Stage 1: Get Appropriate Data

Description of data source and data

The data set refers to customers records of a grocery firm in the last 2 years, provided by Dr. Omar Romero-Hernandez and extracted from Kaggle (Kaggle, 2021).

The data set is in csv. format which contains 27 variables and 2240 rows in total (Appendix 1). There are 2 CLASS type and 25 VAR type variables (Appendix 2). Within the 25 VAR type variables, 7 of them are binary variables.

There is only 1 variable (Income) has 1.071429% of missing value which is 24 rows of observations.

$$\frac{Number\ of\ observation\ that\ contains\ missing\ value}{2240} \times 100 = 1.071429$$

$$Number\ of\ observation\ that\ contains\ missing\ value = 24\ rows$$

Description of variables

Variable	Description
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 enrolment 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
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
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

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

Stage 2: Define a Business Problem and Goal

Business Problem

How to identify different type of customers by grouping customers based on their spending on each product in order to provide suitable marketing strategies for each type of customers?

Business Goal

As customers are considered as an important asset of a business, the grocery firm has to generate insights about their existing customers. By segmenting customers into different segments and examining on their dissimilar spending behavior across each segment, the firm is able to design the most suitable marketing strategies for each segment.

Thus, the business goal of the grocery firm is, by considering the similarities of the amount spent on each product in K-mean clustering, the grocery firm is able to identify different type of customers then tailor the best marketing strategies to each type of customers. The marketing strategies aim to boost the customer lifelong value and to improve customer acquisition.

Stage 3: Select Inputs

As the grocery firm's goal is to group customers based on the spending on the variety of products, therefore, the inputs below are selected.

Inputs selected (Numeric)	Reason chosen
Products	
MntWines	To cluster customers based on amount spent on Wines
MntFruits	To cluster customers based on amount spent on Fruits
MntMeatProducts	To cluster customers based on amount spent on Meat
MntFishProducts	To cluster customers based on amount spent on Fish
MntSweetProducts	To cluster customers based on amount spent on Sweet
MntGoldProds	To cluster customers based on amount spent on Gold

Other inputs are rejected as they are not related to the business problem.

Examine on Inputs Missing Value

All of the selected inputs have no missing value (Appendix 2).

Examine on Variable Independence

Variable correlation is used to examine the independence between all selected inputs by using variable clustering node.

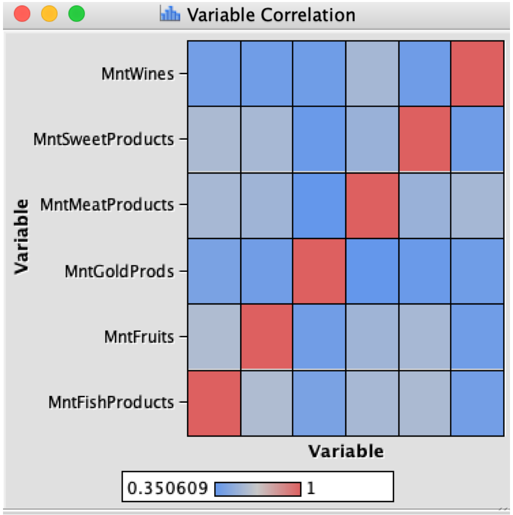


Diagram 1

The chart shows almost blue which means all of the selected inputs are independent to each other. The red box is the correlation between variables and itself which is perfectly correlated, correlation = 1.

Examine on Outliers

By exploring the histogram charts of MntMeatProducts and MntSweetProducts, it is found that the outliers exist in these inputs. Outliers are pointed out in red circle at below.

1. MntMeatProducts

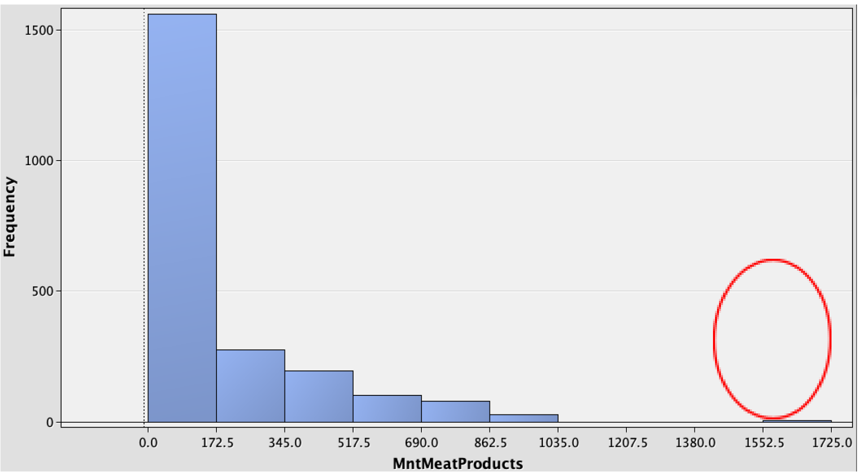


Diagram 2

The outliers exist between 1552.5 and 1725 with a frequency of 5. The outliers might considered as exist by chance. However, we still filter them due to 3 considerations. First, it is a small amount of outliers out of 2440 observations. Second, the outliers are too far away from the data. Third, filtered due to the high-sensitivity nature of K-means to outliers. Thus, the outliers are filtered to prevent disturbing on the final outcome.

The outliers are filtered by adding filter node in the diagram workspace. An upper limit of 1550 is set to eliminate the outliers from 1552.5 to 1725. The Default Filtering Method for interval variable is changed to be “User-Specified Limit”.

2. MntSweetProducts

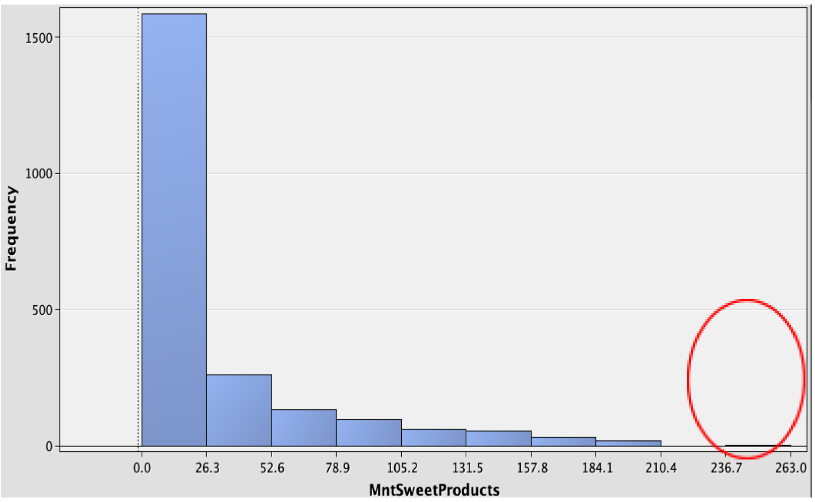


Diagram 3

The outliers exist between 236.7 and 263 with a frequency of 2. In this case, the outliers are treated carefully and did not filter out. This is because the outliers are close to the data. Therefore, the probability of existing by chance is greater. Simply filtering the outliers might generate untrusted insight.

