# DATA608-Module1-inc5000

Michael Y.

2/13/2020

## Contents



---

**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:

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

### Let's preview this data:

```
options(digits=7,scipen=999,width=120)
library(kableExtra)
head(inc) %>%
  kable(format.args = list(big.mark = ",")) %>%
  kable_styling(c("bordered","striped"))
```

| Rank | Name | Growth_Rate | Revenue | Industry | Employees | City | State |
|---|---|---|---|---|---|---|---|
| 1 | Fuhu | 421.48 | 117,900,000 | Consumer Products & Services | 104 | El Segundo | CA |
| 2 | FederalConference.com | 248.31 | 49,600,000 | Government Services | 51 | Dumfries | VA |
| 3 | The HCI Group | 245.45 | 25,500,000 | Health | 132 | Jacksonville | FL |
| 4 | Bridger | 233.08 | 1,900,000,000 | Energy | 50 | Addison | TX |
| 5 | DataXu | 213.37 | 87,000,000 | Advertising & Marketing | 220 | Boston | MA |
| 6 | MileStone Community Builders | 179.38 | 45,700,000 | Real Estate | 63 | Austin | TX |

```r
summary(inc)
```

```
##       Rank                           Name       Growth_Rate         Revenue
##  Min.   :   1   (Add)ventures         :   1   Min.   :  0.340   Min.   :    2000000
##  1st Qu.:1252   @Properties           :   1   1st Qu.:  0.770   1st Qu.:    5100000
##  Median :2502   1-Stop Translation USA:   1   Median :  1.420   Median :   10900000
##  Mean   :2502   110 Consulting        :   1   Mean   :  4.612   Mean   :   48222535
##  3rd Qu.:3751   11thStreetCoffee.com  :   1   3rd Qu.:  3.290   3rd Qu.:   28600000
##  Max.   :5000   123 Exteriors         :   1   Max.   :421.480   Max.   :10100000000
##                 (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:

```
# Insert your code here, create more chunks as necessary
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## <U+2713> ggplot2 3.2.1     <U+2713> purrr   0.3.3
## <U+2713> tibble  2.1.3     <U+2713> dplyr   0.8.3
## <U+2713> tidyr   1.0.0     <U+2713> stringr 1.4.0
## <U+2713> readr   1.3.1     <U+2713> forcats 0.4.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter()     masks stats::filter()
## x dplyr::group_rows() masks kableExtra::group_rows()
## x dplyr::lag()        masks stats::lag()
```

```
# Add a column for revenue per employee at each company
inc2 <- inc %>%
  mutate(RevenuePerEmployee = Revenue / Employees)

### Group by Industry
inc2 %>%
  group_by(Industry) -> inc_byIndustry

### Aggregate by Industry
inc_byIndustry %>% summarise(
  N         = n(),
  AvgGrowth = mean(Growth_Rate,na.rm=T),
  AvgRev    = mean(Revenue,na.rm=T),
  TotalRev  = sum(Revenue,na.rm=T),
  AvgEmpl   = mean(Employees,na.rm=T),
  TotalEmpl = sum(Employees,na.rm=T),
  AvgRevPerEmpl = mean(RevenuePerEmployee,na.rm=T),
  MedRevPerEmpl = median(RevenuePerEmployee,na.rm=T)
) -> Summary_byIndustry
```

```
### Group by state
inc2 %>%
  group_by(State) -> inc_byState

### Aggregate by state
inc_byState %>% summarise(
  Num_Companies         = n(),
  AvgGrowth = mean(Growth_Rate,na.rm=T),
  AvgRev    = mean(Revenue,na.rm=T),
  TotalRev  = sum(Revenue,na.rm=T),
  AvgEmpl   = mean(Employees,na.rm=T),
  TotalEmpl = sum(Employees,na.rm=T),
  AvgRevPerEmpl = mean(RevenuePerEmployee,na.rm=T),
  MedRevPerEmpl = median(RevenuePerEmployee,na.rm=T)
) -> Summary_byState
```

**Display summary by industry**

