MARKETING DATA ANALYSIS

ABSTRACT

This project includes making customer segmentation, building segment profiles, evaluating previous marketing campaigns, and implementing machine learning models to make predictions and handle customer complaint issues. Finally, this project leveraged the business insights from the dataset to derive business growth by making suggestions and creating a data dashboard for decision making

xinwen yao

Table of Contents

Introduce 3

Data preprocessing 3

Data evaluation 3

Data structure 3

Details 4

Data wrangling 4

Handling missing value 4

Handling outliers 4

Data transformation 5

Data restructuring 6

Data mining/ Exploratory data analysis 7

Correlation analysis 7

Multivariate statistical analysis 8

Data normalization 8

Compute the percentage of information represented by different components 9

Score analysis: cluster analysis 10

Multivariate outlier detection 10

Time-based trend analysis 11

Loading analysis 11

Correlated feature analysis 13

Customer segmentation and profile 13

Implement the cluster algorithm 14

Determine the number of clusters 14

Clustering and merging the clusters to dataset 14

Visualize the distribution and population of different segments 15

Purchasing pattern analysis 16

Marketing campaign response 17

Deals purchased 18

Product preference analysis 18

Basic information analysis 19

Customer segments dashboard 20
Complaint prediction 20
Handle imbalance data 20
Model evaluation 21
Conclusion 21

Introduce

The market segment refers to a group of people who share one or more similar characteristics. It is a practice of dividing the market into approachable groups according to several criteria, which is the cornerstone of the subsequent marketing activities.

By developing a comprehensive knowledge of the market segment, the company can leverage this to work out its product, sales, and marketing strategies. Great market segmentation can convey strong marketing messages regarding customers' characteristics. In this day and age, digital advertising is a crucial tool to market products and services. However, the digital platforms do not commit to high response rates and low acquisition costs. With clear segmentation, marketing teams can direct their marketing efforts, such as distributing traffics, to specific ages, genders, locations, buying habits, incomes, and so on, and then attract the right customers.

In addition, the high complaint rate will have a negative impact on the customer satisfaction and then reduce the retention rate. Therefore, I implemented a prediction system based on machine learning and deep learning to improve our customer service and improve the way we serve customers.

Data preprocessing

Data evaluation

Data structure

1. Personal information: ID, Year_birth, Income, Kidhome, Teenhome, Recency, Complain
2. Marketing campaign: AcceptedCmp1, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5, response
3. Consumer goods: MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds, NumDealsPurchases
4. Purchasing habit: NumDealsPurchases, NumWebPurchases, NumCatalogPurchases, NumStorePurchases, NumWebVisitsMonth

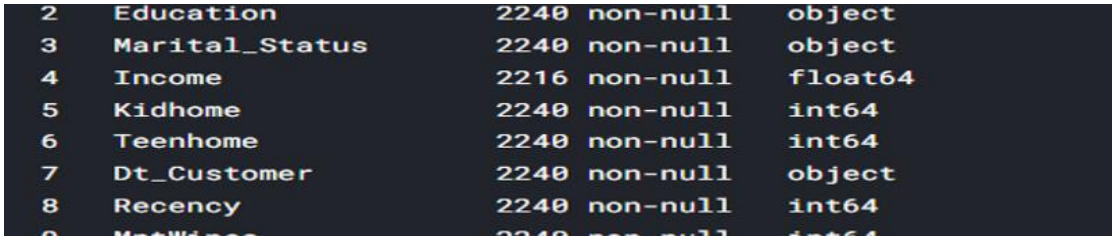
Details

1. There are 2,240 observations with 29 features each.
2. There are missing values in the income column.
3. The dataset has a mix of data types, including numerical data, categorical data and time data.
4. ID is the primary key of this table, we don't have duplicated values
5. Values in Z_CostContact, Z_Revenue are identical.
6. There are synonyms in the text values.
7. There are extreme values in the numerical features.

Data wrangling

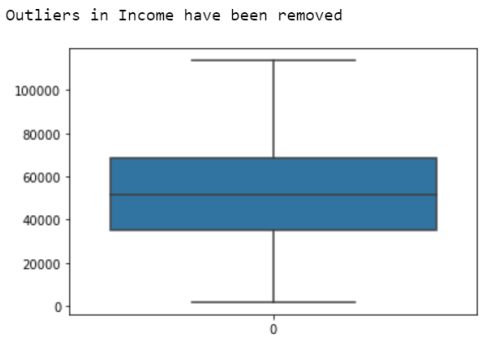
Handling missing value

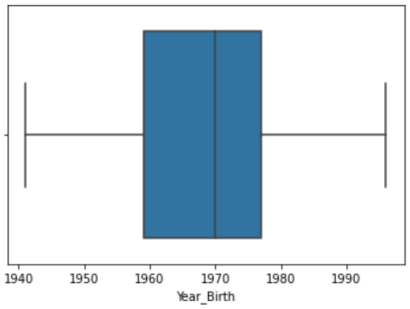
In this case, there are 2,240 observations with 29 features each. Figure 1 shows the overview of this dataset, and it indicates that there are only 2,216 observations in the income column. Considering that the number of missing values is relatively small, the observations with missing values are deleted in this case.



Handling outliers

The outlier can cause serious problems in statistical analysis and have a negative effect on the subsequent machine learning model. The statistical table shows that the max income is significantly larger than the average value and the min value of year_birth does not consistent with common sense. Assuming that the features obey a normal distribution, those observations lie outside three deviations away from mean will be remove.
Result:





Data transformation

Data transformation is the process of changing the format, structure, or values of data.

Date value: The Dt_customer is a datetime value while it was parsed as text in this dataset. To make it align with other features, I took the value relative to the most recent record and transformed it to numerical values.

