Understanding Customer Segmentation :

Sales and Marketing Strategy

Capstone Project:
Unsupervised Learning

MIT / Great Learning/ Karla M. Rodriguez

Contents / Agenda

1. Executive summary.
2. Business Problem.
3. Data Overview.
4. Statistics.
5. Multivariate Analysis.
6. PCA.
7. T-SNE Clustering.
8. K-Means Clustering.
9. K-Medoids Clustering.
10. Conclusion and Recommendations.
11. Appendix.

Executive summary

- To understand customer behavior and characteristics to achieve better marketing
- After implementing several models: K-medoids model by measuring distance to a similar data.
- **Based on the data, We can identify three distinct segments with unique characteristics and behaviors.**

Understand Behavior:

Segment 1: **Highest recency among groups.** second in purchasing power.
Segment 2: **Highest Average income.** Spend the most in products.
Segment 3: **Average spending.** Third purchasing power.

Effective Marketing:

Segment 1: Email and Web marketing.
Segment 2: In store product demonstrations.
Segment 3: Promote promotional deals.

Personalized Communication:

Segment 1: Highest to use web.
Segment 2: Highest response to marketing campaigns.
Segment 3: Highest response promotional deals.

Business Problem

- The objective of this analysis is to employ unsupervised learning techniques to identify optimal customer segments based on the provided data.

- By applying dimensionality reduction and clustering techniques we can effectively identify and eliminate redundant customer features.

- We can then identify valuable insight by tailoring to their specific needs and improve customer engagement with effective marketing.

Data Overview: Numbers... Numbers

- We have 2240 entries in this dataset.
- The customer data includes information such as demographics, purchasing preferences , response to campaigns & complains.
- There were some missing values that were replaced with average values.
- **Transformed data to numbers(integers and floats) to check for relevant features.**

```
customer.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2240 entries, 0 to 2239
Data columns (total 27 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   ID                   2240 non-null   int64
 1   Year_Birth           2240 non-null   int64
 2   Education            2240 non-null   int64
 3   Marital_Status       2240 non-null   float64
 4   Income               2240 non-null   float64
 5   Kidhome              2240 non-null   int64
 6   Teenhome             2240 non-null   int64
 7   Dt_Customer          2240 non-null   float64
 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  Response             2240 non-null   int64
dtypes: float64(3), int64(24)
memory usage: 490.0 KB
```

Statistics

Demographics:

- **Year_birth:** Customers are in **average 54**
- **Income:** median :51,381. Better to consider the **median as its affected by outliers.**
- Kids: min: 0 max: 2 med: 0 most customer with **no children at home.**
- Recency: Average is higher indicating more consistency and predictability

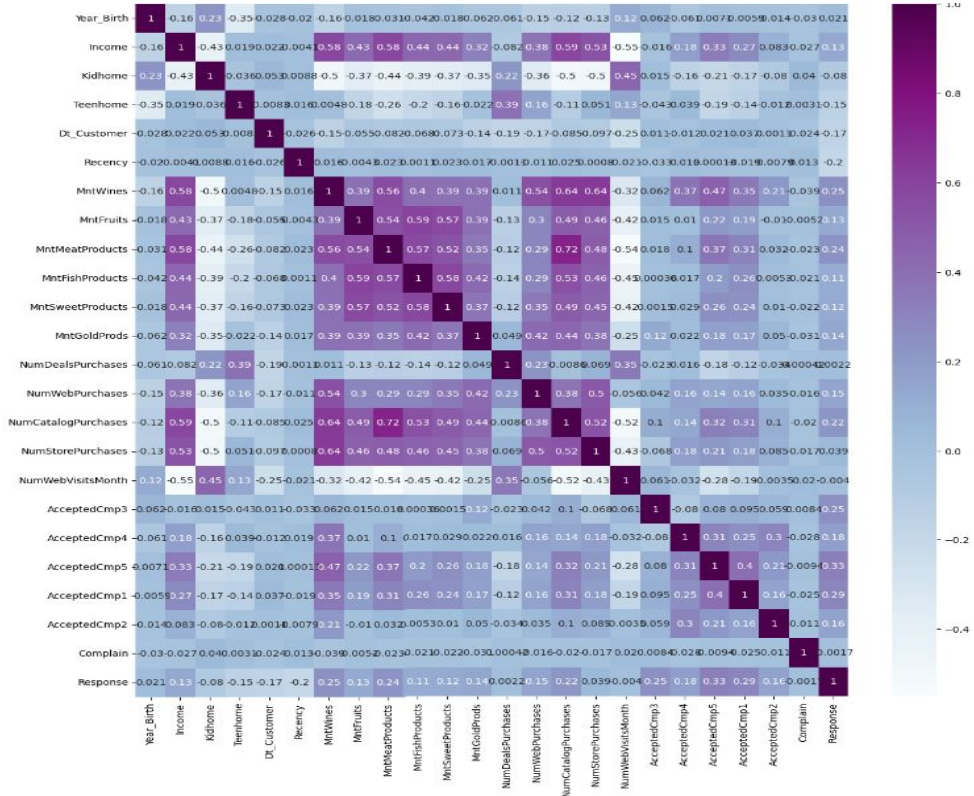
Products: Great Variability Across.

