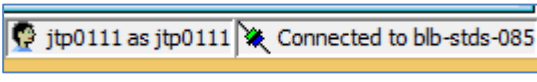


Jonathan Perry
DSCI 4520 Final Exam Report

It is my understanding that this is a take-home exam I am supposed to work on just by myself, without any collaboration with my classmates or any other person. On my honor, I neither have received, nor will I give information and/or aid related to this exam, in a way that constitutes academic dishonesty.

Jonathan Perry, 12/7/14

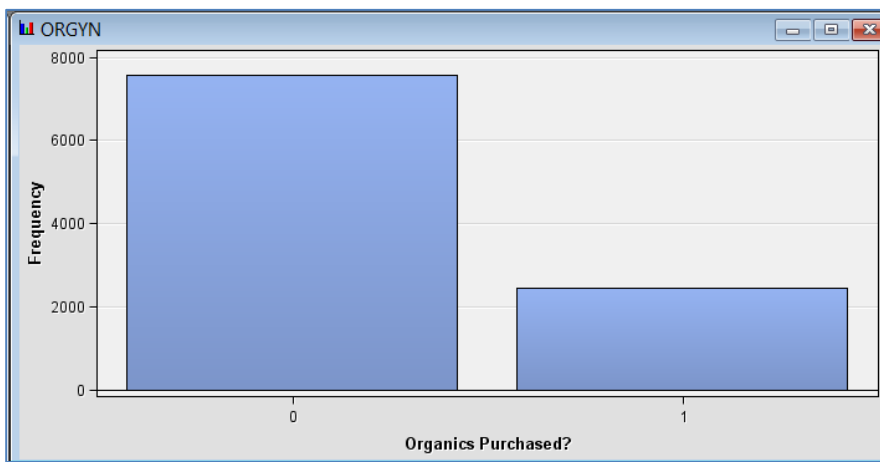
Problem 1

S1.1 

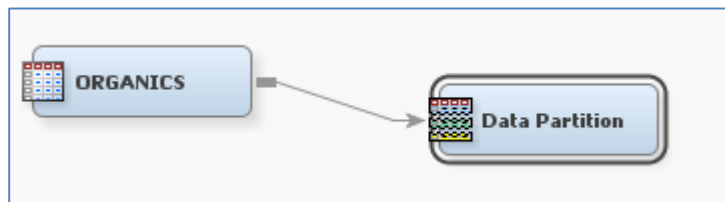
S1.2

```
1 1 The SAS System 18:52 Saturday, December 6, 2014
2
3 NOTE: Copyright (c) 2002-2012 by SAS Institute Inc., Cary, NC, USA.
4 NOTE: SAS (r) Proprietary Software 9.4 (TS1M1)
5 Licensed to UNIVERSITY OF NORTH TEXAS - T&R, Site 70080564.
6 NOTE: This session is executing on the X64_8PRO platform.
```