	Dt_Customer	Days
0	04-09-2012	92
1	08-03-2014	938
2	21-08-2013	591
3	10-02-2014	998
4	19-01-2014	742

Categorical value: Categorical variables in the dataset are ‘Education', 'Marital_Status'.

Education feature:

Explore unique values in the Education feature:

```
The unique values in education is:
 Graduation    1113
PhD            475
Master         364
2n Cycle       198
Basic           54
Name: Education, dtype: int64
```

Standardize the name of different educational level:

```
The unique values in education:
 Graduate         1113
Postgraduate      839
Undergraduate     252
Name: Education, dtype: int64
```

Encode Education feature:

```
Education feature after tranformation
 [0 0 0 ... 0 1 1]
```

Marital_Status:

Explore unique values in the Marital_Status:

```
The unique values in marital_status is:
 Married     854
Together    568
Single      469
Divorced    230
Widow        76
Alone         3
Absurd        2
YOLO          2
Name: Marital_Status, dtype: int64
```

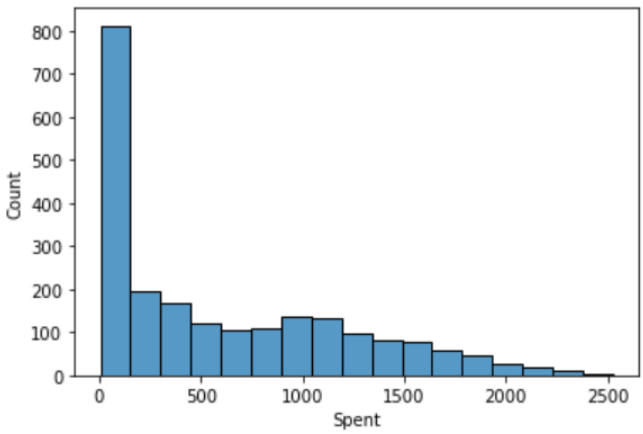
Standardize the name of different marital status and encode them:

```
The uniqual value
 Married     854
Together    568
Single      476
Divorced    306
Name: Marital_Status, dtype: int64
Marital feature after tranformation
 [2 2 3 ... 0 3 1]
mapping rule {'Divorced': 0, 'Married': 1, 'Single': 2, 'Together': 3}
```

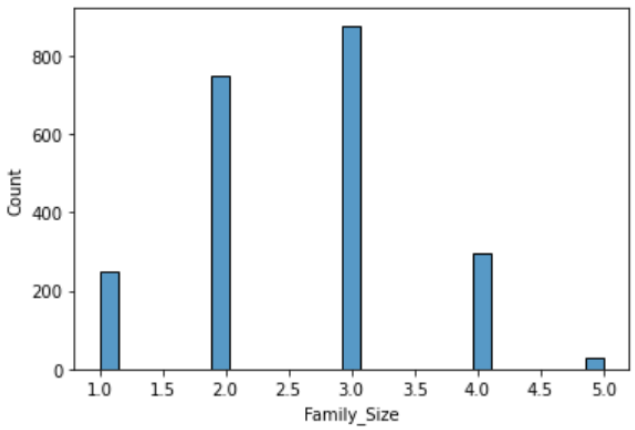
Data restructuring

New features can be created by merging and extracting from the raw dataset.

Considering that the spendings on different products is a good indicator of customer purchasing power, I computed the total spendings on various items.



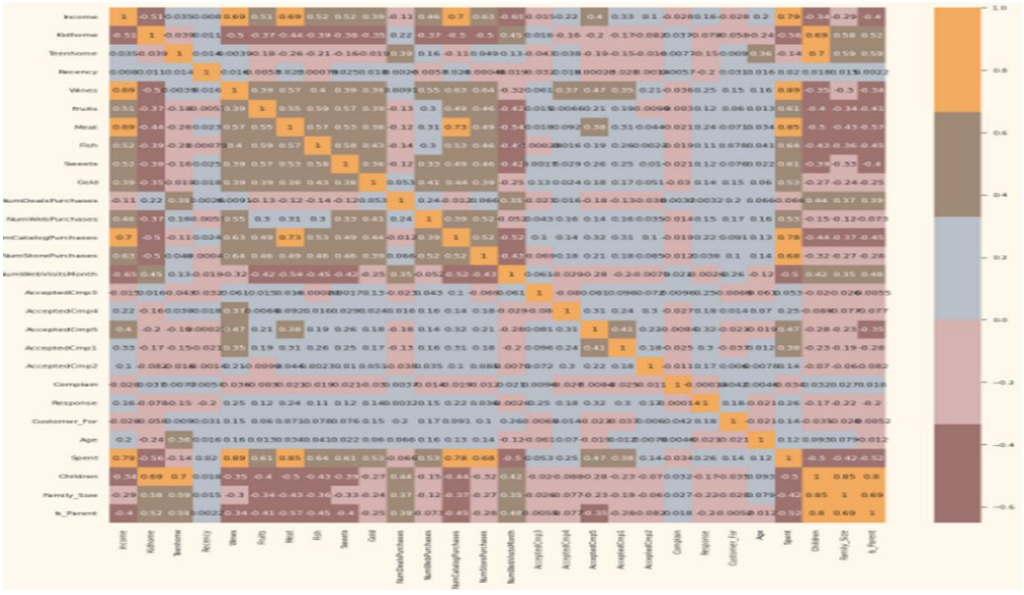
The family size can be extracted according to 'Kidhome', 'Teenhome' and 'Marital'.



Data mining/ Exploratory data analysis

Correlation analysis

High correlated data may degrade the analyzing model's accuracy and decrease the computation speed as well.



Some features are high correlated with each other, especially the amount spent on different product. So, it is necessary to reduce the number of features but maintain the information in the dataset.

