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/inc5000\_data.csv", header= TRUE)

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:

# Insert your code here, create more chunks as necessary  
sapply(inc, class)

## Rank Name Growth\_Rate Revenue Industry Employees   
## "integer" "factor" "numeric" "numeric" "factor" "integer"   
## City State   
## "factor" "factor"

cor(inc[,c(1,3,4,6)],use="complete.obs")

## Rank Growth\_Rate Revenue Employees  
## Rank 1.00000000 -0.397587831 0.082234923 0.0541156  
## Growth\_Rate -0.39758783 1.000000000 0.006328055 -0.0178689  
## Revenue 0.08223492 0.006328055 1.000000000 0.2779332  
## Employees 0.05411560 -0.017868897 0.277933213 1.0000000

Rank is weakly negatively correlated with Growth Rate,Number of employees are weakly positively correlated with Revenue. other correlatinos are very weak here. Rank,Growth Rate,Revenue,Employ are numeric and other variables are categorical.

## 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.

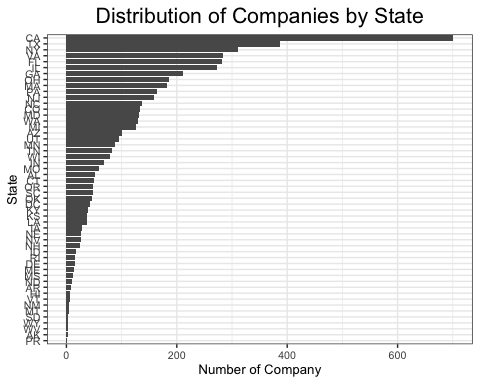
# Answer Question 1 here  
library(ggplot2)  
library(dplyr)

##   
## Attaching package: 'dplyr'

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

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

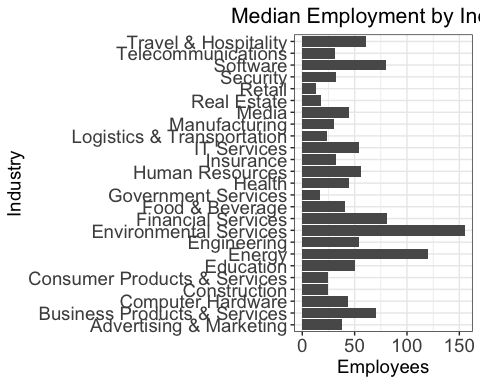
count(inc, State) %>%   
ggplot(aes(x = reorder(State,n),y=n)) +  
 geom\_col()+ coord\_flip()+ggtitle("Distribution of Companies by State")+xlab("State")+ylab("Number of Company")+theme\_bw()+theme(  
 plot.title = element\_text(size = 16,hjust = 0.5),  
 axis.text = element\_text(size =8),  
 axis.title = element\_text(size = 10)  
)



## 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.

# Answer Question 2 here  
# NY has 3rd most companies in the data  
  
NY\_df <- subset(inc, State=="NY")  
NY\_df <- NY\_df[complete.cases(NY\_df),]  
ggplot(NY\_df, aes(x=factor(Industry), y=Employees)) + stat\_summary(fun.y="median", geom="bar")+coord\_flip()+ggtitle("Median Employment by Industry")+xlab("Industry")+theme\_bw()+theme(  
 plot.title = element\_text(size = 16,hjust = 0.5),  
 axis.text = element\_text(size =14),  
 axis.title = element\_text(size = 14)  
)



## 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(inc, aes(x=factor(Industry), y=Revenue)) + stat\_summary(fun.y="mean", geom="bar")+coord\_flip()+ggtitle("Average Revenueby by Industry")+xlab("Industry")+theme\_bw()+theme(  
 plot.title = element\_text(size = 16,hjust = 0.5),  
 axis.text = element\_text(size =14),  
 axis.title = element\_text(size = 14)  
)

