DATA 608 HW 1

Albert Gilharry

September 6, 2018

```
library("dplyr")
library("ggplot2")
library("scales")
```

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:

```
# Insert your code here, create more chunks as necessary
print(paste0("There are ", nrow(inc[!complete.cases(inc),]), " incomplete cases in this data set."))
```

```
## [1] "There are 12 incomplete cases in this data set."
```

```
print(paste0("There are ", length(unique(inc$City)), " cities represented in this data set."))
```

```
## [1] "There are 1519 cities represented in this data set."
```

```
print(paste0("There standard deviation of Growth_Rate is ", round(sd((inc$Growth_Rate)),4), "."))
```

```
## [1] "There standard deviation of Growth_Rate is 14.1237."
```

```
print(paste0("There standard deviation of Revenue is ", round(sd((inc$Revenue)),4), "."))
```

```
## [1] "There standard deviation of Revenue is 240542281.1359."
```

```
print(paste0("There standard deviation of Employees is ", round(sd((na.omit(inc$Employees))) ,4), "."))
```

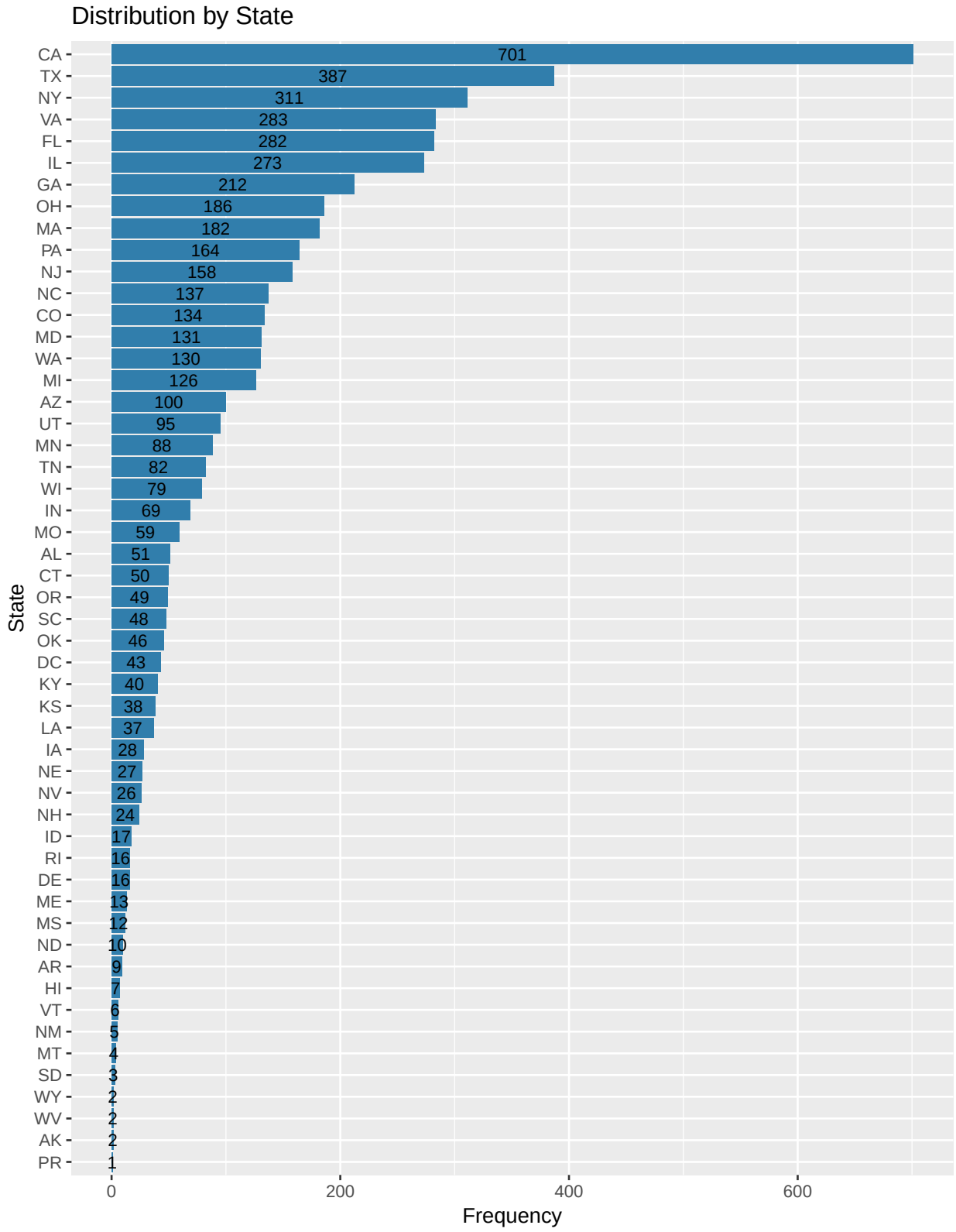
```
## [1] "There standard deviation of Employees is 1353.1279."
```

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.

