Grocery Store Customer Personality Analysis

Technological advances allow retailers to collect real-time data on their customers to determine their target demographics. Information on variables such as age, income, and items purchased are collected whenever a transaction is processed. With the collected data, corporations can find patterns in consumer profiles. These patterns identify which types of clients are more likely to buy specific products. This information is crucial for increasing sales through the use of targeted advertising campaigns.

Executive Summary

Predicting Purchases

What influenced a purchase:

- Income
- Number of teens at home
- Number of kids at home
- Graduating high school
- Having a Master's Degree
- Having a PhD

Wine

Significant Variables

- Income
- Number of kids at home

Fruits

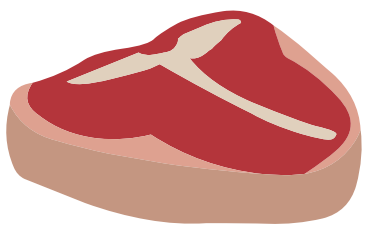
Significant Variables

- Income
- Number of teens at home
- Number of kids at home
- Graduating high school
- Having a Master's Degree
- Having a PhD

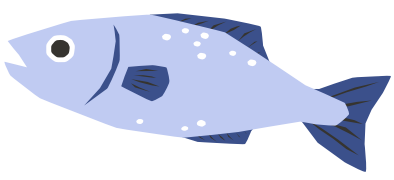
Meat

Significant Variables

- Income
- Number of teens at home
- Number of kids at home



Fish



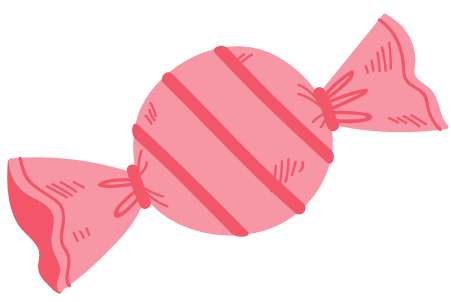
Significant Variables

- Income
- Number of teens at home
- Number of kids at home
- Having a PhD

Sweets

Significant Variables

- Income
- Number of teens at home
- Number of kids at home
- Age
- Graduating from high school
- Having a Masters Degree
- Having a PhD



Targeting Factors

Detailed Report on Customer Personality Analysis

Ciccone Morales, Isabelle (260950847)

Farrell, Cara-Li (261051787)

Gadoua, Jeremy (260983904)

Siou, Mark (261036146)

Yu, Elaine (261030551)

McGill University

INSY 336 - Data Handling & Coding for Analytics

Professor Dongliang Sheng

6 December, 2022

Detailed Report on Customer Personality Analysis

Problem Outline

Technological advances reshaping the digital marketing landscape allow retailers to collect real-time data on their consumers to determine their target demographics. Data such as age, income, and items purchased are collected whenever a transaction is processed. With the collected data, corporations can find patterns in consumer profiles to determine what demographics are more likely to buy specific products. This information is crucial for developing more effective targeted advertising campaigns to increase sales. This analysis, therefore, focuses on answering the following question: What are the key factors in determining whether someone is the ideal customer for purchasing a specific product?

The Dataset

Our dataset contains customer data based on 2-year spending across different product categories such as meat, sweets, wine, fruits and fish. Overall, there are 29 variables in the dataset and 2240 observations. Key variables include income, education, kids at home, teens at home, marital status, the amount spent in each category and recency in terms of purchases.

Approach Overview & Detailed Data Analysis Strategy

To better understand the types of ideal customers, we performed logistic regressions for specific products (i.e.: wine, fruits, meat, fish, and sweets) and performed a multilinear regression in order to observe the relationship between the variables and total spent. We then dummified each product based on whether the customer bought more than the median amount, with 1 being yes and 0 no. We chose these specific variables as the insights gained from our analysis could help grocers in particular time discounts, offer promotions to specific groups and adjust their product offerings.

Before doing so, some key assumptions had to be made. First, is that all data is on the individual level. Second, "education basic" refers to those who did not finish high school and "education graduation" refers to those who finished high school. When it comes to pre-processing the data, we replaced nonsensical values such as "absurd" and "YOLO" for marital status with "None" and added columns for age and total spent. Also, we had to annualize product spending as the product spending in the file was over the course of 2 years. We also made some visualizations describing our dataset (See Appendix A). In doing so, we found that most of the sample is aged between 40 and 50. Also, the product with the highest amount of spending is Wine. Interestingly, we also found that as family size increases food spending decreases. We also found that the majority of the sample has a high school level of education.

Findings

Total amount spent (Appendix B). We first started with a multiple linear regression to predict how much a customer would spend, based on various demographic factors. The model is shown to be significant as the P-value of the F-stat is 0.00, which is ≤ 0.05. The equation, with all the significant variables, is as follows:

$$TotalSpent = 0.006Income - 92.38Teenhome - 177.56Kidhome + 56.97EducationMaster + 65.35EducationGraduation + 83.40EducationPhD$$

For income, it is found that there is a statistically significant and positive relationship between this variable and the total amount spent ($\beta = 0.006,\ p \leq 0.05$). Specifically, the estimated effect of income is found as 0.006, implying that a unit increase in income increases the total amount spent by \$0.006.

For the number of teens at home, it is found that there is a statistically significant and negative relationship between this variable and the total amount spent

($\beta = -92.38,\ p < 0.05$). Specifically, the estimated effect of income is found as -92.38 and this implies a unit increase in the number of teens at home decreases the total amount spent by $92.38.

For the number of kids at home, it is found that there is a statistically significant and negative relationship between this variable and the total amount spent ($\beta = -177.56,\ p \leq 0.05$). Specifically, the estimated effect of the number of kids at home is found as -177.56 and this implies a unit increase in the number of kids at home decreases the total amount spent by $177.56.

