HASKAYNE
SCHOOL OF BUSINESS

UNIVERSITY OF CALGARY
HASKAYNE SCHOOL OF BUSINESS

Predictive Models in Business Analytics – Assignment 4

Due: Wednesday December 7, 2022 at 23:59pm

Group Number:	**01**

Last Name *(in alphabetical order)*	**First Name**
Jovanovic	Daniel
Sehgal	Rupin

QUESTION 1 – WHOLESALE CUSTOMERS

Data File: WHOLESALE CUSTOMERS.CSV

The dataset contains information regarding clients of a wholesale distributor in Portugal. It includes the annual spending in monetary units (m.u.) on diverse product categories.

a. Before we perform any clustering, let's first explore the data and get a better understanding of the distributions for the annual spending on each of the product categories. We will do so by generating the summary statistics in Excel. In the Excel file, perform the Descriptive Statistics functionality, and report the mean, standard deviation, median, mode and skewness for the annual spending on each of the product categories.

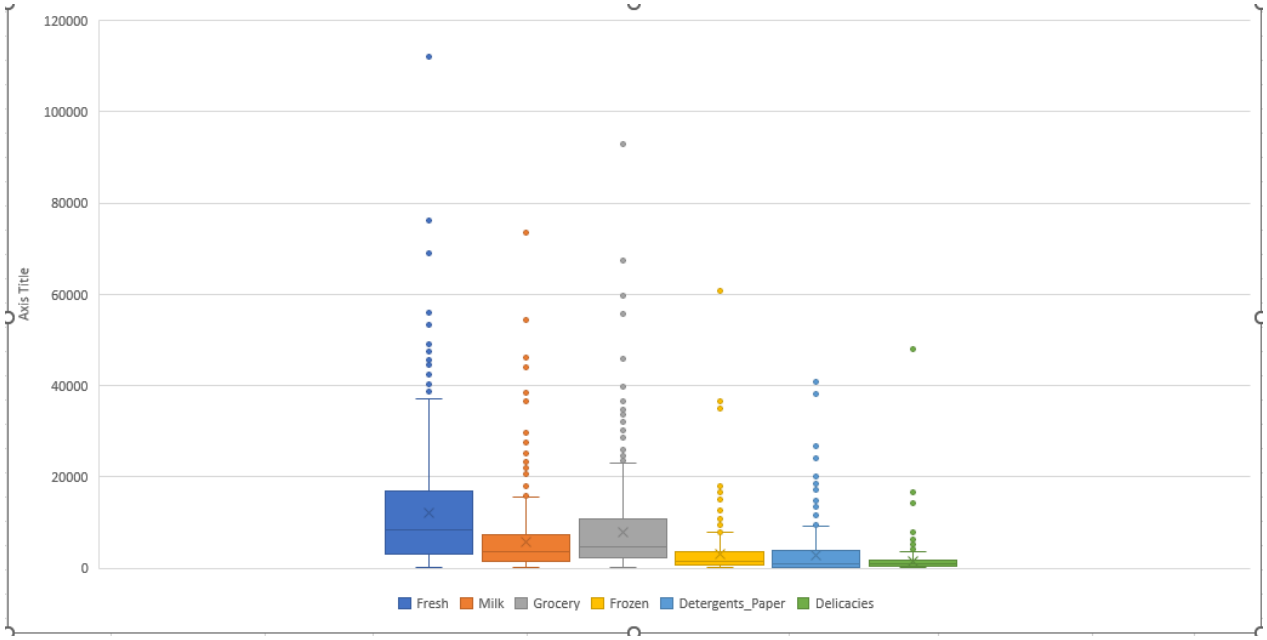
Product Category	**Mean**	**St. Dev.**	**Median**	**Mode**	**Skewness**
Fresh	12000.29773	12647.32887	8504	3366	2.561322752
Milk	5796.265909	7380.377175	3627	1196	4.053754849
Grocery	7951.277273	9503.162829	4755.5	2062	3.58742869
Frozen	3071.931818	4854.673333	1526	425	5.907985692
Detergents/Paper	2881.493182	4767.854448	816.5	918	3.631850631
Delicacies	1524.870455	2820.105937	965.5	834	11.15158648

What can you conclude about the annual spending on each of the product categories based on these statistics?

The product category fresh has the largest spending on average (mean), followed by Grocery, Milk, Frozen, Detergents/Paper, and Delicacies. Furthermore, fresh has the most variance in terms of spending as well, as it has the highest standard deviation. Not only that, but the central tendency (median) of fresh was also the highest when compared to all the other categories. Additionally, the most frequently found number in the dataset also belonged to the fresh category as it has the highest mode. Yet, it has the lowest skewness out of any category. This leads us to believe that the fresh category has the most symmetrical distribution around its central values, although it still has a positive value which indicates a slightly longer right tail. In terms of standard deviation, the category Fresh is followed by Grocery , Milk, Frozen, Detergents/Paper and Delicacies, from largest to smallest. Similarly, Grocery also displayed the second largest median, followed by Milk, Frozen, Delicacies, and Detergents/Paper. Grocery once again displayed the second largest mode, with the third largest again being Milk, followed by Detergents/Paper, Delicacies, and Frozen. Although all of the product categories showed a positive skew, the highest skew was seen within the Delicacies, suggesting that it has the most asymmetrical distribution around its probability distribution. This was followed by Frozen, Milk, Detergents/Paper, Grocery and Fresh, as previously mentioned.

b. Besides the summary statistics, we can also use box-and-whisker plots to compare distributions. In the accompanying Excel file, create a chart with the six box-and-whisker plots regarding the distribution of the annual spending (one for each product category).

Screenshots of the box-and-whisker plots (make sure they are readable after printing):



What can you conclude/observe about the distribution of the annual spending on each of the product categories by looking at the box-and-whisker plots?

The product category Fresh seems to have most expansive interquartile range as indicated by its larger box size. It also has a higher median value, as shown by the line in the centre of the interquartile range box, and it also has the largest mean which can be seen by the "x" inside the box as well. Furthermore, this graph confirms that Fresh has the highest standard deviation which is indicated by the wide range between the whiskers. However, it would appear that Fresh is also a product category that shows some of the most frequent and severe outliers. Overall, from the box-and-whisker plots we can confirm that Frozen is the product category where the majority of annual spending is done. Additionally, the product category with the smallest interquartile range is Delicacies, while the category with the least extreme outliers seems to be Detergents/Paper. However, all of the product categories shown display a relatively large quantity of outliers.

c. Finally, we want to calculate the correlation matrix to analyze any (linear) relationships between the attributes. In the accompanying Excel file, create the correlation matrix of the annual spending between the different product categories.

Screenshot of the correlation matrix (make sure it is readable after printing):

	Fresh	*Milk*	*Grocery*	*Frozen*	*Detergents_Paper*	*Delicacies*	*Channel*	*Region*
Fresh	1							
Milk	0.100509768	1						
Grocery	-0.011853875	0.728335118	1					
Frozen	0.345881457	0.123993759	-0.040192737	1				
Detergents_Paper	-0.101952938	0.661815679	0.924640691	-0.131524906	1			
Delicacies	0.244689969	0.406368316	0.205496511	0.390947465	0.069291297	1		

What can you conclude/observe about the correlation (or linear relationships) of the annual spending between the different product categories?

From the given correlation matrix, we can conclude that the annual spending between Grocery and Detergents/Paper are the two most positively correlated product categories, with a correlation value of 0.9246. This is followed by Milk and Grocery (0.7283) along with Milk and Detergents/Paper (0.6618), which are the second and third most positively correlated product categories, respectively. As for the largest negative correlations, the most negatively correlated product categories are Frozen and Detergents/Paper with a value of -0.1315, followed by Fresh and Detergents/Paper (-0.1019) along with Grocery and Frozen (-0.040). However, none of these top three values suggest a necessarily strong negative correlation. The remaining pairs of product categories show relatively weak correlations as well.

d. Now that we have a better understanding of the data, we can start to cluster the customer profiles based on their annual spending in the different product categories (i.e., do NOT include

the sales channel or the customer location/region for the clustering task). For this question, use k-means clustering where you group the data in 5 clusters. To do so, use 100 runs of at most 250 iterations in each run. Use the cosine similarity as the numerical measure (since all attributes have numerical values). Don't forget to include the local random seed value of 1992 to initialize the random locations of the centroids for each run. Since all attributes are the annual spending in a certain product category, there is no need to normalize the data at this point (we will explore this later in the assignment).

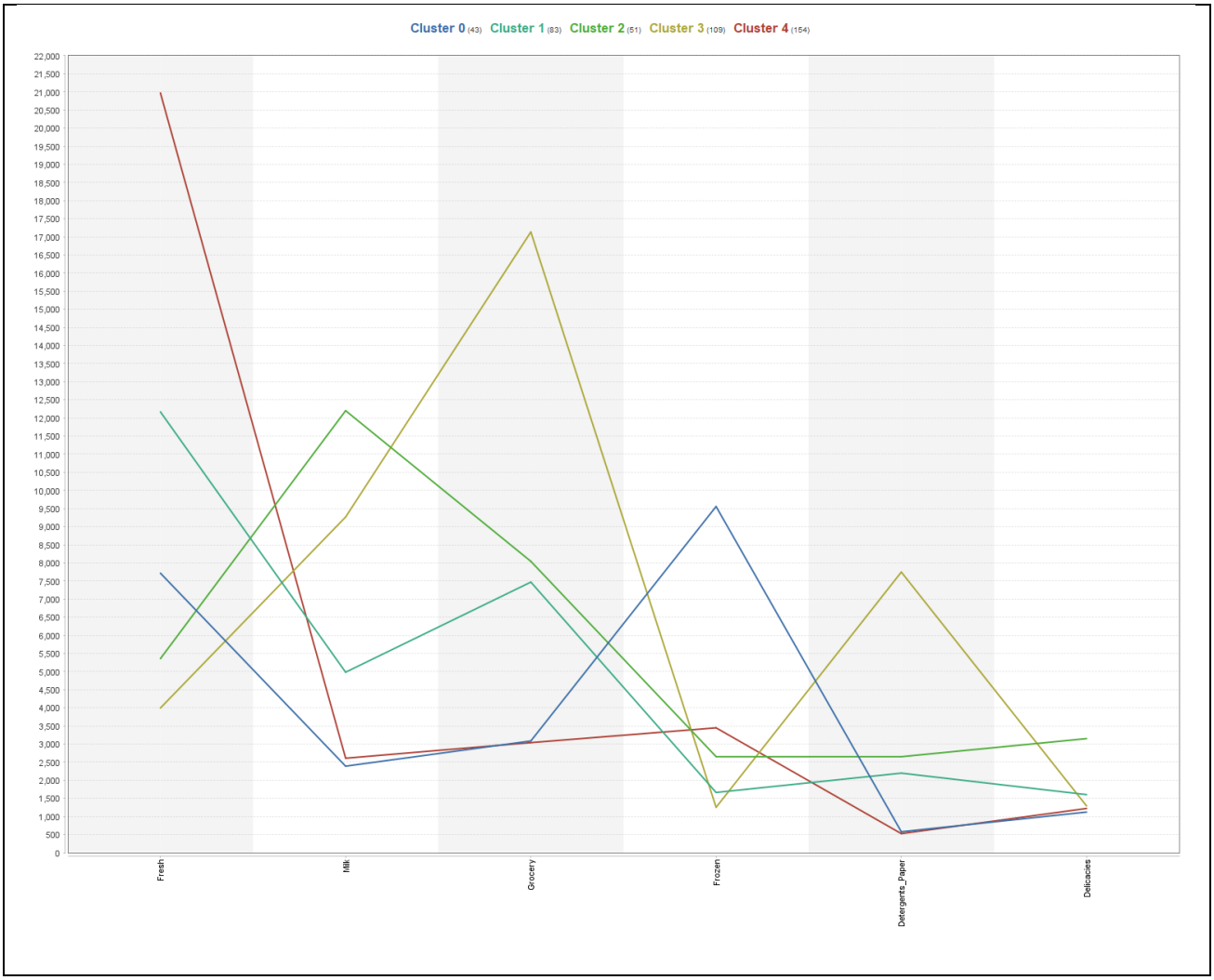
Save the process to create the clusters as "A4_Q1d_e.rmp" in the process folder of your Local Repository. For each cluster, report the number of customers (or records) in the cluster as well as the average distance to the centroid within the cluster. Also report the overall (or weighted) average distance to the centroids and the sum of the squares for the item distribution measure.

