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:

```
require(dplyr)
```

```
## Loading required package: 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
```

```
require(tidyr)
```

```
## Loading required package: tidyr
```

```
require(knitr)
```

```
## Loading required package: knitr
```

```
require(kableExtra)
```

```
## Loading required package: kableExtra
```

```
## Warning: package 'kableExtra' was built under R version 3.6.3
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
require(kable)
```

```
## Loading required package: kable
```

```
## Warning in library(package, lib.loc = lib.loc, character.only = TRUE,
## logical.return = TRUE, : there is no package called 'kable'
```

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
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:

```
RevState<-inc%>%
          #filter(complete.cases(.))%>%
          group_by(State)%>%
          summarise(Total_Rev=sum(Revenue))%>%
          mutate(Total_Rev)%>%
          arrange(desc(Total_Rev))

RevState
```

```
## # A tibble: 52 x 2
##    State   Total_Rev
##    <fct>       <dbl>
##  1 IL    33244300000
##  2 CA    23457900000
##  3 TX    22164200000
##  4 NY    18260400000
##  5 OH    12786600000
##  6 FL    10610300000
##  7 NC     9258500000
##  8 VA     8667700000
##  9 MI     7805800000
## 10 WI     7296600000
## # ... with 42 more rows
```

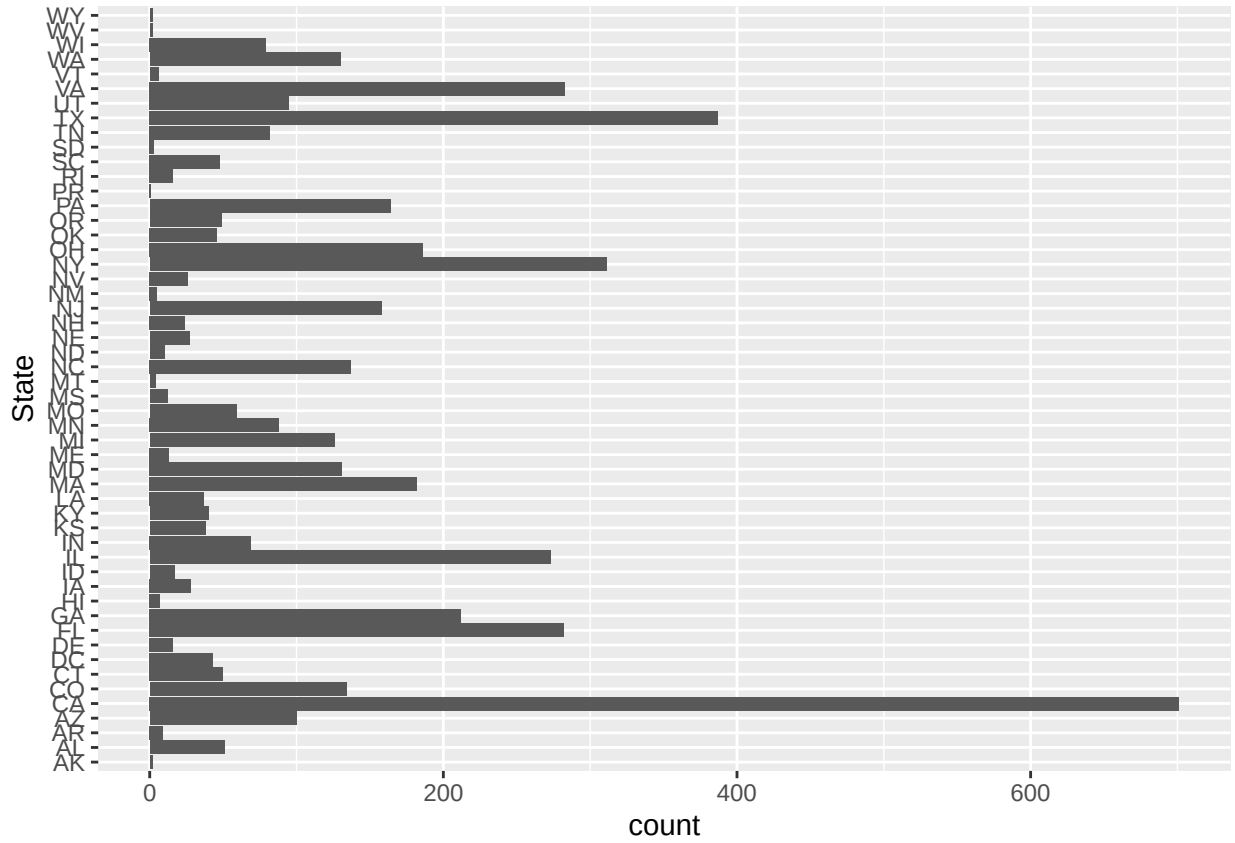
I explore which state have the highest revenue. According to the table, Illinois has the highest Revenue, and California is the next.

