# Data608 Module1

*Chunhui Zhu*

*September 3, 2018*

```
suppressMessages(suppressWarnings(library('tidyverse')))
suppressMessages(suppressWarnings(library('dplyr')))
suppressMessages(suppressWarnings(library('ggplot2')))
suppressMessages(suppressWarnings(library('scales')))
```

## Principles of Data Visualization and Introduction to ggplot2

### Raw Data

**"These data are 5,000 fastest growing companies in the US, as compiled by Inc. magazine."**

```
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."**

-1.The summaries for data table give general information for each columns of data. However, these information do not necessary have relavent amoung the columns, also it doesn' include the number of oberservations for the whold data table.

-2.It is easy to see 'Employees' has 12 missing data in the column.

-3.For numerical data type of columns, summaries show their basic statistic information s.t. mean, and 3 quartile ranges. Comparing with min and max, it is telling the distribution is bias on left or right tails. Sometimes, this information is meaningless, s.t. in 'Rank' in this table.

-4.For categorial data type of columns, summeries give the frequency of categries in each columns. For examples in 'City', New York has 160 obervations, following is Chicago 90 obervations, then Austin 88 obervations.

-5.Since there are 5000 in 'Rank' and only 4995 in 'Name', it tells at least 5 companies names are duplicate in the data table.

-6.For the relation between 'Rank' to other columns, it is not able to observe from the summeries.

**"Use the space below to add some more relevant non-visual exploratory information you think helps you understand this data:"**

-1.This 'inc' table has 5001 observations and 8 fileds.

```
dim(inc)
```

```
## [1] 5001    8
```

-2. Check the information of 12 missing values in 'Employees'. The company 'Frist flight Solutions' ranking at 183th might has the concern. This company has fast growth rate at 22.32%. And the company 'Heartland Business Systems' has 156 millions which has the lagest 'Revenue' amoung them.

```
data_missing<-inc[!complete.cases(inc),]
data_missing
```

```
##      Rank                              Name Growth_Rate   Revenue
## 183   183           First Flight Solutions       22.32   2700000
## 1063 1064                        Popchips        3.98  93300000
## 1123 1124                      Vocalocity        3.72  42900000
## 1652 1653                     Higher Logic        2.36   6000000
## 1685 1686      Global Communications Group        2.30   3600000
## 2196 2197              JeffreyM Consulting        1.68  12100000
## 2742 2743               Excalibur Exhibits        1.27   9900000
## 3000 3001       Heartland Business Systems        1.12 156300000
## 3978 3978                             SSEC        0.68  80400000
## 4112 4112 Carolinas Home Medical Equipment        0.64   3300000
```

```
## 4566 4566                          Oakbrook        0.48   8900000
## 4968 4968                    Popcorn Palace        0.35   5500000
##                            Industry Employees          City State
## 183    Logistics & Transportation        NA  Emerald Isle    NC
## 1063              Food & Beverage        NA San Francisco    CA
## 1123           Telecommunications        NA       Atlanta    GA
## 1652                     Software        NA    Washington    DC
## 1685           Telecommunications        NA     Englewood    CO
## 2196 Business Products & Services        NA      Bellevue    WA
## 2742 Business Products & Services        NA       houston    TX
## 3000                   IT Services        NA  Little Chute    WI
## 3978                 Manufacturing        NA       Horsham    PA
## 4112                        Health        NA      Matthews    NC
## 4566                  Real Estate        NA       Madison    WI
## 4968              Food & Beverage        NA Schiller Park    IL
```

-3.To drop 'Employees' column, it is easy to use aggreate functions in r to find the top 10 revenues by industries and states.

