BA820: Final Project Paper

Team A4: Aleks Lazowski, Yuhan Wang, Eunjin Jeong, Geech Hor Hout, Niming Wang

1. Business Problem:

We will be conducting a Customer Personality Analysis on the company's customers and their products. The business problem is how we can help companies better understand their customers. Looking at the customer's specific needs, behaviors and concerns will allow companies to adjust and modify products according to these demographics. Companies can use this analysis to avoid spending money on products in the wrong target segments and be efficient in their product analysis. The goal of this project is to conduct supervised and unsupervised machine learning in order to create segmentation for existing and make potential predictions for future customers. This will allow the analysis to continuously be used for the present and future.

2. Dataset:

Dataset: Kaggle(link) - **Customer Personality Analysis**

The dataset is for understanding customer personalities and helping businesses to target customers. The dataset contains 29 attributes and 2240 rows. The attributes can be classified into 4 categories; People, Products, Promotions, Place.

- **People**: It consists of customer's personal information including ID, birth year, education level, marital status, income, household composition, enrollment and complaint status.
- **Products**: It consists of customer's purchase history data on products. It's about the amount spent on wine, fruit, meat, fish, sweets, and gold in the last 2 years.
- **Promotions**: There are attributes indicating the number of purchases made with discount, and the number of offers accepted in the campaign.
- **Place**: It has attributes indicating where customers made purchases. It contains the number of purchases on the website, catalog, stores, and the number of visits to the company's website in the last month.

- **Data Preprocessing and Feature Engineering:**

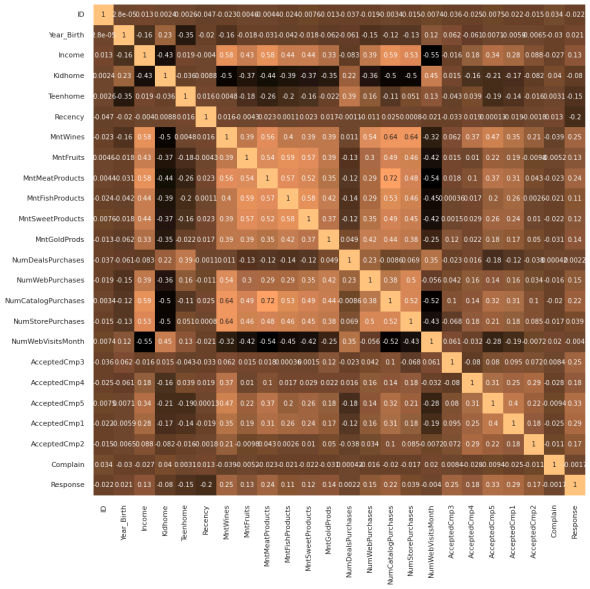
We started our data cleaning by separating the imported CSV dataset into different columns to allow data analysis. In the Income column, we had 24 missing values so we replaced the missing values with the median. There are two columns that consist of the same values (Z_CostCount and Z_Revenue). Therefore, we deleted both of them because they would have 0 variance and will not have any significant impact on our analysis in the future. We also removed extreme outliers for income that are greater than 600,000 in the income column and 3 outliers for

customers that are above 120 years old for a more balanced distribution. We combined 8 different variables that include ‘YOLO’, ‘absurd’, and ‘alone’ in the Marital_Status column to merge them into one term ‘single’.

For prediction and clustering purposes, we ended up changing marital status to having a partner or not. This makes it easy for companies to generalize the customer segmentation. For education categories, we merged all degrees into two categories, undergraduate and postgraduate degrees. We created 3 new variables called ‘Age’ which were calculated by subtracting this year to the customer's year of birth, ‘spent’ which added all products bought, and ‘children’ which added teenhome and kidhome together.

3. Exploratory Data Analysis:

- **Heatmap**

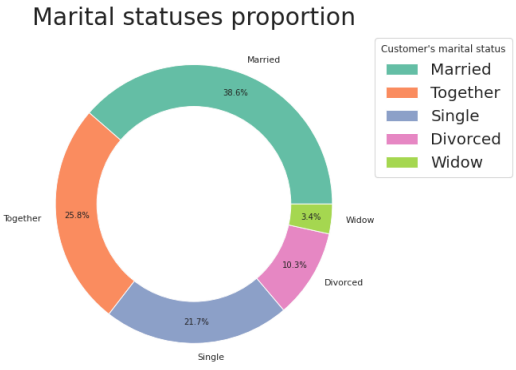


Based on the heatmap, there is a positive correlation between Income and features related to Product and Purchases. There is a negative relationship between Kidhome and features related to Product and Purchases.