For having a Master's degree, it is found that there is a statistically significant and positive relationship between this variable and the total amount spent ($\beta = 56.97,\ p \leq 0.05$). Specifically, the estimated effect of having a Master's degree is found as 56.97 and this implies having a Master's degree increases the total amount spent by $56.97.

For having graduated from high school, it is found that there is a statistically significant and positive relationship between this variable and the total amount spent ($\beta = 65.35,\ p \leq 0.05$). Specifically, the estimated effect of having graduated from high school is found as 65.35, which implies that graduating from high school increases the total amount spent by $65.35.

For having a Ph.D., it is found that there is a statistically significant and positive relationship between this variable and the total amount spent ($\beta = 83.40,\ p \leq 0.05$). Specifically, the estimated effect of having a Ph.D. is found as 83.40 and this implies that having a Master's degree in income increases the total amount spent by $ 83.40.

Wine (Appendix C). The logistic regression containing all significant variables for wine is as follows:

$$log(\frac{\pi}{1-\pi}) = -\ 5.36\ +\ 8.93\times\ 10^{-05}Income-\ 1.30Kidhome$$

Using the F-statistic, we can see that the model is significant at the 0.05 level of significance (P-value of F-statistic $4.41\times\ 10^{-210} \leq\ 0.05$). Diving into the variables, 2 of them were significant with P-values less than 0.05.

Income has a coefficient of $8.93\times\ 10^{-05}$ meaning that, by holding other variables fixed, one unit increase in income means that the log odds of being the ideal wine customer increase by $8.93\times\ 10^{-05}$, multiplies the odds by $e^{8.93\times10^{-05}}$ which is equal to 1.0001 and lastly, the odds increase by$(e^{1.00\times10^{-04}}\ -\ 1)\ \times\ 100\ =\ 0.01\%$ for every dollar increase in income.

The number of kids at home (Kidhome) has a coefficient of -1.30 meaning that, by holding other variables fixed, one unit increase in the number of kids at home means that the log odds of being the ideal wine customer decrease by 1.30, multiplies the odds by $e^{-1.30}$ which is equal to 0.27 and lastly, the odds decreases by $(e^{1.30}\ -\ 1)\ \times\ 100\ =\ 266.93\%$ for every increase in the number of kids at home.

Furthermore, we calculated the scores for wine to test the average accuracy of the training and test data which tells us the percentage of correct predictions. The training and test data scores are respectively 86.91% and 85.86%. These scores are relatively high and reinforce our claim the model is somewhat significant and reliable.

Lastly, we created a confusion matrix to evaluate our predictions. We found that 296 customers from our sample have been predicted to not be the ideal wine consumer and it turned out to be true, and 46 customers from our sample have been predicted to not be the ideal wine

consumer and it turned out to be false. Also, 48 customers from our sample have been predicted to be the ideal wine consumer and it turned out to be false, and 275 customers from our sample have been predicted to be the ideal wine consumer and it turned out to be true.

Fruits (Appendix D). The logistic regression containing all significant variables for fruit is as follows:

$$log(\frac{\pi}{1-\pi}) = -1.39 + 6.83\times 10^{-05} Income - 1.04 Teenhome - 1.07 Kidhome - 0.99 Education\_Master - 0.89 Education\_Graduation - 1.84 Education\_PhD$$

Using the F-statistic, we can see that the model is significant at the 0.05 level of significance (P-value of F-statistic $1.72 \times 10^{-151} \leq 0.05$). Diving into the variables, 6 of them were significant with P-values less than 0.05.

Income has a coefficient of 6.83×10^{-05} meaning that, by holding other variables fixed, one unit increase in income means that the log odds of being the ideal fruits customer increase by 6.83×10^{-05}, multiplies the odds by $e^{6.83\times 10^{-05}}$ which is equal to 1.00 and lastly, the odds increase by$(e^{6.83\times 10^{-05}} - 1) \times 100 = 0.01\%$ for every dollar increase in income.

The number of teens at home (Teenhome) has a coefficient of -1.04 meaning that log odds of being the ideal fruits customer decrease by -1.04 log odds, multiplies the odds by $e^{-1.04} = 0.3534$ and decreases the odds of buying fruit by $(e^{1.04} - 1) \times 100 = 182.92\%$

The number of kids at home (Kidhome) has a coefficient of -1.07 meaning that, by holding other variables fixed, one unit increase in the number of kids at home means that the log odds of being an ideal fruits customer decrease by 1.07, multiplies the odds by $e^{-1.07}$ which is

equal to 0.3430 and, the odds decrease by$(e^{1.07} - 1) \times 100 = 191.54\%$ for every unit increase in the number of kids at home.

To continue, having a master's degree (Education_Master) has a coefficient of -0.99, meaning that, by holding other variables fixed, having a master's degree means that the log odds of being an ideal fruits customer decrease by 0.99, multiplies the odds by $e^{-0.99}$ which is equal to 0.38 and lastly, the odds decrease by $(e^{0.99} - 1) \times 100 = 169.12\%$ when having a master's degree.

Having a high school degree (Education_Graduation) has a coefficient of -0.89 meaning that, by holding other variables fixed, having a high school diploma means that the log odds of being an ideal fruits customer decrease by 0.89, multiplies the odds by $e^{-0.89}$ which is equal to 0.41 and lastly, the odds decrease by$(e^{0.77} - 1) \times 100 = 143.51\%$ when having a high school degree.

Finally, having a PhD (Education_PhD) has a coefficient of -1.84, meaning that, by holding other variables fixed, having a Ph.D. means that the log odds of being an ideal fruits customer decrease by 1.84, multiplies the odds by, $e^{-1.84}$ which is equal to 0.16 and lastly, the odds decrease by$(e^{1.76} - 1) \times 100 = 542.37\%$ when having a PhD.

