

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.

Explore - PEOJDAT.CATALOG2010

File View Actions Window

Sample Properties

Property	Value
Rows	48356
Columns	98
Library	PEOJDAT
Member	CATALOG2010
Type	DATA
Sample Method	Random
Fetch Size	Max
Fetched Rows	20000
Random Seed	12345

Sample Statistics

Obs #	Variable	Label	Type	Percent	Minimum	Maximum	Mean	Number
1	BOTHPA	check & card only	CLASS	0	.	.	2	
2	BCPPAYM	card only	CLASS	0	.	.	4	
3	3MTHPA	payment	CLASS	0	.	.	2	
4	4PCPAYM	check only	CLASS	0	.	.	2	
5	5STATE	state	CLASS	0	.	.	51	
6	6ZIP	zip code	CLASS	0	.	.	128+	
7	7ACTBUY	num qtrr... VAR	VAR	0	0	10	0.98995	
8	8BUYPROP	% quartile ... VAR	VAR	0	0	1	0.98992	
9	9CART	count of cart	VAR	0	1	23	3.765	
10	10COUNTY	county c...	VAR	0	10	999	432.759	
11	11CUST	ID customer	VAR	0	1	48354	24450.64	
12	12DAYLAST	days sinc...	VAR	0	0	7887	1180.701	
13	13DEPT01	women... VAR	VAR	0	0	69	6.68	
14	14DEPT02	women... VAR	VAR	0	0	20	0.2885	
15	15DEPT03	women... VAR	VAR	0	0	54	1.0619	
16	16DEPT04	women... VAR	VAR	0	0	47	0.67875	
17	17DEPT05	women... VAR	VAR	0	0	28	0.53875	

PEOJDAT.CATALOG2010

Obs #	custom	num qr...	check	% quar...	numbe...	card only	days si...	avg \$ d...	total \$...	\$ last 2...	\$ tot n...	avg \$ n...	DTBUY...	DTBUY...	lifetime...	payme...	months...	doles...
1	0.0001	31	16%	80	146	\$15.99	\$319.70	\$47.75	\$15.34	\$306.75	25Mar08	15Aug89	20XBOT	5				
2	0.0001	41	33%	140	256	\$15.99	\$57.33	\$57.33	\$15.99	\$57.33	15Apr08	15Aug89	17CK	19	\$			
3	0.0005	40	36%	141	564	\$109.91	\$127.55	\$159.60	\$127.55	\$159.60	15Apr08	15Dec97	13CK	43	\$			
4	0.0007	00	0%	50	1312	\$64.67	\$194.00	\$0.00	\$64.67	\$194.00	14Jan05	20Jul09	3CK	43	\$			
5	0.0012	31	60%	30	101	\$56.88	\$739.50	\$246.00	\$56.88	\$739.50	09May08	14Oct03	13XBOT	3				
6	0.0001	00	0%	81	738	\$15.99	\$100.00	\$100.00	\$15.99	\$100.00	21Jun08	21Jun08	20CC	26				
7	0.0014	21	100%	30	137	\$46.79	\$194.90	\$194.90	\$46.79	\$194.90	03Apr08	03Apr08	4XBOT	4				
8	0.0017	10	8%	61	1317	\$33.27	\$99.80	\$0.00	\$33.27	\$99.80	09Jan05	15Aug95	3CC	43				
9	0.0018	30	23%	81	50	\$26.17	\$471.11	\$25.85	\$26.17	\$471.11	29Jun08	15Dec95	18CC	8				
10	0.0001	00	0%	71	173	\$33.27	\$299.00	\$69.00	\$33.27	\$299.00	03Aug08	25Aug09	20CC	8				
11	0.0020	00	0%	10	4424	\$27.67	\$83.00	\$0.00	\$27.67	\$83.00	08Jul98	25Feb95	3DK	145				
12	0.0021	30	19%	110	313	\$36.66	\$623.15	\$87.95	\$36.66	\$623.15	09Oct07	15Aug92	17CK	10				
13	0.0023	00	0%	20	3112	\$7.59	\$22.00	\$0.00	\$7.59	\$22.77	09Feb08	25Sep08	3CK	102				
14	0.0001	21	14%	30	189	\$97.48	\$205.85	\$0.00	\$97.48	\$205.85	05Dec08	15Dec94	18XBOT	6				
15	0.0027	10	8%	81	1525	\$55.47	\$1,275.75	\$0.00	\$54.38	\$1,250.80	15Jun04	15Aug95	23CC	50				
16	0.0028	10	8%	20	1303	\$72.30	\$216.90	\$0.00	\$72.30	\$216.90	23Jan05	15Dec95	3CK	43				
17	0.0030	10	6%	41	71	\$86.05	\$240.15	\$56.70	\$86.05	\$240.15	08Feb08	10Feb91	3CC	2				

Task View

File View Actions Window

Start Task View

Windows Taskbar

4/30/2018 12:22 PM

Creating a Pie Chart:

Sequence of steps to create pie chart is given below:

Explore - PEO/DAT.CATALOG2010

File View Actions Window

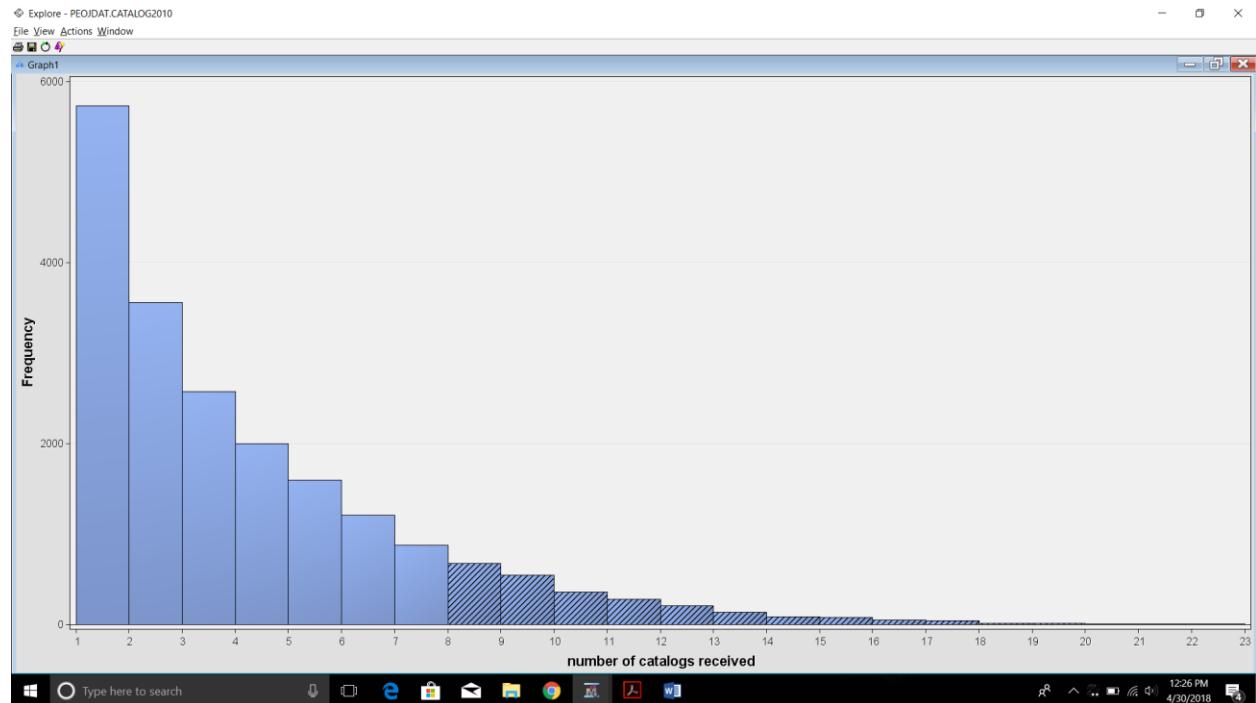
Graph2

0
1

response target = 0
Frequency = 10954

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.

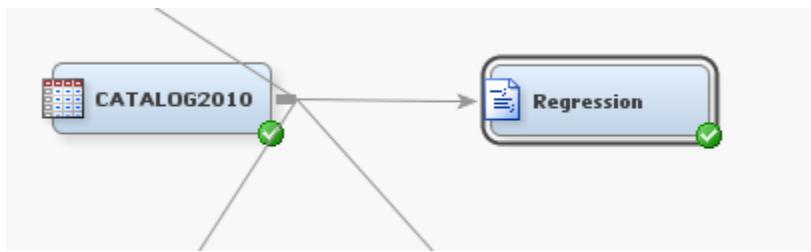
TASK 1: Exploratory analysis of CATALOG2010 Dataset.

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:

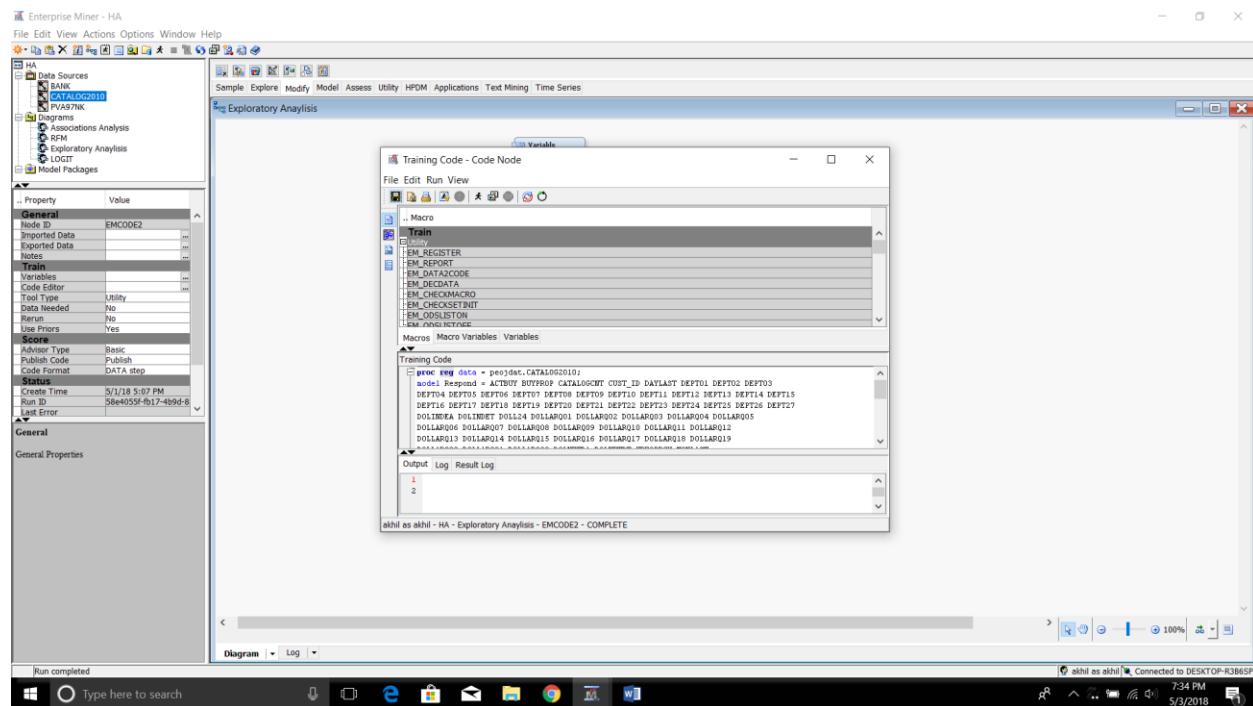
```

proc reg data = peojdat.CATALOG2010;

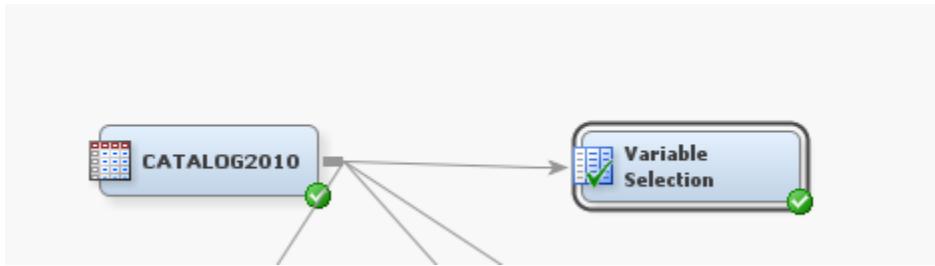
model Respond = ACTBUY BUYPROP CATALOGCNT CUST_ID DAYLAST DEPT01 DEPT02 DEPT03
DEPT04 DEPT05 DEPT06 DEPT07 DEPT08 DEPT09 DEPT10 DEPT11 DEPT12 DEPT13 DEPT14
DEPT15DEPT16 DEPT17 DEPT18 DEPT19 DEPT20 DEPT21 DEPT22 DEPT23 DEPT24 DEPT25
DEPT26 DEPT27 DOLINDEA DOLINDET DOLL24 DOLLARQ01 DOLLARQ02 DOLLARQ03 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
```

/ vif tol collin;

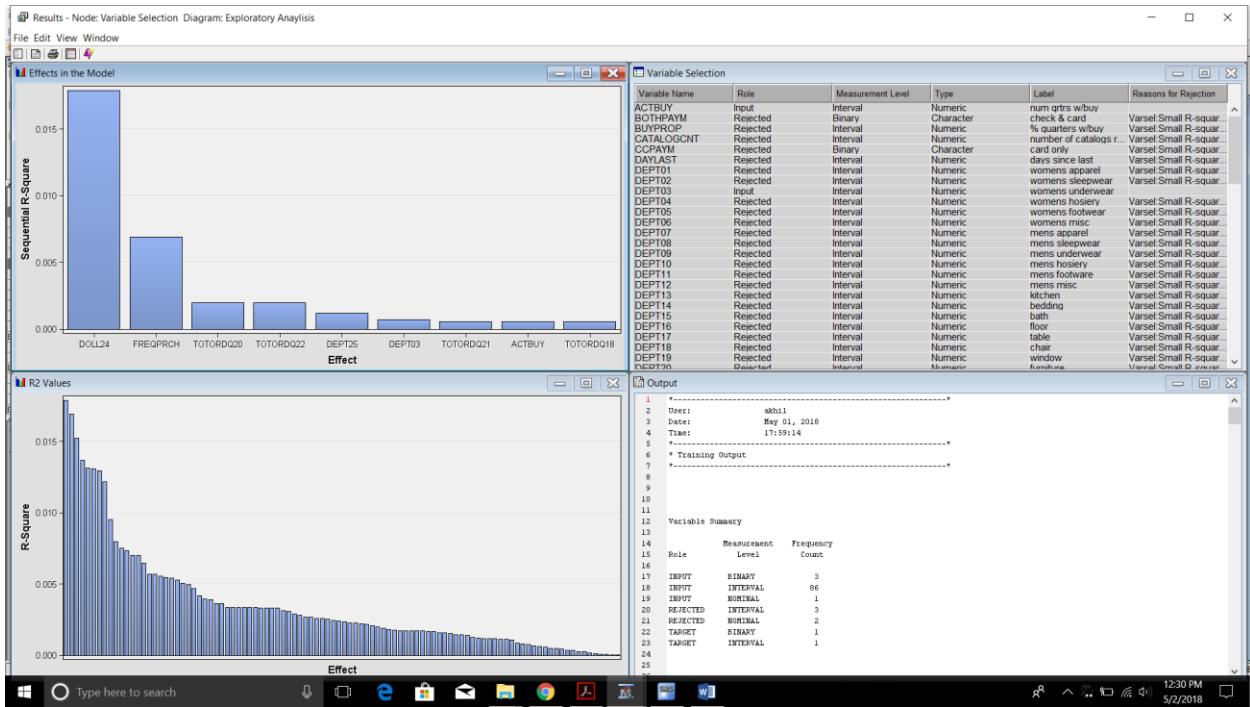
run;



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:



Results - Node: Variable Selection Diagram: Exploratory Analysis						
File Edit View Window						
Variable Selection						
Variable Name	Role	Measurement Level	Type	Label	Reasons for Rejection	
ACTBUY	Input	Interval	Numeric	num qfrs wibuy		
CUSTOMM	Rejected	Binary	Character	% cust & care	Var sel. Small R-square value	
BUYPROP	Rejected	Interval	Numeric	% quest wibuy	Var sel. Small R-square value	
CATALOGCNT	Rejected	Interval	Numeric	number of catalogs received	Var sel. Small R-square value	
CCPAYM	Rejected	Binary	Character	card only	Var sel. Small R-square value	
DATASAST	Rejected	Interval	Numeric	days since last	Var sel. Small R-square value	
DEPT01	Rejected	Interval	Numeric	womens apparel	Var sel. Small R-square value	
DEPT02	Rejected	Interval	Numeric	womens sleepwear	Var sel. Small R-square value	
DEPT03	Rejected	Interval	Numeric	womens underwear	Var sel. Small R-square value	
DEPT04	Rejected	Interval	Numeric	womens wear	Var sel. Small R-square value	
DEPT05	Rejected	Interval	Numeric	womens footwear	Var sel. Small R-square value	
DEPT06	Rejected	Interval	Numeric	womens misc	Var sel. Small R-square value	
DEPT07	Rejected	Interval	Numeric	mens apparel	Var sel. Small R-square value	
DEPT08	Rejected	Interval	Numeric	mens sleepwear	Var sel. Small R-square value	
DEPT09	Rejected	Interval	Numeric	mens underwear	Var sel. Small R-square value	
DEPT10	Rejected	Interval	Numeric	mens hosiery	Var sel. Small R-square value	
DEPT11	Rejected	Interval	Numeric	mens footware	Var sel. Small R-square value	
DEPT12	Rejected	Interval	Numeric	mens misc	Var sel. Small R-square value	
DEPT13	Rejected	Interval	Numeric	kitchen	Var sel. Small R-square value	
DEPT14	Rejected	Interval	Numeric	bedding	Var sel. Small R-square value	
DEPT15	Rejected	Interval	Numeric	bath	Var sel. Small R-square value	
DEPT16	Rejected	Interval	Numeric	floor	Var sel. Small R-square value	
DEPT17	Rejected	Interval	Numeric	table	Var sel. Small R-square value	
DEPT18	Rejected	Interval	Numeric	chair	Var sel. Small R-square value	
DEPT19	Rejected	Interval	Numeric	window	Var sel. Small R-square value	
DEPT20	Rejected	Interval	Numeric	furniture	Var sel. Small R-square value	
DEPT21	Rejected	Interval	Numeric	light	Var sel. Small R-square value	
DEPT22	Rejected	Interval	Numeric	household	Var sel. Small R-square value	
DEPT23	Rejected	Interval	Numeric	beauty	Var sel. Small R-square value	
DEPT24	Rejected	Interval	Numeric	health	Var sel. Small R-square value	
DEPT25	Input	Interval	Numeric	food		
DEPT26	Rejected	Interval	Numeric	gift	Var sel. Small R-square value	
DEPT27	Rejected	Interval	Numeric	outdoor	Var sel. Small R-square value	
DOLINDEA	Rejected	Interval	Numeric	avg d demand	Var sel. Small R-square value	
DOLINDET	Rejected	Interval	Numeric	total d demand	Var sel. Small R-square value	
DOLDET	Input	Interval	Numeric	ts last 24 months		
DOLLAR1	Rejected	Interval	Numeric	tot \$ 3.02	Var sel. Small R-square value	
DOLLARQ01	Rejected	Interval	Numeric	tot \$ 932	Var sel. Small R-square value	
DOLLARQ02	Rejected	Interval	Numeric	tot \$ 932	Var sel. Small R-square value	
DOLLARQ03	Rejected	Interval	Numeric	tot \$ 933	Var sel. Small R-square value	
DOLLARQ04	Rejected	Interval	Numeric	tot \$ 9324	Var sel. Small R-square value	
DOLLARQ05	Rejected	Interval	Numeric	tot \$ 9401	Var sel. Small R-square value	
DOLLARQ06	Rejected	Interval	Numeric	tot \$ 9402	Var sel. Small R-square value	
DOLLARQ07	Rejected	Interval	Numeric	tot \$ 9403	Var sel. Small R-square value	
DOLLARQ08	Rejected	Interval	Numeric	tot \$ 9404	Var sel. Small R-square value	
DOLLARQ09	Rejected	Interval	Numeric	tot \$ 9501	Var sel. Small R-square value	
DOLLARQ10	Rejected	Interval	Numeric	tot \$ 9502	Var sel. Small R-square value	
DOLLARQ11	Rejected	Interval	Numeric	tot \$ 9503	Var sel. Small R-square value	
DOLLARQ12	Rejected	Interval	Numeric	tot \$ 9504	Var sel. Small R-square value	
DOLLARQ13	Rejected	Interval	Numeric	tot \$ 9601	Var sel. Small R-square value	
DOLLARQ14	Rejected	Interval	Numeric	tot \$ 9602	Var sel. Small R-square value	
DOLLARQ15	Rejected	Interval	Numeric	tot \$ 9603	Var sel. Small R-square value	
DOLLARQ16	Rejected	Interval	Numeric	tot \$ 9604	Var sel. Small R-square value	
DOLLARQ17	Rejected	Interval	Numeric	tot \$ 9701	Var sel. Small R-square value	
DOLLARQ18	Rejected	Interval	Numeric	tot \$ 9702	Var sel. Small R-square value	
DOLLARQ19	Rejected	Interval	Numeric	tot \$ 9703	Var sel. Small R-square value	

