1. Dataset Background

The dataset for this project comes from a well-established company in the retail food industry in Brazil, which has a long history of providing a wide variety of products, including wines, rare meats, exotic fruits, specially prepared fish and sweet products. The primary objective of this project is to maximize profits by analyzing customer characteristics and identify those most likely to purchase a new gadget during an upcoming direct marketing campaign. The project aims to enhance future marketing strategies of the company and maximize profit for upcoming campaigns.

2. Dataset Explanation

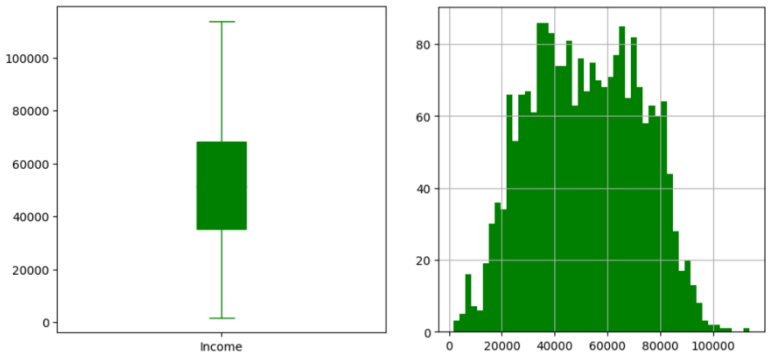
To achieve the aforementioned objective, the company wants a predictive model. In fact, a pilot campaign involving about 2000 random customers was carried out to facilitate model construction, labeling those who accepted the pilot offer and those who did not, which eventually resulted in our dataset. Some of the important known variables provided are: customer's responses to previous campaigns (campaign 1, 2, 3, 4, 5) as well as the pilot campaign, level of education (high school graduate / university graduate / postgraduate / Master/ PhD) , marital status (single / in a relationship / married / divorced / widowed), number of kids as well as teenagers at home, yearly household income, age, amount spent on different types of products (wines, meats, fruits, fish, sweet products) in the last 2 years, number of purchases made through different channels (website, stores, catalogue), days of being customer, and number of days since last purchase. Meanwhile, some of the unknown variables that might affect our result are customer's job, health status, address, and eating habits. In short, this data set is believed to be helpful in assisting company's decision making because given customer's reaction to the pilot campaign and other characteristics, one might be able to identify the relationship between them.

3. General Customer Behavior

Some basic data visualization and exploration have been done on all the customers in this data set to obtain a brief understanding regarding the customer behavior.

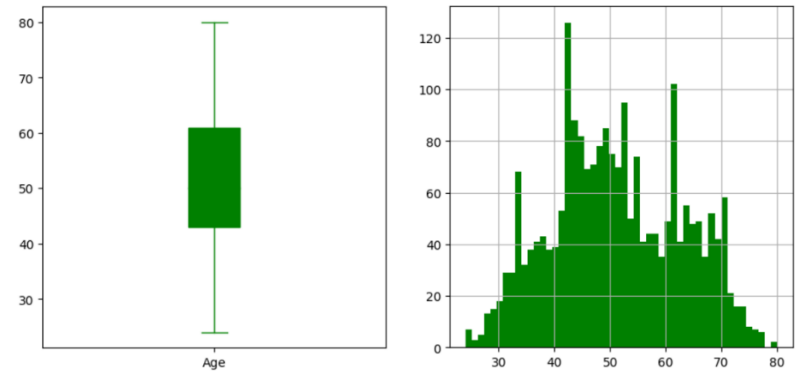
Demographic Characteristics

Income



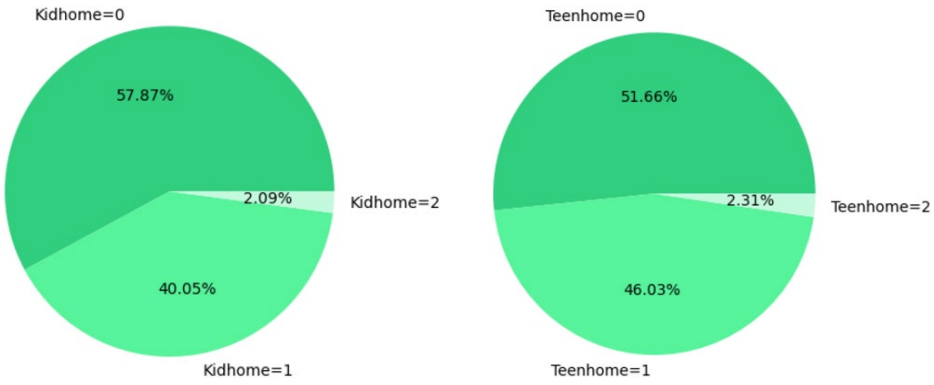
It is observed that the majority of customers have a yearly household income of around $40000 to $60000 with a mean of $51622

Age



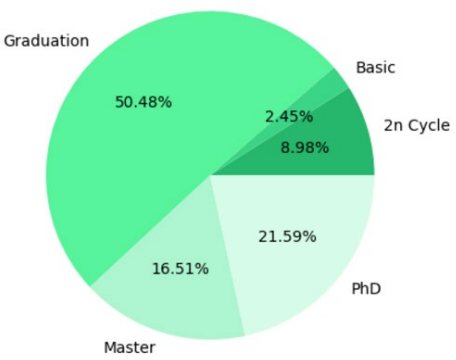
The age of customers is around 40 to 60 years old with a mean of 51.

