Customer Personality Analysis

Big Data Fundamentals Coursework

Student: Nguyen Thi Thu

ID: ntb21167

Table of Contents

List of Figures

List of Tables

1. Introduction

Customer analysis is crucially important for strategy development in industrial markets. Analyses such as customer's behaviours and habits, or supplier/customer relationships can help to optimise the marketing and purchasing strategy [1]. The former leader of Microsoft's home entertainment and mobile business commented on the failure of Zune that their advertisements only appeal to a very small segment of the music space, but they do not attract the majority of music listeners [2]. Microsoft failed to understand customer needs and evaluate the potential of the market and the target segment. Therefore, they could not give them a good reason to buy Zune over the iPod.

Lack of understanding about customers is the most common reason for business failure, regardless of the type of business model or type of products. On the other hand, careful market research on both prospects and existing customers will help businesses to define target customer segments and play a significant role in product development, production, marketing tactics, and financial planning.

This report examines the dataset of a retail company, where their customers are diverse in demographic characteristics, needs and shopping habits. Extracting significant insights may allow the company to classify customers, modify its products based on its target customers and market the product only on that particular segment.

2. Dataset

The chosen dataset is for Customer Personality Analysis that is available on the Kaggle website (https://www.kaggle.com/imakash3011/customer-personality-analysis). The original dataset includes 2240 rows and 29 columns. Each row represents a unique customer. Meanwhile, 29 columns represent different information about customers as follows:

- Customer profile (10 columns):
  - Unique ID
  - Year of birth
  - Education level
  - Marital status

- Yearly household income
- Number of children household
- Number of teenagers household
- Date of customer's enrolment
- Number of days since last purchase
- Customer complain

- Customers' expenditure on products in last 2 years (6 columns):
  - Amount spent on wine
  - Amount spent on fruits
  - Amount spent on meat
  - Amount spent on fish
  - Amount spent on sweets
  - Amount spent on gold

- Customers' response to Marketing campaigns (7 columns):
  - Number of discount purchases
  - Response to the 1st campaign
  - Response to the 2nd campaign
  - Response to the 3rd campaign
  - Response to the 4th campaign
  - Response to the 5th campaign
  - Response to the last campaign

- Place where customers purchase products (4 columns):
  - Number of purchases made through the company's website
  - Number of purchases made using a catalogue
  - Number of purchases made directly in stores
  - Number of visits to company's web site in the last month
- And 2 other undefined columns

There are 24 rows in the dataset that include missing data. These rows and some columns that contain undefined or un-useful information are removed. The column containing the customer's year of birth is converted to customer age. Six columns that contain customers' responses to marketing campaigns are combined into one. Two columns containing information about the number of children and teenagers in customers' households are also combined into one. The final dataset that is used for the analysis has dimensions of 2216 rows x 18 columns. The summary of the final dataset is provided in Table 1.

Table 1. Summary of the tidied dataset

Group of information	Number of columns
Customer profile	6
Expenditure on products	6
Place of purchase	4
Response to Marketing campaigns.	2

3. Summary statistics

In this section, a statistical summary of the dataset is presented. Due to the limited number of words, the section will focus on the overall analysis of the customers' shopping and accepting offers from marketing campaigns. Because the main goal of the analysis is to support adjustments of products and marketing campaigns to increase sales and use the marketing budget effectively.

3.1 Customer's education versus expenditure

Fig. 1 shows spending amounts of customers from different education levels on different types of products. The Graduation group spend the most with nearly 700,000 which is equivalent to 51.6 % of the company's revenue. Meanwhile, the PhD and Master groups spend nearly 33,000 and 22,000, respectively. On the other hand, Wines and Meat are two products that account the most in the total revenue. The higher the education level, the higher the percentage of spending on Wines. The percentage of spending on Wines of their total spending of Graduate, Master, PhD, 2n Cycle, Basic groups are 60.1%, 54.6%, 45.8%, 40.5%,

8.8%, respectively. Meanwhile, this proportion for meat of Basic group is 12.9%, this figure of other groups are between 25% - 29%.

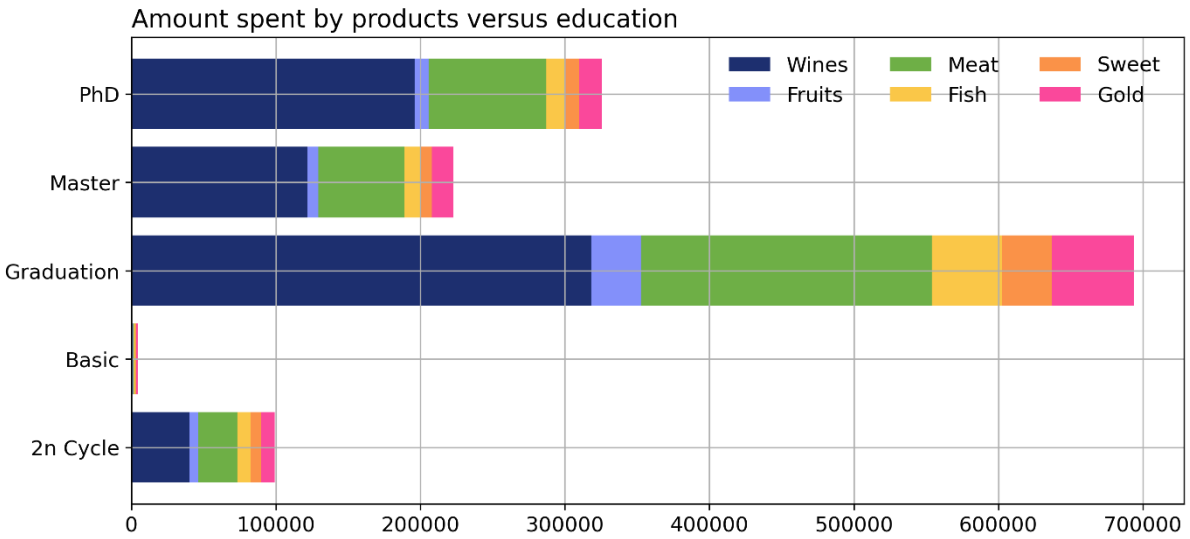
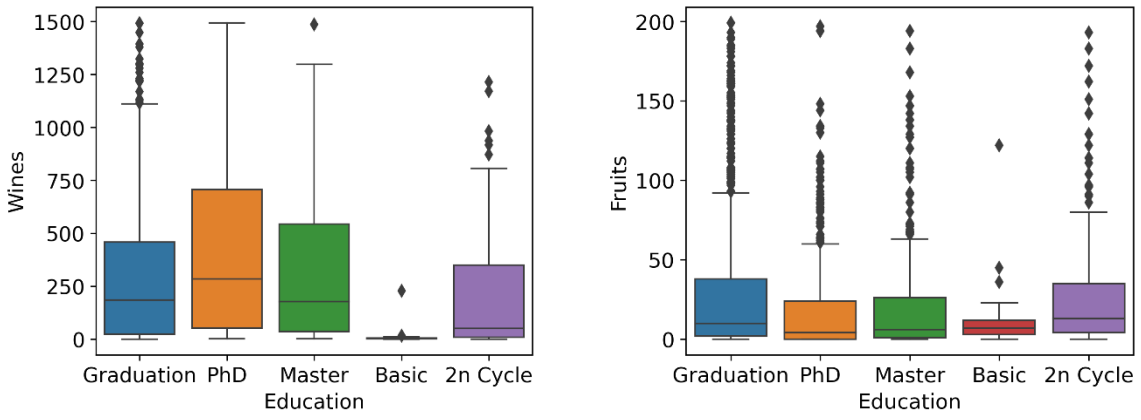


Fig. 1. Spending on products versus education

The difference in revenue contribution of the different qualification groups is due to the difference in the number of customers and the individual spending in each group. As shown in Fig. 2, Graduation, PhD and Master groups are three most important customer segments. these groups account for 50%, 21% and 16%, respectively, of total number of customers. The Graduation is the group that their average spending is highest on most products, except for Wines. The Basic group is the group with the least revenue contribution and the lowest per capita consumption per product.



