Data Processing

Your name Blanca Martinez Perez

Introduction

# Data Gathering ………………………………………………………….………………………….. 3

# Data Cleaning. ……………………………………………………………………………………… 4

# Data Summary. …………………………………………………………………..………………… 4

My 10 project ideas.

1. Identify of what is the segment that consume Babington product? ……………………….5
2. Which are the 5 best seller products? …………………………………………..………..….5
   1. Correlation of trends on the sales …………….………………..…………….………. 6

2.2 Correlation of trends on the profit …………………………….…………….……..… 7

2.3 Hipotesis to analyse the profit for the top 5 products…………….…….…….…….. 8

1. Which are the 10 bottom list products? …………………………………………………….. 8
2. Identify months of slow sales …………….….…………………………………………….… 9
3. When is the highest profit by sales? …………………………………………………….….. 11
4. Which are the top 5 States that generates more sales? …………….……………….…… 12
5. Which are the top 5 cities that bring more sales? ……………………………………..12, 13
6. Which state and region generate less sales? …………………….…………………………13
7. What is the most popular shipping method? ……………………..………………………… 14
8. What is the shipping method for the best seller by city? …………………..……………… 14

CONCLUSIONS for my all projects ideas…………… …………………………………………15, 16

SUMARRY of my analysis. ……………………………………………………………………...…. 17

# Introduction

## What is your project?

My project is based in Babington suppliers

## What will you be doing?

I will analyse my project ideas. The main ideas of my project are based to understand the sales in relationship with the product most or the least consumed, the period of time where sales are higher or the quietest, where are the potential customers of Babington suppliers and finally the shipping method most popular among the consumers.

## Why are you doing this project?

To find solutions or improve strategies that can contribute to increase the sales in the company

# **Data Gathering**

## What data source will you use? (Primary or Secondary)

Secondary source

## What are benefits of primary or secondary data source?

The primary is collected directly from the data source, so the benefits means that the data is more accurate than the secondary data. That means that the researcher is the ownership of the data and he or she can make it available publicly or even sell it.

Whereas the secondary data is the data that has been collected by a third party, so it it easily accessible to the researchers compared to primary data, which the time to collect it it is little compare to primary data, make it possible to carry out long studies without having to wait a long time to make conclusions and it helps to generate new insights to primary data.

## What are drawbacks of primary or secondary data source?

The main disadvantage of the primary source it is difficult to collect and required to invest a lot of time due to its complexity and commitment.

The main disadvantage of secondary source is that the data may not be authentic and reliable, so researchers may need to deal with irrelevant data before finding the required data.

## 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 the data is protected?)

Because DATA is complying with UK GDPR.

I have used for my project customer ID to represent customer rather than use customer name. In this case Babington have obtained the permission from the customers to use the data in line with GDPR and the Data Protection Act 2018. Babington confirm that they follow the data procedure regulations to stored and process the data.

# Data Cleaning

## Identify problems with data (give examples where you found the following issues)

First of all the data has been cleaned through the tool Power query editor to tidy up all the information to avoid work with errors as remove the error column by column and apply other criteria to be able to apply formulas.

For that I have proceeded to do the next steps:

### Missing items or columns of data

Null or missing were found in Customer ID, Ship Date, Ship Mode, Segment, State, Postal Code, Regions, Category

### Duplicate data records

Duplicates were founded mainly in customer ID and also in the post code. They were also founded in the category and sub-category data. They were removed from the data. The benefit of applying pivottables it is that you make sure that the duplicates are completed removed from the data. I have eliminated duplicates by using pivottables.

### Transposed data into different columns

### Customer ID and customer name has being separated by columns. Specifically in the column customer name the symbol dash has been deleted. Order date and ship date has been both standarised as a date to have the same format.

### Calculation errors

Errors have been found so have been subsequently deleted. To do that empty records have been removed all the empty records by columns. The majority of errors were detected in the postcode category.

### Symbols within fields

I haven’t found symbols within fields.

### Different data scales within the same data columns

There were different formats in various columns. The main column where different formats were founded were in ship mode for example data in First Class comes with different formats. The column need to have standardization otherwise they are counting as two separate things . We also founded on country and city. All of this data has been removed before the analysis.

## How did you resolve the issues identified in (A)

# Data Summary

The next data analysis will contain a summary of my ten projects ideas that I will show point by point.

