**DSCI 5240- DATA MINING**

**FINAL PROJECT**

**GROUP 10**

**Sai Varun Teja Mudumba**

**Shashank Ghanta**

**Pranav Modem**

**Urmila Ponnagandla**

**Gouthami Bokka**

**Executive Summary:**

**Objective:**

Main objective of the project is to apply various data mining techniques to analyze the data and build data models and finding out the parameters that highly influence the energy consumption. For this project we considered the data set on Energy prediction in which we predict the energy consumption based on the input parameters like room temperature, humidity, number of appliances.

**About Dataset:**

The Dataset contains the measurement of temperature, humidity, pressure, air temperature and wind speed which are important parameters in the prediction. The dataset contains 29 columns and more than 10,000 rows and this set doesn’t have any null values.

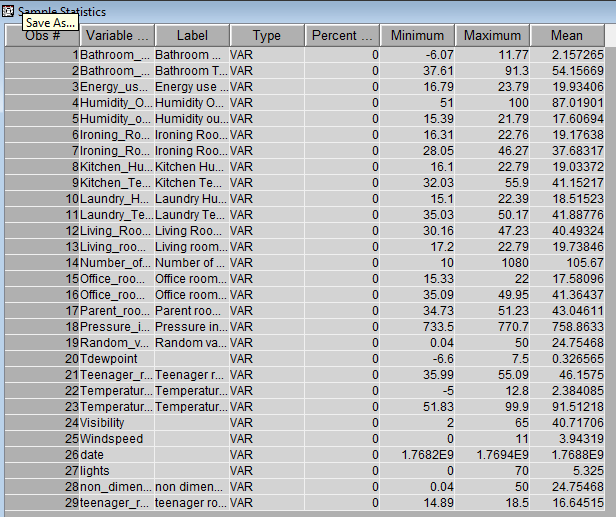
**Data Collection:**

The data in this dataset is information about readings collected for every 10 mins for about 4.5 months. Zigbee wireless sensor network is to record the temperature and humidity levels of each room in the house which sends information every 3.3 seconds. This information is averaged for every 10 minutes. Weather report from the public data set from Reliable Prognosis is downloaded for the surrounding temperature and humidity levels and merged with the recorded data set. We got the dataset from Kaggle website.

**Data Preparation**

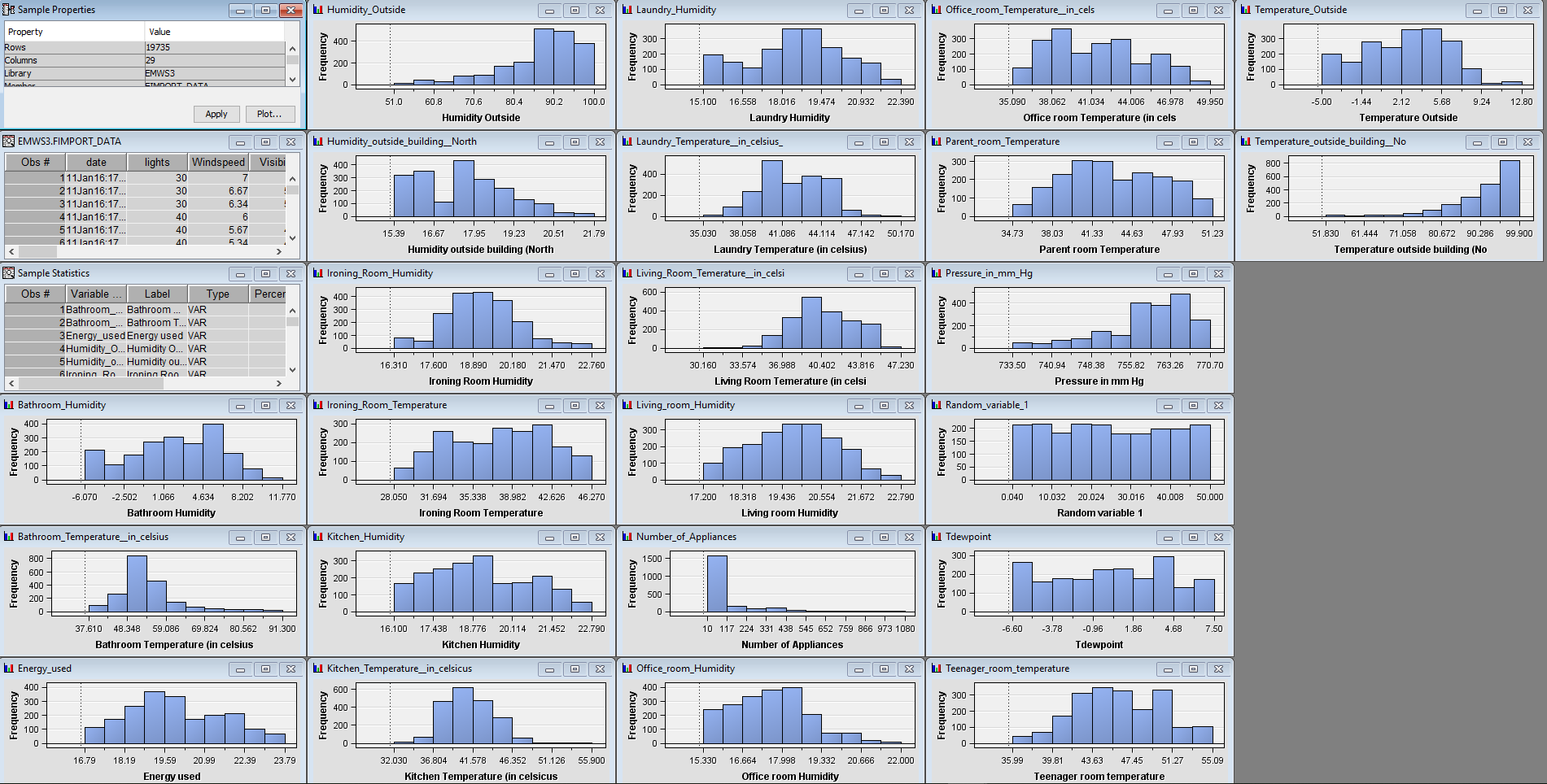
Our data set contains 29 Columns and more than 15000 rows. The data Initially we determined our dependent and independent variables in our data. We concluded that Energy used variable will be our main dependent variable in our data. This variable tells us the energy consumed based on the number of appliances, lights and various factors such as room temperatures and humidity.

