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:

```
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
```

```
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:

A little bit of extra exploration is make easier to read names of the columns and understand the number of rows in the dataset

```
# Insert your code here, create more chunks as necessary
names(inc)
```

```
## [1] "Rank"        "Name"        "Growth_Rate" "Revenue"     "Industry"
## [6] "Employees"   "City"        "State"
```

```
nrow(inc)
```

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

Another important is understand where missing values are located since they might affect or skew our visualizations

```
colSums(is.na(inc))
```

```
##        Rank        Name Growth_Rate     Revenue    Industry   Employees
##           0           0           0           0           0          12
##        City       State
##           0           0
```

```
sum(is.na(inc$Employees))
```

```
## [1] 12
```

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.

```
library(Hmisc)
```

```
## Warning: package 'Hmisc' was built under R version 3.5.3
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Warning: package 'survival' was built under R version 3.5.2
```

```
## Loading required package: Formula
```

```
##
## Attaching package: 'Hmisc'
```

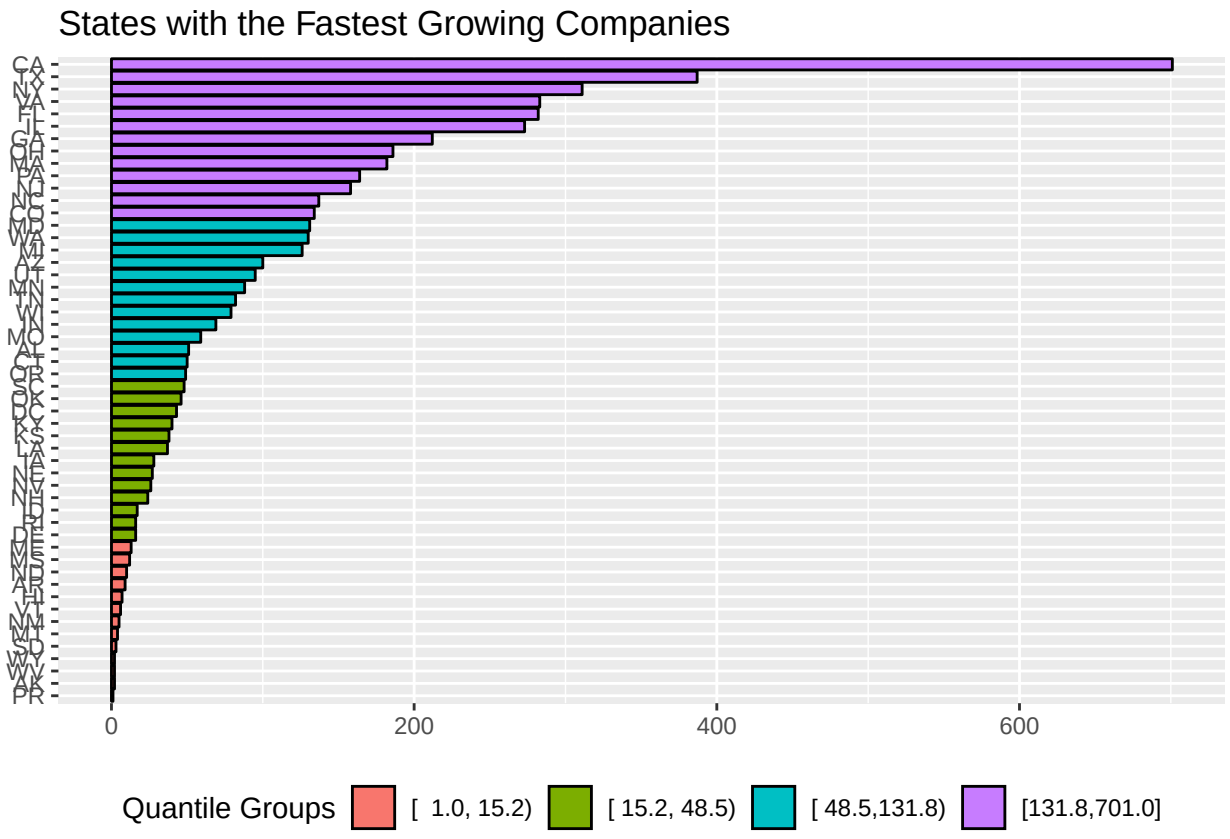
```
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
# Answer Question 1 here
qrtile <- inc %>% count(State) %>% arrange(desc(n))
```

