DATA608 HW1

Bin Lin

2017-9-10

```
#install.packages("ggplot2")
#install.packages("dplyr")
#install.packages("stringi")

suppressWarnings(library(ggplot2))
suppressWarnings(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
```

```
raw_data <- read.csv("https://raw.githubusercontent.com/blin261/608/master/inc5000_data.csv")
str(raw_data)
```

```
## 'data.frame':    5001 obs. of  8 variables:
##  $ Rank       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Name       : Factor w/ 5001 levels "(Add)ventures",..: 1770 1633 4423 690 1198 2839 4733 1
468 1869 4968 ...
##  $ Growth_Rate: num  421 248 245 233 213 ...
##  $ Revenue    : num  1.18e+08 4.96e+07 2.55e+07 1.90e+09 8.70e+07 ...
##  $ Industry   : Factor w/ 25 levels "Advertising & Marketing",..: 5 12 13 7 1 20 10 1 5 21
 ...
##  $ Employees  : int  104 51 132 50 220 63 27 75 97 15 ...
##  $ City       : Factor w/ 1519 levels "Acton","Addison",..: 391 365 635 2 139 66 912 1179 131
1418 ...
##  $ State      : Factor w/ 52 levels "AK","AL","AR",..: 5 47 10 45 20 45 44 5 46 41 ...
```

