

School of InfoComm Technology

**Data Exploration & Analysis Assignment**

Diploma in DS

April 2023 Semester

**ASSIGNMENT 2**

(40% of DEA Module)

10th July 2023 – 13th August 2023

**Submission Deadline:**

**Presentation: 13th August 2023 (Sunday), 11:59PM**

**Report and files: 13th August 2023 (Sunday), 11:59PM**

|  |  |  |
| --- | --- | --- |
| **Tutorial Group** | **:** | **T01 / T02 / T03** |
| **Student Name** | **:** | Akul Arun |
| **Student Number** | **:** | S10243523 |

**Penalty for late submission:**

10% of the marks will be deducted every calendar day after the deadline.

**NO** submission will be accepted after 20th August 2023 (Sunday), 11:59PM.

Table of Contents

[1. Overview 3](#_Toc142864031)

[2. Assignment Part 1: Creating Dashboards to Address Business Questions 4](#_Toc142864032)

[2.1 Data Cleaning in Power BI 4](#_Toc142864033)

[2.1.1 Date in Sales Table 5](#_Toc142864034)

[2.1.2 Store Name in Stores Table 5](#_Toc142864035)

[2.1.3 Promotions Name in Promotion Table 6](#_Toc142864036)

[2.2 Data Modelling 7](#_Toc142864037)

[2.2.1 Calendar Table 7](#_Toc142864038)

[2.2.2 Sorting Tables 8](#_Toc142864039)

[2.2.3 Schema Diagram 9](#_Toc142864040)

[2.3 Feature Engineering 9](#_Toc142864041)

[2.3.1 Profit 9](#_Toc142864042)

[2.3.2 Profit Margin 9](#_Toc142864043)

[2.4 Dashboard 1: Which period has the most sales? 10](#_Toc142864044)

[2.4.1 Dashboard Overview 10](#_Toc142864045)

[2.4.2 Dashboard 10](#_Toc142864046)

[2.4.3 Dashboard Information & Insights 10](#_Toc142864047)

[2.4.4 Measures Created for Dashboard 1 11](#_Toc142864048)

[2.5 Dashboard 2: What are the Top 5 most popular products for the various sales channels? 12](#_Toc142864049)

[2.5.1 Dashboard Overview 12](#_Toc142864050)

[2.5.2 Dashboard 13](#_Toc142864051)

[2.5.3 Dashboard Information & Insights 13](#_Toc142864052)

[2.5.4 Measures created for Dashboard 2 14](#_Toc142864053)

[2.6 Dashboard 3: What are the Top & Bottom 5 Branches in terms of profit? 14](#_Toc142864054)

[2.6.1 Dashboard Overview 14](#_Toc142864055)

[2.6.2 Dashboard 14](#_Toc142864056)

[2.6.3 Dashboard Information & Insights 14](#_Toc142864057)

[2.6.4 Measures created for Dashboard 3 15](#_Toc142864058)

[2.7 Dashboard 4: How can TechScape reduce costs by manufacturer contracts? 16](#_Toc142864059)

[2.7.1 Dashboard Overview 16](#_Toc142864060)

[2.7.2 Dashboard 17](#_Toc142864061)

[2.7.3 Dashboard Information & Insights 17](#_Toc142864062)

[2.7.4 Measures Created for Dashboard 4 17](#_Toc142864063)

[3. Assignment Part 2: Further Analytics 18](#_Toc142864064)

[3.1 Univariate Analysis 18](#_Toc142864065)

[3.1.1 Sales Univariate Analysis 18](#_Toc142864066)

[3.1.2 Profit Univariate Analysis 19](#_Toc142864067)

[3.1.3 Sales Quantity Univariate Analysis 19](#_Toc142864068)

[3.1.4 Product Category Univariate Analysis 20](#_Toc142864069)

[3.2 Bivariate Analysis 21](#_Toc142864070)

[3.2.1 Categorical vs Categorical Variables 21](#_Toc142864071)

[3.2.1.1 Sub Product Category by Product Category 21](#_Toc142864072)

[3.2.1.2 Branches by Region 22](#_Toc142864073)

[3.2.1.3 Stores by Channel Type 22](#_Toc142864074)

[3.2.2 Numerical vs Numerical Variables 23](#_Toc142864075)

[3.2.2.1 Heatmap showing Correlation of Numerical Variables 23](#_Toc142864076)

