Data608_HW1_AGoldberg

Andrew Goldberg

2/5/2017

```
knitr::opts_chunk$set(echo = TRUE)
```

```
dat <- read.csv('https://raw.githubusercontent.com/charleyferrari/CUNY_DATA608/master/lecture1/Data/inc5
summary(dat)
```

```
##       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      Employees      
##  Min.   :2.000e+06   IT Services                 : 733   Min.   :    1.0  
##  1st Qu.:5.100e+06   Business Products & Services: 482   1st Qu.:   25.0  
##  Median :1.090e+07   Advertising & Marketing     : 471   Median :   53.0  
##  Mean   :4.822e+07   Health                      : 355   Mean   :  232.7  
##  3rd Qu.:2.860e+07   Software                    : 342   3rd Qu.:  132.0  
##  Max.   :1.010e+10   Financial Services          : 260   Max.   :66803.0  
##                      (Other)                     :2358   NA's   :12       
##             City          State     
##  New York     : 160   CA     : 701  
##  Chicago      :  90   TX     : 387  
##  Austin       :  88   NY     : 311  
##  Houston      :  76   VA     : 283  
##  San Francisco:  75   FL     : 282  
##  Atlanta      :  74   IL     : 273  
##  (Other)      :4438   (Other):2764
```

```
head(dat)
```

```
##   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
```

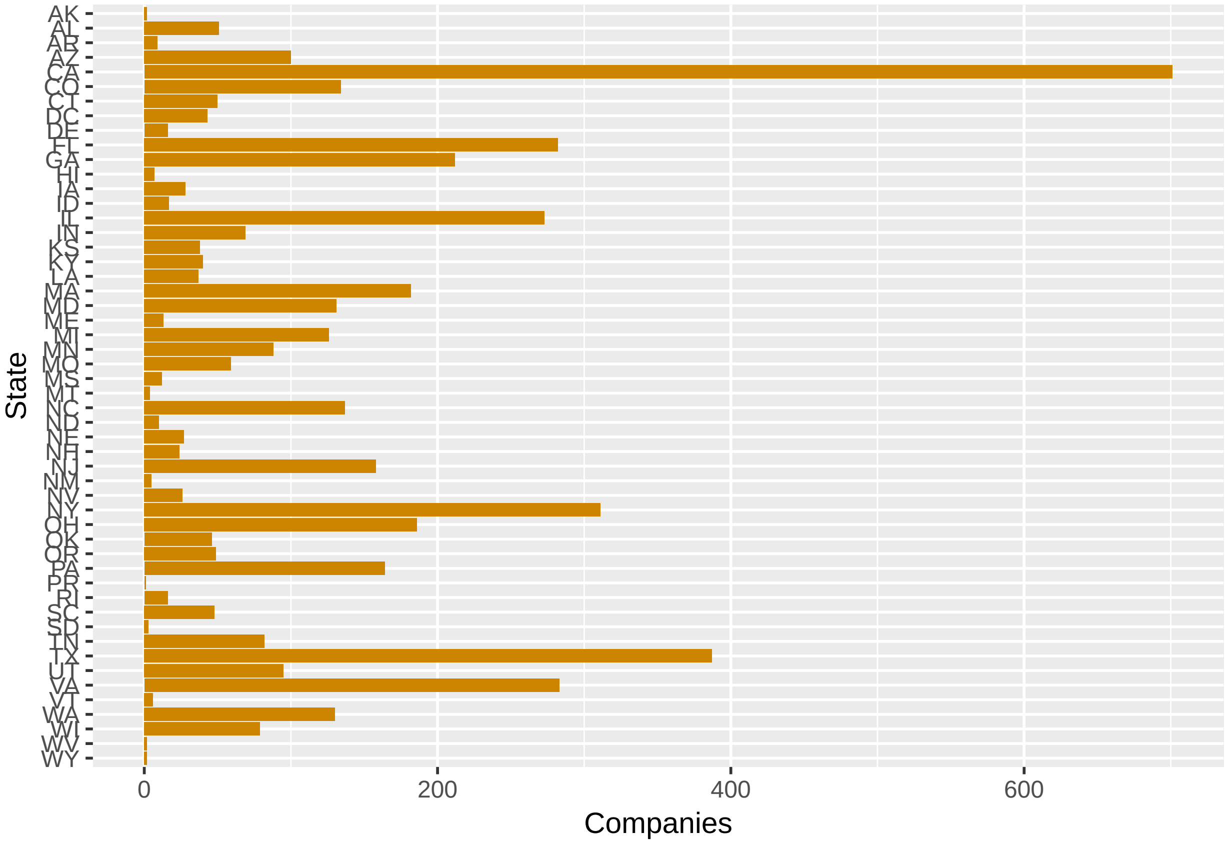
```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)

#1. Create a graph that shows the distribution of companies in the dataset by State (ie how many are in

#count companies by state
companyState <- dat %>%
  select(Name, State) %>%
  group_by(State) %>%
  summarise(Companies = n())
companyState
```

```
## # A tibble: 52 × 2
##     State Companies
##    <fctr>     <int>
## 1      AK         2
## 2      AL        51
## 3      AR         9
## 4      AZ       100
## 5      CA       701
## 6      CO       134
## 7      CT        50
## 8      DC        43
## 9      DE        16
## 10     FL       282
## # ... with 42 more rows
```

```
#reverse order of state names
companyState <- within(companyState, State <- ordered(State, levels = rev(State)))

#plot barchart
ggplot(companyState, aes(y = Companies, x = State)) + geom_bar(fill = 'orange3', stat='identity') + coo
```



```
#2. Create a plot of average employment by industry for companies in this state

#find state with third most companies
orderedStates <- companyState %>%
  arrange(desc(Companies))
thirdState <- orderedStates[3,][,1]
dat$State[dat$State == toString(thirdState)]
```

