Basic Data Visualization

Academic Integrity

I hereby declare that the following document is created by myself by strictly following St. Clair College's Academic Integrity policies.

R Version:

```
R.version.string
```

```
## [1] "R version 4.1.1 (2021-08-10)"
```

R Studio version

RStudio 2021.09.0+351 "Ghost Orchid" Release (077589bcad3467ae79f318afe8641a1899a51606, 2021-09-20) for Windows Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) QtWebEngine/5.12.8 Chrome/69.0.3497.128 Safari/537.36

List of R packages used and their versions:

```
packageVersion("readr")
```

```
## [1] '2.0.2'
```

```
packageVersion("tidyverse")
```

```
## [1] '1.3.1'
```

```
packageVersion("plotly")
```

```
## [1] '4.9.4.1'
```

About the Data set

Name of the data set used in this project - "*Forbes 2000*"

Owner: Forbes - An American Business magazine.

Link to dataset: https://vincentarelbundock.github.io/Rdatasets/csv/HSAUR/Forbes2000.csv (https://vincentarelbundock.github.io/Rdatasets/csv/HSAUR/Forbes2000.csv)

Documentation about the data set: https://vincentarelbundock.github.io/Rdatasets/doc/HSAUR/Forbes2000.html (https://vincentarelbundock.github.io/Rdatasets/doc/HSAUR/Forbes2000.html)

Summary about the data set: The Forbes 2000 list is a ranking of the world's biggest companies, measured by sales, profits, assets and market value for the year 2004.

Format of the dataset: A data frame with 2000 observations on the following 8 variables.

Description about each variables in the dataset:

rank: The ranking of the company.

name: The name of the company.

country: A factor giving the country the company is situated in.

category: A factor describing the products the company produces.

sales: The amount of sales of the company in billion USD.

profits: The profit of the company in billion USD.

assets: The assets of the company in billion USD.

marketvalue: The market value of the company in billion USD.

Data set reference: http://www.forbes.com/ (http://www.forbes.com/)

Modifications to the original data

No modifications were made to the original data set.

Working with the dataset

Load the csv data sheet for processing.

```
library(readr)
Forbes2000 <- read_csv("C:/Forbes2000.csv")
```

```
## New names:
## * `` -> ...1
```

```
## Rows: 2000 Columns: 9
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): name, country, category
## dbl (6): ...1, rank, sales, profits, assets, marketvalue
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Loading the tidyverse package to make use of the ggplot and other library functionality within it.

```
install.packages("tidyverse",repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/KBaska201/Documents/R/win-library/4.1'
## (as 'lib' is unspecified)
```

```
## package 'tidyverse' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\KBaska201\AppData\Local\Temp\RtmpGYVysv\downloaded_packages
```

```
#Used the `repos = "http://cran.us.r-project.org"` to avoid error while Knitting to HTML
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v dplyr   1.0.7
## v tibble  3.1.4     v stringr 1.4.0
## v tidyr   1.1.4     v forcats 0.5.1
## v purrr   0.3.4
```

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

A glimpse of the dataset helps to find out the number of observation and variables in it and to check which variables are continuous and categorical.

```
glimpse(Forbes2000)
```

```
## Rows: 2,000
## Columns: 9
## $ ...1        <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,~
## $ rank        <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,~
## $ name        <chr> "Citigroup", "General Electric", "American Intl Group", "E~
## $ country     <chr> "United States", "United States", "United States", "United~
## $ category    <chr> "Banking", "Conglomerates", "Insurance", "Oil & gas operat~
## $ sales       <dbl> 94.71, 134.19, 76.66, 222.88, 232.57, 49.01, 44.33, 135.82~
## $ profits     <dbl> 17.85, 15.59, 6.46, 20.96, 10.27, 10.81, 6.66, 7.99, 6.48,~
## $ assets      <dbl> 1264.03, 626.93, 647.66, 166.99, 177.57, 736.45, 757.60, 1~
## $ marketvalue <dbl> 255.30, 328.54, 194.87, 277.02, 173.54, 117.55, 177.96, 11~
```

Assigning to a shorter name for convenience and ensuring data is same.

