

Data Handling and Business Intelligence

Student name: Albena Yosifova-Stefanova

Student ID: 2014829

Module Title: Data Handling and Business Intelligence

Assignment Title: Assessment 2

Lecture`s name: Joy Botu

Table of Contents

Part 1	3
Evaluation of excel for pre-processing	3
Evaluation of excel for analysis	4
Evaluation of excel for visualisation.....	5
Pivot table	6
If function.....	9
Percentage share of attributes in each region.....	12
Part 2	13
Vanilla Flavour liking	13
Male and Female students.....	14
Mean and median	15
Clustering.....	16
Data mining and text mining methods	18
Data Mining	18
Text mining.....	19
Comparison of SPSS and Excel	20
References.....	21

Part 1

Evaluation of excel for pre-processing

In it, the excel tool is used for pre-processing the data which is available to the user in the raw form so that data which is used for analysis can be used easily as it will be free from the errors which are present in it as in this the cleaning of the data also happens. It also helps in knowing if there is any missing value in the data and it is known when the function is applied is by going to a special dialogue box and in it selecting the blank from which all the values which are blank in the data can be known and those can be excluded from the analysis and this will result in the accurate analysis of the data as when missing values are there the tool count those values also and they are included in the analysis which affects the output of the data and also the decisions which are made with the help of it (Narayan and Nayak, 2021).

The available data can sometimes be noisy so to know if the data is noisy excel has various methods to find out it and regression is one such method in which the independent variables are identified so that the dependent variable can be predicted with it. Apart from it, the cluster analysis is another such method through which it can be identified as in it the clusters are identified and each cluster is analysed to find such data another method to find it is the method of binning in which data which is there is divided into various groups and each group is analysed separately. Another way through which pre-processing of the data in excel is done is by keeping only those data which is needed in the analysis and removing all the data which is extra and the data which is kept is pre-processed so that if there are any outliers or missing values in it can be identified beforehand (Wiyono, Christian and Abidin, 2021).

Evaluation of excel for analysis

The tool excel has various features for analysing the data which is there as from a very easy method the advanced method is available for the analysis. To know what is the total of the row or column the sum function is there and like these functions, many functions are available for the analysis. The other advanced method is there in this tool to perform the analysis of those data which are large and the analysis of that data cannot be done with those methods which are available this table is used as the data is reduced in it. Also, in this table the data can be entered into a row or column depending on the needs and the filter option is also there in which the data is selected which is needed to be shown in the table and also in it, the values which are there in row or column can be shown according to the sum, average, count and through many other things. Also, to conduct the analysis which is high level there is an option of going to the data analysis tab in which there are various options like correlation, regression, histogram, descriptive statistics, f test, z test and many other options are there (Cosma, 2019).

The other things through which analysis can be performed are with the help of other functions like sorting and filtering in which the data can be selected according to the users using the filter option and that data can be arranged in descending or ascending order using the sorting option. Also, the nested functions are there in this tool through which two or more functions can be used together (Suwannahong, et.al., 2021).

