

# CAB330 - Data and Web Analytics

## Case Study 2

Group Name: UbiquiData

Team Number: 17

Prepared by:

Mitchell Sugden	09136321
Michael Lee	08433518
Harrison Jubb	08433534

11th October, 2015

Weighting: 25%

# Part 1: Descriptive Data Mining: Clustering

## The data description

Name	Description
Location Number	Numerical code for the store (occurs only once in the table)
DEALER CODE	Text identifier for the store (also only occurs once in the table)
REPORT DATE	Date of the data extraction
HATCH	the number of hatch back model cars sold by the store
SEDAN	The number of sedan model cars sold by the store
WAGON	the number of station wagon model cars sold by the store
UTE	the number of utility/tray back model cars sold by the store
K_SALES_TOT	The total sales for the store (thousands of dollars)

The MODEL-CAR-SALES data set gives the number of four different car models sold at stores. Each row represents an individual store. There are eight columns in the data set. The first and second column contain the store identification number and store label code, the third is the date that the report was generated, and the next four columns contain the number of each type of model sold. The last column is a derived variable that shows the total sales for each store. The sales numbers are over a specified time period.

The company has noticed that stores seem to have an overall preference for certain combinations of model types, with some stores referencing a predominance of sales of two model types; e.g. Hatch and Sedan or Hatch and Wagon, thus creating segments in their market. They want to find the minimum number of product sale segments, to allow development of advertising to match the sales in stores of each segment.

The task is to conduct k-mean clustering on this data set, and find and describe the minimum number of effective clusters. Perform the following tasks on this dataset.

## Task 1. Data Preparation.

1.1. Set up a new project, a new diagram and a data source using the data set **MODEL\_CAR\_SALES**

1.2. Create a graphical report of all variables using the Explore menu. Plot the distribution of the variables. Are there any unusual data values? Are there missing values that should be filtered out or replaced? (add a couple of screenshots here)

There exists some Dealerships in the data that are missing all sales numbers. This was found by examining the *Sample Table* in the results of the *Graph Explore* node.

LOCATION_NUMBER	REPORT_DATE	DEALER_CODE	UTE	HATCH	WAGON	SEDAN	K__SALES_TOT
4	30/04/2013	Euro-103					
24	30/04/2013	Euro-123					
50	30/04/2013	Euro-149					
108	30/04/2013	Euro-201					
173	30/04/2013	Euro-250					
174	30/04/2013	Euro-291					
175	30/04/2013	Euro-292					
176	30/04/2013	Euro-293					
177	30/04/2013	Euro-294					
198	30/04/2013	Euro-295					
199	30/04/2013	Euro-294					
200	30/04/2013	Euro-295					
298	30/04/2013	Euro-374					
299	30/04/2013	Euro-375					
300	30/04/2013	Euro-376					
643	30/04/2013	Euro-688					
644	30/04/2013	Euro-689					
645	30/04/2013	Euro-69					
646	30/04/2013	Euro-70					
665	30/04/2013	Euro-89					
666	30/04/2013	Euro-90					
667	30/04/2013	Euro-91					
28	30/04/2013	Euro-127	6		1071	257	823
352	30/04/2013	Euro-423	6		1055	227	953
530	30/04/2013	Euro-585	7		839	312	1075
632	30/04/2013	Euro-678	7		1007	249	992
							452
							451
							468
							473

Figure 1: Results of Sample Table using all variables

The initial *Cluster Plot* created using a *Variable Clustering* node shows some relevance between the models *Sedan* and *Hatch*, and *Ute* and *Wagon*.

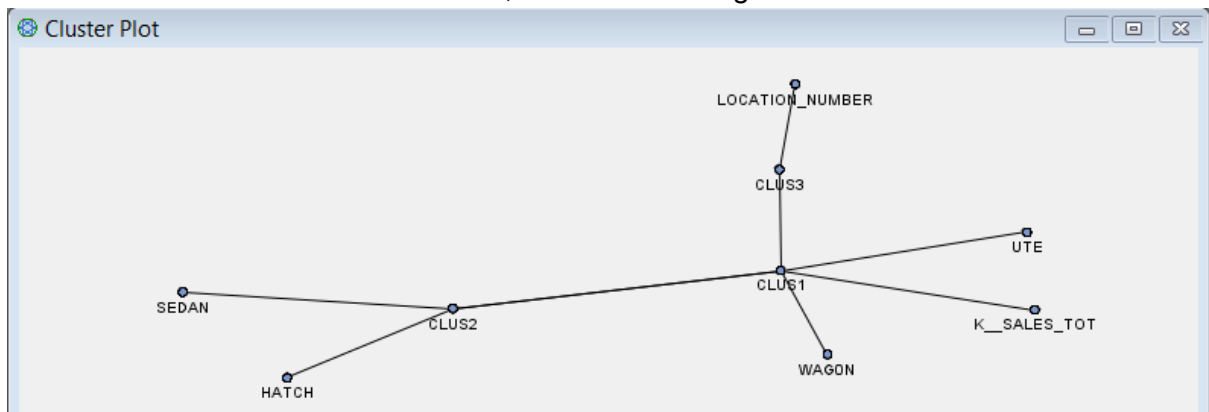


Figure 2: Cluster Plot using all variables

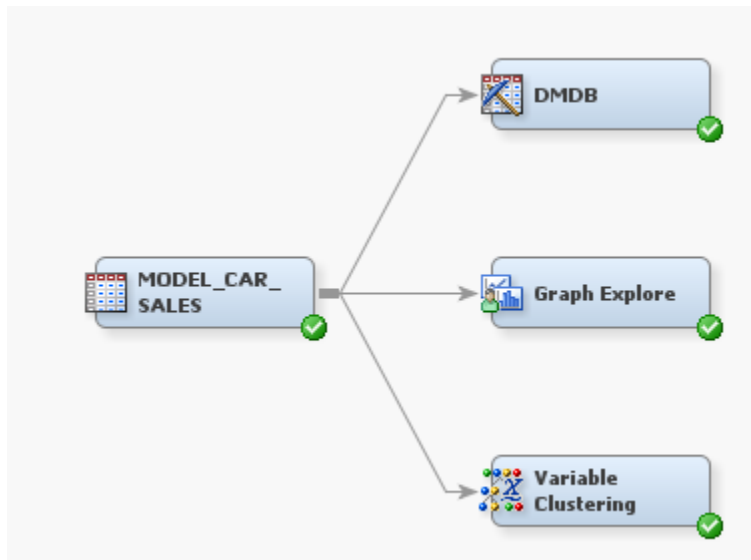


Figure 3: The data diagram after task 1.2.

1.3. One of the model types is underperforming in sales, and the company has decided to drop it from manufacturing. Use the charts to identify which of the products from Hatch, Sedan, Ute and Wagon is to be dropped. Now onwards, the selected (and pruned) product should not be part of analysis.

