DATA 608 - Assignment 1

Rachel Greenlee

```
library('dplyr')
```

```
## Warning: package 'dplyr' was built under R version 4.0.4
```

```
##
## 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('ggplot2')

mpeach <- "#FBAA82"
mteal <- "#73A2AC"
mdarkteal <- "#0B5D69"
mgray <- "#4C4C4C"
```

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         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
```

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:

Taking a look at the top 10 and bottom 10 ranked companies.

```
head(inc, 10)
```

```
##    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
## 7     7         Value Payment Systems      174.04 2.550e+07
## 8     8          Emerge Digital Group      170.64 2.390e+07
## 9     9                     Goal Zero      169.81 3.310e+07
## 10   10                      Yagoozon      166.89 1.860e+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
## 7            Financial Services        27     Nashville    TN
## 8       Advertising & Marketing        75 San Francisco    CA
## 9  Consumer Products & Services        97     Bluffdale    UT
## 10                       Retail        15       Warwick    RI
```

```
tail(inc, 10)
```

```
##      Rank                     Name Growth_Rate  Revenue
## 4992 4992 Salem Metal Fabricators        0.35 7.40e+06
## 4993 4993          The PI Company        0.35 2.00e+06
## 4994 4994            RFB Holdings        0.35 7.20e+06
## 4995 4995      Sterling Computers        0.35 1.66e+08
## 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
## 4992                Manufacturing        50         Middleton    MA
## 4993 Business Products & Services         6 North Little Rock    AR
## 4994              Human Resources        27      Downer Grove    IL
## 4995          Government Services        98           Norfolk    NE
## 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
```

A quick look at the locations of these companies and the frequency in each state. Only 134 in my current state of Colorado, and a mere 79 in my home state of Wisconsin.

```
inc %>% count(State, sort = TRUE)
```

```
##    State   n
## 1     CA 701
## 2     TX 387
## 3     NY 311
## 4     VA 283
## 5     FL 282
## 6     IL 273
## 7     GA 212
## 8     OH 186
## 9     MA 182
## 10    PA 164
## 11    NJ 158
## 12    NC 137
## 13    CO 134
## 14    MD 131
## 15    WA 130
## 16    MI 126
## 17    AZ 100
## 18    UT  95
## 19    MN  88
## 20    TN  82
## 21    WI  79
```

```
## 22    IN  69
## 23    MO  59
## 24    AL  51
## 25    CT  50
## 26    OR  49
## 27    SC  48
## 28    OK  46
## 29    DC  43
## 30    KY  40
## 31    KS  38
## 32    LA  37
## 33    IA  28
## 34    NE  27
## 35    NV  26
## 36    NH  24
## 37    ID  17
## 38    DE  16
## 39    RI  16
## 40    ME  13
## 41    MS  12
## 42    ND  10
## 43    AR   9
## 44    HI   7
## 45    VT   6
## 46    NM   5
## 47    MT   4
## 48    SD   3
## 49    AK   2
## 50    WV   2
## 51    WY   2
## 52    PR   1
```

Using similar code we can look at the most common industries on the ranking.

```
inc %>% count(Industry, sort = TRUE)
```

```
##                        Industry   n
## 1                   IT Services 733
## 2  Business Products & Services 482
## 3       Advertising & Marketing 471
## 4                        Health 355
## 5                      Software 342
## 6            Financial Services 260
## 7                 Manufacturing 256
## 8  Consumer Products & Services 203
## 9                        Retail 203
## 10          Government Services 202
## 11              Human Resources 196
## 12                 Construction 187
## 13   Logistics & Transportation 155
## 14              Food & Beverage 131
## 15           Telecommunications 129
## 16                       Energy 109
```

