Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

Data Handling and Business Intelligence

Table of Contents

Part-1 3

- Need of Preprocessing data in excel: 3
- Visualising the data in excel 5
- Uses of pivot table: 6
- If condition in excel: 8

Part-2 10

- Male and female data 11
- Mean and median data 12
- Quick Clustering 14
- Method of Data mining and text mining: 15
- Advantages and Disadvantages of SPSS 16
- Advantages and Disadvantages of excel tool: 17

References: 18

Part-1

- **Need of Preprocessing data in excel:**

The data analysis process helps to make an easy evaluation process of the data, the important task in the data analysis is to convert all the row data in excel. In the steps of data analysis, the first process is to clean all the data and find the missing data for the evaluation process. Pre-processing makes it an easy process to evaluate. In the analysis of data, missing values do not create an impact on the evaluation. (Al-Yousif, *et.,al,*2021). Excel is the best tool for handling and preprocessing data. Some basic functions are used for pre-processing the data like If, IFERROR, LEFT, RIGHT, MID, FIND, etc. all these functions and methods help to clean the structure data. This transferable data is easily understandable. (Al-Yousif, *et.,al*,2021)

Data preprocessing is important because it checks all the row data before applying it in the excel tool and for mining. In the process of pre-processing accuracy of the data, Completeness of the data, Consistency of the data, timeliness of the data. All the data must be updated before starting the analysis of the data. After that check the believability of the data. If data is not trustable then the evaluation is not complete. (Al-Yousif, *et.,al*,2021)

Then finally start the interpretation of the data for understanding.

Some major tasks follow in the data pre-processing. Data cleaning, Data reduction, Data transformation, Data integration. Excel tool helps to examine all the data. Because excel is a brilliant data tool that is helpful for analytics in any type of project. (Al-Yousif, *et.,al*,2021)

An additional way to pre-process the data is to compare the data in different phases like finding the data requirements, gathering the data, processing the data, cleaning all the data, Analysis all the data, and communicating of data. Analysis of the data is a valuable skill because it helps to make better judgments. In excel, the Pivot table is the tool that is used for the analysis. Also, so many varieties of tools are available for the data. The conditional formation, formula addicting, and quick analysis of the function are available for the pre-proceedings with the data. Data preprocessing is helping to find the missing values and the noisy data or errors base data cannot use in machine learning. Only pre-processed data is used in machine learning. Importing,

Encoding, finding, and splitting are all the techniques for pre-processing the data. (Al-Yousif, *et.,al*,2021)

- **Analysing the data in Excel:**

Excel is the most famous tool for analysing data. In the pivot table, and histogram so many tools are available for evaluating the data. Also, conditional formatting and presentation in charts help to make a good presentation. So many different features are available in excel for analysing the data. Excel is a tool that empowers people to understand data in natural language. Also, Excel helps with visual summaries, pattern, and trends data. Excel is also used for business data like sales data, profit data etc. Excel analysis gives the future growth to the business. A pivot table allows the summarise data as counts, sums, and averages in excel. Excel is a time-saving tool because it transfers all the data quickly and builds the reports. Excel is also used for analysing static data. Addition, subtraction, multiply and divide all functions are easily accessible in excel. T-tests, correlation, and also the part of the excel. In the excel output of data is presented in a proper way. (Chaparro-Pelaez, et.,al, 2019)

In the analysis of Excel, the mean and standard deviation is the part of the descriptive analysis. The data evaluation tool so many options are available for analysing the data like columns are available for inputting the range. Mean and standard deviation are also analysed in excel and differentiate the data in the pivot table. Correlation data is also analysed in excel. Two sample tests paired t-test, linear regression, simple frequencies. Some more functions are available which are helpful for analysing the data like rank system, trend, outlier, majority data, etc. Merage functions are also available in Excel. In excel millions of cells are available for analysing the data. Also Excel gives access to take the data from other locations. Data cleaning option is the basic function of Excel. The excel number is also changed in the text function. (Chaparro-Pelaez, et.,al, 2019)

