International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August, 2019

Automation of Longitudinal Time Series Data Visualization using Plot Specifications obtained from XL file

S.R. Sneha, G. Rohini, Kamal Srinivasan

Abstract*: Longitudinal Time Series data visualization plays important role in all sector of business decision making [9]. With enormous amount of complex data [11] from cloud and business requirement, number of graphs needed for decision making increased many folds. Generating enormous number of plots manually with more human input is tedious, time consuming and error prone. To avoid these issues, suitable visualization techniques with solid design principles become very important. We conceptualized and designed a novel method for automation of these processes. R-GGPLOT2[7] package and XL specifications file were primarily used to achieve this goal. We here show as how we can create multiple plots from time series data, plots specifications-XL file and R package GGPLOT2[7] in a single run. Since all required information are entered in XL sheet, R function can be run with no modification. Multiple plots can be generated by using enormous data available in production and service sectors such as finance, healthcare, transportation and food industries etc.

***Keywords*: Automation, Data visualization, Excel plot generation, GGPLOT2, R.**

I. INTRODUCTION

Time series data for example, stock/bond/commodities/currency, price over time fluctuates lot in every minute leading to generation of huge data/day. This data has enormous categorical information such as portfolio, name of stock/bonds and performance, earning/price ratio etc. Similarly, all other service and production industries have enormous amount of data stored in cloud-based databases. Due to advancement in cloud-based databases, data is stored in well-organized way with ease of accessibility to permitted users. These huge data are very resourceful to make correct decision in fast pace environment. Decision can be made by statistics obtained from data or visualization. However, the benefits of using visualization such as scatter, series, box plot, pie chart, 2 and 3 panel plots, 3 D plots, regression plots and availability of powerful open source software packages are superior over just tables based on statistics in many ways [4]. Precisely, in the data communication, graphics are powerful because human can remember images well, however human tend to fail to recollect numbers. They can understand pictures better than tables that contain rows and columns. Moreover, graphs provide patterns, trends, exceptions in data and reveal relationships among whole sets of values [2]. There are tools available online for data visualization. However, these are confined with limited options making customization impossible. Due to complexity of data [11] and requirements, it is more advantageous to develop tools in house that makes customization possible. Customization such as change between different types of visualization, color, symbols, groups, different data input and graph output formats. All these can be done in user friendly, less or no code intensive methods.

Some of the broad features of visualization include: Selecting features: to highlight specific features of data which has high influence on output. Explore: Explore selected features and analyze further by changing various options. Reconfigure: to re-arrange the data in the visualization. Encode: to change the type of visualization. Abstract/elaborate: to dig deeper and reveal more details, or to abstract or hide details. Filter: to restrict what is displayed in the data visualization according to certain conditions, such as a time range or a specific region within a map. Connect: to reveal relationships between items or features in the visualization [1].

User friendly Automation of visualization are very important wherever big data is dealt. Generation of graphs one by one is not efficient and involves more human modification which is error prone and time consuming. Automation and scalability in generating multiple plots in single run avoid these problems. It makes many tasks easier, such as: selecting various features of data, aesthetics and geometrics of graphs, title, footnote, formatting and saving tasks [6]. This article conceptualises, design and illustrates the detailed analysis as how to communicate data effectively for decision making with full automation right from XL sheet [5]. It is organized as follows: Section 2 describes about the general applications of data visualization in various sectors. Sections 3 deals with methodology used to plot multiple graphs from XL sheet and data. Section 4 explains about conclusion and results.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author
S.R.Sneha, Student, Department of Computer Science and Engineering, St. Joseph's Institute of Technology, OMR, Chennai – 119 India.
G.Rohini, Professor, Department of Electronics and Communication Engineering, St. Joseph's Institute of Technology, OMR, Chennai – 119, India.
Kamal Srinivasan, Senior Statistical Programmer, Regeneron Pharmaceuticals, Tarrytown, New York, USA..

II. APPLICATIONS OF DATA VISUALIZATION

Data visualization plays major role in all kind of prominent sectors. As an illustration, this paper demonstrates how principled visualization algorithms can be used to understand and explore a large data set created in the early stages of drug discovery in health care sector, prediction in stock market, productivity in agricultural field, selling of food products details, growth rate in automobile industries and ranking of education institutions etc.

