Consumer Personality Analysis

Group number – 6

Mengke Jiang
Tanisha Arora
Akash Das
Siyu Chen
Chi-Han-Chi

Table of Contents

Abstract

Any organization, whether small or big, relies on its customers for growth. Customers have different needs and wants. Rather than treating all customers using the same marketing strategy, dividing the entire customer base into smaller, more manageable segments for better brand planning and targeted marketing is a more effective and cost-saving way to optimize its resources and create bigger values.

In this project, we attempt to segment our customers based on certain criteria such as Income, Marriage Status, and Kids. We will use two clustering techniques, K-Mean and K-Prototypes to cluster these groups of customers into smaller segments with similar features. And further using association Rule to answer which products leads to buying of other product.

1. Introduction

A) Motivation

Customers are the make-or-break factor for an organization, without them the business will cease to exist. Apart from all the modernization that has happened in the last few decades, one thing that has remained constant is the consumers acting as the driving forces of a company, and they have becoming more and more important as many forward-thinking companies has adopted a customer-centric value. As a result, using scientific techniques to understand customers' needs and wants is key to design marketing strategies and product recommendations.

B) Business Implications

Customer personality analysis is a way to reach out to the ideal customers of the organization. It affects the organization in two ways:

a. Reduced cost and increased revenue as business focuses on customers based on their segments which reduces the trial marketing expenses.

b. Increased customer loyalty and brand value as customers in one segment are offered the exact same things they want which increase their positive shopping experience and satisfaction level.

The nature of the customer segments is analyzed by the composition of different personality traits that make each segment. Thus, the personality of the customers in each customer segment helps the organization to formulate differentiated strategies to advertise products to only that segment who are the most likely to purchase them. Thus, making customer segments based on personalities allows the marketing and advertising team and to find customers that have the same traits as the respective segments in terms of different criteria such as Income, education etc. and targeting these customers will garner additional revenue than general customers. The goal of the segmentation is to foresee the needs of customers, get to know their interests, lifestyles, priorities, and spending habits to maximize the value of customers to businesses.

2. Data

A) Data Acquisition:

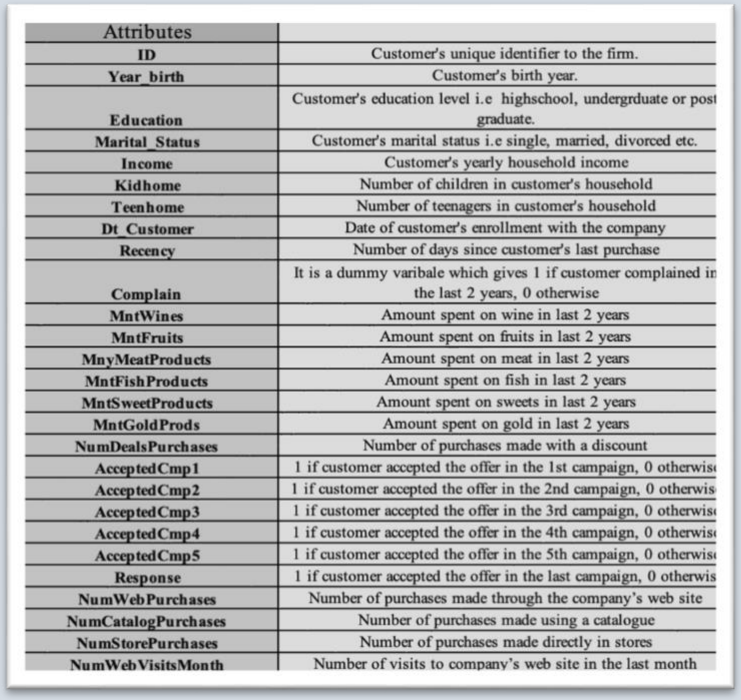
Data was collected from an open-source data competition website called Kaggle. The dataset contains 2240 rows and 29 columns and includes customers demographic characteristics and shopping information from a grocery firm from 2012-2014. Here is the dataset’s link: https://www.kaggle.com/imakash3011/customer-personality-analysis.

B) The Original Dataset’s Variables Description:

The dataset contains 29 attributes which can be classified into 4 groups:

1. Customer’s demographic features. (Birth year, education, marital status, etc.)
2. Buying product information. i.e., customer’s expenses on different categories of products. (Wine, fruit, meat, fish, etc.)
3. Promotion information i.e., the response of customers to different campaigns (Campaign 1 to 5, response)
4. Sales channel. i.e., where customers purchased products from (website, catalogue, in-store)

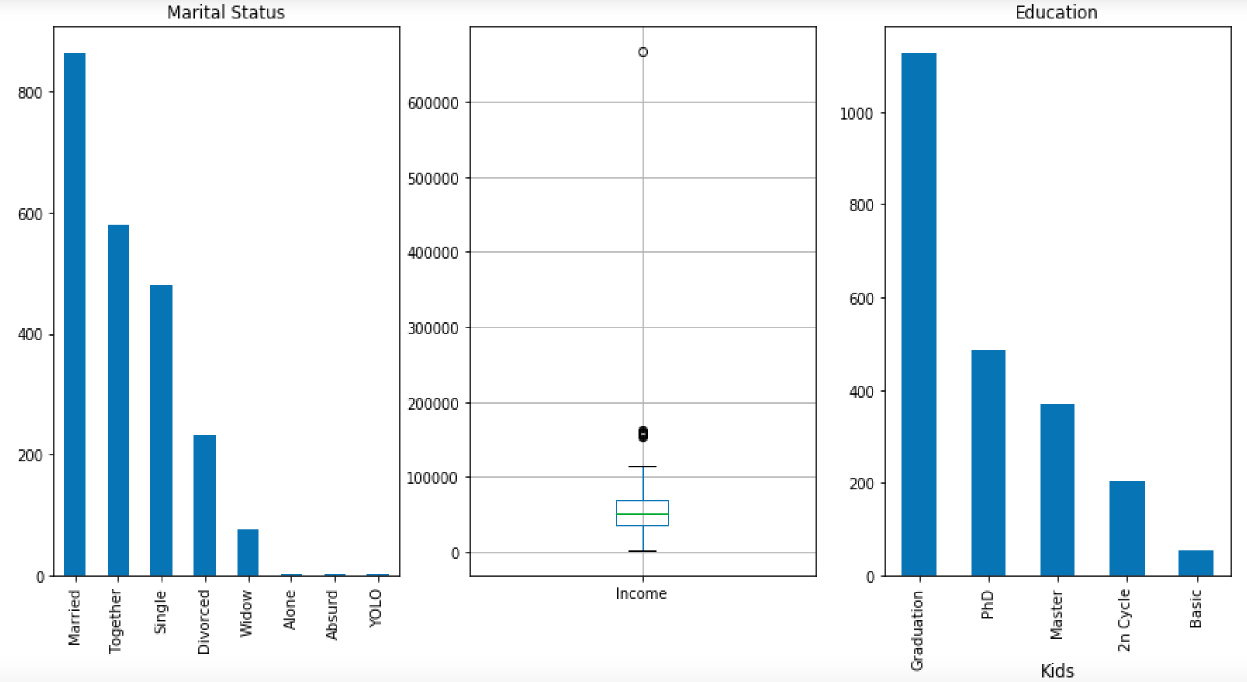
The detailed descriptions of all attributes are as follows:

