Data 608 : Module1

Principles of Data Visualization and Introduction to ggplot2

Ramnivas Singh

02/14/2022

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", heade
r= 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         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
```

```
# Insert your code here, create more chunks as necessary
# Add required libraries
library(DT)
library(tidyverse)
library(ggplot2)
```

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
# Adding a table to seach and sort data in tabular form.
datatable(inc,
          options = list(pageLength = 10,searching = TRUE, filter=TRUE,
                         pageLength = FALSE),rownames = FALSE)
```

Show 10 entries

Search:

Rank	Name	Growth_Rate	Revenue	Industry	Employees	City	State

Rank	Name	Growth_Rate	Revenue	Industry	Employees	City	State
1	Fuhu	421.48	117900000	Consumer Products & Services	104	El Segundo	CA
2	FederalConference.com	248.31	49600000	Government Services	51	Dumfries	VA
3	The HCI Group	245.45	25500000	Health	132	Jacksonville	FL
4	Bridger	233.08	1900000000	Energy	50	Addison	TX
5	DataXu	213.37	87000000	Advertising & Marketing	220	Boston	MA
6	MileStone Community Builders	179.38	45700000	Real Estate	63	Austin	TX
7	Value Payment Systems	174.04	25500000	Financial Services	27	Nashville	TN
8	Emerge Digital Group	170.64	23900000	Advertising & Marketing	75	San Francisco	CA
9	Goal Zero	169.81	33100000	Consumer Products & Services	97	Bluffdale	UT
10	Yagoozon	166.89	18600000	Retail	15	Warwick	RI

Showing 1 to 10 of 5,001 entries

Previous 1 2 3 4 5 … 501 Next

```
# Next I like calculate standard deviation and IQR to understand data skewness
sd(inc$Revenue)
```

```
## [1] 240542281
```

```
sd(inc$Growth_Rate)
```

```
## [1] 14.12369
```

```
sd(inc$Employees, na.rm = TRUE)
```

```
## [1] 1353.128
```

```
#Lets do IQR in case the data is skewed
IQR(inc$Revenue)
```

```
## [1] 23500000
```

```
IQR(inc$Growth_Rate)
```

```
## [1] 2.52
```

```
IQR(inc$Employees, na.rm = TRUE)
```

