Predictive Analytics SAS Eminer Report

Prepared By:
Priya Kumari (800964015)
Shweta Rajaram Patil(800989198)

Course: Business Intelligence and Analytics
University of North Carolina at Charlotte

Part 1- Predictive Analytics_Exploratory

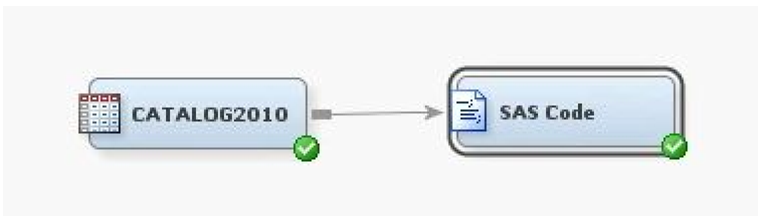
Task 1- Additional Analysis:

a) Finding Correlation Between Variables

Using Pearson Correlation Coefficient to find correlation between variables.

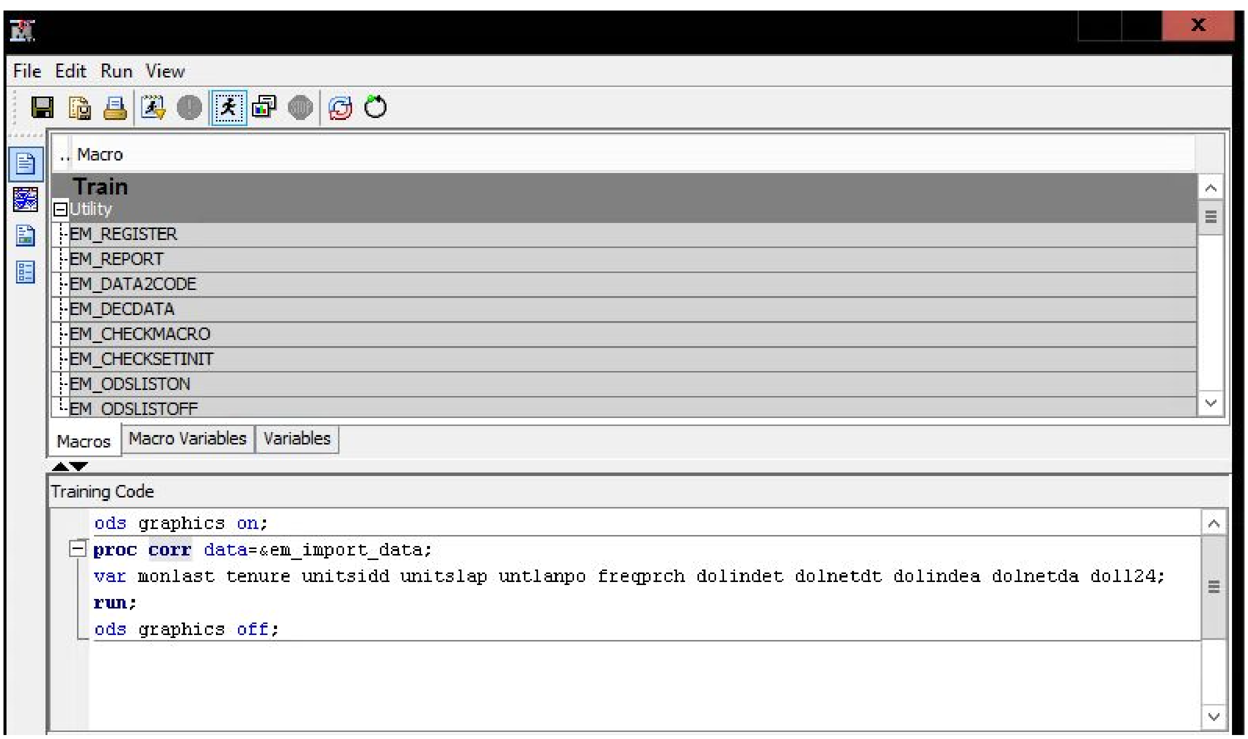
Steps to Execute:

1. Drag datasource CATALOG2010 into the Diagram workspace. Then click on Utility→ SAS code and drag the node into the workspace. Connect datasource to sas code node.



2. Select the SAS code node and in the left property panel, click on ellipsis next to code editor in TRAIN panel.

3. Write the sas code under Training code space and run the program.



The code for PROC CORR-

Training Code

```
ods graphics on;
proc corr data=&em_import_data;
var monlast tenure unitsidd unitslap untlanpo freqprch dolindet dolnetdt dolindea dolnetda doll24;
run;
ods graphics off;
```

Result :

Pearson Correlation Coefficients, N = 48356
Prob > |r| under H0: Rho=0

	MONLAST	TENURE	UNITSIDD	UNITSLAP	UNTLANPO	FREQPRCH	DOLINDET	DOLNETDT	DOLINDEA	DOLNETDA	DOLL24
MONLAST	1.00000	0.44650	-0.23683	0.28623	-0.17294	-0.20852	-0.19440	-0.18917	-0.02626	-0.00908	-0.36171
months since last		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0459	<.0001
TENURE	0.44650	1.00000	0.27977	0.13977	-0.18520	0.46937	0.33358	0.33915	-0.06631	-0.04341	-0.05180
months since 1st	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
UNITSIDD	-0.23683	0.27977	1.00000	-0.12612	0.34385	0.80447	0.88118	0.87736	0.20577	0.20924	0.53989
tot units demand	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
UNITSLAP	0.28623	0.13977	-0.12612	1.00000	-0.23436	-0.06290	0.07040	0.06737	0.49833	0.48301	0.00361
avg price/unit	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.4271
UNTLANPO	-0.17294	-0.18520	0.34385	-0.23436	1.00000	-0.01602	0.17865	0.17731	0.50953	0.50070	0.23678
avg units/order	<.0001	<.0001	<.0001	<.0001		0.0004	<.0001	<.0001	<.0001	<.0001	<.0001
FREQPRCH	-0.20852	0.46937	0.80447	-0.06290	-0.01602	1.00000	0.81540	0.81239	-0.01152	-0.00472	0.40266
lifetime orders	<.0001	<.0001	<.0001	<.0001	0.0004		<.0001	<.0001	0.0113	0.2988	<.0001
DOLINDET	-0.19440	0.33358	0.88118	0.07040	0.17865	0.81540	1.00000	0.99395	0.32696	0.32287	0.57729
total $ demand	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001
DOLNETDT	-0.18917	0.33915	0.87736	0.06737	0.17731	0.81239	0.99395	1.00000	0.31815	0.33505	0.56632
avg $ net demand	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001
DOLINDEA	-0.02626	-0.06631	0.20577	0.49833	0.50953	-0.01152	0.32696	0.31815	1.00000	0.95318	0.34512
avg $ demand	<.0001	<.0001	<.0001	<.0001	<.0001	0.0113	<.0001	<.0001		<.0001	<.0001
DOLNETDA	-0.00908	-0.04341	0.20924	0.48301	0.50070	-0.00472	0.32287	0.33505	0.95318	1.00000	0.32652
tot $ net demand	0.0459	<.0001	<.0001	<.0001	<.0001	0.2988	<.0001	<.0001	<.0001		<.0001
DOLL24	-0.36171	-0.05180	0.53989	0.00361	0.23678	0.40266	0.57729	0.56632	0.34512	0.32652	1.00000
$ last 24 months	<.0001	<.0001	<.0001	0.4271	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	

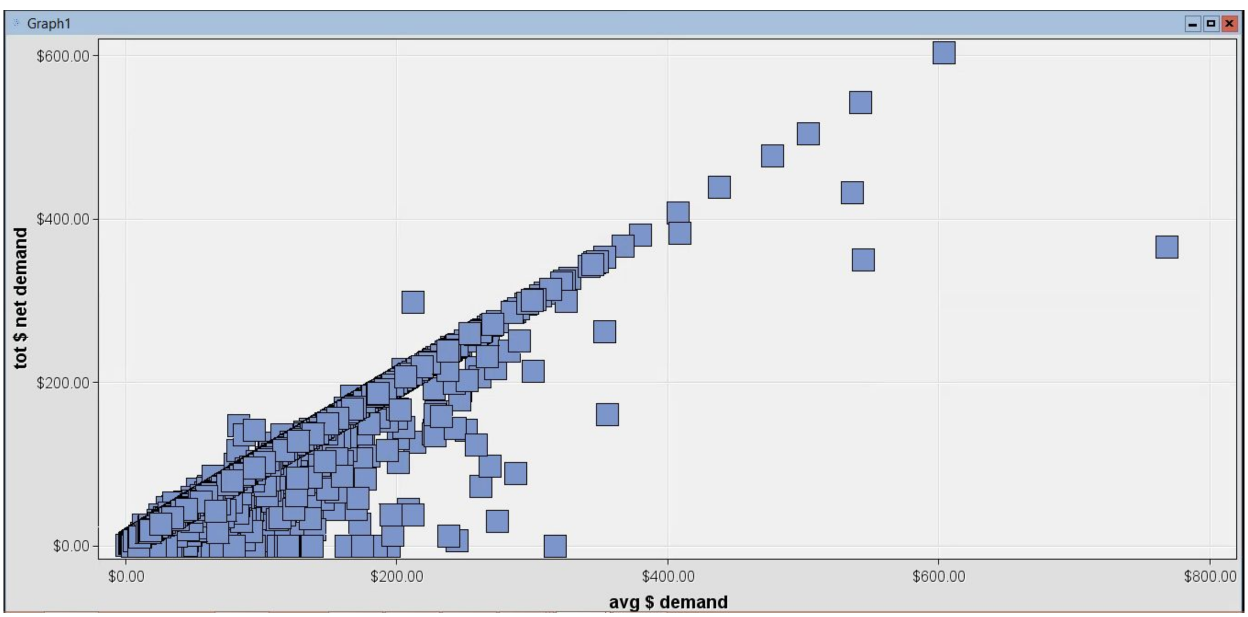
Looking at the above Pearson correlation statistics for pairs of analysis variables, we can say that:

- Variables DOLINDEA(avg $ demand) and DOLNETDA (tot $ net demand)are highly correlated and one of the variables can be dropped.

```
The CORR Procedure

   2  Variables:    DOLINDEA DOLNETDA


                                              Simple Statistics

Variable           N          Mean       Std Dev           Sum       Minimum       Maximum    Label

DOLINDEA       48356      47.74947      37.75177       2308973       1.00000     768.85000    avg $ demand
DOLNETDA       48356      45.30110      36.40940       2190580             0     768.50000    tot $ net demand


Pearson Correlation Coefficients, N = 48356
         Prob > |r| under H0: Rho=0

                      DOLINDEA      DOLNETDA

DOLINDEA               1.00000       0.95318
avg $ demand                          <.0001

DOLNETDA               0.95318       1.00000
tot $ net demand        <.0001
```

Pearson correlation coefficient for these variable is 0.95 which is a very high value and indicates significant linear relationship between the two.

Scatterplot :

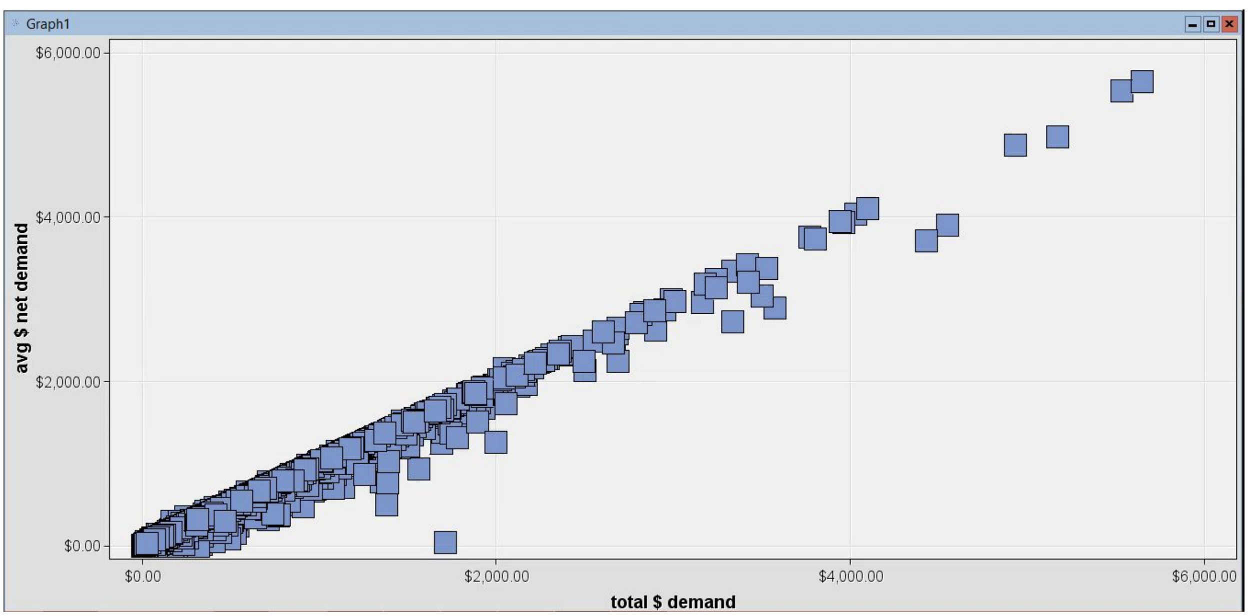


- Variables DOLNETDT(avg $ net demand) and DOLINDET (total $ demand) are highly correlated and one of the variables can be dropped.

```
The CORR Procedure

   2  Variables:    DOLNETDT DOLINDET


                                              Simple Statistics

Variable           N          Mean       Std Dev          Sum       Minimum       Maximum    Label

DOLNETDT       48356     187.85917     302.35363      9084118             0          8029    avg $ net demand
DOLINDET       48356     196.67031     314.09097      9510190       1.00000          7979    total $ demand


Pearson Correlation Coefficients, N = 48356
         Prob > |r| under H0: Rho=0

                      DOLNETDT      DOLINDET

DOLNETDT               1.00000       0.99395
avg $ net demand                      <.0001

DOLINDET               0.99395       1.00000
total $ demand          <.0001
```

Pearson correlation coefficient for these variable is 0.99 which is a very high value and indicates significant linear relationship between the two.

