INDIAN INSTITUTE OF TECHNOLOGY BOMBAY

WIDS

Consumer Behaviour Analysis

Submitted By:Dinesh Pondreti

Roll no. 190010052

Contents

1 Objective **2**

2 Data **2**
2.1 Attributes 2
2.2 Data type of each attribute 3

3 Data Pre Processing **3**
3.1 Finding Null Values 3
3.2 Categorical Classification of columns 4
3.3 Unification of columns 5
3.4 Checking Outliers 5
3.5 One Hot Encoding 6

4 EDA **7**
4.1 Correlation among Attributes 7
4.2 Distribution of Income vs Education 7
4.3 Distribution of Expenses vs Education 8
4.4 Distribution of Expenses vs Marital Status 10
4.5 Distribution of Expenses vs Age 10
4.6 Mode of Purchase vs Age Category 12
4.7 Most Sold Products 13
4.8 Campaigns 15

5 Dimensionality Reduction and Clustering Analysis **15**
5.1 High Spending Customers 17
5.2 Low Spending Customers 18
5.3 Age and Marital Status in Clustering 19

6 Apriori Algorithm **20**
6.1 Implementation 20
6.2 Results 20

7 Overall Conclusions **21**

1 Objective

To study , understand the consumer buying patterns and perform clustering and market basket analysis.

2 Data

2.1 Attributes

The dataset has 27 columns

People:

- ID: Customer's unique identifier
- Year Birth: Customer's birth year
- Education: Customer's education level
- Marital Status: Customer's marital status
- Income: Customer's yearly household income
- Kids: Number of children in customer's household
- Teen: Number of teenagers in customer's household
- Customer Dt: Date of customer's enrollment with the company
- Last purchase: Number of days since customer's last purchase
- Complain: 1 if customer complained in the last 2 years, 0 otherwise

Products:

- Wines: Amount spent on wine in last 2 years
- Fruits: Amount spent on fruits in last 2 years
- MeatProducts: Amount spent on meat in last 2 years
- FishProducts: Amount spent on fish in last 2 years
- SweetProducts: Amount spent on sweets in last 2 years
- GoldProds: Amount spent on gold in last 2 year

Promotion:

- 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 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
- Response: 1 if customer accepted the offer in the last campaign, 0 otherwise

Place:

- WebPurchases: Number of purchases made through the company's web site
- CatalogPurchases: Number of purchases made using a catalogue
- StorePurchases: Number of purchases made directly in stores
- WebVisitsMonth: Number of visits to company's web site in the last month

2.2 Data type of each attribute

	column_name	dtype	column_name	dtype	column_name	dtype
0	ID	int64	Wines	int64	StorePurchases	int64
1	Year_Birth	int64	Fruits	int64	WebVisitsMonth	int64
2	Education	object	MeatProducts	int64	AcceptedCmp3	int64
3	Marital_Status	object	FishProducts	int64	AcceptedCmp4	int64
4	Income	float64	SweetProducts	int64	AcceptedCmp5	int64
5	Kids	int64	GoldProds	int64	AcceptedCmp1	int64
6	Teen	int64	NumDealsPurchases	int64	AcceptedCmp2	int64
7	Customer_Dt	object	WebPurchases	int64	Complain	int64
8	Last_purchase	int64	CatalogPurchases	int64	Response	int64

Figure 1: column data types

3 Data Pre Processing

3.1 Finding Null Values

In order to check the null values in the dataset, df.isnull.sum() function is used in pandas library. After using this function it is observed that only income column has 24 null values. All of these null values are replaced with the mean of income column after removing potential outliers.

3.2 Categorical Classification of columns

From the table in the figure 1 it is seen that the columns 'Education', Marital Status' and 'Customer Dt' are object data types.Education and Marital Status comes under nominal data scale whereas Customer Dt is ordinal data scale. The Education column has 8 categories which are shown in below table:

Category	Total Count
Married	864
Together	580
Single	480
Divorced	232
Widow	77
Alone	3
Yolo	2
Absurd	2

We can see here that the categories named married , together sounds similar and the categories single, divorced, widow, alone sounds alike. The last two categories are meaningless in the marital status. Here I assumed last two categories are also single and these are stacked with single. So the 8 categories column now becomes two categories with categories named 'Single' and 'Married'.The transformed column categorical count is shown in below table.

Category	Total Count
Married	1444
Together	796

Similarly Education column has 5 categories named and those are listed in below table

Category	Total Count
Basic	54
Graduation	1127
2n Cycle	203
Master	370
PhD	486

Here the 5 categories classified into two categories, customers who have done only basic or graduation is stacked into level 1 Education category and the rest of the people grouped into level 2 Education. Here is the table after transforming into two categories.

Category	Total Count
level 1 Education	1181
level 2 Education	1059

The 'Customer Dt' column is also object type, which tells date on which customer enrolled in the company customers list.This date is transformed into number of days

from the date of enrollment till 01-01-2022. All these are done easily by using datetime library. Similarly the date of birth column is also is converted into number of years. Using this age we can categorize the customers into three different categories which is shown in below table.

Category	Age
Young adult	18 <= age <=30
Adult Education	31 <= age <= 45
Adult	Above 45

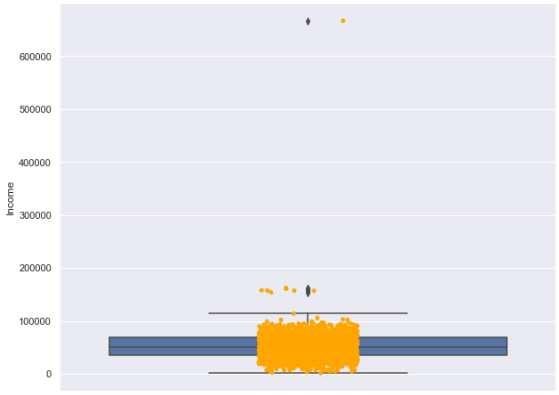
Based on the above criteria a separate column Age Cat is constructed.

