Improving data literacy using liteRate: an R Shiny Apps for visualizing and exploring data published on BPS-Statistics Indonesia's Website

Erika Siregar[1]; Aris Prawisudatama[1]

[1] BPS-Statistics Indonesia {erika@bps.go.id; arisp@bps.go.id}

Abstract:

The Independent Review of UK Economic Statistics stated that "the longer a decision-maker has to wait for the statistics, the less useful they are likely to be". This statement is not only related to how fast the data is available, but also how fast the data can be understood. Official statistics data are generally considered of high quality. Related to this, BPS-Statistics Indonesia strives to satisfy the public's need for data through its official website (https://bps.go.id/). As the leading portal in presenting trusted data in Indonesia, the BPS website provides complete data that covers various areas, subjects, and domains. However, the visualization of these data seems to be lagging. These data are presented in the form of tables and static graphs that are monotonous and lack of interactivity which discourages users from exploring the data more.

To tackle this gap, we introduce liteRate: an interactive web-based visualization/exploration tool that is based on the Shiny R. LiteRate aims at improving user experience and strengthening public ability in understanding the data published on BPS' website. It utilizes web scraping and headless browsers to parse tables and convert them into ready-to-visualize dataframes. By using this tool, public users will be empowered to quickly gain insight, see patterns, correlation, outlier, and view statistics across a variety of topics and areas. The users can play around with tables using various available widgets (slider, dropdown, etc.) and directly produce the visualization. There are numerous visualization options available, from scatter plot to boxplot, and from heatmap to word cloud. The user can also do further exploration by changing colors, add/remove legend, faceting, zooming, panning, and annotating/highlighting important features. Moreover, users can perform simple descriptive computations and download the resulted output into various formats.

Our paper will explore the use of Shiny R and other javascript libraries to create an exploratory tool that offers two prominent features: interactivity and customization. Effective visualization techniques and interactive formats will help BPS to reach out to more users and encourage them to engage themselves in making sense of data. This will eventually increase statistical literacy in Indonesia as well as the whole world and realize the Statistics that left no one behind.

Keywords: visualization; interactivity; shiny R; user experience; insight mining

1.Introduction

Nowadays, statistical literacy has become a concern of every National Statistical Office (NSO) around the globe. The ability to understand statistics is a prerequisite for successful communication with public users. Especially amid the present-day hoax and fake news, the need to reinforce ability in understanding figures and evaluating information is deemed increasingly urgent.

Statistical literacy is generally defined as the ability of an individual or a group to understand and comprehend statistics (UNECE, 2012). It includes the ability to read and communicate the meaning of data. Lack of statistical understanding can lead to numerous misinterpretations of official data which can be observed in media reports, daily newspaper articles, and in direct contact with our users.

As the leading portal in presenting trusted data in Indonesia, the BPS-Statistics Indonesia (BPS) website provides complete data that covers various subjects and areas. However, these data are presented in the form of tables and static graphs that are monotonous and lack interactivity, which discourages users from exploring the data more. Simplification of tables and charts becomes necessary as the volume of quantitative information increases. The better people are informed about how to evaluate figures, and the reliability of official statistics, the easier they can assess the meaning and quality of the data. Hans Rosling's Gapminder, for example, has changed the way data can be presented.

There are already numerous products from different sources related to innovative tools for data communication. However, few are from official trusted sources. Questions emerge around how the official statistics community should best invest in this endeavour (Ferligoj, 2015). One way to achieve this is by creating a digital tool that can explain statistics in a clearer and easier way (Corselli-Nordblad & Gauckler, 2018). Organizations such as Eurostat (2020), United Nations Economic Commission for Europe (UNECE), and International Association for Statistical Education (IASE) have developed a range of innovations to promote statistical literacy such as the International Statistical Literacy Project (ISLP). Inspired by these breakthroughs, we introduce **liteRate**, an interactive web-based tool for rapid visualization of data published on the BPS website.

LiteRate is an excellent way of showcasing data, which unveils the beauty of data by converting uninteresting figures into attractive and interactive visualizations. LiteRate targets users of various knowledge levels including students, academics, journalists, and policymakers. The users can intuitively engage themselves in customizing the visualization using the provided advanced widgets, such as slider, dropdown, and checkbox. This combination gives the public users the ability to gain a simple overview of complicated subjects as well as spot quick facts and any interesting features in a graph. Hence, increasing data literacy.

