Data Processing

Mina Kakhodaverdi

Contents

[Introduction 2](#_Toc114124815)

[ What is your project? 2](#_Toc114124816)

[ What will you be doing? 2](#_Toc114124817)

[ Why are you doing this project? 2](#_Toc114124818)

[Data Gathering 3](#_Toc114124819)

[ What data source will you use? (Primary or Secondary) 3](#_Toc114124820)

[ What are benefits of primary or secondary data source? 3](#_Toc114124821)

[ What drawbacks of primary or secondary data source? 3](#_Toc114124822)

[ How are you compliant with GDPR and/or DPA? (Does your data contain personal data that can identify a person? If so, how will you ensure that data is protected?) 3](#_Toc114124823)

[Data Cleaning 3](#_Toc114124824)

[A) Identify problems with data (give examples where you found the following issues) 3](#_Toc114124825)

[ Missing items or columns of data 3](#_Toc114124826)

[ Duplicate data records 3](#_Toc114124827)

[ Transposed data into different columns 3](#_Toc114124828)

[ Calculation errors 3](#_Toc114124829)

[ Symbols within fields 3](#_Toc114124830)

[ Different data scales within the same data columns 3](#_Toc114124831)

[B) How did you resolve the issues identified in 3](#_Toc114124832)

[Data Summary 3](#_Toc114124833)

-Introduction

Babington Supplies Ltd has submitted a data collection for analysis in order to provide insight on a number of important business areas that covers a wide range of data from the business order system, including information about the customers, orders, products and geographic details. The data anylising will help in future planning and desiocion making in order to grow.

The aim of data analysis is to provide information on the following areas;

- The highest selling and lowest selling products.

-The highest number of sales, revenue, and profit made in different categories.

-The most successful sub-category of product and customer segment in terms of highest product sales, profit, and the geographical success of products in terms of highest product revenue.

-Data Gathering and Classification

The provided data was structured in an Excel format, including quantitative and qualitative data, such as numbers, prices, order dates covering four years. Additionally, qualitative nominal data in the form of product names, categories and sub-categories, customer segments, and geographical locations were included. Since the data was collected from the company's order system, it is considered secondary data that was gathered to fulfill customer orders.

Compliance

The data provided was ethically, morally, and legally assessed for the analysis objective, as it contains personal information that could identify individuals or groups of individuals. The company has also obtained specific consent from its customers in compliance with GDPR and the Data Protection Act 2018, ensuring that any customer data without consent has been removed from the data set.

The data is stored securely in the organization's IT systems and is processed, shared, and destroyed in compliance with the legal regulations. The company Privacy notice outlines the retention period and reasons for retention, and the data is destroyed accordingly. Access to the original data is restricted to authorized personnel within the organization.

To ensure data accuracy and completeness, the company carried out processes to remove any inaccurate or incomplete data. Moreover, the data used has been processed to minimize duplication of personal information.

-Primary and secondary data

Both primary and secondary data sources have their own benefits in data analysis.

Benefits of primary data source:

Relevance: Primary data is collected specifically for the research question, which makes it more relevant to the research objectives.

Control: The researcher has complete control over the research process, including the type of data collected, the method of collection, and the timing of data collection.

Accuracy: Since the researcher is directly involved in collecting primary data, the possibility of errors or bias is reduced, resulting in more accurate data.

Benefits of secondary data source:

Time and cost savings: Secondary data is often readily available and can be accessed quickly and at a lower cost than collecting primary data.

Large sample size: Secondary data sources may provide a large sample size, which can increase the statistical power of the analysis.

Historical data: Secondary data sources can provide historical data, which can be useful in studying trends over time.

- Disadvantages of primary or secondary data sources

Both primary and secondary data sources have some potential drawbacks.

Drawbacks of primary data sources may include:

Time-consuming and expensive: Collecting primary data can be a time-consuming and expensive process, especially if a large sample size is needed.

Limited scope: Primary data is often collected from a small sample size, which may not accurately represent the larger population.

Biased results: The results of primary data collection may be influenced by factors such as researcher bias, leading to biased results.

Incomplete or inaccurate data: There is a risk of incomplete or inaccurate data if the data collection process is not carefully designed and implemented.

Drawbacks of secondary data sources may include:

Lack of control over data quality: Since the data has been collected by someone else, there may be issues with data quality, accuracy, or completeness.

Data may not be relevant: Secondary data may not be relevant to the specific research question being investigated.

Data may not be current: Secondary data may be outdated, which can be a problem for certain types of analysis.

Data may not be accessible: Some sources of secondary data may be difficult or expensive to access, which can limit its usefulness for analysis.

-Data Cleansing

The data was cleaned before analysis was carried out to ensure its accuracy and reliability. This involved several steps:

-Removing null (blanks) or missing values from the Customer ID, Ship Date, Ship Mode, Segment, State, Postal Code, Regions, and Category fields to prevent issues with linking tables and loading data into pivot tables and charts.

-Duplicate text in the Category and Sub-Category fields was removed to prevent inaccurate grouping of data or slicers.

-Inconsistent data scales in the Country and City fields were removed to prevent false and misleading results from analysis.

-The Order Date and Ship Date fields were standardized to a consistent date-only format to ensure accurate and consistent analysis of timeline data.