- Wines: max: 1,493, med: 173.50, mean: 303. Close average
- Fruits: amount spent is, max:199, med: 8, mean: 26, std: 39.
- Meat: amount spent is, max: 1,725 median: 67. **Greater variability, wider range of price diversity.**
- Fish: The amount spent is, max:259, med:12, mean:37, std:54. **Similar values as meat**
- Gold products: The amount spent is maximum:362, med:24, mean:44.

Behavior:

- Deals Purchases: max:15, med:2, mean:2.3,,std: 1.93, surprising is not more
- Web Purchases: max:27,
- med:4, mean:4, std:2.77, I would have thought it sold more!
- Catalog Purchases: The number of is, max:28 med:2 mean:2.66 std:2.9. Similar to web
- Store Purchases: The number of is max:13, med:5, mean:5.9, std:3.2. **Higher mean can be influenced by outliers**
- Web visits: The number of is, max:20 med:6 mean:5.3 std: 2.4. Consistent and predictable range.
- Surprisingly the response to the campaigns have been not successful, 0 everywhere max of 1 in 2240 entries!

Multivariate Analysis



Using heatmap as a guide to highlight the strongest correlated values might give us a stronger sense of the most important features. Listed below:

- **Catalog and meat** has a high correlation number of **(0.72)**
- **Wines and store purchases** has a medium high positive correlation **(0.64)**
- **Wines and catalog** also a medium high positive correlation **(0.64)**
- **Fruits and fish moderate high (0.59)**
- **Income and catalog moderate high(0.59)**
- Kid and web visits a month also moderate(0.45)
- **Also interesting to note the campaign 1 and 5 have a low to moderate range (-.001-0.47)**

PCA: Reduced to 80% of data 9 PC groups maximum variance.

This model helped identifying important features.

PCA Observations: positive correlation & The magnitude of the value indicates the strength of the correlation.

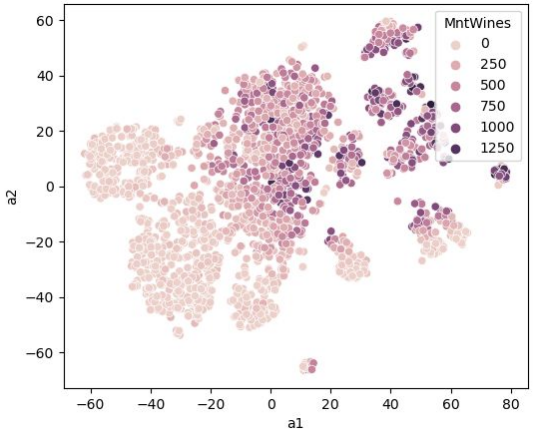
Looking at the PCs from 1-9. The positive and negative correlation breakdown as follows:

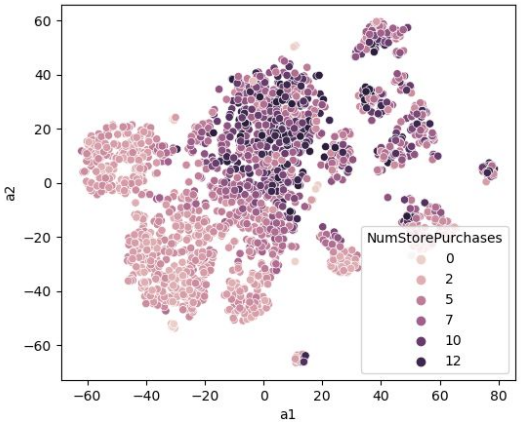
- **Demographics: Income(0.30), date enrolled(0.38) and marital status(0.64)**
- Purchasing Products: amount sold: wines(0.32), fruits(0.28), meat(0.32),fish(0.29),and sweets(0.28)
- **Purchasing platform: catalog purchases(0.33), store purchases(0.29), deal purchases**
- Response: PC2 web visits(0.32), **Accepted campaign 4(0.4), Accepted campaign 5(0.32) have the highest numbers.**
- **negative values in gray indicate a negative correlation.**

