INTERNATIONAL UNIVERSITY
VIETNAM NATIONAL UNIVERSITY, HCM CITY
School of Computer Science & Engineering

PROJECT ASSIGMENT

Topic: Customer Personality Analysis

Advisor: Assoc. Prof. Dr. Nguyễn Thị Thúy Loan
Course: Data Analysis

Group members:

Nguyễn Hoàng Anh Tú	ITDSIU20090
Đặng Quốc Anh Duy	ITDSIU20015

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENT

We would like to express our deep and sincere gratitude to our research supervisor, Assoc. Prof. Dr. Nguyễn Thị Thúy Loan for allowing us to do research and providing invaluable guidance throughout this research. It was a great privilege and honor to work and study under her guidance.

Our completion of this project could not have been accomplished without the support of each other. Furthermore, we highly appreciated that we could finally complete this project within 10 weeks. Here is table of contribution of us during this time.

No.	Full Name	Student ID	Contribution
1	Đặng Quốc Anh Duy	ITDSIU20015	50%
2	Nguyễn Hoàng Anh Tú	ITDSIU20090	50%

CHAPTER 1: INTRODUCTION

I. Objective

Nowadays, thanks to the rapid growth of users' datasets, companies tend to explore these values and convert these data to gain profits from customers. Since using data to target customers has never been more controversial, more and more brands apply this method to get business tactics.

Customer Personality Analysis is one of the fast-growing fields that data analysts often deal with in their daily routine. It is a detailed analysis of customers to reckon the potential target group. Furthermore, customer analysis helps a business to know the favor of its customers and makes it easier for them to modify products according to the specific needs, behaviors, and concerns of different types of customers.

As a result, our project derived inspiration from that issue. The objectives of our project are to classify customer segments combined with the evaluation of consumption levels of customers and their buying habits so that company can figure out the strategies to run campaigns of marketing more effectively.

Figure 1 : Marketing Strategy

II. Questionnaires

From the lecturer's requirements and the purpose of this assignment, we have some first questions before going through the following in-depth analysis steps. These questions will be answered in turn in the following sections.

- How many possible data points are required for the analysis and detection?
- How can we tackle the errors of the original dataset?
- How do some factors such as customers' information, promotion, purchasing places affect the - money spent on given products?
- How can we cluster customers into particular groups?
- What is the main type of analysis that this project was performed?

III. Dataset

As for the first question above, **how many possible data points are required for the analysis and detection?**

The original data set consists of twenty-nine columns and two thousand two hundred forty rows **[1]**. However, we consider that we only use, analyze, and evaluate based on twenty-one columns and most of the rows in this assignment. We will define the exact number of rows after preprocessing data. These twenty columns of data consist of three main kinds of information: customers' information, customers' expenditure on products, purchasing method, and promotion. The first is data about customer information, including an ID number, year of birth, education level, marital status, number of children, date of customer's enrollment with the company, and number of days since customer's last purchasing.

ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency
5524	1957	Graduation	Single	58138	0	0	04-09-2012	58
2174	1954	Graduation	Single	46344	1	1	08-03-2014	38
4141	1965	Graduation	Together	71613	0	0	21-08-2013	26
6182	1984	Graduation	Together	26646	1	0	10-02-2014	26
5324	1981	PhD	Married	58293	1	0	19-01-2014	94

Next, the information on customer spending across six product categories includes wine, meat, gold, fish, sweet, and fruit.

MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds
635	88	546	172	88	88
11	1	6	2	1	6
426	49	127	111	21	42
11	4	20	10	3	5
173	43	118	46	27	15

Finally, there is information about how customers buy, the number of orders purchased with promotions.

NumDealsPurchases	NumWebPurchase	NumCatalogPurchase	NumStorePurchases	NumWebVisitsMonth
3	8	10	4	7
2	1	1	2	5
1	8	2	10	4
2	2	0	4	6
5	5	3	6	5

This is our original dataset: https://docs.google.com/spreadsheets/d/1VwhidcDr0C-VjZh1CcC6uZmzAyRgXfU_ITdTzw_pNmQ/edit?usp=sharing

IV. Tools and resources.

- Programming Language: Python with Google Colaboratory.
- Data Visualization: Python, Power BI.
- Supportive Analyzing: Excel, RapidMiner, Google Sheet
- Resources: Kaggle

CHAPTER 2: PREPROCESSING

I. Introduction

In the era of big data, more and more datasets are collected to get information about customers' consumption. However, before using data to get valuable profits, data analysts need to clean up data so that they can fix any data that is inaccurate, duplicated, and irrelevant and remove null values to make sure that all data are valid. In this part, we will give an overall answer to the question in the questionnaire part above: **How can we tackle the errors of the original dataset?**

In the beginning, we will drop nine columns of data that are not mentioned in the dataset part above from the original dataset. We divide our cleaning procedure into 3 steps. First, data will be merged from different types into some large groups or changed into a new group name. Next, we will figure out the problem of missing values, and then find solutions to tackle that situation. Finally, we will look for outliers to ensure that values are true and ready to explore.

II. Procedure

1. Merging and Changing

- ID: We consider that this column is not related to our analytic results, so we drop this column.
- Age: Since the customer's enrollment with the company was between 2012 and 2014, we assume that the data was collected in January 2015 for the sake of simplicity. We create a new column named "Age", then we add data to this new one by taking 2015 minus each number in Year_Birth column and after that, we drop the original column.
- Children: We merge Kidhome and Teenhome columns into the new one called "Children". The values in this column are equal to the sum of the two previous columns.

- Marital_Status: We see that there are seven different values in this column, so we convert them into only two appropriate values. " In relationship " replaces "Married" and "Together" values and "Single" replaces the others.
- Total_spending: The customer's spending is given separately for different products. Thus, we sum them up to calculate the total spending of the customers.
- Month_customer: We calculate the number of months that customers are affiliated with their company by the formula: month with company = 12 * (2015 - Year) + (12 - month). The data of year and month are taken from the "Dt_Customer" series.
- Education: "2n Cycle" and "Master" have the same meaning, "2n Cycle" will be transformed to "Master" to observe the data more effectively.

Based on these new columns, especially "Total_spending" and "Month_customer" columns, we will classify people into 4 different groups. Moreover, the other columns will help us to optimize and minimize unnecessary data.

2. Missing Value

As we all know, this dataset contains 2240 rows which are relatively large for analysis; however, there are 24 rows with missing values in the income column and the proportion of these rows only account for roughly 0,01% in total. Therefore, we decide to drop these rows instead of filling them with other values because the values of these rows are negligible, and it cannot affect the result.

3. Outliers

- Income: From the data, while most people's earnings fluctuate under 80.000, there still are some customers whose income is more than 120.000$ and some of them even more than 600.000$. They are clearly the outliers in the data, so we will leave them out.

- Age: There are three customers aged above 100, which is unlikely to happen. As a result, we also drop those customers from the database.

CHAPTER 3: EXPLORING DATA

In this part, we will determine the answer to this question:

How do some factors such as customers' information, promotion, purchasing places affect the money spent on given products?

I. Clients' Information

1. Marital Status

One thousand four hundred twenty-two clients were in the relationship, covering nearly two-thirds, whereas single was the marital status of remainders.

2. Child Status

There were 628 buyers who did not have any children. Besides, 1112 purchasers have one child, more than half of the total number of clients, compared to 415 buyers with two children and 50 remaining with three children.

3. Education

In the original dataset, the education column has four different kinds: "Ph.D.", "Master", "Graduation", and "Basic". Graduation is the educational level of over a half of total clients, with 1113 people (50.5%), while just 54 (2.4%) clients have the level as basic. Furthermore, 562 (25.5%) and 476 (21.6%) clients are at Master and Ph.D. level respectively.

4. Age

The youngest clients are 19 years old, while the oldest is 75. The data of clients' age is separated into four groups (Teen, Adult, Middle Age Adult, and Senior Adult) to mine deeply in the next part below:

- "Teen" are customers whose ages are under 20.

Only two clients belong to the "Teen" group, making up a tiny percent. (0.1%)

- "Adult" are customers whose age is between 20 and 39.

It seems that a pretty large number of clients belong to the "Adult" group, with 683 people. (31%)

- "Middle Age Adult" are customers whose age is between 40 and 59.

