

Student Name: Hena Rodriques

Student ID: 1921350

Module Code: LCBB5000

Module Title: Data Handling & Business Intelligence Assignment-Part-2

Module Lecturer's Name: Joy Botu

Part 1

Many academic studies in the fields of crime data mining and criminal investigation have been conducted, and data mining has been the focus of many of these studies. By recognizing and avoiding potential difficulties, criminal investigation researchers hope to reduce, prevent, and improve the overall quality of criminal investigations. Here are a few notable research publications in the relevant field that you might be interested in. Some authors described crime variables that analysed and forecasted criminal activity using the relationship between computer science and criminal justice. There is also an in-depth explanation of some of the difficulties that law enforcement authorities have in detecting crime patterns and trends on a consistent basis. Clustering is the most often utilized method. The classification system returns the results of crimes that have occurred as well as those that have been solved. The use of a data mining strategy, which is to say, analysing the crimes as well as the data mining techniques used, is done in five steps in the criminal analysis process.

Since 1997, some of the authors have been employing a broad data mining system based on knowledge gathered through the COPLINK project, with the help of University of Arizona professors and the Tucson and Phoenix police departments. During this time, the authors got support from the Tucson and Phoenix police agencies. These unique data mining methods for separating structured and unstructured data also contain a variety of other creative data mining methods for separating structured and unstructured data. The classifications for the various data mining approaches, as well as the data mining methodologies and classifications for the various data mining components. The contemporary Indian e-governance system has piqued their curiosity. This article provides a brief overview of an interactive query-based interface used by the National Crime Record Bureau (NCRB) as a tool for criminal inquiry and investigation. In addition, this essay examines the Indian police system, as well as key components of the new Crime Criminal Information System (CCIS) and Common Integrated Police Application (CIPA), as well as their current status and shortcomings, and describes many crime prevention data mining methodologies, with the authors employing specific tactics in each of the following steps: data

extraction and gathering, pre-processing, grouping, and classification. The WEKA tool described in [9] was used to create visualizations in this study, along with Google's marker clustering (GMAPI) and crime-prone zones on the Indian map, all of which were sourced from Google. Random forest and cross-validation are used to validate the dataset that has been chosen.

Investigators used clustering algorithms to improve their capacity to detect and prevent unlawful activities, hence enhancing their effectiveness. They were able to save the data while also predicting the consequences thanks to this technology and a clustering method. The goal was for government officials to be able to forecast crime and offenders based on historical data and geographical location using the findings of this study. The authors also look at a number of previously created systems, as well as the benefits and drawbacks of the algorithm they utilized. Some of the writers went on a study in which they employed forecasting and prediction methods to try to find patterns of criminal behaviour by forecasting and predicting crime. The research was compiled using data from criminal record databases such as India's National Crime Records Bureau (NCRB), the FBI, and the Central Intelligence Agency (CIA). Classification techniques such as the Bayesian Network algorithm and the regression algorithm are used when predicting data and anticipating the eventual end-product. These algorithms are used to create a visual representation of the final product. For data analysis and prediction, Weka was employed, as well as prediction models such as the Auto-regressive Integrated Moving Average (ARIMA) and Artificial Neural Network (ANN) (ANN). The authors discovered six visualization modules and two machine learning algorithms, K-Nearest Neighbour and Naive Bayes, for identifying sites that are more likely to be targeted by criminals. It was demonstrated in this paper how to use a variety of tools and packages to create a web-based crime prediction mapping and visualization application in R.

Some of the authors looked into some of the challenges that have come up in the field of criminal data research. They discussed, among other things, the research methodology, which was centered on crime kinds and local law enforcement, as well as the description and interpretation of crime data mining, which was based on methodologies, technology, and other challenges. They employed the decision tree approach and the Apriori algorithm to identify and categorize the data in the correct order of occurrence, both of which are explained in detail below. They discussed how their tests went and what they learned from them on the pre-process tab of the spreadsheet.

One of the most important steps of this research was identifying, correcting, or eliminating and replacing missing, erroneous, or duplicate data records. The merging of datasets, which was one of the most crucial aspects of this study, was also completed during this period. This stage must be completed ahead of time in order to anticipate calculations and models working properly, as well as to collect consistent data. At the outset of the study, the data was manually pre-processed in Excel before being used for analysis.