Evaluation of excel for visualisation

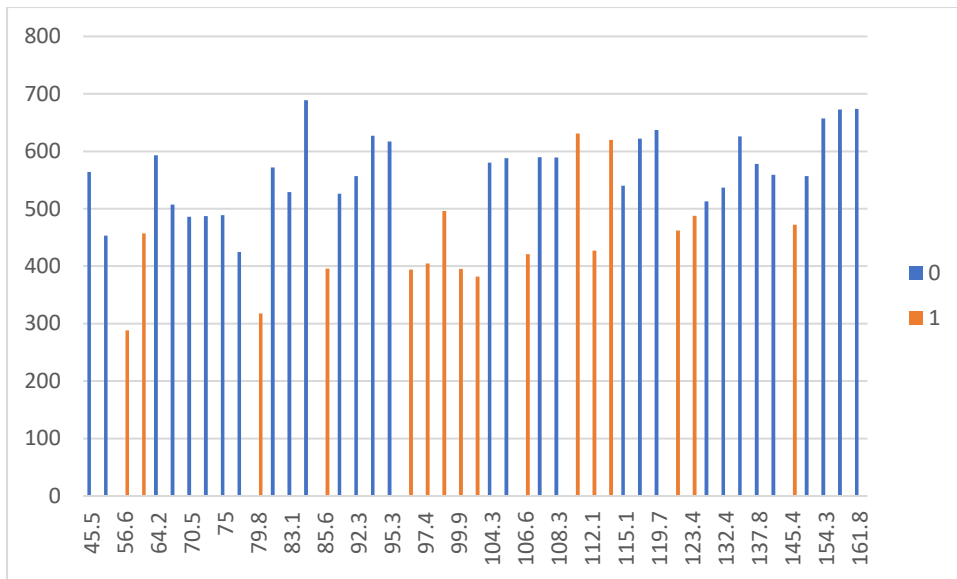
For conducting the visualisation of data which is given or obtained from the primary or secondary source excel is the best tool for it as it contains many features and has a variety of charts under it. In excel, the choice of the chart which is chosen is dependent upon the data and also what is needed to be described with the help of that data so that the right selection can happen as every chart has its characteristics and the type of charts which are there range from simple charts which are pie chart, bar chart, line chart and many more to the complex charts which are Gantt charts, heat maps, box and whisker and many other varieties of charts. To visualise the data using this tool the steps that are needed to be followed are very simple first the data is needed to be selected and after that click the graph which is needed by going on the insert and selecting it and if the user does not understand which graph to use then they can go to the recommended charts and select from there (Mohammad, Nuar and Yahya, 2021).

In this tool there is the feature that the graph which is made can be edited that is the design of it can be changed in which the background can be edited using various background colours, the title can be added along with many other things which can be added to the graph which are trendline, data labels, axis labels and many other things. Also, by making a simple column chart the overline grid chart can be made easier with the help of it which the data is divided in the column to showcase various segments and if the chart type is needed to be changed it can simply be done by just right click and selecting the change chart and choosing the other chart option from it. Also, if any type of relationship or differences are needed to be showcased then it can be done with the help of combo charts in which two or more chart types can be made easily in one graph (Maknunah, et.al., 2021).

Pivot table

Sum of Wage	Column Labels		
Row Labels	0	1	Grand Total
45.5	564		564
52.3	453		453
56.6		288	288
60.3		457	457
64.2	593		593
67.6	507		507
70.5	486		486
73.2	487		487
75	489		489
78.1	425		425
79.8		318	318
82.3	572		572
83.1	529		529
84.9	689		689
85.6		396	396
88	526		526
92.3	557		557
94.3	627		627
95.3	617		617
96.8		394	394
97.4		405	405
98.7		496	496
99.9		395	395
103		382	382
104.3	580		580
105.9	588		588
106.6		421	421

107.2	590		590
108.3	589		589
109.4		631	631
112.1		427	427
114.3		620	620
115.1	540		540
117.2	622		622
119.7	637		637
121.6		462	462
123.4		488	488
127.2	513		513
132.4	537		537
135.5	626		626
137.8	578		578
140.8	559		559
145.4		472	472
149.3	557		557
154.3	657		657
157.7	673		673
161.8	674		674
Grand Total	17641	7052	24693

