BANA 7047 Individual Case 2

Honor Policy: Individual cases should be treated as a take-home exam, which should be the sole work of each student. You can only discuss with the instructor if you have any question.

Last Name _____**Valles**_____

First Name _____**Adrian**_____

M#_____**M07593734**_____

Please use your M# to set the seed to draw a random sample.

Signature _____**Adrian**_____

European employment data

Background and Goal

The data are the percentage employed in different industries in Europe countries during 1979. The purpose of examining this data is to get insight into patterns of employment (if any) amongst European countries in 1970s.

Approach

To ensure that the we find patterns of employment amongst European countries, we will use 2 different clustering methods. Using R-Studio, we will try different number of clusters using k-means clustering as well as hierarchical clustering to determine which method and what number of cluster, better identify the patterns.

Major findings

After comparing the 2 methods using different number of clusters, we can conclude that both methods perform almost identically and that it makes more sense to apply 2 or 3 clusters.

As seen in **table A**, applying 2 clusters leads a distribution of a group of countries that heavily rely on the agricultural sector (group 1) and another group of countries who heavily rely on manufacturing and services industries (group 2). Based on the name of the countries group 2 countries are more developed countries than the ones in group 1

Group	Agr	Min	Man	PS	Con	SI	Fin	SPS	TC
1	44.5	1.5	19.6	0.7	6.6	7.3	3.4	11.2	5.2
2	13.1	1.2	28.8	1	8.5	14.3	4.1	22.1	6.9

Group 1						
Greece	Turkey	Poland	Rumania	Yugoslavia		
Group 2						
Belgium	Denmark France	WGermany	Ireland	Italy	Luxembourg	Netherlands
UK	Austria Finland	Norway	Portugal	Spain	Sweden	Switzerland
Bulgaria	Czechoslovakia EGermany	Hungary	USSR			

Table A – Summary of the European employment data given 2 clusters

When a third cluster is added, a new group emerges as a group of countries that has a mixed model. This means they rely on both agriculture and manufacturing/services. This can be seen in **table B**.

Group	Agr	Min	Man	PS	Con	SI	Fin	SPS	TC
1	52.3	0.9	14.1	0.6	5.3	7.7	4.9	9.4	4.6
2	8.2	1	29.1	1	8.4	16.3	5	24.2	6.9
3	25	1.8	28.1	0.9	8.7	9.6	2.1	17.1	6.7

Table B - Summary of the European employment data given 3 clusters

It is important to remark that both k-means and hierarchical clustering obtained the same grouping for 3 clusters. For a more in-depth analysis, please go to the full report.

Cincinnati Zoo data

Background and Goal

The data is composed by 2 datasets related to the Cincinnati zoo food consumption. In the first one we have the amount of each type of food bought at the Cincinnati zoo during 6 days. The second dataset contains the products bought by the zoo visitors between 2010 and 2011. The purpose of examining this data is to get insight into patterns of food purchases amongst Cincinnati zoo visitors during 2010 and 2011.

Approach

To ensure that the we find patterns of food purchases amongst the Cincinnati zoo visitors, we will use a clustering method and association rules. Using R-Studio, we will try different number of clusters using k-means clustering as well as association rules in order to make a grouping of the products as well as to find what food products are related with each other.

Major findings

By applying a 2-clusters grouping using the k-means method, we can say divide the food products into popular (group 1) and unpopular products (group 2), as seen in **Table C.** Group 1 contains 8 products and group 2 contains 47 products.

Group.1	Oct,10	Nov,10	Dec,10	Jan,11	Feb,11	Mar,11
1	1342.125	462.75	522.375	81.125	182.25	548.375
2	205.7234	92.53191	131.2979	20.12766	32.80851	82.85106

Group 1	Group 2
Bottled Water	Cheese
IceCreamCone	Alchohol
Medium Drink	Burger
Small Drink	Capri Sun
Snack	Cheese Fries Basket
Souvenir Drink	Cheeseburger Basket
Whole/Slice Cheese	Chicken Nugget Basket
Whole/Slice Pepp	Chicken Tender Basket

Table C- Summary from the k means analysis with 2 clusters

For Association rules, we use the apriori algorithm using a minimum support of 0.3% and a minimum confidence value of 50%. We obtain 40 different rules, 11 of which have a size of 2, 26 rules with a size of 3 and 3 rules with a size of 4. **Table B**, shows the rules that have a support value larger than 0.01.

lhs		rhs	support	confidence	lift	count
{Hot Chocolate Souvenir RefillFood}	=>	{Hot Chocolate SouvenirFood}	0.01499266	0.5596869	13.18097	286
{ToppingFood}	=>	{Ice Cream ConeFood}	0.02856993	0.9981685	8.947868	545
{Chicken TendersFood}	=>	{French Fries BasketFood}	0.01729922	0.7586207	7.771992	330
{CheeseburgerFood}	=>	{French Fries BasketFood}	0.01687985	0.7931034	8.125264	322
{GatoradeFood,Slice of PeppFood}	=>	{Slice of CheeseFood}	0.01011743	0.5830816	3.620724	193
{Medium DrinkFood,Slice of PeppFood}	=>	{Slice of CheeseFood}	0.01362969	0.5273834	3.274858	260
{Bottled WaterFood,Slice of PeppFood}	=>	{Slice of CheeseFood}	0.01069407	0.5151515	3.198903	204

Table B- Association rules with support higher than 0.01

For a more in-depth analysis, please go to the full report.

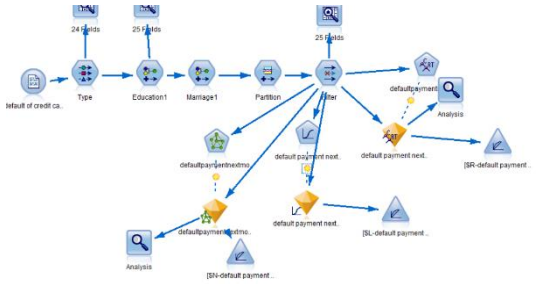
Credit card default data (SPSS)

Background and Goal

Predicting credit card payment default is critical for the successful business model of a credit card company. In this project, we showcase our solution by using a public dataset which contains 30,000 credit card accounts. Among the total 30,000 accounts, 6636 accounts (22%) are cardholders with default payments. The response variable of this study is default payment (Yes=1, No=0). Our goal is to build an accurate classifier to predict if a credit card account will default or not.