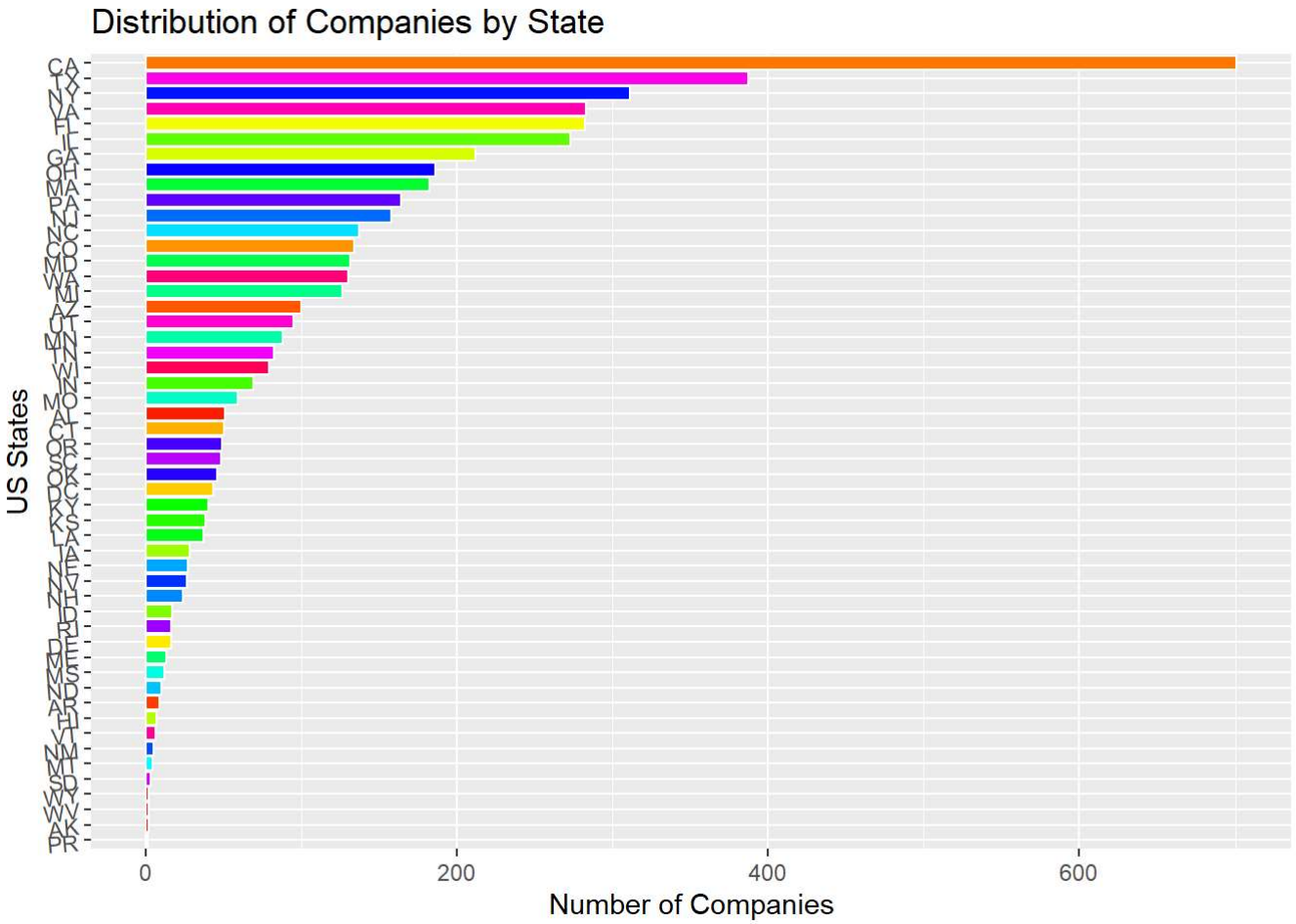
```
## [1] 107
```

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
# Ask : Give distributions of categorical data: States or Industries.
# We will use a bar which seems to be an intuitive way to present the information. Also bar will provide an comparable view
 in plot

