Hantz_Angrand_Data608_project1

Hantz Angrand

February 9, 2019

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(tidyverse)
```

```
## -- Attaching packages -------------------------------------------------------------- tidyverse 1.2.1 -
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.6
## v tidyr   0.8.1     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
```

```
## -- Conflicts ----------------------------------------------------------------- tidyverse_conflicts() -
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
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
names(inc)
```

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

```
#removing Na from the dataset
inc_na<-na.omit(inc)
head(inc_na)
```

```
##   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
```

Aggregate to get the frequency of employee by industry

```
#indeed_skillaggr<-aggregate(read_indeed_url$Count,by=list(Category=read_indeed_url$Skills), FUN=sum)
#indeed_skillaggr

inc_na_aggr<-aggregate(inc_na$Employees, by=list(Categgory=inc_na$Industry), FUN=sum)
inc_na_aggr
```

```
##                       Categgory      x
## 1       Advertising & Marketing  39731
## 2  Business Products & Services 117357
## 3             Computer Hardware   9714
## 4                  Construction  29099
## 5  Consumer Products & Services  45464
## 6                     Education   7685
## 7                        Energy  26437
## 8                   Engineering  20435
## 9        Environmental Services  10155
## 10           Financial Services  47693
## 11              Food & Beverage  65911
## 12          Government Services  26185
## 13                       Health  82430
## 14              Human Resources 226980
## 15                    Insurance   7339
## 16                  IT Services 102788
## 17   Logistics & Transportation  39994
## 18                Manufacturing  43942
## 19                        Media   9532
## 20                  Real Estate  18893
## 21                       Retail  37068
## 22                     Security  41059
## 23                     Software  51262
## 24           Telecommunications  30842
## 25         Travel & Hospitality  23035
```

Revenue by state

```
#skills_count<-read_indeed_url %>%
#  group_by(Skills) %>%
#  summarise(Total=sum(Count)) %>%
#  arrange(desc(Total))

#skills_count

revenue_by_state<-inc_na %>%
  group_by(State) %>%
  summarise(Total=sum(Revenue)) %>%
  arrange(desc(Total))

revenue_by_state
```

```
## # A tibble: 52 x 2
##    State       Total
##    <fct>       <dbl>
##  1 IL    33238800000
##  2 CA    23364600000
##  3 TX    22154300000
##  4 NY    18260400000
##  5 OH    12786600000
##  6 FL    10610300000
##  7 NC     9252500000
##  8 VA     8667700000
##  9 MI     7805800000
## 10 WI     7131400000
## # ... with 42 more rows
```

```
#skills_city<-read_indeed_url %>%
 # group_by(Skills,City) %>%
#  summarise(Total=sum(Count)) %>%
#  arrange(desc(Total))

#skills_city

rev_by_ind_state<-inc_na %>%
  group_by(Industry, State) %>%
  summarise(Total=sum(Revenue)) %>%
  arrange(desc(Total))