Approach

To ensure that the model predicts well for data not used to build the model, we use model validation. We will build different models (e.g., GLM, CART and neural network) using SPSS, compare the performance of these models, and select the best-performing model based on the area under the curve.

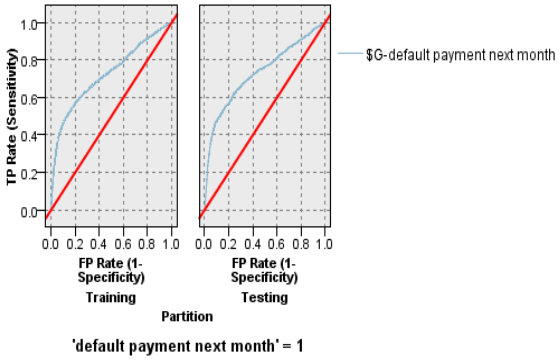


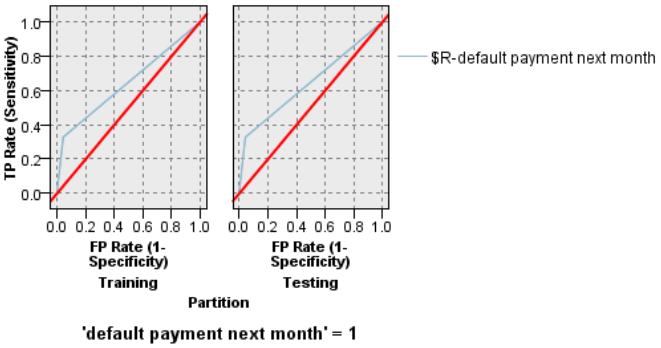
Major findings

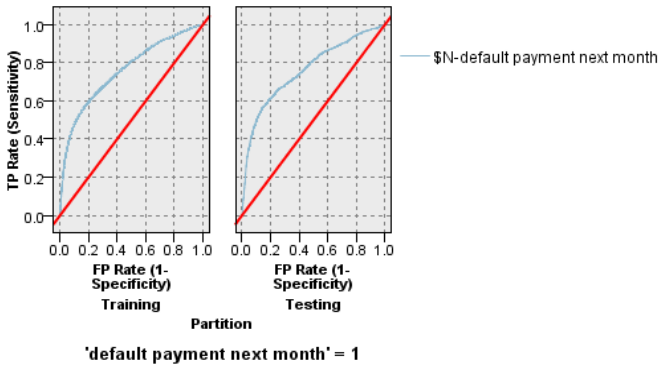
After comparing the performance of the 3 different approaches, we can conclude that Neural network reports the lowest AUC followed by GLM and CART, as it can be seen in the **tables**.

Tables: Area under the curve summary (top left) , ROC curve of GLM (top right), ROC curve of CART(bottom left) and ROC curve of Neural network (bottom right)

	AUC	
Model	In sample	Out of sample
GLM	0.72	0.73
Decision tree	0.64	0.64
Neural network	0.76	0.76







European employment data full report

Before starting with the clustering stage, let's provide a summary of the employment in the different sectors. As seen in **Figure 1,** manufacturing is the field that employs, on average, a larger percentage of people in Europe

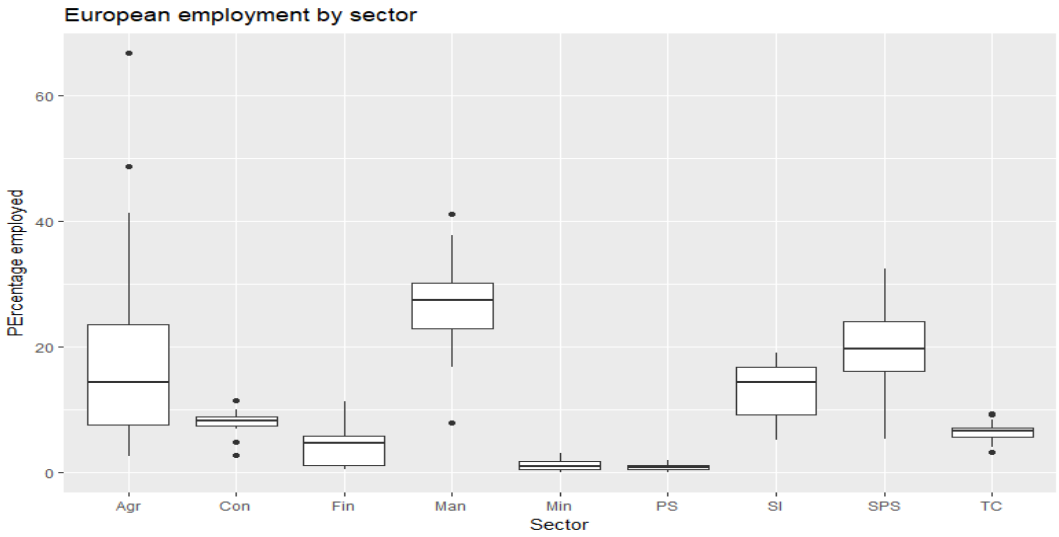


Figure 1-Percentage of employment in Europe by sector

Now that we understand slightly better the type of data we are dealing with, we can start with the clustering stage. We will explore K-means and hierarchical clustering but given that this dataset is composed only of 27 observations we expect hierarchical clustering to perform better.

K-means

For k-means the first thing we want to know is how many clusters we want to split the data on. On this case, we will determine the number of cluster based on the sum of squares method. As it can be seen in **Figure 2**, different number of clusters are associated with a different within groups sum of squares. Logically, as the number of clusters goes up, sum of squares goes down.

