STAT 450 Project

Customer Personality Analysis

Submitted to
Dr. Rebecca Le

Prepared by
Evan Cabrera
Sean Cunniff
Gregory Lent
Luis Osorio

May 7, 2023

I. Background and Introduction

Businesses serve a large and varied demographic of customers. As data science expands, more information is being collected by businesses about customers every day. To maintain customer happiness and increase profitability, it is useful to provide targeted advertisements to customers. Targeted advertisements allow customers to not have their time wasted, and for businesses to not waste money on meaningless promotions. The decision about how to best advertise to customers can be complicated, but being able to identify specific groups and needs is a good first step.

This leads to our main research question of interest: How can customer personalities be used to divide customers into groups? If there are no prior assumptions being made about group numbers or structures, an excellent technique to answer this question is via Cluster Analysis. Cluster analysis is a form of unsupervised classification, in which variables are grouped into clusters based upon similarities or distances. Furthermore through our process of data exploration we developed a secondary research question: What is the impact of education and children on the total purchase amount and income?

II. Data and Exploratory Analysis

To answer our research question, we utilized the Customer Personality Analysis dataset found on Kaggle.com. This dataset contains information about 2240 customers. Each row of data includes several types of information including demographic information, purchase history, purchase format, and how the customer has responded to marketing outreach.

We began our investigation into the data with exploratory data analysis, and basic visualization of the data set which can provide further insights. We examined the overall shape of the data, investigated outliers in numeric variables, and calculated new variables that were believed to be helpful for future analysis.

Our initial exploration found that the data was composed of 2240 customers across 27 variables. We examined the data for absent or missing values, and we found 24 customers without income data; these observations were dropped from the data set. Furthermore, we plotted several numeric variables (see Appendix A**)** which we suspected might contain outliers, and through this process determined that one customer's income was significantly larger than other customers, and three customers had ages believed to be outliers. These four customers were removed from the dataset. Leaving the final number of observations at 2212.

Age and "time as customer" are provided in the dataset as dates, so they have been given an appropriately calculated value in years. Two additional variables were calculated due to suspected usefulness: total spent by a customer, and total number of advertisements responded to. The final part of our data exploration included constructing a correlation plot of our numeric variables to receive a general understanding of how our data might be related (see Appendix A).

III. Model and Results

Cluster Analysis

Clustering is an unsupervised learning technique which classifies observations of an unlabeled dataset into groups. There are several different types of Clustering algorithms which all use different distance metrics. For our specific dataset we used K-Means clustering, which is a non-hierarchical clustering method that uses the smallest centroid distance to classify the observation. A notable feature of K-Means clustering is that it is very sensitive to the number of clusters chosen.

In order to determine the optimal number of clusters, we considered the Elbow Method and the Silhouette Score. The Elbow method uses Inertia, or within-cluster sum of squares, which calculates the sum of squared distances between each data point and the centroid of its assigned cluster. As the number of clusters increases, the within-cluster sum of squares will decrease. These points are plotted in a graph, and the elbow of the graph is chosen, hence the name. However, for this dataset, the Elbow Method proved to be an impractical choice.

Instead, we used the Silhouette Score, a criterion which takes on a value between -1 and 1. This score provides a measure of how well each data point fits into its assigned cluster based on both how close it is to other points in the cluster and how far it is from points in the neighboring clusters. A score close to 1 indicates that the data point is well-matched to its own cluster, while a score close to -1 indicates a poor fit. A score close to 0 indicates that the point is roughly equally close to points in its own cluster and those in neighboring clusters. Using the Silhouette Score metric we discovered that the optimal number of clusters for our model was two with a Silhouette Score of 0.286.

