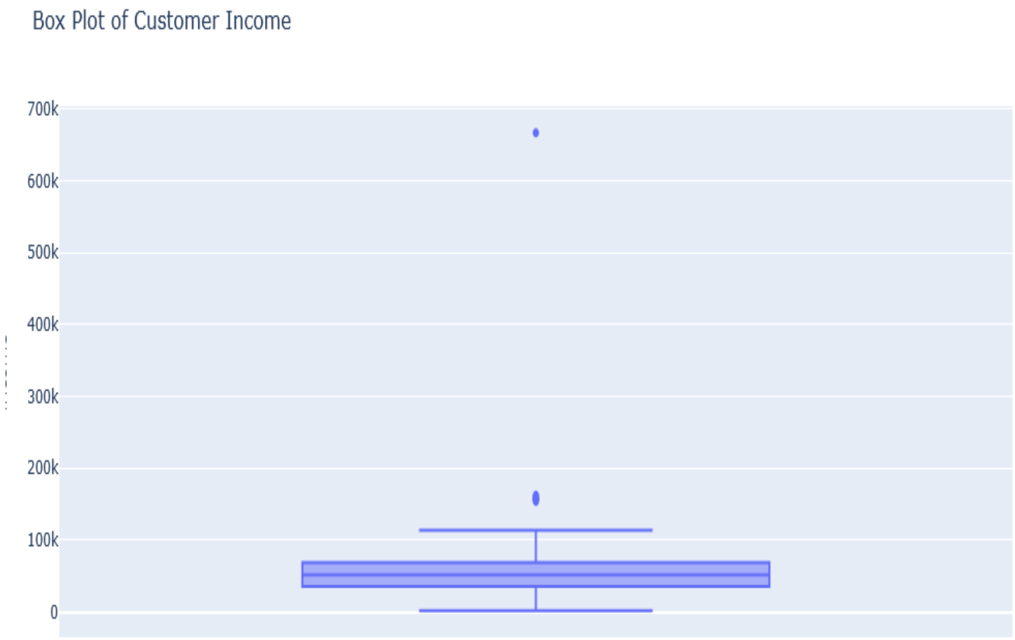
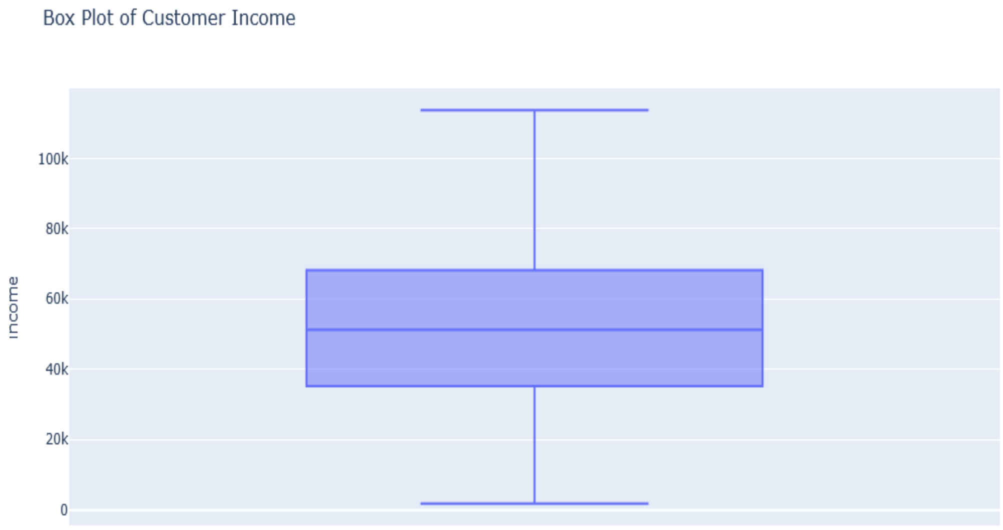
Emmanuel Pedernal

Data cleaning and Corr Matrix

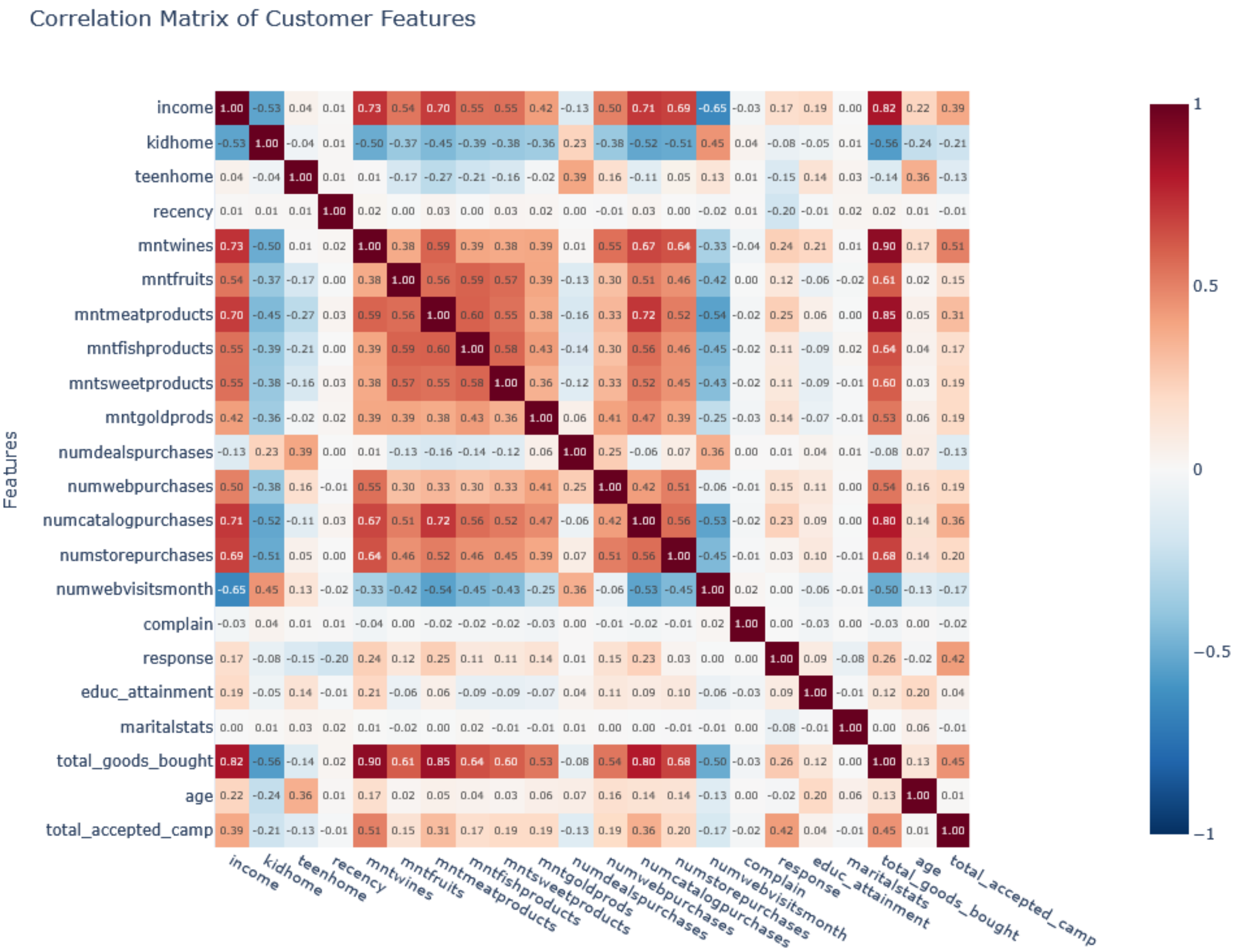
- Removed {'z_costcontact': 1, 'z_revenue': 1} due to having single value across all rows
- Changed format of dt_customer column to YYYY-MM-DD
- Added *age* column year_birth – current year
- Binned
    - Education_attainment to 0,1,2,3
    - Marital_status to binary 0 = no partner, 1 = with partner
    - Responses to single column with total number of responses
    - Total goods from (mntwines,mntgold,mntfruits…)
- Drop the following
    - 24 missing values from Income feature 1% of total values
    - ['education','marital_status','id','year_birth','dt_customer','acceptedcmp1', 'acceptedcmp2', 'acceptedcmp3', 'acceptedcmp4', 'acceptedcmp5']
- Applied a function to remove outlier for each feature (IQR)



