Course Report: Unsupervised Learning

Nan He

1 Main objective of the analysis

Acknowledgement: this is a course project for IBM Machine Learning professional certificate. The project notebook can be accessed here.

In this report, the data I will be using is a Customer Personality Analysis data set from Kaggle, the link for the data set is here. The goal of this report is to segment the customer using unsupervised learning techniques.

The target includes:

- Find clusters of customers that share similar characteristics.
- Interpret the clustering results, create portraits for each group of customers.
- Evaluate the performance of the clustering algorithms.

2 Description of the data set

The data set is a surveying result of 2140 customers for a grocery store. The feature columns are in four categories:

- People

  ID: Customer's unique identifier
  Year_Birth: Customer's birth year
  Education: Customer's education level
  Marital_Status: Customer's marital status
  Income: Customer's yearly household income
  Kidhome: Number of children in customer's household
  Teenhome: Number of teenagers in customer's household
  Dt_Customer: Date of customer's enrollment with the company
  Recency: Number of days since customer's last purchase
  Complain: 1 if the customer complained in the last 2 years, 0 otherwise

- Products

  MntWines: Amount spent on wine in last 2 years
  MntFruits: Amount spent on fruits in last 2 years
  MntMeatProducts: Amount spent on meat in last 2 years
  MntFishProducts: Amount spent on fish in last 2 years
  MntSweetProducts: Amount spent on sweets in last 2 years
  MntGoldProds: Amount spent on gold in last 2 years

- Promotion

  NumDealsPurchases: Number of purchases made with a discount
  AcceptedCmp1: 1 if customer accepted the offer in the 1st campaign, 0 otherwise
  AcceptedCmp2: 1 if customer accepted the offer in the 2nd campaign, 0 otherwise
  AcceptedCmp3: 1 if customer accepted the offer in the 3rd campaign, 0 otherwise
  AcceptedCmp4: 1 if customer accepted the offer in the 4th campaign, 0 otherwise
  AcceptedCmp5: 1 if customer accepted the offer in the 5th campaign, 0 otherwise
  Response: 1 if customer accepted the offer in the last campaign, 0 otherwise

- Place

  NumWebPurchases: Number of purchases made through the company's website
  NumCatalogPurchases: Number of purchases made using a catalogue
  NumStorePurchases: Number of purchases made directly in stores
  NumWebVisitsMonth: Number of visits to company's website in the last month

The data types are:

```
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)
```

Figure 1: Initial data types.

3 Data exploration, cleaning, and feature engineering

The data need to be cleaned. Here is the list of actions I took:

- Drop "ID", "Z_Revenue", "Z_CostContact". Since they are not related with our goal.
- Fill NaN values in "Income" with the median of income. The assumption is that a random people have the highest possibility to have incomes around the median.
- Encode "Martial Status" and "Education" to numeric variables. To reduce the number of possible values, I will only consider "Single" and "Not Single" for "Martial Status", and "Low", "Medium", "High" for "Education". All original values will be assigned into those new categories.
- Create "Customer_days" to replace the "Dt_Customer". This will be done by first convert "Dt_Customer" to pandas date-time format, then use the latest date in the data minus every

date-time values. The resulting "Customer_days" will be a new feature with value between 0 and 720 days.

The data after cleaning contains 25 features, of which one is float64, and others are all int64. Here is the correlation heat-map for all numerical variables:

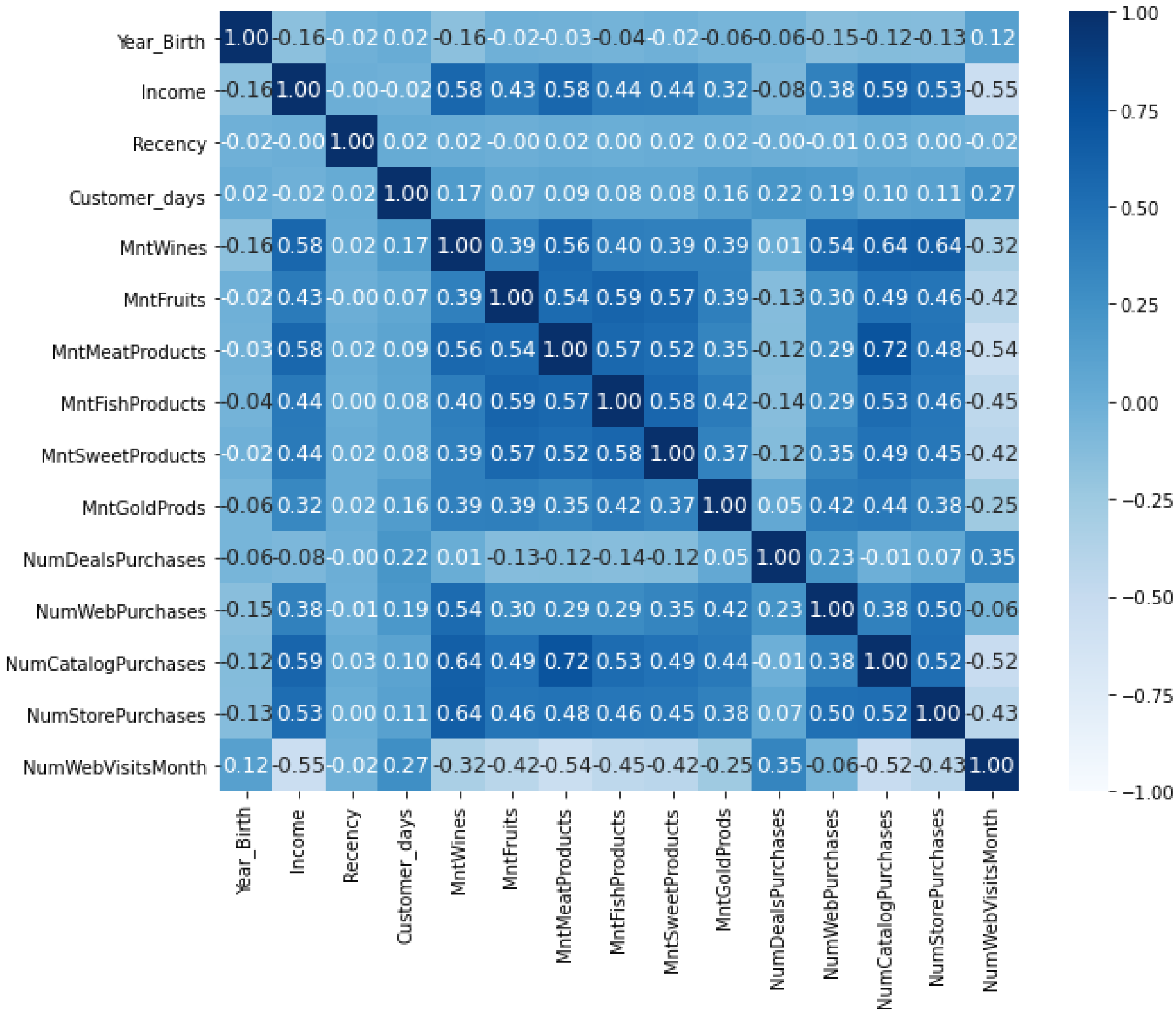

Figure 2: Correlation Heatmap.

There are some interesting observations here, for example:

- Time-related features, including "Year_Birth", "Recency", and "Customer_days", have very weak correlation with other features.
- "Income" is positively correlated with every type of consumption, but negatively correlated with "Year_Birth", "NumDealsPurchases", and "NumWebVisitsMonth".

4 Dimensionality Reduction

In this section, I will start building an unsupervised model to segment the customers. Since there are 25 features, a direct clustering will end up into curse of dimensionality. There are also many encoded categorical features, whose distance do not have actual meanings. Therefore, it is important to first perform PCA to transform the features into principle components. The steps include:

- Scale the data using RobustScalar, since there are many outliers.
- Compute and transform the features using PCA, truncated at 2 principle components.

The resulting data are shown below:

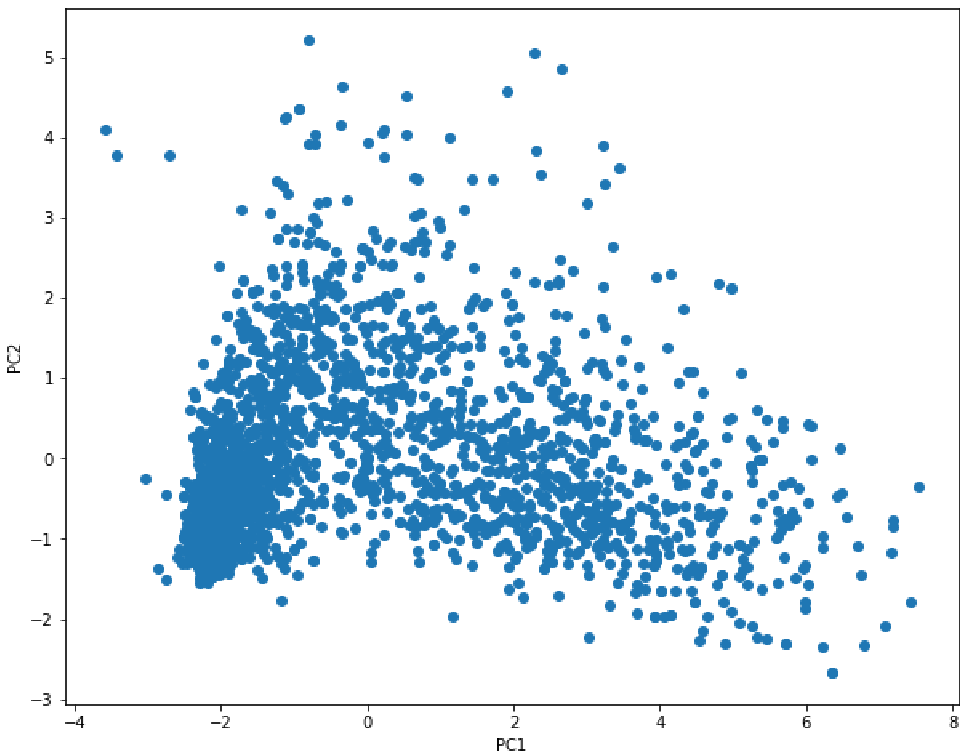

Figure 3: Data distribution after PCA.

5 Testing different clustering models

Here I will test multiple clustering models. The clustering results are plotted for every parameters below. The model will be evaluated using

1) Silhouette score: higher value indicates better defined clusters.

2) Davies-Bouldin score: lower value indicates better cluster separation:

- **Kmeans clustering**, $k = 4$ is decided using elbow method.

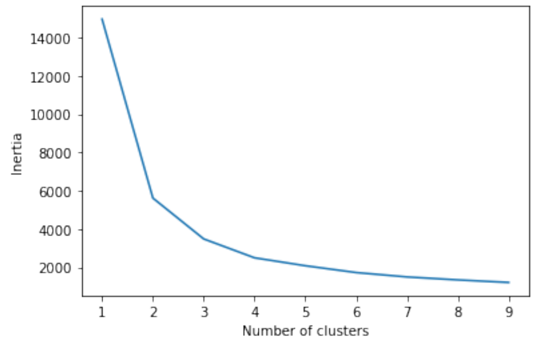

Figure 4: Kmeans elbow curve.

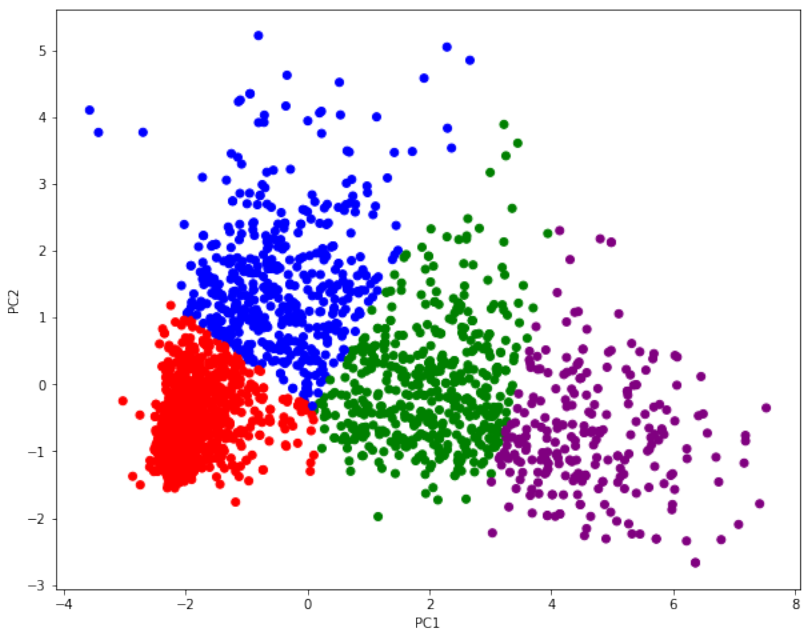

Figure 5: Kmeans clustering result.

- **Single-linkage agglomerative clustering**, using $k = 4$.

