MBAD 6201

BUSINESS INTELLIGENCE AND ANALYTICS

PROJECT REPORT ON PREDICTIVE ANALYSIS USING SAS MINER

By:
VENKATA AKHILESH (800989790)

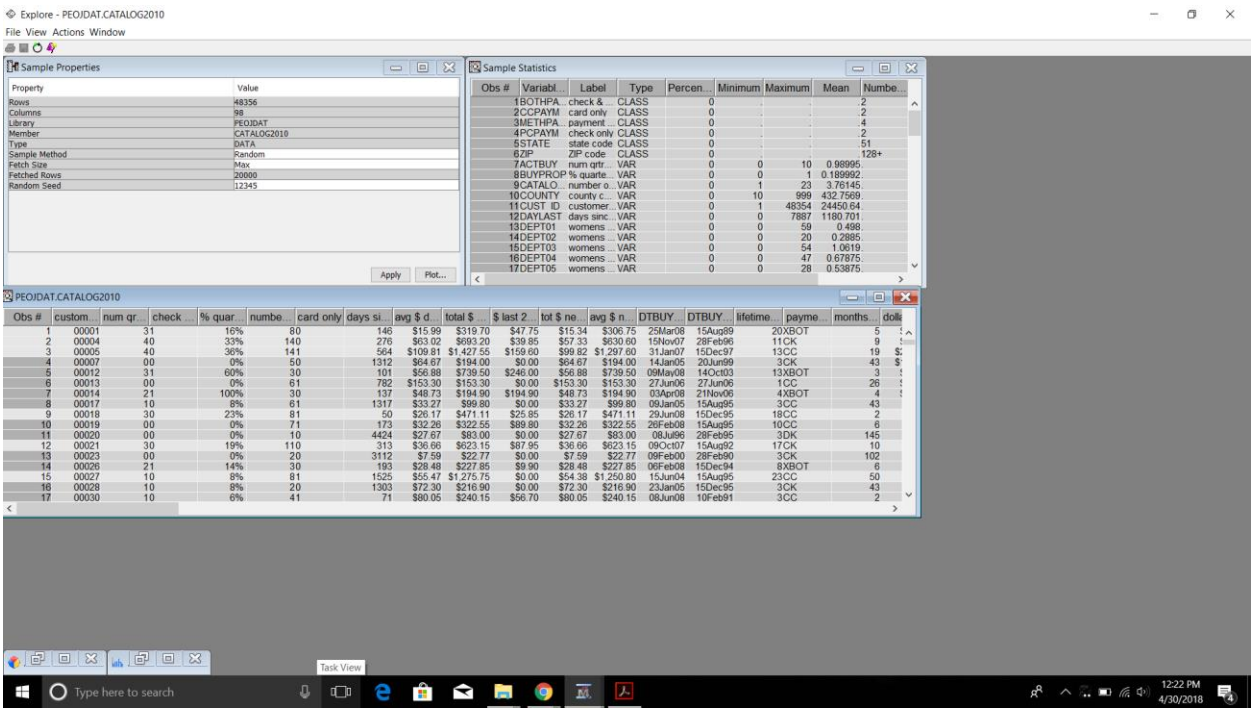
HARSHA VARDHAN GANTI (800989176)

Contents:

- Part 1 – Performing the Predictive Analysis – Exploratory. 3
- Task 1 – Exploratory Analysis of CATALOG2010 dataset. 5
- Task 2 – RFM analysis using PVA97NK dataset. 17
- Task 3 – Decision Trees with Organic Dataset. 28
- Task 4 – LOGIT. 37

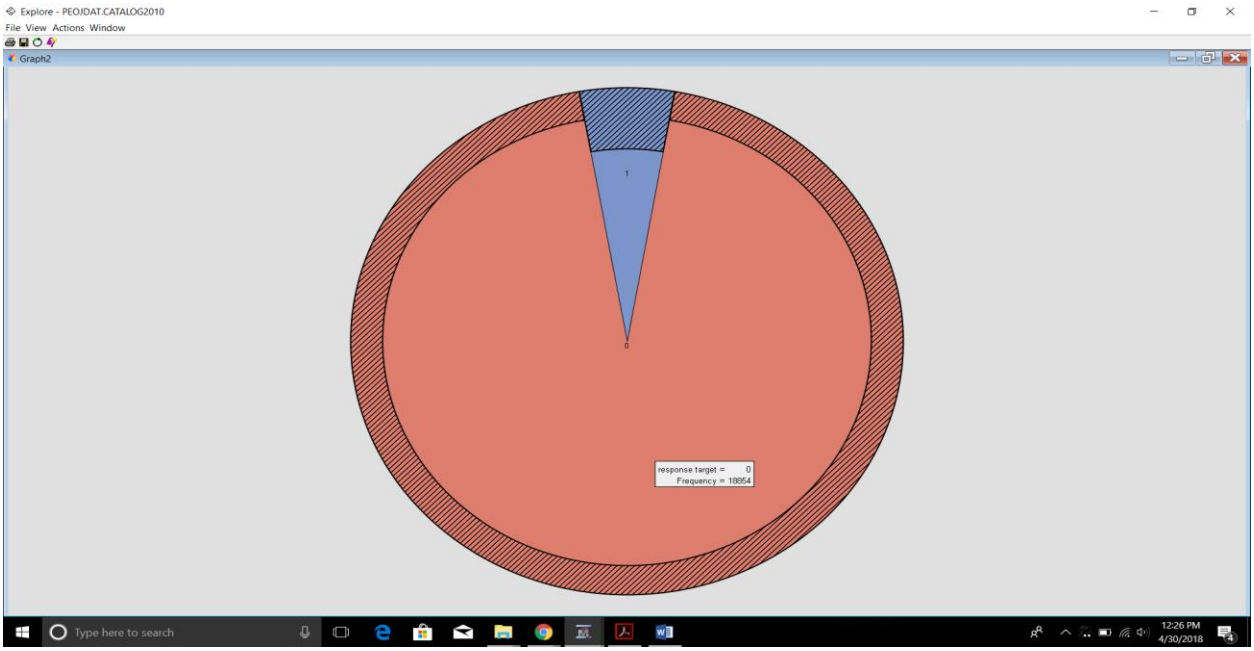
PART 1 - Predictive Analysis

After doing all the basic steps as instructed in the PDF, the next step is to explore data with graphs. To do this, we need to right click on the Catalog2010 data set and click on explore. Go to actions and select plot to plot various graphs according to the requirement.



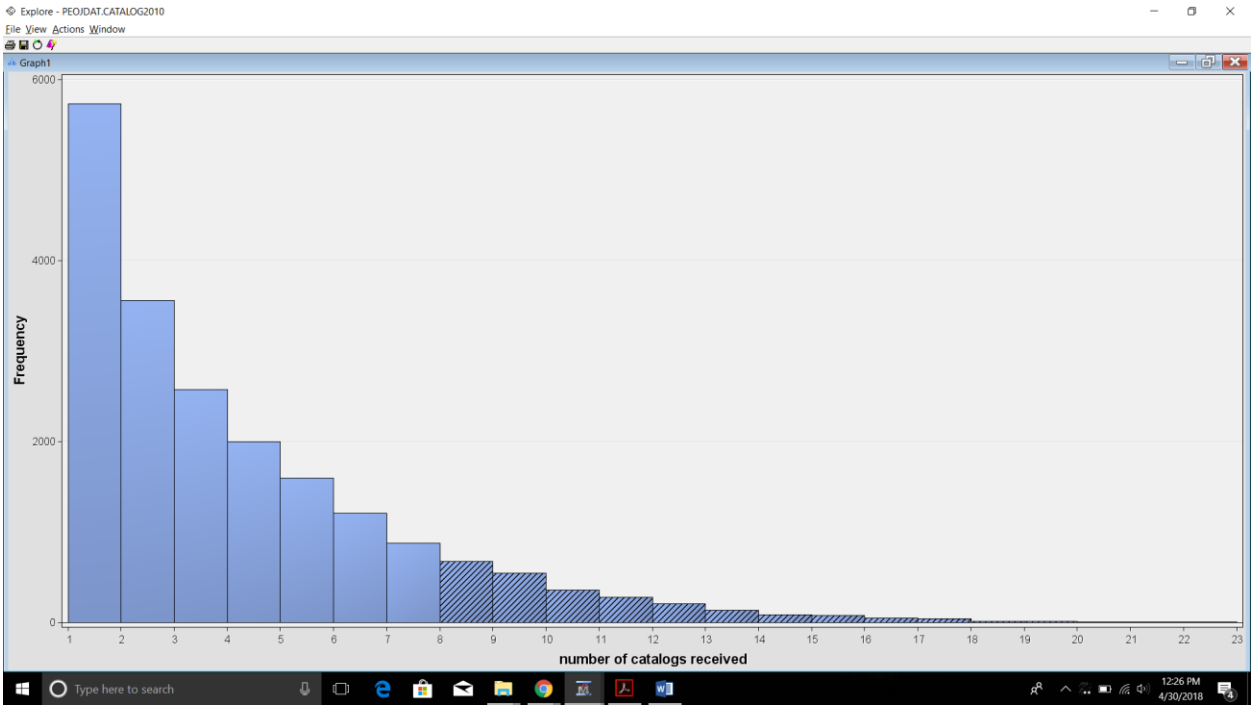
Creating a Pie Chart:

Sequence of steps to create pie chart is given below:



Steps to create a Histogram is shown below:

Select actions -> plot. Click on histogram. Select role -> X for CATALOGCNT variable and click finish.



When we select an area in one graph the entire region corresponding to that selected region will be highlighted in all the other plotted graphs as well.

TASK 1: Exploratory analysis of CATALOG2010 Dataset.

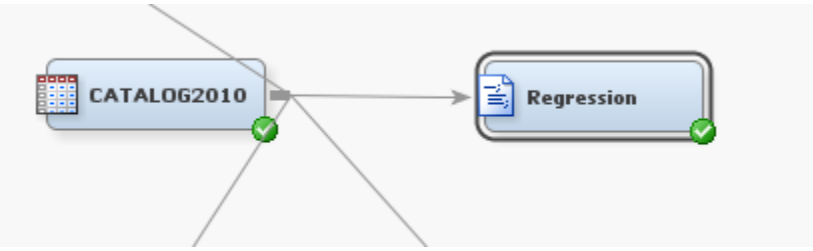
Perform analysis to find out:

1) **Which variables are highly correlated? If you find such variables you can suggest dimension reduction by dropping one of the variables.**

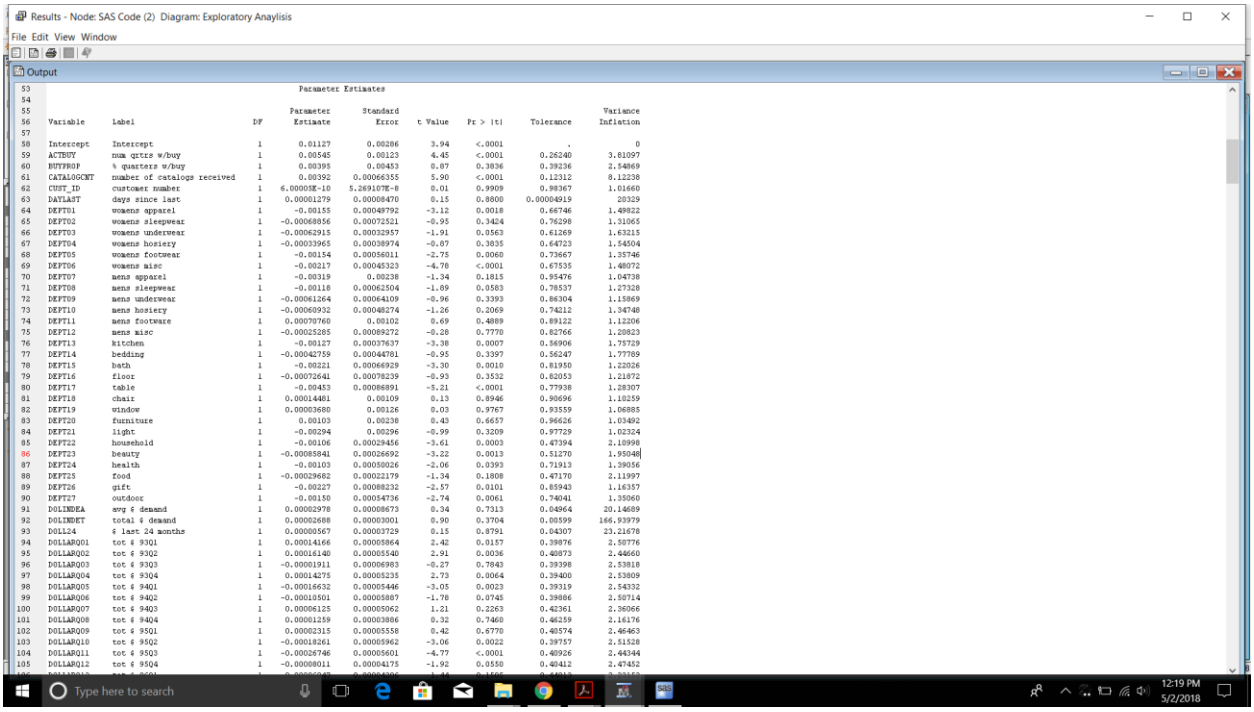
- To find out correlation among the variables, we used four methods as listed below:

a) Regression by using PROC Reg in SAS Code Node.
b) Variable Selection Method.
c) StatExplore block.
d) Correlation, by using the PROC CORR in SAS code node.

a) **Regression**



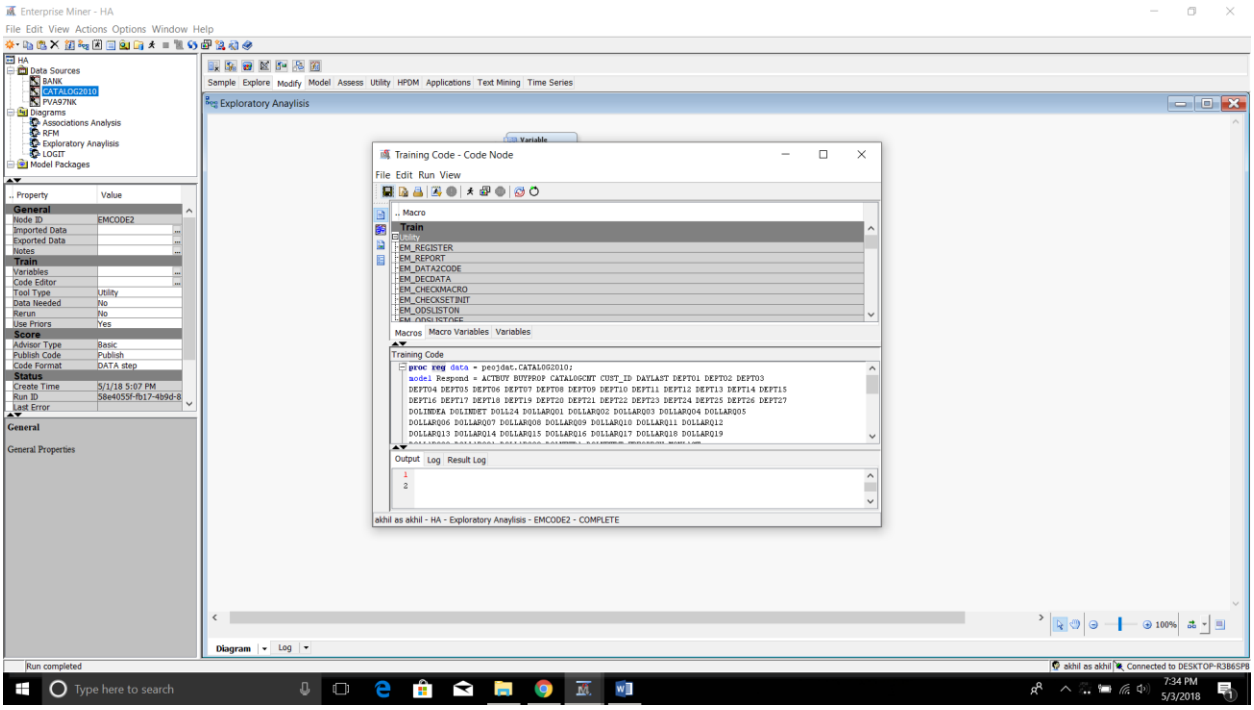
We performed regression to observe the correlation of the variables based on VIF values.



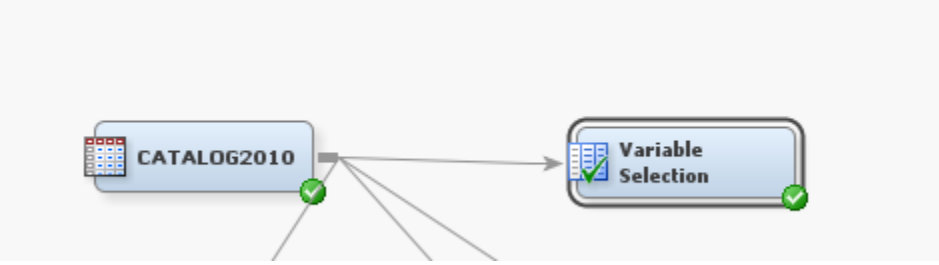
From the above obtained results summary, we can see that we have 9 variables that have a VIF value of greater than 10 and a VIF value of greater than 10 states that these variables are highly correlated.

➔ SAS Code for Regression:

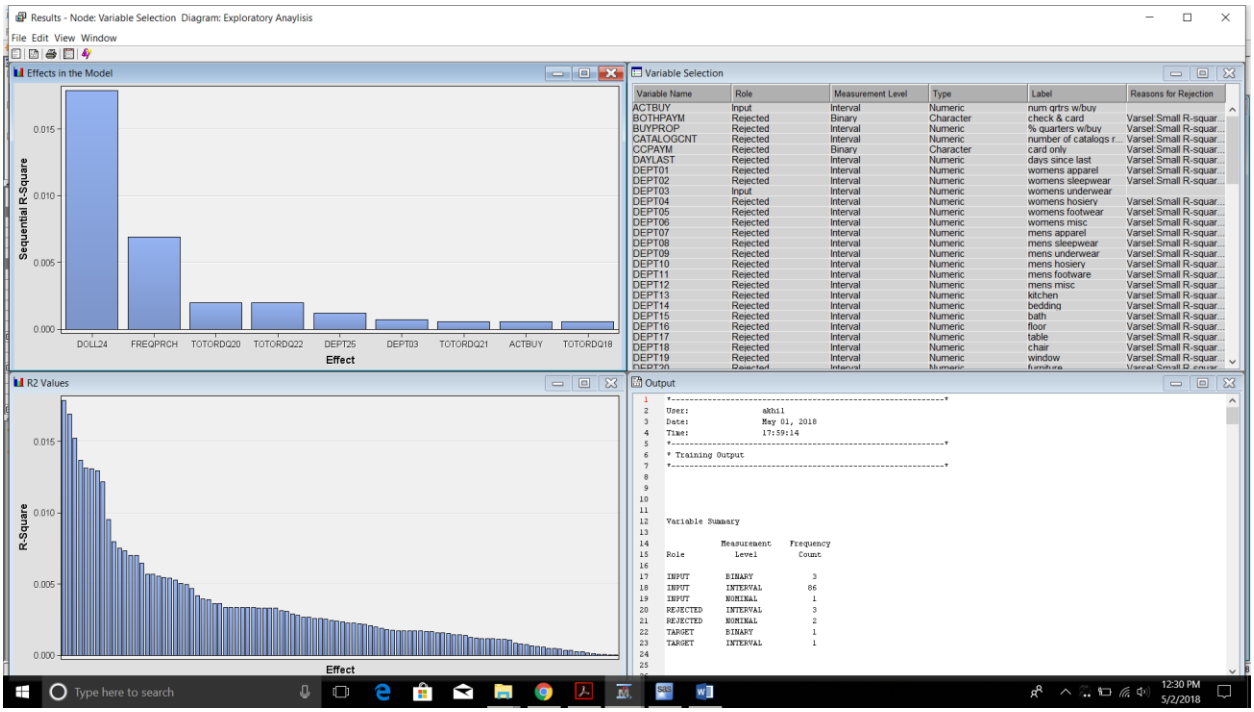
```
proc reg data = peojdat. CATALOG2010;

model Respond = ACTBUY BUYPROP CATALOGCNT CUST_ID DAYLAST DEPT01 DEPT02 DEPT03

DEPT04 DEPT05 DEPT06 DEPT07 DEPT08 DEPT09 DEPT10 DEPT11 DEPT12 DEPT13 DEPT14
DEPT15DEPT16 DEPT17 DEPT18 DEPT19 DEPT20 DEPT21 DEPT22 DEPT23 DEPT24 DEPT25
DEPT26 DEPT27 DOLINDEA DOLINDET DOLL24 DOLLARQ01 DOLLARQ02 DOLLARQ03 DOLLARQ04
DOLLARQ05 DOLLARQ06 DOLLARQ07 DOLLARQ08 DOLLARQ09 DOLLARQ10 DOLLARQ11
DOLLARQ12 DOLLARQ13 DOLLARQ14 DOLLARQ15 DOLLARQ16 DOLLARQ17 DOLLARQ18
DOLLARQ19 DOLLARQ20 DOLLARQ21 DOLLARQ22 DOLNETDA DOLNETDT FREQPRCH MONLAST
ORDERSIZE TENURE TOTORDQ01 TOTORDQ02 TOTORDQ03 TOTORDQ04 TOTORDQ05
TOTORDQ06 TOTORDQ07 TOTORDQ08 TOTORDQ09 TOTORDQ10 TOTORDQ11 TOTORDQ12
TOTORDQ13 TOTORDQ14 TOTORDQ15 TOTORDQ16 TOTORDQ17 TOTORDQ18 TOTORDQ19
TOTORDQ20 TOTORDQ21 TOTORDQ22 UNITSIDD UNITSLAP UNTLANPO

/ vif tol collin;

run;
```