The next 10 points are based on my project idea explain each one with an explanation of the data analysis to know more about the highest or lowest selling and best seller or less popular products and so on, variables that help us to understand more the number of sales.

**My 10 project ideas.**

**1. Identify of what is the segment that consume Babington product?**

**Chart, pie chart

Description automatically generated**

**WHY?**

This point it is benefit for the company to know who are the public that are interested to buy the company’s products.

The public of Babington is divided by consumers, corporate and Home Office, being the consumers the amount who represent the main target for Babington providing half of the profit of sales. Corporate is the second target contributing a 30% of the total sales and Home Office just a 19% of the sales.

1. **Which are the 5 best seller products?**

**Chart

Description automatically generated**

**WHY?**

* 1. This information is good for the company to keep on top of selling the same category and specific products. To help working in the same line and to find out what makes them so attractive, also to follow the same line of products when it comes to research in new products to launch for the future.

The total of products that generate higher benefit in the company are copiers on the top followed by phones, accessories, paper and blinders.

Descriptive statistics

After apply a descriptive statistics about the top ten products, the data gave me a reflection

on my hypothesis for the top 10 products. With this tool, I was able to analyse the total sales value in comparison with the profit after selling this products. The data tells me that on average the total amount of sales of the top ten products is around 849.000£ with a total profit of 21.000£. It also tells me that sales total shows that the data is positively skewed whereas the skewed of analysis in the profit shows that the data is negatively skewed. That aspect is something to have in consideration in my hypothesis. For that reason, I would like to make a hypothesis to prove if selling less products, the profit amount in relationship with the sales still have a positive value.

First of all, I identify correlation of trends for the top 10 products. For that, I calculated the trend based on the sales and the trend based on the profit to get the next conclusions.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **2.1 CORRELATION OF TRENDS on the SALES** | |  | | | |  | |  |  |  |  |
| In the table above we can see a pick in sales for phones and another one in blinders.  **2.2 CORRELATION OF TRENDS on the PROFIT**    The table above shows the trendline in terms of the profit based on the sales of the top 10 products. What I have identified here is that the graphs represent up and downs in profit in relationship with the products that generates more and less profit, which we can see there are sales picks in phones, accessories, and copiers, but take down to 0 the number of profit for tables, machines and bookcases. |  | |  | |
|  | |  | |
|  | |  | |
|  | |  |  | | |
|  | |  |  | | |
|  | |  |  | | |

The conclusion on the trendline graphs that we can see next visualised that the products tables, bookcases and machines generate a profit below 0. It means in conclusion that these products don’t generate total profit.

**2.3 HIPOTHESIS to analyse the profit for the top 10 products.**

* + What I have done to extract the information in the table above, is to compare the total amount of sales and profit with the top 10 products and top 7 products to prove my hypothesis.

|  |  |  |
| --- | --- | --- |
|  | sales | profit |
| top 10 products | **£8,486,841.30** | **£213,641.49** |
| top 7 products | **£8,117,848.19** | **£307,468.10** |
|  |  |  |
| The total amount of sales is £8,486,841.30 making a profit of £213,641.49. | | | |
| If the total amount of sales decreases to £8,117,848.19 because we sell less products,  but the total profit is higher than previous: £307,468.10 even selling less products, | | | | |
| it means that my hypothesis is positive. | |  |
| In other words, If top 7 products profit is higher than top 10 products, my hypothesis holds true. | | | |
| So after conclusions my hypothesis is true.  **HYPOTHESIS** | |  |

Top 7 products profit > top 10 products profit = yes

Top 7 products (£307,468.10) profit > top 10 products (£213,641.49) profit = hypothesis holds true

**Conclusion:**

In conclusion it makes sense to eliminate the three products: tables, machines and bookcases of the top ten products due that these don’t make increase the total profit in relationship with the sales.

In addition, the data that proves my hypothesis, and we can see it in the previous table that even if we don’t sale these three products: tables, machines and bookcases, the number of profit in relationship with the products we sell still increasing. So for the next year, I would get rid of these three products in my list of items to sell.

1. **Which are the 10 bottom list products?**

**Chart, bar chart

Description automatically generated**

**WHY?**

This information is good for the company to start thinking in get rid of them and start thinking to introduce new trendy products in the market in order to bring more profit to the company and keep on the development of new products.