```
## 17                  Real Estate  96
## 18                   Education  83
## 19                 Engineering  74
## 20                    Security  73
## 21        Travel & Hospitality  62
## 22                       Media  54
## 23      Environmental Services  51
## 24                   Insurance  50
## 25           Computer Hardware  44
```

Question 1

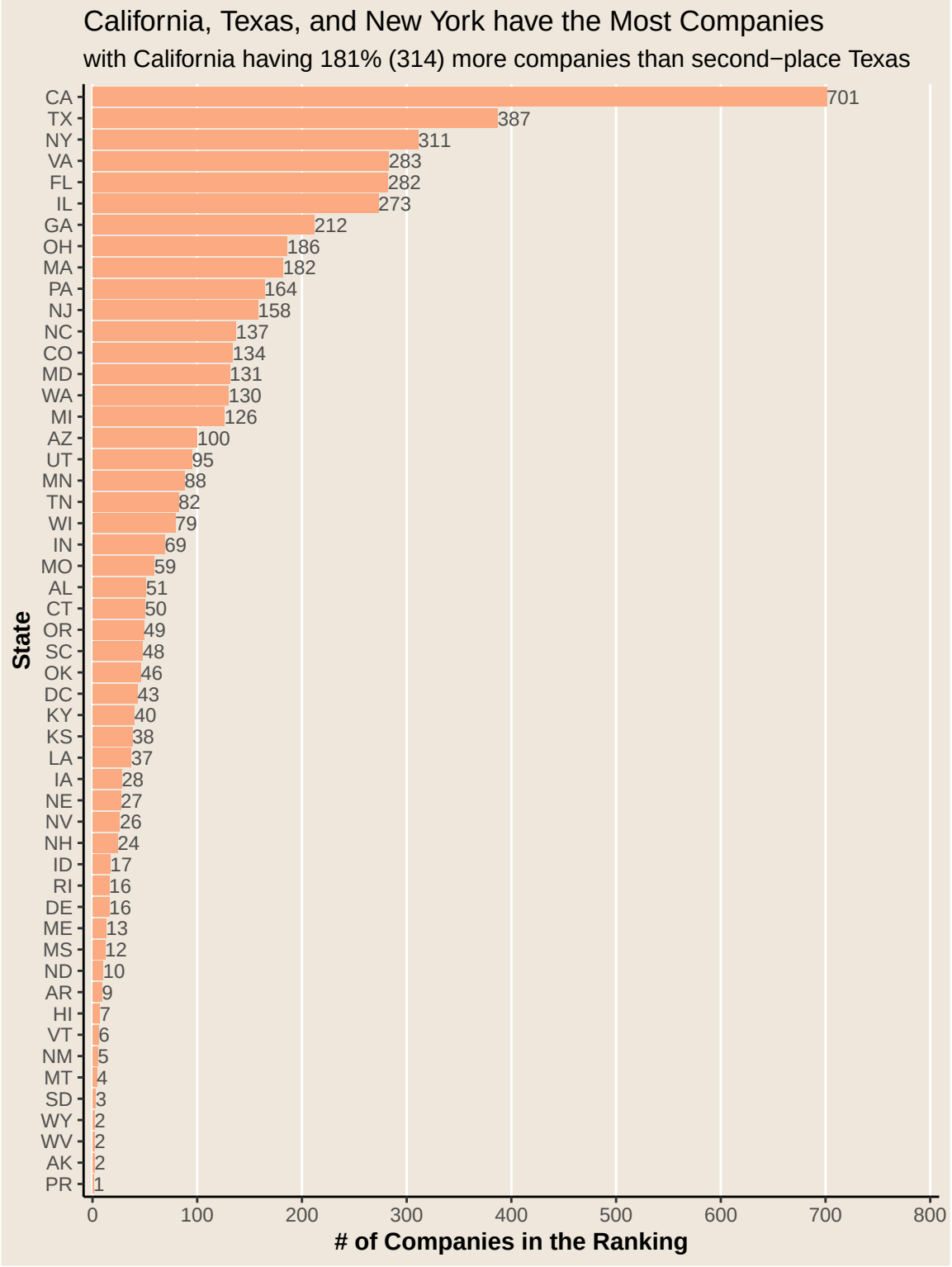
Create a graph that shows the distribution of companies in the dataset by State (ie how many are in each state). There are a lot of States, so consider which axis you should use. This visualization is ultimately going to be consumed on a 'portrait' oriented screen (ie taller than wide), which should further guide your layout choices.

