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:

```
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:

1) Import the required packages

```
library(ggplot2)
library(kableExtra)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:kableExtra':
##
##     group_rows

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

If we analyze the growth rates, min is 0.340 and it maximum growth rate is 421.480. Let's see how many companies experianced growth rates of 50 or higher:

```
inc %>% dplyr::filter(Growth_Rate >= 50) %>% summarise(n = n())
```

```
##    n
## 1 59
```

Therefore, there were 59 companies with growth greater than or equal to 50.

And below is the list of those companies:

```
kable(inc %>% dplyr::filter(Growth_Rate >= 50)) %>% kable_styling()
```

Rank	Name	Growth_Rate	Revenue	Industry	Employees	C
1	Fuhu	421.48	1.179e+08	Consumer Products & Services	104	F
2	FederalConference.com	248.31	4.960e+07	Government Services	51	D
3	The HCI Group	245.45	2.550e+07	Health	132	J
4	Bridger	233.08	1.900e+09	Energy	50	A
5	DataXu	213.37	8.700e+07	Advertising & Marketing	220	B
6	MileStone Community Builders	179.38	4.570e+07	Real Estate	63	A
7	Value Payment Systems	174.04	2.550e+07	Financial Services	27	N
8	Emerge Digital Group	170.64	2.390e+07	Advertising & Marketing	75	S
9	Goal Zero	169.81	3.310e+07	Consumer Products & Services	97	B
10	Yagoozon	166.89	1.860e+07	Retail	15	W
11	OBXtek	164.33	2.960e+07	Government Services	149	T
12	AdRoll	150.65	3.410e+07	Advertising & Marketing	165	S
13	uBreakiFix	141.02	1.700e+07	Retail	250	O
14	Sparc	128.63	2.110e+07	Software	160	C
15	LivingSocial	123.33	5.360e+08	Consumer Products & Services	4100	W
16	Amped Wireless	110.68	1.430e+07	Computer Hardware	26	C
17	Intelligent Audit	105.73	1.450e+08	Logistics & Transportation	15	R
18	Integrity Funding	104.62	1.110e+07	Financial Services	11	S
19	Vertex Body Sciences	100.10	1.180e+07	Food & Beverage	51	c
20	BlueKai	92.45	2.680e+07	Advertising & Marketing	107	C
21	Level 11	90.44	9.600e+06	IT Services	30	S
22	Patient Conversation Media	87.82	9.800e+06	Health	41	A
23	Wingspan Portfolio Advisors	87.69	7.700e+07	Financial Services	1016	C
24	Vets First Choice	85.85	1.460e+07	Business Products & Services	74	P
25	Timberhorn	85.16	1.380e+07	IT Services	150	B
26	BeenVerified	84.43	1.370e+07	Consumer Products & Services	17	N
27	Trada	81.01	1.260e+07	Advertising & Marketing	75	B
28	Kony	77.86	5.110e+07	Software	1100	O
29	OneSource Virtual	73.53	2.350e+07	IT Services	260	I
30	Sailthru	73.22	8.100e+06	Advertising & Marketing	79	N
31	Innovolt	72.48	8.000e+06	Business Products & Services	30	A
32	Provider Power	72.24	5.680e+07	Energy	50	A
33	Zurple	71.12	7.700e+06	Software	32	C
34	US Logistics	70.87	4.830e+07	Logistics & Transportation	10	C
35	McAfee Institute	70.63	1.500e+07	Education	15	F
36	Now Communications	67.64	7.000e+06	Consumer Products & Services	110	T
37	YellowHammer	67.40	1.800e+07	Advertising & Marketing	27	N
38	Conductor	67.02	7.100e+06	Advertising & Marketing	89	N
39	Intellect Resources	65.54	3.000e+07	Health	675	C
40	Phunware	65.27	8.200e+06	Software	92	A
41	NSR Solutions	63.92	1.010e+07	Government Services	252	F
42	Pangea Properties	62.18	2.830e+07	Real Estate	264	C
43	Field Nation	60.31	2.770e+07	Business Products & Services	15	M
44	The Joint	59.62	7.400e+06	Health	14	S
45	Silver Spring Networks	58.67	1.967e+08	Energy	566	F
46	ThinkLite	55.25	8.500e+06	Energy	14	N
47	29 Prime	54.43	1.380e+07	Advertising & Marketing	152	I
48	Cinium Financial Services	53.65	5.900e+06	Financial Services	32	F
49	Solar Alliance of America	53.37	1.730e+07	Energy	4	S
50	Saratoga Roofing & Construction	53.28	2.930e+07	Construction	106	C
51	Madwire Media	52.54	6.900e+06	Advertising & Marketing	94	I
52	Eventus Solutions Group	52.21	6.400e+06	Business Products & Services	22	F
53	LabTech Software	52.08	2.310e+07	Software	152	T
54	Altitude Digital	51.62	1.140e+07	Advertising & Marketing	21	I
55	DSFederal	51.51	9.400e+06	Government Services	70	C
56	Pinnacle Strategies	51.34	7.000e+06	Business Products & Services	17	F
57	Lead5 Media	50.16	3.560e+07	Advertising & Marketing	33	N
58	[illegible]	50.15	9.600e+06	Advertising & Marketing	125	I

2) Now let's see how many distinct companies exists in our dataset:

```
kable(inc %>% dplyr::group_by(Industry) %>% dplyr::summarise(n =n()) %>% arrange(desc(n))) %>% kable_sty
```

Industry	n
IT Services	733
Business Products & Services	482
Advertising & Marketing	471
Health	355
Software	342
Financial Services	260
Manufacturing	256
Consumer Products & Services	203
Retail	203
Government Services	202
Human Resources	196
Construction	187
Logistics & Transportation	155
Food & Beverage	131
Telecommunications	129
Energy	109
Real Estate	96
Education	83
Engineering	74
Security	73
Travel & Hospitality	62
Media	54
Environmental Services	51
Insurance	50
Computer Hardware	44

3) Let us calculate the median revenue:

```
inc %>% dplyr::summarise(min=min(Revenue), median=median(Revenue), max=max(Revenue))
```

```
##     min   median      max
## 1 2e+06 10900000 1.01e+10
```