```
options(digits = 10)
Summary_byIndustry %>%
  kable(format.args = list(big.mark = ","),digits=2) %>%
  kable_styling(c("bordered","striped"))
```

| Industry | N | AvgGrowth | AvgRev | TotalRev | AvgEmpl | TotalEmpl | AvgRevPerEmpl | MedRevPerEmpl |
|---|---|---|---|---|---|---|---|---|
| Advertising & Marketing | 471 | 6.23 | 16,528,662.42 | 7,785,000,000 | 84.35 | 39,731 | 306,036.31 | 202,061.86 |
| Business Products & Services | 482 | 3.52 | 54,705,186.72 | 26,367,900,000 | 244.49 | 117,357 | 359,096.93 | 200,000.00 |
| Computer Hardware | 44 | 4.09 | 270,129,545.45 | 11,885,700,000 | 220.77 | 9,714 | 817,702.22 | 516,477.27 |
| Construction | 187 | 3.37 | 70,450,802.14 | 13,174,300,000 | 155.61 | 29,099 | 465,682.41 | 280,000.00 |
| Consumer Products & Services | 203 | 8.78 | 73,676,847.29 | 14,956,400,000 | 223.96 | 45,464 | 466,068.10 | 313,043.48 |
| Education | 83 | 3.64 | 13,726,506.02 | 1,139,300,000 | 92.59 | 7,685 | 296,453.51 | 166,666.67 |
| Energy | 109 | 9.60 | 126,344,954.13 | 13,771,600,000 | 242.54 | 26,437 | 1,554,655.82 | 283,211.68 |
| Engineering | 74 | 1.98 | 34,222,972.97 | 2,532,500,000 | 276.15 | 20,435 | 201,119.95 | 164,840.86 |
| Environmental Services | 51 | 2.07 | 51,741,176.47 | 2,638,800,000 | 199.12 | 10,155 | 283,607.34 | 179,723.50 |
| Financial Services | 260 | 5.44 | 50,580,384.62 | 13,150,900,000 | 183.43 | 47,693 | 394,230.87 | 214,858.48 |
| Food & Beverage | 131 | 3.64 | 98,559,541.98 | 12,911,300,000 | 510.94 | 65,911 | 618,382.85 | 231,372.55 |
| Government Services | 202 | 7.24 | 29,748,019.80 | 6,009,100,000 | 129.63 | 26,185 | 243,596.00 | 165,323.89 |
| Health | 355 | 4.86 | 50,319,436.62 | 17,863,400,000 | 232.85 | 82,430 | 325,198.89 | 174,166.67 |
| Human Resources | 196 | 3.30 | 47,173,979.59 | 9,246,100,000 | 1,158.06 | 226,980 | 395,972.25 | 149,547.51 |
| Insurance | 50 | 2.01 | 46,758,000.00 | 2,337,900,000 | 146.78 | 7,339 | 474,966.36 | 225,622.97 |
| IT Services | 733 | 3.33 | 28,214,597.54 | 20,681,300,000 | 140.42 | 102,788 | 270,494.21 | 163,914.89 |
| Logistics & Transportation | 155 | 4.34 | 95,745,161.29 | 14,840,500,000 | 259.70 | 39,994 | 794,810.91 | 425,024.15 |
| Manufacturing | 256 | 2.30 | 49,546,875.00 | 12,684,000,000 | 172.32 | 43,942 | 453,524.01 | 231,250.00 |
| Media | 54 | 4.37 | 32,266,666.67 | 1,742,400,000 | 176.52 | 9,532 | 307,143.79 | 261,458.33 |
| Real Estate | 96 | 7.75 | 30,892,708.33 | 2,965,700,000 | 198.87 | 18,893 | 434,515.57 | 253,571.43 |
| Retail | 203 | 6.18 | 50,529,064.04 | 10,257,400,000 | 182.60 | 37,068 | 412,554.86 | 312,755.10 |
| Security | 73 | 3.39 | 52,230,136.99 | 3,812,800,000 | 562.45 | 41,059 | 283,391.38 | 158,744.39 |
| Software | 342 | 5.02 | 23,802,923.98 | 8,140,600,000 | 150.33 | 51,262 | 225,989.25 | 155,319.15 |
| Telecommunications | 129 | 2.88 | 56,855,813.95 | 7,334,400,000 | 242.85 | 30,842 | 449,259.60 | 284,000.00 |
| Travel & Hospitality | 62 | 2.35 | 47,283,870.97 | 2,931,600,000 | 371.53 | 23,035 | 414,788.11 | 224,404.76 |

**Display summary by state**

