1. **Objective & Description of Dataset**

Customer segmentation is a powerful technique for dividing a customer base into groups of individuals who are similar in specific ways relevant to marketing, such as age, gender, interests, and spending habits. By doing so, a business can tailor its marketing efforts to each group more effectively, resulting in more efficient and targeted marketing strategies.

Our project objective is to predict how much money a customer is likely to spend on products. This would be done by analyzing the customer's past purchasing behavior, as well as demographic information, such as age, income, and education level. By using machine learning algorithms such as clustering, linear regression, and classification, we would like to estimate how much money a customer will likely spend in the future based on these factors.

The data set consists of 2240 data points and 29 attributes. It can be categorized into 4 subsets, the first one is customer’s information which contains ID, Year_Birth , Education, Marital_Status , Income, Kidhome, Teenhome , Dt_Customer, Recency, Complain. The second one is Products which contains Mntwines , MntFruits , MntMeatProducts , MntFishProducts , MntFishProducts , MntGoldProds.The third one is Promotion which includes NumDealsPurchases, AcceptedCmp1, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5, and Response. The fourth one is Place which includes NumWebPurchases , NumCatalogPurchases , NumStorePurchases , NumWebVisitsMonth.

```
Informations Of The Dataset :

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2240 entries, 0 to 2239
Data columns (total 29 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   ID                   2240 non-null   int64
 1   Year_Birth           2240 non-null   int64
 2   Education            2240 non-null   object
 3   Marital_Status       2240 non-null   object
 4   Income               2216 non-null   float64
 5   Kidhome              2240 non-null   int64
 6   Teenhome             2240 non-null   int64
 7   Dt_Customer          2240 non-null   object
 8   Recency              2240 non-null   int64
 9   MntWines             2240 non-null   int64
 10  MntFruits            2240 non-null   int64
 11  MntMeatProducts      2240 non-null   int64
 12  MntFishProducts      2240 non-null   int64
 13  MntSweetProducts     2240 non-null   int64
 14  MntGoldProds         2240 non-null   int64
 15  NumDealsPurchases    2240 non-null   int64
 16  NumWebPurchases      2240 non-null   int64
 17  NumCatalogPurchases  2240 non-null   int64
 18  NumStorePurchases    2240 non-null   int64
 19  NumWebVisitsMonth    2240 non-null   int64
 20  AcceptedCmp3         2240 non-null   int64
 21  AcceptedCmp4         2240 non-null   int64
 22  AcceptedCmp5         2240 non-null   int64
 23  AcceptedCmp1         2240 non-null   int64
 24  AcceptedCmp2         2240 non-null   int64
 25  Complain             2240 non-null   int64
 26  Z_CostContact        2240 non-null   int64
 27  Z_Revenue            2240 non-null   int64
 28  Response             2240 non-null   int64
dtypes: float64(1), int64(25), object(3)
memory usage: 507.6+ KB
None
```

Fig1: Informations of the dataset

2. **Key statistics**

For the distribution of customers' ages, we create a histogram to visualize the frequency of customers in each age range. And we add a line plot and find out the data is normally distributed. From the graph, we can see most of the customers are in the age range of 36 to 50.

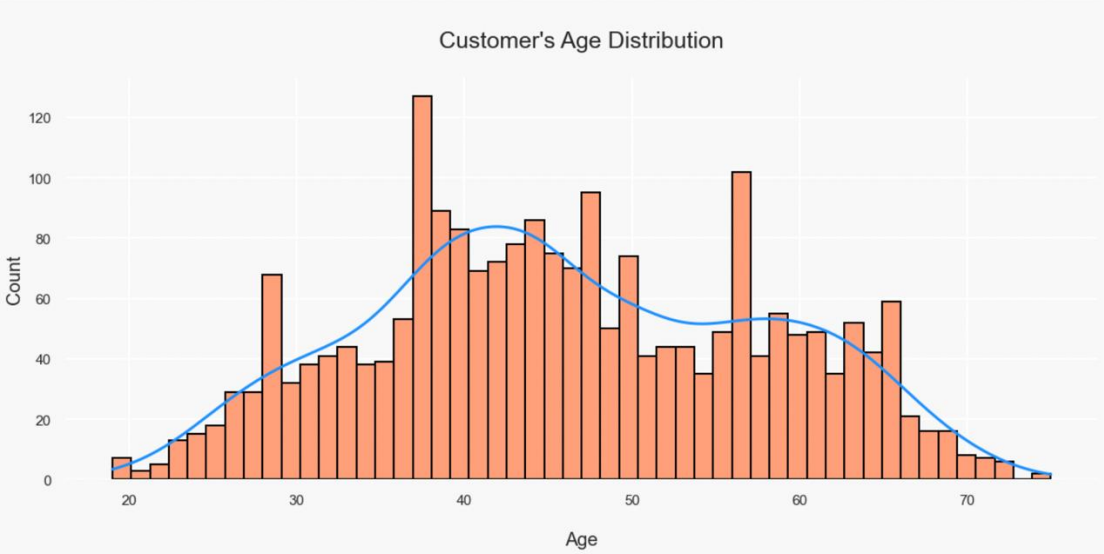

Fig2: The distribution of customer's age

For the distribution of customers' income and spending, we create a histogram to show the frequency of customers in each income and spending range. And we find out that most of the customer's income is between 20k to 80k. However, almost all customers spend only a small portion of their income.

