**CP2 Milestone Report**

**What is the problem you want to solve?**

A company has tracked the results of six marketing campaigns for the same offer. The offer is a discount on a bundle of products including wine, meat, fish, sweets, and a small piece of gold jewelry. With response data from over two thousand customers who received the offer six times, the company would like me to predict customer acceptance rates to view which customers will accept the offer more often, less often, or not at all. A segmentation of its customers has been requested as well.

**Who is your client and why do they care about this problem?**

The client is an anonymous company that sells wine, meat, fish, sweets, and gold products. The client’s customers can purchase their products in-store, online, or from their catalog. The client wants to utilize their historical campaign data to now run a seventh campaign that is more targeted, reducing unnecessary customer contact costs while also limiting any associated decrease in revenue. The segmentation has been requested as the client would like to understand the key features of each customer group to successfully market the current offer and any future offers to their customers.

**Data Wrangling / Feature Engineering:**

Acquiring the marketing dataset was very simple as I simply downloaded it from Kaggle [[1]](#footnote-1). The dataset contained records for 2,240 customers and their individual responses to the six marketing campaigns. For each customer, unique features like their birth year, education, relationship status, income, number of kids, number of teens, enrollment date, purchasing channels, amount of money spent on certain products, and if they have filed any complaints are included.

Upon my first inspection of the dataset, it was apparent that income data was not available/provided for several customers. In order to impute the missing data, I decided to utilize the average income for their respective education and age group. Education was already provided as a categorical feature, but I needed to do some wrangling to engineer the age groups of the customers.

Using each customers birth year, I calculated their age and proceeded to them into “Late-Teens”, “Twenties”, “Thirties”, “Forties”, etc. age groups. Now that I had each customers education and age group, I imputed the missing values based on the average income for that customer’s education and age group.

The relationship status feature contained three categories (“YOLO”, “Alone”, “Absurd”) that I grouped into one “Single” category. For the number of kids and teens a customer has, I combined these into one “Children” feature containing the children count.

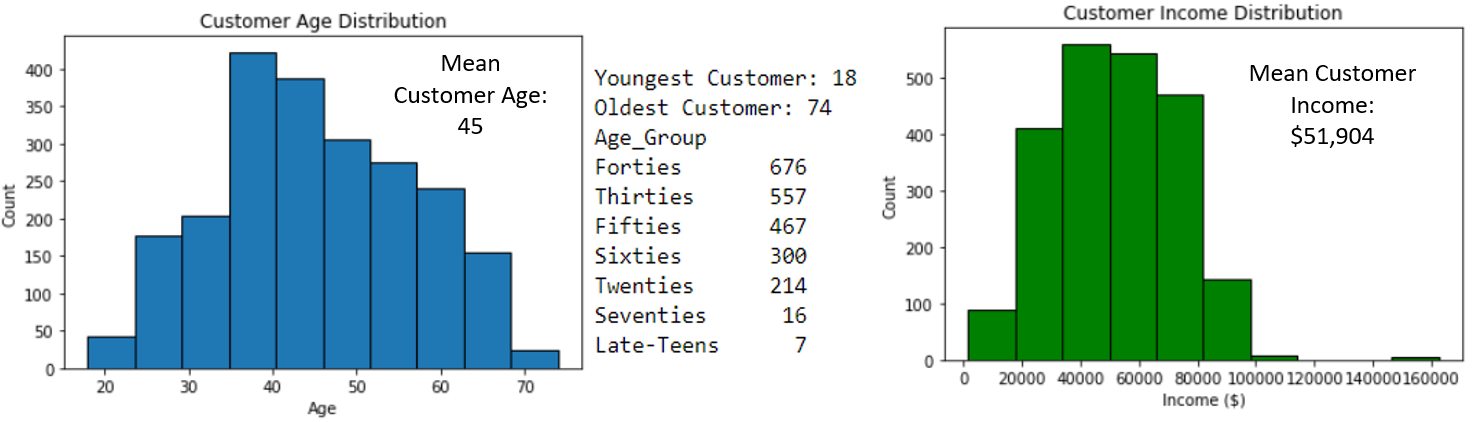
I also thought that engineering some features indicating customer spending habits would prove worthwhile. The first of these features I calculated was average purchase frequency (“Avg\_Purch\_Freq”). Subtracting each customer’s enrollment date from the assumed data collection date, I calculated “Days\_Enrolled”. The days enrolled were then divided it by the total number of purchases (“Number\_Purchases”).