2. Literature Review

2.1 R

R is a free and open-source language and environment that provides a wide variety of statistical and graphical techniques (The R Foundation, 2020). It compiles and runs on variety platforms (Windows, UNIX, and MacOS. We choose R in this project because it is a well-developed and effective programming language that provides services such as effective data handling, integrated collection of intermediate tools for data analysis, graphical facilities, and highly extensible (de Vries & Meys, 2015). Some of the R libraries used to build liteRate are ggplot (Wickham, H. et al., 2020) and rplotly, rpivotTable (Martoglio, E., 2018), and esquisse (Meyer, F., 2020).

2.2 Shiny

Shiny is an R package that allows the development of interactive web apps entirely in R, providing a powerful framework for disseminating official statistics (RStudio, 2020). Shiny is a combination of a dashboard, apps, and interactive document, which is commonly used for data exploratory and analysis. The main important features of Shiny are the user interactivity and widget control. Furthermore, a Shiny app can also be extended with CSS themes, htmlwidgets, shinyWidgets, and JavaScript actions. R Shiny to build a service that enables users to create on-the-fly data visualization which is presented in a dashboard format.

2.3 RSelenium, JQuery, and Docker

RSelenium (Kim, 2020) is an R wrapper for Selenium 2.0 Remote WebDriver. It provides a range of tools and libraries that enable and support the automation of web browsers locally or remotely. It allows developers to simulate common activities performed by end-users such as entering text into fields, selecting drop-down values, checking boxes, and clicking links in documents. It also provides many other controls such as mouse movement and arbitrary JavaScript (JS) execution (Selenium, 2020). JS execution is conducted inside the web browser. This step involves a series of additional processes including interaction with Document Object Model (DOM) (Robie, J. & Research, T., 2014) and making Hypertext Transfer Protocol (HTTP) request by utilizing JQuery. JQuery is a fast, small, and feature-rich JavaScript library that simplifies HTML document traversal, manipulation, and event handling (JQuery Foundation, 2020).

Crawling data with RSelenium requires the server to run execution in parallel. To handle this challenge, we utilize containers provided by Docker. Docker is a full development platform to build, run, and share containerized applications (Docker Inc., 2020). It provides a way to run applications securely isolated in a container, packaged with all its dependencies and libraries.

2.4 RSQLite

RSQLite is an R package that embeds the SQLite (SQLite Consortium, 2020) database engine in R (Müller, K., 2020). We use SQLite in this project because it is a simple, public-domain, single-user, embedded (serverless), and very light-weight database engine that implements table creation, updating, insertion, and selection operations, plus transaction management.

3. Methodology

3. 1 Data Crawling and Extract, Transform, Load (ETL)

LiteRate works by parsing the data available on BPS website and making them ready-to-use for further visualization and analysis. It takes the data published on BPS website as input by scanning all pages containing data and scraping (extracting) the data (the figures including the metadata) from the webpage. Pages containing data can be accessed from the menu located on the page's left-hand sidebar (Figure 1). This data is classified into three groups, which are "*Social and Population"*, *"Economy and Trade"*, and *"Agriculture and Mining"*. Each group has several related subjects that are presented on different pages and tables.