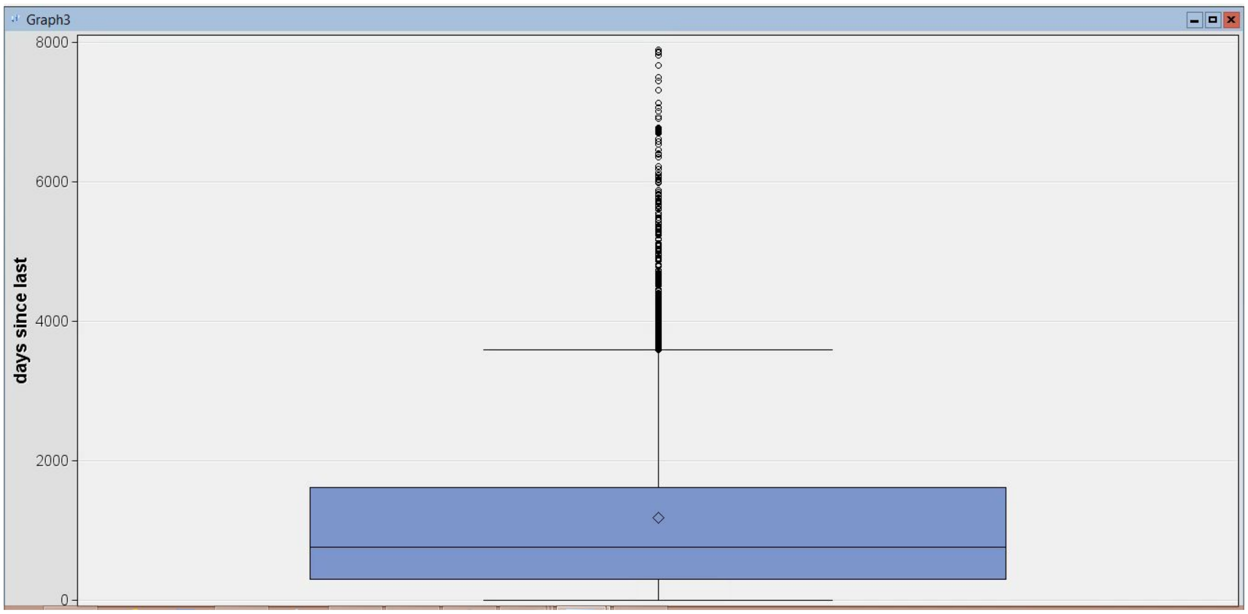
Scatterplot:



b) Outliers : In SAS Eminer, outliers can be determined using inbuilt Boxplot functionality or writing sas code. Both the methods are explored below for different variables.

-Right click on CATALOG2010 datasource and select Explore.
-Click on Plot and Select Box.
The Boxplot consists of the smallest observation, lower quartile (Q1), median, upper quartile (Q3), and largest observation; in addition, the boxplot indicates which observations, if any, are considered unusual, or outliers.

DAYLAST:

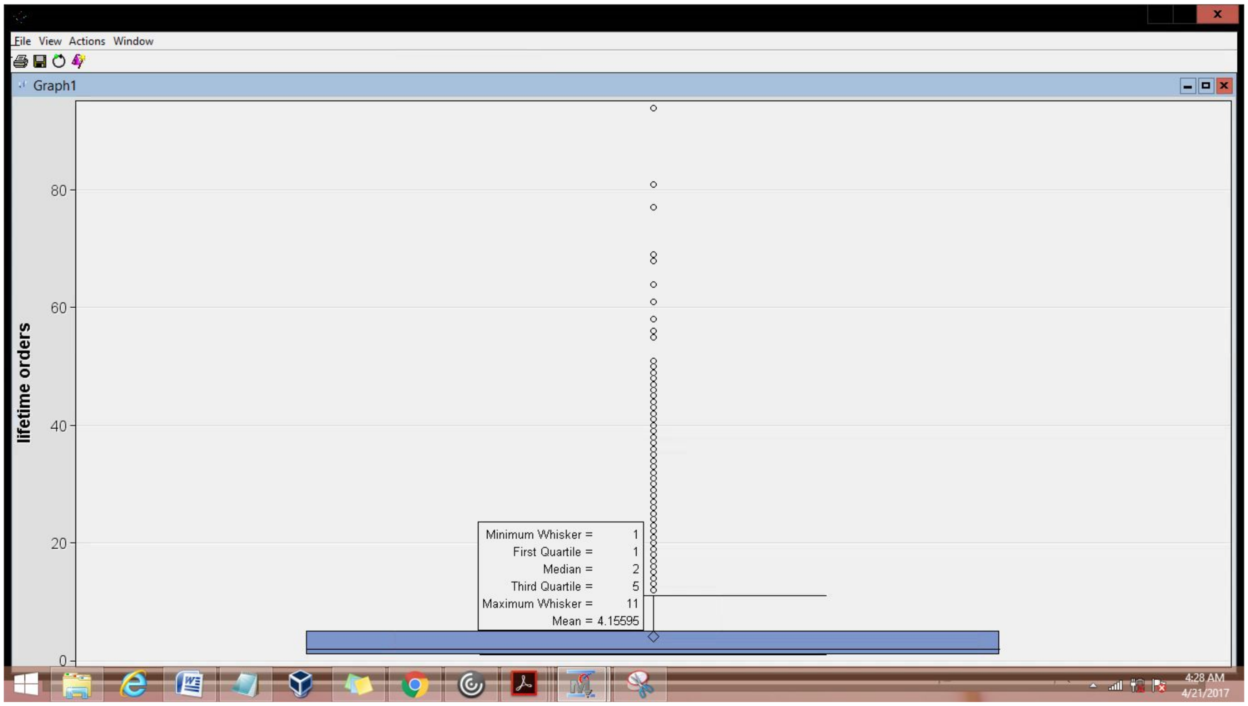


Minimum Whisker =	0
First Quartile =	292
Median =	761
Third Quartile =	1610
Maximum Whisker =	3588
Mean =	1180.7009

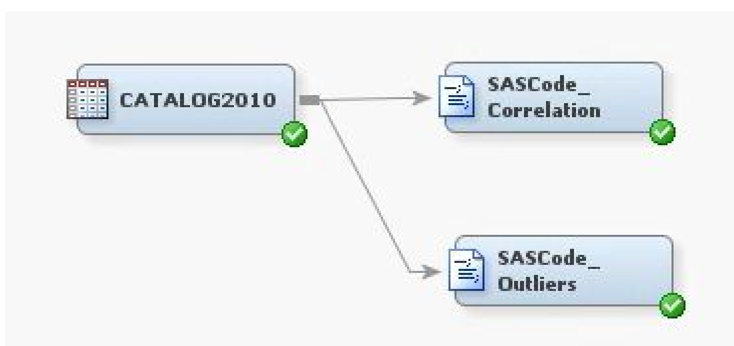
For DAYLAST variable, cutoff value for outliers is 3588.

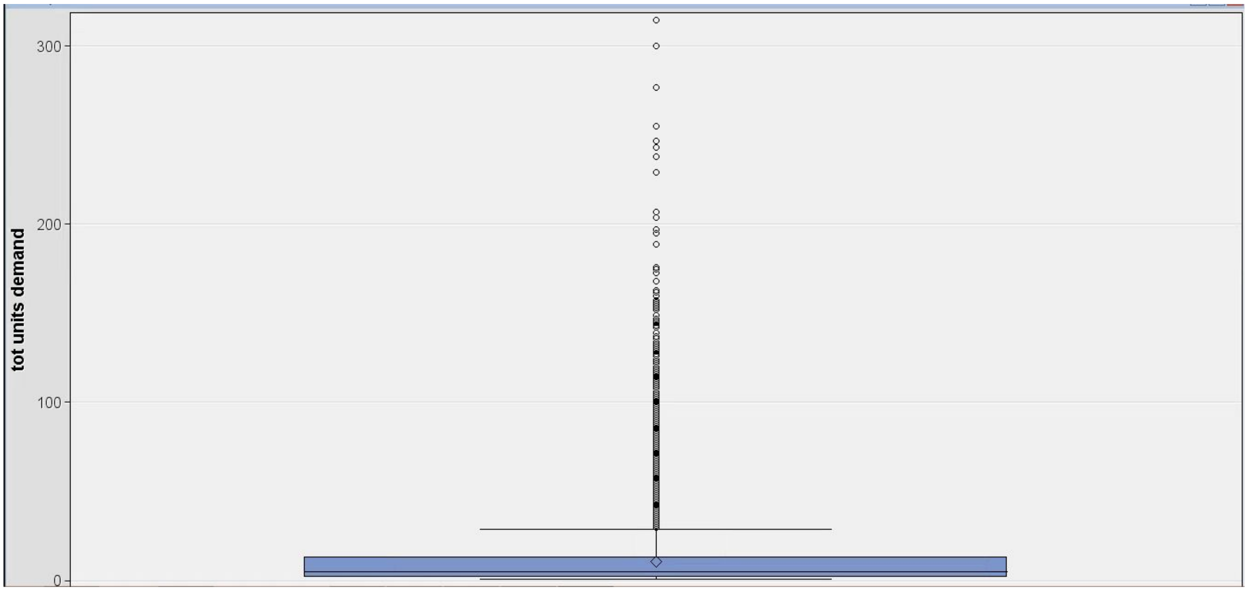
FREQPRCH:

This variable has high number of outliers. Cutoff value for outliers is 11. Many observations lie beyond the maximum whisker value.


File View Actions Window
Graph1
lifetime orders
80
60
40
20
0
Minimum Whisker = 1
First Quartile = 1
Median = 2
Third Quartile = 5
Maximum Whisker = 11
Mean = 4.15595
4:28 AM
4/21/2017

UNITSIDD:


CATALOG2010
SASCode_ Correlation
SASCode_ Outliers



SAS code :

```
proc univariate data=&em_import_data plot;
var unitsidd;
run;
```

Results:

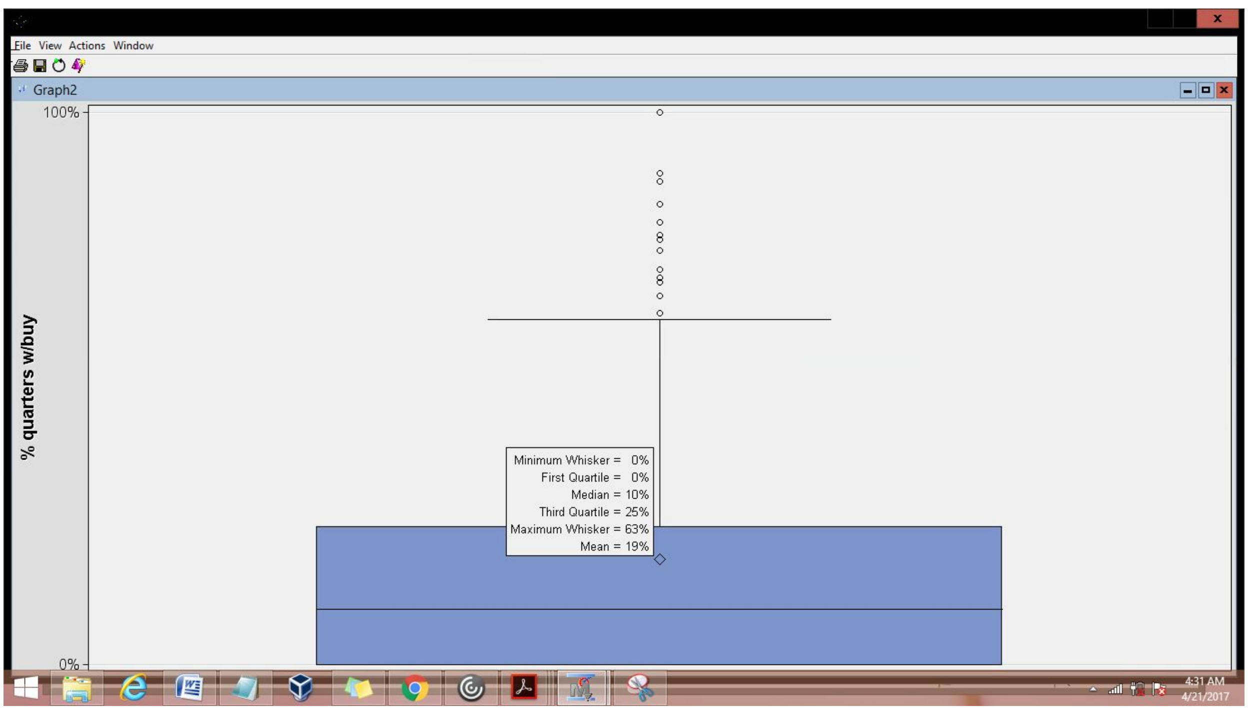
```
                    Histogram                        #  Boxplot                 Normal Probability Plot
 430+*                                               3     *        430+                                             *
    .                                                                  |
    .                                                                  |
    .                                                                  |
    .                                                                  |
    .*                                               3     *           |                                             *
    .*                                               2     *           |                                             *
 290+*                                               3     *        290+                                             *
    .*                                               3     *           |                                             *
    .*                                               5     *           |                                             *
    .*                                              11     *           |                                             *
    .*                                               9     *           |                                             *
    .*                                              16     *           |                                             *
    .*                                              25     *           |                                             *
 150+*                                              30     *        150+                                             *
    .*                                              58     *           |                                             *
    .*                                             107     *           |                                             *
    .*                                             220     *           |                                           ***
    .*                                             505     *           |                                        ****
    .**                                           1269     0           |                                     ****+++++
    .******                                       4804     0           |                              +++++*******
  10+************************************************* 41283  +--+--+    10+****************************************
     ----+----+----+----+----+----+----+----+---                           +----+----+----+----+----+----+----+----+----+
     * may represent up to 861 counts                                          -2        -1         0        +1        +2
```

The variable has high number of outliers with cutoff value of 29.