-



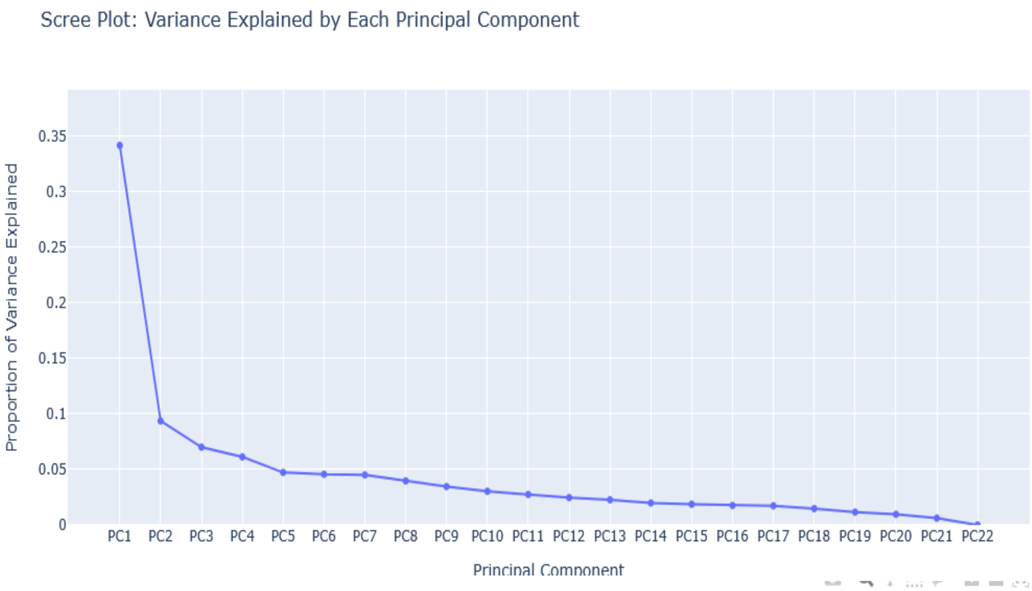
- Applied StandardScaler() to features
- Get the corr matrix to see if features are correlated


Correlation Matrix of Customer Features

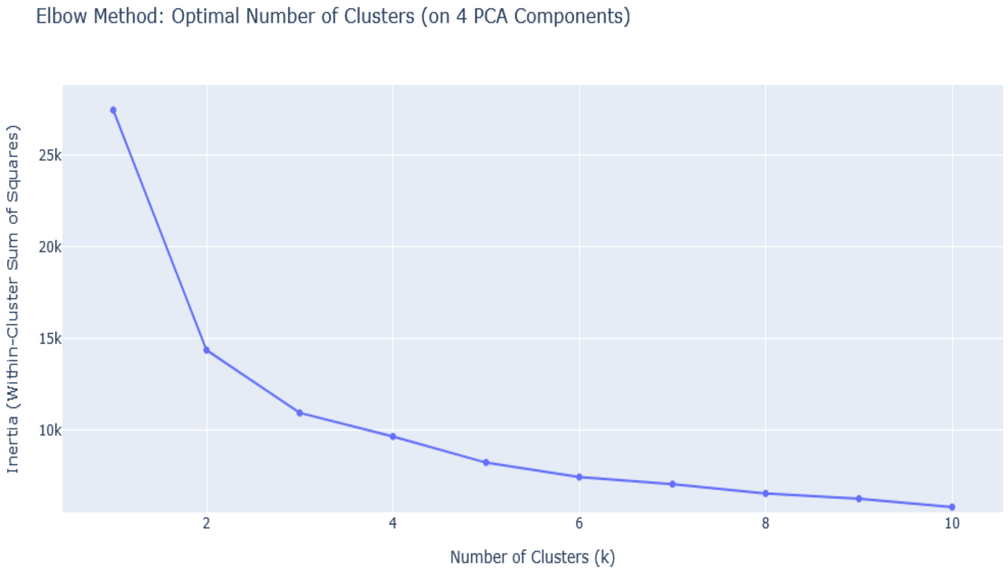
	income	kidhome	teenhome	recency	mntwines	mntfruits	mntmeatproducts	mntfishproducts	mntsweetproducts	mntgoldprods	numdealspurchases	numwebpurchases	numcatalogpurchases	numstorepurchases	numwebvisitsmonth	complain	response	educ_attainment	maritalstats	total_goods_bought	age	total_accepted_camp
income	1.00	-0.53	0.04	0.01	0.73	0.54	0.70	0.55	0.55	0.42	-0.13	0.50	0.71	0.69	-0.65	-0.03	0.17	0.19	0.00	0.82	0.22	0.39
kidhome	-0.53	1.00	-0.04	0.01	-0.50	-0.37	-0.45	-0.39	-0.38	-0.36	0.23	-0.38	-0.52	-0.51	0.45	0.04	-0.08	-0.05	0.01	-0.56	-0.24	-0.21
teenhome	0.04	-0.04	1.00	0.01	0.01	-0.17	-0.27	-0.21	-0.16	-0.02	0.39	0.16	-0.11	0.05	0.13	0.01	-0.15	0.14	0.03	-0.14	0.36	-0.13
recency	0.01	0.01	0.01	1.00	0.02	0.00	0.03	0.00	0.03	0.02	0.00	-0.01	0.03	0.00	-0.02	0.01	-0.20	-0.01	0.02	0.02	0.01	-0.01
mntwines	0.73	-0.50	0.01	0.02	1.00	0.38	0.59	0.39	0.38	0.39	0.01	0.55	0.67	0.64	-0.33	-0.04	0.24	0.21	0.01	0.90	0.17	0.51
mntfruits	0.54	-0.37	-0.17	0.00	0.38	1.00	0.56	0.59	0.57	0.39	-0.13	0.30	0.51	0.46	-0.42	0.00	0.12	-0.06	-0.02	0.61	0.02	0.15
mntmeatproducts	0.70	-0.45	-0.27	0.03	0.59	0.56	1.00	0.60	0.55	0.38	-0.16	0.33	0.72	0.52	-0.54	-0.02	0.25	0.06	0.00	0.85	0.05	0.31
mntfishproducts	0.55	-0.39	-0.21	0.00	0.39	0.59	0.60	1.00	0.58	0.43	-0.14	0.30	0.56	0.46	-0.45	-0.02	0.11	-0.09	0.02	0.64	0.04	0.17
mntsweetproducts	0.55	-0.38	-0.16	0.03	0.38	0.57	0.55	0.58	1.00	0.36	-0.12	0.33	0.52	0.45	-0.43	-0.02	0.11	-0.09	-0.01	0.60	0.03	0.19
mntgoldprods	0.42	-0.36	-0.02	0.02	0.39	0.39	0.38	0.43	0.36	1.00	0.06	0.41	0.47	0.39	-0.25	-0.03	0.14	-0.07	-0.01	0.53	0.06	0.19
numdealspurchases	-0.13	0.23	0.39	0.00	0.01	-0.13	-0.16	-0.14	-0.12	0.06	1.00	0.25	-0.06	0.07	0.36	0.00	0.01	0.04	0.01	-0.08	0.07	-0.13
numwebpurchases	0.50	-0.38	0.16	-0.01	0.55	0.30	0.33	0.30	0.33	0.41	0.25	1.00	0.42	0.51	-0.06	-0.01	0.15	0.11	0.00	0.54	0.16	0.19
numcatalogpurchases	0.71	-0.52	-0.11	0.03	0.67	0.51	0.72	0.56	0.52	0.47	-0.06	0.42	1.00	0.56	-0.53	-0.02	0.23	0.09	0.00	0.80	0.14	0.36
numstorepurchases	0.69	-0.51	0.05	0.00	0.64	0.46	0.52	0.46	0.45	0.39	0.07	0.51	0.56	1.00	-0.45	-0.01	0.03	0.10	-0.01	0.68	0.14	0.20
numwebvisitsmonth	-0.65	0.45	0.13	-0.02	-0.33	-0.42	-0.54	-0.45	-0.43	-0.25	0.36	-0.06	-0.53	-0.45	1.00	0.02	0.00	-0.06	-0.01	-0.50	-0.13	-0.17
complain	-0.03	0.04	0.01	0.01	-0.04	0.00	-0.02	-0.02	-0.02	-0.03	0.00	-0.01	-0.02	-0.01	0.02	1.00	0.00	-0.03	0.00	-0.03	0.00	-0.02
response	0.17	-0.08	-0.15	-0.20	0.24	0.12	0.25	0.11	0.11	0.14	0.01	0.15	0.23	0.03	0.00	0.00	1.00	0.09	-0.08	0.26	-0.02	0.42
educ_attainment	0.19	-0.05	0.14	-0.01	0.21	-0.06	0.06	-0.09	-0.09	-0.07	0.04	0.11	0.09	0.10	-0.06	-0.03	0.09	1.00	-0.01	0.12	0.20	0.04
maritalstats	0.00	0.01	0.03	0.02	0.01	-0.02	0.00	0.02	-0.01	-0.01	0.01	0.00	0.00	-0.01	-0.01	0.00	-0.08	-0.01	1.00	0.00	0.06	-0.01
total_goods_bought	0.82	-0.56	-0.14	0.02	0.90	0.61	0.85	0.64	0.60	0.53	-0.08	0.54	0.80	0.68	-0.50	-0.03	0.26	0.12	0.00	1.00	0.13	0.45
age	0.22	-0.24	0.36	0.01	0.17	0.02	0.05	0.04	0.03	0.06	0.07	0.16	0.14	0.14	-0.13	0.00	-0.02	0.20	0.06	0.13	1.00	0.01
total_accepted_camp	0.39	-0.21	-0.13	-0.01	0.51	0.15	0.31	0.17	0.19	0.19	-0.13	0.19	0.36	0.20	-0.17	-0.02	0.42	0.04	-0.01	0.45	0.01	1.00

