Title: “A report on R and Power BI discussing practical work in both pieces of software, the impacts of Big Data in Organisations and how a Power BI solution can be used in business”.

Exeter College Hele Road

Report assignment

JACK BURKEY

2023

**Contents**

[Task 1: Creating a bar chart that shows fires burned greater than 1,000 acres between 2010 and 2016 from the San Diego Dataset …….. page 2](#_Toc100689073)

[Task 1.1: Creating a histogram that pipes the data frame,uses a bin size of 500 and shows only fires burned above 1,000 acres in the time range of the San Diego dataset …… page 5](#_Toc100689074)

[Task 1.2: A box plot that shows fires burned between 1000 and 5000 acres, grouped by organisation …… page 7](#_Toc100689075)

[Task 2: Creating a Linear model using the Using R package alongside statistical tests such as the summary() and lm() function: ……… page 9](#_Toc100689077)

Task 3: Managing the Building Data Model relationships in Power BI: ….. page 12

[Task 4:Editing relationships in Power BI: …… page](#_Toc100689080) 19

[Task 5:](#_Toc100689081) Creating the Visualizing Data dashboard: …… page 21

[Task 6.1: What are the different types of Big Data challenges faced by organisations historicaly and currently? ….……. page 23](#_Toc100689083)

[Task 6.2:What are the different types of Big Data Analytics methods utlisied by organisations historically and currently? …………. page](#_Toc100689083) 26

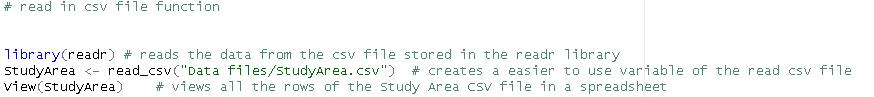
[Task 7: AGGORA Power BI Solution business plan:](#_Toc100689085) …….. page 29

***Task 1:* Creating a bar chart that shows fires burned greater than 1,000 acres between 2010 and 2016 from the San Diego dataset:**

**Creating the bar chart:**

1. Firstly, before I started any work on any part of the assignment, I created a plan for this particular task, as well as the assignment as a whole. This set out a structure for the task, which incorporated the course’s handbook, the assignment brief and my own intuition.
2. Next, I started to follow the assignment brief and created a function that read the CSV ( comma separated value) **(D. Johnson, 2022**) **(1)** file in R Studio which was given as part of the data file folder provided by the course. Shown below are screenshots of the code with comments about the function of each line and a snapshot of the study area table:

**Figure 1: a screenshot that shows the code for the CSV file inside R studio**



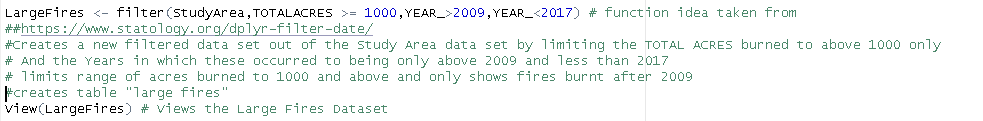
**Figure 2: a screenshot that shows a snapshot of the table of the Study area CSV file inside R Studio**

A picture containing text, menu, number, document

Description automatically generated

1. After this, I created a variable called “Large fires” by filtering the Study area CSV file to only include fires larger than 1000 acres (the filter function was taken from **(SFU, 2023) (2))**. Screenshots of this stage are shown below:

**Figure 3: A screenshot of the code used to filter the Study Area dataset into the Large Fires Dataset:**



(Statology , 2022)(3)(R bloggers , 2022)(4)

(Reference 3 refers to the filter function shown by the URL in code)

(Reference 4 refers to view function in code screenshot)

**A picture containing text, menu, number, screenshot

Description automatically generatedFigure 4: A screenshot of a snapshot of the “large fires” inside R Studio table :**

1. The next step involved converting this large fire dataset into a bar chart. Shown below is a screenshot of the code used to transform this data into a bar chart, alongside the bar chart itself:

A picture containing text, screenshot, font

Description automatically generated**Figure 5: A screenshot of the code used to transform the large fire data into a bar chart**

(A.Bajak , 2017) (5)

(Stack Overflow, 2013) (6)

(Statology , 2021) (7)

(Reference 5 refers to barplot function although some code was taken from my own class notes)

(Reference 6 refers to options Scipen function)

(Reference 7 refers to a website that helped me to discover the Scipen function for changing the axis scale on the barplot, as well as for the xlim and ylim functions)

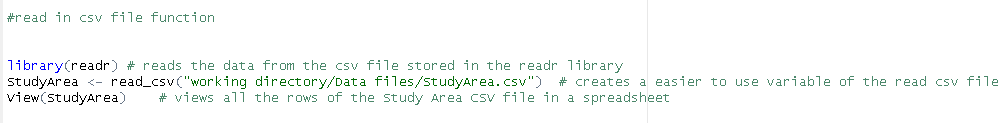
**Figure 6: A bar chart to show: “The acres burned between 2010 and 2016 that are greater than 1,000 acres in the Study Area Dataset”**

A picture containing sketch

Description automatically generated

**Task 1.1: Creating a histogram that pipes the data frame, uses a bin size of 500 and shows only fires burned above 1,000 acres in the time range of the San Diego dataset:**

**Figure 7: Screenshot of the code that uses the same read dataset function as task 1 to create the Study Area dataset:**



**Figure 8: Screenshot of the code that uses the same filter function as task 1 to create the Large Fires dataset:**

A picture containing text, screenshot, font, algebra

Description automatically generated

A screenshot of a computer code

Description automatically generated with low confidence

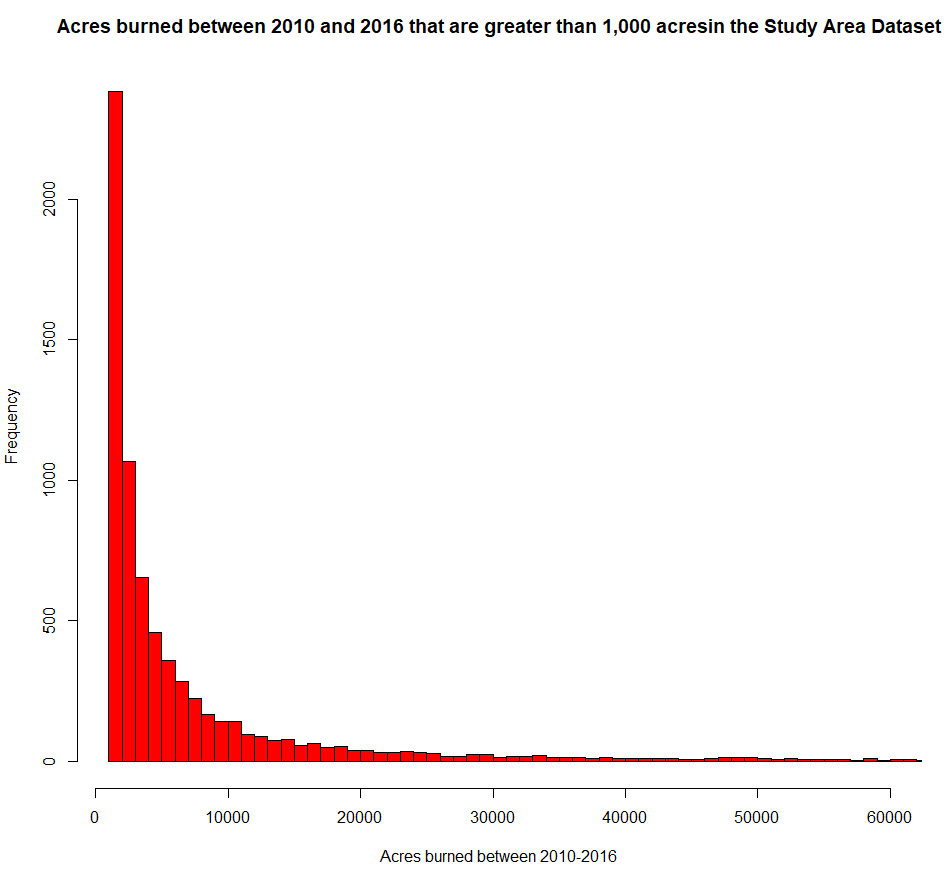
**Figure 9: Screenshot of the Histogram function code:**