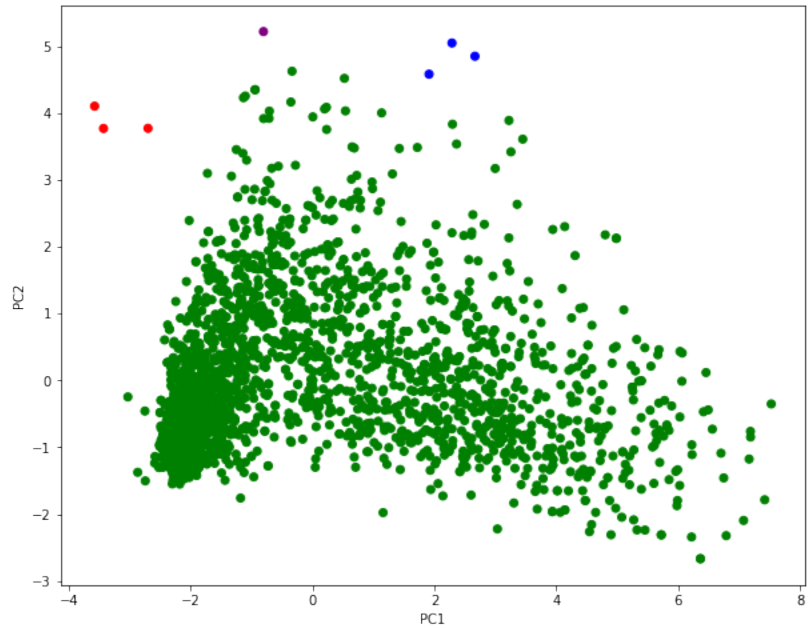

Figure 6: Single-linkage agglomerative clustering result.

- **Average-linkage agglomerative clustering**, using $k = 4$.

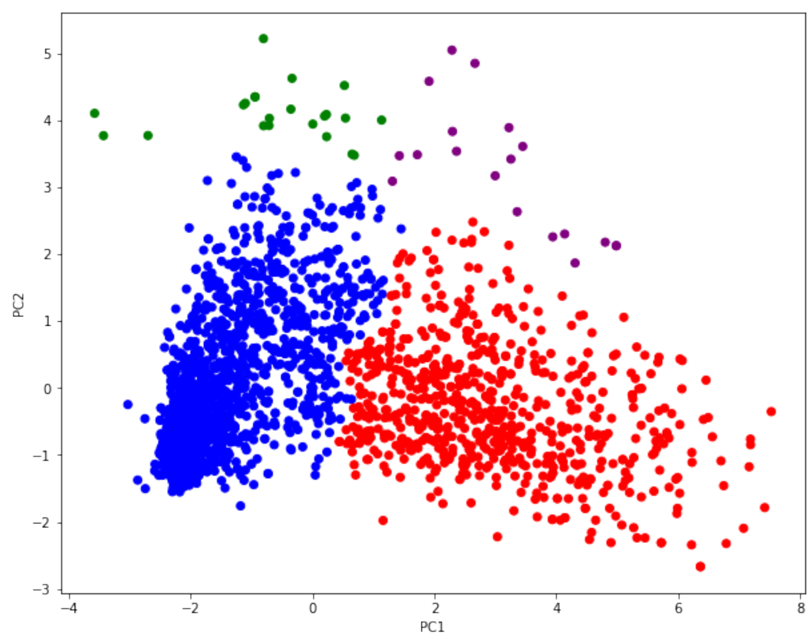

Figure 7: Average-linkage agglomerative clustering result.

- **Ward-linkage agglomerative clustering**, using $k = 4$.

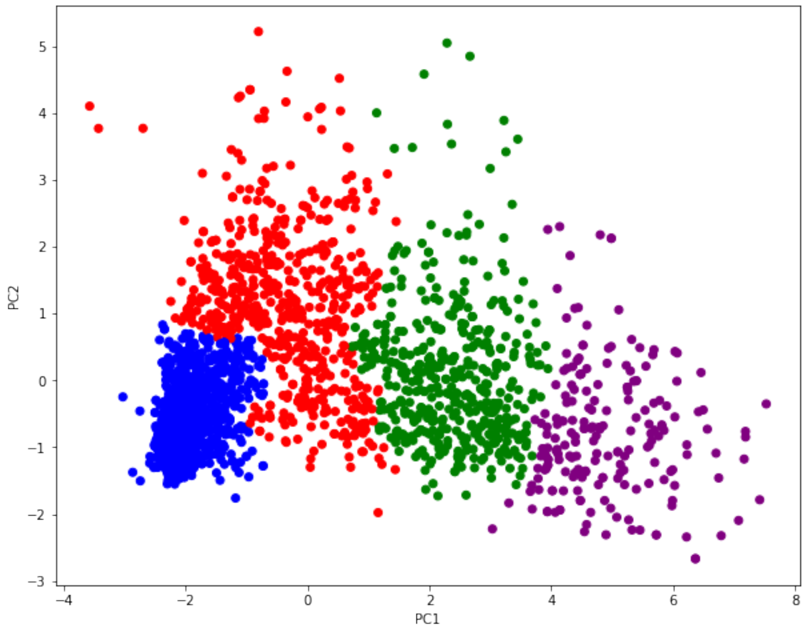

Figure 8: Ward-linkage agglomerative clustering result.