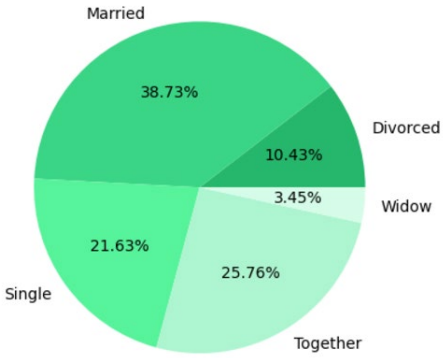
Family Structure



It is found that most customers have 0 to1 children as well as 0 to 1 teenagers in their household.

Education Level & Marital Status





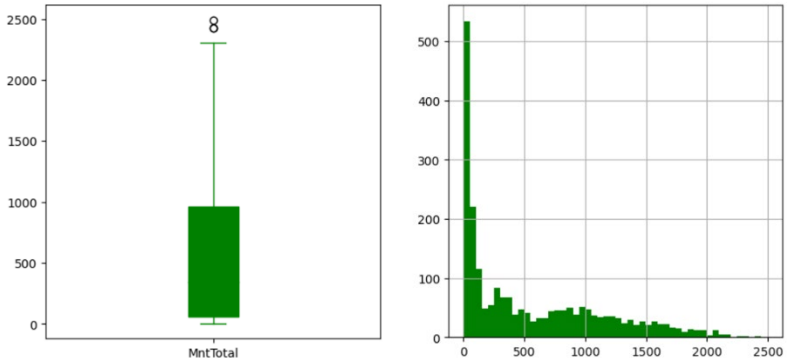
The majority of customers are un university graduated, followed by PhD. On the other hand, most of the customers are married, followed by being in a relationship (Together).

Short Conclusion

One can say that the customers of this company tend to be graduated, married, middle-aged, and with an yearly household income around $40000 to $60000.

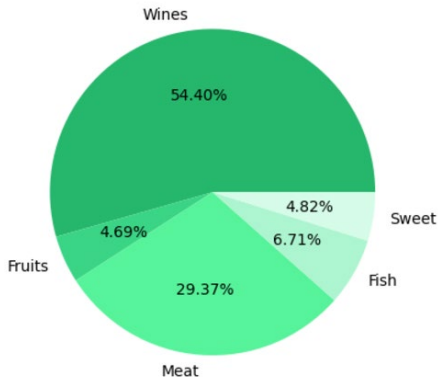
Spending Habits

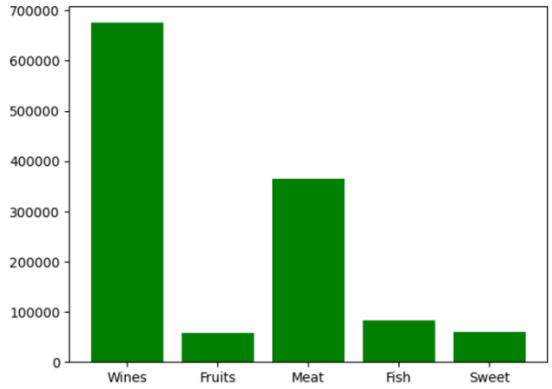
Total Amount Spent by Each Customer



The total amount spent by each customer in the last 2 years is around $0 to $1000 with a mean of $563.

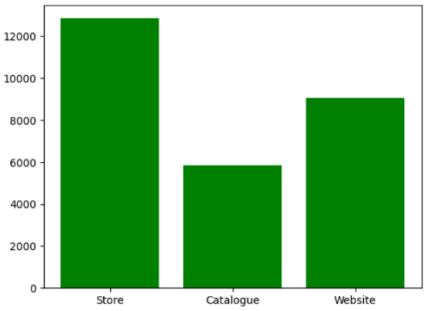
Amount Spent on Different Products by All Customers

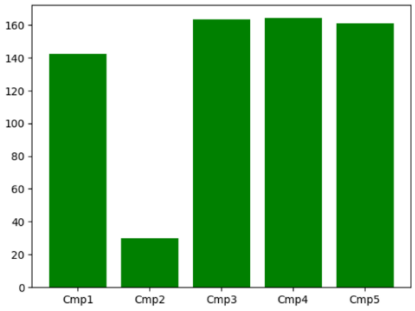




It is observed that wine is the most spent product in the last 2 years while the second most spent is meat.

