


STUDENT NAME AND ID NUMBER	Suren Raj Tuladhar 202135813
Qualification	Pearson BTEC Level 5 Higher National Diploma in Digital Technologies (Cyber Security)
Academic Year	2021
Unit Number and Name	Unit 5 Big Data & Visualization
Unit Leader	Himanshu Babbar
Unit Lecturer	Dr. Brinitha Raji
Assignment Title	Big data analysis, visualization and decision making in retail
Type of Assignment	Business Report
Weighting	100%
Issue Date	14 th March 2022
Formative Submission Date	14 th April 2022
Summative Submission Date	2 th May 2022
Assessor	Dr. Brinitha Raji
IV	Himanshu Babbar
Resubmission Date	25 th May 2022

Student Declaration

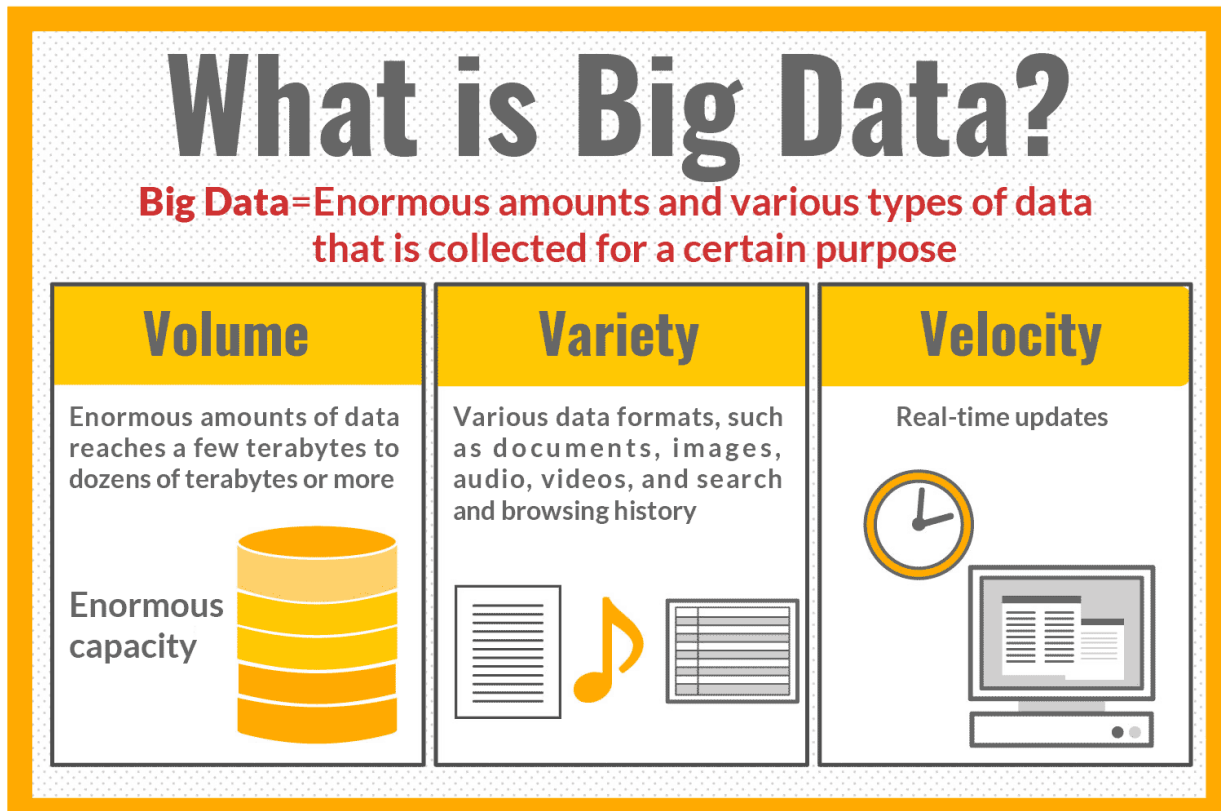
This is to confirm that this submission is my own work, produced without any external help except acceptable support from my lecturer. It has not been copied from any other person's work (published or unpublished) and has not previously been submitted for assessment either at GBS or elsewhere. I confirm that I have read and understood the '[GBS Academic Good Practice and Academic Misconduct: Policy and Procedure](#)' available on Moodle.

I confirm I have read and understood the above Student Declaration.

Student Name (print)	Suren Raj Tuladhar
Signature	
Date	2 nd May 2022

Introductions:

Big data is the collection of data that is huge in volume. The collected volume of data is growing with time. The gathered volume of data will be organized, and processed. Big data volume is high in volume, velocity and variety of information assets. (What is Big data in simple words? Application and perspectives of big data, 2022)



Here, below are some examples of big data in the real world.

Big Data Usage In the real world

1. Banking and Financial Services:

Big data plays an important role in this sector. Using these data, a company can easily analyse the data for high productivity their company. Hence banking and financial services are also no exceptions.

2. Government: Government agencies collect huge amounts of data. For example here in UAE, The UAE Government collects the data of every people who visited the country. The People are required to run medical tests and events required to give their biometrics such as retina scans and fingerprints. Such type of data needs to be processed and analysed very cautiously.

3. Transportations: Big data is the backbone of the GPS application which we use mostly in our smartphones. Similarly, the aeroplanes and airlines generated a huge amount of such data, including fuel efficiency, passengers' weight and cargo, and weather conditions for better route planning and safety of passengers.

4. Healthcare: Big data is the major backbone of the healthcare industry. From sample collection of an individual to the collections of the reports. Technology advancement has taken this game to the next level. Hence, This data plays an important part in real-time predicting and prevention of serious medical conditions by maintaining the e-health records of every individual.

5. Cyber security: With big power comes bigger responsibility. Here big data can make a business vulnerable to cyber-attacks. Historically, we have witnessed numerous ransom ware attacks, and attempts of unauthorized access. etc. This has made the company suffer an intrusion or data theft. Hence, such kind of data can be used to prevent such attacks on the company in future.

(An Introduction to Big Data Concepts and Terminology | DigitalOcean, 2022)

6V's of the Big Data

Volume:

Volume is one of the V's of big data. This scale of information helps in big data systems. Which is higher in magnitude compared to the traditional datasets. Which further demands more processing and storage life cycle.

Velocity:

The collected volume of information in big data needs to travel across the system. The information which is on the continuous flow in the system from different sources is often processed in real-time to receive updates on the current system. The data is constantly added, massaged processed and analyzed to keep up with the flow of new information.

Variety:

The collected data could be of various types. It could be structured, semi-structured or unstructured. These wide ranges of data often make space for unique problems in the case of big data.

Veracity:

Veracity can be referred to as the degree to which big data can be trusted.

Variability: Variation in the data leads to wide variation in quality. Additional resources may be needed to identify, process or filter low-quality data to make it more useful.

Value: The ultimate challenge of big data is delivering value. Sometimes, the systems and processes in place are complex enough that using the data and extracting actual value can become difficult.

(An Introduction to Big Data Concepts and Terminology | DigitalOcean, 2022)

Data Lake:

A data lake allows you to store all your data. It doesn't matter if you're data is structured or unstructured – in volume. In data lake data is typically stored in a raw format without first being processed or structured. From there, it can be further optimized for the purpose at hand, be it a dashboard for interactive analytics, downstream machine learning, or analytics applications. It uses the ELT (Extract Load Transform process. This is ideal for those who want in-depth analysis

Data Warehouse:

A data warehouse is the collection of data that is much more refined, smaller and relational. In other words, unlike Data Lake we can say that it is the optimized and processed data that is structured and semi-structured. Hence, these data are stored in the proprietary format for further easily usage. Data Warehouse uses ETL(Extract Transform Load) process which is ideal for operational users. (Data Lakes vs. Data Warehouses: The Co-existence Argument | Qubole, 2022)

Comparing Data lake vs Warehouse, Data Lake is ideal for those who want in-depth analysis whereas Data Warehouse is ideal for operational users.

Dieter Ram`s Principles

Innovative: Working on the technology means being innovative. Also, technological development often offers new options for innovative design. Hence the innovation can push the design to the next level.

Makes a product useful: The product should be useful not only functional but it has to satisfy certain criteria. It should justify why it is being bought in the first place. A good design emphasizes the usefulness of a product.

Aesthetic: The quality of the product should be integrated with its usefulness and its beauty as well. Because the products we use every day affect our person and our well being.

Make a product understandable: A product should be self-explanatory. Just like a masterpiece of art. It should portray its value as the buyer glance through it.

Unobtrusive: A product should filling a purpose just like tools. These products aren't any decorative objects or works of art. Hence the design should be natural and neutral as possible.

Honesty: Honesty is the best policy and it is also applicable in this scenario as well. The product should not manipulate the client with false promises.

Long-lasting: Unlike, Fashion the design should stay and last many years in this modern-day society.

Thorough down to the last details: No arbitrary to complain. Craftsmanship and detailing should be done very precisely and accurately to show respect towards the consumers.

Environmentally Friendly: Design should be simple and make an important contribution to the environment. The preservation of the environment shouldn't be neglected. Hence, a design should be something that minimizes physical, visual and other resources throughout the lifecycle of the product.

Little as possible: Sometimes the less is better. Simplicity should be the 1st priority, along with the purity of the design should be able to connect with the heart of customers.

Data-Driven Decision Making

Step 1- Define Objective

Make you identify the roles and people for the project

Make a good understanding of DDDM to everyone and the value of DDDM in your company

Define a specific goal from the project. Be specific about your achievements. Such as “ We seek 40% growth in customer satisfaction”. (How to Use Data-Driven Decision-Making to Fuel Growth, 2022)

Step 2- Frame the Assignment

Know the ins and outs of the area of your business that have the biggest impact

declare the key points in the form of a hypothesis whether the statement could be proven or disproven

Step 3- Data Process

Make sure you have the data you needed

Qualitative vs Quantitative

Number of sources of data you need

Reliable and comparable.

Tally the worth of the project

The period for the achievement of the data

Process and status of data collections

Step 4- Filling Data Holes.

An alternative source for data.

Budget allocation.

Legal approvals and process

Research the service provider which is needed and execute them.

Step 5 Data Collections

The job assigned for the data collection and management roles.
Protocols and processes for data collections.
Preliminary process of data and making it clean.

Step 6 Data Analysis

Test your hypothesis using the analytics
Make different models based on changing scenarios
Identify potential change actions

Step 7 Present Finding and decision

Visualization of your raw data
Make concrete benefits of proposed changes
Make decisions
Make an implementation plan
Meeting with key stakeholders.
(Quick Guide to Data-Driven Management | Smartsheet, 2022)

Roles and responsibilities.

