



CUSTOMER SEGMENTATION & OPPORTUNITY ANALYSIS

MIS 382N: Marketing Analysis - I

Yash Warty, Shivarjun Sarkar, Mahika Bansal, Vivek Mehendiratta

Background



US Retail Landscape

Retail Grocery Sales (US Market)

- 2019: \$683 Bn
- 2020: \$760 Bn (+11% YoY)

Retail Ad Spend (US Market)

- 2019: \$1.51 Billion
- 2020: \$1.62 Billion (+7% YoY)



Need for Customer Centric Strategies



COVID-19 has upended the retail landscape:

- 47% of millennials are now purchasing groceries online
- 2/3 customers reported switching to cheaper brands during the pandemic
- Meat sales increased ~35% during COVID-19; frozen food increased 21%
- Fresh produce sales increased 11% in 2020 to ~\$70Bn.

With changes in customer behavior, retailers are looking for innovative techniques to improve footfall in stores and retain shoppers!

Key Objectives




Increase customer retention



Increase frequency of purchases



Improve promotional targeting



Introduction & Methodology

Introduction

About the data:

The data consists of 2240 customer records detailing their engagement for retail purchase (Kaggle) with ~30 attributes:

Demographics

- Birth Year
- Education level
- Marital status
- Income
- Family members (kids or teens at home)

Engagement History

- Enrollment for retail purchase
- Recency of last purchase
- Complaints (Binary)

Purchase History

- Sales of:
 - Wines
 - Fruits
 - Meat Products
 - Fish products
 - Sweet Products
 - Gold

Purchase Channels

- Purchases via:
 - Web
 - Catalog
 - Stores
- Website Visits in the last month

Promotions Available

- Purchases on deals
- Campaigns availed (past 5 campaigns; binary)
- Cost and revenue associated with last campaign