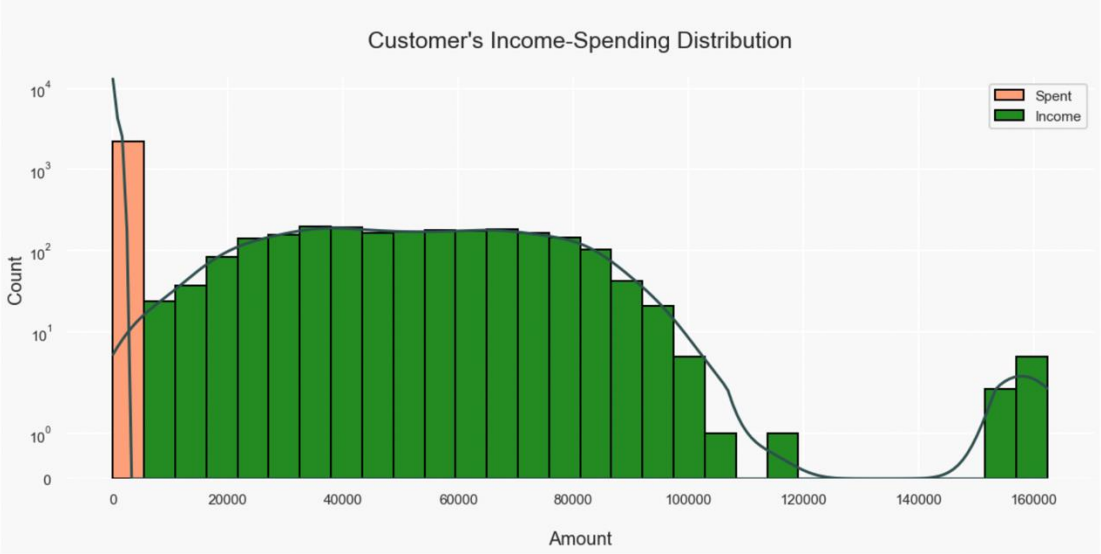

Fig3: The distribution of customers' income and spending

For the customer's education level, we create a pie chart to visualize it. There are three education levels: 1) Graduate 2) Postgraduate 3) Undergraduate. There are 50.41% of customers completed graduate, 38.2% of customers are on the postgraduate level, and 11.39% of customers are at the undergraduate level.

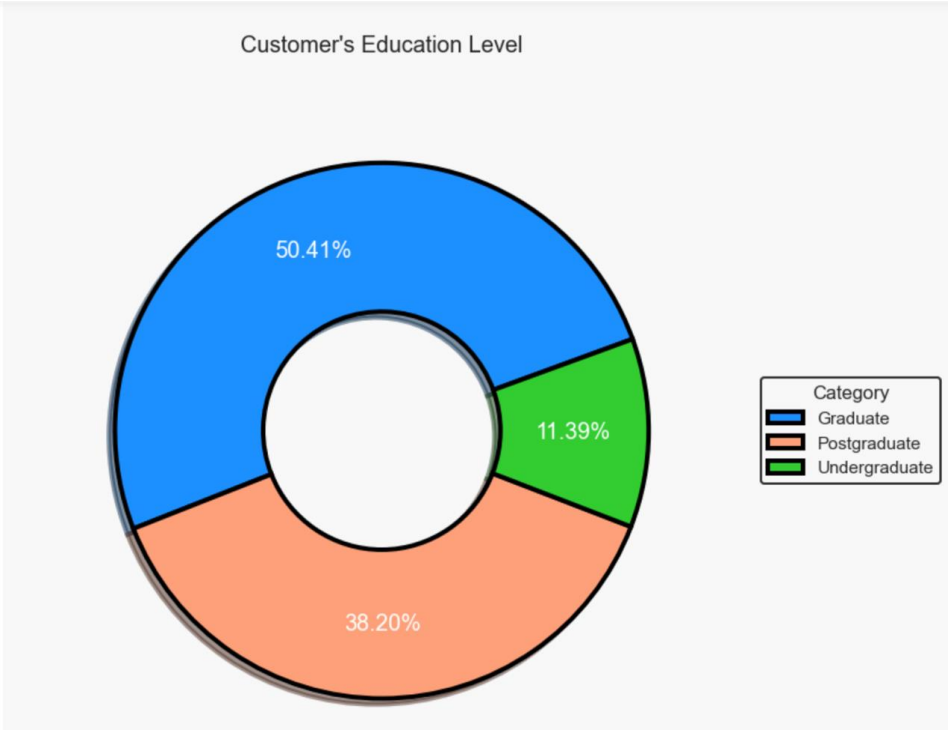

Fig4: Customers' Education Level

To visualize the distribution of customers' income and spending on basis of education level, we create two separate histograms to show the frequency of customers in each income and spending range. We find out most of the graduate-level customers' income range is between 20k to 85k and their spending is between 0 to 2k. Most of the postgraduate-level customers' income range is between 30k to 80k and their spending is between 0 to 2k. And most of the undergraduate-level customers' income range is between 10k to 80k, their spending is between 0 to 1.4k.

