

SC1015 Mini Project

LAB GROUP: DSF3

LIM YAN KAI (U2121179C)

SEAH JUN SHENG (U2122264F)

ISAAC HONG ZHANG JIE (U2120148D)



The Story of Mall X



Practical
MOTIVATION



Sample
COLLECTION



Practical
MOTIVATION



Sample
COLLECTION



Practical
MOTIVATION



Sample
COLLECTION



Customer Segmentation

Practical
MOTIVATION



Sample
COLLECTION





Key Questions to Consider

1. What products should be advertised?
2. Which purchasing platform (e.g. physical stores, website) to advertise the chosen products?
3. Should there be any promotions as part of the advertisement?

Our Dataset





RODOLFO SALDANHA · UPDATED 2 YEARS AGO

Marketing Campaign

Boost the profit of a marketing campaign

kaggle

Our Dataset

Size of dataset: 2240

#	Year_Birth	Education	Marital_Status	# Income	# Kidhome	# MntWines	# NumFruits	# NumWebP... 8
1957	Graduation	Single	58138	0	635	88		
1954	Graduation	Single	46344	1	11	1		1
1965	Graduation	Together	71613	0	426	49		8
1984	Graduation	Together	26646	1	11	4		2
1981	PhD	Married	58293	1	173	43		5
1967	Master	Together	62513	0	528	42		6
1971	Graduation	Divorced	55635	0	235	65		7
1985	PhD	Married	33454	1	76	10		4
1974	PhD	Together	38351	1	14	8		3
1959	PhD	Together	5648	1	28	8		1
1983	Graduation	Married		1	5	5		1
1976	Basic	Married	7500	0	6	16		2
1959	Graduation	Divorced	63633	0	194	61		3
1962	Master	Divorced	59354	1	233	2		6
1987	Graduation	Married	17323	0	3	14		1
1946	PhD	Single	82808	0	1866	22		7



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
- Kidhome: Number of children in customer's household
- Teenhome: Number of teenagers in customer's household
- Dt_Customer: Date of customer's enrollment with the company
- Recency: Number of days since customer's last purchase
- Complain: 1 if the customer complained in the last 2 years, 0 otherwise
- Products

Products

- MntWines: Amount spent on wine in last 2 years
- MntFruits: Amount spent on fruits in last 2 years
- MntMeatProducts: 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

Our Variables

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

- NumWebPurchases: Number of purchases made through the company's website
- NumCatalogPurchases: Number of purchases made using a catalogue
- NumStorePurchases: Number of purchases made directly in stores
- NumWebVisitsMonth: Number of visits to company's website in the last month

Data Preparation & Exploratory Analysis



Dropping Useless columns



Drop unneeded columns

First, we decided to drop the following columns as they were not relevant.

- ID
- NumDealsPurchases (we dropped this as this includes discounted items not from promotional campaigns)
- Complain
- Z_CostContact
- Z_Revenue
- Dt_customer
- Recency



Dropping Rows with Null Values

Shows that all rows with null values
new_df4[new_df4.isnull().any(axis=1)]

	Age	Education	Marital_Status	Income	Kidhome	Teenhome	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGold
10	31	Graduation	Married	NaN	1	0	5	5	6	0	2	
27	28	Graduation	Single	NaN	1	0	5	1	3	3	263	
43	55	PhD	Single	NaN	0	0	81	11	50	3	2	
48	63	Graduation	Single	NaN	2	1	48	5	48	6	10	
58	32	Graduation	Single	NaN	1	0	11	3	22	2	2	
71	41	2n Cycle	Married	NaN	1	0	25	3	43	17	4	
90	57	PhD	Married	NaN	2	1	230	42	192	49	37	
91	57	Graduation	Single	NaN	1	1	7	0	8	2	0	
92	41	Master	Together	NaN	0	0	445	37	359	98	28	
128	53	PhD	Married	NaN	0	1	352	0	27	10	0	
133	51	Graduation	Married	NaN	0	1	231	65	196	38	71	
312	25	Graduation	Married	NaN	0	0	861	138	461	60	30	
319	44	Graduation	Single	NaN	1	2	738	20	172	52	50	
1379	44	Master	Together	NaN	0	1	187	5	65	26	20	
1382	56	Graduation	Together	NaN	1	1	19	4	12	2	2	
1383	50	2n Cycle	Single	NaN	1	1	5	1	9	2	0	
1386	42	PhD	Together	NaN	1	0	25	1	13	0	0	
2059	45	Master	Together	NaN	1	1	375	42	48	94	66	
2061	33	PhD	Single	NaN	1	0	23	0	15	0	2	
2078	43	Graduation	Married	NaN	1	1	71	1	16	0	0	
2079	60	Master	Together	NaN	0	1	161	0	22	0	0	
2081	59	Graduation	Single	NaN	0	1	264	0	21	12	6	
2084	71	Master	Widow	NaN	0	0	532	126	490	164	126	
2228	36	2n Cycle	Together	NaN	0	0	32	2	1607	12	4	



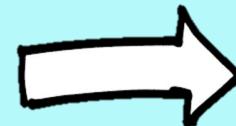
Modifying existing rows

Year_Birth



Age

AcceptedCmp1-5
&
Response



Receptiveness

Identifying our Features and Responses

Feature variables

- Age
- Income
- Kidhome
- Teenhome
- Education
- Marital_Status

Response variables

- MntWines
- MntFruits
- MntMeatProducts
- MntFishProducts
- MntSweetProducts
- MntGoldProds
- NumWebPurchases
- NumCatalogPurchases
- NumStorePurchases
- NumWebVisitsMonth
- Receptiveness

Identifying the different variable types



Continuous Numerical Variables

- Age
- Income
- MntWines
- MntFruits
- MntMeatProducts
- MntFishProducts
- MntSweetProducts
- MntGoldProds

Discrete Numerical Variables

- Kidhome
- Teenhome
- NumWebPurchases
- NumCatalogPurchases
- NumStorePurchases
- NumWebVisitsMonth