"Middle Age Adult" is a group with the most impressive number of people, with 1160 accounting for over a half. (52.6%)

- "Senior Adult" are customers whose age is higher or equal to 60.

Three hundred sixty customers belong to the "Senior Adult" group. (16.3%)

➔ Even though the "Middle Age Adult" group is the largest group, the most popular age is not belonging to this group. In fact, the most common age is 39, with 88 clients having this age, whereas in reverse, just one client is 75 years old at that time.

II. Income

1. Overview

Analyzing the salary of customers is one of the most essential parts of this analysis and plays a vital role in making decisions on which group of customers must be focused on. In general, the average income of the customer was approximately 51600$ while the highest one was double with 113734$ per period. However, the percentage of customers who earn less than the average was nearly 50,6%, accounting for one-half of the total. As a result, due to this imbalance, our project makes efforts to explore more insights from this column.

2. The correlation between income and educational level

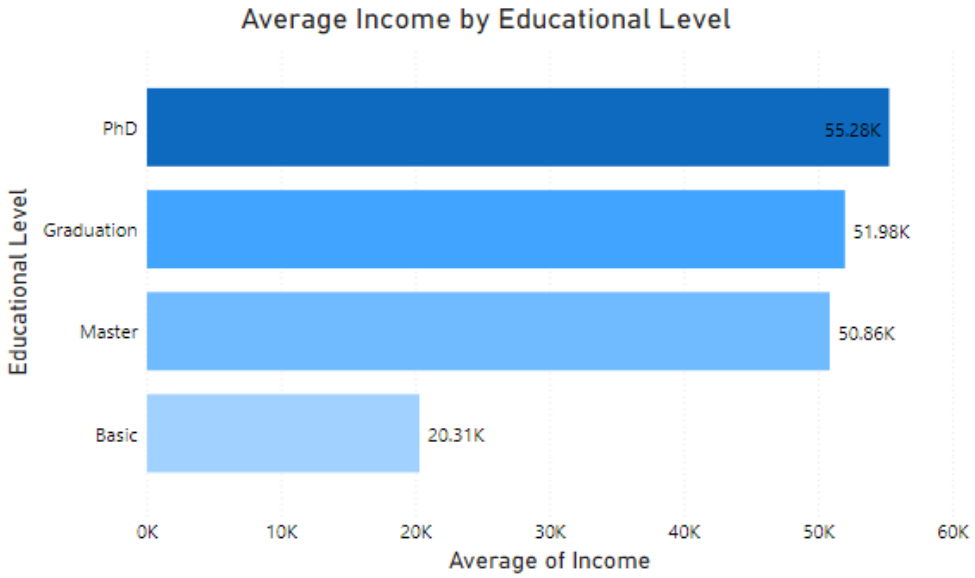

Figure 2: Average income by education

The graph above indicates the mean income of customers in the year 2015 based on their educational level. While PhD. has received the highest income (55,28K $) compared to the others, the average salary of people who were in basic level was only 20,31K $. The remaining groups were graduation and master, which earned 51,98K $ and 50,86K $ respectively. Consequently, there was a huge gap in the income of different groups, especially Ph.D. versus basic level. In this case, the first, three groups should be our target to run the business.

3. The relationship between total spending and income

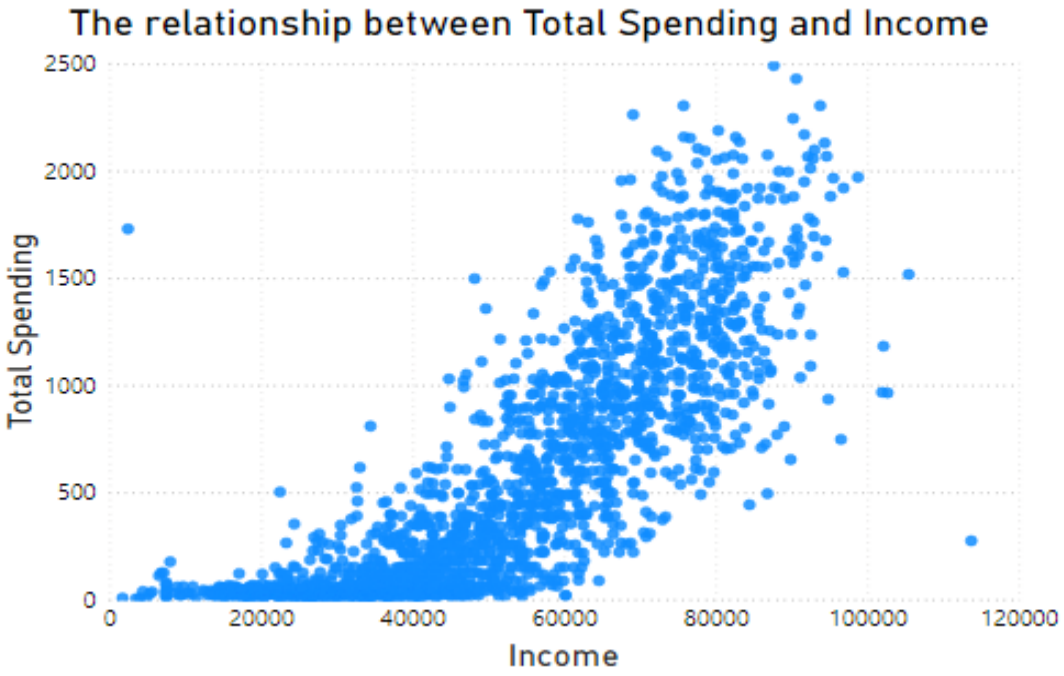

Figure 3: The relationship between total spending and income

In this part, a scatter plot is used to figure out the relationship between total spending and income since it can help us to check the similarity of customers in the wide range of income. Most of the dots fluctuated from over 50K dollars to nearly 100K dollars, which means that people who get high salaries will spend more money than others whose income was low. Therefore, we can admit that the more income they get, the greater money they will spend.

III. Expenditure

1. General Information

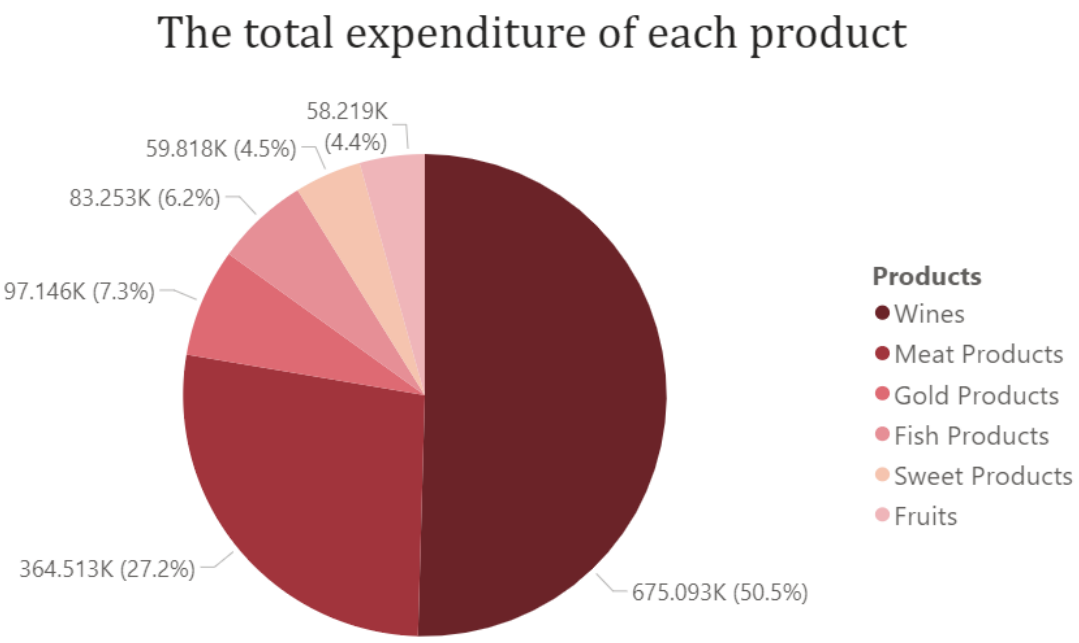

Figure 4: Total expenditure of each product

- From the table, we have six types of products. They are wines, fruits, meat products, fish products, sweet products, and gold products.
- The minimum value of all kinds of products is equal to 0.
- The total spending of clients on these products above was 1 338 042.
    - Wines:
- The sum of money spent on wines was 675 093.
- This product spent the most considerable amount of money, accounting for 50.5%.
- The highest spending on these products was 1 493 while the average one was approximately 306.16.
    - Meat products:

- The sum of money spent on meat products was 364 513.
- Meat products stand right after wines, with 27.2% of the total.
- The highest spending on these products was 1 725, higher than the highest spending on wines even though wine spending was the largest. In addition, the average one was approximately 165.31.
    - Gold products:
- The sum of money spent on gold products was 97 146.
- This product was spent the third large amount of money, making up 7.3%
- The highest spending on this product was 321, just one-fifth of the highest expenditures on meat products, while the average was approximately 44.05.
    - Fish products:
- The sum of money spent on fish products was 83 253.
- This product was spent the third small amount of money, accounting for 6.2%
- The highest spending on this product was 259, while the average one was approximately 37.75.
    - Sweet products:
- The sum of money spent on sweet products was 59.818.
- This product ranks 3rd, accounting for only 4.5% of the total.
- The highest spending on this product was 262, higher than the fish products' one. Also, the average one was approximately 27.12.
    - Fruits:
- The sum of money spent on fruits was 58 219.

- This product was spent the smallest amount of money, just 4.4% of total
- The highest spending on this product was 199, while the average one was approximately 26.4

2. The correlation between children and products

The number of customers who have had children is 1577, while the number of other customers is 628. Thus, the percentage of customers without children accounts for only 28.48%. Therefore, all expenditures of these customers are assumed to be transformed according to the following formula so that purchasing power can be compared with customers who have children more accurately.

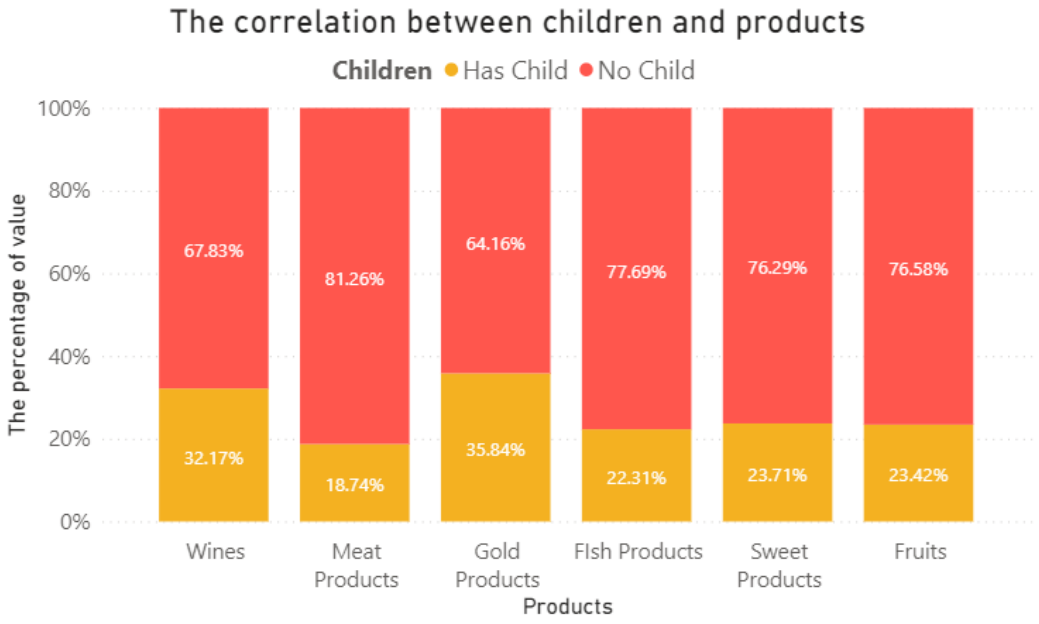

Figure 5:The correlation between children and products

$$\textbf{\textit{Assumed expenditure}} = \textbf{\textit{Real expenditure}} \times \frac{0.7152}{0.2848}$$

From the children part in customers' information above, the number of customers who have had children is 1577, while the number of other customers is 628. Thus, the percentage of customers without children accounts for only 28.48%. Therefore, all expenditures of these customers are assumed to be transformed according to the following formula so that purchasing power can be compared with customers who have children more accurately.

This bar stacked chart illustrates the correlation between having children and expenditure with childless clients' data converted based on the formula above. We can easily observe that even if the number of childless clients is just around 28.5% per total quantity of recorded clients, their purchasing power is enormous in all types of products, especially as they consume a significant amount of meat products, as forth times as the remaining clients. Furthermore, fish products, sweet products, and fruits were consumed considerably, over three times as clients who have children. Clients who have children spent nearly one-third as much as the others on fruits.

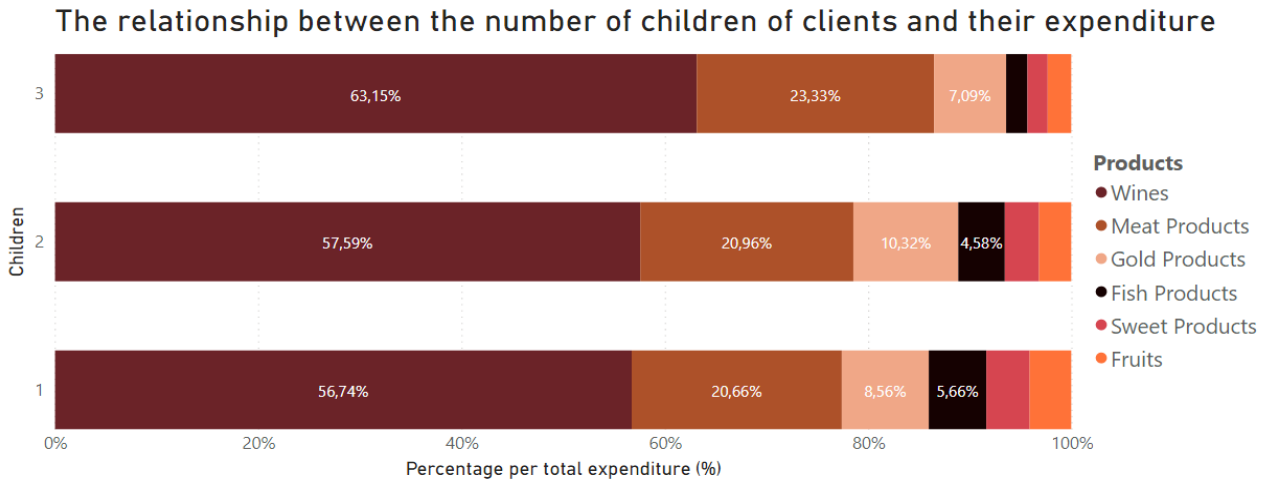

Figure 6: The relationship between the number of children of clients and their expenditure

We will analyze deeply inside clients who have children. We knew that most clients have one child, and clients have three children occupying a small percentage.

From the chart giving information about the relationship between the number of clients' children and expenditure, we can know which kind of products is most prevalent in each type of client having one child, two or three children.

The expenditure in both wines and meat products has a linear relationship with the number of children; the more children that clients have, the more spending on these products, 63.15% per total expenditure of clients having three children compared to 56.74% of the one-child one.

On the other hand, the linear relationship also occurs in the expenditures on fish products, sweet products, and fruits but the reverse trend. The spending on these three kinds of products in the total spending of one-child clients is double that of three-child clients.

Conclusion:

To sum up, the child is a factor that has a clear relationship with clients' spending. From this point, we can use this feature when we cluster clients in several groups in the Cluster part below.

