BANA 7047 – Prof. Yan Yu
Individual Case II

Last Name ___Newkold_____

First Name ___Tess_____

M#____11434445_____
Please use your **M#** to set the seed to draw a random sample.

Signature _____
Please use this as your cover page and follow **exactly** the required format below. Points will be deducted otherwise.

Please submit a WORD copy via blackboard with file name 7047-00X case#-last 4 digits of your M#.docx.

Case reports: Please have a cover page including above; **one-page executive summary** on Page 1, clearly stating:

- Goal and Background -- What is the problem?
- Approach -- What have you done?
- Major findings -- What do you find and what is your conclusion?

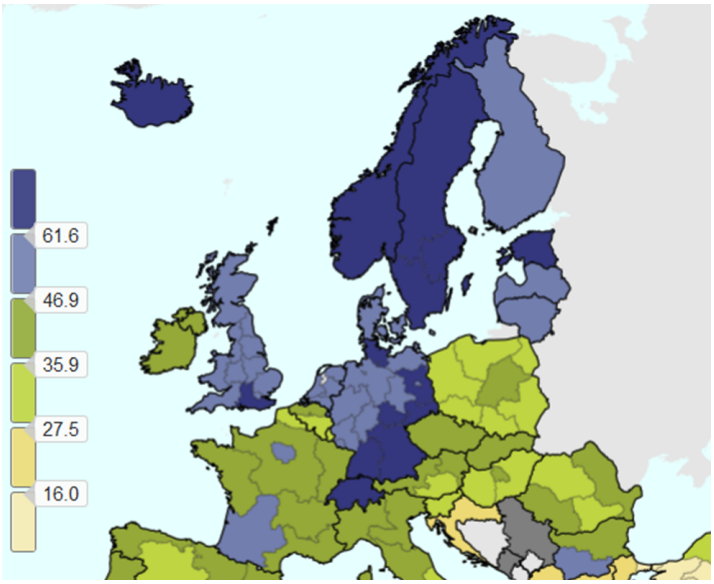
Organize, report, and interpret your **major** outputs with labeled figures and tables in your detailed report. Please do NOT include R code, data, and raw outputs.

Case Three

Data Mining II

Tess Newkold

04-16-2019



European Employment Data

Clustering

Cincinnati Zoo Data

Association Rules

Classification using SPSS Modeler

European Employment Executive Summary

Goal:

Our goal is to evaluate the data in a few different clustering methods, and to evaluate how to choose the appropriate number of clusters for the data.

Approach and Major Findings:

To do this, the data is clustered by many numbers to visually see what looks best (Figure 1). Then a more mathematically sound way to pick the number of clusters is evaluated (Figure 2). I came to the conclusion that in this data what appears visually to be the correct number of clusters is the same as the number that there should be.

Figure 1. Cluster Plot with two, three, four, and five clusters respectively.

