Customer Analytics in Retail Supermarket Chain

BY:

AKANKSHA SHARMA

NAMRITA GUPTA

Introduction & Background

- Dunnhumby is a worldwide retail chain spread across 29 countries with the revenue of $3 billion.
- The business of retail industry depends largely upon the customers they serve. So, it becomes very important to build and sustain relationship with customers.
- In order to manage customer relationships successfully, segmentation of customers becomes a crucial part.
- Customer segmentation would help in increase in company's profit apart from providing selective marketing strategies as per customers segment.

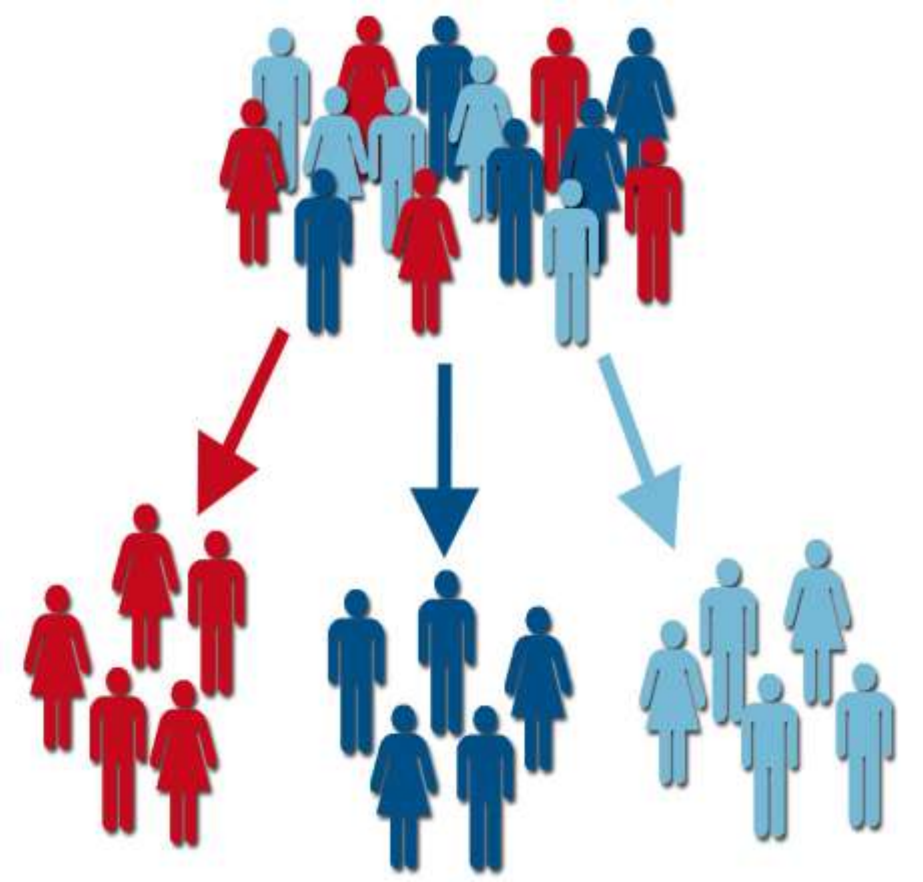
https://en.wikipedia.org/wiki/Dunnhumby
https://www-sciencedirect-com.proxy.libraries.rutgers.edu/science/article/pii/S0957417411010761

Approach: Data Preparation and Cleaning

- Transactional data consist of purchase details of 2500 households for period of 2 years whereas demographic data consist of 800 households only.
- We are planning to aggregate the transactional data at household level (customer level) and then merge it with demographic data to make it to 800 households only.
- The transaction data didn't have any missing values, but we are planning to perform exploratory data analysis on this dataset to find out the probable outliers, if any using boxplot.
- We will also look at the distribution of all the attributes to check for any skewness of data.
- Three attributes from demographic data will be removed as each contain more than 20% missing values. However absence of 2 of these attributes viz. HH_COMP_DESC and KID_CATEGORY_DESC can be compensated with HOUSEHOLD_SIZE_DESC attribute which has no "Unknown" values.

Approach: Learnings and Approach

- The goal is to form clusters which distinctly represent the customer purchase behaviour.
- Feature selection in WEKA, will be performed using Linear Regression as classifier which is Supervised learning. The class of this regression data would be the customer spend.
- Unsupervised learning will be used for clustering the customers.
- K-means and EM clustering algorithms will be applied on this merged dataset and then resulting clusters will be compared.



https://www.profiletree.com/customer-segmentation/

Approach: Evaluation Methods

- The resulting clusters should distinctly represent customer spend and involved demographic attributes.
- Thus our evaluation would be focused on gauging whether the clusters formed are representative of the average sales value of that cluster, for example all customers associated with high sales value should be the part of one cluster.
- Classes to Clusters Evaluation method would be employed by using class as "Customer spend".
- The correctness of the clusters would be decided on the basis of whether the instances belonging to a particular class falls under a same cluster or not. This would be shown by incorrectly clustered instances percentage in WEKA.

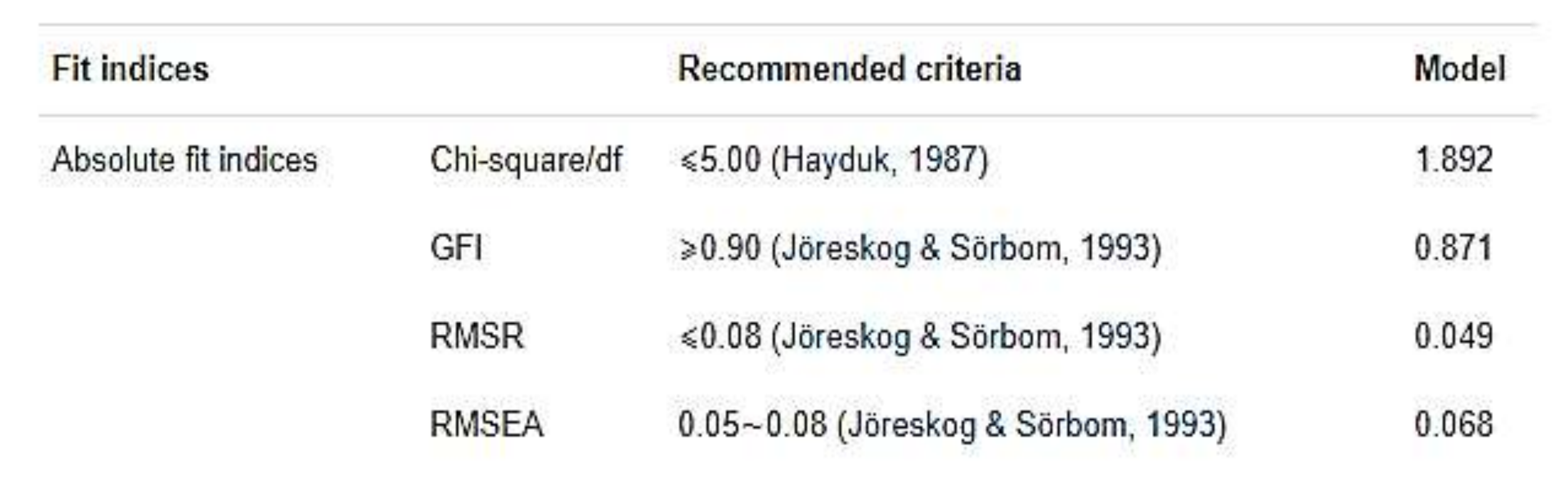
Literature Review

- One of our primary literature review *Segmenting customers in online stores based on factors that affect the customer's intention to purchase* highlighted the importance of clustering techniques in segmenting the data of online customers.
- According to literature these segments can be utilized in designing customized marketing offers. Before performing segmentation important features, that determine the customer purchase intention, were selected using Structural equation Model (SEM).
- K-means and SOM (Self-organized maps) are two commonly used clustering methods for customer segmentation.

https://www-sciencedirect-com.proxy.libraries.rutgers.edu/science/article/pii/S0957417411010761
https://doi.org/10.1016/j.eswa.2011.07.114

Literature Review

- K-means – In Weka, the fitness of data to the model can be measured by minimising sum of squared error.



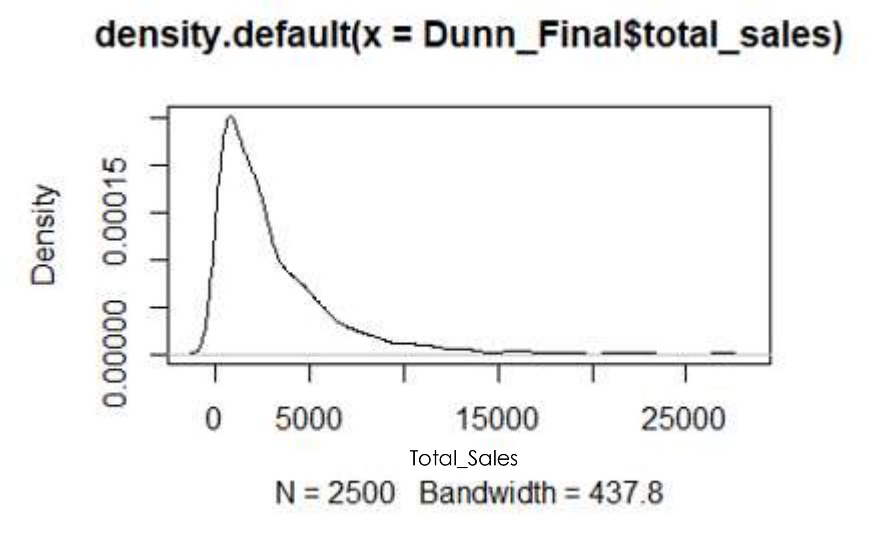
Fit indices		Recommended criteria	Model
Absolute fit indices	Chi-square/df	⩽5.00 (Hayduk, 1987)	1.892
	GFI	⩾0.90 (Jöreskog & Sörbom, 1993)	0.871
	RMSR	⩽0.08 (Jöreskog & Sörbom, 1993)	0.049
	RMSEA	0.05~0.08 (Jöreskog & Sörbom, 1993)	0.068

- Expectation maximization – The fitness of data to the model can be measured by loglikelihood.

https://www-sciencedirect-com.proxy.libraries.rutgers.edu/science/article/pii/S0957417411010761
https://pdfs.semanticscholar.org/eca2/5eb78be04ffe09b029dd1d36f5ba66749f29.pdf

Results and Discussion – Data Preparation

- The initial distribution of customer spend showed right skewed result.