In comparison with the previous point, the main products that generate less profit for Babington in terms of sales are tables which has an impact in the number of sales making a profit of above 17.000. Therefore, tables should be an item to consider to have or not in stock for the next year. The next 2 itemsbookcases and supplies are another two that don’t generate profit for Babington.

1. **Identify months of slow sales**

There is a slow period of months between January and February every year being specially significant in the period of February represented in the table which are the lowest sales month of every year being February of 2014 the lowest month in the total period of sales.

The graph shows a noticeable increase of sales between 2014 to 2016 and a slight decrease in the last year, 2017.

Graphical user interface, table

Description automatically generated with medium confidence

Whereas the highest months of the year it is represented in the table below with the value of the highest month, we can see that sales start to increase from September and there is a pick of sales in winter time between November and December reaching in November of 2017 the highest number of sales.

Graphical user interface, application, table

Description automatically generated

If we compare both graphs in the graphic below, the lowest period of sales to the highest along the years, it shows that after a higher period of sales in winter time, the sales go down between January and February. This can be interpreted in the way customer increase their sales in winter and slow down the sales after the “hard January” when the budget for the next year slow down. This can be related to the fact that sales in generally increases in Christmas and slow down in the months afterwards.

In summary, there is a big contrast between those slow months in sales during January and February to highest in terms of sales so as the graph shows, the lowest in sales can reach lower than 20,000£ whereas the highest months in sales can go up to 500,000£, 26 times higher.

Table

Description automatically generated

**WHY?**

* 1. This information is good for the company in order to know when is the most quiet period as a good information for the warehouse department and relevant information of the stock and staff working within the department. It is also a good way to advance information for the finance department. And also good to bring incomes from other services of the company if it’s necessary.

1. **When is the highest profit by sales?**

**WHY?**

This point help to the company to have stock ready for this period of time of the year and advance when will be the highest profit by the sales in the company. Also it’s interesting to invest in advertising campaigns to keep on or improve the sales.

The highest profit that the company makes by sales during the period between 2014 and 2017 was in December of 2016 with a profit of 16,724£. This data in relationship with the previous point to compare with slow and highest sales; there is a connection with the profit linked to the seasonal period of sales between November and December.

Graphical user interface, table

Description automatically generated

1. **Which are the top 5 States that generates more sales?**

Chart, pie chart

Description automatically generated

**WHY?**

To find deeper who is the target of the company in terms of the sales.

The top 5 states that generates more sales by order of colours are California in the first position distinguished by blue with a 40% of the total sales value, followed by New York in orange with a 28% of the contribution in sales and the las positions Washington, Texas and Michigan.

1. **Which are the top 5 cities that bring more sales?**

**WHY?**

* 1. This information is a benefit for the company to know which is our target and so on. It is also good to know the tools we need to reach that market, whatever is sales department or advertising online campaigns, …

**Chart, pie chart

Description automatically generated**

The top cities 5 cities that bring more sales noticeable by colours in the chart distinguished New York leading the first position contributing with a 34% of the total sales by cities, followed by Los Angeles in orange with a 23%, Seattle in grey, San Francisco in yellow and Philadelphia with a less proportion.

1. **Which state and region generate less sales**?

**Chart

Description automatically generated**

**WHY?**

This information is beneficial for the company to see the difference of sales in the area and to find out who is not your potencial target for future advertising campaigns,…

The South region leads the total of the less percentage of sales of the regions of United States. In the table above we can see that the blue line that cross over the South region represent the sales of this area that has a limited of less proportion in sales above 20,000£ in the State of Kentucky that represents the lowest contribution in sales from the South and the rest of the regions.

1. **What is the most popular shipping method?**

**A picture containing graphical user interface

Description automatically generated**

Standard Class is the most popular shipping mode method by the total number of customer who represent the 60% of the total sales.

**WHY?**

To help to understand the company what is the most popular delivery methods and also find out which are the less popular. This is also a good information to know what are the most delivery methods for the company and doble check in consequence, the delivery companies to work for.

1. **What is the shipping method for the best seller by city?**

**Chart

Description automatically generated**

**WHY?**

This information is good to know what is the most common shipping method among the best seller and see if there is a relation between the best seller and the shipping method they use.

The above table represent the most popular shipping methods used by the best seller cities: the one that makes a higher profit.

The conclusion after the table means that the cities that provide a higher number in profit of the total sales, cities such as New York, Los Angeles, Seattle, Lafayette or San Francisco, has in common that they all use **Standard Class** as the principal shipping method. Therefore, standard class represent the method used in more than the 50% of the total sales.

