Maria Cristina Aarmeniei Pricope

30245077

Business Intelligence

Assignment 1

Table of Contents

[Section One - Introduction 2](#_Toc182899458)

[Section Two - Data Extraction and Transformation 2](#_Toc182899459)

[Section Three - Data Analysis and Visualisation 3](#_Toc182899460)

[Section Four - Applying Techniques for a Business Scenario 4](#_Toc182899461)

[Section Five - Insights and Recommendations 13](#_Toc182899462)

# Section One - Introduction

This report aims to evaluate the sales data of a mid-size e-commerce company that has experienced fluctuating sales throughout the year of 2023 and would like to understand the reasons behind the customer behaviour. The data will be handled by an existing data analyst employee within the company whose main goal is to identify patterns in sales.

To achieve the goal, an excel spreadsheet has been provided which includes three sheets, each containing a table, as follows:

1. Sales Data: this table contains date of purchase, product category information, product identification number, product name, quantity sold, revenue in USD and order ID;
2. Customer Data: this table contains information about the amount of customer purchases, first purchase date, total purchases, the revenue resulted from the sales and region;
3. Marketing Campaign Data: contains information about different sales campaigns, the periods in which they run, the discount they offered, and which product categories were promoted during the marketing campaigns.

This report aims to follow Business Intelligence (BI) practices to gather data from multiple sources and produce insights to aid in understanding the vast ecosystem of a business. BI is a technology driven process for analysing data and presenting actionable information to help executives and end users make informed business decisions; it uses facts, not intuition.

This report will explore the core components of BI, which are:

1. Data collection: gather data from different sources
2. Data analysis: using tools and methods to analyse the data
3. Information presentation: providing visualisation or reports for decision making

BI is important in modern business because it creates strategic decision-making using data driven decisions, providing facts rather than relying on intuition. BI provides companies with a competitive edge by understanding customer preferences, optimising operations and improve products or services.

BI process flow follows four important stages, which are also regarded as the backbone of effective data, these are:

1. Data extraction: data is gathered from a variety of sources
2. Data transformation: cleaning the data to ensure accuracy and consistency ensuring that the data is meaningful
3. Data analysis: data is loaded in the BI application to analyse
4. Presentation: creating visualisation to portray the story behind the data

To analyse the data, Python will be used together with libraries to import the data, perform numerical operations, data manipulation, transform the data into a temporary SQL database and create visual representation of the data.

# Section Two - Data Extraction and Transformation

This section aims to explore the first two stages of the BI flow, also known as data extraction and data transformation.

1. Data extraction is the process of gathering data from various internal and external sources like this involves collecting data from databases(SQL DB, data warehouses), web services (APIs from third party services) or even through scrapping unstructured data(from logs, documents or websites) aiming to capture all the facets of the business operations. In the section below we will explore APIs and SQL as a data extraction tools:

SQL – Structured Query Language

SQL is used to manage and manipulate databases, access and retrieve data from relational database systems which store data in tables that can be linked by common attributes. SQL is a structured query language to manage and extract data. SQL is using the key features SLECT\*FROM to extract specific data – it allows you to specify exactly which data you want to retrieve; JOIN operations to combine data from multiple tables, merge data based on related columns. SQL is best for structured data; databases and it excels in environments where data is organised in pre-defined schemes such as tables and rows. It allows for complex queries including filtering, sorting and joining data, making it ideal for a deep analytical task on structured databases.

SQL is easy to use as the syntax is straightforward and makes it easy to learn especially for those with technical background; it excels in environments where data is stored in pre-defined formats.

APIs – Application Programming Interfaces

APIs are a set of rules and protocols that allow different software applications to communicate which each other, often used to access data from other services. APIs provide a programmable way to access real time data directly from the source which can be then integrated in a business intelligence process. APIs are incredibly useful for accessing dynamic, real-time data from external platforms; provide a gateway for a vast unstructured data, like live financial data, social media feeds, integrating with cloud-based services; integrate data from various websites with minimum modifications – very complex to handle Aps and data integrity issues. Example of use of APIs: a company can analyse patterns and delivery routes based on live weather to avoid delays.

