

CUSTOMER PERSONALITY ANALYSIS

ABSTRACT:

Based on detailed data of company A' s customers, this report analysis and predict **customer response** to promotion campaigns, **customer preference** to wines and meat products and **amount of customer spent**. According the prediction, the report provides a **recommendation system** to realize personalized recommendation and promotion. Meanwhile, **customer cluster** has been done to help company A segment their customers into **3** groups and corresponding suggestions are also provided.

CONTENTS:

- 1.Introduction
- 2.Data cleaning and processing
- 3.Descriptive statistic
- 4.Part1: consumer behavior prediction
 - 4.1 generalized logistic regression
 - 4.2 regression tree
 - 4.3 random forest
 - 4.4 model comparison
- 5.Part2: consumer preference prediction
 - 5.1 prediction on wines
 - 5.2 prediction on meat
 - 5.3 prediction on total spend
- 6.Recommendation system based on the prediction
- 7.Part3: customer clustering
 - 7.1 K-means clustering
 - 7.2 market segment and suggestion
- 8.Appendix

1.INTRODUCTION

DATA BACKGROUND:

Resource link: <https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis>

Problem statement:

This dataset is a detailed analysis of a Company A's customers. It helps a business to better understand its customers and makes it easier for them to modify products according to the specific needs, behaviors and concerns of different types of customers. It collected **29** variables of **2240** customers, the attributes of the variables are attached to the appendix. There are 3 problems set to help Company A achieve better promotion and recommendation performance.

Problem1: How customers response to the promotion campaign? Is the campaign efficient? (Part 1)

Problem2: What's customers' preference? Which product should Company A recommend to them? (Part 2)

Problem3: How to segment these customer? (Part 3)

2.DATA CLEANING AND PROCESSING

DATA CLEANING

DATA READING AND REMOVE NA: remove 24 rows with missing data

REDESIGN VARIABLES:

Age: change Year_Birth to age

Dt_Customer: change to days the customer has been to the company

Marital_Status: 8 levels to 2 levels; combine 'Absurd Alone Divorced Single Widow YOLO' to 1; combine 'Married. Together' to 2

Education: 5 levels to 4 levels; combine '2n Cycle' & 'basic' to 'undergraduate'

ADD VARIABLES:

ttlspend: total money spent on the 6 products

ttlnum: total numbers of purchase the customer made

child: = kid home + teen home

familymember: = child + Marital_Status

CREATE DUMMY VARIABLES:

Education : to Education_Master; Education_PhD; Education_Undergraduate

REMOVE VARIABLES:

Remove: ID, Year_Birth, Kidhome, Teenhome, Z_CostContact, Z_Revenue (figure 1)

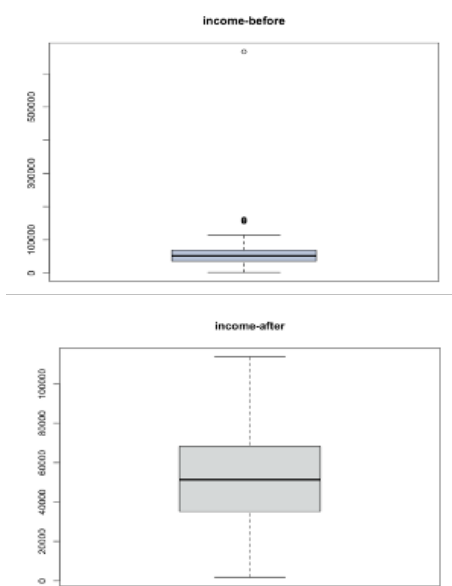
After data cleaning, we now have data of **2216** rows and **30** variables.

DATA PROCESSING:

REMOVE OUTLINERS:

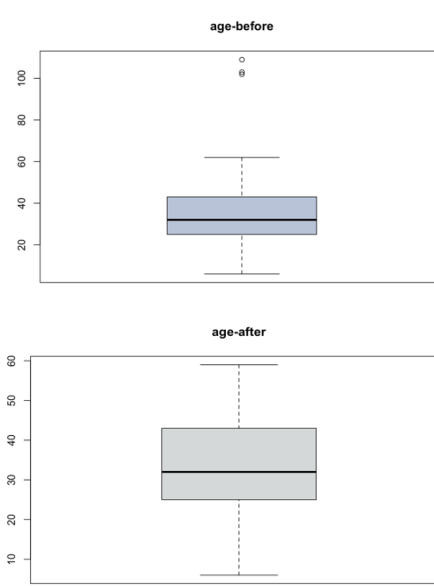
Income:

remove those whose income $\geq 15,000$



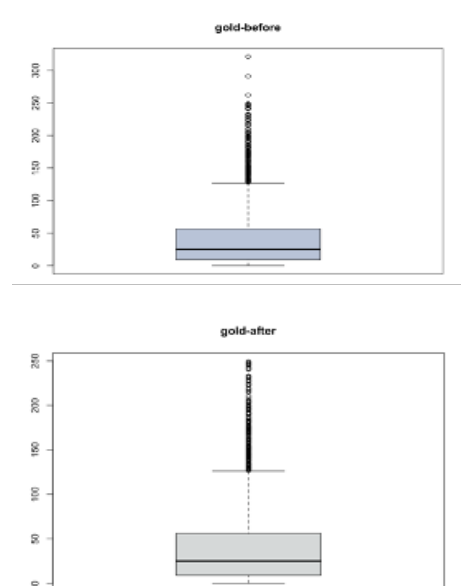
Age:

remove those whose age ≥ 100



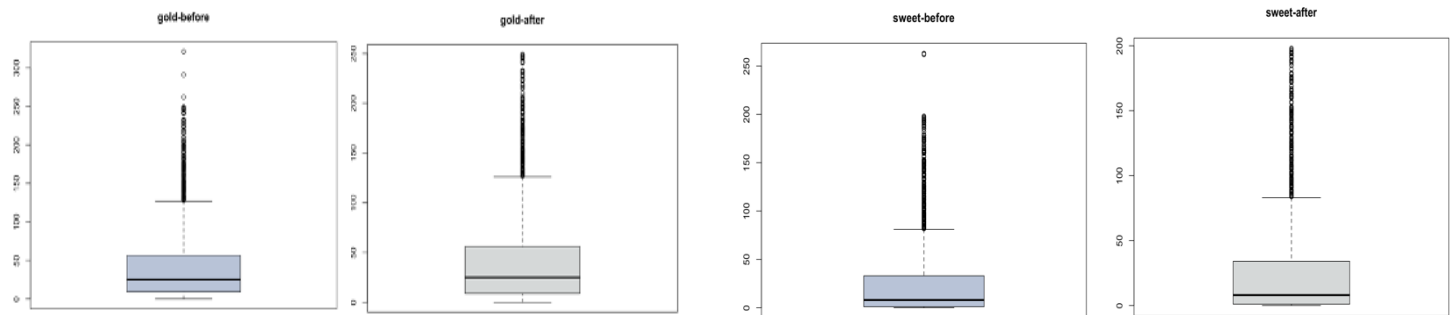
Gold:

remove 'MntGoldProds' ≥ 250



Gold: remove those whose 'MntGoldProds' ≥ 250

Sweat: remove those whose 'MntSweetProducts' ≥ 200



After data processing, we now have data of 2198 rows and 30 variables, which can be divided into **3** groups to help our further analysis.

1. Demographic (11) :

age, Education(3 dummies), Marital_Status, Income, child, family, Dt_Customer, Recency, complain

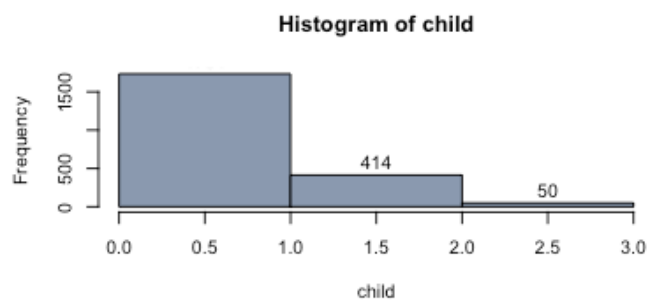
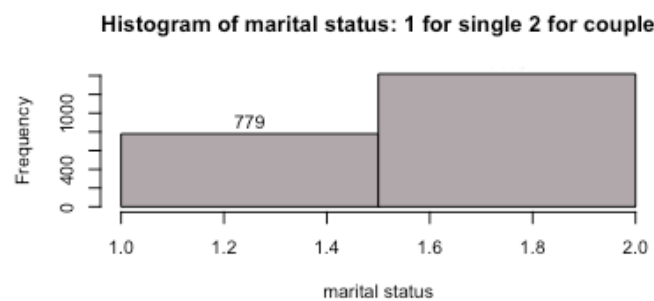
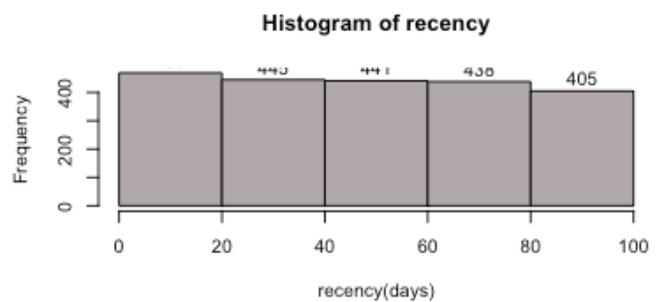
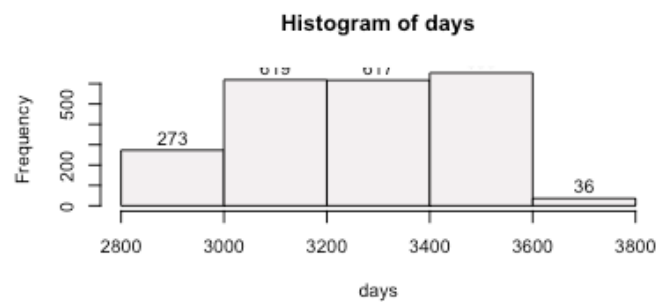
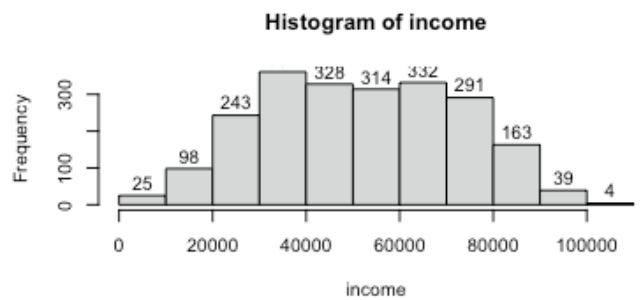
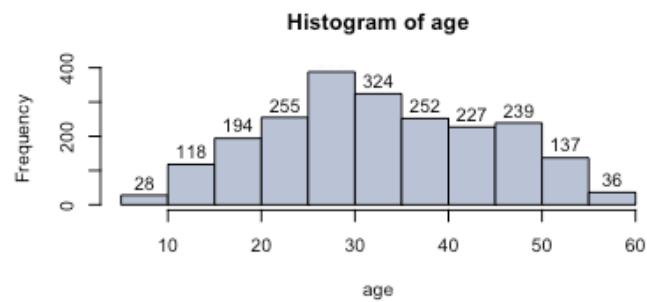
2. Behaviors :