Furthermore, to test the average accuracy of the training and test data, which tells us the percentage of correct predictions, we calculated the scores for fruit. The training and test data scores are 79.75% and 78.19%. These scores are not as high as other models for different products, but they still reinforce our claim that the model is somewhat significant and reliable.

Lastly, we created a confusion matrix to evaluate our predictions. We found that 295 customers from our sample have been predicted to not be the ideal fruits consumer and it turned

out to be true, and 93 customers from our sample have been predicted to not be the ideal fruits consumer and it turned out to be false. Also, 52 customers from our sample have been predicted to be the ideal fruits consumer and it turned out to be false, and 225 customers from our sample have been predicted to be the ideal fruits consumer and it turned out to be true.

Logit Regression Results

Dep. Variable:	Meat	No. Observations:	1551
Model:	Logit	Df Residuals:	1542
Method:	MLE	Df Model:	8
Date:	Tue, 06 Dec 2022	Pseudo R-squ.:	0.4523
Time:	17:57:52	Log-Likelihood:	-588.79
converged:	True	LL-Null:	-1075.1
Covariance Type:	nonrobust	LLR p-value:	1.249e-204

	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.5677	1.089	-4.194	0.000	-6.702	-2.433
Income	9.608e-05	5.68e-06	16.926	0.000	8.5e-05	0.000
Teenhome	-0.4728	0.150	-3.158	0.002	-0.766	-0.179
Kidhome	-1.1232	0.156	-7.201	0.000	-1.429	-0.817
Age	-0.0105	0.007	-1.510	0.131	-0.024	0.003
Education_Master	0.9556	1.053	0.908	0.364	-1.108	3.019
Education_Graduation	1.0089	1.049	0.962	0.336	-1.047	3.064
Education_PhD	0.6088	1.058	0.576	0.565	-1.464	2.681
Marital_Status_Single	-0.0021	0.155	-0.013	0.989	-0.305	0.301

Meat (Appendix E). The logistic regression containing all significant variables for meat is as follows:

$$log(\tfrac{\pi}{1-\pi}) = -\ 4.57\ +\ 9.61\times\ 10^{-05} Income\ -\ 0.47 Teenhome$$

$$-\ 1.12 Kidhome$$

Using the F-statistic, we can see that the model is significant at the 0.05 level of significance (P-value of F-statistic $1.25e^{-204} \leq 0.05$). Diving into the variables, 3 of them were significant with P-values less than 0.05.

Income has a coefficient of $9.61\times\ 10^{-05}$ meaning that, by holding other variables fixed, one unit increase in income means that the log odds of being an ideal customer for meat increase

by 9.61×10^{-05}, multiplies the odds by $e^{9.61\times10^{-05}}$ which is equal to 1.0001 and lastly, the odds increase by$(e^{9.61\times10^{-05}} - 1) \times 100 = 0.0096\%$ for every dollar increase in income.

The number of teens at home (Teenhome) has a coefficient of -0.47 meaning that, by holding other variables fixed, one unit increase in income means that the log odds for purchasing meat decrease by -0.47, multiplies the odds by $e^{-0.47} = 0.625$ and decreases the odds of being an ideal customer for meat by $(e^{0.47} - 1) \times 100 = 60\%$.

The number of kids at home (Kidhome) has a coefficient of -1.12 meaning that, by holding other variables fixed, one unit increase in the number of kids at home means that the log odds of being an ideal customer for meat decrease by 1.12, multiplies the odds by $e^{-1.12}$ which is equal to 0.33 and lastly, the odds decrease by$(e^{1.12} - 1) \times 100 = 206.5\%$ for every increase in the number of kids at home.

Furthermore, to test the average accuracy of the training and test data, which tells us the percentage of correct predictions, we calculated the scores for meat. The training and test data scores are respectively 85.56% and 85.71%. These scores are relatively high and reinforce our claim the model is somewhat significant and reliable.

Lastly, we created a confusion matrix to evaluate our predictions. We found that 306 customers from our sample have been predicted to not be the ideal meat consumer and it turned out to be true, and 58 customers from our sample have been predicted to not be the ideal meat consumer and it turned out to be false. Also, 37 customers from our sample have been predicted to be the ideal meat consumer and it turned out to be false, and 264 customers from our sample have been predicted to be the ideal meat consumer and it turned out to be true.

Fish (Appendix F). The logistic regression containing all significant variables for fish is as follows:

$$log(\frac{\pi}{1-\pi}) = -\ 0.97\ +\ 5.68\times\ 10^{-05}Income\ -\ 1.16Teenhome$$

$$-\ 1.21Kidhome\ -\ 1.42Education\_PhD$$

Using the F-statistic, we can see that the model is significant at the 0.05 level of significance (P-value of F-statistic $6.14\times\ 10^{-137} \leq\ 0.05$). Diving into the variables, 4 of them were significant with P-values less than 0.05.

Income has a coefficient of $5.68\times\ 10^{-05}$ meaning that, by holding other variables fixed, one unit increase in income means that the log odds of being the ideal fish customer increase by $5.68\times\ 10^{-05}$, multiplies the odds by $e^{5.68\times10^{-05}}$ which is equal to 1.0000 and lastly, the odds increase by$(e^{5.68\times10^{-05}}\ -\ 1)\ \times\ 100\ =\ 0.0057\%$ for every dollar increase in income.

The number of teens at home (Teenhome) has a coefficient of -1.16 meaning that log odds of being the ideal fish customer decrease by 1.16 log odds, multiplies the odds by $e^{-1.16}$ which is equal to 0.3135 and decreases the odds of buying fish by $(e^{1.16}\ -\ 1)\ \times\ 100\ =\ 218.99\%$.