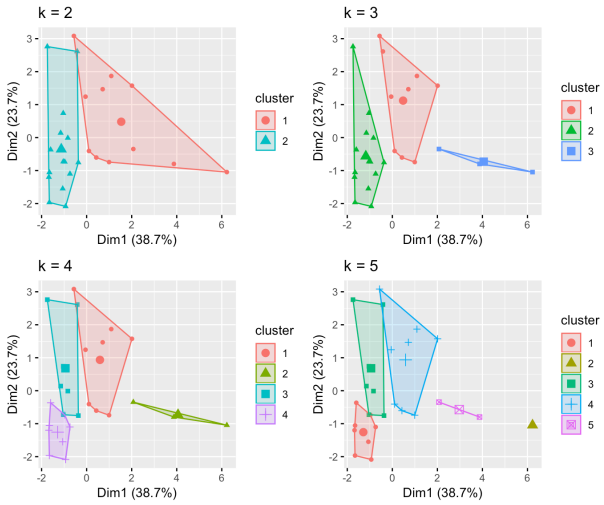
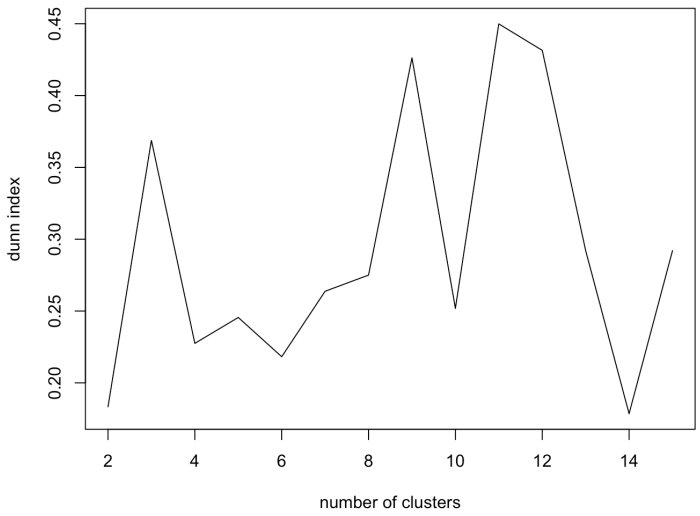


Figure 2. Graph that shows the dunn index that shows the optimal number of clusters at the highest peak



Cincinnati Zoo Executive Summary

Goal:

Our goal is to evaluate the associations that are made in purchase transactions at the Cincinnati zoo. In doing this we can see associations and better help the Zoo know what people are buy together and give them insights into ways to increase revenue.

Approach and Major Findings:

To evaluate the associations in the purchase transactions at the Cincinnati Zoo I first looked at the item frequency. In this data the most frequently bought item is bottled water, which makes sense why places make the price of bottled water so high. To visualize the associations, you can look at many rules, or just a few at a time. The way I think is most understandable is viewing fewer rules at a time. This is show in Figure 2 here, you can see that there are strong associations between chicken nuggets and pink lemonade, this sounds like a child's dream lunch. There is also a strong association between hot chocolate and the souvenir cup. This information is very valuable to the zoo as they can take these associations and market them together, or just knowing things are correlated may help inform them on what items to keep available. It may to helpful for the zoo to look at these association rules each year to see how things are changing and how customers are viewing things in association.

Figure 1. Item frequency chart

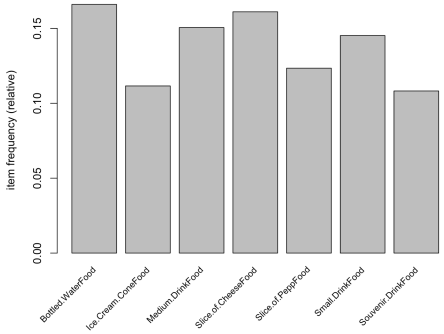
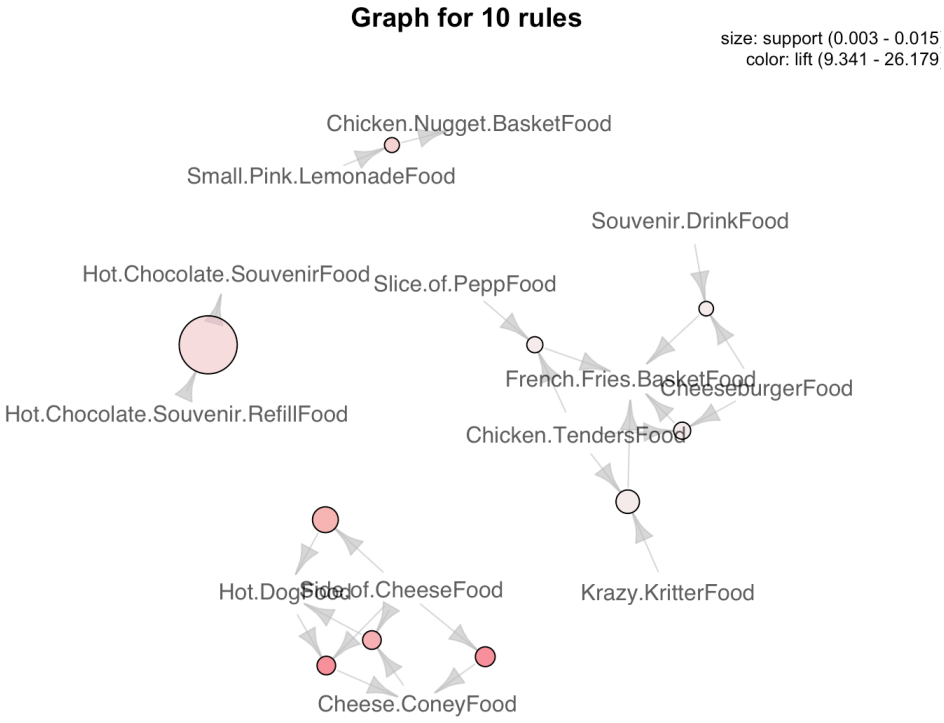


