Dashboard Pre-Processing Data (DPD) as Data Analysis System with Technological Innovation to Perform Pre-Processing Quantitative Data

Mashuri Mashuri[1*], Albertus Eka Putra Haryanto[1], Yola Argatha Manik[1], Dinar Sukma Dewi[1], Tegar Primadana Putra[1]

***Abstract*— In essence, data in real-life always needs to be pre-processed or better known as pre-processing data. Pre-processing data is one of the early techniques for converting raw data from various sources into cleaner information that can be used for further analysis. There are three types of pre-processing data, missing values, checking outlier data, and identifying the types of distribution in the data. Currently, statistical software that offers to be used in pre-processing data analysis has been widely and is quite familiar. However, the user is often can not run the analysis quickly. Therefore, the idea is to create and develop an application or dashboard that can be used to solve these problems. The application are designed using the updated technology, where users can use menus that are designed practically in order to pre-processing data. The application this at is offered and trying to be developed is called "DPD (Dashboard Pre-Processing Data)". This application serves as a tool to pre-process data quickly and efficiently. In addition, with this application, it's expected that users can identify missing values, data outliers, and some types of data distribution, so users can determine the analysis method that will be used on the research data they have.**

***Keywords*— Outlier data, Distribution of the data, Pre-processing data, Missing value**

I. INTRODUCTION

There is a fact, data in real life always needs to be pre-processed or better known as pre-processing data that used before analyzing with a certain method. Pre-processing data is an initial technique for converting raw data from various sources into cleaner information that can be used for further analysis. Pre-processing helps improve the data's precision and performance and prevents errors in data analysis. Three general problems can be solved in data pre-processing activities, there are missing values, data noise, and data distribution. Missing value is inaccurate data due to missing information which causes the information contained in it to be irrelevant. Missing value often occurs when there are problems in the collection process, such as errors in data entry or problems in using biometrics [1]. In addition, the occurrence of missing values is because of several things, there is failure to use equipment, inconsistency with other recorded data, data not being entered due to misunderstandings in the data collection process, besides certain data are not considered important at the time of data entry [2]. Checking the noise of data is also considered important in pre-processing data. Noise data or also known as outlier data is data that is outside of the average limit. The occurrence of data outliers due to human error (data input error) or other errors in data collection. In addition to missing values and outlier data, data pre-processing activities can also be identified regarding the type of distribution or distribution of data. Identification of the type of distribution or distribution is important to determine the analysis method by describing the data held. Data analysis is generally divided into linear data analysis (assuming normal distribution) and non-linear data analysis (data tends not to follow a normal distribution).

Much statistical software is familiar and can be used in pre-processing data, such as Minitab, SPSS, WEKA, and other software. However, users are often unable to run in an efficient time pre-processing data. This is impractical for users whose decision-making takes a fast time in the data analysis process. In its development, R software developed an R-Studio application which is an Integrated Development application system for open source and free. R-Studio software in its development consists of two types: R-Studio desktop and R-Studio Browser. On the other hand, R-Studio is also developing an interactive website application or dashboard with packages from Shiny. Shiny is a work tool that can be used to create a website-based application. Shiny is a package from the Website Based Framework that is used in the R programming language. Shiny is often used because it is simple and easy to use and developed by making basic website applications easy and attractive. Shiny itself has components that are divided into server parts, where the server itself is a user who can operate several programs with a function as a simulator, various kinds of appropriate data analysis according to the user's choice, where the results of the analysis will be sent to the output section. The advantage of using Shiny is also the use of a user interface with a panel to set input in the form of data, variables, and models.

From the variety of those problems above, an idea emerged to make a DPD system design. DPD stands for Dashboard Pre-Processing Data. This dashboard serves as a tool for pre-processing the data quickly and efficiently. Researchers first designed this dashboard and offers

[1] Department of Industrial Mechanical Engineering, Institut Teknologi Sepuluh Nopember, Kampus ITS Sukolilo, Surabaya, 60111, Indonesia. E-mail: mashuri@its.ac.id

several features and menus expected to assist users in pre-processing data. With this application, it is expected that users can identify missing values, data outliers, and types of data distribution so that users can determine the analytical method to be used on the research data they have.