The other customer spending feature I calculated was the average amount spent per purchase (“Avg\_Spend ($)”). I first summed the amount of money each customer spent on wine, meat, fish, sweets, and gold products and divided by the total number of purchases the customer made.

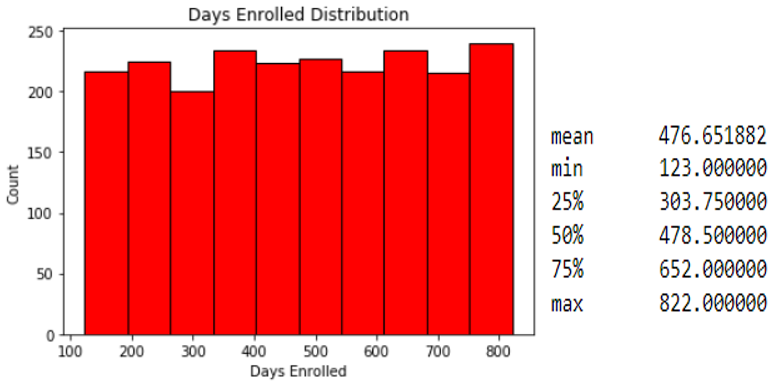
To provide some target variables to later predict, I engineered the “Accept\_One” feature (indicating if the customer accepted at least one of the six offer attempts) and the “Accept\_Rate” feature (indicates the number of offers accepted out of six. Ex: “(3/6)”) for building classifiers. The “Cust\_Accept (%)” was also engineered to use for later analysis.

**Data Storytelling / Inferential Statistics**

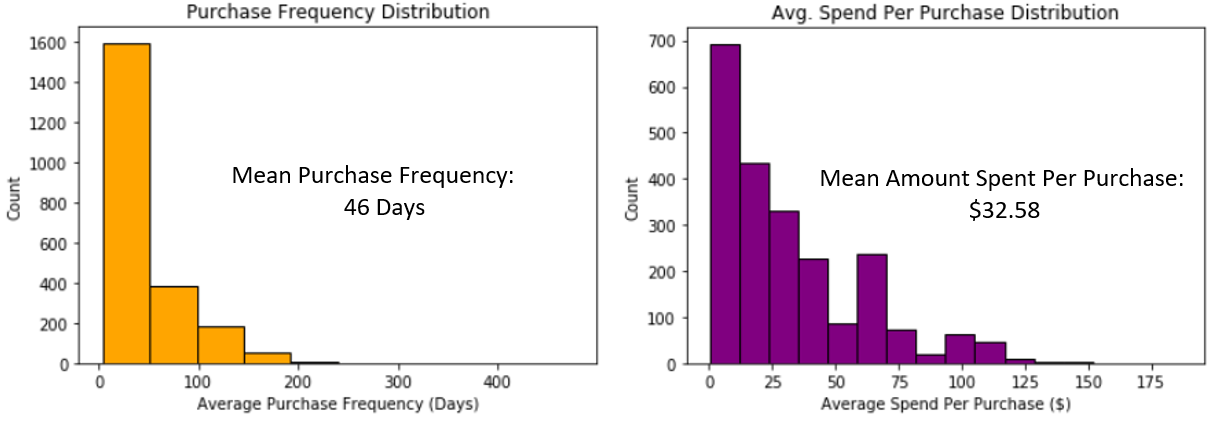
After wrangling the data into a format suitable for analysis, I began to try and get to know the customers a bit better through their feature distributions. First feature I explored was age. In doing so, I found that the dataset contained three customers born in 1893, 1899, and 1900. I believe its safe to assume the customers are no longer with us, so I dropped them from the dataset. I then proceeded to group the age groups by count and plot the distribution of the customer age. I also provided below the income distribution of the customers.



