DATA 608 - Project 1

Joshua Sturm

02/08/2018

```r
# Load packages
packages <- c("tidyverse")
invisible(lapply(packages, library, character.only = T))
```

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:

```r
inc <- read.csv("https://raw.githubusercontent.com/charleyferrari/CUNY_DATA_608/master/module1/Data/inc5
```

And lets preview this data:

```r
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
```

```r
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:

```r
glimpse(inc) # View number of rows and columns, variable types
```

```
## Observations: 5,001
## Variables: 8
## $ Rank        <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,...
## $ Name        <fct> Fuhu, FederalConference.com, The HCI Group, Bridge...
## $ Growth_Rate <dbl> 421.48, 248.31, 245.45, 233.08, 213.37, 179.38, 17...
## $ Revenue     <dbl> 1.179e+08, 4.960e+07, 2.550e+07, 1.900e+09, 8.700e...
## $ Industry    <fct> Consumer Products & Services, Government Services,...
## $ Employees   <int> 104, 51, 132, 50, 220, 63, 27, 75, 97, 15, 149, 16...
## $ City        <fct> El Segundo, Dumfries, Jacksonville, Addison, Bosto...
## $ State       <fct> CA, VA, FL, TX, MA, TX, TN, CA, UT, RI, VA, CA, FL...
```

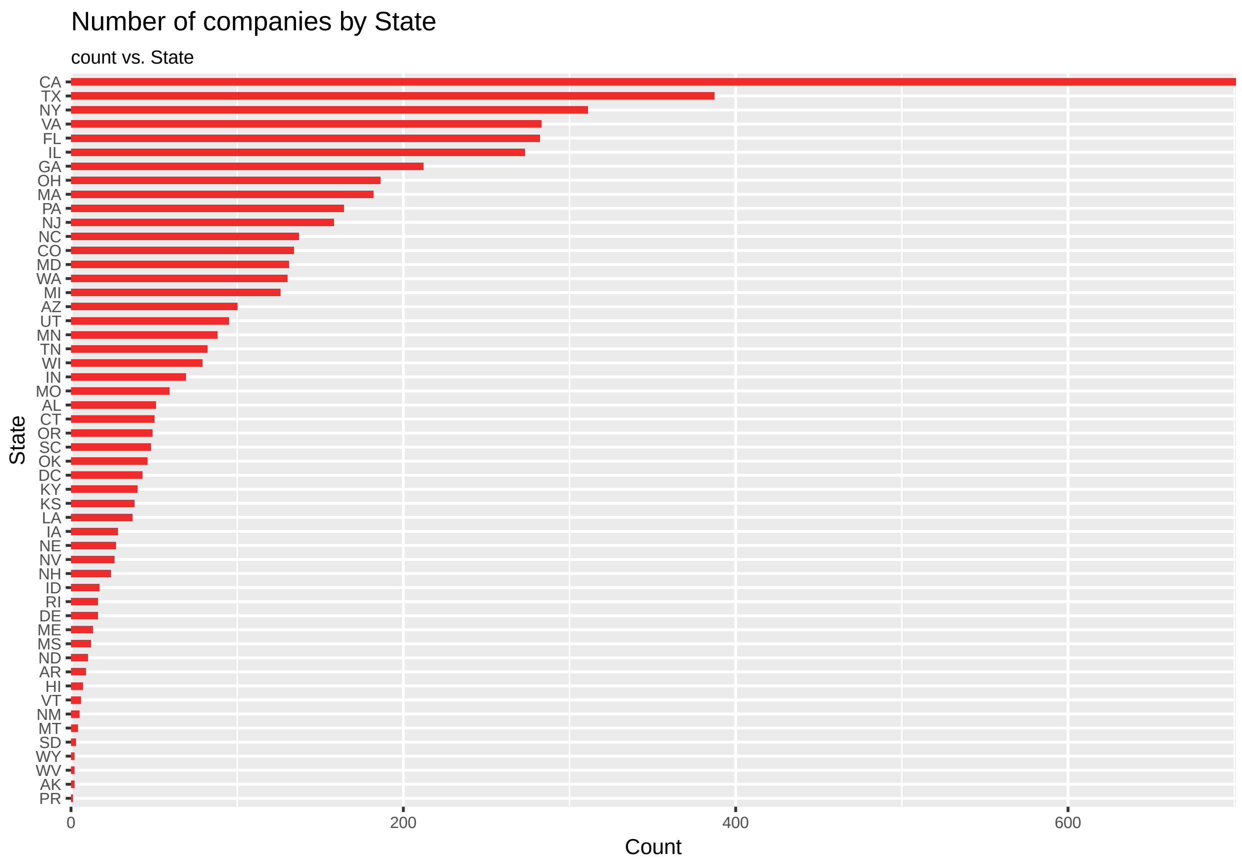
Question 1

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.

Since we'll be displaying the output on a portrait screen, we want to flip the coordinates, and use the y axis.

