Assignment 1

CUNY MSDS, Data 608

Tamiko Jenkins

TFeb 09, 2020

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:

```
inc <- read.csv("https://raw.githubusercontent.com/charleyferrari/CUNY_DATA_608/master/module1/Data/inc5
```

And lets preview this data:

```
head(inc)
```

```
##   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
```

```
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      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
```

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


# Import display libraries and set display options
library("knitr")
library("rmarkdown")
knitr::opts_chunk$set(comment = NA)

# Import main libraries
library("ggplot2")
library("tidyverse")
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
## v purrr   0.3.3
```

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

```
library("dplyr")


library("psych")
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
describe(inc)
```

```
##             vars    n        mean           sd    median     trimmed
## Rank           1 5001     2501.64      1443.51 2.502e+03     2501.73
```

```
## Name*          2 5001     2501.00      1443.81 2.501e+03     2501.00
## Growth_Rate    3 5001        4.61        14.12 1.420e+00        2.14
## Revenue        4 5001 48222535.49 240542281.14 1.090e+07 17334966.26
## Industry*      5 5001       12.10         7.33 1.300e+01       12.05
## Employees      6 4989      232.72      1353.13 5.300e+01       81.78
## City*          7 5001      732.00       441.12 7.610e+02      731.74
## State*         8 5001       24.80        15.64 2.300e+01       24.44
##                    mad     min        max      range  skew kurtosis         se
## Rank            1853.25 1.0e+00 5.0000e+03 4.9990e+03  0.00    -1.20      20.41
## Name*           1853.25 1.0e+00 5.0010e+03 5.0000e+03  0.00    -1.20      20.42
## Growth_Rate        1.22 3.4e-01 4.2148e+02 4.2114e+02 12.55   242.34       0.20
## Revenue     10674720.00 2.0e+06 1.0100e+10 1.0098e+10 22.17   722.66 3401441.44
## Industry*          8.90 1.0e+00 2.5000e+01 2.4000e+01 -0.10    -1.18       0.10
## Employees         53.37 1.0e+00 6.6803e+04 6.6802e+04 29.81  1268.67      19.16
## City*            604.90 1.0e+00 1.5190e+03 1.5180e+03 -0.04    -1.26       6.24
## State*            19.27 1.0e+00 5.2000e+01 5.1000e+01  0.12    -1.46       0.22
```

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.

Arrange data

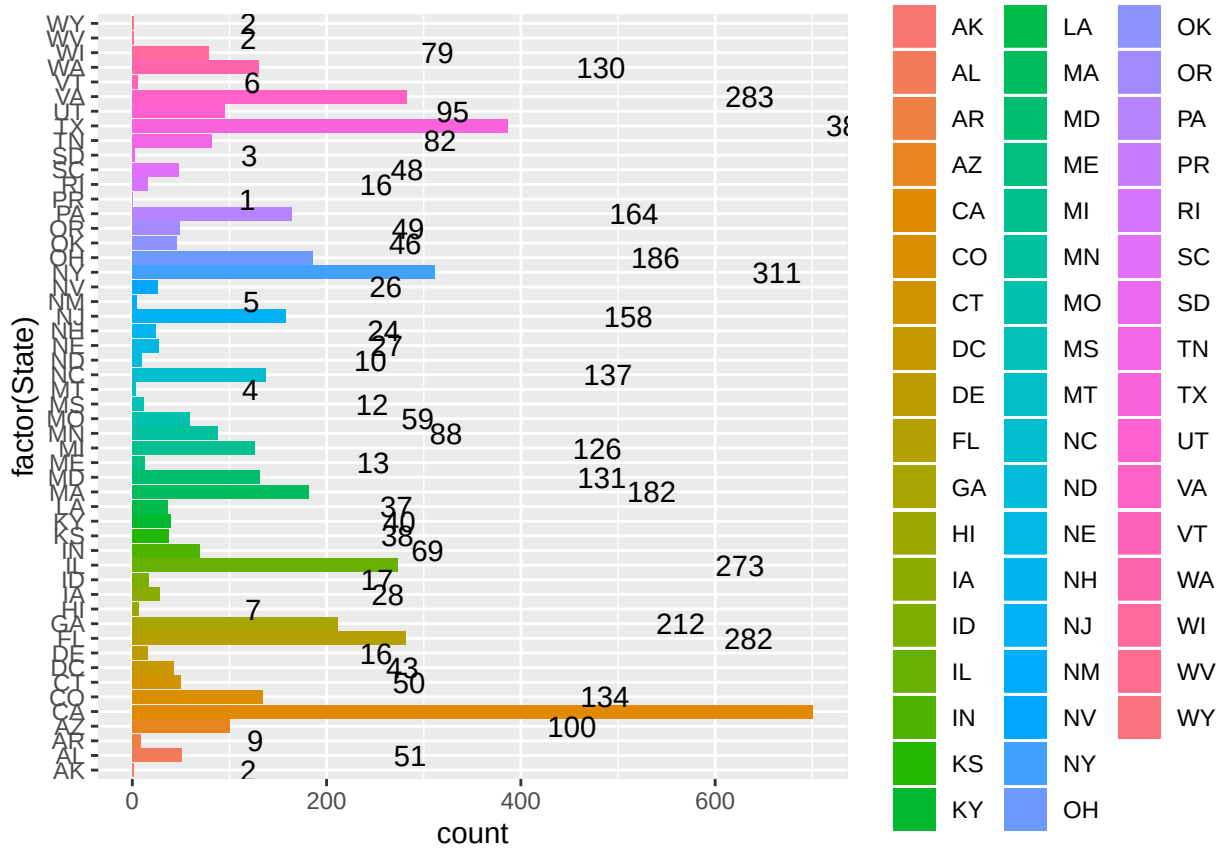
```
# Answer Question 1 here