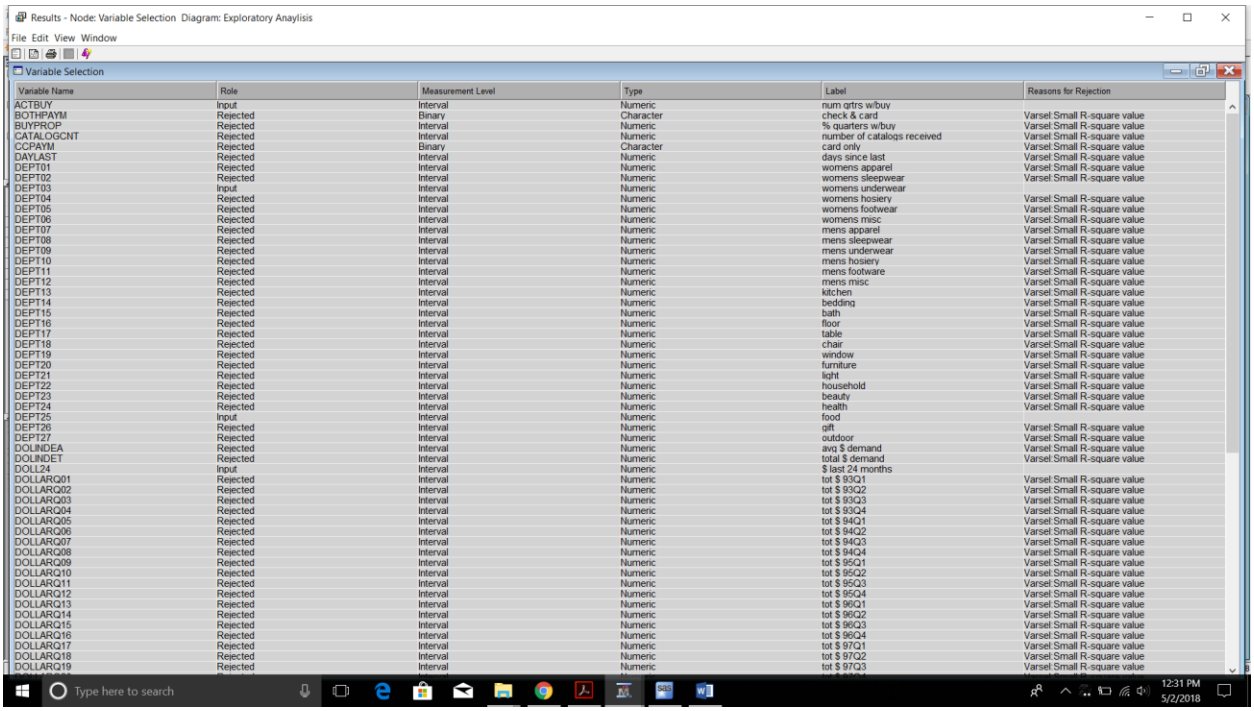


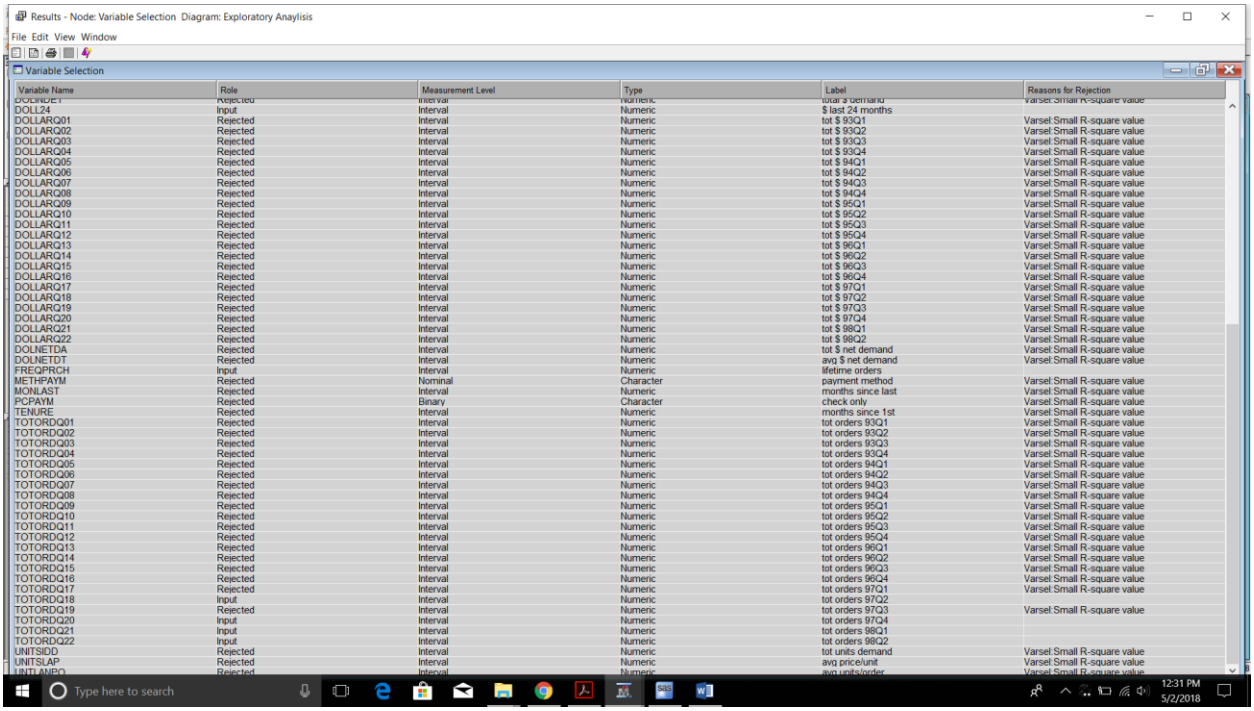
b) **Variable Selection Method**



- The variable separation node can be pulled out from the Explore block of the SEMMA tools. This node is then connected to the data source CATALOG2010.
- When we run the variable separation node, we can see the output results as posted below:

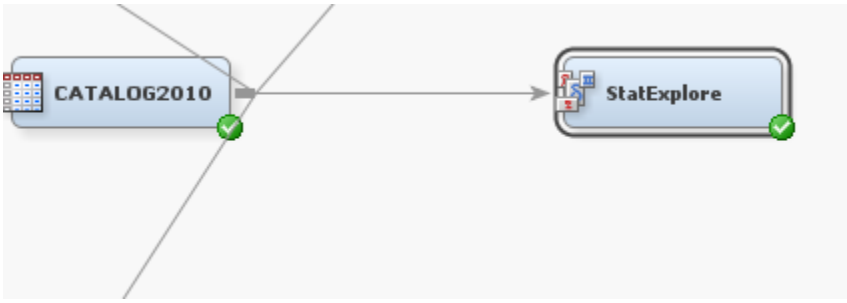




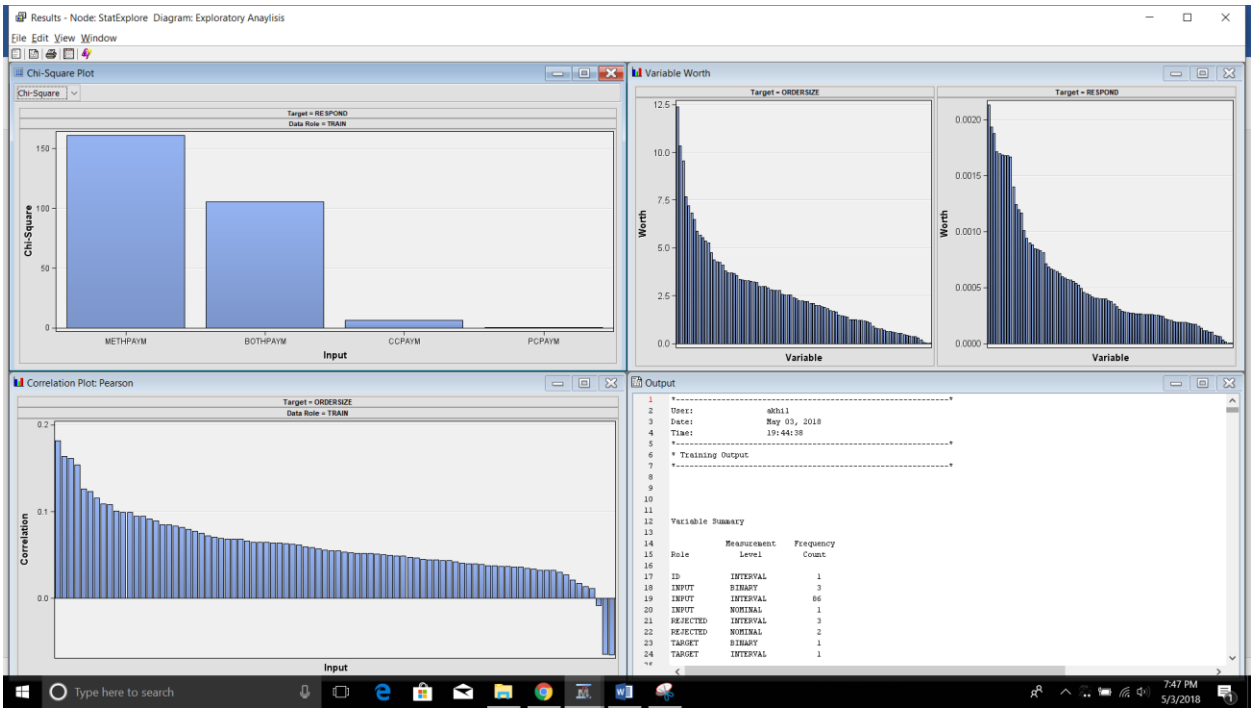


- The outputs for variable separation node are shown above.
- By using this method, we dropped down the number of acceptable variables to 7.
- The other variables that hold a rejected value, have high correlation with the target variables.

c) StatExplore method:

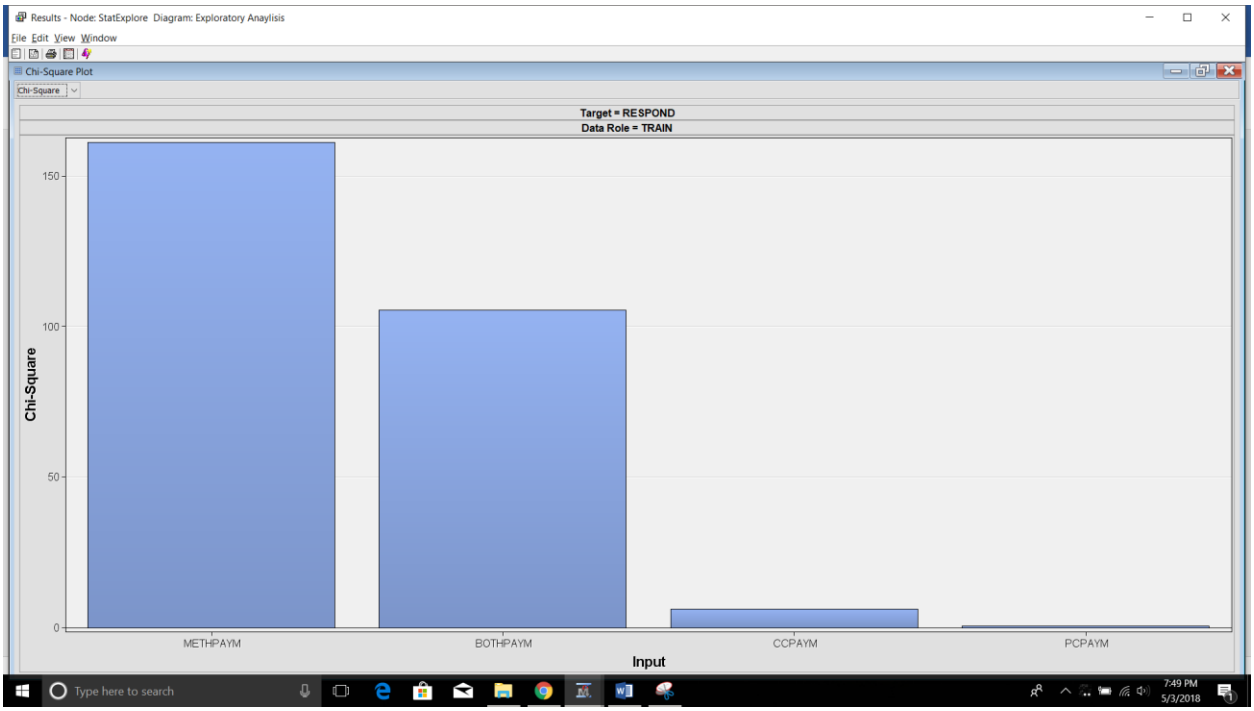


- The StatExplore block can be pulled out of the Explore tab that is a part of the SEMMA tools.
- This block is connected to the CATALOG2010 node as shown above.
- Then we run the StatExplore node to observe the results.
- The results are obtained as shown below:



The different tiles of the above results block are

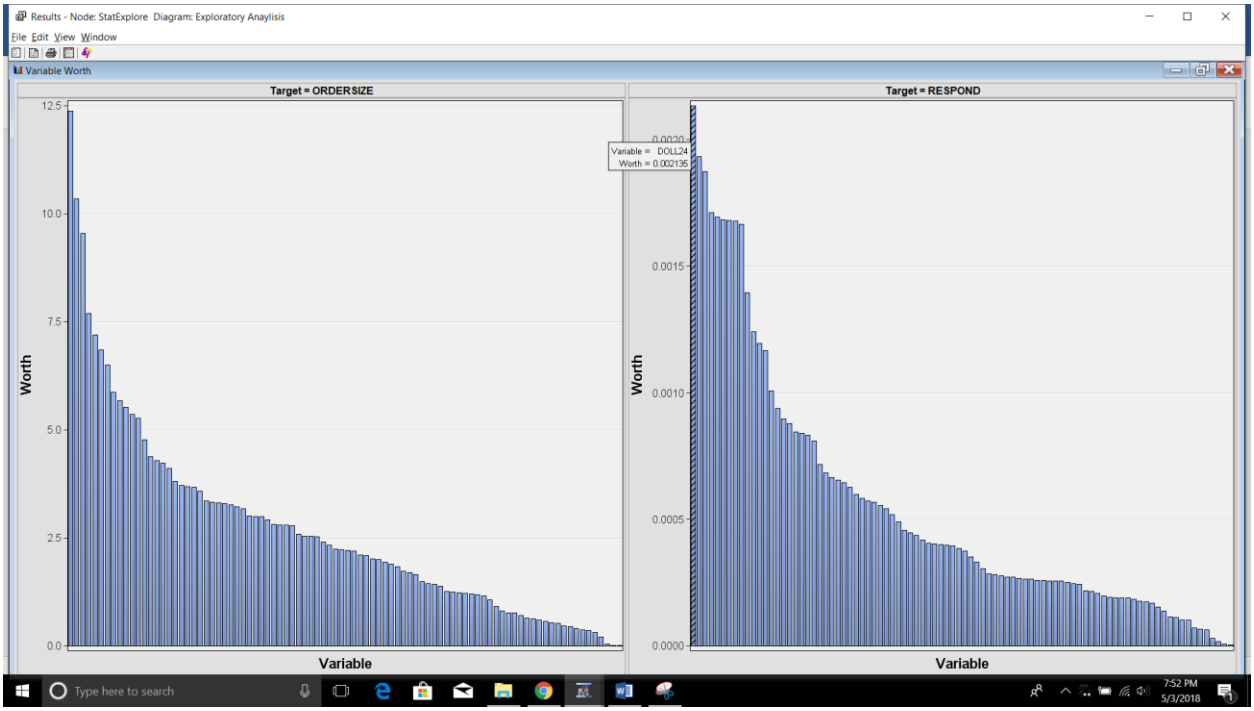
i) Chi-square plot.



ii) Variable Worth Plot:

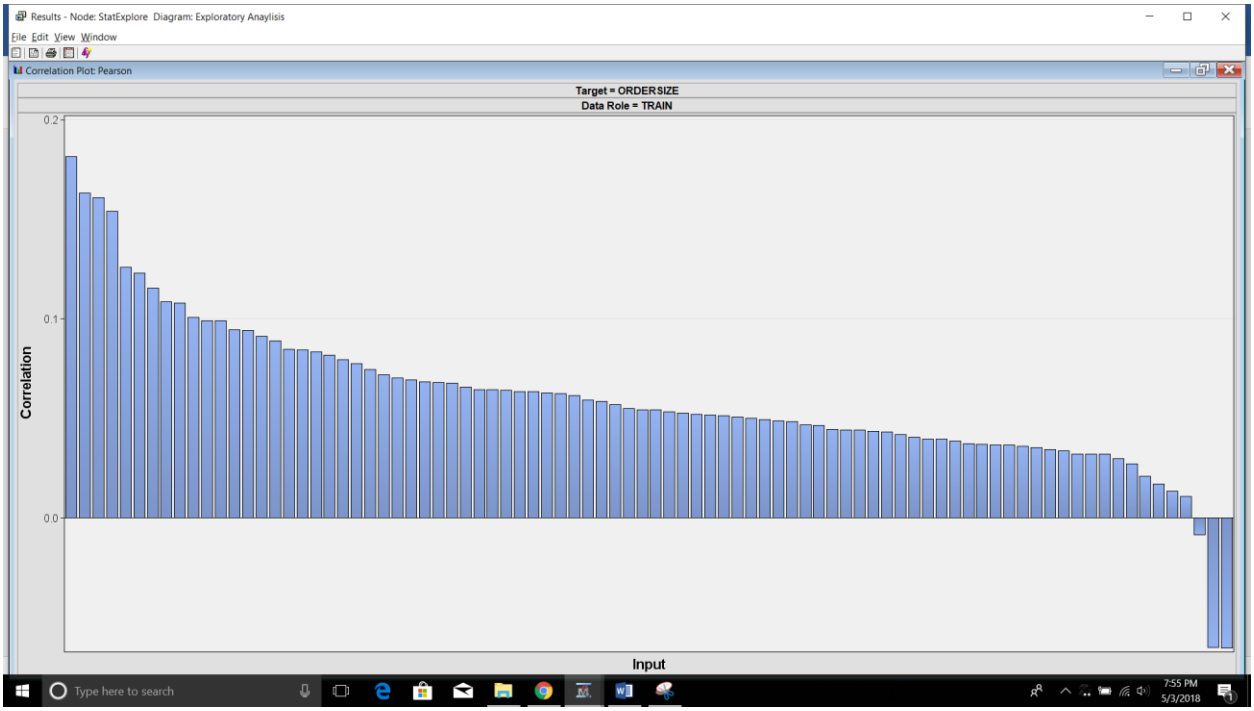
This plot orders the variables based on their true worth in predicting the target variables.

Here we have two targets ORDERSIZE and RESPOND. Hence, we obtained 2 different variable worth for both the targets.



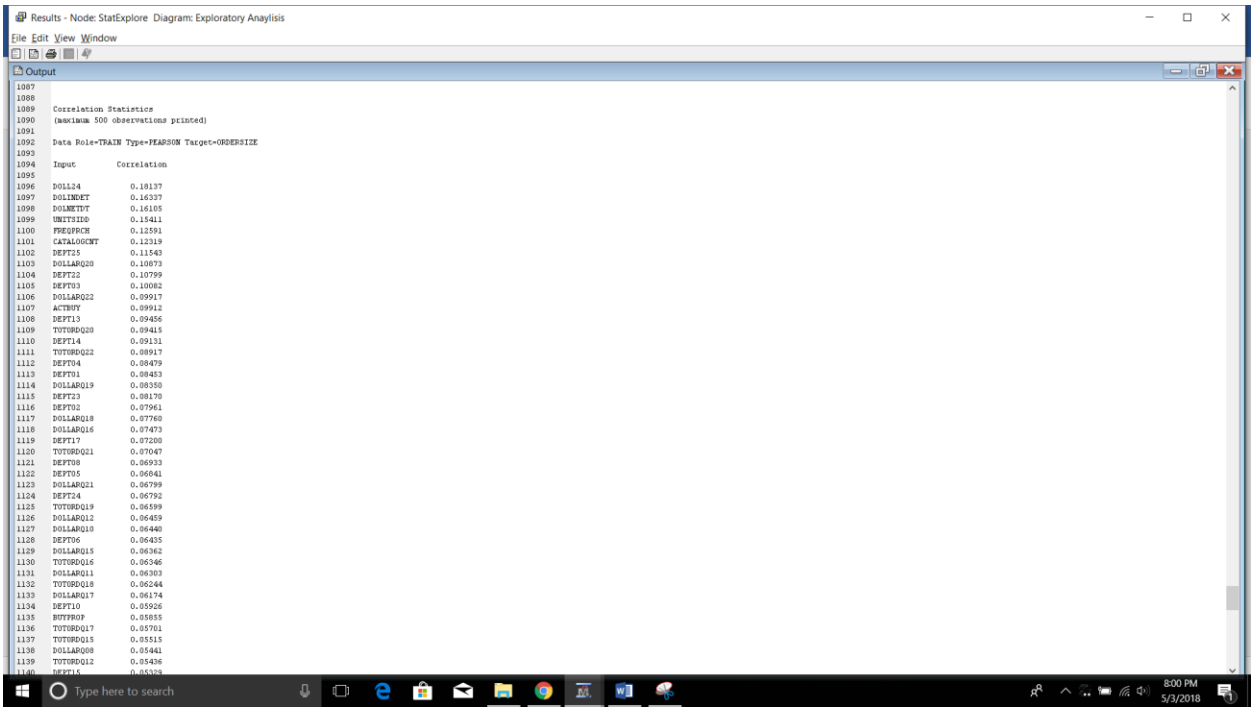
- By placing the curser on the longest bars of the graphs, we can see the variable name to which that bar belongs to.
- When the target variable is ORDERSIZE, DOLL24 has the highest worth.
- When the target variable is RESPOND, again DOLL24 happens to be the variable with highest worth.

iii) Correlation plot:



- From the above graph, we can infer a fact that for almost all the variables have positive correlation with the target variable (ORDERSIZE), with a few of them holding a value near to 0.2.
- However, there are these three variables UNITSLAP, MONLAST and DAYLAST, which have values less than 0. This implies that these 3 variables are negatively correlated with the ORDERSIZE.

iv) Analysis output:



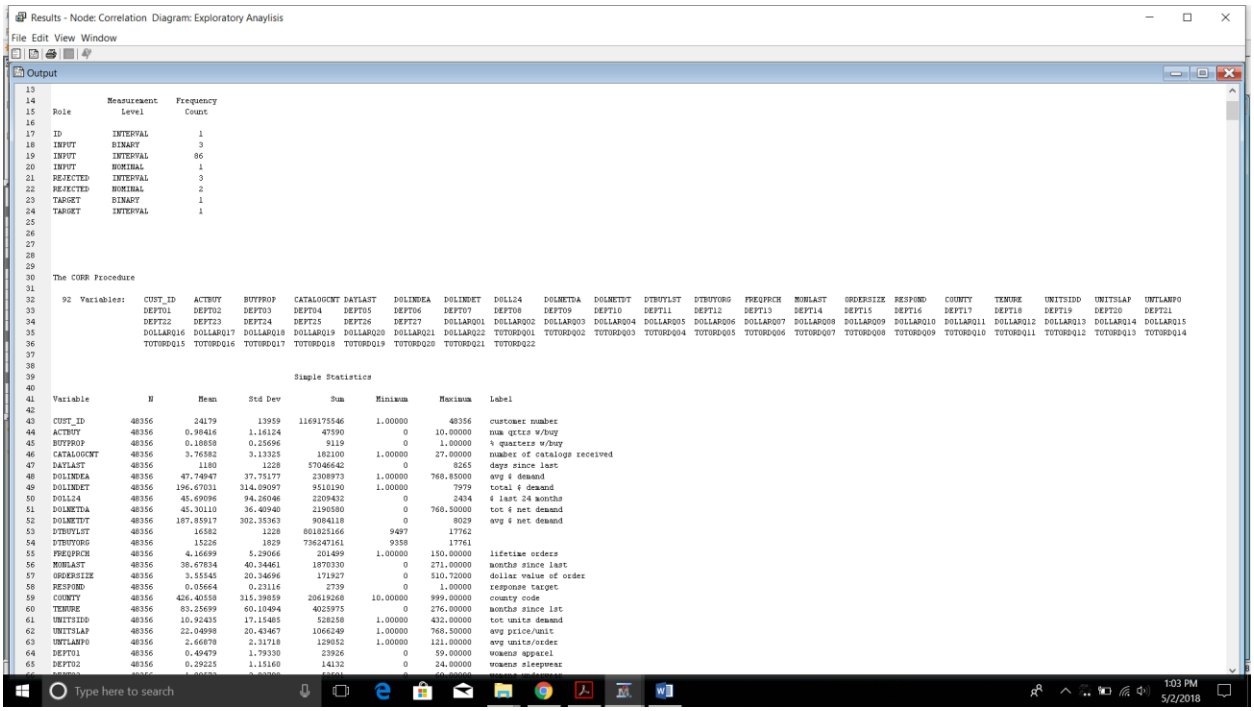
- In the above analysis result, we can see the variables and their corresponding correlation values.
- As we can observe, the DOLL24 variable has a high correlation value, compared to all the other variables. Thus, this variable can be removed for dimension reduction.

