SI 618 Project Part 2 Report

Name: Yang Shen
Uniqname: lizshen

1 Motivation

As an enthusiastic consumer myself, I am very interested in customer personality, which is related to many factors, such as social class, age, education level, etc.. For sales industry, analyzing customer personality helps to understand production needs and adjust production capacity. For advertising industry, it helps to adjust the propaganda strategy to achieve a more targeted publicity effect.

This project will perform exploratory analysis on a customer information dataset and try to explore the customer personality. There are three main problems:

1. Find out the relationship between the consumption amount on each type of product (6 types) and variables of personal information (age, marital status, education level, etc.). Which factor has the most impact? And how does consumption amount distribute depending on this factor?
2. Find out the characteristics of people who tend to accept promotion.
3. Create clusters to show customer segments.

2 Data Source

`https://www.kaggle.com/imakash3011/customer-personality-analysis`

The dataset I use is Customer Personality Analysis dataset. The data is in a csv file, but is separated by tabs instead of comma. There are four types of columns, while each row represents one customer:

1. Personal info, including Year_Birth, Education, Income, etc.
2. Consumption amount: amount of money spent on 6 types of products, including MntWines, MntFruits, etc.
3. How the customer purchase things: number of purchases in different ways, including NumWebPurchases, NumCatalogPurchases, etc.
4. Promotion: the number of the customer accepting promotion campaign, etc.

There are in total 2240 records.

The enrollment date of each customer is recorded in the dataset. The latest is 2014-12-06, the earliest is 2012-01-08.

3 Methods

3.1 Q1: Find out the relationship between the consumption amount on each type of product and variables of personal information.

3.1.1 Manipulation

I used two sets of columns, one is the Personal info, one is the Consumption amount.

I created a new dataframe for this question, with the two column sets mentioned. I modified some columns to be more readable and convenient to analyze. For example, 'Year_Birth' was turned into 'Age', 'Kidhome' and 'Teenhome' were summed as 'Children', 'Marital_Statis' was turned into 'Partner' (0 or 1), etc.

3.1.2 Missing/Incomplete/Noisy Data

In 'Income' there exists missing data. These rows are dropped at the very beginning.

Also I created a very general pairplot to observe the outliers, and got rid of them, including rows with Age $>$ 100 or Income $>$ 600000.

3.1.3 Challenges

There are not many obviously correlated relationships.

The consumption amount of different types are in different columns, so to plot them on the same diagram, I would need to melt the dataframe and set the types as hue.

3.2 Q2: Find out the characteristics of people who tend to accept promotion.

3.2.1 Manipulation

I used two sets of columns, one is the Personal info, one is the Promotion.

I created a new dataframe for this question, with the two column sets mentioned. I modified some columns to be more readable and convenient to analyze. For example, the columns for promotion campaign 1-6 are summed as 'AcceptedTtl', meaning total number of accepted promotions; income are put into 3 bins, as 'Income_cut', etc.

3.2.2 Missing/Incomplete/Noisy Data

Processed in Question 1.

3.2.3 Challenges

It is sure that some variables are dependent (e.g. Family size & the number of accepted promotions), but it is hard to summarize a certain pattern from the correlation value and mosaic plot. I solve this by creating cross tables and describe the pattern the best as I can.

3.3 Q3: Create clusters to show customer segments.

3.3.1 Manipulation

I created two new columns: 'NumTtlPurchases' and 'MntPerPurchase'. Then I extracted four numerical columns which are most likely co-variant to perform further clustering.

3.3.2 Missing/Incomplete/Noisy Data

Simply drop all NaN and inf rows.

3.3.3 Challenges

The number of cluster was a bit hard to determine, since the elbow method is ambiguous. I initially set the number of cluster at 4 or 5 by elbow method, and finally decide to make 4 clusters by trial and error.

4 Analysis and Results

4.1 Q1: Find out the relationship between the consumption amount on each type of product and variables of personal information.

4.1.1 Work Flow

1. Extract the necessary columns for this question and create a new table.
2. Create a pair plot and a correlation table for this table.
3. Find the most correlated variables. Here it is the 'Income' and 'MntWines', 'MntFruits',..., 'MntGoldProds', 'MntTtl' (columns in Consumption Amount).
4. Perform an ANOVA to test the relation ship between 'Income' and 'MntTtl'.
5. Melt the table, set different types of products as hue. Use jointplot and ggplot with geom_smooth() to check the relationship between 'Income' and 'MntWines', 'MntFruits', ..., 'MntGoldProds'.