Figure 2. Visual way to see the associations of items for a small number of items



SPSS Classification Executive Summary

Goal:

Our goal is to evaluate the performance of generalized additive model, CART model, neural networks in the SPSS modeler. Shown by AUC and ROC curves for in-sample and out-of-sample performance.

Approach and Major Findings:

A summary table for the AUC for in-sample and out-of-sample performance is shown in Figure 1. A visual representation of those corresponding ROC curves are shown in Figure 2. A diagram of the SPSS modeler is shown in Figure 3. From this analysis it appears that the generalized linear model is best at predicting default between these three models.

Figure 1. Summary of area under the curve values for in-sample and out-of-sample performance.

	AUC In-Sample	AUC Out-of-Sample
Generalized Linear Model	0.807	0.877
Classification Tree	0.731	0.768
Neural Network	0.822	0.807

Figure 2. ROC curves for GLM, CART, and Neural Networks respectively.

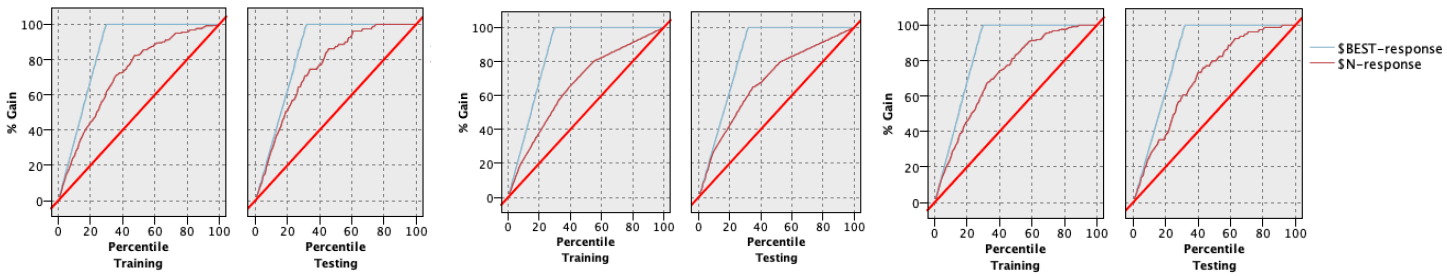
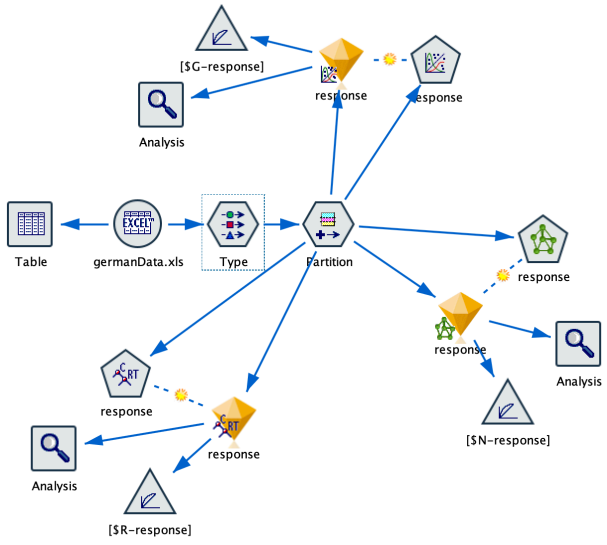