In the middle of the heatmap, we can see the correlation between products and the number of purchases.

- **Education level**

We created a pie chart for different Education levels.
50% of our customers have a bachelor’s degree while the rest have a master’s and Ph.D. Therefore, we can say that most of our customers have at least a college-level education.



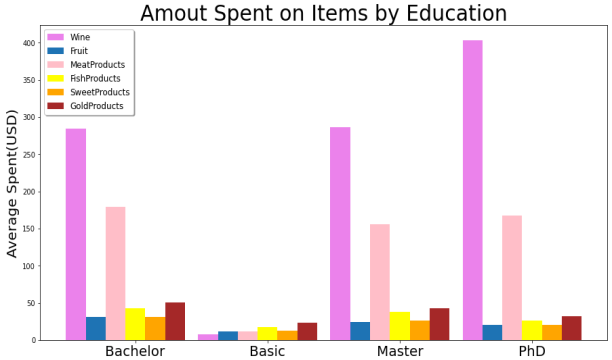
- **Age distribution of the customers:**

The average age of our customers is 52 and the range of the age distribution is between 25 to 81 years old. Majority of our customers are in their 40s to 50s, accounting for 60% of the total.

- **Marital Status of the customers**

According to the pie chart, 39% of the customers are married and 26% have partners. Therefore, we can say that 75% of our customers have partners or family.

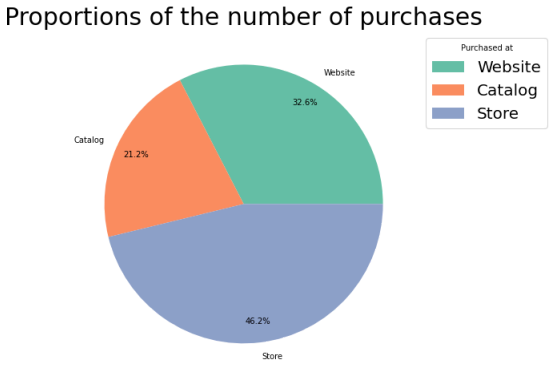
- **Averaging spending of the customers based on their Education level**



We explored the total amount spent on each category of product by the customers. Consumers spent most on wine, followed by meat, gold, and fish. After that, we created a bar chart showing the average amount spent on different products categories by education level. The graph showed that the customers who have Ph.D. tend to spend most on wine products while customers who are within basic education level do not spend much in general compared to those with higher degrees. The second most popular product is meat and it is bought by all the customers.

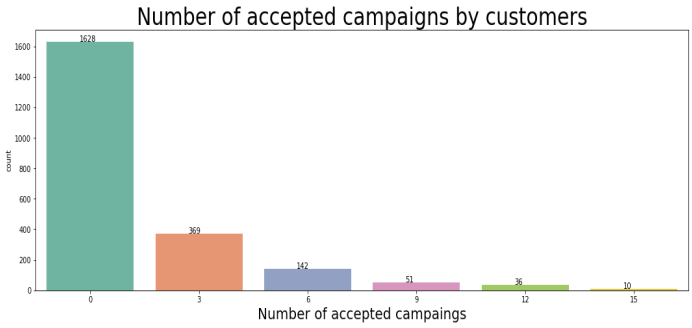
- **Number of Purchases**

We also explored the customer preferred method for their purchases using 'NumWebPurchase', 'NumCatalogPurchases', 'NumStorePurchases' columns. We split their shopping type into three categories which are in-store, catalog and website. Based on the graph, we found that 46% of our customers in the dataset prefer in-store shopping. Another point that we can assume is that our customers also like online shopping as much too.



