

Student Name: Ronju Costa

Student ID: 1921242

Module Code: LCBB5000

Assignment Title: Assignment-2

Module Title: Data Handling & Business Intelligence (Current Trends of Business intelligence.)

Module Lecturer's Name: Anupam Mazumdar-GP1

Amitave Banik-GP2

Part 1

Many academic research have been undertaken in the domains of crime data mining and criminal investigation, and data mining has been the subject of many of these studies. The purpose of criminal investigation researchers is to decrease, prevent, and improve the overall quality of criminal investigations by identifying and avoiding potential problems. Here are a few examples of noteworthy research publications in the related topic that you might find interesting. Some authors described crime variables that took use of the relationship between computer science and criminal justice in order to analyze and forecast criminal activity. An in-depth discussion is also provided on some of the issues that law enforcement officials have in spotting criminal patterns and trends on an effective basis. The approach of clustering is the one that is most commonly

used. The outcomes of crimes that have happened, as well as those that have been solved, are returned by the classification system. There are five steps in the criminal analysis process that involve the use of a data mining approach, which is to say, analyzing the crimes, as well as the data mining tools that were employed.

Also, some the authors have been using a general data mining system based on familiarity gained from the COPLINK project with the support of scholars at the University of Arizona and the assistance of the police departments of Tucson and Phoenix since 1997. The authors have received assistance from the police departments of Tucson and Phoenix during this time. These unique data mining methods for differentiating between structured and unstructured data which also includes a number of other novel data mining methods for distinguishing between structured and unstructured data. The data mining techniques and classifications for the various data mining components, as well as the classifications for the various data mining approaches. There are particularly interested in the current Indian e-governance system. A brief explanation of an interactive query-based interface, which is used by the National Crime Record Bureau (NCRB) as a tool for criminal inquiry and investigation is provided in this article. As an added bonus, this essay provides a thorough examination of 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, pattern prediction, and visualisation. To construct visualizations in this study, the WEKA tool described in [9] was used in conjunction with Google's marker clustering (GMAPI) and crime-prone zones on the Indian map, both of which were sourced from Google. Validation of the dataset that has been selected is accomplished through the use of random forest and cross-validation.

Clustering algorithms were employed to increase the ability of investigators to predict and prevent illegal activity from occurring, hence increasing their effectiveness. This technology, along with a clustering strategy, allowed them to keep the data while also anticipating the consequence that would occur. It was intended that by using the outcomes of this research, government employees would be able to forecast crime and offenders based on historical data and physical location. The authors also examine numerous previously developed systems, as well

as the pros and limitations of the algorithm that was used in their research. Some of the authors embarked on a study in which they sought to identify patterns of criminal behavior by forecasting and predicting crime, and they used forecasting and prediction tools to do this. Information from criminal record databases, such as India's National Crime Records Bureau (NCRB), the FBI, and the Central Intelligence Agency, was used to compile their report (CIA). When predicting data and forecasting the final end-product, classification techniques such as the Bayesian Network algorithm and the regression algorithm are employed. These algorithms are used to visualize the ultimate end-product. Weka, as well as prediction models such the Auto-regressive Integrated Moving Average (ARIMA) and Artificial Neural Network (ANN), were used for data analysis and prediction (ANN). Six visualization modules and two machine learning approaches, namely K-Nearest Neighbor and Naive Bayes, were found by the authors in for recognizing places that are more likely to be targeted by criminals. In this work, it was presented how to build a web-based crime prediction mapping and visualization application in R by leveraging a range of tools and packages.

Some authors investigated some of the issues that have been identified in the field of crime data research. Among other things, they spoke about the research methodology, which was focused on crime types and local law enforcement, as well as the description and interpretation of crime data mining, which was based on methods, technology, and other difficulties. To identify and categorize the data in the right sequence of occurrence, they have used the decision tree approach and the Apriori algorithm, both of which are described in detail below. On the pre-process tab of the spreadsheet, they spoke about how their experiments turned out and what they learned from them.

Identifying, correcting, or removing and replacing of missing, erroneous, or duplicate data records was one of the most essential stages of this investigation. It was also during this phase that the merging of datasets was accomplished, which was one of the most important phases of this study. In order to anticipate calculations and models functioning well, as well as to collect consistent data, it is essential to complete this stage in advance. The data was manually pre-processed in Excel before being utilized for analysis at the start of the project.