It is critical to understand how to identify and obtain the best features from unemployment, homelessness, and employees earning less than the minimum wage based on location and race attributes (which describe the target variable crime count) in order to improve the accuracy of the crime prediction model and eliminate multi collinearity. In order to better comprehend the linear link between a range of variables and the frequency of crimes, a vast number of crime studies have looked at Pearson Correlation Coefficients. The findings of this investigation confirm that there is a linear relationship between them, as demonstrated in a previous study. It was made feasible to use Excel's capabilities, such as the pivot table. IF functions and lookups were used as part of the data preparation process to clean, organize, and prepare the dataset for the study. Descriptive analysis and data discovery approaches were utilized with great care and precision to determine whether or not the datasets had missing values.

We performed person correlation tests using default and customized parameters, as well as accuracy evaluations of the train and test runs, using the dataset. We may be able to gain some pretty fascinating insights by looking at what was identified as a result of the correlation study. In terms of the statistics, it supplies us with some fascinating details. For the purposes of this inquiry, three different approaches were used. Pearson Correlation, Multiple Linear Regression, and K-means Clustering are all examples of data analysis techniques. The Pearson Correlation Coefficient can be used to determine whether or not two persistent variables have a statistical link or correlation. It is typically regarded as the most effective method for determining the relationship between revenue components because it is dependent on the technique employed for covariance estimation. It provides information about the strength of the association or link as well as the extent to which it exists. The correlation coefficient method of analysis was used in this study to determine and assess the strength of potential relationships between, among other things, unemployment and the London crime rate, homelessness and the London crime rate, and

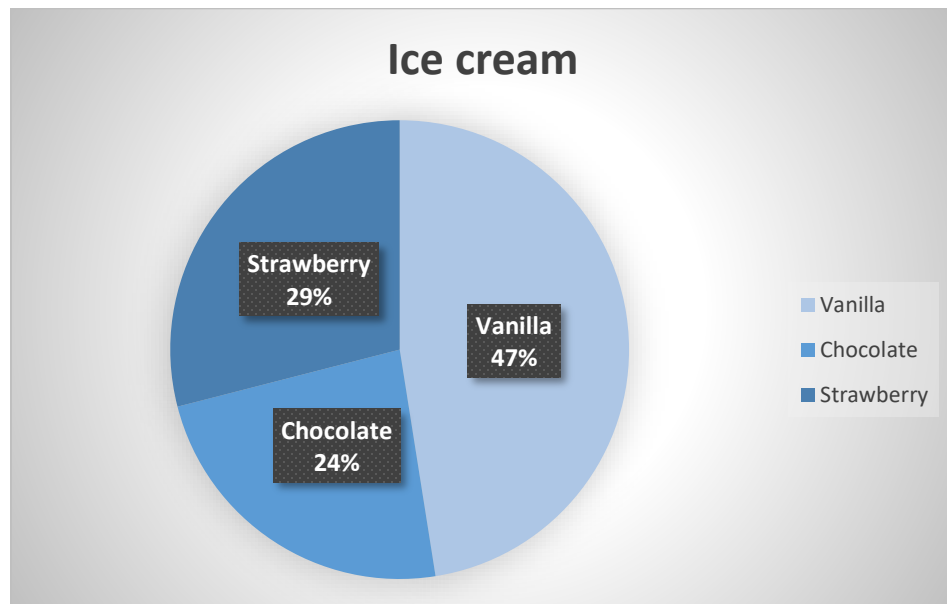
employees earning less than the minimum wage and the London crime rate. Several It's a type of linear regression that differs from standard linear regression in that it uses several instances of variables rather than just one.

The association between two or more independent factors, such as unemployment, homelessness, and employees earning less than the federal minimum wage per hour, and one dependent variable, such as crime rate statistics, was investigated using the multiple linear regression approach. This investigation will be carried out using the chosen K-means clustering method of analysis. This method takes data points like crime rate numbers as input and organizes them into k clusters based on how similar they are. The process of combining crime statistics from several sources into a single report is known as assembling. Decision trees, random forests, and kernel neural networks are some of the other alternatives that could be used in the research (kNNs). All of the known classification methods, as well as the Random forest approach, are among these options. In order for these models to be effectively used, the relevant data types and analyses would have to be separated in order for the model to make the correct forecast. The difference edge is a fundamental baseline method to feature selection that has been around for a long time and is still in use today. Certain attributes are eliminated because their variance does not exceed a predetermined limit. This strategy excludes all zero-variance features from consideration, which are those that have the same value across all samples. Because the data collected had missing values and unwanted columns, this feature selection technique was the most appropriate in this investigation.