	Number of customers	**Avg. within centroid distance**
Cluster 1	43	-156259602.929
Cluster 2	83	-86941436.729
Cluster 3	51	-455742127.916
Cluster 4	109	-294772685.391
Cluster 5	154	-260785360.625

Avg. within centroid distance	-248793906.687
Sum of squares	0.242

e. In this question, we want to get a better understanding of the customer profiles for each of the five clusters that you generated in the previous question. Update the file of Question 1d to include the *Cluster Model Visualizer operator*. Include a copy of the Centroid Chart. Use the Overview and include screenshots of the scatter plots for each of the clusters to describe the customer profiles in the clusters (i.e., clearly describe the spending behavior of customers in each cluster and support your argumentation with the output from RapidMiner, where you include screenshots).

Screenshot of the Centroid Chart (make sure it is readable after printing):



Screenshot of the Overview (make sure it is readable after printing):

Number of Clusters: 5

Cluster 0 43

Frozen is on average **212.83%** larger, **Detergents_Paper** is on average **80.01%** smaller, **Grocery** is on average **61.22%** smaller

Cluster 1 83

Frozen is on average **46.23%** smaller, **Detergents_Paper** is on average **23.80%** smaller, **Milk** is on average **14.17%** smaller

Cluster 2 51

Milk is on average **111.57%** larger, **Delicacies** is on average **106.77%** larger, **Fresh** is on average **55.38%** smaller

Cluster 3 109

Detergents_Paper is on average **169.02%** larger, **Grocery** is on average **115.54%** larger, **Fresh** is on average **66.74%** smaller

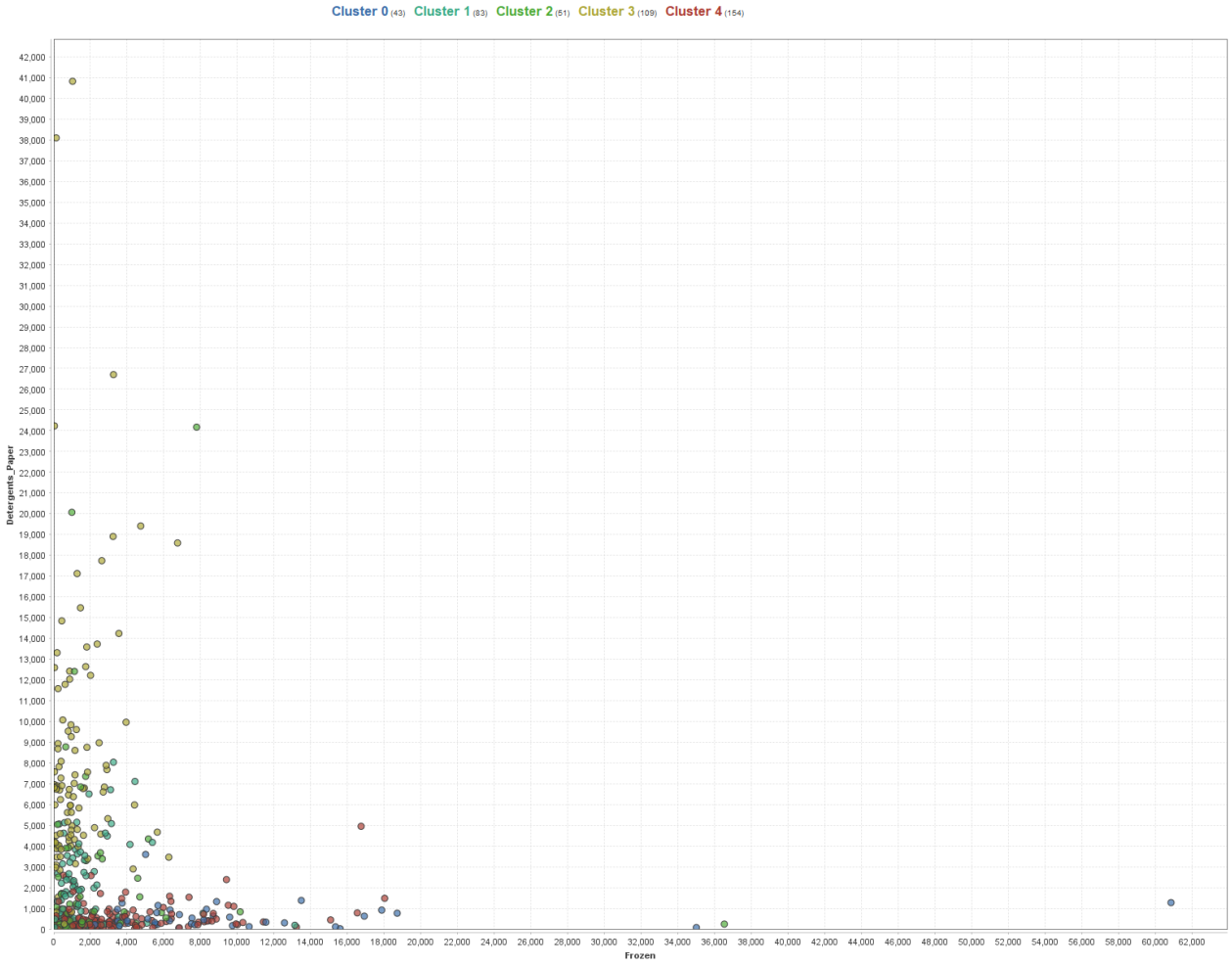
Cluster 4 154

Detergents_Paper is on average **81.81%** smaller, **Fresh** is on average **74.78%** larger, **Grocery** is on average **61.80%** smaller

Description of cluster 1:

Within Cluster 1, if we ignore extreme outliers, we can see that customers seem to spend similar amounts on Detergents/Paper and Frozen products. However, based on the cluster's behaviour, it would appear that customers who spend higher amounts on Detergents/Paper tend to spend lower amounts on Frozen products, and vice versa. More specifically, customers who spend approximately $8,000 per year or more on Detergents/Paper tend to spend less than $2,000 per year on Frozen product. This pattern is most apparent in "cluster 3". Likewise, individuals who spent more than roughly $6,000 on Frozen products seems to spend less than approximately $2,000 on Detergents/Paper. This trend appears to be most relevant for "cluster 0" and "cluster 4".

Screenshot of scatterplot to support your motivation (make sure it is readable after printing):

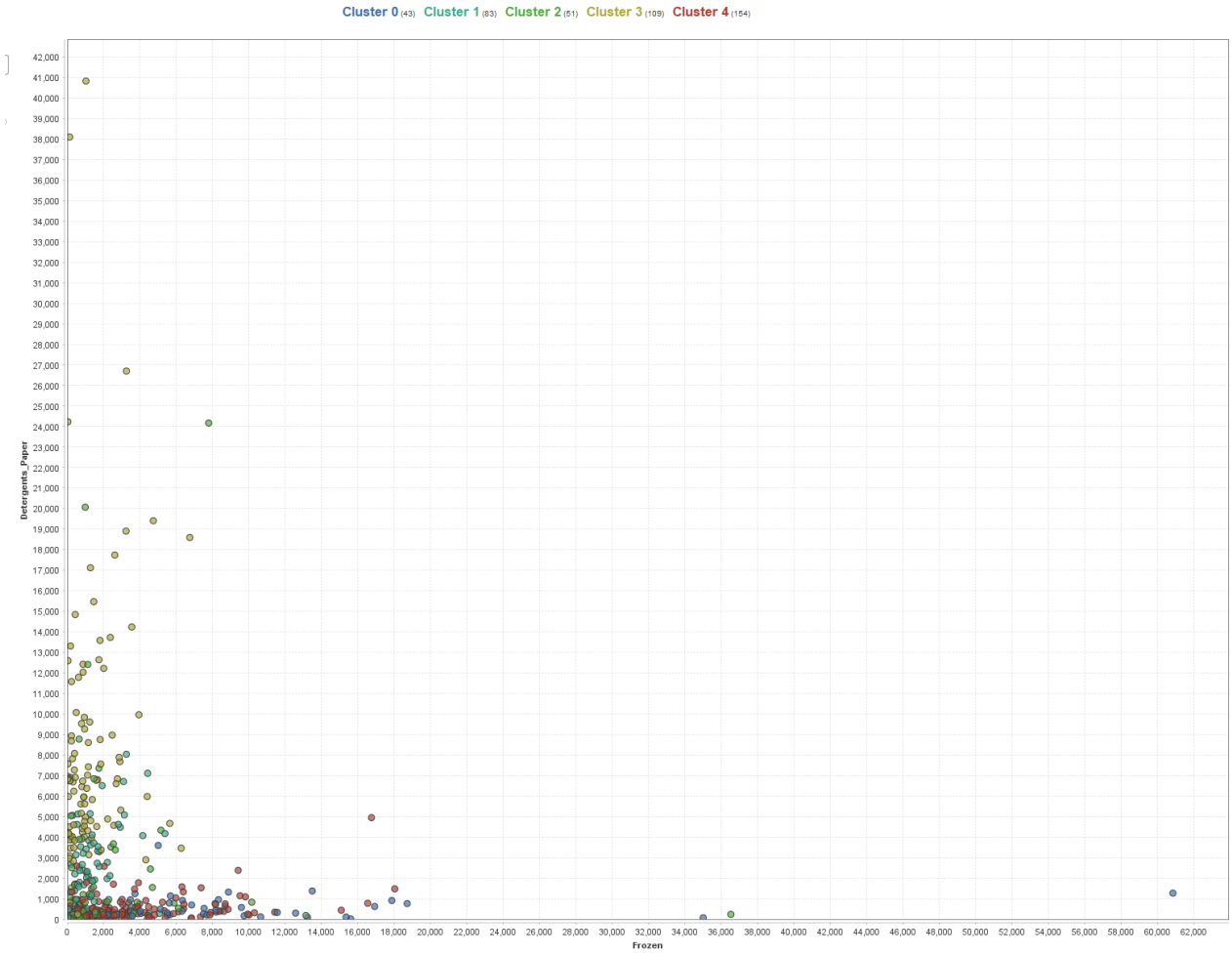


Description of cluster 2:

When observing cluster 2, we noticed that it is essentially identical to cluster 1. As a result, we can draw very similar conclusions on this cluster as we did with the prior. The same

pattern is evident, which shows that spending on Detergents/Paper and Frozen products is similar, and the more people spend on Detergents/Paper the less they tend to spend on Frozen products, especially when their spending on Detergents/Paper is relatively large. This same trend can be seen the other way around as well.

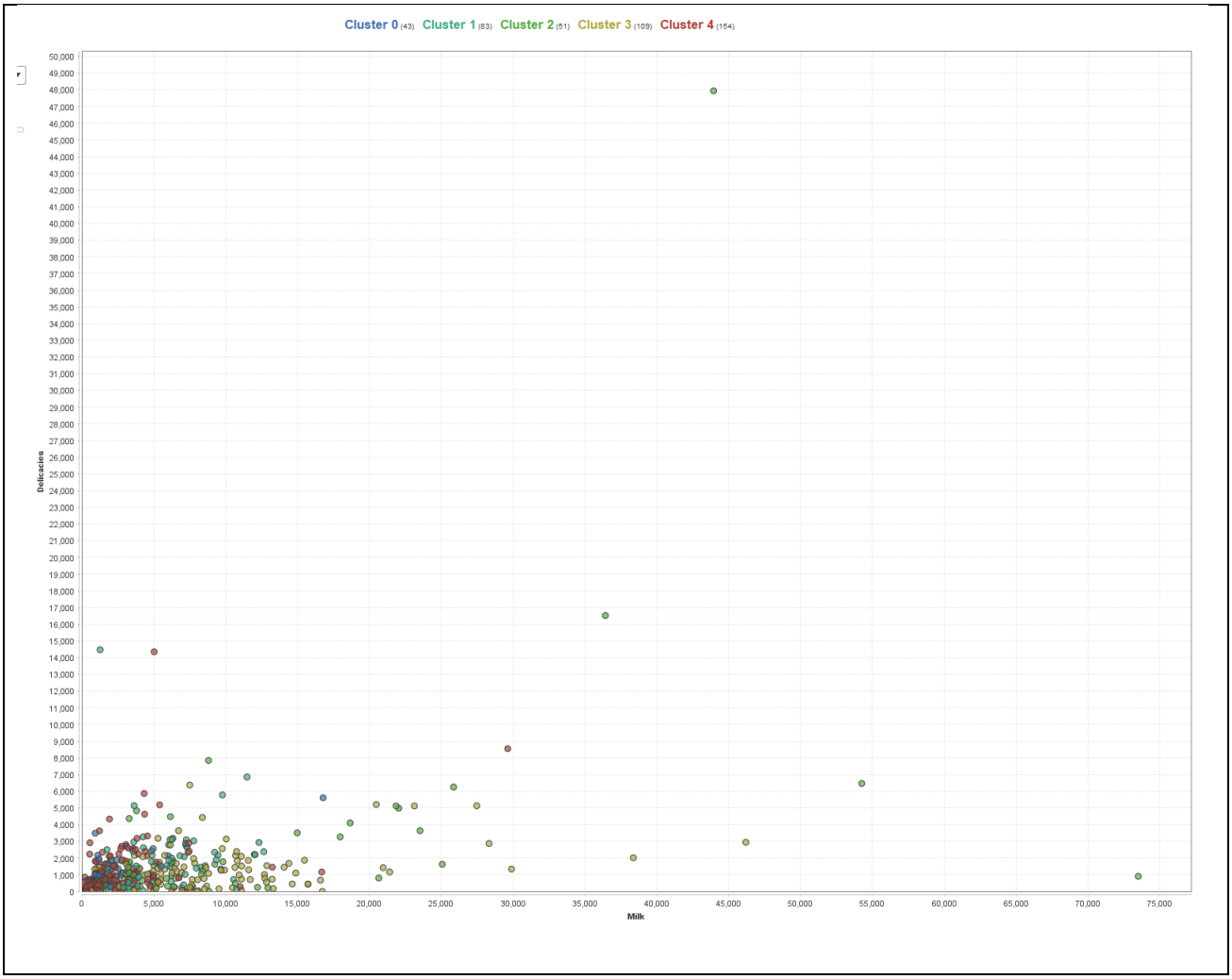
Screenshot of scatterplot to support your motivation (make sure it is readable after printing):



