Cluster Analysis Assignment

Data Mining

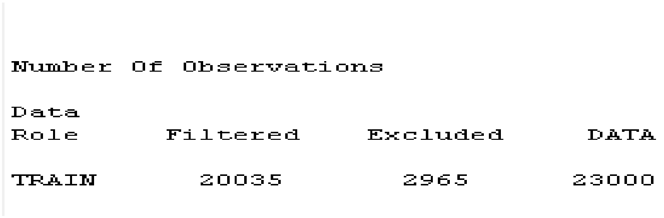
AMBA

Mohamed AlaaElDin Mohamed Awad

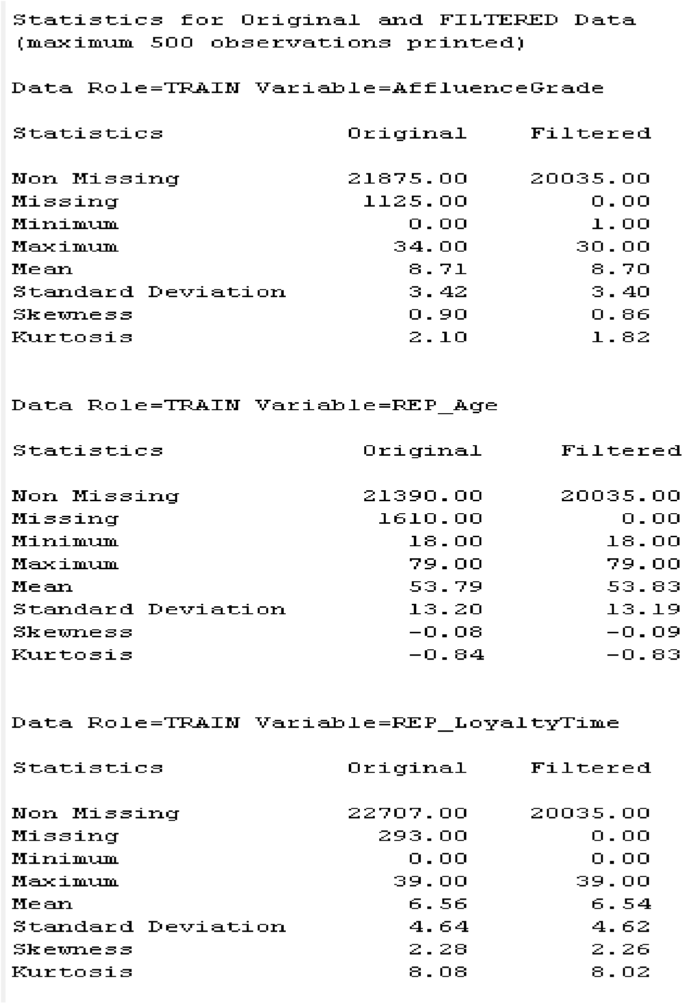
ID: 5910215012

Clustering Analysis (Super Market Campaign)

To cluster the customers into segments, it was necessary first to filter the data from any outliers.