[3.2.2.2 Correlation of Sales Quantity and Sales Amount 24](#_Toc142864077)

[3.2.2.3 Correlation of Unit Cost and Unit Price 24](#_Toc142864078)

[3.2.3 Categorical vs Numerical Variables 25](#_Toc142864079)

[3.2.3.1 Sales Amount by Product Category 25](#_Toc142864080)

[3.2.3.2 Sales Quantity by Product Category 25](#_Toc142864081)

[3.2.3.3 Employee Count by Store 26](#_Toc142864082)

[3.3 Multivariate Analysis 26](#_Toc142864083)

[3.4 Data Mining 28](#_Toc142864084)

[4. Summary 29](#_Toc142864085)

[4.1 Data Cleaning 29](#_Toc142864086)

[4.2: Creating Dashboards to Address Business Questions 29](#_Toc142864087)

[4.3: Further Analytics 29](#_Toc142864088)

[4.4 Business Implications 29](#_Toc142864089)

[4.5 Thoughts 29](#_Toc142864090)

# 1. Overview

This report will be providing detailed recounts on the actions I took as the member of the Business Analytics team of TechScape Singapore while carrying out an assignment. The main goal of our team is to provide Business Intelligence reports and analysis to assist the management team in making decisions for the company so the assignment that I completed had to achieve this goal. As part of my assignment, I had to use data consolidated by the various departments of TechScape Singapore to create dashboards and perform further analytics to gain useful insights that could help the management team make more informed decisions.

This report will be covering the contents of the dashboards created for TechScape Singapore and it will also be covering content about further analytics based on the dataset from TechScape Singapore. The report will be broken into two parts, one part will be covering the dashboards and the second part will be covering further analytics such as the use of advanced analytical techniques such as univariate, bivariate, multivariate analysis, data mining, etc. Dashboard creation and further analytics will all be done on Microsoft Power BI.

Since this report will be based on a dataset provided by TechScape Singapore, I will introduce TechScape to provide a better understanding of the dataset that will be used. TechScape Singapore which was founded a few years ago is a leading technology company that provides a wide range of electronics and smart solutions. From computers and smart devices to televisions and phones, TechScape offers a diverse range of gadgets to cater to different demands. Currently, the company has a few branches in different regions of Singapore and is planning to build more branches to expand its business operations. Recently, there has been a large increase in the use of online shopping and due to that TechScape has launched their online store so customers can buy their products through several channels.

Now that we have some context on what TechScape Singapore does, let us have a better look at the datasets that I will use for this assignment. I was provided with a total of 8 datasets and 1 csv file for the data dictionary. The 8 datasets were also given in the form of csv files and there are CSV files containing different types of information. The ‘Sale.csv’ file has some of the most important information such as sales amount and quantity sold. There are 3 CSV files, ‘Product.csv’, ‘ProductCategory.csv’, and ‘ProductSubCategory.csv’ contain information about the products sold by TechScape. ‘Stores.csv’ and ‘Geography.csv’ contain information about the names of the stores and their locations. ‘Channel.csv’ contains the data for the different channels that the company sells through and ‘Promotions.csv’ contains information about the promotions offered by TechScape such as the discount offered in each promotion.

# 2. Assignment Part 1: Creating Dashboards to Address Business Questions

## 2.1 Data Cleaning in Power BI

Before creating any visualization, I first had to make sure that the data was clean and that there were no inconsistencies. I also had to make sure that the columns were in the correct data type after loading the datasets into Power BI. I used Power Query Editor in Power BI to clean the data.

Before performing data cleaning, I explored the data by looking at the type of values in each table and the columns in each of the tables.

## 2.1.1 Date in Sales Table

A screenshot of a computer

Description automatically generated

Power BI identified the Date column as a text column when it is supposed to be a date column and I suspected this was due to the underscores so I decided to replace them with slashes as dates are commonly expressed with slashes instead of underscores.

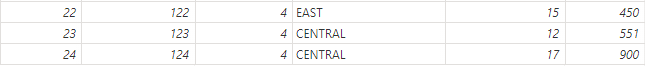
After replacing the underscores with slashes, I changed the data type of the column to Date type.

A screenshot of a computer

Description automatically generated

Now, the values are right-justified and the symbol next to Date also shows a calendar instead of alphabets.

