Luther Nicholaus
CIS 607
Unit 5 Assignment

Software Used

The software used for data analysis are Microsoft Excel and SPSS. Excel is used to divide the primary dataset into two chunks (80% to 20% samples). The later is established on the random sampling technique. SPSS aids cluster analysis of the dataset.

Dataset

The dataset used for analysis is titles "Wholesale Customers Data Set". It is obtained from the UCI machine learning repository. It informs the annual spending on various product categories by customers of a wholesale supplier. The variables in the dataset are Channel, Region, Fresh, Milk, Grocery, Frozen, Detergents_Paper, and Delicatessen (Cardoso, 2011).

Attribute	Information
Region	Area the client is located
Channel	Channel of industry for the client
Fresh	Annual spending on fresh products
Milk	Annual spending on milk products
Grocery	Annual spending on Grocery products
Frozen	Annual spending on frozen products
Detergents_Paper	Annual spending on detergents & paper products
Delicatessen	Annual spending on Delicatessen products

Input Variables used for Cluster Analysis

The input variables for the analysis are Fresh, Milk, Grocery, Frozen, Detergents_Paper, and Delicatessen. These variables relate to the amounts spent on various product categories by each of the wholesaler's clients. This will aid the grouping of clients/customers based on their purchasing patterns. The grouping will help marketers identify and predict the class of new

customers, hence prescribing appropriate strategies for maximum conversion, profitability, satisfaction, and retention.

The K-Means clustering algorithm is implemented in SPS using the selected variables. Based on observation from the ANOVA table, multiple tests using different number of clusters are performed to determine the optional number of clusters in which there is no significant difference in means in the clusters formed. 3 clusters satisfied the low significance level prerequisite.

Results and Business Meaning

The final cluster centers for the three clusters are displayed on Fig. 1 and Fig. 2. Cluster 1 includes clients that are high spenders. These customers spend more on Milk, Grocery, and Detergents, and Delicatessen product categories. Cluster 2 contains clients that are low spenders. Cluster 3 contains clients that are high spenders, buying relatively more Fresh and Frozen products. The ANOVA table (Fig. 3) indicates that there is statistical significance for all the variables on determining which cluster a client was grouped into.

Using the Fig. 4, one can determine that frequency of clients for each cluster. The wholesaler has a lot of clients that are low spenders (relatively). Most of the high spending clients spend more towards the Fresh products category. This conclusion follows the fact that there are more clients on cluster 3 than cluster 1.

Final Cluster Centers

	Cluster		
	1	2	3
Fresh	8761	7317	30836
Milk	17318	3892	4313
Grocery	26761	5410	5475
Frozen	2797	2319	5081
Detergents_Paper	12107	1862	953
Delicassen	3428	1119	1955

Fig. 1: Cluster Centers

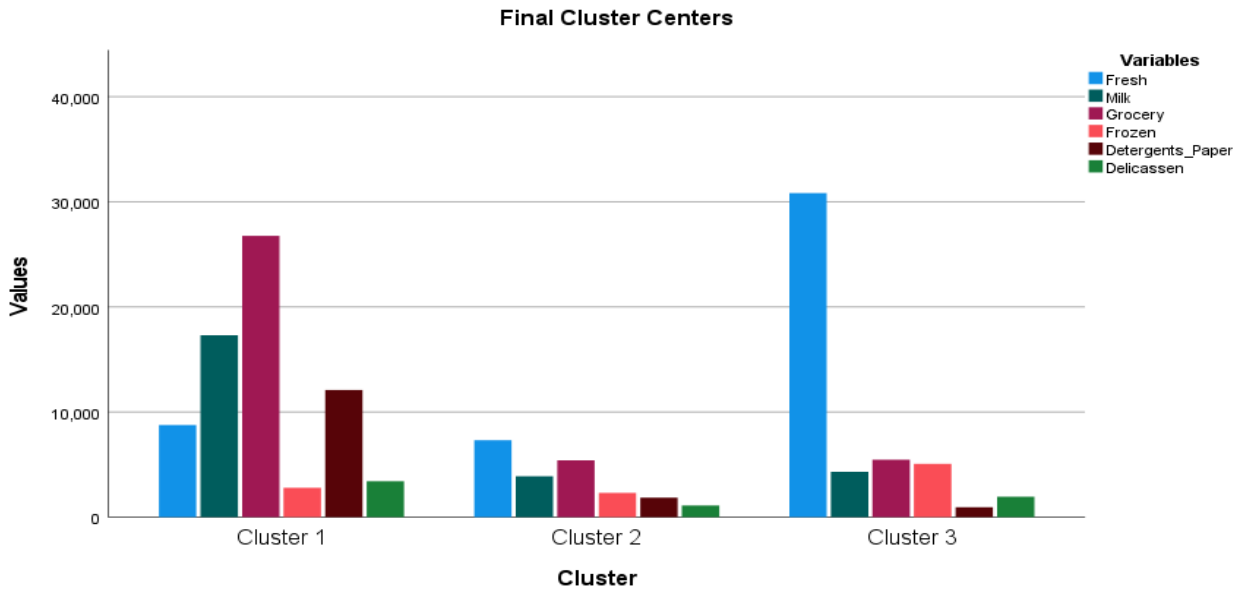

Fig. 2: Cluster Centers Visual

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Fresh	1.356E+10	2	60759968.75	349	223.182	<.001
Milk	3297538997	2	22533660.54	349	146.338	<.001
Grocery	8421290581	2	32845218.66	349	256.393	<.001
Frozen	184580610.4	2	12592303.20	349	14.658	<.001
Detergents_Paper	2028340003	2	9695066.992	349	209.214	<.001
Delicassen	102198593.3	2	8686963.152	349	11.765	<.001

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Fig. 3: ANOVA table showing the significant levels.

Number of Cases in each Cluster

Cluster	1	42.000
	2	250.000
	3	60.000
Valid		352.000
Missing		.000

Fig. 4: Number of clients in each cluster.

Prediction using the Smaller Sample Dataset.

Using the cluster centroids, the prediction of six records is presented below. With a high confidence level, as shown in the ANOVA table, a marketer is expected to make the right prediction classifying customers. This could yield improved marketing efficiency, hence increased sales, profitability, and customer satisfaction.

Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen	Prediction Cluster
3	31812	1433	1651	800	113	1440	Cluster 3 (High Spender)
3	45640	6958	6536	7368	1532	230	Cluster 3 (High Spender)
3	16165	4230	7595	201	4003	57	Cluster 3 (High Spender)
3	1406	16729	28986	673	836	3	Cluster 1 (High Spender)
3	4591	15729	16709	33	6956	433	Cluster 1 (High Spender)
1	5396	7503	10646	91	4167	239	Cluster 2 (Low Spender)
3	22925	73498	32114	987	20070	903	Cluster 1 (High Spender)
3	2647	2761	2313	907	95	1827	Cluster 2 (Low Spender)

Fig. 5: Predictive Analysis.

References

Cardoso, M. (2011). *UCI Machine Learning Repository*. Retrieved from https://archive.ics.uci.edu/ml/datasets/wholesale+customers