Customer Personality Analysis

CREATED BY: Uma Rajagopalan

Introduction

Our team's dataset name is '**Customer Personality Analysis**'. This type of analysis helps businesses identify the customers' profiles needed for certain products. Since analyzing a completely new customer profile for every new product can be costly, by understanding customers' buying behavior, marketing agencies can customize products or sale deals according to the customer's needs and behaviors.

The dataset contains customers' data and their buying information from local grocery stores. Some common attributes in the dataset are customer's education, income, amount of meat or wine, or even customer's purchase channel (Figure 1). Within the dataset, there are 29 columns and 2240 rows, and 29 columns. The goal is to utilize customer personality analysis for targeted advertising.

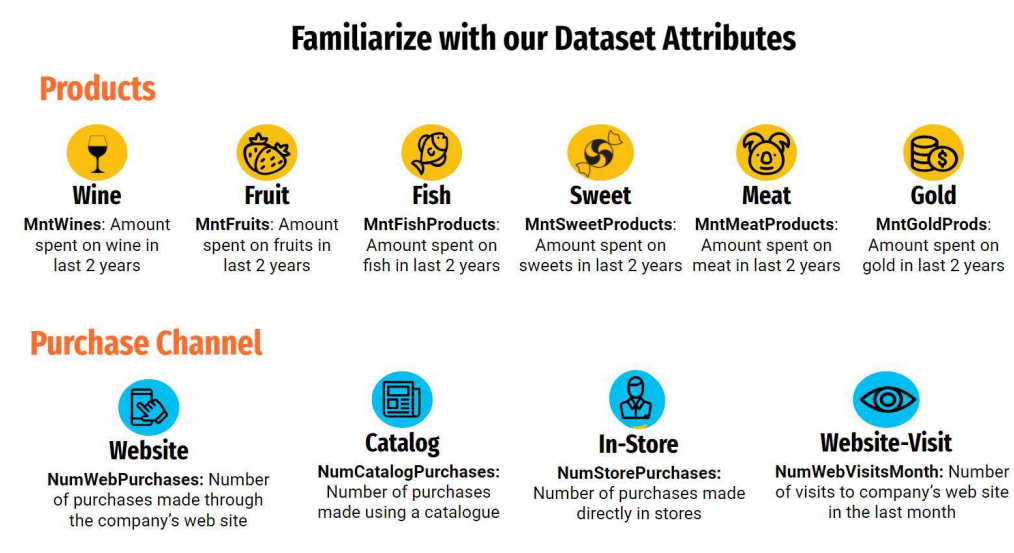


Figure 1: Attributes in the dataset

Motivation: Our team was drawn to this dataset because all of us are avid consumers who enjoy shopping and are passionate about strategic marketing. Our team members are interested in learning about the process of analyzing a customer's profile and how that analysis will help increase targeted marketing efficiency.

Driving Question: Since our goal is to learn the customer's buying behaviors to increase our marketing effectiveness, our driving question is:

"How would we improve marketing strategy by analyzing customers demographics?"

Data Cleaning

Data cleaning is the process of fixing or removing incorrect, incorrectly formatted, duplicate, or incomplete data within a dataset.

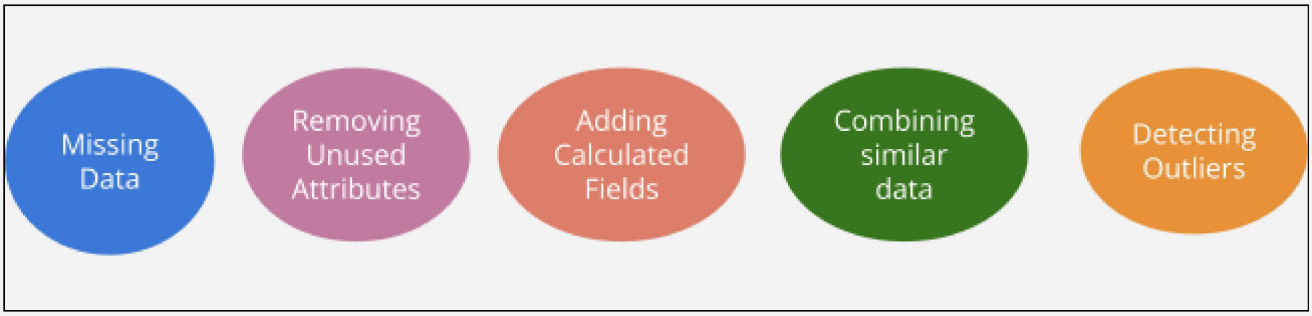


Figure 2: Parts of the cleaning process

1. Missing Data:

To find the missing data, we subtract the count of non-empty rows from the total rows of the data frame.