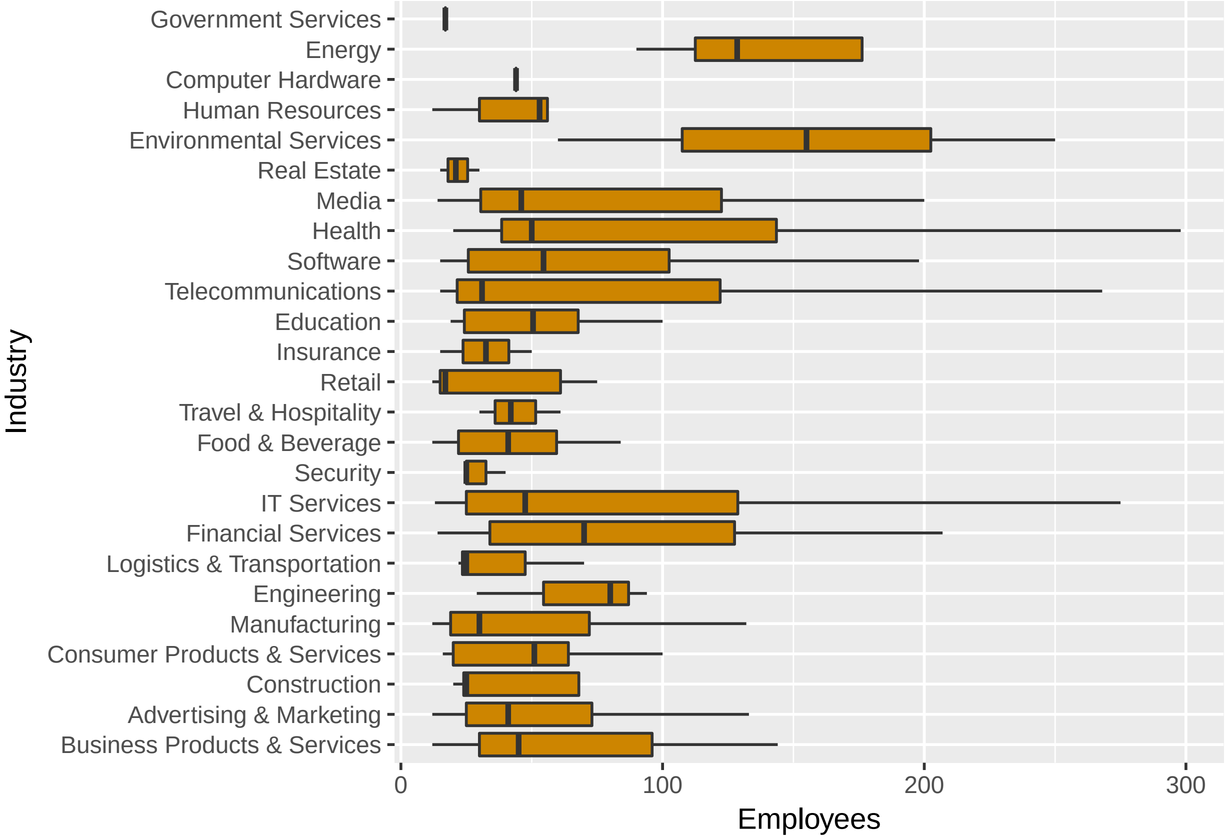
```
## factor(0)
## 52 Levels: AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA ... WY
```

```
#filter by third state
avgEmploymentIndustry <- dat %>%
  filter(State == toString(thirdState[[1]]))
avgEmploymentIndustry <- avgEmploymentIndustry[complete.cases(avgEmploymentIndustry),]

#reverse order of industry names
avgEmploymentIndustry <- within(avgEmploymentIndustry, Industry <- ordered(Industry, levels = rev(Indus
```

```
## Warning in `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels)
## else paste0(labels, : duplicated levels in factors are deprecated
```

