R Notebook

Code ▾

Neil Shah HW1: DATA 608

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:

Hide

```
inc <- read.csv("https://raw.githubusercontent.com/charleyferrari/CUNY_DATA_608/master/module1/D
ata/inc5000_data.csv", header= TRUE)
```

And lets preview this data:

```
head(inc)

  Rank                         Name Growth_Rate   Revenue                     Industry Employees
City State
1    1                         Fuhu      421.48 1.179e+08 Consumer Products & Services       104
El Segundo    CA
2    2       FederalConference.com      248.31 4.960e+07          Government Services        51
Dumfries    VA
3    3                The HCI Group      245.45 2.550e+07                       Health       132
Jacksonville    FL
4    4                      Bridger      233.08 1.900e+09                       Energy        50
Addison    TX
5    5                       DataXu      213.37 8.700e+07      Advertising & Marketing       220
Boston    MA
6    6 MileStone Community Builders      179.38 4.570e+07                  Real Estate        63
Austin    TX
>
```

```
summary(inc)

  Rank                          Name        Growth_Rate          Revenue
 Min.   :   1   (Add)ventures         :   1   Min.   :  0.340   Min.   :2.000e+06
 1st Qu.:1252   @Properties           :   1   1st Qu.:  0.770   1st Qu.:5.100e+06
 Median :2502   1-Stop Translation USA:   1   Median :  1.420   Median :1.090e+07
 Mean   :2502   110 Consulting        :   1   Mean   :  4.612   Mean   :4.822e+07
 3rd Qu.:3751   11thStreetCoffee.com  :   1   3rd Qu.:  3.290   3rd Qu.:2.860e+07
 Max.   :5000   123 Exteriors         :   1   Max.   :421.480   Max.   :1.010e+10
                (Other)               :4995
                         Industry      Employees                  City          State
 IT Services                 : 733   Min.   :    1.0   New York     : 160   CA     : 701
 Business Products & Services: 482   1st Qu.:   25.0   Chicago      :  90   TX     : 387
 Advertising & Marketing     : 471   Median :   53.0   Austin       :  88   NY     : 311
 Health                      : 355   Mean   :  232.7   Houston      :  76   VA     : 283
 Software                    : 342   3rd Qu.:  132.0   San Francisco:  75   FL     : 282
 Financial Services          : 260   Max.   :66803.0   Atlanta      :  74   IL     : 273
 (Other)                     :2358   NA's   :12        (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:

These summaries provide robust statistics on the overall columns in the data-set. They give us a cursory view of the entire data-set and alert us to overall trends, outliers and serve as a baseline to start out analysis.

As a financial profession–I really like to look at skewness to give me an idea of how a distriubtion might lean.

I found this package called Performance Analytics here (https://rviews.rstudio.com/2017/12/13/introduction-to-skewness/) and used it’s skew function.

Let’s apply this to the numerical categories in inc.

```
library('PerformanceAnalytics')

> skewness(inc$Growth_Rate)
[1] 12.55327
>
> skewness(inc$Revenue)
[1] 22.1811
> skewness(inc$Employees)
[1] 29.81938
```

Notice that all of these values have a positive skew–meaning that they have tails to the right–this is interesting and might point to possible outliers. From a robust statistics side–we might need to look at median instead of mean to get an idea of variability.

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

First lets get some imports

```
library(ggplot2)
library(zeallot)
```

Now let's make a table and group by State

```


state <- inc %>% group_by(State) %>% summarize(Count = n())
```

Take a look at it

```
head(state)

> head(state)
# A tibble: 6 x 2
  State Count
  <fct> <int>
1 AK        2
2 AL       51
3 AR        9
4 AZ      100
5 CA      701
6 CO      134
```

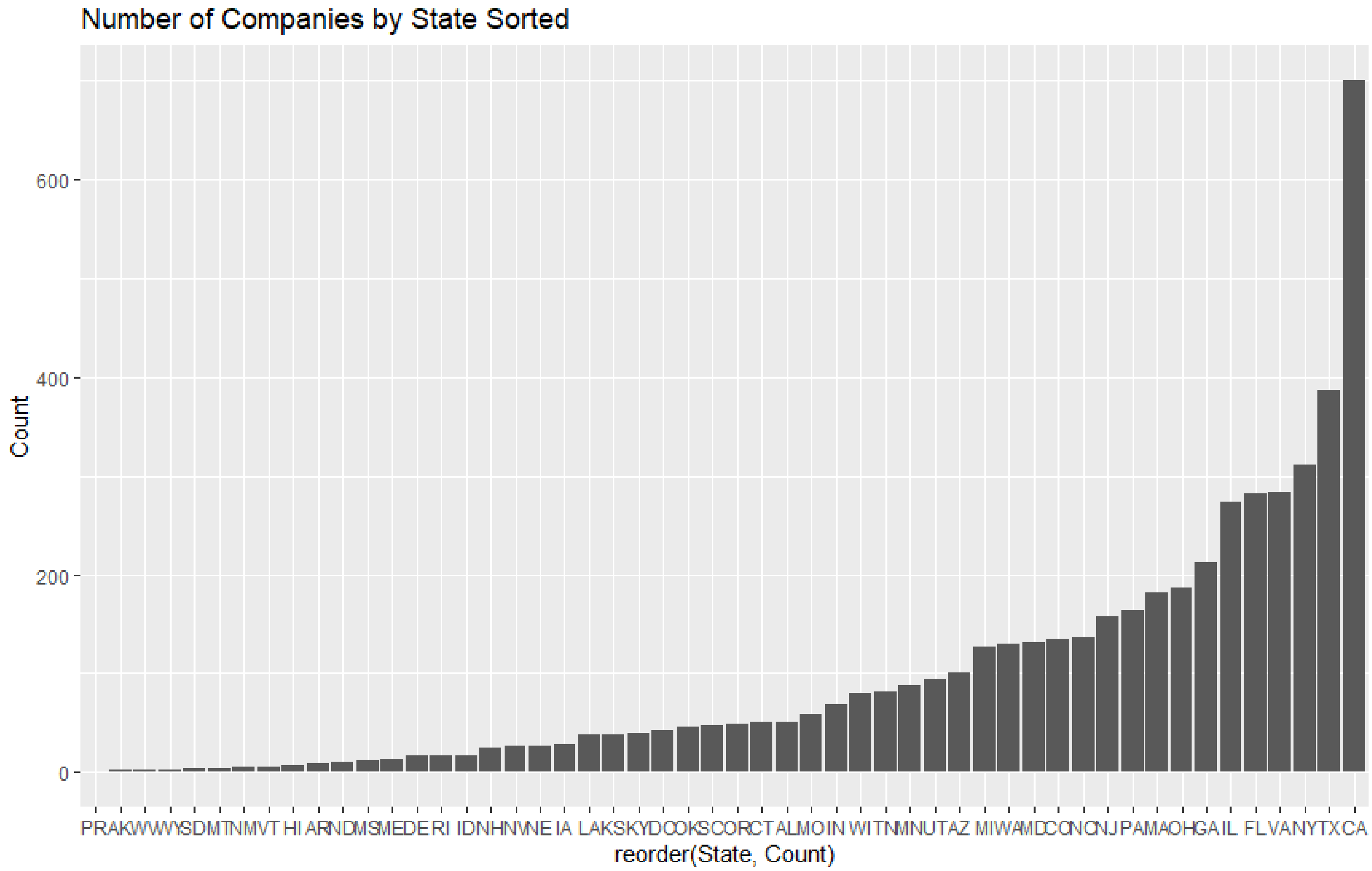
Ok let's sort this out