Customer's Number of Purchases Made with Different Channels & Campaigns' Number of Offers Accepted by Customers





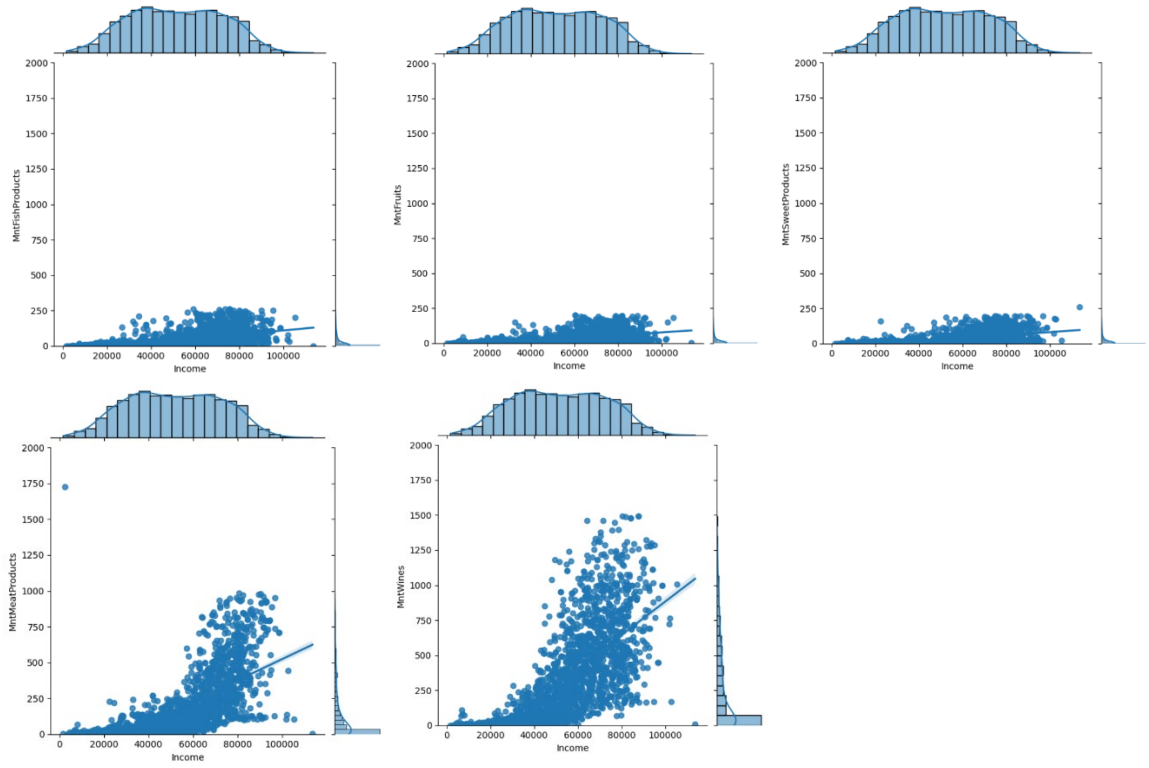
It is found that physical store is the most popular channel, receiving the most purchases, followed by company's website. Additionally, campaign 3, 4, and 5 are the top three successful campaigns, generating the most offers accepted.

Short Conclusion

To conclude, wines and meat are the top two products customers spent on, plus, customers prefer campaign 3, 4, and 5, as well as purchasing in stores. Last but not least, they usually spent around $0 to $1000 in total in the last 2 years.

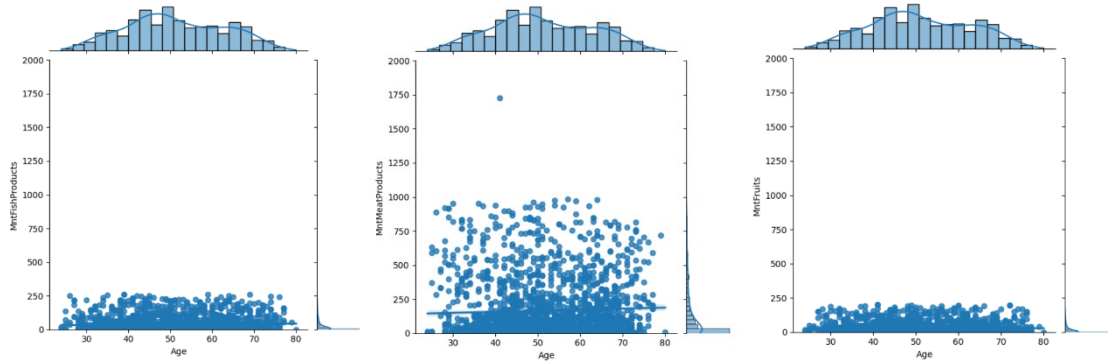
Regression Analysis

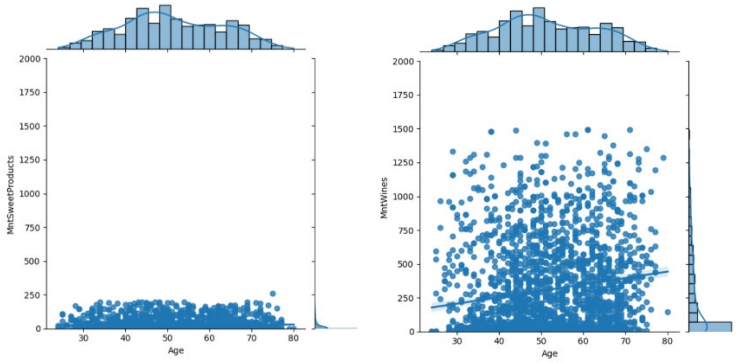
Customer's Income VS Amount Spent on Different Products



It is found that lines of regression for all types of products are upward sloping, while the ones for wines and meat have greater slopes than the others. This implies that customer's amount spent on all products increases as their income increase, especially for wines and meat.