4.1.2 Results and Visualization

After Step 2 in the above work flow, it is shown that the most noticeable relationship is between 'Income' and the columns in Consumption Amount (other corr values are too small), with corr value all above 0.5.

So I performed an ANOVA test between 'Income' and 'MntTtl' (Step 4). It turned out that the R-squred is 0.996, very high in the OLS regression result. And the p value in the ANOVA test is $7.417894 * 10^{-130}$, almost 0. Therefore it is definite that 'Income' and 'MntTtl' are positively linearly correlated. Then I melted the table to obeserve each category's relationship with 'Income' (Step 5). First I created a joint plot (Figure 1). Although the points kind of block each other, we can still observe that the amount of wines and meat products grows much more than other categories as the income grows.

Then I made a ggplot with geom_smooth() with the linear model (Figure 2). In this figure, it is more obvious that people tend to buy more wines and meat products as the income increases, while the purchases for other types of products are relatively stable.

4.2 Q2: Find out the characteristics of people who tend to accept promotion.

4.2.1 Work Flow

1. Extract the necessary columns for this question and create a new table.
2. Create a correlation table for this table.
3. Find the most correlated variables. As a result, I will analyze around these variables: Family_Size, Income, NumDealsPurchases, AcceptedTtl.

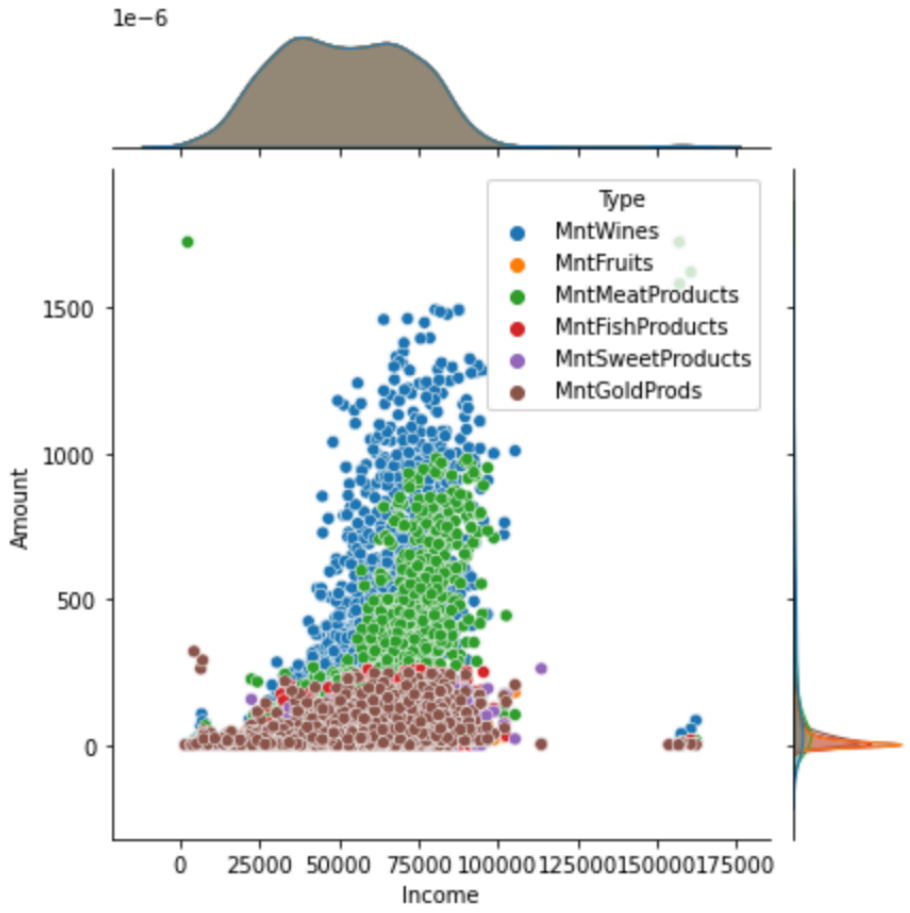

Figure 1: Jointplot of income and consumption amount.