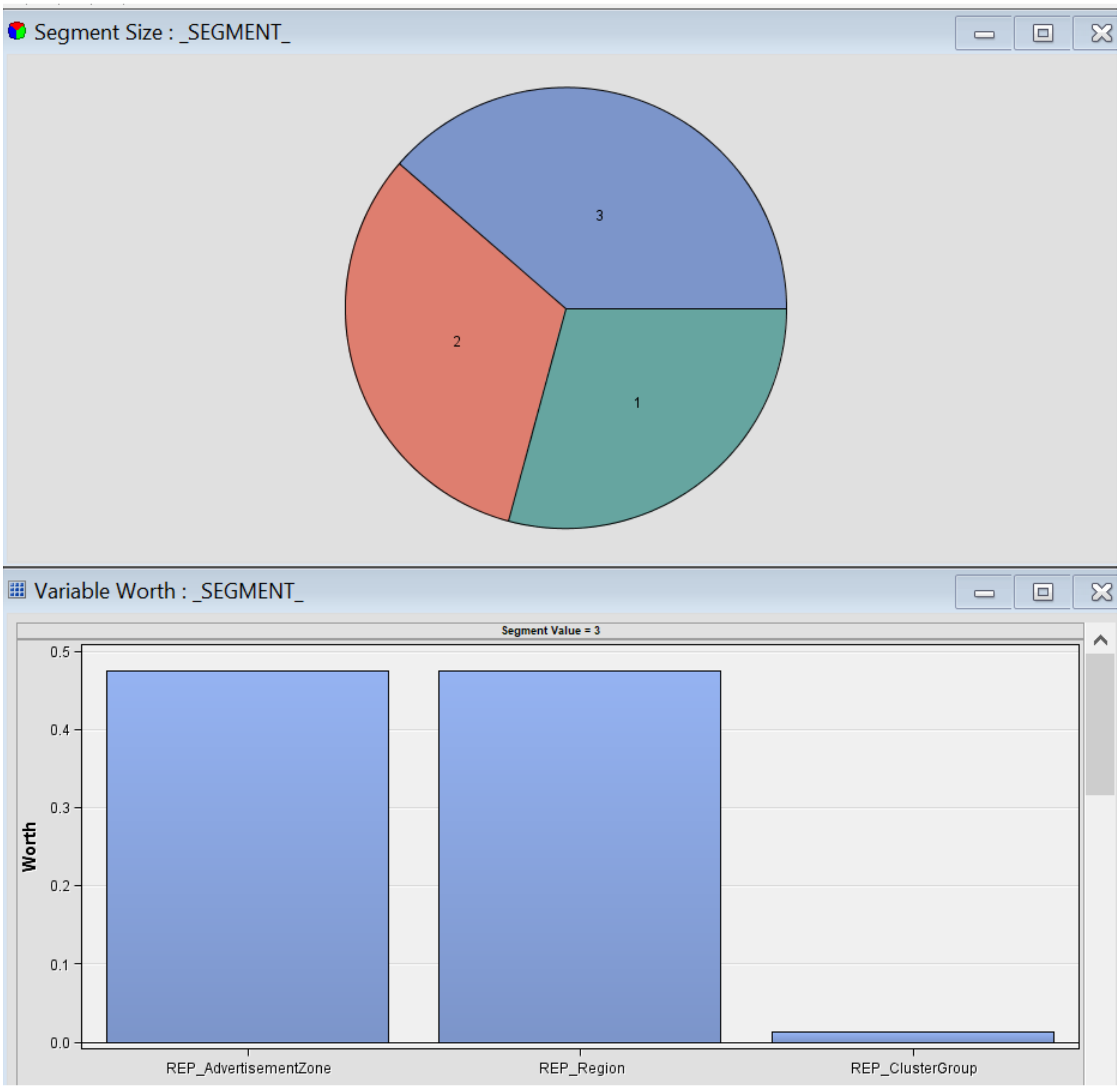
```
Number Of Observations

Data
Role      Filtered     Excluded      DATA

TRAIN       20035        2965       23000
```