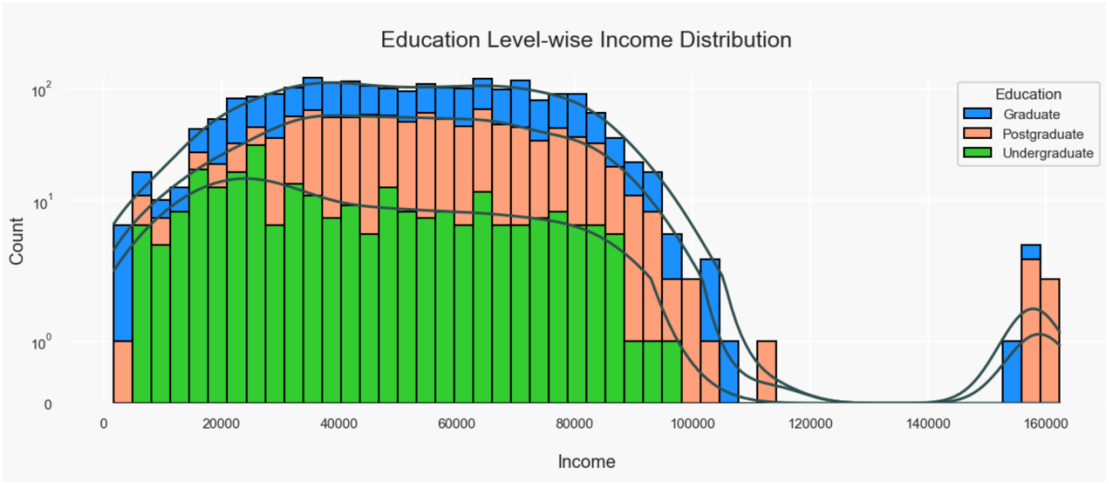

Fig5: Education Level-wise Income Distribution

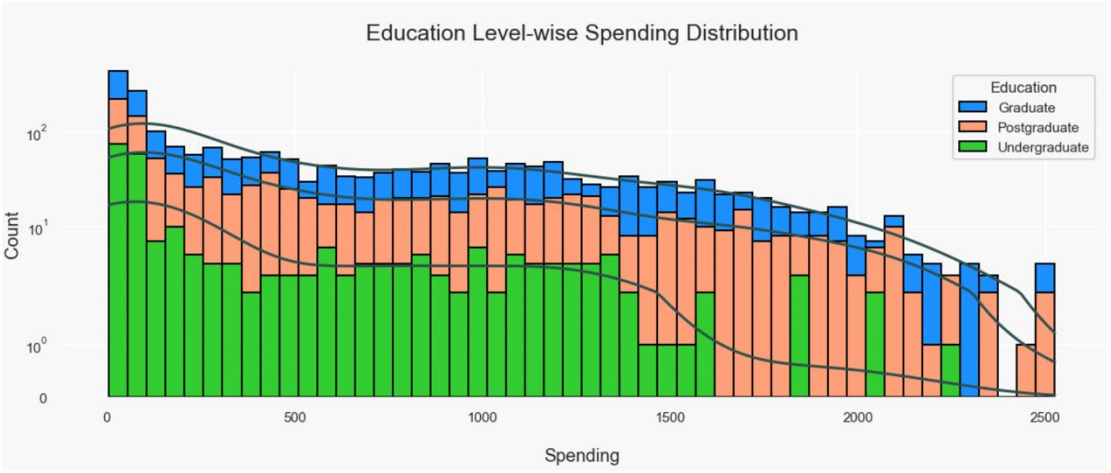

Fig6: Education Level-wise Spending Distribution

For the distribution of customers' children, we create a bar chart to show the number of children that the customers have in each category. There are 632 customers who have no child. There are 1114 customers who have one child which is the most. And there are 416 customers who have two children, and 50 customers who have three children.

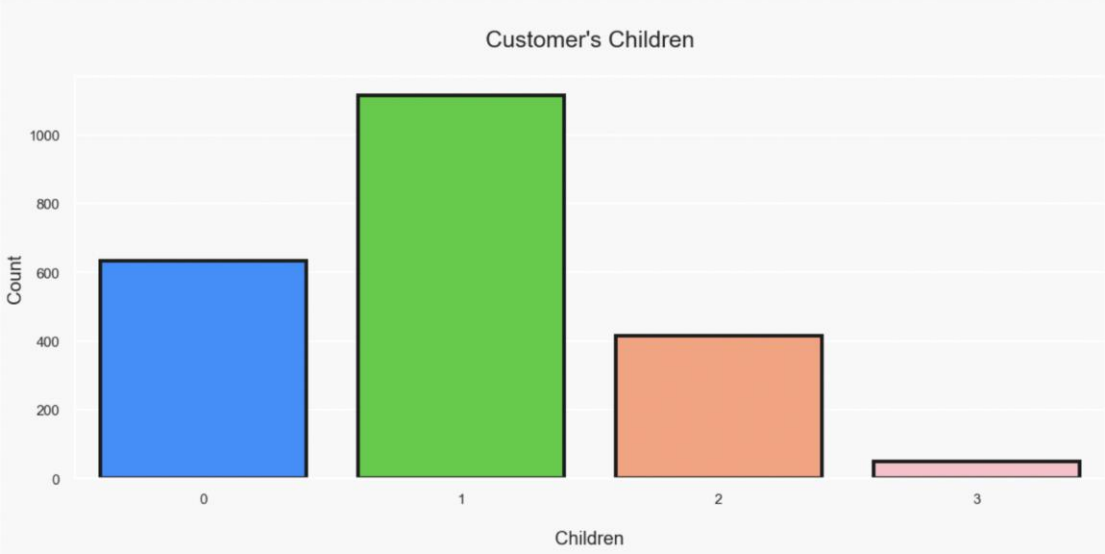

Fig7: The distribution of customer's children

For the distribution of customers' income and spending on basis of marital and parental status, we have plotted two separate scatter charts, one with income on the x-axis, marital status on the y-axis, and one with spending on the x-axis, marital status on the y-axis. We use different colors to represent the categories. We find out that most of the customers live together. And the customers who live together and are parents earn much and spend much.