```r
# Group by state, and take the count
state.count <- inc %>%
  count(State)

ggplot(state.count, aes(x=reorder(State, n), y=n)) +
  geom_bar(stat="identity", fill="firebrick2", width=0.5) +
  coord_flip() +
  labs(title="Number of companies by State",
       subtitle="count vs. State",
       x = "State",
       y = "Count") +
  theme_grey(base_size = 8) +
  scale_y_continuous(expand=c(0,0))
```



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.

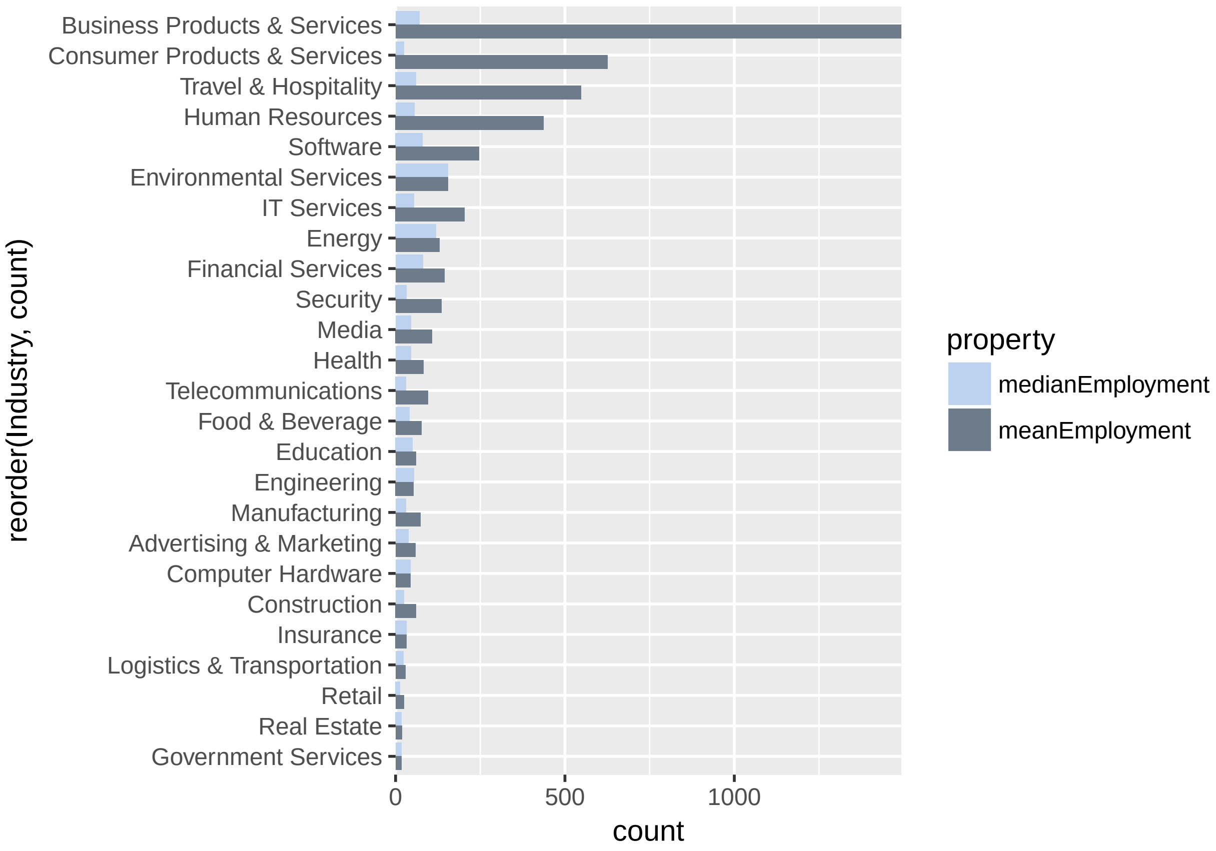
```
inc <- inc[complete.cases(inc),]

find.third.state <- state.count %>%
  arrange(desc(n))
find.third.state <- find.third.state$State[[3]]

third.state <- filter(inc, State == find.third.state) %>%
  filter(complete.cases(.))