```
f2000 <- Forbes2000
glimpse(f2000)
```

```
## Rows: 2,000
## Columns: 9
## $ ...1         <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,~
## $ rank         <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,~
## $ name         <chr> "Citigroup", "General Electric", "American Intl Group", "E~
## $ country      <chr> "United States", "United States", "United States", "United~
## $ category     <chr> "Banking", "Conglomerates", "Insurance", "Oil & gas operat~
## $ sales        <dbl> 94.71, 134.19, 76.66, 222.88, 232.57, 49.01, 44.33, 135.82~
## $ profits      <dbl> 17.85, 15.59, 6.46, 20.96, 10.27, 10.81, 6.66, 7.99, 6.48,~
## $ assets       <dbl> 1264.03, 626.93, 647.66, 166.99, 177.57, 736.45, 757.60, 1~
## $ marketvalue <dbl> 255.30, 328.54, 194.87, 277.02, 173.54, 117.55, 177.96, 11~
```

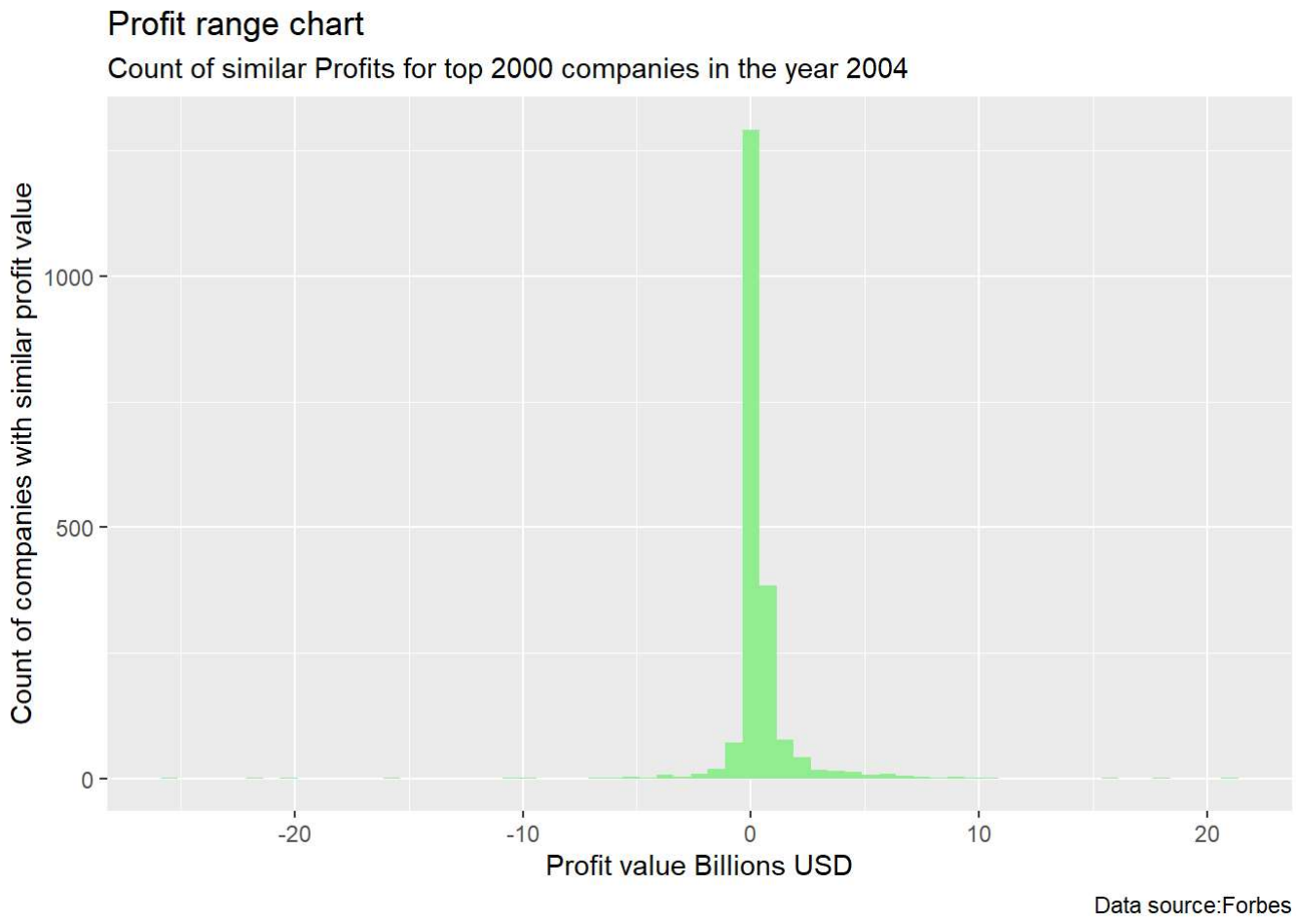
9 plots

Q1: ○ Two plots displaying the distribution of a single continuous variable

- The below graph displays the overall count of companies with similar profit levels in the data set.

```
q1 <- ggplot(f2000, aes(profits))
q1 + geom_histogram(binwidth = 0.75, fill="lightgreen") + labs(x="Profit value Billions USD", y=
"Count of companies with similar profit value", subtitle="Count of similar Profits for top 2000
 companies in the year 2004", title="Profit range chart", caption = "Data source:Forbes")
```

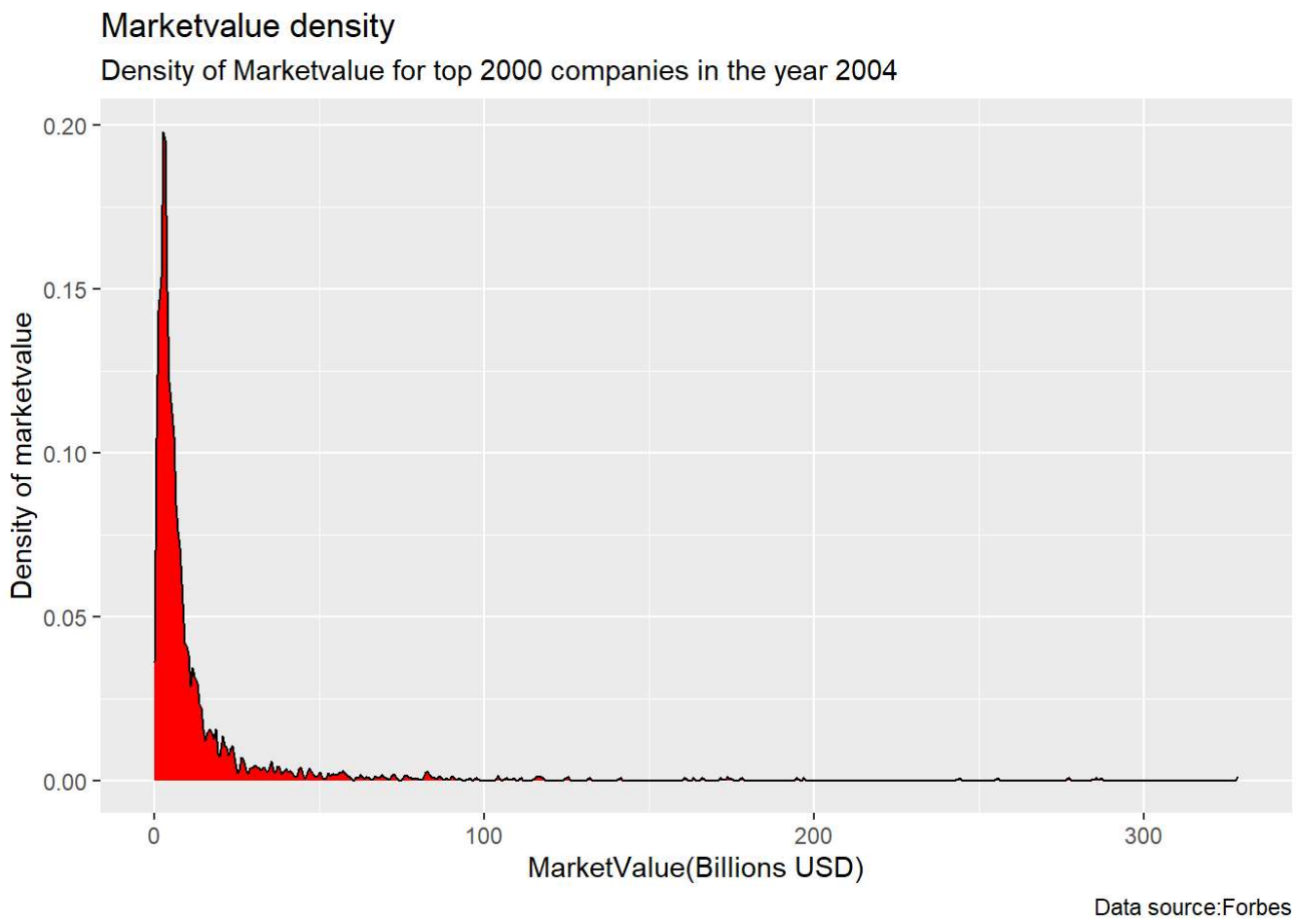
```
## Warning: Removed 5 rows containing non-finite values (stat_bin).
```



From the graph it is clear that most companies have profit level less that 2 billion and only few companies have profit more that 5 billion, also could see a some companies with negative profit levels.

- The below graph displays the density plot of marketvalue of all the top 2000 companies from the data set.

