



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

LCBB5000

Assessment 2

Data Handling and Business Intelligence

Student name	Renata Szabo
Student ID no.	2009640
Lecturer	Joy Botu
Programme	BA Business Management

Table of Contents

Introduction.....	2
Part 1	3
Basic Functions	3
IFERROR.....	3
IF	4
Minimum and Maximum	4
Pivot Table.....	5
Part- 2.....	6
K-Means Cluster analysis	6
Data Mining and Text Mining	10
Advantages and disadvantages of using SPSS and Excel.....	10
Conclusion	11
Reference	12

List of Tables

Table 1: Descriptive Statistics	7
Table 2: Initial Cluster Centers	7
Table 3:Iteration History	8
Table 4: Final Cluster Centers	8
Table 5: Sistance between final cluster data	9
Table 6: Number of Cases in each Cluster.....	10

Introduction

Data handling is a method of confirming that data is achieved, stored and disposed of in a secure and safe mode to a research project conclusion. Business intelligence (BI) is a process of technical infrastructure that is stored, analyzed and collected by the companies produced activities data. As per Jayakrishnan et al. (2018), BI is a term for data process analysis, data mining and performance benchmarking. The issue of developing these techniques in the business is for a large amount of data to be accurately processed, analysed and visualised. The main aim of this report is to identify the critically evaluated data warehouses, data mining and BI in business data with the help of using predictive analytics software. The report will be here divided into two parts where one part helps to analyze, visualize and process the data and the other part will help to understand the text mining and analysis techniques of business. Therefore the completion of this report will be helpful to demonstrate the ability to identify the business data handling techniques.

Part 1

Ms excel helps the user for understanding this software and perform the basic activities. It helps the user for organizing, quantitative data visualisation, and calculation of data; even if they are a manager or senior manager, they can evaluate the information to make important decisions. As per Zhong and Jiang (2019), The evaluation formula helps to identify and resolve the problems. With the help of evaluation, the user can use formulas and in the nested function and calculate the query with a successful result. The dataset is in excel is prepossessed to check the noisy data, missing values and other different functions for executing the algorithm. The crime rate was determined with the help of $=27/6.27$ formula, where 27 is deterring the crime number and 6.27 crore explaining to the per million population. The data visualization helped to present data in a graphical format. It helps in the data decision-making for analytics presented visually so they can resolve the complex concepts and identify the new patterns of resolutions. As per Dai et al. (2019), Pre-processing helps to handle the structured data and where functions and techniques make it easier to clean the structured data. A few of the functions are described here, along with the examples:

Basic Functions

IF ERROR

The function IFERROR helps in error handling for the errors that are #NAME?, #NUM, #VALUE, etc. it is used as the 1 and 0, for example

=IFERROR (1/0,-1000)

IF

According to Fu and Fu (2019), the IF function helps to check the condition and return the specified value such as “true” and “false”. According to this condition, the value returns the corresponding content, and it helps to use the conditional evaluations of formula and values of applications. It helps to judge effectively and monitor the data based on project standard limit of surface water environmental quality standards. The IF function is editable, and the results were based on the evaluation of monitored data and its surface water monitoring data. It is used to simplify a large amount of data and evaluate the process of reducing errors and provide standard results. For example, in the table of crime_rate using the IF function in figure 1 with the help of =IF (A2>80, "True", "False") then here, if the value is greater than 80, then print the truth and if the value is not greater than 80 then print false.

Condition 1 IF Function
False
False
False
False
False
False
False
False
False
False
False
True
True
True
True
True

Figure 1: Use of IF function

Minimum and Maximum

The Max function helps to determine the largest numeric number of the cell range and whereas the just opposite Min function is refereeing to the smallest number of the cell range. For example, in the table of crime rates, the min and max range is evaluated with the help of clicking on the cell and then the AutoSum button of the listed arrow, then selecting the Min

and Max. After that, if it is necessary, then select the particular cell and click on the enter key. After that, the Min and Max formula displays the Largest and the smallest number from the specified range. For example, each pivot table of Crime_data and figure 2 uses the Min and Max function for determining the Minimum and Maximum value of the cell.