II. LITERATURE REVIEW

A. Missing Value

The missing value is an event that the information that is not available for a subject (case). Missing data occur for several reasons, including information about an object that is not provided. It is difficult to find, even if the information does not exist. Missing value causes empty cells in one or more variables [4].

B. Data Outliers

Several things cause outliers, including errors in data collection, errors in entering data, or data outliers. New data can be said to be an outlier if the data is outside the range of the lower and upper limits of an interquartile in a data structure. The steps to detect outlier data are as follows [4].

1. Standardize the data.
2. If the sample size is 80, then the data is said to be an outlier if the standard score is outside the limit of ± 2.5.
3. If the sample size is> 80, then the data is said to be an outlier if the standard score is outside the limit of ± 3.

C. Opportunity Continuous Distribution

A continuous probability distribution is a sample space containing infinite sample points equal to the number of points on a line [5].

The condition for the continuous probability distribution is that if the function f(x) is a solid probability function of the continuous random variable Xt defined over the set of all real numbers R_t if:

1. $f(x) \geq 0$ for all $x \in R$
2. $\int f(x)dx = 1$
3. $P(a < x < b) = \int f(x)dx$

One type of continuous probability distribution is the normal distribution. It is said that the data is normally distributed if the experiment whose random variable x is determined by the parameter μ and σ^2. If x is a normal random variable with a mean μ and σ^2 [6].

$$N(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \text{ for } -\infty < x < \infty \quad (1)$$

Description:

x = Normal continuous random variable
μ = Mean
σ = Standard deviation
π = 3.14159
e = 2.71828

The normality test of the data can be seen from the D value obtained from the results of the Kolmogorov Smirnov - Liliefors test, namely by comparing the D value with the D value The hypothesis of the normality test is [6].

H_0 : Data is normally distributed.
H_1 : Data is not normally distributed.

$$D = \sup_x Fn(x) - Fo(x) \quad (2)$$

Critical area: Reject H_0 if the pvalue of D < Dα
Description:
Fo(x) : Theoretical cumulative frequency distribution
Fn(x) : Sample cumulative frequency distribution

D. Mean

Mean is sum of values in a data divided by the number of data. The mean is obtained by adding up all the values and dividing by the number of individuals. In statistics, it is often called the Arithmetic Mean. This measure is easy to calculate by utilizing all the available data. If there is a group of data, then to call the numerical measure as a representative of the data, the arithmetic average is often used [6].

The average formula is

$$\overline{X} = \frac{\sum_{i=1}^{n} Xi}{n} = \frac{X_1+X_2+....+X_n}{n} \quad (3)$$

Description:

$\overline{X}$ = Mean or average
n = Amount of data to be processed
$\sum_{i=1}^{n} Xi$ = Amount of data obtained
X_n = Data of the n[th] term.

E. Median

The median is a set of data sorted from smallest to largest, and vice versa, from the largest to the smallest data. This means the datum that divides the data sorted into two equal parts [6]. The following is the formula to find the median of single data with an odd number of data, Here is the formula for finding single data with an even number of data.

$$Med = X_{\left(\frac{n}{2}\right)} \quad (4)$$

Description:

Med = Median
$X_{\left(\frac{n}{2}\right)}$ = Term data of half the sum of n
$X_{\left(\frac{n}{2}+1\right)}$ = Term data of half the sum of n+1

F. Maximum and Minimum Value

The maximum value is the largest in a data set or the value that has the highest and largest level of other values. While the minimum value is the value that has the smallest or lowest level of other values [5].

