DATA 608 HW1

Hui Han

February 12, 2019

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
```

```
str(inc)
```

```
## '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 28
39 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 91
2 1179 131 1418 ...
##  $ State      : Factor w/ 52 levels "AK","AL","AR",..: 5 47 10 45 20 45 44 5 46 4
1 ...
```

1. Create a graph that shows the distribution of companies in the dataset by State (i.e. how many are in each state). There are a lot of States, so consider which axis you should use assuming I am using a ‘portrait’ oriented screen

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
require(dplyr)
```

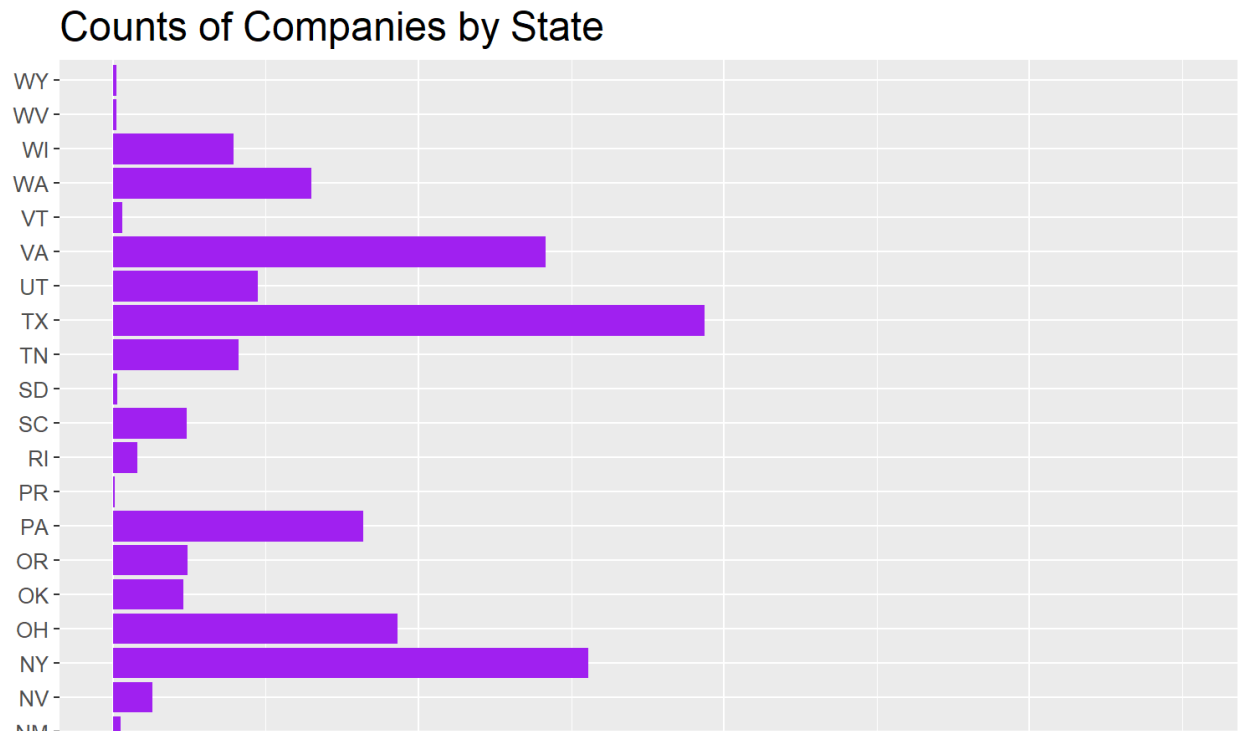
```
## Loading required package: dplyr
```