(a) (b)

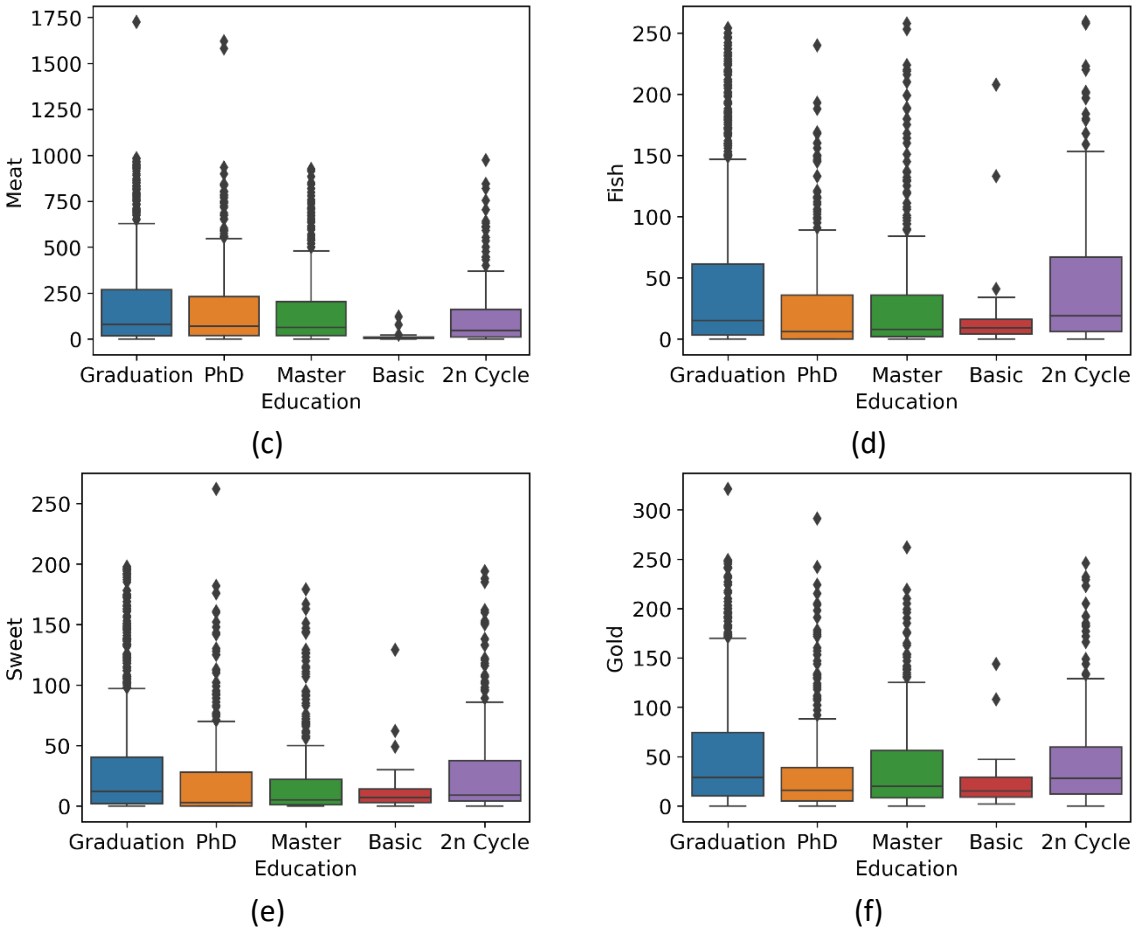


Fig. 2. Boxplots for spending on products versus education levels

3.2. Number of kids versus expenditure

Fig. 3 show the expenditure of customers grouped by the number of children. The group has 0 and 1 child contributes most of the company's revenue. Specifically, they contributed 700,000 and 530,000, respectively, equivalent to 52% and 39.4% of total revenue, which are nearly 7 times and 5.5 times higher than the group with 2 children. The group with 3 children is the group that contributes the least.

However, the group with no child accounts for 28% of the total number of customers while the group with 1 child accounts for 50%. What make the significant contribution of childless group is due to the average spending on all products is much higher than other groups as shown in Fig. 4. Each childless customer spends nearly 500 on average for wine, 30 for fruits, 370 for meat, 60 for fish, 40 for sweet, and 45 for Gold. Meanwhile, the group with 1 child spends about 150 for wins, 10 for fruits, 50 for meat, 10 for fish, 10 for sweet, and 2 for gold. These groups of customers are two most important customer segments, regardless of in terms of number of customers, revenue contribution or average expenditure.