```
q1_1 <- ggplot(f2000, aes(marketvalue))
q1_1 + geom_density(adjust=0.15, fill="red") + labs(x="MarketValue(Billions USD)", y="Density of
marketvalue", subtitle="Density of Marketvalue for top 2000 companies in the year 2004", title=
"Marketvalue density", caption = "Data source:Forbes")
```

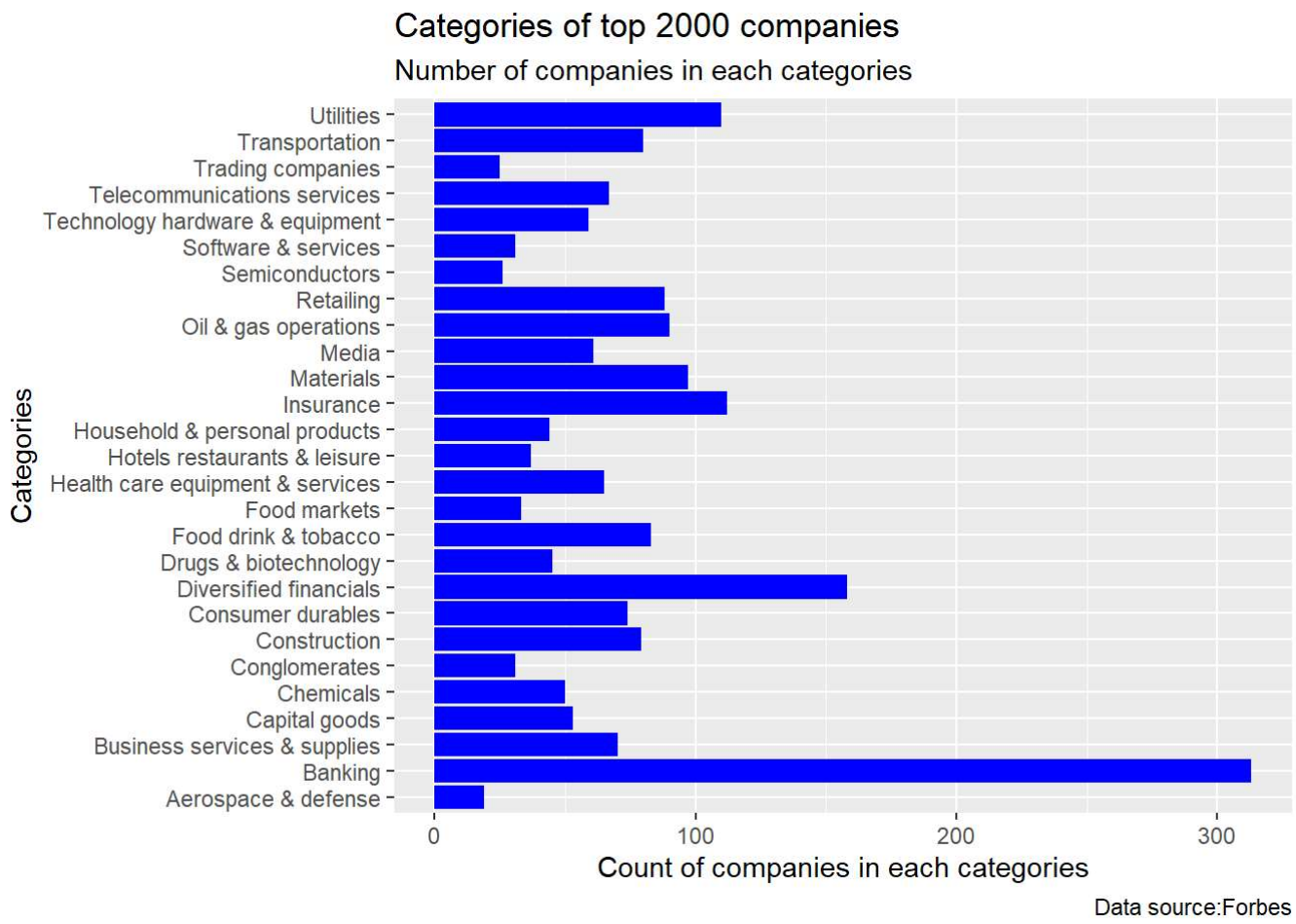


The graph clearly shows us most companies has the market value less than 50 billion USD.

Q2: ○ Two plots displaying information about a single categorical variable

- This plot describes the count of companies in each categories.