Multivariate statistical analysis

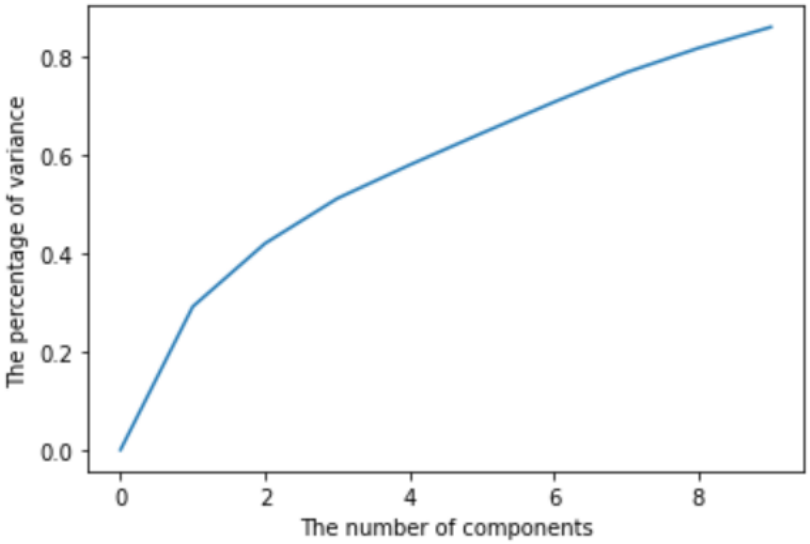
The principal component analysis is a data processing technique that is used in exploratory data analysis. It is commonly used for dimensionality reduction by projecting each data point onto only the first few principal components to obtain lower-dimensional data while preserving as much of the data's variation as possible. This inherent ability to reduce dimension is one of the attributes that make PCA an attractive method for data preprocessing. However, it is necessary to find out the best hyperparameter before applying PCA. In the case, I will compute the percentage of variance explained each of the selected components to determine the best components.

Data normalization

The features have different units, and those variables may not contribute equally to the analysis and might end up creating a bias. So, it is important to center and scale the data
Centers and scales a variable to mean 0 and standard deviation

	Education	Income	Kidhome	Teenhome	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProd
count	2.204000e+03	2.204000e+03	2.204000e+03	2.204000e+03	2.204000e+03	2.204000e+03	2.204000e+03	2.204000e+03	2.204000e+03	2.204000e
mean	-5.399996e-17	-9.335709e-17	1.570633e-16	5.850331e-16	1.164122e-16	2.014924e-17	-7.525740e-17	-5.150649e-17	-5.672011e-17	-1.536379
std	1.000227e+00	1.000227e+00	1.000227e+00	1.000227e+00	1.000227e+00	1.000227e+00	1.000227e+00	1.000227e+00	1.000227e+00	1.000227e
min	-8.919427e-01	-2.408737e+00	-8.237186e-01	-9.311609e-01	-1.694211e+00	-9.074366e-01	-6.640226e-01	-7.594833e-01	-6.890636e-01	-6.599325
25%	-8.919427e-01	-7.932586e-01	-8.237186e-01	-9.311609e-01	-8.641859e-01	-8.363204e-01	-6.137467e-01	-6.860073e-01	-6.343378e-01	-6.356173
50%	-8.919427e-01	-1.551258e-02	-8.237186e-01	-9.311609e-01	4.236744e-04	-3.799919e-01	-4.629190e-01	-4.472105e-01	-4.701603e-01	-4.654103
75%	5.718262e-01	8.046825e-01	1.038308e+00	9.061521e-01	8.650333e-01	5.948918e-01	1.655295e-01	3.059177e-01	2.230334e-01	1.667869
max	2.035595e+00	2.998672e+00	2.900334e+00	2.743465e+00	1.729643e+00	3.516580e+00	4.338428e+00	7.162140e+00	4.035599e+00	5.710670e

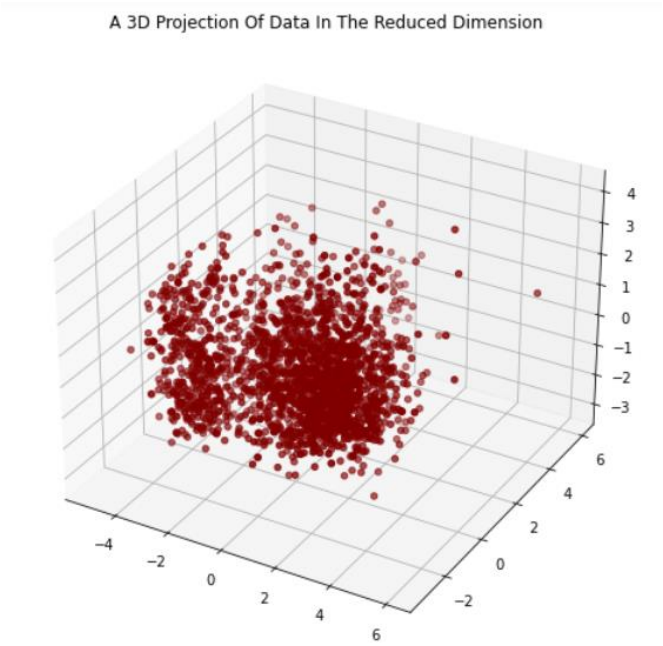
Compute the percentage of information represented by different components.