S1.3



S1.4

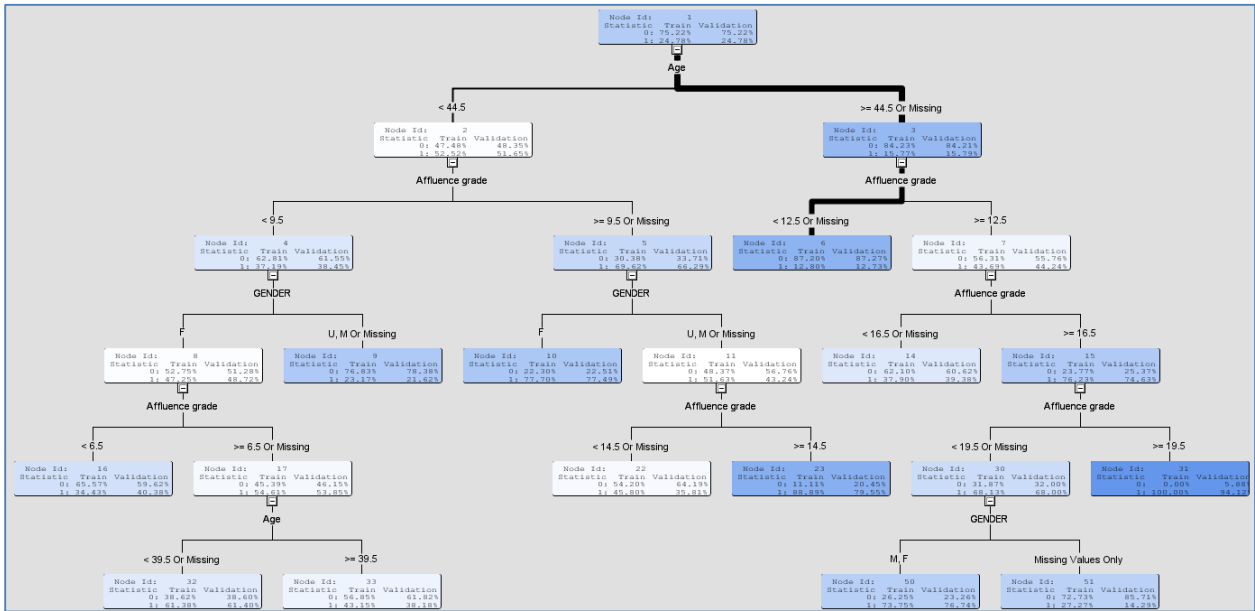


Q1.1 The proportion of individuals who purchased organic products is 2434, which is out of 10,000 so 24% of individuals have purchased organic products.

Q1.2 I don't have any reservations about any of the other variables in the data set that aren't already marked as rejected. I believe all other variables should be included as input variables that aren't already marked as rejected for my analysis.