state_counts <- inc$State %>% table() %>% as.data.frame(stringsAsFactors=FALSE)
colnames(state_counts) <- c('State', 'Frequency')
ggplot(state_counts, aes(x=reorder(State, Frequency),y=Frequency, color=State)) +
  geom_bar(stat='identity', color = 'white', fill=rainbow(52)) +labs(title="Distribution of Companies by State", x="US State
s", y="Number of Companies")+
  coord_flip() +theme(axis.text.y = element_text(angle = 5))
```

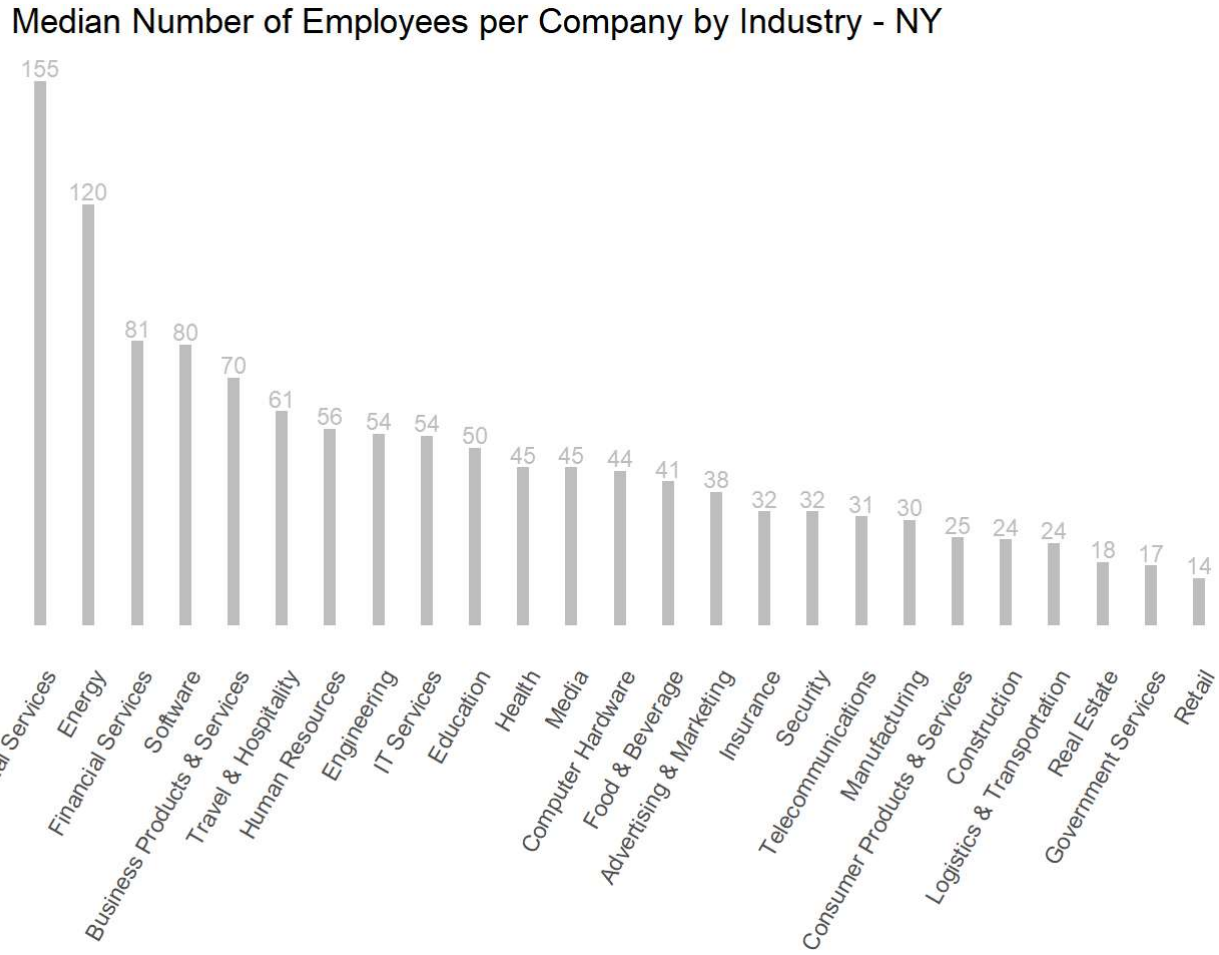


Quesiton 2

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
# By plot in question 1, we can see NY is 3rd state with most of the companies.
NY <- subset(inc, State=="NY")
NY <- NY[complete.cases(NY), ]

ggplot(NY %>% group_by(Industry) %>%
         summarise(`Median Employees` = median(Employees))) +
  geom_col(
    aes(x=reorder(Industry, -`Median Employees`), y = `Median Employees`),
    fill = "grey",
    width = 0.25) +
  geom_text(
    aes(x = Industry, y = `Median Employees`, label=round(`Median Employees`, digits = 0)),
    vjust=-0.25,
    size=3,
    color="gray") +
  theme(axis.text.x = element_text(angle = 60, hjust = 1),
        axis.text.y = element_blank(),
        axis.title=element_blank(),
        axis.ticks = element_blank(),
        panel.grid = element_blank(),
        panel.background = element_blank(),
        plot.margin = margin(1, 1, 5, 45)
        ) +
  labs(title = "Median Number of Employees per Company by Industry - NY")
```



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
inc %>%
  filter(complete.cases(.)) %>%
  group_by(Industry) %>%
  summarise(Employees_n = sum(Employees),
            Revenue_n = sum(Revenue)) %>%
  mutate(Revenue_Per_Employee = Revenue_n / Employees_n) %>%
  ggplot(., aes(x= Revenue_Per_Employee, y = reorder(Industry, Revenue_Per_Employee))) +
  geom_bar(stat = "identity", ) +
  labs(title = "Distribution of Revenue/ Employee by Industry", y = "Industry in NY",
       x = "Revenue/ Employee in NY")
```