third.state.table <- group_by(third.state, Industry) %>%
  summarize(meanEmployment = mean(Employees),
            medianEmployment = median(Employees)
  ) %>%
  gather(property, count, meanEmployment, medianEmployment)
```

```
ggplot(third.state.table, aes(x=reorder(Industry, count), y=count)) +
  geom_bar(stat="identity", position="dodge", aes(fill=property)) +
  coord_flip() +
  scale_fill_manual(values=c("lightsteelblue4", "lightsteelblue2"), guide=guide_legend(reverse=T)) +
  scale_y_continuous(expand=c(0,0))
```



`Business Products & Services` appears to be an outlier. If we check the average difference between industry means, we can somewhat regulate the outlier.

```
state.outlier <- third.state.table %>%
  filter(property == "meanEmployment")
state.outlier <- state.outlier[-c(2),] # drop the outlier
mean(diff(sort(state.outlier$count))) # calculate the average difference
```

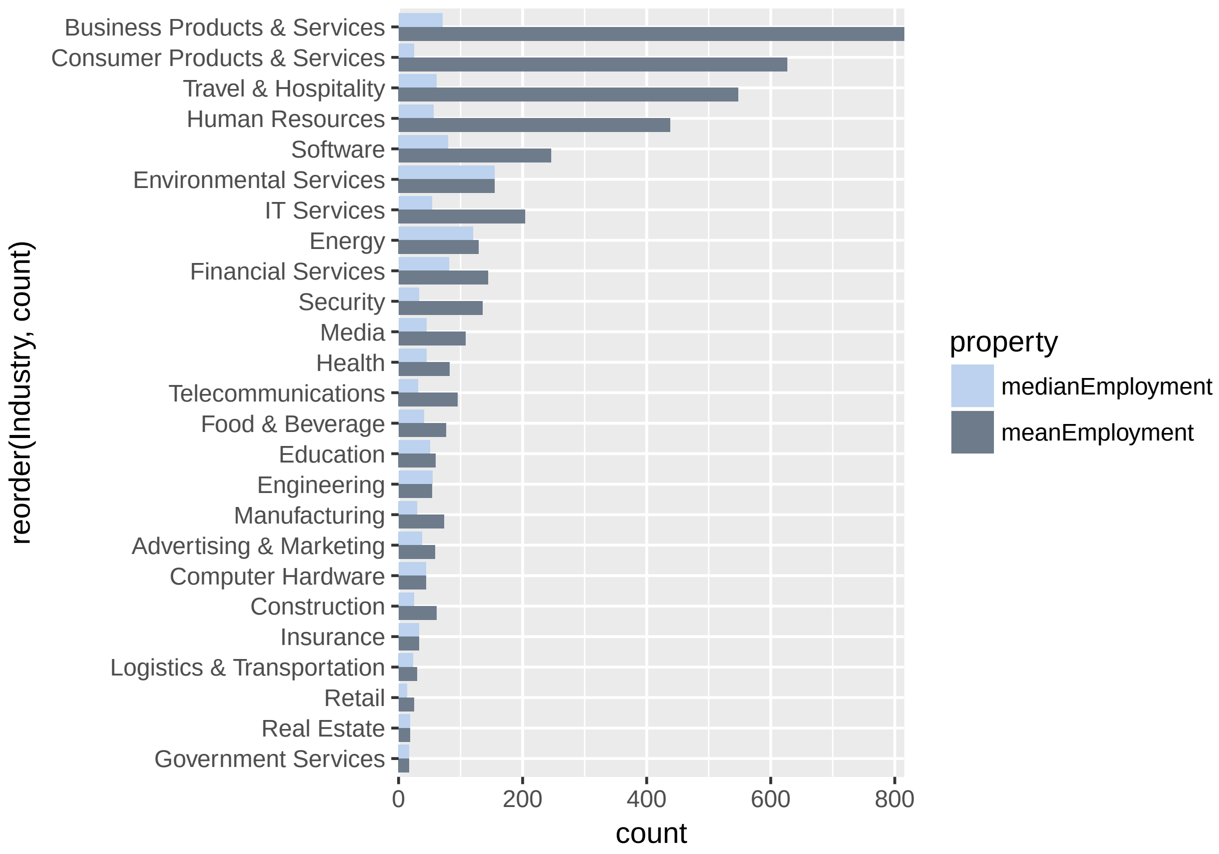
```
## [1] 26.49105
```

```
max(diff(sort(state.outlier$count)))
```

```
## [1] 191.6224
```

We see the average difference between consecutive leading industries is 26.49105, with a max of 191.6224. With this in mind, I think we can cap the outlier at ~200 more than the second-highest.

