Data Science With Excel

1stSigit Setiyanto, 2nd Ismail Setiawan
[1,2]Information Systems and Technology, Universitas 'Aisyiyah Surakarta
[1,2]Jl. Ki Hajar Dewantara No.10, Jawa, Kec. Jebres, Kota Surakarta, Jawa Tengah 57146
[1]sigitsiak@gmail.com, [2]ismail@aiska-university.ac.id

Abstract*—The stages in data science consist of several stages, one of which is data preparation. At this stage, many things are done so that the dirty data becomes clean data that is ready for modeling. Many applications offer data science convenience in terms of processing data. One of them is excel, this application from Microsoft can perform data processing so that the data is ready for modeling. However, there are limitations in using excel. The maximum number of rows that excel has is only 1,048,576 and the number of columns is 16,384. However, if you process data of no more than 1 million rows, excel can still handle it by using features such as error detection, removing duplicate data, correcting error values, detecting outlier values, handling missing data and validating data. This study shows some of these features along with examples of their use.*

Keywords : *Data Science; Excel; Data Preparation.*

I. INTRODUCTION

In the world of data science, a term called data cleaning is known. The definition of data cleaning itself is related to data quality. As already knows, the quality of the data greatly affects the results of the analysis. No matter how good and up-to-date the analysis is, if the quality is poor, even the results will not be satisfactory. Data quality can be ensured through a procedure called data cleaning. Data cleansing is the process of ensuring the accuracy, consistency, and usability of data in a data set [1][2]. The secret is to detect data errors or corruption and correct or delete the data as needed. Combining multiple data sources at the same time can result in duplicate or mislabeled data. In this situations, data sanitization is also necessary to avoid more complex problems. Here are some reasons why data cleaning is mandatory: Eliminate errors and inconsistencies that arise when multiple data sources are collected on a single dataset. Improving work efficiency because this process will make it easier for developers and data processing teams to find what to expect based on the data. A lower error rate will also bring customer satisfaction and reduce the team's workload. Helps developers map several different data functions. This process will also make developers more familiar with the usefulness of the data and get to know where the data comes from.

II. RESEARCH METHODS

Data cleaning methods carried out with excel on this activity includes [3] :

Detecting Errors; The first step that must be done is to monitor error or corrupt notifications [4]. Errors are sometimes not visible when editing a document. The error is only seen and feels annoying when printing the document because it turns out that there is an error output that appears. Actually, to find the error cell, you don't need to look at the contents of every cell contained in workshet. Errors can be found using the Go To facility found in Microsoft Excel.

Here's how to find cells that are error or still have errors: Press the key F5 or press the CTRL+G.

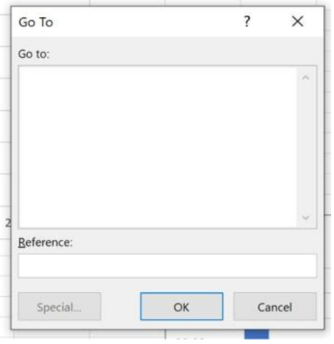
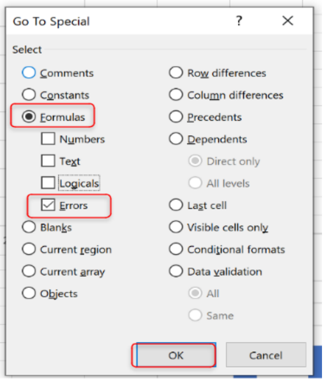


Figure 1. Go to feature visualization
In the Go To dialog

Click the Special button then in the Go To Special dialog select the Formulas option then provide a check syndication only in the Errors option.