4) Let's calculate number of distinct cities:

```
cities <- inc %>% group_by(City) %>% summarise((n=n()))
nrow(cities)
```

```
## [1] 1519
```

Below are the top 10 companies and the cities they are located in:

```
kable(inc %>% group_by(City) %>% summarise(n=n()) %>% arrange(desc(n)) %>% top_n(10)) %>% kable_styling
```

```
## Selecting by n
```

City	n
New York	160
Chicago	90
Austin	88
Houston	76
San Francisco	75
Atlanta	74
San Diego	67
Seattle	52
Boston	43
Dallas	42
Denver	42

5) Now let's see the employee minimum and maximum range in the companies listed in the dataset.

```
kable(inc %>% dplyr:: summarise(min=min(Employees, na.rm = TRUE), median=median(Employees, na.rm = TRUE
```

min	median	max
1	53	66803

6) Number of states that are distinct in the given dataset:

```
distinct_states <- inc %>% group_by(State) %>% summarise(n=n())
nrow(distinct_states)
```

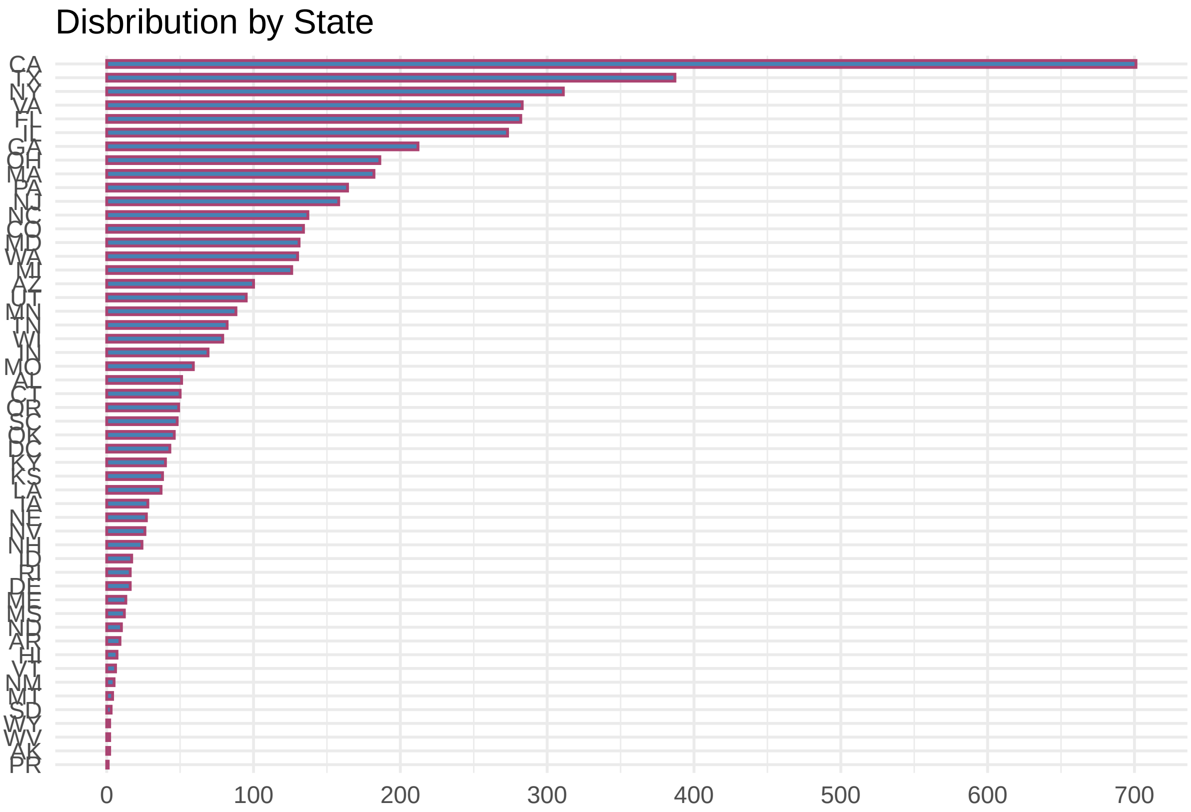
```
## [1] 52
```

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

## distribution of companies in the dataset by State
companies_byState <- inc %>% group_by(State) %>% summarise(n=n()) %>% arrange(desc(n))

plt1 <-
  ggplot(data = companies_byState[1:52,], aes(x=reorder(State,n), y=n)) +
  geom_bar(stat="identity", width=0.5, color="#AA4371", fill="steelblue",
           position=position_dodge()) +
    #geom_text(aes(label=round(n, digits=2)), hjust=1.3, size=3.0, color="white") +
    coord_flip() +
    scale_y_continuous(breaks=seq(0,700,100)) +
    ggtitle("Disbribution by State") +
    xlab("") + ylab("") +
    theme_minimal()
```

```
## Plotting the graph

plt1
```



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

## state with the 3rd most companies in the data set

kable(inc %>% group_by(State) %>% summarise(n=n()) %>% arrange(desc(n)) %>% top_n(3)) %>% kable_styling
```

```
## Selecting by n
```

State	n
CA	701
TX	387
NY	311

As shown above, the state with third most companies in the dataset is New York.

Let's evaluate the company cases:

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

Now, let's find the median number of employees in each industry for NY state.

```
kable(company_cases%>%filter(State=='NY')%>%group_by(Industry)%>%summarise(min=min(Employees),median=me
```

Industry	min	median	max	var
Business Products & Services	4	70.5	32000	3.894641e+07
Consumer Products & Services	5	25.0	10000	5.835802e+06
Travel & Hospitality	6	61.0	2280	6.974669e+05
Human Resources	7	56.0	2081	4.634787e+05
IT Services	8	54.0	3000	2.241769e+05
Software	15	80.0	1271	1.404907e+05
Security	25	32.5	450	4.415000e+04
Media	4	45.0	602	3.099560e+04
Financial Services	14	81.0	483	2.299190e+04
Environmental Services	60	155.0	250	1.805000e+04
Food & Beverage	5	41.0	383	1.390028e+04
Energy	5	120.0	294	1.106670e+04
Telecommunications	6	31.0	316	1.064462e+04
Manufacturing	11	30.0	307	8.048231e+03
Health	2	45.0	298	7.505141e+03
Construction	10	24.5	219	6.392000e+03
Advertising & Marketing	2	38.0	270	3.872536e+03
Education	19	50.5	200	2.359516e+03
Engineering	11	54.5	94	1.583000e+03
Logistics & Transportation	1	23.5	70	8.430000e+02
Retail	3	13.5	75	6.378736e+02
Insurance	15	32.5	50	6.125000e+02
Real Estate	7	18.0	30	9.425000e+01
Computer Hardware	44	44.0	44	NA
Government Services	17	17.0	17	NA

The data above shows the min, median, and max number of employees for each industry in NY. It is ordered from highest to lowest variability.