Categorical Variables

- Education
- Marital_Status
- Receptiveness



Decoding Kidhome & Teenhome

0



zero

1



one

2



two

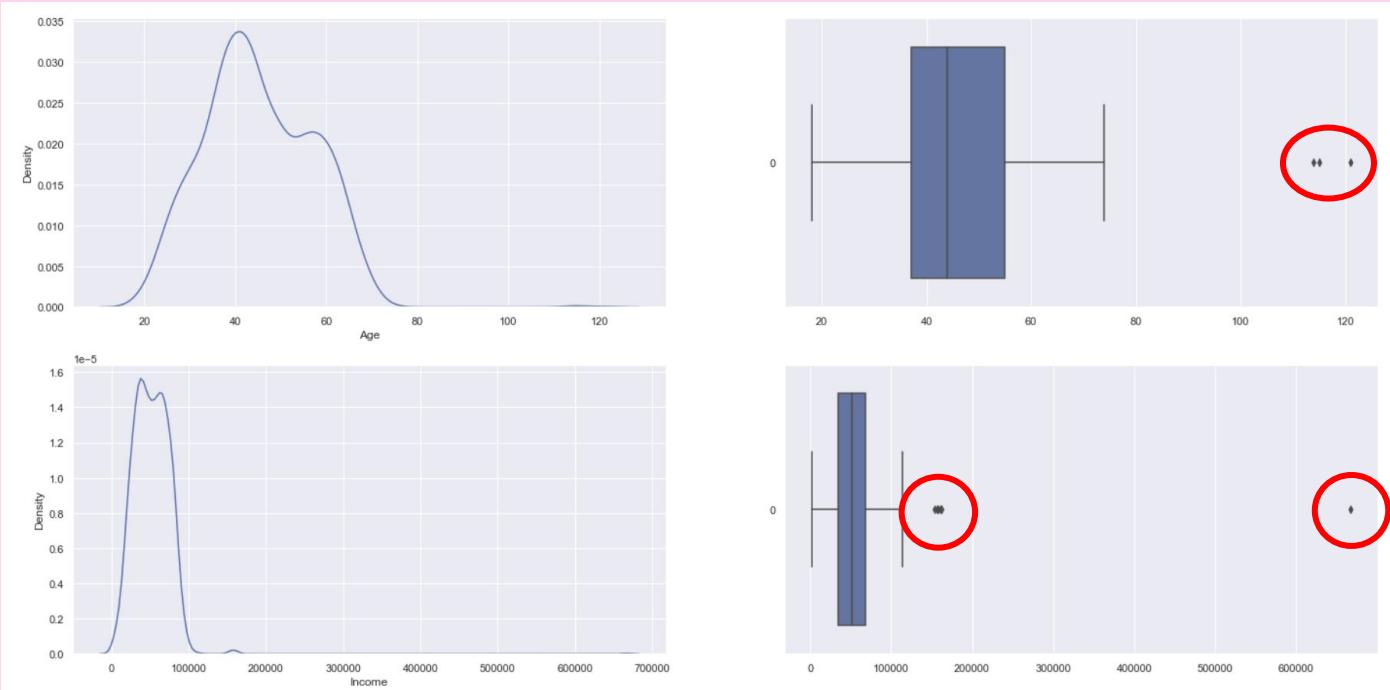
Uni-variate analysis



Exploratory ANALYSIS

Statistical DESCRIPTION

Income & Age

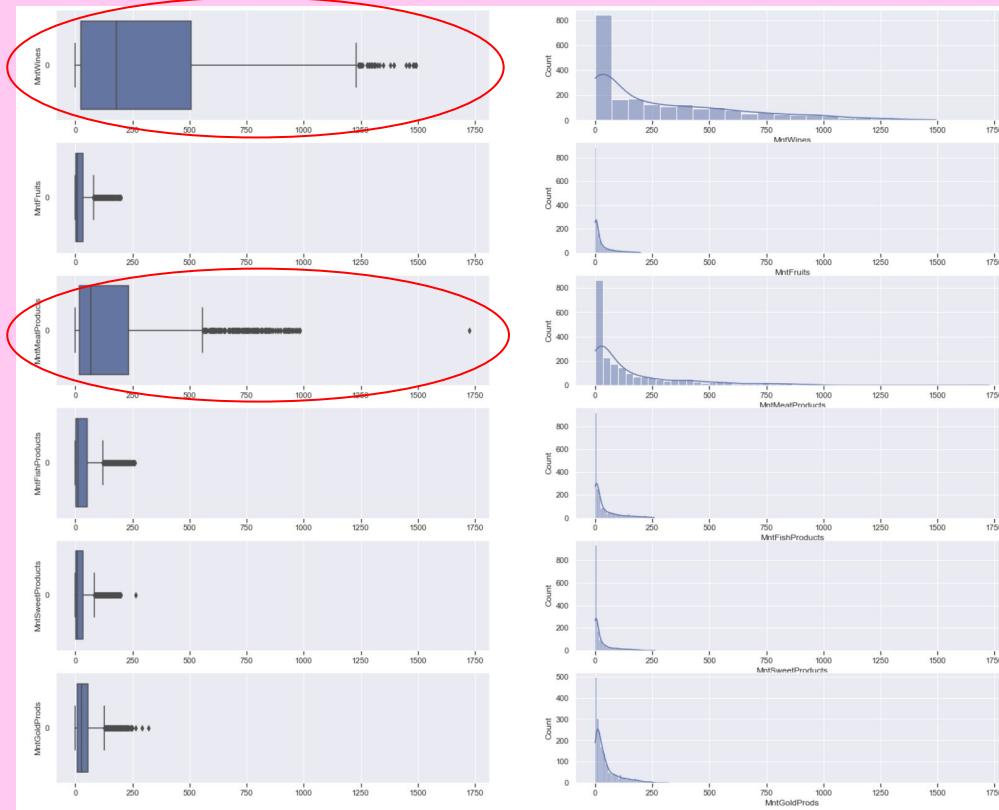


Exploratory ANALYSIS

Statistical DESCRIPTION

Statistical DESCRIPTION

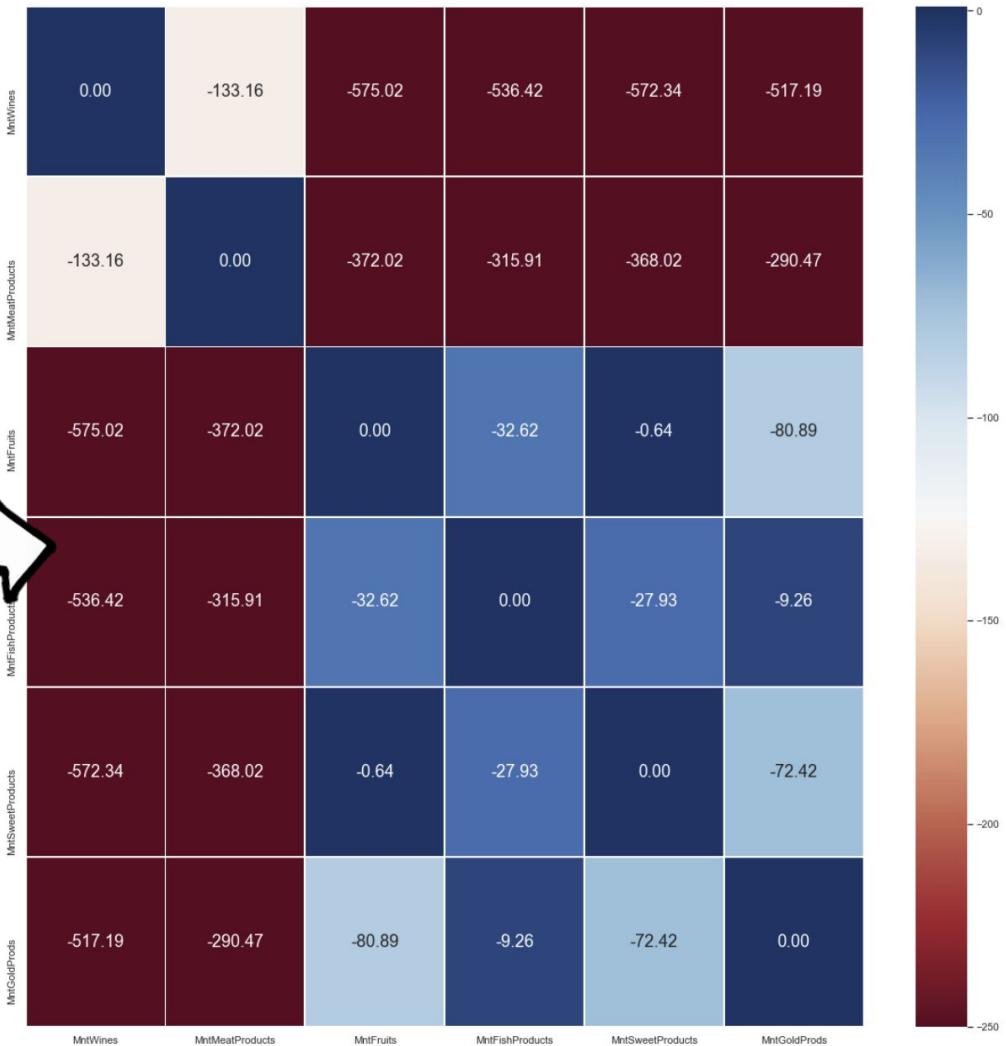
Amount Spent on various goods



Exploratory ANALYSIS

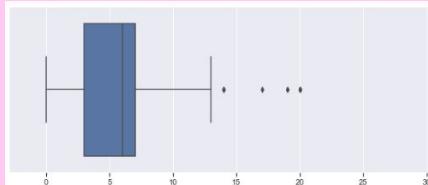
Statistical DESCRIPTION

Dark Red indicates
greater
difference in
distribution!



Number of Purchases on different platforms & Number of Web visits

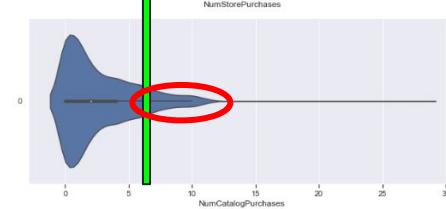
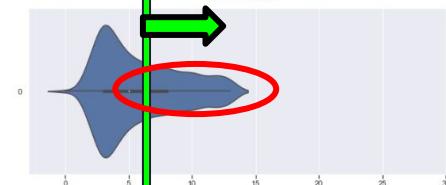
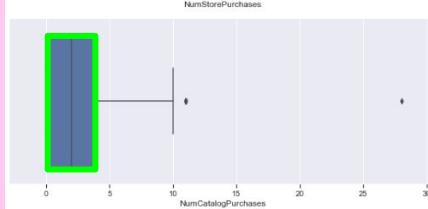
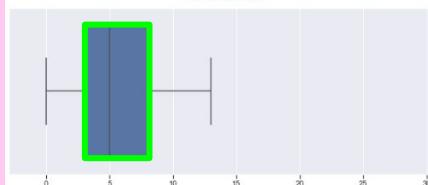
2



1



3





Uni-variate analysis of Categorical Variables



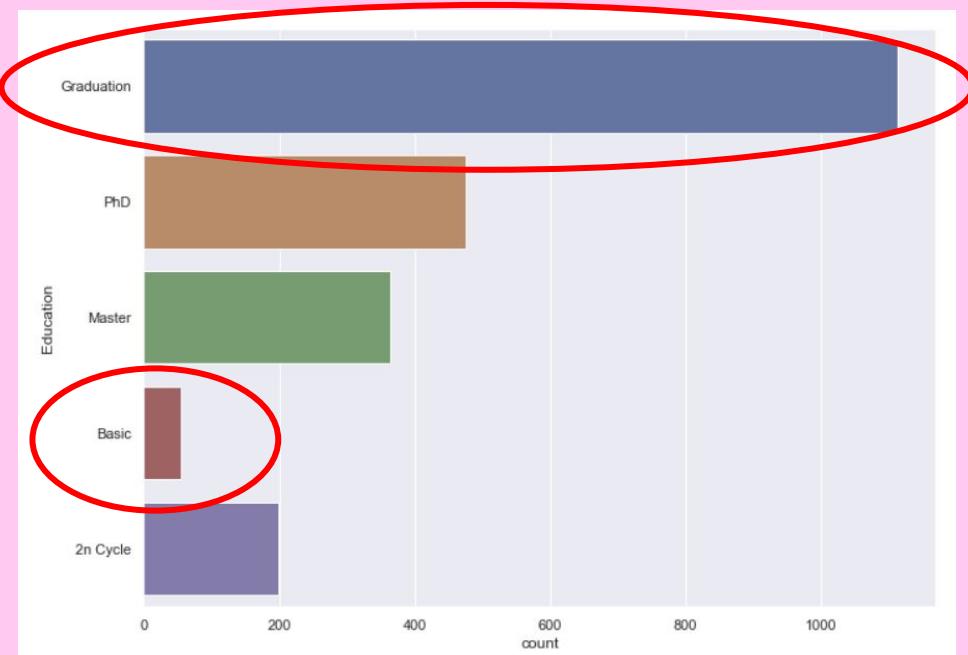
Imbalances



Education

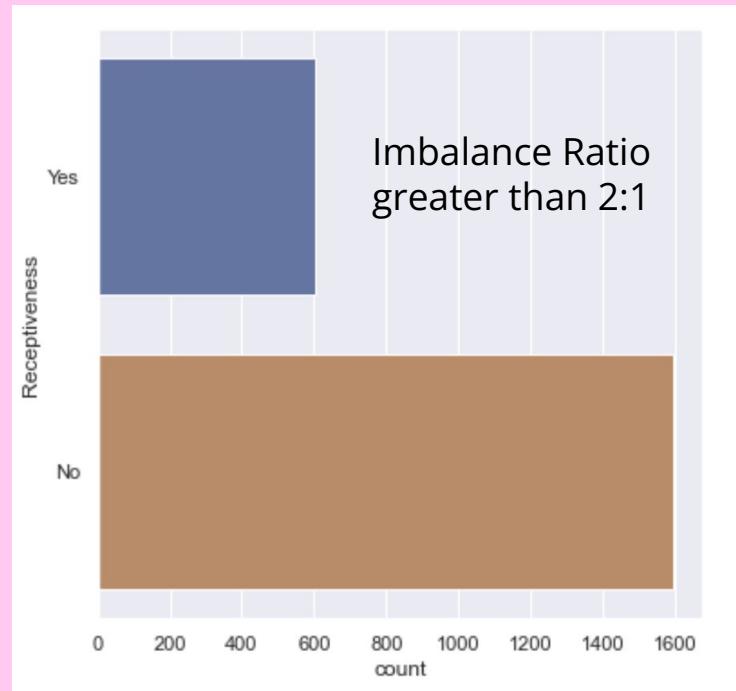
(Categorical Cluster Feature)

```
Education:  
Graduation    1113  
PhD           476  
Master         364  
2n Cycle       198  
Basic          54  
Name: Education, dtype: int64
```



Receptiveness:
No 1597
Yes 601
Name: Receptiveness, dtype: int64

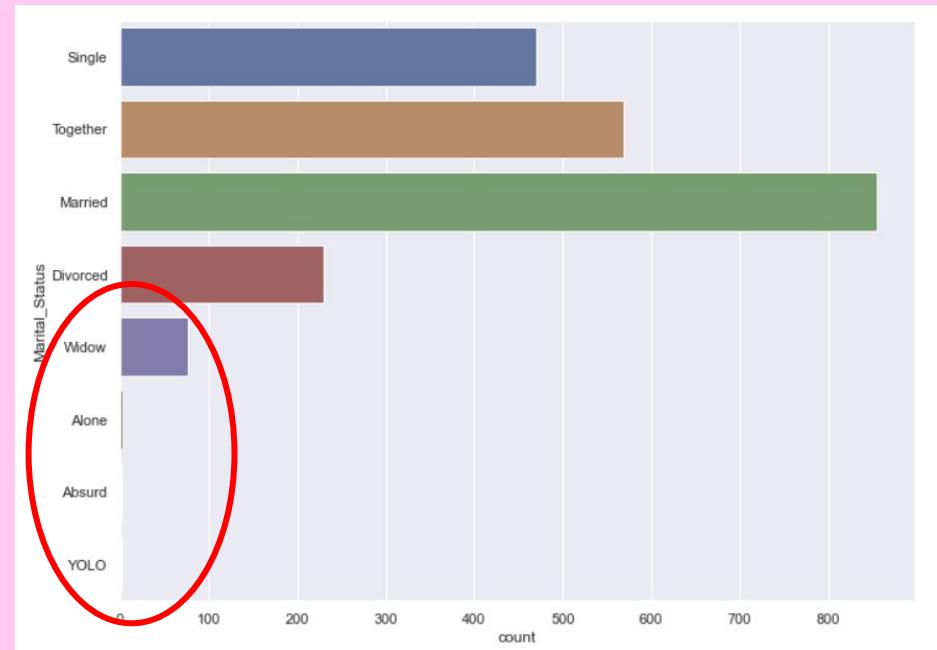
Receptiveness



Marital Status

(Categorical Cluster Feature)

```
Marital_Status:  
Married      854  
Together     568  
Single       470  
Divorced     230  
Widow        76  
Alone         3  
Absurd        2  
YOLO          2  
Name: Marital_Status, dtype: int64
```