Reduce the number of dimensions from 29 to 4.

```
       col1      col2      col3      col4
0 -2.960244  0.371266  3.415092  1.099200
1  1.999957 -0.603824 -1.615073 -1.088615
2 -1.959395 -0.351803 -0.325593 -0.009587
3  2.216569 -1.351923 -0.893777 -0.754990
4  0.423776  0.444305 -0.586775  0.840523
The dataset has been reduced to four dimensions
```

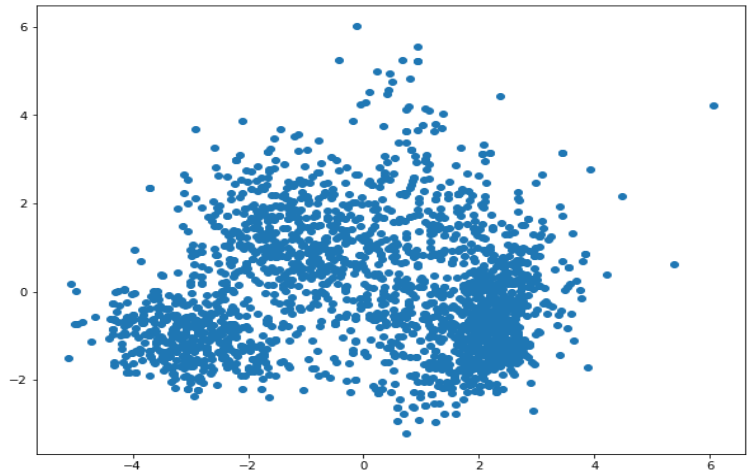
Score analysis: cluster analysis



There is a large number of customers in common, which should be our target customers. And the score distribution also indicated that this dataset is separable. However, there is still a few numbers of data point away from the center, which should be investigate deeper to determine whether or not they should be deleted.

Multivariate outlier detection

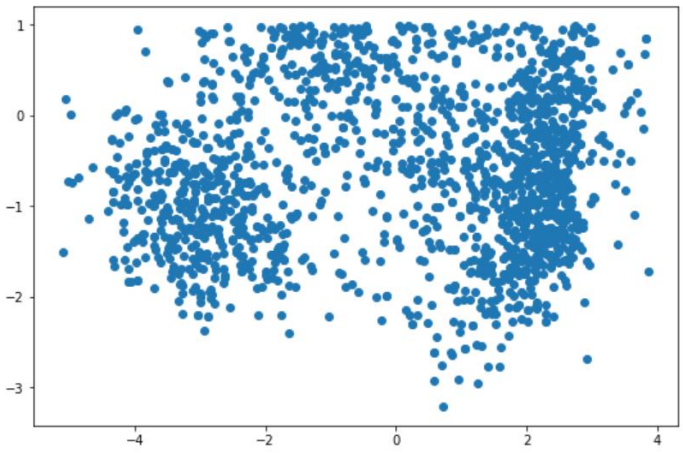
Considering that the first two directions represent the greatest variation in the data, it is more intuitive to investigate the multivariate outliers in two dimensions.



After analyzing the data point that far way from the origin, I found that those customers have an above average number of children, an above average family size, a below average spent and so on. It shows that those customers are supposed to spend more because they have a large family size. But it turns out that they even spent less. Therefore, I consider those data

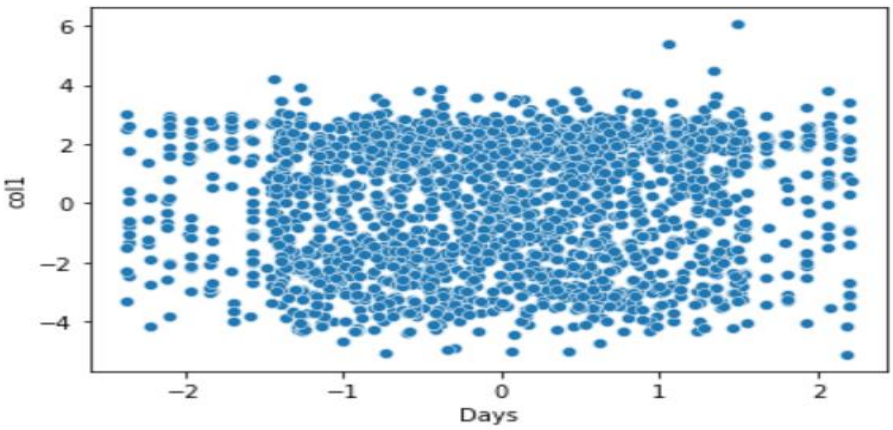
points are outliers because it is not a reasonable situation.

Data after cleaning:



Time-based trend analysis

The buying pattern and customers profile may change with time, so I want to see if there any time-based pattern in customers. The strong and consistent time-based trends in the raw data will be reflected in the scores.

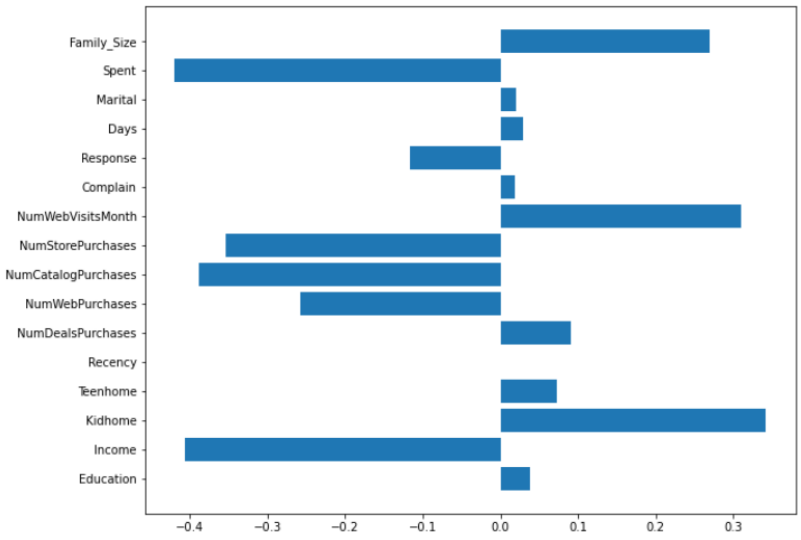


Those samples appear to be evenly spread, which means customers' profile doesn't change with time. However, there are a few datapoint have the relatively high score1 and they fall within the same period of time.