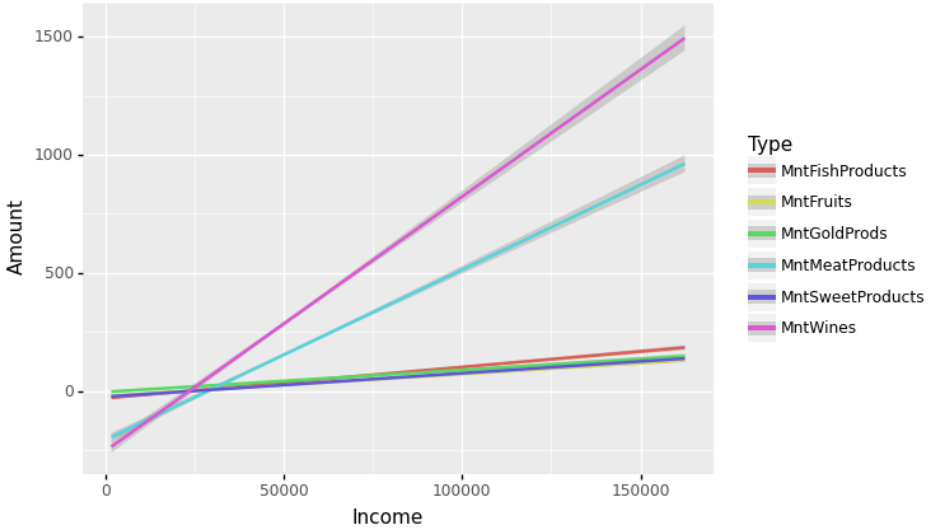

Figure 2: ggplot of income and consumption amount.

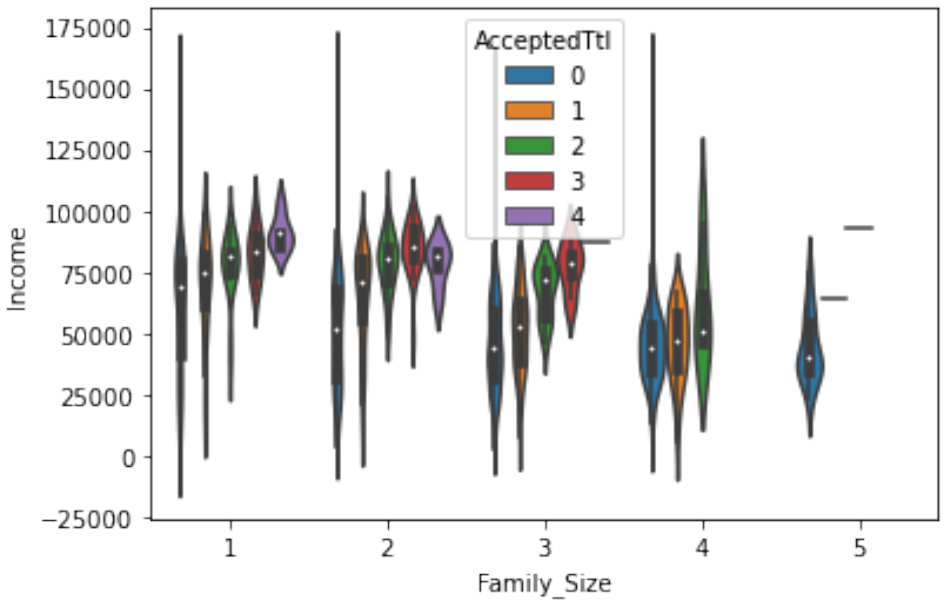

Figure 3: Violin plot.

4. Create a violin plot to see how number of promotion accepted and income are related with the family size varying.

5. Use mosaic plot and crosstable to analysis the relationship of Income vs. AccepetedTtl, Income vs. NumsDealsPurchases, Family_Size vs. AccepetedTtl, Family_Size vs. NumsDealsPurchases.

4.2.2 Results and Visualization

The result of the correlation table shows that there are likely positive relation between Income and AccepetedTtl, Family_Size and NumsDealsPurchases, negative relation between Income and NumsDealsPurchases, Family_Size and AccepetedTtl. After Step 4 (Figure 3), it is observed that people with higher income in small family accept more promotions.

Then I cut 'Income' using qcut, put the data into 3 bins, renamed as 'low', 'middle', 'high'.

Income_cut vs. AccepetedTtl

From Figure 4, it is obvious that the propotion in the mosaic plot is not even. From low income to high income, the propotion of more than 1 accepted promotion grows visibly.

From the crosstab, we can tell people with high income are most likely to accept promotions. Also, the p value is $1.025852855584581 * 10^{-52}$, proving income and total number of accepting promotions are significantly dependent.

