Data 608-01 Spring 2021: Homework #1

Adam Rich

February 14, 2021

Additional Summaries

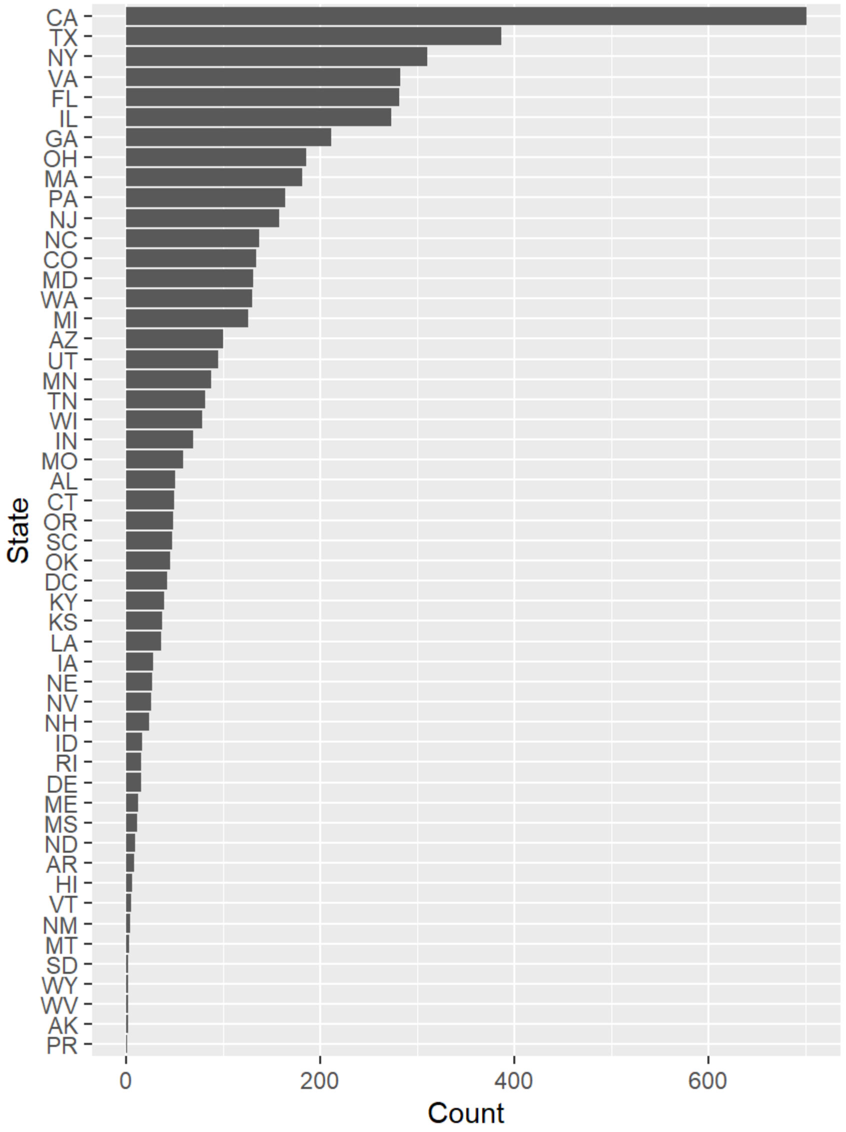
```
inc <- readr::read_csv("inc5000_data.csv")
```

```
# Number of unique industries represented
length(unique(inc$Industry))
## [1] 25

# Unique States
# 50 + DC + PR
length(unique(inc$State))
## [1] 52
table(inc$State)
##
##  AK  AL  AR  AZ  CA  CO  CT  DC  DE  FL  GA  HI  IA  ID  IL  IN  KS  KY  LA  MA
##   2  51   9 100 701 134  50  43  16 282 212   7  28  17 273  69  38  40  37 182
##  MD  ME  MI  MN  MO  MS  MT  NC  ND  NE  NH  NJ  NM  NV  NY  OH  OK  OR  PA  PR
## 131  13 126  88  59  12   4 137  10  27  24 158   5  26 311 186  46  49 164   1
##  RI  SC  SD  TN  TX  UT  VA  VT  WA  WI  WV  WY
##  16  48   3  82 387  95 283   6 130  79   2   2

# Top 5 industries by count
inc %>%
  group_by(Industry) %>%
  summarize(Count = n()) %>%
  arrange(desc(Count)) %>%
  head(5)
## `summarise()` ungrouping output (override with `.groups` argument)
## # A tibble: 5 x 2
##   Industry                     Count
##   <chr>                        <int>
## 1 IT Services                    733
## 2 Business Products & Services   482
## 3 Advertising & Marketing        471
## 4 Health                         355
## 5 Software                       342

# Top 5 States by count
inc %>%
  group_by(State) %>%
  summarize(Count = n()) %>%
  arrange(desc(Count)) %>%
  head(5)
## `summarise()` ungrouping output (override with `.groups` argument)
## # A tibble: 5 x 2
##   State Count
##   <chr> <int>
## 1 CA      701
## 2 TX      387
## 3 NY      311
## 4 VA      283
## 5 FL      282

# Top 5 companies by revenue
inc %>%
  mutate(RevenueBillions = Revenue / 1e9) %>%
  select(Name, RevenueBillions) %>%
  arrange(desc(RevenueBillions)) %>%
  head(5)
## # A tibble: 5 x 2
##   Name          RevenueBillions
##   <chr>                   <dbl>
## 1 CDW                      10.1
## 2 ABC Supply                4.7
## 3 Coty                      4.6
## 4 Dot Foods                 4.5
## 5 Westcon Group             3.8
```