Select
○ Comments
○ Constants
● Formulas
□ Numbers
□ Text
□ Logicals
☑ Errors
○ Blanks
○ Current region
○ Current array
○ Objects
○ Row differences
○ Column differences
○ Precedents
○ Dependents
● Direct only
○ All levels
○ Last cell
○ Visible cells only
○ Conditional formats
○ Data validation
● All
○ Same

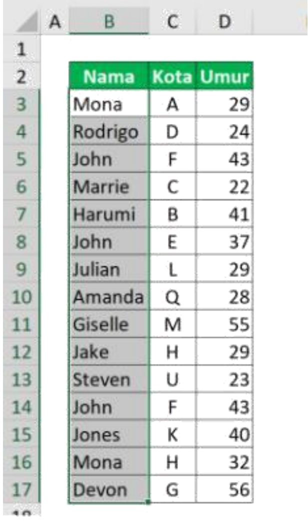
Figure 2. Go to special view visualization

Then click the OK button, all error cells in the worksheet will be blocked. cells that have errors can be marked by giving color / fill to the cell, for example, it becomes yellow (Yellow).

Remove Duplicate Data Or Unnecessary Data One of the excel features that can be used to find double data in excel or the same data, namely data that has duplicates is the Conditional Formatting feature [5].

Conditional formatting is one of the features of Microsoft Excel that is used to format cells according to the value of the particular cell in question.

The easiest way to find out double data in excel or double and duplicate data in excel is to use the conditional formatting feature. Selection of the range of data you want to find duplicate data.



	Nama	Kota	Umur
3	Mona	A	29
4	Rodrigo	D	24
5	John	F	43
6	Marrie	C	22
7	Harumi	B	41
8	John	E	37
9	Julian	L	29
10	Amanda	Q	28
11	Giselle	M	55
12	Jake	H	29
13	Steven	U	23
14	John	F	43
15	Jones	K	40
16	Mona	H	32
17	Devon	G	56

Figure 3. Data range selection

Choose menu: Home--Conditional Formatting--Highlights Cells Rules--Duplicates Values

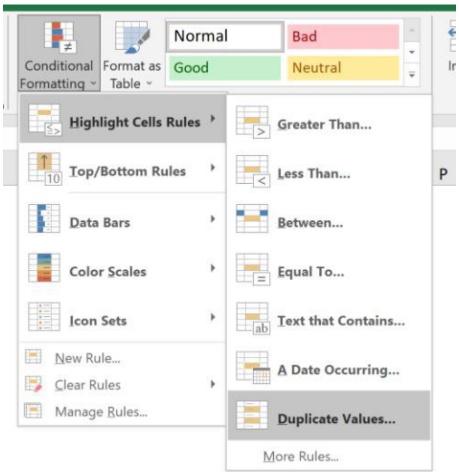
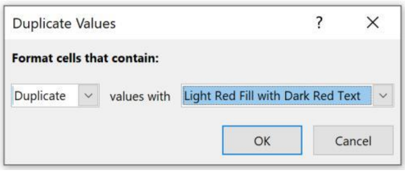


Figure 4. Visualization of the duplicate value menu.

Then specify the desired cell format can be seen as in Figure 5.



Duplicate Values

Format cells that contain:

Duplicate values with Light Red Fill with Dark Red Text

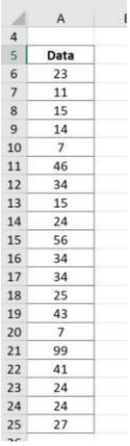
Figure 5. Duplicate value property settings Click OK, and Finish

Fix Structure Errors; Excel basically groups data into numeric and text. Numeric is to the right of the cell while the text is to the left of the cell. Data that fall into the numerical category can then perform number operations [6].

Whereas text cannot be performed number operations. For example, in the phone number data, although the value is in the form of a number, it cannot be done number operations such as multiplied, divided, added or subtracted. Therefore, phone number data is classified as text data. To fix the data structure error, you can use 2 functions, namely replace and subtitue.

Unwanted Outlier Filter In the process of data processing, sometimes data appears that at first glance appears out of sync or far apart using other data. This is what is claimed to use outliers or outliers [7]. It's okay to remove the outlier found but include a clear reason. Because, outlier filtering is indeed able to help the performance of the data that is being worked on. Even so, keep in mind that the emergence of outliers does not mean that the theory being worked on is wrong [8]. Quite the contrary, the existence of an outlier can be used as an indicator to choose the validity of the data.

Presented a data as follows to see the outlier of the data that mungin has.