Data Analyst:

The Data Analyst is the stepping stone in the field. This data-related job starts as a Data analyst that can be considered an entry-level job as an individual. A good statistical knowledge would be sufficient to be qualified for this role. However, strong technical skill is heartily welcome in this field. Despite these fundamentals, companies expect to be good at data handling, modelling and reporting techniques along with a strong understanding of their business.

Data Engineers:

The data engineers do the development construction and maintenance of the data architectures. Conducting tests, handling error logs and building robust pipelines on a large scale is the major responsibility of the data engineers. Also, the ability to handle raw and unstructured data, providing ideas for improvements in the quality and efficiency of data along with the data process and its development for data modelling, mining and production.

Data Scientist:

The data scientist's task is to perform data processing which includes data transformation along with data cleaning. The task is done by using various machine learning tools which help to forecast the data maintaining the classified pattern of the data. It also helps to increase the performance and maintain the accuracy of machine learning algorithm by fine-tuning and performance optimization

Data Visualization Specialist:

The role and responsibilities of the data visualization specialist are to make complex data more accessible and understandable. In other words, these data should be much more presentable to the stakeholder of the company. The main role is to process the data and deliver it in the most useful and appealing for the user or the stakeholders of the company.

Challenges Of Data Scientists

Multiple Data Sources: Companies nowadays have a lot of data. They have started to use various kinds of software and mobile application like ERPs and CRM for the collection and management of data related to their customer business, or employees. To collect the data in the first place requires loss of manual entries of data which is time-consuming and also can lead to errors and repetitions which may lead to poor decisions.

Data Security: The term data science in business is used to identify business opportunities, and improve overall business performance. Hence, the term data security can also be related to an umbrella that includes all security measures along with the tools applied to analytics and data processes. Hence, I have also listed some of the data security breaches involved.

- Attack on data system
- Ransomware
- Theft

The companies need to follow the three fundamentals of data security i.e

- Confidentiality
- Integrity
- Accessibility

If the 3 fundamentals are correctly placed then the pyramid of data security will be invulnerable.

Lack of Clarity in Business Problems: Clarity of the business is very important for the data specialist. As he will be the response for steering the direction of the business. One should be clear-minded even if the expectation from data science implementation is not aligned with the end goals. Data scientists should follow a proper workflow before starting any analysis.

Undefined KPIs & Metrics

There is often high hope from data science among the management team which may lead to misguided and unrealistic expectations. This may create pressure on the data scientist that may end up affecting their performance. Moreover, a well-defined metrics to measure the accuracy of analysis generated by the data scientist. Also, Proper business KPIs to analyze the impact would be appreciated by the data scientist.

Difficulty In Finding Skilled Data Scientists

The shortage of data scientists is another major issue that is faced by companies across the globe. A person with depth-knowledge and domain expertise is hard to find which also include a deep understanding of ML and AI algorithms and specialist. Finding someone with the art of storytelling through data, along with problem-solving capabilities is quite rare.

Getting Value Out of Data Science.

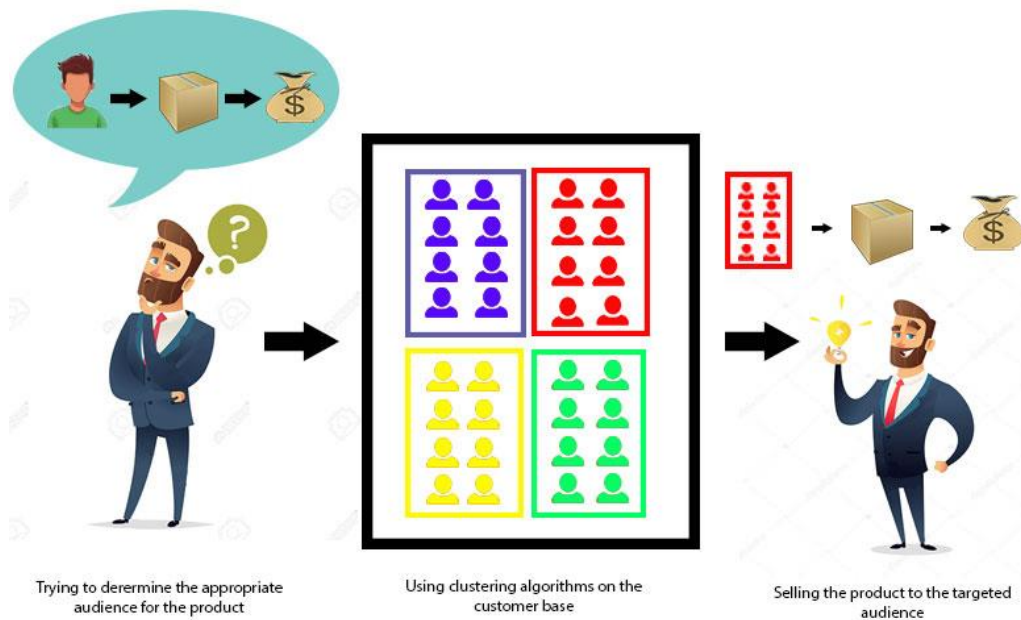
The main objective of the data analysis is to make the right decision. To do so data experts need to be much more agile and in sync with business during the decision-making process.

Statistical & Graphical Techniques Used In Industry

When it comes to the statistical techniques required for big data & visualization, I have listed some of the statistical techniques for predictive analysis

1. Cluster Analysis
2. Regression Analysis
3. Machine Learning
4. Data mining
5. Sentimental analysis

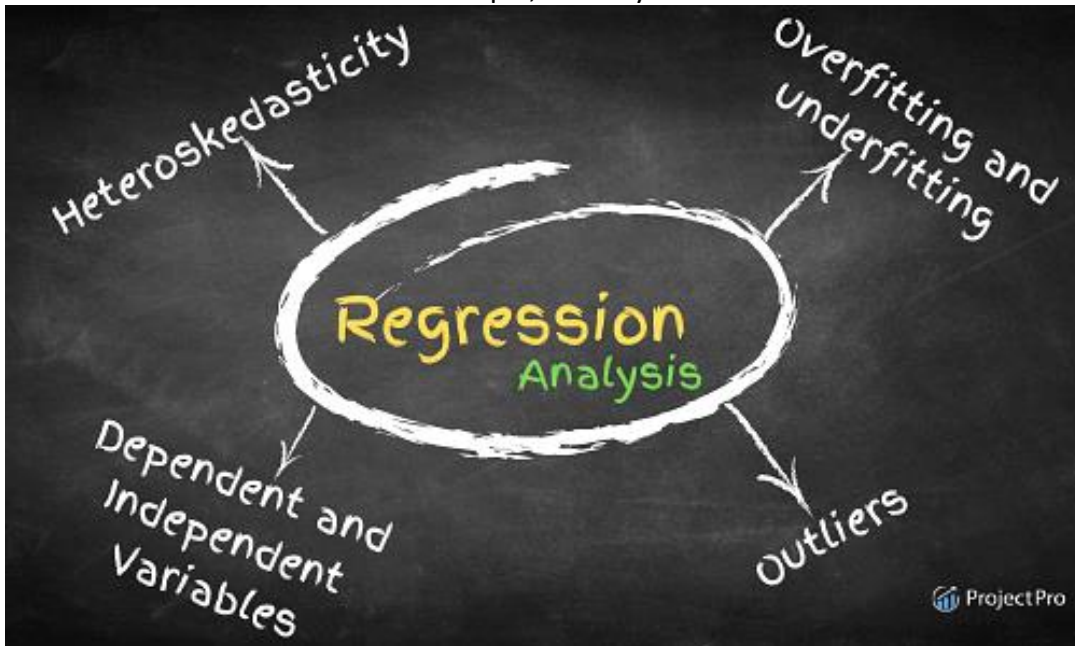
1. **Cluster Analysis:** The analysis process is based on an unsupervised Machine Learning Algorithm that is made of a group of data points into a cluster so that the objects stay in the same group. Hence, clustering helps to splits data into several subsets in which each subset has similar data to each other which are clusters. The primary objective of this analysis is to sort different data points into groups. This means the data points with similar data points are within one cluster whereas dissimilar data points are in another cluster



For example, In the above figure, we can see that sells most of its products, and the customer based clustering algorithm is used. Which are categorized into 4 different colours. Hence, depending on the colour the manager can decide the targeted audience to whom it can sell it.

- 2. Regression Analysis:** This analysis's main objective is to estimate the correlation between a set of variables by analyzing trends and patterns which might impact the dependent variables. This is especially useful to make predictions and forecast trends for the future.

For example, let's say in an eCom

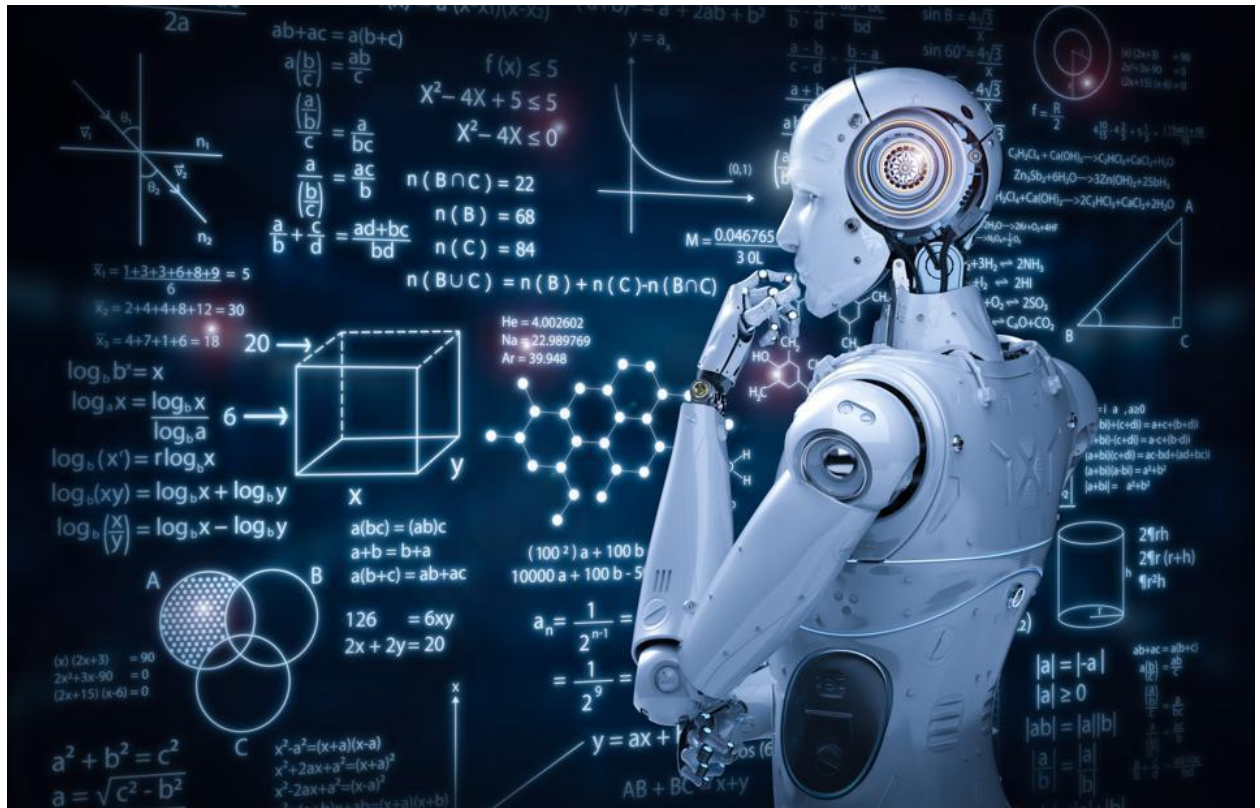


merce company. We wanted to examine the correlation between (a) Budget allocation for social marketing & (b) Sales revenue.