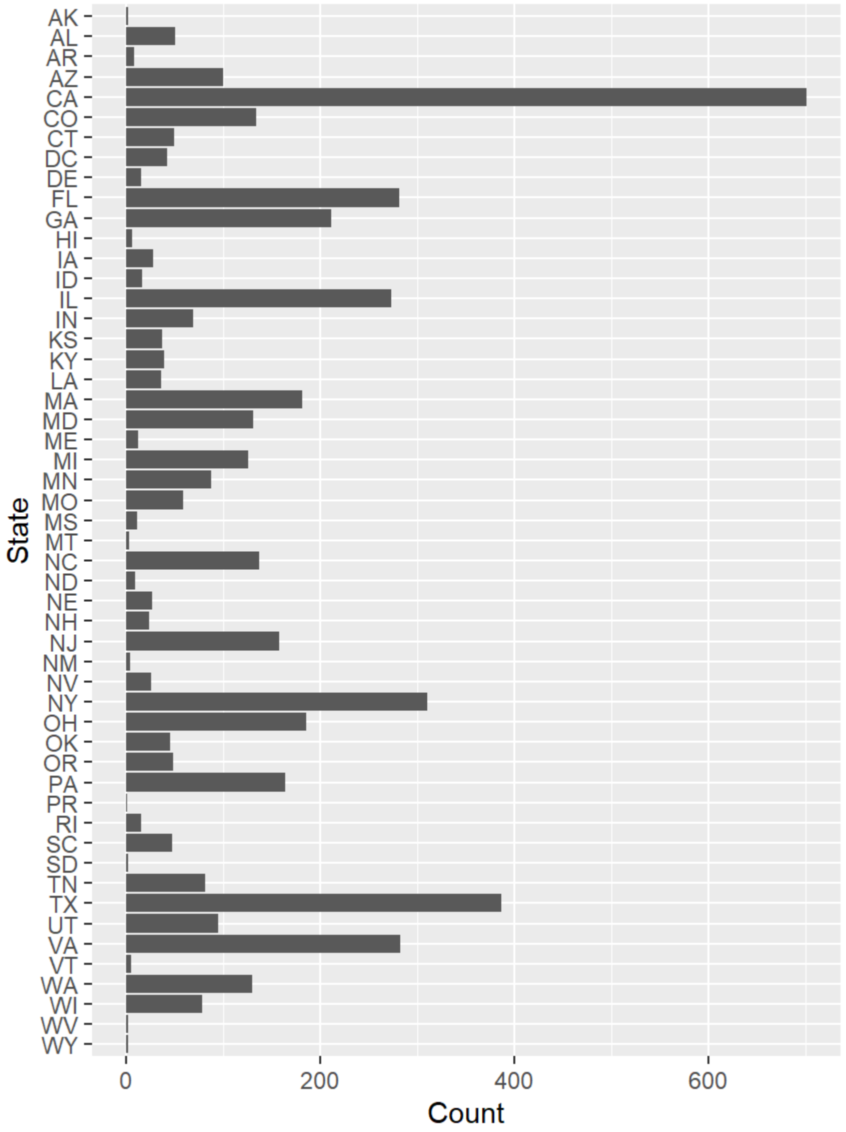
Question 1

- Show this two different ways
- By count – makes it easier to compare similar counts over so many levels
- But it makes it hard to find, so having an option to show it alphabetically would also be good

```
# Sorted by count for easier comparison
inc %>%
  mutate(State = factor(
    inc$State,
    levels = names(sort(table(inc$State), decreasing = FALSE)))) %>%
  group_by(State) %>%
  summarize(Count = n()) %>%
  ggplot() +
  aes(x = State, y = Count) +
  geom_col() +
  coord_flip()
## `summarise()` ungrouping output (override with `.groups` argument)
```



