HW_1- DATA VISUALIZATION

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:

Rank

This is the rank of the company, presumably some combination of revenue and growth rate. A visual analysis of this will be best to decipher how the ranking is done.

Growth Rate

This is the rate at which the companies are growing. Let us note that smaller companies will be more sensitive to change in size.

Revenue

This will be how much money the company is pulling in. keep in mind this isn't profit. We will see if this is relevant later.

City/State

There might be some correlation between location, we should be careful about developing causal relationships with location because it might be due to factors outside this dataset

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

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

```
#number of industries
inc$Industry<-as.factor(inc$Industry)
inc%>%count(Industry, sort=TRUE)
```

```
##                        Industry   n
## 1                   IT Services 733
## 2  Business Products & Services 482
## 3       Advertising & Marketing 471
## 4                        Health 355
## 5                      Software 342
## 6             Financial Services 260
## 7                  Manufacturing 256
## 8  Consumer Products & Services 203
## 9                        Retail 203
## 10          Government Services 202
## 11              Human Resources 196
## 12                 Construction 187
```

```
## 13   Logistics & Transportation 155
## 14               Food & Beverage 131
## 15           Telecommunications 129
## 16                       Energy 109
## 17                  Real Estate  96
## 18                    Education  83
## 19                  Engineering  74
## 20                     Security  73
## 21          Travel & Hospitality  62
## 22                        Media  54
## 23       Environmental Services  51
## 24                    Insurance  50
## 25             Computer Hardware  44
```

My first instinct is to say that "newer" industries make up a higher percentage of fast growing companies because of `IT Services`, but note that Software is in the middle of the pack. It is also hard to argue that `Business Products & Services` and `Advertising` are "new".

Maybe you could say that companies that provide services as opposed to products top the list.

summary by industry

Note that the percent of the list taken up by an industry does not translate to revenue. Keep in mind that revenue does not equal profit. A company providing a service might make more profit because they don't have to account for the cost of a product, and they might have less employees to pay for in order to produce the product. These things will be more easily investigated with graphs.

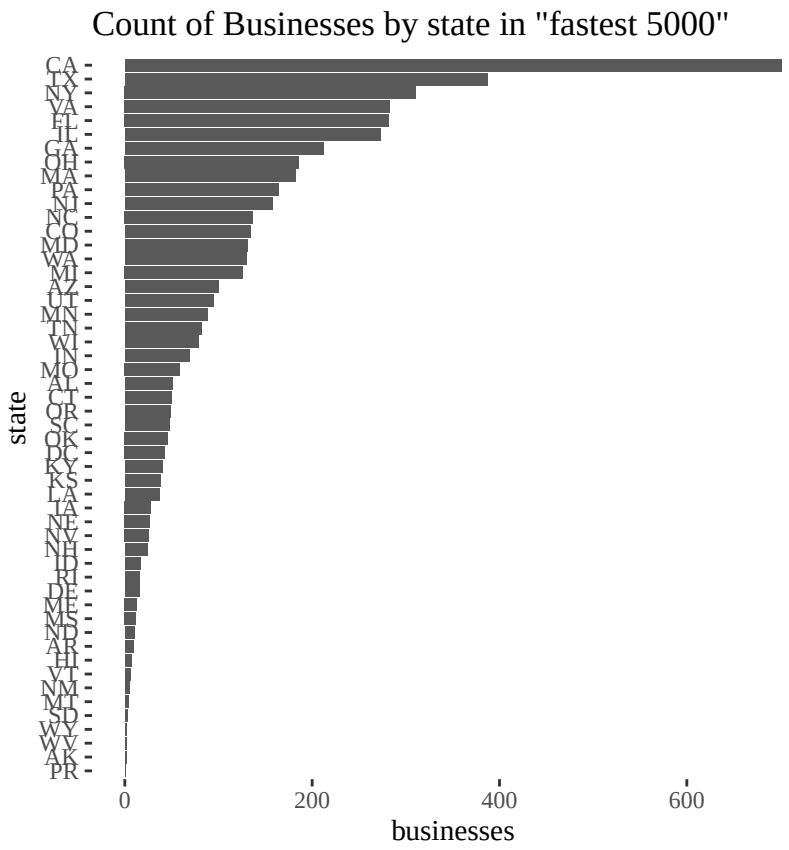
```
inc%>%group_by(Industry)%>%summarize(growth_rate=round(mean(Growth_Rate),1),revenue=mean(Revenue), empl
```

```
## # A tibble: 25 x 4
##    Industry                      growth_rate    revenue employees
##    <fct>                               <dbl>      <dbl>     <dbl>
##  1 Computer Hardware                     4.1 270129545.      221.
##  2 Energy                                9.6 126344954.      242.
##  3 Food & Beverage                       3.6  98559542.       NA
##  4 Logistics & Transportation            4.3  95745161.       NA
##  5 Consumer Products & Services          8.8  73676847.      224
##  6 Construction                          3.4  70450802.      156.
##  7 Telecommunications                    2.9  56855814.       NA
##  8 Business Products & Services          3.5  54705187.       NA
##  9 Security                              3.4  52230137.      562.
## 10 Environmental Services                2.1  51741176.      199.
## # ... with 15 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.