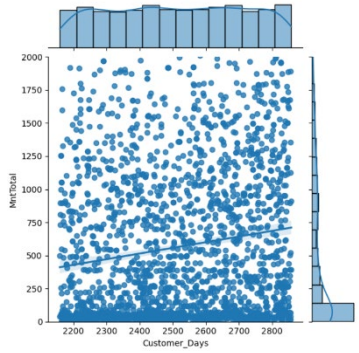
Customer's Age VS Amount Spent on Different Products





It is observed that the lines of regression for fish, meat, fruits, and sweet products are nearly horizontal, but significantly upward sloping for wines. Hence, we say that customer's amount spent on different products barely changes as their age changes, except for wines, which increases obviously.

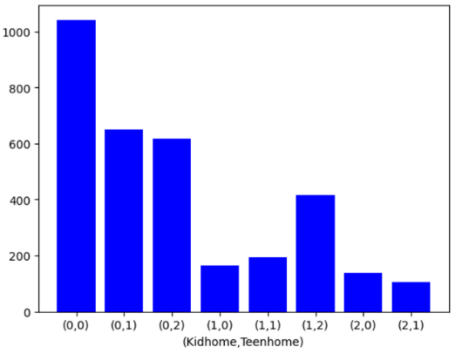
Time of Being Customer VS Total Amount Spent



The line of regression is upward sloping, thus the longer time someone has become a customer of the company, the more he spent on purchasing in the last 2 years.

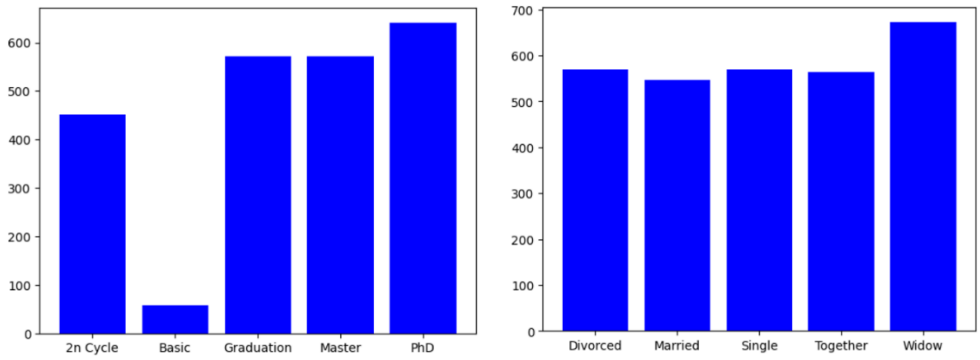
Simple Comparison Analysis

Total Amount Spent by Customers with Different Number of Children and Teenagers



Those with no children nor teenagers spent the most in the last 2 years, followed by those with no children but one teenager, and two teenagers.

Total Amount Spent by Customers with Different Education Level and Marital Status



We find that those widowed and with PhD education level usually spent more than the others in the last 2 years.

Short Conclusion

Those with no children nor teenagers, PhD education level, widowed tend to have a greater total amount spent.

Correlation Analysis

Column Pairs with Large Correlation

Income & Total Amount Spent	0.82
Income & Amount Spent on Wines	0.73
Income & Amount Spent on Meat	0.70

Column Pairs with Small Correlation

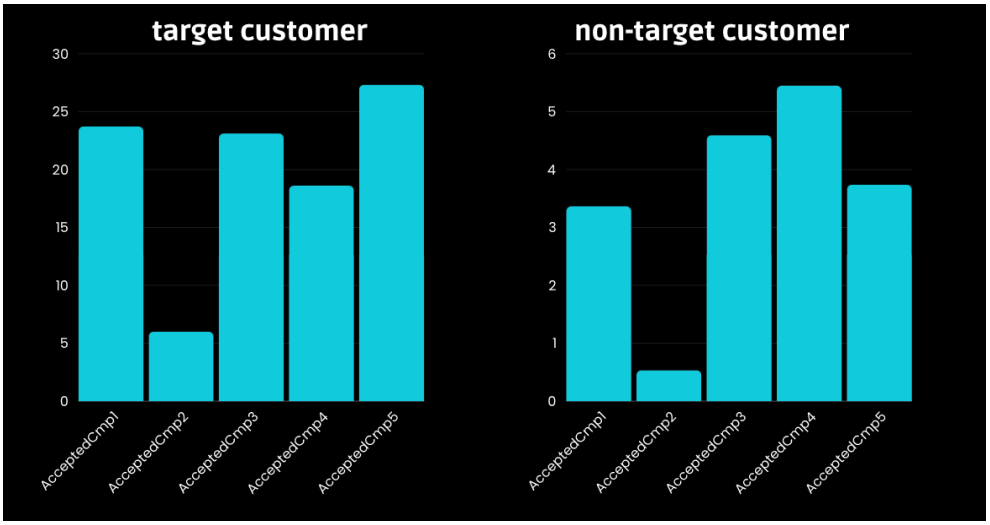
Number of Children & Total Amount Spent	-0.55
Number of Children & Income	-0.53
Number of Children & Amount Spent on Wines	-0.50

The result implies that the income and number of children are both important factors affecting customer's spending, the former has a positive impact on it while the latter has a negative one. Moreover, this is in align with our conclusion made earlier from regression analysis and comparison analysis.