In pharma industry, during drug discovery process, time series data of drug concentration in animals and patients are represented as various type of multiple plots to know the efficacy of the drug. Modeling and simulation in this process also rely on graphs. In this sector, many graphs are generated for regulatory submission for drug approval. All these can be achieved just by entering required plot specifications in XL sheet.

For past two decades, middle class, house hold, and retired people actively involved in share market operations. In such a transaction, investments are made with the help of rumors, own assumptions and other decisions without proper data. Because of less market knowledge and insufficient data, investors are not able to understand the high market volatility, stock market crash, lack of capital investment skills. Hence appropriate stock selections need data visualization as a tool as most of market data are available freely online.

In education sector, the ability to access student progress during the semester which then alerts advisors to intervene promptly without reach to students who are under performing. The plot representation of student grades and other metrices can easily identify the performance leading to timely investigation on underperforming students. Graphs can generate insights that help transform programs, curriculum, student outcomes and more in ways that deliver desired results faster.

III. METHODOLOGY

All specifications required for generating graphs are mentioned in XL file and any part of plot can be manipulated straight from XL file. These specifications supply all parameters to R-GGPLOT [7,8,3]. Each row of the XL file represents one plot with all parameters and index number. Users can select parameters from list below in XL file. It can be scatter, series, mean or box plots. Data file has all required parameters. In R function [7,8], user needs to input only range of plot or specific plots in accordance with index number of corresponding plots from XL file and then one single click run. This R function runs in a loop taking specifications mentioned in the XL file and data from data file. Individual or multiple plots are produced as entered in for loop of R function. So, this function can handle scalability to many numbers of plots. After choosing specifications for the plots from XL [5] file and input data, simply sourcing and running R code can produce plot in docx and pdf. format. Docx file can be opened as MS office-word file [10]. This function also renders executing program directory, name of R user, date and R-version as footnote. Title and footnote are printed as editable text which can be post processed.

Steps in method:

- Define all specifications mentioned below for plot in XL file.
- Prepare the data file according to XL file.
- Specify index range or specific index numbers in R-GGPLOT2 function.
- Run.

IV. RESULTS AND DISCUSSION

Content of XL specifications (Options)

Name = output name
Tit = title of the plot
Fot = footnote
Grp = grouping of categorical data
Clr = color of the legend and data
X, Y = X and Y coordinate values
Sbset = sub setting condition
Scl = scale of the axis (linear or log scale)
Xlabel, Ylabel = X and Y axis label
Type = Scatter or series plot or box plot
Indata = Input data directory
Inxlspec = Input XL specification file directory
Output= output data directory
Format = output file formats, pdf or docx.

The above-mentioned details are depicted in Table- I.

Using this methodology, around 150 plots can be visualized at a time. Here we have considered the following series (Fig.1), scatter (Fig. 2), mean (Fig. 3) and box (Fig. 4) plots which are included in XL speciation file and R-GGPLOT 2 function. However, we can scale up too many different types and number of plots after incorporating relevant specifications in XL file and modifications in R function.

Fig.5 represents the flow of various process carried out in this work.

Table- I: XL Parameters specification file

Index	name	tit	fot	grp	clr	x	y	sbset	scl	xlabel	ylabel	geoms	type	indata	inxlspec	output	format
1	p1	price	note	stock	stock	time	price	stock == 'ms'	lg	days	price	series	mn	path	path	path	pdf
2	p2	sale	note	stock	stock	time	price	stock == 'fb'	ln	days	price	scatter	ind	path	path	path	docx
3	p3	value	note	stock	stock	stock	price		ln	stock	price	box	box	path	path	path	pdf