In this case, sale revenue tends to be the dependent variable whereas social media spending is the independent variable. So we want to determine whether the investment in social media is going to be fruitful or will backfire. Using this regression analysis we can find out whether the relationship between these two dependent and independent variables is working in their favour or working against each other.

- 3. Sentiment Analysis:** This analysis solely depends upon the customer's feedback. Think of which there are sectors where the numbers and spreadsheets do make the actual predictions. For such kinds of circumstances, we have to appreciate our customer's emotions and their feedback to set milestones for the future.

5. **Machine Learning:** Machine learning is an application that enables AI systems to learn and improve from experience without needing to be programmed. As a result, it can focus on developing computer programs that can access data and learn from it by themselves.



7. **Data mining:** Data mining's main goal is to identify patterns and relationships which can help to solve issues in the business with the help of data analysis. It is also used to predict future trends and make much more crucial business decisions.

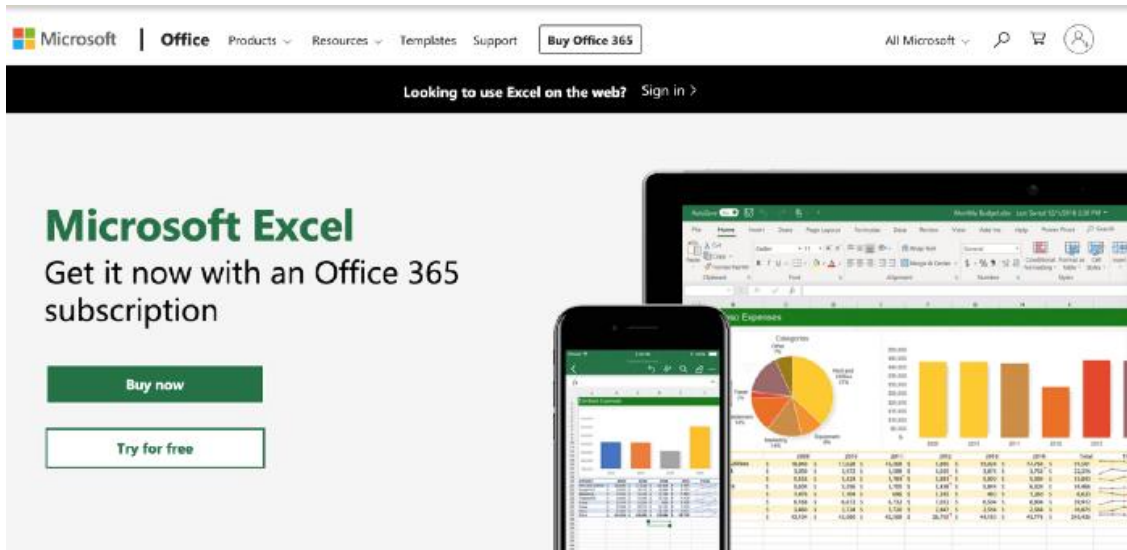


For the data mining ecosystem to exist we have to make sure that its vital elements need to be in place and running. The above picture illustrates clearly the 5 elements i.e. Business Goal, Identify Data, Prepare Data, Evaluate Data, and Present data. Hence, by deep analysis, the company can accomplish its milestone soon.

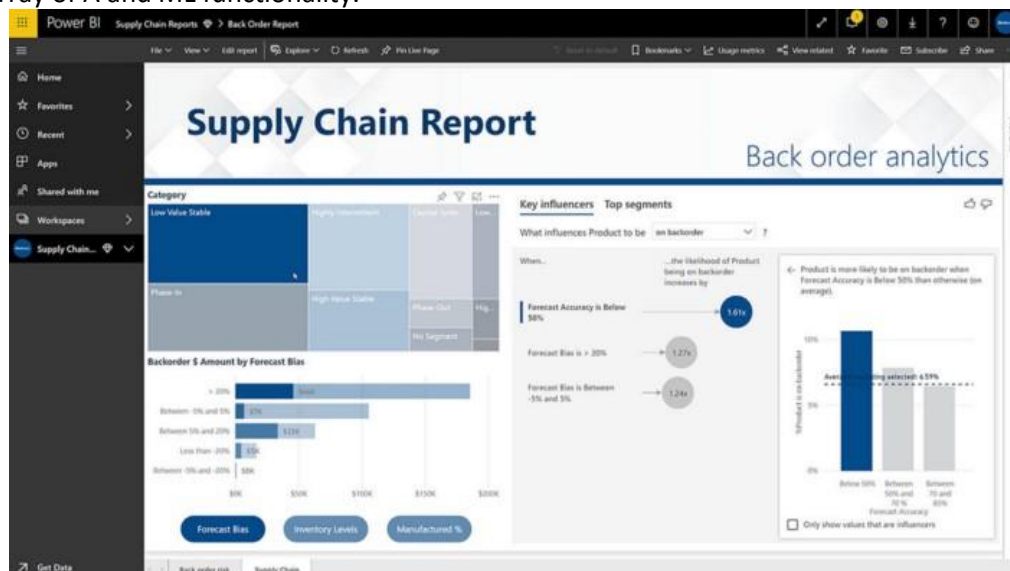
Data Analytics Tools

Tools are the most important things to get the task done. Here I have listed the most common use data analytics tools.

1. **Excel:** The nostalgic analytics tool remains the favourite for most of us to use. The versatile nature of this tool has won the heart of the user. It works best for small data and along with the plugins it can handle millions of data. Being one of the most used Data Analysis Tools it is also categorized as the stepping stone to analyses.

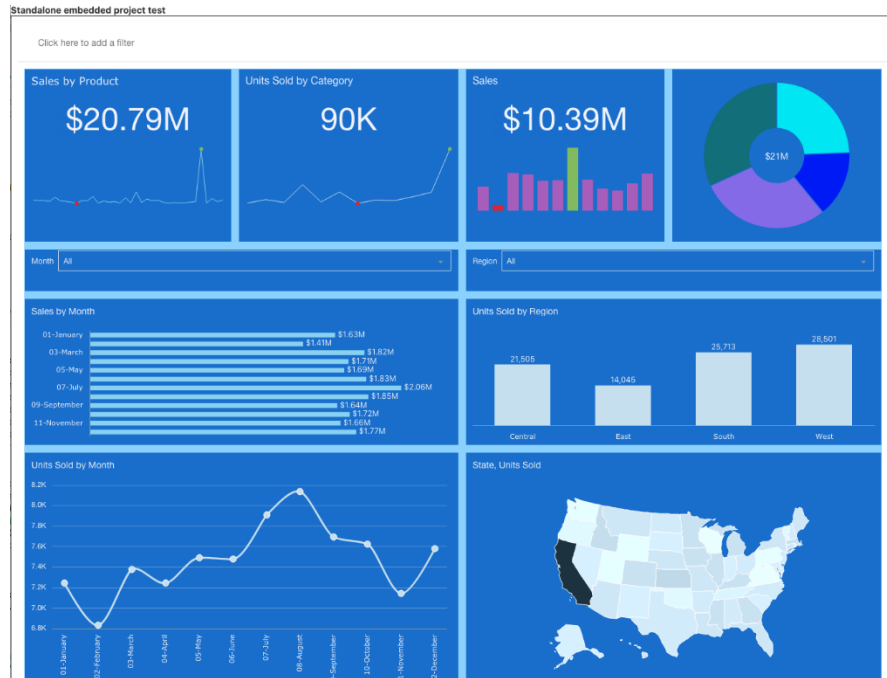


2. **Microsoft Power Bi:** The power BI was initially used to be the plugin for excel. Hence it's been in the market since the initial phase of data revolutions. A newer version of the Power BI is now capable of Machine Learning. This makes the pocket of power bi very deep enabling it to handle the array of A and ML functionality.



3.

4. **Oracle Analytics:** Oracle is the must-have tool for an organization, With AI-based insights generations based on the input data it easily forecasts future outcomes with just a few clicks. Moreover, it can deliver a promising outcome with its machine learning algorithms. Since it is a cloud-based analytics application it is target large enterprises. (Big data analytics: from big data to smart decisions, 2022)



5. **Python:** This analytical tool is widely used by data scientists. It is a much more powerful tool than Excel and Power BI tools. This professional tool uses a high-level programming language that is much more capable than excel and Power Bi. It is often used for statistical analysis and predictive analysis. The 3rd party package can easily be integrated with python for machine learning and data visualizations.

