

# Understanding Our Customers

## Customer Personality Segmentation Making Sense of Unstructured Data

28 June, 2025

# Contents / Agenda

- Business Problem Overview
- Data Overview
- EDA - Univariate Analysis
- EDA - Multivariate Analysis
- K Means Clustering
- Cluster Profiling and Analysis
- Business Recommendations

# Business Problem Overview

- Our business's customer base is rapidly growing
- Targeted and personalized approaches is essential for sustaining a competitive edge
- Understanding customers personalities, lifestyle and purchasing habits can help us what deliver to our customers
- Our objective and aim should be to:
  1. develop personalized marketing campaigns
  2. Create retention strategies for high value customers
  3. Optimize resource allocation, such as inventory management, pricing strategies and store layouts.

# Data Overview (continued next slide)

First 8 Customers

	ID	Year_Birth	Education	Marital_Status
0	5524	1957	Graduation	Single
1	2174	1954	Graduation	Single
2	4141	1965	Graduation	Together
3	6182	1984	Graduation	Together
4	5324	1981	PhD	Married
5	7446	1967	Master	Together
6	965	1971	Graduation	Divorced
7	6177	1985	PhD	Married

Last 8 Customers

	ID	Year_Birth	Education	Marital_Status
2232	8080	1986	Graduation	Single
2233	9432	1977	Graduation	Together
2234	8372	1974	Graduation	Married
2235	10870	1967	Graduation	Married
2236	4001	1946	PhD	Together
2237	7270	1981	Graduation	Divorced
2238	8235	1956	Master	Together
2239	9405	1954	PhD	Married

Total Customers = 2240

Data Categories = 30

# Data Overview

Our collected data types range from:-

## Customer Information

- ID
- Birth Year
- Education
- Martial\_Status
- Income
- Kid home
- Teen home
- Customer enrollment date
- Last purchase
- Complaint within last two years

## Spending Information

- Amount spent on wine
- Amount spent on fruits
- Amount spent on meat
- Amount spent on fish
- Amount spent on sweets
- Amount spent on gold products

## Purchase and Campaign Interaction

- Purchases made during discount
- Response to campaign 1
- Response to campaign 2
- Response to campaign 3
- Response to campaign 3
- Response to campaign 4
- Response to campaign 5

## Shopping Behaviour

- Web Purchases
- Purchases made from catalogue
- Store purchases
- Company website purchases

# Statistical Summary (Continued next slide)

	count	mean	std	min	25%	50%	75%	max
ID	2240.0	5592.159821	3246.662198	0.0	2828.25	5458.5	8427.75	11191.0
Year_Birth	2240.0	1968.805804	11.984069	1893.0	1959.00	1970.0	1977.00	1996.0
Income	2216.0	52247.251354	25173.076661	1730.0	35303.00	51381.5	68522.00	666666.0
Kidhome	2240.0	0.444196	0.538398	0.0	0.00	0.0	1.00	2.0
Teenhome	2240.0	0.506250	0.544538	0.0	0.00	0.0	1.00	2.0
Recency	2240.0	49.109375	28.962453	0.0	24.00	49.0	74.00	99.0
MntWines	2240.0	303.935714	336.597393	0.0	23.75	173.5	504.25	1493.0
MntFruits	2240.0	26.302232	39.773434	0.0	1.00	8.0	33.00	199.0
MntMeatProducts	2240.0	166.950000	225.715373	0.0	16.00	67.0	232.00	1725.0
MntFishProducts	2240.0	37.525446	54.628979	0.0	3.00	12.0	50.00	259.0
MntSweetProducts	2240.0	27.062946	41.280498	0.0	1.00	8.0	33.00	263.0
MntGoldProds	2240.0	44.021875	52.167439	0.0	9.00	24.0	56.00	362.0
NumDealsPurchases	2240.0	2.325000	1.932238	0.0	1.00	2.0	3.00	15.0
NumWebPurchases	2240.0	4.084821	2.778714	0.0	2.00	4.0	6.00	27.0

# Statistical Summary

<b>NumWebVisitsMonth</b>	2240.0	5.316518	2.426645	0.0	3.00	6.0	7.00	20.0
<b>AcceptedCmp3</b>	2240.0	0.072768	0.259813	0.0	0.00	0.0	0.00	1.0
<b>AcceptedCmp4</b>	2240.0	0.074554	0.262728	0.0	0.00	0.0	0.00	1.0
<b>AcceptedCmp5</b>	2240.0	0.072768	0.259813	0.0	0.00	0.0	0.00	1.0
<b>AcceptedCmp1</b>	2240.0	0.064286	0.245316	0.0	0.00	0.0	0.00	1.0
<b>AcceptedCmp2</b>	2240.0	0.013393	0.114976	0.0	0.00	0.0	0.00	1.0
<b>Complain</b>	2240.0	0.009375	0.096391	0.0	0.00	0.0	0.00	1.0
<b>Z_CostContact</b>	2240.0	3.000000	0.000000	3.0	3.00	3.0	3.00	3.0
<b>Z_Revenue</b>	2240.0	11.000000	0.000000	11.0	11.00	11.0	11.00	11.0
<b>Response</b>	2240.0	0.149107	0.356274	0.0	0.00	0.0	0.00	1.0

**Average Household Income is \$52247.25**

# Missing Data

- Missing data in the Income data type.
- 24 missing values in the income categories
- **Solution:** We remove those customers from the database

Missing values before treatment:

ID	0
Year_Birth	0
Education	0
Marital_Status	0
Income	24
Kidhome	0
Teenhome	0
Dt_Customer	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
Z_CostContact	0
Z_Revenue	0
Response	0
dtype:	int64

Missing values after treatment:

ID	0
Year_Birth	0
Education	0
Marital_Status	0
Income	0
Kidhome	0
Teenhome	0
Dt_Customer	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
Z_CostContact	0
Z_Revenue	0
Response	0