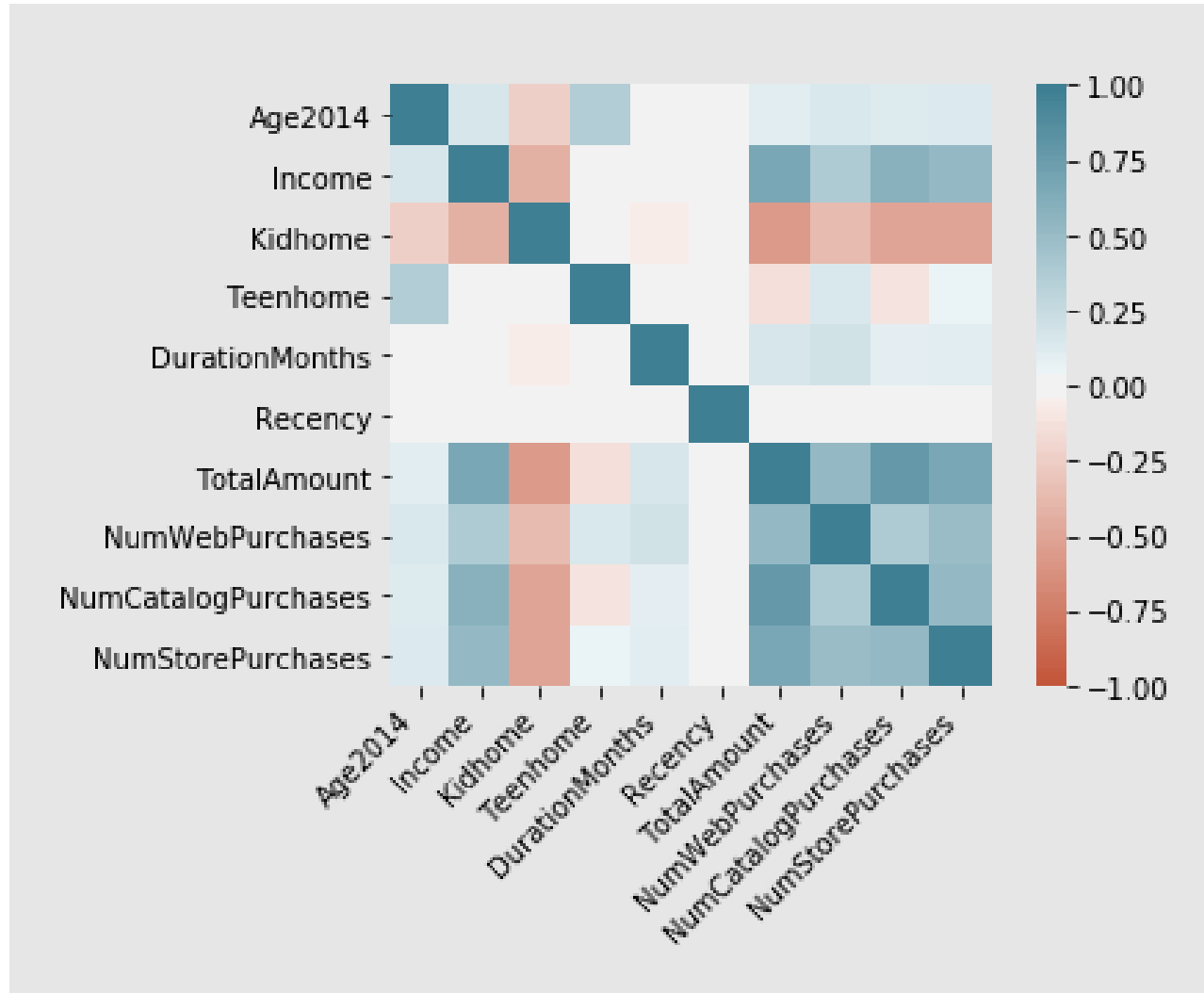
Methodology



A person wearing a blue and white striped apron is holding a cardboard box filled with a variety of fresh fruits and vegetables. The produce includes bunches of green asparagus, several bright red tomatoes, green apples, yellow bananas, orange carrots, a head of green lettuce, yellow onions, and a bunch of green herbs. The scene is set against a dark background, with a small yellow rectangular graphic element in the top left corner.

Preliminary Analysis

Correlation Analysis

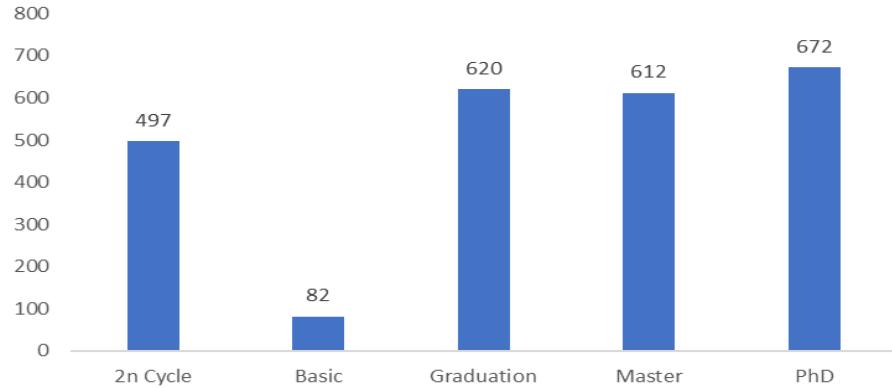


Key Observations:

- Higher income translates to higher purchases, especially in catalog purchases
- With kids at home, disposable income tends to decrease slightly and thus lower purchases consistently in all the purchasing channels
- People with kids at home seem to be in the younger age brackets
- Purchases within different channels seem to be correlated, thus showing no substitution

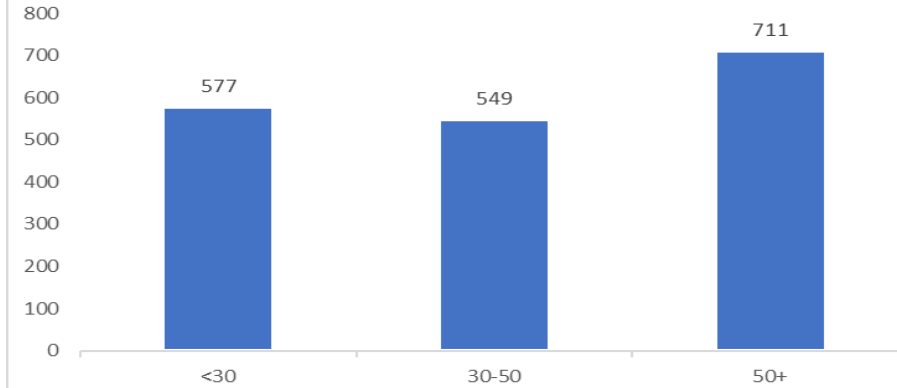
Exploratory Data Analysis

Avg Sales by Education



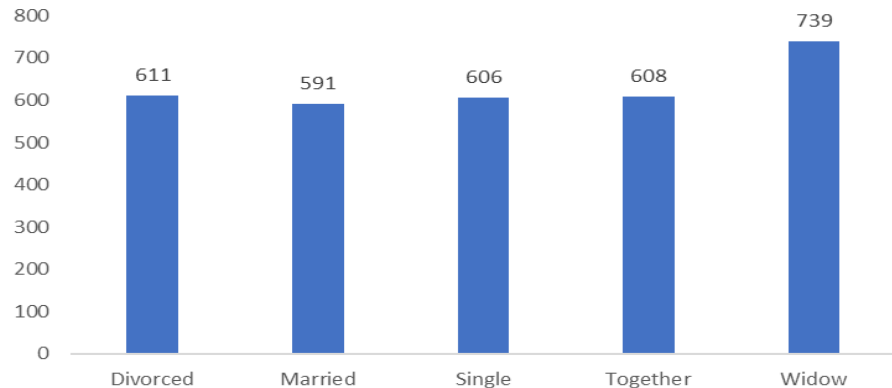
Average sales tends to increase with education level, highest is that of PhD at 672