PCA

- Scree plot and Elbow method to confirm number of clusters (4)

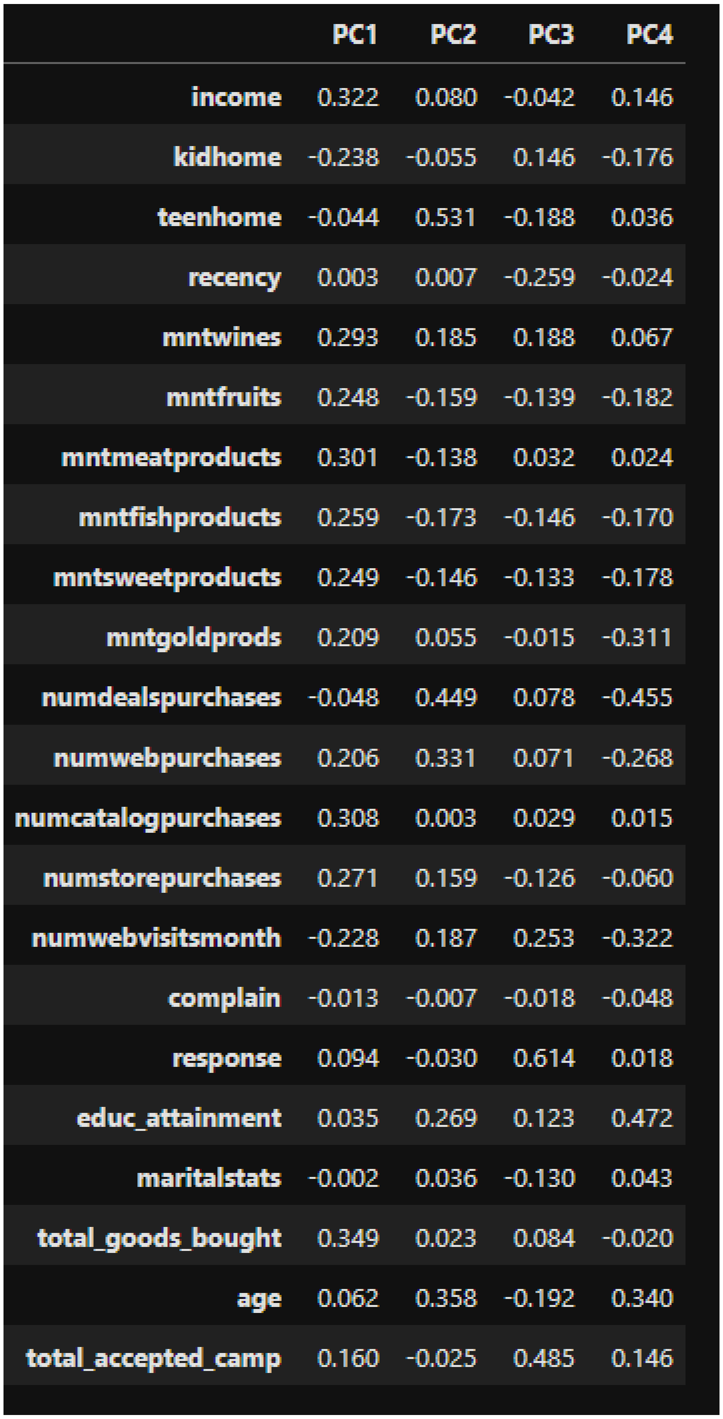

Scree Plot: Variance Explained by Each Principal Component
Proportion of Variance Explained
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
PC1
PC2
PC3
PC4
PC5
PC6
PC7
PC8
PC9
PC10
PC11
PC12
PC13
PC14
PC15
PC16
PC17
PC18
PC19
PC20
PC21
PC22
Principal Component

-


Elbow Method: Optimal Number of Clusters (on 4 PCA Components)
Inertia (Within-Cluster Sum of Squares)
25k
20k
15k
10k
2
4
6
8
10
Number of Clusters (k)

-

PCA loadings



	PC1	PC2	PC3	PC4
income	0.322	0.080	-0.042	0.146
kidhome	-0.238	-0.055	0.146	-0.176
teenhome	-0.044	0.531	-0.188	0.036
recency	0.003	0.007	-0.259	-0.024
mntwines	0.293	0.185	0.188	0.067
mntfruits	0.248	-0.159	-0.139	-0.182
mntmeatproducts	0.301	-0.138	0.032	0.024
mntfishproducts	0.259	-0.173	-0.146	-0.170
mntsweetproducts	0.249	-0.146	-0.133	-0.178
mntgoldprods	0.209	0.055	-0.015	-0.311
numdealspurchases	-0.048	0.449	0.078	-0.455
numwebpurchases	0.206	0.331	0.071	-0.268
numcatalogpurchases	0.308	0.003	0.029	0.015
numstorepurchases	0.271	0.159	-0.126	-0.060
numwebvisitsmonth	-0.228	0.187	0.253	-0.322
complain	-0.013	-0.007	-0.018	-0.048
response	0.094	-0.030	0.614	0.018
educ_attainment	0.035	0.269	0.123	0.472
maritalstats	-0.002	0.036	-0.130	0.043
total_goods_bought	0.349	0.023	0.084	-0.020
age	0.062	0.358	-0.192	0.340
total_accepted_camp	0.160	-0.025	0.485	0.146

PCA1 : "High rollers"

total_goods_bought (0.349)

income (0.322)

mntmeatproducts, mntwines, mntfruits, mntfishproducts, mntsweetproducts (≈ 0.25–0.30)

numcatalogpurchases (0.308)

numstorepurchases (0.271)

PC1 customers have spent the most and have varried product purchases;

* customers are from high income bracket due to high total product purchases and income

* Buys product from multiple channels

# Ideal targets for premium products, subscription programs, or loyalty tiers

PCA2 : "Couponing young family"

teenhome (0.531)

numdealspurchases (0.449)

age (0.358)

numwebpurchases (0.331)

educ_attainment (0.269)

PC2 customers are from families with teenagers

* active in utilizing deals * Most Purchases are over the internet (numwebpurchases)

* has moderate educational attainment

# Could be target for mobile deals/limited time offers, bundle/family offers (within age bracket), PRICE driven

PCA3 : "Marketing Engagers"

response (0.614)

total_accepted_camp (0.485)

numwebvisitsmonth (0.253)

PC3 Customers are highly responsive to marketing/ mostly interacts through web (email/ads/website)