```
q2 <- ggplot(f2000, aes(x=category))
q2 + geom_bar(fill="blue") + labs(x="Categories", y="Count of companies in each categories", sub
title="Number of companies in each categories", title="Categories of top 2000 companies", captio
n = "Data source:Forbes") + coord_flip()
```

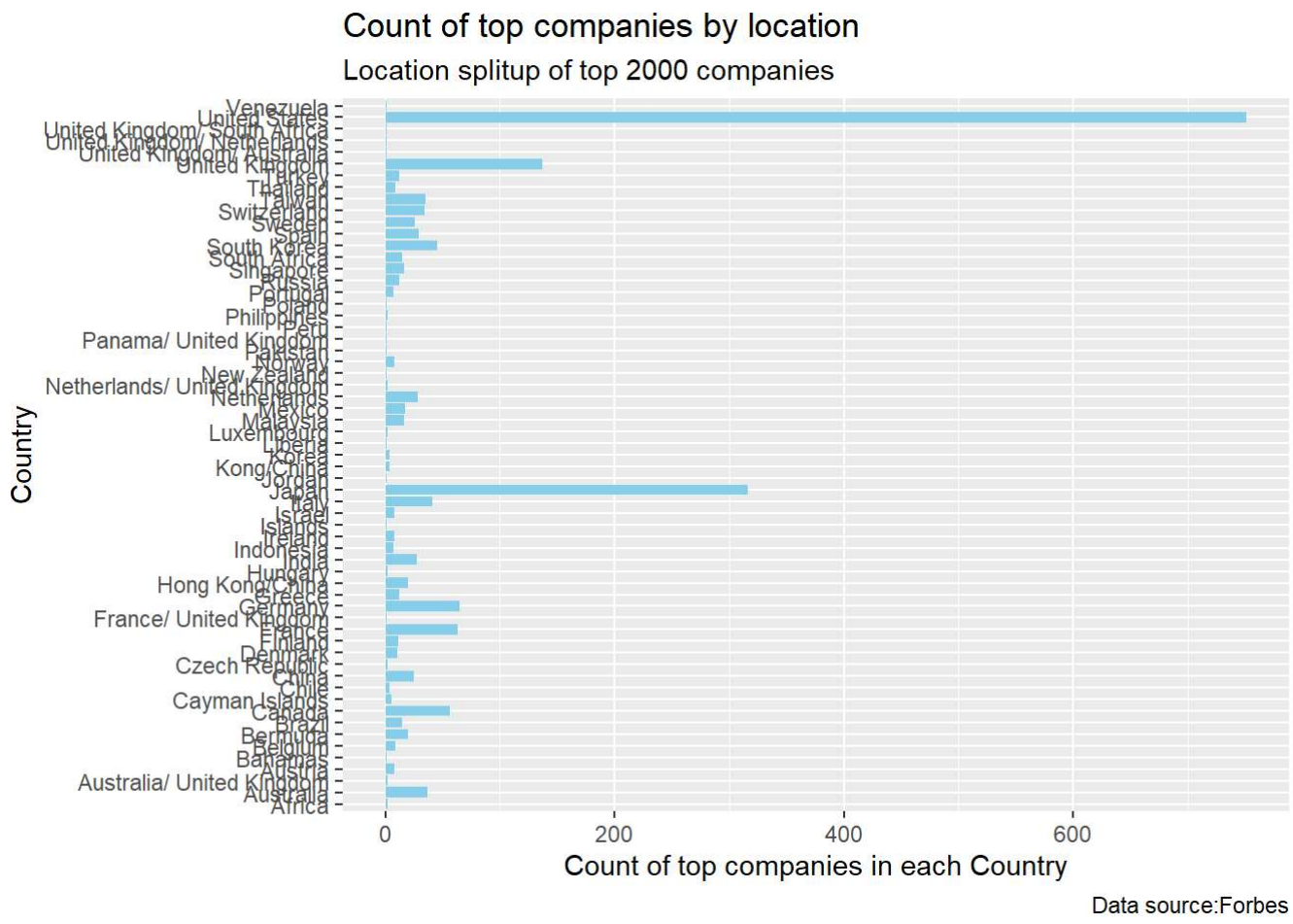


On looking at the graph it is clear that Banking sector contributes to the most number of companies from the data set.

- The below shows the location wise distribution of the top 2000 companies.

```
q2_1 <- ggplot(f2000, aes(x=country))
q2_1 + geom_histogram(fill="skyblue", stat="count") + labs(x="Country", y="Count of top companie
s in each Country", subtitle="Location splitup of top 2000 companies", title="Count of top compa
nies by location", caption = "Data source:Forbes") + coord_flip()
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