3.3 Unification of columns

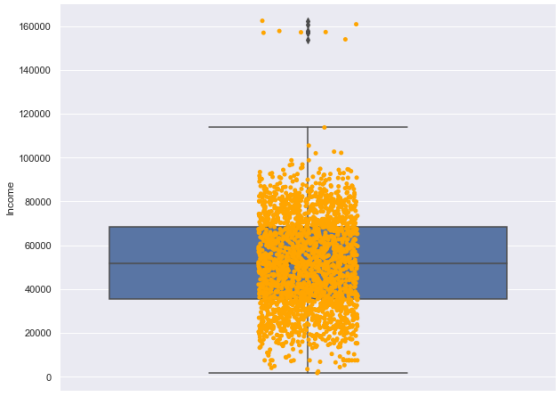
To perform clustering algorithm on the given data we can merge some similar columns and get a combined number which essentially gives all information about it. A new column is created called tot exp which is total sum of 'Wines', 'Fruits','MeatProducts', 'FishProducts', 'SweetProducts'and 'GoldProds'. In the same way tot cam is created which is the total sum of 'AcceptedCmp3', 'AcceptedCmp4', 'AcceptedCmp5', 'AcceptedCmp1', 'AcceptedCmp2', 'Complain', 'Response'. The column tot purchases is created which is total sum of 'NumDealsPurchases', 'WebPurchases', 'CatalogPurchases', 'StorePurchases'. Kids Teen is another column which tells total number of teens and kids in a particular customer household.

3.4 Checking Outliers

When outliers are present it drastically affects the mean; hence, they should be identified and removed before running any algorithm on the given dataset. The 'Income' seems to have outliers that are beyond the interquartile range , but all of these outliers are not so far from the median of the income, ,although there is one outlier that is very far from the interquartile range that is removed. The following two box plots indicate the distribution of customer's income before and after removing outliers.



(a) Box Plot of Income



(b) Box Plot after removing outlier

3.5 One Hot Encoding

The following table depicts the dataset attributes and their datatypes after data preprocessing and object type attributes are one hot encoded.

column_name	dtype	column_name	dtype
ID	int64	tot_purchases	int64
Year_Birth	int64	tot_cam	int64
Income	float64	Marital_S_Married	uint8
Customer_Dt	int64	Marital_S_Single	uint8
Last_purchase	int64	Education_level 1 Education	uint8
WebVisitsMonth	int64	Education_level 2 Education	uint8
Complain	int64	Age_Cat_Adult	uint8
tot_exp	int64	Age_Cat_Old Adult	uint8
Kids_Teen	int64	Age_Cat_Young Adult	uint8

Figure 3: column data types

4 EDA

4.1 Correlation among Attributes

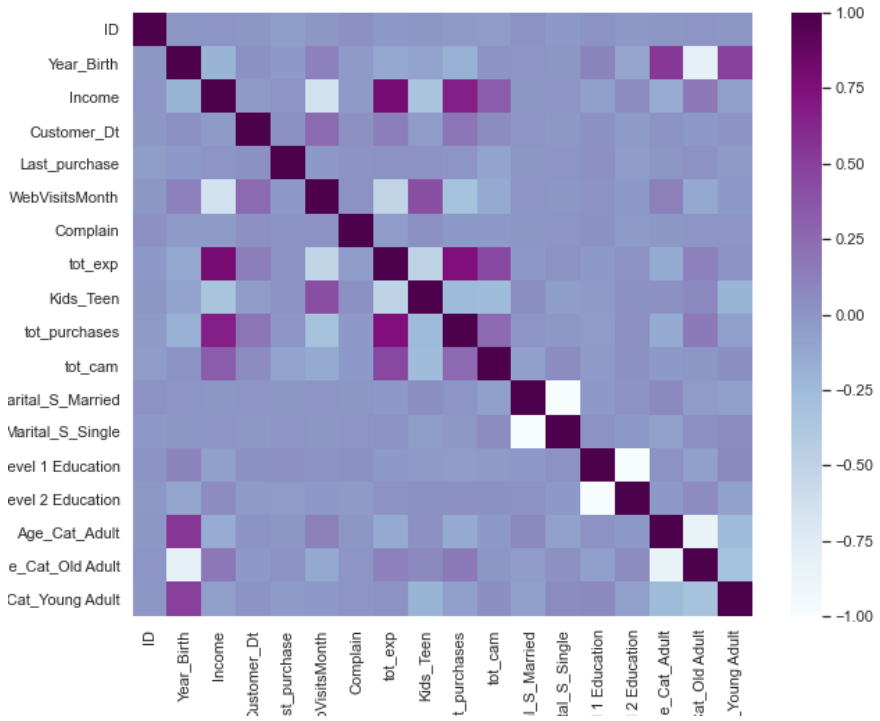
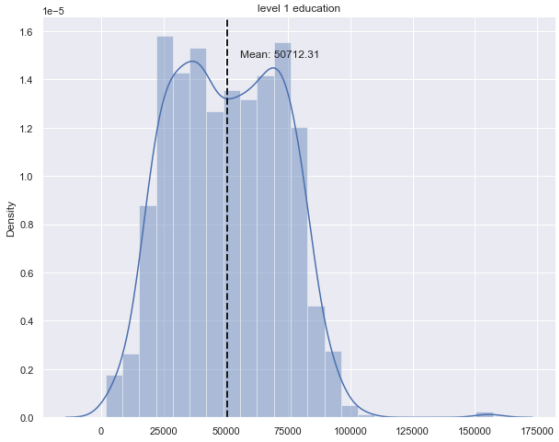


Figure 4: heat map

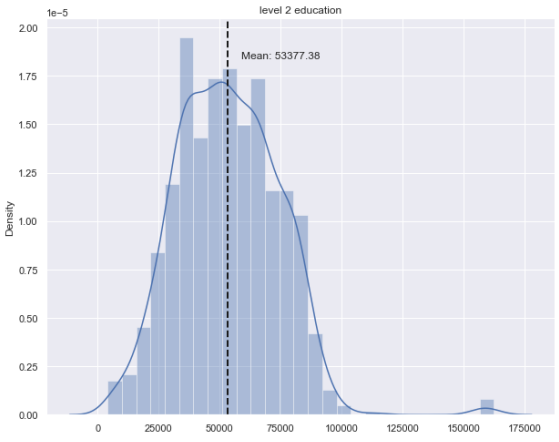
It is observed that total purchases is highly correlated with total expenses , so tot purchases is removed.