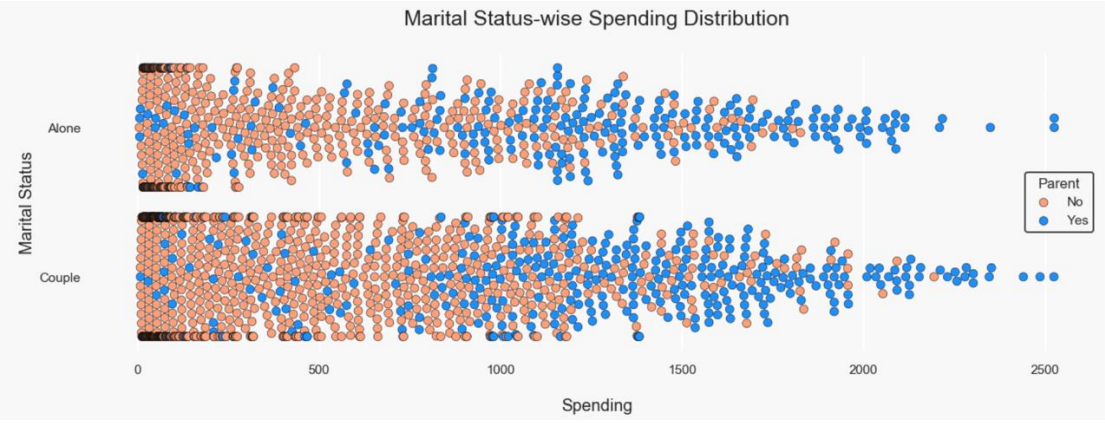

Fig8: Marital Status-wise Spending Distribution

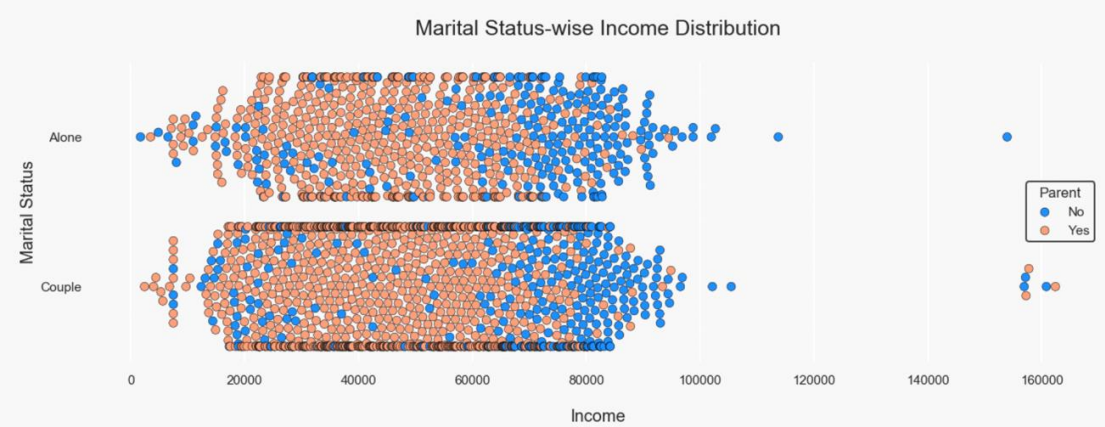

Fig9: Marital Status-wise Income Distribution

For the correlation of features, we create a correlation matrix heatmap to show the correlation between all variables. We have found out there is a high correlation between customers' income and spending. Also, there is a high correlation between buying wine and meat and buying through catalogs and stores with income and spending. Besides, there is a high correlation between buying meat and buying through the catalog.

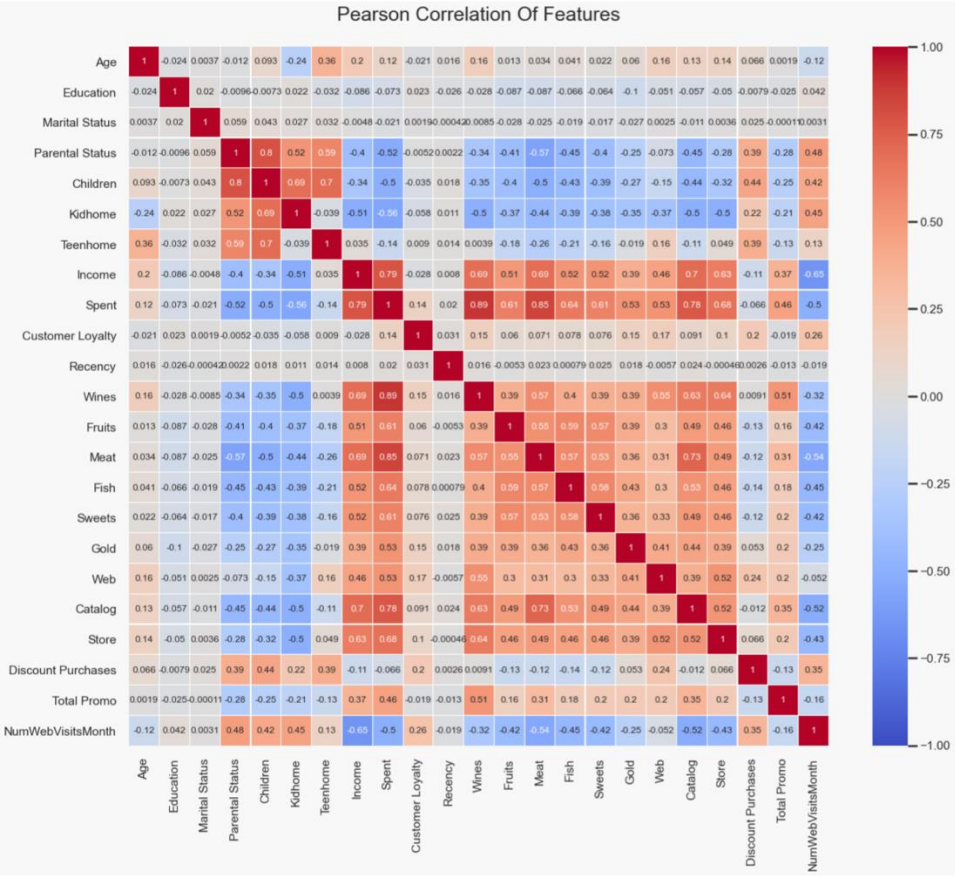

Fig10: Correlation of Features

3. **Linear Regression**

In addition, we performed a linear regression analysis with objective of identifying and predicting the factors that affect the amount of money people spend on a product. We conducted six linear regression models, with each model having the number of money spent on a single product as the dependent variable(y-variable) are the customer's personal information, promotions, and places as the independent variables(x-variables.)