G. Box-plot

Box-plot (also called a box-and-whisker diagram) is a box (a square-shaped box). Box plots are one way of describing descriptively in the form of graphs of numerical data that includes the following five measures:

- The smallest observation value,
- The lowest quartile or first quartile (Q1), which truncates 25% of the lowest data
- The median (Q2) or the middle value,
- The highest quartile or third quartile (Q3), which truncates 25% of the largest data
- The largest observation value.

The box plot also shows, if there is an outlier value from the observation, then the box plot can be used to show differences between populations without using the assumption of the underlying statistical distribution. Therefore, box plots are classified as non-parametric statistics. The distance between parts of the box indicates the degree of dispersion and skewness in the data. In its depiction, box plots can be described horizontally or vertically [5].

H. QQ Plot

In addition to histograms, a tool that be used to plot of the data distribution is can called a Quantile-Quantile Plot (QQ-Plot) can be used to plot variables more accurately based on the value of quantile data. QQ Plot is a scatter plot that compares the empirical distribution with the fitted distribution about the dimension value of a variable (eg empirical quantile value). QQ plots can also plot well if the dataset is obtained from a known population [7].

I. Histogram

A histogram is a one of chart that represents about a tabulation of data organized by size. Generally, this tabulation of data is called the frequency distribution. The histogram represents the characteristics of the data divided into category. In the histogram, the frequency shows the observed value of each category. If the data follows a normal distribution, then the histogram shape will resemble a "bell". This shows that a lot of data are in the average value. The shape of the histogram that is skewed or not symmetrical shows the amount of data that is not within the average value, but its value is within the upper or lower limit [6].

J. R-Shiny

R-Shiny is a Web-based R application that has a framework for creating applications on websites using R code designed for data science without knowledge of HTML, CSS, or JavaScript. On the other hand, Shiny is not limited to making simple applications, Shiny has been equipped with user interface components that can be easily customized or extended, where the server uses reactive programming to enable it to create any back-end logic you want. Shiny is an R package allowing you to create rich and interactive web applications easily. Shiny makes it easy to display and publish via the website so that anyone can use it. Provides a set of meticulously accurate user interface functions that generate the HTML, CSS, and JavaScript needed for common tasks. This shows that the user does not need to know the HTML/CSS/JS details. Shiny is often used for dashboards that track important high-level performance indicators, while facilitating tracing to metrics that require further investigation, communicating complex models to non-technical audiences with informative visualizations and interactive sensitivity analysis, providing data analysis for common workflows, replacing email requests with the Shiny app that allows people to upload data and perform simple statistical analyses. R Shiny can host standalone applications on web pages or embed them in R Markdown documents or build dashboards. When using Shiny apps, there is no need to have a server or know how to configure a firewall to deploy and manage cloud applications. Plus, no hardware, installation, or annual purchase contracts are required. Shiny apps run in a protected environment and are always SSL encrypted. The Standard and Professional packages offer user authentication, preventing anonymous visitors from being able to access the application [8]. Shiny apps run in a protected environment and are always SSL encrypted. The Standard and Professional packages offer user authentication, preventing anonymous visitors from being able to access the application [8]. Shiny apps run in a protected environment and are always SSL encrypted. The Standard and Professional packages offer user authentication, preventing anonymous visitors from being able to access the application [8].

III. Implementation Method

The application developed is "DPD (Dashboard Pre-Processing Data)". This application is in the form of a dashboard that can be accessed via the link provided, which is made using the R-Shiny platform. This dashboard serves as a tool to pre-process data quickly and efficiently. With this application, it is expected that users can identify and identify missing values, data outliers, and types of data distribution, so that users can determine the analytical method to be used on the research data they have. In its use, the dashboard can be accessed via the following link https://intip.in/DPD2021ATCKel1.

Figure 1 shows the initial view on the dashboard.