4.2 Distribution of Income vs Education

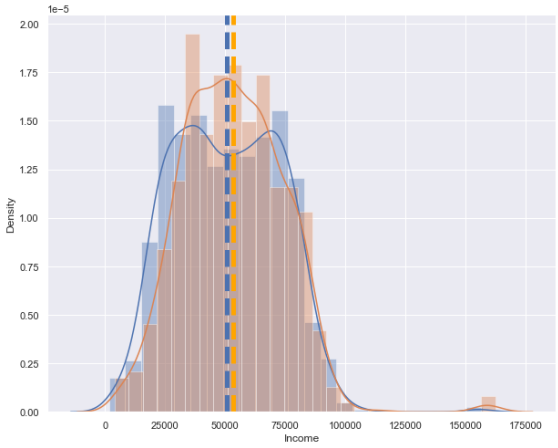
As we know that the education attribute has two categories, one is level 1 education and the other is level 2 education. We want to know whether people who had done level 2 education earning more than level 1 education. We need to incorporate some statistical methods to prove the above hypothesis. T test is used to compare the two different distributions. The following graphs indicate the distribution of income in each category:



(a) Level 1 Education



(b) Level 2 Education



(c) Level 1 and Level 2 Education

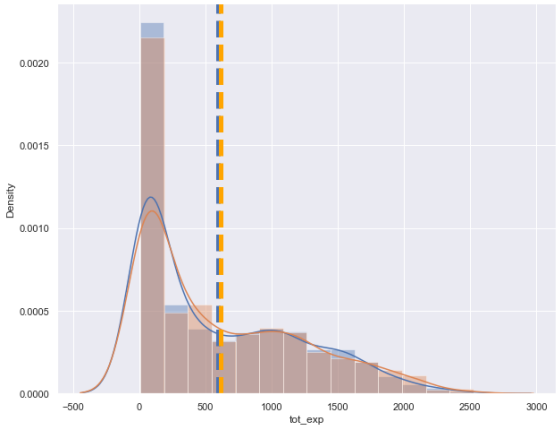
I took a sample size of 50 and performed a t test with significance level 0.05, we get a p value of 0.064 which is greater than the significance level it indicates that both the population has same mean.So we can say that average income of the given data set is around 52k irrespective of level 1 or level 2 education.

4.3 Distribution of Expenses vs Education

Here also we can do same analysis to check whether education has any effect on expenses or not. The difference is we use a different test called Mann-Whitney U Test which is heavily used in place of t test when there is normal distribution violation.

(a) Level 1 Education Expenses

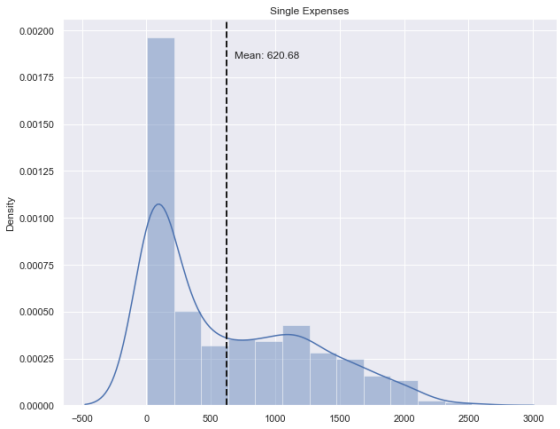
(b) Level 2 Education Expenses



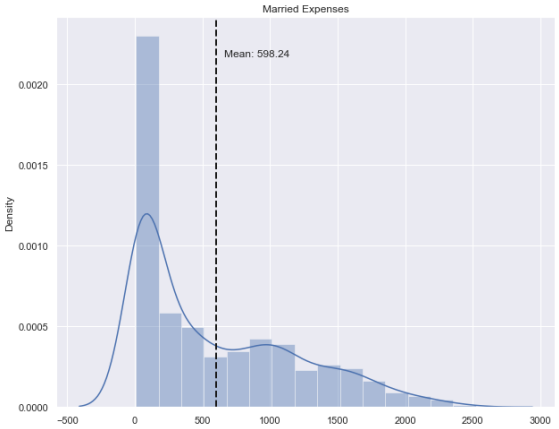
(c) Level 1 and Level 2 Expenses

For a sample size of 50, the test gives a p value of 0.4 which is greater than significance level (0.05), which tells that on an average the expenses are 606 rupees.

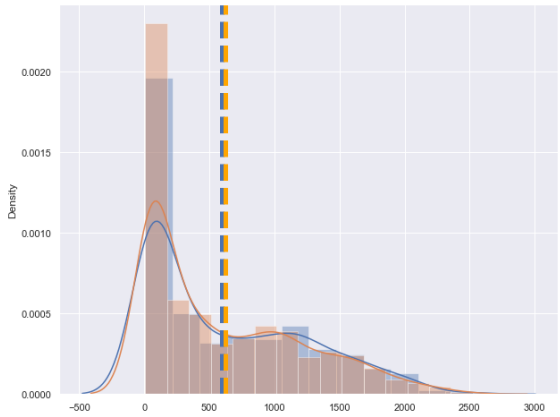
4.4 Distribution of Expenses vs Marital Status



(a) Single Expenses



(b) Married Expenses



(c) Single and Married Expenses

Mann-Whitney U Test is performed. It gave a p value of 0.2 which means the two distributions are same with a spending average of 606 rupees.

4.5 Distribution of Expenses vs Age

We try to see whether the attribute age has any leverage to effect the expenses. Before we go into distributions related to them let's have l0ok at age histogram which is shown below.