Figure 3. Diagram of the SPSS modeler.



European Employment Report

K-means Clustering

To cluster the European employment data, I first visualized the distance matrix shown in Figure 1. The k-means clustering analysis puts 15 countries in one cluster and 11 countries into the second cluster. I then looked at how the data looks with two, three, four and five clusters shown in Figure 3. Just by eye sight in the plot I would say three clusters appears to be the best choice. To see the data points according to the first two principal components that explains most of the variance is shown in Figure 2.

Figure 1. Distance matrix

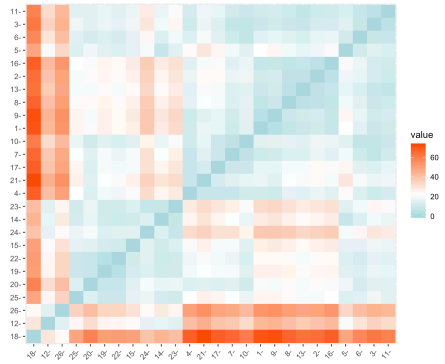


Figure 2. Principal Component Analysis (PCA)

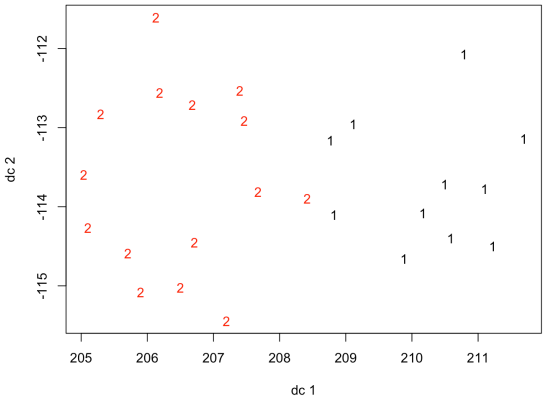
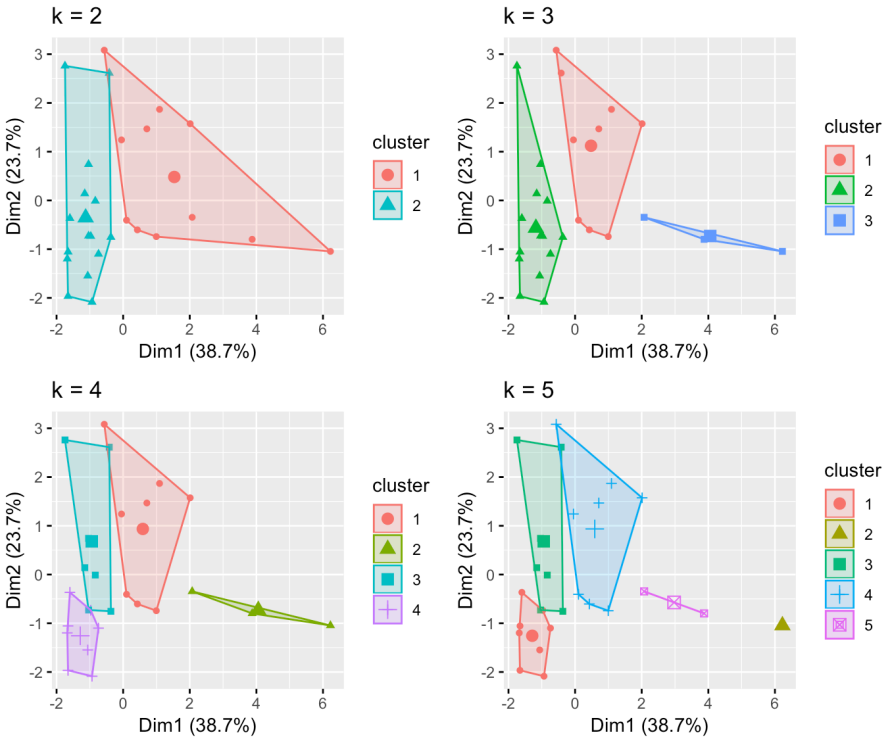
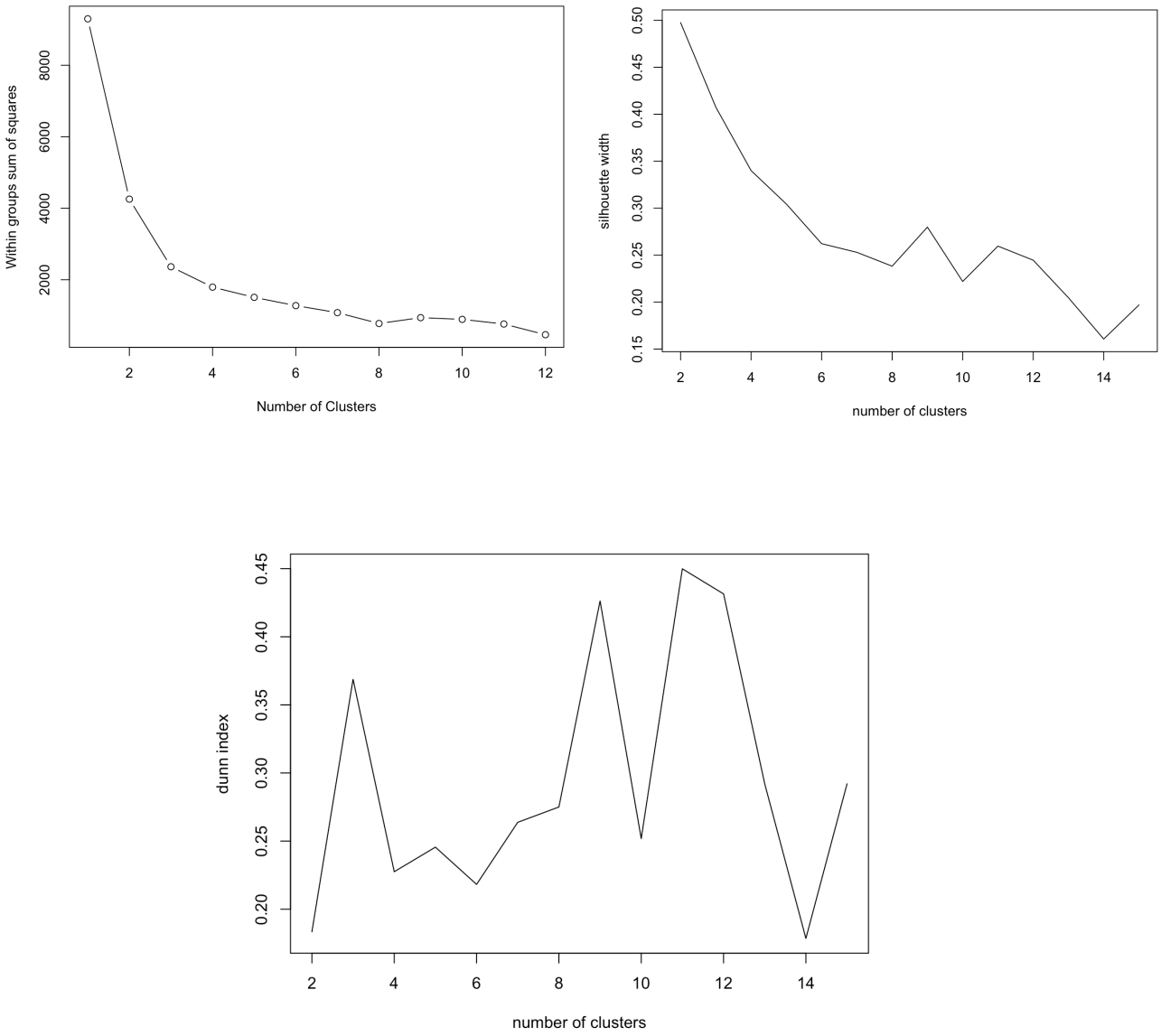


