HMW 1- Data 608

OMER OZEREN

Table of Contents

### Principles of Data Visualization and Introduction to ggplot2

# Required libraries  
library(dplyr)  
library(ggplot2)  
library(scales)  
# Turn off scientific notation  
options(scipen=999)

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(paste0("https://raw.githubusercontent.com/charleyferrari/CUNY\_DATA\_608/",  
 "master/module1/Data/inc5000\_data.csv"), header= TRUE)

And lets preview this data:

knitr::kable(head(inc))

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Rank | Name | Growth\_Rate | Revenue | Industry | Employees | City | State |
| 1 | Fuhu | 421.48 | 117900000 | Consumer Products & Services | 104 | El Segundo | CA |
| 2 | FederalConference.com | 248.31 | 49600000 | Government Services | 51 | Dumfries | VA |
| 3 | The HCI Group | 245.45 | 25500000 | Health | 132 | Jacksonville | FL |
| 4 | Bridger | 233.08 | 1900000000 | Energy | 50 | Addison | TX |
| 5 | DataXu | 213.37 | 87000000 | Advertising & Marketing | 220 | Boston | MA |
| 6 | MileStone Community Builders | 179.38 | 45700000 | 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   
## Min. : 2000000 IT Services : 733   
## 1st Qu.: 5100000 Business Products & Services: 482   
## Median : 10900000 Advertising & Marketing : 471   
## Mean : 48222535 Health : 355   
## 3rd Qu.: 28600000 Software : 342   
## Max. :10100000000 Financial Services : 260   
## (Other) :2358   
## Employees City State   
## Min. : 1.0 New York : 160 CA : 701   
## 1st Qu.: 25.0 Chicago : 90 TX : 387   
## Median : 53.0 Austin : 88 NY : 311   
## Mean : 232.7 Houston : 76 VA : 283   
## 3rd Qu.: 132.0 San Francisco: 75 FL : 282   
## Max. :66803.0 Atlanta : 74 IL : 273   
## 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.

Maximum number of employees seems high. Consider top ten companies based on employee count.

knitr::kable(head(inc[order(-inc$Employees),c(2,5:8)],10))

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Name | Industry | Employees | City | State |
| 2344 | Integrity staffing Solutions | Human Resources | 66803 | Wilmington | DE |
| 4577 | Sutherland Global Services | Business Products & Services | 32000 | Pittsford | NY |
| 1868 | Universal Services of America | Security | 20000 | Santa Ana | CA |
| 3456 | The Seaton Companies | Human Resources | 18887 | Chicago | IL |
| 2870 | PrideStaff | Human Resources | 17057 | Fresno | CA |
| 2313 | Infiniti HR | Human Resources | 17000 | Olney | MD |
| 4655 | CareersUSA | Human Resources | 14451 | Boca Raton | FL |
| 1487 | Sprouts Farmers Market | Consumer Products & Services | 13200 | Phoenix | AZ |
| 4140 | Cornerstone Staffing Solutions | Human Resources | 13071 | Pleasanton | CA |
| 3650 | Genco | Logistics & Transportation | 10800 | Pittsburgh | PA |

Check for unique company names (that companies are not duplicated in the data).

dupNames <- group\_by(inc, Name) %>%  
 summarize(Count=n()) %>%  
 filter(Count>1)  
cat("Number of duplicate company names:",nrow(dupNames))  
## Number of duplicate company names: 0

Consider all industries.