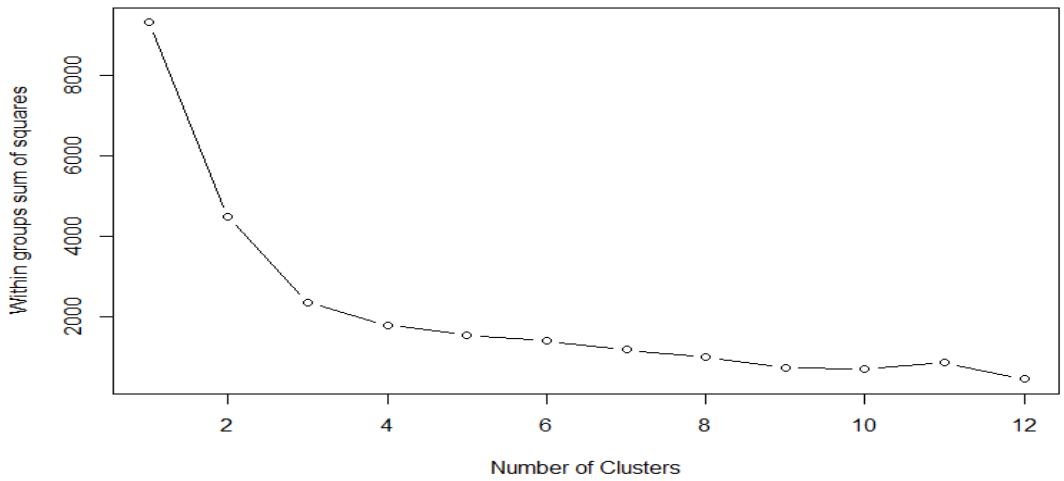


Figure 2 – Number of clusters vs within groups sum of squares

We will start by selecting 3 clusters given that this value represents the "elbow" of the graph, given that the improvement in sum of squares for more than 3 clusters is slower.

When we split the data into 3 clusters taking into consideration all the sectors, we obtain the following distribution, as shown in **Figure 3**

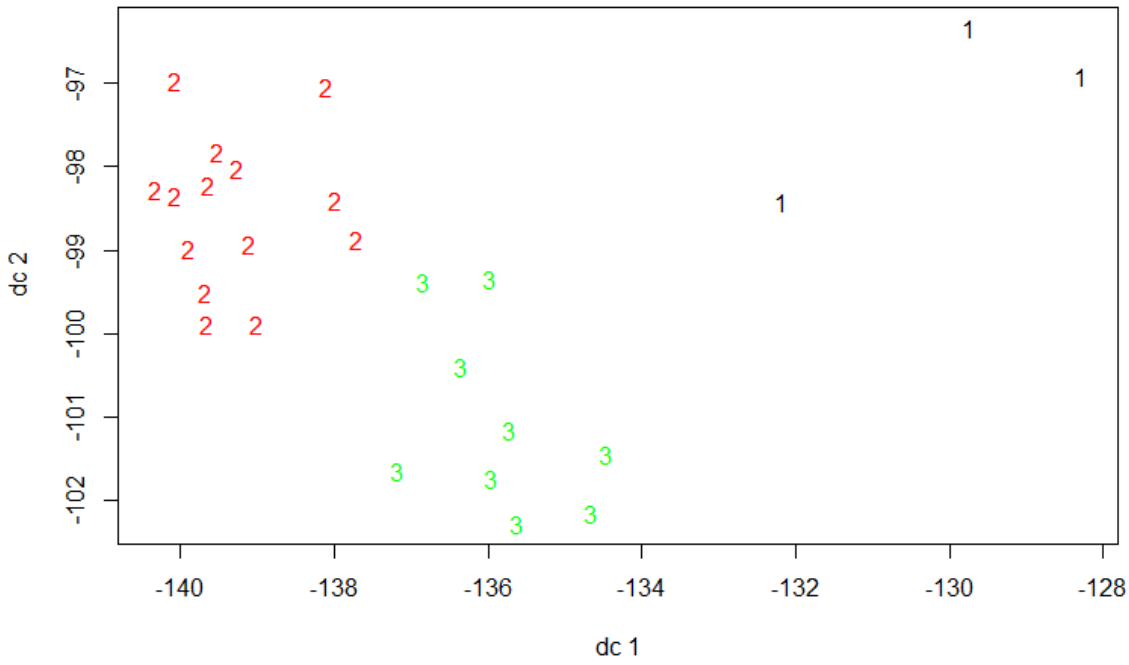


Figure 3 – Groups distribution

Where the countries in each group are the following:

Group 1						
Greece	Turkey	Yugoslavia				
Group 2						
Belgium	Denmark	France	WGermany	Italy	Luxembourg	Netherlands
UK	Austria	Finland	Norway	Sweden	Switzerland	EGermany
Group 3						
Ireland	Portugal	Spain	Bulgaria	Czechoslovakia		
Hungary	Poland	Rumania	USSR			

Table 1- Countries by group

We can also provide a summary statistic by group, as seen in **table 2**

Group	Agr	Min	Man	PS	Con	SI	Fin	SPS	TC
1	52.3	0.9	14.1	0.6	5.3	7.7	4.9	9.4	4.6
2	8.2	1	29.1	1	8.4	16.3	5	24.2	6.9
3	25	1.8	28.1	0.9	8.7	9.6	2.1	17.1	6.7

Table 2 – Summary statistic by group

Therefore, bases on this we can already get the next insights:

- Group 1 relies heavily on Agriculture. It makes sense as Greece, Turkey and Yugoslavia have a Mediterranean weather that benefit this type of industry.
- Group 2 are central European countries, whose industry heavily on manufacturing and services.

- Group 3 combines similarities with both groups 1 and 2 given that while, they don't rely heavily on any industry as Group 1 and 2 do, they have high percentages in Agriculture, manufacturing and services.

Now that we have been able to understand the distribution of the European countries, we can see that we are grouping based on agriculture reliance vs manufacturing and services reliance. We are therefore going to do a k-means with 2 clusters to see if we can better capture this. The summary of the k-means given 2 clusters can be seen below in **Table 3**

Group 1						
Greece	Turkey	Poland	Rumania	Yugoslavia		
Group 2						
Belgium	Denmark France	WGermany	Ireland	Italy	Luxembourg	Netherlands
UK	Austria Finland	Norway	Portugal	Spain	Sweden	Switzerland
Bulgaria	Czechoslovakia EGermany	Hungary	USSR			

Group	Agr	Min	Man	PS	Con	SI	Fin	SPS	TC
1	44.5	1.5	19.6	0.7	6.6	7.3	3.4	11.2	5.2
2	13.1	1.2	28.8	1	8.5	14.3	4.1	22.1	6.9

Table 3-Summary of the k-means given 2 clusters

As we predicted, the biggest difference between the groups lies on the high reliance of group 1 in Agriculture, and the high reliance of group 2 in manufacturing and services. It is interesting to observe that group 1 countries are developing and rising countries while the ones in group 2 are developed and "rich" countries for the most part. In this case both a 2 and 3 clusters split seem to be appropriate.