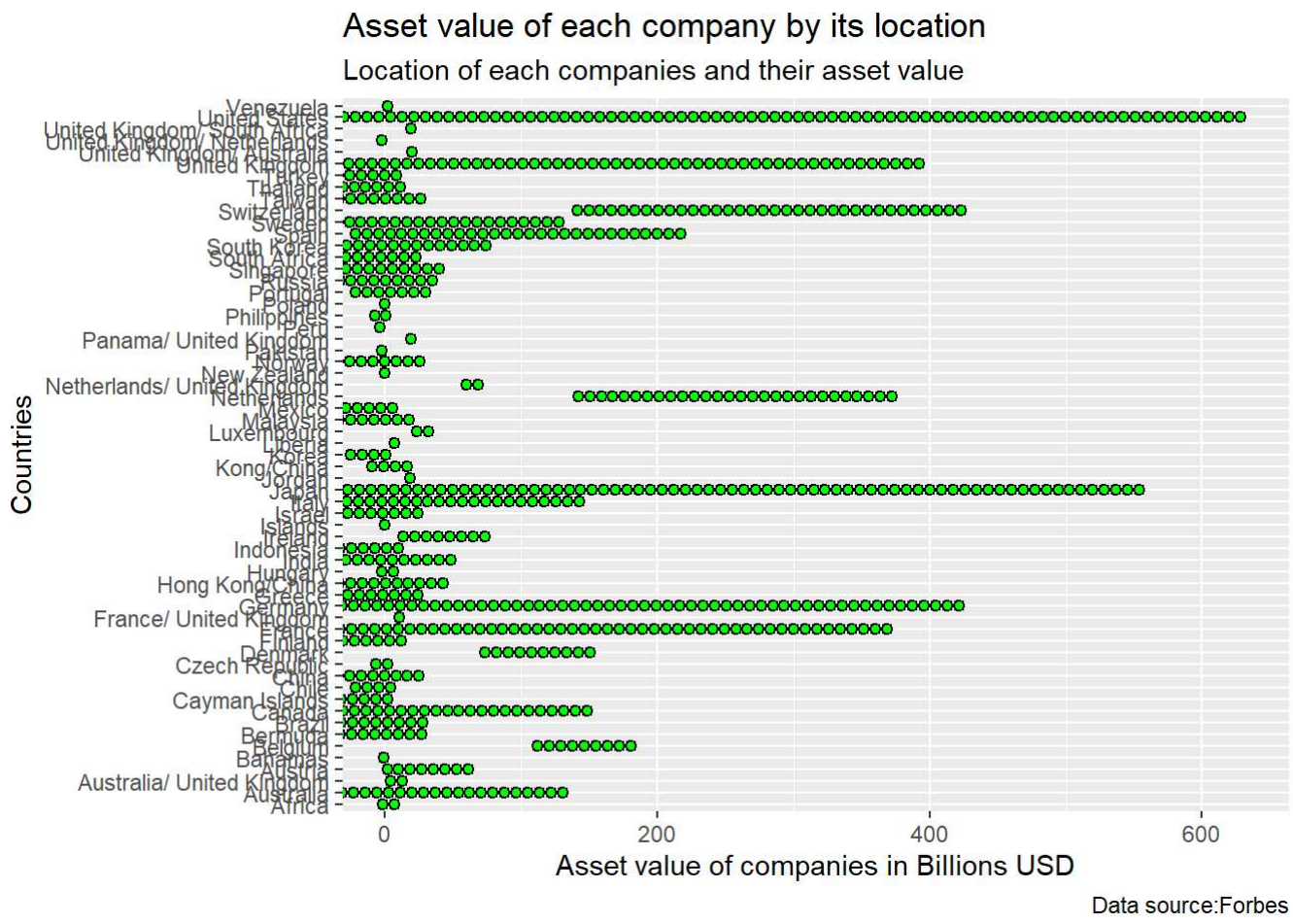


It is clear from the graph that United states has the most number of companies in the data set followed by Japan.

Q3: ○ One plot displaying information about both a continuous variable and a categorical variable

- This plot represents the asset value of each company and its location where it is situated.

```
q3 <- ggplot(f2000, aes(assets, country))
q3 + geom_dotplot(binaxis = "y", stackdir="down", binwidth = 1, fill="green") + labs(x="Asset va
lue of companies in Billions USD", y="Countries", subtitle="Location of each companies and their
asset value", title="Asset value of each company by its location", caption = "Data source:Forbe
s")
```