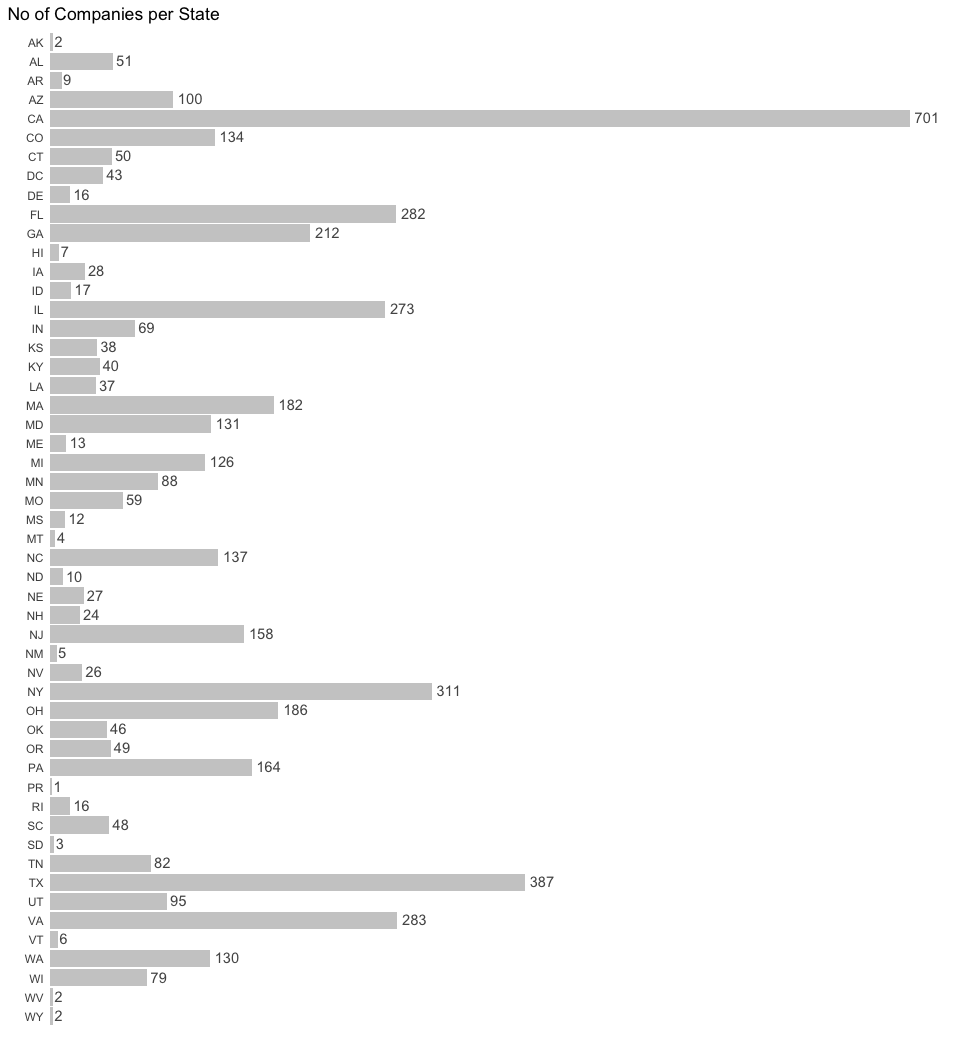
knitr::kable(group\_by(inc, Industry) %>% summarize(Count=n()) %>% arrange(desc(Count)))

|  |  |
| --- | --- |
| Industry | Count |
| IT Services | 733 |
| Business Products & Services | 482 |
| Advertising & Marketing | 471 |
| Health | 355 |
| Software | 342 |
| Financial Services | 260 |
| Manufacturing | 256 |
| Consumer Products & Services | 203 |
| Retail | 203 |
| Government Services | 202 |
| Human Resources | 196 |
| Construction | 187 |
| Logistics & Transportation | 155 |
| Food & Beverage | 131 |
| Telecommunications | 129 |
| Energy | 109 |
| Real Estate | 96 |
| Education | 83 |
| Engineering | 74 |
| Security | 73 |
| Travel & Hospitality | 62 |
| Media | 54 |
| Environmental Services | 51 |
| Insurance | 50 |
| Computer Hardware | 44 |

## 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.

# Get a list of counts by state  
stateCount <- group\_by(inc, State) %>%  
 summarize(Count=n())  
# Plot results  
ggplot(data = stateCount, aes(x = State, y = Count)) +   
 geom\_bar(stat="identity", fill="#CCCCCC") +   
 geom\_text(aes(label=Count), hjust=-0.2, vjust=0.4, color="#555555") +  
 scale\_x\_discrete(limits = rev(levels(stateCount$State))) +  
 coord\_flip() +   
 ggtitle("No of Companies per State") + labs(x = "", y = "") +  
 theme(panel.background = element\_blank(),  
 axis.ticks = element\_blank(),  
 axis.text.x = element\_blank(),  
 axis.text.y = element\_text(margin = margin(r=-30)))