- Products (7): MntWine, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds, ttlspend
- Promotion (7): NumDealsPurchases, AcceptedCmp1, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5, Response

3. Distribution (5): NumWebPurchases, NumCatalogPurchases, NumStorePurchases, NumWebVisitsMonth, ttlNum

3.DESCRPTIVE STATISTIC

Group1 : Demographic



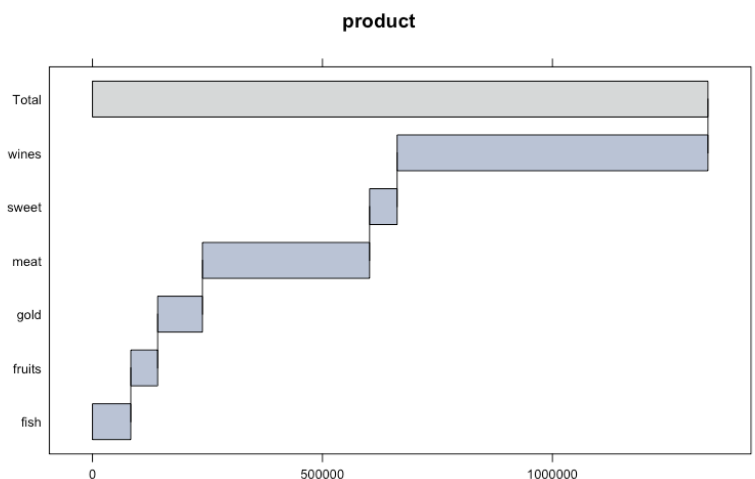
Group2.a: Behaviors--Product

We divide 6 products into 2 groups

according to the purchase amount:

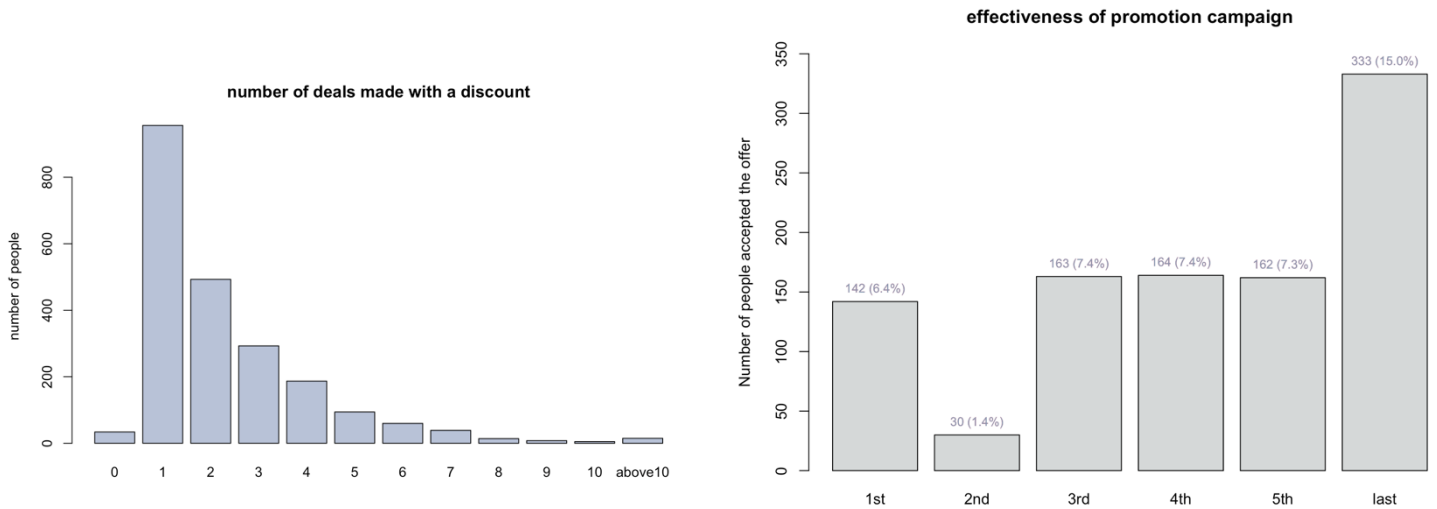
ace product: wines & meat

normal product: sweet & gold & fruits & fish

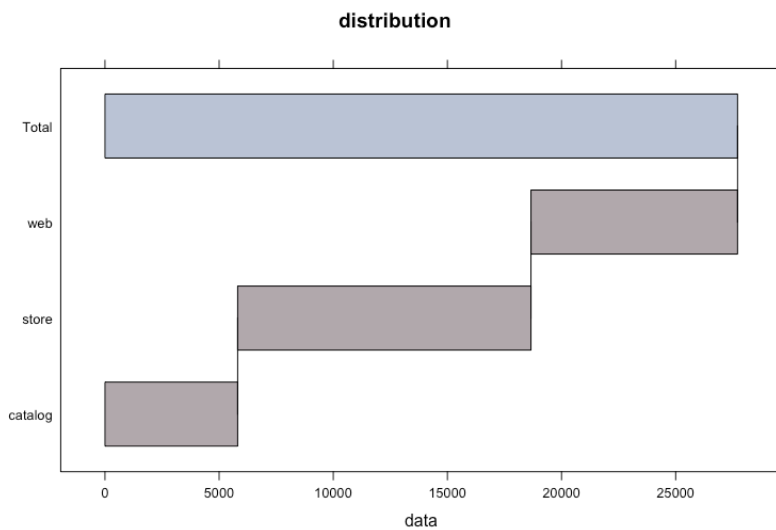


Group2.b: Behaviors—Promotion

Percentage: numbers of those who accepted the offer / total amount



Group 3: Distribution



The traditional **store** distribution is the most contributed one, followed by websites and catalog.

