R Notebook - Data 608 Module 1

Joseph Simone

```
library(ggplot2)
library(tidyverse)
library(tidyr)
library(dplyr)
library(knitr)
library(grid)
library(gridExtra)
library(latex2exp)
library(kableExtra)
library(ggthemes)
```

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
```

```
str(inc)
```

```
## 'data.frame':    5001 obs. of  8 variables:
##  $ Rank       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Name       : Factor w/ 5001 levels "(Add)ventures",..: 1770 1633 4423 690 1198 2839 4733 1468 186
##  $ Growth_Rate: num  421 248 245 233 213 ...
##  $ Revenue    : num  1.18e+08 4.96e+07 2.55e+07 1.90e+09 8.70e+07 ...
##  $ Industry   : Factor w/ 25 levels "Advertising & Marketing",..: 5 12 13 7 1 20 10 1 5 21 ...
##  $ Employees  : int  104 51 132 50 220 63 27 75 97 15 ...
##  $ City       : Factor w/ 1519 levels "Acton","Addison",..: 391 365 635 2 139 66 912 1179 131 1418 .
##  $ State      : Factor w/ 52 levels "AK","AL","AR",..: 5 47 10 45 20 45 44 5 46 41 ...
```

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:

Data Exploration

Upon Examining the summary and data values of this set, it was time to dive deeper in and set apart some Catergories from the DataFrame. First, before answering the questions, let's explore the Revenue, Indrusties, Employees, Cities, and States that make up the Data Set regarading the 5,000 fastest growing companies in the US, as compiled by Inc. magazine.

Growth Rate

I was curious to see which companies experienced a growth rate of 100 or higher. This was due to the growth rate growth from 0.340 to 431.480.

```
inc %>% dplyr::filter(Growth_Rate >= 100) %>% summarise(n = n())
```

```
##    n
## 1 19
```

```
kable(inc %>% dplyr::filter(Growth_Rate >= 100)) %>% kable_styling()
```

Rank	Name	Growth_Rate	Revenue	Industry	Employees	Cit
1	Fuhu	421.48	1.179e+08	Consumer Products & Services	104	El
2	FederalConference.com	248.31	4.960e+07	Government Services	51	Du
3	The HCI Group	245.45	2.550e+07	Health	132	Jac
4	Bridger	233.08	1.900e+09	Energy	50	Ad
5	DataXu	213.37	8.700e+07	Advertising & Marketing	220	Bo
6	MileStone Community Builders	179.38	4.570e+07	Real Estate	63	Au
7	Value Payment Systems	174.04	2.550e+07	Financial Services	27	Na
8	Emerge Digital Group	170.64	2.390e+07	Advertising & Marketing	75	Sar
9	Goal Zero	169.81	3.310e+07	Consumer Products & Services	97	Blu
10	Yagoozon	166.89	1.860e+07	Retail	15	Wa
11	OBXtek	164.33	2.960e+07	Government Services	149	Ty
12	AdRoll	150.65	3.410e+07	Advertising & Marketing	165	Sar
13	uBreakiFix	141.02	1.700e+07	Retail	250	Or
14	Sparc	128.63	2.110e+07	Software	160	Ch
15	LivingSocial	123.33	5.360e+08	Consumer Products & Services	4100	Wa
16	Amped Wireless	110.68	1.430e+07	Computer Hardware	26	Ch
17	Intelligent Audit	105.73	1.450e+08	Logistics & Transportation	15	Ro
18	Integrity Funding	104.62	1.110e+07	Financial Services	11	Sar
19	Vertex Body Sciences	100.10	1.180e+07	Food & Beverage	51	col

There were 19 companies that experienced a growth rate of 100 or more.

Revenue

```
inc %>% dplyr::summarise(min=min(Revenue), median=median(Revenue), max=max(Revenue))
```

```
##     min   median      max
## 1 2e+06 10900000 1.01e+10
```

The revenue ranges from 2 million to about 10 billion. The median revenue is about 11 million.

Industries

```
kable(inc %>% dplyr::group_by(Industry) %>% dplyr::summarise(n=n()) %>% arrange(desc(n))) %>% kable_styl
```

Industry	n
IT Services	733
Business Products & Services	482
Advertising & Marketing	471
Health	355
Software	342
Financial Services	260
Manufacturing	256
Consumer Products & Services	203
Retail	203
Government Services	202
Human Resources	196
Construction	187
Logistics & Transportation	155
Food & Beverage	131
Telecommunications	129
Energy	109
Real Estate	96
Education	83
Engineering	74
Security	73
Travel & Hospitality	62
Media	54
Environmental Services	51
Insurance	50
Computer Hardware	44

There are 25 distinct industries.

Employees

```
kable(inc %>% dplyr::summarise(min=min(Employees, na.rm = TRUE), median=median(Employees, na.rm = TRUE)
```

min	median	max
1	53	66803

The number of employees range from 1 to 66,803. The median employee size is 53. In addition, there are some companies whom have a total number of employees as zero.

Cities

