

University of Wales Trinity Saint David



Module name: Data Handling & Business Intelligence

Module code: LCBB 5000

Assignment Title: Assessment 2-Data Analysis & Visualisation

Student name: Romyana Georgieva Yordanova

Student ID: 2028611

Submission date: 10/01/2023

Lecturer name: Gasso Wilson

Contents

INTRODUCTION 3

Part 1 3

Evaluation of use of excel for pre processing, analysing and visualising the data 3

Demonstration of excel functions 4

Part 2 6

2.1 Liking of vanilla flavour ice cream among students 6

2.2 Population of male and female among students 6

2.3 Mean and median of participants who like chocolate and strawberry flavour ice cream 6

2.4 Clustering of data through SPSS 8

2.5 Explanation of most common data mining and text mining methods to be used 15

2.6 Discussion of advantages and disadvantages of SPSS and Excel 16

CONCLUSION 18

REFERENCES 19

INTRODUCTION

Data handling refers to the process that make sure that the data on which research is done is stored well, which can be used to evaluate the performance of the business over a period of time (Atitallah and et.al. 2020). It involves formulation of policies and procedures to manage data electronically and non electronically. Business intelligence is also a part of data handling, it refers to the process of collecting, storing and analysing data provided by an organisation that help the management to make decisions on the basis of the data provided. This report covers the excel function such as If, Lookup, Pivot table using the data provided of superstores and evaluation of usage of excel for pre-processing, analysis and visualisation of data. In the part 2 of this report, there is shown data through pie, bar or histogram and also there is to be shown explanation of data mining and methods thereof and advantages or disadvantages of using SPSS and Excel.

Part 1

Evaluation of use of excel for pre processing, analysing and visualising the data

Excel is a software which is used to analysis of data through presentation of data in a structured format. This software was developed for the complex arithmetical calculations, presentation of information through graphs, pivot table which makes the information presentable and easy to understand. It refers to the spreadsheet which includes usage of grid of cells which makes it sheet of rows and columns. The excel is used for presentation of statistical, financial and other data in a systematic way (Bansal and Kumar, 2020).

Excel is referred to as a tool for pre processing and handling the data shown in structured manner. In excel, there are many functions and techniques which makes it easier to filter structured data. There are many functions which are used for pre processing, which will be discussed later in this report. Pre-processing in excel can be done through collecting data and presenting them in a chronological and structured format. Core usage of excel is to analyse the data shown, it can be analysed through presentation of data as per the need of the user of excel. It enables the user to derive the data category wise which is useful for him. The data can be sorted through pivot table and many other functions. As far as the visualisation is concerned, the presentation of data is essential to understand which makes the data useful for the users. The information presented in tabular format, sometimes becomes difficult to understand which can be understood easily through charts and graphs which makes it useful for the laymen (Bekken, 2019).

Demonstration of excel functions using charts and graphs

In excel there are many functions which are to be performed to extract the information from the tables prepared in the spreadsheet, some of such functions are described below:

- *Lookup Function:*

Under the lookup function, information related to particular column or row can be extracted which has been entered in a table array. This function can be performed through the formula “=LOOKUP (lookup value, array)”. Lookup functions can be used for find the value in the table on a horizontal or vertical base (Choi and et.al. 2018).

To find the value given in the top row of a table or array, then it is advisable for the user to use the Hlookup formula when the comparison is to be made among the rows. The formula of horizontal lookup is “=HLOOKUP (lookup value, table array, row index number, [range lookup]).

On the other hand, to find the value among the column in a table array, then there arises a need to find out the values virtually which takes into account the formula of virtual lookup which is “=VLOOKUP (lookup value, range of lookup value, column number, [TRUE] or [FALSE])”. True refers to approximate match whereas false represents exact match of the information needed.

- *Pivot Table:*

It is an useful and interactive manner to summarize the large amount of data. Under pivot table, user becomes able to set up and change the data summary structured by bringing into and dropping the areas graphically and this rotation is known as pivot, which has given this table its name (Choudhury and Pal, 2019). A pivot table can be created through selection of cells for which, there is a need to create a pivot table and then from the ribbon select an option of the pivot table and then click ok, which will make a pivot table. Under pivot table, the first row is considered as heading of the column which consists of all the information in a column.