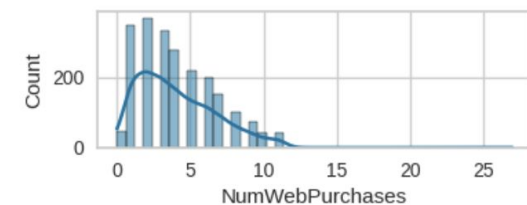
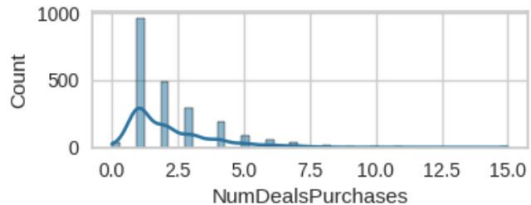
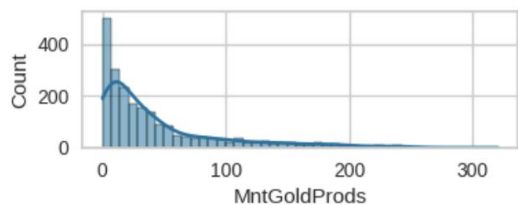
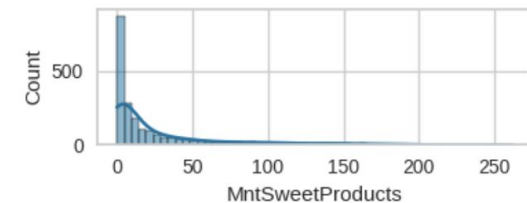
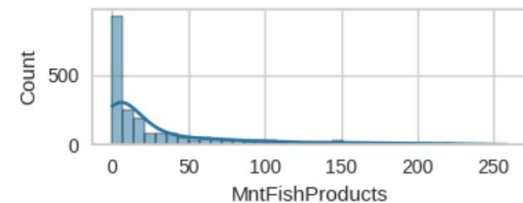
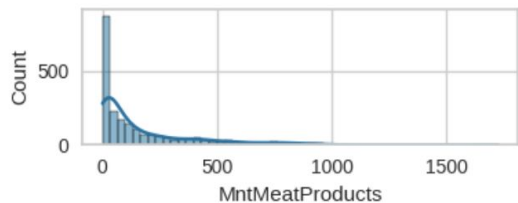
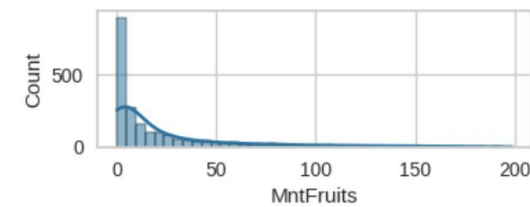
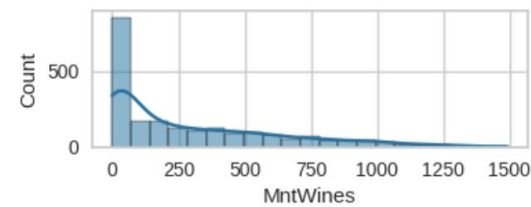
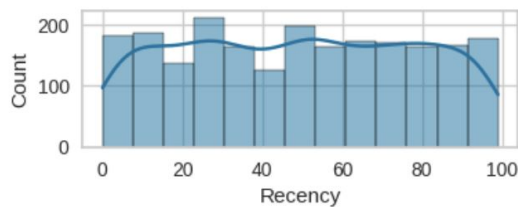
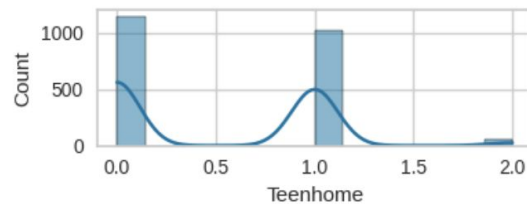
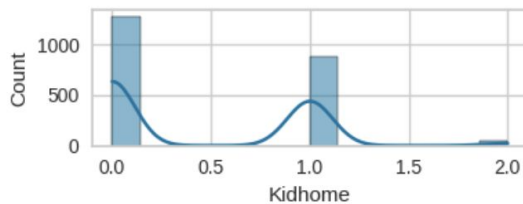
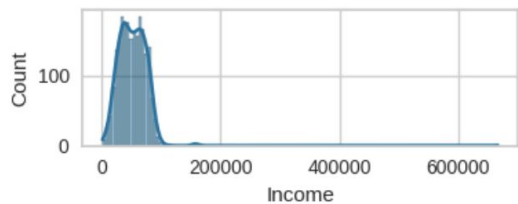
# Duplicates?

The data holds no duplicates

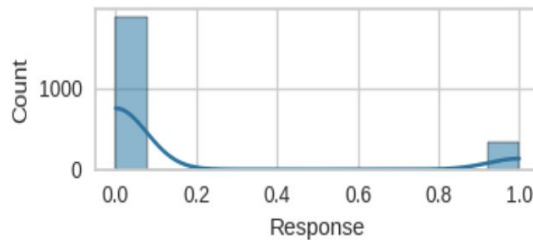
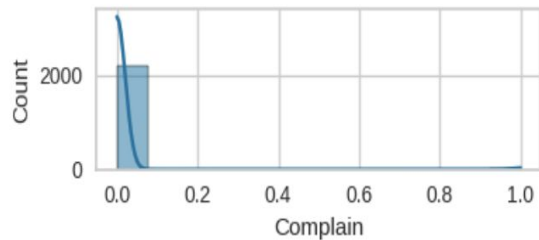
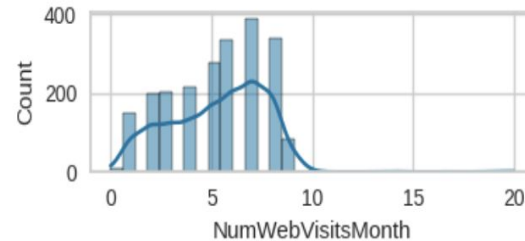
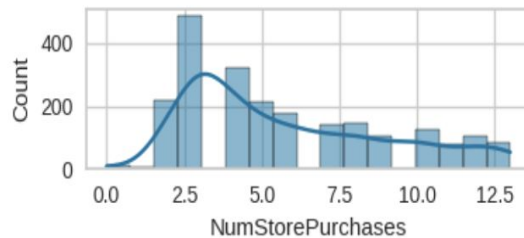
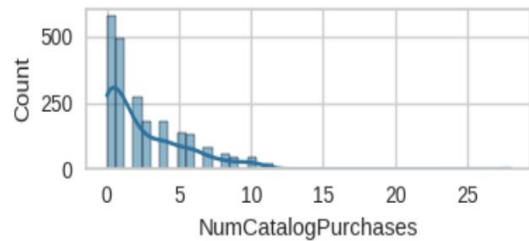
```
data.duplicated().sum()
```

```
np.int64(0)
```