4. Comparison between target and non-target customer

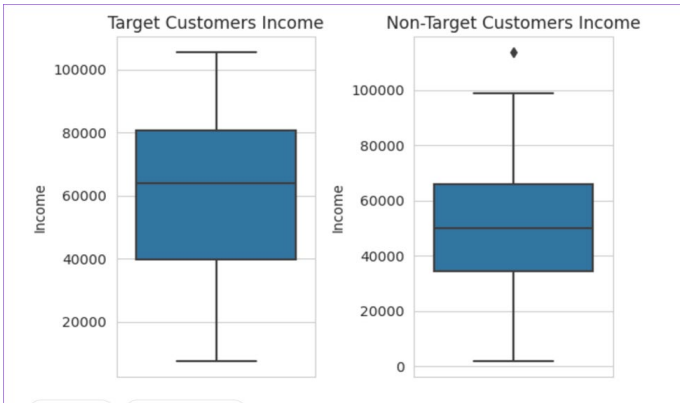
In analyzing customer behavior for the upcoming marketing campaign, it is essential to compare the characteristics of target customers, who are likely to purchase the new gadget, with non-target customers. This comparison provides valuable insights that can understand unique characteristics of target customers, which can help us build a better model to predict who might become our target customers in the future. By identifying these characteristics, we can refine marketing strategies and maximize profits for the upcoming marketing campaign.

Campaign acceptance rates



Campaign acceptance rates reveal distinct differences between target customers and non-target customers. In specific, the target customers have significantly higher acceptance rates across various marketing campaigns compared to non-target customers. Plus, target customers show a strong preference for campaign 5, while non-target customers show more interest for campaign 4. This indicates that target customers have a higher engagement level in marketing campaigns and are more responsive to promotional offers.

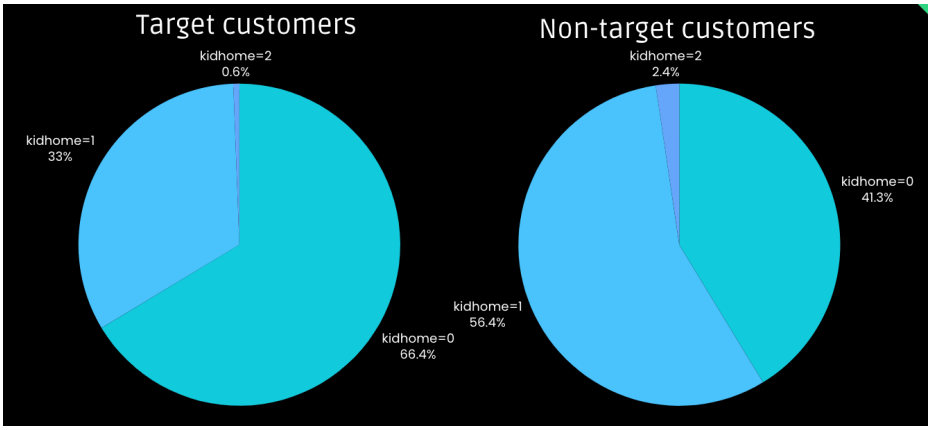
Income



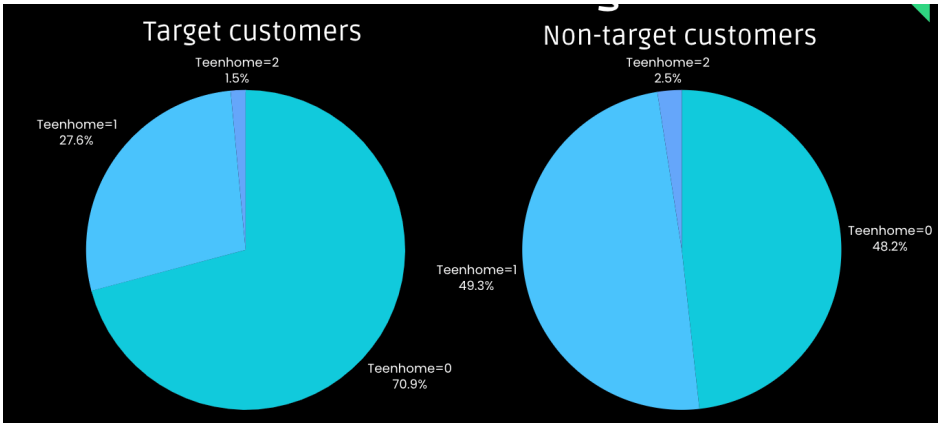
For income levels, target customers generally earn more than non-target customers. The median income of target customers is about $64,000, while many non-target customers fall below this level.