```
## Warning: The `printer` argument is deprecated as of rlang 0.3.0.
## This warning is displayed once per session.
```

```
qrtile <- qrtile %>%  mutate(quant = cut2(qrtile$n,quantile(qrtile$n, include.lowest=TRUE)))
# https://stackoverflow.com/questions/11728419/using-cut-and-quartile-to-generate-breaks-in-r-function
ggplot(qrtile, aes(x = reorder(State, n), y = n)) +
  geom_bar(aes(fill = quant), color="black", stat = "identity") +
  coord_flip() +
  ggtitle("States with the Fastest Growing Companies") +
  labs(y= NULL, x = NULL) +
  scale_fill_discrete(name = "Quantile Groups") +
  theme(legend.position="bottom")
```



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.

```
# Answer Question 2 here
#Wjo is the third state?
qrtile$State[3]
```

```
## [1] NY
## 52 Levels: AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA ... WY
```

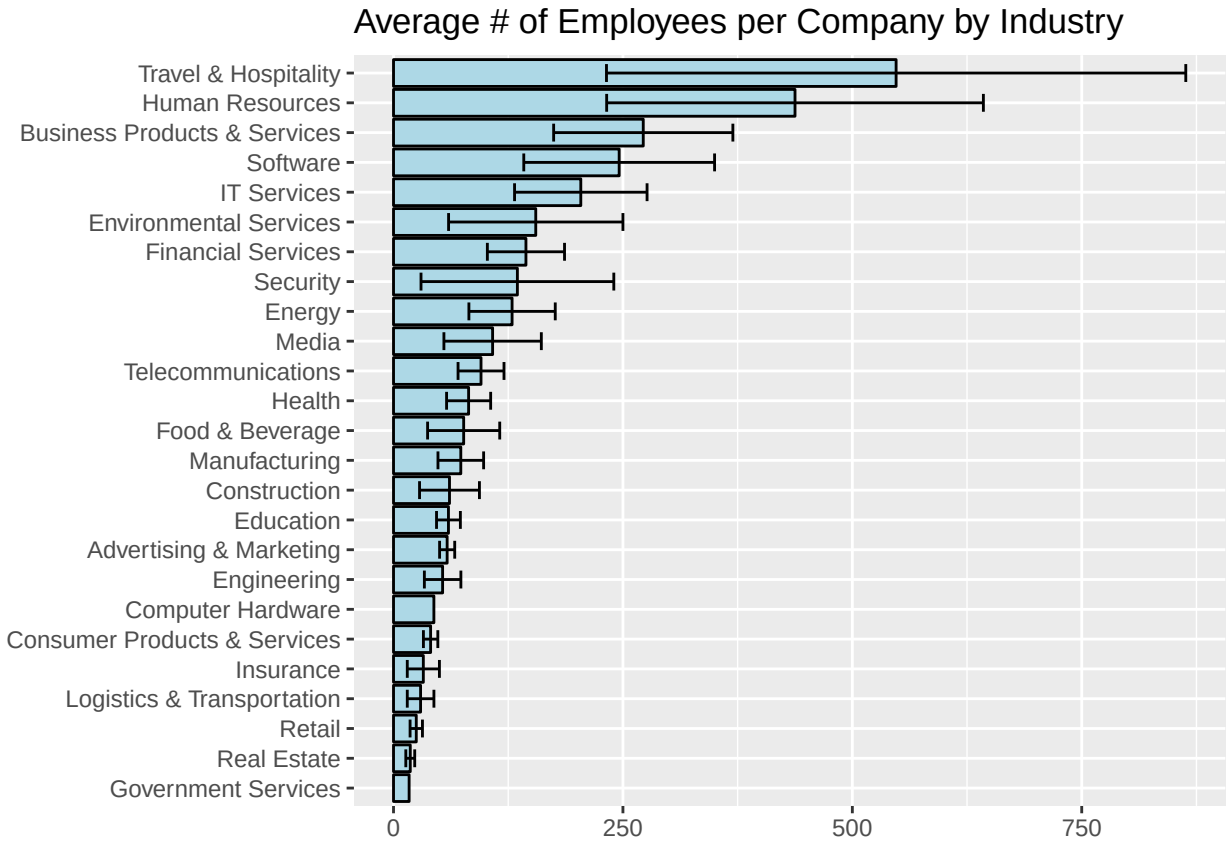
```
ny <- inc %>%
  mutate(cases = complete.cases(inc)) %>%
  filter(cases=="TRUE") %>%
  filter(State == "NY") %>%
  #looks to see if values are more than 2 standard deviations from the mean.I take care of outliers
  filter(!(abs(Employees - mean(Employees)) > 2*sd(Employees))) %>%
  group_by(Industry)%>%
  #Find the mean and standard error
  summarise(mean = mean(Employees),
            n = length(Industry),
            se = sd(Employees)/sqrt(n))