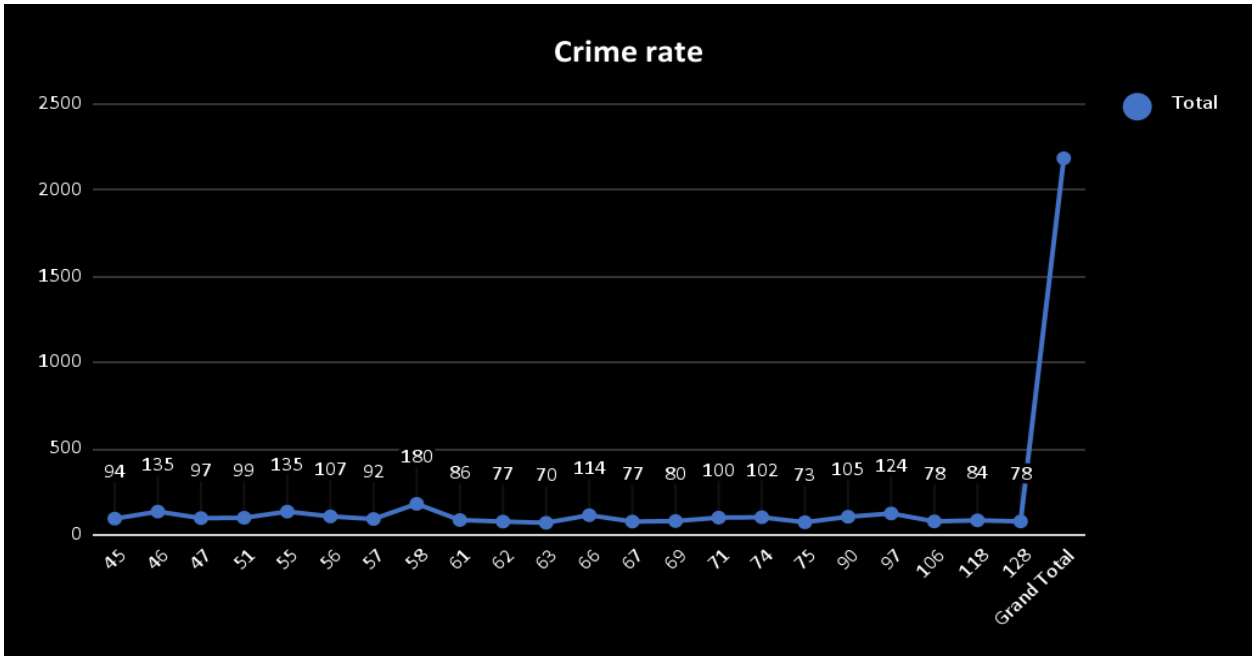
- **Visualising the data in excel**

There are so many tools available in the market for the visualisation of the data and evaluating the data of business. Microsoft Excel is the tool that helps make the analysis process. If someone has no access to Microsoft Excel then the analyst can work on google sheets or some similar option for the analysis. Excel is the visualisation tool that is used by professionals for the analysis of data. The most typical work in excel is to put the row data in a manual way. So many steps are involved in Excel for analysis. Excel visualisation helps to improve the performance of the business after the analysis because excel presents all the data in a durable way. (Chaparro-Pelaez, et.,al, 2019)

Effective visualisation in excel to use charts and graphs for the presentation and qualitative and quantitative analysis of all the data is easily analysed in the excel tool. Pie charts, histograms, Area charts, Scatter charts, and bar charts all are tools which are helping to visualise the data. But some techniques are also available in excel like Gantt charts, Timelines, Heat maps, Bullet graphs, and highlighting the table helps to analyse the data. Data insert in excel from the main menu and choose the file which wants to analyse the data. Some different types of ways are available for the visualisation in excel. V lookup and HLookup functions are also used for the analysis of the data. In the V lookup function data are transferred from the different sheets. Filter option is also available in excel which helps to analyse only those data which needs for the analysis. (Chaparro-Pelaez, et.,al, 2019)