```
df_drop_Employees<-inc[c(-6)]
dim(df_drop_Employees)
```

```
## [1] 5001    7
```

```
#top 10 revenues by industries
df_drop_Employees %>% group_by(Industry) %>% summarise(Revenue = sum(Revenue)) %>% arrange(desc(Revenue)
```

```
## Warning: package 'bindrcpp' was built under R version 3.3.3
```

```
## # A tibble: 10 x 2
##                         Industry     Revenue
##                           <fctr>       <dbl>
##  1 Business Products & Services 26367900000
##  2                  IT Services 20681300000
##  3                       Health 17863400000
##  4 Consumer Products & Services 14956400000
##  5   Logistics & Transportation 14840500000
##  6                       Energy 13771600000
##  7                 Construction 13174300000
##  8           Financial Services 13150900000
##  9              Food & Beverage 12911300000
## 10                Manufacturing 12684000000
```

```
#top 10 revenues by State
df_drop_Employees %>% group_by(State) %>% summarise(Revenue = sum(Revenue)) %>% arrange(desc(Revenue)) %
```

```
## # A tibble: 10 x 2
##     State     Revenue
##    <fctr>       <dbl>
##  1     IL 33244300000
##  2     CA 23457900000
##  3     TX 22164200000
##  4     NY 18260400000
##  5     OH 12786600000
##  6     FL 10610300000
##  7     NC  9258500000
##  8     VA  8667700000
```

```
##  9     MI  7805800000
## 10     WI  7296600000
```

-4.To remove missing values in 'Employees' column and use same aggreate functions in r to find the top 10 revenues by industries and states.

```
df_remove_na<-na.omit(inc)
dim(df_remove_na)
```

```
## [1] 4989    8
```

```
#top 10 revenues by industries
top_10_Industry<-df_remove_na %>% group_by(Industry) %>% summarise(Revenue = sum(Revenue)) %>% arrange(
top_10_Industry
```

```
## # A tibble: 10 x 2
##                         Industry     Revenue
##                           <fctr>       <dbl>
##  1 Business Products & Services 26345900000
##  2                  IT Services 20525000000
##  3                       Health 17860100000
##  4 Consumer Products & Services 14956400000
##  5   Logistics & Transportation 14837800000
##  6                       Energy 13771600000
##  7                 Construction 13174300000
##  8           Financial Services 13150900000
##  9              Food & Beverage 12812500000
## 10                Manufacturing 12603600000
```

```
#top 10 revenues by State
top_10_state<-df_remove_na %>% group_by(State) %>% summarise(Revenue = sum(Revenue)) %>% arrange(desc(R
top_10_state
```