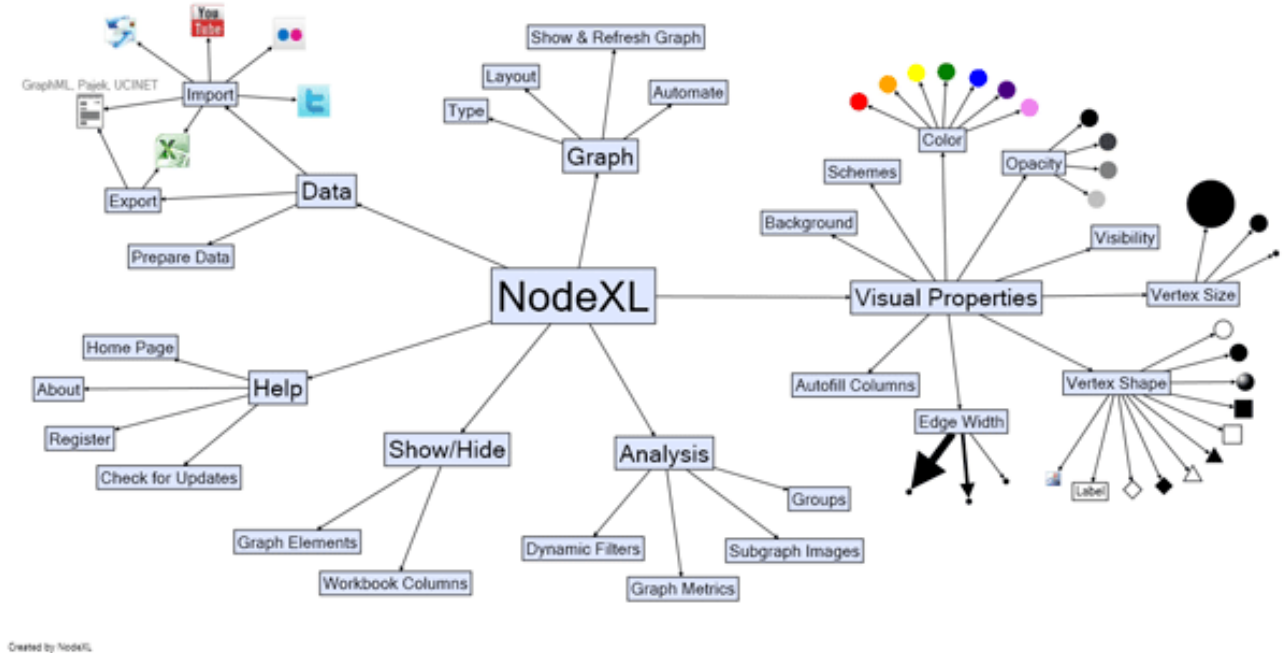


6. **Qlik:** It is the best choice for the organization which seeks to use ML and AI for data analysis. It utilizes AI and ML. The easier nature of this tool enables the sales rep and mid-level staff to execute the software data mining. Being a cloud-based application it can be also analyzed the data from different locations.

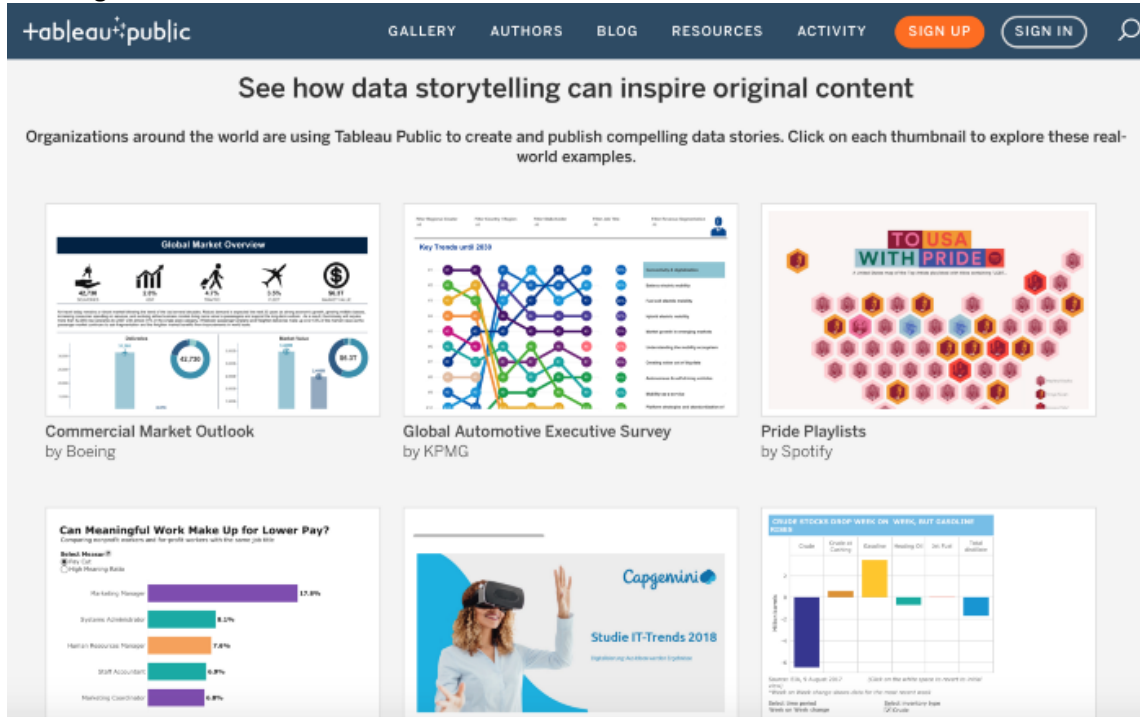


Visualization Tool

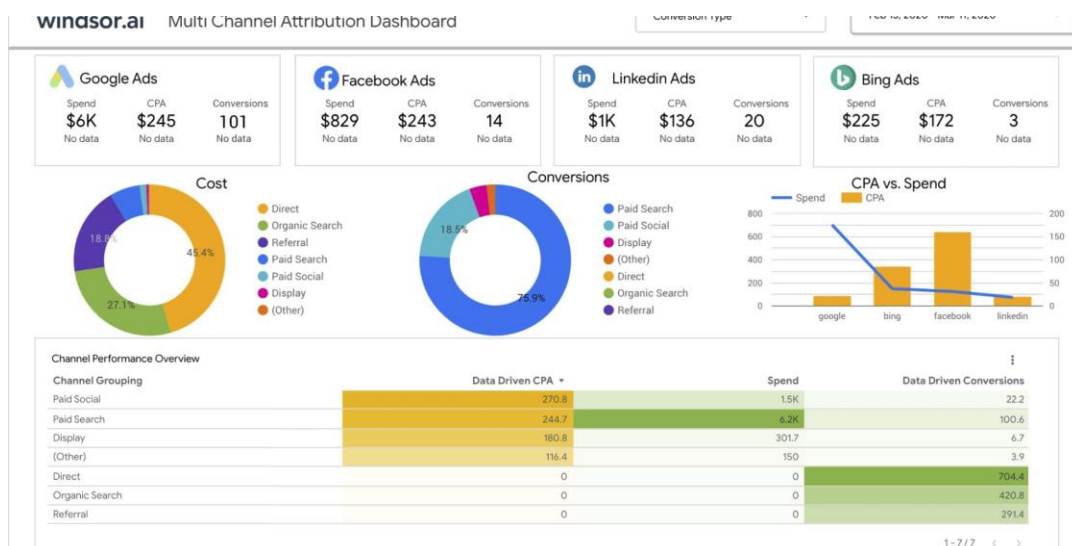
1. **NodeXL:** It is open-source network analysis and visualization software. It is one of the best statistical tools for data analysis which can include advanced network metrics, access to social media network data importers and other automation. (Brockman, 2022)



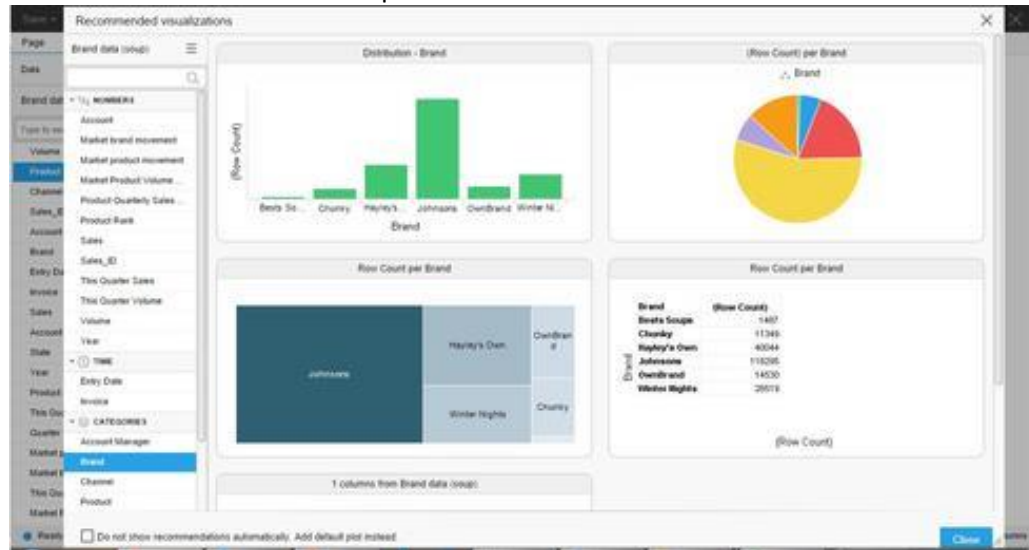
- Tableau:** This is one top data analytics software tools in the market. It has built a large and enthusiastic user base thanks to the depth and quality of the output that it can provide to the user. This tool collects the multiple data inputs, further allowing the user to combine them and then displaying them in a very presentable way in the dashboard which enhances visual data mining.



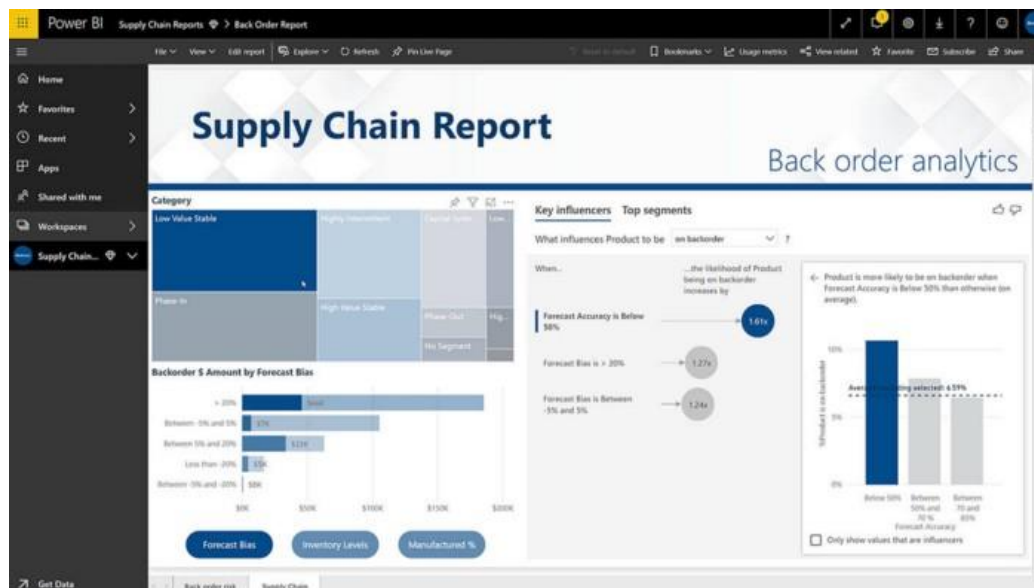
- Google Data Studio:** It is free dashboarding and data visualization tool which is integrated with Google Analytics, Goggle Ads and Google Big Query. Also, help to analyze the sentimental analysis which can further help the company for better productivity and concert prediction and output by understanding customer's conversion and retentions and the reason behind it through a survey for example. And then successfully present the outcome in a presentable and simple form.