Figure 3. Cluster Plot with two, three, four, and five clusters respectively.



Looking at the graphs and determining how many clusters there should be is not the most accurate way. There are many ways to determine the number of clusters there should be in a set of data. I have shown three different graphs that validate that three clusters is the best number of cluster to proceed with. The first shows that the sum of squares is smallest around three clusters without becoming over fitted, with not much further reduction in distance between clusters. The second graph shows the silhouette width decreasing drastically when there is three clusters, and in the third graph that shows the dunn index that shows the optimal number of clusters at the highest peak. All three of these graphs confirm that the number of clusters should be three.

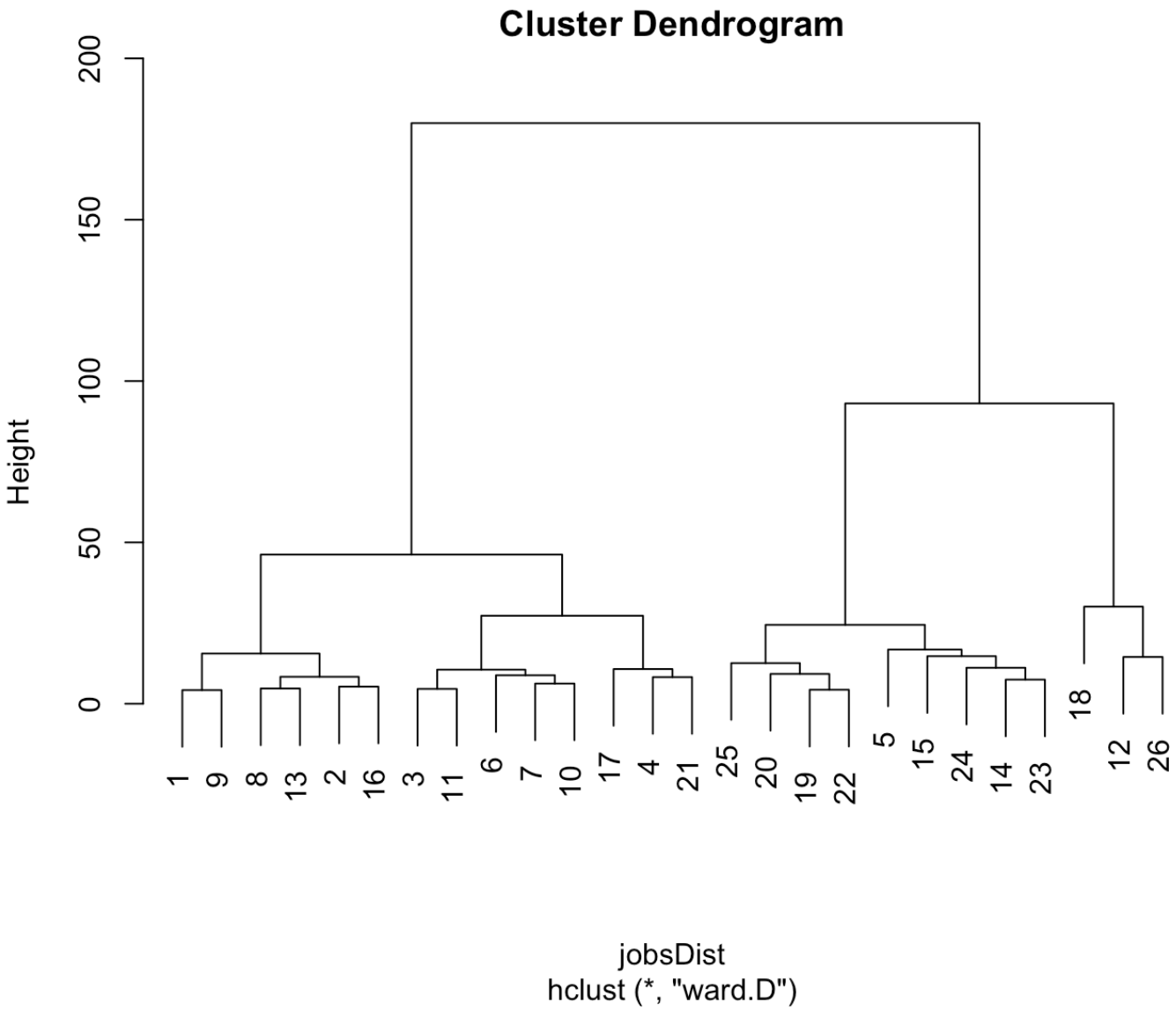
Figure 4. Three graphs showing that the optimal number of clusters should be three.



