R Notebook

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(ggplot2)
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
```

```
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:

```
# Looking at the structure of the dataset

str(inc)
```

```
## '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 1468 186
##  $ 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 ...
```

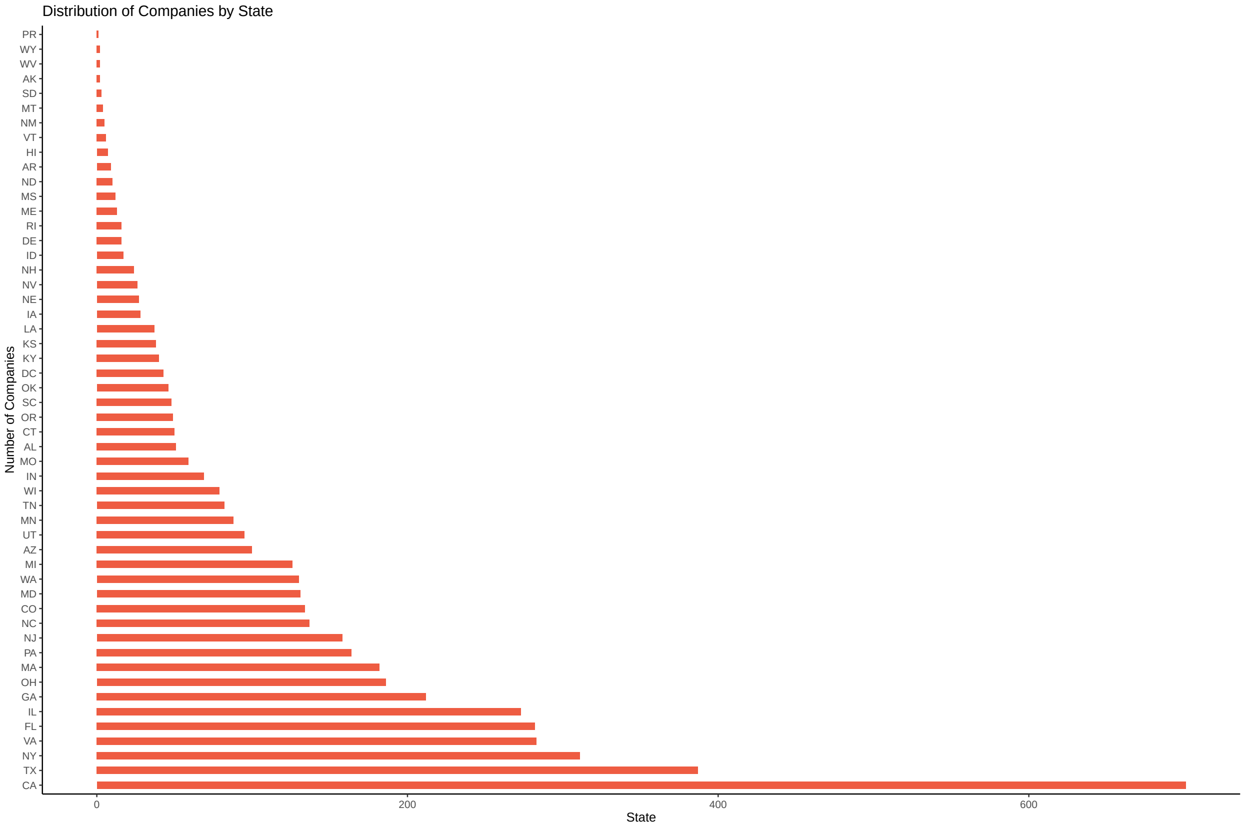
Question 1

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 1 - Option 1

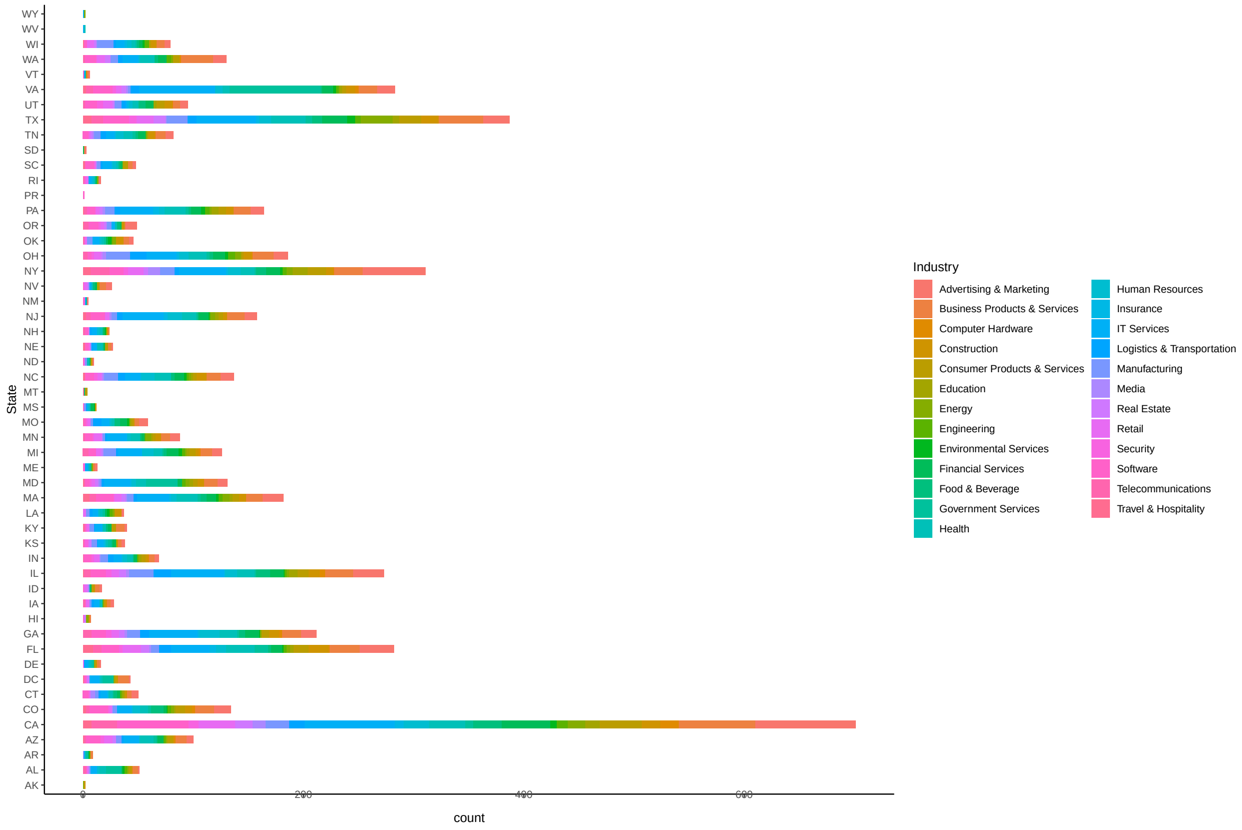
```
# Frequency of each state in the dataframe
state_inc <- inc %>%
  group_by(State) %>%
  summarize(Freq=n()) %>%
  arrange(desc(Freq))

# create a barplot of the new dataframe
theme_set(theme_classic())
ggplot(state_inc, aes(x=reorder(State, -Freq), y=Freq))+
  geom_bar(stat="identity", width=0.5, fill="tomato2")+
  coord_flip()+
  labs(title="Distribution of Companies by State")+
  xlab("Number of Companies")+
  ylab("State")+
  theme(axis.text.x=element_text(vjust=0.6))
```



Answer 1 Option 2

```
#create a barplot as option 2
ggplot(inc, aes(State))+
  geom_bar(aes(fill=Industry), width=0.5)+
  coord_flip()+ # unfortunately i couldnt get the flip work (also couldnt get the reorder work in this
  theme(axis.text.x = element_text(vjust=05))
```