In these two clusters, the first cluster contained 864 observations and the second cluster contained 1344 observations. To investigate what separates these two clusters, we examined the criteria used by the K-Means algorithm to find the difference between these two groups. To do this, we examined the differences between the means of each variable between the two clusters. In order to quantify these differences, we conducted a t-test using a significance level of 0.05/20 (the number of predictors) = 0.0025 testing the differences between the means of each predictor (See Appendix A). Since the t-test only works for continuous variables, we used a Chi-Squared Test of Homogeneity for the categorical variables in our dataset. These tests showed that income, children, amount spent on meat in the last 2 years, amount spent on fish in the last 2 years, number of purchases made using a catalog, and total amount spent differed significantly between the two clusters.

MANOVA

In order to quantify the effect of education and children on income and spending, a two-way MANOVA (Multivariate Analysis of Variance) was performed. The customers were grouped according to education level (high school, college degree, or postgraduate degree), as well as whether they have any children at home, for a total of six groups. The response variable was a

vector containing yearly income and total spent in the store in the last two years. Additionally, a post-hoc Tukey's HSD test was performed to test the individual interactions between the grouping variables.

When testing for the multivariate normality assumption of the MANOVA model, it was found that income does not follow a normal distribution, so the assumption of multivariate normality was not satisfied (see Appendix A). Additionally, the assumption of homogeneous variance-covariance was not satisfied for education level or number of children (see Appendix A).

Both income and total spent were shown to be significantly affected by both education and presence of children ($p < 0.0001$, see Appendix A). Additionally, most interactions between variables were significant in the post-hoc tests. For income, insignificant interactions include the presence of children for those with a high school or college education, and the difference between a graduate and postgraduate college degree for customers with and without children. For total spent, insignificant interactions include the presence of children for those with a high school or college education, the difference between those without children and a high school education and those with children and a postgraduate degree, and the difference between a graduate and postgraduate college degree for customers with and without children (see Appendix A). All other differences between groups were significant.

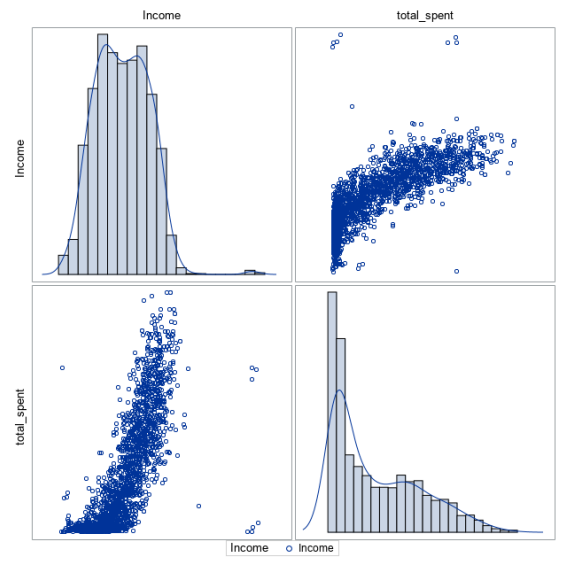
IV. Discussion and Conclusion

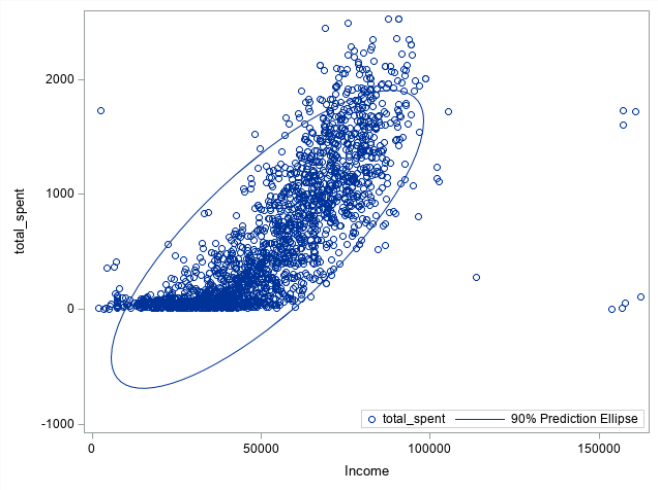
After cleaning the data and exploring the surface level associations, we used a K-means clustering method to group customers into clusters. The number of clusters to be used within our K-means method was specifically picked following consideration of the Elbow Method and Silhouette Score for the number of clusters to use. Comparing the column means between our clusters we were able to eliminate insignificant variables from the clusters and develop a characterization of the clusters.

