Demystifying Customer Behavior: Insights from Data Analysis

STATGR5291 Advanced Data Analysis

Team 2

Team Members:

Tianyi Zhu (tz2538)
Hanyun Hu (hh2951)
Sally Wang (hw2917)
Zekai Shen (zs2578)
Amanda Wang (zw2764)

Contents

1 Introduction

1.1 Background Information

Data Analysis plays a significant role in making business decisions. One of the analysis technique is customer personality analysis. It helps a business identify ideal customers and better understand their customer pool when marketing. For example, instead of advertising high-end wines to all customers in their database, the business can promote the product only to those who are more likely to purchase. In this way, the business can reduce their advertising costs, which might lead to an increase in profits.

Broadly speaking, the use of customer personality analysis has become increasingly popular in recent years due to the growth of online shopping and social media. The vast amount of data lays the foundation for detailed customer analysis and profiling. Within the same industry, data are usually segmented by their demographic, behavior, lifestyle, psychographic, value, etc. By breaking down customer data, businesses can gain insights into customer preferences, behaviors, and needs, and thus they can tailor products and strategies accordingly. In a nutshell, customer data analysis is essential in developing effective market strategies and building strong customer relationships, which are beneficial for businesses in the long run.

1.2 Objectives

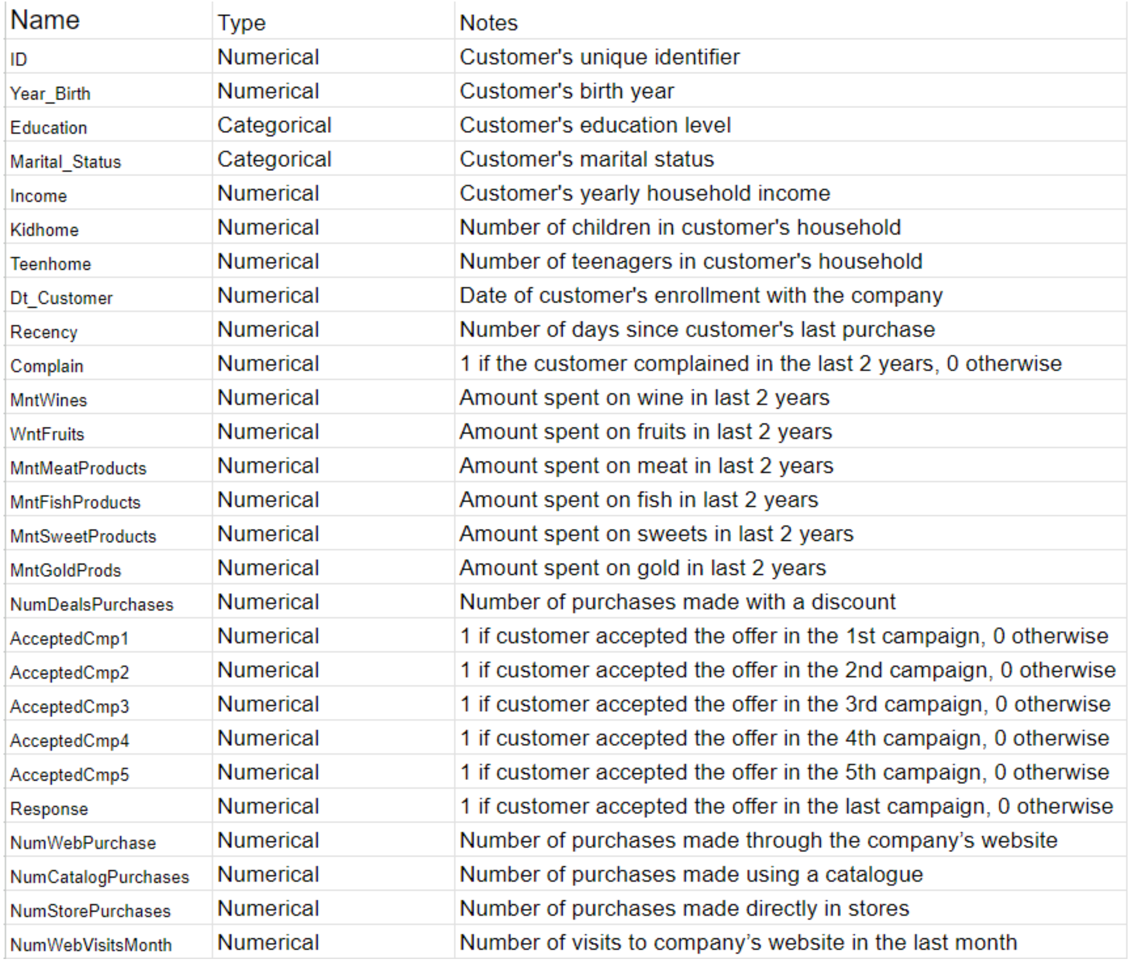
For this project, we are trying to analyze customer personality in order to provide marketing strategies for companies. Customer personality analysis is an essential tool that enables businesses to gain a closer insight of their ideal customers. By analyzing the specific needs and behaviors of different customer segments, companies can improve their products and services accordingly. This not only enhances customer satisfaction but also helps businesses build their marketing strategies more effectively. In order to better understand customer personality, we will find the answer for the following two research questions.

Research Question 1 : In what ways can we effectively cluster our customers based on their demographic and behavioral characteristics to gain a better understanding of their needs?

