DATA-INK (Visualization of fastest-growing companies in the US)

Subhalaxmi Rout

2021-02-14

```
# Load libraries
library(ggplot2)
library(stats)
library(DT)
library(dplyr)
library(psych)
library(visdat)
```

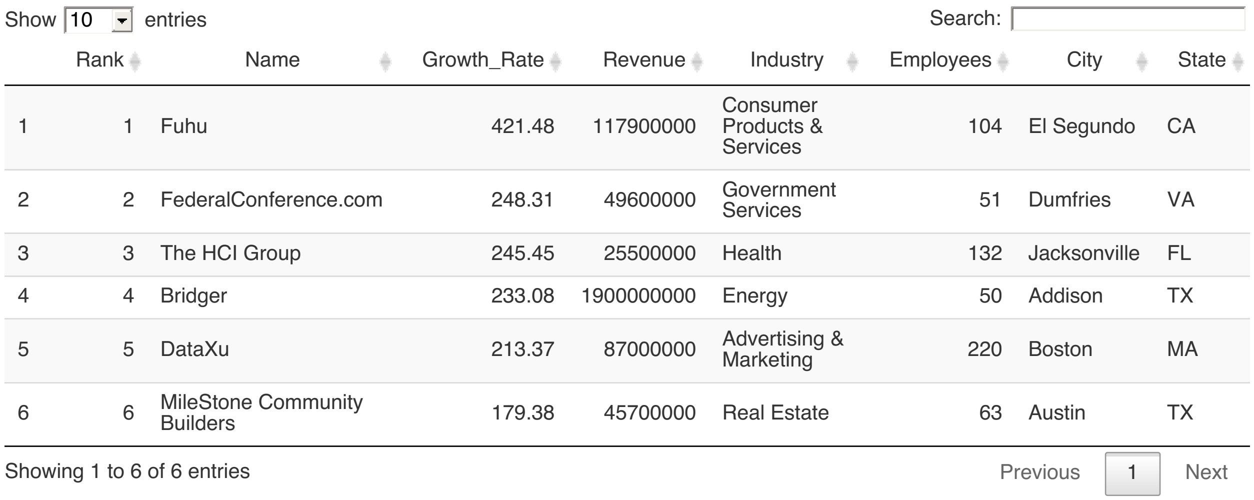
Principles of Data Visualization and Introduction to ggplot2

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:

```
DT::datatable(head(inc))
```



Show 10 entries Search:

	Rank	Name	Growth_Rate	Revenue	Industry	Employees	City	State
1	1	Fuhu	421.48	117900000	Consumer Products & Services	104	El Segundo	CA
2	2	FederalConference.com	248.31	49600000	Government Services	51	Dumfries	VA
3	3	The HCI Group	245.45	25500000	Health	132	Jacksonville	FL
4	4	Bridger	233.08	1900000000	Energy	50	Addison	TX
5	5	DataXu	213.37	87000000	Advertising & Marketing	220	Boston	MA
6	6	MileStone Community Builders	179.38	45700000	Real Estate	63	Austin	TX

Showing 1 to 6 of 6 entries

```
summary(inc)
```

```
##       Rank          Name            Growth_Rate        Revenue
##  Min.   :   1   Length:5001        Min.   :  0.340   Min.   :2.000e+06
##  1st Qu.:1252   Class :character   1st Qu.:  0.770   1st Qu.:5.100e+06
##  Median :2502   Mode  :character   Median :  1.420   Median :1.090e+07
##  Mean   :2502                      Mean   :  4.612   Mean   :4.822e+07
##  3rd Qu.:3751                      3rd Qu.:  3.290   3rd Qu.:2.860e+07
##  Max.   :5000                      Max.   :421.480   Max.   :1.010e+10
##
##    Industry           Employees          City               State
##  Length:5001        Min.   :    1.0   Length:5001        Length:5001
##  Class :character   1st Qu.:   25.0   Class :character   Class :character
##  Mode  :character   Median :   53.0   Mode  :character   Mode  :character
##                     Mean   :  232.7
##                     3rd Qu.:  132.0
##                     Max.   :66803.0
##                     NA's   :12
```

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:

Lets have a look on datatypes and structure od data.