3. The correlation between age and products

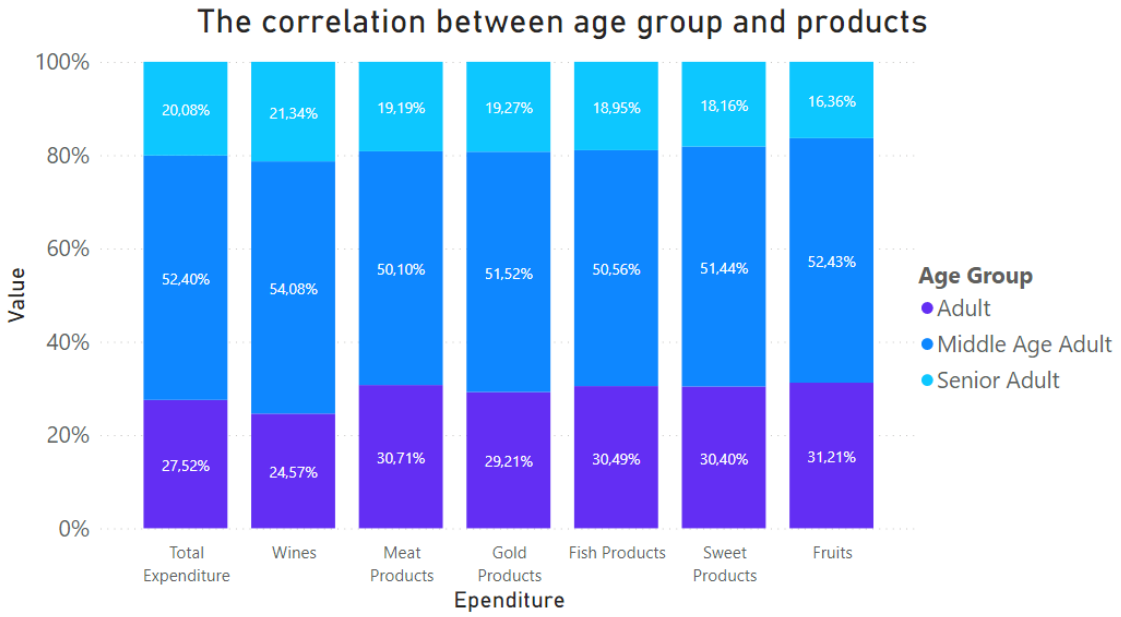

Figure 7: The correlation between age and products

- "Teen" age group only consists of two buyers, so in this part, we just analyze based on three other age groups.

- Wines seem to be the best seller toward three analyzed age groups (Adult, Middle Age Adult, Senior Adult). However, clients below 40 years did not buy as much as the others, with around 45.06% in total spending compared with 52.09% and 53.62% in total spending of "Middle Age Adult" and "Senior Adult" age groups. The correlation between spending on wines and age can be considered linear.

- "Middle Age Adult" group paid the slightest attention to meat products; it is similar to the expenditure of the "Senior Adult" group on fruits. Nevertheless, meat products and fruits occupied a significant amount in total bills of the "Adult" group, especially fruits.

Conclusion:

To sum up, there are also have some highlight points on the expenditure of recorded clients and their age. Nearly all linear relationships happen in the chart above; the older clients become, the more spending they allocate on wines, while the opposite trend is apparent in most other products. We can use clients' age information in the Cluster part.

4. The correlation between education and products

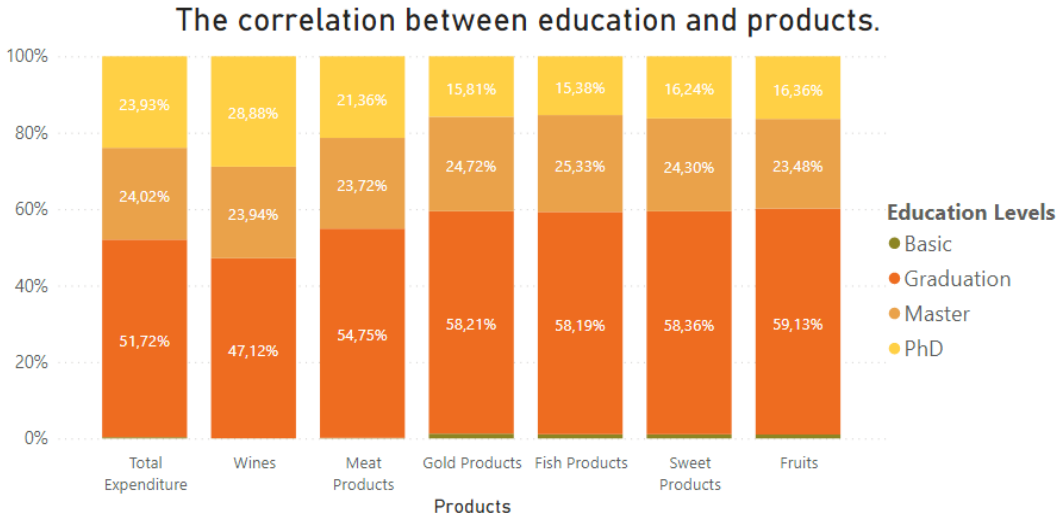

Figure 8: The correlation between education and products

Firstly, the "basic" level group spent insignificantly, the number of people who belong to this level occupying 2.4% in the total number of clients; however, the expenditure of this group just was slightly over 0.3%.

Secondly, the expense of Ph.D. and Master educational level groups around 24% per total while the number of people who are Ph.D. less than Master group 86 people.

Thirdly, the individual who has Ph.D. contributed a large amount of money on wines, whereas this product was not bought much by graduated people. Nevertheless, the remaining products were purchased considerably by individuals who graduated, especially fruits, while the opposite was true for the Ph.D. group.

Finally, as for the master education level group tends to buy fish products, but on the other hand, these products were spent least by the Ph.D. group.

Conclusion:

Although we have some critical points in this part to make the decision on the marketing campaign, there are not as clear as the two parts above. We will not use the correlation between clients' expenditure and education level to cluster clients into several groups and analyze them more deeply.

IV. Places

1. Overview

As you know, we are now living in the 21st century with a variety of developments thanks to the 4.0 technological revolution. In fact, more and more paying methods are created to serve the needs of this situation such as buying at the store, shopping online via mobile app, and purchasing products through the web platform. As a result, a huge amount of data is generated every day, and this data should be taken into consideration to take insights from it. In this project, we focus on the four main types of purchasing things, including:

- Number of purchases made through the company's website
- Number of purchases made using a catalog

- Number of purchases made directly in stores
- Times of visiting a web page

In the previous part – preprocessing data, the number of paying products had been merged from 3 columns into 1 by taking the sum of these methods of purchasing. After analyzing, the number of purchasing products ranged from 0 to 32 times per period while paying online accounts for the largest percent.

2. The correlation between total spending and the total purchases

In this part, we focus on analyzing the total spending and the total number of purchasing times so that it could be helpful to distinguish the fourth main group of customers in clustering stage.

The chart below looks like the normal distribution with the mean times of purchasing is 18 along with the highest spending in this case was 118K dollars per period. Although there are some customers who purchases many bills, the budget for their bills were negligible. For instance, while consumers purchased 30 times per period, the spending for these times were approximately between 3K-5K dollars. In contrast, some customers only purchased 5-6 times per period, but their budget were more than three-times compared with the previous group of customers.

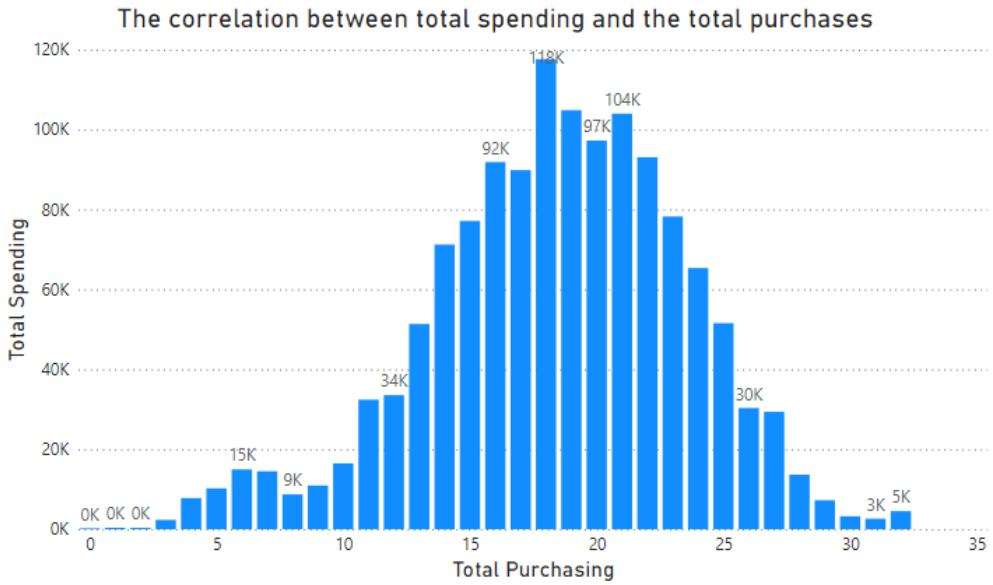

Figure 9: The correlation between total spending and the total purchases

V. Promotion

1. Overview

In doing business, running on the campaign plays a crucial role in marketing products to customers. From the collective data, data analysts are expected to get deep into these values to take advantage; therefore, marketing officers can base on these values to choose the right product to make a discount. For instance, with the products that have difficulty in selling, marketing officers will know how to apply discounts with a willingness to boost the selling profits.

