



Coursework (Written Report)

Programme: BA in Business Management

Module Name: Data Handling and Business Intelligence

Module Code: LCBB5000

Lecturer: Joy Botu

Assessment 2

Student Name: Hristo Haralampiev Hristov

Student ID: 2010799

Contents

Introduction	4
Analysis and Discussion.....	5
2.1 Current trends in data warehousing, business intelligence, and data mining	5
2.2 Assessment Component: 2	8
Part:1.....	8
2.2.1 Data Pre-processing	8
2.3 Assessment Component: 2	11
Part-2.....	11
Data Analysis with SPSS	11
Part: a.....	15
Cluster Analysis	15
Part: b.....	17
Most Common data mining methods	17
Part: c.....	18
Advantages and disadvantages of using SPSS and Excel	18
Conclusion.....	20
References	21

Introduction

Process that ensuring research data is properly and securely stored, retained, or discarded throughout and after the end of a research study is known as data handling. It entails the development of rules and methods for managing data that is handled both digitally and manually. It is critical for ensuring the integrity of research data since it handles concerns such as privacy, security, and the preservation of research findings. A good data management plan may also lead to more efficient and cost-effective data storage, access, and deletion (Muthu, n.d.). Business intelligence (BI) is a term that refers to the technology, applications, methods, and procedures that are utilized to evaluate and visually portray pertinent business data. The main objective is to assist in making better business decisions, which are critical to a company's success. This evaluation is about both of these topics, where data handling is utilized for data and its analysis, and where business intelligence will be conveyed.

Analysis and Discussion

2.1 Current trends in data warehousing, business intelligence, and data mining

Data warehousing is the act of gathering and arranging data from many sources in order to provide meaningful business insights. It is often used to link and analyse business data from a variety of sources, and it is at the heart of data processing and display for management decision-making. The current trends of data warehousing are discussed below:

Open-Source DBMS for Deploying the DW: Any database system having a codebase that is free to view, download, alter, share, and reuse is considered a free software database. Open-sourced licenses enable developers to build applications based on current database technology. Open-source DBMSs continue to be utilized in both exploratory and more structured techniques. At the moment, open-source warehouses are uncommon, usually lower than traditional ones, and need a higher level of manual assistance. Some systems, on the other hand, are designed expressly for data warehousing.

Data Warehouse as a Service and Cloud: SaaS (Software-as-a-Service)-based Data Warehousing consists of programs that expedite data administration, boost the effectiveness of mission essential systems, and organize the information stored. It offers an online platform for managing information, data collecting, analysis, forecasting, and storage, among other things. Data warehousing as a service aggregates data from numerous sources into a single database, resulting in homogenous data that can be more readily managed. Businesses examine the information generated from this data to ensure compliance with the regulatory regulations. As a result, this web-based solution organizes information, records interactions, and automates data storage. Such solutions are either part of company processes or are necessary to address unique business requirements for a project, division, or department (Rizzi and Golfarelli, 2016).

In-Memory DBMS: In-memory DMBS (IMDBMS) technology provide extraordinarily rapid requested data and data commitment times, increasing the likelihood that analytics and operational applications can use the same database. Data analytic modelling, data management methods, and network operators inside a middle tier will start to emerge as the dominant strategy, compelling more conventional row-based suppliers to concurrently adjust to column methods and in-memory. Business Intelligence (BI) options will appear fairly soon,

using IMDBMSs with executive solutions, and will swiftly become targets for major companies.

Business intelligence (BI) refers to the use of software solutions to turn data into useable information that supports a corporation in making strategic and operational decisions. BI systems access and analyse data sets and display sharing data in presentations, highlights, dashboards, graphs, charts, and maps to give customers with accurate insight into the condition of the company (Fruhlinger and Pratt, 2022).

Literacy in Data: The capacity to read, interpret, and explain data in context, including an understanding of data sources and formats, analytical methods and methodologies applied, and the ability to express the use case, application, and value of the data, is referred to as "data literacy." Businesses are starting to see the value of incorporating data analytics across their organisation. Rather of segregating analysis and decision-making, firms are increasingly attempting to prioritize findings. Businesses want data to guide every choice they make, from goal formulation to strategy development to execution. To increase data literacy and inspire workers to value data, businesses must establish a data-first culture. Through training and programs, teams throughout businesses must learn how to interpret data and apply insights to their responsibilities.

