R Notebook

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:

```
# Maximum and Minimum Growth Rate
(Growth_max <- inc[which.max(inc$Growth_Rate),])
```

```
##   Rank Name Growth_Rate   Revenue                      Industry Employees
## 1    1 Fuhu      421.48 117900000 Consumer Products & Services       104
##         City State
## 1 El Segundo    CA
```

```
(Growth_min <- inc[which.min(inc$Growth_Rate),])
```

```
##      Rank  Name Growth_Rate  Revenue                      Industry
## 4996 4996 cSubs        0.34 13400000 Business Products & Services
##      Employees     City State
## 4996        19 Montvale    NJ
```

```
# Maximum and Minimum Revenue
(Revenue_max <- inc[which.max(inc$Revenue),])
```

```
##      Rank Name Growth_Rate  Revenue          Industry Employees
## 4788 4788  CDW        0.41 1.01e+10 Computer Hardware      6800
##              City State
## 4788 Vernon Hills    IL
```

```
(Revenue_min <- inc[which.min(inc$Revenue),])
```

```
##     Rank                    Name Growth_Rate Revenue            Industry
## 245  246 Cardinal Point Captains       17.65   2e+06 Government Services
##     Employees     City State
## 245        30 Carlsbad    CA
```

```
# Maximum and Minimum Employees
(Employees_max <- inc[which.max(inc$Employees),])
```

```
##      Rank                         Name Growth_Rate   Revenue
## 2344 2345 Integrity staffing Solutions        1.55 278200000
##             Industry Employees       City State
## 2344 Human Resources     66803 Wilmington    DE
```

```
(Employees_min <- inc[which.min(inc$Employees),])
```

```
##     Rank         Name Growth_Rate Revenue                     Industry
## 413  414 Merch Makers       10.85 2100000 Consumer Products & Services
##     Employees City State
## 413         1 Ames    IA
```

Loading libraries

```
suppressMessages(if (!require('dplyr')) install.packages('dplyr'))
suppressMessages(if (!require('ggplot2')) install.packages('ggplot2'))
suppressMessages(if (!require('outliers')) install.packages('outliers'))
suppressMessages(if (!require('sqldf')) install.packages('sqldf'))
```

```
## Warning: package 'sqldf' was built under R version 3.6.2
```

```
## Warning: package 'gsubfn' was built under R version 3.6.2
```