In order to show the median number of employees, a box plot could be plotted.The plot will also display range of data and outliers. Number of distinct industries are 25. Let's use the table above to show that companies that have higher variability in employee number is ones with higher maximum number of employees.

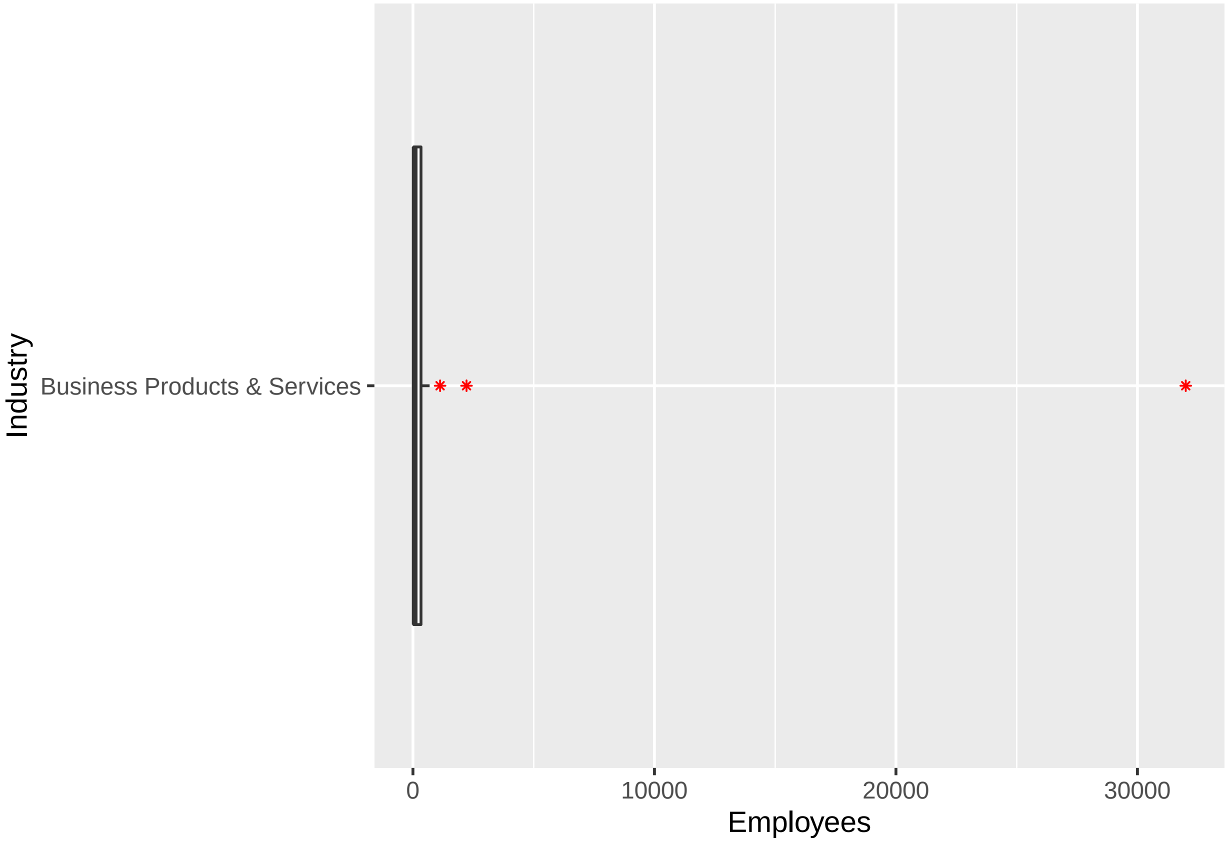
Below, the industries are grouped together.

```
Business_Products_Services <- c('Business Products & Services')
Consumer_Products_Services <- c('Consumer Products & Services')
group_2 <- c('Travel & Hospitality', 'Human Resources', 'IT Services', 'Software')
group_3 <- c('Security', 'Media', 'Financial Services',  'Environmental Services', 'Food & Beverage')
group_4 <- c('Energy', 'Telecommunications', 'Manufacturing', 'Health', 'Construction')
group_5 <- c('Advertising & Marketing', 'Education', 'Engineering', 'Logistics & Transportation', 'Retai
group_6 <- c('Insurance', 'Real Estate', 'Computer Hardware', 'Government Services')
```