```
Summary_byState[1:26,] %>%
  kable(format.args = list(big.mark = ","),digits=2) %>%
  kable_styling(c("bordered","striped"))
```

| State | Num_Companies | AvgGrowth | AvgRev | TotalRev | AvgEmpl | TotalEmpl | AvgRevPerEmpl | MedRevPerEmpl |
|---|---|---|---|---|---|---|---|---|
| AK | 2 | 4.80 | 171,500,000.00 | 343,000,000 | 1,264.00 | 2,528 | 154,669.98 | 154,669.98 |
| AL | 51 | 2.41 | 25,907,843.14 | 1,321,300,000 | 125.35 | 6,393 | 249,273.67 | 194,059.41 |
| AR | 9 | 1.67 | 8,333,333.33 | 75,000,000 | 55.11 | 496 | 180,752.59 | 154,761.90 |
| AZ | 100 | 4.62 | 55,015,000.00 | 5,501,500,000 | 342.81 | 34,281 | 314,132.28 | 190,814.85 |
| CA | 701 | 5.90 | 33,463,480.74 | 23,457,900,000 | 230.31 | 161,219 | 407,764.93 | 222,967.03 |
| CO | 134 | 4.95 | 31,270,149.25 | 4,190,200,000 | 198.78 | 26,438 | 418,053.48 | 176,086.96 |
| CT | 50 | 4.99 | 49,486,000.00 | 2,474,300,000 | 139.78 | 6,989 | 595,983.46 | 218,750.00 |
| DC | 43 | 8.30 | 76,344,186.05 | 3,282,800,000 | 219.55 | 9,221 | 309,673.37 | 195,652.17 |
| DE | 16 | 2.42 | 42,300,000.00 | 676,800,000 | 4,284.00 | 68,544 | 261,811.20 | 98,250.08 |
| FL | 282 | 5.85 | 37,625,177.30 | 10,610,300,000 | 217.10 | 61,221 | 409,142.84 | 191,153.85 |
| GA | 212 | 3.52 | 30,510,849.06 | 6,468,300,000 | 163.73 | 34,546 | 417,074.41 | 216,666.67 |
| HI | 7 | 6.79 | 99,485,714.29 | 696,400,000 | 88.71 | 621 | 1,001,869.05 | 366,428.57 |
| IA | 28 | 1.76 | 123,142,857.14 | 3,448,000,000 | 405.14 | 11,344 | 337,214.61 | 218,436.29 |
| ID | 17 | 2.65 | 231,523,529.41 | 3,935,900,000 | 342.18 | 5,817 | 675,024.18 | 271,428.57 |
| IL | 273 | 3.74 | 121,773,992.67 | 33,244,300,000 | 379.65 | 103,266 | 462,003.12 | 216,447.09 |
| IN | 69 | 4.79 | 50,105,797.10 | 3,457,300,000 | 184.01 | 12,697 | 281,487.24 | 164,516.13 |
| KS | 38 | 3.63 | 40,752,631.58 | 1,548,600,000 | 229.61 | 8,725 | 480,676.68 | 167,715.73 |
| KY | 40 | 2.06 | 33,232,500.00 | 1,329,300,000 | 138.60 | 5,544 | 435,194.31 | 198,015.87 |
| LA | 37 | 1.94 | 56,648,648.65 | 2,096,000,000 | 288.35 | 10,669 | 271,525.50 | 166,666.67 |
| MA | 182 | 5.42 | 33,174,725.27 | 6,037,800,000 | 135.62 | 24,682 | 336,959.11 | 200,000.00 |
| MD | 131 | 4.98 | 25,193,893.13 | 3,300,400,000 | 308.69 | 40,439 | 269,938.18 | 183,428.57 |
| ME | 13 | 16.21 | 12,476,923.08 | 162,200,000 | 67.62 | 879 | 293,313.56 | 157,777.78 |
| MI | 126 | 2.24 | 61,950,793.65 | 7,805,800,000 | 292.90 | 36,905 | 294,574.44 | 153,887.46 |
| MN | 88 | 3.82 | 57,256,818.18 | 5,038,600,000 | 210.61 | 18,534 | 510,810.49 | 186,144.07 |
| MO | 59 | 2.50 | 45,164,406.78 | 2,664,700,000 | 293.15 | 17,296 | 408,161.70 | 265,217.39 |
| MS | 12 | 5.64 | 43,766,666.67 | 525,200,000 | 460.92 | 5,531 | 611,781.75 | 288,038.28 |