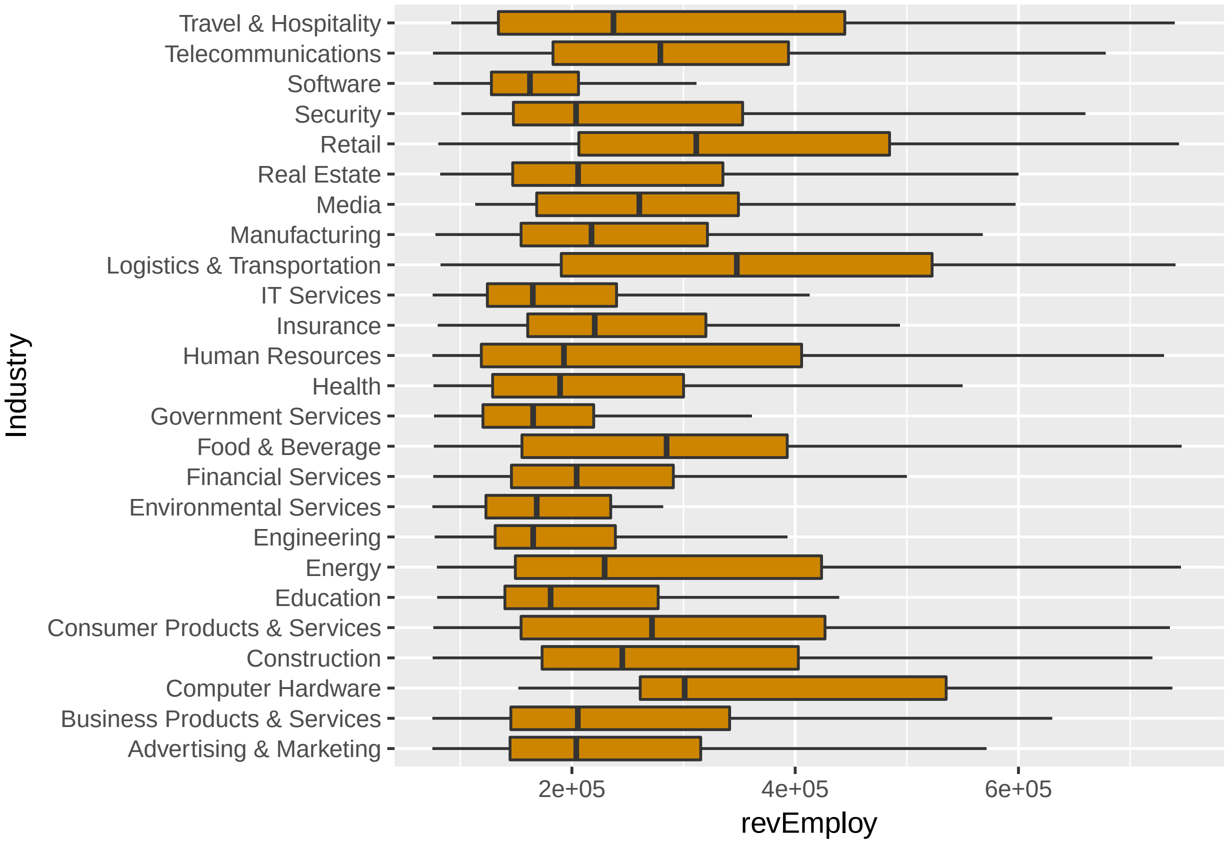
```
#define limits to exclude outliers
outlierLimits <- as.numeric(quantile(avgEmploymentIndustry$Employees, c(0.1, 0.9)))

#plot barchart
ggplot(avgEmploymentIndustry, aes(y = Employees, x = Industry)) + geom_boxplot(outlier.shape=NA, fill =
```

```
## Warning in `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels)
## else paste0(labels, : duplicated levels in factors are deprecated

## Warning in `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels)
## else paste0(labels, : duplicated levels in factors are deprecated

## Warning: Removed 62 rows containing non-finite values (stat_boxplot).
```



```
#3 Now imagine you work for an investor and want to see which industries generate the most revenue per

#calculate revenue per employee
revenueEmployee <- dat %>%
  select(Revenue, Employees, Industry) %>%
  mutate(revEmploy = Revenue / Employees)

#remove incomplete cases
revenueEmployee<- revenueEmployee[complete.cases(revenueEmployee),]

#define limits to exclude outliers
outlierLimits <- as.numeric(quantile(revenueEmployee$revEmploy, c(0.1, 0.9)))

#plot barchart
ggplot(revenueEmployee, aes(y = revEmploy, x = Industry)) + geom_boxplot(outlier.shape=NA, fill = 'orang
```

```
## Warning: Removed 997 rows containing non-finite values (stat_boxplot).
```


Industry
Travel & Hospitality
Telecommunications
Software
Security
Retail
Real Estate
Media
Manufacturing
Logistics & Transportation
IT Services
Insurance
Human Resources
Health
Government Services
Food & Beverage
Financial Services
Environmental Services
Engineering
Energy
Education
Consumer Products & Services
Construction
Computer Hardware
Business Products & Services
Advertising & Marketing
2e+05
4e+05
6e+05
revEmploy