R Notebook

Alejandro D. Osborne

```
library(knitr)
library(ggplot2)
library(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
```

```
library(scales)
library(forcats)
```

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:

We want to see all industries as well as the amount of rows that are encompassed within those industries -

```
knitr::kable(group_by(inc, Industry) %>% summarize(Count=n()) %>% arrange(desc(Count)))
```

Industry	Count
IT Services	733
Business Products & Services	482
Advertising & Marketing	471
Health	355
Software	342
Financial Services	260
Manufacturing	256
Consumer Products & Services	203
Retail	203
Government Services	202
Human Resources	196
Construction	187
Logistics & Transportation	155
Food & Beverage	131
Telecommunications	129
Energy	109
Real Estate	96
Education	83
Engineering	74
Security	73
Travel & Hospitality	62
Media	54
Environmental Services	51
Insurance	50
Computer Hardware	44

```
nrow(inc)
```

```
## [1] 5001
```

We're given that this ranks the top 5000 companies, but nrows tells us more is there! Let's hope we don't have to sift through the entire database to find out why, I am going to skip to the end and see what happens -

```
tail(inc)
```

```
##      Rank               Name Growth_Rate  Revenue
## 4996 4996              cSubs        0.34 1.34e+07
## 4997 4997          Dot Foods        0.34 4.50e+09
## 4998 4998 Lethal Performance        0.34 6.80e+06
## 4999 4999   ArcaTech Systems        0.34 3.26e+07
## 5000 5000                INE        0.34 6.80e+06
## 5001 5000               ALL4        0.34 4.70e+06
##                          Industry Employees         City State
## 4996 Business Products & Services        19     Montvale    NJ
## 4997              Food & Beverage      3919 Mt. Sterling    IL
## 4998                       Retail         8   Wellington    FL
## 4999           Financial Services        63       Mebane    NC
## 5000                  IT Services        35     Bellevue    WA
## 5001       Environmental Services        34    Kimberton    PA
```

Perfect (and quite luckily)...there are two companies ranked at 5000 so that kills that mystery. Now I'm most interested in the growth rate perspective, more importantly, just how the distribution falls numerically.

```
rateofGrowth = seq(min(inc$Growth_Rate),max(inc$Growth_Rate),by=(max(inc$Growth_Rate) - min(inc$Growth_
GrowthRaterange = paste(head(rateofGrowth,-1), rateofGrowth[-1], sep=" - ")
GRFrequency = hist(inc$Growth_Rate, breaks=rateofGrowth, include.lowest=TRUE, plot=FALSE)
data.frame(range = GrowthRaterange, frequency = GRFrequency$counts)
```

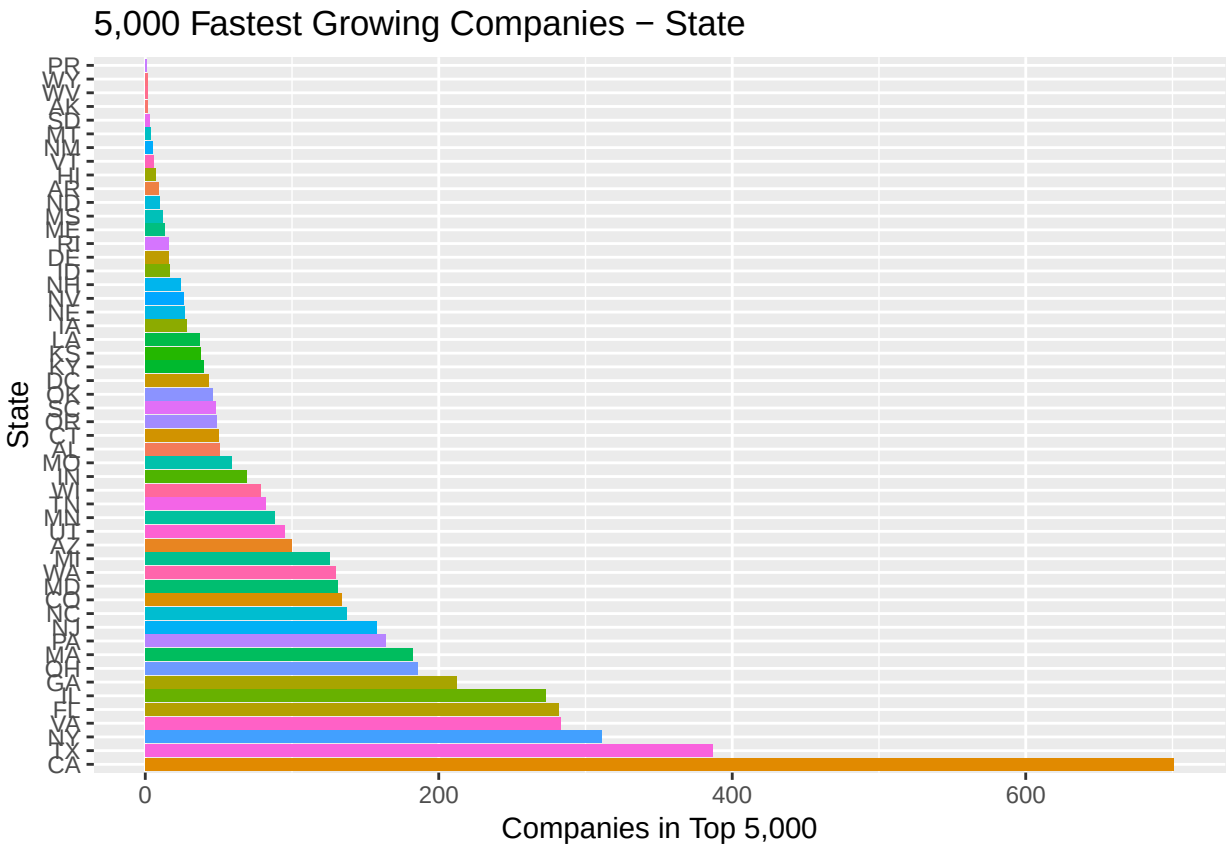
```
##                range frequency
## 1      0.34 - 42.454      4927
## 2    42.454 - 84.568        49
## 3   84.568 - 126.682        11
## 4  126.682 - 168.796         5
## 5   168.796 - 210.91         4
## 6   210.91 - 253.024         4
## 7  253.024 - 295.138         0
## 8  295.138 - 337.252         0
## 9  337.252 - 379.366         0
## 10  379.366 - 421.48         1
```

