Customer Personality Analysis using the Unsupervised Learning Methods

1. Introduction

The purpose of this technical analysis report is to analyze customer personalities in the data set provided by Kaggle. It is a detailed analysis of the company's ideal customers. The results of the analysis can help companies better understand their customers, and allow companies to modify their products according to the specific needs, concerns and behaviours of different types of customers (Patel, 2021).

By exploring different types of customer segments and explaining why they are a segment using the methods of unsupervised learning (K-Means, Agglomerative, K-Medoids), we will identify the key characteristics of their clusters.

2. Exploratory Data Analysis

(a) Data Exploration

Despite the customer ID, there are 28 features with a total of 2240 data points in this data set. Two of the features are categorical features that only contain text data, namely “Education” and “Marital_Status”. Looking closer at the data set, we found that most features can be categorized into four different subsets: Customer Information, Product Information, Product Promotion, and Purchase Locations. Features “Z_CostContact” and “Z_Revenue” have been removed because they simply contain duplicate values, and do not contribute to customer personality analysis.

- i) Customer Information

ID	Customer's unique identifier
Year_Birth	Customer's birth year
Education	Education Qualification of customer
Marital_Status	Marital Status of customer

Income	Customer's yearly household income
Kidhome	Number of children in customer's household
Teenhome	Number of teenagers in customer's household
Dt_Customer	Date of customer's enrollment with the company
Recency	Number of days since customer's last purchase
Complain	1 if customer complained, 0 otherwise

- ii) Product Information

MntWines	Amount spent on wine
MntFruits	Amount spent on fruits
MntMeatProducts	Amount spent on meat
MntFishProducts	Amount spent on fish
MntSweetProducts	Amount spent on sweets
MntGoldProds	Amount spent on gold

- iii) Product Promotion

NumDealsPurchases	Number of purchases made with a discount
Response	1 if customer accepted the offer in the last campaign, 0 otherwise

AcceptedCmp1	1 if customer accepted the offer in the 1st campaign, 0 otherwise
AcceptedCmp2	1 if customer accepted the offer in the 2nd campaign, 0 otherwise
AcceptedCmp3	1 if customer accepted the offer in the 3rd campaign, 0 otherwise
AcceptedCmp4	1 if customer accepted the offer in the 4th campaign, 0 otherwise
AcceptedCmp5	1 if customer accepted the offer in the 5th campaign, 0 otherwise

- iv) Purchase Locations

NumWebPurchases	Number of purchases made through company's website
NumCatalogPurchases	Number of purchases made using catalog
NumStorePurchases	Number of purchases made in stores
NumWebVisitsMonth	Number of visits to company's website

(b) Feature Engineering

For precise data summary and further analysis, we need to modify some features. For instance, we used Year_Birth to calculate the age of customers, and added a feature named “Age” to our data set. The customer’s age is equal to 2021 minus the customer’s year of birth. We then deleted Year_Birth because Age can replace it.

We also calculated the total expenditure of customers on all types of products, and created a new column called “Spending”. Moreover, we added one more feature to indicate the number of kids and teenagers in the customer’s family, namely “Num_Children”. Lastly, we added another

feature namely “Enrollment_Time” to indicate the customer’s enrollment time with the company, and removed Ct_Customer. The entire modifications are as follows:

- Features added: Age, Spending, Num_Children, Enrollment_Time

Age	Customer's age
Spending	Total amount spent on all products
Num_Children	Number of kids or teenagers in customer's household
Enrollment_Time	Customer's enrollment time with the company

- Features removed: ID, Birth_Year, Dt_Customer, Z_CostContact, Z_Revenue

We remove ID, Z_CostContact and Z_Revenue because they are redundant and useless to customer personalities. We also removed Year_Birth and Dt_Customer because we already have Age and Enrollment_Time to replace them.

- Feature modified: Marital_Status

Marital_Status is a categorical feature that contains a list of eight different marital status: divorced, widow, alone, absurd, YOLO, single, together, married. However, there are only two major types of marital status: with or without a partner. Therefore, we divided all marital status into two categories: “Single” and “Pair”, in order to replace the original status.

That is, if the customer has a marital status of “Together” or “Married”, we then replace the status with “Pair” since the customer has a partner; Otherwise the status is replaced with “Single”.

Original text data in Marital_Status	**Updated text data in Marital_Status**
Divorced, Widow, Alone, Absurd, YOLO, Single	Single
Together, Married	Pair

(c) Numeric Summaries

After feature processing, the data set now contains 28 features, with Age, Spending, Num_Children and Enrollment_Time added, while ID, Year_Birth, Z_CostContact, Z_Revenue and Dt_Customer removed. There are 26 numerical features and 2 categorical features. A quick look at the specific values of these numerical features are as follows.

	Income	Kidhome	...	Spending	Age	Num_Children	Enrollment_Time
count	2216.00000	2216.00000		2216.00000	2216.00000	2216.00000	2216.00000
mean	52247.25135	0.44179		607.07536	52.17960	0.94720	102.08404
std	25173.07666	0.53690		602.90048	11.98555	0.74906	6.74782
min	1730.00000	0.00000		5.00000	25.00000	0.00000	90.30000
25%	35303.00000	0.00000		69.00000	44.00000	0.00000	96.30000
50%	51381.50000	0.00000		396.50000	51.00000	1.00000	102.15000
75%	68522.00000	1.00000		1048.00000	62.00000	1.00000	107.93333
max	666666.00000	2.00000		2525.00000	128.00000	3.00000	113.60000

Figure 1: Data set description of numeric features

The description table above only contains numeric features. For the categorical features Education and Marital_Status, we will see if the counts are even across the categories.

```
Graduation    1116
PhD            481
Master         365
2n Cycle       200
Basic           54
Name: Education, dtype: int64
```

```
Pair      1430
Single     786
Name: Marital_Status, dtype: int64
```

Figure 2: Counts for customer's education level and marital status

Both categories are imbalanced. The data shows that the number of customers with graduation level is much more than other education levels, and the number of customers with basic education level is significantly less than the others. In addition, the number of customers with partners is far more than those without a partner.

(d) Graphing Summaries

i) Correlation among features

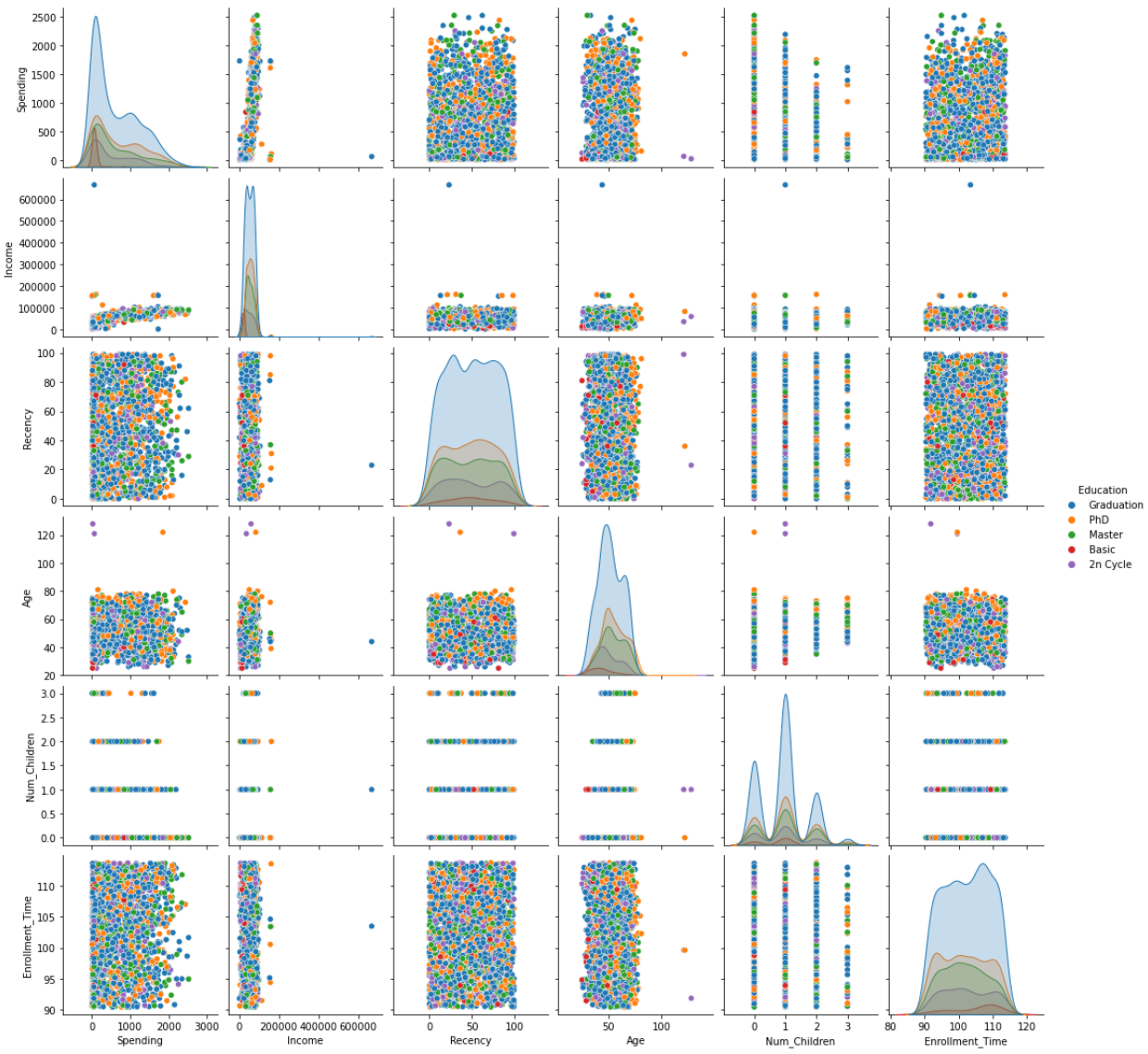


Figure 3: Relationships between variables of customer characteristic features (sorted by education level)