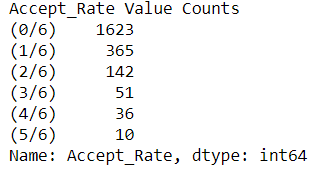
It would also be interesting to view the distribution for “Days\_Enrolled” as this would give us an idea of how long the customers included in the marketing campaigns have been enrolled and the client’s acquisition rate for these customers.



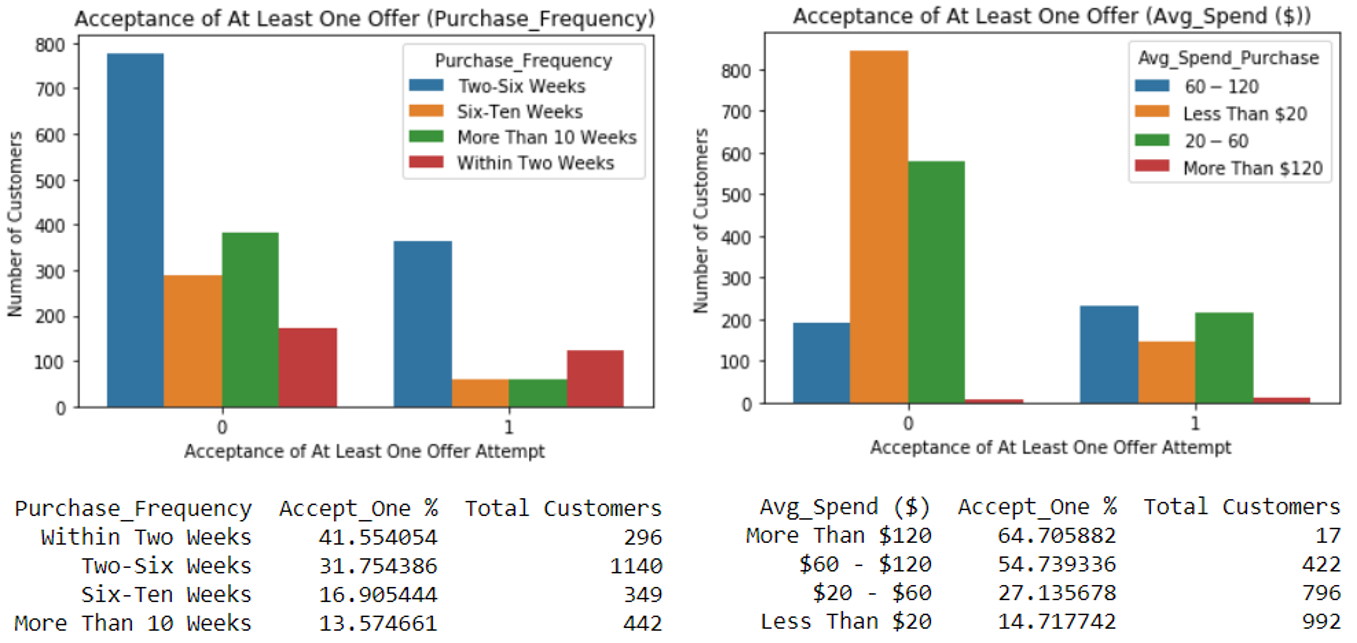
We see that the acquisition rate of these customers was relatively steady. Now let’s take a look at the spending habits of the customers through their purchase frequency and average spend per purchase distributions.



In analyzing the success of the six marketing campaigns, I have provided the following value counts of the customer acceptance rate.

