R Notebook

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/mast
er/module1/Data/inc5000_data.csv", header= TRUE)
```

Packages Used

```
require(tidyverse)
```

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:

```
# a table of counts of industry
inc %>% group_by(Industry) %>% tally() %>% arrange(desc(n))
```

```
## # A tibble: 25 x 2
##    Industry                         n
##    <fct>                        <int>
##  1 IT Services                    733
##  2 Business Products & Services   482
##  3 Advertising & Marketing        471
##  4 Health                         355
##  5 Software                       342
##  6 Financial Services             260
##  7 Manufacturing                  256
##  8 Consumer Products & Services   203
##  9 Retail                         203
## 10 Government Services            202
## # ... with 15 more rows
```

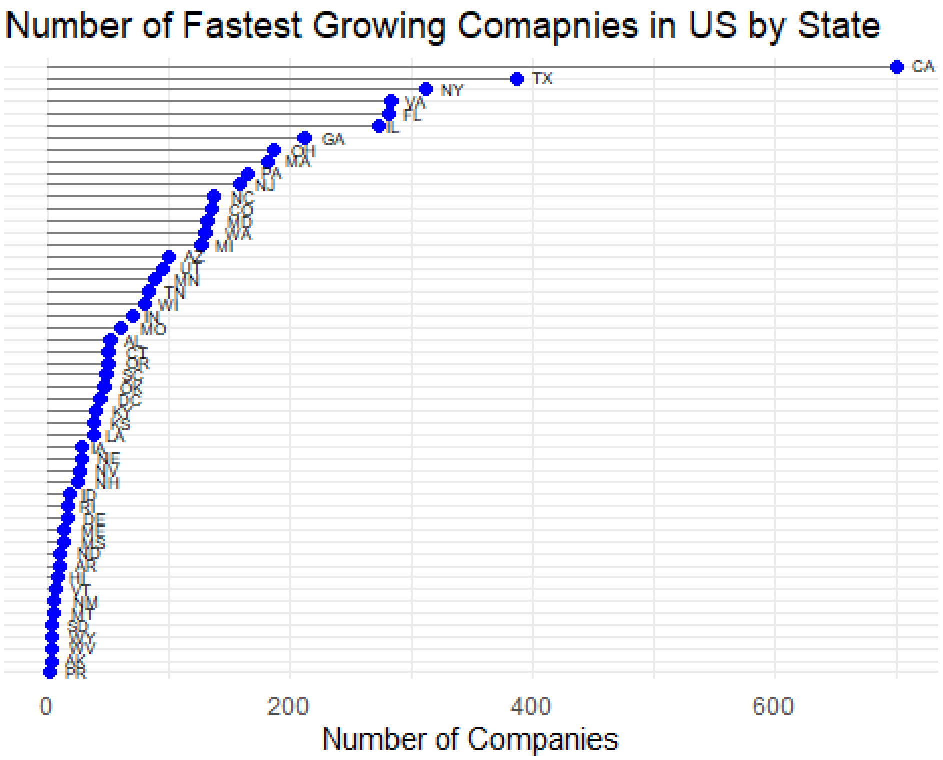
```
# table of total revenue by industry
inc %>% group_by(Industry) %>% summarize(TotalRev=sum(Revenue)) %>%
arrange(desc(TotalRev))
```

