

Name: Zheko Valchanov

Student ID: 2024650

Data Handling and Business Intelligence
Assignment 2
BA Business Management
LCBB5000

Table of Contents

Introduction:	2
Data warehouse:	3
Business intelligence:	4
Data mining:	6
Part 1:	7
Data Pre-processing:	7
Data Analysis & Visualisation:	7
Part 2:	9
1.....	9
2.....	10
3.....	11
a)K-means clustering in SPSS.....	12
b) Data mining methods	14
c) Advantages & Disadvantages	15
Conclusion:	16
References:	17

Introduction:

Business Management is the procure which makes use of technology to analyse data and provide managers, executives and other related personals with useful information so that they may make better business choices. BI procedures gather data from IT systems and other sources, prepare it for analysis, perform queries on it, and then produce BI dashboards, visualisations, and reports for business users to utilise for operational and strategic planning (Data Analysis in Business Intelligence for Mathematical Prediction of Online Sale, 2016). There are several tools for processing, analysing and visualising data. In this report Microsoft Excel and IBM SPSS have been chosen as the data management programs. An introduction to current data warehouse, business intelligence, and mining technologies was included in the first part of the report as these are the topics this report based on. Excel an analytical tool is used to get a deeper grasp of the subject matter. SPSS, another data analysis tool, was employed in this case due to the ease with which different trends and patterns may be conducted and presented. SPSS's advantages and weaknesses were then explained in greater detail.

Data warehouse:

Having access to a vast volume of data in a data warehouse aids in decision-making. Relational databases system, transactional systems and others all regularly feed data into data warehouses. Analytical tools such as SQL clients, BI tools and others are used by data engineers, decision makers and business researchers also with analysts to get access to the data (Data Warehouse, 2022).

Tiers make up the structure of a data warehouse. Reporting, analysis, and data mining are all part of the front-end customer experience. The analytics engine is located in the intermediate tier and is responsible for accessing and analysing the data. The database server sits at the very bottom of the architecture, and it's here that data is loaded and saved. Object stores like Amazon S3 are used to store data that is rarely accessed, while SSD disks are used to store data that is often accessed (such as emails).

Multiple databases can be housed in a single data warehouse. Tables and columns are used to organize data in databases. Integer, data field, or text can be used to describe the data in each column. Schemas, or folders, can be used to organize tables in a database. Schema-defined tables are used to hold data that is ingested. Access and analysis of data are determined by query tools based on the structure of the data.

The following are some of the advantages of having a data warehouse:

- Decision-making based on solid information.
- Data culled from a variety of sources.
- Inquiry into the past
- Consistency, reliability, and quality of data
- Improves the performance of both systems by separating analytics processing from transactional databases.

In order to analyse vast amounts of information, a data warehouse is a specialized tool. Keeping track of financial transactions, for example, requires the usage of a database.

While a data warehouse can store organized and semi-structured data, it cannot store unstructured data in the same place as a data lake. Tables of data are required in a data warehouse, and this is

where the schema comes in. SQL can be used to query the data if the data is presented in a tabular fashion. It should be noted, however, that data in tabular format is not required for all applications. Data that is "semi-structured" or "totally unstructured" can nevertheless be accessed by some applications, such as big data analytics, full text search, and machine learning.

Data warehouses (DWHs) can be divided into three primary categories:

As a starting point, the EDW (Enterprise Data Warehouse): EDW is a centralized data storage facility. It serves as a decision-making tool for the entire company. A single strategy to data organization and presentation is provided. It also provides the option to classify data based on subject and grant access based on those classifications.

Operational Data Store: There are two types of operational data stores: operational data warehouse (ODW) and operational data store (ODS). When using ODS, the data warehouse is constantly being reloaded. As a result, it is frequently chosen for mundane tasks like archiving employee records.

The Data Mart: A data mart is a subset of the data warehouse. Sales, finance, sales, or finance are just a few examples of industries that might benefit from it. Data can be collected directly from sources in an independent data mart.

Business intelligence:

The use of business intelligence (BI) enables firms to swiftly identify actionable insights for making strategic decisions by analysing historical and current data. To make this happen, business intelligence tools process big data sets from numerous sources and provide conclusions in visual ways that are easy to grasp and communicate (What Is Business Intelligence | Microsoft Power BI, 2022).