Avg Sales by Age



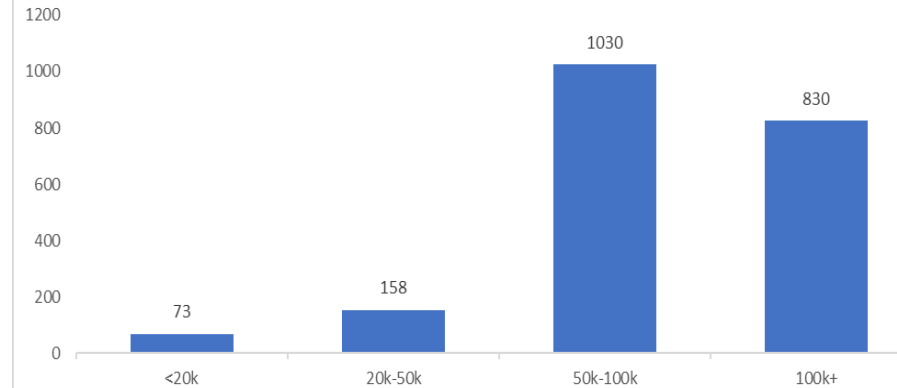
Average sales tends to be highest for older age brackets (50+) at 711

Avg Sales by Marital Status



Average sales tend to be highest for the widowed segment, and lowest for married

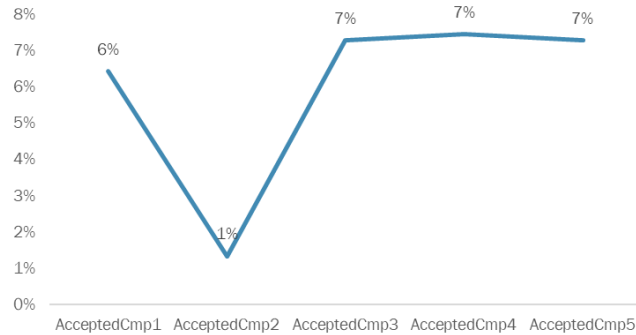
Avg Sales by Income



Average sales tend to be highest (1030) for people with income between 50-100k

Exploratory Data Analysis

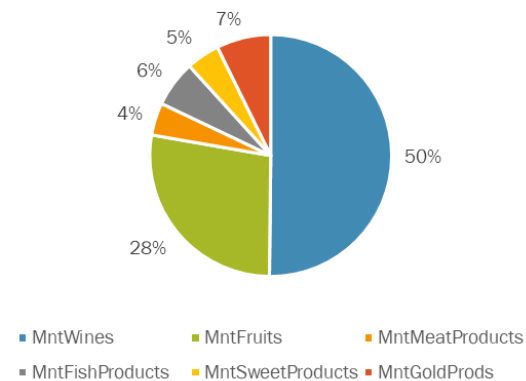
Success Rate - Past Campaigns



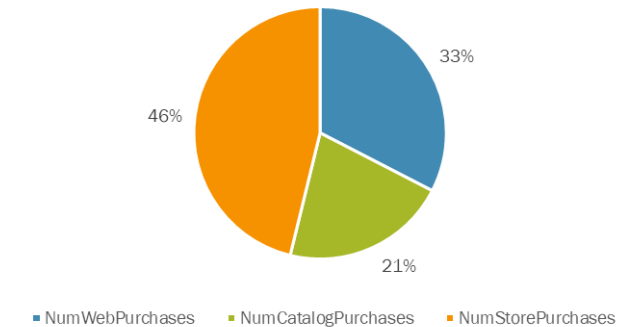
- Success rate of previous campaigns around 6-7% on average
- Campaign 2 achieved only 1% conversion of all the customer records
- With such rates, targeting is imperative for better response and profitability

- Almost 50% of the share of retail sales is from Wines
- Fruits are second with 28% share of sales
- Fish, meat, sweet products and gold seem to have almost similar shares of sales (4-7%)

Amount Spent on Products



Purchase Medium



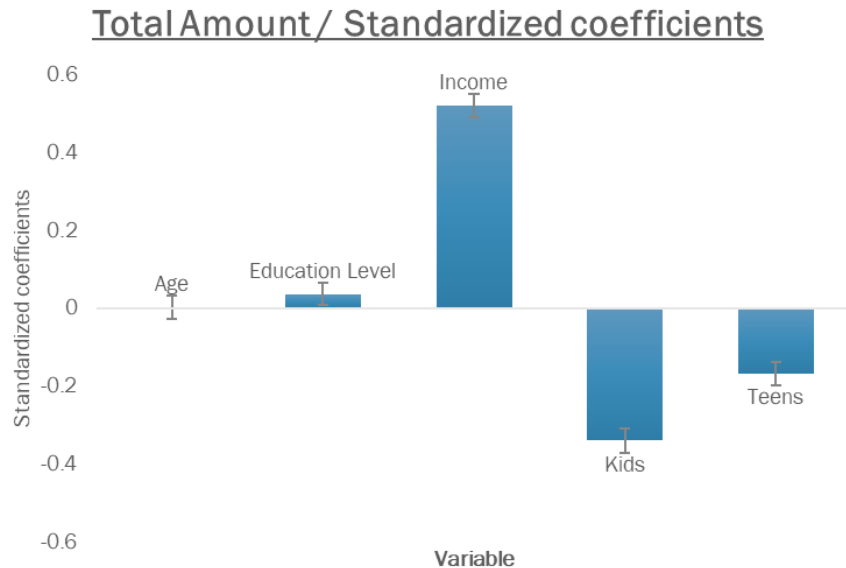
- In-store purchases tend to be the most popular with almost half of total sales
- Web Purchases are the next highest with 33% and this will increase due to increased online shopping with Covid-19