Hierarchical clustering

The next step is to explore the hierarchical clustering. After we have defined the distance matrix we can produce the Cluster Dendrogram using Ward's method, as seen in **Figure 4.**

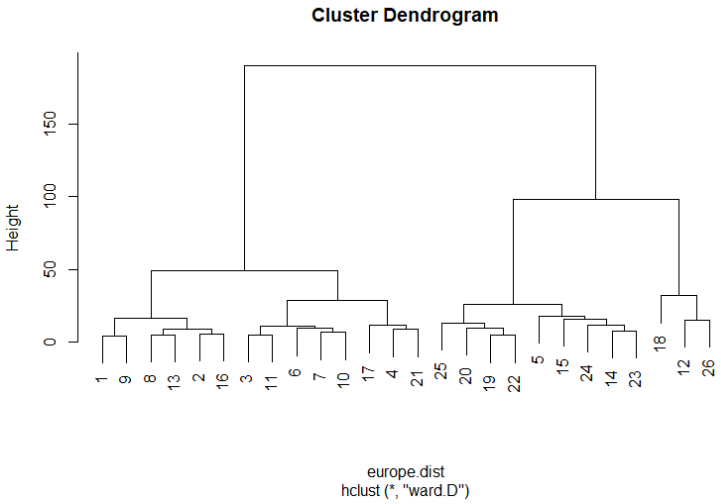


Figure 4 – Cluster Dendogram

We can now cut the dendrogram at a certain clusters level and obtain cluster membership. We will first cut the dendrogram at 3 clusters to see if we get similar results than K-means. Surprisingly, we obtain exactly the same grouping than with K-means, as seen in **Table 4.**

Group	Agr	Min	Man	PS	Con	SI	Fin	SPS	TC
1	52.3	0.9	14.1	0.6	5.3	7.7	4.9	9.4	4.6
2	8.2	1	29.1	1	8.4	16.3	5	24.2	6.9
3	25	1.8	28.1	0.9	8.7	9.6	2.1	17.1	6.7

Group 1						
Greece	Turkey	Yugoslavia				
Group 2						
Belgium	Denmark	France	WGermany	Italy	Luxembourg	Netherlands
UK	Austria	Finland	Norway	Sweden	Switzerland	EGermany
Group 3						
Ireland	Portugal	Spain	Bulgaria	Czechoslovakia		
Hungary	Poland	Rumania	USSR			

Table 3-summary of the Hierarchical clustering with 3 groups

Given that we obtain similar results, we have decided to study one different alternative. In this case, we will cut the dendrogram in 4 clusters. **Table 4** summarizes the 4-cluster distribution.

Group	Agr	Min	Man	PS	Con	SI	Fin	SPS	TC
1	6.1	0.57	25	0.9	8.2	16.7	6	29.3	7.3
2	9.8	1.3	32.1	1	8.6	16	4.3	20.3	6.6
3	25	1.8	28	0.9	8.7	9.6	2.1	17.1	6.7
4	52.3	0.9	14.1	0.6	5.3	7.7	4.9	9.4	4.6

Group 1								
Belgiur	Denmark	Netherland	UK	Norway	Sweden			
Group 2								
France	WGermar	Italy	Luxembourg	Austria	Finland	Switzerland	EGermany	
Group 3								
Irelanc	Portugal	Spain	Bulgaria	Czechoslovakia	Hungary	Poland	Rumania	USSR
Group 4								
Greece	Turkey	Yugoslavia						

Table 4 – 4-cluster distribution

From the 4-cluster distribution we can get the following insights:

- Group 1 relies heavily on services
- Group 2 relies heavily on manufacturing and services
- Group 3 covers diverse industries and relies in both agriculture, and services /manufacturing
- Group 4 relies heavily on agriculture

In my opinion when we split the clusters into 2 and 3 is when we get the better results. It is important to remark as well that in this case, k means and hierarchical clustering perform almost identical.

Cincinnati zoo data full report

Clustering

We will start our analysis of the Cincinnati zoo data with clustering before we jump into the association rules. For this, we will use "qry_Food_by_Month.xls" which is a dataset containing 55 rows and 7 columns. Below, in table 4, we can see the first 10 observations of the dataset.

	NickName	Oct, 10	Nov, 10	Dec, 10	Jan, 11	Feb, 11	Mar, 11
1	Cheese	343	66	99	37	4	105
2	Alchohol	131	79	232	12	18	49
3	Bottled Water	1448	410	577	59	165	507
4	Burger	188	86	103	19	40	73
5	Capri Sun	32	2	0	0	1	0
6	Cheese Fries Basket	37	55	59	3	33	65
7	Cheeseburger Basket	662	274	292	51	93	266
8	Chicken Nugget Basket	529	106	15	0	0	72
9	Chicken Tender Basket	395	299	298	61	132	298
10	Chili Cheese Sandwich	74	6	3	7	16	14

Table 5 – Head of the "qry_Food_by_Month.xls" dataset.

K-MEANS CLUSTERING

When using k-means, the clustering analysis starts by determining the optimal number of clusters that we need. **Figure 5**, shows the within groups sum square associated with each number of clusters.