(Stack Overflow, 2015) (8)

( Reference 8 refers to the use of xlim, xlab, and ylab/ylim in the code )

**Figure 10: Screenshot of the Histogram created by Histogram function code**

A histogram to show the acres burned between 2010 and 2016 that are greater than 1,000 acres in the Study Area Dataset:



**Task 1.2: A box plot that shows fires burned between 1000 and 5000 acres, grouped by organisation:**

A picture containing text, screenshot, font, line

Description automatically generated**Figure 11 : Screenshot the of read function code:**

A screenshot of a computer code

Description automatically generated with low confidence**Figure 12: Screenshot of the Filter Function and creation of new dataset code:**

(Statology , 2022) (10)

(Statology, 2023) (11)

(Reference 10 refers to the piping operator (%>%) in the Large Fires 2 function)

(Reference 11 refers to the between filter function of TOTALACRE

**Figure 13: screenshot of the boxplot code**

A screenshot of a computer

Description automatically generated with low confidence

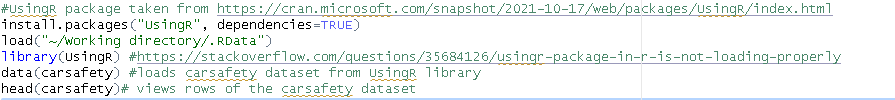
**Figure 14: A box plot to show the total acres burned per record by each different organisation between 1000 and 5000 acres:**

A picture containing diagram, rectangle, square, technical drawing

Description automatically generated

**Task 2: Creating a Linear model using the Using R package alongside statistical tests such as the summary() and lm() function:**

**Figure 15: Screenshot the of view Car safety file function, data function and head function**



(J Verzani, 2022)(12.1)

(Stack Overflow, 2015) (12.2)

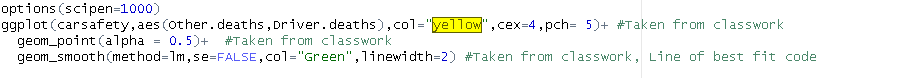
(reference 12.1 refers to finding about how to install the UsingR package)

(reference 12.2 refers to troubleshooting the UsingR code)

A black text on a white background

Description automatically generated with low confidence**Figure 16: Screenshot of the first six rows of Car safety dataset**

**Figure 17: Screenshot of the code for linear model comparing type of car with driver deaths**



(Statology, 2022) (13)

(Reference 13 refers to troubleshooting the aes x,y function in ggplot for the linear model, I created the data frame shown below with the summary and lm functions with the information from this source )

Figure 18: A linear model comparing the driver deaths with other deaths with a line of best fit:

A graph with a green line

Description automatically generated with low confidence

**Figure 19: Screenshot of the code passing the lm and summary function comparing significance levels of Other.Deaths and Driver.Deaths of the carsafety dataset:**

A picture containing text, font, screenshot, line

Description automatically generated

(Rdocumentation , Copyright 2023) (13.1)

(Reference 13.1 refers to the model and summary(model) functions passed)

**Figure 20: Screenshot of the mathematical results including significance codes and standard errors produced by the passing of the lm ,data, model and summary functions:**

A picture containing text, font, screenshot, number

Description automatically generated

**Task 3: Managing the Building Data Model relationships in Power BI:**

**Figure 21: Screenshot of the general table relationships of the Power BI dataset “Building the Data Model”:**

A screenshot of a computer

Description automatically generated with medium confidence

Shown in figure 21 are the default relationships of the data model dataset when loaded into Power BI from local folder. It also shows that there are three, many to one, relationships. Figures 22, 23 and 24 show the specific keys used.

