

Data handling and business intelligence

Module code: LCBB5000

Student number : 2024897

Hristina Rageva

Table of Contents

Task 1	3
1.1 Evaluation of Microsoft Excel	3
1.2 Exercise of if function	5
1.3 Exercise of Pivot table and charts	6
Task 2	9
2.1 Gender.....	9
2.2 Lunch.....	9
2.3 Mean and median	10
2.5 Cluster.....	11
2.6 Data mining and text mining methods.....	13
2.7 Advantages and Disadvantages of Using SPSS and Excel.....	14
References	16

Task 1

1.1 Evaluation of Microsoft Excel

Excel is a powerful programme that may be used for a broad range of data preparation tasks, such as finding and removing duplicates, applying filters to data, and fixing errors. By employing filters such as date, value, or language, users of Excel's powerful filtering features are able to isolate pertinent data for the purpose of doing additional analysis. Excel is a terrific tool for doing activities involving simple data processing and analysis, but it is possible that it is not the ideal option for operations that are more sophisticated (Hussain and Abbas, 2020). Before the data can be analysed, they need to go through a step called pre-processing, in which they are sorted, cleansed, and converted. Only after this can the data be used. Excel's various features and functionalities include the ability to find and get rid of duplicates, filter data, and correct errors. The built-in functions and formulae of Excel may, among other things, format data, standardise data, and filter out information that is extraneous to the task at hand. Formulas and functions like this may be found under the Data tab (Mehta and Singh, 2020). Excel, on the other hand, may not be the best choice for data processing activities that are more involved.

The process of examining information for hidden meanings, structures, and linkages is known as data analysis. Excel provides users with a wide variety of statistical tools and capabilities, some examples of which are regression analysis, hypothesis testing, and correlation analysis. However, users may come into circumstances in which they are required to apply more advanced statistical methods or computer languages in order to carry out certain data analysis activities. The process of examining data in order to discover previously unknown patterns, connections, and insights is known as data analysis (Roque and Reis, 2019). The statistical capabilities in Excel may be used to test hypotheses and analyse associations, and the program's pivot tables can make the analysis of large datasets considerably easier. Excel-based spreadsheets may also benefit from pivot tables. Excel is a flexible programme that can be used to create both simple and intricate models for financial analysis and projection. Excel's statistical toolkit includes correlation analysis, hypothesis testing, and regression analysis, and the program's pivot tables make it simple to summarise and evaluate big datasets (Manupati and Anisetti, 2020). Regression analysis is also included in Excel's statistical toolset. Excel provides users access to capabilities like these. Excel may also be used to assist in the creation of more complicated financial models and forecasting

models on occasion. It's possible that one'll need to resort to different statistical approaches or computer languages in order to complete data analysis assignments that are more complicated.

The act of displaying data in a manner that can be easily comprehended by making use of graphical representations such as charts and graphs is referred to as data visualisation. Excel provides users with a wide variety of formatting choices, labelling systems, and colour schemes, allowing them to create charts that are uniquely suited to their needs and preferences. Users are able to apply various styles to cells depending on user-defined criteria when using conditional formatting, and Sparklines may be used to show data trends. Both of these features are available in Excel. Excel has a wide variety of chart types, including bar charts, line charts, and pie charts, to name just a few of them (Kumar and Saini, 2020). Excel's in-built visualisation capabilities are helpful for creating basic charts and visualisations, but more complex data visualisation projects may need the use of more powerful data visualisation platforms like as Tableau or Power BI. Excel's in-built visualisation features are advantageous for producing simple charts and visualisations. The discipline of displaying data in a visual format in order to facilitate improved comprehension is referred to as data visualisation. Customers have the ability to personalise their charts by picking from the wide variety of formats, labelling methods, and colour schemes that are offered by Excel. Conditional formatting tools shape cells according to user-defined criteria, while Sparklines display data patterns (Koo and Kim, 2020). Both types of tools are available in Microsoft Excel. Excel also enables users to create several kinds of charts, including line charts, bar charts, and pie charts, amongst others. On the other hand, more complex data visualisation projects can call for enterprise-grade software.

