1. Wonderful Wines of the World

In the world of today it is important to understand and research all aspects of existing databases. Digitalization is not a strategy, it is a necessity for every company. One aspect that has seen a lot of development is the path to create new customers. WonderFul Wines of the World (WWW) has contacted us to get a better understanding of its customer base and how to target and create new customers.

WWW is an established wine seller that specializes in finding unique wineries from around the world. The goal of WWW is to deliver unique and interesting wines to their customers. Their operations started as a small store, but soon busted out of its seams. Currently they are operating from ten small stores across major cities in the USA. And besides physical stores, customers are also able to buy via phone or on the website. The customers can choose from several hundreds of carefully selected wines, which are presented every six weeks in a catalog. For 4 years WWW has been building a database and has gathered in total over 350.000 customers. These customers buy besides the wines also all kinds of accessories like wine racks or cork extractors. WWW wants to use their database and create a better understanding of their customer. WWW sees a lot of potential in creating new marketing strategies. Currently they do mass-marketing and do not differentiate in any way and see great opportunity to expand their marketing strategy.

WWW has provided a sample of 10.000 customers from its “active” database. Active means that these customers have purchased something in the last 18 months. The objective is to segment this active database in clusters on two levels. One segmentation is based on engagement and value of the customer and the second segmentation is based on the buying behavior of customers. These two segmentations should be represented in a single frame so that WWW understands the value of each customer and understands what kind of wines they will be the most interested to buy.

First we need to understand the given dataset and explore if the data given is relevant and coherent in the situation. After that we will prepare different feature sets for the two segmentations. For the modeling part we will use a variety of data mining techniques to create clusters. Lastly we will combine the findings of the two segmentations and suggest business applications from the derived knowledge.

2. Data understanding

The delivered dataset by WWW consists of 10.001 records. Each record has 18 features and all of the features have numerical values. When checking for missing values we found that the last record had a missing value for the feature Custid. Looking into this record it was clear that all other values in this record were the mean value of each column. For this reason the record had no added value and was deleted from the dataset. All features had the datatype float64. You could have argued for integer as prefered data type for some of the features, but it is not necessary to change this. Looking into the statistical summary of the dataset there were no extraordinary values, meaning that all of the values were in expected boundaries. Also there are no duplicate records found in the dataset.

In the provided description of the dataset it is stated that the columns Dryred, Drywh, SweetWh, Sweetred and Dessert combined should add up to 100. Each feature stance for a percentage of the sales corresponding to a wine type. The mean of these features add up to 100, but looking towards the individual records we found that only 7300 records are adding up to 100. After this finding, we created a boundary for the total features. We gave an error margin of 5%, with the lower bound being 95 and the higher bound being 105. Applying this gave a result that 100% of the records were in the margin and so there are no records deleted by this coherence check.

The sixth wine type, exotic, created some difficulties in the interpretation of its values. Just like the other wine type values it is an integer that represents a percentage of something. As we assumed that the other wine type feature values are interpreted as a percentage of total sales by the customer, this cannot hold up as an interpretation for the sixth wine type exotic. Due to the relevance of the wine type for WWW (WWW specializes in delivering unique (exotic) wines) we will keep the feature in our analysis and will try to give an appropriate interpretation in our conclusion.

3. Data preparation

After the data exploration we ended up with a dataset with 10.000 records and 18 features. The data preparation will be split in two parts. We are doing two different segmentations, which need different feature sets.

3.1 Data preparation for Customer value segmentation

See Appendix A for visualizations

For behavioral segmentation we used the features that best represent this segment, this includes 9 of 17 features provided. Before starting the analysis, we constructed a correlation matrix that shows how every variable is correlated with the other, the correlation matrix can be found in appendix A. We can use this matrix to omit a variable if it was highly correlated with another. In our case we can observe that the monetary feature is highly correlated with lifetime value of the customer (LTV) feature (0.99), and the frequency feature(0.93), meaning that multicollinearity is present. This negatively affects the clustering algorithms, however due to the importance of the monetary feature we decided not to omit the feature. Before clustering we standardized the data using the standard scaler, this is done so the data can be transformed to a similar unit of measurement, so we can have meaningful analysis.

3.2 Data preparation for Behavior segmentation

See Appendix B for visualizations

For this segmentation we will use features that give a possible insight into the customer buying behavior. The six different wine types are the more obvious features that we considered in the dataset. But beside these features we also found 3 other interesting features to take into account, namely Perdeal (% purchases bought on discount), WebPurch (% of purchases made on website) and Webvisit (Average visits to website per month). As stated before, all features had no unexpected values and had no missing values. The next step was to look for any correlation between the features. According to the correlation matrix found in appendix B, we found that WebVisits and WebPurchase are highly correlated (0.88) with each other. We chose to keep WebPurchase over Webvisit, since WebVisit could be seen as an explanation for customers doing a purchase on the web. Also the value of WebPurchase are similar as those of the other features which could benefit slightly during the modeling. Due to the relatively high correlation between Dryred and almost all other wine types (Correlations between -0.38 & -0.66), we created one feature set without the feature Dryred. Before the modeling we also scaled our dataset with the StandardScaler method. Each value is now centered around the mean with a unit standard deviation. Meaning positive values are a unit standard deviation above the mean and vice versa. In our testing phase we also tried to leave some features out in the modeling. Due to the relatively high correlation between Dryred and almost all other wine types (Correlations between -0.38 & -0.66),

4. Modeling

4.1 Customer value segmentation

Hierarchical Clustering for Customer Value

See Appendix C and D for visualizations

Our strategy was to take a bottom up approach to build the dendrogram. This allows us to have the number of clusters built out for us and we can then split the y threshold for the ideal amount of clusters. We tested other methods, but the ward method was the most effective. With the data scaled with standard scaling, it was input and the dendrogram built out. You can see there three ideal clusters are formed. From there, we took the cluster labels, grouped them back into an unscaled dataframe with the same features. Using bar plots, with the cluster as the x-axis, and the feature on the y-axis, we can draw some basic conclusions for customer value.

For instance, we can see that the youngest age cluster spends less money and are less frequent purchasers. They also purchase the most wines on discount. The oldest age cluster, averaging around 68 years old, is far and away the biggest spender, with the most income in order to spend that money on wine. They also don’t seem to spend a lot on discounted wines, as their purchases reflect a small percentage purchased on discount. However, the percentage of the youngest cluster purchases is mostly made online. Whereas the largest purchasing cluster made less than 40% of their purchases online.

The other cluster’s age averages about 51. They’re really in between the two clusters as far as spending and value goes. They don’t spend quite as much as the oldest cluster, but still spend more as a percentage than the oldest cluster in online purchases

The conclusions are clear to see, the older customers have a higher income, spend more money, and are more frequent customers thus making them more valuable customers. The question here is, how can we get them to spend more of their money towards online purchases?