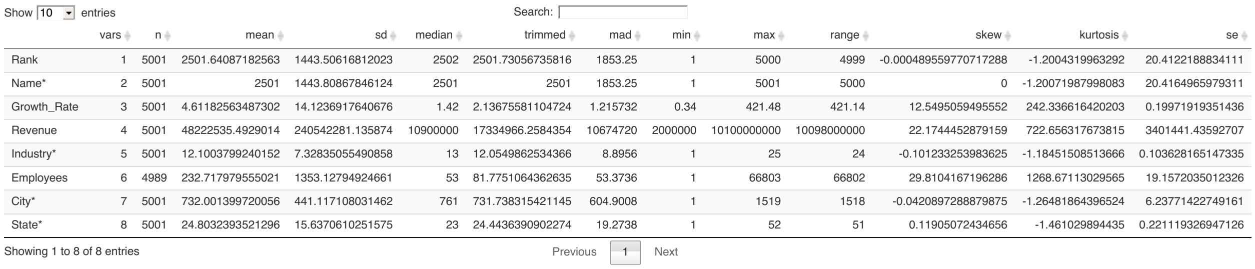
```
# Insert your code here, create more chunks as necessary
glimpse(inc)
```

```
## Rows: 5,001
## Columns: 8
## $ Rank        <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, ...
## $ Name        <chr> "Fuhu", "FederalConference.com", "The HCI Group", "Brid...
## $ Growth_Rate <dbl> 421.48, 248.31, 245.45, 233.08, 213.37, 179.38, 174.04,...
## $ Revenue     <dbl> 1.179e+08, 4.960e+07, 2.550e+07, 1.900e+09, 8.700e+07, ...
## $ Industry    <chr> "Consumer Products & Services", "Government Services", ...
## $ Employees   <int> 104, 51, 132, 50, 220, 63, 27, 75, 97, 15, 149, 165, 25...
## $ City        <chr> "El Segundo", "Dumfries", "Jacksonville", "Addison", "B...
## $ State       <chr> "CA", "VA", "FL", "TX", "MA", "TX", "TN", "CA", "UT", "...
```

State, City, Industry, and Name are in character type. Growth_Rate and Employees have a double type, and Rank has integer type.

Describe() shows summary statistics of data.

```
DT::datatable(describe(inc))
```



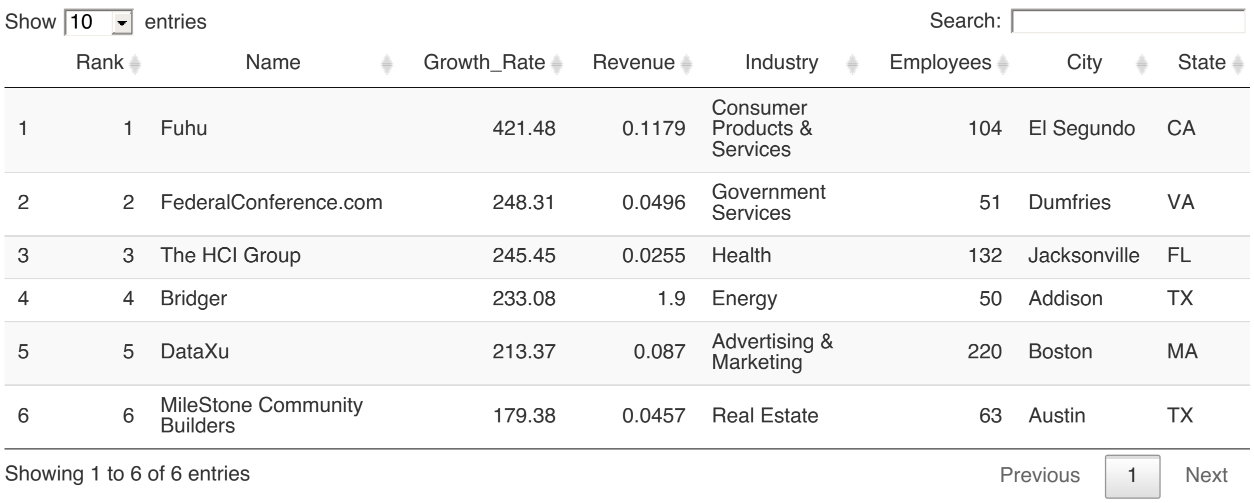
Show 10 entries Search:

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Rank	1	5001	2501.64087182563	1443.50616812023	2502	2501.73056735816	1853.25	1	5000	4999	-0.000489559770717288	-1.2004319963292	20.4122188834111
Name*	2	5001	2501	1443.80867846124	2501	2501	1853.25	1	5001	5000	0	-1.20071987998083	20.4164965979311
Growth_Rate	3	5001	4.61182563487302	14.1236917640676	1.42	2.13675581104724	1.215732	0.34	421.48	421.14	12.5495059495552	242.336616420203	0.19971919351436
Revenue	4	5001	48222535.4929014	240542281.135874	10900000	17334966.2584354	10674720	2000000	10100000000	10098000000	22.1744452879159	722.656317673815	3401441.43592707
Industry*	5	5001	12.1003799240152	7.32835055490858	13	12.0549862534366	8.8956	1	25	24	-0.101233253983625	-1.18451508513666	0.103628165147335
Employees	6	4989	232.717979555021	1353.12794924661	53	81.7751064362635	53.3736	1	66803	66802	29.8104167196286	1268.67113029565	19.1572035012326
City*	7	5001	732.001399720056	441.117108031462	761	731.738315421145	604.9008	1	1519	1518	-0.0420897288879875	-1.26481864396524	6.23771422749161
State*	8	5001	24.8032393521296	15.6370610251575	23	24.4436390902274	19.2738	1	52	51	0.11905072434656	-1.461029894435	0.221119326947126