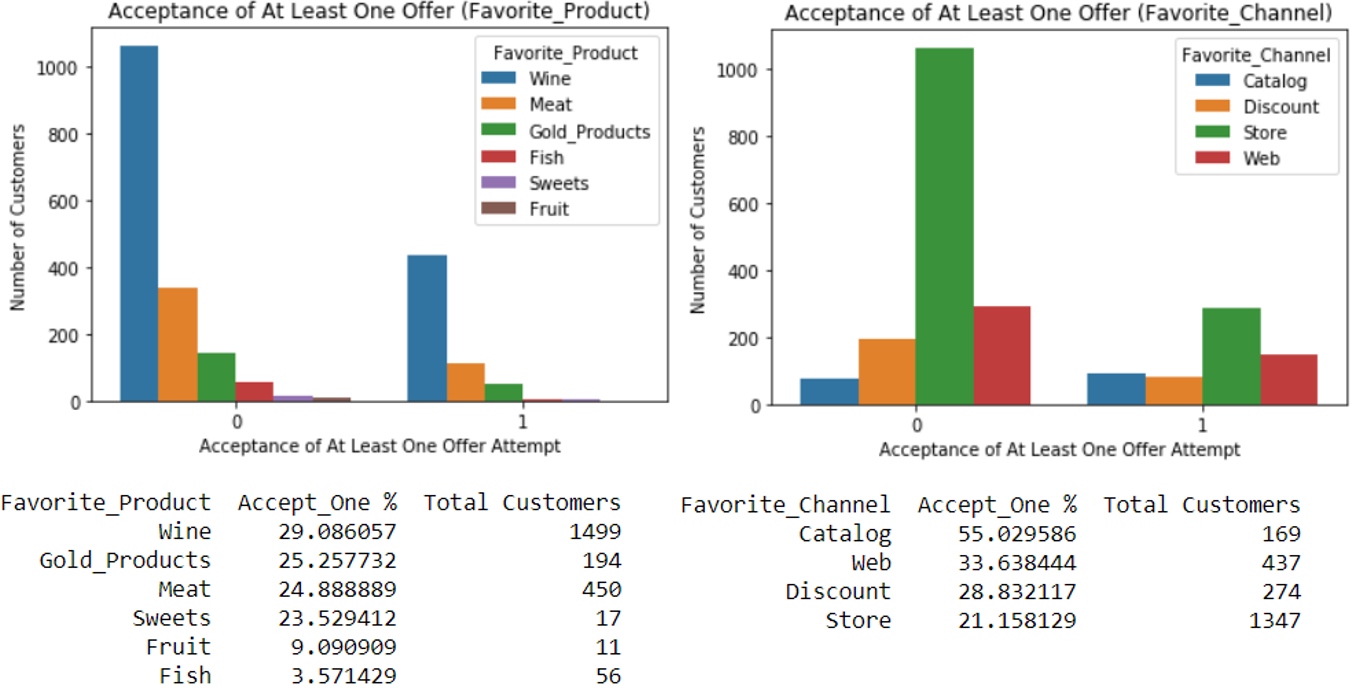


Roughly 73% of the customers did not accept the offer at least once during the six marketing campaigns. In total, the offer was accepted 996 times out of 13,362 attempts for an overall acceptance percentage of 7.45%. After previously viewing the distributions for “Purchase\_Frequency” and “Avg\_Spend ($)”, lets see if we can spot any interesting trends between these features and the acceptance of at least one of the six offer attempts. Below I’ll begin with “Purchase Frequency”.



Above we see that the offer was increasingly more popular with those who purchase products from the company more frequently. Let’s take a look at “Avg\_Spend ($)”. The offer was increasingly more popular among those who spend more per purchase. These customers favor buying in bulk, possibly a calculated effort to save on costs. When presenting the discount offer on a bundle, the customers who spend more per purchase seemed to favor the opportunity to save more.