```
> state %>% arrange(desc(Count))
# A tibble: 52 x 2
   State Count
   <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


state <- state %>% arrange(desc(Count))
```

Now let's plot it

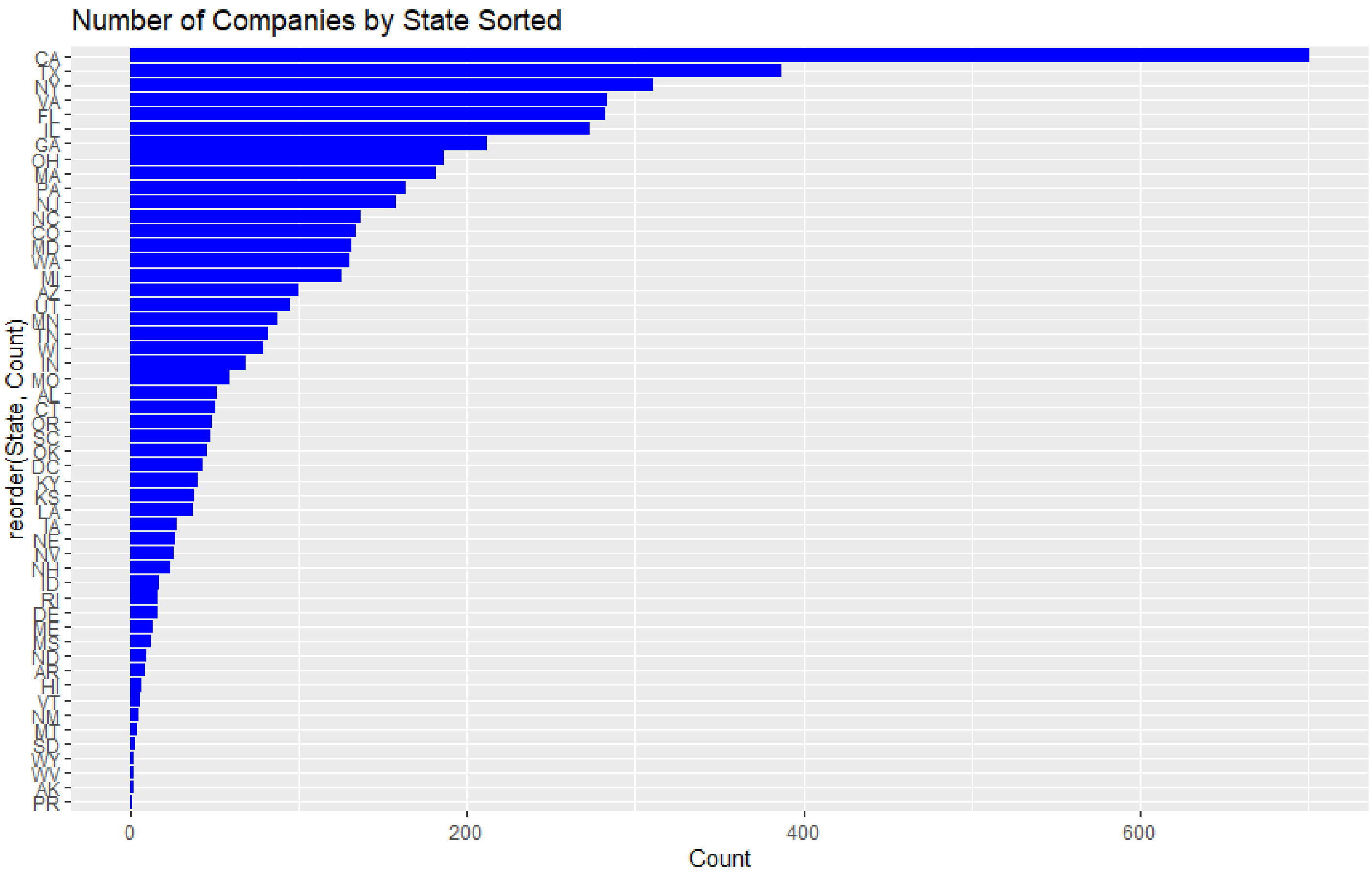
```
>ggplot(state, aes(x = reorder(State, Count), y = Count)) +geom_bar(stat = "identity") +ggtitle
('Number of Companies by State Sorted')
```



Number of Companies by State Plot

Alright–now let's clean it up, add some color and fix the axis so we can see the labels.

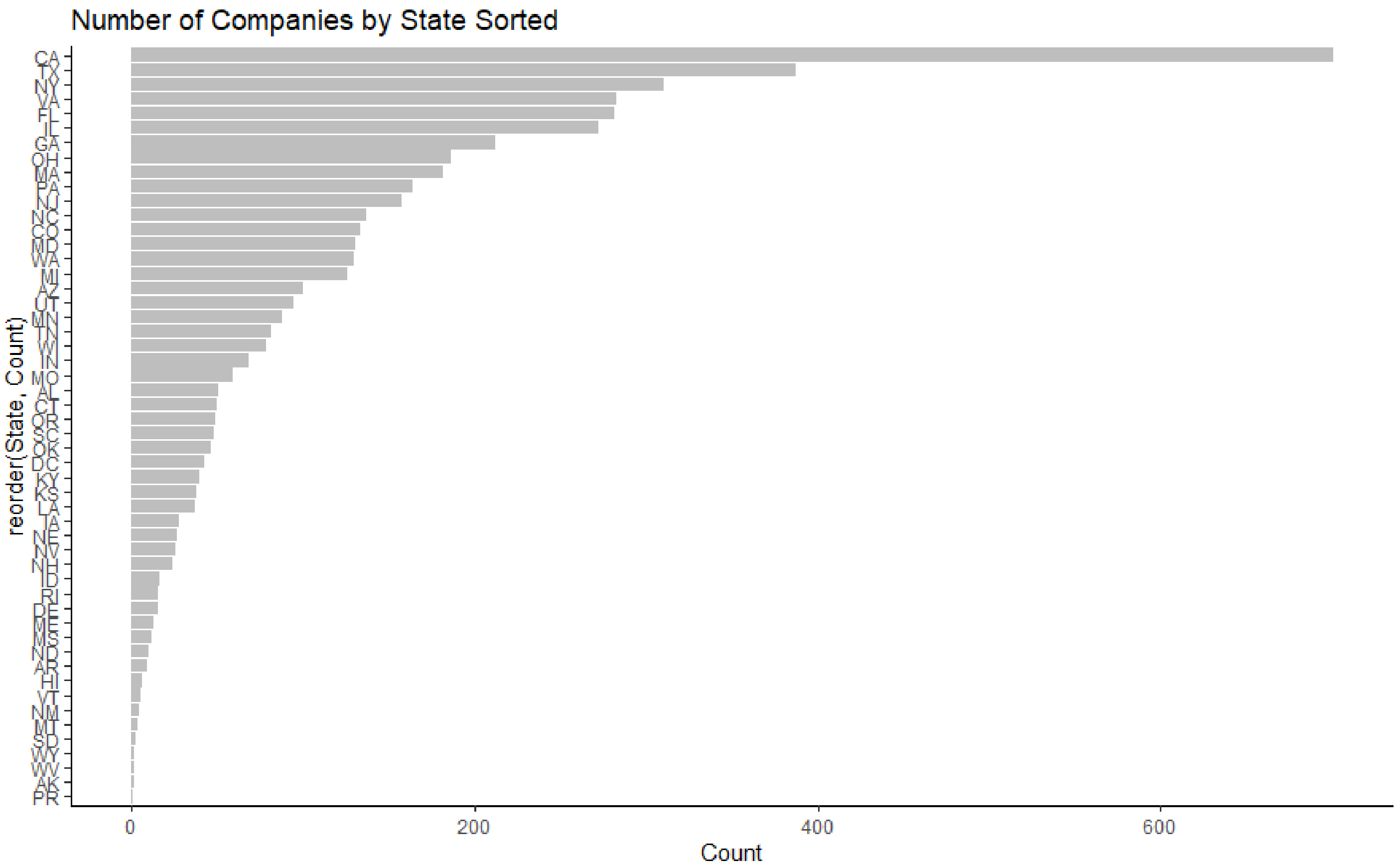
```
ggplot(state, aes(x = reorder(State, Count), y = Count)) +geom_bar(stat = "identity", fill='blu
e') +ggtitle('Number of Companies by State Sorted')+coord_flip()
```



Number of Companies by State Plot

Since I just learned about date-ink ratio–let’s try to apply it here. I am referencing the methods from Felix Fans Reference Site (https://felixfan.github.io/ggplot2-remove-grid-background-margin/)