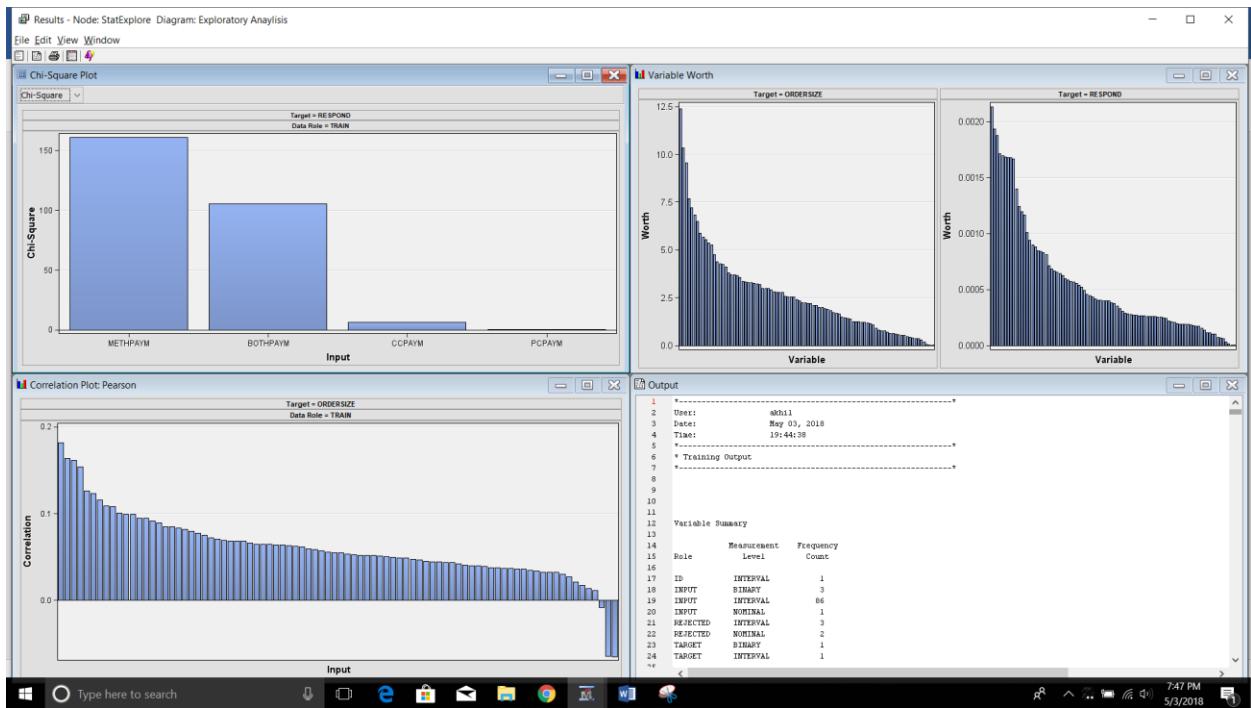
Results - Node: Variable Selection Diagram: Exploratory Analysis						
File Edit View Window		Variable Selection				
Variable Name	Role	Measurement Level	Type	Label		Reasons for Rejection
DOLLARQ01	Rejected	Interval	Numeric	total 3 years or less	\$ less than 24 months	VarSel:Small R-square value
DOLLARQ02	Rejected	Interval	Numeric	total \$ 93Q1	\$ total 93Q2	VarSel:Small R-square value
DOLLARQ03	Rejected	Interval	Numeric	total \$ 93Q2	\$ total 93Q3	VarSel:Small R-square value
DOLLARQ04	Rejected	Interval	Numeric	total \$ 93Q3	\$ total 93Q4	VarSel:Small R-square value
DOLLARQ05	Rejected	Interval	Numeric	total \$ 94Q1	\$ total 94Q2	VarSel:Small R-square value
DOLLARQ06	Rejected	Interval	Numeric	total \$ 94Q2	\$ total 94Q3	VarSel:Small R-square value
DOLLARQ07	Rejected	Interval	Numeric	total \$ 94Q3	\$ total 94Q4	VarSel:Small R-square value
DOLLARQ08	Rejected	Interval	Numeric	total \$ 94Q4	\$ total 95Q1	VarSel:Small R-square value
DOLLARQ09	Rejected	Interval	Numeric	total \$ 95Q1	\$ total 95Q2	VarSel:Small R-square value
DOLLARQ10	Rejected	Interval	Numeric	total \$ 95Q2	\$ total 95Q3	VarSel:Small R-square value
DOLLARQ11	Rejected	Interval	Numeric	total \$ 95Q3	\$ total 95Q4	VarSel:Small R-square value
DOLLARQ12	Rejected	Interval	Numeric	total \$ 95Q4	\$ total 96Q1	VarSel:Small R-square value
DOLLARQ13	Rejected	Interval	Numeric	total \$ 96Q1	\$ total 96Q2	VarSel:Small R-square value
DOLLARQ14	Rejected	Interval	Numeric	total \$ 96Q2	\$ total 96Q3	VarSel:Small R-square value
DOLLARQ15	Rejected	Interval	Numeric	total \$ 96Q3	\$ total 96Q4	VarSel:Small R-square value
DOLLARQ16	Rejected	Interval	Numeric	total \$ 96Q4	\$ total 97Q1	VarSel:Small R-square value
DOLLARQ17	Rejected	Interval	Numeric	total \$ 97Q1	\$ total 97Q2	VarSel:Small R-square value
DOLLARQ18	Rejected	Interval	Numeric	total \$ 97Q2	\$ total 97Q3	VarSel:Small R-square value
DOLLARQ19	Rejected	Interval	Numeric	total \$ 97Q3	\$ total 97Q4	VarSel:Small R-square value
DOLLARQ20	Rejected	Interval	Numeric	total \$ 97Q4	\$ total 98Q1	VarSel:Small R-square value
DOLLARQ21	Rejected	Interval	Numeric	total \$ 98Q1	\$ total 98Q2	VarSel:Small R-square value
DOLLARQ22	Rejected	Interval	Numeric	total \$ 98Q2	\$ total 98Q3	VarSel:Small R-square value
DOLNETD1	Rejected	Interval	Numeric	total \$ 98Q3	\$ total 98Q4	VarSel:Small R-square value
DOLNETDT	Rejected	Interval	Numeric	avg \$ net demand	total \$ net demand	VarSel:Small R-square value
FREQPRCH	Input	Interval	Numeric	lifetime orders	total \$ net demand	VarSel:Small R-square value
RETRNPRYM	Rejected	Character	Character	payment method	months since last	VarSel:Small R-square value
MONLAST	Rejected	Interval	Numeric	months since last	check only	VarSel:Small R-square value
PCPAYM	Rejected	Binary	Character	months since last	months since 1st	VarSel:Small R-square value
TENURE	Rejected	Interval	Numeric	months since 1st	total orders	VarSel:Small R-square value
TOTORQ001	Rejected	Interval	Numeric	total orders	total orders 93Q1	VarSel:Small R-square value
TOTORQ002	Rejected	Interval	Numeric	total orders	total orders 93Q2	VarSel:Small R-square value
TOTORQ003	Rejected	Interval	Numeric	total orders	total orders 93Q3	VarSel:Small R-square value
TOTORQ004	Rejected	Interval	Numeric	total orders	total orders 93Q4	VarSel:Small R-square value
TOTORQ005	Rejected	Interval	Numeric	total orders	total orders 94Q1	VarSel:Small R-square value
TOTORQ006	Rejected	Interval	Numeric	total orders	total orders 94Q2	VarSel:Small R-square value
TOTORQ007	Rejected	Interval	Numeric	total orders	total orders 94Q3	VarSel:Small R-square value
TOTORQ008	Rejected	Interval	Numeric	total orders	total orders 94Q4	VarSel:Small R-square value
TOTORQ009	Rejected	Interval	Numeric	total orders	total orders 95Q1	VarSel:Small R-square value
TOTORQ010	Rejected	Interval	Numeric	total orders	total orders 95Q2	VarSel:Small R-square value
TOTORQ011	Rejected	Interval	Numeric	total orders	total orders 95Q3	VarSel:Small R-square value
TOTORQ012	Rejected	Interval	Numeric	total orders	total orders 95Q4	VarSel:Small R-square value
TOTORQ013	Rejected	Interval	Numeric	total orders	total orders 96Q1	VarSel:Small R-square value
TOTORQ014	Rejected	Interval	Numeric	total orders	total orders 96Q2	VarSel:Small R-square value
TOTORQ015	Rejected	Interval	Numeric	total orders	total orders 96Q3	VarSel:Small R-square value
TOTORQ016	Rejected	Interval	Numeric	total orders	total orders 96Q4	VarSel:Small R-square value
TOTORQ017	Rejected	Interval	Numeric	total orders	total orders 97Q1	VarSel:Small R-square value
TOTORQ018	Input	Interval	Numeric	total orders	total orders 97Q2	VarSel:Small R-square value
TOTORQ019	Rejected	Interval	Numeric	total orders	total orders 97Q3	VarSel:Small R-square value
TOTORQ020	Input	Interval	Numeric	total orders	total orders 97Q4	VarSel:Small R-square value
TOTORQ021	Input	Interval	Numeric	total orders	total orders 98Q1	VarSel:Small R-square value
TOTORQ022	Input	Interval	Numeric	total orders	total orders 98Q2	VarSel:Small R-square value
UNITSIDE	Rejected	Interval	Numeric	total units demand	avg price per unit	VarSel:Small R-square value
BINMAP	Rejected	Interval	Numeric	avg price per unit	units/sales/order	VarSel:Small R-square value
UNITANPO	Rejected	Interval	Numeric	units/sales/order		

- 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:

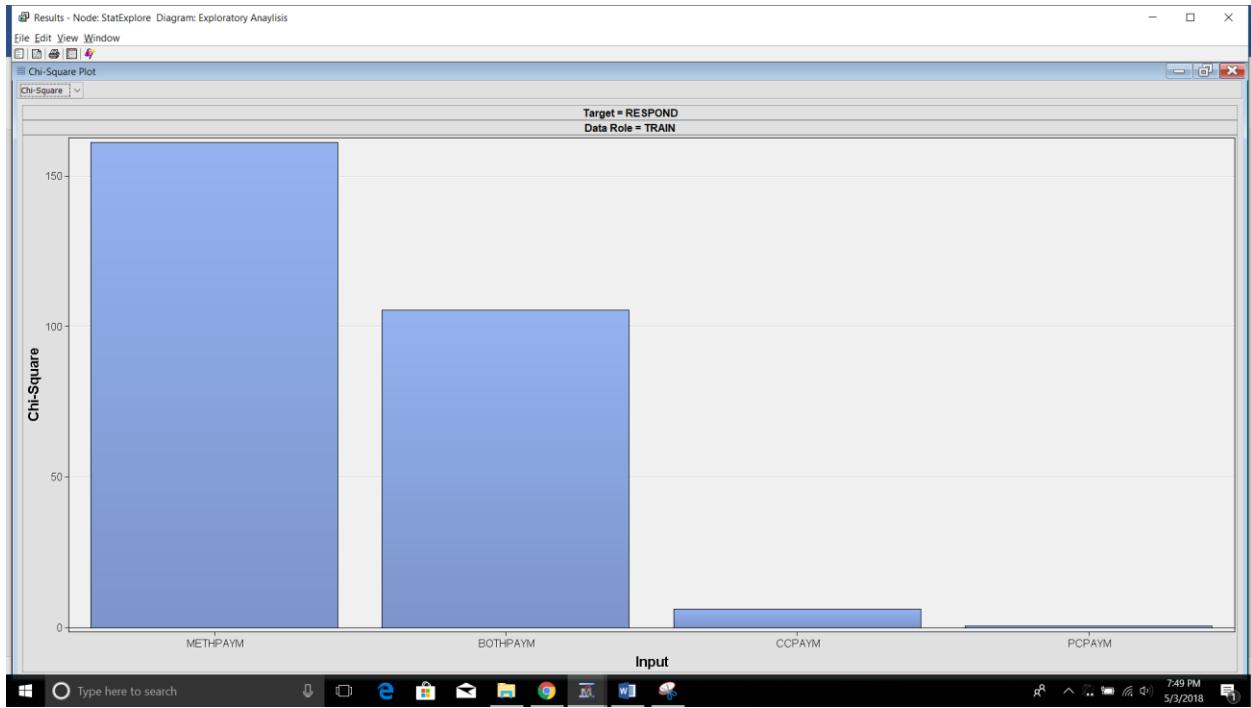


- The StatExplore block can be pulled out of the Explore tab that is a part of the SEMMA tools.
- This block is connected to the CATALOG2010 node as shown above.
- Then we run the StatExplore node to observe the results.
- The results are obtained as shown below:



The different tiles of the above results block are

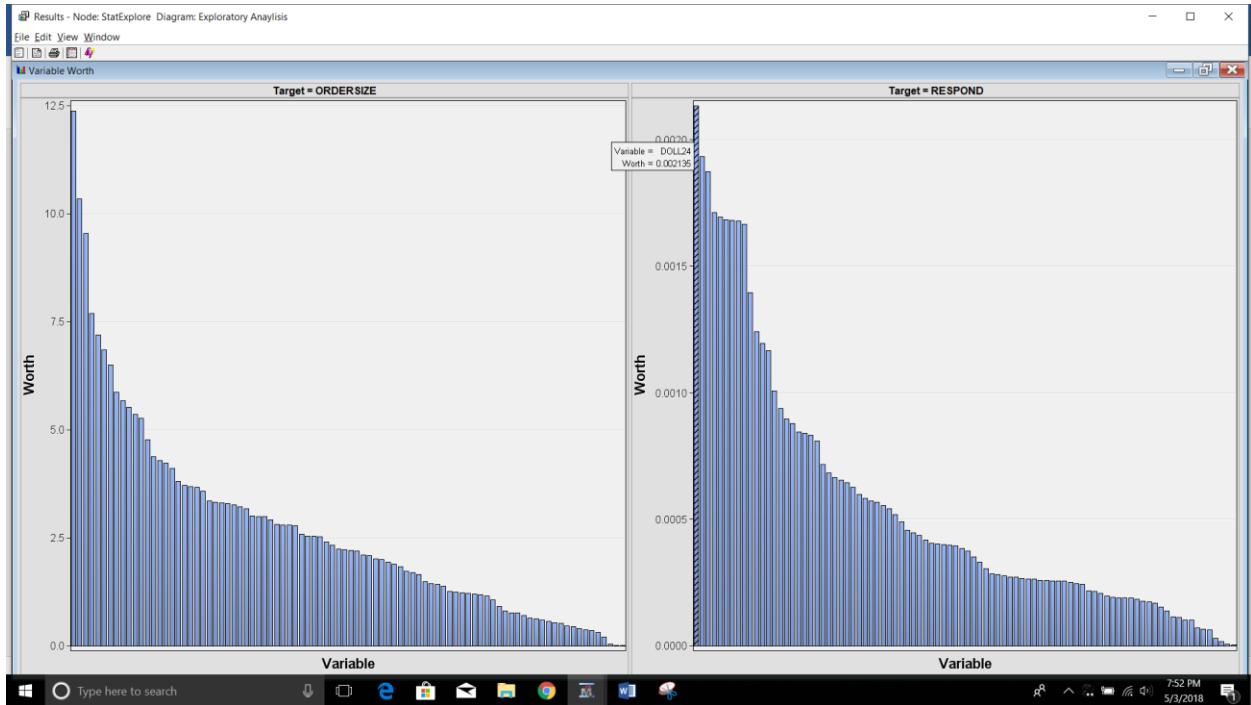
- i) Chi-square plot.



ii) Variable Worth Plot:

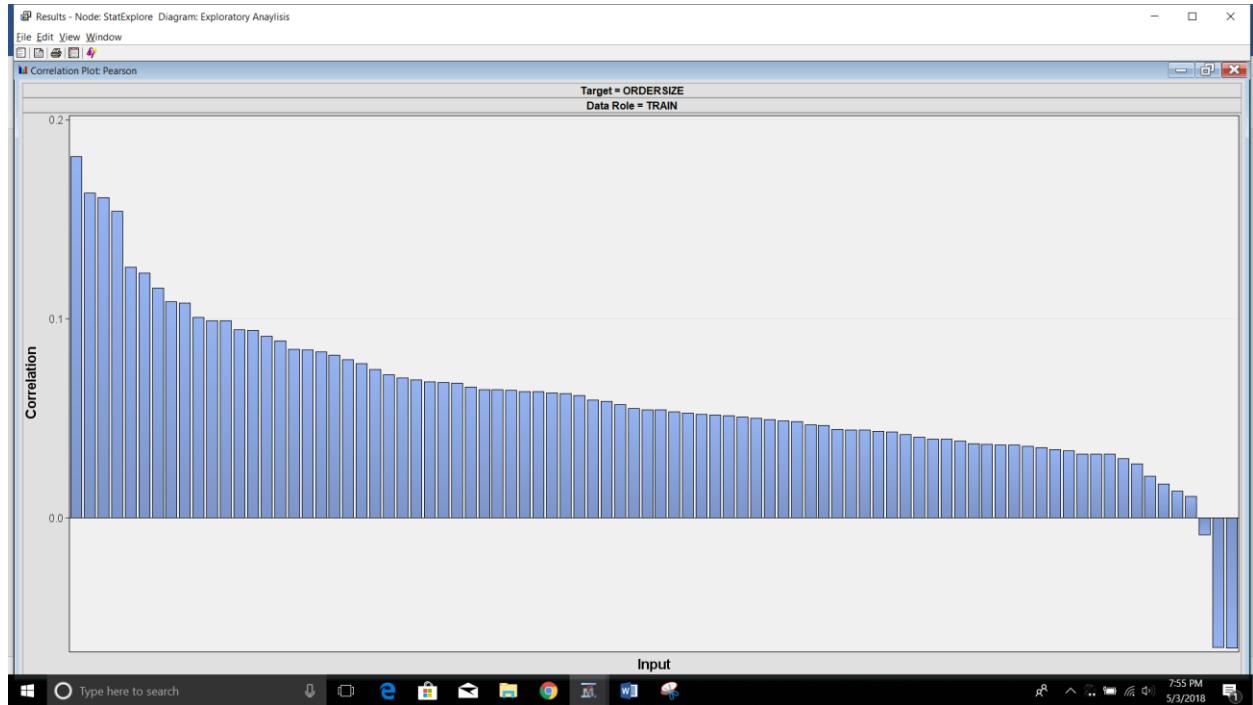
This plot orders the variables based on their true worth in predicting the target variables.

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



- By placing the cursor 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:

Results - Node: StatExplore Diagram: Exploratory Analysis

File Edit View Window

Output

```

1087
1088
1089 Correlation Statistics
1090 (maximum 500 observations printed)
1091
1092 Data Role=TRAIN Type=PEARSON Target=ORDERSIZE
1093
1094 Input Correlation
1095
1096 DOLL24 0.18137
1097 DOL14T 0.14337
1098 DOLMNTDF 0.1160
1099 UNIT3ID0 0.15411
1100 FREQPCH 0.12591
1101 CATALOG2010 0.12319
1102 DEPT12 0.11543
1103 DEPT13QDM 0.10373
1104 DEPT12 0.10799
1105 DEPT03 0.10082
1106 DOLLARQ02 0.09917
1107 ACTR1 0.09712
1108 DEPT13 0.09456
1109 TUT09PQ03 0.09415
1110 DEPT11 0.09131
1111 TUT09PQ02 0.08910
1112 DEPT13 0.08479
1113 DEPT04 0.08453
1114 DOLLARQ19 0.08350
1115 DEPT23 0.08179
1116 DEPT10 0.07961
1117 DOLLARQ18 0.07760
1118 DOLLARQ16 0.07473
1119 DEPT17 0.07200
1120 TUT09PQ01 0.07047
1121 DEPT15 0.06933
1122 DEPT05 0.06841
1123 DOLLARQ12 0.06799
1124 DEPT24 0.06792
1125 TUT09PQ18 0.06599
1126 DOLLARQ03 0.06459
1127 DOLLARQ10 0.06440
1128 DEPT06 0.06435
1129 DOLLARQ15 0.06362
1130 TUT09PQ17 0.06246
1131 DOLLARQ11 0.06200
1132 TUT09PQ18 0.06244
1133 DOLLARQ17 0.06174
1134 DEPT19 0.05926
1135 DEPT09 0.05895
1136 TUT09PQ17 0.05701
1137 TUT09PQ15 0.05515
1138 DOLLARQ08 0.05441
1139 TUT09PQ12 0.05436
1140 DEPT15 0.05373

