UCL
SCHOOL OF
MANAGEMENT

MSIN0041: MARKETING SCIENCE

MARKETING SOFT DRINKS

December 2021

WORD COUNT: 2,000

Prepared by

François Barone - 19113943
Bhavya Gupta - 19073155
Oliver H. Haulund - 19021869
Marek Istok - 19025338
Markus Keiblinger - 19020414
Gabrielius Valiunas - 19020857

Table of Contents

1 Introduction

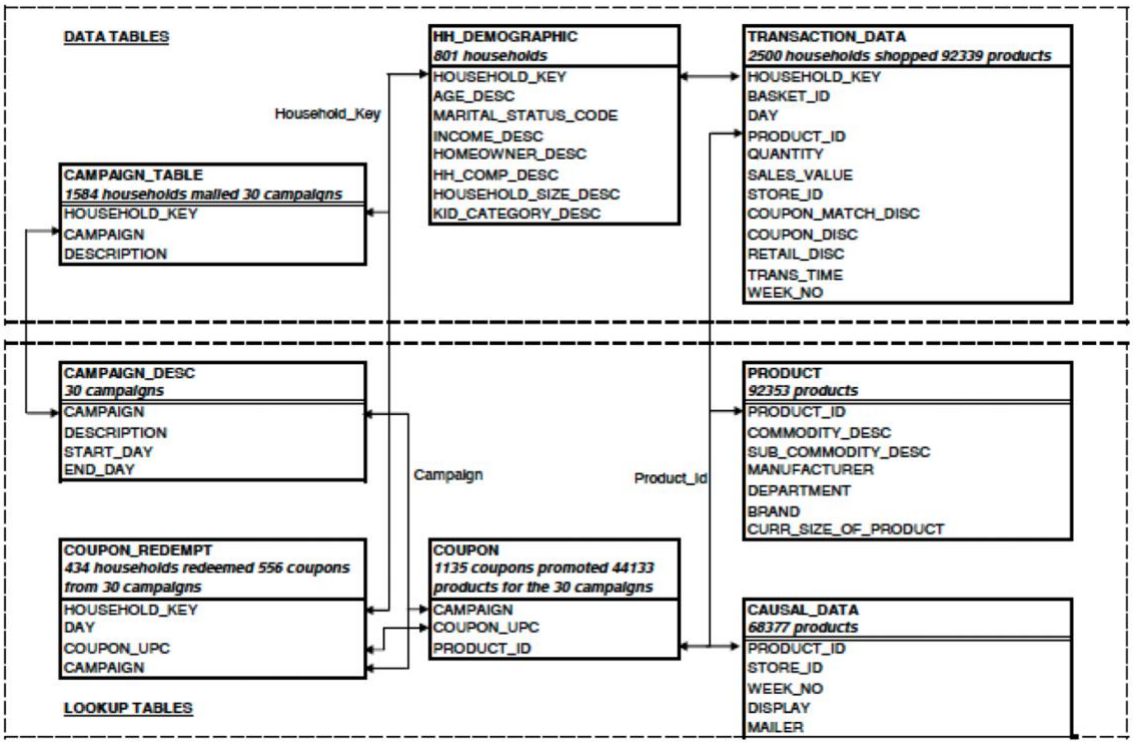
In recent decades, data analytics has become a main driver in decision making of supermarkets around the world (McKinsey Global Institute, 2011). Alongside other factors, it has played a significant role in increasing sales of U.S. grocery retailers almost twofold between 2000 and 2020 (Blázquez, 2021).

For instance, Walmart collects 2.5 petabytes of unstructured data from 1 million users every hour (Weitzel, 2019). Retailers use data analytics to determine pricing, product placement, promotions, and personalised advertisements. Actionable data-driven insights increase the effectiveness of strategies employed, most often increasing profits. The report seeks to explore the “The Complete Journey” dataset from Dunnhumby to discover insights that would allow the given supermarket to understand its pricing better and make data-driven optimisation decisions (Dunnhumby, 2014).

2 Overview of Data

The real-world data from a U.S. retailer includes transactions of 2,500 households over two years and provides demographic data for 801 (for a full data dictionary see Appendix 5.1). Calculating total revenue reveals that the retailers is among the 30 largest in the country.

Figure 1 – Data tables and their connections



In the data provided, most households consist of 45-54 years olds (Figure 2) with the largest income brackets being $35-49k and $50-79k (Figure 3).