Loading analysis

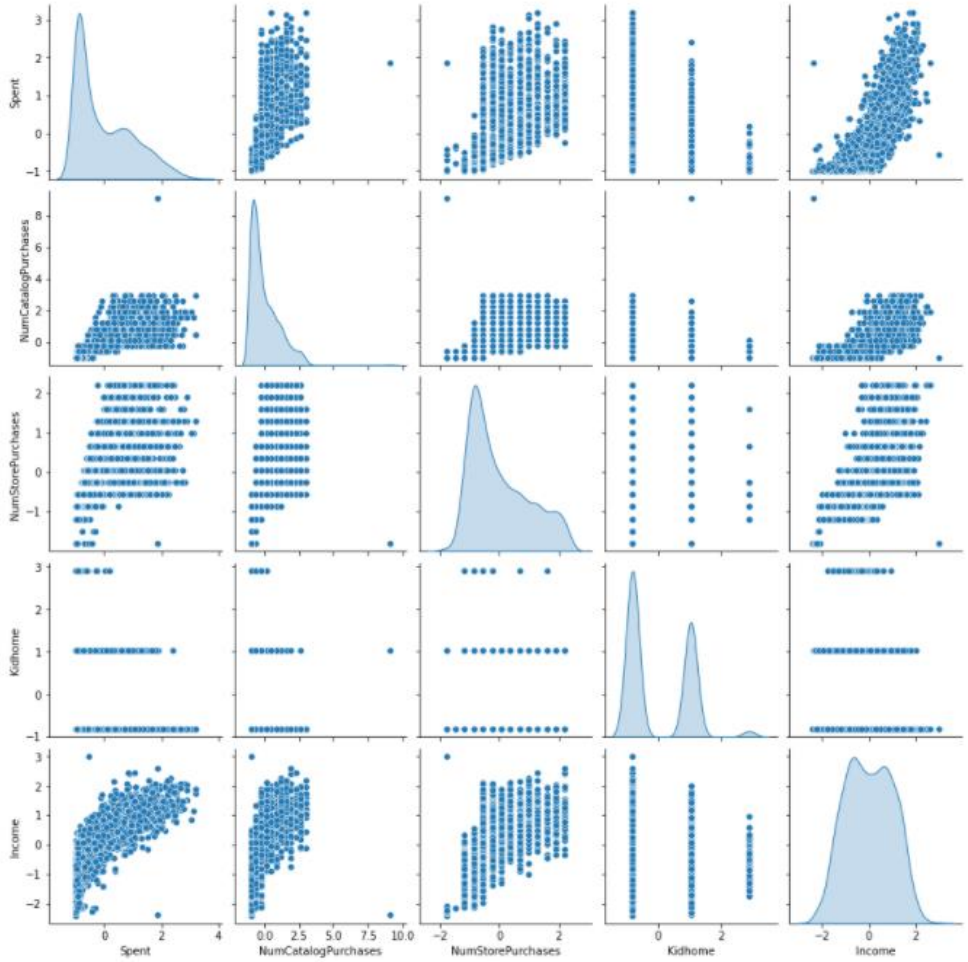
In order to divide the customer better, I'd like to investigate which features cause the

difference among customers. One way to locate unimportant variables in the model is by finding which variables which have small weights in all components. And then, these variables can generally be removed, as they show no correlation to any of the components or with other variables. In another word, those variables have little importance or relevance in understanding the total variability in the system. Therefore, I'd like to investigate those features have high weights, which may cause in variance.



Strongly correlated variables, will have approximately the same weight value when they are positively correlated. While negatively correlated variables will appear diagonally opposite each other. As shown above:'Spent','NumCatalogPurchases', 'NumStorePurchases', 'Kidhome', and 'Income' have strong relationship with each other, either positive or negative. They are strong indicators to differentiate a customer from the other.

Correlated feature analysis



1. There is a strong positive correlation between income and spent.
2. People with high income purchase more kind of consumer products.
3. High income customers are more like purchasing I the store.
4. High income people have less children at home.
5. The more children at home, the less spent will be. Because people will reduce their consumption on our product and save money for children's products.

Customer segmentation and profile

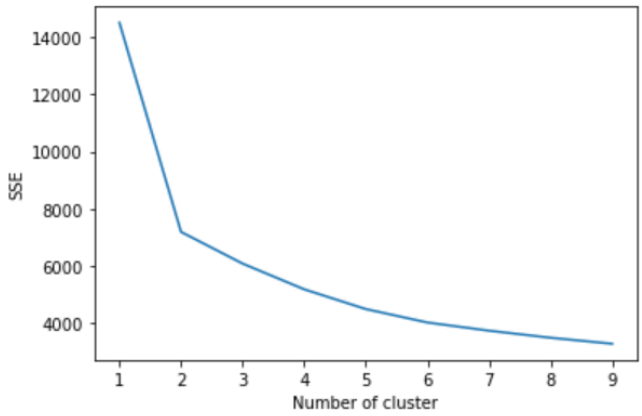
After running the PCA, the observations that are similar will fall close to each other. In addition, the multivariate outliers were removed to ensure the consistency of the data. I'm going to run

the K-means model to segment the customers automatically.