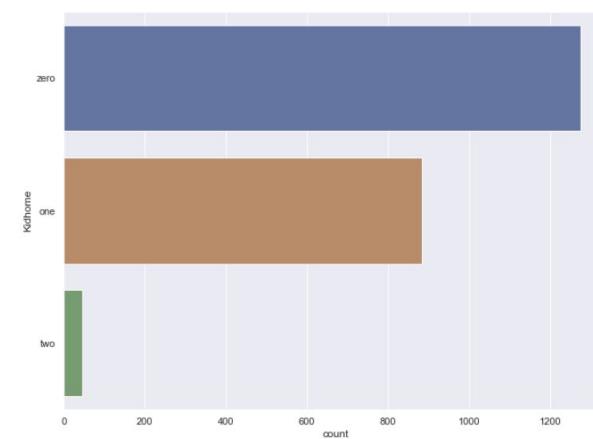


Kidhome & Teenhome (Categorical Cluster Features)

Kidhome:

zero	1276
one	883
two	46

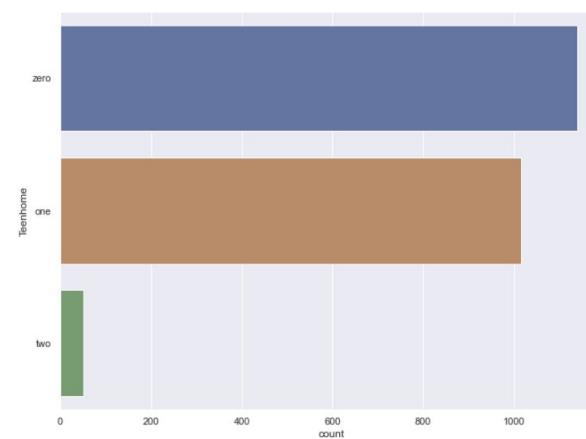
Name: Kidhome, dtype: int64



Teenhome:

zero	1139
one	1015
two	51

Name: Teenhome, dtype: int64



Combining 'one' and
'two' for Kidhome
and Teenhome into
a single level

```
No Kids      1276  
Have Kids    929  
Name: Kidhome, dtype: int64
```

```
No Teens     1139  
Have Teens   1066  
Name: Teenhome, dtype: int64
```

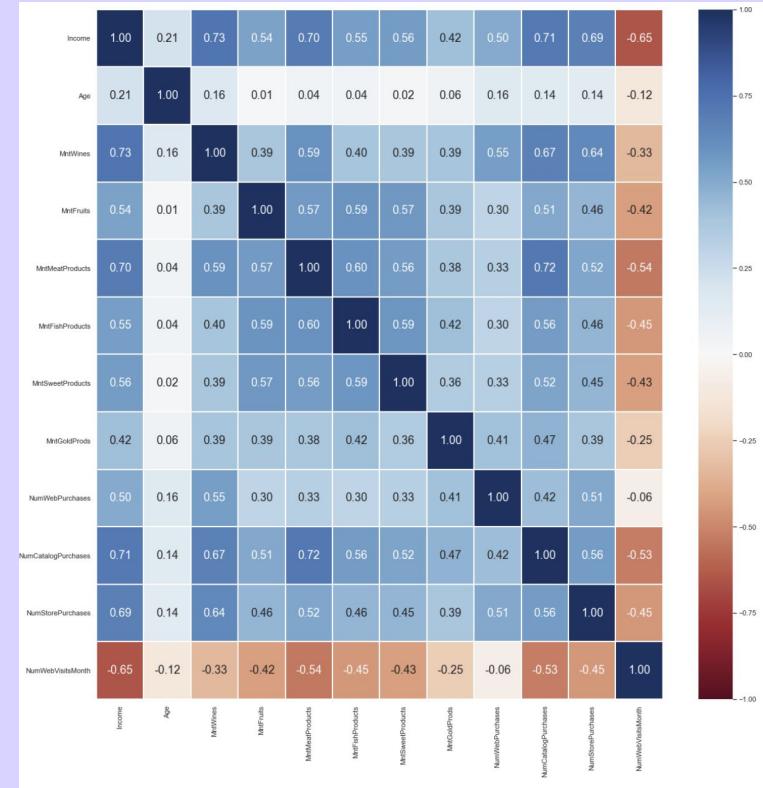
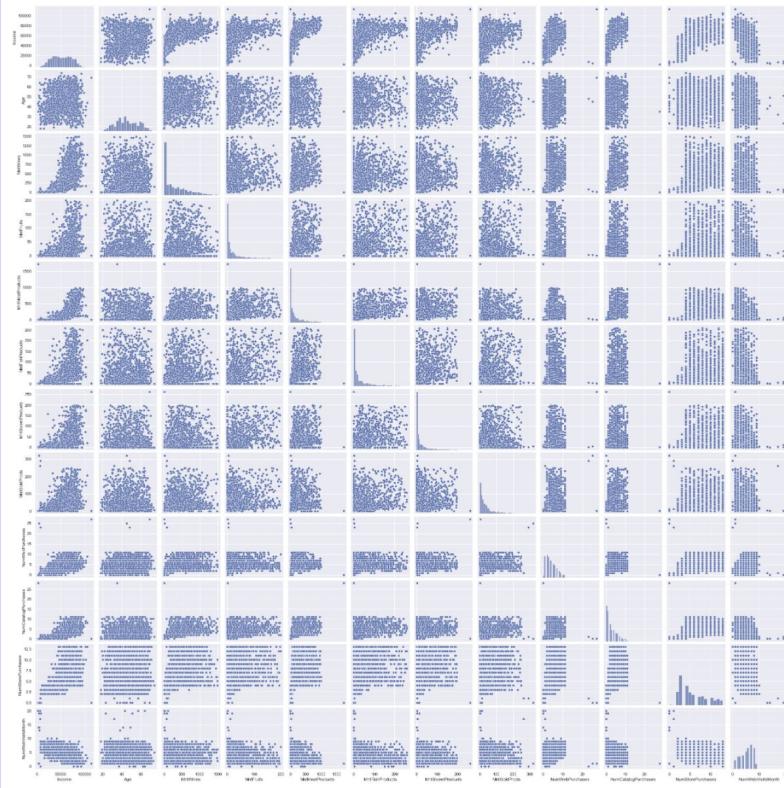
Analytic Visualisation & In-Depth Analysis



Bi-variate analysis of **Numerical** Variables



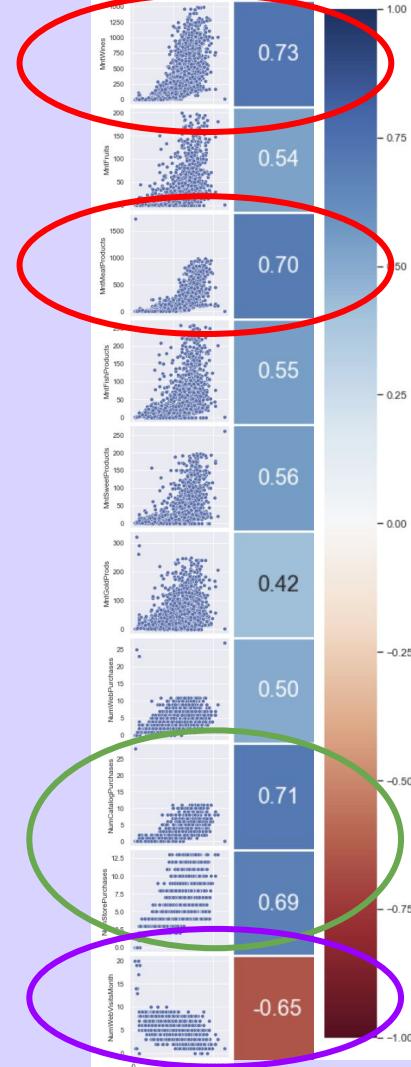
Income & Age against ALL Numerical Response variables





Income

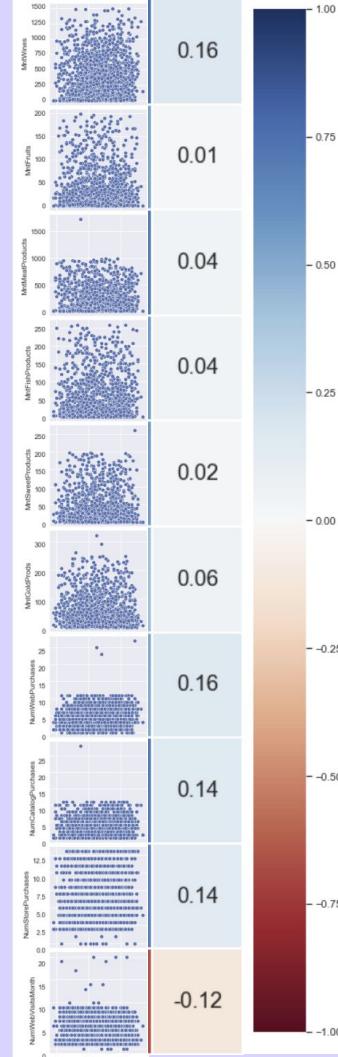
- High positive correlation with amount spent on **Wine and Meat products**
- High positive correlation with number of store and catalogue purchases
- Moderate **negative** correlation with number of website visits per month





Age

- Weak to almost negligible correlation with any of the numerical response variables
- This might mean that it isn't a good clustering feature
- However, since effective adverts are designed to appeal to the age-group of the target audience, we decided to still keep it as a clustering feature