4. **TIBCO:** It provides natural languages and AI-powered data insight. It can publish reports to both mobile and desktop applications. It also can build the predictive analytics models by providing point and click features.



5. **Power BI:** The power BI was initially used to be the plugin for excel. Hence it's been in the market since the initial phase of data revolutions. A newer version of the Power BI Is now capable of Machine Learning. This makes the pocket of power bi very deep enabling it to handle the array of A and ML functionality.



Analytical And Visualization Tool An Organization Requires.

The organization requires a good analytical and Visualization tool to process and beautify their data. Hereafter, they can use the processed data to ensure crucial decisions for their organization.

Here, for this project, I will use “**Microsoft SQL Server Management Studio 18**”. Being a well-known data tool in the market it will be much more relevant to use in the given scenario. Followed by for the visualization, I will be using “**Power BI**” since it compliments each other quite well.

Business Requests And Planning

Client: AJ Retails

Changes requested:

- Complete Analysis of their Sales
- Improvement of Online Sales & Reports
- Evolve their static report into a visual dashboard.
- Quantity of the product which they have sold, co-related to the time

System requirements:

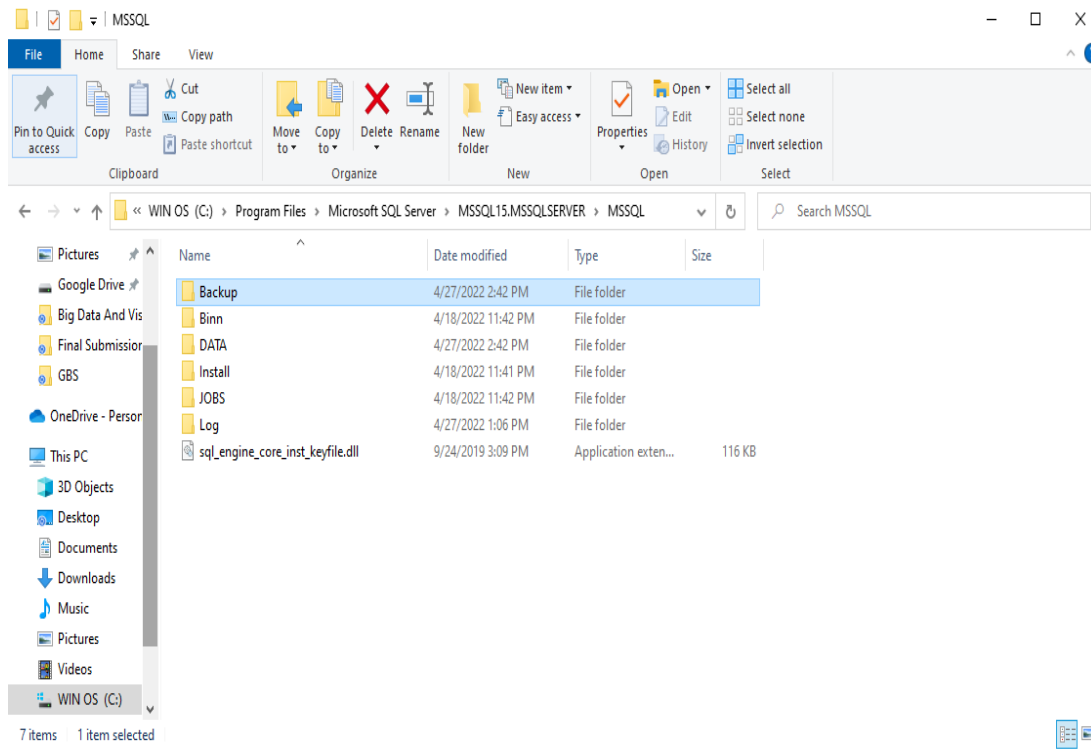
- PC with 128 Gb Of SSD, 8 Gb Min Of RAM, i5 CPU
- That supports Win 10 or higher
- SQL Server Management Studio 18
- SQL Server Profiler 18
- Power BI
- Microsoft Excel

Business Requests And Planning

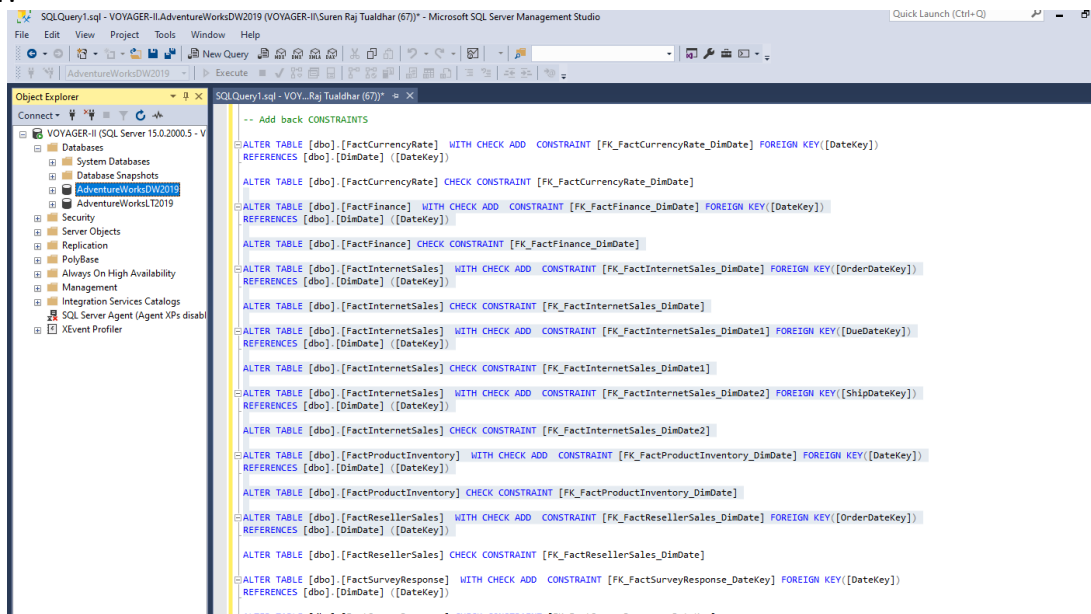
Role	Business Requests	User value	Acceptance criteria
Sales & Marketing Department	Sales Analysis	With this data, the Sales and marketing department will be able to analyse the total sales report. With the help of it would be much easier to make the future decision.	The data should be updated daily.
Sales & Marketing Department, E-commerce	The request is to make the report for online sales so that the company could out stand on the e-commerce platform as well. Also, to shift from static reports to a visual dashboard.	With this data, the corresponding department will be able to understand the sales pattern and algorithm. As static reports could be understood well by the technical people only The dynamic dashboards approach will give a visual picture of the static reports which will be much simpler and easy to understand.	The data should be updated daily and the accuracy of the data should be maintained.
Sales & Marketing Department	The request is to find out the amount of product the company has sold to the client over a while.	With this, the company would be able to understand the algorithm behind the product wise sales.	The data should be up to date and accuracy should be strongly maintained.
Sales & Marketing Department	The request is to find the customer wise sales report.	With this, the company will be able to understand the sales report customer wise. Which will further help the company to make the better decision for coming up with better promotions and offer.	The sales data should be updated daily
Sales & Marketing Department	The request is to find the sale report over time.	With this, the company will be able to understand the sales report over time. Which will further help the company to make the better decision for placing the coming up with better promotions and offer in the year's calendar.	The sales data should be updated daily and accuracy should be maintained.

Here, as the rudimentary requirement of our client “**AJ Retails**” to obtain the data of sales in respect to Calendar, Internet sales, Product Category, & Customer City. I have chosen to work with respective tables which provide me with the needed data to execute the analysis and to portray the outcome for the visualization.

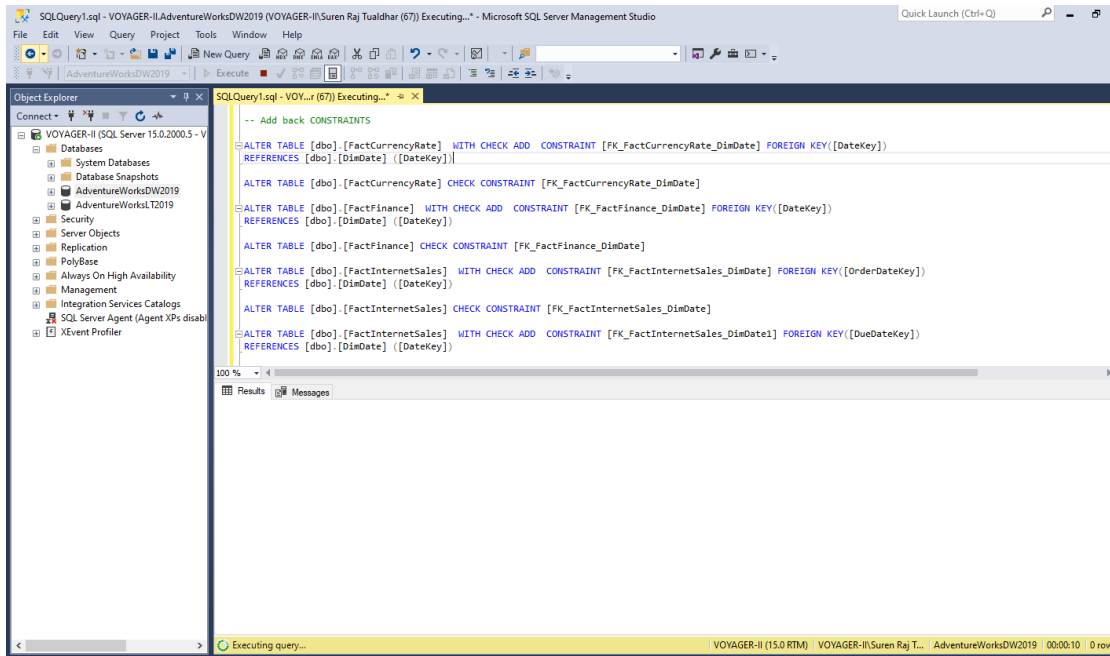
After acquiring the data, it is uploaded into the SQL database. For this project, I have used the pre-existing backup file, which is illustrated clearly in the picture below.