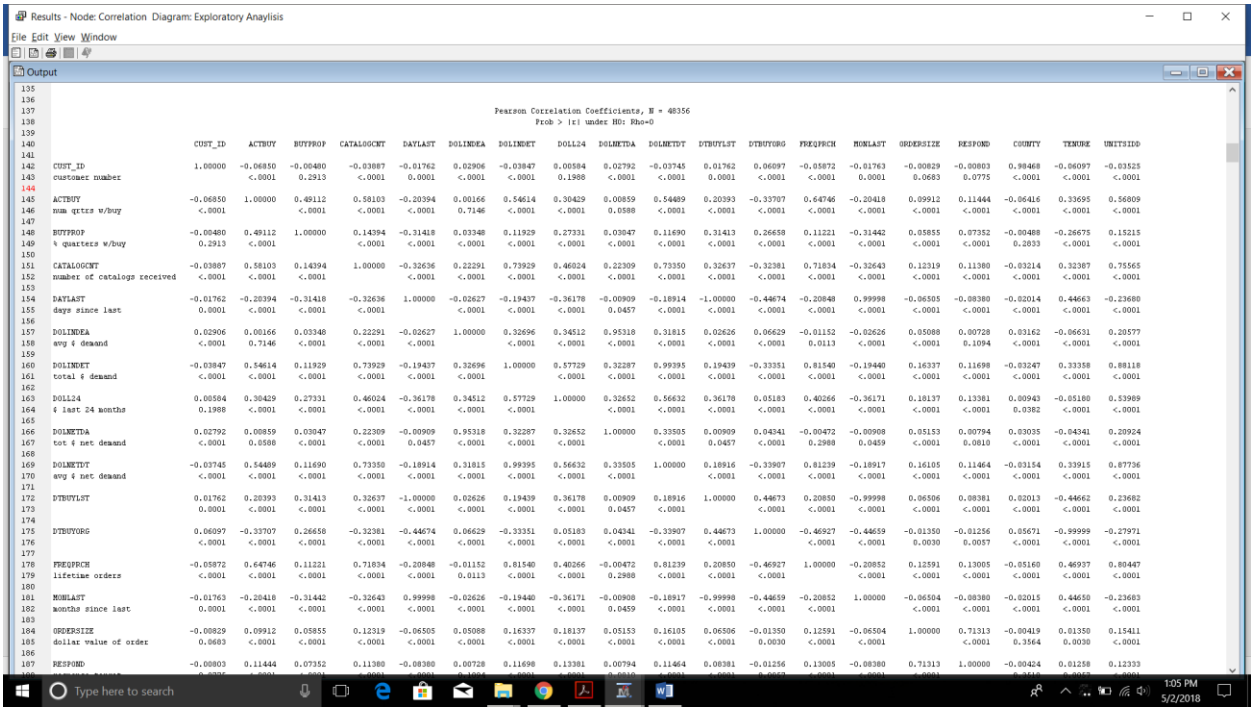
d) **CORRELATION method:**

- Correlation method is performed to observe the correlation of the variables.

- Here we used SAS code node from the utility block of the SEMMA tools.
- *SAS code for correlation:*

proc corr data = peojdat.CATALOG2010;
run;



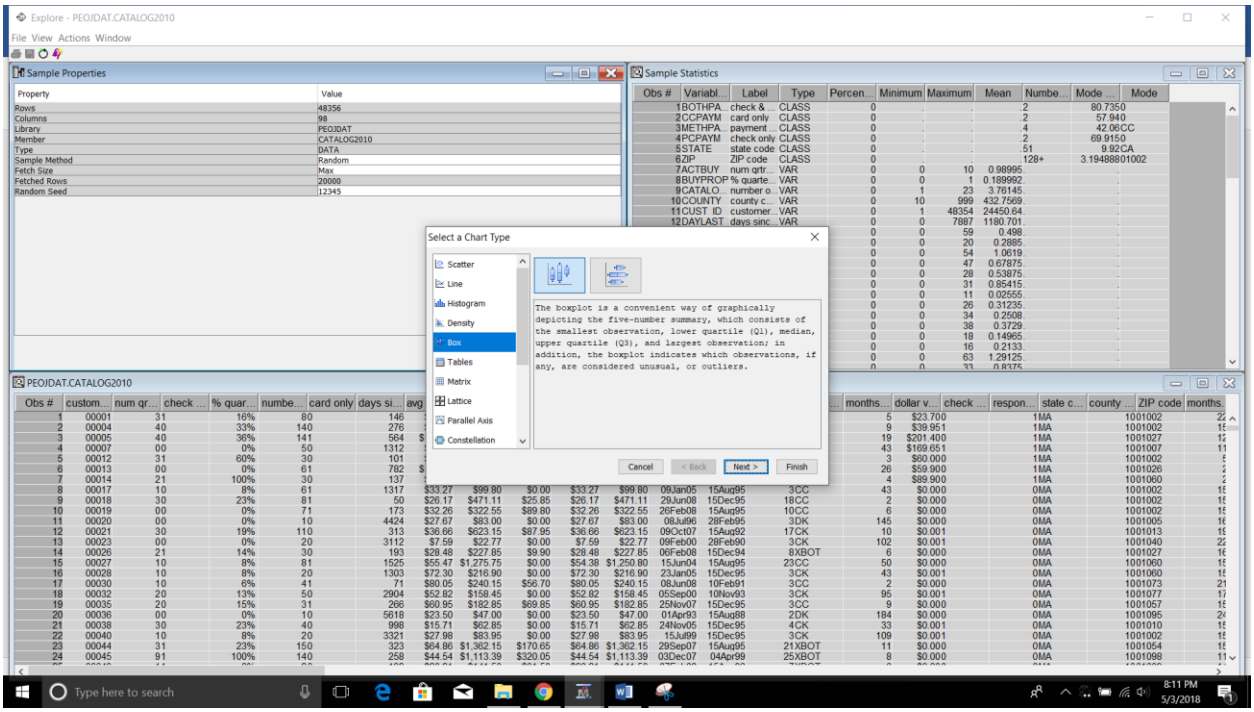


The correlation analysis is shown above.

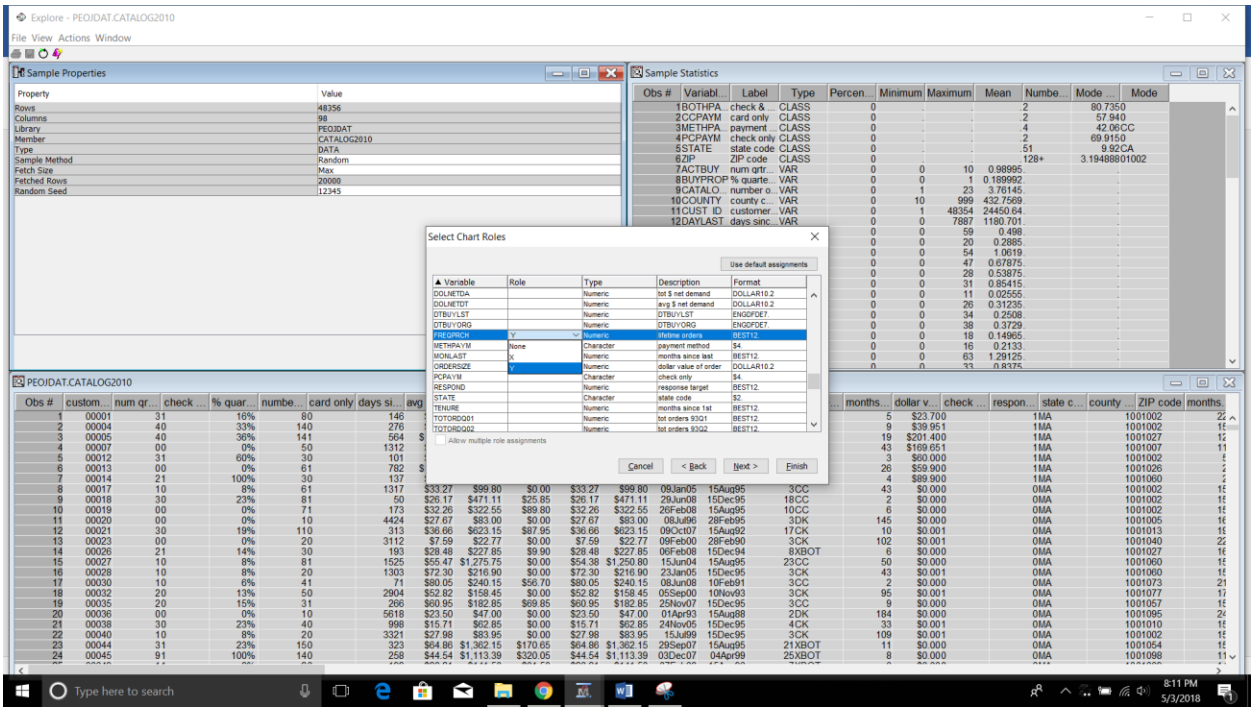
2) You can study if there are any outliers in the variables.

- The outliers can be determined by using a box plot.
- To plot the boxplot, we can follow steps as described here:

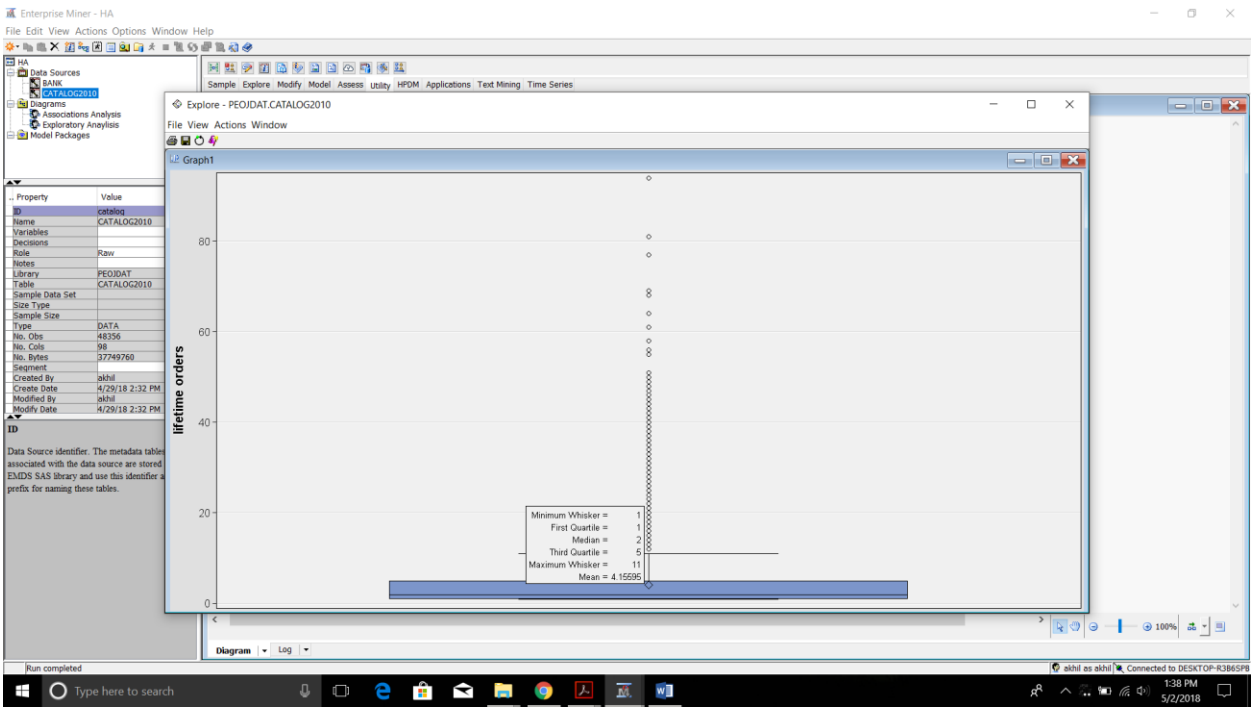
Right click on CATALOG2010 data set -> select Explore -> go to Actions -> select Plot -> search for Box option as shown below



- In this step, we assign the FREQPRCH variable to the Y axis and click finish.

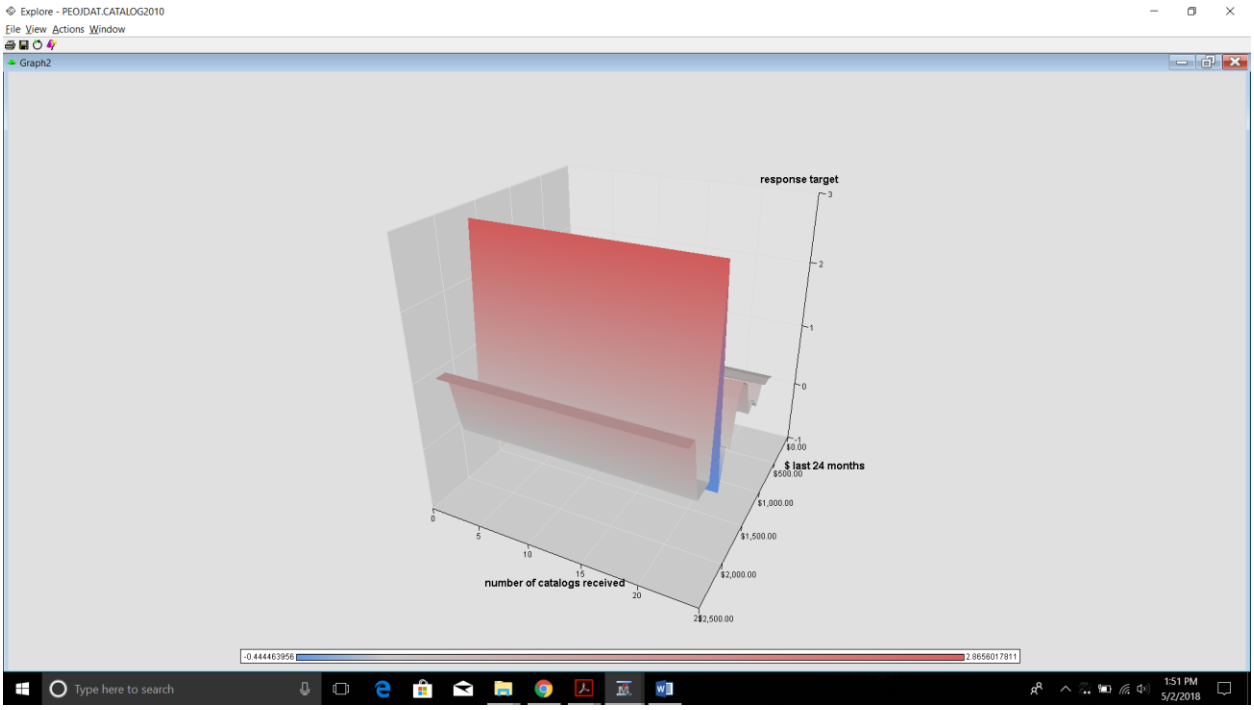


- The output with boxplot is shown below.



The above box plot analysis can be used to determine the outliers. From this we can say that for lifetime orders greater than the Maximum whisker value of 11 can be ignored as outliers.

3) 3D plot



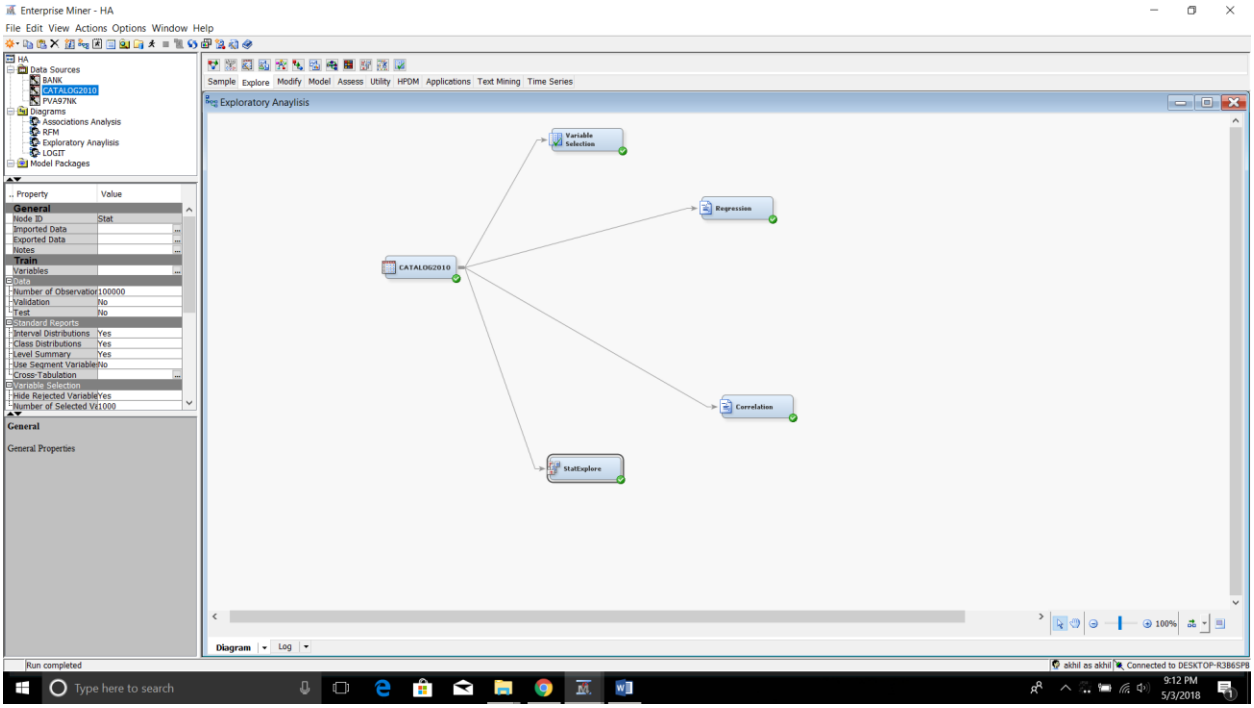
-As we hover around the 3-D plot we can see the values for different variables at each point.

-We can change the 3D plot based on desired variables.

-The red part of the 3D plot indicates the amount of money spent in the past 24 months is more and the blue part indicates that the amount of money spent is less.

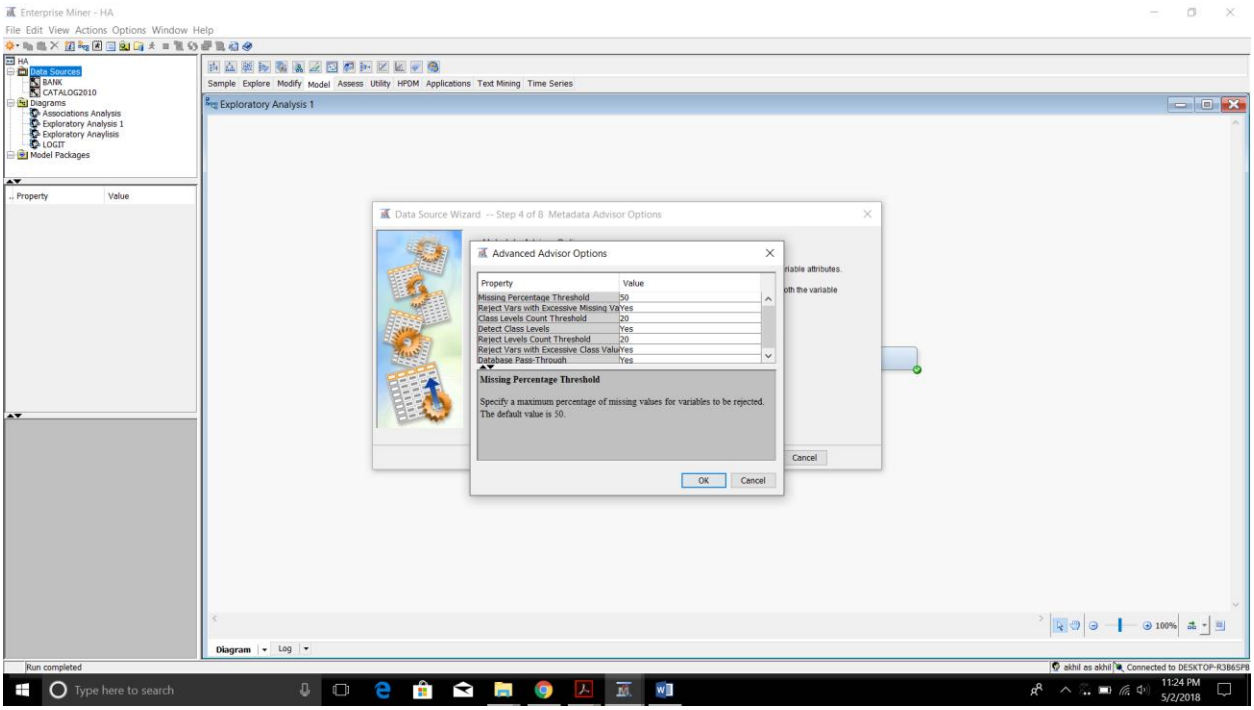
- This implies, as the number of catalogs received increases, the money spent also increases.

➔ The final Diagram

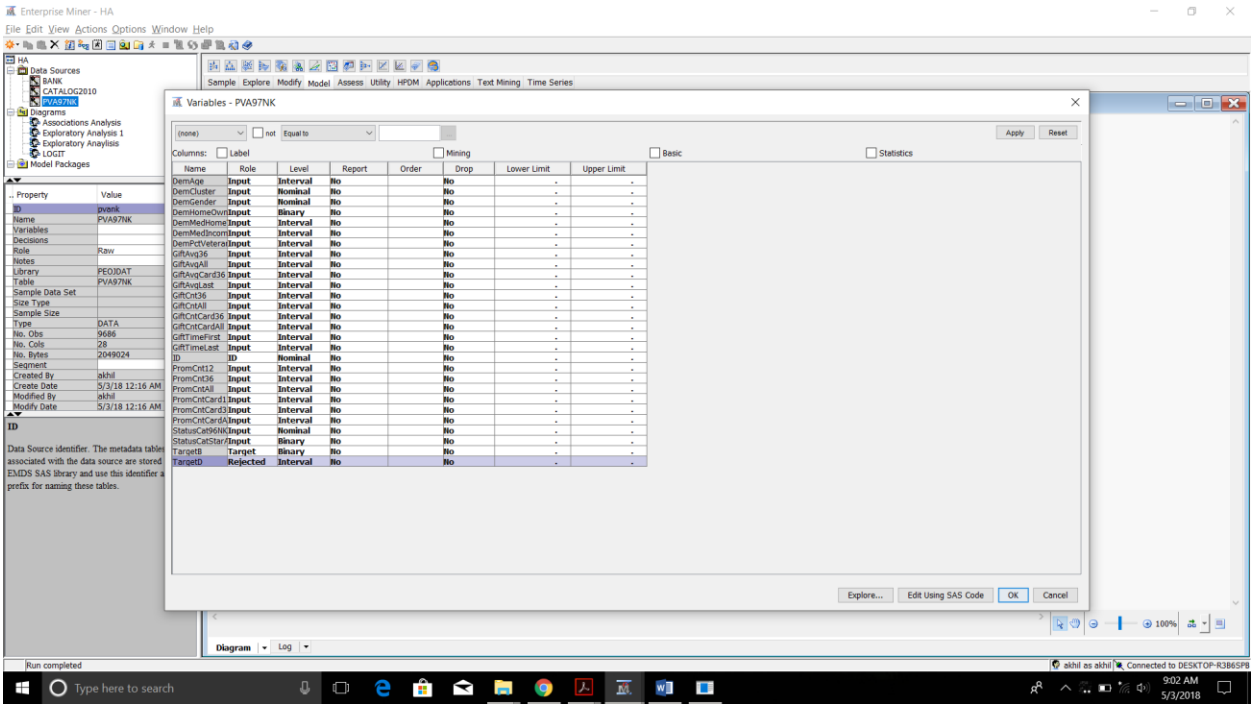


TASK 2 – RFM Analysis using PVA97NK

- Creating the PVA97NK data source.
- Changing class level count from 20 to 5.
- Changing the reject level from 20 to 80.