When analyzing and obtaining 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), it is critical to understand how to identify and obtain the best features from these variables in order to improve the accuracy of the crime prediction model and eliminate multicollinearity. A large number of crime studies have analyzed Pearson Correlation Coefficients in attempt to better understand the linear link between a variety of variables and the frequency of crimes. This work's data confirms that there is a linear link between them, as demonstrated in earlier study. The use of Excel's features, such as the pivot table, was made possible. In order to clean, organize, and prepare the dataset for the study, IF functions and lookups were employed as part of the data preparation process. In order to establish whether or not the datasets had missing values, descriptive analysis and data discovery methodologies were used with great care and precision.

Using the dataset, we conducted person correlation tests with default and customized parameters, as well as evaluations of the train and test runs' accuracy. By looking at what was discovered as a consequence the correlation analysis, we may get some really fascinating insights. In terms of the data, it provides us with some quite interesting information. Three distinct methodologies were taken into account for the objectives of this investigation. Data analysis techniques such as Pearson Correlation, Multiple Linear Regression, and K-means Clustering are all examples of this. It is possible to establish whether or not there is a statistical relationship or correlation between two persistent variables using the Pearson Correlation Coefficient. Because it is dependent on the technique used for covariance estimation, it is often recognized as the most effective way for determining the relationship between revenue components. It offers information on the extent to which the association or link is present and how strong it is. In this study, the correlation coefficient method of analysis was used to determine and assess the strength of potential relationships between unemployment and the London crime rate, homelessness and the crime rate in London, and employees earning less than the minimum wage in relation to the crime rate in London, among other things. Several It is a sort of linear regression that goes beyond the ordinary linear regression in that it uses numerous instances of variables rather than just one.

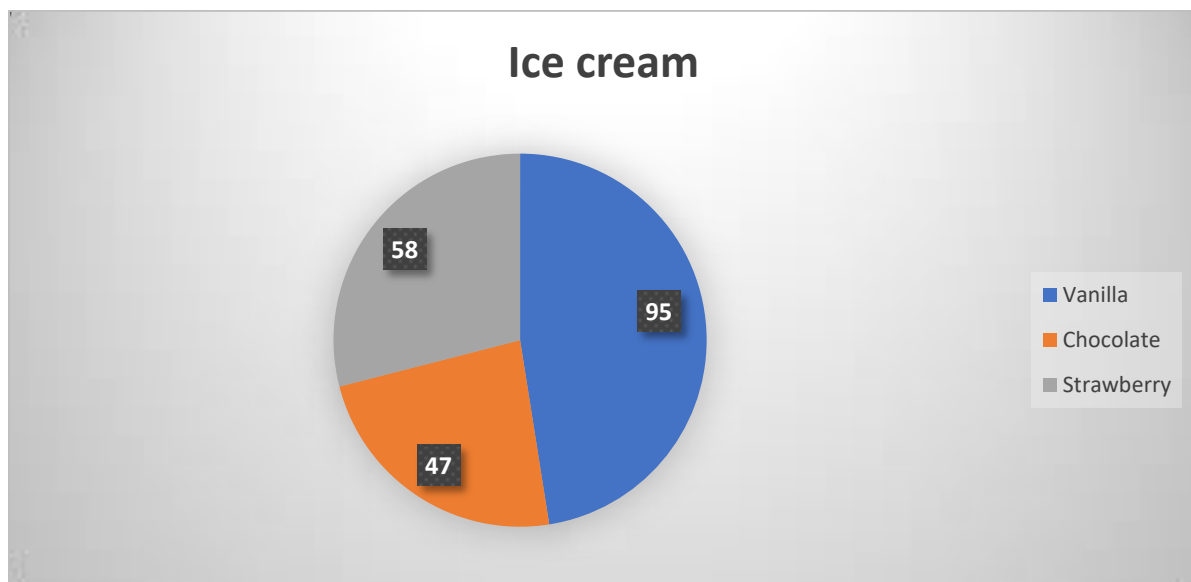
The multiple linear regression approach of analysis was used to examine the relationship between two or more independent factors, such as unemployment, homelessness, and employees earning less than the federal minimum wage per hour, and one dependent variable, such as crime

rate statistics. This examination will be carried out using the K-means clustering method of analysis, which was chosen. This method takes as input data points such as crime rate statistics and arranges them into k clusters depending on how similar they are to one another. This is the process of compiling crime statistics from several sources into a single report. There are a variety of alternative possibilities that may be employed in the study including decision trees, random forests, and kernel neural networks (kNNs). Among these alternatives are all of the many classification methods available, as well as the Random forest approach. The essential data types and analysis would have to be separate in order for the model to create the right forecast in order for these models to be effectively applied. It has been around for a long time, yet the difference edge is a fundamental baseline approach of feature selection that is still in use today. There are certain qualities that are removed because their variance does not surpass a predefined limit. Every one of the zero-variance characteristics, which are those that have the same value across all samples, is excluded from consideration by this method. In this inquiry, this strategy to feature selection was the most suitable due to the fact that the data that was chosen had missing values as well as undesirable columns.