Considering the significance of children in the cluster, and our previous expectation of the importance of education, we decided to quantify the influence of these two variables on total amount spent and income. Our multivariate normality assumptions failed, deeming our MANOVA findings unreliable. However, we were able to determine that income and total spending were shown to be significantly affected by the presence of children and education level.

In conclusion, we have found that our multivariate statistical methods of Clustering and MANOVA to be useful in the analysis of customer data. These methods were able to break customers into groups, and further define relationships and interactions between variables. The results from using either of these methods could potentially help businesses to better provide services to customers based upon customer shopping personalities.

Appendix A: Graphs and Results





The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
195.095649	3	<.0001

The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
189.723961	6	<.0001

Dependent Variable: Income

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	228499477082	45699895416	126.63	<.0001
Error	2206	796130399741	360893200.25		
Corrected Total	2211	1.0246299E12			

Dependent Variable: total_spent

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	239953327.5	47990665.5	188.15	<.0001
Error	2206	562689096.5	255072.1		
Corrected Total	2211	802642424.0			

children	new_education	total_spent LSMEAN	LSMEAN Number
0	College Degree	1107.43438	1
0	High School	136.17647	2
0	Post Graduate Degree	1156.18644	3
1	College Degree	426.86918	4
1	High School	56.81081	5
1	Post Graduate Degree	406.47727	6

Least Squares Means for effect children*new_educati
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: Income

i/j	1	2	3	4	5	6
1		<.0001	0.5070	<.0001	<.0001	<.0001
2	<.0001		<.0001	<.0001	0.7364	<.0001
3	0.5070	<.0001		<.0001	<.0001	<.0001
4	<.0001	<.0001	<.0001		<.0001	0.1507
5	<.0001	0.7364	<.0001	<.0001		<.0001
6	<.0001	<.0001	<.0001	0.1507	<.0001	

Least Squares Means for effect children*new_educati
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: total_spent

i/j	1	2	3	4	5	6
1		<.0001	0.2318	<.0001	<.0001	<.0001
2	<.0001		<.0001	0.0190	0.5918	0.0292
3	0.2318	<.0001		<.0001	<.0001	<.0001
4	<.0001	0.0190	<.0001		<.0001	0.4281
5	<.0001	0.5918	<.0001	<.0001		<.0001
6	<.0001	0.0292	<.0001	0.4281	<.0001	

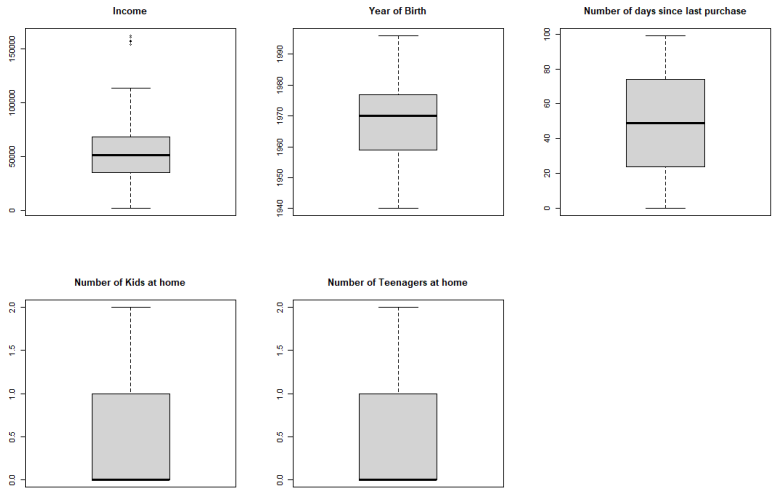
children	new_education	_FREQ_	Income	total_spent
0	College Degree	320	66407.93125	1107.434375
0	High School	17	21590.235294	136.17647059
0	Post Graduate Degree	295	67425.488136	1156.1864407
1	College Degree	795	46438.657862	426.86918239
1	High School	37	19716.324324	56.810810811
1	Post Graduate Degree	748	47829.627005	406.47727273