Showing 1 to 8 of 8 entries

Revenue looks very big nubmer, devide the revenue by 10^9, the result will be in billion.

```r
inc$Revenue <- sapply(inc$Revenue, function(x) x / 1000000000)
DT::datatable(head(inc))
```



Show 10 entries

Search:

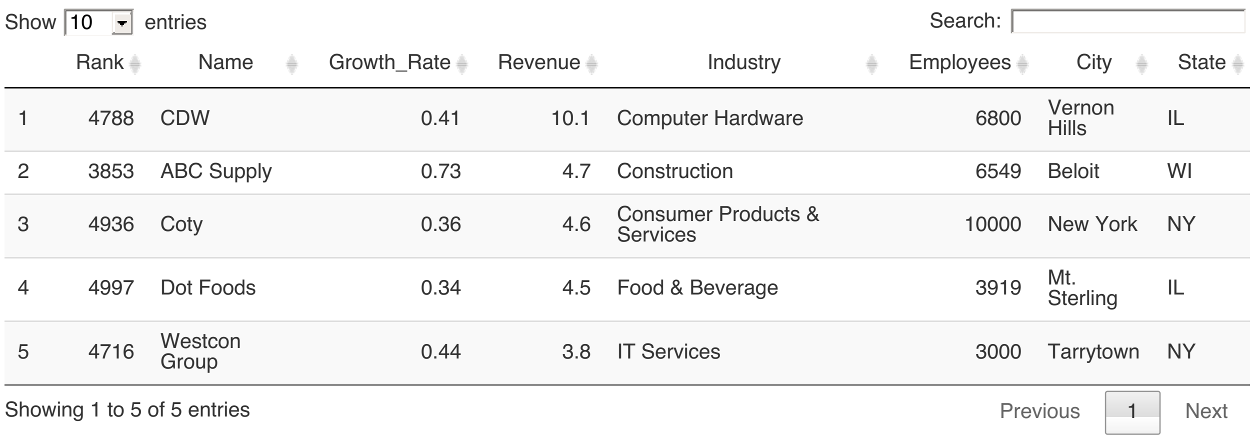
	Rank	Name	Growth_Rate	Revenue	Industry	Employees	City	State
1	1	Fuhu	421.48	0.1179	Consumer Products & Services	104	El Segundo	CA
2	2	FederalConference.com	248.31	0.0496	Government Services	51	Dumfries	VA
3	3	The HCI Group	245.45	0.0255	Health	132	Jacksonville	FL
4	4	Bridger	233.08	1.9	Energy	50	Addison	TX
5	5	DataXu	213.37	0.087	Advertising & Marketing	220	Boston	MA
6	6	MileStone Community Builders	179.38	0.0457	Real Estate	63	Austin	TX

Showing 1 to 6 of 6 entries

Previous 1 Next

Top 5 and bottom 5 revenue generated company

```r
DT::datatable(inc %>% arrange(desc(Revenue)) %>% head(5))
```



Show 10 entries

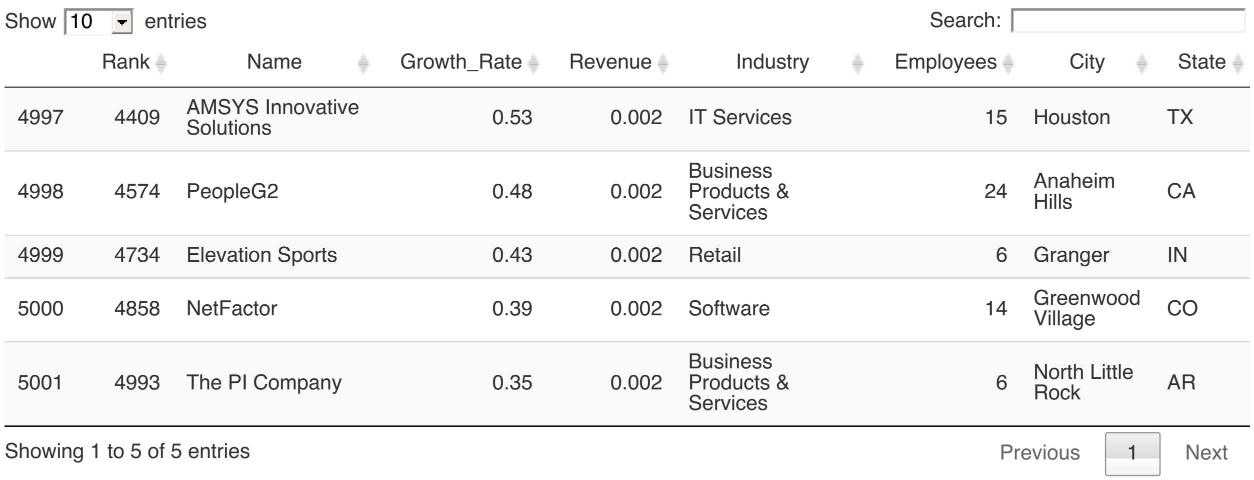
Search:

	Rank	Name	Growth_Rate	Revenue	Industry	Employees	City	State
1	4788	CDW	0.41	10.1	Computer Hardware	6800	Vernon Hills	IL
2	3853	ABC Supply	0.73	4.7	Construction	6549	Beloit	WI
3	4936	Coty	0.36	4.6	Consumer Products & Services	10000	New York	NY
4	4997	Dot Foods	0.34	4.5	Food & Beverage	3919	Mt. Sterling	IL
5	4716	Westcon Group	0.44	3.8	IT Services	3000	Tarrytown	NY

Showing 1 to 5 of 5 entries

Previous 1 Next

```r
DT::datatable(inc %>% arrange(desc(Revenue)) %>% tail(5))
```



Show 10 entries

Search:

	Rank	Name	Growth_Rate	Revenue	Industry	Employees	City	State
4997	4409	AMSYS Innovative Solutions	0.53	0.002	IT Services	15	Houston	TX
4998	4574	PeopleG2	0.48	0.002	Business Products & Services	24	Anaheim Hills	CA
4999	4734	Elevation Sports	0.43	0.002	Retail	6	Granger	IN
5000	4858	NetFactor	0.39	0.002	Software	14	Greenwood Village	CO
5001	4993	The PI Company	0.35	0.002	Business Products & Services	6	North Little Rock	AR

Showing 1 to 5 of 5 entries

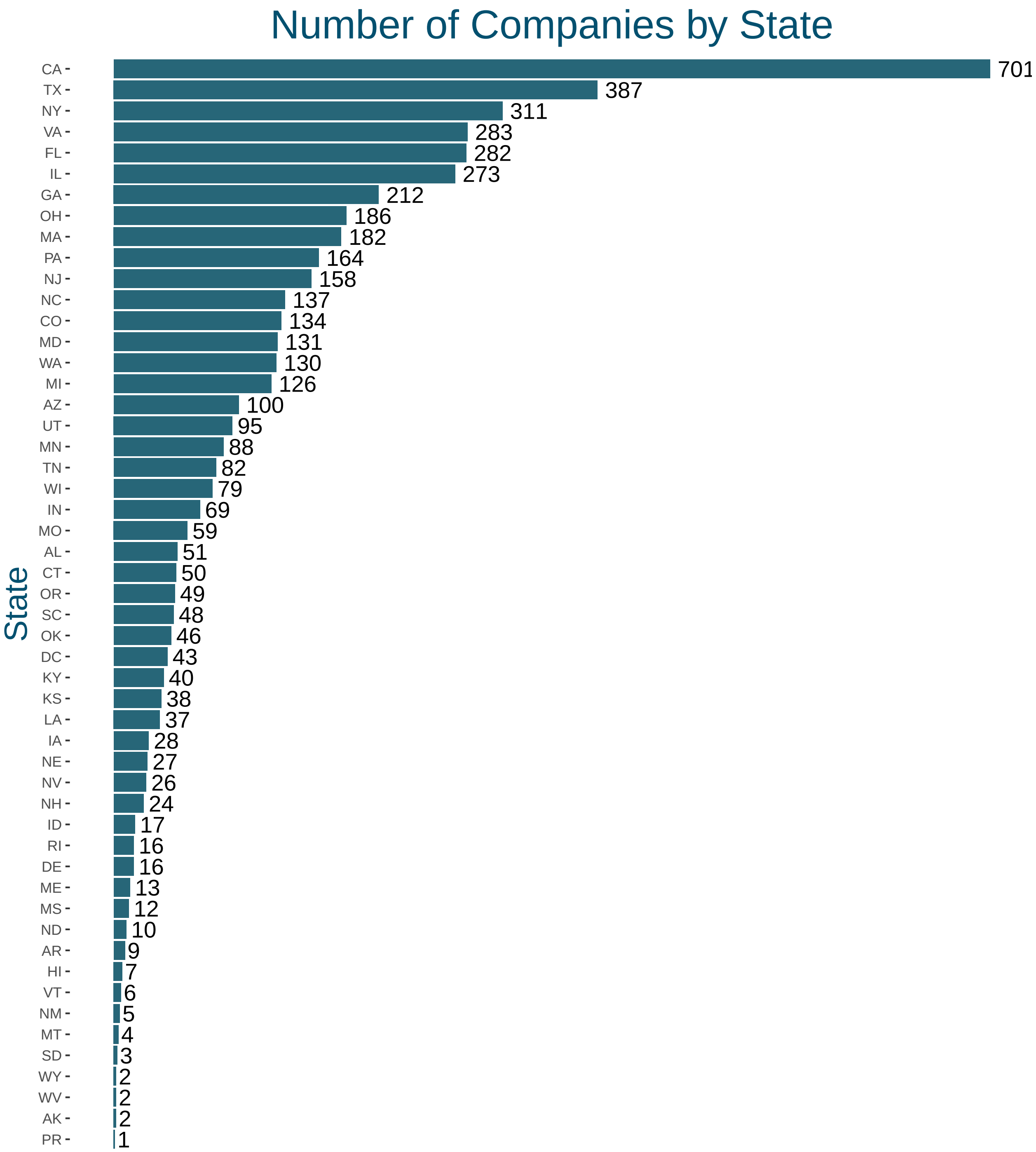
Previous 1 Next

Above table shows, Computer Hardware industry generate the high revenue and Business Products & Services generates the lowest revenue.