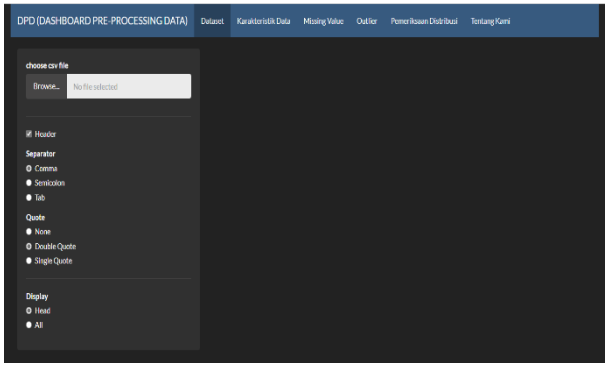

Figure 1. Initial View

Figure 1 is the initial view of the “DPD (Dashboard Pre-Processing Data)” application. There are 6 menu’s that users can access: datasets, data characteristics, missing values, outliers, distribution checks, and about us.

The dataset menu is the initial menu for uploading data for pre-processing. The data used is data in CSV (Comma Separated Values) format. The separator function is used to separate data from one another so that the data is neat, the separator is adjusted to the data that has been uploaded, the user can change the separator to a comma, semicolon, or tab, adjust the data so that it is displayed properly by the dashboard. The display function is to display the uploaded data. Users can change it with the first two options header. Headers are used to display some of the top data. The second is all, used to display all available data. As an illustration, used data on the percentage of households that have perennials by province and region in 2013, 2014 and 2017 which were obtained from BPS. The research variables are defined in Table 1 as follows.

Table 1. Research Variables

Variable	Description
Province	34 Provinces in Indonesia
X_1	Percentage of Households Owning Perennials in Urban Areas in 2013
X_2	Percentage of Households Owning Perennials in Rural Areas in 2013
X_3	Percentage of Households Owning Perennials in Urban Areas in 2014
X_4	Percentage of Households Owning Perennials in Rural Areas in 2014
X_5	Percentage of Households Owning Perennials in Urban Areas in 2017
X_6	Percentage of Households Owning Perennials in Rural Areas in 2017

IV. Results and Discussion

Figure 2 is a menu display on the dataset that can be seen on the dashboard after uploading data.

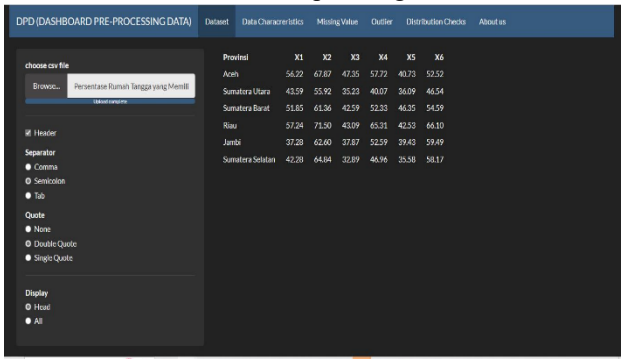

Figure 2. Display After Uploading Data

User can see the characteristics of the data first, which contains information related to Minimum, Q_1, Median, Mean, Q_3, Max, and NA's. The NA's value displayed on the data characteristics menu shows the number of each research variable related to the missing value. Figure 3 is a display of the data characteristics of the percentage of households that have perennials.

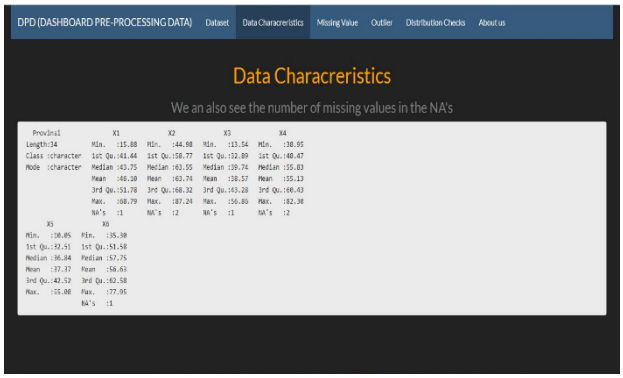

Figure 3. Data Characteristics

Figure 3 shows there are 7 research variables, where 1 variable is a categorical variable (character), and 6 are quantitative variables. If viewed from the missing value, it can be seen that the X_1 variable has 1 missing value, X_2 has 2 missing values, X_3 variable has 1 missing value, X_4 variable has 2 missing values, and X_6 variable has 1 missing value. variable X_5 has no missing value. On the Dashboard, the user can observe and see which variables and rows there are missing values. Figure 4 is a display of raw data containing missing values. In this case, the observations written with the word true are observations with missing values.