Part 2

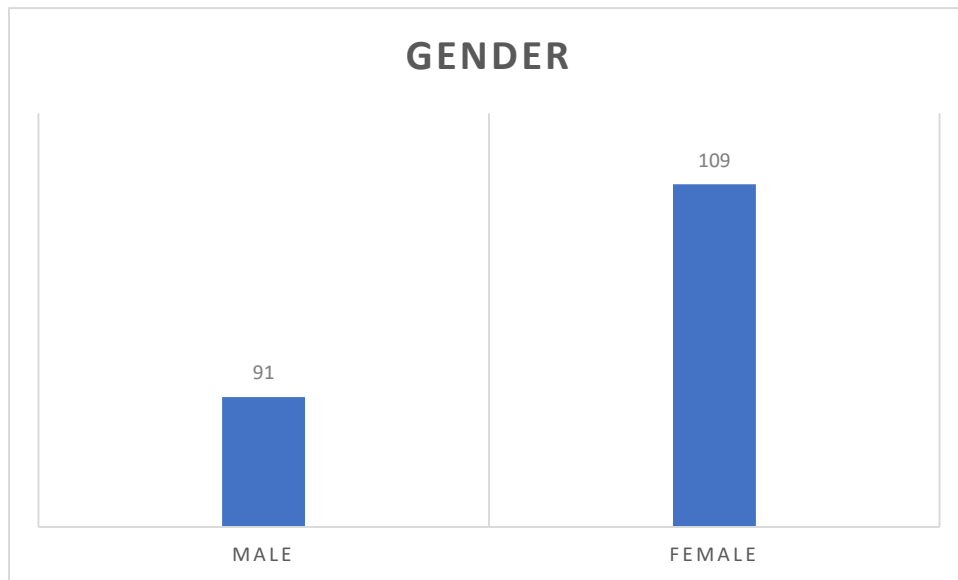
1) How many students like vanilla flavour?

95 students like vanilla.



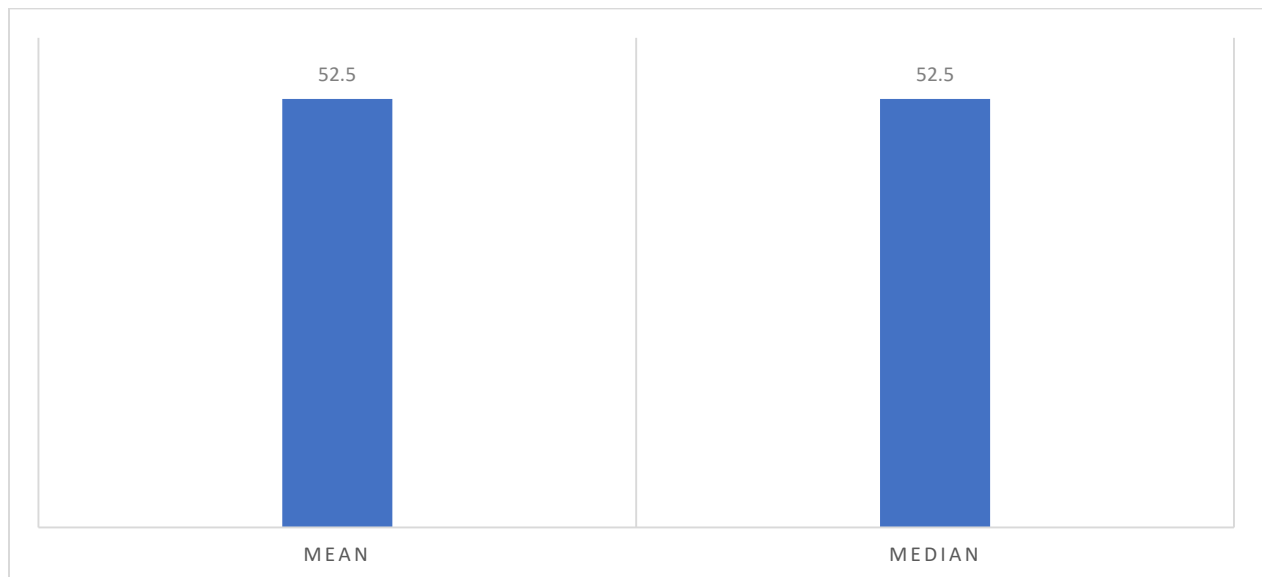
2) How many students are Male and Female?