**A screenshot of a computer

Description automatically generatedFigure 22: A screenshot showing the, many to one, relationship between the DimCustomer and FactInternetSales tables**

As seen in Figure 22, there is a, many to one, relationship between the Dim Customer table and the FactInternetSales table, more specifically using the Primary key Customer key and Customer key foreign key.

**Figure 23: A screenshot showing the, many to one, relationship between the Dim Sales territory table and the Fact Internet Sales table:**

A screenshot of a computer

Description automatically generated

Figure 23 shows the, many to one, relationship between the Dim Sales Territory table to FactInternetSales table, more specifically between the Primary key Sales Territory and the Foreign key Sale Territory.

A screenshot of a computer

Description automatically generated**Figure 24: A screenshot showing the, many to one, relationship between the Dim Product and FactInternetSales table**

Figure 24 shows the, many to one, relationship between the DimProduct table to FactInternetSales table, more specifically between the Primary key Product Key and the foreign key Product key.

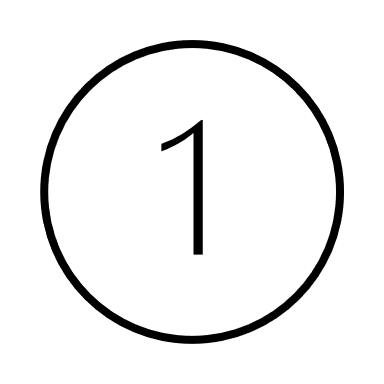
In this dataset, there are no one-to-one relationships shown in the model view between the dataset’s different tables, or in fact any other types of relationships.

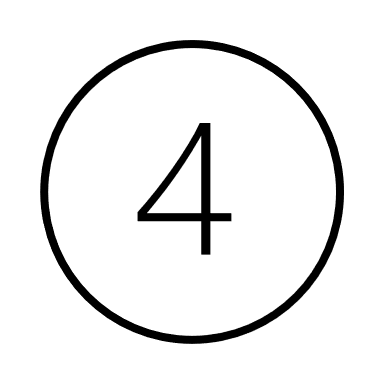
**Manage relationship editor:**

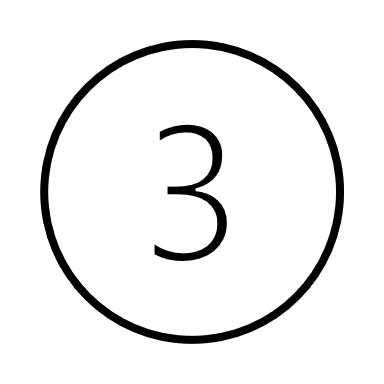
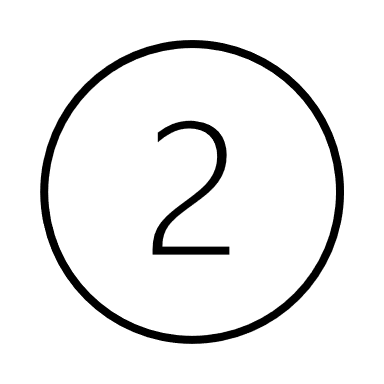
**Figure 25: A annotated screenshot of the manage relationship editor**

A screenshot of a computer

Description automatically generated with medium confidence



z



**1**: Current relationships, where the active status from table column and to table column details are shown in text format.

**2**: New… (relationship) : By clicking this button a popup appears which is shown in figure 26. Here you can select tables and columns that are related, as well as editing cardinality, cross filter and other options to characterise the intended relationship.

**Figure 26: A screenshot of the New… popup**

A screenshot of a computer

Description automatically generated

**3:** Edit… : Similarly to the new relationship button, the edit button brings up a popup, shown in figure 27. Here you have very similar options to the new relationship popup, as the appropriate tables and columns can be selected as well as the cardinality, cross filter direction and other options.