Description of cluster 3:

For cluster 3, we can see a wider range of items across the x-axis. This behaviour suggests that individuals seem to spend more money on Milk compared Delicacies. Furthermore, those who spend a relatively large amount on Milk will tend to spend a much lower amount on Delicacies. In other words, individuals who spend a lot on Delicacies will still tend to spend quite a bit on Milk as well, but this same phenomenon is not very prevalent the other way around. This is particularly true for those within “cluster 3”.

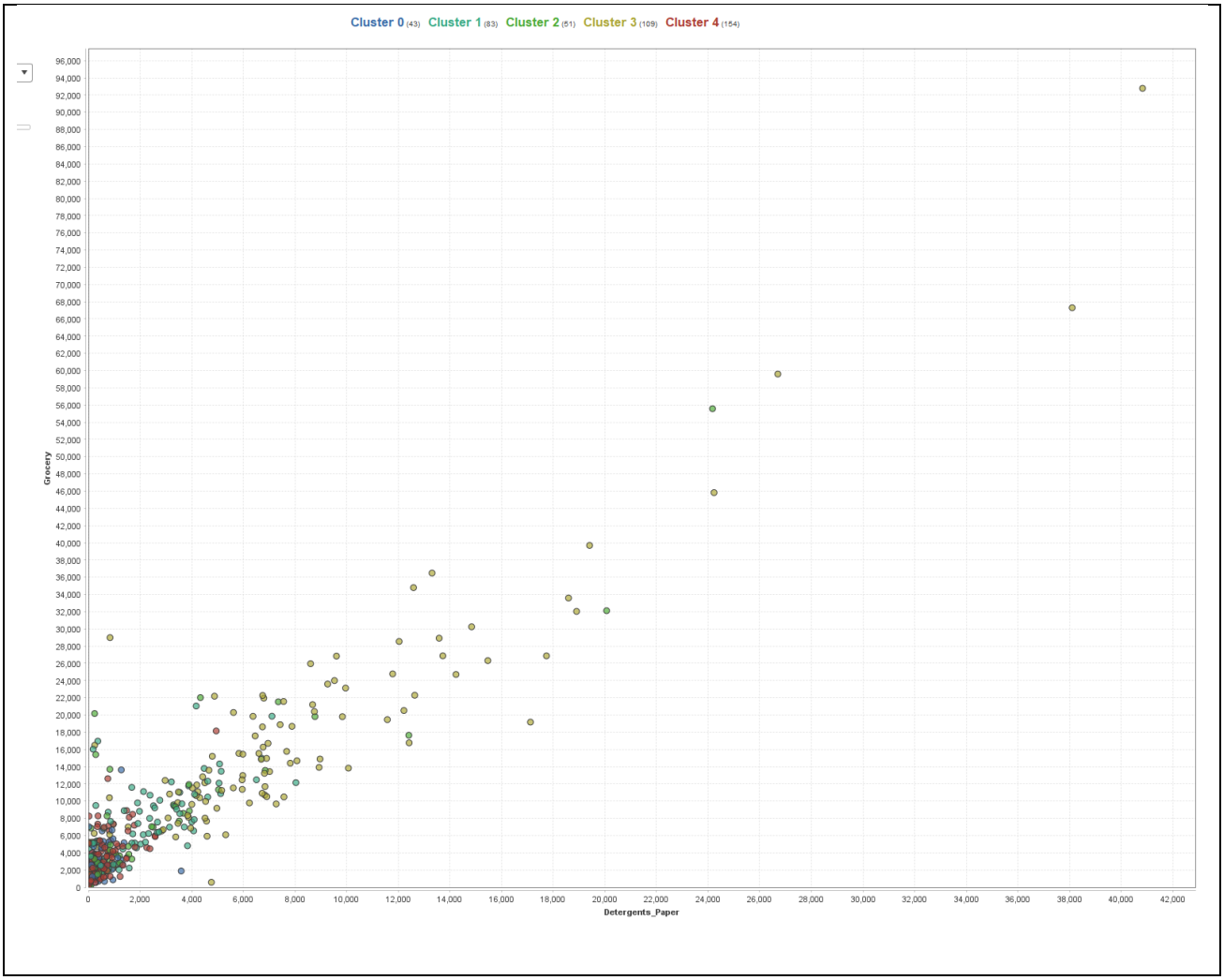
Screenshot of scatterplot to support your motivation (make sure it is readable after printing):



Description of cluster 4:

The fourth cluster shows a strong positive relationship between spending on Groceries and Detergents/Paper. This scatterplot indicates that those who spend a large amount on Groceries will also tend to spend a lot on Detergents/Paper. Similarly, those who spend a relatively small amount on Groceries will also tend to also spend lower amounts on Detergents/Paper as well. This scatterplot seems to confirm our conclusions in the correlation matrix above: the product categories Grocery and Detergents/Paper have a very high positive correlation. However, those within “cluster 0”, “cluster 2”, “cluster 4” and to a lesser extend “cluster 1”, don’t seem to spend a large amount on either Grocery products or Detergent/Paper.

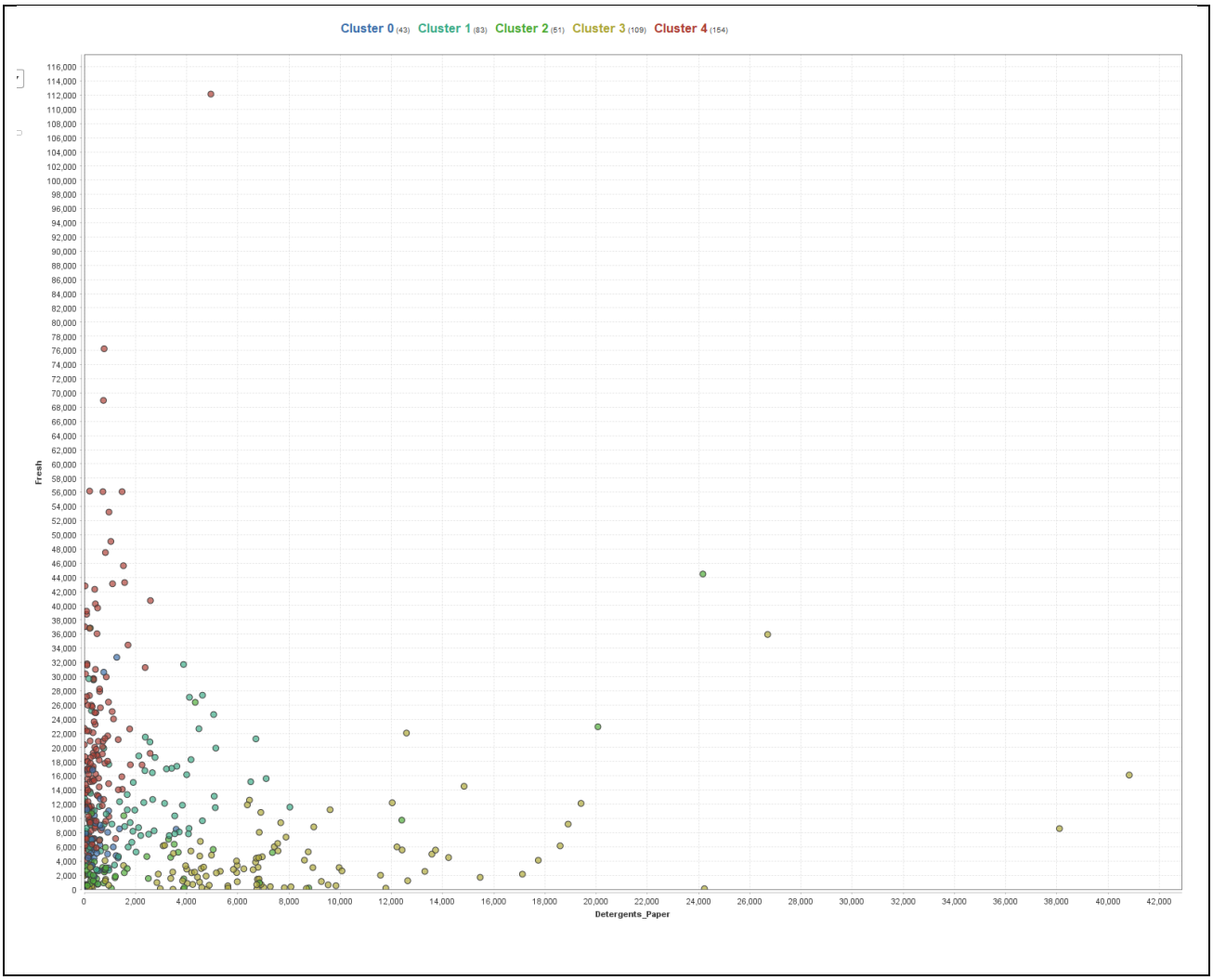
Screenshot of scatterplot to support your motivation (make sure it is readable after printing):



Description of cluster 5:

With regards to cluster 5, it appears as though individuals spend more on Fresh products when compared to Detergents/Paper. Furthermore, customers who spend higher amounts on Fresh products will tend to spend lower amounts on Detergents/Paper, which is especially apparent when observing the red dots which represent “cluster 4”. Although it can also be concluded that those who will spend a lot on Detergents/Paper will tend to spend less on Fresh products, this pattern is not as strong as the one previously mentioned. In fact, this pattern is predominantly only driven by “cluster 3”.