rev_by_ind_state
```

```
## # A tibble: 798 x 3
## # Groups:   Industry [25]
##    Industry                     State       Total
##    <fct>                        <fct>       <dbl>
##  1 Computer Hardware            IL    10261300000
##  2 Energy                       TX     7800800000
##  3 Food & Beverage              IL     6239000000
##  4 Business Products & Services IL     5733100000
##  5 Construction                 WI     4847200000
##  6 IT Services                  NY     4826200000
##  7 Consumer Products & Services NY     4799300000
##  8 Government Services          VA     3822300000
##  9 Consumer Products & Services NC     3507100000
## 10 Financial Services           CA     3444200000
## # ... with 788 more rows
```

Question 1

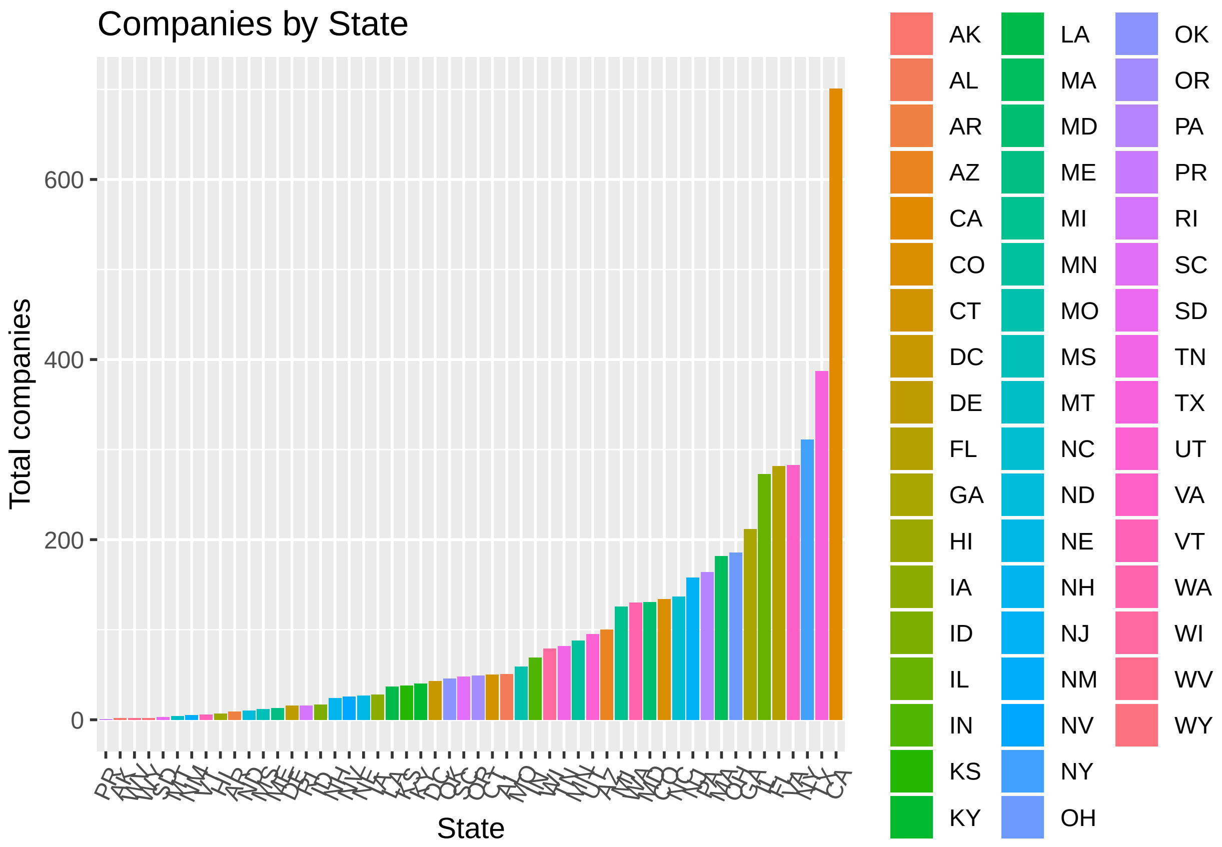
Create a graph that shows the distribution of companies in the dataset by State (ie how many are in each state). There are a lot of States, so consider which axis you should use. This visualization is ultimately going to be consumed on a 'portrait' oriented screen (ie taller than wide), which should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

```
# Answer Question 1 here
inc_state<-inc %>%
  group_by(State) %>%
  summarise(Total=n()) %>%
  arrange(desc(Total))