A business intelligence process has four main processes that it takes in order to turn raw data into easily digestible insights for the entire organization. The first three steps—data gathering, analysis, and visualization—set the foundation for the last step of decision-making. Firms used to have to do a lot of their analysis by hand before using BI, but these technologies now automate a lot of the procedures, saving businesses both money and labour.

First phase: Collect and transform data from various sources

The extract, transform, and load (ETL) approach is commonly used by business intelligence products to combine structured and unstructured data from many sources. In order to make it easier for apps to evaluate and query the data, this data is first changed and redesigned.

Second phase: Identify patterns and discrepancies.

Automation is commonly used in data mining and data discovery to swiftly evaluate large amounts of data in search of patterns and outliers that reveal information about the present status of a business. There are a number of different data modelling and analytic techniques available in BI systems that may be used to investigate data, anticipate future trends, and make suggestions.

Third phase: data visualization

Business intelligence reports often make use of data visualisations to make conclusions easier to grasp and disseminate. A real-time perspective of the company is provided via graphs, dashboards, charts, and mapping that are interactive.

Fourth phase: Take immediate action based on new information

The capacity to quickly move from insights to action is made possible by putting current and historical data in context with business activities and processes. Business intelligence allows for immediate and long-term strategy changes that minimize inefficiencies, respond to market developments, fix supply difficulties, and address customer complaints.

The following are some of the unique advantages that firms gain from BI:

- Operational processes have become more efficient.
- Customers' habits and shopping habits are revealed.
- Financial and operational success can be accurately measured.
- Identifiable standards derived from both past and present data.
- Data irregularities and consumer difficulties are immediately alerted to you.
- The ability to communicate real-time data analysis throughout many divisions.

Data analysts and IT users were the primary users of business intelligence tools in the past. A new breed of self-service BI tools has made it possible for everyone from CEOs to operations staff alike to access business intelligence data.

Business Intelligence (BI) Developments

Take note of the upcoming developments in business intelligence and analytics.

Human cognition is no longer required to do complex tasks, according to Gartner. Data analysis and dashboard reporting can be done in real time with this functionality.

When BI software, social media, and other cutting-edge technologies are merged, teams are better able to work together and share information in order to make better decisions as a group.

It is possible to embed BI software, or at least parts of it, into another business application, so increasing and expanding the reporting capabilities of that application.

On the near future, BI software will be available in the cloud, and as a result, more companies will make the switch. According to their estimates, cloud-based analytics spending will rise 4.5 times quicker in the next few years.

Data mining:

Data mining methods may be used in the workplace to transform unstructured data into meaningful information. Businesses may get a better understanding of their clients and develop more cost-effective marketing strategies through using software to search for patterns in massive amounts of data. Data mining requires effective data collecting, storage, and processing.

- Data mining is the process of looking for patterns and trends in a vast amount of data.
- Businesses can employ data mining for a wide range of purposes, from figuring out what their customers want to buy to detecting fraud and spam.
- Data mining tools employ the information provided by the user to identify patterns and relationships in the data.
- Data mining techniques are employed by social media businesses in order to benefit from their consumers.
- The use of data mining to influence preferences has recently been criticized due to the fact that users are often unaware that data mining is taking place with their personal information.

The data mining procedure consists of five stages. Collecting data and putting it into a database is the first stage. A database or cloud service is used to store and manage acquired data. If you're a business analyst, you may access and arrange data in a number of different ways. Graphs and tables make it simple to communicate data with others, and application software does this by categorising it based on the user's input.

Part 1:

Using Microsoft excel and crime rate dataset we will pre-process the data, and then visualize the data after analysing.