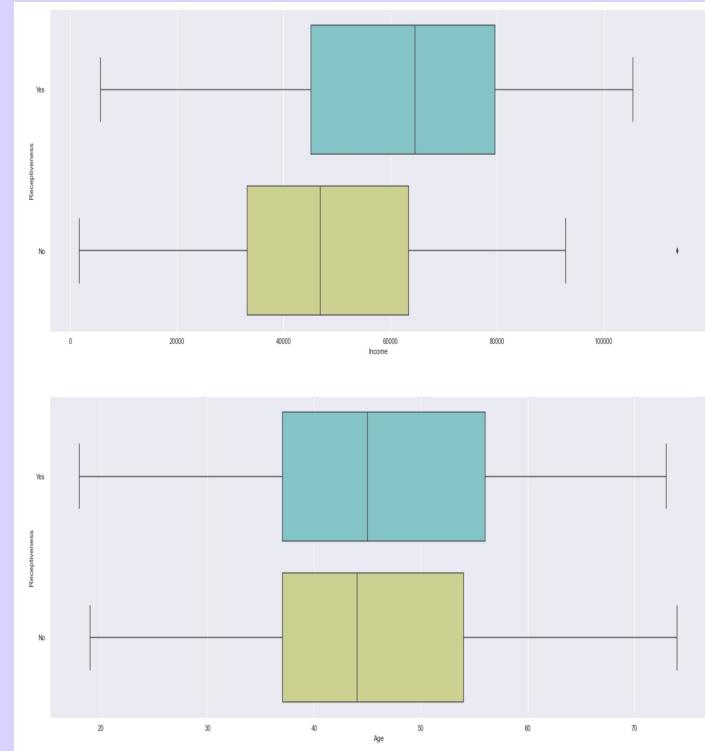
Income & Age against Receptiveness

Income:

The higher one's income, the more receptive one is to promotions

Age:

Weak to negligible relationship between one's age and his receptiveness to promotions





Bi-variate analysis of **Categorical** Variables

Analytic
VISUALIZATION

Pattern
RECOGNITION

1. Kidhome



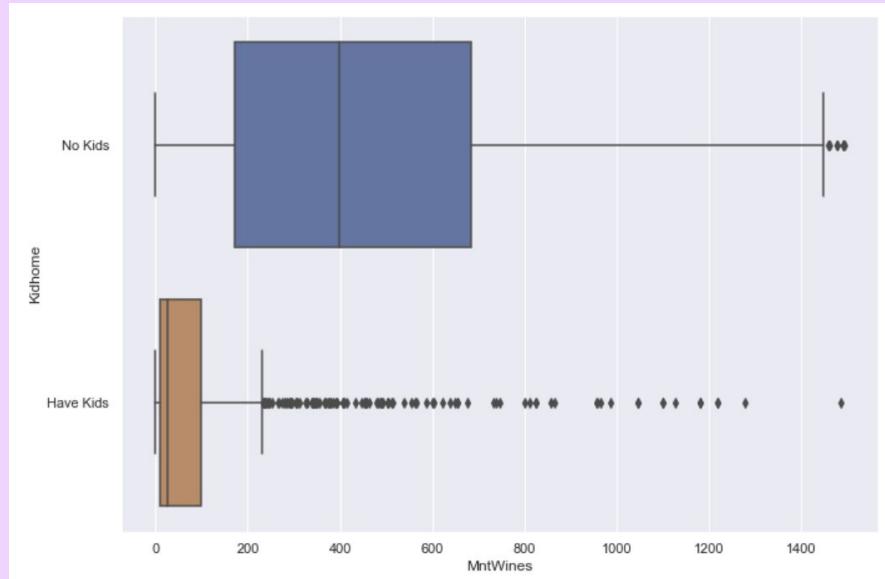
Kidhome against Amount Spent on products (Numerical Response Variables)

- Amount purchased tends to be **higher for all products** among people **without kids**.



Amount Spent on Wine (MntWines)

- Most notably, people without kids tend to spend a lot more on wine
- This can be seen from the greatest separation between the Interquartile ranges of both boxplots





Kidhome against Number of Purchases on different platforms & Website visits (Numerical Response Variables)

- Number of purchases on all platforms higher among people **without kids**
- People with kids tend to **visit the online store more frequently**

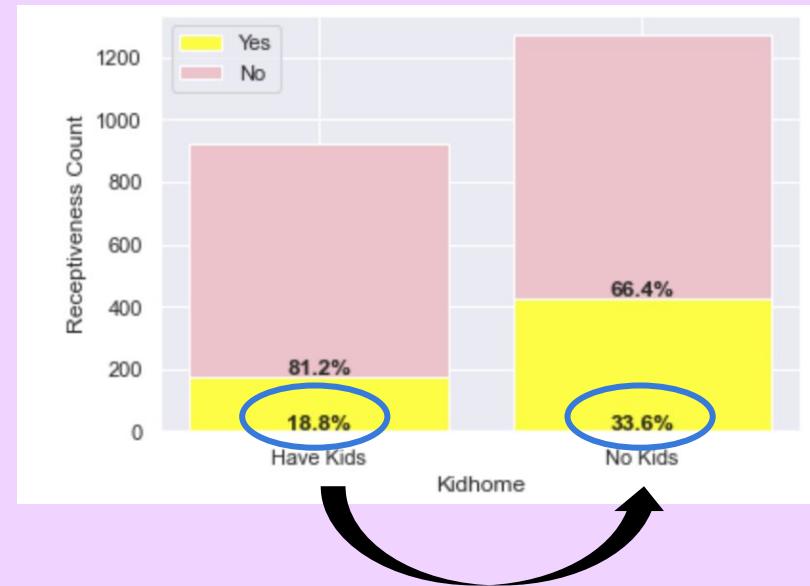




Kidhome against Responsiveness

(Categorical Response Variable)

- People without kids tend to be more receptive to promotions
- This can be seen from the increase in proportion of people receptive to promotions



Analytic
VISUALIZATION

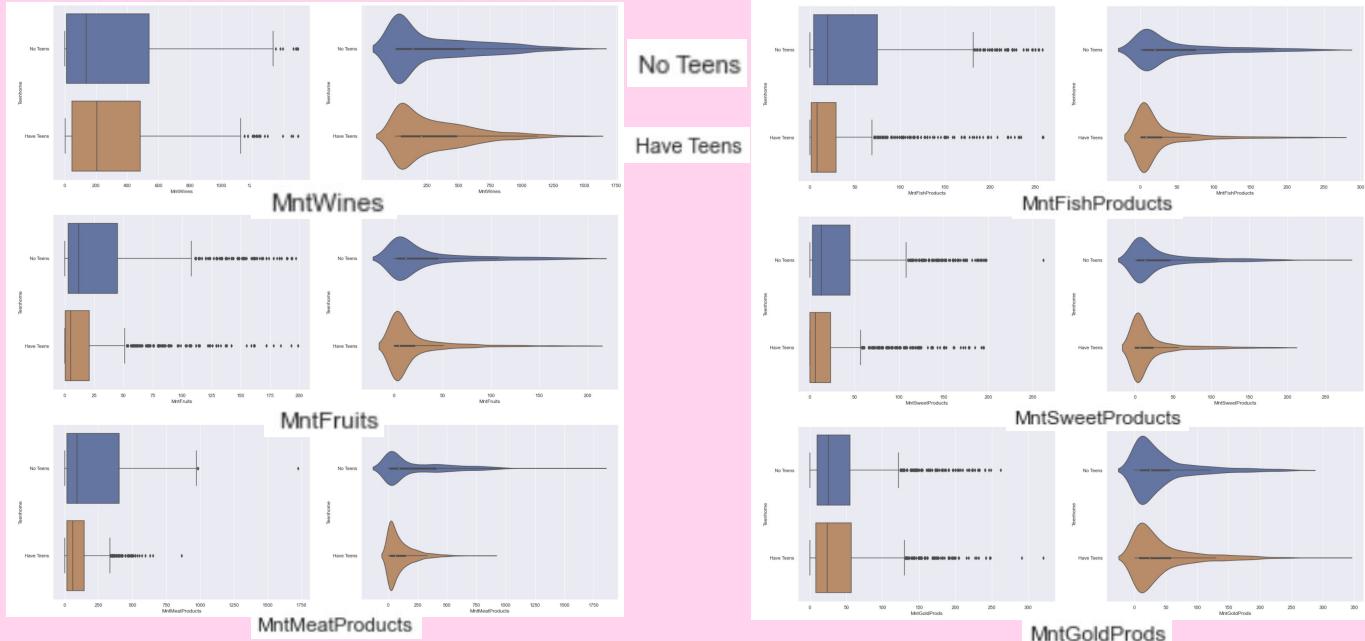
Pattern
RECOGNITION

2. Teenhome



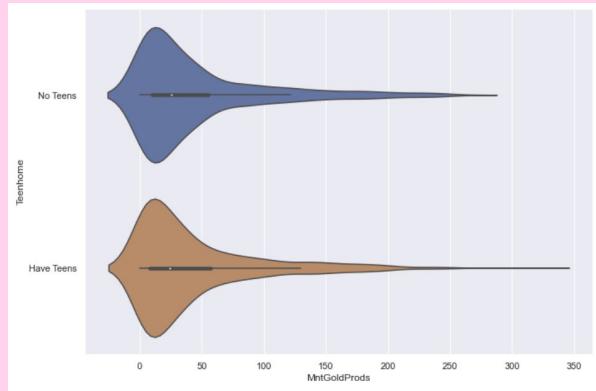
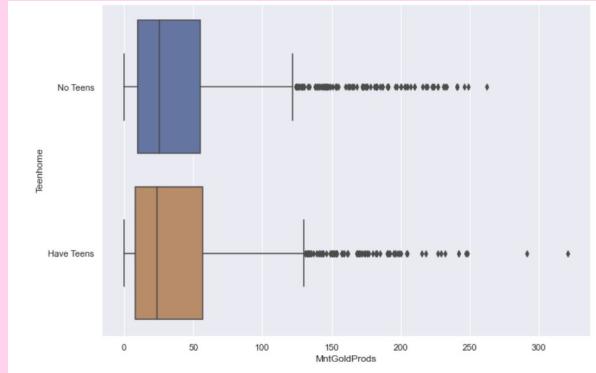
Teenhome against Amount Spent on products (Numerical Response Variables)

- Generally, the amount spent on all products are **higher among people without teens**



Amount Spent on Gold (MntGoldProds)

- The amount spent on Gold (MntGoldProds) are **similar** for both levels
- This can be seen from **almost identical** distributions and box plots for both levels