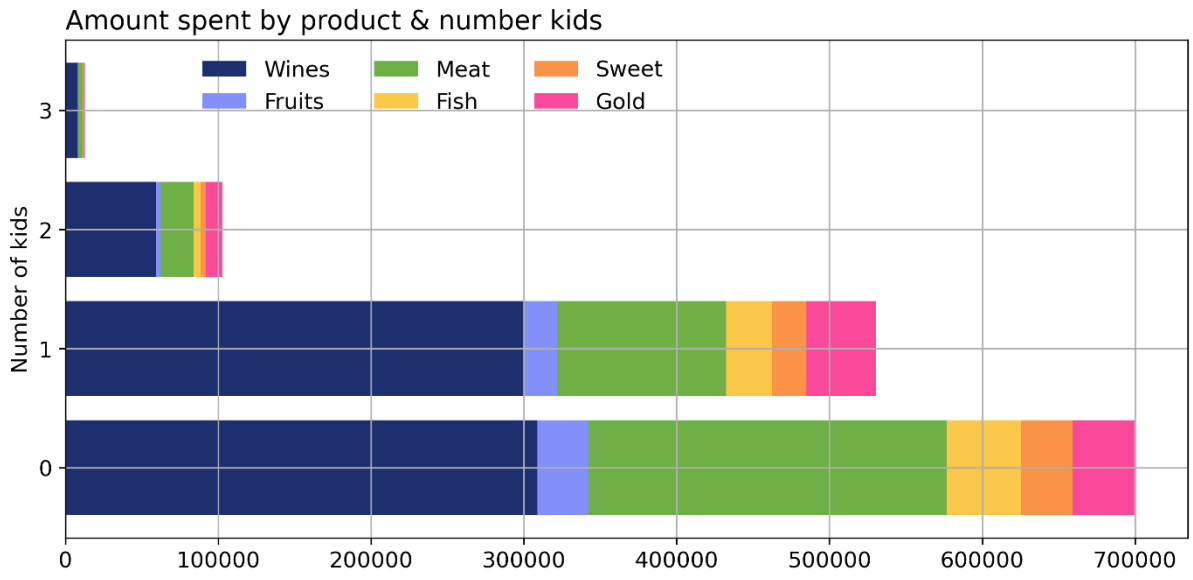


Fig. 3. Spending on products versus number of kids

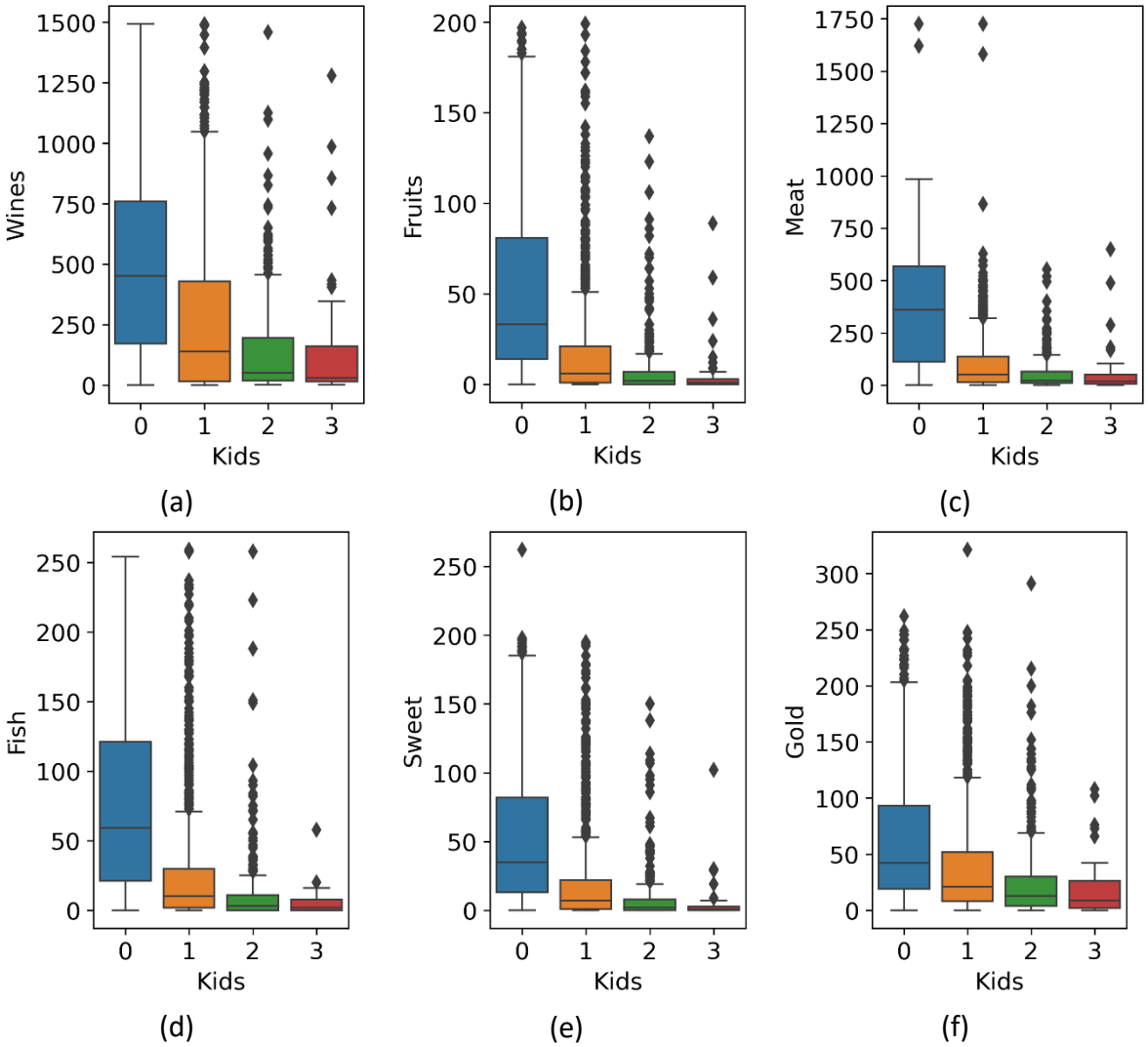


Fig. 4. Boxplots for spending on products versus number of kids

3.3 Marital status versus expenditure

In this section, the spending of customers grouped by their marital status is investigated. As shown in Fig. 5, two groups Together and Married spend the most with more than 500,000 and 350,000, respectively. This is because these two groups are the majority of customers.

The number of customers with the marital status of Alone, Absurd, and YOLO spend very little Therefore, customers are grouped into 2 main groups, Relationship (including Together and Married) and Single (including other groups). As shown in Fig. 6, the average spending of each person in the two groups of relationship and single is quite similar in all products. In both groups, Wine and Meat are the two most consumed items.

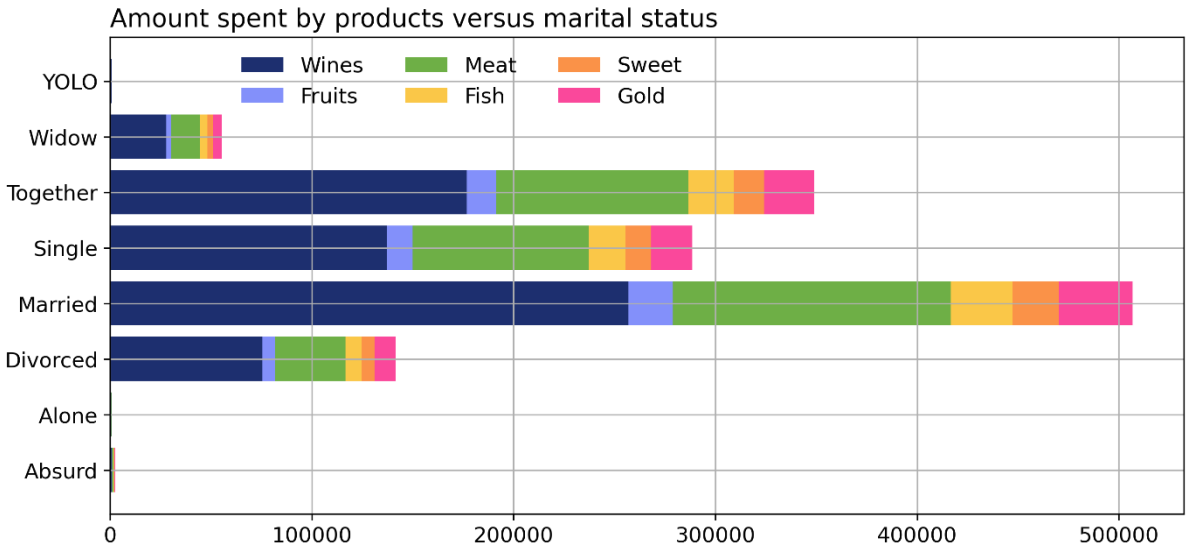
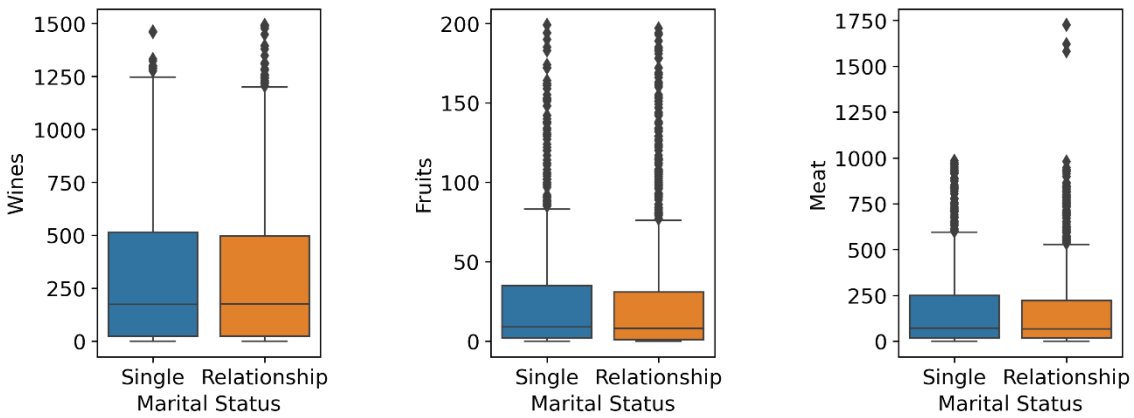
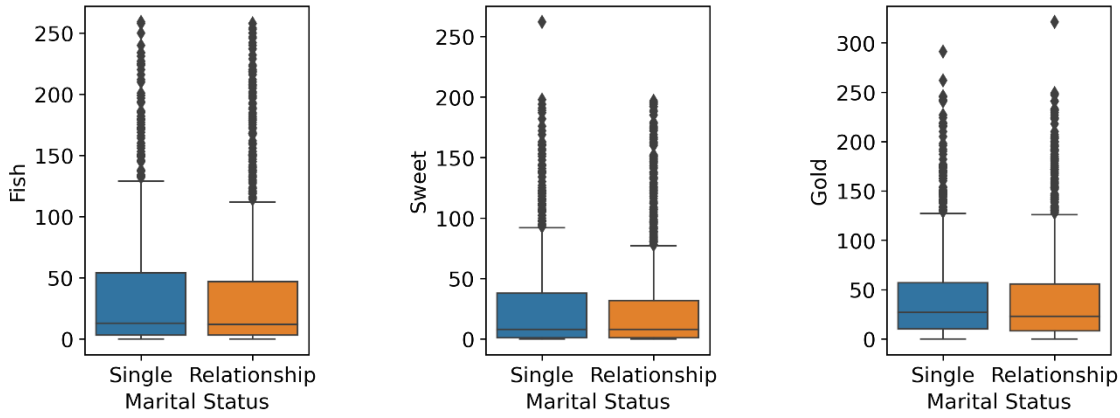


Fig. 5. Spending on products versus marital status





(d) (e) (f)

Fig. 6. Boxplots for spending on products versus marital status

3.4. Customer personality and Marketing campaigns

The company has implemented six marketing campaigns, all acceptance offers results have been combined and aggregated in Table 2 and Fig. 7. As can be seen from Table 2 and Fig. 7, the average number of acceptance offers of most groups, regardless of aspects such as marital status, education, or the number of kids, is only about 0.5 times/campaign or less. The most interesting case is the childless group with an average acceptance rate is 0.86 times/campaign.

