R Notebook : HW1

Samriti Malhotra

Feb 09, 2020

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:

```
library(ggplot2)
library(dplyr)
library(kableExtra)
library(sqldf)
```

```
inc <- read.csv("https://raw.githubusercontent.com/charleyferrari/CUNY_DATA_608/master/module1/Data/inc5
```

And lets preview this data:

```
kable(head(inc)) %>%
  kable_styling(bootstrap_options = c("striped","hover","condensed","responsive"),full_width   = F,posi
  row_spec(0, background ="gray")
```

Rank	Name	Growth_Rate	Revenue	Industry
1	Fuhu	421.48	1.179e+08	Consumer Products & Services
2	FederalConference.com	248.31	4.960e+07	Government Services
3	The HCI Group	245.45	2.550e+07	Health
4	Bridger	233.08	1.900e+09	Energy
5	DataXu	213.37	8.700e+07	Advertising & Marketing
6	MileStone Community Builders	179.38	4.570e+07	Real Estate

```
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:

```
# Insert your code here, create more chunks as necessary
tibble::glimpse(inc)
```

```
## Observations: 5,001
## Variables: 8
## $ Rank        <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,...
## $ Name        <fct> Fuhu, FederalConference.com, The HCI Group, Bridge...
## $ 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    <fct> Consumer Products & Services, Government Services,...
## $ Employees   <int> 104, 51, 132, 50, 220, 63, 27, 75, 97, 15, 149, 16...
## $ City        <fct> El Segundo, Dumfries, Jacksonville, Addison, Bosto...
## $ State       <fct> CA, VA, FL, TX, MA, TX, TN, CA, UT, RI, VA, CA, FL...
```

```
### Top 10 companies by Growth rate.
top10_by_Growth_Rate = inc %>% arrange(desc(Growth_Rate)) %>% head(10)

kable(top10_by_Growth_Rate) %>%
  kable_styling(bootstrap_options = c("striped","hover","condensed","responsive"),full_width   = F,posi
```

Rank	Name	Growth_Rate	Revenue	Industry
1	Fuhu	421.48	1.179e+08	Consumer Products & Services
2	FederalConference.com	248.31	4.960e+07	Government Services
3	The HCI Group	245.45	2.550e+07	Health
4	Bridger	233.08	1.900e+09	Energy
5	DataXu	213.37	8.700e+07	Advertising & Marketing
6	MileStone Community Builders	179.38	4.570e+07	Real Estate
7	Value Payment Systems	174.04	2.550e+07	Financial Services
8	Emerge Digital Group	170.64	2.390e+07	Advertising & Marketing
9	Goal Zero	169.81	3.310e+07	Consumer Products & Services
10	Yagoozon	166.89	1.860e+07	Retail

```
# Top 10 Industry by Revenue
inc = inc[complete.cases(inc), ]
industry = inc %>%
  group_by(Industry) %>%
  count(Industry)%>%
  arrange(desc(n))

industry_rev = inc %>%
  group_by(Industry) %>%
  summarise(TotalRev_industry=sum(Revenue))  %>%
arrange(desc(TotalRev_industry))

industry_rev$TotalRev_industry = sapply(industry_rev$TotalRev_industry, function(x) paste(round((x / 1e9