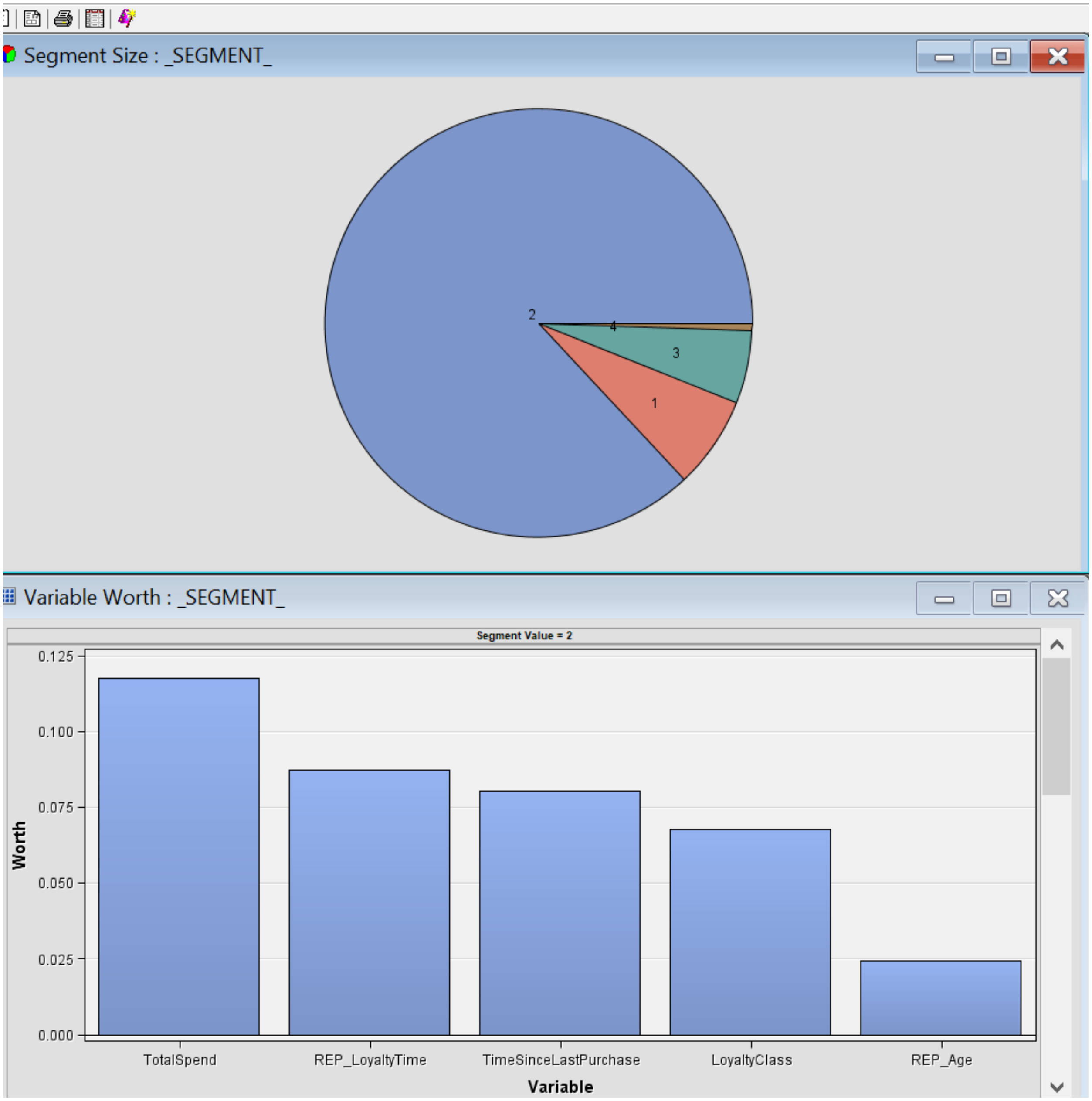


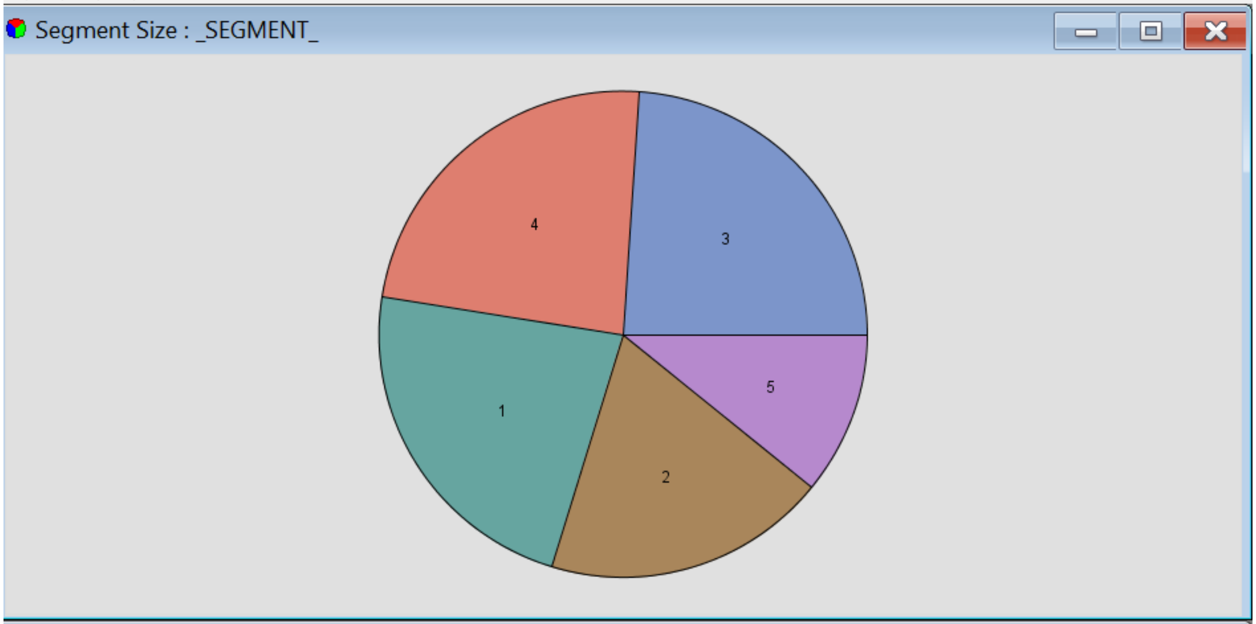
```
Statistics for Original and FILTERED Data
(maximum 500 observations printed)

Data Role=TRAIN Variable=AffluenceGrade

Statistics             Original    Filtered

Non Missing            21875.00    20035.00
Missing                 1125.00        0.00
Minimum                    0.00        1.00
Maximum                   34.00       30.00
Mean                       8.71        8.70
Standard Deviation         3.42        3.40
Skewness                   0.90        0.86
Kurtosis                   2.10        1.82


Data Role=TRAIN Variable=REP_Age

Statistics               Original      Filtered

Non Missing              21390.00      20035.00
Missing                   1610.00          0.00
Minimum                     18.00         18.00
Maximum                     79.00         79.00