1.2 Exercise of function

Hotel Reservation (2) [Compatibility Mode] - Excel (Product Activation Failed)

File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do... Sign in Share

Clipboard Font Alignment Number Styles Cells Editing

Calibri 11 A A

B I U Wrap Text

General

Conditional Formatting Format as Table Cell Styles

Insert Delete Format

Σ AutoSum Fill Clear

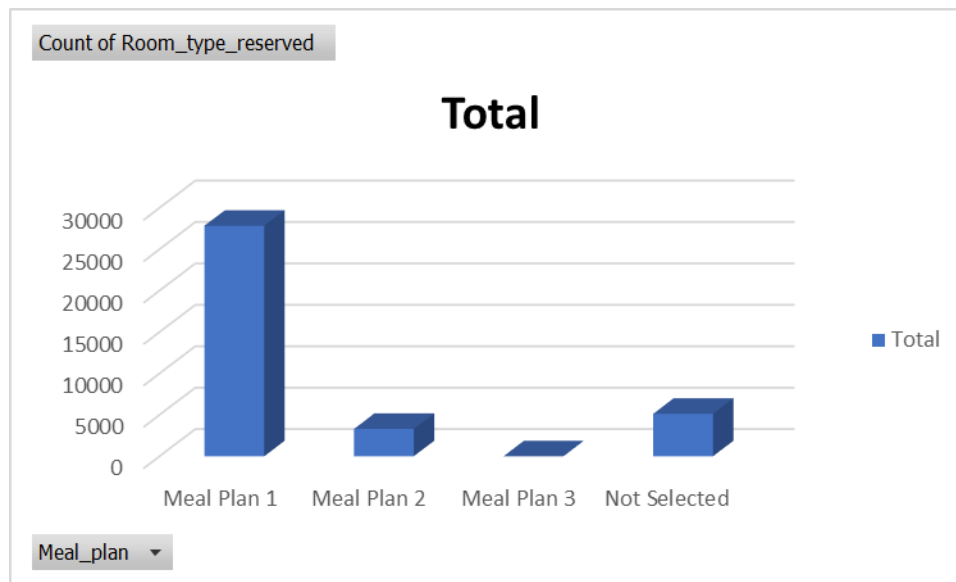
Sort & Find & Filter Select

T2 X ✓ fx =IF(I2>10, "Lead time is more than 10 days", "Lead time is less than 10 days")

	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
	Room_type_reserved	Lead_time	Arrival_year	Arrival_month	Arrival_date	Market_type	Repeated_guest	Previous_Cancellation	Previous_bookings	Average_price	Special_requests	Booking_status	If function		
1															
2	Room_Type 1	224	2017	10	2	Offline	0	0	0	65	0	Not_Canceled	Lead time is more than 10 days		
3	Room_Type 1	5	2018	11	6	Online	0	0	0	106.68	1	Not_Canceled	Lead time is less than 10 days		
4	Room_Type 1	1	2018	2	28	Online	0	0	0	60	0	Canceled	Lead time is less than 10 days		
5	Room_Type 1	211	2018	5	20	Online	0	0	0	100	0	Canceled	Lead time is more than 10 days		
6	Room_Type 1	48	2018	4	11	Online	0	0	0	94.5	0	Canceled	Lead time is more than 10 days		
7	Room_Type 1	346	2018	9	13	Online	0	0	0	115	1	Canceled	Lead time is more than 10 days		
8	Room_Type 1	34	2017	10	15	Online	0	0	0	107.55	1	Not_Canceled	Lead time is more than 10 days		
9	Room_Type 4	83	2018	12	26	Online	0	0	0	105.61	1	Not_Canceled	Lead time is more than 10 days		
10	Room_Type 1	121	2018	7	6	Offline	0	0	0	96.9	1	Not_Canceled	Lead time is more than 10 days		
11	Room_Type 4	44	2018	10	18	Online	0	0	0	133.44	3	Not_Canceled	Lead time is more than 10 days		
12	Room_Type 1	0	2018	9	11	Online	0	0	0	85.03	0	Not_Canceled	Lead time is less than 10 days		
13	Room_Type 4	35	2018	4	30	Online	0	0	0	140.4	1	Not_Canceled	Lead time is more than 10 days		
14	Room_Type 1	30	2018	11	26	Online	0	0	0	88	0	Canceled	Lead time is more than 10 days		
15	Room_Type 1	95	2018	11	20	Online	0	0	0	90	2	Canceled	Lead time is more than 10 days		
16	Room_Type 1	47	2019	10	20	Online	0	0	0	94.5	2	Not_Canceled	Lead time is more than 10 days		
17	Room_Type 1	256	2018	6	15	Online	0	0	0	115	1	Canceled	Lead time is more than 10 days		
18	Room_Type 1	0	2017	10	5	Offline	0	0	0	96	0	Not_Canceled	Lead time is less than 10 days		
19	Room_Type 1	1	2017	8	10	Online	0	0	0	96	1	Not_Canceled	Lead time is less than 10 days		
20	Room_Type 1	99	2017	10	30	Online	0	0	0	65	0	Canceled	Lead time is more than 10 days		
21	Room_Type 1	12	2019	10	4	Offline	0	0	0	72	0	Not_Canceled	Lead time is more than 10 days		