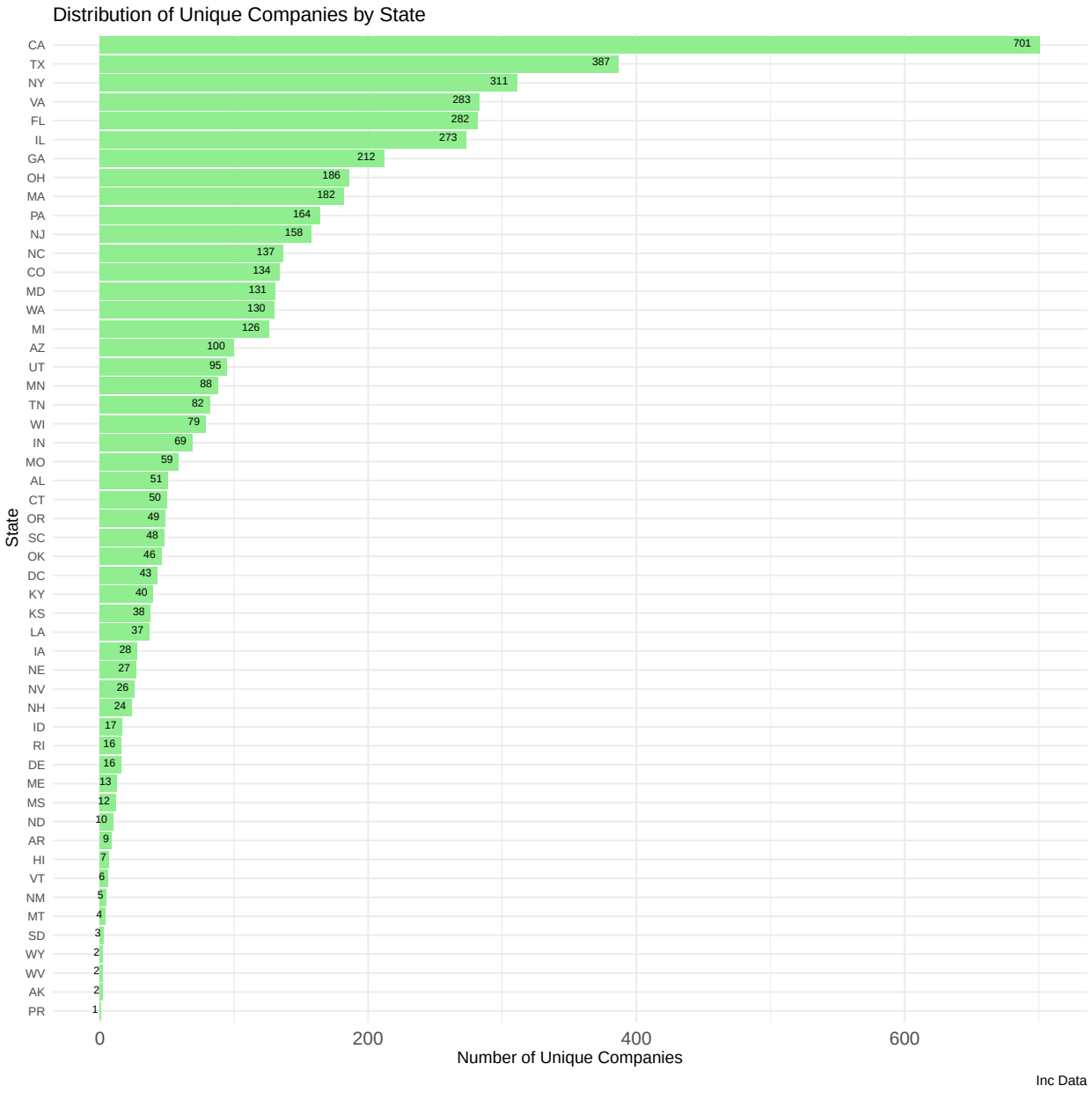
```
## Warning: package 'proto' was built under R version 3.6.2
```

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.

```
# Here we are using sqldf for writing the Sql code for data manipulation
q1 <- sqldf("select
          State, count(distinct Name) as number_companies
          from inc
          group by State")

ggplot(q1, aes(x=reorder(State,number_companies),number_companies))+
  geom_bar(stat="identity", fill="LightGreen")+
  geom_text(aes(label=round(number_companies, digits=2)), vjust=0.2, size=2.5, position=position_dodge(
  theme_minimal()+
  theme(axis.text.x=element_text(size=12, vjust=0.5))+
  theme(axis.text.y=element_text(size=8, vjust=0.5))+
  labs( x="State", y="Number of Unique Companies")+
  coord_flip()+
  labs(caption="Inc Data")+
  ggtitle("Distribution of Unique Companies by State")
```