Research Question 2 : How can we analyze the behavior of customers with each identified cluster and among all clusters, and what insights can we gather to improve overall customer satisfaction?

1.3 Data Description

To answer the research questions, we found the data set from Kaggle [1], and the data set was provided by Dr. Omar Romero-Hernadez. The data set has 2240 rows recording information of 2240 customers and 27 columns containing various information about people, products, promotion, and place. More information on these variables can be found in Figure 1.



Name	Type	Notes
ID	Numerical	Customer's unique identifier
Year_Birth	Numerical	Customer's birth year
Education	Categorical	Customer's education level
Marital_Status	Categorical	Customer's marital status
Income	Numerical	Customer's yearly household income
Kidhome	Numerical	Number of children in customer's household
Teenhome	Numerical	Number of teenagers in customer's household
Dt_Customer	Numerical	Date of customer's enrollment with the company
Recency	Numerical	Number of days since customer's last purchase
Complain	Numerical	1 if the customer complained in the last 2 years, 0 otherwise
MntWines	Numerical	Amount spent on wine in last 2 years
WntFruits	Numerical	Amount spent on fruits in last 2 years
MntMeatProducts	Numerical	Amount spent on meat in last 2 years
MntFishProducts	Numerical	Amount spent on fish in last 2 years
MntSweetProducts	Numerical	Amount spent on sweets in last 2 years
MntGoldProds	Numerical	Amount spent on gold in last 2 years
NumDealsPurchases	Numerical	Number of purchases made with a discount
AcceptedCmp1	Numerical	1 if customer accepted the offer in the 1st campaign, 0 otherwise
AcceptedCmp2	Numerical	1 if customer accepted the offer in the 2nd campaign, 0 otherwise
AcceptedCmp3	Numerical	1 if customer accepted the offer in the 3rd campaign, 0 otherwise
AcceptedCmp4	Numerical	1 if customer accepted the offer in the 4th campaign, 0 otherwise
AcceptedCmp5	Numerical	1 if customer accepted the offer in the 5th campaign, 0 otherwise
Response	Numerical	1 if customer accepted the offer in the last campaign, 0 otherwise
NumWebPurchase	Numerical	Number of purchases made through the company's website
NumCatalogPurchases	Numerical	Number of purchases made using a catalogue
NumStorePurchases	Numerical	Number of purchases made directly in stores
NumWebVisitsMonth	Numerical	Number of visits to company's website in the last month

Figure 1: Descriptions of each attribute of raw data set[1]

1.4 Methodology

- Data Collection: The data for this research was downloaded from Kaggle. It consists of 2240 individuals whose information were recorded. The sample is considered a representation of the population.

- Data Analysis: The data collected was analyzed first using descriptive statistics and then some machine learning methods. Descriptive statistics such as mean, standard deviation, and frequency distribution were used to summarize the data. Machine learning methods such as Principle Component Analysis (PCA) and K-means clustering (unsupervised) were used to identify significant variables and explore hidden patterns within the data that may not be apparent through visual inspection.

- Limitations: One limitation of this study is the use of sampling method. The source lacks information on how the data was collected. This limits the generalizability of the findings to the population being studied if bias exists. We deal with this issue in the Assumptions section.

2 Assumptions

A few assumptions have been made while cleaning the data set and building the model for customer behaviors analysis, the most significant of which are stated below:

- For the Education column containing 'Graduation', 'PhD', 'Master', 'Basic', '2n Cycle'. Among them '2n Cycle' is not a commonly used term in education. According to our research, '2n Cycle' is often considered part of basic education and includes programs like bachelors degrees and master degrees. So we consider '2n Cycle' as basic education. "Graduation" is also a vague term and we determine it as basic education.
- Our customer behavioral analysis model is trained based on our data set. It might be argued that other important variables can be crucial to analyze customer behavior and they might have a huge impact on our results. We are only focusing on customer behavioral variables from our data set.
- We assume the data is a good representative of the population and it is collected in an unbiased way. Each customer on the data set is independent and identically distributed.

3 Data Cleaning

In this section, we prepared the data for further analysis. We checked for uniqueness and missing values; we dealt with numerical and categorical variables separately, and finally, we created new metrics that would add to modeling in the next section.

3.1 Check Uniqueness

The first thing we did was to check if duplicated data existed to make sure that every record was unique. Our results showed there were no duplicate data in the table.

3.2 Missing Values

To check for missing values, we used .dropna() to delete rows with any value missing. We discovered that the only column with missing values was the column "Income" and that the number of values missing was 24. In other words, there were 24 customers for which we have no data on their incomes. To resolve the issue, we could either replace them with the average income of all customers or the median income. However, since only a small percentage of the total number of customers has missing values (1.07%), we remove these rows for the purpose of convenience. With the rest of the data for 2216 customers, we are still able to develop an efficient customer segmentation model, which will be shown later.

3.3 Categorical Variable Assessment

With 2216 remaining customers, we conduct the follow-up analysis by first dealing with categorical variables. Two categorical variables are "Education" and "Marital_Status". We assess each variable and determine if we should regroup them in a more meaningful way.

- Education: Unique values for this variable are: "Ph.D.", "Master", "2n Cycle", "Graduation", and "Basic". We group the first three categories under "Higher Education", while the last category is under "Basic Education".
- Marital_Status: Unique values for this variable are: "Single", "Together", "Married", "Divorced", "Widow", "Alone", "Absurd", and "YOLO". We group "Together" and "Married" under a new category called "Pair", while the rest are regrouped into "Alone".