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.

```r
# Answer Question 1 here
inc %>% group_by(State) %>% count() %>%
  ggplot() + aes(x = reorder(State, n) , y = n, fill = n) +
  ggtitle('Number of Companies by State') +
  xlab('State') +
  geom_bar(fill="#276678", stat = "identity") +
  coord_flip() + geom_text(aes(label = n), size = 5,  hjust=-0.20) +
  theme(panel.background = element_rect(fill = "white", color = NA),
        plot.title = element_text(hjust = 0.5, size = 25, colour = "#03506f"),
        axis.title.y = element_text(size = 20, colour = "#03506f"),
        axis.title.x=element_blank(),
        axis.text.x = element_blank(),
        axis.ticks.x=element_blank()
        )
```



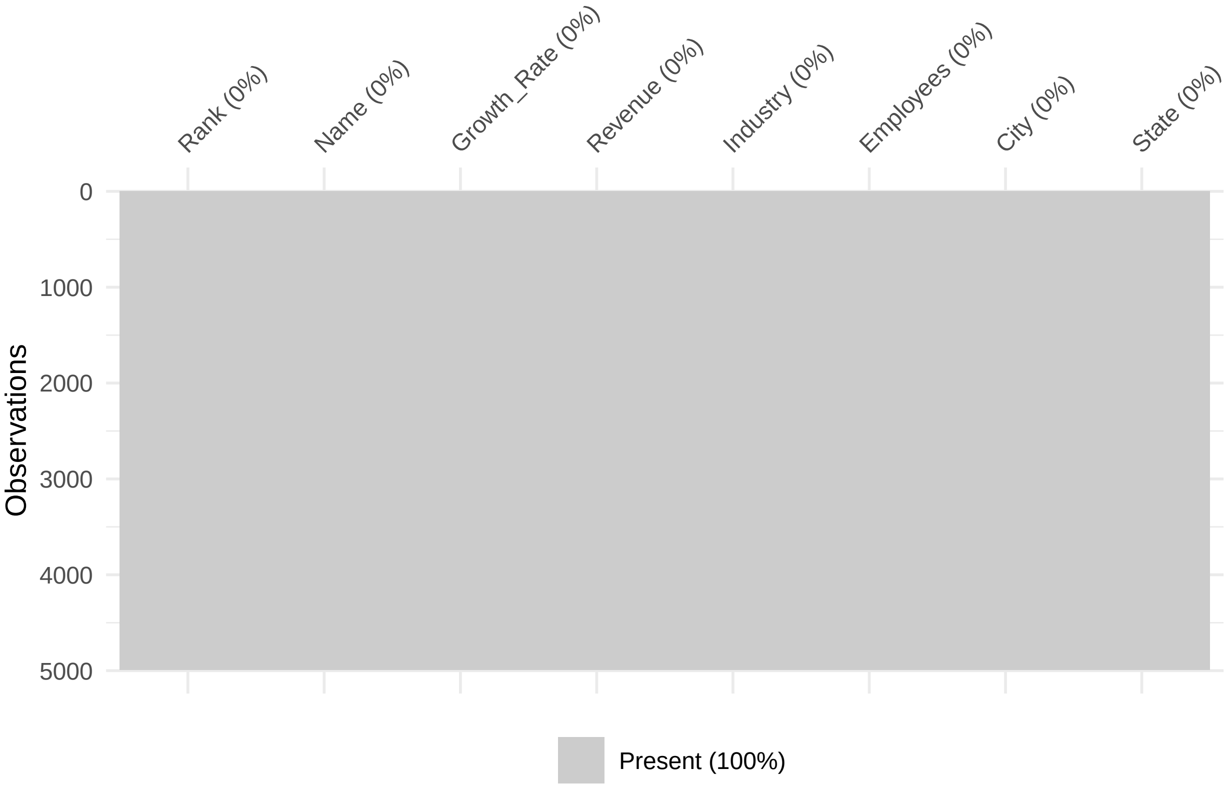
Above viz shows California and Texas has more companies than other states.