Income_cut vs. NumDealsPurchases

From the crosstab, we can tell People with middle and low income are more likely to make purchases with discount. Also, the p value is $3.894225196354892 * 10^{-67}$, proving income and number of purchases with discount are significantly dependent.

Family_Size vs. AccepetedTtl

From Figure 7, it is obvious that the propotion in the mosaic plot is not even. From low income to high income, the propotion of more than 1 accepted promotion decreases

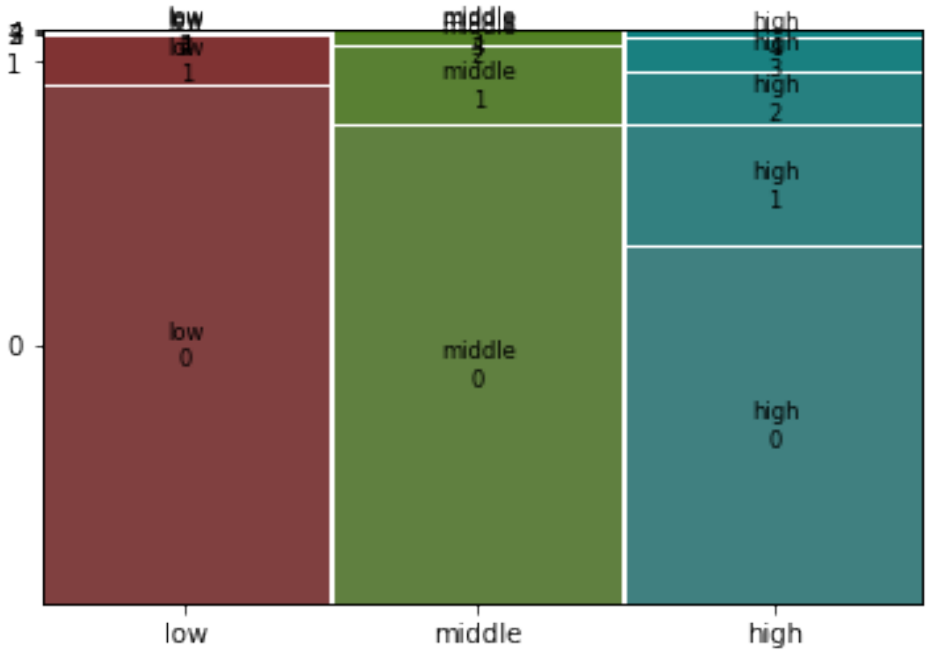

Figure 4: Mosaic plot of Income_cut vs. AccepetedTtl.

AcceptedTtl	0	1	2	3	4
Income_cut					
low	666	63	1	0	0
middle	615	100	13	2	0
high	473	159	67	42	11

Figure 5: Crosstab of Income_cut vs. AccepetedTtl.

NumDealsPurchases	0	1	2	3	4	5	6	7	8	9	10	11	12	13	15
Income_cut															
low	7	305	202	118	60	16	9	6	1	2	0	0	0	0	4
middle	0	188	177	114	89	62	49	24	10	5	4	2	3	3	0
high	37	464	114	61	38	16	2	9	3	1	1	3	0	0	3

Figure 6: Crosstab of Income_cut vs. NumDealsPurchases.