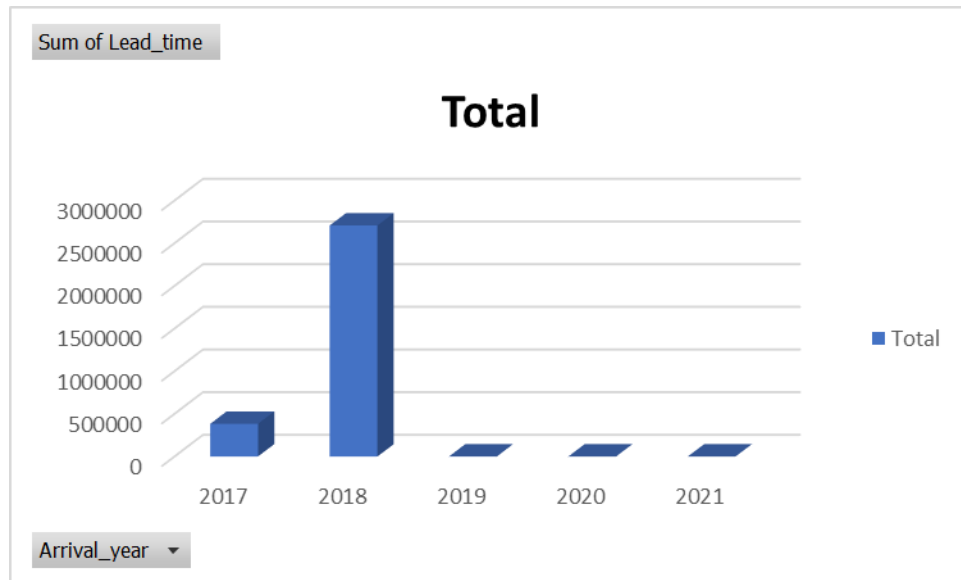
One may use an if statement like the one shown below in an Excel formula to check whether the number in cell I2 is bigger than 10. A result of "Lead time is more than 10 days" is given by the algorithm if this holds true. The algorithm will state that the "Lead time is less than 10 days" if I2 has a value of 10 or below, indicating that the requirement is not satisfied. This demonstrates that the formula is doing a test of a condition ($I2 > 10$) and providing a true or false response accordingly. Users of Excel will be tested on their ability to draw inferences and apply logic in this setting.