```
# create 2-variable df of States & counts
state_freq <- inc %>% count(State, sort = TRUE)

# sort df by count so graph displays ordered and not alphabetically
state_freq$State <- factor(state_freq$State,
                  levels = state_freq$State[order(state_freq$n, decreasing = FALSE)])

# set plot theme for assignment
my_plot_theme <- list(
  theme_classic() +
  theme(plot.background = element_rect(fill = "#EFE7DB"),
        panel.background = element_rect(fill = "#EFE7DB"),
        panel.grid.major.x = element_line(color = "white"),
        axis.title.y = element_text(face = "bold"),
        axis.title.x = element_text(face = "bold")))

# plot generation
ggplot(data = state_freq, aes(x = n, y = State)) +
  geom_col(fill=mpeach) +
  labs(title = "California, Texas, and New York have the Most Companies",
       subtitle = "with California having 181% (314) more companies than second-place Texas",
       x="# of Companies in the Ranking",
       y="State") +
  scale_x_continuous(limits = c(0, 800),
                     expand = c(.01, 0.5),
                     breaks = seq(0, 800, 100)) +
  geom_text(
    aes(x = n, label = n),
    size = 3,
    color = mgray, hjust = 0) +
  my_plot_theme
```



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

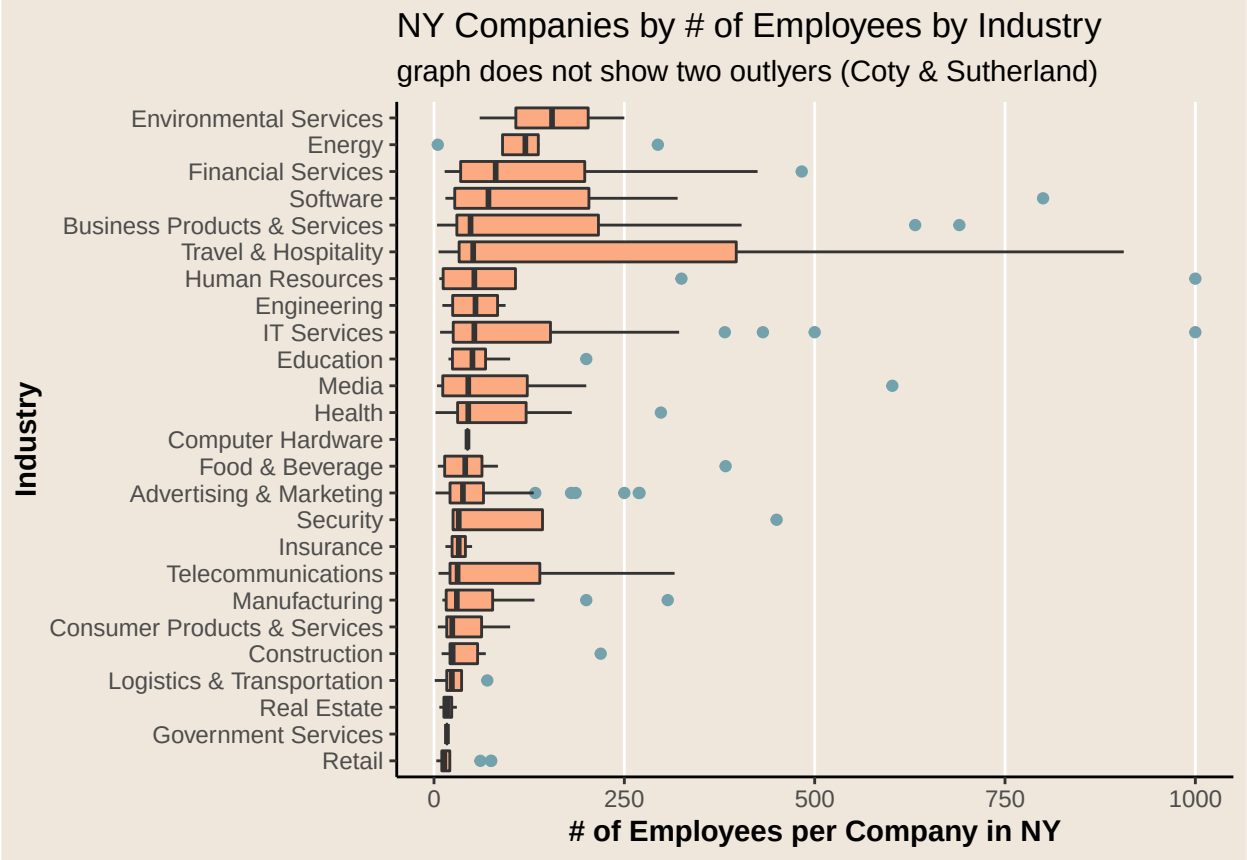
shows the average and/or median employment by industry for companies in this state (only use cases with full data, use R's `complete.cases()` function.) In addition to this, your graph should show how variable the ranges are, and you should deal with outliers.

Showing a boxplot for each industry is the best way to see the median employment level as well as the range. This dataset had one moderate outlyer and one extreme outlyer that I acknowledged were cutoff in the subtitle to the graph. I'd argue a healthy supplement to this graph would be a table showing the n's for each industry, which I produced below. For example, while Environmental Services appears to have the highest median number of employees, it's from a sample of only 2 companies.