Figure 2 - Age distribution for households

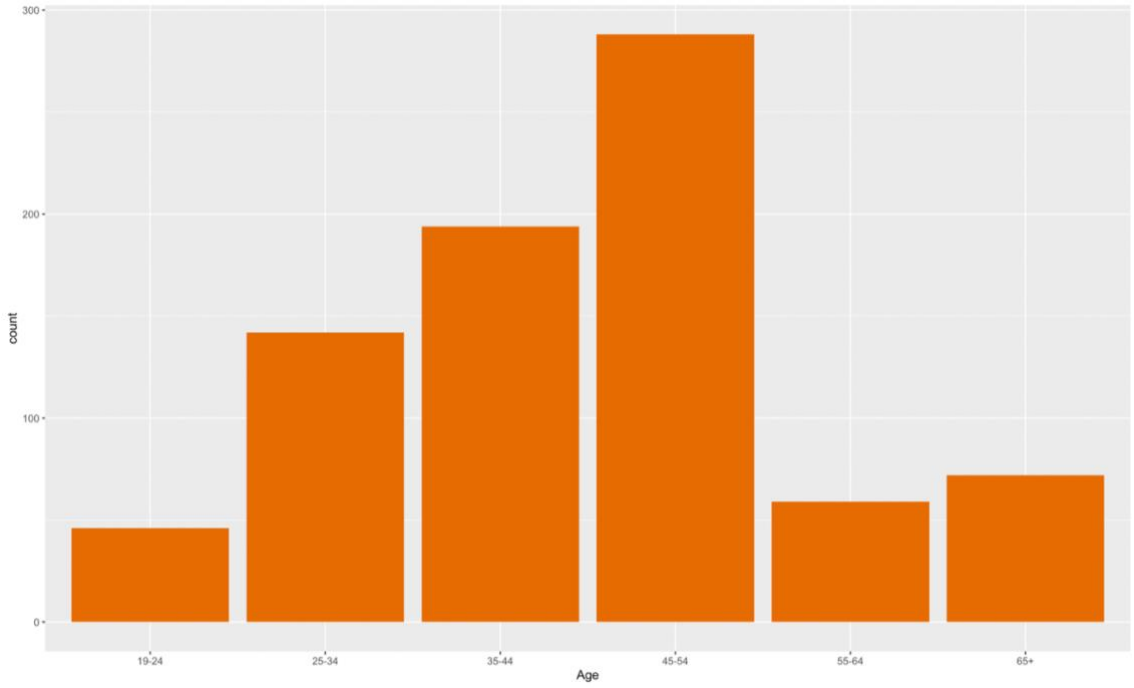
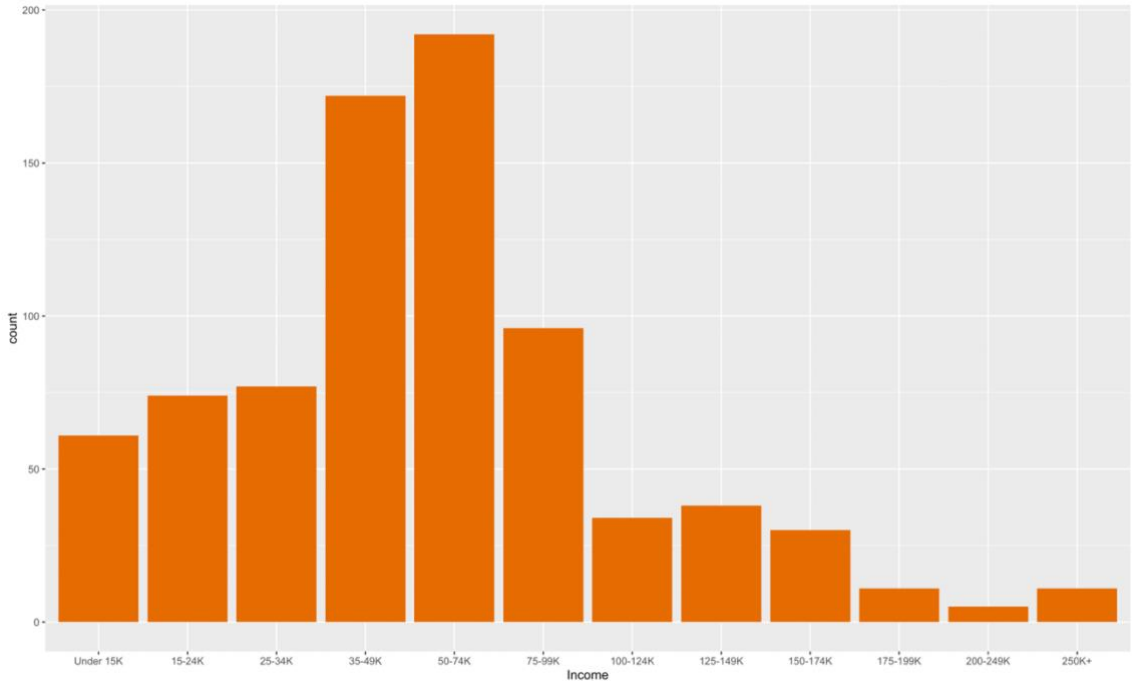