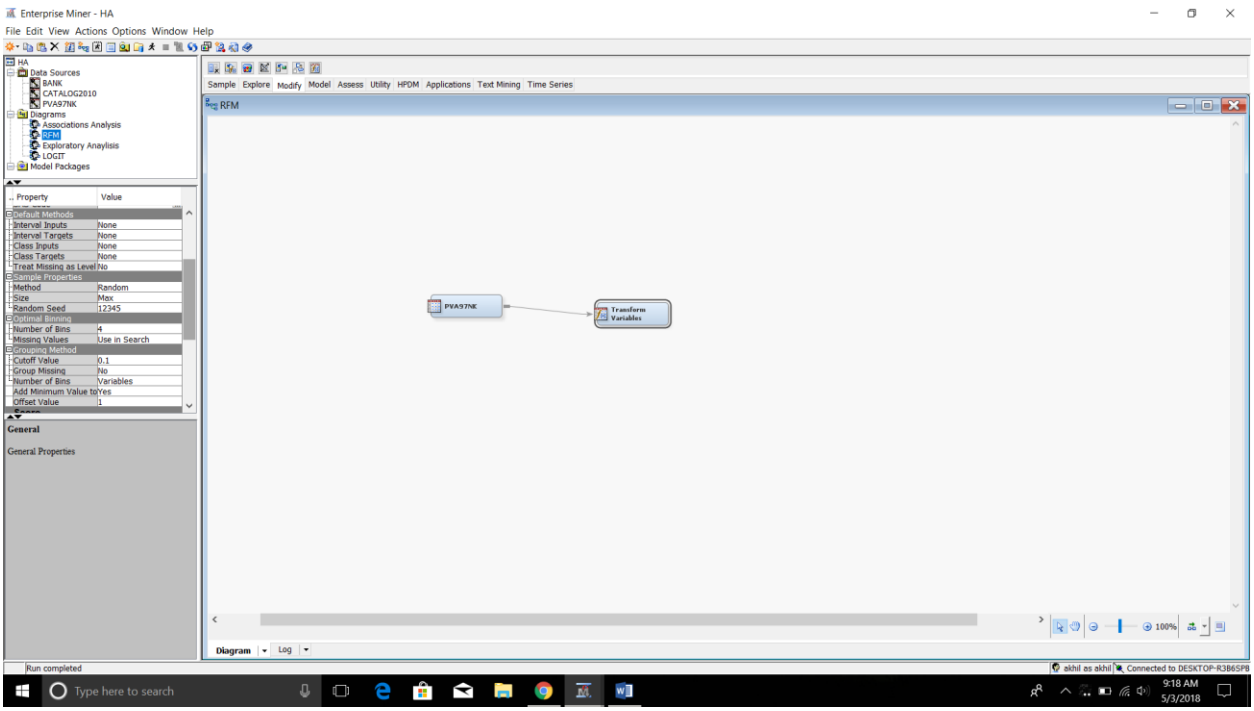


- Changing the TargetD variable role to rejected.



B) Create a new diagram and transform the R, F and M variables as described previously to create four bins of each variable. Concatenate them to create RFM variable.

- First drag and drop PVA97NK dataset from the data sources to the diagram.
- Next, we drag the transform variables block available in the modify block of the SEMMA tool bar.



While assigning the transform variables block we change the sampling properties of this block as shown below:

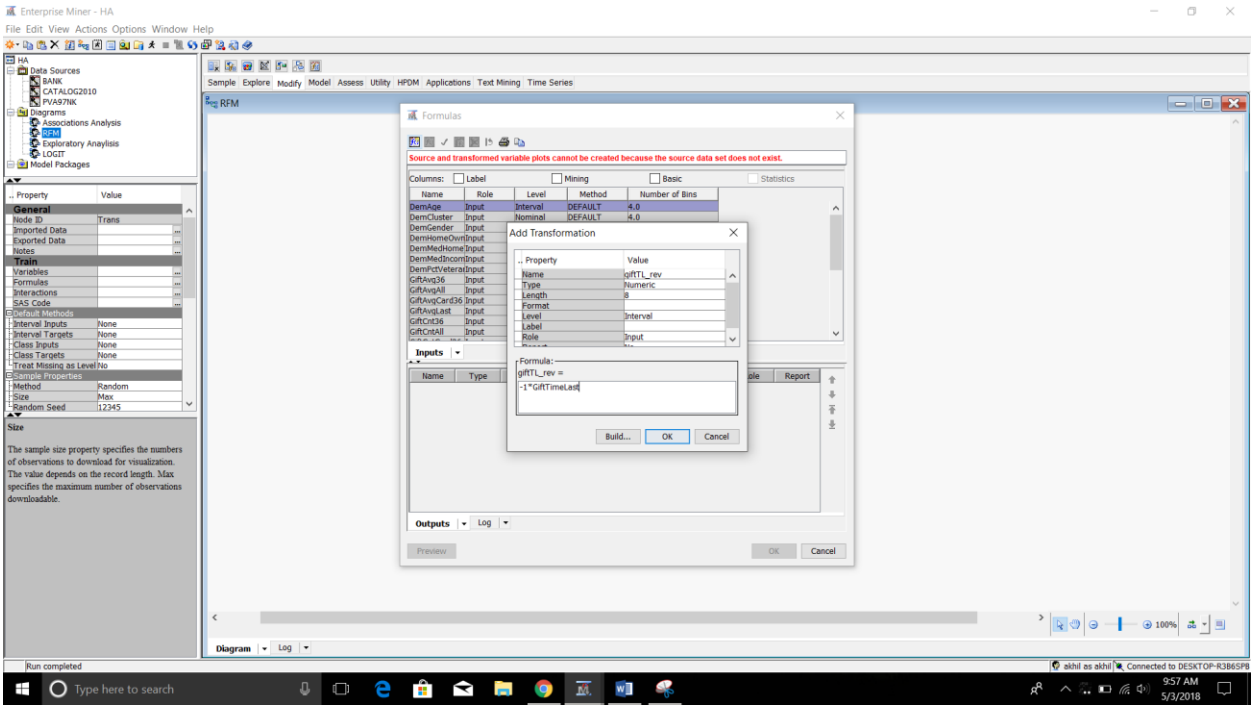
Initially

Sample Properties	
Method	First N
Size	Default
Random Seed	12345

After changing the properties

Sample Properties	
Method	Random
Size	Max
Random Seed	12345

- Next step is setting up the formula for Gift Time Last.

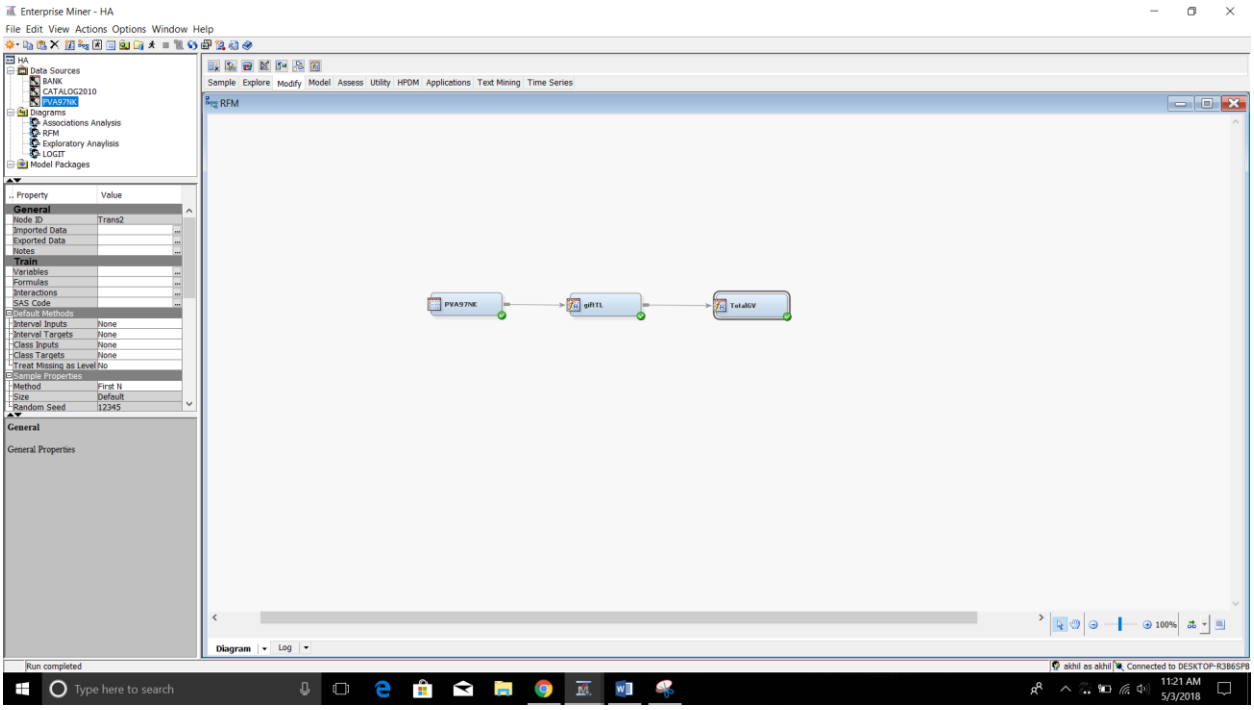


Setting up the formula is done by selecting the options box right next to the formulas option available in the properties block. We define a variable called **giftTL** and assign the Gift Time Last variable to this variable as shown below:

giftTL = -1*GiftTimeLast

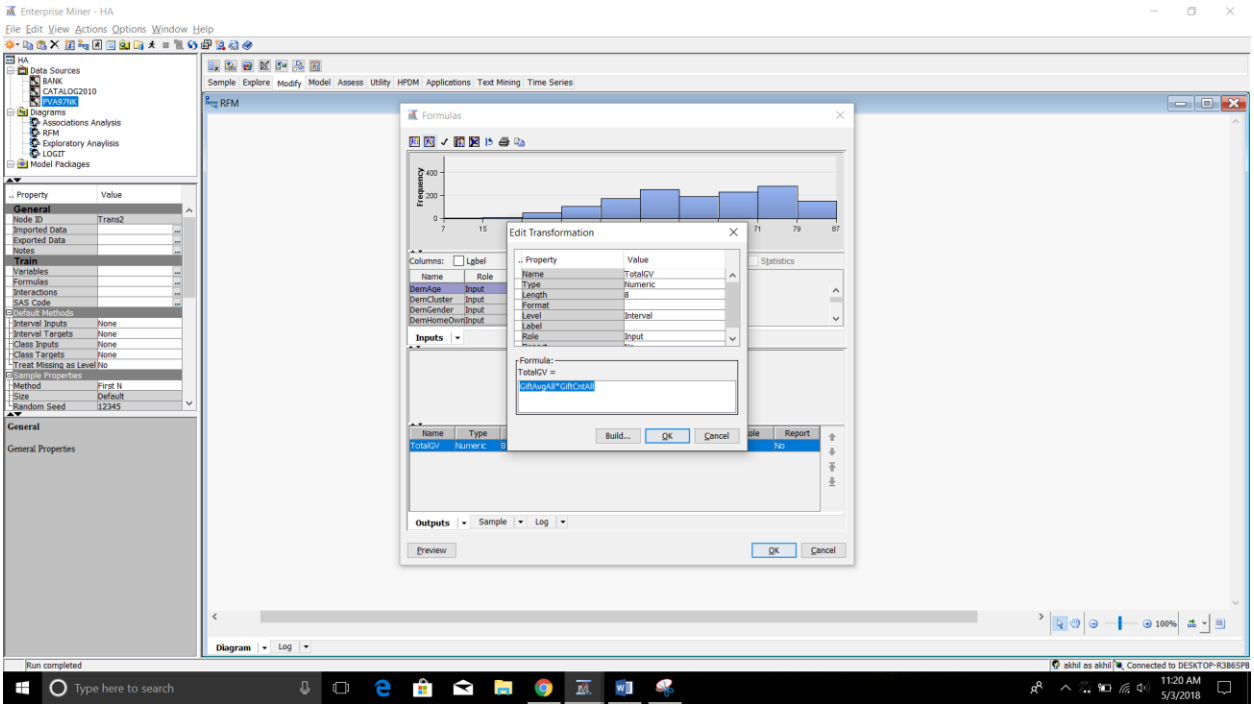
This process of multiplying -1 to the variable is done to reverse code the variable.

- We add another transform variable block (for Monetary Value) and name it as Total Gift Value (TotalGV). We use this block to see if the giftTL_rev variable is created and check for the status of this variable to be set as input for the analysis. Also, we add one more variable called Total GV, to calculate the total gift value.



- The total gift value (TotalGV) is formulated as shown below:

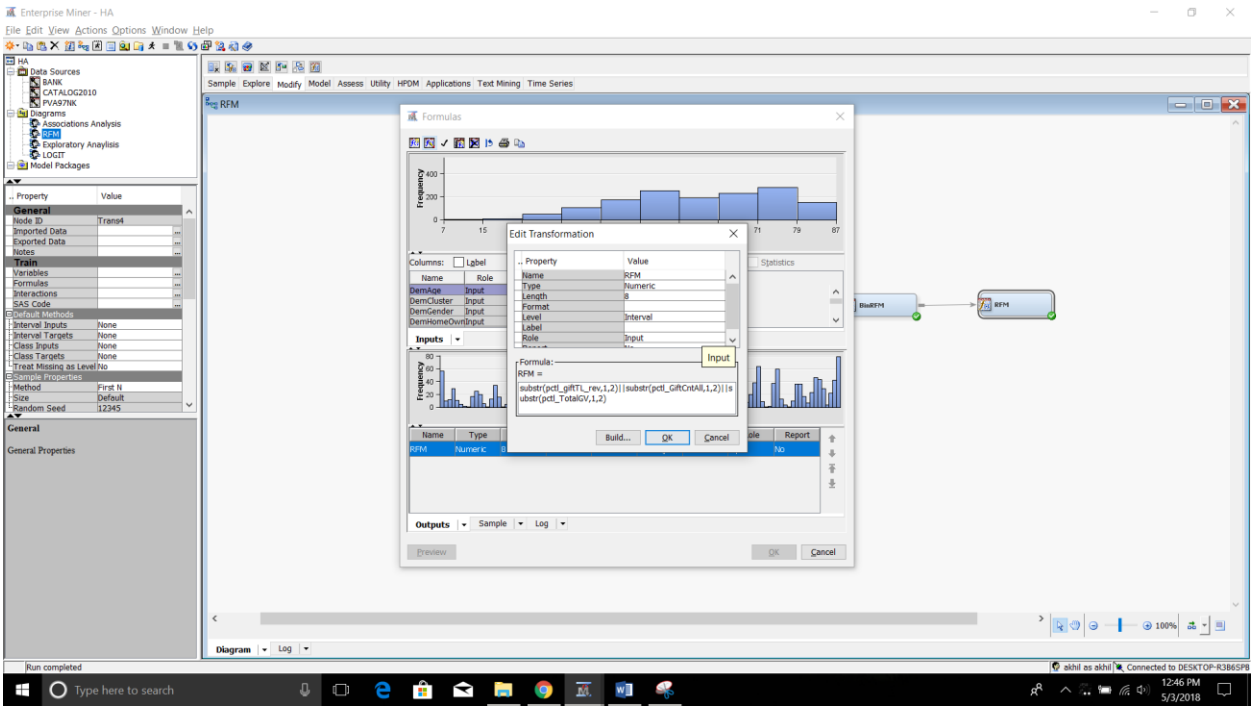
TotalGV = GiftAvgAll*GiftCntAll



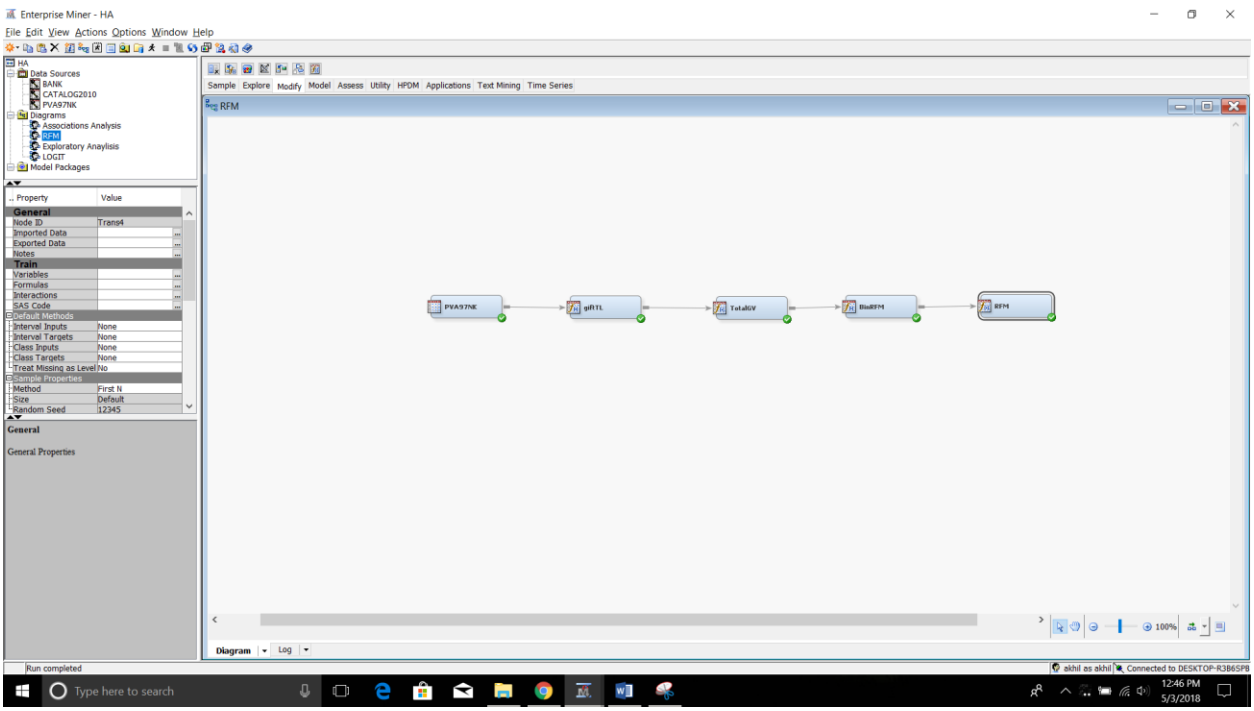
- Now we add the RFM block. The RFM block is formulated as shown below:

RFM = substr(pctl_giftTL_rev,1,2)||substr(pctl_GiftCntAll,1,2)||substr(pctl_TotalGV,1,2)

"||" this symbol is used in the syntax to concatenate the 2 substrings into a single string.

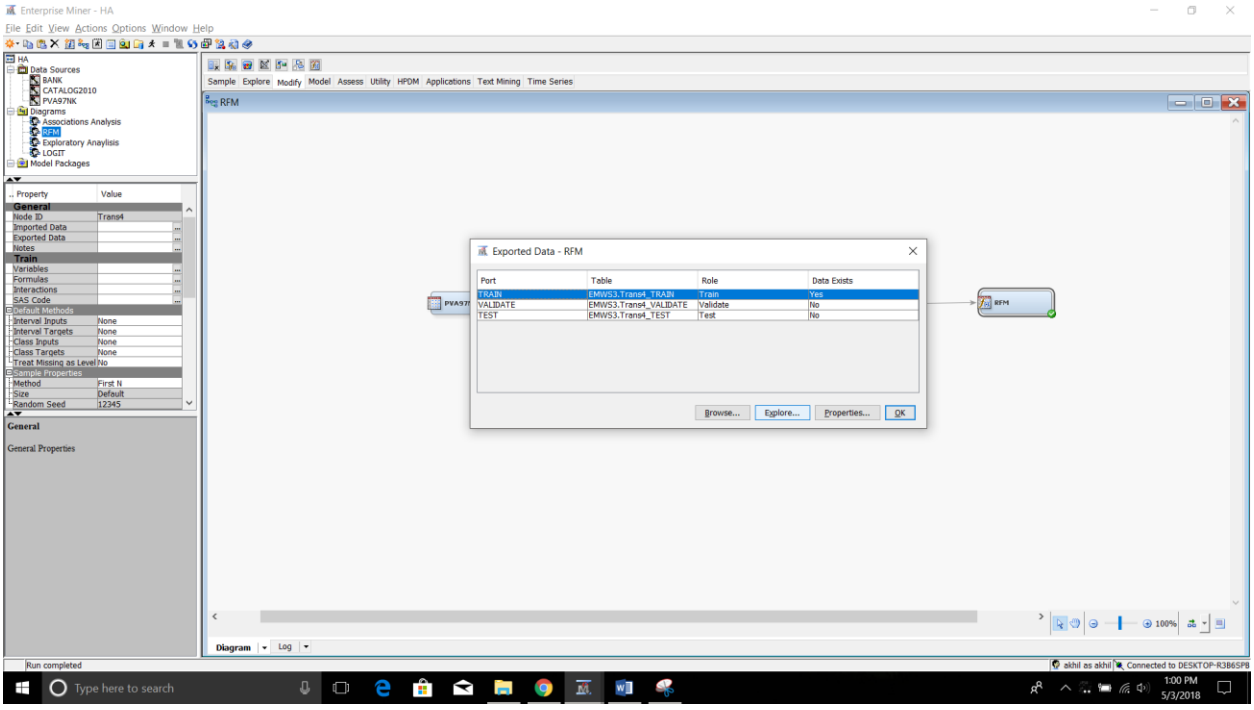


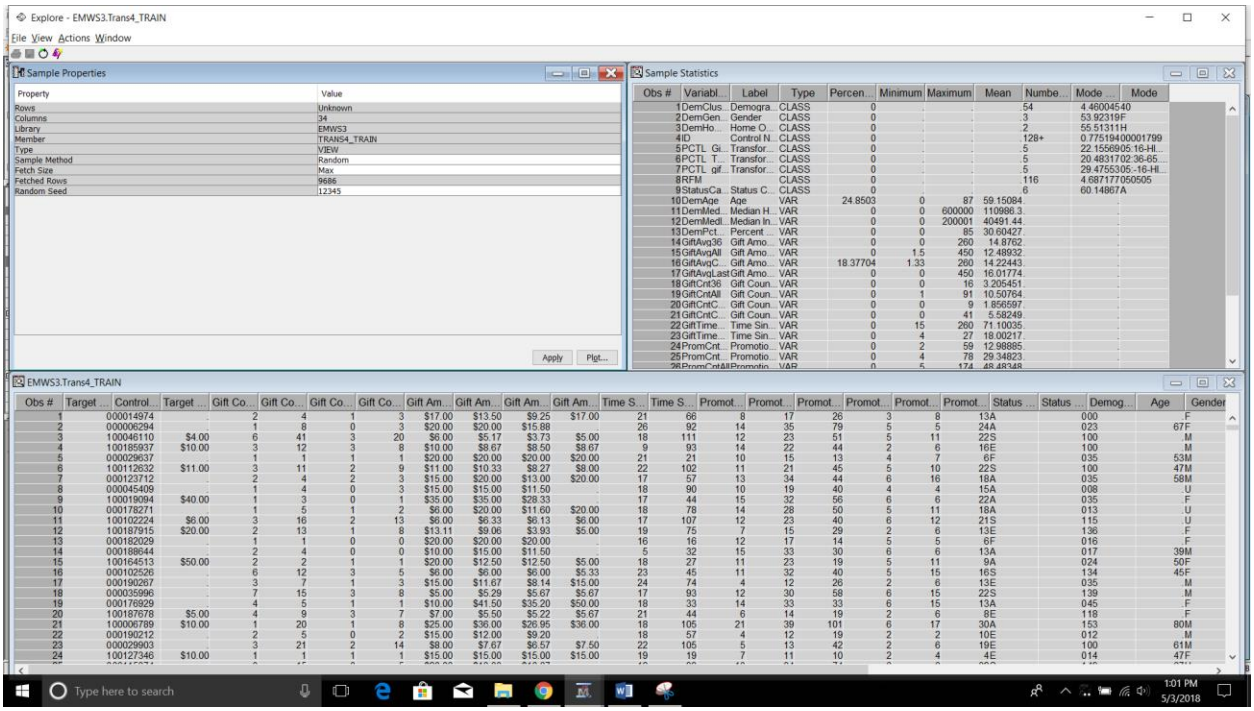
- The final block diagram for the RFM analysis is as shown below:



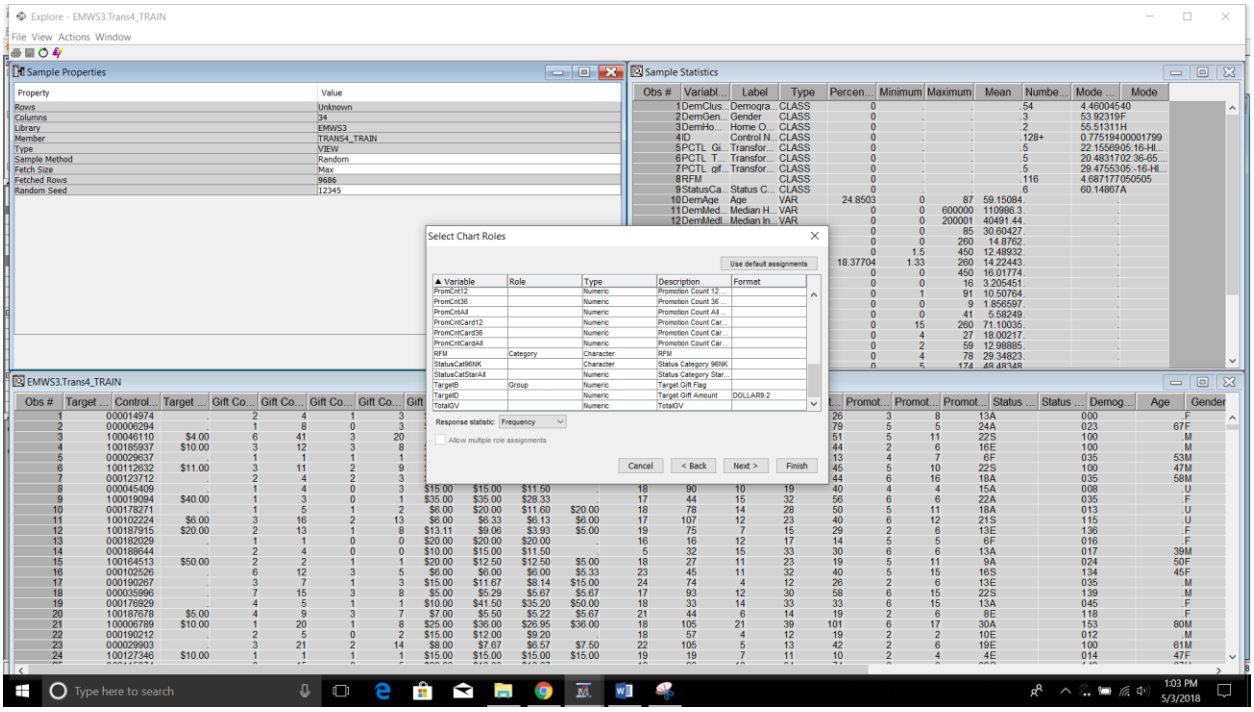
c) Explore the data and perform graphical RFM analysis using a grouped pie chart and a stacked bar chart.