K-means Clustering for Customer Value

See Appendix J, K, L and M for visualizations

To start the K-means clustering we constructed an inertia or an elbow method, it consists of plotting the explained variation as a function of the number of clusters, we can use it to find the optimal number of clusters. Unlike the dendrogram used in the hierarchical clustering that showed 3 clusters, The inertia shows that 2 clusters are the optimal number of clusters. We also calculated the silhouette score for the clusters, and it further confirms that 2 is the optimal number of clusters, since 2 clusters achieved the highest score, this can be shown in appendix J.

Since the K-means algorithm usually suffers to perform when there are many numbers of features, we used PCA To improve our clusters and reduce the dimensionality. After plotting the cumulative variance against the number of components, we were able to observe that 5 principal components were able to explain more than 95% of the variance when using all the 9 features, this is shown in appendix K. The Inertia and Silhouette score also showed that 2 clusters were optimal even after the PCA transformation. Compared to using the feature set without the dimensionality reduction, we can notice an increase of 2% in the silhouette score from using the data without the reduction.

We also used T-SNE to compare its results to the PCA, since both are features reduction methods, the silhouette score increased by 7% compared to the PCA, the Inertia and Silhouette score also showed that 2 clusters were still optimal after the T-SNE transformation, as shown in appendix L. We performed the K-means clusters based on the previous results, and plotted the 2 clusters that were created in 2D, this can be shown in appendix M. We can clearly see the distinct clusters, since they are not overlapping.

4.2 Behavior segmentation

See Appendix F, G and H for visualizations

For the segmentation of buying behavior we used data mining techniques like Hierarchical Clustering (HC), Kmeans and DBScan. In the testing phase we found out that DBScan was not a prefered method. The method created mostly two or three clusters. But the clusters consisted of a small percentage of the total number of records. We did not find this a prefered outcome and focused more on the other techniques.

Our strategy for behavior segmentation was similar to customer value segmentation. A bottom up approach where we used the dendrogram of the hierarchical clustering to determine the amount of clustering we are using for Kmeans. The linkage method used for the hierarchical clustering is determined by calculating the R-squared for different amounts of clusters and the different linkage methods. In appendix E you can see the results of this and based on this we will use the Ward method. As we see from the dendrogram in appendix F we conclude to use 3 clusters. To evaluate the amount of clusters used we also looked into an elbow graph. In appendix G you can see that an amount of two or three clusters is the prefered amount, which confirms our earlier findings in appendix F.

Lastly we performed the K-means with K = 3. In appendix H we see the table with mean values for each feature grouped by the different clusters.

Cluster 1: Buys mainly online (56%) and buys the most in discount (55%). Buys relatively few Dryred compared to the others clusters, but much more of Sweetred (2-5 times), Sweetwh (2-5 times), Dessert (2-5 times) and Exotic (2.5 - 4 times). This cluster consists of 2302 records (23% of total).

Cluster 2: buys mainly online (50%) and 40% in discount. They buy almost only Dryred wines (71%) and have little interest in wines like Sweetred, Sweetwh or Dessert. This cluster consists of 4351 records (44% of total).

Cluster 3: Buys mainly in physical stores (only 23% online) and buys very few in discount (8%). They buy relatively the most Drywh (35%) and have the least interest in Exotic wines (9%). This cluster consists of 3347 records (33% of total).

4.3 Customer Value & Behavior segmentation combined

See Appendix N,O,P and Q for visualizations

After combining our two different segmentation categories, we produced table N, from which we were able to derive six distinct customer segmentations, the distribution and density of which can be seen in appendices O and P respectively.

After grouping the data frame by the different permutations of clusters and dropping unnecessary features we were able to produce a stacked radar graph which displayed a graph for every combination of customer value and behavior, see appendix Q, which would then be unraveled and evaluated for marketing purposes amongst others.

5. Evaluation

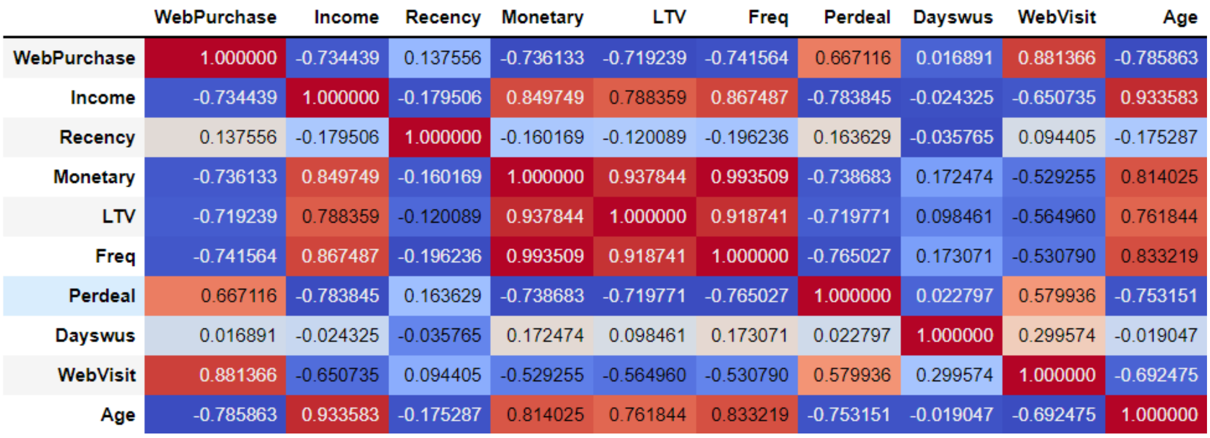
See Appendix R for visualizations

Fortunately, we were able to draw strong conclusions from the largest three out of six combined segments, segments (0,0), (1,0) and (2,1) which combined accounted for 81.17% of our customers. These segments are contained in appendices R1-3 and detail the very different characteristics of our customer base. Our recommendation would be to create bespoke marketing strategies to address the needs of each of these segments For instance, our cluster that is made up of predominantly young customers with the lower income, overwhelmingly prefer dry red wines purchased online and on discount. For this, we would suggest implementing a social media campaign and Google Ads for discounted dry red wines aimed towards this younger demographic.