The model that is underperforming in sales in the *Ute*. This is obvious from looking through the table data as the number of *Utes* sold is commonly much lower than the amount of the next lowest seller. This is backed up by the *DMDB* node. Looking at the summary of the variables we found that *UTE* has a much lower maximum and mean than the other models.

37	Variable	Label	Missing	N	Minimum	Maximum	Mean	Deviation	Skewness	Kurtosis
38										
39	HATCH		22	653	648	2927	1914.08	352.370	-0.17953	0.6192
40	K_SALES_TOT		22	653	450	965	904.43	75.273	-3.87045	16.6057
41	LOCATION_NUMBER		0	675	1	675	338.00	195.000	0.00000	-1.2000
42	SEDAN		22	653	823	2715	1850.86	290.073	-0.17594	0.1802
43	UTE		22	653	6	209	97.18	32.251	0.44854	0.7522
44	WAGON		22	653	2	1224	446.99	212.045	0.36534	0.6163
--										

Figure 4: Output log from the DMDB node using all variables

1.4. Assign appropriate model roles to the variables. Justify if you have rejected any variable for the clustering process? (add a screenshot here).

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
DEALER_CODE	Rejected	Nominal	No		No	.	.
HATCH	Input	Interval	No		No	.	.
K_SALES_TOT	Rejected	Interval	No		No	.	.
LOCATION_NUMBER	Rejected	Interval	No		No	.	.
REPORT_DATE	Rejected	Interval	No		No	.	.
SEDAN	Input	Interval	No		No	.	.
UTE	Rejected	Interval	No		Yes	.	.
WAGON	Input	Interval	No		No	.	.

Figure 5: Variable roles

This data clustering process will be using undefined data therefore no target is required.

**DEALER\_CODE** - Not needed for clustering purposes.

**K\_SALES\_TOT** - Not needed as the price for each model will vary significantly and is unknown. Also without Ute being included the total sales will not solely represent the remaining models.

**LOCATION\_NUMBER** - Not needed for clustering analysis

**REPORT\_DATE** - Same for every Dealership. Redundant

**UTE** - Dropped and rejected from task 1.3.

### **1.5 Include a Filter node and configure to remove instances with missing values if any.**

Running a filter node with the default settings has successfully removed the 22 missing values for each model. This was found by reading the output log.

## Task 2. Building the cluster

2.1. Build a default clustering model using the clustering node. Examine the results:

(a) What was the optimal number of clusters?

Note: The default Clustering node uses **Standardization** as default.

The default clustering node built a solution with 49 clusters. This was found by examining the CCC plot.

(b) Which of the variables were found important in determining the clusters?

*HATCH* was the most important variable with an importance on 1.0. *SEDAN* and *WAGON* also had high importance with values of 0.90 and 0.79 respectively.

Add the screenshots of (1) the CCC plot:

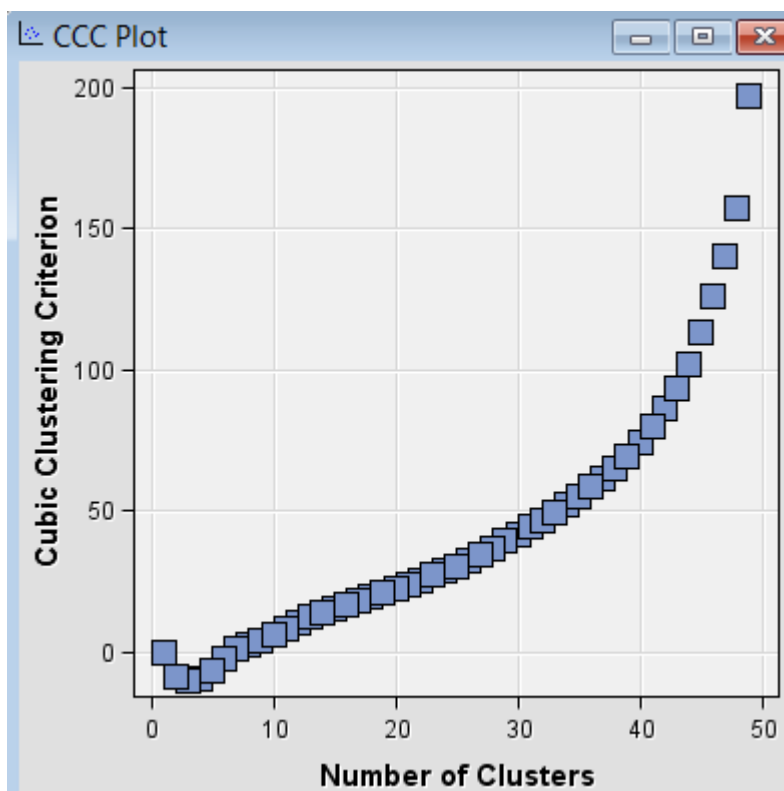


Figure 6: CCC Plot using default clustering node

(2) the cluster result screen:

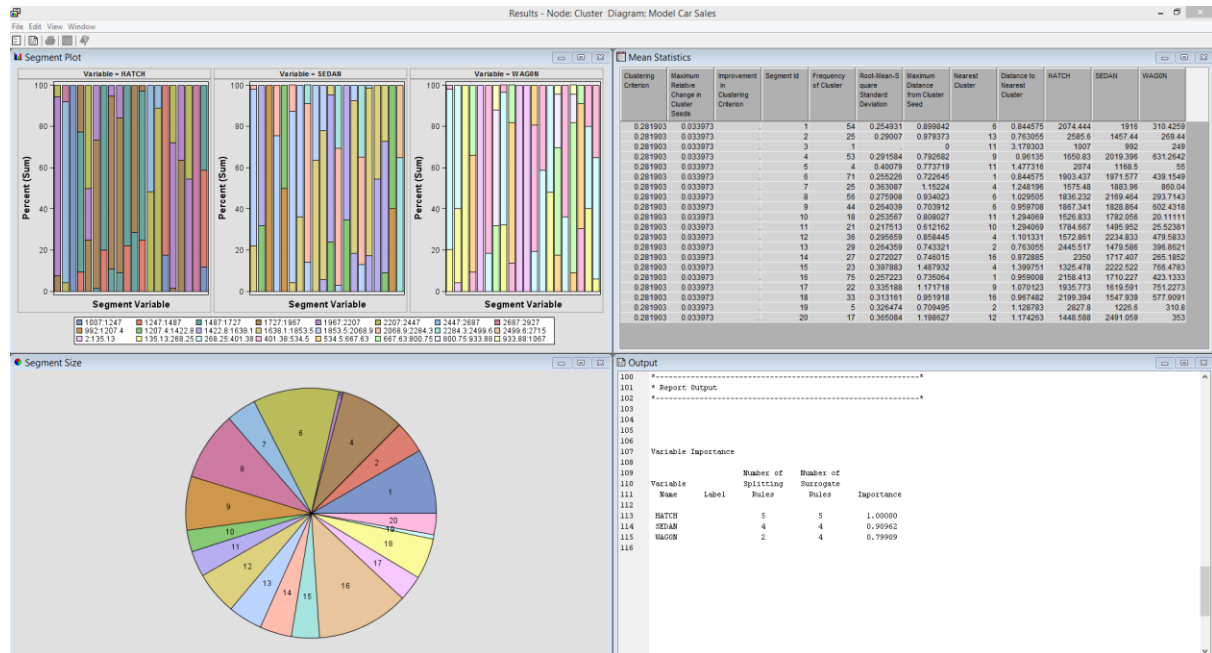


Figure 7: Results screen of default clustering node

(3) the variable importance list:

107	Variable Importance				
108					
109					
110	Variable		Number of	Number of	
111	Name	Label	Splitting Rules	Surrogate Rules	Importance
112					
113	HATCH		5	5	1.00000
114	SEDAN		4	4	0.90962
115	WAGON		2	4	0.79909
116					

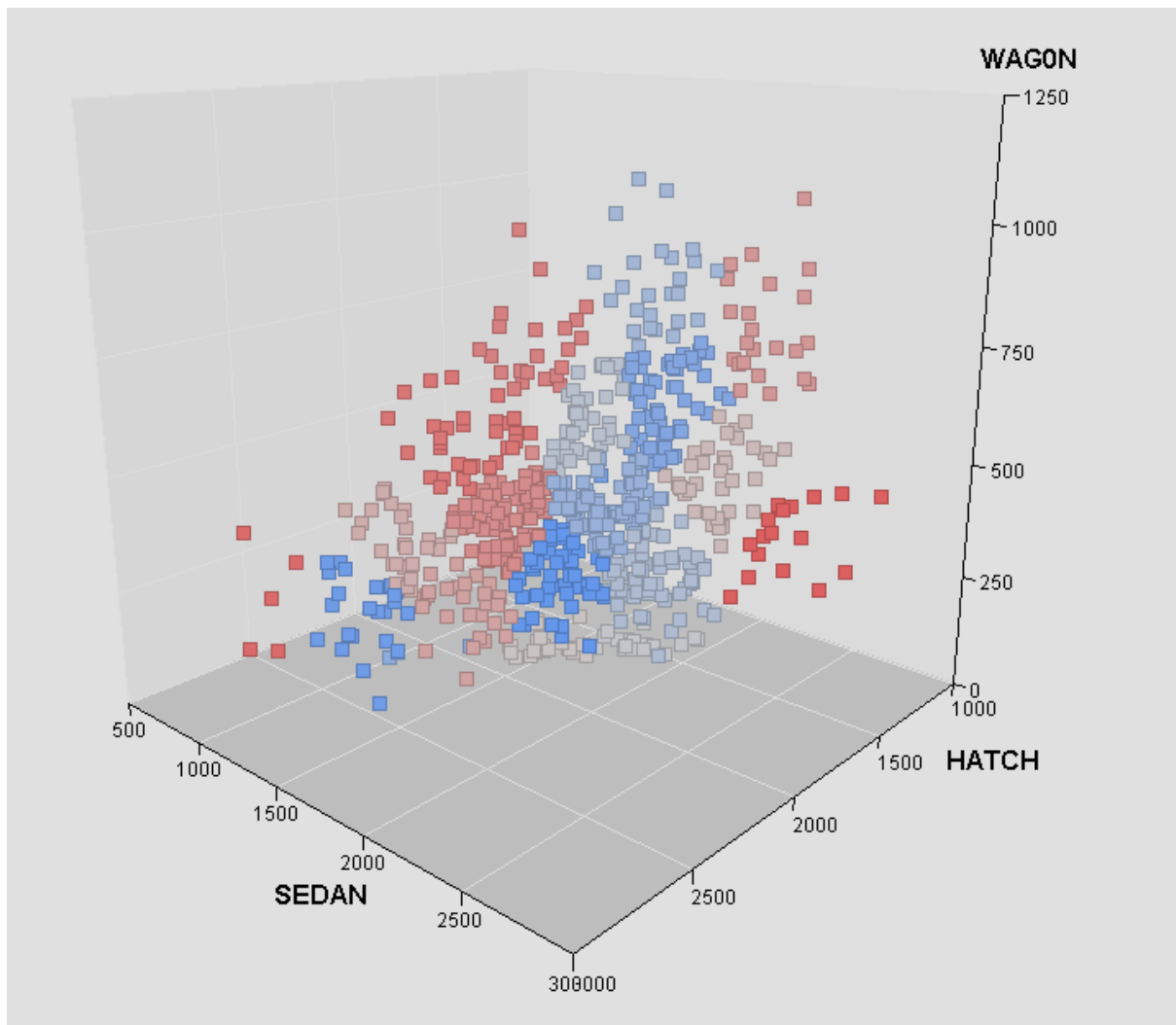
Figure 8: Section of output log taken from default clustering node

Draw a 3 dimensional plot of the clusters, and summarize the characteristics.

**2.2. What is the effect of using the standardization method on the outcome? Does the use of this method enable a better clustering solution in this dataset?**

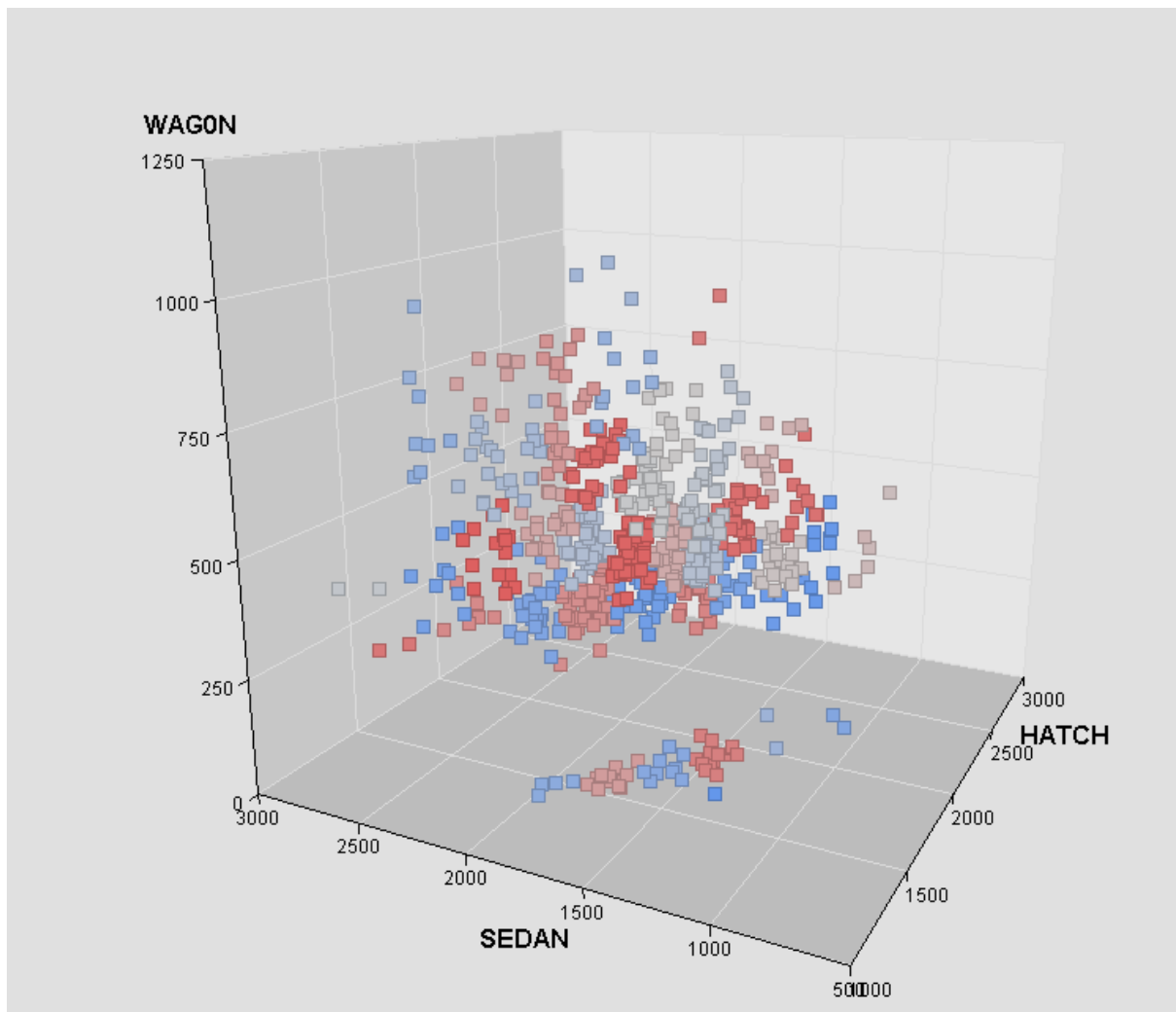
Standardization ensures that the node uses similar measurement scales for efficient clustering. In this dataset this method does allow for a better clustering solution, as all variables are of the same measurement and missing values have been filtered.