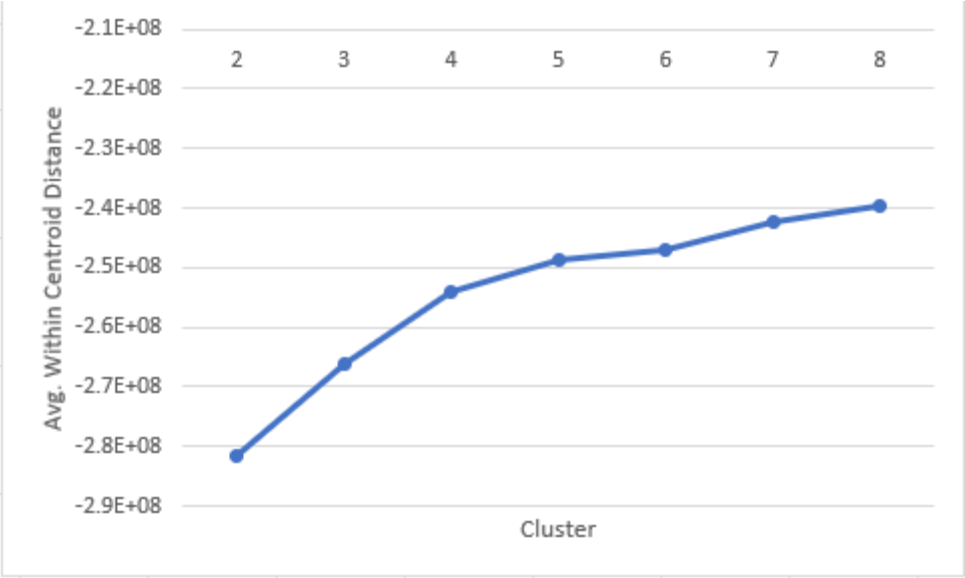
Screenshot of scatterplot to support your motivation (make sure it is readable after printing):

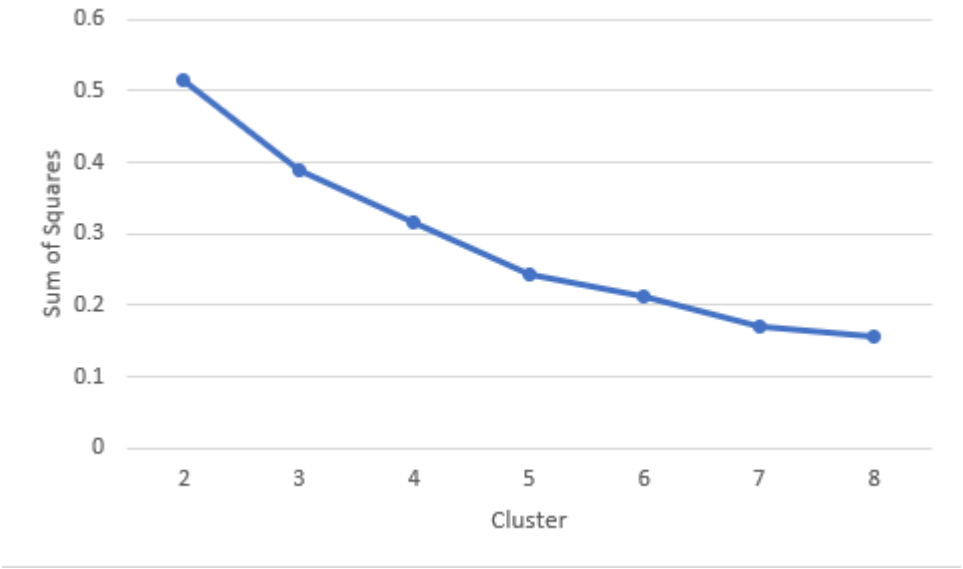


f. So far, we have created 5 clusters. In this question, set the number of k-means clusters from 2 until 8 (i.e., 2, 3, …, 8) and report the overall (or weighted) average distance to the centroids and the sum of the squares. Create two visualizations (in Excel) where these measures are on the y-axis, respectively, and the number of clusters is on the x-axis. Also report the number of customers (or records) in each cluster.

Number of clusters (k)	Number of customers in each cluster {n1; n2; n3; …}	Avg. within centroid distance	Sum of squares
2	257; 183	-281738167.113	0.514
3	204; 65; 171	-266071332.209	0.388
4	115; 197; 83; 45	-254048025.328	0.315
5	43; 83; 51; 109; 154	-248793906.687	0.242
6	57; 136; 106; 73; 46; 22	-247074530.687	0.211
7	55; 115; 84; 54; 22; 67; 43	-242187084.304	0.171
8	67; 55; 34; 114; 71; 51; 22; 26	-239577770.563	0.157

Screenshots of the two visualizations (make sure they are readable after printing):





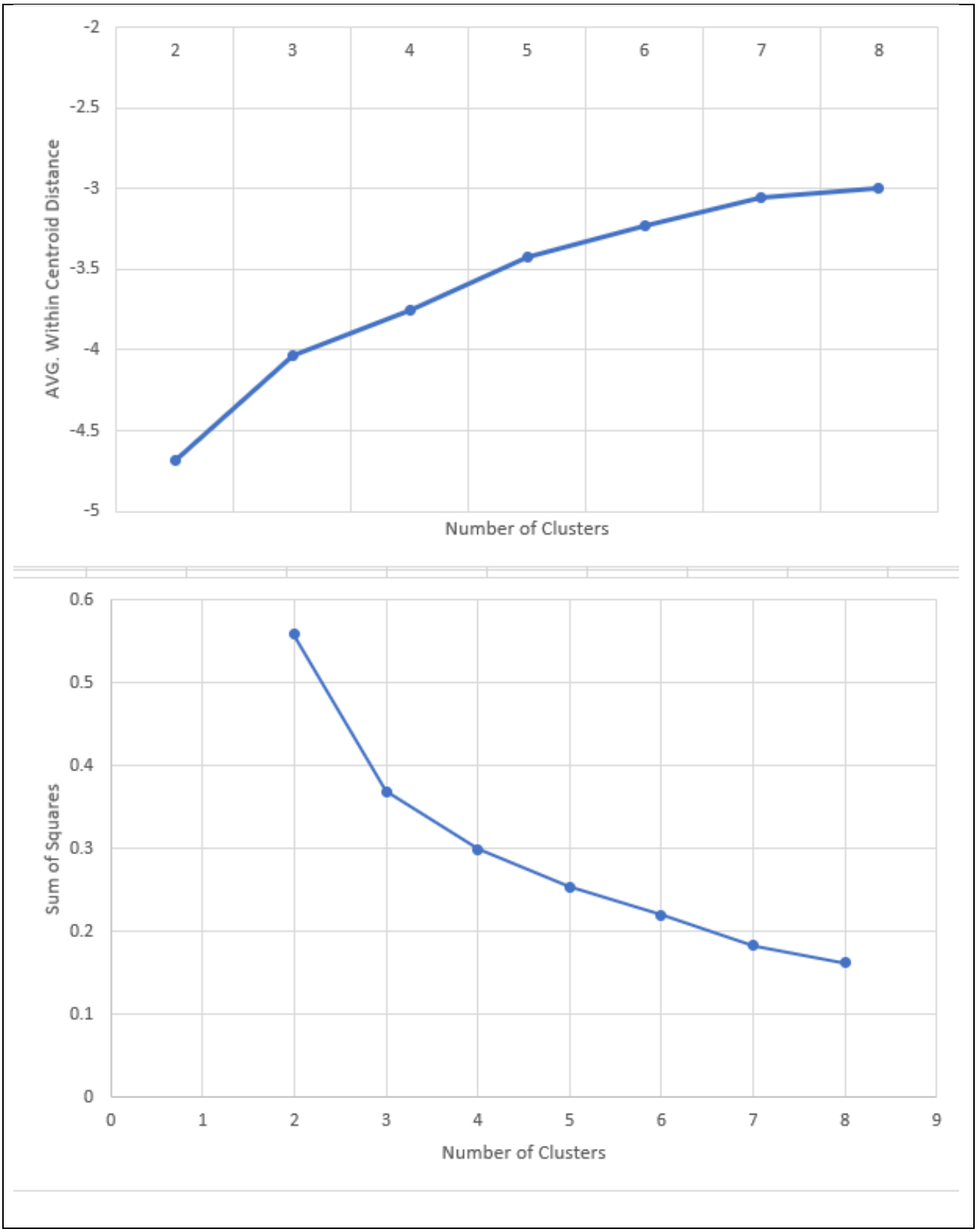
What should be the preferred number of clusters based on the elbow technique? Motivate your answer.

For the graph with "Average within centroid distance" on the y-axis, it would appear as though the preferred number of clusters would be 5, because after this the y-value seems to increase at a much smaller rate. Similarly, when observing the graph that has "Sum of Squares" on the y-axis, the preferred number of clusters is also 5 because the rate of decrease from cluster-to-cluster reduces to a relatively large degree past this point.

g. Save the file of Question 1d/e as "A4_Q1g.rmp" in the process folder of your Local Repository. Normalize the data first (with a z-transformation), before you cluster the data (use the same parameter values for the clustering process). Redo the previous question (only Question 1f) and determine the preferred number of clusters now that the data is normalized. Motivate your answer. Also report the number of customers (or records) in each cluster.

Number of clusters (k)	**Number of customers in each cluster {n1; n2; n3; …}**	**Avg. within centroid distance**	**Sum of squares**
2	295; 145	-4.687	0.558
3	103; 124; 213	-4.032	0.369
4	53; 116; 86; 185	-3.752	0.299
5	84; 49; 56; 176; 75	-3.428	0.254
6	24; 48; 69; 77; 64; 158	-3.227	0.220
7	137; 68; 65; 47; 43; 58; 22	-3.056	0.183
8	123; 69; 41; 30; 60; 20; 37; 60	-2.994	0.162

Screenshots of the two visualizations (make sure they are readable after printing):



What should be the preferred number of clusters based on the elbow technique? Motivate your answer.

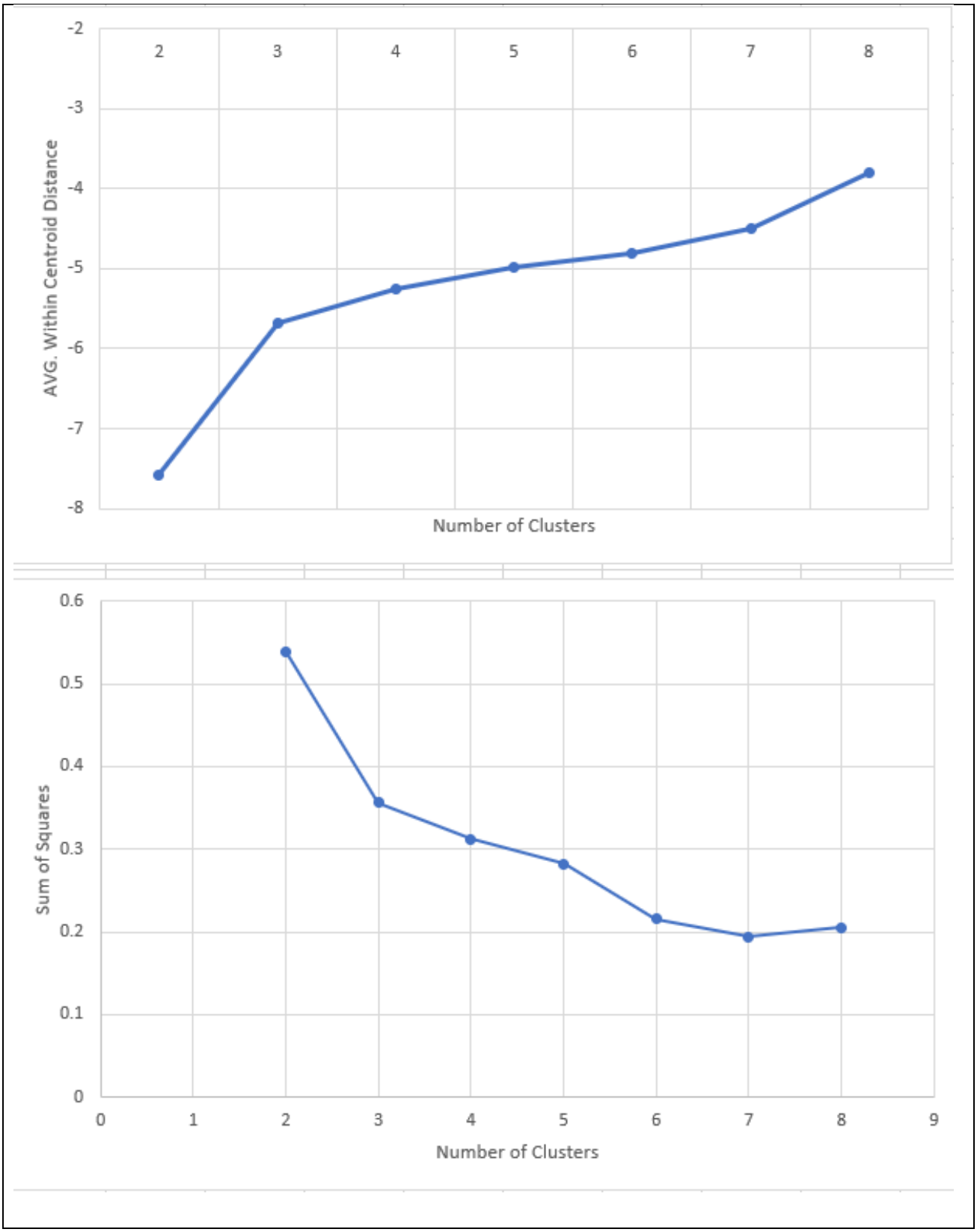
According to the elbow technique, it appears as though the preferred number of clusters that should be chosen is three, based off our observations of both graphs. As we have learned in class, the elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of picking the elbow of the curve as the number of clusters to use. According to the graphs above, the clear choice is three as it has the sharpest change out of any of the other points, and the rate of decrease (sum of squares) slows down drastically after 3 clusters.