**Figure 27: A screenshot of the Edit… popup:**

A screenshot of a computer

Description automatically generated with medium confidence

**4:** Delete : When clicking the delete button on a selected relationship, a warning message popup appears, shown below in figure 28:

**A screenshot of a computer

Description automatically generatedFigure 28: A screenshot of the Delete popup:**

**Task 4: Editing relationships in Power BI**

Shown below in Figure 29 is a screenshot of the creation of a new relationship between the FactInternetSales table and the DimDate table. This is through the OrderDate Foreign key and FullDateAlternateKey Primary key. This screenshot is a filled out version of figure 26 and therefore has the same options and details.

**Figure 29: A annotated screenshot of the new relationship popup which creates a new relationship between the FactInternetSales and DimDate tables:**

****A screenshot of a computer

Description automatically generated

This new relationship can now be seen in model view, shown below in figure 30:

**Figure 30: A annotated screenshot of the new relationship created in model view in Power BI:**

****A screenshot of a computer

Description automatically generated

Annotations 1-3 continue over figure 29-30.

**1:** Here, the appropriate tables have selected in the dropped down menu, as well as the appropriate columns being selected from the table generated directly.

**2:** The options at the bottom of the page include cardinality which changes the type of relationship (e.g. many to many, one to one etc.), cross filter direction ( single or both), as well as unchecks or checks for an active relationship, referential integrity and a security filter.

**3:** Finally, the, many to one, relationship created can be seen in model view in Power BI.

**Task 5: Creating the Visualizing Data dashboard:**

A selection of the key data visualisation techniques used in the Visualising Data dashboard:

**Figure 31: A table showing the parameters chosen to create a relevant and concise data visualisation for sales and other metrics per territory**

A picture containing text, screenshot, font, number

Description automatically generated

**Figure 32: A graph that shows Internet sales and other metrics in each territory(country)**

A picture containing text, screenshot, colorfulness, font

Description automatically generated

**Figure 33:**

**A graph that shows Internet sales and profit per age of person:**

A picture containing screenshot, plot, text, line

Description automatically generated

**Figure 34: Slicer for filtering data on each visualisation for a specific territory**

A screenshot of a computer

Description automatically generated

**Task 6.1: What are the different types of Big Data(BD) challenges faced by organisations historically and currently?**

Before delving into what challenges Big Data (BD) poses to organisations, a definition of what BD is will help answer the question. BD can be described as the three Vs and “**In terms of data management challenges, due to increasing volume, velocity and variety of data, cannot be solved with traditional databases”** (Amazon Web Service, 2023)(14). So, just by exploring the definition, it can be ascertained that the core of the challenges that BD poses to organisations stem from the three Vs, making it a difficult data source to work with. BD also creates a range of other technological, social, political, and ethical challenges which will be explored in this report.

Firstly, volume is a key challenge faced by many organisations, because without the right storage, whether it be cloud based or physical servers, an entire company’s data could be corrupted or potentially lost. This means that without the right storage parameters put in place to deal with the volume of data that BD creates there may be problems. Volume represents the “Big” in Big Data, and appropriate computing resources are needed to effectively utilise the full potential that BD offers. This is because although **“The size of data that is being processed can be unlimited, the speed of processing operations is constant.”** (Adrian, 2013)(15) So, how an organisation currently manages BD, in Petabytes in many cases, depends solely on its computing infrastructure. In the past, volume has been less of a challenge for organisations as datasets were not created or taken from BD, and also datasets stored and used were a lot smaller over the last 25 years. For instance, in 1999, the total datasets on the Internet amounted to 1.5 exabytes of data and 1 Gigabyte was considered BD. In comparison, in 2019 the entire world’s datasets of data amount to 10-50 zettabytes and is estimated by 2025 to grow to 150-200 zettabytes (Chojecki, 2019) (16). Therefore, for organisations to use the volumes of data that BD produces to their advantage, an organisation must have expensive, powerful, and expansive networks of physical/cloud servers at their disposal.