```

800 PM
5/3/2018

- 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;
```

```
Results - Node: Correlation Diagram: Exploratory Analysis
File Edit View Window
Output
13
14     Measurement Frequency
15     Role Level Count
16
17     ID INTERVAL 1
18     INPUT BINARY 3
19     INPUT INTERVAL 86
20     INPUT NOMINAL 1
21     REJECTED INTERVAL 3
22     REJECTED NOMINAL 2
23     TARGET BINARY 1
24     TARGET INTERVAL 1
25
26
27
28
29
30 The CORR Procedure
31
32
33  92 Variables: CUST_ID ACTBHF BHTP909 CATALOGUQ DIASTC DOLLAR0A DOLLAR0C DOLLAR14 DOLLAR16 DOLLAR17 DOLLAR18 DOLLAR19 DOLLAR20 DOLLAR20Q DOLLAR20Q2 DOLLAR20Q3 DOLLAR20Q4 DOLLAR20Q5 DOLLAR20Q6 DOLLAR20Q7 DOLLAR20Q8 DOLLAR20Q9 DOLLAR20Q10 DOLLAR20Q11 DOLLAR20Q12 DOLLAR20Q13 DOLLAR20Q14 DOLLAR20Q15
34  DOLLARQ16 DOLLARQ17 DOLLARQ18 DOLLARQ19 DOLLARQ20 DOLLARQ21 DOLLARQ22 TUTOR0Q01 TUTOR0Q02 TUTOR0Q03 TUTOR0Q04 TUTOR0Q05 TUTOR0Q06 TUTOR0Q07 TUTOR0Q08 TUTOR0Q09 TUTOR0Q10 TUTOR0Q11 TUTOR0Q12 TUTOR0Q13 TUTOR0Q14
35  TUTOR0Q15 TUTOR0Q16 TUTOR0Q17 TUTOR0Q18 TUTOR0Q19 TUTOR0Q20 TUTOR0Q21 TUTOR0Q22
36
37
38
39
40
41 Simple Statistics
42
43 Variable N Mean Std Dev Sum Minimum Maximum Label
44
45 CUST_ID 48356 24179 13959 1169175546 1.00000 48356 customer number
46 ACTBHF 0.99416 1.16124 47390 0 10.00000 num grts w/bag
47 BUYPROP 48356 0.18658 0.25696 9119 0 1.00000 % quarters w/buy
48 CATALOGUQ 48356 3.76505 3.13325 16232 1.00000 27.00000 number of catalogs received
49 DIASTC 48356 1.180 1.228 578642 0 2.616 avg size last
50 DOLLAR0A 48356 47.74047 37.37377 21289 0 122.00000 avg 1 demand
51 DOLLAR0C 48356 91.67031 314.69097 9510190 1.00000 7979 total # demand
52 DOLLAR14 48356 45.69996 94.26099 209432 0 2434 # last 24 months
53 DOLLAR16 48356 45.69996 36.43548 209432 0 760.58888 tot 4 net demand
54 DOLLAR17 48356 107.35917 302.35363 9084118 0 8029 avg 4 net demand
55 DOLLAR18 48356 13582 1224 681232166 9497 17762
56 DOLLAR19 48356 15226 1829 738472161 9358 17761
57 DOLLAR20 48356 20.44059 201.499 10.00000 190.00000 lifetime orders
58 DOLLAR20Q 48356 59.44059 100.44059 0 27.00000 max since last
59 DOLLAR20Q2 48356 0.05664 0.23116 2739 0 101.00000 dollar value of order
60 DOLLAR20Q3 48356 10.00000 10.00000 999.00000 response target
61 DOLLAR20Q4 48356 0.51000 0.444 4925.00000 county code
62 DOLLAR20Q5 48356 10.00000 10.00000 27.00000 state
63 DOLLAR20Q6 48356 17.15405 5282358 1.00000 432.00000 tot units demand
64 DOLLAR20Q7 48356 22.04998 20.43467 1066249 1.00000 768.50000 avg price/unit
65 DOLLAR20Q8 48356 2.37118 128052 0 121.00000 avg units/order
66 DPT01 48356 0.49479 1.79330 0 59.00000 women apparel
67 DPT02 48356 0.39223 1.15360 1432 0 24.00000 women wear
68 DPT03 48356 0.49479 1.79330 0 59.00000 men apparel
69 DPT04 48356 0.39223 1.15360 1432 0 24.00000 men wear
70 DPT05 48356 0.49479 1.79330 0 59.00000 children apparel
71 DPT06 48356 0.39223 1.15360 1432 0 24.00000 children wear
72 DPT07 48356 0.49479 1.79330 0 59.00000 men accessories
73 DPT08 48356 0.39223 1.15360 1432 0 24.00000 men shoes
74 DPT09 48356 0.49479 1.79330 0 59.00000 women accessories
75 DPT10 48356 0.39223 1.15360 1432 0 24.00000 women shoes
76 DPT11 48356 0.49479 1.79330 0 59.00000 unisex apparel
77 DPT12 48356 0.39223 1.15360 1432 0 24.00000 unisex shoes
78 DPT13 48356 0.49479 1.79330 0 59.00000 children accessories
79 DPT14 48356 0.39223 1.15360 1432 0 24.00000 children shoes
80 DPT15 48356 0.49479 1.79330 0 59.00000 men sportswear
81 DPT16 48356 0.39223 1.15360 1432 0 24.00000 women sportswear
82 DPT17 48356 0.49479 1.79330 0 59.00000 unisex sportswear
83 DPT18 48356 0.39223 1.15360 1432 0 24.00000 men activewear
84 DPT19 48356 0.49479 1.79330 0 59.00000 women activewear
85 DPT20 48356 0.39223 1.15360 1432 0 24.00000 unisex activewear
86 DPT21 48356 0.49479 1.79330 0 59.00000 men swimwear
87 DPT22 48356 0.39223 1.15360 1432 0 24.00000 women swimwear
88 DPT23 48356 0.49479 1.79330 0 59.00000 unisex swimwear
89 DPT24 48356 0.39223 1.15360 1432 0 24.00000 men outerwear
90 DPT25 48356 0.49479 1.79330 0 59.00000 women outerwear
91 DPT26 48356 0.39223 1.15360 1432 0 24.00000 unisex outerwear
92 DPT27 48356 0.49479 1.79330 0 59.00000 men bags
93 DPT28 48356 0.39223 1.15360 1432 0 24.00000 women bags
94 DPT29 48356 0.49479 1.79330 0 59.00000 unisex bags
95 DPT30 48356 0.39223 1.15360 1432 0 24.00000 men hats
96 DPT31 48356 0.49479 1.79330 0 59.00000 women hats
97 DPT32 48356 0.39223 1.15360 1432 0 24.00000 unisex hats
98 DPT33 48356 0.49479 1.79330 0 59.00000 men belts
99 DPT34 48356 0.39223 1.15360 1432 0 24.00000 women belts
100 DPT35 48356 0.49479 1.79330 0 59.00000 unisex belts
101 DPT36 48356 0.39223 1.15360 1432 0 24.00000 men ties
102 DPT37 48356 0.49479 1.79330 0 59.00000 women ties
103 DPT38 48356 0.39223 1.15360 1432 0 24.00000 unisex ties
104 DPT39 48356 0.49479 1.79330 0 59.00000 men socks
105 DPT40 48356 0.39223 1.15360 1432 0 24.00000 women socks
106 DPT41 48356 0.49479 1.79330 0 59.00000 unisex socks
107 DPT42 48356 0.39223 1.15360 1432 0 24.00000 men shirts
108 DPT43 48356 0.49479 1.79330 0 59.00000 women shirts
109 DPT44 48356 0.39223 1.15360 1432 0 24.00000 unisex shirts
110 DPT45 48356 0.49479 1.79330 0 59.00000 men jackets
111 DPT46 48356 0.39223 1.15360 1432 0 24.00000 women jackets
112 DPT47 48356 0.49479 1.79330 0 59.00000 unisex jackets
113 DPT48 48356 0.39223 1.15360 1432 0 24.00000 men pants
114 DPT49 48356 0.49479 1.79330 0 59.00000 women pants
115 DPT50 48356 0.39223 1.15360 1432 0 24.00000 unisex pants
116 DPT51 48356 0.49479 1.79330 0 59.00000 men shorts
117 DPT52 48356 0.39223 1.15360 1432 0 24.00000 women shorts
118 DPT53 48356 0.49479 1.79330 0 59.00000 unisex shorts
119 DPT54 48356 0.39223 1.15360 1432 0 24.00000 men shirts
120 DPT55 48356 0.49479 1.79330 0 59.00000 women shirts
121 DPT56 48356 0.39223 1.15360 1432 0 24.00000 unisex shirts
122 DPT57 48356 0.49479 1.79330 0 59.00000 men jackets
123 DPT58 48356 0.39223 1.15360 1432 0 24.00000 women jackets
124 DPT59 48356 0.49479 1.79330 0 59.00000 unisex jackets
125 DPT60 48356 0.39223 1.15360 1432 0 24.00000 men pants
126 DPT61 48356 0.49479 1.79330 0 59.00000 women pants
127 DPT62 48356 0.39223 1.15360 1432 0 24.00000 unisex pants
128 DPT63 48356 0.49479 1.79330 0 59.00000 men shorts
129 DPT64 48356 0.39223 1.15360 1432 0 24.00000 women shorts
130 DPT65 48356 0.49479 1.79330 0 59.00000 unisex shorts
131 DPT66 48356 0.39223 1.15360 1432 0 24.00000 men shirts
132 DPT67 48356 0.49479 1.79330 0 59.00000 women shirts