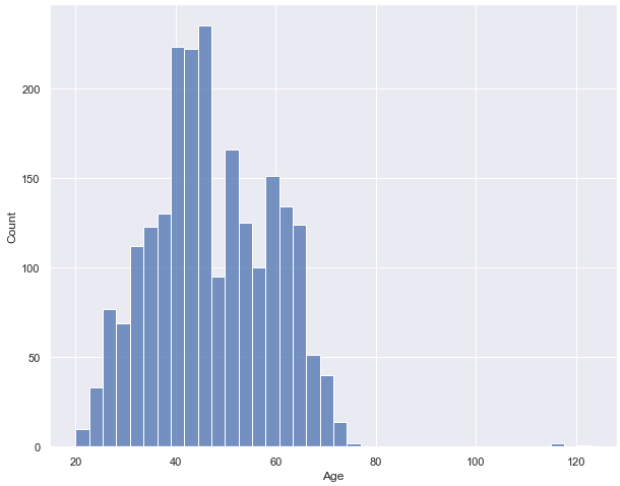
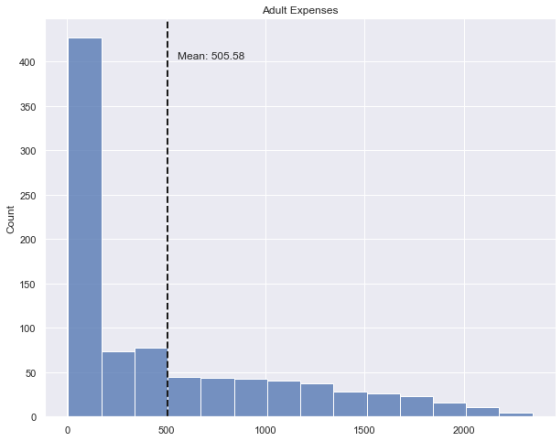
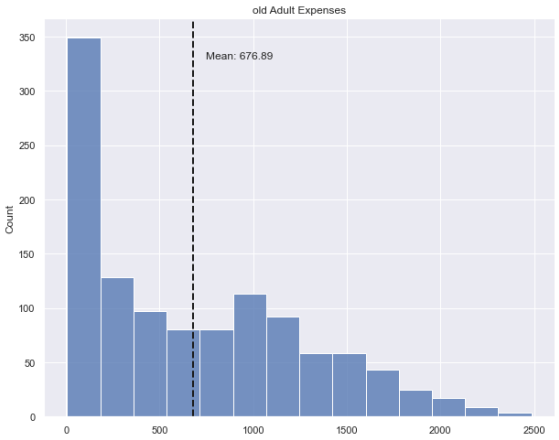


Figure 8: Age Histogram

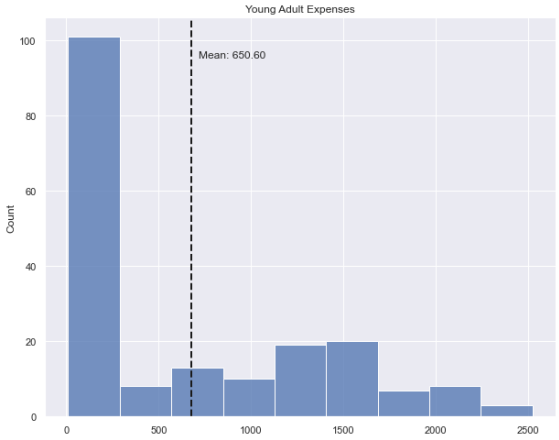
We have three different age categories , each of their distribution is shown below



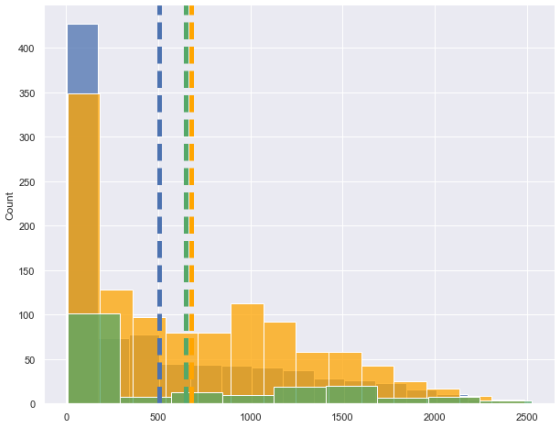
(a) Adult Expenses



(b) Old Adult Expenses



(c) Young Adult Expenses



(d) all age group Expenses

Here we can see that there are 3 different groups, to test the means of these different

groups we can use non parametric test such as kruskal method. The p value obtained here is 0.001 which is less than the significant level which means that there is a difference in means. In order to obtain which particular pair has a difference we further do post hoc analysis which tells about pair wise distinction. The following table shows the p value between all age categories.

	Old Adult	Adult	Young Adult
Old Adult	1.000000e+00	1.143050e-09	0.854224
Adult	1.143050e-09	1.000000e+00	0.010071
Young Adult	8.542240e-01	1.007078e-02	1.000000

Figure 10: posthoc scheffe analysis

We can observe that the pairs (old adult, adult) and (adult, young adult) have p values less than significant level which means that their means are different.

4.6 Mode of Purchase vs Age Category

Now we will have look at on which mode customers are buying most products, the following table and graph tells that store purchases are more in all the age categories followed by web purchases.

	NumDealsPurchases	WebPurchases	CatalogPurchases	StorePurchases
Age_Cat				
Adult	2113	3347	1979	4735
Old Adult	2788	5150	3469	7176
Young Adult	303	650	514	1056