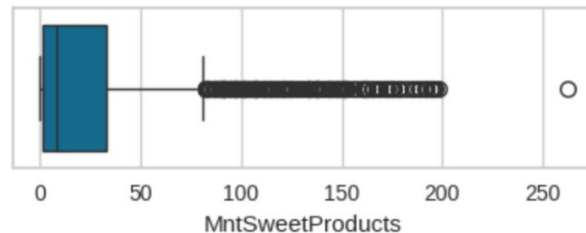
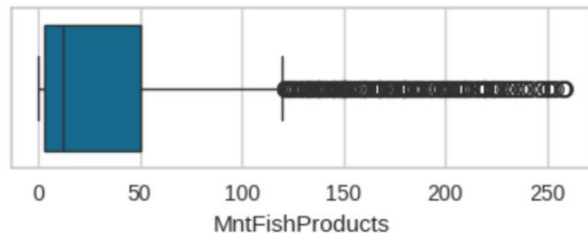
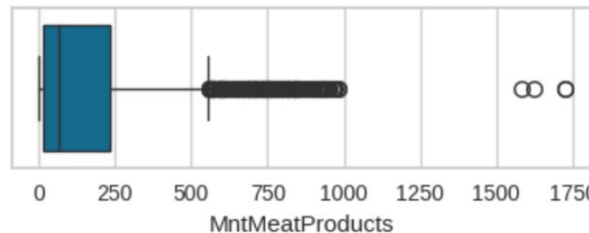
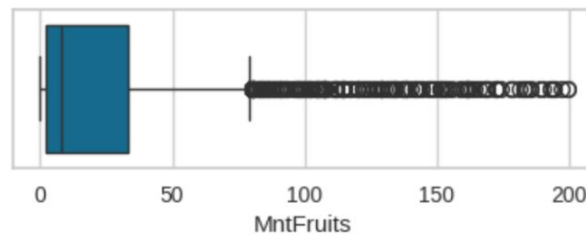
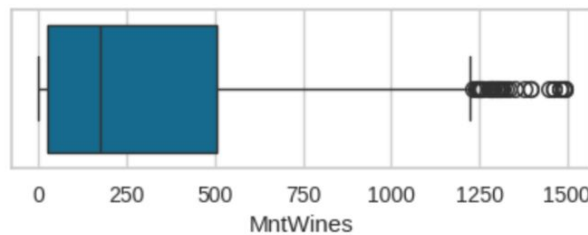
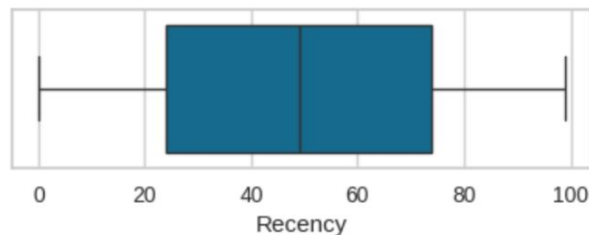
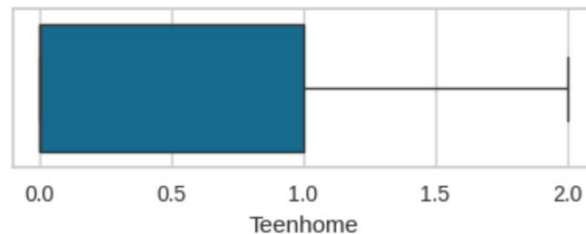
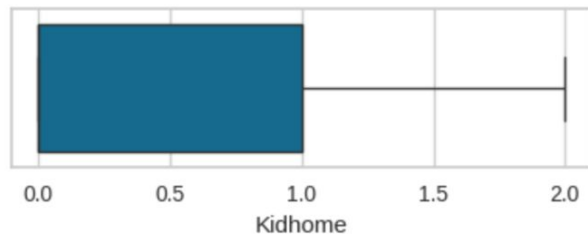
# Univariate Analysis



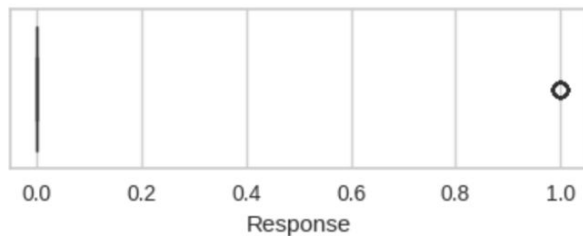
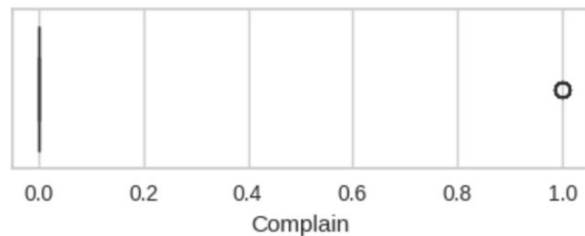
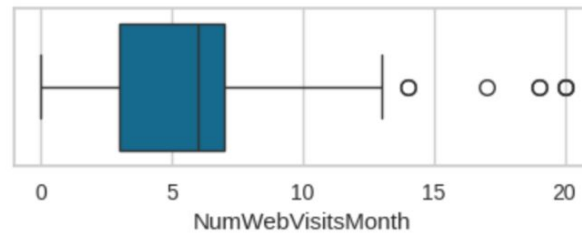
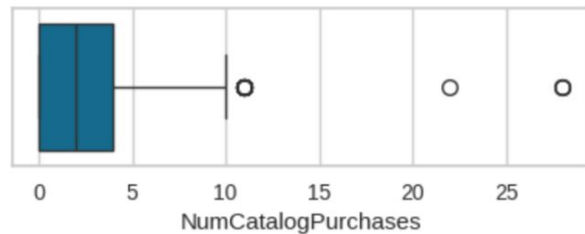
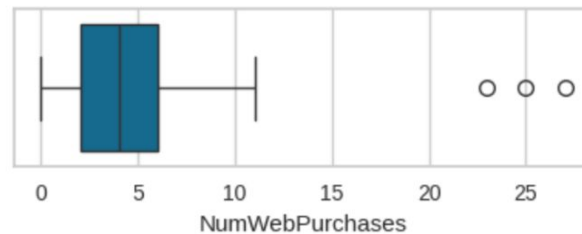
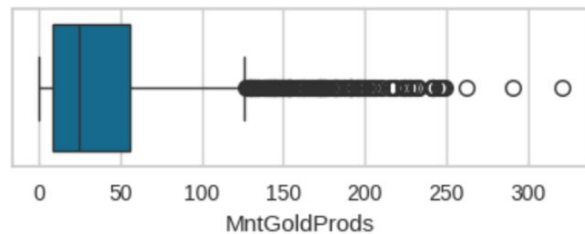
# Univariate Analysis