To show how far data has come in terms of volume here is a range of data units from bits to zettabytes shown below:

* One bit contains a value of 0 or 1.
* Eight bits make a byte.
* Kilobytes (1,000 bytes)
* Megabytes (1000² bytes)
* Gigabytes (1000³ bytes)
* Terabytes (1000⁴ bytes)
* Petabytes (1000⁵ bytes)
* Exabytes (1000⁶ bytes)
* Zettabytes (1000⁷ bytes)

Next, velocity is a key challenge to organisations. Many large organisations and key players in the digital sector such as Google, Apple and Amazon analyse, manage, and run huge velocity BD streams every second. Due to this, no matter how pre-processed and “washed” the datasets are, the velocity at which these BD streams are tapped into by these types of organisations results in the data having a reduced validity of accuracy and also veracity. (S. Yin et al, 2015) (17) Veracity is a subsection of the 2nd V (volume) and can be defined as the accuracy of data and encompasses the following three traits:

1. Precision
2. Trustworthiness
3. Reliability of the data

(Grand Canyon University, 2022) (18)

The final V is variety and is a way BD can be a challenge for organisations. Datasets for scientific, industrial and commercial applications change in size regularly and are becoming critically important to manage as older datasets are compared and combined with new BD data pools. This is because BD comes in many different forms and types, so therefore different accommodations of software, hardware and schemas need to be used to manage the seemingly limitless pool of data BD has. Cloud-related technologies such as Virtualisation, Cloud Computing and Cloud Storage are being used effectively and in a widespread manner in the technological sector. These help reduce physical computing power and employees, but further developments need to be made in complex industries such as smart device manufacturing and power systems which historically are very new BD developments. (S. Yin et al, 2015) (17)

Aside from the three Vs, there are other problems that BD can cause for organisations. For example, when organisations do not limit the type of BD they have access to especially in the public sector when dealing with sensitive information. This has been seen in Mainland China since the 1990’s where BD was and is still used to constantly monitor its own citizens for any wrongdoing where all their personal information is in Government databases. More recently, face recognition software was used in Cardiff to monitor fans at a Beyond concert in alliance with datasets to potentially reveal people with “priority offences”. ( Chris V et al, 2023) (19). This of use of BD such as by South Wales Police and the CCP is politically and socially controversial in many lawmakers and public eyes, and although these occurrences are not the first use of BD to be used in a sceptical way, BD has and is a problem for organisations if they use BD without referring to GDPR and other legal documentation.

Another problem with BD usage by organisations, is the security of large datasets and data streams. There are multiple ways a dataset could be compromised, especially since most big data tools are not designed with security in mind (G Maayan, 2020) (20) and so are very prone to many attacks such as DDOS (Direct Denial Of Service), Malware and Trojan Horses. Furthermore, these problems can occur within live data streams, the Cloud as well as any type of organisation datasets. However, BD Security works to keep all data secure for organisations with the main priority to keep out unauthorised users and intrusions with firewalls, strong user authentication and Intrusion Protection Systems (ISP’s) (E Crockett , 2023) (21). This is very general but in comparison to historic data management, the range of security techniques for BD is much wider ranging and effective due to much stronger cryptographic encryption methods and the techniques mentioned earlier.

In conclusion, It can be said that although some historical BD challenges to organisations such as better security of data and cloud technology have been resolved, there are still many problems to organisations in the present day. Examples of these are the immense pressure BD has on varying organisations computing systems, the commonplace unethical usage of BD by trusted organisations such as governments and emergency services and many others.

**Task 6.2: What are the different types of Big Data Analytics(BDA) methods utilised by organisations historically and currently?**

BDA is paramount for organisations of varying sizes and structures to use so that they can utilise BD to it’s fullest extent. Big Data Analytics can be defined as the “use of advanced analytic techniques against very large, diverse big data sets” (IBM, Copyright 2023) (22). Although BDA has been in usage since the early 1990s as claimed by certain computer scientists (G. Firican, Copuright 2023) (23), it’s usage has only been seen since the early 2000’s when “Big Data” was exceeding 1 Gigabytes per dataset and datasets were needed to be analysed in a much more complex and sophisticated way, such as through the three V’s and further which were discussed earlier in the report.