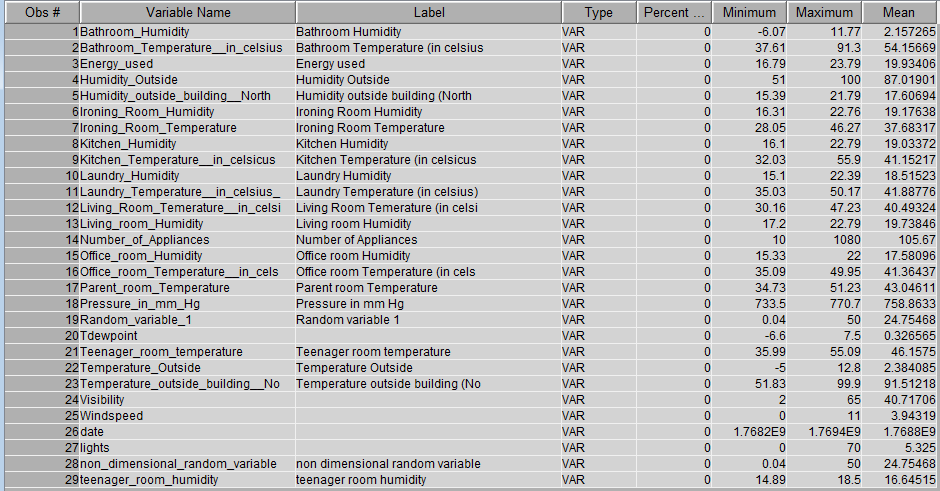
Initially when we uploaded the data to SAS Enterprise Miner, we faced an error related to exceeded maximum target level of 512 due to many unique values in the data set. Our initial data contains decimal values which have more than 3 decimal values due to this the number of unique values increased. So, we used Microsoft excel to eliminate this problem by rounding up the values to 2 decimal values such that we reduce the number of unique rows in our data set. Thus, we can see it in the second figure that the issue of exceeding levels is sorted out with the above solution.



**Data Description**

Our dataset contains around 28 independent variables and more than 15000 observations. Most of the independent variables are room temperatures and humidity levels of different rooms and outside temperature, humidity level, visibility and wind speed. As mentioned in the Data preparation many of the observations have unique values due to the larger decimal values. Below Data directory shows the details about the variables in our dataset. The following pic shows us that there are no unusual data as there are no missing values and the data variation of each variable is plotted as shown.