* Best use for A/B tests

* Personalized offers

# Study behaviour further to undertand market behaviour directly from actual customers

PCA4 : "Selective Customers"

educ_attainment (0.472)

age (0.340)

income (0.146)

NEGATIVE weights

numdealspurchases (-0.455)

numwebvisitsmonth (-0.322)

mntgoldprods (-0.311)

numwebpurchases (-0.268)

PC4 Customers are selective and prone to impulse buying due to negative weights

* Age has high weight and educ_attainment which suggests the customer could be from older group ~30+

# Customers in this bracket that has positive weights could be reached through email/newspaper/reading materials

Segmentation using K-means

3D Plot of Customer Segments (First 4 PCA Components)

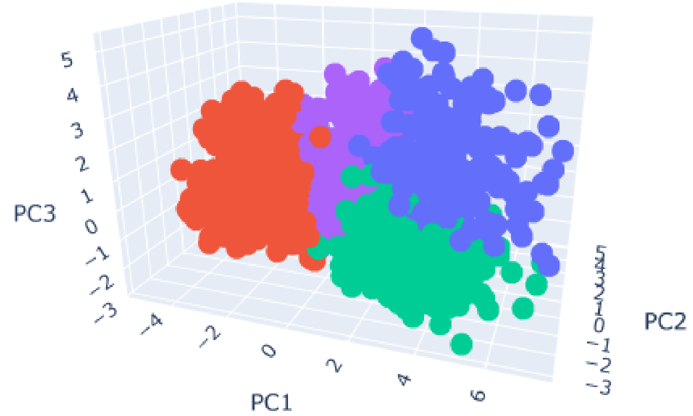
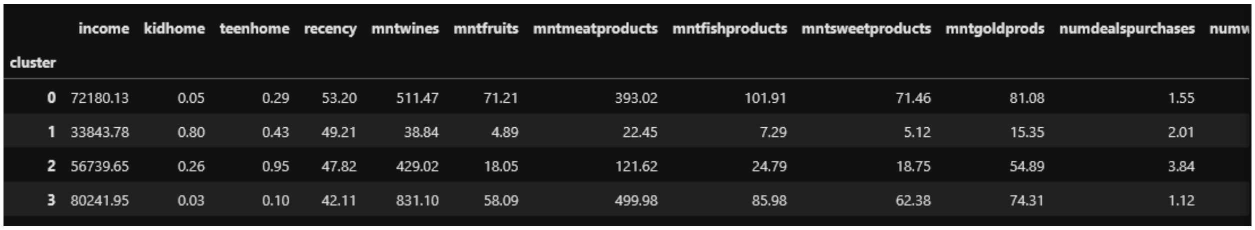


Table of the clusters



	income	kidhome	teenhome	recency	mntwines	mntfruits	mntmeatproducts	mntfishproducts	mntsweetproducts	mntgoldprods	numdealspurchases	numw
cluster												
0	72180.13	0.05	0.29	53.20	511.47	71.21	393.02	101.91	71.46	81.08	1.55	
1	33843.78	0.80	0.43	49.21	38.84	4.89	22.45	7.29	5.12	15.35	2.01	
2	56739.65	0.26	0.95	47.82	429.02	18.05	121.62	24.79	18.75	54.89	3.84	
3	80241.95	0.03	0.10	42.11	831.10	58.09	499.98	85.98	62.38	74.31	1.12	

Insights

**Values are averaged*

Cluster 0 — High rollers - Profile: Wealthy, older customers who purchase frequently and spend broadly across product categories, but rarely respond to promotions.

- Average income of 72,000 USD
- Various Item Spending: wine, meat, gold, fish, and other products
- High Total Goods Bought 1230
- Most Purchases are from in-store and online purchases
- almost no kids/teens

- Older age 57
- Does not use promotions
- Cluster Size: 440

Business application:

- Focus on loyalty rewards/ points
- Offer Cash backs per threshold of purchase
- The grocery can introduce convenient ways limited to this bracket to increase shopping experience.

Cluster 1 — Customers on a budget: Price-sensitive, possibly middle-class larger households.

- Income of 33,000 USD
- low spending
- Low total Goods Bought 94
- Cannot determine preferred way of shopping due to low items bought
- More kids (0.8) and teens (0.43)
- Low promotion response
- Cluster Size: Largest 992

Business application:

- Bundle discounts (kids/teens), essentials packs, and free delivery with thresholds.
- Promote cost-saving offers via in-store or app discounts.

Cluster 2 — General class with teens: older households with teenagers, reasonable spending and moderate promotional responsiveness.

- Income of 56,000 USD
- Most frequent items are wine, meat, and gold
- highly likely to have teen agers (0.95)
- Oldest cluster with ave age of 61
- Moderate responses to promotion (maybe due to teenagers)
- Total goods bought at moderate amount 667
- Cluster Size: 577

Business Application:

- This cluster looks like our general customers to due mix of old age + teenagers + promotion usage
- Target family-oriented promotions or convenience bundles
- Provide multi-channel offers Online/Application
- Emphasize value upgrades like family plan especially for teenagers' products

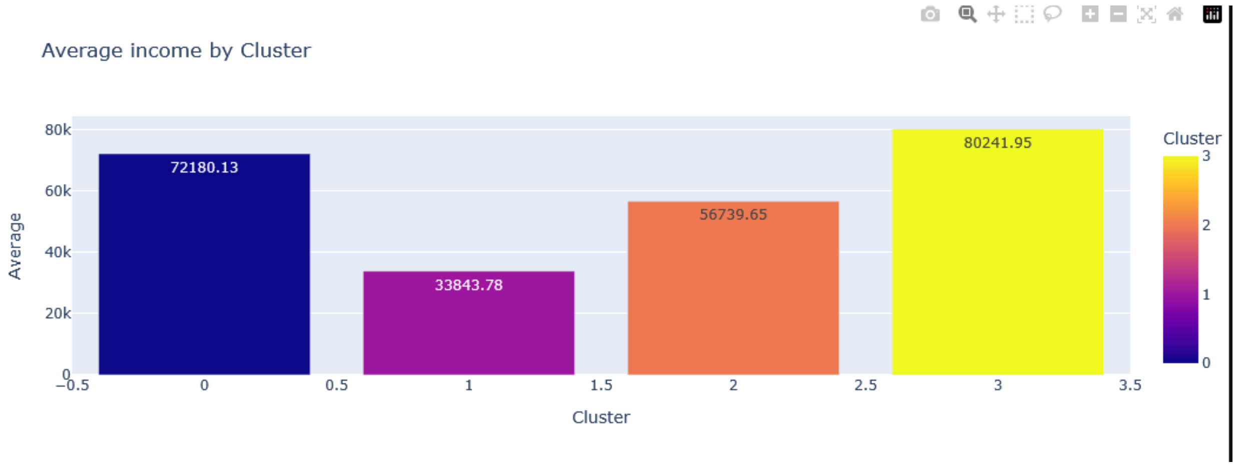
Cluster 3 — Luxury Customers: High-income, promotion-responsive customers. affluent professionals or retirees with no children.

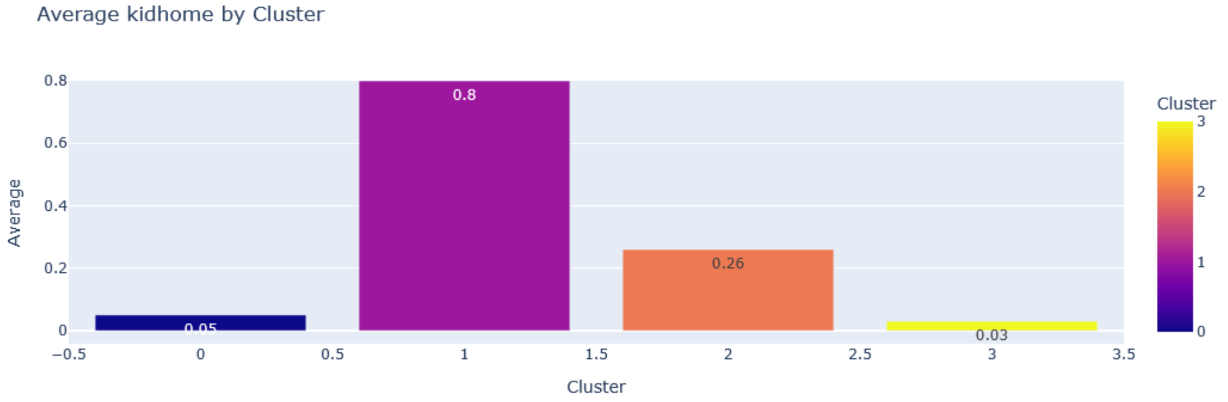
- Highest income of 80,000 USD
- Highest spending in all categories (especially wine and meat)
- Actively uses promotion (.84)
- Highest total goods bought 1612
- No kids/teen
- Median age of 56
- Cluster Size: Smallest (193)