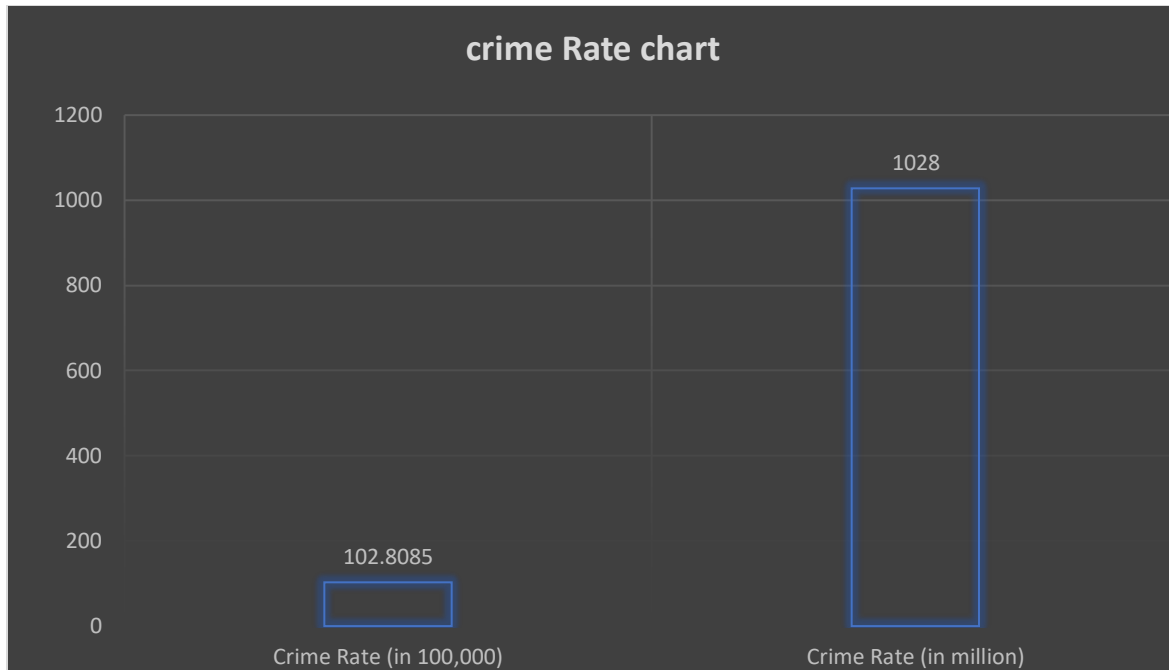
Data Pre-processing:

Processing raw data into a form suitable for data mining analytics is known as data preparation. It is common for raw data to be incomplete and formatted erratically. If the data preparation is inadequate, the project's success is directly related to the quality of the data. Data validation and imputation are both part of the pre-processing process. Analysing the data to determine its completeness and accuracy are the primary objectives of this phase of the data collection process.

Data Analysis & Visualisation:

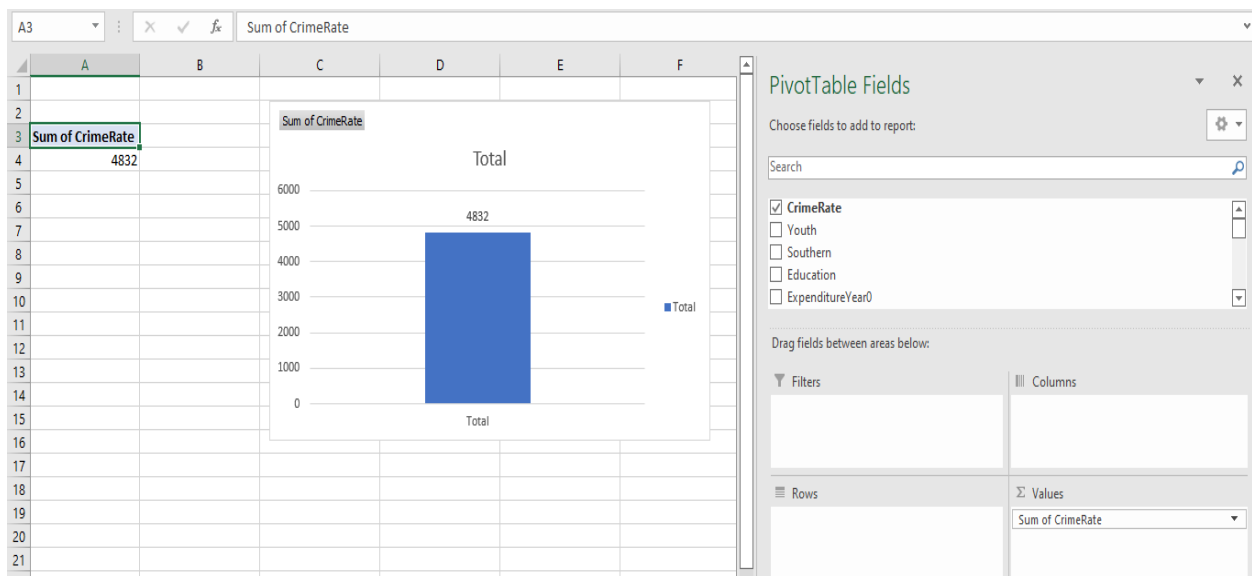
Visualization of data and information in a pictorial or graphical form is known as data visualization. Analysing data sets in order to make decisions based on the information they contain is referred to as data analytics.