Mean                        53.79         53.83
Standard Deviation          13.20         13.19
Skewness                    -0.08         -0.09
Kurtosis                    -0.84         -0.83


Data Role=TRAIN Variable=REP_LoyaltyTime

Statistics             Original    Filtered

Non Missing            22707.00    20035.00
Missing                  293.00        0.00
Minimum                    0.00        0.00
Maximum                   39.00       39.00
Mean                       6.56        6.54
Standard Deviation         4.64        4.62
Skewness                   2.28        2.26
Kurtosis                   8.08        8.02
```

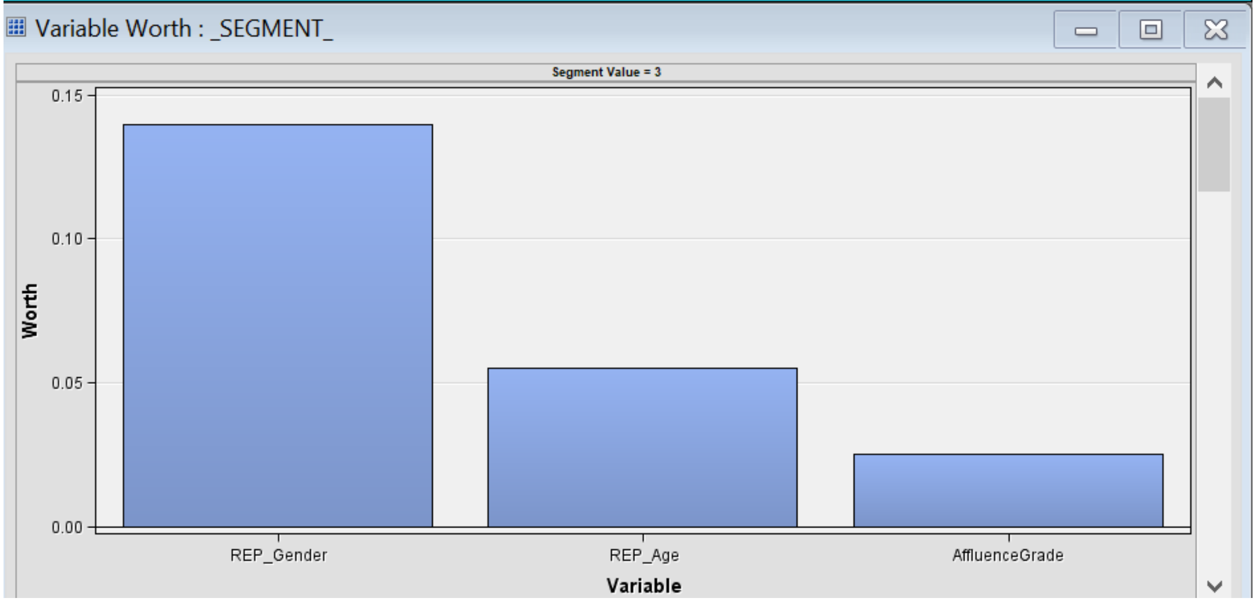
A number of different clusters were tried to find the model that better represents the data and later on building a predictive model with the lowest possible misclassification rate.