# Univariate Analysis



# Univariate Analysis



# Univariate Analysis

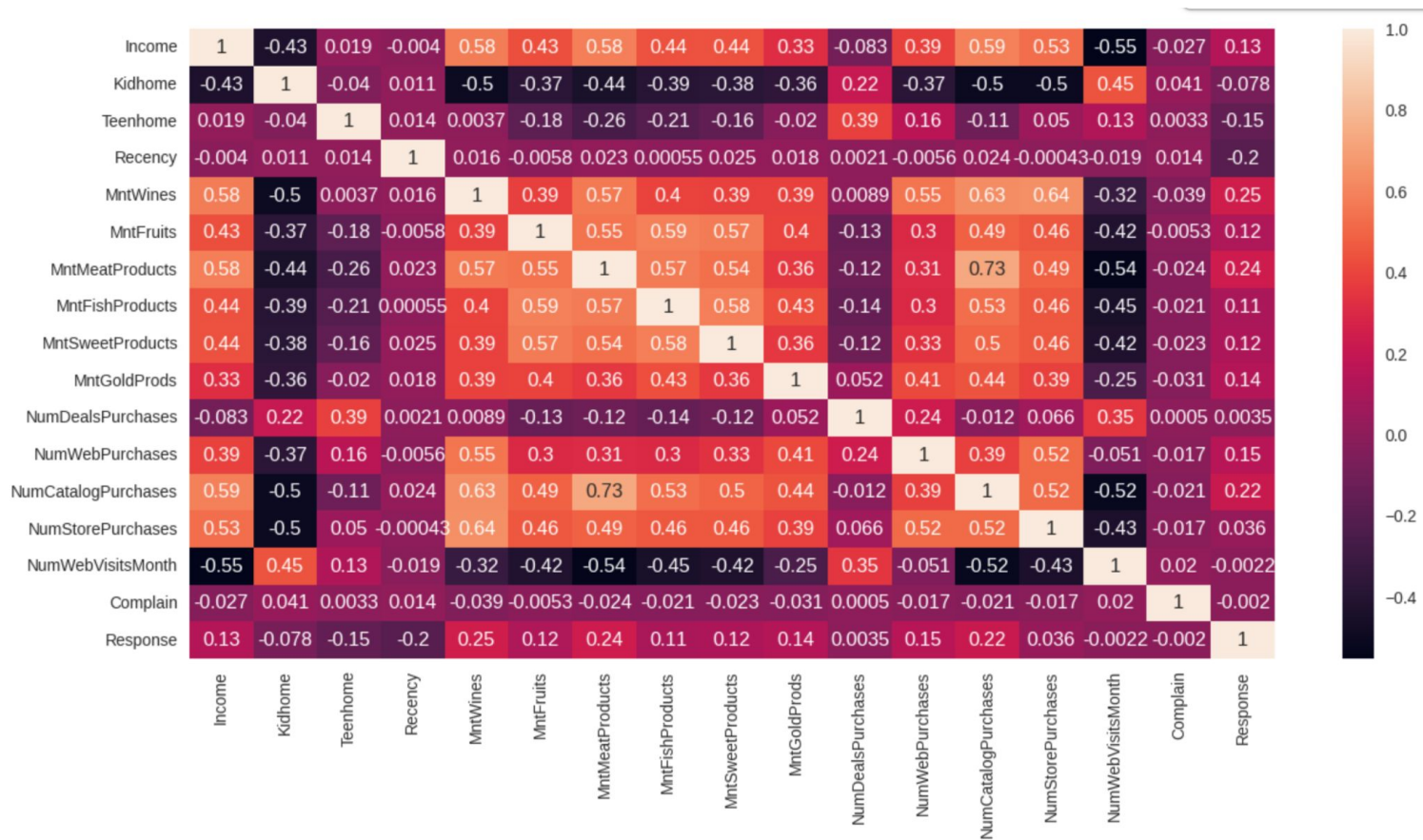
Statistics after duplicates and irrelevant categories have been removed

	count	mean	std	min	25%	50%	75%	max
Income	2216.0	52247.251354	25173.076661	1730.0	35303.0	51381.5	68522.00	666666.0
Kidhome	2216.0	0.441787	0.536896	0.0	0.0	0.0	1.00	2.0
Teenhome	2216.0	0.505415	0.544181	0.0	0.0	0.0	1.00	2.0
Recency	2216.0	49.012635	28.948352	0.0	24.0	49.0	74.00	99.0
MntWines	2216.0	305.091606	337.327920	0.0	24.0	174.5	505.00	1493.0
MntFruits	2216.0	26.356047	39.793917	0.0	2.0	8.0	33.00	199.0
MntMeatProducts	2216.0	166.995939	224.283273	0.0	16.0	68.0	232.25	1725.0
MntFishProducts	2216.0	37.637635	54.752082	0.0	3.0	12.0	50.00	259.0
MntSweetProducts	2216.0	27.028881	41.072046	0.0	1.0	8.0	33.00	262.0
MntGoldProds	2216.0	43.965253	51.815414	0.0	9.0	24.5	56.00	321.0
NumDealsPurchases	2216.0	2.323556	1.923716	0.0	1.0	2.0	3.00	15.0
NumWebPurchases	2216.0	4.085289	2.740951	0.0	2.0	4.0	6.00	27.0
NumCatalogPurchases	2216.0	2.671029	2.926734	0.0	0.0	2.0	4.00	28.0
NumStorePurchases	2216.0	5.800993	3.250785	0.0	3.0	5.0	8.00	13.0
NumWebVisitsMonth	2216.0	5.319043	2.425359	0.0	3.0	6.0	7.00	20.0
Complain	2216.0	0.009477	0.096907	0.0	0.0	0.0	0.00	1.0
Response	2216.0	0.150271	0.357417	0.0	0.0	0.0	0.00	1.0

# Observations

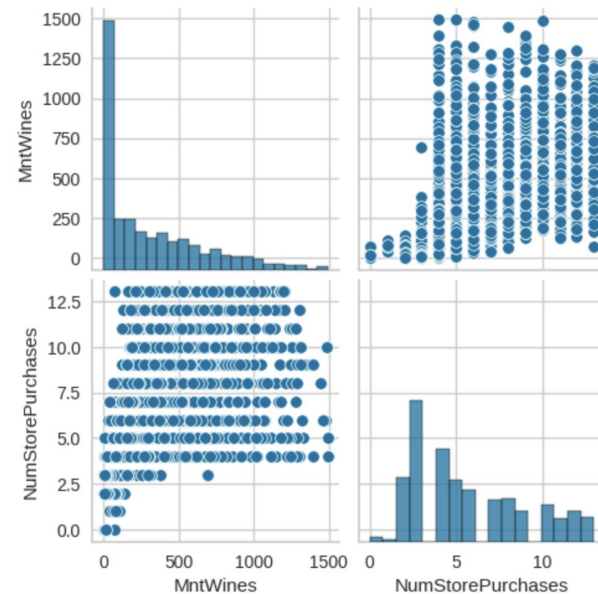
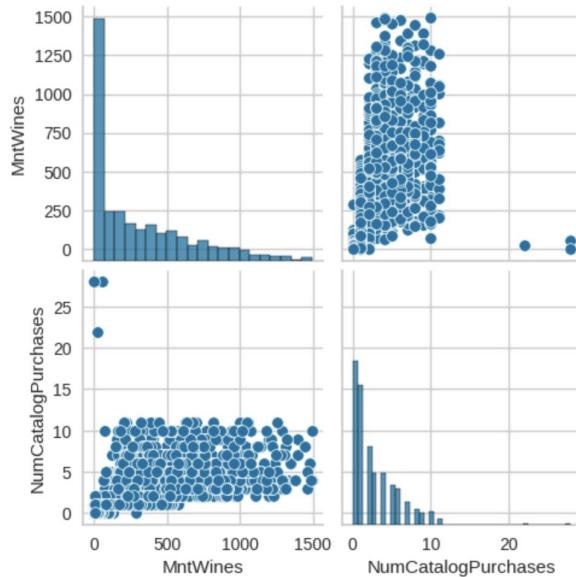
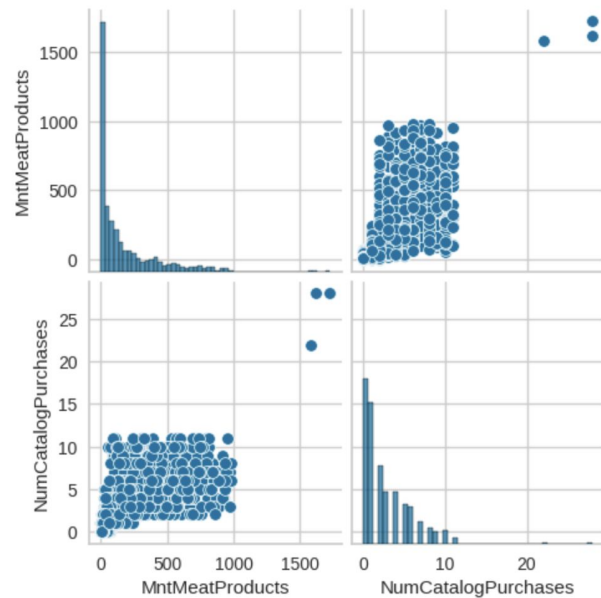
- Income is skewed. The difference between min, mean and max is huge.
- Spending on wines is higher than other food related categories
- Mean web purchases is very close to mean web visits.
- Not a lot of complaints