```
Summary_byState[27:52,] %>%
  kable(format.args = list(big.mark = ","),digits=2) %>%
  kable_styling(c("bordered","striped"))
```

| State | Num_Companies | AvgGrowth | AvgRev | TotalRev | AvgEmpl | TotalEmpl | AvgRevPerEmpl | MedRevPerEmpl |
|---|---|---|---|---|---|---|---|---|
| MT | 4 | 0.76 | 6,150,000.00 | 24,600,000 | 418.25 | 1,673 | 264,333.47 | 246,187.36 |
| NC | 137 | 3.51 | 67,580,291.97 | 9,258,500,000 | 271.74 | 36,685 | 307,454.39 | 181,382.98 |
| ND | 10 | 1.23 | 18,240,000.00 | 182,400,000 | 96.30 | 963 | 230,156.51 | 245,285.17 |
| NE | 27 | 2.08 | 40,696,296.30 | 1,098,800,000 | 141.59 | 3,823 | 382,766.77 | 203,076.92 |
| NH | 24 | 1.51 | 41,691,666.67 | 1,000,600,000 | 120.42 | 2,890 | 427,787.67 | 310,344.83 |
| NJ | 158 | 4.45 | 29,574,683.54 | 4,672,800,000 | 190.90 | 30,162 | 357,740.83 | 179,261.36 |
| NM | 5 | 1.36 | 9,640,000.00 | 48,200,000 | 123.40 | 617 | 133,085.46 | 135,000.00 |
| NV | 26 | 2.33 | 19,915,384.62 | 517,800,000 | 66.35 | 1,725 | 418,628.12 | 213,116.88 |
| NY | 311 | 4.37 | 58,715,112.54 | 18,260,400,000 | 271.29 | 84,370 | 495,851.34 | 200,000.00 |
| OH | 186 | 3.56 | 68,745,161.29 | 12,786,600,000 | 204.31 | 38,002 | 490,292.92 | 230,958.39 |
| OK | 46 | 3.10 | 41,015,217.39 | 1,886,700,000 | 151.65 | 6,976 | 321,872.32 | 256,220.10 |
| OR | 49 | 3.15 | 28,589,795.92 | 1,400,900,000 | 89.78 | 4,399 | 355,951.88 | 192,307.69 |
| PA | 164 | 2.57 | 34,568,902.44 | 5,669,300,000 | 186.45 | 30,392 | 285,522.91 | 195,918.37 |
| PR | 1 | 1.73 | 2,300,000.00 | 2,300,000 | 29.00 | 29 | 79,310.34 | 79,310.34 |