```
## # A tibble: 10 x 2
##     State     Revenue
##    <fctr>       <dbl>
##  1     IL 33238800000
##  2     CA 23364600000
##  3     TX 22154300000
##  4     NY 18260400000
##  5     OH 12786600000
##  6     FL 10610300000
##  7     NC  9252500000
##  8     VA  8667700000
##  9     MI  7805800000
## 10     WI  7131400000
```

-5.To compare the results from steps 3 and 4, the missing data do not effect the rank of Revenue by Industries and States. It could have a assumption that missing data are too samll to effect the Revenue rank for the top 10 industries and states.

-6.To observe if the duplicated company 'Name' exist in 'df_remove_na' table.

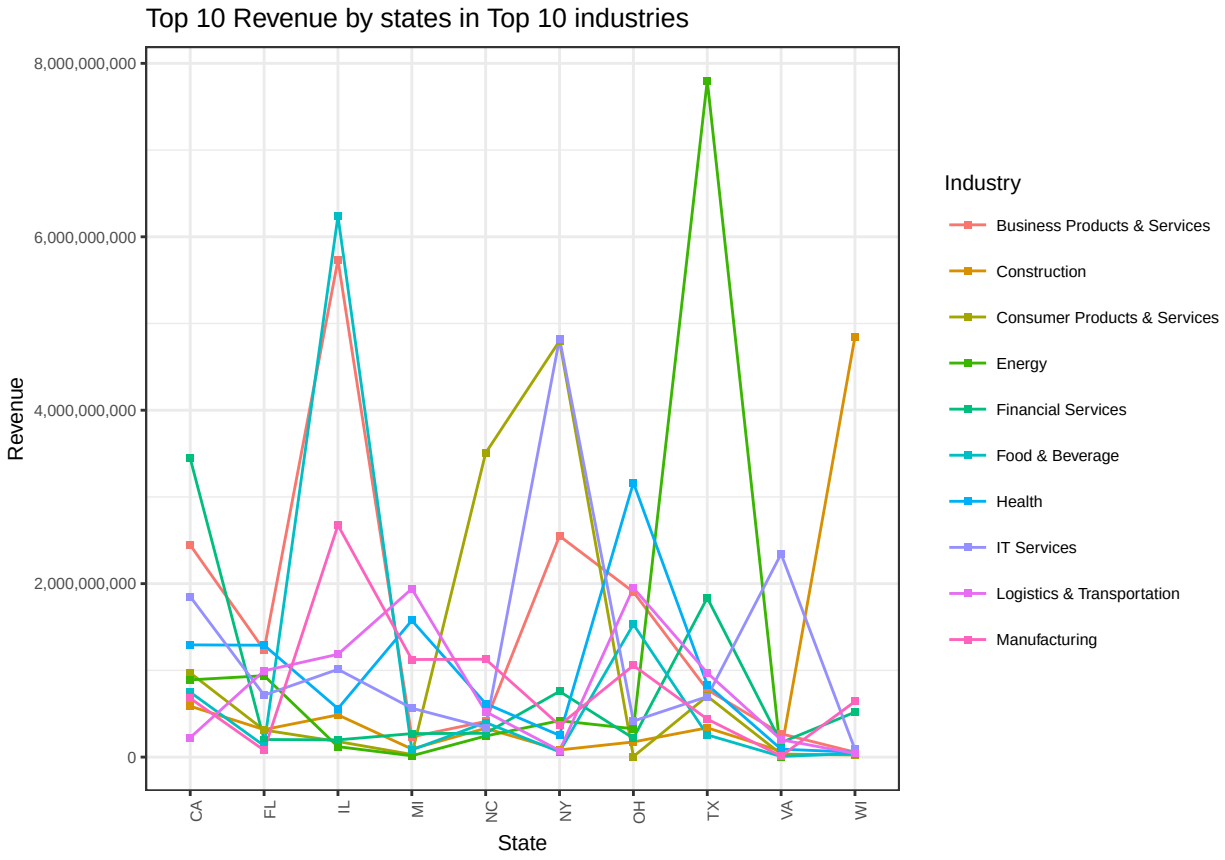
```
#top 10 revenues by State
nrow(df_remove_na[duplicated(df_remove_na$Name),])
```