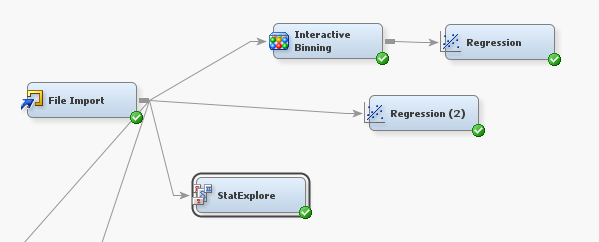


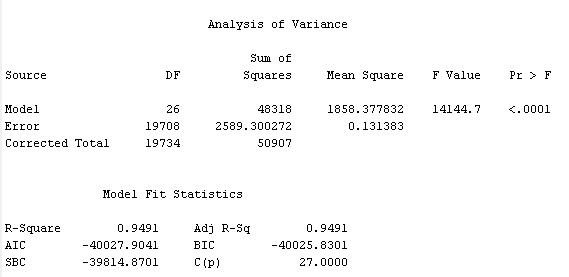


**Models/Enterprise Miner diagrams used**

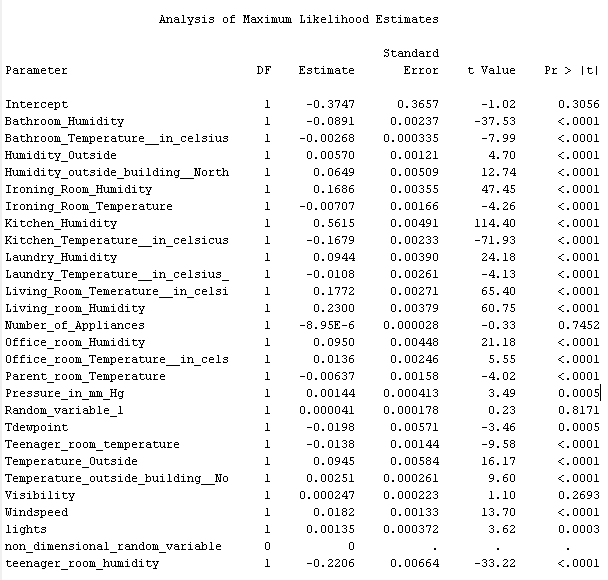
**Linear Regression**

We performed Linear regression on our complete data set to check whether the data model is significant or not and to check the significant variables in the data set. As per our initial regression analysis we found that the model is significant as the Pr > F value is <.0001 and the R-square value is 0.9491 which explains the variance of the analysis.



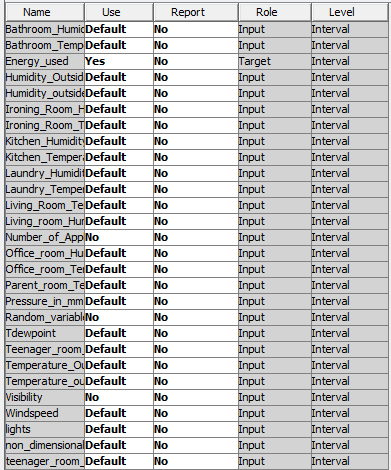


The following table lists out all the parameters used as inputs and states their t value and Pr > |t| value which helps in analyzing the variables which are significant are which are not baased on their pr > |t| value.

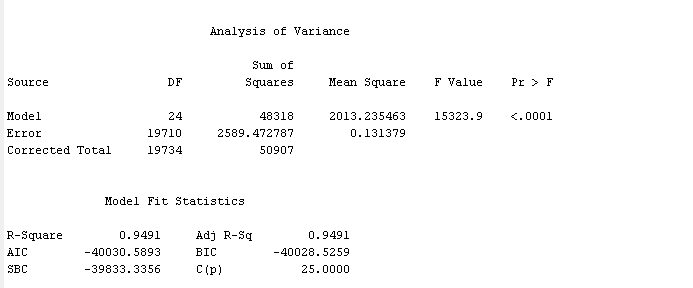


From the above data we found out that the Number\_of\_Appliances, Random\_variable\_1 and Visibility are insignificant as their pr > |t| value is greater than 0.05 which would inflate the R2 value.

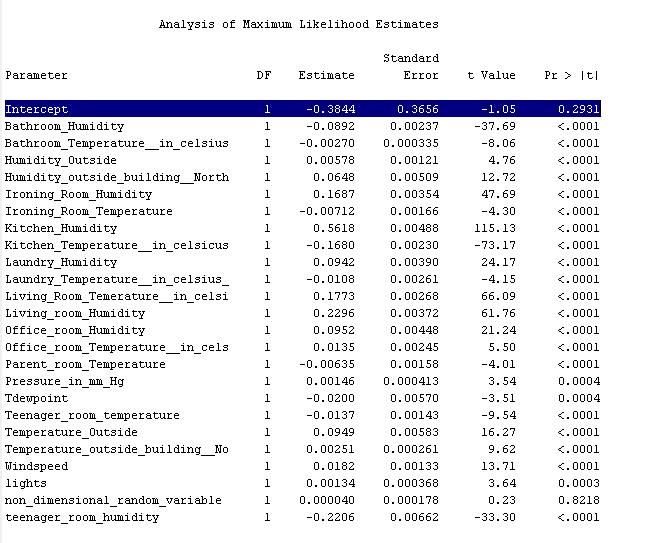
So as a part of further analysis we set the input values of the parameters which has pr > |t| value greater than 0.05 as rejected as shown in the below figure and performed the linear regression model.