-The data was separated into three separate tables showing Customers, Products, and Sales, and all duplicate data was removed to ensure table relationships would work and pivot tables and charts would load correctly.

-The Customer ID and name fields were merged and then transposed into separate columns to enable correct and accurate identification of customers.

-Duplicate data records were removed from the Sales table using the unique identifier of Order ID to establish accurate table relationships.

-The data was checked for any unnecessary symbols within fields and removed to prevent calculation errors or missing values.

-Calculation errors in the amounts shown for discounts and total sales were identified and recalculated to prevent false results in analysis of sales and revenue.

-Data Summary

-Key Metrics

The chart below shows how the company is performing.

The total sales value( £1,967,991.74), total profit(£285,155.10), highest sales value(£17,499.95), lowest sale value(£0.09) and the number of transactions(9946) .

-The Revenue over the 4 years

Table, Excel

Description automatically generated

This graph shows the total profit that’s business has made in each year. The total profit over 4 years is £285,155.10 and the most profitable year is 2017 with £92,689.86. this will help the business to get insights as to improve the business for the upcoming year.

-The monthly revenue

Chart

Description automatically generated

This chart shows the monthly sales. The most sales has been made in November with £300,512,27 following by December with £282,289,54. The sales has been down in February and Januar. This can be improve by promoting the products to improve the sales of that months.

-Profit in different regions

The chart and table below shows the sum of profit in different regions.

Graphical user interface, table

Description automatically generated with medium confidence

This information gives the company an understanding of which regions are generating the most profit. By analyzing this data, they can focus their efforts on regions that are generating the most profit and explore ways to increase profit in regions that are not performing as well. As you can see the central region is the least profitable one and the west region is the most profitable.

-Contribution of each sub-category

Graphical user interface, chart, application, table, Excel

Description automatically generated

In this chart you can see the contribution of each sub-category in total salesover the years. As you can see the biggest contribution are phones with 14.18% contributution to total sales follwog by chairs with 14.08% and storage with 10.67%.

The lowest contributor are Office Supplies and fasteners with 0% contribution to total sales.

**-Correlation between Discount Amount and Sales Values**

|  |  |  |
| --- | --- | --- |
| **Row Labels** | **Total Discount Amount** | **Total Sales Value** |
| First Class | £51,706.84 | £298,930.10 |
| Normal | £3.25 | £13.02 |
| Same day | £20,091.98 | £108,084.17 |
| Second Class | £55,621.01 | £402,248.51 |
| Standard | £193,989.49 | £1,158,715.94 |
| **Grand Total** | **£321,412.57** | **£1,967,991.74** |
| |  |  | | --- | --- | | **Correl** | **0.99718** | |  |  |

The table and chart above show the correlation between sales and discount amuount. As you can see the correlation is almost 0.997 which shows that the total sales for each shipment mode is directly affacted by discount amount. The company can use the information on the correlation between discount amount and sales value to make informed decisions on pricing and discounting strategies to increase sales.

-Top 10 customers

Timeline

Description automatically generated

The company can increase the profit margins by targeting the most profitable customers. As you can see above the customer by customert ID TC-20980 is the customer that has the highest sale. Thr company could generate more sales by offering a subscription service to these top 10 customers and reminding them to buy a new product regularly.

-Bottom 10 products

As you can see these products leasted below in the tables are not delivering good returns in total sales and profit. In the right side table you can see the prodicts that have a good number of sales but they are not profitable.

Graphical user interface, text, application, table, Excel

Description automatically generated

This data will give the company the chance to think about how they can increase sales for these products. For example, the company can promote the products through social media, offer discount or a delayed payment plan. Also, by finding out what products are not delivering good returns they can put their focus on the more profitable products.

**Conclusion and Recomodations**

Base on data analysis provided, we can conclude that the company has achived a total sales value of £1,967,991.74 and the total profit of £285,155.10. The revenue over the four years shows that the most profitable year was 2017 with £92,689,86. This information can be used by the company to improve its business operation for the upcoming year.

The monthly revenue shows that the company had the most sales in November and December however the sales were down in January and February. To improve the sales during these months the company can promote its products to boots sales. This include offering discounts or advertising to increase consumer demands during these months.

Analyse the customer data to identify patterns and preference which can be used to tailor marketing campaigns and products offering to better meet the needs of the customers.

From the data analysis, it is evident that phones, chairs, and storage make the highest contribution to total sales, However, the office supplies and fasteners did not make any contribution to total sales. The company can focus on promoting and increasing the sales of the top-performing sub-categories, while also considering ways to boost the sales of sub-categories like office supplies.

In conclusion, the company can also use the information from the top 10 customers and bottom 10 products to optimize its sales and increase its profits. By focusing on the most profitable customers and products, and taking steps to increase sales of less profitable products, the company can improve its overall performance and stay competitive in the market.

The company can consider adjusting the price to make it more competitive in the market. They can also Improve the product presentation by making sure that the product is presented in an appealing way to attract customers. They can use targeted advertising and social media platforms to promote the product. They can also offer the product as part of a bundle or package deal, or provide discounts to customers who purchase a certain quantity. This can incentivize customers to buy more of the product, increasing sales. Providing free samples of the product can also help customers try it out and develop an interest in it.

By implementing these strategies, they can improve the sales of their low-performing products and increase overall revenue.