Result statistic

OLS Regression Results

Dep. Variable:	MntWines	R-squared:	0.695
Model:	OLS	Adj. R-squared:	0.690
Method:	Least Squares	F-statistic:	165.6
Date:	Mon, 10 Apr 2023	Prob (F-statistic):	0.00
Time:	17:25:20	Log-Likelihood:	-11776.
No. Observations:	1772	AIC:	2.360e+04
Df Residuals:	1747	BIC:	2.374e+04
Df Model:	24		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Education	43.0478	4.788	8.990	0.000	33.656	52.440
Marital_Status	-0.0199	2.602	-0.008	0.994	-5.124	5.084
Income	0.0012	0.000	4.758	0.000	0.001	0.002
Kidhome	-46.3555	11.250	-4.121	0.000	-68.420	-24.291
Teenhome	11.5796	10.412	1.112	0.266	-8.841	32.001
Recency	0.2667	0.160	1.671	0.095	-0.046	0.580
MntFruits	-0.0478	0.155	-0.308	0.758	-0.351	0.256
MntMeatProducts	0.1837	0.034	5.360	0.000	0.116	0.251
MntFishProducts	0.0548	0.115	0.475	0.635	-0.172	0.281
MntSweetProducts	-0.2151	0.149	-1.447	0.148	-0.507	0.076
MntGoldProds	0.2264	0.109	2.086	0.037	0.013	0.439
NumDealsPurchases	-7.0744	2.973	-2.379	0.017	-12.906	-1.243
NumWebPurchases	19.2007	2.192	8.759	0.000	14.901	23.500
NumCatalogPurchases	22.3055	2.588	8.619	0.000	17.230	27.381
NumStorePurchases	31.0121	1.997	15.530	0.000	27.096	34.929
NumWebVisitsMonth	20.4873	2.798	7.321	0.000	14.999	25.976
AcceptedCmp3	45.3046	18.674	2.426	0.015	8.679	81.930
AcceptedCmp4	176.8656	20.133	8.785	0.000	137.379	216.353
AcceptedCmp5	244.9207	21.551	11.365	0.000	202.653	287.189
AcceptedCmp1	45.4200	21.265	2.136	0.033	3.712	87.128
AcceptedCmp2	73.5153	40.404	1.819	0.069	-5.731	152.761
Complain	-28.3903	45.980	-0.617	0.537	-118.572	61.791
Z_CostContact	-9.7984	0.835	-11.738	0.000	-11.436	-8.161
Z_Revenue	-35.9273	3.061	-11.738	0.000	-41.930	-29.924
Response	6.1593	14.990	0.411	0.681	-23.242	35.560
Age	0.7120	0.411	1.731	0.084	-0.095	1.519

Figure 11: Regression Results of wine

Figure 11 shows the linear regression result with the amount of money people spend on wine is set to be the dependent variable. From the result table. The r^2 is 0.695, indicating that 69.5% of the variance in the dependent variable is explained by the independent variables. There are 15 independent variables that have a p-value less than 0.05, meaning that these independent variables have a significant impact on the dependent variable. The RMSE values for the training and test sets are 186.1540 and 158.3761, respectively. This suggests that the model is not overfitting to the training data and is able to generalize well to new data.

OLS Regression Results

Dep. Variable:	MntFruits	R-squared:	0.486
Model:	OLS	Adj. R-squared:	0.479
Method:	Least Squares	F-statistic:	68.91
Date:	Mon, 10 Apr 2023	Prob (F-statistic):	4.33e-232
Time:	17:25:42	Log-Likelihood:	-8466.4
No. Observations:	1772	AIC:	1.698e+04
Df Residuals:	1747	BIC:	1.712e+04
Df Model:	24		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Education	-1.8564	0.755	-2.458	0.014	-3.338	-0.375
Marital_Status	0.2268	0.402	0.564	0.573	-0.562	1.015
Income	5.623e-05	3.77e-05	1.490	0.136	-1.78e-05	0.000
Kidhome	-0.4301	1.746	-0.246	0.805	-3.855	2.995
Teenhome	-2.5823	1.608	-1.606	0.108	-5.736	0.571
Recency	-0.0224	0.025	-0.907	0.365	-0.071	0.026
MntWines	-0.0011	0.004	-0.308	0.758	-0.008	0.006
MntMeatProducts	0.0340	0.005	6.444	0.000	0.024	0.044
MntFishProducts	0.1494	0.017	8.552	0.000	0.115	0.184
MntSweetProducts	0.2159	0.022	9.642	0.000	0.172	0.260
MntGoldProds	0.0775	0.017	4.643	0.000	0.045	0.110
NumDealsPurchases	-0.9200	0.460	-2.002	0.045	-1.821	-0.019
NumWebPurchases	0.3920	0.346	1.133	0.257	-0.286	1.070
NumCatalogPurchases	-0.5182	0.408	-1.270	0.204	-1.318	0.282
NumStorePurchases	1.8190	0.326	5.576	0.000	1.179	2.459
NumWebVisitsMonth	-0.4480	0.439	-1.021	0.307	-1.309	0.413
AcceptedCmp3	2.7317	2.889	0.946	0.345	-2.935	8.398
AcceptedCmp4	-2.6849	3.178	-0.845	0.398	-8.917	3.548
AcceptedCmp5	-4.1710	3.449	-1.209	0.227	-10.935	2.593
AcceptedCmp1	-2.2294	3.289	-0.678	0.498	-8.680	4.222
AcceptedCmp2	-10.2788	6.243	-1.646	0.100	-22.524	1.966
Complain	3.6233	7.104	0.510	0.610	-10.309	17.556
Z_CostContact	0.1691	0.134	1.263	0.207	-0.094	0.432
Z_Revenue	0.6200	0.491	1.263	0.207	-0.343	1.583
Response	0.5957	2.316	0.257	0.797	-3.947	5.138
Age	-0.0324	0.064	-0.509	0.611	-0.157	0.092

Figure 12: Regression Result of Fruits

Figure 12 displays the results of the linear regression analysis, with the amount of money people spend on fruits being the dependent variable. The r^2 value of 0.486 indicates that 48.6% of the variance in the dependent variable can be explained by the independent variables. Four independent variables have p-values less than 0.05, except the amount of money spent on other products, the number of times people shop in stores showing a significant impact on the amount of people spend on fruit. This suggests that people prefer purchasing fruit in-store. The RMSE values for the training set and test set are 28.8 and 27.4, respectively, which are acceptable values for RMSE of linear regression.