6. Appendix

6.A

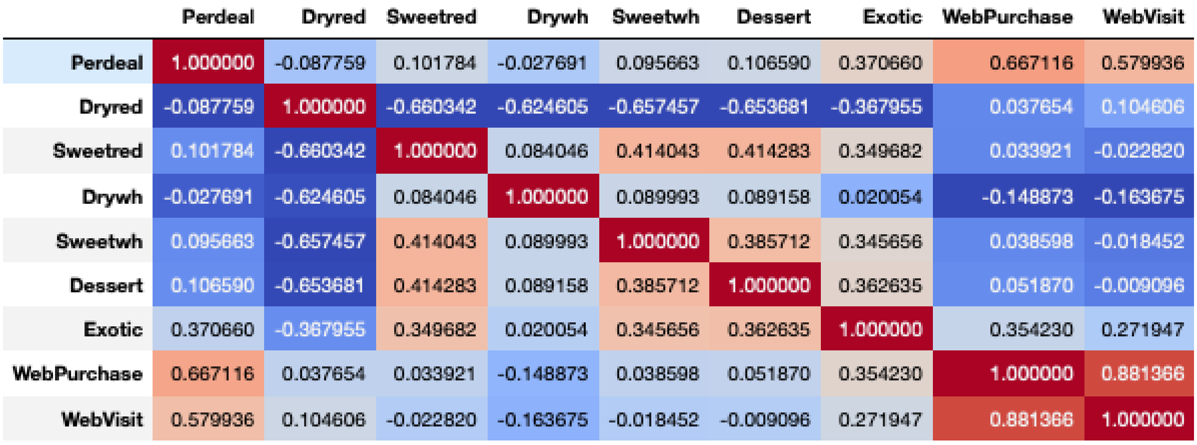
Correlation matrix for the feature set of our customer value segmentation



	WebPurchase	Income	Recency	Monetary	LTV	Freq	Perdeal	Dayswus	WebVisit	Age
WebPurchase	1.000000	-0.734439	0.137556	-0.736133	-0.719239	-0.741564	0.667116	0.016891	0.881366	-0.785863
Income	-0.734439	1.000000	-0.179506	0.849749	0.788359	0.867487	-0.783845	-0.024325	-0.650735	0.933583
Recency	0.137556	-0.179506	1.000000	-0.160169	-0.120089	-0.196236	0.163629	-0.035765	0.094405	-0.175287
Monetary	-0.736133	0.849749	-0.160169	1.000000	0.937844	0.993509	-0.738683	0.172474	-0.529255	0.814025
LTV	-0.719239	0.788359	-0.120089	0.937844	1.000000	0.918741	-0.719771	0.098461	-0.564960	0.761844
Freq	-0.741564	0.867487	-0.196236	0.993509	0.918741	1.000000	-0.765027	0.173071	-0.530790	0.833219
Perdeal	0.667116	-0.783845	0.163629	-0.738683	-0.719771	-0.765027	1.000000	0.022797	0.579936	-0.753151
Dayswus	0.016891	-0.024325	-0.035765	0.172474	0.098461	0.173071	0.022797	1.000000	0.299574	-0.019047
WebVisit	0.881366	-0.650735	0.094405	-0.529255	-0.564960	-0.530790	0.579936	0.299574	1.000000	-0.692475
Age	-0.785863	0.933583	-0.175287	0.814025	0.761844	0.833219	-0.753151	-0.019047	-0.692475	1.000000

6.B

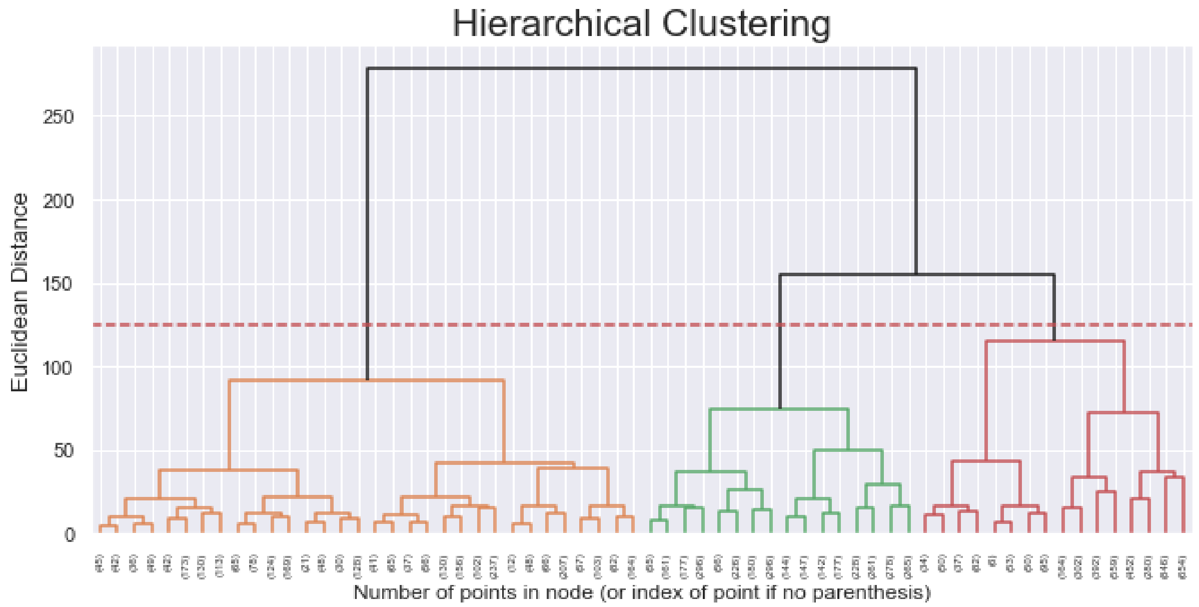
Correlation matrix for the feature set of our buying behavior segmentation



	Perdeal	Dryred	Sweetred	Drywh	Sweetwh	Dessert	Exotic	WebPurchase	WebVisit
Perdeal	1.000000	-0.087759	0.101784	-0.027691	0.095663	0.106590	0.370660	0.667116	0.579936
Dryred	-0.087759	1.000000	-0.660342	-0.624605	-0.657457	-0.653681	-0.367955	0.037654	0.104606
Sweetred	0.101784	-0.660342	1.000000	0.084046	0.414043	0.414283	0.349682	0.033921	-0.022820
Drywh	-0.027691	-0.624605	0.084046	1.000000	0.089993	0.089158	0.020054	-0.148873	-0.163675
Sweetwh	0.095663	-0.657457	0.414043	0.089993	1.000000	0.385712	0.345656	0.038598	-0.018452
Dessert	0.106590	-0.653681	0.414283	0.089158	0.385712	1.000000	0.362635	0.051870	-0.009096
Exotic	0.370660	-0.367955	0.349682	0.020054	0.345656	0.362635	1.000000	0.354230	0.271947
WebPurchase	0.667116	0.037654	0.033921	-0.148873	0.038598	0.051870	0.354230	1.000000	0.881366
WebVisit	0.579936	0.104606	-0.022820	-0.163675	-0.018452	-0.009096	0.271947	0.881366	1.000000