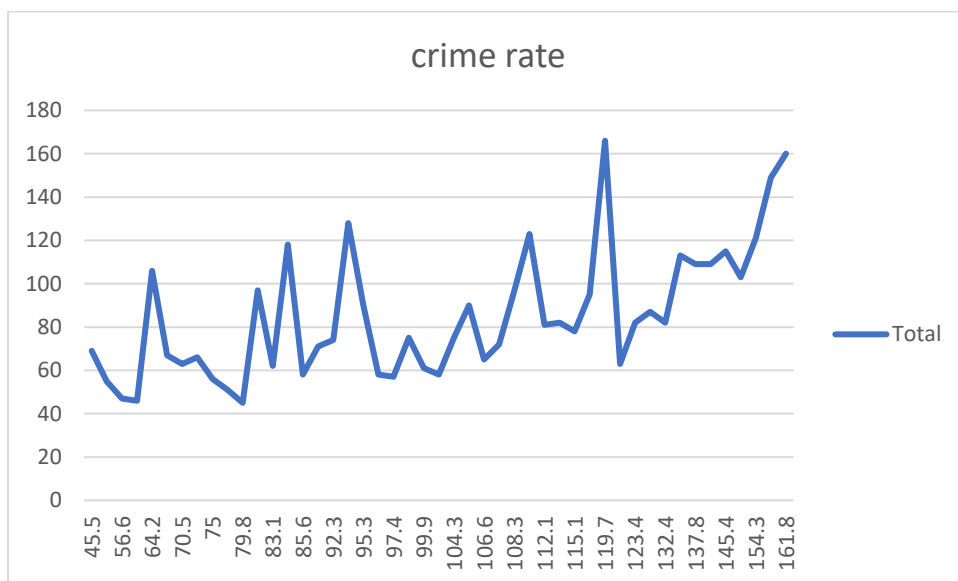


Interpretation: In the column graph and table which have been shown above the crime rate has been given according to the region and the wages which are being earned by the residents of that region. It can be observed that wages of the northern region are more than the southern region as the sum which has come out of both the regions are 17641 and 7052 respectively. Also, the blue colour in the graph is depicting region zero which is the northern and orange is of one which is the southern region. Various crime rates are being given and each crime rate is having a rate of only one region and these rates are from 45.5 to 161.8. The higher wage rate is 674 in the northern region in which the crime rate is 161.8 respectively (Kandpal, 2021).

If function

Row Labels	Sum of ExpenditureYear0	If condition
45.5	69	low
52.3	55	low
56.6	47	low
60.3	46	low
64.2	106	high
67.6	67	low
70.5	63	low
73.2	66	low
75	56	low
78.1	51	low
79.8	45	low
82.3	97	high
83.1	62	low
84.9	118	high
85.6	58	low
88	71	high
92.3	74	high
94.3	128	high
95.3	90	high
96.8	58	low
97.4	57	low
98.7	75	high
99.9	61	low
103	58	low
104.3	75	high
105.9	90	high
106.6	65	low
107.2	72	high
108.3	97	high

109.4	123	high
112.1	81	high
114.3	82	high
115.1	78	high
117.2	95	high
119.7	166	high
121.6	63	low
123.4	82	high
127.2	87	high
132.4	82	high
135.5	113	high
137.8	109	high
140.8	109	high
145.4	115	high
149.3	103	high
154.3	121	high
157.7	149	high
161.8	160	high
Grand Total	3995	



Interpretation: In the line graph and the table are depicted above the status of the crime rate has been given based on the condition if which has been used on expenditure year and the condition

which is used is that if the expenditure that has been done is more than the sixty then it is referred as the high crime rate and if it is low than that then the crime rate is low. Based on it the result which has come out to be is that more values of higher crime rate are there as the values in the low crime rate are only eighteen and rest of the values are of higher expenditure that has been done. Also, it can be observed from the graph that expenditure that is there is fluctuating according to the crime rate and lower expenditure which is 45 is having a crime rate of 79.8 respectively (Narayan and Nayak, 2021).

Percentage share of attributes in each region

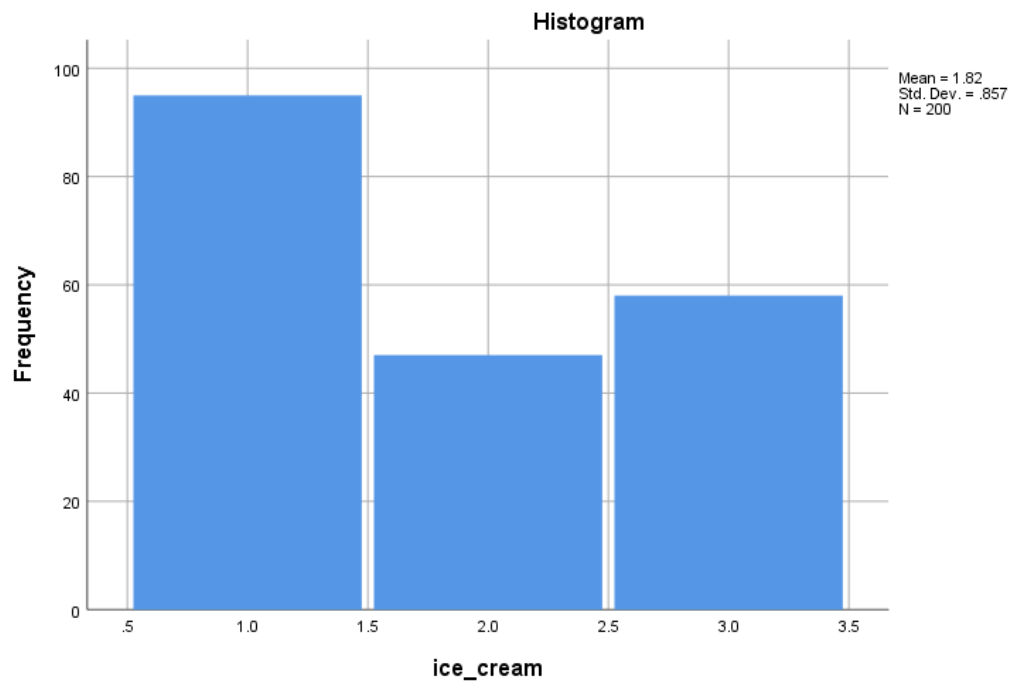
		Northern	Southern
Youth	%	63%	37%
Education	%	68%	32%
ExpenditureYear0	%	72%	28%
Labour Force	%	68%	32%
Males	%	66%	34%
More Males	%	66%	34%
State Size	%	68%	32%
Youth Unemployment	%	67%	33%
Mature Unemployment	%	65%	35%
High Youth Unemployment	%	93%	7%
Wage	%	71%	29%
Below Wage	%	59%	41%

