Module 1 for DATA 608

Thomas Hill

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:

```
library(dplyr)
library(ggplot2)
library(forcats)
```

```
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
```

Looking at the top companies by growth, it apears that the top 5000 companies come from inc.com's 2013 list.

```
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:

The dataset provides eight variables to consider to describe the patterns of growth. Feature 'Name' is an important label for each firm, while 'Industry' may be used to draw useful comparisons. 'City' and 'State' variables offer the possibilty of identifying geographical trends or high growth clusters. 'Rank' and 'Growth_Rate' both describe the same underlying data - the rank given provides a rank of the highest growth companies. Growth rate is in terms of percent, i.e., the top company Fuhu grew 42,148% in 2013 or grew over 400 times its original size. 'Revenue' and 'Employees' provide a measure of each company's size and income.

There are two features I'm most interested in adding. The first is simply revenue divided by employee as a rough estimate of company and sector productivity. Next, I'll consider what I'm calling 'revenue change', or the absolute year-on-year growth for a company. This will allow for direct comparisons of change in revenue irrespective of the company's size, and will consider the growth rate paired with the revenue. My reasoning behind this is because many of the growth rates are exaggerated owing ot the small size of the company. Consistent but mediocre growth in a firm with high capitalization could still be more lucrative in the long run than speculating on smaller, private companies. Revenue growth will be useful in steering away from untested firms.

These variables can be further explored by grouping companies by industry. This will allow identification of fastest growing industries and provide useful rankings Additionally, I'll consider missing values for employees to see if this is significant. Finally, revenue and change in revenue can be expressed in milliions of dollars (MM) for readability reasons.

```
inc$Revenue <- (inc$Revenue)/10e5 #change to millions of dollars
inc$City <- toupper(inc$City) #keep city names a consistent case
inc$Industry <- as.factor(inc$Industry) #change to factor
inc$State <- as.factor(inc$State)#change to factor
inc$City <- as.factor(inc$City) #change to factor
```

```
inc <- inc %>%
  mutate(revenue_change = round(Revenue*(1-(1+Growth_Rate)^-1),2)) %>% # absolute dollar change
 in revenue over the past year, in millions
  mutate(revenue_employee = round((Revenue/Employees)*10e2,2)) %>% #dollars revenue per employe
e, thousands
  mutate(revenue_change_employee = round(10e2*Revenue*(1-(1+Growth_Rate)^-1)/Employees,2)) #chan
ge in revenue per employee, thousands
```

```
summary(inc)
```

```
##       Rank          Name            Growth_Rate         Revenue
##  Min.   :   1   Length:5001        Min.   :  0.340   Min.   :    2.00
##  1st Qu.:1252   Class :character   1st Qu.:  0.770   1st Qu.:    5.10
##  Median :2502   Mode  :character   Median :  1.420   Median :   10.90
##  Mean   :2502                      Mean   :  4.612   Mean   :   48.22
##  3rd Qu.:3751                      3rd Qu.:  3.290   3rd Qu.:   28.60
##  Max.   :5000                      Max.   :421.480   Max.   :10100.00
##
##                          Industry      Employees                  City
##  IT Services                 : 733   Min.   :    1.0   NEW YORK     : 166
##  Business Products & Services: 482   1st Qu.:   25.0   CHICAGO      :  93
##  Advertising & Marketing     : 471   Median :   53.0   AUSTIN       :  89
##  Health                      : 355   Mean   :  232.7   HOUSTON      :  77
##  Software                    : 342   3rd Qu.:  132.0   ATLANTA      :  75
##  Financial Services          : 260   Max.   :66803.0   SAN FRANCISCO:  75
##  (Other)                     :2358   NA's   :12        (Other)      :4426
##      State      revenue_change    revenue_employee  revenue_change_employee
##  CA     : 701   Min.   :   0.52   Min.   :    1.8   Min.   :    0.68
##  TX     : 387   1st Qu.:   2.89   1st Qu.:  125.0   1st Qu.:   65.87
##  NY     : 311   Median :   6.46   Median :  198.7   Median :  112.51
##  VA     : 283   Mean   :  25.16   Mean   :  393.6   Mean   :  246.82
##  FL     : 282   3rd Qu.:  16.50   3rd Qu.:  375.0   3rd Qu.:  222.93
##  IL     : 273   Max.   :2936.88   Max.   :40740.0   Max.   :37837.66
##  (Other):2764                     NA's   :12        NA's   :12
```