Part 2

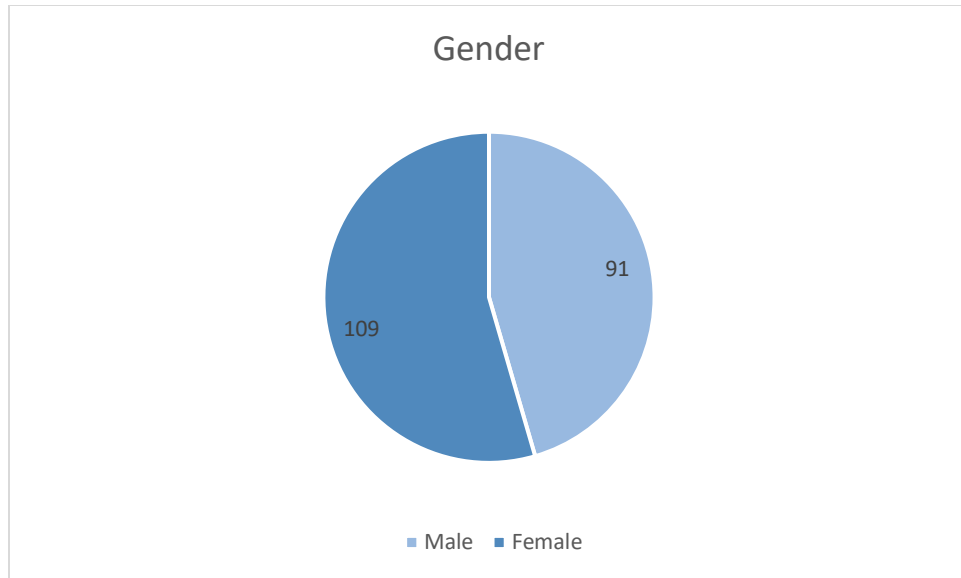
1) How many students like vanilla flavour?

Ans: 95 students like vanilla.



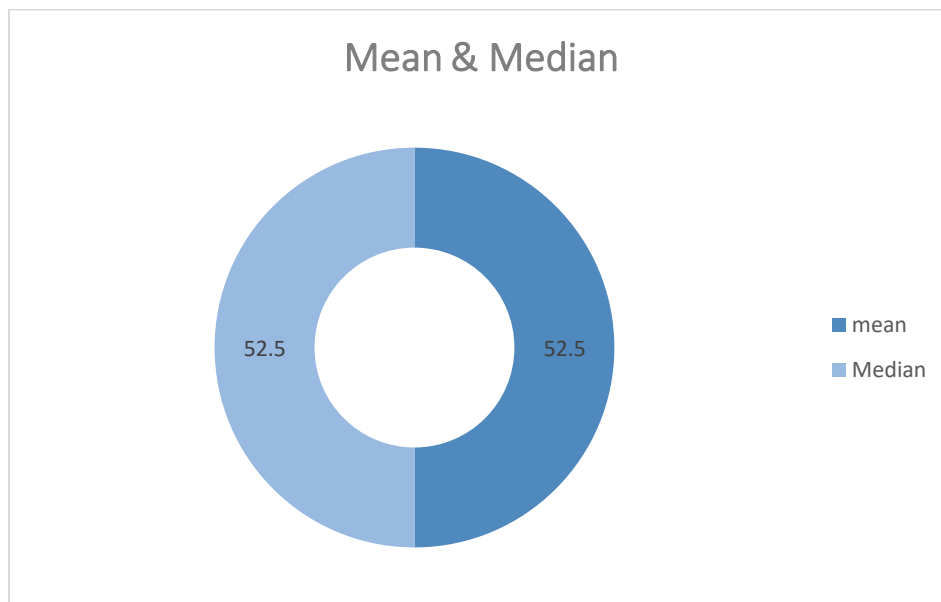
2) How many students are Male and Female?