The presentation of crime rate data:



According to the graph, there are 1028 violations of the law per million persons in the United States.

Pivotal table for sum of crime rate and its bar diagram.



Using VLOOKUP formula, we wanted to find 75 values in the crime rate table. Here is the formula and answer.

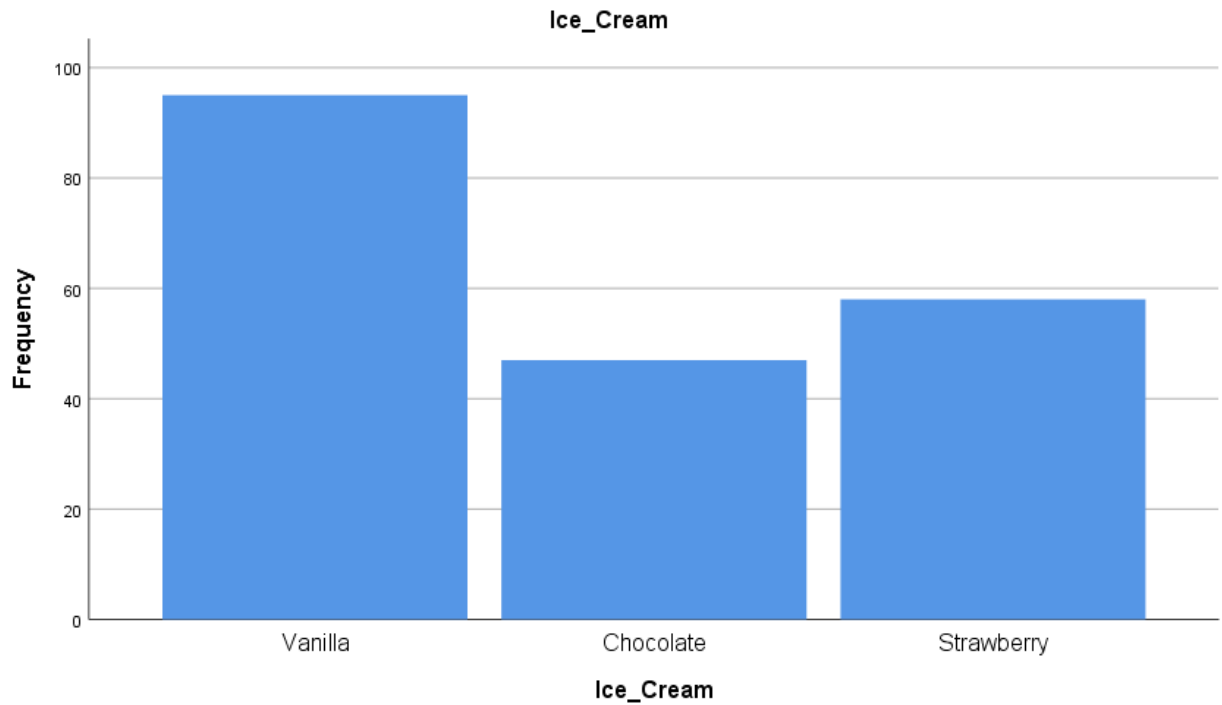
C51									
	A	B	C	D	E	F	G	H	I
40	132.4	152	0	12	82	571	1018	1	
41	135.5	125	0	12.5	113	567	985	0	
42	137.8	141	0	14.2	109	591	985	0	
43	140.8	150	0	12	109	531	964	0	
44	145.4	131	1	12.2	115	542	969	0	
45	149.3	143	0	12.3	103	583	1012	1	
46	154.3	124	0	12.3	121	580	966	0	1
47	157.7	136	0	15.1	149	577	994	0	1
48	161.8	131	0	13.2	160	631	1071	1	
49									
50									
51			75						
52									
53									

Part 2:

1.

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	

According to the SPSS-created figure, 95 participants liked vanilla ice cream, which is also the highest percentage of the three ice cream tastes at 47.5%.