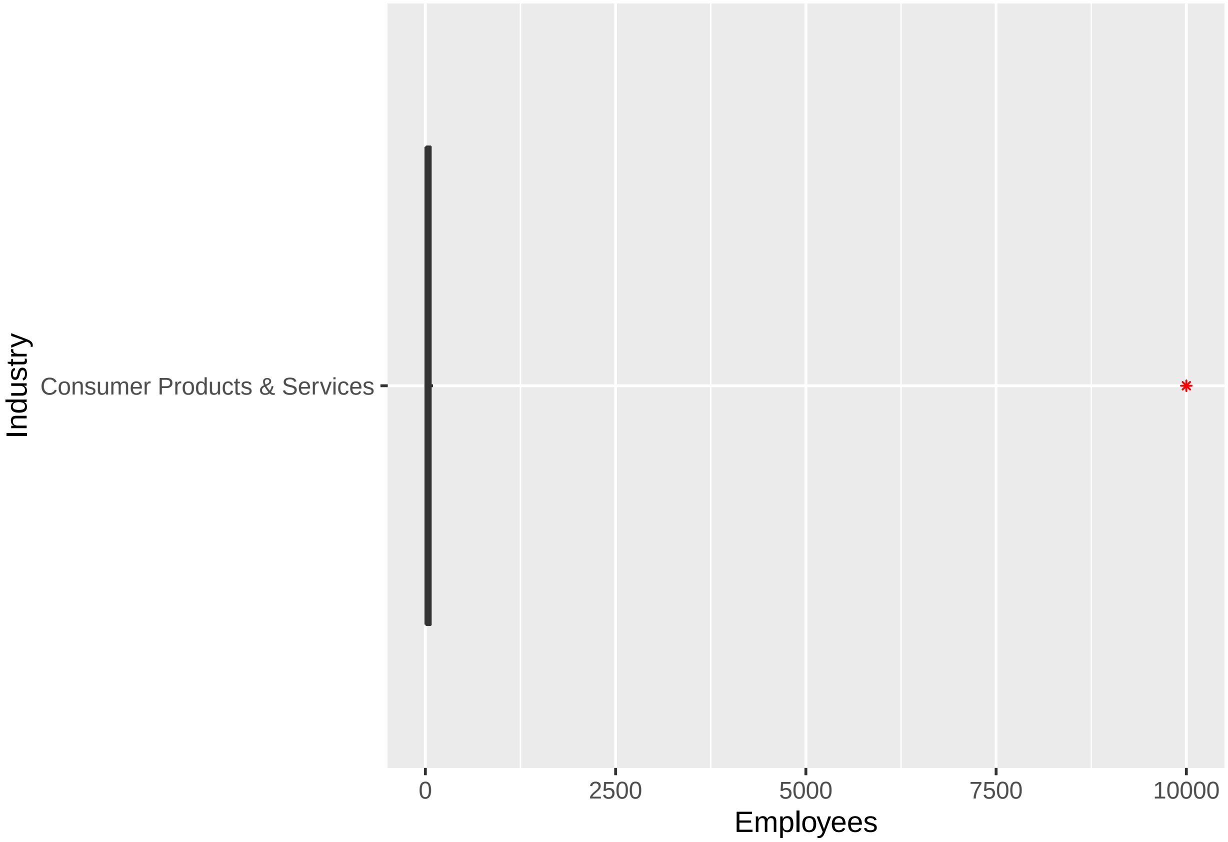
Creating box plots for the respective groups:

```
plt_Business_Products_Services <- ggplot(company_cases %>% filter(State=='NY' & Industry %in% Business_
        coord_flip() +
        geom_boxplot(outlier.colour="red", outlier.shape=8,
             outlier.size=1, notch=FALSE)

plt_Consumer_Products_Services <- ggplot(company_cases %>% filter(State=='NY' & Industry %in% Consumer_
        coord_flip() +
        geom_boxplot(outlier.colour="red", outlier.shape=8,
             outlier.size=1, notch=FALSE)

plt_group_2 <- ggplot(company_cases %>% filter(State=='NY' & Industry %in% group_2), aes(x = Industry, 
        coord_flip() +
        geom_boxplot(outlier.colour="red", outlier.shape=8,
             outlier.size=1, notch=FALSE)

plt_group_3 <- ggplot(company_cases %>% filter(State=='NY' & Industry %in% group_3), aes(x = Industry, 
        coord_flip() +
        geom_boxplot(outlier.colour="red", outlier.shape=8,
             outlier.size=1, notch=FALSE)

plt_group_4 <- ggplot(company_cases %>% filter(State=='NY' & Industry %in% group_4), aes(x = Industry, 
        coord_flip() +
        geom_boxplot(outlier.colour="red", outlier.shape=8,
             outlier.size=1, notch=FALSE)

plt_group_5 <- ggplot(company_cases %>% filter(State=='NY' & Industry %in% group_5), aes(x = Industry, 
        coord_flip() +
        geom_boxplot(outlier.colour="red", outlier.shape=8,
             outlier.size=1, notch=FALSE)

plt_group_6 <- ggplot(company_cases %>% filter(State=='NY' & Industry %in% group_6), aes(x = Industry, 
        coord_flip() +
        geom_boxplot(outlier.colour="red", outlier.shape=8,
             outlier.size=1, notch=FALSE)
```

```
plt_Business_Products_Services
```