4.CONSUMER BEHAVIOR PREDICTION

We want to use data analysis to predict whether the customer will accept the offer when the promotion campaign is made (**Response-dependent variables**) , thus to help decide whether we should give the customer a discount or do the promotion.

We randomly select 30% of the data to be the test set, and the rest 70% to be the training set.

Key features used:

1. Accuracy: % correct prediction
2. Sensitivity: true positive rate
3. AIC: Akaike Information Criterion
4. Specificity: rue negative rate

GENERALIZED LINEAR REGRESSION

1. FULL MODEL:

Formula: Response ~ .

Summary of the regression result :

```
Call:
glm(formula = Response ~ ., family = "binomial", data = train.data)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.36183  -0.41208  -0.20146  -0.08322   2.98000
```

```
Coefficients: (3 not defined because of singularities)
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -15.86489251  2.01357475  -7.879  0.00000000000000033
Marital_Status -1.45107483  0.19723202  -7.357  0.00000000000001878
Income        0.00001098  0.00001151   0.954    0.339836
Dt_Customer   0.00496497  0.00058440   8.496 < 0.0000000000000002
Recency       -0.03401140  0.00370109  -9.190 < 0.0000000000000002
MntWines      -0.00119915  0.00051496  -2.329    0.019879
MntFruits      0.00142554  0.00292663   0.487    0.626194
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 1352.16 on 1537 degrees of freedom
Residual deviance: 773.45 on 1511 degrees of freedom
AIC: 827.45
```

Number of Fisher Scoring iterations: 6

Confusion Matrix :

Confusion Matrix and Statistics

```
Reference
Prediction 0 1
0 548 40
1 25 47

Accuracy : 0.9015
95% CI : (0.8762, 0.9232)
No Information Rate : 0.8682
P-Value [Acc > NIR] : 0.005291
```

```
Kappa : 0.5358
McNemar's Test P-Value : 0.082478
```

```
Sensitivity : 0.54023
Specificity : 0.95637
Pos Pred Value : 0.65278
Neg Pred Value : 0.93197
Prevalence : 0.13182
Detection Rate : 0.07121
Detection Prevalence : 0.10909
Balanced Accuracy : 0.74830
```

'Positive' Class : 1

Key features: AIC: **827.45** Accuracy: **0.9015** Sensitivity: **0.54023** Specificity: **0.95637**

2. REDUCED MODEL:

Formula: Response ~ Marital_Status + Dt_Customer + Recency + MntWines + MntMeatProducts + NumDealsPurchases + NumCatalogPurchases + NumStorePurchases + NumWebVisitsMonth + AcceptedCmp3 + AcceptedCmp4 + AcceptedCmp5 + AcceptedCmp1 + AcceptedCmp2 + child + Education_PhD + Education_undergraduate

Variables selection : backward elimination

	Df	Deviance	AIC
<none>		779.73	815.73
- NumWebVisitsMonth	1	782.70	816.70
- child	1	782.94	816.94
- MntWines	1	783.17	817.17
- NumDealsPurchases	1	786.09	820.09
- Education_undergraduate	1	786.62	820.62
- AcceptedCmp2	1	787.50	821.50
- AcceptedCmp4	1	791.78	825.78
- Education_PhD	1	793.34	827.34
- MntMeatProducts	1	795.92	829.92
- NumCatalogPurchases	1	797.86	831.86
- AcceptedCmp1	1	800.26	834.26
- NumStorePurchases	1	810.80	844.80
- AcceptedCmp5	1	815.96	849.96
- Marital_Status	1	836.68	870.68
- AcceptedCmp3	1	841.56	875.56
- Dt_Customer	1	866.56	900.56
- Recency	1	883.40	917.40

Confusion Matrix :

Confusion Matrix and Statistics

```
Reference
Prediction 0 1
0 548 39
1 25 48

Accuracy : 0.903
95% CI : (0.8779, 0.9245)
No Information Rate : 0.8682
P-Value [Acc > NIR] : 0.003668
```

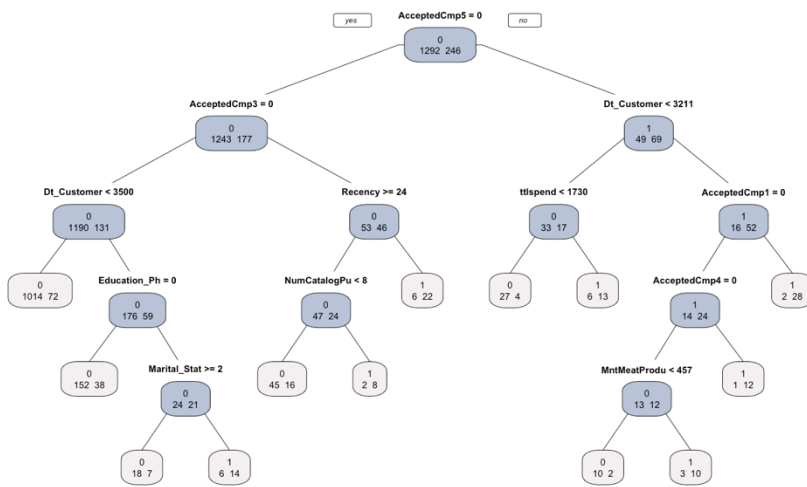
```
Kappa : 0.5453
McNemar's Test P-Value : 0.104163
```

```
Sensitivity : 0.55172
Specificity : 0.95637
```