```
head(raw_data)
```

```
##   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
```

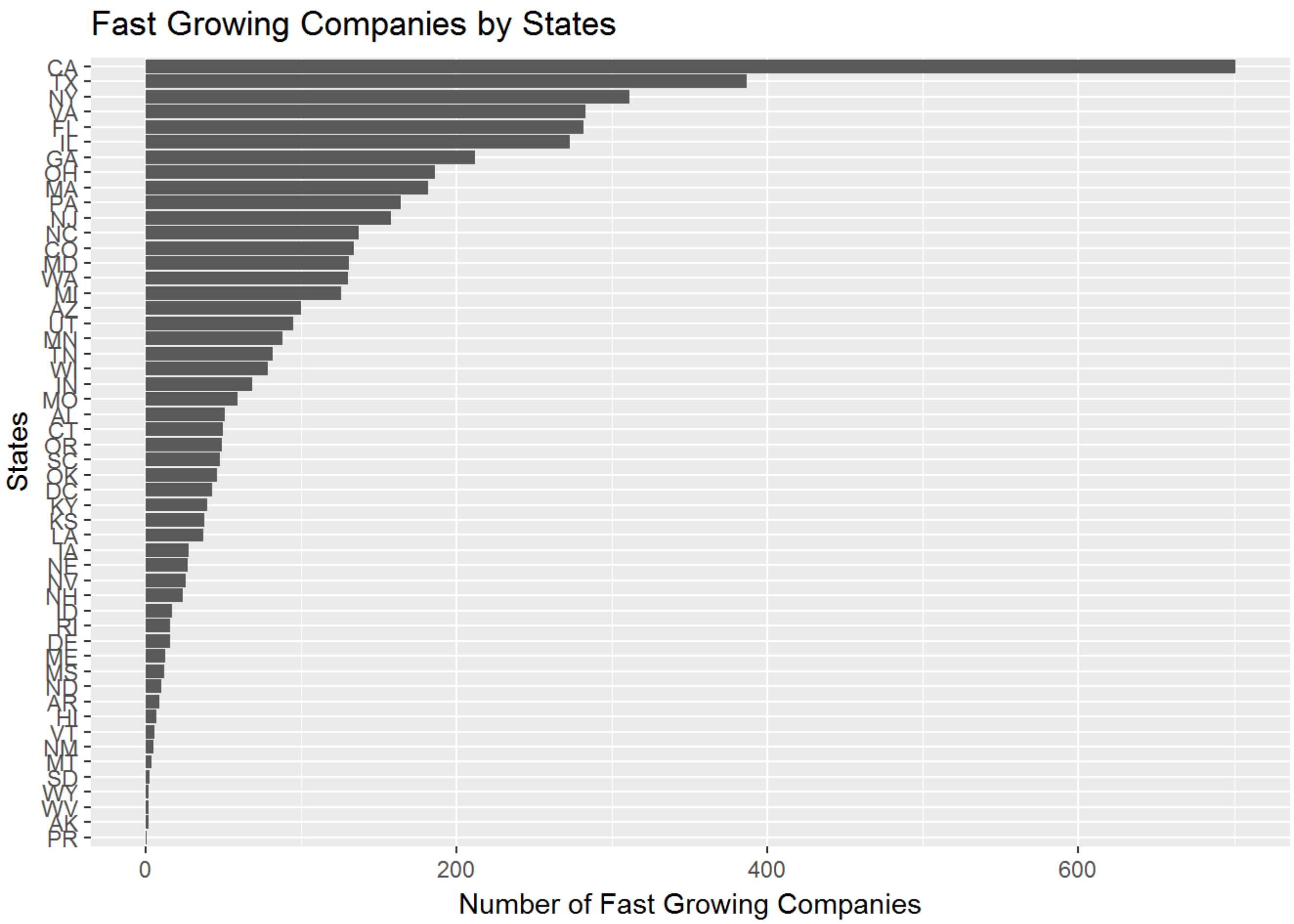
1. Create a graph that shows the distribution of companies in the dataset by State (ie how many are in each state). There are a lot of States, so consider which axis you should use assuming I am using a ‘portrait’ oriented screen (ie taller than wide).

```
state_data <- raw_data %>%
  group_by(State) %>%
  summarize(n = n()) %>%
  arrange(desc(n))

head(state_data)
```

```
## # A tibble: 6 x 2
##    State     n
##   <fctr> <int>
## 1     CA   701
## 2     TX   387
## 3     NY   311
## 4     VA   283
## 5     FL   282
## 6     IL   273
```

```
ggplot(data = state_data, aes(x = reorder(State, n), y =n)) + geom_bar(stat = "identity") + coor
d_flip() + ggtitle("Fast Growing Companies by States") + labs(x = "States", y = "Number of Fast
 Growing Companies")
```



```
ggsave("Figure1.jpg")
```

```
## Saving 7 x 5 in image
```