Advantages: real time data access allowing business to make decisions based on the most current information available to them; this is crucial in any environment where conditions can change rapidly such as supply chain markets; automated integration of data from multiple platforms to streamline workflow, reducing time and effort needed to gather and process data.

2. Data transformation is the process of cleaning and processing the data, formatting consistently to make data meaningful. Data transformation includes cleaning, standardisation and aggregation which are needed to improve quality, accuracy, standardise data and remove invalid or null vales. Inconsistent data formats make it difficult to compare the data against other datasets. Data transformation is the foundation of reliable decision making.

Data cleaning is the process of removing or correcting inaccuracies as these can distort the analysis and mislead the insights and can involve addressing duplicate records, missing values, incorrect formats (dates, currencies etc.). For example, removing duplicate values from the Sales Data table will provide real insights in terms of customer retention or purchasing trends; also, normalising the Quantity Sold in a sales dataset, provides a fair insight of a product’s performance. Other examples of cleaning the data include addressing missing or null values in key metrics like columns Total Purchases and Total Revenue (USD) from the Customer Data table – once these are identified, a decision can be taken into how to address the missing/null values.

Data aggregation is the process of grouping data into larger categories for summarisation. For example aggregate daily sales into monthly sales, to identify peak sales. SQL uses GROUP BY key to aggregate data in databases and group records by specific columns. In Python, Pandas library offers the group by function (total, averages etc). For example, in a Sales Data table, converting the Date column into datetime formats allows for monthly aggregation of data which can reveal peak sales periods but also could help in visualising which sales campaigns performed best.

# Section Three - Data Analysis and Visualisation

This section aims to explore the last two stages of the BI flow, also known as data analysis and presentation.

1. Data Analysis is the process of identifying patterns and trends which could lead to decision making; it can be used to predict outcomes in the future, make informed decisions and increase performance.

In a business scenario, a few analysing methods can be used to extract insights from the data. Descriptive Analysis is a method that can be used to understand sales history, market campaign performance and demographics:

**Comparative Analysis** is a method that can be used to compare datasets or categories, for example identifying the top 5 best performing products, sales by discount percentage.

**Time Series Analysis** can be used to assess data recorded between certain dates or at specific points in time, for example to observe in which period marketing campaigns perform best or to understand sales trends by month.

**Predictive Analytics** is a method that can analyse historical trends and forecast future sales, for example, it could forecast the items that are more likely to perform well in the future.

2. Presentation is the process of turning raw data into visual formats to uncover trends and insights to aid in decision making. This step can be achieved by using visuals in the form of reports, dashboards, images, charts etc. Microsoft Power BI is a free service for employees (not employer) - a widely used tool to create interactive reports and dashboards; it offers seamless integration with other Microsoft apps such as Microsoft Office and it is known to be user friendly. Other services to aid in presentation include Tableau (a visual analytics platform transforming the way people use data), Google Data Studio (a free tool turning data into informative dashboards and reports) and IBM Cognos Analytics (an AI-infused business intelligence solution for smarter data-driven decision-making).

# Section Four - Applying Techniques for a Business Scenario

In this section we will be applying the BI techniques at a mid-sized e-commerce company using the given data *Data Sets for Assignment One.xlsx* which contains three tables for sales, customer and marketing data. Below we will explore techniques to extract, transform, analyse and visualise the data using Python as the development language and performing actions such as create a temporary SQL database to perform queries and create visuals such as charts with the help of libraries. Visual Studio Code is the platform in which we will build and run the code.

**Data Extraction:**

The dataset was loaded into Python using *pandas* library and each sheet containing data was labelled according to the data it contains (sales, customer and marketing) and then separated into individual tables.

A screen shot of a computer program

Description automatically generated

A screenshot of a computer screen

Description automatically generated

**Data Transformation:**

In this section a few steps have been taken in cleaning & processing the data and formatting consistently to make the data meaningful.

Standardising ‘Product Name’ values in the sales table to lowercase, making sure the data is all in one format.