(add a few screenshots here)

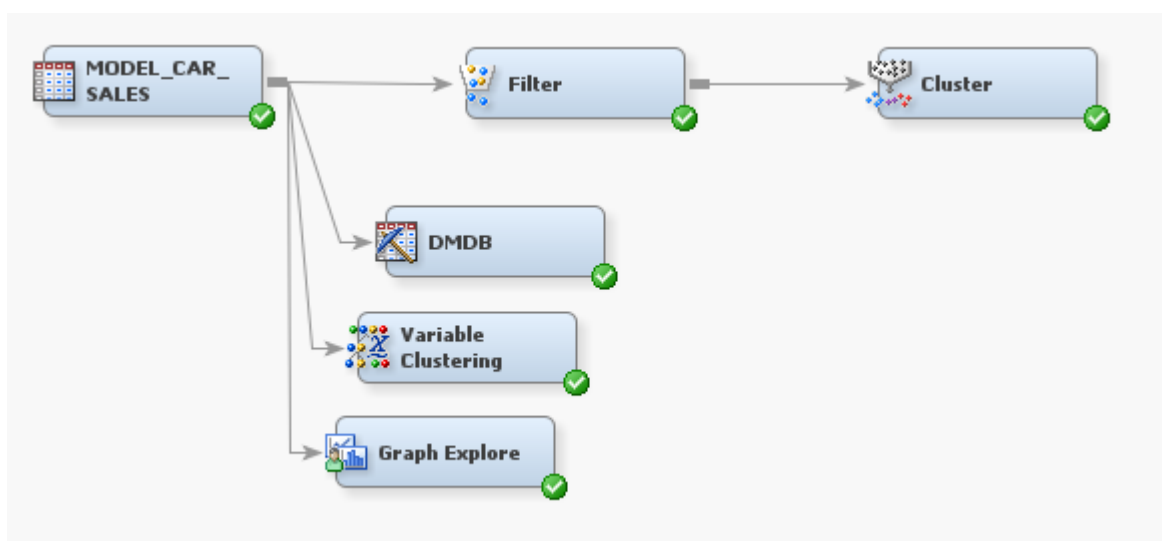


**Figure 9: 3-Dimensional scatter plot using Standardization**





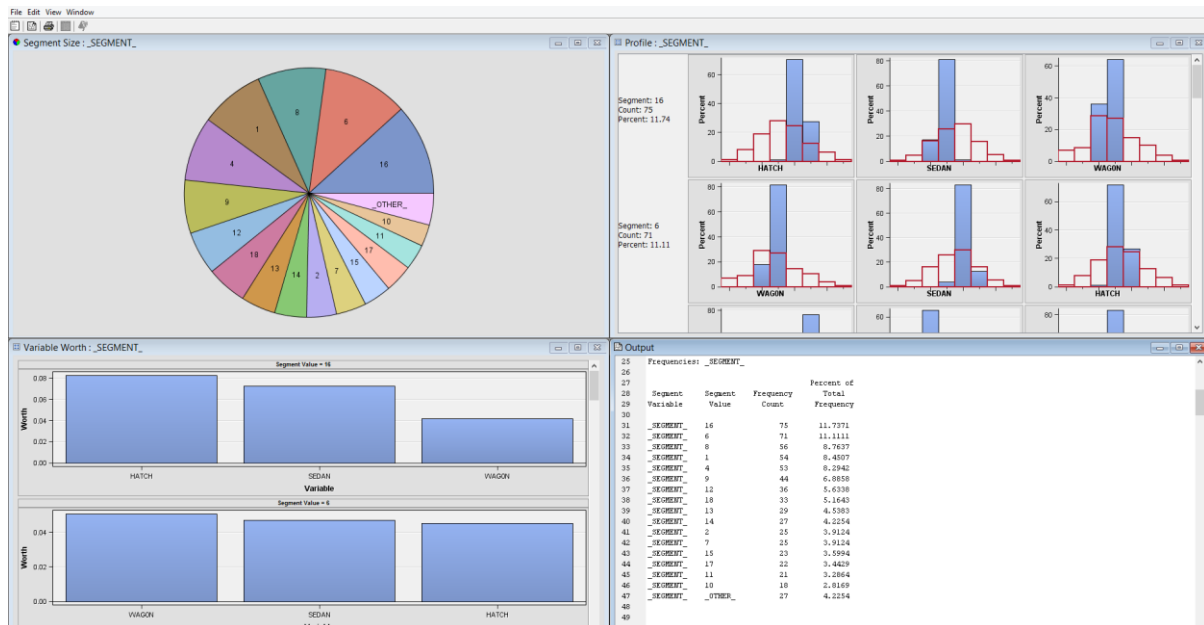
**Figure 10: 3-Dimensional scatter plot without Standardization**



**Figure 11: Data model at step 2.2.**

**2.3. Connect a Segment Profile node to the Cluster node (with the best setting) in order to compare the numbers of different types of model cars sold in each of the clusters. After examining the results from the Segment Profile node, characterize the nature of each cluster by giving it a descriptive label.**

The results of the Segment Profile node show 16 segments each representing different similarities in model sales. By looking at each segment we can see which models are sold most frequently together and identify which segment each dealership belongs to for improved marketing.