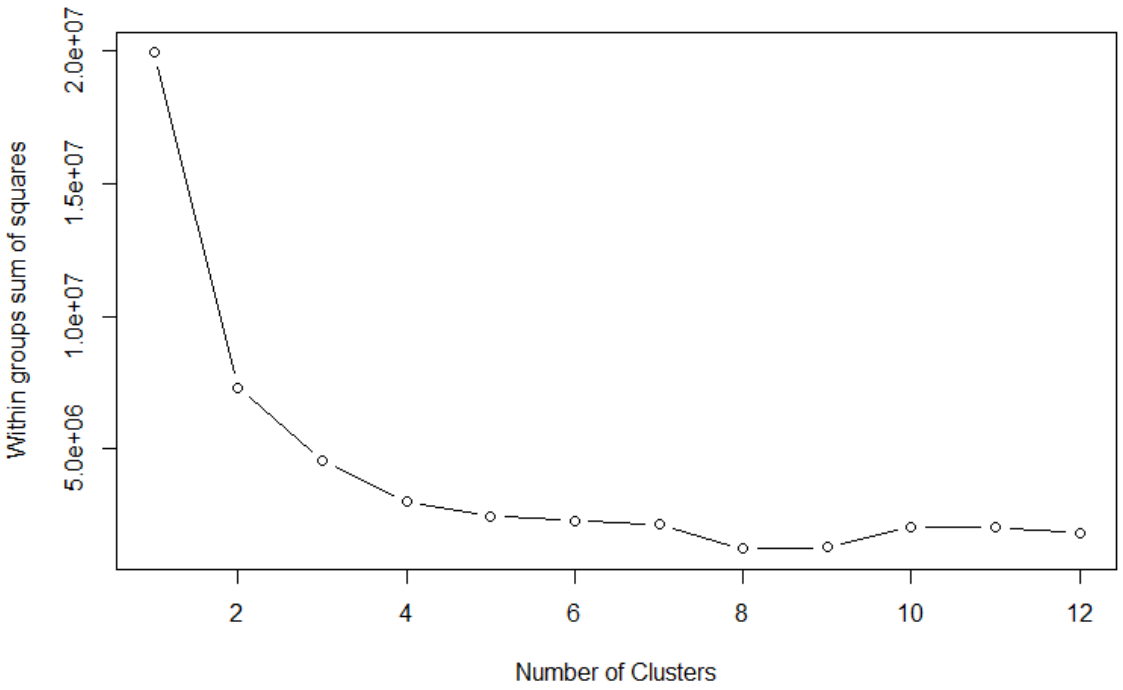


Figure 5- Number of clusters vs within group sum of squares plot

Based on the elbow of the curve, it seems that a cluster between 2, 3 and 4 will work the best.

We will continue our cluster analysis by using k-means on 2 clusters. **Figure 6** shows the distribution split for 2 clusters.

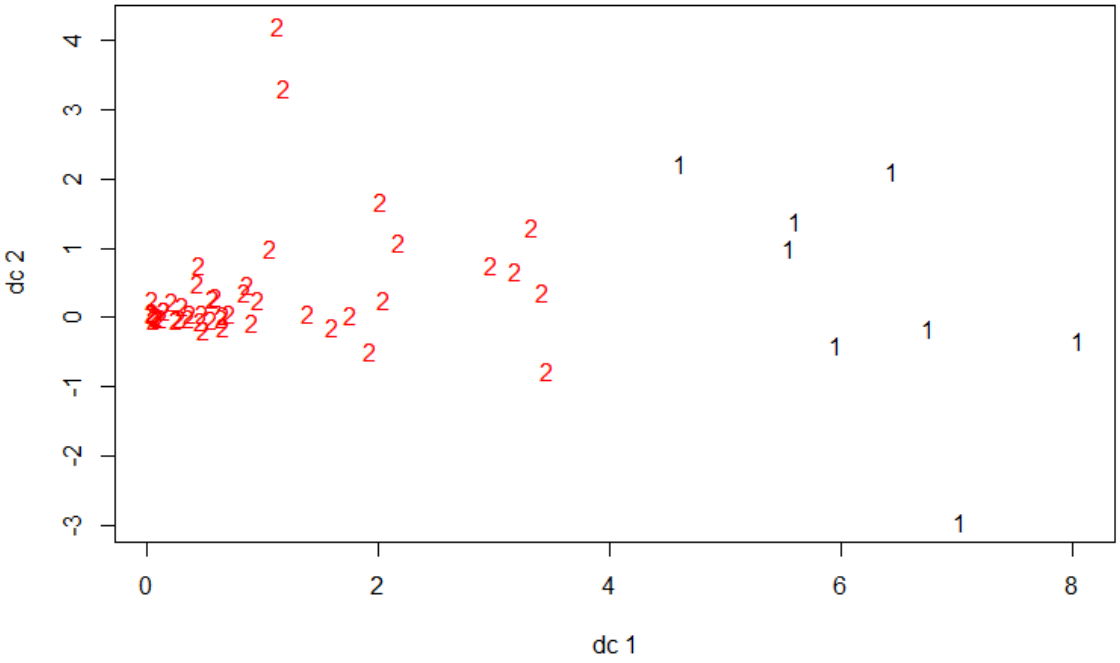


Figure 6 – Distribution split for 2 clusters using k-means

The result of this split is that we have 8 products in group 1 and 47 products in group 2. But how do these groups differ?

As we can see in **table 6,** group 1 products contain those products that have been sold quite a lot during the 6 days and are therefore popular products. On the other side, products from group 2 are less popular products. All the products belonging to group 1 can be seen below as well as some of the products from group 2.

Group.1	Oct,10	Nov,10	Dec,10	Jan,11	Feb,11	Mar,11
1	1342.125	462.75	522.375	81.125	182.25	548.375
2	205.7234	92.53191	131.2979	20.12766	32.80851	82.85106

Group 1	Group 2
Bottled Water	Cheese
IceCreamCone	Alchohol
Medium Drink	Burger
Small Drink	Capri Sun
Snack	Cheese Fries Basket
Souvenir Drink	Cheeseburger Basket
Whole/Slice Cheese	Chicken Nugget Basket
Whole/Slice Pepp	Chicken Tender Basket

Table 6 -Summary from the k means analysis with 2 clusters

Next, we will study k means clustering using 3 clusters. **Figure 7** shows the 3 different groups.