Key features: AIC: **815.73** Accuracy: **0.903**
Sensitivity: **0.55172** Specificity: **0.95637**

DEFAULT REGRESSION TREE

Tree Plot :



Confusion Matrix :

Confusion Matrix and Statistics

		Reference	
Prediction		0	1
		0 556 61	1 17 26

Accuracy : 0.8818
 95% CI : (0.8547, 0.9055)
 No Information Rate : 0.8682
 P-Value [Acc > NIR] : 0.1641

Kappa : 0.3427

Mcnemar's Test P-Value : 0.000001123

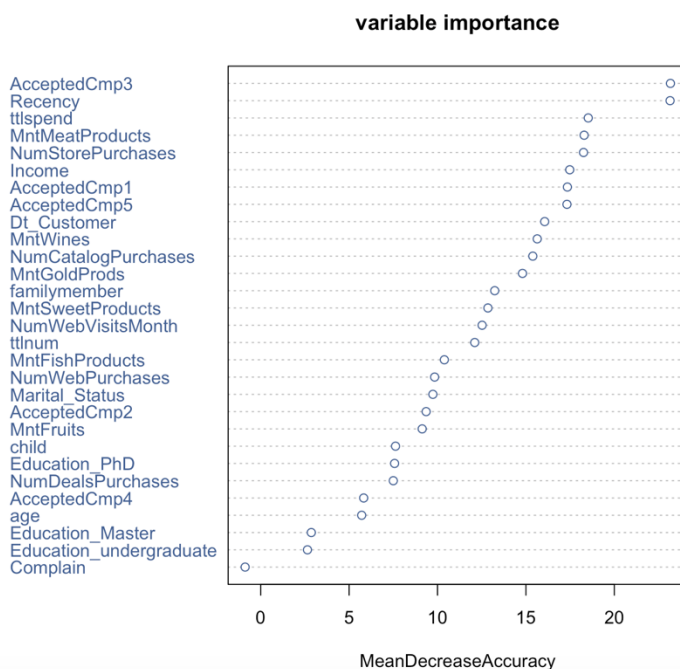
Sensitivity : 0.29885
 Specificity : 0.97033
 Pos Pred Value : 0.60465
 Neg Pred Value : 0.90113
 Prevalence : 0.13182
 Detection Rate : 0.03939
 Detection Prevalence : 0.06515
 Balanced Accuracy : 0.63459

Key features: Accuracy: 0.8818 Sensitivity: 0.29885 Specificity: 0.97033

'Positive' Class : 1

RANDOM FOREST

Variable Importance Plot :



Confusion Matrix :

Confusion Matrix and Statistics

		Reference	
Prediction		0	1
		0 559 53	1 14 34

Accuracy : 0.8985
 95% CI : (0.8729, 0.9205)
 No Information Rate : 0.8682
 P-Value [Acc > NIR] : 0.0105

Kappa : 0.4524

Mcnemar's Test P-Value : 0.000003443

Sensitivity : 0.39080
 Specificity : 0.97557
 Pos Pred Value : 0.70833
 Neg Pred Value : 0.91340
 Prevalence : 0.13182
 Detection Rate : 0.05152
 Detection Prevalence : 0.07273
 Balanced Accuracy : 0.68319

'Positive' Class : 1

Key features: Accuracy: 0.8985 Sensitivity: 0.39080 Specificity: 0.97557

MODEL COMPARISON

We can evaluate the efficiency of these models above by comparing their AIC, accuracy, sensitivity, specificity.

The comparison table:

	Full linear regression	Reduced linear regression	Default tree	Random forest
AIC	827.45	815.73	NA	NA
Accuracy	0.9015	0.903	0.8818	0.8985
Sensitivity	0.54023	0.55172	0.29885	0.39080
Specificity	0.95637	0.95637	0.97033	0.97557

Conclusion:

According to the table, we can see that the reduced linear regression model has the highest accuracy of 90.3%, the highest the sensitivity of 55.2%. It is also among the highest specificity. So we should choose the **reduced linear regression model** as our algorithms to predict consumer behavior.

Using the model, if the customer's 'response' to the promotion campaign is '1', which means promotion is efficient to this customer, than Company A should push the promotion campaign to him/her. Otherwise, Company A shouldn't take the action.

5.CONSUMER PREFERENCE PREDICTION

After predicted whether we should push the promotion campaign, Company A would likely to know which of the 6 kinds of products is the one the customer most likely to purchase.

Model used: reduced linear regression (backward elimination)

Dependent variable (group2.a) :

MntWine, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds, ttlspend

	Wine	Meat	Fruit	Fish	Sweet	Gold	ttlspend
multiple R ²	0.6856	0.6517	0.389	0.4553	0.4171	0.4234	0.8262
adjusted R ²	0.6838	0.6499	0.385	0.4517	0.4129	0.4196	0.8253

After regression of the 7 variables separately, we found that the models of fruit\fish\sweet\gold don't performance well, but the models of wine\meat/ttlspend are quite good. So we notice that the information provided **cannot** predict one's preference on **normal products** (fruit, fish, sweet, gold), but can have a **good** prediction on **ace products** (wines , meat) and the **total amount** one would spend.

Following, let's see the efficient models in details.