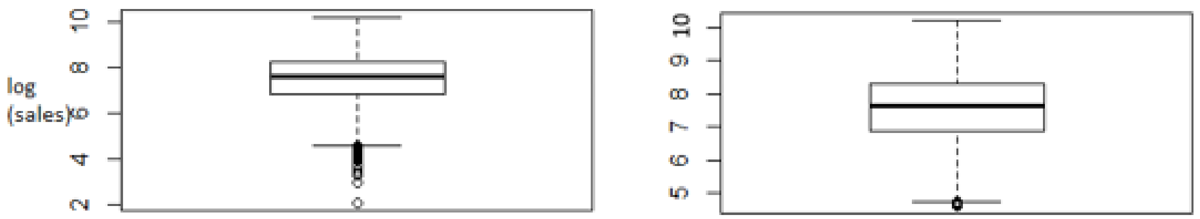


Density plot of "total_sales" before treatment

- On deep diving we found that Gasoline-Reg-Unleaded have very low product price, nearly 0.2 cents and are purchased in very high quantities. So these inaccurate instances were removed from the dataset.

Results and Discussion – Data Preparation

- Then the skewness in data was further reduced by implementing different transformations on customer spend like log, square, square root. Finally, Logarithmic function gave the better results among others in reducing the skewness by a larger degree.
- Box Plot in R was used for detecting the outliers in customer spend and the resulting 40 outliers were removed from the dataset.



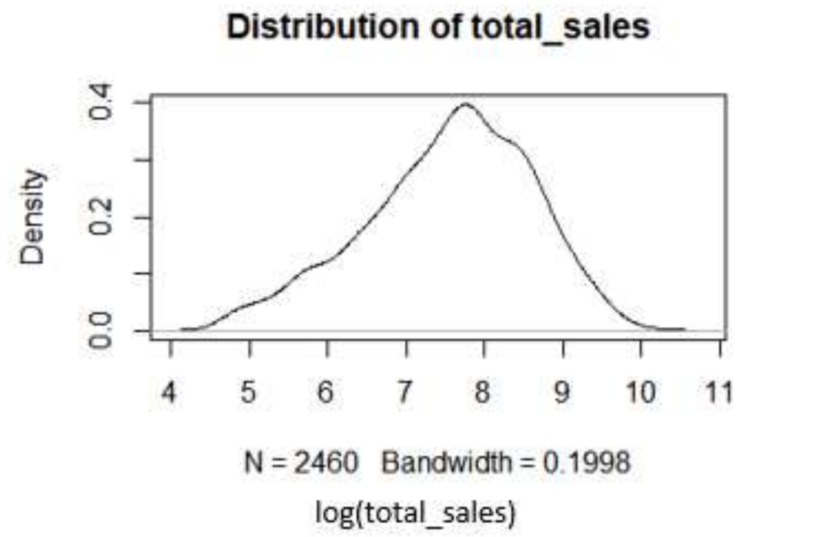
Before outliers treatment

After outliers removal

Outliers detection using Boxplot

Results and Discussion – Data Preparation

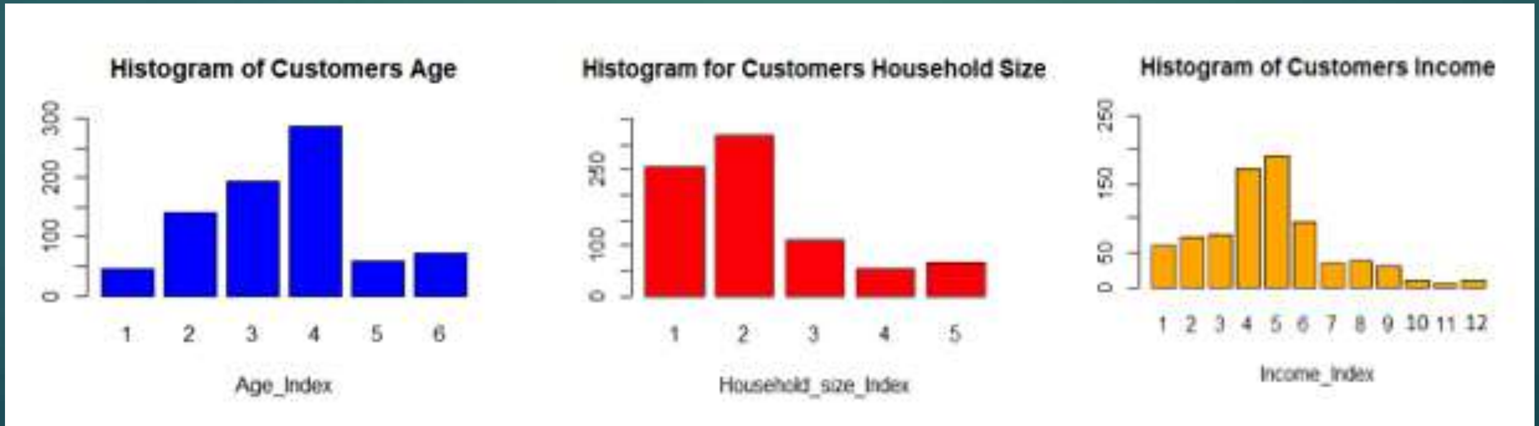
- The final density plot of customer spend showed the normal distribution of customers.



Density plot of "total_sales" after treatment

Results and Discussion – EDA of Demographic data

- The maximum number of customers fell in the age bucket of 45-54 (Age_Index** = 4).
- Customers having the household size of 2 (Household_size_Index** = 2) showed the maximum frequency.
- Most of the customers in dataset have income in the range of 50-74k (Income_Index** = 5).



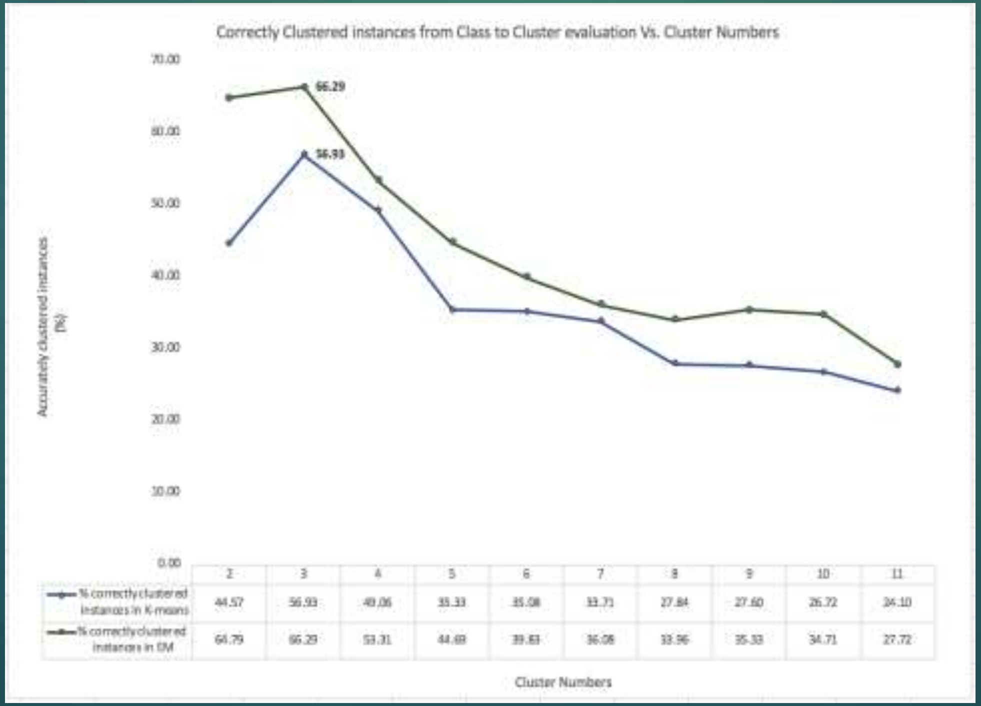
** Refer Appendix (Slide 29) for Index to Values

Results and Discussion – Features Selection

- Feature selection in WEKA was performed using Wrapper Subset.
- In this process, linear regression was used to identify the significant features corresponding to “customer spend” (class).
- Attributes were selected as the input for executing clustering algorithms were number of product purchased, total number of order, total number of visits, number of national products and income.
- From demographic data income was the only attribute that came as significant.

Results and Discussion – Optimal Number of Cluster

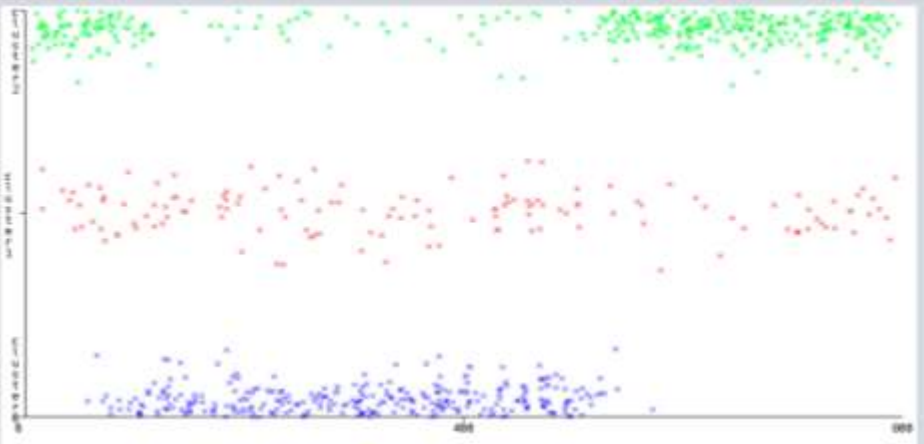
- Graph was plotted between “class to cluster” performance accuracy of clustering algorithms and different cluster numbers ranging from 2 to 11. Both the algorithms performed best at “3” value of cluster number.