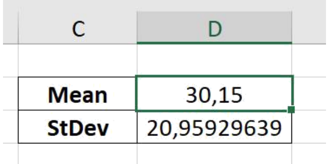


	A
5	Data
6	23
7	11
8	15
9	14
10	7
11	46
12	34
13	15
14	24
15	56
16	34
17	34
18	25
19	43
20	7
21	99
22	41
23	24
24	24
25	27

Figure 6. Data to be checked outlier

The steps to perform outlier detection on the data are create the average or Mean value of a series of numbers in column D5 by using the formula: Average(A6:A25).

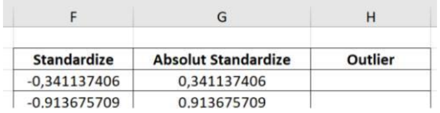
Create a standard deviation value from that line of numbers in column D6 by using the formula: Stdev.s(A6:A25).



C	D
Mean	30,15
StDev	20,95929639

Figure 7. The result of the discovery of the standard deviation value

Type the column labels in Cells F5, G5 and H5, with the labels: Standardize, Absolut Standardize and Outlier.



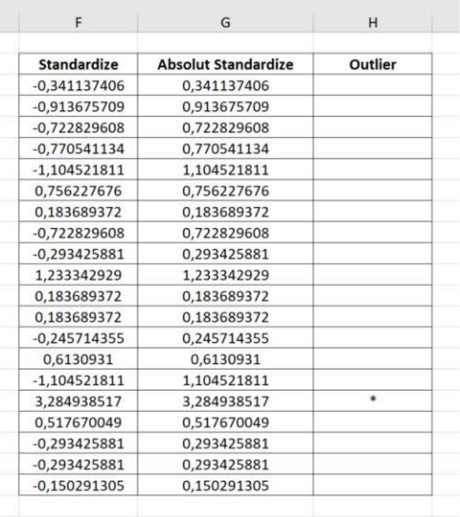
F	G	H
Standardize	**Absolut Standardize**	**Outlier**
-0,341137406	0,341137406	
-0.913675709	0.913675709	

Figure 8. Data labeling

Calculate the standardize value of a series of numbers in cell A6:A25 in cell F6:F25, how to type the formula in cell F6, namely: =STANDARDIZE(A6,D$5,D$6). The value in step 1 is standardized based on the Mean and Standard Deviation values in steps 2 and 3. Copy Cell F6 and Paste cell F7 through F25.

Next, calculate the absolute value of the standardized value in steps 5 and 6 by typing the formula in cell G6, namely: =ABS(F6). Copy cell F6 and paste cell F7 through F25 to determine whether the sample or observation is an outlier or not, then in cell H6, type the formula: =IF(G6>3,"*","").

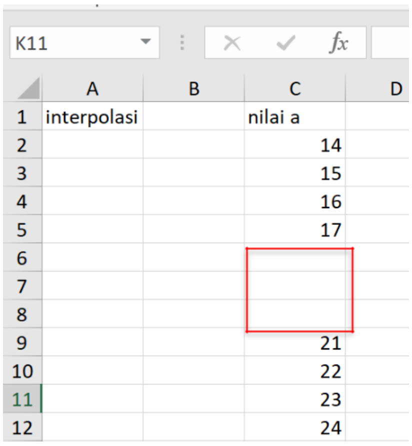
Copy cell H6 and Paste in cell H7 to H25. Look at the results in Cell H7:H25, if there is a * sign, then the observation is an outlier.



F	G	H
Standardize	**Absolut Standardize**	**Outlier**
-0,341137406	0,341137406	
-0,913675709	0,913675709	
-0,722829608	0,722829608	
-0,770541134	0,770541134	
-1,104521811	1,104521811	
0,756227676	0,756227676	
0,183689372	0,183689372	
-0,722829608	0,722829608	
-0,293425881	0,293425881	
1,233342929	1,233342929	
0,183689372	0,183689372	
0,183689372	0,183689372	
-0,245714355	0,245714355	
0,6130931	0,6130931	
-1,104521811	1,104521811	
3,284938517	3,284938517	*
0,517670049	0,517670049	
-0,293425881	0,293425881	
-0,293425881	0,293425881	
-0,150291305	0,150291305	