```
citi <- inc %>% group_by(City) %>% summarise(n=n())
nrow(citi)
```

```
## [1] 1519
```

There are 1,519 cities.

We can oprder them by the top 10 cities, based on the number of companies located there.

States

```
states <- inc %>% group_by(State) %>% summarise(n=n())
nrow(states)
```

```
## [1] 52
```

52 States are included in this Data Set

We can order this the same as the 10 cities with 10 states

```
kable(inc %>% group_by(State) %>% summarise(n=n()) %>% arrange(desc(n)) %>% top_n(10)) %>% kable_stylin
```

```
## Selecting by n
```

State	n
CA	701
TX	387
NY	311
VA	283
FL	282
IL	273
GA	212
OH	186
MA	182
PA	164

```
kable(inc %>% group_by(City) %>% summarise(n=n()) %>% arrange(desc(n)) %>% top_n(10)) %>% kable_styling
```

```
## Selecting by n
```

City	n
New York	160
Chicago	90
Austin	88
Houston	76
San Francisco	75
Atlanta	74
San Diego	67
Seattle	52
Boston	43
Dallas	42
Denver	42

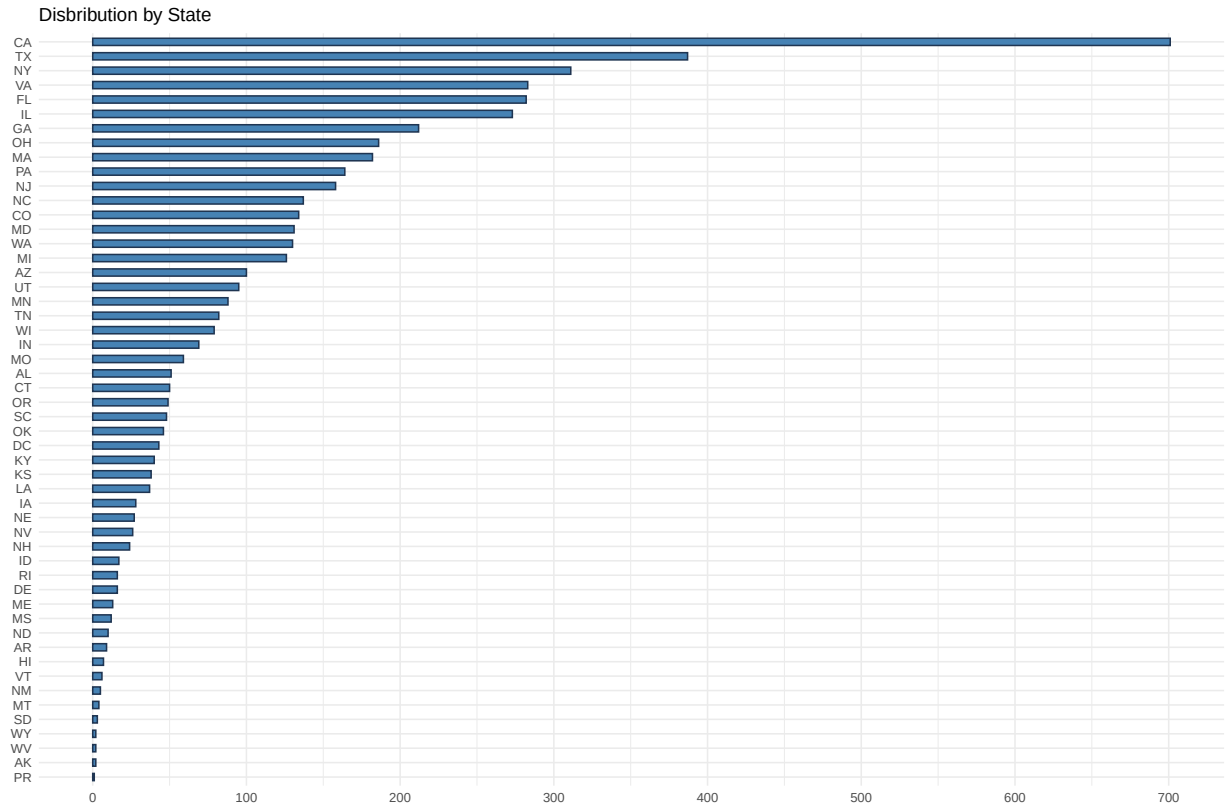
Question 1

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.