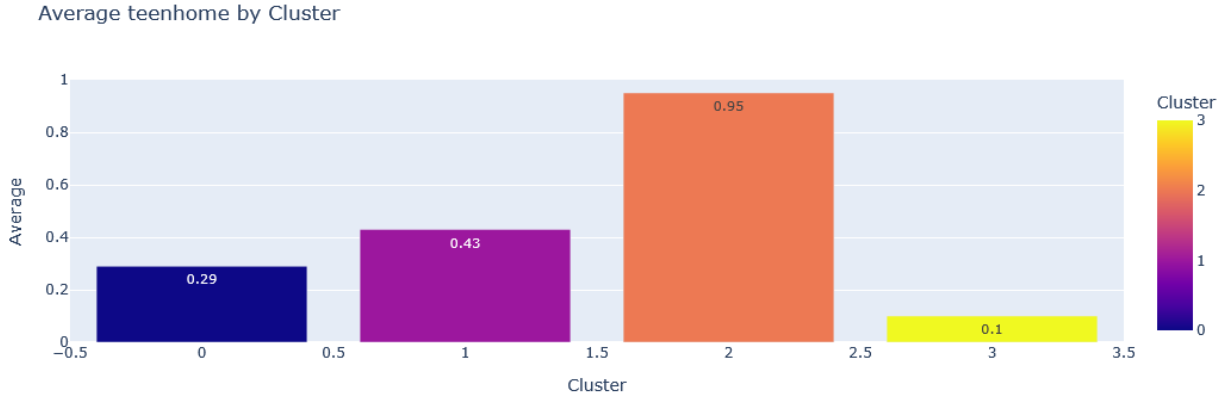
Business Application:

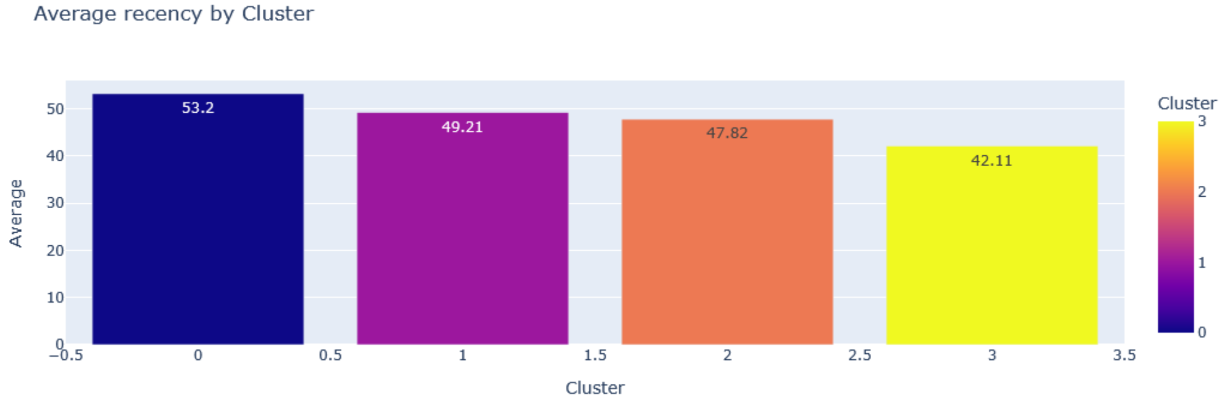
- Endorse exclusive/personalize deals to maximize customer relations
- Implement special tier for shopping convenience like priority of delivery or separate check out
- Offer discount for wine and meat or luxury items
- Lowest cluster size, see if we can market more for this customer type for maximum profit base from the clustering results