Implement the cluster algorithm

Determine the number of clusters

The hyperparameter of K-means is the number of clusters. In order to determine this parameter, both business objective and statistical indicators should be taken into consideration.
Compute the sum of distances of samples to their closest cluster center:

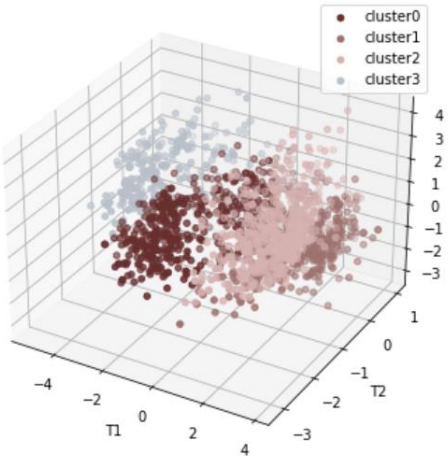


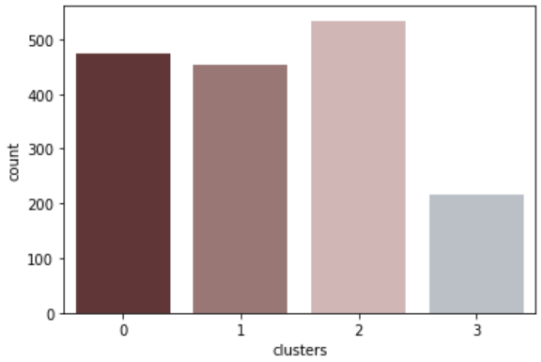
The elbow methodology indicates that the best number of clusters is two. I want to make slightly modification to make each customer segment smaller and the profile for each customer more precise.

Clustering and merging the clusters to dataset

	col1	col2	col3	col4	clusters
0	-2.960244	0.371219	3.414960	1.099925	3
1	1.999956	-0.603942	-1.615439	-1.087780	2
2	-1.959395	-0.351829	-0.325701	-0.009573	1
3	2.216570	-1.351857	-0.893556	-0.755276	2
4	0.423775	0.444275	-0.586913	0.840587	2

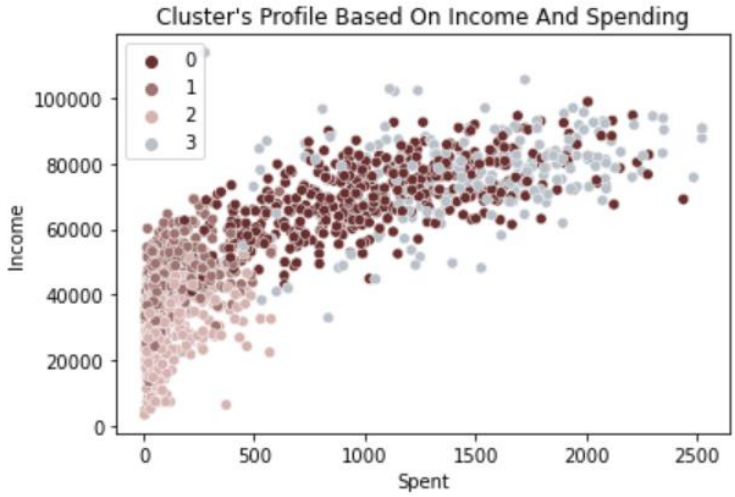
Visualize the distribution and population of different segments

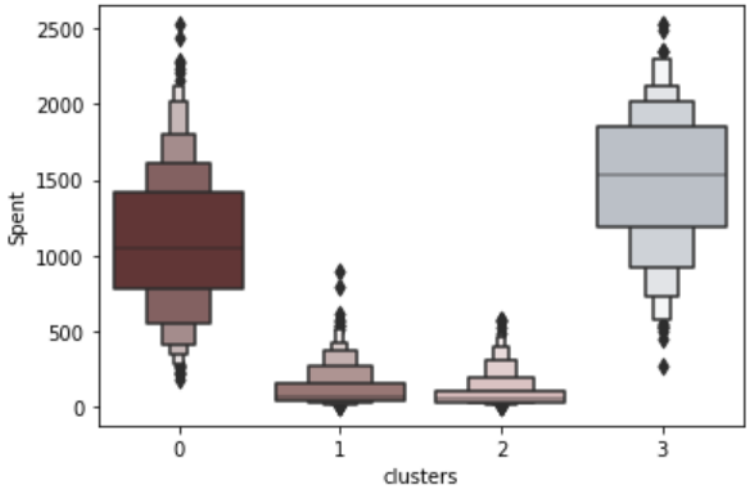




The clusters seem to be reasonably distributed. Cluster2 has the largest number of customers while cluster3 has the smallest number of customers.