**Figure 12: Results of Segment Profile node**

**Segment 1:** Hatch sales are the highest. Similar sales numbers for Sedan and Wagon.

**Segment 2:** Hatch sales are the highest. Sedan and Wagon sales are similar and much lower.

**Segment 3:**

**Segment 4:** Wagon sales are the highest. Similar sales for Hatch and Sedan.

**Segment 5:**

**Segment 6:** Second highest worth segment. Wagon sales are the highest, just below are Sedan and Hatch which are almost identical.

**Segment 7:** Wagon sales are the highest. Hatch sales are low. No Sedan sales in this segment.

**Segment 8:** Third highest worth. Sedan sales are almost twice as much as Wagon or Hatch. Wagon and Hatch sales are similar.

**Segment 9:** Sedan and Wagon sales are similar. Hatch sales are much lower.

**Segment 10:** This segment only contains Wagon sales.

**Segment 11:** Wagon sales are the highest. Sedan sales a much lower. No Hatch sales for this segment.

**Segment 12:** Sedan sales are the highest. Hatch has a medium amount of sales, Wagon sales are low. No clear relationship between two models.

**Segment 13:** Hatch sales are high, followed by Sedan. No Wagon sales in this segment.

**Segment 14:** Hatch sales are high, followed by Wagon. No Sedan sales in this segment

**Segment 15:** Hatch sales are the highest. Wagon and Sedan sales are similar although much lower.

**Segment 16:** Highest worth segment. Hatch and Sedan making up majority of sales. Wagon sales are significantly lower.

**Segment 17:** This segment only contains Wagon sales.

**Segment \_OTHER\_:** This segment has the least amount of worth. It contains the rest of the values from the data set.

## Part 2: Descriptive Data Mining: Association

### The data description

Name	Description
LOCATION	Point of sale device identification number (e.g. for Register 3)
TRANSACTION_ID	Unique transaction identification number for a given sale. A sale may include several products and thus the same transaction id may occur over several rows.
TRANSACTION_DATE	Date of transaction
PRODUCT_NAME	Product Purchased
QUANTITY	Quantity of this product purchased (always set to 1 by a point of sale device)

A store is interested in determining the associations between items purchased from the health and beauty aids department and the stationery department. The store has chosen to conduct a market basket analysis of specific items purchased from these two departments.

The data set contains information on over 400,000 transactions made over the past three months. The following products are represented in the data set:

[Bar soap, Bows, Candy bars, Deodorant, Greeting cards, Magazines, Markers, Pain relievers, Pencils, Pens, Perfume, Photo processing, Prescription medications, Shampoo, Toothbrushes, Toothpaste, Wrapping paper]

The task is to conduct association analysis on this data set. The detailed descriptions of required tasks are as follows.

### Task 3. Data Preparation

**3.1. Open a new diagram and a new data source using the data set POS\_TRANSACTIONS, in the same project as clustering analysis.**

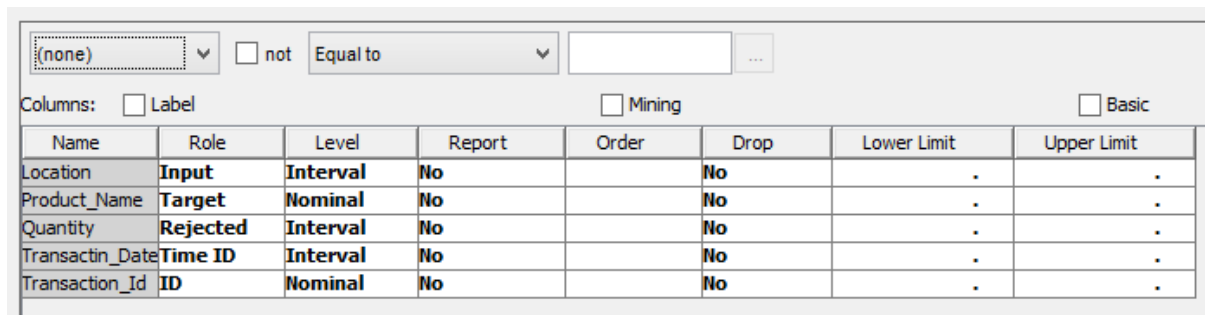
**3.2. Examine the distribution of the variables. Are there any unusual data values?**

The *Product\_Name* feild has a 14 character maximum which is cutting off the names of some products .e Photo Processi-. This was found by using a **Graph Explore** node with **Size: Max** and a **Stat Explore** node. Changing the size on the **Graph Explore** node allowed us to view more tuples and revealed new values for Transaction\_Date which previously only displayed one value.

### Are there missing values that should be replaced?

There does not appear to be any missing values in this dataset. This is not uncommon in transaction records as they are produced automatically (no human error).

### Assign appropriate model roles to the variables. Justify if you have rejected any variable for the association mining process? (screenshots here)



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Location	Input	Interval	No		No	.	.
Product_Name	Target	Nominal	No		No	.	.
Quantity	Rejected	Interval	No		No	.	.
Transaction_Date	Time ID	Interval	No		No	.	.
Transaction_Id	ID	Nominal	No		No	.	.

Figure 13: POS\_TRANSACTIONS variables

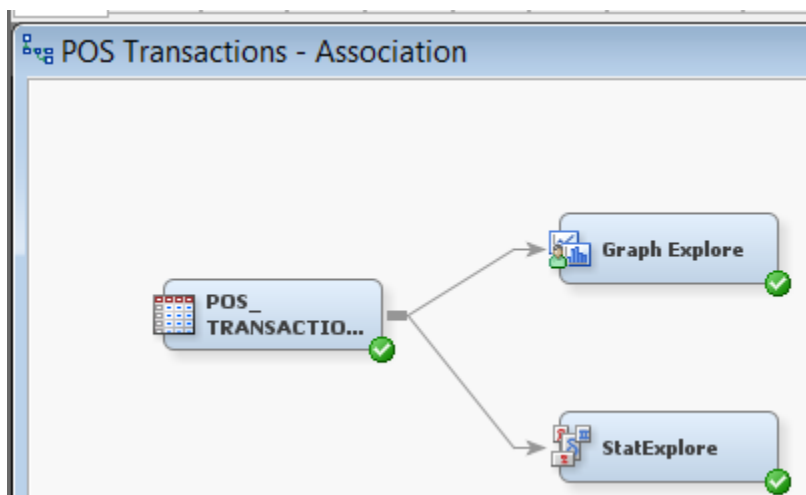


Figure 14: Data exploration of POS\_TRANSACTIONS

Association mining requires one target variable and one or more unique identification variable. It is ideal for the target to have a nominal measurement level.