# Multivariate Analysis





# Multivariate Analysis

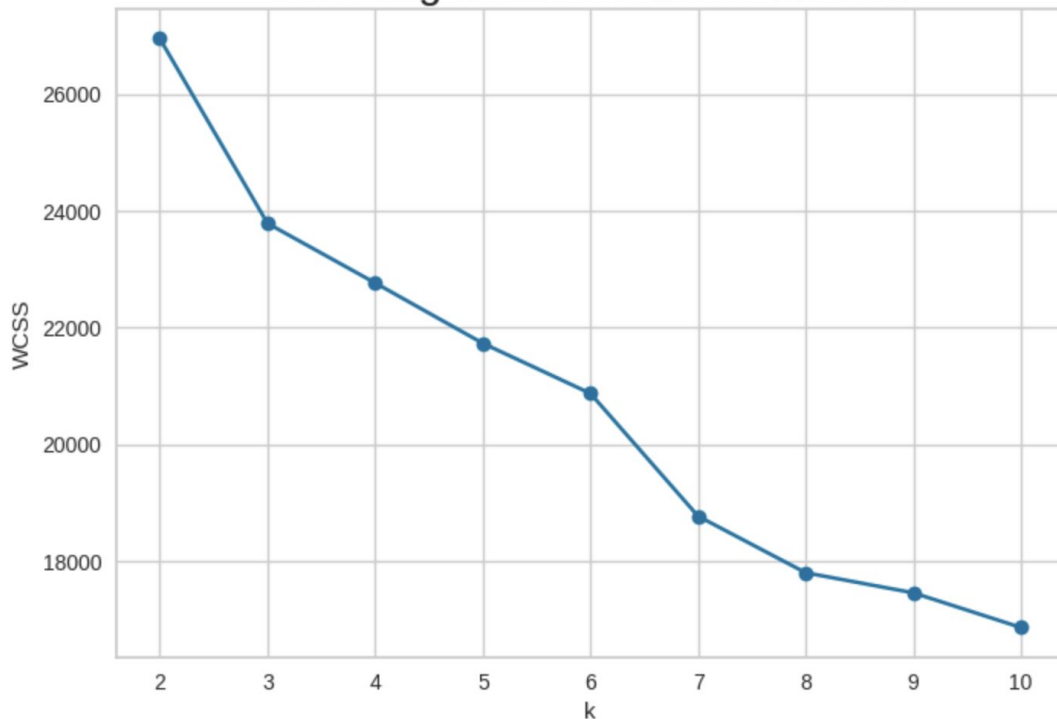


# Observations

- We made correlation table and many pairs have a decent correlation.
  - e.g Sweet Products with Fish, Sweet and Meat products have more than 0.5 correlation
- However, only 3 products have correlation higher than 0.6 which were shown in the pairplots.
  - Amount of Wines vs Number of Catalogue Purchases
  - Amount of Wines vs Number of Store Purchases
  - Amount of Meat Products vs Number of Catalog Purchases

# K Means Clustering:

Selecting k with the Elbow Method

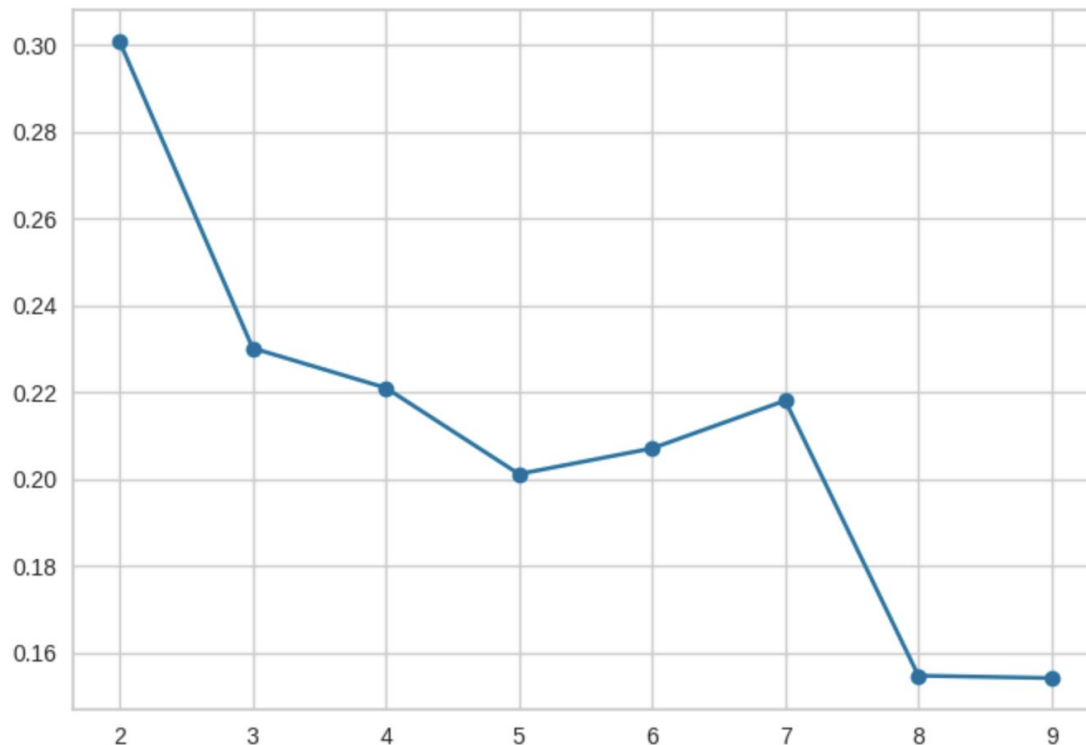


Number of Clusters: 2	WCSS: 26946.202862559545
Number of Clusters: 3	WCSS: 23785.001200435698
Number of Clusters: 4	WCSS: 22764.776258492882
Number of Clusters: 5	WCSS: 21731.597771380926
Number of Clusters: 6	WCSS: 20872.189379414514
Number of Clusters: 7	WCSS: 18771.043023962407
Number of Clusters: 8	WCSS: 17802.70907428246
Number of Clusters: 9	WCSS: 17459.046204021768
Number of Clusters: 10	WCSS: 16864.67859689163