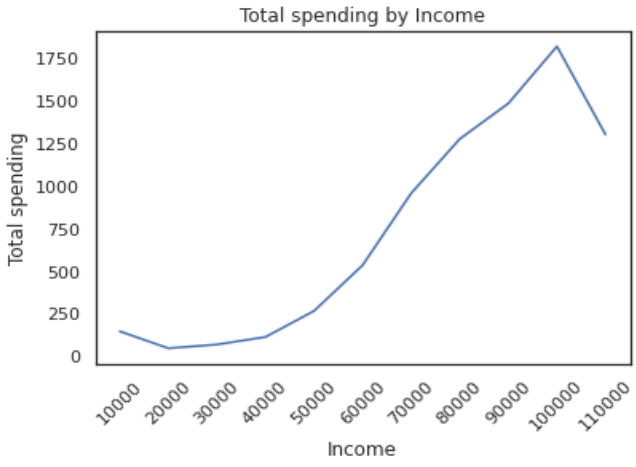
- **Accepted campaign**

Accepted campaign has two values (0 and 1). It's 1 if a customer accepted the offer in the 1st, 2nd, 3rd, 4th and 5th campaign, 0 otherwise. Based on the plot, many customers do not like the campaign and the first campaign is the only successful one. It shows that the campaign is not effective and there is a possibility that the customer does not like the campaign or think that it is sketchy.



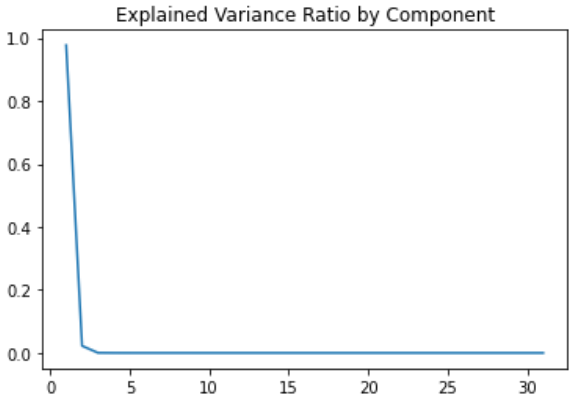
- **Income and spending**

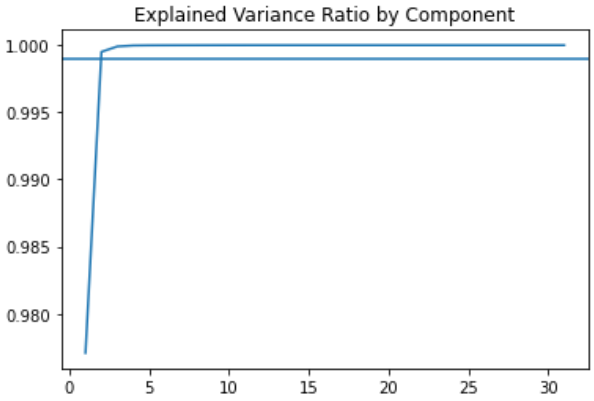
The average income of the customers is $52,000. Most of the customers' income is between $35,000 and $68,000. There are also a few customers who have high income. Based on the correlation analysis of features, income was mostly correlated with the amount spent on products and the number of purchases. To dive deeper into the income column, we created a line plot of income and average total spending, by combining spendings on all products. On average, consumers with higher income tend to spend more. We concluded that income is one of the most important columns.



4. Dimensionality Reduction(PCA)

Since there were 29 columns and correlations between columns, we decided to start our analysis by conducting principal component analysis (PCA). PCA allowed us to reduce the dimensionality of our dataset, increasing interpretability and minimizing information loss, by creating new uncorrelated variables. When we performed PCA, we had to choose how many components it should have and what the explained variance ratio should be. So, we plotted the cumulative explained variance ratio and explained variance by component.