Out of the 29 columns in our dataset, only 1 column ('Income') had 24 missing values. Even though the missing values accounted for 1.07% of the total data, we decided to fill the missing values with the mean of the Income Series to not lose any insights.

2. Removing Unused Attributes

Columns: ID, AcceptedCmp1, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5,Z_CostContact, Z_Revenue are not used in the data analysis process of this project. So, it is better to remove these attributes.

3. Adding Calculated Fields

The column 'Year' is useful to determine the age of the customers in the dataset. So we will use year to form a new column called 'Age'. We will then drop the column 'Year'.

4. Combining Similar Data

The following columns were added to analyze the data.

Marital Status: The column Marital Status had 8 different kinds of relationship statuses of the customers. We brought it down to just two categories: Single, Relationship to properly use the column for our data analysis.

Total Kids: Sum of 'kid home' and 'teen home' columns.

Total Amount spent: Sum of the amount spent on all products

Total Purchases: Sum of purchases made from all buying channels.

5. Detecting Outliers

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.

Through plotting the Numerical values like Income and Age in a box plot, we determined the outliers and removed them. The before and after results look like follows

Income:

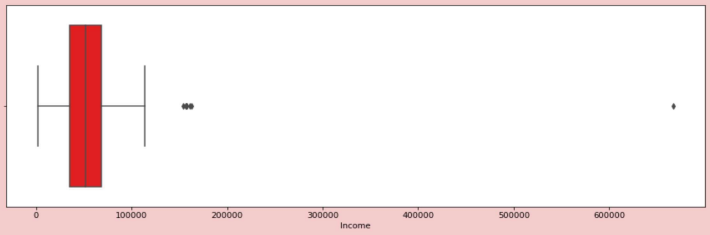
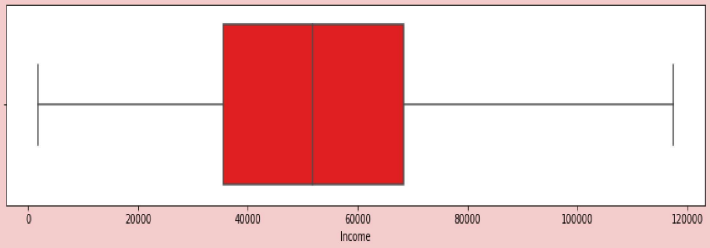
Before	After
	

Figure 3: Before and After removing outliers in Income

Age:

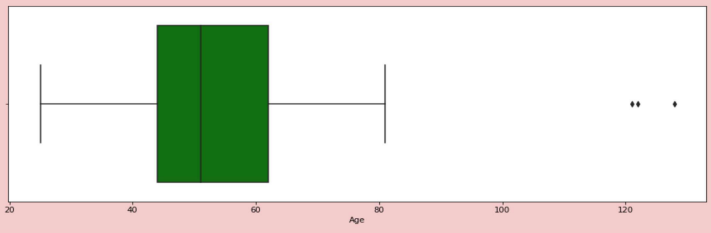
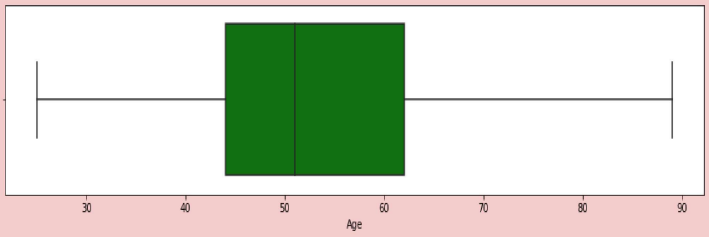
Before	After
	

Figure 4: Before and After removing outliers in Age

Final Data Set

Final Dataset Our Team uses for Data Analysis!
(first 5 rows)

