FACULTY OF COMPUTING

SESSION: 2020/2021 SEMESTER 2

BCM3253

DATA ANALYTICS AND VISUALIZATION

ASSIGNMENT

FADHLIN SAKINA BINTI KHALID

CB20168

SECTION : 01B

1 INTRODUCTION

1.1 What is Data Analytics?

Data analytics is the science of analysing raw data to conclude the given data [1]. The techniques and the processes of data analytics are automated into the mechanical processes and algorithms. The mechanical processes and algorithms will work on the raw data to help in human use. The techniques of data analytics can uncover the trends and metrics that might rather be lost within the mass of data. This data can be used to maximize processes in order to increase the efficiency of certain business or even a system.

1.2 What is Technique of Data Analytics?

There are many techniques for data analytics and the technique that will be discussed here is **cluster analysis**. Cluster analysis is also called classification analysis or numerical taxonomy. In cluster analysis also, there is no advance information about the group or cluster membership for any of the objects. For example, people who provide or sell insurances use cluster analysis to detect false insurance claims. Cluster analysis have two approaches or methods which are non-hierarchal (partitioning) and hierarchal. Hierarchal clustering is unsupervised clustering technique that requires creating clusters in a predefined order. The clusters are ordered in a top to bottom manner. Same or similar clusters are grouped together and are arranged in a hierarchical manner. This method link pairs of clusters until every data object is included in the hierarchy. Meanwhile, the non-hierarchal method is the formation of new clusters by merging or splitting the clusters and the relationship between clusters in undetermined. One of example for non-hierarchal method is k-means algorithm [2]. In term of reliability, non-hierarchal is more reliable than hierarchical clustering. Cluster analysis is a crucial task of data mining exploration and a standard technique of statistical data processing. It is used in many fields such as image analysis, machine learning, data compression and many more [3]. Figure 1 shows the difference of non-hierarchal and hierarchal methods.

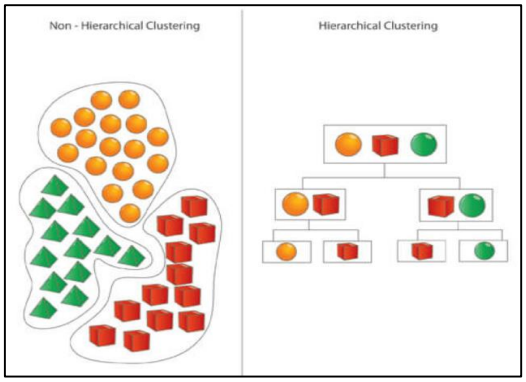

Figure 1

1.3 What is Data Visualization?

Data Visualization is representing data or information in term of graphical which can make the information that going to present easier to understand. The representation of data can be in a graph, chart or other visual format. Nowadays, there are a lot of demands and needs to display huge amount of data in a way that is understandable as organizations generate data every single day [10]. The design, development and application of computer generated graphical representation of the data are very important in data visualization. This is because, a good data visualization will lead to an effective data representation. Then, it will help decision makers to discover trends, patterns, comprehend the information and form an opinion from the data.

1.4 What is Technique of Data Visualization?

There are few techniques of data visualization that we need to know. Some of examples of the techniques are knowing your audience, set your goals, choose the right charts, taking advantage of colour theory and many more [11]. But in this report the technique that going to be discussed is choosing the right type of chart for your data. In data visualization, the technique to choose the right type of chart is important because data is only valuable if we know how to visualize it and give context. There are many charts that will help you in visualize given data. For examples of the chart are line chart, area chart, histogram, Sankey diagram and many more charts. Figure 2 shows the examples of charts. Before choosing the chart, we should consider the type of data that we going to analyse as different data have different way to analyse. For instance, if the data is unstructured data, one of the way to analyse it is by using word cloud.