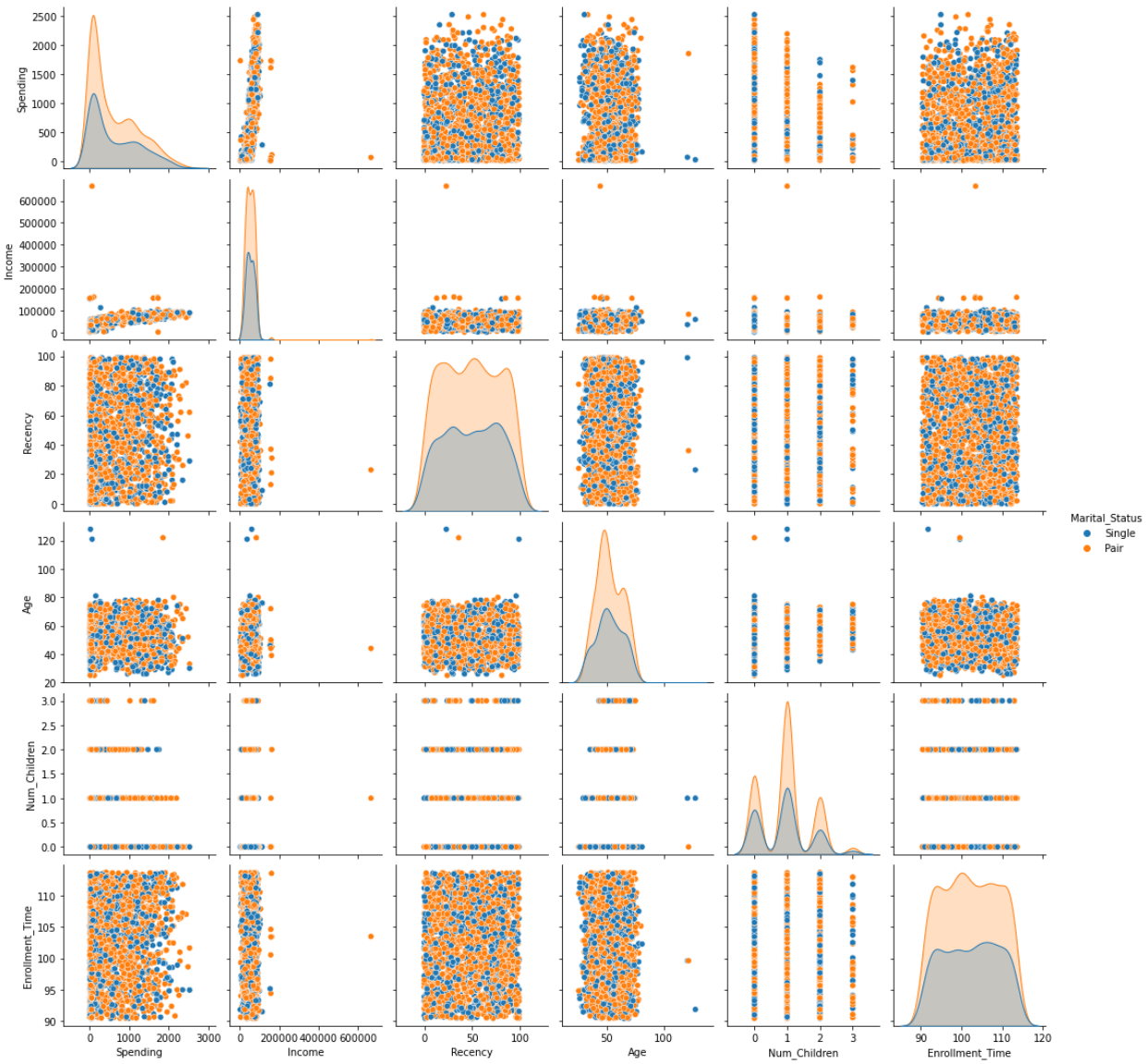


Figure 4: Relationships between variables of customer characteristic features (sorted by marital status)

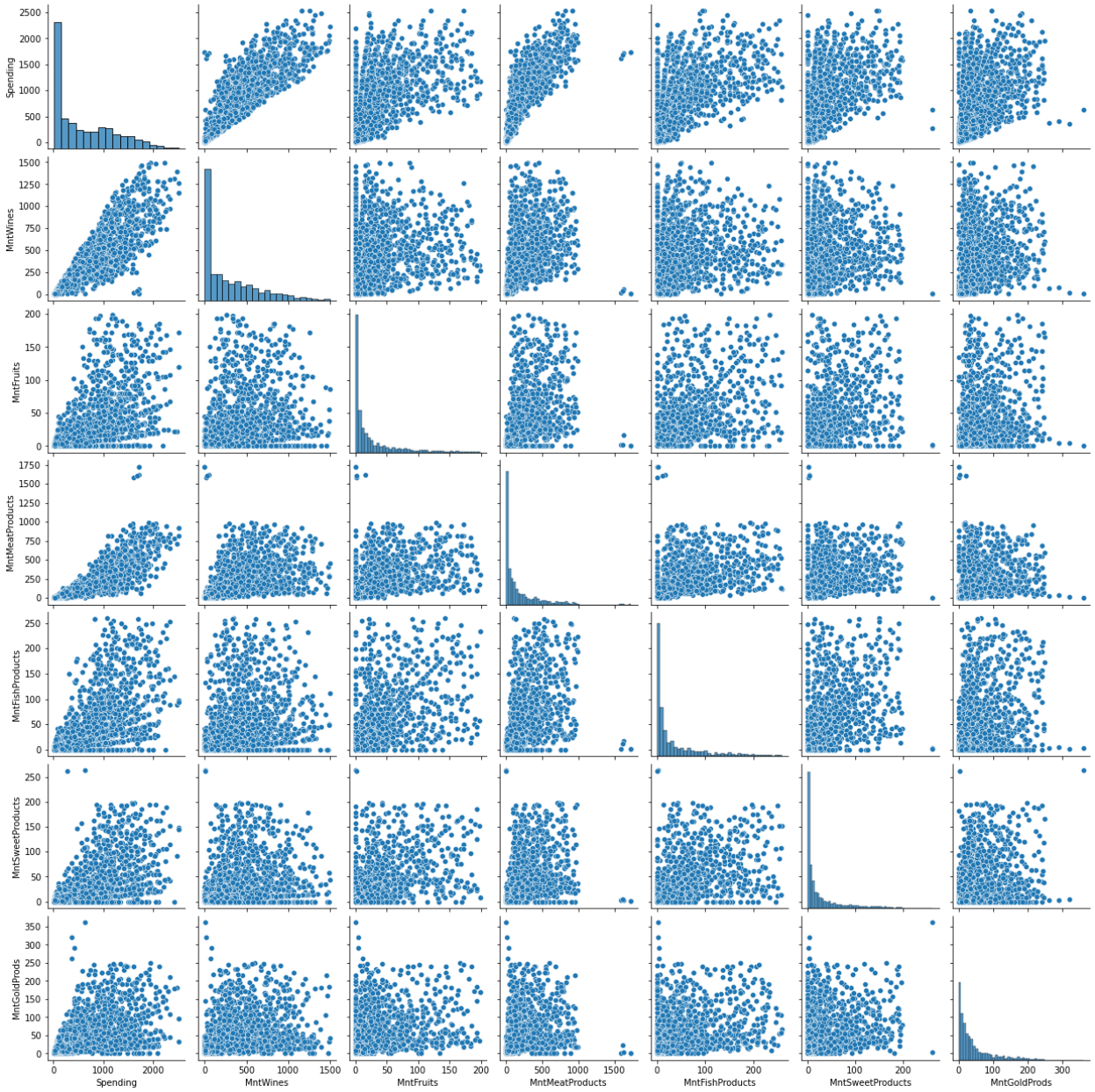


Figure 5: Relationships between variables of spending on different products

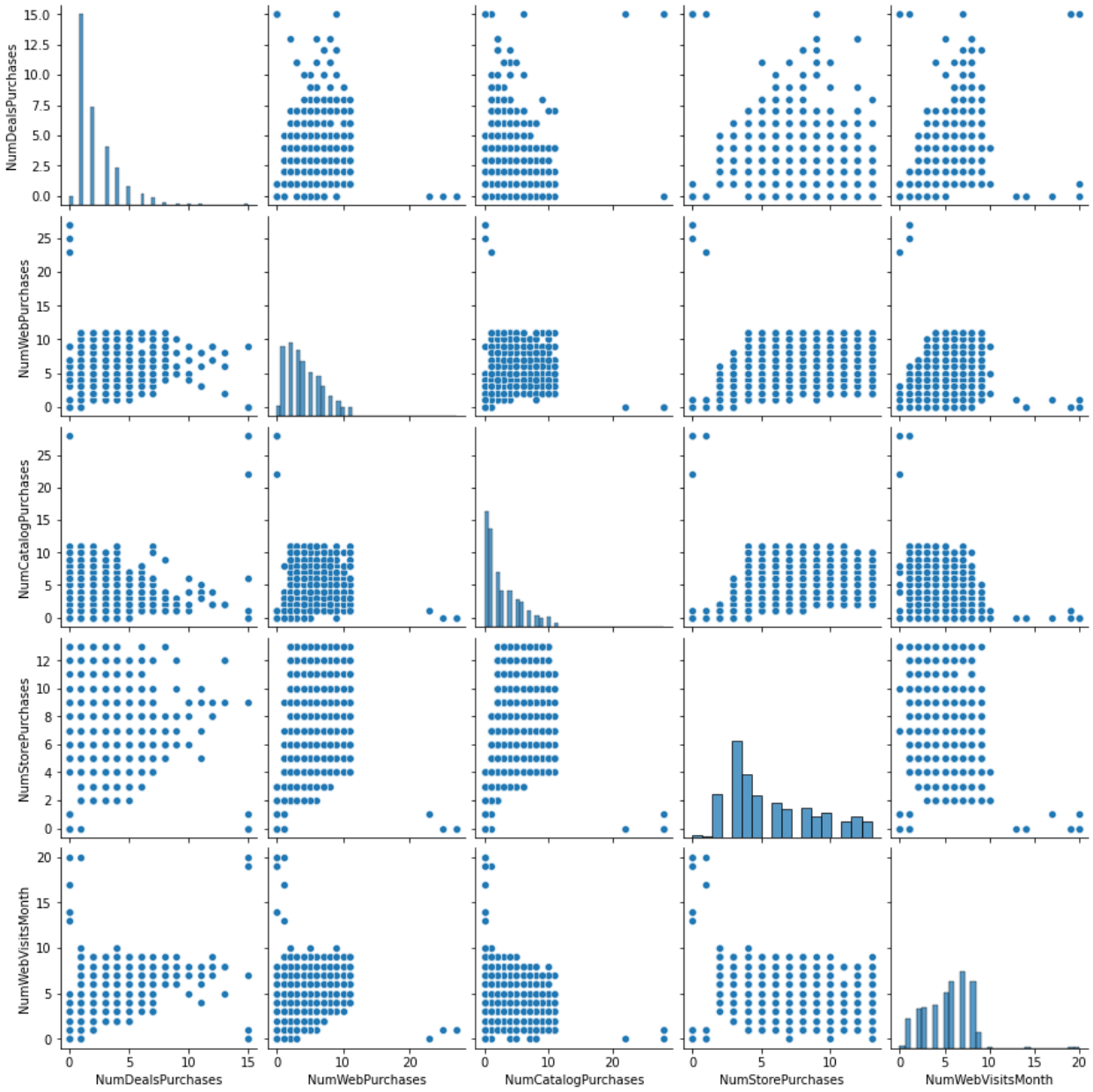


Figure 6: Relationships between variables of purchases in different places

It can be seen from Figure 3 that customers with an education level of Graduation seem to have a higher expenditure and higher annual incomes, followed by PhD and Mater. Customers with Basic education level have the least spending and income. In addition, customers with higher annual incomes tend to spend more. As shown in Figure 4, among the high-income customer groups, the majority are not in a single status. Similarly, in customer groups that have a higher expenditure, most of them have a partner. Besides, there are potential outliers in Age and Income.