## 2.1.2 Store Name in Stores Table



Store names for the last 3 rows of the store table is the region its from. This information is too general so I decided to represent the data in a more specific manner by replacing these 3 values with the branch where the store is from instead of region by using the geography key which is the second column in the picture above.

A white grid with black text

Description automatically generated

The branch names were in the geography table shown above and the common column between the geography table and the sales table was ‘GeographyKey’.

A screenshot of a computer

Description automatically generated

I created a temporary table where I mapped the ‘Geographykey’ value to the new value I wanted.

A screenshot of a computer

Description automatically generated

I joined the temporary table and the stores table using left join.

A screenshot of a computer

Description automatically generated

Lastly, I merged the columns together and deleted the last column and now I have more specific data values.

## 2.1.3 Promotions Name in Promotion Table





The promotion names are different but they happen on the same days of the different years so I thought that the promotion names should be the same as Power BI will take the 2 promotions to be different if they were spelt differently. Hence, I standardized both values to be ‘Holiday Promotion’.





## 2.2 Data Modelling

As part of the data modelling process, I new tables as I felt that I would require them while doing my dashboards and created the connections between the tables.

## 2.2.1 Calendar Table

I thought that I would be using be doing time intelligence calculations and visualizations related to time intervals a lot to create the dashboards. I thought that I should have the calendar table with different types of time intervals and have separate columns for each time interval instead of having just one date column which would only provide basic functionality. This table would also make it more convenient to create the dashboards. There are many other benefits of using a calendar table including eliminating calculated fields, simplifying date scaffolding, dealing with weekends & holidays, etc.

I did not create a calendar table before, so I had to watch tutorials on creating calendar tables. It was a tricky process as I had to use functions that I did not use before but soon I fine-tuned the code that produced a working calendar table.

Code for calendar table:

A close-up of a computer screen

Description automatically generated

Code for calculated columns added to calendar table:

1) SeasonOfYear



2) DayOfWeekType



## 2.2.2 Sorting Tables

When creating visualisations, the days and months and years were not sorted in the format that I was looking for. Hence, I created 2 separate tables, one for day and one for month and I mapped each of the day name and month name to the index I wanted to sort them by for custom sorting.

1. SortedDayOrder

A screenshot of a computer

Description automatically generated

2. SortedMonthOrder

A screenshot of a calendar

Description automatically generated

After creating these 2 tables, I had to change the sorting of the column to ‘Order’ in both tables and I had to use the ‘DayOfWeekName’, ‘MonthName’ and ‘Year’ from these tables when creating my visualisations in order to have my visualisations sorted as I wanted.

## 2.2.3 Schema Diagram

A screenshot of a computer

Description automatically generated

The Sale table is the fact table which is why it is in the middle, and which is why it has the greatest number of connections with other tables. All the tables are linked to each other through their common columns, mostly their primary keys. For example, Promotion Key links the Sale and Promotions table with a 1:Many relationship because Promotions Key is the primary key of the Promotions table and is the foreign key of the Sale table.

## 2.3 Feature Engineering

Feature engineering is the development of new data features from existing data. With this, I could obtain other types of useful data that I could use for my dashboards to show more insights.

For this, I first needed to have a better understanding of the dataset to find out the potential existing data I could use to create new features. These were some general features I created that could be useful in multiple scenarios.

## 2.3.1 Profit

The first feature that I could create was profit. Profit is the difference between the amount earned and the amount spent. I had 2 existing columns in my Sale table called ‘SalesAmount’ for amount earned and ‘TotalCost’ for amount spent. Hence, I created a new column in my Sale table called Profit using the formula, Profit = Sale Amount – TotalCost.



## 2.3.2 Profit Margin

Profit Margin is a common measure of the degree a company makes money. Profit margin is a financial ratio that measures the profit earned by the company in relation to its revenue expressed as a percentage. I created a calculated column for profit margin in the Sale table and changed the format of the values to percentage from numeric.



## 2.4 Dashboard 1: Which period has the most sales?

## 2.4.1 Dashboard Overview

This dashboard focuses on the analysis of the timeline of sales so the important aspects of this dashboard are sales and time. By determining which period has the most sales, the seasonality of sales can be identified, and this will optimize management of inventories, staffing and timing of appropriate business campaigns.