	2	3	4	5	6	7	8	9	10	11
% correctly clustered instances in K-means	44.57	56.93	49.06	35.33	35.08	33.71	27.84	27.60	26.72	24.10
% correctly clustered instances in EM	64.79	66.29	53.31	44.69	39.83	36.08	33.96	35.33	34.71	27.72

Results and Discussion – K-Means Clustering

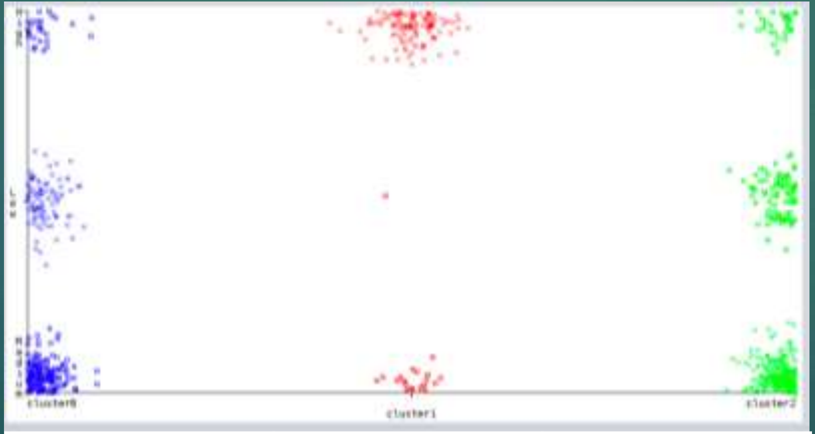
- K-means clustering was performed with value of k as 3 and seed as 10 using WEKA tool.
- The Unsupervised filter of WEKA was used to add resulting cluster numbers in the dataset and evaluated using Class to Cluster feature of WEKA



Cluster 0	Cluster 1	Cluster 2
Low-Valued Cluster	High-Valued Cluster	Medium-Valued Clusters

Results and Discussion – Class to Cluster Evaluation of K-means

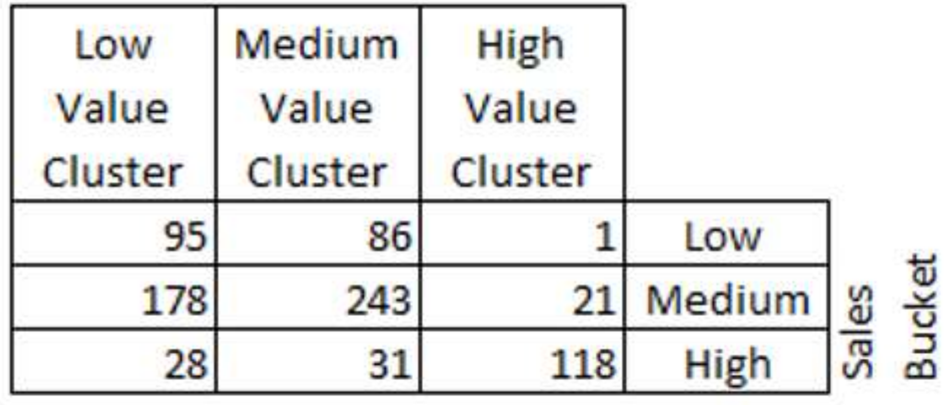
- The output showed three different clusters of customers associated with the three different class labels of “high”, “medium” and “low” value customers as well as the percentage of accurately clustered instances based on class values of customer spend.
- The percentage of correctly clustered instances by K-means was around 57%.



Cluster 0	Cluster 1	Cluster 2
Low-Valued Cluster	High-Valued Cluster	Medium-Valued Clusters

Results and Discussion – Class to Cluster Evaluation of K-means

- Low-valued customers were most incorrectly clustered using K-means algorithm, i.e., 68% customers were inaccurately segmented.
- Majority of these customers which were actually medium valued customer were inaccurately clustered into low-valued segment.

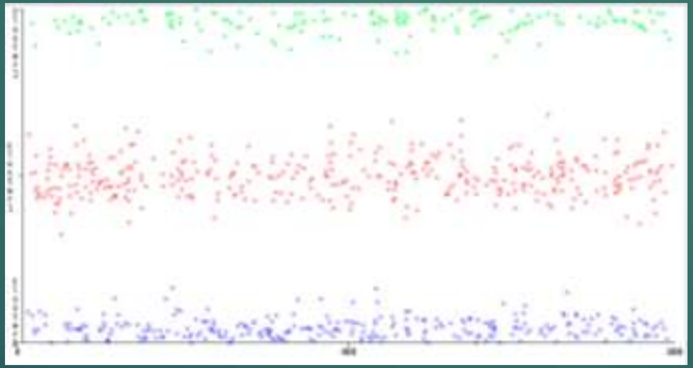


Low Value Cluster	Medium Value Cluster	High Value Cluster	Sales Bucket
95	86	1	Low
178	243	21	Medium
28	31	118	High

Class to Cluster Evaluation Matrix of K-means

Results and Discussion – Expectation Maximization

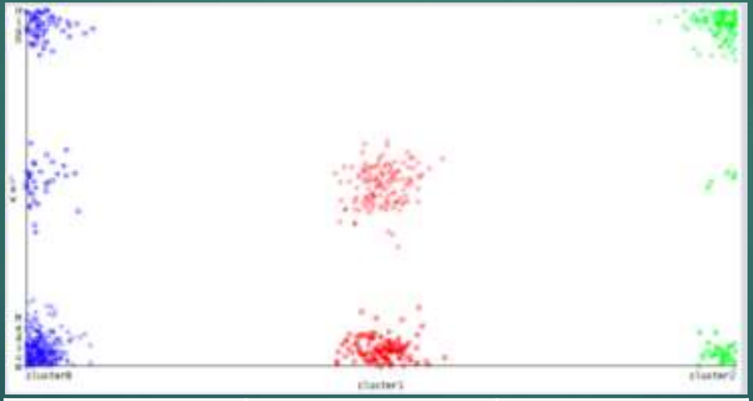
- EM was performed with the seed value as 100 and number of clusters as 3
- Assignment of 800 instances into different clusters are visualized in WEKA.



Cluster 0	Cluster 1	Cluster 2
Low-Valued Cluster	High-Valued Cluster	Medium-Valued Clusters

Results and Discussion – Class to Cluster Evaluation of EM

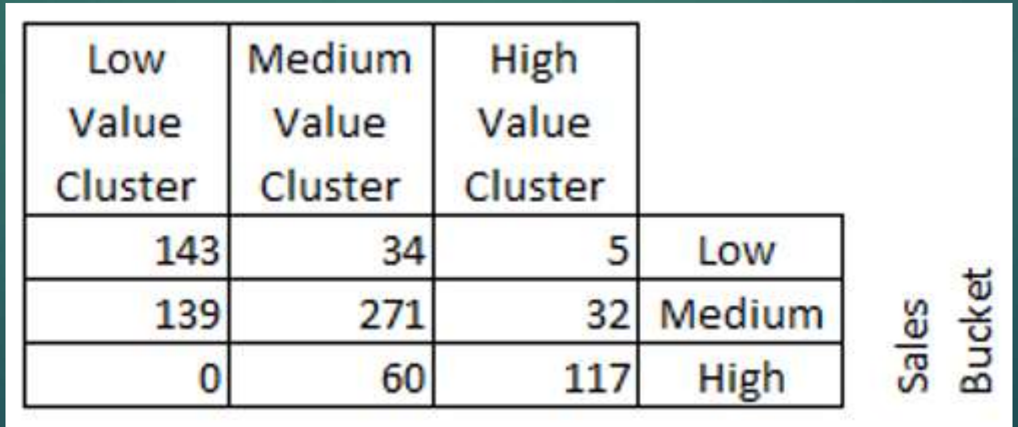
- The Unsupervised filter of WEKA was used to add resulting cluster numbers in the dataset.
- The percentage of correctly clustered instances by EM was around 66%.



Cluster 0	Cluster 1	Cluster 2
Medium-Valued Cluster	Low-Valued Cluster	High-Valued Clusters

Results and Discussion – Class to Cluster Evaluation of EM

- EM showed a relatively higher performance than K-means in terms of correctly clustering the low-valued instances by reducing the inaccurate percentage to 49%.
- Moreover, EM keeps a separate boundary between High and Low-Valued customers as there was “0” inaccurate assignment of actual high valued customer into low valued cluster.

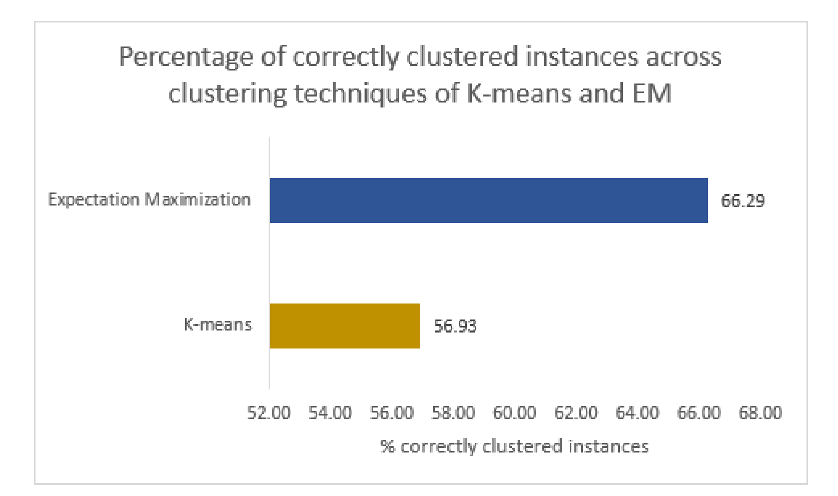


Low Value Cluster	Medium Value Cluster	High Value Cluster	Sales Bucket
143	34	5	Low
139	271	32	Medium
0	60	117	High

Class to Cluster Evaluation Matrix of EM

Results and Discussion – Performance comparison of K-means & EM

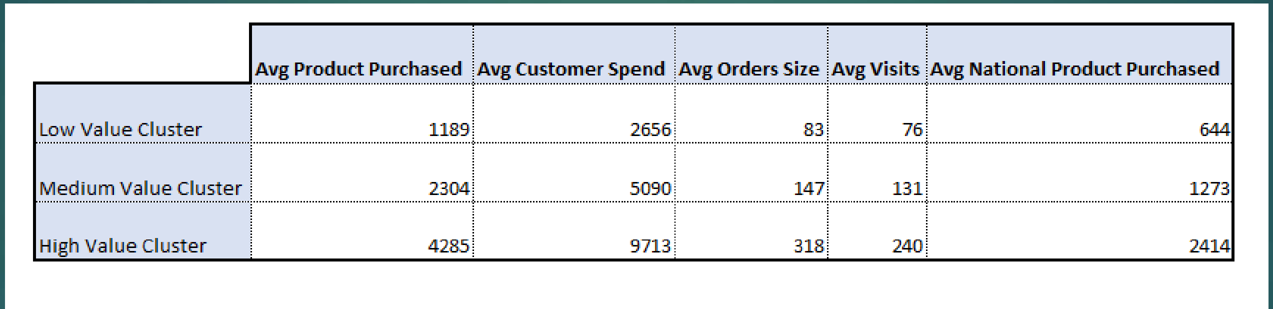
- EM performed better in segmenting the retail customers distinctively based on customer spend with 66% of correctly clustered instances.
- Clusters formed by EM were chosen for Classification Via Clustering.