The number of kids at home (Kidhome) has a coefficient of -1.21 meaning that, by holding other variables fixed, one unit increase in the number of kids at home means that the log odds of being the ideal fish customer decrease by 1.21, multiplies the odds by $e^{-1.21}$ which is equal to 0.2982 and lastly, the odds increase by$(e^{1.21}\ -\ 1)\ \times\ 100\ =\ 235.35\%$ for every increase in the number of kids at home.

Having a Ph.D. (Education_PhD) has a coefficient of -1.42, meaning that the log odds of being the ideal fish customer decrease by 1.42, multiplies the odds by $e^{-1.42}$ which is equal to 0.2417 and lastly, the odds increase by$(e^{1.42} - 1) \times 100 = 313.71\%$.

Furthermore, to test the average accuracy of the training and test data which tells us the percentage of correct predictions, we calculated the scores for fish. The training and test data scores are 79.30% and 76.39%. These scores are not as high as other models for different products, but they still reinforce our claim that the model is somewhat significant and reliable.

Lastly, we created a confusion matrix to evaluate our predictions. We found that 289 customers from our sample have been predicted to not be the ideal fish consumer and it turned out to be true, and 98 customers from our sample have been predicted to not be the ideal fish consumer and it turned out to be false. Also, 59 customers from our sample have been predicted to be the ideal fish consumer and it turned out to be false, and 219 customers from our sample have been predicted to be the ideal fish consumer and it turned out to be true.

Sweets (See Appendix G). The logistic regression containing all significant variables for sweets is as follows:

$$log(\frac{\pi}{1-\pi}) = 5.73\text{x}10^{-05}Income - 0.86Teenhome - 1.28Kidhome - 0.02Age$$

$$-1.69Education\_Master - 1.18\ Education\_Graduation - 2.26Education\_PhD$$

Using the F-statistic, we can see that the model is significant at the 0.05 level of significance (P-value of F-statistic ≤ 0.05). Diving into the variables, 7 of them were significant with P-values less than 0.05.

Income has a coefficient of $5.73\text{x}10^{-05}$ meaning that, by holding other variables fixed, one unit increase in income means that the log odds of being the ideal sweets customer increase

by $5.73\text{x}10^{-05}$, multiplies the odds by $e^{5.73\text{x}10^{-05}}$ which is equal to 1.00006 and lastly, the odds increase by$(e^{6.53\text{x}10^{-05}} - 1) \times 100 = 0.01\%$ for every dollar increase in income.

The number of teens at home (Teenhome) has a coefficient of -0.86 meaning that log odds of being the ideal sweets customer decrease by -0.86 log odds, multiplies the odds by $e^{-0.86} = 0.42$ and decreases the odds of buying sweets by$(e^{0.82} - 1) \times 100 = 136.32\%$.

The number of kids at home (Kidhome) has a coefficient of -1.28, which means that log odds of being the ideal sweets customer decreases by 1.28, multiplies the odds by $e^{-1.28} = 0.28$ and decreases the odds of buying sweets by $(e^{1.28} - 1) \times 100 = 259.66\%$.

Age has been found to be significant for sweets. Age has a coefficient of -0.02, meaning that, by holding all other variables fixed, one unit increase in age decreases the log odds of being the ideal sweets customer by 0.02. The odds decrease by $e^{-0.02}$, it multiplies the odds by 0.9802 and decreases the odds of purchasing sweets by $(e^{0.02} - 1) \times 100 = 202.01\%$.

Having a master's degree (Education_Master) has a coefficient of -1.69 meaning that, by holding other variables fixed, having a master's degree means that the log odds of being the ideal sweets customer decrease by 1.69, multiplies the odds by $e^{-1.69}$ which is equal to 0.18 and lastly, the odds increase by$(e^{1.69} - 1) \times 100 = 441.95\%$ when having a master's degree.

Having a high school degree (Education_Graduation) has a coefficient of -1.18, meaning that, by holding other variables fixed, having a high school diploma means that the log odds of being the ideal sweets customer decrease by 1.18, multiplies the odds by $e^{-1.18}$ which is equal to 0.31 and lastly, the odds decrease by $(e^{1.18} - 1) \times 100 = 225.44\%$ when having a master's degree.

Having a doctorate (Education_PhD) has a coefficient of -2.26, meaning that, by holding other variables fixed, having a Ph.D. means that the log odds of being the ideal sweets customer decrease by 2.26, multiplies the odds by $e^{-2.26}$, which is equal to 0.10 and lastly, the odds decrease by$(e^{2.26} - 1) \times 100 = 858.31\%$ when having a Ph.D. degree.

Furthermore, to test the average accuracy of the training and test data which tells us the percentage of correct predictions, we calculated the scores for sweets. The training and test data scores are respectively 78.78% and 75.18%. These scores are not as high as other models for different products, but they still reinforce our claim that the model is somewhat significant and reliable.

Lastly, we created a confusion matrix to evaluate our predictions. We found that 284 customers from our sample have been predicted to not be the ideal sweets consumer and it turned out to be true, and 109 customers from our sample have been predicted to not be the ideal sweets consumer and it turned out to be false. Also, 56 customers from our sample have been predicted to be the ideal sweets consumer and it turned out to be false, and 216 customers from our sample have been predicted to be the ideal sweets consumer and it turned out to be true.