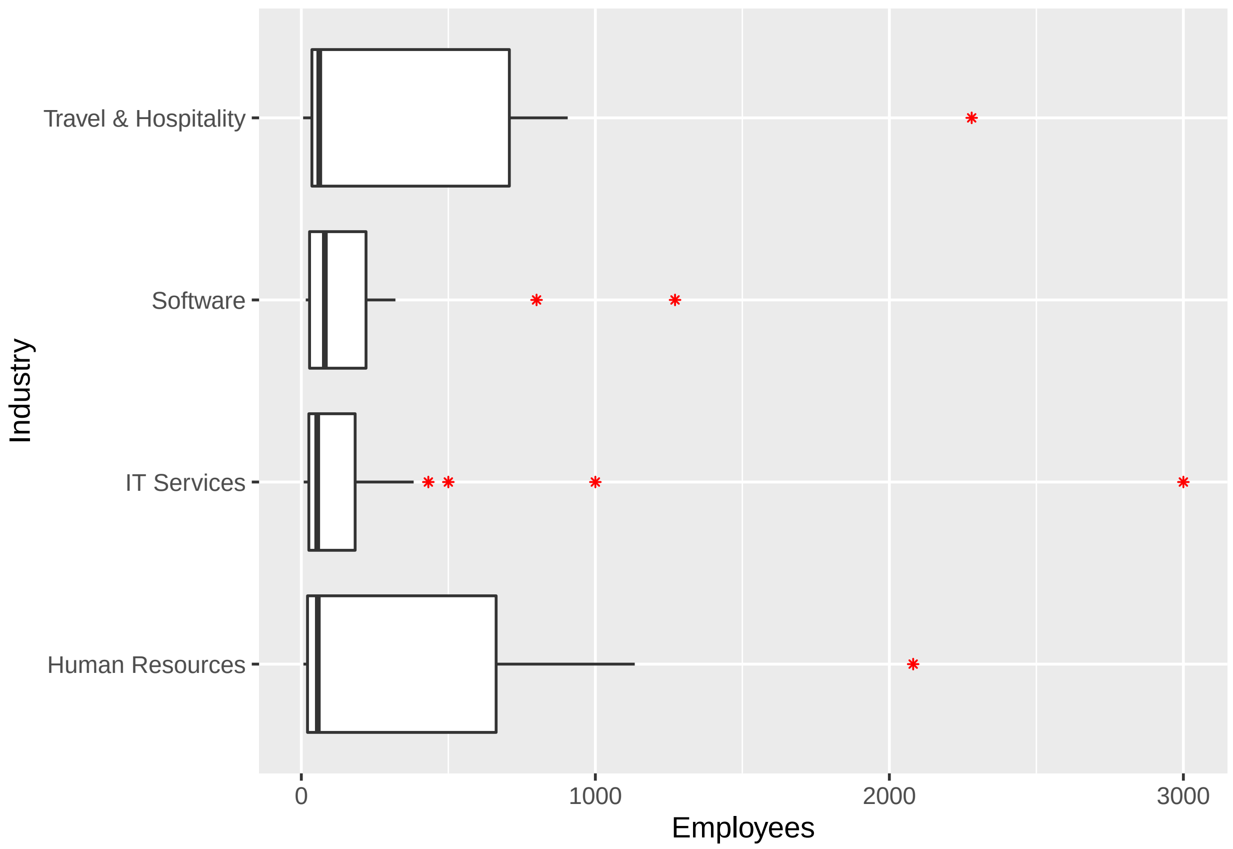
```
plt_Consumer_Products_Services
```



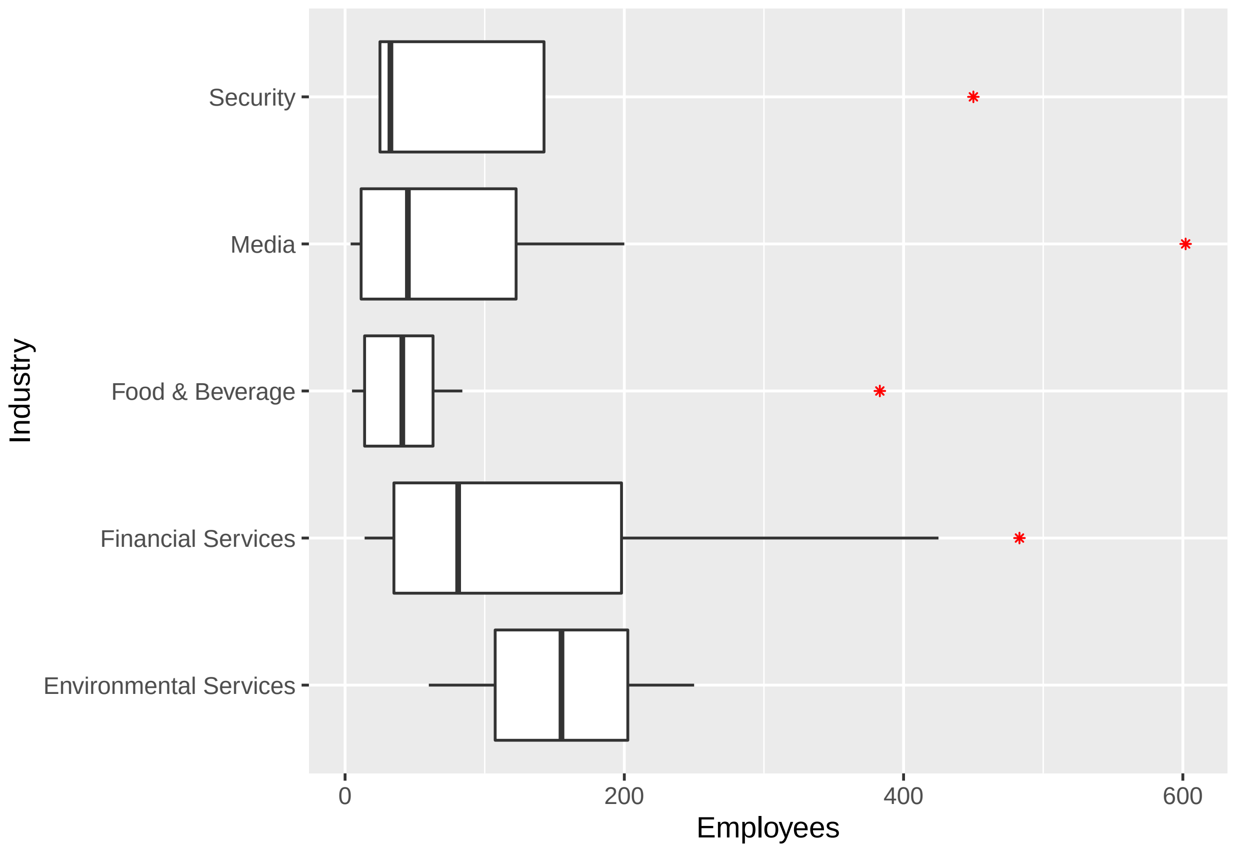
Box plots for remaining industries: x-axis scale for each group is different.

The box plots for 'Business Products & Services' and 'Consumer Products & Services' came out as very small. The outlier data is causing the box plot of these 2 industries to flatten out.