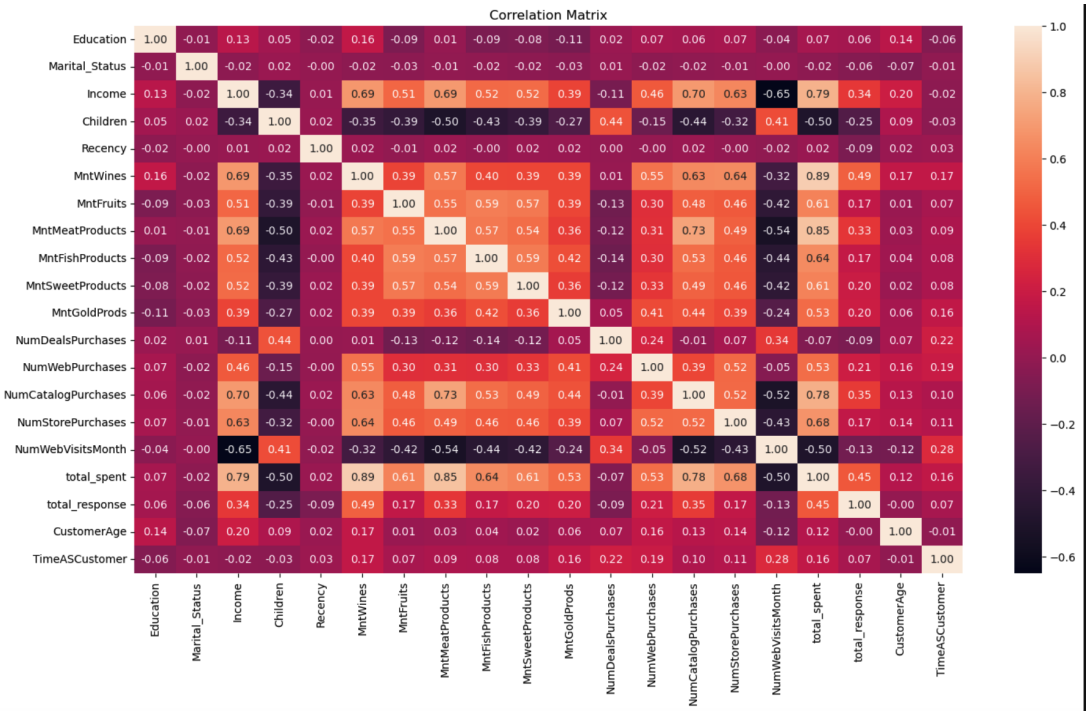
```
── Data Summary ────────────────────────
                           Values
Name                       df
Number of rows             2240
Number of columns          29
_________________________
Column type frequency:
  Date                     1
  factor                   2
  numeric                  26
_________________________
Group variables            None

── Variable type: Date ─────────────────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate min        max        median     n_unique
1 Dt_Customer           0             1 2012-07-30 2014-06-29 2013-07-08      663

── Variable type: factor ───────────────────────────────────────────────────────────────────────────
  skim_variable  n_missing complete_rate ordered n_unique top_counts
1 Education              0             1 FALSE          5 Gra: 1127, PhD: 486, Mas: 370, 2n : 203
2 Marital_Status         0             1 FALSE          8 Mar: 864, Tog: 580, Sin: 480, Div: 232

── Variable type: numeric ──────────────────────────────────────────────────────────────────────────
   skim_variable       n_missing complete_rate        mean          sd   p0     p25    p50    p75   p100
 1 ID                          0             1      5592.        3247.       0  2828.   5458.  8428.  11191
 2 Year_Birth                  0             1      1969.          12.0    1893  1959    1970   1977   1996
 3 Income                     24             0.989 52247.       25173.    1730 35303   51382. 68522  666666
 4 Kidhome                     0             1         0.444       0.538    0     0       0      1      2
 5 Teenhome                    0             1         0.506       0.545    0     0       0      1      2
 6 Recency                     0             1        49.1        29.0      0    24      49     74     99
 7 MntWines                    0             1       304.        337.       0    23.8   174.   504.   1493
 8 MntFruits                   0             1        26.3        39.8      0     1       8     33     199
 9 MntMeatProducts             0             1       167.        226.       0    16      67    232    1725
10 MntFishProducts             0             1        37.5        54.6      0     3      12     50     259
11 MntSweetProducts            0             1        27.1        41.3      0     1       8     33     263
12 MntGoldProds                0             1        44.0        52.2      0     9      24     56     362
13 NumDealsPurchases           0             1         2.33        1.93     0     1       2      3      15
14 NumWebPurchases             0             1         4.08        2.78     0     2       4      6      27
15 NumCatalogPurchases         0             1         2.66        2.92     0     0       2      4      28
16 NumStorePurchases           0             1         5.79        3.25     0     3       5      8      13
17 NumWebVisitsMonth           0             1         5.32        2.43     0     3       6      7      20
18 AcceptedCmp3                0             1         0.0728      0.260    0     0       0      0       1
19 AcceptedCmp4                0             1         0.0746      0.263    0     0       0      0       1
20 AcceptedCmp5                0             1         0.0728      0.260    0     0       0      0       1
21 AcceptedCmp1                0             1         0.0643      0.245    0     0       0      0       1
22 AcceptedCmp2                0             1         0.0134      0.115    0     0       0      0       1
23 Complain                    0             1         0.00938     0.0964   0     0       0      0       1
24 Z_CostContact               0             1         3           0        3     3       3      3       3
25 Z_Revenue                   0             1        11           0       11    11      11     11      11
26 Response                    0             1         0.149       0.356    0     0       0      0       1
```