kable(head(industry_rev , 10)) %>%
```

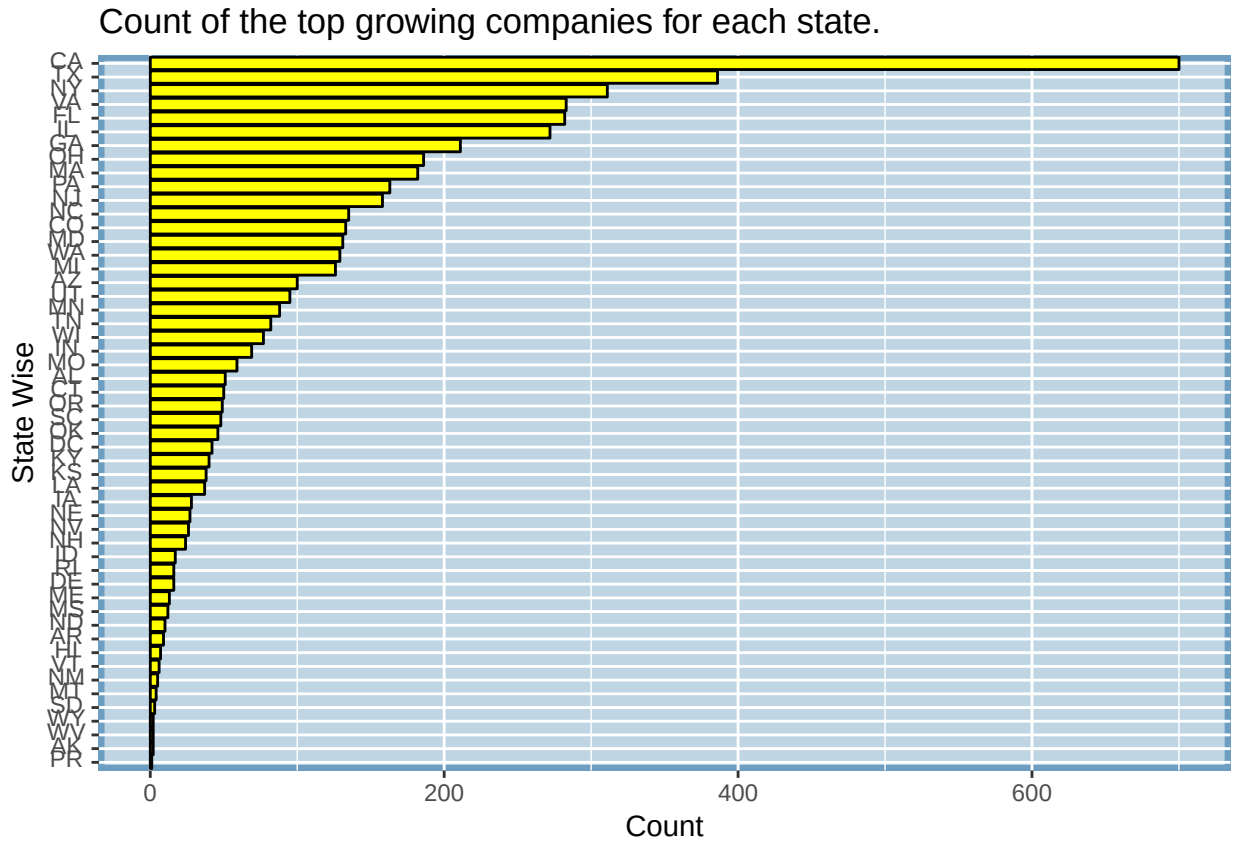
```
  kable_styling(bootstrap_options = c("striped","hover","condensed","responsive"),full_width   = F,posi
```

Industry	TotalRev_industry
Business Products & Services	26.3 billion
IT Services	20.5 billion
Health	17.9 billion