Hierarchical Clustering

Hierarchical clustering is an alternative way to model our data in a cluster fashion. The dendrogram at the three clusters level is shown below in Figure 5.

Figure 5. Dendrogram at 3 clusters level



Cincinnati Zoo Report

Association Rules

For the transaction data at the Cincinnati zoo one method to use is association rules that is good for showing relationships between variables in large databases. This dataset has 190,076 transactions with 118 items that could be purchased each time. The most frequent items are bottled water purchased 3166 times, slice of cheese pizza bought 3072 times, medium drink bought 2871 times, small drink purchased 2769 times, and a slice of pepperoni pizza bought 2354 times. The mean number of items purchased during one transaction is 2.632 with the minimum being 0 and maximum being 15 items purchased. Items that are important in the dataset are shown in Figure 6 below. Figure 7 shows association rules using support and confidence on the axes and left being represented as the color of the points. The following figure is a good visualization of a smaller set of rules that have the highest lift value, this is seem in Figure 8. The last figure is a way to visualize many rules at the same time. The main thing that I observe is that bottled water is the most frequently item purchased, and that there are some strong correlations between chicken nuggets and a small pink lemonade. This seems like a meal a small child would like, so potentially marketing these two options in a combo meal could increase sales further for this combination.

Figure 6. Item frequency chart

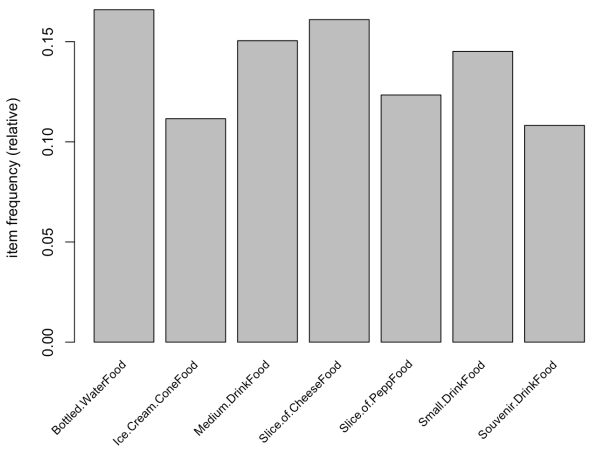


Figure 7. Visualization of association rules using support and confidence on the axes and left being represented as the color of the points.