```
## # A tibble: 25 x 2
##    Industry                        TotalRev
##    <fct>                              <dbl>
##  1 Business Products & Services 26367900000
##  2 IT Services                  20681300000
##  3 Health                       17863400000
##  4 Consumer Products & Services 14956400000
##  5 Logistics & Transportation   14840500000
##  6 Energy                       13771600000
##  7 Construction                 13174300000
##  8 Financial Services           13150900000
##  9 Food & Beverage              12911300000
## 10 Manufacturing                12684000000
## # ... with 15 more rows
```

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.

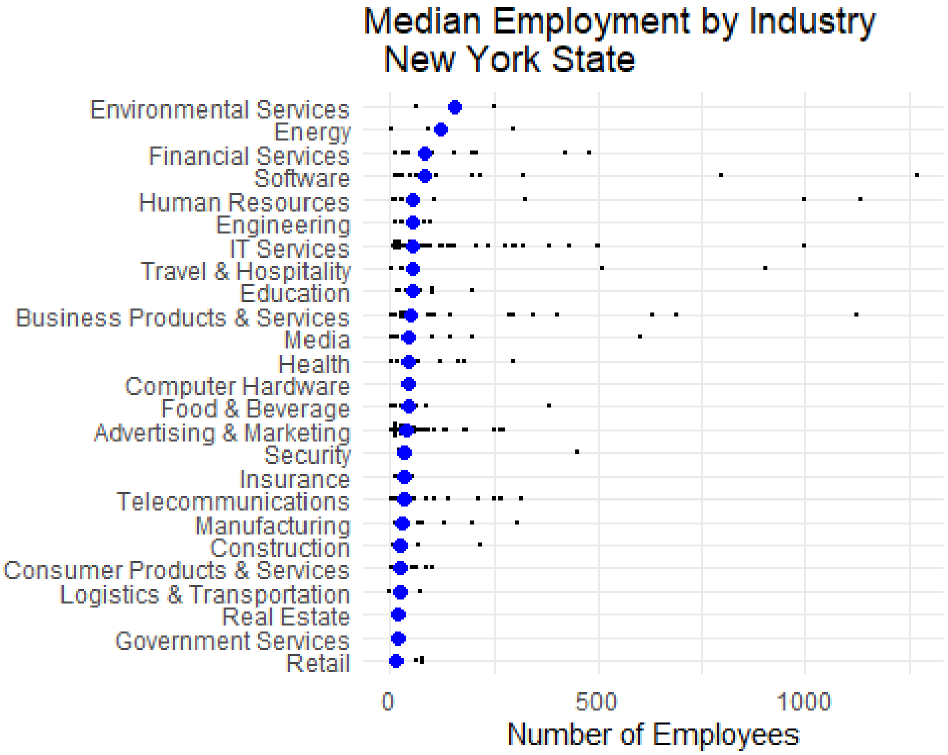
```
inc %>%
  group_by(State) %>%
  tally(sort = T) %>%
  filter(n>0) %>%
  ggplot(aes(x=reorder(State,n),y=n))+
    geom_segment(aes(xend=State,yend=0), color="grey50") +
    geom_point(size=2,color="blue")+
    geom_text(aes(label=State),size = 2, hjust=-.75, vjust=.4) +
    guides(fill=F)  +
    ggtitle("Number of Fastest Growing Comapnies in US by State") +
    labs(y='Number of Companies') +
    coord_flip() +
    theme_minimal()+
    theme(axis.title.y=element_blank(),
        axis.text.y=element_blank(),
        axis.ticks.y=element_blank())
```



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.

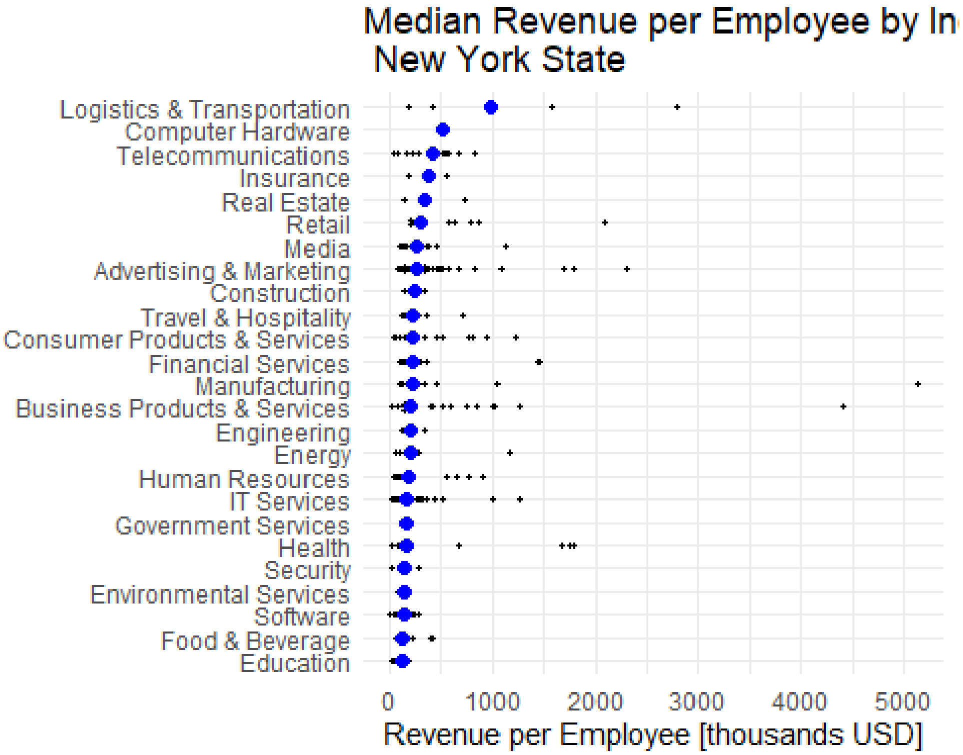
```
# find the state with the third most companies
inc %>% group_by(State) %>% tally() %>% arrange(desc(n)) %>% slice(3)