On the other hand, in the aspect of a customer, many people tend to buy a product when the price is deducted dramatically as well as fit their pocket. Consequently, we decide to analyze the number of deals purchased from the database to check whether it is beneficial in presenting the result.

2. The correlation between purchasing deals and total purchases

From the given dataset, the largest number of deals purchases is 15 times, which make up nearly one-half of the total number of purchases. By comparison to the total, we can admit that most of the purchasing times are related to the promotion. Therefore, giving discounts to customers can be considered as one of the ideal options in selling products since it can not only meet the requirements from customers about the high-quality of products but also be affordable to any group of customers.

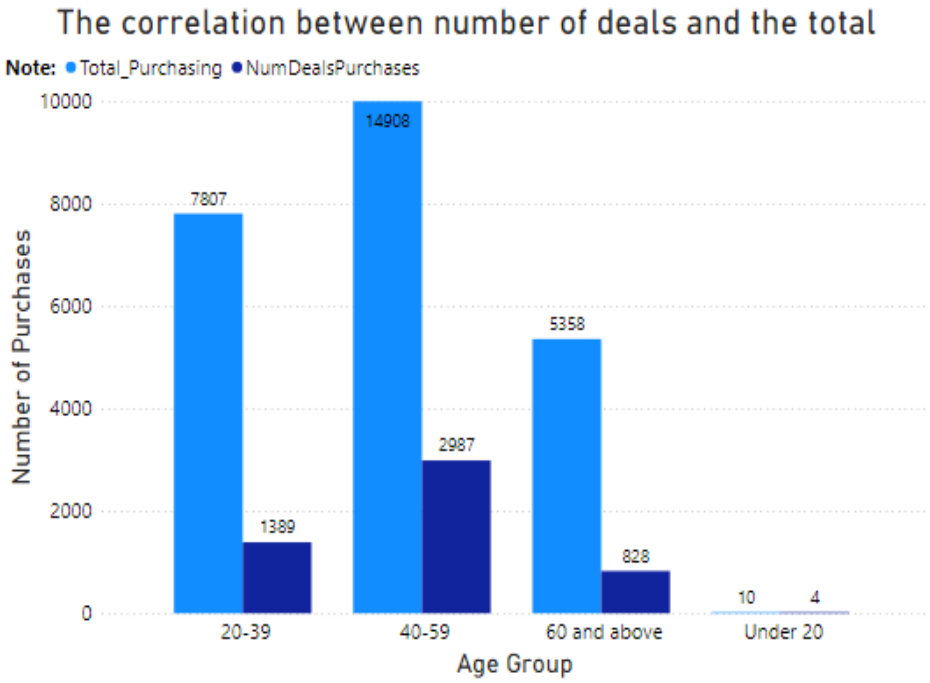

Figure 10: The correlation between number of deals and the total

As you can see, people who are in the group 40-59 (Middle Age Adult) purchased products the most in total, with nearly 15000 times whereas the group under 20 only spend money when there was a fabulous deal from the company. In addition, the total purchases of the adult group

(20-39) made up one-half compared to the Middle Age Adult group while the figure of the Senior Adult group was nearly one-third. Furthermore, deals accounted for nearly 20% of the total, which is an important process in selling products. As a result, this bar chart is useful in arranging levels between different groups of ages.

CHAPTER 4: CLUSTERING

I. Introduction

The primary purpose of this project is to localize and classify customers into small groups with similar characteristics to offer advertising campaigns or incentives for each customer group to promote business and boost revenue. So, how can we cluster customers into particular groups?

Although it is customer segmentation, we do not use classification, a supervised learning concept in machine learning. We use clustering, an unsupervised learning concept **[2]**. Additionally, we use the K-Means clustering algorithm for this project. **[3]**

Based on these analyses in many previous parts, we will use five buyers' information to optimize and have the best result after fitting it to the model. First, customer income and total spending are the two vital factors for customer clustering. Besides, the customer's registration date with the company is also helpful information. Finally, the age of the customer and the number of their children are strongly related to the customer's spending; as demonstrated above, the linear correlations are shown very clear, so we also use these two pieces of information in our analysis.

1. Elbow Method

We do not know in advance whether how many groups of customers are best relevant, so we rely on the Elbow method because of its ease of understanding **[4]**. After applying the model and plotting them into a line chart by using Python in Google Colaboratory, where the point illustrating the number of clusters equal to 4 is the most like an elbow, we decided to choose the number of customer groups to divide as 4.

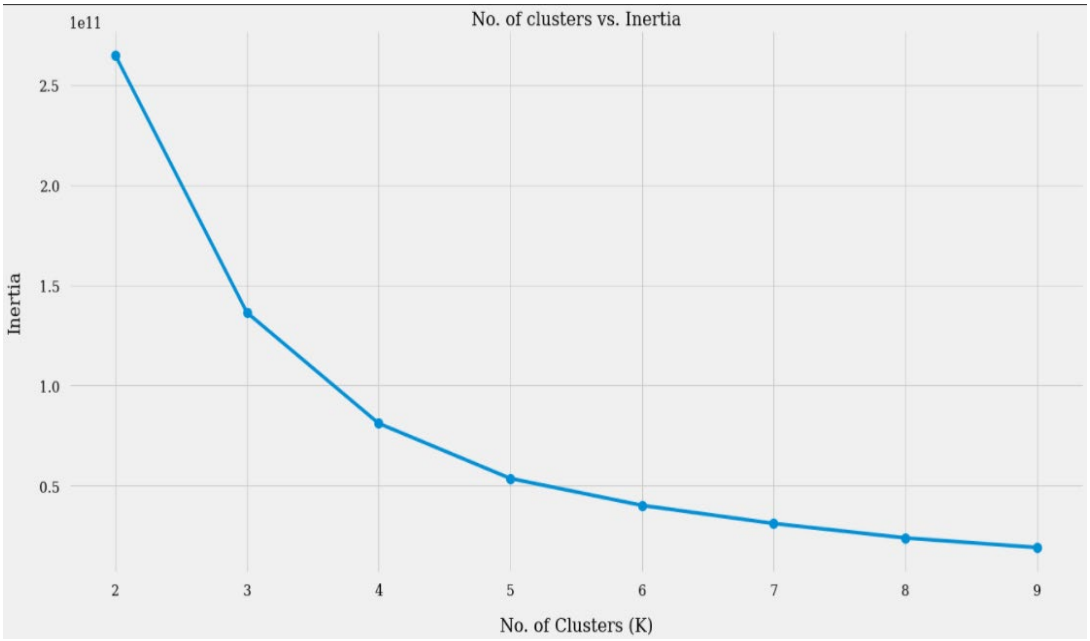

Figure 11: Elbow Method

2. Cluster identification

After choosing the number of clusters to use, we apply it to the model. We'll create two box charts based on a customer's income and total spending to label these customer groups.