From Figure 5, we can see that in general, customers who spend less on a type of product seem to spend less on other types of products. From Figure 6, customers who do not often use

catalogs to make purchases seem to purchase less online. Those who like to shop in stores might not purchase on websites very often. But customers who like to purchase on websites also like to purchase in stores. Generally speaking, more customers tend to purchase in stores.

ii) Distribution plots

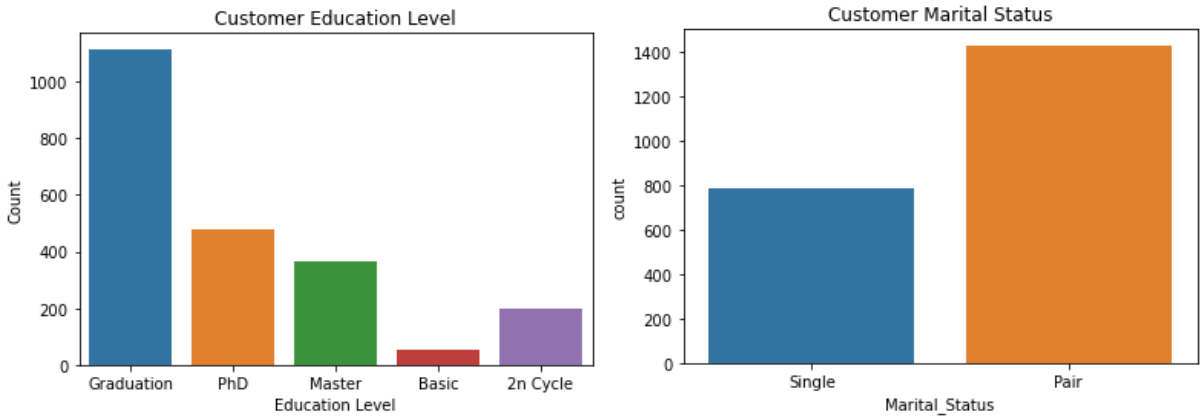


Figure 7-8: Counts for customer's education level; Counts for customer's marital status

The number of customers with an education level of Graduation is significantly more than the others, followed by PhD and Mater. When it comes to marital status, the number of customers with a partner is far more than that of single customers.

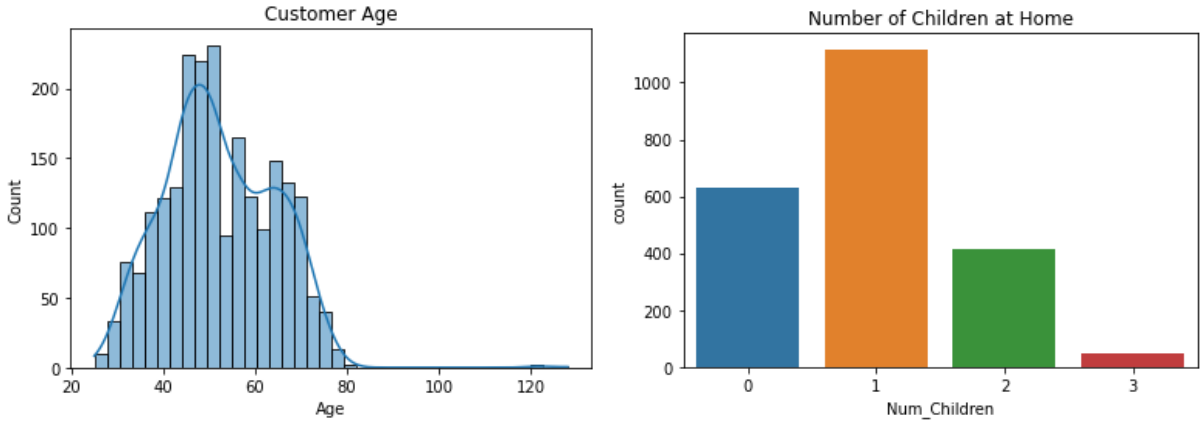


Figure 9-10: Counts for customer age; Counts for number of children at home

From Figure 9, most of the customers are middle-aged people, and many of them are in their 40s to 50s. From Figure 10, most customers have at least one child.

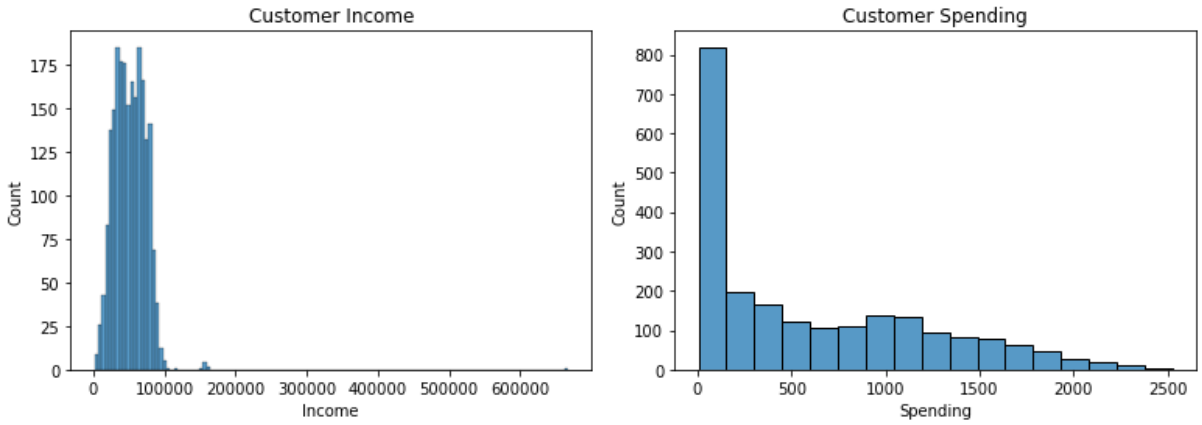


Figure 11-12: Counts for customers' annual income; Counts for customer total spending

As shown in Figure 11, both high-income and low-income earners are in the minority. Most of the customers' annual income is approximately between 35000 and 70000. From Figure 12, the number of customers spending below 500 is significantly more than others.

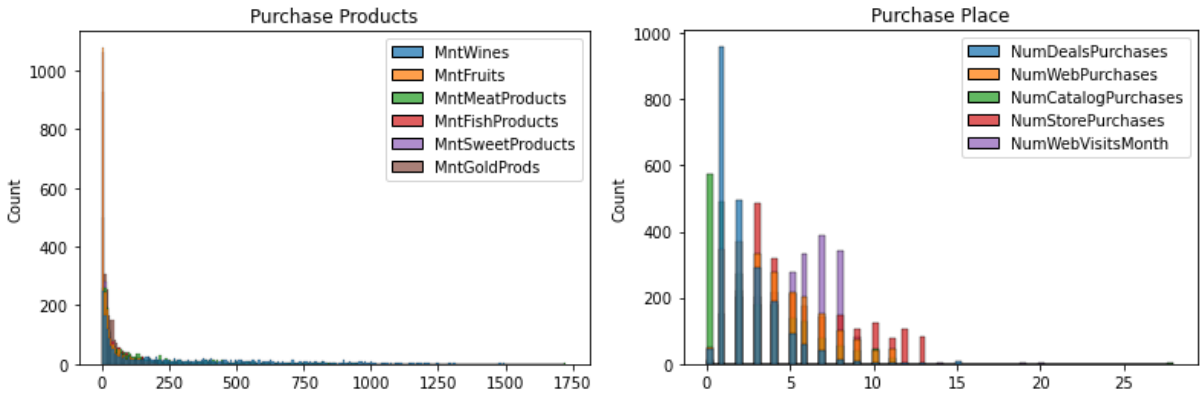


Figure 13-14: Counts for products purchased; Counts for purchase locations

From Figure 13, there are a large number of customers who do not consume fish, fruit or sweets at all. Customers tend to spend more on wine or meat. From Figure 14, the number of customers who tend to purchase in stores or on websites is the largest.

