Data608-Module-01-HomeWork

Vinayak Kamath

02/14/2021

```
#Loading additional libraries
library(dplyr)
library("ggplot2")
```

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
##  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:

1. *=> Using the str function we can see there are 50001 observations in the data set and there are 8 variables:*

```
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 2839 4733 1468
##  $ 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 14
##  $ State      : Factor w/ 52 levels "AK","AL","AR",..: 5 47 10 45 20 45 44 5 46 41 ...
```

2. *=> We can see that their are 52 unique values in the variable State and this includes Puerto Rico (PR) and Washington D.C. (DC) in addition to the 50 states:*

```
str(inc$State)
```

```
##  Factor w/ 52 levels "AK","AL","AR",..: 5 47 10 45 20 45 44 5 46 41 ...
```

3. *=> Not all observations have count of employees in it. There are 12 observations with NA values:*

```
summary(inc$Employees)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     1.0    25.0    53.0   232.7   132.0 66803.0      12
```

```
filter(inc, is.na(Employees))
```

```
##    Rank                              Name Growth_Rate   Revenue
## 1   183           First Flight Solutions       22.32   2700000
## 2  1064                         Popchips        3.98  93300000
## 3  1124                      Vocalocity         3.72  42900000
## 4  1653                     Higher Logic        2.36   6000000
## 5  1686      Global Communications Group        2.30   3600000
## 6  2197              JeffreyM Consulting        1.68  12100000
## 7  2743               Excalibur Exhibits        1.27   9900000
## 8  3001       Heartland Business Systems        1.12 156300000
## 9  3978                             SSEC        0.68  80400000
## 10 4112 Carolinas Home Medical Equipment        0.64   3300000
## 11 4566                         Oakbrook        0.48   8900000
## 12 4968                   Popcorn Palace        0.35   5500000
##                        Industry Employees          City State
## 1    Logistics & Transportation        NA  Emerald Isle    NC
## 2               Food & Beverage        NA San Francisco    CA
## 3            Telecommunications        NA       Atlanta    GA
## 4                      Software        NA    Washington    DC
## 5            Telecommunications        NA     Englewood    CO
## 6  Business Products & Services        NA      Bellevue    WA
## 7  Business Products & Services        NA       houston    TX
## 8                   IT Services        NA  Little Chute    WI
## 9                 Manufacturing        NA       Horsham    PA
## 10                       Health        NA      Matthews    NC
## 11                  Real Estate        NA       Madison    WI
## 12              Food & Beverage        NA Schiller Park    IL
```

4. *=> The max value fro the variable `Rank` shown is 5000; where as there are 5001 observations; telling the variable `Rank` is not unique:*

```
summary(inc$Rank)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1    1252    2502    2502    3751    5000
```

```
data1 <- inc %>%
  group_by(Rank) %>%
  summarise(n = n()) %>%
  filter(n != 1)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
data1
```

```
## # A tibble: 2 x 2
##    Rank     n
##   <int> <int>
## 1  3424     2
## 2  5000     2
```

```
# we can see there are 2 records each for the Rank 3424 and 5000
```