| RI | 16 | 16.03 | 46,981,250.00 | 751,700,000 | 185.25 | 2,964 | 337,965.65 | 183,571.43 |
| SC | 48 | 6.06 | 30,993,750.00 | 1,487,700,000 | 111.42 | 5,348 | 317,050.73 | 222,162.16 |
| SD | 3 | 1.41 | 5,900,000.00 | 17,700,000 | 253.67 | 761 | 72,534.37 | 67,647.06 |
| TN | 82 | 4.95 | 35,870,731.71 | 2,941,400,000 | 177.88 | 14,586 | 458,637.47 | 190,109.88 |
| TX | 387 | 6.02 | 57,271,834.63 | 22,164,200,000 | 235.14 | 90,765 | 499,383.92 | 196,168.41 |
| UT | 95 | 6.31 | 36,038,947.37 | 3,423,700,000 | 200.29 | 19,028 | 356,986.20 | 172,195.12 |
| VA | 283 | 4.88 | 30,627,915.19 | 8,667,700,000 | 126.03 | 35,667 | 298,791.42 | 174,545.45 |
| VT | 6 | 1.30 | 46,200,000.00 | 277,200,000 | 178.17 | 1,069 | 275,514.23 | 237,422.71 |
| WA | 130 | 4.00 | 27,156,923.08 | 3,530,400,000 | 135.01 | 17,416 | 255,218.89 | 181,374.72 |
| WI | 79 | 2.69 | 92,362,025.32 | 7,296,600,000 | 201.92 | 15,548 | 390,960.27 | 202,127.66 |
| WV | 2 | 0.62 | 15,650,000.00 | 31,300,000 | 120.00 | 240 | 109,674.55 | 109,674.55 |
| WY | 2 | 19.14 | 34,750,000.00 | 69,500,000 | 53.50 | 107 | 568,315.02 | 568,315.02 |

## Question 1

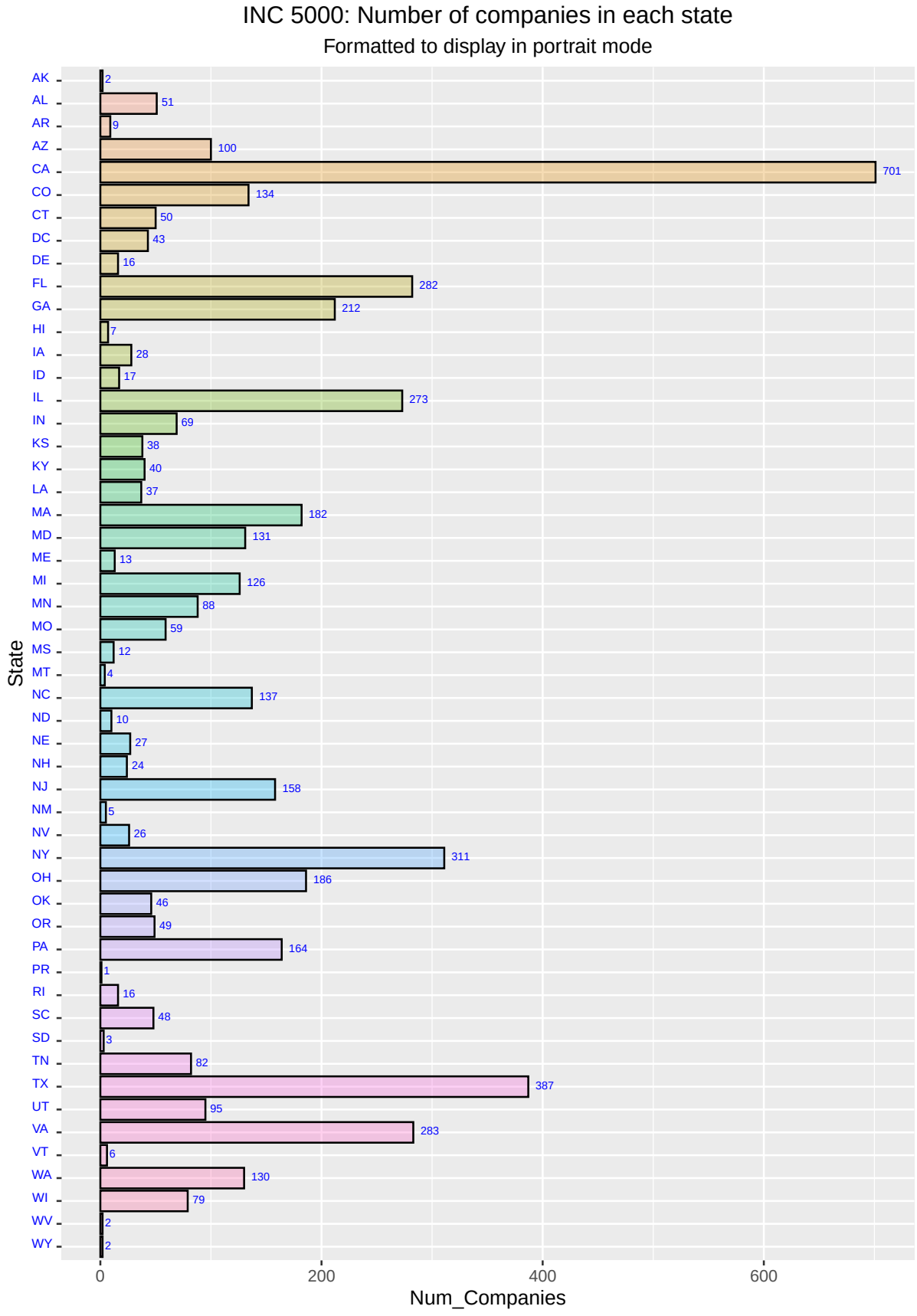
Create a graph that shows the distribution of companies in the dataset by State (ie how many are in each state). There are a lot of States, so consider which axis you should use. This visualization is ultimately going to be consumed on a 'portrait' oriented screen (ie taller than wide), which should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