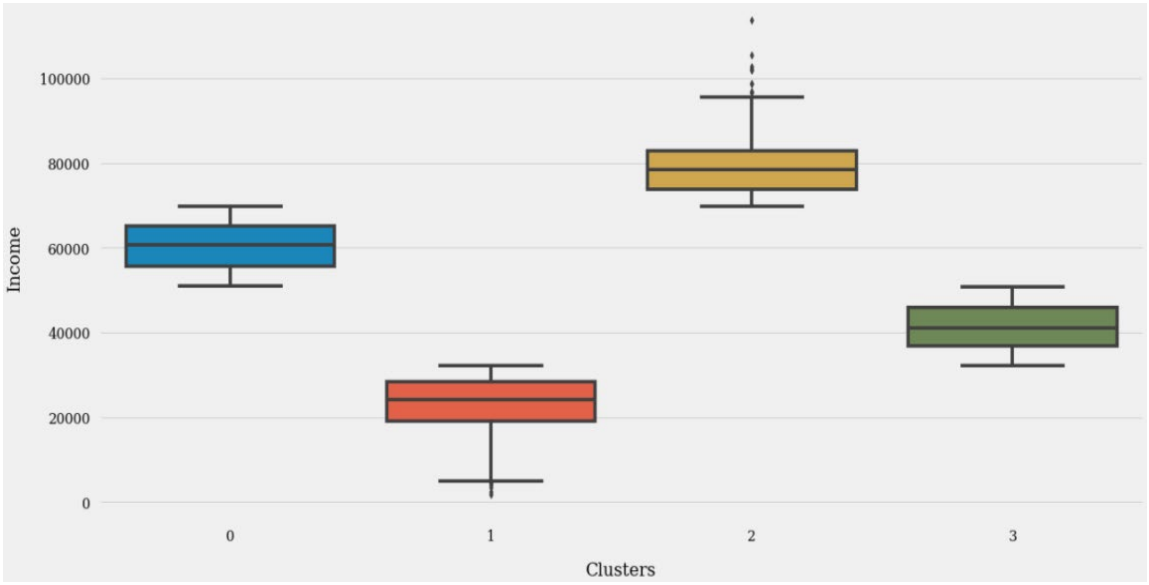

Figure 12: Cluster identification 1

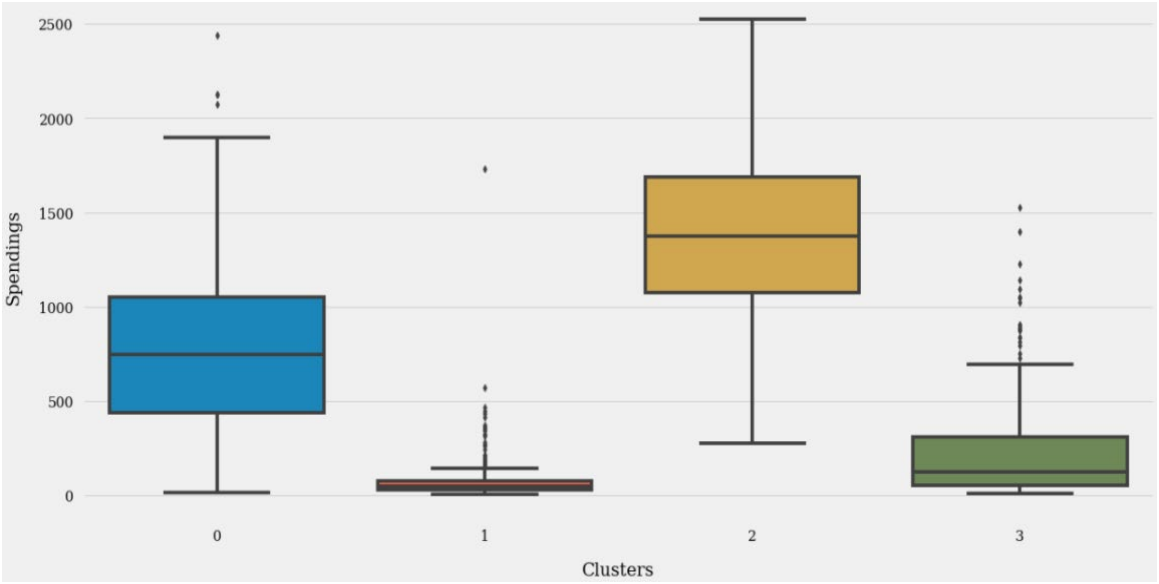

Figure 13: Cluster Identification 2

Based on these two box plots, we can easily spot the difference between customer clusters. Group 2 has the highest income, with an average value of about 80,000, and is also the group with the most spending, while group 0 ranks second in terms of income and expenditure. Group 1 and

3 are the two groups with the lowest and medium-low spending, respectively; their spending is like their income rank.

From the above data, we can label these customer clusters based on their income and total spending. The customer group with the highest income and spending will be marked as the platinum customer group. Based on income and total spending ratings from high to low, the remaining three groups will be labeled as gold, silver, and bronze customers, respectively.

3. Customer Distribution

Customer distribution by group is vital to the companies since it could differentiate between various groups of customers. As you know, we depend on many criterions to dispense customers into four main groups like their age, their spending habits, etc.

The silver customer group is the group with the largest number of customers, with six hundred forty-two people. A little less is the golden customer group with six hundred and thirteen people. The customer group with the least number of people is the bronze customer group, with only four hundred and forty-five people; the other group is the platinum group with five hundred and five people.

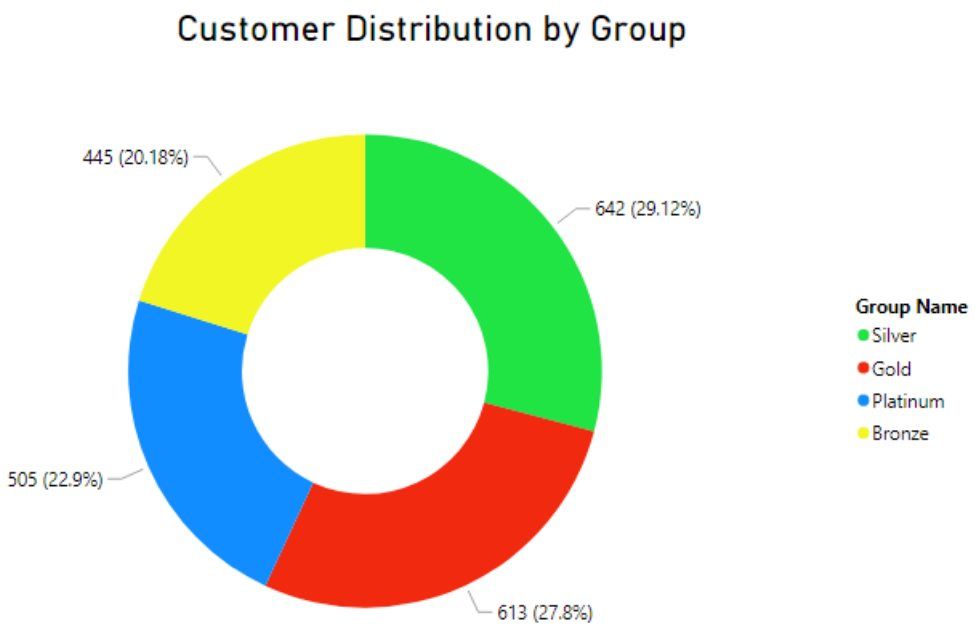

Figure 14:Customer Distribution by Group

In this donut chart, most of the consumers lie in the silver and gold group, accounting for 29.12% and 27.8% of total respectively. However, with the bronze group, this figure is slightly over 20% while platinum members make up nearly 23% in total. Therefore, we can conclude that groups of customers are evenly distributed, and there is no significant divergence between these groups after clustering.

We will look at some main features of these four client clusters:

- Platinum customers have income over 69 867, and people above 50 870 will be clustered as gold clients. In addition, if they have an income higher than 32 146, they will become silver

clients, and the remaining are bronze customers. The average income of these groups has a gap of around 20 000 for each group.

- There are not many apparent differences in client clusters' month registration and age. The average of each group differentiates nearly one unit; for example, the average age of the platinum group is forty-seven years compared to forty-six of the silver group.
- As for the number of children of each client cluster, most of the platinum clients do not have any children, while most of the gold clients have one child, and most of the silver clients have two children.