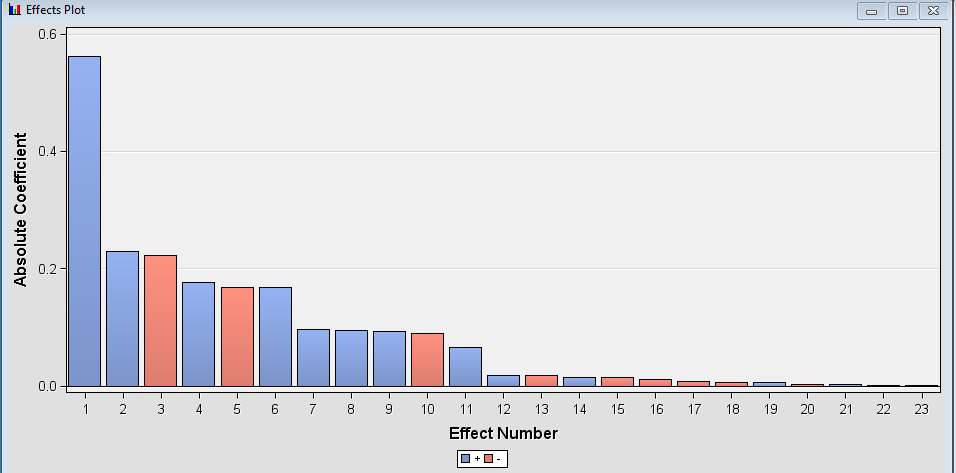
We found that the model is significant as the Pr > F value is <0.0001 and the R2 value is 0.9491 which is same as the previous value.

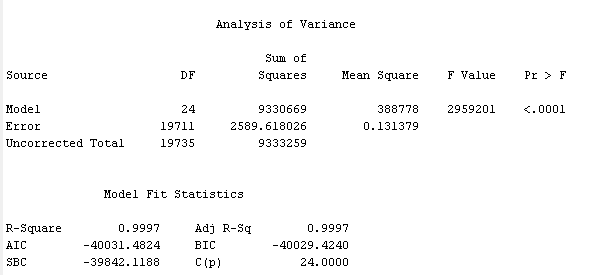


From this test we found the variables which will impact the analysis based on the Pr > |t| value



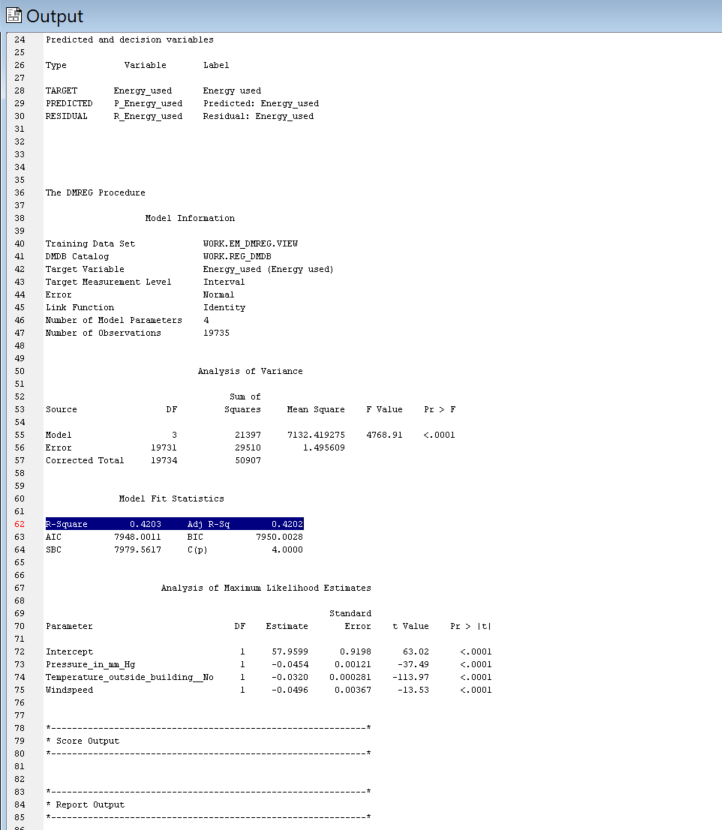
The Absolute Coefficient and Effect Number plot is as shown below:

  
Though the overall models in the above cases are significant we see that the Intercept in the models seem to be insignificant which triggered us with trying out building the regression model by suppressing the Intercept. So, by running the regression with only the significant inputs and by suppressing the Intercept we get the model statistics as follows.