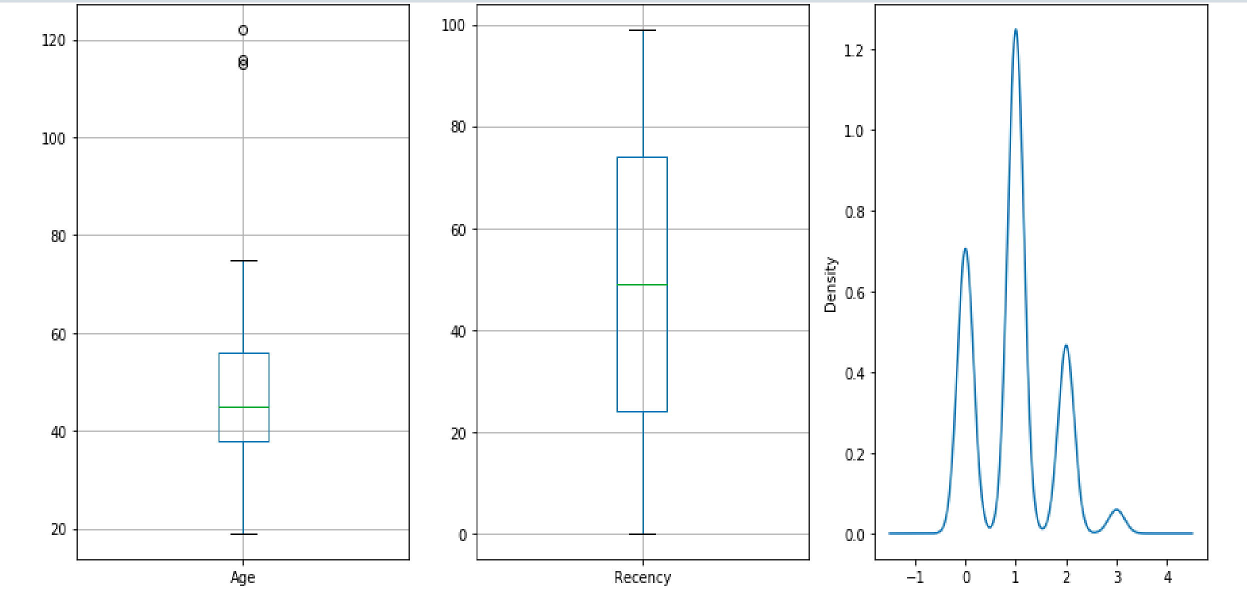


Attributes	
ID	Customer's unique identifier to the firm.
Year_birth	Customer's birth year.
Education	Customer's education level i.e highschool, undergrduate or post graduate.
Marital_Status	Customer's marital status i.e single, married, divorced etc.
Income	Customer's yearly household income
Kidhome	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	It is a dummy varibale which gives 1 if customer complained in the last 2 years, 0 otherwise
MntWines	Amount spent on wine in last 2 years
MntFruits	Amount spent on fruits in last 2 years
MnyMeatProducts	Amount spent on meat in last 2 years
MntFishProducts	Amount spent on fish in last 2 years
MntSweetProducts	Amount spent on sweets in last 2 years
MntGoldProds	Amount spent on gold in last 2 years
NumDealsPurchases	Number of purchases made with a discount
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 otherwis
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
Response	1 if customer accepted the offer in the last campaign, 0 otherwis
NumWebPurchases	Number of purchases made through the company’s web site
NumCatalogPurchases	Number of purchases made using a catalogue
NumStorePurchases	Number of purchases made directly in stores
NumWebVisitsMonth	Number of visits to company’s web site in the last month

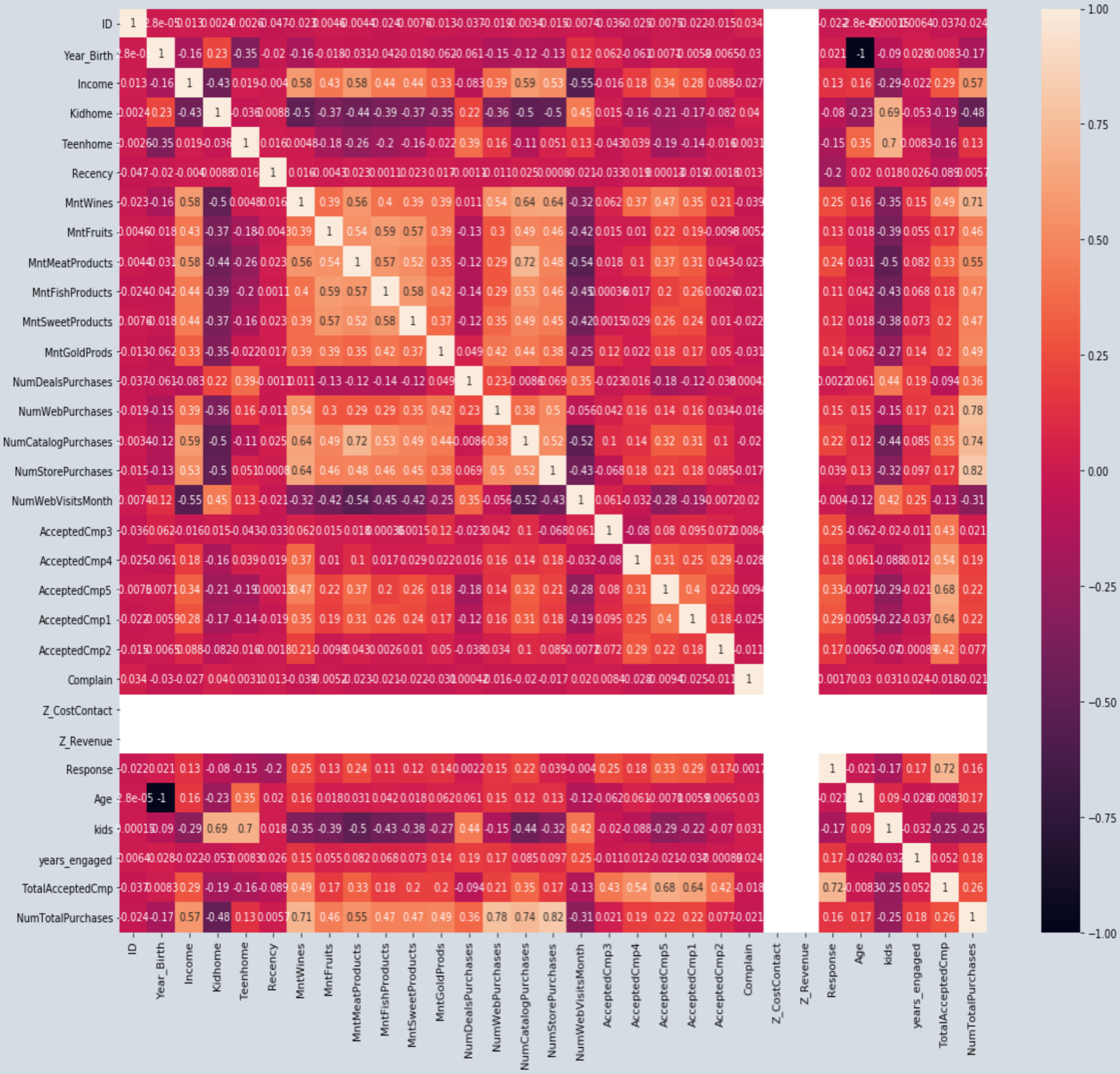
C) Data Visualization

Before preprocessing the data and running our unsupervised learning analysis, we did an exploratory data analysis to visualize customers' main demographic features and product buying patterns.



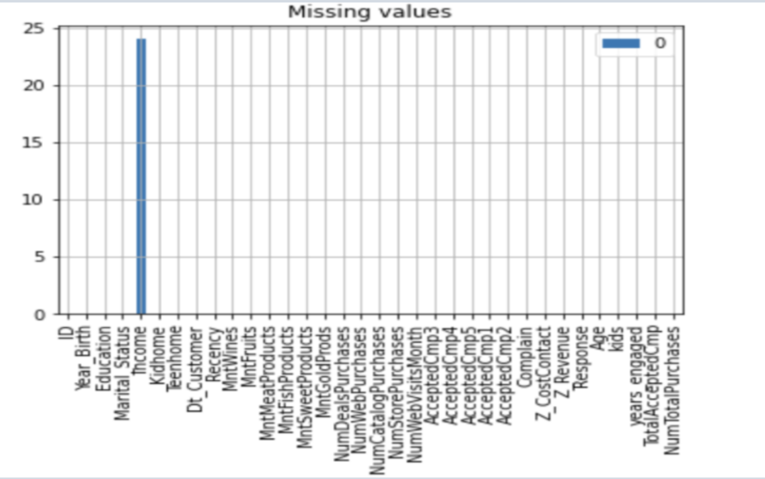


The above box plots tells that there are some outliers in our dataset which should be removed while preprocessing the data in the next step. We observe that customers' Income Level, Age, and Education are diverse. The shopping recency are spreading far away from each other, which implies that customers' diverse demographic characteristics influence their shopping patterns. Running the correlation analysis among all variables tells us that there exist many negligible correlations between some data points. But there are also some unignorable correlations between some variables which should dive deeper into analysis.