To achieve that, different numbers of clusters were tried: 3 clusters, 4 clusters, and 5 clusters. After that the node “Segment profile” was executed for each of the clusters to have an idea about who are the customers in each cluster.

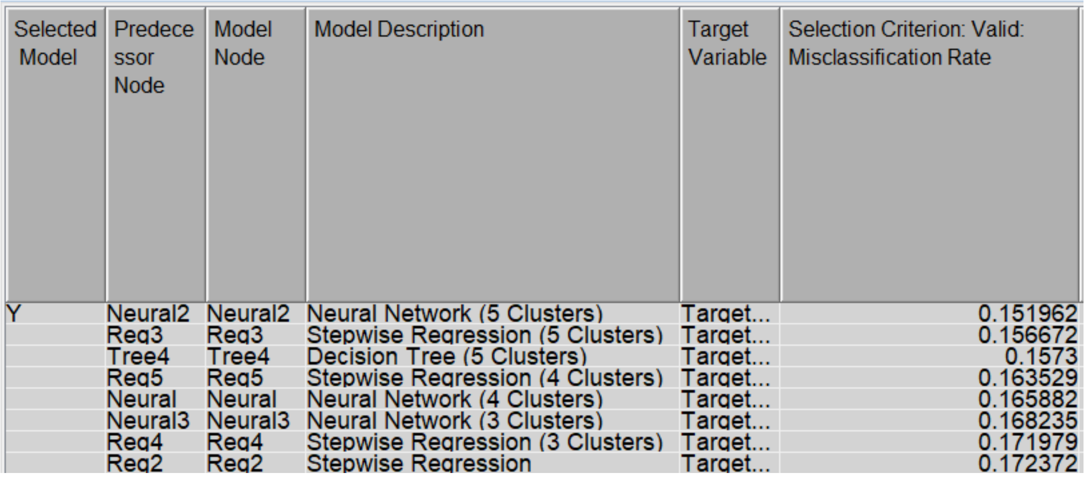



Segment Size : _SEGMENT_
2
4
3
1
Variable Worth : _SEGMENT_
Segment Value = 2
Worth
0.125
0.100
0.075
0.050
0.025
0.000
TotalSpend
REP_LoyaltyTime
TimeSinceLastPurchase
LoyaltyClass
REP_Age
Variable


Segment Size : _SEGMENT_
4
3
1
5
2


Variable Worth : _SEGMENT_
Segment Value = 3
Worth
0.15
0.10
0.05
0.00
REP_Gender
REP_Age
AffluenceGrade
Variable