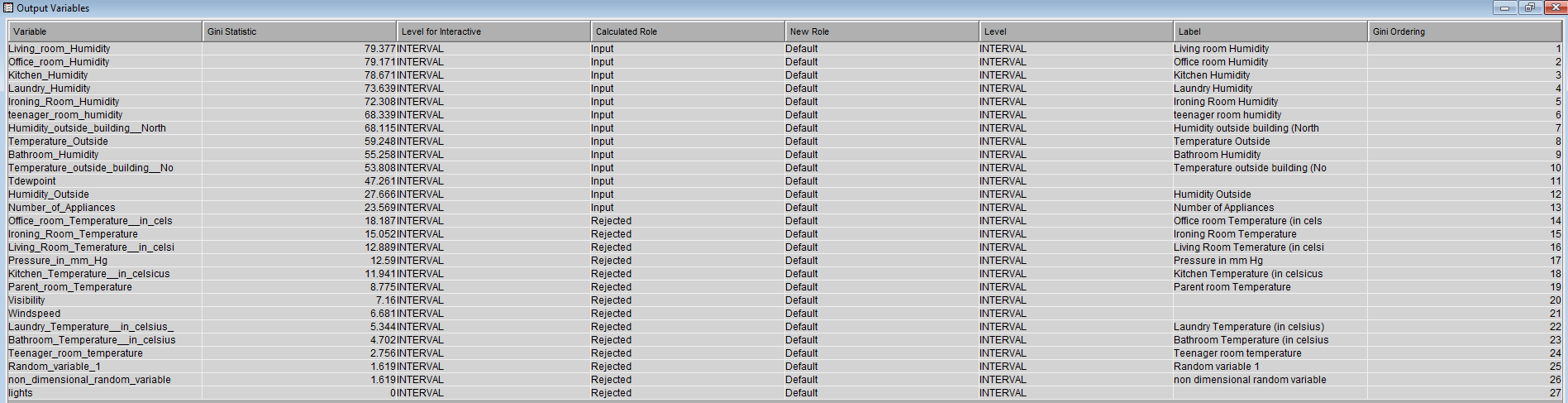
The results show that the model is significant as the Pr> F value is <0.001and it shows that the R2 value is 0.9997 which means that it explains 99.97 % of the variance.

We also found out that variables such as Pressure, Air temperature and Windspeed are important parameters of the prediction and make a significant impact in predicting the target variable. Because when we executed the linear regression only with these 3 parameters as inputs, rejecting all the other inputs, they alone explain 42% of the variance of the model.