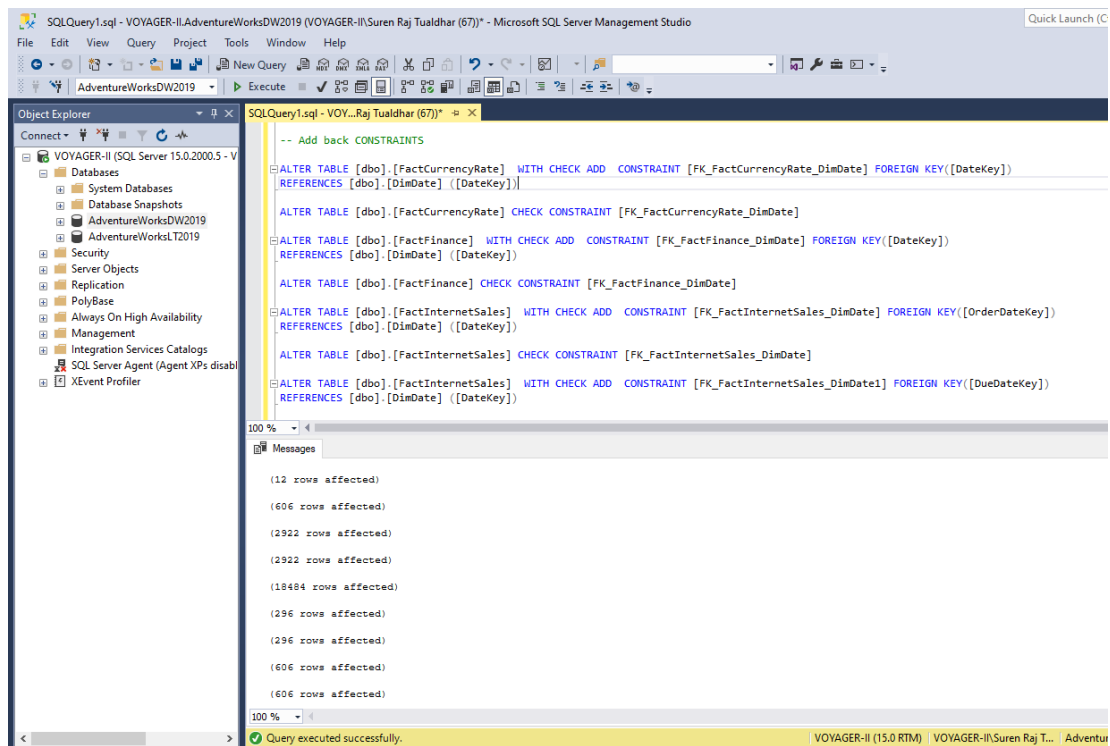
Here, In SQL Server Management Studio I have restored the backup files for the “Backup” folder.



Here, I have used the “Github Script”, So that I can take and filter data dated from 2019- 2022.



In the above diagram, we can see that the “Github Script” is being executed.



The above diagram illustrates that the “Github Script” has been successfully executed.

Difference between Dimension and Fact Table

Dimension Table	Fact Table
It is the collection of data where much more details are given such as Student Name, Class, Student Id, and Course.	It's a limited collection where much-needed fact data are stored, for example, Student Id Number, Student Grade, Etc.

SQL Queries that are used to obtain refined data

The screenshot displays the Microsoft SQL Server Enterprise Manager interface. The Object Explorer on the left shows the database structure, including tables like dbo.AdventureWorks, dbo.DatabaseLog, and various dimension tables. The central pane shows a SQL query script for 'SQLQuery3.sql' with the following content:

```
/****** Script for SelectTopNRows command from SSMS ******/
SELECT distinct
    [CalendarYear]
FROM [AdventureWorksDW2019].[dbo].[DimDate]
```

The bottom pane shows the results of the query, displaying a list of years from 2010 to 2017, with 18 rows in total. The status bar at the bottom indicates 'Query executed successfully.' and '18 rows'.

Here, In the above diagram, I have to use the “**Distinct Query**” to filter the Year from repeating.

The screenshot shows the Microsoft SQL Server Management Studio interface. The Object Explorer on the left displays the database structure, including tables like DimDate, DimCustomer, and DimProduct. The central query window shows a SQL query for the 'Date' table, which includes various date and time components. The Results pane at the bottom displays the output of the query, showing columns like DateKey, Date, Day, Month, MonthShort, MonthNo, Quarter, and Year. The status bar at the bottom indicates that the query was executed successfully, returning 1,461 rows.

```

SELECT
-- Cleansed DIM_Date Table --
[DateKey],
[FullDateAlternateKey] AS Date,
--[DayNumberOfWeek],
[EnglishDayNameOfWeek] AS Day,
--[SpanishDayNameOfWeek],
--[FrenchDayNameOfWeek],
--[DayNumberOfMonth],
--[DayNumberOfYear],
--[WeekNumberOfYear],
[EnglishMonthName] AS Month,
Left([EnglishMonthName], 3) AS MonthShort, -- Useful for front end date navigation and front end graphs.
--[SpanishMonthName],
--[FrenchMonthName],
[MonthNumberOfYear] AS MonthNo,
[CalendarQuarter] AS Quarter,
[CalendarYear] AS Year --[CalendarSemester],
--[FiscalQuarter],
--[FiscalYear],
--[FiscalSemester]

```

DateKey	Date	Day	Month	MonthShort	MonthNo	Quarter	Year
1	2019-01-01	Tuesday	January	Jan	1	1	2019
2	2019-01-02	Wednesday	January	Jan	1	1	2019
3	2019-01-03	Thursday	January	Jan	1	1	2019
4	2019-01-04	Friday	January	Jan	1	1	2019
5	2019-01-05	Saturday	January	Jan	1	1	2019
6	2019-01-06	Sunday	January	Jan	1	1	2019
7	2019-01-07	Monday	January	Jan	1	1	2019
8	2019-01-08	Tuesday	January	Jan	1	1	2019
9	2019-01-09	Wednesday	January	Jan	1	1	2019
10	2019-01-10	Thursday	January	Jan	1	1	2019
11	2019-01-11	Friday	January	Jan	1	1	2019
12	2019-01-12	Saturday	January	Jan	1	1	2019
13	2019-01-13	Sunday	January	Jan	1	1	2019
14	2019-01-14	Monday	January	Jan	1	1	2019

Here, In the above diagram, I have used Query for data cleansing of the Date table.

The screenshot shows the Microsoft SQL Server Management Studio interface. The Object Explorer on the left displays the database structure, including tables like DimCustomer, DimProduct, and DimSalesTerritory. The central query window shows a SQL query for the 'Customer' table, which includes various customer attributes. The Results pane at the bottom displays the output of the query, showing columns like CustomerKey, First Name, Last Name, Full Name, Gender, DateFirstPurchase, and Customer City. The status bar at the bottom indicates that the query was executed successfully, returning 18,484 rows.

```

-- Cleansed DIM_Customers Table --
SELECT
c.customerkey AS CustomerKey,
-- ,[GeographyKey]
-- ,[CustomerAlternateKey]
-- ,[Title]
c.firstname AS [First Name],
-- ,[MiddleName]
c.lastname AS [Last Name],
c.firstname + ' ' + lastname AS [Full Name],
-- Combined First and Last Name
-- ,[NameStyle]
-- ,[BirthDate]
-- ,[MaritalStatus]
-- ,[Suffix]
CASE c.gender WHEN 'M' THEN 'Male' WHEN 'F' THEN 'Female' END AS Gender,
-- ,[EmailAddress]
-- ,[YearlyIncome]
-- ,[TotalChildren]
-- ,[NumberChildrenAtHome]
-- ,[EnglishEducation]

```

CustomerKey	First Name	Last Name	Full Name	Gender	DateFirstPurchase	Customer City
11000	Jon	Yang	Jon Yang	Male	2019-01-19	Rockhampton
11001	Eugene	Huang	Eugene Huang	Male	2019-01-15	Seaford
11002	Ruben	Torres	Ruben Torres	Male	2019-01-07	Hobart
11003	Christy	Zhu	Christy Zhu	Female	2018-12-29	North Ryde
11004	Elizabeth	Johnson	Elizabeth Johnson	Female	2019-01-23	Wollongong
11005	Julio	Ruiz	Julio Ruiz	Male	2018-12-30	East Brisbane
11006	Janet	Avarez	Janet Avarez	Female	2019-01-24	Matraville
11007	Marco	Mehta	Marco Mehta	Male	2019-01-09	Warrambool
11008	Rob	Verhoff	Rob Verhoff	Female	2019-01-25	Bendigo
11009	Shannon	Carlson	Shannon Carlson	Male	2019-01-27	Henvey Bay
11010	Jacquelyn	Suarez	Jacquelyn Suarez	Female	2019-01-14	East Brisbane
11011	Curtis	Lu	Curtis Lu	Male	2018-12-30	East Brisbane
11012	Lauren	Walker	Lauren Walker	Female	2021-03-16	Brimston
11013	Ian	Jenkins	Ian Jenkins	Male	2021-04-13	Lebanon

Here, in the above diagram, I have used Query for data cleansing of the Customer table.