- Pie chart and stacked bar chart are used for the analysis of the RFM model that we created above.
- In order to do this, we first select the RFM node.
- Next, we go to the properties block and select the ellipsis next to the Exported data option.
- We select train followed by selecting the Explore option to see the report as shown below:

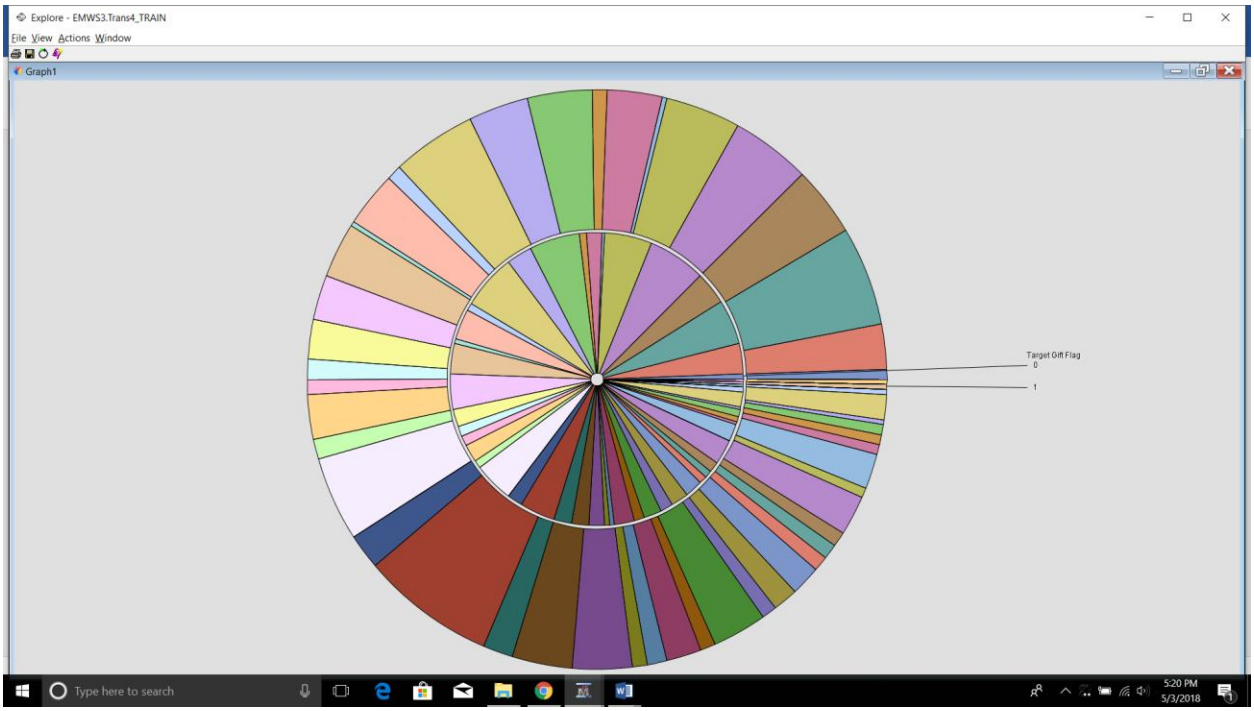




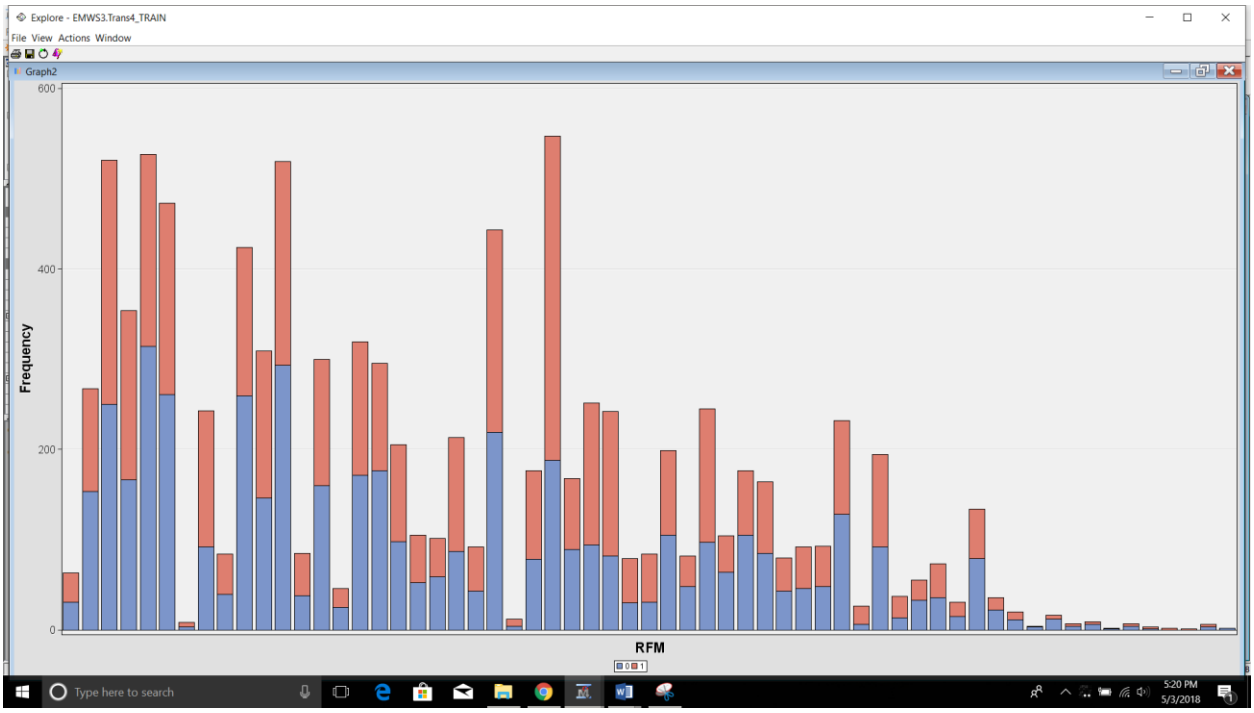
- To observe the plots, we click on actions followed by selecting on the plot option that is available there.

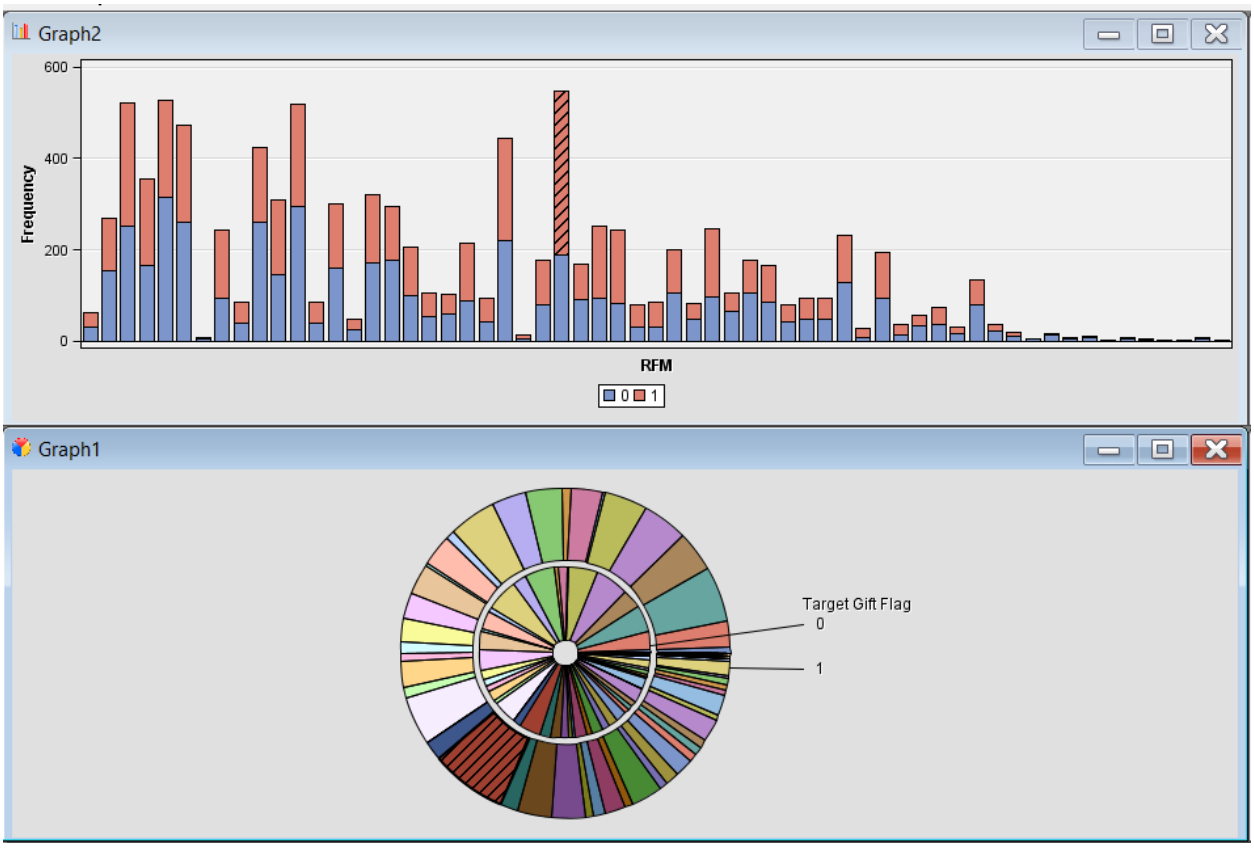


- Before we plot the graph, we set the RFM as a category variable and the TargetB as a group variable as shown in the screen shot above.
- The pie chart plot is as shown below:



- The stacked bar graph can also be obtained by following the same steps as mentioned for pie chart. Here also, we give the RFM variable a role of category and the TargetB variable a role of group.





- When we select a part of the bar graph, the corresponding part pertaining to those values will also get highlighted in the pie chart.

- **Analysis:**

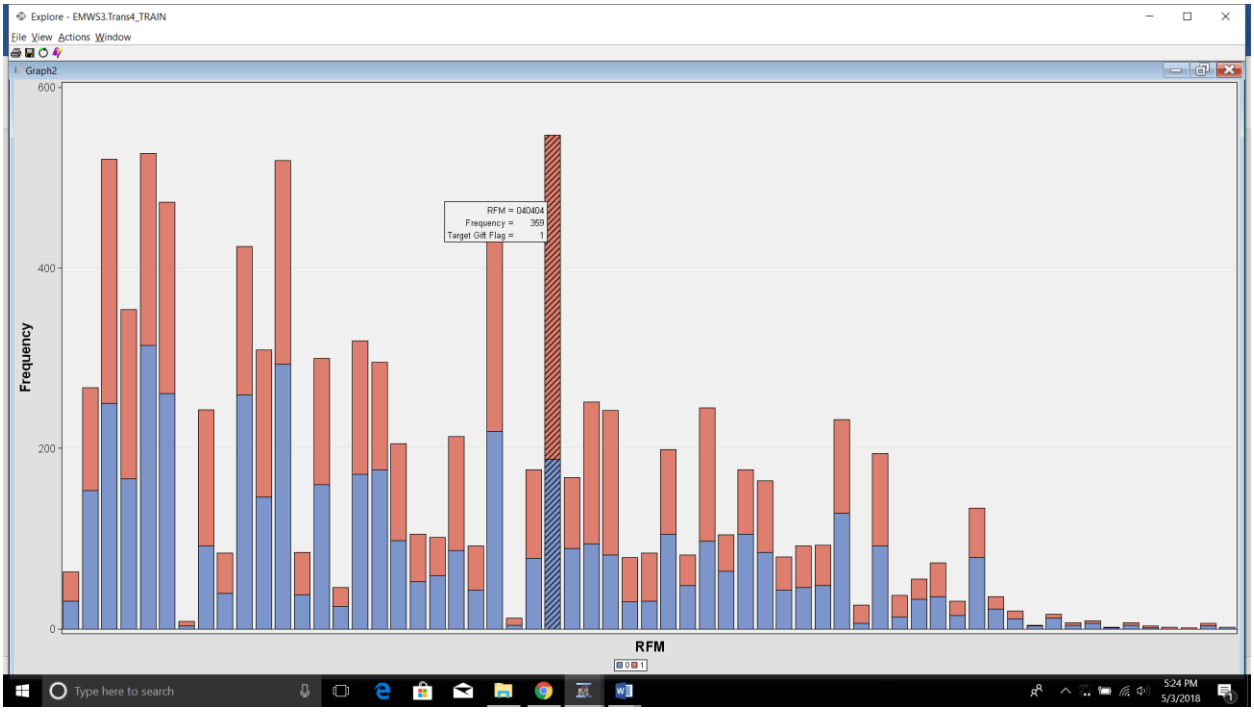
- The bar graph infers 2 things:

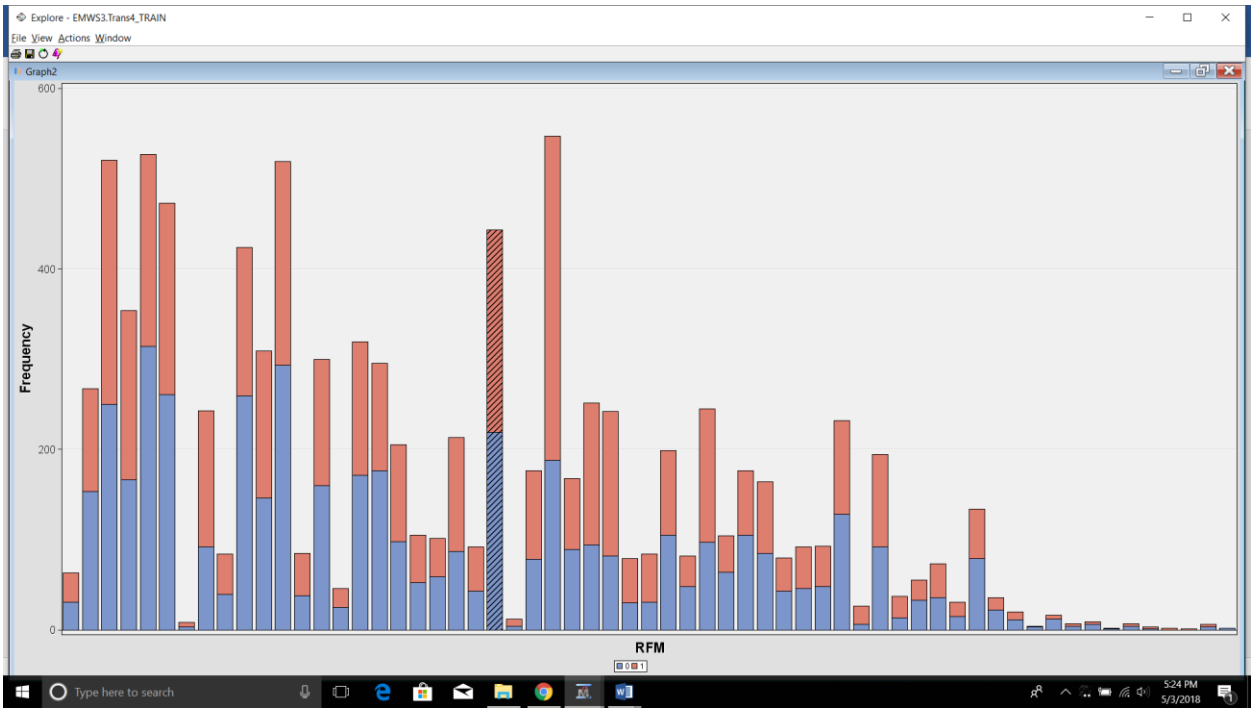
a) 0 – implies the amount of people who did not receive any gift. It is marked as blue.

b) 1 - implies the amount of people who received gifts and they are marked red.

- The entire bar indicates the total donations made.

D) Calculate the Response rate for 040404 and 030303 group?





Response rate for 040404 is 359/ (359+188) = 65.63.

Response rate for 030303 is 224/ (224+219) = 50.65.

e) Each promotional mailing (request for gift) costs $2.3 and the average donation is about $21. What is the break-even response rate for this promotion? Do any of the RFM cells exceed this response rate? Remember to account for the fact that in the population, 95% of mailings are not responded to, while this sample is oversampled to 50% responders and 50% non-responders.

Solution:

Given cost of each mail is $2.3 and the average donation is $21.

Break-even response is calculated using this formula

(Cost of each mail / Average Donation)

= 2.3 / 21 = 0.109 = 10.9%

According to the given information we can see that the response rate is around 0.052, which is calculated as a fraction of responded mailings to non-responded mailings (i.e.., 5% /95%).

Based on the break-even value obtained, we can consider a bin that has value greater than 10.9% or say 11%.

The response rate for each of the bins can be calculated by using the formula given here:

(Frequency value of a bin with flag 1 * 0.052) / [(Frequency value of a bin with flag 1* 0.052) + frequency value when target flag value is 0].

When we computed the response rates using the above formula, we found out that bin 040301 has the highest response rate of 14.77%, based on the calculation shown below:

(20*0.052) / [(20*0.052) + 6] = 14.77%.

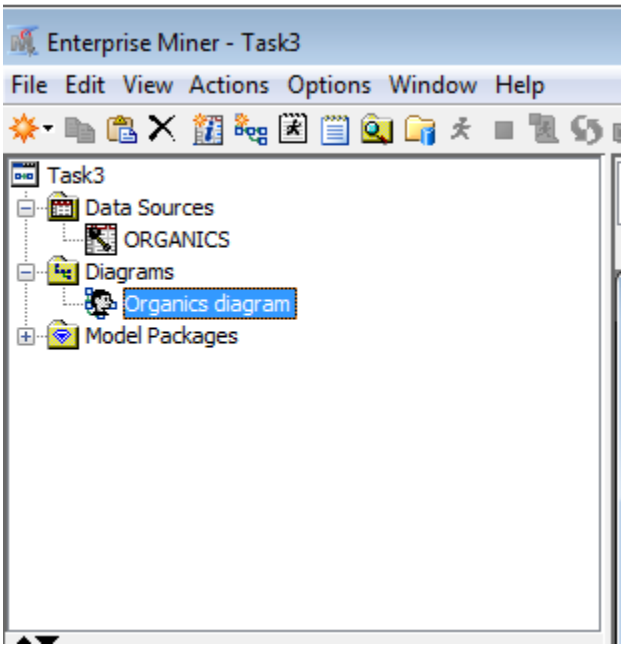
Thus, we can consider this bin 040301.

Part-2: Predictive Modelling, Decision Tree

Task 3 - Data Sources: ORGANICS

1)

a) **Create a new diagram named organics.**



b) **Define the data set ORGANICS as a data source for the project .**

1) set the roles for the analysis variable as shown above .

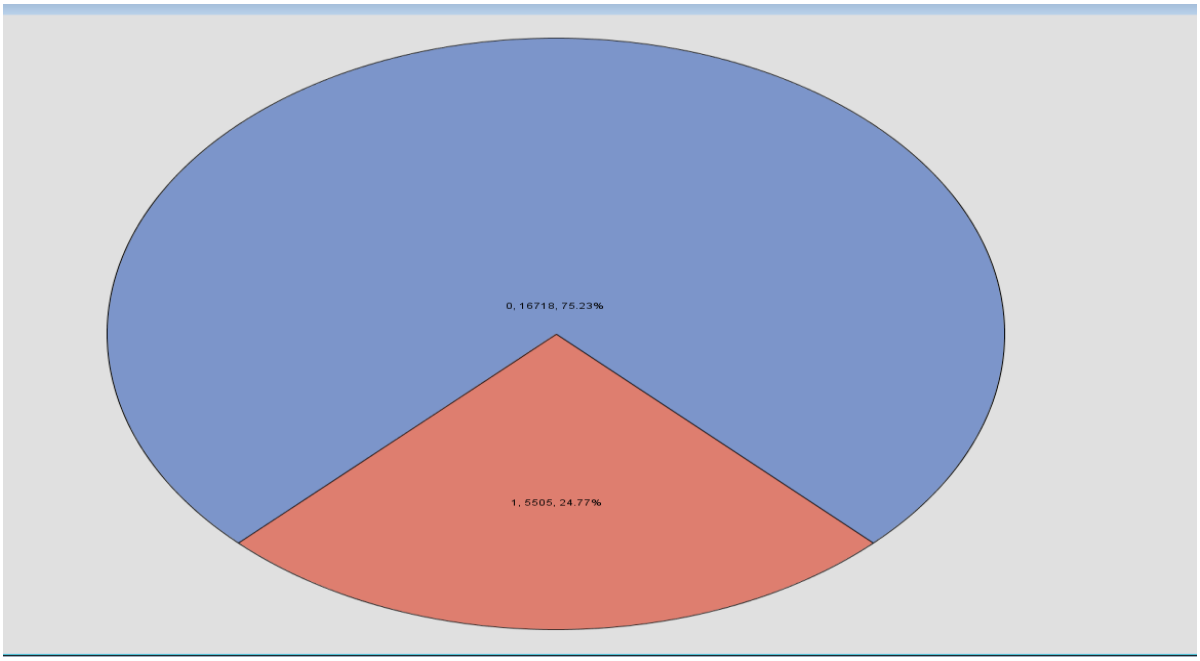
Variables - Ids

(none) not Equal to

Columns: Label Mining

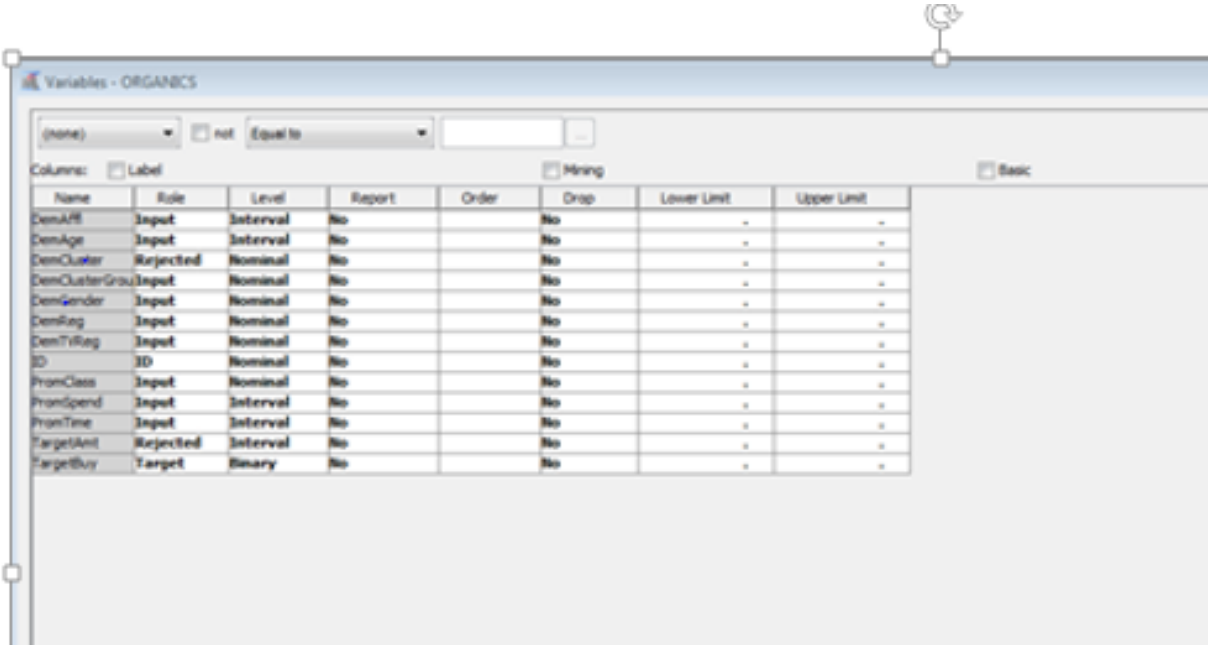
Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
DemAffl	Input	Interval	No		No	.	.
DemAge	Input	Interval	No		No	.	.
DemCluster	Rejected	Nominal	No		No	.	.
DemClusterGrou	Input	Nominal	No		No	.	.
DemGender	Input	Nominal	No		No	.	.
DemReg	Input	Nominal	No		No	.	.
DemTVReg	Input	Nominal	No		No	.	.
ID	ID	Nominal	No		No	.	.
PromClass	Input	Nominal	No		No	.	.
PromSpend	Input	Interval	No		No	.	.
PromTime	Input	Interval	No		No	.	.
TargetAmt	Rejected	Interval	No		No	.	.
TargetBuy	Target	Binary	No		No	.	.