- **Uses of pivot table:**

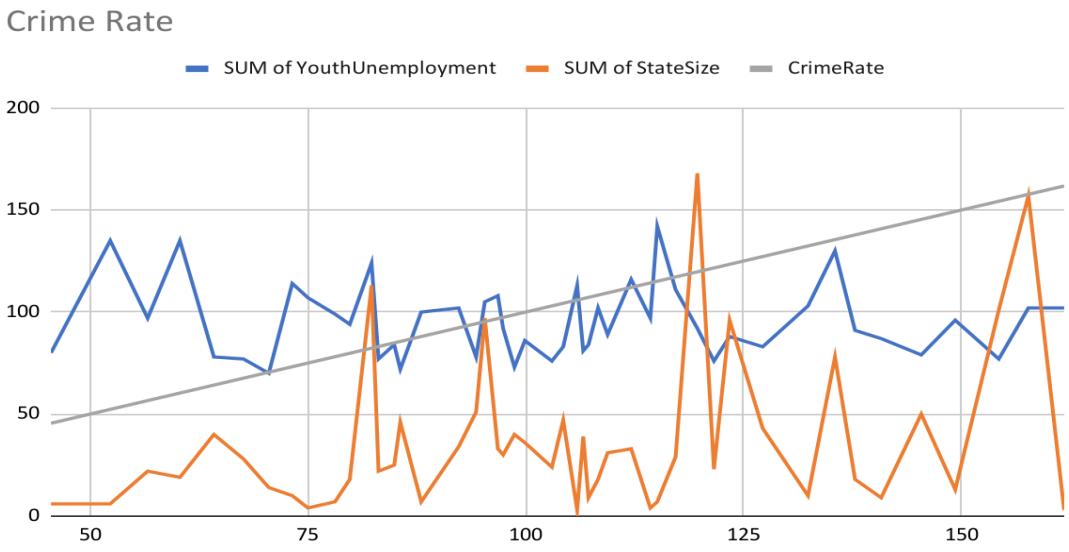
ExpenditureYear0	Sum of YouthUnemployment
45	94
46	135
47	97
51	99
55	135
56	107
57	92
58	180
61	86
62	77
63	70
66	114
67	77
69	80
71	100
74	102
75	73
90	105
97	124
106	78
118	84
128	78
Grand Total	**2187**



The above table gives the information of crime rate and all the data prepared with the help of pivot table. The pivot has four options that are filter, axis, legend and fourth one is values. (Nayak, et.,al,2016). Grand total of youth unemployment is 2187 which helps to understand the data of unemployment. 180 is the highest value of youth unemployment rate. This pivot table summarises all the data in a systematic way. Pivot tables represent the data in a detailed manner or querying a large amount of the data in a systematic way. Filtering, conditionally, grouping sorting formatting system available in the pivot table, all moving the column to row or row to column function available in the pivot table. The above chart is prepared with the help of the crime datasheet and in this sheet, other data is also available for analysis. (Nayak, et.,al,2016).

- **If condition in excel:**

Crime rate	Sum of Youth Unemployment	Sum of StateSize	
103	76	24	low
104.3	83	47	low
105.9	113	3	high
106.6	81	39	low
107.2	84	9	low
108.3	102	18	high
109.4	89	31	low
112.1	116	33	high
114.3	97	4	low
115.1	142	7	high
117.2	111	29	high
119.7	92	168	low
121.6	76	23	low
123.4	88	96	low
127.2	83	43	low
132.4	103	10	high
135.5	130	78	high
137.8	91	18	low
140.8	87	9	low
145.4	79	50	low
149.3	96	13	low
154.3	77	101	low
157.7	102	157	high
161.8	102	3	high
Grand Total	2356	1047	