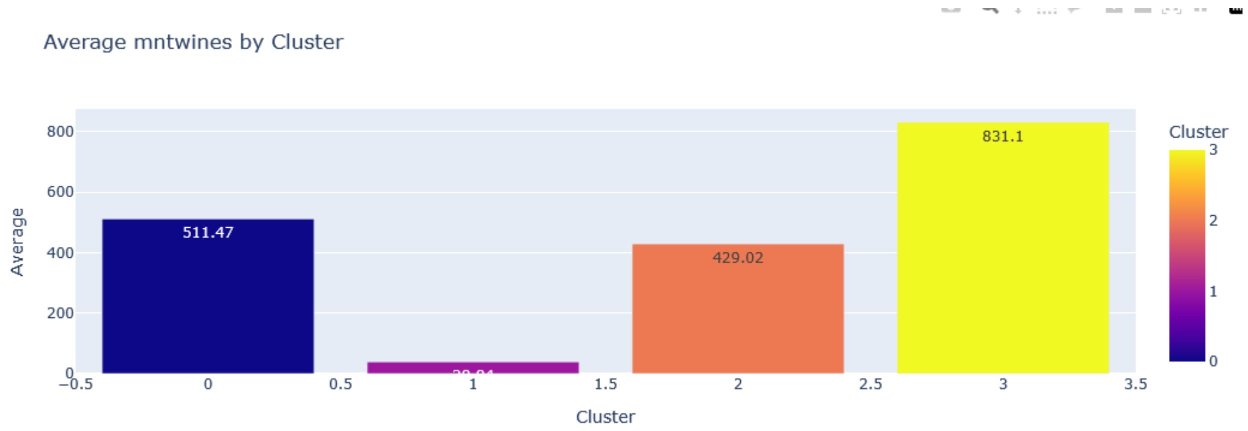
Plots per feature per cluster



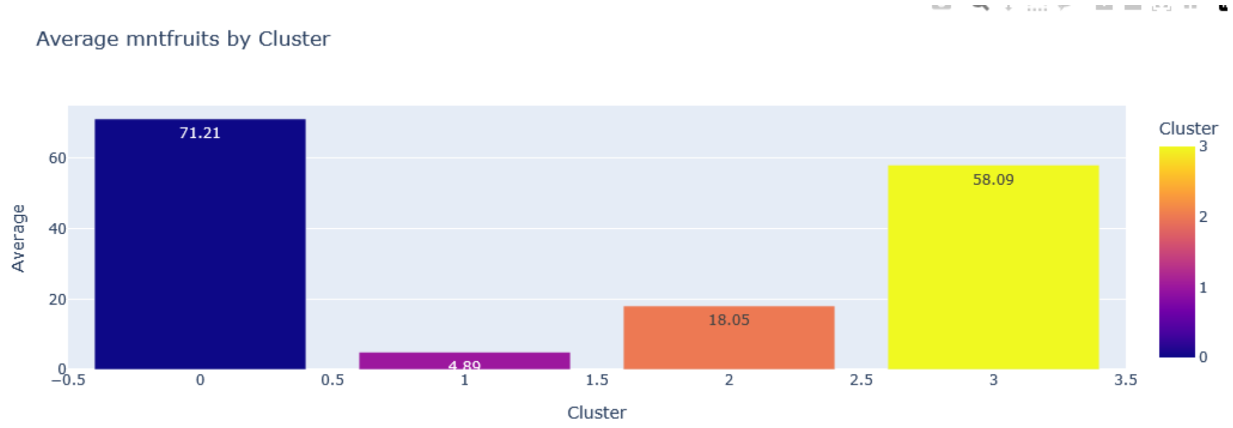

Average kidhome by Cluster


Average teenhome by Cluster

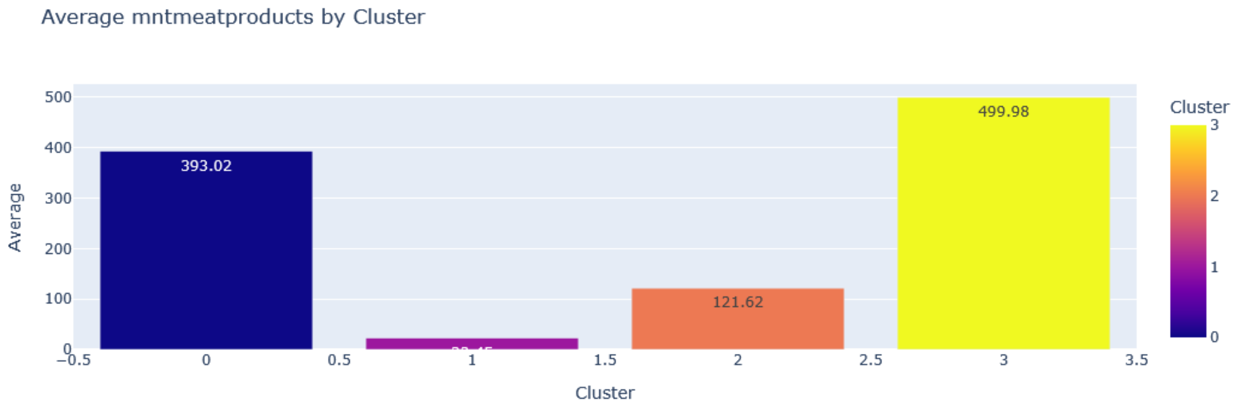

Average recency by Cluster



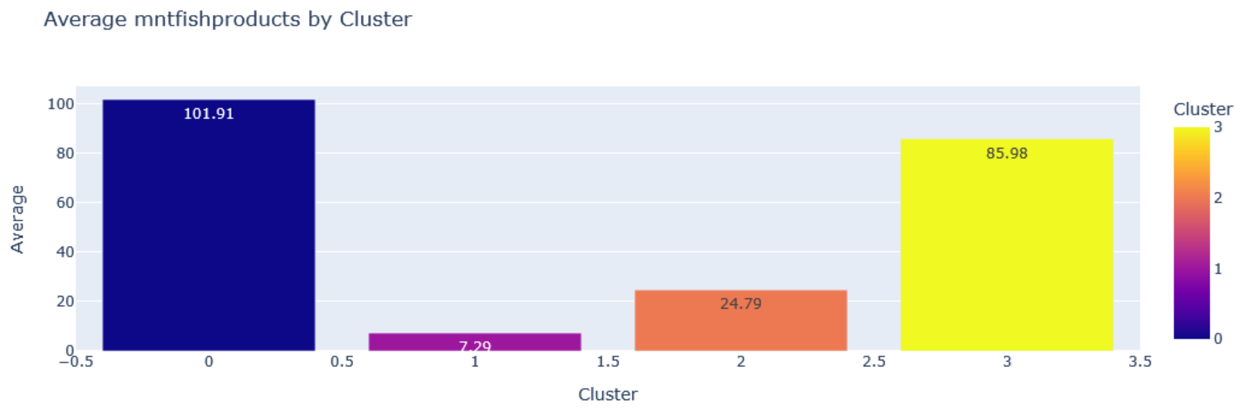
Average mntwines by Cluster

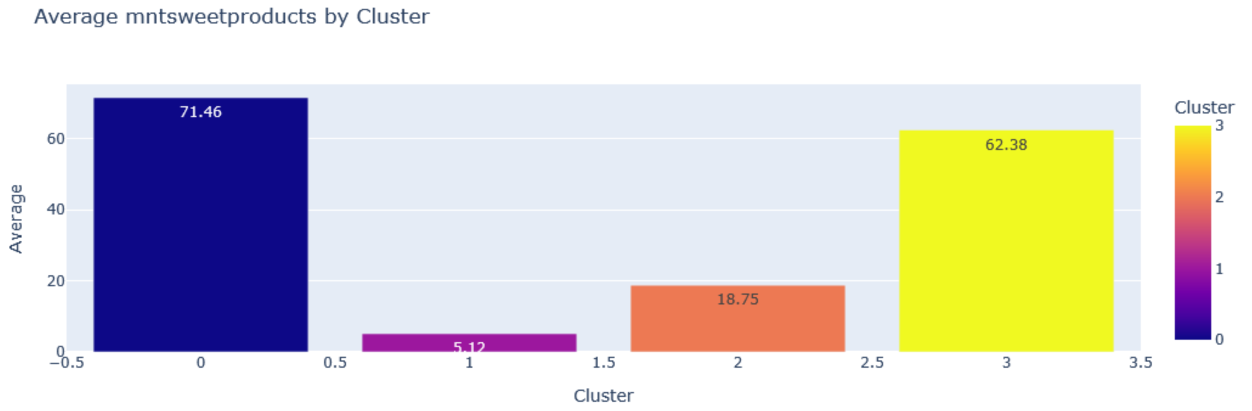


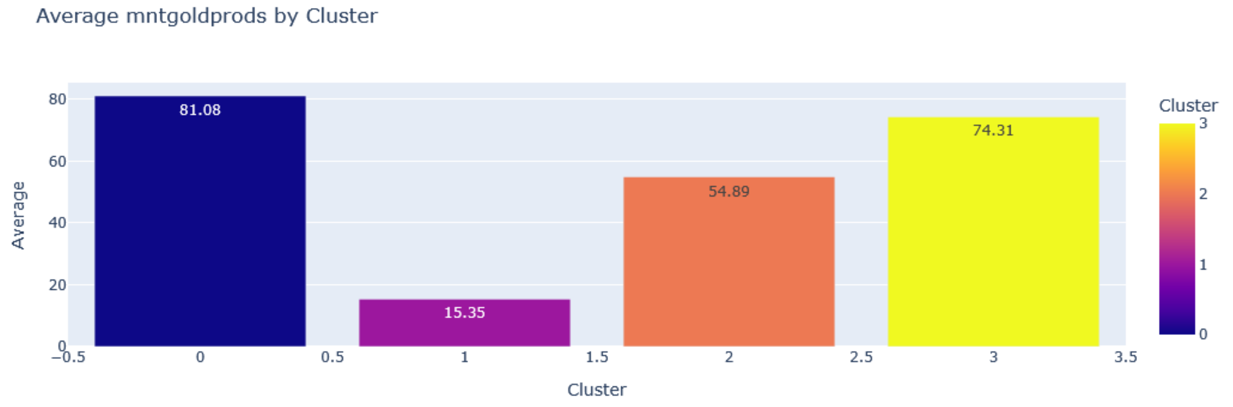
Average mntfruits by Cluster



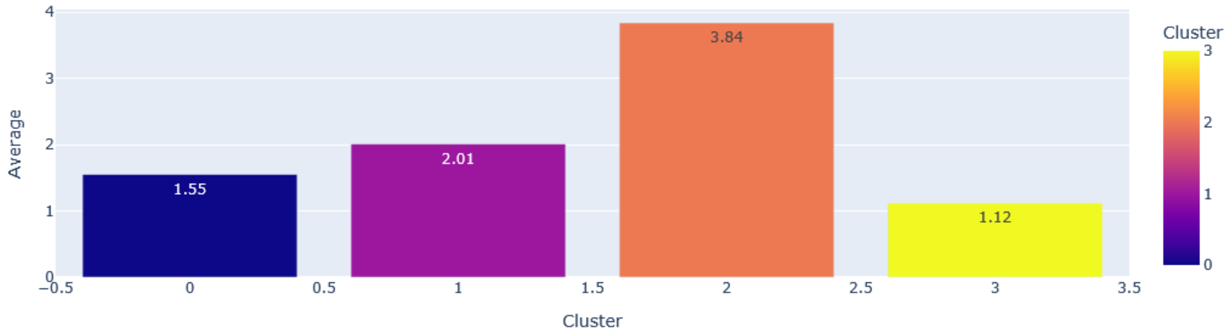
Average mntmeatproducts by Cluster