- **Complete-linkage agglomerative clustering**, using $k = 4$.

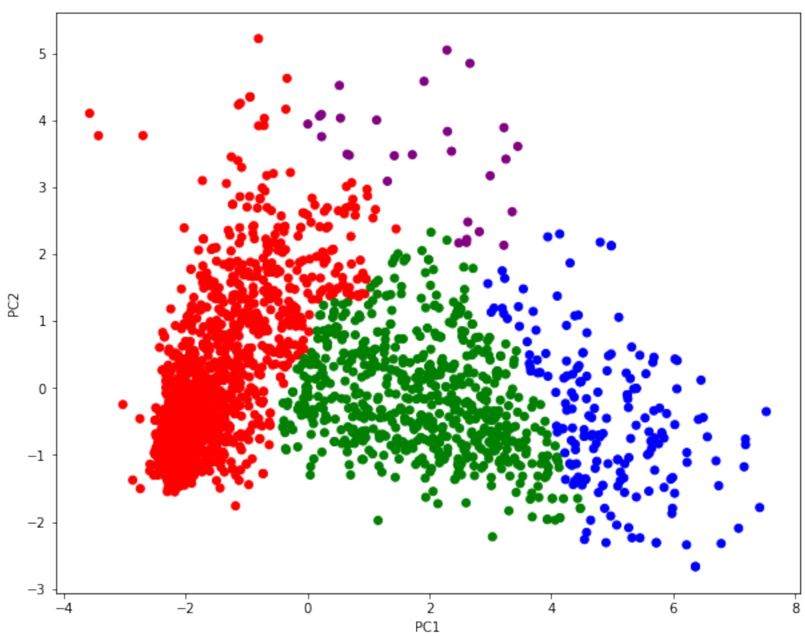

Figure 9: Complete-linkage agglomerative clustering result.

- **DBSCAN**, using $\epsilon = 0.1$.

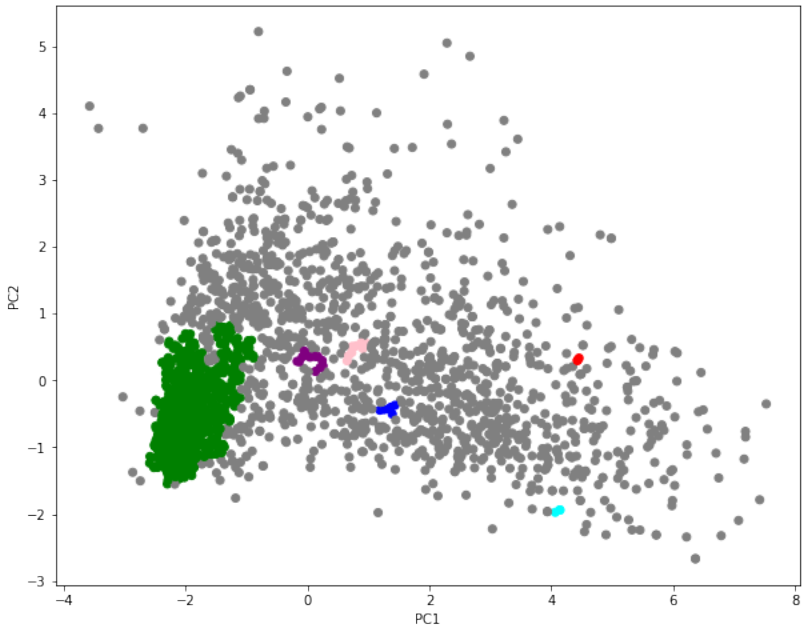

Figure 10: DBSCAN result 1, not all clusters are colored.

- **DBSCAN**, using $\epsilon = 0.5$.