Decision points:

* The plot is sorted by state (rather than by count) since viewers may be interested in a particular state (their home state?) and it is easier to find a state this way. Additionally, viewers are generally used to seeing US states listed in alphabetical order.
* It may be interesting to know exact values, so they are added to corresponding bars.
* With values, gridlines are redundant.
* Background is not necessary.
* Tick marks are not necessary.
* Default color was too dark.
* Various color themes were considered - gradient from highest to lowest count, highlighting top 3, 5 or 10 states, etc. This was deemed unnecessary.

## 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.

# Top 3 states  
knitr::kable(arrange(stateCount, desc(Count)) %>% top\_n(3))

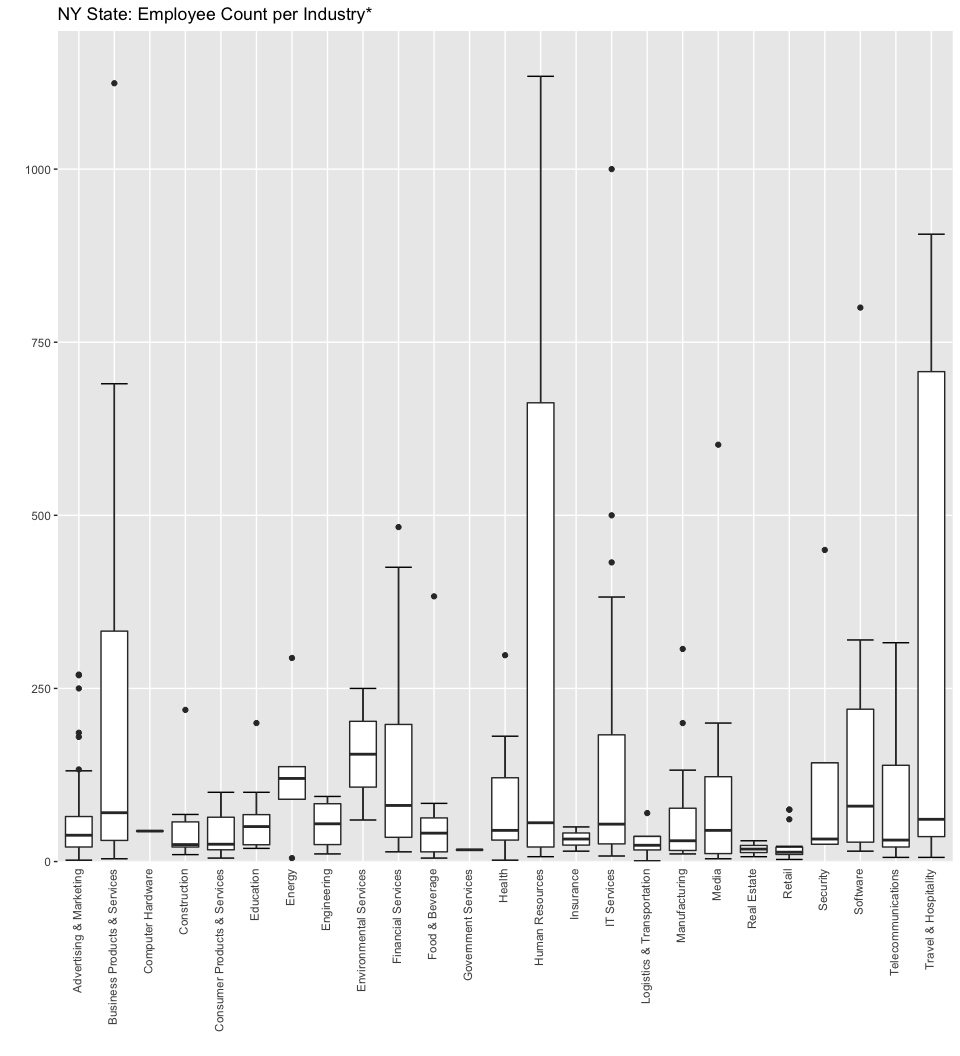
|  |  |
| --- | --- |
| State | Count |
| CA | 701 |
| TX | 387 |
| NY | 311 |