Interpretation: The above table shows the uses of if condition in the excel tool and present on chart this chart prepare with help of line chart and present the crime rate. The above table has three variables: first is crime rate, second is sum of youth unemployment and third one is sum of state size and fourth one is representing the frequency level on the basis of the high and low systems. (Gupta, and Shringare, 2018) The highest crime rate is 161.8 and the sum of youth is 2356 and the size of the state is 1047 all the data are prepared with the help of if condition in the excel tool. First in this table prepare your data with the help of the pivot table. pivot table is the tool that is available in excel and helps diversify all the data. and present in a systematic manner. In this chart the blue line represents the crime rate, the red line presents the sum of youth unemployment and the green line represents the sum of state size. (Gupta, and Shringare,2018)

Part-2

- **Vanilla Flavour**

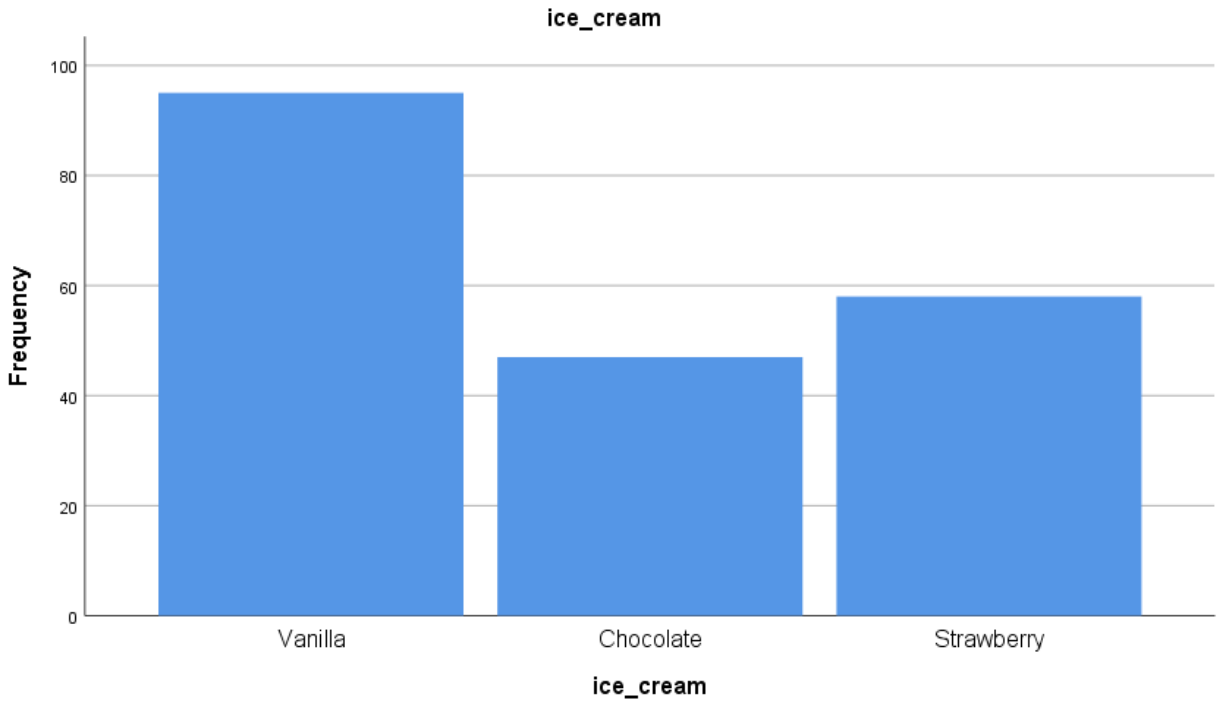
Statistics

ice_cream

N	Valid	200
	Missing	0

ice_cream

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Vanilla	95	47.5	47.5	47.5
	Chocolate	47	23.5	23.5	71.0
	Strawberry	58	29.0	29.0	100.0
	Total	200	100.0	100.0	



The above chart prepared in the SPSS tool represents the ice cream flavour that is vanilla, chocolate and strawberry. The highest use of ice cream is vanilla and second one is strawberry. The highest frequency variable is vanilla that is 95% and lowest frequency is chocolate that is 47%.