To check whether the dataset contains any missing values, we did a bar chart of missing values of all the variables given as follows –

These null values will be removed during data preprocessing.

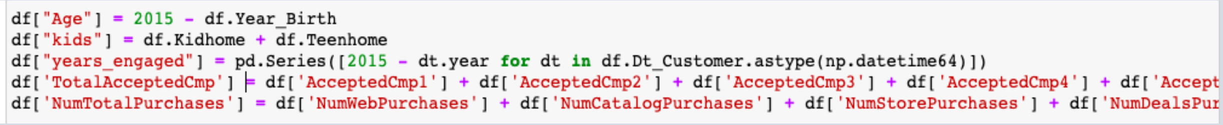


D) Data Cleaning

To enhance the performance of our analysis and to prepare the data to be fitted into our unsupervised model, we pre-processed the data by doing the following –

a. Data cleaning
b. Data transformation
c. Variables dimension reduction.

- First, we created five new variables: “Age”, “Kids”, “years_engaged”, “TotalAcceptedCmp”, “NumTotalPurchases” by combining several variables together to better summarize their demographic features and buying patterns.

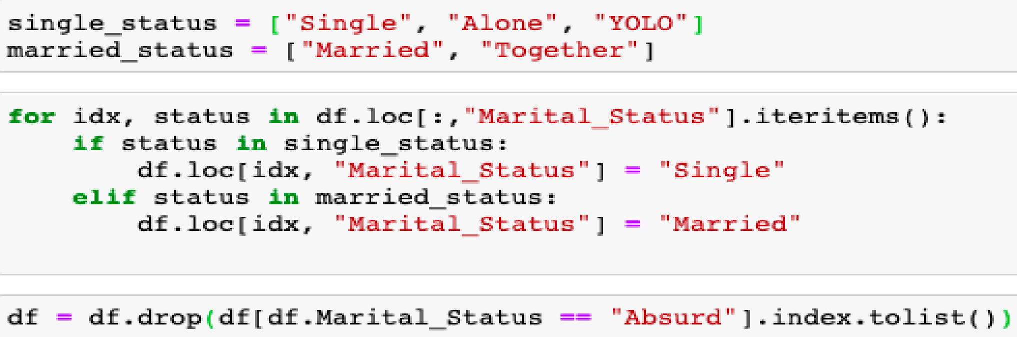


```
df["Age"] = 2015 - df.Year_Birth
df["kids"] = df.Kidhome + df.Teenhome
df["years_engaged"] = pd.Series([2015 - dt.year for dt in df.Dt_Customer.astype(np.datetime64)])
df['TotalAcceptedCmp'] = df['AcceptedCmp1'] + df['AcceptedCmp2'] + df['AcceptedCmp3'] + df['AcceptedCmp4'] + df['Accept
df['NumTotalPurchases'] = df['NumWebPurchases'] + df['NumCatalogPurchases'] + df['NumStorePurchases'] + df['NumDealsPur
```

Age: the age of each customer
Kids: add the number of kid and teen in a home together
Years_engaged: how many years past since customers first enrolled with the company
NumTotalPurchases: total number of purchases from different purchase sources

TotalAcceptedCmp: total number of campaigns customers accepted

- Second, we removed the “Absurd” category from “Marital_Status” variable and combined “Married” and “Together” categories into a new variable: “Married”, and combining “Single”, “Alone”, and “YOLO” into a new variable: “Single”.

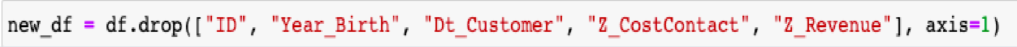


```
single_status = ["Single", "Alone", "YOLO"]
married_status = ["Married", "Together"]
```

```
for idx, status in df.loc[:,"Marital_Status"].iteritems():
    if status in single_status:
        df.loc[idx, "Marital_Status"] = "Single"
    elif status in married_status:
        df.loc[idx, "Marital_Status"] = "Married"
```

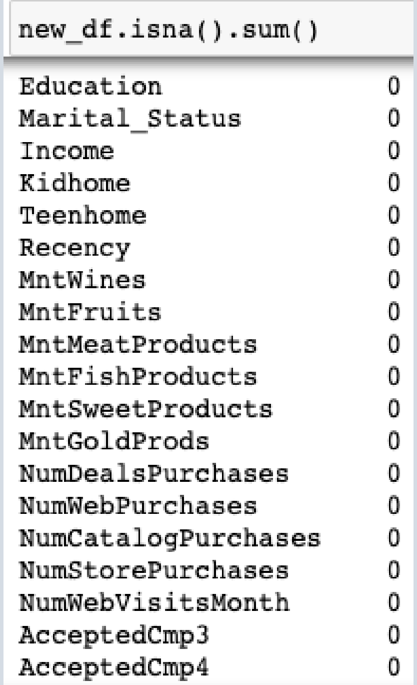
```
df = df.drop(df[df.Marital_Status == "Absurd"].index.tolist())
```

- Third, we dropped the non-adding-value variables: “ID”, “Year_Birth”, “Dt_Customer”, “Z_CostContact”, and “Z_Revenue” from the dataset.



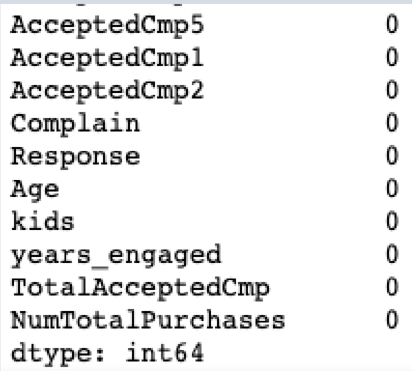
```
new_df = df.drop(["ID", "Year_Birth", "Dt_Customer", "Z_CostContact", "Z_Revenue"], axis=1)
```

- Fourth, we removed all the missing values from the original dataset.



```
new_df.isna().sum()
```

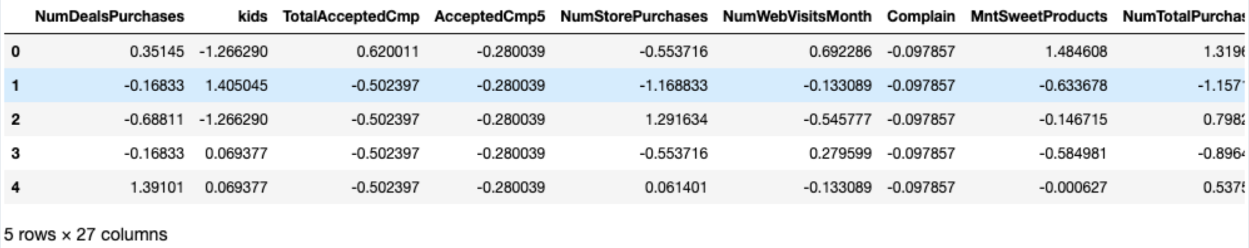
```
Education              0
Marital_Status         0
Income                 0
Kidhome                0
Teenhome               0
Recency                0
MntWines               0
MntFruits              0
MntMeatProducts        0
MntFishProducts        0
MntSweetProducts       0
MntGoldProds           0
NumDealsPurchases      0
NumWebPurchases        0
NumCatalogPurchases    0
NumStorePurchases      0
NumWebVisitsMonth      0
AcceptedCmp3           0
AcceptedCmp4           0
```



```
AcceptedCmp5            0
AcceptedCmp1            0
AcceptedCmp2            0
Complain                0
Response                0
Age                     0
kids                    0
years_engaged           0
TotalAcceptedCmp        0
NumTotalPurchases       0
dtype: int64
```

Because K-Means model works best on continuous variables, we dropped categorical variables “Education” and “Marital_Status” from the dataset.