As result, the number of observations is changed from 2240 to 2235 (Appendix 3).

Examine on Inputs Skewness and Kurtosis

After creating histogram graph for each selected input, we found that they are all skewed.

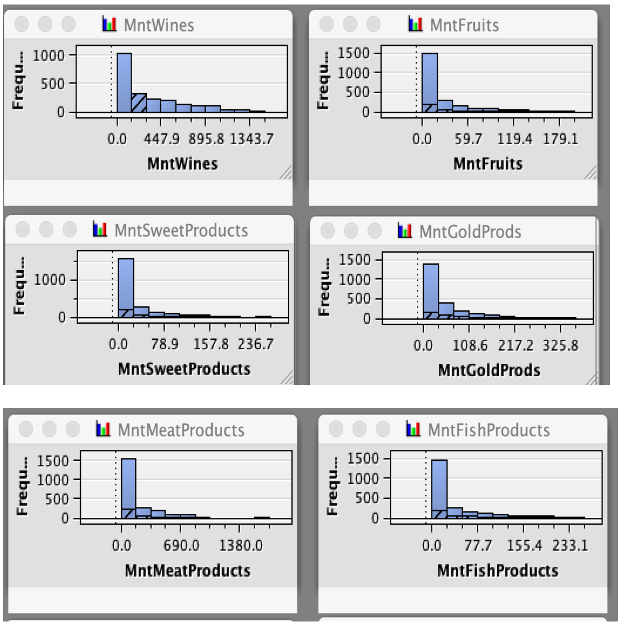


Diagram 4

According to SAS Help Center (n.d), transform variable node functions to transform the skewed interval variable using standard transformation. Therefore, a transform variable node is added and applying log transformation to the selected inputs. The default method of interval inputs is set to "Yes" (Appendix 4).

Inputs	Problem	Remedies
MntWines	Positively-skewed	Log Transformation
MntFruits	Positively-skewed	Log Transformation
MntMeatProducts	Positively-skewed	Log Transformation
MntFishProducts	Positively-skewed	Log Transformation
MntSweetProducts	Positively-skewed	Log Transformation
MntGoldProds	Positively-skewed	Log Transformation

Log transformation creates a new variable by taking the natural log of each original input variable. Therefore, after applying log transformation on the variables, a new set of variables is created where the name of variables is started with "LOG_" (Appendix 5).

Besides, the skewness and kurtosis of selected inputs has reduced to relatively symmetrical distribution (2 decimal places is taken in this table) (Appendix 5). The absolute transformed values of skewness are smaller than the original data set.

Variables	Original Skewness, Kurtosis	Transformed Skewness, Kurtosis
MntWines	1.17, 0.59	-0.55, -0.84
MntFruits	2.10, 4.04	0.08, -1.13
MntMeatProducts	1.73, 2.34	-0.09, -1.07
MntFishProducts	1.92, 3.08	-0.05, -1.09
MntSweetProducts	2.13, 4.36	0.08, -1.15
MntGoldProds	1.89, 3.54	-0.34, -0.41

At this stage, all the inputs selected fulfil the requirement of K-mean clustering where they are meaningful to the analysis objective, relatively independent, limited in number, numeric, and low kurtosis, and skewness.

Stage 4: Construct Clusters

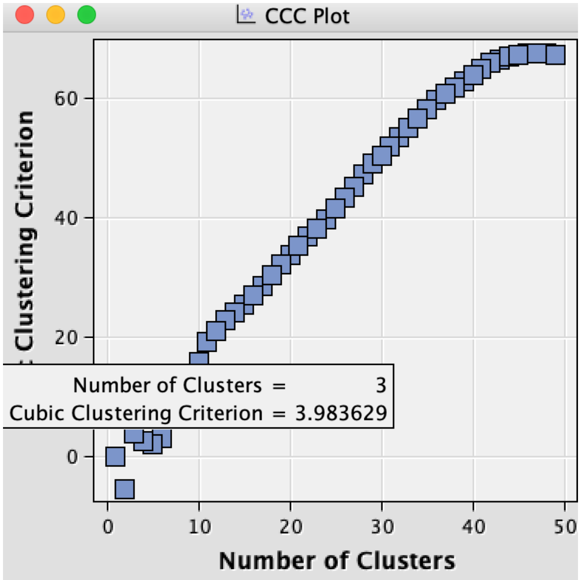
According to article "Customer-centric and Consumer-driven Brands" (Roll, 2017), in real world, it can be said that the success of a business is driven by 3 main abilities of a firm. First, the ability to retain loyal customers. Second, the ability to convert average spender customers to active customers. Third, the ability to stimulate low spending customers. Therefore, it can be assumed that we can classify the customers into 3 main types.

As the grocery firm's goal is to identify different type of customers to perform marketing approaches, thus, the customers can be clustered by applying the concept.

Please note that this is not the final decision on how the firm is going to name the different cluster group. It is just an assumption made in order to decide the number of clusters (k) where we usually do not know at the initial stage and more likely to adopt a repeating try and error process by using different k value. Thus, at this stage, the number of clusters chosen is 3, k=3.

In order to ensure that the number of clusters is optimal, there are two alternative cross checking methods.

First, CCC plot from Cluster Node can be used to cross checking the optimal number of clusters. The process of creating CCC plot is running the cluster node, go to View, click Summary Statistics, and choose CCC plot.



Number of Clusters = 3
Cubic Clustering Criterion = 3.983629

Diagram 5

According to Fernandez (n.d), values of CCC larger than 2 or 3 indicates good clusters. For k=3, CCC = 3.98 which is larger than 3. This indicates that 3 is an optimal number of clusters. Therefore, we decide to choose k = 3 in this project.