Unpaired T-Test Results for our 2 Clusters:

Our hypothesis test:

$H_0 : \mu_{Cluster_1,i} = \mu_{Cluster2,i}$, where i = all Columns

$H_a : \mu_{Cluster_1,i} \neq \mu_{Cluster2,i}$, where i = all Columns

	Columns	T-Statistics	P-Value	Hypothesis
0	Income	3.213192	0.001335	Reject Null
1	Children	-4.251294	0.000022	Reject Null
2	Recency	0.515392	0.606342	Fail to Reject Null
3	MntWines	2.664487	0.007780	Fail to Reject Null
4	MntFruits	2.181306	0.029289	Fail to Reject Null
5	MntMeatProducts	3.409790	0.000665	Reject Null
6	MntFishProducts	3.432960	0.000611	Reject Null
7	MntSweetProducts	2.964397	0.003073	Fail to Reject Null
8	MntGoldProds	2.946625	0.003256	Fail to Reject Null
9	NumDealsPurchases	-2.270911	0.023263	Fail to Reject Null
10	NumWebPurchases	2.868567	0.004172	Fail to Reject Null
11	NumCatalogPurchases	3.915758	0.000094	Reject Null
12	NumStorePurchases	2.345334	0.019118	Fail to Reject Null
13	NumWebVisitsMonth	-2.045779	0.040923	Fail to Reject Null
14	total_spent	3.693768	0.000228	Reject Null
15	total_response	2.650108	0.008124	Fail to Reject Null
16	CustomerAge	0.559338	0.576002	Fail to Reject Null
17	TimeASCustomer	-1.358044	0.174616	Fail to Reject Null