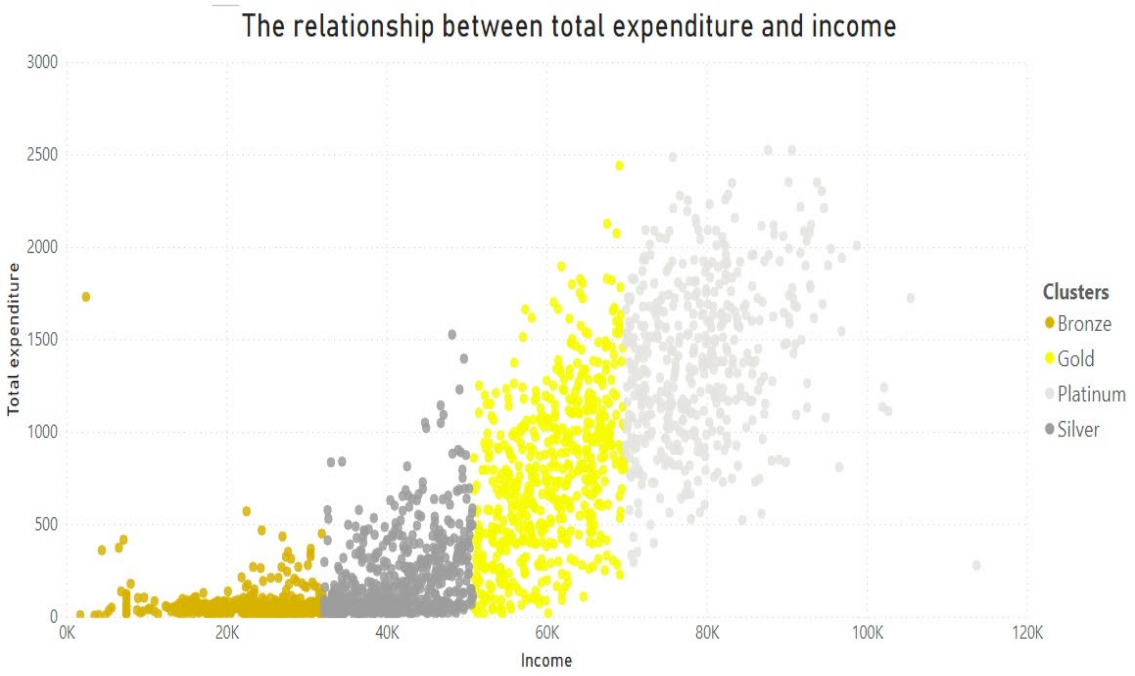

Figure 15: The relationship between total expenditure and income

The graph above shows the relationship between income and total expenditure. In fact, this graph was built in the income section above; the difference is that the points had colors according to their labels.

CHAPTER 5: DATA ANALYSIS

I. Client clusters' expenditure.

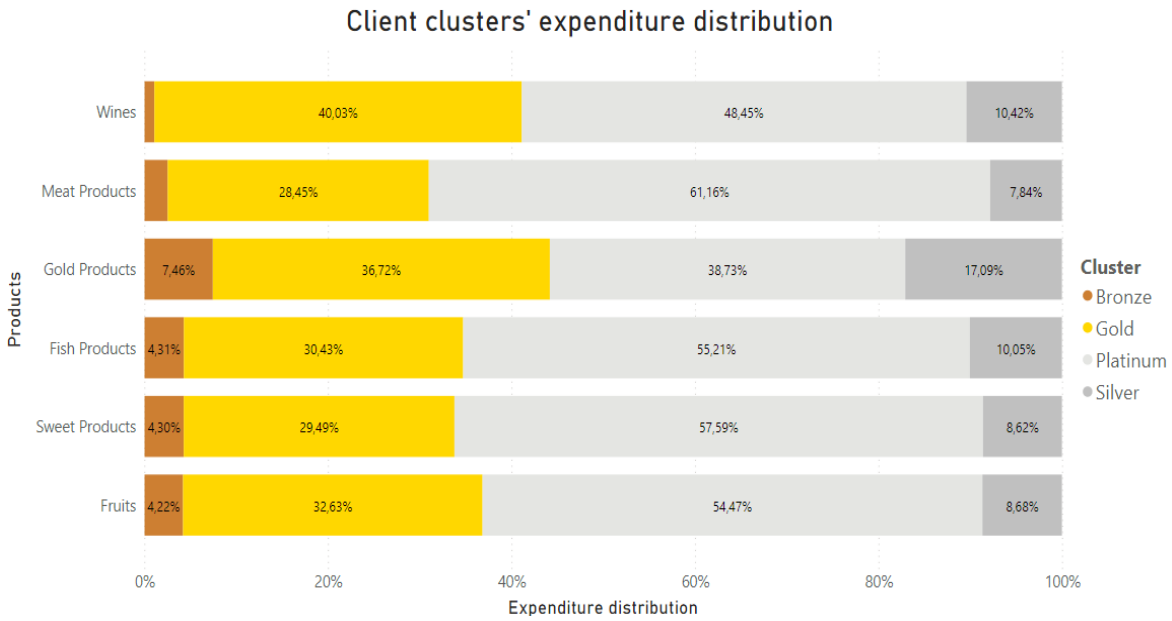

Figure 16: Client clusters' expenditure

Although platinum customers contribute the most to total spending on gold products, it is the least compared to other products. In contrast, the remaining three customer groups, especially silver and bronze customers, spent pretty much on these products. As for meat products, both gold and silver customers spent the least amount of the rest of the products, while platinum customers spent a significant amount, over two-fifths of the total money spent on these types of products belonging to this group.

Bronze clients spent the least on wine, and the other three groups spend a fair amount on this product. The remaining three products do not have many highlight points and have a relatively stable distribution.

II. Purchasing Habits

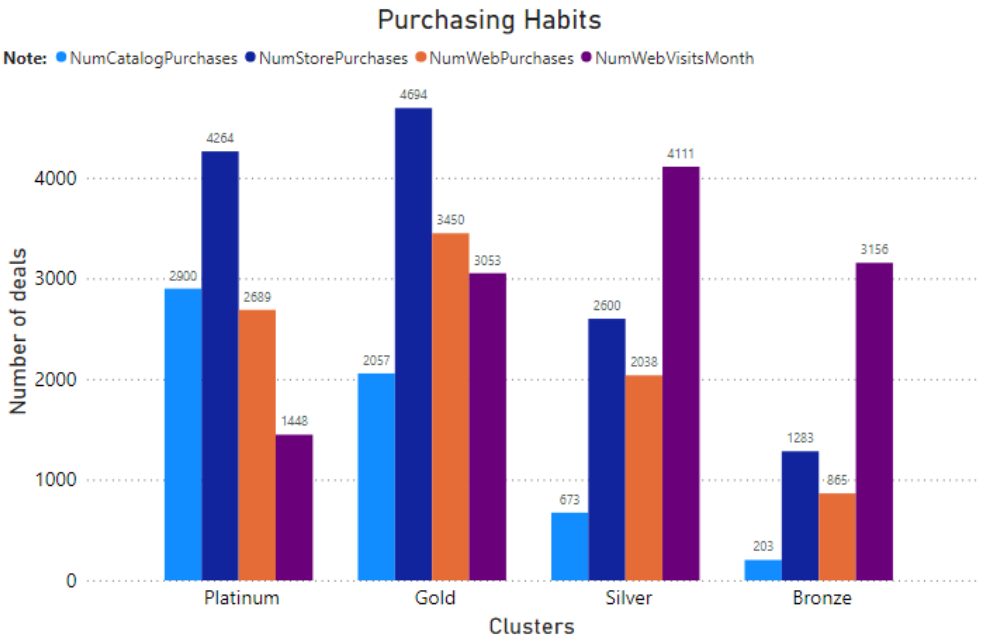

Figure 17: Purchasing Habits

Both Gold and Platinum groups are interested in purchasing products at the store, with over 4200 deals while the remaining groups are Silver and Bronze, only spending for 2600 and 1283 deals respectively. In contrast, the great deal of visiting times that these two groups spent on web platforms is enormous whereas Platinum members have least web visits. Moreover, most web purchases were conducted by Gold and Platinum members while this figure decreased dramatically from the silver group to the Bronze one.

Another insight that is particular to be mentioned is that Gold and Platinum customers also preferred to buy products through catalog. It means that advertisements like paper, video by marketing campaigns are helpful in selling products.

To conclude, companies should continue to run on business based on web platform and stores and try to approach more types of customers to create beneficial values for the company.