```
nlevels(inc$Industry)
```

```
## [1] 25
```

```
nlevels(inc$City) #unique city names, double-counts commmon names (e.g,. Portland, Springfield)
```

```
## [1] 1425
```

```
nlevels(inc$State) #uniqure state names, 50 states + DC + PR
```

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

Recoding several variables as features offers a little more insight into which companies are growing. The industry with the most representation in the top 5000 is Information Technology, encompassing almost 15% of the firms. There are 25 different industries represented on the list. Likewise, 14% of the companies are located in one state: California. Beyond this, there is at least one fast-growing company in each state, including Washingotn DC and Puerto Rico.

For my engineered features, revenue_change appears to be skewed to the right, with mean 50% higher than median. Considering revenue and its change relative to number of employees provides more question than answers, as there appear to be some companies generating massive revenue with few employees.

```
inc[is.na(inc$Employees),]
```

```
##      Rank                              Name Growth_Rate Revenue
## 183   183           First Flight Solutions       22.32     2.7
## 1063 1064                         Popchips        3.98    93.3
## 1123 1124                       Vocalocity        3.72    42.9
## 1652 1653                     Higher Logic        2.36     6.0
## 1685 1686      Global Communications Group        2.30     3.6
## 2196 2197              JeffreyM Consulting        1.68    12.1
## 2742 2743               Excalibur Exhibits        1.27     9.9
## 3000 3001       Heartland Business Systems        1.12   156.3
## 3978 3978                             SSEC        0.68    80.4
## 4112 4112 Carolinas Home Medical Equipment        0.64     3.3
## 4566 4566                         Oakbrook        0.48     8.9
## 4968 4968                   Popcorn Palace        0.35     5.5
##                          Industry Employees          City State revenue_change
## 183    Logistics & Transportation        NA  EMERALD ISLE    NC           2.58
## 1063              Food & Beverage        NA SAN FRANCISCO    CA          74.57
## 1123           Telecommunications        NA       ATLANTA    GA          33.81
## 1652                     Software        NA    WASHINGTON    DC           4.21
## 1685           Telecommunications        NA     ENGLEWOOD    CO           2.51
## 2196 Business Products & Services        NA      BELLEVUE    WA           7.59
## 2742 Business Products & Services        NA       HOUSTON    TX           5.54
## 3000                  IT Services        NA  LITTLE CHUTE    WI          82.57
## 3978                Manufacturing        NA       HORSHAM    PA          32.54
## 4112                       Health        NA      MATTHEWS    NC           1.29
## 4566                  Real Estate        NA       MADISON    WI           2.89
## 4968              Food & Beverage        NA SCHILLER PARK    IL           1.43
##      revenue_employee revenue_change_employee
## 183                NA                     NA
## 1063               NA                     NA
## 1123               NA                     NA
## 1652               NA                     NA
## 1685               NA                     NA
## 2196               NA                     NA
## 2742               NA                     NA
## 3000               NA                     NA
## 3978               NA                     NA
## 4112               NA                     NA
## 4566               NA                     NA
## 4968               NA                     NA
```