Interpretation: In the table which is showcased above it can be depicted that attributes related to the crime are given in which they are being depicted in percentage form where they are distributed in two regions which are northern and southern and it can be inferred that percentage share of each attribute is more in the northern region and less in the southern region. Also, only one attribute is there whose share is low than ten per cent which is high youth unemployment in the southern region where the share is only seven per cent and when compared with all attributes this attribute is having a greater share in the northern region which is ninety-three per cent (Wiyono, Christian and Abidin, 2021).

Part 2

Vanilla Flavour liking

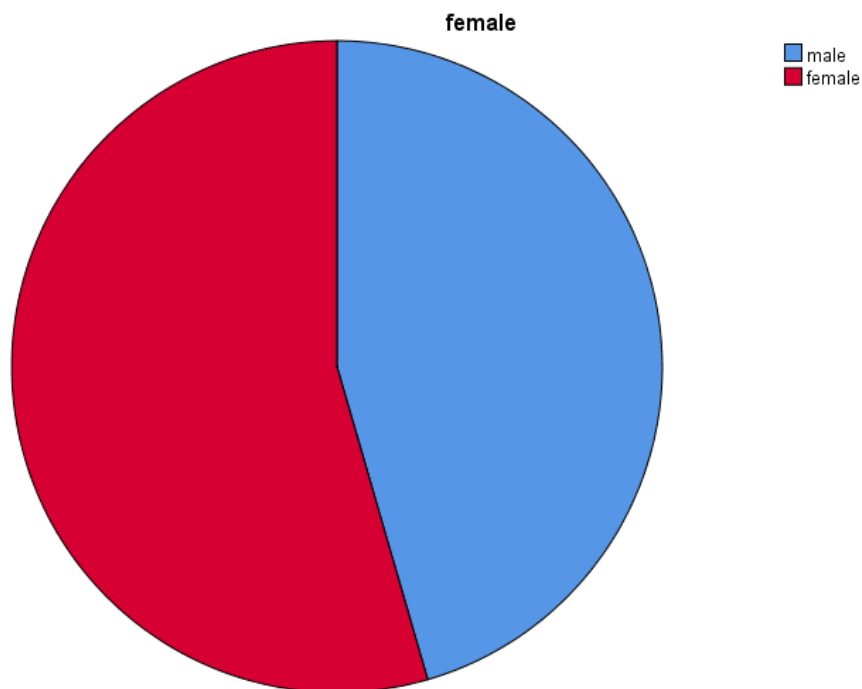
ice_cream					
		Frequenc y	Percent	Valid Percent	Cumulative Percent
Valid	vanilla	95	47.5	47.5	47.5
	chocolate	47	23.5	23.5	71.0
	strawberr y	58	29.0	29.0	100.0
	Total	200	100.0	100.0	



Interpretation: The liking of the flavours by the participants are different in the ice cream as they had the choice of three flavours and from it, they had to choose one flavour which was liked by them which the vanilla received 95 respondents' choice as they were liking this flavours and rest were liking two other flavours which were chocolate and strawberry (Gupta and Shringare, 2018).