133 DPT68 48356 0.39223 1.15360 1432 0 24.00000 unisex shirts
134 DPT69 48356 0.49479 1.79330 0 59.00000 men jackets
135 DPT70 48356 0.39223 1.15360 1432 0 24.00000 women jackets
136 DPT71 48356 0.49479 1.79330 0 59.00000 unisex jackets
137 DPT72 48356 0.39223 1.15360 1432 0 24.00000 men pants
138 DPT73 48356 0.49479 1.79330 0 59.00000 women pants
139 DPT74 48356 0.39223 1.15360 1432 0 24.00000 unisex pants
140 DPT75 48356 0.49479 1.79330 0 59.00000 men shorts
141 DPT76 48356 0.39223 1.15360 1432 0 24.00000 women shorts
142 DPT77 48356 0.49479 1.79330 0 59.00000 unisex shorts
143 DPT78 48356 0.39223 1.15360 1432 0 24.00000 men shirts
144 DPT79 48356 0.49479 1.79330 0 59.00000 women shirts
145 DPT80 48356 0.39223 1.15360 1432 0 24.00000 unisex shirts
146 DPT81 48356 0.49479 1.79330 0 59.00000 men jackets
147 DPT82 48356 0.39223 1.15360 1432 0 24.00000 women jackets
148 DPT83 48356 0.49479 1.79330 0 59.00000 unisex jackets
149 DPT84 48356 0.39223 1.15360 1432 0 24.00000 men pants
150 DPT85 48356 0.49479 1.79330 0 59.00000 women pants
151 DPT86 48356 0.39223 1.15360 1432 0 24.00000 unisex pants
152 DPT87 48356 0.49479 1.79330 0 59.00000 men shorts
153 DPT88 48356 0.39223 1.15360 1432 0 24.00000 women shorts
154 DPT89 48356 0.49479 1.79330 0 59.00000 unisex shorts
155 DPT90 48356 0.39223 1.15360 1432 0 24.00000 men shirts
156 DPT91 48356 0.49479 1.79330 0 59.00000 women shirts
157 DPT92 48356 0.39223 1.15360 1432 0 24.00000 unisex shirts
158 DPT93 48356 0.49479 1.79330 0 59.00000 men jackets
159 DPT94 48356 0.39223 1.15360 1432 0 24.00000 women jackets
160 DPT95 48356 0.49479 1.79330 0 59.00000 unisex jackets
161 DPT96 48356 0.39223 1.15360 1432 0 24.00000 men pants
162 DPT97 48356 0.49479 1.79330 0 59.00000 women pants
163 DPT98 48356 0.39223 1.15360 1432 0 24.00000 unisex pants
164 DPT99 48356 0.49479 1.79330 0 59.00000 men shorts
165 DPT100 48356 0.39223 1.15360 1432 0 24.00000 women shorts
166 DPT101 48356 0.49479 1.79330 0 59.00000 unisex shorts
167 DPT102 48356 0.39223 1.15360 1432 0 24.00000 men shirts
168 DPT103 48356 0.49479 1.79330 0 59.00000 women shirts
169 DPT104 48356 0.39223 1.15360 1432 0 24.00000 unisex shirts
170 DPT105 48356 0.49479 1.79330 0 59.00000 men jackets
171 DPT106 48356 0.39223 1.15360 1432 0 24.00000 women jackets
172 DPT107 48356 0.49479 1.79330 0 59.00000 unisex jackets
173 DPT108 48356 0.39223 1.15360 1432 0 24.00000 men pants
174 DPT109 48356 0.49479 1.79330 0 59.00000 women pants
175 DPT110 48356 0.39223 1.15360 1432 0 24.00000 unisex pants
176 DPT111 48356 0.49479 1.79330 0 59.00000 men shorts
177 DPT112 48356 0.39223 1.15360 1432 0 24.00000 women shorts
178 DPT113 48356 0.49479 1.79330 0 59.00000 unisex shorts
179 DPT114 48356 0.39223 1.15360 1432 0 24.00000 men shirts
180 DPT115 48356 0.49479 1.79330 0 59.00000 women shirts
181 DPT116 48356 0.39223 1.15360 1432 0 24.00000 unisex shirts
182 DPT117 48356 0.49479 1.79330 0 59.00000 men jackets
183 DPT118 48356 0.39223 1.15360 1432 0 24.00000 women jackets
184 DPT119 48356 0.49479 1.79330 0 59.00000 unisex jackets
185 DPT120 48356 0.39223 1.15360 1432 0 24.00000 men pants
186 DPT121 48356 0.49479 1.79330 0 59.00000 women pants
187 DPT122 48356 0.39223 1.15360 1432 0 24.00000 unisex pants
188 DPT123 48356 0.49479 1.79330 0 59.00000 men shorts
189 DPT124 48356 0.39223 1.15360 1432 0 24.00000 women shorts
190 DPT125 48356 0.49479 1.79330 0 59.00000 unisex shorts
191 DPT126 48356 0.49479 1.79330 0 59.00000 men shirts
192 DPT127 48356 0.39223 1.15360 1432 0 24.00000 women shirts
193 DPT128 48356 0.49479 1.79330 0 59.00000 unisex shirts
194 DPT129 48356 0.49479 1.79330 0 59.00000 men jackets
195 DPT130 48356 0.39223 1.15360 1432 0 24.00000 women jackets
196 DPT131 48356 0.49479 1.79330 0 59.00000 unisex jackets
197 DPT132 48356 0.39223 1.15360 1432 0 24.00000 men pants
198 DPT133 48356 0.49479 1.79330 0 59.00000 women pants
199 DPT134 48356 0.39223 1.15360 1432 0 24.00000 unisex pants
200 DPT135 48356 0.49479 1.79330 0 59.00000 men shorts
201 DPT136 48356 0.39223 1.15360 1432 0 24.00000 women shorts
202 DPT137 48356 0.49479 1.79330 0 59.00000 unisex shorts
203 DPT138 48356 0.39223 1.15360 1432 0 24.00000 men shirts
204 DPT139 48356 0.39223 1.15360 1432 0 24.00000 women shirts
205 DPT140 48356 0.39223 1.15360 1432 0 24.00000 unisex shirts
206 DPT141 48356 0.49479 1.79330 0 59.00000 men jackets
207 DPT142 48356 0.39223 1.15360 1432 0 24.00000 women jackets
208 DPT143 48356 0.49479 1.79330 0 59.00000 unisex jackets
209 DPT144 48356 0.39223 1.15360 1432 0 24.00000 men pants
210 DPT145 48356 0.49479 1.79330 0 59.00000 women pants
211 DPT146 48356 0.39223 1.15360 1432 0 24.00000 unisex pants
212 DPT147 48356 0.49479 1.79330 0 59.00000 men shorts
213 DPT148 48356 0.39223 1.15360 1432 0 24.00000 women shorts
214 DPT149 48356 0.49479 1.79330 0 59.00000 unisex shorts
215 DPT150 48356 0.39223 1.15360 1432 0 24.00000 men shirts
216 DPT151 48356 0.39223 1.15360 1432 0 24.00000 women shirts
217 DPT152 48356 0.39223 1.15360 1432 0 24.00000 unisex shirts
218 DPT153 48356 0.49479 1.79330 0 59.00000 men jackets
219 DPT154 48356 0.39223 1.15360 1432 0 24.00000 women jackets
220 DPT155 48356 0.49479 1.79330 0 59.00000 unisex jackets
221 DPT156 48356 0.39223 1.15360 1432 0 24.00000 men pants
222 DPT157 48356 0.49479 1.79330 0 59.00000 women pants
223 DPT158 48356 0.39223 1.15360 1432 0 24.00000 unisex pants
224 DPT159 48356 0.49479 1.79330 0 59.00000 men shorts
225 DPT160 48356 0.39223 1.15360 1432 0 24.00000 women shorts
226 DPT161 48356 0.49479 1.79330 0 59.00000 unisex shorts
227 DPT162 48356 0.39223 1.15360 1432 0 24.00000 men shirts
228 DPT163 48356 0.39223 1.15360 1432 0 24.00000 women shirts
229 DPT164 48356 0.39223 1.15360 1432 0 24.00000 unisex shirts
230 DPT165 48356 0.49479 1.79330 0 59.00000 men jackets
231 DPT166 48356 0.39223 1.15360 1432 0 24.00000 women jackets
232 DPT167 48356 0.49479 1.79330 0 59.00000 unisex jackets
233 DPT168 48356 0.39223 1.15360 1432 0 24.00000 men pants
234 DPT169 48356 0.49479 1.79330 0 59.00000 women pants
235 DPT170 48356 0.39223 1.15360 1432 0 24.00000 unisex pants
236 DPT171 48356 0.49479 1.79330 0 59.00000 men shorts
237 DPT172 48356 0.39223 1.15360 1432 0 24.00000 women shorts
238 DPT173 48356 0.49479 1.79330 0 59.00000 unisex shorts
239 DPT174 48356 0.39223 1.15360 1432 0 24.00000 men shirts
240 DPT175 48356 0.39223 1.15360 1432 0 24.00000 women shirts
241 DPT176 48356 0.39223 1.15360 1432 0 24.00000 unisex shirts