Year	Sum of Sales	Sum of Profit
2009	4209896.846	434096.02
2010	3560087.045	364917.33
2011	3429944.981	380310.5
2012	3715671.953	342444.13

- *IF function:*

This is the function which is popular in excel that enables the users to make logical comparison between the value and expectation. If function may have two results,

where the first result may be true or yes whereas the second one may be false or no. There is a formula is provided for the excel function which is “IF (cell number, expectation, “Yes”, ”No”). In this method formula is used of SUMIF to find out yearly sales and profits

Year	Total Sales	Total Profit
2009	4209896.8	434096
2010	3560087	364917.3
2011	3429945	380310.5
2012	3715672	342444.1

Part 2

2.1 Liking of vanilla flavour ice cream among students

To find out the value of a particular column, the user needs to use the formula of “COUNTIF” On the basis of calculation in the excel by using the formula “COUNTIF”, it is noted that 95 students like vanilla flavour of ice cream.

COUNT	FAV.FLAVOR
95	Vanilla
47	Chocolate
58	Strawberry

2.2 Population of male and female among students

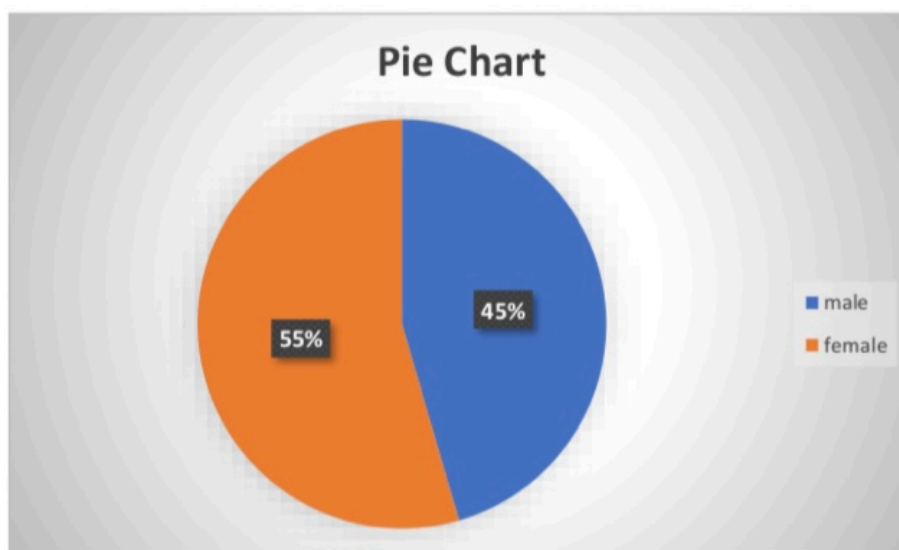
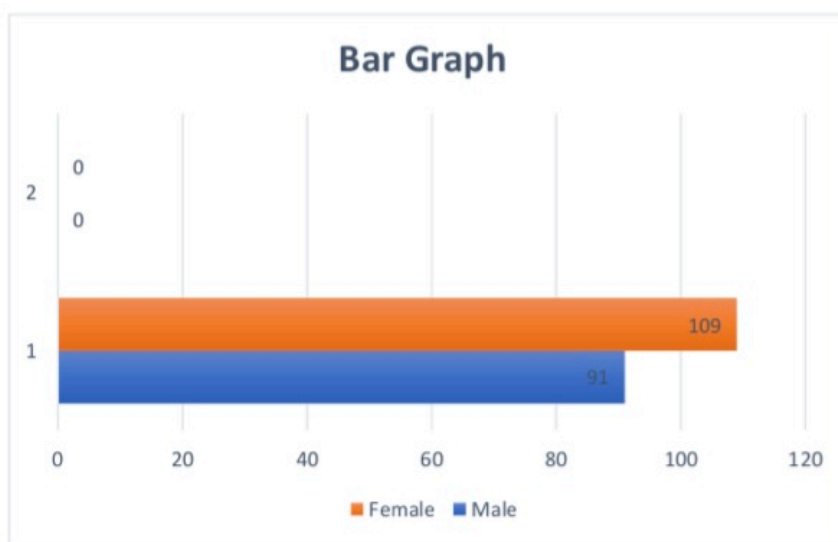
Using the formula of COUNTIF in this as well, there is shown bifurcation between numbers of male and female. It indicates that males are, 91 and females are 109.

91	Male
109	Female

2.3 Mean and median of participants who like chocolate and strawberry flavour ice cream

On basis of calculation of mean it is found out that mean of participants like chocolate and strawberry flavour ice cream are dl and ff respectively. The median of participants like both the flavours of ice cream are 63 and 101.5 respectively.