BUYPROP: Similarly, we can check for outliers in BUYPROP variable. Buyprop too has high number of outliers. Cutoff value for outliers is 63%

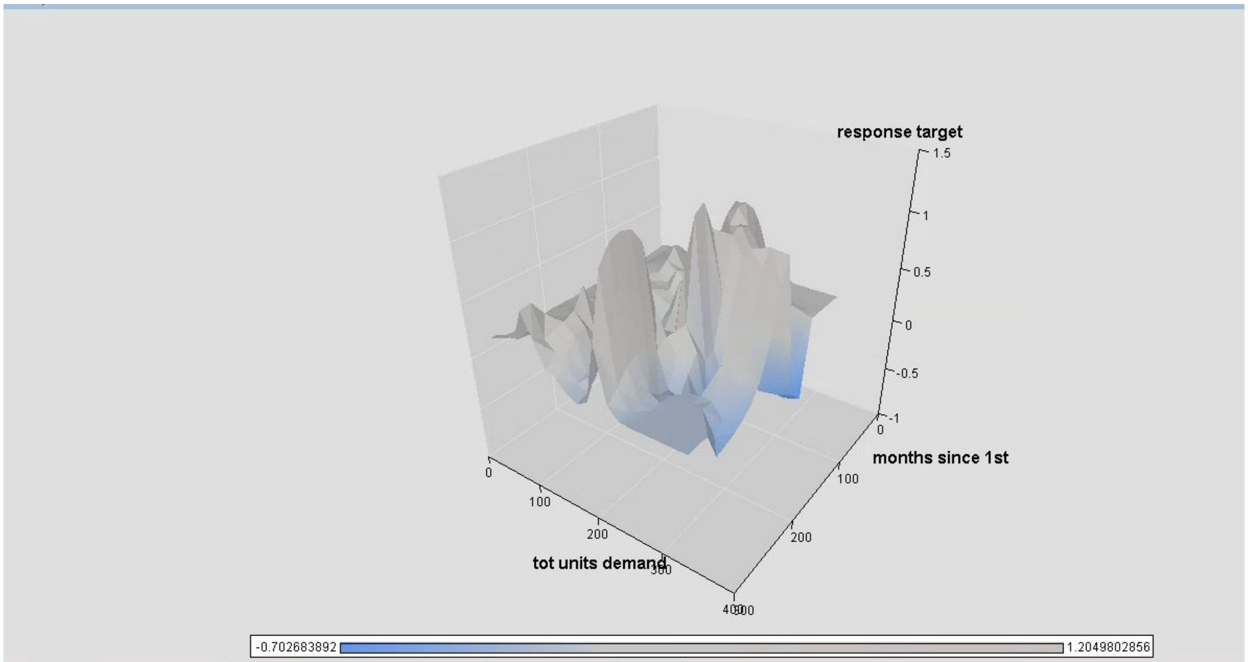
```
proc univariate data=&em_import_data plot;
var buyprop;
run;
```

```
                 Histogram                      #  Boxplot              Normal Probability Plot
1.025+*******                                2708     0      1.025+                          **********
     .                                                            |
0.925+*                                         2     0      0.925+                          *
     .*                                        16     0           |                          *
0.825+*                                        58     0      0.825+                          *       ++
     .*                                        92     0           |                          *     ++
0.725+*                                        38     0      0.725+                          *   ++
     .**                                      431     0           |                         ** ++
0.625+*                                       214     0      0.625+                        ** +
     .*                                       121     |           |                        *++
0.525+*******                                2838     |      0.525+                     ****
     .*                                        77     |           |                     *+
0.425+***                                     900     |      0.425+                   +**
     .**                                      409     |           |                 ++ *
0.325+********                               2903     |      0.325+               ++ ***
     .********                               3094  +-----+        |             ++ ***
0.225+********                               3158  |     |   0.225+           ++ ***
     .*******                                2809  |  +  |        |         ++ ***
0.125+************                           4696  *-----*   0.125+       ++ ***
     .***********                            4294  |     |        |     ++ ***
0.025+************************************************ 19498  +-----+   0.025+************************
      ----+----+----+----+----+----+----+----+--               +----+----+----+----+----+----+----+----+----+
      * may represent up to 407 counts                               -2        -1         0        +1        +2
```



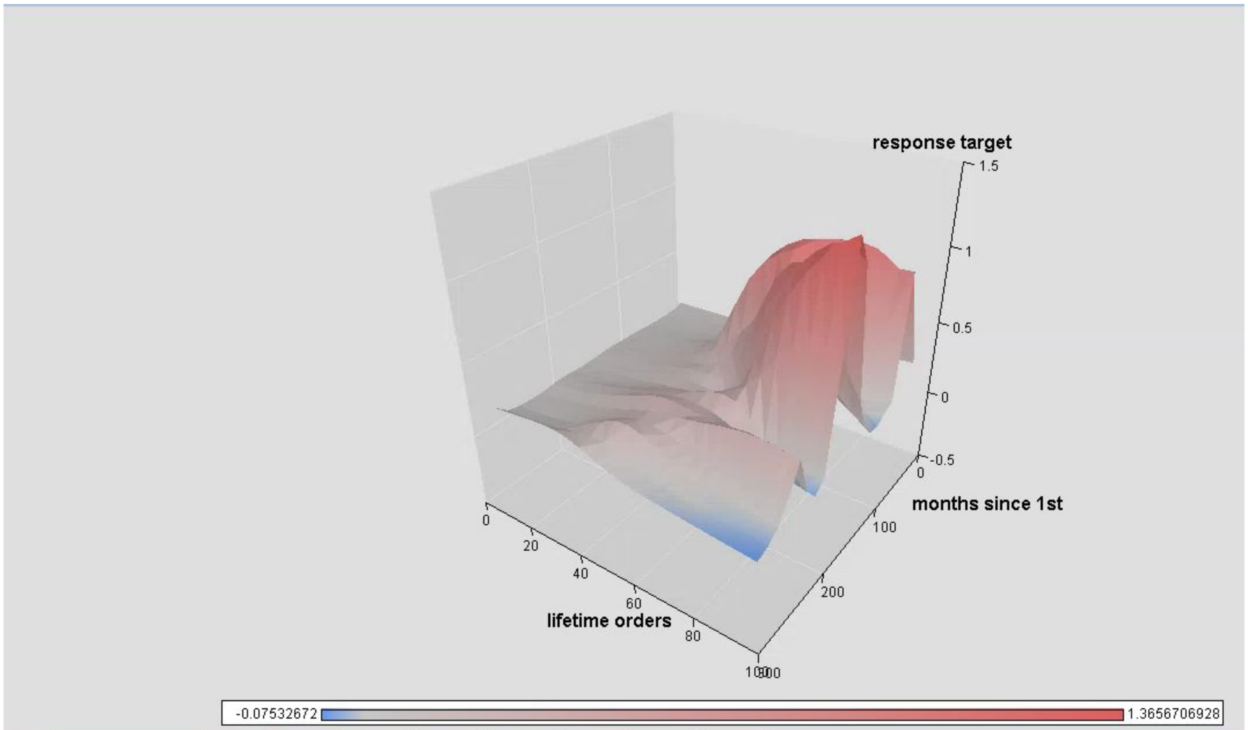
c) **3D Charts to show how independent variables impact dependent variable.**

From the below 3D chart, lets figure out how independent variables: TENURE(label: months since first) and UNITSIDD (tot Units demands) affect our target variable: RESPOND(response target). TENURE and UNITSIDD are not correlated variables.



From the chart above, we can say that for the population with tenure less than 200 months and total units demand in between 100 to 300, response target is positive.

Second 3D chart showing how TENURE and FREQPRCH(lifetime orders) affect the response target RESPOND.



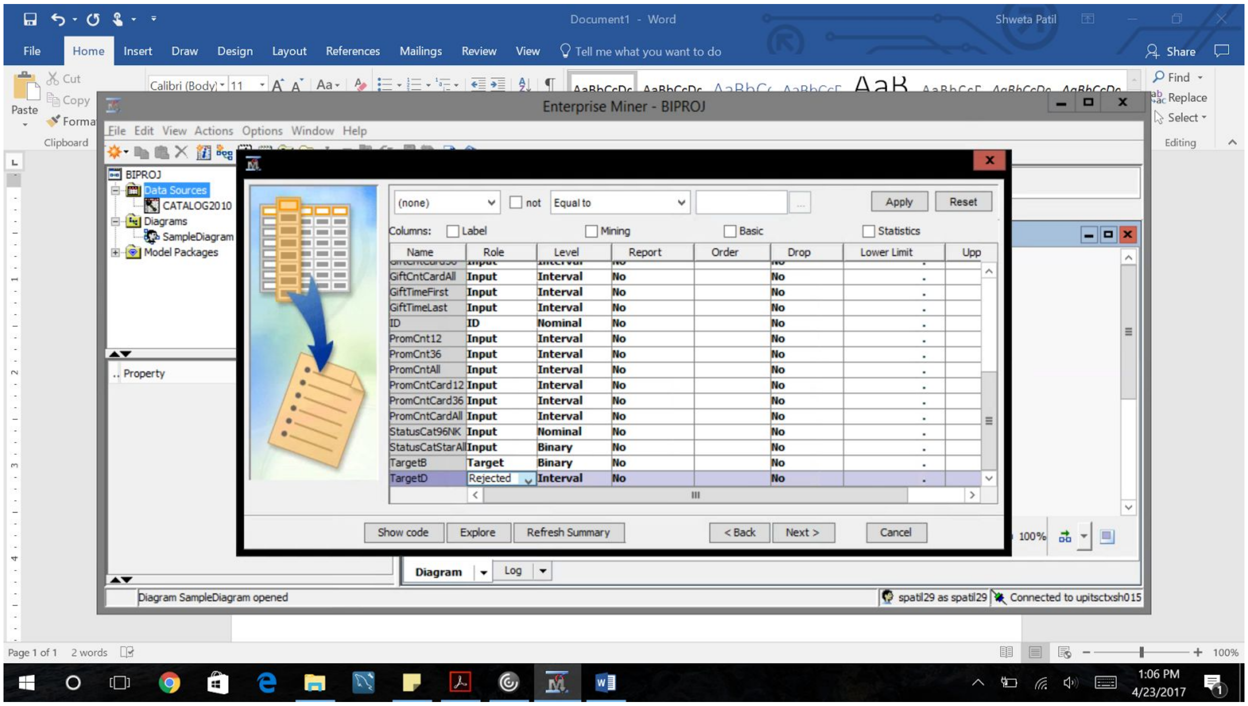
The Chart clearly states that response target is largely negatively related to both the independent variables but for tenure <= 100 months and FREQPRCH >40, the response target is positive.

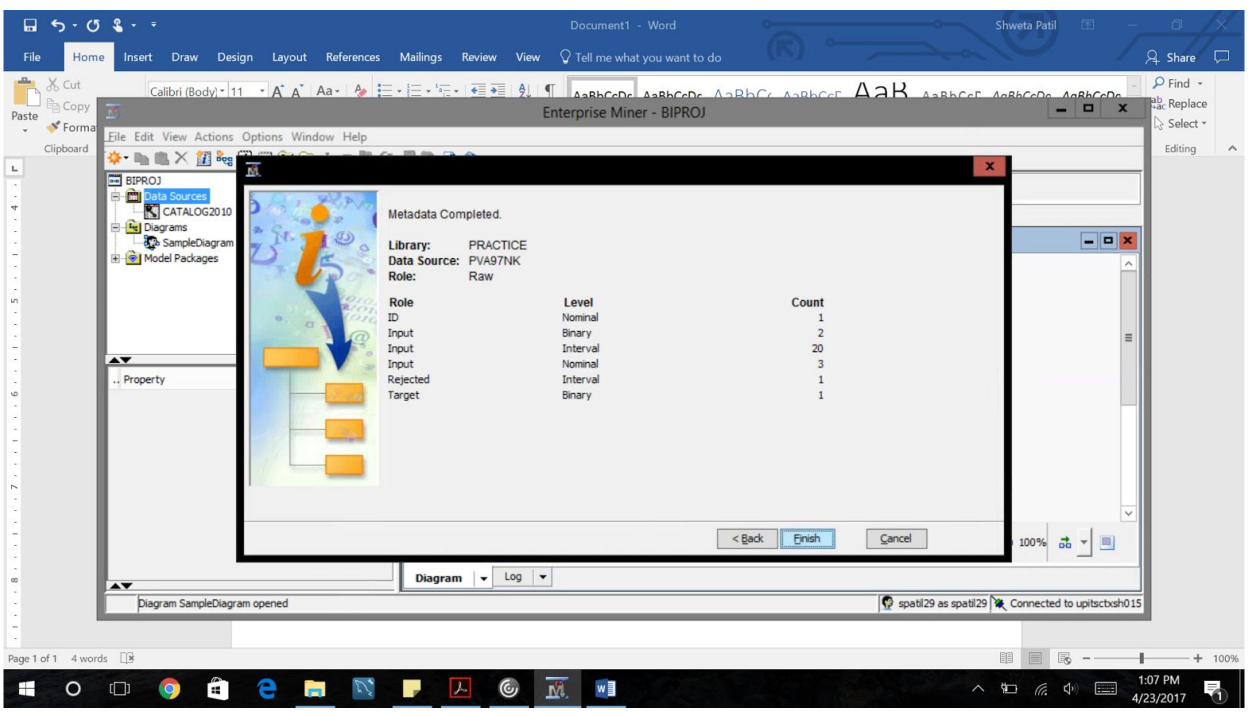
Task 2: RFM Analysis of Charity Direct Mail Data