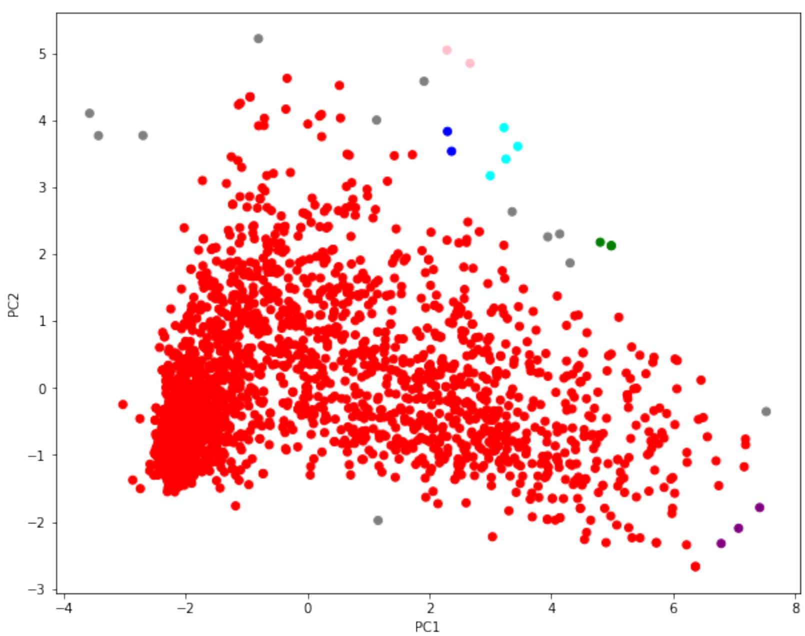

Figure 11: DBSCAN result 2, not all clusters are colored.

The scores for each model is listed below.

Method	Silhouette score	Davies-Bouldin score
Kmeans	0.497	0.786
Single-linkage	0.325	0.499
Average-linkage	0.500	0.716
Ward-linkage	0.466	0.833
Complete-linkage	0.457	0.771
DBSCAN $\epsilon = 0.1$	-0.083	1.951
DBSCAN $\epsilon = 0.5$	0.186	1.639

6 Final model choices and analysis

From the scatter plots and the scores, the observations are

- Kmeams works well in this case, creating very balanced clusters.
- Single-linkage agglomerative clustering provide the best separation between clusters (lowest Davies-Bouldin score), however, its clusters are very dispersed and ill-defined.
- In my case, DBSCAN struggles to find correct clusters no matter what ϵ I chose. One possible reason is that the data itself has various densities for different regions, which plaguing density-based methods.
- The best performance comes from Average-linkage agglomerative clustering, which will be the model I use for the following analysis.

7 Key Findings and Insights

Let's dive into the results of Average-linkage agglomerative clustering, and see what this segmentation tells us. I put the predicted labels back to the dataframe, and investigate the distribution of different features among clusters.

- Income.

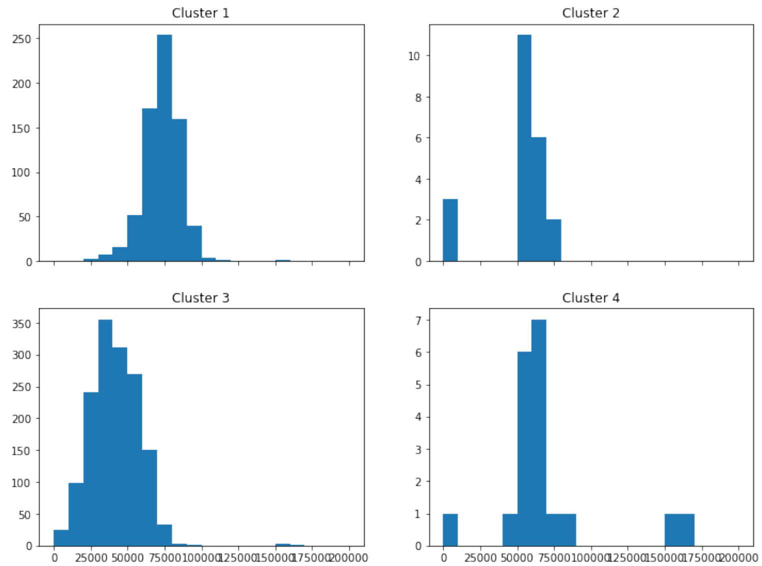

Figure 12: Income of different clusters.

Since only cluster 1 and 3 have significant number of labels, in the following plot, I will focus only on cluster 1 and 3.

- Education.

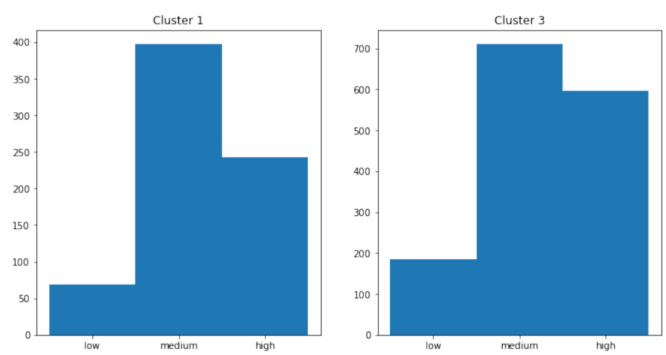

Figure 13: Education level of different clusters.

