Data608-HW1

Armenoush Aslanian-Persico

Principles of Data Visualization and Introduction to ggplot2

0. Setup

```
library(DT)
library(ggplot2)
library(jsonlite)
library(knitr)
library(plyr)
library(dplyr)
library(RCurl)
library(reshape2)
library(rmarkdown)
library(stringr)
library(tidyr)
```

1. Show Companies by States.

Create a graph that shows the distribution of companies in the dataset by State (ie how many are in each state).

1.2. Load Data from URL

```
companiescsv<-getURL("https://raw.githubusercontent.com/charleyferrari/CUNY_DATA608/master/lecture1/Data/inc5000_data.csv")

df1<- data.frame(read.csv(text=companiescsv))

datatable(df1, options = list( pageLength = 5, lengthMenu = c(5, 10),  initComplete = JS(
    "function(settings, json) {",
    "$(this.api().table().header()).css({'background-color': '#00838f', 'color': '#fff'});", "}"),
rownames=TRUE))
```

Show 5 entries Search:

	Rank	Name	Growth_Rate	Revenue	Industry	Employees	City	State
1	1	Fuhu	421.48	117900000	Consumer Products & Services	104	El Segundo	CA
2	2	FederalConference.com	248.31	49600000	Government Services	51	Dumfries	VA
3	3	The HCI Group	245.45	25500000	Health	132	Jacksonville	FL
4	4	Bridger	233.08	1900000000	Energy	50	Addison	TX
5	5	DataXu	213.37	87000000	Advertising & Marketing	220	Boston	MA

Showing 1 to 5 of 5,001 entries Previous 1 2 3 4 5 … 1001 Next

1.3. Group Companies by State

```
#show data structure
str(df1)
```

```
## 'data.frame':    5001 obs. of  8 variables:
##  $ Rank       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Name       : Factor w/ 5001 levels "(Add)ventures",..: 1770 1633 4423 690 1198 2839 4733 1468 1869 4968
 ...
##  $ Growth_Rate: num  421 248 245 233 213 ...
##  $ Revenue    : num  1.18e+08 4.96e+07 2.55e+07 1.90e+09 8.70e+07 ...
##  $ Industry   : Factor w/ 25 levels "Advertising & Marketing",..: 5 12 13 7 1 20 10 1 5 21 ...
##  $ Employees  : int  104 51 132 50 220 63 27 75 97 15 ...
##  $ City       : Factor w/ 1519 levels "Acton","Addison",..: 391 365 635 2 139 66 912 1179 131 1418 ...
##  $ State      : Factor w/ 52 levels "AK","AL","AR",..: 5 47 10 45 20 45 44 5 46 41 ...
```

```
#group companies by state
groupbystate <- ddply(df1, .(State), summarize, count = length(Rank))
```

```
## Warning in split_indices(as.integer(splitv), attr(splitv, "n")):
## '.Random.seed' is not an integer vector but of type 'NULL', so ignored
```

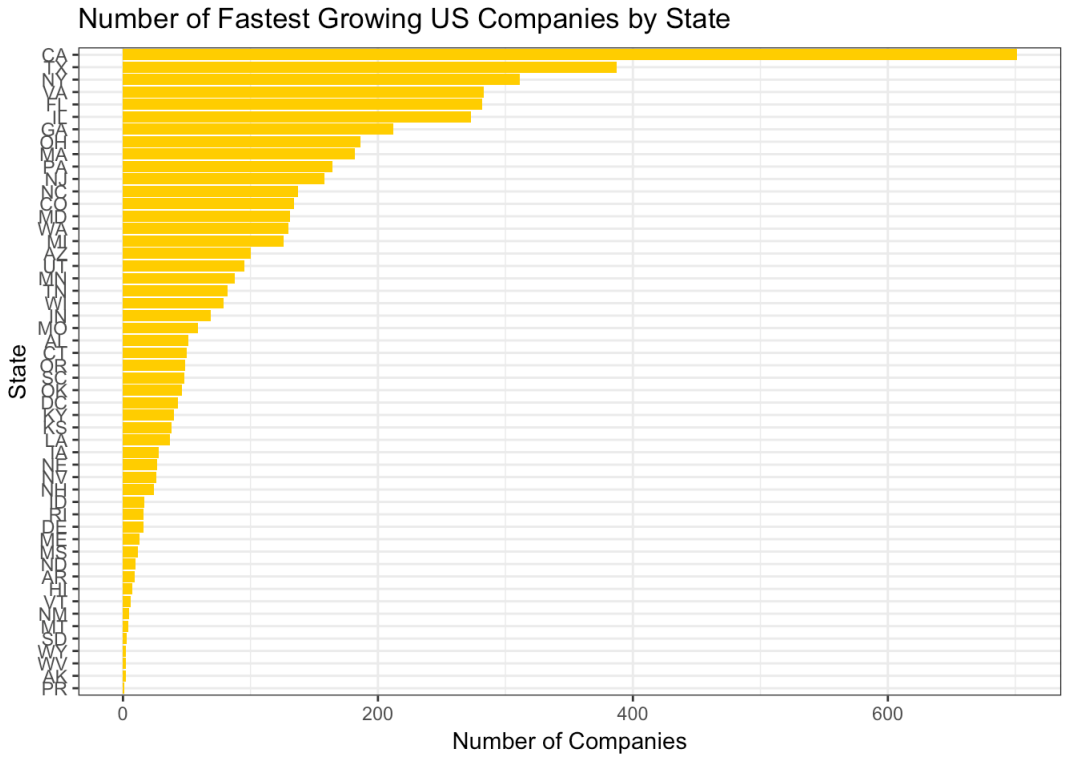
```
orderbystate <- groupbystate[order(groupbystate$count), ]

#order by state with most companies
orderbystate$State <- factor(orderbystate$State, levels = orderbystate$State)
```