```
## [1] 0
```

Great! In the following analysis, we could directly use 'df_remove_na' table.

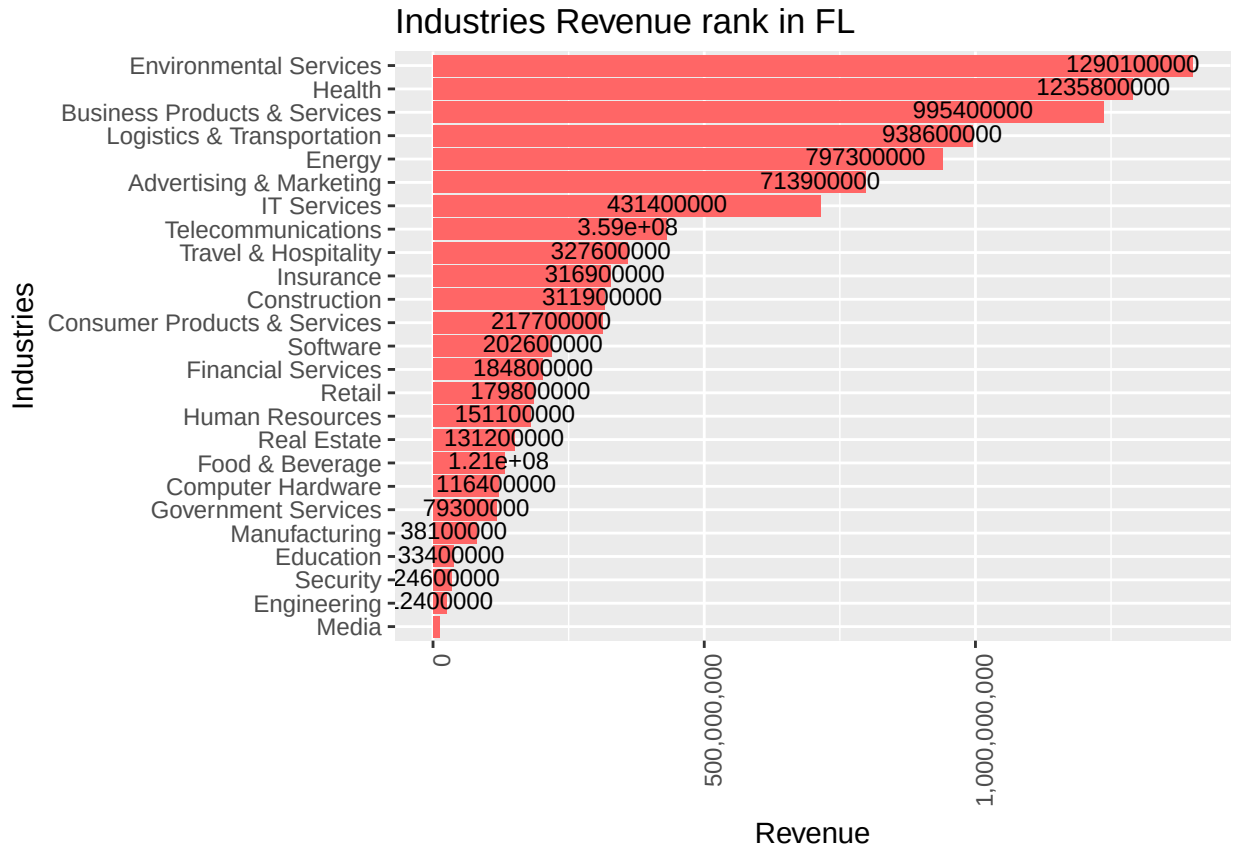
-7.Observe “Top 10 Revenue by states in Top 10 industries”

```
df1<-df_remove_na %>% group_by(State,Industry) %>% summarise(Revenue = sum(Revenue))
df2<-df1[df1$State %in% top_10_state$State,]
df3<-df2[df2$Industry %in% top_10_Industry$Industry,]

ggplot(data=df3, aes(x=State, y=Revenue, group = Industry, colour = Industry)) +
     geom_line() +
    geom_point(size=1, shape=15, fill="white")+
    ggtitle("Top 10 Revenue by states in Top 10 industries") +theme_bw(base_size = 8) +
  theme(axis.text.x=element_text(angle=90,hjust=1))+scale_y_continuous(labels = comma)
```




-8. FL has everything low revenue in top 10 revenue industries. Let’s take a look for revenue ranking by industries in FL.

```
fl_df<- df1[df1$State=='FL',]%>%  subset(select=c('Industry','Revenue')) %>% arrange(desc(Revenue))
fl_df<-fl_df[order(fl_df$Revenue,decreasing = TRUE),]