We proceed with the data set after categorical values are reorganized.

3.4 Numerical Variable Assessment and Outliers

To assess numerical values, we create a box plot for each variable, and then we remove outliers using the Interquartile Range (IQR) method.

The Interquartile Range (IQR) is a statistical measure that describes the spread or dispersion of a dataset. It is calculated as the difference between the third quartile (Q3) and the first quartile (Q1) of a dataset. It is commonly used in box plots to visualize the distribution of a dataset and identify potential outliers.

We take column **"Age"** as an exmaple. Since we are given "Year_Birth" in the original data set, we use 2023 (current year) minus the birth year for each customer to calculate their ages. We then draw the box plot for "Age", as shown in Figure 2. From this, we calculate the IQR for the "Age" column. An observation is considered an outlier if it falls below Q1 - 1.5× IQR or above Q3 + 1.5×IQR. We remove any data that falls outside of this range.

Similarly, we follow the same rule and identify outliers for other numerical variables, including "Dt_Customer", "Income", "Children", "Recency". After this step, we then proceed to feature engineering.

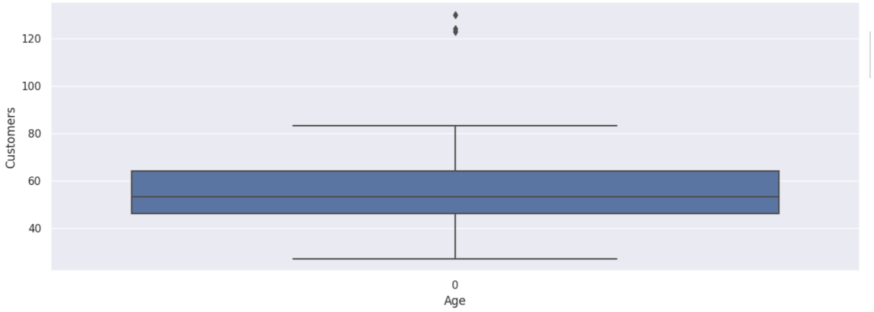


Figure 2: Plot showing the outliers in column "Age" [1]

3.5 Feature Engineering

In this section, we select certain variables and prepare them in a way that is suitable for modeling later on. We create the following features:

- Age = Current year-Year_Birth
- Total_Sales = MntWines + MntFruits + MntMeatProducts + MntFishProducts + MntSweetProducts + MntGoldProds
- Total_Purchases = NumDealsPurchases + NumWebPurchases + NumCatalogPurchases + NumStorePurchases
- Avg_Purchase = Total_Sales/Total_Purchases
- Deal_Share = NumDealsPurchases/Total_Purchases
- Web_Share = NumWebPurchases/Total_Purchases
- Catalog_Share = NumCatalogPurchases/Total_Purchases
- Store_Share = NumStorePurchases/Total_Purchases
- Family_Size = number of marriage + number of children

With these new features, we delete existing features that are not necessary. These features will reappear in modeling.

3.6 Adding Columns

To better analyze our data, we added a column named "Seniority". Our data contains consumer purchase date in between 2012 to 2014 and the most recent purchase date is 12/06/2014. "Seniority" is calculated by subtracting consumer's purchase date by 12/06/2014 and divide the resulting days by 30 to give "Seniority" an unit in month. Therefore, "Seniority" means how active each consumer is by month.

We have complete the steps of data cleaning and can utilize the refined data set to build our model.

3.7 EDA

In this section, we tried to examine and understand if there was any pattern, trend, and relationship within the data set, so we could better understand the characteristics of the data and check if there was anything that might inform decision making later on.
In order to further understand customers, we took a closer look at their age and education.To begin with, we wanted to understand how many customers within each age groups. We created a histogram showed that 15 customers were 19 to 30 years old, 283 customers were 31 to 40 years old, 608 customers were 41 to 50 years old, 611 customers were 51 to 60 years old, and 686 customers were older than 61 years old. Moreover, we were interested in the relationship between customers' age and purchasing behavior. We noticed that customers within age group 19 to 30 and age group 61+ had made the most purchases, which was about 16 to 17 purchases, while customers within age group 31 to 40 had made the least purchases which was about 12 purchases. Furthermore, we discovered that more than half of the customers had basic education, while a smaller portion of the customers had high education. From our discoveries, we believed our data was a good sample of the population.

We also created a correlation matrix plot to see the strength and direction of the linear relationship between two variables.