Appendix B: Code

Python Code can be found here in ipynb and pdf format along with the data sets.

```
proc import out = customer
     datafile = "&path/cleaned.csv"
     dbms = csv replace;
run;

 * Cleaning data;
data customer;
     set customer;
     if kidhome = 0 and teenhome = 0 then children = 0;
     else children = 1;
     if education = '2n Cycle' or education = 'Master' or education = 'PhD' then new_education = 'Post Graduate Degree';
     if education = 'Basic' then new_education = 'High School';
     if education = 'Graduation' then new_education = 'College Degree';
run;

 * Testing normality;
proc glm data=customer;
     class children new_education;
     model income total_spent = children new_education children*new_education;
     output out=resids r=income total_spent;
run;

ods graphics on;
proc sgscatter data=resids;
     matrix income total_spent/
     group=income ellipse=(type=mean) diagonal=(histogram kernel);
run;
ods graphics off;

ods graphics on;
proc sgplot data=resids;
     scatter x=income y=total_spent;
     ellipse x=income y=total_spent/ alpha=0.1;
     keylegend/ location=inside position=bottomright;
run;
ods graphics off;
```

```
 * Testing homogeneity of variance-covariance;
proc discrim data=customer pool=test;
     class children;
     var income total_spent;
 run;

proc discrim data=customer pool=test;
     class new_education;
     var income total_spent;
 run;

 * Running MANOVA with Tukey's HSD Post-hoc test;
proc glm data=customer;
     class children new_education;
     model income total_spent = children new_education children*new_education;
     manova h = children*new_education;
     lsmeans children*new_education / pdiff;
     output out=resids;
 run;

 * Generating means table;
proc means data=customer;
   class children new_education;
   var income total_spent;
   output out=means_table mean=;
 run;

data means_table;
   set means_table;
   if _TYPE_ = 3;
   drop _TYPE_;
 run;

proc print data=means_table noobs;
 run;
```

```
library(tidyverse)
library(skimr)
library(corrplot)
library(gridExtra)


## Data Importation and Correction

df = read.csv("./marketing_campaign.csv")
glimpse(df)

df = df |> mutate(
    Dt_Customer = as.Date(Dt_Customer,"%d-%m-%Y"),
    Education = as.factor(Education),
    Marital_Status = as.factor(Marital_Status))


## Sumamry Statistics
df |> skim_without_charts()
df = df |> drop_na()


#Boxplots
par(mfrow=c(2,3))
boxplot(df$Income,main="Income")
boxplot(df$Year_Birth, main="Year of Birth")
boxplot(df$Recency,main="Number of days since last purchase")
boxplot(df$Kidhome,main="Number of Kids at home")
boxplot(df$Teenhome,main="Number of Teenagers at home")
par(mfrow=c(1,1))


#Filtering outliers
df = df |>
    filter(Income < 600000 & Year_Birth > 1920)

#Correlation Plot of numerics
correlations = df |>
    select_if(is.numeric) |>
    select(-c(AcceptedCmp1,AcceptedCmp2, AcceptedCmp3,AcceptedCmp4,AcceptedCmp5,
              Response,Complain,Z_CostContact,Z_Revenue,Year_Birth,ID)) |>
    cor()
corrplot(correlations, method = 'color', order = 'alphabet',type = 'lower', diag = FALSE)

## Calculated fields
df = df |>
    group_by(ID) |>
    mutate(
        total_spent = sum(MntWines,MntFruits,MntMeatProducts,MntFishProducts,
                          MntSweetProducts,MntGoldProds),
        total_response = sum(AcceptedCmp1,AcceptedCmp2,AcceptedCmp3,
                             AcceptedCmp4,AcceptedCmp5,Response),
        CustomerAge = 2021 - Year_Birth,
        TimeASCustomer = (as.Date("2021-01-01") - Dt_Customer)/365,
    )


## Export the data
write.csv(df,"./cleaned.csv")
```