DataSource: PVA97NK

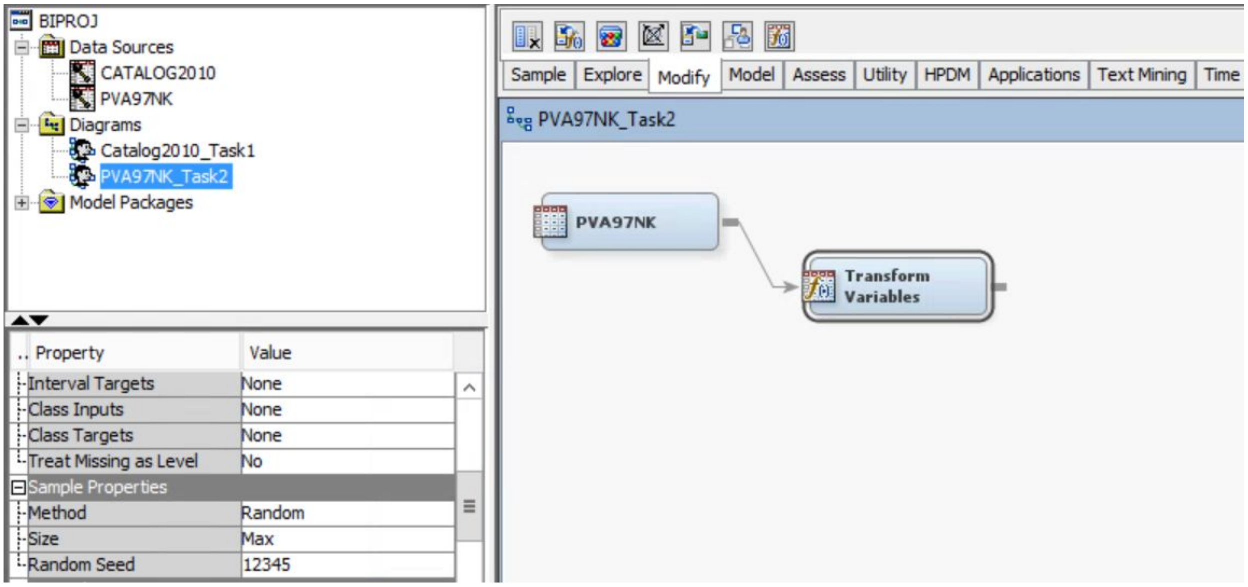
a)

TargetID rejected:

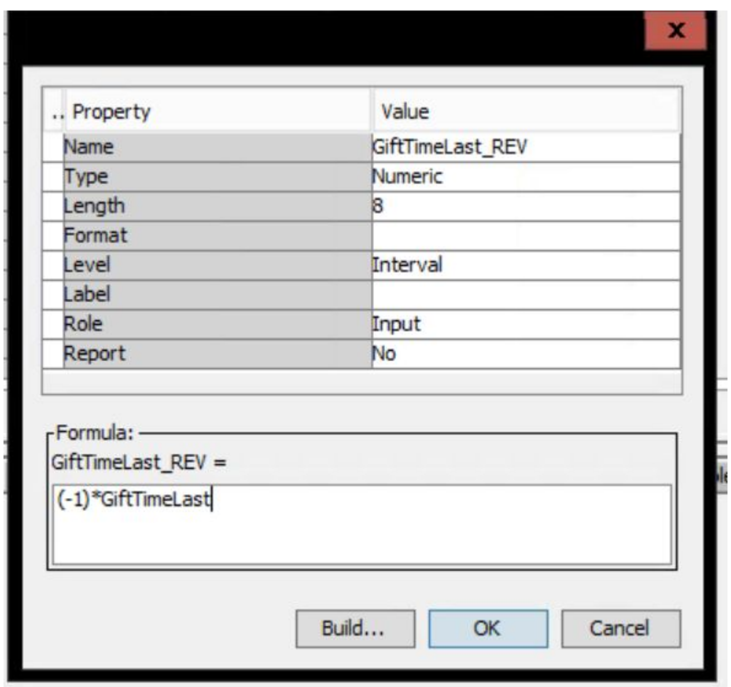




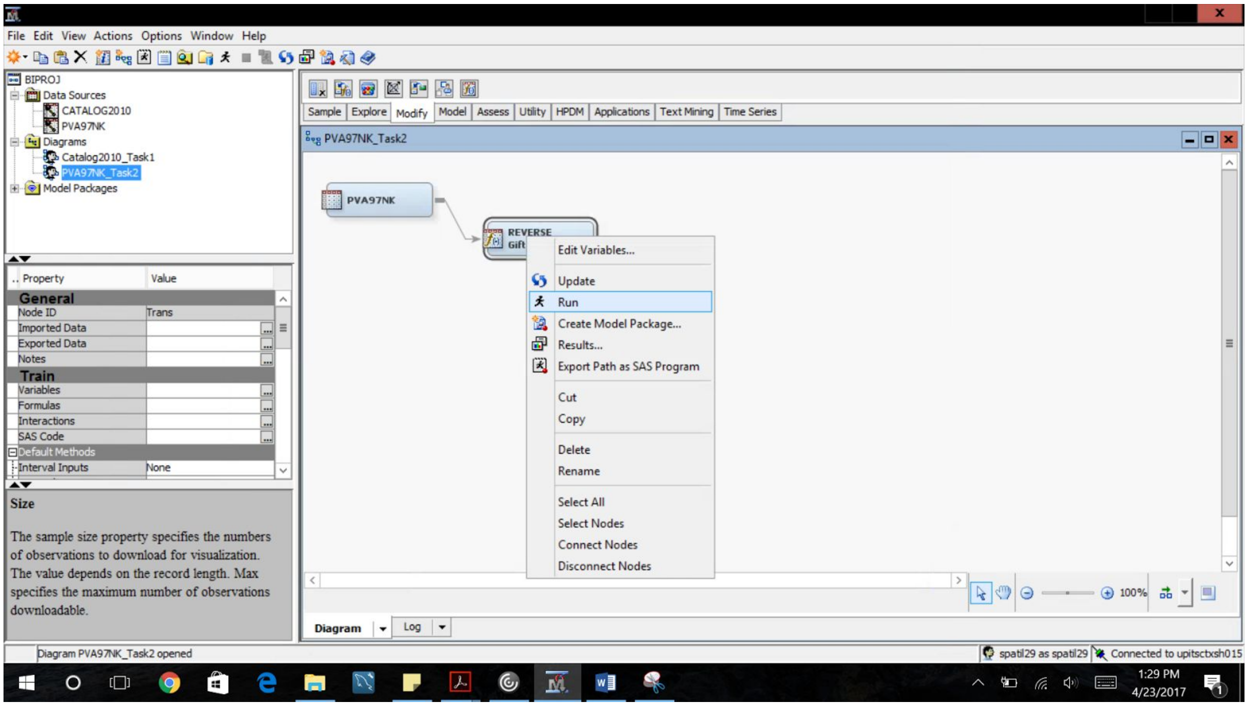
b)



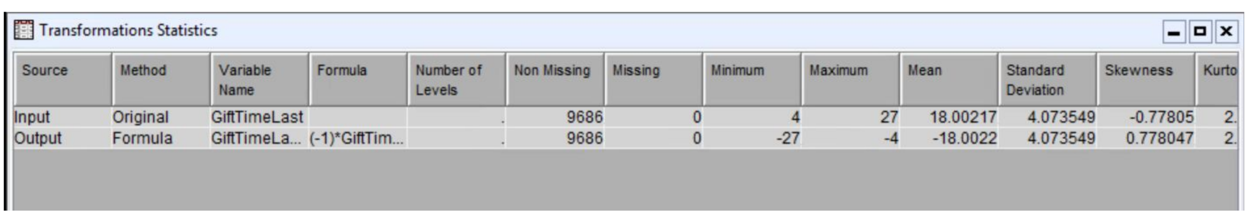
For Recency:



Run the transformation for Recency as (-1)*GiftTimeLast



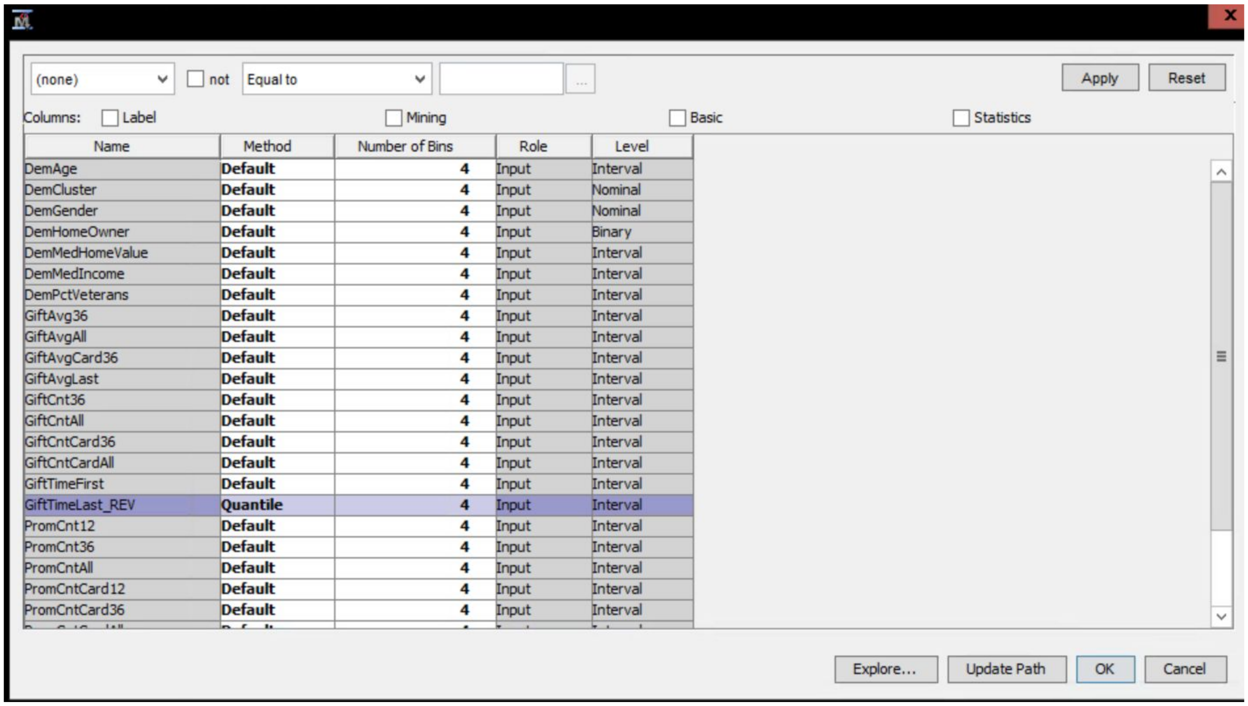
Output:



Transformations Statistics

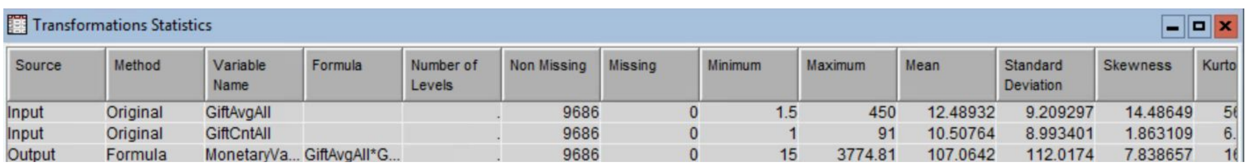
Source	Method	Variable Name	Formula	Number of Levels	Non Missing	Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurto
Input	Original	GiftTimeLast		.	9686	0	4	27	18.00217	4.073549	-0.77805	2.
Output	Formula	GiftTimeLa...	(-1)*GiftTim...	.	9686	0	-27	-4	-18.0022	4.073549	0.778047	2.