2) Examine the disturbution of the target variable.what is the propotion of individuals who purchased organic products?



From the pie chart we can see that, proportion of individuals who purchased organic parts (orange=1) is 24.77%

3) The variable Demcluster to rejected.



4) As noted above, only TargetBuy is used for this analysis, and it should have a role of Target. Can TargetAmt be used as an input for a model that is used to predict TargetBuy? Why or why not?

A) Not necessarily. TargetAmt is the number of organic products purchased by customers and TargetBuy is a binary variable that gives whether a purchase is made or not. Obviously, TargetBuy is only present for variable that contain a value>0 in TargetAmt. It could be redundant variable for analysis. There is a high correlation between the independent variable TargetAmt and outcome TargetBuy. Therefore, rejecting this variable from analysis would result in an unbiased model.

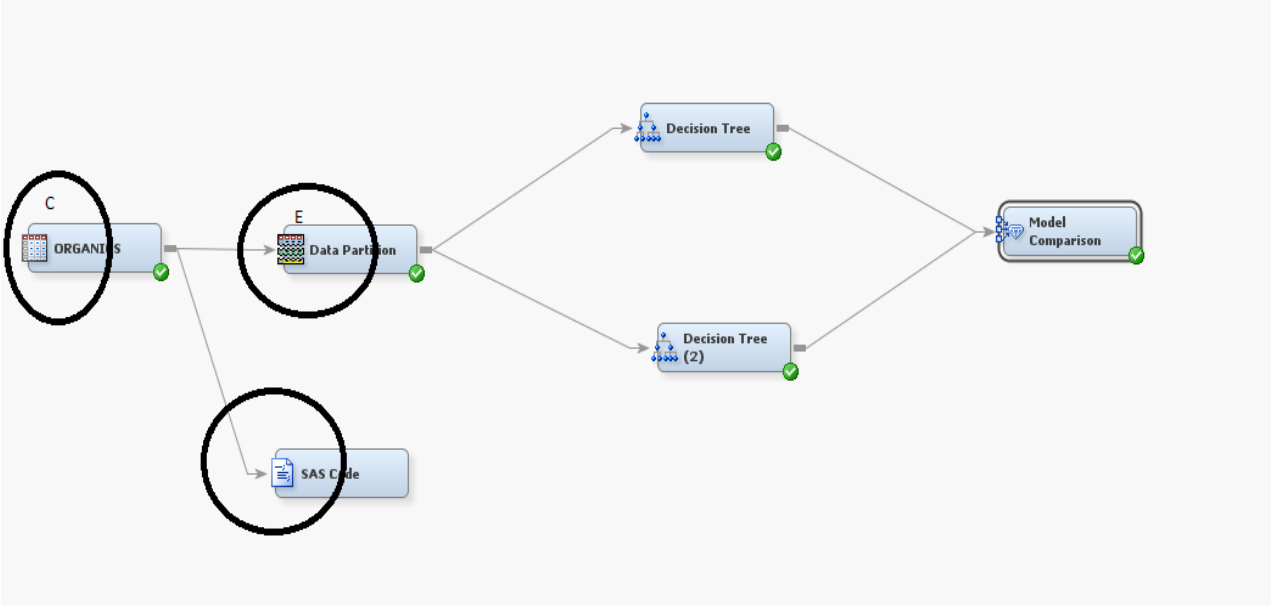
5) **Data source definition is implemented from beginning.**

c. Add the ORGANICS data source to the Organics diagram workspace.

d. Add a Data partition node to the diagram and connect it to the data source node. Assign 65% of the data for data training and 35 % for data validation.

.. Property	Value
General	
Node ID	Part
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	65.0
Validation	35.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	5/2/18 1:25 PM
Run ID	c38f3133-a212-46b1-ada8-

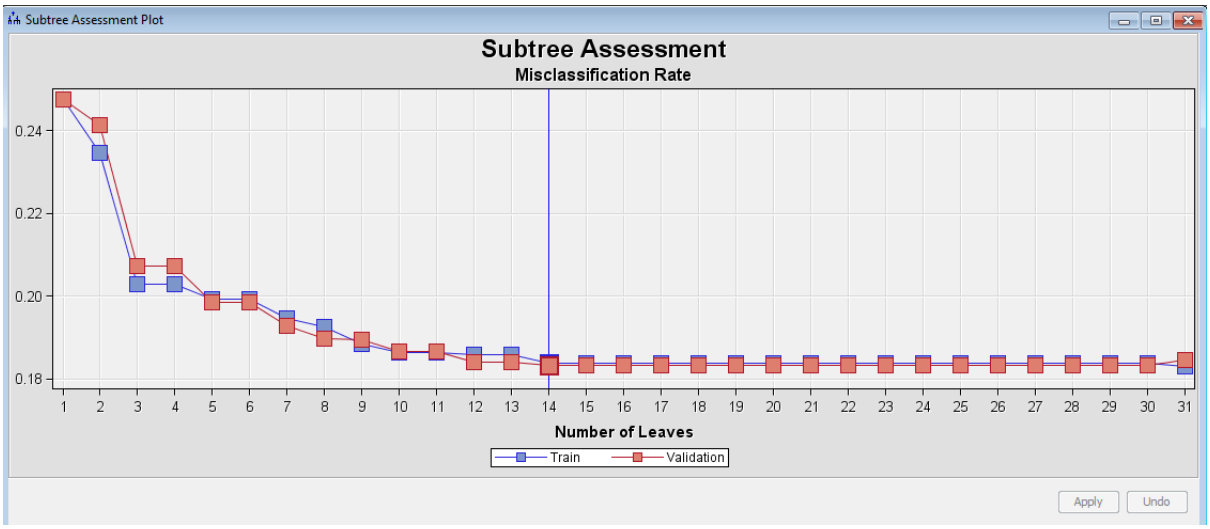
e. Add a decision tree node to the workspace and connect it to the data partition node.



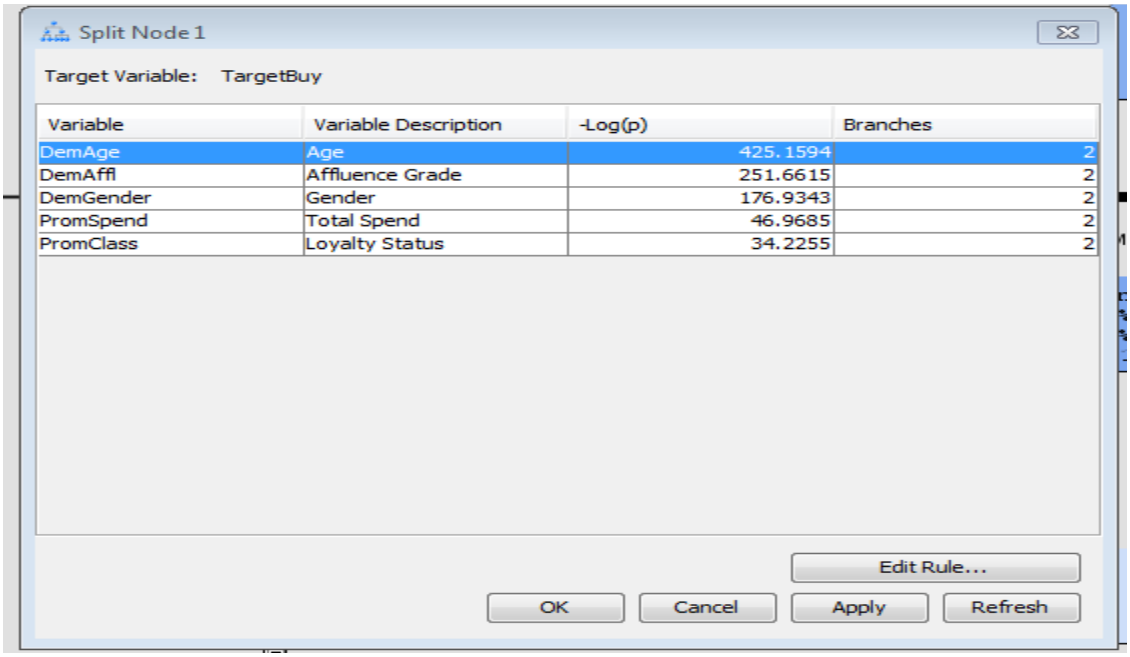
f. Create a decision tree model autonomously. Use Misclassification as the model assessment statistic.

1. How many leaves are in the optimal tree?

14 leaves.



2. Which variables are used in the first split? What were the completing splits for this split?



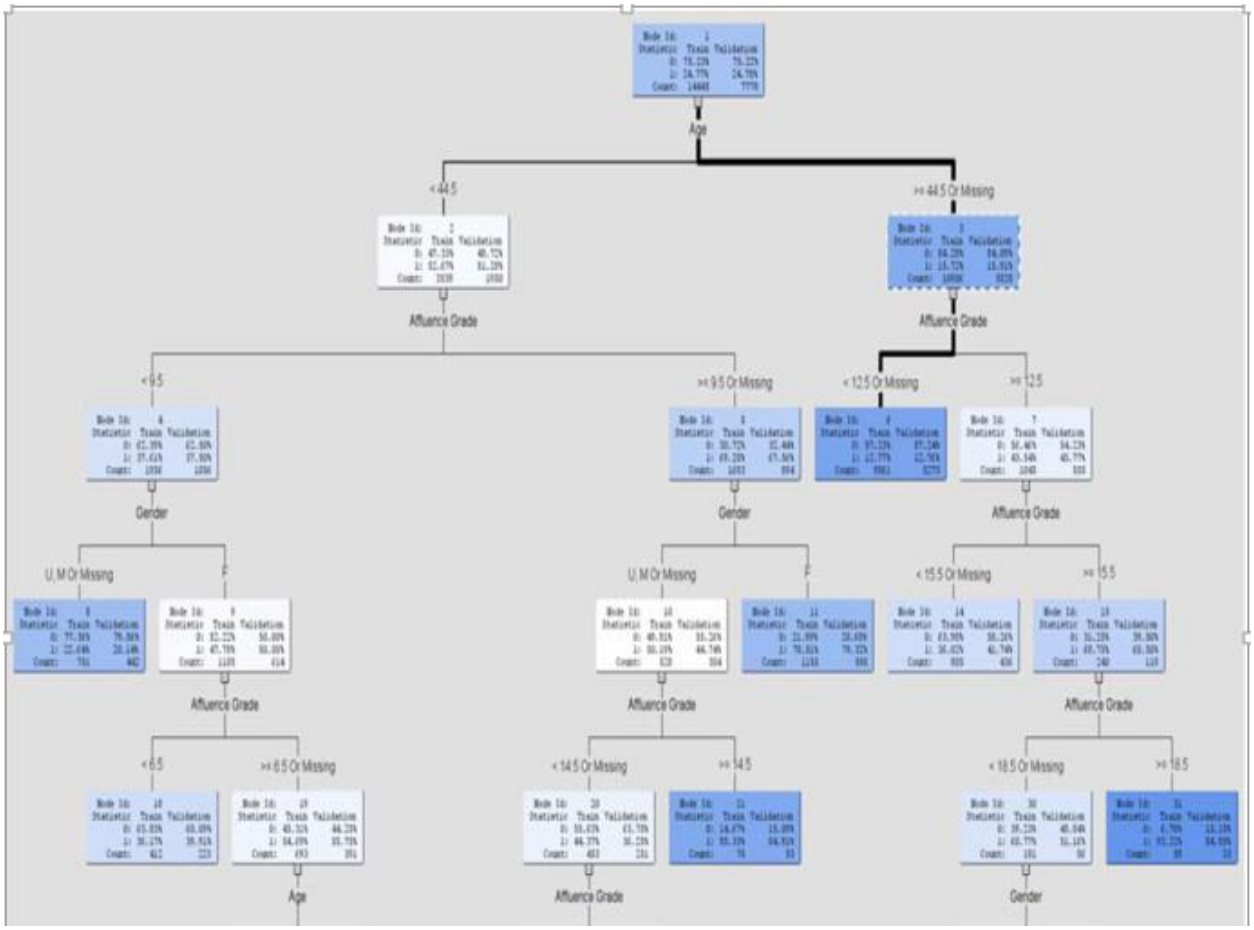
Age variable was used for the split. Age < 44.5 and Age >=44.5 or Missing.

Competing Splits: Affluence Grade<12.5 or Missing and Affluence Grade >=12.5.

Competing is not a right word as these two variables have less information than that of AGE.

3. Which variables were used for the second split for all branches from first split?

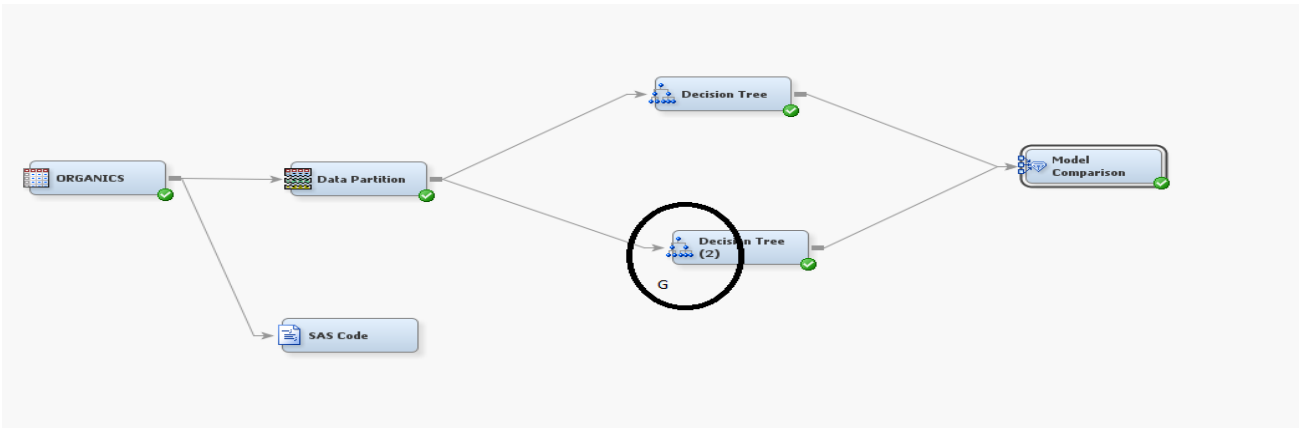
The variable **AFFLUENCE GRADE** is used for second split across all branches.



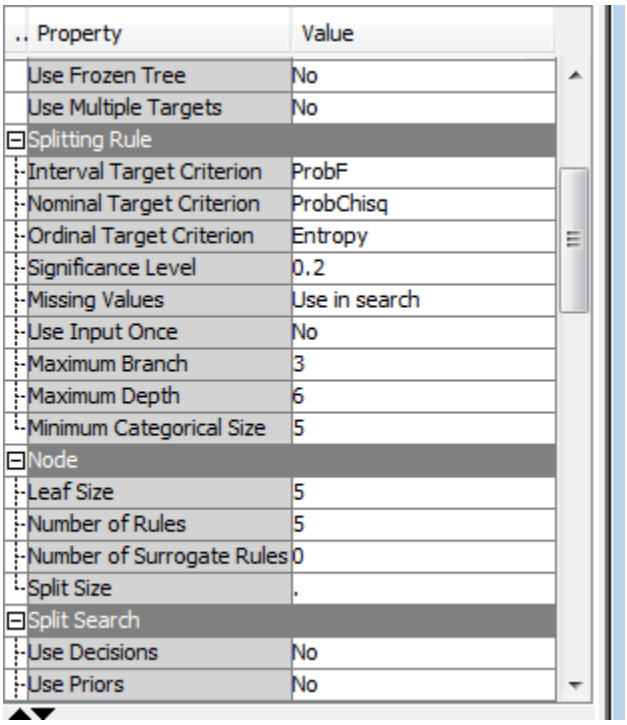
4. Discuss the results and provide your insights?

A) Age is the root node that we got in this decision tree. If the age is less than 44.5, the chance of target variable being 1 is more than 50 %. When the age is more than 44.5, the chances of target being 0 is 15%. More the **AFFLUENCE GRADE**, more chances of getting variable 1. Overall, we can say that the model isn't an overfitting one as the error rate decreased with increase in number of leaves.

g. **Add a second decision tree node to the diagram and connect it to the data partition node.**

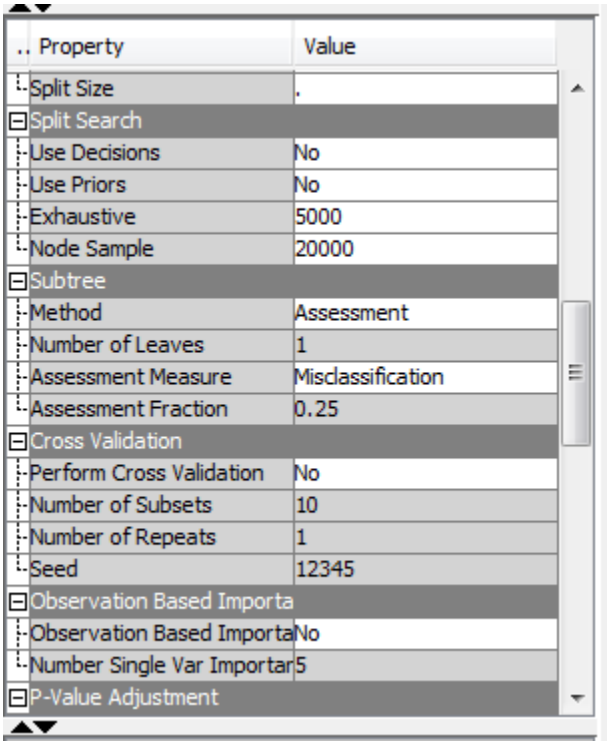


1. In the properties panel of the new decision tree node, change the maximum number of branches from a node to 3 to allow for three-way splits.



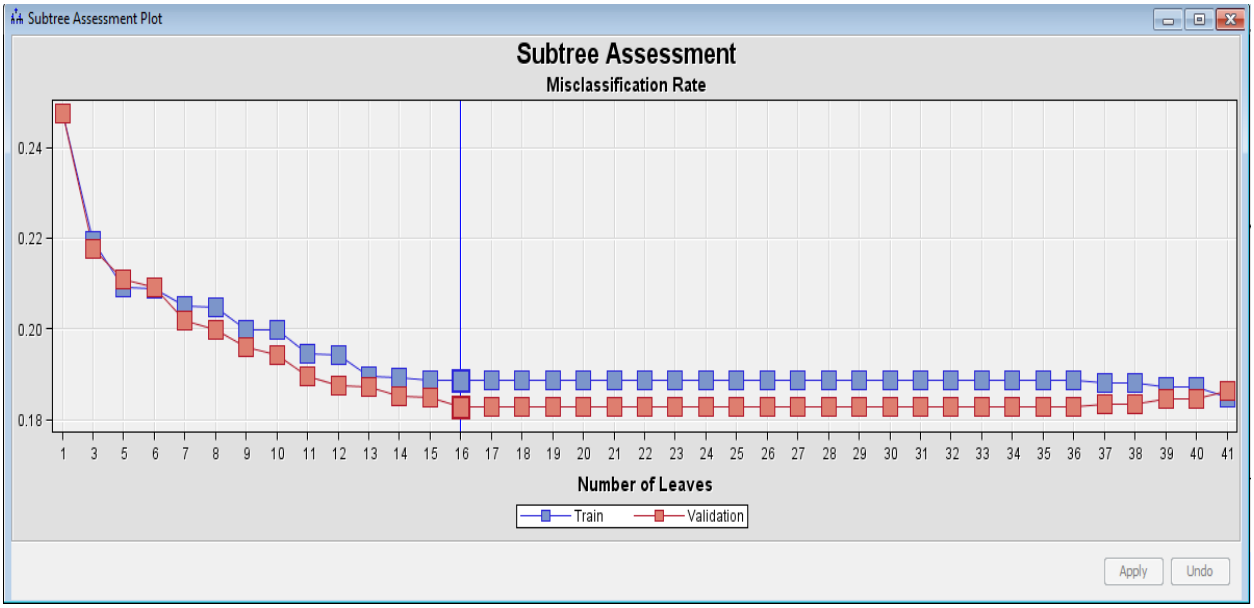
Property	Value
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	3
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No

