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/inc5000_data.csv", header= TRUE)
```

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         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:

```
# Insert your code here, create more chunks as necessary
```

```
# using unique on Industry we see there are 25 different industries represented here
unique(inc["Industry"])
```

```
##                          Industry
## 1   Consumer Products & Services
## 2             Government Services
## 3                          Health
## 4                          Energy
## 5         Advertising & Marketing
## 6                     Real Estate
## 7              Financial Services
## 10                         Retail
## 14                       Software
## 16              Computer Hardware
## 17    Logistics & Transportation
## 19                Food & Beverage
## 21                    IT Services
## 24  Business Products & Services
## 35                     Education
## 50                  Construction
## 64                 Manufacturing
## 90            Telecommunications
## 110                     Security
## 137               Human Resources
## 153         Travel & Hospitality
## 174                        Media
## 531       Environmental Services
## 532                   Engineering
## 552                     Insurance
```

```
# similarly doing the same on state but just using the count we can see all states are r
epresented including Washington DC and Puerto Rico
nrow(unique(inc["State"]))
```

```
## [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.

```
library(dplyr)
```

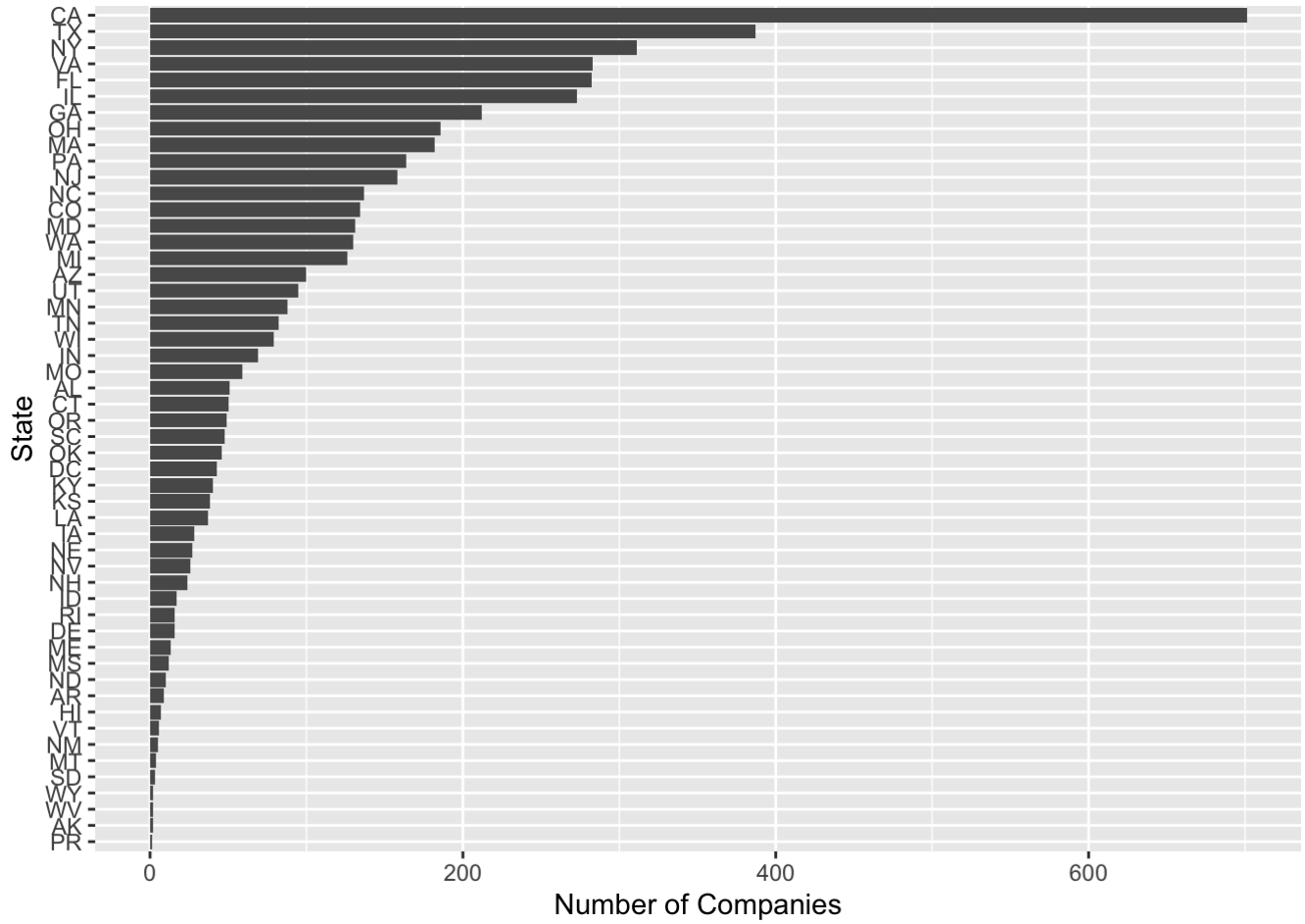
```
##
## Attaching package: 'dplyr'
```

