Case study on Cincinnati Zoo Dataset

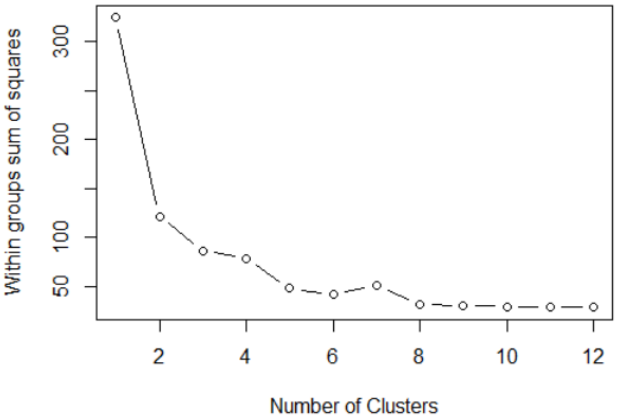
-Udyot Kumar

Executive Summary

Cincinnati Zoo Dataset

Background and MAJOR FINDINGS

This dataset was given to us by Cincinnati zoo and here we have first found the optimal number of clusters by using the group sum of squares, Dunn index and avg. silhouette method.



The number of clusters came to 3 as there is a steep slope till 3.

And after that I deduced some association rules:

Here I have applied the rule that Lift ratio should be greater than 10

	lhs		Rhs	support	confidence	lift
1	**{Small.Pink.LemonadeFood}**	**=>**	**{Chicken.Nugget.BasketFood}**	0.003355001	0.5925926	16.03446
4	**{Side.of.CheeseFood}**	**=>**	**{Cheese.ConeyFood}**	0.004665548	0.6846154	25.91215
5	**{Side.of.CheeseFood}**	**=>**	**{Hot.DogFood}**	0.006290627	0.9230769	21.60566
8	**{Hot.Chocolate.Souvenir.RefillFood}**	**=>**	**{Hot.Chocolate.SouvenirFood}**	0.014992661	0.5596869	13.18097
13	**{Cheese.ConeyFood,Side.of.CheeseFood}**	**=>**	**{Hot.DogFood}**	0.004351017	0.9325843	21.82819
14	**{Hot.DogFood,Side.of.CheeseFood}**	**=>**	**{Cheese.ConeyFood}**	0.004351017	0.6916667	26.17903

These are among the strongest association as higher the lift ratio higher the association. So, the zoo office should put Chicken nugget nearby Pink lemonade or hot Dog near a side of cheese to increase sales. Many rules can be derived from the above table which can be used to increase the sales of the food data.

So, like this any number of inferences can be deduced using the above table.

1) Clustering and Association on Cincinnati Zoo DATA

A) Introduction:

Cincinnati Zoo

Founded in 1873. Officially opened in 1875. Second oldest in the Nation after Pennsylvania Zoo. Zoo houses over 300 animals and over 3,000 plant species.
Reptile house is the oldest Zoo building in the country, dating from 1875.
Zoo serves over a million visitors each year.

Here the main goal is to:

- Identify useful and or/hidden information in the data collected by the zoo.
- To Study buying and/or visiting behavior of Zoo members.

In the food dataset, there are 55 food items.

Firstly, I will do the clustering where I will find the groups or sets in which the data (food) is clustered and after that I will apply association rules to find insights on the customer buying patterns which can help the Cincinnati Zoo to elevate their sales.

B) K-MEANS CLUSTERING

K- means clustering algorithm tries to cluster data based on their similarity. Here the algorithm tries to find pattern in the data. Initially we should specify the number of clusters we want our data to grouped into.

