Consumer Market Segmentation – Clustering using RapidMiner

Mehul Parmar
Graduate Student, Management Information Systems
University of Illinois at Chicago
Date: 10/24/2015

DATA MINING - PART A

The given data is collected from market research across different cities and towns in India to track consumer purchase behavior in consumer goods.

Problem Definition:

- Segment market based on two key variables:
  - Purchase Behavior (Brand loyalty, volume, frequency)
  - Basis for Purchase (price, selling proposition)
- Enable clients to deploy more effective promotional campaigns based on effective market segmentation.

General Description:

The given data consists of multiple transactions whereby each transaction represents a household (Unique ID). The data was clean by performing a series of operation like Filter Example, Generate Attributes, and Normalize.

Answer 1-a:

In order to identify clusters using K-means approach, the following variables were selected to construct the model:

Table 1: Purchase Behavior Variables

Attribute Name	Description	Type
Avg. Price	Avg. price (rupees per 100 gram)	Real
Brand Runs	Number of runs (streaks) of purchasing the same brand	Integer
Brand_Royalty	Whether customer is Brand loyal or not (Max_Br_Cd > 30%)	Binomial
Max_Br_Cd	Maximum Brand loyalty amongst the given Brands	Numeric
No. of Trans	Number of transactions	Integer
No. of Brands	Number of brands purchased	Integer
Others 999	Brand Codes other than (55, 272, 286, 24, 481, 352, 5)	Numeric
Total Volume	Volume of product purchased (grams)	Integer
Value	Value in paise (100 paise = 1 rupee)	Numeric

The model was constructed with different values of k. The following table summarizes the model details for different values of k:

Table 2: Cluster models for Purchase Behavior

	K = 2	K = 3	K = 4	K = 5
Cluster Model	Cluster 0 = 260 Cluster 1 = 340	Cluster 0 = 211 Cluster 1 = 249 Cluster 2 = 140	Cluster 0 = 163 Cluster 1 = 231 Cluster 2 = 151 Cluster 3 = 55	Cluster 0 = 142 Cluster 1 = 145 Cluster 2 = 40 Cluster 3 = 223 Cluster 4 = 50
Within Centroid Distance	Avg. within Centroid Distance = 6.429 Avg. within Centroid Distance (Cluster 0) = 5.029 Avg. within Centroid Distance (Cluster 1) = 7.499	Avg. within Centroid Distance = 5.171 Avg. within Centroid Distance (Cluster 0) = 4.827 Avg. within Centroid Distance (Cluster 1) = 4.416 Avg. within Centroid Distance (Cluster 2) = 7.844	Avg. within Centroid Distance = 4.558 Avg. within Centroid Distance (Cluster 0) = 3.866 Avg. within Centroid Distance (Cluster 1) = 3.872 Avg. within Centroid Distance (Cluster 2) = 4.744 Avg. within Centroid Distance (Cluster 3) = 8.896	Avg. within Centroid Distance = 4.227 Avg. within Centroid Distance (Cluster 0) = 3.872 Avg. within Centroid Distance (Cluster 1) = 3.405 Avg. within Centroid Distance (Cluster 2) = 7.741 Avg. within Centroid Distance (Cluster 3) = 3.788 Avg. within Centroid Distance (Cluster 4) = 6.769
Between Cluster Distance	1 – 2 = 3.226	1 – 2 = 3.296 1 – 3 = 3.067 2 – 3 = 3.880	1 – 2 = 3.372 1 – 3 = 2.779 1 – 4 = 4.396 2 – 3 = 3.826 2 – 4 = 3.948 3 – 4 = 3.360	1 – 2 = 2.425 1 – 3 = 4.683 1 – 4 = 3.465 1 – 5 = 4.425 2 – 3 = 2.847 2 – 4 = 3.375 2 – 5 = 3.437 3 – 4 = 5.609 3 – 5 = 3.763 4 – 5 = 3.517

- For k=2, there is a significant difference between the Others 999 (brand code other than the given brand columns), Brand royalty and Maximum brand loyalty (%) of the customer. Thus, the people who buy more from the Others 999 segment have lesser brand royalty and those who buy less from the Others 999 segment have more brand royalty.
- For k=3, the cluster sizes do not vary too much. People having large brand runs and large number of transactions are less brand loyal. Rest of the criteria remain same as those observed in clustering model with K =2.

- For k=4 and k=5, we observe that the difference in the cluster sizes is very large. Also, it divides the data into more number of groups that have varying patterns but decreased number of observations indicates lesser authenticity as compared to the previous models.

Answer 1-b:

The variables used to describe basis-for-purchase are as follows:

Table 3: Basis-for-purchase Variables

Attribute Name	Description	Type
Pur Vol No Promo - %	Percent of volume purchased under no-promotion	Numeric
Pur Vol Other Promo %	Percent of volume purchased under other promotions	Numeric
Pur Vol Promo 6 %	Percent of volume purchased under Promotion Code 6	Numeric
Max_SP	Maximum selling proposition amongst different brands	Real
Pr_Cat 1	Any Premium soap	Numeric
Pr_Cat 2	Any Popular soap	Numeric
Pr_Cat 3	Any economic/carbolic soap	Numeric
Pr_Cat 4	Any sub-popular soap	Numeric

Table 4: Cluster Models for Basis-for-purchase

	K = 2	K = 3	K = 4	K = 5
Cluster Model	Cluster 0 = 105 Cluster 1 = 495	Cluster 0 = 398 Cluster 1 = 118 Cluster 2 = 84	Cluster 0 = 78 Cluster 1 = 19 Cluster 2 = 208 Cluster 3 = 295	Cluster 0 = 60 Cluster 1 = 111 Cluster 2 = 16 Cluster 3 = 286 Cluster 4 = 127
Within Centroid Distance	Avg. within Centroid Distance = 5.615 Avg. within Centroid Distance (Cluster 0) = 11.022 Avg. within Centroid Distance (Cluster 1) = 4.468	Avg. within Centroid Distance = 4.565 Avg. within Centroid Distance (Cluster 0) = 2.728 Avg. within Centroid Distance (Cluster 1) = 5.856 Avg. within Centroid Distance (Cluster 2) = 11.455	Avg. within Centroid Distance = 4.126 Avg. within Centroid Distance (Cluster 0) = 6.790 Avg. within Centroid Distance (Cluster 1) = 14.221 Avg. within Centroid Distance (Cluster 2) = 6.074 Avg. within Centroid Distance (Cluster 3) = 1.398	Avg. within Centroid Distance = 3.125 Avg. within Centroid Distance (Cluster 0) = 6.930 Avg. within Centroid Distance (Cluster 1) = 5.256 Avg. within Centroid Distance (Cluster 2) = 14.880 Avg. within Centroid Distance (Cluster 3) = 1.358 Avg. within Centroid Distance (Cluster 4) = 1.965

Between Cluster Distance	1 – 2 = 3.084	1 – 2 = 2.689 1 – 3 =3.367 2 – 3 = 3.940	1 – 2 = 4.463 1 – 3 = 3.422 1 – 4 = 3.403 2 – 3 = 4.959 2 – 4 =5.105 3 – 4 = 2.216	1 – 2 = 4.218 1 – 3 = 4.837 1 – 4 = 3.751 1 – 5 = 3.791 2 – 3 = 5.515 2 – 4 =2.821 2 – 5 = 3.257 3 – 4 = 5.445 3 - 5 = 5.499 4 – 5 = 2.600

- K=2: There are two distinct clusters which can be seen with one having four times the observations of the second one. There is high distinction amongst the percent of volume purchased by different means. The other three factors do not significantly differentiate amongst the clusters. Thus people in one cluster buy items which are a part of other promotion and promotion 6 and people in the other cluster buy items which are under no promotion category.
- K=3: When a new cluster emerges it is similar to the second cluster but people in this cluster buy items which are under economic/carbolic pricing group.
- K=4: When we increase the cluster size the clusters are similar and can be distinguished as two major clusters since the other 3 are not differentiated by a large margin.
- K=5: It has similar observation like K = 4 signifying that increase in K is not distinguishing the clusters more.

Answer 1-c:

In order to identify clusters using K-means approach that describe both purchase behavior and basis for purchase, all the variables listed in Tables 1 and 3 were used.

The model was constructed with different values of k. The following table summarizes the model details for different values of k:

Table 5: Cluster models for Purchase Behavior

	K = 2	K = 3	K = 4	K = 5
Cluster Model	Cluster 0 = 340 Cluster 1 = 260	Cluster 0 = 169 Cluster 1 = 226 Cluster 2 = 205	Cluster 0 = 62 Cluster 1 = 169 Cluster 2 = 156 Cluster 3 = 213	Cluster 0 = 58 Cluster 1 = 116 Cluster 2 = 190 Cluster 3 = 82 Cluster 4 = 154
Within Centroid Distance	Avg. within Centroid Distance = 12.957 Avg. within Centroid Distance (Cluster 0) = 13.84 Avg. within Centroid Distance (Cluster 1) = 11.79	Avg. within Centroid Distance = 11.612 Avg. within Centroid Distance (Cluster 0) = 10.733 Avg. within Centroid Distance (Cluster 1) = 10.668	Avg. within Centroid Distance = 10.699 Avg. within Centroid Distance (Cluster 0) = 16.332 Avg. within Centroid Distance (Cluster 1) = 10.561	Avg. within Centroid Distance = 9.642 Avg. within Centroid Distance (Cluster 0) = 15.439 Avg. within Centroid Distance (Cluster 1) = 8.533

		Avg. within Centroid Distance (Cluster 2) = 13.370	Avg. within Centroid Distance (Cluster 2) = 10.332 Avg. within Centroid Distance (Cluster 3) = 9.346	Avg. within Centroid Distance (Cluster 2) = 8.461 Avg. within Centroid Distance (Cluster 3) = 9.268 Avg. within Centroid Distance (Cluster 4) = 9.951
Between Cluster Distance	1 – 2 = 3.505	1 – 2 = 3.868 1 – 3 = 3.144 2 – 3 = 3.764	1 – 2 = 3.709 1 – 3 = 4.123 1 – 4 = 4.960 2 – 3 = 3.279 2 – 4 = 3.773 3 – 4 = 3.802	1 – 2 = 4.211 1 – 3 = 5.063 1 – 4 = 4.198 1 – 5 = 4.056 2 – 3 = 4.196 2 – 4 = 3.790 2 – 5 = 3.781 3 – 4 = 4.080 3 – 5 = 3.861 4 – 5 = 3.681

The above table shows that

- For K = 2, the two clusters have large observations within them with significant differences in Brand loyalty, Maximum Brand loyalty (%), Brand Codes other than the given Brand columns (Others 999) as well the Selling proposition (Max_SP).
- Also, for K = 3, the same variables differ amongst the 3 generated clusters. The new cluster generated has the highest values for these attributes. Moreover, the Centroid Plot view (Figure 19) tells us that Other 999 (Brands other than the given Brand columns) and Brand Loyalty have an inverse relation.
- For K = 4, one of the clusters has a very small cluster size suggesting that the given model is not feasible to consider for clustering of the given data.
- Based on the distance computations, the K-means clustering with K = 5 gives the best results.

Answer 1-d:

In order to select the set of variables which we think will be useful for addressing the segmentation needed, we compared the different characteristics of the clusters formed by these variables. The clusters were compared based on the following characteristics:

- **Cluster Size:** Data points should be evenly distributed across all Clusters.
- **Average within Centroid Distance:** Clusters should have lower values of average within centroid distance signifying tight coupling within the cluster.
- **Variations between clusters in terms of the attributes that define the model:** There should be distinguishable variations between the various clusters formed in a model from the point of view of the selected model attributes.

Based on all these factors we found that the set of variables which describes the Purchase behavior provides best segmentation with evenly distributes clusters, lower values of average within centroid distance, and distinguishable variations between the various clusters formed in a model compared to other set of variables. So, we pick the set of variables which describes the Purchase behavior for obtaining segmentation using the k-medoids, kernel k-means, agglomerative clustering, and DBSCAN clustering.

k-medoids:

The model was constructed with different values of k. The following table summarizes the model details for different values of k:

Table 6: k-medoids Cluster models for Purchase Behavior

	K = 2	K = 3	K = 4	K = 5
Cluster Model	Cluster 0 = 296 Cluster 1 = 304	Cluster 0 = 136 Cluster 1 = 188 Cluster 2 = 276	Cluster 0 = 78 Cluster 1 = 117 Cluster 2 = 155 Cluster 3 = 250	Cluster 0 = 57 Cluster 1 = 117 Cluster 2 = 139 Cluster 3 = 47 Cluster 4 = 240
Within Centroid Distance	Avg. within Centroid Distance = 9.05 Avg. within Centroid Distance (Cluster 0) = 10.01 Avg. within Centroid Distance (Cluster 1) = 8.108	Avg. within Centroid Distance = 7.210 Avg. within Centroid Distance (Cluster 0) = 5.350 Avg. within Centroid Distance (Cluster 1) = 9.568 Avg. within Centroid Distance (Cluster 2) = 6.520	Avg. within Centroid Distance = 6.146 Avg. within Centroid Distance (Cluster 0) = 10.871 Avg. within Centroid Distance (Cluster 1) = 3.981 Avg. within Centroid Distance (Cluster 2) = 7.081 Avg. within Centroid Distance (Cluster 3) = 5.104	Avg. within Centroid Distance = 5.644 Avg. within Centroid Distance (Cluster 0) = 8.820 Avg. within Centroid Distance (Cluster 1) = 3.981 Avg. within Centroid Distance (Cluster 2) = 5.392 Avg. within Centroid Distance (Cluster 3) = 11.145 Avg. within Centroid Distance (Cluster 4) = 4.819
Between Cluster Distance	1 – 2 = 3.84	1 – 2 = 2.862 1 – 3 = 3.840 2 – 3 = 4.124	1 – 2 = 4.710 1 – 3 = 4.521 1 – 4 = 4.423 2 – 3 = 2.862 2 – 4 = 3.840 3 – 4 = 4.124	1 – 2 = 4.710 1 – 3 = 4.521 1 – 4 = 3.960 1 – 5 = 4.243 2 – 3 = 2.862 2 – 4 = 5.719 2 – 5 = 3.840 3 – 4 = 5.000 3 – 5 = 4.124 4 – 5 = 4.607

The above table shows that

- For K = 2, the two clusters have large observations within them with significant differences in Brand loyalty, Maximum Brand loyalty (%), Brand Codes other than the given Brand columns (Others 999).

- Also, for K = 3, the same variables differ amongst the 3 generated clusters. The new cluster generated has the highest values for these attributes. Moreover, the Centroid Plot view (Figure 19) tells us that Other 999 (Brands other than the given Brand columns) and Brand Loyalty have an inverse relation.
- For K = 4, one of the clusters has a very small cluster size suggesting that the given model is not feasible to consider for clustering of the given data.
- Based on the distance computations, the K-means clustering with K = 5 gives the best results but two of the clusters has a very small cluster size suggesting that the given model is not feasible to consider for clustering of the given data.

kernel k-means:

As k-means using kernel clustering is a non-centroid based cluster model we made use of the Cluster Density Performance operator to evaluate the model. The cluster density performance operator calculates the average distance between points in a cluster and multiplies this by the number of points minus 1.

Kernel k-means uses kernels to estimate the distance between objects and clusters. We set the kernel type as dot operator as it was giving a better cluster distribution compared to other kernel types. The model was constructed with different values of k. The following table summarizes the model details for different values of k:

Table 7: Kernel k-means Clustering Cluster models for Purchase Behavior

	K = 2	K = 3	K = 4	K = 5
Cluster Model	Cluster 0 = 283 Cluster 1 = 317	Cluster 0 = 249 Cluster 1 = 152 Cluster 2 = 199	Cluster 0 = 100 Cluster 1 = 144 Cluster 2 = 145 Cluster 3 = 211	Cluster 0 = 85 Cluster 1 = 166 Cluster 2 = 90 Cluster 3 = 108 Cluster 4 = 151
Within Cluster Distance	Avg. within Cluster Distance = 3887.940 Avg. within Cluster Distance (Cluster 0) = 3313.644 Avg. within Cluster Distance (Cluster 1) = 4400.60	Avg. within Cluster Distance = 2058.061 Avg. within Cluster Distance (Cluster 0) = 2198.135 Avg. within Cluster Distance (Cluster 1) = 2340.351 Avg. within Cluster Distance (Cluster 2) = 1667.173	Avg. within Cluster Distance = 1441.377 Avg. within Cluster Distance (Cluster 0) = 1445.321 Avg. within Cluster Distance (Cluster 1) = 1277.166 Avg. within Cluster Distance (Cluster 2) = 1073.347 Avg. within Cluster Distance (Cluster 3) = 1804.488	Avg. within Cluster Distance = 1038.626 Avg. within Cluster Distance (Cluster 0) = 1445.844 Avg. within Cluster Distance (Cluster 1) = 1157.633 Avg. within Cluster Distance (Cluster 2) = 797.012 Avg. within Cluster Distance (Cluster 3) = 819.015

				Avg. within Cluster Distance (Cluster 4) = 979.649

The above table shows that

- For K = 2, we get an evenly distributed cluster size. However, within the cluster distance is large for both the clusters, which means that points within the clusters are not that close to each other. This suggests that the given model is not feasible to consider for clustering of the given data.
- For K = 3, also we get evenly distributes cluster size but within the cluster distance is large.
- For K = 4, also we get evenly distributes cluster size but within the cluster distance is large.
- For K=5, we don't get an evenly distributes cluster and also within the cluster distance is large, suggesting that the given model is not feasible to consider for clustering of the given data.

Agglomerative clustering:

As Agglomerative clustering is a non-centroid based cluster model we made use of the Cluster Density Performance operator to evaluate the model. Also, as Agglomerative clustering is a hierarchical clustering technique we made use of the Flatten Clustering operator to obtain the number of clusters of our choice.

It is a bottom-up strategy of Hierarchical clustering. Three different strategies are supported by this operator: single-link, complete-link and average-link. We select the mode parameter as Complete link for our model and measure type as Mixed Measures and mixed measure parameter as MixedEuclideanDistance which gives us the best performance for this modeling technique. The result of this operator is a hierarchical cluster model, providing distance information to plot as a dendrogram.