Figure 9. Outlier value detection results

Handling lost data; Missing Value is the loss of some data that has been obtained. In the world of data science, missing value is closely related to the process of data disputes (data wrangling) before later data analysis and prediction will be carried out [9][10]. Data wrangling is an activity of uniformizing data or cleaning data (cleaning data) from dirty (raw) data to data that will be ready to be used for analysis. The gross (raw) data in question is data that is indicated that there is still a uniformity of format, missing values appear in the data, and there are still additional suffixes, prefixes and others. Usually, a data scientist spends 60% of his time in carrying out this process. Because the facts show that 75% of the data owned by the company is gross data. In excel the process of correcting missing values can use linear interpolation techniques.



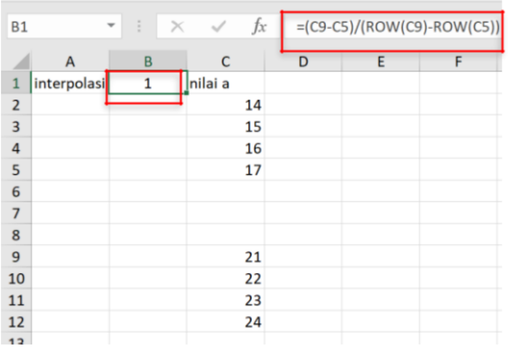
K11

	A	B	C	D
1	interpolasi		nilai a	
2			14	
3			15	
4			16	
5			17	
6				
7				
8				
9			21	
10			22	
11			23	
12			24	

Figure 10. Missing value position

Cells B6, B7 and B8 are missing or unfilled. The possible cause is incorrect input or the system is not validated so that the value in the kolm can be filled in blanks. To solve the situation, you can use the formula =(NumbersExpected–NumbersNexted)/(ROW(NumbersInst)–ROW(NumbersNexted)

When entered in excel the formula is as follows

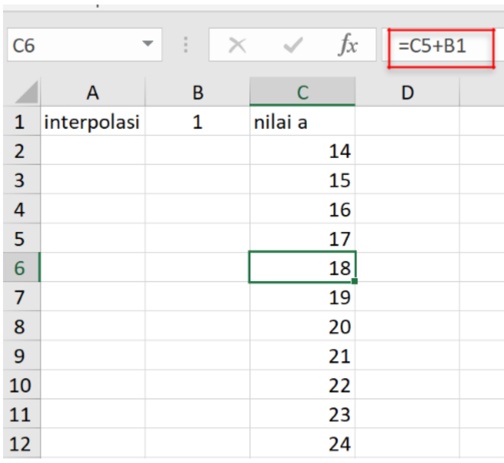


B1 =(C9-C5)/(ROW(C9)-ROW(C5))

	A	B	C	D	E	F
1	interpolasi	1	nilai a			
2			14			
3			15			
4			16			
5			17			
6						
7						
8						
9			21			
10			22			
11			23			
12			24			

Figure 11. Interpolated position after searching using the formula

Next enter the value for cell c6 with the formula c5 + B1 (the interpolation value found), then the result is as follows



C6 =C5+B1

	A	B	C	D
1	interpolasi	1	nilai a	
2			14	
3			15	
4			16	
5			17	
6			18	
7			19	
8			20	
9			21	
10			22	
11			23	
12			24	

Figure 12. The process of inputting missing numbers by summing the smallest numbers with the interpolated values obtained

Data Validation; The last step of data cleaning is validation. In microsoft excel limit the value or text entered in a cell or range in accordance with certain criteria that are desired. For example, a cell can only be filled with the numbers 1-10, limited to a certain list of text, can only be filled with a date format, and so on. For this kind of need, Microsoft Excel provides a feature called "Data Validation" or excel data validation [11].

How to make data validation by using data validation in excel to limit the contents of a cell and or excel range, is as follows:

Select the Cell/Range for which you will set the validation data.

Select the Data Validation Menu on the Data--Group Data Tools Tab.