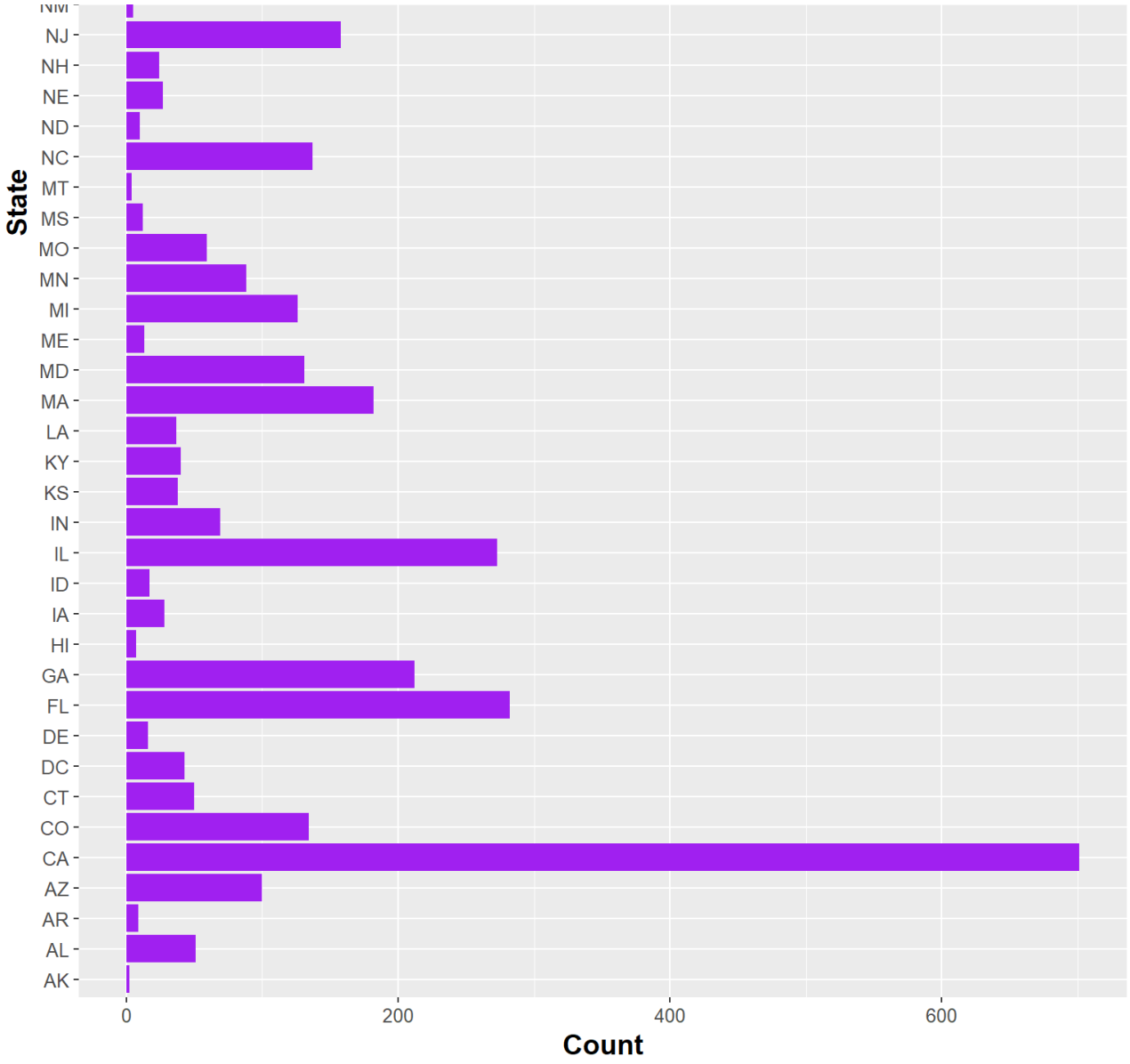
```
##
## Attaching package: 'dplyr'
```

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

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

```
p <- ggplot(inc, aes(factor(State))) + geom_bar(fill="purple")
p <- p + coord_flip()
p <- p + theme(text = element_text(size=12), axis.title=element_text(size=14,face="bol
d"))
p <- p + labs(title = "Counts of Companies by State", x= "State", y= "Count")
p <- p + theme(plot.title = element_text(size=18))
p
```





2.For the State with the 3rd most companies, create a plot of average employment by industry for companies in this state (only use cases with full data. Your graph should show how variable the ranges are, and exclude outliers.

```
counts <- as.data.frame(table(inc$State))
colnames(counts) <- c("State", "Count")
head(counts)
```

```
##   State Count
## 1    AK     2
## 2    AL    51
## 3    AR     9
## 4    AZ   100
## 5    CA   701
## 6    CO   134
```

Find the 3rd most companies by state

```
x <- sort(counts$Count, TRUE)[3]
filter(counts, Count == x)
```