## 2.4.2 Dashboard

A screenshot of a computer

Description automatically generated

## 2.4.3 Dashboard Information & Insights

For the time series aspect of this dashboard, I focused on year, quarter, season and month. I decided to choose year as the global filter for this dashboard as year is the highest level of the time measure available and can be used to filter month and season. Using the cards at the top, I showed the months with the most sales and least sales. A part of the dashboard on the left, below the filter, compares the differences between the peak and off-peak months. and the bottom, I compared the seasons.

While comparing the seasons, I was surprised to find out the difference in the performance of months of the same season. For example, for the autumn season, the sales performance across November is significantly higher than the other 2 months. The same can be seen for December in Winter. This shows that our sales tend to pick up at the end of the year. To maximize revenue around this period, the company can increase the selling prices of the products. The company should also improve marketing for the month of March as the sales performance of it is extremely weak compared to April and May in Spring.

## 2.4.4 Measures Created for Dashboard 1

I could not filter the month in the cards by topN so I had to create custom measures. I created 2 measures, ‘Peak Month’ and ‘Off Peak Month’.

1. Peak Month

A computer screen shot of a program

Description automatically generated

The code basically filters the top month with the greatest amount of revenue. Hence, I used MAXX function.

2. Off Peak Month

A screenshot of a computer program

Description automatically generated

The code for off peak is like that of peak month but the difference is now I want the bottom month with lowest revenue which is why I used MINX function here.

## 2.5 Dashboard 2: What are the Top 5 most popular products for the various sales channels?

## 2.5.1 Dashboard Overview

Through this dashboard, I aim to identify the most appropriate target products for TechScape to efficiently and effectively use marketing activities to engage customers.

This dashboard focuses on the number of products sold for each channel and compares the different channels, so the aspect that this dashboard focuses on is the number of products sold. Hence, I have enabled the channels to be the top-level filter for more effective comparison and the visuals show the effect of various types of factors that can affect product sales.

## 2.5.2 Dashboard

A screenshot of a video game

Description automatically generated

## 2.5.3 Dashboard Information & Insights

The global filter for this chart is the year filter and filters all the charts. I also added a filter for channel type for exploring each channel type in depth, but this filter does not affect the 4 funnel charts.

The funnel charts show the top 5 highest selling products for each of the channel types. As seen from the dashboard, cellphone accessories are the highest selling products by a large margin for all 4 channels.

After comparing the area charts, we can see that April and October have the biggest return quantities for stores even though the number of products sold is like the other months. TechScape should investigate why this is and solve this.

For all the channel types, there is less than 1% of products returned out of the number of products sold, which is good as it shows high customer satisfaction. The channel type with the highest number of returns is catalog with 0.9% of products returned.

A blue circle with white text

Description automatically generated

As seen from the product performance table at the bottom right, camcorders generate almost 10,000 dollars per customer on average and have a higher profit margin than the other products. However, it is only the third most popular product across all the channels. Hence, TechScape should try to promote camcorders more to increase revenue.

## 2.5.4 Measures created for Dashboard 2

Most Popular Product Category:



To find the product category with the greatest number of items sold, I used the TOPN function and chose the first value when ordered in descending order.

## 2.6 Dashboard 3: What are the Top & Bottom 5 Branches in terms of profit?

## 2.6.1 Dashboard Overview

This dashboard monitors the top and bottom 5 branches in terms of gross profit, enabling TechScape to make informed decisions, optimize resources, drive important, and align its strategies to maximize overall sales and profitability.

This dashboard focuses on the profits for the top and bottom five branches, so the aspect that this dashboard focuses on is profit and the features of the branches. I have created a top-level filter to easily filter the visuals to show information about the top 5 branches or the bottom 5 branches.

## 2.6.2 Dashboard

A screenshot of a computer

Description automatically generated

## 2.6.3 Dashboard Information & Insights

TechScape is also looking into expanding their business operations so this dashboard will provide insights on where to open the branch locations and what kind of branch to open such as the region and the area size of the branch to build.

I have also focused my attention on the number of customers and the number of employees for each of the branch locations as these factors can affect the profit earned by each branch location. Gaining insights about employees can help with the planning of human resources.

## 2.6.4 Measures created for Dashboard 3

Top and Bottom Filter:

This filter is one of the most complex filter I have created and it involves the use of 5 measures and 1 table.

Table

1) Top&Bottom Table

The table is a simple table just to create the filter names.