Quesiton 2

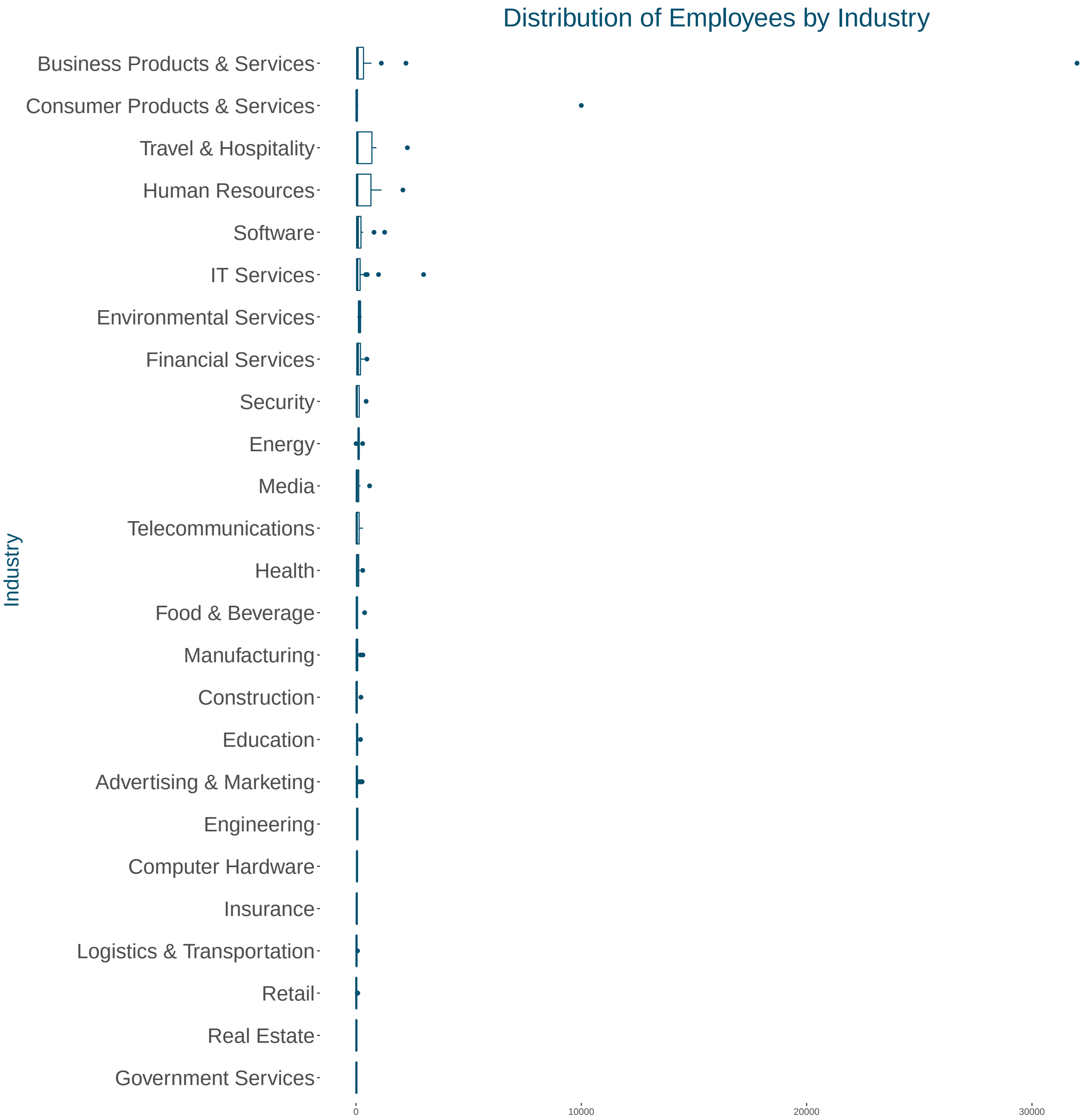
Lets dig in on the state with the 3rd most companies in the data set. Imagine you work for the state and are interested in how many people are employed by companies in different industries. Create a plot that shows the average and/or median employment by industry for companies in this state (only use 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.

From summary() we got to know there are 12 NAs present in Employees column. Use `complete.cases()` to get the data with out NAs. Below graph shows after apply `complete.cases()` no missing values present in data.

```
# Answer Question 2 here
inc <- inc[complete.cases(inc),]
visdat::vis_miss(inc)
```