```
summary(inc[inc$Employees < 24,])
```

```
##       Rank          Name              Growth_Rate         Revenue
##  Min.   :  10   Length:1153        Min.   :  0.340   Min.   :  2.000
##  1st Qu.: 925   Class :character   1st Qu.:  0.880   1st Qu.:  2.800
##  Median :2089   Mode  :character   Median :  1.780   Median :  4.300
##  Mean   :2244                      Mean   :  5.124   Mean   :  8.024
##  3rd Qu.:3479                      3rd Qu.:  4.720   3rd Qu.:  7.600
##  Max.   :4998                      Max.   :166.890   Max.   :303.000
##  NA's   :12                        NA's   :12        NA's   :12
##                          Industry     Employees               City
##  IT Services                 :141   Min.   : 1.00   NEW YORK     :  45
##  Advertising & Marketing     :136   1st Qu.:10.00   SAN DIEGO    :  20
##  Business Products & Services:111   Median :15.00   SAN FRANCISCO:  19
##  Retail                      : 87   Mean   :14.27   AUSTIN       :  18
##  Consumer Products & Services: 74   3rd Qu.:19.00   ATLANTA      :  16
##  (Other)                     :592   Max.   :23.00   (Other)      :1023
##  NA's                        : 12   NA's   :12      NA's         :  12
##      State     revenue_change    revenue_employee   revenue_change_employee
##  CA     :180   Min.   :  0.520   Min.   :   86.96   Min.   :   35.29
##  NY     : 90   1st Qu.:  1.680   1st Qu.:  205.00   1st Qu.:  117.82
##  FL     : 81   Median :  2.680   Median :  344.44   Median :  210.02
##  TX     : 76   Mean   :  5.344   Mean   :  681.48   Mean   :  458.68
##  OH     : 57   3rd Qu.:  4.730   3rd Qu.:  631.25   3rd Qu.:  421.93
##  (Other):657   Max.   :198.520   Max.   :40740.00   Max.   :23103.64
##  NA's   : 12   NA's   :12        NA's   :12         NA's   :12
```

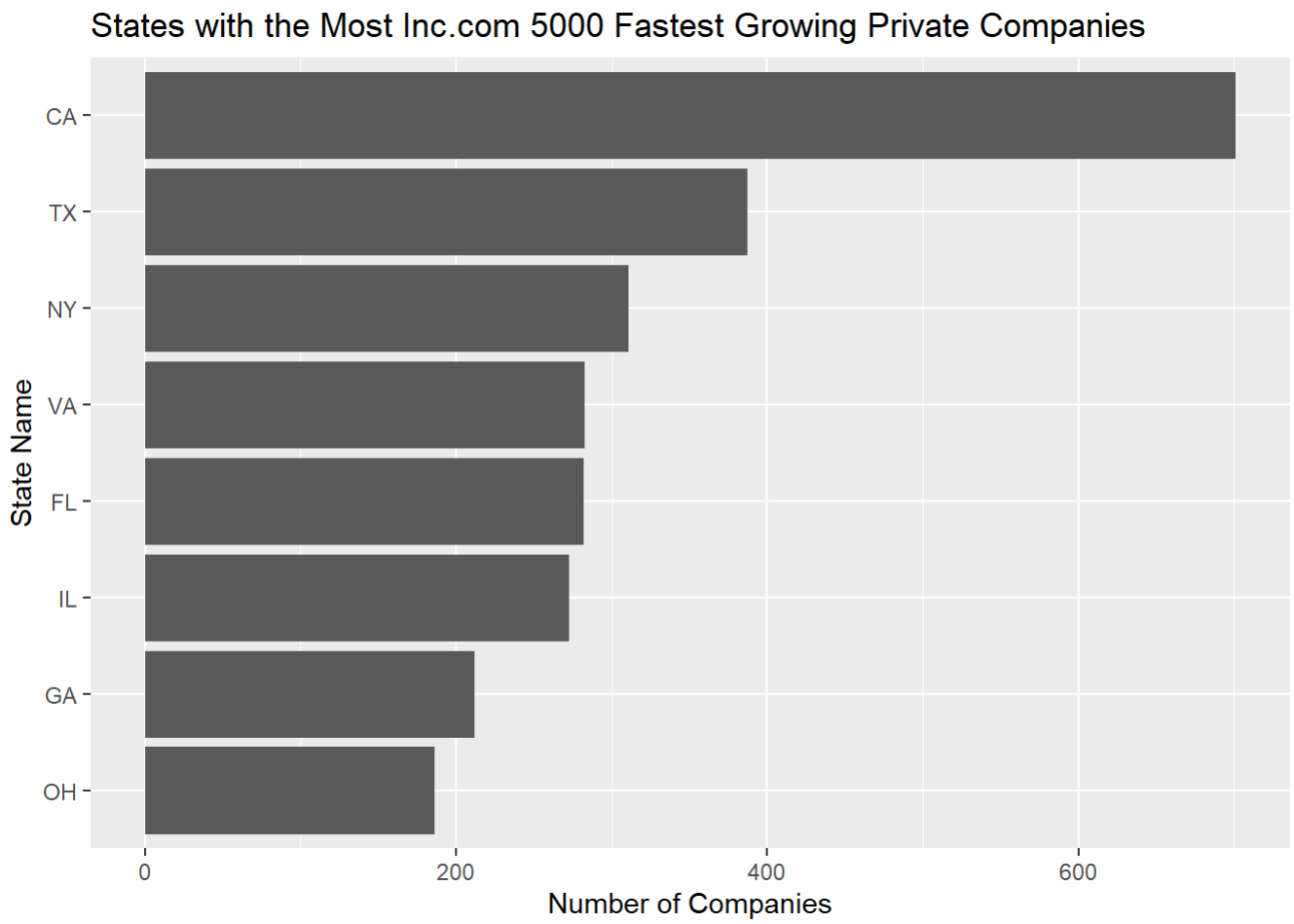
For missing values, no immediate patterns can be identified. Looking at the smallest 25% of companies, these companies appear to have lower revenue but conversely much higher revenue per employee. Absolute change in revenue is also lower than the top 5000 but higher on a per-emplmoyee basis. The bottom 25% also may represent more retail and fewer health-related companies.

Question 1

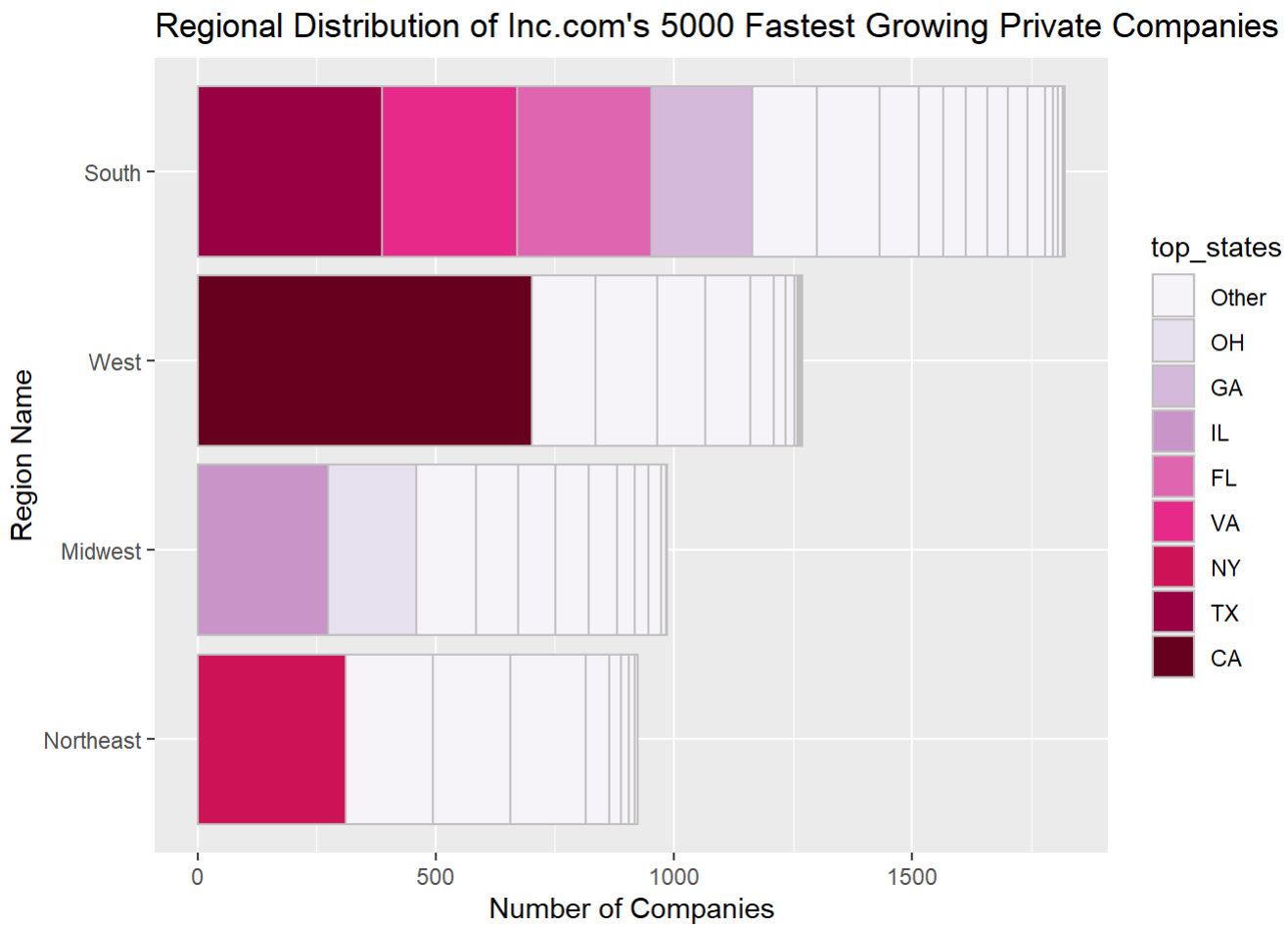
Create a graph that shows the distribution of companies in the data set by State (ie how many are in each state). There are a lot of States, so consider which axis you should use. This visualization is ultimately going to be consumed on a ‘portrait’ oriented screen (ie taller than wide), which should further guide your layout choices.