#### Transaction\_ID

Model Role: ID

Measurement: Nominal

#### Product\_Name

Model Role: Target

Measurement: Nominal

### Rejected Variables:

**Quantity** - This variable has been rejected from processing as it is set to "1" by default. As well as this, multiple purchases of the same item show as multiple tuples, so this value is irrelevant.

## Task 4. Association Mining

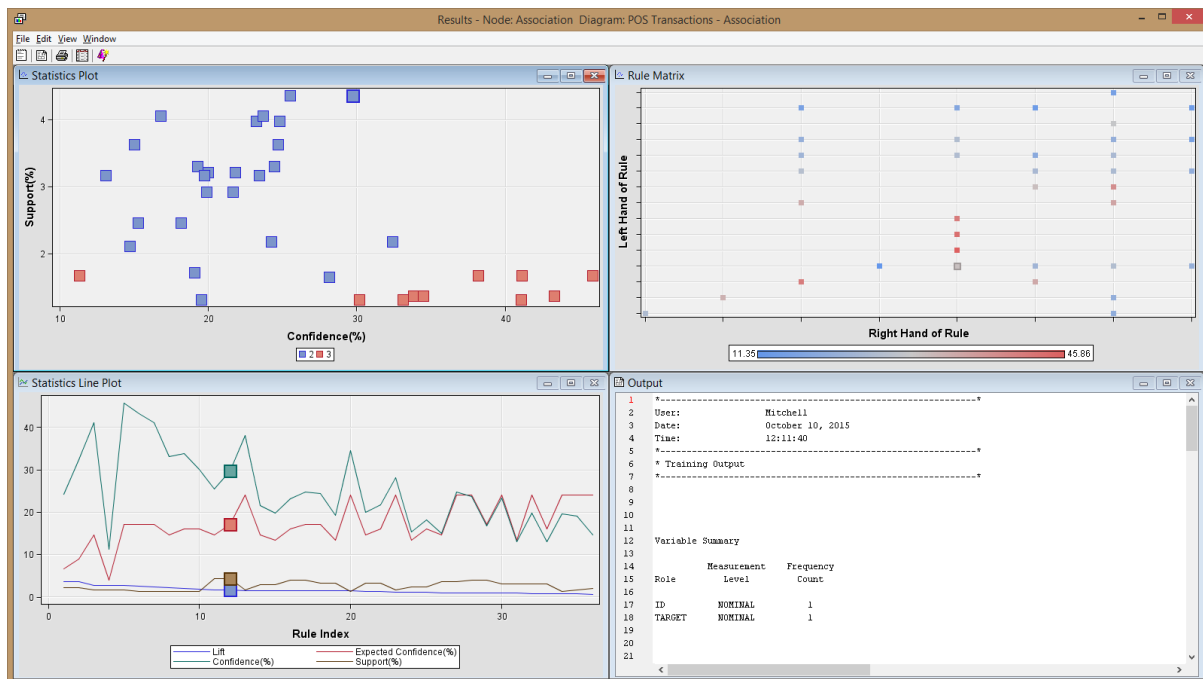
### 4.1. Perform association analysis. Examine the results of the association analysis.

**What is the highest lift value for the resulting rules? Which rule has this value?**

Rules are listed in descending order of lift, so we can easily find the rule with the highest lift by looking at **View → Rules → Rule Description**. Lift values can be found by looking at the Association Report in the Output log.

The highest lift value for the rules is 3.6. This is for the rules **Perfume ⇒ Toothbrush** and **Toothbrush ⇒ Perfume**. This means that a customer who has purchased perfume or a toothbrush is 3.6 times more likely to buy the second item than a customer chosen at random.

**Discuss each of the result node windows - Statistics Plot, Statistics Line Plot, Rule Matrix, and Output – succinctly? Interpret them to discuss the rule-set obtained. (add screenshots here)**



**Figure 15: Results of Association diagram on POS\_TRANSACTIONS data.**

## Statistics Plot

Visual representation of the rules based on their Support and Confidence. Clicking on a rule provides more information and shows its position on the Statistics Line Plot.

## Statistics Line Plot

The Statistics Line Plot shows the confidence, expected confidence, lift and support for each rule. The rules are ordered by their index number, which is in descending order of lift.

## Rule Matrix

The Rule Matrix positions rules based on the left hand side, and the right hand side. i.e with the rule Perfume  $\Rightarrow$  Toothbrush, Perfume is the left hand side, and Toothbrush is the right hand side. It uses a heat map to show the confidence of each rule.

This is a useful visual tool as it allows us to easily identify that there is a high confidence among rules that have Candy Bar on the right hand side.

## Output

The Output window shows the statistics for each rule and the results of the SAS data mining process including any errors.

## Dataset Observations

From interpreting the results we have found:

- Perfume  $\Rightarrow$  Toothbrush has the highest lift. This makes sense as they are both hygiene products.
- Pencils  $\Rightarrow$  Toothpaste has the lowest lift. This makes sense as stationary and hygiene products are less likely to be purchased together.
- Greeting Cards  $\Rightarrow$  Candy Bar has the highest support. This means that 29.72% of people who buy Greeting Cards also buy a Candy Bar. This happens in 4.37% of all purchases.
- Magazine & Greeting Cards  $\Rightarrow$  Candy Bar has the highest confidence.