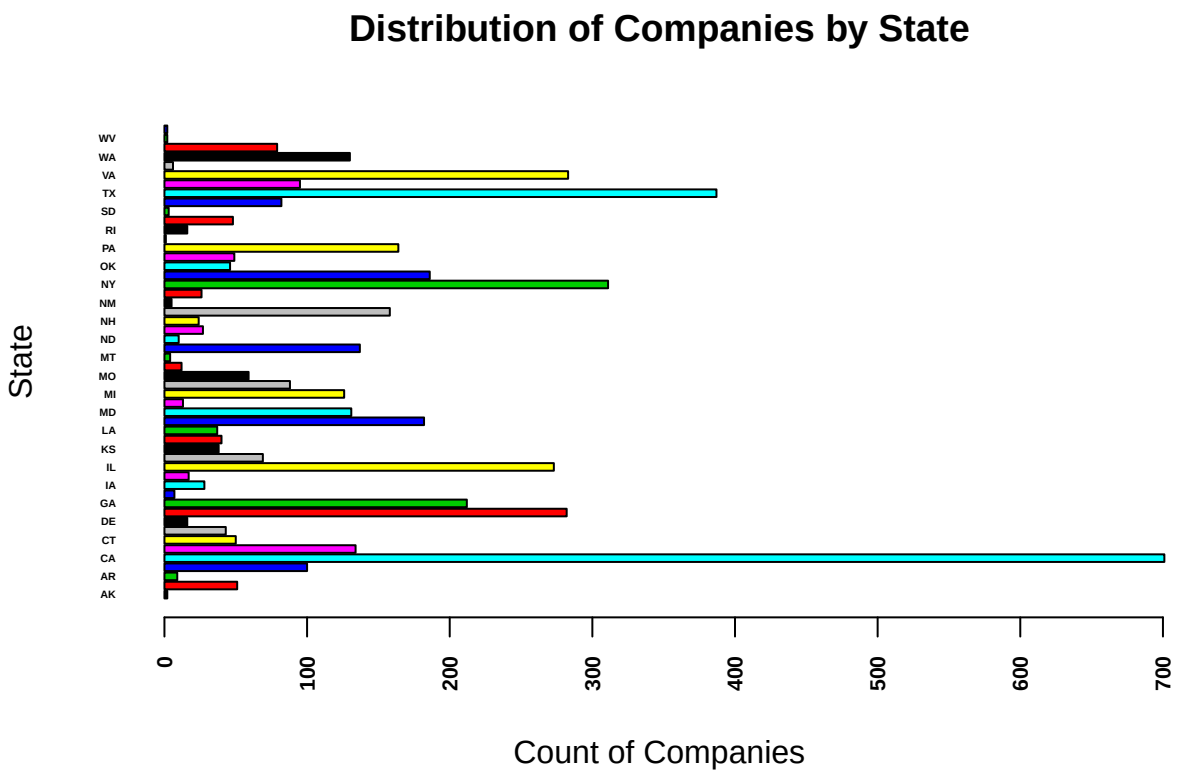
Question 1

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.

```
# Grouping the data to the count by State
data1 <- inc %>%
  group_by(State) %>%
  summarise(n = n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
# Plot the bar chart
barplot(data1$n,names.arg=data1$State,xlab="Count of Companies",ylab="State",col=data1$State, main="Dis
        ,  las=2 , cex.names=.3,  space =0.2, font=2 , cex.axis = .7)
```



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.

```
# R's `complete.cases()` function to get the observations with full data. 12 records that had NA will b
complete_data <- filter(inc, complete.cases(inc) )
str(complete_data)
```

```
## 'data.frame':    4989 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 186
##  $ 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 ...
```

```
# Grouping  by State to get the state with 3rd most companies in it:
complete_data_3rd_state <- complete_data %>%
  group_by(State) %>%
  summarise(n = n()) %>%
  mutate(ranks = order(order(n, decreasing=T))) %>%
  filter(ranks == 3)
```