Also, the Absurd group has the highest accepted rate with 1.5 times/campaign. However, as mentioned in section 3.3, the revenue contribution of this group is negligible. From these observations, it can be concluded that it is necessary to rethink the implementation of 6 campaigns to avoid wasting resources through adjusting the marketing strategy and targeted audience.

Table 2. Mean of response versus marital status, education and number of kids

Marital status		Education		Number of kids	
Group	Mean of response	Group	Mean of response	Group	Mean of response
Absurd	1.500000	2n Cycle	0.360000	0	0.864139
Alone	0.666667	Basic	0.148148	1	0.301701
Divorced	0.500000	Graduation	0.439964	2	0.245192
Married	0.417736	Master	0.432877	3	0.160000
Single	0.513800	PhD	0.550936		
Together	0.396161				
Widow	0.592105				
YOLO	0.500000				

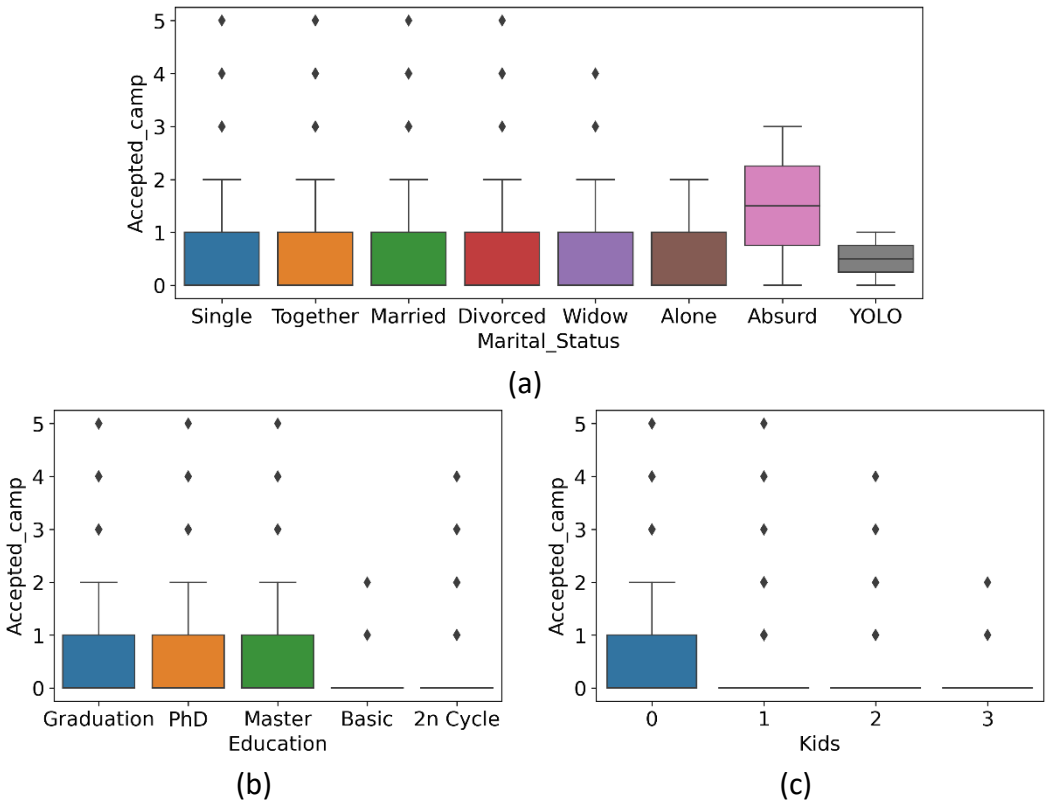


Fig. 7. Number of accepted campaigns versus (a) marital status, (b) education, (c) kids

3.5. Correlation between variables

Since there are many attributes (18 columns) that cannot be analysed in this report due to the limited number of words, this section will analyse the correlation of all variables in the dataset using a heatmap as shown in Fig. 8. As can be seen from Fig. 8, there is a strong correlation between Catalog purchases and Meat. It means that people who buy products by catalogue method also spend more money on meat. There are also clear correlations Catalog_Purchases and Wines, Store_purchase and Wines. People buy wines also frequently buy at stores and catalogues. A weaker correlation, but also positive, exists between fruits and Fish. On the other hand, there are some clear negative correlations between Web visits and Income, meat, catalogue purchase.

