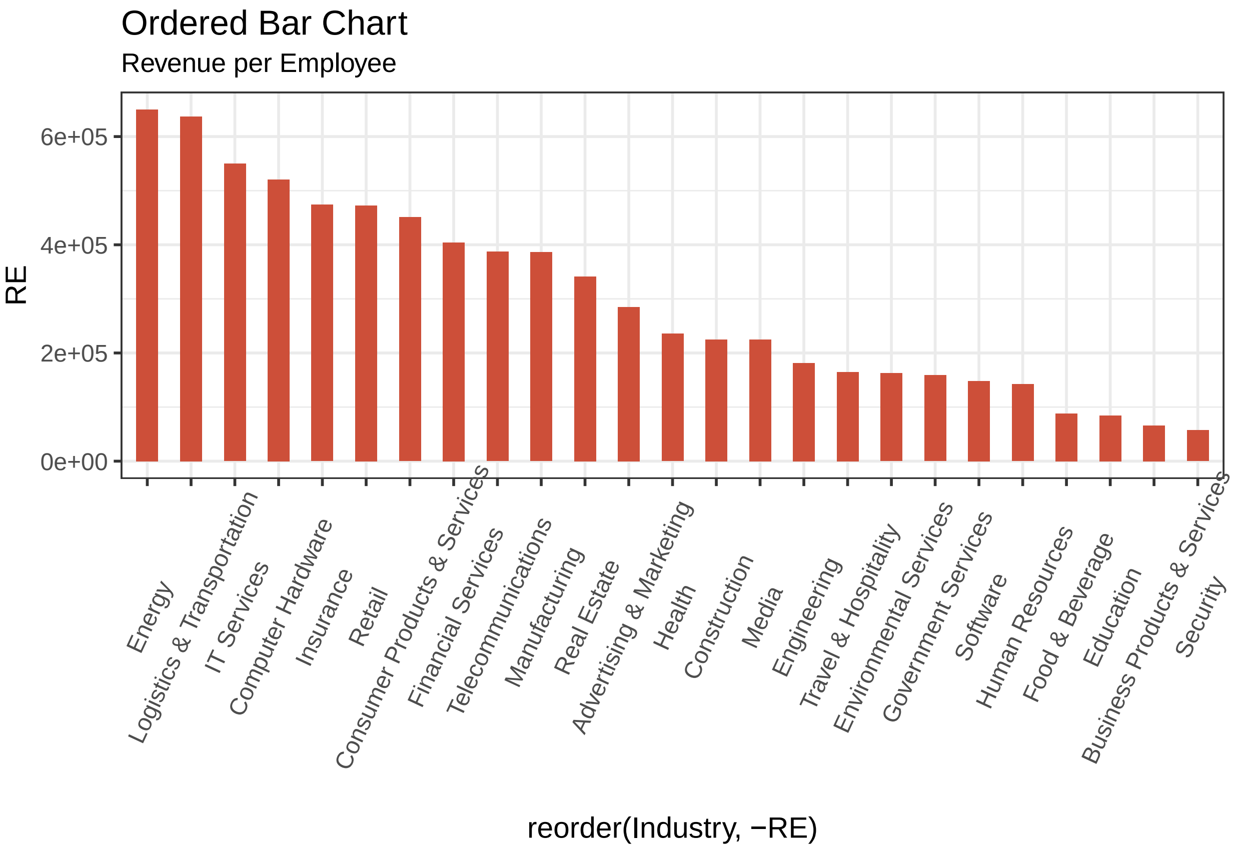
Question 3

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.

```
inc6 = inc[which(inc$State == "NY"),]
inc6 = inc6 %>% filter(complete.cases(.))
inc6 <- subset(inc6,select = c("Revenue","Employees","Industry"))
```

```
c = ddply(inc6,.(Industry),summarise,Revenue = sum(Revenue), Employees = sum(Employees))
c["RE"] <- (c$Revenue/c$Employees)
```

```
ggplot(c, aes(x= reorder(Industry, -RE), y=RE)) +
  geom_bar(stat="identity", width=.5, fill="tomato3") +
  labs(title="Ordered Bar Chart",
       subtitle="Revenue per Employee",xlabel = "Industry",ylabel = "Revenue per Industry") + theme(axi
```



Energy Industry generates most Revenue per employee.