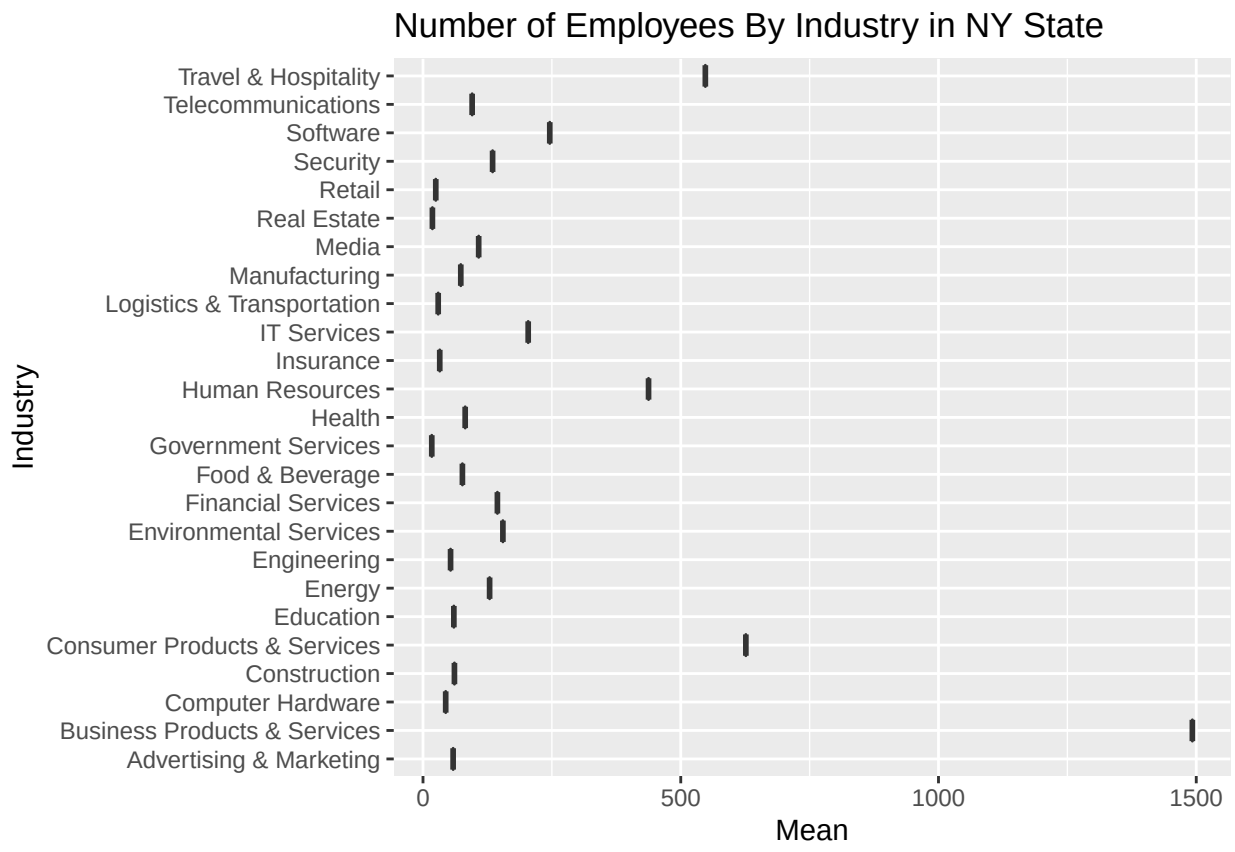

```
ggsave('Q1.png')
```

```
## Saving 6.5 x 4.5 in image
```

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.

```
q2 <- inc[complete.cases(inc), ]
q2 <- subset(inc, State == "NY")
q2 <- group_by(q2, Industry) %>% summarize(m = mean(Employees), max= max(Employees), min = min(Employees
  na.omit()
upper <- q2$max
lower <- q2$min
ggplot(q2, aes(x = Industry, y =m, ymax=max,  ymin = min, lower = lower, upper= upper)) + geom_boxplot(
  labs(title="Number of Employees By Industry in NY State", y = "Mean")
```




```
ggsave('Q2.png')
```

```
## Saving 6.5 x 4.5 in image
```