Some variables such as “Income” have a much higher scaling which would significantly impact the model accuracy, so we did a re-scaling for all continuous variables.
Then we dropped outliers from it when the z-score is greater than 3. Below is part of the new dataset after scaling and removing categorical variables and outliers.



	NumDealsPurchases	kids	TotalAcceptedCmp	AcceptedCmp5	NumStorePurchases	NumWebVisitsMonth	Complain	MntSweetProducts	NumTotalPurchas
0	0.35145	-1.266290	0.620011	-0.280039	-0.553716	0.692286	-0.097857	1.484608	1.319
1	-0.16833	1.405045	-0.502397	-0.280039	-1.168833	-0.133089	-0.097857	-0.633678	-1.157
2	-0.68811	-1.266290	-0.502397	-0.280039	1.291634	-0.545777	-0.097857	-0.146715	0.798
3	-0.16833	0.069377	-0.502397	-0.280039	-0.553716	0.279599	-0.097857	-0.584981	-0.896
4	1.39101	0.069377	-0.502397	-0.280039	0.061401	-0.133089	-0.097857	-0.000627	0.537

5 rows × 27 columns

3. Model Exploration & Model Evaluation & Model Analysis

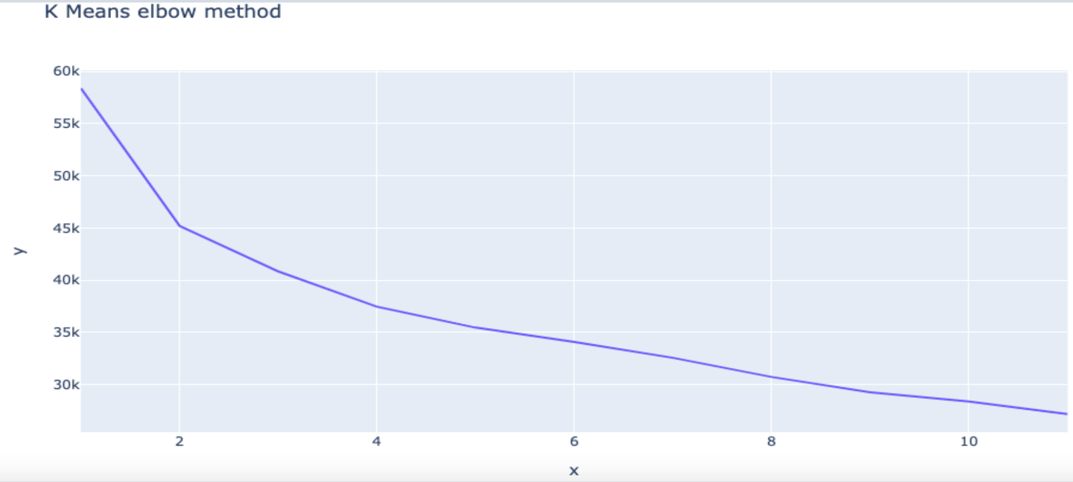
A) K-Means

K-Means clustering is one of the simplest and most popular unsupervised methods uses continuous data for machine learning. As K-Means clustering is one of the unsupervised learning, we do not need to tell the machine which attribute/pattern we want to know, after we do the clustering, the machine will tell us its discovery. In this case, we want to learn more about the different type of customers, K-Means clustering is suitable.
Reasons why K-Means is suitable for our analysis is as follows.

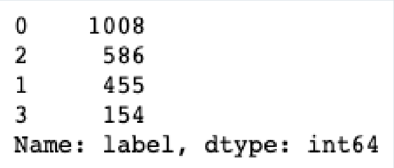
- Scales to large datasets
- Guarantee convergences
- Accuracy
- Easy to interpret

Before building the model, we must determine the number of our clustering.

We used elbow method to determine the optimal cluster number, which is 4, for building K-Means clustering model.

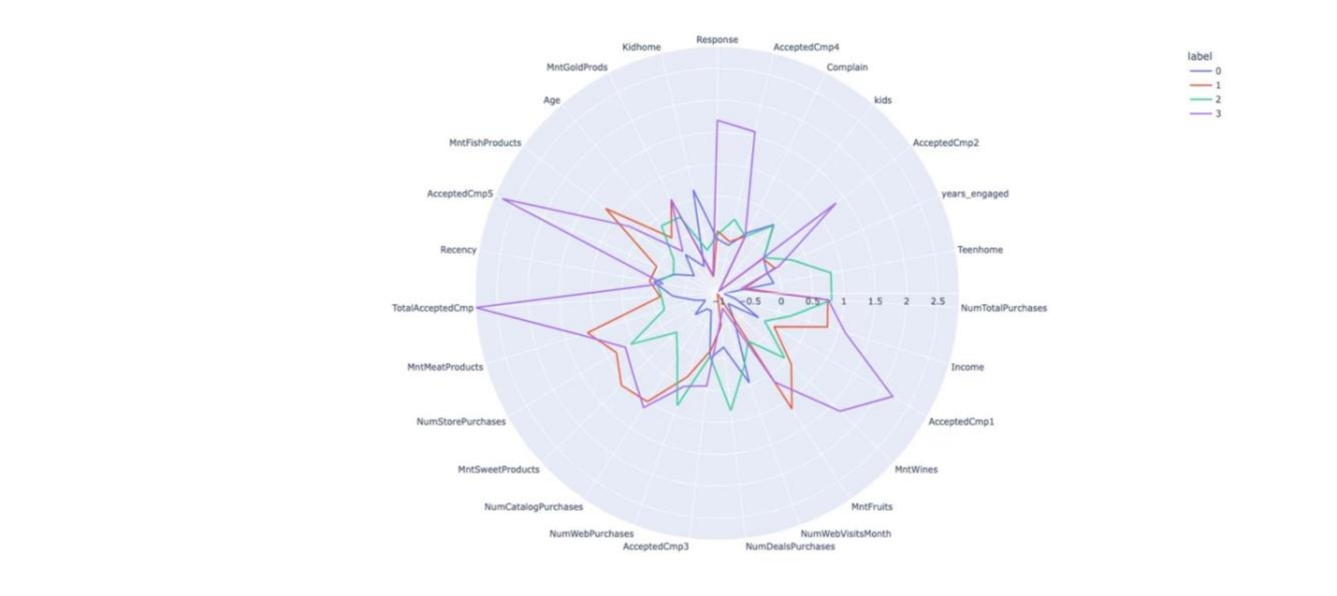


The total number of customers in each segmentation:



```
0    1008
2     586
1     455
3     154
Name: label, dtype: int64
```

After looking through the number of each segmentation, we can know that there are mainly three types of customers, which contains more than 99% of customers. As we have 27 columns it is difficult for us to display the relationship in a 2-D or 3-D graph. However, we have deduced a way to summarize the result by drawing a circular graph that will have all 27 variables on it. Those will represent the average value for each of the variables.