For all remaining questions, we keep the normalization before the clustering.

h. Save the file of Question 1g as "A4_Q1h.rmp" in the process folder of your Local Repository. Update the process to include k-Medoids clustering instead of k-Means clustering. Use the same parameter values for the clustering process as in the previous questions. Redo the previous question and determine the preferred number of clusters now that a different clustering technique is performed. Motivate your answer. Also report the number of customers (or records) in each cluster.

Number of clusters (k)	**Number of customers in each cluster {n1; n2; n3; …}**	**Avg. within centroid distance**	**Sum of squares**
2	281; 159	-7.587	0.538
3	116; 200; 124	-5.677	0.356
4	41; 96; 109; 194	-5.258	0.312
5	17; 87; 181; 46; 109	-4.983	0.282
6	46; 76; 17; 108; 55; 138	-4.800	0.216
7	76;23;53;46;17;131;94	-4.491	0.194
8	9;46;70;53;17;159;28;58	-3.802	0.205

Screenshots of the two visualizations (make sure they are readable after printing):



What should be the preferred number of clusters based on the elbow technique? Motivate your answer.

Applying the elbow technique to both of our graphs makes it clear that the preferred number of clusters that should be chosen is three. The prudent choice is three clusters as it has the sharpest change out of any of the other points, and the rate of decrease (sum of squares), along with the rate of increase (Avg. within centroid distance) slows down drastically after 3 clusters.

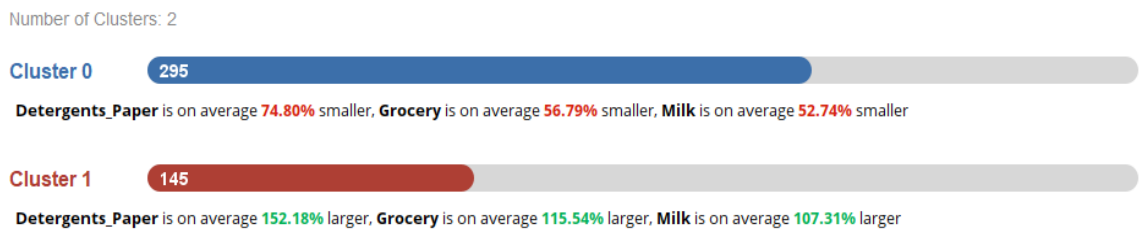
i. Save the file of Question 1g as "A4_Q1i.rmp" in the process folder of your Local Repository. Update the process to include the *x-Means* operator. Use the same parameter values for the clustering process as in the previous questions, and let the number of clusters with the regular k-Means clustering technique (not fast k-Means) vary between 2 until 8.

How many clusters are selected?

The number of clusters selected is 2.

Based on the results of the *Cluster Model Visualizer* operator, use the Overview and include screenshots of the scatter plots for each of the clusters to describe each customer profile in the clusters (i.e., clearly describe the spending behavior of customers in each cluster and support your argumentation with the output from RapidMiner, where you include screenshots).

Screenshot of the Overview (make sure it is readable after printing):

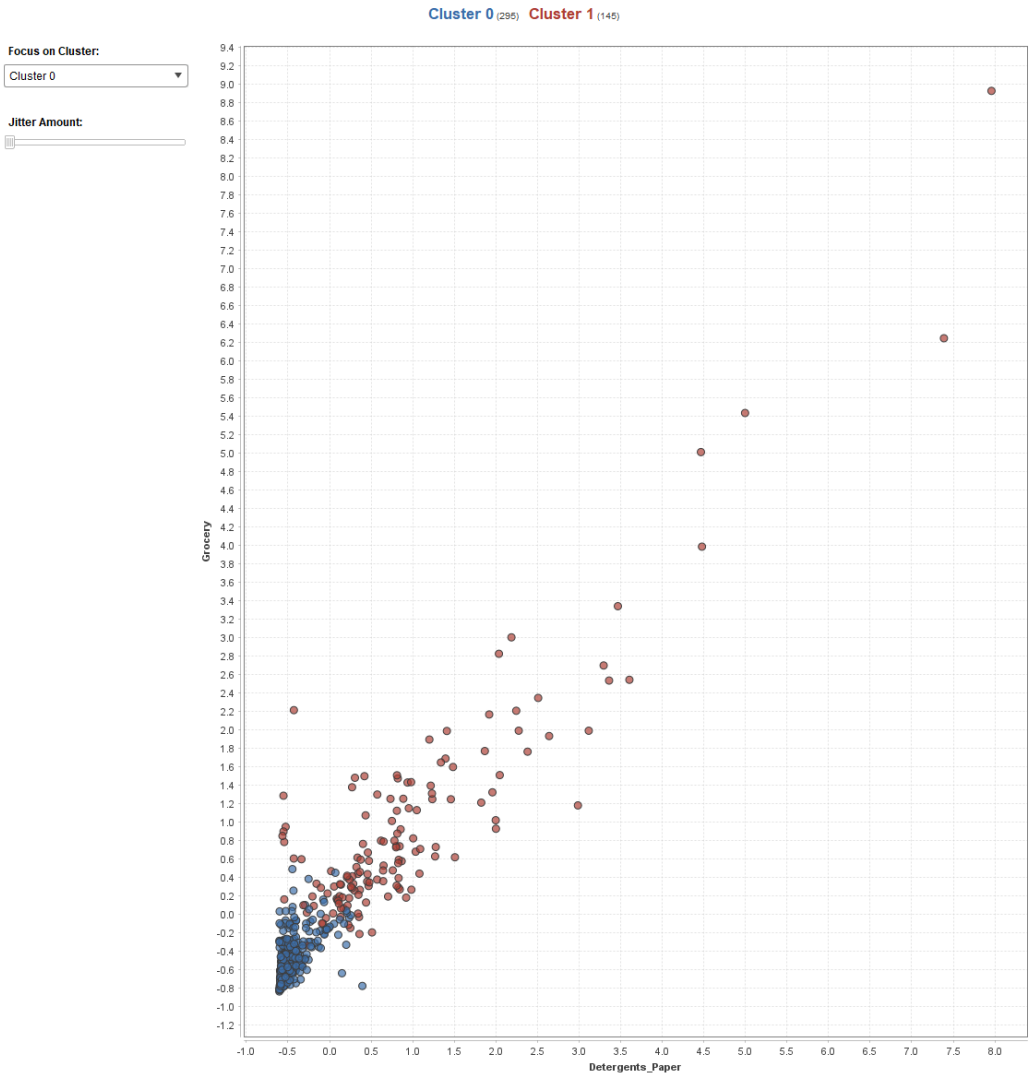


(Repeat this textbox for each cluster)

Description of cluster 0:

When focusing on cluster 0, the spend between the categories seems to show a positive relationship. "Cluster 0" is highly concentrated and has little variation, whereas "cluster 1" has drastically higher levels of spend, with high levels of variability. Spenders in both clusters spend relatively the same amount on both detergent/paper and groceries as seen by the linear distribution of spenders in the scatter plot. Furthermore, "cluster 0" spends less on both categories, and "cluster 1" spenders spend far more on both.

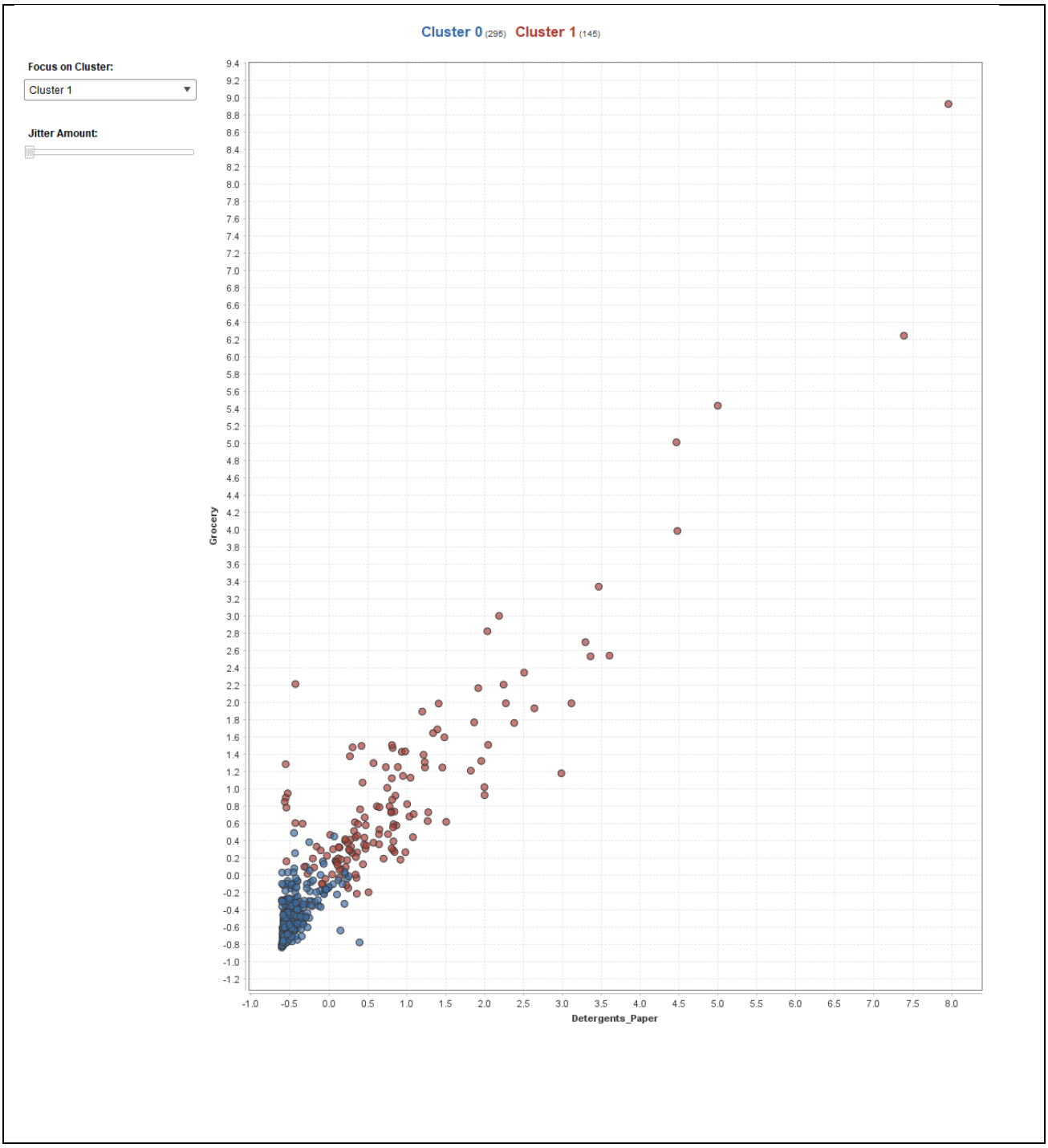
Screenshot of scatterplot to support your motivation (make sure it is readable after printing):



Description of cluster 1:

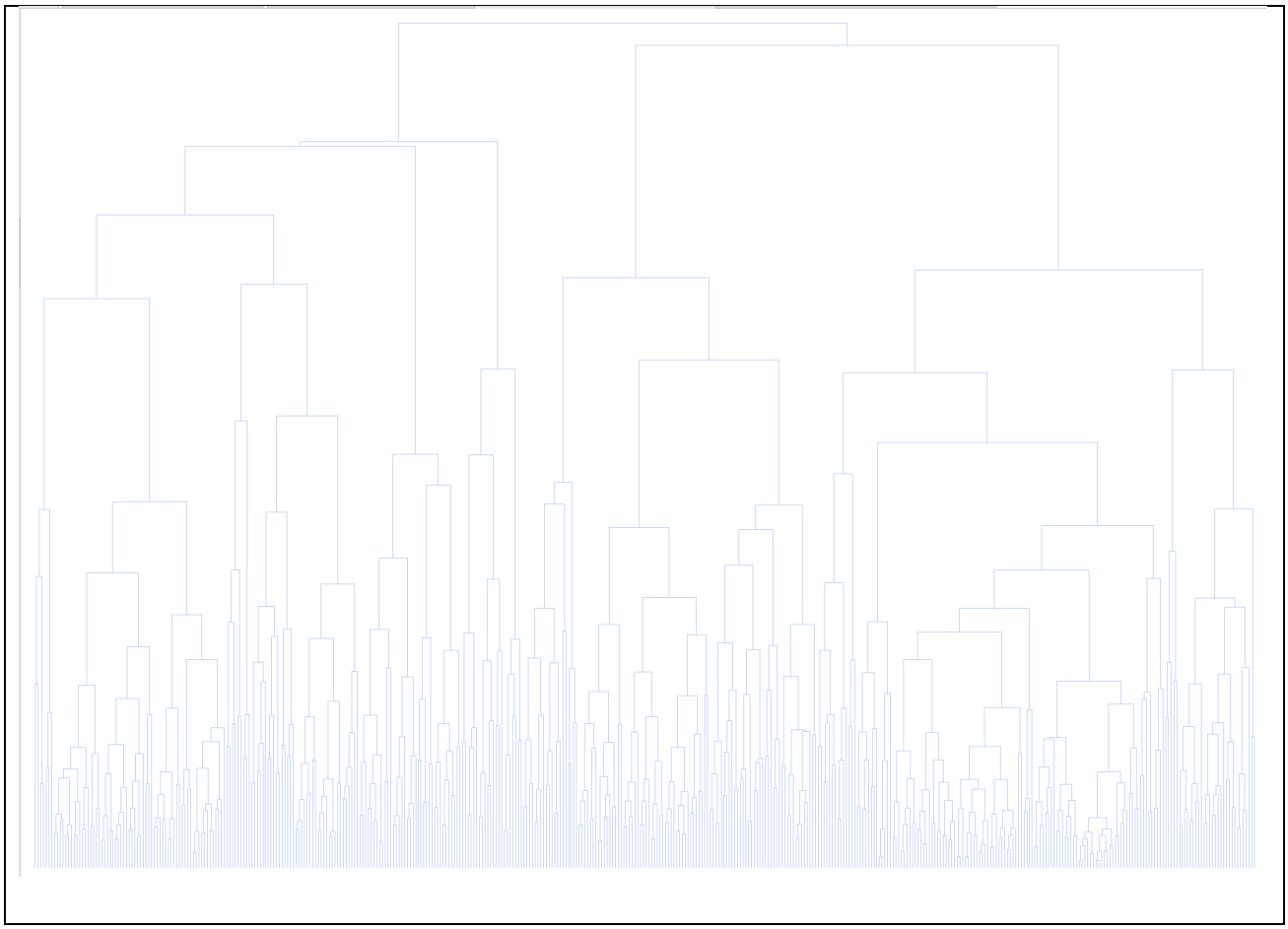
When focusing on cluster 1, the spend between the categories seems to show a positive relationship. “Cluster 0” is has a high level of spenders and has little variation, whereas “cluster 1” has increasing levels of spend, with high levels of variability. Spenders in both clusters spend similar amounts on Detergent/Paper and Groceries as seen by the linear increase of spenders in the scatter plot. Furthermore, “cluster 0” spends less on both categories, and “cluster 1” spenders spend far more on both.

Screenshot of scatterplot to support your motivation (make sure it is readable after printing):



j. Save the file of Question 1g as "A4.Q1j.rmp" in the process folder of your Local Repository. Update the process to include the bottom-up (or agglomerative) clustering as hierarchical clustering technique to create a dendrogram, and use the *Flatten Clustering* operator to create actually clusters. Use the same distance measures as in the previous questions, and use CompleteLink as mode in the *Agglomerative Clustering* operator.

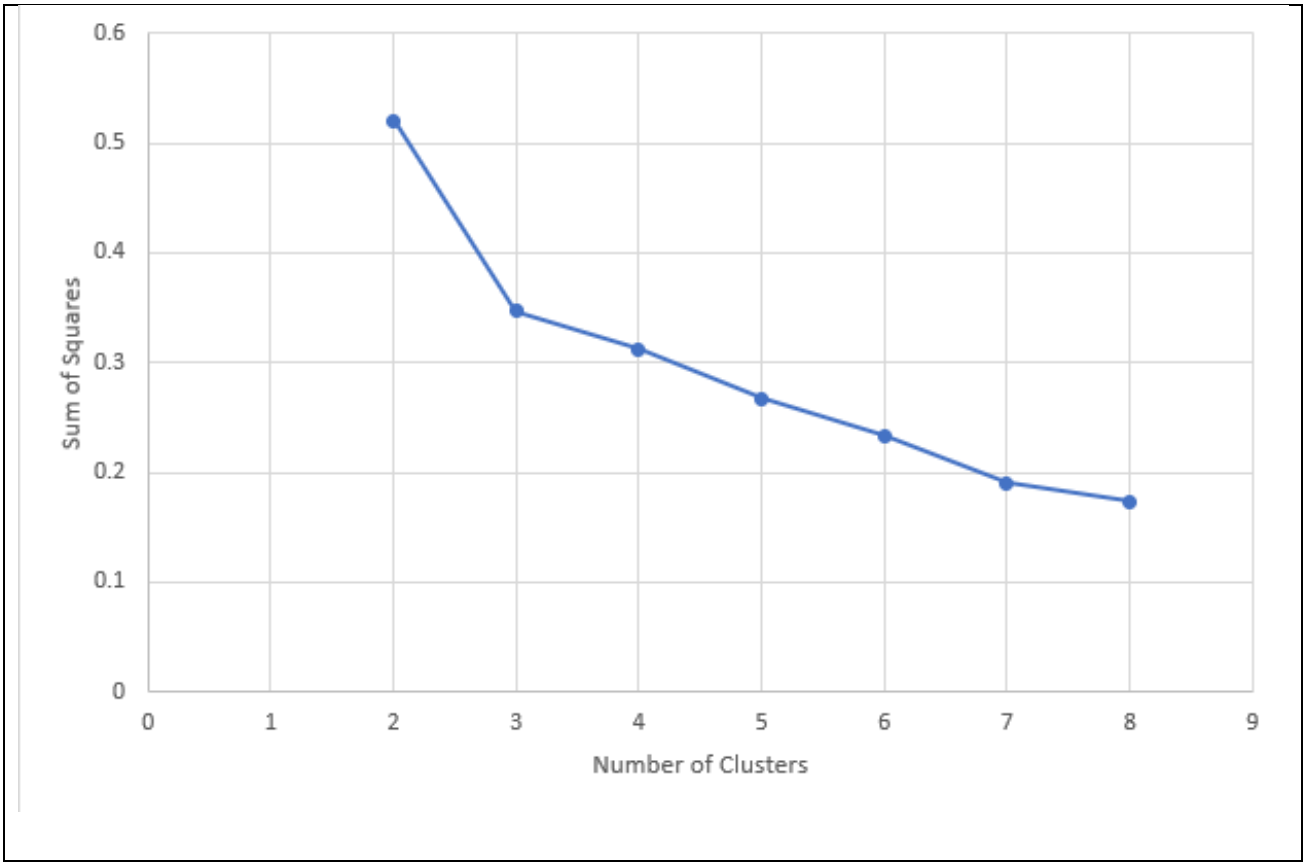
Screenshot of the dendrogram (make sure it is readable after printing):



Note that you can not find any performance measures anymore regarding distances to the centroid. Set the number of clusters from 2 until 8 (i.e., 2, 3, …, 8) in the *Flatten Clustering* operator and (only) report the overall sum of the squares. Create a visualization where this measure is on the y-axis, and the number of clusters is on the x-axis. Also report the number of customers (or records) in each cluster.

Number of clusters (k)	**Number of customers in each cluster {n1; n2; n3; …}**	**Sum of squares**
2	264; 176	0.520
3	176; 106; 158	0.347
4	154; 106; 158; 22	0.312
5	117;158;22;106;37	0.267
6	158;106;48;37;69;22	0.233
7	106;69;48;37;22;125;33	0.190
8	48;69;33;86;22;125;20;37	0.173

Screenshot of the visualization (make sure it is readable after printing):



What should be the preferred number of clusters based on the elbow technique? Motivate your answer.

Using the elbow technique on the graph above makes it evident that the preferred number of clusters that should be chosen is three. The prudent choice is three clusters as it has the sharpest change out of any of the other points, and the rate of decrease (sum of squares) slows down drastically after 3 clusters.

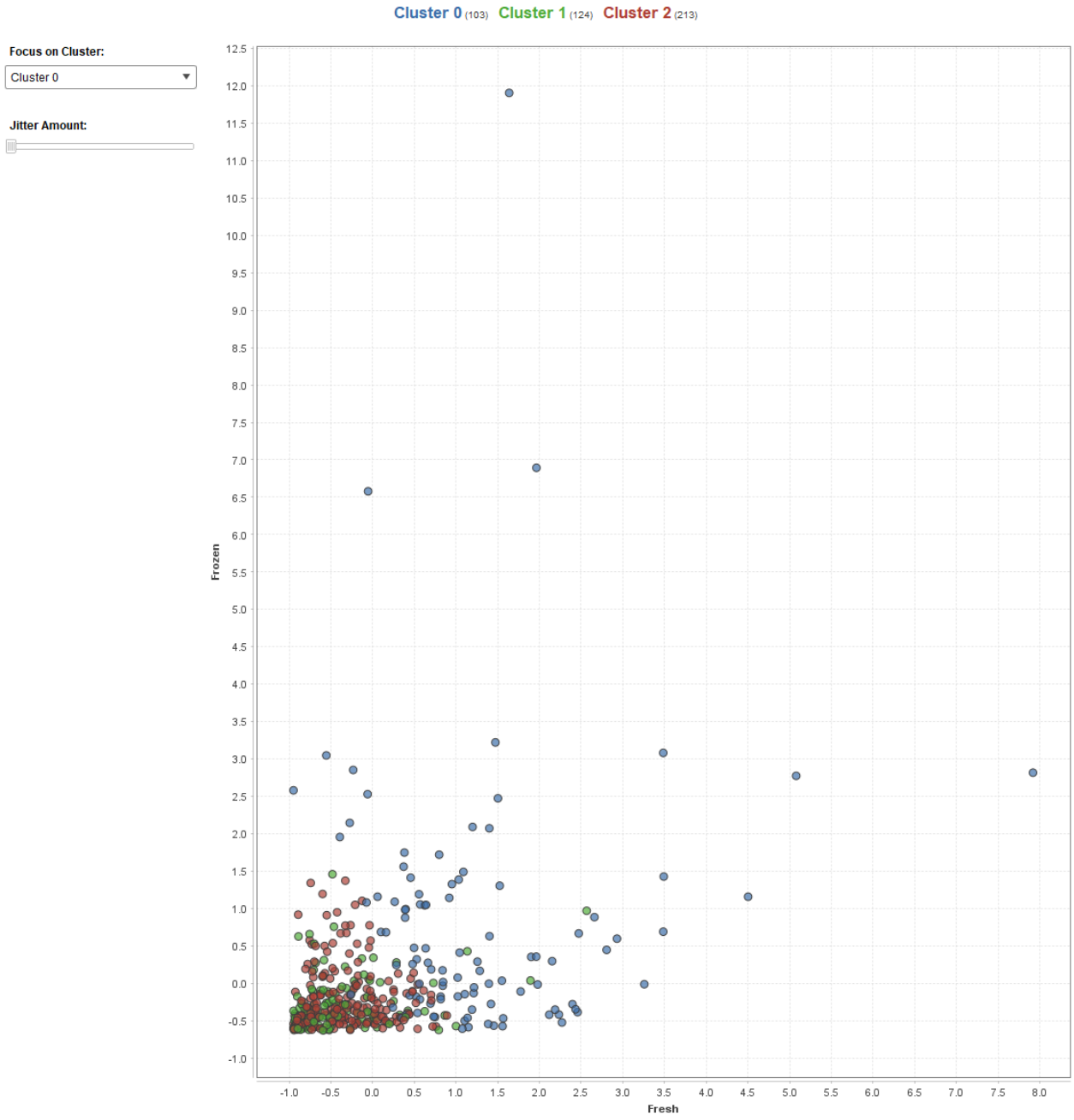
For the number of clusters that you selected, include screenshots of the scatter plots for each of the clusters to describe each customer profile in the clusters (i.e., clearly describe the spending behavior of customers in each cluster and support your argumentation with the output from RapidMiner, where you include screenshots).