Checking for missing values in a few key fields from all three tables, for example counting missing values in Start Date and End Date columns from the marketing table. The output indicates the sum of missing values in all indicated fields are 0.

A screen shot of a computer code

Description automatically generated

A black screen with white text

Description automatically generated

In the ‘Region’ column from the customer table, any missing values will be replaced with the value ‘Unknown’, this is because we can still perform other queries for example revenue, total purchases, total revenue etc. without knowing the value for column ‘Region’.

Any missing values in columns ‘Product Name’ and ‘Quantity Sold’ from the sales table and ‘Start Date’ and ‘End Date’ from the marketing table will be removed from analysis, this is because these are key fields that could lead to misinterpretation or wrong conclusions.

A screen shot of a computer

Description automatically generated

In the final analysis, duplicate values will not be considered, because they can lead to misleading conclusions, as such, duplicate values will be removed from all three tables.

A screenshot of a computer code

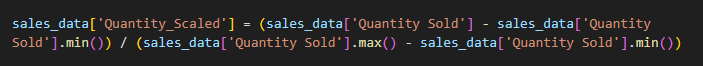
Description automatically generated

From sales table, ‘Quantity Sold’ values are converted to integers and from the values in ‘Revenue (USD)’ any comma (,) or dollar ($) sign are removed, and the values are converted to floats – the same is done for ‘Total Revenue (USD)' in the customer table.

A screen shot of a computer

Description automatically generated

Normalising ‘Quantity Sold’ data from the sales table to ranges between 0 and 1 to prepare the data for consistency in analysis and machine learning algorithms.



A screen shot of a computer

Description automatically generated

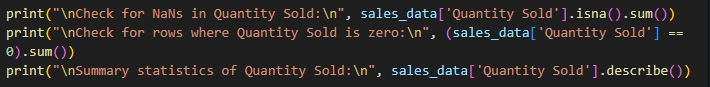
Standardising date formats to *datetime* format to be able to use the data with date specific methods, for example to transform data from days into monthly data (as shown further down when aggregating daily sales into monthly sales).

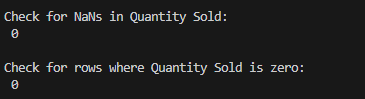
A screen shot of a computer code

Description automatically generated

Check and count null values in the ‘Quantity Sold’ column and zero values in the ‘Quantity Sold’ column

from the sales table, finally display the results.





From the sales table, aggregating daily sales (‘Date’) into monthly sales(‘Month’), extracting month from date and displaying the total quantity per month, total revenue per month and average quantity sold per month in order to understand the sales per month.

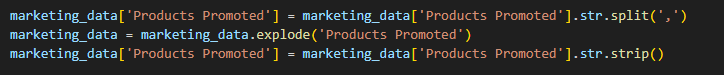
A computer code with many colorful text

Description automatically generated with medium confidence

A screenshot of a computer screen

Description automatically generated

Looking at the ‘Products Promoted’ data in the marketing table, it is noticed that there are sometimes multiple categories of products involved in the marketing campaigns; in order to be able to effectively compare the marketing table with the sales table, we will need to split the ‘Products Promoted’ into one product category per entry allowing us to join the column with the ‘Product Category’ column in the sales table and determine exactly which products sold with a discount. The delimiter used in the ‘Products Promoted’ is comma, so this will be used as a delimiter using the *split* function, *explode* function will be used to split the list into rows and then the *strip* function fill be used to remove any remaining spaces.



**Data analysis:**

The section below goes through creating a temporary SQL database in memory to perform various queries to analyse data.

A temporary SQL database is created using the *sqlite3* library in Python.



The data is loaded into SQL tables using the *.to\_sql* function.

A computer code on a black background

Description automatically generated

Customer retention rate is calculated by looking at a period of one year (calculating the most recent date by using the max() function and subtracting one year by using *pd.DateOffset(years=1) )* function, then the year is halved into two periods of six months including initial customers and returning customers; the retention calculation below checks for any customers from the initial period of six months that are also present in the returning customers from the remaining six months.