Quantiles:



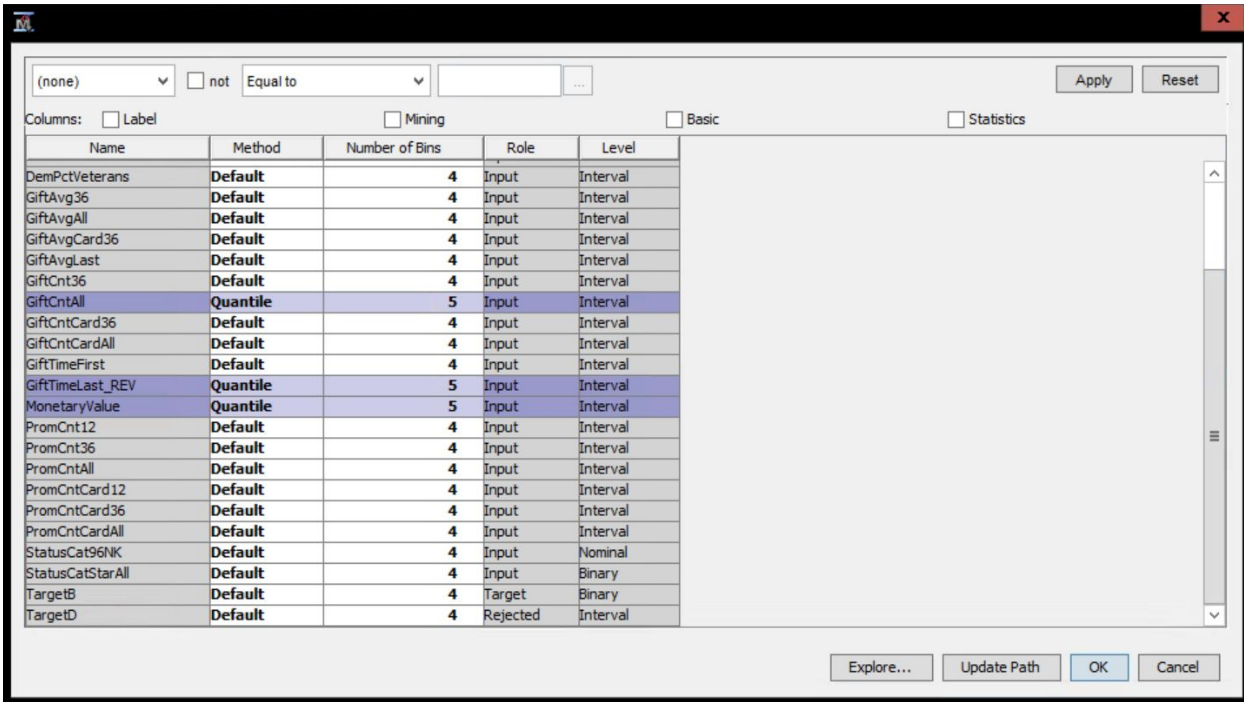
Name	Method	Number of Bins	Role	Level
DemAge	**Default**	4	Input	Interval
DemCluster	**Default**	4	Input	Nominal
DemGender	**Default**	4	Input	Nominal
DemHomeOwner	**Default**	4	Input	Binary
DemMedHomeValue	**Default**	4	Input	Interval
DemMedIncome	**Default**	4	Input	Interval
DemPctVeterans	**Default**	4	Input	Interval
GiftAvg36	**Default**	4	Input	Interval
GiftAvgAll	**Default**	4	Input	Interval
GiftAvgCard36	**Default**	4	Input	Interval
GiftAvgLast	**Default**	4	Input	Interval
GiftCnt36	**Default**	4	Input	Interval
GiftCntAll	**Default**	4	Input	Interval
GiftCntCard36	**Default**	4	Input	Interval
GiftCntCardAll	**Default**	4	Input	Interval
GiftTimeFirst	**Default**	4	Input	Interval
GiftTimeLast_REV	**Quantile**	4	Input	Interval
PromCnt12	**Default**	4	Input	Interval
PromCnt36	**Default**	4	Input	Interval
PromCntAll	**Default**	4	Input	Interval
PromCntCard12	**Default**	4	Input	Interval
PromCntCard36	**Default**	4	Input	Interval

MonetaryValue = GiftAvgAll*GiftCntAll

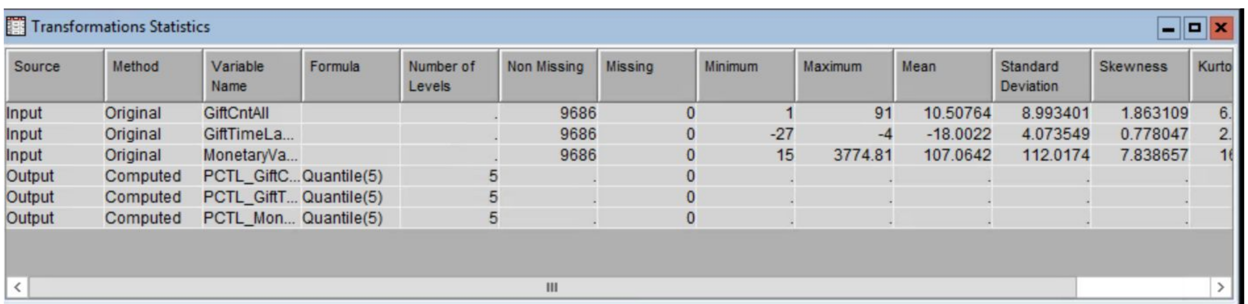


Source	Method	Variable Name	Formula	Number of Levels	Non Missing	Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurto
Input	Original	GiftAvgAll		.	9686	0	1.5	450	12.48932	9.209297	14.48649	5
Input	Original	GiftCntAll		.	9686	0	1	91	10.50764	8.993401	1.863109	6.
Output	Formula	MonetaryVa...	GiftAvgAll*G...	.	9686	0	15	3774.81	107.0642	112.0174	7.838657	1

Frequency:



Name	Method	Number of Bins	Role	Level
DemPctVeterans	Default	4	Input	Interval
GiftAvg36	Default	4	Input	Interval
GiftAvgAll	Default	4	Input	Interval
GiftAvgCard36	Default	4	Input	Interval
GiftAvgLast	Default	4	Input	Interval
GiftCnt36	Default	4	Input	Interval
GiftCntAll	Quantile	5	Input	Interval
GiftCntCard36	Default	4	Input	Interval
GiftCntCardAll	Default	4	Input	Interval
GiftTimeFirst	Default	4	Input	Interval
GiftTimeLast_REV	Quantile	5	Input	Interval
MonetaryValue	Quantile	5	Input	Interval
PromCnt12	Default	4	Input	Interval
PromCnt36	Default	4	Input	Interval
PromCntAll	Default	4	Input	Interval
PromCntCard12	Default	4	Input	Interval
PromCntCard36	Default	4	Input	Interval
PromCntCardAll	Default	4	Input	Interval
StatusCat96NK	Default	4	Input	Nominal
StatusCatStarAll	Default	4	Input	Binary
TargetB	Default	4	Target	Binary
TargetD	Default	4	Rejected	Interval



Source	Method	Variable Name	Formula	Number of Levels	Non Missing	Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurto
Input	Original	GiftCntAll		.	9686	0	1	91	10.50764	8.993401	1.863109	6.
Input	Original	GiftTimeLa...		.	9686	0	-27	-4	-18.0022	4.073549	0.778047	2.
Input	Original	MonetaryVa...		.	9686	0	15	3774.81	107.0642	112.0174	7.838657	1
Output	Computed	PCTL_GiftC...	Quantile(5)	5	.	0	.	.	.	.	.	.
Output	Computed	PCTL_GiftT...	Quantile(5)	5	.	0	.	.	.	.	.	.
Output	Computed	PCTL_Mon...	Quantile(5)	5	.	0	.	.	.	.	.	.

Investigate Binned Values:

	Gift Count All Mont...	GiftTimeLast_REV	MonetaryValue	Transform...	Transform...	Transform...
0	4	-21	37	02:3-6	02:-21--18	02:36-65.01
0	8	-26	127.04	03:6-10	01:low--21	04:99.96-1...
0	41	-18	152.93	05:16-high	03:-18--17	05:150.96-...
2	12	-9	102	04:10-16	05:-16-high	04:99.96-1...
9	1	-21	20	01:low-3	02:-21--18	01:low-36
4	11	-22	90.97	04:10-16	01:low--21	03:65.01-9...
8	4	-17	52	02:3-6	04:-17--16	02:36-65.01
0	4	-18	46	02:3-6	03:-18--17	02:36-65.01
0	3	-17	84.99	02:3-6	04:-17--16	03:65.01-9...
0	5	-18	58	02:3-6	03:-18--17	02:36-65.01
0	16	-17	98.08	05:16-high	04:-17--16	03:65.01-9...
0	13	-19	51.09	04:10-16	02:-21--18	02:36-65.01
0	1	-16	20	01:low-3	05:-16-high	01:low-36
0	4	-5	46	02:3-6	05:-16-high	02:36-65.01
0	2	-18	25	01:low-3	03:-18--17	01:low-36
0	12	-23	72	04:10-16	01:low--21	03:65.01-9...
0	7	-24	56.98	03:6-10	01:low--21	02:36-65.01
5	15	-17	85.05	04:10-16	04:-17--16	03:65.01-9...
0	5	-18	176	02:3-6	03:-18--17	05:150.96-...
0	9	-21	46.98	03:6-10	02:-21--18	02:36-65.01
7	20	-18	539	05:16-high	03:-18--17	05:150.96-...
0	5	-18	46	02:3-6	03:-18--17	02:36-65.01
0	21	-22	137.97	05:16-high	01:low--21	04:99.96-1...
0	1	-19	15	01:low-3	02:-21--18	01:low-36

RFM = substr(pctl_GiftTimeLast_REV,1,2)||substr(pctl_GiftCntAll,1,2)||substr(pctl_MonetaryValue,1,2)

File Edit View Window

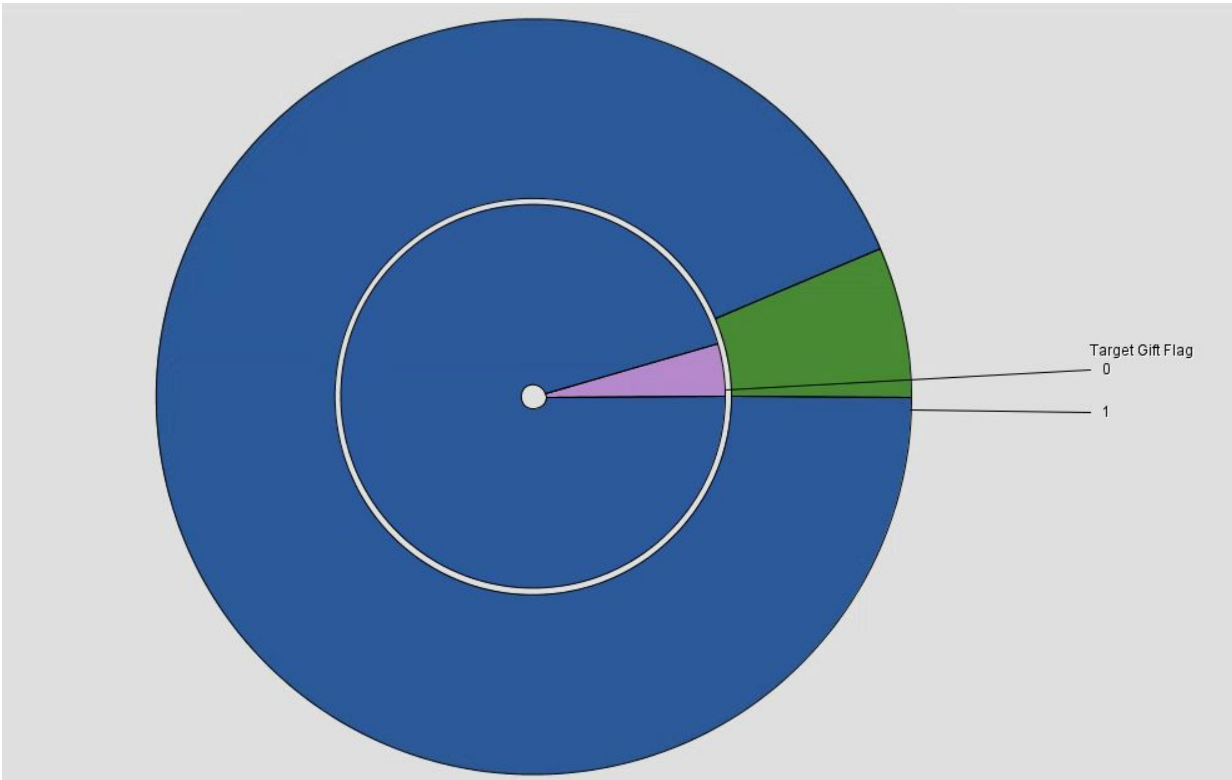
Log ...sformations Statistics

Source	Method	Variable Name	Formula	Number of Levels	Non Missing	Missing	Minimum	Maximum	Mean	Standard Deviation	Sk
Input	Original	PCTL_GiftC...		5	.	0	.	.	.	.	
Input	Original	PCTL_GiftT...		5	.	0	.	.	.	.	
Input	Original	PCTL_Mon...		5	.	0	.	.	.	.	
Output	Formula	RFM	substr(pctl_GiftTimeLast_...	116	.	0	.	.	.	.	