Family structure

Number of kids

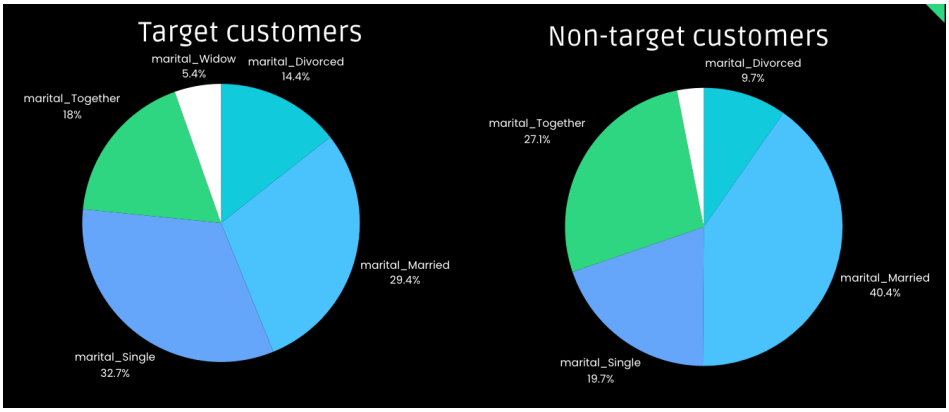


Number of teenagers



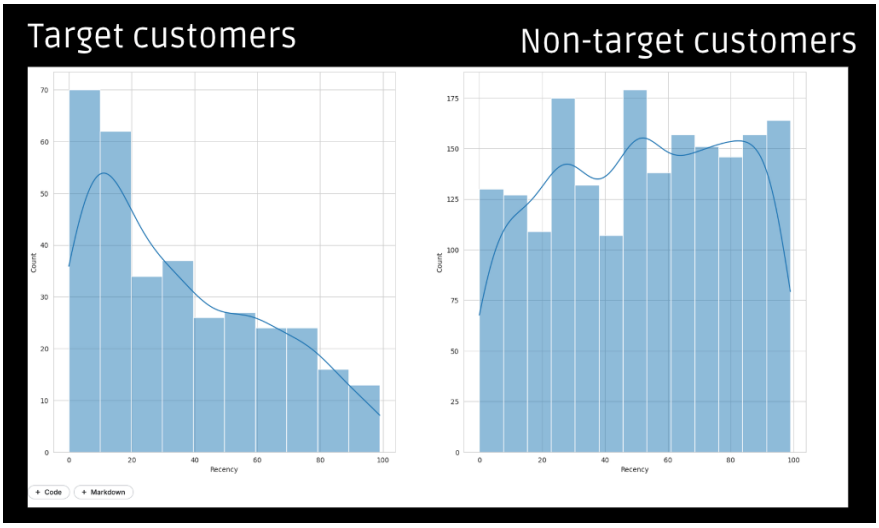
Family structure also differentiates the two groups. A large percentage of target customers do not have kids or teenagers; about 66.4% and 70.9% of target customers do not have kids and teenagers respectively.

Marital status



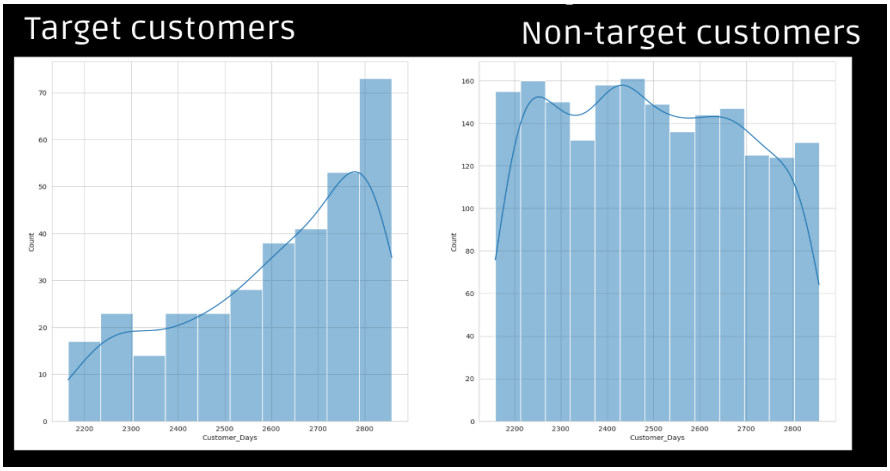
Marital status is an important factor to separate target and non-target consumers. Among target customers, about 32.7% are single and 29.4% are married, while about 19.7% of non-target customers are single and 40.4% are married.

Recent purchase



Recent purchase activity shows that target customers make purchases more frequently.

Customer days



Customer days, which represent the duration of time a customer has been with the company, provide important insights into customer loyalty and engagement. Target customers generally have longer customer days compared to non-target customers. This indicates that they have maintained a longer relationship with the company.