1.4 Show Graph of Companies by State.

```
#Create graph of companies by state
Figure1 <-
  ggplot(orderbystate, aes(x = State, y = count)) +
  geom_bar(stat = "identity", fill="#ffca28")+
  coord_flip() +
  theme_bw()+
  ggtitle("Number of Fastest Growing US Companies by State") +
  labs(x="State",y="Number of Companies")

Figure1
```




```
ggsave("Figure1.jpg")
```

```
## Saving 7 x 5 in image
```

2. Show Average Employment by Industry in the Third Ranked State.

Let's 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 employ. Create a plot of average employment by industry for companies in this state (only use cases with full data (user R's complete.cases() function). Your graph should show how variable the ranges are, and exclude outliers.

2.1 Sort States by Count of Companies

```
allstates <- as.data.frame(table(df1$State))
colnames(allstates) <- c("State", "Count")

# sort by descending count
sortstates <- allstates[order(-allstates$Count),]
head(sortstates)
```

```
##    State Count
## 5     CA   701
## 45    TX   387
## 35    NY   311
## 47    VA   283
## 10    FL   282
## 15    IL   273
```

2.2 Select State with Third Most Companies

```
#subset third highest value
x <- sort(allstates$Count, TRUE)[3]
filter(allstates, Count == x)
```

```
##   State Count
## 1    NY   311
```

```
#filter by resulting state
nys <- filter(df1, State == "NY")

#subset complete cases only
nys <- nys[complete.cases(nys),]

#show data structure
str(nys)
```

```
## 'data.frame':    311 obs. of  8 variables:
##  $ Rank       : int  26 30 37 38 48 70 71 124 126 153 ...
##  $ Name       : Factor w/ 5001 levels "(Add)ventures",..: 529 3822 4972 1037 912 19 2608 3591 3684 3668 ...
##  $ Growth_Rate: num  84.4 73.2 67.4 67 53.6 ...
##  $ Revenue    : num  13700000 8100000 18000000 7100000 5900000 27900000 6900000 11500000 9800000 15400000 ...
##  $ Industry   : Factor w/ 25 levels "Advertising & Marketing",..: 5 1 1 1 10 1 1 24 21 25 ...
##  $ Employees  : int  17 79 27 89 32 75 42 28 17 42 ...
##  $ City       : Factor w/ 1519 levels "Acton","Addison",..: 929 929 929 929 1135 929 929 929 574 162 ...
##  $ State      : Factor w/ 52 levels "AK","AL","AR",..: 35 35 35 35 35 35 35 35 35 35 ...
```

2.3 Calculate Average Employment by Industry

```
#Create summary columns for mean and count
nysjobs <- ddply(nys, .(Industry), summarize,
    meanemployees = round(mean(Employees),0),
    sumemployees = sum(Employees),
    countemployers = length(Employees),
    meanrevenueperemployee = round(mean(Revenue/Employees),2),
    meangrowthrate = round(mean(Growth_Rate),2)
    )

#Show industries with highest average employee count
nysjobs <- nysjobs[order(nysjobs$meanemployees, decreasing = TRUE),]

#Show calculated data
datatable(nysjobs, options = list( pageLength = 5, lengthMenu = c(5, 10),  initComplete = JS(
    "function(settings, json) {",
    "$(this.api().table().header()).css({'background-color': '#1565c0', 'color': '#fff'});", "}"),
rownames=TRUE))
```