**After looking at the graph**  
4 clusters is the appropriate amount

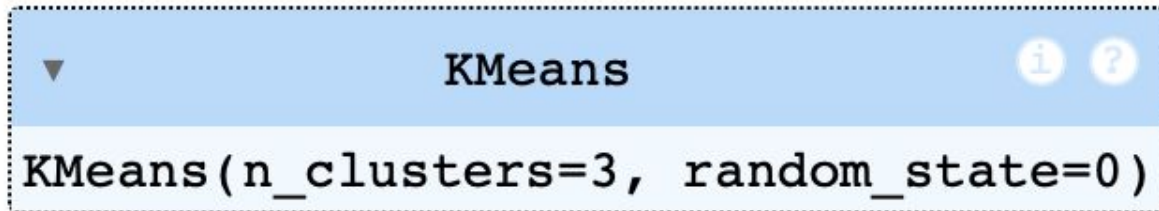
# K-means clustering

- Silhouette score suggests that the right number is  $k=3$ .
- Our initial mark was on  $k=4$  but  $k=3$  because the silhouette score gives better insight into the cluster separation



# How much time does it take for the model to fit the data?

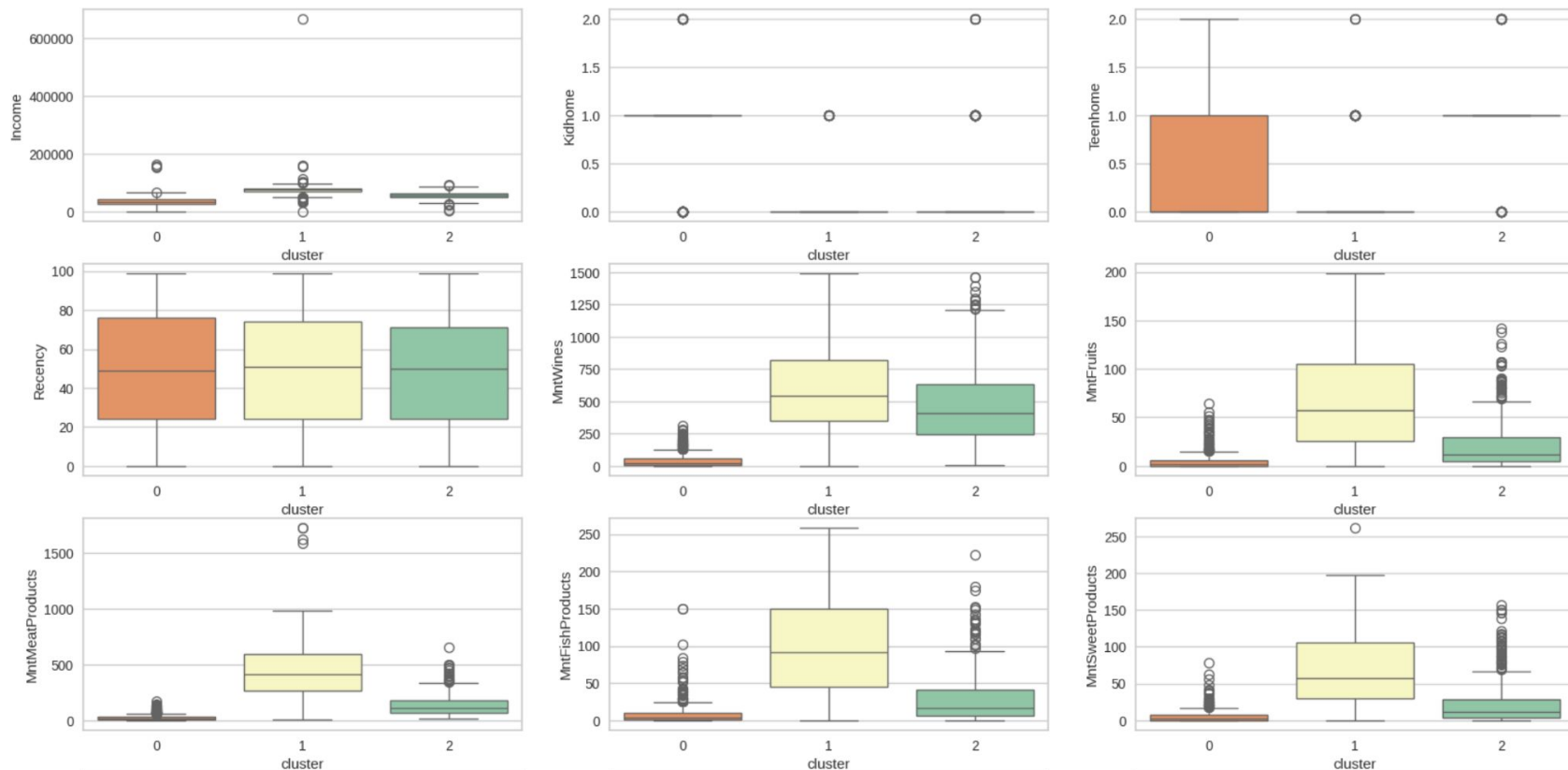
```
CPU times: user 6.4 ms, sys: 0 ns, total: 6.4 ms  
Wall time: 6.07 ms
```