- I will create a histogram sorted by frequencies of companies in each state. My assumption is that this graphic is primarily intended for business purposes and having it sorted this way will better inform investment and other business decisions, by highlighting possible saturated markets and identifying possible areas for opportunities.

```
# Answer Question 1 here
group_by(inc, State) %>%
  summarize( count = n() ) %>%
  arrange(desc(State)) %>%
  ggplot(aes( x = reorder(State,count), y = count, label = count, fill= State )) +
  geom_bar( stat='identity', show.legend = F, fill="#317EAC" ) +
  geom_text(size = 3, position = position_stack(vjust = 0.5)) +
  labs(title = "Distribution by State", x = "State", y = "Frequency" ) +
  coord_flip()
```


Distribution by State
State
CA 701
TX 387
NY 311
VA 283
FL 282
IL 273
GA 212
OH 186
MA 182
PA 164
NJ 158
NC 137
CO 134
MD 131
WA 130
MI 126
AZ 100
UT 95
MN 88
TN 82
WI 79
IN 69
MO 59
AL 51
CT 50
OR 49
SC 48
OK 46
DC 43
KY 40
KS 38
LA 37
IA 28
NE 27
NV 26
NH 24
ID 17
RI 16
DE 16
ME 13
MS 12
ND 10
AR 9
HI 7
VT 6
NM 5
MT 4
SD 3
WY 2
WV 2
AK 2
PR 1
0
200
400
600
Frequency

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.

- The results from the previous question showed that the state with the third most companies is NY. I will plot the median because the variable `Employees` has some very large outliers, but is this also the case for NY?