```
> ggplot(state, aes(x = reorder(State, Count), y = Count)) +geom_bar(stat = "identity", fill='gr
ey') +ggtitle('Number of Companies by State Sorted')+coord_flip() + theme(panel.grid.major = ele
ment_blank(), panel.grid.minor = element_blank(),
+ panel.background = element_blank(), axis.line = element_line(colour = "black"))
```

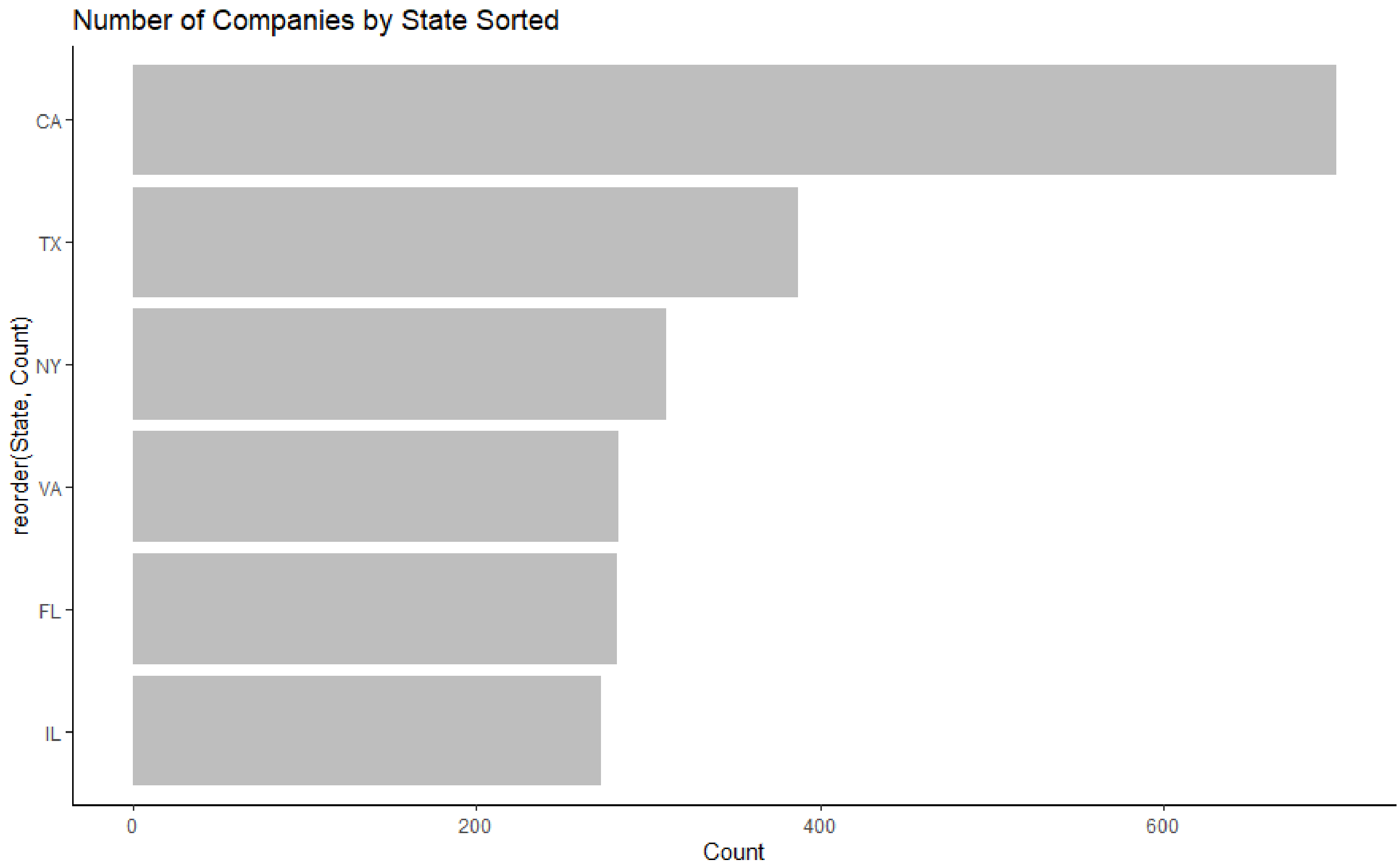


Number of Companies by State Plot

There–much cleaner! So it seems that the top states are CA, TX and then NY.

Let’s zoom in on the top values to focus onthem