inc_state
```

```
## # A tibble: 52 x 2
##    State Total
##    <fct> <int>
##  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
## # ... with 42 more rows
```

Graph distribution of companies by state

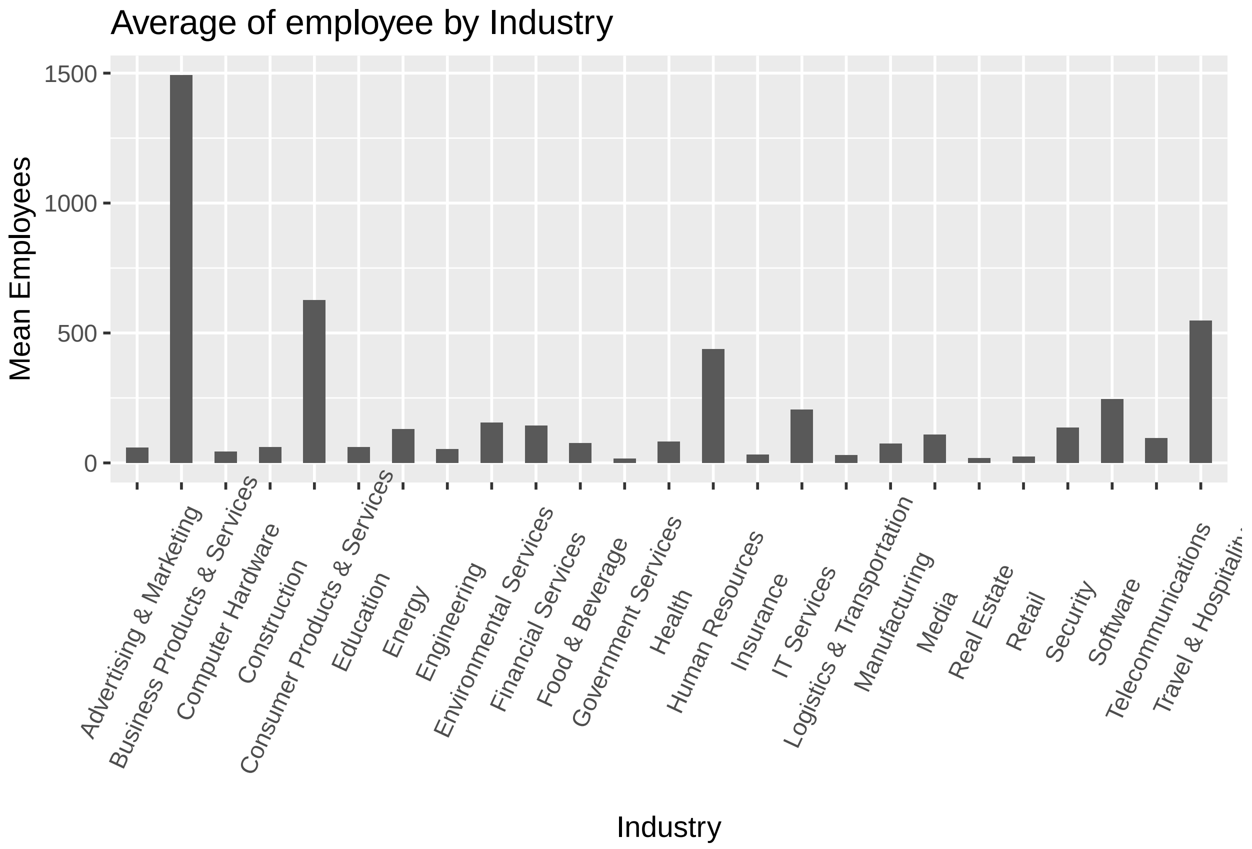
```
#g<-ggplot(inc_state,aes(x=reorder(State,Total, height=1),y=Total))+ geom_bar(stat='identity')+
  #coord_flip()+
#  labs(title='Frequency of Companies by State') +
 # xlab('State')+
 # ylab('Total')
#g