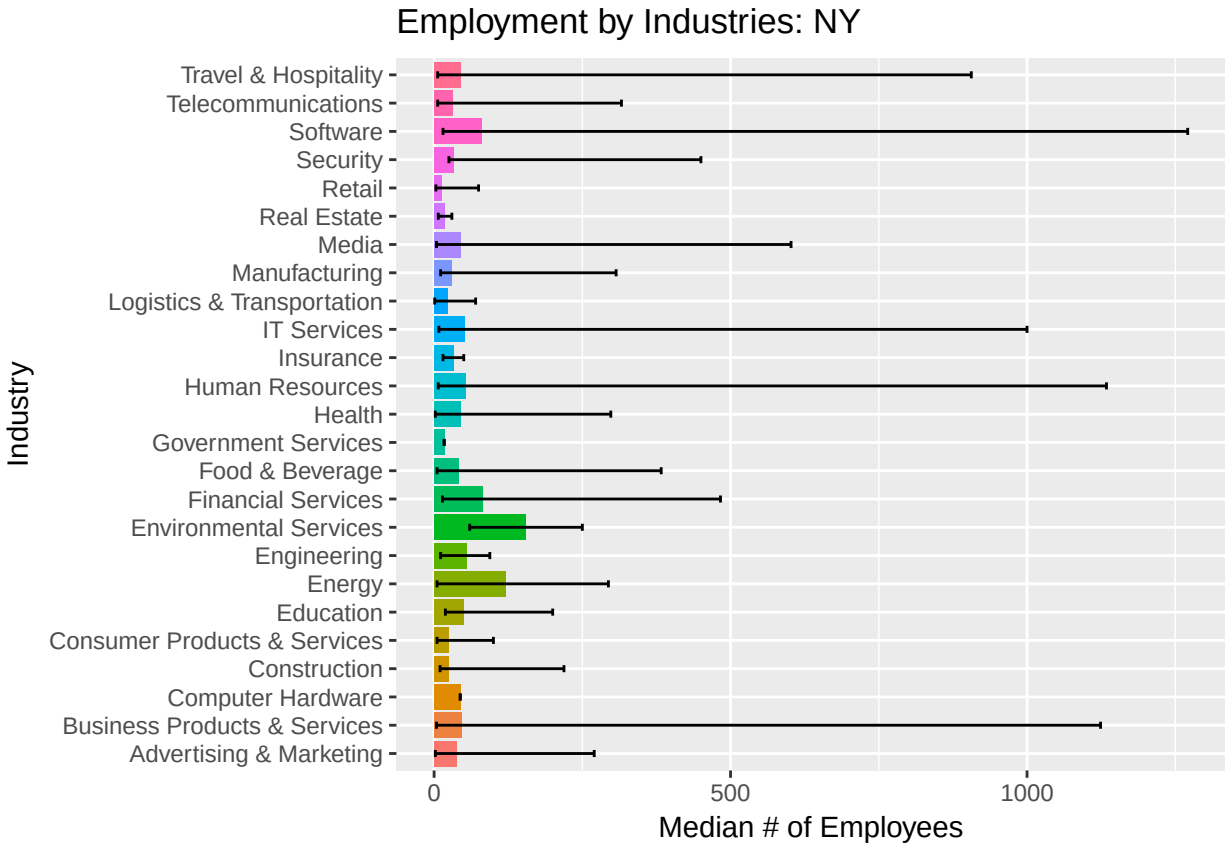
```
complete_cases_ny <- inc[complete.cases(inc),] %>%
  filter(State == 'NY')  %>%
  arrange(desc(Employees))

  head(complete_cases_ny)
```

```
##   Rank                        Name Growth_Rate   Revenue
## 1 4577 Sutherland Global Services        0.48 5.976e+08
## 2 4936                       Coty        0.36 4.600e+09
## 3 4716             Westcon Group         0.44 3.800e+09
## 4 3899  Denihan Hospitality Group        0.71 2.808e+08
## 5 4363              TransPerfect         0.55 3.413e+08
## 6 1499       Sterling Infosystems        2.66 2.149e+08
##                        Industry Employees      City State
## 1 Business Products & Services     32000 Pittsford    NY
## 2 Consumer Products & Services     10000  New York    NY
## 3                  IT Services      3000 Tarrytown    NY
## 4         Travel & Hospitality      2280  New York    NY
## 5 Business Products & Services      2218  New York    NY
## 6              Human Resources      2081  New York    NY
```

- It turns out that NY has only 2 companies with more 10000 or more employees. I will replace cases with employees greater than 1 standard deviation with the median. These outliers are likely candidates to be removed but there is not sufficient context to justify removing them, so they will remain. This will result in reduced ranges, albeit minimally. Below is a bar chart with error bars attached, depicting the ranges.

```
  # Answer Question 2 here
  complete_cases_ny$Employees[complete_cases_ny$Employees > 1*sd(complete_cases_ny$Employees)] <- media
  complete_cases_ny %>%
  group_by(Industry) %>%
  summarize(med = median(Employees), mx = max(Employees), mn = min(Employees), avg = mean(Employees)) %
  ggplot(aes(x = Industry, y = med,  fill=Industry, label = med)) +
  geom_bar( stat='identity', show.legend = F ) +
  geom_errorbar(aes(ymin = mn, ymax = mx), position = "dodge", width = 0.25) +
  labs(title = "Employment by Industries: NY", x = "Industry", y = "Median # of Employees" ) +
  coord_flip()
```