Figure 11: All mode Purchases data

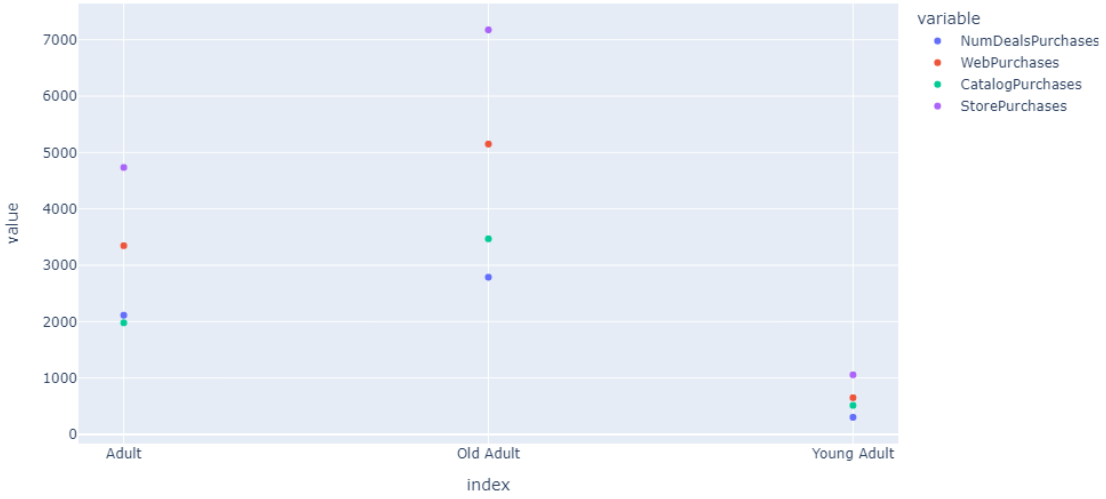


Figure 12: Scatter plot of Purchases vs Age Category

4.7 Most Sold Products

There are 6 different products in the table. Out of this, the most sold products are wines, meat products, and gold across all age categories. The below graph depicts the expenditure on spending on every age group.

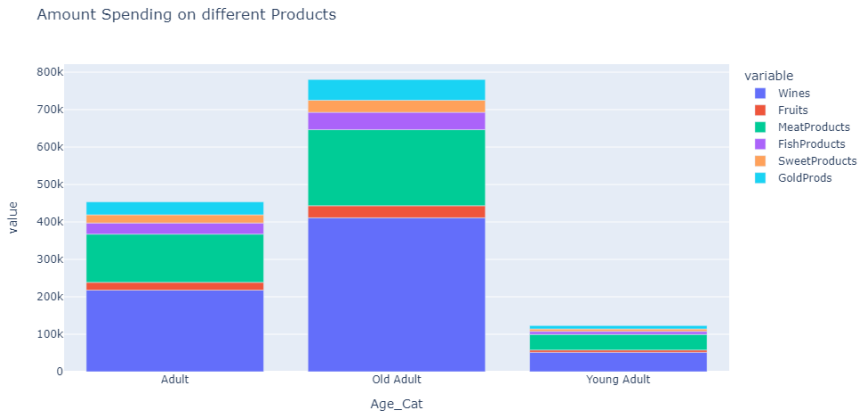
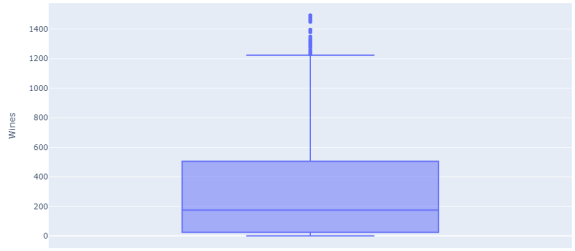


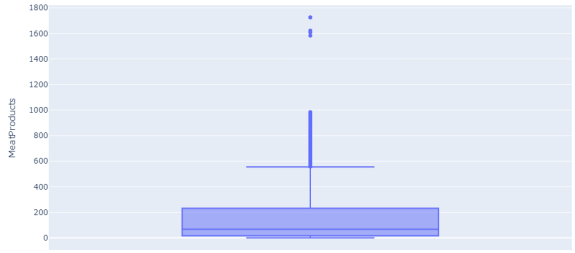
Figure 13: Expediture on every product

Again we can classify this spending on the basis of income. Most of the spending in the top products sold done by those customers whose income is more than the average income. The following graphs depicts this behaviour.

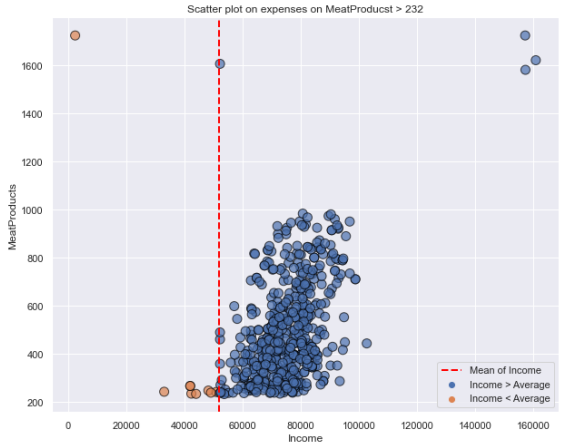


(a) Wine Expenses Box Plot

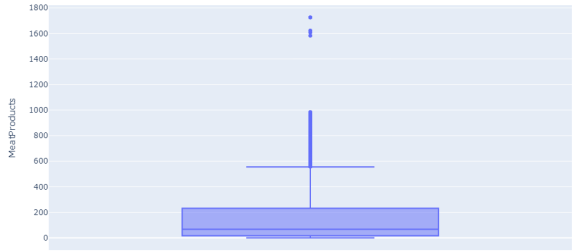
(b) income vs Wine Scatter Plot



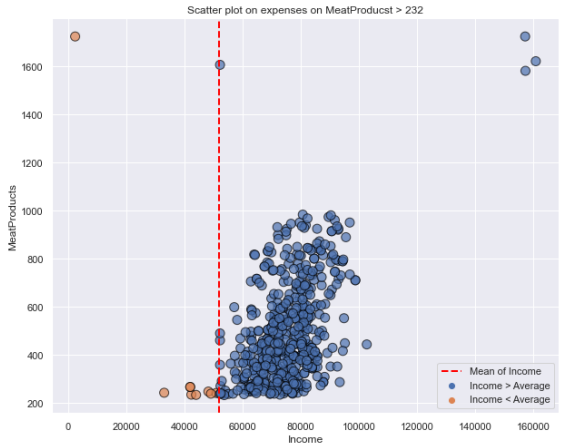
(a) Meat Expenses Box Plot



(b) income vs meat Scatter Plot



(a) Gold Expenses Box Plot



(b) income vs Gold Scatter Plot

4.8 Campaigns

They were total of 6 campaigns in which campaign 6(last one) is the most successful campaign covering 33 % customers and after that campaign 4 is the second most successful one. The following pie chart show a clear picture of campaigns division:

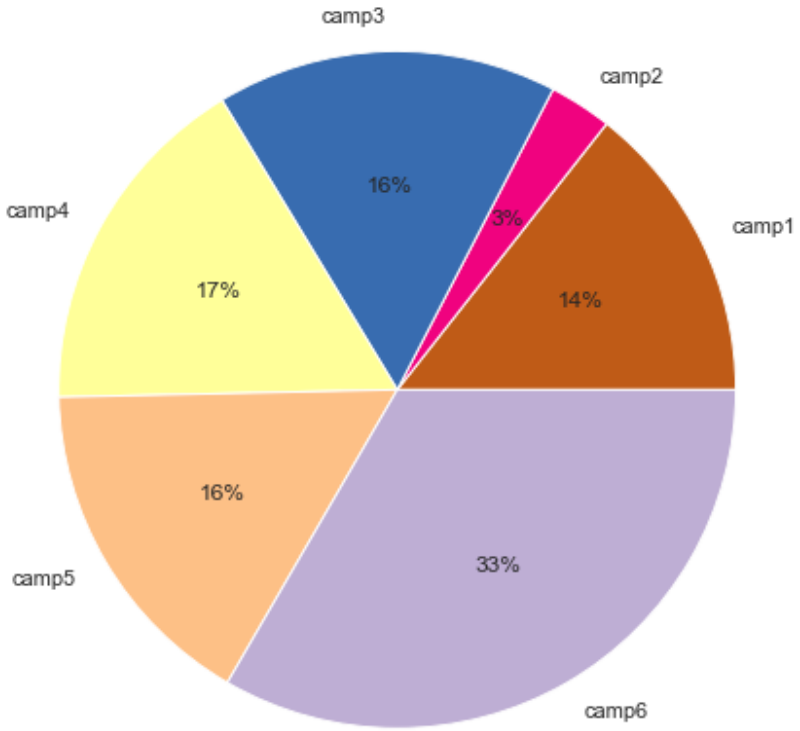


Figure 17: Campaigns Pie Chart

5 Dimensionality Reduction and Clustering Analysis

The principal component analysis is used to minimize the dimensions of the data. The columns utilized in PCA are age, marital status, income, and expenses. For further analysis, the top three eigenvalues of PCA are used. The maximum silhouette score is achieved when the number of clusters is 8. Consequently, I ran the k means procedure on n = 8 clusters. The clusters are clearly distinguished, as shown in the plot below. Although other clustering models are employed, their silhouette scores are lower than the k means.

Model	**Silhouette Score**
K means Clustering	0.60
Heirarchical Clustreing	0.58
Gaussian Mixture	0.35

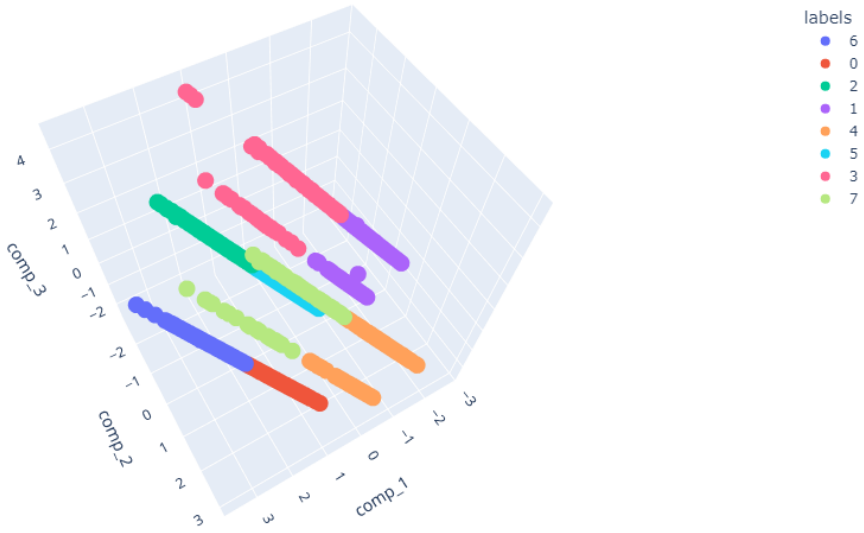


Figure 18: Kmeans Clustering of Customers using PCA

			count
labels	Age_Cat	Marital_S	
0	Old Adult	Single	229
1	Adult	Married	409
	Young Adult	Married	63
2	Adult	Married	1
		Single	1
	Old Adult	Married	325
3	Adult	Married	206
	Young Adult	Married	39
4	Adult	Single	189
	Young Adult	Single	52
5	Old Adult	Married	398
6	Old Adult	Single	199
7	Adult	Single	89
	Old Adult	Married	1
	Young Adult	Single	35