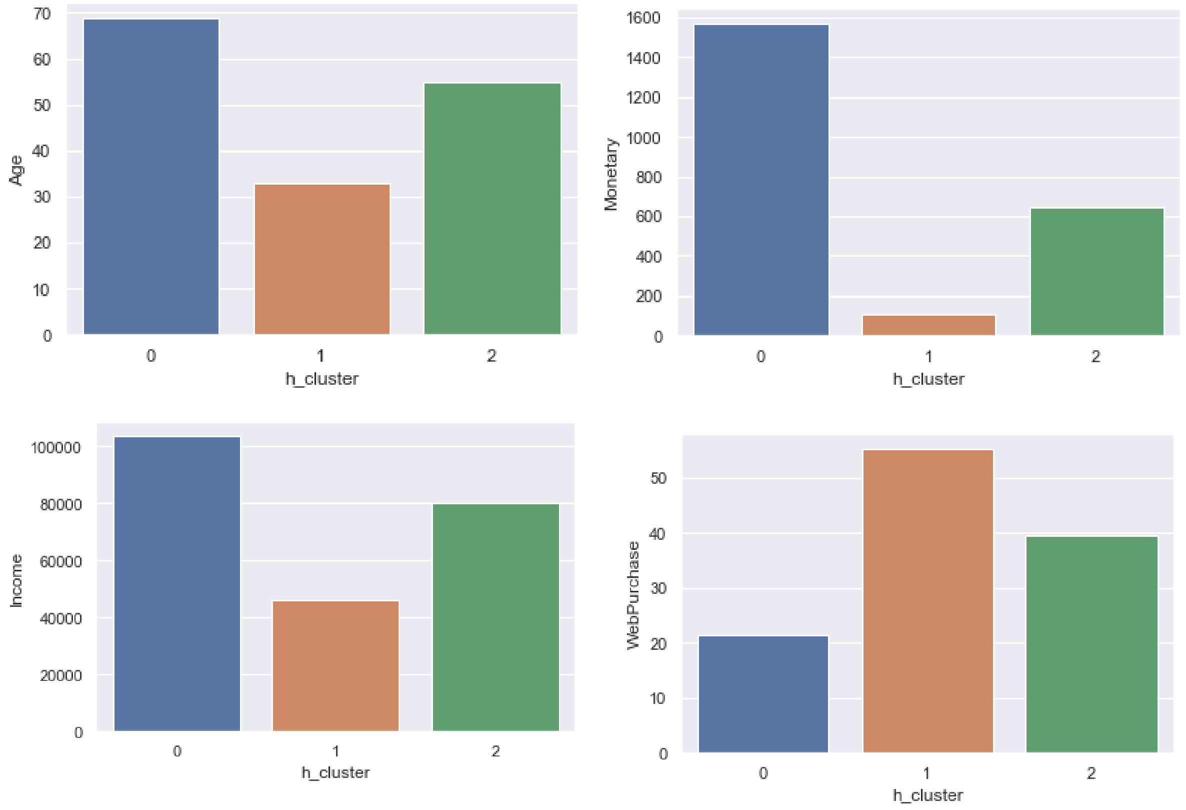
6.C

Dendrogram to show hierarchical clusters - Value segmentation



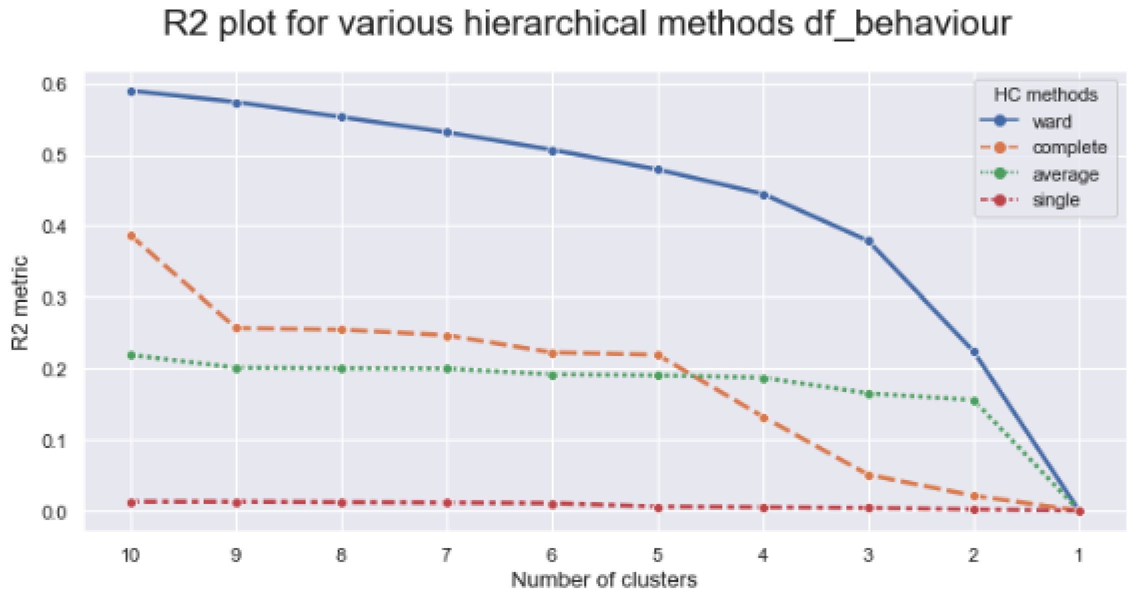
6.D

Bar plots to show findings in hierarchical clusters for value segmentation



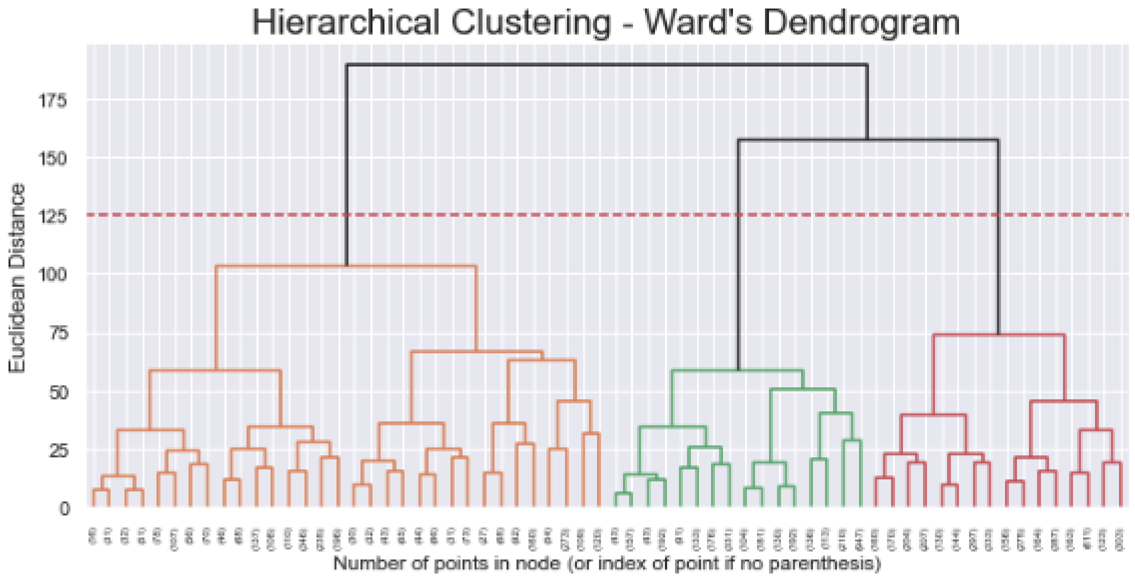
6.E

R2 plot for various linkage methods - Behavior Segmentation