Show 5 **entries** **Search:**

	Industry	meanemployees	sumemployees	countemployers	meanrevenueperemployee	meangrowthrate
2	Business Products & Services	1492	38804	26	527816.95	2.04
5	Consumer Products & Services	626	10647	17	382942.57	7.96
25	Travel & Hospitality	548	3834	7	282089.82	4.97
14	Human Resources	438	4813	11	337366.32	2.97
23	Software	246	3197	13	143749.03	1.15

Showing 1 to 5 of 25 entries

Previous 1 2 3 4 5 Next

```
#Create industry factor
nys$Industry <- factor(nys$Industry, levels = nysjobs$Industry)

#Show nys data
datatable(nys, options = list( pageLength = 5, lengthMenu = c(5, 10),  initComplete = JS(
    "function(settings, json) {",
    "$(this.api().table().header()).css({'background-color': '#283593', 'color': '#fff'});", "}"),
rownames=TRUE))
```

Show 5 **entries** **Search:**

	Rank	Name	Growth_Rate	Revenue	Industry	Employees	City	State
1	26	BeenVerified	84.43	13700000	Consumer Products & Services	17	New York	NY
2	30	Sailthru	73.22	8100000	Advertising & Marketing	79	New York	NY
3	37	YellowHammer	67.4	18000000	Advertising & Marketing	27	New York	NY
4	38	Conductor	67.02	7100000	Advertising & Marketing	89	New York	NY
5	48	Cinium Financial Services	53.65	5900000	Financial Services	32	Rock Hill	NY

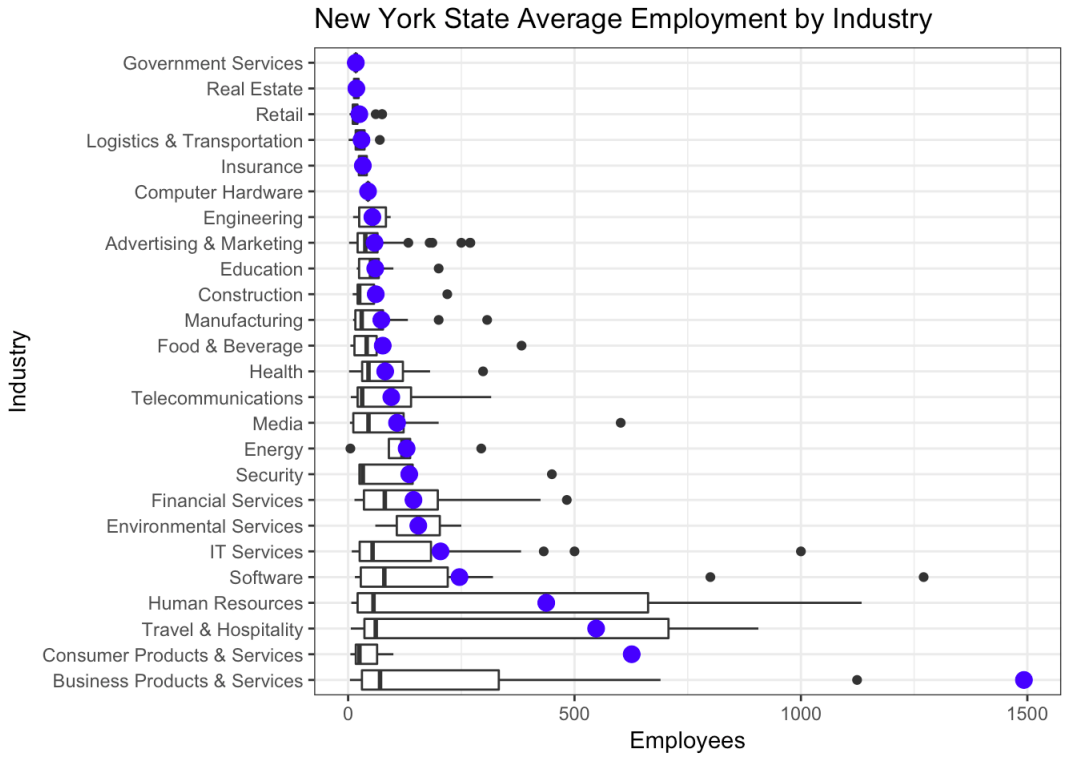
Showing 1 to 5 of 311 entries

Previous 1 2 3 4 5 … 63 Next

2.4 Show Graph of Average Employment by Industry.

```
#Create graph of jobs by industry
Figure2 <-
ggplot(nys, aes(x = Industry, y = Employees)) +
  geom_boxplot() +
  stat_summary(fun.y = mean, geom = "point", shape = 20, size = 5, color = "blue") +
  theme_bw()+
  ggtitle("New York State Average Employment by Industry")

#select limits to hide outliers
Figure2<- Figure2 + coord_flip(ylim = c(0,1500))
Figure2
```




```
ggsave("Figure2.jpg")
```

```
## Saving 7 x 5 in image
```