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

ggplot(inc,aes(x=State))+geom_bar()+coord_flip()
```



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

inc2 <- inc%>%
  group_by(State)%>%
  count(State)%>%
  arrange(desc(n))

inc2
```

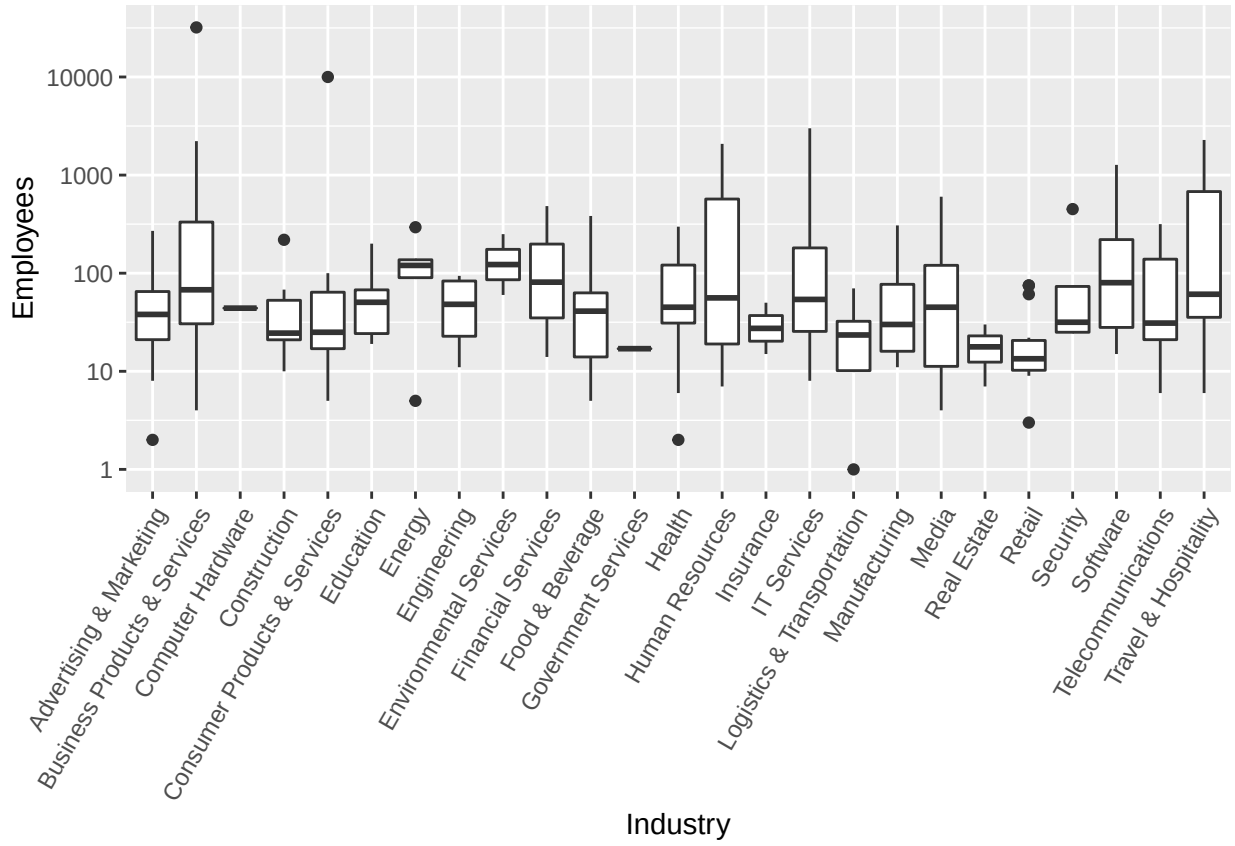
```
## # A tibble: 52 x 2
## # Groups:   State [52]
##    State     n
##    <fct> <int>
##  1 CA      701
##  2 TX      387
##  3 NY      311
##  4 VA      283
```

```
##  5 FL      282
##  6 IL      273
##  7 GA      212
##  8 OH      186
##  9 MA      182
## 10 PA      164
## # ... with 42 more rows
```

The table indicate the 3rd most companies is NY, so we will extract the data from NY.

```
inc3 <-inc %>%
       filter(State == 'NY') %>%
       filter(complete.cases(.))


ggplot(inc3, aes(x = Industry,y = Employees)) + geom_boxplot() + scale_y_continuous(trans='log10') + th
```