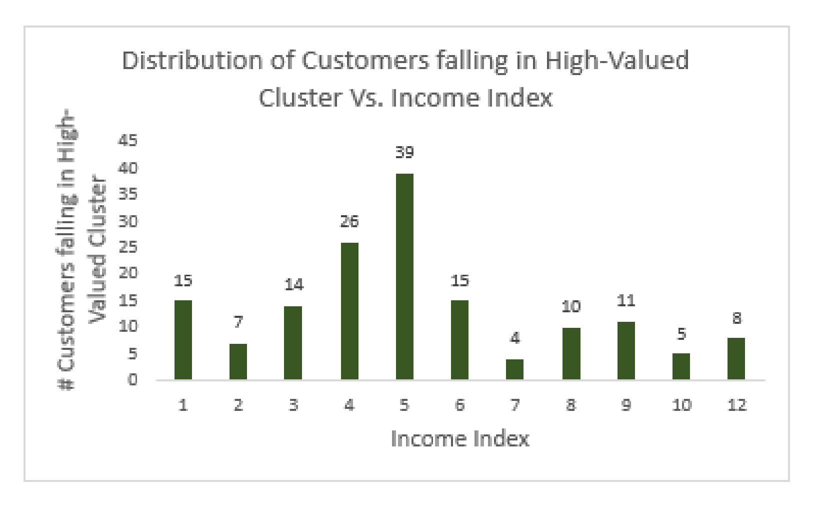
Results and Discussion – Clusters Interpretation

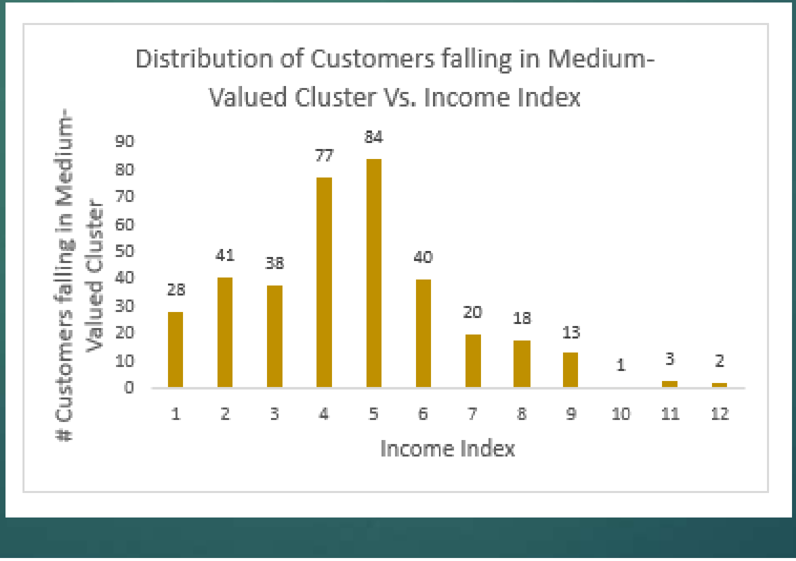
- Low-Valued Cluster – Customers falling in this cluster had the least number of product purchased, customer spend, number of orders they have made, number of visits as well as quantity of products with Brand National purchased as shown in below fig().
- Medium-Valued Cluster – Customers falling in this cluster made purchase having products quantity intermediate between that of customers falling into Low and High valued clusters. Moreover, their spend values, number of orders made, their number of visits as well as quantity of products with Brand National purchased in 2 years fell in range intermediate between that of customers in low and high valued clusters.
- High-Value Cluster - Customers falling in this cluster had the maximum number of product purchased, customer spend, number of orders they have made, number of visits as well as quantity of products with Brand National purchased.

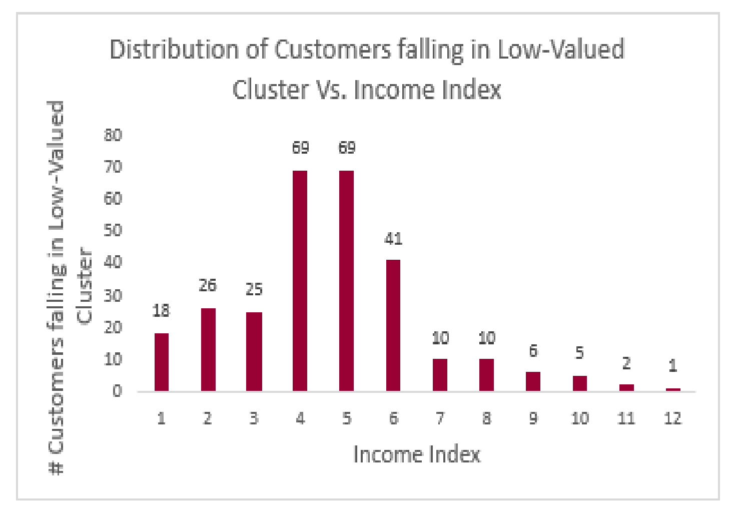
Results and Discussion – Clusters Interpretation

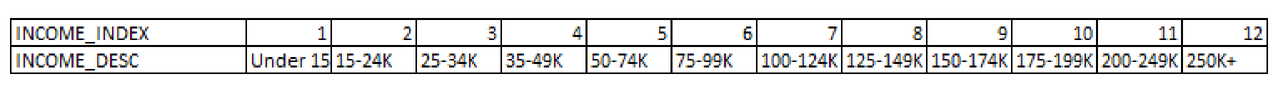


	Avg Product Purchased	Avg Customer Spend	Avg Orders Size	Avg Visits	Avg National Product Purchased
Low Value Cluster	1189	2656	83	76	644
Medium Value Cluster	2304	5090	147	131	1273
High Value Cluster	4285	9713	318	240	2414









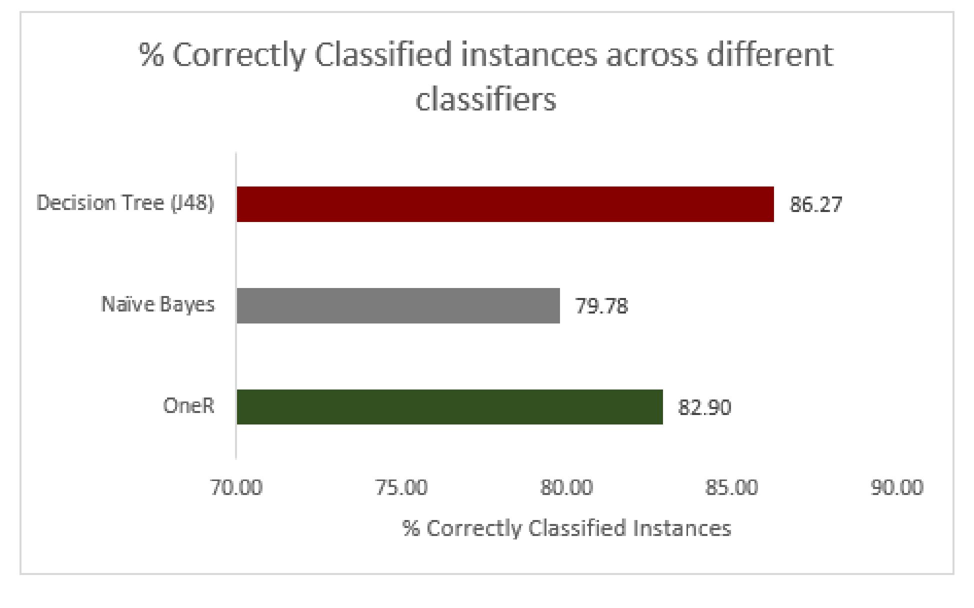
INCOME_INDEX	1	2	3	4	5	6	7	8	9	10	11	12
INCOME_DESC	Under 15	15-24K	25-34K	35-49K	50-74K	75-99K	100-124K	125-149K	150-174K	175-199K	200-249K	250K+

Results and Discussion – Clusters Interpretation

- Low-Valued Cluster – Only 12% of the total customer falling in low-valued cluster were having the income above 99K. Thus, customers falling in this cluster are not high-income customers.
- Medium-Valued Cluster – Among all the customers falling in medium valued cluster ,16% were having the income above 99K. Thus, customers falling in this cluster have a greater number of high-income customers than that present in low-valued cluster, whereas when compared with high-valued cluster, this number is lower.
- High-Valued Cluster – Among all the customers falling in this cluster 25% were having the income above 99K. Thus, customers falling in this cluster are having maximum number of high-income customers.

Results and Discussion – Classification Via EM Clustering

- J48 performed best among all classifiers with prediction accuracy of 86%



Conclusion

- EM yielded better clustering results than K-means.
- Few number of attributes of transaction data and only one income attribute of demographic data can yield a fairly high correctly clustered instances.
- Income attribute associated with customer spend behaviour.
- Clusters formed can be utilized in predicting the sales of new customers.
- Customer segmentation can be treated as an intermediate step for designing personalized marketing strategy. Each customer segment would then be analyzed at customer level and based on the his/her spend behavior one-to-one marketing offer would be rolled out.
- Another important aspect is Store Layout which has a significant impact on sales of products in retail supermarkets. Sequential Pattern Mining can be used to find out frequently occurring product categories as well as the products within them.

thank you
danke
謝謝
ngiyabonga
شكراً جزيلاً
рахмат
Баярлалаа
спасибо
teşekkür ederim
gracias
tapadh leat
dank je
mersi
barka
welalin
tack
faafetai lava
vinaka
спасибі
blagodaram
kiitos
dankie
misaotra
malondo
paldies
хвала
asante
manana
bedankt
enkosi
bayarlalaa
grazie
hvala
māuruuru
köszönöm
dziękuję
mochchakkeram
djiere dieuf
go raibh maith agat
mamnun
choorakaloodioun
gracies
sulpay
taiku
sobodi
dėkui
obrigado
mesi
sagolun
sukriya
kop khun krap
ありがとう
najis tuke
kam sah hamnida
terima kasih
tanemirt
rahmet
grazie
arigatō
takk
dakujem
trugarez
diolch
dhanyavadagalu
shukriya
мерси
감사합니다
xièxie
ευχαριστώ
merci

Appendix

References

Primary

1.Aloysius, G., & Binu, D. (2012). An approach to products placement in supermarkets using PrefixSpan algorithm. Journal of King Saud University ¨C Computer and Information Sciences, 25(1), 77–87. https://doi.org/10.1016/j.jksuci.2012.07.001

2.Hong, T., & Kim, E. (2011). Segmenting customers in online stores based on factors that affect the customer's intention to purchase. Expert Systems With Applications, 39(2), 2127–2131. https://doi.org/10.1016/j.eswa.2011.07.114

3. Sharma, N., Bajpai, A., & Litoriya, M. R. (2012). Comparison the various clustering algorithms of weka tools. facilities, 4(7), 78-80.