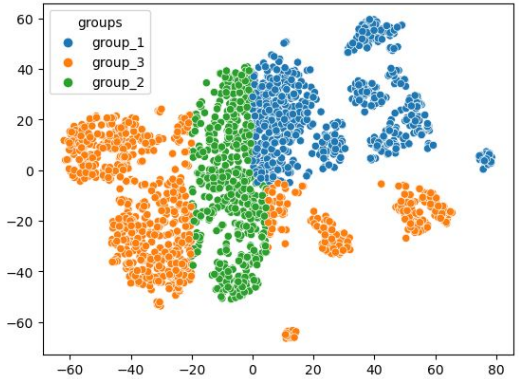
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Year_Birth	0.060000	0.000000	-0.120000	-0.360000	0.580000	-0.010000	0.140000	0.140000	-0.280000
Marital_Status	0.020000	0.040000	-0.010000	0.060000	0.640000	0.340000	-0.470000	-0.240000	0.130000
Income	0.300000	-0.050000	0.040000	-0.170000	0.080000	-0.100000	0.070000	0.040000	-0.080000
Dt_Customer	-0.040000	-0.200000	0.380000	-0.210000	0.210000	-0.250000	0.110000	0.350000	0.220000
Recency	0.000000	-0.100000	-0.050000	-0.240000	-0.250000	0.730000	-0.210000	0.390000	-0.180000
MntWines	0.320000	0.220000	-0.070000	-0.160000	-0.030000	-0.010000	0.030000	0.050000	-0.050000
MntFruits	0.280000	-0.210000	-0.040000	0.170000	-0.050000	0.030000	-0.040000	-0.120000	0.140000
MntMeatProducts	0.320000	-0.090000	0.060000	0.080000	-0.050000	0.040000	-0.000000	-0.070000	-0.170000
MntFishProducts	0.290000	-0.210000	-0.030000	0.150000	-0.040000	0.040000	-0.070000	-0.110000	0.110000
MntSweetProducts	0.280000	-0.180000	-0.030000	0.130000	-0.090000	0.050000	-0.070000	-0.110000	0.090000
MntGoldProds	0.230000	-0.010000	-0.220000	0.170000	0.030000	0.050000	-0.050000	0.200000	0.240000
NumDealsPurchases	-0.040000	0.210000	-0.510000	-0.080000	-0.010000	-0.130000	0.060000	0.090000	-0.060000
NumWebPurchases	0.220000	0.160000	-0.390000	-0.070000	0.030000	-0.110000	0.050000	0.120000	0.070000
NumCatalogPurchases	0.330000	-0.020000	-0.030000	0.030000	0.030000	0.020000	0.040000	0.100000	-0.080000
NumStorePurchases	0.290000	-0.040000	-0.200000	-0.200000	-0.030000	-0.100000	0.040000	-0.020000	0.150000
NumWebVisitsMonth	-0.250000	0.320000	-0.270000	0.120000	-0.120000	0.050000	-0.040000	-0.040000	0.000000
AcceptedCmp3	0.020000	0.180000	0.070000	0.470000	0.160000	0.060000	0.150000	0.630000	0.090000
AcceptedCmp4	0.100000	0.400000	0.140000	-0.380000	-0.110000	0.020000	-0.070000	-0.150000	0.050000
AcceptedCmp5	0.200000	0.250000	0.320000	-0.010000	-0.100000	0.070000	0.030000	-0.020000	-0.210000
AcceptedCmp1	0.180000	0.230000	0.270000	0.050000	-0.100000	0.010000	0.030000	0.040000	-0.260000
AcceptedCmp2	0.060000	0.350000	0.180000	-0.170000	-0.050000	0.100000	-0.070000	0.030000	0.680000
Complain	-0.010000	-0.010000	-0.020000	0.020000	0.120000	0.460000	0.800000	-0.270000	0.150000
Response	0.120000	0.400000	0.120000	0.380000	0.190000	-0.030000	-0.010000	-0.150000	-0.200000

T-SNE: preserves relationship and reveals patterns.