Average mntfishproducts by Cluster

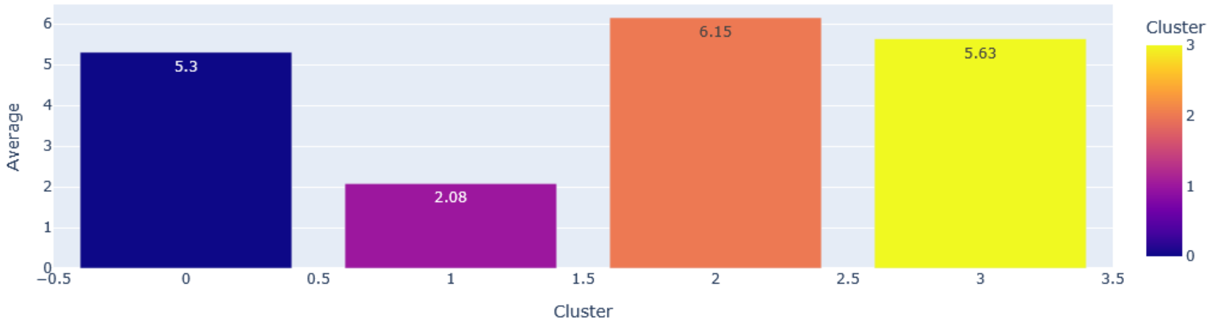

Average mntsweetproducts by Cluster


Average mntgoldprods by Cluster

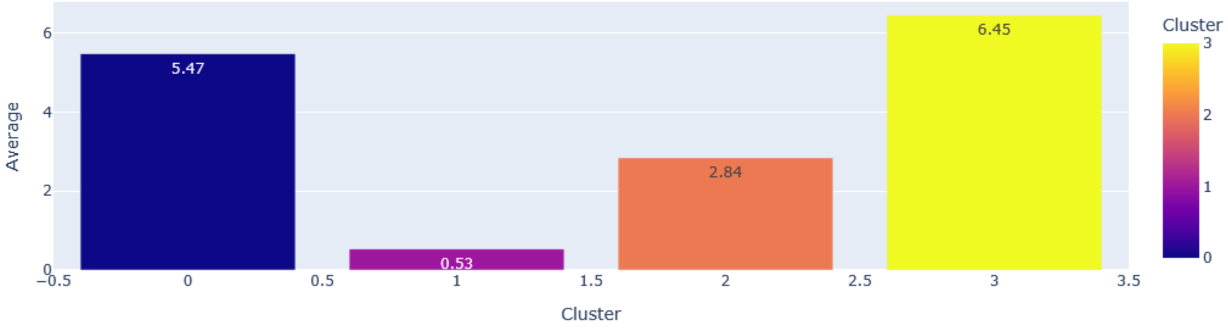
Average numdealspurchases by Cluster

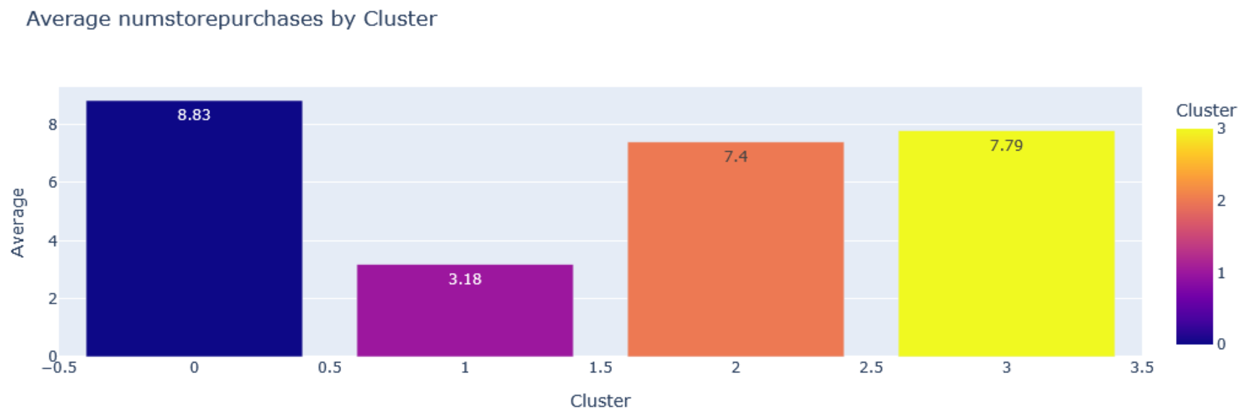


Average numwebpurchases by Cluster

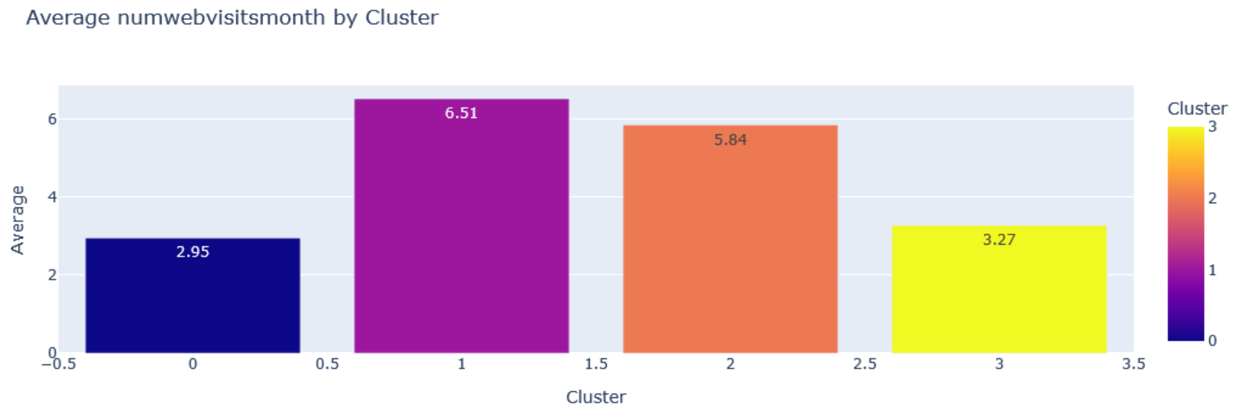


Average numcatalogpurchases by Cluster

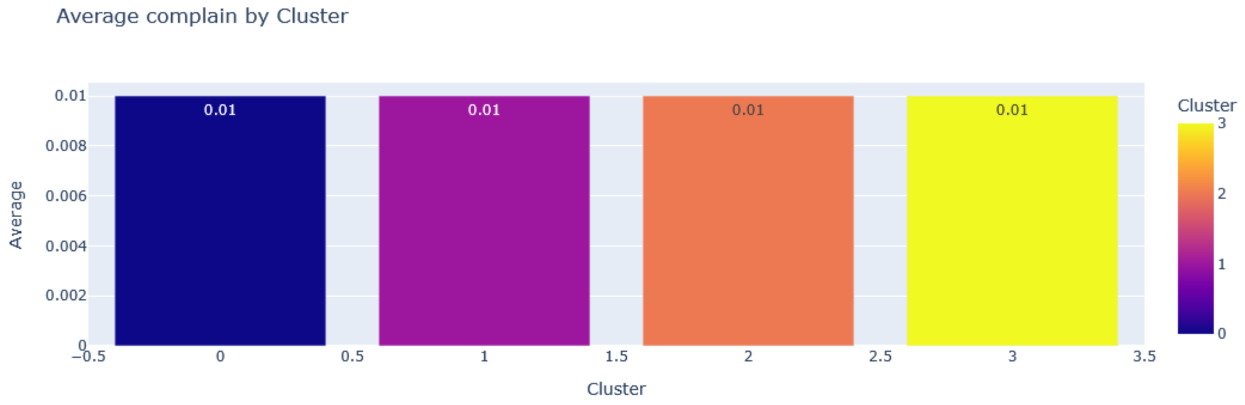




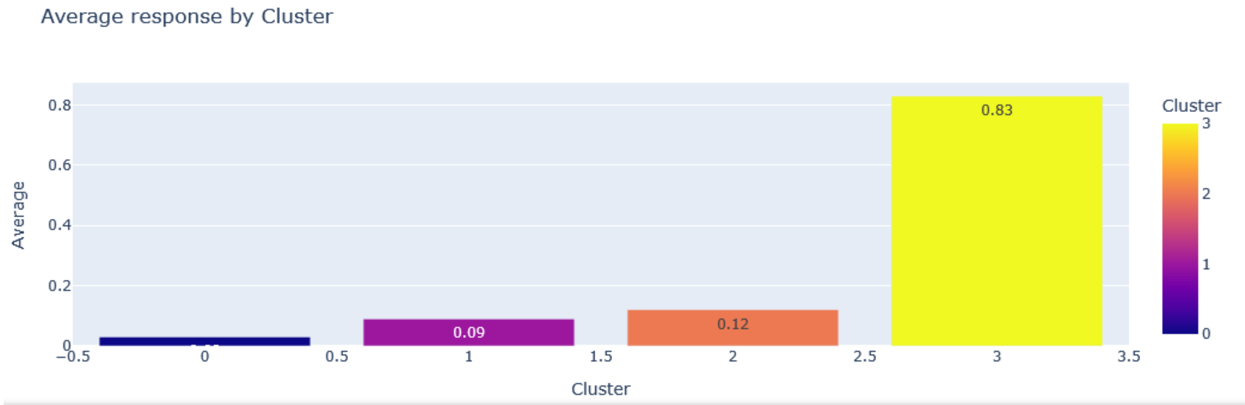