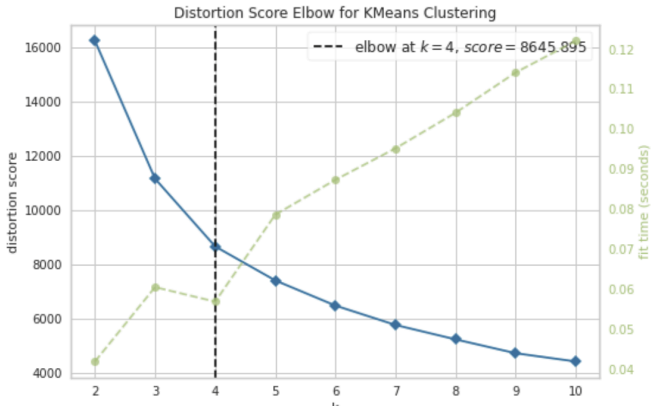


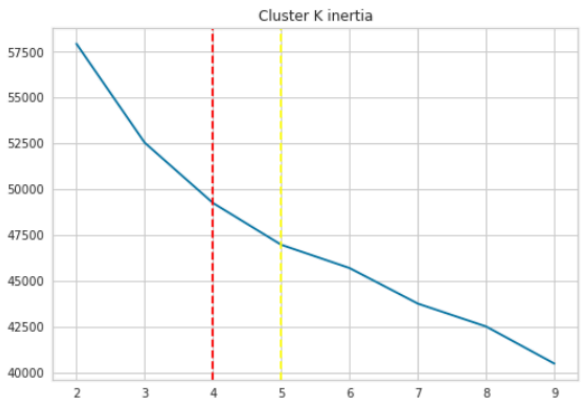
As those two graphs shown above for Variance Ratio by component, the number of components by the elbow method is around 3 because explained variance ratio dropped significantly at the elbow curve. The explained variance ratio by 3 components is 99.9%, which means that we should choose 3 as the number of components and add the explained variance ratio of each component to reach the number 99.9% (bigger than 80%) to avoid overfitting.

5. Clustering Analysis:

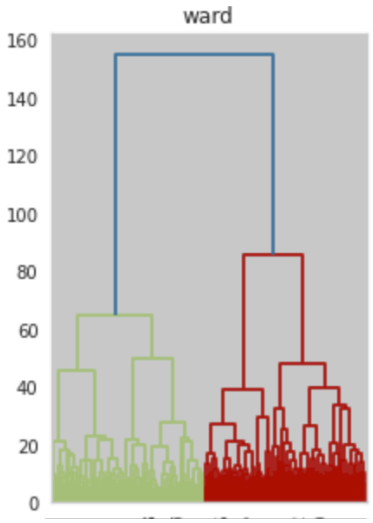
- **KMeans**

We chose to implement 3 different clustering models and test their accuracy scores to determine the model that should be used to form clusters. The first method chosen was KMeans. KMeans





randomly selects the first centroid from the dataset and computes the distance of all points. It then picks a point as the new centroid that has maximum probability proportional to this distance. These steps are then repeated until the number of k centroids have been sampled. In our model, we used the distortion score from the elbow method and the K inertia graph to find the optimal amount of clusters. In this case, 4 clusters were chosen. Furthermore, the silhouette score, which is used to calculate the goodness of a clustering technique, ended up being close to 37% which was the least most accurate model.

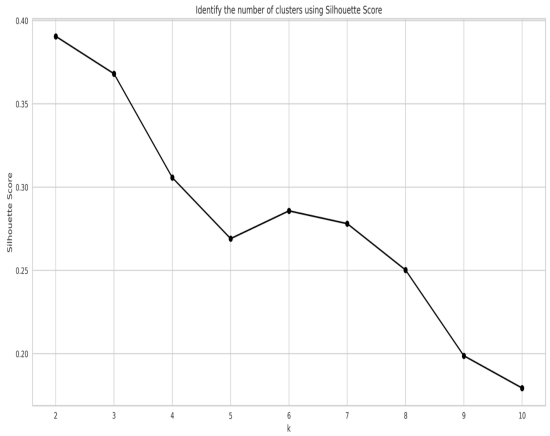


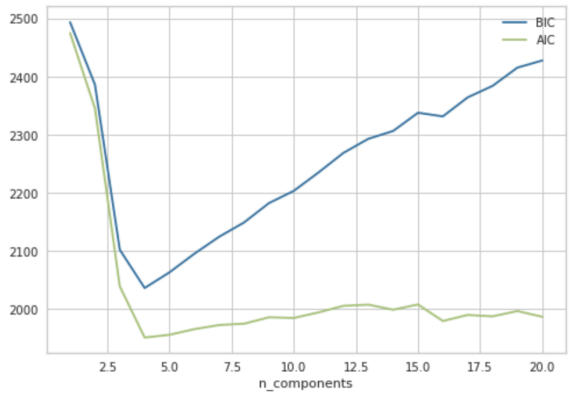
- **Hierarchical Clustering - Agglomerative**