inc_state <- inc %>%
  arrange(State)

inc_state_count  <- inc_state %>%
  count(State, sort=TRUE) %>%
  rename(Count=n) %>%
  arrange(Count) %>%
  mutate(State = factor(State, State))
```

Chart the default order and flip chart

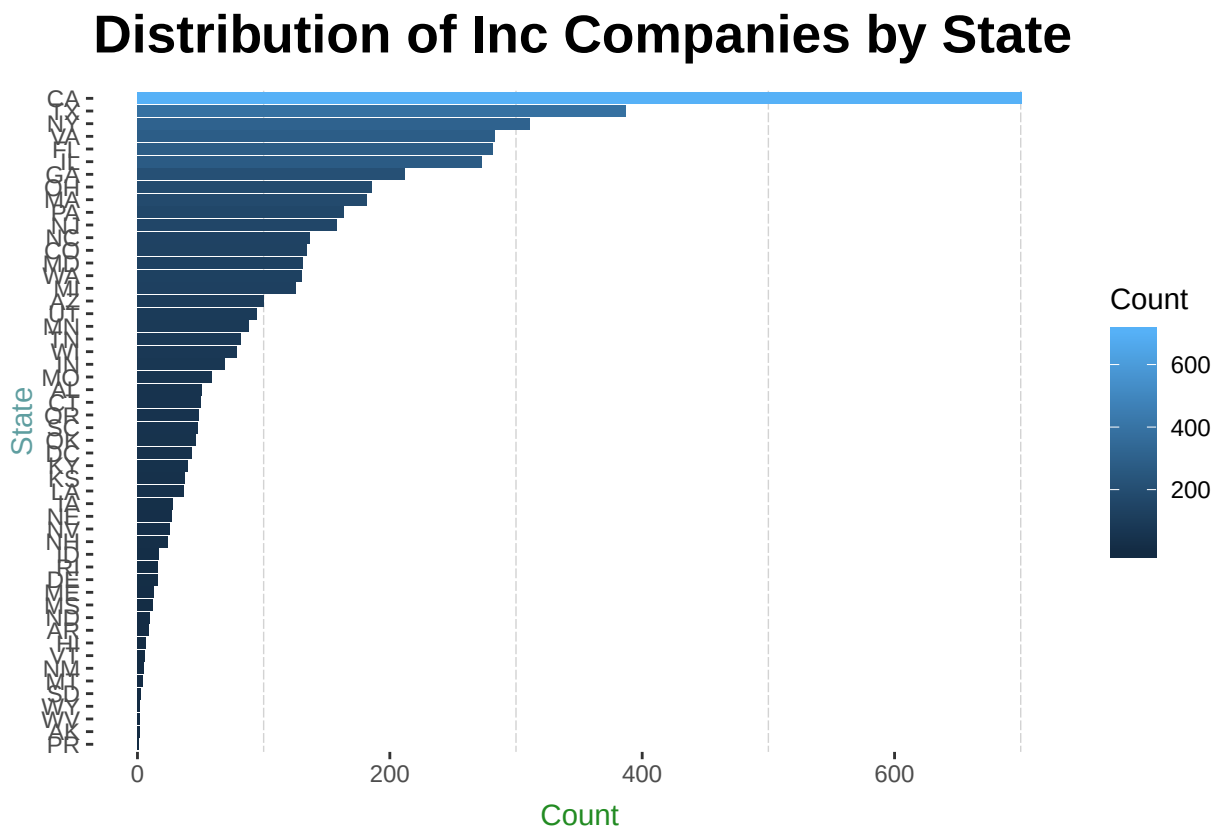
```
ggplot(inc_state, aes(x = factor(State), fill = State)) +
  geom_bar() +
  geom_text(aes(label = ..count..), stat = "count", hjust = -6.5, colour = "black") +
  coord_flip()
```



Arrange by State Count and Remove Labels, Emphasize Count range