Row Labels	Count of CrimeRate		
45.5	1		
52.3	1		
56.6	1		
60.3	1	Minimum	45.5
64.2	1	Maximum	161.8
67.6	1		
70.5	1		
73.2	1		
75	1		
78.1	1		
79.8	1		
82.3	1		
83.1	1		
84.9	1		

Figure 2: Minimum and Maximum Value of the cell

Pivot Table

The Pivot table is used to reorganize, count, sort, total, summarise and evaluate the average of store data from a table. The table was used to transform the rows into colours and columns into rows. It is used to group the data with the help of columns and use advanced calculations on them (support. microsoft, 2020). The table here allows extracting data easily from the large amount of data set. For example, the pivot table is executed on the crime_data table and figure 3 with the help of some steps that are

- Select the cell of the data set and then click on the Insert tab, tables group and click on the pivot table.
- A dialogue box will appear for the default location that is created on a new pivot table in a new worksheet, then click on OK.
- Then drag the fields for exporting each product and then sort and filter it according to the needs.
- For changing the summary of calculation right, click on the field value and click on ok and change it.
- Then building the pivot table

Row Labels	Count of Southern
0	31
1	16
Grand Total	47

Figure 3: Pivot Table

Part- 2

According to the given scenario of part 2, 95 students like the vanilla flavour.

- Male: 91
- Female: 109
- According to the findings, the pie chart in figure 4 represents the count of males (0) and females (1).

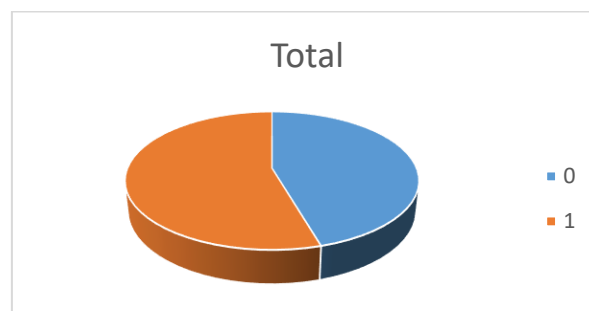


Figure 4: Count of male and female

As per Holladay et al. (2021), the Mean is defined with the mean formula =AVERAGE(A3:A4) in the table of Ice_cream XSL, and the mean value of students who like chocolate and strawberry flavour that is 2.5 and the median value is determined with the help of =MEDIAN(A: A) so, the median value is 2.

K-Means Cluster analysis

The cluster analysis is an exploratory analysis that helps to identify the structure of data. The cluster analysis helps to find the analysis of group cases, and SPSS where provides three methods of analysis that is K-Means Cluster, Two-step cluster and Hierarchical cluster. As per Sane et al. (2020), the SPSS cluster can handle the nominal, scale, ordinal and binary interval-related data. Clustering is a process of creating a group of abstract values into the classes of the same object. The cluster was treated as one group when doing the cluster analysis and dividing the data into groups based on the same data and assigning to the group data labels. The K-Means is the unsupervised method of combining the unlabelled data set into the different cluster sets where K helps to define the number of clusters that are needed to be created, and for example, if the K=3 then the 3 clusters are needed.

As per the relevant practical example here in table 1, descriptive statistics are defined for explaining the coefficients that are summarized in the given data set. Here it is represented as per the sample of the population, and 200 peoples gender is described that the minimum value is 0 in 2nd column, the maximum value here represented as 1 in 3rd column, the mean value is described herein 4th column .54 and standard deviation described as .499 in 5th column. As per the same sequence here, all answers are provided in front of the question.

Table 1: Descriptive Statistics

Descriptive Statistics					
	(1) N	(2) Minimum	(3) Maximum	(4) Mean	(5) Std. Deviation
What is your gender?	200	0	1	.54	.499
What is Your favourite flavour of ice cream?	200	1	3	1.82	.857
Video data	200	26	74	51.85	9.901
Puzzle data	200	26	71	52.40	10.736
Valid N (listwise)	200				

The initial cluster centre represents the variable value of the k well-spaced observation. It helps to represent the arithmetic mean value of all data related to the cluster. Table 2 here represents the value of the initial cluster centre that is represented in the 1st, 2nd and 3rd columns. Such as here represented the gender, favourite flavour of ice cream, video data and puzzle data.

