JIGSAW ACADEMY
THE ONLINE SCHOOL OF ANALYTICS

Data Visualization Case Study

Use the data from the HTML page "The World's Most Valuable Brands List - Forbes.html" saved at (**/Data Science R/Assignments/Graded Assignments/Topic 6.2 Data Visualization in R**). You will need to read the data pertaining to "Most Valuable Brands" from this HTML file and create a dataframe out of it using appropriate functions and libraries.(Use the library XML, RCurl)

Once the data is in a data.frame format, you will need to create a visualization for 4 sectors:

(a) Technology

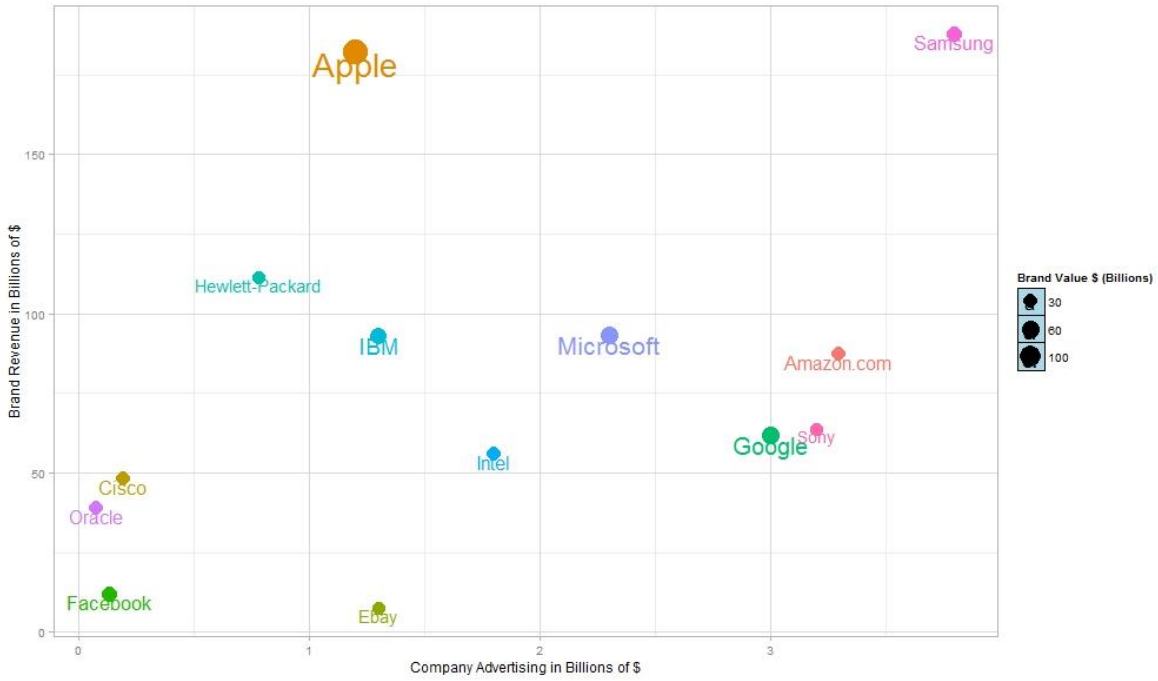
(b) Luxury

(c) Automotive

(d) Financial Services

The visualizations should look like the ones below:

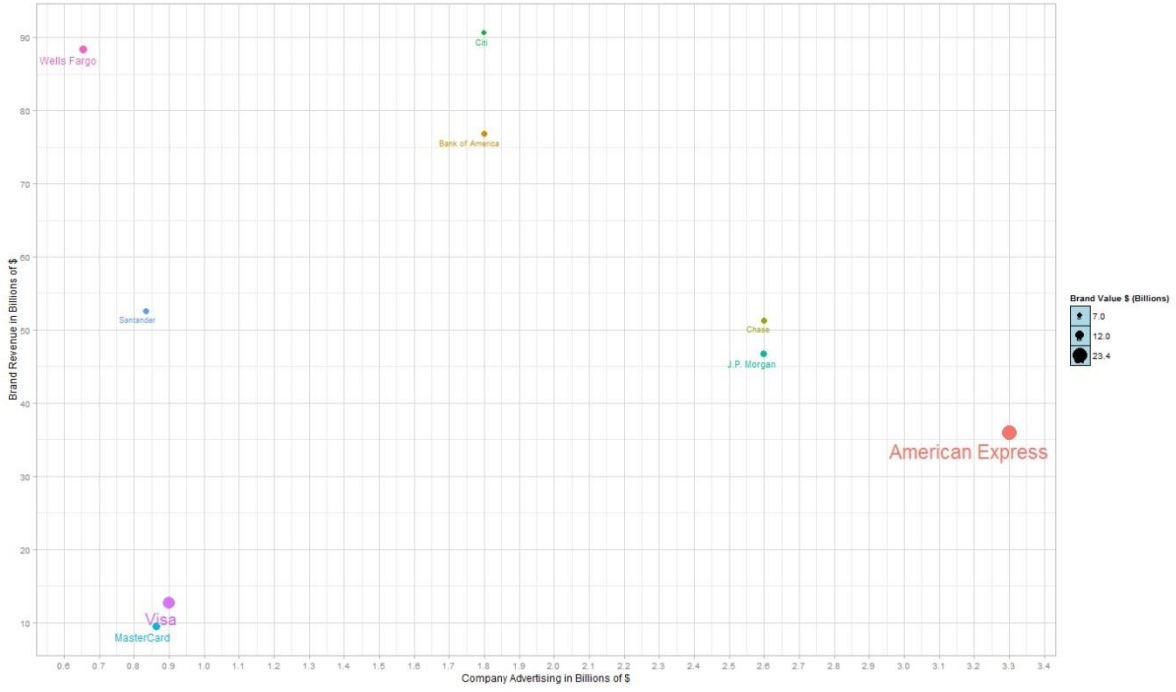
Technology



JIGSAW ACADEMY
THE ONLINE SCHOOL OF ANALYTICS

Luxury

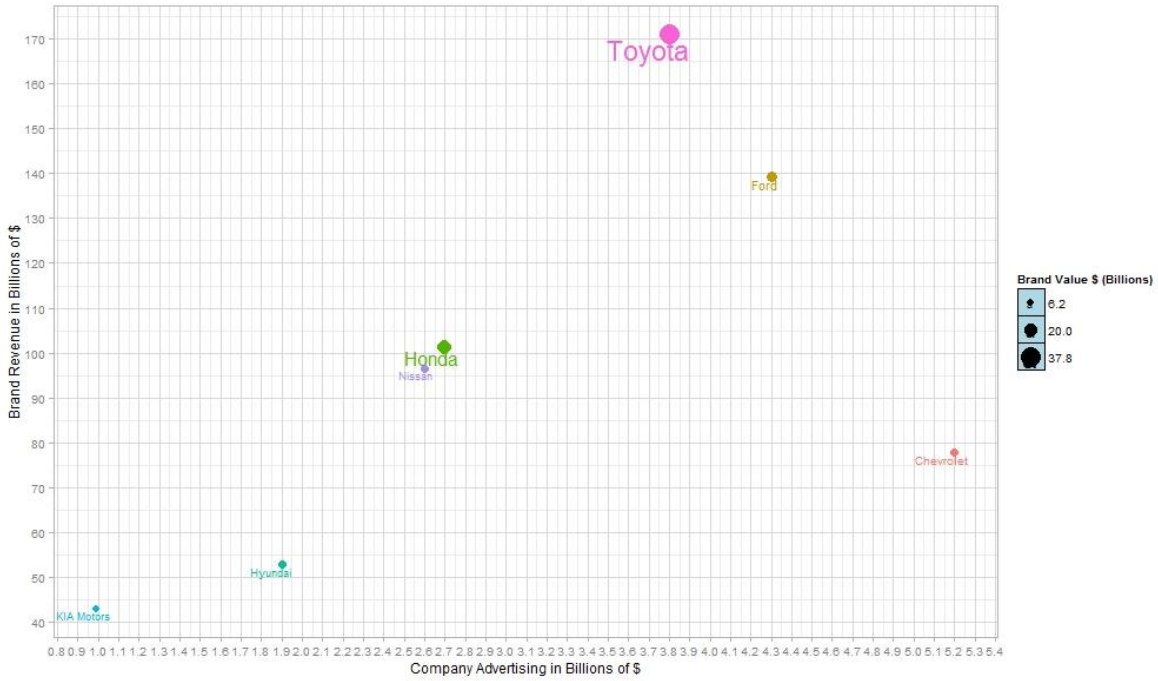
Financial



JIGSAW ACADEMY

THE ONLINE SCHOOL OF ANALYTICS

Automotive



Notes and Hints to the assignment:

(i) The data source is an html file, you can use relevant functions in libraries XML or RCurl to read the html file inside R and parse the appropriate table.

(ii) The html table inside this file has information on top 100 brands by value

(iii) Once you are successful in parsing this table and converting it into a dataframe, make sure that all the columns have appropriate data types. Eg, Column such as "Brand Value" will by default be a column of type "factor". The reason is the presence of symbols such as "$" and " B". You will need to remove the unnecessary symbols such as "#", "$", "B" and "M" etc.

(iv) To clean the columns you will have to rely on string manipulation functions such as sub, gsub, grepl etc. (You can do this cleaning task outside of R also if you are comfortable in tools such as excel, VB etc)

(v) While cleaning some of the columns, keep in mind for some rows (this is true only for certain columns) the monetary values are recorded in Billions of dollars as well as Millions of dollars (symbols B and M refer to Billion and Million respectively). You will need to keep this in mind so that all the values in a column are on same scale after you clean the data.

(vi) The sample graphs above have been created using ggplot2(). First understand the grammar behind these graphs before you set out to write the code. Also notice these sample graphs are highly modified and by default you will not get a similar visualization.

(vii) Some of the defaults that have been changed while creating these graphs are:

(a) Theme (http://docs.ggplot2.org/dev/vignettes/themes.html) Refer to the link to learn more about themes in ggplot2

(b) Legends (Some legends have been suppressed and some have been modified, refer to this link for more details, http://www.cookbook-r.com/Graphs/Legends_(ggplot2)/)

(c) X axis and Y axis Scales (Change of X and Y axis scales can be done using scale_x_continous and scale_y_continous functions in ggplot2)

(d) Text annotations (http://docs.ggplot2.org/current/geom_text.html)

Make sure you are comfortable with following topics before you attempt the assignment:

(i) Data import/export from web sources (particularly html files)

(ii) Data manipulation (particularly string manipulation)

(iii) Data visualization (Grammar of Graphics)

Following links on grammar of graphics are excellent resources:

https://ramnathv.github.io/pycon2014-r/visualize/ggplot2.html

http://ggplot2.org/