```
> ggplot(head(state), aes(x = reorder(State, Count), y = Count)) +geom_bar(stat = "identity", fi
ll='grey') +ggtitle('Number of Companies by State Sorted')+coord_flip() + theme(panel.grid.major
= element_blank(), panel.grid.minor = element_blank(),
+ panel.background = element_blank(), axis.line = element_line(colour = "black"))
>
```



Top by State Plot

This makes some sense to me given that these states have the highest populations.

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.

So based on our data-set we are looking at good ole NY

First let’s sort out our dataset for NY only and by industry

```
> inc %>% filter(State=='NY') %>%filter(complete.cases(.)) %>% group_by(Industry)
# A tibble: 311 x 8
# Groups:   Industry [25]
    Rank Name                       Growth_Rate  Revenue Industry                     Employees C
ity      State
   <int> <fct>                            <dbl>    <dbl> <fct>                            <int> <
fct>     <fct>
 1    26 BeenVerified                      84.4 13700000 Consumer Products & Services        17 N
ew York  NY
 2    30 Sailthru                          73.2  8100000 Advertising & Marketing             79 N
ew York  NY
 3    37 YellowHammer                      67.4 18000000 Advertising & Marketing             27 N
ew York  NY
 4    38 Conductor                         67.0  7100000 Advertising & Marketing             89 N
ew York  NY
 5    48 Cinium Financial Services         53.6  5900000 Financial Services                  32 R
ock Hill NY
 6    70 33Across                          45.0 27900000 Advertising & Marketing             75 N
ew York  NY
 7    71 LiveIntent                        44.8  6900000 Advertising & Marketing             42 N
ew York  NY
 8   124 Quantum Networks                  29.4 11500000 Telecommunications                  28 N
ew York  NY
 9   126 Renegade Furniture Group          29.3  9800000 Retail                              17 H
ewlett   NY
10   153 Regal Wings                       25.1 15400000 Travel & Hospitality                42 B
rooklyn  NY
# ... with 301 more rows```
```

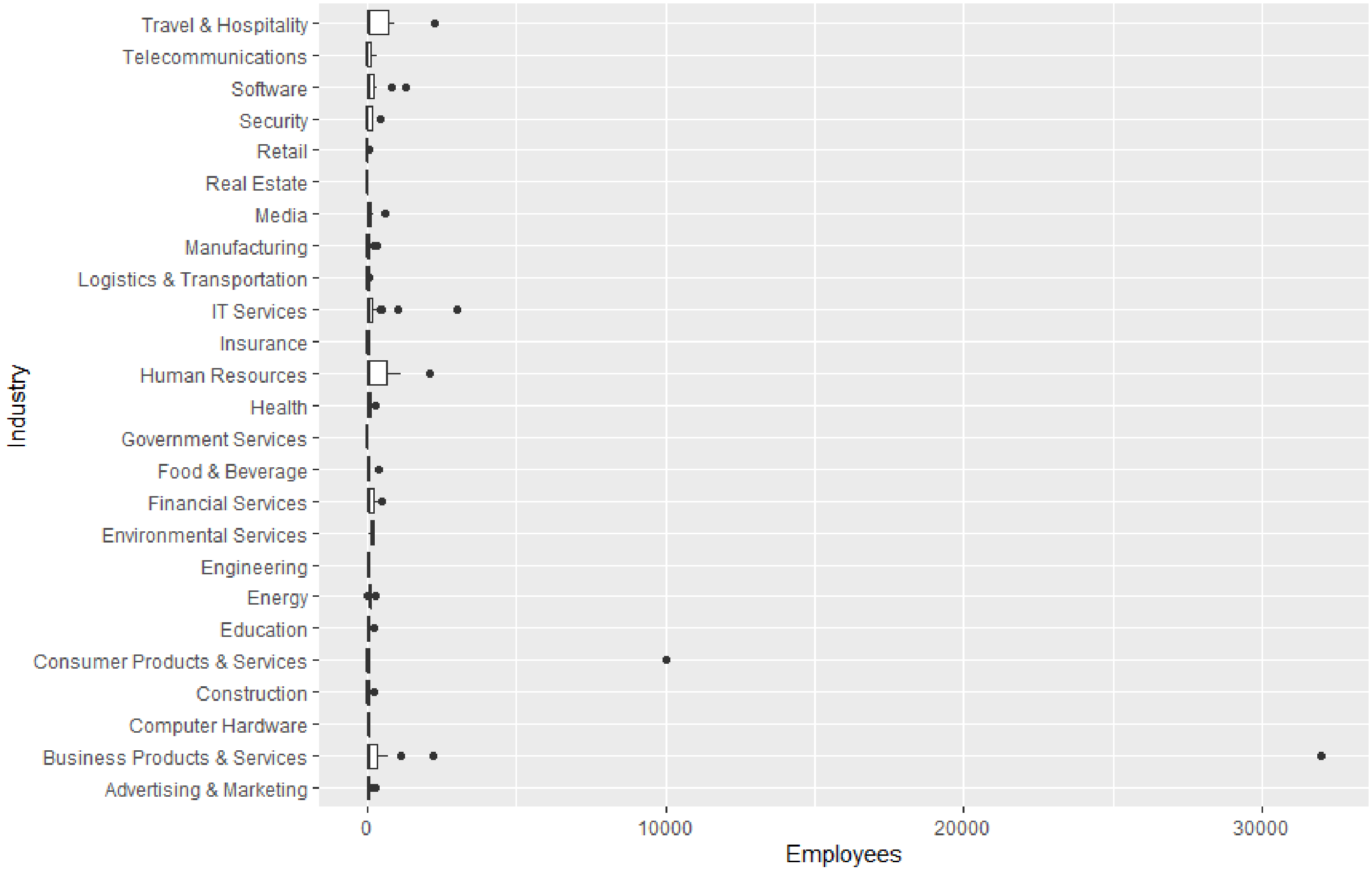
I like to do a quick summary statistics to explore the data-set