Figure 3 - Income distribution of households



2.1 Category selection

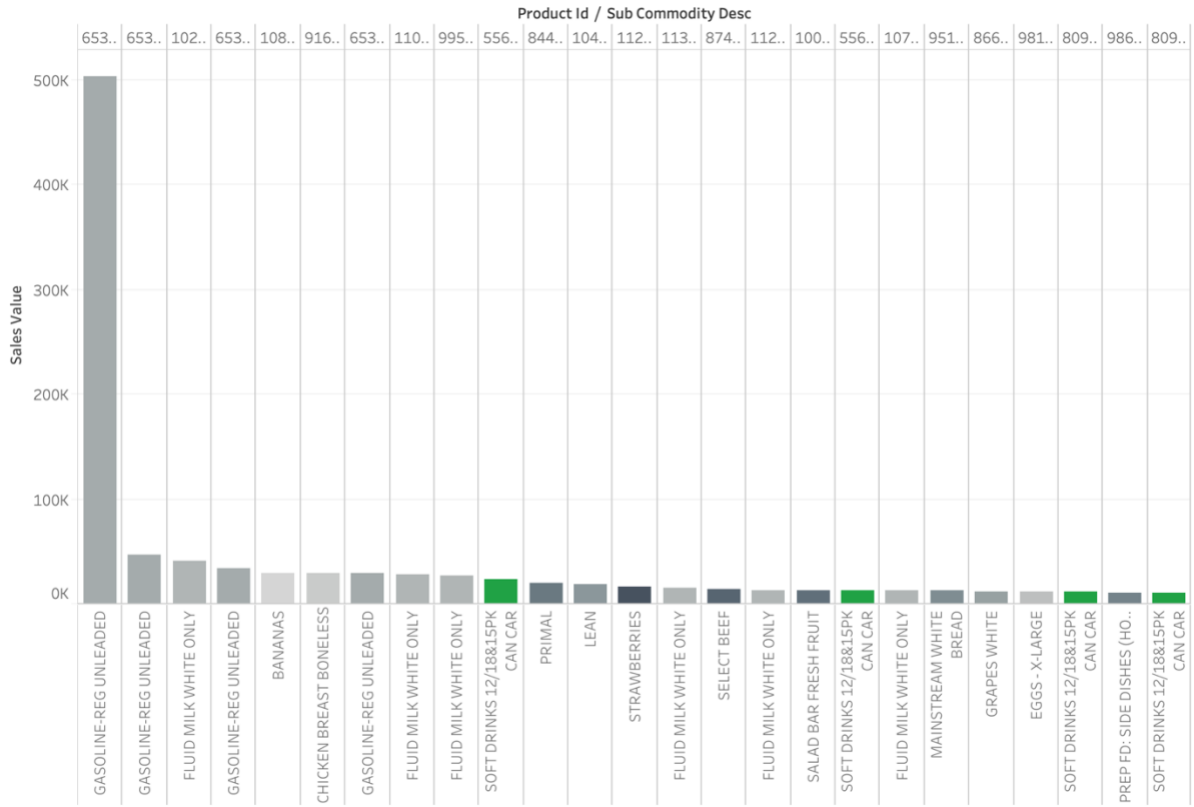
Figure 4 presents three reasons for focusing on 12oz soft drinks.

Figure 4 - Reasons for focus on soft drinks

They are one of the top three categories generating the most revenue amongst (appendix 5.2) and have the lowest year-over-year growth of 9.78%, which is also lower than the 28.25% overall average (appendix 5.3).
4 out of 7 chosen products fall within the top 25 of 92353 products generating the most sales (Figure 4).
Two different brands were represented within the 7 selected products, which can help us understand differences across brands.

The 12oz soft drink category makes up a significant part of the overall revenue generated by the retailer and there is room for further growth. Research questions in this report focus on discovering actionable insights which would lead to increased profits.

Figure 5 - Top 25 revenue-generating products



2.2 Research Questions

2.2.1 Question 1: "Can customer prices be optimised to increase profits?"

According to a study conducted in March 2018, about 75% of internet users consider discounts and coupons important for digital purchasing decision (Chevalier, 2021), drawing attention to the potential of these two marketing tools. Using sales data over time, the dataset can be used to calculate the price sensitivity of customers and determine the optimal cost markup. It can then be compared to the current pricing to identify potential improvements.

2.2.2 Question 2: "How can potential differences in Customer Life-Time-Value (CLV) across customer clusters be used to inform future marketing campaigns?"

To understand how much the retailer could spend on acquiring new customers, it is important to know the expected cash flows from each customer. The dataset allows for a granular analysis down to the individual customer. However, this report will investigate the differences on a community level, as it allows for easier targeting of future campaigns. As there is no data on costs, the analysis will not present the CLV in the traditional profit terms; rather, it will be based on revenue.

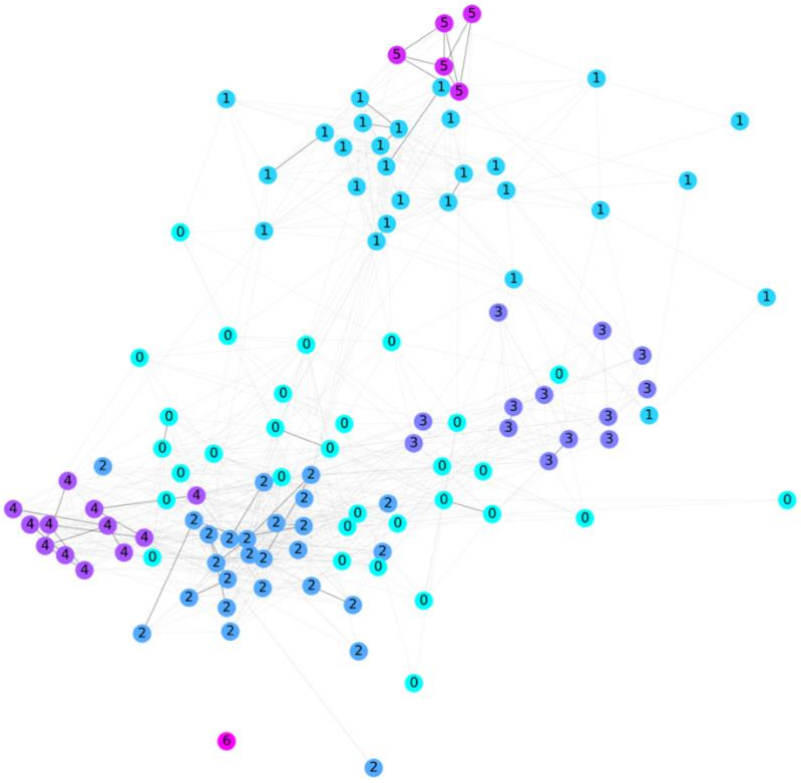
3 Methods, Findings and Implications

3.1 Question 1

To investigate the retailer's use of discounts through loyalty cards and coupons, and whether there is room to optimise profits, we first find the price elasticity of each selected product in distinct store communities.