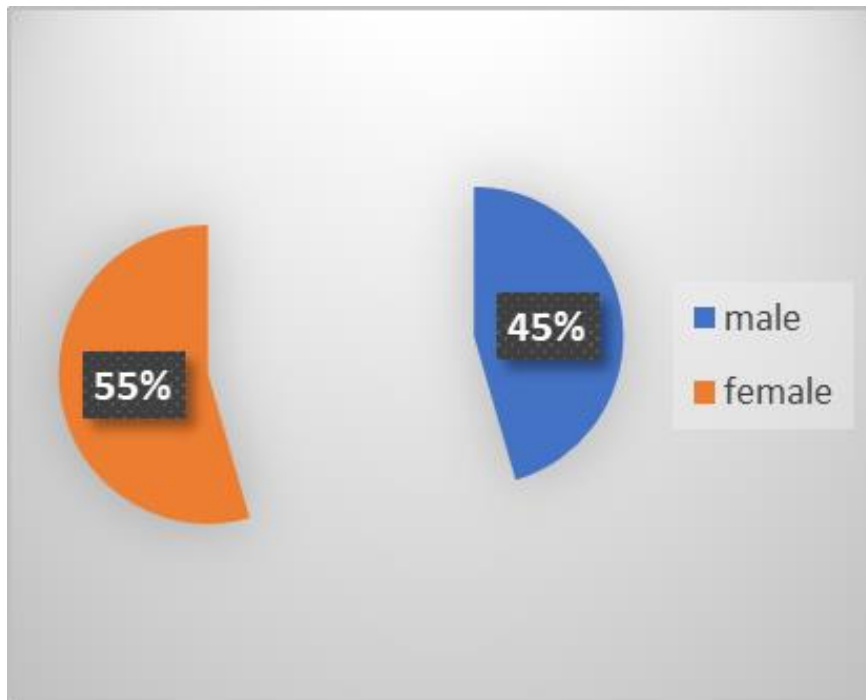


We can see the various ice cream flavours and their frequency by using a bar graphical depiction.

2.

The frequency table for the categorical variable gender and its two subcategories, male and female, was created using SPSS. In order to make the percentage of females and males clearer, a pie graph is also employed.

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	



3.

The following tables indicate the mean and median of people who favour chocolate and strawberry flavour. Mean and median can be seen to be close to one another. The two definitions of "chocolate flavour" are extremely similar. Compared to strawberry, chocolate has a median of 47.

Statistics			
		Video	Puzzle
N	Valid	47	47
	Missing	0	0
Mean		47.70	47.32
Median		47.00	46.00
Ice_Cream = Chocolate			

Statistics			
		Video	Puzzle
N	Valid	58	58
	Missing	0	0

Missing	0	0
Mean	55.45	57.14
Median	58.00	61.00
Ice_Cream = Strawberry		

a) K-means clustering in SPSS

We have divided the whole data into two cluster size. There is initial and final iteration history, ANOVA and the number of clusters in each size in the output. We see both clusters are not same sized. We have used dendrogram to approximately determine the cluster size. From ANOVA table we can say puzzle has more effect on variable.

Initial Cluster Centers		
	Cluster	
	1	2
video	29	74
puzzle	26	66

Iteration History ^a		
	Change in Cluster Centers	
Iteration	1	2
1	21.986	18.375
2	.739	.631
3	1.084	.969
4	.215	.219
5	.285	.262
6	.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 6. The minimum distance between initial centers is 60.208.

Final Cluster Centers		
	Cluster	
	1	2
video	45	58
puzzl e	44	60

Distances between Final Cluster Centers		
Cluster	1	2
1		20.770
2	20.770	

ANOVA						
	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
video	8700.783	1	54.579	198	159.415	<.001
puzzl e	12815.244	1	51.116	198	250.709	<.001

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Number of Cases in each Cluster		
Cluster	1	95.000
	2	105.000
Valid		200.000
Missing		.000

b) Data mining methods

1) Pattern mining and association analysis are often used techniques.

This is a form of data mining approach that aims to identify patterns in a dataset by analysing it. Using the database, it will hunt for intriguing correlations and links between the various entries and identify an overall pattern.

To give an example, consider "Shopping Basket Analysis," which involves determining "which things buyers are likely to purchase together at the store?" Bread and butter are two examples.

Marketing, cross-selling, and product placement are all examples of applications.

2) Correlational Analysis is the second step.