PREDICTION OF MNTWINES (AMOUNT SPENT ON WINE IN LAST 2 YEARS)

Formula: wine ~ Income + Dt_Customer + NumWebPurchases + NumCatalogPurchases + NumStorePurchases + NumWebVisitsMonth + child + Education_Master + Education_PhD

Summary of the regression result :

Call:

```
lm(formula = wine ~ Income + Dt_Customer + NumWebPurchases +
    NumCatalogPurchases + NumStorePurchases + NumWebVisitsMonth +
    child + Education_Master + Education_PhD, data = train.data)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-792.03 -109.09  -10.91   82.04  913.26
```

Coefficients:

```
              Estimate Std. Error t value      Pr(>|t|)
(Intercept)  -907.2200163   84.0498614  -10.794 < 0.0000000000000002 ***
Income         0.0083699    0.0004682   17.878 < 0.0000000000000002 ***
Dt_Customer   0.1189641    0.0263968    4.507 0.00000708376403745 ***
NumWebPurchases  9.4816695    2.6884575    3.527  0.000433 ***
NumCatalogPurchases 29.9296427    2.8445045   10.522 < 0.0000000000000002 ***
NumStorePurchases 18.0999787    2.2487494    8.049 0.00000000000000166 ***
NumWebVisitsMonth 33.3301922    3.4682962    9.610 < 0.0000000000000002 ***
child        -39.5864064    7.6390088   -5.182 0.00000024861529493 ***
Education_Master 62.1502598    13.3327646    4.661 0.00000341291978295 ***
Education_PhD  95.5450785    12.3179283    7.757 0.00000000000001584 ***
---
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 189.1 on 1528 degrees of freedom

Multiple R-squared: 0.6856, Adjusted R-squared: 0.6838

F-statistic: 370.3 on 9 and 1528 DF, p-value: < 0.00000000000000022

Predicted	Actual	Residual
415	173	-242
202	76	-126
160	14	-146
353	194	-159
244	84	-160
714	1012	298
590	867	277
-153	8	161
-38	6	44
-203	3	206

PREDICTION OF MNTMEAT (AMOUNT SPENT ON MEAT IN LAST 2 YEARS)

Formula: meat ~ Income + Dt_Customer + NumWebPurchases + NumCatalogPurchases + NumStorePurchases + age + child + Education_PhD

Predicted	Actual	Residual
178	118	-60
12	56	44
-8	24	32
303	480	177
127	38	-89
287	498	211
184	86	-98
-36	10	46
-2	14	16
-113	10	123

Summary of the regression result :

Call:

```
lm(formula = meat ~ Income + Dt_Customer + NumWebPurchases +
    NumCatalogPurchases + NumStorePurchases + age + child + Education_PhD,
    data = train.data)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-513.14  -71.94  -10.38   48.67  522.39
```

Coefficients:

```
              Estimate Std. Error t value      Pr(>|t|)
(Intercept)  -302.2482157   57.6574482   -5.242  0.000000181 ***
Income         0.0053157    0.0002779   19.131 < 0.0000000000000002 ***
Dt_Customer   0.0809520    0.0169061    4.788  0.000001845 ***
NumWebPurchases -3.7055854    1.6105025   -2.301  0.02153 *
NumCatalogPurchases 22.3429734    1.9010772   11.753 < 0.0000000000000002 ***
NumStorePurchases -2.6732388    1.4950113   -1.788  0.07396 .
age          -0.9418859    0.2935538   -3.209  0.00136 **
child        -67.3331085    5.0875429  -13.235 < 0.0000000000000002 ***
Education_PhD -24.1966276    8.1075249   -2.984  0.00289 **
---
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 128.3 on 1529 degrees of freedom

Multiple R-squared: 0.6517, Adjusted R-squared: 0.6499

F-statistic: 357.6 on 8 and 1529 DF, p-value: < 0.00000000000000022

PERDICTION OF TTLSPEND (TOTAL AMOUNT SPENT ON COMPANY A' S PRODUCTS)

Formula: $\text{ttlspend} \sim \text{Income} + \text{Dt_Customer} + \text{NumWebPurchases} + \text{NumCatalogPurchases} + \text{NumStorePurchases} + \text{NumWebVisitsMonth} + \text{age} + \text{child}$

Summary of the regression result :

Call:

```
lm(formula = ttlspend ~ Income + Dt_Customer + NumWebPurchases +
    NumCatalogPurchases + NumStorePurchases + NumWebVisitsMonth +
    age + child, data = train.data)
```

Residuals:

Min	1Q	Median	3Q	Max
-1310.64	-145.14	-19.55	117.00	1153.34

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1270.0414017	114.4179994	-11.100	< 0.0000000000000002 ***
Income	0.0148579	0.0006367	23.336	< 0.0000000000000002 ***
Dt_Customer	0.2585922	0.0356874	7.246	0.0000000000000678 ***
NumWebPurchases	10.5902475	3.6136451	2.931	0.00343 **
NumCatalogPurchases	68.2663174	3.8579205	17.695	< 0.0000000000000002 ***
NumStorePurchases	19.8649863	3.0719240	6.467	0.000000000134489 ***
NumWebVisitsMonth	20.3805091	4.7074912	4.329	0.000015924753669 ***
age	-1.5727677	0.5829754	-2.698	0.00706 **
child	-137.5313042	10.4031413	-13.220	< 0.0000000000000002 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 254.2 on 1529 degrees of freedom

Multiple R-squared: 0.8262, Adjusted R-squared: 0.8253

F-statistic: 908.6 on 8 and 1529 DF, p-value: < 0.0000000000000002

Predicted	Actual	Residual
1468	1617	149
960	716	-244
1551	1315	-236
502	317	-185
206	131	-75
331	302	-29
25	81	56
200	67	-133
-336	31	367
960	1319	359

6.RECOMMENDATION SYSTEM BASED ON THE PREDICTION

We assume that if **the predicted amount spent** is above the **0.75 quantile** of the original dataset, then the customer has preference on the product, and we should recommend this product to him/her.