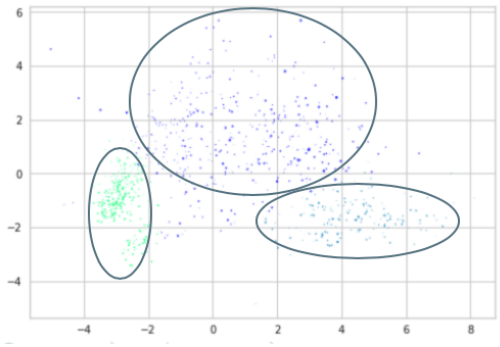
The second model used was hierarchical clustering. Agglomerative clustering is a general family of clustering algorithms that build nested clusters by merging data points successively. We used all 4 linkage methods for comparison. Ward was the clearest dendrogram and the most effective. In that dendrogram, we see that the optimal amount of clusters is around 3 which seems reasonable as it follows this distance method. The silhouette score for hierarchical clustering ended up at 42% which was the second most accurate model of the 3.

- **Gaussian Mixture Method**

Lastly, we tested Gaussian mixture method. GMM assumes that there are a certain number of Gaussian distributions, and each distribution represents a cluster. This model tends to group data points belonging to the same distribution. To identify the optimal number of components, we used silhouette score, Bayesian Information Criteria and Akaike's Information Criteria to give us optimal number of components. We compared the average of all 3 statistical anlysis' and found the optimal number to be 3 components. Eventually, we tested and predicted the model and got an accuracy score of 44% which was the most accurate model.



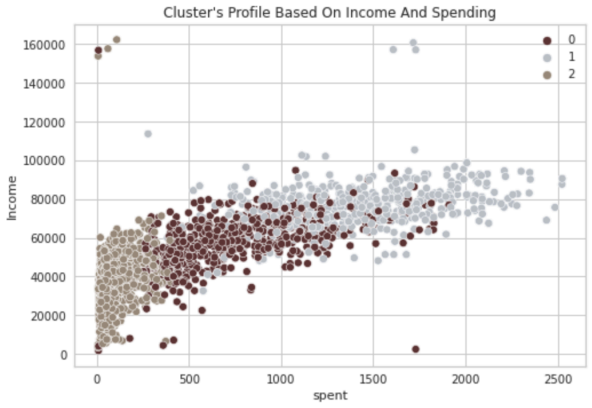


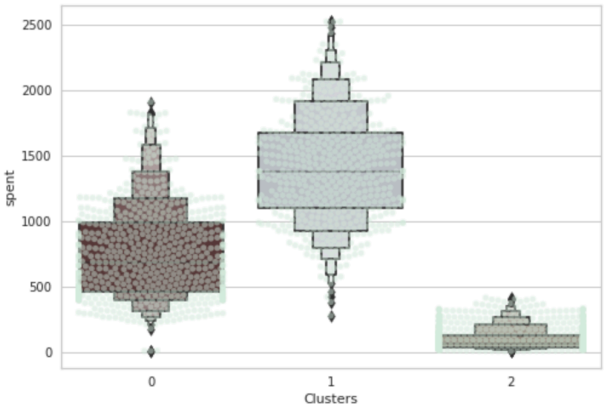


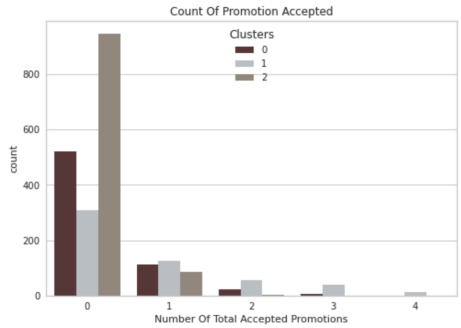
GMM was the most accurate model because we think it was able to best cluster the data points better the other two models. The picture on the left shows the 3 clusters created from our PCA reduction. The advantage of Gaussian mixture model is that it can handle very oblong clusters. The shape of our clusters is are indeed less circular and more oval shaped. This also indicates why KMeans was our worst model because it uses spherical clusters based on the centroid created.