52.5	Mean
63	Median of Chocolate
101.5	Median of Strawberry



2.4 Clustering of data through SPSS

Table 1: -

Initial Cluster Centres			
	Cluster		
	1	2	3
Gender	0.00	1.00	1.00
Favourit	2.00	1.00	3.00

The first cluster is Vanilla, the second is Chocolate, and the third is Strawberry, and their values are based on two factors, namely gender and favourite flavour, as seen in the initial cluster table above.

Table 2: -

Iteration History			
Iteratio	Change in Cluster		
	1	2	3
1	0.298	0.000	0.000
2	0.363	0.000	0.500
3	0.000	0.000	0.000

Because of little or no change in the cluster centres, convergence was achieved. Any centre can move its absolute coordinates up to a maximum of 0.000. 3.0 is the most recent iteration. The first centres must be separated by a minimum of 1.414.

The number of iterations and changes in the cluster centres have been determined from the above table. A change of.298 was seen in the first cluster, a change from.363 to.500 in the second cluster, and no change was observed in the third cluster (Currie and et.al. 2019).

Table 3:-

Cluster Membership		
Case Num	Cluster	Distance
1	1	0.605
2	2	0.000
3	3	0.500
4	3	0.500
5	1	0.605
6	1	0.605
7	1	0.605
8	1	0.605
9	1	0.605
10	1	0.605
11	1	0.605
12	1	0.605
13	3	0.500
14	3	0.500
15	1	0.605
16	1	0.605
17	3	0.500
18	1	0.605
19	3	0.500
20	1	0.605
21	1	0.605
22	1	0.605
23	1	0.605
24	3	0.500
25	1	0.605
26	1	0.605
27	3	0.500

28	1	0.605
29	3	0.500
30	1	0.605
31	1	0.605
32	3	0.500
33	3	0.500
34	1	0.605
35	3	0.500
36	3	0.500
37	1	0.605
38	3	0.500
39	1	0.605
40	1	0.605
41	1	0.605
42	1	0.605
43	1	0.605
44	1	0.605
45	3	0.500
46	1	0.605
47	3	0.500
48	1	0.605
49	1	0.605
50	1	0.605
51	1	0.605
52	3	0.500
53	1	0.605
54	1	0.605
55	1	0.605
56	3	0.500
57	1	0.605
58	1	0.605
59	1	0.605
60	1	0.605
61	1	0.605
62	3	0.500
63	1	0.605

64	1	0.605
65	3	0.500
66	1	0.605
67	1	0.605
68	1	0.605
69	1	0.605
70	1	0.605
71	1	0.605
72	3	0.500
73	1	0.605
74	1	0.605
75	1	0.605
76	1	0.605
77	3	0.500
78	3	0.500
79	1	0.605
80	3	0.500
81	3	0.500
82	1	0.605
83	1	0.605
84	1	0.605
85	1	0.605
86	3	0.500
87	3	0.500
88	3	0.500
89	1	0.605
90	3	0.500
91	1	0.605
92	1	0.605
93	3	0.500
94	1	0.828
95	1	0.828
96	1	0.828
97	2	0.000
98	3	0.500
99	1	0.828

100	3	0.500
101	3	0.500
102	3	0.500
103	1	0.828
104	3	0.500
105	2	0.000
106	2	0.000
107	3	0.500
108	1	0.828
109	1	0.828
110	3	0.500
111	2	0.000
112	1	0.828
113	3	0.500
114	1	0.828
115	3	0.500
116	2	0.000
117	3	0.500
118	3	0.500
119	1	0.828
120	1	0.828
121	1	0.828
122	2	0.000
123	2	0.000
124	2	0.000
125	1	0.828
126	3	0.500
127	2	0.000
128	2	0.000
129	3	0.500
130	1	0.828
131	2	0.000
132	1	0.828
133	2	0.000
134	2	0.000
135	1	0.828

136	3	0.500
137	2	0.000
138	2	0.000
139	2	0.000
140	2	0.000
141	1	0.828
142	3	0.500
143	2	0.000
144	2	0.000
145	2	0.000
146	3	0.500
147	2	0.000
148	2	0.000
149	1	0.828
150	1	0.828
151	1	0.828
152	3	0.500
153	2	0.000
154	2	0.000
155	2	0.000
156	2	0.000
157	2	0.000
158	2	0.000
159	2	0.000
160	3	0.500
161	1	0.828
162	2	0.000
163	3	0.500
164	1	0.828
165	2	0.000
166	1	0.828
167	2	0.000
168	1	0.828
169	2	0.000
170	1	0.828
171	1	0.828