```
# Answer Question 1 here

## Select just the State name and the number of companies
tempgrid <- Summary_byState  %>% select (State,Num_Companies)

## Reverse the grid so the results will display alphabetically (rather than backwards)
tempgrid <- tempgrid[rev(tempgrid$State),]


ggplot(tempgrid, aes(x=State, y=Num_Companies, fill=State))+
geom_col(position=position_dodge(+.5),color="black", alpha=0.3) +
geom_text( aes(label=Num_Companies), hjust=-0.4, vjust=+0.3, color="blue",
   size=2) +
ggtitle(label="INC 5000: Number of companies in each state",
        subtitle="Formatted to display in portrait mode") +
scale_x_discrete(limits = rev(levels(tempgrid$State)))+
theme(axis.text.y = element_text(angle = 0,
                                 hjust = +0.1, vjust=+0.1,
                                 size=7,color="blue"))+
theme(plot.title = element_text(hjust = 0.5))+
theme(plot.subtitle = element_text(hjust = 0.5))+
theme(legend.position="none")+
coord_flip()
```


INC 5000: Number of companies in each state
Formatted to display in portrait mode
State
AK 2
AL 51
AR 9
AZ 100
CA 701
CO 134
CT 50
DC 43
DE 16
FL 282
GA 212
HI 7
IA 28
ID 17
IL 273
IN 69
KS 38
KY 40
LA 37
MA 182
MD 131
ME 13
MI 126
MN 88
MO 59
MS 12
MT 4
NC 137
ND 10
NE 27
NH 24
NJ 158
NM 5
NV 26
NY 311
OH 186
OK 46
OR 49
PA 164
PR 1
RI 16
SC 48
SD 3
TN 82
TX 387
UT 95
VA 283
VT 6
WA 130
WI 79
WV 2
WY 2
0
200
400
600
Num_Companies

## Question 2

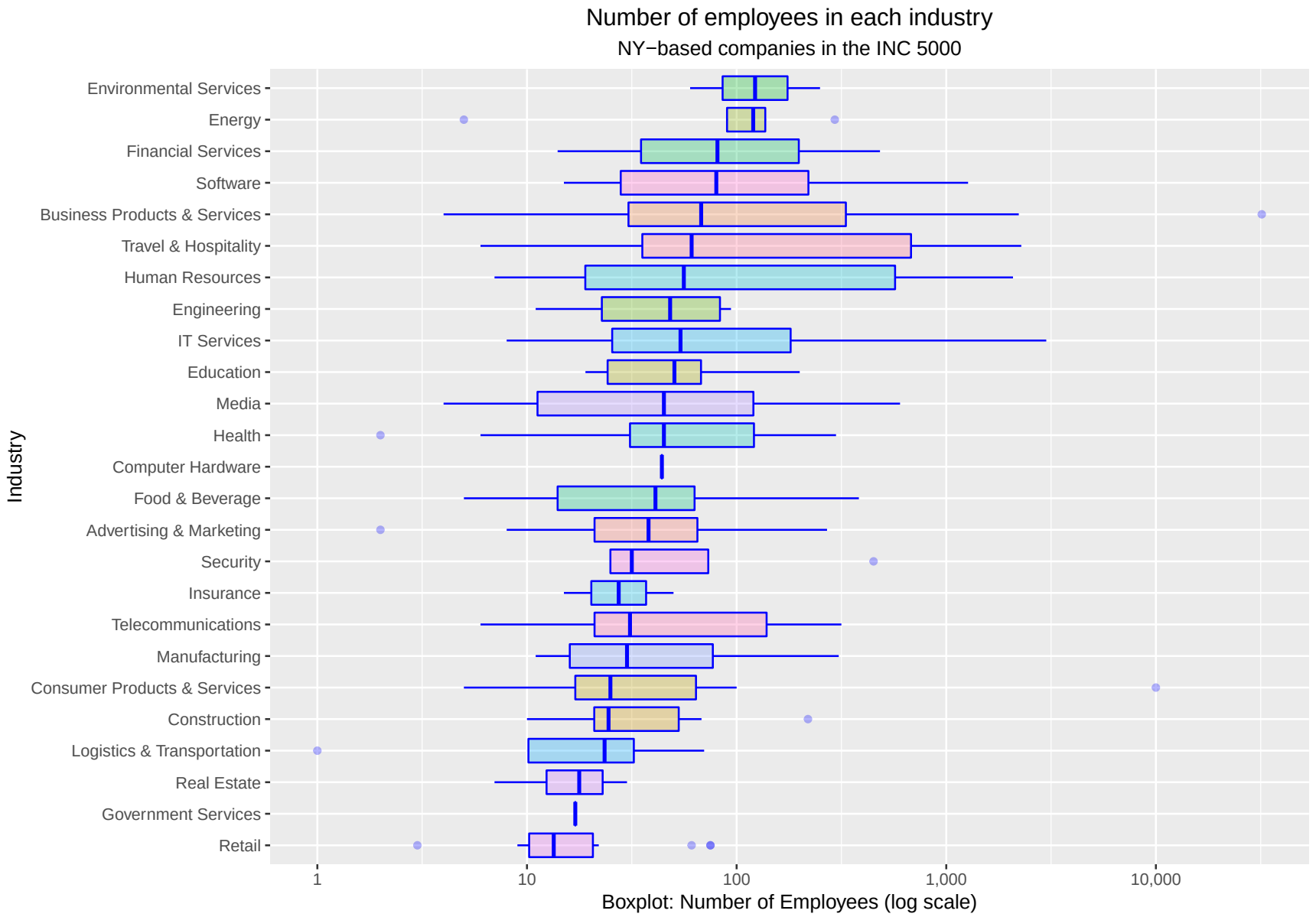