A screen shot of a computer program

Description automatically generated



**SQL queries**

1. Query to identify the top 5 products from sales that attract the most revenue

A screenshot of a computer code

Description automatically generated

1. Query to identify purchases by region from customer table:

A screen shot of a computer screen

Description automatically generated

A screen shot of a computer screen

Description automatically generated

1. Query to observe which products sell the best because of marketing campaigns:

A computer screen with text

Description automatically generated

A black screen with white text

Description automatically generated

1. Query to observe which marketing discounts sold more products and the revenue associated with them:

A screenshot of a computer program

Description automatically generated

A black screen with white text

Description automatically generated

1. Query to check the customer who spent the most and their region:

A computer screen shot of a black background

Description automatically generated



**Visualisation:**

In this section below the data is transformed into visualisation using *matplotlib* library to aid in presentation and decision-making.

A pie chart is created using *plt.pie* function to display the percentages of total purchases in each region using the data in the customer table; the pie chart is ideal because it shows the share each region has in revenue percentages.

A screen shot of a computer program

Description automatically generated

A pie chart with different colored circles

Description automatically generated

Checking the sales by discount by looking at the total amount of ‘Quantity Sold’ and ‘Date’ fields in the sale and matching these with fields from the marketing table (‘Start Date’, ‘End Date’ and ‘Discount Offered (%)’); the results are showcased in a bar chart which clearly shows which discount percentage had the best customer response.

A screen shot of a computer program

Description automatically generated

Output:

A graph of sales by discount percentage

Description automatically generated

For visualising the monthly sales trends, a graph chart was created using dual axis: axis 1 showing the ‘Total Sold’ data and axis 2 showing the ‘Total Revenue (USD)’ data. The x axis is attributed to ‘Month’. Colours are used to differentiate between the data: ‘Total Sold’ is set to use a blue line and 'Total Revenue (USD)' is set to displaying a red line graph. Line chart was used in this instance to illustrate monthly sales trends and observe market changes over time; line graphs provide clear view of upward and downward movement in sales, making easy to see growth and decline and utilise forecasting based on historical data.

A computer screen shot of text

Description automatically generated

Output:

A graph with red and blue lines

Description automatically generated

Using similar techniques to display a chart concerning the top 5 most sold products:

A screen shot of a computer code

Description automatically generated

Output:

A graph of different colored bars

Description automatically generated

# Section Five - Insights and Recommendations

After following Business Intelligence best practices, it is concluded that the data as it was provided didn’t include null, zero values or duplicates, however, for good practices and for future analysis, steps should still be followed to ensure the data is relevant.

The analysis discovered there are customers who bring a considerable revenue (such as Lucy Brown, who spent $4800) – while the retention analysis showed that there are no return customers, the business could use loyalty programmes or targeted marketing to retain customers and encourage purchases in the future.

The monthly trends show periods with peak sales and revenue(for example, January, total revenue $14000 which could be due to the New Year Deals campaign) but also periods showing lower sales (February and April); this variation could be assessed in terms of promotional campaigns but also regarding certain months, which could help to plan and predict sales patterns.

Also, from a product perspective, the analysis shows that sneakers and smartwatches sell the best, which could be a reflection of the market interest – this analysis could help the business focus on having inventory available for products in demand; on the other side, it seems that Home Goods had lower sales indicating that there is an area that requires improvement.

When it comes to region, it appears that the Middle East and North America are the leading regions in terms of revenue, however regions like Europe and South America indicate an opportunity to boost sales.

Marketing campaigns such as the Spring Sale (20% discount) sold the most units (especially for Clothing, with a total of 30 units sold), which could predict that future sales campaigns such as Black Friday (50% discount) and Holiday Bonanza (40% discount) could have stronger sales potential, not necessarily higher revenue. Other campaigns like Weekend Specials were not successful, which indicate that it could be due to the low discount offered (only 5% discount).

For future analysis, machine learning algorithms could be used to uncover sales patterns related to products, categories or regions.