EDA with R

Maria A Ginorio

9/11/2022

## **Principles of Data Visualization and Introduction to ggplot**

### Required Packages

library(tidyverse)

### Data

data about the 5,000 fastest growing companies in the US, as compiled by Inc. magazine

skimr::skim\_without\_charts(inc\_df)

Data summary

|  |  |
| --- | --- |
| Name | inc\_df |
| Number of rows | 5001 |
| Number of columns | 8 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 4 |
| numeric | 4 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| Name | 0 | 1 | 2 | 51 | 0 | 5001 | 0 |
| Industry | 0 | 1 | 5 | 28 | 0 | 25 | 0 |
| City | 0 | 1 | 4 | 22 | 0 | 1519 | 0 |
| State | 0 | 1 | 2 | 2 | 0 | 52 | 0 |

**Variable type: numeric**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
| Rank | 0 | 1 | 2501.64 | 1443.51 | 1.0e+00 | 1.252e+03 | 2.502e+03 | 3.751e+03 | 5.0000e+03 |
| Growth\_Rate | 0 | 1 | 4.61 | 14.12 | 3.4e-01 | 7.700e-01 | 1.420e+00 | 3.290e+00 | 4.2148e+02 |
| Revenue | 0 | 1 | 48222535.49 | 240542281.14 | 2.0e+06 | 5.100e+06 | 1.090e+07 | 2.860e+07 | 1.0100e+10 |
| Employees | 12 | 1 | 232.72 | 1353.13 | 1.0e+00 | 2.500e+01 | 5.300e+01 | 1.320e+02 | 6.6803e+04 |

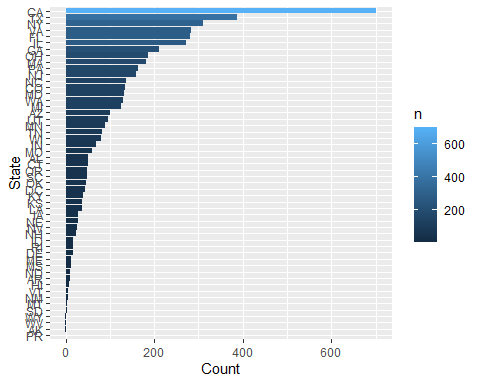
#### Question 1

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.

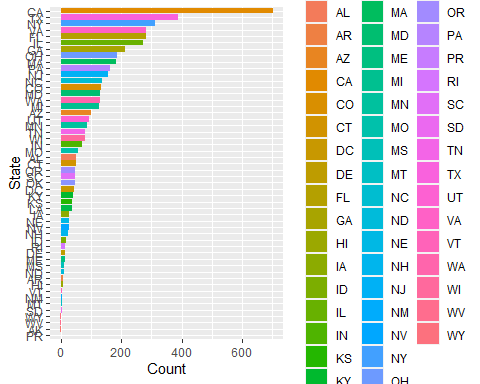
inc\_dist <- inc\_df %>%   
 group\_by(State) %>%   
 count(State) %>%   
 arrange(desc(n))

g <- ggplot(inc\_dist, aes(y=reorder(State, n), x=n))

g + geom\_col(aes(fill = n)) +  
 xlab("Count") +   
 ylab("State")



g + geom\_col(aes(fill = State)) +  
 xlab("Count") +   
 ylab("State")



#### 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.

NY\_state <- inc\_df %>%   
 filter(State=="NY")  
  
skimr::skim(NY\_state)

Data summary

|  |  |
| --- | --- |
| Name | NY\_state |
| Number of rows | 311 |
| Number of columns | 8 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 4 |
| numeric | 4 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