	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	MntFruits	MntMeatProducts
0	Graduation	Single	58138.0	0	0	04-09-2012	58	635	88	546
1	Graduation	Single	46344.0	1	1	08-03-2014	38	11	1	6
2	Graduation	Relationship	71613.0	0	0	21-08-2013	26	426	49	127
3	Graduation	Relationship	26646.0	1	0	10-02-2014	26	11	4	20
4	PhD	Relationship	58293.0	1	0	19-01-2014	94	173	43	118

MntFishProducts	MntSweetProducts	MntGoldProds	NumDealsPurchases	NumWebPurchases
172	88	88	3	8
2	1	6	2	1
111	21	42	1	8
10	3	5	2	2
46	27	15	5	5

NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth	Complain	Response	Age	Total_Kids	Total_Amount_Spent	Total_Purchases
10	4	7	0	1	64.0	0	1617	25
1	2	5	0	0	67.0	2	27	6
2	10	4	0	0	56.0	0	776	21
0	4	6	0	0	37.0	1	53	8
3	6	5	0	0	40.0	1	422	19

Exploratory Data Analysis

After making the dataset suitable for our data analysis, we plotted a heatmap to get some insights on how our variables correlate and better understand their relationships.

The heatmap shows some strong correlations (either positive or negative) among customers' demographics and the products and channels they use. For example, we can see that customers' income strongly correlates to products bought such as wine and meat. We can also see that income correlates positively with store purchases and negatively with web purchases. On the other hand, we can see that customers' age doesn't show any strong correlation with either products or channels.

The insights gained from this step guided our data analysis and we were able to concentrate on the meaningful relationships and further analyze and quantify them.