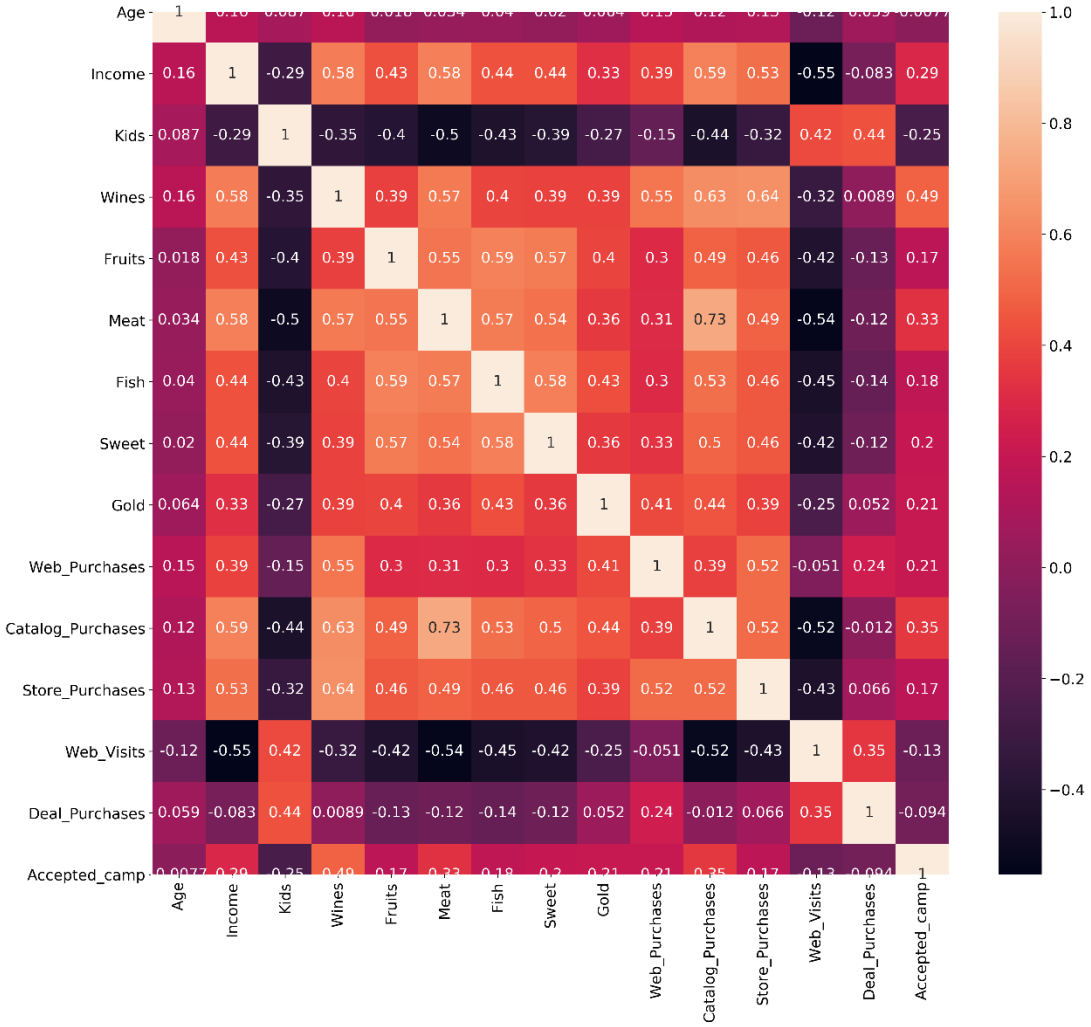


Fig. 8. Heatmap for whole dataset

4. Unsupervised analysis

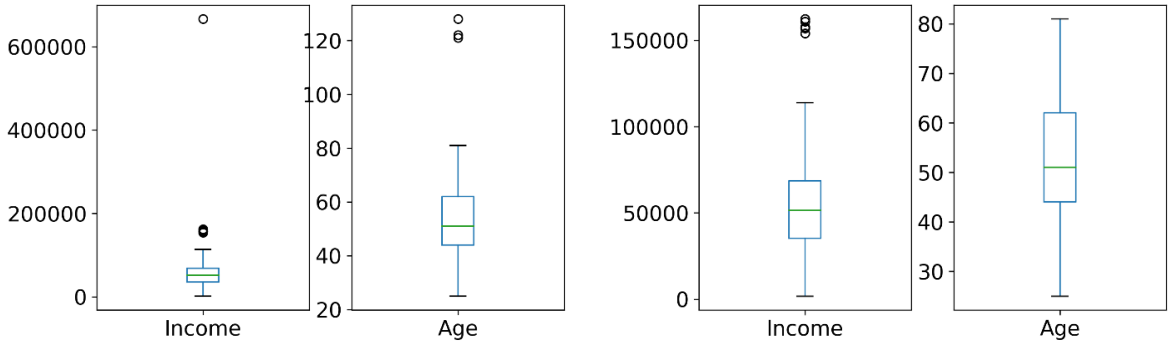
The objective of this section is to find customer segments based on their profiles, expenditure, and shopping habits by using KMeans and Hierarchical clustering. Challenging work is that the tidied dataset still has 16 columns including both charterers and numerical data types. The columns that contain characters data should be converted to numerically encoded labels. There are also some customers with extremely high income or very old (>120 years old). The data points corresponding to these customers should be considered as outliers and need to be removed. In addition, instead of directly using customer age, which is converted from their year of birth, for clustering, it is necessary to group them in different ranges of age. Therefore,

in this section, first, the preparation of data is described in section 4.1. Next, the clustering is presented in section 4.2.

4.1. Preparing data

4.1.1. Remove outliers

By using boxplots for all columns of the dataset, it is observed that there are some extreme outliers in the attributes “Income” and “Age”. As shown in Fig. 9(a), there is one customer who has an extremely high income (over 600000) and three customers who are over 120 years old. These data points are considered extreme outliers which may distort statistical analyses. Therefore, only customers which age is smaller than 100 and income smaller than 200000 are kept. The boxplot for Income and Age columns after removing outliers are shown in Fig. 9(b).



(a) (b)

Fig. 9. Boxplot for Income and Age (a) before (b) after removing outliers

4.1.2. Making age groups for customers

It is also observed that the customers’ ages vary in a wide range from 25 to 81. Therefore, in this section, the attribute “Age” of customers will be grouped into 6 groups:

- Customers with age ≤ 30
- Customers with $30 <$ age ≤ 40
- Customers with $40 <$ age ≤ 50
- Customers with $50 <$ age ≤ 60
- Customers with $60 <$ age ≤ 70
- Customers with age > 70

4.1.3. Scaling data

The dataset contains features highly varying in magnitudes, units, and range. Therefore, to ensure the accuracy of clustering, it is important to scale numerical data.

4.1.4. LabelEncoder

In the dataset, three columns contain characters including 'Education', 'Marital_Status2', 'Age'. To apply these columns for clustering as well as classifications, it is essential to convert string labels in these columns to encoded labels.

4.2. Clustering

In this section, KMeans clustering will be applied. The determination of the number of clusters, k, is presented in Section 4.2.1. Next, the optimisation for options used in KMeans clustering is presented in Section 4.2.2. The clustering results are presented in Section 4.2.3.

4.2.1. Determination of the number of clusters, k, for KMeans clustering

To find an appropriate k value, the well-known Elbow method [3] is used. Specifically, k is assumed to be in a range from 2 to 20. For each value of k, KMeans clustering is conducted, and the sum of square error (SSE) is obtained. The sum of square errors is stored in a vector and then plotted against values of k as shown in Fig. 10. As can be seen from Fig. 10, the elbow point is not clear and the value of k can be 4, 5, or 6. Therefore, to ensure a suitable value of k, a Hierarchical Clustering is also conducted as shown in Fig. 11. As shown in Fig. 11, the horizontal line shown in purple shows that the number of clusters $k = 4$ is also a possible option. Therefore, by combining findings from elbow method hand Hierarchical Clustering dendrogram, the number of clusters $k = 4$ is chosen.