```
companies_by_state <- inc %>%
  group_by(State) %>%
  summarize(n_companies = n()) %>% #get number of companies per state
  arrange(desc(n_companies)) %>%
  mutate(State = fct_reorder(State,n_companies)) %>% #rearrange by state in descending order
  mutate(Region = recode_factor(State, VA = 'South', WV = 'South', AR = 'South', DE = 'South', D
C = 'South', FL = 'South', GA = 'South', MD = 'South', NC = 'South', SC = 'South', AL = 'South',
KY = 'South', MS = 'South', TN = 'South', AK = 'South', LA = 'South', OK = 'South' , TX = 'Sout
h', PR = 'South', #south
AZ = 'West', CO = 'West', ID = 'West', NM = 'West', MT = 'West', UT = 'West', NV = 'West', WY =
'West', AK = 'West', CA = 'West', HI = 'West', OR = 'West', WA = 'West',
IA = 'Midwest', IL = 'Midwest', MI = 'Midwest', OH = 'Midwest', WI = 'Midwest', IA = 'Midwest',
 KS = 'Midwest', MN = 'Midwest', MS = 'Midwest', NE = 'Midwest', SD = 'Midwest', ND = 'Midwest',
IN = 'Midwest', MO = 'Midwest',#midwest
CT = 'Northeast', ME = 'Northeast', MA = 'Northeast', NH = 'Northeast', RI = 'Northeast', VT =
'Northeast', NJ = 'Northeast', NY = 'Northeast', PA = 'Northeast' #northeast
)) %>%
  mutate(top_states = recode_factor(State, CA= 'CA', TX = 'TX' , NY = 'NY', VA = 'VA', FL = 'FL'
, IL = 'IL', GA = 'GA', OH = 'OH', .default = "Other")) %>%
  mutate(top_states = fct_reorder(top_states, n_companies))#rearrange by state in descending ord
er
```