```
# filter for just NY data
# filter for complete cases (they all were w/o NAs)
# plot to show median across industries in NY

inc %>%
  filter(State == "NY", complete.cases(.)) %>%
  ggplot(aes(x = reorder(Industry, Employees, median), y = Employees)) +
    geom_boxplot(fill = mpeach, outlier.color = mteal) +
    coord_flip() +
    labs(title = "NY Companies by # of Employees by Industry",
       subtitle = "graph does not show two outlyers (Coty & Sutherland)",
       x="Industry",
       y="# of Employees per Company in NY") +
    ylim(NA, 1000) +
    my_plot_theme
```

```
## Warning: Removed 9 rows containing non-finite values (stat_boxplot).
```




```
# show counts sorted high to low of companies within each industry in NY
inc %>%
  filter(State == "NY") %>%
  count(Industry, sort = TRUE)
```

```
##                        Industry  n
## 1       Advertising & Marketing 57
## 2                    IT Services 43
## 3  Business Products & Services 26
## 4  Consumer Products & Services 17
## 5            Telecommunications 17
## 6                     Education 14
## 7                        Retail 14
## 8             Financial Services 13
## 9                        Health 13
## 10                Manufacturing 13
## 11                     Software 13
## 12               Human Resources 11
## 13                         Media 11
## 14               Food & Beverage  9
## 15         Travel & Hospitality  7
## 16                 Construction  6
## 17                       Energy  5
## 18                  Engineering  4
## 19   Logistics & Transportation  4
## 20                  Real Estate  4
```

```
## 21                    Security  4
## 22       Environmental Services  2
## 23                   Insurance  2
## 24           Computer Hardware  1
## 25          Government Services  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.

```
inc %>%
  filter(complete.cases(.)) %>%
  group_by(Industry) %>%
    summarize(rev_tot = sum(Revenue), emp_tot = sum(Employees)) %>%
    mutate(rev_per_emp = rev_tot/emp_tot) %>%
  ggplot(aes(x = reorder(Industry, rev_per_emp), y = rev_per_emp)) +
    geom_bar(stat = "identity", fill = mpeach) +
    coord_flip() +
    labs(title = "Revenue per Employee by Industry",
       subtitle = "total revenue of industry / total count employees of industry",
       x = "Industry",
       y = "Revenue per Employee") +
    my_plot_theme
```