OLS Regression Results

Dep. Variable:	MntMeatProducts	R-squared:	0.683
Model:	OLS	Adj. R-squared:	0.679
Method:	Least Squares	F-statistic:	157.1
Date:	Mon, 10 Apr 2023	Prob (F-statistic):	0.00
Time:	17:26:04	Log-Likelihood:	-11124.
No. Observations:	1772	AIC:	2.230e+04
Df Residuals:	1747	BIC:	2.244e+04
Df Model:	24		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Education	1.2270	3.390	0.362	0.717	-5.423	7.877
Marital_Status	-5.0785	1.797	-2.826	0.005	-8.604	-1.553
Income	0.0009	0.000	5.545	0.000	0.001	0.001
Kidhome	-10.4943	7.821	-1.342	0.180	-25.835	4.846
Teenhome	-60.4387	7.064	-8.556	0.000	-74.293	-46.585
Recency	0.1723	0.110	1.560	0.119	-0.044	0.389
MntWines	0.0880	0.016	5.360	0.000	0.056	0.120
MntFruits	0.6828	0.106	6.444	0.000	0.475	0.891
MntFishProducts	0.4176	0.079	5.267	0.000	0.262	0.573
MntSweetProducts	0.1256	0.103	1.220	0.223	-0.076	0.327
MntGoldProds	-0.1650	0.075	-2.196	0.028	-0.312	-0.018
NumDealsPurchases	6.4646	2.056	3.145	0.002	2.432	10.497
NumWebPurchases	-1.7121	1.550	-1.105	0.269	-4.752	1.328
NumCatalogPurchases	30.8174	1.674	18.410	0.000	27.534	34.100
NumStorePurchases	-1.2457	1.474	-0.845	0.398	-4.137	1.646
NumWebVisitsMonth	-11.1647	1.948	-5.730	0.000	-14.986	-7.343
AcceptedCmp3	-46.2353	12.901	-3.584	0.000	-71.539	-20.932
AcceptedCmp4	-58.8401	14.172	-4.152	0.000	-86.635	-31.045
AcceptedCmp5	73.2519	15.360	4.769	0.000	43.125	103.379
AcceptedCmp1	8.1126	14.739	0.550	0.582	-20.795	37.020
AcceptedCmp2	-46.6063	27.974	-1.666	0.096	-101.472	8.260
Complain	-22.2101	31.829	-0.698	0.485	-84.636	40.216
Z_CostContact	2.1566	0.598	3.607	0.000	0.984	3.329
Z_Revenue	7.9077	2.193	3.607	0.000	3.607	12.208
Response	40.4238	10.332	3.912	0.000	20.159	60.689
Age	-0.2675	0.285	-0.939	0.348	-0.826	0.291

Figure 13: Regression Result of Meat

Figure 13 displays the results of the linear regression result with y-variables be the amount of money people spend on meat. The r^2 value of 0.683. There are 15 independent variables that have p-value less than 0.05, including income, number of teens at home, number of purchases made through the company's website, etc. Meaning there is strong relationship between these independent variables to the amount of money people spend on meat. The RMSE of the training and test set are 129 and 125 respectively.

OLS Regression Results

Dep. Variable:	MntFishProducts	R-squared:	0.521
Model:	OLS	Adj. R-squared:	0.514
Method:	Least Squares	F-statistic:	79.06
Date:	Mon, 10 Apr 2023	Prob (F-statistic):	5.41e-258
Time:	17:26:16	Log-Likelihood:	-8987.0
No. Observations:	1772	AIC:	1.802e+04
Df Residuals:	1747	BIC:	1.816e+04
Df Model:	24		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Education	-4.0939	1.010	-4.052	0.000	-6.076	-2.112
Marital_Status	-0.1466	0.539	-0.272	0.786	-1.204	0.911
Income	3.976e-05	5.06e-05	0.785	0.433	-5.96e-05	0.000
Kidhome	0.2581	2.343	0.110	0.912	-4.337	4.853
Teenhome	-7.9560	2.150	-3.700	0.000	-12.173	-3.739
Recency	-0.0299	0.033	-0.902	0.367	-0.095	0.035
MntWines	0.0024	0.005	0.475	0.635	-0.007	0.012
MntFruits	0.2689	0.031	8.552	0.000	0.207	0.331
MntMeatProducts	0.0374	0.007	5.267	0.000	0.023	0.051
MntSweetProducts	0.2590	0.030	8.579	0.000	0.200	0.318
MntGoldProds	0.1327	0.022	5.952	0.000	0.089	0.176
NumDealsPurchases	-1.4134	0.616	-2.293	0.022	-2.622	-0.205
NumWebPurchases	0.7466	0.464	1.610	0.108	-0.163	1.656
NumCatalogPurchases	1.5585	0.546	2.853	0.004	0.487	2.630
NumStorePurchases	1.3977	0.440	3.175	0.002	0.534	2.261
NumWebVisitsMonth	-1.2644	0.588	-2.150	0.032	-2.418	-0.111
AcceptedCmp3	-5.4994	3.874	-1.419	0.156	-13.098	2.099
AcceptedCmp4	-6.0929	4.261	-1.430	0.153	-14.450	2.265
AcceptedCmp5	-18.9414	4.606	-4.112	0.000	-27.976	-9.907
AcceptedCmp1	15.5329	4.397	3.532	0.000	6.909	24.157
AcceptedCmp2	-3.5383	8.381	-0.422	0.673	-19.976	12.900
Complain	-6.0029	9.529	-0.630	0.529	-24.693	12.687
Z_CostContact	0.4511	0.179	2.515	0.012	0.099	0.803
Z_Revenue	1.6539	0.658	2.515	0.012	0.364	2.944
Response	-3.2260	3.106	-1.039	0.299	-9.318	2.866
Age	0.1090	0.085	1.278	0.202	-0.058	0.276

Figure 14: Regression Result of Fish

Figure 14 illustrates the results of the result with the dependent variable be the amount of money people spend on fish. The r^2 value is 0.521, represents that approximately half of the variance of the dependent variables can be explained by the independent variables. There are 13 independent variables that have p-value less than 0.05. The RMSE of the training and test set are 39 and 32.2 respectively suggesting that the model has an acceptable level of accuracy in predicting the amount of money people spend on fish.

Figure 15: Regression Result of Sweets

Figure 15 illustrates the results with the dependent variable be the amount of money

people spend on sweets. The r^2 value is 0.475. There are 12 independent variables that have p-value less than 0.05. The RMSE of the training and test set are 30 and 28 respectively.

Figure 16: Regression Result of Gold

Figure 16 displays the results of the linear regression analysis with the amount of money people spend on gold as the dependent variable. The R-squared value is 0.366 indicates that only one-third of the variance in the independent variable can be explained by the independent variables. However, the model can still identify which variables have a significant impact on the amount of money people spend on gold. Seventeen variables have p-value less than 0.05, indicating their significant impact. For the RMSE, it is 41 for the training test and 45 for the test set.