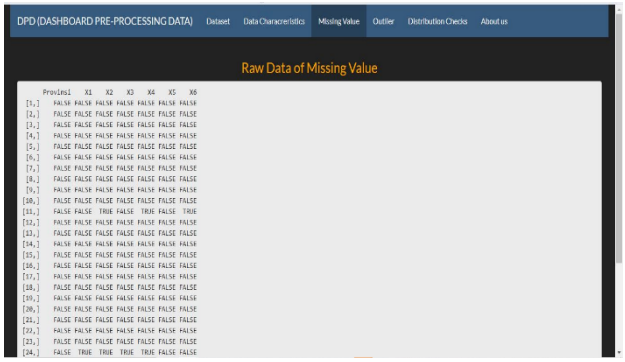

Figure 4. Raw Data Missing Value Tampilan Display

Figure 4 shows that the missing value in the data on the percentage of households that have missing is found in the 11th observation of the variables X_2, X_4, and X_6; observations in the 24th at the variables X_1, X_2, X_3, and X_4. The cause of missing values in the 11th observation of variables X_2, X_4, and X_6 is that DKI Jakarta Province does not have rural areas, because DKI Jakarta Province is a province that only has urban areas. Meanwhile, the cause of the missing value from the observations of the 24 variables X_1, X_2, X_3, and X_4 is in these observations, which are observations for the Province of North Kalimantan, where in 2013 and 2014, the Province of North Kalimantan has not yet been divided and is still a province. with East Kalimantan Province.

In outlier detection, the user can see the outlier data on the variables entered by the user himself. The results of the outlier data can be seen from the visualization of the box plot and the results of outlier detection by testing the Z value. Figure 5 and Figure 6 are the display of the box plot and the results of outlier testing on the X_1 variable.