Summary & Conclusion

In conclusion, when companies try to identify their ideal consumer base, there are some common significant predictive variables across all the product types. The variables of income, number of kids at home and number of teens at home are all significant indicators to determine whether or not an individual is a potential ideal customer (See Appendix H). The only exception was wine, where having a teen at home was insignificant. After running our analysis, we can now identify the ideal customer more closely. For example, suppose you're a Sweets seller. In

that case, you should target a younger audience with slightly higher incomes and no kids to have the highest chance of successfully increasing sales through marketing campaigns.

Furthermore, we identified that a customer's marital status was insignificant in predicting whether someone is the ideal customer. Thus, companies should not focus on targeting consumers based on their marital status. Overall, our analysis demonstrates the importance of customer segmentation and targeting. Identifying the ideal customer for different product types allows the marketing team to be more effective and efficient in their advertising campaigns. Tailoring ads to consumers who are more willing to purchase will yield higher returns, saving the company both time and resources.

Limitations

A limitation in our models is in coding the dummy variables. For each product (i.e. wine, fruits, meat, fish, and sweets), we took the median in order to determine a threshold for whether it would be coded as 1 (they are an ideal customer for targeting for the specific product) or 0 (they are not an ideal customer for targeting for the specific product). Therefore, by splitting the data in half, it only took the highest-spending customers. Our dummified variables were skewed towards people with more income, which explains why income had a positive relationship in all our logistic regressions.

Additionally, we should keep in mind that our analysis does not necessarily predict consumer purchasing decisions per se, as we all need to grocery shop and we will not stop doing so, regardless of any factors. However, our analysis does successfully predict which factors will ensure that companies target their ideal customer profiles, keeping in mind that this doesn't mean other customer profiles will not purchase these products. Rather that they would be less inclined to be persuaded by these marketing campaigns.

Technical Appendix

Dataset: https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis

Descriptions of Variables Used

Customer Background

1. *ID: Customer's unique identifier*
2. *Year_Birth: Customer's birth year*
3. *Education: Customer's education level*
    a. *Education_Master: Customer has a Master's Degree*
    b. *Education_Graduation: Customer has a High School Diploma*
    c. *Education_PhD: Customer has a Ph.D.*
    d. *Education_Basic: Customer did not finish high school*
4. *Marital_Status: Customer's marital status*
    a. *Marital_Status_Single: Customer is single, alone, divorced, or widowed*
    b. *Marital_Status_Together: Customer is married, or together*
5. *Income: Customer's yearly household income*
6. *Kidhome: Number of children in customer's household*
7. *Teenhome: Number of teenagers in customer's household*

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*

Link to our code:

https://colab.research.google.com/drive/1CvVeTcltyeN_qG6IMstikysRmcVshjwK?usp=sharing

Appendix A

Figure 1. Age Distribution of Sample

Note: Most of the customers in this sample appear to be in their late 40s or early 50s.

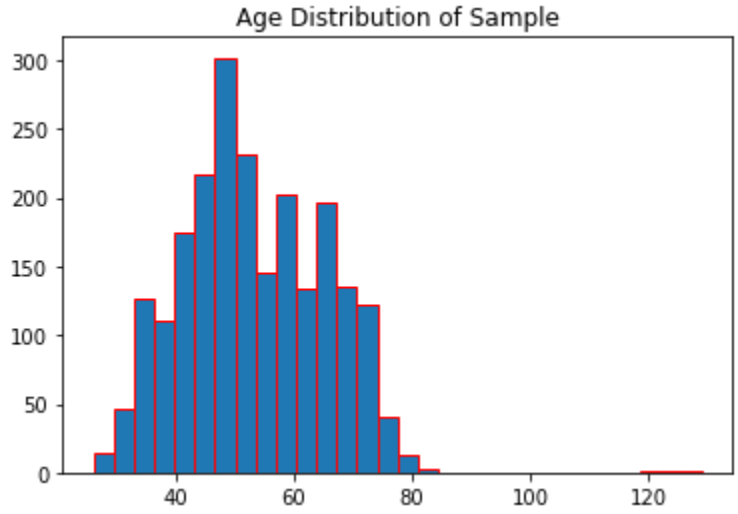


Figure 2. Average Yearly Spending in Each Product Category

Note: Most of the customers in this sample appear to spend more on wine and meat.

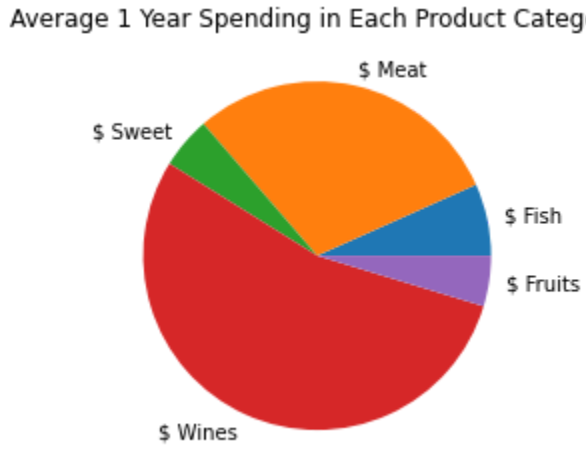


Figure 3. Spending by Family Size

Note: Most of the customers in this sample appear to spend less on groceries the more people there are in their household.