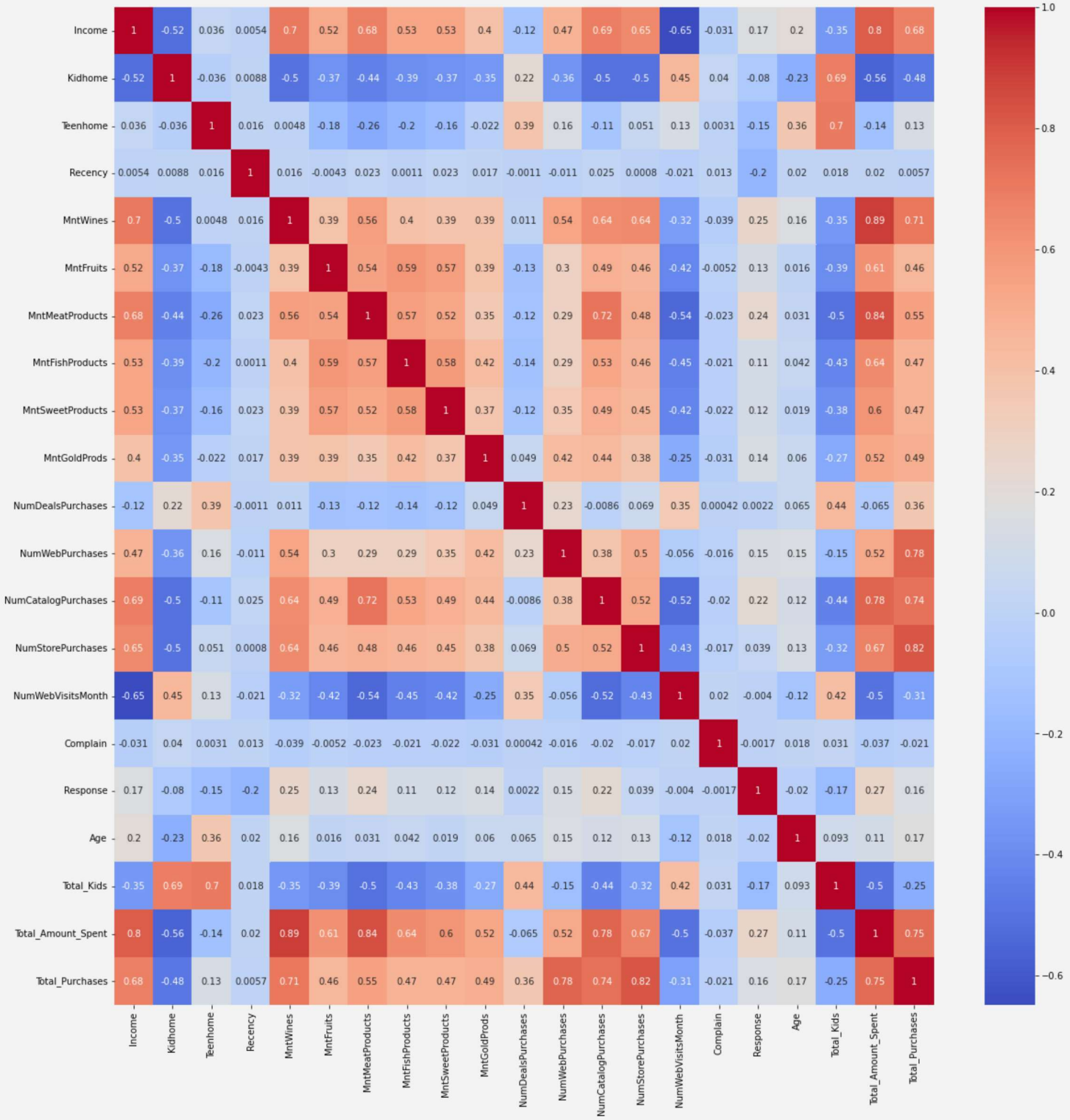


Figure 5: Heat map representing the correlation of all variables

Data Analysis & Findings

We will dive in now to analyze how customers' demographics affect their buying options and channels they use to buy. This will give suggestions to the marketing team to improve the marketing strategy by implementing targeted advertising. By understanding customers' buying behavior, marketing can customize products or sale deals according to the customer's needs and behaviors. We will use different methods and plots to answer how specific customers' demographics affect their buying options.

1. How does education impact products bought?

For determining how education impacts products bought, we analyzed education affects buying behavior at comprehensive and granular levels.

The following bar graphs represent how education impacted buying trends:

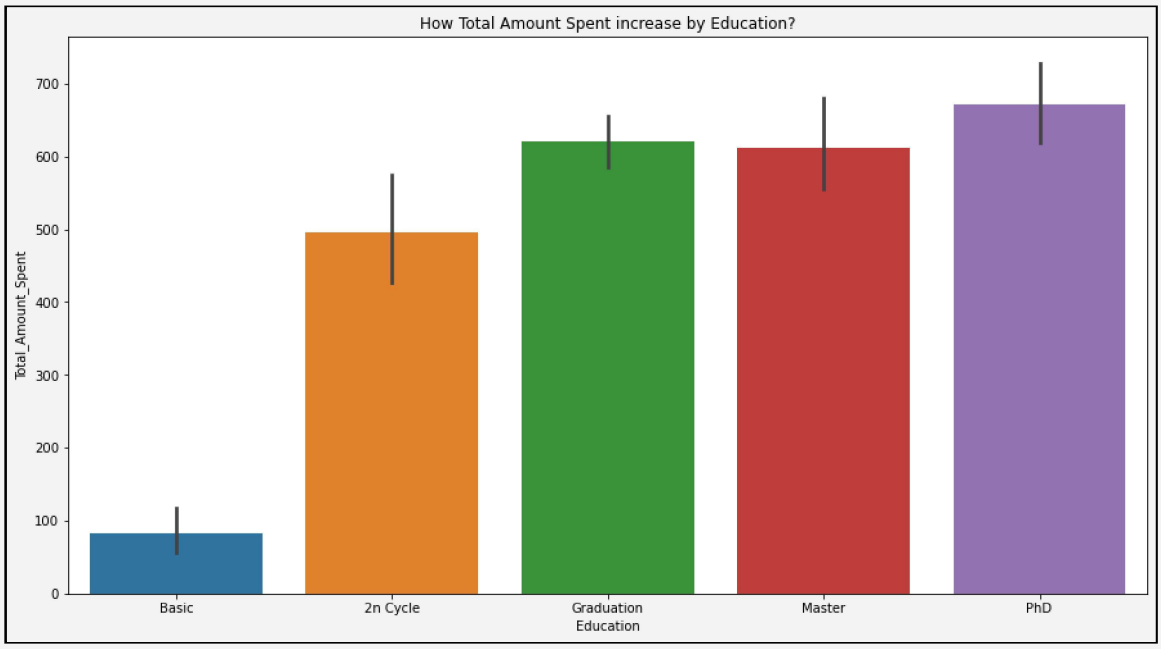


Figure 6: Education vs total amount spent

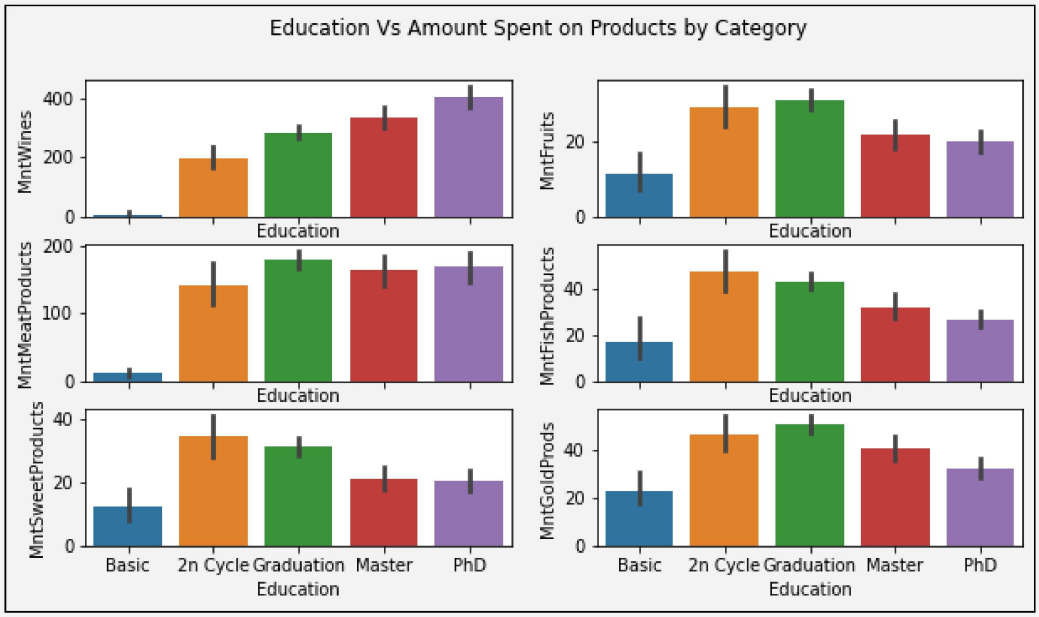


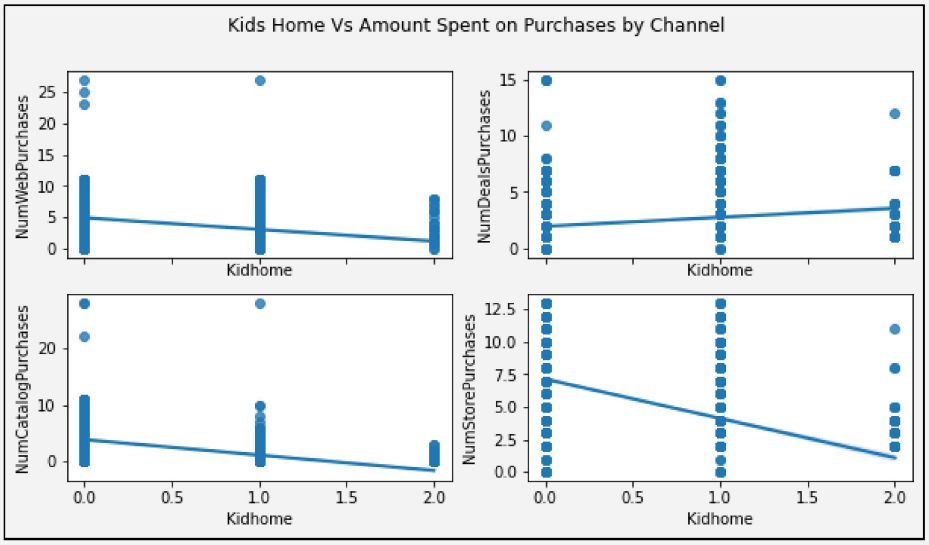
Figure 7: Education vs amount spent on products (by each category)

People with basic education tend to spend less in every product category. As the level of education increases, the amount spent on all product categories also increases (Figure 4).

Education vs Individual Products:
*The higher the education, the higher the amount spent on Wines and Meat
*The higher the education, the lower the amount spent on Sweets, Fish
*Customers with Graduate education tend to buy more Meat, Gold, and Fruits than the other levels.