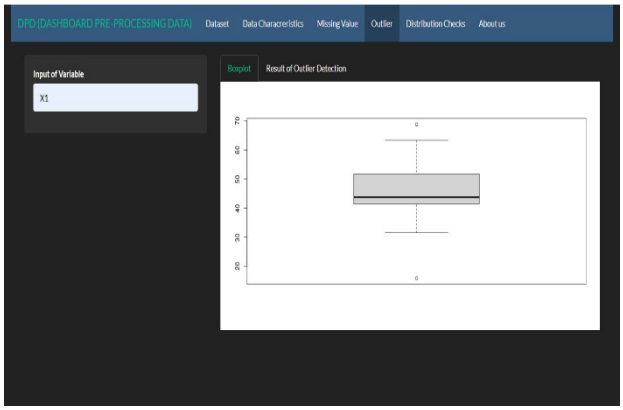

Figure 5. Box Plot on Variable X_1

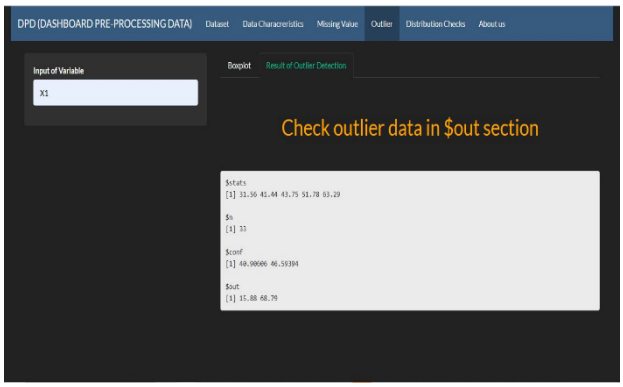

Figure 6. Outlier Detection Results on Variable X_1

Figures 5 and 6 are display of the outlier menu. On the outlier menu, there are two sub menu's, visualization of box plot and outlier detection results. The box plot sub-menu displays the uploaded data box plot. Users can enter which variables will be visualized and tested in the "Enter Variable" input box. The results obtained based on the box plot visualization that there are two observations on the X_1 variable which are included in the outlier group. This can be seen from two plots, each below or above the Interquartile Range. In the dashboard, in addition to visualization, the user can see the outlier value in the part of menu the outlier detection results in the outlier detection sub-menu. The part of menu the outlier detection results represents about the results of outlier detection which can be seen in the $out output. It is said that there is no outlier value if the result is 0.

In addition to detecting missing values and outliers, the user can check the type of distribution and test the data whether it is normally distributed (linear) or not normally distributed (non-linear). In the distribution check menu, there are 3 sub menus, namely histogram to display the visualization of data distribution, QQ Plot to display data distribution plots based on quantile values associated with the normality line, and test results to test whether the data is normally distributed using the Kolmogorov Smirnov test. Figures 7, 8, and 9 are the display of the distribution check on the X_1 variable that has been inputted by the user in the previous menu (Figure 5).