A top-down view of several antique silver spoons arranged on a dark, textured surface. The spoons are filled with various spices: one with coarse white salt, another with a mix of red and white peppercorns, and a third with bright orange-red powder. Scattered around the spoons are other spices like star anise, cinnamon sticks, and small red berries. A piece of twine is tied around the handles of the spoons.

Data Modelling & Insights

Key Drivers for Retail Purchase

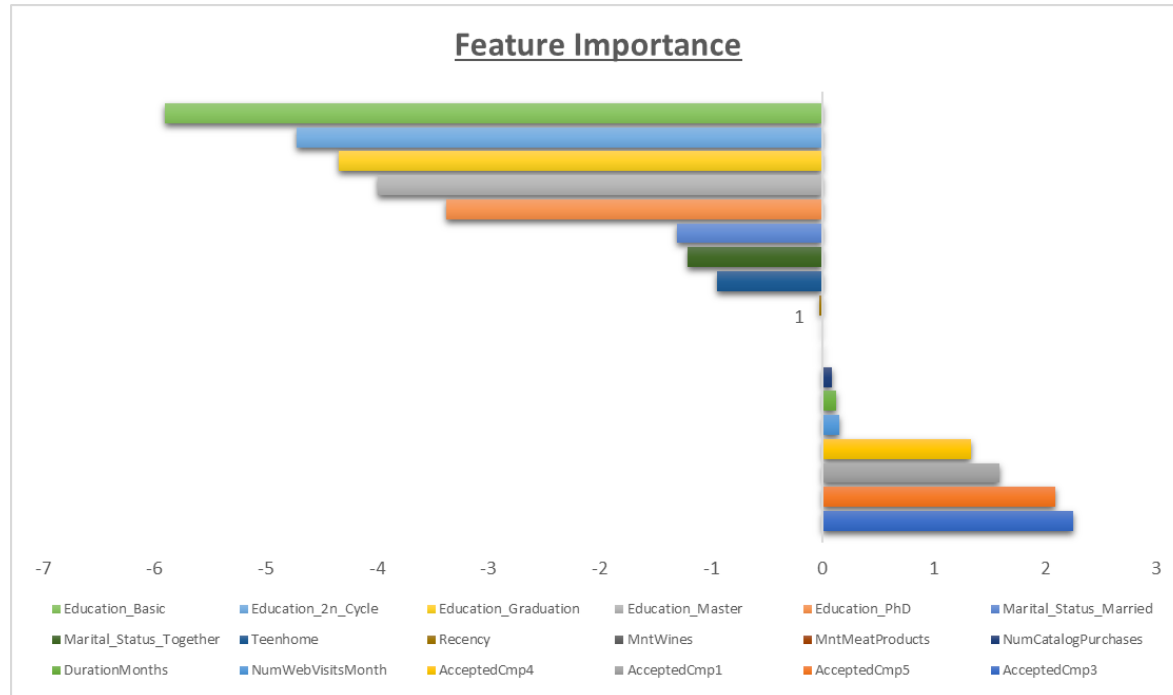
Source	Value	Standard error	t	Pr > t	Lower bound (95%)	Upper bound (95%)
Age	0.002	0.016	0.136	0.892	-0.029	0.033
Education Level	0.036	0.014	2.515	0.012	0.008	0.065
Income	0.519	0.016	33.012	<0.0001	0.489	0.550
Kids	-0.339	0.016	-21.418	<0.0001	-0.371	-0.308
Teens	-0.167	0.015	-11.098	<0.0001	-0.197	-0.138



Family size, Education and Income primarily affect purchase decisions

- Age (though not statistically significant), Education and Income are positively correlated with total amount spent and deals purchased, whereas number of kids and teens are inversely correlated
- Wines move in tandem with overall spend, whereas fruit, fish, sweet & gold product purchases are dominated by 2n-cycle and graduated segments
- Increasing number of web visits along with income led to higher web purchases, whereas kids and teens contributed to lower purchase probabilities
- Along with wines, fish & sweets, web & store purchases are highly dependent on fruit purchases. Additionally, catalog purchases rely more on meat purchases

Targeting via List Scoring



Response in last campaign was most correlated with:

- Customer Demographics:
 - **Positive*:** Kids at home (price sensitivity)
 - **Negative:** Education Level (highly educated people avail for more deals; awareness might be an issue), teens at home, marital status (together & married)
- Past Purchase/ engagement:
 - **Positive:** Previous campaigns' response (especially Campaign 3 & 5), web visits and duration for which the person has been engaged for; catalog and meat purchases
 - **Negative:** Wine purchases, recency (older the last purchase, less likely to respond to campaigns – retention can be a concern)

Taking into account probabilities of response by logistic regression, cost of campaign and revenue expected:

Cases	Cost per Customer	Revenue per Customer	Actual Response	Customers in Focus	Customers%	Total Cost	Total Revenue	Revenue%	Profitability
No Targeting	3	11	334	2237	100%	6711	3674	100%	-83%
Targeting	3	11	334	381	17%	1143	2387	65%	52%

By targeting 17% of the base, we can capture 65% of the revenue and become profitable

Targeting with RFM Analysis

Recency

Row Labels	Average of Bought Offer
1	0.347490347
2	0.244541485
3	0.276119403
4	0.237704918
Grand Total	0.278

Class 1 of recent purchasers are most likely to accept our promotions (~35%)

Frequency

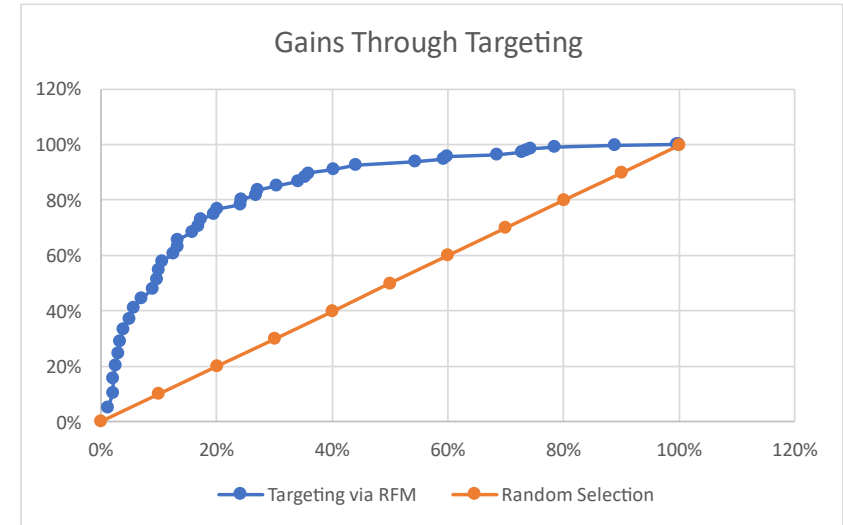
Row Labels	Average of Bought Offer
1	0.411764706
2	0.424137931
3	0.292775665
4	0.143939394
Grand Total	0.278

Based on frequency, Class 1 & 2 customers are equally likely to accept promotions (~42%)

Monetary

Row Labels	Average of Bought Offer
1	0.846153846
2	0.5546875
3	0.329004329
4	0.162790698
Grand Total	0.278

~85% of our highest spenders are likely to use promotions while making purchases



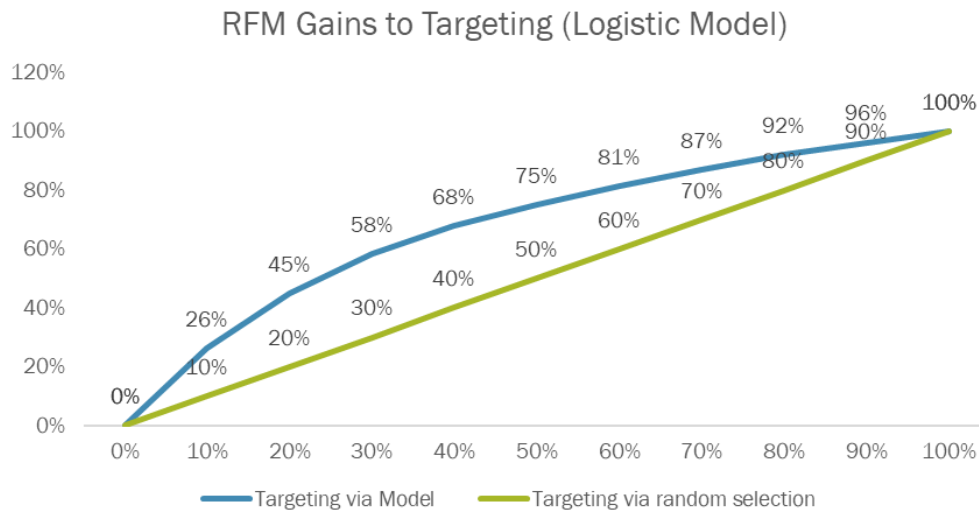
- By segmenting customers based on RFM, we notice windfall gains from targeting.
- From our holdout set of 1240 observations, the overall customers using promotions were 154 (~12%).
- Targeting achieves **90%** of revenues from random selection by targeting **~40%** of customers.
- Hence, rather than sending broad promotions, retailers could benefit from tailoring offers to specific segments