From the chart above, we ranked different variables from smallest to largest. Below is the ranking of all the segments based on certain attributes-

Summary Statistics-
Segments' personal characteristics:

Income: segment 0 < segment 2 < segment 1 < segment 3
Age: segment 0 < segment 3 < segment 1 < segment 2
Kidhome: segment 1 < segment 3 < segment 2 < segment 0
Teenhome: segment 3 < segment 1 < segment 0 < segment 2

Segments' different channels' purchase amounts:

TotalAcceptedCmp: segment 0 < segment 2 < segment 1 < segment 3
NumberTotalPurchases: segment 0 < segment 1 < segment 3 < segment 2
NumCatalogPurchases: segment 0 < segment 2 < segment 1 < segment 3
NumWebPurchases: segment 0 < segment 1 < segment 3 < segment 2
NumDealsPurchases: segment 3 < segment 1 < segment 0 < segment 2
NumWebVisitsMonth: segment 1 < segment 3 < segment 2 < segment 0

Segments' buying categories:

MntFruits: segment 0 < segment 2 < segment 3 < segment 1
MntWines: segment 0 < segment 2 < segment 1 < segment 3
MntSweetProducts: segment 0 < segment 2 < segment 3 < segment 1
MntMeatProducts: segment 0 < segment 2 < segment 1 < segment 3
MntGoldProducts: segment 0 < segment 2 < segment 1 < segment 3

MntFishProducts: segment 0 < segment 2 < segment 3 < segment 1
MntMeatProducts: segment 0 < segment 2 < segment 1 < segment 3

Customer satisfaction:

segment 1 < segment 3 < segment 2 < segment 0

Segment wise Analysis -

For business implementation, we'll take three segments which cover 99% of the customers i.e., label 1, label 2 and label 0.

Basis of choosing these three labels –

- Total Purchases.
- Customer population in each segment.

Segment 1

This group is the middle-aged group having the highest income out of all three segments which is why their purchase of fruit, sweets, fish, and all other products is the highest. These are the customers who do not care much about the deals and their web visits is the least. They do most of their purchase from the catalog or go directly to the store to purchase their products. Their campaign acceptance is also the least indicating that they either don't like the campaign or are not interested in the products offered in that.

Assumption about the segment-

Given the above features, we consider them as single or just married people. They have high income, but they may not have free time to enjoy their life. Their web visits being the least could be because of the reason that they are working population which leaves them no time to go through the websites or campaigns.

Segment 2

This group is the aged population having both smaller kids and teenagers at home. They are the second highest earners out of the three segments. Their website views are the highest which could be because of the reason that they have no time left to go to store to purchase the products. Even though their income is not as high as segment 1, they still purchase wine and gold indicating that they are both enjoying their present life and saving for future purposes.

Assumption about the segment-

Given the features, we could consider them as medium-to-high income family who tend to invest in low-risk products to save money by taking deals. Their income being low could be because of reasons like basic education, single source of income etc.

Segment 0

This is the younger generation who visit the website the greatest number of the times. They have smaller kids at home and are not satisfied with the service i.e. complain the most. Their gold products consumption is the least indicating that they are not planning. Their website visits are the highest, but their consumption is the least through web purchase showing that either that they face difficulty accessing the website content or they are waiting for more deals to come as they accept campaigns and deals provided to them.

Assumption about the segment-

Given the features, we can assume that this segment is either studying and single or is widow/ divorced as their income is not that much.

By evaluating the graph, we learn the traits of each type of customers. For segment 0, they have the lowest income and age, but has more kids than other groups. Although their accepted campaign numbers and total purchase amount are the lowest, the number of website visits are the highest. We could imply that they are the most price-sensitive group. However, compared to other campaigns, segment 0 has a preference to campaign 3. In the future, if we want to increase segment 0's spending, we can hold more campaigns like campaign 3.

Limitations of using K means clustering model –

K means does not consider the categorical value we have in our dataset such as Education, marital status and does an incomplete segmentation of the customers. Also, each time the model is run, it produces different results which can prove inefficient for an organization. Therefore, to have proper segments we will use another model that considers categorical value too.

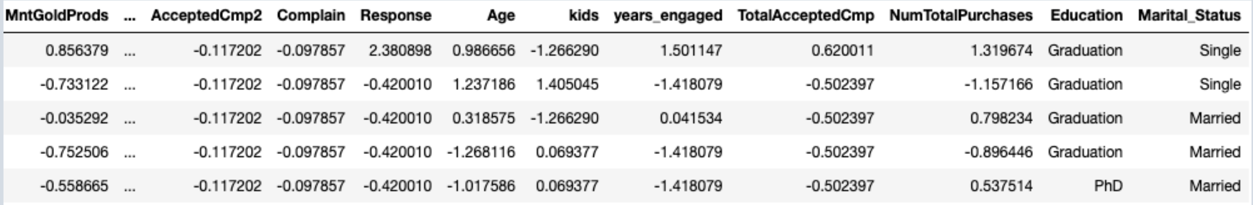
Although we were able to derive very useful business insights from the above analysis, there are still some gaps left that will help us further. For example,

1. We still can't say which segment is highly educated or which is not?
2. How is education level impacting the shopping behavior?
3. How marital status is affecting the shopping behavior of customers?
4. How can we relate marital status with education level to create campaigns specifically?

Though we have the respective variables, we are not able to utilize them as K-Means won't be able to properly work with categorical values. Even after creating dummy values, it can only be converted into discretized variable which itself won't change the output of K-Means

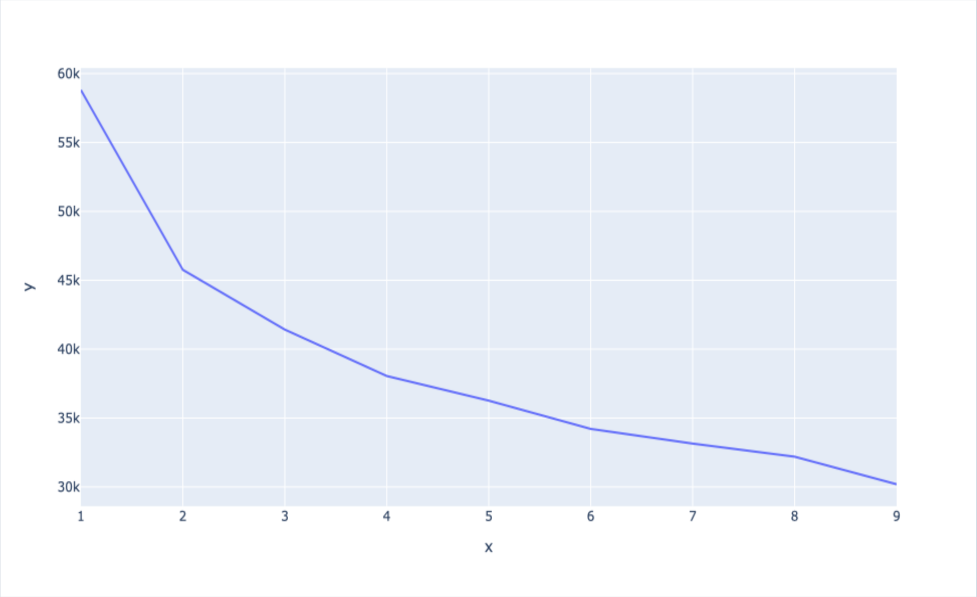
B) K-Prototype

K-Prototype is the combination of K-Means and K-Mode machine learning algorithms. The reason why we are using K-Prototype is because K-Means cannot deal with categorical data. K-means can only handle continuous data. By using K-Prototype, we can utilize important categorical variables "Education" and "Marital_status" to build our model and learn some new insights about how these two variables influence the customer clustering.

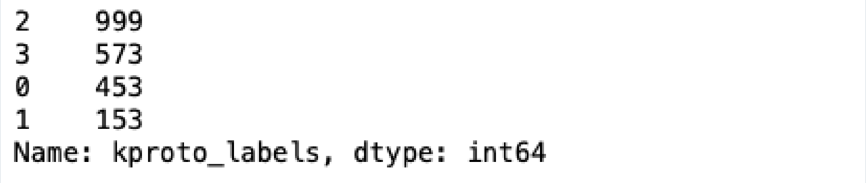


MntGoldProds	...	AcceptedCmp2	Complain	Response	Age	kids	years_engaged	TotalAcceptedCmp	NumTotalPurchases	Education	Marital_Status
0.856379	...	-0.117202	-0.097857	2.380898	0.986656	-1.266290	1.501147	0.620011	1.319674	Graduation	Single
-0.733122	...	-0.117202	-0.097857	-0.420010	1.237186	1.405045	-1.418079	-0.502397	-1.157166	Graduation	Single
-0.035292	...	-0.117202	-0.097857	-0.420010	0.318575	-1.266290	0.041534	-0.502397	0.798234	Graduation	Married
-0.752506	...	-0.117202	-0.097857	-0.420010	-1.268116	0.069377	-1.418079	-0.502397	-0.896446	Graduation	Married
-0.558665	...	-0.117202	-0.097857	-0.420010	-1.017586	0.069377	-1.418079	-0.502397	0.537514	PhD	Married

There are two reasons why we still choose 4 as our new clustering number. The first reason is that we use elbow method to determine the k value. We can look at the following graph showing that the elbow occurs when clustering number equal to 4. The other reason is that we would like to compare the differences between K-Means and K-Prototype models, knowing more about how the two categorical variables has influenced the model.