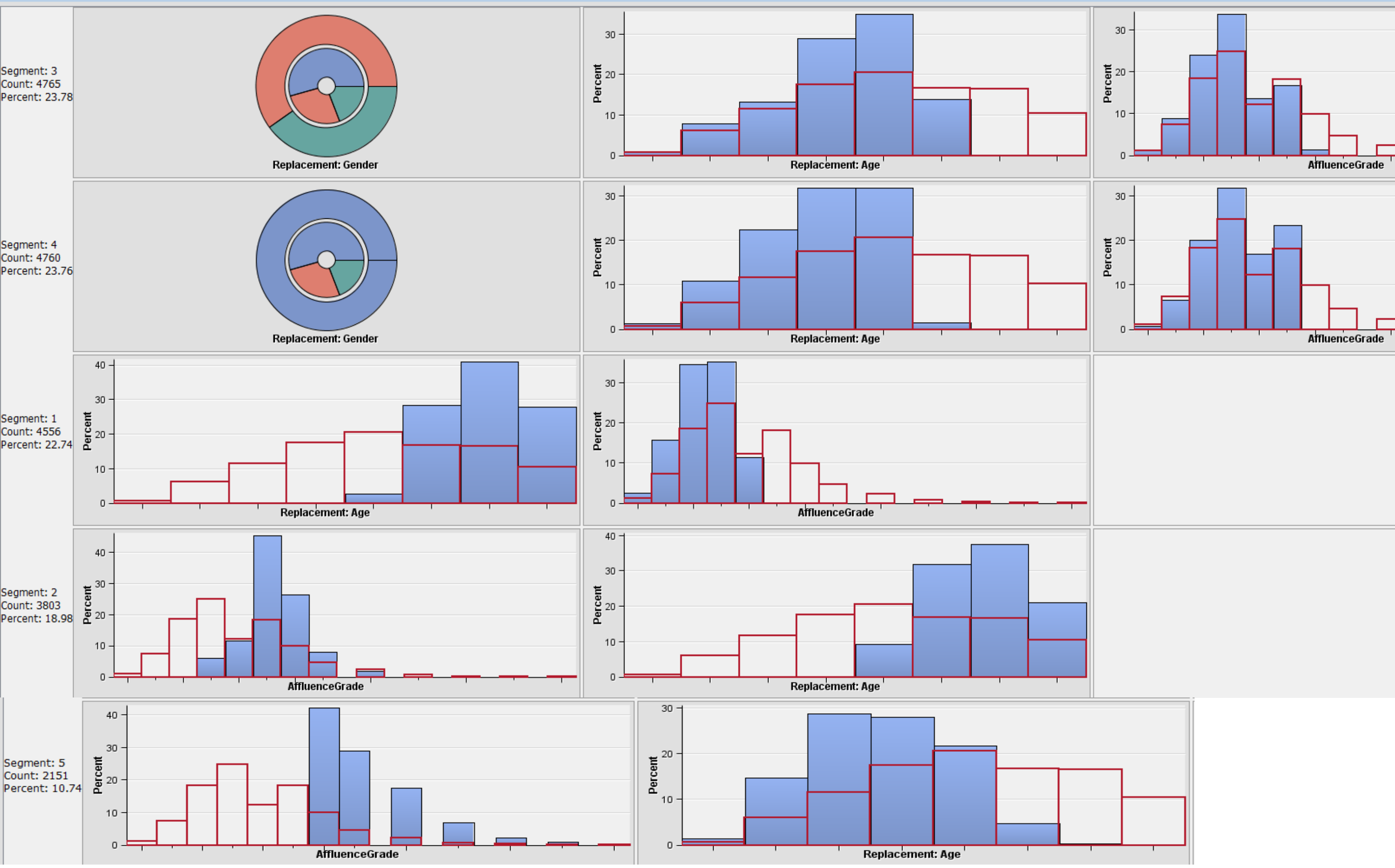
After running predictive modeling for each cluster, it appeared that the model with 5 clusters achieved the lowest misclassification rate of all models.



Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Selection Criterion: Valid: Misclassification Rate
Y	Neural2	Neural2	Neural Network (5 Clusters)	Target...	0.151962
	Reg3	Reg3	Stepwise Regression (5 Clusters)	Target...	0.156672
	Tree4	Tree4	Decision Tree (5 Clusters)	Target...	0.1573
	Reg5	Reg5	Stepwise Regression (4 Clusters)	Target...	0.163529
	Neural	Neural	Neural Network (4 Clusters)	Target...	0.165882
	Neural3	Neural3	Neural Network (3 Clusters)	Target...	0.168235
	Reg4	Reg4	Stepwise Regression (3 Clusters)	Target...	0.171979
	Reg2	Reg2	Stepwise Regression	Target...	0.172372

For neural network model for the 5 cluster segments, many different numbers of hidden units were tried, the lowest misclassification rate came out from the neural network that contains 14 hidden units, the misclassification rate was 0.151962.

The different segments from the 5 clusters model involve the following variables:



Segment 3(Male and working): Contains 4765 customers, represents 23.78 % and can be characterized by Gender(60% are Males), Replacement Age(21-59), Affluence Grade (1.8-12). Those people would need more discounts and sales to become more loyal.