# Take a look at the outliers we eliminated
test <- inc %>%
  mutate(cases = complete.cases(inc)) %>%
  filter(cases=="TRUE") %>%
  filter(State == "NY") %>%
  arrange(desc(Employees))


ggplot(ny, aes(x = reorder(Industry, mean), y = mean)) +
  geom_bar(fill = "Lightblue", color="black", stat = "identity") +
  geom_errorbar(aes(ymin=mean-se, ymax=mean+se), width=0.6) +
  ggtitle("Average # of Employees per Company by Industry") +
  labs(y= NULL, x = NULL) +
  guides(fill=FALSE) +
  coord_flip()
```

```
## Warning: Removed 2 rows containing missing values (geom_errorbar).
```




```
#########################################################
library(scales)
qtile2<-inc %>%
  filter (State == "NY")

head(qtile2)
```

```
##   Rank                       Name Growth_Rate  Revenue
## 1   26                BeenVerified       84.43 13700000
## 2   30                    Sailthru       73.22  8100000
## 3   37                YellowHammer       67.40 18000000
## 4   38                   Conductor       67.02  7100000
## 5   48 Cinium Financial Services       53.65  5900000
## 6   70                    33Across       44.99 27900000
##                         Industry Employees      City State
## 1 Consumer Products & Services        17  New York    NY
## 2      Advertising & Marketing        79  New York    NY
## 3      Advertising & Marketing        27  New York    NY
## 4      Advertising & Marketing        89  New York    NY
## 5           Financial Services        32 Rock Hill    NY
## 6      Advertising & Marketing        75  New York    NY
```

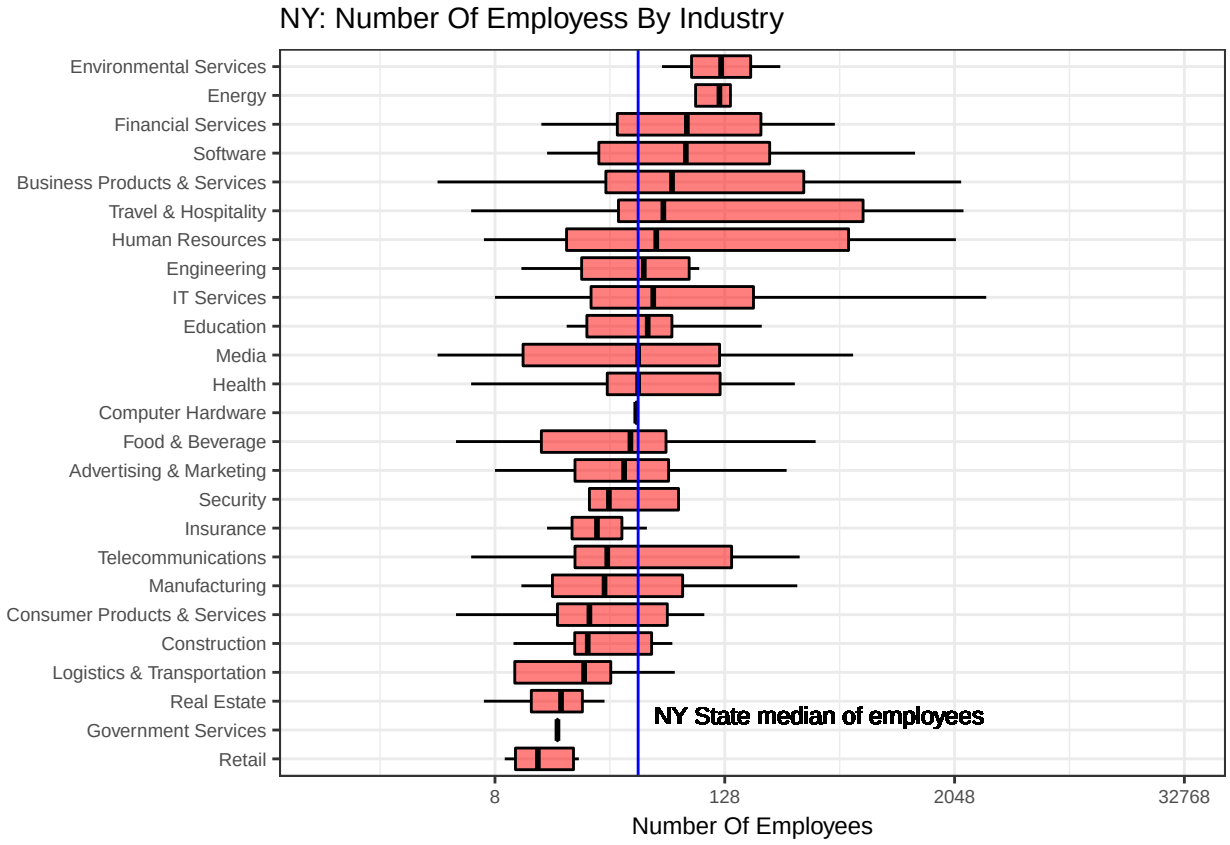
```
qtile2 <- qtile2[complete.cases(qtile2$Industry), ]
qtile2 <- qtile2[complete.cases(qtile2$Employees), ]
ny_median<-median(qtile2$Employees)