The model was constructed with different values of k by setting the number of clusters parameter in Flatten Clustering operator. The following table summarizes the model details for different values of k:

Table 8: Agglomerative Clustering Cluster models for Purchase Behavior

	K = 2	K = 3	K = 4	K = 5
Cluster Model	Cluster 0 = 587 Cluster 1 = 13	Cluster 0 = 565 Cluster 1 = 13 Cluster 2 = 22	Cluster 0 = 228 Cluster 1 = 13 Cluster 2 = 337 Cluster 3 = 22	Cluster 0 = 13 Cluster 1 = 22 Cluster 2 = 337 Cluster 3 = 203 Cluster 4 = 25
Within Cluster Distance	Avg. within Cluster Distance = 2182.784 Avg. within Cluster Distance (Cluster 0) = 2229.860 Avg. within Cluster Distance (Cluster 1) = 57.16	Avg. within Cluster Distance = 1963.451 Avg. within Cluster Distance (Cluster 0) = 2080.832 Avg. within Cluster Distance (Cluster 1) = 57.116 Avg. within Cluster Distance	Avg. within Cluster Distance = 892.938 Avg. within Cluster Distance (Cluster 0) = 707.122 Avg. within Cluster Distance (Cluster 1) = 57.116 Avg. within Cluster Distance	Avg. within Cluster Distance = 820.343 Avg. within Cluster Distance (Cluster 0) = 57.116 Avg. within Cluster Distance (Cluster 1) = 75.367 Avg. within Cluster Distance

		(Cluster 2) = 75.367	(Cluster 2) = 1104.267 Avg. within Cluster Distance (Cluster 3) = 75.367	(Cluster 2) = 1104.267 Avg. within Cluster Distance (Cluster 3) = 572.125 Avg. within Cluster Distance (Cluster 4) = 61.030

The above table shows that

- For K = 2, we don't get an evenly distributes cluster and also within the cluster distance is large, suggesting that the given model is not feasible to consider for clustering of the given data.
- For K = 3, also we don't get an evenly distributes cluster and also within the cluster distance is large, suggesting that the given model is not feasible to consider for clustering of the given data.
- For K = 4, also we don't get an evenly distributes cluster and also within the cluster distance is large, suggesting that the given model is not feasible to consider for clustering of the given data.
- For K=5, we don't get an evenly distributes cluster and also within the cluster distance is large, suggesting that the given model is not feasible to consider for clustering of the given data.

DBSCAN:

As DBSCAN is a non-centroid based cluster model we made use of the Cluster Density Performance operator to evaluate the model. We changed the various values of min points in the DBSCAN operator by keeping the measure type as Mixed Measures and mixed measure parameter as MixedEuclideanDistance which gives us the best performance for this modeling technique.

The following table summarizes the model details for different values of min points:

Table 9: DBSCAN Cluster models for Purchase Behavior

	min points = 10	min points = 12	min points = 20
Cluster Model	Cluster 0 = 369 Cluster 1 = 9 Cluster 2 = 85 Cluster 3 = 51 Cluster 4 = 27 Cluster 5 = 11 Cluster 6 = 28 Cluster 7 = 10 Cluster 8 = 10	Cluster 0 = 447 Cluster 1 = 71 Cluster 2 = 21 Cluster 3 = 25 Cluster 4 = 36	Cluster 0 = 579 Cluster 1 = 21
Within Cluster Distance	Avg. within Cluster Distance = 1013.274 Avg. within Cluster Distance (Cluster 0) = 1595.000 Avg. within Cluster Distance (Cluster 1) = 9.356	Avg. within Cluster Distance = 1401.149 Avg. within Cluster Distance (Cluster 0) = 1855.300 Avg. within Cluster Distance (Cluster 1) = 116.940	Avg. within Cluster Distance = 2204.768 Avg. within Cluster Distance (Cluster 0) = 2283.951 Avg. within Cluster Distance (Cluster 1) = 21.600

	Avg. within Cluster Distance (Cluster 2) = 150.082 Avg. within Cluster Distance (Cluster 3) = 84.190 Avg. within Cluster Distance (Cluster 4) = 36.142 Avg. within Cluster Distance (Cluster 5) = 11.049 Avg. within Cluster Distance (Cluster 6) = 34.933 Avg. within Cluster Distance (Cluster 7) = 9.802 Avg. within Cluster Distance (Cluster 8) = 10.092	Avg. within Cluster Distance (Cluster 2) = 25.045 Avg. within Cluster Distance (Cluster 3) = 29.408 Avg. within Cluster Distance (Cluster 4) = 50.169	

- For min points = 10, we get 9 clusters with uneven cluster size and even the avg. distance within the cluster is high, suggesting that the given model is not feasible to consider for clustering of the given data.
- For min points = 12, we get 5 clusters with uneven cluster size and again the avg. distance within the cluster is high, suggesting that the given model is not feasible to consider for clustering of the given data.
- For min points = 20, we get just 2 clusters with uneven cluster size and again the avg. distance within the cluster is high, suggesting that the given model is not feasible to consider for clustering of the given data.

Based on the analysis of various clustering technique by using the various performance criteria, it can be concluded that k-medoids and k-means clustering technique gives better results with evenly distributed cluster, small average within the centroid distance and distinguishable variations between the various clusters formed in a model. For both these models we see K=3 gives the best result. However, k-medoids outperforms k-means with a better Average within the centroid distance. Also, the medoid as used by k-medoids is roughly comparable to the median which is more robust to outliers than the arithmetic mean.

Best Model: k-medoids with K = 3 for Purchasing Behavior

Answer 2:

In order to choose the best segmentation, the following key elements were considered:

- Purchase behavior
- Basis-for-purchase
- Both

The different characteristics of the clusters formed by these elements were compared to select the best model. The clusters were compared based on the following characteristics:

- **Cluster Size:** Data points should be evenly distributed across all Clusters.
- **Average within Centroid Distance:** Clusters should have lower values of average within centroid distance signifying tight coupling within the cluster.
- **Variations between clusters in terms of the attributes that define the model:** There should be distinguishable variations between the various clusters formed in a model from the point of view of the selected model attributes.

If we consider both the basis-for-purchase and purchasing behavior for the best segmentation, the model with K = 5 gives the least Average within cluster distance, however the model is not evenly distributed with clusters of varying size as well the within centroid distance for different clusters has a wide range.

Based on the distance calculation, purchasing behavior with K =5 seems the best model. However, since that model has abnormal distribution (cluster size) across different clusters as well as minimum variations amongst clusters with respect to Brand royalty, the model with K = 3 seems the best Model.

Best Model: k-medoids with K = 3 for Purchasing Behavior
Characteristics:

- Evenly distributed cluster size.
- Small Average within centroid distance signifying tightly coupled clusters.
- Evident variations between the clusters in terms of the model attributes. (Refer Figure 27 and 28 in Part B)
- Clusters:
    - Cluster 0: These consumers are the least brand loyal as they purchase most of their products from the other category. Also, their total volume of products purchased is the low amongst the clusters. Other characteristics include low average value, high average volume by transaction and high percent of volume purchased under promotion Code 6.
    - Cluster 1: These consumers are not brand loyal as they purchase most of their products from the Other category. Also, their total volume of products purchased is the lowest amongst the clusters. Other characteristics include highest no. of brands, low no. of transactions, lowest average value, highest percent of volume purchased under promotion Code 6, lowest average volume by transaction and highest average price.
    - Cluster 2: These consumers are the ones that represent the target customers for CRISA clients regarding their brand promotions. They have high brand loyalty with most of the products belonging to the Brands specified in the data columns (Br Cd. 144,55, 286 etc.). They have lowest value for Other brands, lowest percent of volume purchased under promotion Code 6, highest average volume by transaction and highest percent of volume purchased under no-promotion.

  In addition, the above clusters have nearly the same selling Affluence Index range and percent of volume purchased under other promotions.

Answer 3:

After deciding on the “best” cluster size and the cluster variable we consider building a decision tree to obtain rules and descriptions for the different clusters of the model.

The following variables were selected to compute the decision tree:

- Avg. Price
- Brand Runs
- Brand_Royalty
- Max_Br_Cd
- No. of Trans
- No. of Brands
- Others 999
- Total Volume
- Value

The leaf node of the tree is the cluster number. Thus, each household in the dataset will be placed in the corresponding cluster leaf nodes based on the generated rules.

The decision tree (Part B - Figure 25) was built on the criteria of **information gain** with the following parameters:

Minimum size of split: 10
Minimum leaf size: 10
Minimum gain: 0.1
Maximum Depth: 20
Confidence: 0.25

The main characteristics of the tree are:

- The root element for the decision tree is **Max_Br_Cd** (Brand Loyalty %).
- The other key attributes at the top of the tree are No. of Brands, Brand Runs and Avg_Price.
- The effectiveness of the tree in classifying the data can be viewed in terms of the Decision tree Accuracy performance. The performance of this tree is 94.50% (Refer to Part B – Figure 45) which is a good indication of the effectiveness of the Decision tree.
- Also, the root nodes for the tree are relatively large with almost pure nodes, thereby avoiding the issue of Over-fit to the current dataset.
- The tree also helps in defining the clusters with more concrete rules / definitions as compared to the cluster model which provides only an approximation (high, low, medium) to interpret the results.
- Sample rule: Households whose Max_Br_Cd is less than or equal to -0.240 and No_of Brands is less than or equal to -0.87 and Avg_Price is less than and equal to 0.264 then cluster_0.

DATA MINING - PART B

Answer 1-a:

- **K= 2**

Figure 1: Centroid Plot View K = 2

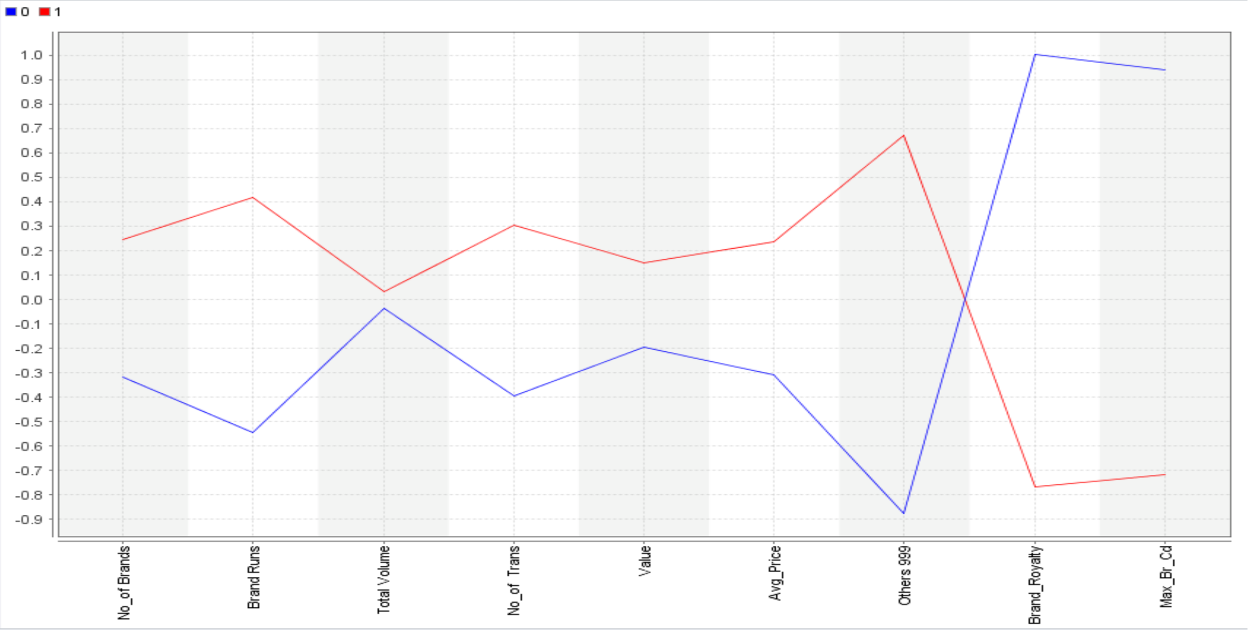
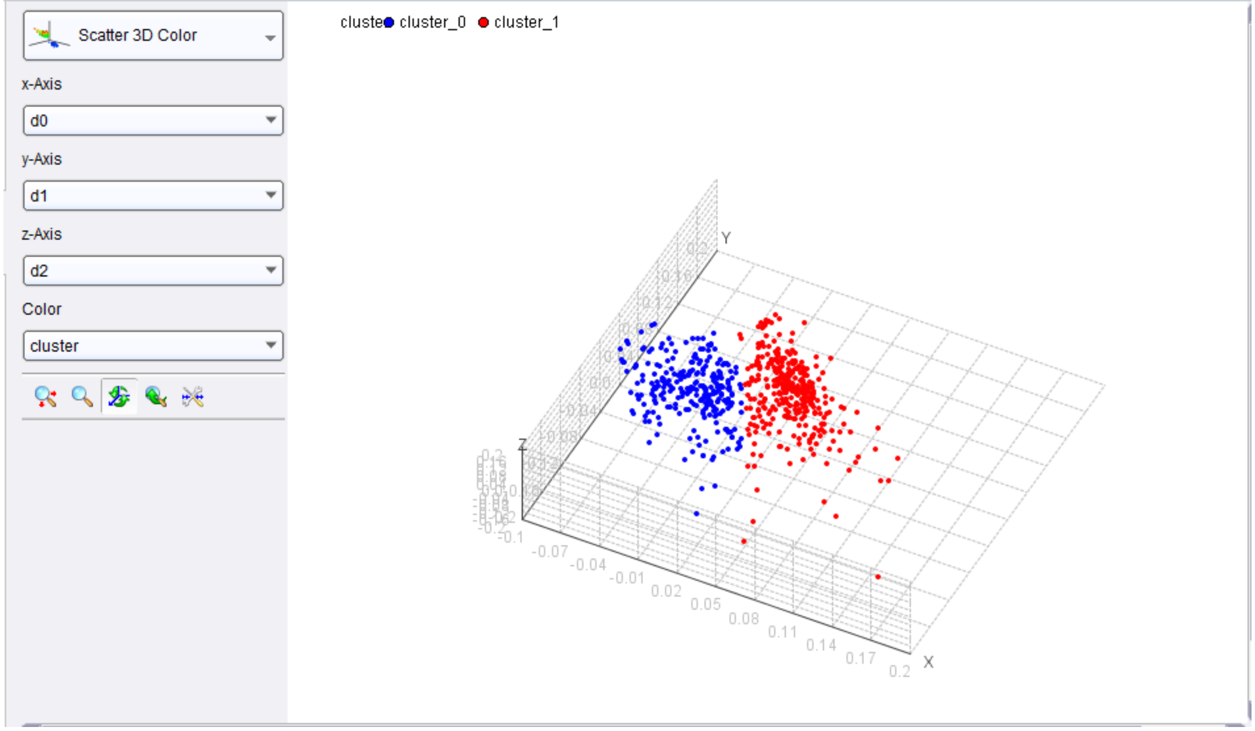


Figure 2: 3D Scatter chart K = 2


Scatter 3D Color

x-Axis

d0

y-Axis

d1

z-Axis

d2

Color

cluster

- **K= 3**

Figure 3: Centroid Plot View K = 3

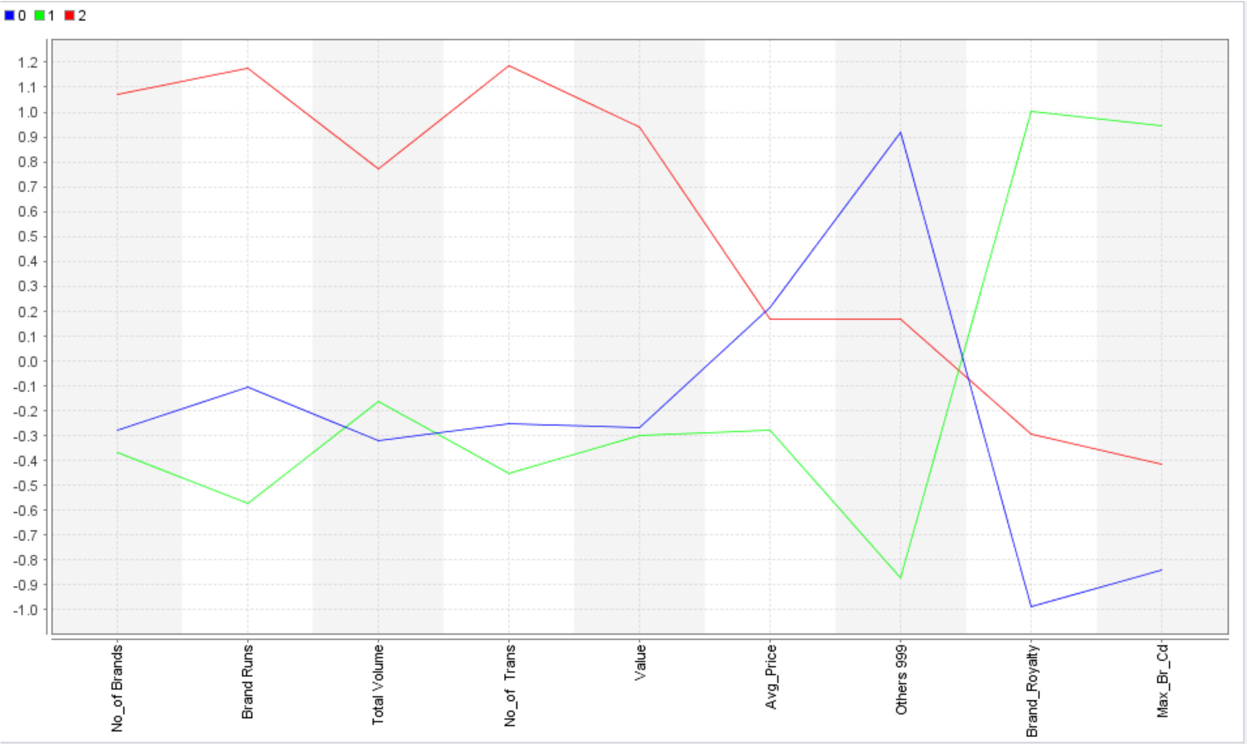
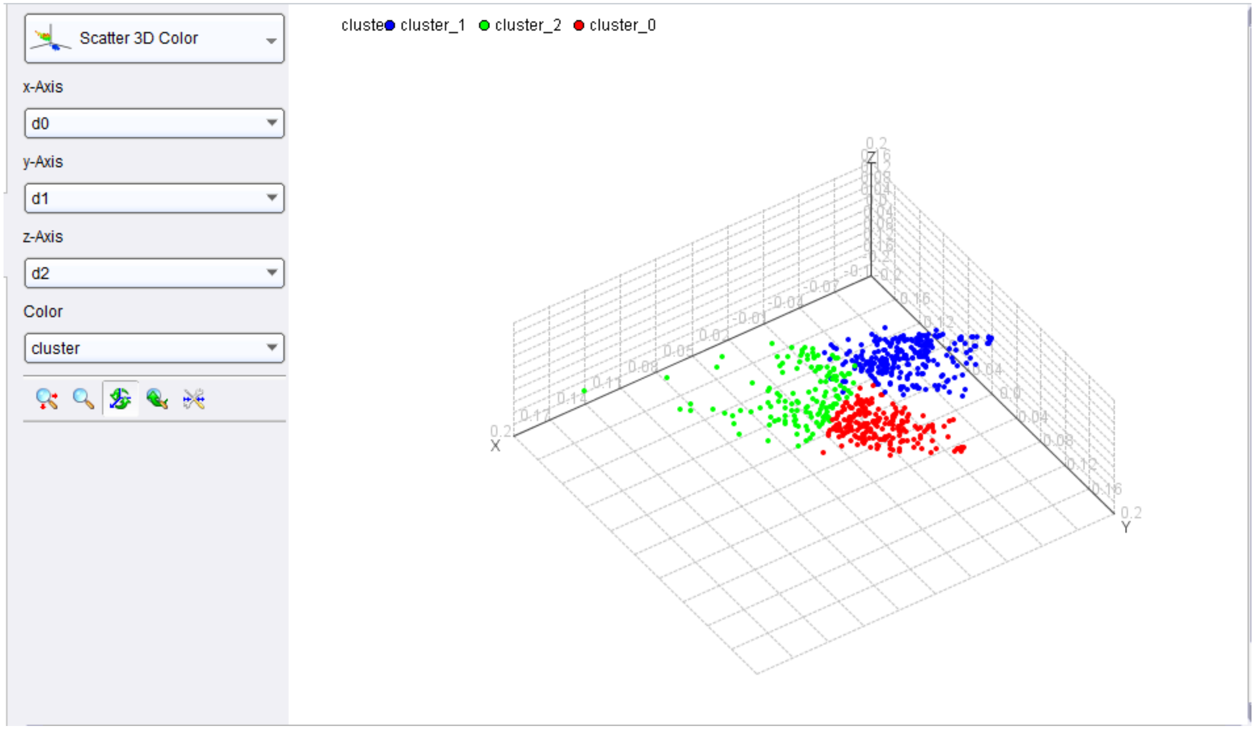