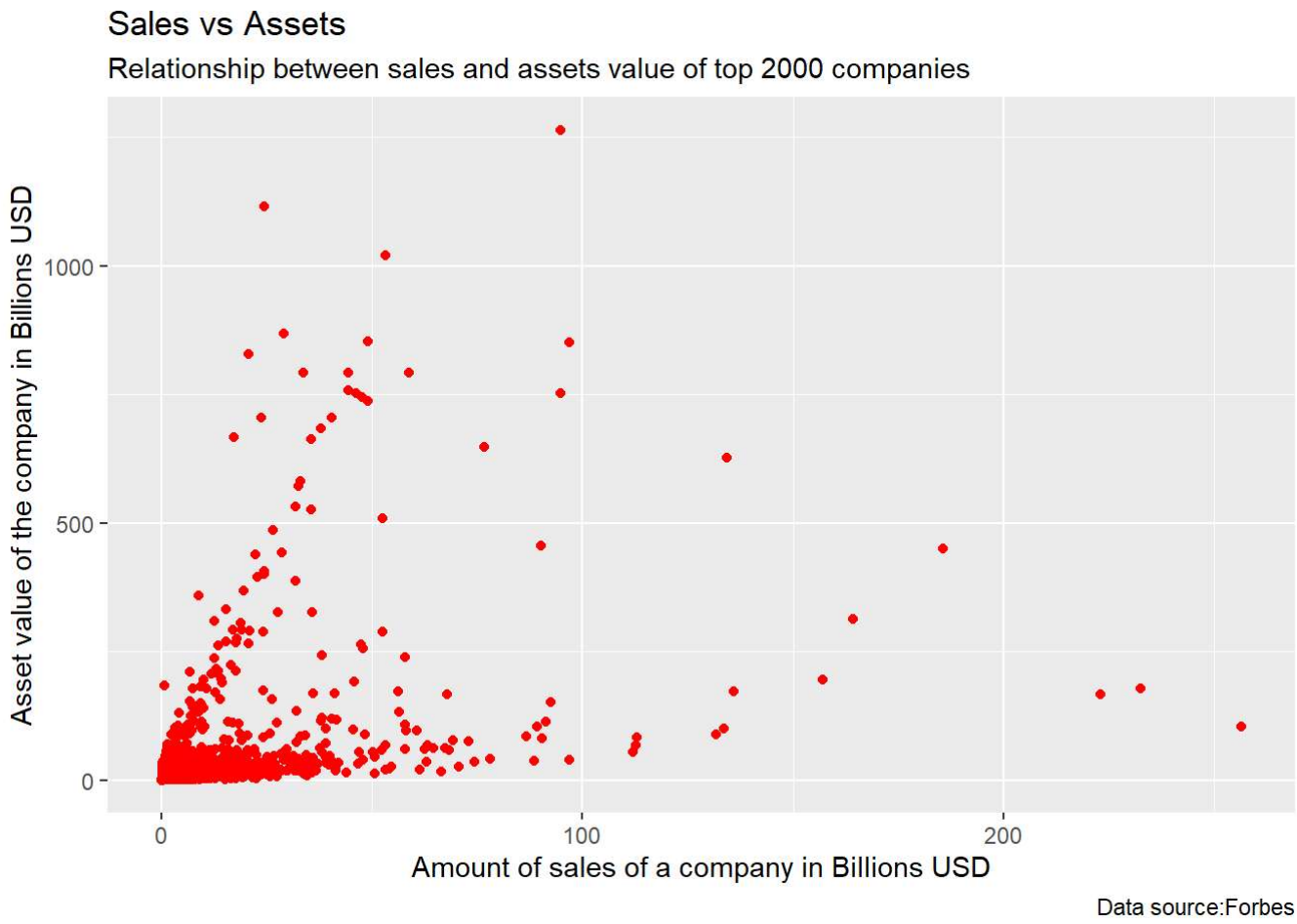


Looks like the United states has companies in all asset values.

Q4: ○ Two plots should display information that shows a relationship between two variables

- The below plot describes the relation between Sales and Asset value of the top 2000 companies.

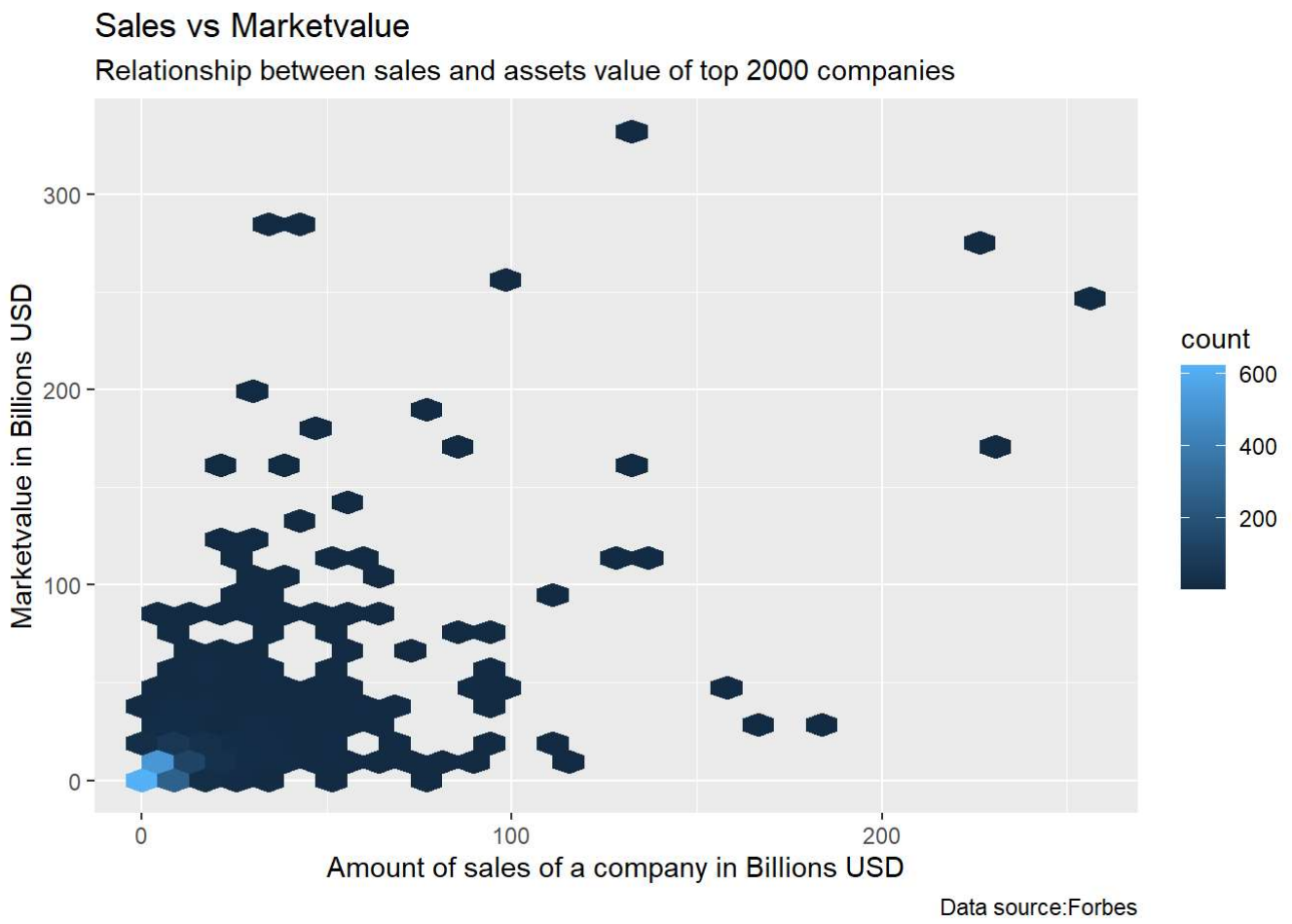
```
q4 <- ggplot(f2000, aes(sales, assets))
q4 + geom_point(color="red") + labs(x="Amount of sales of a company in Billions USD", y="Asset v
alue of the company in Billions USD", subtitle="Relationship between sales and assets value of t
op 2000 companies", title="Sales vs Assets", caption = "Data source:Forbes")
```



Looks like even if the sales value is higher only a little amount of companies has a higher asset values.