To find the store communities, stores were linked together based on households that shop in both stores. Two store nodes are connected when there is a household which shopped in both stores. The weight of the edge between these two nodes is proportionate to the number of shared households. Once the store network is constructed, we use the Clauset-Newman-Moore greedy modularity maximizing algorithm to split the stores into different communities (NetworkX Developers, 2018).

Figure 6 - Stores grouped by customer shopping overlap for multiple stores



Communities of stores were used to segment customers, as this makes the segments more homogenous: customers shopping in the same stores likely share some, if not many, characteristics. The segmentation method also provides a good platform for targeting with coupons and loyalty cards.

Figure 7 - Demographical differences between store communities

	Avg. age	Avg. income	Revenue/household	Avg. household size	Avg. # of children
Com 0	37.73	88477	510.71	2.35	0.64
Com 1	37.77	57887	912.92	2.28	0.56
Com 2	39.40	68693	699.57	2.36	0.71
Com 3	39.62	79161	296.53	2.56	0.87
Com 4	40.64	79895	1101.43	2.35	0.74
Com 5	38.31	47838	547.40	1.64	0.15
Com 6	44.21	50684	468.18	2.21	0.47
Total	38.78	72566	760.12	2.31	0.64

After adjusting the data to fit our needs by merging it and adjusting prices marginally away from 0, regress log of weekly quantities on log of weekly prices for each community.

Figure 8 - Price elasticities of each product in each store community

	Com 0	Com 1	Com 2	Com 3	Com 4	Com 5	Com 6
Prod 1	-2.663	-2.963	-3.586	-3.343	-3.074	-1.249	-1.350
Prod 2	-3.522	-3.851	-2.430	-2.291	-4.570	-1.490	0.057
Prod 3	-2.701	-3.226	-3.280	NA	-3.295	-0.019	-0.928
Prod 4	-2.643	-3.878	-3.106	NA	-3.523	-2.477	-1.123
Prod 5	-2.441	-2.803	-3.625	-2.124	-2.955	-0.558	-0.470
Prod 6	-3.099	-3.477	-2.766	-1.563	-2.671	-0.465	0.000
Prod 7	-1.910	-3.136	-2.470	NA	-2.246	-1.089	-0.275

Figure 9 - The significance of the price elasticity for each product in each community

	Com 0	Com 1	Com 2	Com 3	Com 4	Com 5	Com 6
Prod 1	0	0	0	0	0	0.00	0.00
Prod 2	0	0	0	0	0	0.03	0.40
Prod 3	0	0	0	NA	0	0.88	0.01
Prod 4	0	0	0	NA	0	0.00	0.01
Prod 5	0	0	0	0	0	0.09	0.87
Prod 6	0	0	0	0	0	0.21	NaN
Prod 7	0	0	0	NA	0	0.04	0.59

It is important to note that some of the price elasticities in Figure 8 for communities 5 and 6 are not significant, as seen in Figure 9. This is largely a result of few weeks with transactions data for these communities, as seen in Figure 10 below.

Figure 10 - Weeks with transactions in dataset for each product and store community

	Com 0	Com 1	Com 2	Com 3	Com 4	Com 5	Com 6
Prod 1	101	100	98	90	98	58	23
Prod 2	95	92	98	82	91	12	7
Prod 3	95	97	94	0	95	60	26
Prod 4	88	97	93	0	92	81	28
Prod 5	76	72	85	44	71	27	7
Prod 6	85	71	58	30	73	14	5
Prod 7	75	70	78	0	62	34	12

From these results, it is evident that customers' sensitivity to price for the same products differs significantly from community to community, as seen in Figure 8. As the number of stores in each community is a good proxy of density of stores, and density of stores is a good proxy for competition that may drive more elasticity. Part of the difference in elasticities between communities can, therefore, be explained by the different number of stores in each community

as seen in Figure 11. The differences are also likely driven by demographic differences, such as those seen in Figure 7.

Figure 11 - Number of stores in each community

	Com 0	Com 1	Com 2	Com 3	Com 4	Com 5	Com 6
Number of stores	31	27	26	13	12	5	1

The differences in price elasticities suggest that the retailer should deploy community-individualised customer prices, unless the marginal cost of each product differs between communities. More accurately, the cost markup formula ($markup = -\frac{1}{1+\beta_{ii}}$) suggests that the retailer's customer prices are only optimal if the markup on the marginal cost of each product in each community is that found in Figure 12 below. One limitation of this method is that it only works for elastic products, hence all inelastic elasticities have been allocated "NA".

Figure 12 - Optimal markup on marginal cost for each product in each store community

	Com 0	Com 1	Com 2	Com 3	Com 4	Com 5	Com 6
Prod 1	0.601	0.509	0.387	0.427	0.482	4.016	2.857
Prod 2	0.397	0.351	0.699	0.775	0.280	2.041	NA
Prod 3	0.588	0.449	0.439	NA	0.436	NA	NA
Prod 4	0.609	0.347	0.475	NA	0.396	0.677	8.130
Prod 5	0.694	0.555	0.381	0.890	0.512	NA	NA
Prod 6	0.476	0.404	0.566	1.776	0.598	NA	NA
Prod 7	1.099	0.468	0.680	NA	0.803	11.236	NA

Considering that we do not know the retailer's marginal cost for these products, our recommendation is for the retailer to use the markups provided in Figure 12 to set their prices for each product in each community. As an example, if the retailer's marginal cost for product 1 in community 0 is \$2, the profit-maximising optimal price that the retailer should use will be $\$2 + \$2 * 0.601 = \$3.202$. We would however recommend treating all markups larger than 1 with some concern, as the closer the elasticities get to -1, the more sensitive the markups become.