Figure 2

2 ILLUSTRATION ON HOW EACH TECHNIQUE WORKS USING PROCESS FLOW (ALGORITHM)

2.1 Data Analytics Technique – Cluster Analysis

Figure 3 shows the flowchart on how cluster analysis work. The first step of cluster analysis is to decide on the clustering variables and in this step it is should relate with the given data. The second step is to decide on the clustering procedure. This step can be done by reviewing the data set and the theoretical grounding for the research data. It can be either hierarchical method if the data set is small or do not have theoretical grounding for knowing the number of clusters and it can non-hierarchical if the data set is larger or we already have theoretical grounding for knowing the number of clusters. The next step is depending on what procedure that we chose. If hierarchical method is chosen then the next step are going to be selecting a measure of similarity or dissimilarity and choose a clustering algorithm based on data behaviour. Then decide on the number of clusters by evaluating dendogramsdendrograms (hierarchal algorithms only) and other plot behaviour. The last step of hierarchal methods is to validate and interpret the cluster solution. For non-hierarchal methods, a few steps is different. It skips the two steps from hierarchal methods and immediately goes to the deciding on the number of clusters and the last step which is validating and interpreting the cluster solution [4].

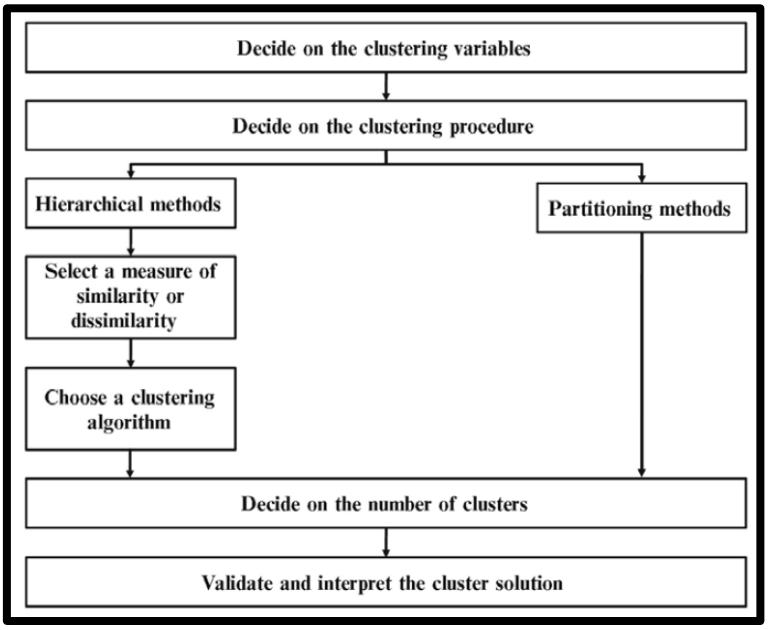

Figure 3

2.2 Data Visualization Technique – Choose the right type of chart

Figure 4 shows the steps on how we can choose the right type for our data. The first step is to analyse given data carefully. Then from the data, we should make a conclusion either it is a structured data or unstructured data. This is important as both structured data and unstructured data have different types of chart to visualize. Figure 5 show the best way to think in order to choose chart. After that, we need to choose the best type chart that suitable with our data in order to have an accurate information from it. The last step is of course to visualize it. There are few tools that we can use to visualize data and one of it is Tableu.