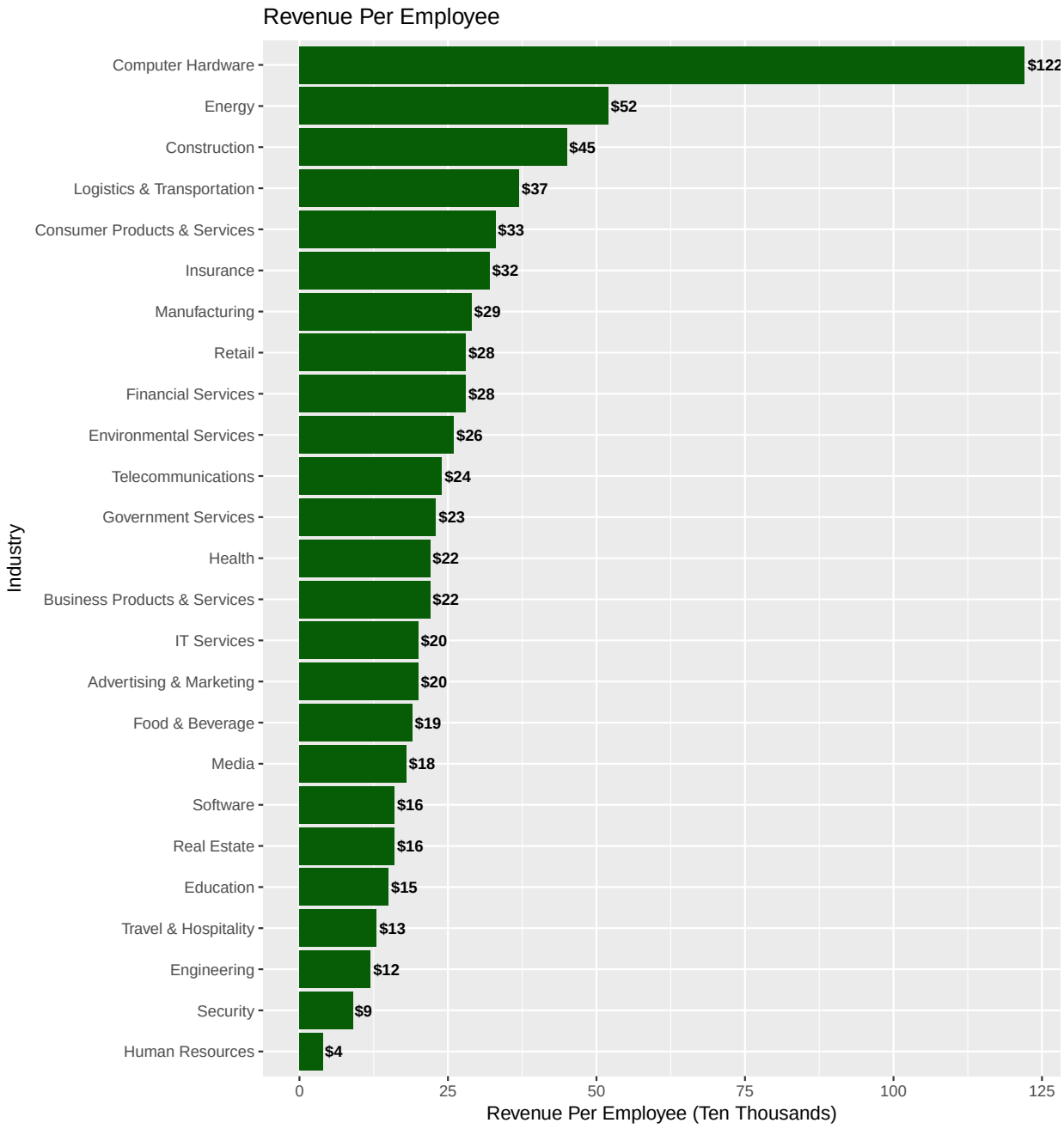
Question 3

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.

- I will use only complete cases. I wil also use the entire data set, assuming this question is independent of Question 2.

```
# Answer Question 3 here
revenues <- inc[complete.cases(inc),] %>%
  group_by(Industry) %>%
  summarize(RPE = round((sum(Revenue)/sum(Employees))/10000))

  ggplot(revenues, aes(x = reorder(Industry,RPE), y = RPE,  fill=Industry, label = RPE)) +
  geom_bar( stat='identity', show.legend = F, fill="#065d06" ) +
  #scale_y_continuous(labels=dollar_format(suffix="OK")) +
  labs(title = "Revenue Per Employee", x = "Industry", y = "Revenue Per Employee (Ten Thousands)" ) +
  geom_text(data = revenues, size = 3, vjust = 0.5, color = "#000000", hjust=-0.1, label = dollar(reven
  coord_flip()
```


Revenue Per Employee
Computer Hardware
Energy
Construction
Logistics & Transportation
Consumer Products & Services
Insurance
Manufacturing
Retail
Financial Services
Environmental Services
Telecommunications
Government Services
Health
Business Products & Services
IT Services
Advertising & Marketing
Food & Beverage
Media
Software
Real Estate
Education
Travel & Hospitality
Engineering
Security
Human Resources
$122
$52
$45
$37
$33
$32
$29
$28
$28
$26
$24
$23
$22
$22
$20
$20
$19
$18
$16
$16
$15
$13
$12
$9
$4
Industry
0
25
50
75
100
125
Revenue Per Employee (Ten Thousands)