ggplot(data=inc_state, aes(x=reorder(State, Total),y=Total, fill=State)) +
    geom_bar(stat= "identity") +
    #guides(fill=FALSE) +
    xlab("State") + ylab("Total companies") +
    ggtitle("Companies by State") +
    theme(axis.text.x = element_text(angle=65, vjust=0.6))
```



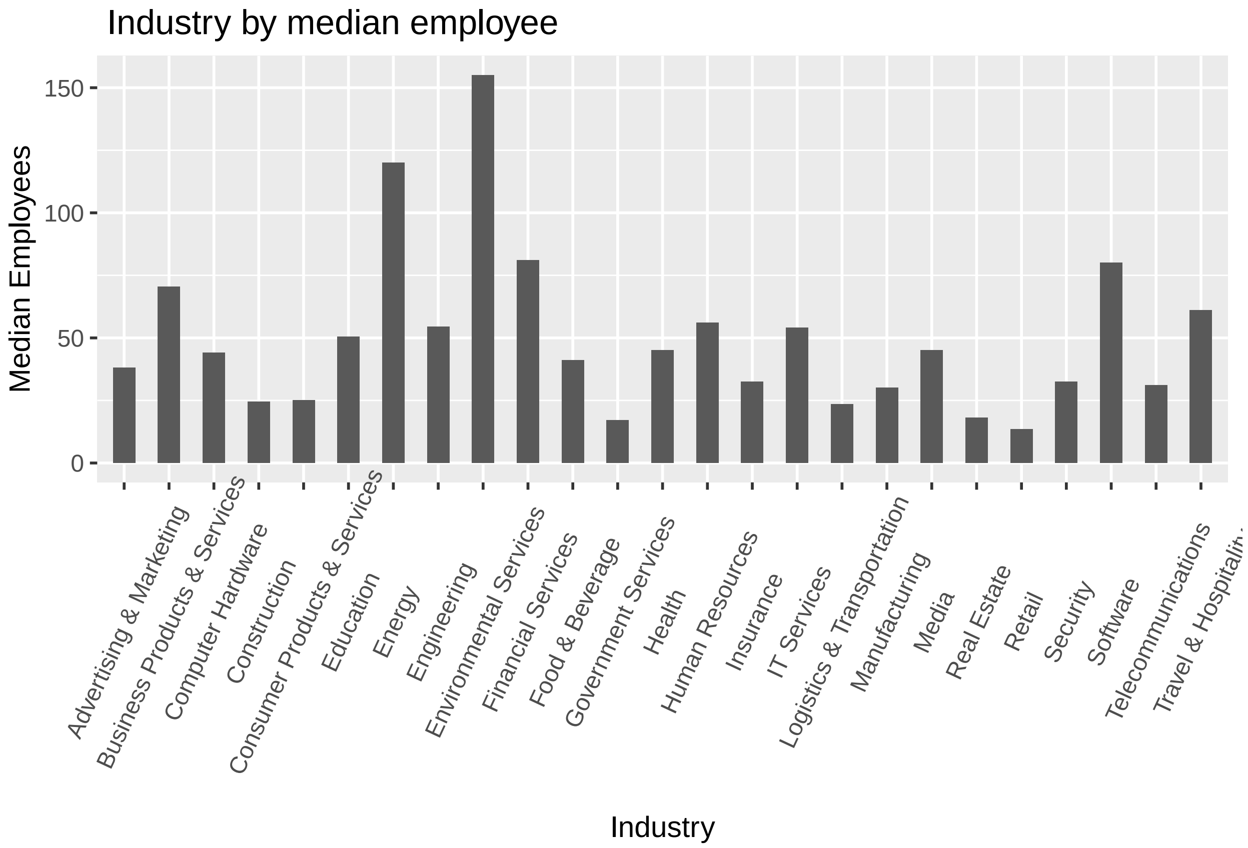
Quesiton 2

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.

```
# Answer Question 2 here
inc_select<-  inc%>%select(c(State, Industry, Employees))

inc_select<-inc_select[complete.cases(inc_select),]

inc_mean<-inc_select %>%
  filter(State == 'NY') %>%
  group_by(Industry) %>%
  summarise(mean=mean(Employees), median=median(Employees))