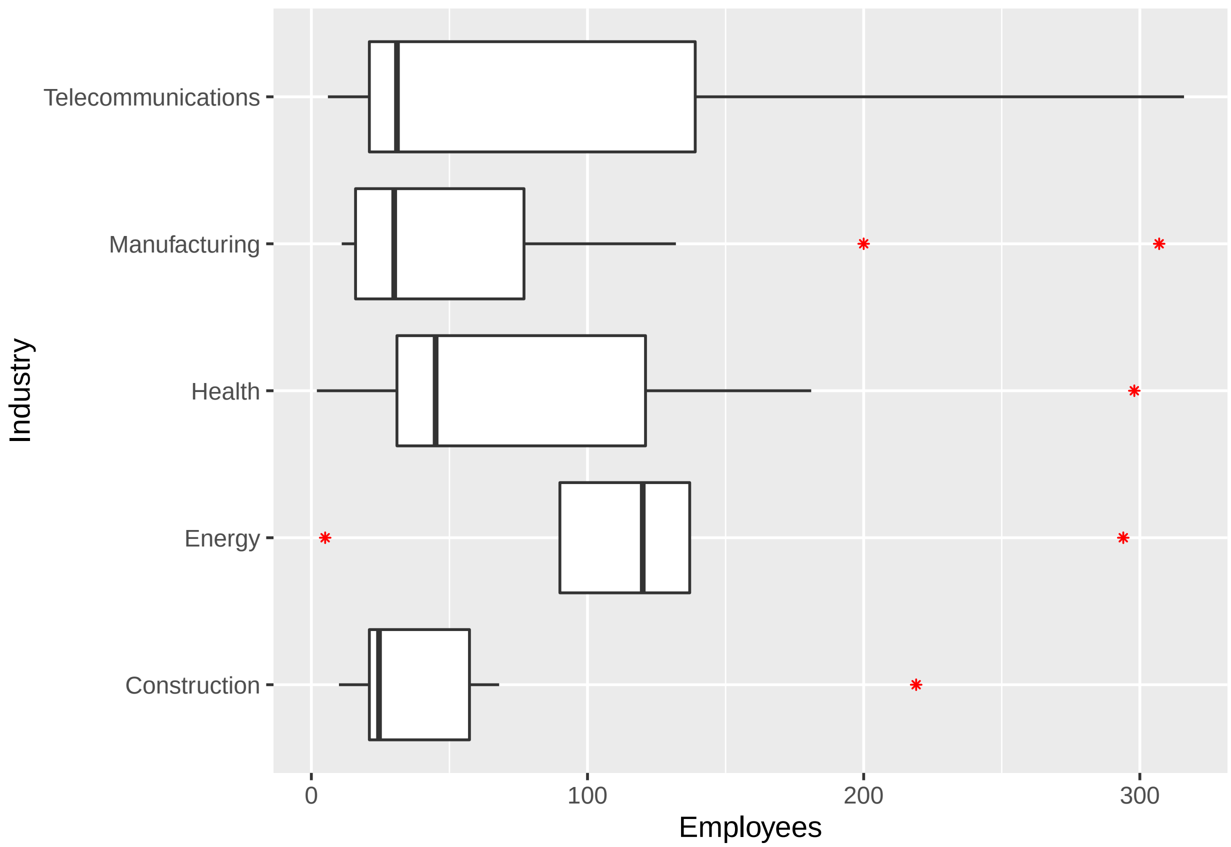
```
plt_group_2
```


Travel & Hospitality
Software
IT Services
Human Resources
Industry
0
1000
2000
3000
Employees

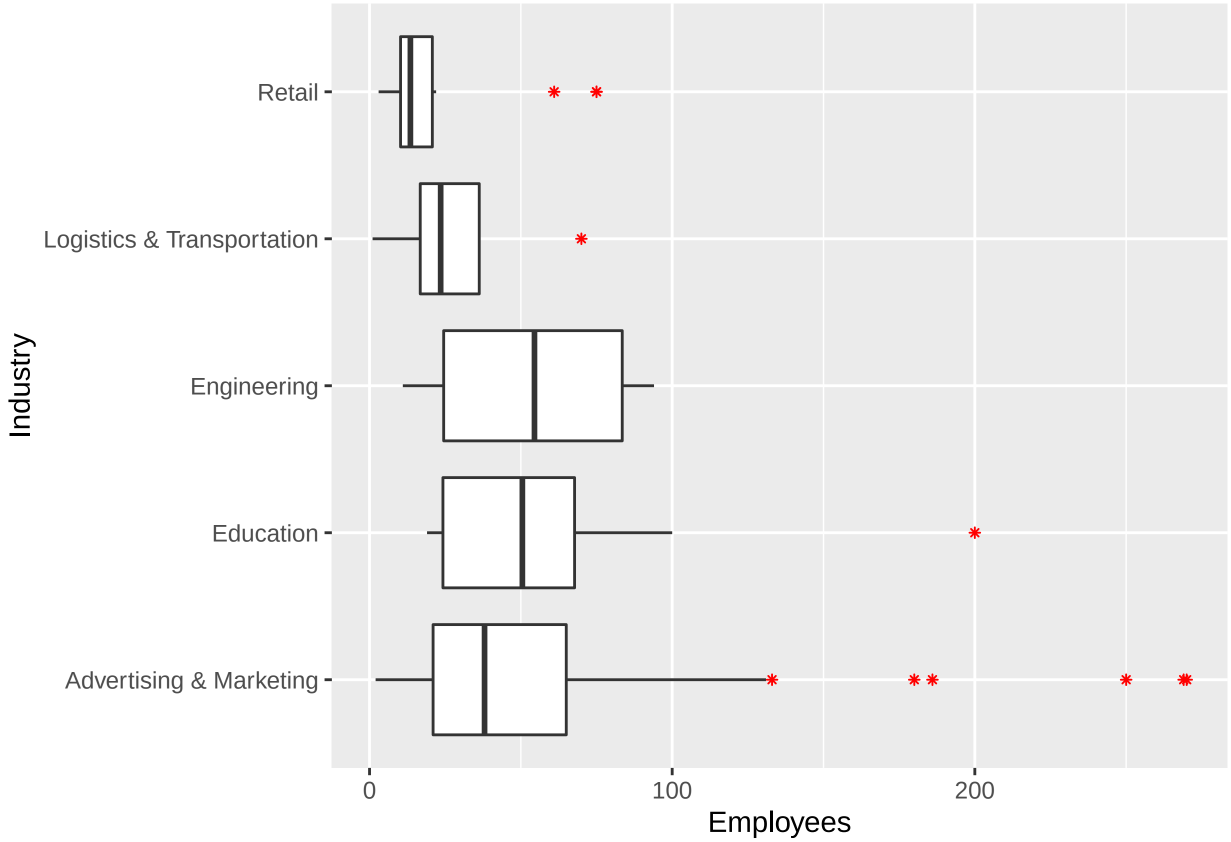
```
plt_group_3
```


Security
Media
Food & Beverage
Financial Services
Environmental Services
Industry
0
200
400
600
Employees