Table 2: Initial Cluster Centers

	Cluster		
	1	2	3
Zscore: What is your gender?	-1.09170	-1.09170	.91142
Zscore: What is Your favourite flavour of ice cream?	1.38273	-.95099	1.38273
Zscore: Video data	-2.61088	2.03517	1.73217
Zscore: Puzzle data	-.96919	-1.99380	1.73206

Iteration history help to perform the programme of clustering process on each step which includes the log likelihood and restricted log-likelihood. It is the early iteration where cluster centres are shifted quite a lot. The iteration was repeated on each step again and again. In table 3 here, grouped iteration data that is not explicitly labelled in data is described in the 1st, 2nd and 3rd columns.

Table 3: Iteration History

Iteration	Change in Cluster Centers		
	1	2	3
1	2.267	2.290	1.906
2	.241	.225	.114
3	.194	.369	.145
4	.125	.121	.076
5	.111	.089	.088
6	.081	.054	.091
7	.024	.046	.044
8	.023	.000	.027
9	.000	.000	.000

In table 4, final cluster centres are described with the mean of each variable in each cluster. All of the values here described by the total number are divided by the member of the cluster. Here gender is evaluated in 1st, 2nd and 3rd columns where data is centroid as (.42874, -1.09170, .20610). All questions' final cluster data is represented in the same sequences; also, the result is described in Figure 5.

Table 4: Final Cluster Centers

	Cluster		
	1	2	3
Zscore: What is your gender?	.42874	-1.09170	.20610
Zscore: What is Your favourite flavour of ice cream?	-.33242	-.67196	.82395
Zscore: Video data	-.79164	.58603	.54576
Zscore: Puzzle data	-.66955	-.00128	.78354