- **Profiling the Clusters**

After looking at the distribution of the clusters, we wanted to start to profile these customers in each cluster to gain an understanding of who the company is trying to target. We created a customer profile based on income and spending in the first graph. It is noticeable that cluster 0 and 1 have the highest income and spend the most money on products. While cluster 2 is considered low spending and has low income. The graph on the right further illustrates the amount spent on products like wine, meat, gold and fish. Cluster 1 spent the most and cluster 0 wasn't too far behind.

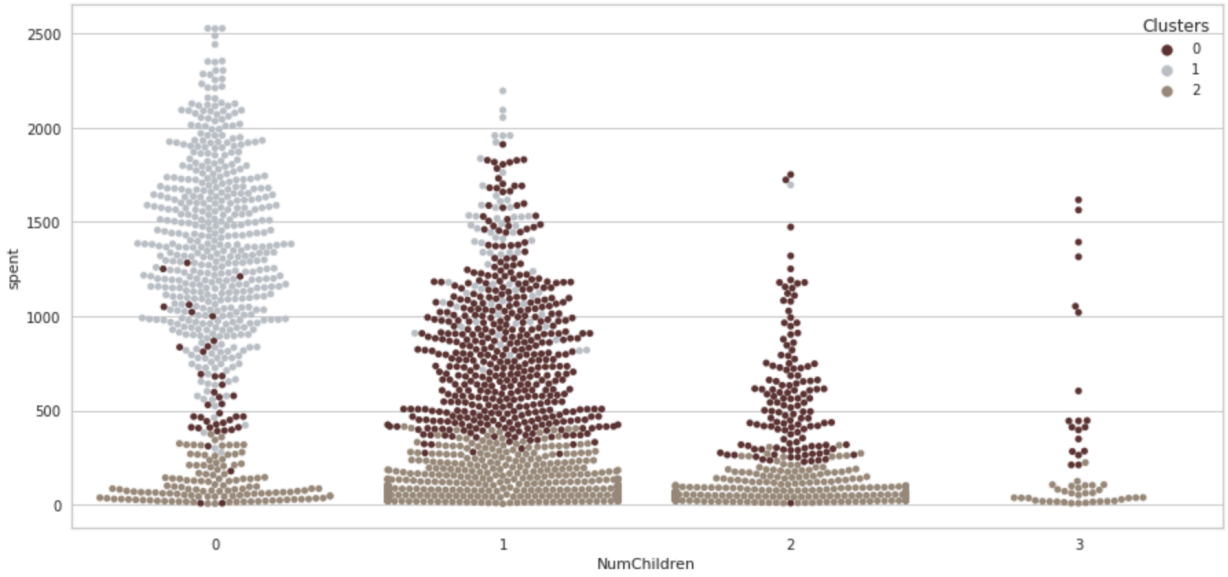






Our group also examined how the clusters responded to all 5 different campaigns created by companies. It is clear that there is limited participation in most of the campaigns. From here, it can be thought that the campaigns are not well planned and do not go to the right target market. The company would have to re-work their targeting plan based on customer behavior. To solve this problem, a better understanding of their customer base would help.

To take a preliminary look at the clusters, we looked at 7 different variables: education, marriage status, age, number of children, web purchases, store purchases and catalog purchases. One example in the figure below shows which customers have children. Cluster 1, clearly spends the most which was explained earlier. But in the graph below, we see that most of these customers do not have any children. Cluster 2 on the other hand spends the least, and most of the time has children. This is surprising because we would expect parents with more children to spend more money due to a bigger household. The age range mainly consisted of customers above the age of 18. One cluster that jumped out was cluster 0 where it consisted of mainly people over the age of 40. This is something the company can target when sending our promotions and different campaigns. Another trend found between all 3 clusters was that most were postgraduate customers. This makes sense because these are the customers that are more likely to have made more money is due to extra schooling. Those who have basic education were mainly in group 2 which follows the trend of limited spending.



- **Full Cluster Profile:**

<u>Cluster 0:</u>

- Biggest group with 1 child
- Biggest age group are parents
- Highest group that appeal to web purchases
- Highest group at doctoral level of education.