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 2 Option 1

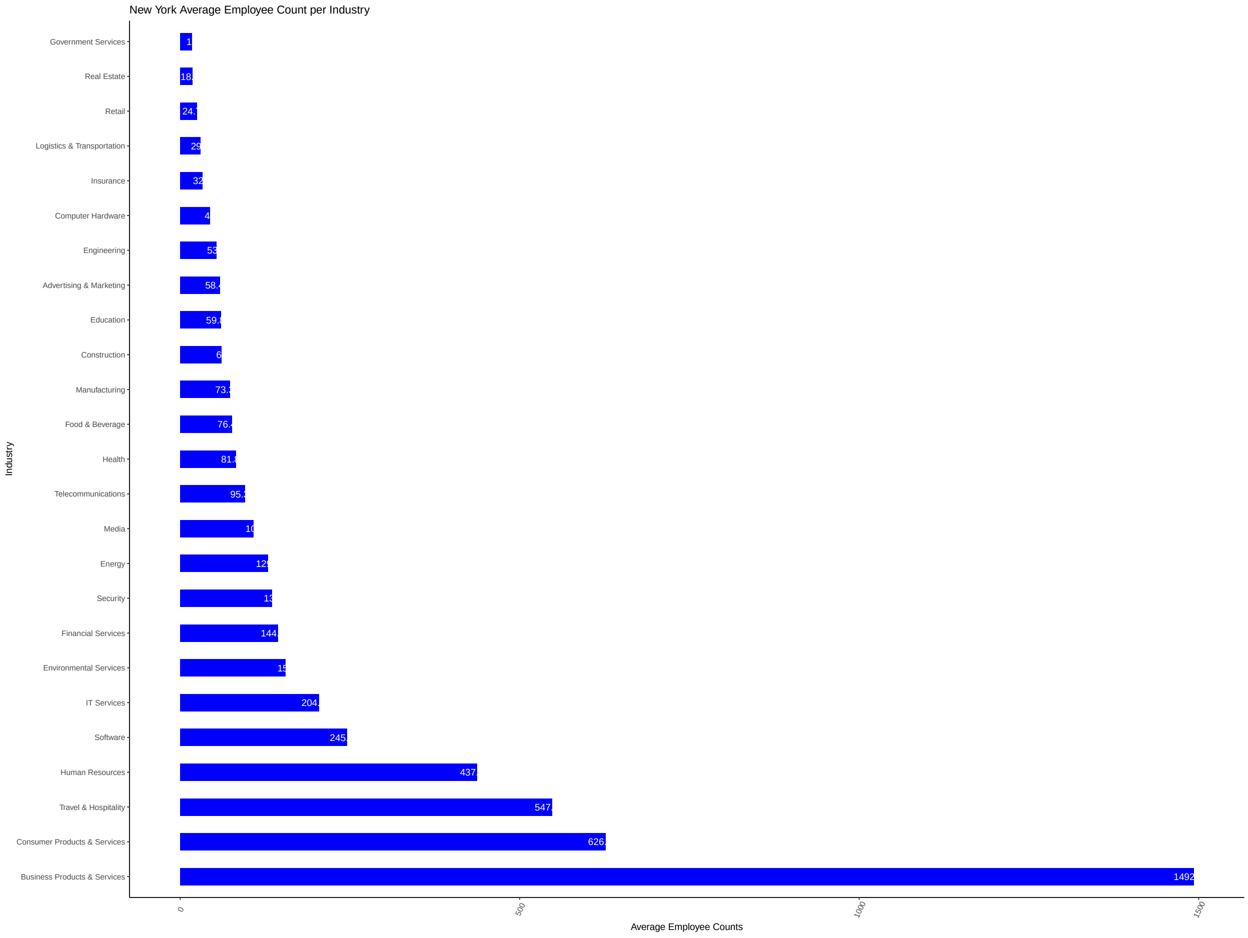
```r
# Filter the third state(NY)
ny <- inc %>%
  filter(State=="NY")

# Use complete.cases() for full available data
ny <- ny[complete.cases(ny), ]

# group by industry and mean of employees
ny <- ny %>%
  group_by(Industry) %>%
  summarize(avgemp=mean(Employees, na.rm = TRUE))

# create a bar plot
theme_set(theme_classic())
```