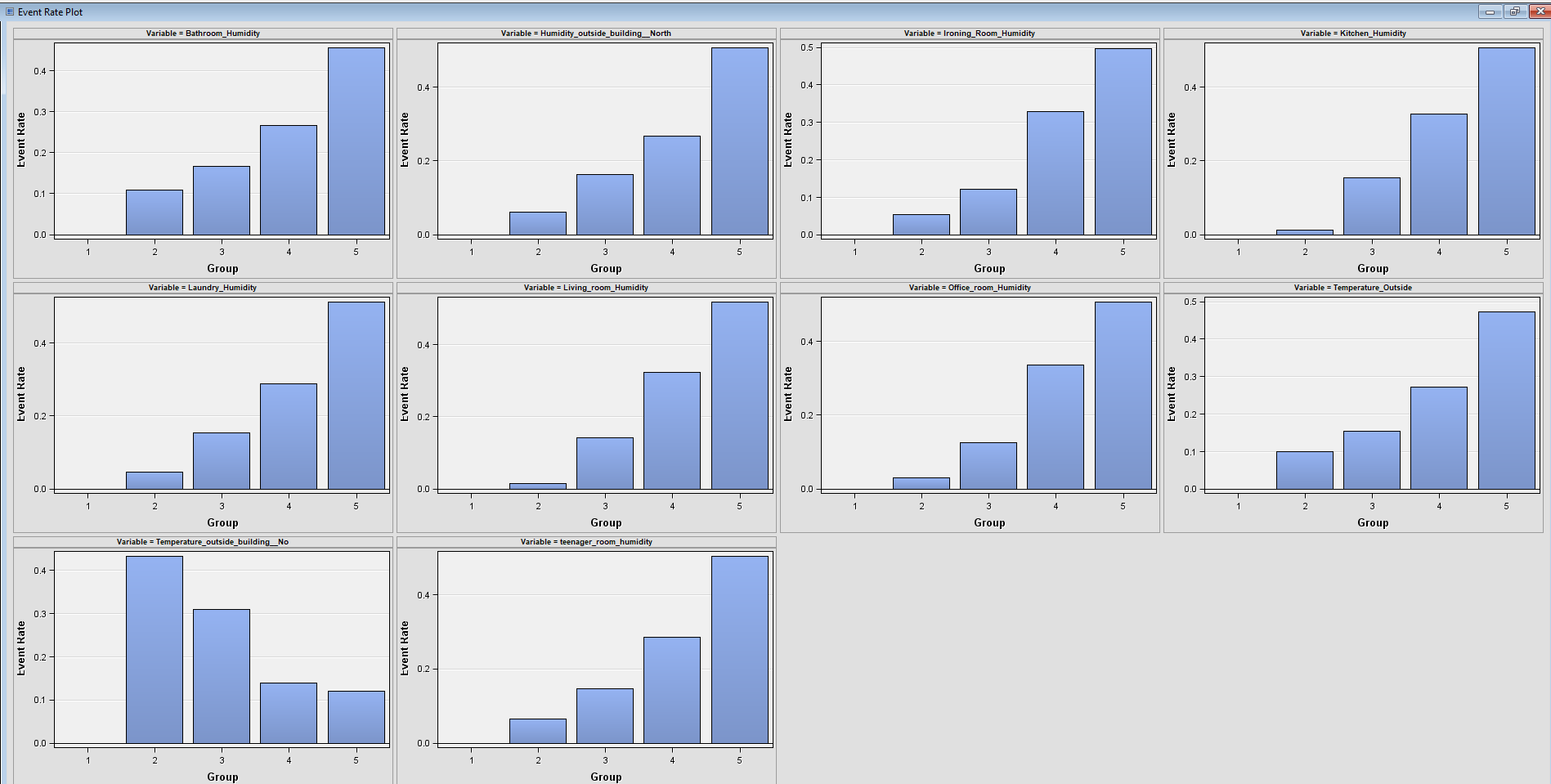


**Linear Regression Using Interactive Binning**

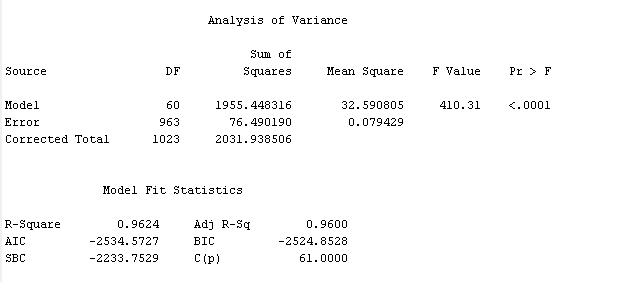
Interactive binning is an inbuilt function in SAS Enterprise miner which Groups variable values in to classes that can be used as inputs for the predictive modeling. Executing the interactive binning the systems automatically rejected some of the input variables based on the Gini statistics and other parameters. So, we wanted to try out this feature and look at the statistics imposed by it in predicting the target variable. The results obtained are as follows