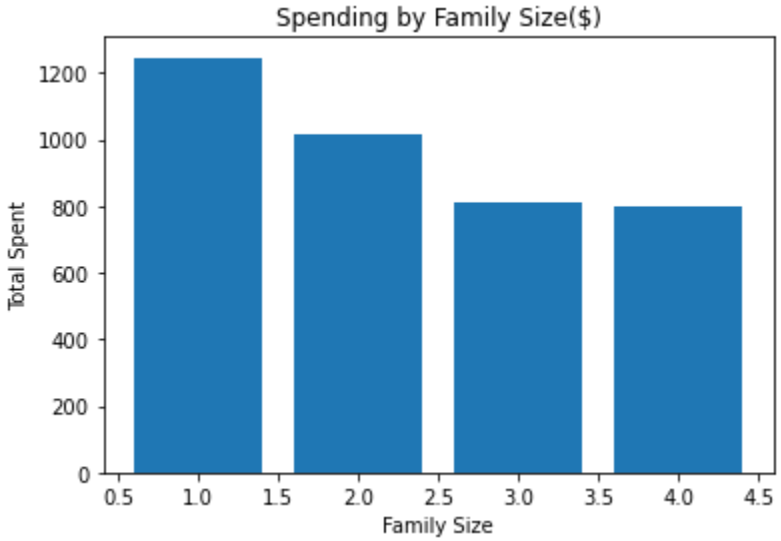
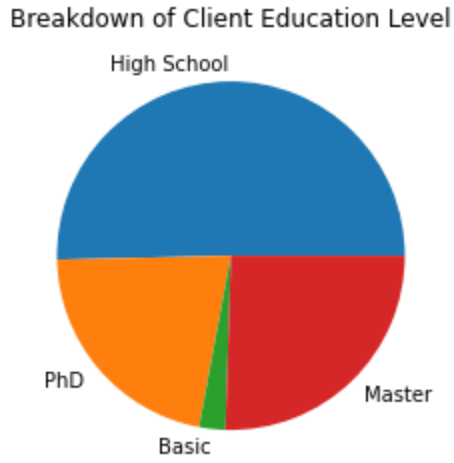


Figure 4. Breakdown of Client Education Level

Note: Most of the customers in this sample appear to be educated, most having at least a high school diploma.



Appendix B

Table 1. Multiple Linear Regression Results for the Total Amount Spent

OLS Regression Results

Dep. Variable:	Total_Spent	R-squared:	0.562
Model:	OLS	Adj. R-squared:	0.560
Method:	Least Squares	F-statistic:	353.6
Date:	Tue, 06 Dec 2022	Prob (F-statistic):	0.00
Time:	17:57:47	Log-Likelihood:	-14783.
No. Observations:	2216	AIC:	2.958e+04
Df Residuals:	2207	BIC:	2.964e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	21.6492	32.350	0.669	0.503	-41.790	85.089
Income	0.0060	0.000	32.527	0.000	0.006	0.006
Teenhome	-92.3763	8.042	-11.487	0.000	-108.147	-76.606
Kidhome	-177.5647	8.549	-20.771	0.000	-194.329	-160.800
Age	0.1340	0.376	0.357	0.721	-0.603	0.871
Education_Master	56.9673	27.968	2.037	0.042	2.122	111.813
Education_Graduation	65.3459	27.421	2.383	0.017	11.573	119.119
Education_PhD	83.4036	28.468	2.930	0.003	27.576	139.231
Marital_Status_Single	3.4778	8.515	0.408	0.683	-13.220	20.176

Omnibus:	1957.887	Durbin-Watson:	1.974
Prob(Omnibus):	0.000	Jarque-Bera (JB):	535049.662
Skew:	-3.343	Prob(JB):	0.00
Kurtosis:	78.829	Cond. No.	7.64e+05

Appendix C

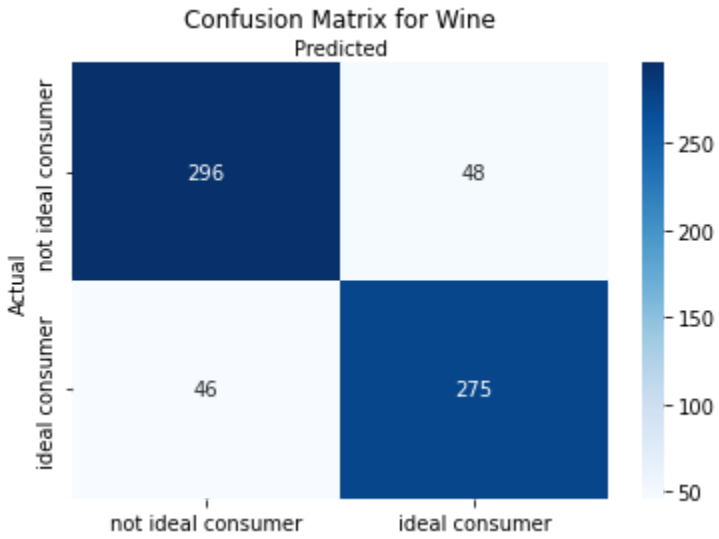
Table 2. Logistic Regression Results for Wine

Logit Regression Results

Dep. Variable:	Wines	No. Observations:	1551
Model:	Logit	Df Residuals:	1542
Method:	MLE	Df Model:	8
Date:	Tue, 06 Dec 2022	Pseudo R-squ.:	0.4641
Time:	17:57:47	Log-Likelihood:	-575.99
converged:	True	LL-Null:	-1074.9
Covariance Type:	nonrobust	LLR p-value:	4.407e-210

	coef	std err	z	P>|z|	[0.025	0.975]
const	-5.3585	1.097	-4.883	0.000	-7.509	-3.208
Income	8.932e-05	5.45e-06	16.398	0.000	7.86e-05	0.000
Teenhome	0.2647	0.150	1.769	0.077	-0.029	0.558
Kidhome	-1.2950	0.155	-8.352	0.000	-1.599	-0.991
Age	0.0081	0.007	1.179	0.238	-0.005	0.022
Education_Master	0.5611	1.059	0.530	0.596	-1.515	2.637
Education_Graduation	0.6317	1.055	0.599	0.549	-1.436	2.699
Education_PhD	0.9184	1.061	0.865	0.387	-1.162	2.999
Marital_Status_Single	-0.0538	0.157	-0.343	0.732	-0.362	0.254