Observations: 3 groups. Similar cluster. Different intensity.





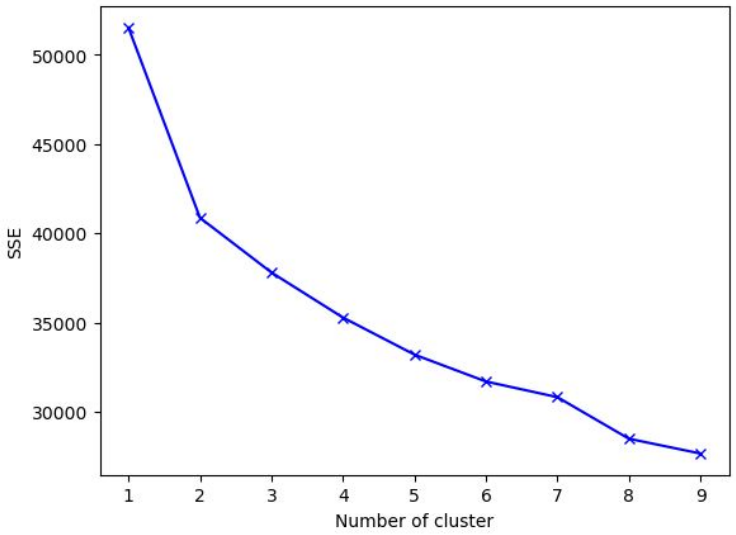


Cluster can help identify distinct clusters or groups within the data that **share similar characteristics or patterns. Looking at the 3 groups from 8 data points identified**.

- **Demographics: Birth & marital status boxplots imply similar tendency of distribution with a few outliers** for group 3 Income groups are also similar with very few outliers. **Date enrolled group 1,2 are similar whereas group 3 is high on the opposite scale.**
- **Purchasing Products:** amount sold: **wines,fruits, meat & fish products group 1 consumes the most followed by group 2 and on the third group 3 with many outliers.**
- **Response:** Recency is the same across all three groups with a similar median and no outliers.

K-Means Clustering: aims to partition data.

Two groups emerge



K-means Observations:

- **Income:** Income appears to have a significant difference between group 0 and group 1, with group 1 having a much higher mean and median income. **This feature could be relevant in understanding the purchasing power and financial status of the two groups.**
- **Wines, Fruits, Meat Products, Fish Products, Sweet Products, and Gold Products:** These features represent the amount spent on various product categories. Group 1 has substantially higher mean and median values for all of these categories compared to group 0. **These features can be relevant in understanding the differential spending patterns and preferences of the two groups.**
- **Web Purchases, Catalog Purchases, and Store Purchases:** These features indicate the number of purchases made through different channels. Group 1 has higher mean and median values for all three categories, suggesting a higher overall purchase activity. These features **can be relevant in understanding the shopping behavior of the two groups.**
- Accepted campaigns 1-5 : These features represent whether the customers accepted specific marketing campaigns. **Group 1 has higher mean and median values for all these features, indicating a higher response rate to the marketing campaigns.**

K-Medoids Clustering: Clouds parted. 3 distinct groups emerge. The best model.

KMed_segments	0	1	2
Year_Birth	54.567715	47.962590	59.179361
Marital_Status	2.198358	2.106475	2.140049
Income	74249.363201	30283.575540	51215.842138
Dt_Customer	2012.978112	2013.080576	2013.028256
Recency	49.235294	42.856115	54.335381
MntWines	603.521204	21.810072	275.778870
MntFruits	63.748290	4.221583	11.527027
MntMeatProducts	405.928865	16.451799	80.835381
MntFishProducts	90.239398	6.359712	16.796069
MntSweetProducts	65.398085	4.345324	12.033170
MntGoldProds	83.002736	12.575540	35.864865
NumDealsPurchases	1.677155	1.794245	3.359951
NumWebPurchases	5.592339	1.785612	4.694103
NumCatalogPurchases	5.642955	0.328058	1.977887
NumStorePurchases	8.566347	2.860432	5.798526
NumWebVisitsMonth	3.303694	6.968345	5.713759
AcceptedCmp3	0.082079	0.082014	0.056511
AcceptedCmp4	0.114911	0.002878	0.099509
AcceptedCmp5	0.210670	0.000000	0.011057
AcceptedCmp1	0.170999	0.001439	0.022113
AcceptedCmp2	0.027360	0.002878	0.008600
Complain	0.006840	0.011511	0.009828
Response	0.240766	0.107914	0.101966
KMeans_Labels	0.751026	1.998561	1.669533
count_in_each_segment	731.000000	695.000000	814.000000