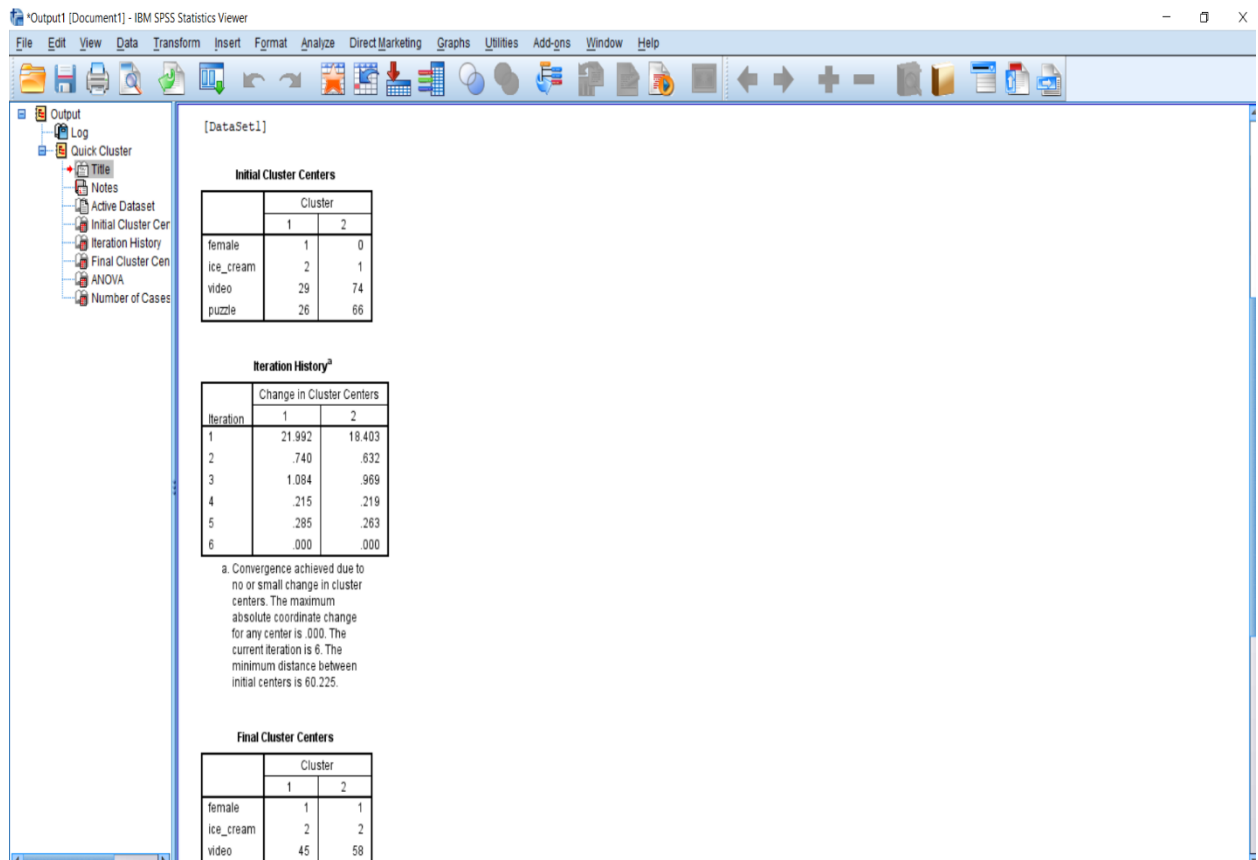
Ans: 91 students are male and 109 students are female.