One way that historical data is used in BDA, is through predictive analysis. For example, by using datasets from the past, future trends in consumer behaviour and trends can be predicted by using statistical methods to outline the historical datasets values of mean spend per person, the mode type of clothing bought each year and so on. (A Mosavi et al, 2013) (25).This kind of BDA method has been used since the early 2000’s in digital form, but this kind of business analytics across big data sets has been carried out from the 1950’s onwards in spreadsheet form to “uncover insights and trends” (R. Nadikattu, 2020) (28). Furthermore, through predictive analysis of historical datasets, organisations can be at a competitive advantage over other businesses in sectors such as retail, hospitality and many work sectors. Likewise, In a survey conducted by Ventana research in 2013, 67% of businesses aimed to use predictive analytics to create more strategic marketing campaigns ,as well as 68% of these businesses sighting a competitive advantage as the prime benefit of predictive analysis. (S.Millard , 2013) (26) (J.Zakir et al , 2015) (27). Another way BDA has been used in the past is shown to an extent through the articles published by different organisations since 2010. Big Data Analytics in the mainstream has been a relatively recent development, and so publications on the topic have been relatively limited up until 2010 when organisations started to use big data and analyse it in denser quantities. If you were to query the results of “big data analytics” in an Academic research website such as Google Scholar or any Academic institution’s online library such as Plymouth Primo, you would find the results from 2010 to the present day would follow a steep curve line graph model. Shown below is a diagram of what an academic paper’s query found out with the keywords “big data analytics”:

**Figure 1: A diagram to show the growth of Big Data Analytics publications published by different organisations from 2010-2018**

A picture containing text, plot, line, diagram

Description automatically generated

Source for diagram and 2010 paragraph: (P. Mikalef et al, Date not given) (29)

Indirectly, this graph shows that organisations have been carrying research and using Big Data Analytics in Big Data in large volumes only recently, but due to exponential growth in this field, BDA has soared in use cases by organisations in the last 10 years.

Following on from this, in modern times, there are many BDA methods used by organisations today. One such example was the use of BDA during the COVID-19 Pandemic from 2019-2021, where due to the high risk and first time that modern civilisation had seen a hazardous pandemic before, policy makers, governmental figures and many organisations collected and analysed huge swathes of BD and analysed it through BDA. Leading to 26 million cases of COVID, the Pandemic was ultimately ended by the publics usage of smart devices (Worldometer, Accessed 2020 by Author) (30). This was because before the smart phone age of the 2010s and onwards, there was not such a global and untapped wealth of public data available for governments and organisations to monitor the pandemic at every step and in every country. Furthermore, to this day, organisations continue to data mine digital mobile devices for a vast amount of unstructured and structured data for decision-making by governments as well as business campaigns among many other reasons (J. Sheng et al , 2020) (31). Another modern BDA method used by organisations in the present day is the use of BDA by the retail industry globally. A survey analysed by the Journal of Retailing and Consumer Services in 2022 revealed varied reasons why BDA was used in retail in developed and developing countries. For example, in developed countries such as the UK, security concerns, external support, top management support and rational decision making culture were the top reasons for BDA adoption. In contrast, in developing countries such as the United Arab Emirates, BDA adoption was primarily done because of organisations in the retail industry wanting an advantage due to competition intensity and firm size in relation to other retail businesses. Therefore, this shows that in retail and indeed all other BDA use cases, a “one-size-fits-all” analogy is not correct in showing the use of BDA by organisations in sectors such as retail (M. Youssef et al, 2022) (32).