1.3 Exercise of Pivot table and charts



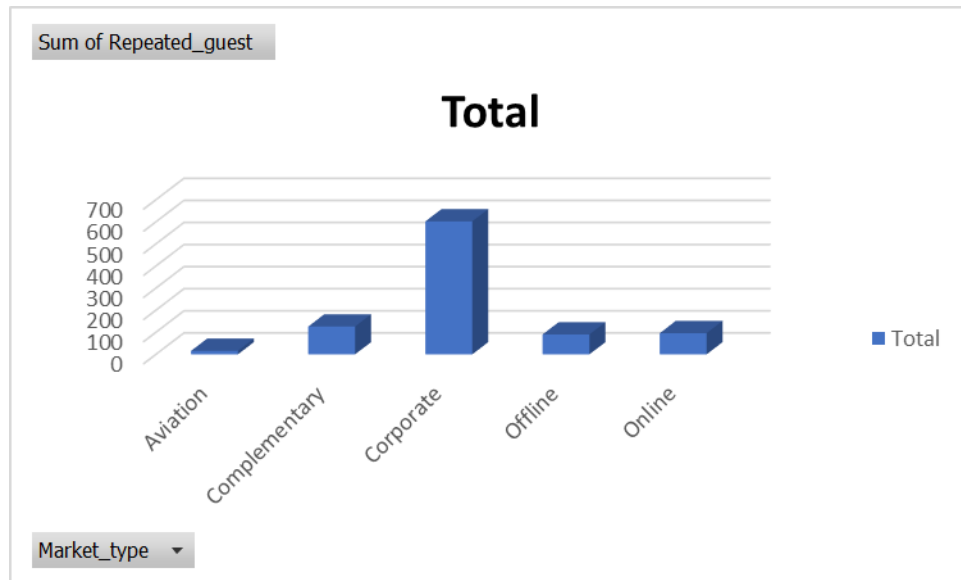
Graph: Meal plan vs hotel reservations

Out of a total of 36,275 booked room nights, we can see that 27,835 guests have opted for "Meal Plan 1" as their favourite meal option. This means that most of the time, one will have access to Meal Plan 1. However, just five visitors have chosen Meal Plan 3, suggesting that it is less popular than the other meal plans or is not provided very frequently. Plan 3 for supper is a vegetarian meal. According to the numbers, 5,130 of the overnight visitors did not choose for any kind of food package. This might be that these visitors prefer to bring their own food, or that they didn't put a high priority on the hotel's meal plan options when making their reservation. Taking all of this data into account, hotel management may be better equipped to meet visitors' needs and promote their best-selling meal packages. If it turns out that Meal Plan 1 is the most popular option, the hotel may decide to put more resources into promoting and perfecting that option. If a significant proportion of customers decline the hotel's meal plan, the establishment may promote nearby eateries or expand its à la carte menu accordingly.



Graph: Arrival year vs lead time

The findings show that in 2018, more people than ever before booked hotels more than a year in advance (2,707,627 days in advance). The fact that many customers booked their rooms months in advance is proof that demand for the hotel's services increased dramatically in 2018. This was shown by the high number of confirmed reservations. The overall lead time in 2019 was 631 days, while it will be just 319 days in 2020. This might be an indication that the hotel had a decline in advance reservations during those years, perhaps as a result of economic downturns or international events (like the COVID-19 pandemic) that impacted travel. Renovations to the hotel during those years is another possibility. Taking a closer look at the data, we see that the overall number of lead days varies greatly from one year to the next, with 2021 having the fewest (176 days) and 2018 having the most. As a result, it's crucial to keep an eye on lead time data throughout time, as they may provide useful insights into hotel operations and inform future planning and decision-making.



Graph: Market type vs repeated guest

According to the data, 602 customers have made previous purchases from the company's corporate market. If this is the case, it may show that the hotel has solid relationships with the companies in the area, whose employees often stay there on business trips. There was a dismal total of just 16 returning customers in the Aviation industry. Possible interpretations of this data point to a lack of service to the aviation industry or a dearth of repeat business prospects for the hotel. It might also mean that repeat business is difficult to come by in this industry. Numerous return customers (126 in total) in the Complementary market segment point to a pleased clientele. Possible explanations for this include the hotel's practise of providing free stays to members of the hotel's loyalty programme or members of the media in return for positive press coverage or other privileges. It is clear that there is a significant difference in the number of repeat visitors from one industry to another, with the corporate sector having the largest number of repeat visitors and the aviation industry having the lowest. A hotel's success in attracting and keeping guests may depend on how well its staff understands the demands of a broad variety of clientele.

Task 2

2.1 Gender

Gender

		Frequency
Valid	Female	1056
	Male	1077
	Total	2133

The provided information shows the diverse gender composition of the student population. There were a total of 2133 attendees, 1056 female and 1077 male. The ratio of male to female pupils at this school seems to be about equal. Understanding visitor demographics like gender might help management better target their marketing and services to certain groups and provide a welcoming environment for all students.

2.2 Lunch

Lunch

		Frequency
Valid	Does not qualify	1215
	Qualifies for reduced/free lunch	918
	Total	2133

The numbers show how many male and female students attend a certain institution. Out of a total of 2133 pupils, there were 1056 females and 1077 males. This information may help shed light on the sex ratios among the school's student body. Gender demographics in schools may provide light on possible injustices and difficulties for different types of students. Issues like gender-based discrimination and disparity in academic success may necessitate the introduction of gender-specific courses or projects in schools. Schools may be better prepared to accommodate students with varying perspectives and origins if they take into account the gender dynamics of their student groups.