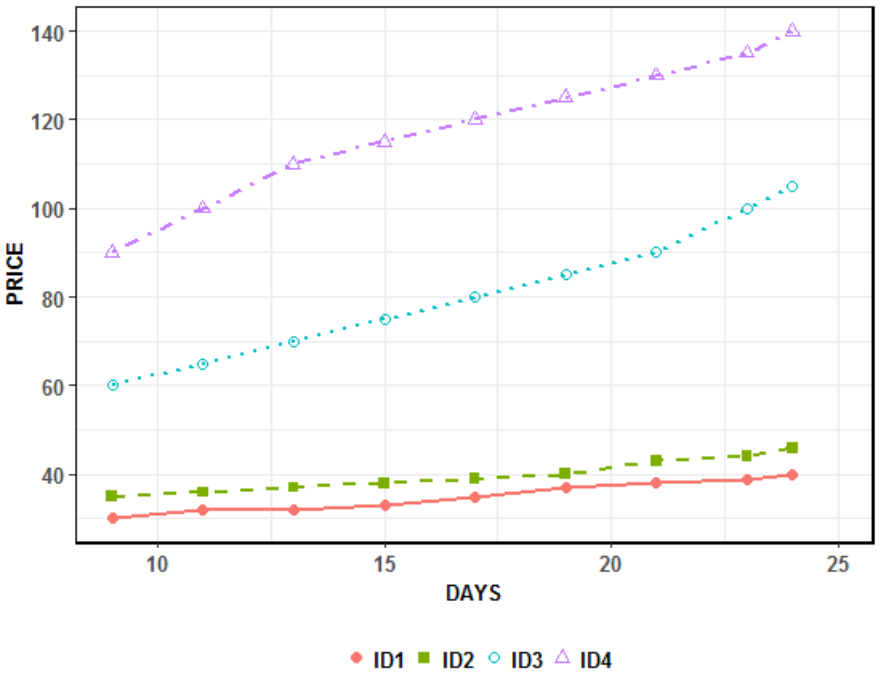

Fig.1. Output Series Plot

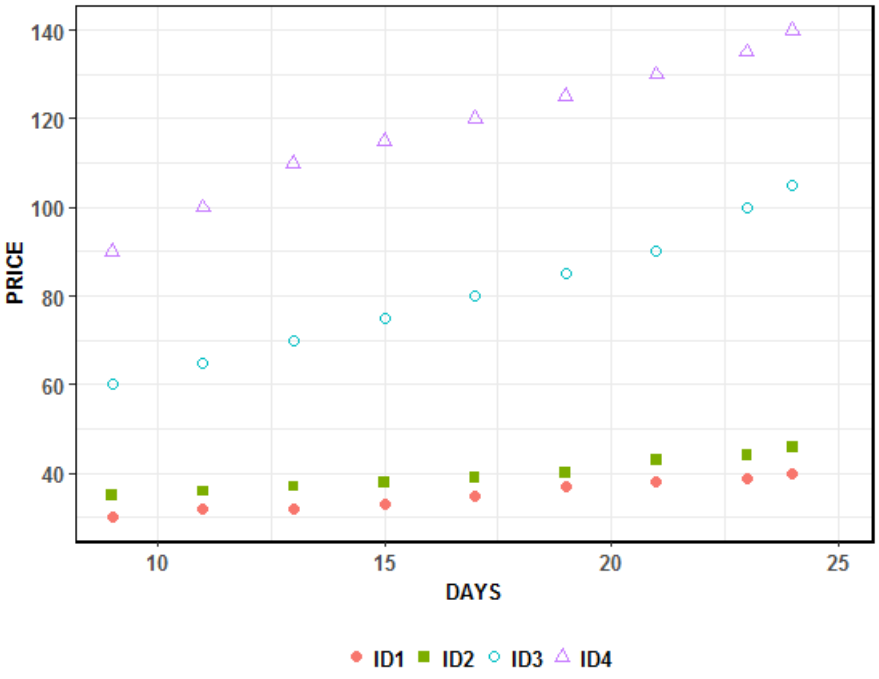

Fig. 2. Output Scatter Plot

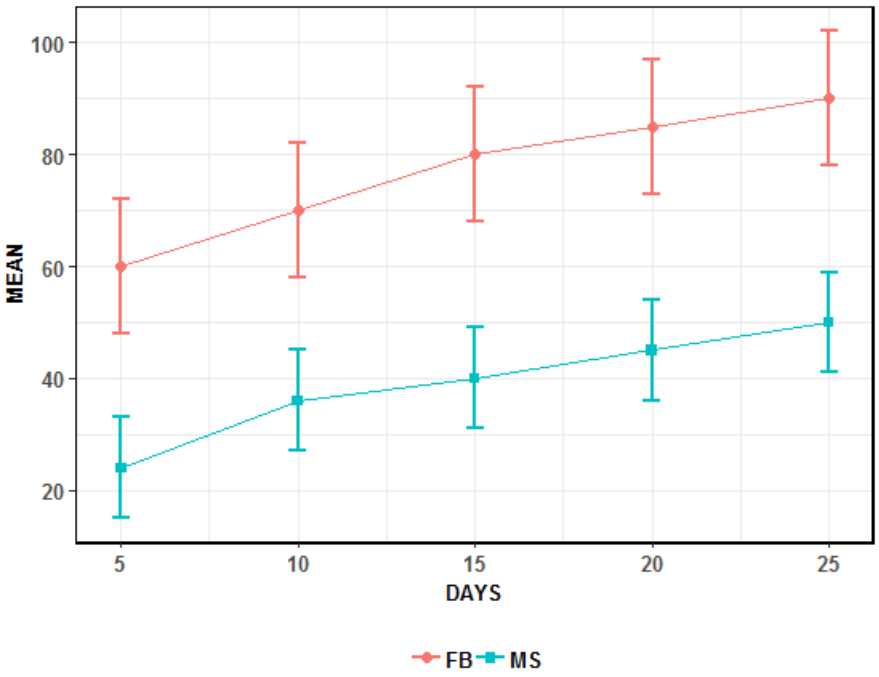

Fig. 3. Mean Plot

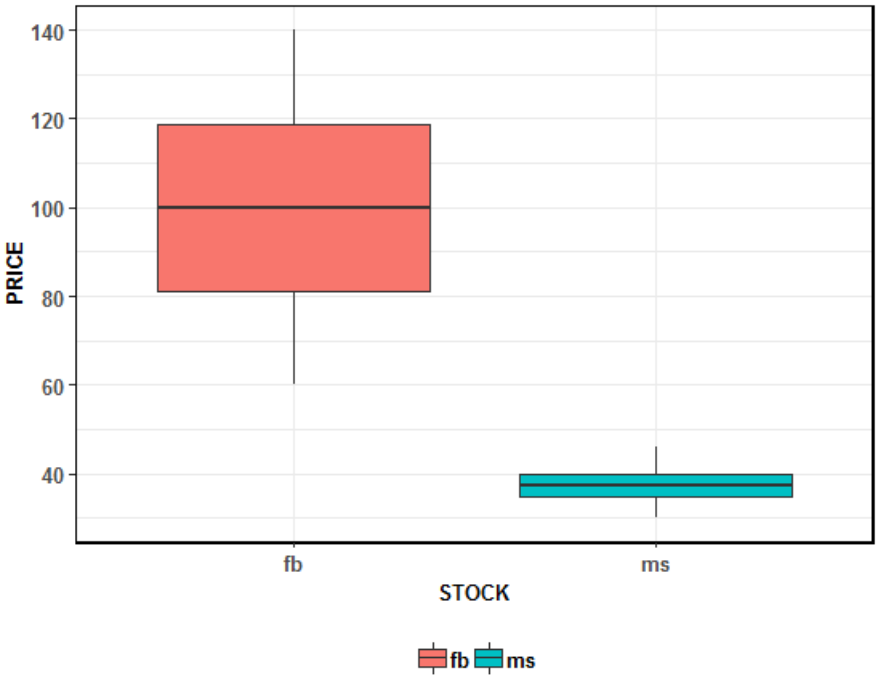

Fig. 4. Box plot

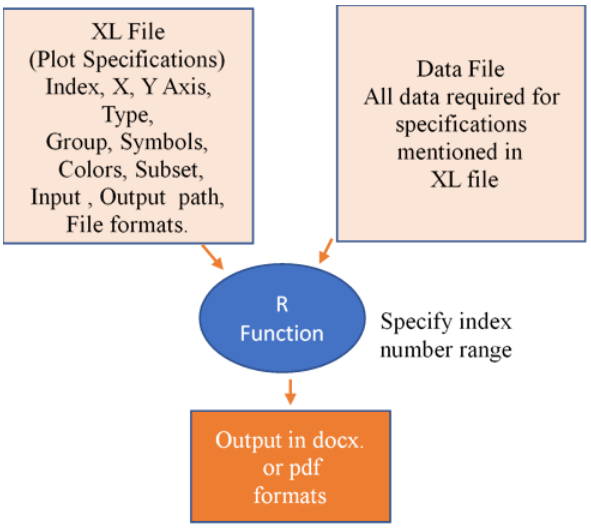

Fig.5. Flow Chart

V. CONCLUSION

The methodology described in this article will be useful in any sector of business which mostly depends on data for decision making process. Thanks to existing advanced cloud technology, every business is storing big data [11] in cloud and retrieving it anytime for deeper statistical analysis and visualization. The R function described here can be used to generate and visualize multiple plots in one click using specifications for the plot from XL file and data. So that user minimally modifies R-code, only index number of plots as whether it is individual or range of plots. Many outputs can thus be easily generated with minimal human input and faster manner which reduces human error and is time saving. Due to availability big data [11] and free data in other online resources for all business sectors, decision making can be easier using plotting system described in this article.

REFERENCES

1. Yi, J. S., Kang, Y. A., Stasko, J., & Jacko, J. A. (2007). Toward a Deeper Understanding of the Role of Interaction in Information Visualization. IEEE Transactions on Visualization and Computer Graphics, 13(6), 1224-1231.