Secondly, running segment profile node to get the final result right after deciding number of clusters is an alternative method to see whether the number of clusters chosen is suitable to the data set or not. This is because we usually do not know the optimal number of clusters at the initial stage, we are more likely adopting try and error repeating process. We try prespecified k value and running segment profile node to see whether the segments graph looks good with interpretable information and without a lot of empty clusters in a segment. In this case, segmentation graph with k=3 generates graph without empty cluster. Therefore, it is believed that the number of clusters chosen k=3 is a suitable number due to all the above reasons.

There are several settings in properties panel of cluster node to be examined.

1. “Standardization” is selected as Internal Standardizations. This can reduce the effect of the inputs that have large variances on the distance measure (SAS Help Center, n.d). Besides, standardization ensure the Euclidean distance between each data points can be calculated to produce meaningful clustering by standardizing the measurement scale of all selected inputs.

2. Lastly, select “User Specify” for Specification Method and change the maximum number of clusters to 3. The number of clusters is set to 3 due to the reasons discussed above which are firstly, k=3 is the most suitable for this data set after performing CCC Plot. Second, it suits the interest to solve the business problem with the support from research article.

After running the cluster node, a cluster diagram window appears with segment plot, mean statistics, segment size and output are generated. Segment size and mean statistics are focused on this stage.

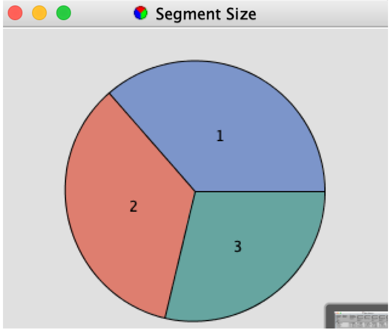
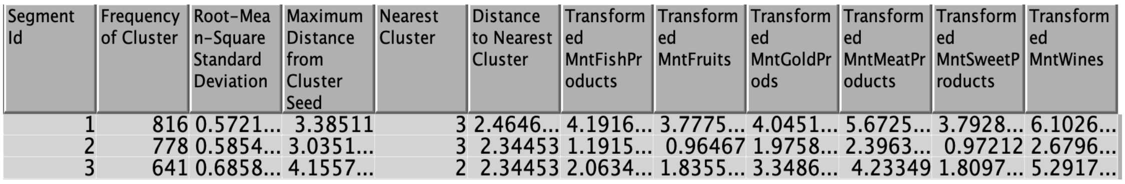


Diagram 6

1. Segment size
   By referring to diagram 6, the number of segments is consistent with the chosen number of clusters which is 3. It is found that segment 1 and 2 looks similar in size while segment 3 is slightly smaller. It indicates that the sample data in segments 1 and 2 does not vary much in term of number while the sample data in segment 3 is the least.

2. Mean Statistics
   We can get to know the hints of how the data are grouped together.



Segment Id	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	Transformed MntFishProducts	Transformed MntFruits	Transformed MntGoldProds	Transformed MntMeatProducts	Transformed MntSweetProducts	Transformed MntWines
1	816	0.5721...	3.38511	3	2.4646...	4.1916...	3.7775...	4.0451...	5.6725...	3.7928...	6.1026...
2	778	0.5854...	3.0351...	3	2.34453	1.1915...	0.96467	1.9758...	2.3963...	0.97212	2.6796...
3	641	0.6858...	4.1557...	2	2.34453	2.0634...	1.8355...	3.3486...	4.23349	1.8097...	5.2917...

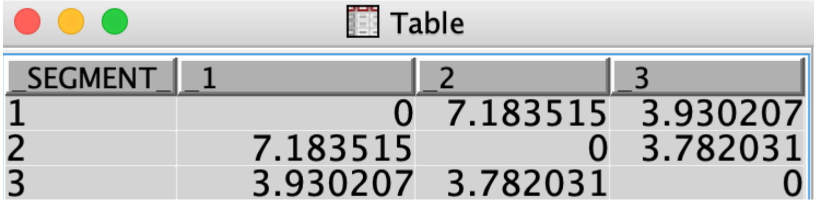
Diagram 7

The frequency of cluster is consistent with segment size where segment 1 is the largest, followed by segment 2 and 3.

In overview, firstly, in segment 1, amount spent on all products are the highest among the three segments. Secondly, in segment 2, the amount spent on wines is relatively high although the amount spent on other products is the lowest among the three segments.

Thirdly, it is interesting that unlike segment 1 and 2 where the amount of spent on each product are quite similar (high for all products or low for all products), segment 3 is having low and high spending amount on different product. For example, the amount spent on fruits and sweet products are considered low while the amount spent on wines is considered quite high.

3. Cluster Distance



SEGMENT	_1	_2	_3
1	0	7.183515	3.930207
2	7.183515	0	3.782031
3	3.930207	3.782031	0

Diagram 8

The table indicates that distance between segment 1 and 2 is the largest while the distance between segment 1 and 3 & segment 2 and 3 is similar.

Stage 5: Segment Profiling

In this stage, segment profile node is added into diagram workspace to perform segment profiling. The segment profile node is connected to cluster node as the diagram shown below.

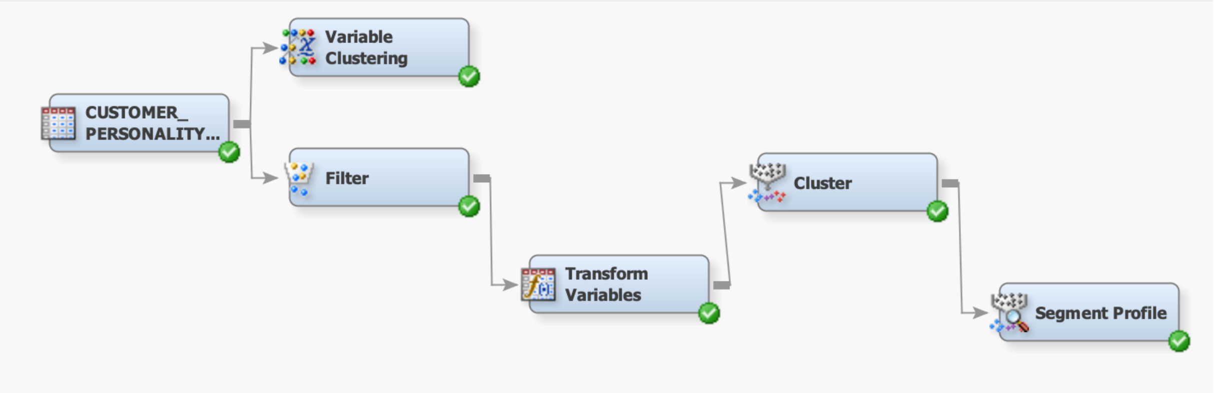


Diagram 9

By examining on segment, we are not only able to know how the data set is differentiated via looking at the characteristics of each segment but also can understand how the customers segment behave differently compared to population.