ggplot(data=fl_df, aes(x=reorder(Industry, Revenue), y=Revenue)) +
  geom_bar(stat="identity",fill = "#FF6666")+
  geom_text(aes(label=Revenue), vjust=-1, color="black", size=3)+
  ggtitle("Industries Revenue rank in FL") +
  xlab("Industries") +theme(axis.text.x=element_text(angle=90,hjust=1))+
  scale_y_continuous(labels = comma)+
  coord_flip()
```

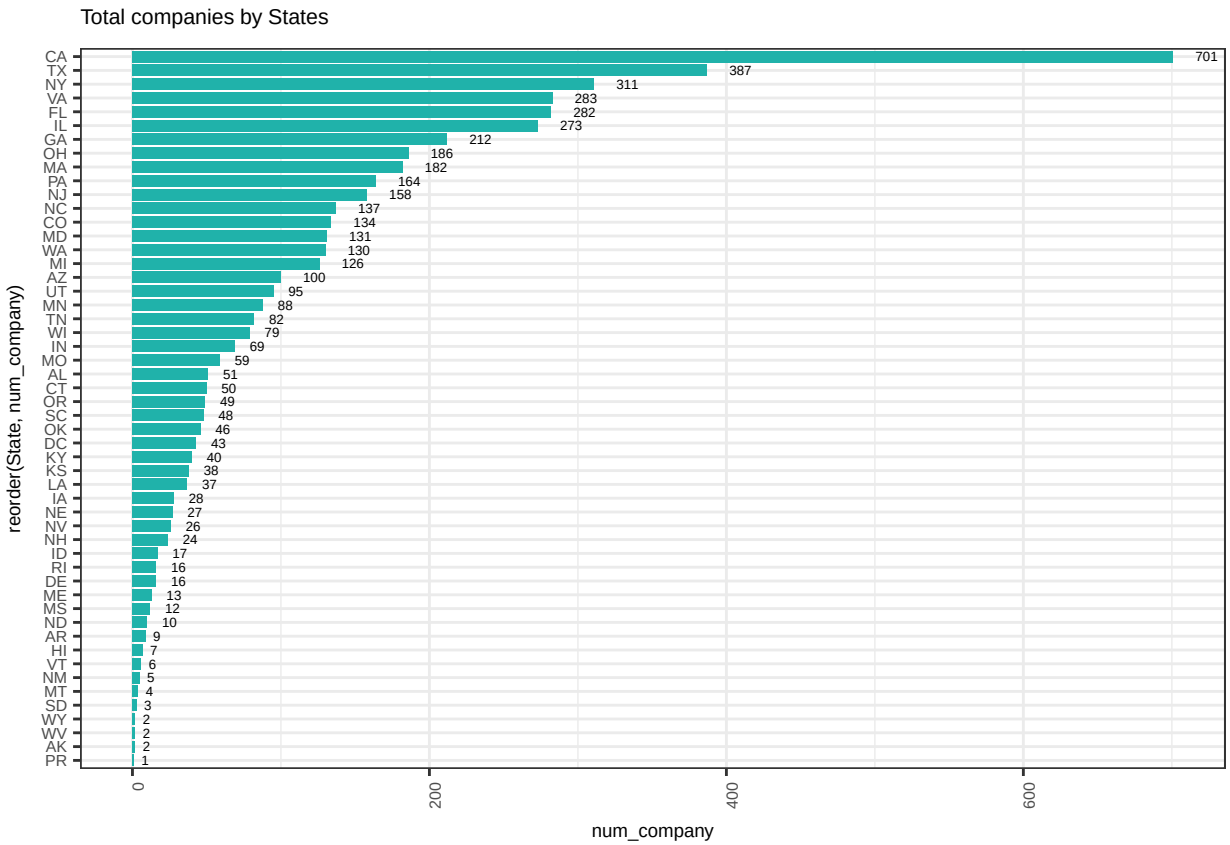



## Question 1

**"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."**

```
total_company_byState<-inc%>% group_by(State) %>% count(Name) %>% summarise(num_company = sum(n)) %>% a
total_company_byState<-total_company_byState[order(total_company_byState$num_company,decreasing = TRUE)
```