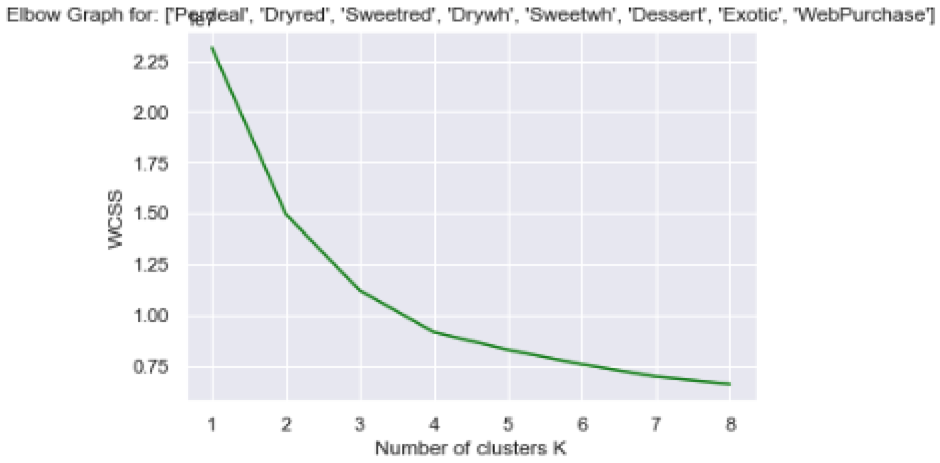
6.F

Dendrogram with Ward’s Linkage method - Behavior Segmentation



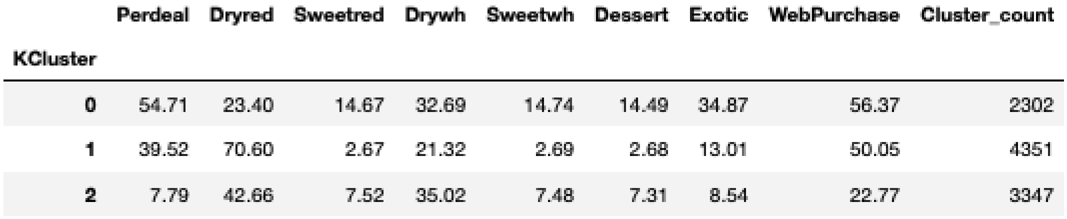
6.G

Elbow Graph for determining amount of clusters - Behavior Segmentation



6.H

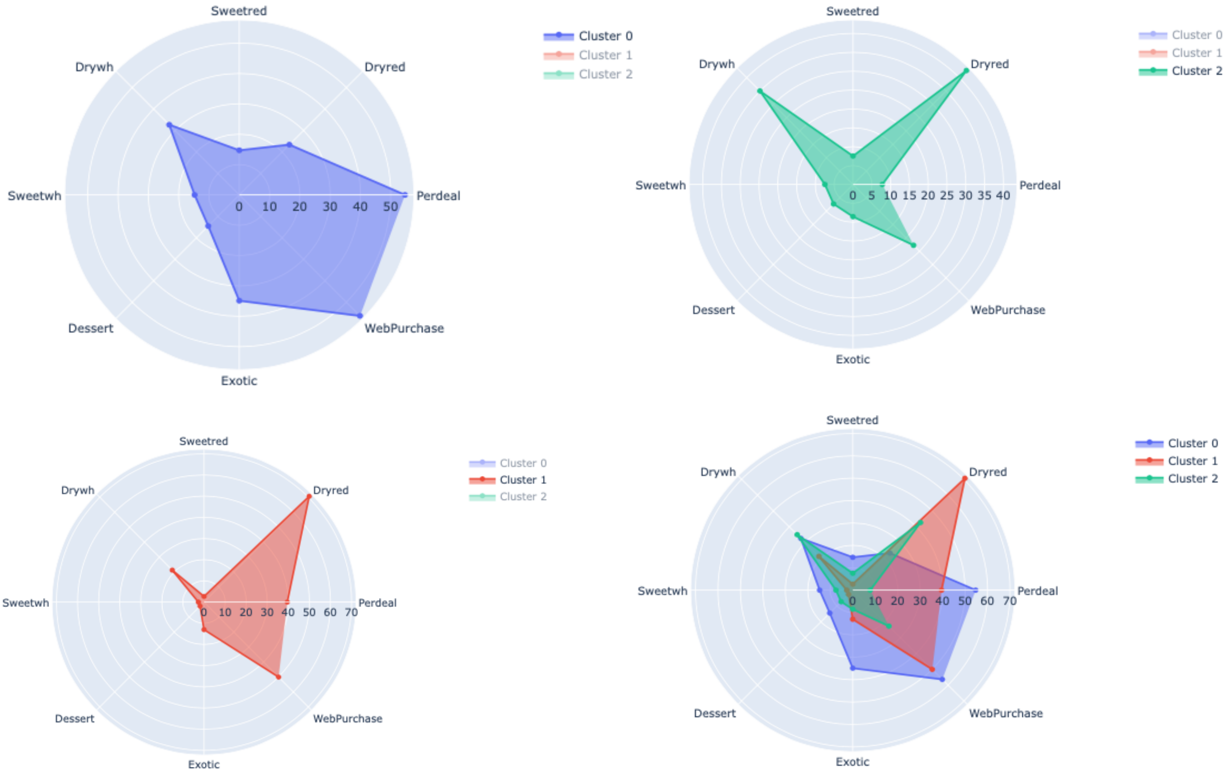
table with mean of each column per cluster - Behavior Segmentation