Output									
22	Association Report								
23									
24									
25		Expected							
26		Confidence	Confidence	Support	Lift	Transaction	Rule		
27	Relations	(%)	(%)	(%)		Count		Left Hand of Rule	Right Hand of Rule
28									
29	2	6.74	24.26	2.18	3.60	4364.0	Perfume $\Rightarrow$ Toothbrush	Perfume	Toothbrush
30	2	9.00	32.40	2.18	3.60	4364.0	Toothbrush $\Rightarrow$ Perfume	Toothbrush	Perfume
31	3	14.69	41.11	1.67	2.80	3333.0	Magazine & Candy Bar $\Rightarrow$ Greeting Cards	Magazine & Candy Bar	Greeting Cards
32	3	4.05	11.35	1.67	2.80	3333.0	Greeting Cards $\Rightarrow$ Magazine & Candy Bar	Greeting Cards	Magazine & Candy Bar
33	3	17.10	45.86	1.67	2.68	3333.0	Magazine & Greeting Cards $\Rightarrow$ Candy Bar	Magazine & Greeting Cards	Candy Bar
34	3	17.10	43.33	1.37	2.53	2744.0	Toothpaste & Magazine $\Rightarrow$ Candy Bar	Toothpaste & Magazine	Candy Bar
35	3	17.10	41.07	1.32	2.40	2635.0	Toothpaste & Greeting Cards $\Rightarrow$ Candy Bar	Toothpaste & Greeting Cards	Candy Bar
36	3	14.69	33.12	1.32	2.25	2635.0	Toothpaste & Candy Bar $\Rightarrow$ Greeting Cards	Toothpaste & Candy Bar	Greeting Cards
37	3	16.04	33.85	1.37	2.11	2744.0	Magazine & Candy Bar $\Rightarrow$ Toothpaste	Magazine & Candy Bar	Toothpaste
38	3	16.04	30.18	1.32	1.88	2635.0	Greeting Cards & Candy Bar $\Rightarrow$ Toothpaste	Greeting Cards & Candy Bar	Toothpaste
39	2	14.69	25.53	4.37	1.74	8732.0	Candy Bar $\Rightarrow$ Greeting Cards	Candy Bar	Greeting Cards
40	2	17.10	29.72	4.37	1.74	8732.0	Greeting Cards $\Rightarrow$ Candy Bar	Greeting Cards	Candy Bar
41	3	24.13	38.17	1.67	1.58	3333.0	Greeting Cards & Candy Bar $\Rightarrow$ Magazine	Greeting Cards & Candy Bar	Magazine
42	2	14.69	21.67	2.92	1.48	5848.0	Pencils $\Rightarrow$ Greeting Cards	Pencils	Greeting Cards
43	2	13.49	19.91	2.92	1.48	5848.0	Greeting Cards $\Rightarrow$ Pencils	Greeting Cards	Pencils
44	2	16.04	23.26	3.98	1.45	7956.0	Candy Bar $\Rightarrow$ Toothpaste	Candy Bar	Toothpaste
45	2	17.10	24.80	3.98	1.45	7956.0	Toothpaste $\Rightarrow$ Candy Bar	Toothpaste	Candy Bar
46	2	17.10	24.47	3.30	1.43	6603.0	Pencils $\Rightarrow$ Candy Bar	Pencils	Candy Bar
47	2	13.49	19.31	3.30	1.43	6603.0	Candy Bar $\Rightarrow$ Pencils	Candy Bar	Pencils
48	3	24.13	34.49	1.37	1.43	2744.0	Toothpaste & Candy Bar $\Rightarrow$ Magazine	Toothpaste & Candy Bar	Magazine
49	2	14.69	20.00	3.21	1.36	6416.0	Toothpaste $\Rightarrow$ Greeting Cards	Toothpaste	Greeting Cards
50	2	16.04	21.84	3.21	1.36	6416.0	Greeting Cards $\Rightarrow$ Toothpaste	Greeting Cards	Toothpaste
51	2	24.13	28.14	1.65	1.17	3291.0	Photo Processi $\Rightarrow$ Magazine	Photo Processi	Magazine
52	2	13.49	15.31	2.46	1.13	4912.0	Toothpaste $\Rightarrow$ Pencils	Toothpaste	Pencils
53	2	16.04	18.20	2.46	1.13	4912.0	Pencils $\Rightarrow$ Toothpaste	Pencils	Toothpaste
54									
55									

Figure 16: Output log showing Association Report for all rules.

**4.2. You are particularly interested in products that individuals purchase when they come to your store to buy Candy Bar. How many rules are in the subset? Based on the rules, what are the other products these individuals are most likely to purchase? (add screenshots here)**



Output							
22	Association Report						
23							
24							
25		Expected					
26		Confidence	Confidence	Support		Transaction	
27	Relations	(%)	(%)	(%)	Lift	Count	Rule
28							
29	2	6.74	24.26	2.18	3.60	4364.0	Perfume ==> Toothbrush
30	2	9.00	32.40	2.18	3.60	4364.0	Toothbrush ==> Perfume
31	3	14.69	41.11	1.67	2.80	3333.0	Magazine & Candy Bar ==> Greeting Cards
32	3	4.05	11.35	1.67	2.80	3333.0	Greeting Cards ==> Magazine & Candy Bar
33	3	17.10	45.86	1.67	2.68	3333.0	Magazine & Greeting Cards ==> Candy Bar
34	3	17.10	43.33	1.37	2.53	2744.0	Toothpaste & Magazine ==> Candy Bar
35	3	17.10	41.07	1.32	2.40	2635.0	Toothpaste & Greeting Cards ==> Candy Bar
36	3	14.69	33.12	1.32	2.25	2635.0	Toothpaste & Candy Bar ==> Greeting Cards
37	3	16.04	33.85	1.37	2.11	2744.0	Magazine & Candy Bar ==> Toothpaste
38	3	16.04	30.18	1.32	1.88	2635.0	Greeting Cards & Candy Bar ==> Toothpaste
39	2	14.69	25.53	4.37	1.74	8732.0	Candy Bar ==> Greeting Cards
40	2	17.10	29.72	4.37	1.74	8732.0	Greeting Cards ==> Candy Bar
41	3	24.13	38.17	1.67	1.58	3333.0	Greeting Cards & Candy Bar ==> Magazine
42	2	14.69	21.67	2.92	1.48	5848.0	Pencils ==> Greeting Cards
43	2	13.49	19.91	2.92	1.48	5848.0	Greeting Cards ==> Pencils
44	2	16.04	23.26	3.98	1.45	7956.0	Candy Bar ==> Toothpaste
45	2	17.10	24.80	3.98	1.45	7956.0	Toothpaste ==> Candy Bar
46	2	17.10	24.47	3.30	1.43	6603.0	Pencils ==> Candy Bar
47	2	13.49	19.31	3.30	1.43	6603.0	Candy Bar ==> Pencils
48	3	24.13	34.49	1.37	1.43	2744.0	Toothpaste & Candy Bar ==> Magazine
49	2	14.69	20.00	3.21	1.36	6416.0	Toothpaste ==> Greeting Cards
50	2	16.04	21.84	3.21	1.36	6416.0	Greeting Cards ==> Toothpaste
51	2	24.13	28.14	1.65	1.17	3291.0	Photo Processi ==> Magazine
52	2	13.49	15.31	2.46	1.13	4912.0	Toothpaste ==> Pencils
53	2	16.04	18.20	2.46	1.13	4912.0	Pencils ==> Toothpaste

**Figure 17: Output log highlighting rules for Candy Bar**

There are 3 rules that exist for customers that come to the store to purchase a Candy Bar. This is only for customers who bought Candy Bars, not customers who bought “Candy Bars & another”.

Candy Bar ==> Greeting Cards  
Candy Bar ==> Toothpaste  
Candy Bar ==> Pencils

Customers who purchase a Candy Bar are most likely to purchase Greeting Cards, Toothpaste, or Pencils.

## Part 3: Text Mining

### Task 5 Text Mining

Given a set of documents, apply SAS Text Miner.

Create a new diagram and a new data source node using the data set **EMCODE\_TRAN**, in the same project.

**Answer the following questions:**

**Perform text mining (clustering of documents in table view or in tree view).**

**Examine the results of text mining.**

```
11
12 Variable Summary
13
14           Measurement   Frequency
15 Role      Level        Count
16
17 TEXT      NOMINAL       1
18 URL       NOMINAL       1
19
20
21 *-----*
22 * Score Output
23 *-----*
24
25
26
27 The FREQ Procedure
28
29           Frequency      Percent      Cumulative      Cumulative
30 cluster      Frequency      Percent      Frequency      Percent
31 -----
32           1           7      41.18           7      41.18
33           2          10      58.82          17     100.00
34
```

**Figure 18: Output log of Text Cluster node**

Clusters	
Cluster ID	Descriptive Terms
1	isby +fighter +operation +war +fighter operation' +review air +include +photo security technology ...
2	bluetooth 'bluetooth wireless technology' +device wireless +phone +user +access mobile +address +issue data information ...