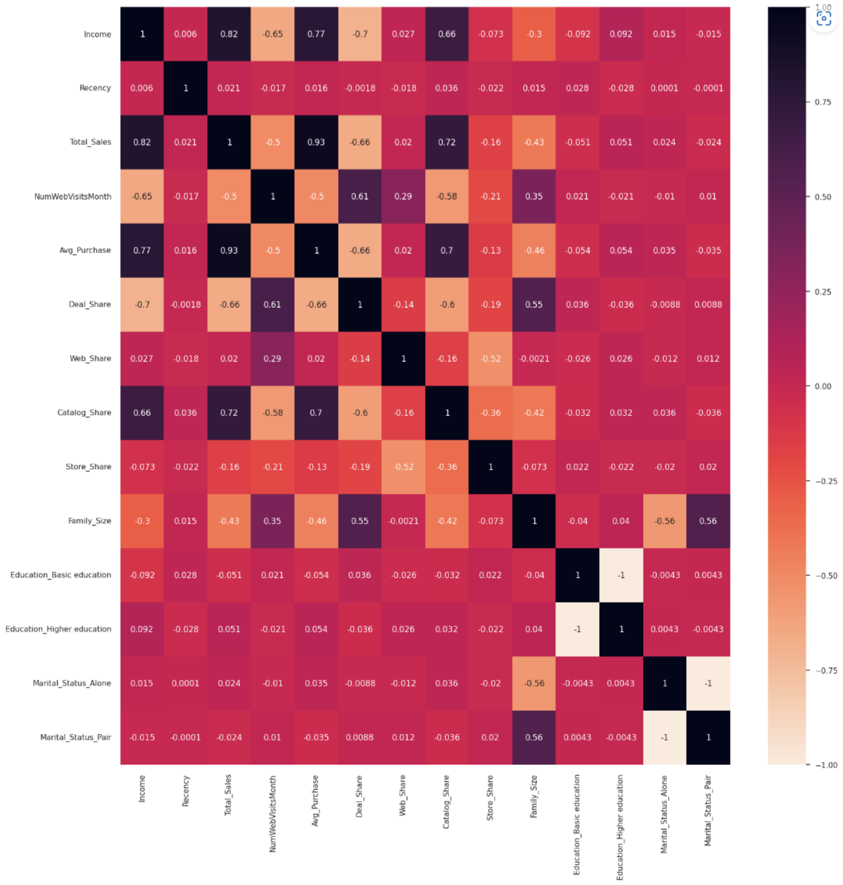


Figure 3: Correlation Matrix

4 Model

4.1 Principle Component Analysis

4.1.1 Definition

Principle Component Analysis(PCA) is a statistical technique for analyzing large data set by applying Singular Value Decomposition of the data to project it to a lower dimensional space. The technique can also be described as Linear dimensionality reduction. PCA technique helps us with data dimensions reduction, variable selection, and data visualization.[2] Algorithm of PCA is:[5]

- Compute empirical covariance matrix of the data (with d dimensions):

$$\hat{\sum}_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu_n})(x_i - \hat{\mu_n})^T$$

  where $\hat{\mu_n}$ is the empirical mean of data

$$\hat{\mu_n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Compute its eigenvalues $\lambda_1, . . . , \lambda_D$ and (normalized) eigenvectors $\varepsilon_1, . . . , \varepsilon_D$.
- Choose the d largest eigenvalues, say, $\lambda_{j1}, . . . , \lambda_{jd}$.
- Define subspace as V := span$\{\varepsilon_{j1}, . . . , \varepsilon_{jd}\}$.
- Project data onto V: For each xi, compute

$$x_i^v = \sum_{j=1}^{d}\langle x_i, \lambda_j\rangle\lambda_j$$

4.1.2 Application to Data Set

We first applied Python implemented PCA function to fit our data and plot the variance ratio of each component. By definition, the higher the variance ratio, the more important the component is.

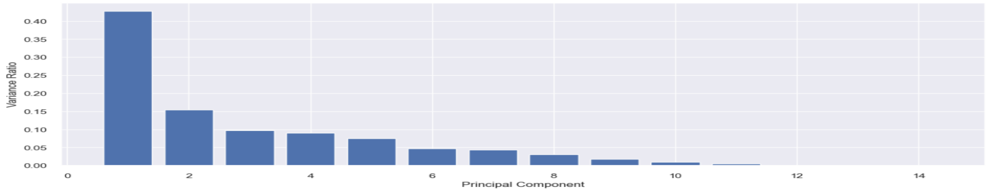
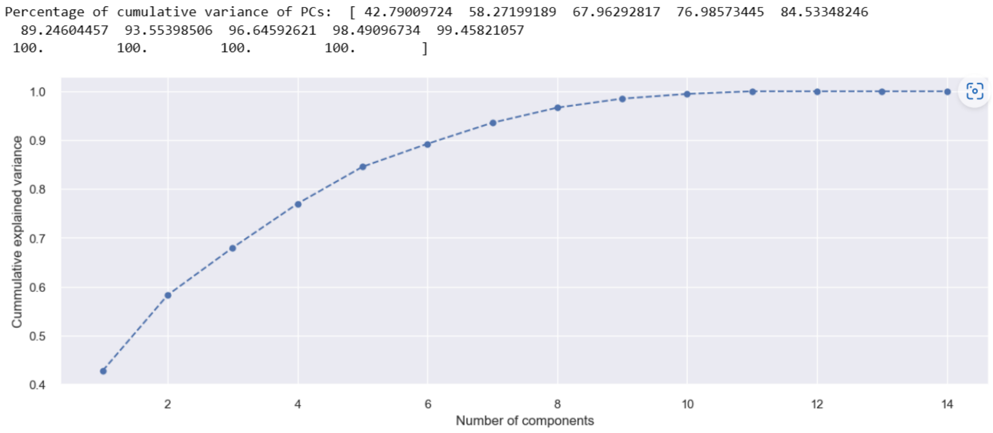