Figure 19: Clusters Formed

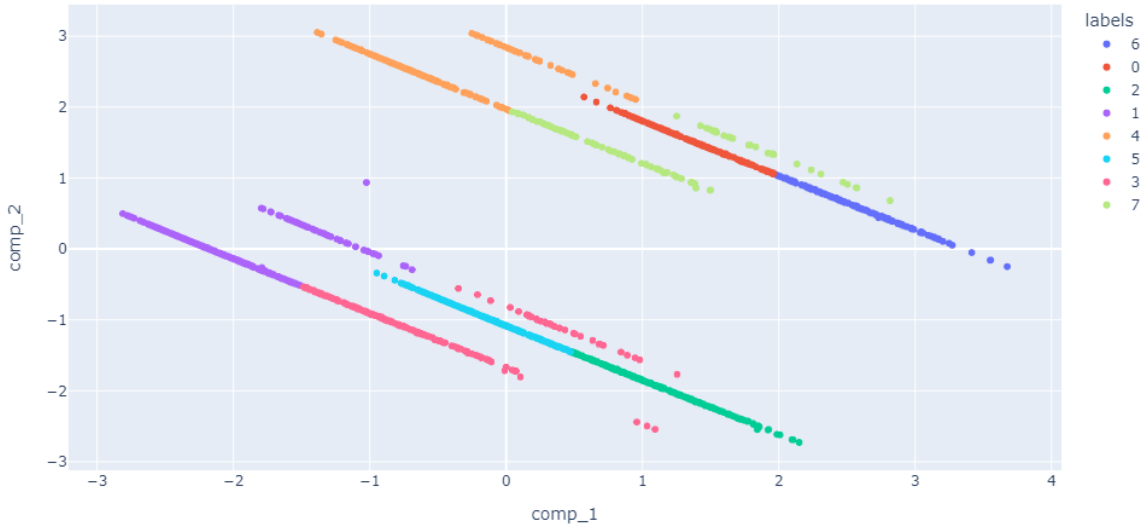


Figure 20: 2D plot of Clustering using first two components of PCA

5.1 High Spending Customers

The creation of the clusters is due to the columns used in the PCA. We can identify high-spending consumers from these groupings. Customers with classifications 2, 3, 6 and 7 are high spenders across all age groups. This can be further separated into married and single categories. Compared to other customers in this high-income bracket, older single adults are the ones who spend the most. These customers have a lot of purchasing power. For customer acquisition, the business can use unique offers targeted at these customers via email.

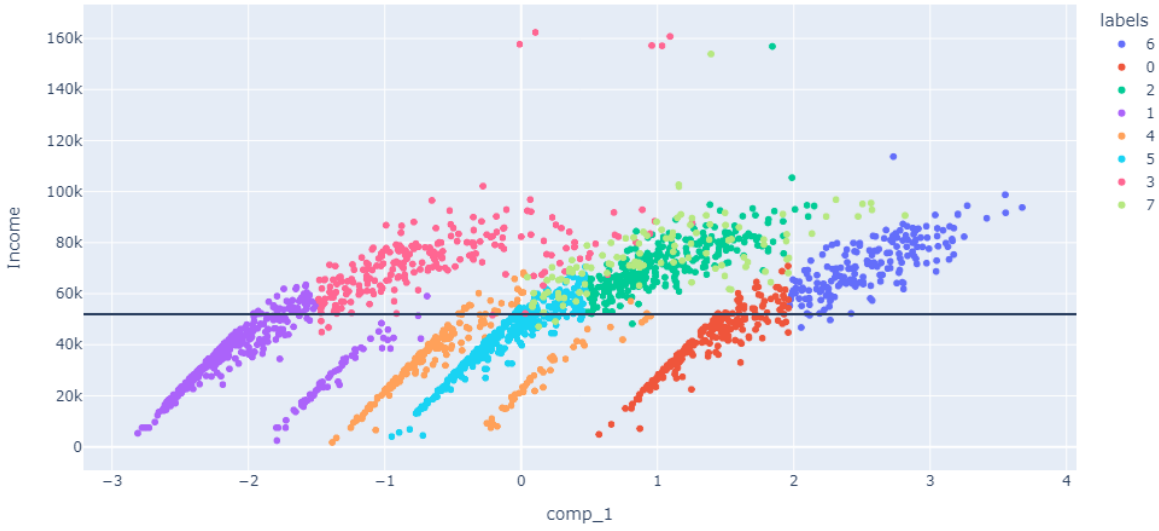


Figure 21: Income Clustering

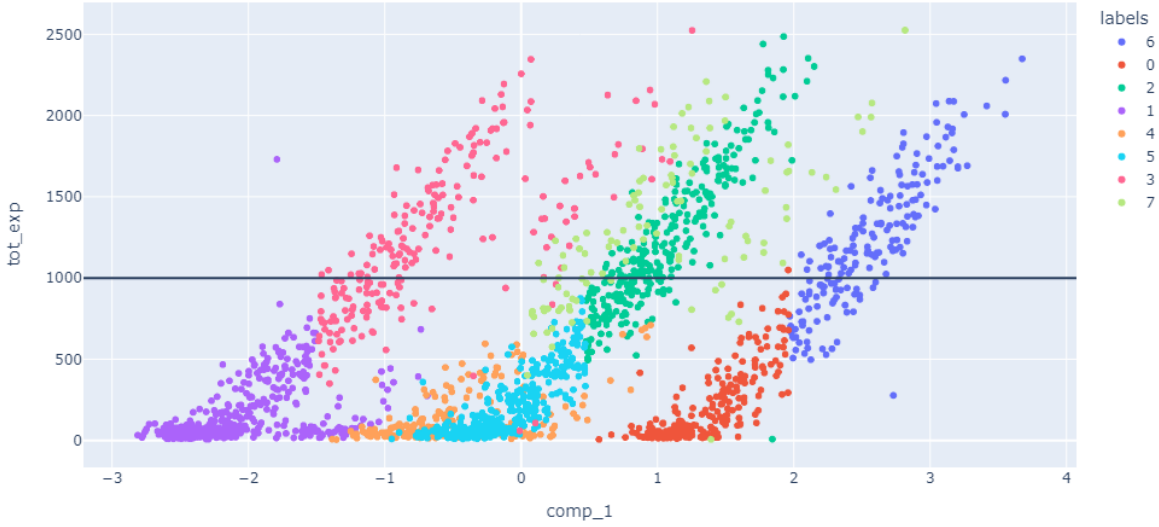


Figure 22: Expenses Clustering

5.2 Low Spending Customers

The remaining labels, 0, 1, 4, and 5, are for clients with lower-than-average incomes.These people spending score is low.

5.3 Age and Marital Status in Clustering

Here, it is clear that the entire clustering model divides into two groups based on wealth, but clustering based on other characteristics, such as age and marital status, influences the offers that should be made and how to build lasting relationships with all of these potential buyers. The below plots depicts how the expenses are with respect to age and the marital status.