Correlation Analysis is nothing more than an application of Association Rules in a more sophisticated manner. When the parameters of support and confidence are used, consumers may still see patterns that are uninspiring.

In support of the assertion above, an example can be given: of the 1000 transactions examined, 600 only had bread; 750 had butter; 400 had both bread and butter; Suppose there is a 30% minimum level of support and a 60% minimum level of confidence in the association rule being followed.

The support value of 40% and the confidence level of 66% meet the requirement. The likelihood of purchasing butter is 75 percent, which is higher than the average of 66 percent. There is a

correlation between bread and butter since the purchase of one would lead to a drop in bread-and-butter purchases. The results can be misleading.

A further interestingness metric, correlation analysis, has been added to the support and confidence shown in the preceding example, which will aid in discovering interesting patterns (Illiashenko, 2019).

3) Induction of Decision Trees

Classification Analysis includes the Decision Trees Induction technique. An easy-to-understand tree-like form, a decision tree is simple and rapid. An attribute is tested on a node, and the result of the test is shown by branching out from that node. The label of the node reflects the class.

A tuple's attribute values are checked against the decision tree from the root to the leaf node. Due to the lack of specialized expertise required, decision trees are popular. Multiple dimensions can be represented with these. Classification rules can be simply derived from decision trees.

Decision trees are used in a variety of fields, including medical, manufacturing, and astronomy.

c) Advantages & Disadvantages

Comparatively speaking, the following are the benefits of utilizing SPSS as a software package:

- In addition, SPSS is a powerful statistical software package.
- A large number of complicated statistical tests are built in.
- Results are straightforward to decipher.
- Data tables may be seen quickly and easily.
- Potentially expandable

LIMITATIONS

- For students, SPSS might be a pricey buy.
- To make the most of all the tools at your disposal, you'll usually need to invest in further training.
- Microsoft Excel's graphing tools are more complex than those of OpenOffice.org.

Advantages of excel:

For the most part, spreadsheets are used to collect and organize data, which is one of their most basic functions. Spreadsheets are popular because they eliminate the need for the user to perform any calculations. Calculations involving multiple variables can be performed with ease once a formula is created and a set command is entered into the application. Nowadays, many people in a workplace need to be able to view the same documents because of the way we work together. Spreadsheets created in Microsoft Excel can be shared, but only one person can make changes at a time.

Limitations of excel:

Because only the data that the user chooses to analyse is included, it's possible that other pertinent information that might influence a choice will be overlooked. Calculations must be entered into the spreadsheet as formulas, which might be a hassle for certain users. Each sort of calculation has its own syntax that must be learned in order to do it. Another problem of spreadsheets is the absence of data security. As a result, spreadsheets are more susceptible to data corruption and mismanagement than other types of files. Access to financial data may still be compromised even when password-protected (Albright, Winston and Zappe, 2010).

Conclusion:

People may access information whenever and wherever they need it thanks to BI software, which goes beyond spreadsheets to provide a complete image of the company. However, finding a BI strategy that matches a company's capabilities, experience, and budget can be particularly difficult for small and mid-sized businesses. BI systems often begin with periodic reports or a reactive approach to corporate events. According to such reporting, historical data is frequently used. However, it is possible to move on from that. Predictive decision-making will become more common in the future. There are no limits when it comes to analysing client perceptions as well as the overall health of a firm thanks to Big Data, mobile computing, internal databases, and the cloud. It's impossible to keep up with the ever-changing nature of business intelligence (BI).

References:

Albright, S., Winston, W. and Zappe, C., 2010. *Data analysis and decision making with Microsoft Excel*. Mason, Ohio: South-Western.

Amazon Web Services, Inc. 2022. *What is a Data Warehouse? | Key Concepts | Amazon Web Services*. [online] Available at: <<https://aws.amazon.com/data-warehouse/>> [Accessed 15 April 2022].

Illiashenko, K., 2019. THE APPLICATION OF DATA MINING METHODS IN ACCOUNTING. *Pryazovskyi Economic Herald*, (6(17).

International Journal of Science and Research (IJSR), 2016. Data Analysis in Business Intelligence for Mathematical Prediction of Online Sale. 5(4), pp.2020-2024.

Powerbi.microsoft.com. 2022. *What Is Business Intelligence | Microsoft Power BI*. [online] Available at: <<https://powerbi.microsoft.com/en-us/what-is-business-intelligence/>> [Accessed 15 April 2022].