Figure 5. Confusion Matrix for Wine



Appendix D

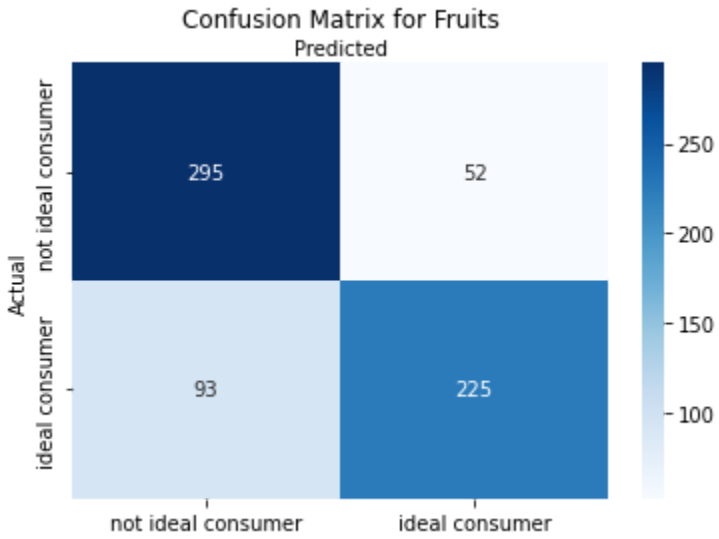
Table 3. Logistic Regression Results for Fruits

Logit Regression Results

Dep. Variable:	Fruits	No. Observations:	1551
Model:	Logit	Df Residuals:	1542
Method:	MLE	Df Model:	8
Date:	Tue, 06 Dec 2022	Pseudo R-squ.:	0.3377
Time:	17:57:49	Log-Likelihood:	-711.98
converged:	True	LL-Null:	-1075.0
Covariance Type:	nonrobust	LLR p-value:	1.719e-151

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.3913	0.500	-2.781	0.005	-2.372	-0.411
Income	6.838e-05	4.56e-06	15.006	0.000	5.95e-05	7.73e-05
Teenhome	-1.0412	0.139	-7.501	0.000	-1.313	-0.769
Kidhome	-1.0768	0.144	-7.459	0.000	-1.360	-0.794
Age	-0.0002	0.006	-0.030	0.976	-0.012	0.012
Education_Master	-0.9947	0.425	-2.342	0.019	-1.827	-0.162
Education_Graduation	-0.8920	0.416	-2.146	0.032	-1.707	-0.077
Education_PhD	-1.8406	0.440	-4.181	0.000	-2.703	-0.978
Marital_Status_Single	-0.0677	0.139	-0.488	0.625	-0.339	0.204

Figure 6. Confusion Matrix for Fruits



Appendix E

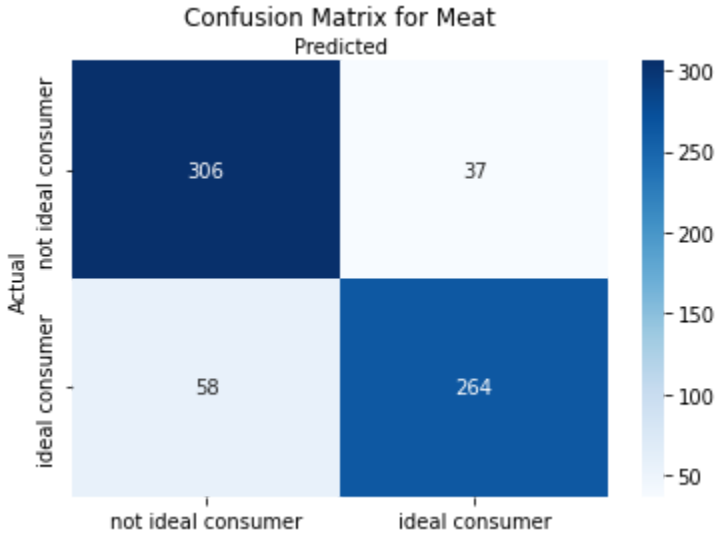
Table 4. Logistic Regression Results for Meat

Logit Regression Results

Dep. Variable:	Meat	No. Observations:	1551
Model:	Logit	Df Residuals:	1542
Method:	MLE	Df Model:	8
Date:	Tue, 06 Dec 2022	Pseudo R-squ.:	0.4523
Time:	17:57:52	Log-Likelihood:	-588.79
converged:	True	LL-Null:	-1075.1
Covariance Type:	nonrobust	LLR p-value:	1.249e-204

	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.5677	1.089	-4.194	0.000	-6.702	-2.433
Income	9.608e-05	5.68e-06	16.926	0.000	8.5e-05	0.000
Teenhome	-0.4728	0.150	-3.158	0.002	-0.766	-0.179
Kidhome	-1.1232	0.156	-7.201	0.000	-1.429	-0.817
Age	-0.0105	0.007	-1.510	0.131	-0.024	0.003
Education_Master	0.9556	1.053	0.908	0.364	-1.108	3.019
Education_Graduation	1.0089	1.049	0.962	0.336	-1.047	3.064
Education_PhD	0.6088	1.058	0.576	0.565	-1.464	2.681
Marital_Status_Single	-0.0021	0.155	-0.013	0.989	-0.305	0.301