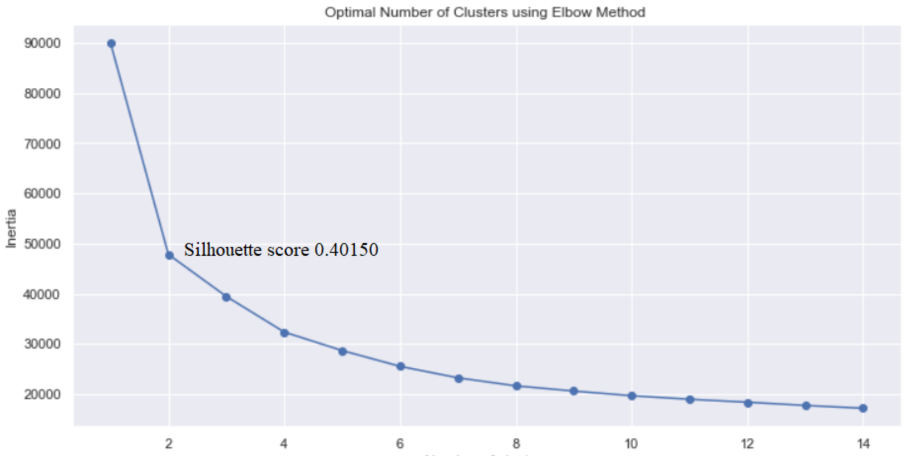
	Perdeal	Dryred	Sweetred	Drywh	Sweetwh	Dessert	Exotic	WebPurchase	Cluster_count
KCluster									
0	54.71	23.40	14.67	32.69	14.74	14.49	34.87	56.37	2302
1	39.52	70.60	2.67	21.32	2.69	2.68	13.01	50.05	4351
2	7.79	42.66	7.52	35.02	7.48	7.31	8.54	22.77	3347

6.I

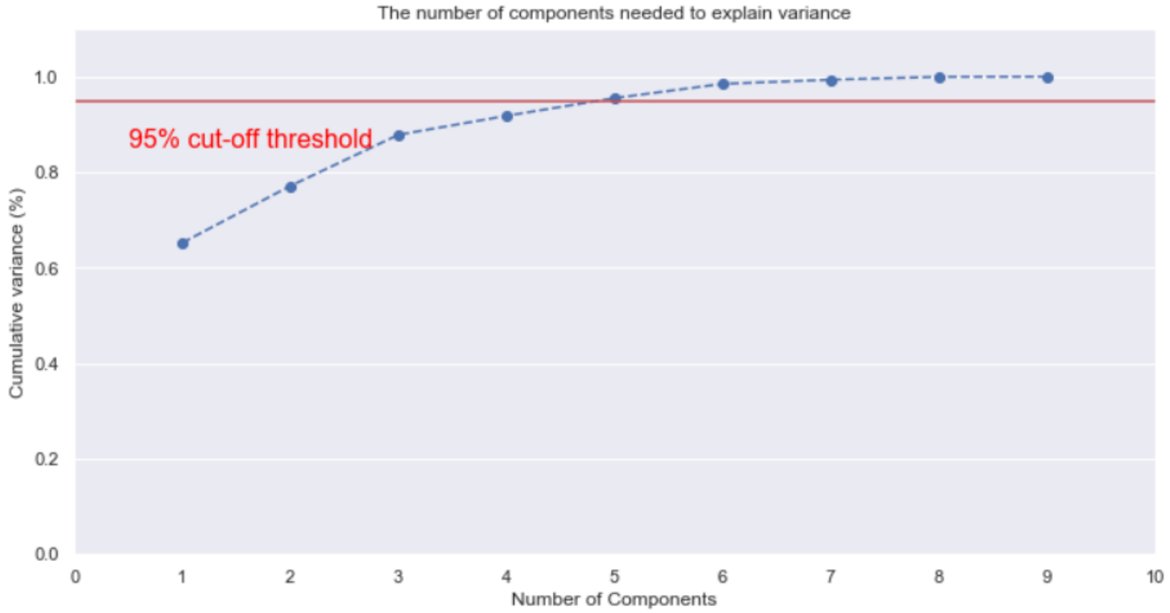
Radar graphs of results in appendix H - Behavior Segmentation