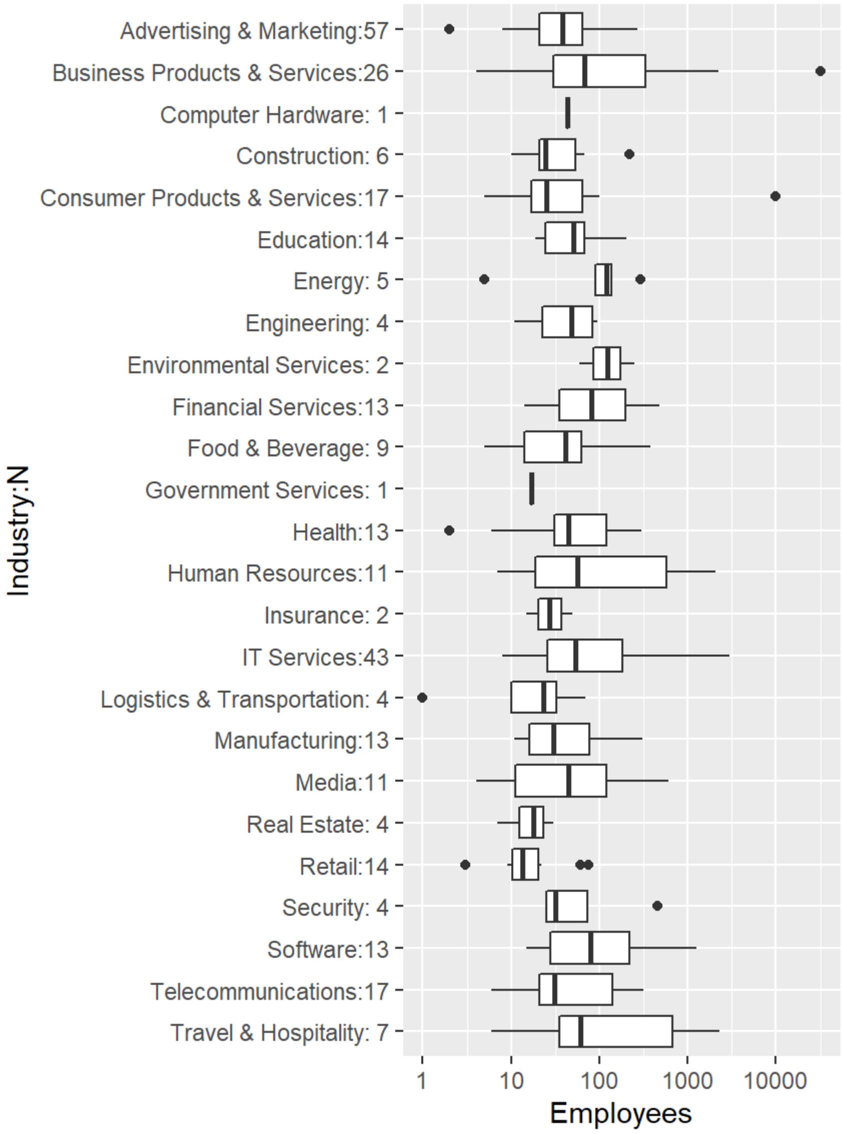
```
# Sorted alphabetically for easier finding
inc %>%
  mutate(State = factor(
    inc$State,
    levels = sort(unique(inc$State), decreasing = TRUE))) %>%
  group_by(State) %>%
  summarize(Count = n()) %>%
  ggplot() +
  aes(x = State, y = Count) +
  geom_col() +
  coord_flip()
## `summarise()` ungrouping output (override with `.groups` argument)
```



Question 2

- Sort industries alphabetically
- Use a boxplot to show variability
- Outliers are shown as individual points
- Show x axis on a log-scale because there are some very large companies
- Add the company count per industry to the label