The total number of customers in each segmentation:



```
2    999
3    573
0    453
1    153
Name: kproto_labels, dtype: int64
```

We can see that there are no big differences on the clustering number between K-Means and K-Prototype models. From the chart below, we ranked segments from smallest to largest for each important variable.

From the chart below, we ranked different variables from smallest to largest. Below is the ranking of all the segments based on certain attributes-

Summary Statistics-
Segments’ personal characteristics:

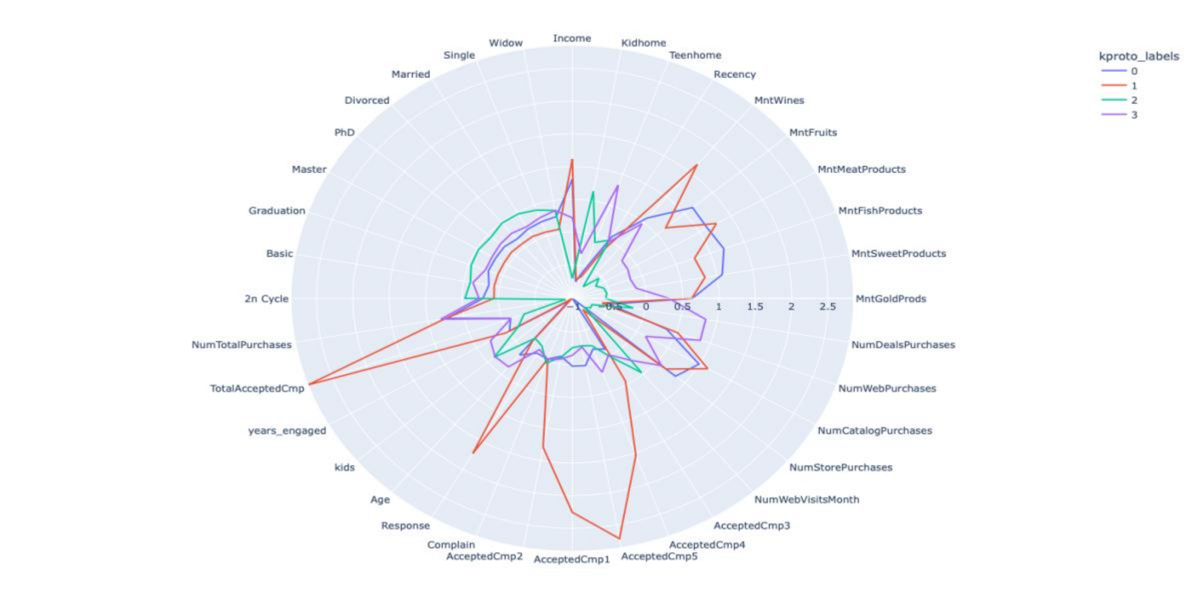
Income: segment 2 < segment 3 < segment 0 < segment 1
Age: segment 2 < segment 1 < segment 0 < segment 3
Kidhome: segment 0 < segment 1 < segment 3 < segment 2
Teenhome: segment 1 < segment 0 < segment 2 < segment 3

Segments’ different channels’ purchase amounts:

TotalAcceptedCmp: segment 2 < segment 3 < segment 0 < segment 1
NumberTotalPurchases: segment 2 < segment 0 < segment 1 < segment 3
NumCatalogPurchases: segment 2 < segment 3 < segment 0 < segment 1
NumWebPurchases: segment 2 < segment 0 < segment 1 < segment 3
NumDealsPurchases: segment 1 < segment 0 < segment 2 < segment 3
NumWebVisitsMonth: segment 0 < segment 1 < segment 3 < segment 2

Segments’ buying categories:
MntFruits: segment 2 < segment 3 < segment 1 < segment 0
MntWines: segment 3 < segment 2 < segment 0 < segment 1
MntSweetProducts: segment 2 < segment 3 < segment 1 < segment 0
MntMeatProducts: segment 2 < segment 3 < segment 0 < segment 1
MntGoldProducts: segment 2 < segment 3 < segment 1 < segment 0
MntFishProducts: segment 2 < segment 3 < segment 1 < segment 0
MntMeatProducts: segment 2 < segment 3 < segment 0 < segment 1

Customer satisfaction:
segment 2 < segment 3 <= segment 0 < segment 1



Segment wise Analysis -

For business implementation, we'll take three segments which cover 99% of the customers i.e. label 2, label 3 and label 0.

Basis of choosing these three labels –

- Total Purchases.
- Customer population in each segment.

Compared to K-Means clustering, K-Prototypes clustering does not change much. But after doing K-Prototypes clustering, we can know more about the customers.

Customer segment in K-Means	=	Customer segment in K-Prototype
0	=	2
1	=	0
2	=	3
3	=	1

As we can see in the above table, the correspondence of clusters in K- means to that of K-Prototype.

Segment 2 –
Reason for the features mentioned in segment 0 of K means-

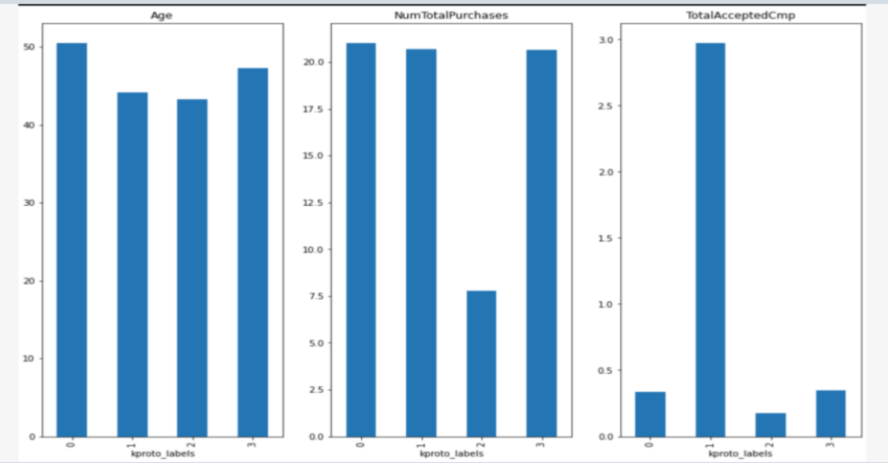
K Prototypes provides us the reason that our assumption that this segment is either studying and single or is widow/ divorced as their income is not that much is true.

Segment 3–
Reason for the features mentioned in segment 1 of K means-

K Prototypes provides us the reason that our assumption that this segment is single or is working population is true. They have high income, but they may not have free time to enjoy their life. Their web visits being the least could be because of the reason that they are working population which leaves them no time to go through the websites or campaigns.