Procedures of segment profiling:

1. Change the setting in the properties panel in order to generate preferred segment graph
2. Running the segment profile node for first time to see whether the number of clusters defined generate insightful graph

3. If it is not an insightful graph, then redecide number of clusters

For step 1, there are the settings applied to the properties panel:

1. At the Train Properties (General), the number of midpoints is set to "8" to specify the number of bins in the histogram which shows the distribution of sample and population data.
2. At the Train Properties (Input Variables), the number of inputs is set to "6" as there are 6 variables selected (Transformed MntFishProducts, Transformed MntMeatProducts, Transformed MntFruits, Transformed MntSweetProducts, Transformed MntWines, and Transformed MntGoldProds).

For step 2, there are 3 segments created after running the segment profile node. The variable for each segment is sorted from left to right based on their importance within each cluster (Chuang, 2020). The distribution for the sample data of the segment is shown in blue while the distribution for the population data is shown in red. The graphs generated is insightful and interpretable to answer the business questions.

Segment 1

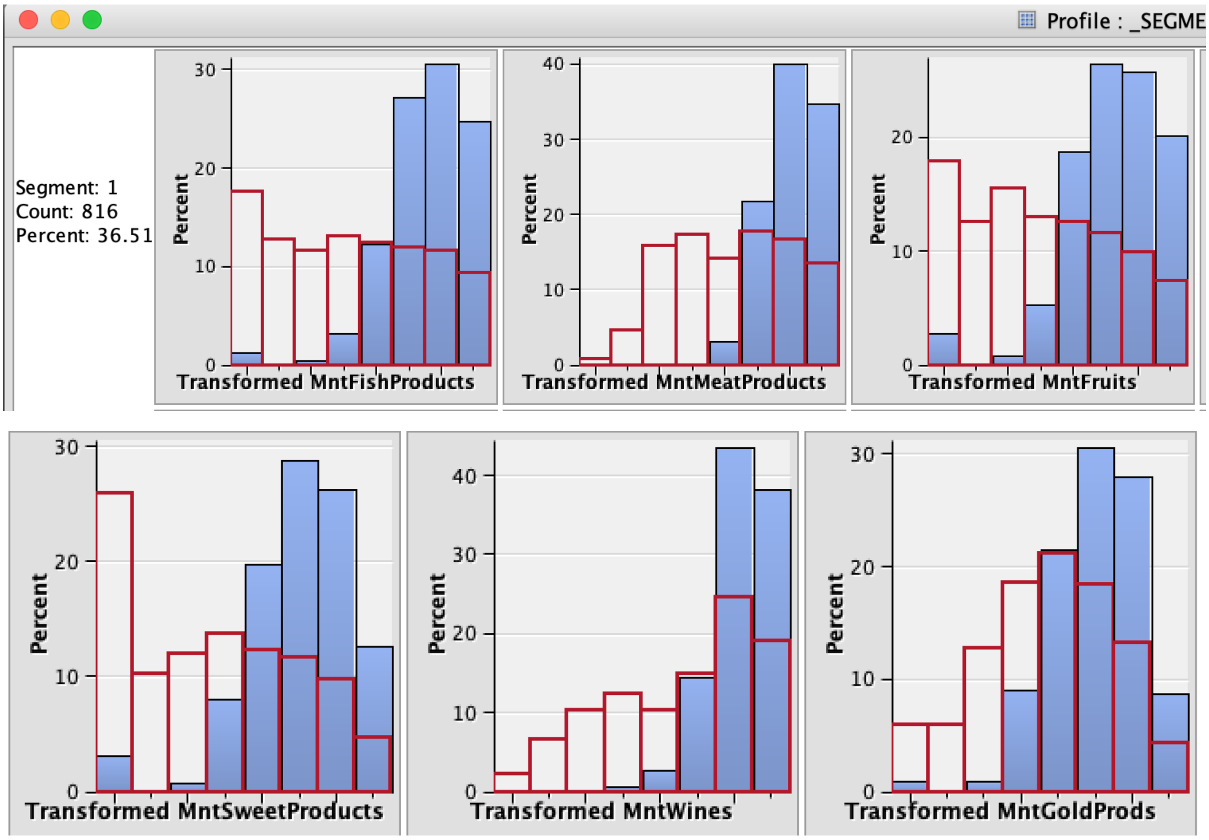


Diagram10

Interpretation

The segment 1 represents 36.51% of the population.

Compared to the overall distribution, segment 1 has a higher amount spent on all products. The product where the customers in this segment spent the most money on is wines, followed by meat, fish, fruits, sweet and gold. The percentage of the customers spend within both top right volume ranges – volume ranges 7 and 8 are 81%, 74%, 54%, 45%, 38% and 35% approximately.

Overall, this segment has more customers spend in higher volume ranges than the overall average. Thus, we can say that the amount spent on all product is higher than average.

Characteristics

- High spending on all types of products
- Wines is the most popular and profitable product where the customers spend the most money on, followed by meat, fish, fruits, sweet, and gold.

Customer Type

We can categorize this group of customers as loyal customer who usually revisits the selling platform to purchase any products they want. They contribute most of the portion of sales among all type of customers.