```
ggplot(companies_by_state[companies_by_state$top_states != 'Other',], aes(top_states,n_companie
s)) + #omit states listed as 'other'
  geom_bar(stat = 'identity') +
  coord_flip() + #change axis
  labs(title = 'States with the Most Inc.com 5000 Fastest Growing Private Companies') +
  ylab(label = 'Number of Companies') +
  xlab(label = 'State Name')
```




```
ggplot(companies_by_state, aes(fill = top_states, Region,n_companies)) + #all data included
  geom_bar(stat = 'identity', color = 'grey') + #use grey to distinguish between states within t
he 'other' column
  scale_fill_brewer(palette='PuRd') +
  coord_flip()  + #change axis
  scale_x_discrete(limits = rev) + #largest regions on top, largest states closest to y-axis
  labs(title = 'Regional Distribution of Inc.com\'s 5000 Fastest Growing Private Companies', col
or = 'State Name') +
  ylab(label = 'Number of Companies') +
  xlab(label = 'Region Name')
```



To visualize companies by state in the most concise way possible, I provided two bar graphs. The first graph provides the top 8 states in terms of number of high-growht companies. This graph, complete with a coord flip to show bar graphs horizontally, only provides ~50% of the top 5,000 companies. This a large loss of information and context.

To compensate for this, the second graph also utilizes color as a third element of data to distinguish the top 8 states. These states are part of four larger regions of the United States, as defined by the Census Bureau: Northeast, South, Midwest, and West. While these regions are very diverse, this provides a way of comparing states with their geographical counterparts. To interpret the second graph, the largest comntributors are the most saturated, while the 'Other' column is white in each region. This allows for comparisons between geographical regions, as well as showing the relative contribution of a state to its region. For example, California makes up over half of the West's fastest growing companies. This second graph also illustrates an important takeaway that's not immediately obvious from the top cities and states: the South is not only the largest contributor to fastest growing compnaies, but also contains four of the top eight. I also added grey borders in each state to add an intuition about how many states are in each region - the South for instance has many states but only some of them contribute appreciably to the largest growth companies tally.

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.