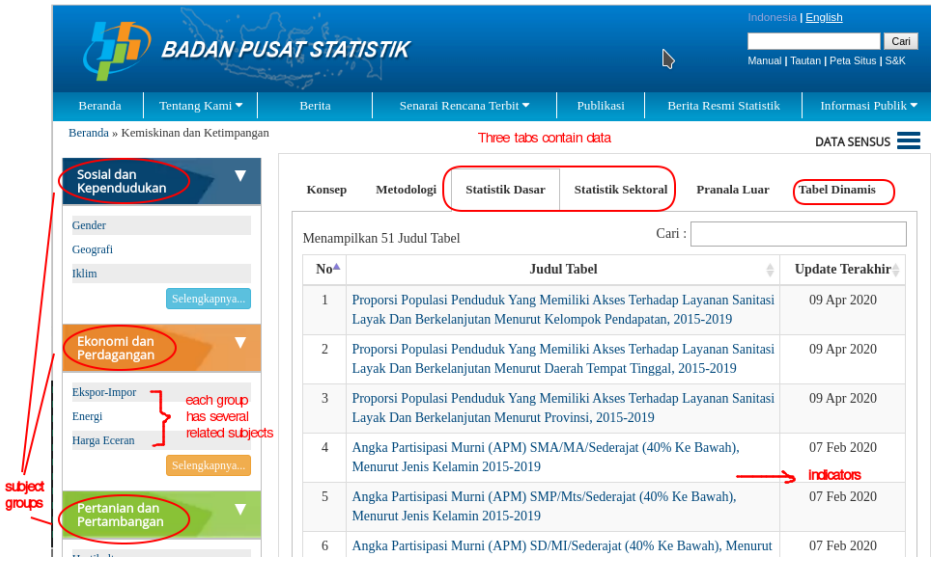

Figure 1. *The Data Section on BPS Website*

Each subject page has six tabs, which are *concept*, *methodology*, *basic statistics*, *sectoral statistics*, *external link*, and *dynamic table*. We can ignore the tabs *concept*, *methodology*, and *external link* since they do not contain data. For the tabs basic statistics and sectoral statistics, there are two types of data format available: static table and dynamic table. The static table has irregular format with non-standardized row and column names, hence we exclude it from this research. On the other hand, the dynamic table has consistent structures (rows and columns) and standardized metadata that enables us to automate content and variable extraction. Furthermore, we determine that tables presented in the tabs basic statistics and sectoral statistics are just subsets of tab dynamic table. Therefore, we will crawl data from the tab *dynamic table* only.

The tab *dynamic table* comprises several sections that are subject, indicator, characteristic, time, and area (Figure 2). Each section is dynamically loaded using AJAX request. However, BPS website is protected from Cross-Site Request Forgery (CSRF) (Wasson, M., 2012) that prevents outside access via Application Programming Interface (API). We solve this challenge using Selenium (Figure 3). Selenium opens the dynamic table page, gets the CSRF code, and makes an AJAX request by including the CSRF code in the payload request. Since each subject in dynamic tables has a different collection of variables (indicators, characteristics, time, and area), it is necessary to create a request for each variable combination using the AJAX request. The final result is stored to the RSQLite database. Figure 4 and 5 show the data crawling flow chart and a result example, respectively.