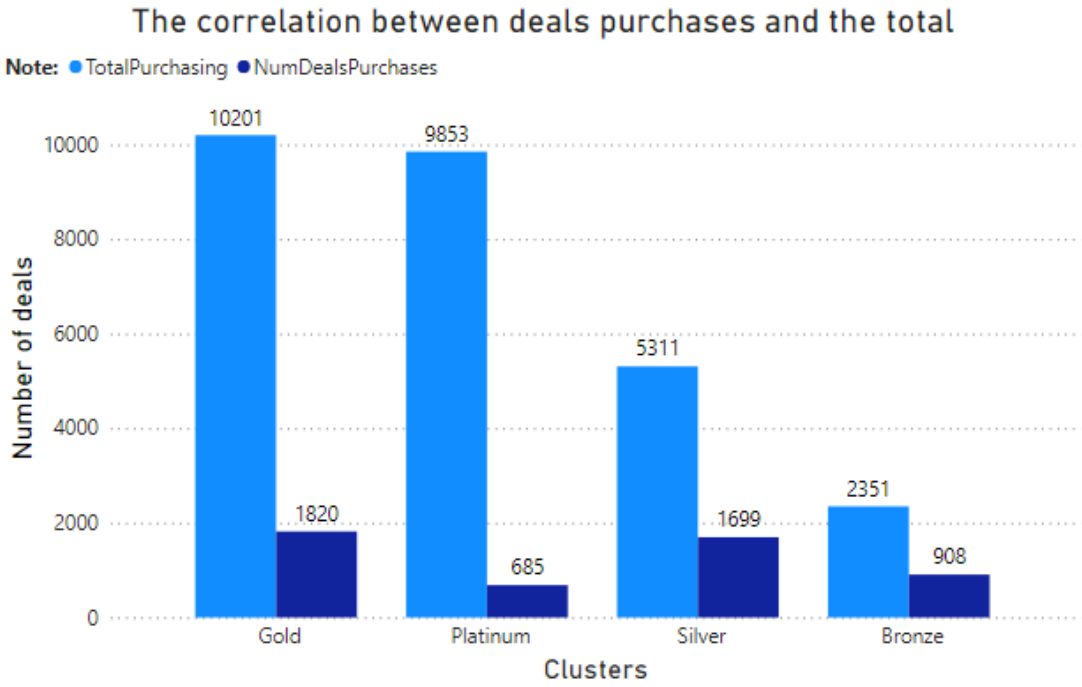

Figure 18:The correlation between deals purchases and the total

Deals purchase of these categories must be considered as the gold and silver customers are alike in total deals. While the figure of the previous groups remains high in total, the following groups are bronze and platinum, consumed 908 and 685 deals respectively. Although the number of deals purchases, which made up about 10% of total, are quite small quantities, this kind of offer is crucial since clients from different groups can have the opportunities to be exposed to many products that seems to be luxurious. Moreover, Platinum groups bought enormous products (9853 deals in total) even though their deals purchase were only 685 deals.

CHAPTER 6: INTERPRETING THE RESULTS

I. Result

1. Origin

After analyzing recorded data from clients, we will list some vital points.

- Wine and meat are the most popular products amongst all the groups.
- Most of the consumers have enrolled as company members for an extended period.
- Although the number of customers without children is not much, their purchasing power is immense; meat products were bought four times more purchasing power than customers with children. As for customers with children, the more children they have, the more they spend on wine and meat products.
- Age influences customer purchases. For example, older people spent more on wine, while adults bought fruits more than those above forty years.
- Education level also affects consumer spending. For example, from the analysis in Chapter 3, we see that doctoral customers have a very high demand for wine.
- Middle-Aged Adult Group purchased products the most in all buying methods
- Deals accounted for nearly 20% of the total, which is an important strategy to encounter some new group of customers.
- Places have great impact on buying habits because of its promotion to the clients
- Different group of age will spend money distinctively.
- A wide range of customer segment is taken into consideration through histogram chart. For example, some purchase 18 orders with the total of 118K dollars while others place up to 30 orders with the paid only between 3K to 5K dollars.

2. After clustering

After using the model to cluster customers into 4 groups, we can also summarize the salient features of each group.

- These customer groups are best distinguished by their income, with the average income of each group varying by nearly 20,000 from more than 20,000 for the bronze group to nearly 80,000 for the platinum group.
- The majorities of platinum customers are those who do not have children, while the gold customer group is dominated by customers with one child, while the silver customer group is the majorities of customers with two children.
- Customers with incomes below 70 000 tend to spend a lot on gold products, while the rest consume large amounts of meat, fish, and sweets.
- Most of the clients with incomes higher than 30,000 spend a lot on wines.
- Customer groups are clustered mainly by their income and other factors such as age, total spending, number of children, the month of membership. Companies can base on those characteristics to come up with appropriate strategies.
- A uniform distribution of different group of customers was expressed through the donut chart. It means each group make up about one-fourth of the total distribution.
- Web purchase is the most popular method of buying product.
- The huge number of customers suft the webpage regularly to find products. For instance, the total times that customers stayed on the webpage were up to 11768 times per period.
- The Platinum and Gold members contributed the most money through web platform and store

We already have the answer to the last question posed at the beginning: **What are the main types of analysis that this project was performed?**

There are four types of analytics: descriptive analytics, diagnostic analytics, predictive analytics, and prescriptive analytics [5]. In this assignment, the primary kind of analysis we use is descriptive analysis. Based on old dataset from 2015, we focus on answering the question: what happened in the past? We also use the remaining three types of analysis, but not many and not too complicated.

Figure out anomalies in our dataset:

After looking through the report again to check, we recently found that there are two anomalies in the correlation between income and total spendings, which we didn't realize during our analysis process. The first anomaly is on the left hand-side where the income was 2447$ over two years while their spending was up to 1730$.This means that their spending makes up approximately 70% of their income. However, another anomaly is opposite to the first one with the huge amount of income, but their spending is quite small. Its income and total spendings are 113734$ and 277$, respectively. So, we can consider these points as anomalies and unexpected behaviors in our dataset, and they can lead to some suspicions whether these data are valid or not. This can be considered as an outstanding discovery from our dataset. However, the duration of this project is limited, we hope that we can take it as a helpful lesson to improve in future project.

II. Conclusion

The purpose of this project is to analyze data recorded since 2015 about customer information as well as their spending on specific items so that companies can come up with appropriate marketing campaigns. Bringing the right products to the right people in need is

beneficial to increasing companies' revenue and helps customers have more incentives in their spending. This project helped us gain a lot of practical knowledge such as visualization of graphs using Power BI, learn more about clustering model in machine learning, know more commands to process data frames using Python programming language, use Excel or RapidMiner to analyze the data. Above all, this is a valuable opportunity for us to do a complete report, trying to analyze like a real data analyst. To be honest, as we are quite new to the world of data science, especially in data analytics, we do not have enough experiences to handle some limitations: more optimizing the clustering model since in this project, the customers are grouped mainly based on their income; or there are many other columns of data that we have not yet discovered that can be processed to use for their evaluation and analysis; or we cannot detect some anomalies in our process until checking the final result of analysis. Hopefully, throughout this course, we can have more practical knowledge to handle these shortcomings in the future.

III. References

[1] Patel, A. (2021, August 22). Customer personality analysis. Kaggle. Retrieved January 2, 2022, from https://www.kaggle.com/imakash3011/customer-personality-analysis

[2] Clustering in machine learning. GeeksforGeeks. (2021, September 22). Retrieved January 2, 2022, from https://www.geeksforgeeks.org/clustering-in-machine-learning/

[3] yasirhussain1987. (2021, September 14). Customer segmentation using K-means. Kaggle. Retrieved January 2, 2022, from https://www.kaggle.com/yasirhussain1987/customer-segmentation-using-k-means

[4] Elbow method¶. Elbow Method - Yellowbrick v1.3.post1 documentation. (n.d.). Retrieved January 2, 2022, from https://www.scikit-yb.org/en/latest/api/cluster/elbow.html

[5] Maz. (2021, August 29). Descriptive - diagnostic - predictive - prescriptive analytics,Maz Nguyen. Retrieved January 2, 2022, from https://madzynguyen.com/descriptive-diagnostic-predictive-prescriptive-analytics/