2. Let's dig in on the State with the 3 rd 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 employ. Create a plot of average employment by industry for companies in this state (only use cases with full data (user R's complete.cases() function). Your graph should show how variable the ranges are, and exclude outliers.

```
third_state <- state_data[3, 1]
#typeof(third_state)
ny_data <- filter(raw_data, State == unlist(third_state))
```

```
## Warning: package 'bindrcpp' was built under R version 3.3.3
```

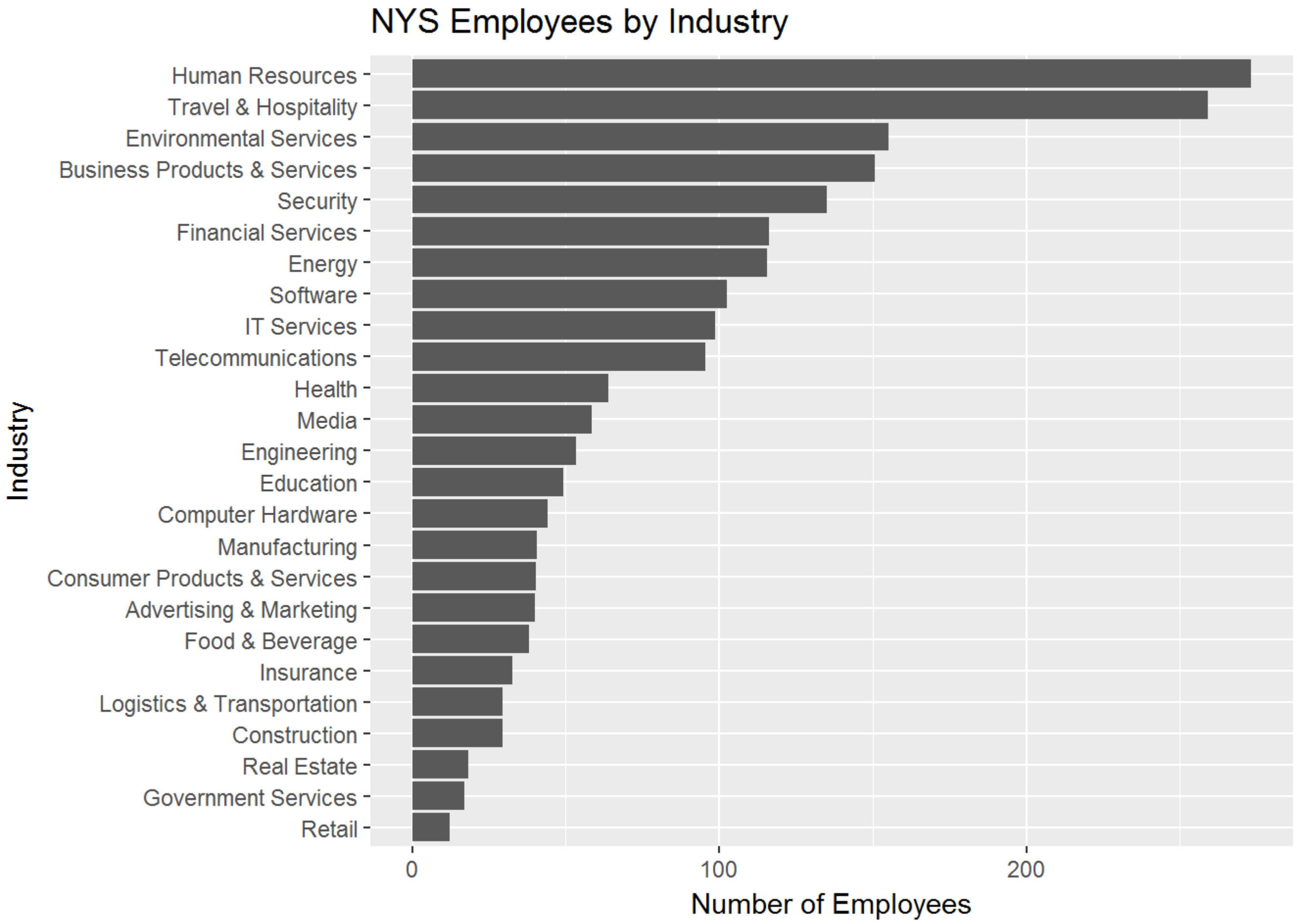
```
ny_data <- ny_data[complete.cases(ny_data), ]
```

```
industry_data <- ny_data %>%
  group_by(Industry) %>%
  filter(!Employees %in% boxplot.stats(Employees)$out) %>%
  summarize(average_emp = sum(Employees) / n())

head(industry_data)
```

```
## # A tibble: 6 x 2
##                         Industry average_emp
##                           <fctr>       <dbl>
## 1      Advertising & Marketing    40.05882
## 2 Business Products & Services   150.52174
## 3            Computer Hardware    44.00000
## 4                 Construction    29.40000
## 5 Consumer Products & Services    40.43750
## 6                    Education    49.07692
```

```
ggplot(data = industry_data, aes(x = reorder(Industry, average_emp), y = average_emp)) + geom_ba
r(stat = "identity") + coord_flip() + ggtitle("NYS Employees by Industry") + labs(x =
"Industry", y = "Number of Employees")
```



```
ggsave("Figure2.jpg")
```