Segment 2

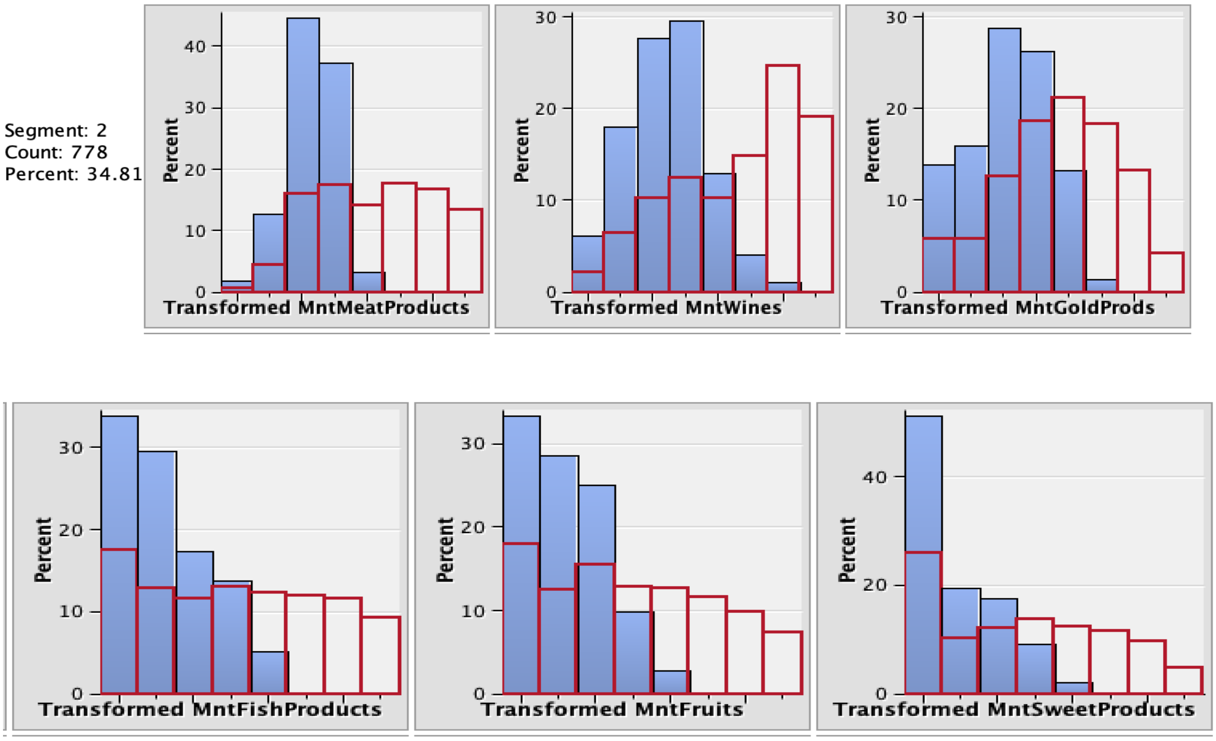


Diagram11

Interpretation

The segment 2 represents 34.81% of the population.

In overview, majority of customers' spending on all product is below the overall average.

The dispersion of amount spent on fish, fruits and sweet is smaller than the overall distribution. This indicates that it is very concentrated that the customers did not purchase or very less likely to purchase on these products from this grocery firm. Besides, customers spent slightly more on wines and meat which are approximately 41% and 40% in the volume ranges 4 and 5.

Overall, this segment has more customers spend in lower volume ranges than the overall average. Thus, we can say that the amount spent on all products is lower than average.

Characteristics

- Low spending on all types of products
- Spending on wines and meat is slightly higher compared to the spending on other product
- Obviously very less likely to purchase fish, fruits, and sweet product.

Customer Type

We can categorize this group of customers as wandering customers. Wandering customers are also known as “Just Looking Around” customers. They visit store for no specific purpose and just casually browsing. This causes low amount of spent on all products than overall average. They generally will leave alone without buying anything but sometimes they may buy the product that caught their eyes or what they like. This is supported by the scenario that customers spent slightly more on wines and meat.

Segment 3

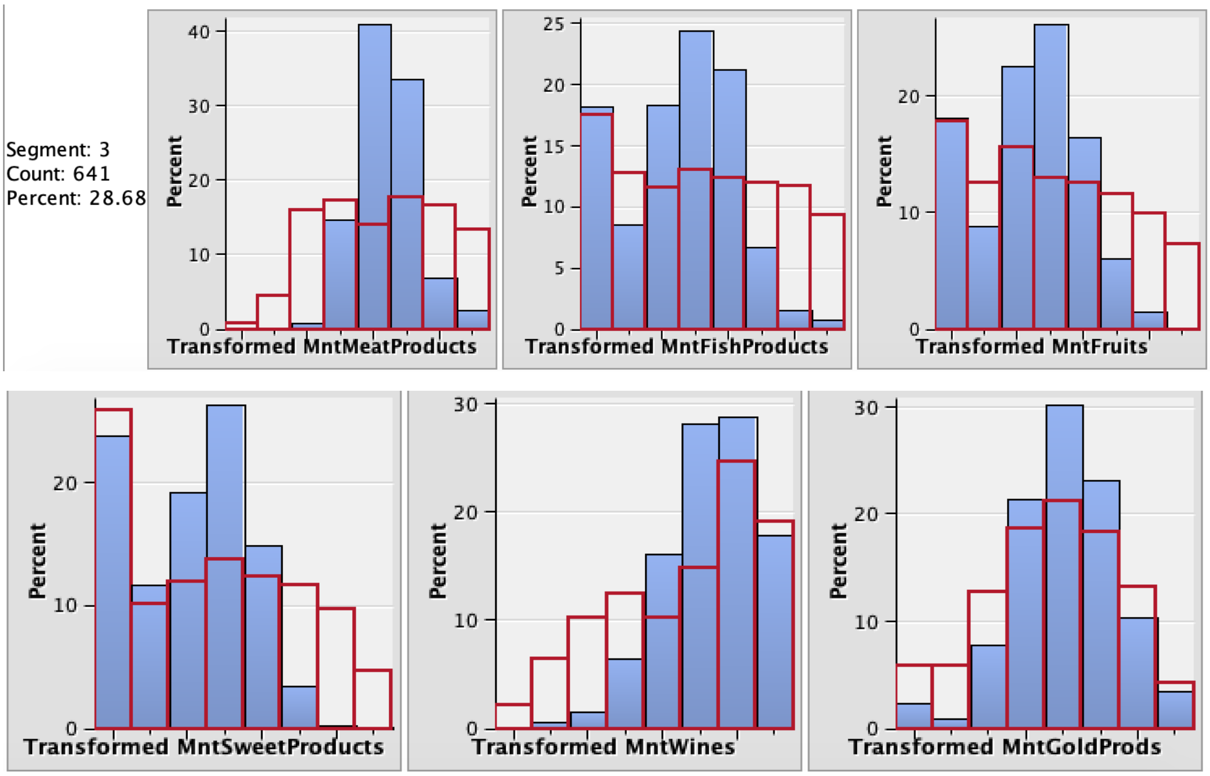


Diagram12

Interpretation

The segment 3 represents 28.68% of the population.