Consumer Products & Services	15 billion
Logistics & Transportation	14.8 billion
Energy	13.8 billion
Construction	13.2 billion
Financial Services	13.2 billion
Food & Beverage	12.8 billion
Manufacturing	12.6 billion

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.

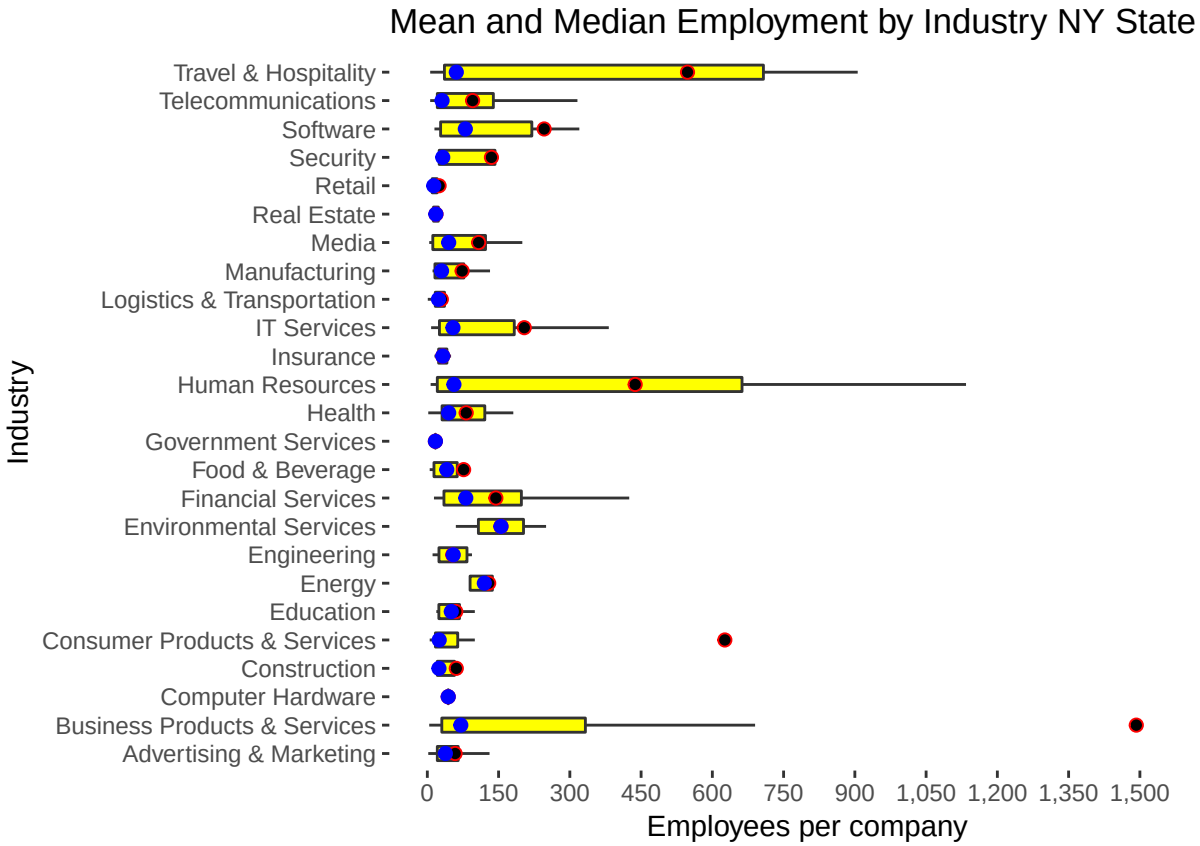
```
# Answer Question 1 here