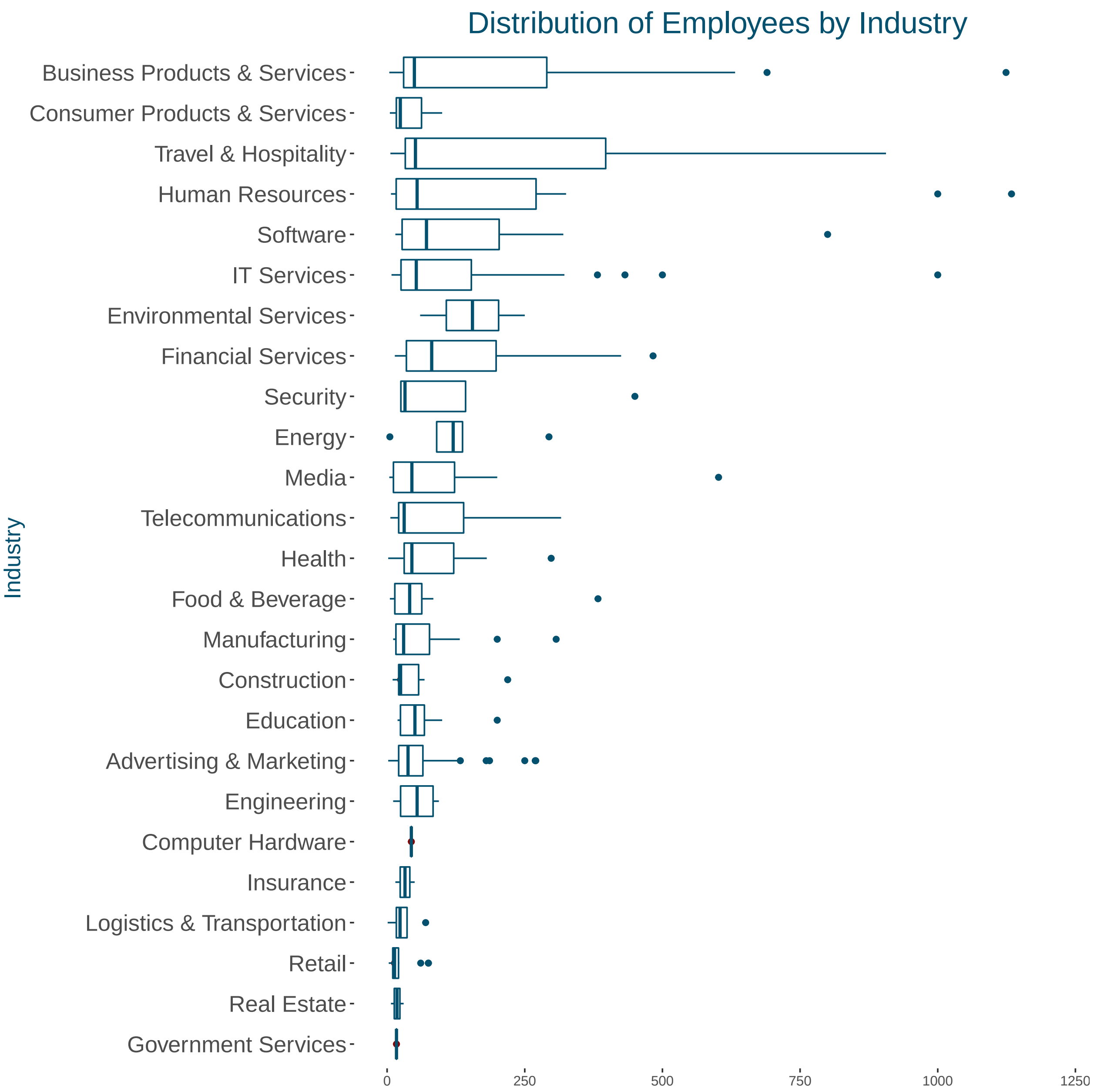


```
data_NY <- inc %>% filter(State == "NY")

data_NY %>% ggplot() +
  aes(x = reorder(Industry, Employees), y = Employees) +
  geom_boxplot(color = '#03506f') +
  ggtitle('Distribution of Employees by Industry') +
  xlab('Industry') +
  coord_flip() +
  theme(panel.background = element_rect(fill = "white", color = NA),
        plot.title = element_text(hjust = 0.5, size = 25, colour = "#03506f"),
        axis.title.y = element_text(size = 20, colour = "#03506f"),
        axis.title.x=element_blank(),
        axis.text.y = element_text(size = 20)
        )
```



Some industies have high outliers due to this visualisation is not clear. Lets exclued outliers above 1200.