Purchasing pattern analysis



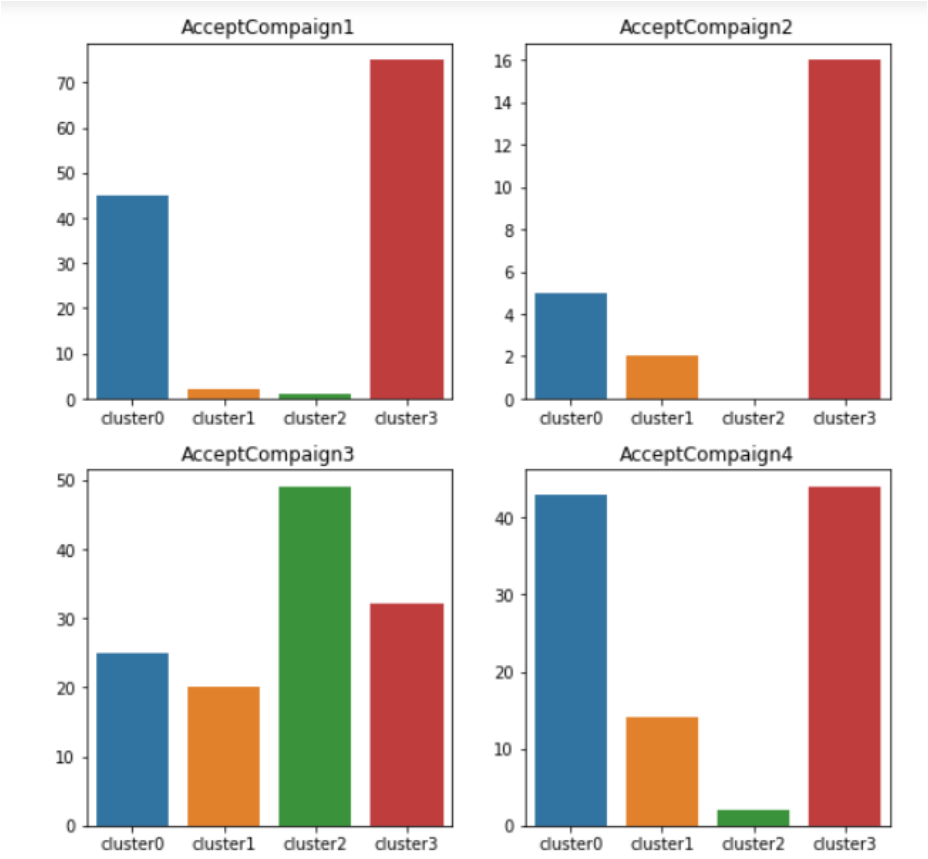


Result:

	Population	Income	Purchasing power
Cluster0	Medium	High	High
Cluster1	Medium	Medium	Medium
Cluster2	Most	Low	low
Cluster3	Least	High	High

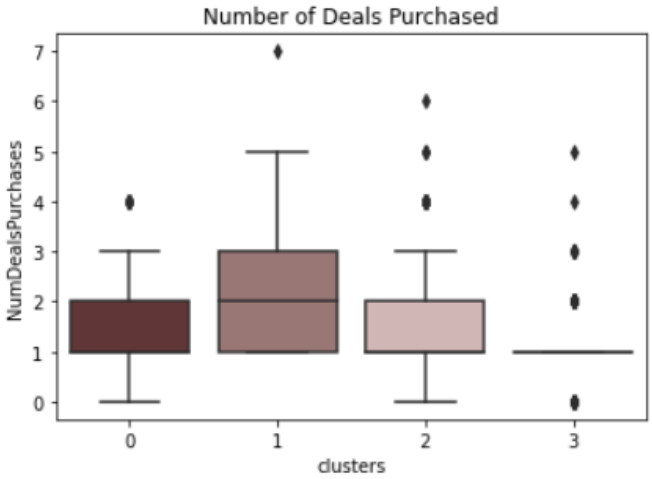
Marketing campaign response

I'd like to investigate how each segment response to different marketing campaigns.



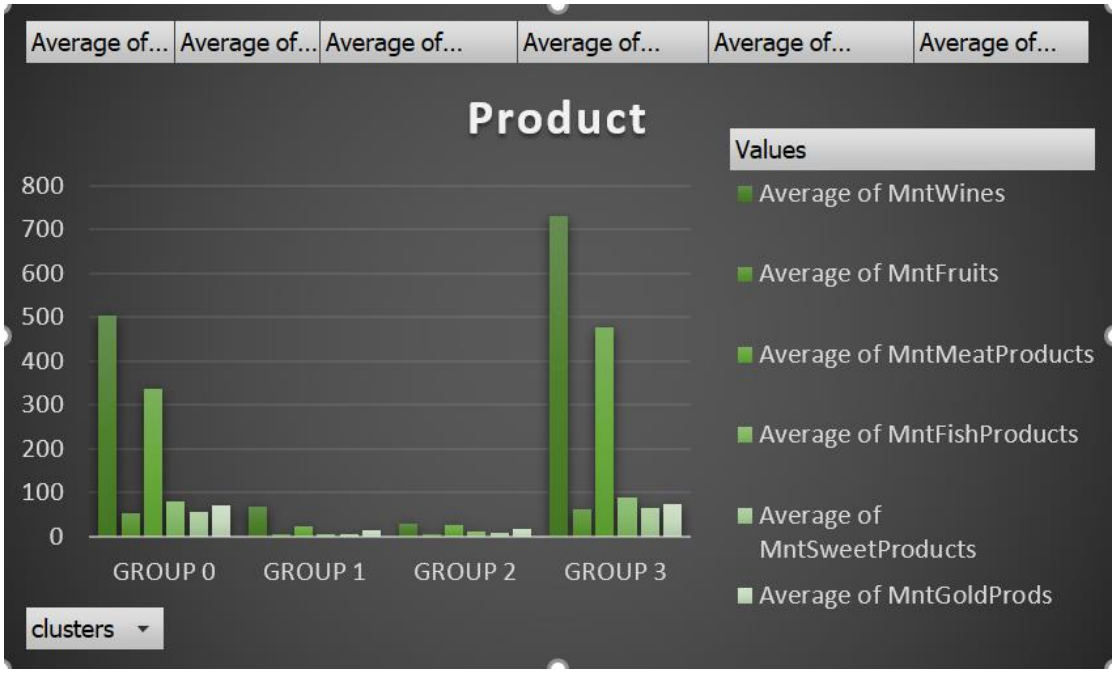
They show that customers in cluster3 are actively participate in marketing activities. Customers in cluster2 are less interested in marketing campaign while they show a strong interest on marketing campaign3. Except campaign3, there are no patterns show that a specific cluster prefer a specific campaign. Perhaps better-targeted and well-planned campaigns are required to boost sales.