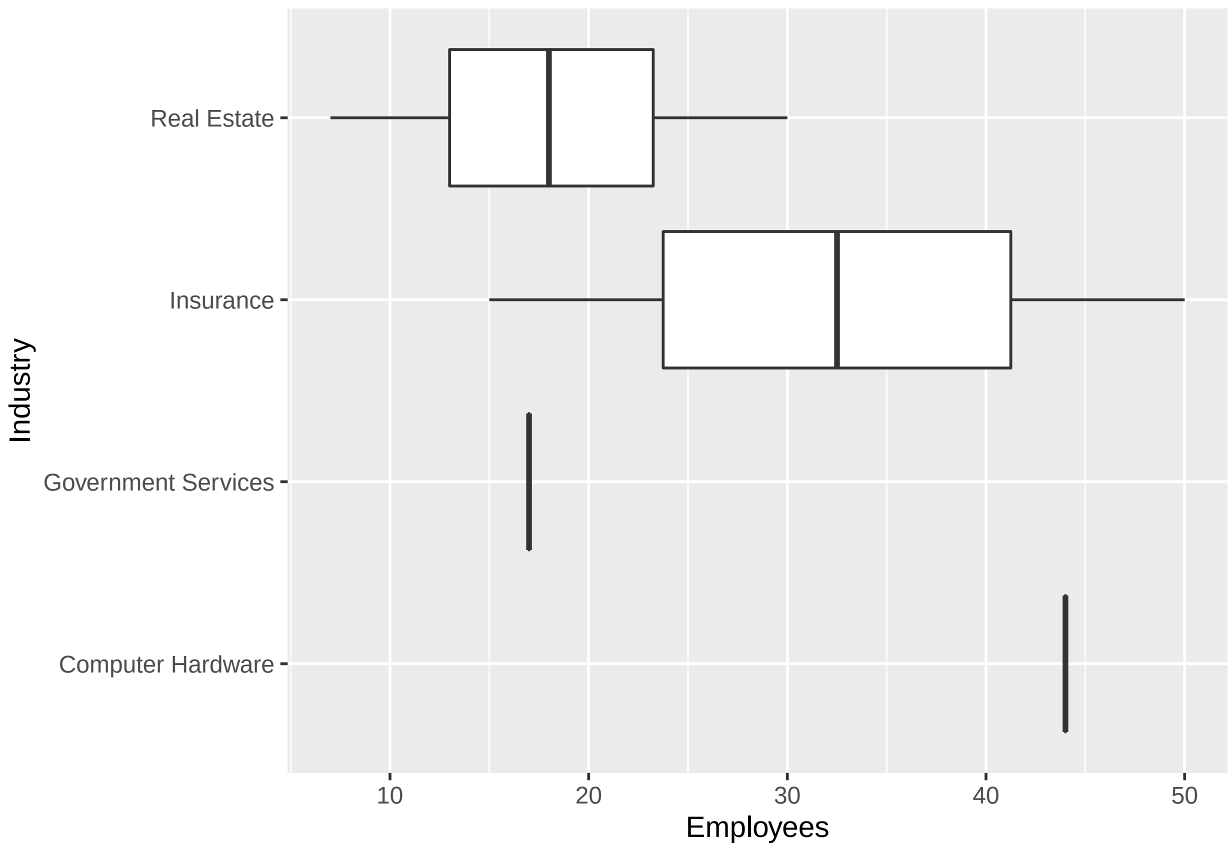
```
plt_group_4
```


Telecommunications
Manufacturing
Health
Energy
Construction
Industry
0
100
200
300
Employees

```
plt_group_5
```


Retail
Logistics & Transportation
Engineering
Education
Advertising & Marketing
Industry
0
100
200
Employees

```
plt_group_6
```



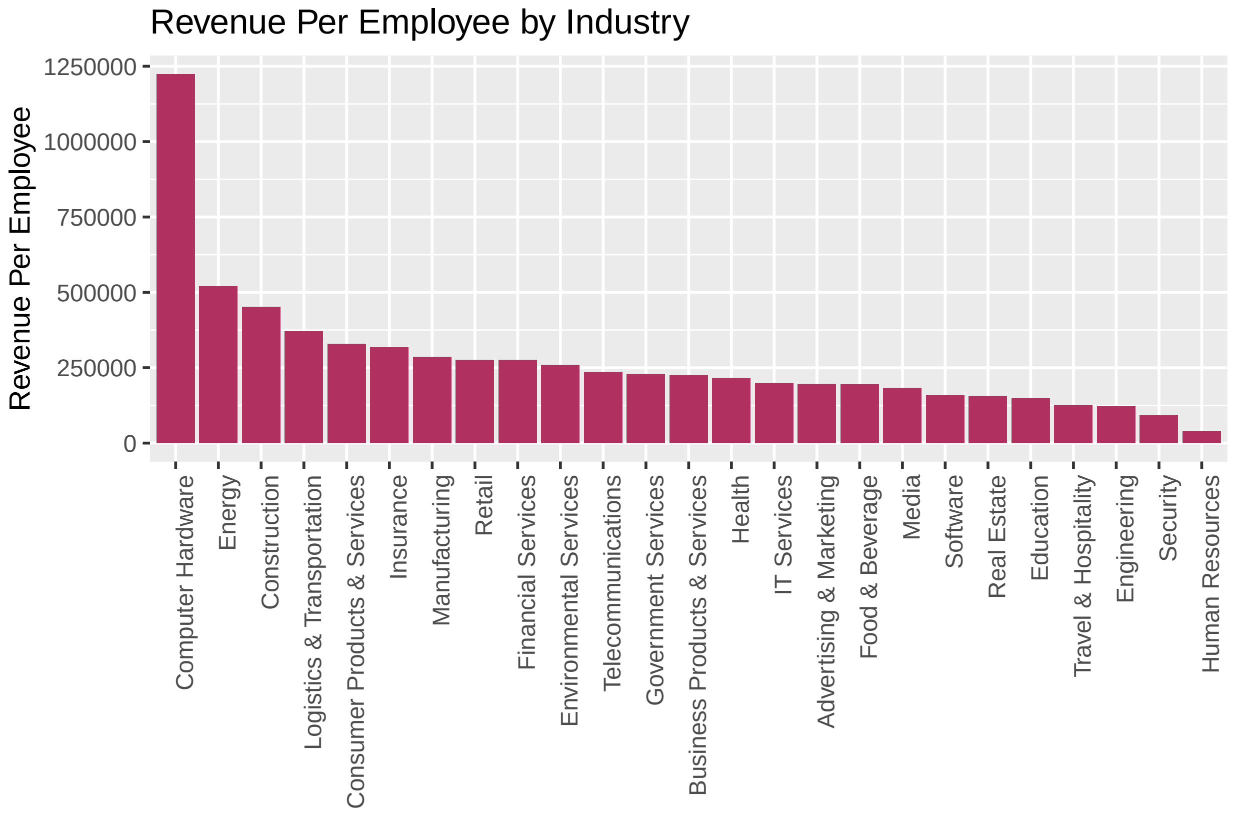
Question 3

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.