- **Male and female data**

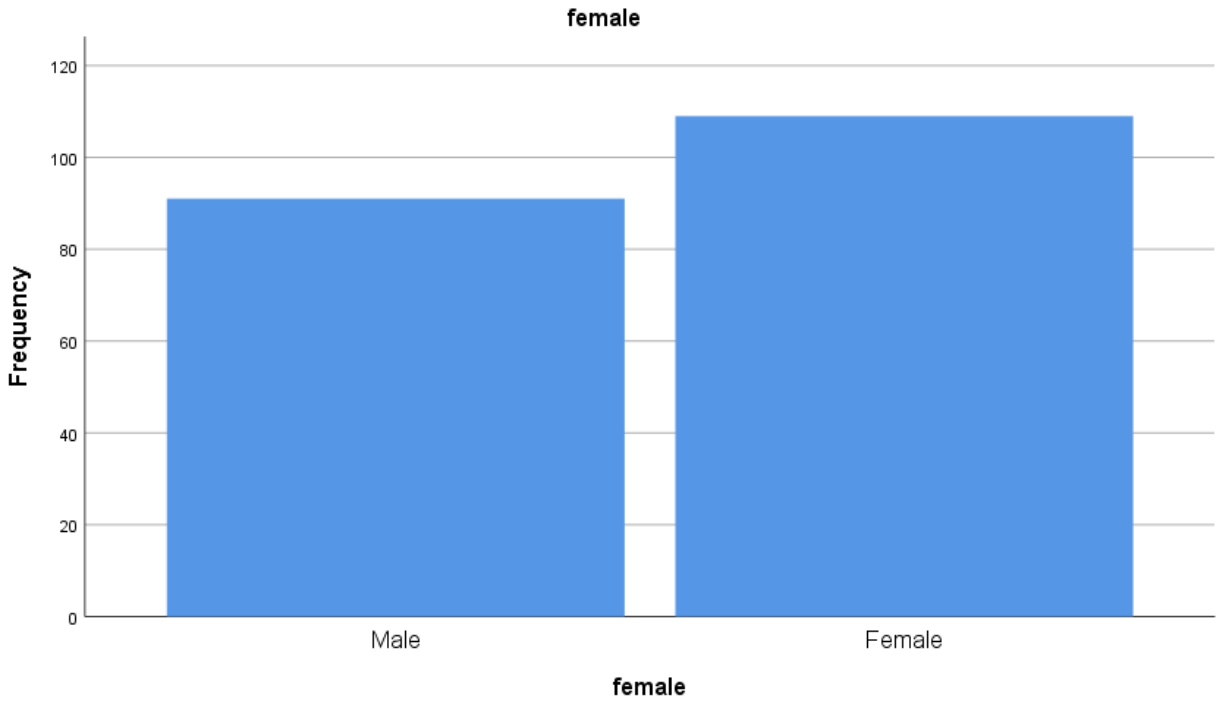
Statistics

female

N	Valid	200
	Missing	0

female

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	91	45.5	45.5	45.5
	Female	109	54.5	54.5	100.0
	Total	200	100.0	100.0	



This data analysis depends on the male and female data . The above chart is prepared with the help of 200 participants . This data shows the female is more than the male. No missing values are available in this research. The male frequency is 91 and female frequency is 109. The present of male percentage is 45.5% and the female percentage is 54.5%. This all the data is prepared with the help of SPSS tool. SPSS tools help to evaluate all the data separately. (Purwanto, et.,al, 2021)