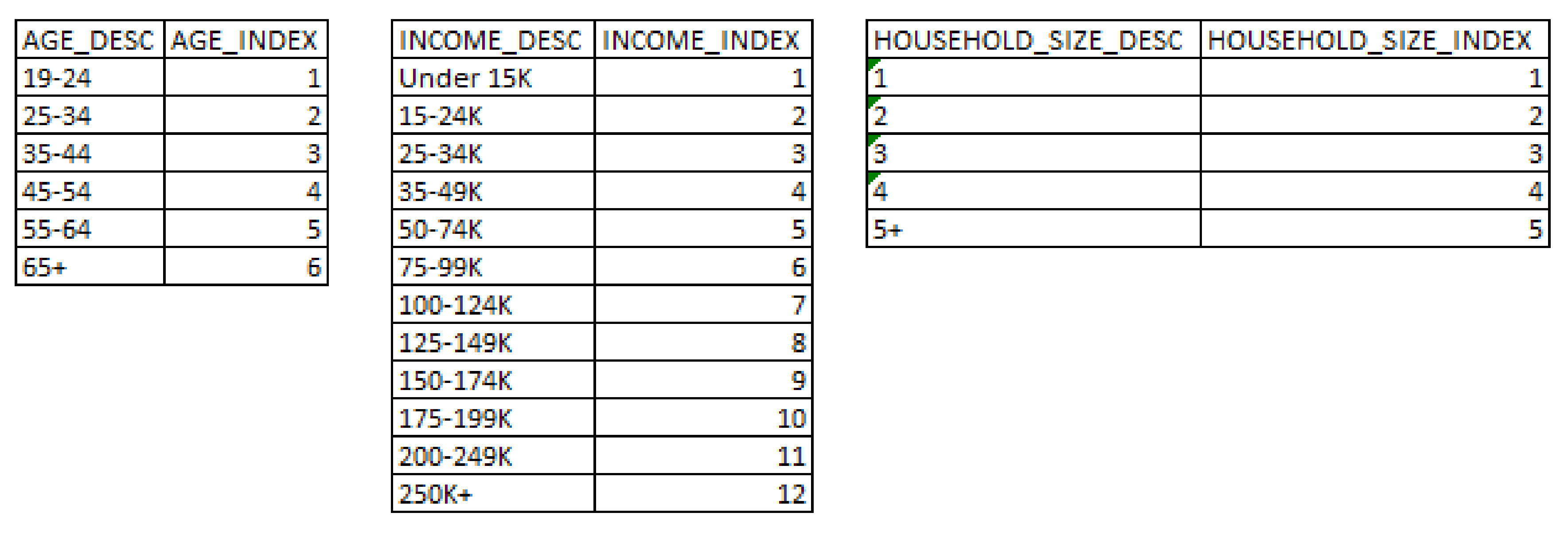
Secondary

4. Lockshin, L. S., Spawton, A. L., & Macintosh, G. (1997). Using product, brand and purchasing involvement for retail segmentation. Journal of Retailing and Consumer services, 4(3), 171-183.

5.Katsaras, N., Wolfson, P., Kinsey, J., & Senauer, B. (2001). Data mining: A segmentation analysis of US grocery shoppers. *St. Paul, MN: The University of Minnesota, The Retail Food Industry Center, Working Paper*, 01-01.

6.Ziafat, H., & Shakeri, M. (2014). Using Data Mining Techniques in Customer Segmentation. *Journal of Engineering Research and Applications*, 4(9), 70-79.

Values for Demographic attributes Indices

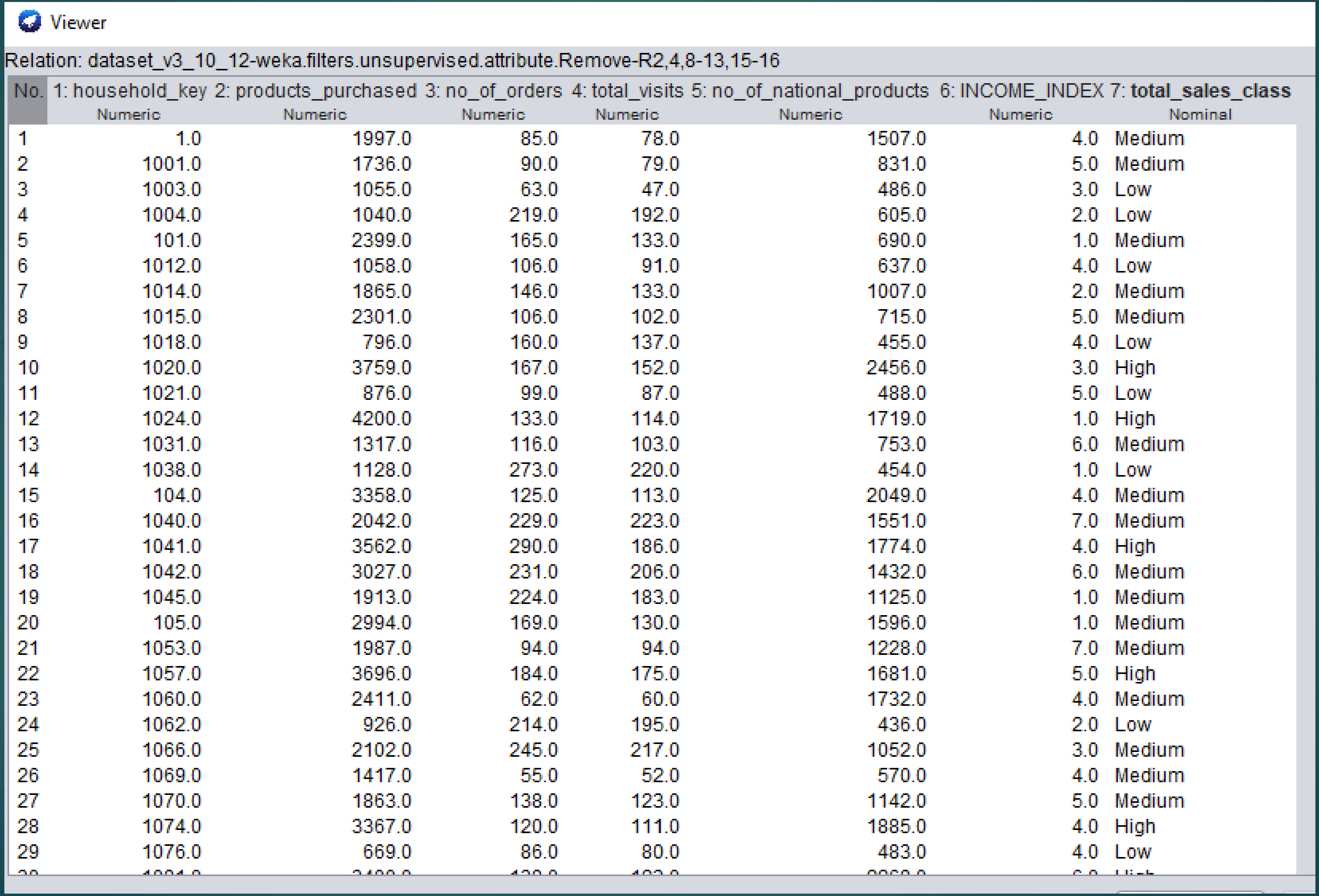


AGE_DESC	AGE_INDEX
19-24	1
25-34	2
35-44	3
45-54	4
55-64	5
65+	6

INCOME_DESC	INCOME_INDEX
Under 15K	1
15-24K	2
25-34K	3
35-49K	4
50-74K	5
75-99K	6
100-124K	7
125-149K	8
150-174K	9
175-199K	10
200-249K	11
250K+	12