To check how the retailer's current pricing matches these optimal markups, the marginal cost of each product is first assumed to be the lower whisker of the distribution of its prices, as we believe they would not sell the products at losses. Moreover, we assume the marginal costs are identical across all communities. These values are found in Figure 13 and Figure 14 below.

Figure 13 - Distribution of customer prices of each product

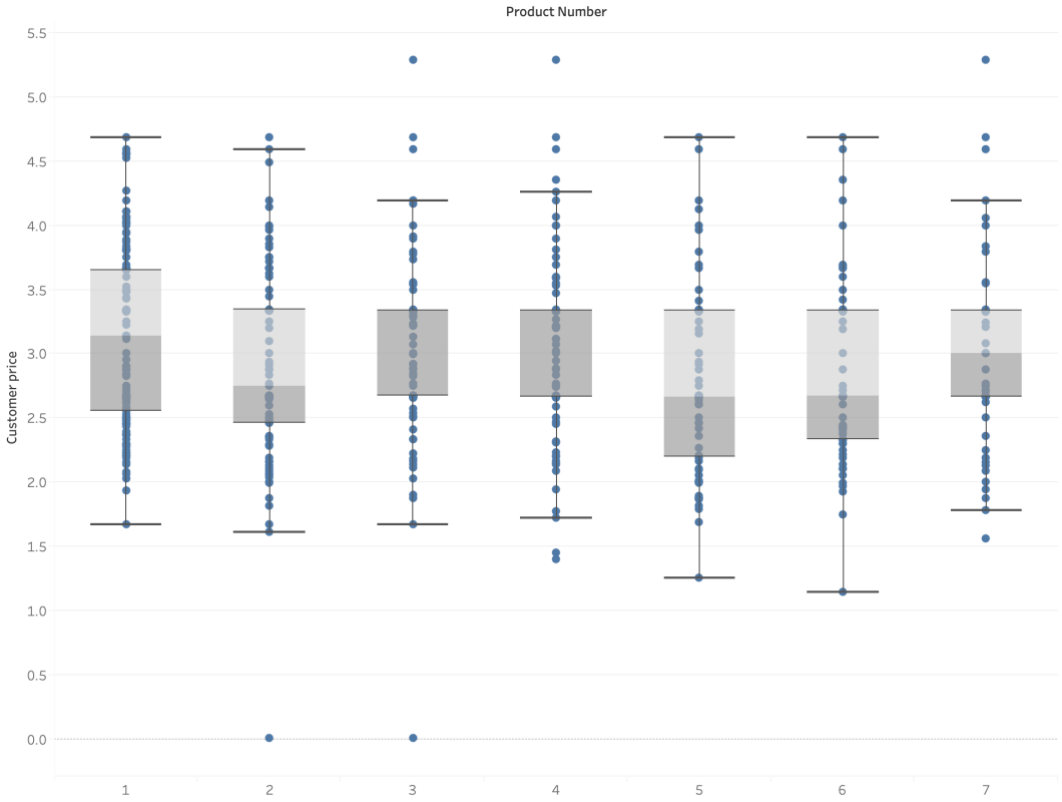


Figure 14 - Exact assumed marginal cost of each product

	Prod 1	Prod 2	Prod 3	Prod 4	Prod 5	Prod 6	Prod 7
Marginal cost	1.67	1.61	1.67	1.73	1.25	1.15	1.78

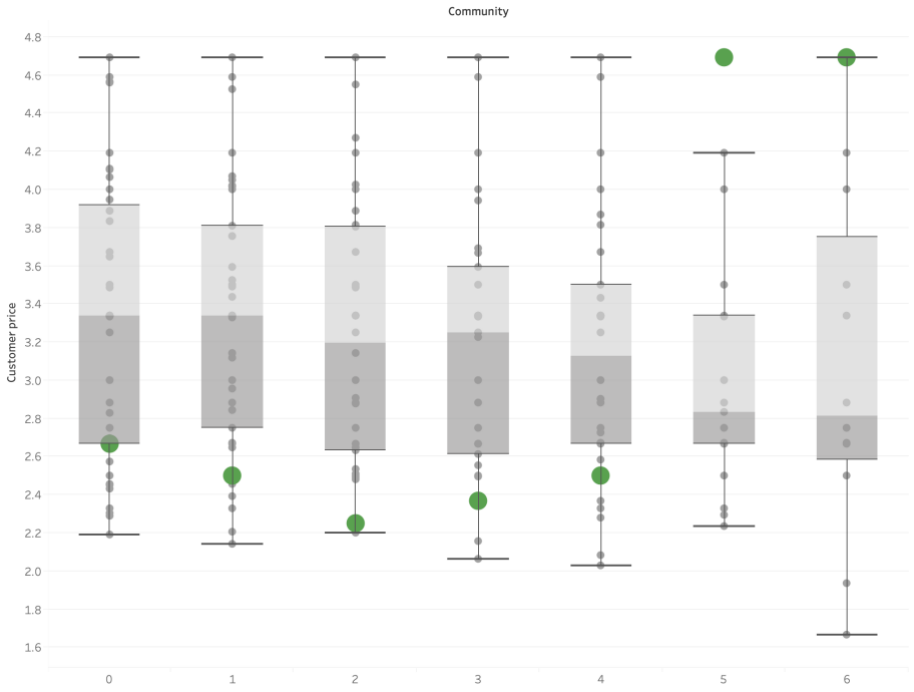
Using these and the optimal markups from Figure 12, without accounting for the voiced limitations, optimal prices are found per product in each community to be those in Figure 15.