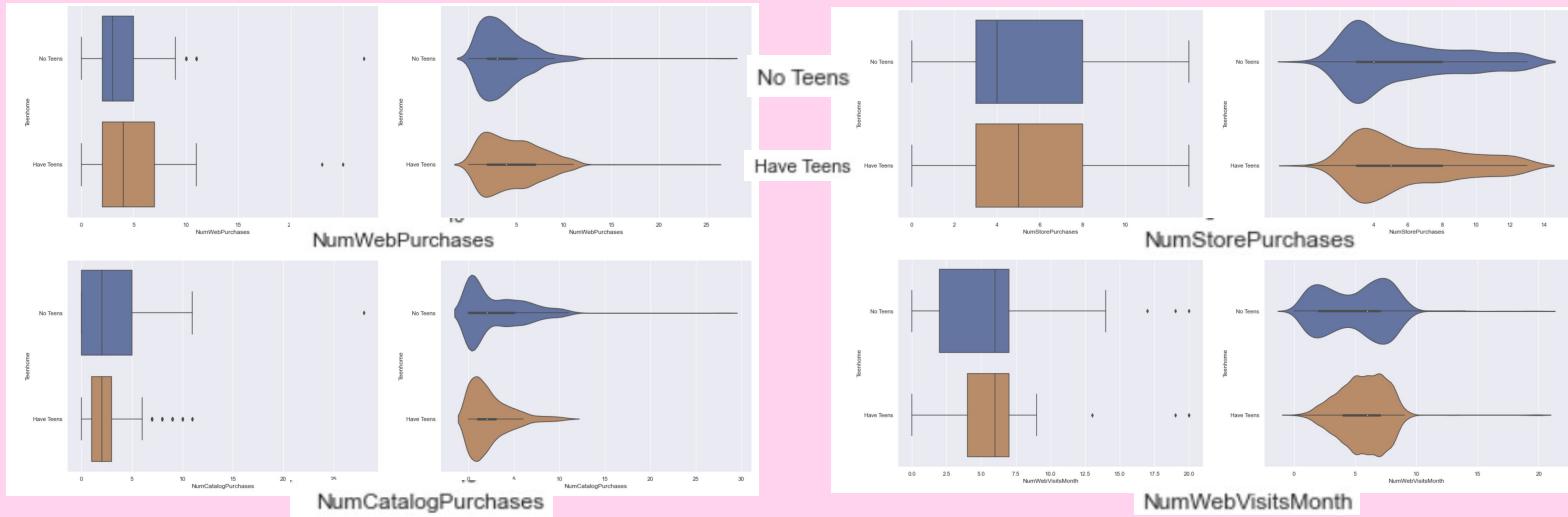
Amount Spent on Wine (MntWines)

- People with teens tend to purchase slightly more wine (MntWines) than people without
- This is seen from the slightly higher median in the box plot for 'Have Teens' compared to that of 'No Teens'



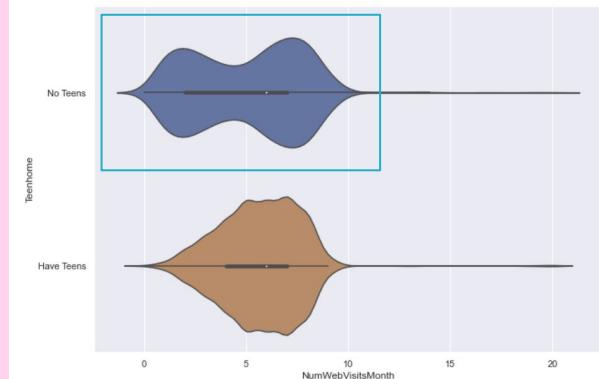
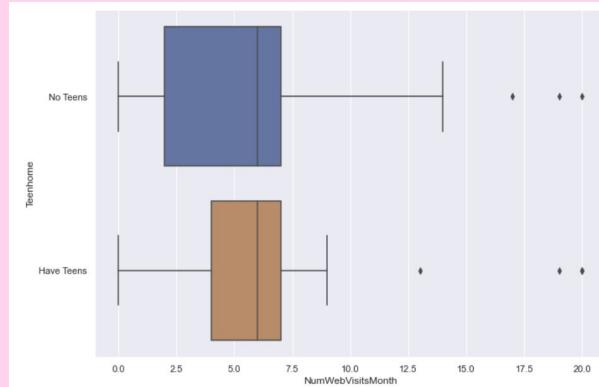
Teenhome against Number of Purchases on different platforms & Website visits (Numerical Response Variables)

- Generally, the number of purchases between both levels are **largely the same** with the exception of Number of Website Visits per month (NumWebVisits)



Number of Website Visits per month (NumWebVisitsMonth)

- **Bimodal distribution** for 'No Teens' as evinced from violin plot
- A possible indication of **two separate groups** of people without teens
 - one group tends to visit the mall's website more often

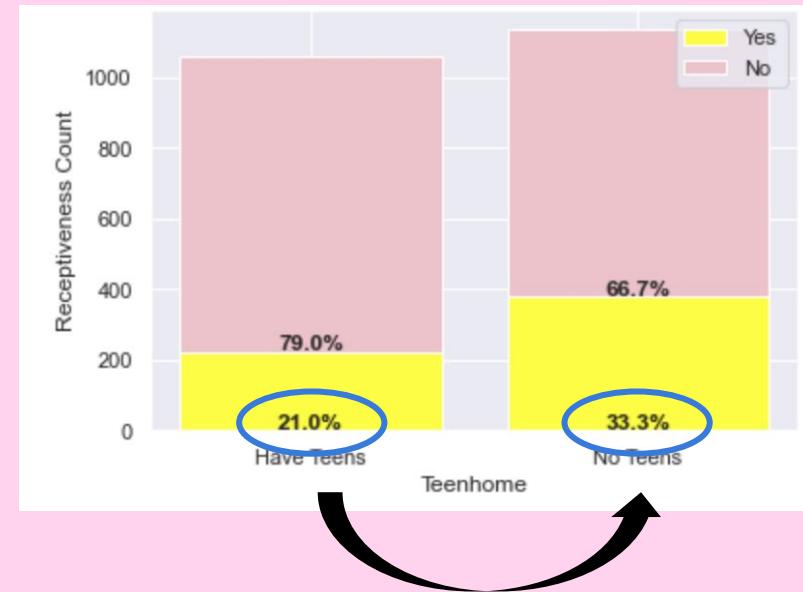




Teenhome against Responsiveness

(Categorical Response Variable)

- People without teens also tend to be more receptive to promotions
- This can be seen from the increase in proportion of people receptive to promotions



Analytic
VISUALIZATION

Pattern
RECOGNITION

3. Education



Education against Amount Spent on products

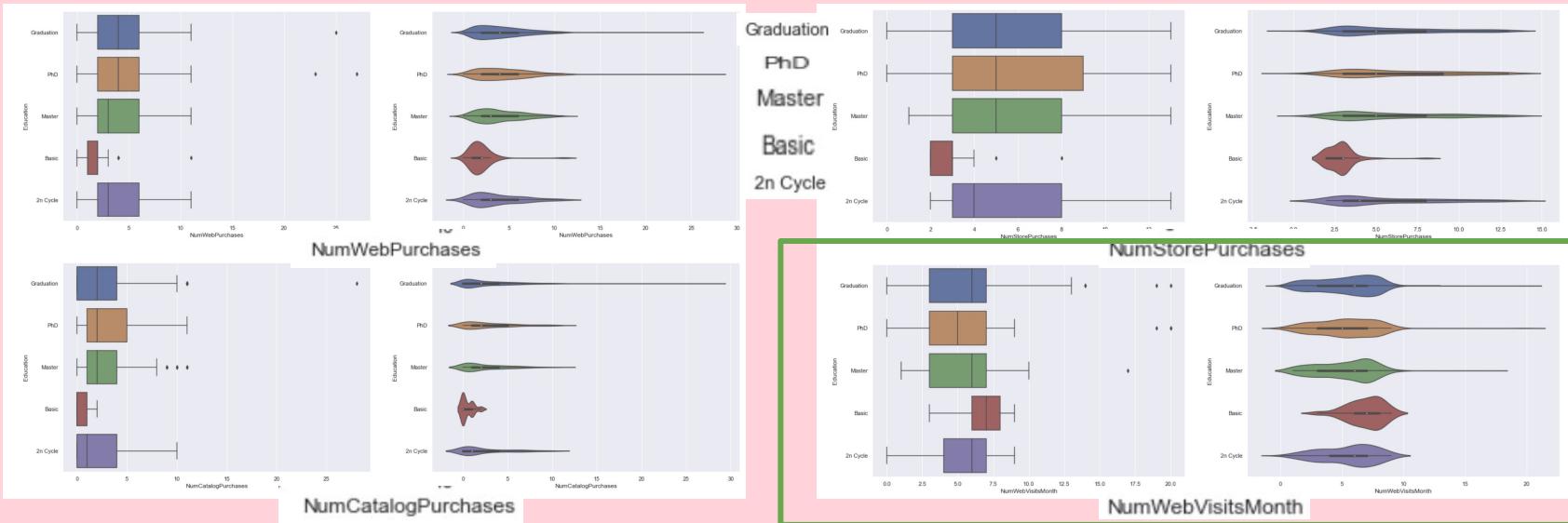
(Numerical Response Variables)

- People with 'Basic' Education **spend and purchase the least compared to other levels**



Education against Number of Purchases on different platforms & Website visits (Numerical Response Variables)

- Although spending the least on all platforms, people with 'Basic' education visit the online website most often

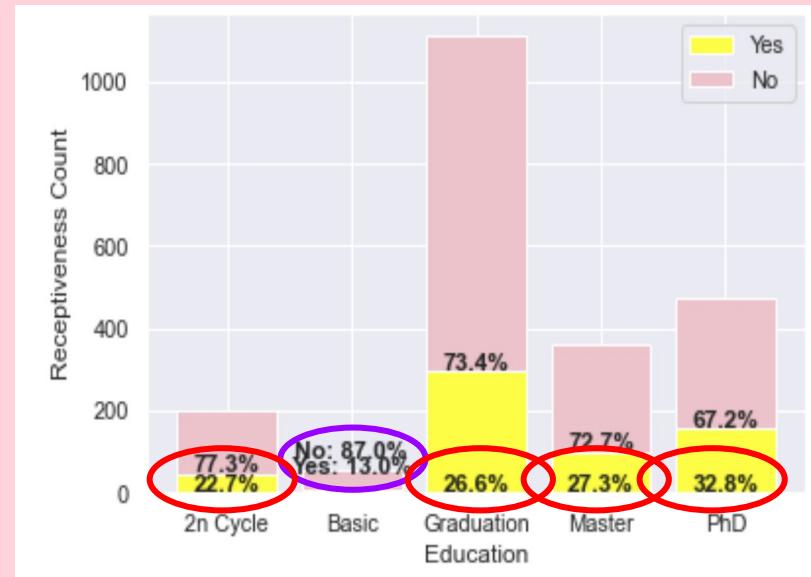




Teenhome against Responsiveness

(Categorical Response Variable)

- People with 'Basic' Education tend to be **less receptive to promotions** compared to other education levels
- This can be seen from how only **13% of data points** with 'Basic' are receptive compared to between **22.7%** to **32.8%** in the other Education levels





HOWEVER...