```
companies_by_state %>%
  arrange(desc(n_companies)) %>%
  head(3)
```

```
## # A tibble: 3 x 4
##   State n_companies Region    top_states
##   <fct>       <int> <fct>     <fct>
## 1 CA            701 West      CA
## 2 TX            387 South     TX
## 3 NY            311 Northeast NY
```

```
# Answer Question 2 here
```

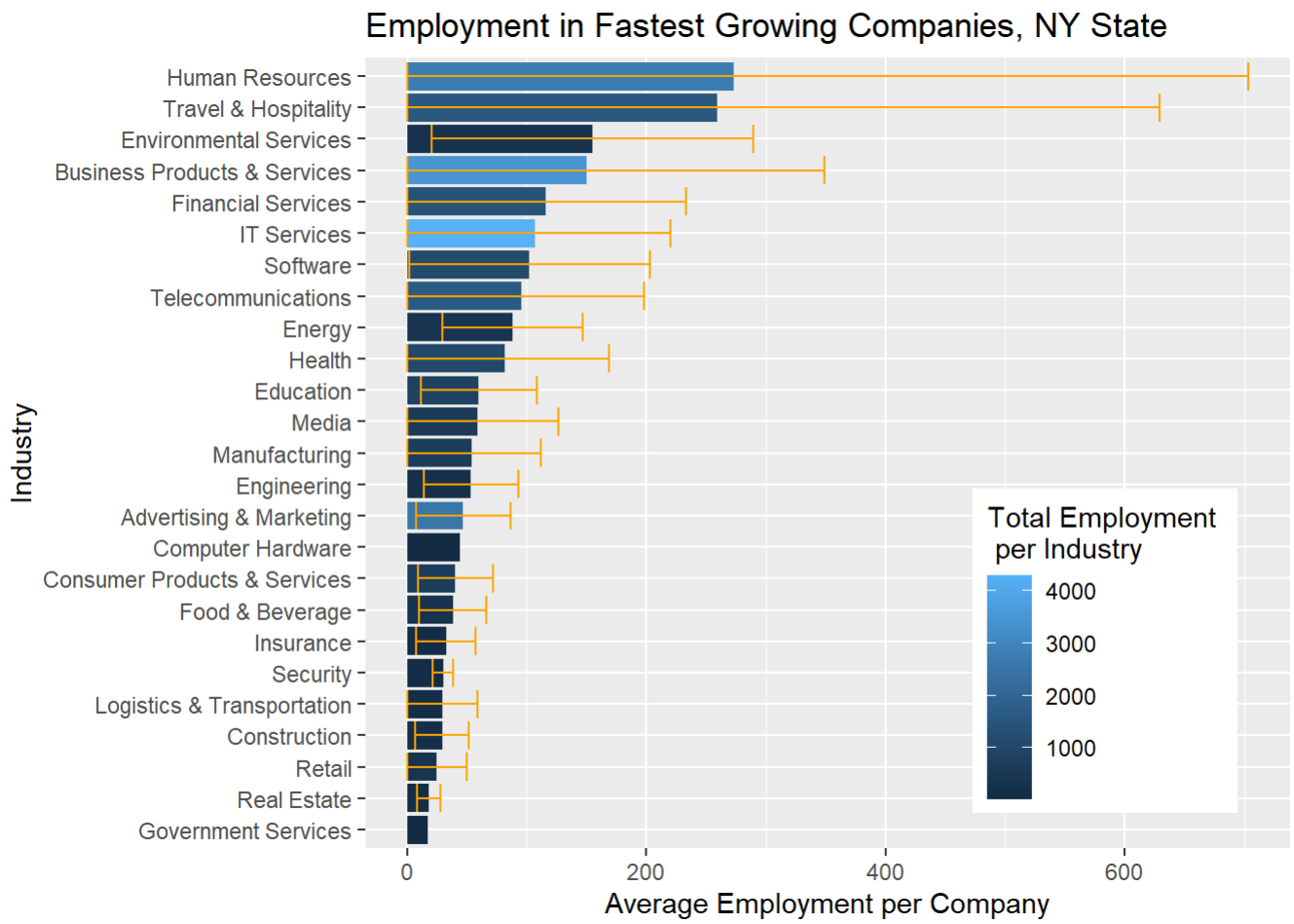
The state with the third largest companies is New York.