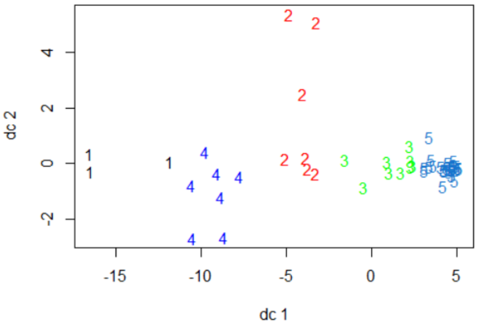


Fig 10: K-Means with 5 clusters

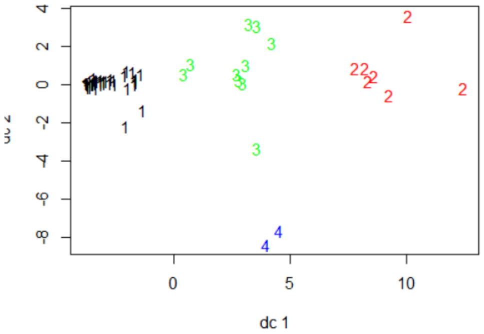


Fig 11: K-means with 4 clusters

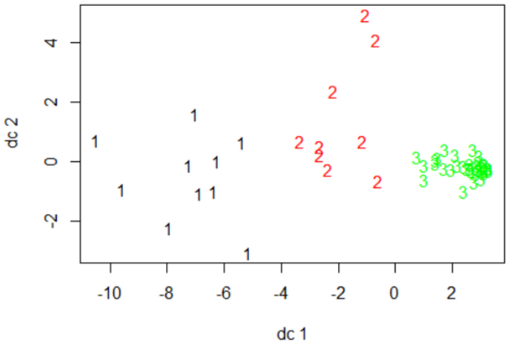


Fig 12: K-Means with 3 clusters

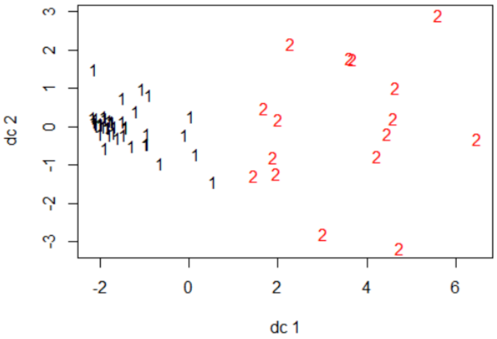


Fig 13: K-means with 2 clusters

C) Determining the optimum number of clusters

Here I will determine the optimal number of clusters based on 3 methods:

1.) Group sum of squares method

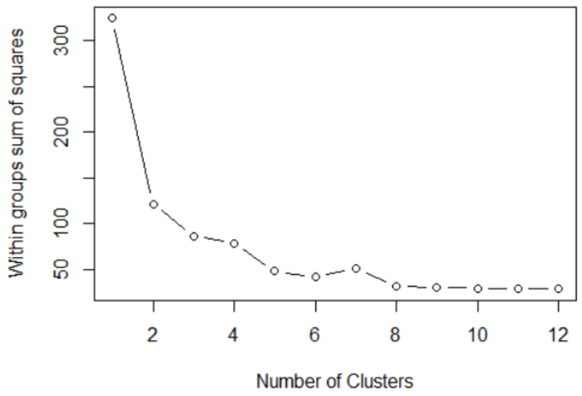
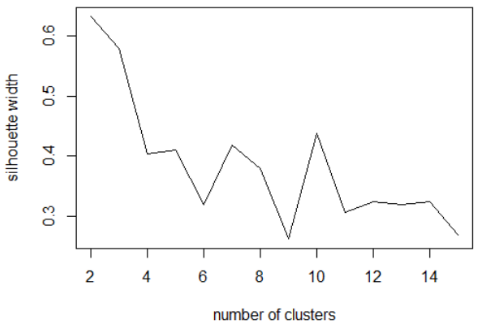


Fig 13: Group sum of squares

With the help of this plot we can see that there is a steep drop from 1 to 3 and from 3 to 4 the drop is not that steep. So, based on this graph I will prefer that there are 3 clusters.



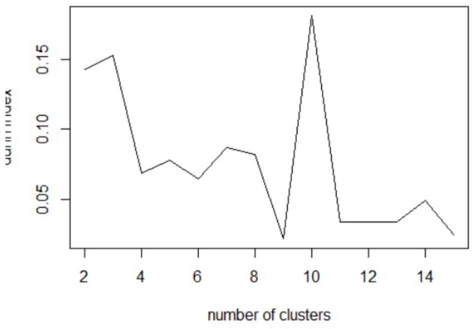


Fig 14: Average Silhouette method

Fig 15: Dunn-Index

On the basis on Dunn index pplot we can clearly see that there is an increase from 2 to 3 so we will select 3 as the number of clusters. We will not select 10 because the data has just 55 food items and clustering them into 10 parts will not make much sense.