<u>Cluster 1:</u>

- Biggest spending group, and most attractive to companies
- Almost all customers have no children
- Equal age distribution

- Majority of customers have graduated bachelors and are in masters or other postgraduate programs
- Highest group that appeal to in store purchases

Cluster 2:

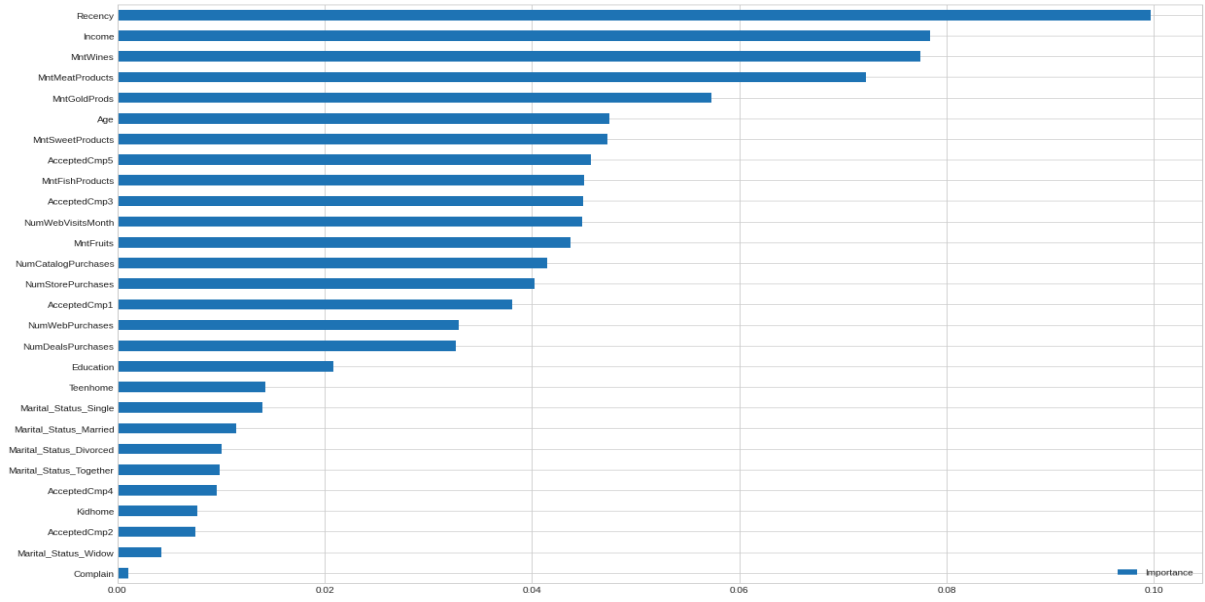
- Biggest group of parents with either 1 or 2 children
- Age is mainly between 30-55
- Least amount of money spent from catalog purchases
- High number of customers that are single. Could indicate again why limited money is spent.

6. Prediction Analysis

`Before we started our prediction analysis, we did some further data cleaning. We dropped some useless columns and changed all the columns into numeric types. We dropped the 'ID', 'DT_Customer', 'Year_Birth' columns, which won't have any influence on our prediction.

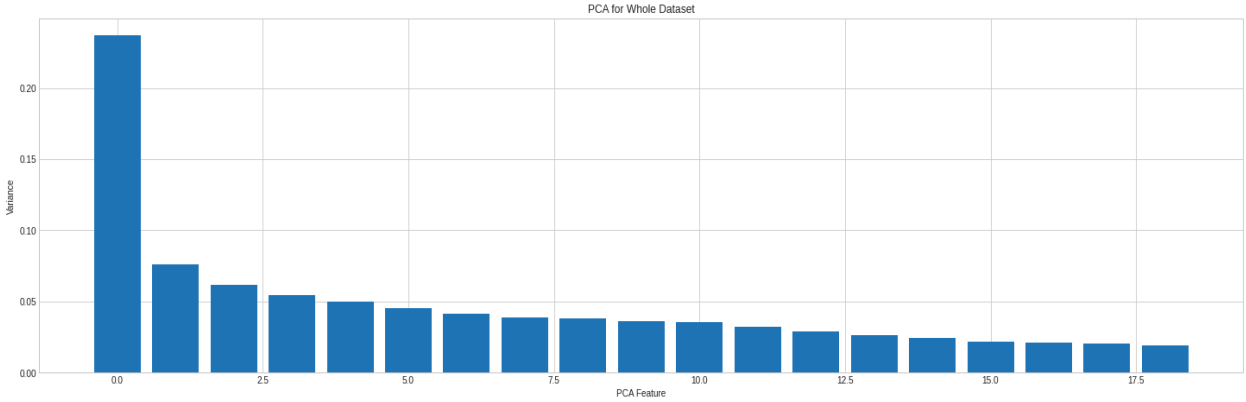
- **Feature Selection**

When our group was doing feature selection, we chose to use Random Forest to help us select valid features. We obtained the following graph by calculating the feature importance of each column. We chose the first 18 columns as our predictive features. And calculated the sum of their importance as 0.9102521421188535.