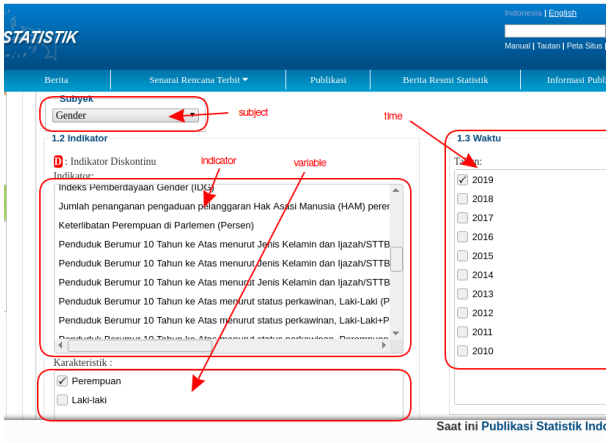

Figure 2. *The dynamic table page*

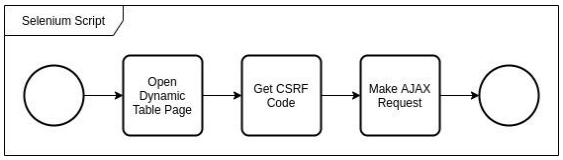

Figure 3. *AJAX Request Dengan CSRF*

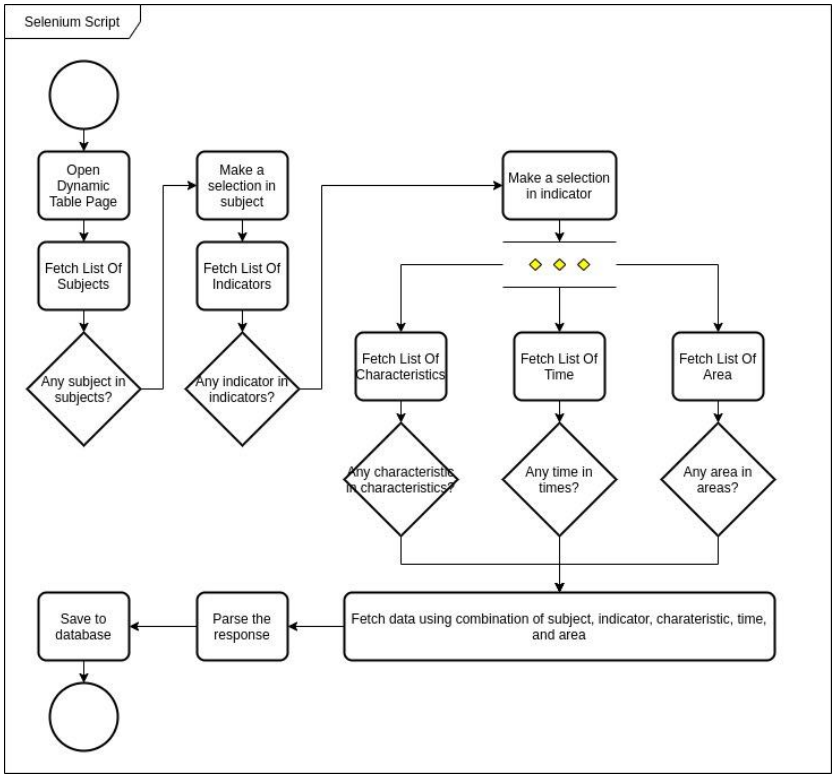

Figure 4. *Flowchart Data Crawling*

3.2 Building The Literate Shiny App

The data resulted from the ETL process is visualized in a dashboard format that is built by utilizing the libraries Shiny, shinyWidgets, rpivotTable, and rplotly. The widgets pivot tables and plotters enable users to easily explore data, selecting variables, and automatically detect and give a suggestion about the right chart type to use according to the data type and axis chosen. Users can also change different types of plot charts according to their needs.

Data obtained through ETL is still in unpivoted format (long table). Thus, we have to create a service in which users can conduct an on-the-fly pivoting process using the tidyr library. A selector is provided on the left-hand sidebar as well as the *drag and drop* features for variable selection and visualization components (title, legend, etc.). This feature is created using the esquisse and pivottables.js libraries. The codes used in building liteRate are available on GitHub https://github.com/erikaris/liteRate.

Provinsi/Kabupaten/Kota	Angka Harapan Hidup (AHH) Menurut Kabupaten/Kota dan Jenis Kelamin (Tahun)
	Perempuan
	2019
ACEH	71.85

Figure 5. *Result Single Combination of Data Crawling*

4. Result

The result of this research is a comprehensive Shiny app named liteRate. In general, there are three major features offered by liteRate:

1. Fancy and complicated widgets.
   Utilize shinywidgets to implement tabs, dropdown, slider, textbox, and radio button.
2. Interactivity, customization, and aesthetic aspect.
   LiteRate enables users to follow their intuition in exploring the data, which will trigger the improvement of statistical literacy. Available interaction and customization includes defining plot axis, modifying table (sort, filter), changing color, and customizing graph.
3. Various visualization options.
   LiteRate provides pie chart, bar graph, boxplot, scatter plot, line graph, histogram, area chart.

Figure 6 and 7 provide examples of pivot table and visualization that can be created using liteRate.

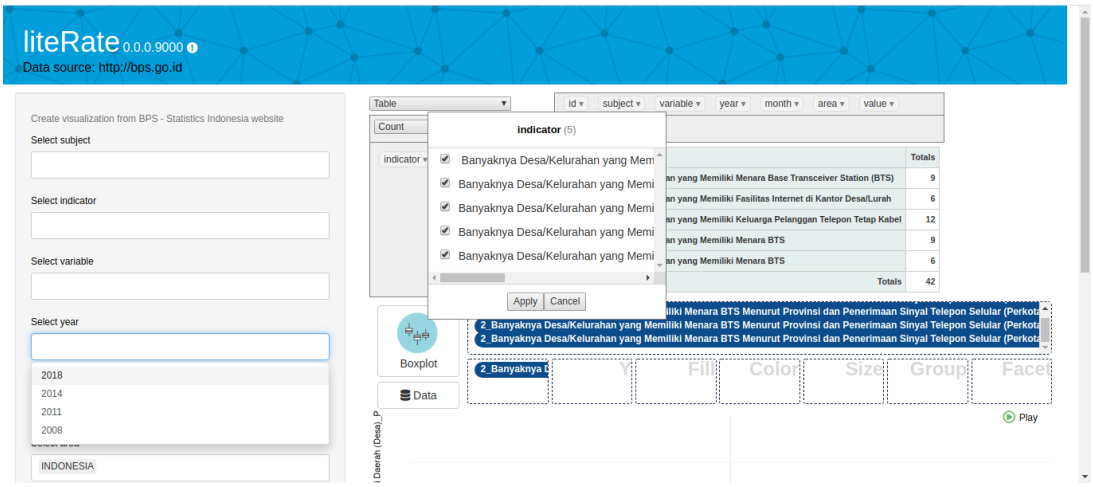

Figure 6. *Pivot table in liteRate*

5. Discussion, Conclusion, and Recommendation

We provide a new way to investigate data by providing a non-conventional tool to visualize BPS-Statistics Indonesia data that enables users to understand the information clearly and directly through some underexplored point of views. Although currently still in the prototype phase, we believe that liteRate could be a breakthrough for BPS in creating an appetite for their data and helping to develop citizens' understanding on how to use data.