2. Create a decision tree model using Misclassification as the model assessment statistic.



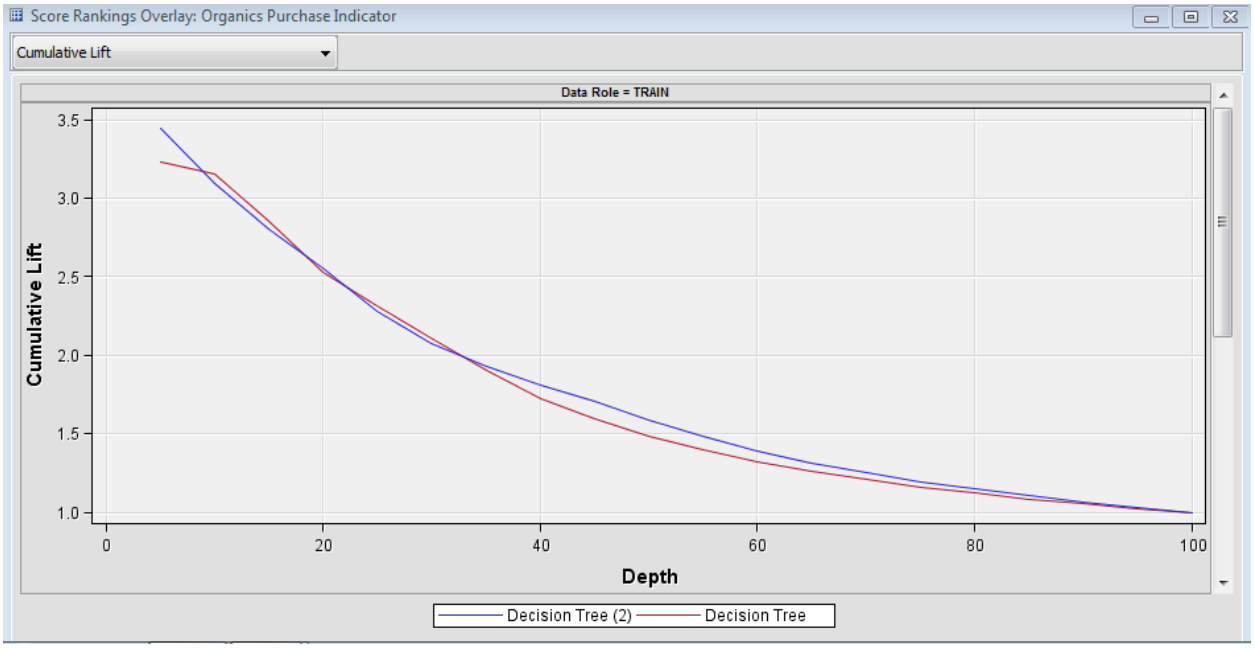
Property	Value
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Misclassification
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation Based Importa	
Observation Based Importa	No
Number Single Var Importar	5
P-Value Adjustment	

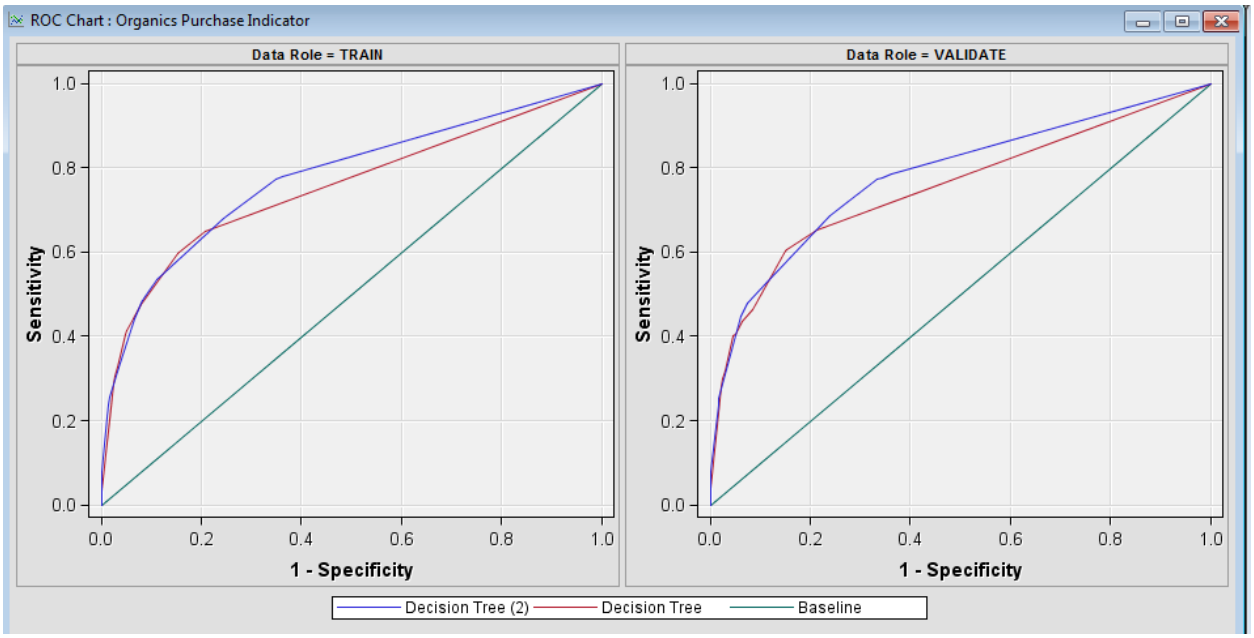
3. How many leaves are in the optimal tree?



With Decision tree 3 splits and misclassification as model assessment statistic it resulted in **42 leaves.**

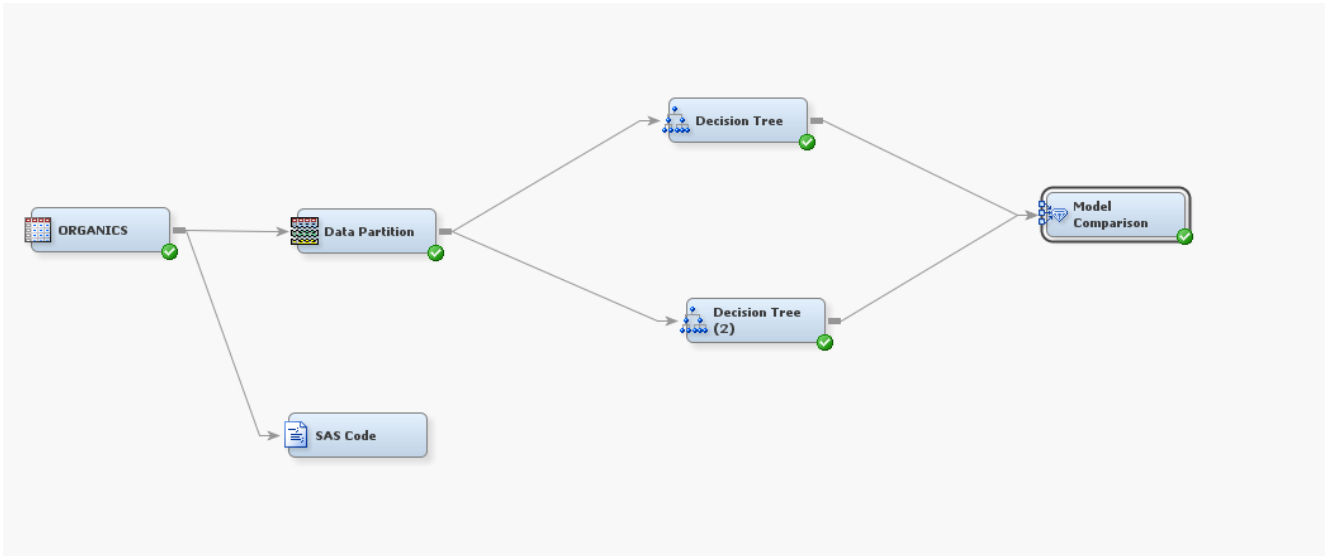
h. Based on Misclassification rate, which of the decision tree models appears to be better?





TREE 2 is better when compared to TREE 1, as the area under the curve of TREE 1 is higher than that of TREE 2, and the blue line (i.e the TREE 2) is much closer to 1, which also proves that TREE 2 is better than TREE 1.

- **<u>Final Model</u>**



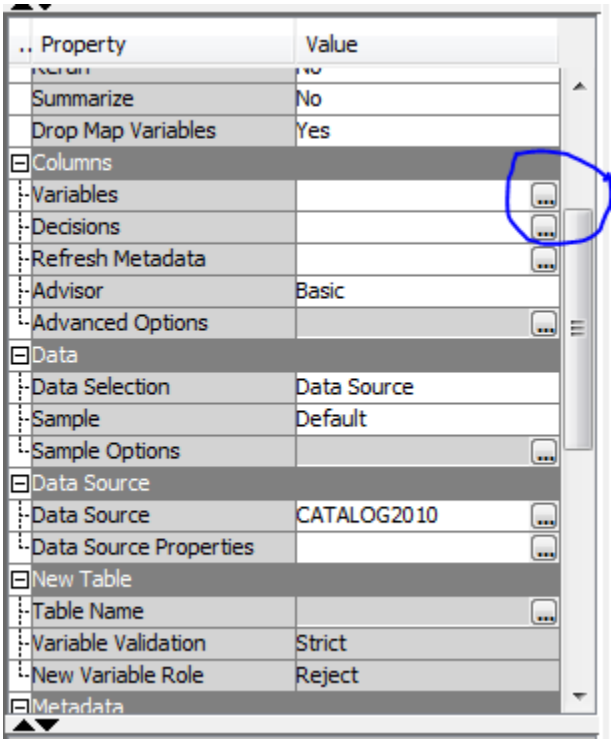
PART 3: Predictive Modelling using LOGIT (Logistic Regression)

TASK 4 - LOGIT

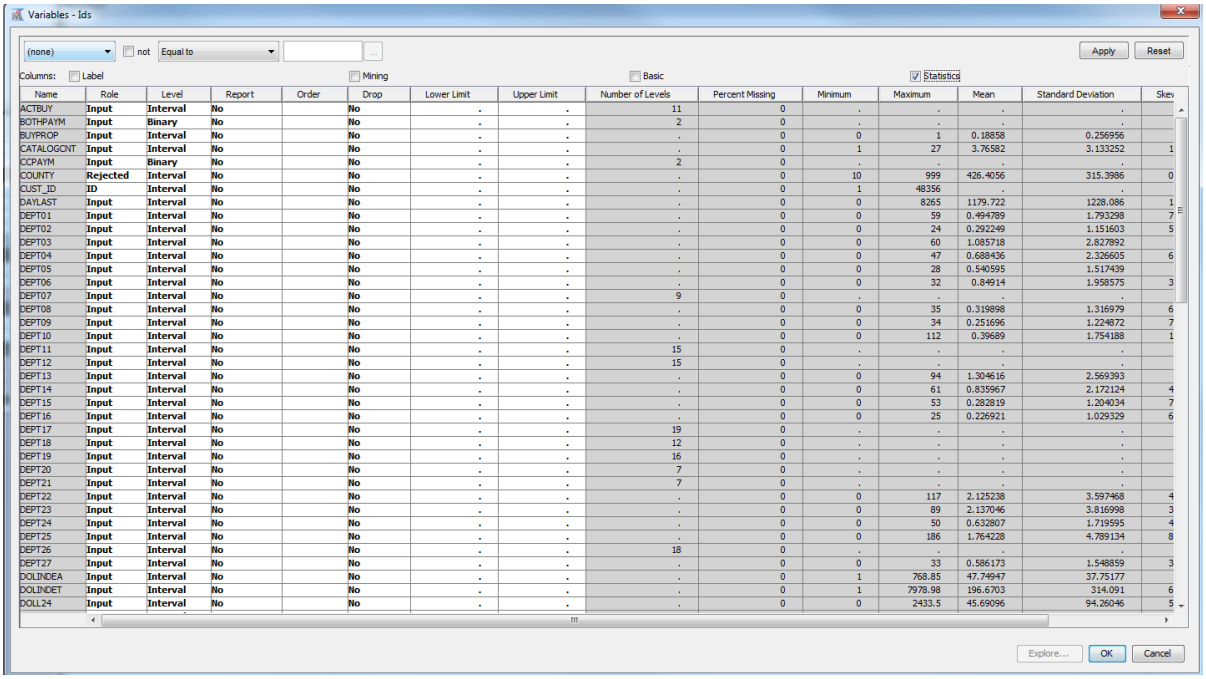
The steps used for LOGIT is similar to that of Decision tree, except that we add Variable Clustering node to reduce redundancy and Regression node is used to select relevant inputs.

Variable clustering is to group variables according to similarities between them.

1. **Review the CATLOG2010 data set. From the panel on left, click the ellipsis next to Variables.**

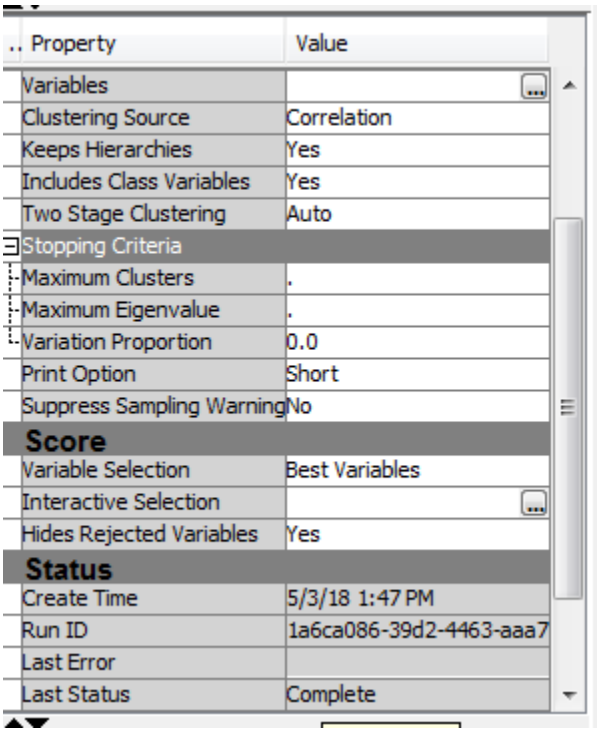


2. Compute some basic statistics.



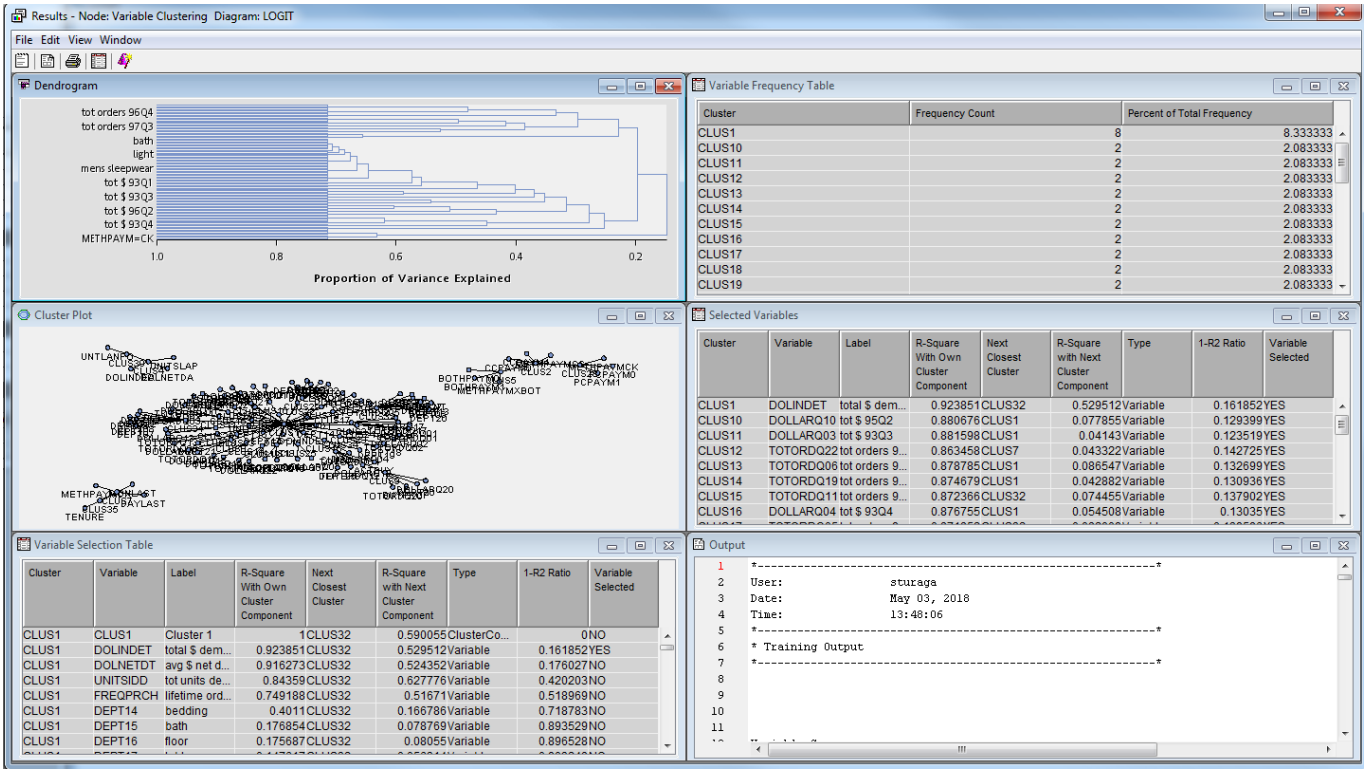
Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit	Number of Levels	Percent Missing	Minimum	Maximum	Mean	Standard Deviation	Skew
ACTBUY	Input	Interval	No		No	.	.	11	0	.	.	.	.	
BOTHPAYM	Input	Binary	No		No	.	.	2	0	.	.	.	.	
BUYPROP	Input	Interval	No		No	.	.	.	0	0	1	0.18858	0.256956	
CATALOGCNT	Input	Interval	No		No	.	.	.	0	1	27	3.76582	3.133252	1
CCPAYM	Input	Binary	No		No	.	.	2	0	.	.	.	.	
COUNTY	Rejected	Interval	No		No	.	.	.	0	10	999	426.4056	315.3986	0
CUST_ID	ID	Interval	No		No	.	.	.	0	1	48356	.	.	
DAYLAST	Input	Interval	No		No	.	.	.	0	0	8265	1179.722	1228.086	1
DEPT01	Input	Interval	No		No	.	.	.	0	0	59	0.494789	1.793298	7
DEPT02	Input	Interval	No		No	.	.	.	0	0	24	0.292249	1.151603	5
DEPT03	Input	Interval	No		No	.	.	.	0	0	60	1.085718	2.827892	
DEPT04	Input	Interval	No		No	.	.	.	0	0	47	0.688436	2.326605	6
DEPT05	Input	Interval	No		No	.	.	.	0	0	28	0.540595	1.517439	
DEPT06	Input	Interval	No		No	.	.	.	0	0	32	0.84914	1.958575	3
DEPT07	Input	Interval	No		No	.	.	9	0	.	.	.	.	
DEPT08	Input	Interval	No		No	.	.	.	0	0	35	0.319898	1.316979	6
DEPT09	Input	Interval	No		No	.	.	.	0	0	34	0.251696	1.224872	7
DEPT10	Input	Interval	No		No	.	.	.	0	0	112	0.39689	1.754188	1
DEPT11	Input	Interval	No		No	.	.	15	0	.	.	.	.	
DEPT12	Input	Interval	No		No	.	.	15	0	.	.	.	.	
DEPT13	Input	Interval	No		No	.	.	.	0	0	94	1.304616	2.569393	
DEPT14	Input	Interval	No		No	.	.	.	0	0	61	0.835967	2.172124	4
DEPT15	Input	Interval	No		No	.	.	.	0	0	53	0.282819	1.204034	7
DEPT16	Input	Interval	No		No	.	.	.	0	0	25	0.226921	1.029329	6
DEPT17	Input	Interval	No		No	.	.	19	0	.	.	.	.	
DEPT18	Input	Interval	No		No	.	.	12	0	.	.	.	.	
DEPT19	Input	Interval	No		No	.	.	16	0	.	.	.	.	
DEPT20	Input	Interval	No		No	.	.	7	0	.	.	.	.	
DEPT21	Input	Interval	No		No	.	.	7	0	.	.	.	.	
DEPT22	Input	Interval	No		No	.	.	.	0	0	117	2.125238	3.597468	4
DEPT23	Input	Interval	No		No	.	.	.	0	0	89	2.137046	3.816998	3
DEPT24	Input	Interval	No		No	.	.	.	0	0	50	0.632807	1.719595	4
DEPT25	Input	Interval	No		No	.	.	.	0	0	186	1.764228	4.789134	8
DEPT26	Input	Interval	No		No	.	.	18	0	.	.	.	.	
DEPT27	Input	Interval	No		No	.	.	.	0	0	33	0.586173	1.548859	3
DOLINDEA	Input	Interval	No		No	.	.	.	0	1	768.85	47.74947	37.75177	
DOLINDET	Input	Interval	No		No	.	.	.	0	1	7978.98	196.6703	314.091	6
DOLL24	Input	Interval	No		No	.	.	.	0	0	2433.5	45.69096	94.26046	5

Cluster variables according to their similarities. In the properties panel, change the Include Class Variable property to **YES** and the variable selection property to **BEST VARIABLES.**



Property	Value
Variables	
Clustering Source	Correlation
Keeps Hierarchies	Yes
Includes Class Variables	Yes
Two Stage Clustering	Auto
Stopping Criteria	
Maximum Clusters	.
Maximum Eigenvalue	.
Variation Proportion	0.0
Print Option	Short
Suppress Sampling Warning	No
Score	
Variable Selection	Best Variables
Interactive Selection	
Hides Rejected Variables	Yes
Status	
Create Time	5/3/18 1:47 PM
Run ID	1a6ca086-39d2-4463-aaa7
Last Error	
Last Status	Complete

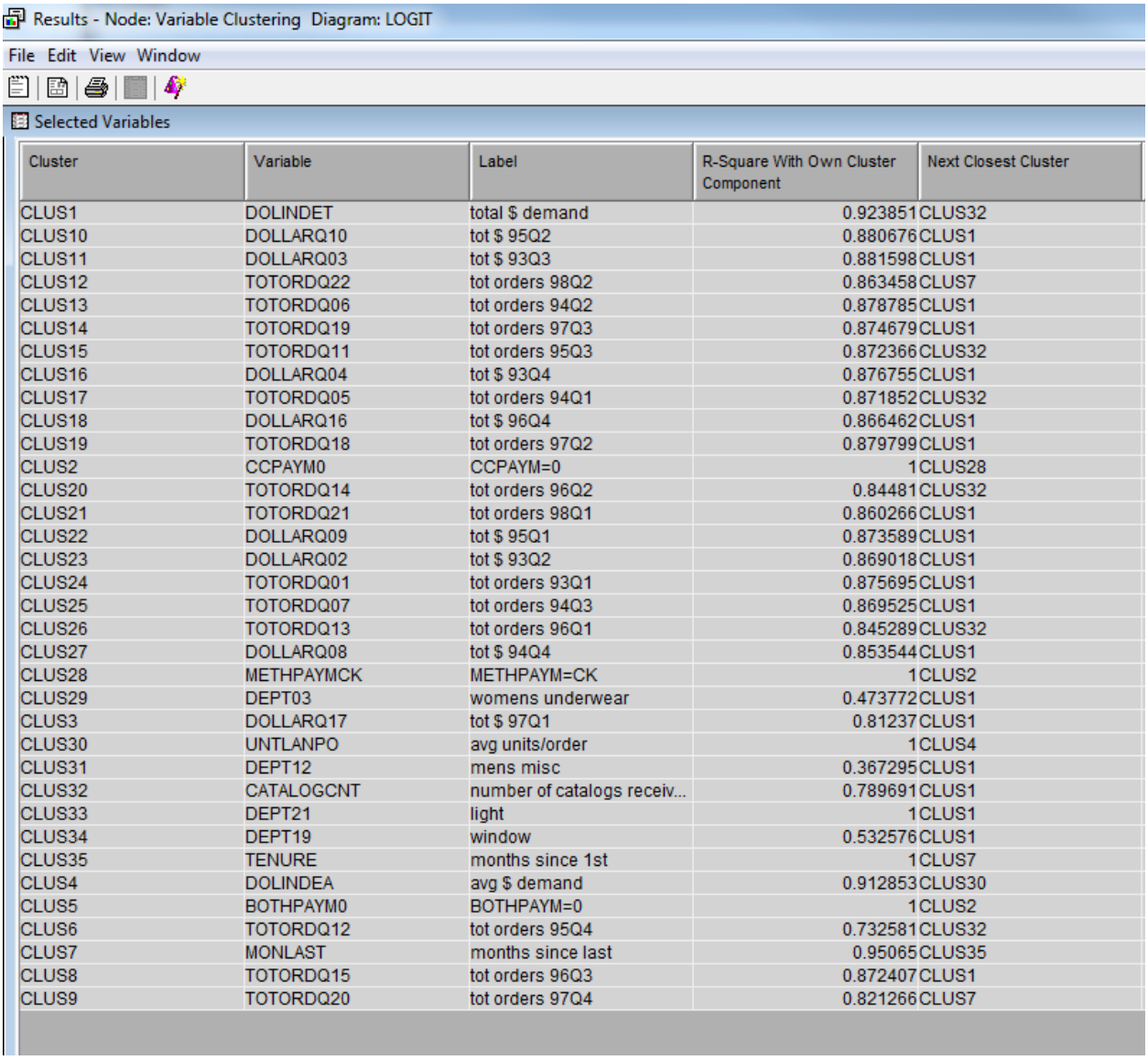
CLUSTERING OUTPUT:



AN OUTPUT WINDOW SHOWING A SUMMARY OF THE FINAL CLUSTER SOLUTION:

Number of Clusters	Explained by Clusters	Variation Explained by Clusters	Explained by a Cluster	Eigenvalue in a Cluster	R-squared for a Variable	Ratio for a Variable
1	14.269655	0.1486	0.1486	5.075727	0.0043	
2	18.878088	0.1966	0.1578	3.699895	0.0106	0.9916
3	21.903078	0.2282	0.1877	2.748968	0.0111	0.9918
4	24.167157	0.2517	0.1877	2.623513	0.0111	0.9916
5	26.705701	0.2782	0.1893	2.073863	0.0113	0.9914
6	28.466963	0.2965	0.1958	2.023150	0.0113	0.9927
7	30.338115	0.3160	0.1958	1.906582	0.0113	1.0351
8	31.993414	0.3333	0.2145	1.829092	0.0117	1.0203
9	33.791416	0.3520	0.2145	1.734839	0.0117	1.1004
10	35.400976	0.3688	0.2252	1.661294	0.0119	1.0955
11	37.045861	0.3859	0.2333	1.645777	0.0120	1.0950
12	38.514796	0.4012	0.2333	1.641035	0.0120	1.0950
13	40.075066	0.4174	0.2462	1.578818	0.0122	1.0903
14	41.568625	0.4330	0.2462	1.555217	0.0122	1.0903
15	43.111311	0.4491	0.2462	1.548926	0.0122	1.0903
16	44.584612	0.4644	0.2462	1.545298	0.0122	1.0903
17	46.100224	0.4802	0.2538	1.537361	0.0124	1.0892
18	47.580873	0.4956	0.2538	1.522834	0.0124	1.0892
19	48.978020	0.5102	0.2538	1.518141	0.0124	1.0892
20	50.489323	0.5259	0.2538	1.507744	0.0124	1.0892
21	51.997062	0.5416	0.2538	1.501807	0.0124	1.0892
22	53.498395	0.5573	0.2538	1.499339	0.0124	1.0892
23	54.982973	0.5727	0.2615	1.482932	0.0126	1.0875
24	56.453832	0.5881	0.2693	1.480621	0.0125	1.0875
25	57.934102	0.6035	0.2693	1.474760	0.0125	1.0875
26	59.408841	0.6188	0.2693	1.400954	0.0125	1.0875
27	60.681938	0.6321	0.2693	1.327516	0.0125	1.0875
28	62.009454	0.6459	0.2693	1.255397	0.0125	1.0875
29	63.032289	0.6566	0.2862	1.237242	0.0133	1.0759
30	63.821130	0.6648	0.2862	1.203431	0.0133	1.0759
31	64.942290	0.6765	0.2631	1.137188	0.0138	1.0682
32	65.829529	0.6857	0.2631	1.043253	0.0167	1.0147
33	66.816439	0.6960	0.2631	1.043149	0.0326	1.0143
34	67.792590	0.7062	0.2631	1.014065	0.0326	1.0077
35	68.641746	0.7150	0.2631	0.984193	0.0326	0.9841

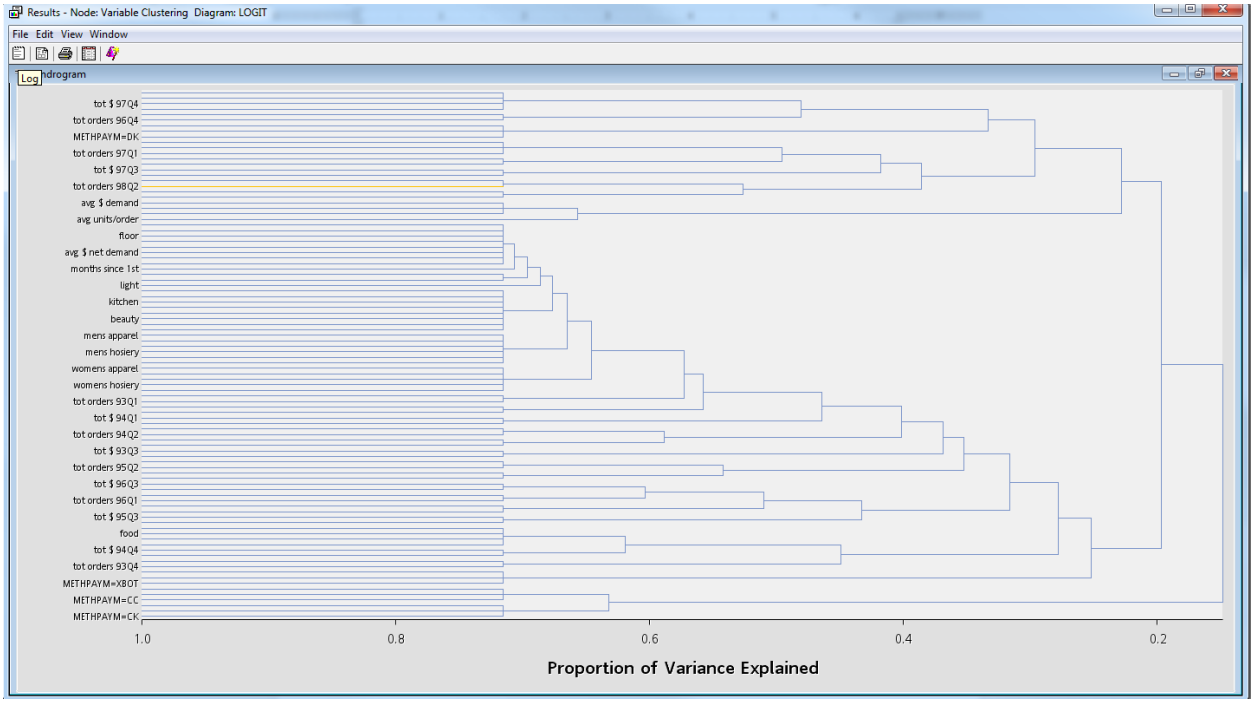
Cluster explains 70.81% of variation in data. Containing 35 clusters.



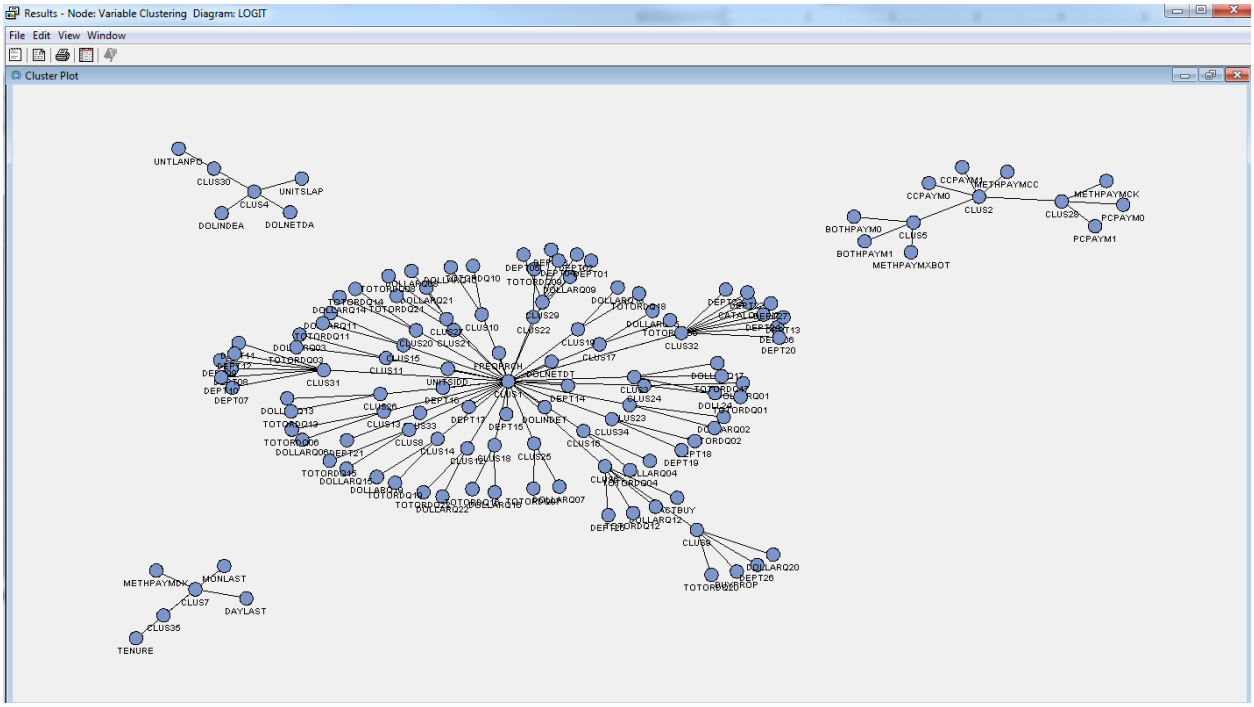
Cluster	Variable	Label	R-Square With Own Cluster Component	Next Closest Cluster
CLUS1	DOLINDET	total $ demand	0.923851	CLUS32
CLUS10	DOLLARQ10	tot $ 95Q2	0.880676	CLUS1
CLUS11	DOLLARQ03	tot $ 93Q3	0.881598	CLUS1
CLUS12	TOTORDQ22	tot orders 98Q2	0.863458	CLUS7
CLUS13	TOTORDQ06	tot orders 94Q2	0.878785	CLUS1
CLUS14	TOTORDQ19	tot orders 97Q3	0.874679	CLUS1
CLUS15	TOTORDQ11	tot orders 95Q3	0.872366	CLUS32
CLUS16	DOLLARQ04	tot $ 93Q4	0.876755	CLUS1
CLUS17	TOTORDQ05	tot orders 94Q1	0.871852	CLUS32
CLUS18	DOLLARQ16	tot $ 96Q4	0.866462	CLUS1
CLUS19	TOTORDQ18	tot orders 97Q2	0.879799	CLUS1
CLUS2	CCPAYM0	CCPAYM=0	1	CLUS28
CLUS20	TOTORDQ14	tot orders 96Q2	0.84481	CLUS32
CLUS21	TOTORDQ21	tot orders 98Q1	0.860266	CLUS1
CLUS22	DOLLARQ09	tot $ 95Q1	0.873589	CLUS1
CLUS23	DOLLARQ02	tot $ 93Q2	0.869018	CLUS1
CLUS24	TOTORDQ01	tot orders 93Q1	0.875695	CLUS1
CLUS25	TOTORDQ07	tot orders 94Q3	0.869525	CLUS1
CLUS26	TOTORDQ13	tot orders 96Q1	0.845289	CLUS32
CLUS27	DOLLARQ08	tot $ 94Q4	0.853544	CLUS1
CLUS28	METHPAYMCK	METHPAYM=CK	1	CLUS2
CLUS29	DEPT03	womens underwear	0.473772	CLUS1
CLUS3	DOLLARQ17	tot $ 97Q1	0.81237	CLUS1
CLUS30	UNTLANPO	avg units/order	1	CLUS4
CLUS31	DEPT12	mens misc	0.367295	CLUS1
CLUS32	CATALOGCNT	number of catalogs receiv...	0.789691	CLUS1
CLUS33	DEPT21	light	1	CLUS1
CLUS34	DEPT19	window	0.532576	CLUS1
CLUS35	TENURE	months since 1st	1	CLUS7
CLUS4	DOLINDEA	avg $ demand	0.912853	CLUS30
CLUS5	BOTHPAYM0	BOTHPAYM=0	1	CLUS2
CLUS6	TOTORDQ12	tot orders 95Q4	0.732581	CLUS32
CLUS7	MONLAST	months since last	0.95065	CLUS35
CLUS8	TOTORDQ15	tot orders 96Q3	0.872407	CLUS1
CLUS9	TOTORDQ20	tot orders 97Q4	0.821266	CLUS7

Above picture shows the 35 clusters.

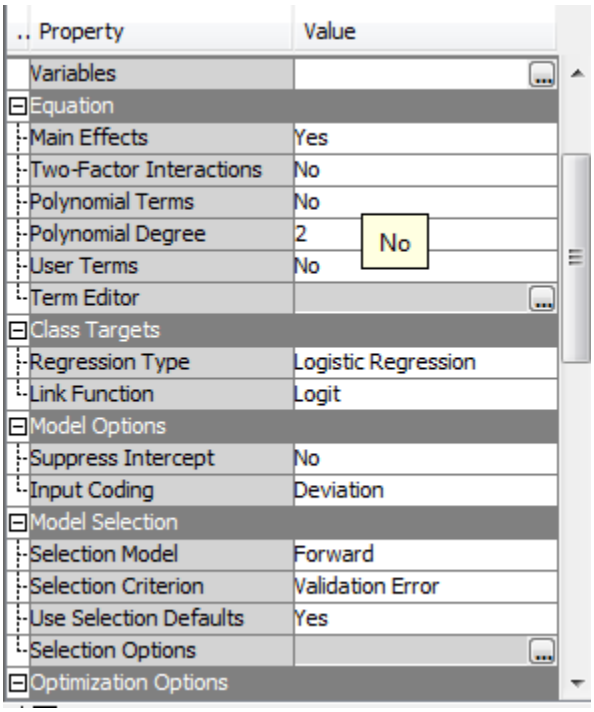
Dendrogram of cluster.



Cluster:



➢ LOGISTIC REGRESSION MODEL:



Summary of Forward selection:

Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq	Validation Error Rate
1	DOLINDET	1	1	418.2287	<.0001	6878.3
2	TOTORDQ20	1	2	178.2565	<.0001	6847.2
3	MONLAST	1	3	113.3610	<.0001	6781.4
4	TOTORDQ22	1	4	47.4870	<.0001	6751.4
5	CATALOGCNT	1	5	36.9828	<.0001	6727.5
6	TOTORDQ18	1	6	19.9779	<.0001	6719.0
7	TOTORDQ21	1	7	14.9769	0.0001	6712.7
8	TOTORDQ12	1	8	13.5709	0.0002	6701.1
9	TOTORDQ19	1	9	11.8344	0.0006	6702.1
10	DEPT03	1	10	10.4403	0.0012	6701.4
11	CCPAYMO	1	11	9.3003	0.0023	6709.1
12	TOTORDQ05	1	12	6.4600	0.0110	6716.5
13	DOLLARQ09	1	13	5.3211	0.0211	6717.9

T-test Analysis:

Since the DOLLARQ17 & BOTHAPAYM0 has a P-value>0.0001, they are rejected from the model.

Data Role=VALIDATE Target Variable=RESPOND Target Label=response target

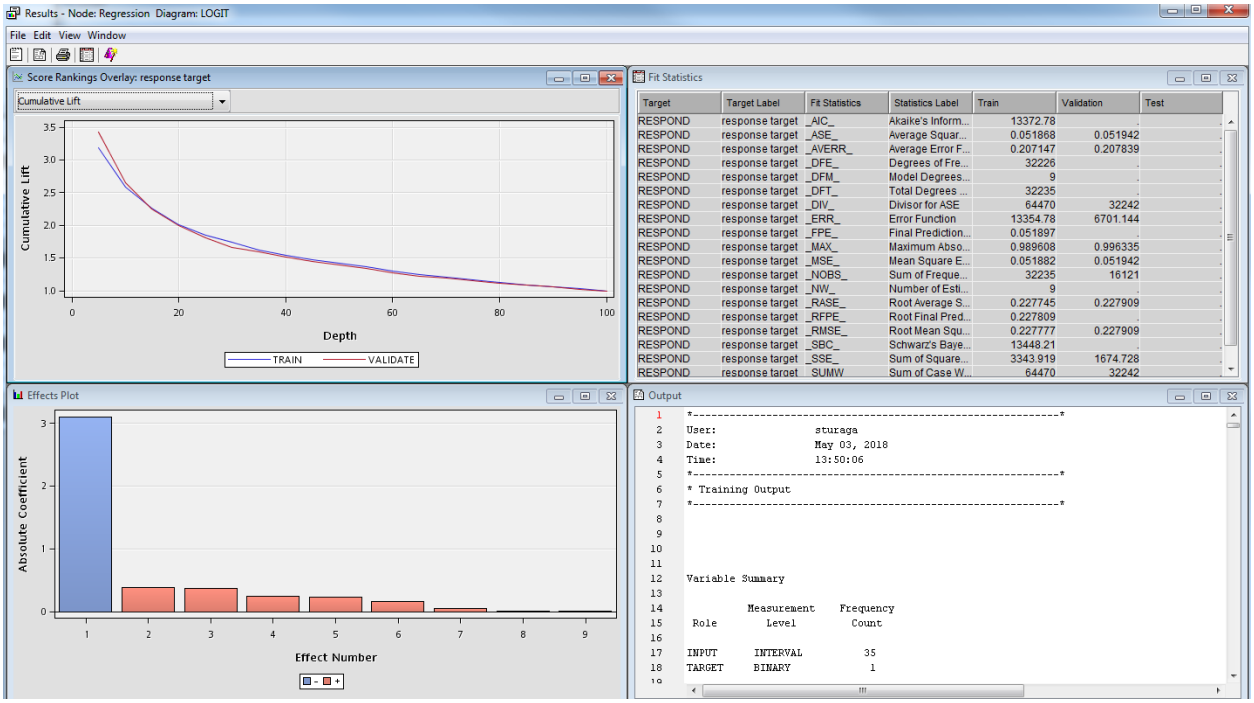
Depth	Gain	Lift	Cumulative Lift	% Response	Cumulative % Response	Number of Observations	Mean Posterior Probability
5	243.140	3.43140	3.43140	19.4548	19.4548	807	0.18015
10	164.623	1.86007	2.64623	10.5459	15.0031	806	0.10221
15	125.304	1.46617	2.25304	8.3127	12.7739	806	0.08419
20	99.622	1.22546	1.99622	6.9479	11.3178	806	0.07390
25	80.710	1.05039	1.80710	5.9553	10.2456	806	0.06746
30	66.643	0.96286	1.66643	5.4591	9.4480	806	0.06347
35	59.719	1.18169	1.59719	6.6998	9.0555	806	0.05996
40	52.065	0.98474	1.52065	5.5831	8.6215	806	0.05518
45	44.409	0.83156	1.44409	4.7146	8.1875	806	0.05136
50	39.379	0.94098	1.39379	5.3350	7.9022	806	0.04757
55	33.472	0.74403	1.33472	4.2184	7.5674	806	0.04496
60	27.639	0.63461	1.27639	3.5980	7.2366	806	0.04298
65	22.366	0.59085	1.22366	3.3499	6.9377	806	0.04116
70	19.097	0.76591	1.19097	4.3424	6.7523	806	0.03973
75	15.680	0.67838	1.15680	3.8462	6.5586	806	0.03797
80	11.869	0.54708	1.11869	3.1017	6.3426	806	0.03586
85	8.378	0.52520	1.08378	2.9777	6.1446	806	0.03348
90	6.552	0.75497	1.06552	4.2804	6.0411	806	0.03016
95	2.268	0.25166	1.02268	1.4268	5.7982	806	0.02514
100	0.000	0.56896	1.00000	3.2258	5.6696	806	0.01863

From the above results we can see that the posterior probability is best for 5% depth in the data set.

Interpretation from the table:

We can clear say interpret that as we go deeper into the data set the probability decreases.

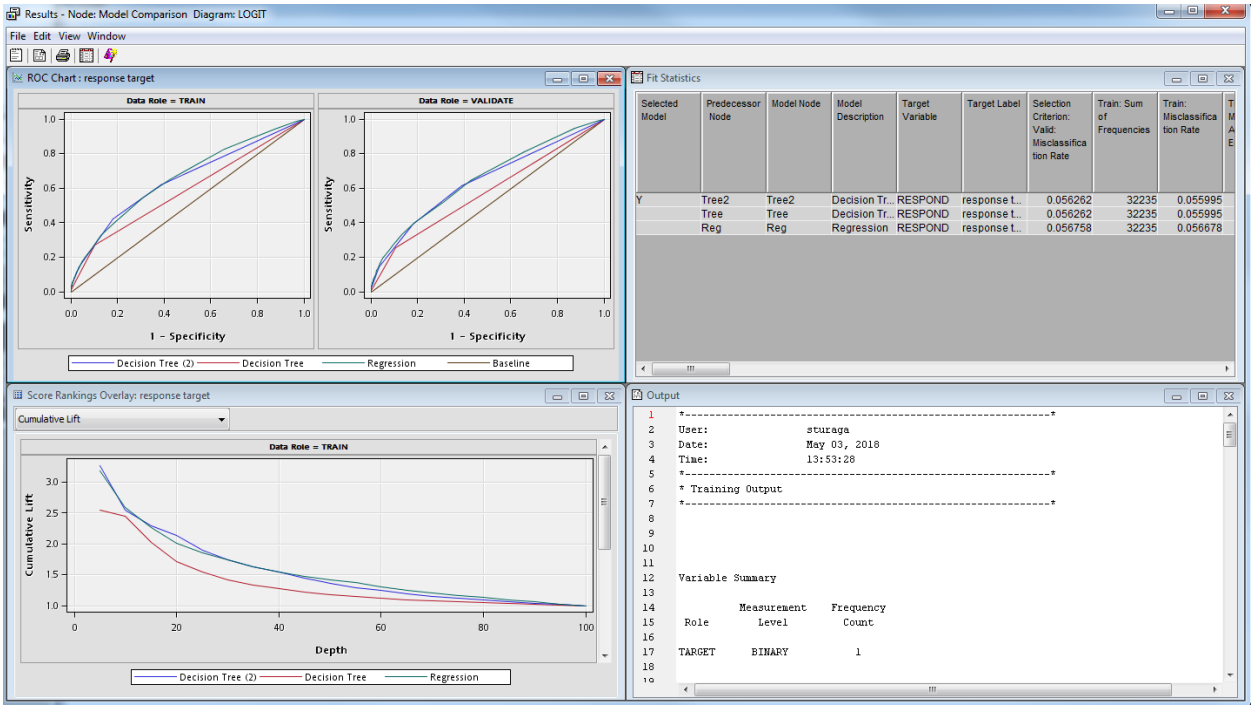
➢ COMPLETE OUTPUT OF REGRESSION:



➔ MODEL COMPARISION RESULTS:

Data Role=Valid

Statistics	Tree2	Tree	Reg
Valid: Kolmogorov-Smirnov Statistic	0.22	0.15	0.22
Valid: Average Squared Error	0.05	0.05	0.05
Valid: Roc Index	0.64	0.58	0.65
Valid: Average Error Function	.	.	0.21
Valid: Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff	0.05	0.09	0.06
Valid: Cumulative Percent Captured Response	24.16	22.63	26.48
Valid: Percent Captured Response	8.53	10.97	9.30
Valid: Divisor for VASE	32242.00	32242.00	32242.00
Valid: Error Function	.	.	6701.14
Valid: Gain	141.49	126.21	164.62
Valid: Gini Coefficient	0.28	0.15	0.31
Valid: Bin-Based Two-Way Kolmogorov-Smirnov Statistic	0.22	0.15	0.22
Valid: Kolmogorov-Smirnov Probability Cutoff	0.04	0.12	0.05
Valid: Cumulative Lift	2.41	2.26	2.65
Valid: Lift	1.71	2.20	1.86
Valid: Maximum Absolute Error	0.96	0.95	1.00
Valid: Misclassification Rate	0.06	0.06	0.06
Valid: Mean Square Error	.	.	0.05
Valid: Sum of Frequencies	16121.00	16121.00	16121.00
Valid: Root Average Squared Error	0.23	0.23	0.23
Valid: Cumulative Percent Response	13.69	12.83	15.00
Valid: Percent Response	9.67	12.45	10.55
Valid: Root Mean Square Error	.	.	0.23
Valid: Sum of Squared Errors	1676.25	1693.19	1674.73
Valid: Sum of Case Weights Times Freq	.	.	32242.00



From the results and graphs we can see that Decision tree 3 and regression give almost similar predictions, as the area under the curve is similar and the lines are closer to 1.