The final model looks like:

c) **Graphical RFM Analysis**

Grouped Pie Chart:



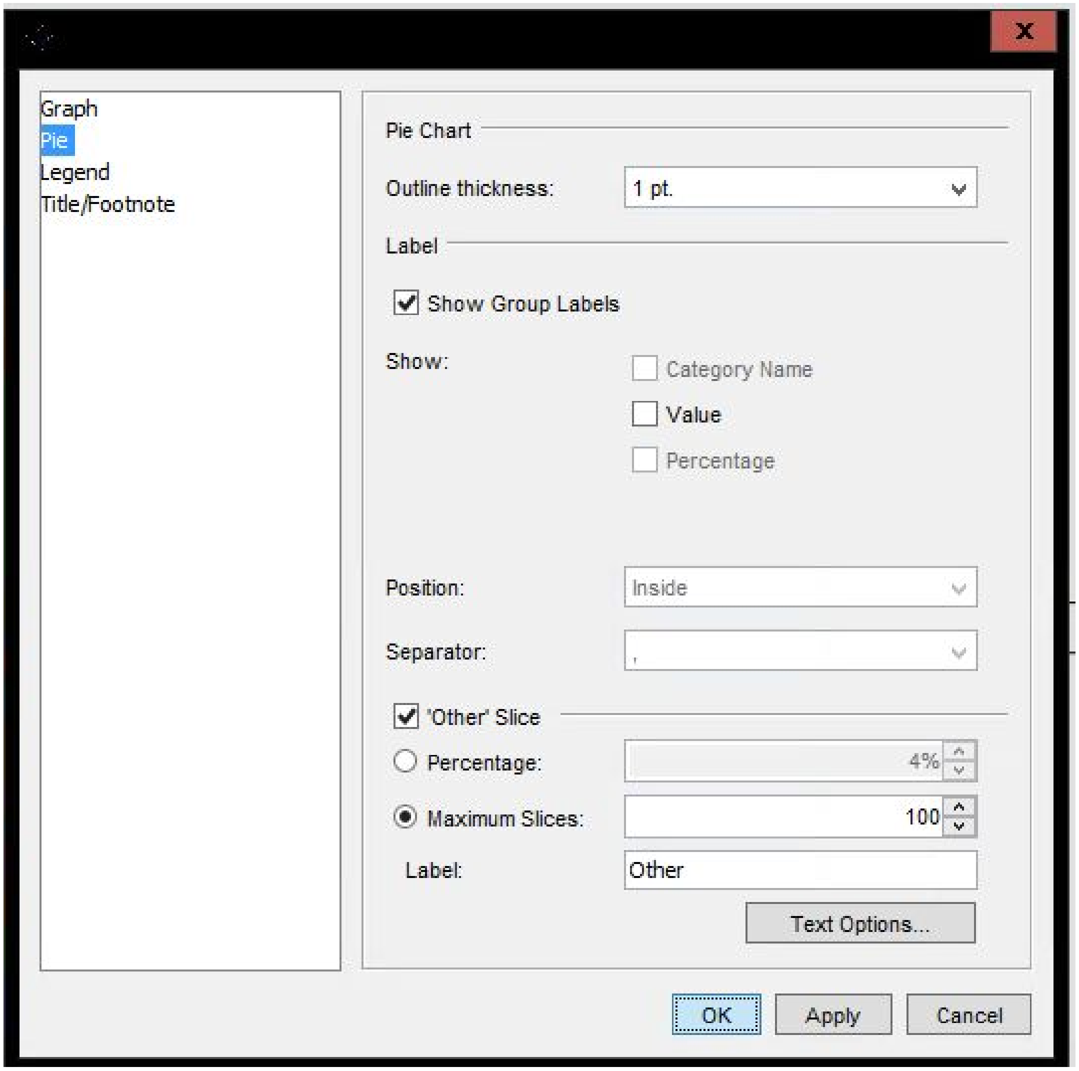
Target Variable: TargetB (Target Gift Flag)

Inner Circle: Target Gift Flag = 0

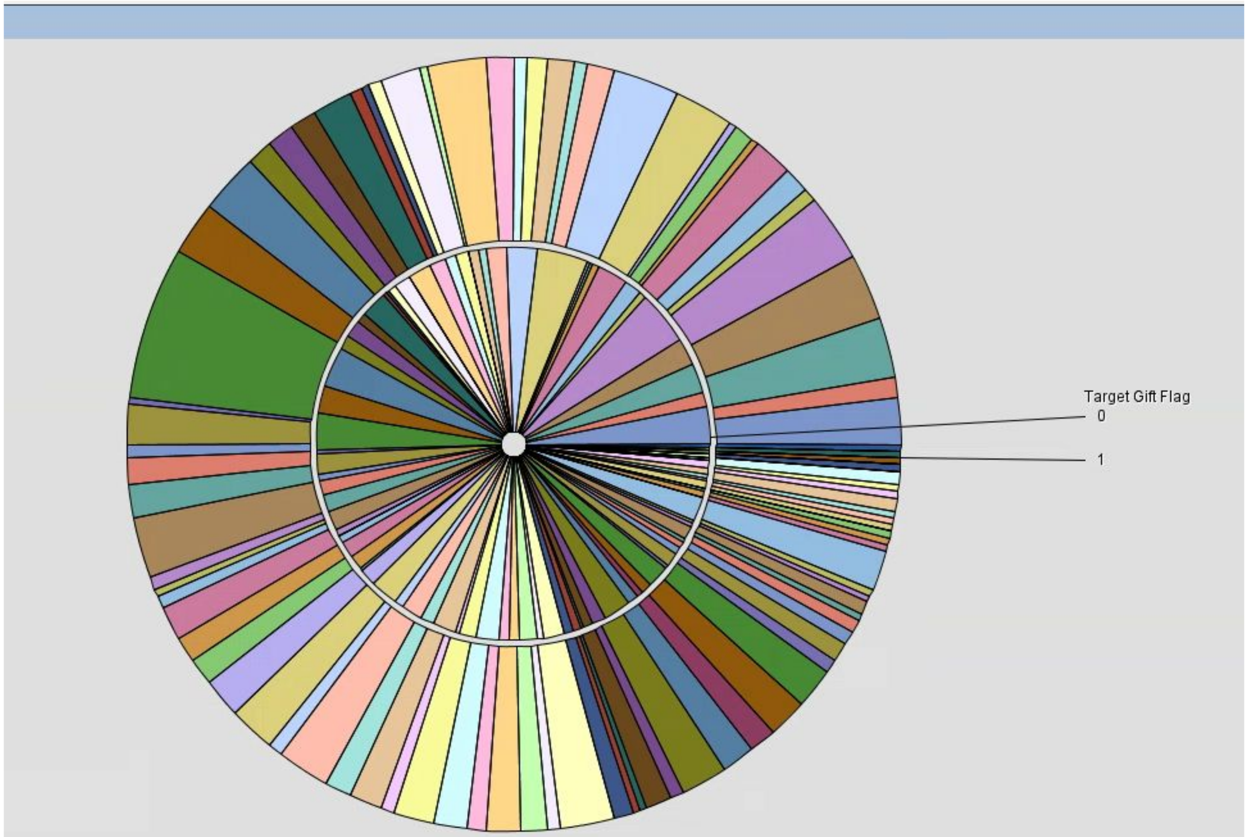
Outer Circle: Target Gift Flag = 1

A major part of pie chart is labelled for RFM=“Other”. To visualize more slices, we can edit the Graph Properties. Unchecking the ‘Other’ Slice checkbox does not show other slices/groups because there are too many categories to be fitted into the graph.

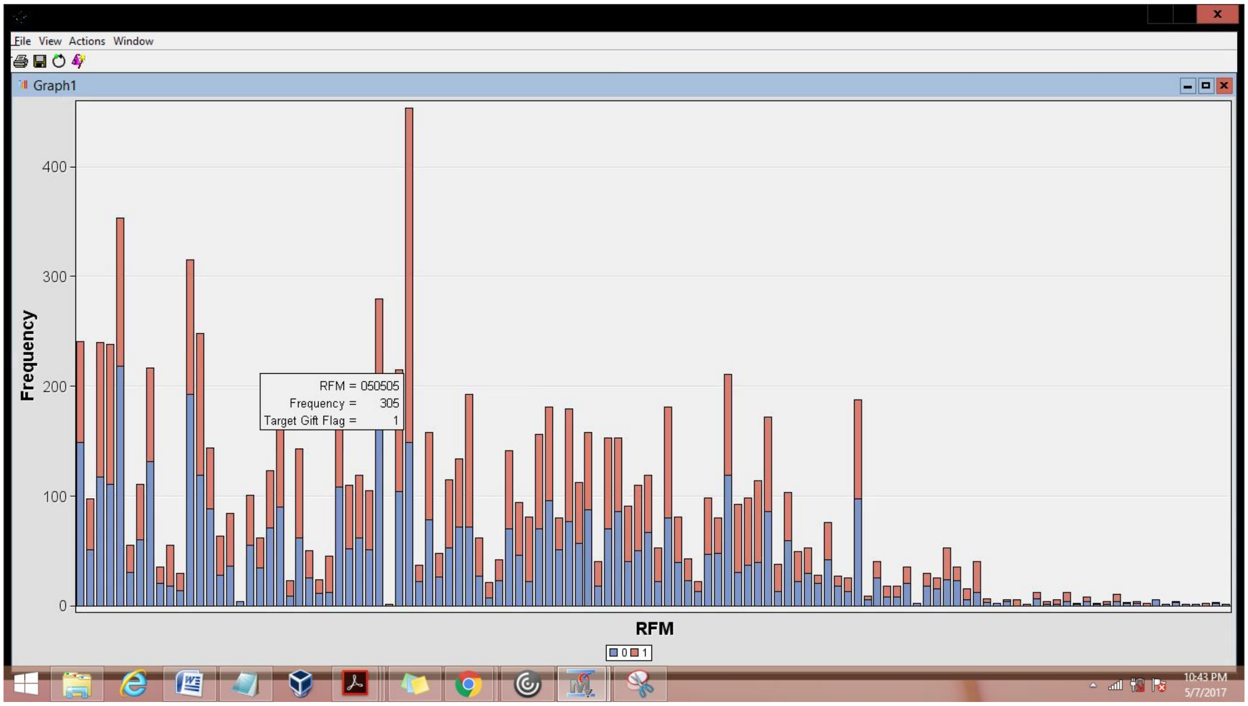
So,we set maximum slices values: 100 as shown below:



Resulting Pie Chart:

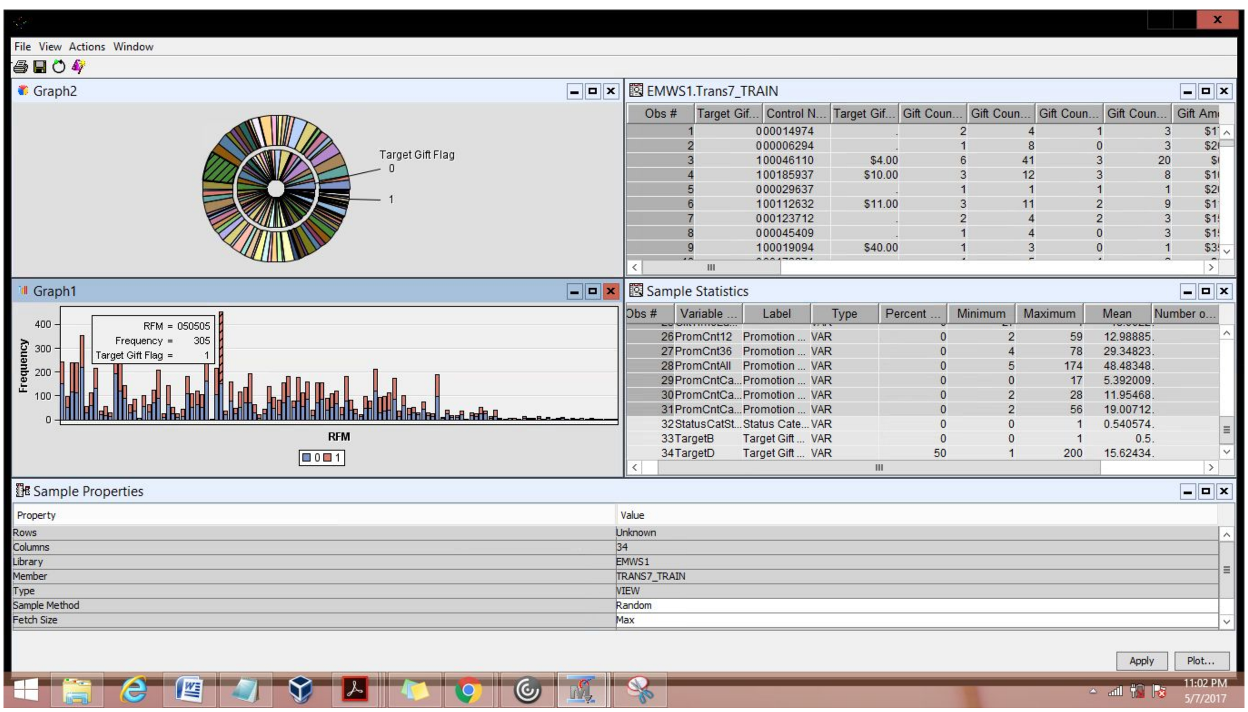


Stacked Bar Chart – for the ease of reading observations and the trend of RFM versus dependent variables:



Brick Red Part of the bar chart shows the proportion where the Dependent variable, Target Gift Flag = 1, whereas the blue part shows Target Gift Flag = 0.

Tile View of grouped pie chart and a stacked bar chart:



The RFM Category 050505 has highest target gift value =1.

Frequency is 305: Target Gift Flag = 1

Frequency is 149: Target Gift Flag = 0

d) Break-even Response Rate = $\frac{\text{Current Cost of Promotion for each gift}}{\text{Average Donation}} = 1.5/15 = 0.1 \rightarrow 10\%$

Therefore, Profitable RFM cells are those with response rate > 10%
Most of the RFM categories will exceed this response rate. This may be because of oversampling as we are working with a sample of original data which consists of 50% responders and 50% non responders.

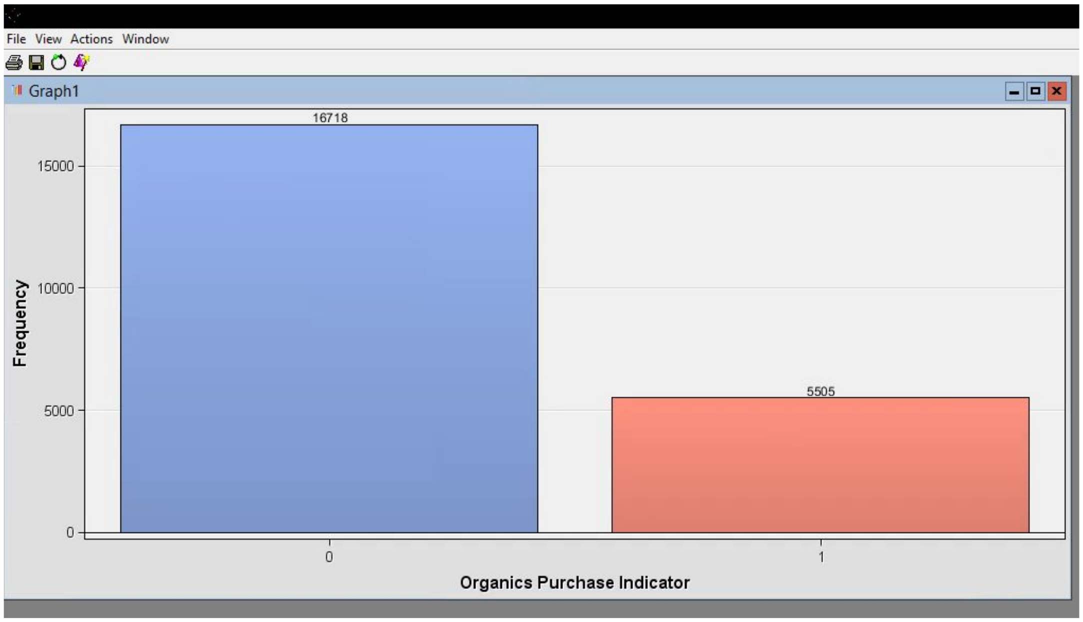
Part 2 -Predictive Modeling_Decision Tree : Task 3

Data Source: ORGANICS

1) Set the roles for the analysis variables as shown above.