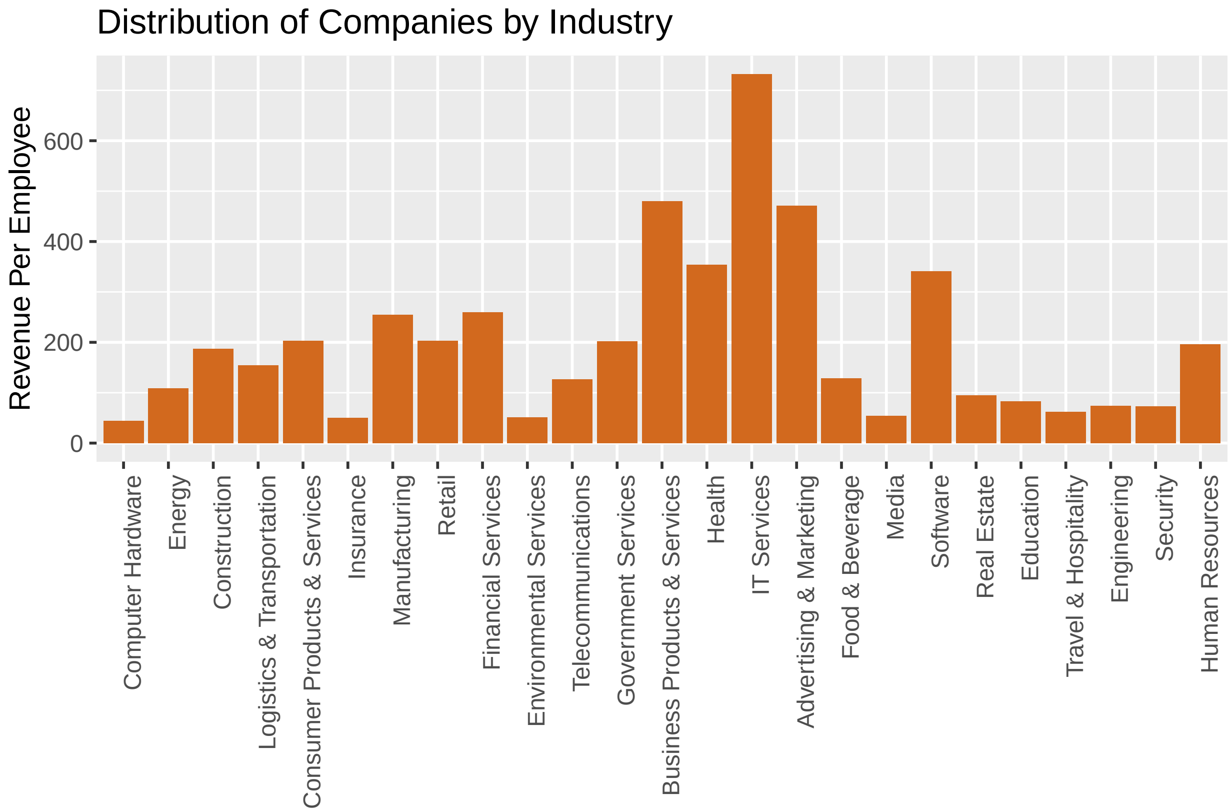
```
# Answer Question 3 here

revenue_perEmployee <-
company_cases %>% group_by(Industry) %>% summarise(count=n(), total_revenue=sum(Revenue), total_employe

kable(revenue_perEmployee) %>% kable_styling()
```

Industry	count	total_revenue	total_employees	revenue_perEmployee
Computer Hardware	44	11885700000	9714	1223563.93
Energy	109	13771600000	26437	520921.44
Construction	187	13174300000	29099	452740.64
Logistics & Transportation	154	14837800000	39994	371000.65
Consumer Products & Services	203	14956400000	45464	328972.37
Insurance	50	2337900000	7339	318558.39
Manufacturing	255	12603600000	43942	286823.54
Retail	203	10257400000	37068	276718.46
Financial Services	260	13150900000	47693	275740.67
Environmental Services	51	2638800000	10155	259852.29
Telecommunications	127	7287900000	30842	236297.91
Government Services	202	6009100000	26185	229486.35
Business Products & Services	480	26345900000	117357	224493.64
Health	354	17860100000	82430	216669.90
IT Services	732	20525000000	102788	199682.84
Advertising & Marketing	471	7785000000	39731	195942.71
Food & Beverage	129	12812500000	65911	194390.92
Media	54	1742400000	9532	182794.80
Software	341	8134600000	51262	158686.75
Real Estate	95	2956800000	18893	156502.41
Education	83	1139300000	7685	148249.84
Travel & Hospitality	62	2931600000	23035	127267.20
Engineering	74	2532500000	20435	123929.53
Security	73	3812800000	41059	92861.49
Human Resources	196	9246100000	226980	40735.31

```
plt_revenue_perEmployee_a<- ggplot(data=revenue_perEmployee, aes(x=reorder(Industry,-revenue_perEmployee
     geom_bar(stat="identity", fill="maroon") +
     theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
     ggtitle("Revenue Per Employee by Industry") +
     ylab("Revenue Per Employee") +
     xlab("")

plt_revenue_perEmployee_b <- ggplot(data=revenue_perEmployee, aes(x=reorder(Industry,-revenue_perEmploye
     geom_bar(stat="identity", fill="chocolate") +
     theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
     ggtitle("Distribution of Companies by Industry") +
     ylab("Revenue Per Employee") +
     xlab("")

plt_revenue_perEmployee_a
```



```
plt_revenue_perEmployee_b
```



The code above plots the revenue per employee as a bar chart sorted by revenue per employee from highest to lowest.

A second bar chart plot is generated that shows the distribution of companies by industry sorted by revenue per employee from highest to lowest, it uses same order as the first plot.