Data Quality Management (DQM): Data quality is critical for assuring the veracity of analytics findings. It offers a context-specific method for increasing the fitness of data utilized for analysis and decision making. The objective is to gain insights on the health of that data by using diverse procedures and techniques to increasingly large and complicated data sets. Bad data can lead to inaccurate insights, which can lead to poor business decisions. Data of high quality is often accurate, consistent, comprehensive, timely, unique, and verifiable.

Automation and the API Economy: Though developing a software solution appears to be less expensive than acquiring one, organisations frequently struggle with development logistics. When you include in the need to gather data from on-premise systems, the cloud, streaming apps, and other sources, it gets daunting. For a shorter time to understanding, business apps should connect smoothly across interfaces, networking, and clouds. Companies can grow the capability of software by connecting it with other programs. While traveling from Facebook to an eCommerce site without quitting the software, combination of different characteristics

is common. It's really the API economy at work, and it's influencing the present and future of analytics, with BI solutions lecture given to users through their business apps.

Data mining is a technique for forecasting outcomes by monitoring and analysing dataset activity, trends, and correlations. It may also be utilized to create income, save costs, develop consumer relationships, and make other business choices. Companies can use data mining to learn about what their consumers are interested in or want to buy, as well as for detecting fraud and malware scanning. Current trends in data mining are given below:

Biological Sequence Mining: Nucleotide or amino acid sequences make up biological sequences. Biological data study compared, aligns, classifies, and analyses biological elements in bioinformatics and contemporary biology. The examination of biological sequences is critical in bioinformatics and contemporary biology. This type of analysis is divided into two tasks: pairwise sequence matching and multiple sequence matching. It is consequently critical to recover all of the general laws that data mining can recover from records in order to comprehend the genetic foundation of all biological activities (Han and Pei, 2012).

Symbolic Sequence Mining: A symbolic series is an orderly collection of components that can be represented with or without a temporal sense. The series may be used for a variety of purposes, such as consumer purchasing sequences, online clickstreams, application implementation sequences, biological sequences, and so on. The process of mining frequent patterns entails recognizing the substrings that appear often in one or more series. Several scalable algorithms have been created as a consequence of extensive study in this field.

Time Series in Mining: In a mining time series, a defined number of data points are collected at a specific moment or events gathered across repeated measures of time. Values or data are often measured at regular intervals such as hourly, monthly, or daily. Regular periods are also documented in time information, as are time-series elements such as trend, seasonality, cycle, and irregularity.

2.2 Assessment Component: 2

Part:1

2.2.1 Data Pre-processing

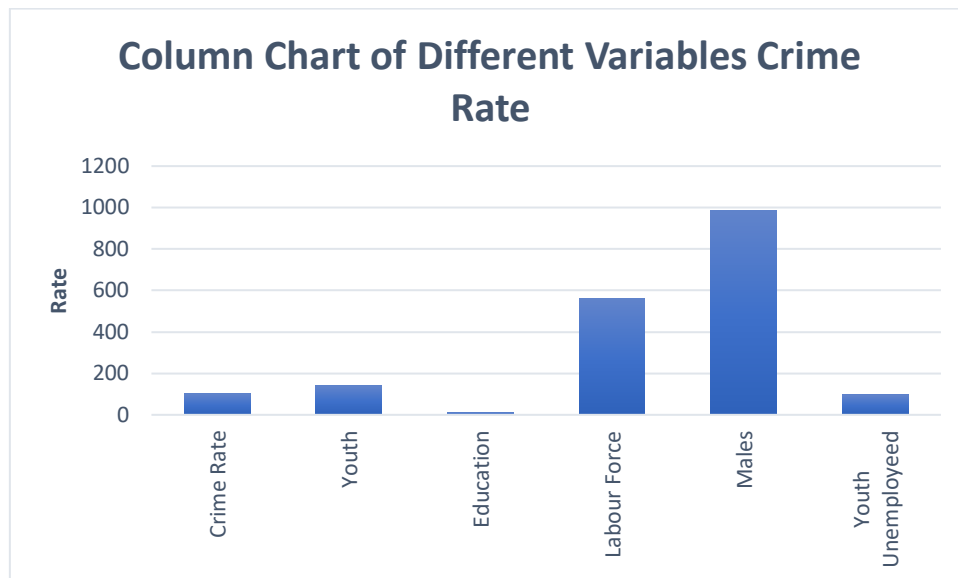
In a nutshell, data processing is the process of pre-processing data before going on to data analysis, and it is a step that converts unprocessed data into a format that computers and algorithms can comprehend and utilize to provide the best and most effective results. This section converts raw, real-world data in the form of text, photographs, video, and audio into useful data. Quality, pre-processed data is more crucial than even the most sophisticated algorithms, to the point where machine learning models trained on bad data may actually be damaging to the study (Garcia and Herrera, 2015). During this stage, the data has been processed and, in the way, the unnecessary data has been removed to make the analysis less confusing.