Problem 2

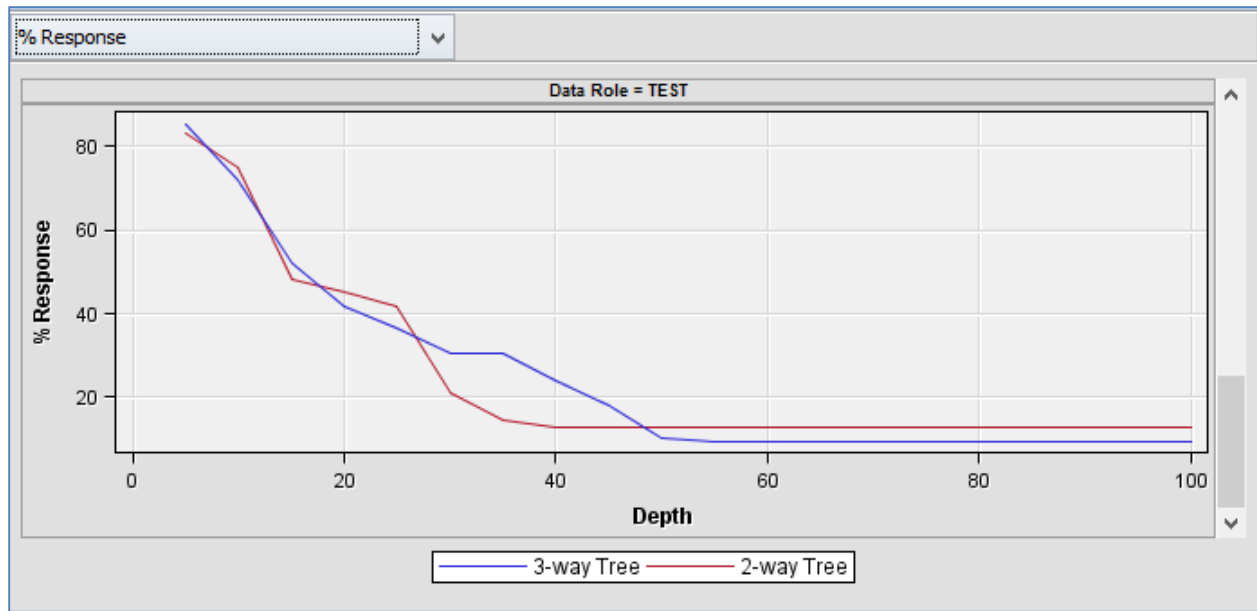
S2.1



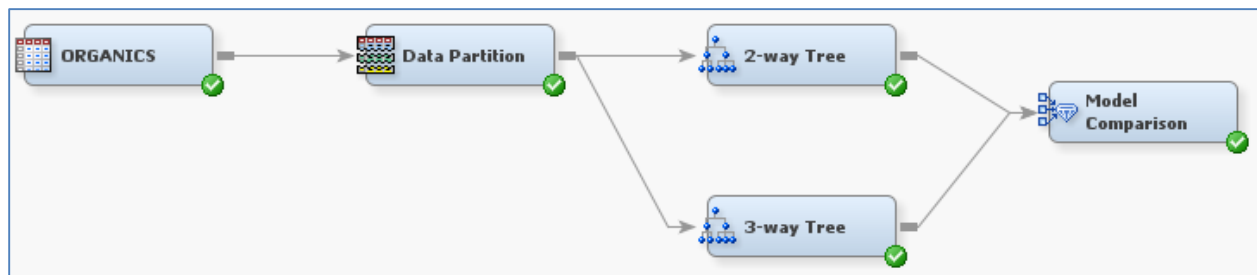
S2.2

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
AGE	Age	1	1.0000	1.0000	1.0000
AFFL	Affluence gr...	4	0.7830	0.6889	0.8798
GENDER		3	0.3497	0.4124	1.1794
BILL	Total Amou...	0	0.0000	0.0000	-
CLASS	Customer L...	0	0.0000	0.0000	-
NGROUP	Neighborho...	0	0.0000	0.0000	-
LTIME	Years as L...	0	0.0000	0.0000	-
TV_REG	TV Region	0	0.0000	0.0000	-
REGION	Geographic...	0	0.0000	0.0000	-

S2.3



S2.4



Q2.1 There are 3 leaves based on the validation data set.

Q2.2 The variable age was used for the first split.

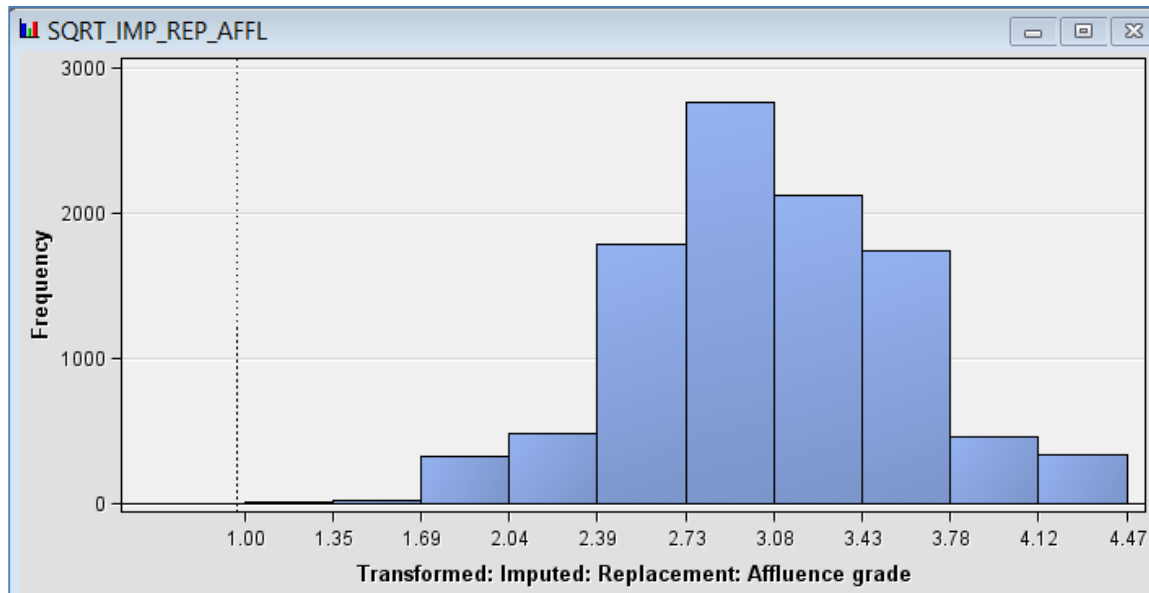
Q2.3 There are 6 leaves based on the validation data set.