- This hexplot graph gives us a clear view of the number of companies which has good sales and marketvalue.

```
q4_1 <- ggplot(f2000, aes(sales, marketvalue))
q4_1 + geom_hex() + labs(x="Amount of sales of a company in Billions USD",y="Marketvalue in Bill
ions USD", subtitle="Relationship between sales and assets value of top 2000 companies", title=
"Sales vs Marketvalue", caption = "Data source:Forbes")
```

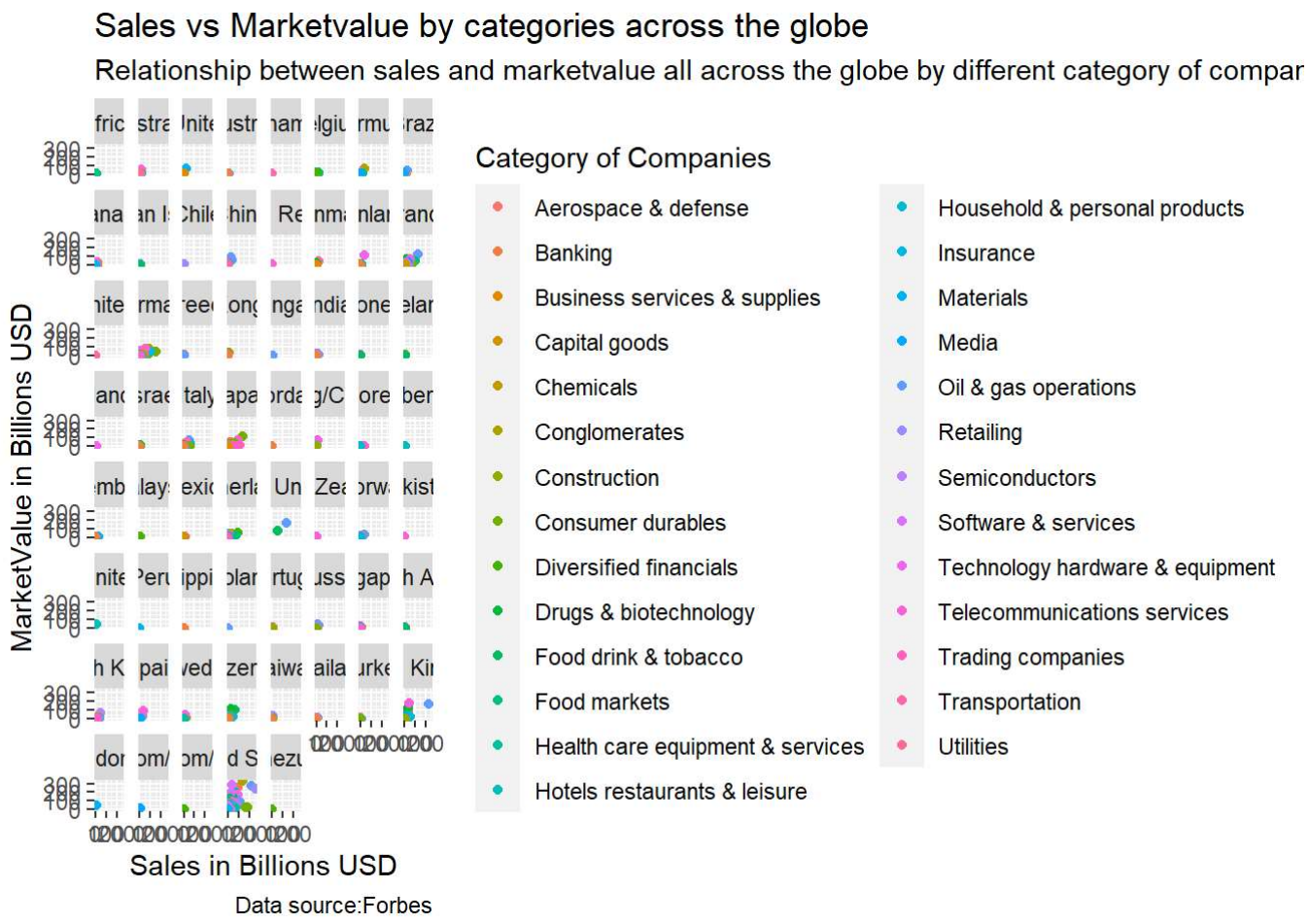


From the plot it looks like there are more companies with lesser sales and market values when compared with companies with higher sales and marketvalue.

Q5: One plot should use faceting and display information about 4 variables

- The below graphs explains about the category of companies in each country and its correlation between its sales and marketvalue.

```
ggplot(f2000, aes(sales, marketvalue, color=category)) + geom_point() + facet_wrap(~country) + l
abs(x="Sales in Billions USD", y="MarketValue in Billions USD", color="Category of Companies", t
itle="Sales vs Marketvalue by categories across the globe", subtitle = "Relationship between sal
es and marketvalue all across the globe by different category of companies", caption = "Data sou
rce:Forbes")
```



From the graph it is clear that United States has the highest number of companies with higher sales and market value followed by Japan and so on.