(Repeat this textbox for each cluster)

Description of cluster 0: When focusing on cluster 0, it is evident that cluster 1 and 2 have very similar levels of spend. Additionally, both the aforementioned clusters have low, but similar levels of spend for fresh and frozen foods. Furthermore, they both have low variation in terms of spend, and they are highly concentrated in a small range of spend. “Cluster 0” seems to have a lot higher levels of spend and shows more variability. Subsequently, “cluster 0” also spends similar amounts of

both categories, but can lean towards one category or the other in certain instances. However, the spending for customers in “cluster 0” seems to be quite sporadic.

Screenshot of scatterplot to support your motivation (make sure it is readable after printing):

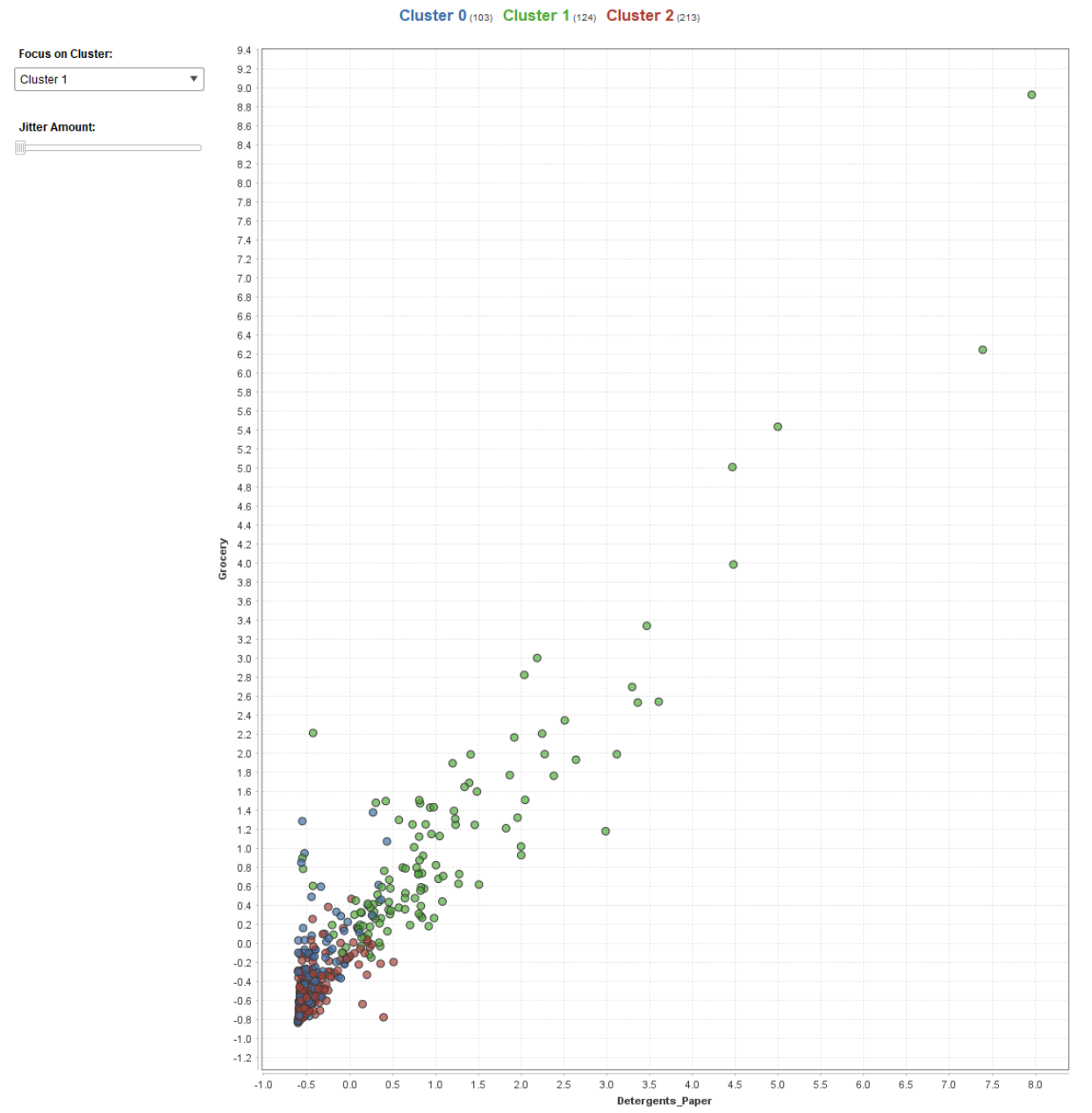


Description of cluster 1:

As for cluster 1, when we focus on this group, we can see a fairly linear spend across both product categories. Customers in “cluster 1” seem to spend significantly more compared to those featured in other clusters within the scatterplot. In other words, “Clusters 0 and 2” have similar spend to one another and have generally a lower spend when compared to

“cluster 1”. Overall, there is a very strong positive correlation between spending on Grocery and Detergents/Paper, especially with regards to those in “cluster 1”. This signifies that people who spend a lot of money on Groceries will also spend a similar amount on Detergents/Paper.

Screenshot of scatterplot to support your motivation (make sure it is readable after printing):

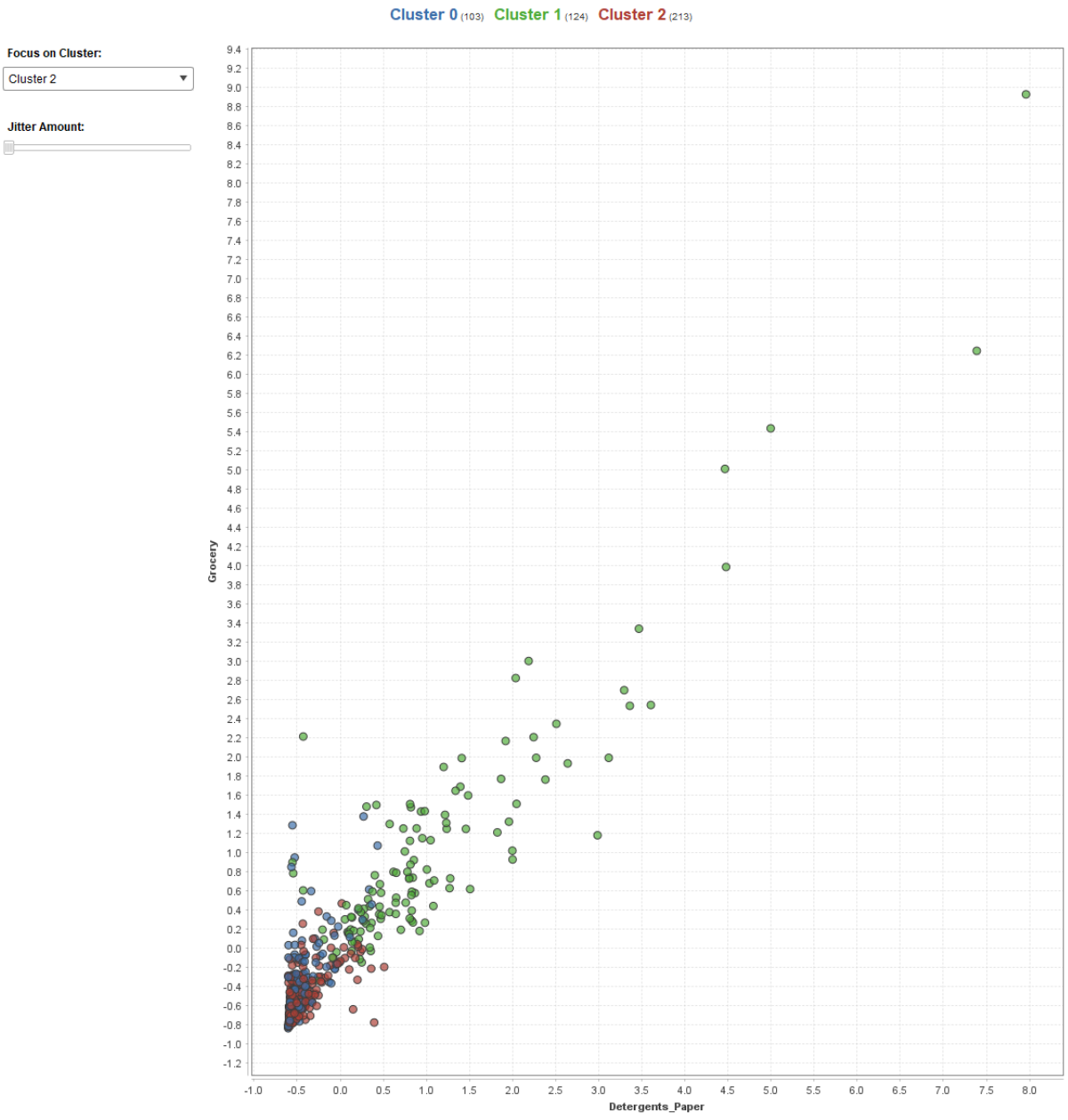


Description of cluster 2:

With regards to cluster 2, we can see quite similar spending behaviours between both categories. When focusing on cluster 2, clusters “0” and “2” are highly concentrated between the normalized spend of ~-0.6 to +0.6 for detergents/paper and between ~-0.8 to +0.4.

Conversely, “cluster 1” is the opposite, only significantly appearing for spending higher than ~+0.5 for detergents/paper and grocery. “Cluster 1” seems to spend more, and has significantly more variance in its spending when compared to the other two clusters shown in the scatter plot. The three clusters show a strong positive relationship between spending on Groceries and Detergents/Paper, which indicates that people who spend a lot of money on Groceries will also spend a similar amount on Detergents/Paper.

Screenshot of scatterplot to support your motivation (make sure it is readable after printing):

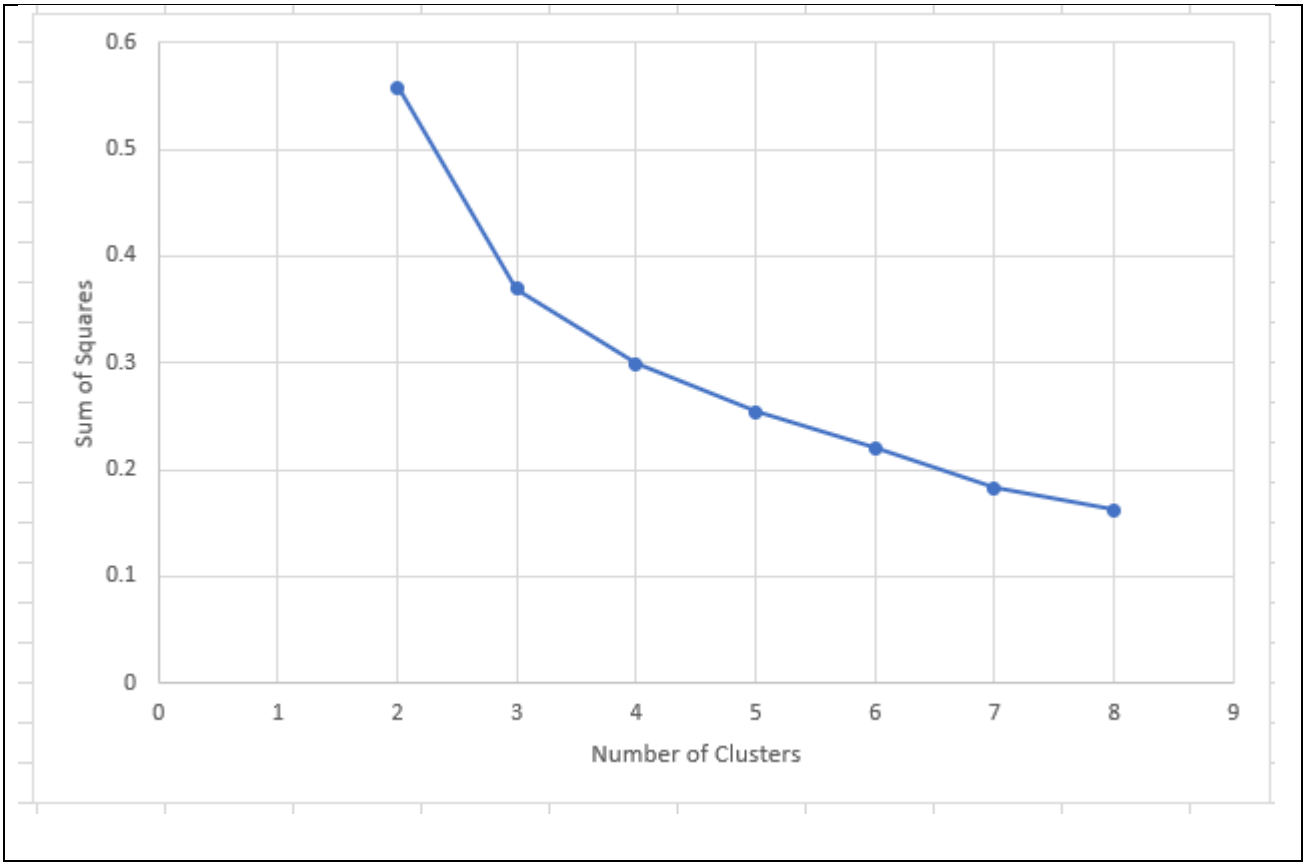


k. Save the file of Question 1j as "A4.Q1k.rmp" in the process folder of your Local Repository. Update the process to include the top down (or divisive) clustering as hierarchical clustering technique, where the k-Means clustering technique is used as the inner flat clustering technique. For the hierarchical clustering use a maximum depth of 4 and set the maximum leaf size to 150. For the k-Means clustering, use the same parameter values for the clustering process as in the previous questions. Make sure to set the number of clusters consistent throughout your clustering process. After the hierarchical clustering, use the *Flatten Clustering* operator to create actually clusters.