```
ny_inc <- inc[complete.cases(inc),] %>%
  filter(State == 'NY')

ny_industry <- ny_inc %>%
  group_by(Industry) %>%
  mutate(iqr = 1.5*(quantile(Employees,.75) - quantile(Employees,.25))) %>% #define each other's
IQR
  ungroup() %>% #ungroup to filter outliers
  filter(((Employees < quantile(Employees,0.75) + 1.5*iqr) & (Employees > quantile(Employees,.25
) - 1.5*iqr)) | iqr == 0)%>% #filter outliers using (Q1 + 1.5IQR, Q3 + 1.5IQR) criteria
  mutate(mean_employee = mean(Employees)) %>% #get mean for residual calculation
  mutate(sqr_resid = (Employees - mean_employee)^2)  %>% #get squared residuals for standard dev
iation while ungrouped
  group_by(Industry) %>% #regroup to generate industry statistics
  summarize(n_company = n(), n_employee = sum(Employees), avg_employee = mean(Employees), med_em
ployee = median(Employees), sd_employee = sd(Employees))  %>% #summarize industry statistics
  arrange(desc(avg_employee)) %>%
  mutate(Industry = fct_reorder(Industry, avg_employee)) #arrange so that largest average indust
ry is on top



ggplot(ny_industry, aes(fill=n_employee, Industry, avg_employee)) + #this graph will show each i
ndustry versus average employees, colored by total employees in the industry in NY state
  geom_bar(aes(Industry, avg_employee), stat = 'identity') + #bar length defined by average empl
oyee in industry
  geom_errorbar(aes(x = Industry, ymin=ifelse(avg_employee-sd_employee<0,0,avg_employee-sd_emplo
yee), ymax = avg_employee+sd_employee), color = 'orange') + #error bars +/- standard deviation
  coord_flip() + #show industry on y-axis
  labs(title = 'Employment in Fastest Growing Companies, NY State', fill = 'Total Employment \n
 per Industry') +
  ylab(label = 'Average Employment per Company') +
  theme(legend.position = c(0.8,0.25))
```



Payrolls are largest in the Human Resources and Travel industries. Surprisingly, real estate is not well representated. However, variance and thereby error bars are large, owing to vast differences in staffing at each company. This remains true even after removal of outliers. Finally, IT services is the largest industry represented in the fastest growing companies in NY state.

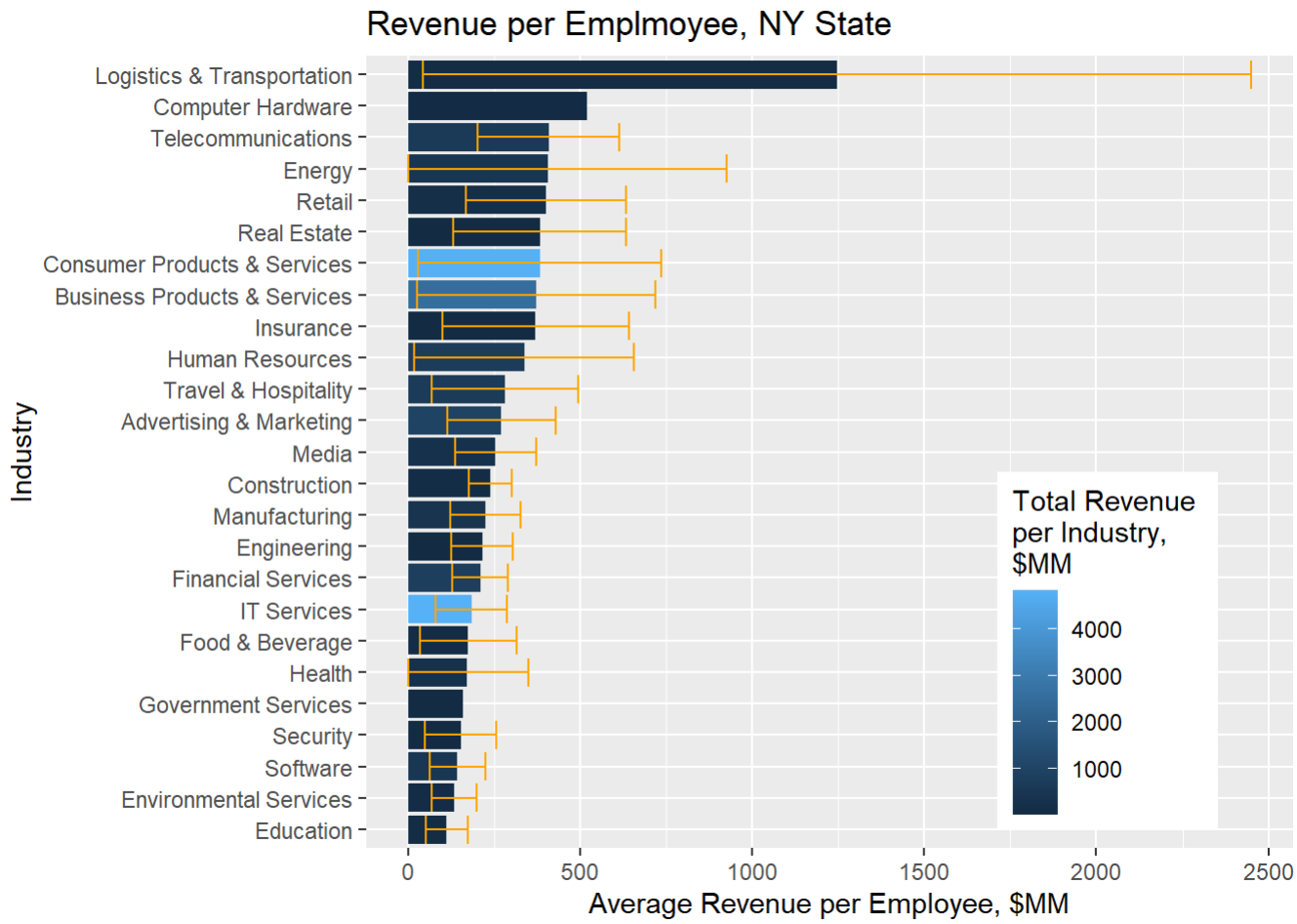
Question 3

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.