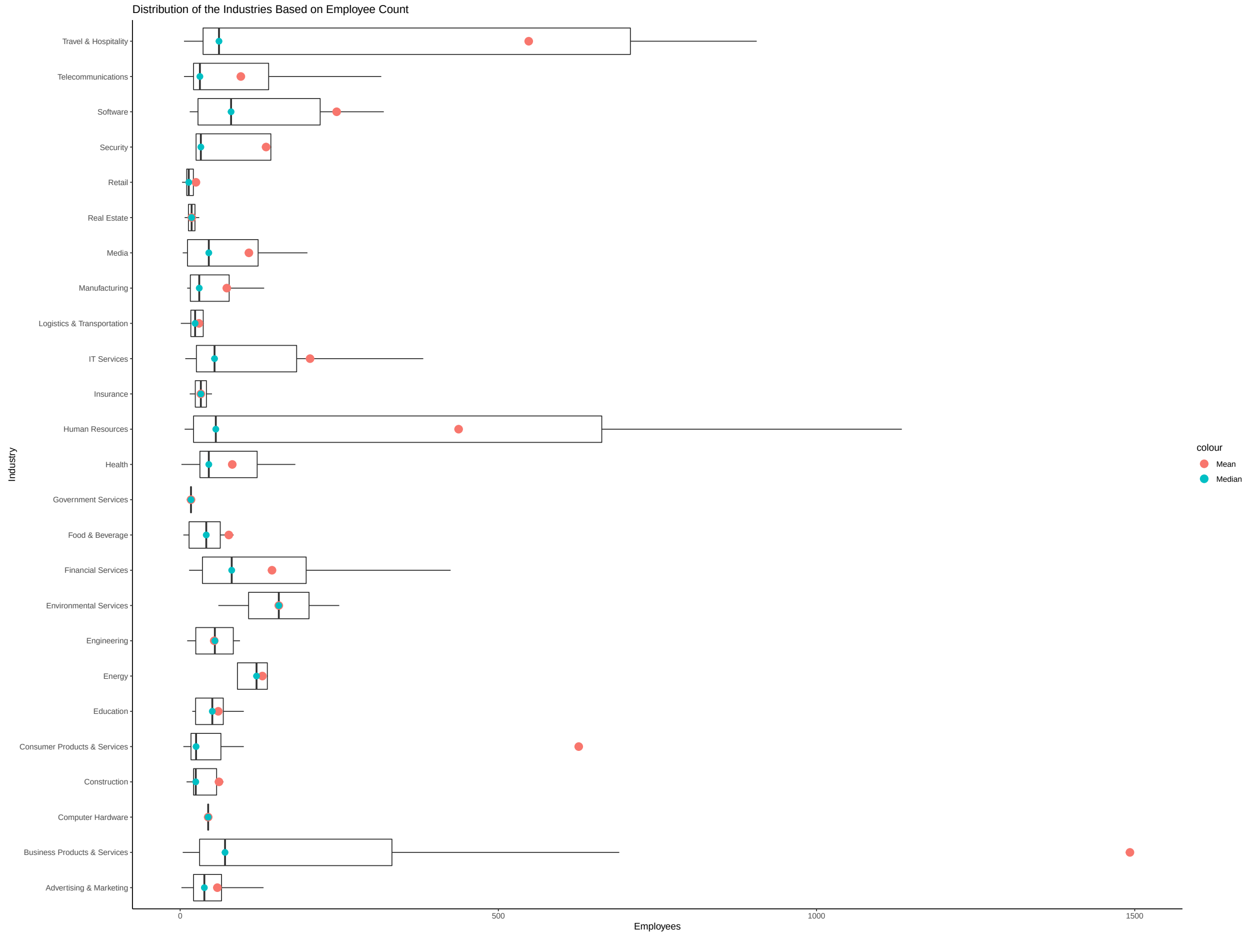
```
ggplot(ny, aes(x=reorder(Industry, -avgemp), y=avgemp))+
  geom_bar(stat="identity", width=0.5, fill="Blue")+
  coord_flip()+
  labs(title="New York Average Employee Count per Industry")+
  xlab("Industry")+
  ylab("Average Employee Counts")+
  theme(axis.text.x=element_text(angle = 65, vjust=0.6))+
  geom_text(aes(y=avgemp, label=round(avgemp,3)), color="white", size=4)
```



Answer 2 Option 2

```
# Filter the third state(NY) for full available data
ny_2 <- inc[complete.cases(inc), ]%>%
  filter(State=="NY")

# boxplot with mean and median of employees count within industry
ggplot(ny_2, aes(x=Industry, y=Employees))+
  geom_boxplot(outlier.colour = NA)+
  coord_flip(ylim=c(0,1500))+
  stat_summary(fun.y="mean", size=4, geom = "point", aes(color="Mean"))+
  stat_summary(fun.y="median", size=3, geom = "point", aes(color="Median"))+
  labs(title="Distribution of the Industries Based on Employee Count")+
  xlab("Industry")+
  ylab("Employees")
```