Short conclusion

In conclusion, target customers have significantly higher campaign acceptance rates, particularly favoring campaign 5, higher income levels and more likely do not have kids or teenagers. They are also more likely to be single and have been loyal customers for a longer duration.

Other factors, such as channel preferences, product preferences, education levels and age, did not show significant differences between the two groups. The analysis of these factors is included in the appendix for further reference.

5. Model construction

Based on the previous analysis, the model construction focuses on selecting customers as features that meet at least one of the following criteria. We use median as a key threshold for several criteria because it is less affected by outliers compared to mean.

Customers have accepted at least one marketing campaign, excluding campaign 4 and 2.
Customers have an income higher than the median income of target customers ($64,090).
Customers do not have kids or teenagers.

Customers are single.
Customers' latest purchase days should be within the last 30 days. 30 days is the median of the latest purchase days in target customers.
They have been a customer of the company for longer than the median duration of target customers (2651 days)

We chose to exclude campaign 4 from our analysis because it is the most popular campaign among non-target customers. We also exclude campaign 2 because it is not an importance factor for classifying customers as it is unpopular among all customers.

There are two models particularly useful for classification, logistic regression model and random forest classifier. Logistic regression models the probability that a given input belongs to a particular category using a logistic function, which outputs values between 0 and 1. In contrast, the random forest classifier is good at handling complex datasets with multiple variables and can capture non-linear relationships.

We adopted the logistic regression model because it is particularly suitable for our needs. It provides clear insights into the likelihood of customers belonging to the target customers. To avoid data leakage, we separate the data into training and testing sets.

The logistic regression model achieved an accuracy of 88%, better than the random forest classifier, which had an accuracy of 85%. Additionally, logistic regression has lower false positive and false negative rates, with 40 false positives compared to 74 in random forest classifier, and 167 false negatives compared to 174. This indicates that our approach using logistic regression is effective in accurately identifying target customers.

6. **Conclusion**

In this project, we analyzed the customers comprehensively, from their demographic characteristics, spending habits, to unique features the target customers (i.e. those who are likely to purchase the new gadget in the upcoming marketing campaign) have. Our findings show that the key difference between target customers and non-target customers will be that target customers usually have a higher campaign acceptance rate, a higher income level, no children nor teenagers, and are single.

Based on what we have observed, we constructed our predictive model using logistic regression and managed to obtain an accuracy of 88%, which is better than what random forest classifier can obtain, 85%. Plus, when compared with random forest classifier, our

model also shows lower false positive and false negative rate, implying that it is able to identify the target customers effectively.

To sum up, not only does our project enhance company's understanding on the customer base by providing a comprehensive analysis, but it also helps the company better identify their target customers in the upcoming campaign by constructing a reliable model. By making use of all these, the company will be able to design a consummate marketing strategy to use in the coming event and thus maximize their profits.

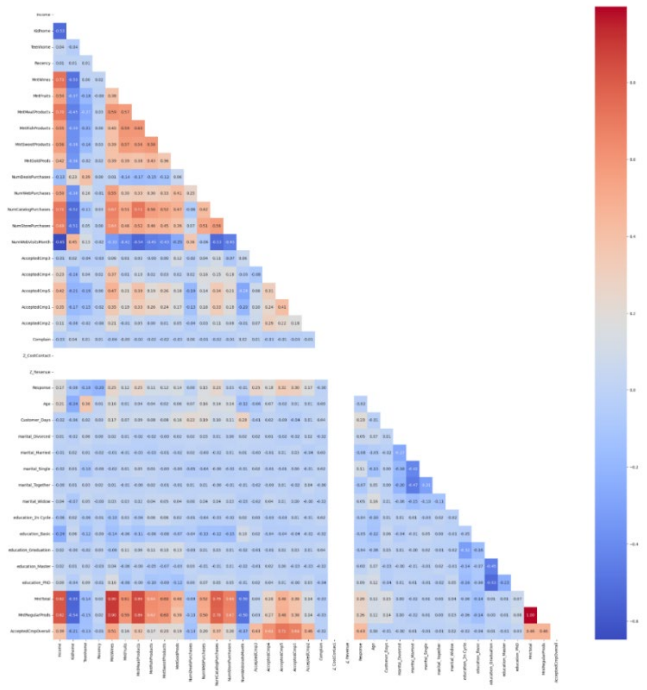
Last but not least, we have also obtained a meaning experience in getting insights from a large dataset by making use of different types of analysis methods, such as regression analysis and correlation analysis. Additionally, we have also practiced on model construction and model performance evaluation. We believe that we are now more capable of dealing with different kinds of datasets.