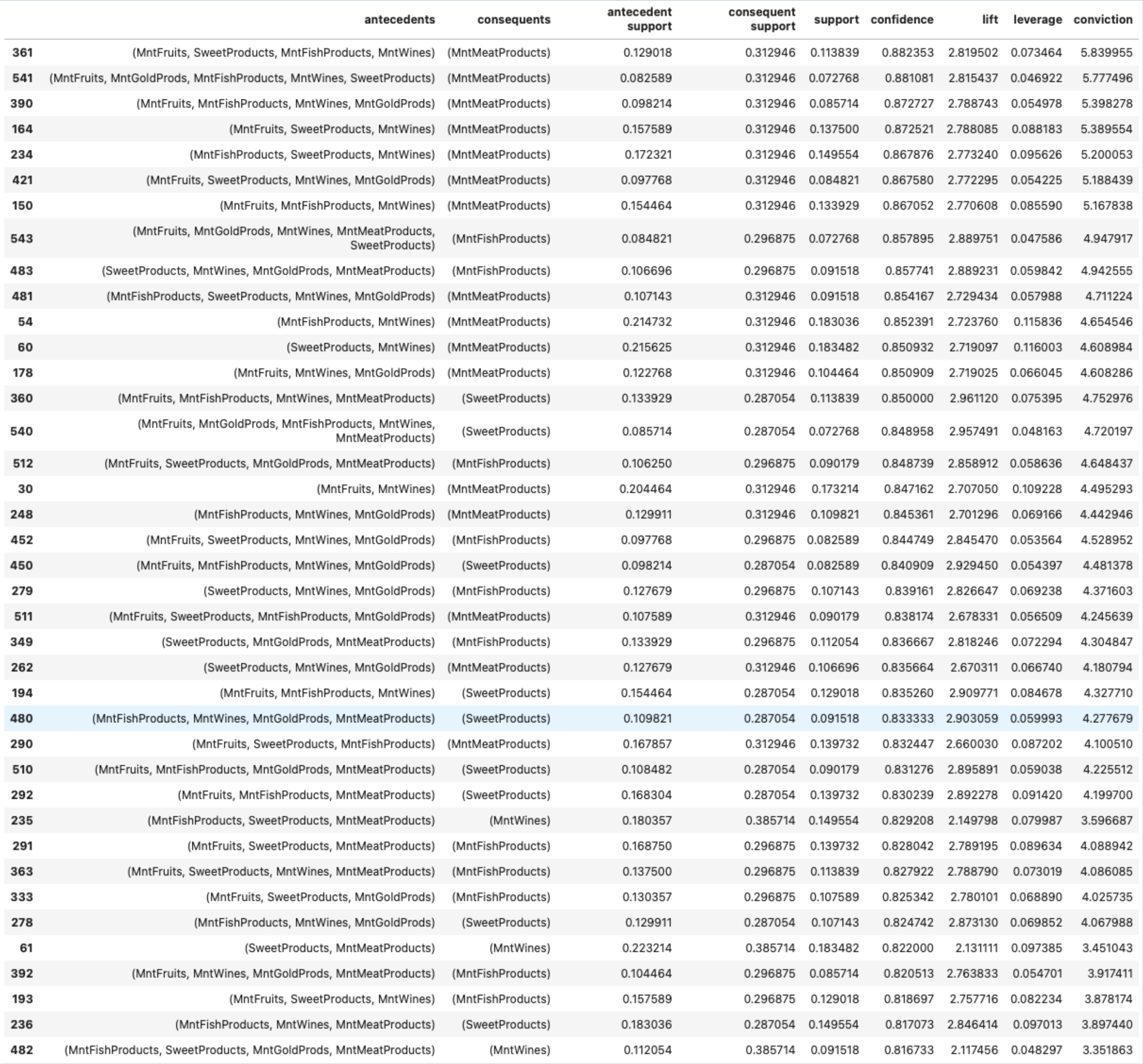
We can know that the richest segment does not have pattern of their education and marital status. Most surprisingly, segment 2 has the lowest income, but they have the high level of education. Segment 1 which is the richest segment, have a higher probability of being a PhD. And same as our thought in K-means clustering, segment 3 which we considered as medium-to-high income family and tend to invest in low-risk products to save money is also fit to our thought. This type of customers has a higher chance to be divorced or widow, but they have children to raise, therefore they have to invest in low-risk products, they could not afford the high risk.

Our above analysis can also be deduced to the graphs below -



C) Association Rule Mining

Association rule mining is a technique to discover potential relationships among different items. When it comes to association rule mining, the dataset “beer and diaper” pops into people’s mind. In our case, the situation is similar. We want to use specific customer segments and target them on the basis of certain products which are bought together. After doing association rule mining, we can know what kind of promotions or bundles are fit to what type of customers. It could help company increase profit by creating new selling bundles and promotional campaigns.