- **Mean and median data**

Statistics

ice_cream

N	Valid	200
	Missing	0
Mean		1.82
Median		2.00

ice_cream

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Vanilla	95	47.5	47.5	47.5
	Chocolate	47	23.5	23.5	71.0
	Strawberry	58	29.0	29.0	100.0
	Total	200	100.0	100.0	

Interpretation; Table represents variables of ice cream that name is vanilla, chocolate, strawberry. The blue colour presents the vanilla flavour and Red colour represents chocolate, and the green colour presents strawberry flour. In this research, the table mean value is 1.82 and the median value is 2.00. The highest frequency value is 47.5 as vanilla. (Kippers, et.,al, 2018)

- ## Quick Clustering

Initial Cluster Centres

	Cluster		
	1	2	3
ice_cream	2	1	3

Iteration History[a]

Iteration	Change in Cluster Centres 1	2	3
1	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centres. The maximum absolute coordinate change for any centre is .000. The current iteration is 1. The minimum distance between initial centres is 1.000.

Final Cluster Centres

	Cluster 1	2	3
ice_cream	2	1	3

Number of Cases in each Cluster

Cluster	1	47.000
	2	95.000
	3	58.000
Valid		200.000
Missing		.000

The above table gives the information for the variable which is available in this research. This research is based on the cluster and variables of ice cream. Three types of clusters are available in this research that are 1, 2, 3, and all have different variables that are 47, 95, and 58 respectively. In this research no missing values are available. (Nayak, et al,2016). Three types of tables are available in this research that are clusters head, interaction history, and interaction history. All the tables help to make the coordination between all the tables. The highest cluster is 95 and the second highest cluster is 58 and the final cluster is 47.

- ## Method of Data mining and text mining:

The data mining process helps with the association, classification, outlier detection, regression, clustering, and prediction of the data; it is a process that gives the conclusion of the data. The data mining process analyzes the different categories with different sections. Data mining tools give information and knowledge from a different set of data. Uses interface, pattern evaluation, data mining engine, database, and data cleaning is the process. (Nayak, et al, 2016)

Different types of mining methods:

This is all the data that help to resolve the issue of the problem:

- **Prediction:** this method is used for the analysis of the future data and predict the data. All the predictions are based on past experience and present trends data. The prediction method helps to make a combination with the other methods like relation, trend analysis, classification, and pattern matching. For example, if a sales manager wants to improve the business sale then the analysis is based on past data. (Chaparro-Pelaez, et al, 2019)
- **Pattern tracking:** This method is used for the analysis of frequent data for a certain period of time. Like the winter clothing business, when winter business comes that time all the businessmen purchase winter clothes and this analysis is based on the previous pattern.
- **Decision tree:** This method works like the tree structure like every branch of the data denotes the outcome of the test. Terminal mentions the label of class.
- **Classification:** In this method, all the data is divided into different groups and analysed the behaviour of the data.

Text mining method:

This method evaluates all the unstructured text data. This is the process of conducting the mining of the data and finding insights into the data. This mining method is complete with the help of a mining tool.

- **Information Retrieval:** In this process, all the data is connected with the pattern and all the information is relevant to the actual data set. IR also tracks the behaviour of the user.

Like, google search engine is the best example of IR because the google engine works according to the behaviour of the user.

- **Natural language processing:** NLP is a summarization of all the data and tagging the part of speech, analysis of sentiments, and Text categorization.

  Clustering: Clustering helps to classify all the data with the algorithms and patterns with the implementation. In the process of clustering, all the data is unlabeled textual. Clustering methods are used to sort the objects defined by a set of numerical properties in such a way that the objects within a group are more similar than the objects in different groups. Clustering is a method to find a set of clusters that best fits without any class information. There are three fundamental criteria to investigate the clustering analysis: external criteria, internal criteria, and relative criteria (Chaparro-Pelaez et al., 2019)

Advantages and Disadvantages of SPSS

Advantage:

- SPSS is statistical software that compares all the data with each other. Many types of statistical tests are available in the SPSS tool. Result interpretation depends on the relative data. In the SPSS all the data can be expanded.
- SPPS gives the measurement of the data in one row. Also, SPPS is a highly flexible tool.