Figure 4: Variance Ratio vs Principal Component

We then plotted cumulative explained variance which helps us to quantify how much regression line is useful to predict or model our data. Besides, In social sciences setting, a cumulative variance explained should not be less than 50%.[3]



```
Percentage of cumulative variance of PCs:  [ 42.79009724  58.27199189  67.96292817  76.98573445  84.53348246
  89.24604457  93.55398506  96.64592621  98.49096734  99.45821057
 100.        100.        100.        100.        ]
```

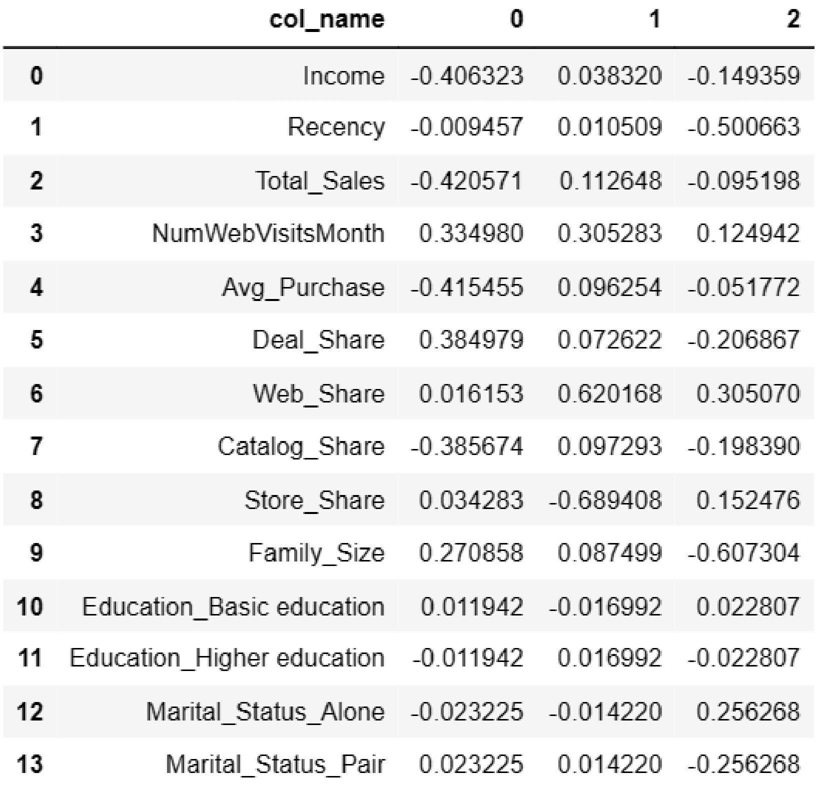
Figure 5: Cumulative Explained Variance vs Principal Component

The number of components we chose ultimately depends on our preference. Typically, it is recommended that the total variance explained by all components should be between 70% to 80%, which would equate to roughly 6 components in this case. Despite this, for the present example, only 3 components will be selected, which should preserve a little over 55% of the variance. While this may not be ideal in social sciences, it is still a reasonable level of variance retention. PC1 is responsible for around 36% of the overall variation in the data set and has the greatest impact on it.[3]

4.1.3 PCA Results

Eventually, we chose three components and fitted the model with our data with the 3 selected components using pca.fit(). pca.transform() calculated the resulting components scores for elements in our data set. We then created a data frame of PC components with scores.

In general, the biggest absolute value of each PC Score demonstrates the most significant variable. From our data frame, The variables Income and Total_Sales make the most significant contribution to PC1 (Avg_Purchase is also based on Total_Sales). The variables Store_Share and Web_Share are the primary contributors to PC2.



	col_name	0	1	2
0	Income	-0.406323	0.038320	-0.149359
1	Recency	-0.009457	0.010509	-0.500663
2	Total_Sales	-0.420571	0.112648	-0.095198
3	NumWebVisitsMonth	0.334980	0.305283	0.124942
4	Avg_Purchase	-0.415455	0.096254	-0.051772
5	Deal_Share	0.384979	0.072622	-0.206867
6	Web_Share	0.016153	0.620168	0.305070
7	Catalog_Share	-0.385674	0.097293	-0.198390
8	Store_Share	0.034283	-0.689408	0.152476
9	Family_Size	0.270858	0.087499	-0.607304
10	Education_Basic education	0.011942	-0.016992	0.022807
11	Education_Higher education	-0.011942	0.016992	-0.022807
12	Marital_Status_Alone	-0.023225	-0.014220	0.256268
13	Marital_Status_Pair	0.023225	0.014220	-0.256268

Figure 6: PC Scores Data Frame

4.2 K-Means Clustering

4.2.1 Definition

K-means clustering is a machine learning algorithm which could be simplified as grouping similar data points together and discovering underlying patterns. To achieve that, K-means looks for a fixed number of clusters in a data set. A cluster refers to a collection of data points aggregated together because of certain similarities. [4]
Algorithm of K-Means is:[5]

- Randomly choose K "cluster centers" (the "means") $\mu_1, \ldots, \mu_K \in \mathbb{R}^d$
- Iterate until convergence (j = iteration number):
- Assign each x_i to the closest (in Euclidean distance) mean:

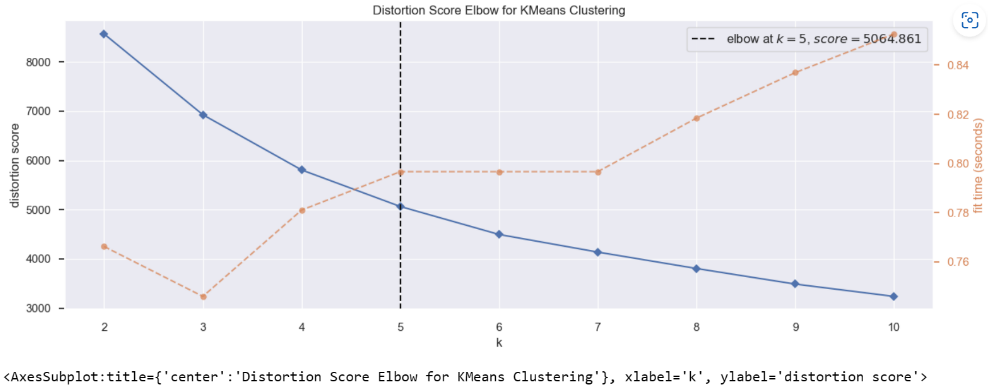
$$m_i^{j+1} := argmin_{k \in \{1,\ldots,K\}} \|x_i - \mu_k^j\|$$

- Recompute each μ_k^j as the mean of all points assigned to it:

$$\mu_k^{j+1} := \frac{1}{|i|m_i^{j+1} = k|} \sum_{i:m_i^{j+1}=k} x_i$$

4.2.2 Elbow method

We first used Kmeans() function to assign each data randomly to one of the K clusters. The function computes the center of each cluster called centroid and reassign data to their closest centroid until the cluster assigned for each data is no longer changing. Then we applied Elbow Method to determine the optimal value of K. We used function KElbowVisualizer() with inout of our Kmeans() model with a maximum range of k=10 values. We calculated the sum of the squared distance between each point and the centroid in a cluster. We called the calculation WCSS (Within-Cluster Sum of Squares) and plotted it with the K values.



```
<AxesSubplot:title={'center':'Distortion Score Elbow for KMeans Clustering'}, xlabel='k', ylabel='distortion score'>
```

Figure 7: Elbow Plot

As K value increases, the WCSS value decreases. From our graph, we could see that the blue line rapidly changes when K=5 and creates an elbow shape. From that point, the blue line moves gently to the X-axis. K=5 is the optimal value of K. The dashed green line represents the amount of time to train the clustering model per K.

We then combined our data with the Kmeans model with K=5. Add the cluster label to our original data frame. Then we plotted the histograms of 5 clusters. The figure shows the distribustion of clusters.

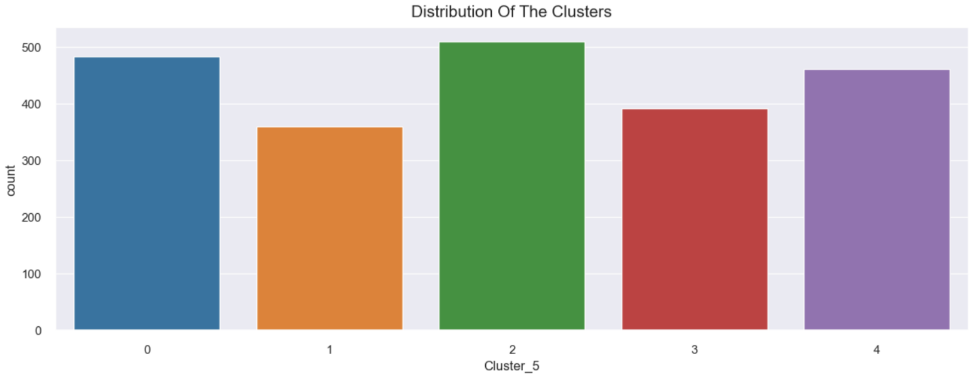
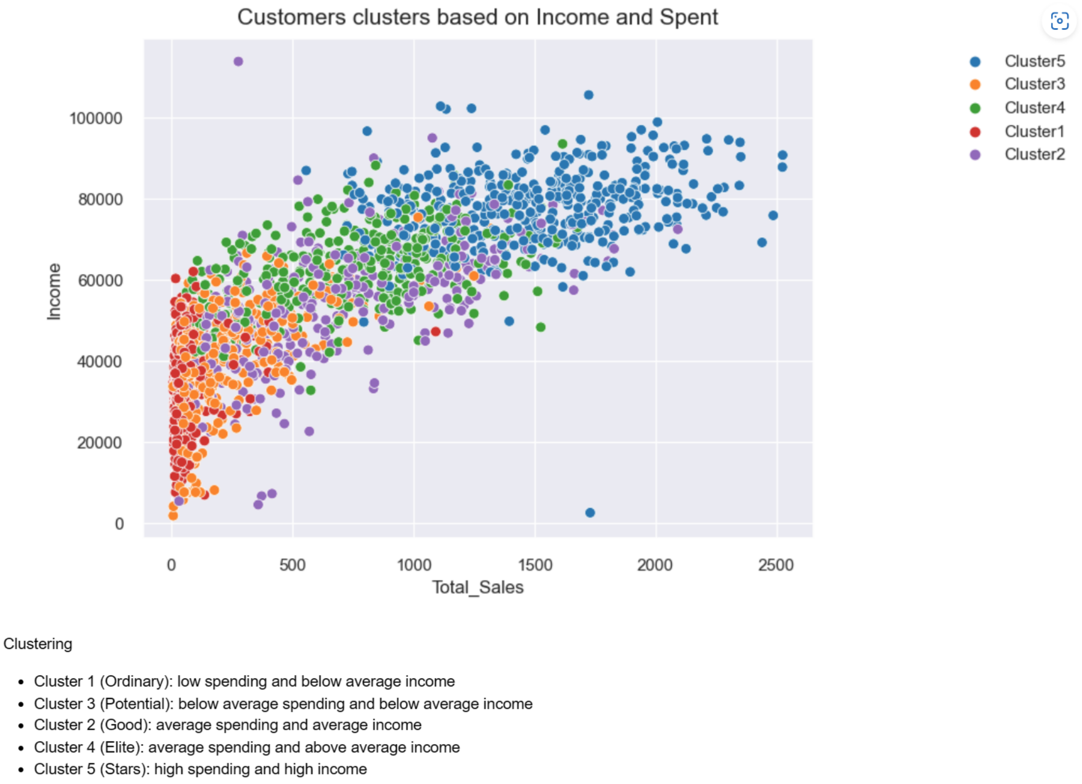