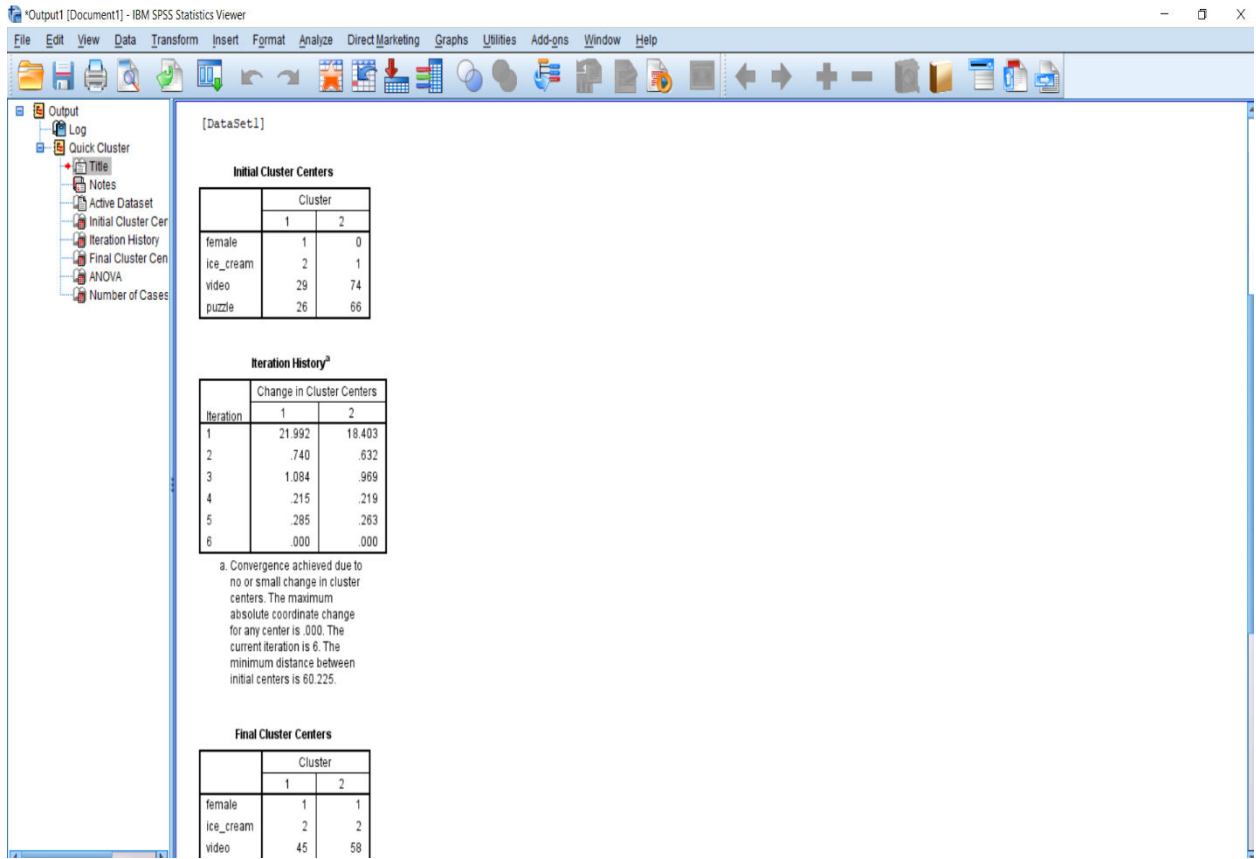
91 students are male and 109 students are female.