Observations:

- **Year_Birth: Segment 0 consists of older customers compared to the other segments.**
- Marital_Status: The distribution of marital statuses is relatively balanced across all segments.
- **Income: Segment 0 has the highest average income**, indicating a difference in purchasing power among the segments.
- **Date of enrollment: Segment 1 has more recent customers** compared to the other segments.
- **Recency: Segment 1 has the most recent interactions** or purchases on average.
- **Wines: Segment 0 spends the most on average for wines, followed by segment 2 and segment 1.**
- Fruits, Meat Products, Fish Products, Sweet Products, Gold Products: Segment 0 generally spends the most on average for these product categories.
- Deals Purchases: Segment 2 has the highest average number of purchases made through promotional deals.
- Web Purchases, Num Catalog Purchases, Num Store Purchases: Segment 0 has the highest average number of purchases across all channels.
- **Web visits a month: Segment 1 has the highest average number of monthly web visits, indicating higher online engagement**.
- All accepted campaigns: The acceptance rates of marketing campaigns vary among the segments.
- **Complain: Segment 1 has the highest average complaint rate.**
- **Response: Segment 0 has the highest average response rate to marketing campaigns.**
- **KMeans_Labels: Segments 0 and 2 have a higher proportion of customers labeled as 1 or 2, while segment 1 has a higher proportion of customers labeled as 0.**
- Count in each segment: Segment 2 has the highest number of customers, followed by segment 0 and segment 1.

Conclusion and Recommendations

Actionable Insights and Recommendation

- We can identify three distinct segments with unique characteristics and behaviors.
- By understanding the segment interest, we can advise on:
- ATL(tv, radio, print ads, billboards)
- BTL(direct mail, email, marketing)
- Engagement through personalized communications and targeted campaigns.

Conclusion and Recommendations

Segment 1:

- In this segment, customers are typically older and do not have children at home.
- They exhibit a median purchasing power compared to the other segments
- highest recency.
- While their wine spending is third highest among the segments,
- **they show the highest average number of web visits per month, suggesting higher online engagement.**
- **Interestingly, Segment 1 also has the highest average complaint rate.**
- In terms of product preferences, spends the most on meat products. **Email and web marketing might be the best way to approach this segment. Preferred customer service treatment.**

Conclusion and Recommendations

Segment 2:

Comprises older customers, as indicated by the highest average birth year.

Their higher income allows them to indulge in wine, fish, fruits, and sweets.

They exhibit a relatively balanced distribution of marital statuses.

With the highest average income

This segment demonstrates a greater purchasing power.

Spends the most on wines and has the highest average spending on various product categories such as fruits, meat products, fish products, sweet products, and gold products.

They make the highest average number of purchases across all channels.

Additionally, Segment 2 shows the highest average response rate to marketing campaigns.

Conclusion and Recommendations

Actionable Insights and Recommendation

- we can identify three distinct segments with unique characteristics and behaviors.
- By understanding the segment interest, we can advise on: ATL(tv, radio, print ads, billboards)
- BTL(direct mail, email, marketing)
- Engagement through personalized communications and targeted campaigns.

Segment 3:

Customers have a relatively balanced distribution of marital statuses.

However, they exhibit a lower purchasing power compared to the other segments.

They rank second in wine spending

Higher response to promotional deals, as indicated by the highest average number of deal purchases.

Appeal to sales, discounts and deals might be a good idea to approach this group as they respond high to promotional deals.

Appendix

- MIT/ Great Learning/ with the collaboration of student Karla M. Rodriguez
- **Practical statistics for data scientist** book by Peter and Andrew Bruce & Peter Gedeck. Chapter 7 (unsupervised learning)
- PN wing logo: marketing logo.
- Vision edge: marketing entrepreneur logo .
- Freepik: Thank you logo.