Episode 3 - SQL Script - DIM_Products.sql - VOYAGER-IIAdventureWorksDW2019 (VOYAGER-II:Suren Raj Tualldhar (53)) - Microsoft SQL Server Management Studio

Object Explorer: AdventureWorksDW2019

Query: Episode 3 - SQL Script - DIM_Products Table --

```

SELECT
    p.[ProductKey],
    p.[ProductAlternateKey] AS [ProductItemCode],
    -- ,[ProductSubcategoryKey],
    -- ,[WeightUnitMeasureCode],
    -- ,[SizeUnitMeasureCode],
    p.[EnglishProductName] AS [Product Name],
    pc.[EnglishProductCategoryName] AS [Sub Category], -- Joined in from Sub Category Table
    pc.[EnglishProductCategoryName] AS [Product Category], -- Joined in from Category Table
    -- ,[SpanishProductName],
    -- ,[FrenchProductName],
    -- ,[StandardCost],
    -- ,[FinishedGoodsFlag],
    p.[Color] AS [Product Color],
    -- ,[SafetyStockLevel],
    -- ,[ReorderPoint],
    -- ,[ListPrice],
    p.[Size] AS [Product Size],
    -- ,[SizeRange],
    -- ,[Weight]

```

ProductKey	ProductItemCode	Product Name	Sub Category	Product Category	Product Color	Product Size	Product Line	Product Model Name	Product Description	Product Status
1	AR-5381	Adjustable Race	NULL	NULL	NA	NULL	NULL	NULL	NULL	Current
2	BA-8327	Bearing Ball	NULL	NULL	NA	NULL	NULL	NULL	NULL	Current
3	BE-2349	BB Ball Bearing	NULL	NULL	NA	NULL	NULL	NULL	NULL	Current
4	BE-2908	Headset Ball Bearings	NULL	NULL	NA	NULL	NULL	NULL	NULL	Current
5	BL-2036	Blade	NULL	NULL	NA	NULL	NULL	NULL	NULL	Current
6	CA-5965	LL Crankarm	NULL	NULL	Black	NULL	NULL	NULL	NULL	Current
7	CA-6738	HL Crankarm	NULL	NULL	Black	NULL	NULL	NULL	NULL	Current
8	CA-7457	HL Crankarm	NULL	NULL	Black	NULL	NULL	NULL	NULL	Current
9	CB-2903	Chaining Bolts	NULL	NULL	Silver	NULL	NULL	NULL	NULL	Current
10	CN-6137	Chaining Nut	NULL	NULL	Silver	NULL	NULL	NULL	NULL	Current
11	CR-7833	Chaining	NULL	NULL	Black	NULL	NULL	NULL	NULL	Current
12	CR-9981	Crown Race	NULL	NULL	NA	NULL	NULL	NULL	NULL	Current
13	CS-2812	Chain Stays	NULL	NULL	NA	NULL	NULL	NULL	NULL	Current
14	DC-8732	Decal 1	NULL	NULL	NA	NULL	NULL	NULL	NULL	Current

Query executed successfully. 14 rows affected.

Here, In above diagram I have used Query for data cleansing of the Product table.

Episode 3 - SQL Script - FACT_InternetSales.sql - VOYAGER-IIAdventureWorksDW2019 (VOYAGER-II:Suren Raj Tualldhar (63)) - Microsoft SQL Server Management Studio

Object Explorer: AdventureWorksDW2019

Query: Episode 3 - SQL Script - FACT_InternetSales Table --

```

SELECT
    [ProductKey],
    [OrderDateKey],
    [DueDateKey],
    [ShipDateKey],
    [CustomerKey],
    -- ,[PromotionKey],
    -- ,[CurrencyKey],
    -- ,[SalesTerritoryKey],
    [SalesOrderNumber],
    -- ,[SalesOrderLineNumber],
    -- ,[RevisionNumber],
    [OrderQuantity],
    [UnitPrice],
    [ExtendedAmount],
    [UnitPriceDiscountPct],
    [DiscountAmount],
    [ProductStandardCost],
    [TotalProductCost],
    [SalesAmount] -- ,TaxAmt

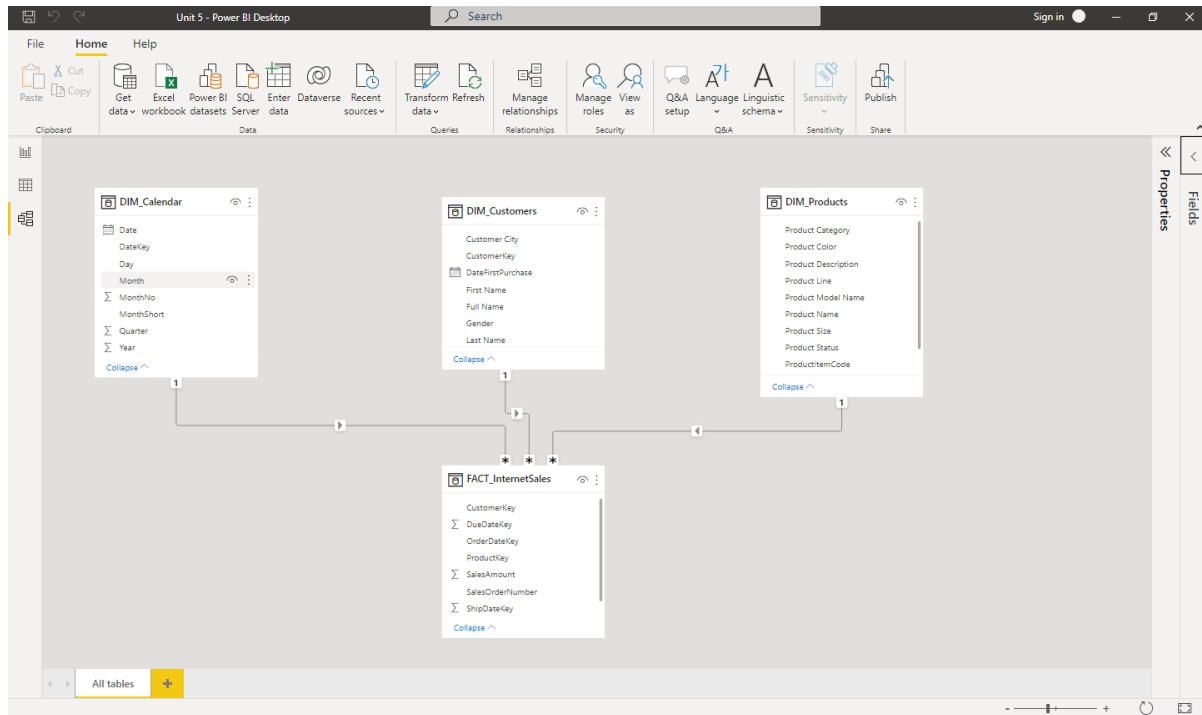
```

ProductKey	OrderDateKey	DueDateKey	ShipDateKey	CustomerKey	SalesOrderNumber	SalesAmount
1	381	20200101	20200113	20200108	16942	1000.4375
2	375	20200101	20200113	20200108	15114	2181.5625
3	369	20200101	20200113	20200108	15116	2443.35
4	337	20200101	20200113	20200108	20576	782.99
5	370	20200101	20200113	20200108	13059	2443.35
6	370	20200101	20200113	20200108	13085	2443.35
7	352	20200101	20200113	20200108	20186	2071.4196
8	337	20200101	20200113	20200108	15199	782.99
9	377	20200101	20200113	20200108	21200	2181.5625
10	387	20200102	20200114	20200109	19172	1000.4375
11	356	20200102	20200114	20200109	11484	2071.4196
12	369	20200102	20200114	20200109	13582	2443.35
13	373	20200102	20200114	20200109	13779	2181.5625
14	371	20200102	20200114	20200109	24778	2181.5625

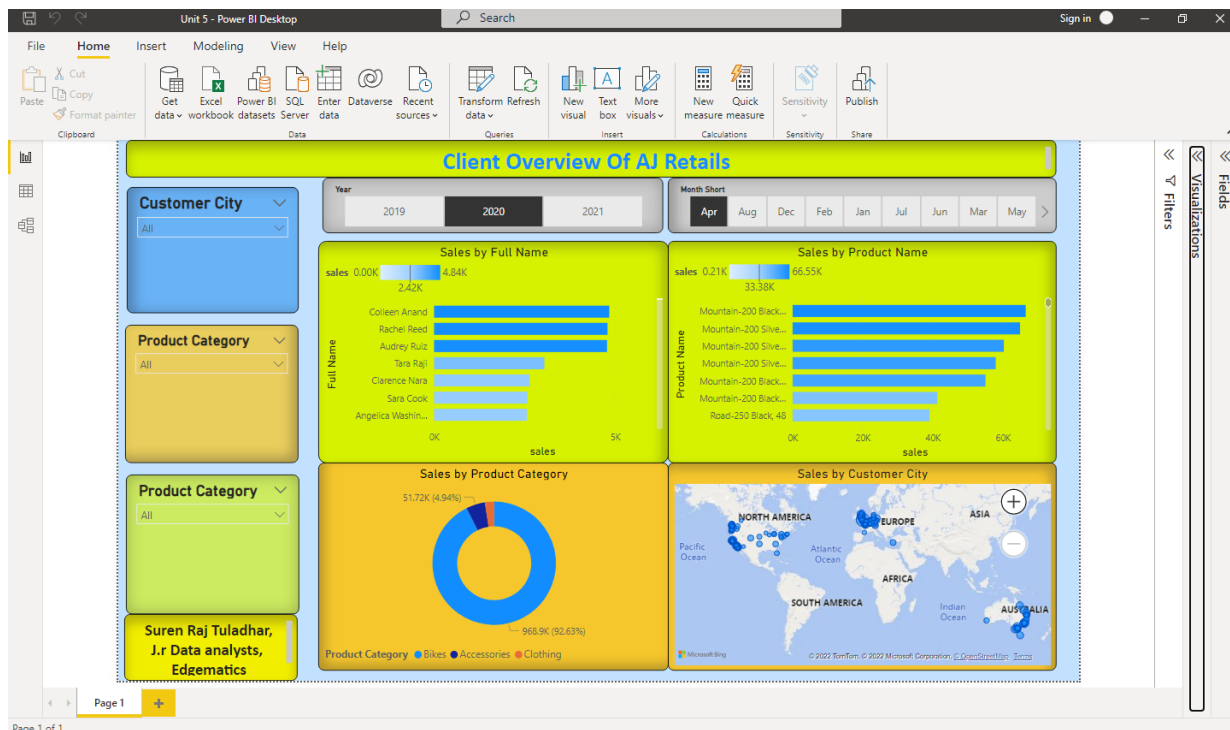
Query executed successfully. 14 rows affected.

Here, in the above diagram, I have used Query for data cleansing of the Sales table.