2.2.2 Data Analysis and Visualization

Data analysis and data visualization may have opposing viewpoints, yet they can also coexist together. This enables the creation of visual analytics that portray complicated data sets in an understandable manner. Successful data presentation allows audiences to get deeper insights quicker, allowing for improved decision-making and more effective analysis. This makes it simpler to display complicated data sets in an easy-to-understand and follow format, and it assists audiences in identifying insights, trends, gaps, and areas that need additional examination. This results in enhanced clarity and comprehension of the facts (Fekete et al., 2019).

Type	Result
Crime Rate	102.4393617
Youth	139.5531915
Education	12.39680851
Labour Force	563.3617021

Males	984.9468085
Youth Unemployed	96.45744681



The solution of crime rate with the “IF” function is presented below:

Type	Result with IF function
Crime Rate	102.4393617
Youth	139.5531915
Education	12.39680851
Labour Force	563.3617021
Males	984.9468085
Youth Unemployed	96.45744681

Another method of calculation of crime rate is pivot table and presented below as:

Average of CrimeRate	Average of CrimeRate10	Average of Youth	Average of Youth10	Average of Education	Average of Education10	Average of Males	Average of Males10	Average of YouthUnempl oyment	Average of YouthUnempl oy10
102.8085106	102.0702128	138.5744681	140.5319149	12.39148936	12.40212766	983.0212766	986.8723404	95.46808511	97.44680851
Crime Rate		Youth		Education		Males		Youth Unemployed	
102.4393617		139.5531915		12.39680851		984.9468085		96.45744681	

2.3 Assessment Component: 2

Part-2

Data Analysis with SPSS

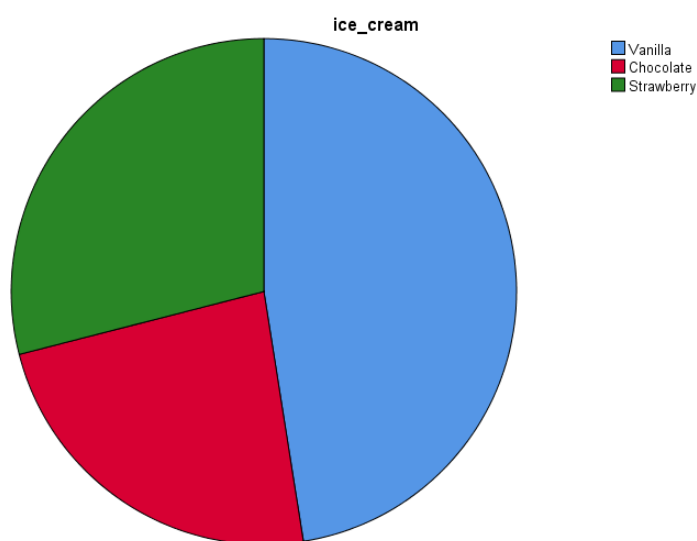
Part:1

The first goal is to identify the number of students who prefer vanilla flavour. The frequency feature under Descriptive Statistics function in the analysis choice is utilized, and the result is found as 95 out of 200 participants who like vanilla flavour.

The graphical demonstration is also being provide at the end of the data analysis by using SPSS.