inc %>% count(State) %>%
  ggplot(aes(x=reorder(State, n), y=n)) +
  geom_bar(stat = 'identity',fill="yellow",colour="black") +
  theme(axis.text.y = element_text(angle = 0, hjust = 0.5, vjust = 0.3),panel.background = element_rect
  coord_flip() +
  xlab("State Wise") +
  ylab("Count") +
  ggtitle("Count of the top growing companies for each 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.

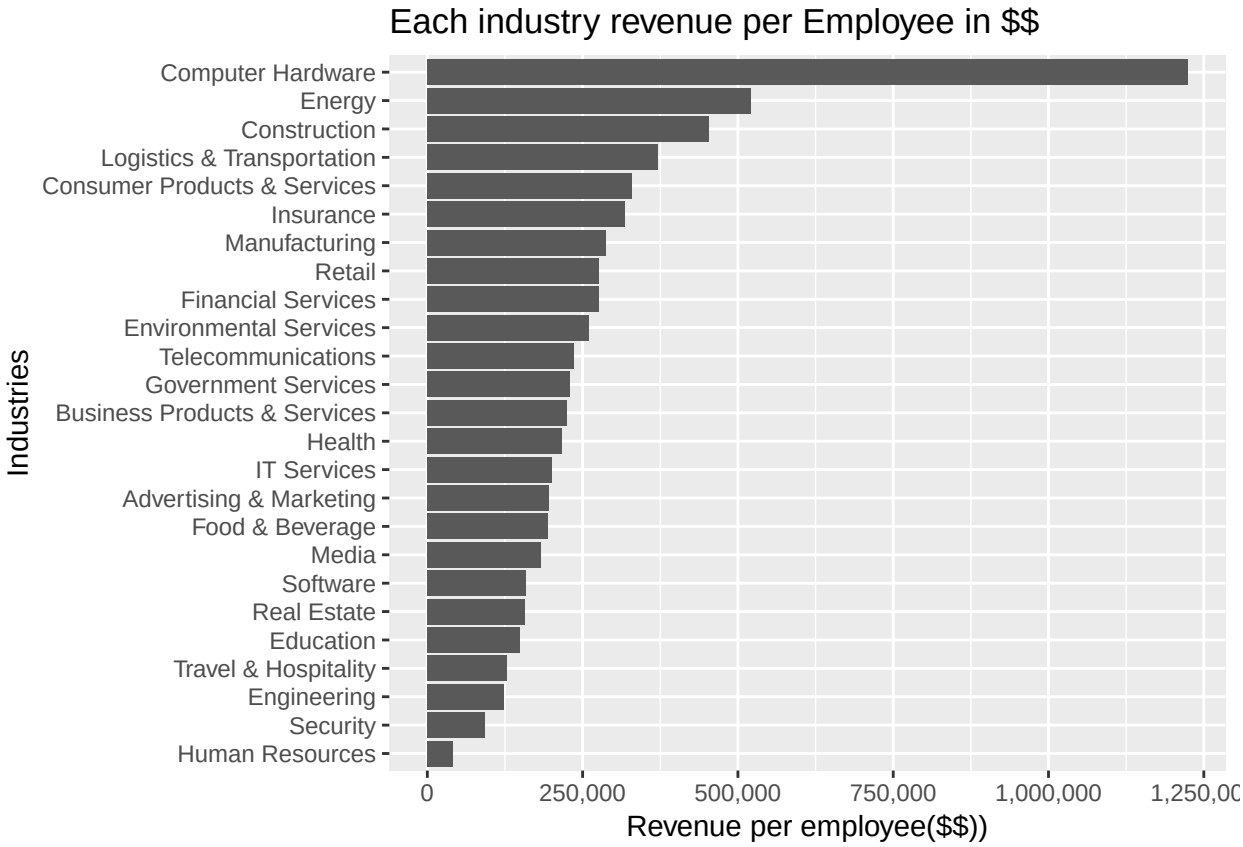
```
# Answer Question 2 here
inc_comp<- inc[complete.cases(inc), ]
q2 <- sqldf("select *from inc_comp where State = 'NY'")
ggplot(q2, aes(x=Industry, y=Employees)) +
    geom_boxplot(width=.5, fill="yellow", outlier.colour=NA) +
    stat_summary(aes(colour = "mean"), fun.y = mean, geom="point", fill="black",
                 colour="red", shape=21, size=2, show.legend=TRUE) +
    stat_summary(aes(colour = "median"), fun.y = median, geom="point", fill="blue",
                 colour="blue", shape=21, size=2, show.legend=TRUE) +
    coord_flip(ylim = c(0, 1600), expand = TRUE) +
    scale_y_continuous(labels = scales::comma,
                       breaks = seq(0, 1500, by = 150)) +
    xlab("Industry") +
    ylab("Employees per company") +
    ggtitle("Mean and Median Employment by Industry NY State") +
    theme(panel.background = element_blank(), legend.position = "top")
```



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

ind_rev_emp = inc %>%
  group_by(Industry) %>%
  summarise(TotalRev_industry_emp=sum(Revenue) / sum(Employees))  %>%
arrange(desc(TotalRev_industry_emp))

ggplot(ind_rev_emp, aes(x=reorder(Industry, TotalRev_industry_emp), y=TotalRev_industry_emp)) +
  geom_bar(stat = 'Identity') +
  coord_flip() +
  xlab("Industries") +
  ylab("Revenue per employee($$))") +
  ggtitle("Each industry revenue per Employee in $$") +
  scale_y_continuous(labels = scales::comma)
```


Each industry revenue per Employee in $$
Industries
Computer Hardware
Energy
Construction
Logistics & Transportation
Consumer Products & Services
Insurance
Manufacturing
Retail
Financial Services
Environmental Services
Telecommunications
Government Services
Business Products & Services
Health
IT Services
Advertising & Marketing
Food & Beverage
Media
Software
Real Estate
Education
Travel & Hospitality
Engineering
Security
Human Resources
0
250,000
500,000
750,000
1,000,000
1,250,00
Revenue per employee($$))