**CONCLUSIONS FOR ALL MY PROJECTS IDEAS**

in relationship with what I wanted to analyse in each point.

**1.**  **Identify of what is the segment that consume Babington product?**

With this conclusion, we can determinate that the target and potential buyers of Babington is the public who consume more than 50% of the total sales.

**2. Which are the 10 best seller products?**

The total of products that generate higher benefit in the company are copiers on the top followed by phones, accessories, paper and blinders. The top ten products produce on average around 849.000£ with a total profit of 21.000£.

Trying to find out the profit that generates less profit, we can see that tables and bookcases generate profit below 0 so I wanted to confirm this data in my hypothesis. To be able to do that, I compared the top 10 products sales and profit with the top 7 products sales, to get the conclusion that not selling the products that don’t generate profit we increase the number of total profit. So my hipothesis holds true. Furthermore , it make sense to eliminate the three products: tables, machines and bookcases of my top 10 products due that these don’t make increase the total profit in relationship with the sales.

So after this analysis we know which are the best seller products to keep on sales for the year after, and also the best thing to do is to eliminate the last three products that don’t generates profit for Babington.

**3.Which are the 10 bottom list products?**

In my previous analysis, I generally assume that tables should be a product to consider in the sales stock for next year. But after my hipothesis and I confirm in the previous point, I would eliminate tables, machines and bookcases of my top 10 products of the sales item for the next season because don’t generate profit for the company.

It will be also a good idea to start to research trends products in the marketing and also study competitors and consumer preference of products, so if we eliminate these 3 products of our sales items we can consider to replace them for another 3 products that we select after the research.

**4. Identify months of slow sales**

There is a significant slow period of sales in winter at the beginning of every year between the months of January and February. In contrast there is a high difference of sales with the last period of the year between November and December.

While the first stage of the year has an average of “lowest sales” in 57,000£, the last stage of the year lead the value of “highest sales” in average of 400,000£, around 8 times in number of sales which is a considerable difference.

**5. When is the highest profit by sales?**

The highest profit for the company by sales was in December of 2016 with a profit of 16,724£. The highest profit by sales coincide with the seasonal highest sales in autumn that the company produce between November and December.

From 2014 and 2017, the sales increase gradually reaching a pick in December 2016 and decrease slightly from 2017.

**6.** **Which are the top 5 States that generates more sales?**

The target of the company in terms of sales are the states of California and New York, followed by Washington, Texas and Michigan.

**7. Which are the top 5 cities that bring more sales?**

The top 5 cities that bring more sales to Babington are New York in the first position, followed by Los Angeles, Seattle, San Francisco and Philadelphia.

**8. Which state and region generate less sales?**

The South region lead the number of less sales with the state of Kentucky reaching the number of less sales. This analysis is good to have it in consideration when we launch the next advertising campaign. So after this analysis we know that the South region generate less sales in comparison to the West that generate the highest in number of sales.

**9. What is the most popular shipping method?**

With this analysis we can see that Standard class is the most popular shipping method and the one in average chose by he majority of customer, the 60% of them, in comparison with the less popular, The same day method. This is a good information to double check the delivery companies the company is working with and make sure the company is happy with their service to have in consideration for the next year sales.

**10. What is the shipping method for the best seller by city?**

The cities that provide a higher number in sales such us New York, Los Angeles, Seatle, … has in common that they use Standard Class as the main shipping method.

**SUMARRY of my analysis**

The previous analysis reflects the main ideas of my project which comes to the conclusion that the top product that generate the highest number of sales are copiers whereas the sale of tables decrease the profit.

Autumn, at the beginning of every year is the season with the lowest period of sales whereas winter in the highest period of sales. At the same time the profit in terms of sales coincide with the highest period in winter season, which is a good exercise to make decision in advance in terms of stock, budget, advertising and so on to organise for the year after.

My analysis explains the areas where the potential customer are based. The West region of United States makes a higher profit being California the top sate and the city of New York the one that produce more sales, in contrast of the South which is the region that produce less sales.

To complete my analysis, the shipping method most popular chosen by the majority of customer that represent a 60% of the total sales is Standard Class, and also the most common method used by the cities that contribute to a higher profit as it is the top one, is New York city. The cities that provide a higher number in sales use Standard Class as the main shipping method.