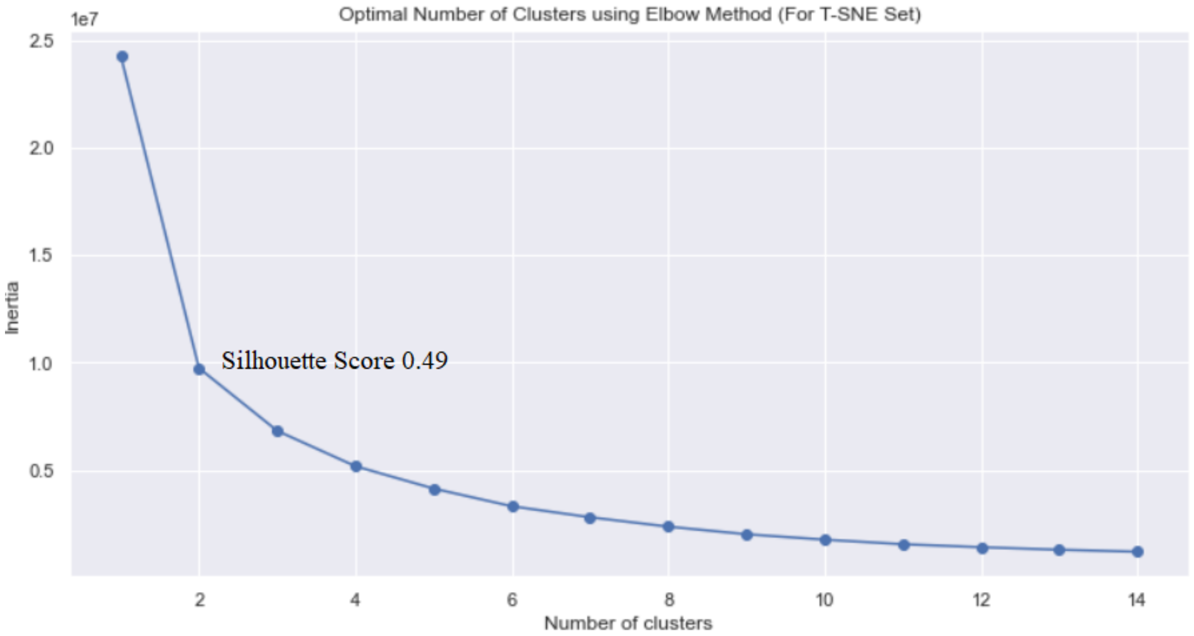
6.J



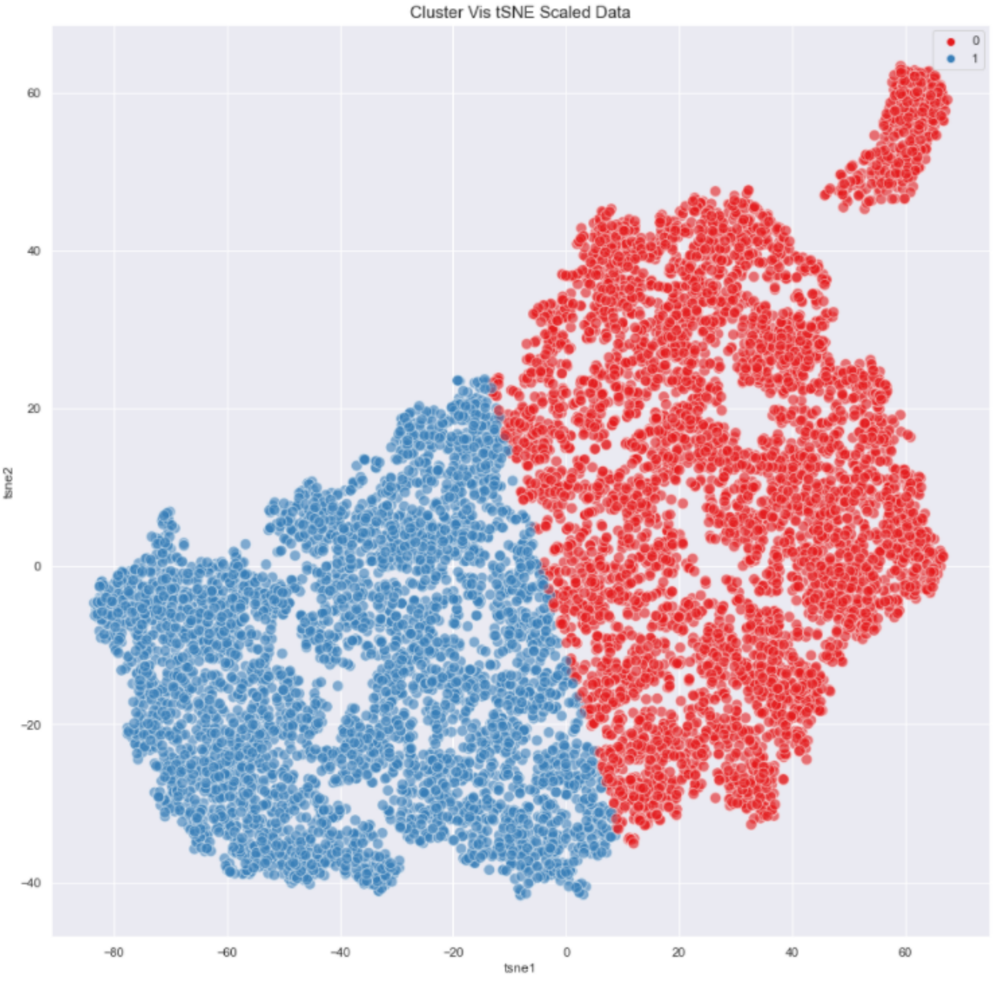
6.K


The number of components needed to explain variance
Cumulative variance (%)
1.0
0.8
0.6
0.4
0.2
0.0
95% cut-off threshold
0
1
2
3
4
5
6
7
8
9
10
Number of Components

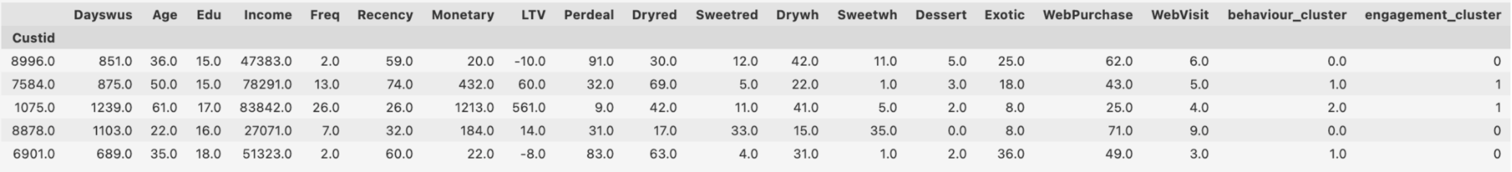
6.L


Optimal Number of Clusters using Elbow Method (For T-SNE Set)
1e7
2.5
2.0
1.5
1.0
0.5
Inertia
Silhouette Score 0.49
2
4
6
8
10
12
14
Number of clusters