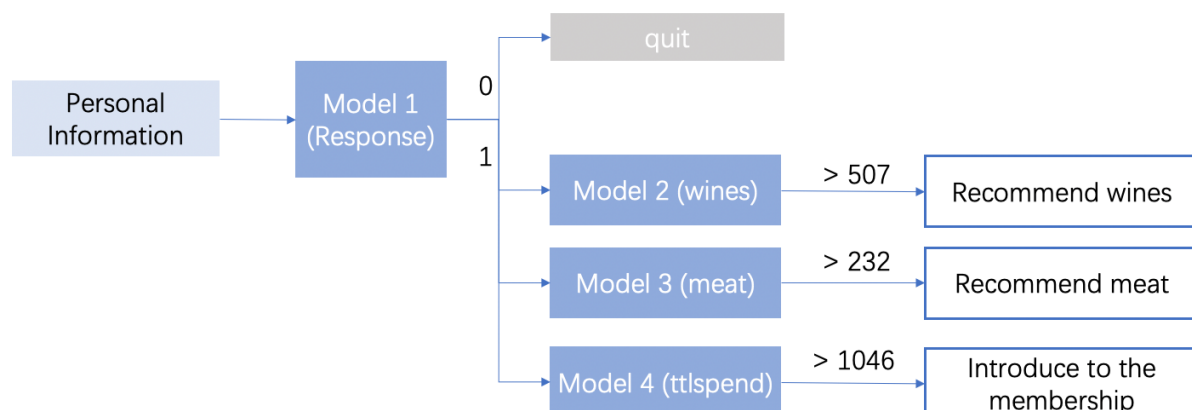
Models used: Model 1: reduced generalized linear regression of **Response**

Model 2: reduced linear regression of **MntWines (0.75 quantile=507)**

Model 3: reduced linear regression of **MntMeatProducts (0.75 quantile=232)**

Model 4: reduced linear regression of **ttlspend (0.75 quantile=1046)**

Recommendation logistic:



7.CUSTOMER CLUSTERING & MARKET SEGMENT

K-MEANS CLUSTERING

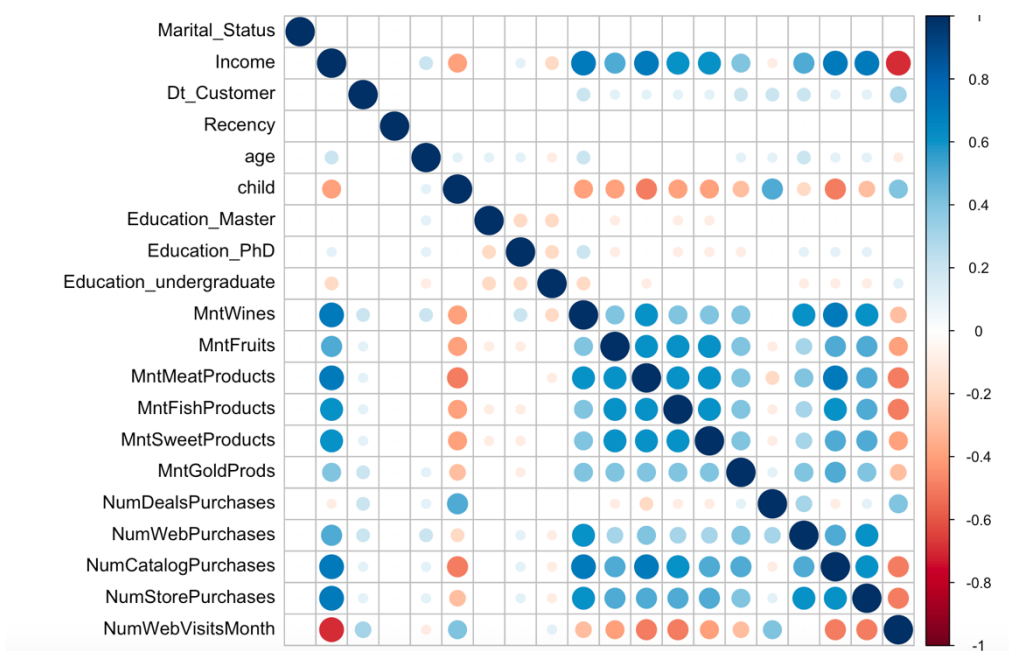
SELECT VARIABLES:

Group1: age, Education(3 dummies), Marital_Status, Income, child, family, Dt_Customer, Recency (10)

Group2.a: MntWine, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds (6)

Group3: NumWebPurchases, NumCatalogPurchases, NumStorePurchases, NumWebVisitsMonth (4)

Correlation table:



Kappa: 25.29492

After standardizing the data, the kappa of it is 25.29492.

Correlations between the variables are not so great.