Lets dig in on the state with the 3rd most companies in the data set. Imagine you work for the state and are interested in how many people are employed by companies in different industries. Create a plot that shows the average and/or median employment by industry for companies in this state (only use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

### Determine which state has the 3rd most companies
whichRow <- order(Summary_byState$Num_Companies,decreasing = T)[3]
whichState <- as.character(Summary_byState$State)[whichRow]
numCompanies  <- as.integer(Summary_byState[whichRow,"Num_Companies"])
print(paste0("The state with the 3rd most companies is ",
             whichState, " with ", numCompanies, " companies."))
```

```
## [1] "The state with the 3rd most companies is NY with 311 companies."
```

```
### Subset the dataset with companies just from that state
inc2 %>% filter(State==whichState) -> inc3
### Filter out any cases with NA values  [ for this example, there are no such cases]
inc3 <- inc3[complete.cases(inc3),]
ggplot(inc3, aes(x = reorder(x=Industry,
                             X=Employees,
                             FUN = median),
                 y = Employees,
                 fill=Industry)) +
  geom_boxplot(color="blue", alpha=0.3) +
  scale_y_log10(label=scales::comma_format(accuracy = 1)) +
  labs(x="Industry",
       y="Boxplot: Number of Employees (log scale)",
       title = "Number of employees in each industry",
       subtitle = "NY-based companies in the INC 5000")+
  theme(plot.title = element_text(hjust = 0.5))+
  theme(plot.subtitle = element_text(hjust = 0.5))+
  theme(legend.position="none")+
  coord_flip()
```