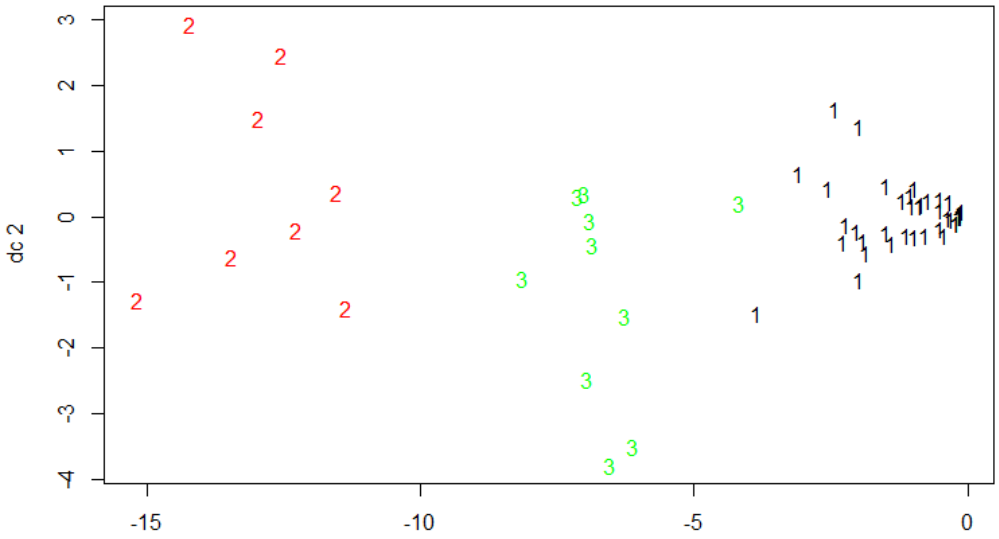


Figure 7 – Distribution split for 3 clusters using k-means

The result of this split is that we have 37 products in group 1 and 8 products in group 2 and 10 products in group 3. But how do these groups differ?

As we can see in **table 7,** group 1 products contain those products that have been sold quite a lot during the 6 days and are therefore popular products (same group 1 as with 2 clusters). On the other side, products from group 2 are less popular products. The difference in this case is that there is a group 3 of products that are semi-popular. All the products belonging to group 1 and 3 can be seen below as well as some of the products from group 2.

Group.1	Oct,10	Nov,10	Dec,10	Jan,11	Feb,11	Mar,11
1	1342.1	462.8	522.4	81.1	182.2	548.4
2	136.9	44.3	45.9	7.8	13.5	40.4
3	460.3	270.9	447.1	65.6	104.1	240.1

Group 1
Bottled Water
IceCreamCone
Medium Drink
Small Drink
Snack
Souvenir Drink
Whole/Slice Cheese
Whole/Slice Pepp

Group 2
Cheese
Alchohol
Burger
Capri Sun
Cheese Fries Basket
Chicken Nugget Basket
Chili Cheese Sandwich
Coffee/HotTea
Coney
Diet Dr Pepper

Group 3
Cheeseburger Basket
Chicken Tender Basket
Chips
French Fries Basket
Gatorade
Hot Chocolate
Hot Chocolate Souvenir
Hot Dog Basket
Krazy Kritter
Soft Pretzel

Table 7 -Summary from the k means analysis with 3 clusters

We could do k-means with 4 clusters, but the only addition would be a new group of products. Let's therefore skip to the next topic on this dataset: Discriminant analysis.

Associations rules

Before starting the analysis, let's take a look at some of the most popular items. **Figure 8** shows the products that are in more than 10% of the transaction of the Cincinnati zoo.

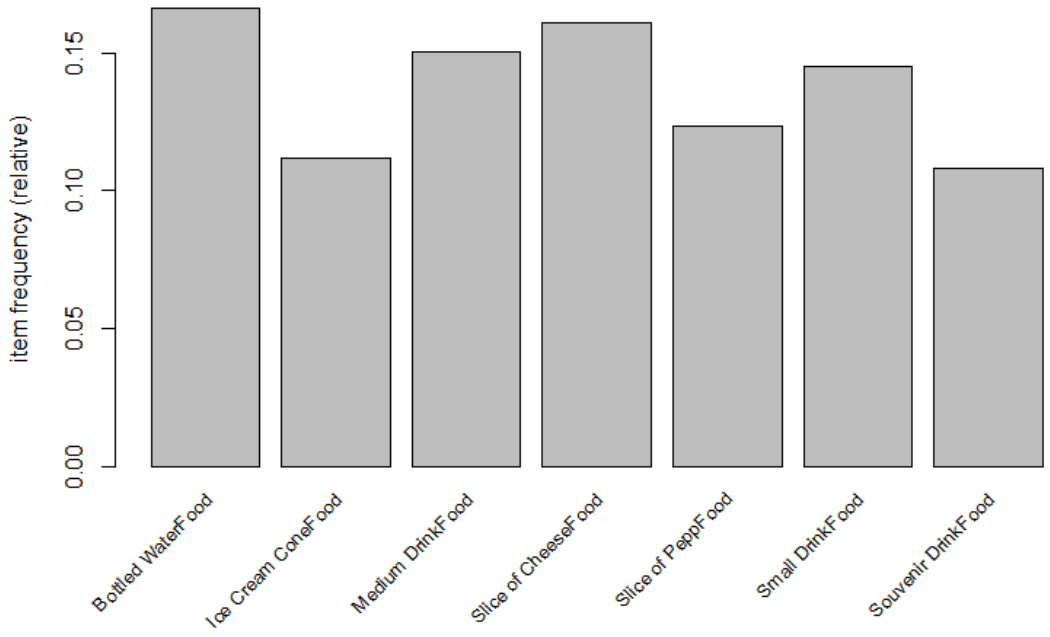


Figure 8- Products with a support value higher than 0.1

Now that we have an initial understanding of some of the products, we can start creating association rules. We will create these through the apriori algorithm using a minimum support of 0.3% and a minimum confidence value of 50%, which are the default values.