2. Show me the Numbers: Designing Tables and Graphs to Enlighten Second Edition, Analytics press, 2012, Burlingame, CA-94011.
3. Introduction to Open Data Science, The Ocean Health Index Team 2017-11-27
4. The Visual Display of Quantitative Information, Edward R Tufte, Second Edition
5. Workplace Automation: Making Data Visualization Smarter by Arden Manning, Trends and Yseop Savvy ,16 August, 2016 |
6. The Importance of Data Visualization and Automation in the Realm of Big Data Published on August 24, 2015
7. Wickham H., 2009, ggplot2: elegant graphics for data analysis, New York, Springer.
8. A grammar of data manipulation: Dplyr, https://dplyr.tidyverse.org/
9. Almost everything you need to know about time series, https://towardsdatascience.com.
10. Officer-package-R , https://davidgohel.github.io/officer/.
11. Big data analytics, www.techopedia.com.

AUTHORS PROFILE

S.R.Sneha currently pursuing final year of her undergraduate in Computer Science and Engineering at St. Joseph's Institute of Technology, Chennai. She has done few on line courses related to data science. She has actively participated in many inter college technical events. She is a member of IEI (Institute of Engineers India). Her area of interest includes Artificial Intelligence, data science and statistical analysis. At present she is working on prediction algorithms and data visualization.

G. Rohini received the B.E. degree in Electronics and Communication Engineering from PSG College of Technology, Coimbatore in 1992. She received the M.E. degree in Applied Electronics from College of Engineering, Anna University, Chennai in 2002 and PhD in VLSI Design and Testing, from College of Engineering, Guindy, Anna University, Chennai. She has twenty-two years of teaching and four years of industrial experience. She is a corporate member of IETE, Member of IEEE, and Life Member of ISTE. Her research interests include VLSI Design and design for testability, High speed low power VLSI architectures, algorithms for VLSI signal processing and data science.

Kamal Srinivasan received his BSc degree (Biochemistry) from PSG Arts and Science Coimbatore, MSc (Biochemistry) from Bharathidasan University and PhD in Protein Engineering from Center for Biotechnology, Anna University, Chennai. He has also received three Post-Doctoral Fellowship from University of Kentucky, University of Tennessee Health Science Center and Vanderbilt University, USA respectively. He has twenty years of experience in teaching, research and industry. Currently he is working as a senior statistical programmer at Regeneron Pharmaceuticals, Tarrytown, New York, USA. His interests include data visualization, statistical analysis and machine learning.