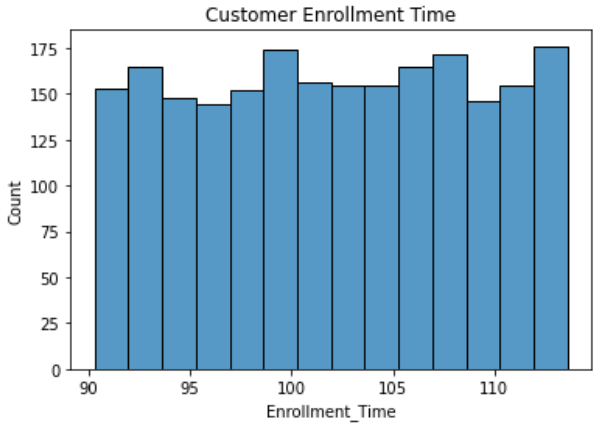


Figure 15: Counts for customer's enrollment time with the company

As it is seen from Figure 15, customers' enrollment time with the company is evenly distributed. There seems to be no obvious sign of relationship between the number of customers and their enrollment time with the company.

iii) More scatter plots

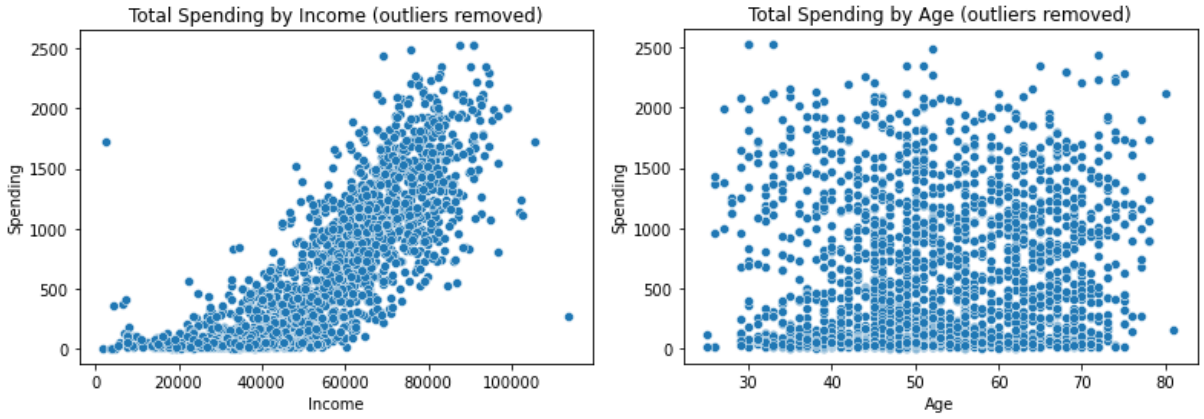


Figure 16-17: Customers' total spending by annual income with outliers removed; Customers' total spending by age with outliers removed

As we can see from Figure 16 and Figure 17, there is evidence that as the annual income increases, customers' total spending on products will also increase substantially. Besides, the total spending of middle-aged customers (aged 45-55) seems to be higher than that of other people, followed by elderly customers.

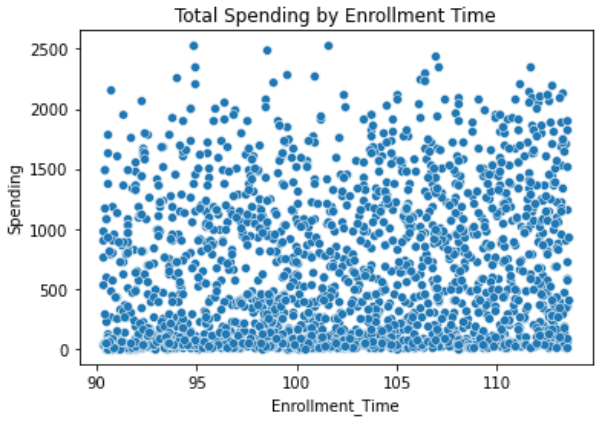


Figure 18: Customers' total spending by their enrollment time with company

From Figure 18, there is no obvious correlation between customers' enrollment time with the company and their total spending. It seems that customers with longer enrollment time tend to spend slightly more, but it is not significant.

3. Data Preparation

(a) Data Cleaning

- Missingness and N/A

There is no N/A in this data set, but there are 24 missing values in Income. The percentage of missingness is calculated as 24/2240 = 1.07%. The probability of missingness in Income is very small, thus we can ignore it.

- Outliers

From the above correlation plots, we found that there are potential outliers in Age and Income. We constructed boxplots and distribution plots for these two features, and computed their whiskers to investigate potential outliers. Any points that fall outside of the range of the whiskers might be outliers.

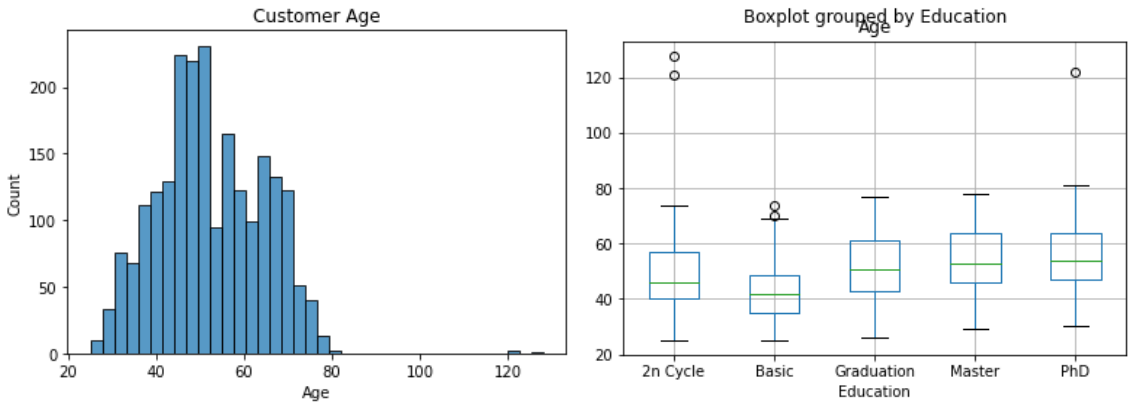


Figure 19-20: Distribution plot of customer age; Boxplot of customer age by education

As is seen from the fitures, there might be a few outliers in Age, indicating that there are 3 incredible customers over 120 years old. We calculated the whisker of Age to check whether these data points are outliers.

IQR of Age = 75% quantile of Age - 25% quantile of Age = 18

25% quantile of Age - 1.5 * IQR of Age = 17

75% quantile of Age + 1.5 * IQR of Age = 89

Therefore, any point outside the range of [17,89] in Age is considered an outlier. Using this method, we removed 3 outliers in Age because they show that the 3 customers are over 89 years old.

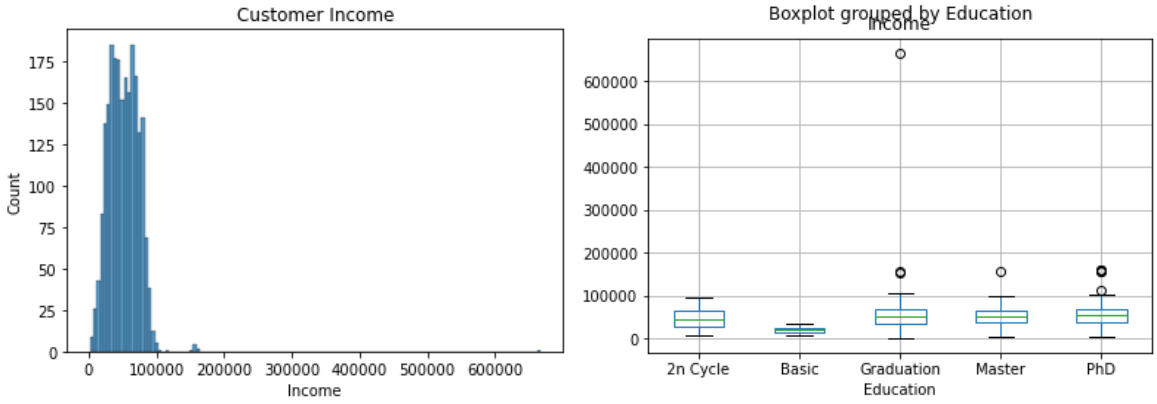


Figure 21-22: Distribution plot of customer income; Boxplot of customer income by education level

There seems to be some outliers in Income as well, one of which has an income higher than $600000, and the other four are close to $200000.

IQR of Income = 75% quantile of Income - 25% quantile of Income = 33219

25% quantile of Income - 1.5 * IQR of Income = -14525.5

75% quantile of Income + 1.5 * IQR of Income = 118350.5

The annual income cannot be negative. Thus any value in Income greater than $118350.5 is considered an outlier. Using this method, we removed a total of 8 outliers in Income.

(b) Dealing with categorical features

We used Label Encoding to process these two categorical features that contain text data: Education and Marital_Status. Label Encoding converted the categories to numeric equivalents as follows:

Education	Education Encoded
2n Cycle	0
Basic	1
Graduation	2
Master	3
PhD	4

Figure 23: 2n Cycle = 0, Basic = 1, Graduation = 2, Master = 3, PhD = 4

Marital_Status	Marital_Status Encoded
Pair	0
Single	1

Figure 24: Pair = 0, Single = 1

Education and Marital_Status are encoded from categorical features to numeric features. Thus, after the Label Encoding process, all features in our data set are now numerical.

(c) Scaling

We performed Min-Max scaling on the data set so that no variables will dominate our clusters in the subsequent modeling. The formula used is as follows:

$$x_T = \frac{x_0 - min(x_0)}{max(x_0) - min(x_0)} \; where \; x_0 \; is \; the \; original \; data$$

This scaling method transfers the original columns to make all observations fall in the range of [0,1]. In this way, all features will be equally treated during the modeling process.

4. Modeling

In order to analyze customer personalities, we used unsupervised learning methods to fit customer data into clusters. The three methods used for clustering are: K-Means Clustering, Agglomerative Clustering, and K-medoids Clustering.

Before starting to build the model, we first determined how many clusters will be appropriate to make the models perform well. If there are too few clusters, people who act differently may be put in the same cluster; But if there are too many clusters, people who act similarly might be put in different clusters. To choose the best number of clusters, we ran a few K-Means models with different numbers of clusters to compare their score (inertia).

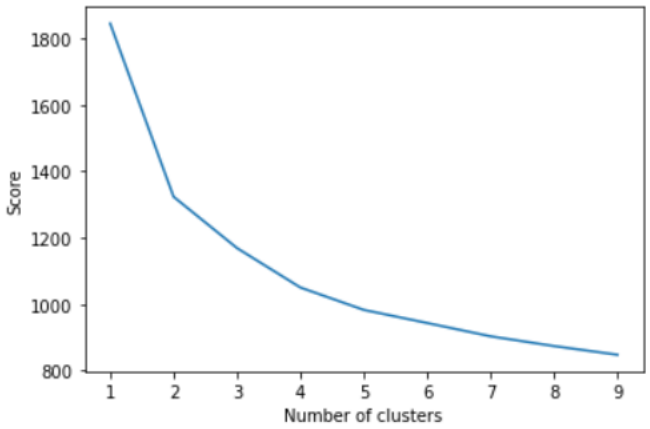


Figure 25: Relationship between inertia and the number of clusters in K-Means

From the graph, n = 4 seems to be a reasonable Elbow point, hence the K-Means model is fitted with 4 clusters.

Here are some processing details of each model:

(a) Process of K-Means Clustering

- As mentioned above, a K-Means model will run with 4 clusters;
- The algorithm will run 10 times with different starting cluster means;
- In each run, the algorithm will run at most 300 iterations;

- Categorical features are removed from the scaled data set. Since K-Means cannot handle categorical data well, only numerical explanatory variables are included in the modeling.