```
third.state.edited <- third.state.table
third.state.edited[2,3] <- 815
ggplot(third.state.edited, aes(x=reorder(Industry, count), y=count)) +
  geom_bar(stat="identity", position="dodge", aes(fill=property)) +
  coord_flip() +
  scale_fill_manual(values=c("lightsteelblue4", "lightsteelblue2"), guide=guide_legend(reverse=T)) +
  scale_y_continuous(expand=c(0,0))
```

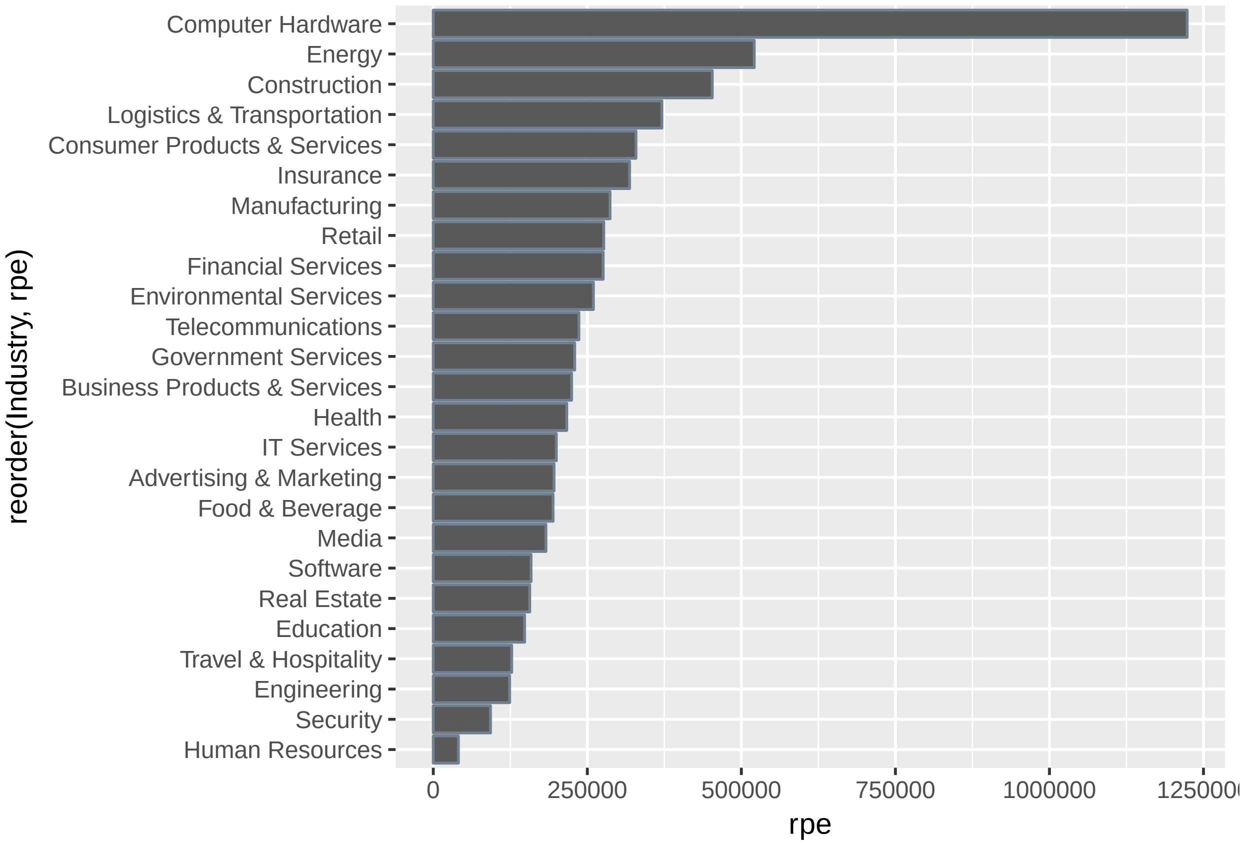


Question 3

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.

```
most.profitable <- inc %>%
  group_by(Industry) %>%
  summarize(rpe = sum(Revenue)/sum(Employees))

ggplot(most.profitable, aes(x=reorder(Industry, rpe), y=rpe)) +
  geom_bar(stat="identity", colour="slategrey") +
  coord_flip()
```



Once again, we have an outlier; in this case, it's `Computer Hardware`. Using the same method as in problem 2, we'll cap the outlier.