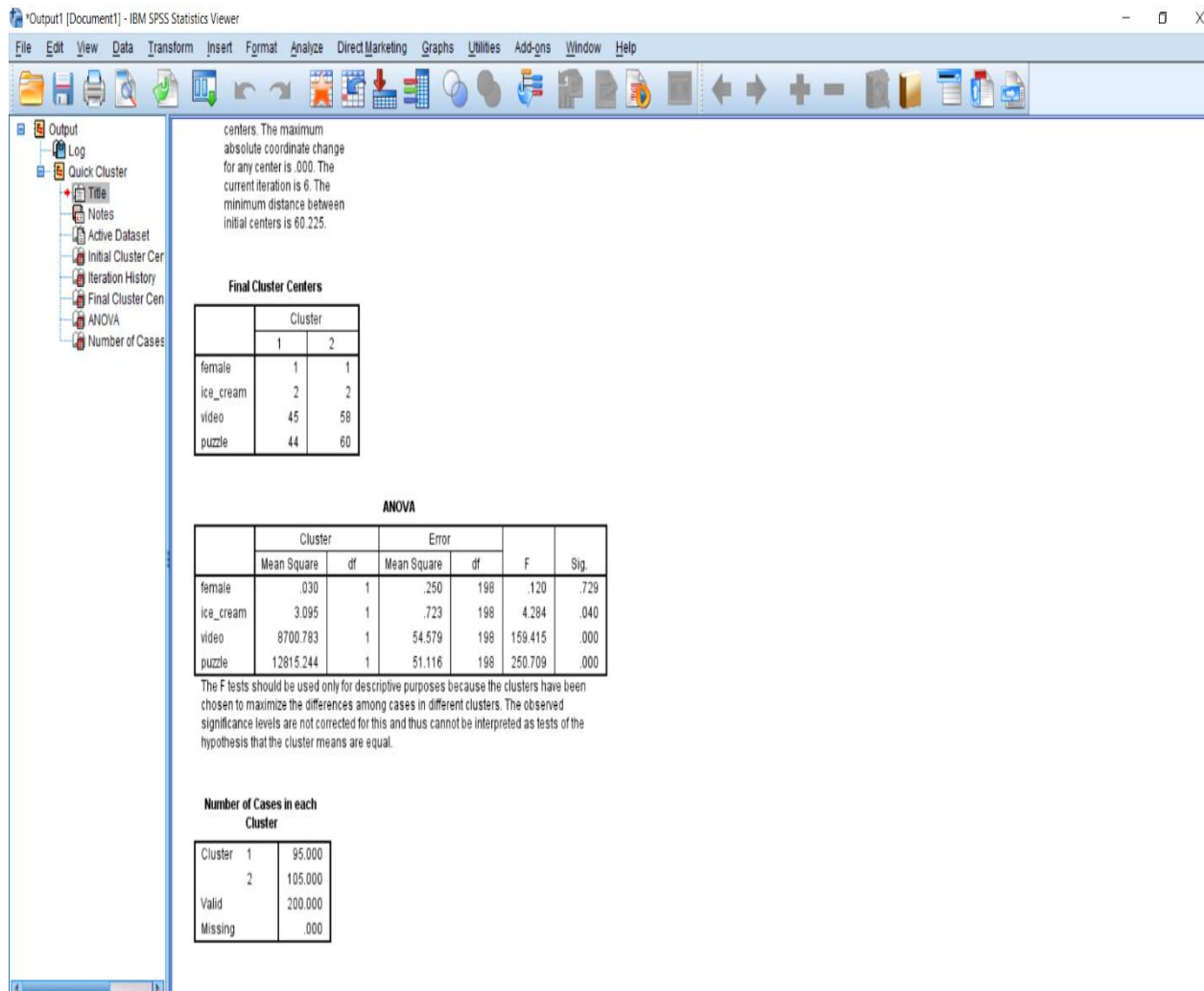


3) What are the Mean and Median of participants who likes chocolate and strawberry flavour?

Ans: The mean and median is 52.5 who likes chocolate and strawberry flavour.







K-means clustering uses "centroids," which are K randomly generated points in the data, to assign each data point to the centroid that is closest to it. The centroid is adjusted to the average of all the points that have been assigned to it once each point has been assigned.

b. At a manufacturing factory, unintentional leaks of hazardous material into the environment were causing worry; as a result, the company was in danger of being forced to close or undertake costly upgrades. The facility's workforce was able to discover warning signals of previous tragedies and prevent them from happening again using data mining techniques. This resulted in cost savings for the corporation in terms of money and reputation, as well as new jobs and increased tax revenue for the local community.

A business discovered while conducting Internet research that by integrating customer registration data from loyalty programs with first-time purchase data, it was possible to anticipate new

customers' predicted long-term spending potential. These enhancements allowed the company to better focus marketing efforts and increase return on investment, allowing it to stay afloat and grow through challenging economic times.