In conclusion, There are varied way organisations adopt methodologies to utilise Big Data Analytics within their systems. This differs between developing countries and developed countries, between global and local areas, as well as in different sectors and big data information subjects. In comparison to the past, modern day BDA deals with much larger datasets and with new datasets coming online each day in a variety of sectors, organisations need to keep developing methodologies to utilise and manage BD for their use cases and to stay competitive in their respective industries.

**Task 7: AGGORA Power BI business solution audit and analysis:**

AGGORA Group is a successful company that specialises in catering equipment solutions with a company turnover of approximately £25 million a year. Due to the business being active for over 10 years as well as other factors, there are some improvements that the company wants for its operations which I will highlight in this business audit through the planned use of Power BI. However, before I delve into further facts and figures about the AGGORA Group and a potential BI solution, there are five key points that the group should implement to use the BI solution successfully:

1. There should be a clear and well-defined scope for these improvements by the Group so that each relevant part of the business is getting the required benefits it needs from the BI solution.
2. Data Governance should be thought about from the beginning. This is because no matter how good the BI solution is, if it is not safely secured from both internal and external threats by user hierarchies, firewalls and such, the Group can’t depend on the BI solution and wouldn’t be able to become as much as a data driven culture as it can be.
3. The AGGORA group should have a reliable and trustworthy implementation body or team to make sure the solution is well incorporated into the business and that it can be used effectively
4. The BI solution and further solutions should be on a business roadmap so that AGGORA Group can make business goals and grow as a business
5. Finally, there should be a readiness assessment for this solution. If the AGGORA group does not have the capabilities or manpower to effectively use the BI solution , then soe rethinking will be required

(R. Vanvleet, 2020) (33)

Once these prerequisites for the BI solution are put in place, the following can take place:

Audit methodology for the implementation of the necessary Power BI solution for AGGORA Group:

1. Firstly, the solution needs to be available for all 150 members of the AGGORA Group but primarily needs to be managed by the data and security team, management and developer teams.
2. The first improvement to the current Power BI dashboard that the AGGORA Group wants is to have a data visualisation technique to show “Analysing Client Key Performance Indicators” which include data on whether Service Level Agreements are met, how satisfied customers are and so on. Due to the yes or no nature of this data, the most effective and simplistic way to show this data would be to use a pie chart/donut chart as these show clear sections of percentages of the data for each group e.g. are SLA’s met (yes/no), how satisfied are customers (very satisfied, satisfied, unsure, unsatisfied, very unsatisfied).
3. The next improvement listed is for a data visualisation that shows the performance and productivity of the Group’s engineers. The best visualisation technique for this is the stacked area chart. This is because as well as again the chart being simplistic to understand and gather meaningful conclusions from, it can effectively compare and show the performance and productivity of the groups engineers over time from the Group’s 10 year history.
4. Thirdly, the AGGORA Group would like a data visualisation for showing financial figures on their current dashboard. Here, a simple table of figures would be appropriate so that the profit margins, annual turnover and other figures could be displayed in a effective, easy to understand yet simplistic way.
5. Lastly, and the most complex improvement to undertake, is AGGORA group want a asset management element to their dashboard which shows large volumes of data in a concise way and includes measures such as service history and overall spend. To show these results effectively and simplistically on the dashboard, a slicer, a gauge and any other appropriate charts can be used as long as the slicer condenses the data using a relevant field for example per Country or Sales Region.
6. Although not a requirement, two problems the company has had in the past ( “laborious report generation” and “legacy IT issues that affected the efficiency of report generation”) are also resolved by Power BI. Report generation in Power BI is fast, simple and efficient. Furthermore, Power BI has access to many different data pools such as Excel, SQL Server and online sources, which aside from Tableau, is the only reporting and data visualisation software that offers this. However, Power BI can be as complicated as the user makes it with reports, so if AGGORA wanted an alternative and simple only reporting software with the same number of data pools to choose from, a good recommendation would be Tableau. This is because Tableau still has a easy to use User Interface but has less views and elements to it, which in turn could improve accessibility to the Power Bi solution as more employees could use and implement the BI solution.