```
> summary(inc %>% filter(State=='NY') %>%filter(complete.cases(.)) %>% group_by(Industry))
      Rank                            Name      Growth_Rate        Revenue          
Industry
 Min.   :  26   1st Equity          :  1   Min.   : 0.350   Min.   :2.000e+06   Advertising & Ma
rketing     : 57
 1st Qu.:1186   33Across            :  1   1st Qu.: 0.670   1st Qu.:4.300e+06   IT Services     
: 43
 Median :2702   5Linx Enterprises   :  1   Median : 1.310   Median :8.800e+06   Business Product
s & Services: 26
 Mean   :2612   Access Display Group:  1   Mean   : 4.371   Mean   :5.872e+07   Consumer Product
s & Services: 17
 3rd Qu.:4005   Adafruit            :  1   3rd Qu.: 3.580   3rd Qu.:2.570e+07   Telecommunicatio
ns          : 17
 Max.   :4981   AdCorp Media Group  :  1   Max.   :84.430   Max.   :4.600e+09   Education       
: 14
                (Other)             :305                                        (Other)         
:137
   Employees              City         State    
 Min.   :    1.0   New York :160   NY     :311  
 1st Qu.:   21.0   Brooklyn : 15   AK     :  0  
 Median :   45.0   Rochester:  9   AL     :  0  
 Mean   :  271.3   Buffalo  :  5   AR     :  0  
 3rd Qu.:  105.5   Fairport :  5   AZ     :  0  
 Max.   :32000.0   new york :  5   CA     :  0  
                   (Other)  :112   (Other):  0  
```

Two things that I want to point out–1) State only has values for NY, which is good! That means my filter by NY worked out and 2) looking at the statistical summary of Employees–the max is 32000 which is well above the IQR ranges; we definitely are going to have outliers!

The easiest way to display variance, median and spread is a boxplot; let’s do that. To make things easier I’ll save the modified dataframe.