Figure 4: 3D Scatter chart K = 3



Scatter 3D Color

x-Axis

d0

y-Axis

d1

z-Axis

d2

Color

cluster

- **K = 4**

Figure 5: Centroid Plot View K = 4

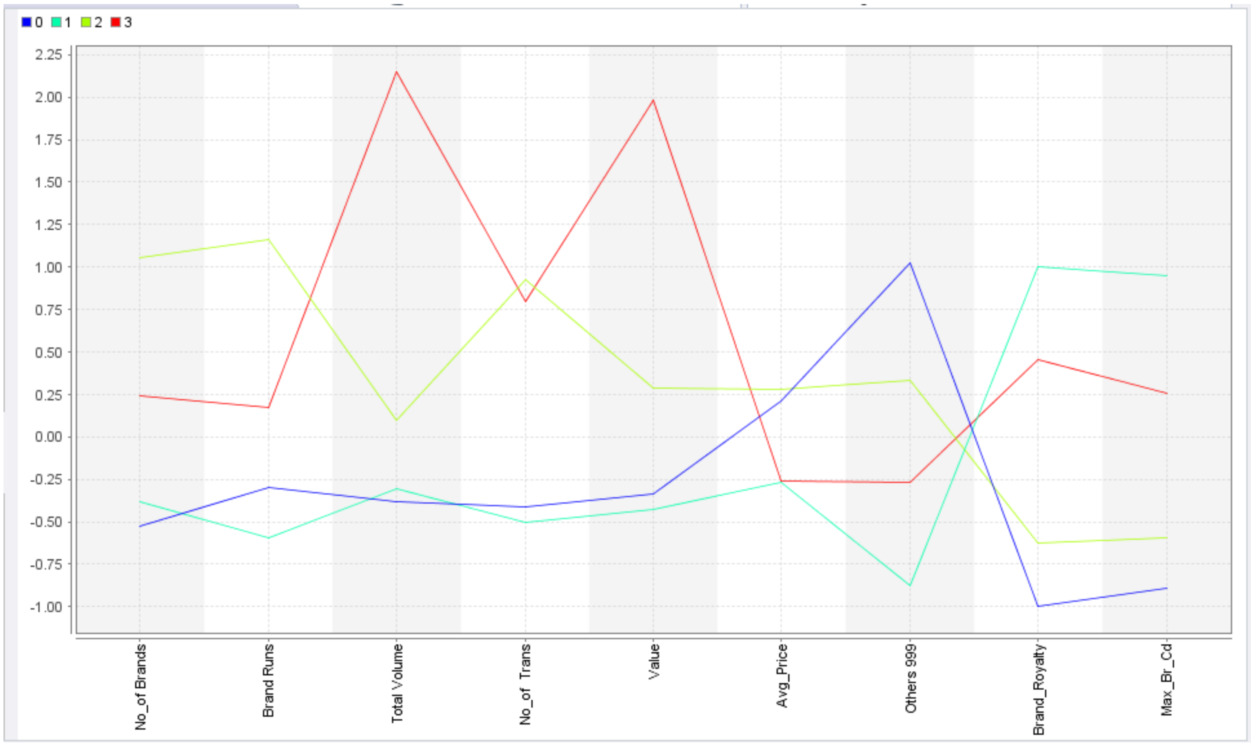
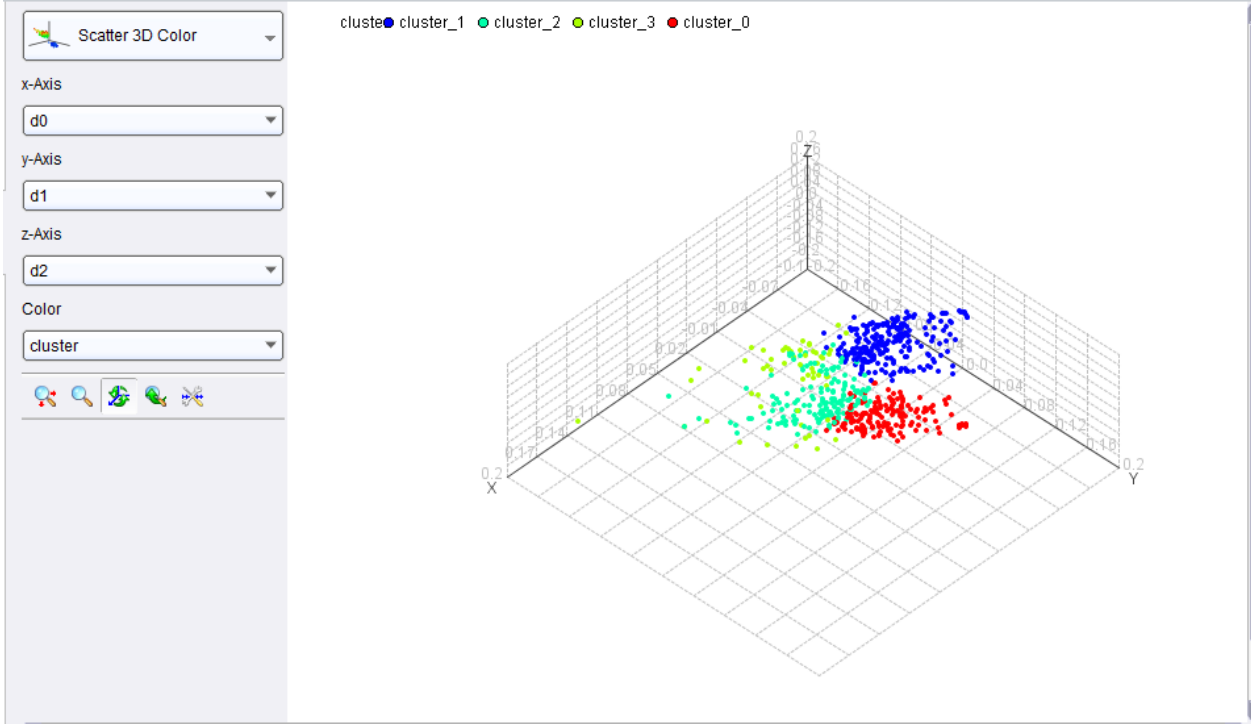


Figure 6: 3D Scatter Chart K = 4



Scatter 3D Color

x-Axis

d0

y-Axis

d1

z-Axis

d2

Color

cluster

- **K = 5**

Figure 7: Centroid Plot View K = 5

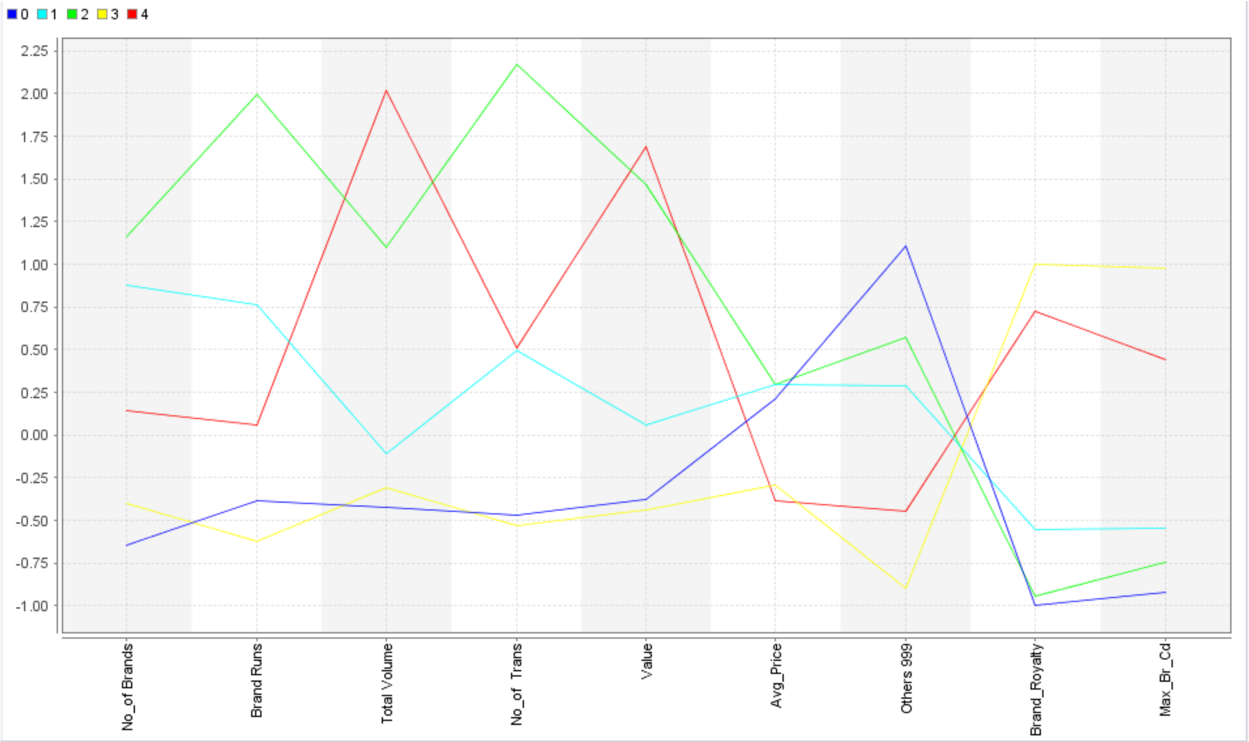
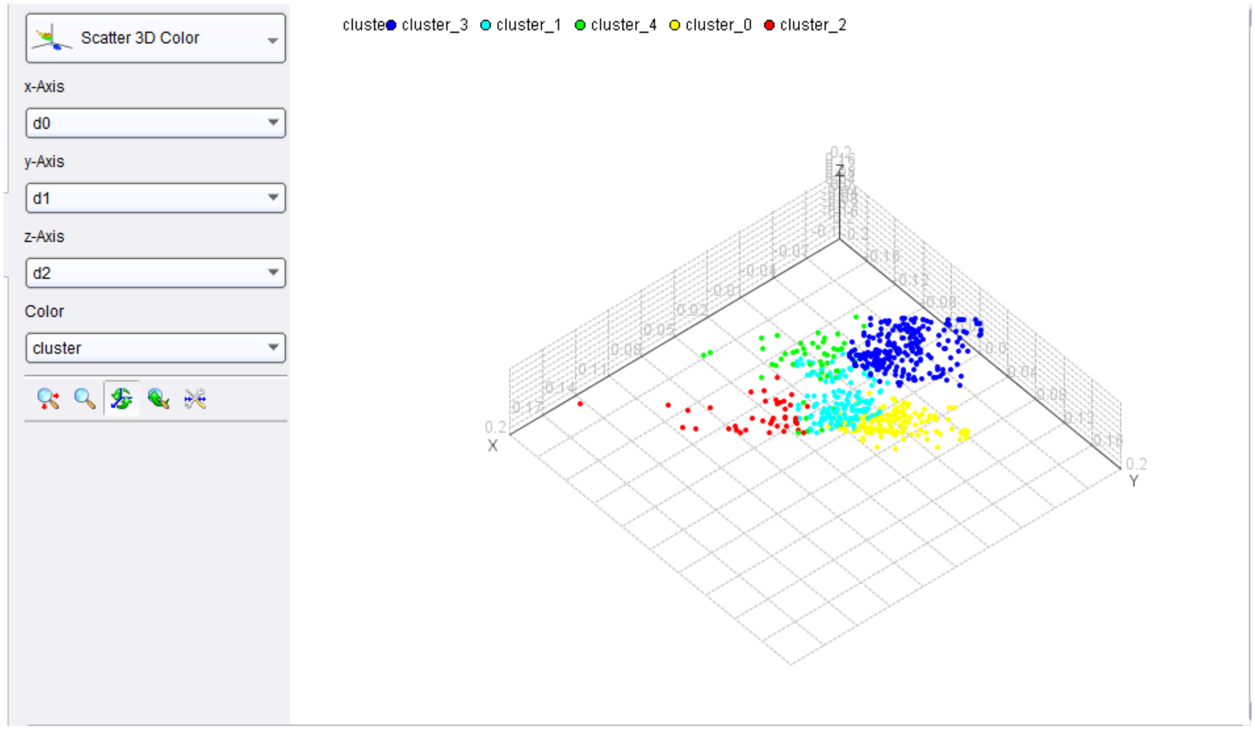


Figure 8: Centroid Plot View K = 5


Scatter 3D Color

x-Axis

d0

y-Axis

d1

z-Axis

d2

Color

cluster

cluste● cluster_3 ○ cluster_1 ○ cluster_4 ○ cluster_0 ● cluster_2
Z
X
Y
0.2
0.2

Answer 1-b:

- **K=2**

Figure 9: Centroid Plot View K = 2

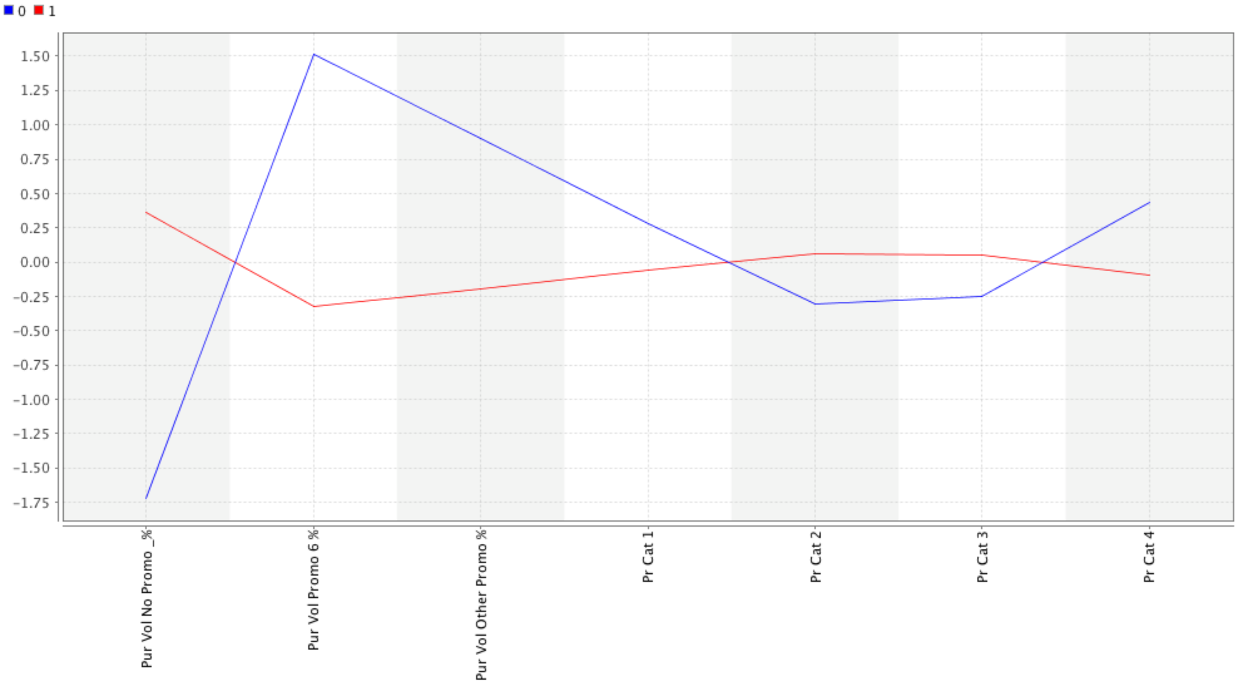
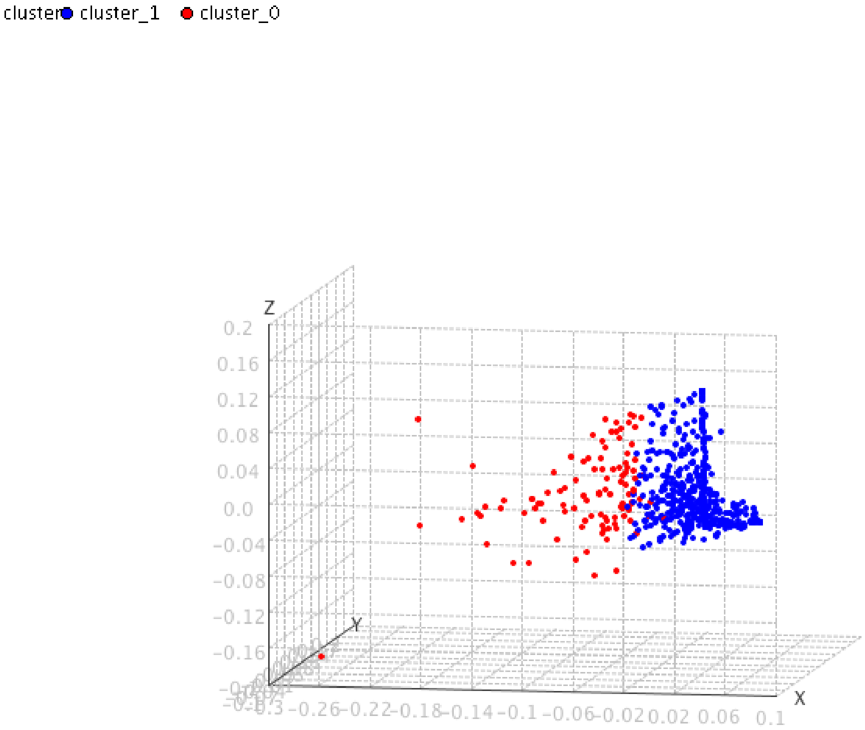