Q2.4 The variables age was used for the first split.

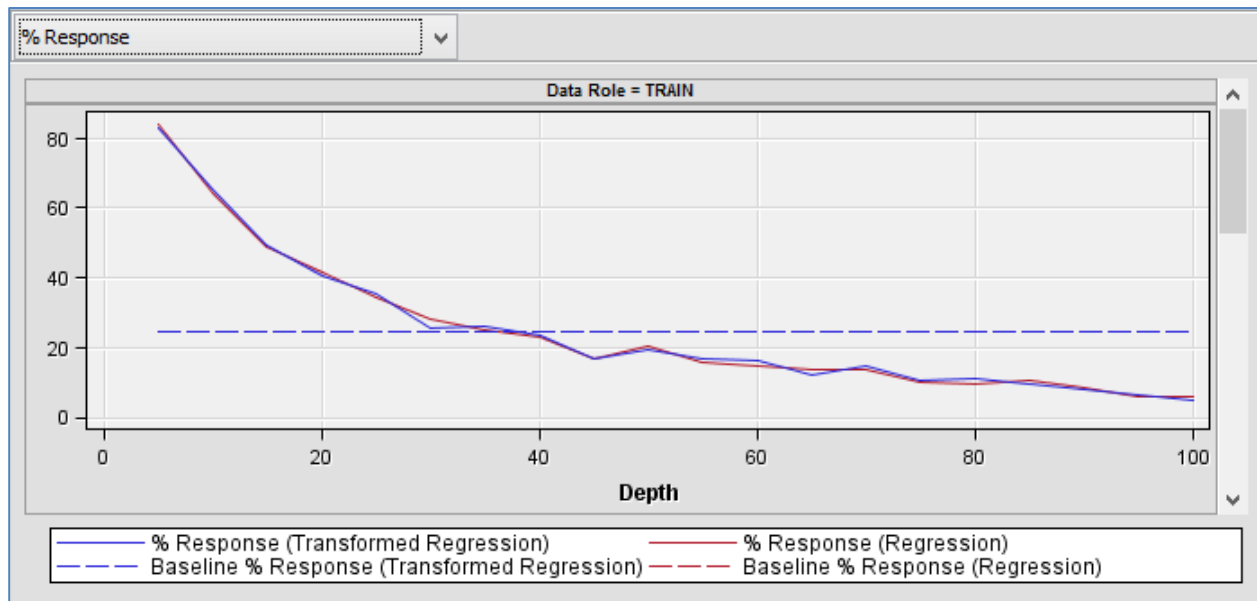
Q2.5 From the model comparison node, the 2-way Tree appears to be better decision tree model.

Problem 3

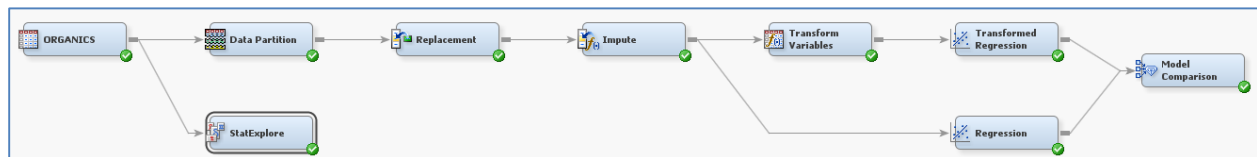
S3.1



S3.2



S3.3



Q3.1 Approximate percentages of missing values for class variables is 4116, and for interval variables is 2874 according to the StatExplore node. This means about 31% of our dataset has missing values. Imputation is needed for regression because SAS Enterprise Miner altogether ignores observations that contain missing values, which in the end will decrease the size of your training data set. The same is not true for decision trees; surrogate splitting rules enable you to use values of other input variables to perform a split for observations with missing values so it was not a mistake to skip imputation before generating the decision trees in problem 2.