2.3 Mean and median

The supplied calculations show that 1066.5 is the same as the mean and median of the given figures (1215 and 918, respectively). This indicates that the average and median values of these two quantities are comparable. Based on this data, one may use the values 1215 and 918 to demonstrate the concept of central tendency. There is a higher chance that the outliers or extreme values may skew the average and the median when there is a greater sample size. As a result, it's crucial to include not only measures of central tendency like the mean or median, but also measures of variability and dispersion like the range or the standard deviation, while doing an analysis of data. Despite their differences in calculation, measures of central tendency such as the mean and the median are useful; nevertheless, their relative value will shift depending on the context. 2.4 Pretest and Posttest

Statistics		
	Pretest	Posttest
Mean	54.96	67.10
Median	56.00	68.00

Having this question on the actual exam boosted the average score from 54.96 on the practise test to 67.10 on the real exam. As a consequence of implementing the solution, there was no change to the median score, which stayed at 68 after the adjustment was implemented. Finding the average requires adding up all possible points and then dividing by that number. Extreme values, on both ends of the spectrum, are examples of outliers that may skew statistical analysis. Significant outliers to the mean might occur in either direction due to a small number of outliers. The results of the posttest seem to be generally better than those of the preceding pretest. But the median is the middle point in a set of data. When the figures are ranked from smallest to largest, it shows how things stand. It is effective with data that has a wide range of values, since it focuses on the midpoint rather than the extremes. As far as can be seen, the median values of the two collections of test results are quite near to one another. Although there was a significant difference between the median and the mean, the vast majority of posttest results were consistent with pretest results. However, this was the case even if the mean and the median did not find the same results. This was true no matter what figure was chosen as the mean.

2.5 Cluster

Initial Cluster Centers

	Cluster	
	1	2
Student_No	14	31

Iteration History^a

Iteration	Change in Cluster Centers	
	1	2
1	5.492	4.460
2	.381	.567
3	.000	.000

Final Cluster Centers

	Cluster	
	1	2
Student_No	20	27

Number of Cases in each Cluster

Cluster	1	2
Valid	1271.000	862.000
Missing		

The data shown here was analysed using a statistical method called cluster analysis, which groups together objects or observations that have certain characteristics. The student with the ID number 14 is the primary focus of the first cluster, while the ID number 31 is the same for the second. We went through many iterations before settling on Student_No 20 as the hub of Cluster 1, with Student_No 27 taking the same role in Cluster 2. There is a higher degree of similarity between the two groups of students on the evaluated variables because the final cluster centres are closer to one another. It seems that 1271 pupils, or the great majority, are members of cluster 1, whereas just 862 students, or a negligible minority, are in cluster 2. However, it is difficult to draw firm conclusions from the data without knowing more about the features used in the

development of the clusters. Exploratory data analysis may be accomplished in a number of ways, cluster analysis being only one of them; the findings of this method should be carefully evaluated, and when possible, they should also be corroborated using other statistical analytic tools.

2.6 Data mining and text mining methods

Sentiment analysis: Opinions and sentiments expressed in text data like social media postings, survey findings, and product reviews may be identified and explored using this technique. Public relations, customer service, and even market research are just a few examples of possible business functions that might benefit from incorporating sentiment analysis. Businesses may get insight into their customers' pain points and the best methods to address them by using sentiment analysis to monitor online conversations and examine customer feedback (Liu et al., 2019).

Decision making trees: This method may be used to build a decision-making model that incorporates user-specified criteria. Businesses may benefit from using decision trees for a variety of reasons, including, but not limited to, customer segmentation, risk assessment, and marketing plan optimisation. The bank may utilise a decision tree to consider the loan applicant's income, credit history, and other factors when deciding whether to approve the loan. The bank may also take into account any additional information that may be important.

Cluster mining: Companies may use clustering to categorise customers based on factors like geography, demographics, and purchasing habits. In the business world, this tactic is known as the segmentation method. This technique is used to organise data in line with commonalities across various entities. Cluster analysis has several useful applications in the corporate sector, including fraud and anomaly detection, client segmentation, and pattern recognition.