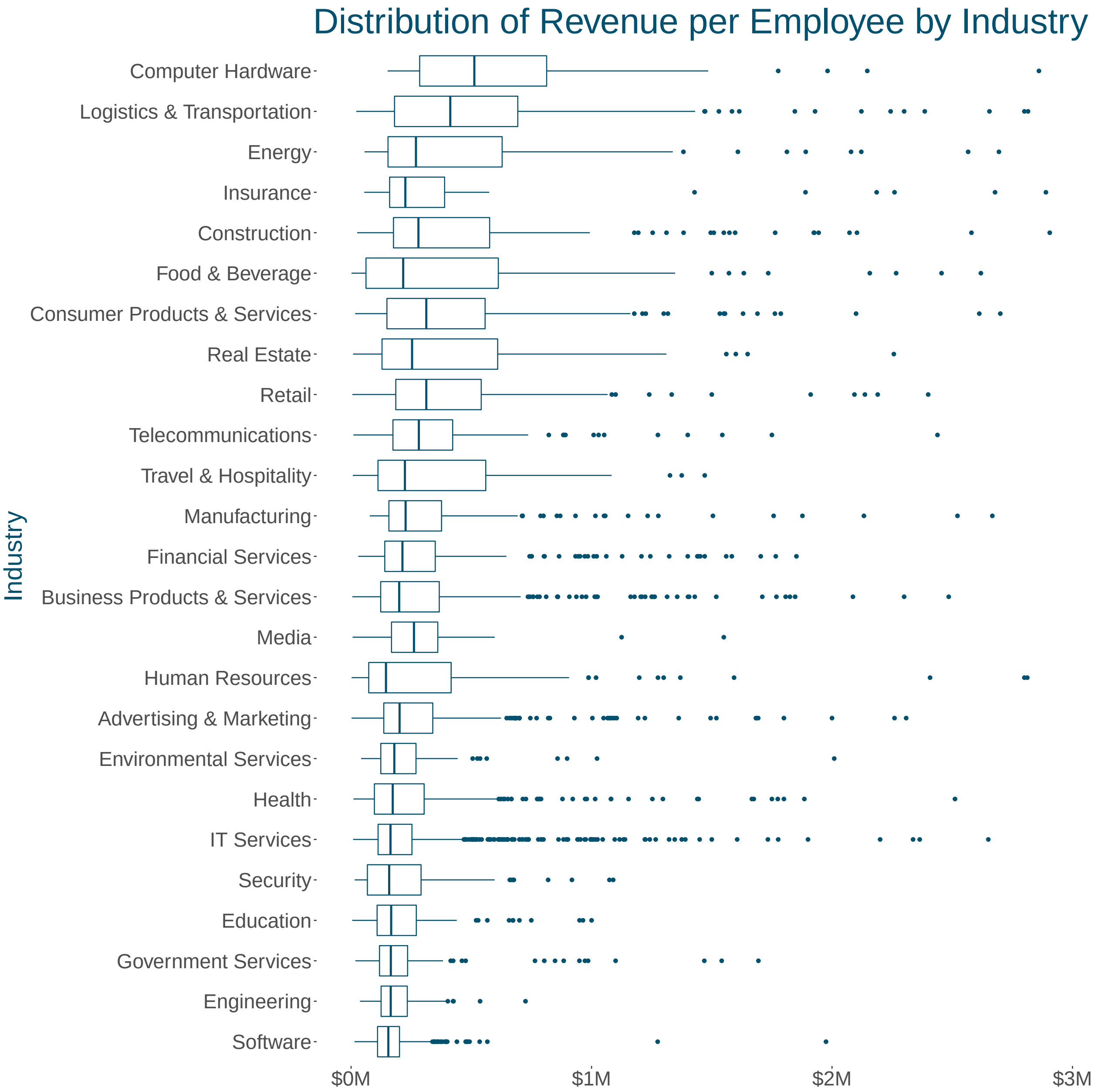
```
data_NY %>% ggplot() +
  aes(x = reorder(Industry, Employees), y = Employees) +
  stat_summary(fun.y=median, colour="darkred", geom = "point") +
  geom_boxplot(color = '#03506f') +
  ggtitle('Distribution of Employees by Industry') +
  xlab('Industry') +
  ylim(0,1200) +
  coord_flip() +
  theme(panel.background = element_rect(fill = "white", color = NA),
        plot.title = element_text(hjust = 0.5, size = 20, colour = "#03506f"),
        axis.title.y = element_text(size = 15, colour = "#03506f"),
        axis.title.x=element_blank(),
        axis.text.y = element_text(size = 15)

        )
```



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.

```r
# Answer Question 3 here
#visdat::vis_miss(inc)

inc %>% group_by(Industry) %>%
  mutate(Employee_Revenue = round((Revenue * 1e9) / Employees,0)) %>%
  filter (Employee_Revenue <= 3000000) %>%
  ggplot() + aes(x = reorder(Industry, Employee_Revenue), y = Employee_Revenue) +
  geom_boxplot(color = "#03506f") +
  coord_flip() +
  ggtitle('Distribution of Revenue per Employee by Industry') +
  xlab('Industry') +
  scale_y_continuous(breaks = c(0, 1000000, 2000000, 3000000),label = c("$0M", "$1M", "$2M", "$3M")) +
  theme(panel.background = element_rect(fill = "white", color = NA),
        plot.title = element_text(hjust = 0.5, size = 35, colour = "#03506f"),
        axis.title.y = element_text(size = 25, colour = "#03506f"),
        axis.title.x=element_blank(),
        axis.text.y = element_text(size = 20),
        axis.text.x = element_text(size = 20)
        )
```



The above boxplot shows, Computer Hardware, and Logistics and Transportation generates a high revenue per employee. However, Software and Engineering produce low revenue. (Note: Applied filter on Employee_Revenue less than or eaual to 3M due to large value outlier)