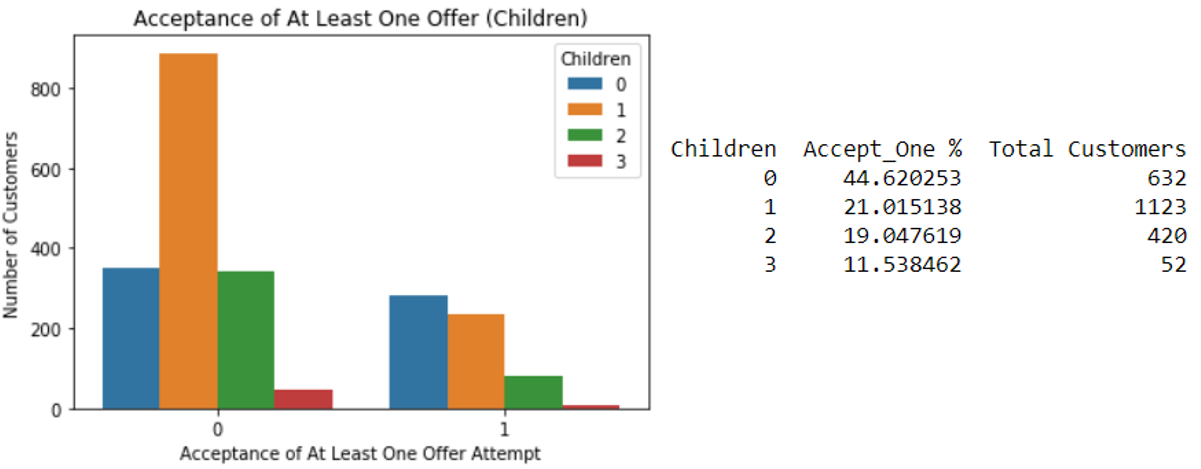
We have seen how spend and purchase frequency affects acceptance, but what products are these customers buying and through which of the client’s available channels?



Above we see that the offer was most popular with those who enjoy wine products, which is encouraging as roughly 50% of the revenue generated from these all of these customers’ purchases was from wine products. With meat products generating the second largest revenue for the client, the acceptance rate among those who favor meat was decent. What is really interesting is that the offer was very unpopular among those who like fish products. This may be something for the client’s marketing team to explore further.

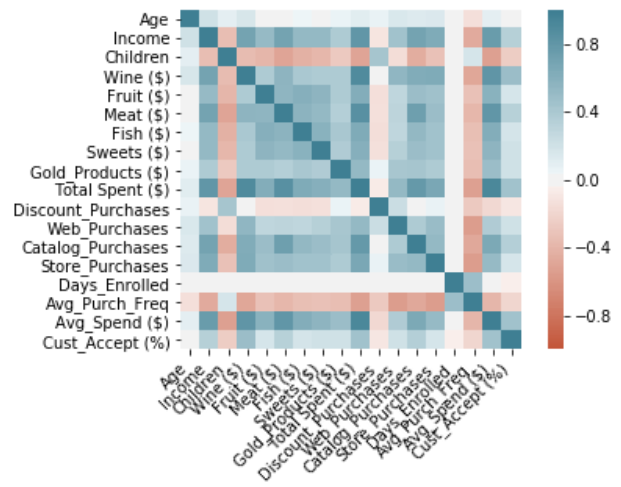
Roughly 39% of customer purchases were made in the store, representing the favorite channel among all the channels with web purchases (27%) and catalog purchases (18%) coming in second and third, respectively. Interestingly enough, it appears that the offer was quite successful among those who prefer purchasing products through the client’s catalog as 55% of customer who favor purchasing through the catalog accepted the offer.

In analyzing interesting trends among customer features, another one that stuck out was how the offer was more popular among those with no children.



Roughly 45% of customers who had no children accepted the offer at least once. The offer seemed to decrease in popularity with an increase in the number of children a customer has.

We can take a look at the correlation matrix for the numerical features included in this study. From the figure, we’ll see what features positively or negatively influenced acceptance percentage and the correlation between the features.



It appears that income, money spent on wine products ("Wine ($)"), total amount spent, number of catalog purchases, and the average amount of money spent per purchase were the greatest drivers of increasing a customer’s offer acceptance percentage. These features contained positive correlation with each other, prompting me to later drop “Total Spent ($)” to improve the performance of my classifiers.

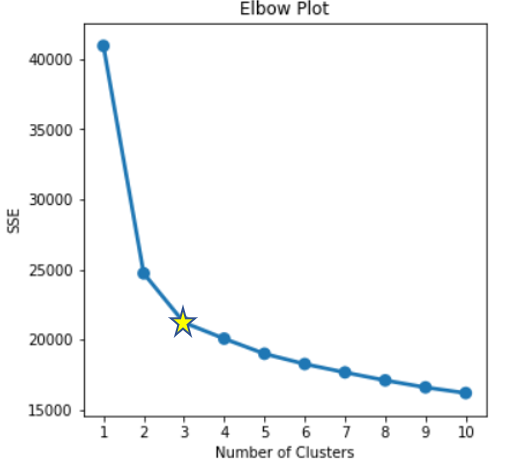
The number of children per customer mostly decreased correlation across the board. Only the number of discount purchases, age, and average purchase frequency contained positive correlation with the number of children, which all make some sense. Average purchase frequency also contained negative correlation with many of the features.