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	



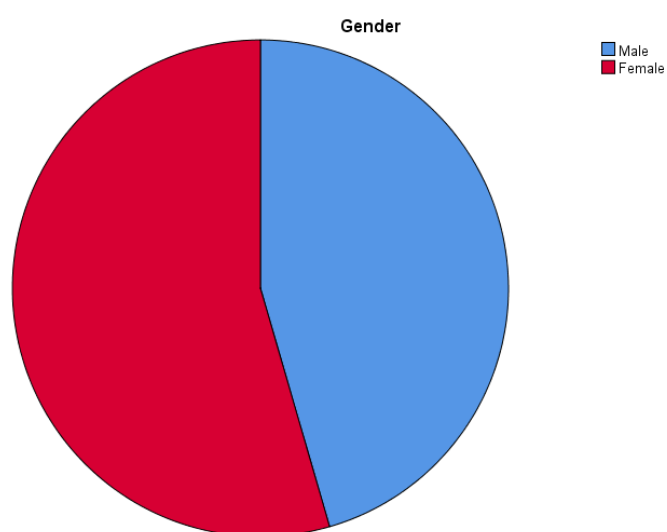
Part:2

For the task of how many students are Male and Female, the descriptive statistics is used for calculating or generating frequency table of male and female students and the result is found as 91 male and 109 females out of 200 students. As in the original data file the variable was defined as female and if the data there were both male and female students, the variables are changed into gender. The end result is presented below:

Gender

		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	

The graphical presentation is also being provided with using pie chart.



Part 3

In the third task the mean and median of participants who likes chocolate and strawberry flavours are asked to find out and for the result the same procedure is followed but instead of generating the frequency table the mean and median under the statistics of Descriptive statistics is clicked to calculate the values and the end result is presented in the below:

Statistics^a

ice_cream

N	Valid	47
	Missing	0
Mean		2.00
Median		2.00

a. ice_cream = Chocolate

Statistics^a

ice_cream

N	Valid	58
	Missing	0
Mean		3.00
Median		3.00

a. ice_cream = Strawberry

The graphical presentation hasn't been provided as the graph is not useable and doesn't demonstrate anything.

Part: a

Cluster Analysis

The roots of cluster analysis are based on one of the most fundamental, simple, and frequently undetected methods of thinking and learning: arranging items into comparable groups. This procedure employs a variety of techniques and approaches to create clusters of a similar type. It is also used in statistical analysis as part of data management.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
ice_cream	200	1	3	1.82	.857
video	200	26	74	51.85	9.901
puzzle	200	26	71	52.40	10.736
Valid N (listwise)	200				

Initial Cluster Centers

	Cluster	
	1	2
Zscore(ice_cream)	-.95099	.21587
Zscore(video)	2.23717	-2.30787
Zscore(puzzle)	1.26632	-2.45953

Iteration History ^a		
Iteration	Change in Cluster Centers	
	1	2
1	2.062	2.152
2	.123	.165
3	.072	.079
4	.074	.083
5	.055	.058
6	.035	.035
7	.000	.000
a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 7. The minimum distance between initial centers is 5.992.		

Final Cluster Centers

	Cluster	
	1	2
Zscore(ice_cream)	.27364	-.27916
Zscore(video)	.65316	-.66635
Zscore(puzzle)	.71944	-.73397

Number of Cases in each Cluster

Cluster	1	101.000
	2	99.000
Valid		200.000
Missing		.000

Part: b

Most Common data mining methods

Data mining methods are used in advertising, marketing, fraud prevention, new breakthroughs, design and development, healthcare, and education to identify hidden and unexpected links among data. Furthermore, associations utilize data mining methods to address business challenges such as boosting revenues, attracting new consumers, enhancing cross-selling and up-selling, and raising the Return on Investment (ROI) from marketing initiatives. As just a consequence, the Organisations consistently generate outcomes that keep firms ahead of the pack.

Analysis of Regression: All of these data mining approaches can aid in the analysis of various data sets from various viewpoints. A regression analysis is the process of detecting and assessing the connection between variables in statistics. It may also help to comprehend the relationship between the dependent variable's characteristic value and the independent variables, where one variable varies with the changing of the other. It denotes that one variable is dependent on another, but not the other way around, and is frequently used in forecasting and prediction.

Analysis of clustering: Observational data of multiple variables for each person are used in cluster analysis to sort similar results into groupings. Cluster analysis and discriminant analysis are conceptually connected. It examines the systematic method of finding groups and clusters in data so that the amount of link between two items can be determined whether they belong to the same group and if it is larger or lesser. It may be used to develop a consumer profile in order to make better business decisions based on customer behaviour.