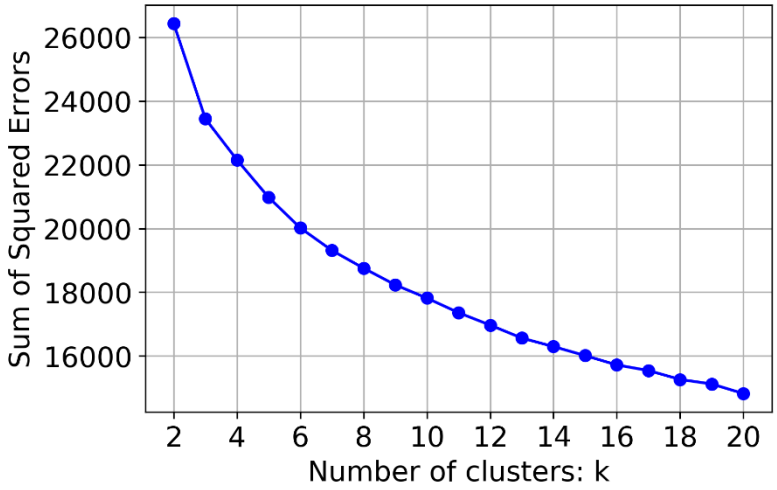


Fig. 10. Elbow plot: SSE versus k

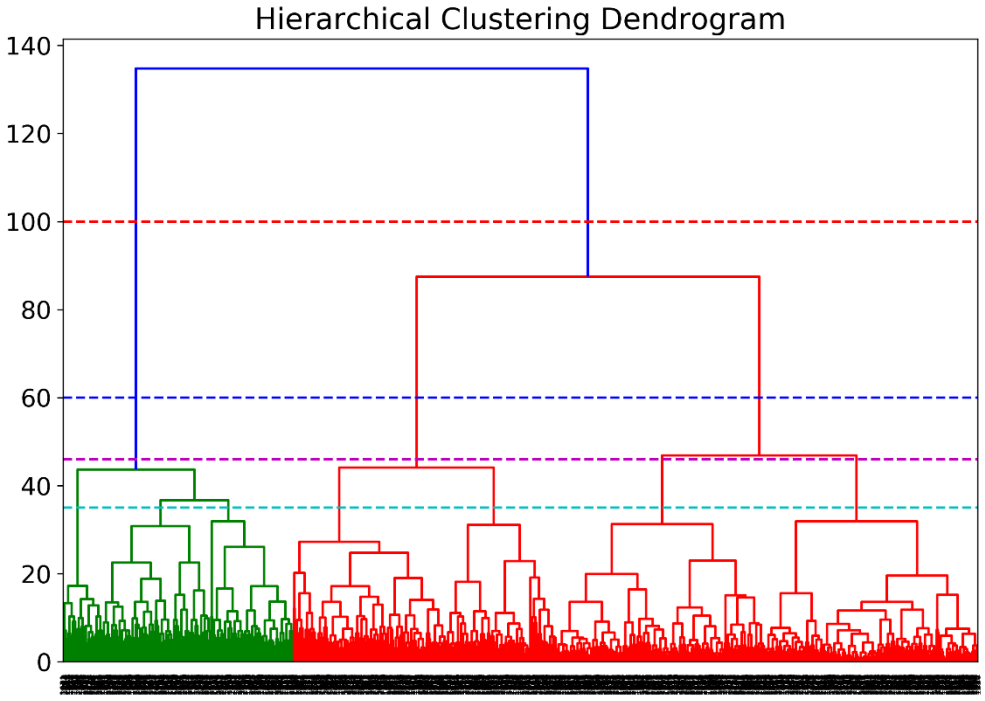


Fig. 11. Hierarchical clustering dendrogram

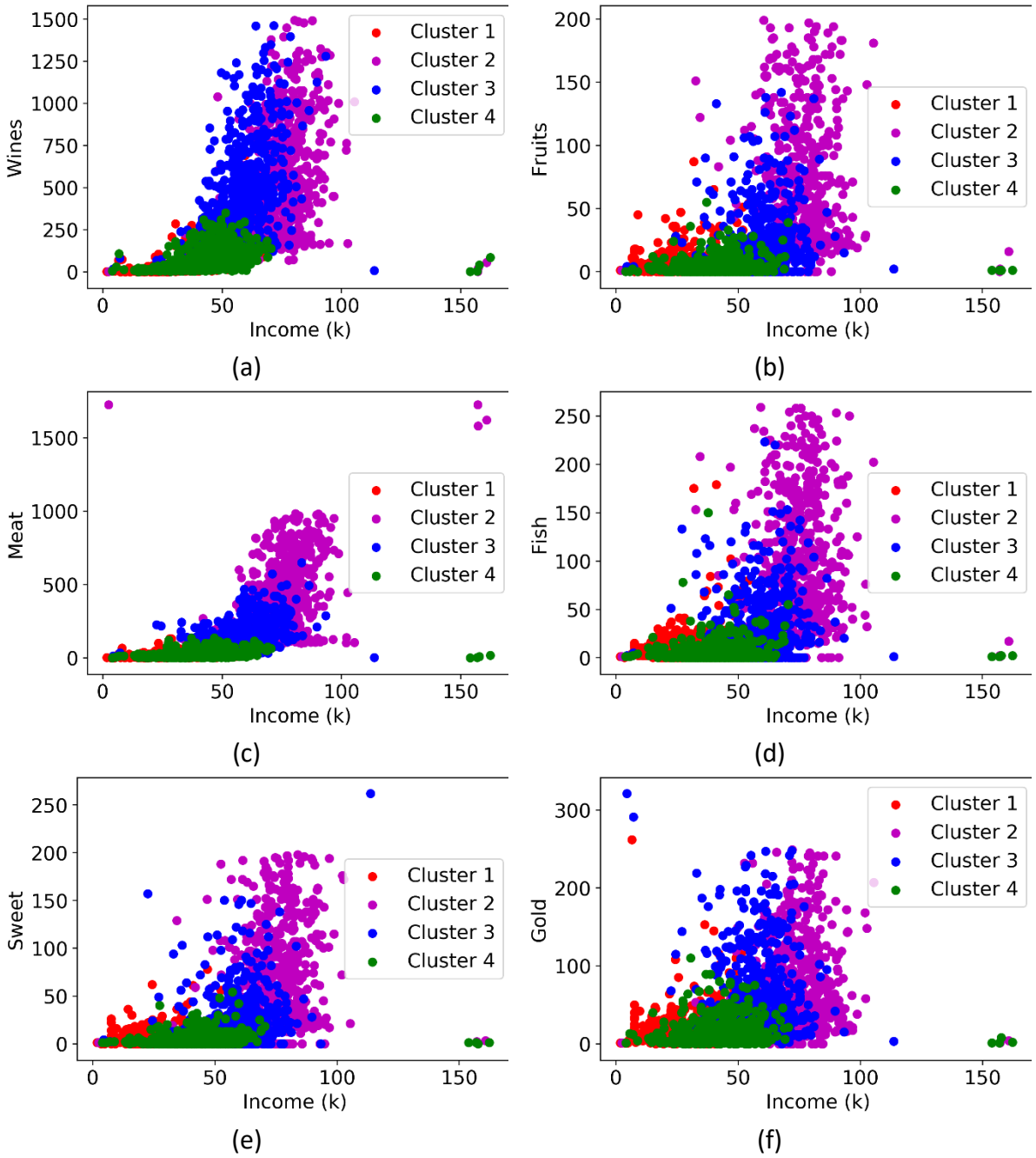
4.2.2. Optimise options for KMeans clustering

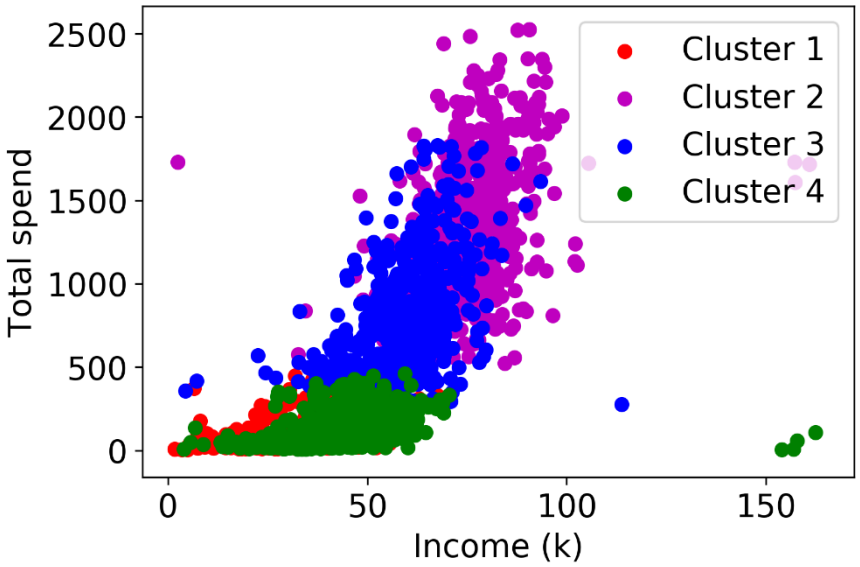
After choosing the number of clusters, it is necessary to optimise options for KMeans clustering analysis. Three algorithms: 'auto', 'full', 'elkan', and three values of "n_init": 10, 15, 20 are considered as possible options for KMeans clustering. Here, "n_init" is the number of times the k-means algorithm will be run with different centroid seeds [4]. By combining three possible algorithms with three possible values of "n_init", we have 9 possible KMeans clustering models. For each clustering, the sum of square error (SSE) is obtained and stored in a vector. Finally, by comparing the SSE values, the best option (Algorithm = 'full', n_init=20) is chosen to correspond to the minimum value of obtained SSE. It is noted that, in this study, all possible 9 options of KMeans clustering give nearly the same value of SSE.

4.2.3. Clustering results

After choosing the number of clusters ($k = 4$) and optimising options for clustering (Algorithm = 'full', n_init=20), a KMeans clustering for the whole dataset is conducted. Fig. 12 shows the plots of Income versus spending amount for different types of products as well as total spending. Here, total spending is calculated by summing the spending amount for all types of products. As can be seen from Fig. 12, clusters 2 and 3 spend much more than clusters 1 and 4 in all products. As a result, the total spending of cluster 2 and 3 are also much higher

than those of cluster 1 and 4. Moreover, the major of clusters 2 and 3 have incomes higher than around 52000. Meanwhile, the major of clusters 1 and 4 have incomes lower than around 52000.





(g)

Fig. 12. Plots of Income versus spending on (a) Wines, (b) Fruits, (c) Meat, (d) Fish, (e) Sweet, (f) Gold, (g) Total spend of 4 clusters.

5. Supervised learning

The purpose of this section is to classify customers based on their features, expenditure and shopping habits. Four methods including Logistic Regression, SVC, Decision Tree, and Gaussian Naive Bayes are used. To reduce the sensitivities of models' scores on the train-test splitting process, the Cross-validation technique is used.

5.1. Cross-validation technique

In this study, instead of splitting the dataset into training and testing datasets one time, the Cross-validation technique [5, 6] is used. Specifically, the dataset is divided into 10 smaller datasets as demonstrated in Fig. 13. Nine of ten datasets will be randomly selected as training data, meanwhile, the remaining will be the testing data. Therefore, there will have 10 possible cases of training and testing datasets. The classification models will be conducted based on these 10 cases of datasets, and the scores of each case will be output. To determine the accuracy of the classification model, the mean value of 10 scores is calculated.