## # A tibble: 1 x 2
##   State     n
##   <fct> <int>
## 1 NY      311

# new dataset with only NY full cases
inc.NY <- inc %>% filter(complete.cases(.), State=='NY')

# get a list of top comapnies
inc.NY %>% arrange(desc(Employees)) %>% select(Name,Employees) %>% head()

##                          Name Employees
## 1 Sutherland Global Services     32000
## 2                        Coty     10000
## 3               Westcon Group      3000
## 4  Denihan Hospitality Group      2280
## 5                TransPerfect      2218
## 6        Sterling Infosystems      2081

# Dotplot with outliers removed
# Blue dots represent median values - small black dots are observations
inc.NY %>%
  filter(Employees<2000) %>% # removing outliers
  group_by(Industry) %>% #
  ggplot(aes(x=reorder(Industry, Employees,FUN=median), y=Employees)) +
    geom_dotplot(dotsize = 20, binaxis="y", binwidth = .5, stackdir =
"center") +
    stat_summary(fun.y=median, geom="point", size=2, color="blue") +
    labs(y='Number of Employees', title="Median Employment by Industry \n
New York State") +
    theme_minimal() +
    theme(axis.title.y=element_blank()) +
    coord_flip()
```



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.

```
# calculate and display the metric rev/emp and plot
# one outlier was removed

inc.NY %>%
  group_by(Industry) %>%
  mutate(RevPerEmp=Revenue/Employees/1000) %>%
  arrange(desc(RevPerEmp)) %>%
  filter(RevPerEmp < 40000) %>%
  ggplot(aes(x=reorder(Industry,RevPerEmp,FUN=median),y=RevPerEmp)) +
    geom_dotplot(dotsize=100, binaxis="y", binwidth = .5, stackdir =
"center") +
    stat_summary(fun.y=median, geom="point", size=2, color="blue") +
    labs(y='Revenue per Employee [thousands USD]', title="Median Revenue per
Employee by Industry \n New York State") +
    theme_minimal() +
    theme(axis.title.y=element_blank())+
    coord_flip()
```


Median Revenue per Employee by In
New York State
Logistics & Transportation
Computer Hardware
Telecommunications
Insurance
Real Estate
Retail
Media
Advertising & Marketing
Construction
Travel & Hospitality
Consumer Products & Services
Financial Services
Manufacturing
Business Products & Services
Engineering
Energy
Human Resources
IT Services
Government Services
Health
Security
Environmental Services
Software
Food & Beverage
Education
0
1000
2000
3000
4000
5000
Revenue per Employee [thousands USD]