```
ggplot(data=total_company_byState,aes(x=reorder(State, num_company), y=num_company)) +
  geom_bar(stat="identity",width=0.8,fill = "lightseagreen")+
  geom_text(aes(label=num_company), hjust=-1, color="black", size=1.8)+
  ggtitle("Total companies by States") +
  theme_bw() +
  theme(text = element_text(size=7),
        axis.text.x = element_text(angle=90, hjust=1)) +
  coord_flip()
```



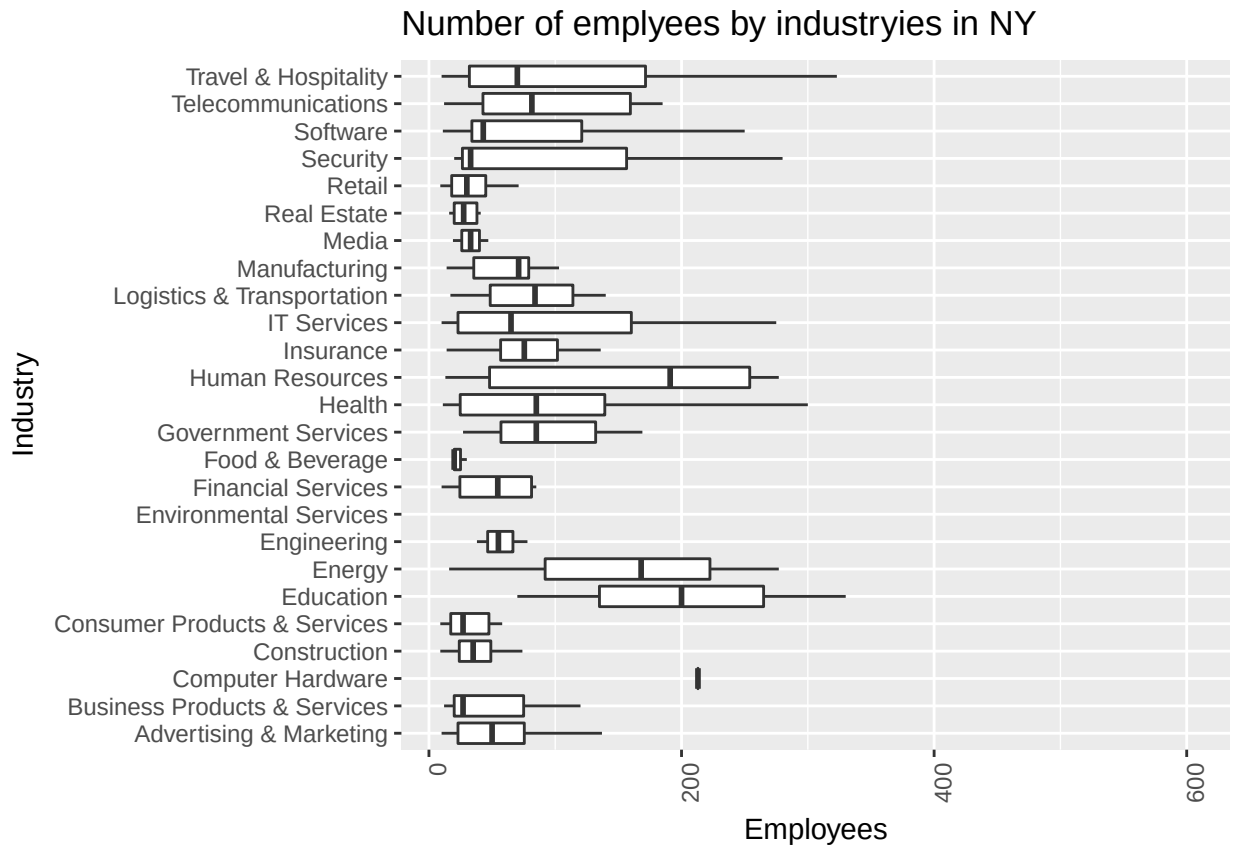

## Question 2

**"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."**

```
complete_df<-inc[complete.cases(inc),]
ny_df<- complete_df[complete_df$State=='FL',] %>% subset(select=c('Name','Industry','Revenue','Employees
```

```
ggplot(ny_df, aes(Industry, Employees)) +
  geom_boxplot(outlier.shape = NA) +
  ggtitle("Number of emplyees by industryies in NY ")+
  scale_y_continuous(limits = quantile(ny_df$Employees, c(0.05, 0.95)))+
  theme(axis.text.x=element_text(angle=90,hjust=1))+
  coord_flip()
```