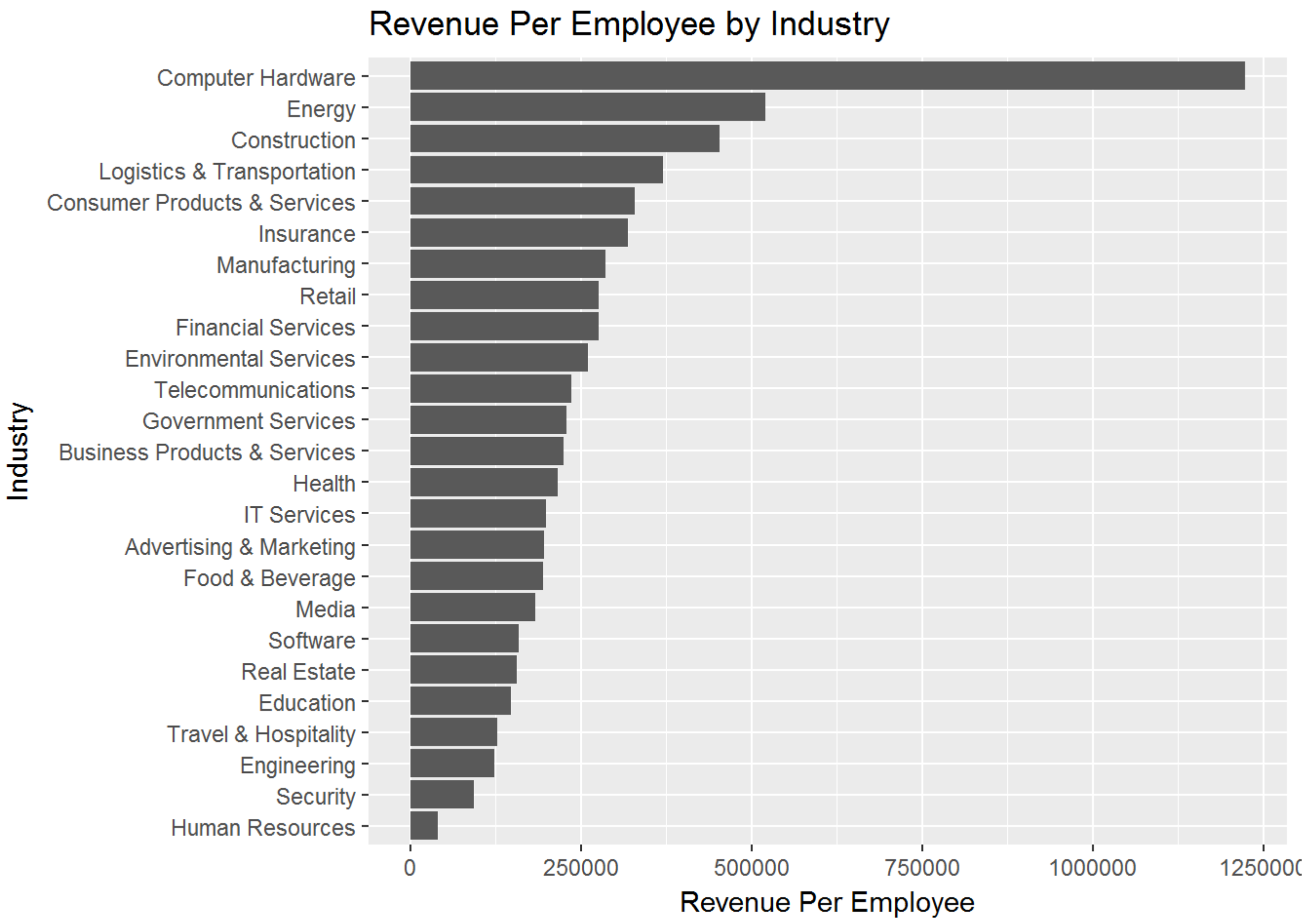
```
## Saving 7 x 5 in image
```

3. Now imagine you work for an investor and want to see which industries generate the most revenue per employee. Create a chart makes this information clear.

```
new_raw_data <- raw_data[complete.cases(raw_data), ]
revenue_data <- new_raw_data %>%
  group_by(Industry)%>%
  summarise(average_rev=(sum(Revenue)/ sum(Employees)))
head(revenue_data)
```

```
## # A tibble: 6 x 2
##                         Industry average_rev
##                           <fctr>       <dbl>
## 1      Advertising & Marketing    195942.7
## 2 Business Products & Services    224493.6
## 3            Computer Hardware   1223563.9
## 4                 Construction    452740.6
## 5 Consumer Products & Services    328972.4
## 6                    Education    148249.8
```

```
ggplot(data = revenue_data, aes(x = reorder(Industry, average_rev), y = average_rev)) +
geom_bar(stat = "identity") + coord_flip() + ggtitle("Revenue Per Employee by Industry") +
labs(x = "Industry", y = "Revenue Per Employee")
```



```
ggsave("Figure3.jpg")
```

```
## Saving 7 x 5 in image
```