Name	Role	Level
DemAffl	**Input**	**Interval**
DemAge	**Input**	**Interval**
DemCluster	**Rejected**	**Nominal**
DemClusterGrou	**Input**	**Nominal**
DemGender	**Input**	**Nominal**
DemReg	**Input**	**Nominal**
DemTVReg	**Input**	**Nominal**
ID	**ID**	**Nominal**
PromClass	**Input**	**Nominal**
PromSpend	**Input**	**Interval**
PromTime	**Input**	**Interval**
TargetAmt	**Rejected**	**Interval**
TargetBuy	**Target**	**Binary**

2) Examine the distribution of the target variable. What is the proportion of individuals who purchased organic products?



Number of individuals who purchased organic products: 5505
Number of individuals who did not purchased organic products : 16718

% of individuals who purchased organic products= 5505/(5505+16718)= 24.77%

3) Set the model role for DemCluster to Rejected.

Name	Role	Level
DemAffl	Input	Interval
DemAge	Input	Interval
DemCluster	Rejected	Nominal
DemClusterGrou	Input	Nominal

4) Can TargetAmt be used as an input for a model that is used to predict TargetBuy? Why or why not?

Code to find correlation between TARGETAMT and TARGETBUY.

Training Code

```
ods graphics on;
proc corr data=&em_import_data;
var targetbuy targetamt;
run;
ods graphics off;
```

Result:

```
The CORR Procedure

   2  Variables:    TargetBuy TargetAmt


                                                Simple Statistics

Variable           N          Mean       Std Dev           Sum       Minimum       Maximum    Label

TargetBuy      22223       0.24772       0.43170          5505             0       1.00000    Organics Purchase Indicator
TargetAmt      22223       0.29474       0.56283          6550             0       3.00000    Organics Purchase Count


     Pearson Correlation Coefficients, N = 22223
              Prob > |r| under H0: Rho=0

                                   Target        Target
                                      Buy           Amt

TargetBuy                         1.00000       0.91261
Organics Purchase Indicator                      <.0001

TargetAmt                         0.91261       1.00000
Organics Purchase Count            <.0001
```

Pearson correlation coefficient value between the variables: 0.91261

TargetAmt(Organics Purchase Count) is highly correlated with TargetBuy(Organic Purchase Indicator). Additionally, from the dataset, we have a certain rule :

Organics Purchase Indicator	Organics Purchase Count
0	0
0	0
1	1
1	1
0	0
0	0
0	0
0	0
0	0
1	1
0	0
1	1
0	0
0	0
1	2
0	0
0	0
1	1
1	1
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0
1	1
1	1
0	0
1	1
0	0
0	0
0	0
0	0
0	0
1	1

For TargetAmt=0, TargetBuy=0 and for TargetAmt=1, TargetBuy=1. Therefore, TargetAmt can be used as input to predict TargetBuy.

5)

f) Create a decision tree model autonomously. Use average square error as the model assessment statistic.

Partitioning the dataset:

Data Set Allocations	
Training	50.0
Validation	50.0
Test	0.0

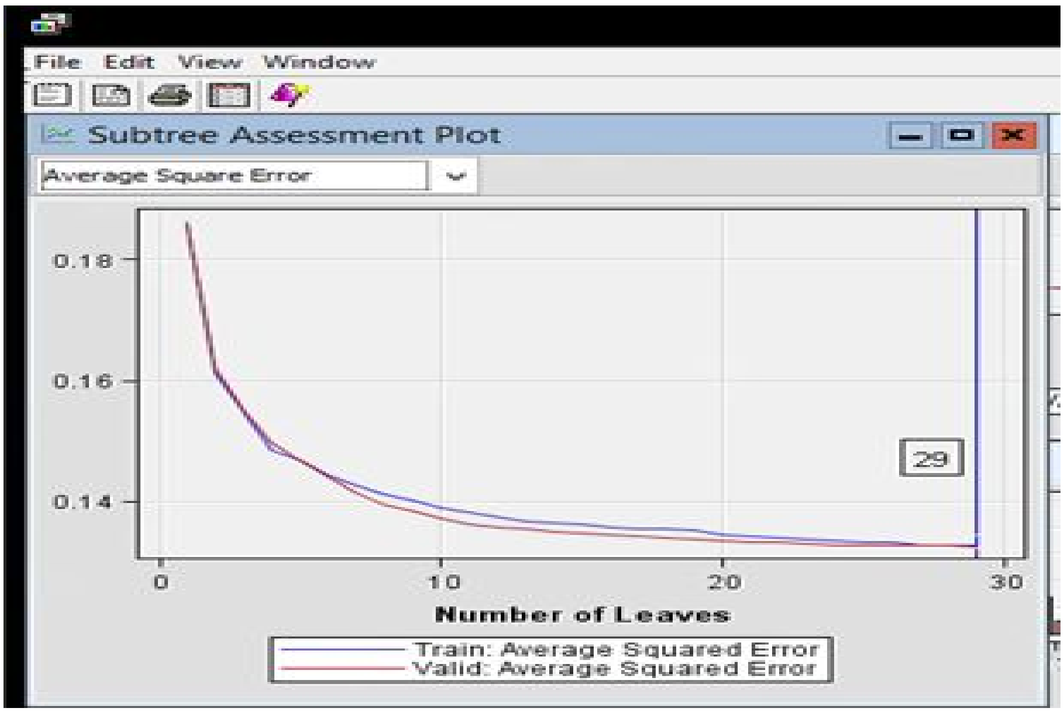
```
Partition Summary

                                         Number of
Type            Data Set                Observations

DATA        EMWS3.Ids2_DATA                22223
TRAIN       EMWS3.Part_TRAIN               11112
VALIDATE    EMWS3.Part_VALIDATE            11111
```

```
Summary Statistics for Class Targets

Data=DATA

             Numeric    Formatted   Frequency
Variable      Value       Value       Count      Percent              Label

TargetBuy       0           0         16718      75.2284    Organics Purchase Indicator
TargetBuy       1           1          5505      24.7716    Organics Purchase Indicator


Data=TRAIN

             Numeric    Formatted   Frequency
Variable      Value       Value       Count      Percent              Label

TargetBuy       0           0          8359      75.2250    Organics Purchase Indicator
TargetBuy       1           1          2753      24.7750    Organics Purchase Indicator


Data=VALIDATE

             Numeric    Formatted   Frequency
Variable      Value       Value       Count      Percent              Label

TargetBuy       0           0          8359      75.2318    Organics Purchase Indicator
TargetBuy       1           1          2752      24.7682    Organics Purchase Indicator
```

For Tree1:

Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Average Square Error
Assessment Fraction	0.25

1) How many leaves are in the optimal tree?



The optimal tree based on average square error has **29 leaves.**

2) Which variable was used for the first split? What were the competing splits for this first split?

Split Node 1

Target Variable: TargetBuy

Variable	Variable Description	-Log(p)	Branches
DemAge	Age	323.299	2
DemAffl	Affluence Grade	200.188	2
DemGender	Gender	133.1391	2
PromSpend	Total Spend	32.9677	2
PromClass	Loyalty Status	23.3334	2

DemAge(Age) was used for the first split.

As we know that higher the logworth value, better the split. Dem Age (Age)has the highest logworth, followed by DemAffl (Affluence Grade) and DemGender(Gender).
So, we can say that **DemAffl and DemGender** were the competing splits for this first split.

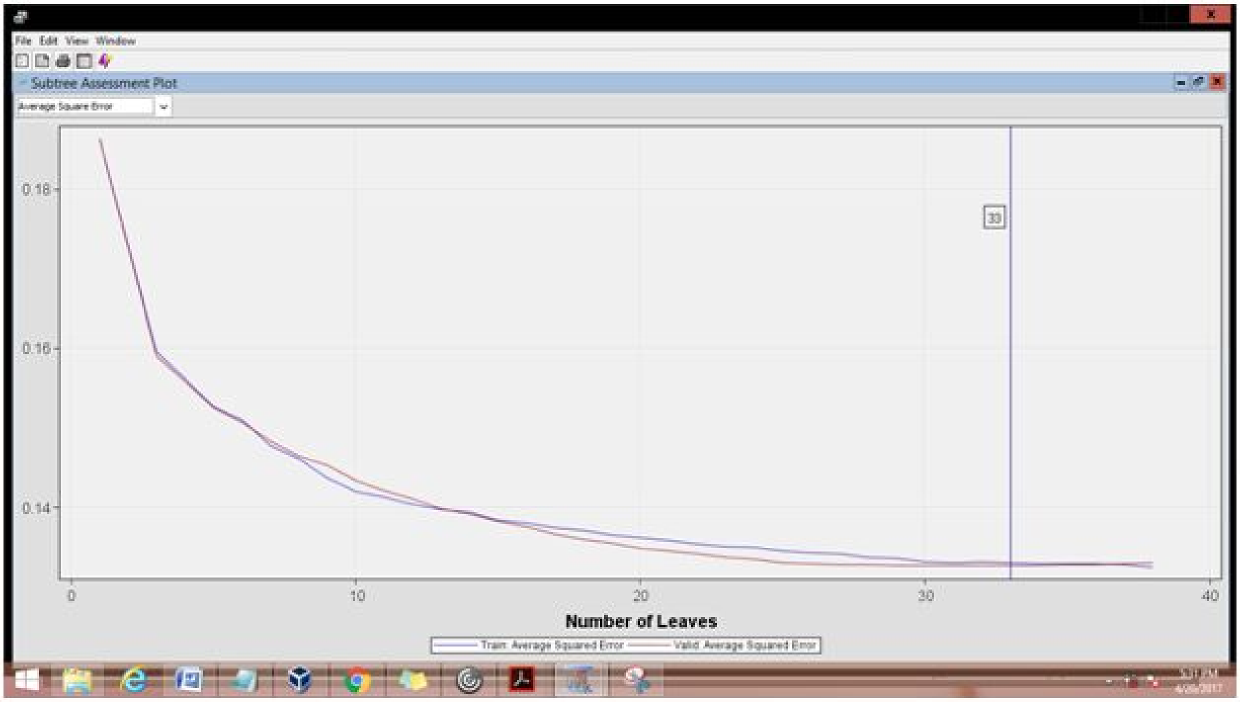
g) Add a second Decision Tree node to the diagram and connect it to the Data Partition node.
1) In the Properties panel of the new Decision Tree node, change the maximum number of branches from a node to 3 to allow for three-way splits.

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

2) Create a decision tree model using average square error as the model assessment statistic.

Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Average Square Error
Assessment Fraction	0.25

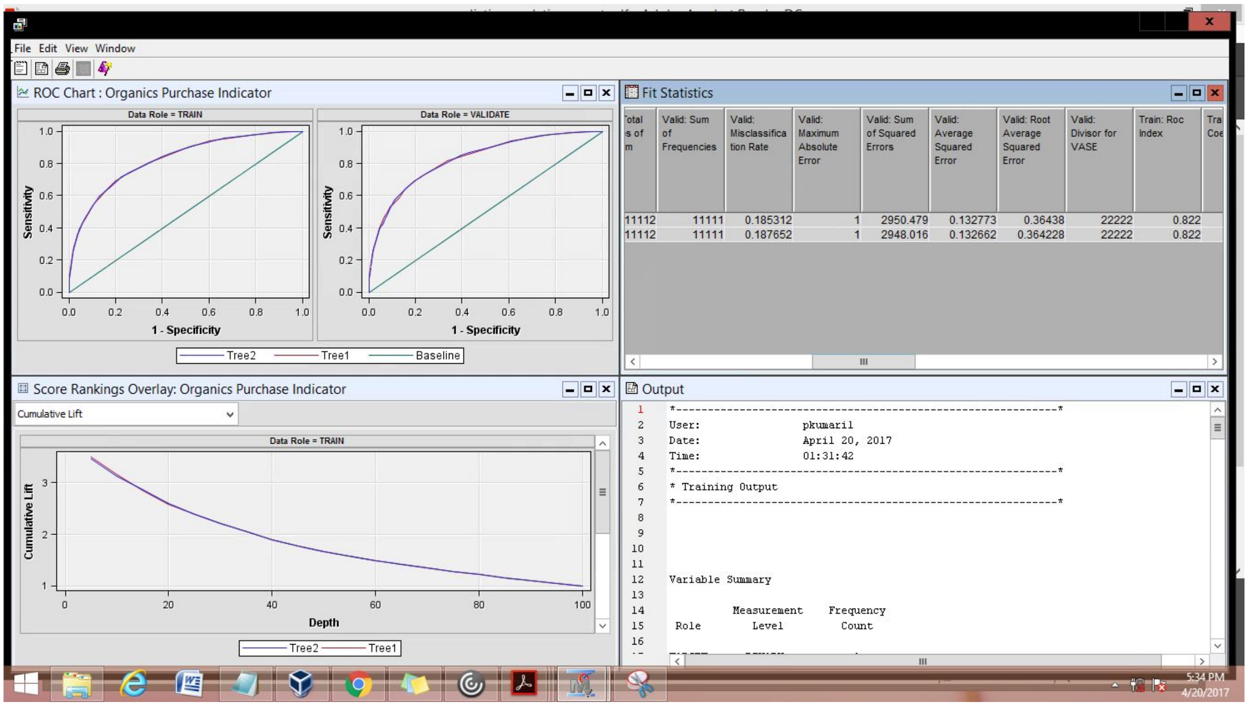
3) How many leaves are in the optimal tree?