- Marital Status.

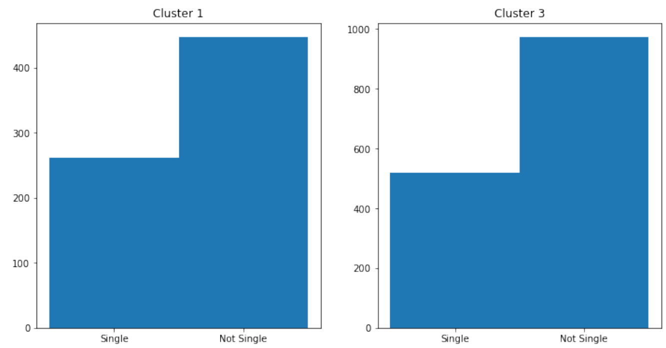

Figure 14: Marital Status of different clusters.

- Year of birth.

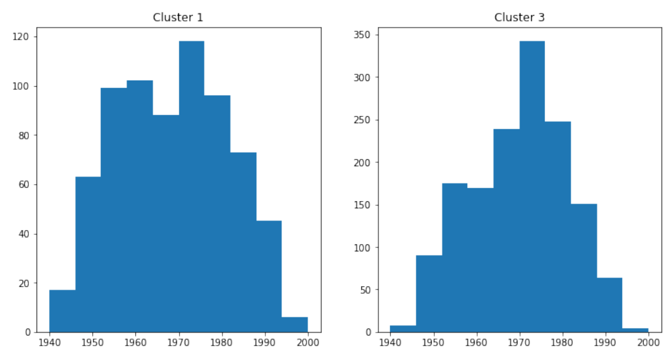

Figure 15: Birth year of different clusters.

- Meat purchases.

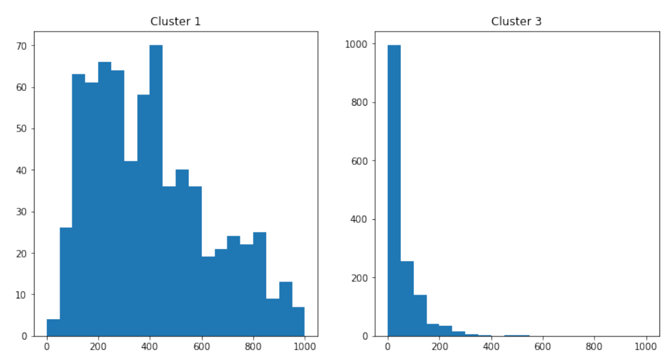

Figure 16: Meat purchases of different clusters.

• Wine purchases.

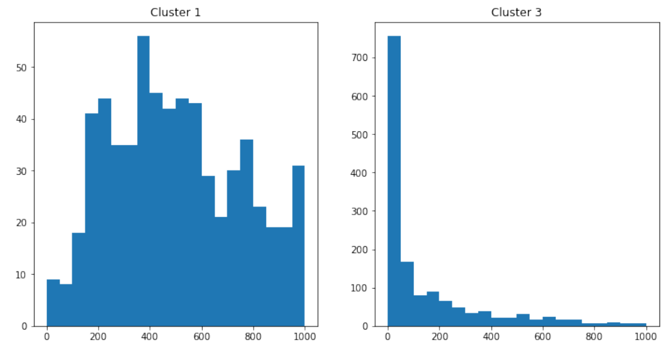

Figure 17: Wine purchases of different clusters.

• Fruit purchases.

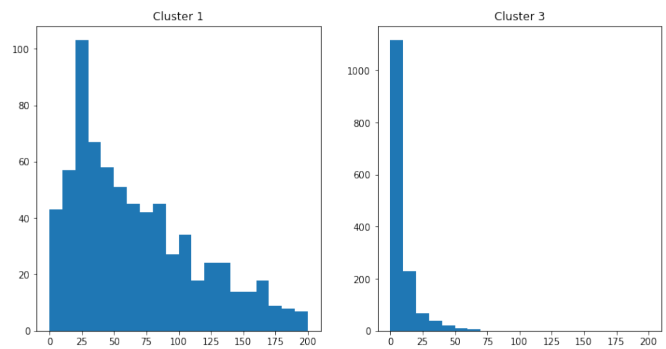

Figure 18: Fruit purchases of different clusters.

• Number of purchases using deals.