Male and Female students

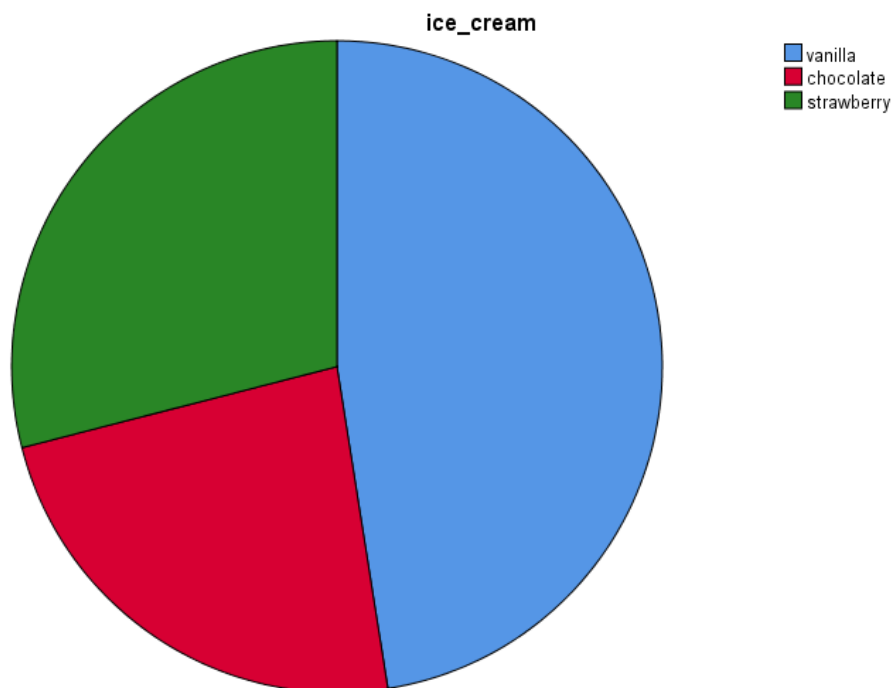
female					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	male	91	45.5	45.5	45.5
	female	109	54.5	54.5	100.0
	Total	200	100.0	100.0	



Interpretation: In the survey was conducted was gender-neutral as both the genders were allowed to participate with which the females were more as they were 109 and males were forming only 45.5 per cent as they were 91 only from the 200 participants (Aziz and Aftab, 2021).

Mean and median

Statistics		
ice_cream		
N	Valid	200
	Missin g	0
Mean		1.82
Median		2.00



Interpretation: In the survey, the mean and median have come out to be respectively 1.82 and 2.00 of the participants who were liking these flavours which are chocolate and strawberry and the mean and median values are good as the difference between them is low so it signifies that data of both these flavours are centred equally (Lemenkova, 2019).

Clustering

Initial Cluster Centers		
	Cluster	
	1	2
female	0	1
puzzle	26	71

Iteration History		
Iteration	Change in Cluster Centers	
	1	2
1	14.067	12.242
2	.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 2. The minimum distance between initial centres is 45.011.		

Final Cluster Centers		
	Cluster	
	1	2
female	0	1
puzzle	40	59

Number of Cases in each Cluster		
Cluster	1	68.000

	2	132.000
Valid		200.000
Missing		.000

Interpretation: In the tables which are shown above the cluster analysis has been performed with the help of k means clustering in which the clusters which are taken are two here which are the female and the puzzle. In cluster one of the initial cluster tables, the value in females is zero and, in the puzzle, there are 26 values in cluster two the female is having only one value and the puzzle is having 71 values in it. In the table of iteration history, clusters one and two are having an iteration of 0.00 in the second and one the iteration in cluster one is 14.067 and in cluster two it is 12.242. The changes which have taken place in clusters one and two in the final cluster centre are that females are having zero in clusters one and one in cluster two whereas the puzzle is having the value of 40 in cluster one and a value of 59 in cluster two. The final number of cases which have come out in each cluster are 68 in cluster one and 132 in cluster two (Er Kara and Oktay Firat, 2018).

Data mining and text mining methods

Data Mining

It is a method of extracting helpful data or information from an incredible different scale quantity of accumulation (or massive data). The gap between cognition and intake has been reduced by victimization varied data processing instruments. It may be mentioned as information discovery from large data sets for the processing of further functionality (Alton, 2017).