lower <- min(qtile2$Employees)
```

```
upper <- max(qtile2$Employees)

qtile2_test<-ggplot(qtile2, aes(reorder(Industry, Employees, FUN=median), Employees)) +
    geom_boxplot(outlier.shape = NA,  color = "black", fill = "red", alpha = 0.5) +
    scale_y_continuous(trans = log2_trans(), limits = c(lower, upper)) +
    geom_hline(yintercept = ny_median, color="blue") +
    geom_text(aes(2.5,400,label = "NY State median of employees"), size = 3)+
    coord_flip() +
    ggtitle ("NY: Number Of Employess By Industry") + ylab("Number Of Employees")+
    theme_bw()+
    theme(axis.title.y=element_blank())+
    theme(text = element_text(size = 9, color = "black"))

qtile2_test
```



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

rev <- inc %>%
  mutate(cases = complete.cases(inc)) %>%
```

```
  filter(cases=="TRUE") %>%
  mutate(rev_emp = Revenue/Employees) %>%
  #looks to see if values are more than 2 standard deviations from the mean.
  filter(!(abs(rev_emp - mean(rev_emp)) > 2*sd(rev_emp))) %>%
  group_by(Industry)%>%
  #Find the mean and standard error
  summarise(Revenue_Employee = sum(Revenue)/sum(Employees),
            n = length(Industry),
            se = sd(Revenue/Employees)/sqrt(n))


ggplot(rev, aes(x = reorder(Industry, Revenue_Employee), y = Revenue_Employee)) +
  geom_bar(fill = "lightblue", color="black", stat = "identity") +
  geom_errorbar(aes(ymin=Revenue_Employee-se, ymax=Revenue_Employee+se), width=0.6) +
  ggtitle("Average Revenue per Employee by Industry") +
  labs(y= NULL, x = NULL) +
  guides(fill=FALSE) +
  scale_y_continuous(labels = scales::comma) +
  coord_flip()
```