Q3.2 IMP_REP_AFFL, IMP_REP_AGE, IMP_REP_GENDER are the variables included in the final model. IMP_REP_GENDERF appears to be the most important variable with an absolute coefficient of .2884.

Q3.3 Yes transformations of the data are warranted because you transform input variables to make the usual assumptions of regression more appropriate for the input data. Transformations on certain variables can be used to stabilize variance, improve additivity, remove nonlinearity, and counter non-normality. This can lead to a better model fit.

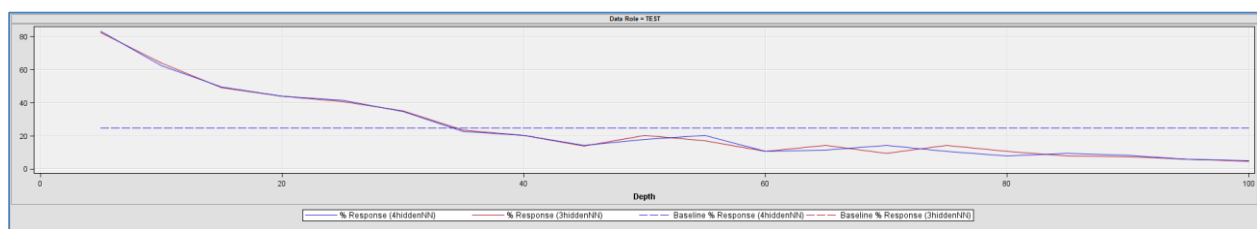
Q3.4 This appears to result in a more normal distribution as the variable is not as skewed due to the square root transformation.

Q3.5 IMP_REP_AGE, IMP_REP_GENDER, SQRT_IMP_REP_AFFL are the variables included in the final model. The variable that appears to be the best model is SQRT_IMP_REP_AFFL with an absolute coefficient of 1.6028.

Q3.6 According to the fit statistics window, the best model appears to be the transformed regression model.

Problem 4

S4.1



Q4.1 Imputation of missing values is needed in a neural network because it acts the same way as a regression model. Both models ignore missing values, which is bad because this will decrease the amount of data that you use in the models and lower their predictive power.

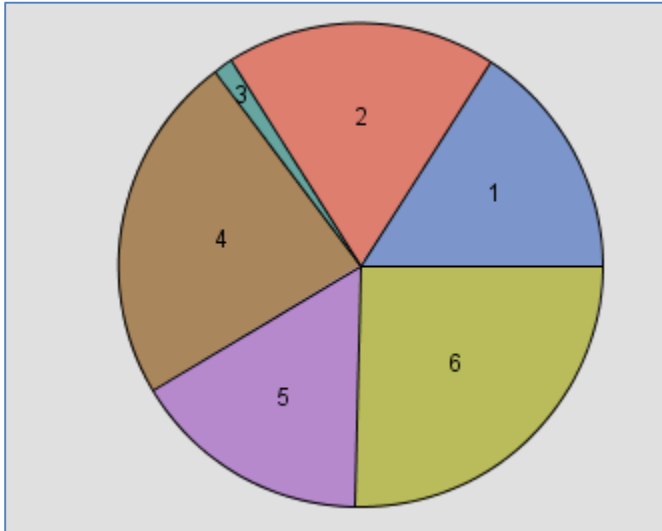
Q4.2 Not always, because neural network models can create transformations of inputs. But when you have input distributions with low kurtosis and low skewness this leads to a more stable model.