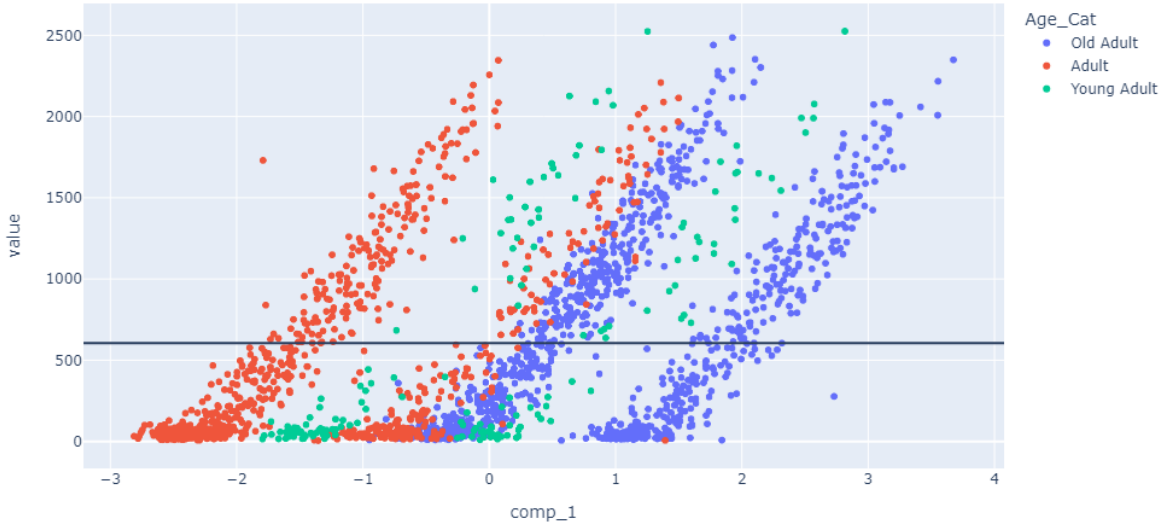


Figure 23: Clustering wrt Age Category

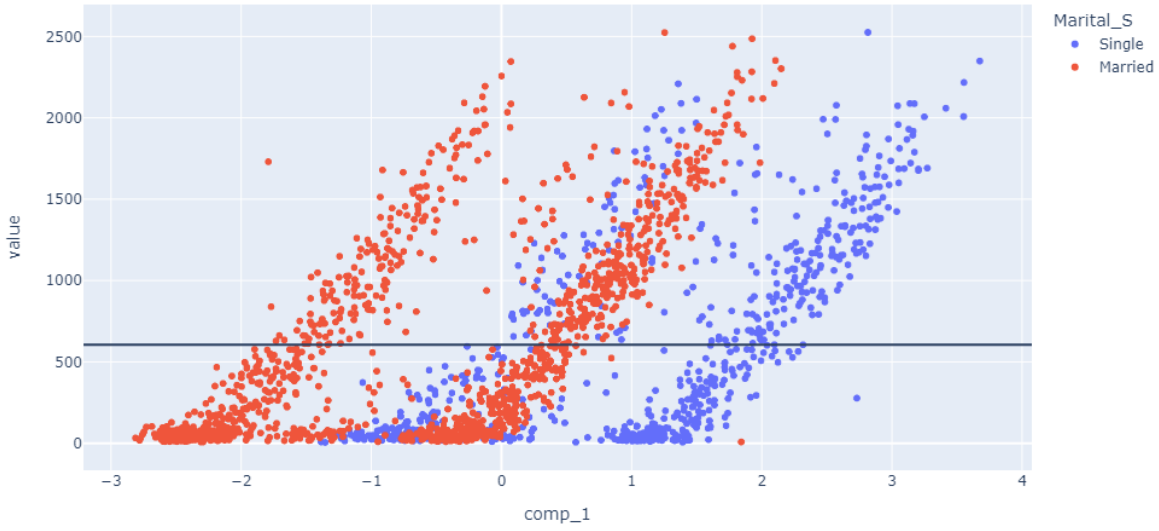


Figure 24: Clustering wrt Marital Status

6 Apriori Algorithm

6.1 Implementation

The idea here is that using clusters from K-means, we can try to figure out what the top characteristics are in some of the clusters. To do this, we categorize some of the columns and use them in the apriori algorithm for some of the clusters. Age, Marital Status, and Education are already categorical in the data; we can try to add some more features such as the number of children and teenagers in the household, the frequency of web visits (more than average or less than average), and income (more than average or less than average).These columns are made categorical, and apriori is applied.

6.2 Results

	Left Hand Side	Right Hand Side	Support	Confidence	Lift
0	3_person	Married	0.008	1.0	125.0
1	3_person	Old Adult	0.008	1.0	125.0
2	Married	Old Adult	0.008	1.0	125.0
3	3_person	Married	0.008	1.0	125.0
4	3_person	Married	0.008	1.0	125.0
5	3_person	level 2 Education	0.008	1.0	125.0
6	Married	level 2 Education	0.008	1.0	125.0
7	3_person	Married	0.008	1.0	125.0

(a) Cluster 2

	Left Hand Side	Right Hand Side	Support	Confidence	Lift
0	Adult	Single	0.003058	0.5	163.500000
1	Adult	Single	0.003058	0.5	163.500000
2	Adult	Single	0.003058	0.5	163.500000
3	Single	no_person	0.003058	1.0	3.892857
4	Married	1_person	0.003058	1.0	4.671429
5	Adult	Single	0.003058	0.5	163.500000

(b) cluster 7

	Left Hand Side	Right Hand Side	Support	Confidence	Lift
0	level 2 Education	no_person	0.033195	0.571429	4.590476
1	level 2 Education	Single	0.033195	0.571429	4.590476

(c) Cluster 4

7 Overall Conclusions

1. The top most products sold are wines, meat and gold.
2. The most successful campaign was camp 6.
3. K-means clustering is heavily linked to the income levels, customers with more income are spending more than average on all of the top customers.
4. Age and marital status helps in determining type of advertisements to be given to the customers.
5. Clusters 2,3, 6, 7 are high spending groups in which most of people earn more than the average income.The rest are low spending groups.
6. The top traits of cluster 2 are Adult and Single. These individuals have very high spending power. The top traits of cluster 7 are married and having 3 kids or teen, or combination of both , even these groups also have high spending potential. Managers can make reliable strategies focusing on these potential buyers.