Question 3

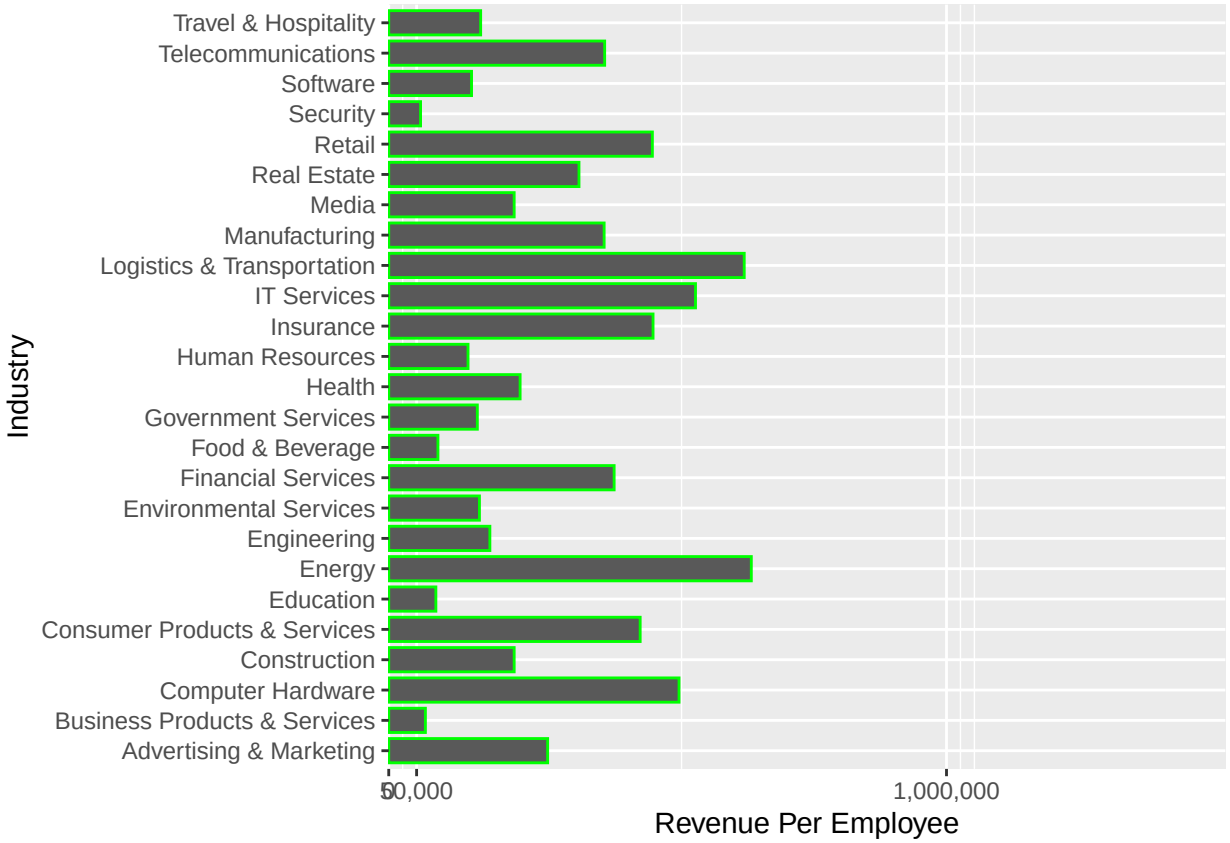
Now imagine you work for an investor and want to see which industries generate the most revenue per employee. Create a chart that makes this information clear. Once again, the distribution per industry should be shown.

```r
# Answer Question 3 here

inc4<-inc3 %>%
      filter(complete.cases(.))%>%
      group_by(Industry)%>%
      summarise(totalRev=sum(Revenue), totalEmp=sum(Employees))%>%
      mutate(RevPerEmp=totalRev/totalEmp)%>%
      arrange(desc(RevPerEmp))
inc4
```

```
## # A tibble: 25 x 4
##    Industry                       totalRev totalEmp RevPerEmp
##    <fct>                             <dbl>    <int>     <dbl>
##  1 Energy                        419900000      646   650000
##  2 Logistics & Transportation     75200000      118   637288.
##  3 IT Services                  4826200000     8776   549932.
##  4 Computer Hardware              22900000       44   520455.
##  5 Insurance                      30800000       65   473846.
##  6 Retail                        164000000      347   472622.
##  7 Consumer Products & Services 4799300000    10647   450765.
##  8 Financial Services            758100000     1876   404104.
##  9 Telecommunications            627500000     1621   387107.
## 10 Manufacturing                 368000000      953   386149.
## # ... with 15 more rows
```

```r
ggplot(inc4, aes(x=Industry, y=RevPerEmp))+
  geom_bar(stat = "identity", position=position_dodge(), colour="green", width = 0.8)+coord_flip()+
 scale_y_continuous(expand = c(0, 0), limits = c(0, 1500000), breaks = c(0, 50000,     1000000), labels
    ylab("Revenue Per Employee")
```


Travel & Hospitality
Telecommunications
Software
Security
Retail
Real Estate
Media
Manufacturing
Logistics & Transportation
IT Services
Insurance
Human Resources
Health
Government Services
Food & Beverage
Financial Services
Environmental Services
Engineering
Energy
Education
Consumer Products & Services
Construction
Computer Hardware
Business Products & Services
Advertising & Marketing
Industry
50,000
1,000,000
Revenue Per Employee