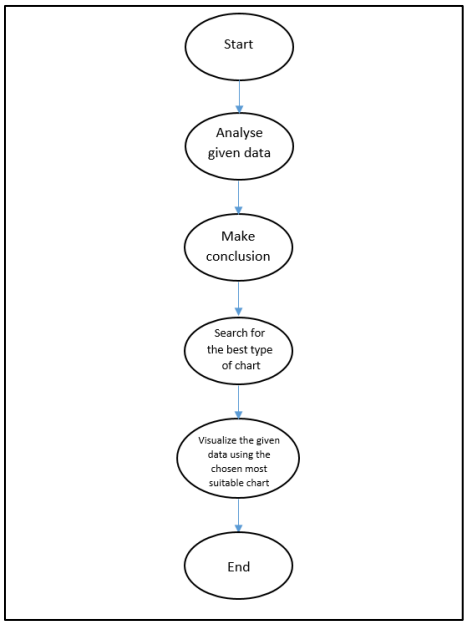

Figure 4

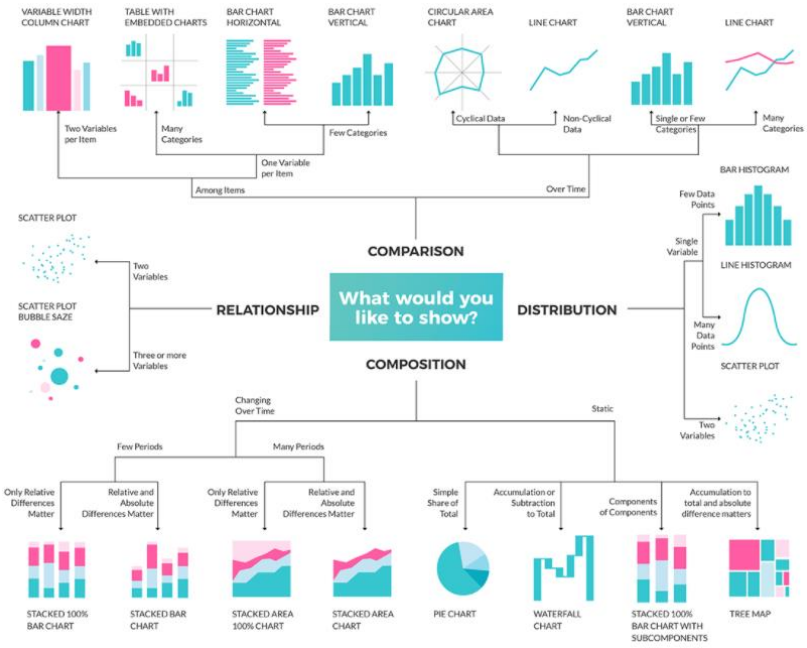

Figure 5

3 ADVANTAGES AND LIMITATIONS FOR EACH TECHNIQUE

3.1 Technique of Data Analytics

3.1.1 Advantages

Cluster analysis have many advantages when used it. The first advantages that could get from it is can cut down classification time. In this case, the advantages of cluster analysis in machine learning is going to be talk about. The first advantages that we could get by using cluster analysis in machine learning is can cut down classification time. If we got a small datasets then it is workable to use manual annotation and organization. However, when a bigger datasets is used, manually dividing and annotating data can be a hard process. Depended on the algorithm that one used to clustering, it can cut down annotation and classification time as it is less interested in specific outcomes and more concerned with the categorization itself. For example, speech recognition algorithms produce millions of data points and it would take hundreds of hours to fully annotate. Thus, a lot of time will waste. But by using clustering algorithms, it can reduce the total work time and give faster answers. The second advantage of cluster analysis is can help in searching for anomalies or peculiarity in data. Algorithms like density-based spatial clustering of applications with noise (DBSCAN) can help to find clusters that are closely positioned and mark outliers in datasets. The last advantage is easy to understand [5]. When used cluster analysis as technique in data analytics, it can help to identify classes of the data given. For instance, there might be some confusion on how many main subsets are there in the datasets.

3.1.2 Limitations

Cluster analysis also have its limitations and it might play a big role either this technique can be used for given dataset. The first limitation of cluster analysis is getting different results as there are a lot of methods for clustering [5]. This problem might happened because of the differences criterion for merging cluster including cases. It is crucial to choose which method suit the best for an information. The second limitation is that while using hierarchical approach, missing data could be a problem. This is because many hierarchical software with values that are missing in the data. The last limitation is when there are too many data types. With so many data types, to compute a distance matrix can be difficult. The reason is because, there is no straightforward formula which can compute a distance where variables are both numeric and qualitative.