```
ny <- inc %>% filter(State=='NY') %>%filter(complete.cases(.)) %>% group_by(Industry)
ggplot(ny,aes(x=Industry,y=Employees))+geom_boxplot()+coord_flip()
```

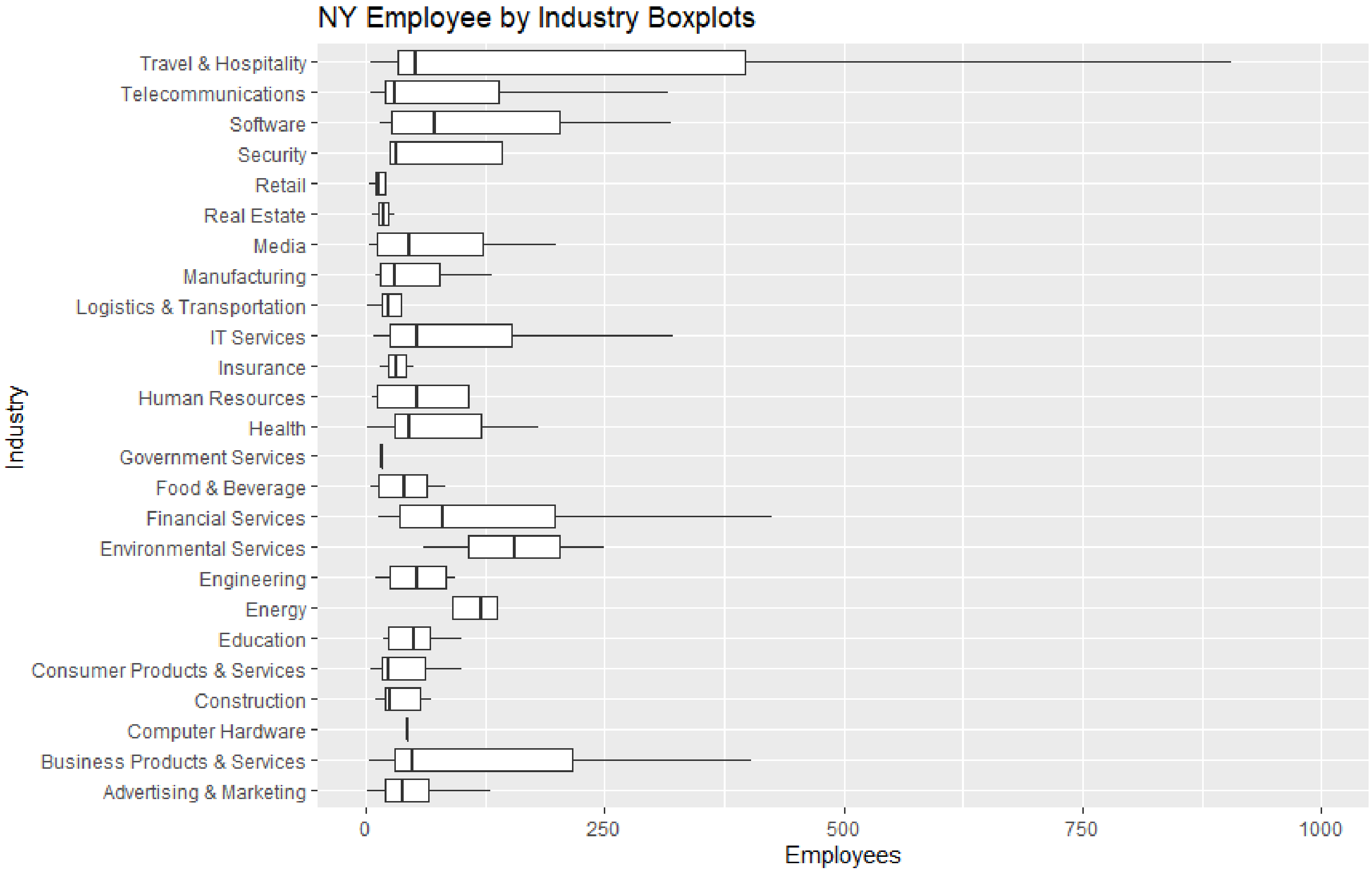


Box Plot

Looks like we have some serious outliers!

So I could extract the outliers and remove them but what is easier is just to cut my axis and hide the outliers.

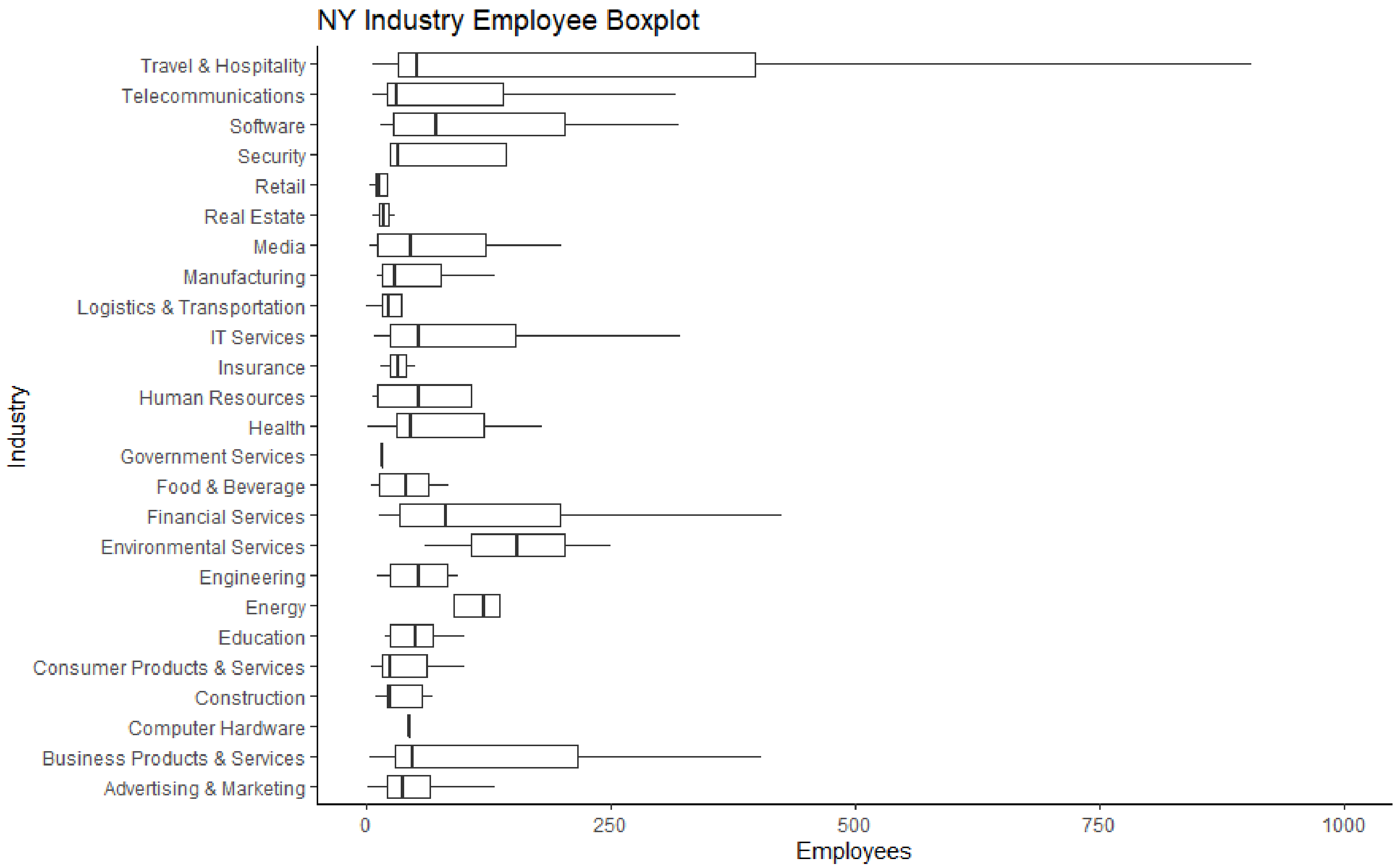
```
> ggplot(ny,aes(x=Industry,y=Employees))+geom_boxplot(outlier.shape=NA)+ ggtitle('NY Industry Em
ployee Boxplot') + coord_flip()+ylim(0,1000)
```



Box Plot

Now let’s combine everything like we did before and make the plot readable via our Data to Ink ratio method.

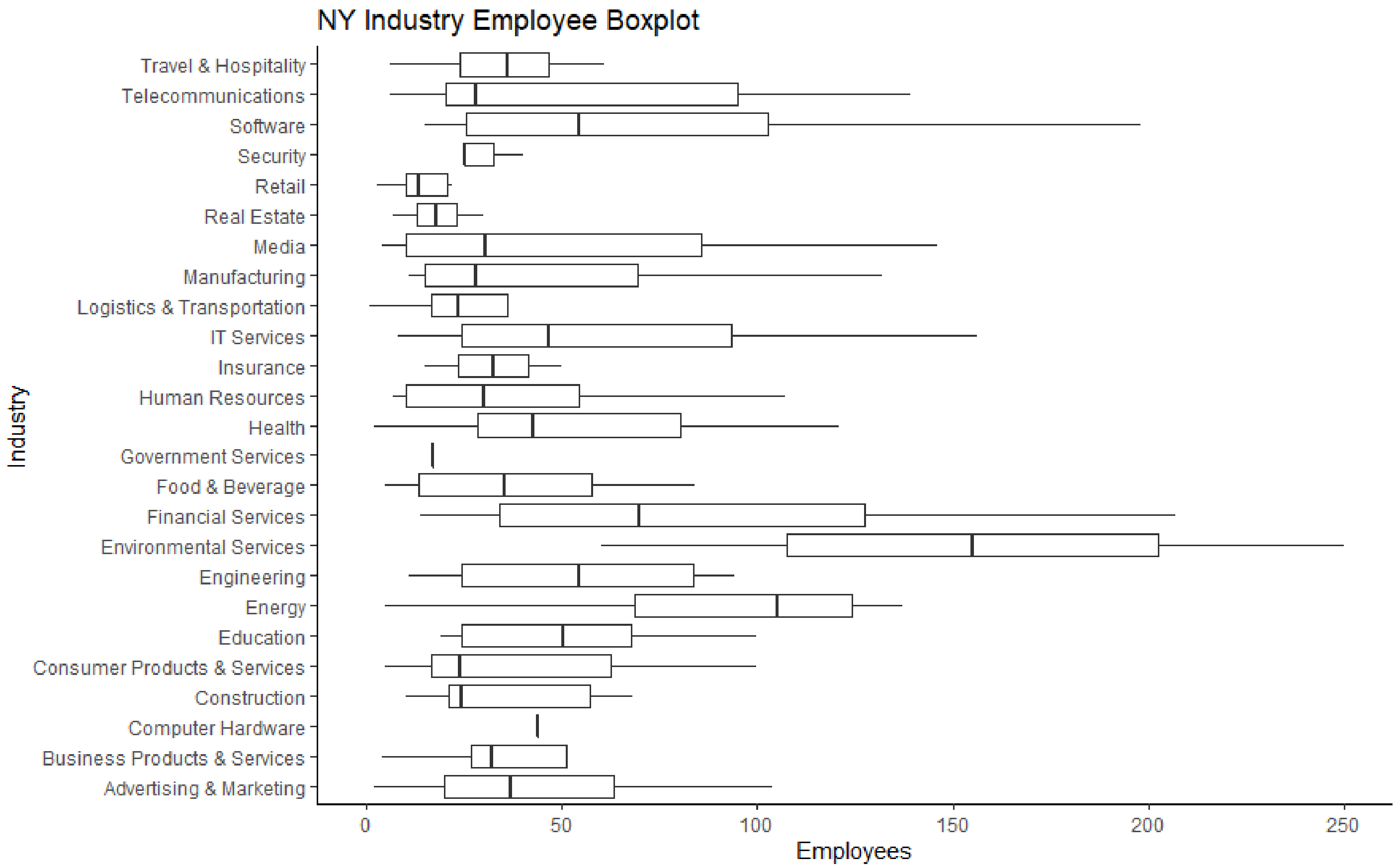
```
> ggplot(ny,aes(x=Industry,y=Employees))+geom_boxplot(outlier.shape=NA)+ ggtitle('NY Industry Em
ployee Boxplot') + coord_flip()+ylim(0,1000) + theme_bw() + theme(panel.border = element_blank
(), panel.grid.major = element_blank(),
+ panel.grid.minor = element_blank(), axis.line = element_line(colour = "black"))
```



Clean Box Plot

Much nicer!

I’m going to include one more zoomed in plot just to see more granuarity.



Clean Box Plot

So looking over the data just some quick observations:
The Travel and Hospitality indstry has the largest spread/variability, given the whisker range/IQR range.
Computer hardware and Government Services have the most narrow spreads
The median for the NY industries are all below 250
Government services has the lowest median employees.
Environmental Services ahs the highest median employees.

Fascinating–ths would be a cool study to dig down further.

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.

First let’s define a metric called medianemp defined by Revenue/Employees.