# Get NY industry employee counts  
nyInd <- filter(inc, State=="NY") %>%  
 select(Industry, Name, Employees)  
# Check if any NAs  
cat("Number of NAs:",sum(is.na(nyInd$Employees)))  
## Number of NAs: 0

There are no NAs for NY data. complete.cases() is not necessary.

Rather than discarding a few large outliers, which skew averages, below plots display **median** values.

# Plot  
ggplot(aes(x=Industry, y=Employees), data = nyInd) +   
 stat\_boxplot(geom ='errorbar') +  
 geom\_boxplot() +   
 coord\_cartesian(ylim = c(0,1200)) +   
 scale\_y\_continuous(breaks=c(0,250,500,750,1000), expand = c(0,.05)) +   
 ggtitle("NY State: Employee Count per Industry\*") + labs(x = "", y = "") +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1, vjust=0.3),  
 axis.ticks.x = element\_blank(),  
 panel.grid.minor.y = element\_blank())



**\*The following companies are not displayed on above plot, but are included in industry representation:**

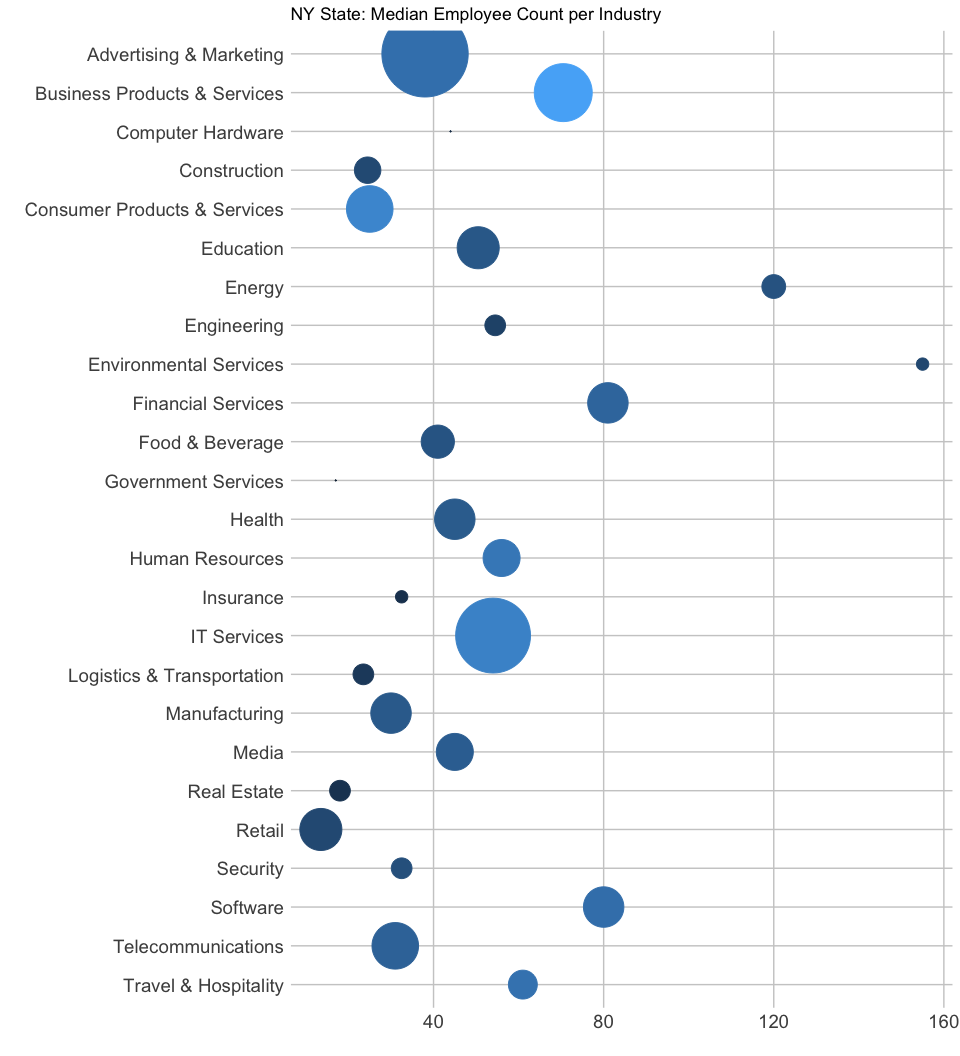
|  |  |  |
| --- | --- | --- |
| Industry | Name | Employees |
| Business Products & Services | Sutherland Global Services | 32000 |
| Consumer Products & Services | Coty | 10000 |
| IT Services | Westcon Group | 3000 |
| Travel & Hospitality | Denihan Hospitality Group | 2280 |
| Business Products & Services | TransPerfect | 2218 |
| Human Resources | Sterling Infosystems | 2081 |
| Software | OpenLink | 1271 |