- **PCA Transformation**

We calculated PCA for both datasets and graphed the Variance for each feature as below. And the final shape of PCA data we got is (2236, 19).



- **Method Selection**

In this experiment, we chose Logistic Regression and Boosting Tree to do the comparison. Furthermore, we compared the PCA dataset with the original dataset which had a better prediction effect.

Boosting Tree performed the best among all the models in all datasets. LR was the worst model because it was too simple for this classification task. Boosting Tree had some outliers in raw datasets and feature selection datasets, which indicated this algorithm might not be stable in these datasets.

In conclusion, in this classification task, we could use Feature Selection Dataset + Boosting Tree.

- **Prediction Results**

We did the prediction model with the Boosting Tree and the Feature Selection Dataset. The overall test accuracy of the model is 0.873. But diving deep into the score report, the model performs quite well in recognizing negative samples(0), but poorly in recognizing positive samples.

The test MCC is 0.399, which indicates that the model may not be good at finding positive samples in the test set. While we find the Train MCC is 0.989, this result shows there is an overfitting problem in the model. We tried to simplify the model and decrease our train MCC, but the test MCC showed little improvement. This result might indicate the predictors we use in this dataset might not predict 'Response' very well.

7. Business Problem - Conclusions & Recommendations

- **Conclusion:**

Based on our research findings, most of the market campaigns were not very effective and had low acceptance rates from their customer base. In this case, companies need to adjust their market segmentation and understand their target customer better. The unsupervised methods offer insights on what the company should focus on to design a more effective promotional campaign for the right target group with a goal to optimize profit. Based on the clustering methods, we found the optimal number of customer groups to be 3. Each cluster has different spending patterns based on beahviors and characteristics. We determined profiles for each cluster based on age, income, spending, and multiple other factors. Each profile categorizes customers and allows the company to visualize what specific customer base they are targeting. To build a more successful promotional campaign, companies should use these profiles to help sell their products and have a higher success rate of promotions accepted.

For Supervised Machine learning, after applying the model on the feature selection dataset, PCA dataset and the raw dataset with both logistic regression and boosting tree, the Boosting Tree model performed the best. LR is the worst model because it is too simple for this classification task. With Boosting Tree, there were a few outliers in the raw datasets. So in this classification task, we could use Feature Selection Dataset + Boosting Tree. The test MCC is 0.399, which indicates that the model may not be a good fit for prediction. We found the Train MCC to be 0.98. This result showed that there might be an overfitting problem in the model. We also tried to simplify the model to decrease train MCC. The test MCC still can not show much improvement. This result might indicate that the predictors used might not be a good fit to predict 'Response' very well.

- **Recommendations:**

After having to find insights on the target market, the company can better design marketing to meet the demand and match the characteristics of the customer clusters. Below are the suggestions to what the company should do in order to effectively reach their customers' demand. These recommendations will improve the acceptance rate of campaign from customers.

- **In-store campaigns for the latest new products**

  In-store shopping is the most preferred option for customers. To increase the profit of promotions, we suggest designing campaigns for the latest and high price items such as wine, gold, and meat. We also suggest targeting customers with higher purchasing power with higher income and spending because they are most likely to accept the offer and be willing to buy new products in stores.

- **A website campaign for specific deals**

  We recommend using a website campaign that utilizes specific deals. Website is notably preferred by customers with high spending and average income. These customers also tend to have a child and are highly educated. Therefore, we suggest especially targeting these customers for the website campaign. To increase campaign acceptance, we suggest specific deals such as free shipping and discount for daily necessity products. By attracting more customers, we can also expect to turn these customers into loyal customers.

- **Eliminate campaigns based on catalog ads**

  We recommend eliminating campaigns based on catalog ads to save costs for unnecessary campaigns. Customers are less likely to purchase products from catalogs according to our analysis.