Q4.3 By default settings, 3 hidden neurons will be used.

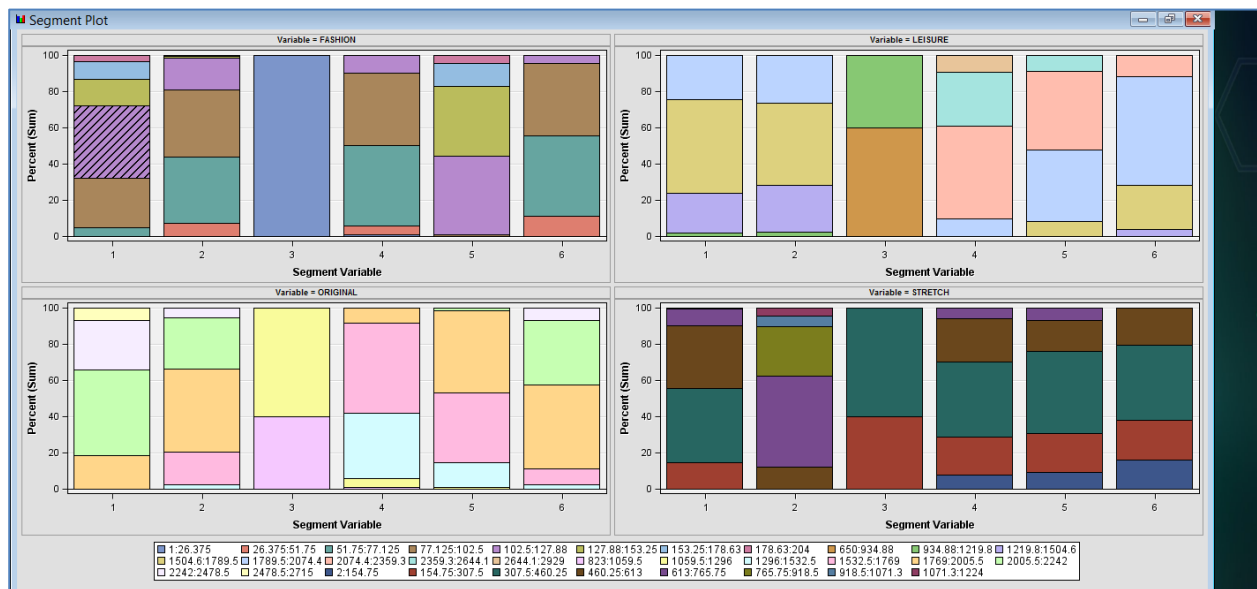
Q4.4 According to the fit statistics window, the best model appears to be the 4hiddenNN model.

Problem 5

S5.1



S5.1



S5.3

Mean Statistics												
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	FASHION	LEISURE	ORIGINAL	STRETCH
0.630825	0.006552	.	1	110	0.661401	2.849547	6	1.736146	117.5455	1646.718	2183.045	454.6545
0.630825	0.006552	.	2	124	0.655498	2.72785	1	1.974937	82.79839	1639.169	1925.202	752.7661
0.630825	0.006552	.	3	10	0.395138	1.039016	6	4.859655	4.2	840	1039.6	315.8
0.630825	0.006552	.	4	160	0.618889	2.525709	6	1.901759	78.11875	2314.669	1561.5	384.4563
0.630825	0.006552	.	5	111	0.644488	3.040225	4	1.913854	133.3784	2080.333	1731.045	365.2432
0.630825	0.006552	.	6	174	0.616007	2.460391	1	1.736146	74.86207	1875.592	1971.172	329.7471

Q5.1 You should standardize the data because K-means clustering is very sensitive to variable scale. When you have variables with very large values, they'll have a large effect on the formed clusters. So, when your data set contains variables that have different measurement scales you want to standardize the variables to a similar scale so we can correctly compare each variable.