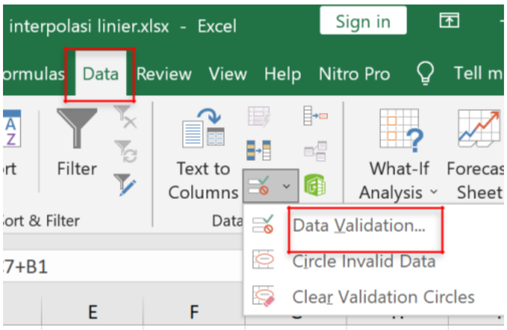
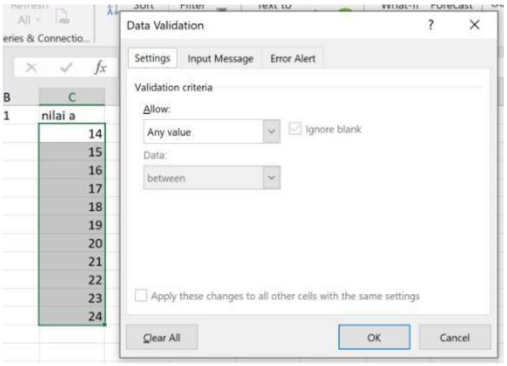


Figure 13. Data validation menu

The above steps can also be done with the keyboard shortcut Alt + A + V + V.

After the Data Validation option box appears, setting the data validation settings or limiting the contents of the desired cells.



Allow: Any value

☑ Ignore blank

Data: between

□ Apply these changes to all other cells with the same settings

Figure 14. Visualization of the data validation menu

The above step can also be done with the keyboard shortcut Alt + A + V + V. After the Data Validation option box appears, setting data validation settings or limiting the contents of the desired cell. Finally Select/Click OK to complete and apply the data validation settings.

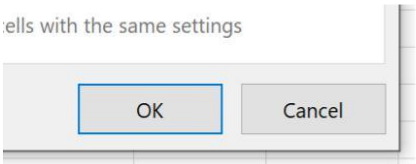
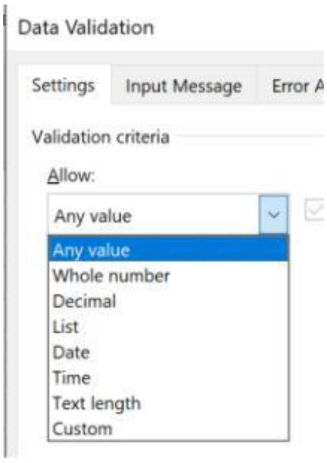


Figure 15. Click ok to continue

III. RESULT AND ANALYSIS

As previously discussed, there are many factors that can affect the results of analysis in data science. One of the keys to success lies in the early stages, namely data preparation which takes the longest and is the most difficult to do. Data Preparation is the process of preparing data so that it is ready to be processed for the next stage [12]. Many things can be found at the data preparation stage that can hinder the successful implementation of data science. For example: poor data quality (missing value, duplicate data, and incomplete data) and unbalanced datasets (there is data that is too dominant, so data science activities cannot make predictions correctly) [5][7]. Errors at the data preparation stage will affect the results data science analysis. In addition to data preparation, the data exploration stage is no less complex.

This paper will focus on the data validation process with the features in Excel. Criteria Settings on Data Validation When viewed in more detail in figure 14 Visualization of the data validation menu, there is an allow menu. If clicked, it will bring up 8 validation criteria menus in excel.



Data Validation

Settings | Input Message | Error A

Validation criteria

Allow:

Any value

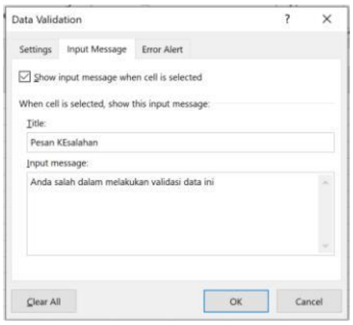
- Any value
- Whole number
- Decimal
- List
- Date
- Time
- Text length
- Custom