6.M

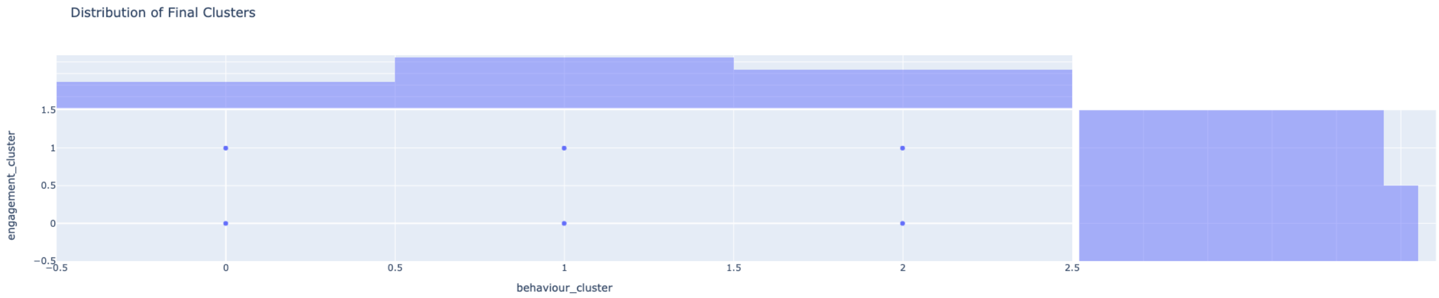


6.N

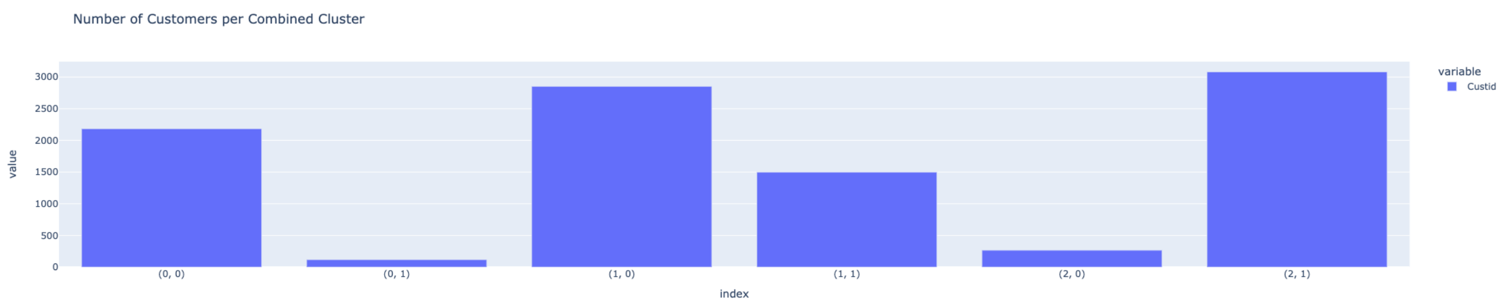


	Dayswus	Age	Edu	Income	Freq	Recency	Monetary	LTV	Perdeal	Dryred	Sweetred	Drywh	Sweetwh	Dessert	Exotic	WebPurchase	WebVisit	behaviour_cluster	engagement_cluster
Custid																			
8996.0	851.0	36.0	15.0	47383.0	2.0	59.0	20.0	-10.0	91.0	30.0	12.0	42.0	11.0	5.0	25.0	62.0	6.0	0.0	0
7584.0	875.0	50.0	15.0	78291.0	13.0	74.0	432.0	60.0	32.0	69.0	5.0	22.0	1.0	3.0	18.0	43.0	5.0	1.0	1
1075.0	1239.0	61.0	17.0	83842.0	26.0	26.0	1213.0	561.0	9.0	42.0	11.0	41.0	5.0	2.0	8.0	25.0	4.0	2.0	1
8878.0	1103.0	22.0	16.0	27071.0	7.0	32.0	184.0	14.0	31.0	17.0	33.0	15.0	35.0	0.0	8.0	71.0	9.0	0.0	0
6901.0	689.0	35.0	18.0	51323.0	2.0	60.0	22.0	-8.0	83.0	63.0	4.0	31.0	1.0	2.0	36.0	49.0	3.0	1.0	0

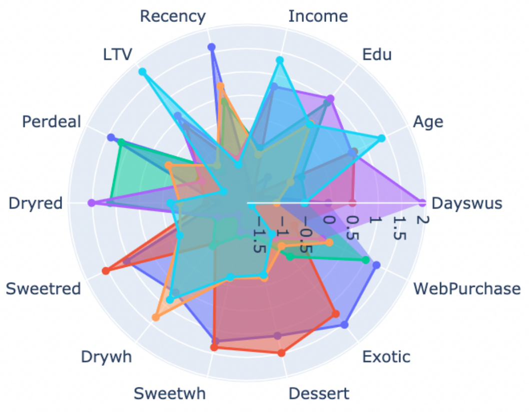
6.O



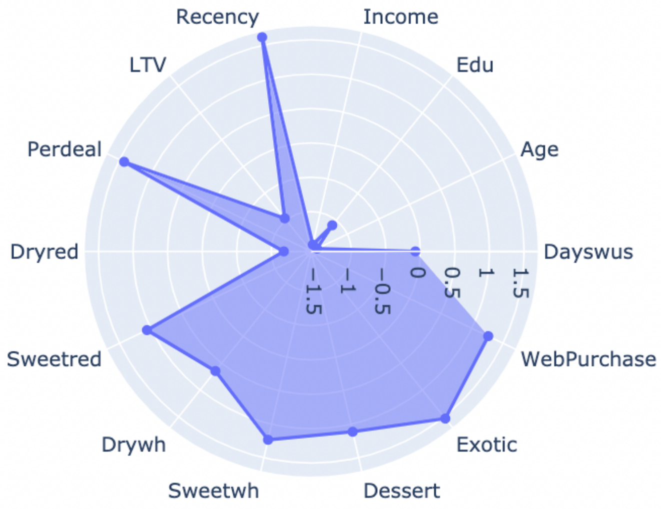
6.P


Number of Customers per Combined Cluster

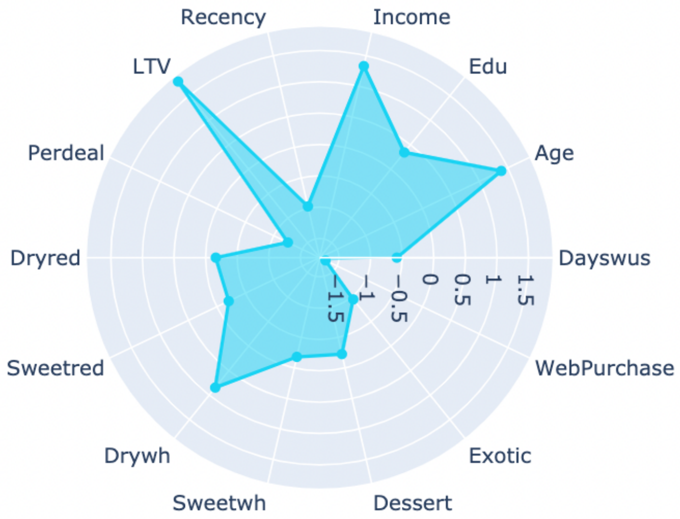
6.Q



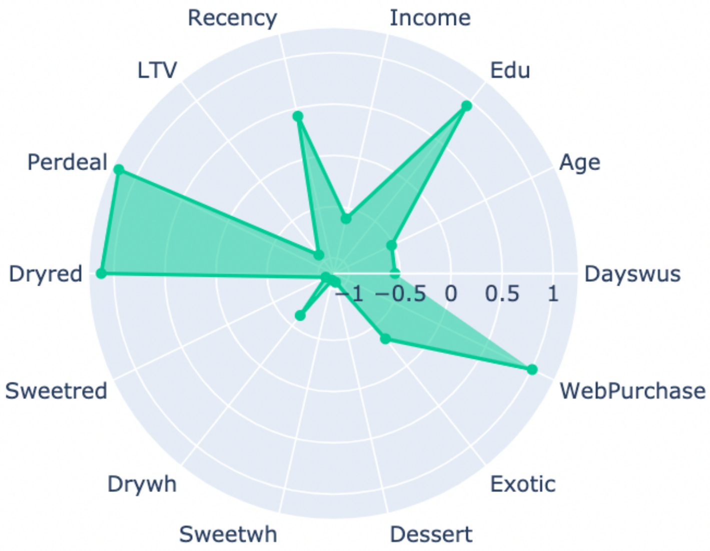
6.R



6.R.2


Recency
Income
Edu
Age
Dayswus
WebPurchase
Exotic
Dessert
Sweetwh
Drywh
Sweetred
Dryred
Perdeal
LTV
−1.5
−1
−0.5
0
0.5
1
1.5

6.R.3


Recency
Income
LTV
Edu
Perdeal
Age
Dryred
Dayswus
−1
−0.5
0
0.5
1
Sweetred
WebPurchase
Drywh
Exotic
Sweetwh
Dessert