In summary, the linear regression models' overall performance is mediocre, with r^2 value ranging from 0.366 to 0.695. However, each model with a different spending amount for each product reveals that some independent variables have a significant effect on the dependent variable. We found that the amount spent on other products always has a p-value less than 0.05, suggesting that if people spend more on one product, they are likely to spend more on another product as well. The RMSE values are overall not too high, indicating that the predicted values are generally close to the actual values.

4. **K means clustering**

The aim of clustering is to group customers into distinct clusters based on their characteristics and behaviors, so that businesses and companies can have a better understanding of distinct groups of customers and make suitable strategies to develop products and services that meet the need of different customers. For example, if a business identifies a group of customers who consistently purchase luxury products and

they are young, companies can use this information to create marketing campaigns that focus on the branded goods. Meanwhile, if a company discovers a market of buyers that are price-sensitive, they can develop marketing campaigns that emphasize the accessibility of their product. In this case, we use k means clustering because it is relatively easy to understand and interpret, we will separate the customers into k groups by using elbow method.

Evaluation:

When deciding on the final classification, several different factors related to this issue will be considered. In essence, these factors are attributes or characteristics. The greater the number of functions, the harder it is to utilize. Since they are related, many of these characteristics are redundant. I will therefore perform dimensionality reduction on the characteristics prior to subjecting them to a classifier. We use PCA reduction to reduce the dimensionality and increase the interpretability of the dataset. I will reduce the dimensionality to three.

After using standard scaler to scale and standardize the features of a dataset. We fit PCA on the scaled data and plot the graph of data after its dimension is reduced to three

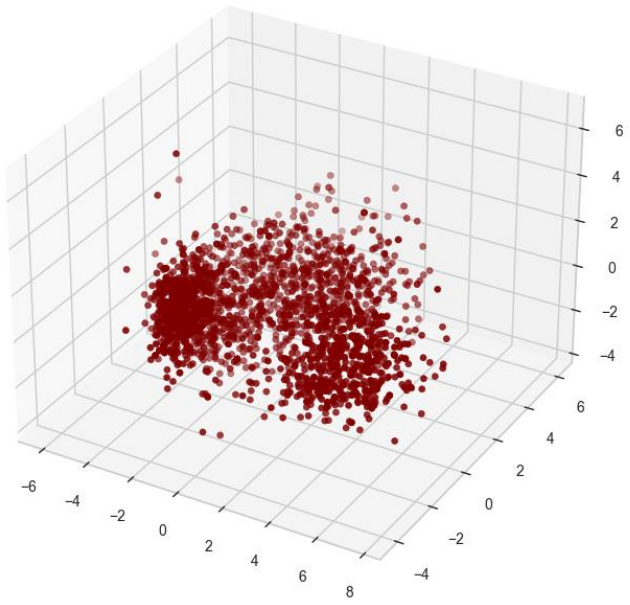

Fig17: 3D projection of data after dimension reduction

After the reduction of dimensionality, I will perform k means clustering using elbow method, that is to determine the optimal number of k by finding the elbow point, and evaluate the clusters in a 3D projection

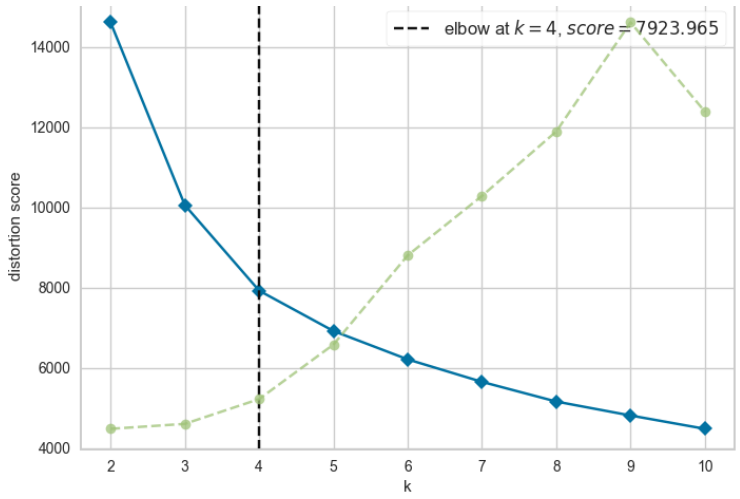

Fig18: Distortion score of elbow method for k means clustering

The k means algorithm is trying to create 10 clusters, it will be run on the data with k value ranging from 1 to 10, and the inertia for each k value will be calculated. From figure 18, the elbow point is 4 where the rate of decrease in inertia starts to level off, it is indicated that 4 is the optimal number of clusters. I will be fitting k means clustering model to get the final clusters in 3d projection with a scatter plot, and the result is shown in figure 19.

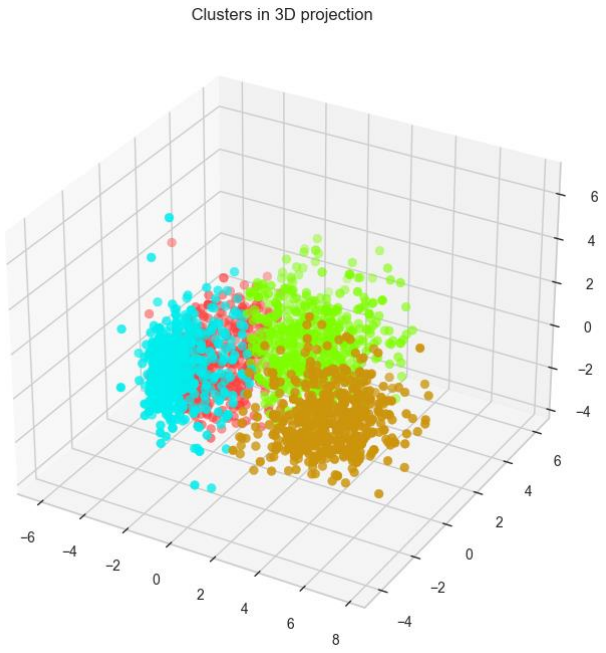

Fig19: Clusters in 3D projection

As mentioned before, our goal is to determine the characteristics and nature of formed clusters, therefore, exploratory data analysis will be used to examine the data in relation to the clusters and draw conclusions depending on what we find. We will start by examining the distribution of the groups among the clusters. Firstly, we go through the clusters based on their spending and income.