In overview, majority of customers are less likely to spend on fish, fruits, and sweet. The percentage of the customers spend on fish, fruits, and sweet within volume ranges 1 and 2 are 26%, 26%, and 34% approximately. Besides, the customers moderate likely to spend on meat and gold. The percentage of customers spend on meat and gold within volume ranges 5 and 6 are 74% and 53% approximately. Lastly, the customers high likely to spend on wines. The percentage of customers spend on wine within volume ranges 7 and 8 is 45% approximately.

It is surprisingly to find out that the customers in this segment did not have same purchasing behavior on each product, instead, there are different spending levels on different products. For example, majority of customers less likely to spend on fish, fruits, and sweet; moderately spend on meat and gold; and highly spend on wines.

Overall, this segment has more customers spend in lower volume ranges than the overall average for fish, fruits, and sweet. Besides, this segment has more customers spend in middle volume ranges for meat and gold. Lastly, this segment has more customers spend in higher volume ranges for wines. Thus, we can say that the amount spent on fish, fruits, and sweet products is lower than average while amount spent on wines is higher than average.

Characteristics

- Different spending level on different products
- High spending on wines, middle spending on meat and gold and low spending on fish, fruits, and sweet product.

Customer Type

We can categorize this group of customers as need-based customer. Need-based customers are those who have intention and preference to buy specific products from same firm to which they are habitual (Indeed Editorial Team, 2021). Thus, they behave differently onto different products where amount spent on specific product is high and amount spent on specific product is low. In this case, the specific products that tends to be purchase by the customers are wines, meat, and gold.

Stage 6: Implement Segmentation

After segment profiling, we have identified 3 types of customers. They are loyal customers, wandering customers and need-based customers. This has answered the business problem as different types of customers have been identified. Next, appropriate marketing strategies should be customized for each type of customers to improve the business performance and saving costs of the grocery firm.

Segment 1: Loyal customers

Loyal customer is the most valuable asset of the grocery firm, and they should be given the most concern on. This is because they have already exhibited brand loyalty on the firm and make the firm as preference grocery store by revisiting the stores frequently and spending the highest amount on all of the products. In result, they are less likely to be attracted by other competitors. Individual attention is their main concern as it makes them feel like they are valued and tends to stay loyal to the firm.

First, implementing loyalty tiers and reward programs is a practical way to retain the loyal customer. For example, a digital mobile passes shopping card is only introduced to every loyal customer. The customer will earn points every payment by scanning the QR code of the card. When the points collected reach a certain amount, the customers can level up to higher loyalty tier with greater rewards benefits than the one before such as larger amount of e-vouchers are rewarded. The customers could use the e-voucher and get a loyalty discount. As wines is the most popular product within the loyal customers, the e-voucher given could be a wine voucher since wines is pricy. This gives incentive to the loyal customers to purchase more products and spend more from the firm in order to collect the points for wine voucher.

Second, highest spending on wines among all the products might indicate that most of the loyal customers are wines drinkers. Therefore, wines sales associate should be employed for both in store and online website to introduce variety of wines to the customers and assist the customers to choose the wines that suit their preferences. As the customers satisfy with the professional and excellent sales services, they will be more likely to purchase repeatedly from the firm. This can boost the customers loyalty.

Segment 2: Wandering customer

Wandering customer is also known as “window shopper” who usually browsing and do not have particular product or brand in mind (Indeed Editorial Team, 2021). They are not looking for a specific product at most of the time and what attract them is the shopping experience itself and the interaction with other people. However, sometimes, they might purchase something that gain their attention at that moment.

By referring to the graph segment above, it is suggested that the grocery firm should not invest too much time, energy, and money to deal with this customer group due to their smallest portion of contribution to the sales revenue and the nature of without intention to buy any products.

Although dealing with wandering customer is a challenging task, but it is not an impossible task. First, wandering customer enjoys interacting with people, sales associate is suggested to be employed to increase their probability to make purchase. By referring to the segment graph above, the wandering customers still spent a portion of money on purchasing wines.

Thus, wines sales associate should be employed. Therefore, the wines sales associate that actually employed for dealing with the loyal customers can serve the wandering customers as well. This saves the costs of the firm where two types of customers could be approached by same employees.

Second, it is recommended that the grocery firm could offer in store wines testing package. This might attract the attention of wandering customers and they might show interest on purchasing wines. The wines sales associate could explain the wines knowledge and provide insightful information to them. Thus, they might purchase after testing which will increase the sales revenue of grocery firm.

Segment 3: Need – based customer

Need – based customer is the customers that only intends to purchase specific products (Corporate Finance Institute, n.d). They usually purchase according to their needs and want the best products that meet their needs. They usually have already decided on what to buy in mind before shopping.

By referring to the segment graph above, most of the need-based customers spent more on wines, meat, and gold products. Therefore, it can be said that the grocery firm's wines, meat, and gold products are preferred by the need – based customers. Wines, meat, and gold products are their target products.

First, the grocery firm should ensure high quality of wines, meat, and gold products are provided. This is because the customer only purchases based on their needs without any brand preference, quality is their only concern. For example, variety of wines and fresh meat should be provided. Otherwise, they will switch to other grocery firm easily if their quality expectation has not been met. Furthermore, it is crucial that the photo of variety of wines and fresh meat has been uploaded onto grocery firm's website to aware the customers as need-based customers usually browsing through website to see whether their desired products are available.

After ensuring having rich variety of wines and fresh meat product on shelves, advertising is taking the turn now. Undoubtedly, social media platform is an effective tool in this era of digitalization to promote products. The grocery firm is recommended to consistently create the advertisement posts for wines and meat with some eye-catching slogan with emphasize on quality. For example, "The True Enjoy" slogan can be used to promote wines. Besides, "Meat The Taste Better", "The Best Meat In Town", and "100% Original" slogan can be used to promote the fresh meat. By promoting on social media with emphasize on quality, it can catch the attention of need-based customers.