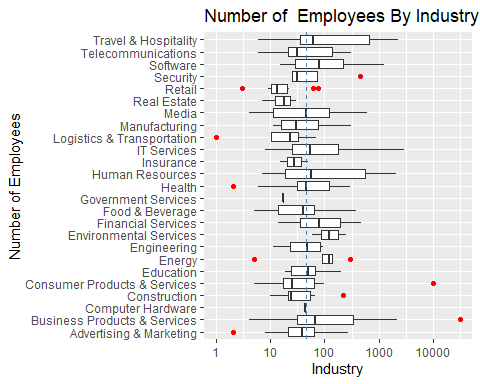
**Variable type: character**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| Name | 0 | 1 | 2 | 44 | 0 | 311 | 0 |
| Industry | 0 | 1 | 5 | 28 | 0 | 25 | 0 |
| City | 0 | 1 | 4 | 16 | 0 | 90 | 0 |
| State | 0 | 1 | 2 | 2 | 0 | 1 | 0 |

**Variable type: numeric**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| Rank | 0 | 1 | 2611.69 | 1515.31 | 2.6e+01 | 1185.50 | 2.702e+03 | 4.005e+03 | 4.981e+03 | ▆▆▆▆▇ |
| Growth\_Rate | 0 | 1 | 4.37 | 10.08 | 3.5e-01 | 0.67 | 1.310e+00 | 3.580e+00 | 8.443e+01 | ▇▁▁▁▁ |
| Revenue | 0 | 1 | 58715112.54 | 341922076.20 | 2.0e+06 | 4300000.00 | 8.800e+06 | 2.570e+07 | 4.600e+09 | ▇▁▁▁▁ |
| Employees | 0 | 1 | 271.29 | 1916.18 | 1.0e+00 | 21.00 | 4.500e+01 | 1.055e+02 | 3.200e+04 | ▇▁▁▁▁ |

NY\_state %>%   
 filter(complete.cases(.)) %>% # complete cases only  
 group\_by(Industry) %>%   
 select(Industry, Employees) %>%  
 ggplot(mapping=aes(x=Industry, y=Employees)) +  
 geom\_boxplot(outlier.color="red") +  
 coord\_flip() +  
 geom\_hline(yintercept = median(NY\_state$Employees),  
 color="steelblue",  
 linetype="dashed") +   
 scale\_y\_log10() +   
 xlab("Number of Employees") +   
 ylab("Industry") +   
 ggtitle(label = "Number of Employees By Industry In NY")

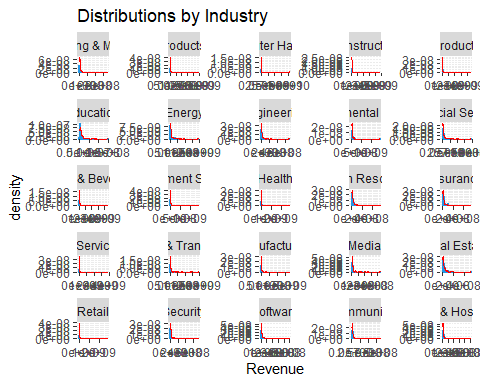


#### Question 3

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.

industry\_dist <- inc\_df %>%   
 select(Revenue, Industry)

ggplot(industry\_dist, aes(Revenue)) +   
 geom\_histogram(aes(x=Revenue, y = ..density..),   
 colour = 4, bins = 30) +  
 geom\_density(aes(x=Revenue), color = "red") +  
 facet\_wrap(~Industry, scales = "free") +   
 ggtitle(label = "Distributions by Industry")



#ggsave("industry\_distribution.png")

inc\_df %>%  
 group\_by(Industry) %>%  
 summarize(total\_rev = sum(Revenue),   
 total\_emp = sum(Employees),   
 rev\_per\_emp = total\_rev/total\_emp) %>%  
 arrange(desc(rev\_per\_emp)) %>%  
 na.omit() %>%  
 ggplot(aes(x=reorder(Industry, rev\_per\_emp), y=rev\_per\_emp)) +  
 geom\_bar(aes(fill = rev\_per\_emp), stat="identity" ) +  
 coord\_flip() +   
 xlab("Revenue / Employee") +   
 ylab("Industry") +   
 ggtitle(label = "Revenue Per Employee By Industry")