```
mp.edited <- most.profitable
mean(diff(sort(mp.edited$rpe)))
```

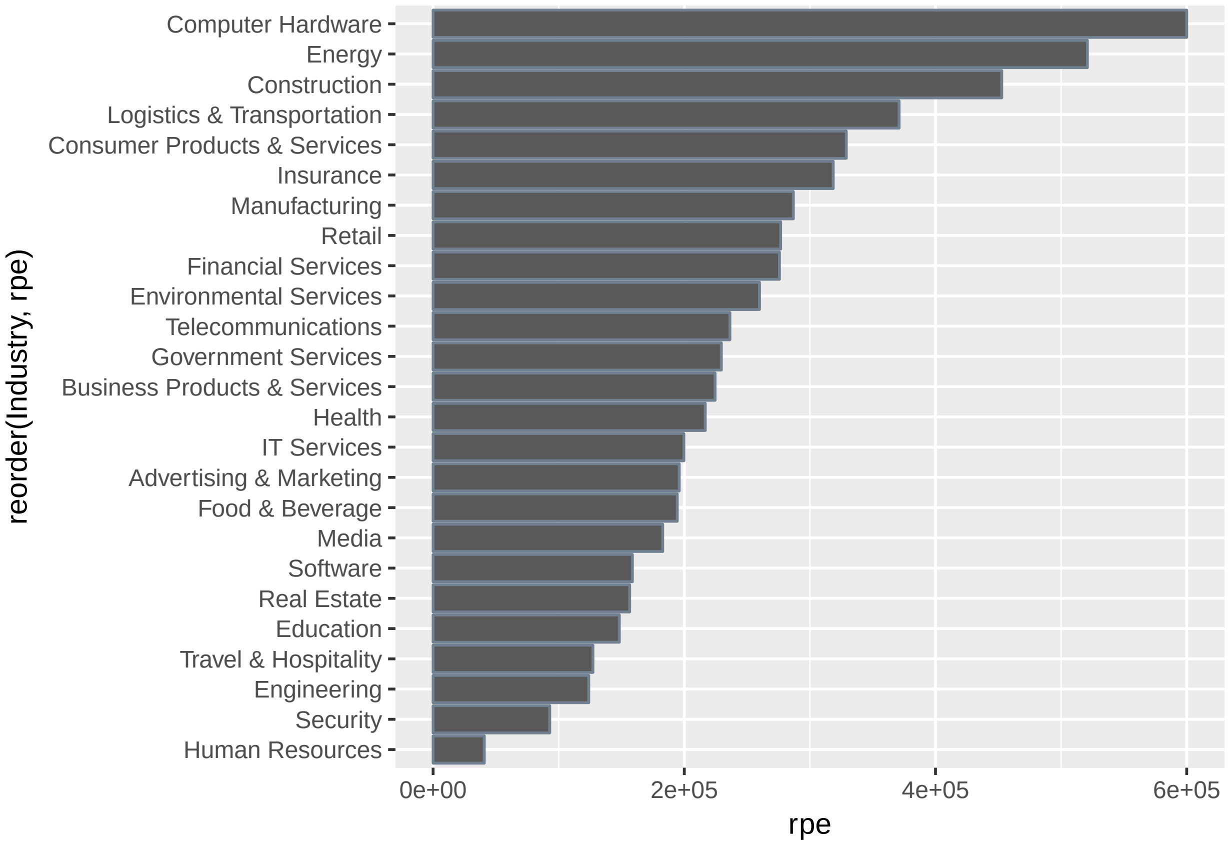
```
## [1] 49284.53
```

```
max(diff(sort(mp.edited$rpe)))
```

```
## [1] 702642.5
```

```
mp.edited[3,2] <- 600000

ggplot(mp.edited, aes(x=reorder(Industry, rpe), y=rpe)) +
  geom_bar(stat="identity", colour="slategrey") +
  coord_flip()
```


Computer Hardware
Energy
Construction
Logistics & Transportation
Consumer Products & Services
Insurance
Manufacturing
Retail
Financial Services
Environmental Services
Telecommunications
Government Services
Business Products & Services
Health
IT Services
Advertising & Marketing
Food & Beverage
Media
Software
Real Estate
Education
Travel & Hospitality
Engineering
Security
Human Resources
reorder(Industry, rpe)
0e+00
2e+05
4e+05
6e+05
rpe