- 'Basic' is **skewed** with only 54 counts
- **Statistically unreliable** as current distribution may not reflect actual distribution
- Other levels of education are all **considered highly educated** enough to make similar spending decisions



Decided **not** to use it as a clustering feature



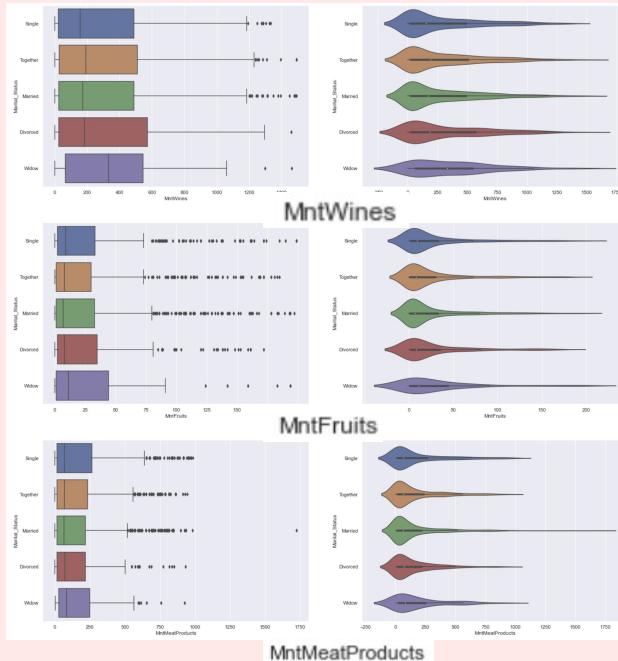
4. Marital Status



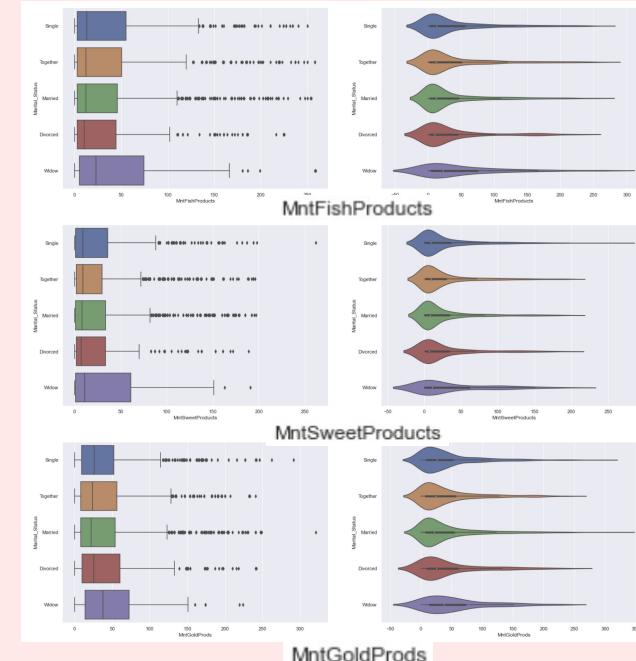
Marital Status against Amount Spent on products

(Numerical Response Variables)

- Except for 'Widow', Marital Status **does not seem to affect the amount spent** on different products much
- This can be seen from the **largely similar distributions and box plots** for all levels except for 'Widow' which is **slightly different**



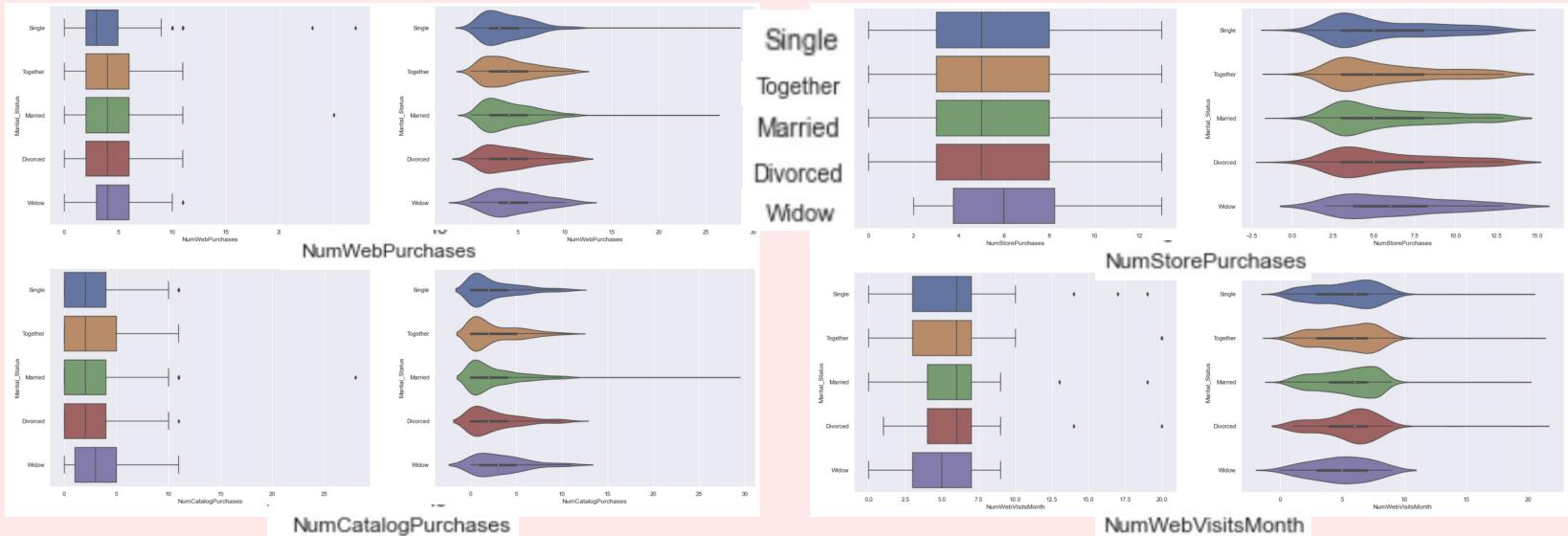
Single
Together
Married
Divorced
Widow



Single
Together
Married
Divorced
Widow

Education against Number of Purchases on different platforms & Website visits (Numerical Response Variables)

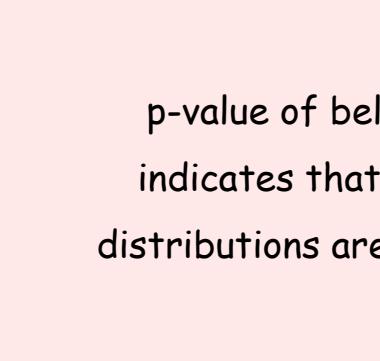
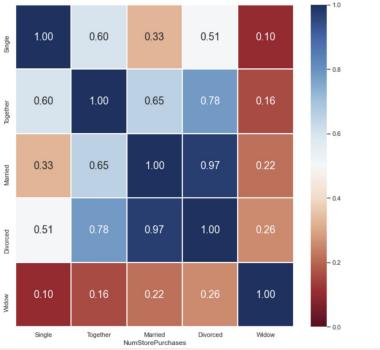
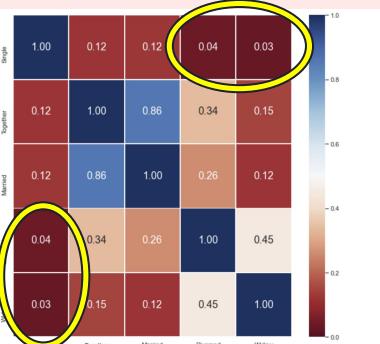
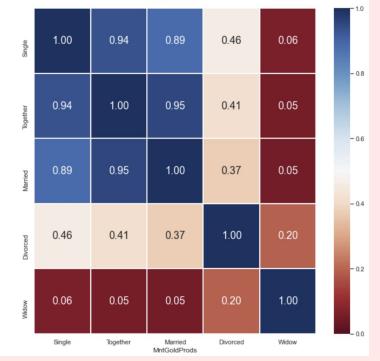
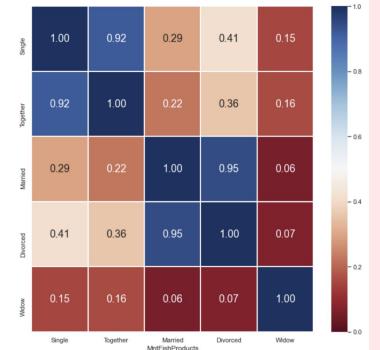
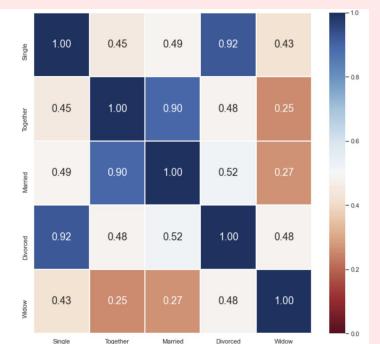
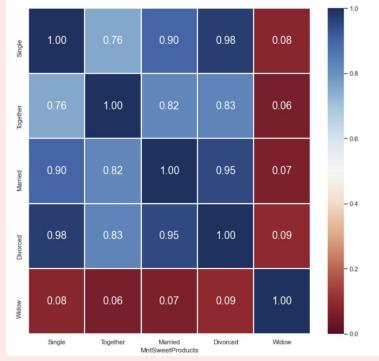
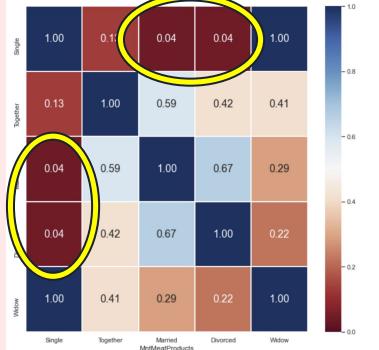
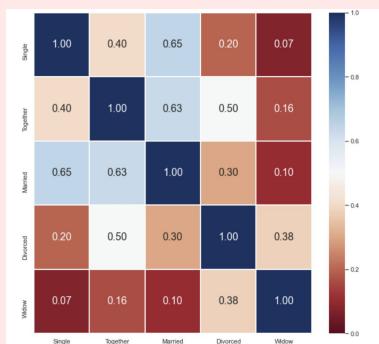
- Likewise, Marital Status **do not seem to be related to these response variables** as seen from the largely similar distributions and box plots, with the **exception of 'Widow'** which is also **slightly different**



Statistical DESCRIPTION



Exploratory ANALYSIS



**p-value of below 0.05
indicates that the two
distributions are different**

- The lowest p-value (0.03) is seen between 'Widow' and 'Single' for "NumWebPurchases"
- Low p-values of 0.04 also seen between levels other than 'Widow' but it isn't far from 0.05

A closer look

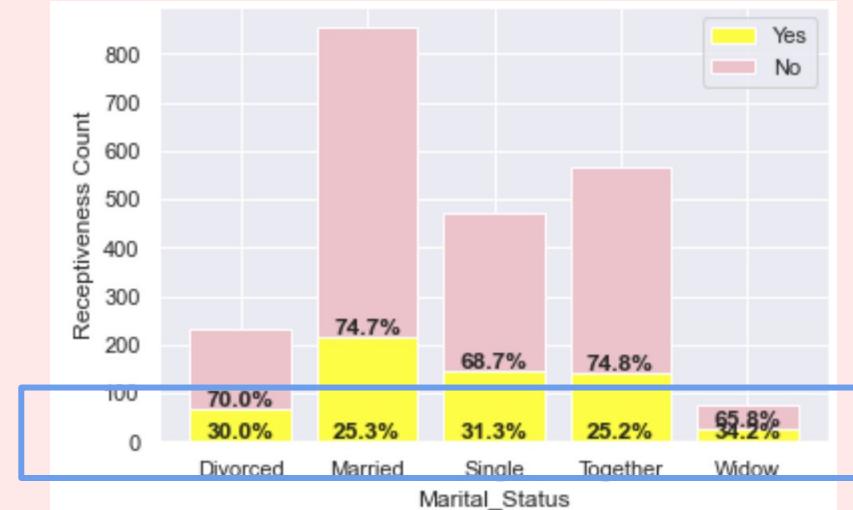




Marital Status against Responsiveness

(Categorical Response Variable)

- No clear relationship between one's marital status and receptiveness towards promotions
- This can be seen from a largely similar proportion of data that are receptive to promotions in each level





BAD CLUSTERING FEATURE!