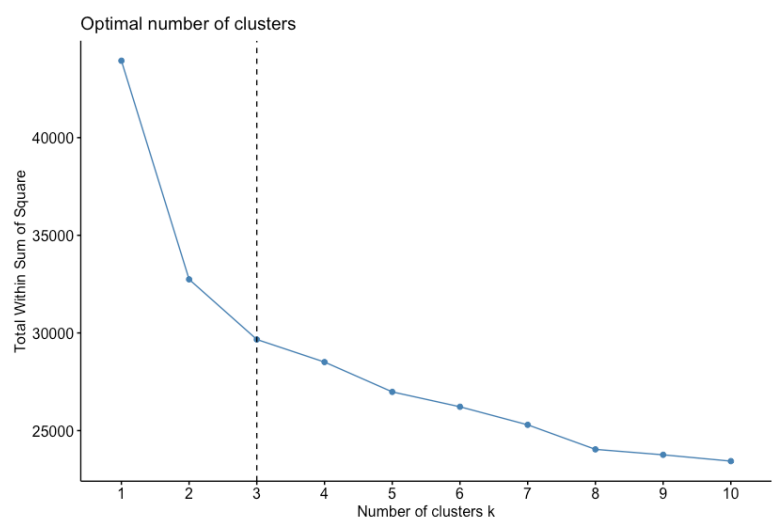
K-MEANS CLUSTERING:

1. Determine cluster number

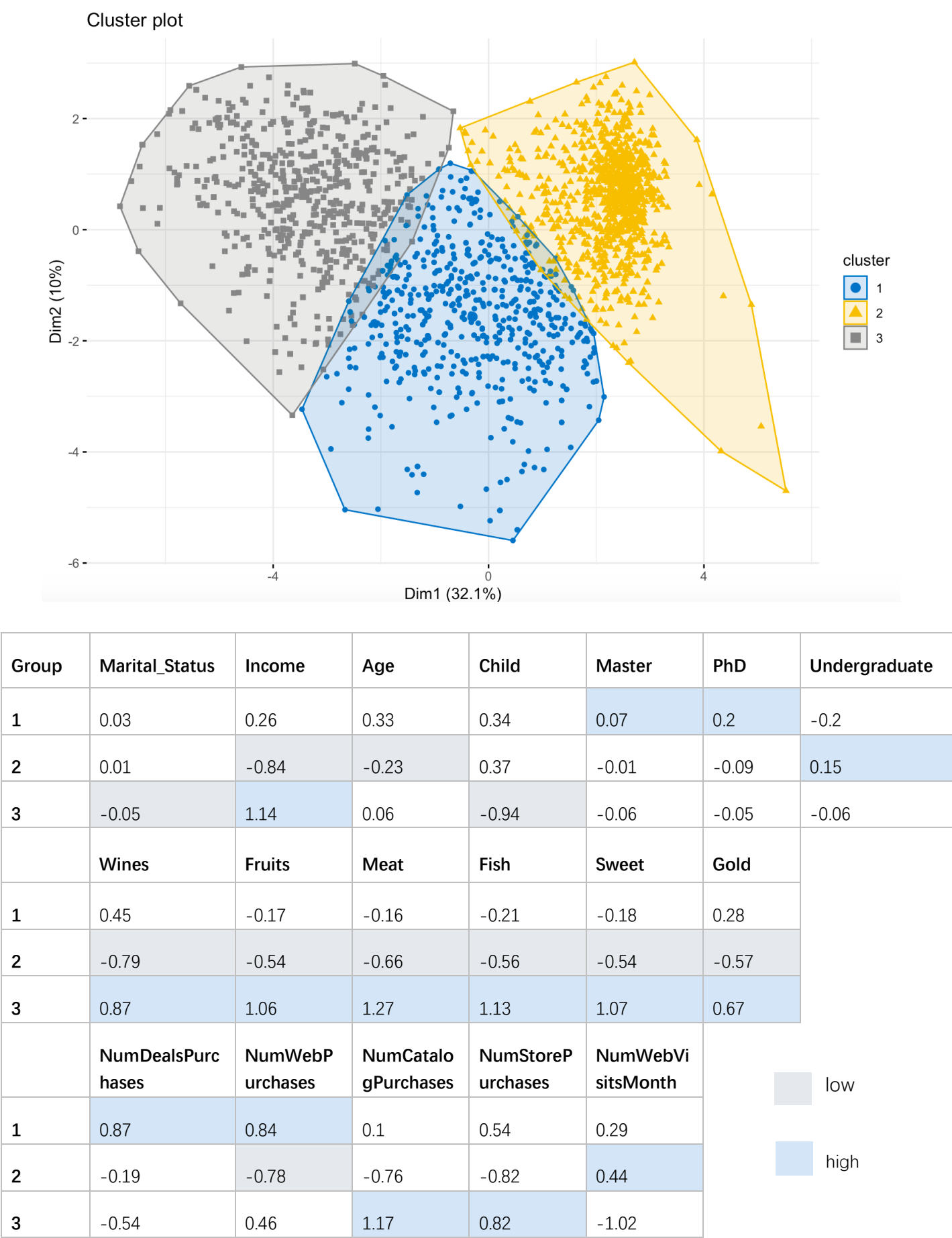
Looking at the elbow plot, we can see that it's decline tends to be gender from 3.

So, we choose to cluster the customers into **three** groups.

Elbow plot:



2. K-means clustering



MARKET SEGMENT AND SUGGESTION

CUSTOMER PROFILES OF THE THREE GROUPS:

Group 1:

Demographic: **middle** income; **high** education level

Behaviors: prefer **wines and gold** to other products

Distributions: prefer making purchases through **websites** and using **discount**

Group 2:

Demographic: **low** income; **younger** generation; most **undergraduate**

Behaviors: don't like most products of company A

Distributions: like visiting the website but seldom make purchase

Group 3:

Demographic: **single; high** income; without child

Behaviors: like most products of company A

Distributions: prefer making purchase at **stores** and through catalogue

SUGGESTION TO COMPANY A

	Group 1	Group 2	Group 3
Attribute	Potential consumer base	not target consumer	Loyal consumer
How to improve	<ul style="list-style-type: none">• Provide discount of wines and gold products• Focusing on website distribution	Drop	<ul style="list-style-type: none">• Improve in-store experience• Establish membership system

8.APPENDIX

Attributes of original variables:

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