Q6: ○ creative plot: an opportunity to explore what's possible and get creative Installing Plotly library to create a interactive graph.

```
install.packages("plotly",repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/KBaska201/Documents/R/win-library/4.1'
## (as 'lib' is unspecified)
```

```
## package 'plotly' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\KBaska201\AppData\Local\Temp\RtmpGYVysv\downloaded_packages
```

```
#Used the `repos = "http://cran.us.r-project.org"` to avoid error while Knitting to HTML
library(plotly)
```

```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

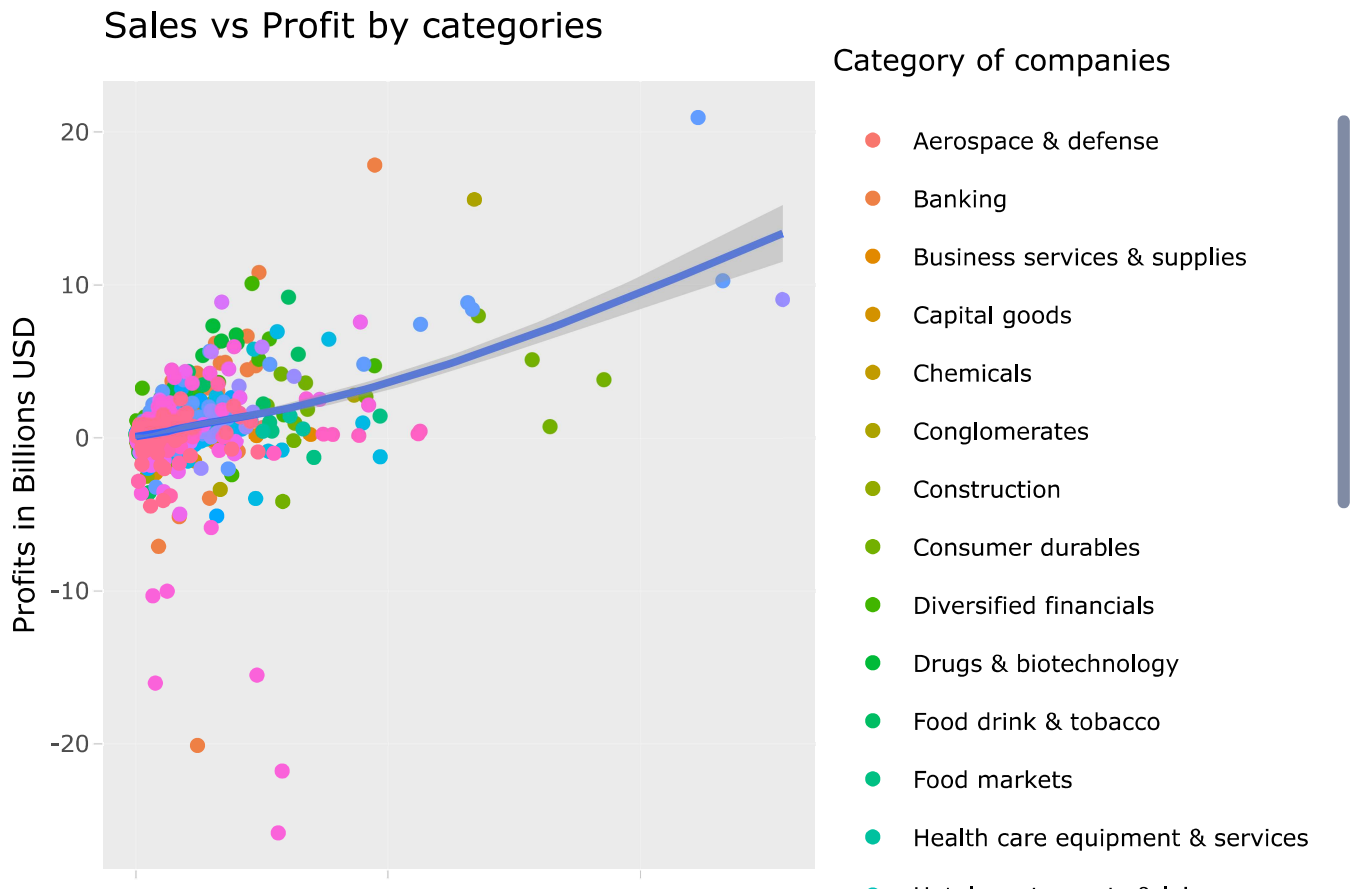
```
## The following object is masked from 'package:graphics':
##
##     layout
```

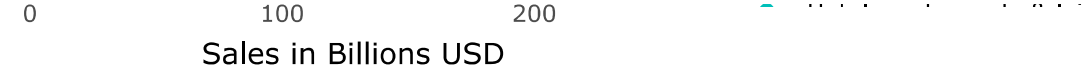
- This interactive plot helps us to understand the relation between sales and profits for all the companies in the data set.

```
q6 <- ggplot(f2000, aes(sales, profits))
q6_1 <- q6 + geom_point(aes(color=category)) + geom_smooth() + labs(x="Sales in Billions USD",y="Profits in Billions USD", color="Category of companies" , title="Sales vs Profit by categories", caption = "Data source:Forbes")
ggplotly(q6_1)
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 5 rows containing non-finite values (stat_smooth).
```





From the graph it is clear that even if there is good amount of sales the profit levels of some companies are negative and in general as the smooth plot describes most of the companies with good sales have a upward profit trend.

References

https://ggplot2.tidyverse.org/index.html (https://ggplot2.tidyverse.org/index.html)

https://github.com/rstudio/cheatsheets/blob/master/data-visualization-2.1.pdf (https://github.com/rstudio/cheatsheets/blob/master/data-visualization-2.1.pdf)

https://plotly.com/ggplot2/line-and-scatter/ (https://plotly.com/ggplot2/line-and-scatter/)

https://plotly.com/r/plotly-fundamentals/ (https://plotly.com/r/plotly-fundamentals/)

https://stackoverflow.com/questions/33969024/install-packages-fails-in-knitr-document-trying-to-use-cran-without-setting-a (https://stackoverflow.com/questions/33969024/install-packages-fails-in-knitr-document-trying-to-use-cran-without-setting-a)

Q&A

■ In what ways do you think data visualization is important to understanding a data set?

- Data visualization helps to convey information in simpler manner. It also helps to find correlation between different variables. Visualization helps to communicate information faster and easier than explaining it by other means.

■ In what ways do you think data visualization is important to communicating important aspects of a data set?

- Data visualization helps to communicate key factors with ease to any person from any background. It helps to picture the complete data in simpler form without having to explain anything about it. It also helps us to find relationship between multiple variables to get a different understanding about the available data.

■ What role does your integrity as an analyst play when creating a data visualization for communicating results to others?

- Self integrity plays a major role in maintaining the transparency of data to the audience. A slight modification to any of the data might create a serious impact to the overall integrity, so maintaining a strict integrity is essential for every data analyst. A good visualization should convey the right information without distorting the original facts.

■ How many variables do you think you can successfully represent in a visualization? What happens when you exceed this number?

- I believe we can represent a maximum of 4 variables clearly in a visualization and when this number exceeds the graph becomes overwhelming with information and thus becomes difficult for the audience to understand. In simple terms visualization should not be overloaded with information.