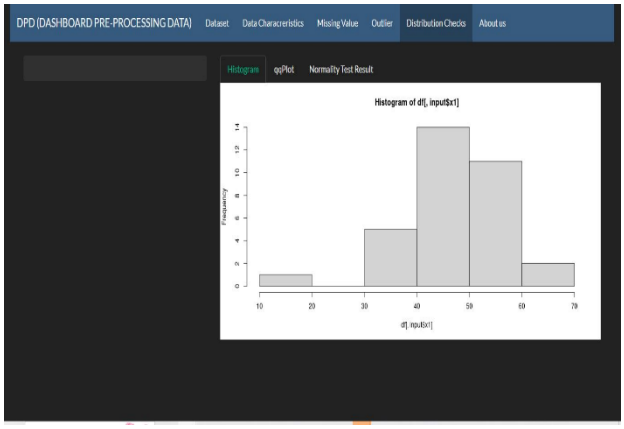

Figure 7. Visualization of Histogram Results on Variable X_1

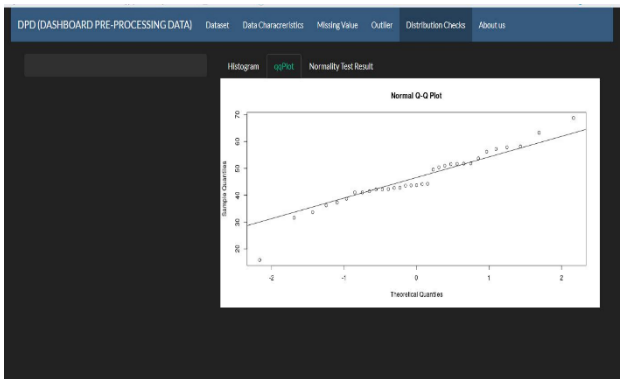

Figure 8. Visualization of QQ Plot Results on Variable X_1

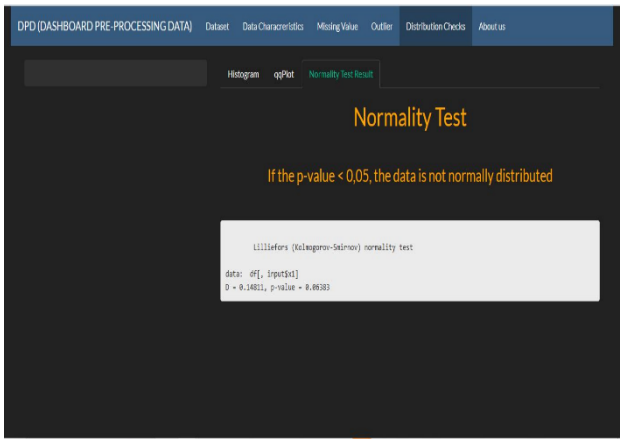

Figure 9. Normality Test Results on Variable X_1

Figure 7 is a display of the distribution check menu. the histogram of the uploaded data is based on the variables previously entered in the variable input box available on the outlier menu (Figure 5). The histogram represents the characteristics of the data divided by category. The frequency of the histogram represents about the observed value of each category. If the data follows a normal distribution, then the histogram shape will resemble a "bell". This shows that a lot of data are in the average value. The histogram's skewed or not symmetrical shape shows the amount of data that is not within the average value. Based on the results of histogram visualization on the X_1 variable, it was found that the tendency of the data to follow the normal distribution pattern.

Figure 8 is a represent about QQ Plot of the data. The sub-menu displays of the QQ Plot of the variables that have been inputted in the variable input box available on the outlier menu (Figure 5). QQ Plot represents that the distribution or distribution of the data, if the data distribution follows a normal distribution, then the data distribution will follow a linear reference line. However, if the data distribution does not follow a linear reference line, then the data tends not to be normally distributed. Based on the results the visualization of QQ Plot on the X_1 variable, it was found that the tendency of the data to follow a normal distribution. This is because the tendency of the formed QQ plot tends to follow the linear reference line.

Figure 9 displays the test results sub-menu, where the sub-menu displays the results of the normal distribution of uploaded data for the variables that have been inputted in the variable input box available on the outlier menu (Figure 5). If the data follows a normal distribution, the p-value > 0.05; if the p-value < 0.05, the data is not normally distributed. Based on results of the normal distribution test on a X_1 variable, it was found that the P-value was 0.0683 which was more than 0.05. The conclusion obtained is that the X_1 variable is a normally distributed.

V. Conclusion and Suggestions

DPD (Dashboard Pre-Processing Data) is a dashboard that used to pre-process data efficiently and accurately, where users can identify and identify missing values, data outliers, and types of data distribution before conducting further analysis. Where, this application offers users the ability to upload and input data for pre-processing.

The advice given for developing this application is the need to develop the visualization of this application to make it more attractive. Application development is required, so the output can be downloaded by the user in a more effective and efficient file form. In addition, application development is needed so that the application can be used to see the trend of data distribution, in addition to the normal distribution test.

References

[1] N. Galuh Importance of Pre-Processing in Statistical Processing. https://www.dqlab.id/importance-preprocessing-dalam-pengolahan-data-statistik. [On line]. Available: dqlab, www.dqlab.id [Accessed on October 19, 2021].

[2] Rosyid. Data Preprocessing Data Mining: Concepts and Techniques. [On line]. Available: http://rosyid.lecturer.pens.ac.id/dataMining/Data%20Preprocessing.pdf [Accessed on October 19, 2021].

[3] T. I Made. Presentation and Data Processing using the R. Application, 2014. [E-book] Available: https://repository.unej.ac.id/bitstream/handle/123456789/59392/PromoBuku2014.pdf?sequence=1&isAllowed=y

[4] J. Richard, and DW Wichern. Applied Multivariate Statistical Analysis, 5th edition. New Jersey: Prentice Hall Inc., 2002.

[5] RE Walpole (In), & IP Sidhi (Ed.), Introduction to Statistics. Jakarta: PT Gramedia Pustaka Utama, 2012.

[6] RE Walpole (In), & IP Sidhi (Ed.), Introduction to Statistics. Jakarta: PT Gramedia Pustaka Utama, 2007.

[7] Didi. Chapter V : R for Statistical Processing & Analysis., 2021. [Online]. Available: http://didi.staff.gunadarma.ac.id/Downloads/files/13709/BabV.pdf [Accessed on October 19, 2021].

[8] U. Abdullah. Getting to Know R Shiny Closer, 2021 [Online]. Available: https://www.abdumar.com/2021/03/mengenal-r-shiny-more-close.html [Accessed on October 19, 2022