242 DPT177 48356 0.49479 1.79330 0 59.00000 men jackets
243 DPT178 48356 0.39223 1.15360 1432 0 24.00000 women jackets
244 DPT179 48356 0.49479 1.79330 0 59.00000 unisex jackets
245 DPT180 48356 0.39223 1.15360 1432 0 24.00000 men pants
246 DPT181 48356 0.49479 1.79330 0 59.00000 women pants
247 DPT182 48356 0.39223 1.15360 1432 0 24.00000 unisex pants
248 DPT183 48356 0.49479 1.79330 0 59.00000 men shorts
249 DPT184 48356 0.39223 1.15360 1432 0 24.00000 women shorts
250 DPT185 48356 0.49479 1.79330 0 59.00000 unisex shorts
251 DPT186 48356 0.39223 1.15360 1432 0 24.00000 men shirts
252 DPT187 48356 0.39223 1.15360 1432 0 24.00000 women shirts
253 DPT188 48356 0.39223 1.15360 1432 0 24.00000 unisex shirts
254 DPT189 48356 0.49479 1.79330 0 59.00000 men jackets
255 DPT190 48356 0.39223 1.15360 1432 0 24.00000 women jackets
256 DPT191 48356 0.49479 1.79330 0 59.00000 unisex jackets
257 DPT192 48356 0.39223 1.15360 1432 0 24.00000 men pants
258 DPT193 48356 0.49479 1.79330 0 59.00000 women pants
259 DPT194 48356 0.39223 1.15360 1432 0 24.00000 unisex pants
260 DPT195 48356 0.49479 1.79330 0 59.00000 men shorts
261 DPT196 48356 0.39223 1.15360 1432 0 24.00000 women shorts
262 DPT197 48356 0.49479 1.79330 0 59.00000 unisex shorts
263 DPT198 48356 0.49479 1.79330 0 59.00000 men shirts
264 DPT199 48356 0.39223 1.15360 1432 0 24.00000 women shirts
265 DPT200 48356 0.39223 1.15360 1432 0 24.00000 unisex shirts
266 DPT201 48356 0.49479 1.79330 0 59.00000 men jackets
267 DPT202 48356 0.39223 1.15360 1432 0 24.00000 women jackets
268 DPT203 48356 0.49479 1.79330 0 59.00000 unisex jackets
269 DPT204 48356 0.39223 1.15360 1432 0 24.00000 men pants
270 DPT205 48356 0.49479 1.79330 0 59.00000 women pants
271 DPT206 48356 0.39223 1.15360 1432 0 24.00000 unisex pants
272 DPT207 48356 0.49479 1.79330 0 59.00000 men shorts
273 DPT208 48356 0.39223 1.15360 1432 0 24.00000 women shorts
274 DPT209 48356 0.49479 1.79330 0 59.00000 unisex shorts
275 DPT210 48356 0.49479 1.79330 0 59.00000 men shirts
276 DPT211 48356 0.39223 1.15360 1432 0 24.00000 women shirts
277 DPT212 48356 0.39223 1.15360 1432 0 24.00000 unisex shirts
278 DPT213 48356 0.49479 1.79330 0 59.00000 men jackets
279 DPT214 48356 0.39223 1.15360 1432 0 24.00000 women jackets
280 DPT215 48356 0.49479 1.79330 0 59.00000 unisex jackets
281 DPT216 48356 0.39223 1.15360 1432 0 24.00000 men pants
282 DPT217 48356 0.49479 1.79330 0 59.00000 women pants
283 DPT218 48356 0.39223 1.15360 1432 0 24.00000 unisex pants
284 DPT219 48356 0.49479 1.79330 0 59.00000 men shorts
285 DPT220 48356 0.39223 1.15360 1432 0 24.00000 women shorts
286 DPT221 48356 0.49479 1.79330 0 59.00000 unisex shorts
287 DPT222 48356 0.49479 1.79330 0 59.00000 men shirts
288 DPT223 48356 0.39223 1.15360 1432 0 24.00000 women shirts
289 DPT224 48356 0.39223 1.15360 1432 0 24.00000 unisex shirts
290 DPT225 48356 0.49479 1.79330 0 59.00000 men jackets
291 DPT226 48356 0.39223 1.15360 1432 0 24.00000 women jackets
292 DPT227 48356 0.49479 1.79330 0 59.00000 unisex jackets
293 DPT228 48356 0.39223 1.15360 1432 0 24.00000 men pants
294 DPT229 48356 0.49479 1.79330 0 59.00000 women pants
295 DPT230 48356 0.39223 1.15360 1432 0 24.00000 unisex pants
296 DPT231 48356 0.49479 1.79330 0 59.00000 men shorts
297 DPT232 48356 0.39223 1.15360 1432 0 24.00000 women shorts
```

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

The screenshot shows the SAS Enterprise Guide interface with the following components visible:

- Sample Properties pane:** Shows basic information about the data set, including Rows (48356), Columns (98), Member (CATALOG2010), and Type (DATA).
- Sample Statistics pane:** Displays a table of observations (Obs #) and variables (Variable, Label, Type, Percent, Minimum, Maximum, Mean, Number, Mode). The table includes columns for various demographic and financial variables like ZIP code, state code, county c., and dollar value.
- Select a Chart Type dialog:** A modal window titled "Select a Chart Type" is open, showing options like Scatter, Line, Histogram, Density, Box, Tables, Lattice, Parallel Axis, and Constellation. The "Box" option is selected.
- Description of a boxplot:** A text box within the dialog provides a brief explanation of what a boxplot is and how it represents the five-number summary.
- Data preview pane:** Shows a preview of the CATALOG2010 data set with columns like Obs #, custom..., num qr..., check ..., % quar..., number..., card only, days si..., avg, months..., dollar v..., check ..., respon..., state c..., county ..., ZIP code, and months.
- System tray:** Shows standard Windows icons for file operations, task switching, and system status.

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

- The output with boxplot is shown below.

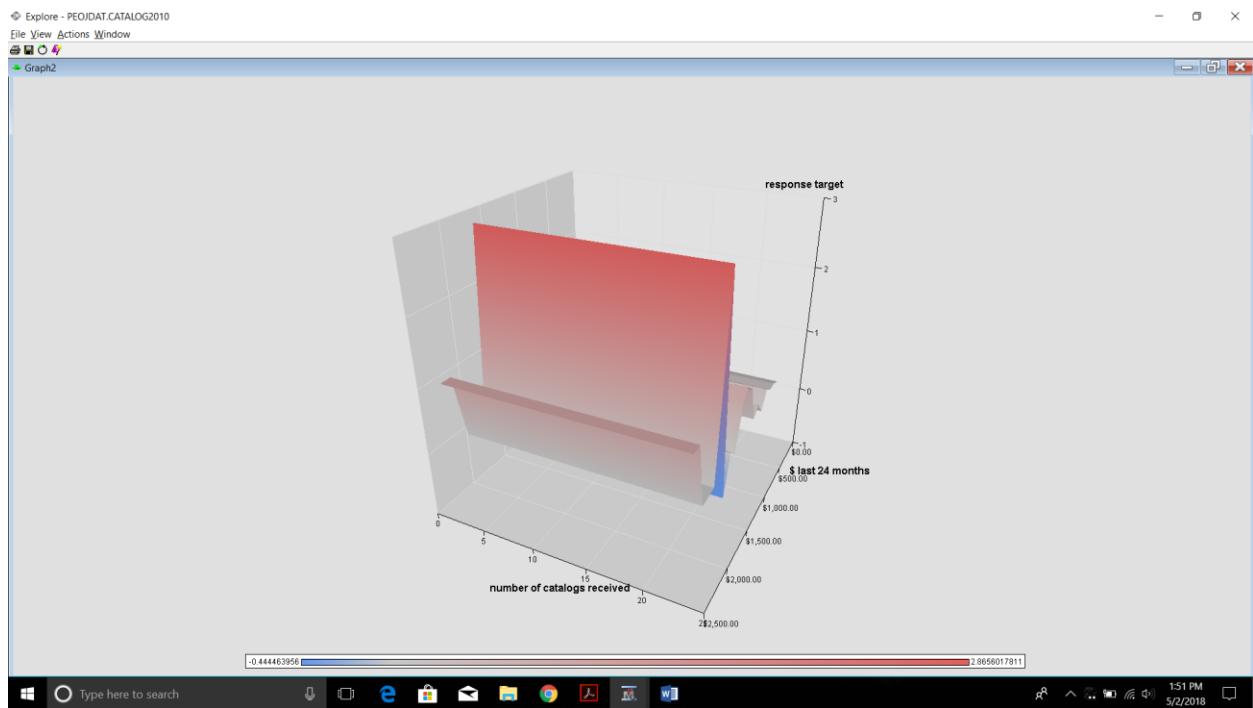
The screenshot shows the Enterprise Miner - HA interface. The left sidebar displays a tree structure with 'Data Sources' expanded, showing 'BANK' and 'CATALOG2010'. The 'CATALOG2010' node is selected. The main workspace contains a 'Sample' window titled 'Explore - PEOJDAT.CATALOG2010'. A box plot is displayed for the variable 'lifetime orders'. The y-axis ranges from 0 to 80. The box plot statistics are as follows:

Minimum Whisker =	1
First Quartile =	2
Median =	2
Third Quartile =	6
Maximum Whisker =	7
Mean =	4.15595

The 'Diagram' tab is selected at the bottom of the workspace.

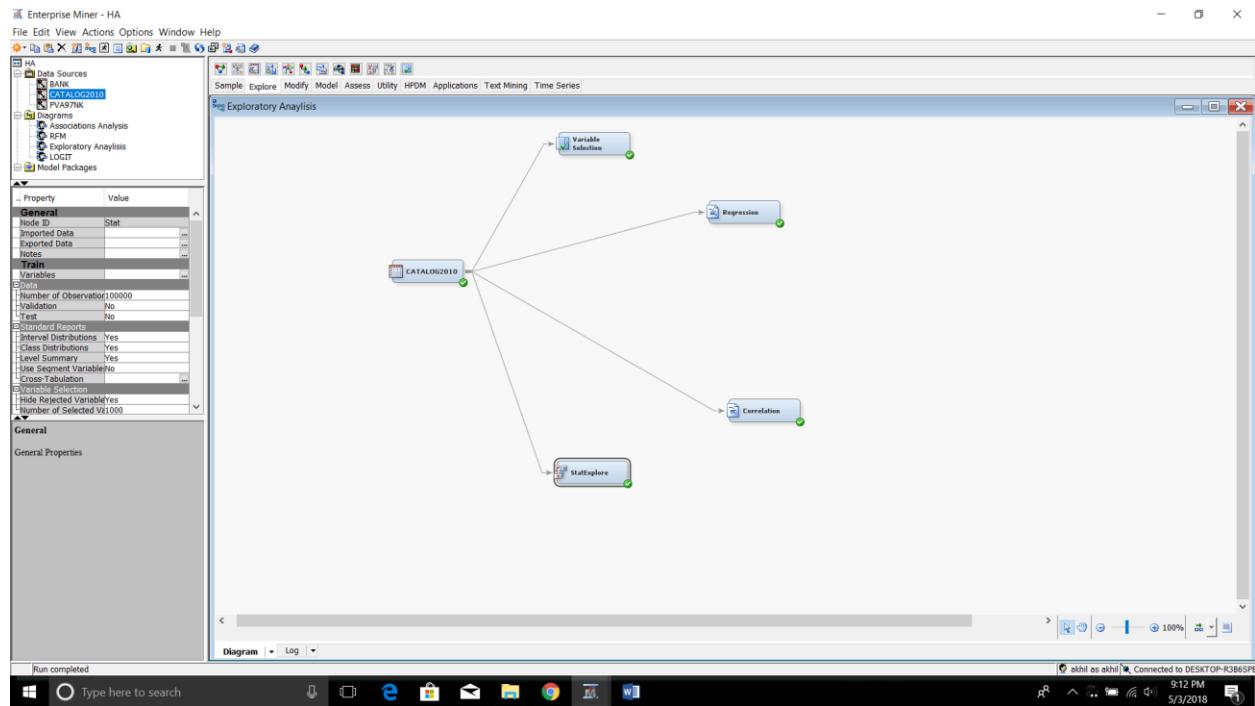
The above box plot analysis can be used to determine the outliers. From this we can say that for lifetime orders greater than the Maximum whisker value of 11 can be ignored as outliers.

3) 3D plot



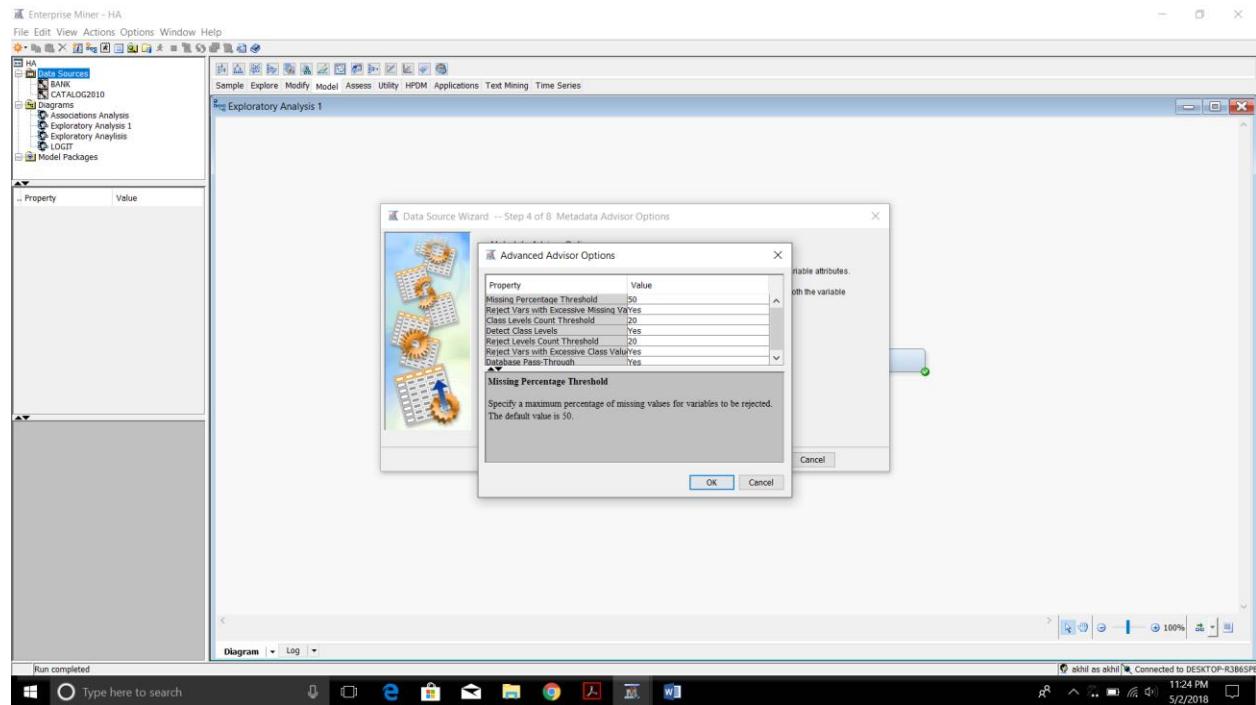
- As we hover around the 3-D plot we can see the values for different variables at each point.
- We can change the 3D plot based on desired variables.
- The red part of the 3D plot indicates the amount of money spent in the past 24 months is more and the blue part indicates that the amount of money spent 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.

Enterprise Miner - HA

File Edit View Actions Options Window Help

Data Sources: BANK, PVA97NK, PVA97C2010

Diagrams: Associations Analysis, Exploratory Analysis, LOGIT, Model Packages

Variables - PVA97NK

Filter: (none) not Equal to

Columns: Label

Mining

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Demographic	Input	Interval	No	No	No	-	-
DemGender	Input	Nominal	No	No	No	-	-
pvenk	Input	Binary	No	No	No	-	-