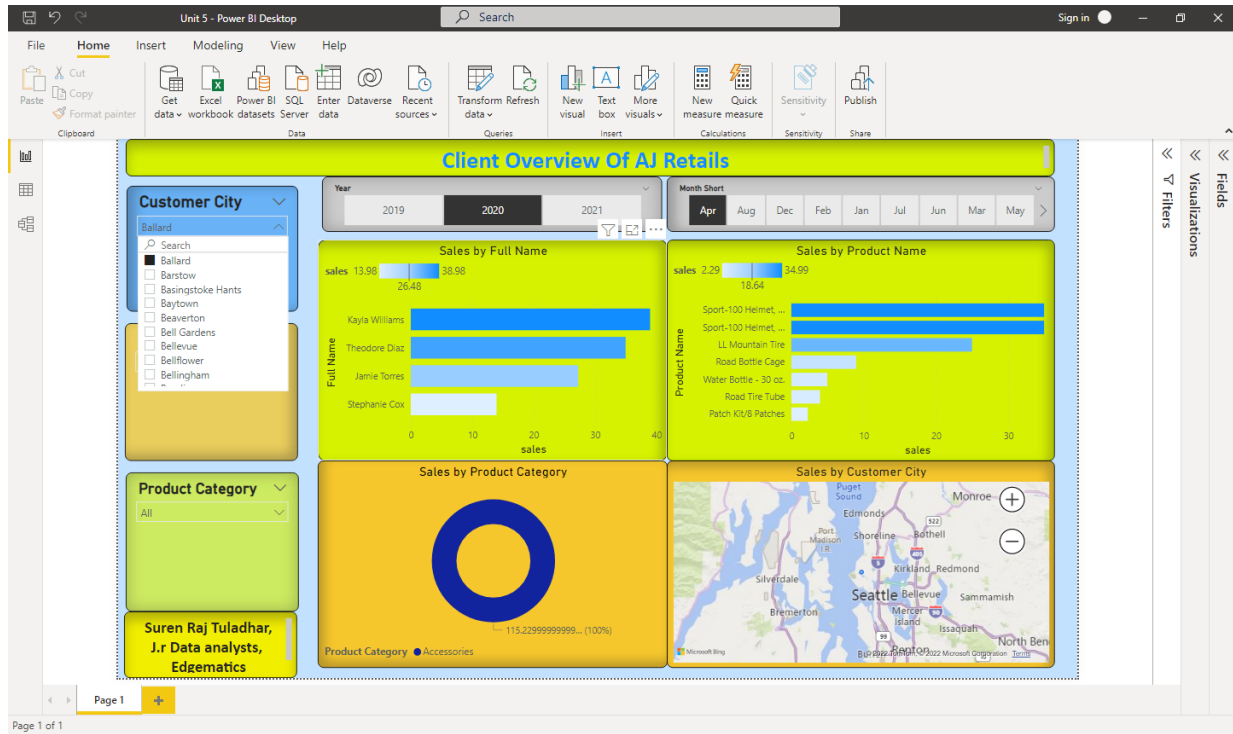
Power BI For Visualization



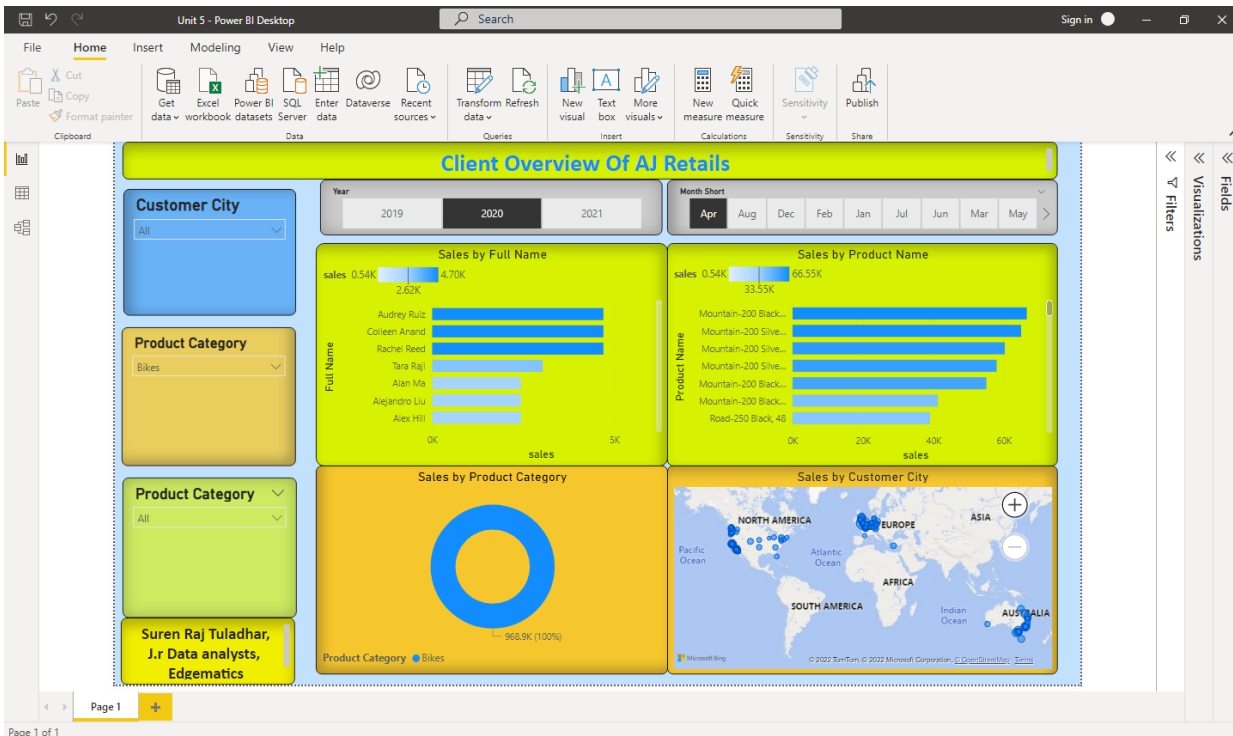
Here, I have used Power BI for the visualization of data. Moreover, I have also used the 4 tables i.e. Calendar, Customer, Products and Internet Sales as it was the fundamental requirement of the client i.e. AJ Retails



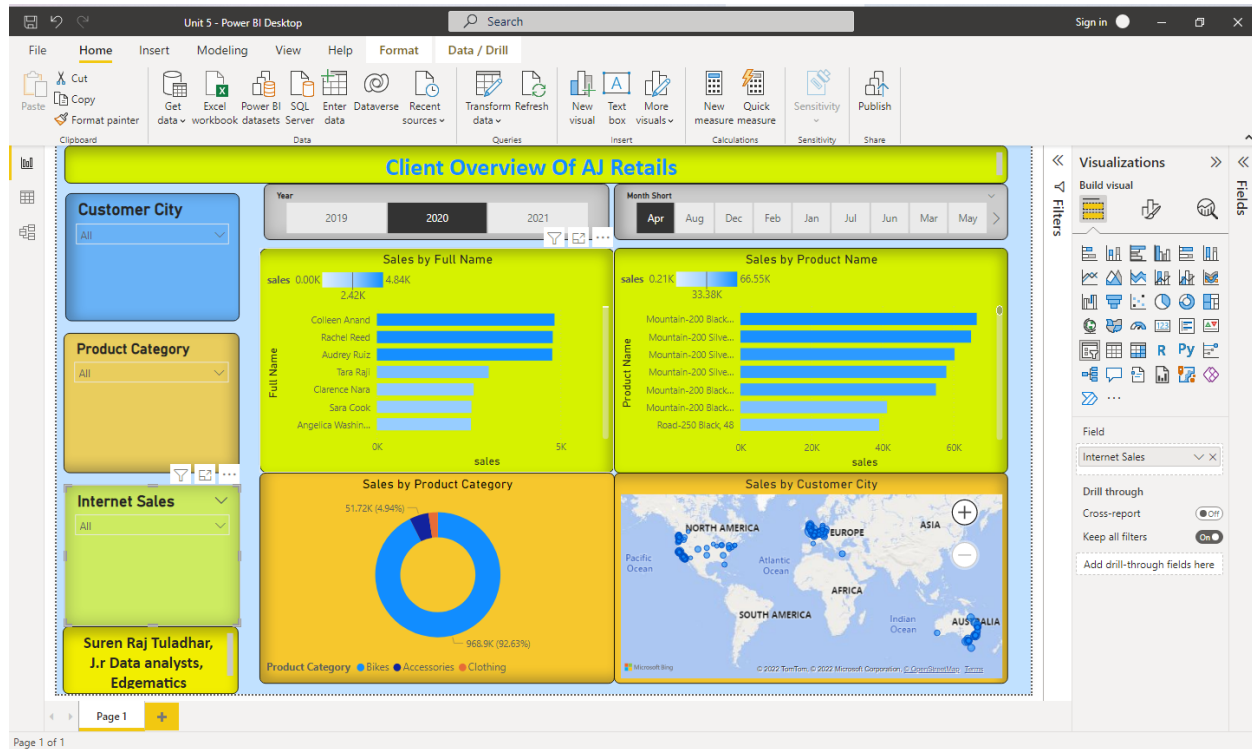
In the above diagram, we can see the data of AJ retails of April 2020,



In the above diagram, we can see the data filter. I have selected to filter the data by the Customer city.



In the above diagram, we can see the data filter. I have selected to filter the data by the Product Category



Here, In above diagram represent the dynamic dashboard as the part of the visualization of data which was requested by the AJ Retails. In which, the data can be viewed as per the fundamental requirements i.e. Calendar, Customer, Products and Internet Sales.

Reference:

- Blockchain Journal. 2022. *What is Big data in simple words? Application and perspectives of big data*. [online] Available at: <<https://blockchainjournal.news/what-is-big-data-in-simple-words-application-and-perspectives-of-big-data/>> [Accessed 2 May 2022].
- Smartsheet. 2022. *Quick Guide to Data-Driven Management | Smartsheet*. [online] Available at: <<https://www.smartsheet.com/data-driven-decision-making-management>> [Accessed 2 May 2022].
- Superoffice.com. 2022. *How to Use Data-Driven Decision-Making to Fuel Growth*. [online] Available at: <<https://www.superoffice.com/blog/data-driven-decision-making/>> [Accessed 2 May 2022].
- Brockman, A., 2022. *6 Tips for Creating Effective Data Visualizations (with Examples)*. [online] Blog.csgsolutions.com. Available at: <<https://blog.csgsolutions.com/6-tips-for-creating-effective-data-visualizations>> [Accessed 2 May 2022].
- Qubole. 2022. *Data Lakes vs. Data Warehouses: The Co-existence Argument | Qubole*. [online] Available at: <<https://www.qubole.com/data-lakes-vs-data-warehouses-the-co-existence-argument/>> [Accessed 2 May 2022].
- i-SCOOP. 2022. *Big data analytics: from big data to smart decisions*. [online] Available at: <<https://www.i-scoop.eu/big-data-action-value-context/big-data-analytics-from-big-data-to-smart-data-and-decisions/>> [Accessed 2 May 2022].
- GeeksforGeeks. 2022. *Big Data Analytics Life Cycle - GeeksforGeeks*. [online] Available at: <<https://www.geeksforgeeks.org/big-data-analytics-life-cycle/>> [Accessed 2 May 2022].
- Digitalocean.com. 2022. *An Introduction to Big Data Concepts and Terminology | DigitalOcean*. [online] Available at: <<https://www.digitalocean.com/community/tutorials/an-introduction-to-big-data-concepts-and-terminology>> [Accessed 2 May 2022].
- Digitalocean.com. 2022. *An Introduction to Big Data Concepts and Terminology | DigitalOcean*. [online] Available at: <<https://www.digitalocean.com/community/tutorials/an-introduction-to-big-data-concepts-and-terminology>> [Accessed 2 May 2022].