Outer detection: This data mining method is concerned with the examination of data objects in a data collection that do not correspond to a typical behaviour or pattern. Outlier mining and Outlier analysis are other terms for the same thing. An outlier is a data point that deviates significantly from the rest of the database. Outlier detection is extremely important in the field of data mining. It is useful in a variety of domains, including credit or debit card fraud detection, network disruption detection, and outlier detection in wireless communication data (Regin et al., 2021).

Part: c

Advantages and disadvantages of using SPSS and Excel

Advantage of using SPSS

- SPSS is a sophisticated statistical software package that includes a basic spreadsheet format for data entry that is intuitive and straightforward to use.
- Like a built-in feature, several advanced statistical tests are provided.
- SPSS offers a wide selection of data documentation to assist maintain uniformity in data entry.
- The conclusions drawn is simple.
- IBM SPSS includes scatterplots, boxplots, and histograms to assist in identifying trends in data.
- For data analysis, IBM SPSS provides a wide selection of extremely configurable statistical models such as the general linear model and a variety of logistic regression (alchemer.com, 2021).

Disadvantage of using SPSS

- The main drawback of SPSS is that it cannot handle very huge data sets.
- Regarding academic application, SPSS lags far behind SAS, R, and possibly others that are more scientific than statistical in nature for current data analysis.
- The list selections are often the most basic of an assessment, and it makes it very easy to perform an incorrect analysis.
- It's ridiculously costly and requires a license, so one can't just purchase it and keep it.

Advantages of Excel

- One of the benefits of MS Excel is the availability of templates and macros. These help users to save time by automating repetitive activities, whereas templates are preprogramed files containing placeholder data for convenient use.

- Microsoft allows to add items from one program into another, making it easy to move data across tools.
- This can aid in the creation of expert presentations by providing slide layouts that allow for quick navigating between slides with the press of a button.
- Utilizing Microsoft Excel offers users more control over the data.
- It has a number of formulae and calculating capabilities that make bookkeeping extremely simple for accounting experts.
- One advantage of utilizing Microsoft Excel is that it can be accessed from any computer equipped with a Web browser (Donovan, 2022.).

Disadvantages of Excel

- MS Excel can't manage more than a few thousand rows without grinding to a halt, and it also has file size restrictions.
- MS Excel might be challenging to master due to the amount of options available in the menu.
- While using MS Excel, there are varying levels of difficulty for data analysis since different analyses may be on separate menu bars.
- It lacks rapid autofill functionality.
- Viruses may be added to an Excel file through macros (Donovan, 2022).

Conclusion

Analytics, mining, data visualisation, and best practises are used to help firms make better data-driven decisions via the usage of business intelligence (BI). In fact, modern business intelligence is defined as having a complete view of the organisation 's activities and being able to use that data to drive change, reduce inefficiencies, and quickly adapt to market or supply changes.

References

alchemer.com, 2021. What is SPSS and How Does it Benefit Survey Data Analysis?. Available at: <https://www.alchemer.com/resources/blog/what-is-spss/> (Accessed: 24 April 2022).

Donovan, M., 2022. Advantages and Disadvantages of MS Excel. [online] Blog.suitebriar.com. Available at: <https://blog.suitebriar.com/advantages-and-disadvantages-of-ms-excel> [Accessed 22 April 2022].

Fekete, J.D., Fisher, D., Nandi, A. and Sedlmair, M., 2019. Progressive data analysis and visualization.

Fruhlinger, J. and Pratt, M., 2022. What is business intelligence? Turning data into business insights | CIO. [online] CIO. Available at: <https://www.cio.com/article/272364/business-intelligence-definition-and-solutions.html> [Accessed 22 April 2022].

Garcia, S., Luengo, J. and Herrera, F., 2015. Data preprocessing in data mining (Vol. 72, pp. 59-139). Cham, Switzerland: Springer International Publishing.

Han, J., Kamber, M. and Pei, J., 2012. Data mining trends and research frontiers. Data Min, pp.585-631.

Regin, R., Rajest, S.S. and Singh, B., 2021. Spatial Data Mining Methods Databases and Statistics Point of Views. *Innovations in Information and Communication Technology Series*, pp.103-109.

Rizzi, S. and Golfarelli, M., 2016. What Time is it in the Data Warehouse?. In *International Conference on Data Warehousing and Knowledge Discovery* (pp. 134-144). Springer, Berlin, Heidelberg.