By using those values for support and confidence we obtain 40 different rules. 11 rules with a size of 2, 26 rules with a size of 3 and 3 rules with a size of 4. A summary statistics of these 40 rules can be seen below in **table 8.**

support	confidence	lift	count
Min. :0.003093	Min. :0.5032	Min. : 3.125	Min. : 59
1st Qu.:0.003670	1st Qu.:0.5902	1st Qu.: 5.766	1st Qu.: 70
Median :0.004561	Median :0.7586	Median : 8.235	Median : 87
Mean :0.006604	Mean :0.7399	Mean : 8.966	Mean :126
3rd Qu.:0.007182	3rd Qu.:0.8737	3rd Qu.: 9.107	3rd Qu.:137
Max. :0.028570	Max. :1.0000	Max. :26.179	Max. :545

Table 8-Summary statistics of the 40 association rules

We can also see in **table 9** some of the association rules found with size 2

lhs		rhs	support	confidence
{Small Pink LemonadeFood}	=>	{Chicken Nugget BasketFood}	0.003355001	0.5925926
{Grilled Chicken SandwichFood}	=>	{French Fries BasketFood}	0.003721954	0.6698113
{FloatFood}	=>	{Ice Cream ConeFood}	0.007024533	0.7089947
{Side of CheeseFood}	=>	{Cheese ConeyFood}	0.004665548	0.6846154
{Side of CheeseFood}	=>	{Hot DogFood}	0.006290627	0.9230769
{BurgerFood}	=>	{French Fries BasketFood}	0.004613126	0.6616541

Table 9-Some of the 40 association rules with size 2

Let's now look in **table 10** at the association rules with size 4.

lhs		rhs	support	confidence	lift	count
{Krazy KritterFood,						
Medium DrinkFood,						
Slice of PeppFood}	=>	{Slice of chheseFood	0.003250157	0.5535714	3.437477	62
{Medium DrinkFood,						
Slice of PeppFood,						
Small DrinkFood}	=>	{Slice of chheseFood	0.003145313	0.6	3.725781	60
{Medium DrinkFood,						
Slice of CheeseFood,						
Small DrinkFood}	=>	{Slice of PrppFood	0.003145313	0.5172414	4.191545	60

Table 10-All three association rules with size 4

Finally, let's take a look to the association rules with a support higher than 0.01 (table 11), confidence higher than 0.9 (table 11) and lift higher than 20 (table 12).

lhs		rhs	support	confidence	lift	count
{Hot Chocolate Souvenir RefillFood}	=>	{Hot Chocolate SouvenirFood}	0.01499266	0.5596869	13.18097	286
{ToppingFood}	=>	{Ice Cream ConeFood}	0.02856993	0.9981685	8.947868	545
{Chicken TendersFood}	=>	{French Fries BasketFood}	0.01729922	0.7586207	7.771992	330
{CheeseburgerFood}	=>	{French Fries BasketFood}	0.01687985	0.7931034	8.125264	322
{GatoradeFood,Slice of PeppFood}	=>	{Slice of CheeseFood}	0.01011743	0.5830816	3.620724	193
{Medium DrinkFood,Slice of PeppFood}	=>	{Slice of CheeseFood}	0.01362969	0.5273834	3.274858	260
{Bottled WaterFood,Slice of PeppFood}	=>	{Slice of CheeseFood}	0.01069407	0.5151515	3.198903	204

Table 11- Association rules with support higher than 0.01

lhs		rhs	support	confidence	lift	count
{Side of CheeseFood}	=>	{Hot DogFood}	0.006290627	0.9230769	21.60566	120
{ToppingFood}	=>	{Ice Cream ConeFood}	0.028569931	0.9981685	8.947868	545
{Cheese ConeyFood,Side of CheeseFood}	=>	{Hot DogFood}	0.004351017	0.9325843	21.82819	83
{Bottled WaterFood,ToppingFood}	=>	{Ice Cream ConeFood}	0.004036486	1	8.964286	77
{CheeseburgerFood,Chicken TendersFood}	=>	{French Fries BasketFood}	0.003931642	0.9615385	9.850863	75
{Chicken TendersFood,Krazy KritterFood}	=>	{French Fries BasketFood}	0.005661564	0.9557522	9.791584	108
{Chicken TendersFood,Slice of PeppFood}	=>	{French Fries BasketFood}	0.003669532	0.9210526	9.43609	70
{CheeseburgerFood,Souvenir DrinkFood}	=>	{French Fries BasketFood}	0.003250157	0.9117647	9.340936	62

Table 12- Association rules with confidence higher than 0.9

lhs		rhs	support	confidence	lift	count
{Side of CheeseFood}	=>	{Cheese ConeyFood}	0.004665548	0.6846154	25.91215	89
{Side of CheeseFood}	=>	{Hot DogFood}	0.006290627	0.9230769	21.60566	120
{Cheese ConeyFood,Side of CheeseFood}	=>	{Hot DogFood}	0.004351017	0.9325843	21.82819	83
{Hot DogFood,Side of CheeseFood}	=>	{Cheese ConeyFood}	0.004351017	0.6916667	26.17903	83

Table 13 – Association rules with lift higher than 20

Next, we can take a look at the scatter plot of all 40 association rules given the support and confidence values. This can bee seen in **Figure 9.**

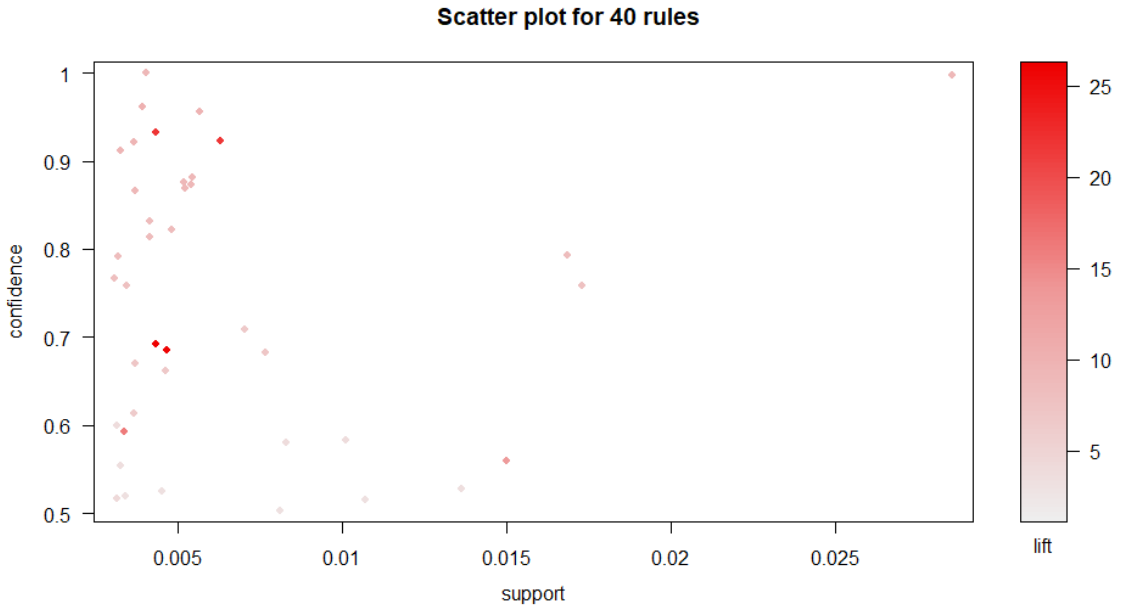


Figure 9 – Scatter plot of all 40 association rules.