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

Remove incomplete cases

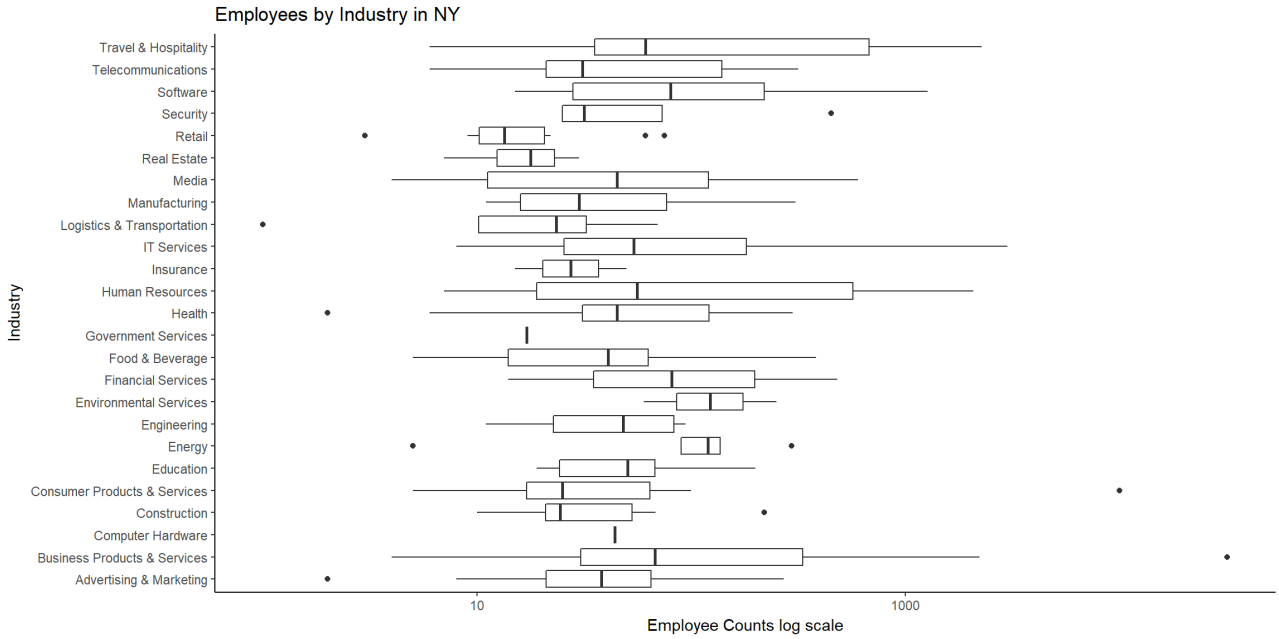
```
ny_inc <- filter(inc, State == "NY")
ny_inc <- ny_inc[complete.cases(ny_inc),]
glimpse(ny_inc)
```

```
## Observations: 311
## Variables: 8
## $ Rank        <int> 26, 30, 37, 38, 48, 70, 71, 124, 126, 153, 174, 21...
## $ Name        <fctr> BeenVerified, Sailthru, YellowHammer, Conductor, ...
## $ Growth_Rate <dbl> 84.43, 73.22, 67.40, 67.02, 53.65, 44.99, 44.85, 2...
## $ Revenue     <dbl> 13700000, 8100000, 18000000, 7100000, 5900000, 279...
## $ Industry    <fctr> Consumer Products & Services, Advertising & Marke...
## $ Employees   <int> 17, 79, 27, 89, 32, 75, 42, 28, 17, 42, 99, 119, 2...
## $ City        <fctr> New York, New York, New York, New York, Rock Hill...
## $ State       <fctr> NY, NY, NY, NY, NY, NY, NY, NY, NY, NY, NY, NY, N...
```

try box plot

```
ny_inc <- ny_inc[c("Industry","Employees")]
IM <- aggregate(ny_inc$Employees, by=list(ny_inc$Industry),
  FUN=mean, na.rm=TRUE)
colnames(IM) <- c("Industry","EmployeeMean")

p <- ggplot(ny_inc,aes(ny_inc$Industry, ny_inc$Employees))+geom_boxplot()+theme_classic
()+scale_y_log10()+labs(title="Employees by Industry in NY", x="Industry", y="Employee
Counts log scale")
p+coord_flip()
```



3. Generate a chart showing which industries generate the most revenue per employee.

```
#remove incomplete cases
inc_rev <- inc[complete.cases(inc),]
# Create a new column rev_per_em = revenue/employee using mutate
inc_rev <- inc_rev %>% mutate(rev_per_em = Revenue / Employees)
glimpse(inc_rev)
```

```
## Observations: 4,989
## Variables: 9
## $ Rank        <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,...
## $ Name        <fctr> Fuhu, FederalConference.com, The HCI Group, Bridg...
## $ Growth_Rate <dbl> 421.48, 248.31, 245.45, 233.08, 213.37, 179.38, 17...
## $ Revenue     <dbl> 1.179e+08, 4.960e+07, 2.550e+07, 1.900e+09, 8.700e...
## $ Industry    <fctr> Consumer Products & Services, Government Services...
## $ Employees   <int> 104, 51, 132, 50, 220, 63, 27, 75, 97, 15, 149, 16...
## $ City        <fctr> El Segundo, Dumfries, Jacksonville, Addison, Bost...
## $ State       <fctr> CA, VA, FL, TX, MA, TX, TN, CA, UT, RI, VA, CA, F...
## $ rev_per_em  <dbl> 1133653.8, 972549.0, 193181.8, 38000000.0, 395454....
```

make a plot

```
p2 <- ggplot(inc_rev) +  geom_bar(aes(Industry, rev_per_em, fill = Industry), position
= "dodge", stat = "summary", fun.y = "mean", fill="purple")
p2 <- p2 + coord_flip()
p2 <- p2 + theme(legend.position="none")
p2 <- p2 + theme(text = element_text(size=12), axis.title=element_text(size=14,face="bo
ld"))
p2 <- p2 + labs(title = "Average Revenue per Employees by Industry", x= "Industry", y=
"Average Revenue per Employees")
p2 <- p2 + theme(plot.title = element_text(size=18))
p2
```