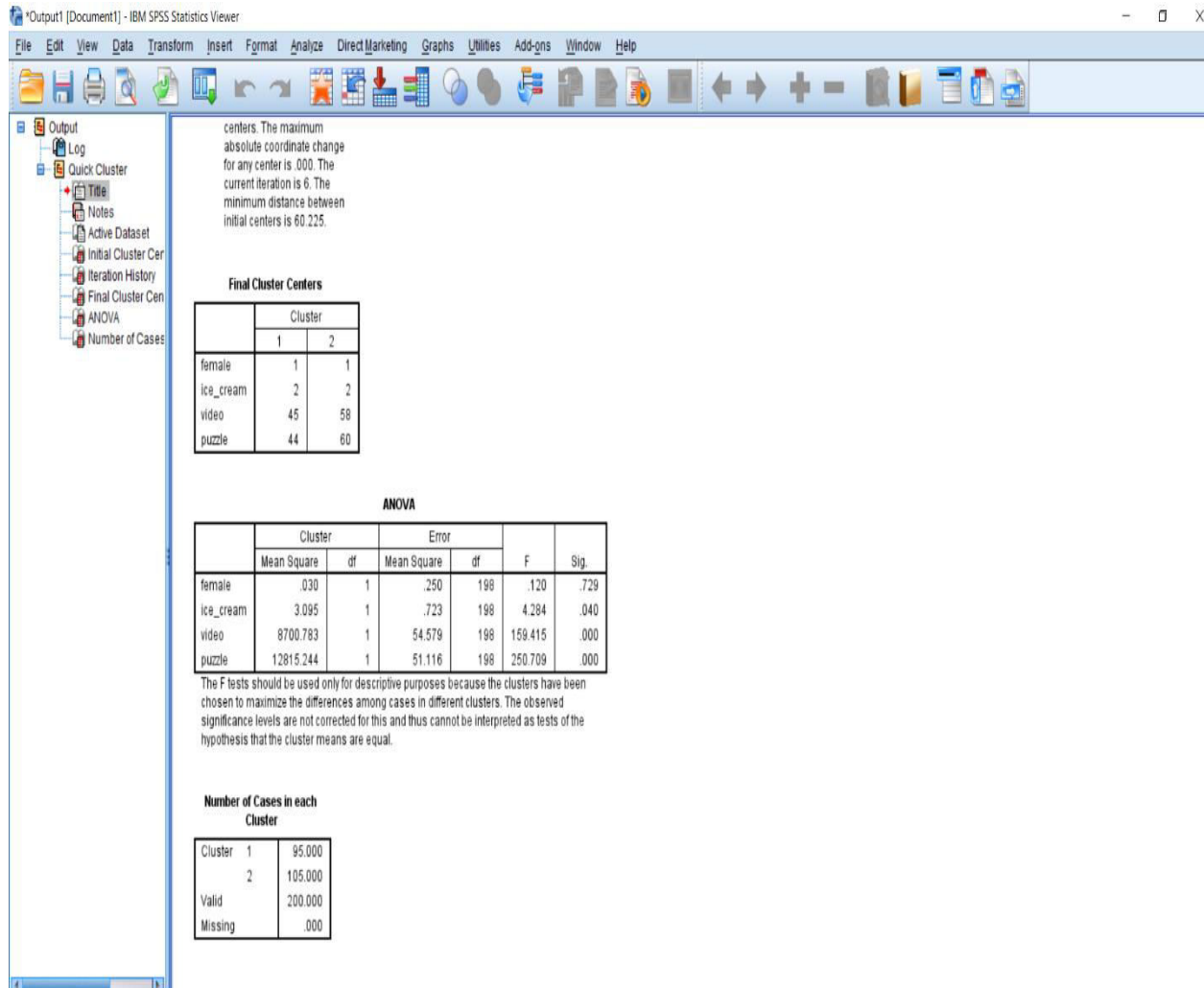


3) What are the Mean and Median of participants who likes chocolate and strawberry flavour?

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







K-means clustering makes use of "centroids," which are K separate randomly-initiated points in the data, and assigns each data point to the centroid that is closest to that data point. Following the assignment of each point, the centroid is changed to the average of all of the points that have been allocated to it.

- b. Unintentional leaks of dangerous material into the surrounding environment were causing concern at a manufacturing plant; as a result, the facility was in risk of being forced to close or undergo highly expensive renovations. By applying data mining techniques, the facility's crew was able to identify warning signs of these disasters and prevent them from reoccurring in the future. In terms of money and reputation, this resulted in cost savings for the company, while also creating new jobs and increasing tax revenue for the local community.

In the course of performing Internet research, a retailer discovered that by combining customer registration data from loyalty programs with first-time purchase data, it was feasible to predict the predicted long-term spending potential of new customers. These improvements enabled the business to more effectively focus marketing efforts and enhance return on investment, which helped to keep the company alive and growing through the difficult economic times.