Number of employees in each industry
NY−based companies in the INC 5000
Industry
Environmental Services
Energy
Financial Services
Software
Business Products & Services
Travel & Hospitality
Human Resources
Engineering
IT Services
Education
Media
Health
Computer Hardware
Food & Beverage
Advertising & Marketing
Security
Insurance
Telecommunications
Manufacturing
Consumer Products & Services
Construction
Logistics & Transportation
Real Estate
Government Services
Retail
1
10
100
1,000
10,000
Boxplot: Number of Employees (log scale)

## 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.

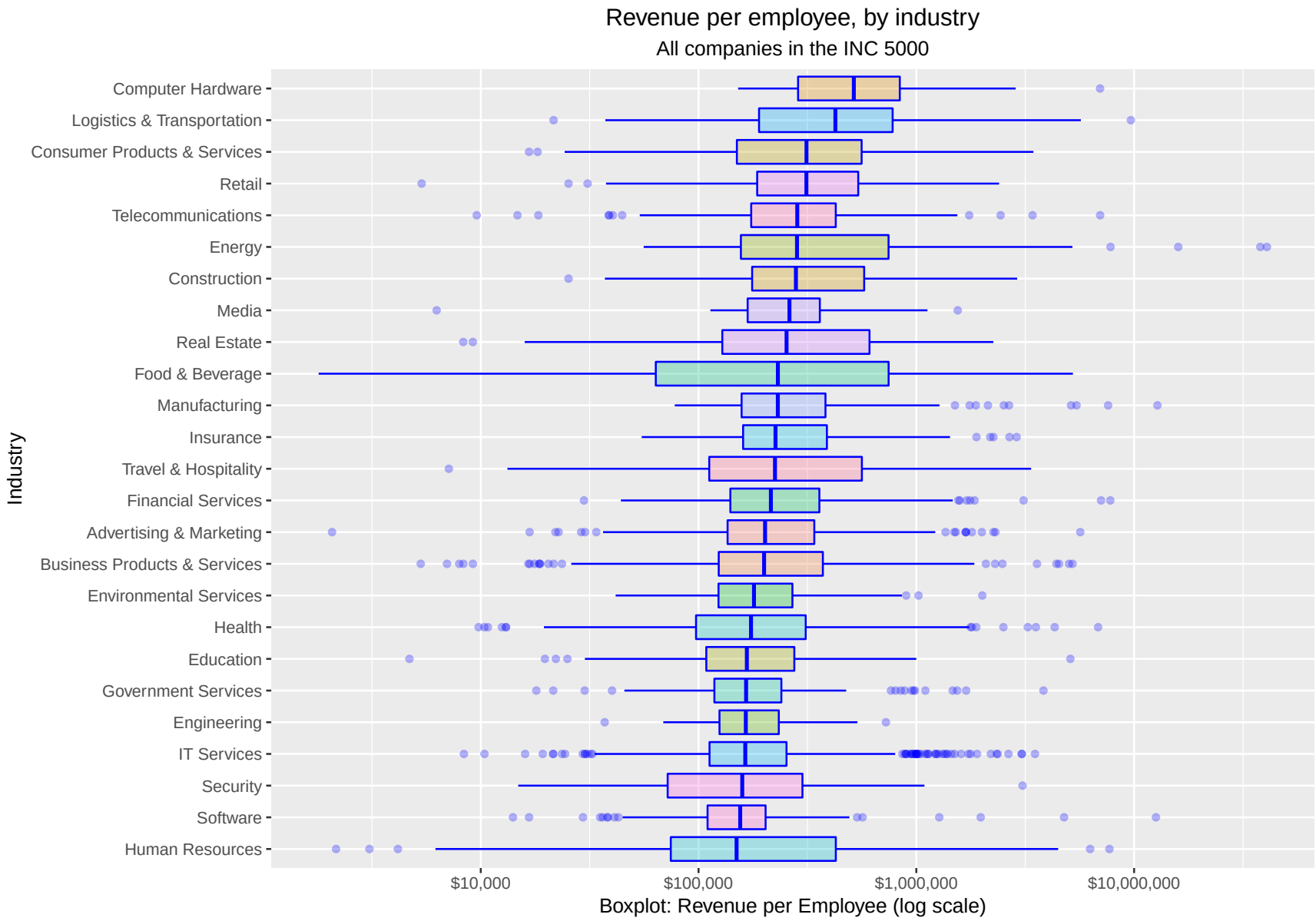
```
# Answer Question 3 here

# Add a column for revenue per employee at each company
inc4 <- inc %>% mutate(RevenuePerEmployee = Revenue / Employees)

# There are 12 companies for which the number of employees is unknown.
# Drop them.
inc4 <- inc4[complete.cases(inc4),]



ggplot(inc4, aes(x = reorder(x=Industry,
                             X=RevenuePerEmployee,
                             FUN = median),
                 y = RevenuePerEmployee,
                 fill=Industry)) +
  geom_boxplot(color="blue", alpha=0.3) +
  scale_y_log10(label=scales::dollar_format(accuracy = 1)) +
  labs(x="Industry",
       y="Boxplot: Revenue per Employee (log scale)",
       title = "Revenue per employee, by industry",
       subtitle = "All companies in the INC 5000")+
  theme(plot.title = element_text(hjust = 0.5))+
  theme(plot.subtitle = element_text(hjust = 0.5))+
  theme(legend.position="none")+
  coord_flip()
```


Revenue per employee, by industry
All companies in the INC 5000
Industry
Computer Hardware
Logistics & Transportation
Consumer Products & Services
Retail
Telecommunications
Energy
Construction
Media
Real Estate
Food & Beverage
Manufacturing
Insurance
Travel & Hospitality
Financial Services
Advertising & Marketing
Business Products & Services
Environmental Services
Health
Education
Government Services
Engineering
IT Services
Security
Software
Human Resources
$10,000
$100,000
$1,000,000
$10,000,000
Boxplot: Revenue per Employee (log scale)