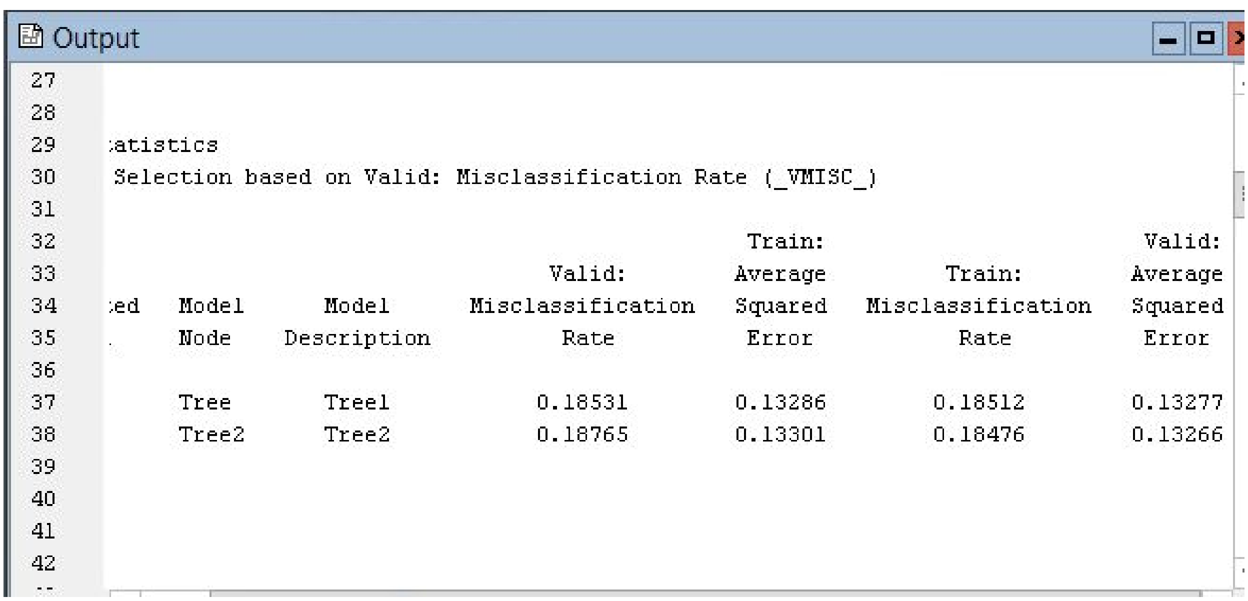


The optimal tree based on average square error has **33 leaves.**

h. Based on average square error, which of the decision tree models appears to be better?

Using Model Comparison we can do this analysis.



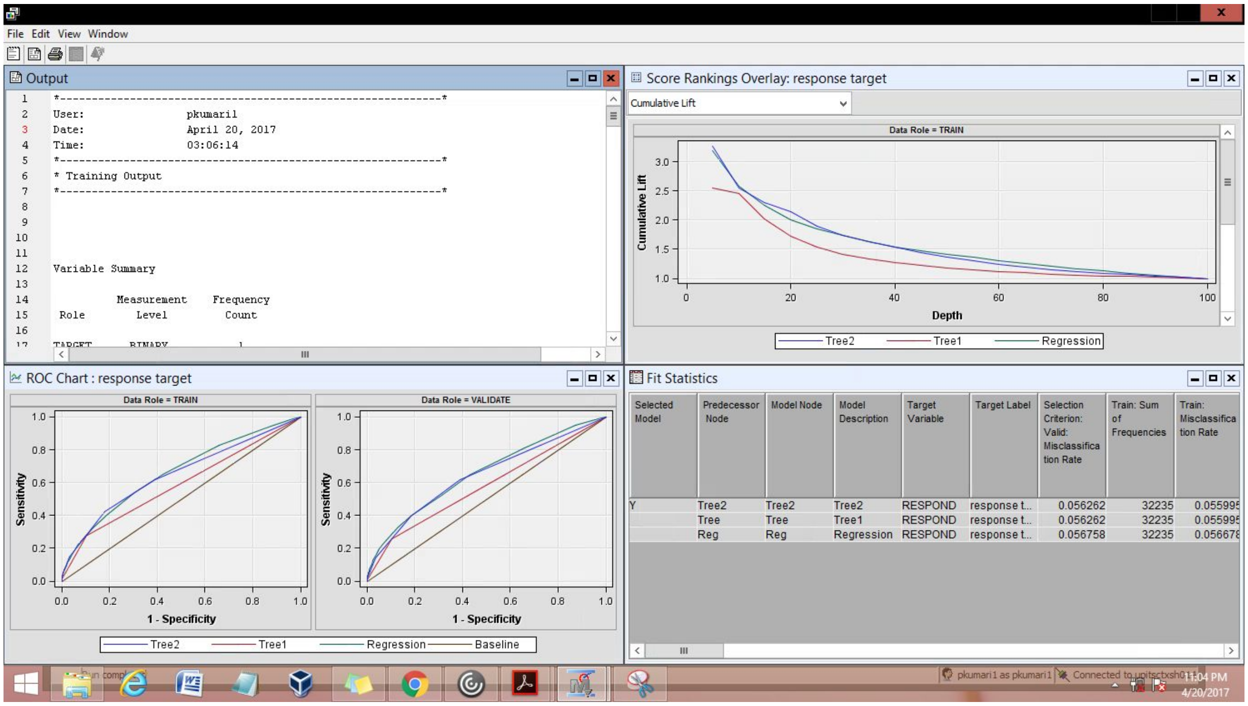


The first tree and the second tree have almost same lift and same ROC statistics. Moreover, average squared error for both the trees is approximately same on validation data. As Tree 2 has higher number of leaves than Tree 1, Tree 2 model may perform better than Tree 1 model.

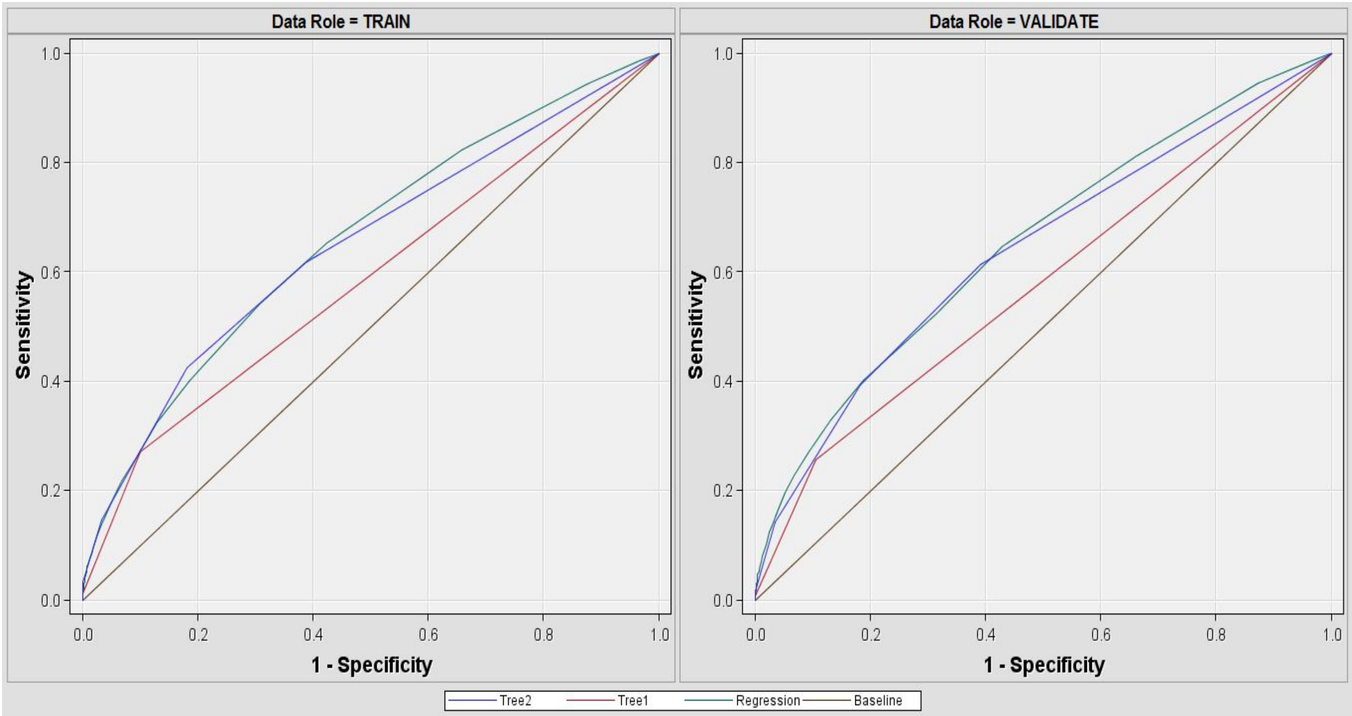
Part 3-Predictive Modeling_LOGIT

Objective:Compare the logistic regression and decision tree models (Tree1 and Tree2)

Result window of Model Comparison node:



ROC chart window:



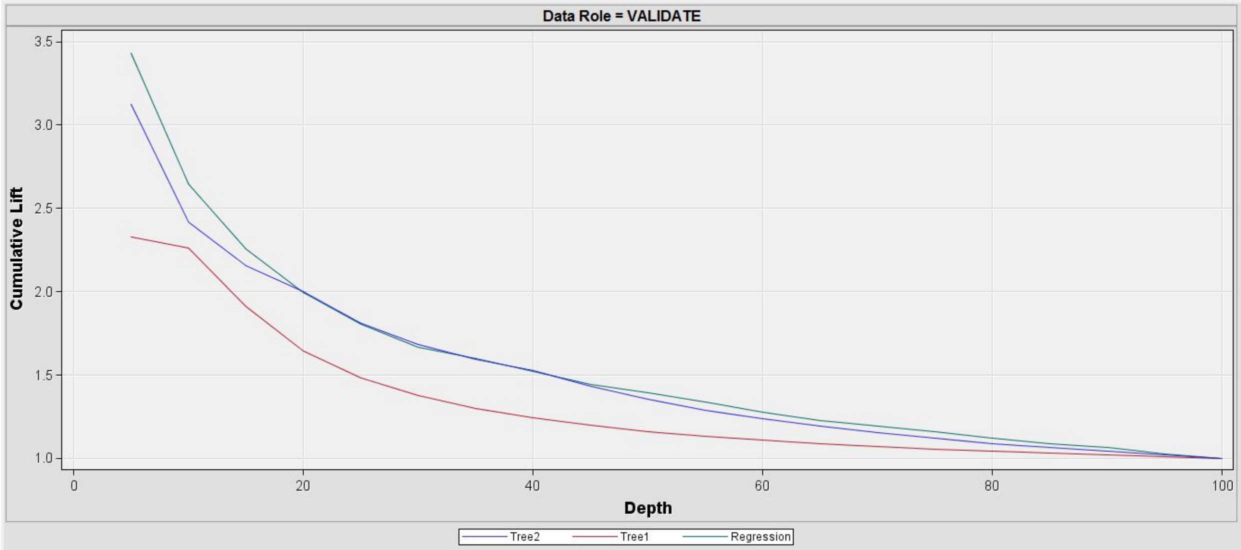
Based on ROC statistics, the logistic regression and Tree2 models perform similarly on the validation data set. Regression model is slightly better than Tree2 model.

Score Ranking Overlay window:

At the 20th percentile, the lift is 1.644 for Tree1 model and 1.9962 for Regression model on the validation data set. This means that if the catalog company mailed to the top 20 percent of its customers based on the predicted probabilities, then they would obtain approximately 2 times more responders compared to a 20-percent random sample of the customers.

The performance of Tree 2 model and logistic regression model is similar. However, if the catalog company mailed to the top 15 percent of its customers, then regression gives better result than Tree 2 model.

At 15th percentile,
Tree 2: Cumulative Lift ->2.155
Logistic Regression: Cumulative Lift-> 2.253



Analysis Goal: The mail-order catalog retailer wants to save money on mailing and increase revenue by targeting mailed catalogs to customers who are most likely to purchase in the future.

Since the retailer wants to target customers who are more likely to make the purchase, the type of prediction here is **'Decision'**. Retailer needs to make decision about whom to mail the catalogs.So, we need to focus on **minimizing misclassification rate and maximizing the Kolmogorov-Smirov statistic**.

From the output window,

Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Tree2	Tree2	0.056262	0.051392	0.055995	0.051990
	Tree	Tree1	0.056262	0.052118	0.055995	0.052515
	Reg	Regression	0.056758	0.051868	0.056678	0.051942

For Validation dataset, misclassification rate for regression model and Tree 2 model is similar but both performs better than Tree1 model. The result is favored by looking at the average squared rate. Low average squared rate (ASE) suggests good model. ASE for Tree2 and Regression models are same and lower than Tree 1 model.

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

In conclusion, looking at the summary table above, we can say that Tree 2 and Regression model are better than Tree1 model. The performance of Tree 2 model and Logistic Regression model is similar for decision making.**Therefore, either Tree 2 model or Regression model can be used.**