Comparing RFM with Logistic Regression

Source	Value	Standard error	Wald Chi-Square	Pr > Chi²
Intercept	-1.265621602	0.215780728	34.40190147	4.48285E-09
overall_purchase	-0.052818508	0.01835659	8.279197555	0.004010185
Recency	-0.008442057	0.002742828	9.473259579	0.002084886
overall_spend	0.002077	0.000223761	86.1599036	0

- On running a regression model for promotions bought w.r.t. the RFM market data, we notice that all of the variables are significant, as denoted by the p-value

- From our total dataset of 1240 customer profiles, about 349 are likely to accept our promotions and make purchases
- Targeting with logistic regression achieves 90% of revenue from ~70% of customers
- This further reiterates that tailoring promotional materials might lead to a higher engagement and purchase rate



Customer Segmentation & Clustering

Class	Education Level	Income	Kids	Teens	Total Amount	Sum of weights	Within-class variance
1	3.645	53018.300	0.384	0.875	482.969	487.000	19983099.989
2	3.550	69267.108	0.110	0.551	1107.060	564.000	25884731.284
3	2.842	21246.689	0.772	0.181	72.200	360.000	38367315.447
4	3.554	37453.409	0.790	0.513	150.299	552.000	19810454.056
5	3.577	89210.506	0.063	0.138	1489.067	253.000	1503011829.048

- On running a clustering analysis for our data, we obtained 5 distinct customer classes:
 - Class 1: Highly Educated, Middle Income families. They could be price conscious as spending amounts are low.
 - Class 2: Highly Educated, High Income families. They spend more as they may be buying higher quantities.
 - Class 3: Less Educated, Low-Income families. They could be spending less due to financial constraints.
 - Class 4: Highly Educated, Middle Income families. They spend moderately as they may buy private label products.
 - Class 5: Highly Educated, High Income customers. They do not have families and may spend on luxury items like gourmet foods, expensive wines, etc.

Retailers must choose their customer niche and target accordingly.

Statistical Significance of Clusters

Cluster 1 vs Rest

We observe Cluster 1 is significantly different from cluster 2,3,4. However the p-values for cluster 5 seem quite high and cannot be considered

Source	Value	Standard error	t	Pr > t
Intercept	-1050.471	452.827	-2.320	0.021
Education Level	-4.065	15.728	-0.258	0.796
Income	0.028	0.004	8.130	<0.0001
Kids	-218.021	29.438	-7.406	<0.0001
Teens	-131.472	34.300	-3.833	0.000
Class	236.933	295.030	0.803	0.422
S2	0.000	0.000		
Education x S2	47.550	22.073	2.154	0.031
Income x S2	-0.004	0.005	-0.921	0.357
Kids x S2	-31.085	52.604	-0.591	0.555
Teens x S2	-93.125	43.858	-2.123	0.034

Source	Value	Standard error	t	Pr > t
Education Level	-0.013	0.035	-0.375	0.708
Income	1.383	0.117	11.787	<0.0001
Kids	-0.351	0.033	-10.737	<0.0001
Teens	-0.211	0.038	-5.557	<0.0001
Class	1.345	0.213	6.316	<0.0001
S3	0.000	0.000		
Education x S3	0.047	0.074	0.640	0.522
Income x S3	-0.993	0.104	-9.528	<0.0001
Kids x S3	0.273	0.048	5.705	<0.0001
Teens x S3	0.097	0.032	3.058	0.002

Source	Value	Standard error	t	Pr > t
Intercept	-1024.662	179.928	-5.695	<0.0001
Education Level	-4.065	10.538	-0.386	0.700
Income	0.028	0.002	12.135	<0.0001
Kids	-218.021	19.724	-11.054	<0.0001
Teens	-131.472	22.981	-5.721	<0.0001
Class	211.124	54.244	3.892	0.000
S4	0.000	0.000		
Education x S4	0.780	14.832	0.053	0.958
Income x S4	-0.016	0.003	-5.029	<0.0001
Kids x S4	117.210	28.194	4.157	<0.0001
Teens x S4	72.112	29.703	2.428	0.015

Source	Value	Standard error	t	Pr > t
Intercept	-1444.140	262.447	-5.503	<0.0001
Education Level	-4.065	16.570	-0.245	0.806
Income	0.028	0.004	7.717	<0.0001
Kids	-218.021	31.014	-7.030	<0.0001
Teens	-131.472	36.135	-3.638	0.000
Class	630.602	58.215	10.832	<0.0001
S5	0.000	0.000		
Education x S5	17.026	29.095	0.585	0.559
Income x S5	-0.031	0.004	-8.319	<0.0001
Kids x S5	20.250	109.410	0.185	0.853
Teens x S5	-0.670	74.334	-0.009	0.993

Statistical Significance of Clusters

Cluster 2 vs Rest

We observe Cluster 2 is significantly different from cluster 3 & 4.