```
## Warning: Removed 30 rows containing non-finite values (stat_boxplot).
```

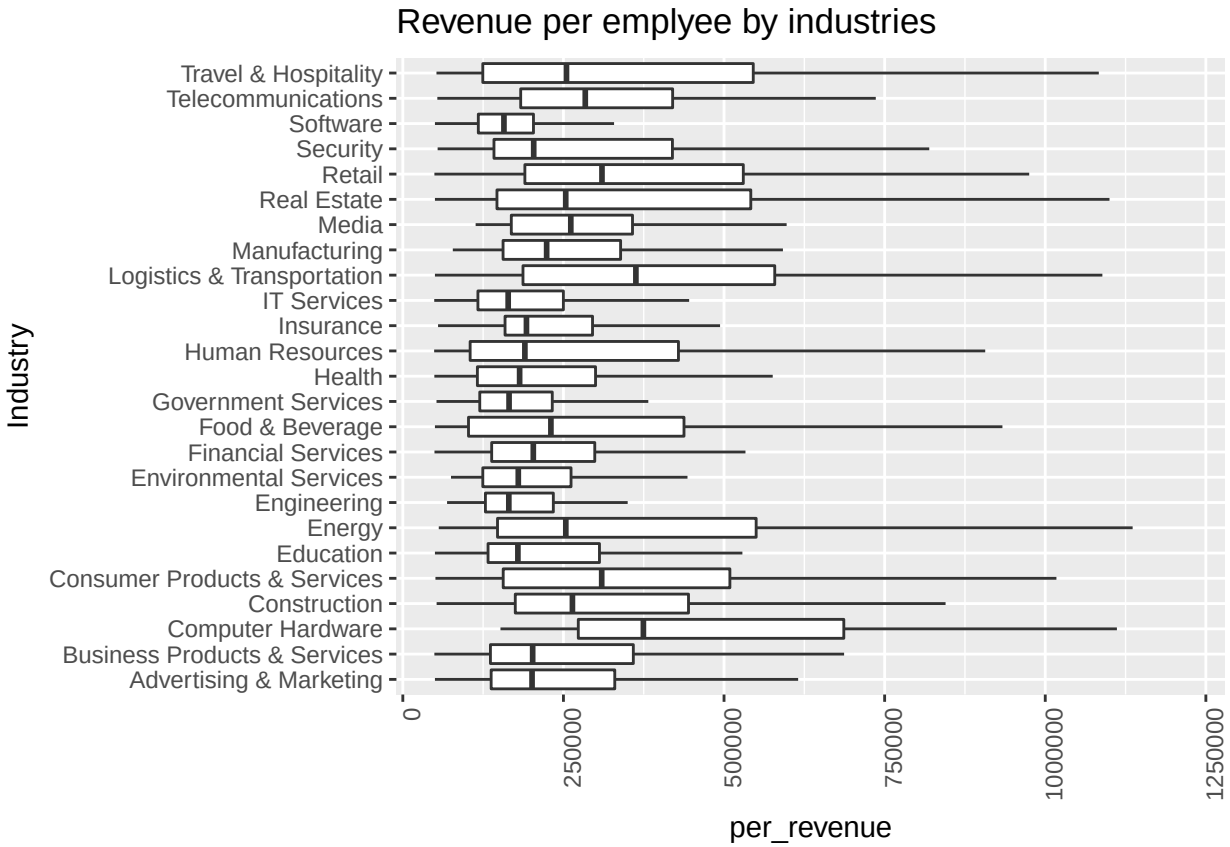



## Question 3

**"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."**

```
df<-complete_df%>% transform(per_revenue = Revenue / Employees)

ggplot(df, aes(Industry, per_revenue)) +
  geom_boxplot(outlier.shape = NA) +
  ggtitle("Revenue per emplyee by industries")+
  scale_y_continuous(limits = quantile(df$per_revenue, c(0.05, 0.95)))+
  theme(axis.text.x=element_text(angle=90,hjust=1))+coord_flip()
```

```
## Warning: Removed 499 rows containing non-finite values (stat_boxplot).
```


Revenue per emplyee by industries
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
0
250000
500000
750000
1000000
1250000
per_revenue