Average numstorepurchases by Cluster



Average numwebvisitsmonth by Cluster

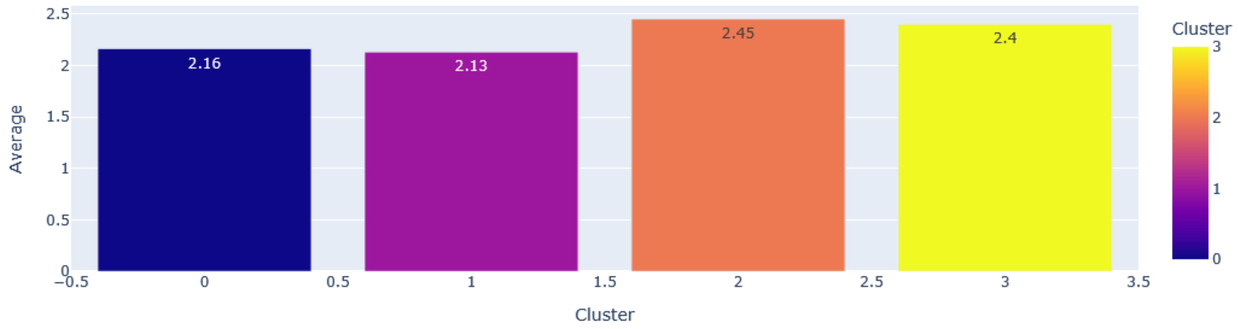


Average complain by Cluster

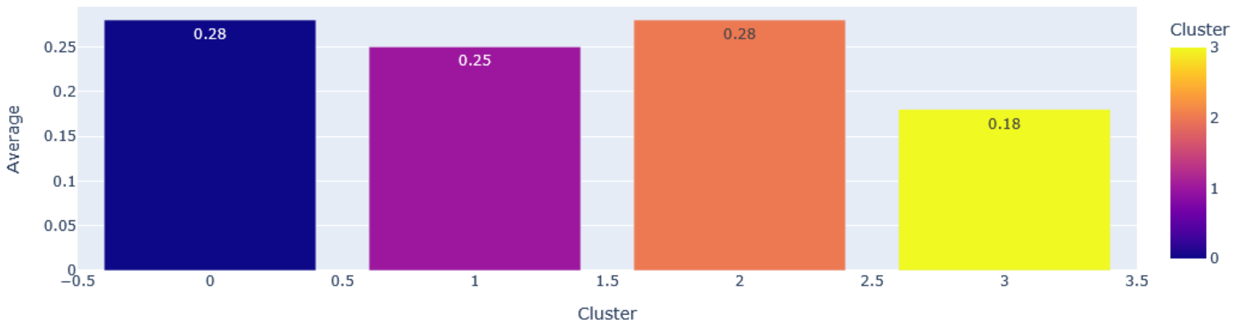


Average response by Cluster

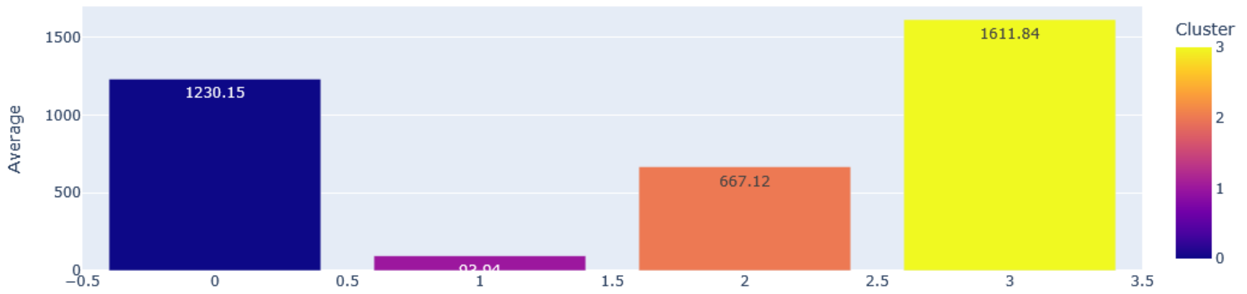
Average educ_attainment by Cluster

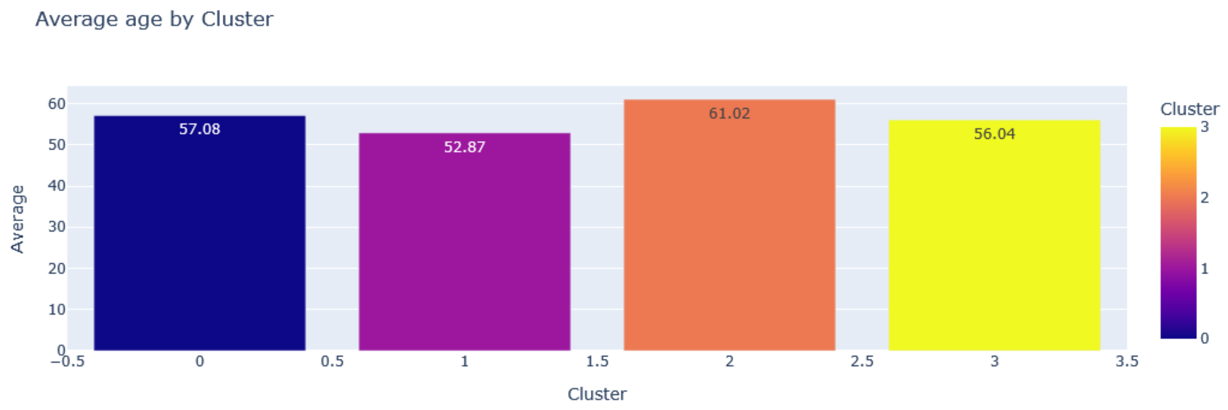


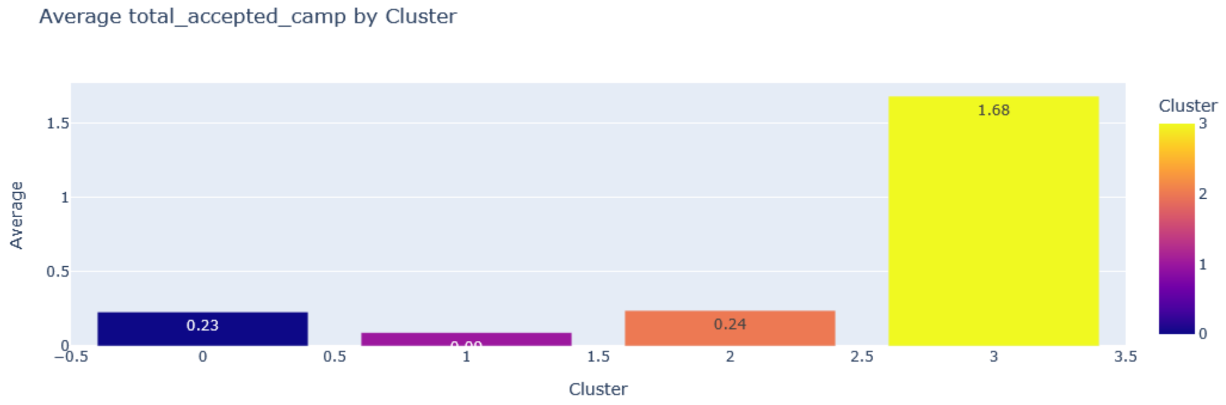
Average maritalstats by Cluster



Average total_goods_bought by Cluster




Average age by Cluster


Average total_accepted_camp by Cluster