Decision points:

* Boxplots by definition display meadian values as well as give general idea about outliers and the spread and variability of data. However, they require the viewer to have general idea about boxplots.
* Number of industries is not too high and it is more common to have numbers on the *y* axis, so boxplots are drawn vertically.
* Major gridlines and light background help define the plot.
* The plot is zoomed in to 0 to 1200 range (leaving a few outliers off the plot) in order to make information more legible.

One problem with boxplots is that they do not give a sense of how many data points there are for each industry. 25 industries range from 1 to 57 companies each and employee count ranges from 17 to 38,804. The plot below tries to address it. It represents median values. **Size is relative to number of companies per industry and color is relative to number of employees per idustry.** Consider *Human Resources* industry. A smaller point indicates few companies (in fact 11), but lighter color indicates relatively large number of employees (4,813). Because the plot is meant to be illutrative, legends are omitted. Perhaps, hover functionality to display actual values would be a good addition.

# Summarize NY data  
nyIndSum <- group\_by(nyInd, Industry) %>%  
 summarise(Median = median(Employees), TotalEmp = sum(Employees), Count = n())  
# Plot  
ggplot(aes(x = Industry, y = Median, size = Count, color = log(TotalEmp)),   
 data = nyIndSum) +   
 geom\_point(show.legend = FALSE) +  
 scale\_size(range = c(0, 30)) +  
 scale\_x\_discrete(limits = rev(levels(nyIndSum$Industry))) +  
 coord\_flip() +   
 ggtitle("NY State: Median Employee Count per Industry") + labs(x = "", y = "") +  
 theme(axis.ticks = element\_blank(),  
 axis.text = element\_text(size = 14),  
 panel.grid.major = element\_line(color = "#CCCCCC"),  
 panel.background = element\_blank(),  
 panel.grid.minor.x = element\_blank())