Thirdly, in order to retain the need – based customers, the grocery firm should ensure wines, meat, and gold products are arranged properly according to the products type in store for ease of finding purpose. This is because the need-based customers only looking for their needs, if the products are not available, they tend to leave stores quickly. Therefore, it is suggested that the boards which written "Alcoholic Beverage", "Meat", and "Gold Product" should be provided for customers. Besides, the grocery firm could carry out in detail analysis to analyze what is the particular wines, meat, and gold product they preferred. Thus, the particular wines, meat, and gold product could be placed at eye level to entice the customers. This will satisfy the need – based customers as it saves their time looking for products and shorten their shopping time.

Conclusion

In overall, by examining on all three types of customers, it can be found that wines and meat are the products where all the customers spent the most on. Therefore, it is crucial that the quality and stock of these two products should always be guaranteed as they contribute the most on the firm's sales. In results, all the customers will be happy with their purchasing experience and tends to build long-term relationship with the grocery firm.

As a conclusion, the grocery firm should implement the customized strategies that derived from segmentation results for each type of customers in order to maximize the profit margin, improve customer acquisition and boost customer lifelong value which are the firm's business goal.

Stage 7: Limitation and Future Works

There are 1 limitation from the algorithm and 1 limitation from data in this project.

First, the limitation from K-means algorithm is the data set must be low skewness and kurtosis due to its high sensitivity nature to outliers. This is because in K-means, the centroid of kth clusters containing the mean values of all data points (Chauhan, 2022). Thus, the centroids will be influenced by outliers (Gan & Ng, 2017). The ideal data set is the data set with perfect symmetric variables where the skewness and kurtosis = 0. Therefore, any skewed variables need to be transformed before clustering. It can be said that, sometimes, by simply transforming skewed variables is not considered as a good approach as the transformed data might not be relevant for the original data (Feng, Wang, Lu, Chen, He, Lu & Tu, 2014). It can be said that log-transformation implication might act like throwing the critical piece of data and produce a misleading outcome.

In order to address this limitation, another unsupervised learning clustering algorithm which known as Density-based spatial clustering of applications with noise (DBSCAN) is recommended to cluster the data set with skewed distribution. This is because DBSCAN has a notion of noise which is robust to skewed data set. Similar to K-means, DBSCAN groups points that are close to each other based on Euclidean distance measurement. However, what makes DBSCAN stands out is the outliers are marked as points in low-density region which cause the outliers could not make great influence on the outcome (Prado, 2017). Thus, it can be said that DBSCAN is able to identify clusters of different shapes and size from a skewed data set with noise and outliers.

Secondly, the limitation from the data set is the data set is outdated and it is a limited sample size data. The data is outdated as the amount spent on all products by the customers is collected from the past 2 years. Besides, the finding generated from a limited sample size data might be misleading. It is difficult to determine that if the finding is true. Thus, the marketing strategies recommendations might not be suitable to apply on the current situation.

In order to address this limitation, the grocery firm has to collect a large new data set from the current customers records. Thus, a finding generated will tends to be more related to the current situation and the marketing strategies suggested will be more applicable.

Appendix

Appendix 1

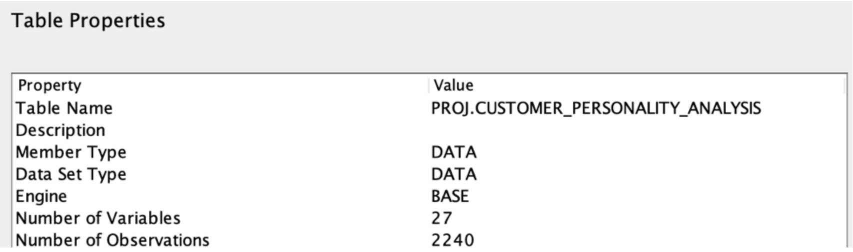
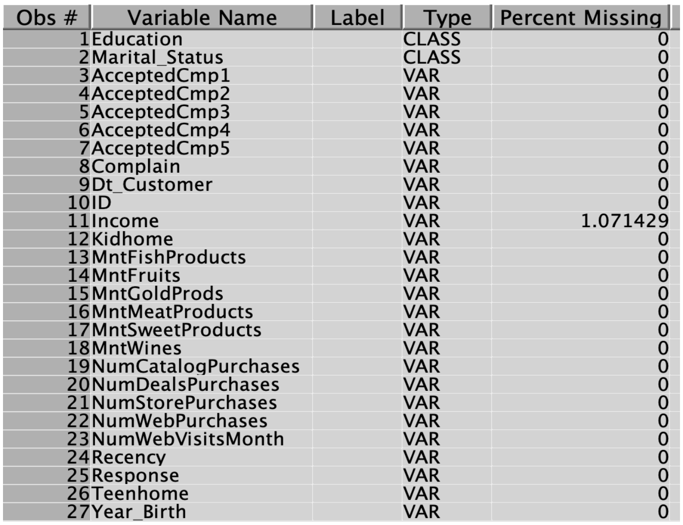


Table Properties

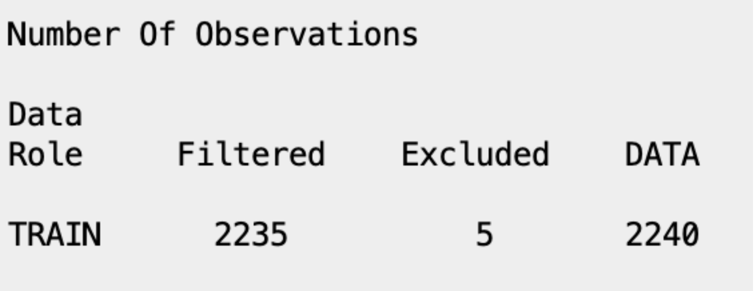
Property	Value
Table Name	PROJ.CUSTOMER_PERSONALITY_ANALYSIS
Description	
Member Type	DATA
Data Set Type	DATA
Engine	BASE
Number of Variables	27
Number of Observations	2240