- Marital Status **does not seem to have any relationship** with the response variables as seen earlier
- Marital Status is also **very imbalanced** ('Widow' and 'Divorced' are the minority classes)
- Conclusions drawn for 'Widow' and 'Divorced' may be unreliable



Decided **not** to use it as a clustering feature

Machine Learning (Clustering)





K-Prototypes

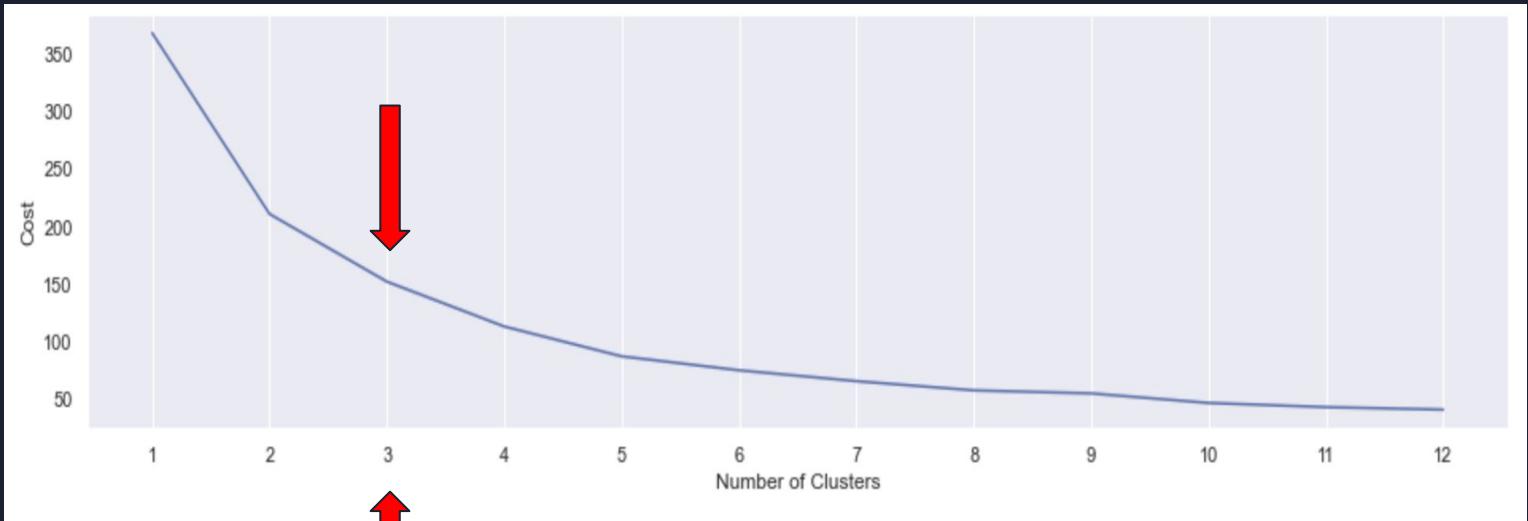


The Best k-value

Machine
LEARNING



Algorithmic
OPTIMIZATION



Core-Analysis





Customer Profiles of each Cluster

Analysis of clusters

Plotting our clusters on a 2-D graph shows some degree of separation among the clusters

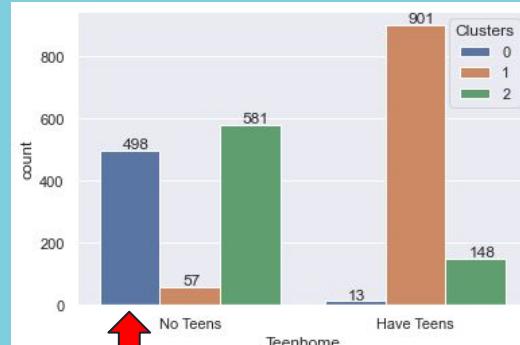
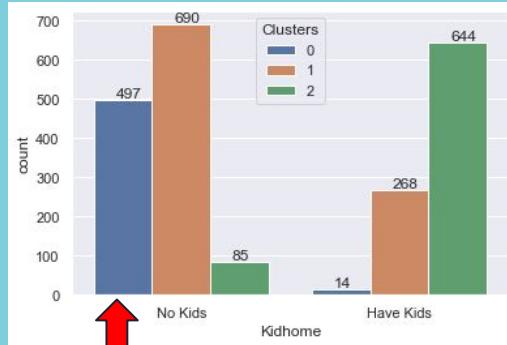
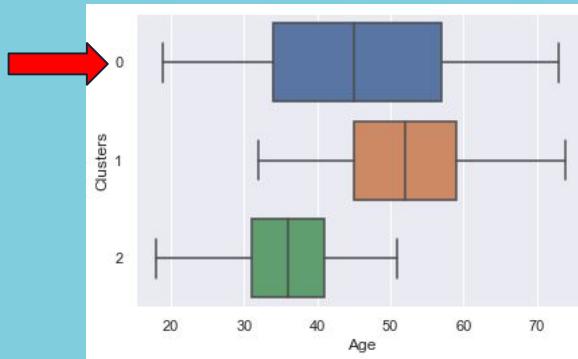
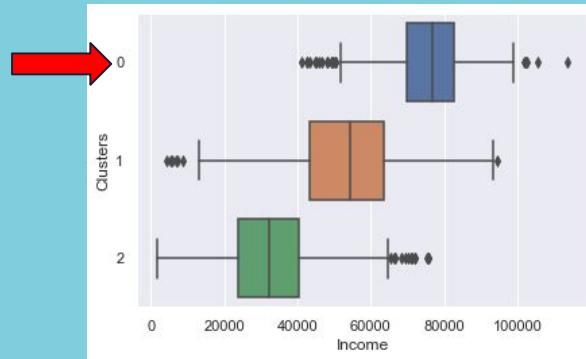
However, since we used 4 features for clustering, using a 2-D graph does not seem to exhibit the actual 'Distance' between the clusters

We will analyse the clusters against each cluster variable individually to have a better understanding of the different clusters





Analysis of clusters





Customer Profiles of each Cluster

Cluster 0:

No Children (Kids or Teens)
Rich

Cluster 1:

Only have Teens
Middle-Income
30 Years old and above

Cluster 2:

Only have Kids
Low-Income
50 Years old and below



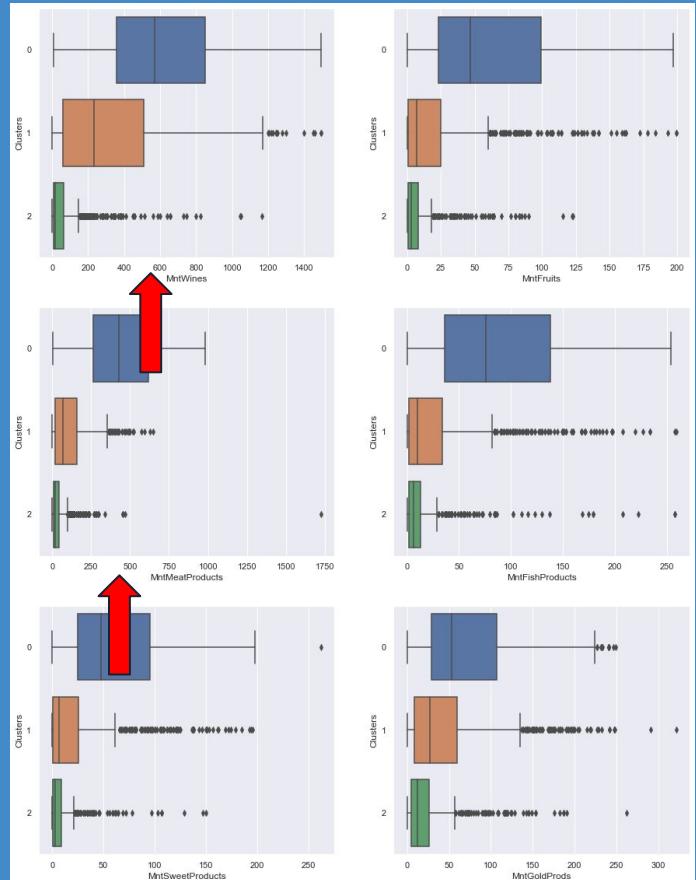
Purchasing behaviour for each cluster



Cluster 0 (Rich & No Children)

Purchase the most products in all categories

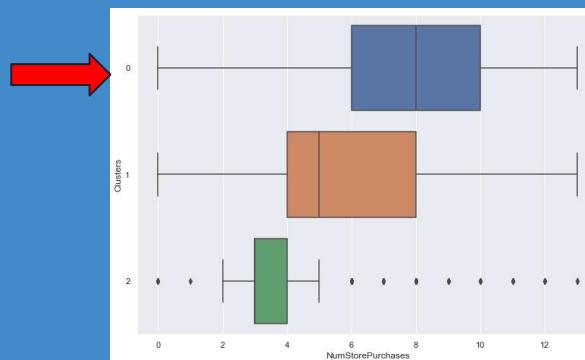
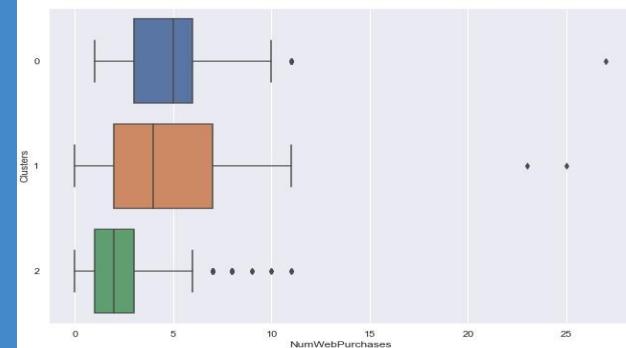
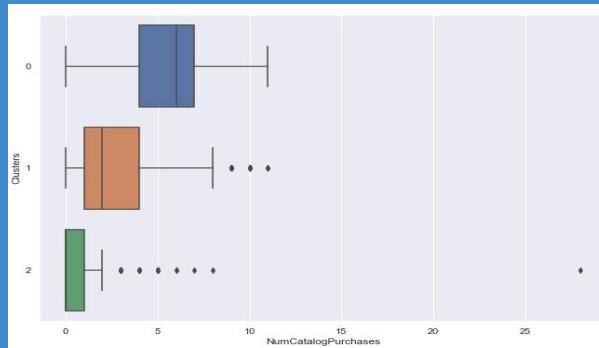
Disproportionately higher amount for Wine and Meat products (Absolute terms)





Cluster 0 (Rich & No Children)