Using data mining, a major insurer was able to identify offices that were significantly more efficient at processing common claim types than other comparable offices in the same industry. It was with this information that they were able to conduct an analysis of productive offices and identify best practices that could be reproduced elsewhere, resulting in higher productivity and enhanced customer service for the business.

Despite the fact that text mining is a relatively new subject of computer science, its use has grown in recent years as the amount and significance of unstructured data available has expanded considerably in both quantity and relevance.

Using text mining, it is possible to make enormous volumes of unstructured data accessible and useable, producing value while also achieving ROI from unstructured data management, as evidenced by the usage of text mining in Risk Management Software and Cybercrime applications, among other places.

It is the process of collecting useful information and knowledge from textual data using techniques such as classification, entity extraction (entities extracted from text), sentiment analysis, and other related approaches. Or to put it another way: being able to identify insights and patterns in even massive volumes of unstructured data in the business sector is a result of this. It is precisely this ability to dismiss any non-relevant information while still delivering answers that has led to its broad adoption, particularly among major corporations.

C.

Advantages	Disadvantages
SPSS	
• Excel files may be imported with relative ease. • The user-friendly "plug and chug" function takes care of all the computations for you.	• Requires some training • A lot of options; need to know how to select appropriate options for the analysis you would like to run • Need to be able to read and

• It is common practice in SPSS to keep the output separate from the data by saving all of the results in a separate file that is distinct from the file in which the data is included. When you use SPSS, you won't have to be concerned about accidentally overwriting any other information on your computer. • Even while Microsoft Excel is a fantastic tool for data organization, employing a particularly created analytic and statistical program such as SPSS may give you with more in-depth, quicker, and more accurate data analysis than using Microsoft Excel alone. • When it comes to organizing and managing your data, the SPSS program provides the user with a great deal of flexibility and customization. As a result, data analysis is faster and more accurate since the program remembers the position of the variables and cases.	appropriately interpret output • Potential problem = too easy to run analyses without understanding them • May be expensive • Limited data visualization capabilities
Excel	
• Table: The table in Excel has several features, including the following:- • Tables are simple to build, and they are also dynamic. • Tables include columns that have been computed. • Tables are pre-configured with more than 50 different table styles. • Tables are equipped with slicers. • Tables include references that are organized. • The heads of the columns in the Excel table are always visible. • Excel's built-in User form may be used to recover data that has been erased. • Tables are simple to choose from and refer to. • Work at the office: • Excel is used in the office to create financial reports and to perform everyday accounting tasks. You may keep all of your vital information at the	• Attack by a virus: As you are aware, Excel has macros. The use of macros allows anybody to inject a virus into an Excel file, which in turn causes your machine to get infected with the virus. • Learning this program is more harder than learning any other software, including Microsoft Word or any other software. In the program, there are formulae that must be learned, and knowing formulas is difficult for a rookie. • Pricey: The cost of this program is rather expensive. There are numerous alternatives to it, for example, OpenOffice is a free application that you may use to create spreadsheet files. OpenOffice is also available for download. • Errors in data handling: If you use the wrong formula, your data will be incorrect. Determining whether or not this program has mistakes is also tough.

workplace. Later on, you'll be able to provide that sorted facts to your manager. Credit and debit sheets can be created on an annual basis as well. • Excel password protection: You may password secure your Excel file so that no other user can access your data in the spreadsheet program. Additionally, the security of this program is superior than that of other applications. • A large amount of data: This program allows you to store an enormous amount of data. This program has the capability of storing all annual data. In addition, you may arrange data by utilizing the charts and other visualization tools that are included with this program. With your data, you have the ability to store millions of rows. • Use on any device: This app is compatible with Android, iOS, Windows, and Mac computers. There is also an online version of this program that you may use to cooperate with other users and to share information with them. • Printing and sharing: You can also print and share through email. If your supervisor or another college need access to your information, it is simple to transmit it to them. Using their mobile device or PC, they may access the same file and data. • Free templates are accessible on this program, and there are many of them. Tables and graphs may be created using a variety of templates, as an example. • A single Excel file contains numerous spreadsheets, which can be saved together as a single document. The ability to create financial and inventory data reports is built into every spreadsheet.	Because this software has the capability of storing millions of entries, identifying mistakes in it becomes more difficult. • Becomes unresponsive: If you save a large amount of data in this program, the software will become unresponsive. When you utilize the program, it will respond to your actions slowly, and you will have to spend more time entering or viewing the data in it.
--	--