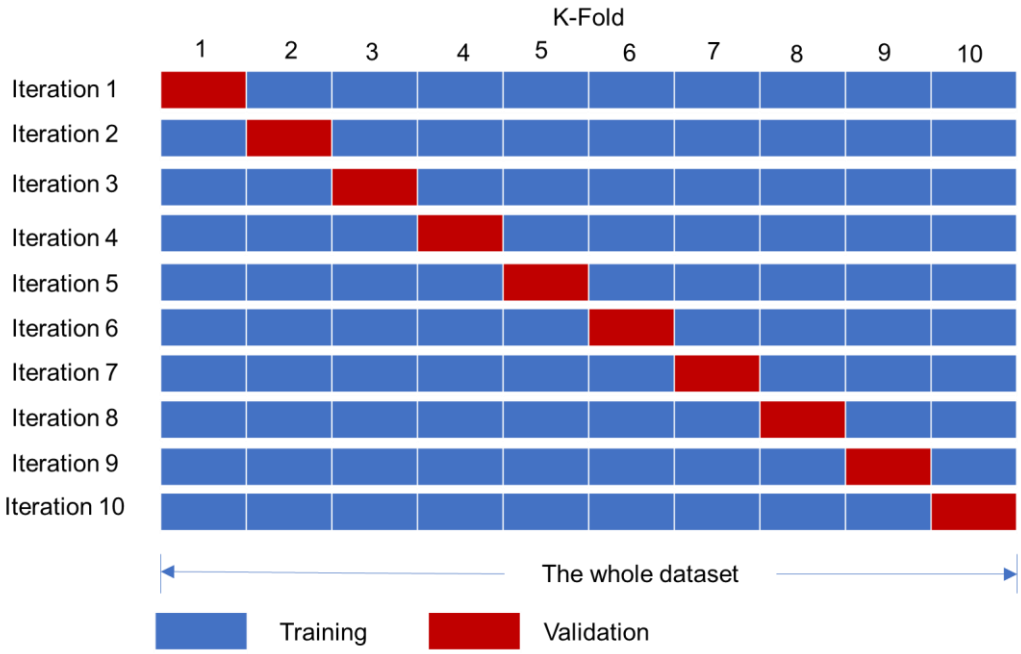


Fig. 13. Demonstration of training and testing datasets in cross-validation method with 10 folds.

5.2. Determining the output for classification models

As mentioned earlier, the overall purposed of classification analysis is to classify customers based on their behaviours and shopping habits. Therefore, the inputs (X) include all columns related to spending on products, places of purchases, Deal_Purchases, and Accepted_camp as: 'Wines', 'Fruits', 'Meat', 'Fish', 'Sweet', 'Gold', 'Web_Purchases', 'Catalog_Purchases', 'Store_Purchases', 'Web_Visits', 'Deal_Purchases', 'Accepted_camp'. However, there is no specific known output (y) in this dataset. Therefore, trial classifications with different possible outputs were conducted to find best suitable output.

The procedure to find the best option for output is shown in Fig. 14. Specifically, in each section, different classification models including Logistic Regression, Support Vector Classification [7], Decision Tree, and Gaussian Naïve Bayes are used. As mentioned in Section 5.1, the Cross-validation technique is used to estimate the mean score of each model. The best model with the best score will be found in each section. Finally, by comparing the best models obtained in all sections, the final best model with the best score will be determined. Details of this work is presented in Appendix A.

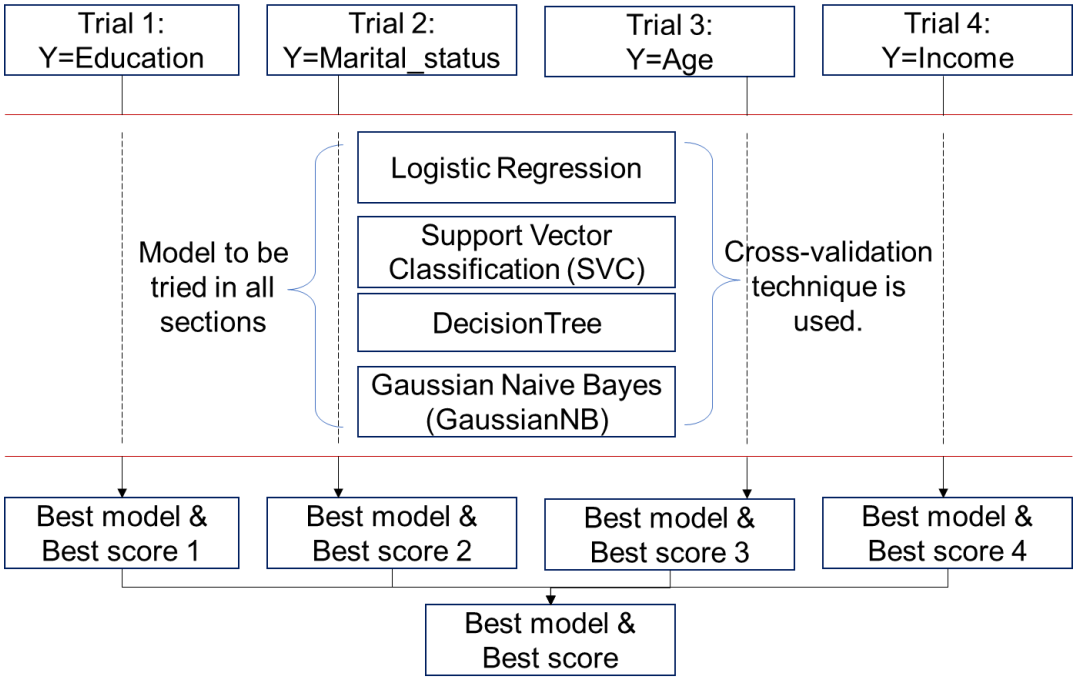


Fig. 14. Flowchart for optimisation of best classification model

As presented in appendix A, the most suitable output for the classification is the customer income. The best model is Logistic Regression with a mean score of 0.915. Therefore, it can be concluded that there is a strong relationship between customer income and their spending and shopping habits.

5.3. Classification results

To print out the detailed results such as confusion matrix and model scores, the data is split one time with 70% for training and 30% for testing. The Logistic regression model is rerun and the obtained model scores confusion matrix are given in Table 3.

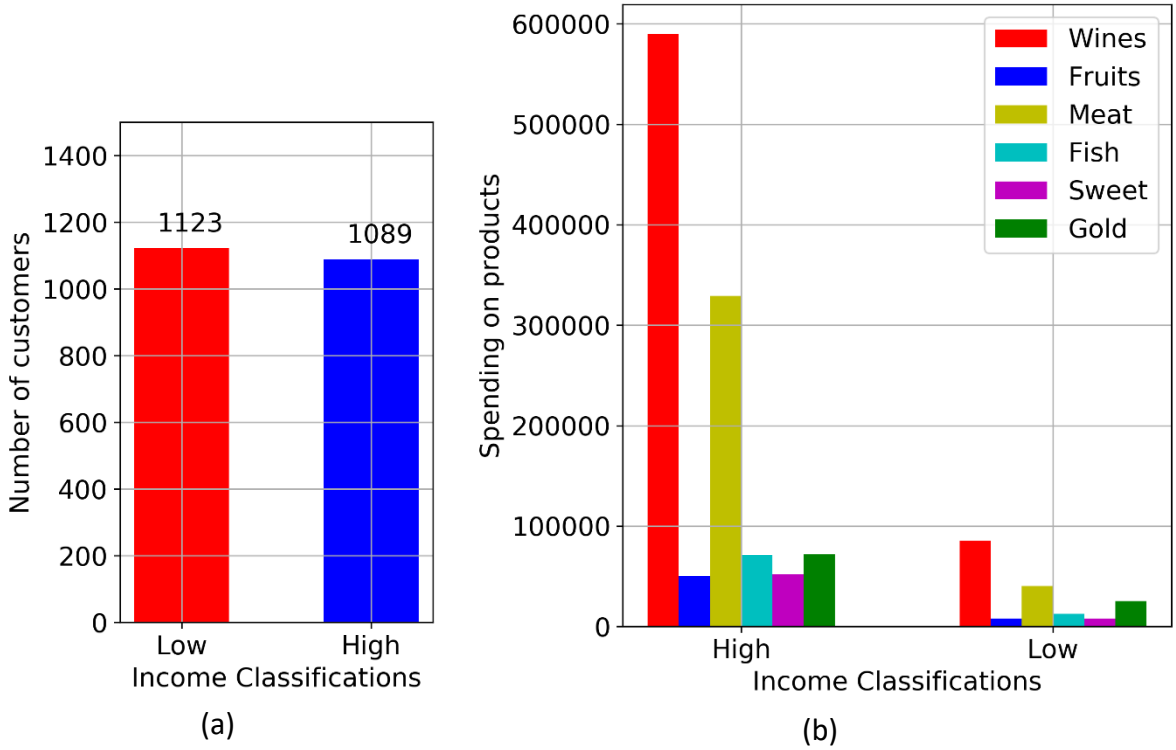
Table 3. Classification results using Logistic regression