Using data mining, a large insurer was able to discover offices that were substantially more efficient than other offices in the same industry in processing common claim types. They were able to do an analysis of productive offices and find best practices that could be replicated elsewhere using this knowledge, resulting in increased productivity and improved customer service for the company.

Despite the fact that text mining is a relatively new field of computer science, its popularity has surged in recent years as the volume and importance of unstructured data has increased dramatically in both quantity and relevancy.

As proven by the use of text mining in Risk Management Software and Cybercrime applications, among other areas, it is possible to make vast volumes of unstructured data accessible and usable, providing value while also attaining ROI from unstructured data management.

It's the process of extracting usable information and knowledge from textual data utilizing techniques like classification, entity extraction (text entities), sentiment analysis, and other related approaches. To put it another way, the ability to find patterns and insights in even huge amounts of unstructured data in the business sector is a result of this. Its ability to filter out irrelevant data while still delivering answers is what has led to its widespread adoption, particularly among major organizations.

C. Advantages and disadvantages of SPSS and excel are given below:

SPSS:

Advantages

- The SPSS program offers a great degree of versatility and customization when it comes to organizing and managing your data. As a result, because the computer remembers the position of the variables and cases, data analysis is faster and more accurate.
- Excel files can be imported rather easily.
- All computations are handled for you via the user-friendly "plug and chug" function.
- In SPSS, it's standard practice to keep the output separate from the data by saving all of the results in a different file from the one that contains the data. You won't have to worry about accidentally overwriting any other data on your PC if you use SPSS.
- While Microsoft Excel is a great tool for data organization, using a specially designed analytic and statistical program like SPSS can provide you with more in-depth, faster, and accurate data analysis than using Excel alone.

Disadvantages

- Expensive
- It necessitates some training.
- The ability to read and interpret output is required.
- There are a lot of possibilities; you'll need to know how to choose the best ones for the analysis you want to do.
- Data visualization capabilities are limited.

Excel

Advantages:

- Tables are simple to create and dynamic as well.
- Columns that have been computed are included in tables.
- More than 50 different table styles are pre-configured on the tables.
- Slicers are available on the tables.
- Tables include well-organized references.
- In the Excel table, the column headings are always visible.

- The built-in User form in Excel can be used to retrieve data that has been accidentally deleted.
- Tables are easy to navigate and refer to.
- In the office, Excel is used to prepare financial reports and perform routine accounting activities. You can retain all of your important documents at work. You'll be able to give your manager with those sorted facts afterwards. Annual credit and debit sheets are also possible to make.
- You can password protect your Excel file so that no one else can see your spreadsheet data. Furthermore, this program's security is superior than that of other programs.
- This program allows you to store a large amount of data. This application can save all of the data from the previous year. Additionally, you can use the charts and other visualization tools supplied with this program to organize data. You can store millions of rows of data with your data.
- Any gadget can be used: Android, iOS, Windows, and Mac PCs are all supported by this software. This program also has an online version that you may use to collaborate with other users and share information.
- You can print and distribute via email as well. It is simple to send your information to your supervisor or another colleague if they require it. They can access the same file and data from their mobile device or PC.
- This application provides free templates, and there are a lot of them. Tables and graphs, for example, can be made using a variety of templates.
- A single Excel file can contain many spreadsheets that can be saved as one document. Every spreadsheet includes the option to generate financial and inventory data reports.

Disadvantages

- This program will become unresponsive if you save a huge amount of data in it. When you use the application, it will take longer for it to respond to your actions, and you will have to spend more time inputting or viewing data.
- As you are probably aware, Excel has macros. Anyone may use macros to introduce a virus into an Excel sheet, infecting your computer in the process.

- Data handling mistakes: If you employ the incorrect formula, your data will be erroneous. It's also difficult to tell whether or not this program is incorrect. Because this software is capable of storing millions of items, detecting errors becomes more challenging.
- Pricey: The price of this program is quite high. There are various alternatives to it, such as OpenOffice, which is a free tool for creating spreadsheet files. You can also get OpenOffice for free.
- Learning this program is more difficult than learning any other software, even Microsoft Word. There are formulae to learn in the program, and learning formulas is challenging for a novice.