Appendix 2



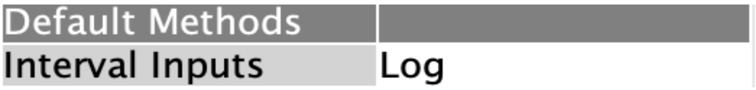
Obs #	Variable Name	Label	Type	Percent Missing
1	Education		CLASS	0
2	Marital_Status		CLASS	0
3	AcceptedCmp1		VAR	0
4	AcceptedCmp2		VAR	0
5	AcceptedCmp3		VAR	0
6	AcceptedCmp4		VAR	0
7	AcceptedCmp5		VAR	0
8	Complain		VAR	0
9	Dt_Customer		VAR	0
10	ID		VAR	0
11	Income		VAR	1.071429
12	Kidhome		VAR	0
13	MntFishProducts		VAR	0
14	MntFruits		VAR	0
15	MntGoldProds		VAR	0
16	MntMeatProducts		VAR	0
17	MntSweetProducts		VAR	0
18	MntWines		VAR	0
19	NumCatalogPurchases		VAR	0
20	NumDealsPurchases		VAR	0
21	NumStorePurchases		VAR	0
22	NumWebPurchases		VAR	0
23	NumWebVisitsMonth		VAR	0
24	Recency		VAR	0
25	Response		VAR	0
26	Teenhome		VAR	0
27	Year_Birth		VAR	0

Appendix 3



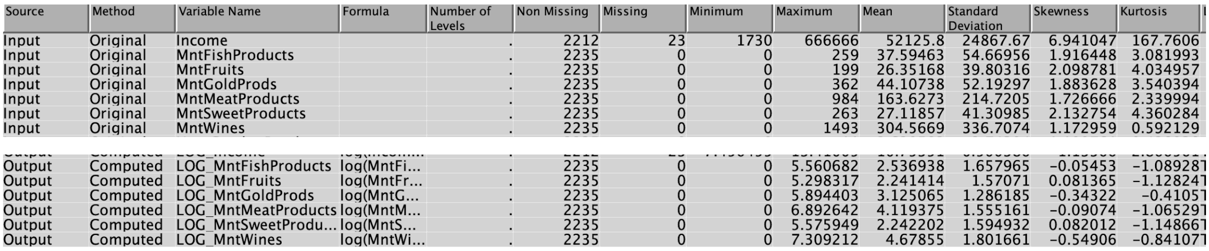
```
Number Of Observations

Data
Role      Filtered    Excluded    DATA

TRAIN       2235          5       2240
```

Appendix 4



Default Methods	
Interval Inputs	Log

Appendix 5



Source	Method	Variable Name	Formula	Number of Levels	Non Missing	Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
Input	Original	Income		.	2212	23	1730	666666	52125.8	24867.67	6.941047	167.7606
Input	Original	MntFishProducts		.	2235	0	0	259	37.59463	54.66956	1.916448	3.081993
Input	Original	MntFruits		.	2235	0	0	199	26.35168	39.80316	2.098781	4.034957
Input	Original	MntGoldProds		.	2235	0	0	362	44.10738	52.19297	1.883628	3.540394
Input	Original	MntMeatProducts		.	2235	0	0	984	163.6273	214.7205	1.726666	2.339994
Input	Original	MntSweetProducts		.	2235	0	0	263	27.11857	41.30985	2.132754	4.360284
Input	Original	MntWines		.	2235	0	0	1493	304.5669	336.7074	1.172959	0.592129
Output	Computed	LOG_MntFishProducts	log(MntFi...	.	2235	0	0	5.560682	2.536938	1.657965	−0.05453	−1.08928
Output	Computed	LOG_MntFruits	log(MntFr...	.	2235	0	0	5.298317	2.241414	1.57071	0.081365	−1.12824
Output	Computed	LOG_MntGoldProds	log(MntG...	.	2235	0	0	5.894403	3.125065	1.286185	−0.34322	−0.4105
Output	Computed	LOG_MntMeatProducts	log(MntM...	.	2235	0	0	6.892642	4.119375	1.555161	−0.09074	−1.06529
Output	Computed	LOG_MntSweetProdu...	log(MntS...	.	2235	0	0	5.575949	2.242202	1.594932	0.082012	−1.14866
Output	Computed	LOG_MntWines	log(MntWi...	.	2235	0	0	7.309212	4.67855	1.801661	−0.54906	−0.84107

References

Chauhan, N. (2022). DBSCAN Clustering Algorithm in Machine Learning. Retrieved from https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html

Chuang, L. (2020). Segmentation Analysis using SAS Miner. Retrieved from https://medium.com/luca-chuangs-bapm-notes/segmentation-analysis-using-sas-miner-43758d8f9863

Corporate Finance Institute. (n.d). Types of Customers. Retrieved from https://corporatefinanceinstitute.com/resources/knowledge/other/types-of-customers/

Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., & Tu, X. M. (2014). Log-transformation and its implications for data analysis. Shanghai archives of psychiatry, 26(2), 105–109. Retrieved from https://doi.org/10.3969/j.issn.1002-0829.2014.02.009

Fernandez, G. (n.d). Statistical Data Mining Using SAS Applications. Retrieved from https://books.google.com.my/books?id=hWnMBQAAQBAJ&pg=PA128&dq=Values+of+the+cubic+clustering+criterion+greater+than+2+or+3+indicate+good+clusters&hl=en&sa=X&ved=2ahUKEwigionGlqr3AhUiTGwGHcjBCugQ6AF6BAgLEAI#v=onepage&q=Values%20of%20the%20cubic%20clustering%20criterion%20greater%20than%202%20or%203%20indicate%20good%20clusters&f=false

Gan, G., & Ng, M. (2017). k-means clustering with outlier removal. Retrieved from https://doi.org/10.1016/j.patrec.2017.03.008

Indeed Editorial Team. (2021). 9 Types of Customers and How To Approach Them. Retrieved from https://www.indeed.com/career-advice/career-development/types-of-customers

Kaggle. (2021). Customer Personality Analysis. Retrieved from https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis

Prado, K. (2017). How DBSCAN works and why should we use it? Retrieved from https://towardsdatascience.com/how-dbscan-works-and-why-should-i-use-it-443b4a191c80

Roll, M. (2017). Customer-centric and Consumer-driven Brands. Retrieved from https://martinroll.com/resources/articles/branding/customer-centric-consumer-driven-brands/

SAS Help Center. (n.d). Data Preparation and Investigation. Retrieved from https://documentation.sas.com/doc/en/emcs/14.3/n06gx08zfc4qdjn17pew99rtx3or.htm