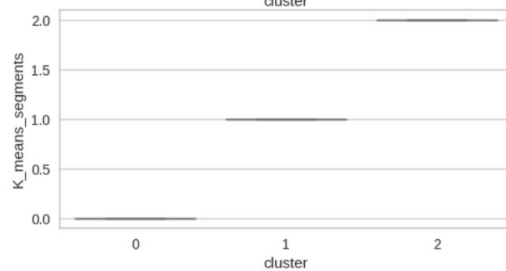
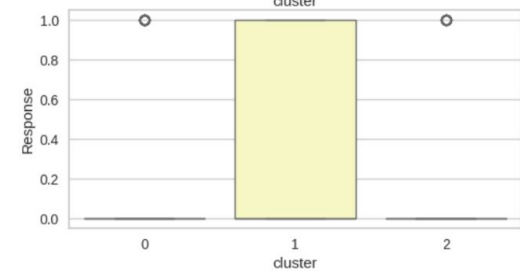
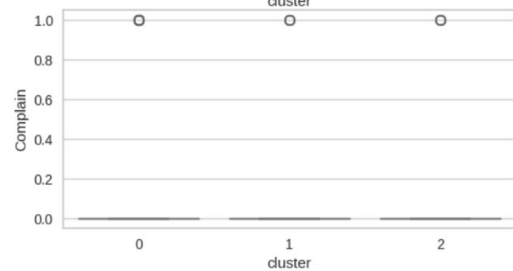
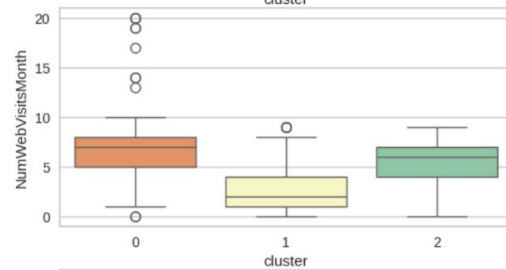
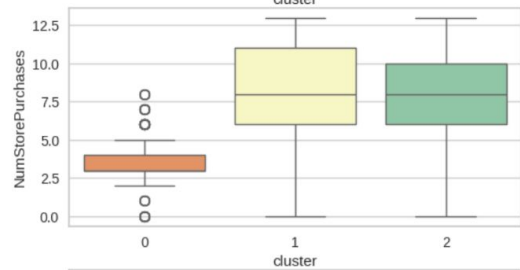
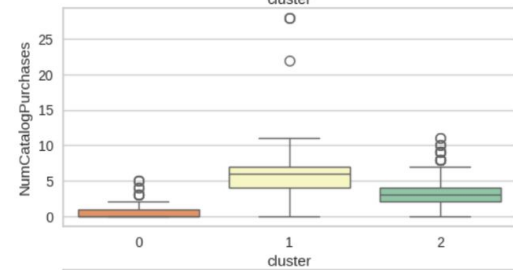
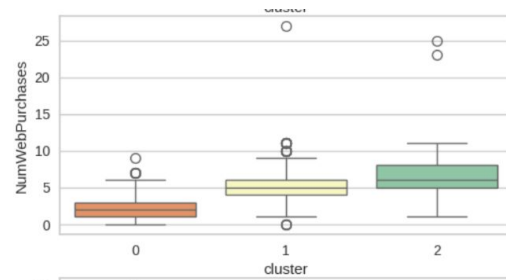
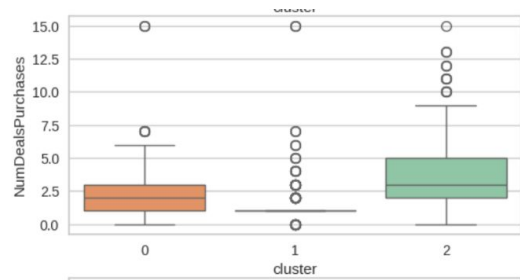
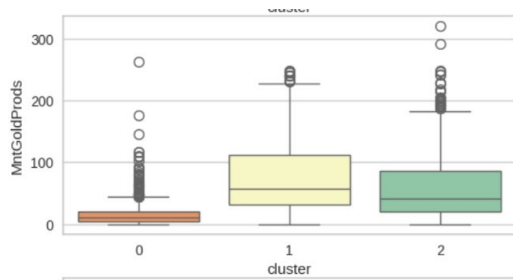
A light blue rectangular box with a dotted border. The top part has a light blue background and contains a downward-pointing triangle icon on the left, the text 'KMeans' in the center, and two circular icons (one with an 'i' and one with a '?') on the right. The bottom part has a white background and contains the text 'KMeans(n\_clusters=3, random\_state=0)' in a monospaced font.

```
▼ KMeans ⓘ ?  
KMeans(n_clusters=3, random_state=0)
```

KMeans model with 3 clusters completed in **6:07 seconds**

# Cluster Profiling and analysis:



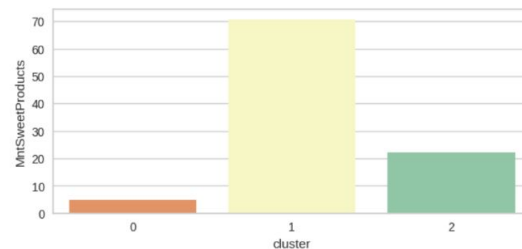
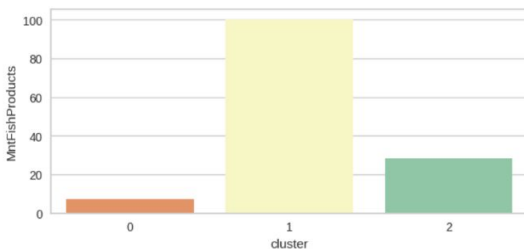
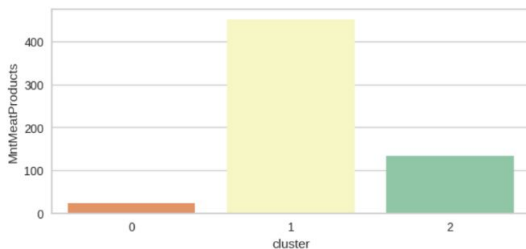
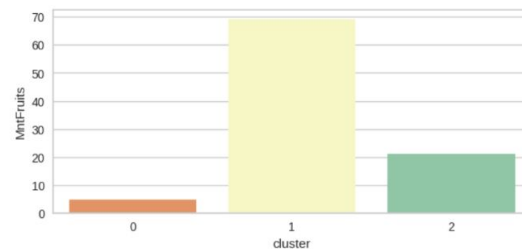
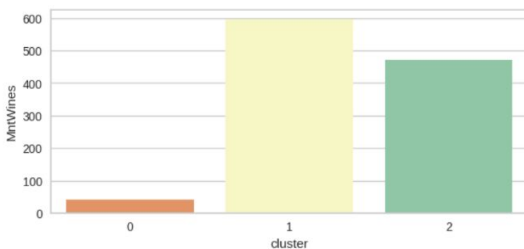
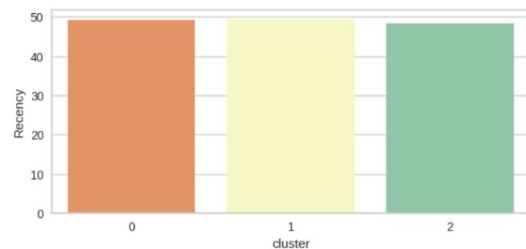
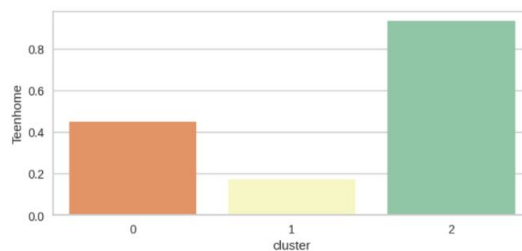
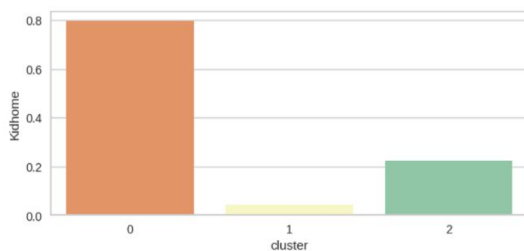
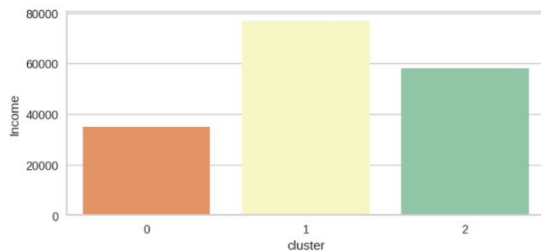