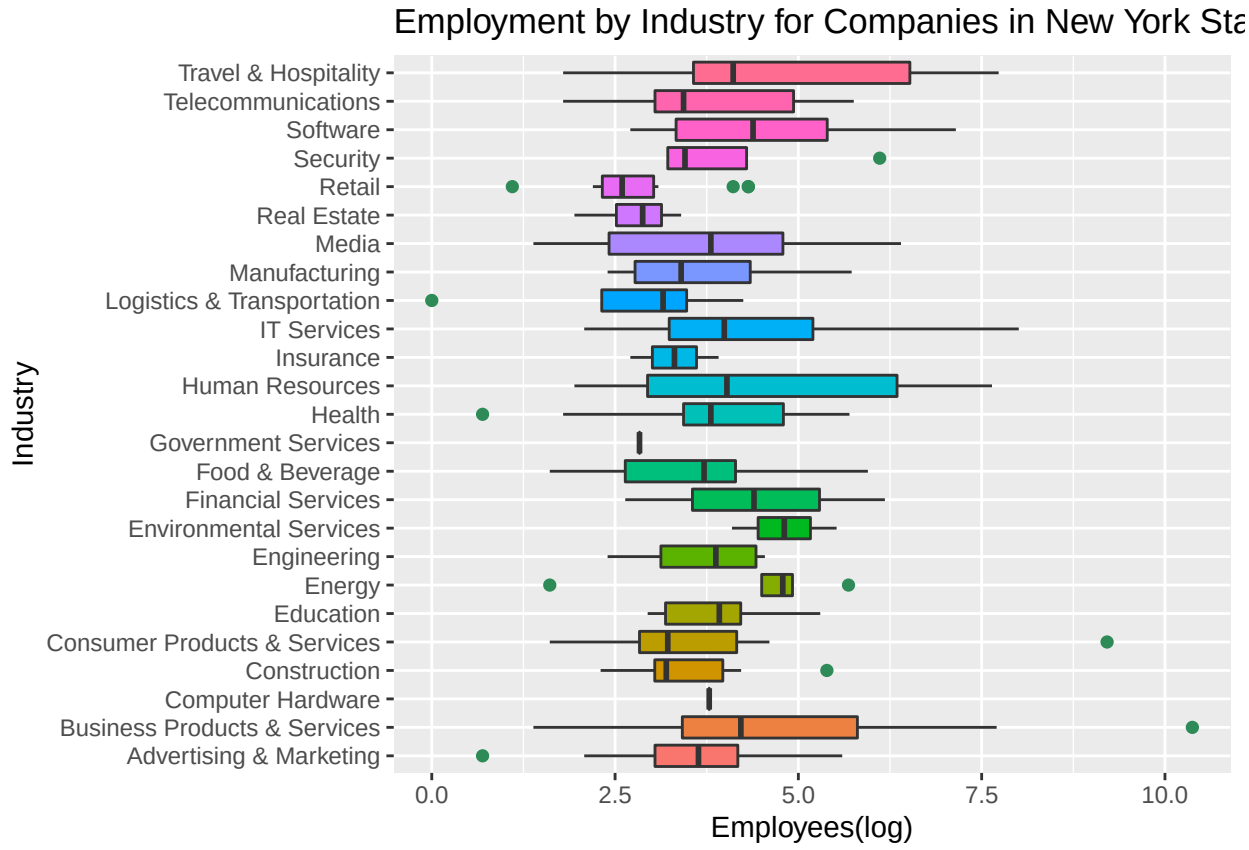
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
# We can see NY as the state with 311 companies in it:
complete_data_3rd_state
```

```
## # A tibble: 1 x 3
##   State     n ranks
##   <fct> <int> <int>
## 1 NY      311     3
```

```
# Getting the data set to plot having total, mean and median of employees grouped by Industry:
data2 <- complete_data %>%
  filter(State == complete_data_3rd_state$State)

# Ggplot2 plot that shows the average and/or median employment by industry for companies in this state:
ggplot(data2, aes(x = Industry, y = log(Employees) , fill= Industry)  ) +
  geom_boxplot(outlier.colour="seagreen", outlier.shape=16, outlier.size=2)    +
  theme(legend.position = "none") +
  coord_flip() +
  labs(title="Employment by Industry for Companies in New York State", y="Employees(log)", x="Industry"
```



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.

```
# Grouping the data to the count by industries
data3 <- complete_data %>%
  group_by(Industry) %>%
  summarise(total_employees = sum(Employees), total_revenue = sum(Revenue)) %>%
  mutate(revenue_per_employee = total_revenue/total_employees)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
# Check the Data:
data3
```

```
## # A tibble: 25 x 4
##    Industry                    total_employees total_revenue revenue_per_employ~
##    <fct>                                 <int>         <dbl>               <dbl>
##  1 Advertising & Marketing               39731    7785000000             195943.
##  2 Business Products & Servic~          117357   26345900000             224494.
```

```
##  3 Computer Hardware                 9714   11885700000          1223564.
##  4 Construction                     29099   13174300000           452741.
##  5 Consumer Products & Servic~      45464   14956400000           328972.
##  6 Education                         7685    1139300000           148250.
##  7 Energy                           26437   13771600000           520921.
##  8 Engineering                      20435    2532500000           123930.
##  9 Environmental Services           10155    2638800000           259852.
## 10 Financial Services               47693   13150900000           275741.
## # ... with 15 more rows
```

```
# Plot the bar chart
barplot(data3$revenue_per_employee,names.arg=data3$Industry,col=data3$Industry, main="Distribution of Re
```