```
inc_state_count %>%
  ggplot(aes(x = State, y = Count, fill=Count)) +
  ggtitle("Distribution of Inc Companies by State") +
  geom_bar(stat = 'identity', width=.9) +
  geom_text(aes(label = ''), vjust = -1, hjust = -4.5, colour = "black") +
  coord_fixed(ratio=4) +
  coord_flip() +
  theme(plot.title = element_text(size=20, face="bold",
    margin = margin(10, 0, 10, 0)),
    axis.title.x = element_text(color="forestgreen", vjust=-0.35),
    axis.title.y = element_text(color="cadetblue" , vjust=0.35),
    panel.background = element_rect(fill = 'white'),
    panel.grid.minor = element_line(colour = "lightgrey")
  )
```

```
Coordinate system already present. Adding new coordinate system, which will replace the existing one.
```



References:

https://r-graphics.org/recipe-bar-graph-grouped-bar

https://stackoverflow.com/a/54504480

http://zevross.com/blog/2014/08/04/beautiful-plotting-in-r-a-ggplot2-cheatsheet-3/

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.

Get the state with the third highest total numbers of Employees of all complete entries: New York

```
# Answer Question 2 here
inc_state_cc <- inc_state %>% filter(complete.cases(.))
state_emp_counts <- inc_state_cc %>% count(State, sort=TRUE)
state_emp_counts %>% head(n=3)
```

```
# A tibble: 3 x 2
  State     n
  <fct> <int>
1 CA      700
2 TX      386
3 NY      311
```

```
third <- state_emp_counts  %>%  slice(3)
third <- third[["State"]][1] %>% toString
```

Find the average number of employees in each industry in New York

```
ny_avg_emp_ind <- inc_state_cc %>%
  filter(State==third) %>%
  group_by(Industry) %>%
  mutate(avg_emp_ind = mean(Employees)) %>%
  arrange(Industry)
```

Example of a company in each NY Industry with its employee average