Figure 15 - Optimal prices in each community

	Com 0	Com 1	Com 2	Com 3	Com 4	Com 5	Com 6
Prod 1	2.67	2.52	2.32	2.38	2.47	8.38	6.44
Prod 2	2.25	2.18	2.74	2.86	2.06	4.90	NA
Prod 3	2.65	2.42	2.40	NA	2.40	NA	NA
Prod 4	2.78	2.33	2.55	NA	2.42	2.90	15.79
Prod 5	2.12	1.94	1.73	2.36	1.89	NA	NA
Prod 6	1.70	1.61	1.80	3.19	1.84	NA	NA
Prod 7	3.74	2.61	2.99	NA	3.21	21.78	NA

Comparing these for product 1 to the mean prices used in transactions each week, as per Figure 16, we can see that the retailer seems to price product 1 higher than what would optimise their profits across community 0 through 4, which is seen by the optimal price points being below the lower quartile of actual prices in each community. As for communities 5 and 6, these are priced much lower than the model recommends, but fall within the reservation taken on markups higher than 100% previously. Therefore, we can conclude that the retailer should decrease the prices for product 1 to maximise the profit from this product alone.

Figure 16 - Actual pricing of product 1 (green dots represent calculated optimal prices)



Finally, comparing the price distributions of each product in Figure 13 with the optimal prices for each product in each community in Figure 15, we have no reason to believe that this overpricing is not happening across the vast majority of products and communities. Therefore, optimising pricing according to our recommendation seems to present a significant opportunity to increase profits from this segment of products for the retailer.

3.2 Question 2

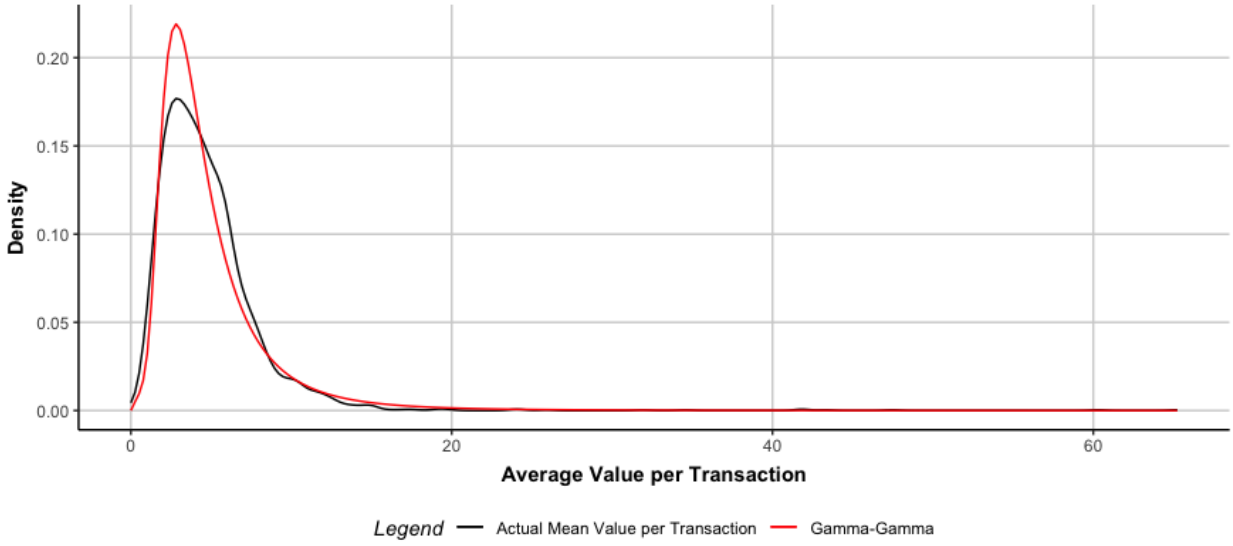
To find the CLV of each customer, we use the CLVTools package provided in R. We fit the BG/NBD model on the data provided, after which we use the estimated model parameters to predict future customer spending, predicting the characteristics in Figure 17.

Figure 17 - Household characteristics predicted by model

Predicted mean spending (calculated using the Gamma- Gamma model)
Conditional Expected Transactions (CET) - number of transactions to expect from a household during the prediction period
Probability of a customer being alive at the end of the period

An estimated spending model object is plotted, which provides a comparison of the estimated and actual density of customer spending.

Figure 18 - Density of average transaction value



Using the values estimated by the model above, we can calculate the CLV of each household, given by $E[CLV] = \sum_{t=0}^{T} m_c * \frac{P(alive|t)}{(1+d)^t}$. Where T is the total amount of periods (we settled on 50 years), m_c is the marginal spend, d is the discount rate (set to 10%), and $P(alive|t)$ is the probability of the customer being active at period t. After calculating the values for each household, we get the CLV as seen in Figure 19.