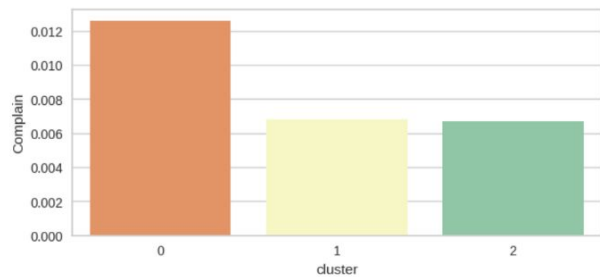
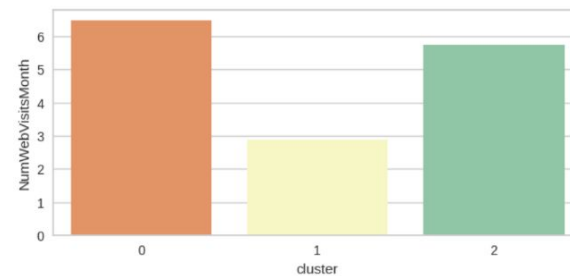
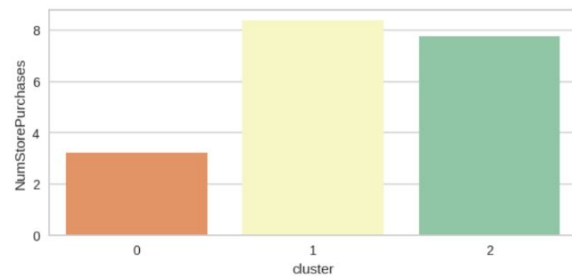
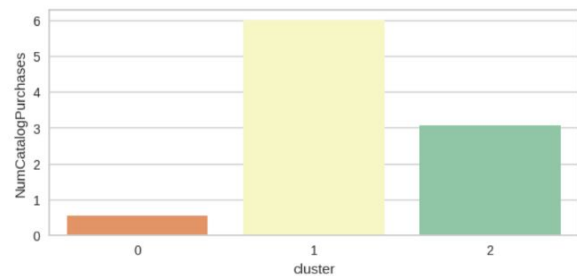
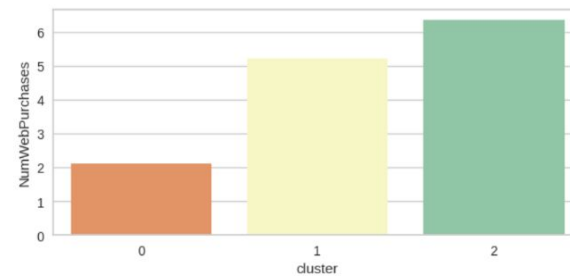
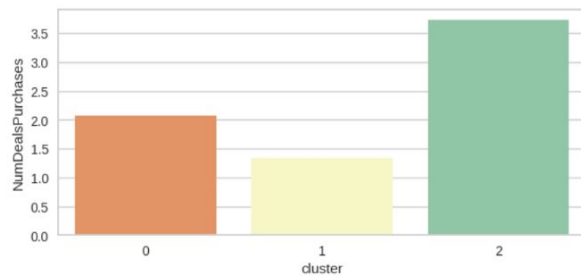
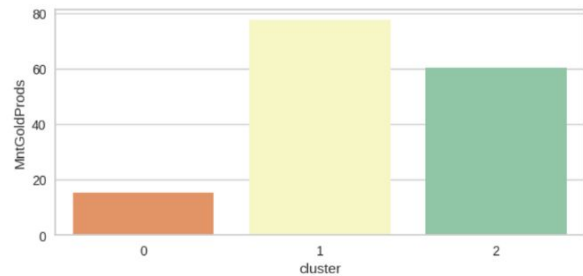
# Cluster Profiling and Analysis - Observations

- Observing these clusters give a few suggestions
- Cluster 0 :-
  - This customer segment is very disengaged, a plausible reason could be the cost. Maybe the customer segment does not have very high spending budget.
- Cluster 1:-
  - This being the high value customer segment. They seem loyal, engaged and profitable.
- Cluster 2:-
  - These are the customers that could be potential customers. Even though they do not show full engagement, they show product interest.



# Clustering Profile and Analysis





# Cluster Profiling and Analysis - Observations

- Observing these clusters give a few suggestions
- Cluster 0 :-
  - Some of the key traits being homes with children and low income among all clusters are there. This shows disengaged and budget sensitive families. They are showing little online behaviour.
- Cluster 1:-
  - This cluster has some higher average income with over 70,000. It has high store and online purchases. These customers are consistent
- Cluster 2:-
  - This cluster has moderate high teenagers presence and lower kid home and tend to be very responsive to deals. They also have a lot of online presence but are not the highest level spenders.

# Business Recommendations:

Based on our overall observations:

- Our customer base is spread across 3 different clusters, lower income with least engagement, moderate income earner and highest income earners.

## **Low Income Earners/Budget conscious families**

- Because they don't respond to campaigns, avoid spending too much on marketing outreach to them.
- Because they tend have children at home, the business could offer family bundles that are not expensive

# Business recommendation

## Moderate income earner

- Because they respond well to deals and have a high online presence, offer online sales and discounts. Offer deals and discounts that are more tailored for their personal experience like offer discount coupons.
- Because these customers are high teen homes, offer discounts on items that might interest teens or children

## High income earners

- Because they are statistically the most profitable, offer loyalty programs to ensure they are constantly returning.
- Because they are the highest spending customers, target future campaigns towards to them keep attracting customers of similar spending habits and households.
- Because they are high earners, offer luxury/exclusive products because they are products they can afford and interests them.

# Thank you

By Taha Asim



**Happy Learning !**