Question 3

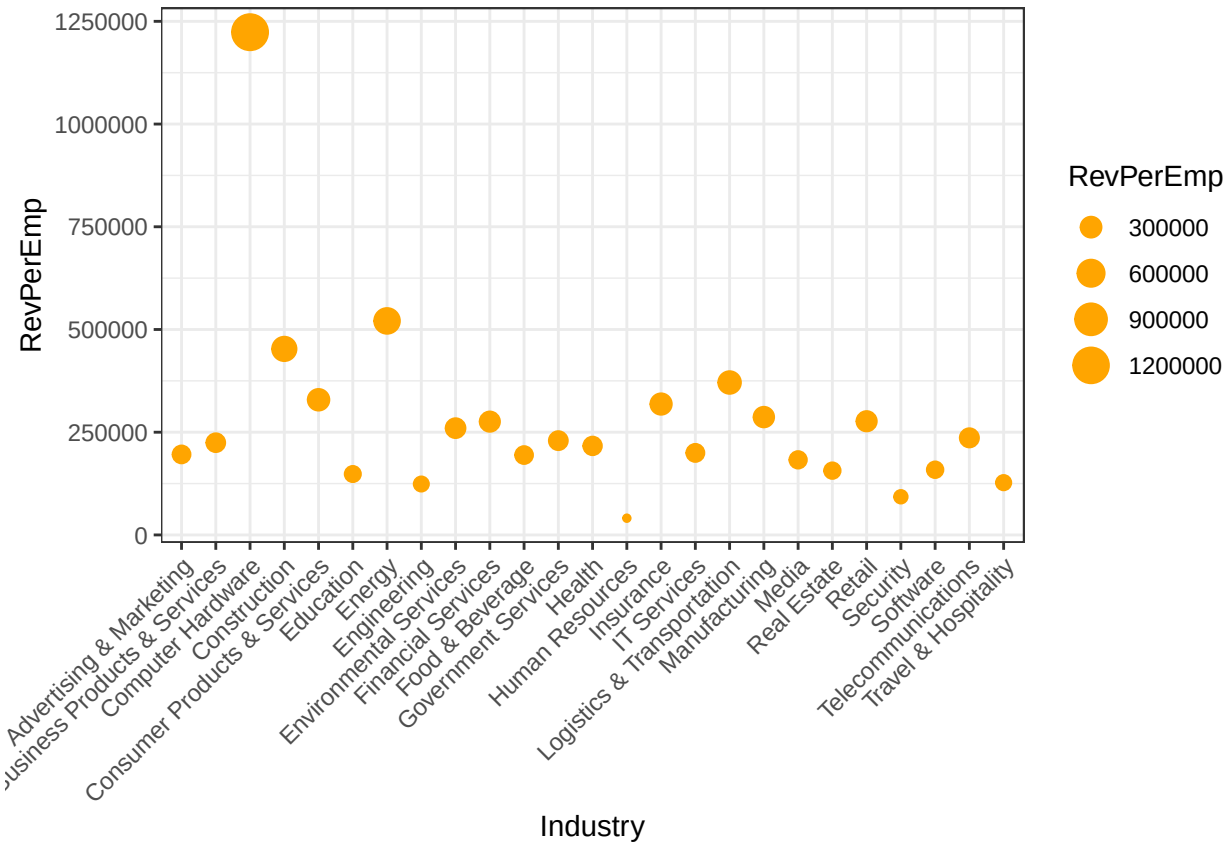
Now imagine you work for an investor and want to see which industries generate the most revenue per employee. Create a chart that makes this information clear. Once again, the distribution per industry should be shown.

```
q3 <- inc[complete.cases(inc), ]
q3 <- q3[, c("Industry", "Revenue", "Employees")] %>% group_by(Industry) %>% summarise_each(funs(sum))
```

```
## Warning: funs() is soft deprecated as of dplyr 0.8.0
```

```
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once per session.
```

```
q3$RevPerEmp <- q3$Revenue / q3$Employees
ggplot(q3, aes(x = Industry, y = RevPerEmp)) +
  geom_point(aes(size = RevPerEmp), color = "Orange") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust=1))
```




```
ggsave('Q3.png')
```

```
## Saving 6.5 x 4.5 in image
```