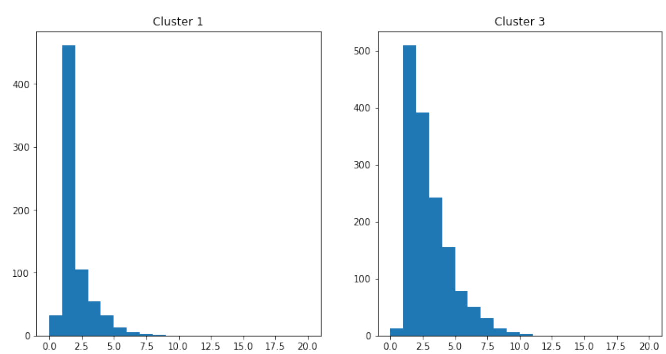

Figure 19: Number of purchases using deals for different clusters.

- Number of purchases through web.

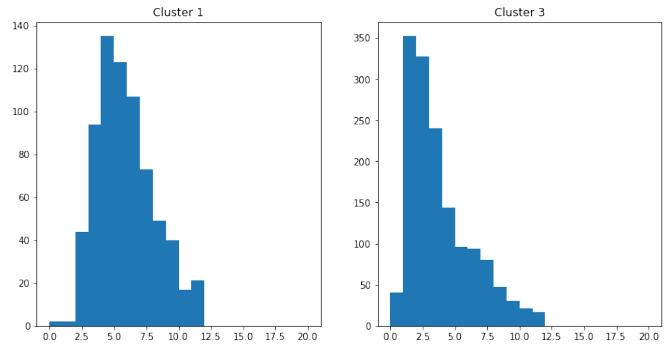

Figure 20: Number of purchases through web for different clusters.

- Number of purchases at store.

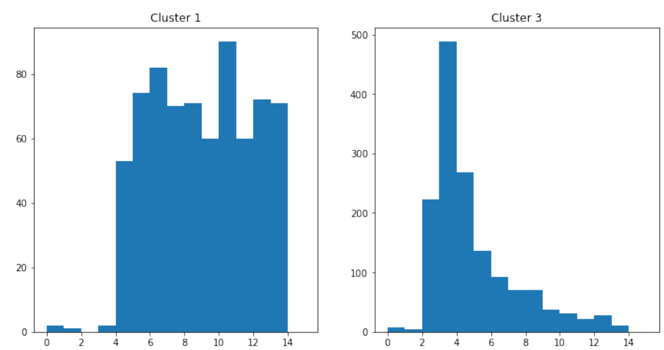

Figure 21: Number of purchases at store for different clusters.

- Number of website visits.

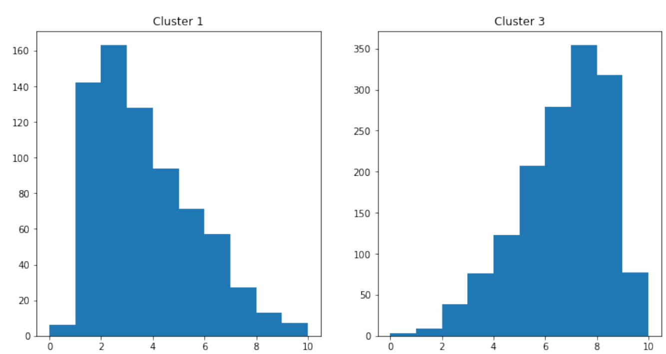

Figure 22: Number of website visits for different clusters.

From those histograms, I find that most customers (98%) fall into two categories, with less than 2% being outliers. The portrait of those two type of customers are:

- Cluster 1: those customers have higher average income, older (in average), purchase more meat, wine, fruits, use fewer deals, and prefer shopping at store.
- Cluster 3: those customers have lower average income, younger, purchase fewer goods, tend to use more deals, prefer shopping through website, and visit the website more frequently.
- Other features are not significantly different among clusters.

8 Summary and suggestions for next steps

In summary, this report perform a customer segmentation for a survey data set. By using PCA and clustering algorithms, I find two major customer types. Those insights will be helpful for the store to target there customers.

One fallacy is that in my report, for simplicity, I truncate the features to only 2 principle components. This action may cause large residuals, where many other contributing factors are ignored. For example, in my clustering, it seems that Education have no impact on the segmentation, which is non-intuitive.

The next steps may be:

- Use different number of principle components and compute the residuals of the decomposition to find a best balance.
- Try kernel PCA and other dimensionality reduction techniques.
- Try to find more balanced segmentations, like 4 medium-sized clusters, which will provide us more information on customer personality.