Figure 19 - CLV model output

	household_key	predicted.spend	CET	PAlive	COMMUNITY	CLV
0	1	6.152	30.669	0.997	5	1,774.100
1	10	2.912	4.834	0.997	2	132.270
2	100	5.483	6.260	0.995	3	315.710
3	1000	3.542	32.270	0.996	2	1,066.730
4	1001	9.712	43.897	0.997	4	4,004.440

Once we have identified the CLVs of each household, we aggregate households into previously found communities, and find the expected CLV in each community. Values for communities 0 to 4 can be taken as significant according to the model, but values found for communities 5 and 6 should not be taken as significant due to the small number of households in these two communities.

Figure 20 - CLVs per community

COMMUNITY	CLV
0	1,336.80
1	1,156.44
2	1,231.79
3	1,250.61
4	1,428.10
5	1,270.87
6	898.86
Total	1,224.78

An upside of considering stores in communities, as opposed to individually, is that we can split the CLV between multiple stores the customer frequents. This should lower marketing costs, as stores do not need to compete between each other for the same customer.

These differences can be used to drive the decision making of the marketing campaigns, as they show the value of acquiring new customers in each community. The information can be used to find or calculate various crucial sets of information - measure ROI, future cash flows of proposed campaigns, adjust marketing budgets in each community according to costs, or further analyse key driving factors in each different community.

4 Directions for Future Research

A limitation of the recommendations in 3.1 above is that these optimal markups do not account for the impact of cross-price elasticities. As the retailer likely wants to maximise profit across all beverages, not only the 7 highest grossing, these cross-price elasticities are essential. The retailer is recommended to complete an actual experiment across the communities to utilize the cross-price elasticities, and to also get more accurate results for communities 5 and 6, which we discuss are less accurate in Figure 12. Moreover, this pricing method does not consider the effect of using discounts and coupons to vary prices from time to time, which the retailer should also analyse to optimise and apply the community-individualised pricing strategy. Naturally, the retailer should do all of this with their actual marginal costs for each product and not those we assumed in 3.1.

Communities 5 and 6 do not have sufficient households and transaction data to provide significantly accurate results for CLV, like for price elasticities. As we are only working with a sample subset of the retailer's data, we would recommend conducting the same analysis for these communities, with a larger subset of the data.

Finally, one important consideration for the retailer to have and potentially analyse is beverages' role in the sales of the wider assortment. It is a known retailer strategy to minimise profit, or even lose money on some products to get customers in the store, where retailers then profit from their other purchases. The retailer should consider if beverages are, or should, play such a role, and how this is, or would, impact our recommendations.

5 Appendix

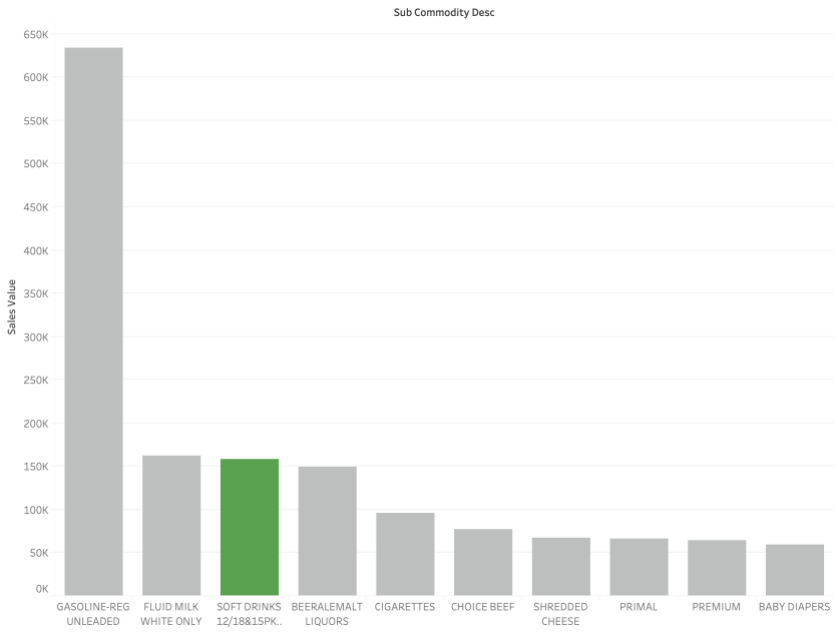
5.1 Data Dictionary

Dataset	Attribute	Data Type	Description
campaign_table and coupon_redempt and transaction_data	household_key	int	Indicates each unique household
transaction_data	BASKET_ID	num	Indicates each unique purchase occasion
coupon_redempt and transaction_data	DAY	int	Day of transaction
causal_data and coupon and product and transaction_data	PRODUCT_ID	int	Indicates each unique purchase occasion
transaction_data	QUANTITY	int	Number of products purchased in a trip
transaction_data	SALES_VALUE	num	Dollars received by the retailer for the sale
causal_data	STORE_ID	int	Indicates each unique store
transaction_data	RETAIL_DISC	num	Discount through retailer's loyalty card program
transaction_data	TRANS_TIME	int	Time of Transaction
causal_data and transaction_data	WEEK_NO	int	Week of transaction
transaction_data	COUPON_DISC	num	Discount through manufacturer's coupon
transaction_data	COUPON_MATCH_DISC	num	Discount through retailer's match of manufacturer coupon
product	MANUFACTURER	int	Manufacturer for each product
product	DEPARTMENT	factor	Groupings for similar products
product	BRAND	factor	Label Brand for the product
product	COMMODITY_DESC	factor	Sub groups for groupings for product
product	SUB_COMMODITY_DESC	factor	Lowest level of groupings for product
product	CURR_SIZE_OF_PRODUCT	factor	Package size
causal_data	display	factor	Display location
causal_data	mailer	factor	Mailer location
coupon and coupon_redempt	COUPON_UPC	num	Indicates each coupon, which is unique to household and campaign
campaign_desc and campaign_table and coupon and	CAMPAIGN	int	Indicates each unique campaign