```
order_df <- inc %>% group_by(State) %>% summarise(n=n()) %>% arrange(desc(n))
```

```
plt <-
  ggplot(data = order_df[1:52,], aes(x=reorder(State,n), y=n)) +
  geom_bar(stat="identity", width=0.5, color="#1F3552", fill="steelblue",
           position=position_dodge()) +
    #geom_text(aes(label=round(n, digits=2)), hjust=1.3, size=3.0, color="white") +
    coord_flip() +
    scale_y_continuous(breaks=seq(0,700,100)) +
    ggtitle("Disbribution by State") +
    xlab("") + ylab("") +
    theme_minimal()
```

```
plt
```



Quesiton 2

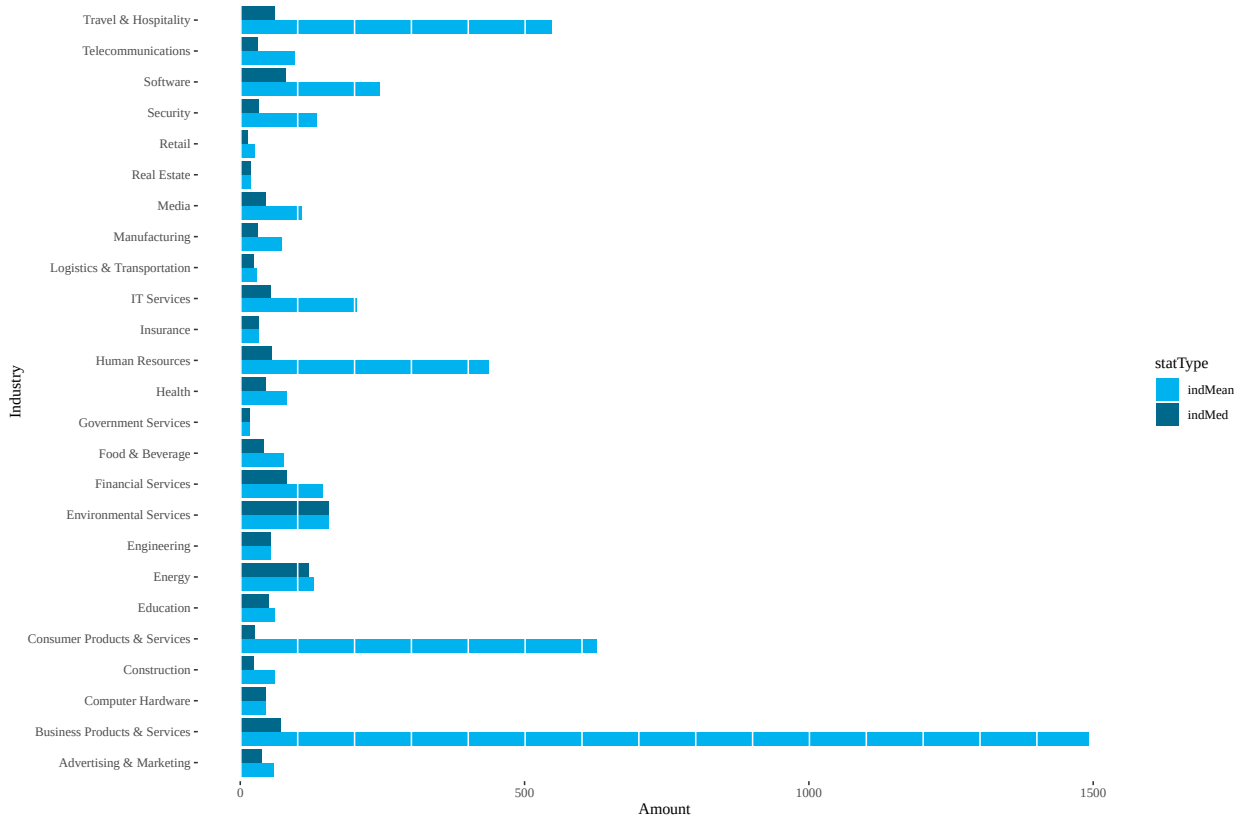
Lets dig in on the state with the 3rd most companies in the data set. Imagine you work for the state and are interested in how many people are employed by companies in different industries. Create a plot that shows the average and/or median employment by industry for companies in this state (only use 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.

```
df <- inc %>%
    group_by(State) %>%
    summarise(bizCount = n()) %>%
    arrange(desc(bizCount))
```

```
states2 <- toString(df$State[3])
```

```
df1 <- inc %>%
    filter(State == states2) %>%
    filter(complete.cases(.)) %>%
    group_by(Industry) %>%
    summarise(indMean = mean(Employees),
              indMed = median(Employees)) %>%
    gather(statType, Amount, indMean, indMed)
```

```
ggplot(data = df1, aes(x = Industry, y = Amount)) +
    geom_bar(stat = 'identity', aes(fill = statType), position = 'dodge') +
    scale_fill_manual(values = c('deepskyblue2', 'deepskyblue4'))+
    geom_hline(yintercept=seq(1, 1500, 100), col="white", lwd=0.5) +
    theme_tufte() +
    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.

```
df2 <- inc %>%
    filter(State == states2) %>%
    filter(complete.cases(.)) %>%
    mutate(RevPerEmp = (Revenue / Employees)/1000) %>%
    group_by(Industry) %>%
    summarise(Mean = mean(RevPerEmp))
```

```
ggplot(data = df2, aes(x = Industry, y = Mean)) +
    geom_bar(stat = 'identity', fill = "#FF6666") +
    theme_tufte()+
    theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
    geom_hline(yintercept=seq(1, 9000, 1000), col="forestgreen", lwd=0.5) +
    ylab('Revenue/Employee ,000 $')
```