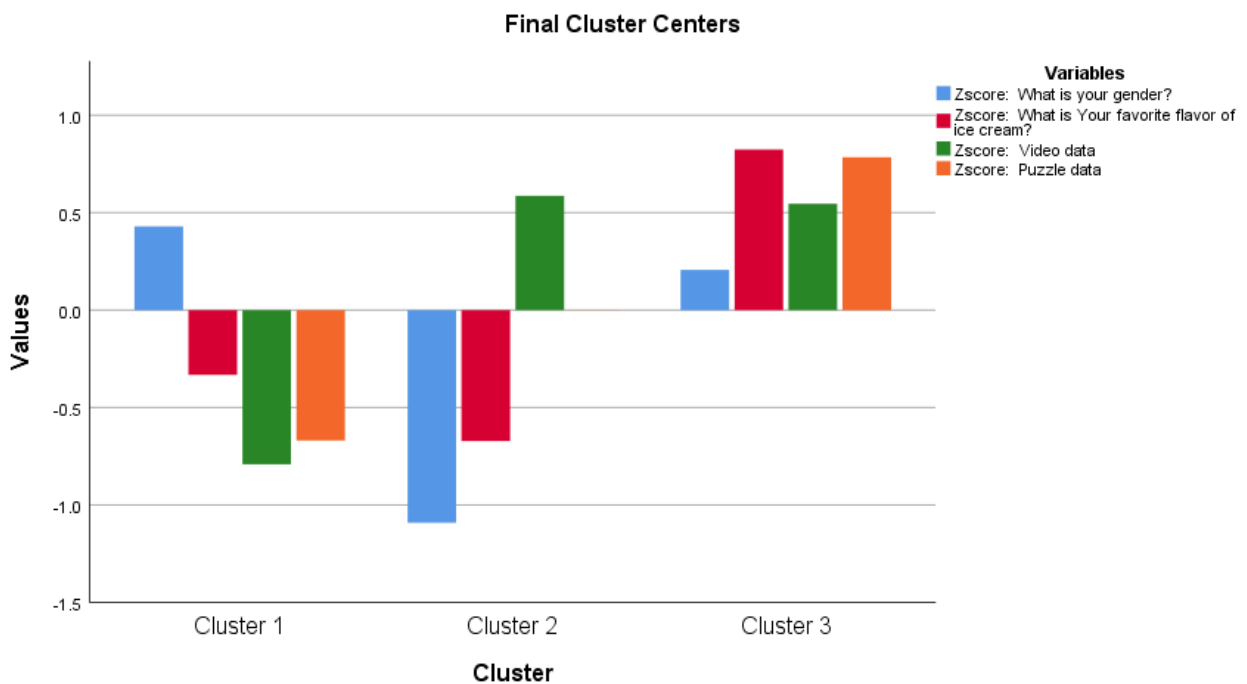


Figure 5: Final cluster centres

The distance between the final cluster centre is described in the 5th table with the help of average linkage clustering. Here the difference between two clusters is defined as the average of pairs of objects, and it is described in the 1st, 2nd and 3rd columns, where the 1st cluster's data is represented in the 2nd column as 2.184 and in the 3rd column 2.299. All three clusters' data here are described in the same sequence. The ANOVA function is also performed here where it is used for examining the differences of mean values in the associated dependent variable that is affecting and controlling the independent variable. Here cluster means square, its degree of freedom and error occurrence value, and its mean square, and its degree of freedom value is described. Then in the last column here, F-test results and significance probability are described. Such as, in the 1st question here cluster mean square value is 36.548, its degree of freedom value is 2, the error mean square value is .639, its degree of freedom value is 197, and the f-test value is 57.186, and the significance probability value is described as .000. All questions results are described in this table in the same sequence.

Table 5: Distance between final cluster data

Distances between Final Cluster Centers			
Cluster	1	2	3
1		2.184	2.299
2	2.184		2.131
3	2.299	2.131	

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Zscore: What is your gender?	36.548	2	.639	197	57.186	.000
Zscore: What is Your favourite flavour of ice cream?	39.072	2	.613	197	63.688	.000
Zscore: Video data	44.481	2	.559	197	79.632	.000
Zscore: Puzzle data	40.399	2	.600	197	67.330	.000

Table 6: Number of Cases in each Cluster

Number of Cases in each Cluster		
Cluster	1	83.000
	2	46.000
	3	71.000
Valid		200.000
Missing		.000

In the table 6th table, each cluster value is described as per the number of cases. Here 3 cluster value is described as 83.000, 46.000 and 71.000. Then the valid value is described where a 200.000 value occurs, and the missing value is .000.

Data Mining and Text Mining

Data Processing is the common data mining and text mining method that is used in the business for data conversation into the suitability of use and wished format. The conversion process was carried out manually and automatically (Saura and Bennett, 2019). The text mining method is also matched to the classical processing of data. The retrieval of data is intended to execute the desirable text that is matching to the data collection. The information extraction was used to fetch the extracted predefined information, which represents the pre-processing of the collected data. For example, data mining is used for the criminal behaviour model to identify and detect it. Data mining is a method of discovering the hidden patterns of a large number of data sets and relationships. It helps to allow reliable future events and predictions based on the categorization and identification of historical data.

Advantages and disadvantages of using SPSS and Excel

Advantage	Disadvantage
The SPSS is a comprehensive statistical software.	The SPSS software is expensive for the students to purchase.

• It is easy software to learn. • The SPSS helps in the interpretation of outcomes relatively easily (Jahromi et al. 2018). • The SPSS software is expendable. • The complex tests are available on future basis outcomes. • Rapidly and quickly displaying to the data tables.	• The feature of the graph is not simple as the MS Excel (Jahromi et al. 2018). • It involves the training to complete all executable features. • It has limited functionality, and it is related to Excel.
---	---

Conclusion

The data handling and business intelligence help to present the data in the graphical format to allow decision-making for the visual data analytics to handle all infrastructure and techniques of business. The techniques helped to store, collect and produce the company's data. The report here was developed to analyze the method of evaluating a large amount of data in business easily. The aim of this work was to demonstrate the ability to identify the current business trends with the help of data mining, BI and analytics software. In the first part of the report, the crime data was determined with the help of excel pre-processing data visualization, analysis and functions. Then in the second part, data and text mining operations with the analysis were described. Therefore, business intelligence and data handling understanding help to enhance the decision-making techniques of the business.