Figure 10: 3D Scatter chart K = 2



- **K=3**

Figure 11: Centroid Plot View K = 3

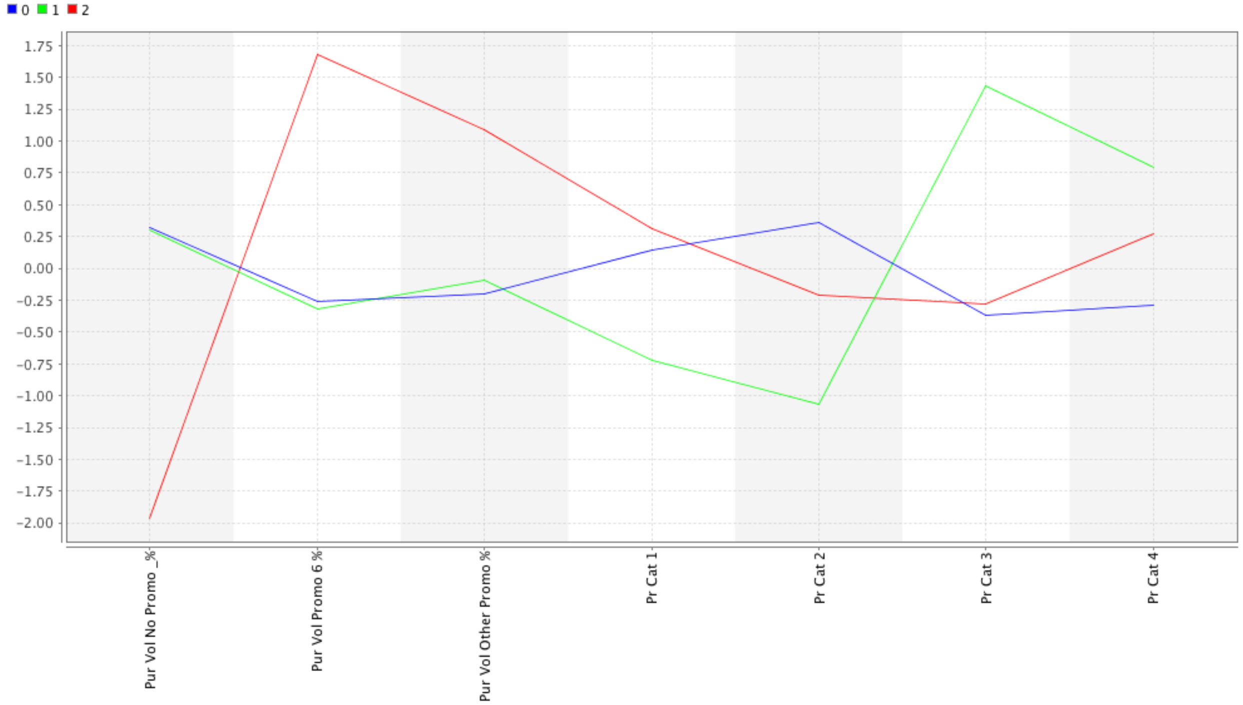
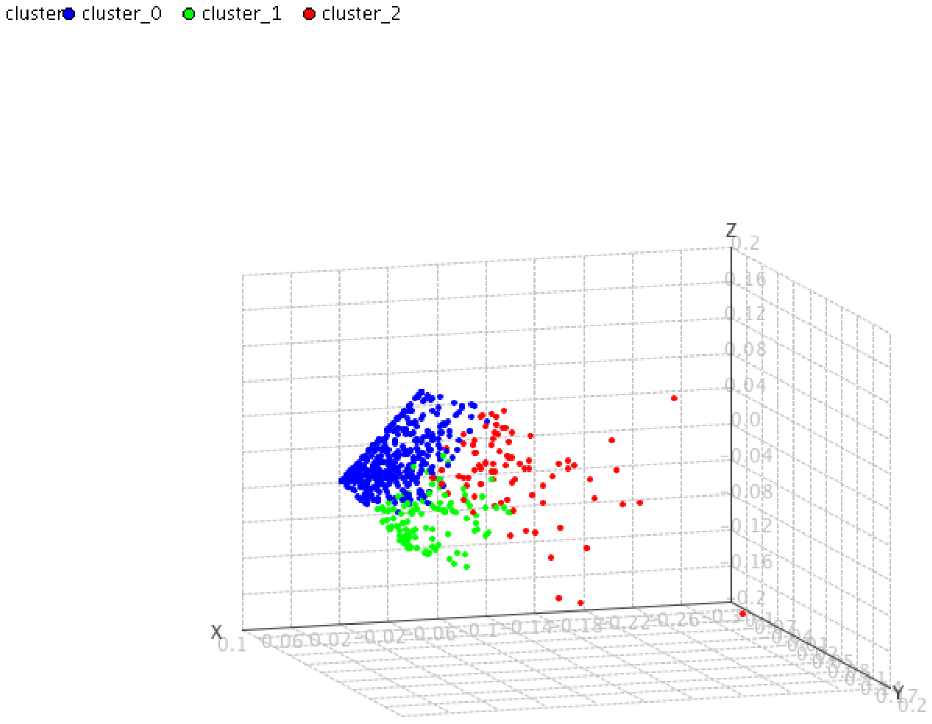


Figure 12: 3D Scatter chart K = 3



- **K=4**

Figure 13: Centroid Plot View K = 4

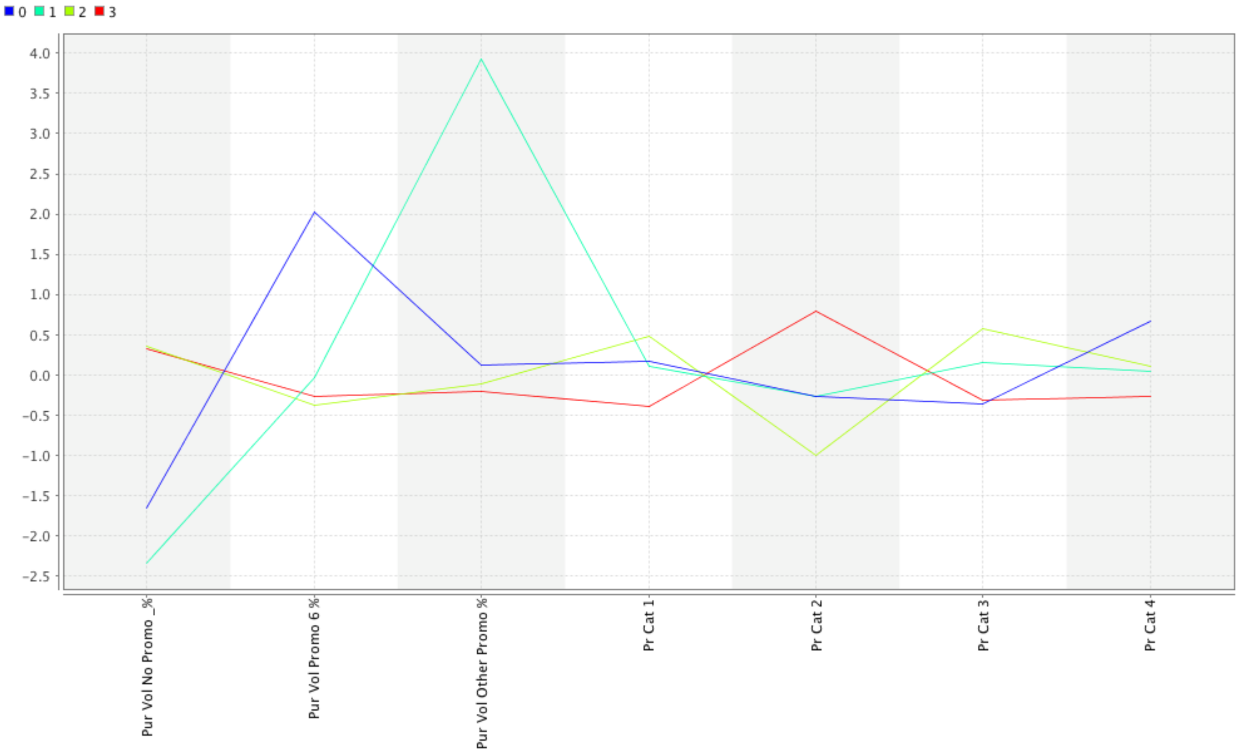
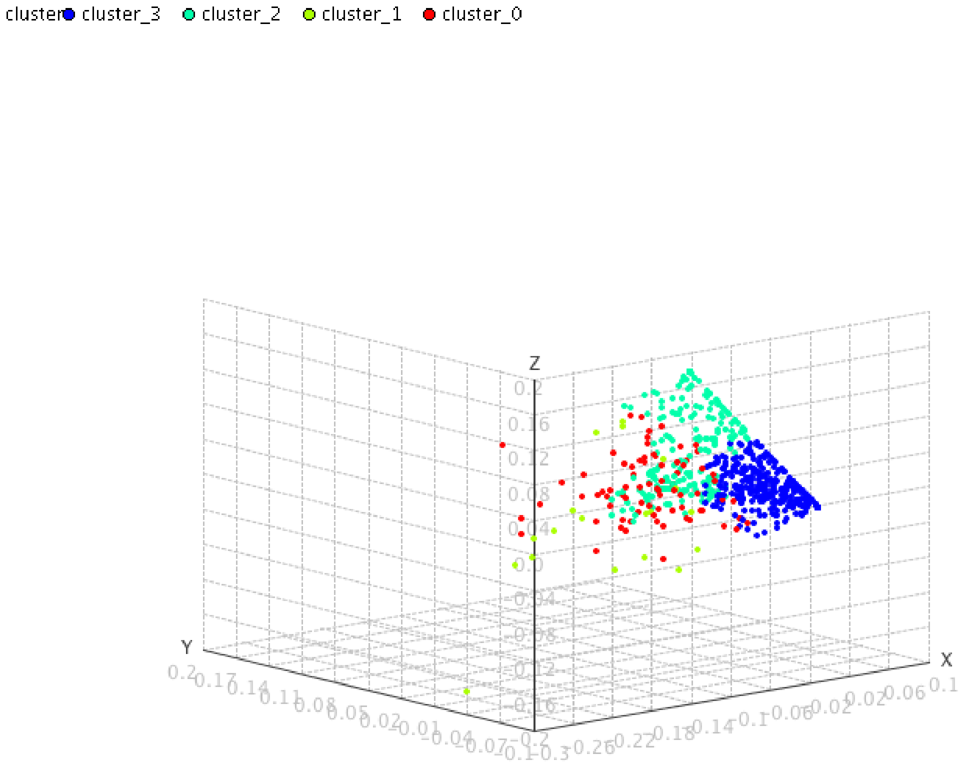


Figure 14: 3D Scatter chart K = 4



- **K=5**

Figure 15: Centroid Plot View K = 5

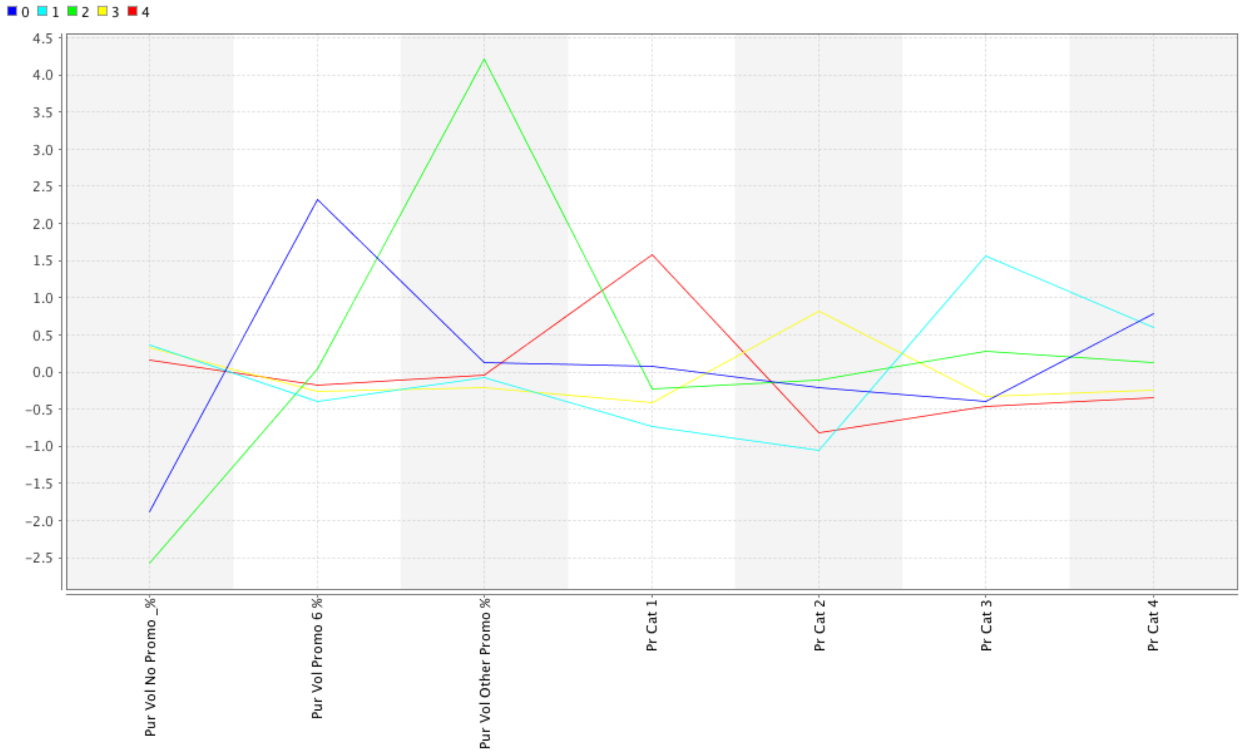
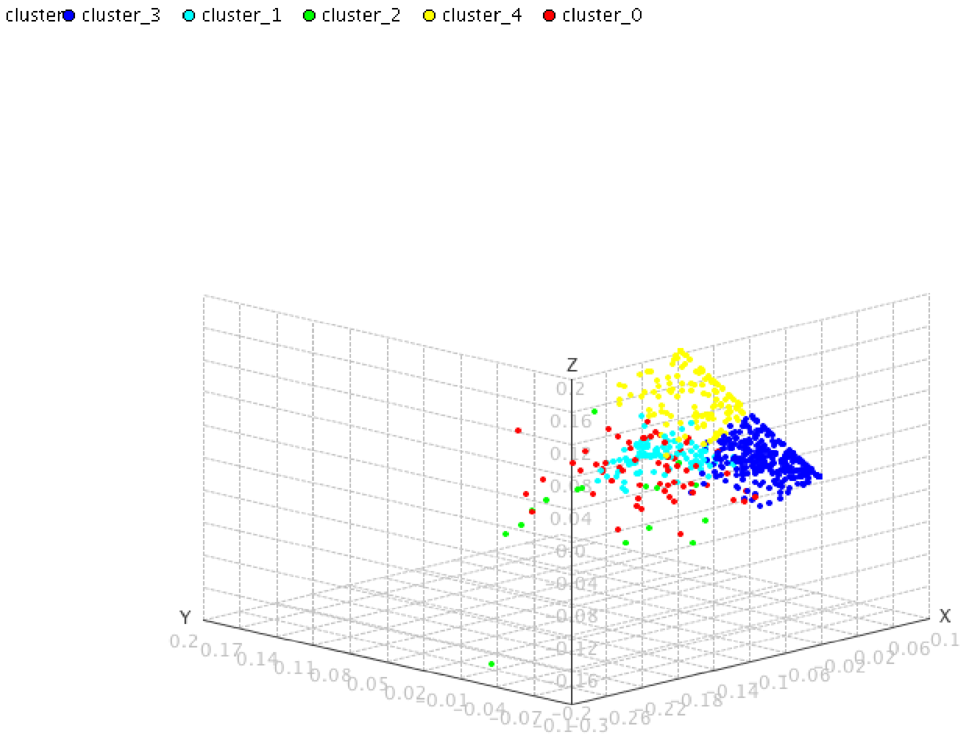


Figure 16: 3D Scatter chart K = 5



Answer 1-c:

- **K= 2**

Figure 17: Centroid Plot View K = 2

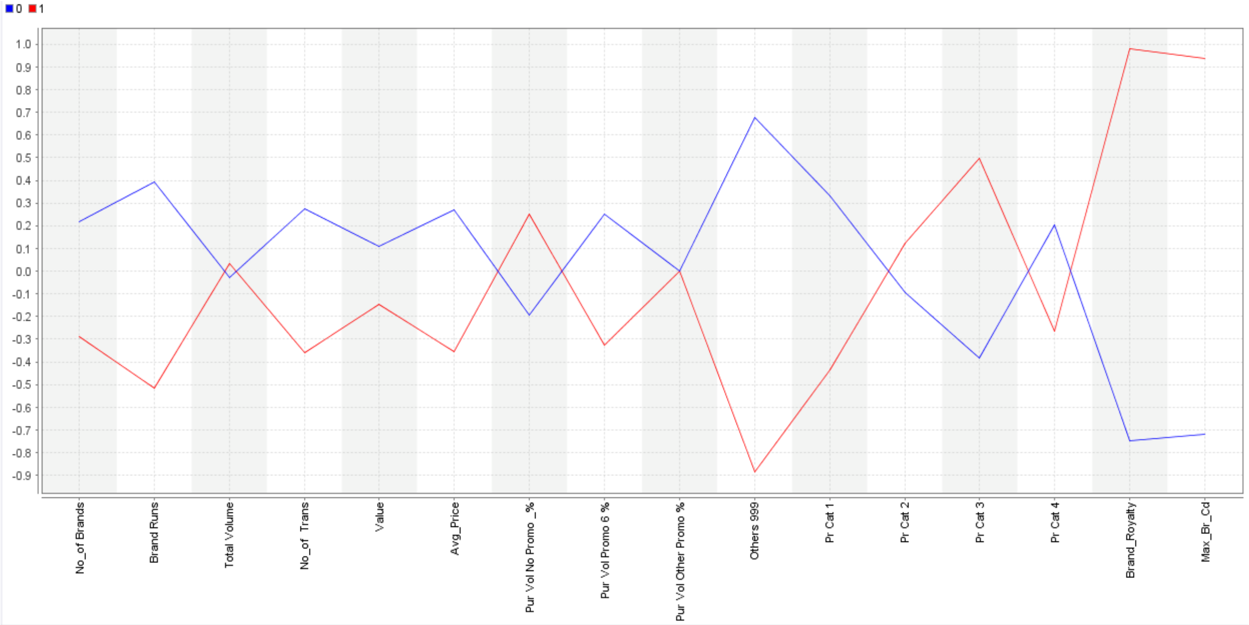
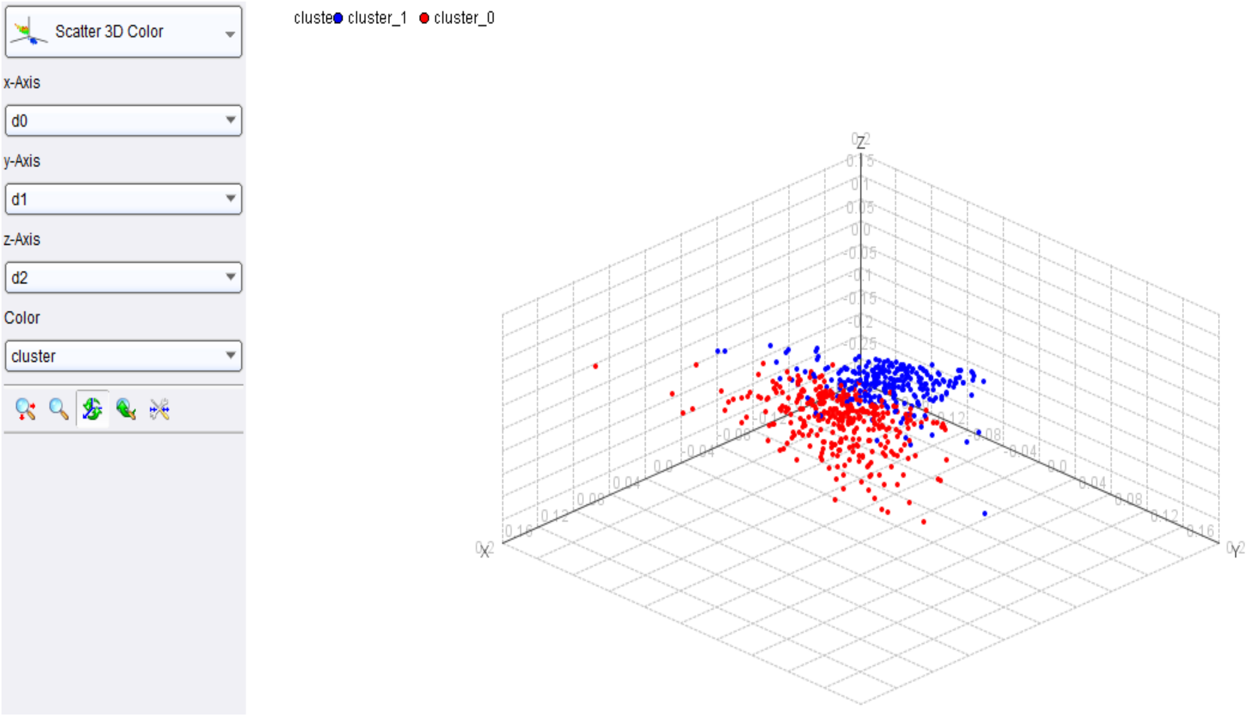