Event rate plot of the input variables are as shown below

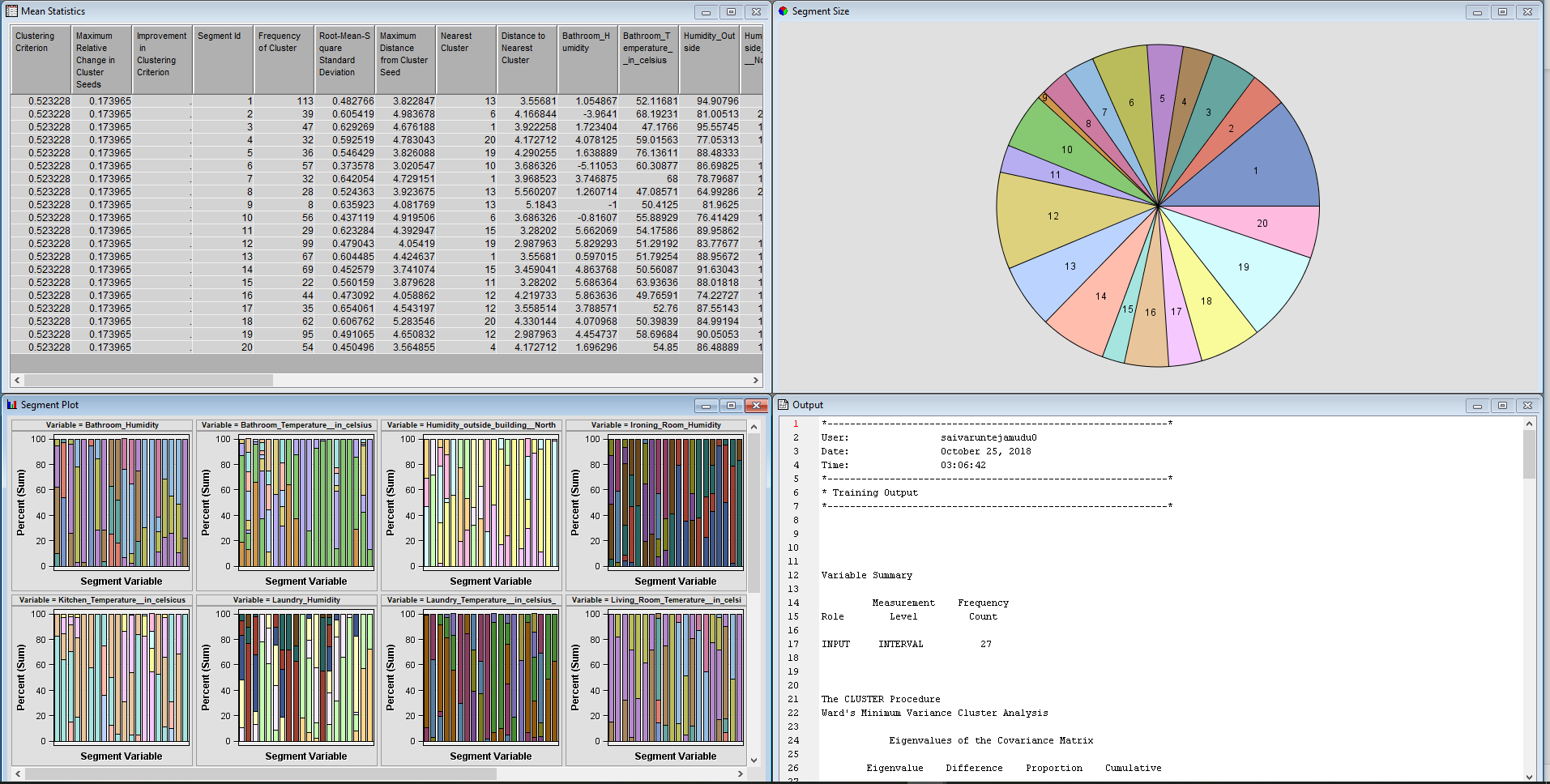
Ii/////

The regression model is significant as the Pr > F value is < 0.0001 and the R-square value is 0.96 which explains the variance of the model.