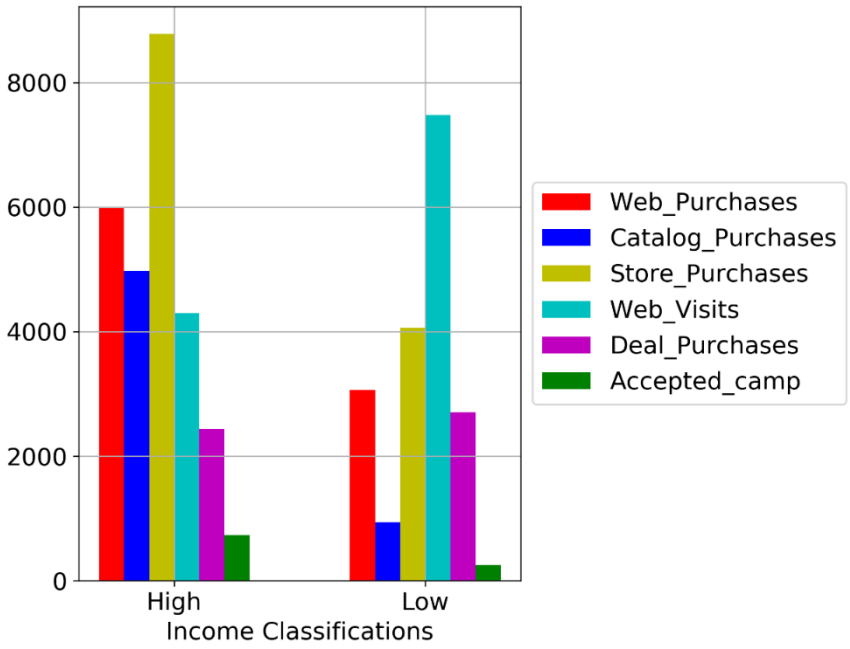
	precision	recall	f1-score	support		Confusion matrix	
0	0.94	0.87	0.9	326		TP = 284	FP = 42
1	0.88	0.95	0.91	338		FN = 18	TN = 320
accuracy			0.91	664			
macro avg	0.91	0.91	0.91	664			
weighted avg	0.91	0.91	0.91	664			

As can be seen from the confusion matrix, there are (284+42+18+320) = 664 data points in the testing data. The prediction results are TP = 284, TN = 320, FP = 42, FN = 18. It means that for precision 0, the model correctly predicts 284 times, wrongly 42 times. For precision 1, the

model correctly predicts 320 times and wrongly 18 times. The f1-score for precision 0 is: 284/(284+42) = 0.871, the f1-score for precision 1 is 320/(320+18)= 0.947. These scores were round as 0.87 and 0.95 as given in the table above.

The classification results are also presented in Fig. 15. As shown in Fig. 15(a), the number of low-income and high-income customers are relatively the same as each other. However, as shown in Fig. 15(b), the spending of high-income customers on all types of products are around 6 times higher than those of low-income customers. As shown in Fig. 15(c), high-income customers also purchase much more than low-income customers in all types of places, and they also accepted more advertising campaigns. However, they visit the web (Web_Visits) and purchase with a discount (Deal_Purchases) less than the other.





(c)

Fig. 15. Visualisation for classified groups (a) number of customers, (b) their spending on products, (c) other shopping habits

6. Conclusion

This report presents a customer personality analysis. First, a statistical summary of all features in the dataset, as well as the data visualisation, was presented. The unsupervised analysis using KMeans clustering for the dataset was conducted. The clustering results showed possible 4 clusters in which two clusters have higher incomes and high spending. Meanwhile, two other clusters have lower incomes and lower spending. Supervised learning was also conducted for the dataset. Various models such as Logistic Regression, SVC, Decision Tree, and Gaussian Naive Bayes were used. To reduce the sensitivities of the models' accuracies with the train-test splitting stage, the Cross-validation technique was used. Various trial outputs for the classifications were tried and the best-optimised classification model was determined. From this classification analysis, it is found that the customers are well classified into two groups of high- and low-income. It was also observed that the numbers of high- and low-income customers are nearly the same, but the amount they spent on products are significantly different. The high-income customers spend much higher than the others on all types of products. The high-income customers use all types of purchases more frequently but visit the company website and take discount purchases less than the others. These

observations can be useful to adjust products and marketing campaigns which can help the company to increase sales and use the marketing budget effectively.

7. Reflection

This coursework is my first data analysis project. My background is in business management. That is the reason why I chose this dataset for the report. My initial expectation is to find clear customer segments and trends in their shopping habits which can be used directly to adjust products and marketing campaigns for the company. However, there was much challenging work that I faced during the analysis.

First of all, the dataset includes many columns (customer's features), meanwhile, the number of rows is not too large. The dataset has few missing data and dealing with missing data was not a big issue. However, combining columns, removing unnecessary columns, converting datatype such as converting year of birth to age and then making age groups are very challenging for me because this is my first data analysis project (except for some practices during the course). In addition, the dataset does not have a specific output. Therefore, deciding which feature will be used for clustering, which feature will be the output for classification are very time consuming which requires many trial analyses. The Python codes submitted with this report is the finally tidied version after many different trials for clustering and classification. Due to time limitations, the clustering and classification results may not be the best. However, I had time to investigate different models and optimise their accuracy.

Appendix A. Trial classifications to find suitable output and best model

A.1. Education is the output

By using the output as the encoded labels in column Education (the encoded labels are 0, 1, 2, 3, 4), the mean score of the Logistic Regression, SVC, Decision Tree, and Gaussian Naive Bayes models are 0.531, 0.537, 0.494, 0.320, respectively. Since the scores of the models are very low, it can be concluded that the assumption of Education is the output is not good and there is no direct relationship between customers' education levels and customer shopping habits.

A.2. Marital status is the output

Similar to the previous case, in this case, the Logistic Regression, SVC, Decision Tree, and Gaussian Naive Bayes models are 0.641, 0.644, 0.596, 0.609, respectively. These scores are slightly higher than those in Section 5.2.1. However, they are still very low scores which indicate that the assumption of using output as Marital status is also not good.

A.3. Age is the output

In this case, the accuracies of the abovementioned models are even lower with scores of 0.324, 0.346, 0.375, 0.289, respectively. Therefore, column Age should not be the target output for the classification analysis.

A.4. Income is the output

Unlike columns Education, Marital_status, and Age which include encoded labels, column Income include numerical data with a very wide range of value as shown in Fig. 9. Therefore, to use this column as the target output for the classifications, it is essential to divide the salaries of customers into groups as we did for the Age column as mentioned in Section 4.1. Moreover, from the clustering results shown in Section 4, customers are relatively well divided into two groups of low income and high income. Therefore, in this section, the customers will be grouped into two groups. Customers with income higher than the mean value of income will be labelled as 0 which mean they are in the high-income group. Otherwise, they will be labelled as 1 which mean they are in the low-income group.

After labelling the income column as mentioned above, the Logistic Regression, SVC, Decision Tree, and Gaussian Naive Bayes classifications are conducted. The obtained mean scores for these models are 0.915, 0.911, 0.896, 0.889, respectively. It can be seen that all four models give quite high accuracies. The best model is Logistic Regression with a mean score of 0.915. Therefore, it can be concluded that the Income column is the best output for the classifications and there is also a strong relationship between customer income and their spending and shopping habits.

Appendix B. Environment and software versions

Python version: Python 3.9.7

jupyter Notebook version: 6.3.0

Libraries: pandas, numpy, matplotlib, seaborn, sklearn, scipy

References

[1] Campbell, N.C. and Cunningham, M.T., 1983. Customer analysis for strategy development in industrial markets. *Strategic Management Journal*, *4*(4), pp.369-380.
[2] Insider, May 2012: https://www.businessinsider.com/robbie-bach-explains-why-the-zune-flopped-2012-5?r=US&IR=T
[3] Joshi, K.D. and Nalwade, P.S., 2013. Modified k-means for better initial cluster centres. *International Journal of Computer Science and Mobile Computing*, *2*(7), pp.219-223.
[4] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
[5] Schaffer, C., 1993. Selecting a classification method by cross-validation. *Machine Learning*, *13*(1), pp.135-143.
[6] Refaeilzadeh, P., Tang, L. and Liu, H., 2009. Cross-validation. *Encyclopedia of database systems*, *5*, pp.532-538.
[7] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html