Also, we have already seen that it is different from 1. However, the p-values for cluster 5 seem quite high again and cannot be considered

Source	Value	Standard error	t	Pr > t
Education Level	0.074	0.023	3.166	0.002
Income	0.955	0.103	9.264	<0.0001
Kids	-0.210	0.033	-6.443	<0.0001
Teens	-0.193	0.021	-9.265	<0.0001
Class	0.551	0.164	3.362	0.001
S3	0.000	0.000		
Education x S3	-0.094	0.052	-1.805	0.071
Income x S3	-0.468	0.067	-6.991	<0.0001
Kids x S3	0.173	0.040	4.286	<0.0001
Teens x S3	0.092	0.020	4.489	<0.0001

Source	Value	Standard error	t	Pr > t
Education Level	0.085	0.036	2.372	0.018
Income	1.202	0.173	6.940	<0.0001
Kids	-0.159	0.033	-4.827	<0.0001
Teens	-0.251	0.036	-6.941	<0.0001
Class	2.201	0.273	8.071	<0.0001
S4	0.000	0.000		
Education x S4	-0.110	0.118	-0.938	0.348
Income x S4	-2.607	0.345	-7.551	<0.0001
Kids x S4	0.015	0.037	0.399	0.690
Teens x S4	0.042	0.037	1.158	0.247

Source	Value	Standard error	t	Pr > t
Intercept	-973.043	385.929	-2.521	0.012
Education Leve	43.485	13.028	3.338	0.001
Income	0.024	0.002	9.767	<0.0001
Kids	-249.106	36.673	-6.793	<0.0001
Teens	-224.597	22.992	-9.768	<0.0001
Class	198.219	109.163	1.816	0.070
S4	0.000	0.000		
Education x S4	-46.770	18.478	-2.531	0.012
Income x S4	-0.012	0.004	-3.227	0.001
Kids x S4	148.295	44.550	3.329	0.001
Teens x S4	165.237	32.968	5.012	<0.0001

Statistical Significance of Clusters

Cluster 3 vs 4/5

We observe Cluster 3 is significantly different from cluster 1, 2, 4 but not from 5

Source	Value	Standard error	t	Pr > t
Education Level	0.043	0.047	0.918	0.359
Income	-0.102	0.073	-1.405	0.160
Kids	-0.100	0.047	-2.116	0.035
Teens	-0.031	0.061	-0.507	0.613
Class	-0.950	0.203	-4.671	<0.0001
S4	0.000	0.000		
Education x S4	-0.119	0.114	-1.041	0.298
Income x S4	1.683	0.217	7.752	<0.0001
Kids x S4	-0.249	0.067	-3.697	0.000
Teens x S4	-0.159	0.068	-2.334	0.020

Cluster 4 vs 5

We observe Cluster 4 is significantly different from cluster 1, 2, 3 but we cannot be sure about 5

Source	Value	Standard error	t	Pr > t
Intercept	-7736.318	668.426	-11.574	<0.0001
Education Level	-3.285	12.979	-0.253	0.800
Income	0.012	0.003	4.284	<0.0001
Kids	-100.811	25.052	-4.024	<0.0001
Teens	-59.360	23.402	-2.537	0.011
Class	1889.038	142.303	13.275	<0.0001
S5	0.000	0.000		
Education x S5	16.246	22.938	0.708	0.479
Income x S5	-0.015	0.003	-5.145	<0.0001
Kids x S5	-96.960	86.677	-1.119	0.264
Teens x S5	-72.782	56.452	-1.289	0.198

Source	Value	Standard error	t	Pr > t
Intercept	-2276.364	219.093	-10.390	<0.0001
Education Level	6.270	15.605	0.402	0.688
Income	-0.002	0.003	-0.615	0.539
Kids	-31.072	33.548	-0.926	0.355
Teens	-9.307	41.984	-0.222	0.825
Class	797.047	56.150	14.195	<0.0001
S5	0.000	0.000		
Education x S5	6.691	25.704	0.260	0.795
Income x S5	-0.001	0.003	-0.371	0.711
Kids x S5	-166.699	95.683	-1.742	0.082
Teens x S5	-122.834	69.574	-1.766	0.078

A man with dark, curly hair is standing in a grocery store aisle. He is wearing a dark blue long-sleeved shirt under a grey zip-up jacket. He is holding a smartphone in his right hand, looking down at it with a frustrated expression. His left hand is scratching the back of his head. The background shows shelves stocked with various fruits and vegetables, including apples and oranges. The lighting is warm and typical of a grocery store.