Figure 8: Cluster Histogram

Since we mentioned in 4.1.3 PCA Results that from our data, the variables Income and Total_Sales make the most significant contribution to PC1. Thus, we plotted clusters of customers based on Income and Total_Sales (we rename as Spent in the plot).



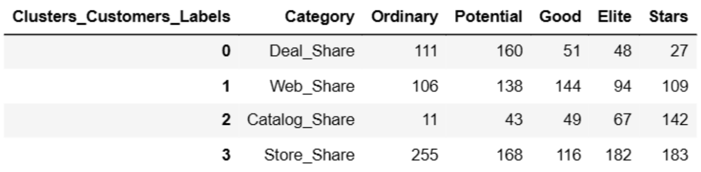
Clustering

- Cluster 1 (Ordinary): low spending and below average income
- Cluster 3 (Potential): below average spending and below average income
- Cluster 2 (Good): average spending and average income
- Cluster 4 (Elite): average spending and above average income
- Cluster 5 (Stars): high spending and high income

Figure 9: Income and Spent Clusters

We discovered that cluster 5 (blue dots) has the highest income and highest total sales, we named it as Stars. The following tier is cluster 4 (green dots) which has the average spending and above average income and we named it as Elite. The third tier is cluster 2 (purple dots) demonstrating average spending and average income and with name Good. The fourth tier is cluster 3 (orange dots) representing below average spending and below average income with name Potential. The last tier is cluster 1 and has low spending and below average income with name Ordinary.

Also, we mentioned in 4.1.3 PCA Results that from our data, the variables Store_Share and Web_Share are the primary contributors to PC2. We used Income and Total_Sales from PC1 for clustering and now use Store_Share and Web_Share from PC2 to analyze which marketing strategy on sales platforms selection. We had four platforms Deal_Share, Wed_Share, Catalog_Share, and Store_Share. We displayed the structure of customers shopping platforms in each cluster.



Clusters_Customers_Labels	Category	Ordinary	Potential	Good	Elite	Stars
0	Deal_Share	111	160	51	48	27
1	Web_Share	106	138	144	94	109
2	Catalog_Share	11	43	49	67	142
3	Store_Share	255	168	116	182	183

Figure 10: Shopping Platform Preference

Plotting the histogram of each cluster would perform a better visualisation for our approach. We use matplotlib to plot the histograms below.