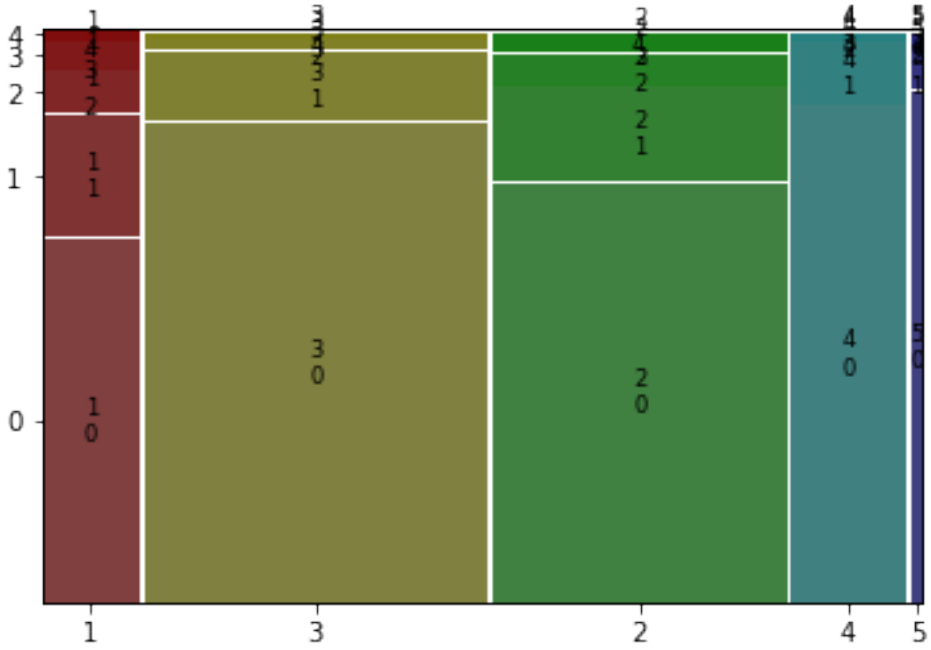

Figure 7: Mosaic plot of Family_Size vs. AccepetedTtl.

AcceptedTtl	0	1	2	3	4
Family_Size					
1	162	54	19	12	5
2	559	125	42	24	4
3	745	109	15	8	2
4	260	32	4	0	0
5	28	2	1	0	0

Figure 8: Crosstab of Family_Size vs. AccepetedTtl.

visibly (Notice here family size of 2 and 3 switched places on the axis).

From the crosstab, we can tell Family with 2 or 3 are most likely to accept promotions. Also, the p value is $2.7733425947691215 * 10^{-15}$, proving family size and total number of accepting promotions are significantly dependent.

Family_Size vs. NumDealsPurchases

From the crosstab, we can tell family with 2 or 3 are most likely to buy with discount. Also, the p value is $1.0690292829941127 * 10^{-115}$, proving family size and number of purchases with discount are significantly dependent.

4.3 Q3: Create clusters to show customer segments.

4.3.1 Work Flow

1. Create new columns: 'NumTtlPurchases' and 'MntPerPurchas'.

NumDealsPurchases	0	1	2	3	4	5	6	7	8	9	10	11	12	13	15
Family_Size															
1	20	199	21	10	0	0	0	0	0	0	0	0	0	0	2
2	18	444	134	73	45	16	15	5	2	0	0	1	0	0	1
3	5	250	259	159	97	54	28	14	6	4	0	1	0	0	2
4	1	57	73	46	43	23	17	11	6	4	5	2	3	3	2
5	0	7	6	5	2	1	0	9	0	0	0	1	0	0	0

Figure 9: Crosstab of Family_Size vs. NumDealsPurchases.

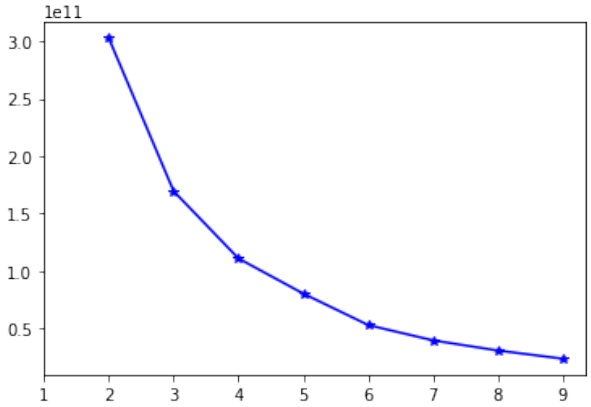

Figure 10: Elbow plot.

2. Use 'Income', 'MntTtl', 'NumTtlPurchases' and 'MntPerPurchas' as features, performing elbow method.

3. Determine the number of cluster from the result of elbow.

4. Assign the cluster number. Perform some analysis.

4.3.2 Results and Visualization

In Figure 10 is the result of the elbow method. It is hard to tell that the ideal number should be 4 or 5. After tried both I decide to use 4. After the cluster numbers are assigned, a 3-d plot is generated in Figure 11 (you can check my source code to see the animation). Due to the 10-page limitation, some analysis based on the clusters are skipped here, you can check my source code where I performed some.

As a result,

Cluster 1: Moderate income, moderate consumption amount per purchase, moderate purchase frequency.

Cluster 2: Very high income, very high consumption amount per purchase, high purchase frquency.

Cluster 3: High income, high consumption amount per purchase, high purchase frquency.

Cluster 4: Low income, low consumption amount per purchase, low purchase frquency.