Reference

- Dai, X., Ji, C. and Wang, H. (2019) 'Application of EXCEL commonly used in navigation data processing', In *IOP Conference Series: Materials Science and Engineering*, 569(5).
<https://iopscience.iop.org/article/10.1088/1757-899X/569/5/052094/meta>
- Fu, H. and Fu, F.E. (2019) 'A Fast Way to Evaluate Environmental Monitoring Results by Excel', In *IOP Conference Series: Earth and Environmental Science*, 252(4).
<https://iopscience.iop.org/article/10.1088/1755-1315/252/4/042030/meta>
- Holladay, J.T., Wilcox, R.R., Koch, D.D. and Wang, L. (2021) 'Review and recommendations for univariate statistical analysis of spherical equivalent prediction error for IOL power calculations', *Journal of Cataract & Refractive Surgery*, 47(1), pp.65-77.
https://journals.lww.com/jcrs/Fulltext/2021/01000/Review_and_recommendations_for_univariate.10.aspx?context=FeaturedArticles
- Jahromi, A.S., Rahmanian, K. and Madani, A. (2018) 'Relation of knowledge about cesarean disadvantages and delivery mode selection in women with first pregnancy; south of Iran', *Journal of Research in Medical and Dental Science*, 6(2), pp.550-6.
https://www.researchgate.net/profile/Abdolreza-Sotoodeh-Jahromi/publication/324835500_Cesarean_Disadvantages_and_Delivery_Mode_Selection_in_Women_with_First_Pregnancy_South_of_Iran/links/5ae6f256aca272ba50821cc9/Cesarean-Disadvantages-and-Delivery-Mode-Selection-in-Women-with-First-Pregnancy-South-of-Iran.pdf
- Jayakrishnan, M., Mohamad, A.K. and Yusof, M.M. (2018) 'Assimilation of Business Intelligence (BI) and Big Data Analytics (BDA) Towards Establishing Organizational Strategic Performance Management Diagnostics Framework: A Case Study', *Journal of Digital Information Management*, 16(1).
https://www.researchgate.net/profile/Mailasan-Jayakrishnan/publication/322821299_Assimilation_of_Business_Intelligence_BI_and_Big_Data_Analytics_BDA_Towards_Establishing_Organizational_Strategic_Performance_Management_Diagnostics_Framework_A_Case_Study/links/5a719d65458515015e64a8be/Assimilation-of-Business-Intelligence-BI-and-Big-Data-Analytics-BDA-To-wards-Establishing-Organizational-Strategic-Performance-Management-Diagnostics-Framework-A-Case-Study.pdf

Sane, S., Bujack, R., Garth, C. and Childs, H. (2020) 'A survey of seed placement and streamline selection techniques', In *Computer Graphics Forum*, 39(3), pp. 785-809. <https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14036>

Saura, J.R. and Bennett, D.R. (2019) 'A three-stage method for data text mining: Using UGC in business intelligence analysis', *Symmetry*, 11(4). <https://www.mdpi.com/443224>

Support.microsoft (2020) *Create a PivotTable to analyze worksheet data*. support.microsoft, [online], Available at: <https://support.microsoft.com/en-us/office/create-a-pivottable-to-analyze-worksheet-data-a9a84538-bfe9-40a9-a8e9-f99134456576>. [Accessed on: 21-March-2022]

Zhong, Y. and Jiang, H. (2019) 'Application of Excel in Financial Management Teaching', In *3rd International Conference on Economics and Management, Education, Humanities and Social Sciences (EMEHSS 2019)* (pp. 186-189). Atlantis Press. <https://www.atlantispress.com/article/125910266.pdf>