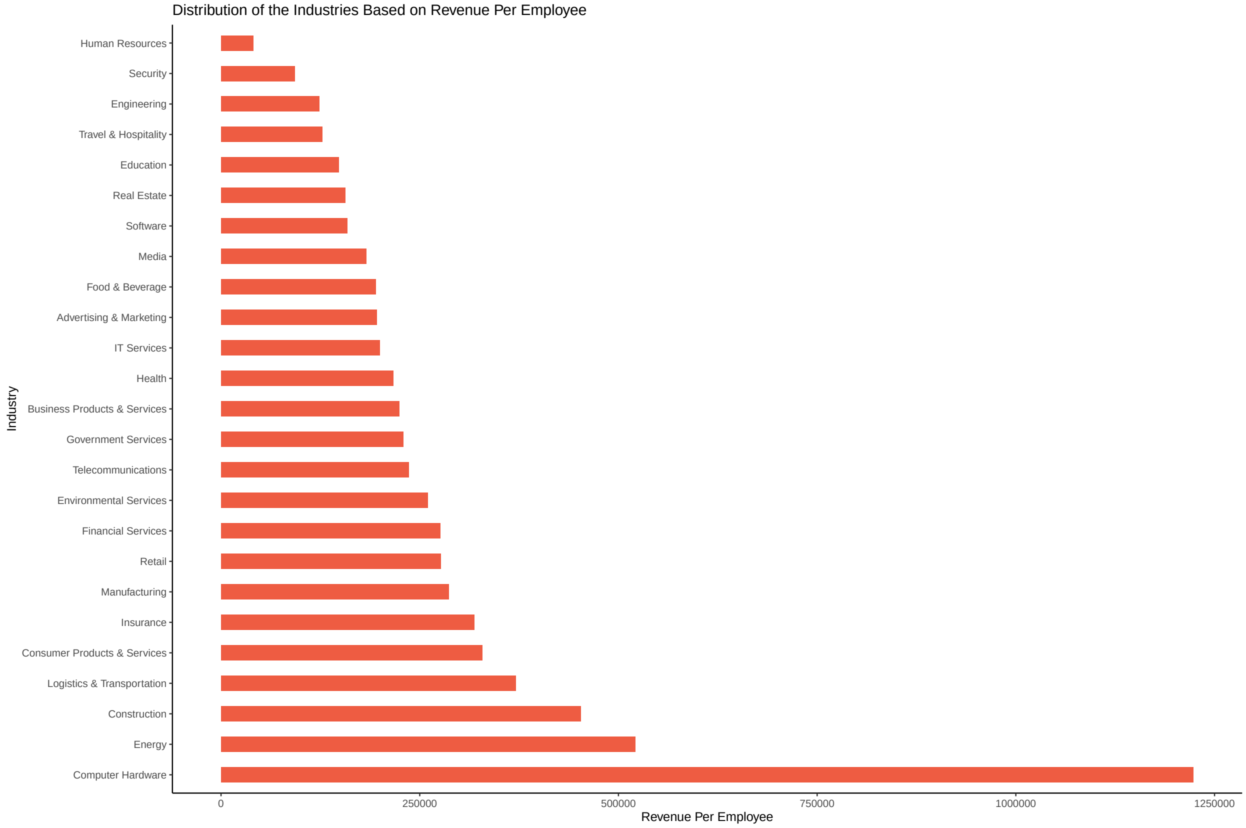


Question 3

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.

```
# calculate total revenue, total employee and revenue for employee
emp_revenue <- inc[complete.cases(inc), ]%>%
  group_by(Industry) %>%
  summarize(total_revenue=sum(Revenue), total_employee=sum(Employees)) %>%
  mutate(rev_per_emp=total_revenue/total_employee)

# create bar plot of industry distribution based on revenue per employee
ggplot(emp_revenue, aes(x=reorder(Industry, -rev_per_emp), y=rev_per_emp))+
  geom_bar(stat="identity", width=0.5, fill="tomato2")+
  coord_flip()+
  labs(title="Distribution of the Industries Based on Revenue Per Employee")+
  xlab("Industry")+
  ylab("Revenue Per Employee")+
  theme(axis.text.x=element_text(vjust=0.6))
```


Distribution of the Industries Based on Revenue Per Employee
Industry
Human Resources
Security
Engineering
Travel & Hospitality
Education
Real Estate
Software
Media
Food & Beverage
Advertising & Marketing
IT Services
Health
Business Products & Services
Government Services
Telecommunications
Environmental Services
Financial Services
Retail
Manufacturing
Insurance
Consumer Products & Services
Logistics & Transportation
Construction
Energy
Computer Hardware
Revenue Per Employee