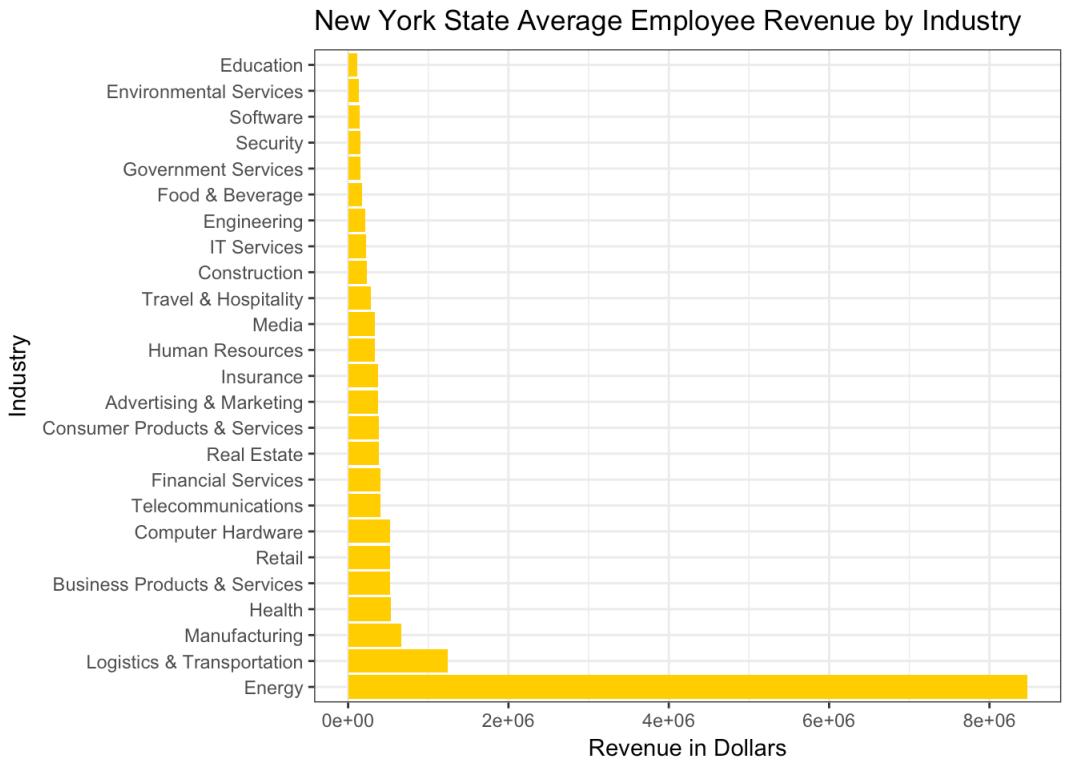
3. Show Highest Revenue Industries.

Now imagine you work for an investor and want to see which industries generate the most revenue per employee. Create a chart makes this information clear.

```
nysjobs1 <- nysjobs[order(nysjobs$meanrevenueperemployee, decreasing = TRUE),]

#Create graph of jobs by revenue
Figure3 <-
ggplot(nysjobs1, aes(x = reorder(Industry,-meanrevenueperemployee), y = meanrevenueperemployee)) +
  geom_bar(stat = "identity", fill="#ffca28")+
  theme_bw()+
  ggtitle("New York State Average Employee Revenue by Industry")+
  labs(y="Revenue in Dollars",x="Industry")

Figure3<- Figure3 + coord_flip()
Figure3
```




```
ggsave("Figure3.jpg")
```

```
## Saving 7 x 5 in image
```

4. Conclusion

California, Texas, New York, Virginia and Florida have the highest number of fastest growing companies in this dataset.

Business Products, Consumer Products, and Travel/Hospitality have the highest average number of employees per company. The boxplots indicate that some outliers influenced the means.

Energy, Transportation, Manufacturing, Health, and Business Products have the highest calculated revenue per employee.

The year and source of this dataset is unknown. If multiple years of data were available, this would be a good source of information about industry growth and job availability trends for job seekers. The growth rate variable may be an indicator of this, but would need defining.