A screenshot of a computer

Description automatically generated

Measures

1) SelectedTopNumber

A screenshot of a computer program

Description automatically generated

This measure defines the number of branches to filter by. In this case, my question focuses on top 5 and bottom 5 branches which is why the number is 5.

2) BranchRankByProfit



This measure defines how the branches are filtered and I have chosen to filter them by gross profit as the question is interested in the profit of the branches.

3) Top5Branches



This measure is for selecting the top 5 branches.

4) Bottom5Branches

A screenshot of a computer code

Description automatically generated

This measure is for filtering the bottom 5 branches.

5) BranchCombinedFilter

A screenshot of a computer code

Description automatically generated

This measure combines the 2 measures of selecting top 5 and bottom 5 branches and maps the measures to the filter names that I previously created in the Top&Bottom table.

## 2.7 Dashboard 4: How can TechScape reduce costs by manufacturer contracts?

## 2.7.1 Dashboard Overview

This dashboard compares the performance of manufacturers in terms of sales and number of products sold. This can help TechScape determine how costs can be reduced either by changes in manufacturer or changes in manufacturer products.

The top half of the dashboard compares the yearly change in sales and number of products sold by each manufacturer. Then, we can see if the products of the manufacturers are able to retain customers through the years which will signify a good manufacturer.

At the bottom, we can learn more about a specific manufacturer by comparing the different products that are sold by the manufacturer. By comparing the number of products sold and number of products returned of each type, we can determine the product quality. By comparing the unit cost and unit price of each product, we can determine which products we are overpaying for.

## 2.7.2 Dashboard

A screenshot of a computer

Description automatically generated

## 2.7.3 Dashboard Information & Insights

We can see that the number of products sold by SY. Ltd is significantly higher than the other manufacturers and it is also increasing with each year, signifying that their products can attract more users. Even though SY. Ltd sells the greatest number of products, EvoTech. Inc Generates the most revenue for us. However, the revenue has been dropping over the years for EvoTech. Inc which is not a good sign. Additionally, AstraCell has been losing revenue over the years and the number of AstraCell products sold has been decreasing which means that AstraCell products are not able to retain customers. Hence, I think we can reduce costs by purchasing less AstraCell products.

When exploring SY. Ltd, I found that even though over cell phones sold the best, the number returned is one of the lowest which means that their cell phones are of very good quality, and we can explore more dealings of cell phones with SY. Ltd. We can also try generating more revenue by increasing the selling price of cell phones because the average selling price of SY. Ltd cellphones are only $15 when the average unit cost is $7 so we can look to increase this gap.

## 2.7.4 Measures Created for Dashboard 4

1) Most Profitable Manufacturer:



I used a TopN filter to get the most profitable Manufacturer by referring to the profit margin of each manufacturer.

2) Return Rate:



I used mathematical calculations to obtain this measure by dividing the number of products returned by number of products sold. Then, I formatted this measure as percentage.

# 3. Assignment Part 2: Further Analytics

## 3.1 Univariate Analysis

## 3.1.1 Sales Univariate Analysis

A screenshot of a computer

Description automatically generated

Firstly, I analysed sales which is one of the most important variables that I have used. It is a numerical variable which is why I have chosen to show its distribution using a histogram. The average amount sales made by the company is around $1767. The distribution is skewed to the right as seen by the positive skewness value from the table. The largest amount of money TechScape has received in a single transaction is $5345.24 while the smallest amount of money is $2.44.

## 3.1.2 Profit Univariate Analysis

A screenshot of a graph

Description automatically generated

Another important measure for businesses is profit. It is a numerical variable which is why I have chosen to show its distribution using a histogram. The average amount profit made by the company is around $592.80. The distribution is skewed to the right as seen by the positive skewness value from the table. The largest amount of profit TechScape has received in a single transaction is $1596.15 while the smallest amount of profit is $0.76.

## 3.1.3 Sales Quantity Univariate Analysis

A screenshot of a graph

Description automatically generated

Another important measure for businesses is sales quantity as it determines the popularity of the company. It is a numerical variable which is why I have chosen to show its distribution using a histogram. The average number of products sold by the company is around 424. The distribution is extremely skewed to the right as seen by the large positive skewness value from the table. The largest number of products TechScape has sold in a single transaction is 2880 while the smallest number is 4.