**Machine Learning:**

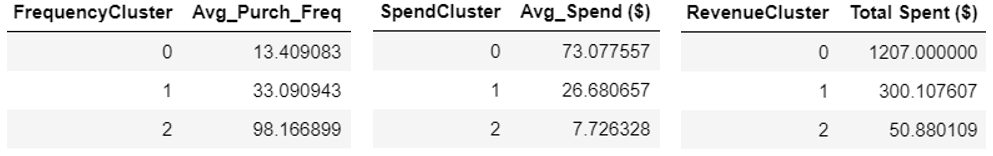
*(Unsupervised – Customer Segmentation)*

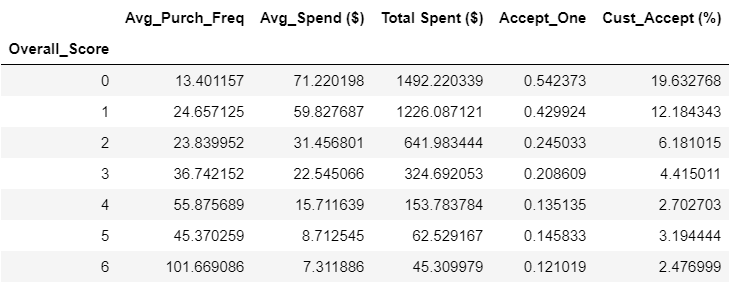
To begin clustering my client’s customers, I first needed to unskew the numerical features with a log(1+x) transform. This particular transform was utilized due to several of the customer features being expressed as count data containing zeros. I then proceeded to use sklearn’s StandardScaler() to scale the numerical data. For the categorical data, I utilized panda’s get\_dummies() method to numerically encode the values.

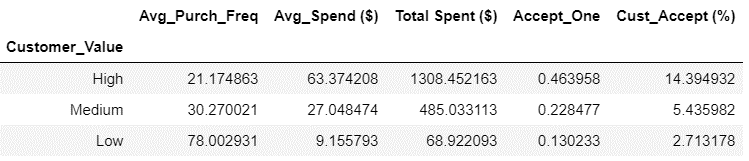
With the data preprocessed, I proceeded to plot an elbow plot to find the ideal number of clusters for kmeans clustering.

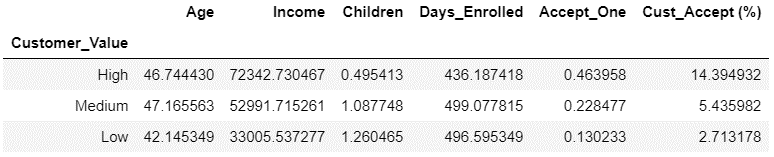


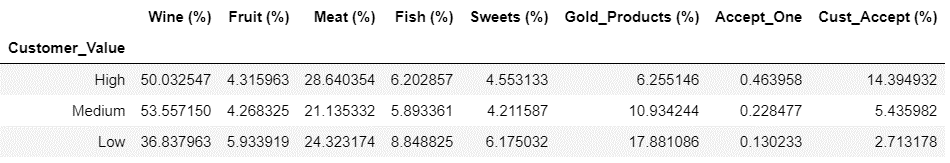
After plotting the elbow plot, three was found to be the optimal number of clusters. This number was convenient in later providing the client with three customer groups describing their value (“Low”, “Medium”, “High”). These customer values were calculated using each customer’s average purchase frequency (“FrequencyCluster”), average spend per purchase (“SpendCluster”), and the total amount of money they have spent purchasing the client’s products (“RevenueCluster”).











1. <https://www.kaggle.com/rodsaldanha/arketing-campaign> [↑](#footnote-ref-1)