(b) Process of Agglomerative Clustering

- In order to compare with the K-Means model, the algorithm of Agglomerative Clustering will also run with 4 clusters;
- The distance metric used is Euclidean;
- Euclidean distance would not properly deal with the categorical data, so only numerical explanatory variables are included in Agglomerative Clustering;
- The model uses Ward linkage which would try to minimize the within cluster variation each time it merges.

(c) Process of K-Medoids Clustering

- K-medoids Clustering is used to check whether the model will perform better if it contains categorical data;
- In order to compare well with the above models, the K-Medoids algorithm will also run with 4 clusters;
- The method used is PAM since it is more accurate;
- The distance metric used is the Gower's metric, which is able to calculate distance for a mix of numerical data and categorical data.

5. Insights

Based on the built models, we will try to find out which features can help us to find customers who will accept the offer in the last campaign.

(a) Insights generated from K-Means Clustering

First in the K-Means model, we have 4 clusters. We tried to find out which cluster has the highest proportion of data whose response is equal to 1. By calculation, there are 611 in cluster 0, 535 in cluster 1, 547 in cluster 2, 522 in cluster 3. The proportions of data that response = 1 in every cluster and the whole data are as follows:

```
#the proportion of data that response =1 in cluster 0
cluster0_Knum/611
```

```
0.11947626841243862
```

```
#the proportion of data that response =1 in cluster 1
cluster1_Knum/525
```

```
0.11238095238095239
```

```
#the proportion of data that response =1 in cluster 2
cluster2_Knum/547
```

```
0.2943327239488117
```

```
#the proportion of data that response =1 in cluster 3
cluster3_Knum/522
```

```
0.07662835249042145
```

```
#the proportion of data that response =1 in the whole data
df_mod['Response'].sum()/len(df_mod['Response'])
```

```
0.1510204081632653
```

Figure 26: The proportion of data whose response = 1 in every cluster

From the calculation, we can find that the data in cluster 2 is more likely to be response = 1. Then we just need to analyze what is the difference between cluster 2 and the others.

Thus, we plot some characteristics for each cluster.

The first part is the four main features: “Spending”, “Age”, ”Num_Children”, and “Enrollment_time”. They are able to connect to the ability of consuming by our common knowledge.

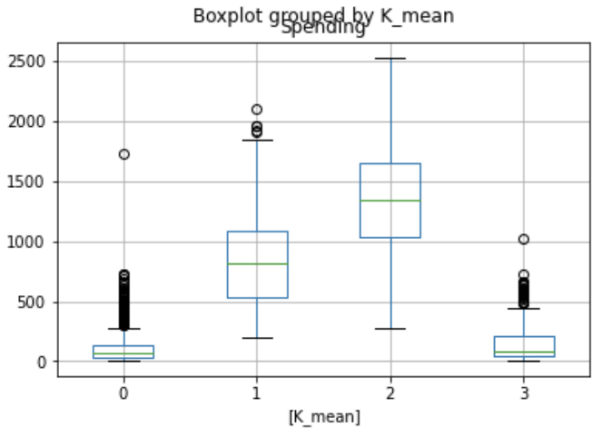


Figure 27: The boxplot of total spending by K_Means

Clearly our target cluster 2 has higher spending than the others. That means people who spend more are more likely to accept the offer in the last campaign.

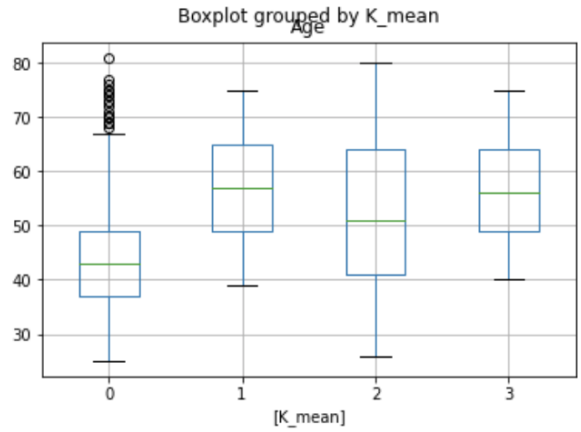


Figure 28: The boxplot of age by K_Means

In this figure, we find there is no significant difference between cluster2 and the others. That means the features “Age” is probably not an important feature.

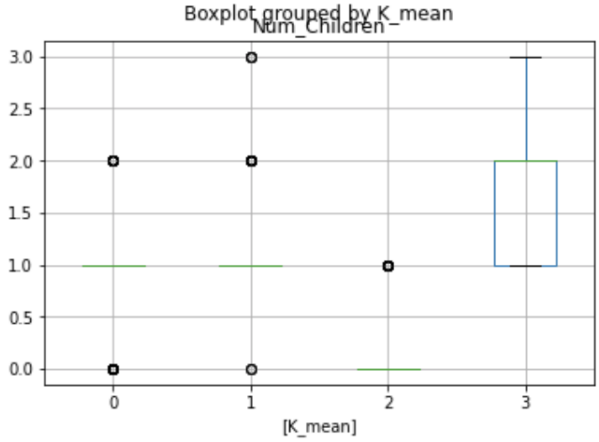


Figure 29: the boxplot of Num_children by K_Means

We can find the Num_Children of cluster 2 is the lowest. Thus, people who have less children are more likely to accept the offer in the last campaign.

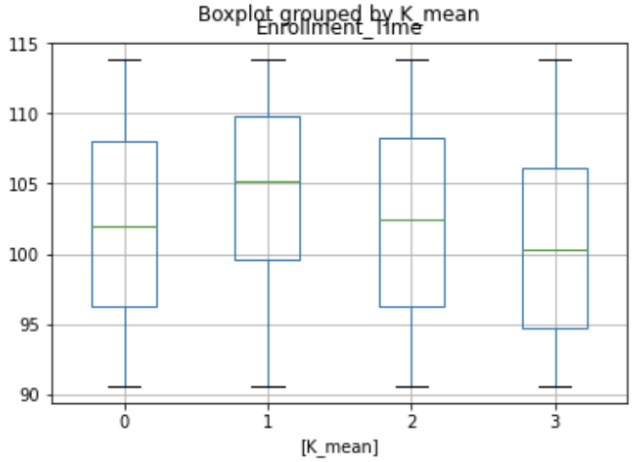


Figure 30: The boxplot of Enrollment_time by K_Means

There is no obvious difference between cluster2 and the others. That means the features "Enrollment_Time" is probably not an important feature, or the interaction between other features are important.

From the four figures, We find the spending and Num_Children will influence our target. Actually more children means the customer needs to spend more money on children which decreases the ability of the customer to spend. So we have a brief idea: "People who have strong consuming ability will accept the offer in the last campaign."

Then we analyze the details of spendings to see whether we can find something.

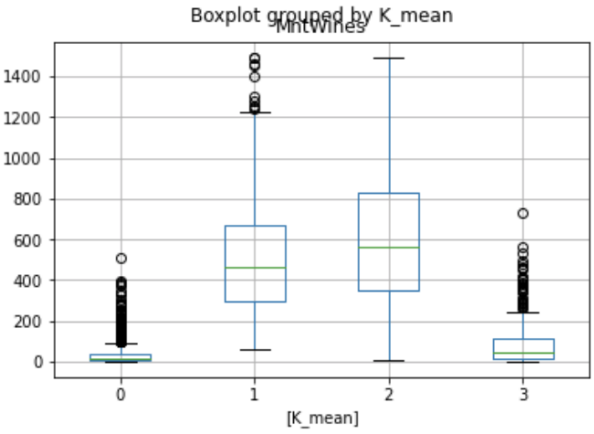


Figure 31: the boxplot of Mntwines by K_Means

From Figure 31, we found that customers in Cluster 2 spend more money on wine.

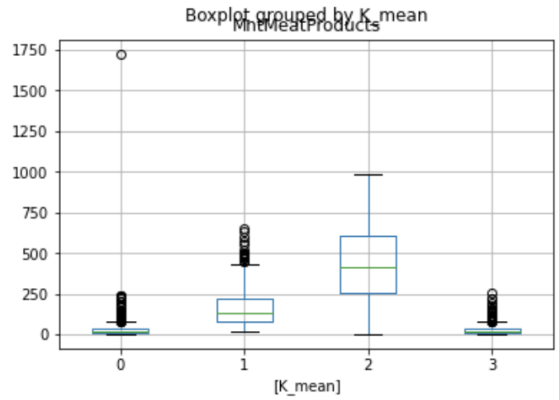


Figure 32: the boxplot of MntMeatProducts by K_Means

From this figure, customers in Cluster 2 spend more money on meat.

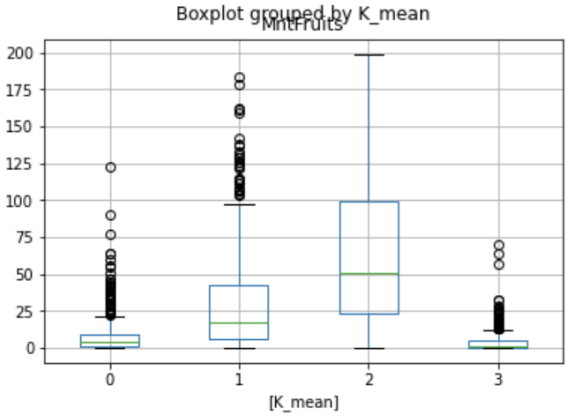


Figure 33: the boxplot of Mntfruits by K_Means

From this figure, customers in Cluster 2 spend more money on fruits.

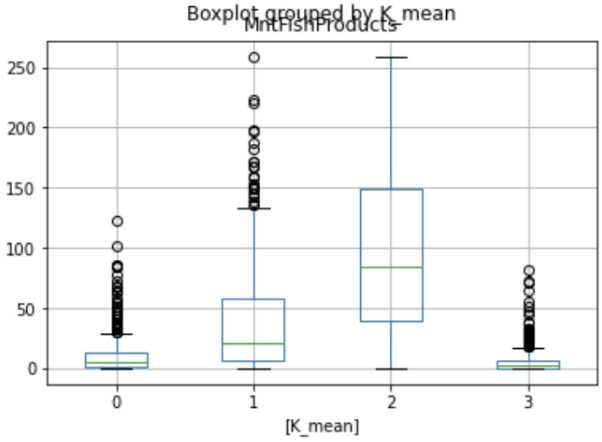


Figure 34: the boxplot of MntfishProducts by K_Means

From this figure, customers in Cluster 2 spend more money on fish.