Figure 18: 3D Scatter chart K = 2



x-Axis
d0

y-Axis
d1

z-Axis
d2

Color
cluster

- **K = 3**

Figure 19: Centroid Plot View K = 3

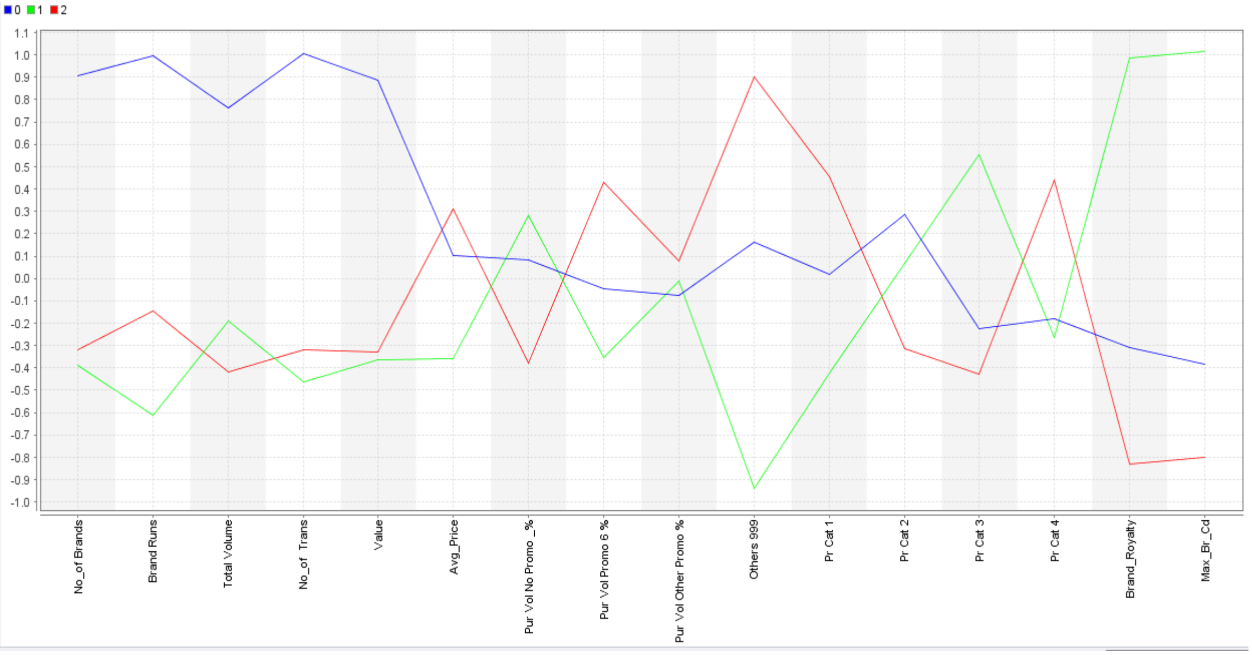
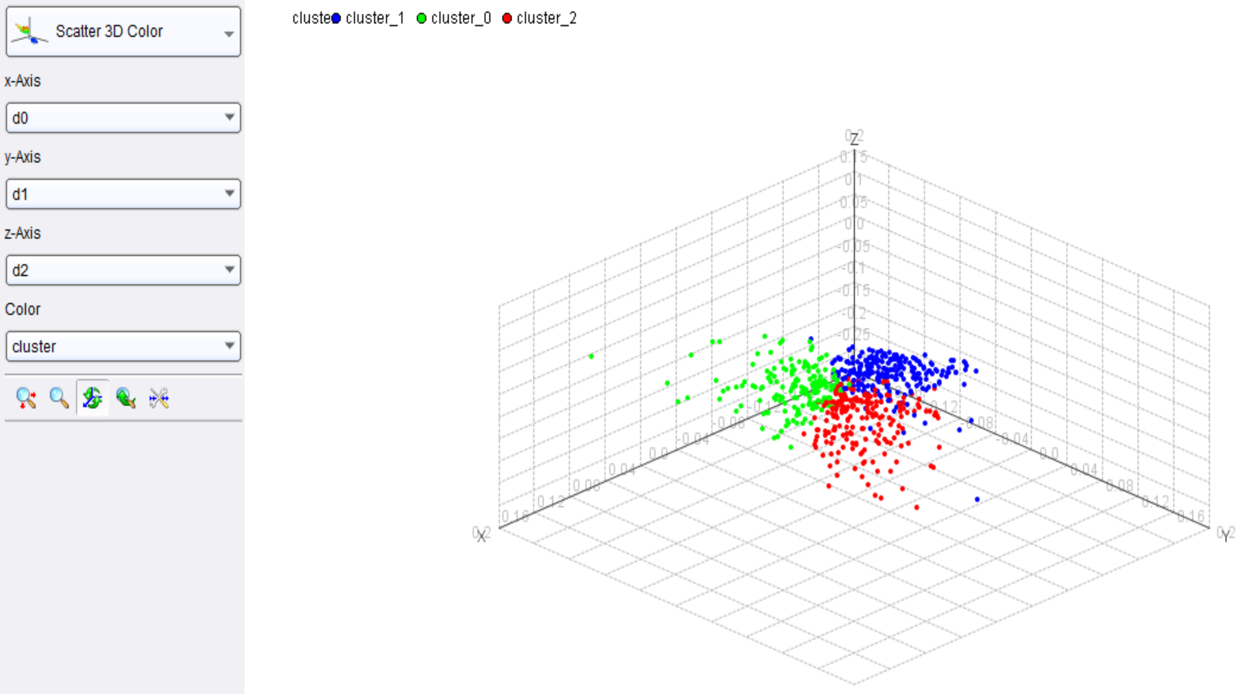


Figure 20: 3D Scatter chart K = 3


Scatter 3D Color

x-Axis

d0

y-Axis

d1

z-Axis

d2

Color

cluster

- **K = 4**

Figure 21: Centroid Plot View K = 4

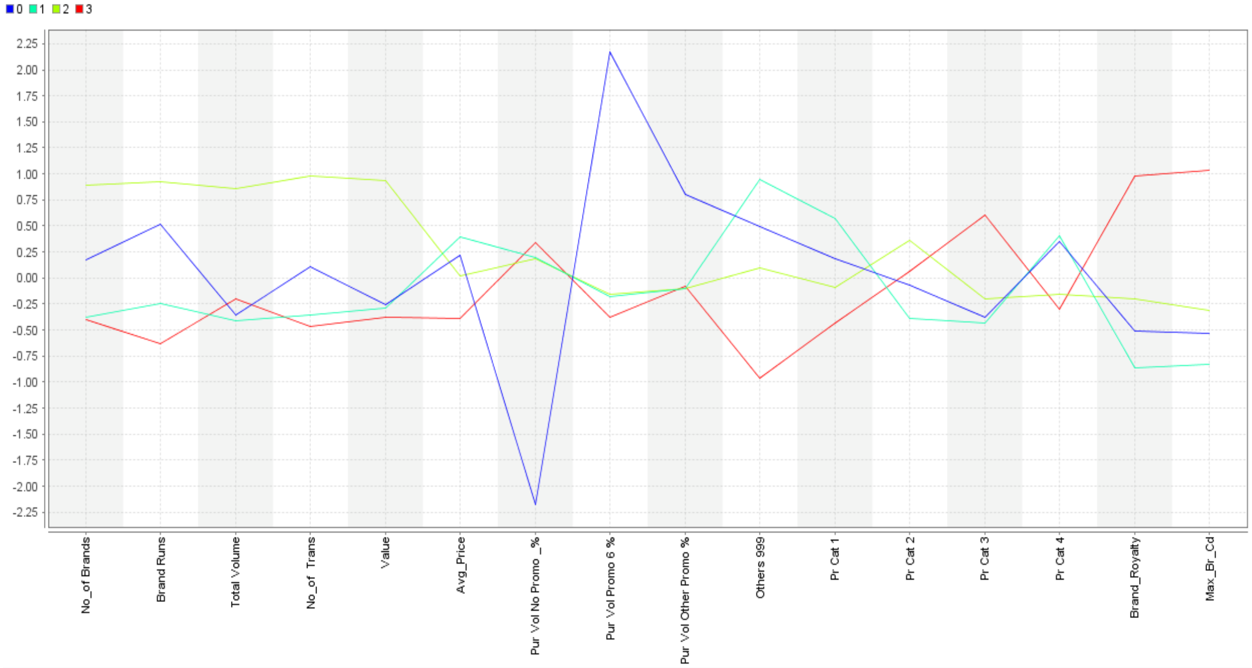
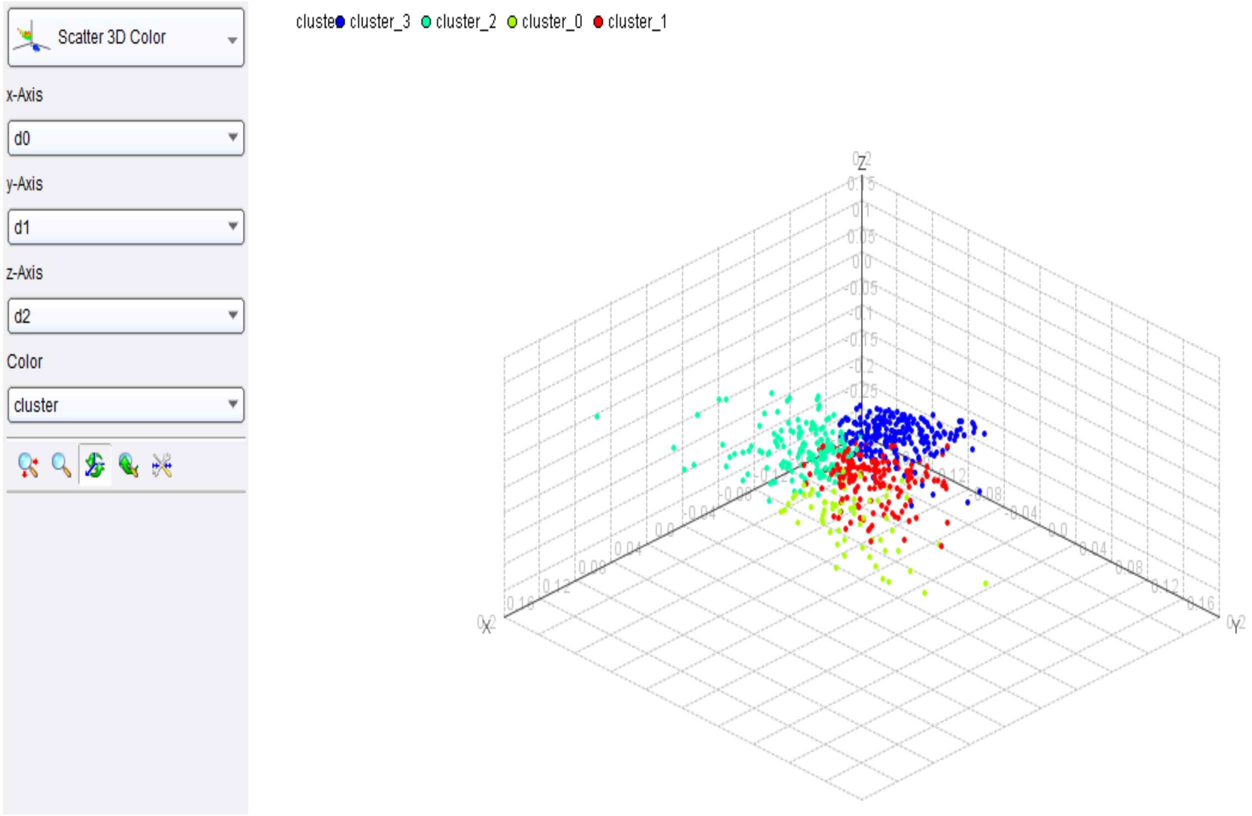


Figure 22: 3D Scatter chart K = 4



x-Axis

d0

y-Axis

d1

z-Axis

d2

Color

cluster

- **K = 5**

Figure 23: Centroid Plot View K = 5

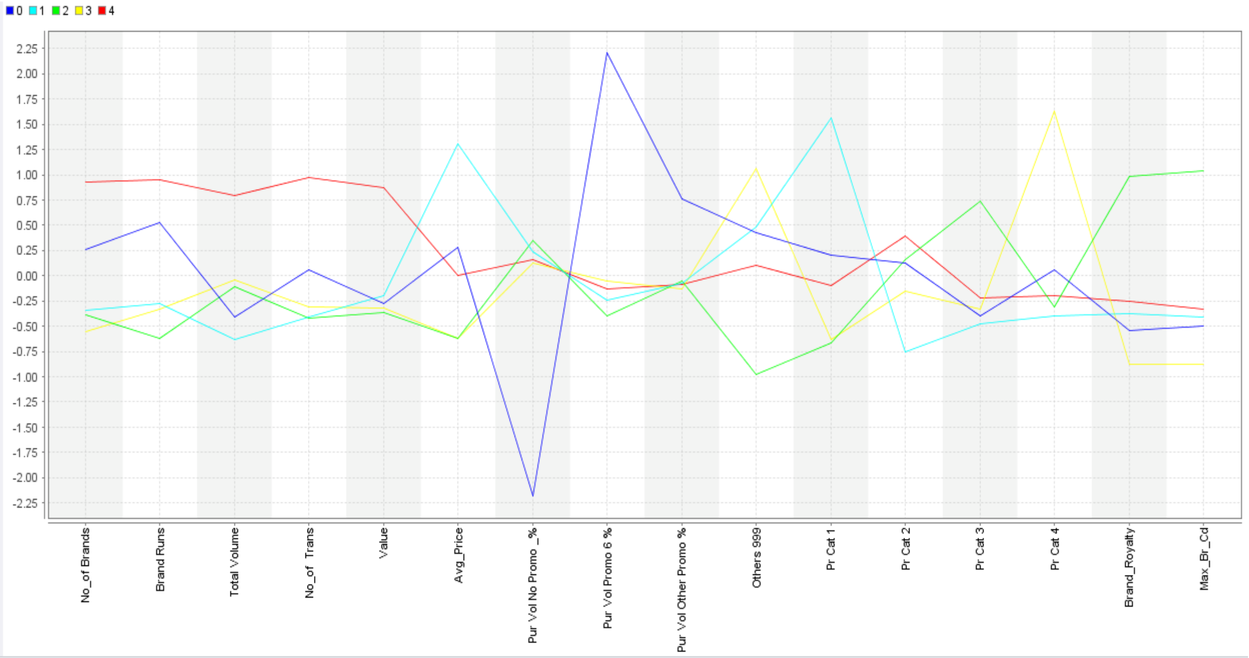
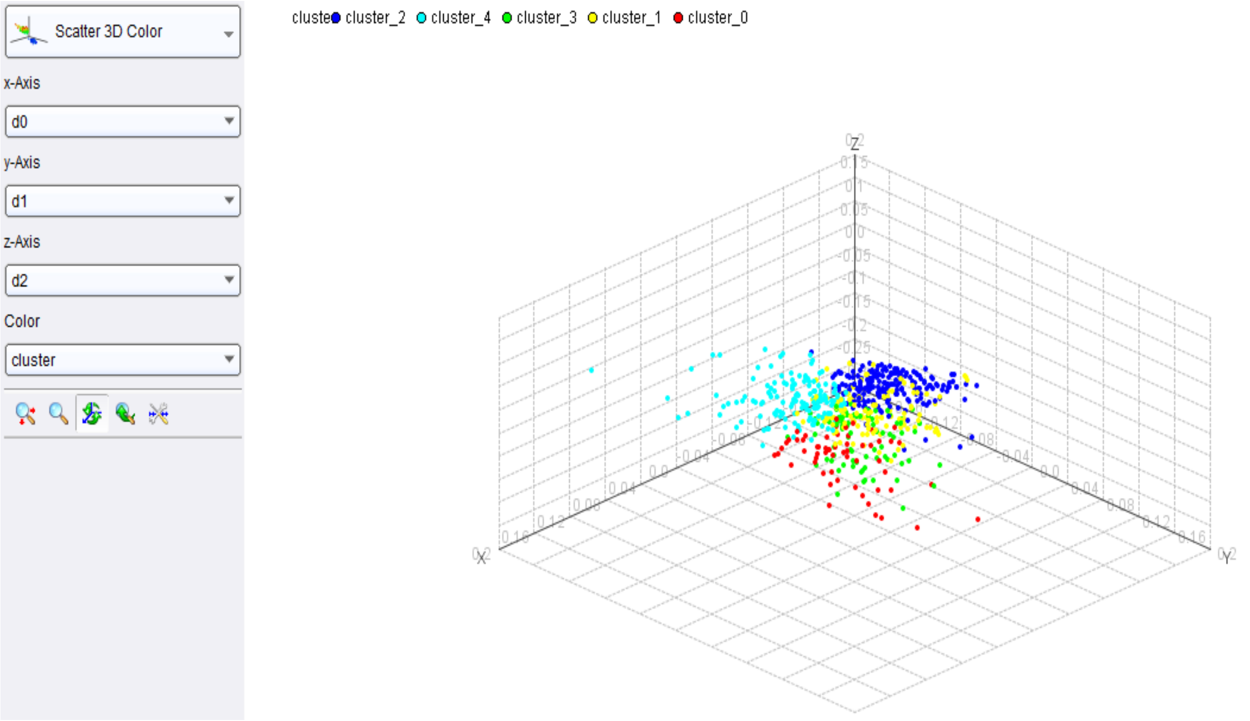