Figure 7. Confusion Matrix for Meat



Appendix F

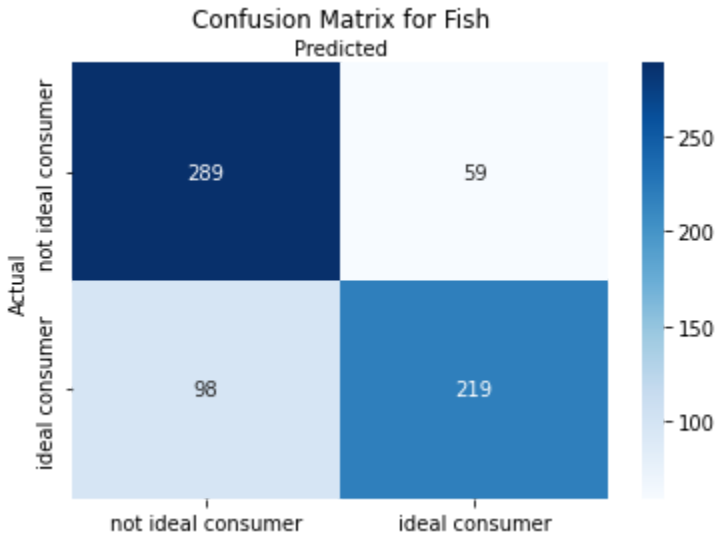
Table 5. Logistic Regression Results for Fish

Logit Regression Results

Dep. Variable:	Fish	No. Observations:	1551
Model:	Logit	Df Residuals:	1542
Method:	MLE	Df Model:	8
Date:	Tue, 06 Dec 2022	Pseudo R-squ.:	0.3063
Time:	17:57:55	Log-Likelihood:	-745.80
converged:	True	LL-Null:	-1075.0
Covariance Type:	nonrobust	LLR p-value:	6.143e-137

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.9703	0.494	-1.963	0.050	-1.939	-0.001
Income	5.682e-05	4.23e-06	13.429	0.000	4.85e-05	6.51e-05
Teenhome	-1.1576	0.136	-8.537	0.000	-1.423	-0.892
Kidhome	-1.2126	0.143	-8.488	0.000	-1.493	-0.933
Age	-0.0028	0.006	-0.470	0.638	-0.015	0.009
Education_Master	-0.5653	0.420	-1.345	0.179	-1.389	0.259
Education_Graduation	-0.4943	0.412	-1.199	0.230	-1.302	0.314
Education_PhD	-1.4214	0.435	-3.265	0.001	-2.275	-0.568
Marital_Status_Single	0.0023	0.135	0.017	0.986	-0.262	0.266

Figure 8. Confusion Matrix for Fish



Appendix G

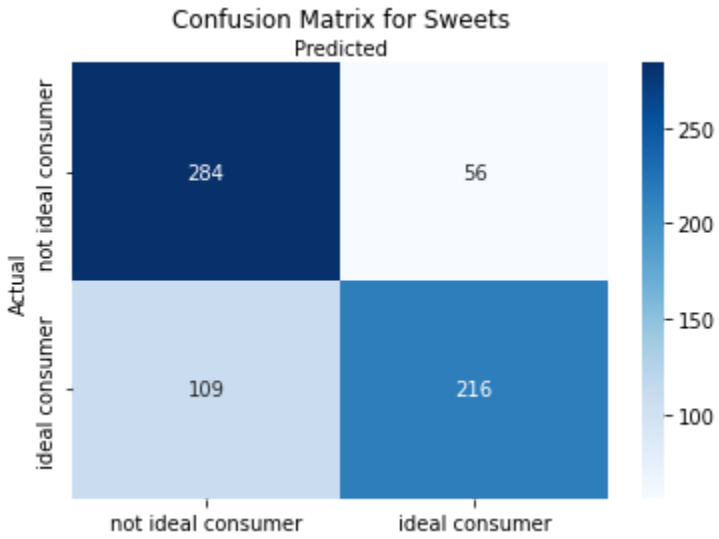
Table 6. Logistic Regression Results for Sweets

Logit Regression Results

Dep. Variable:	Sweet	No. Observations:	1551
Model:	Logit	Df Residuals:	1542
Method:	MLE	Df Model:	8
Date:	Tue, 06 Dec 2022	Pseudo R-squ.:	0.3049
Time:	17:57:57	Log-Likelihood:	-747.31
converged:	True	LL-Null:	-1075.1
Covariance Type:	nonrobust	LLR p-value:	2.722e-136

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.5696	0.477	1.195	0.232	-0.365	1.504
Income	5.729e-05	4.2e-06	13.631	0.000	4.9e-05	6.55e-05
Teenhome	-0.8588	0.131	-6.577	0.000	-1.115	-0.603
Kidhome	-1.2768	0.145	-8.786	0.000	-1.562	-0.992
Age	-0.0192	0.006	-3.131	0.002	-0.031	-0.007
Education_Master	-1.6875	0.396	-4.260	0.000	-2.464	-0.911
Education_Graduation	-1.1757	0.384	-3.059	0.002	-1.929	-0.422
Education_PhD	-2.2605	0.410	-5.510	0.000	-3.065	-1.456
Marital_Status_Single	-0.0487	0.135	-0.362	0.718	-0.313	0.215

Figure 9. Confusion Matrix for Sweets



Appendix H

Table 7. Summary of findings.

	Significant Variables	Training Data Score	Test Data Score
Total Purchases Amount in a Year	- Income - Teenhome - Kidhome - Education_Master - Education_Grduation - Education_PhD	NA	NA
Wine	- Income - Kidhome	86.91%	85.86%
Fruits	- Income - Teenhome - Kidhome - Education_Master - Education_Grduation - Education_PhD	79.75%	78.19%
Meat	- Income - Teenhome - Kidhome	85.56%	85.71%
Fish	- Income - Teenhome - Kidhome - Education_PhD	79.30%	76.39%
Sweets	- Income - Teenhome - Kidhome - Age - Education_Master - Education_Grduation - Education_PhD	78.78%	75.18%