Methods of information mining

Classification

This data processing methodology is engaged to differentiate the things within the knowledge sets into categories or teams. It helps to predict the behaviour of entities among the cluster accurately. it's a ballroom dance process:

- Learning step (training phase): during this, a classification algorithmic rule builds the classifier by analyzing a coaching set.
- Classification step: take a look at knowledge accustomed estimate and the accuracy or preciseness of the classification rules.

For example, a financial institution uses to spot loan candidates at low, medium or high credit risks. Similarly, a medical investigator analyzes cancer knowledge to predict that drugs to visit to the patient.

One commonly seen use case of the classification segment in data mining is done by companies like IBM and Amazon for exploring increasingly large data sets in order to improve the segmentation of their respective markets.

Clustering

Clustering is nearly like classification, however during this cluster are created reckoning on the similarities of information things. totally different teams have dissimilar or unrelated objects. it's additionally referred to as knowledge segmentation because it partitions immense knowledge sets into teams in line with the similarities.

Taking a real-world example of its application, A general store in Illinois conducted a sale of men apparels and clothing, all of its stores showed an incredible increase in their sales except for one that is using radio marketing instead of internet or television marketing.

Text mining

The process of text mining involves varied activities that assist in account data from unstructured text knowledge. Text mining techniques will be explained because the processes that conduct mining of text and find out insights from the info. These techniques deploy varied text mining tools and applications for his or her execution (Guides, 2022).

Even before applying many text mining techniques, one ought to perform text preprocessing. it's the method of cleanup and decoding knowledge into its implementing format. Being a core side of natural language processing, text preprocessing includes the employment of the many techniques like language identification, tokenization, part-of-speech tagging, and plenty of a lot of.

Categorization

Under the categorization methodology, one or a lot of classes of freelance (free format) text documents ar assigned . reckoning on the input-output examples to discriminate new documents , categorization is taken into account as supervised learning methodology. supported the texts content, predefined categories ar assigned to every text documents,

The process of text categorization involves strategies like pre-processing, indexing, spatiality reduction, and classification with the target to coach classifiers on the premise of recognized examples then unrecognized examples would be classified mechanically. Also, text categorization faces the issue of high speciality of feature house. Taking an example of its use case, the categorization of text inside the content of website helps google to crawl upon different websites efficiently for the enhancement of its SEO.

Visualization

Visualization strategies will improve and clarify the analysis of relevant data. so as to stipulate individual documents or clusters of documents, text flags ar practiced to indicate the class of documents and colours are accustomed show document density.

In this methodology, massive matter sources in a very visual hierarchy so a user may move with the documents via diving and scaling. as an example, the govt. uses data image to discover the terrorist networks and to spot crime-information (Li, 2019). Talking about the application

of visualization segment of text mining, company like GE practiced the application of visualization in 2013 which helped them massively in their marketing and messaging efforts.

Comparison of SPSS and Excel

Comparison Criteria	SPSS	EXCEL
Definition	IBM's advanced tool which is formulated for evaluation of large data sets statistically.	Software developed by Microsoft, typically used for manipulation and storage of data.
Use cases	For performing complicated calculations based on stats as per the guidelines of IBM.	Efficient and simple storage of data with facility of performing non-complex formulated calculations.
Real Time applications	Used in the functioning of heavy devices such as supercomputers (Statanalytica, 2019).	Used in maintenance and management of customer/ user data.
Benefits	Exceptional performance and efficiency	Can be used in deduction of useless data.
Applications	Most commonly used in IT based large scale companies.	Mostly used in management of sensitive secure data in both large scale and small scale companies.
Advantages	● Can be used for complex calculations. ● Helps in memory management for companies.	● Can be used to perform simple calculations. ● Helps in filtering out the useful data from a bunch.
Disadvantages	● Can be used only by heavy machines. ● Complex functioning resulting in less usage.	● Cannot be used for complex calculations. ● Memory system is too small.