DemHomeOwnInput	Input	Binary	No	No	No	-	-
DemMedHomeInput	Input	Interval	No	No	No	-	-
Decisions	Input	Interval	No	No	No	-	-
DemPctVeteransInput	Input	Interval	No	No	No	-	-
Role	Raw	Interval	No	No	No	-	-
Income	Input	Interval	No	No	No	-	-
Library	PEO3DAT	Input	Interval	No	No	-	-
Table	PVA97NK	Input	Interval	No	No	-	-
Sample Data Set	Sample	Interval	No	No	No	-	-
Sample Type	DATA	Interval	No	No	No	-	-
Type	DATA	Interval	No	No	No	-	-
No. Obs	9686	Interval	No	No	No	-	-
No. Cards	28	Interval	No	No	No	-	-
No. Bytes	2049024	Interval	No	No	No	-	-
Segment	IN	Nominal	No	No	No	-	-
DemAge	IN	Interval	No	No	No	-	-
Create Date	5/3/18 12:16 AM	Interval	No	No	No	-	-
Modified By	akhil	Interval	No	No	No	-	-
Modified Date	5/3/18 12:16 AM	Interval	No	No	No	-	-
ID							
Data Source identifier. The metadata tables associated with the data source are stored in MDMS SAS library and use this identifier as prefix for naming these tables.							

Explore... Edit Using SAS Code OK Cancel

Diagram | Log | Run completed

Type here to search

ahil as akhil Connected to DESKTOP-R3B6SP8 9:18 AM 5/3/2018

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.

Enterprise Miner - HA

File Edit View Actions Options Window Help

Data Sources: BANK, PVA97NK, PVA97C2010

Diagrams: RFM

PVA97NK → Transform Variables

General Properties

Transform Variables Properties

- Default Methods
- Interval Inputs: None
- Interval Targets: None
- Class: None
- Class Targets: None
- Treat Missing as Level: No
- Example Properties
- Method: Random
- Size: Max
- Random Seed: 12345
- Number of Bins: 4
- Missing Values: Use in Search
- Groupping Method
- Column: 0..1
- Group Missing: No
- Number of Bins: Variables
- Add Minimum Value BYTES
- Offset Value: 1

General

General Properties

Diagram | Log | Run completed

Type here to search

ahil as akhil Connected to DESKTOP-R3B6SP8 9:18 AM 5/3/2018

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

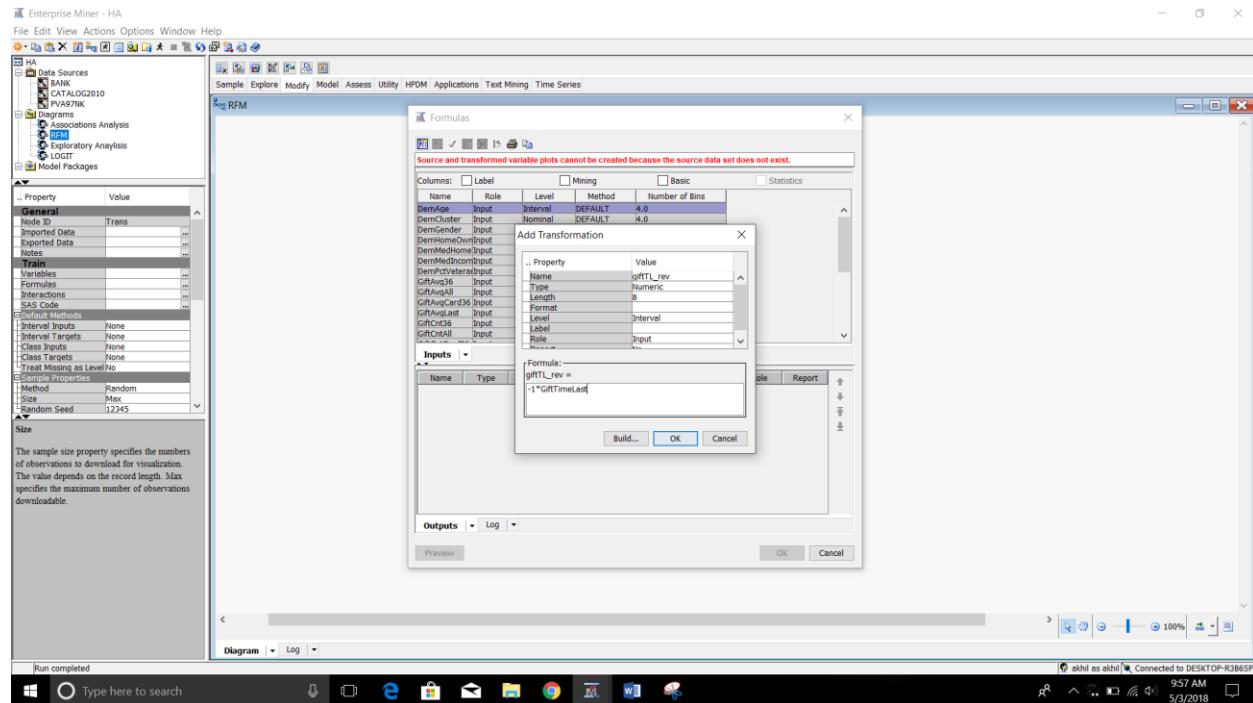
Initially

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.

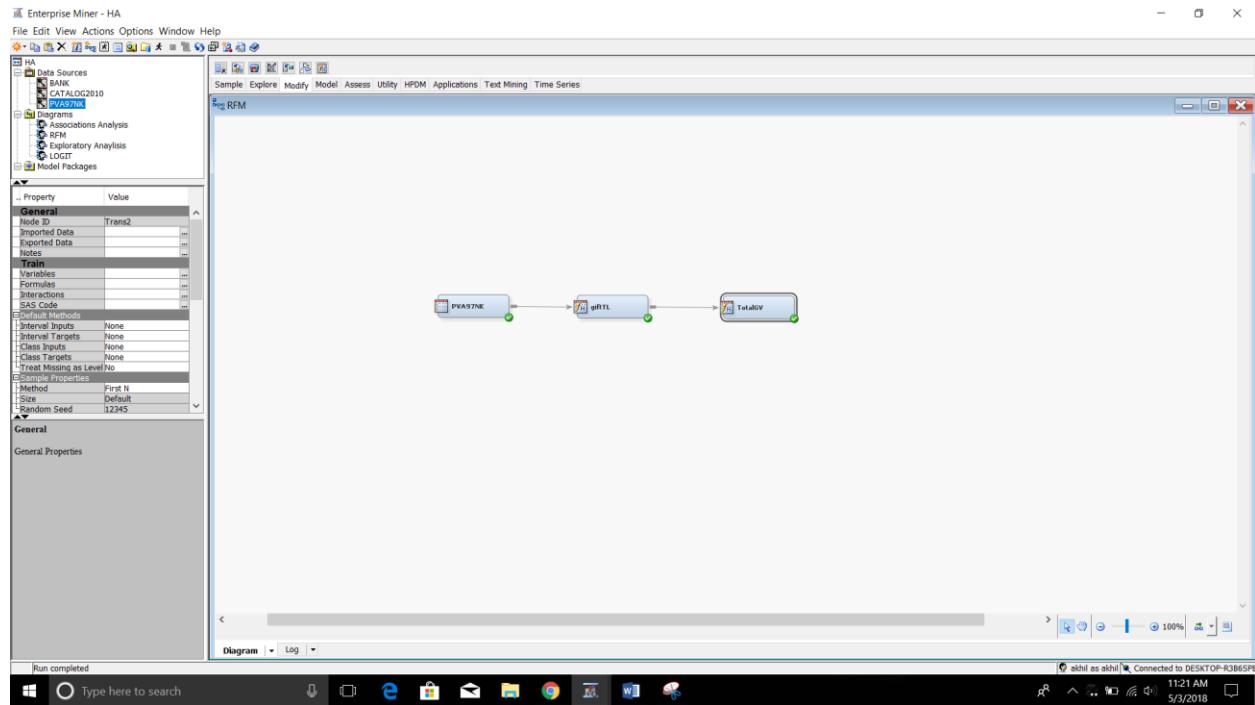


Setting up the formula is done by selecting the options box right next to the formulas option available in the properties block. We define a variable called **giftTL** and assign the Gift Time Last variable to this variable as shown below:

$giftTL = -1 * \text{GiftTimeLast}$

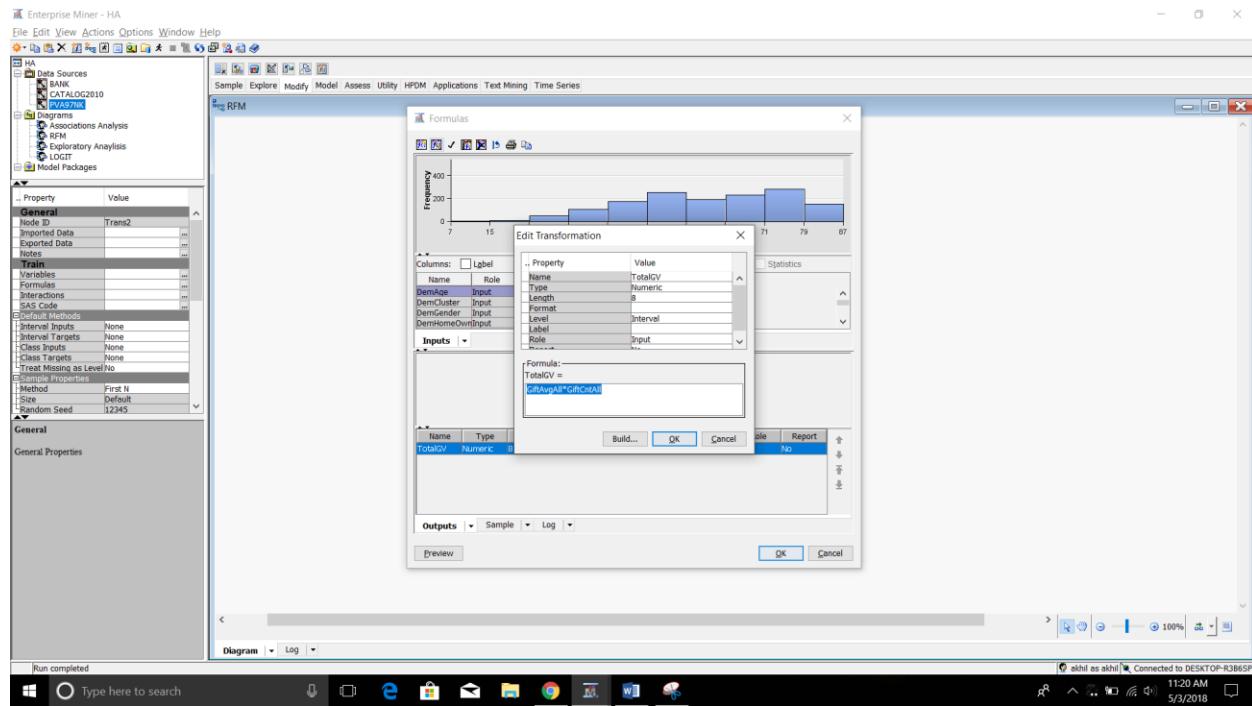
This process of multiplying -1 to the variable is done to reverse code the variable.

- 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:

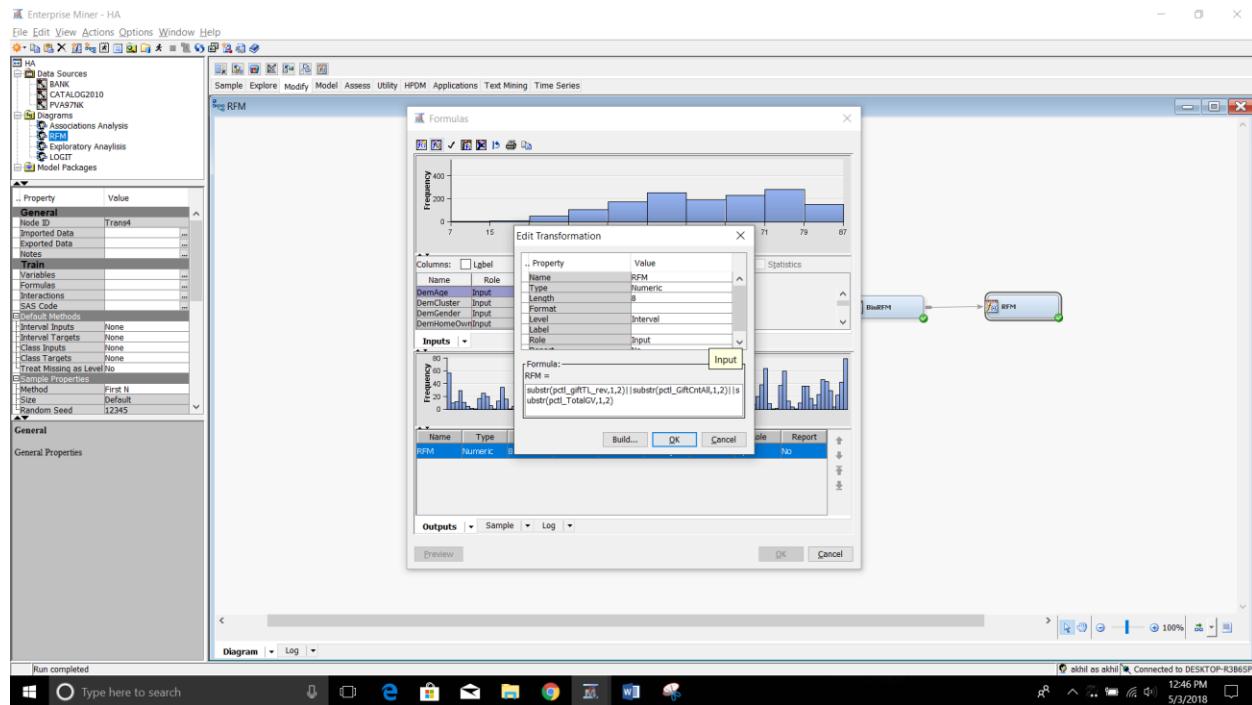
$$\text{TotalGV} = \text{GiftAvgAll} * \text{GiftCntAll}$$



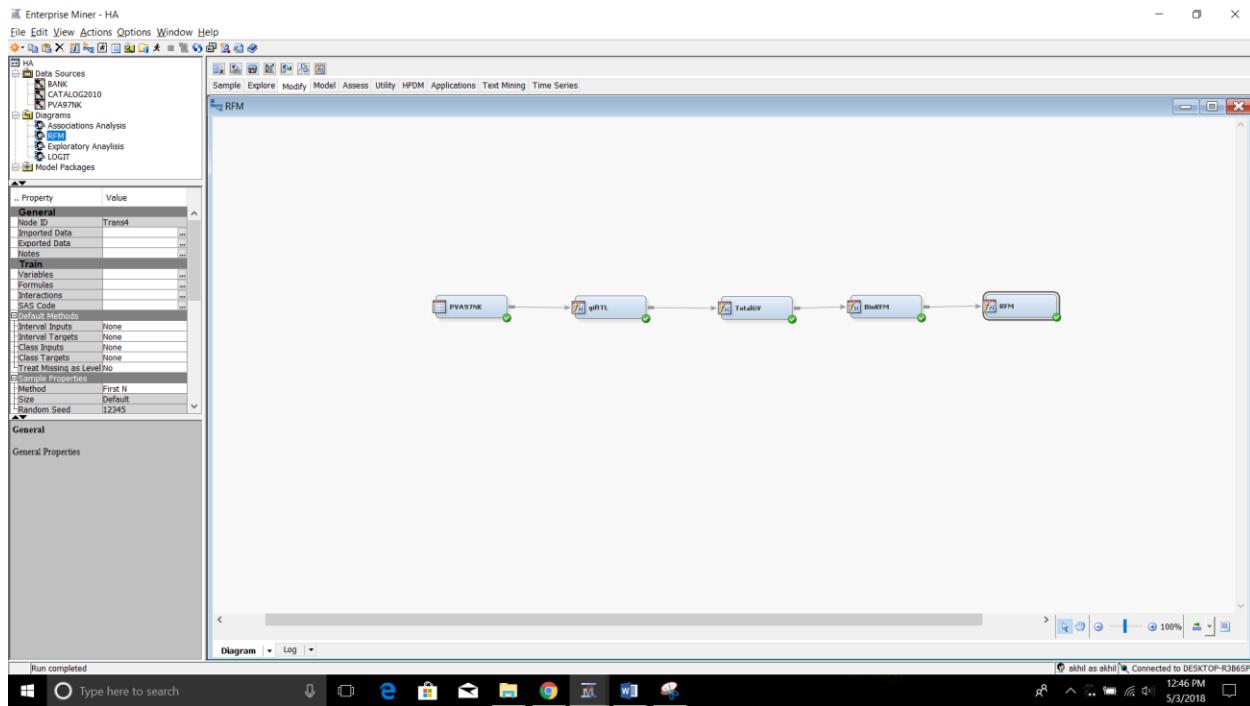
- Now we add the RFM block. The RFM block is formulated as shown below:

$\text{RFM} = \text{substr}(\text{pctl_giftTL_rev}, 1, 2) || \text{substr}(\text{pctl_GiftCntAll}, 1, 2) || \text{substr}(\text{pctl_TotalGV}, 1, 2)$

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

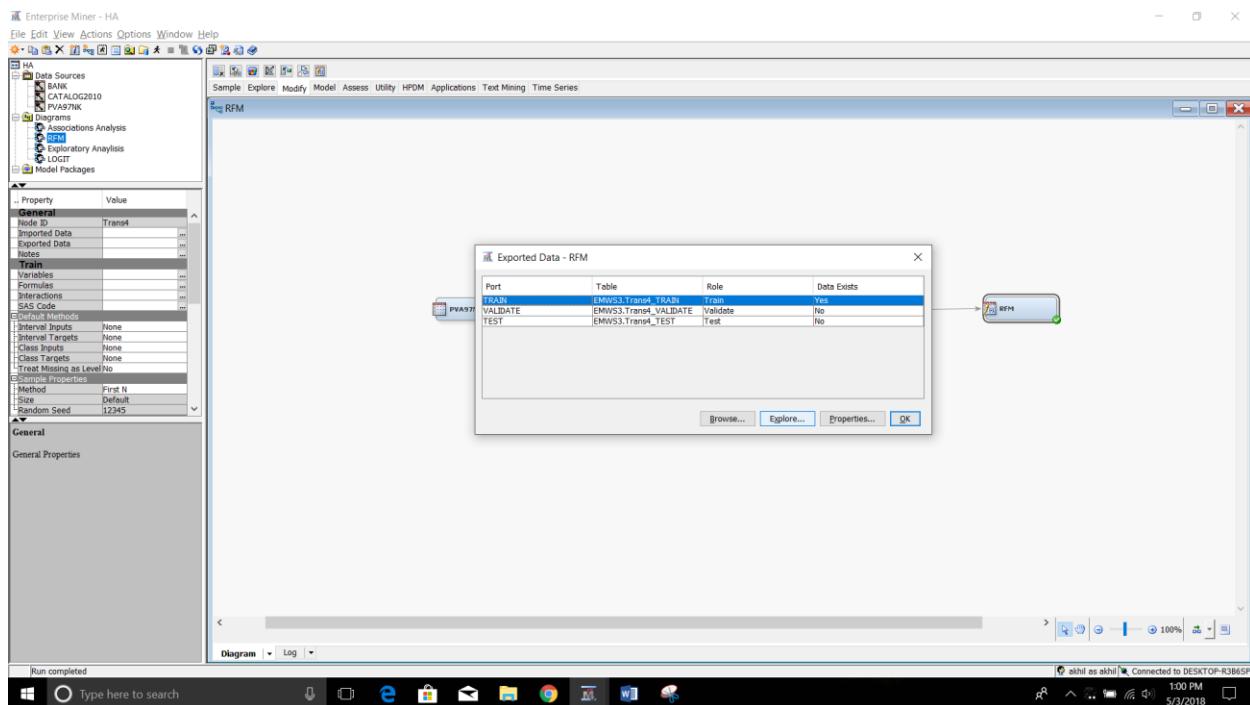


- The final block diagram for the RFM analysis is as shown below:



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.

Explore - EMWS3.Trans4_TRAIN

File View Actions Window

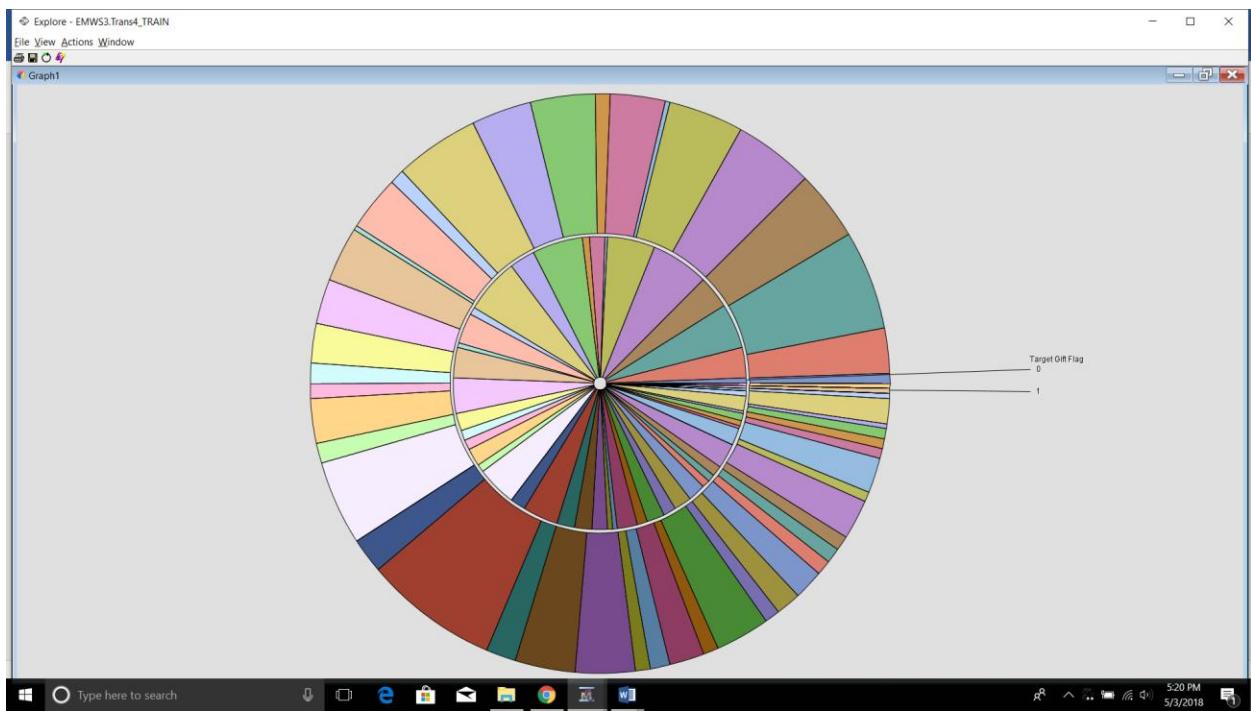
Sample Properties

Property	Value
Rows	Unknown
Columns	34
Library	EMWS3
Member	TRANS4_TRAIN
Type	VIEW
Sample Method	Random
Fetch Size	Max
Fetched Rows	9686
Random Seed	17345

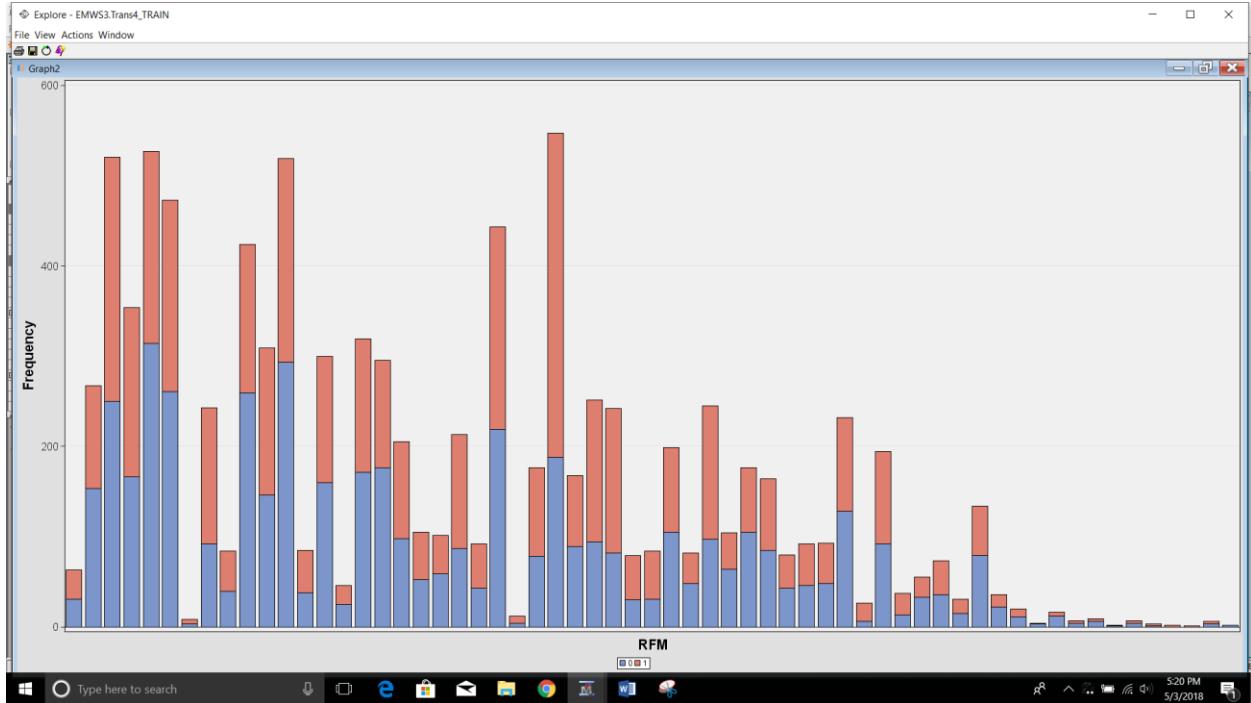
Sample Statistics

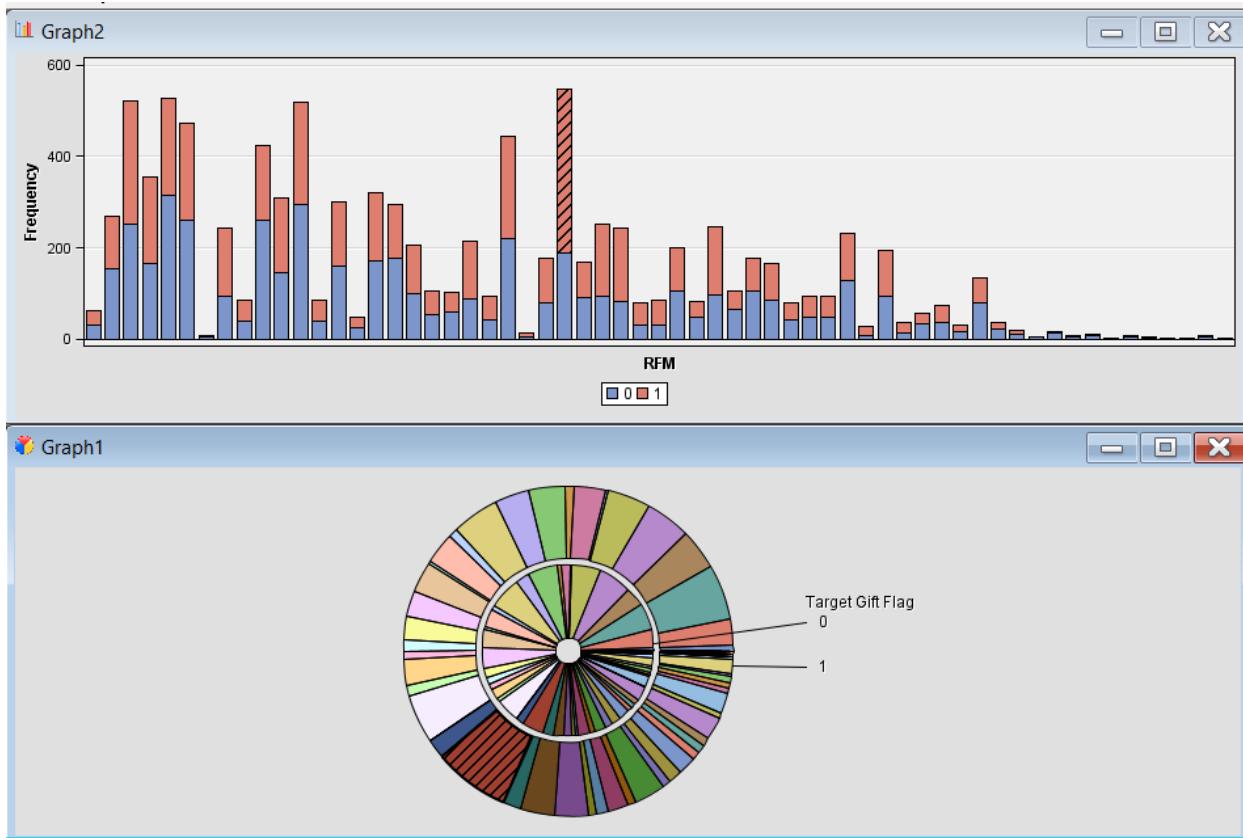
Obs #	Variable	Label	Type	Percent	Minimum	Maximum	Mean	Number	Mode	Mode
1	DemClus	Demogra.	CLASS	0				54	4.46004540	
2	DemGen	Gender	CLASS	0				3	53.923198	
3	DemHo	Home O.	CLASS	0				2	55.310411	
4	ID	Conn ID	CLASS	0				128+	0.751940001799	
5	SPCTL GL	Transfer	CLASS	0				5	22.15569516-HL	
6	SPCTL T	Transfer	CLASS	0				5	20.4891702365	
7	SPCTL GL	Transfer	CLASS	0				5	20.4791702365	
8	BRC	Transfer	CLASS	0				116	4.887177555605	
9	StatusC	Status C.	CLASS	0				6	60.14867A	
10	DemAge	Age	VAR	24.8503	0	87	69.16984			
11	DemMed	Median H.	VAR	0	800000	1198000				
12	DemMedI	Median In.	VAR	0	200000	30.60427				
13	DemMedII	Median In.	VAR	0	0	80	30.60427			
14	DemMedIII	Median In.	VAR	0	0	260	14.8762			
15	DemMedIV	Median In.	VAR	0	0	15	45.49302			
16	DemMedV	Median In.	VAR	0	0	260	14.22443			
17	DemMedVI	Median In.	VAR	0	0	45	16.10774			
18	DemMedVII	Median In.	VAR	0	0	16	3.205451			
19	DemMedVIII	Median In.	VAR	0	0	11	91.16984			
20	DemMedIX	Median In.	VAR	0	0	9	1.856597			
21	DemMedX	Median In.	VAR	0	0	41	5.52849			
22	DemMedXI	Median In.	VAR	0	0	15	269.71.10035			
23	DemMedXII	Median In.	VAR	0	0	4	27.19.10035			
24	DemMedXIII	Median In.	VAR	0	0	2	59.12.98885			
25	DemMedXIV	Median In.	VAR	0	0	4	78.29.34923			
26	DemMedXV	Median In.	VAR	0	0	174	4R.4R.4R			
27	DemMedXVI	Median In.	VAR	0	0	0	0			
28	DemMedXVII	Median In.	VAR	0	0	0	0			
29	DemMedXVIII	Median In.	VAR	0	0	0	0			
30	DemMedXIX	Median In.	VAR	0	0	0	0			
31	DemMedXX	Median In.	VAR	0	0	0	0			
32	DemMedXXI	Median In.	VAR	0	0	0	0			
33	DemMedXXII	Median In.	VAR	0	0	0	0			
34	DemMedXXIII	Median In.	VAR	0	0	0	0			
35	DemMedXXIV	Median In.	VAR	0	0	0	0			
36	DemMedXXV	Median In.	VAR	0	0	0	0			
37	DemMedXXVI	Median In.	VAR	0	0	0	0			
38	DemMedXXVII	Median In.	VAR	0	0	0	0			
39	DemMedXXVIII	Median In.	VAR	0	0	0	0			
40	DemMedXXIX	Median In.	VAR	0	0	0	0			
41	DemMedXXX	Median In.	VAR	0	0	0	0			
42	DemMedXXXI	Median In.	VAR	0	0	0	0			
43	DemMedXXXII	Median In.	VAR	0	0	0	0			
44	DemMedXXXIII	Median In.	VAR	0	0	0	0			
45	DemMedXXXIV	Median In.	VAR	0	0	0	0			
46	DemMedXXXV	Median In.	VAR	0	0	0	0			
47	DemMedXXXVI	Median In.	VAR	0	0	0	0			
48	DemMedXXXVII	Median In.	VAR	0	0	0	0			
49	DemMedXXXVIII	Median In.	VAR	0	0	0	0			
50	DemMedXXXIX	Median In.	VAR	0	0	0	0			
51	DemMedXXXI	Median In.	VAR	0	0	0	0			
52	DemMedXXXII	Median In.	VAR	0	0	0	0			
53	DemMedXXXIII	Median In.	VAR	0	0	0	0			
54	DemMedXXXIV	Median In.	VAR	0	0	0	0			
55	DemMedXXXV	Median In.	VAR	0	0	0	0			
56	DemMedXXXVI	Median In.	VAR	0	0	0	0			
57	DemMedXXXVII	Median In.	VAR	0	0	0	0			
58	DemMedXXXVIII	Median In.	VAR	0	0	0	0			
59	DemMedXXXIX	Median In.	VAR	0	0	0	0			
60	DemMedXXXI	Median In.	VAR	0	0	0	0			
61	DemMedXXXII	Median In.	VAR	0	0	0	0			
62	DemMedXXXIII	Median In.	VAR	0	0	0	0			
63	DemMedXXXIV	Median In.	VAR	0	0	0	0			
64	DemMedXXXV	Median In.	VAR	0	0	0	0			
65	DemMedXXXVI	Median In.	VAR	0	0	0	0			
66	DemMedXXXVII	Median In.	VAR	0	0	0	0			
67	DemMedXXXVIII	Median In.	VAR	0	0	0	0			