```
ny_revenue <- ny_inc %>%
  group_by(Industry) %>%
  summarize(total_revenue = sum(Revenue))

ny_emp_revenue <- ny_inc %>%
  group_by(Industry) %>%
  mutate(iqr = 1.5*(quantile(revenue_employee,.75) - quantile(revenue_employee,.25))) %>% #defin
e each Industry's IQR
  ungroup() %>% #ungroup to filter outliers
  filter(((revenue_employee < quantile(revenue_employee, 0.75) + 1.5*iqr) & (revenue_employee >
 quantile(revenue_employee, 0.25) - 1.5*iqr) )| (iqr == 0))%>% #filter outliers using (Q1 + 1.5I
QR, Q3 + 1.5IQR) criteria
  group_by(Industry) %>%
  mutate(mean_revenue_employee = mean(revenue_employee))  %>% #get mean for residual calculation
  summarize(n_company = n(), avg_revenue = mean(revenue_employee), med_revenue = median(revenue_
employee), sd_revenue = sd(revenue_employee))  %>% #summarize industry statistics
  cbind(ny_revenue$total_revenue) %>%
  mutate(Industry = fct_reorder(Industry, avg_revenue)) #arrange so that largest average industr
y is on top


ggplot(ny_emp_revenue, aes(fill=ny_revenue$total_revenue,Industry, avg_revenue)) + #this graph w
ill show each industry versus average employees, colored by total employees in the industry in N
Y state
  geom_bar(aes(Industry, avg_revenue), stat = 'identity') + #bar length defined by average emplo
yee in industry
  geom_errorbar(aes(x = Industry, ymin=ifelse(avg_revenue-sd_revenue <0, 0, avg_revenue-sd_reven
ue), ymax = avg_revenue+sd_revenue), color = 'orange') + #error bars +/- standard deviation
  coord_flip() + #show industry on y-axis
  labs(title = 'Revenue per Emplmoyee, NY State', fill = 'Total Revenue \nper Industry, \n$MM ')
+
  ylab(label = 'Average Revenue per Employee, $MM') +
  theme(legend.position = c(0.8,0.25))
```



When considering revenue per employee, the must lucrative industry is Logistics & Transportation, with all other industries trailing significantly. Even after removing one large outlier, a second smaller outlier remains. Variance for Logistics in particular is large owing to small payrolls in some of the most lucrative companies. While not evident in our dataset, a brief search on the wbe suggests that these compoanies do consulting work and act as a broker between shipping agencies and prospective customers. While not the most lucrative industries identified, Consumer Products and IT are the largest industries by revenue.