HOUSEHOLD_SIZE_DESC	HOUSEHOLD_SIZE_INDEX
1	1
2	2
3	3
4	4
5+	5

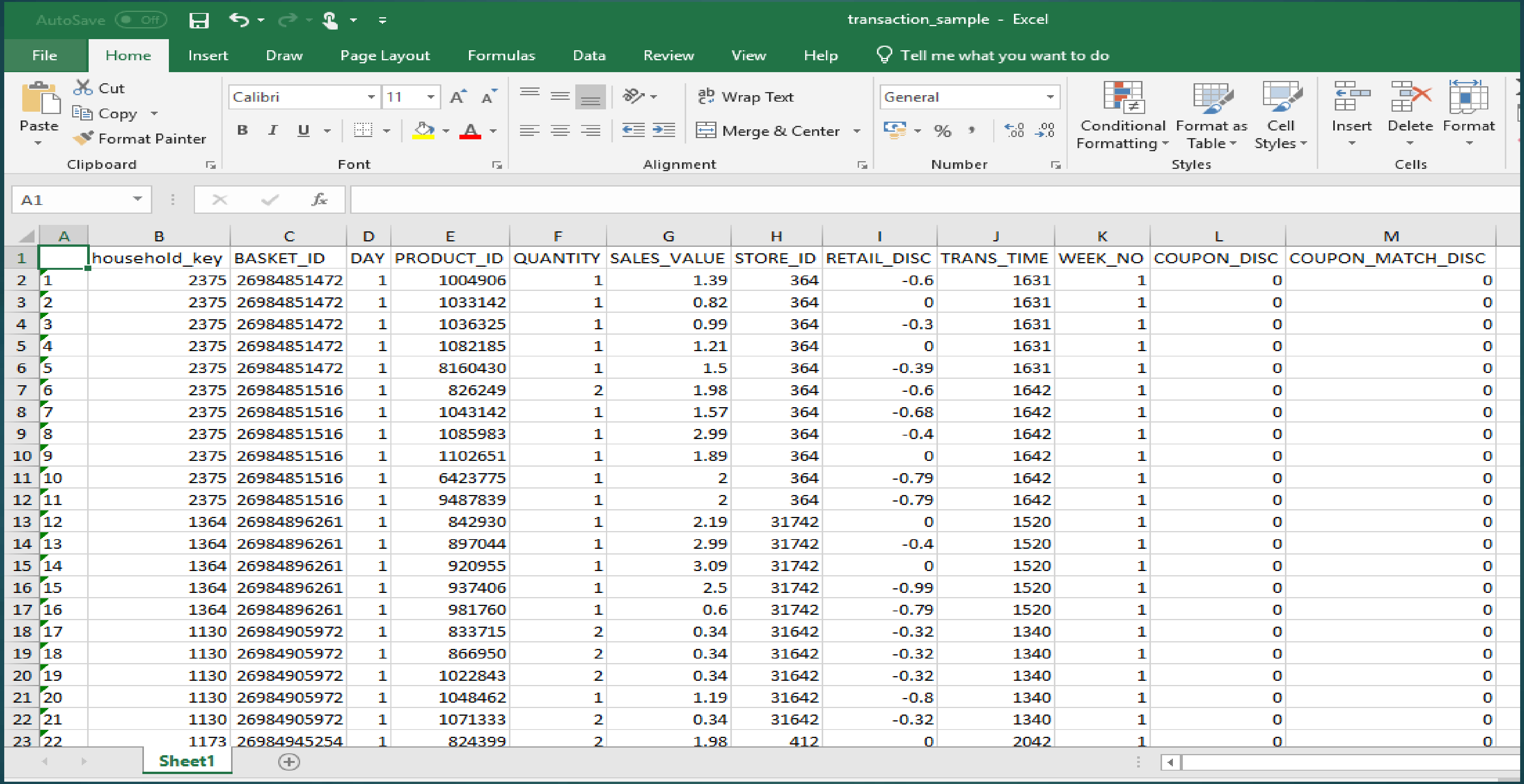
Final Dataset for input to Modelling



Relation: dataset_v3_10_12-weka.filters.unsupervised.attribute.Remove-R2,4,8-13,15-16

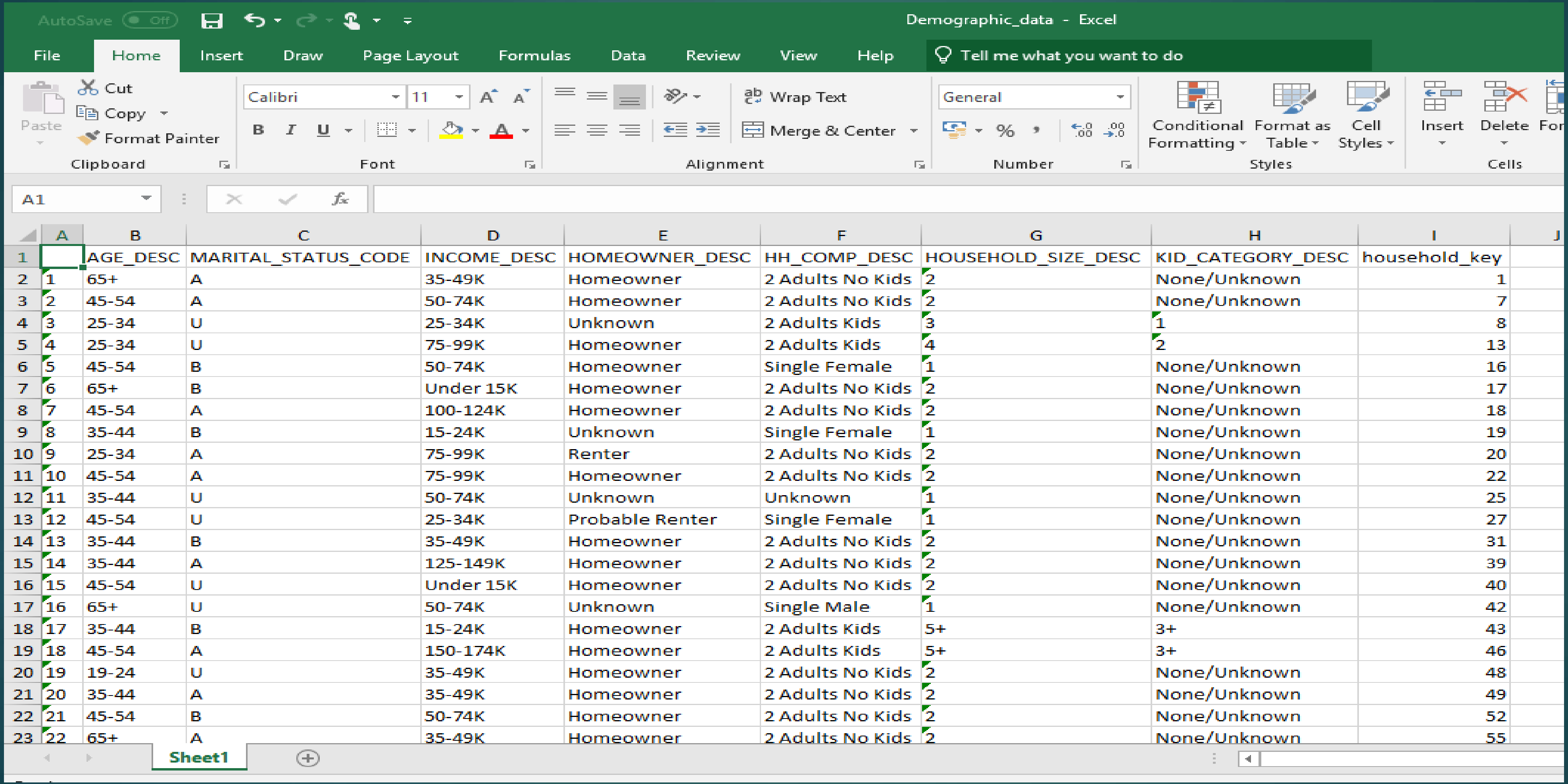
No.	1: household_key	2: products_purchased	3: no_of_orders	4: total_visits	5: no_of_national_products	6: INCOME_INDEX	7: total_sales_class
	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Nominal
1	1.0	1997.0	85.0	78.0	1507.0	4.0	Medium
2	1001.0	1736.0	90.0	79.0	831.0	5.0	Medium
3	1003.0	1055.0	63.0	47.0	486.0	3.0	Low
4	1004.0	1040.0	219.0	192.0	605.0	2.0	Low
5	101.0	2399.0	165.0	133.0	690.0	1.0	Medium
6	1012.0	1058.0	106.0	91.0	637.0	4.0	Low
7	1014.0	1865.0	146.0	133.0	1007.0	2.0	Medium
8	1015.0	2301.0	106.0	102.0	715.0	5.0	Medium
9	1018.0	796.0	160.0	137.0	455.0	4.0	Low
10	1020.0	3759.0	167.0	152.0	2456.0	3.0	High
11	1021.0	876.0	99.0	87.0	488.0	5.0	Low
12	1024.0	4200.0	133.0	114.0	1719.0	1.0	High
13	1031.0	1317.0	116.0	103.0	753.0	6.0	Medium
14	1038.0	1128.0	273.0	220.0	454.0	1.0	Low
15	104.0	3358.0	125.0	113.0	2049.0	4.0	Medium
16	1040.0	2042.0	229.0	223.0	1551.0	7.0	Medium
17	1041.0	3562.0	290.0	186.0	1774.0	4.0	High
18	1042.0	3027.0	231.0	206.0	1432.0	6.0	Medium
19	1045.0	1913.0	224.0	183.0	1125.0	1.0	Medium
20	105.0	2994.0	169.0	130.0	1596.0	1.0	Medium
21	1053.0	1987.0	94.0	94.0	1228.0	7.0	Medium
22	1057.0	3696.0	184.0	175.0	1681.0	5.0	High
23	1060.0	2411.0	62.0	60.0	1732.0	4.0	Medium
24	1062.0	926.0	214.0	195.0	436.0	2.0	Low
25	1066.0	2102.0	245.0	217.0	1052.0	3.0	Medium
26	1069.0	1417.0	55.0	52.0	570.0	4.0	Medium
27	1070.0	1863.0	138.0	123.0	1142.0	5.0	Medium
28	1074.0	3367.0	120.0	111.0	1885.0	4.0	High
29	1076.0	669.0	86.0	80.0	483.0	4.0	Low