Figure 24: 3D Scatter chart K = 5


Scatter 3D Color

x-Axis

d0

y-Axis

d1

z-Axis

d2

Color

cluster

cluste● cluster_2 ● cluster_4 ● cluster_3 ● cluster_1 ● cluster_0
Z
X
Y
0.2
0.15
0.1
0.05
0.0
-0.05
-0.1
-0.15
-0.2
-0.25
0.04
0.08
0.12
0.16

Answer 1-d:

k-medoids:

- **K = 2**

Figure 25: Centroid Plot View K = 2

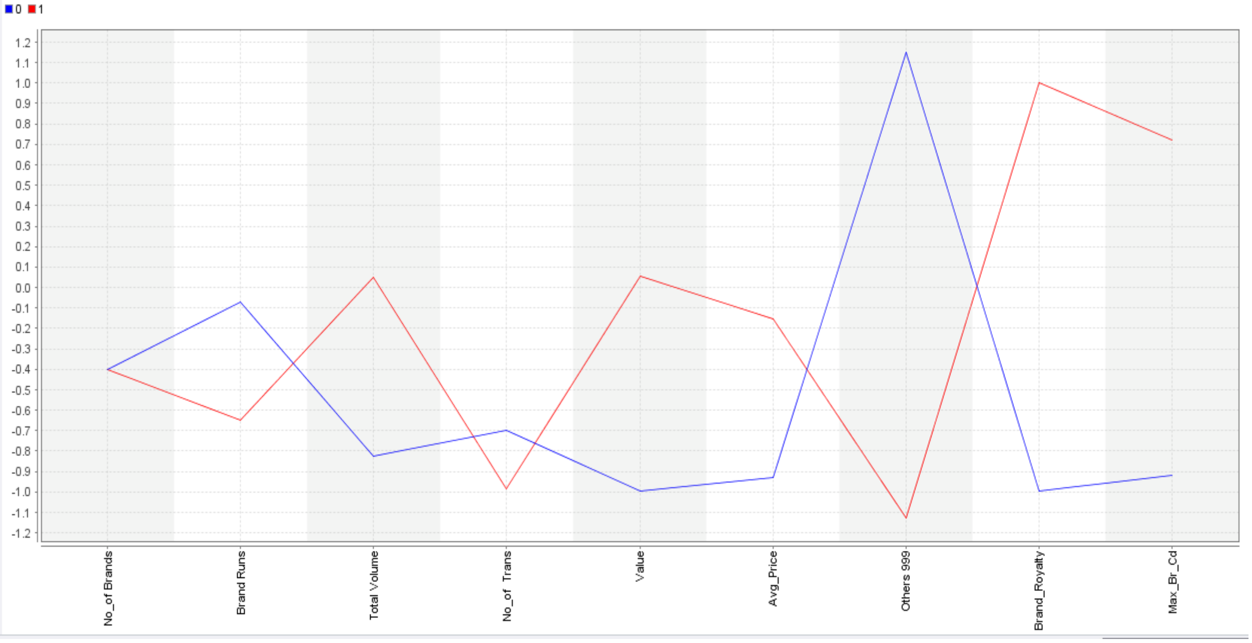
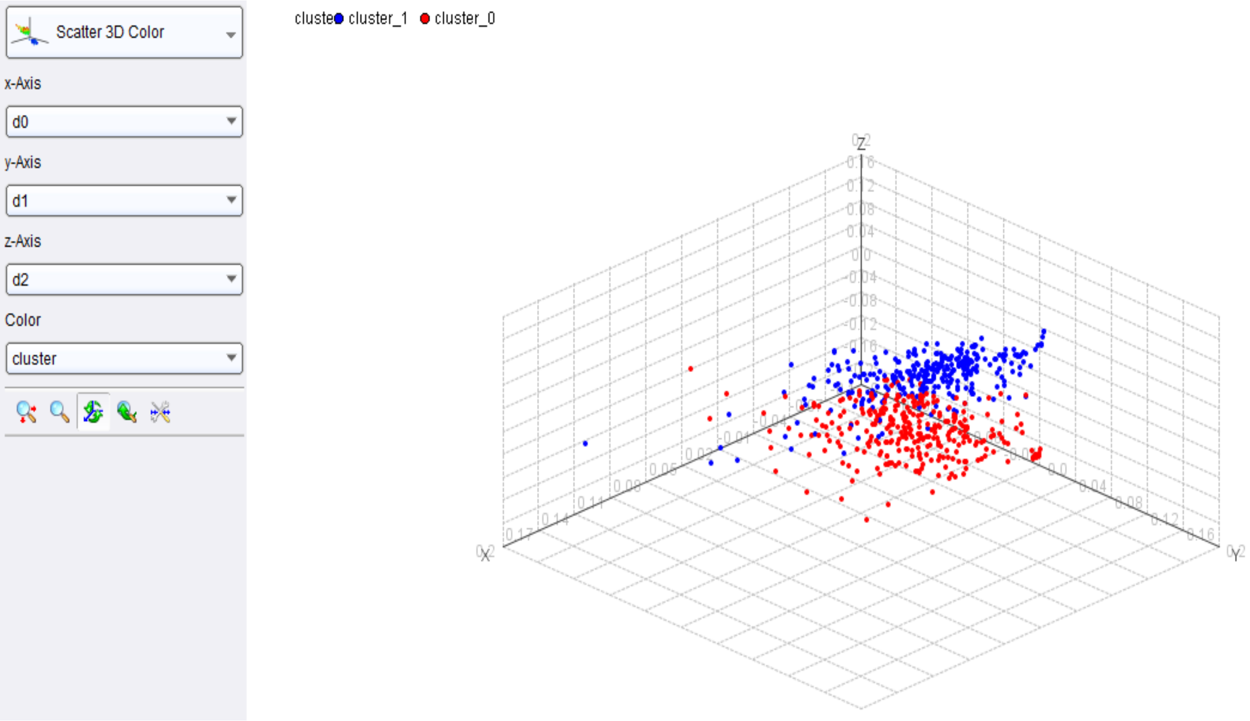


Figure 26: 3D Scatter chart K = 2



Scatter 3D Color

x-Axis

d0

y-Axis

d1

z-Axis

d2

Color

cluster

- **K = 3**

Figure 27: Centroid Plot View K = 3

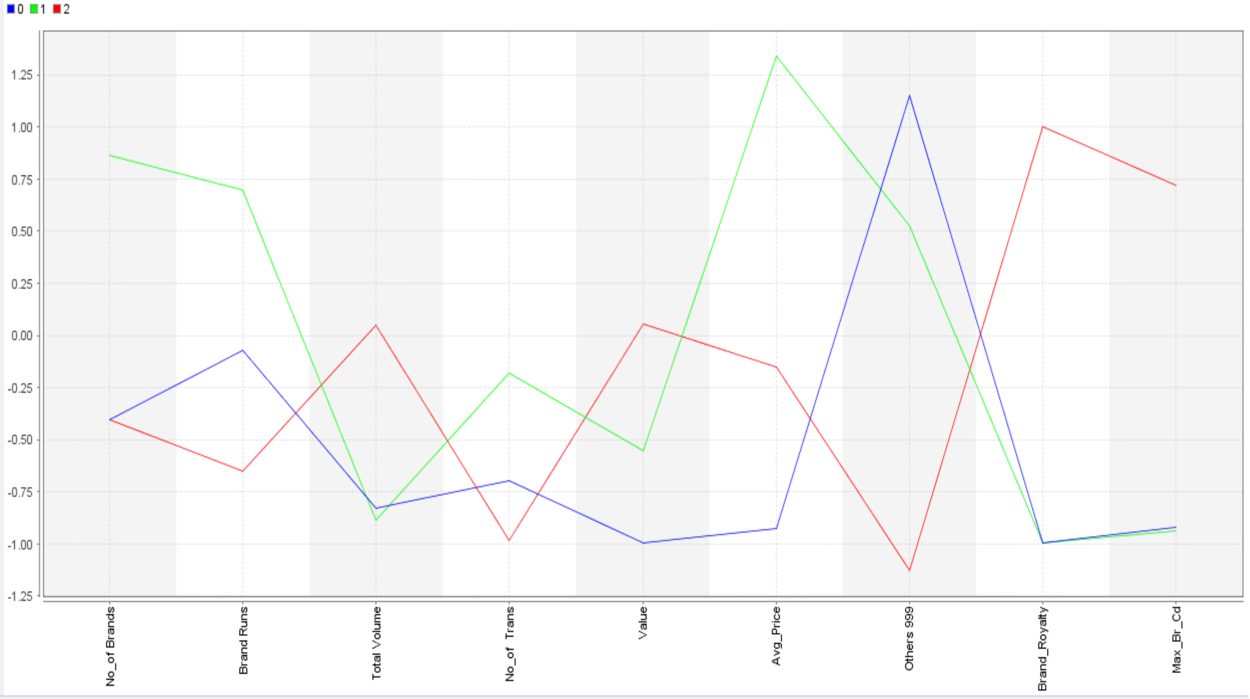
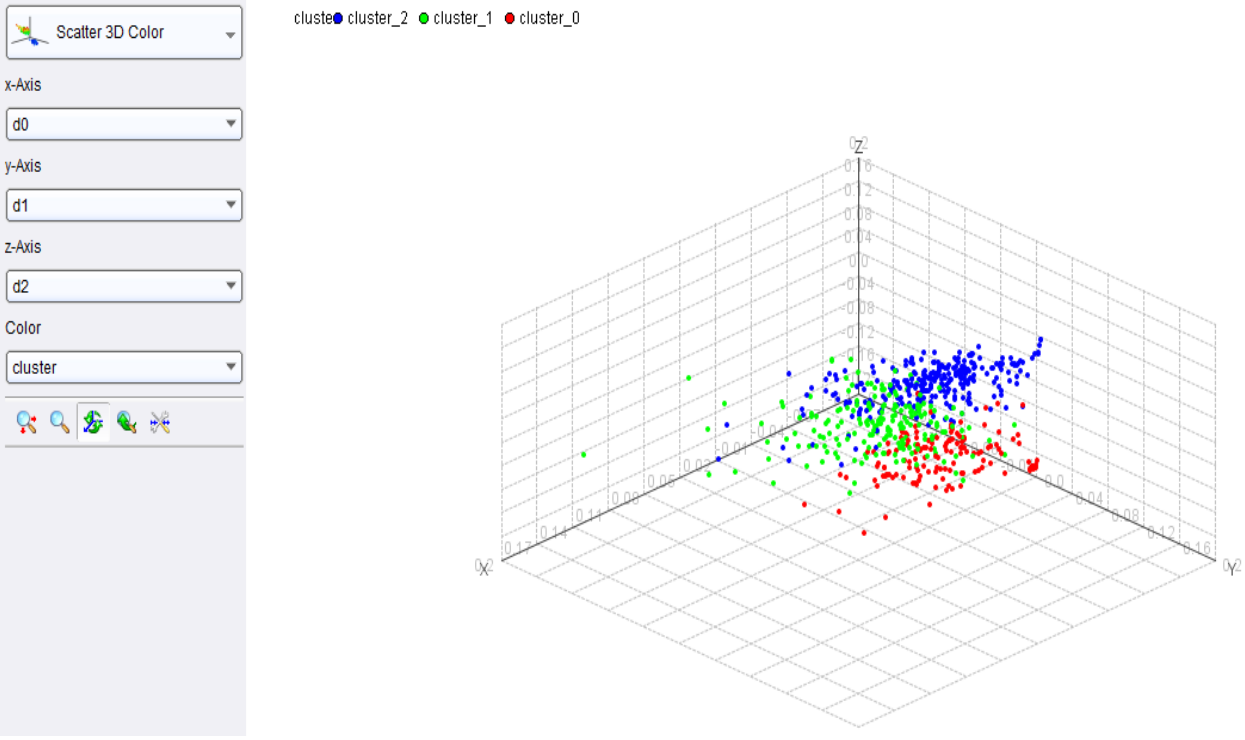


Figure 28: 3D Scatter chart K = 3


Scatter 3D Color

x-Axis

d0

y-Axis

d1

z-Axis

d2

Color

cluster

- **K = 4**

Figure 29: Centroid Plot View K = 4

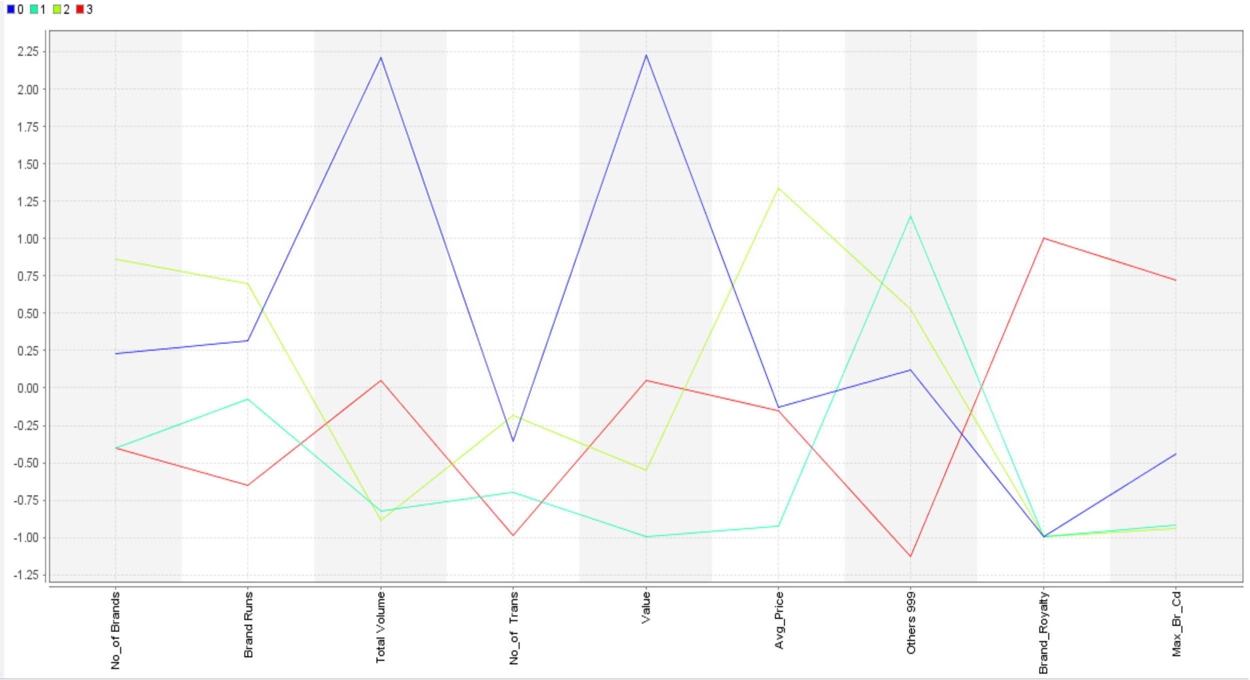
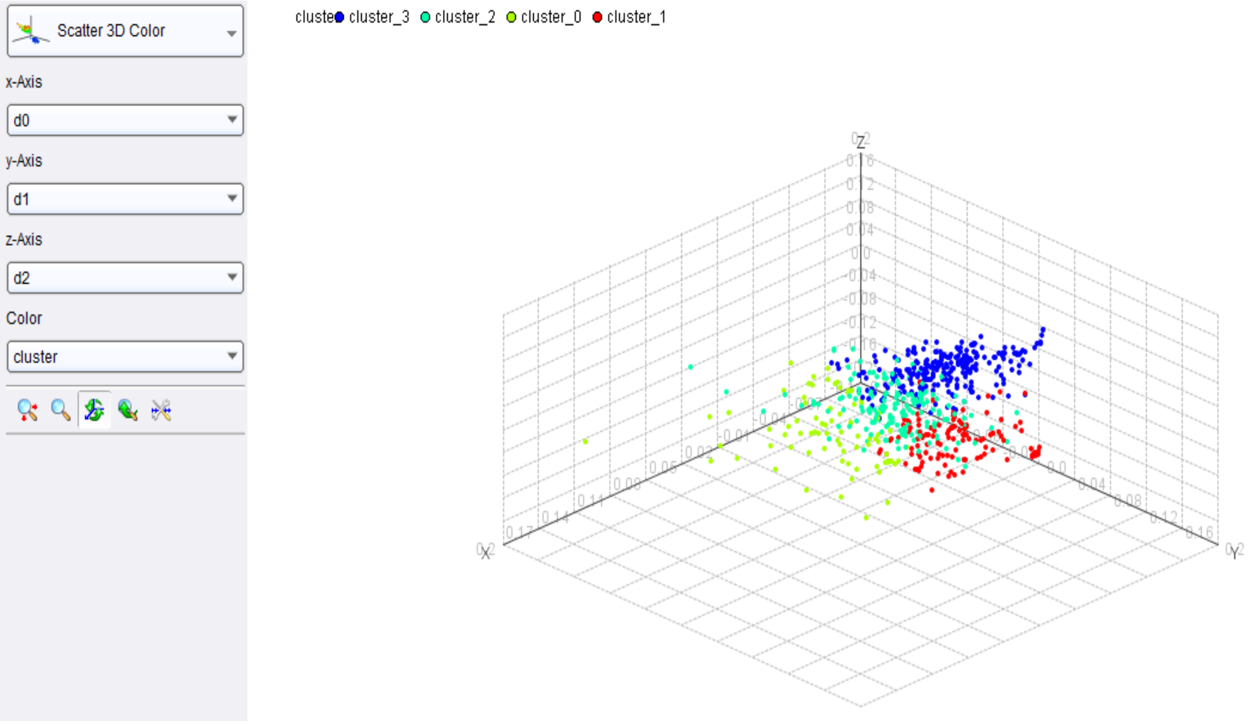