**Figure 19: Descriptive Terms for the Clusters**

**State how many clusters were generated?**

We generated 3 clusters using Expectation-Maximization.

**Name each cluster meaningfully according to the terms that appear in the clusters.**

Cluster 1 has been renamed to Fighter Combat. Cluster 2 has been renamed to Wireless Technology.

**Identify the first six high frequent terms.**

Term	Frequency
security	18
device / devices	17
bluetooth	16
fighter / fighters	16
phone / phones	14
wireless	11

**Table 1: The six most frequent terms.**

## Part 4: Web Mining

### Task 6 Web Mining

You have been provided with a log file in SAS format. This was originally a text file and was processed with the steps required for web usage mining as explained the lecture.

Create a new diagram and a new data source node using the data set **`SORT_USERS_DATA`**, in the same project.

Your task in this part is to apply a data mining operation such as classification or clustering or association mining to the pre-processed data set. Choose a data mining technique and analyse the mining results.

**Answer the following questions:**

#### **Rationale and execution of the data mining operation**

We decided to use Clustering to determine the frequency of visits to different websites based on users requests. Clustering was chosen because the data set was best suited towards clustering just like most web mining data sets.

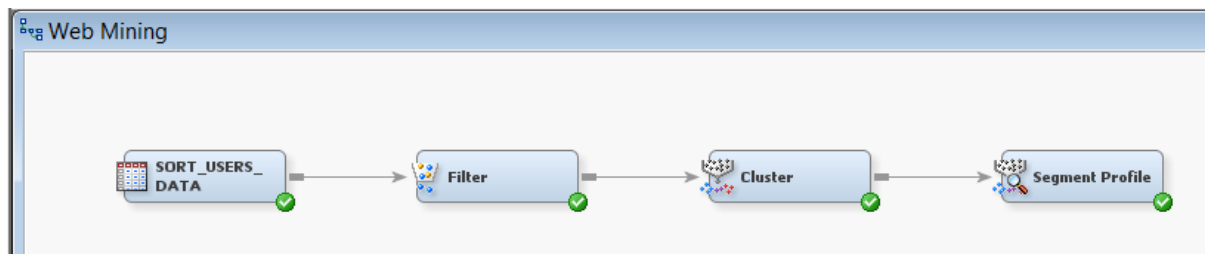
To begin with we rejected the user\_id and session variables. This was because we used the IP variable in place of the user\_id and the session wasn't needed. We then ran the data through a filter applying it to all data sets and not just training data. Once fed into the cluster node we used the request variable as our primary data type for the clustering node.

Our results have returned 7 segments. The segments are grouped by the users requests e.g. one cluster is dedicated to all requests associated with `"/richlands/"` while another cluster is dedicated to all requests for `"/"`. The frequency of the clusters was relatively even across the board.

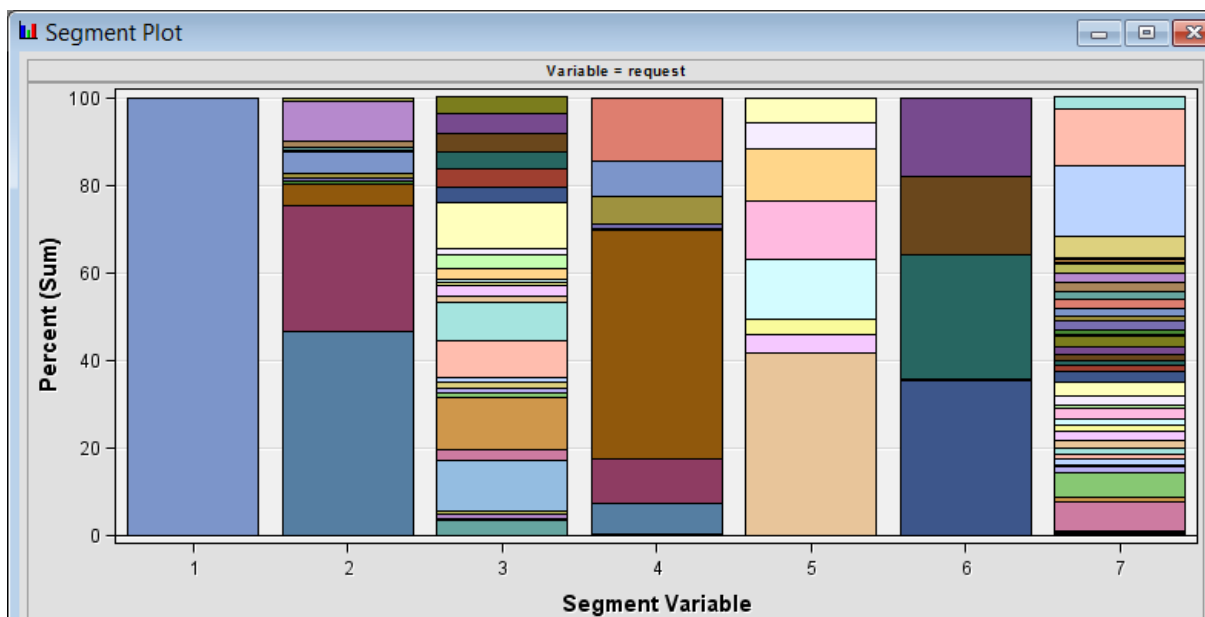
#### **Discussion and Applicability of findings of the method**

These mining applications can provide numerous real-world benefits. By analysing the web usage of users and the location-specific web pages they are loading we can assume each user's location and therefore relevant information to gather about not only the users that are browsing the web site but also the general habits of people from that area.

Possible applications could include specifically targeted experiences for individual users as well as doing more analysis on the overall trends that users from a particular suburb might engage in. We found data was easiest to cluster based on the location-specific requests as more likely than not we can assume someone who is browsing a web page that is related to a suburb is from that suburb or is around that suburb; the location was most relevant out of all of the others.



**Figure 20: Final diagram for Part 4: Web Mining**



**Figure 21: Segment Plot results taken from Web Mining Cluster node**

Clustering Criterion	Maximum Relative Change in Cluster Seeds	Improvement in Clustering Criterion	Segment Id	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster
0.132675	0	0	1	814	0	9.55E-15	7	0.44774
0.132675	0	0	2	848	0.138471	0.253951	3	0.469809
0.132675	0	0	3	706	0.122935	0.206486	2	0.469809
0.132675	0	0	4	993	0.159906	0.299904	3	0.513646
0.132675	0	0	5	680	0.113983	0.193486	7	0.407228
0.132675	0	0	6	1034	0.173282	0.256067	5	0.518181
0.132675	0	0	7	667	0.116087	0.201044	5	0.407228

**Figure 22: Mean Statistics taken from Web Mining Cluster node**