```
ny_avg_emp_ind[!duplicated(ny_avg_emp_ind$avg_emp_ind),]
```

```
# A tibble: 25 x 9
# Groups:   Industry [25]
    Rank Name    Growth_Rate  Revenue Industry Employees City  State avg_emp_ind
   <int> <fct>         <dbl>    <dbl> <fct>        <int> <fct> <fct>       <dbl>
 1    30 Sailth~       73.2   8100000 Adverti~        79 New ~ NY           58.4
 2   264 MSR Pr~       16.3   2400000 Busines~         4 New ~ NY         1492.
 3  2877 Myriad~        1.19 22900000 Compute~        44 New ~ NY           44
 4  1723 Spicer~        2.25  5600000 Constru~        20 Buff~ NY           61
 5    26 BeenVe~       84.4  13700000 Consume~        17 New ~ NY          626.
 6   232 Rethin~       18.3   4100000 Educati~        22 New ~ NY           59.9
 7   609 SmartW~        7.54 38800000 Energy         137 Ball~ NY          129.
 8  4474 Sam Sc~        0.51 16800000 Enginee~        94 New ~ NY           53.5
 9  3661 Enviro~        0.81 45100000 Environ~       250 Syra~ NY          155
10    48 Cinium~       53.6   5900000 Financi~        32 Rock~ NY          144.
# ... with 15 more rows
```

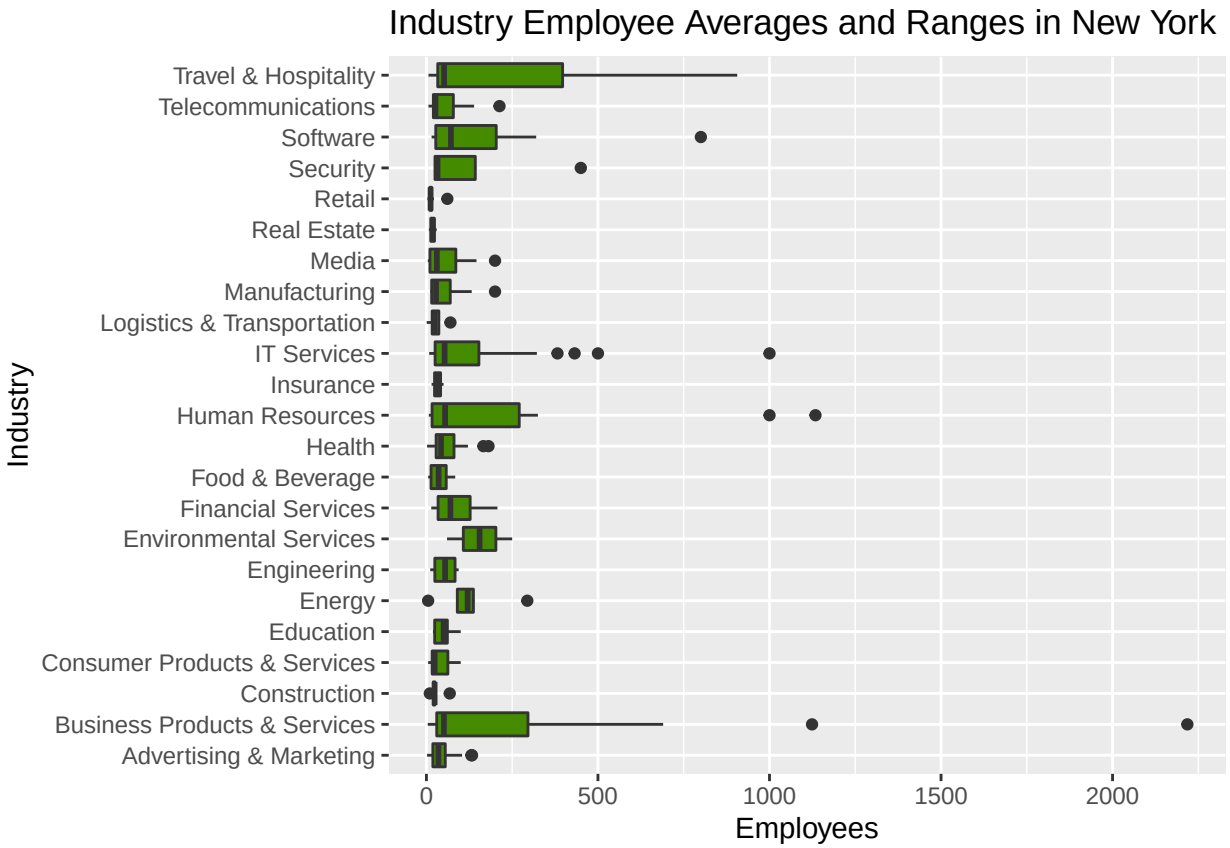
Chart Display

```
#graphics.off()
#par("mar")
#par(mar=c(1,1,1,1))
```

Show boxplots with based on outlier removal equation

```r
ny_emp_counts <- ny_avg_emp_ind %>%
  dplyr::select(Name, Employees, Industry, avg_emp_ind) %>%
  group_by(Industry)  %>%
  filter(!(abs(Employees - median(Employees)) > 2*sd(Employees))) # Remove outliers, needs work


ggplot(ny_emp_counts, aes(x=Industry, y=Employees)) +
  ggtitle("Industry Employee Averages and Ranges in New York") +
  geom_boxplot(fill="chartreuse4") +
#  geom_text(aes(label = avg_emp_ind), colour = "black") +
  coord_flip()
```



References:

http://zevross.com/blog/2014/08/04/beautiful-plotting-in-r-a-ggplot2-cheatsheet-3/

https://stackoverflow.com/questions/28687515/search-for-and-remove-outliers-from-a-dataframe-grouped-by-a-variable

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.

```r
# Answer Question 3 here

ny_emp_rev_ind <- inc_state_cc %>%
  dplyr::select(Employees, Industry, Revenue) %>%
  group_by(Industry)  %>%
  transmute(Rev_Emp=round(sum(Revenue)/sum(Employees))) %>%
  dplyr::select(Industry, Rev_Emp)


# To Do: Sort
ggplot(ny_emp_rev_ind, aes(x=Industry, y=Rev_Emp)) +
  ggtitle("NY Industry Revenue per Employee") +
  geom_bar(stat="identity") +
  coord_flip()
```