Figure 16. Data validation criteria in excel

Information : **Any Value**: There is no limit to the contents of the cell. All values (text or number data) can be entered into cells. **Whole Number**: Limit the contents of the Cell to numeric numbers or numbers only. **Decimal**: Same as Whole Number, it's just that we can set the validation value used to the accuracy of the decimal number. **List**: The contents of the cell are only limited to the list or lists we specify. **Date**: Restricts cell contents to date criteria. **Time**: Limits the contents of a cell to time criteria. **Text Lenght**: Limits the number of lengths of Text characters that can be inserted into excel cells. **Custom**: Define your own data validation criteria by entering an excel formula or a specific formula. After choosing one of the 8 types data validation above in the Data section: for some types of criteria can be subsequent arrangements are made. Menus that appear include, **between**: Only related data resides between two data settings that can inputted in the cell. **Not between**: Only in addition to the related data that is between the two data settings can be entered in the cell. **equal to**: Only data that is the same as in the settings can be inputted in the cell. Similar to the comparison operator "=". **not equal to**: Only data that is not the same as in the settings can be inputted in the cell. Similar to the comparison operator "<>". **greater than**: Only data larger than the data in the settings should be inputted in the cell. Similar to the comparison operator ">". **less than**: Only data smaller/less than the data in the settings is allowed to be inputted in the cell. Similar to the comparison operator "<". **greater than or equal to**: Only data greater than or equal to the data in the settings can be inputted on the cell. Similar to the comparison operator ">=". **less than or equal to**: Only data smaller/less or equal to that of the data in the setting can be inputted in the cell. Similar to the comparison operator "<=". Message input on data validation.

The INPUT MESSAGE settings tab in the "Data validation" option box is used to set the message displayed when a validated cell is selected.

By setting the Input Message, excel will display certain information when the validated cell is active.



Data Validation

Settings | Input Message | Error Alert

☑ Show input message when cell is selected

When cell is selected, show this input message:

Title: Pesan KEsalahan

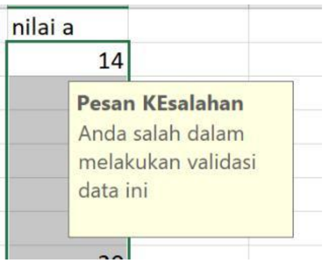
Input message: Anda salah dalam melakukan validasi data ini

Clear All | OK | Cancel

Figure 17. Input message if there is an error

If the Show input message when cell is selected section is checked, then when the cell is validated we select, excel will automatically display the message according to what we are atuar or we write in the Title and Input message sections. And instead leave the default or clear the title and Input message or uncheck not to displays a message when performing data input.

The Title section is the title of the message while the input message is the content of the message to be displayed. If we set the validation data for cell A1 for example and the input message we set as above, then when cell A1 we select it, a message will appear as shown below:



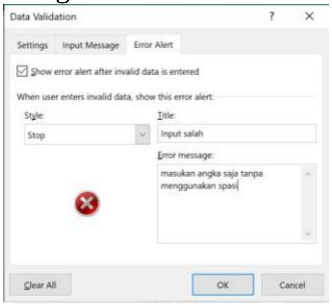
nilai a

14

Pesan KEsalahan
Anda salah dalam melakukan validasi data ini

Figure 18. Error Message that appears when incorrectly validating error alerts in Excel Validation Data

The ERROR ALERT settings tab in the "Data Validation" dialog box is used to set a warning message if the input data we entered in the validated cell does not match the settings we have set.



Data Validation

Settings | Input Message | Error Alert

☑ Show error alert after invalid data is entered

When user enters invalid data, show this error alert:

Style: Stop

Title: Input salah

Error message: masukan angka saja tanpa menggunakan spasi

Clear All | OK | Cancel

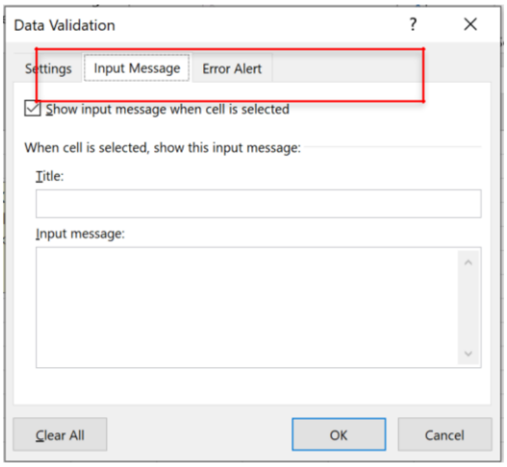
Figure 19. Error message

By activating the Error alert setting, the user will get information or a warning message if the data entered does not match the applied data validation.