172	2	0.000
173	3	0.500
174	3	0.500
175	1	0.828
176	1	0.828
177	2	0.000
178	2	0.000
179	3	0.500
180	1	0.828
181	2	0.000
182	2	0.000
183	3	0.500
184	2	0.000
185	3	0.500
186	1	0.828
187	2	0.000
188	2	0.000
189	3	0.500
190	1	0.828
191	1	0.828
192	3	0.500
193	2	0.000
194	3	0.500
195	2	0.000
196	2	0.000
197	2	0.000
198	2	0.000
199	2	0.000
200	3	0.500

Table 4: -

Final Cluster Centres			
	Cluster		
	1	2	3
Gender	0.34	1.00	0.50
Favour	1.50	1.00	3.00

Table 6: -

Distances between Final Cluster Centres			
Cluster	1	2	3
1		0.828	1.508
2	0.828		2.062
3	1.508	2.062	

The above prepared table shows that when the results of tables 1 and 4 are compared, the third cluster remains the same. The analysis's conclusion is merely descriptive.

ANOVA						
	Cluster		Error		F	Sig
	Mean	df	Mean	df		
Gender	6.994	2	0.181	197	38.698	0.000
Favourite	61.328	2	0.119	197	514.107	0.000

Given that the clusters have been chosen to maximise the differences across instances in various clusters, the F tests should only be used for descriptive purposes. Since this is not taken into account when calculating the observed significance levels, they cannot be construed as tests of the equality of the cluster means (Dezi and et.al. 2018).

According to the aforementioned ANOVA table, preferred flavour has a greater impact on cluster formation than gender.

Number of Cases in each Cluster		
Cluster	1	94.000
	2	48.000
	3	58.000
Valid		200.000
Missing		0.000

As per the above presented data vanilla flavour has significant influence on the forming of cluster and Chocolate the least.

2.5 Explanation of most common data mining and text mining methods to be used

Data and text mining refers to the process of selection and analysis of large amount of dataset (Glistau and Coello Machado, 2018). This process of mining involves identifying, assessing and managing the datasets. Basically, mining process refers to the automation of the whole process of selecting and analysing enormous volumes of text or data to assess in searching, figuring out linkages, sifting through patterns, and drawing relevant conclusions from the data. Some of data mining and text mining techniques which can be utilized are described as follows:

- *IBM Watson:*

Among the features on this platform, which is meant to reveal value from data, are a tone analyser and personality insights. These are all based on interactions through internet platforms (Krishna and et.al. 2021). The first helps analyse the tones and emotions in consumer comments, whilst the second uses language analysis to predict personality characteristics, preferences, and habits.

- *Data mining through Oracle:*

This tool, which is a part of Oracle Advanced analytics, aids analysts in developing and applying models (Lee, 2019). It can also be used to create models that target and segment profiles, detect theft, and pinpoint the initial potency to target while assisting with determining customer behaviour. With its assistance, a variety of activities can be completed, including classification, anomaly detection, regression analysis, prediction, and many more.

- *Data mining by SAS:*

By supporting the user in mining, modifying, and managing diverse data sets from multiple sources, this programme helps in data management and analysis. It also provides users who lack the necessary technical skills with a customised graphical user interface. This tool supports speedy decision-making while also supporting data mining and optimization needs. By filling in null values, this is used to clean, remove errors from the data, and get rid of duplicate and pointless sets.

- *Thematic:*

Thematic Intelligence are used to aggregate phrases that appear to be related after highlighting the context of the input. The next step is to classify those phrases as either negative or positive. Numerous feedbacks show patterns and linkages that can be thoroughly analysed by developing subthemes, classifying customers, or combining metrics to discover their root causes and potential remedies. Thematic catalyst assists in creating dashboards with the information produced from the enormous amount of data. This aids in convincing business associates and management to act swiftly in response to client complaints (Lee, Azamfar and Singh, 2019).

2.6 Discussion of advantages and disadvantages of SPSS and Excel

Statistical Packages for Social Science is a tool that is used by the users for the research, analysis and performing tasks. It functions somewhat like a data mining tool, aiding in the extraction of data or information to carry out different mathematical operations like trigonometry, algebra, or arithmetic (Mosteanu, 2020). The software called SPSS allows users to run a variety of models, from the most basic to the most complex. Usage of this tool has advantages and disadvantages, which are described as below:

Advantages:

- This software saves time of performing tasks which would otherwise be more time consuming.
- Calculations made by SPSS tool is accurate.
- This software is used to examine large and complex data by using samples.
- The SPSS software concentrates on technical and conceptual aspects to analyse.
- SPSS is used to perform tasks efficiently and involves optimal utilization of resources.