Q5.2

Cluster 1 contains stores selling a higher-than-average number of original jeans.

Cluster 2 contains stores selling a higher-than-average number of stretch jeans.

Cluster 3 contains stores selling small numbers of all jean styles.

Cluster 4 contains stores selling a higher-than-average number of leisure jeans.

Cluster 5 contains stores selling a higher-than-average number of fashion jeans.

Cluster 6 contains stores selling a higher-than-average number of original jeans, but lower-than-average number of stretch and fashion.

Problem 6

S6.1

Support(%) ▼	Confidence(%)	Lift	Item 1	Item 2	Rule	Relations	Transpose Rule
4.2168	32.0470	3.5709	Greeting Cards	Candy Bar	Greeting Cards ==> Candy Bar	2Y	
4.2168	23.0120	1.5250	Candy Bar	Greeting Cards	Candy Bar ==> Greeting Cards	2Y	
4.1285	25.7753	1.4066	Toothpaste	Candy Bar	Toothpaste ==> Candy Bar	2Y	
4.1285	22.5301	1.4066	Candy Bar	Toothpaste	Candy Bar ==> Toothpaste	2Y	
3.1240	23.9020	1.3044	Pencils	Candy Bar	Pencils ==> Candy Bar	2Y	
2.6383	20.1858	1.3377	Pencils	Greeting Cards	Pencils ==> Greeting Cards	2Y	
2.1084	32.0470	3.5709	Toothbrush	Perfume	Toothbrush ==> Perfume	2Y	
2.1084	23.4932	3.5709	Perfume	Toothbrush	Perfume ==> Toothbrush	2Y	

S6.2

Confidence(%) ▼	Lift	Item 1	Item 2	Rule	Relations	Transpose Rule
32.0470	3.5709	Toothbrush	Perfume	Toothbrush ==> Perfume	2Y	
27.9444	1.5250	Greeting Cards	Candy Bar	Greeting Cards ==> Candy Bar	2Y	
25.7753	1.4066	Toothpaste	Candy Bar	Toothpaste ==> Candy Bar	2Y	
23.9020	1.3044	Pencils	Candy Bar	Pencils ==> Candy Bar	2Y	
23.4932	3.5709	Perfume	Toothbrush	Perfume ==> Toothbrush	2Y	
23.0120	1.5250	Candy Bar	Greeting Cards	Candy Bar ==> Greeting Cards	2Y	
22.5301	1.4066	Candy Bar	Toothpaste	Candy Bar ==> Toothpaste	2Y	
20.1858	1.3377	Pencils	Greeting Cards	Pencils ==> Greeting Cards	2Y	

S6.3

Lift ▼	Item 1	Item 2	Rule	Relations	Transpose Rule
3.5709	Perfume	Toothbrush	Perfume ==> Toothbrush	2Y	
3.5709	Toothbrush	Perfume	Toothbrush ==> Perfume	2Y	
1.5250	Greeting Cards	Candy Bar	Greeting Cards ==> Candy Bar	2Y	
1.5250	Candy Bar	Greeting Cards	Candy Bar ==> Greeting Cards	2Y	
1.4066	Toothpaste	Candy Bar	Toothpaste ==> Candy Bar	2Y	
1.4066	Candy Bar	Toothpaste	Candy Bar ==> Toothpaste	2Y	
1.3377	Pencils	Greeting Cards	Pencils ==> Greeting Cards	2Y	
1.3044	Pencils	Candy Bar	Pencils ==> Candy Bar	2Y	

Q6.1 The support (%) on the first row is 4.2168, support shows 4.22% of baskets had both greeting cards and candy bars.

Q6.2 The confidence (%) on the first row is 32.0470, confidence shows that 32.0470% of those who bought a tooth brush also bought perfume.

Q6.3 The lift value on the first row is 3.5709, this shows us that we get 3.5709 times more outcomes of the rule that perfume determines toothbrush by only contacting an amount of people based on the baseline percentage.