```
subset(inc, complete.cases(inc)) %>%
    mutate(medianemp = Revenue/Employees)
```

I am going to just store this as a new dataframe to make my life easier

```

inc_investor <- subset(inc, complete.cases(inc)) %>%
    mutate(medianemp = Revenue/Employees)
```

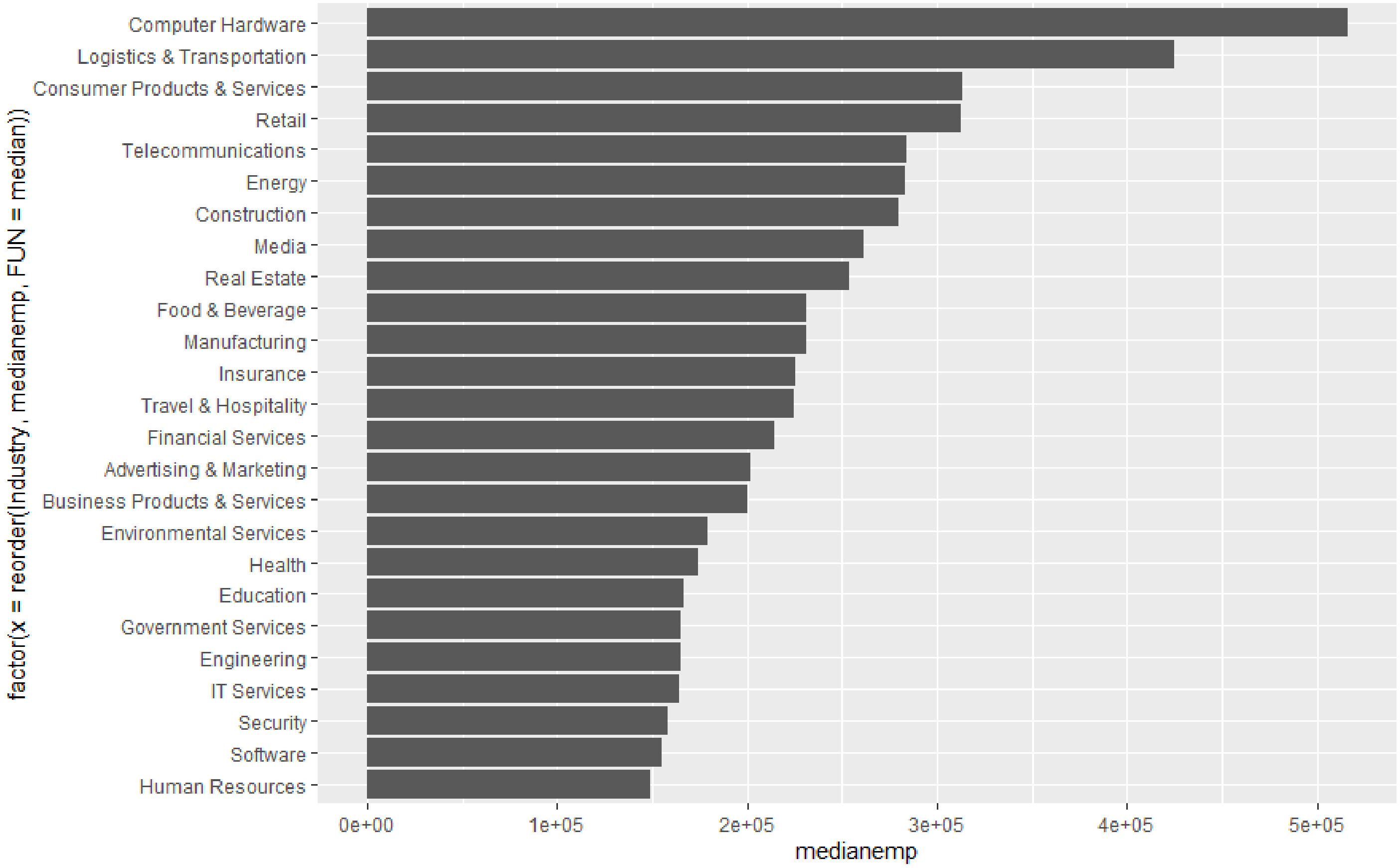
Let’s quickly look at summary statistics