This tells us that 98.5% of the growth taken place is under 43%.

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.

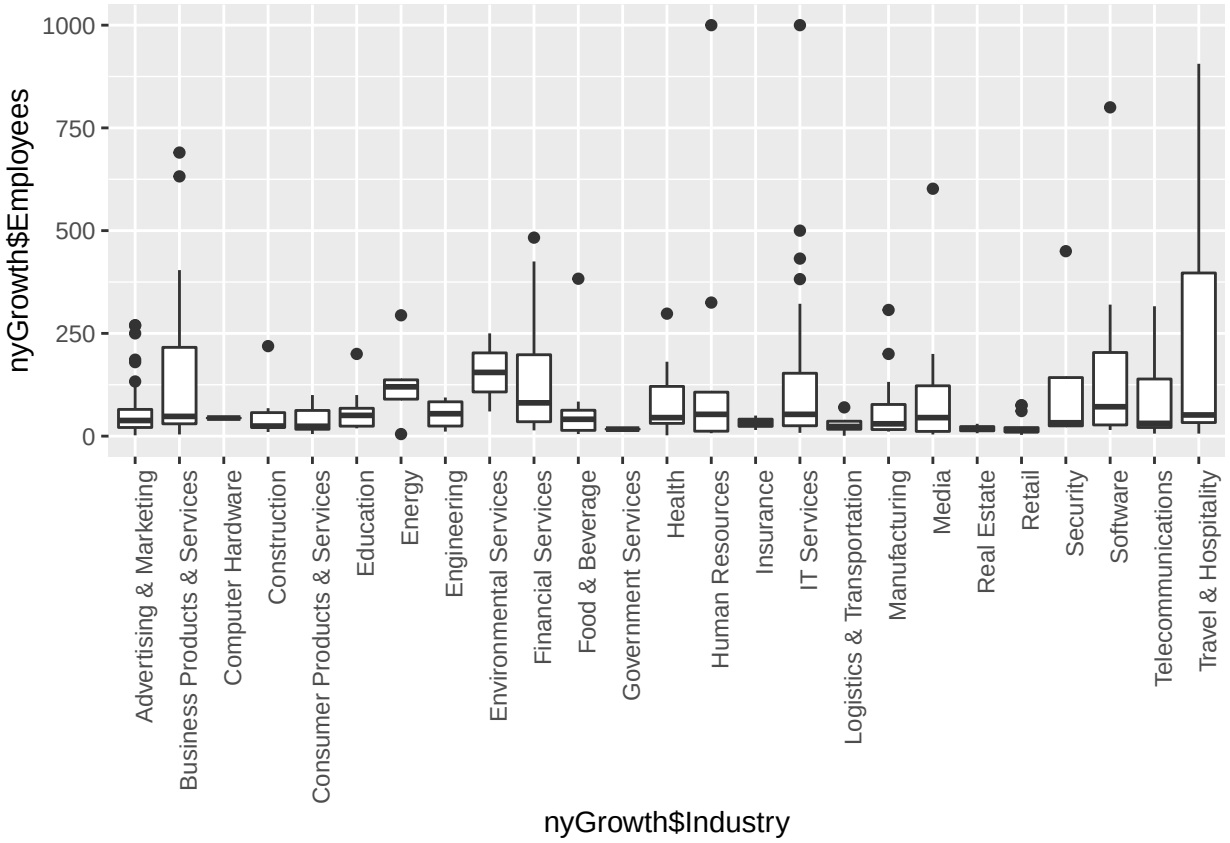
```
g <- ggplot(inc, aes(State))
g + geom_bar(aes(fct_infreq(factor(State)), fill=State), position = position_stack(reverse = TRUE), shou
```