68	DemMedXXXIX	Median In.	VAR	0	0	0	0			
69	DemMedXXXI	Median In.	VAR	0	0	0	0			
70	DemMedXXXII	Median In.	VAR	0	0	0	0			
71	DemMedXXXIII	Median In.	VAR	0	0	0	0			
72	DemMedXXXIV	Median In.	VAR	0	0	0	0			
73	DemMedXXXV	Median In.	VAR	0	0	0	0			
74	DemMedXXXVI	Median In.	VAR	0	0	0	0			
75	DemMedXXXVII	Median In.	VAR	0	0	0	0			
76	DemMedXXXVIII	Median In.	VAR	0	0	0	0			
77	DemMedXXXIX	Median In.	VAR	0	0	0	0			
78	DemMedXXXI	Median In.	VAR	0	0	0	0			
79	DemMedXXXII	Median In.	VAR	0	0	0	0			
80	DemMedXXXIII	Median In.	VAR	0	0	0	0			
81	DemMedXXXIV	Median In.	VAR	0	0	0	0			
82	DemMedXXXV	Median In.	VAR	0	0	0	0			
83	DemMedXXXVI	Median In.	VAR	0	0	0	0			
84	DemMedXXXVII	Median In.	VAR	0	0	0	0			
85	DemMedXXXVIII	Median In.	VAR	0	0	0	0			
86	DemMedXXXIX	Median In.	VAR	0	0	0	0			
87	DemMedXXXI	Median In.	VAR	0	0	0	0			
88	DemMedXXXII	Median In.	VAR	0	0	0	0			
89	DemMedXXXIII	Median In.	VAR	0	0	0	0			
90	DemMedXXXIV	Median In.	VAR	0	0	0	0			
91	DemMedXXXV	Median In.	VAR	0	0	0	0			
92	DemMedXXXVI	Median In.	VAR	0	0	0	0			
93	DemMedXXXVII	Median In.	VAR	0	0	0	0			
94	DemMedXXXVIII	Median In.	VAR	0	0	0	0			
95	DemMedXXXIX	Median In.	VAR	0	0	0	0			
96	DemMedXXXI	Median In.	VAR	0	0	0	0			
97	DemMedXXXII	Median In.	VAR	0	0	0	0			
98	DemMedXXXIII	Median In.	VAR	0	0	0	0			
99	DemMedXXXIV	Median In.	VAR	0	0	0	0			
100	DemMedXXXV	Median In.	VAR	0	0	0	0			
101	DemMedXXXVI	Median In.	VAR	0	0	0	0			
102	DemMedXXXVII	Median In.	VAR	0	0	0	0			
103	DemMedXXXVIII	Median In.	VAR	0	0	0	0			
104	DemMedXXXIX	Median In.	VAR	0	0	0	0			
105	DemMedXXXI	Median In.	VAR	0	0	0	0			
106	DemMedXXXII	Median In.	VAR	0	0	0	0			
107	DemMedXXXIII	Median In.	VAR	0	0	0	0			
108	DemMedXXXIV	Median In.	VAR	0	0	0	0			
109	DemMedXXXV	Median In.	VAR	0	0	0	0			
110	DemMedXXXVI	Median In.	VAR	0	0	0	0			
111	DemMedXXXVII	Median In.	VAR	0	0	0	0			
112	DemMedXXXVIII	Median In.	VAR	0	0	0	0			
113	DemMedXXXIX	Median In.	VAR	0	0	0	0			
114	DemMedXXXI	Median In.	VAR	0	0	0	0			
115	DemMedXXXII	Median In.	VAR	0	0	0	0			
116	DemMedXXXIII	Median In.	VAR	0	0	0	0			
117	DemMedXXXIV	Median In.	VAR	0	0	0	0			
118	DemMedXXXV	Median In.	VAR	0	0	0	0			
119	DemMedXXXVI	Median In.	VAR	0	0	0	0			
120	DemMedXXXVII	Median In.	VAR	0	0	0	0			
121	DemMedXXXVIII	Median In.	VAR	0	0	0	0			
122	DemMedXXXIX	Median In.	VAR	0	0	0	0			
123	DemMedXXXI	Median In.	VAR	0	0	0	0			
124	DemMedXXXII	Median In.	VAR	0	0	0	0			
125	DemMedXXXIII	Median In.	VAR	0	0	0	0			
126	DemMedXXXIV	Median In.	VAR	0	0	0	0			
127	DemMedXXXV	Median In.	VAR	0	0	0	0			
128	DemMedXXXVI	Median In.	VAR	0	0	0	0			
129	DemMedXXXVII	Median In.	VAR	0	0	0	0			
130	DemMedXXXVIII	Median In.	VAR	0	0	0	0			
131	DemMedXXXIX	Median In.	VAR	0	0	0	0			
132	DemMedXXXI	Median In.	VAR	0	0	0	0			
133	DemMedXXXII	Median In.	VAR	0	0	0	0			
134	DemMedXXXIII	Median In.	VAR	0	0	0	0			
135	DemMedXXXIV	Median In.	VAR	0	0	0	0			
136	DemMedXXXV	Median In.	VAR	0	0	0	0			
137	DemMedXXXVI	Median In.	VAR	0	0	0	0			
138	DemMedXXXVII	Median In.	VAR	0	0	0	0			
139	DemMedXXXVIII	Median In.	VAR	0	0	0	0			
140	DemMedXXXIX	Median In.	VAR	0	0	0	0			
141	DemMedXXXI	Median In.	VAR	0	0	0	0			
142	DemMedXXXII	Median In.	VAR	0	0	0	0			
143	DemMedXXXIII	Median In.	VAR	0	0	0	0			
144	DemMedXXXIV	Median In.	VAR	0	0	0	0			
145	DemMedXXXV	Median In.	VAR	0	0	0	0			
146	DemMedXXXVI	Median In.	VAR	0	0	0	0			
147	DemMedXXXVII	Median In.	VAR	0	0	0	0			
148	DemMedXXXVIII	Median In.	VAR	0	0	0	0			
149	DemMedXXXIX	Median In.	VAR	0	0	0	0			
150	DemMedXXXI	Median In.	VAR	0	0	0	0			
151	DemMedXXXII	Median In.	VAR	0	0	0	0			
152	DemMedXXXIII	Median In.	VAR	0	0	0	0			
153	DemMedXXXIV	Median In.	VAR	0	0	0	0			
154	DemMedXXXV	Median In.	VAR	0	0	0	0			
155	DemMedXXXVI	Median In.	VAR	0	0	0	0			
156	DemMedXXXVII	Median In.	VAR	0	0	0	0			
157	DemMedXXXVIII	Median In.	VAR	0	0	0	0			
158	DemMedXXXIX	Median In.	VAR	0	0	0	0			
159	DemMedXXXI	Median In.	VAR	0	0	0	0			
160	DemMedXXXII	Median In.	VAR	0	0	0	0			
161	DemMedXXXIII	Median In.	VAR	0	0	0	0			
162	DemMedXXXIV	Median In.	VAR	0	0	0	0			
163	DemMedXXXV	Median In.	VAR	0	0	0	0			
164	DemMedXXXVI	Median In.	VAR	0	0	0	0			
165	DemMedXXXVII	Median In.	VAR	0	0	0	0			
166	DemMedXXXVIII	Median In.	VAR	0	0	0	0			
167	DemMedXXXIX	Median In.	VAR	0	0	0	0			
168	DemMedXXXI	Median In.	VAR	0	0	0	0			
169	DemMedXXXII	Median In.	VAR	0	0	0	0			
170	DemMedXXXIII	Median In.	VAR	0	0	0	0			
171	DemMedXXXIV	Median In.	VAR	0						

- 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.

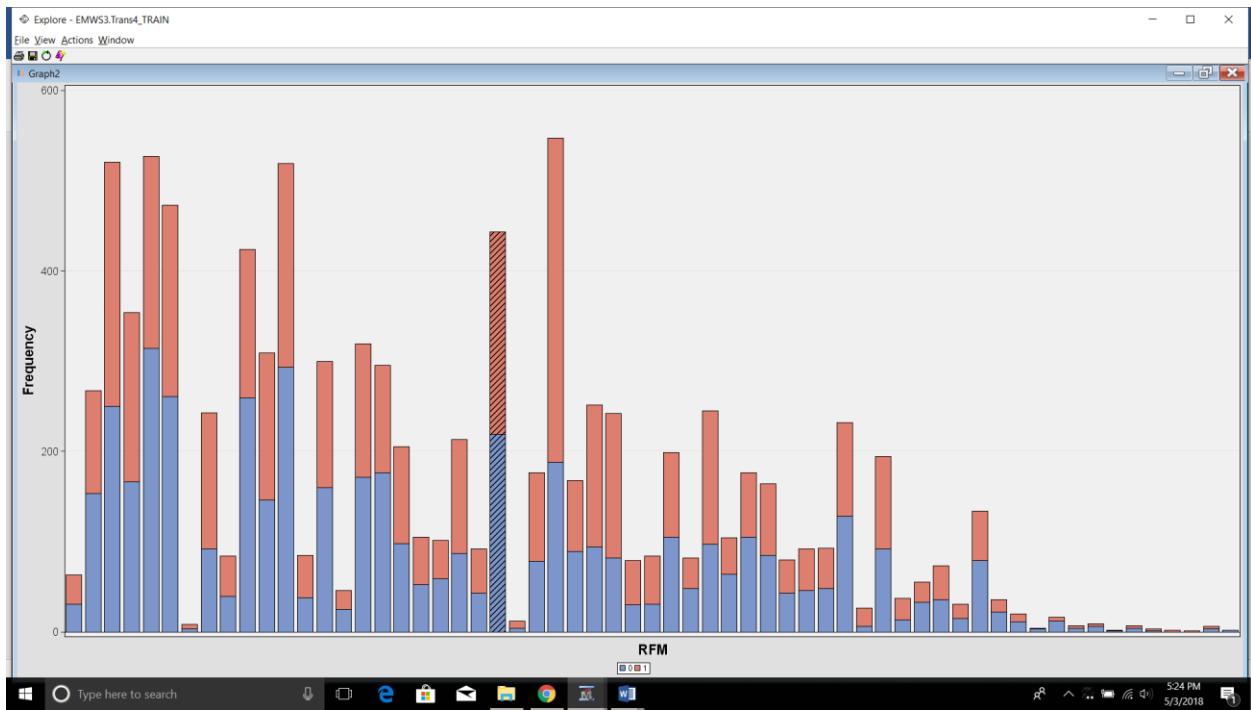
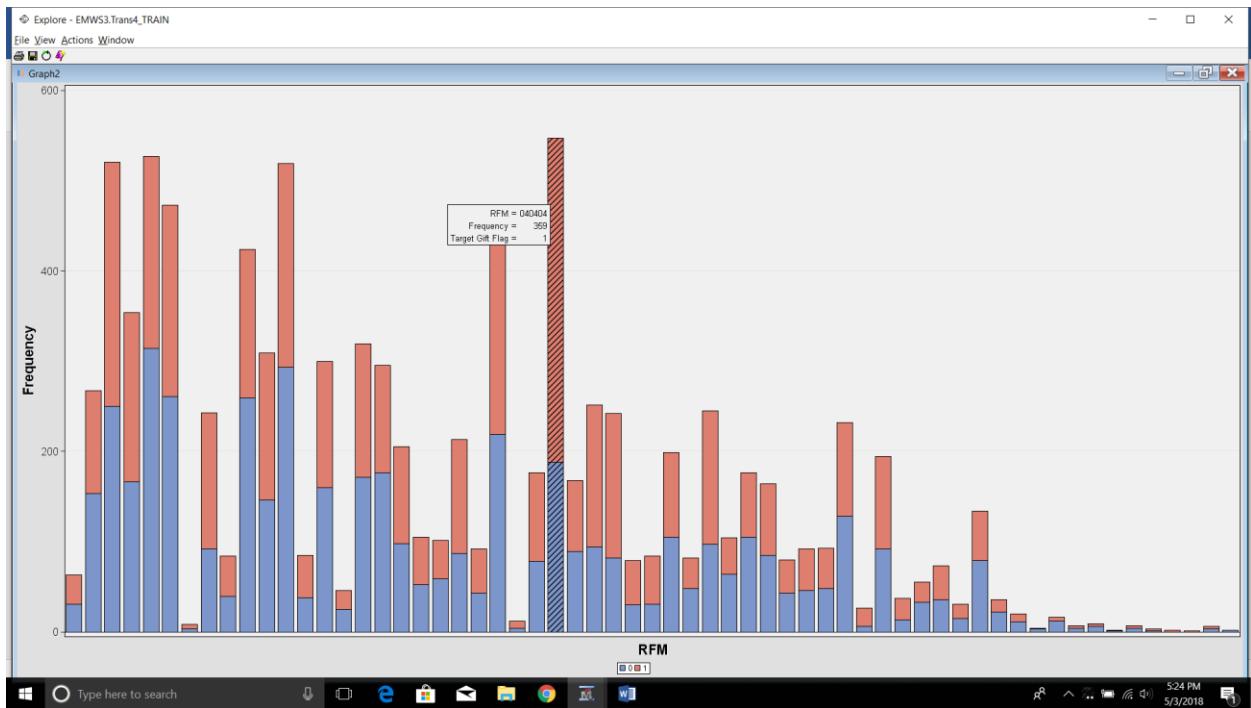




- When we select a part of the bar graph, the corresponding part pertaining to those values will also get highlighted in the pie chart.

- **Analysis:**
- The bar graph infers 2 things:
 - 0 – implies the amount of people who did not receive any gift. It is marked as blue.
 - 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:

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

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:

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

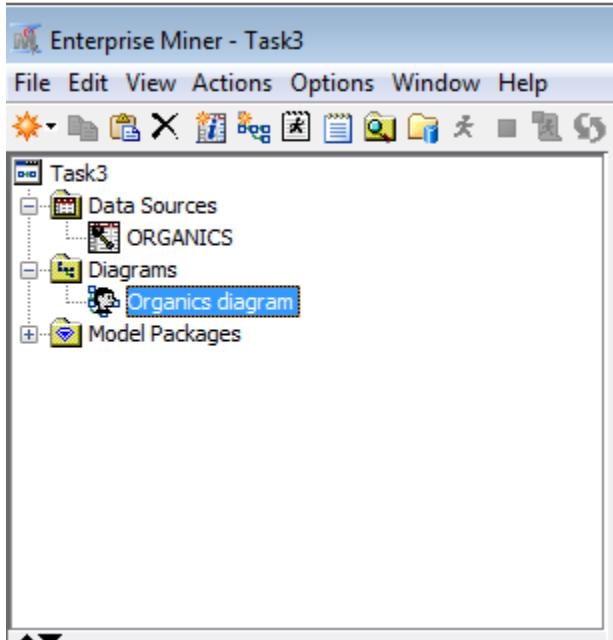
Thus, we can consider this bin 040301.

Part-2: Predictive Modelling, Decision Tree

Task 3 - Data Sources: ORGANICS

1)

- a) Create a new diagram named organics.



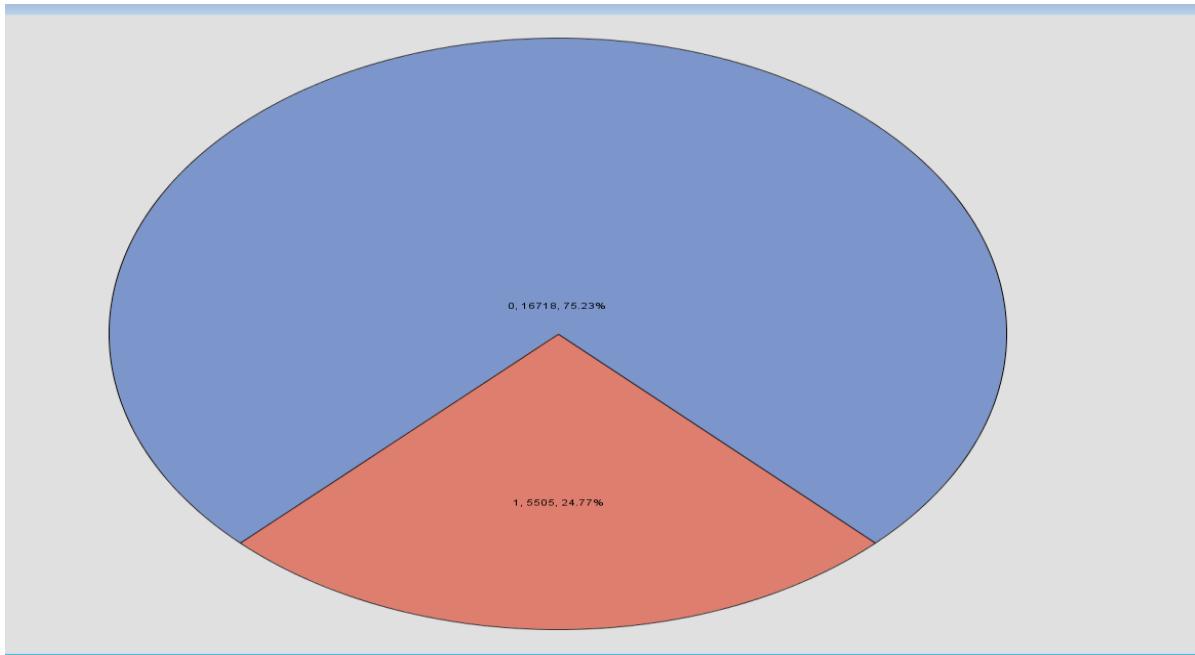
- b) Define the data set ORGANICS as a data source for the project .

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

The screenshot shows the "Variables - Ids" dialog box. At the top, there is a search bar with dropdown menus for "(none)", operators like "not" and "Equal to", and a value input field. Below the search bar are buttons for "Label" and "Mining". The main area is a table with columns: Name, Role, Level, Report, Order, Drop, Lower Limit, and Upper Limit. The table lists 15 variables with their respective roles and levels:

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 distribution of the target variable.what is the proportion 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.

Screenshot of the Orange data mining software interface, specifically the 'Variables - ORGANICS' tab. The table lists the following variables:

Name	Role	Level	Report	Order	Mining	Drop	Lower Limit	Upper Limit
DemAff	Input	Interval	No		No	No	-	-
DemAge	Input	Interval	No		No	No	-	-
DemCluster	Rejected	Nominal	No		No	No	-	-
DemClusterGroup	Input	Nominal	No		No	No	-	-
DemGender	Input	Nominal	No		No	No	-	-
DemReg	Input	Nominal	No		No	No	-	-
DemTriReg	Input	Nominal	No		No	No	-	-
ID	3D	Nominal	No		No	No	-	-