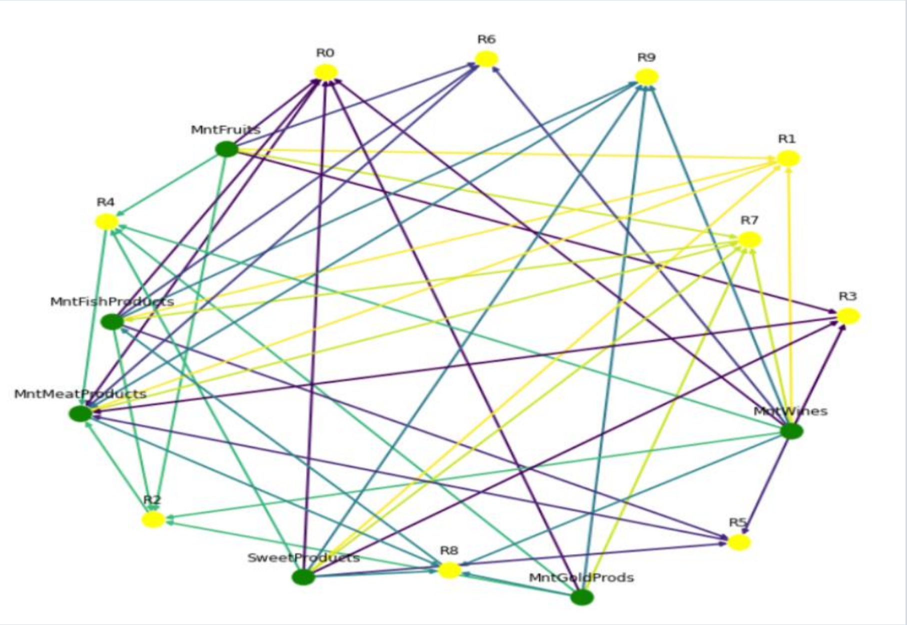
	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
361	(MntFruits, SweetProducts, MntFishProducts, MntWines)	(MntMeatProducts)	0.129018	0.312946	0.113839	0.882353	2.819502	0.073464	5.839955
541	(MntFruits, MntGoldProds, MntFishProducts, MntWines, SweetProducts)	(MntMeatProducts)	0.082589	0.312946	0.072768	0.881081	2.815437	0.046922	5.777496
390	(MntFruits, MntFishProducts, MntWines, MntGoldProds)	(MntMeatProducts)	0.098214	0.312946	0.085714	0.872727	2.788743	0.054978	5.398278
164	(MntFruits, SweetProducts, MntWines)	(MntMeatProducts)	0.157589	0.312946	0.137500	0.872521	2.788085	0.088183	5.389554
234	(MntFishProducts, SweetProducts, MntWines)	(MntMeatProducts)	0.172321	0.312946	0.149554	0.867876	2.773240	0.095626	5.200053
421	(MntFruits, SweetProducts, MntWines, MntGoldProds)	(MntMeatProducts)	0.097768	0.312946	0.084821	0.867580	2.772295	0.054225	5.188439
150	(MntFruits, MntFishProducts, MntWines)	(MntMeatProducts)	0.154464	0.312946	0.133929	0.867052	2.770608	0.085590	5.167838
543	(MntFruits, MntGoldProds, MntWines, MntMeatProducts, SweetProducts)	(MntFishProducts)	0.084821	0.296875	0.072768	0.857895	2.889751	0.047586	4.947917
483	(SweetProducts, MntWines, MntGoldProds, MntMeatProducts)	(MntFishProducts)	0.106696	0.296875	0.091518	0.857741	2.889231	0.059842	4.942555
481	(MntFishProducts, SweetProducts, MntWines, MntGoldProds)	(MntMeatProducts)	0.107143	0.312946	0.091518	0.854167	2.729434	0.057988	4.711224
54	(MntFishProducts, MntWines)	(MntMeatProducts)	0.214732	0.312946	0.183036	0.852391	2.723760	0.115836	4.654546
60	(SweetProducts, MntWines)	(MntMeatProducts)	0.215625	0.312946	0.183482	0.850932	2.719097	0.116003	4.608984
178	(MntFruits, MntWines, MntGoldProds)	(MntMeatProducts)	0.122768	0.312946	0.104464	0.850909	2.719025	0.066045	4.608286
360	(MntFruits, MntFishProducts, MntWines, MntMeatProducts)	(SweetProducts)	0.133929	0.287054	0.113839	0.850000	2.961120	0.075395	4.752976
540	(MntFruits, MntGoldProds, MntFishProducts, MntWines, MntMeatProducts)	(SweetProducts)	0.085714	0.287054	0.072768	0.848958	2.957491	0.048163	4.720197
512	(MntFruits, SweetProducts, MntGoldProds, MntMeatProducts)	(MntFishProducts)	0.106250	0.296875	0.090179	0.848739	2.858912	0.058636	4.648437
30	(MntFruits, MntWines)	(MntMeatProducts)	0.204464	0.312946	0.173214	0.847162	2.707050	0.109228	4.495293
248	(MntFishProducts, MntWines, MntGoldProds)	(MntMeatProducts)	0.129911	0.312946	0.109821	0.845361	2.701296	0.069166	4.442946
452	(MntFruits, SweetProducts, MntWines, MntGoldProds)	(MntFishProducts)	0.097768	0.296875	0.082589	0.844749	2.845470	0.053564	4.528952
450	(MntFruits, MntFishProducts, MntWines, MntGoldProds)	(SweetProducts)	0.098214	0.287054	0.082589	0.840909	2.929450	0.054397	4.481378
279	(SweetProducts, MntWines, MntGoldProds)	(MntFishProducts)	0.127679	0.296875	0.107143	0.839161	2.826647	0.069238	4.371603
511	(MntFruits, SweetProducts, MntFishProducts, MntGoldProds)	(MntMeatProducts)	0.107589	0.312946	0.090179	0.838174	2.678331	0.056509	4.245639
349	(SweetProducts, MntGoldProds, MntMeatProducts)	(MntFishProducts)	0.133929	0.296875	0.112054	0.836667	2.818246	0.072294	4.304847
262	(SweetProducts, MntWines, MntGoldProds)	(MntMeatProducts)	0.127679	0.312946	0.106696	0.835664	2.670311	0.066740	4.180794
194	(MntFruits, MntFishProducts, MntWines)	(SweetProducts)	0.154464	0.287054	0.129018	0.835260	2.909771	0.084678	4.327710
480	(MntFishProducts, MntWines, MntGoldProds, MntMeatProducts)	(SweetProducts)	0.109821	0.287054	0.091518	0.833333	2.903059	0.059993	4.277679
290	(MntFruits, SweetProducts, MntFishProducts)	(MntMeatProducts)	0.167857	0.312946	0.139732	0.832447	2.660030	0.087202	4.100510
510	(MntFruits, MntFishProducts, MntGoldProds, MntMeatProducts)	(SweetProducts)	0.108482	0.287054	0.090179	0.831276	2.895891	0.059038	4.225512
292	(MntFruits, MntFishProducts, MntMeatProducts)	(SweetProducts)	0.168304	0.287054	0.139732	0.830239	2.892278	0.091420	4.199700
235	(MntFishProducts, SweetProducts, MntMeatProducts)	(MntWines)	0.180357	0.385714	0.149554	0.829208	2.149798	0.079987	3.596687
291	(MntFruits, SweetProducts, MntMeatProducts)	(MntFishProducts)	0.168750	0.296875	0.139732	0.828042	2.789195	0.089634	4.088942
363	(MntFruits, SweetProducts, MntWines, MntMeatProducts)	(MntFishProducts)	0.137500	0.296875	0.113839	0.827922	2.788790	0.073019	4.086085
333	(MntFruits, SweetProducts, MntGoldProds)	(MntFishProducts)	0.130357	0.296875	0.107589	0.825342	2.780101	0.068890	4.025735
278	(MntFishProducts, MntWines, MntGoldProds)	(SweetProducts)	0.129911	0.287054	0.107143	0.824742	2.873130	0.069852	4.067988
61	(SweetProducts, MntMeatProducts)	(MntWines)	0.223214	0.385714	0.183482	0.822000	2.131111	0.097385	3.451043
392	(MntFruits, MntWines, MntGoldProds, MntMeatProducts)	(MntFishProducts)	0.104464	0.296875	0.085714	0.820513	2.763833	0.054701	3.917411
193	(MntFruits, SweetProducts, MntWines)	(MntFishProducts)	0.157589	0.296875	0.129018	0.818697	2.757716	0.082234	3.878174
236	(MntFishProducts, MntWines, MntMeatProducts)	(SweetProducts)	0.183036	0.287054	0.149554	0.817073	2.846414	0.097013	3.897440
482	(MntFishProducts, SweetProducts, MntGoldProds, MntMeatProducts)	(MntWines)	0.112054	0.385714	0.091518	0.816733	2.117456	0.048297	3.351863

In the above chart we get the itemset using Apriori Algorithm. As you can see, we only display the product combinations whose confidence are bigger than 80%. The reason why we choose 80% as our threshold confidence level is because when the confidence exceeds 80%, it means significant relationship between them.

Noticeable Pattern-

One noticeable pattern is that people who buy meat also like to buy wines and fruits. People who buy fish also tend to buy meat. Customers who purchase sweet also have a high probability to buy meat, fish, or fruits. Through this association rule, we can use them to customize promotions and campaigns which are more suitable for different segment of customers.

The below graph supplements the relationship found in the association rules-



Combining K prototypes with Association rule -

K prototypes tells us the shopping features that each segment possesses but it will only be useful for the organization if they can take it to the marketing and advertising level.

Therefore, using association rule with the k prototypes segment will provide the organization with the item basket they should market to each segment.

For example-

Associate rule 30 tells us that people who buy fruit and wine tend to buy meat product as well. This item basket can be promoted to segment 0 as this segment tends to buy more fruits.

Association rule 61 tells us that that people who buy sweet products and meat tend to buy more of wine as well. This item basket can be catered to segment 1 as they are already purchasing more of wine and meat.

4. Next Step-

For business use case purposes, we divided the segments into 3 group based on their total purchase i.e., High spenders, average and lowest spenders.

This will not only help the organization know the traits of the customer but will also help them to efficiently use their resources for marketing and campaigns.

High spenders (Segment 3)	**Average spenders (Segment 0)**	**Low spenders (Segment 2)**
Website purchases	Highest purchase of all products	Less consumption of all products
Deal purchases	Catalog purchases	Deal purchases highest
Mediocre income	Highest income earner	Lowest income earner

5. Recommendations for businesses catering to these segments-

High Spenders	**Average Spenders**	**Low Spenders**
Maintain website content	Making catalog content more informative.	More discounts.
Providing them more deals	Maintain campaign directment to them.	Increase the customer care support.
As their campaign acceptance is low, focus on campaigns that cater to them the most.	Using association rule basket like to provide deals for wine with meat.	

- The low spenders are the ones who are least satisfied with the service which could be due to difficulty in accessing website, low customer attention etc. One way to solve this issue is to direct more customer care support to these segments.

- For the average spenders, catalog and campaigns are important so making sure that the campaign content is clear, and the catalog provides all the desired information is of utmost importance. Furthermore, association rule basket will help in identifying the products which can be bought together by this segment.

- For the high spenders there are only two things that organization should focus on, which is giving them heavy discounts/ deals as these are the biggest customers of the company. Also, as this segment does most of purchase from the site, website should be maintained and updated regularly. Providing deals through the website is a good solution for the organization at hand.

The above recommendations and analysis answer our questions of which marketing to be used for each customer segment and also the products that should be marketed together.

6. Conclusion-

Catering to all the customers in the same manner will not only lead to inefficient use of resources by the organization but also to dissatisfaction of the customer in turn leading to revenue loss. Hence, using customer segmentation and association rule together will help the organization to make the necessary changes in an efficient manner.