Figure 30: 3D Scatter chart K = 4


Scatter 3D Color

x-Axis

d0

y-Axis

d1

z-Axis

d2

Color

cluster

- **K = 5**

Figure 31: Centroid Plot View K = 5

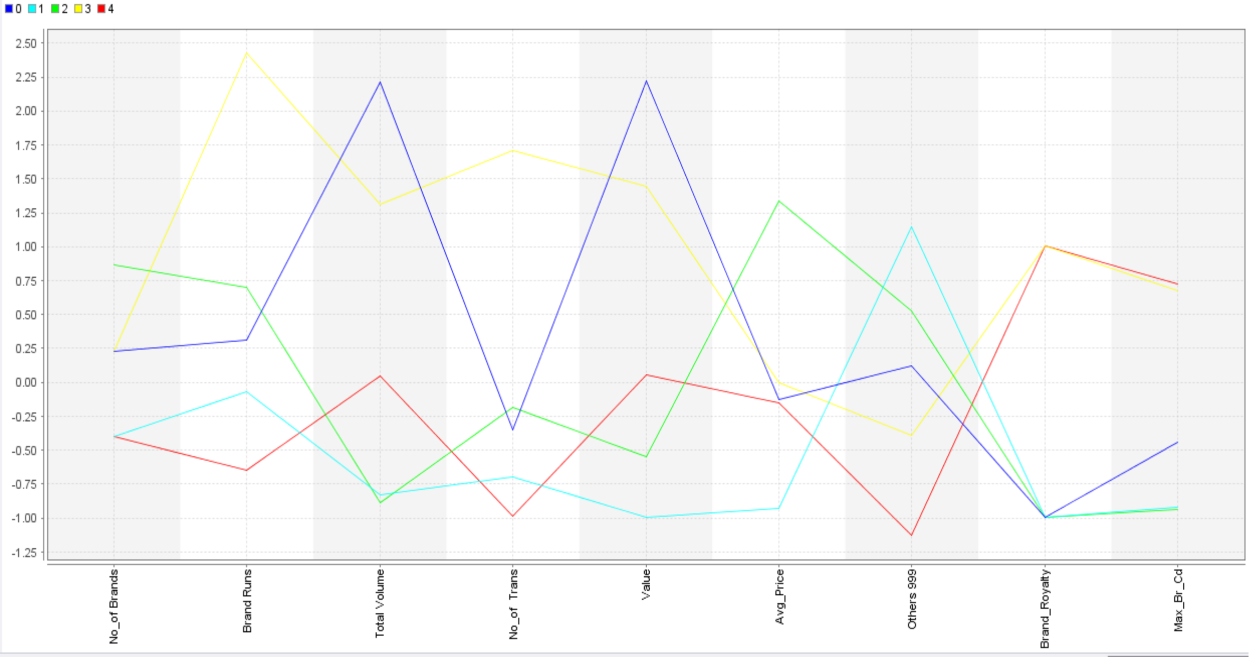
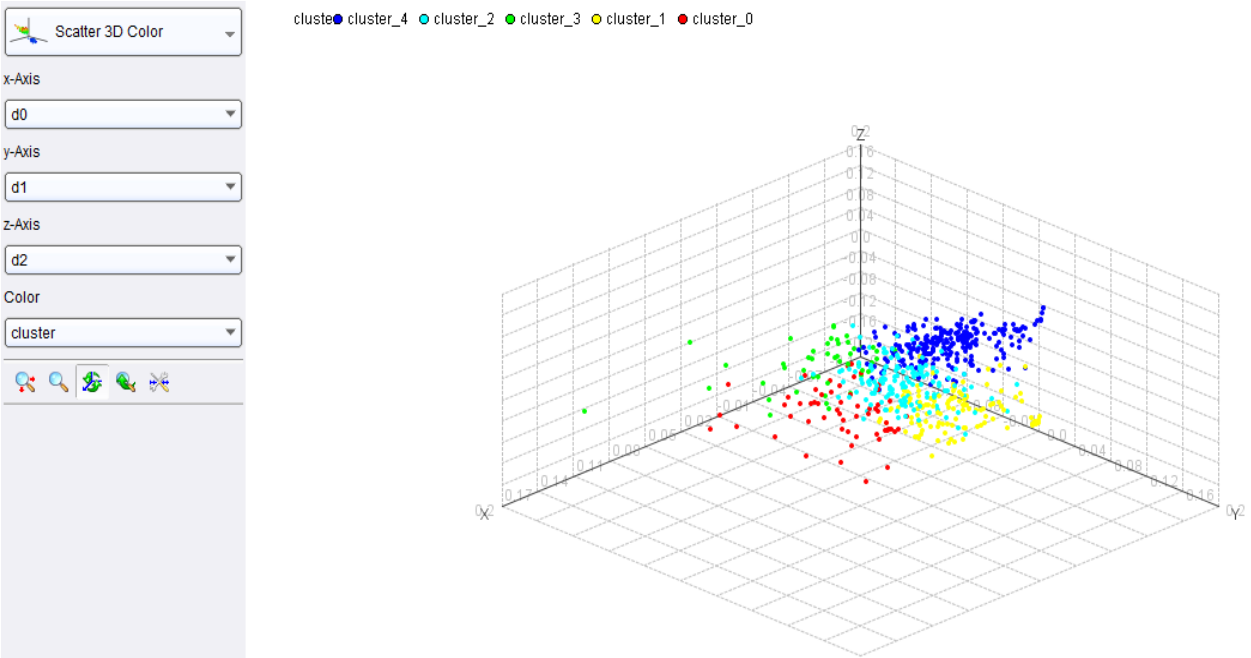


Figure 32: 3D Scatter chart K = 5


Scatter 3D Color

x-Axis

d0

y-Axis

d1

z-Axis

d2

Color

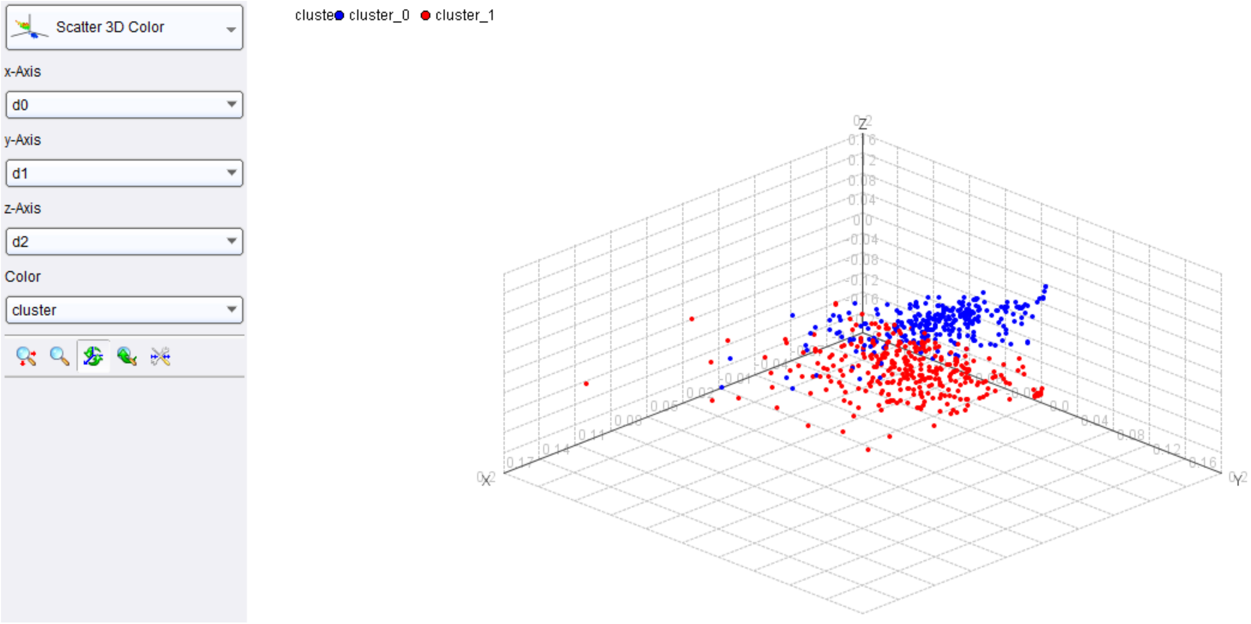
cluster

cluste● cluster_4 ○ cluster_2 ● cluster_3 ○ cluster_1 ● cluster_0
Z
X
Y

kernel k-means:

- **K = 2**

  Figure 33: 3D Scatter chart K = 2



Scatter 3D Color

x-Axis

d0

y-Axis

d1

z-Axis

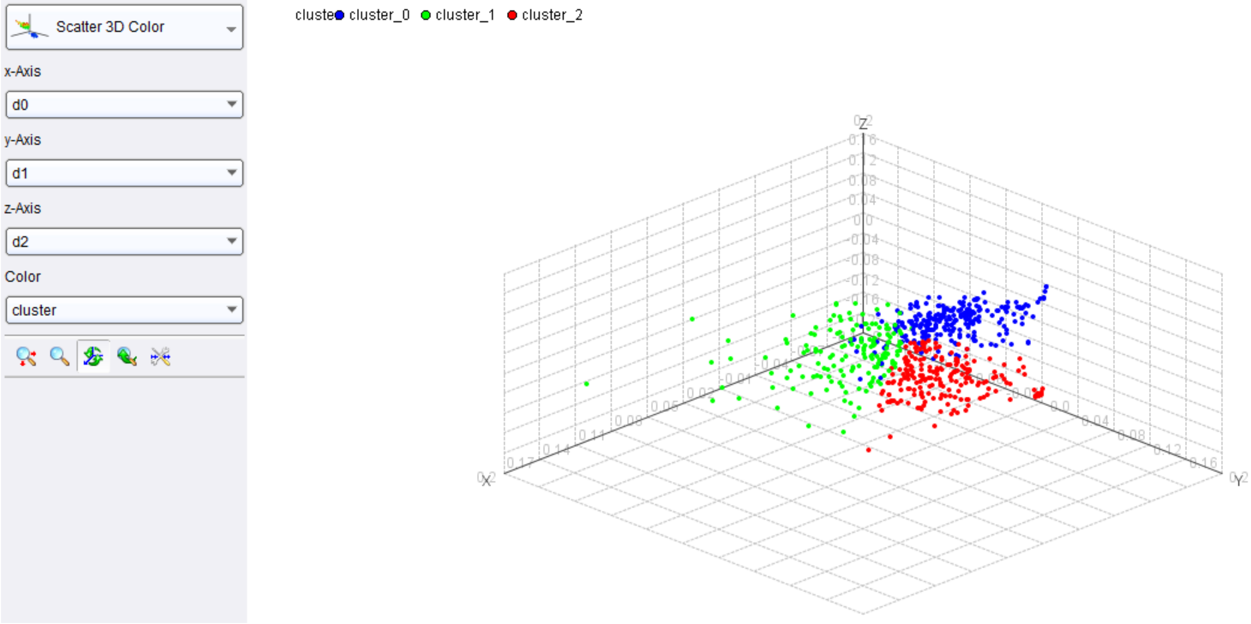
d2

Color

cluster

- **K = 3**

  Figure 34: 3D Scatter chart K = 3



Scatter 3D Color

x-Axis

d0

y-Axis

d1

z-Axis

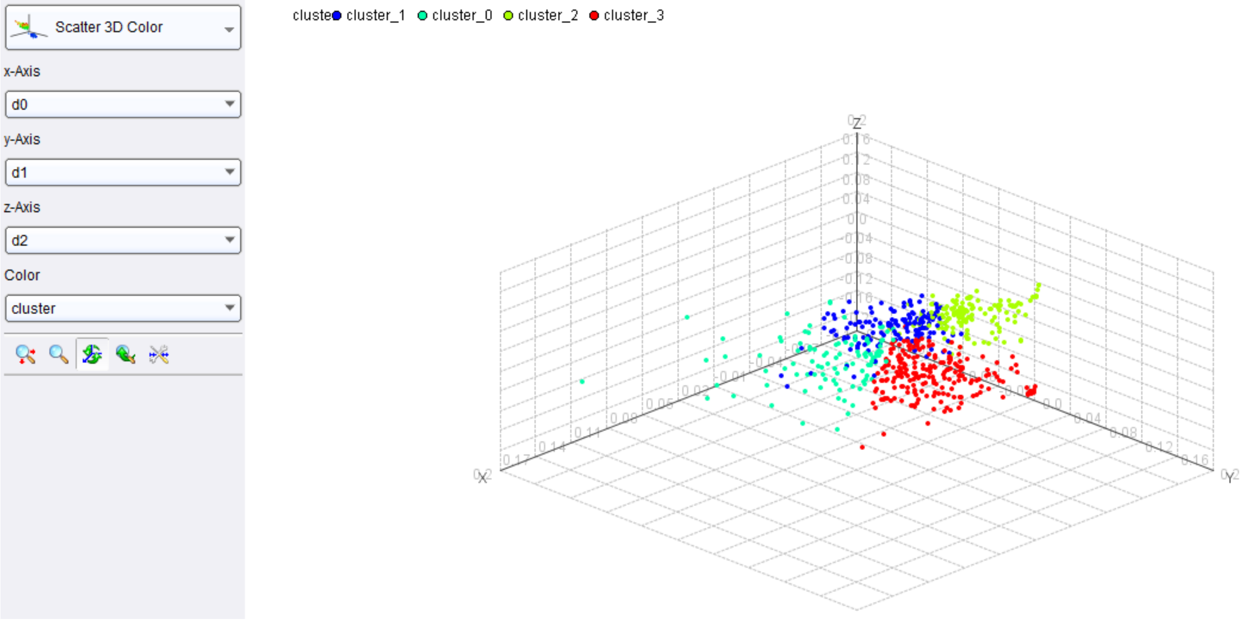
d2

Color

cluster

- **K = 4**

Figure 35: 3D Scatter chart K = 4



Scatter 3D Color

x-Axis
d0

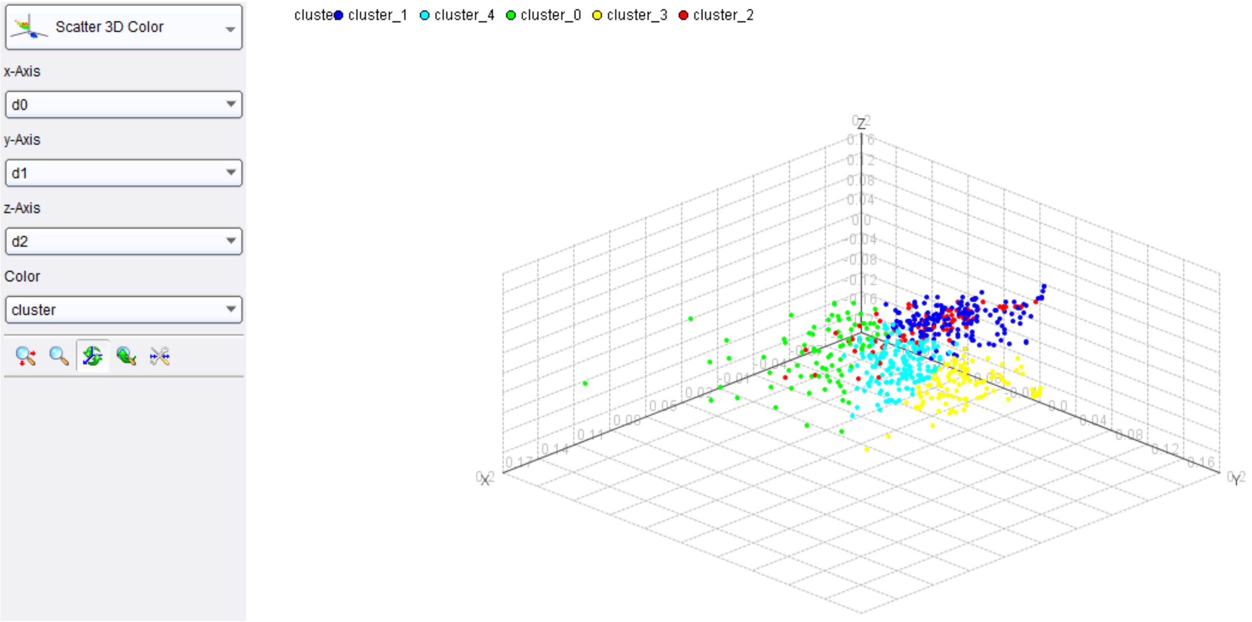
y-Axis
d1

z-Axis
d2

Color
cluster

- **K = 5**

Figure 36: 3D Scatter chart K = 5



Scatter 3D Color

x-Axis
d0

y-Axis
d1

z-Axis
d2

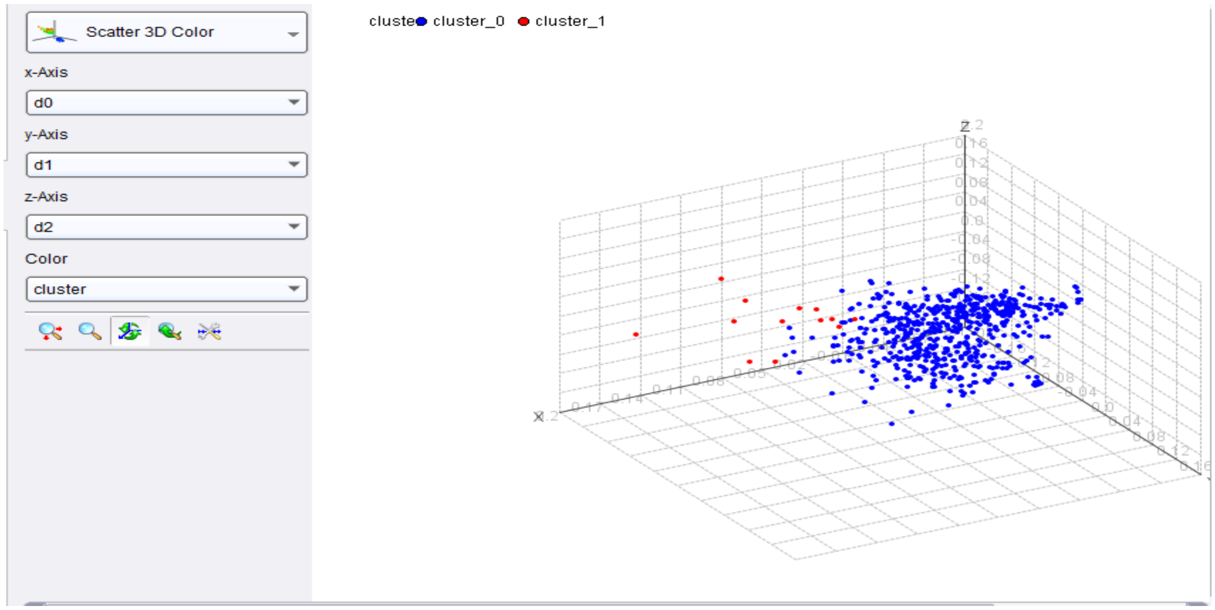
Color
cluster

cluste● cluster_1 ○ cluster_4 ● cluster_0 ○ cluster_3 ● cluster_2
Z
X
Y
0.2
0.16
0.12
0.08
0.04
0.0
-0.04
-0.08
-0.12
-0.16

Agglomerative Clustering:

- **K = 2**

Figure 37: 3D Scatter chart K = 2



Scatter 3D Color

x-Axis

d0

y-Axis

d1

z-Axis

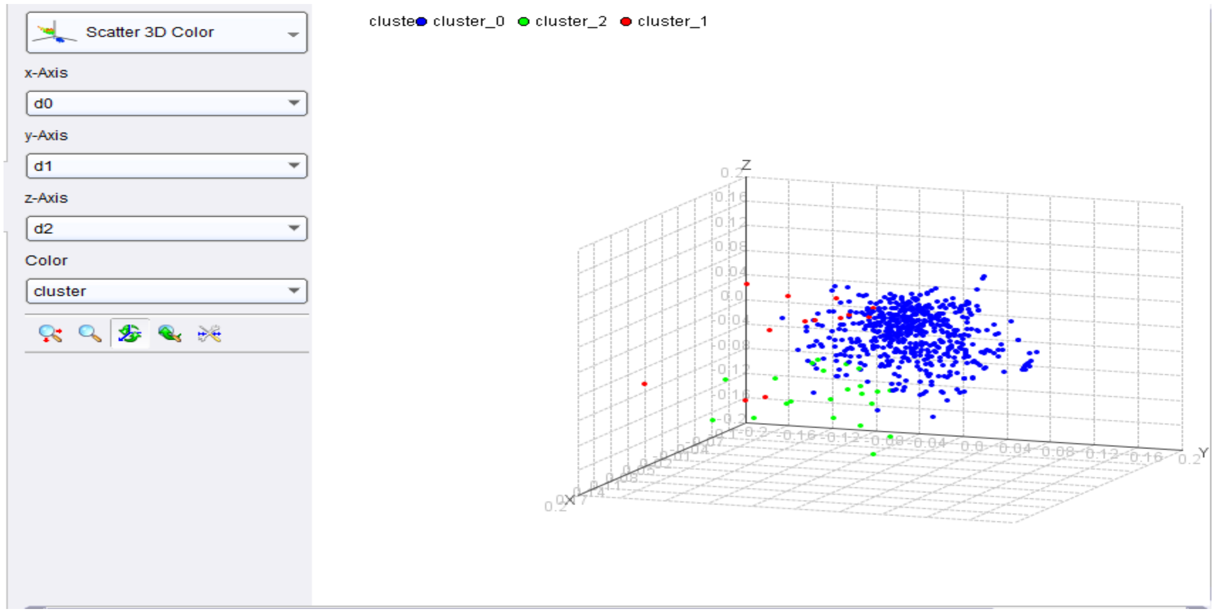
d2

Color

cluster

- **K = 3**

Figure 38: 3D Scatter chart K = 3



Scatter 3D Color

x-Axis

d0

y-Axis

d1

z-Axis

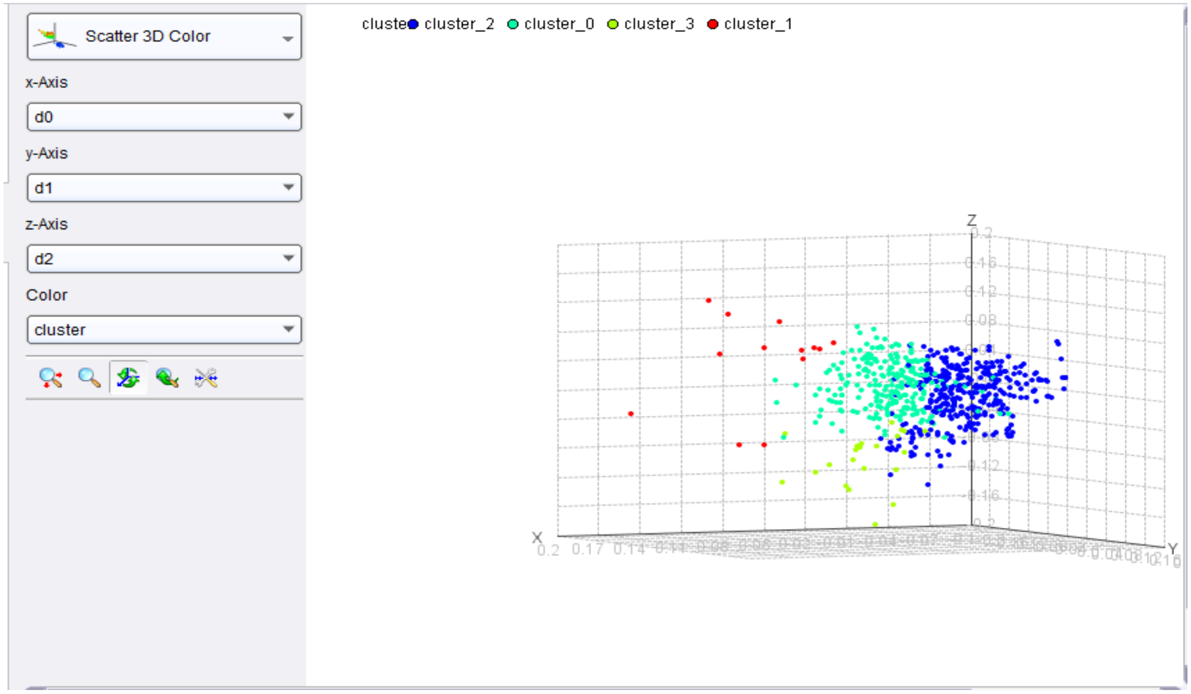
d2

Color

cluster

- **K = 4**

Figure 39: 3D Scatter chart K = 4



Scatter 3D Color

x-Axis

d0

y-Axis

d1

z-Axis

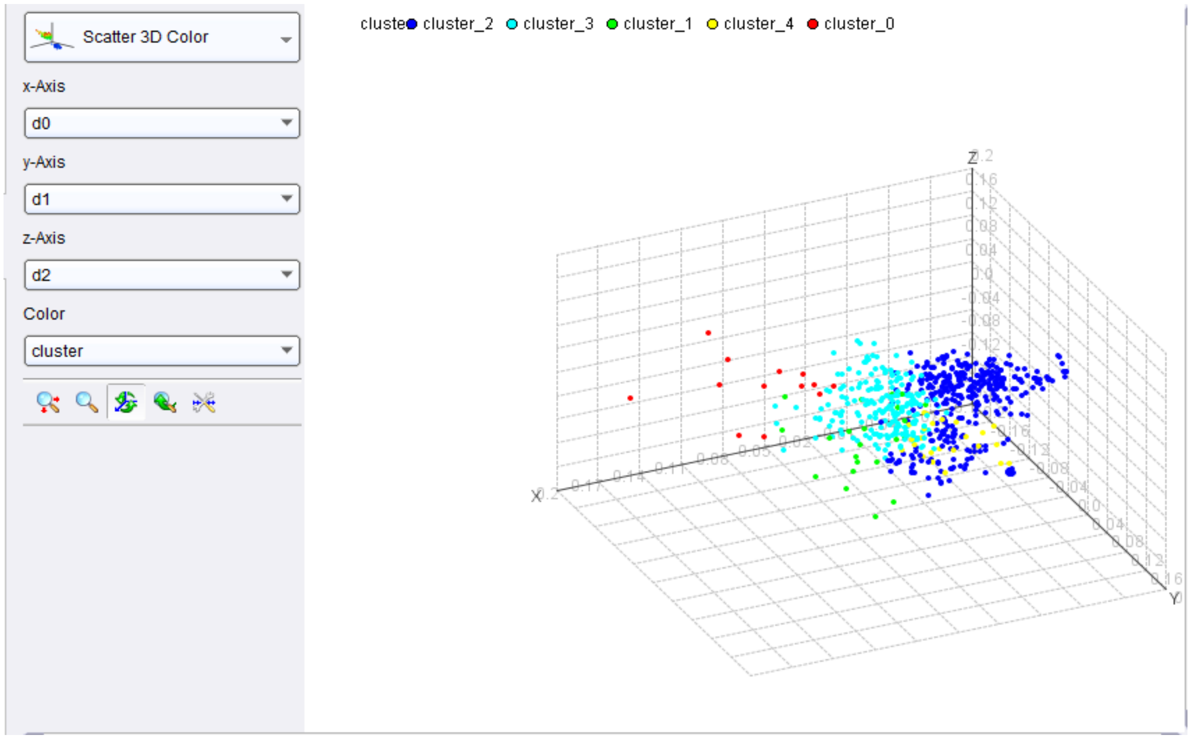
d2

Color

cluster

- **K = 5**

Figure 40: 3D Scatter chart K = 5



Scatter 3D Color

x-Axis

d0

y-Axis

d1

z-Axis

d2

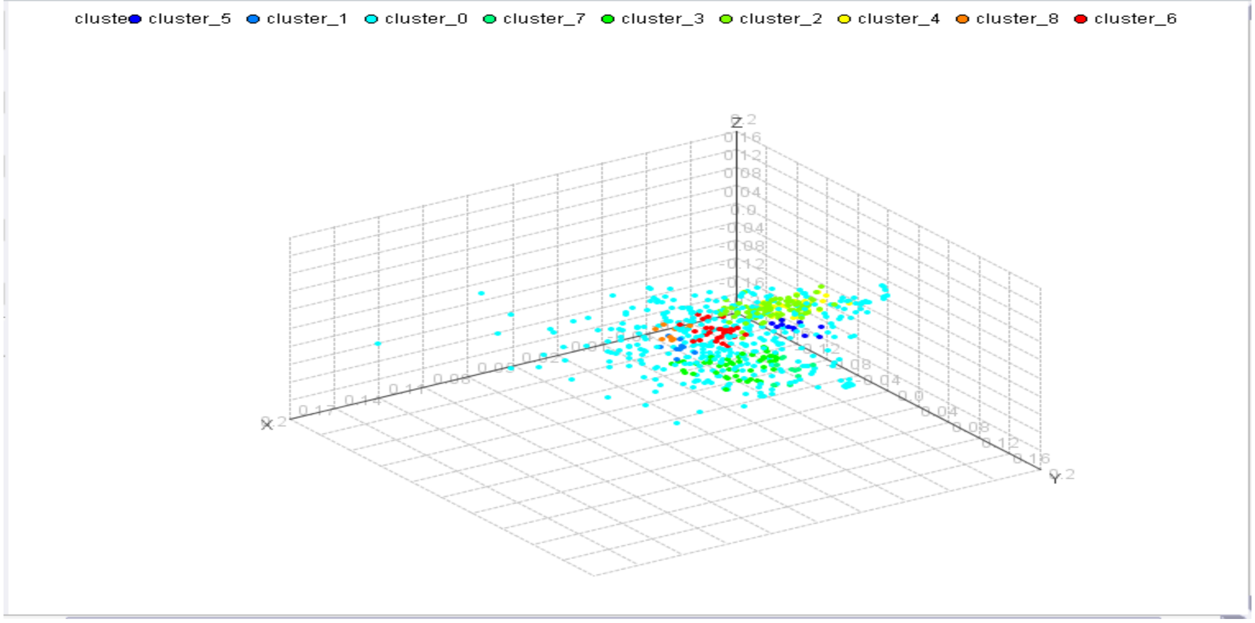
Color

cluster

DBSCAN:

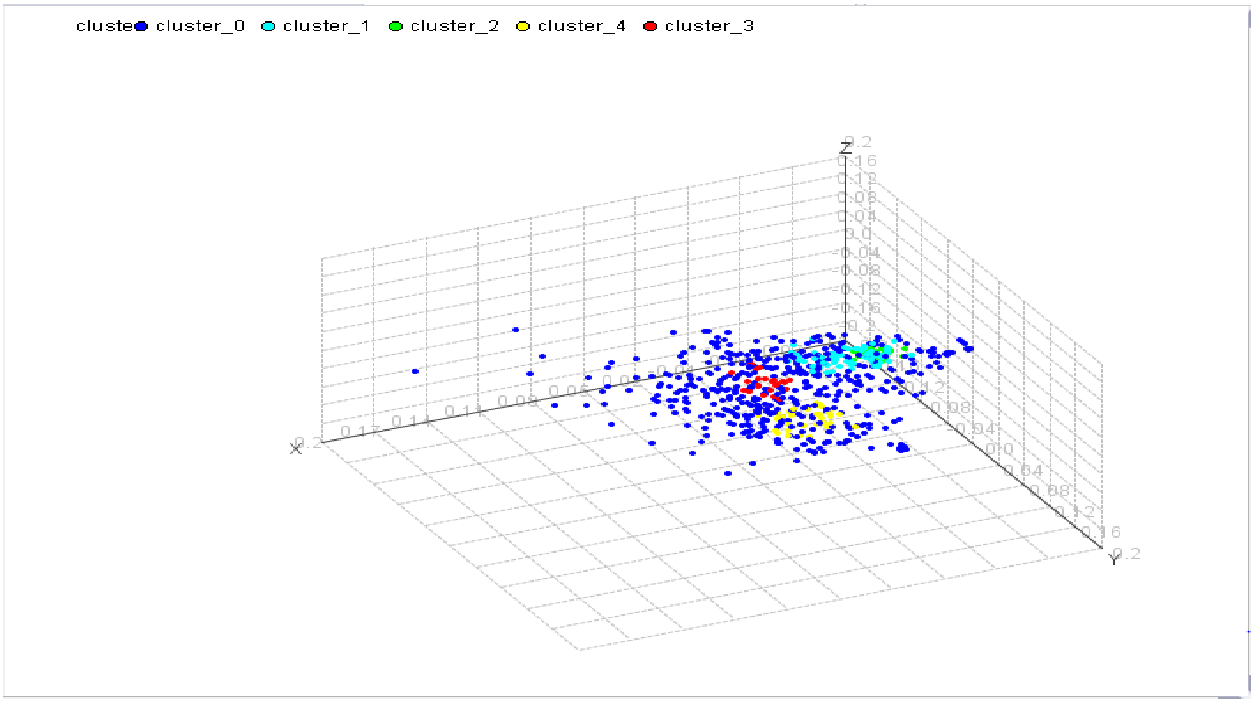
- **min points = 10**

  Figure 41: 3D Scatter chart min points = 10



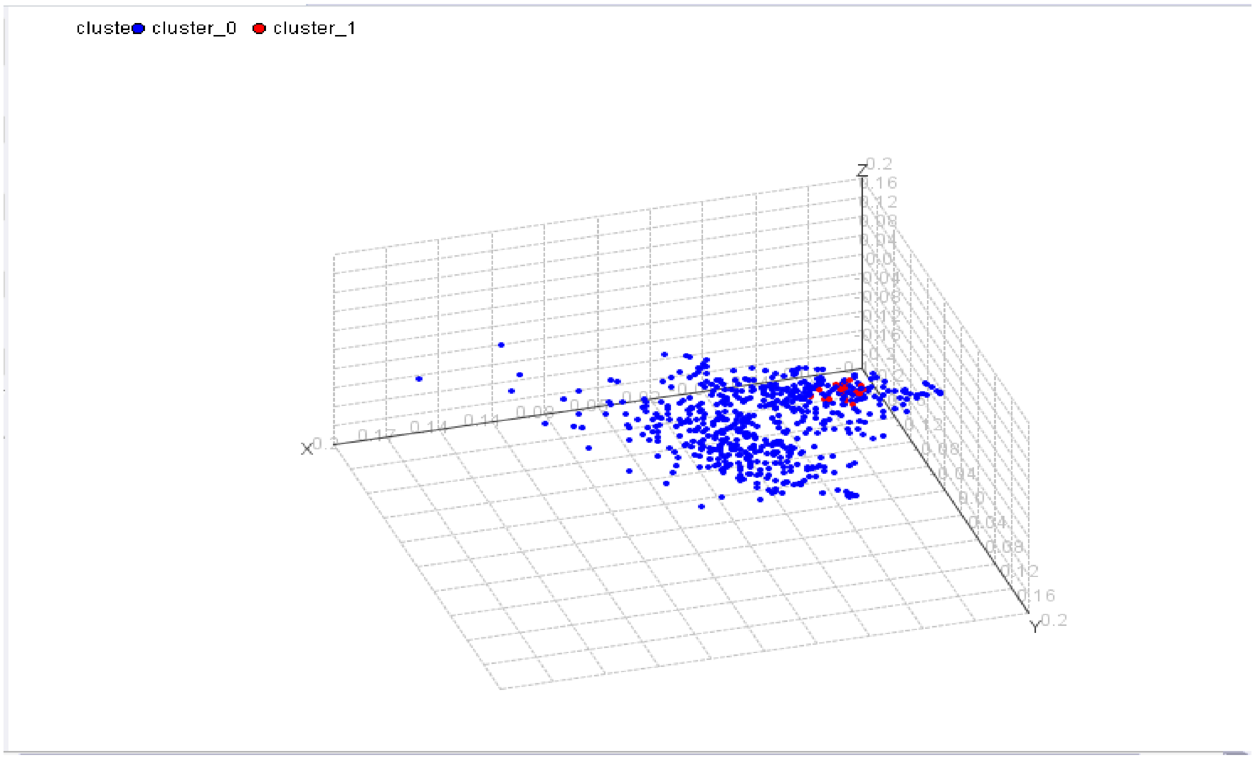
- **min points = 12**

  Figure 42: 3D Scatter chart min points = 12



- **min points = 20**

Figure 43: 3D Scatter chart min points = 20



Answer 3:

Figure 44: Decision tree

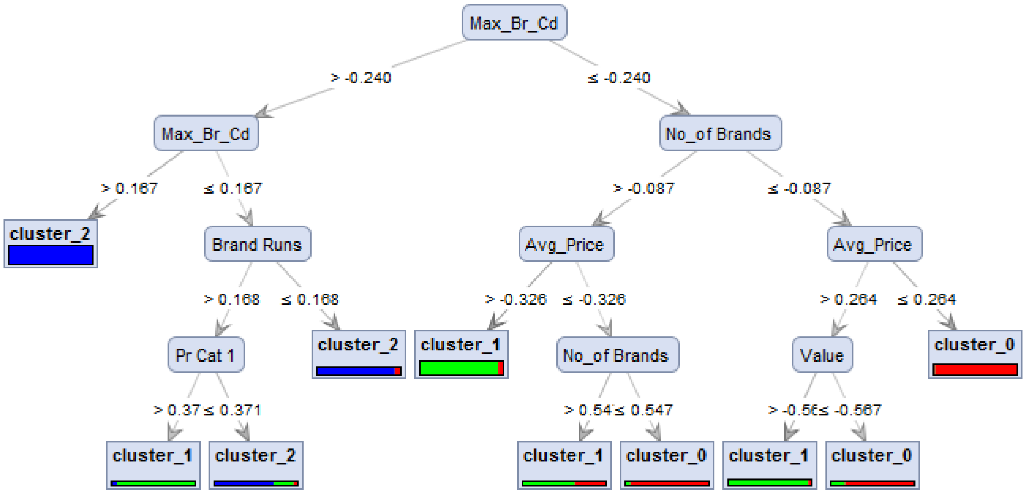
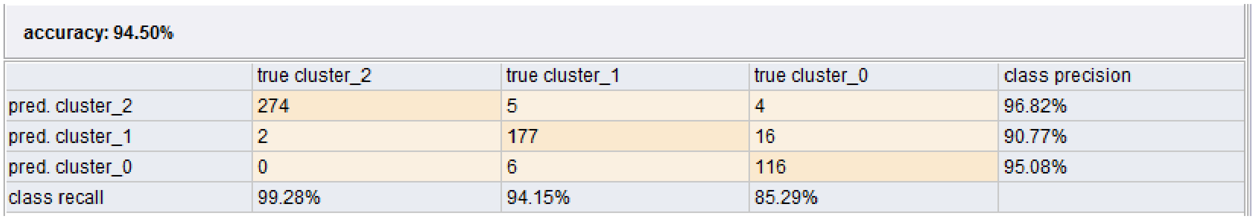


Figure 45: Performance table for Decision tree



accuracy: 94.50%

	true cluster_2	true cluster_1	true cluster_0	class precision
pred. cluster_2	274	5	4	96.82%
pred. cluster_1	2	177	16	90.77%
pred. cluster_0	0	6	116	95.08%
class recall	99.28%	94.15%	85.29%	