Transaction Data



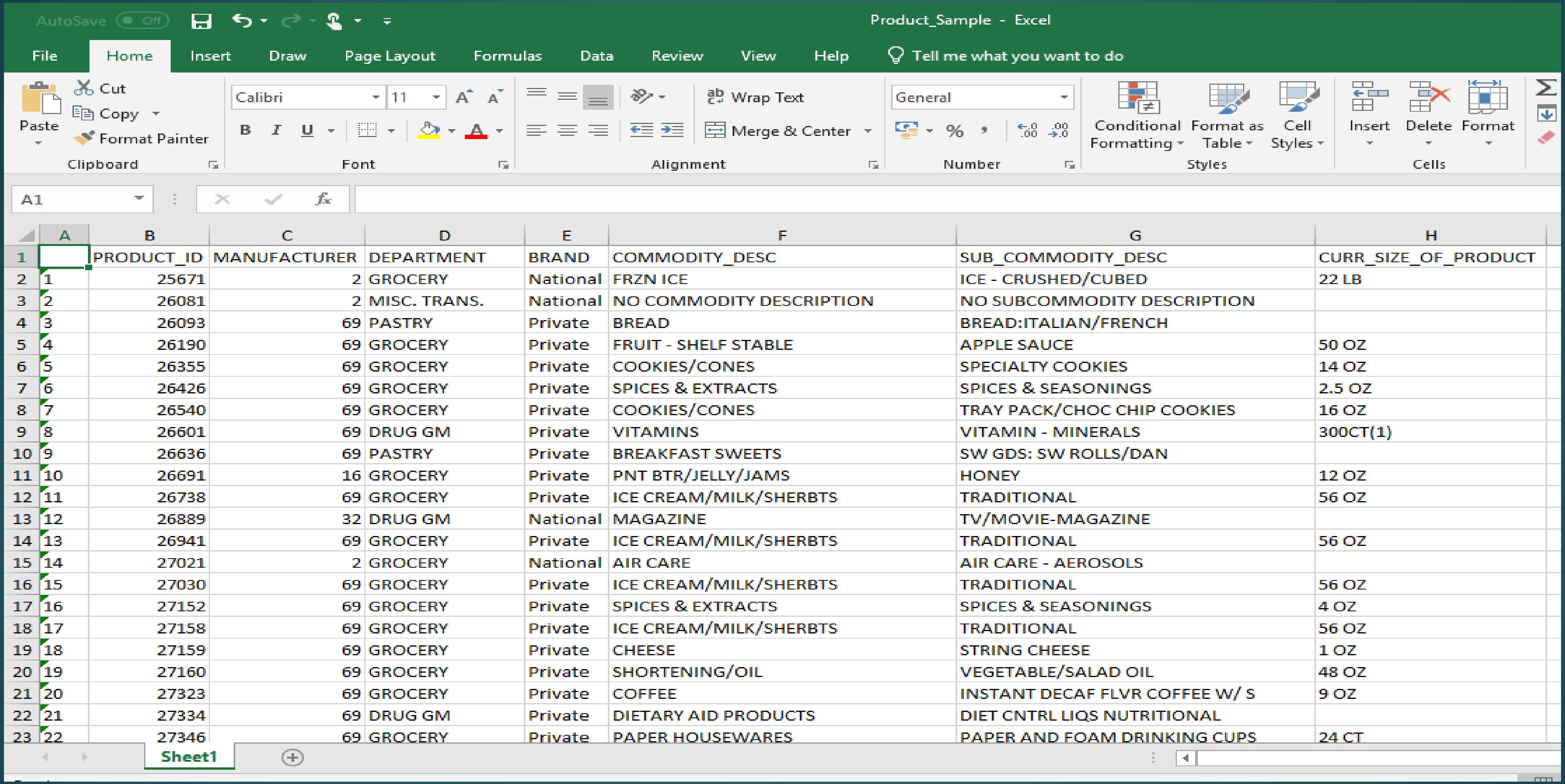
	A	B	C	D	E	F	G	H	I	J	K	L	M
1		household_key	BASKET_ID	DAY	PRODUCT_ID	QUANTITY	SALES_VALUE	STORE_ID	RETAIL_DISC	TRANS_TIME	WEEK_NO	COUPON_DISC	COUPON_MATCH_DISC
2	1	2375	26984851472	1	1004906	1	1.39	364	-0.6	1631	1	0	0
3	2	2375	26984851472	1	1033142	1	0.82	364	0	1631	1	0	0
4	3	2375	26984851472	1	1036325	1	0.99	364	-0.3	1631	1	0	0
5	4	2375	26984851472	1	1082185	1	1.21	364	0	1631	1	0	0
6	5	2375	26984851472	1	8160430	1	1.5	364	-0.39	1631	1	0	0
7	6	2375	26984851516	1	826249	2	1.98	364	-0.6	1642	1	0	0
8	7	2375	26984851516	1	1043142	1	1.57	364	-0.68	1642	1	0	0
9	8	2375	26984851516	1	1085983	1	2.99	364	-0.4	1642	1	0	0
10	9	2375	26984851516	1	1102651	1	1.89	364	0	1642	1	0	0
11	10	2375	26984851516	1	6423775	1	2	364	-0.79	1642	1	0	0
12	11	2375	26984851516	1	9487839	1	2	364	-0.79	1642	1	0	0
13	12	1364	26984896261	1	842930	1	2.19	31742	0	1520	1	0	0
14	13	1364	26984896261	1	897044	1	2.99	31742	-0.4	1520	1	0	0
15	14	1364	26984896261	1	920955	1	3.09	31742	0	1520	1	0	0
16	15	1364	26984896261	1	937406	1	2.5	31742	-0.99	1520	1	0	0
17	16	1364	26984896261	1	981760	1	0.6	31742	-0.79	1520	1	0	0
18	17	1130	26984905972	1	833715	2	0.34	31642	-0.32	1340	1	0	0
19	18	1130	26984905972	1	866950	2	0.34	31642	-0.32	1340	1	0	0
20	19	1130	26984905972	1	1022843	2	0.34	31642	-0.32	1340	1	0	0
21	20	1130	26984905972	1	1048462	1	1.19	31642	-0.8	1340	1	0	0
22	21	1130	26984905972	1	1071333	2	0.34	31642	-0.32	1340	1	0	0
23	22	1173	26984945254	1	824399	2	1.98	412	0	2042	1	0	0

Demographic Data



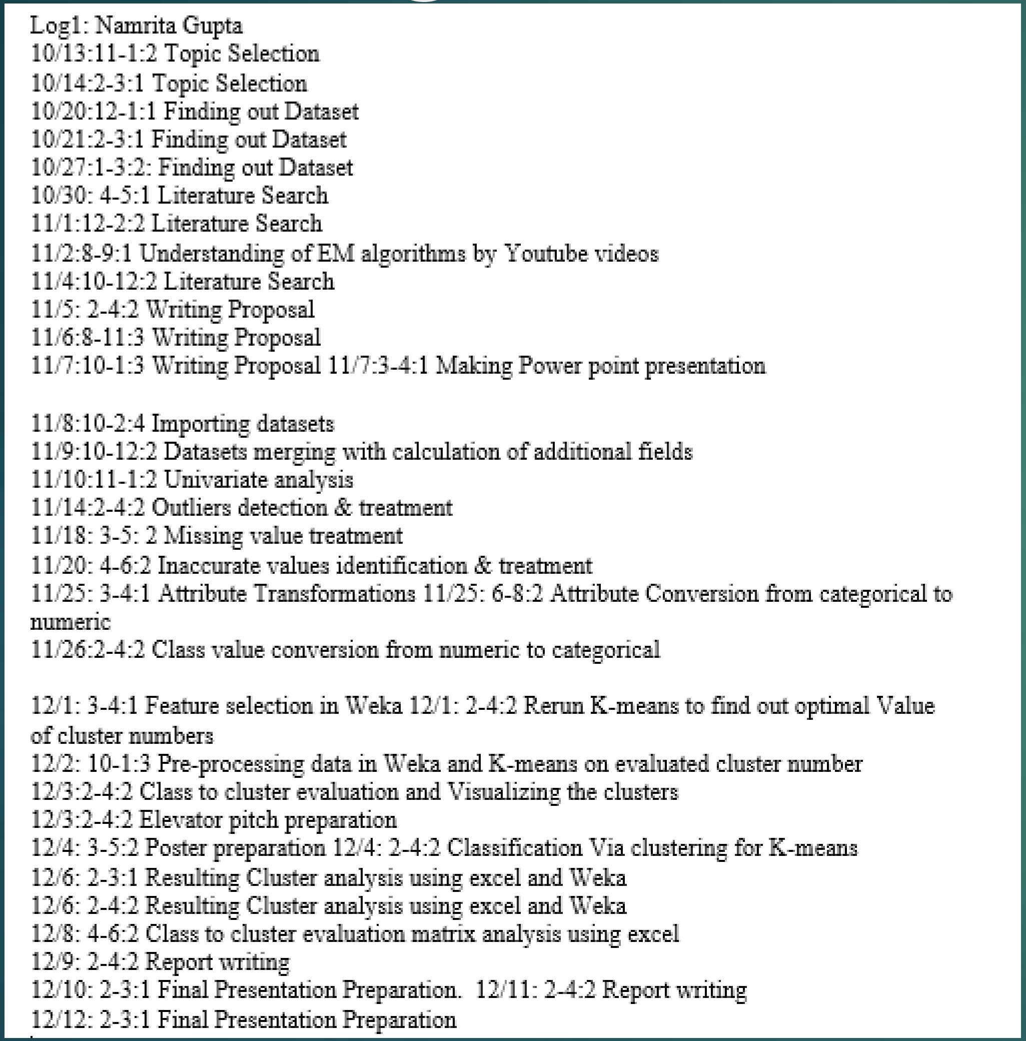
	A	B	C	D	E	F	G	H	I
1		AGE_DESC	MARITAL_STATUS_CODE	INCOME_DESC	HOMEOWNER_DESC	HH_COMP_DESC	HOUSEHOLD_SIZE_DESC	KID_CATEGORY_DESC	household_key
2	1	65+	A	35-49K	Homeowner	2 Adults No Kids	2	None/Unknown	1
3	2	45-54	A	50-74K	Homeowner	2 Adults No Kids	2	None/Unknown	7
4	3	25-34	U	25-34K	Unknown	2 Adults Kids	3	1	8
5	4	25-34	U	75-99K	Homeowner	2 Adults Kids	4	2	13
6	5	45-54	B	50-74K	Homeowner	Single Female	1	None/Unknown	16
7	6	65+	B	Under 15K	Homeowner	2 Adults No Kids	2	None/Unknown	17
8	7	45-54	A	100-124K	Homeowner	2 Adults No Kids	2	None/Unknown	18
9	8	35-44	B	15-24K	Unknown	Single Female	1	None/Unknown	19
10	9	25-34	A	75-99K	Renter	2 Adults No Kids	2	None/Unknown	20
11	10	45-54	A	75-99K	Homeowner	2 Adults No Kids	2	None/Unknown	22
12	11	35-44	U	50-74K	Unknown	Unknown	1	None/Unknown	25
13	12	45-54	U	25-34K	Probable Renter	Single Female	1	None/Unknown	27
14	13	35-44	B	35-49K	Homeowner	2 Adults No Kids	2	None/Unknown	31
15	14	35-44	A	125-149K	Homeowner	2 Adults No Kids	2	None/Unknown	39
16	15	45-54	U	Under 15K	Homeowner	2 Adults No Kids	2	None/Unknown	40
17	16	65+	U	50-74K	Unknown	Single Male	1	None/Unknown	42
18	17	35-44	B	15-24K	Homeowner	2 Adults Kids	5+	3+	43
19	18	45-54	A	150-174K	Homeowner	2 Adults Kids	5+	3+	46
20	19	19-24	U	35-49K	Homeowner	2 Adults No Kids	2	None/Unknown	48
21	20	35-44	A	35-49K	Homeowner	2 Adults No Kids	2	None/Unknown	49
22	21	45-54	B	50-74K	Homeowner	2 Adults No Kids	2	None/Unknown	52
23	22	65+	A	35-49K	Homeowner	2 Adults No Kids	2	None/Unknown	55