References

- Alton, L. 2017. *The 7 Most Important Data Mining Techniques*. Datascience, [online], Available at: <https://www.datasciencecentral.com/the-7-most-important-data-mining-techniques/>. [Accessed on: 06-05-2022]
- Aziz, N. and Aftab, S., 2021. Data Mining Framework for Nutrition Ranking: Methodology: SPSS Modeller. *International Journal of Technology, Innovation and Management (IJTIM)*, 1(1), pp.85-95.
<https://journals.gaftim.com/index.php/ijtim/article/view/16>
- Cosma, E., 2019. Pivot Table: A Tool to Calculate, Summarize, and Analyze Data. *Ovidius University Annals, Economic Sciences Series*, 19(1), pp.161-166.
<https://stec.univ-ovidius.ro/html/anale/RO/wp-content/uploads/2019/08/5-2.pdf>
- Er Kara, M. and Oktay Firat, S.Ü., 2018. Supplier risk assessment based on best-worst method and K-means clustering: a case study. *Sustainability*, 10(4), p.1066.
<https://www.mdpi.com/278900>
- Guides, 2022. *Text mining & text analysis*. Guides, [online], Available at: <https://guides.library.uq.edu.au/research-techniques/text-mining-analysis/types-of-text-analysis>. [Accessed on: 06-05-2022]
- Gupta, P. and Shringare, K., 2018. Analyzing the mobile phone buying behavior using SPSS. *AADYA-National Journal of Management and Technology*, 8, pp.30-37.
<http://ischolar.info/index.php/AADYA/article/view/183159>
- Kandpal, N., 2021. Web Log Mining of Server Log Data of ‘Counselling Website’ Using Microsoft Excel and Implementation of Preprocessing Algorithm. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(12), pp.1402-1415.
<https://turcomat.org/index.php/turkbilmat/article/view/7622>
- Lemenkova, P., 2019. Numerical data modelling and classification in marine geology by the SPSS statistics. *International Journal of Engineering Technologies IJET*, 5(2), pp.90-99. <https://dergipark.org.tr/en/pub/ijet/issue/45163/585830>
- Li, S. 2019. *A Complete Exploratory Data Analysis and Visualization for Text Data*. Towardsdata, [online], Available at: <https://towardsdatascience.com/a-complete-exploratory-data-analysis-and-visualization-for-text-data-29fb1b96fb6a>. [Accessed on: 06-05-2022]
- Maknunah, J., Mumpuni, I.D., Rahmawati, L.S. and Widayanti, R., 2021. Microsoft Excel Training for Data Processing for Vocational High School Students. *Jurnal Dedikasi*, 18(2), pp.20-26.
<https://ejournal.umm.ac.id/index.php/dedikasi/article/download/18305/10106>

- Mohammad, M., Nuar, A. and Yahya, M.S., 2021. A Microsoft Excel Spreadsheet for Selecting Organisational Improvement Initiatives based on Malaysia Business Excellence Framework and Industrial Revolution 4.0. *Research Progress in Mechanical and Manufacturing Engineering*, 2(1), pp.145-151. <https://publisher.uthm.edu.my/periodicals/index.php/rpmme/article/view/1880>
- Narayan, K.A. and Nayak, M.S.D.P., 2021. Need for interactive data visualization in public health practice: Examples from India. *International Journal of Preventive Medicine*, 12, pp.3-9. <https://www.ncbi.nlm.nih.gov/pmc/articles/pmc8106281/>
- StatAnalytica, 2019. *SPSS vs Excel: Which One is The Best Tool For Statistics*. Statanalytica, [online], Available at: <https://statanalytica.com/blog/spss-vs-excel/>. [Accessed on: 06-05-2022]
- Suwannahong, K., Wongcharee, S., Kreetachart, T., Sirilamduan, C., Rioyo, J. and Wongphat, A., 2021. Evaluation of the Microsoft Excel Solver spreadsheet-based program for nonlinear expressions of adsorption isotherm models onto magnetic nanosorbent. *Applied Sciences*, 11(16), p.7432. <https://www.mdpi.com/1226832>
- Wiyono, W., Christian, P.Y. and Abidin, A., 2021. SIMPLE LINIER CORRELATION AND REGRESSION ANALYSIS USING MICROSOFT EXCEL AND ITS APPLICATION IN MARKETING MANAGEMENT. *Akselerator: Jurnal Sains Terapan dan Teknologi*, 2(1), pp.29-40. <https://jurnal.ubd.ac.id/index.php/aksel/article/view/602>