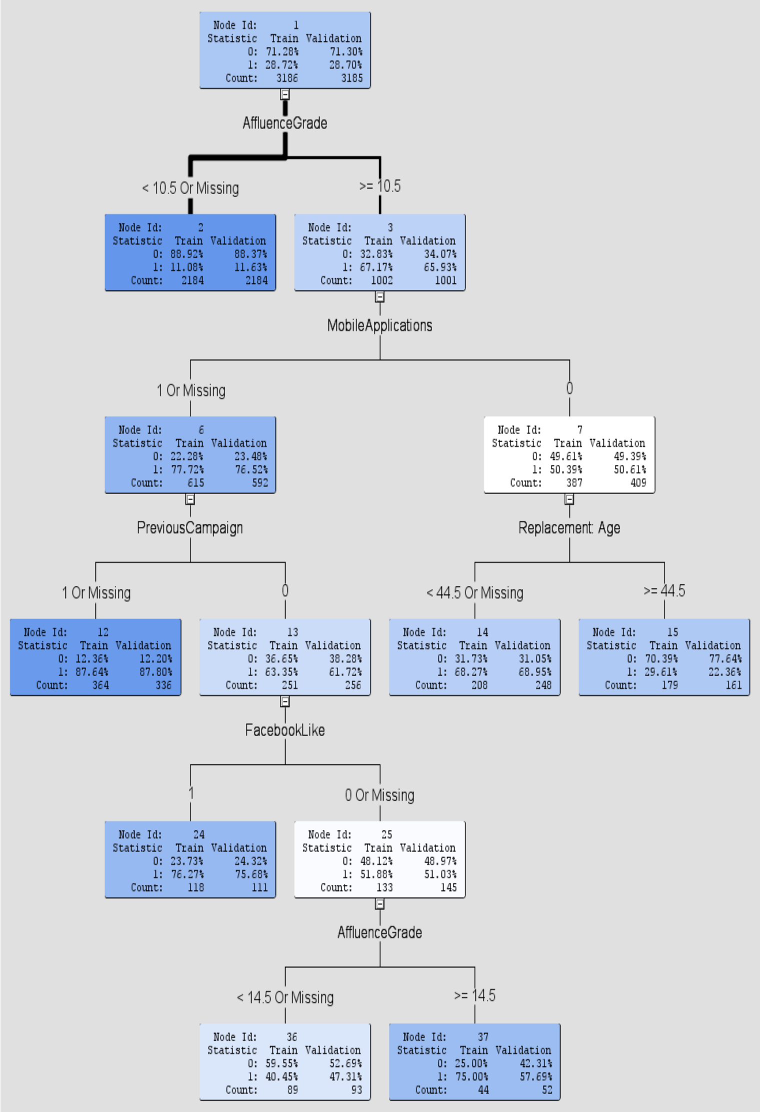
Segment 4(Females and working): Contains 4760 customers, represents 23.76 % and can be characterized by Gender(100% are Females), Replacement Age(21-59), Affluence Grade (1.8-10.6). Those people would need more discounts and sales to become more loyal. Since this group contains females only, they can be targeted by cosmetics products.

Segment 1(Retired and less loyal): Contains 4556 customers, represents 22.74% and can be characterized by age (52-75) and Affluence grade (1.8-8)