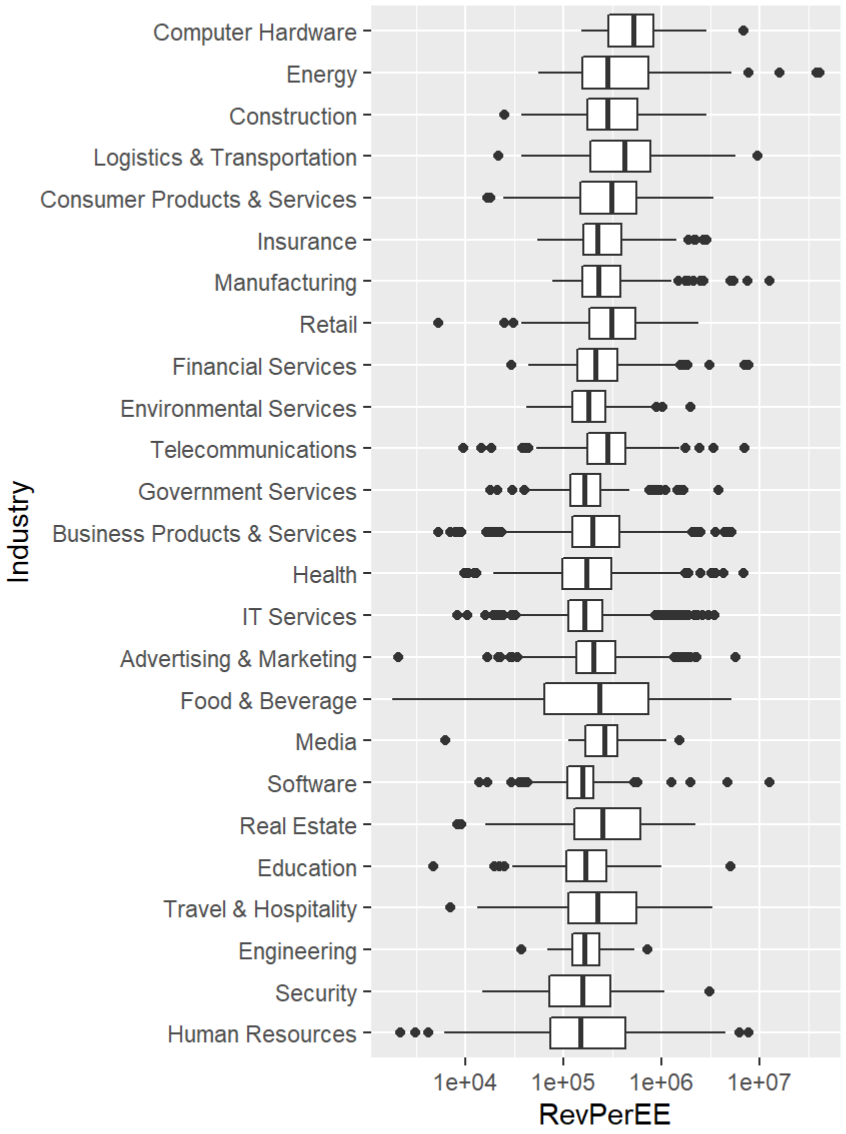
```
# New York is the 3rd most represented State
table(inc$State) %>% sort(decreasing = TRUE) %>% head()
##
##  CA  TX  NY  VA  FL  IL
## 701 387 311 283 282 273

inc_NY <- inc %>%
  filter(State == 'NY', complete.cases(.))

# Add company counts to labels
counts_by_industry <- inc_NY %>%
  group_by(Industry) %>%
  summarize(Count = n())
## `summarise()` ungrouping output (override with `.groups` argument)

counts_by_industry$`Industry:N` <-
  apply(
    X = counts_by_industry[, 1:2],
    MARGIN = 1,
    FUN = function(r) {paste0(r, collapse = ':')})

# Show employee counts on LOG scale because outliers are HUGE
inc_NY %>%
  inner_join(counts_by_industry) %>%
  mutate(`Industry:N` = factor(
      x = `Industry:N`,
      levels = sort(unique(`Industry:N`), decreasing = TRUE))) %>%
  ggplot() +
  aes(x = `Industry:N`, y = Employees) +
  geom_boxplot() +
  scale_y_log10() +
  coord_flip()
## Joining, by = "Industry"
```



Question 3

- Use a boxplot to show variability
- Sort industries by *overall* revenue per employee
- Boxplot shows the variability of the *individual* revenue per employee values

```
ranked <- inc %>%
  filter(complete.cases(.)) %>%
  group_by(Industry) %>%
  summarize(RevPerEE_IndustryWide = sum(Revenue) / sum(Employees)) %>%
  arrange(RevPerEE_IndustryWide)
## `summarise()` ungrouping output (override with `.groups` argument)

# Assuming I should go back to full dataset
inc %>%
  filter(complete.cases(.)) %>%
  mutate(
    RevPerEE = Revenue / Employees,
    Industry = factor(Industry, ranked$Industry)) %>%
  ggplot() +
  aes(x = Industry, y = RevPerEE) +
  geom_boxplot() +
  scale_y_log10() +
  coord_flip()
```


Industry
Computer Hardware
Energy
Construction
Logistics & Transportation
Consumer Products & Services
Insurance
Manufacturing
Retail
Financial Services
Environmental Services
Telecommunications
Government Services
Business Products & Services
Health
IT Services
Advertising & Marketing
Food & Beverage
Media
Software
Real Estate
Education
Travel & Hospitality
Engineering
Security
Human Resources
1e+04
1e+05
1e+06
1e+07
RevPerEE