## 3.1.4 Product Category Univariate Analysis

A blue pie chart with white text

Description automatically generated

I also analysed product category in greater detail to further understand the types of products offered by TechScape. The pie chart shows the distribution and we can see that there is quite a big gap in the distributions of the different product categories. More than half of the products offered by TechScape are computers, and cameras and camcorders. TechScape offers little music, movie and audio products and they only make up about 10% of products sold by TechScape.

## 3.2 Bivariate Analysis

## 3.2.1 Categorical vs Categorical Variables

## 3.2.1.1 Sub Product Category by Product Category

A screenshot of a computer

Description automatically generated

The treemap shows the composition of the number of sub product categories that are part of each product category. This treemap will help us have a greater understanding of the products that we are selling, and we can see the types of products that TechScape to look to include in the future. For example, there are 8 sub-categories for audio, computers and home appliances each but there are 3 product sub-categories for games and music. Hence, if TechScape expands the variety of products that we are selling, we can look into getting more types of products for games and music.

## 3.2.1.2 Branches by Region

A graph with blue rectangular shapes

Description automatically generated

The bar chart shows the number of branches that are part of each region. This chart helps us have a greater geographical understanding of our branches and we can see the region where we can build future stores at. For example, there are already 7 branches in the west, but the north region only has 3 branches. Hence, if TechScape wants to expand their business operations, we should focus on the northern region.

## 3.2.1.3 Stores by Channel Type

A screenshot of a computer

Description automatically generated

The bar chart shows the number of stores that are part of each channel type. This chart helps us have a greater understanding of our selling methods and having a deeper understanding of our selling methods can help us increase sales. For example, TechScape already has 17 stores for the store channel type but there are only 3 online stores. However, in the modern era, there is a rise in online shopping and 3 stores might not be enough for TechScape. Hence, we should set up more online stores to keep up with the current trends.

## 3.2.2 Numerical vs Numerical Variables

## 3.2.2.1 Heatmap showing Correlation of Numerical Variables

A screenshot of a computer screen

Description automatically generated

The top 3 correlations that we can observe are between SalesAmount & UnitCost, SalesAmount & Unit Price, and SalesAmount & TotalCost. Through looking at the correlations, we can find the variables that have the highest correlation with sales and profit and with a deeper understanding of sales and profit, we can more easily increase them. Profit has a high correlation with unit price, so to generate more profit, we can increase the price of products.

## 3.2.2.2 Correlation of Sales Quantity and Sales Amount

A graph of sales

Description automatically generated

I created a scatterplot to further investigate the relationship between sales quantity and sales amount. This is because in the heatmap, the correlation between sales amount and sales quantity correlation was only 0.071, which was very low. I expected these 2 variables to have a high correlation. The scatterplot also shows that there is a weak positive relationship between sales quantity and sales amount. I expected these 2 variables to have a positive relationship as sales quantity increases, the sales amount will increase. We can understand that even though sometimes the sales quantity is low, the sales amount can be significantly higher.

## 3.2.2.3 Correlation of Unit Cost and Unit Price

A graph of a graph with numbers and a line

Description automatically generated with medium confidence

I created scatterplot to see if there were any outliers present for the unit cost and unit price of any of the products. I was looking for products that had high unit costs and low unit price or low unit cost and high unit price, the latter being the more favorable one to see which product can bring us more profit. As seen from the scatterplot, products with higher unit costs tend to have a higher unit price. There are some markers around the middle of the scatter plot that have similar unit costs but different unit prices.

## 3.2.3 Categorical vs Numerical Variables

## 3.2.3.1 Sales Amount by Product Category

A graph of sales and sales

Description automatically generated with medium confidence

This information will aid TechScape in understanding the performance of the product categories. This way TechScape will know what products are most capable of driving in more sales. We can see that computers and cameras have significantly greater sales than the other product categories, both exceeding 1 billion dollars in sales. Hence, TechScape can focus on these product categories to maximise revenue.

## 3.2.3.2 Sales Quantity by Product Category

A graph of products sold of each product category

Description automatically generated

This chart also helps TechScape have a better understanding of product performance, but now in terms of quantity sold. With knowledge of the quantity of products sold, we can determine the popularity of our products. We can see that cell phones are our top selling product category but it ranks fourth in terms of revenue. I think that we can generate more revenue using cell phones and one strategy is increasing the prices of the cell phones.