2. How does the number of children affect the customer's buying channel?

Figure 8: Number of kids at home vs amount spent on purchases by buying channel



To show how the number of kids at home affects how the customers buy and what channels they utilize more, we used a combination of reg plots.

Reg plots show that the greater the number of kids at home, the more purchases are made through deals, catalog and web purchases.

Also, customers with more kids at home tend to visit the store the least.

3. How would having kids or teenagers affect the customer's use of websites?

Since we already discovered that there is an interesting relationship between having kids and buying behavior from the previous question, we want to explore whether parents will utilize the website of the store.

We created two reg plots to show the number of website visits for parents with kids and parents with teenagers.

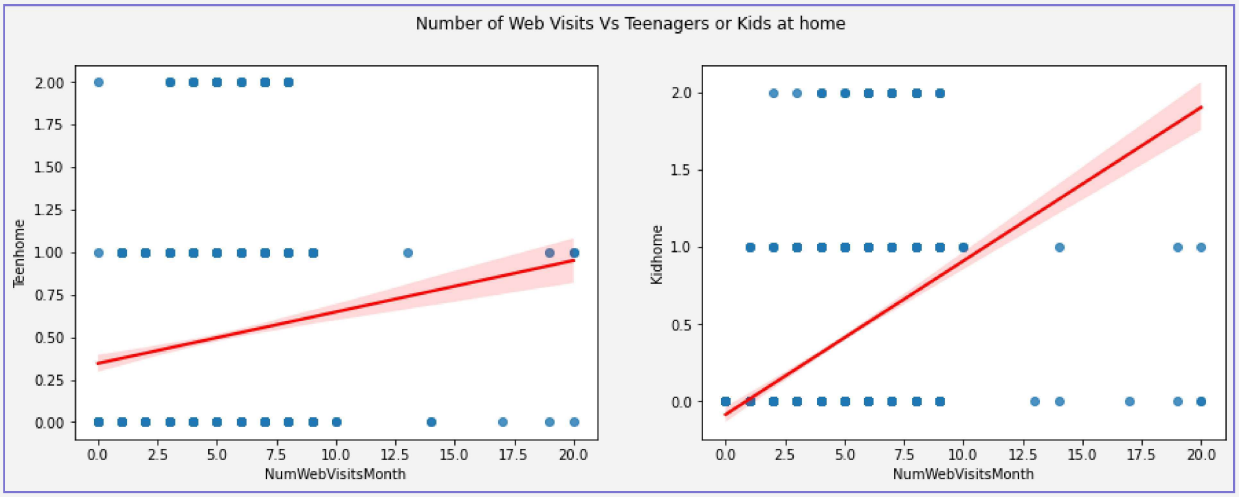


Figure 9: Number of web visits vs number of teenagers or kids at home

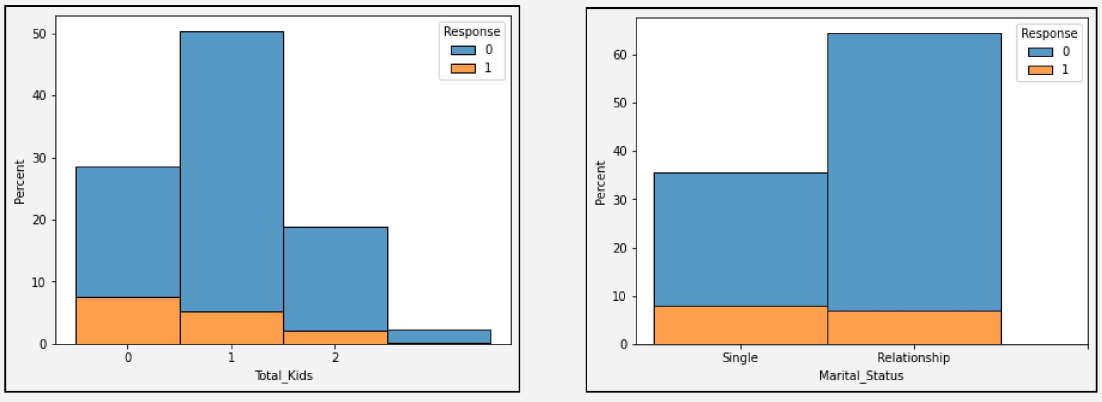
It is clear that customers who have kids tend to visit online stores more than people who have teenage kids.