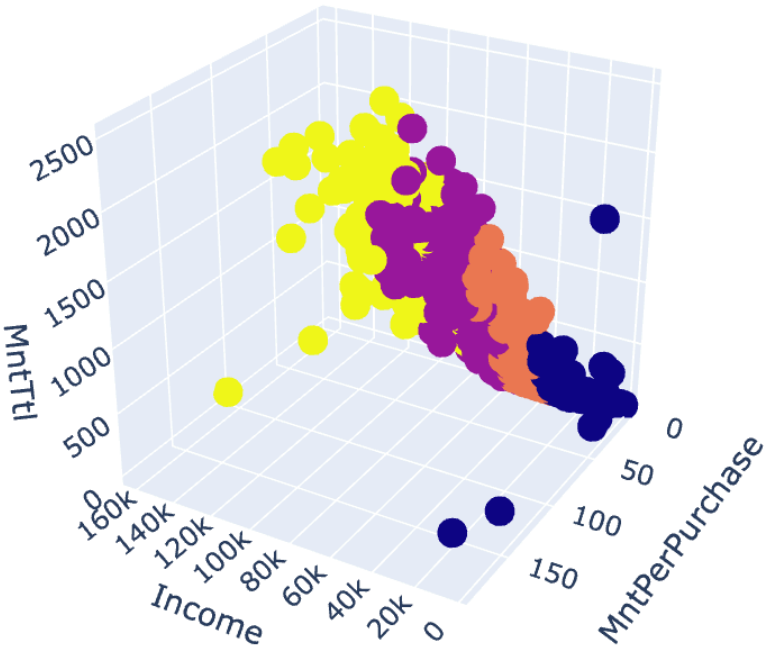

Figure 11: Clusters of customer.

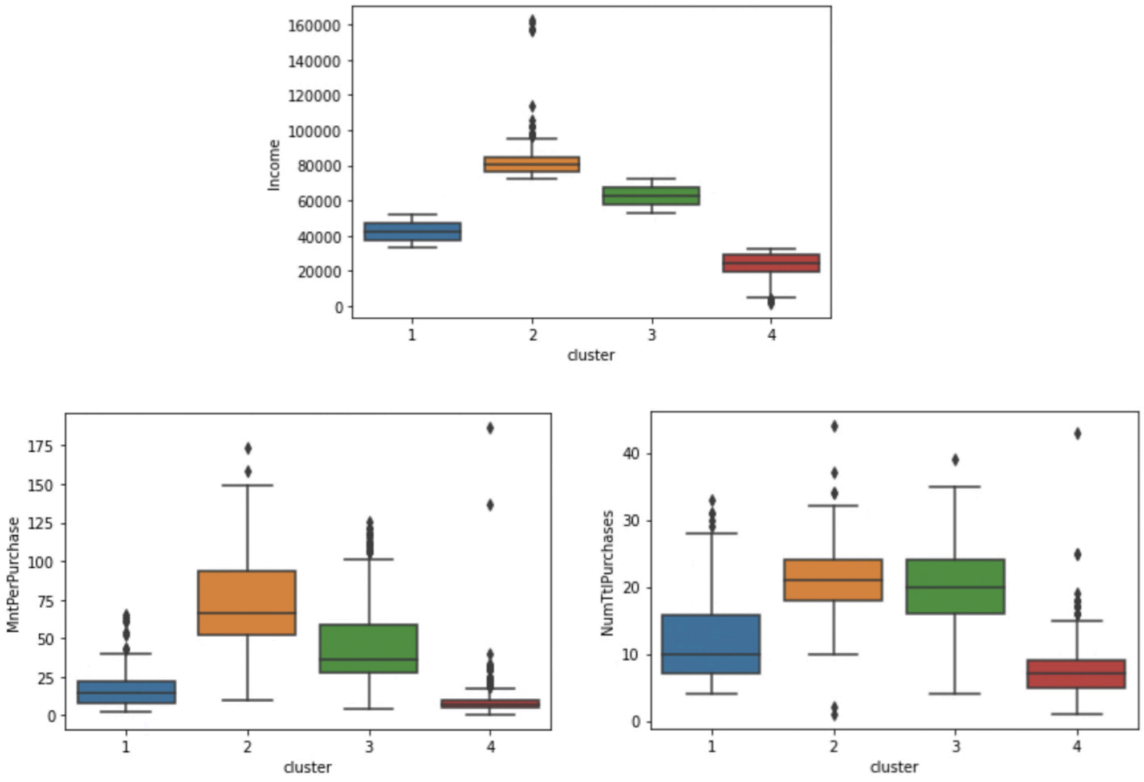

Figure 12: Barplots of Income, MntPerPurchase, NumTtlPurchases vs. cluster.