## 3.2.3.3 Employee Count by Store

A screenshot of a computer

Description automatically generated

The treemap allows us to have more knowledge about our employees and stores and this can help with human resources and planning. For example, now that we know that Queenstown only has 12 employees and Bedok Castle Store has 47 employees, when we hire new employees, we should focus on assigning them to the Queenstown store.

## 3.3 Multivariate Analysis

A graph of colored dots

Description automatically generated with medium confidence

As part of multivariate analysis, I have done PCA to better understand the relationship between AreaSize, EmployeeCount, SalesAmount, SalesQuantity coloured by BranchRegion. As seen from the chart above, there are apparent clusters being formed for the different branch regions. The clusters of West and Central branches formed around each other. There is a cluster for Northeast that is isolated around which does not have any other markers around it. All the clusters show a downward trend, when pc1 increases, pc2 decreases.

To ensure that all variables are on comparable scales, we first standardize the data using the Standard Scaler. The dimensionality is then reduced to two main components using PCA. The colours of the marker show clearer distinction and easy recognition of patterns.

A blue line with black text

Description automatically generated

Additionally, I created a line chart to better understand the effect of the PCA conducted. As expected of PCA dimensionality reduction, the percentage of variance explained decreases as the index number of components increases. The first index started with 40% variance explained and the ending index has around 28% variance explained.

Using multivariate analysis, especially methods like PCA, offers a reliable method for obtaining insightful information from complicated and high-dimensional data. We obtain a thorough understanding of the fundamental patterns and linkages within the data by looking at numerous variables at once. This enables us to make data-driven decisions that eventually improve business outcomes and successfully steer strategic undertakings.

## 3.4 Data Mining

A screen shot of a graph

Description automatically generated

As part of data mining, I decided to perform K-Means clustering as I wanted to learn something new. For the K-Means clustering I decided to compare the sales amount and quantity sold for each specific product, which is why I used product description. Sales amount and sales quantity characteristics offer insightful data about each product's demand and sales performance which can help with inform marketing tactics, inventory management, and product promotion.

For the clustering, I decided to go with a cluster number of 4 why is why there are 4 clusters being shown for the K-Means Clustering scatter plot. For the scaling method, I used the MinMax scaler. On the scatter plot, each marker represents each product description, and they are colour coded according to their cluster. Looking at cluster 1 which is colour coded in green, the products have low quantity sold but they have generated a large amount of sale but the products in cluster 2 which is colour coded in red, there are high quantities sold but they do not generate much revenue. Hence, we should focus on improving the sales of the products in the low scale clusters, the clusters at the left side of the scatterplot.

By employing K-means clustering, we have gained an enriched understanding of the product sales landscape at TechScape Singapore. This approach enables us to categorize items based on their sales performance, revealing concealed insights and trends that can inform astute decisions by the company's leadership. Through this method, TechScape Singapore is now empowered to enhance its product portfolio, elevate customer interactions, and proficiently allocate resources based on the insights derived from the K-means clustering analysis. These findings contribute to fostering the company's growth and accomplishments as it expands its business horizons.

# 4. Summary

The Report explores data analytics in great detail, using a variety of methods to glean insightful information that can guide strategic choices. I have highlighted the key points below.

## 4.1 Data Cleaning

This part describes how to find and fix errors, including duplicate promotions, changes to the data structure, and the establishment of a calendar table for time-based analysis.

## 4.2: Creating Dashboards to Address Business Questions

Four dashboards are created in this section, each of which addresses a different business question. The dashboards are made to offer information on peak selling times, top-selling items across all modes of distribution, branch profitability, and top-selling manufacturers. The organization acquires a better grasp of its sales trends, product performance, branch profitability, and manufacturer partnerships.

## 4.3: Further Analytics

Advanced analytics methods are covered in more detail in the second section. To find hidden patterns and correlations in the data, univariate, bivariate, and multivariate analysis are used. Bivariate and multivariate analyses use Principal Component Analysis (PCA) to show intricate interactions among several variables, while univariate analysis focuses on understanding the distributions and patterns of individual variables. K-Means clustering was also done as part of data mining.

## 4.4 Business Implications

This assignment emphasizes how important data visualization is in converting complicated datasets into clear insights. TechScape Singapore is able to make well-informed decisions, increase profitability, and benefit stakeholders because of the dashboards and analysis built.

## 4.5 Thoughts

Overall, I am grateful that I have done this assignment as I have learnt a lot from this assignment and this assignment included a lot of trial and error, especially when creating the dashboards.