ggplot(inc_mean, aes(x=Industry, y=mean)) +
  geom_bar(stat="identity", width = 0.5) +
  ggtitle("Average of employee by Industry")+
  xlab("Industry")+
  ylab("Mean Employees")+
    theme(axis.text.x = element_text(angle=65, vjust=0.6))
```



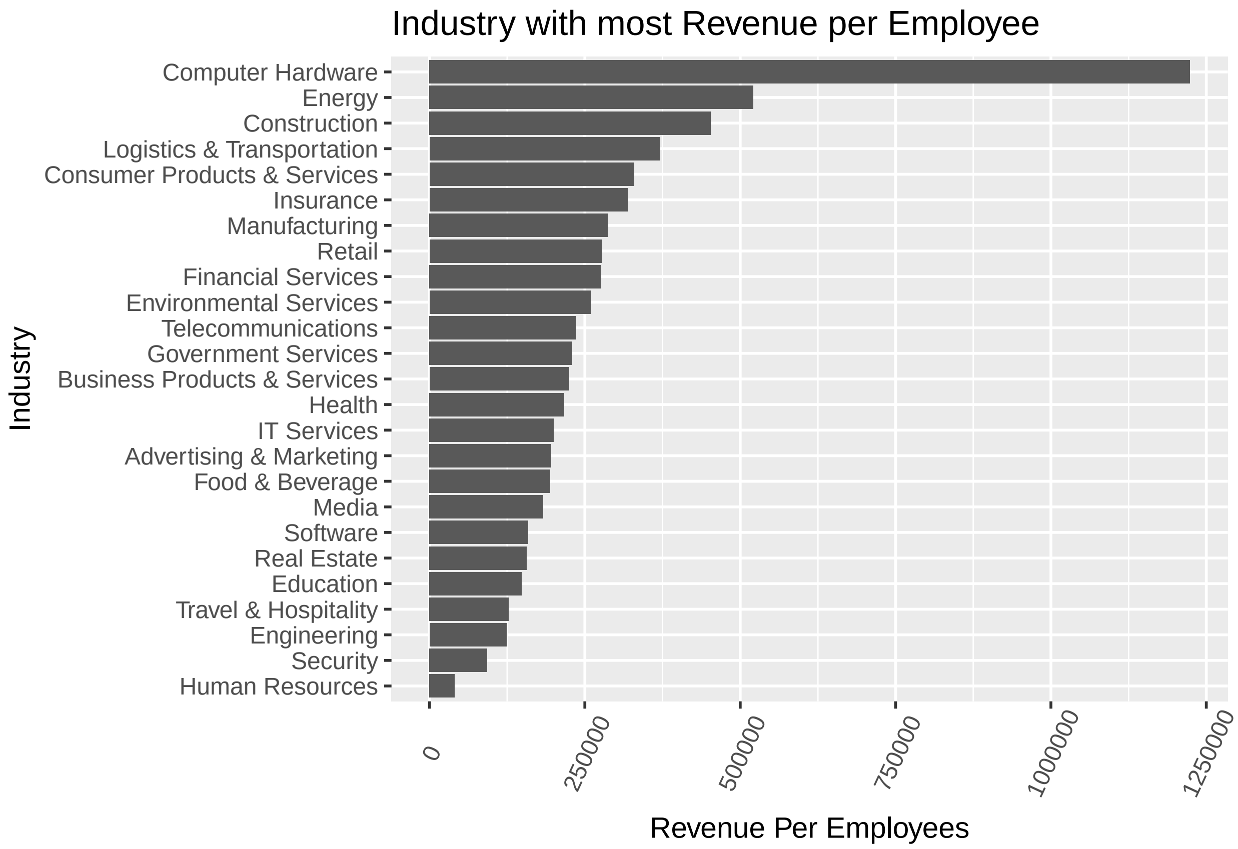
```
ggplot(inc_mean, aes(x=Industry, y=median)) +
  geom_bar(stat="identity", width = 0.5) +
  ggtitle(" Industry by median employee")+
  xlab("Industry")+
  ylab("Median Employees")+
    theme(axis.text.x = element_text(angle=65, vjust=0.6))
```



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
inc_investor<-inc %>%
  select(c(Industry,Revenue, Employees))

inc_investor<-inc_investor[complete.cases(inc_investor),]

inc_rev_total<-inc_investor %>%
  group_by(Industry) %>%
  summarise(Revenue_Total=sum(Revenue),Employees_Total=sum(Employees))

inc_revenue_emp<-transform(inc_rev_total, rev_per_emp= Revenue_Total / Employees_Total)%>%
  arrange(desc(rev_per_emp))

ggplot(inc_revenue_emp, aes(x=reorder(Industry,rev_per_emp), y=rev_per_emp))+
  geom_bar(stat="identity")+
  coord_flip()+
  ggtitle("Industry with most Revenue per Employee")+
  xlab("Industry")+
  ylab("Revenue Per Employees") +
  theme(axis.text.x = element_text(angle=65, vjust=0.6))
```


Industry with most Revenue per Employee
Industry
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
Revenue Per Employees