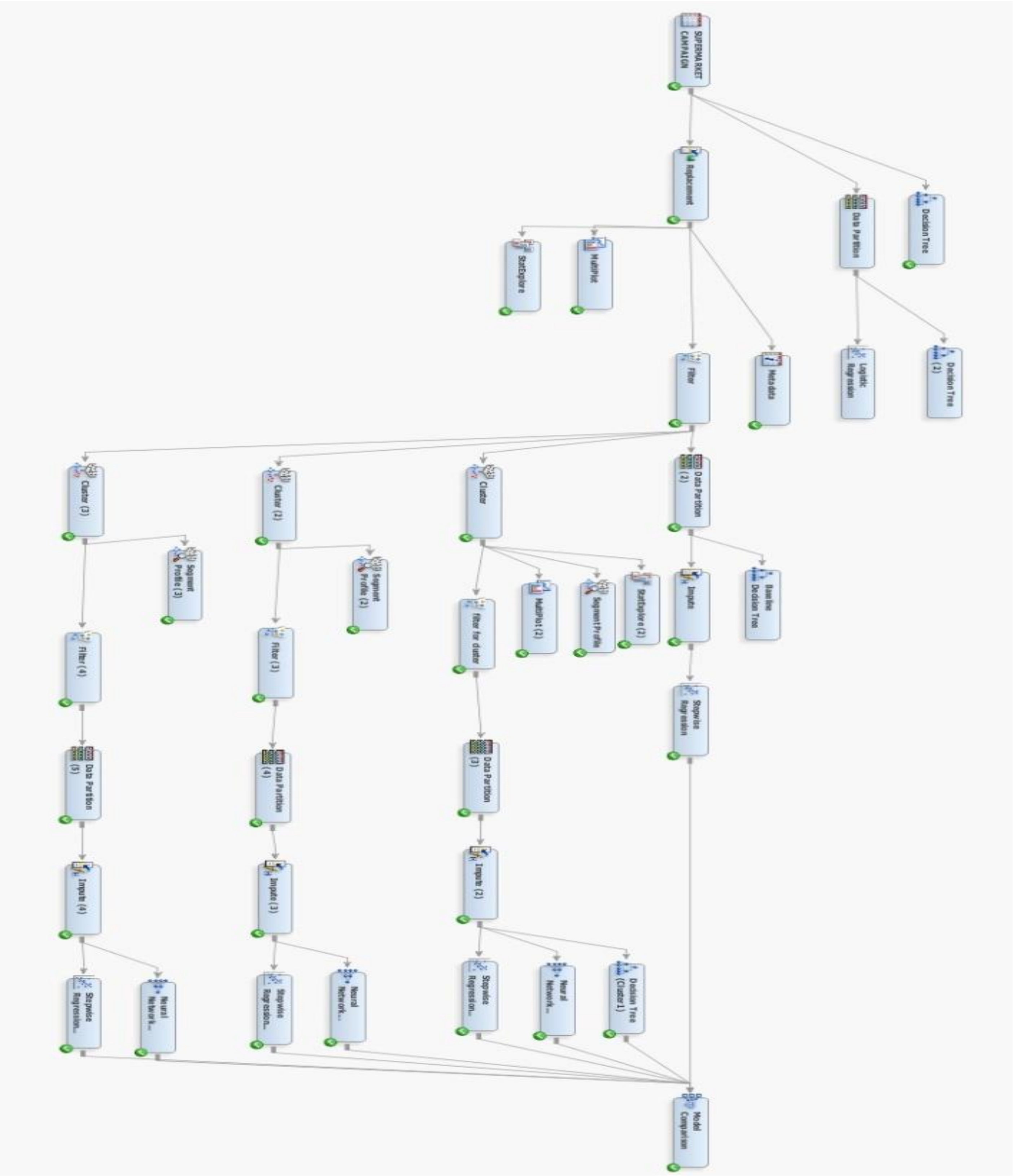
Segment 2 (Retired and more loyal): Contains 3803 customers, represents 19 % and can be characterized by Affluence grade (7-17) and age (52-75) . This segment is transitional, that's, they can become more loyal if you offer them rewards and discounts on certain purchases relative to their age.

Segment 5(Working and the most loyal):): Contains 2152 customers, represents 10.4 % and can be characterized by Affluence grade (12-26) and age (21-59). Since this group contains the most loyal customers, those customers should be retained through various choices such as offering them rewards on their special occasions.

The decision tree for the 5 clusters also show that the affluence grade, having the mobile applications, participation in previous campaign, Facebook likes are important factors in targeting a certain group of customers.



Appendix: Full Diagram


SUPERMARKET CAMPAIGN
Replacement
Data Partition
Decision Tree
StatExplore
MultiPlot
Filter
Metadata
Logistic Regression
Decision Tree (2)
Cluster (3)
Cluster (2)
Cluster
Data Partition (2)
Segment Profile (3)
Segment Profile (2)
MultiPlot (2)
Segment Profile
StatExplore (2)
Impute
Baseline Decision Tree
Filter (4)
Filter (3)
filter for cluster
Stepwise Regression
Data Partition (5)
Data Partition (4)
Data Partition (3)
Impute (4)
Impute (3)
Impute (2)
Stepwise Regression...
Neural Network...
Stepwise Regression...
Neural Network...
Stepwise Regression...
Neural Network...
Decision Tree (Cluster 1)
Model Comparison