For example, if you set a cell that can only be filled with numbers and then we input text into the cell, then this error warning message will appear automatically. If the data input is correct as per the enforced data validation then this message will not appear.

How to Clear Excel Validation Data; If you no longer need data validation, the ways to delete or eliminate data validation are as follows:

Selection or select the cell that we will invalidate the data. Select the Data Validation menu on the Data--Group Data Tools Tab. Select Clear All and then click OK.



☑ Show input message when cell is selected

When cell is selected, show this input message:

Figure 20. Clearing Data Validation Excel

VI. CONCLUSION

The process of becoming a data scientis can be starting with studying the type of data and transofrmation of data according to the purpose of the analysis. Such stages can be done with multiple applications that will facilitate and maintain data reliability. One of the applications that can be used to carry out activities scientis data is excel. Lots of features excel that can be used for perform data analysis especially when data preparation process. Functions, formulas and existing menus can support a scientis data in his work. For data over 1 million in size data rows should use the app else because excel is only capable holds 1,048,576 rows and 16,384 column. Of course it will make it difficult for a person scientis data if it has to work with data which is more than 1 million if using excel.

THANK-YOU NOTE

We would like to thank Universitas 'Aisyiyah Surakarta (www.aiska-university.ac.id) through P3M for funding this research.

REFERENCES

[1] Pandita, R., Parnin, C., Hermans, F., & Murphy-Hill, E. (2018). No half-measures: A study of manual and tool-assisted end-user programming tasks in Excel. 2018 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC), 95–103.J. Breckling, Ed., *The Analysis of Directional Time Series: Applications to Wind Speed and Direction*, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.

[2] Ruel, E., William, W., & Gillespie, B. J. (2018). Data cleaning. The Practice of Survey Research: Theory and Applications, 208–237.

[3] Hossain, E. (2021). MS Excel in Engineering Data. In Excel Crash Course for Engineers (pp. 169–242). Springer.

[4] Huang, Z., & He, Y. (2018). Auto-detect: Data-driven error detection in tables. *Proceedings of the 2018 International Conference on Management of Data*, 1377–1392.

[5] Wang, P., & He, Y. (2019). Uni-detect: A unified approach to automated error detection in tables. Proceedings of the 2019 International Conference on Management of Data, 811–828.

[6] Liu, R., Glover, K. P., Feasel, M. G., & Wallqvist, A. (2018). General approach to estimate error bars for quantitative structure– activity relationship predictions of molecular activity. Journal of Chemical Information and Modeling, 58(8), 1561– 1575.

[7] Wu, Z., Wu, Z., & Rilett, L. R. (2020). Outlier *Record*, filtering. *2674*(10), *Transportation Research* 167–176.

[8] Grech, V. (2018). WASP (Write a Scientific Paper) using Excel–3: Plotting data. *Early Human Development*, *117*, 110–112.

[9] Kaminskyi, R., Kunanets, N., Pasichnyk, V., Rzheuskyi, A., & Khudyi, A. (2018). Recovery Gaps in Experimental Data. COLINS, 108– 118.

[10] Biessmann, F., Salinas, D., Schelter, S., Schmidt, P., & Lange, D. (2018). " Deep"Learning for Missing Value Imputationin Tables with Non-numerical Data. Proceedings of the 27th ACM International Conference on Information and Knowledge Management, 2017–2025.

[11] Sofalvi, S., & Schueler, H. E. (2021). Assessment of Bioanalytical Method Validation Data Utilizing Heteroscedastic Seven-Point Linear Calibration

Curves by EZSTATSG1 Customized Microsoft Excel Template. Journal of Analytical Toxicology, 45(8), 772–779.

[12] Georgieva, P., Nikolova, E., & Orozova, D. (2020). Data Cleaning Techniques in Detecting Tendencies in Software Engineering. 2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO), 1028– 1033.