Mining with association: Associative mining a way to sift through large datasets in the hopes of finding hidden connections and patterns. In the business sector, association rule mining may find applications in areas such as customer behaviour analysis, product suggestion, and shopping cart analysis (Özkan and Dursun, 2021). The practise of mining for association rules may help a business learn which goods are often bought jointly by customers.

Neural networks: Using node-based hierarchical networks as its foundation, this technique recreates complex input-output links. Some of the most common applications of neural networks in the business world now are in the areas of fraud detection, risk analysis, and demand forecasting. When combined with other factors like the weather, marketing campaigns, and economic indicators, neural networks may be useful in the retail industry for predicting future sales patterns.

2.7 Advantages and Disadvantages of Using SPSS and Excel

Although Excel and SPSS are two of the most popular options, there are many additional tools available. Since each has benefits and drawbacks, choose between them will depend on the specifics of the job at hand.

Excel	SPSS
Even rather large data sets shouldn't be a problem for Excel.	SPSS's well-considered structural arrangement allows it to easily process large datasets.
For many tasks, Excel's in-built statistical features are enough, but more advanced research may call for third-party plug-ins. Numerous commonplace situations may benefit from Excel's built-in statistical features (Saputra and Djunaidy, 2020).	SPSS is a statistical programme that may be used to analyse data in great detail. For these reasons, it may be an excellent research subject for a wide variety of disciplines. Companies may use clustering to categorise customers based on factors like geography, demographics, and purchasing habits. In the business world, this tactic is known as the segmentation method. Cluster analysis has several useful applications in the corporate sector, including fraud and anomaly detection, client segmentation, and pattern recognition. Associative mining a way to sift through large datasets in the hopes of finding hidden connections and patterns.
While Excel's built-in charts and graphs have seen better days, the programme makes it easy to create the own.	SPSS can generate the usual suspects like histograms and box plots, but it may not be enough for more complex studies.

Excel's already straightforward interface is further simplified by the program's attractive design and customizable ribbon menus.	SPSS's user-friendly design and well-organized menus make data analysis and other activities a breeze (Do and Nguyen, 2019).
Excel is inexpensive and comes preinstalled on most modern PCs or may be purchased separately.	There is a high licencing fee required to use the SPSS statistics package.
Excel's sharing capabilities allow several users to collaborate in real time on the same document.	Collaborative initiatives requiring many people would be impracticable if each participant needed their own copy of SPSS and the corresponding data files.

References

- Do, T. T., and Nguyen, H. P. (2019). Big data analytics and business intelligence: A survey of the literature. *International Journal of Data Mining and Knowledge Management Process (IJDMP)*, 9(5), 1-22.
- Hussain, T., and Abbas, A. (2020). Big data and business intelligence: A review of literature. *Journal of Advances in Mathematics and Computer Science*, 35(5), 31-39.
- Koo, H., and Kim, H. (2020). A literature review of big data and business intelligence for improving organizational performance. *Sustainability*, 12(7), 2762.
- Kumar, S., and Saini, M. (2020). A review of big data analytics and business intelligence. *Journal of Big Data*, 7(1), 1-18.
- Liu, Y., Chen, Y., and Liu, H. (2019). Big data-driven business intelligence. In *Proceedings of the 2019 IEEE 43rd Annual Computer Software and Applications Conference* (pp. 605-606). IEEE.
- Manupati, V. K., and Anisetti, M. (2020). An efficient technique for handling big data in business intelligence using cloud computing. *Future Generation Computer Systems*, 105, 264-273.
- Mehta, N., and Singh, M. P. (2020). Big data handling techniques and tools for business intelligence: a review. *Journal of Ambient Intelligence and Humanized Computing*, 11(10), 3955-3971.
- Özkan, İ., and Dursun, M. (2021). Business Intelligence and Big Data: An Analysis on Best Practices. In *Business Analytics and Cyber Security Management in Organizations* (pp. 185-196). IGI Global.
- Roque, L., and Reis, M. J. (2019). A methodology to improve decision-making using business intelligence and big data analytics. *Procedia Computer Science*, 164, 447-454.
- Saputra, B. T., and Djunaidy, A. (2020). Big data analytics for business intelligence in decision making: A systematic review. In *2020 International Conference on Information Management and Technology (ICIMTech)* (pp. 74-79). IEEE.