Set the number of clusters from 2 until 8 (i.e., 2, 3, …, 8) in the clustering process and (only) report the overall sum of the squares. Create a visualization where this measure is on the y-axis, and the number of clusters is on the x-axis. Also report the number of customers (or records) in each cluster.

Number of clusters (k)	**Number of customers in each cluster {n1; n2; n3; …}**	**Sum of squares**
2	295;145	0.558
3	213;203;124	0.369
4	185;116;86;53	0.299
5	176;84;56;49;75	0.254
6	158;69;77;24;64;48	0.220
7	137;68;65;47;43;58;22	0.183
8	123;69;41;60;60;20;37;30	0.162

Screenshot of the visualization (make sure it is readable after printing):



What should be the preferred number of clusters based on the elbow technique? Motivate your answer.

According to the graphs above, the elbow technique suggests that the preferred number of clusters that should be chosen is three. The clear choice is three clusters as it has the sharpest change out of any of the other points, and the rate of decrease (sum of squares) slows down drastically after 3 clusters.

Similar to the previous question, based on the number of clusters that you selected, include screenshots of the scatter plots for each of the clusters to describe each customer profile in the clusters.

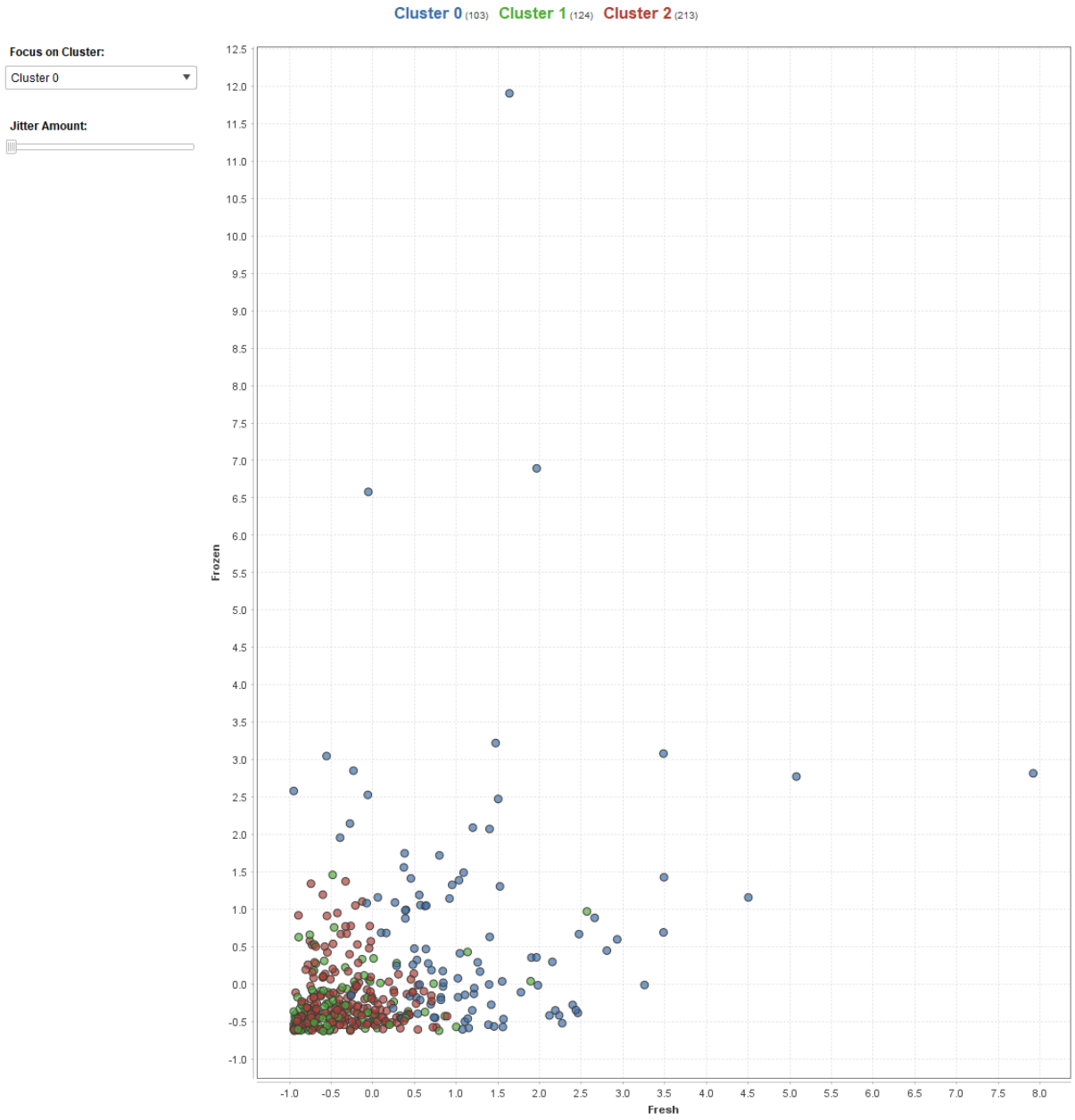
(Repeat this textbox for each cluster)

Description of cluster 0:

When focusing on cluster 0, cluster "1" and 2" are highly concentrated up to ~1.0 for fresh and ~1.5 for frozen. This indicates that "clusters 1 and 2" tend spend less and have very similar trends for spending. They also have less variance in terms of spending as they highly concentrated in the ranges mentioned. Furthermore, cluster 0 only appears after ~1.0 for fresh. It also has increasing levels of spending and variability. This indicates that spenders in

“cluster 0” tend to shop more erratically and spend various amounts on both fresh and frozen items.

Screenshot of scatterplot to support your motivation (make sure it is readable after printing):

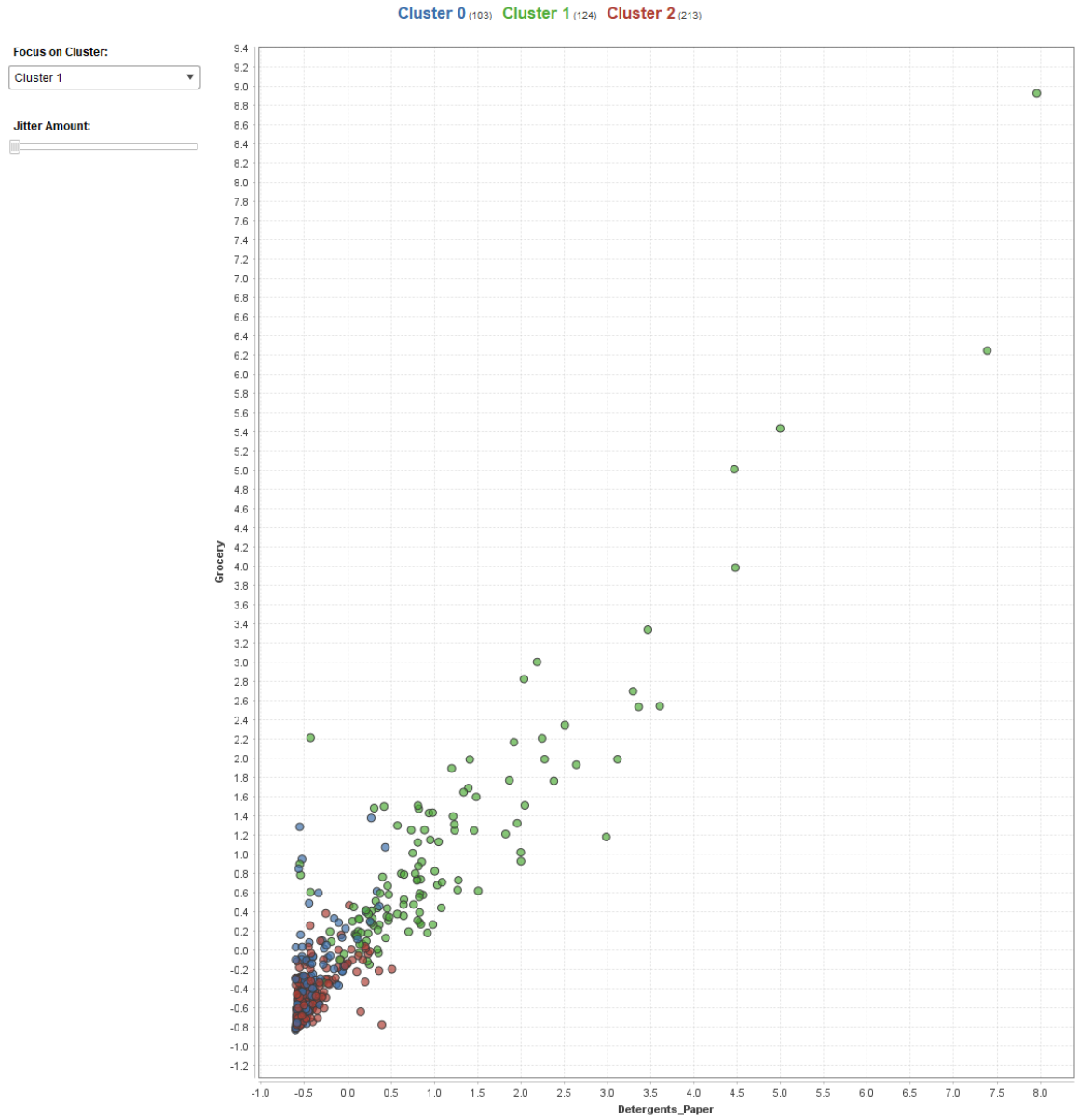


Description of cluster 1:

When focusing on Cluster 1, if we ignore extreme outliers, we can see similar overall spend between both categories. Interestingly, when focusing on cluster 1, “clusters 0 and 2” are highly concentrated between the normalized spend of ~-0.6 to +0.6 for detergents/paper and between ~-0.8 to +0.4. Whereas cluster 1 is the opposite, only significantly appearing for

spend higher than ~+0.5 for detergents/paper and grocery. “Clusters 0 and 2” have similar spend to one another and have generally a lower spend when compared to “cluster 1”, The three clusters show a strong positive relationship between spending on Groceries and Detergents/Paper. This signifies that people who spend a money on Groceries will also spend a similar amount on Detergents/Paper. This scatterplot confirms that Grocery and Detergents/Paper have a very high positive correlation.

Screenshot of scatterplot to support your motivation (make sure it is readable after printing):



Description of cluster 2:

Within Cluster 2, we can observe similar spend between both categories. Interestingly, when focusing on cluster 2, clusters “0” and “2” are highly concentrated between the normalized spend of ~-0.6 to +0.6 for detergents/paper and between ~-0.8 to +0.4. Whereas “cluster 1” is

the opposite, only significantly appearing for spending higher than ~+0.5 for detergents/paper and grocery. Clusters “0” and “2” have very similar spending, and cluster 1 seems to spend more, and has significantly more variance in its spending when compared to the other two clusters. The three clusters show a strong positive relationship between spending on Groceries and Detergents/Paper. This signifies that people who spend a lot of money on Groceries will also spend a similar amount on Detergents/Paper. This scatterplot confirms that Grocery and Detergents/Paper have a very high positive correlation.