4. How do people respond to marketing campaigns based on family demographics (marital status and number of children)?

To answer this question we used two methods: histplot and logistic regression, which we describe in the next section.

We created two hist plots to show how marital status and the number of kids affect the customer's response (0/1) to marketing campaigns. Single status and no kids at home are the traits that show a better response to the campaigns.

Figure 10: Marital status and the number of kids



5. How does income affect total spending?

We want to further analyze how income affects the total spending of the customers and quantify this relationship.

To answer this question we used scatterplot, which will be described in the data modeling section.

The scatter plot clearly shows a linear relationship: as the income increases, the total amount spent on products increases.

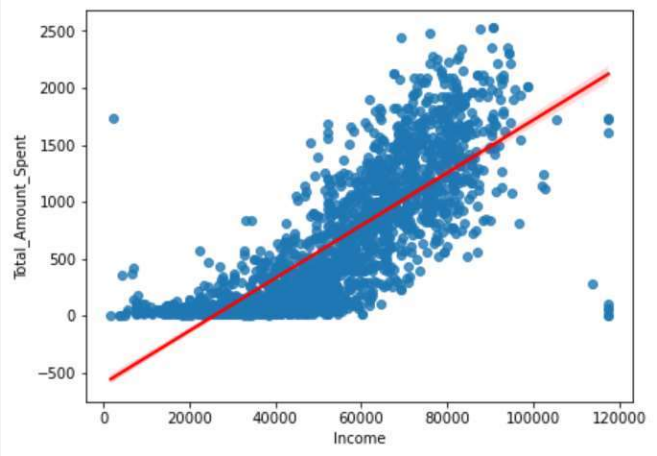


Figure 11: Income vs total amount spent

Data Modeling

1. Logistic Regression

Model: Marketing Campaign Response vs Marital_Status & Total_Kids

Logit Regression Results

Dep. Variable:	Response	No. Observations:	2240
Model:	Logit	Df Residuals:	2237
Method:	MLE	Df Model:	2
Date:	Wed, 08 Dec 2021	Pseudo R-squ.:	0.05991
Time:	22:49:53	Log-Likelihood:	-886.87
converged:	True	LL-Null:	-943.39
Covariance Type:	nonrobust	LLR p-value:	2.849e-25

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.5243	0.106	-14.370	0.000	-1.732	-1.316
Marital_Status[T.Single]	0.8151	0.122	6.686	0.000	0.576	1.054
Total_Kids	-0.6872	0.090	-7.653	0.000	-0.863	-0.511

Figure 12: Logistic regression table between marketing campaign response and marital status with kids

The results from the regression analysis confirm that marital status and the number of children are good predictors for Response to Campaigns. The p-values in the table suggest they are both significant variables for this model.

Single marital status shows a positive effect on the response while the number of children shows a negative effect on the response.

2. Linear Regression

```
model = smf.ols(formula = 'Total_Amount_Spent ~ Income',
                data = MarketingCampaignDF)

results = model.fit()
```

```
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Total_Amount_Spent   R-squared:                       0.646
Model:                            OLS   Adj. R-squared:                  0.646
Method:                 Least Squares   F-statistic:                     4091.
Date:                Wed, 08 Dec 2021   Prob (F-statistic):               0.00
Time:                        22:32:06   Log-Likelihood:                -16351.
No. Observations:                2240   AIC:                         3.271e+04
Df Residuals:                    2238   BIC:                         3.272e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   -593.9632     20.226    -29.366      0.000    -633.627    -554.299
Income         0.0231      0.000     63.964      0.000       0.022       0.024
==============================================================================
Omnibus:                      180.937   Durbin-Watson:                   2.012
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1010.425
Skew:                           0.098   Prob(JB):                    3.88e-220
Kurtosis:                       6.284   Cond. No.                     1.49e+05
==============================================================================
```

Figure 13: Total amount spent vs Income

The result from the regression shows that R-squared = 65%, which means that this model fits most of our data and that income is a good predictor for the total amount spent on products overall. Low p-values suggest that Income is a significant variable for this model.

We can predict customers' total spent based on their income with the following formula:

*Total Spent = -229.5615 + 0.0160 * Income*

Both models provide insights into how customers' demographics affect their buying options and products.