PromClass	Input	Nominal	No		No	No	-	-
PromSpend	Input	Interval	No		No	No	-	-
PromTime	Input	Interval	No		No	No	-	-
TargetAmt	Rejected	Interval	No		No	No	-	-
TargetBuy	Target	Binary	No		No	No	-	-

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.

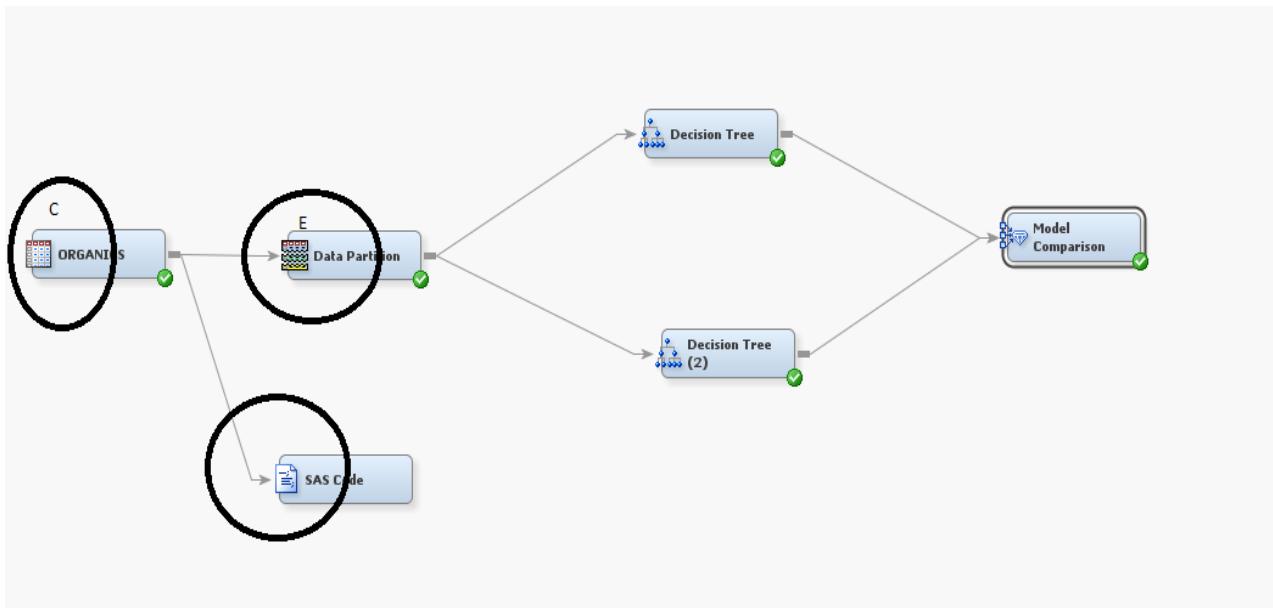
5) Data source definition is implemented from beginning.

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-

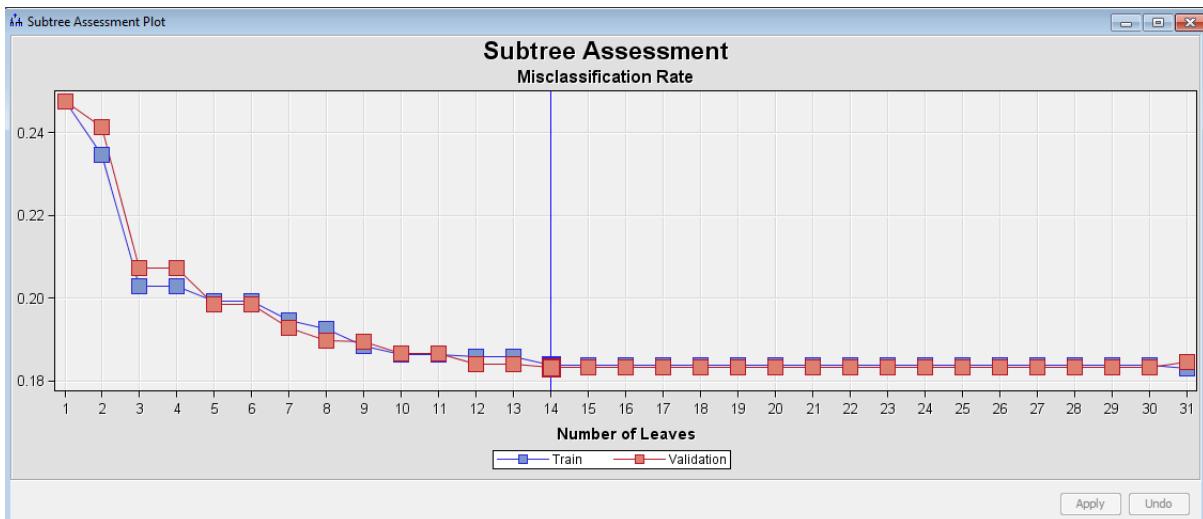
e. Add a decision tree node to the workspace and connect it to the data partition node.



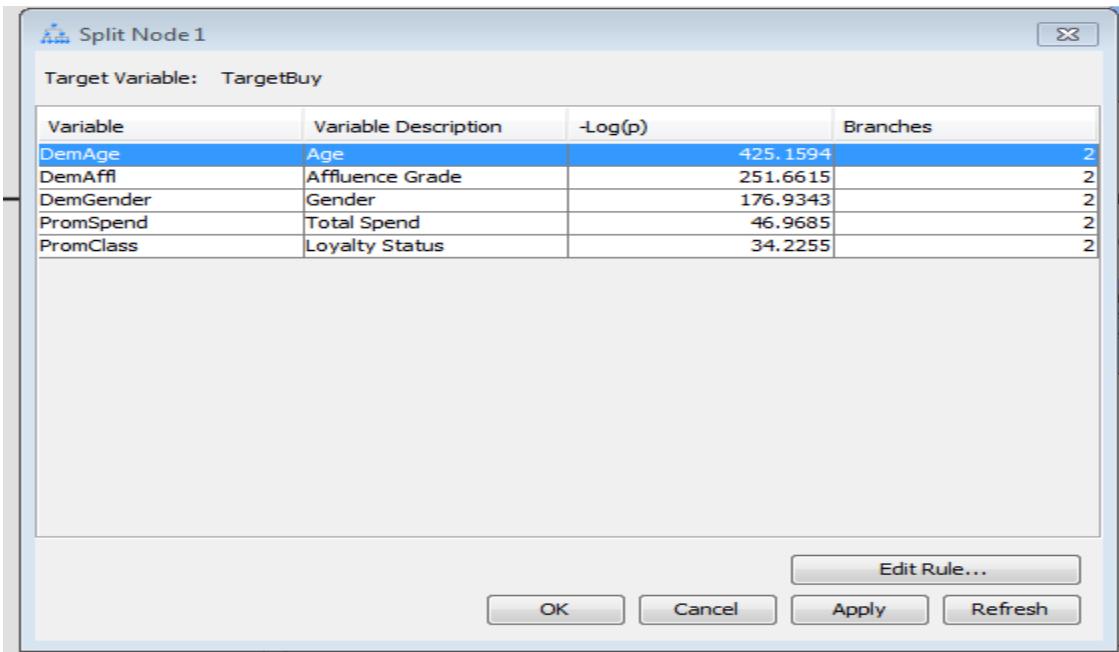
f. Create a decision tree model autonomously. Use Misclassification as the model assessment statistic.

1. How many leaves are in the optimal tree?

14 leaves.



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



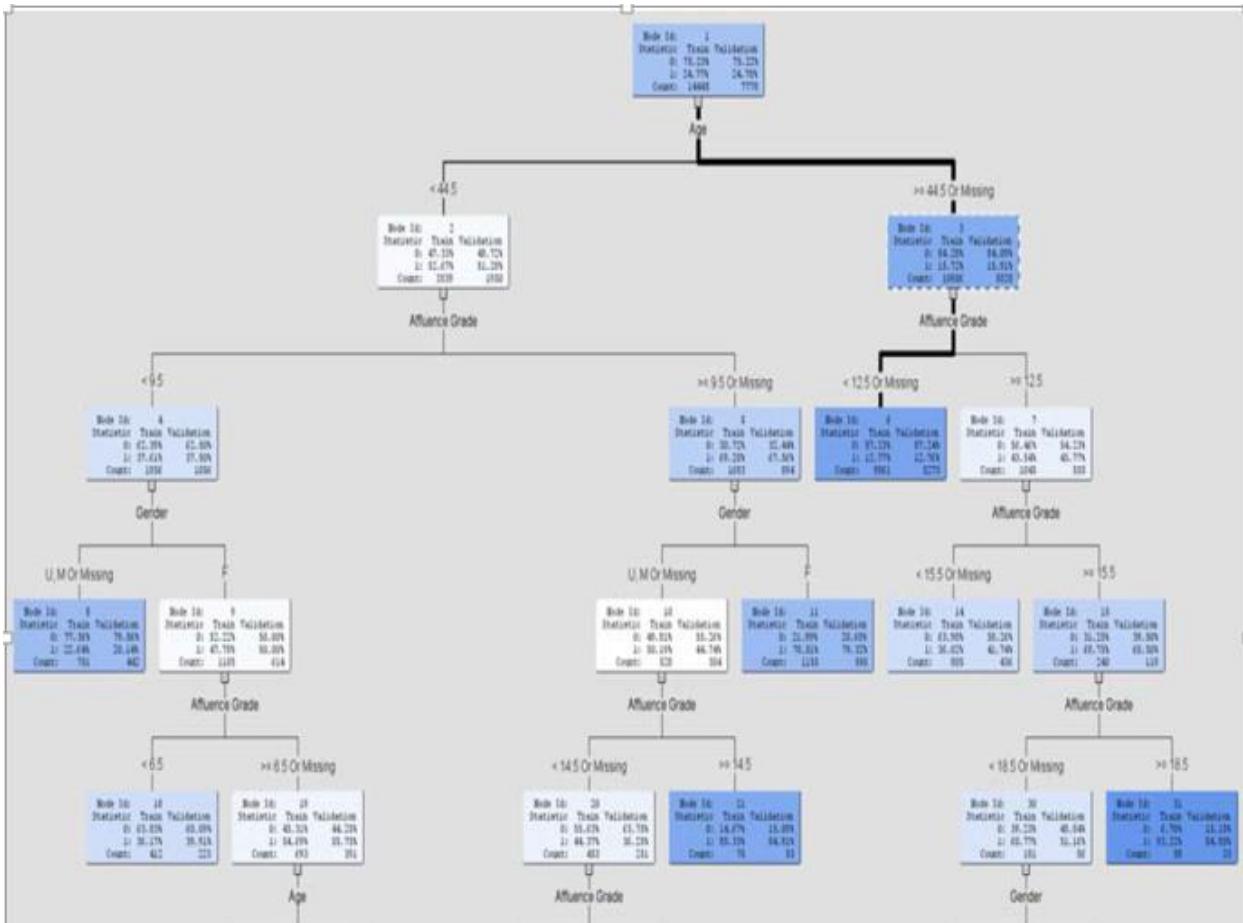
Age variable was used for the split. Age < 44.5 and Age \geq 44.5 or Missing.

Competing Splits: Affluence Grade < 12.5 or Missing and Affluence Grade \geq 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?

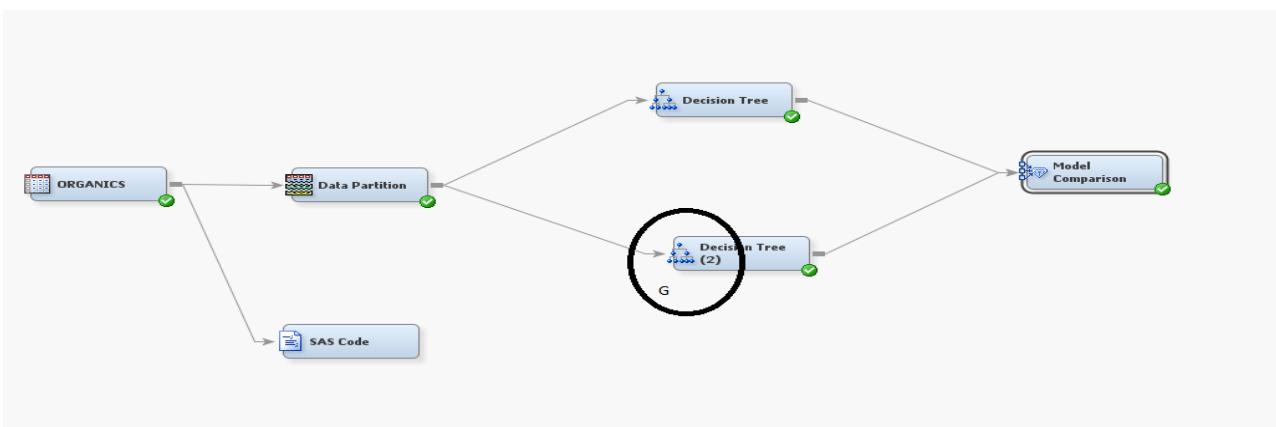
The variable **AFFLUENCE GRADE** is used for second split across all branches.



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.



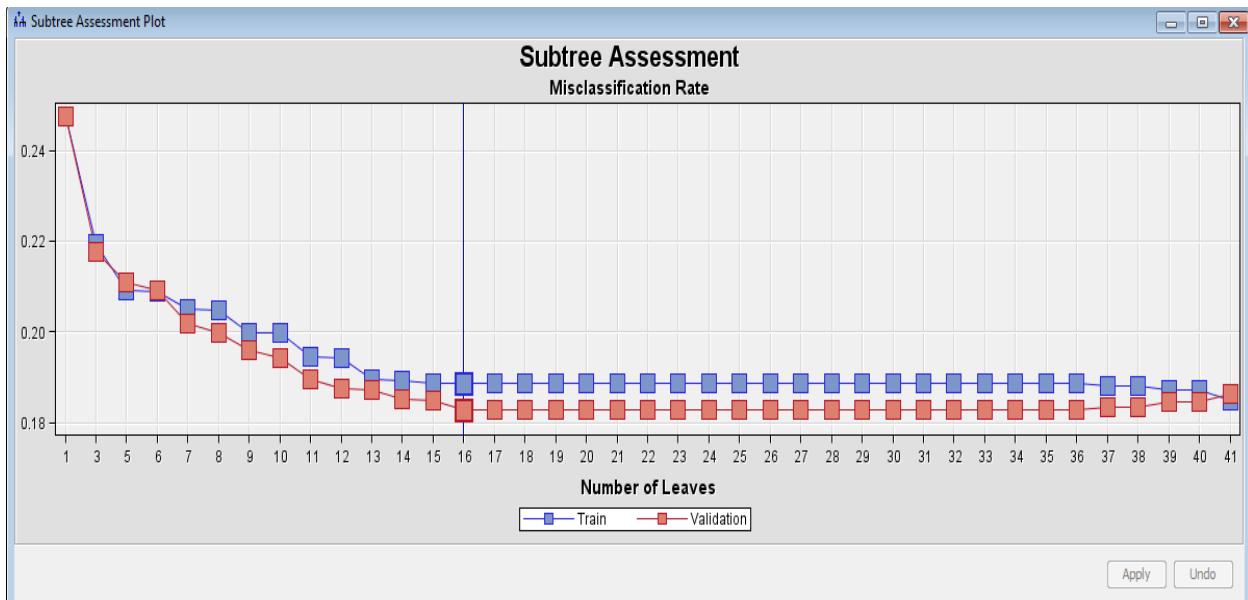
1. In the properties panel of the new decision tree node, change the maximum number of branches from a node to 3 to allow for three-way splits.

.. Property	Value
Use Frozen Tree	No
Use Multiple Targets	No
<input checked="" type="checkbox"/> 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
<input checked="" type="checkbox"/> Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
<input checked="" type="checkbox"/> Split Search	
Use Decisions	No
Use Priors	No

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

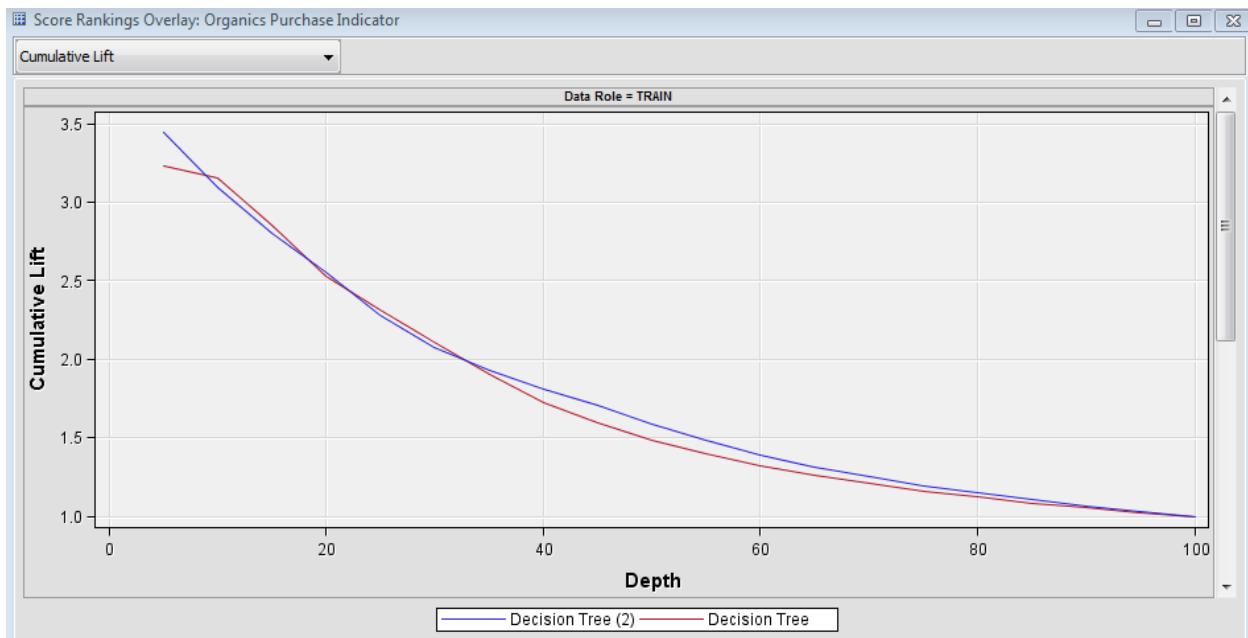
.. Property	Value
Split Size	.
<input checked="" type="checkbox"/> Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
<input checked="" type="checkbox"/> Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Misclassification
Assessment Fraction	0.25
<input checked=""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

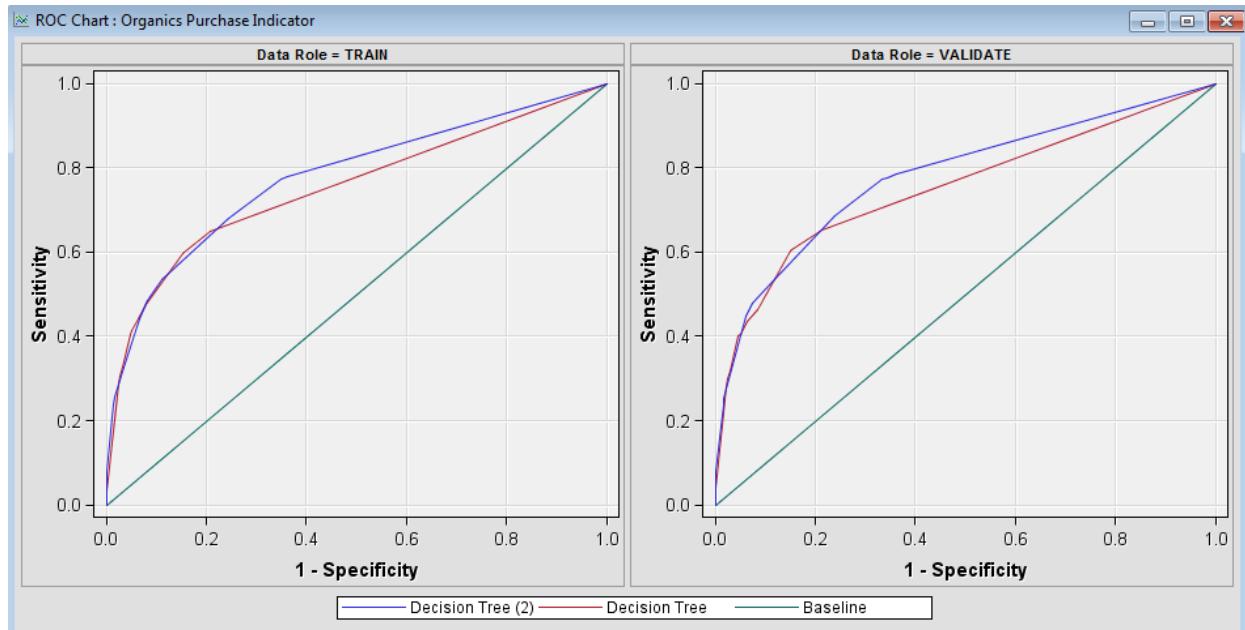
3. How many leaves are in the optimal tree?



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

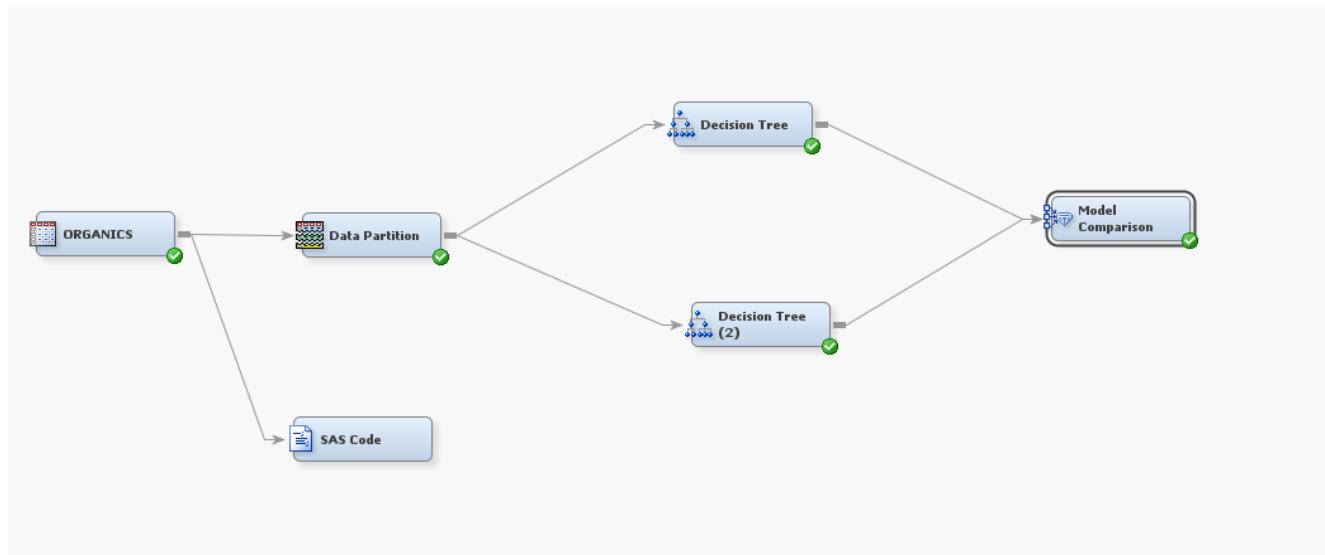
h. Based on Misclassification rate, which of the decision tree models appears to be better?





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.

➤ Final Model



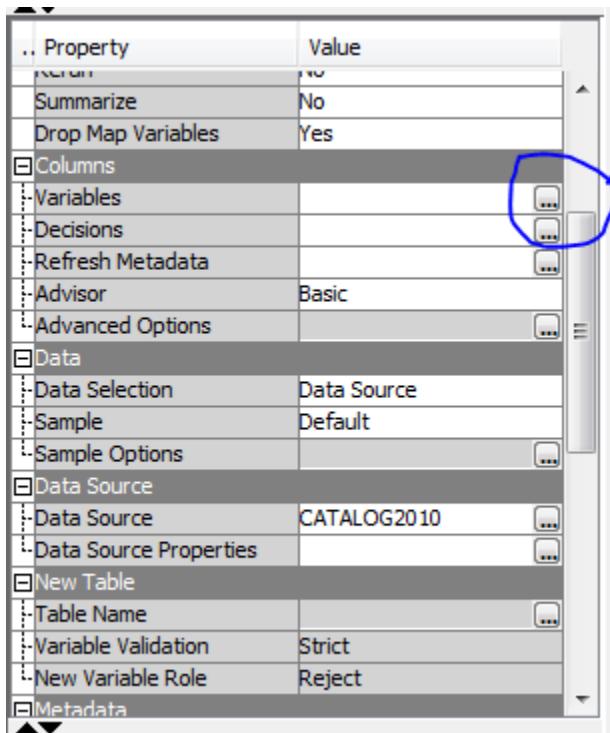
PART 3: Predictive Modelling using LOGIT (Logistic Regression)

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.

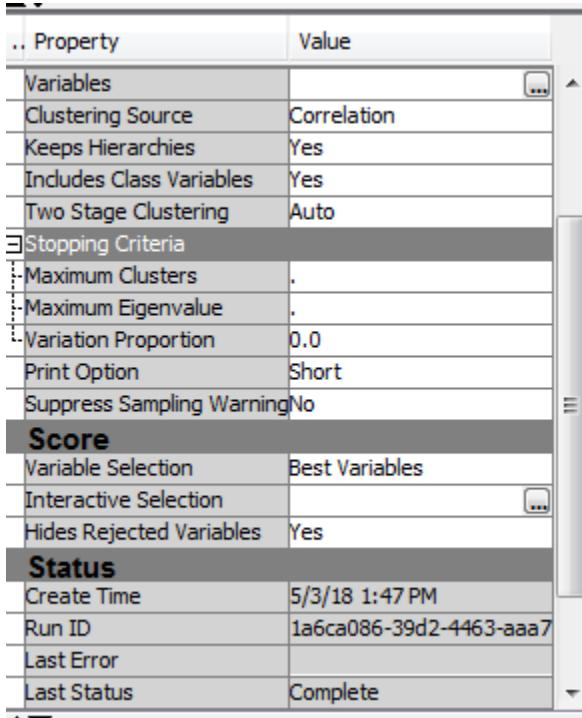
1. Review the CATALOG2010 data set. From the panel on left, click the ellipsis next to Variables.



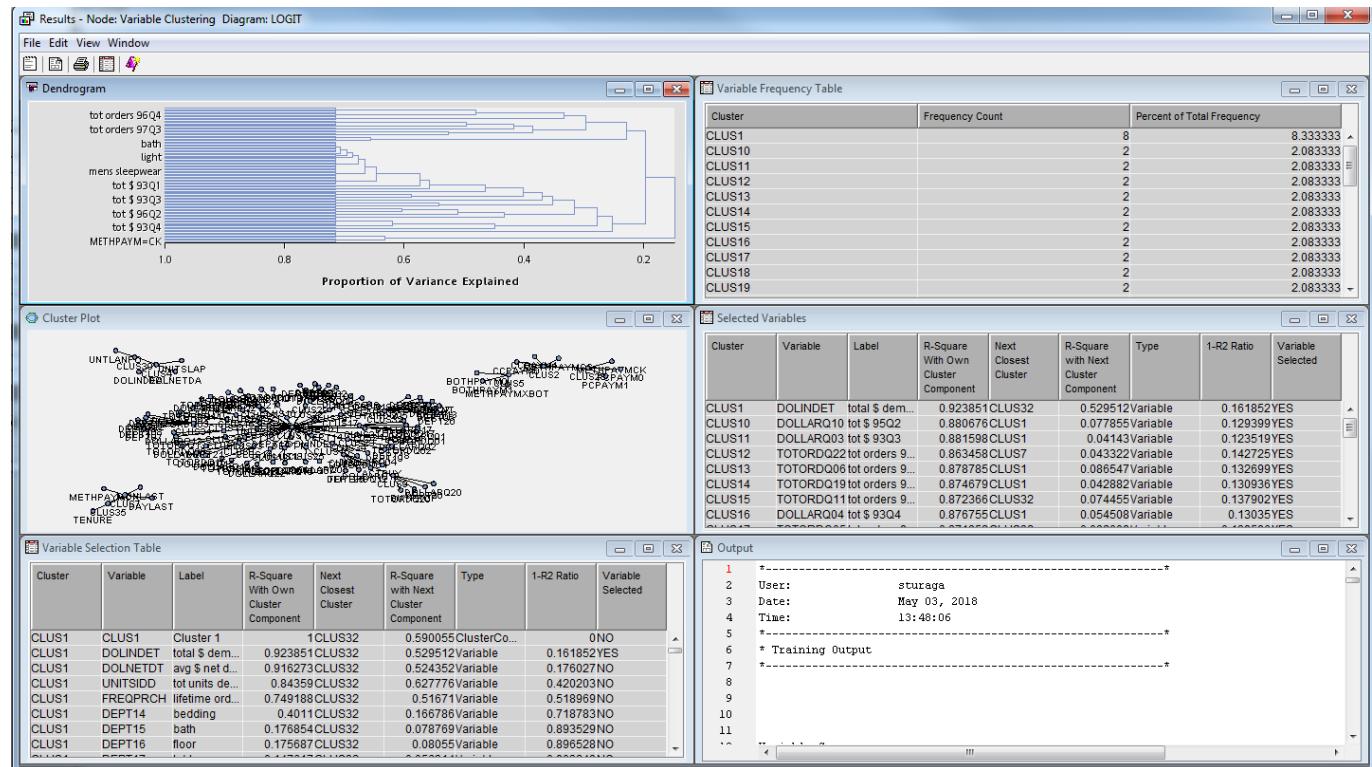
2. Compute some basic statistics.

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit	Number of Levels	Percent Missing	Minimum	Maximum	Mean	Standard Deviation	Skev
ACTBUN	Input	Interval	No	No	.	.	.	11	0
BOTTAWM	Input	Binary	No	No	.	.	.	2	0
BUYPROP	Input	Interval	No	No	.	.	.	0	0	1	0.18858	0.256955	.	
CATALOGNT	Input	Interval	No	No	.	.	.	0	1	27	3.76582	3.133252	1	
CCPAIM	Input	Binary	No	No	.	.	.	2	0
COUNTY	Rejected	Interval	No	No	.	.	.	0	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	7
DEPT01	Input	Interval	No	No	.	.	.	0	0	59	0.494789	1.793298	1	5
DEPT02	Input	Interval	No	No	.	.	.	0	0	24	0.292249	1.151603	.	.
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	1.759188	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	0.839967	2.172124	4	.
DEPT15	Input	Interval	No	No	.	.	.	0	0	53	0.262819	1.204034	7	.
DEPT16	Input	Interval	No	No	.	.	.	0	0	25	0.226921	1.029329	6	.
DEPT17	Input	Interval	No	No	.	.	.	0	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	7976.98	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.



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.

Results - Node: Variable Clustering Diagram: LOGIT

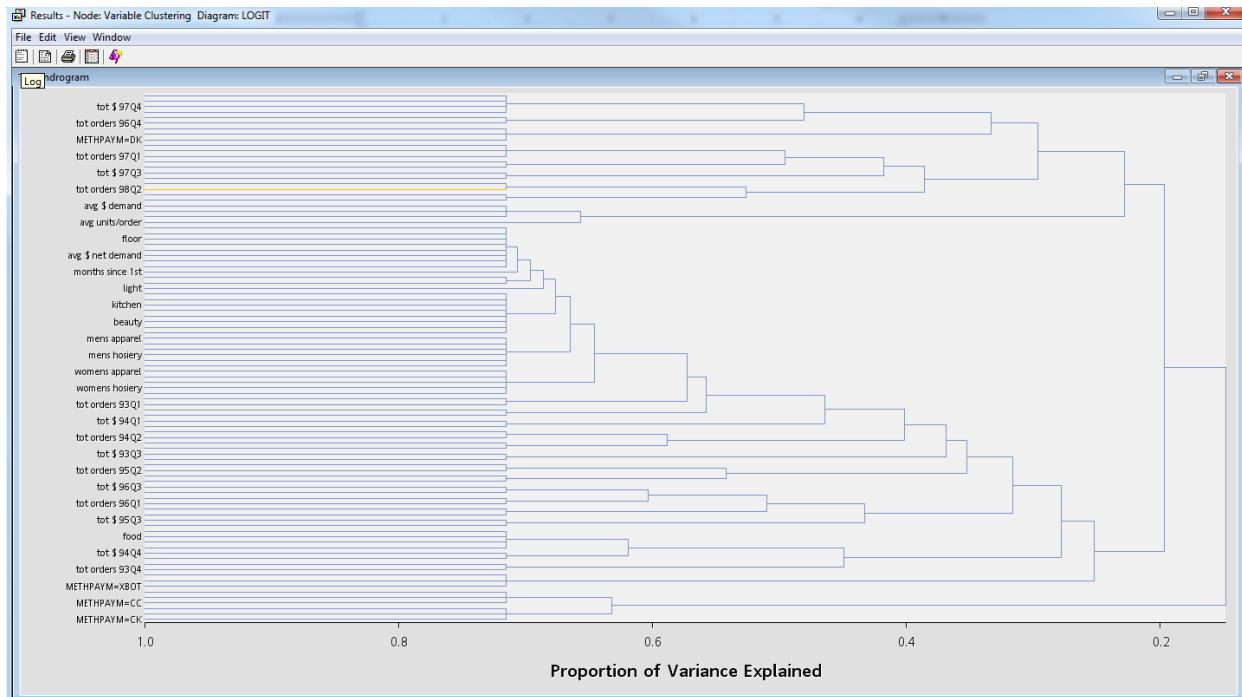
File Edit View Window

Selected Variables

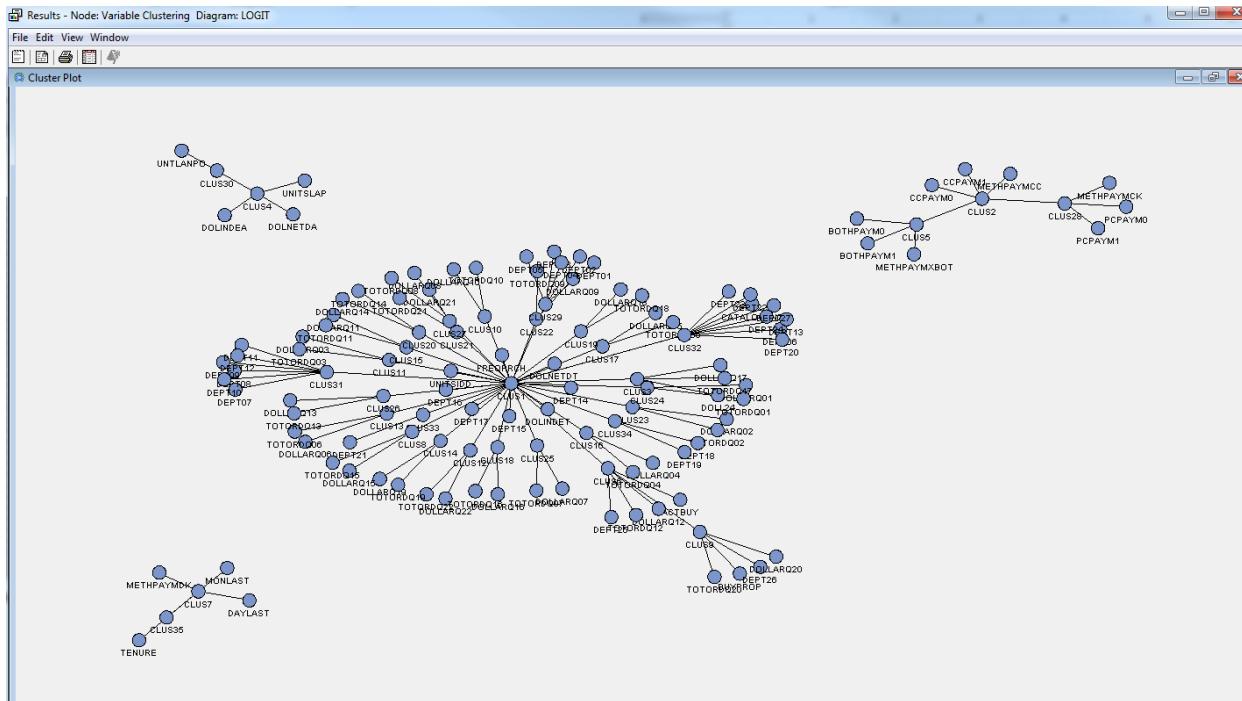
Cluster	Variable	Label	R-Square With Own Cluster Component	Next Closest Cluster
CLUS1	DOLINDET	total \$ demand		0.923851CLUS32
CLUS10	DOLLARQ10	tot \$ 95Q2		0.880676CLUS1
CLUS11	DOLLARQ03	tot \$ 93Q3		0.881598CLUS1
CLUS12	TOTORDQ22	tot orders 98Q2		0.863458CLUS7
CLUS13	TOTORDQ06	tot orders 94Q2		0.878785CLUS1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

Above picture shows the 35 clusters.

Dendrogram of cluster.



Cluster:



➤ LOGISTIC REGRESSION MODEL:

.. Property	Value
Variables	
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Forward
Selection Criterion	Validation Error
Use Selection Defaults	Yes
Selection Options	
Optimization Options	

Summary of Forward selection:

Summary of Forward Selection							
		Effect	Number	Score		Validation	
	Step	Entered	DF	In	Chi-Square	Pr > ChiSq	Error Rate
1181							
1182							
1183							
1184							
1185	Step	Entered	DF	In	Chi-Square	Pr > ChiSq	Error Rate
1186							
1187	1	DOLINDET	1	1	418.2287	<.0001	6878.3
1188	2	TOTORDQ20	1	2	178.2565	<.0001	6847.2
1189	3	MONLAST	1	3	113.3610	<.0001	6781.4
1190	4	TOTORDQ22	1	4	47.4870	<.0001	6751.4
1191	5	CATALOGCNT	1	5	36.9828	<.0001	6727.5
1192	6	TOTORDQ18	1	6	19.9779	<.0001	6719.0
1193	7	TOTORDQ21	1	7	14.9769	0.0001	6712.7
1194	8	TOTORDQ12	1	8	13.5709	0.0002	6701.1
1195	9	TOTORDQ19	1	9	11.8344	0.0006	6702.1
1196	10	DEPT03	1	10	10.4403	0.0012	6701.4
1197	11	CCPAYMO	1	11	9.3003	0.0023	6709.1
1198	12	TOTORDQ05	1	12	6.4600	0.0110	6716.5
1199	13	DOLLARQ09	1	13	5.3211	0.0211	6717.9
1200							
....							

T-test Analysis:

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

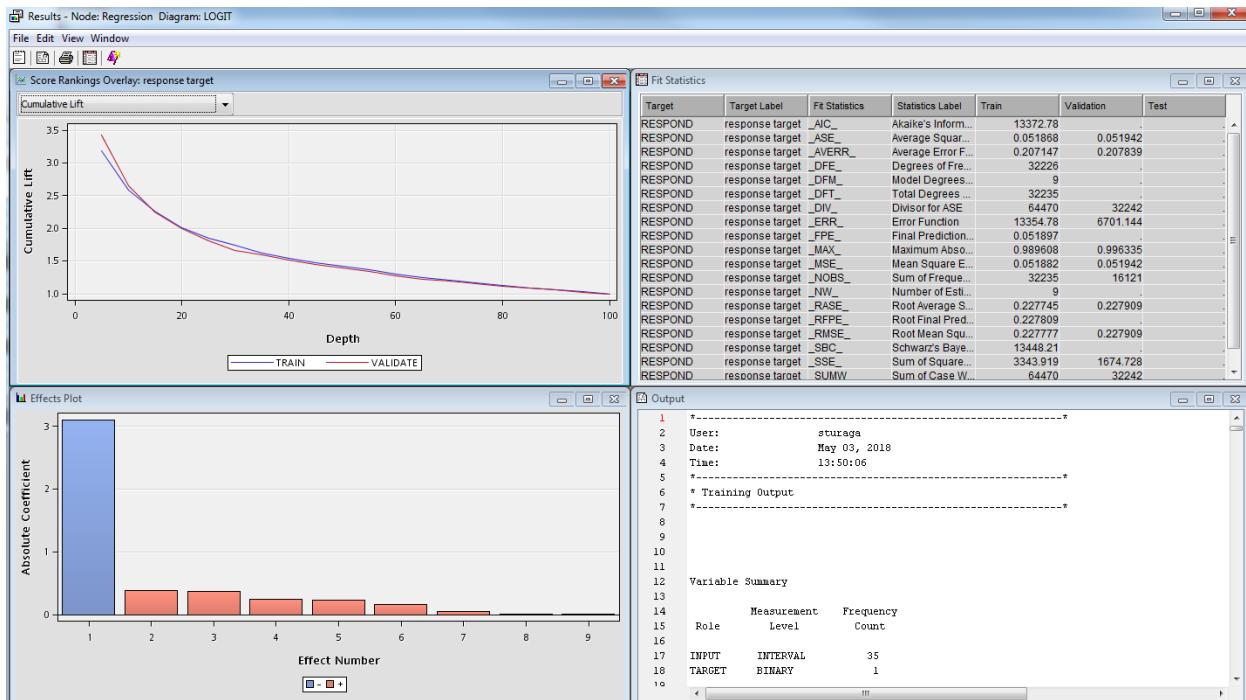
Data Role=VALIDATE Target Variable=RESPOND Target Label=response target								Mean
Depth	Gain	Lift	Cumulative Lift	% Response	Cumulative % Response	Number of Observations	Posterior Probability	
5	243.140	3.43140	3.43140	19.4548	19.4548	807	0.18015	
10	164.623	1.86007	2.64623	10.5459	15.0031	806	0.10221	
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:

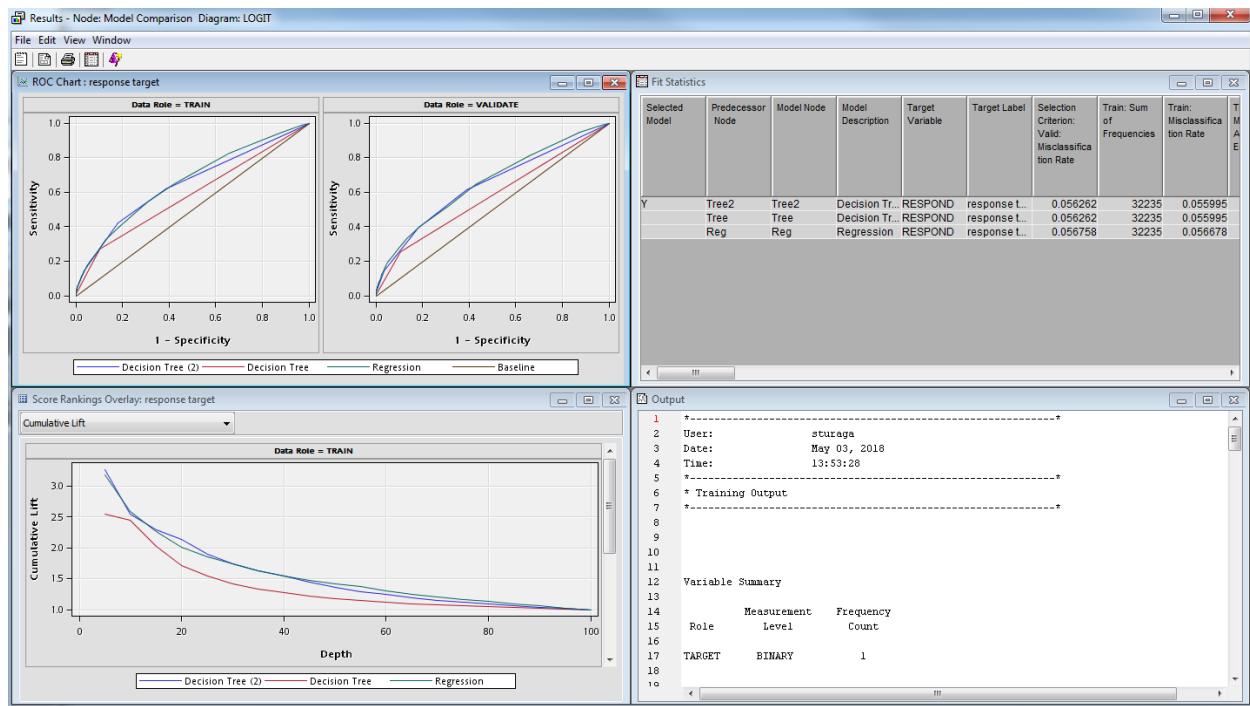
We can clearly say interpret that as we go deeper into the data set the probability decreases.

➤ COMPLETE OUTPUT OF REGRESSION:



➤ MODEL COMPARISION RESULTS:

Data Role=Valid	Tree2	Tree	Reg
Statistics			
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.