coupon_redempt			
campaign_desc	DESCRIPTION	factor	Type of campaign
campaign_desc	START_DAY	int	Start date of campaign
campaign_desc	END_DAY	int	End date of campaign
hh_demographic	AGE_DESC	factor	Approximate age range
hh_demographic	MARITAL_STATUS_CODE	factor	Marital Status
hh_demographic	INCOME_DESC	factor	Household income
hh_demographic	HOMEOWNER_DESC	factor	Homeowner, renter etc
hh_demographic	HH_COMP_DESC	factor	Household composition
hh_demographic	HOUSEHOLD_SIZE_DESC	factor	Size of household
hh_demographic	KID_CATEGORY_DESC	factor	Number of children
Derived	SHELF_PRICE		Shelf price of each product = [sales value - (all discounts applied)]/quantity
Derived	CUSTOMER_PRICE		Price customer pays for each product = [Sales Value + Coupon Discount (since it is negative in the data)]/quantity sold

5.2 Top revenue-generating sub commodity groups

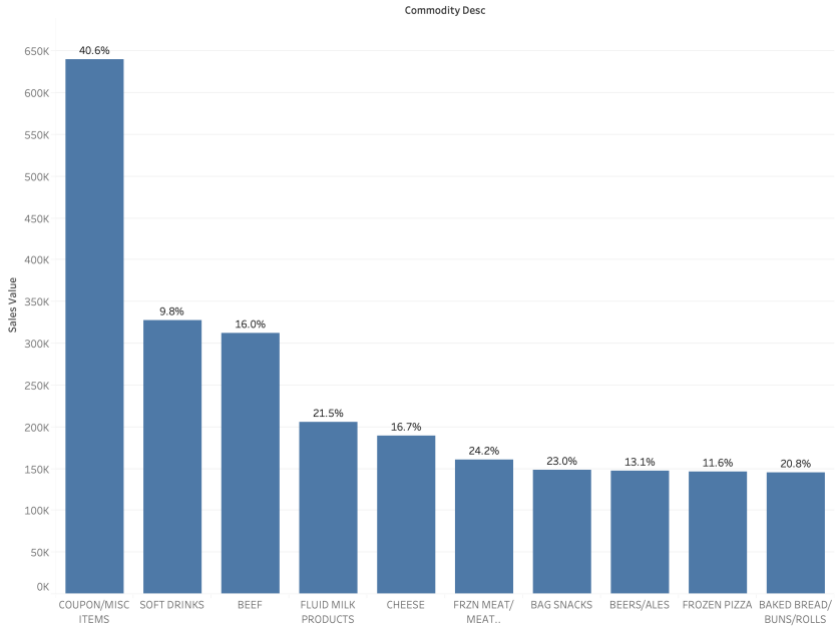
Figure 21 - Top 10 revenue-generating sub commodity groups



These are the ten most revenue-generating sub commodity groups amongst the 2,383 sub commodity groups held by the retailer.

5.3 Year-over-year growth rates of the top 10 revenue-generating commodity groups

Figure 22 - YoY growth rates for sales value of top 10 revenue-generating commodity groups



These sales values per commodity group are the ten highest, and the percentages above each is the YoY growth rate in sales value from the first to the second year of our data. It is evident that “soft drinks” is the only category with single digit growth rates.

6 Bibliography

Blázquez, A., 2021. *Grocery store sales in the United States from 1992 to 2020.* [Online]
Available at: https://www.statista.com/statistics/197621/annual-grocery-store-sales-in-the-us-since-1992/
[Accessed 9 December 2021].

Chevalier, S., 2021. *Importance of discounts and coupons to the overall digital purchasing decisions according to internet users in the United States as of March 2018, by age group.* [Online]
Available at: https://www-statista-com.libproxy.ucl.ac.uk/statistics/824001/users-importance-discounts-coupons-digital-purchasing-decisions/
[Accessed 7 December 2021].

Dunnhumby, 2014. *Dunnhumby - Source files.* [Online]
Available at: https://www.dunnhumby.com/source-files/
[Accessed 17 November 2021].

McKinsey Global Institute, 2011. *Big data: The next frontier for innovation, competition, and productivity.* [Online]
Available at:
https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20digital/our%20insights/big%20data%20the%20next%20frontier%20for%20innovation/mgi_big_data_full_report.pdf
[Accessed 7 December 2021].

NetworkX Developers, 2018. *NetworkX.* [Online]
Available at: https://networkx.org/documentation/networkx-2.2/reference/algorithms/generated/networkx.algorithms.community.modularity_max.greedy_modularity_communities.html
[Accessed 10 December 2021].

Weitzel, A., 2019. *What Was The Secret Behind Walmart's Rapid Ascension to Retail Glory?.* [Online]
Available at: https://coursekey.com/blog/walmart-real-time-data/
[Accessed 5 December 2021].