Purchase from physical stores most out of the 3 platforms

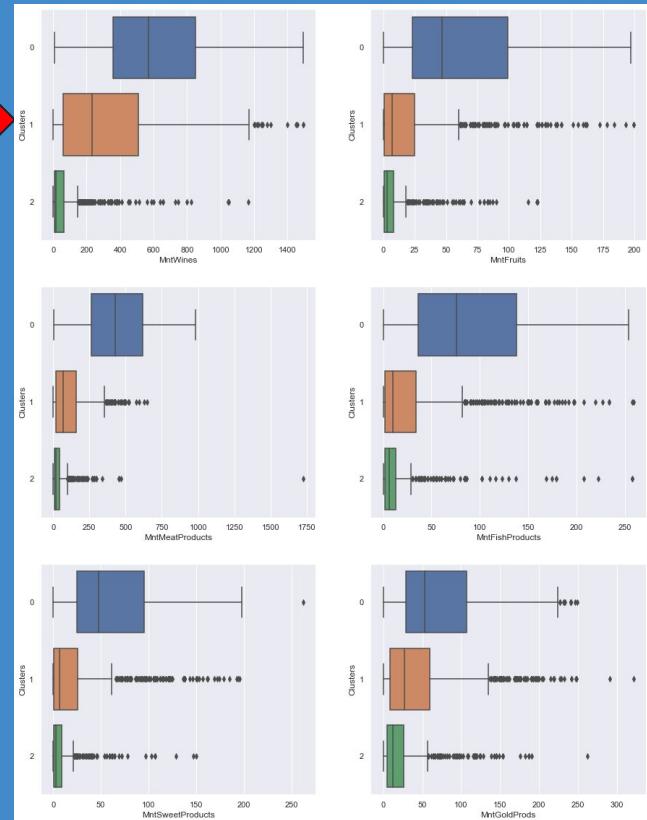




Cluster 1 (Has Teen, >30 Y/O, Middle-income)

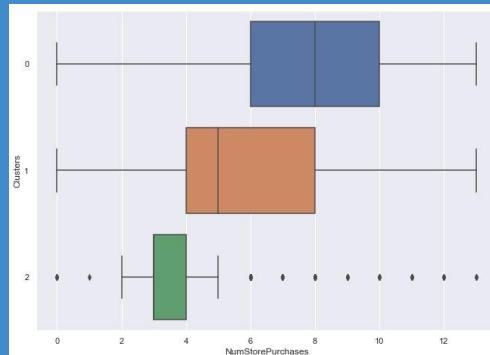
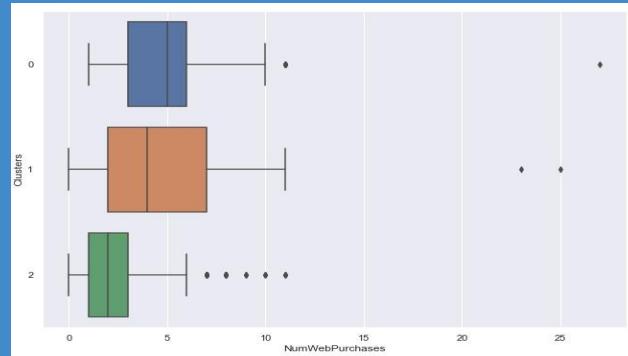
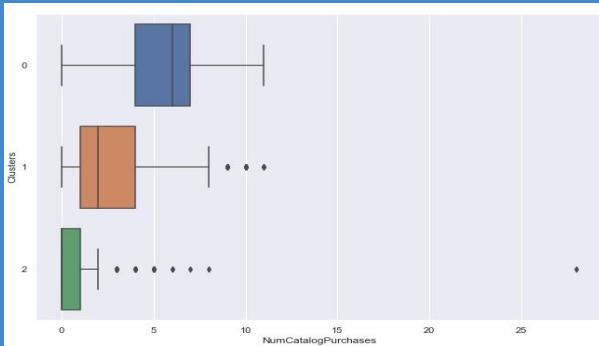
Purchase less than those in cluster 0 but more than those in cluster 2

Buys a lot more wine products out of all the products



Cluster 1 (Has Teen, >30 Y/O, Middle-income)

Similar to cluster 0, cluster 1 purchase from physical stores most out of the 3 platforms

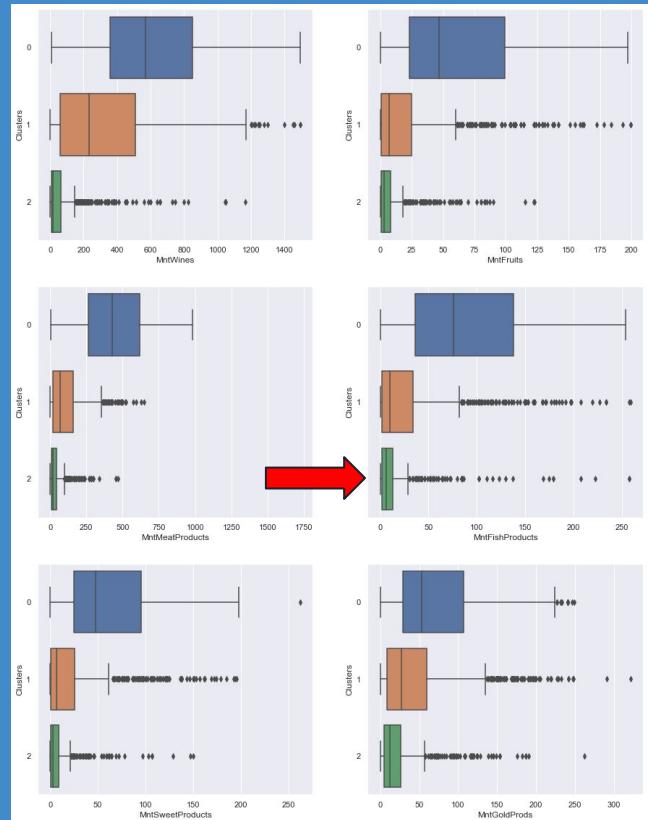


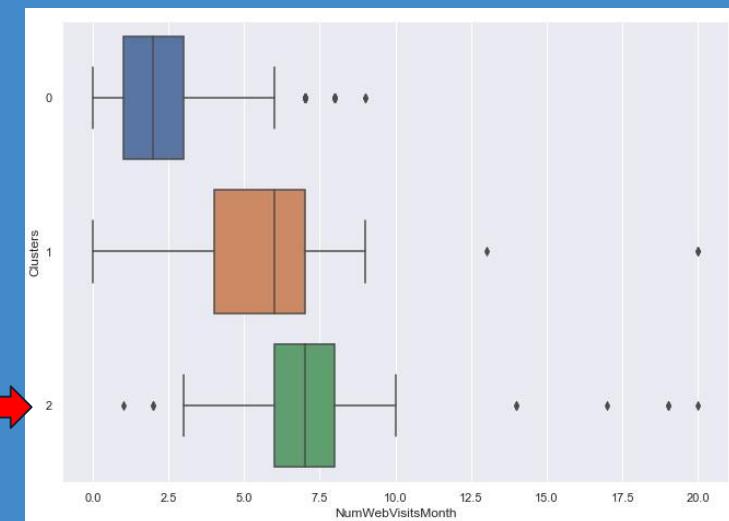
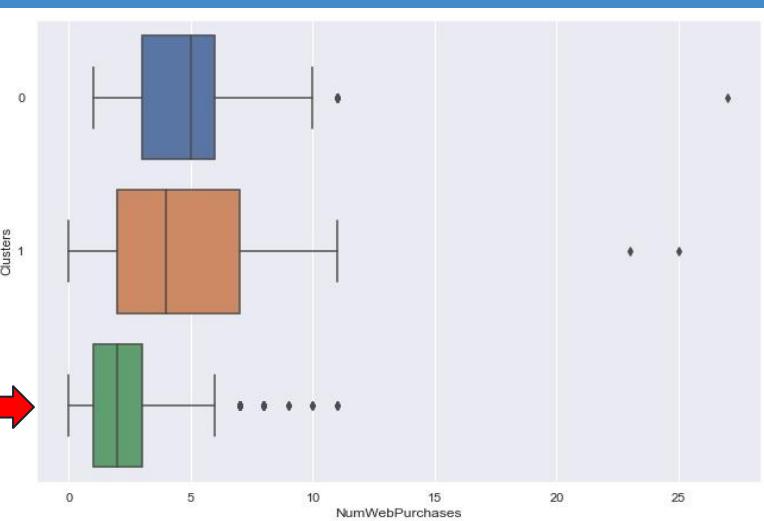


Cluster 2 (Has Kid, <50 Y/O, Low-income)

Purchase the least amount of products in all categories

Out of the necessities, buys the most amount of fish products (although not by much)



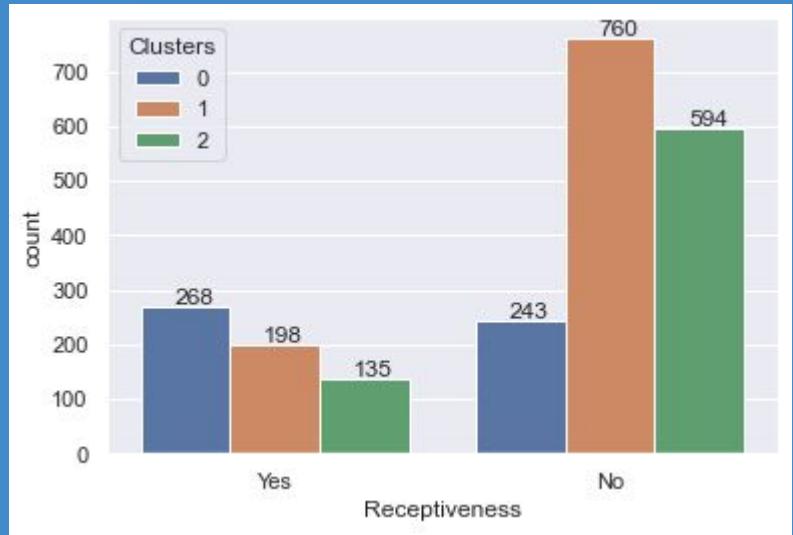


Cluster 2 (Has kid, <50 y/o, Low-income)

Visits the online shop most often but purchase the least from the online shop (tendency to window-shop)

Receptiveness based on clusters

People with higher income are more receptive to promotions
(in line with our previous conclusion)



Cluster 0 - 52% , Cluster 1 - 21% , Cluster 2 - 18%

Conclusion





Possible advertising strategies

1. Carry out advertisements for meat and wine products in physical stores to target clusters 0 and 1
(Make these advertisements cater to people of all age range)

2. Carry out advertisements for fish products in online stores to target cluster 2
(Make these advertisements cater to the younger population)

Main problems answered:

1. What products to advertise
2. Which platform to advertise on
3. Whether promotions should be used



Limitations

1. Ethical considerations:

While it is in line with business' interest to attract customers to maximise profits, advertisements of sensitive/addictive products should be done in a tasteful manner and with accordance with the law. (eg. our advertisements should not promote unhealthy consumption of wine)

2. Our conclusion is limited to the types of products given in the dataset which might not be reflective of all the products sold. We could be missing out on a large opportunity to advertise other products which are not reflected in the dataset



WE GOT PROMOTED!!!



Thank you