It its interesting to see how most of the rules have levels of support lower than 0.01. We can also visualize how 7 of the 40 rules associate with each other as seen in **Figure 10.**

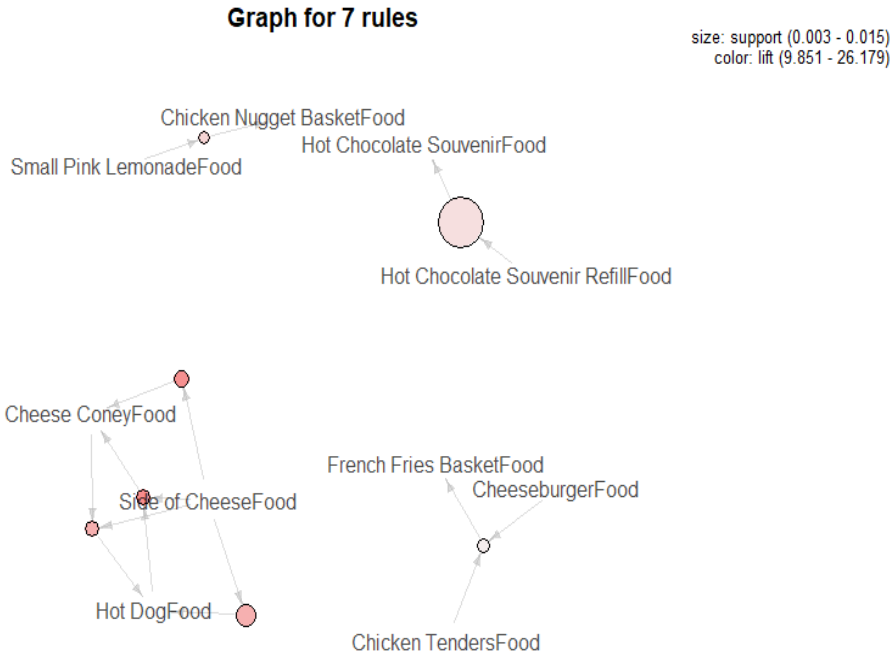


Figure 10 – Association of 7 our of the 40 rules

Finally, in **figure 11** we have plotted the grouped matrix for all 40 rules, where the size of the circles indicate the support and the color indicate the lift.

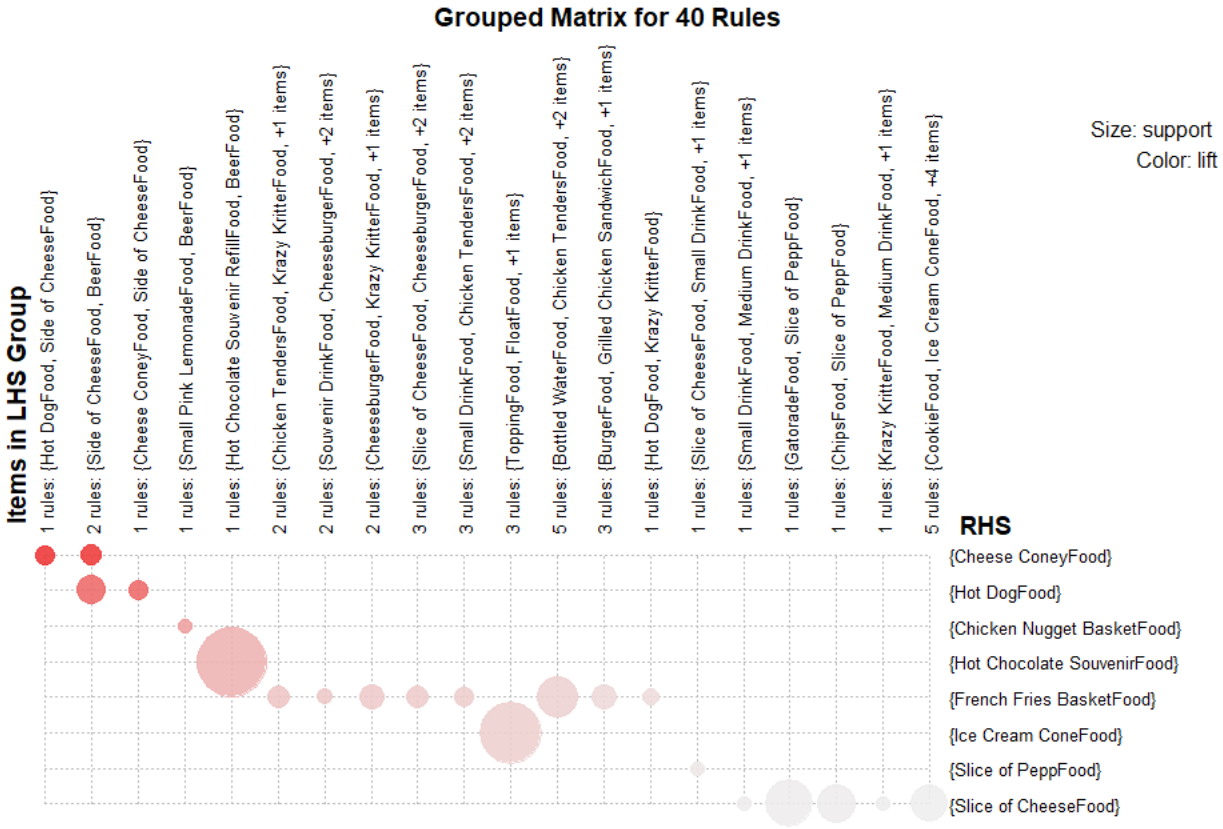


Figure 11- Grouped matrix for all 40 rules