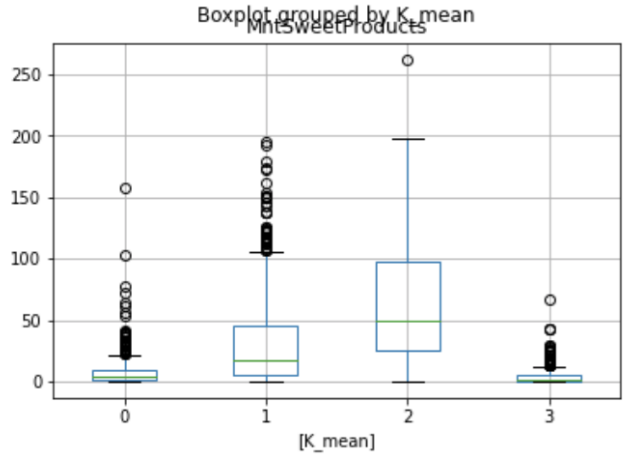


Figure 35: the boxplot of MntsweetProducts by K_Means

From this figure, customers in Cluster 2 spend more money on sweets.

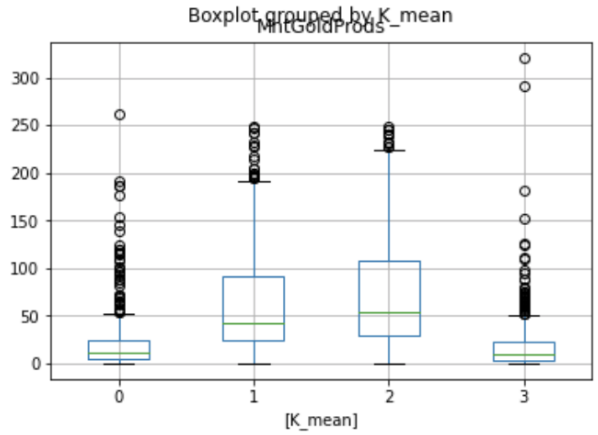


Figure 36: the boxplot of MntGoldProducts by K_Means

From Figure 36, customers in Cluster 2 spend more money on gold.

From the above plots, we cannot find too much information because the total amount of spending of cluster 2 is much higher than the rest. Thus, it is reasonable that customers in cluster 2 spend more money on every aspect than the others. However, we can still find something. We found that cluster 1 also spends a lot on many products, yet the proportion of people with a response equal to 1 is low.

Reviewing the difference between cluster 0 and cluster 1, we found that the spending of cluster 1 is more than cluster 0. However, the age of cluster 1, the enrollment time, and the number of children are higher. Compare cluster 0 and cluster 1, we found that customers who have many children but spend a little are more likely to accept the offer than those who have many children but spend a lot. Actually, that still can be concluded into the consuming ability. People who have the same consuming ability if they spend too much then the rest consuming ability becomes less.

Then we try to analyze the influence of purchasing places.

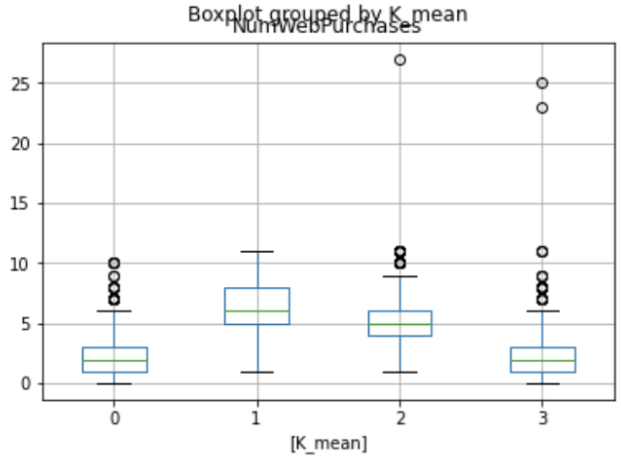


Figure 37: the boxplot of MntwebProducts by K_Means

Number of purchases made through the company's website of cluster 1 is the highest. Recall the result we concluded in the previous part, cluster 1 spent too much. If customers purchase a lot through the company's website, it means they spend too much. Then their consuming ability will decrease, which means that the probability of these customers to accept the offer in the last campaign will also decrease.

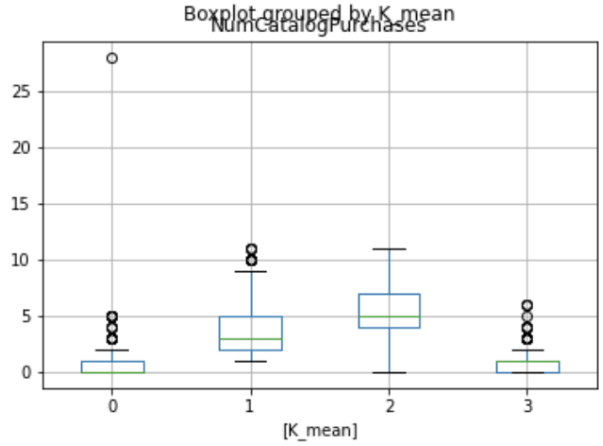


Figure 38: the boxplot of MntCatalogProducts by K_Means

From Figure 38, we found that cluster 2 made more purchases using catalog.

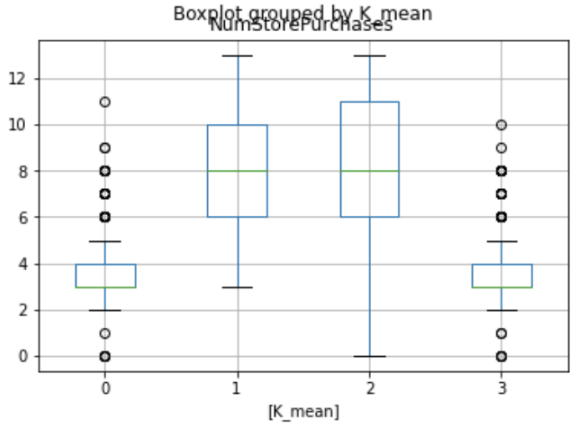


Figure 39: the boxplot of MntstoreProducts by K_Means

Cluster 2 made more purchases in stores.

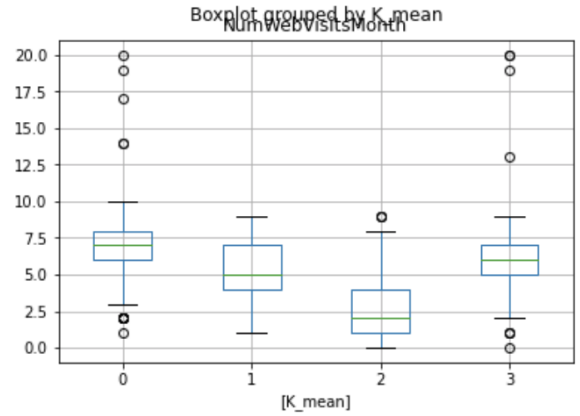


Figure 40: the boxplot of MntWebvisitMonth by K_Means

Number of Cluster 2 visits to the company's website is the least.

These four figures above show that our target cluster 2 makes more purchases in store or using catalog instead of online. The cluster that purchases a lot online is less likely to accept the offer in the last campaign. The reason behind that is people are more likely to spend too much online. We can conclude that too much online shopping is more likely to reduce customers' consuming ability.

Conclusions of the K_Means model:

The company should focus on customers who have:

- Strong consuming ability;
- Low number of children (which is the most important feature that influences the consuming ability directly);
- High spending (for those customers with no children, they would like to spend money while they have strong sunsuming ability).

For people who have weak consuming ability (number of children is large), less spending means rest consuming ability is stronger. Besides, online shopping is more likely to lead people to spend too much that will decrease their consuming ability.

(b) Insights generated from Agglomerative Clustering

Similar to the K-Means model, we also have 4 clusters in the Agglomerative model. We will find out which cluster has the highest proportion of data whose response is equal to 1. By calculation, there are 857 in cluster 0, 540 in cluster 1, 452 in cluster 2, 356 in cluster 3. The proportions of data that response = 1 in every cluster and the whole data are as follows:

```
#the proportion of data that response =1 in cluster 0
cluster0_Anum/857
```

```
0.10035005834305717
```

```
#the proportion of data that response =1 in cluster 1
cluster1_Anum/540
```

```
0.30185185185185187
```

```
#the proportion of data that response =1 in cluster 2
cluster2_Anum/452
```

```
0.10398230088495575
```

```
#the proportion of data that response =1 in cluster 3
cluster3_Anum/356
```

```
0.10393258426966293
```

```
#the proportion of data that response =1 in the whole data
df_mod['Response'].sum()/len(df_mod['Response'])
```

```
0.1510204081632653
```

Figure 41: The proportion of data whose response = 1 in every cluster

From the calculation, we found that the data in cluster 1 is more likely to have response = 1. Then we just need to analyze what is the difference between cluster 1 and the others. The

performance of the Agglomerative model is slightly better than the K_Means model.

Thus, we plotted some characteristics for each cluster.

The first part is the four main features: “Spending”, “Age”, ”Num_Children” and “Enrollment_time”. They are able to connect to the consuming ability by our common knowledge.

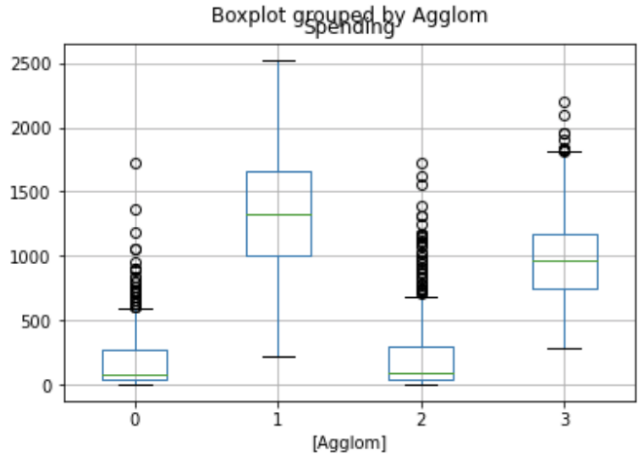


Figure 42: The boxplot of spending by Agglom

From Figure 42, it is evident that our target cluster 1 has a higher spending than the others. It means that customers who spend more tend to accept the offer in the last campaign.