Product Data



	A	B	C	D	E	F	G	H
1		PRODUCT_ID	MANUFACTURER	DEPARTMENT	BRAND	COMMODITY_DESC	SUB_COMMODITY_DESC	CURR_SIZE_OF_PRODUCT
2	1	25671	2	GROCERY	National	FRZN ICE	ICE - CRUSHED/CUBED	22 LB
3	2	26081	2	MISC. TRANS.	National	NO COMMODITY DESCRIPTION	NO SUBCOMMODITY DESCRIPTION	
4	3	26093	69	PASTRY	Private	BREAD	BREAD:ITALIAN/FRENCH	
5	4	26190	69	GROCERY	Private	FRUIT - SHELF STABLE	APPLE SAUCE	50 OZ
6	5	26355	69	GROCERY	Private	COOKIES/CONES	SPECIALTY COOKIES	14 OZ
7	6	26426	69	GROCERY	Private	SPICES & EXTRACTS	SPICES & SEASONINGS	2.5 OZ
8	7	26540	69	GROCERY	Private	COOKIES/CONES	TRAY PACK/CHOC CHIP COOKIES	16 OZ
9	8	26601	69	DRUG GM	Private	VITAMINS	VITAMIN - MINERALS	300CT(1)
10	9	26636	69	PASTRY	Private	BREAKFAST SWEETS	SW GDS: SW ROLLS/DAN	
11	10	26691	16	GROCERY	Private	PNT BTR/JELLY/JAMS	HONEY	12 OZ
12	11	26738	69	GROCERY	Private	ICE CREAM/MILK/SHERBTS	TRADITIONAL	56 OZ
13	12	26889	32	DRUG GM	National	MAGAZINE	TV/MOVIE-MAGAZINE	
14	13	26941	69	GROCERY	Private	ICE CREAM/MILK/SHERBTS	TRADITIONAL	56 OZ
15	14	27021	2	GROCERY	National	AIR CARE	AIR CARE - AEROSOLS	
16	15	27030	69	GROCERY	Private	ICE CREAM/MILK/SHERBTS	TRADITIONAL	56 OZ
17	16	27152	69	GROCERY	Private	SPICES & EXTRACTS	SPICES & SEASONINGS	4 OZ
18	17	27158	69	GROCERY	Private	ICE CREAM/MILK/SHERBTS	TRADITIONAL	56 OZ
19	18	27159	69	GROCERY	Private	CHEESE	STRING CHEESE	1 OZ
20	19	27160	69	GROCERY	Private	SHORTENING/OIL	VEGETABLE/SALAD OIL	48 OZ
21	20	27323	69	GROCERY	Private	COFFEE	INSTANT DECAF FLVR COFFEE W/ S	9 OZ
22	21	27334	69	DRUG GM	Private	DIETARY AID PRODUCTS	DIET CNTRL LIQS NUTRITIONAL	
23	22	27346	69	GROCERY	Private	PAPER HOUSEWARES	PAPER AND FOAM DRINKING CUPS	24 CT

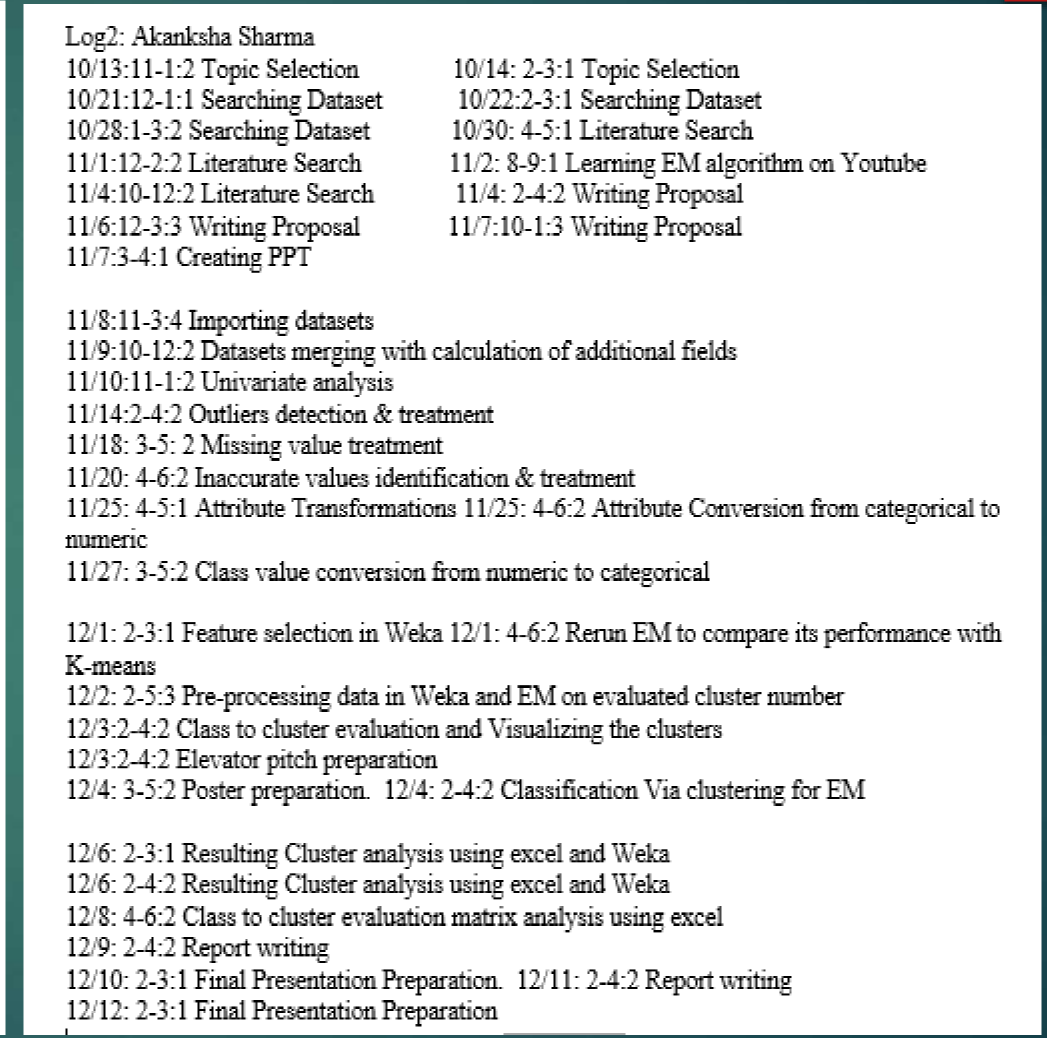
Time Log



Log1: Namrita Gupta
10/13:11-1:2 Topic Selection
10/14:2-3:1 Topic Selection
10/20:12-1:1 Finding out Dataset
10/21:2-3:1 Finding out Dataset
10/27:1-3:2: Finding out Dataset
10/30: 4-5:1 Literature Search
11/1:12-2:2 Literature Search
11/2:8-9:1 Understanding of EM algorithms by Youtube videos
11/4:10-12:2 Literature Search
11/5: 2-4:2 Writing Proposal
11/6:8-11:3 Writing Proposal
11/7:10-1:3 Writing Proposal 11/7:3-4:1 Making Power point presentation

11/8:10-2:4 Importing datasets
11/9:10-12:2 Datasets merging with calculation of additional fields
11/10:11-1:2 Univariate analysis
11/14:2-4:2 Outliers detection & treatment
11/18: 3-5: 2 Missing value treatment
11/20: 4-6:2 Inaccurate values identification & treatment
11/25: 3-4:1 Attribute Transformations 11/25: 6-8:2 Attribute Conversion from categorical to numeric
11/26:2-4:2 Class value conversion from numeric to categorical

12/1: 3-4:1 Feature selection in Weka 12/1: 2-4:2 Rerun K-means to find out optimal Value of cluster numbers
12/2: 10-1:3 Pre-processing data in Weka and K-means on evaluated cluster number
12/3:2-4:2 Class to cluster evaluation and Visualizing the clusters
12/3:2-4:2 Elevator pitch preparation
12/4: 3-5:2 Poster preparation 12/4: 2-4:2 Classification Via clustering for K-means
12/6: 2-3:1 Resulting Cluster analysis using excel and Weka
12/6: 2-4:2 Resulting Cluster analysis using excel and Weka
12/8: 4-6:2 Class to cluster evaluation matrix analysis using excel
12/9: 2-4:2 Report writing
12/10: 2-3:1 Final Presentation Preparation. 12/11: 2-4:2 Report writing
12/12: 2-3:1 Final Presentation Preparation



Log2: Akanksha Sharma
10/13:11-1:2 Topic Selection 10/14: 2-3:1 Topic Selection
10/21:12-1:1 Searching Dataset 10/22:2-3:1 Searching Dataset
10/28:1-3:2 Searching Dataset 10/30: 4-5:1 Literature Search
11/1:12-2:2 Literature Search 11/2: 8-9:1 Learning EM algorithm on Youtube
11/4:10-12:2 Literature Search 11/4: 2-4:2 Writing Proposal
11/6:12-3:3 Writing Proposal 11/7:10-1:3 Writing Proposal
11/7:3-4:1 Creating PPT

11/8:11-3:4 Importing datasets
11/9:10-12:2 Datasets merging with calculation of additional fields
11/10:11-1:2 Univariate analysis
11/14:2-4:2 Outliers detection & treatment
11/18: 3-5: 2 Missing value treatment
11/20: 4-6:2 Inaccurate values identification & treatment
11/25: 4-5:1 Attribute Transformations 11/25: 4-6:2 Attribute Conversion from categorical to numeric
11/27: 3-5:2 Class value conversion from numeric to categorical

12/1: 2-3:1 Feature selection in Weka 12/1: 4-6:2 Rerun EM to compare its performance with K-means
12/2: 2-5:3 Pre-processing data in Weka and EM on evaluated cluster number
12/3:2-4:2 Class to cluster evaluation and Visualizing the clusters
12/3:2-4:2 Elevator pitch preparation
12/4: 3-5:2 Poster preparation. 12/4: 2-4:2 Classification Via clustering for EM

12/6: 2-3:1 Resulting Cluster analysis using excel and Weka
12/6: 2-4:2 Resulting Cluster analysis using excel and Weka
12/8: 4-6:2 Class to cluster evaluation matrix analysis using excel
12/9: 2-4:2 Report writing
12/10: 2-3:1 Final Presentation Preparation. 12/11: 2-4:2 Report writing
12/12: 2-3:1 Final Presentation Preparation