```
library(ggplot2)
library(ggthemes)
#convert State variable to factor
inc$State<-as.factor(inc$State)
df_state<-inc%>%count(State, sort=TRUE)
ggplot(df_state, aes(x=n, y=reorder(State,n)))+geom_bar(stat='identity')+xlab('businesses')+ylab('state
```



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.

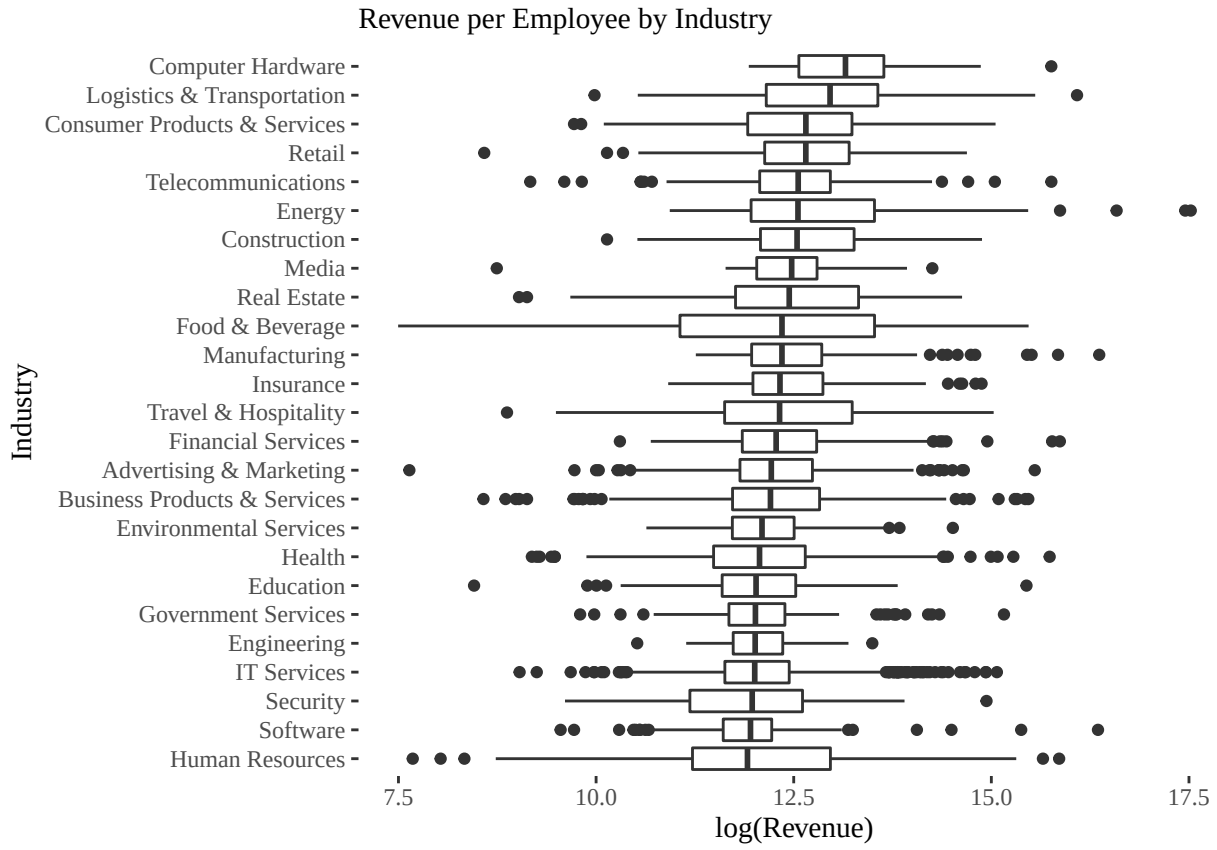
```
# Answer Question 2 here

#find 3rd highest state

state_3<-as.character(df_state$State)[3]
df_state_3<-inc%>%filter(State==state_3)%>%na.omit()

ggplot(df_state_3, aes(x=log(Employees),y=reorder(factor(Industry),Employees, FUN=median)))+geom_boxplo
```

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

#create revenue per employee statistic

df_rev_emp<-inc%>%drop_na(c(Employees,Revenue))%>%select(Industry,Employees,Revenue)%>%group_by(Industry

df_rev_emp<-inc%>%
  drop_na(c(Employees,Revenue))%>%
  select(Industry,Employees,Revenue)%>%
  group_by(Industry)%>%
  summarize(revenue=sum(Revenue),employees=sum(Employees))

df_rev_emp<-inc%>%
  drop_na(c(Employees,Revenue))%>%
  select(Industry,Employees,Revenue)%>%
  mutate(rev_emp=Revenue/Employees)

ggplot(df_state_3, aes(x=log(Employees),y=reorder(factor(Industry),Employees, FUN=median)))+geom_boxplo
```


```
ggplot(df_rev_emp, aes(x=log(rev_emp), y=reorder(factor(Industry),rev_emp, FUN=median)))+geom_boxplot()+
```


Revenue per Employee by Industry
Computer Hardware
Logistics & Transportation
Consumer Products & Services
Retail
Telecommunications
Energy
Construction
Media
Real Estate
Food & Beverage
Manufacturing
Insurance
Travel & Hospitality
Financial Services
Advertising & Marketing
Business Products & Services
Environmental Services
Health
Education
Government Services
Engineering
IT Services
Security
Software
Human Resources
Industry
7.5
10.0
12.5
15.0
17.5
log(Revenue)