3.2 Technique of Data Visualization

3.2.1 Advantages

There are a lot of advantages from having to choose the right type of chart as data visualization techniques. The first advantage is can avoid inaccurate data translation later on. By choosing the most suitable chart for our data, there will be no risk if we might visualize the data wrongly. As we have taken the action of selecting the best chart for our data. The second advantage is client will be more understand of the data. For instance, if we choose line graphs to visualize a change over a set of time periods, client will be understand it easily as they can see the line graph decreasing and increasing or stable. The last advantage is can attract people visually. If we choose a visually represent chart, people would be interested in seeing the chart that we choose. For instance, if we had word cloud as the chart to present the data for survey, people will attract to know the detail about the keyword that have the largest word. Therefore, they will be more focus on the presentation.

3.2.2 Limitations

Technique of data visualization which is choose the right chart also have its own limitations. The first limitation is there are too many options of choosing the right type. Because there are too many formats, we might make a wrong decision in choosing the chart. This can lead to the wrong information at the end of it. Too many choices charts and graphs can come across to be very confusing, lack clarity or irrelevant. The second limitation is that could be conflicted situation with client to choose which chart to use. There might be some problem in choosing the right chart techniques as our client could have different options. So we need carefully choose the right chart by taking both opinion in mind in order to produce a great data visualization. The last limitation is when the charts could have oversimplify the data. Oversimplify data can lead to misleading view of the data. Thus, it is important to analyse our data first before choose the chart that we going to use to represent the data.

SUMMARY

Both data analytics and data visualization have their own impacts towards industry and academic. By choosing the right types of chart, in term of academic the impact is we can easily understand a data. Compared to data in form of essay, data in right chart presentation can make students understand information easily as it do not required much reading from it. Charts are also important in industry as simple table might won't adequately demonstrate important relationships or patterns between data points. Choose the right type of chart also can help to reduce risk of getting wrong information. Cluster analysis have the impact to make data analytics easier to understand. This is because there are many algorithm that can be used in academic and industry factors.

REFERENCES

[1] Jake F. (2020, July 1) *Data Analytics.* Retrieved from https://www.investopedia.com/terms/d/data-analytics.asp

[2] Kapilsparshi (2020, July 16) *Difference between Hierarchical and Non Hierarchical Clustering.* Retrieved from https://www.geeksforgeeks.org/difference-between-hierarchical-and-non-hierarchical-clustering/

[3] John M. (2019, December 22) *What is the purpose of cluster analysis?* Retrieved from https://www.vproexpert.com/sccm_vpro/module_03/module_03.html

[4] Shweta B. (2011, August 29) *Limitation of Cluster Analysis.* Retrieved from http://sibmb-ba.blogspot.com/2011/08/limitation-of-cluster-analysis.html

[5] Explorium Data Science T. (2020, February 3) Retrieved from https://www.explorium.ai/blog/clustering-when-you-should-use-it-and-avoid-it/

[6] Ehlert, Katherine M. Faber, Courtney June Kennedy, Marian S. Benson, Lisa (2017, June) *Utilizing cluster analysis of close-ended survey responses to select participants for qualitative data collection.* Retrieved from https://www.researchgate.net/figure/Flow-chart-for-performing-a-cluster-analysis-This-chart-was-modified-based-on-earlier_fig1_323343064

[10] Mallon, Melissa (2015) Public Services Quarterly. *Data Visualization. 11*(3), 11-12, https://www.researchgate.net/publication/311597028_DATA_VISUALIZATION#:~:text=Data%20visualization%20involves%20presenting%20data,of%20presenting%20large%2C%20complex%20information.

[11] Sandra D. (2018, October 5) *10 Essential Data Visualization Techniques, Concepts & Methods To Improve Your Business – Fast.* Retrieved from https://www.datapine.com/blog/data-visualization-techniques-concepts-and-methods/