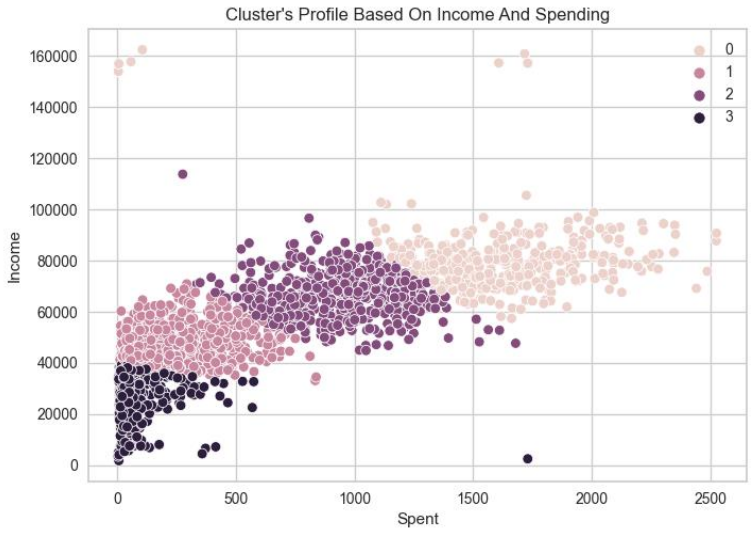

Fig20: Cluster's profile based on income and spending

From figure 20, we can observe the distribution of each customer and they are classified into four different groups. We can draw conclusions based on this figure:

Cluster 0: People with high spending and high income

Cluster 1: People with low spending and medium income

Cluster 2: People with high spending and medium income

Cluster 3: People with low spending and low income

After using the k means clustering algorithm to group the customers into different clusters based on their spending behaviors, it's important to understand the characteristics of each cluster. One way to gain insight into the spending habits of each cluster is to examine the percentage of spending on different products within each cluster.

In this case, the six products in question are wines, fruits, meat, fish, sweets, and gold. By analyzing the percentage of spending on each of these products within each cluster, we can gain insight into the unique spending patterns and preferences of each group.

For example, we might find that customers in one cluster tend to spend a higher percentage of their budget on meat and fish, while customers in another cluster tend to spend more on sweets and fruits. By identifying these patterns, we can gain a better understanding of the different customer segments and tailor our marketing strategies accordingly.

Overall, analyzing the percentage of spending on different products within each cluster is an important step in gaining a deeper understanding of the spending habits and preferences of our customers.

Fig21: Cluster's spending on different product

Based on figure 21, obtained through the analysis of the spending habits of different customer clusters, we can make a number of observations about the spending patterns of each group.

Firstly, it is evident that all of the clusters have the highest spending on wines, followed by meat. This suggests that wines and meat are the most popular products among all of the customer groups.

Secondly, it is noted that all of the clusters have the least spending on fruits and sweets, indicating that these products are less popular among customers in general.

Moreover, we can see that cluster 3 are more willing to purchase gold products and cluster 0 are more willing to purchase meat when compared to other clusters. This finding could be useful in tailoring marketing strategies specifically to this group of customers.

Furthermore, to see what is in these clusters, we create a joint plot to show the relationship between spending to other attributes, that is finding whether customers are parents or not, their family size, their age and the number of children they have.

Fig 22-25 Joint plot with spent to different attributes

Based on the joint plots, we can make several conclusions about the characteristics of each customer cluster:

1. Cluster 0: This group is likely not comprised of parents, as they have high spending levels and high incomes. Their family size is mainly 2, and they do not have any children.
2. Cluster 1: This cluster is likely comprised of parents, as they have lower spending levels and lower incomes. Their family size is mainly 2 to 4, and most of them have only one child.
3. Cluster 2: This group is also likely comprised of parents, as they have higher spending levels and medium incomes. However, their family size and the number of children are not explicitly stated, so we cannot draw any specific conclusions about these attributes.
4. Cluster 3: This group is mostly comprised of parents, with low spending levels and low incomes. Their family size is mainly 3, and most of them have only one child.

Overall, these conclusions provide important insights into the customer segments represented by each cluster. By understanding the unique characteristics of each group, businesses can tailor their marketing and sales strategies to better meet the needs and preferences of each customer segment, which can ultimately lead to increased sales and revenue.

5. **Conclusion**

In conclusion, customer personality analysis is a useful technique for companies seeking to understand their customers and make data-driven decisions. By analyzing customer data such as income, age, marital status, promotion history, and purchase history, companies can gain insights into customer behavior and preferences. This information can be used to make informed marketing decisions, improve customer experiences, and develop new products.

In this project, we used several techniques to perform customer personality analysis, including PCA, K-mean clustering, and linear regression. By using these methods, we were able to identify four distinct customer clusters based on income, family size, child, and spending level. These clusters provided valuable insights into the characteristics and behaviors of different customer groups, which can help companies tailor their marketing strategies to better meet the needs and preferences of each segment.

Furthermore, our linear regression analysis allowed us to identify key variables that were most strongly associated with total spending on items. This information can be used to develop targeted marketing campaigns that are more likely to resonate with customers and drive sales.

Overall, our project demonstrates the importance of customer personality analysis as an important tool for companies seeking to gain a deeper understanding of their customers. By leveraging data and analytics, companies can make better decisions and provide personalized experiences that improve customer satisfaction and drive business growth.

Reference

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

Source code: https://portland-my.sharepoint.com/personal/tszyinlui2-c_my_cityu_edu_hk/_layouts/15/onedrive.aspx?ga=1&id=%2Fpersonal%2Ftszyinlui2%2Dc%5Fmy%5Fcityu%5Fedu%5Fhk%2FDocuments%2Fschool%20work%20file%2FHI%20Gp%20project%20codes