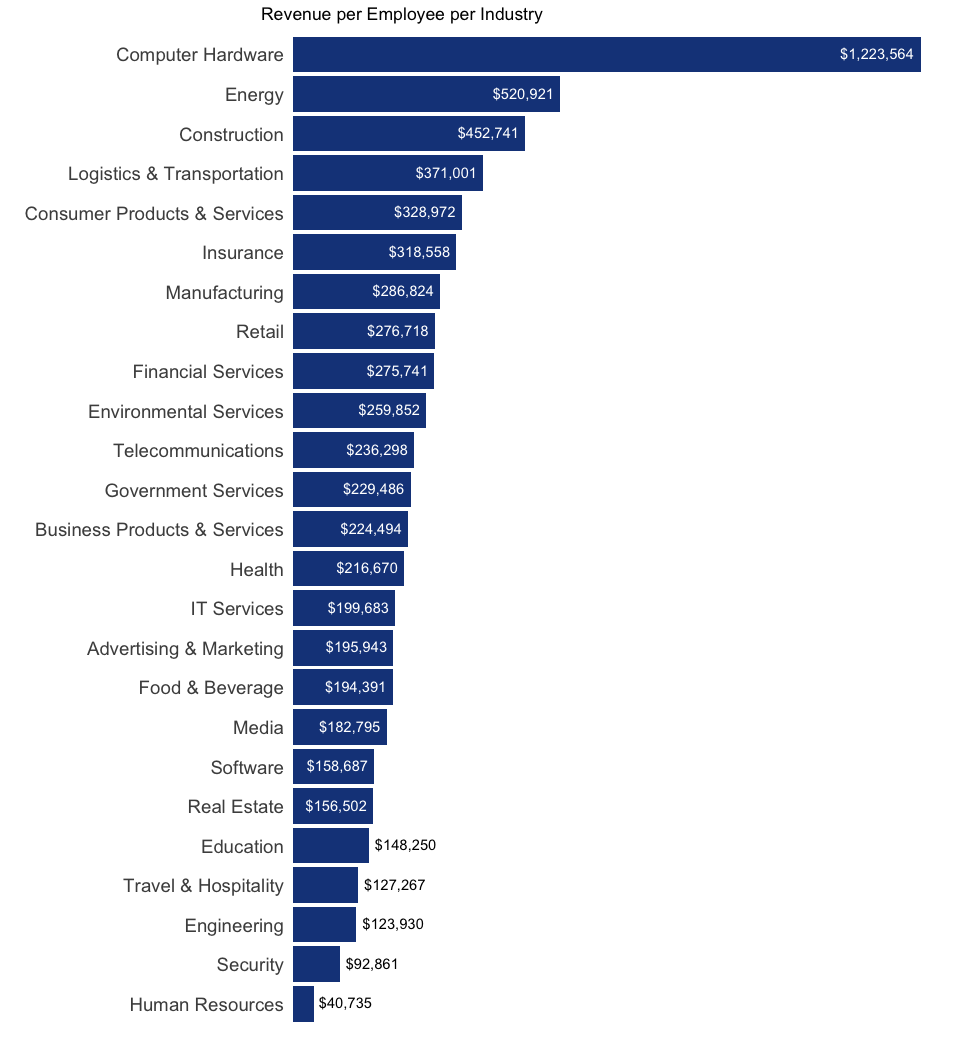


## Question 3

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.

Based on the summary statistics above there are no missing, negative or zero values in the Revenue column.

# Get data  
revenue <- select(inc, Industry, Revenue, Employees) %>%  
 na.omit() %>%  
 group\_by(Industry) %>%  
 summarise(TotalRev = sum(Revenue), TotalEmp = sum(Employees)) %>%  
 mutate(RevEmployee = TotalRev / TotalEmp)  
# Plot results  
ggplot(data = revenue, aes(x = reorder(Industry, RevEmployee), y = RevEmployee)) +   
 geom\_bar(stat="identity", fill="#184489") +  
 geom\_text(data = filter(revenue, RevEmployee>150000),  
 aes(x = Industry, y = RevEmployee, label=dollar\_format()(RevEmployee)),   
 hjust=1.1, vjust=0.4, color="#FFFFFF") +  
 geom\_text(data = filter(revenue, RevEmployee<150000),  
 aes(x = Industry, y = RevEmployee, label=dollar\_format()(RevEmployee)),   
 hjust=-0.1, vjust=0.4, color="#000000") +  
 coord\_flip() +   
 ggtitle("Revenue per Employee per Industry") + labs(x = "", y = "") +  
 theme(panel.background = element\_blank(),  
 axis.ticks = element\_blank(),  
 axis.text.x = element\_blank(),  
 axis.text.y = element\_text(size = 14, margin = margin(r=-20)))



Decision points:

* Similar to plot from question 1.
* Sorted by amount since the focus is likely to be on top/bottom industries.
* Additional embellishments were considered, but deemed unnecessary. Those include varying bar colors or bringing another dimension to show total number of employees. The idea is to have a simple display of revenue per employee. Further analysis of interesting industries can be done to break it up by state or city, average employees per company, etc.