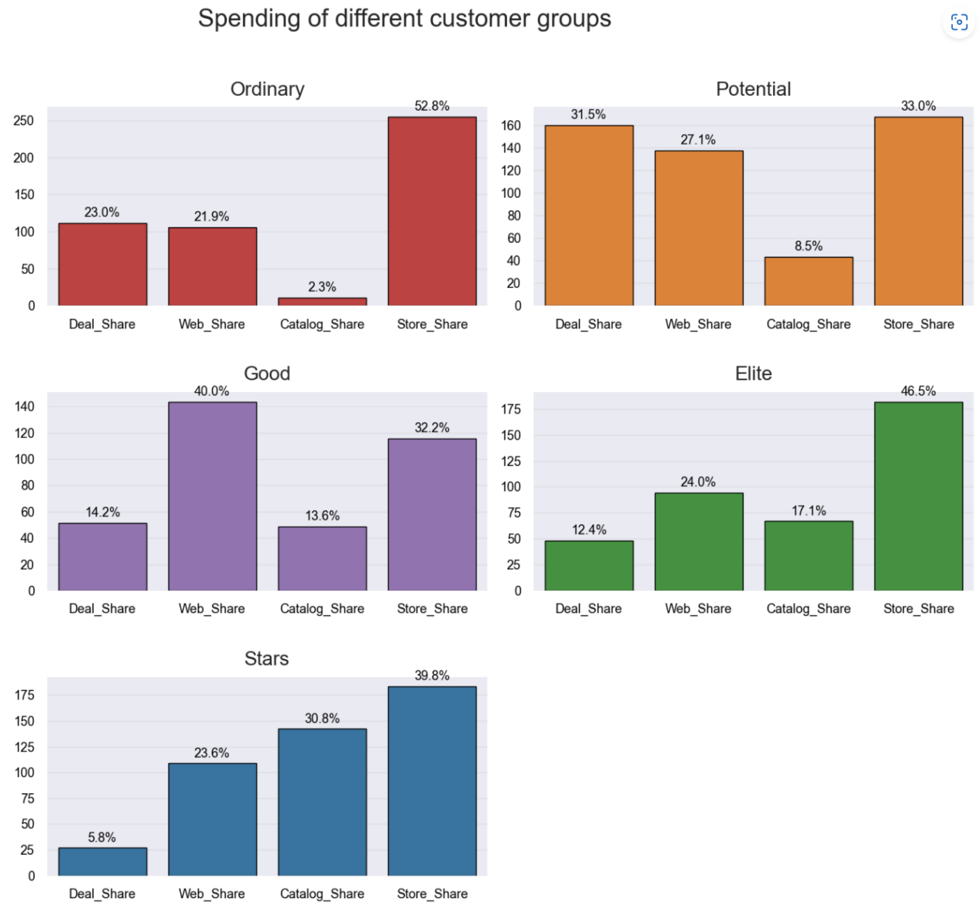


Figure 11: Shopping Platform Preference Visualisation

From the visualization, we could observe that different cluster contains different formations of platform preferences. For example, the customer group with high income and high spending prefer in person store shopping the most (39.8%), shopping using catalogue as second (30.8%), online shopping as third (23.6%), and shopping using discount as last (5.8%).

4.3 Results

From our analysis, we concluded that generally, customers prefer to shop in stores in person. We also summarized targeted customers into three groups based on the shopping platform histogram. The Ordinary and Potential customer groups could be analyzed together since their top two preference platforms are store shopping (Store_Share) and shopping with discounts (Deal_Share). We combined Good and Elite customer groups due to the fact that their top two shopping methods are online shopping (Web_Share) and store shopping (Store_Share). The last group Stars remained its own group since its top two shopping platforms are store shopping (Store_Share) and shopping using catalogues (Catalog_Share). Based on the three groups, we would provide following market strategy suggestions:

- For companies targeting lower income customers, it is recommended that they input most fund into opening more stores and providing more promotions and discounts for customers.
- For customers with average income, companies would focus on developing both physical stores and online shopping websites evenly to increase sales on both platforms.
- Lastly, for the high income customer, companies are recommended to target on attracting customers using catalogues and also opening physical stores.

5 Evaluation

5.1 Potential Disadvantages of Data set

Although we assumed the data set is randomly collected from 2240 customers, the sample size might not be enough for us to fully understand the behavior of the customers. In general, a larger sample size is preferred as it needs to provide more accurate and reliable estimates of true population parameters. For building such complex models, a larger sample size would ensure models give better accuracy.

Furthermore, the data set was collected during 2012 to 2014. During that time, it was very likely that online shopping was not as popular as it is nowadays; the majority of the population still chose to go to the store for purchases. As the development of modern technology and the impact of COVID-19 in previous years, more and more people are familiar with online shopping and would also more likely to choose it as their shopping platform nowadays.

Moreover, this was an observational study. Observation studies may lead to different types of biases, especially sample biases in this case, which can affect the accuracy and the reliability of our result.

Also, although each variable had a context explanation of what each variable was about, some contexts of the variable were still ambiguous, which could be misleading when we make our conclusion.

In addition, it might also be difficult to predict human behaviors as there are lots of uncertainties.

5.2 How we can improve

To begin with, we can increase the sample size while maintaining the quality of the new data. Moreover, we need to pay a closer attention to how we collect data. A survey with unbiased questions is preferred, and it is the best to have the survey available to all customers with some incentives no matter where their purchases were made.

6 Conclusion

Research Question 1 : In what ways can we effectively cluster our customers based on their demographic and behavioral characteristics to gain a better understanding of their needs?

We created five clusters for our customers, which are ordinary, potential, good, elite, and stars. We put customers with low spending and low income as ordinary; low spending and below average income as potential customers; low spending and average income as good customers; high spending and above average income as elite customers; high spending and high income as stars customers. And their distribution is shown in Figure 7.

Research Question 2 : How can we analyze the behavior of customers with each identified cluster and among all clusters, and what insights can we gather to improve overall customer satisfaction?

Among all five clusters, the most common way of making a purchase is directly in the store,which is about 40% , following by making a purchase in the company's website, which is about 26% to 27%. For ordinary, good, and elite customers, they tend not to make a purchase by using a catalogue; while for potential and stars customers, they tend not to make a purchase with a discount. Since purchasing directly in the store and in the company's website are the two major ways the public shop, it is recommended that businesses to pay more attention to and improve the experience of their local stores and company websites.

References

[1] Customer Personality Analysis
AKASH PATEL, 2021
`https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis`

[2] sklearn.decomposition.PCA
scikit-learn developers (BSD License), 2007 - 2023
`https://scikitlearn.org/stable/modules/generated/sklearn.decomposition.PCA.html`

[3] After doing EFA the cumulative% of variance is 49%. Is anything below 60% unacceptable?
ResearchGate GmbH, 2008-2023
`https://www.researchgate.net/post/After-doing-EFA-the-cumulative-of-variance-is-49-Is-anything-below-60-unacceptable`

[4] Understanding K-means Clustering in Machine Learning
Education Ecosystem (LEDU)), 2018
`https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1`

[5] Statistical Machine Learning Lecture 9
Chenyang Zhong, 2023
STAT GR5241 Statistical Machine Learning Lecture Notes