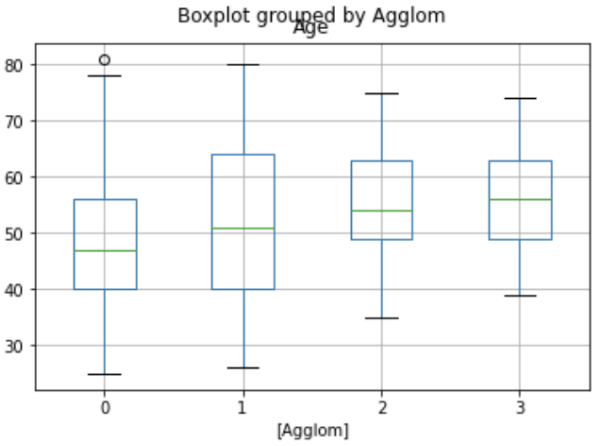


Figure 43: The boxplot of age by Agglom

In this figure, we found that there is no significant difference between cluster 1 and the others. It means that the feature “Age” is probably not an important feature.

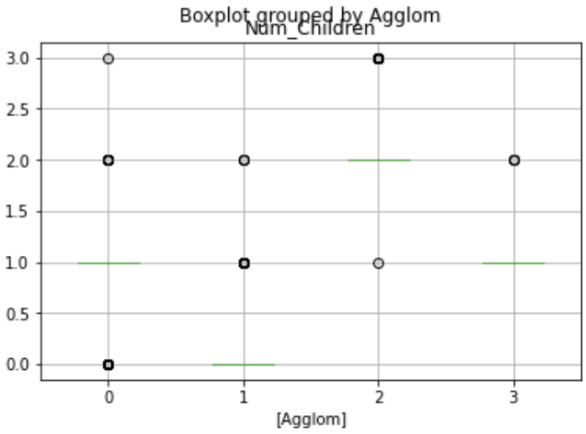


Figure 44: The boxplot of Num_Children by Agglom

From Figure 44, we found that the Num_Children of cluster 1 is the lowest. Thus, people who have less children are more likely to accept the offer in the last campaign.

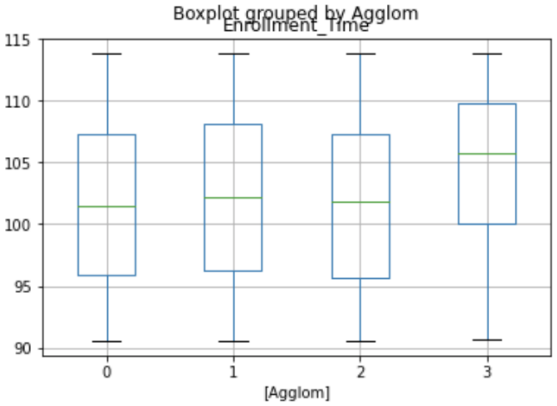
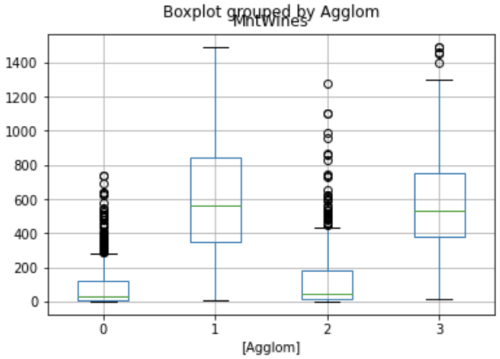


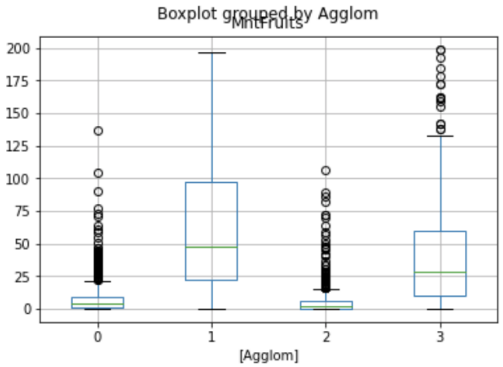
Figure 45: The boxplot of Enrollment_Time by Agglom

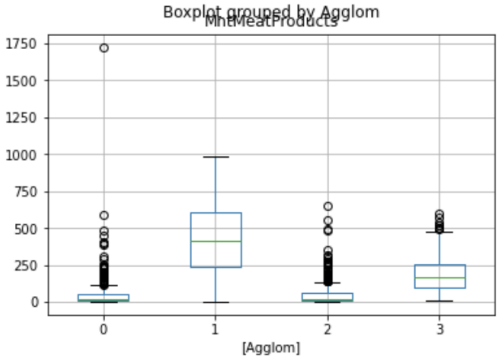
There is no significant difference between cluster 1 and the others. It means that the feature "Enrollment_Time" is probably not an important feature, or the interaction between other features is important.

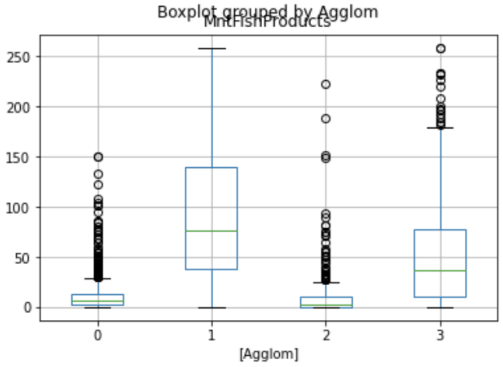
There is evidence that these figures are similar to those of the K-Means model.

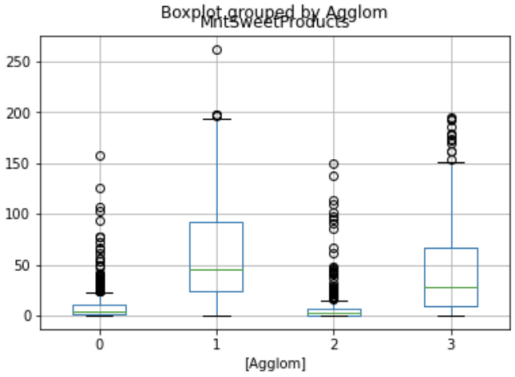
From the four figures above, we found that Spending and Num_Children will influence our target. In fact, more children means that customers need to spend more money on their children, which will reduce their spending power. So we came to the same conclusion as the above K-Means model: people with stronger consuming ability are more inclined to accept the offer in the last campaign.

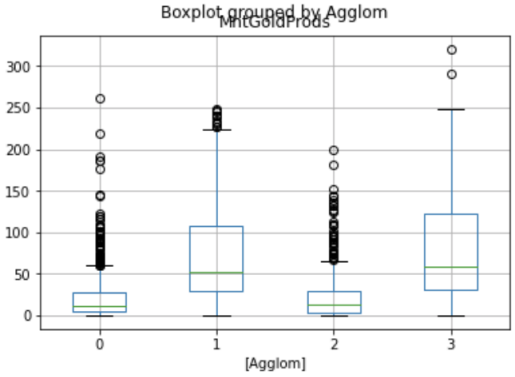

Boxplot grouped by Agglom
MntWines
[Agglom]


Boxplot grouped by Agglom
MntFruits
[Agglom]


Boxplot grouped by Agglom
MntMeatProducts
[Agglom]


Boxplot grouped by Agglom
MntFishProducts
[Agglom]


Boxplot grouped by Agglom
MntSweetProducts
[Agglom]


Boxplot grouped by Agglom
MntGoldProds
[Agglom]

Figures 46-51: The boxplot of MntWines/MntFruits/MntMeatProducts/MntMeatProducts /MntFishProducts/MntSweetProducts/MntGoldProds by Agglom

The above figures are very similar to those of the K_Means model. Therefore, we have reached a similar conclusion: if customers with the same consuming ability spend too much, the rest of the consuming ability will become smaller.

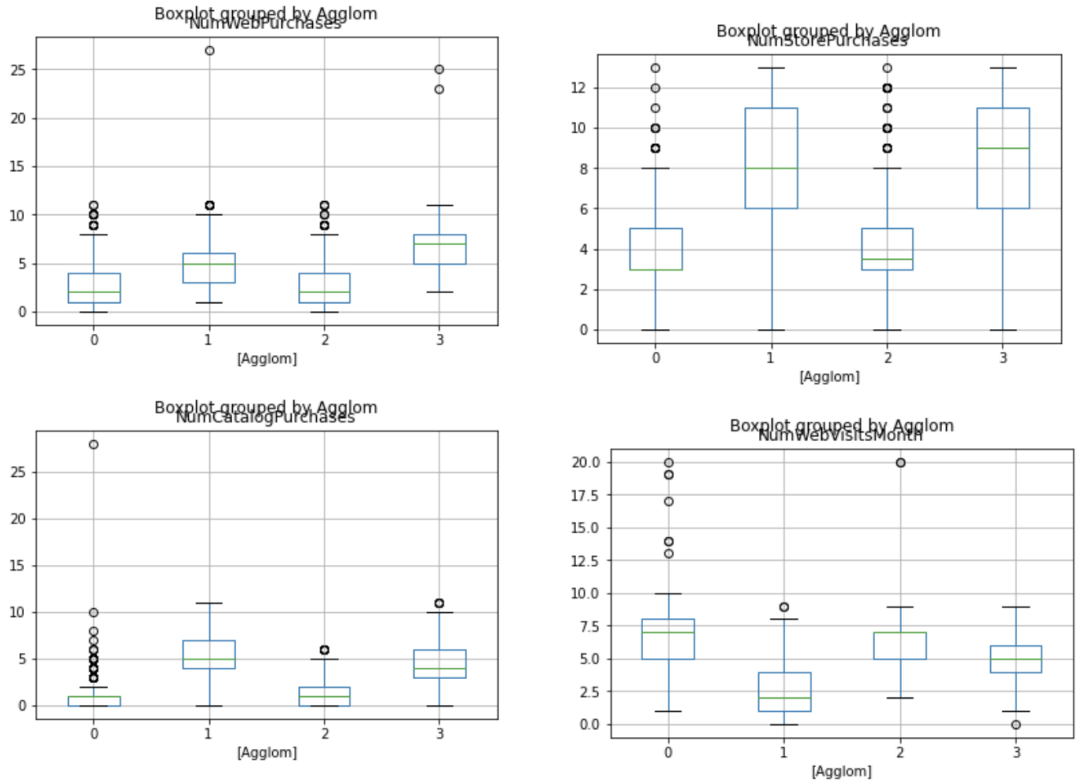


Figure 52-55: The boxplot of NumWebPurchases/NumCatalogPurchases / NumStorePurchases/NumWebVisitsMonth by agglom

Number of purchases made through the company's website of cluster 3 is the highest. Recall the result we concluded in the previous part, cluster 3 spent too much on all products. If customers purchase frequently through the company's website, it means that they have a high total spending. Thus, the consuming ability of these customers will decrease, which means that the probability of these customers to accept the offer in the last campaign will decrease as well.