This research requires further development to provide better contributions and more benefits. To keep improving liteRate, we encourage users to provide feedback to us. Future work following this study would focus on (1) handling static-format tables, (2) providing a feature for conducting simple analysis such as descriptive and regression, (3) collecting logs to analyze user behavior, and (4) publishing liteRate on The Comprehensive R Archive Network (CRAN) to reach a broader user.

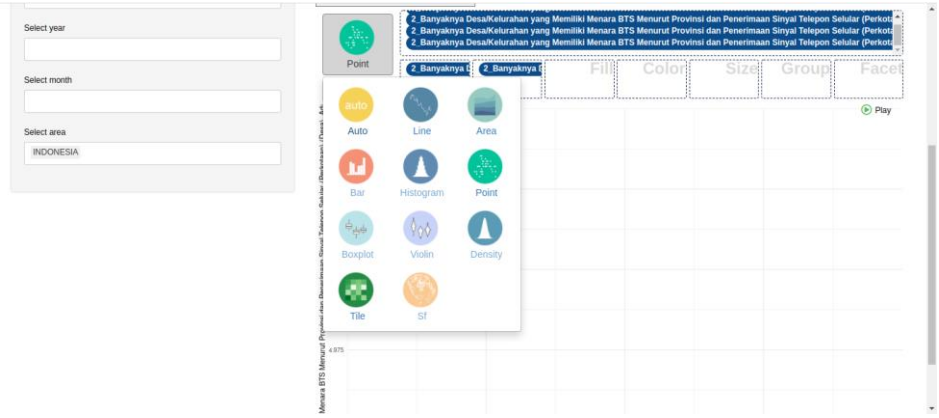

Figure 7. An example of visualization in liteRate

References

Corselli-Nordblad, L., & Gauckler, B. (2018). New tools to improve statistical literacy – developments and projects—An ESS priority and reality. *16th Conference of IAOS*, 8.

de Vries, A., & Meys, J. (2015). *R For Dummies* (2nd ed.). John Wiley & Sons, Inc.

Docker Inc. (2020, April 28). *Get started with Docker for Windows*. Docker Documentation. https://docs.docker.com/docker-for-windows/

Eurostat. (2020). *Visualisations, Mobile Apps & Extraction Tools*. https://ec.europa.eu/eurostat/help/first-visit/tools

Ferligoj, A. (2015). *How to improve statistical literacy? 12*, 1–10.

JQuery Foundation. (2020). *JQuery*. https://jquery.com/

Kim, J. Y. (2020). *RSelenium package | R Documentation*. https://www.rdocumentation.org/packages/RSelenium/versions/1.7.7

Martoglio, E. (2018, January 30). *RpivotTable.* https://cran.r-project.org/web/packages/rpivotTable/vignettes/rpivotTableIntroduction.html

Meyer, F. (2020). *Explore and Visualize Your Data Interactively • esquisse*. https://dreamrs.github.io/esquisse/index.html

Müller, K. (2020). *RSQLite package | R Documentation*. https://www.rdocumentation.org/packages/RSQLite/versions/2.2.0

Robie, J., & Research, T. (2014). *What is the Document Object Model?* https://www.w3.org/TR/WD-DOM/introduction.html

RStudio. (2020). *Shiny*. https://shiny.rstudio.com/

Selenium. (2020, April 18). *The Selenium project and tools: Documentation for Selenium.* https://www.selenium.dev/documentation/en/introduction/the_selenium_project_and_tools/

SQLite Consortium. (2020). *SQLite Home Page*. https://sqlite.org/index.html

The R Foundation. (2020). *R: What is R?* https://www.r-project.org/about.html

UNECE. (2012). *Making Data Meaningful Part 4- A Guide to Improving Statistical Literacy*. https://www.unece.org/fileadmin/DAM/stats/documents/writing/Making_Data_Meaningful_Part_4_for_Web.pdf

Wasson, M. (2012, December 12). *Preventing Cross-Site Request Forgery (CSRF) Attacks in ASP.NET MVC*. https://docs.microsoft.com/en-us/aspnet/web-api/overview/security/preventing-cross-site-request-forgery-csrf-attacks

Wickham, H., Chang, W., & Henry, L. (2020). *Create Elegant Data Visualisations Using the Grammar of Graphics*. https://ggplot2.tidyverse.org/