Disadvantages:

- If there is involved biasness in collection of data, then it will fail to provide an accurate answer.
- It just focuses on providing the answers irrespective of the proper explanation of that answer.
- Being a time saving software, SPSS is quite expensive tool (Singh and et.al. 2020).
- Usage of this tool is not easy, hence it requires a lot of trainings and learnings to understand the features of this tool.
- Where dataset is huge then SPSS will fail to analyse the data.

As far as the advantages and disadvantages are concerned, then those are mentioned below:

Advantages:

- Excel is useful in making complex calculations in no time using the formula given in templates.
- Unlike SPSS software, excel does not involve hard training, hence it can be used with minimum direction.
- Users can use multiple sheets while using excel (Viloria and et.al. 2018).
- Excel provides templates which can be used to make budgets, resume etc.
- Excel can be proved effective software to make data presentable and easy to understand.

Disadvantages:

- At times, it becomes difficult to compile the data of various sheets into one.
- Excel is used by humans hence it is prone to errors or mistakes which can be committed by users.
- MS-excel is not appropriate for collaborating with working (Young and Rogers, 2019).

CONCLUSION

On the basis of above explanation, tables and calculations, it can be concluded that in the time of technology, using the software with proper due care, user can make complex data simple. This report highlighted as how data can be managed and sorted using MS-excel and SPSS software which will make it easy to understand for the users. Thus, software like excel and SPSS are very useful to perform complex tasks.

REFERENCES

Books and Journals

- Atitallah, S. B. and et.al. 2020. Leveraging Deep Learning and IoT big data analytics to support the smart cities development: Review and future directions. *Computer Science Review*. 38. p.100303.
- Bansal, S. and Kumar, D., 2020. IoT ecosystem: A survey on devices, gateways, operating systems, middleware and communication. *International Journal of Wireless Information Networks*. 27(3). pp.340-364.
- Bekken, G., 2019. The algorithmic governance of data driven-processing employment: Evidence-based management practices, artificial intelligence recruiting software, and

- automated hiring decisions. *Psychosociological Issues in Human Resource Management*. 7(2). pp.25-30.
- Choi, C. and et.al. 2018. Intelligent power equipment management based on distributed context-aware inference in smart cities. *IEEE Communications Magazine*. 56(7). pp.212-217.
- Choudhury, S. J. and Pal, N. R., 2019. Imputation of missing data with neural networks for classification. *Knowledge-Based Systems*. 182. p.104838.
- Currie, G. and et.al. 2019. Machine learning and deep learning in medical imaging: intelligent imaging. *Journal of medical imaging and radiation sciences*. 50(4). pp.477-487.
- Dezi, L. and et.al. 2018. The role of big data in shaping ambidextrous business process management: Case studies from the service industry. *Business Process Management Journal*.
- Glistau, E. and Coello Machado, N. I., 2018. Industry 4.0, logistics 4.0 and materials-Chances and solutions. In *Materials Science Forum* (Vol. 919, pp. 307-314). Trans Tech Publications Ltd.
- Krishna, M. and et.al. 2021, October. A Survey on Multimedia Analytics in Security Systems of Cyber Physical Systems and IoT. In *2021 2nd International Conference on Smart Electronics and Communication (ICOSEC)* (pp. 1-7). IEEE.
- Lee, I., 2019. The Internet of Things for enterprises: An ecosystem, architecture, and IoT service business model. *Internet of Things*. 7. p.100078.
- Lee, J., Azamfar, M. and Singh, J., 2019. A blockchain enabled Cyber-Physical System architecture for Industry 4.0 manufacturing systems. *Manufacturing letters*. 20. pp.34-39.
- Mosteanu, N. R., 2020. Artificial Intelligence and Cyber Security—A Shield against Cyberattack as a Risk Business Management Tool—Case of European Countries. *Quality-Access to Success*. 21(175).

- Singh, S. and et.al. 2020. Convergence of blockchain and artificial intelligence in IoT network for the sustainable smart city. *Sustainable Cities and Society*. 63. p.102364.
- Viloria, A. and et.al. 2018, June. Methodology for the design of a student pattern recognition tool to facilitate the teaching-learning process through knowledge data discovery (big data). In *International conference on data mining and big data* (pp. 670-679). Springer, Cham.
- Young, A. and Rogers, P., 2019. A review of digital transformation in mining. *Mining, Metallurgy & Exploration*. 36(4). pp.683-699.