7. Appendix

Dataset comes from Kaggle
https://www.kaggle.com/datasets/jackdaoud/marketing-data

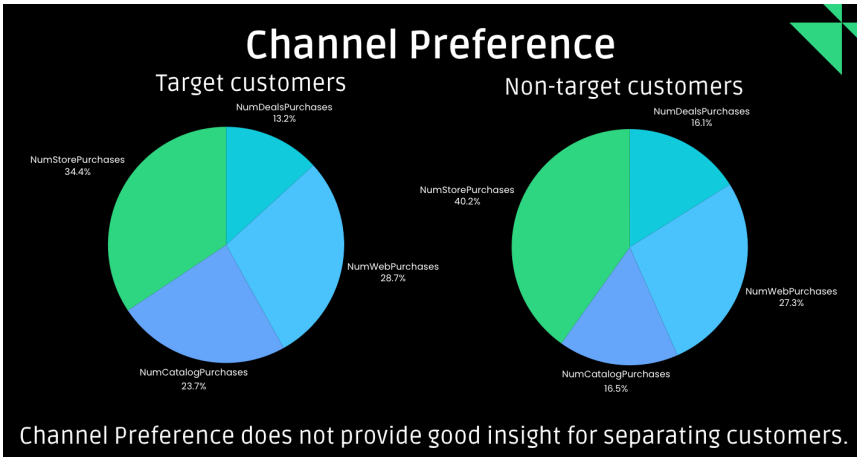
Code for General Customer Behavior section
https://www.kaggle.com/code/mannixyam/ftec4002-project-marketing

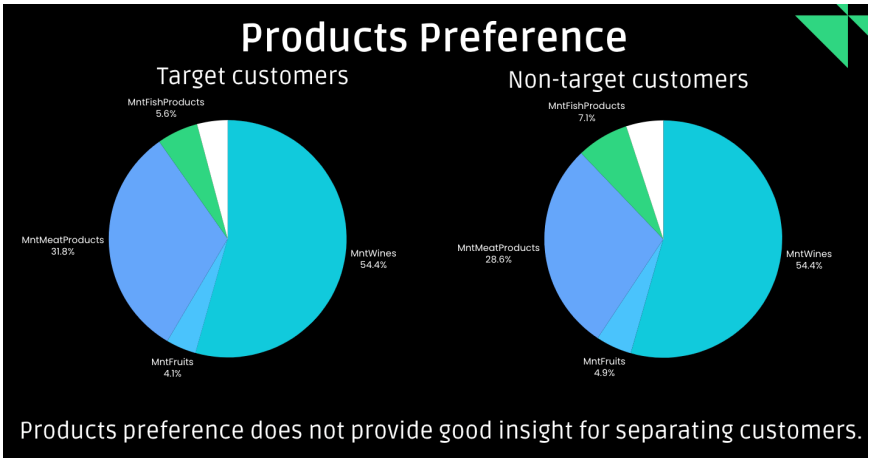
The complete correlation matrix

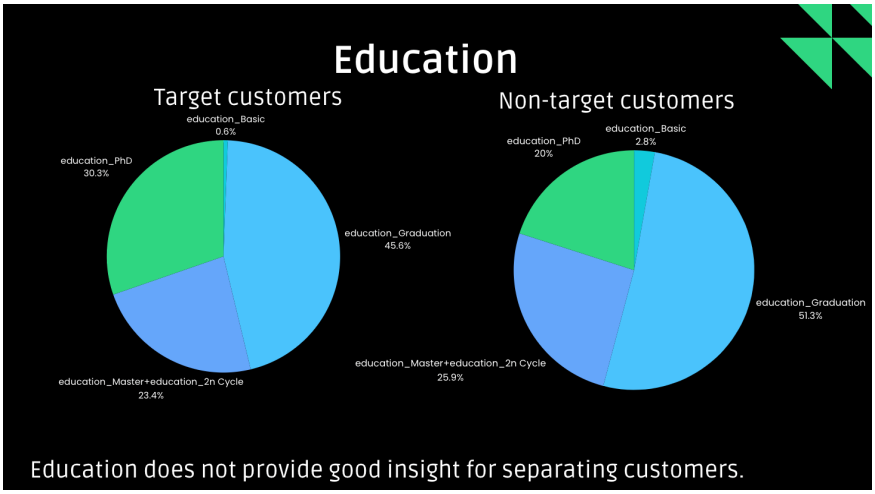


Implementation of target and non-target customers comparison and model construction
https://www.kaggle.com/code/raymondraman/ftec4002-final-project

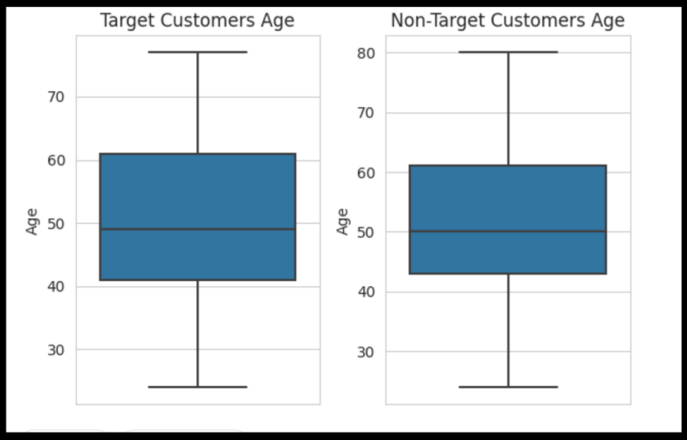
Other factors, such as channel preferences, product preferences, education levels and age, did not show significant differences between the two customer groups.







Age



Customer age is not an important factor in separating them.