Disadvantage:

- SPSS is an expensive tool and the uses of the graph tool are not easily manageable. SPPS is not able to analyse the big data.
- SPSS tools need proper training and this tool has limited access to users. Also, the SPSS tool is a costly tool.

Advantages and Disadvantages of Excel tool:

Advantages:

- This tool is helpful for office work in the finance department and accounting work. Excel helps to organise all the data. Also, in excel, all the files are password

protected. Excel manages a lot of data. The use of excel is an easy process and anywhere use the tool. Excel has multiple spreadsheets in a single excel file.

- Excel has so many options like calculating the columns. The excel tool table is easy to refer to and data select. It is helpful for making presentations and this tool is majorly based on the formula. (Sugiyarti, et.,al, 2018)

Disadvantages:

- In Excel viruses attack is also possible because excel has the macros in it. And the hackers can inject the virus into macros. This software tool is difficult to learn as compared to MS Word. This software has a high price.
- In this tool, also some errors are found like handling errors. This type of error creates difficulty for the software. Also, Excel has no fast autofill features. The most problematic error in Excel is calculation. (Sugiyarti, et.,al, 2018)

References:

AL-YOUSIF, S., JAENUL, A., AL-DAYYENI, W., ALAMOODI, A., JABORI, I., TAHIR, N.M., ALI AMER, A.A., CÖMERT, Z., AL-SHAREEFI, N. and SALEH, A.H., (2021). A systematic review of automated pre-processing, feature extraction and classification of cardiotocography. *PeerJ Computer Science.* Vol.7, p.e452 [Online] Available at: https://www.proquest.com/docview/2518855041?accountid=130472&parentSessionId=k9AhQAb%2BQyI6eRk934whuIA3P4HySDYhgp5XFV9temk%3D&pq-origsite=primo

Nayak, A.S., Kanive, A.P., Chandavekar, N. and Balasubramani, R., (2016). *Survey on pre-processing techniques for text mining.* International Journal of Engineering and Computer Science, 5(6), pp.16875-16879. [Online] Available at: https://www.researchgate.net/profile/Balasubramani-Ramasamy/publication/319716444_Survey_on_Pre-Processing_Techniques_for_Text_Mining/links/5b90b70092851c6b7ec76e55/Survey-on-Pre-Processing-Techniques-for-Text-Mining.pdf

Chaparro-Pelaez, J. Iglesias-Pradas, S. Rodriguez-Sedano, F.J. and Aquila-Natale, E. (2019). *Extraction processing and visualisation of pear assessment data in moodle.* Applied sciences, 10(1), pp.163. [Online] Available at: https://www.researchgate.net/publication/338151458_Extraction_Processing_and_Visualization_of_Peer_Assessment_Data_in_Moodle

Kippers, W.B. Poortman, C.L. Schildkamp, K. and Visscher, A.J. (2018). *Data literacy: What do educators learn and struggle with during a data use intervention?* Studies in Educational Evaluation, pp.21-31. [Online] Available at: https://www-sciencedirect-com.ezproxy.uwtsd.ac.uk/science/article/pii/S0191491X17301700

Gupta, P. and Shringare, K., (2018). *Analysing the mobile phone buying behaviour using SPSS.* Aadya National Journal of Management and Technology, pp.30-37. [Online] Available at: http://ischolar.info/index.php/AADYA/article/view/183159/170335

Sugiyarti, E. Jasmi, K.A. Basiron, B. Huda, M., Shankar, K. and Maseleno, A., 2018. *The decision support system of scholarship grantee selection uses data mining.* International Journal of Pure and Applied Mathematics, 119(15), pp.2239-2249.

https://www.academia.edu/download/57052927/DECISION_SUPPORT_SYSTEM_OF_SCHOLARSHIP_GRANTEE.pdf