Deals purchased



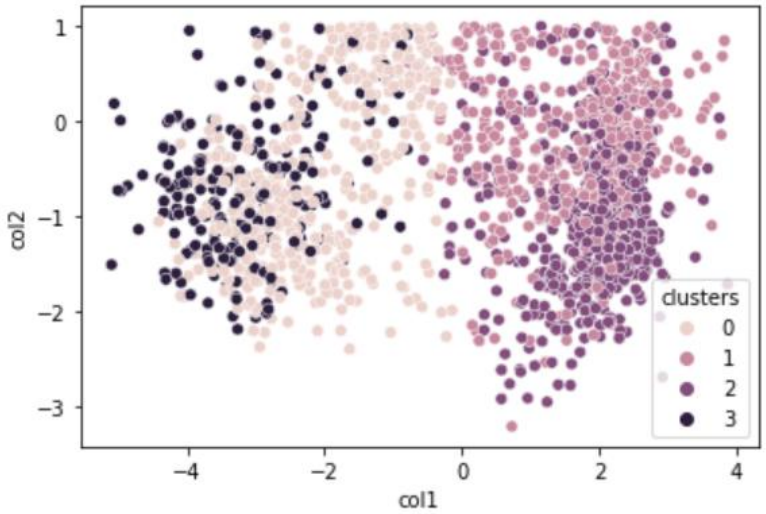
It seems the offers distributed well. Customers accepted one to two offers on average and we saw a small number of people love to look for products on sales.

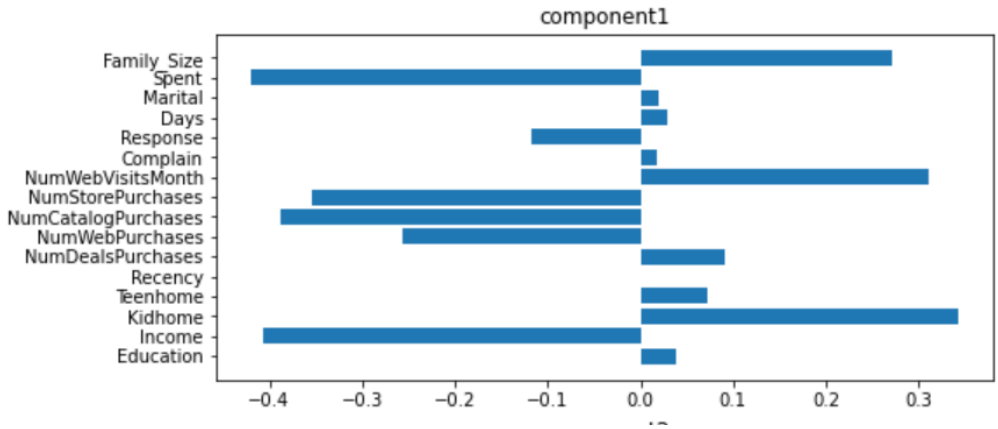
Product preference analysis

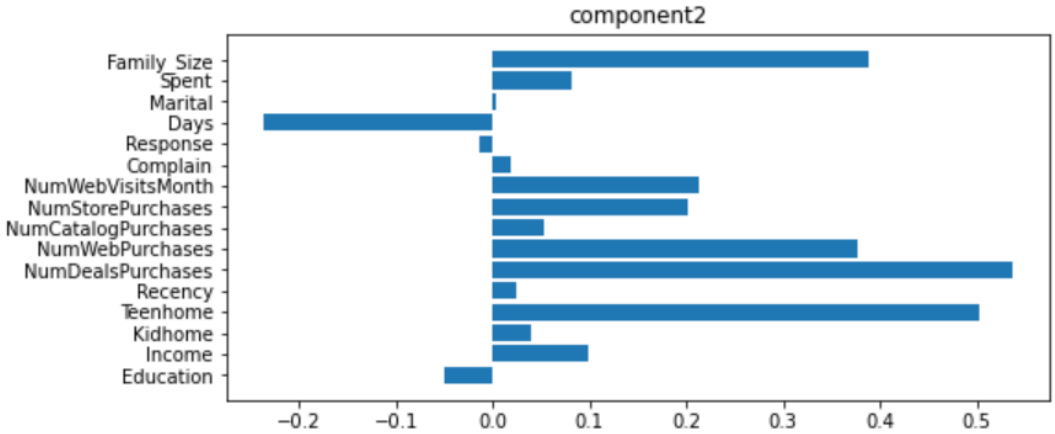


The result aligns with the previous analysis. Customers in cluster0 and cluster3 have strong purchasing power and they spent a lot on wines and meat products.

Basic information analysis







Cluster1 & Cluster2:

They have an above average family size and an above average number of children at home.

However, their purchasing power and income are relatively low. They used to browse the website but didn't purchase frequently. Customers in cluster0 are similar to those in cluster3, but customers in cluster0 enrolled in this system much earlier than those in cluster3

Cluster0 & Cluster3:

Their family size is relatively small and they have a below average number of children at home. They are high income group with high purchasing power. They shop frequently and don't spend much time on website. Compared with cluster1, cluster2 is more extreme.

Customer segments dashboard

Complaint prediction

Handle imbalance data

This is a very skewed dataset with only 20 positive samples and 2184 negative samples. Therefore, it is necessary to enrich the dataset or machine learning model will tend to predict every observation with negative result. In this case, I perform random over sampling make the dataset balanced.

Model evaluation

Model	Accuracy	Confusion matric	Precision	Recall	F1 score
Logistic Regression	0.69	252 182 93 347	0.58	0.73	0.685
Support Vector Machine	0.99	427 7 0 440	0.98	1	0.991
Decision Tree	0.99	431 3 0 440	0.99	1	0.996

Conclusion

The chart above indicates that decision tree achieves the best perform in this case, with the highest accuracy and the highest F1 score. By using this this prediction model, we can know who will make complaint and reach them out in advance.