Conclusions of the Agglomerative model:

The conclusions of the Agglomerative model are similar to the conclusions of K-Means.

The company should focus on customers who have a strong consuming ability, fewer children (which is the most important feature that influences the consuming ability directly), and

high total spending (for customers with no children, they tend to spend money when they have strong consuming ability).

For some customers whose consuming ability is relatively low because of too many children, less spending means that they have stronger consuming ability to accept the offer in the last campaign.

Online shopping is more likely to lead people to spend too much, and thus will decrease their consuming ability.

(c) Insights generated from K-medoids Clustering

In this model, we also consider categorical variables because the metric we used here is grower distance. The conclusion about numerical variables is the same as the conclusion of previous models.

We have 4 clusters. There are 718 in cluster 0, 455 in cluster 1, 575 in cluster 2, 457 in cluster 3. It is similar to the previous steps where we first found the target cluster.

```
#the proportion of data that response =1 in cluster 0
cluster0_Mnum/718
```

```
0.10167130919220056
```

```
#the proportion of data that response =1 in cluster 1
cluster1_Mnum/455
```

```
0.12967032967032968
```

```
#the proportion of data that response =1 in cluster 2
cluster2_Mnum/575
```

```
0.26956521739130435
```

```
#the proportion of data that response =1 in cluster 0
cluster3_Mnum/457
```

```
0.10065645514223195
```

Figure 56: The proportion of data whose response =1 in every cluster

From this figure, we found that cluster 2 has the highest proportion of data whose response is equal to 1.

We found that the performance of the K-Medoids model is worse than the previous models. However, we can use this model to find out whether categorical variables are important

to the customer personality.

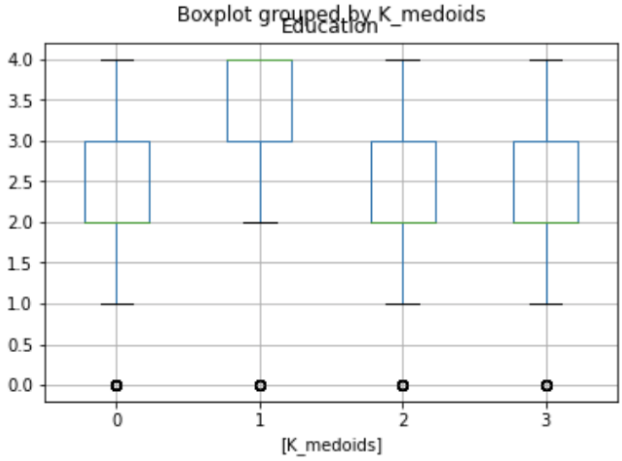


Figure 57: The box plot of Education by K_medoids

Figure 57 shows that education in cluster 2 is not higher than the others. Thus, Education is probably not an important feature.

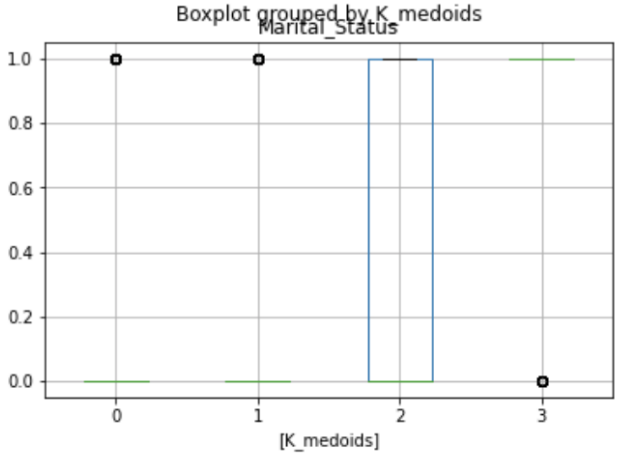


Figure 58: The box plot of Marital_Status by K_medoids

Through this figure, we found that Marital_Status does not influence the response, indicating that “Pair” or “Single” is not important. However, the number of children is an important factor. This is because many customers are married, but they do not have children.

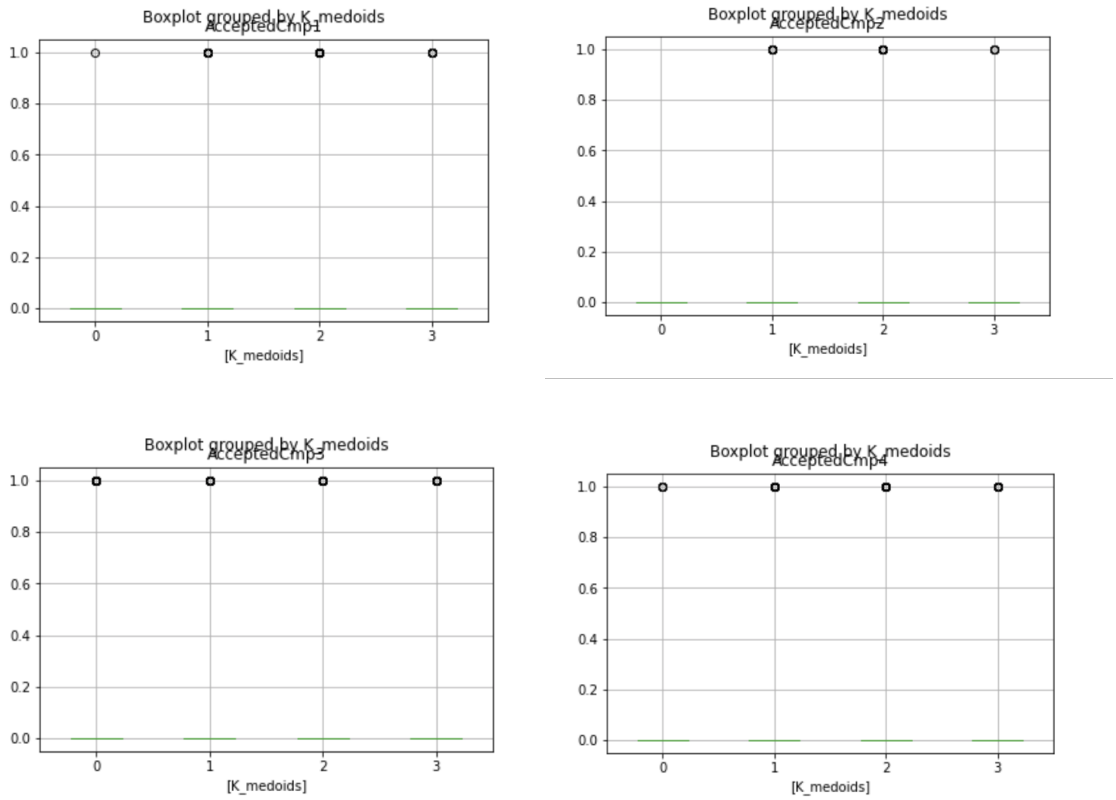


Figure 59-62: The boxplot of AcceptCmp1 2 3 4by K_medoids

From the above four figures, we found that there is no significant difference between cluster 2 and the others in “AcceptedcCmp 1 2 3 4”, which means that whether the customer accepts the offer in the first 4 campaigns will not influence the response.

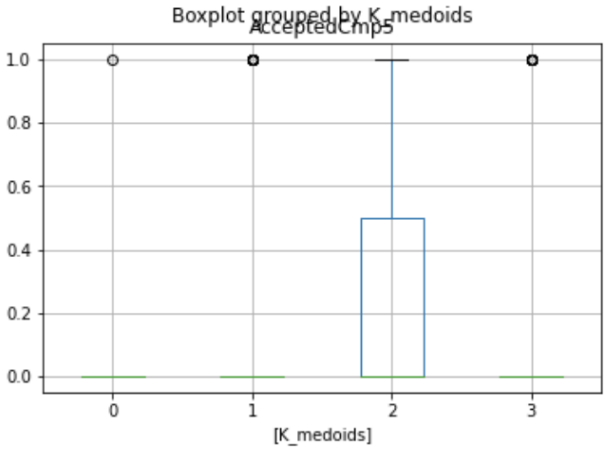


Figure 63: The boxplot of AcceptCmp5 by K_medoids

The only categorical feature that has obvious influence is AcceptedCmp5. Therefore, if customers accept the offer in campaign 5, they are more likely to accept the offer in the last campaign.

Conclusions of the K_Medoids model:

The results of numerical variables are similar to the previous results. We focused on the analysis of categorical features in this model, and found that Marital_Status is not an important feature because marital status cannot directly determine the number of children. The only categorical feature that has significant influence is AcceptedCmp5. Therefore, the company should focus on the customers who accepted the offer in the 5th campaign, which can indicate the customer's willingness to spend.

6. Conclusions

From the above analysis, we conclude that the company should consider customers who have strong consuming ability and are more willing to consume. This is because customers' consuming ability is the most significant factor, followed by their willingness to consume.

We can focus on the features "Num_Children", "Spending", "NumWebPurchase" and "AcceptedCmp5". Customers with fewer children have stronger consuming ability. In addition, high Spending indicates that these customers have more willingness to consume. Besides, high NumWebPurchase means potential waste in life, leading to a decline in consuming ability. AcceptedCmp5 = 1 indicates that customers just accepted the offer in the fifth campaign (which is the most recent campaign), thus they are likely to accept the offer again in a new campaign.

Our best target is people who have strong consuming ability due to fewer children. For these customers, high spending which means they are more willing to consume is an important feature we should consider. However, for customers who have weak consuming ability because they have too many children, we should first consider whether they have the ability to accept our offer. Thus, for those customers, lower spending means they are more likely to have the ability to accept the offer. If we are sure the customers have the ability to accept the offer, we will prefer to see them have more willness to consume.

In conclusion, consuming ability is the most important factor, which can be indicated by the feature "Num_Children", in whether a customer accepts our offer in the last campaign. If the ability of customers is strong enough, we will focus on their willingness to consume, which can be indicated by the features "Spending", "NumWebPurchase" and "AcceptedCmp5".

7. References

Patel, A. (2021, August). Customer Personality Analysis. Retrieved from:

https://www.kaggle.com/imakash3011/customer-personality-analysis?select=marketing_campaign.csv.