Also in the average Silhouette method we can see that 3 has a very high silhouette index. Though 2 has higher but if we compare with other methods of determing the number of clusters 3 is maximum times favored.

D) Analysis of the clusters

Cluster:	1	2	3
No of items:	9	10	36

Here we can notice that Cluster 1 contains 9 items and similarly cluster 2 and 3 contains 10 and 36 items respectively.

Group.1	sales in Oct-10	sales in Nov-10	sales in Dec-10	sales in Jan-11	sales in Feb-11	sales in Mar-11
1	0.1421967	0.6456491	0.993078	0.7531544	0.2723983	0.1838072
2	1.677961	1.7502256	1.0570593	1.2746319	1.7350387	1.8576906
3	-0.5016495	-0.647586	-0.5418971	-0.542353	-0.5500548	-0.561977

Based on above table we can find the mean of sales within each cluster.

E) Hierarchical Clustering

Hierarchical Clustering is an alternate which doesn’t require us to specify the number of clusters beforehand and it builds the hierarchy from bottom-up.

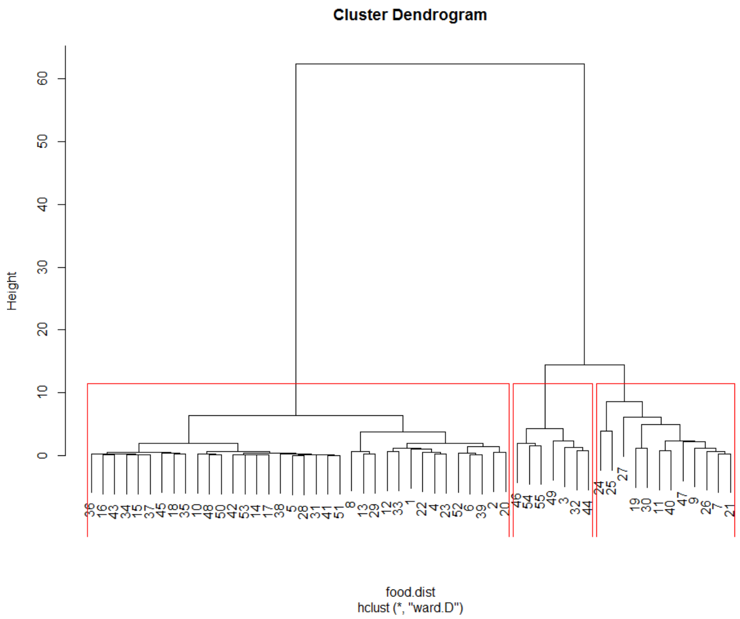


Fig 16: Hierarchical clustering

Here I have created a dendrogram and divided it into 3 clusters.

F) Association Rules

Here the data is food for asscoiation and the dataset contains 19076 observations which are the transactions by the visitors and 118 columns which are the food items available in the zoo.

Summary of the data contains:

Bottled Water	Slice of Cheese	Medium Drink	Small Drink	Slice of Pepp	(Other)
3166	3072	2871	2769	2354	35981

Here we can notice that the visitors have purchased Bottled water the maximum number of times which is understandable as generally people prefer drinking liquid stuff when they are out. After that Slice of cheese and medium drink.

Here is the frequency plot:

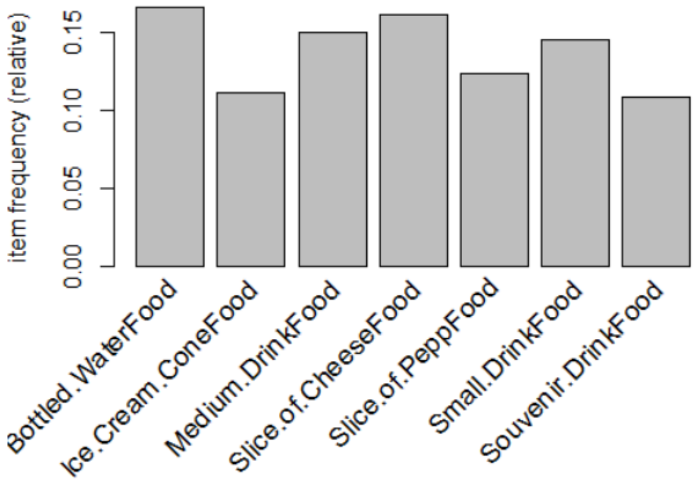


Fig 17: Frequency plot

G) Apriori algorithm

inspect(head(basket_rules)) #lhs: atecedent, rhs: consequent

	lhs		rhs	support	confidence	lift
1	{Small.Pink.LemonadeFood}	=>	{Chicken.Nugget.BasketFood}	0.003355001	0.5925926	16.034463
2	{Grilled.Chicken.SandwichFood}	=>	{French.Fries.BasketFood}	0.003721954	0.6698113	6.862149
3	{FloatFood}	=>	{Ice.Cream.ConeFood}	0.007024533	0.7089947	6.355631
4	{Side.of.CheeseFood}	=>	{Cheese.ConeyFood}	0.004665548	0.6846154	25.912149
5	{Side.of.CheeseFood}	=>	{Hot.DogFood}	0.006290627	0.9230769	21.605663
6	{BurgerFood}	=>	{French.Fries.BasketFood}	0.004613126	0.6616541	6.778579

In the above table, few rules are mentioned which can be interpreted as if a customer has purchased small pink lemonade then he has purchased chicken nugget food also. Similarly, we can deduce relation among other rules. It is understandable as generally people select food items and drink items together.

As there are many rules available I will select a rule where lhs >3 and here we can see that inspect(subset(basket_rules,size(basket_rules)>3))

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

So, from the above table I can deduce that if 3 items have been purchased from the LHS then There's chance that 4th item from RHS is also purchased. Higher the confidence and support better the chance.

Here I am applying the rule that Lift ratio should be greater than 10

	lhs		rhs	support	confidence	lift
1	{Small.Pink.LemonadeFood}	=>	{Chicken.Nugget.BasketFood}	0.003355001	0.5925926	16.03
4	{Side.of.CheeseFood}	=>	{Cheese.ConeyFood}	0.004665548	0.6846154	25.91
5	{Side.of.CheeseFood}	=>	{Hot.DogFood}	0.006290627	0.9230769	21.60
8	{Hot.Chocolate.Souvenir.RefillFood}	=>	{Hot.Chocolate.SouvenirFood}	0.014992661	0.5596869	13.18
13	{Cheese.ConeyFood,Side.of.CheeseFood}	=>	{Hot.DogFood}	0.004351017	0.9325843	21.82
14	{Hot.DogFood,Side.of.CheeseFood}	=>	{Cheese.ConeyFood}	0.004351017	0.6916667	26.17

These are among the strongest association as higher the lift ratio higher the association. So the zoo office should put Chicken nugget nearby Pink lemonade or hot Dog near a side of cheese. Many rules can be derived from the above table which can be use to increase the sales of the food data.

Also now I am trying to find my own rule using the below code:

```
Slice.of.CheeseFood.rhs <- subset(basket_rules, subset = rhs %in% "Slice.of.CheeseFood" &
lift>1.5)
inspect(Slice.of.CheeseFood.rhs)
```

lhs		rhs	support	confidence	lift
{ChipsFood,Slice.of.PeppFood}	=>	{Slice.of.CheeseFood}	0.008282659	0.5808824	3.607068
{GatoradeFood,Slice.of.PeppFood}	=>	{Slice.of.CheeseFood}	0.010117425	0.5830816	3.620724
{Medium.DrinkFood,Slice.of.PeppFood,Small.DrinkFood}	=>	{Slice.of.CheeseFood}	0.003145313	0.6	3.725781

By analyzing this table we can see when slice of cheese is purchased. So We can analyze from the above table that Gatorade and slice of Pepp food led to purchasing of cheese food.

So, in this way we can create multiple association rules and can use them in increasing the sales of food at Cincinnati zoo.