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

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

```
library(ggplot2)
```

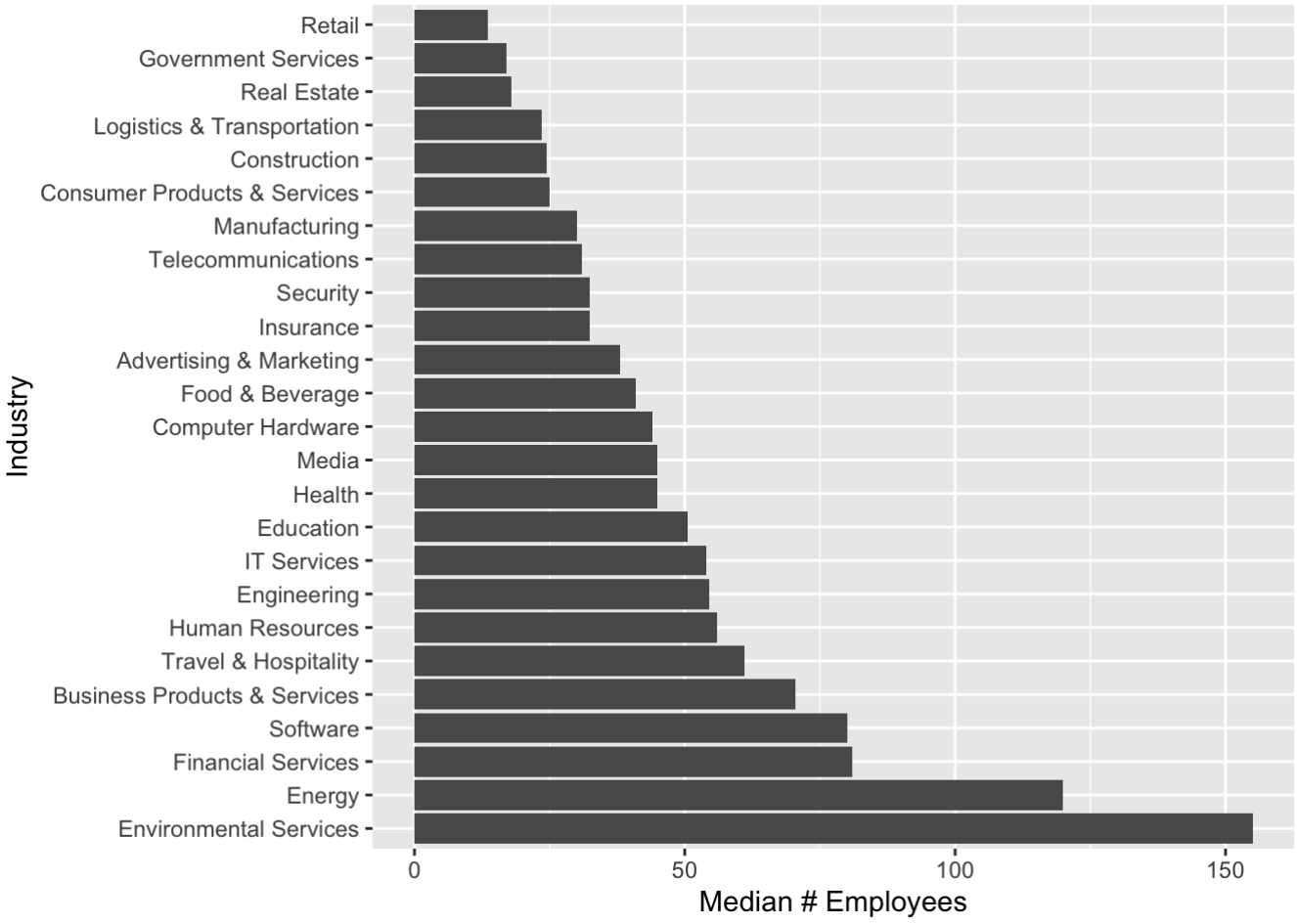
```
# Answer Question 1 here
inc %>%
  count(State, n(), sort=T) %>%
  ggplot(aes(x = reorder(State,n), y = n))+
  geom_col() +
  coord_flip() +
  labs(x="State", y="Number of Companies")
```



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
incNy <- inc %>% filter(complete.cases(.) & State == 'NY')
incNy %>%
  group_by(Industry) %>%
  summarise(medianByInd = median(Employees)) %>%
  ggplot(aes(x=reorder(Industry,-medianByInd), y=medianByInd)) +
  geom_col() +
  coord_flip() +
  labs(x = "Industry", y = "Median # 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.

```
# Answer Question 3 here
inc %>%
  mutate(revPerEmployee = Revenue / Employees) %>%
  group_by(Industry) %>%
  summarise(medianRevPerEmployee = median(revPerEmployee, na.rm=T)) %>%
  ggplot(aes(x=reorder(Industry,-medianRevPerEmployee), y=medianRevPerEmployee)) + geom_
col() + coord_flip() +
  scale_y_continuous(labels=scales::dollar_format()) +
  labs(y = "Median Revenue Per Employee", x = "Industry")
```