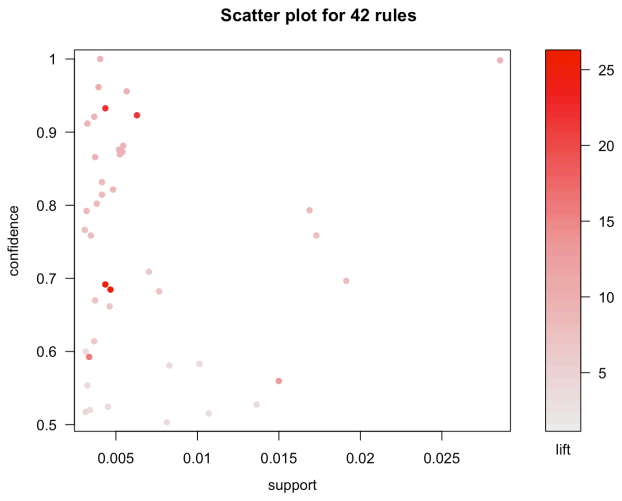


Figure 8. Visualization of a small number of association rules.

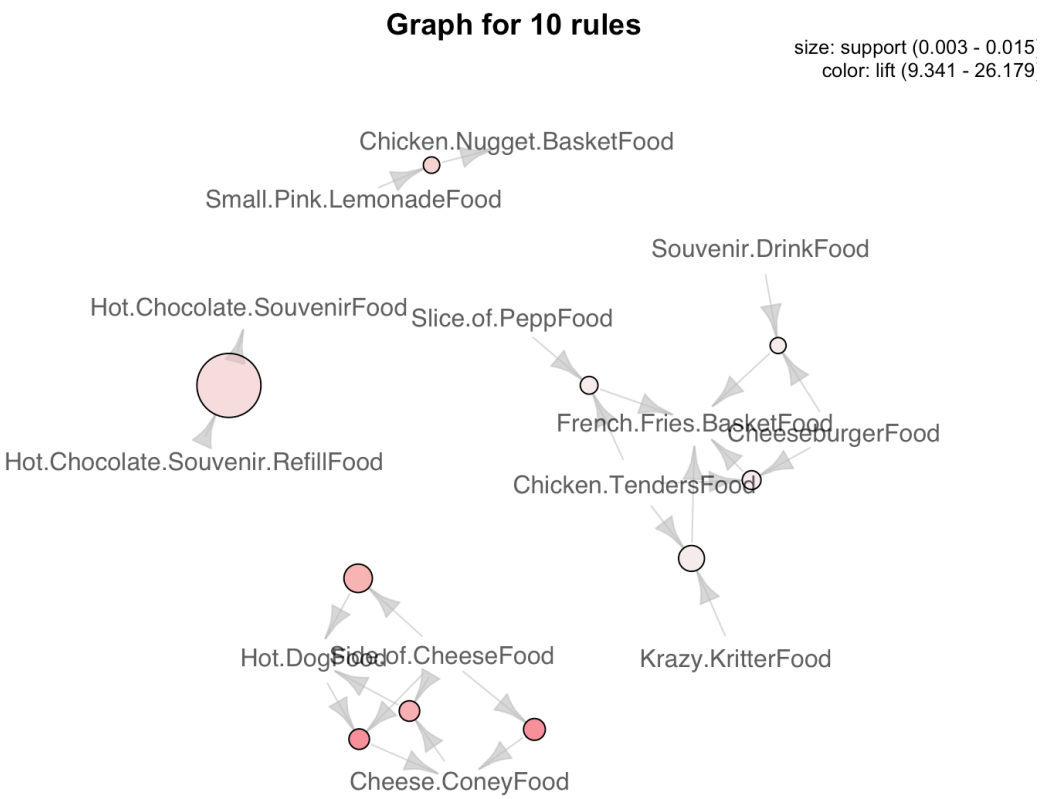


Figure 9. Visualization of a large number of association rules.