```
> summary(inc_case)
      Rank                              Name       Growth_Rate         Revenue          
 Min.   :   1   (Add)ventures         :   1   Min.   :  0.340   Min.   :2.000e+06  
 1st Qu.:1252   @Properties           :   1   1st Qu.:  0.770   1st Qu.:5.100e+06  
 Median :2502   1-Stop Translation USA:   1   Median :  1.420   Median :1.090e+07  
 Mean   :2501   110 Consulting        :   1   Mean   :  4.615   Mean   :4.825e+07  
 3rd Qu.:3750   11thStreetCoffee.com  :   1   3rd Qu.:  3.290   3rd Qu.:2.860e+07  
 Max.   :5000   123 Exteriors         :   1   Max.   :421.480   Max.   :1.010e+10  
                (Other)               :4983                                        
                         Industry      Employees                  City          State        med
ianemp        
 IT Services                 : 732   Min.   :    1.0   New York     : 160   CA     : 700   Min.   
:    1801  
 Business Products & Services: 480   1st Qu.:   25.0   Chicago      :  90   TX     : 386   1st Q
u.:  125000  
 Advertising & Marketing     : 471   Median :   53.0   Austin       :  88   NY     : 311   Media
n :  198658  
 Health                      : 354   Mean   :  232.7   Houston      :  76   VA     : 283   Mean   
:  393613  
 Software                    : 341   3rd Qu.:  132.0   San Francisco:  74   FL     : 282   3rd Q
u.:  375000  
 Financial Services          : 260   Max.   :66803.0   Atlanta      :  73   IL     : 272   Max.   
:40740000  
 (Other)                     :2351                     (Other)      :4428   (Other):2755        
```

Focusing in on medianemp–the metric i defined, it appears that the median is around $200,000 per employee and most of the distribution is under $400,000–however look at that outlier! Let’s investigate.

Let’s plot this out

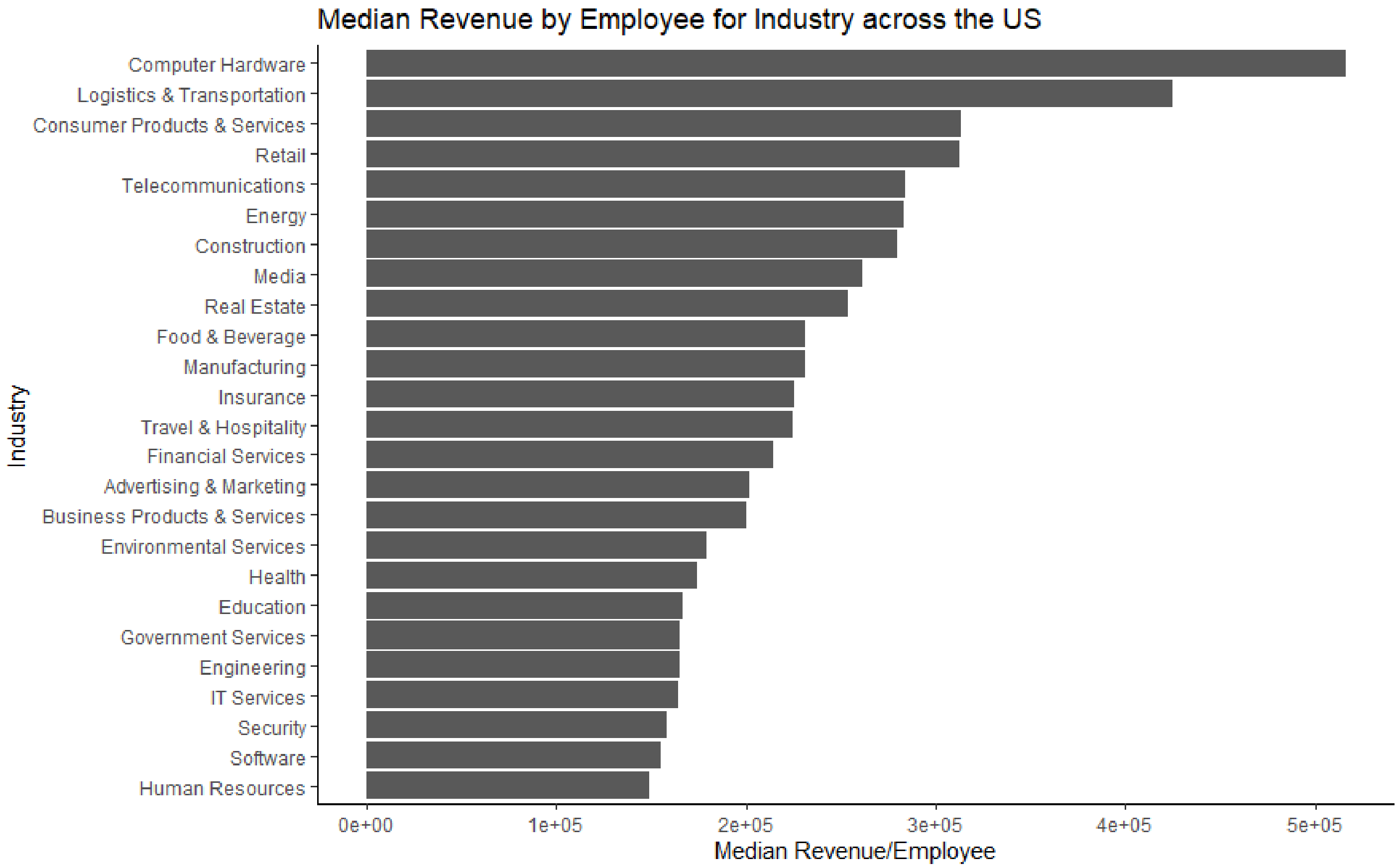
```
> ggplot(inc_investor, aes(factor(x = reorder(Industry, medianemp, FUN = median)))) +
+ stat_summary_bin(aes(y = medianemp), fun.y = "median", geom = "bar") +
+ coord_flip()
```



Median Revenue/Employee by Industry

Now combining our plotting methods from before

```
> ggplot(inc_investor, aes(factor(x = reorder(Industry, medianemp, FUN = median)))) +
+ stat_summary_bin(aes(y = medianemp), fun.y = "median", geom = "bar") + coord_flip() + xlab("In
dustry") + ylab("Median Revenue/Employee")+ ggtitle('Median Revenue by Employee for Industry acr
oss the US') + theme_bw() + theme(panel.border = element_blank(), panel.grid.major = element_bla
nk(),
+ panel.grid.minor = element_blank(), axis.line = element_line(colour = "black"))
```



Median Revenue/Employee by Industry

Much cleaner and clearer!

If we recall from the origianl summary table–the median revenue per employee was around $200,000 but it seems that two industries are out liers (on the high end)–Computer Hardware and Logistics and Transport. As an investor–I don't want average performance since we have to beat the market; Computer Hardware and Logistics Transports seem two industries we should further analyze for invesment opportunities.

This provies a good starting ground for an investment thesis!

Conlusions

In this assignment I was able to load a dataframe, maniuplate it through filtering, produce summary statistics and plot the data in a clear/easy to read fashion.

These skills–while basic–are powerful and will serve as the foundation for my Data Analysis.