Observations & Recommendations

Key Takeaways & Recommendations

Observations



Co-Joint Analysis

- Education, Family Size and Income significantly impact retail amount spent
- 2ncycle & graduated customers seem to have similar purchase patterns



RFM Analysis

- 90% revenue with 40% customers targeted with RFM
- Compared to this, predicting promotion acceptance via logistic regression, 90% revenue achieved from 70% of customers



List Scoring

- Tendency to participate in campaign highly dependent on previous engagement
- Targeting 17% of the base can capture 65% revenue with 52% profitability



Clustering

- Useful for understanding customers further
- Segmented into 5 clusters, out of which 4 statistically significant



Recommendations:

- Inclusive promotions, with offers on products for kids/teens for targeting families with kids
- Tailored promotions to achieve better engagement and footfall by understanding the customer segments and the channels used
- Improving reach of promotions and increasing awareness within customers
- Focus on customer retention and loyalty

Future Scope

The next steps could be as follows:

- Market Basket Analysis for identified customer segments and further stratification based on purchase patterns & channels and further refinement in inventory
- Analysis of individual segments to understand and predict response rate to different campaigns for marketing strategy customization
- Understanding customer churn and creation of strategies for retention and loyalty
- Ideation of real-time promotions to promote impulse buying, cross-sell and up-sell



THANK YOU



LOCATION
Please enter your address.



KEEP
CALM
AND
FIGHT
ON



Questions?

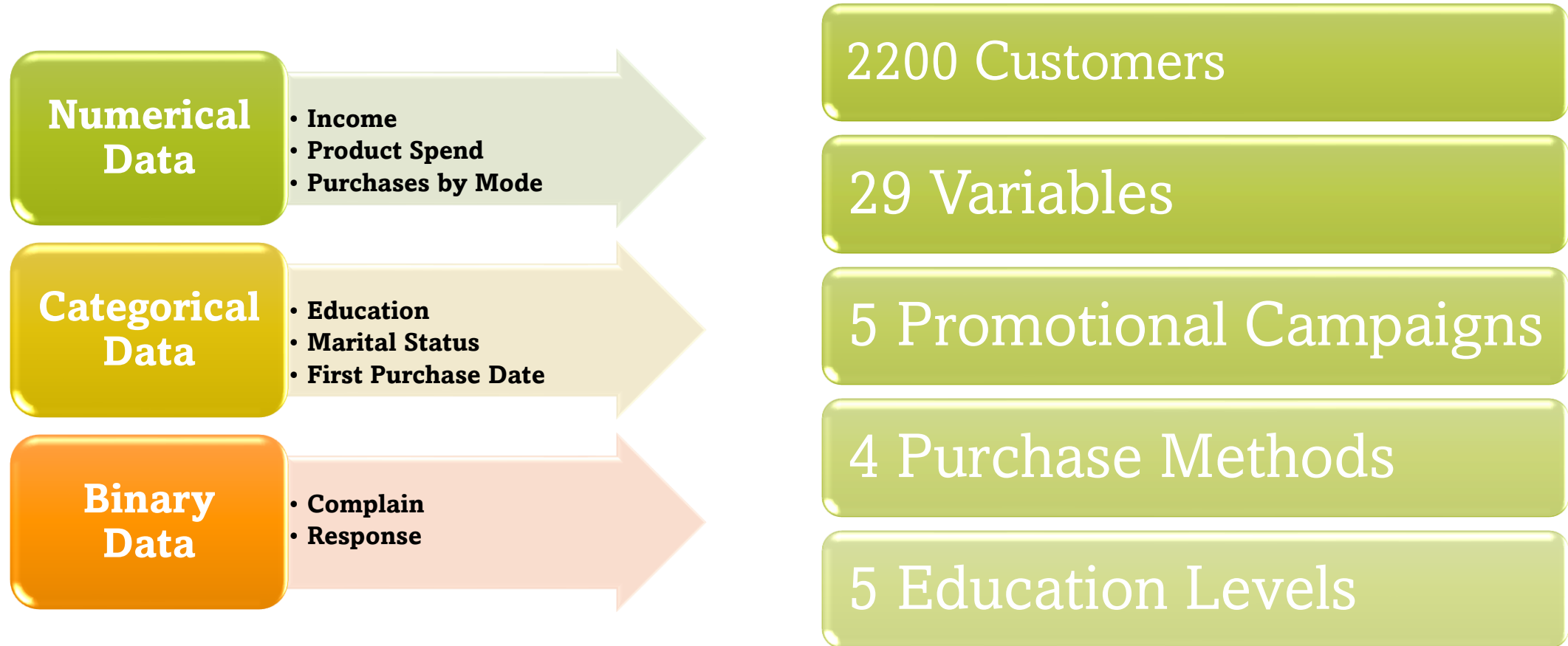
References

- <https://www.nytimes.com/2020/09/08/dining/grocery-shopping-coronavirus.html>
- <https://www.skyfilabs.com/project-ideas/customer-segmentation>
- <https://balancingeverything.com/grocery-shopping-statistics/>

Appendix



Introduction



Dataset Description

- **Customer:**

- 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 customer complained in the last 2 years, 0 otherwise

- **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

Dataset Description

- **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

- **Mode/ Location:**

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