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.

```
nyGrowth <- subset(inc,State=="NY")
ny <- ggplot(nyGrowth, aes(nyGrowth$Industry, nyGrowth$Employees))
ny + geom_boxplot(na.rm = TRUE) + ylim(0,1000) + theme(axis.text.x = element_text(angle = 90, hjust = 1
```

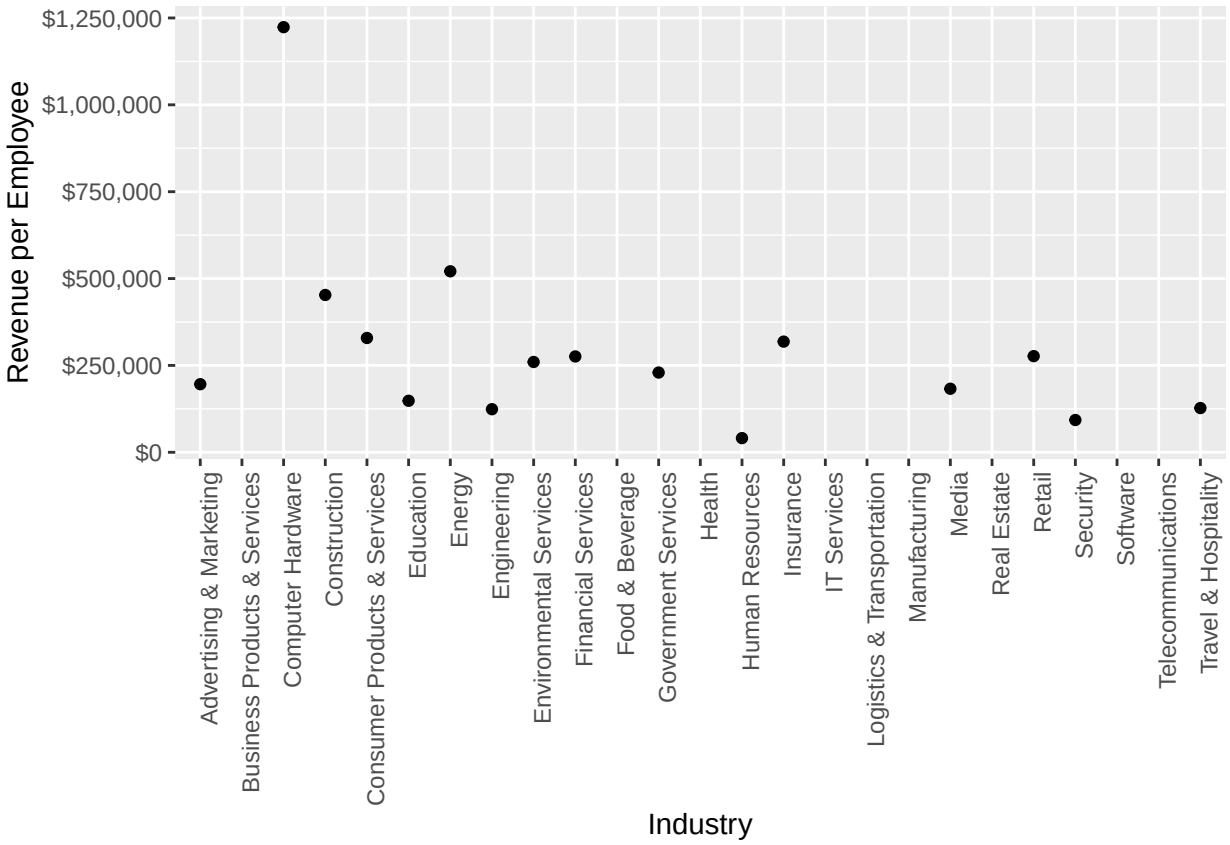


Question 3

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.

```
industry <- group_by(inc, Industry)
RevbyEmp <- summarize(industry, revpempl = sum(Revenue)/sum(Employees))
r <- ggplot(RevbyEmp, aes(RevbyEmp$Industry, RevbyEmp$revpempl))
r + geom_point() + theme(axis.text.x = element_text(angle = 90, hjust = 1)) + scale_y_continuous(labels
```

```
## Warning: Removed 9 rows containing missing values (geom_point).
```


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