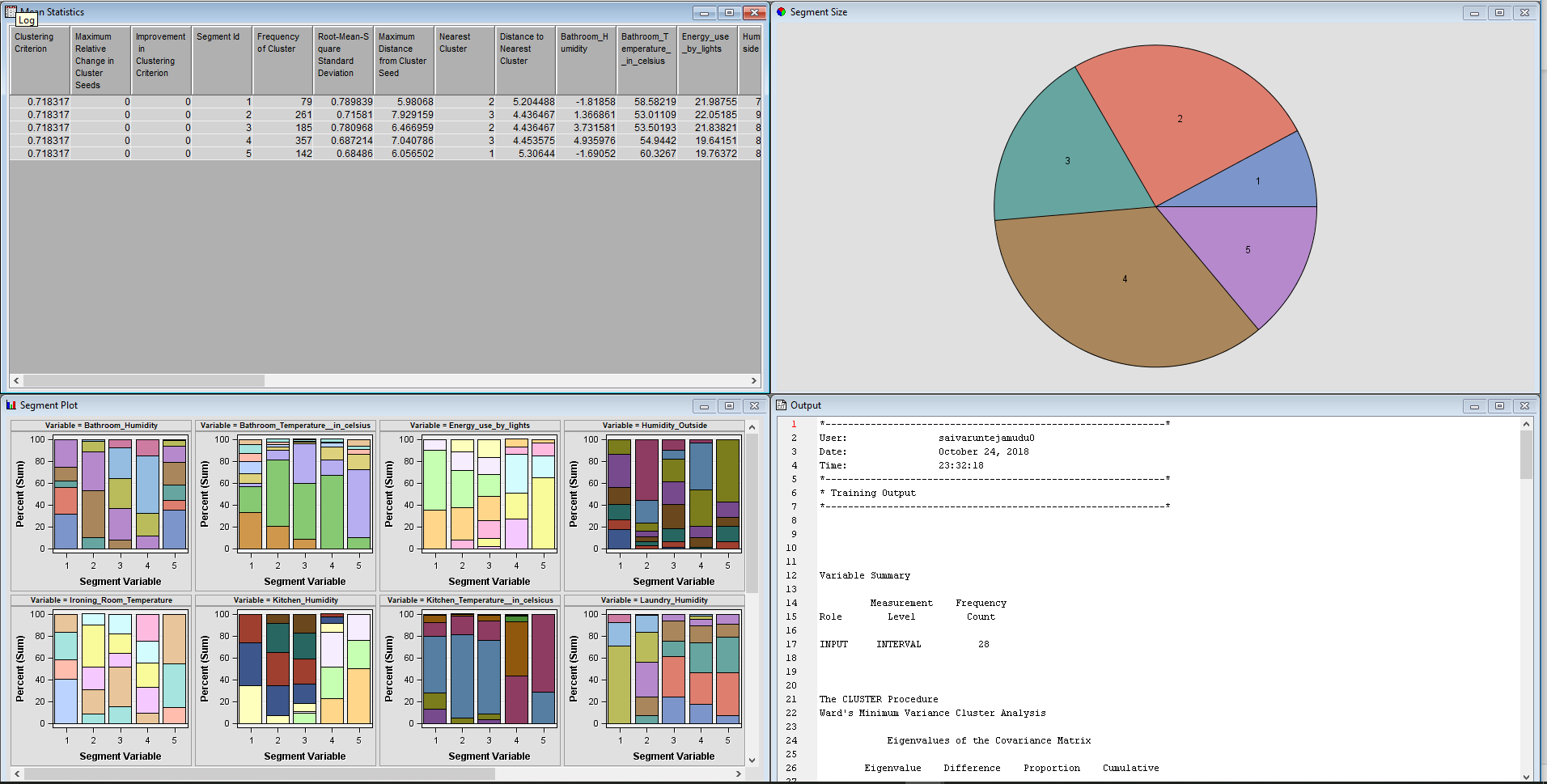


**Cluster Analysis**

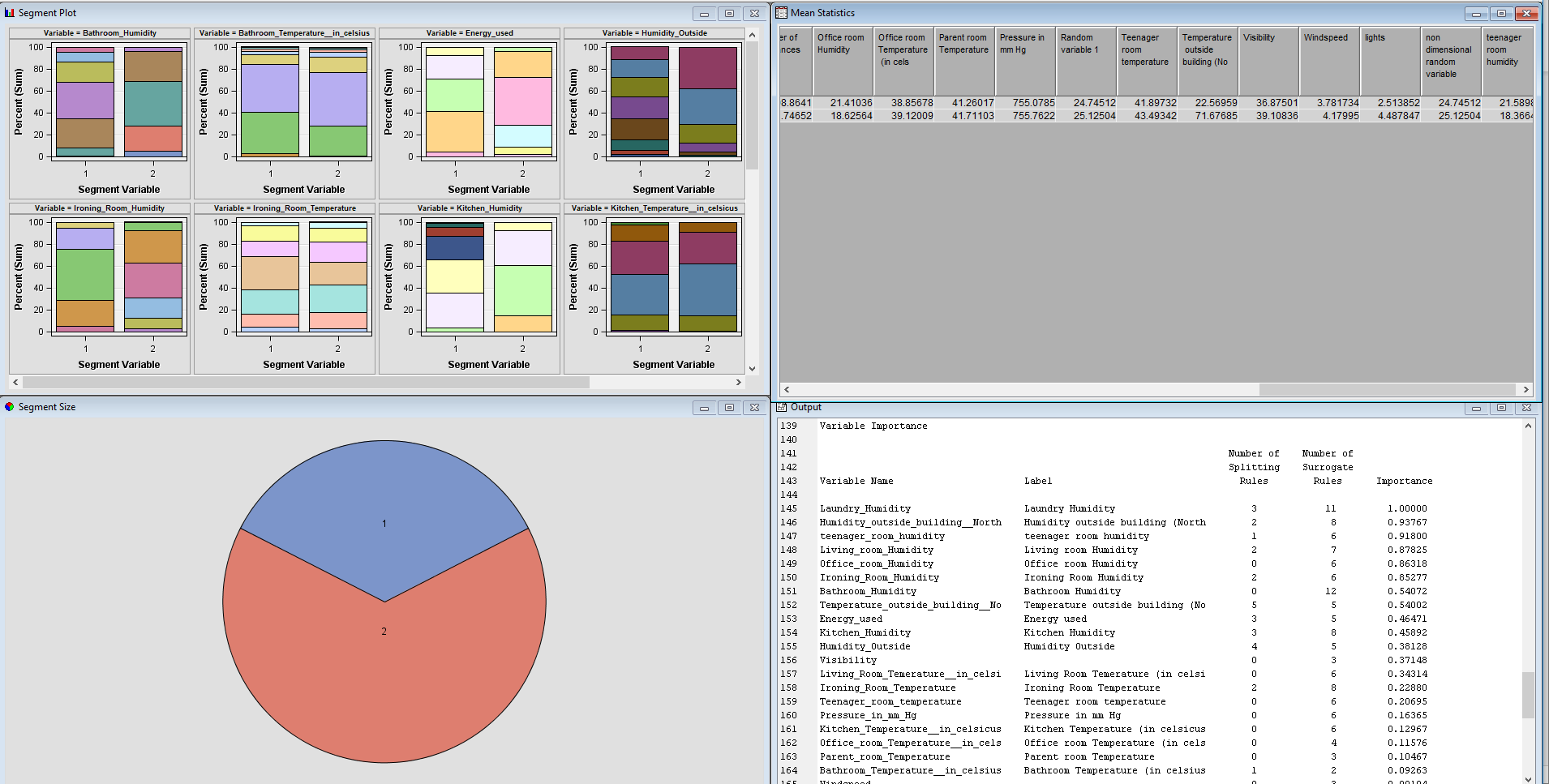
Firstly, we decided to run a cluster analysis on the initial data of 1000 records which initially created 20 clusters which didn’t seem so reasonable in conveying information about the clusters. It is as shown below.



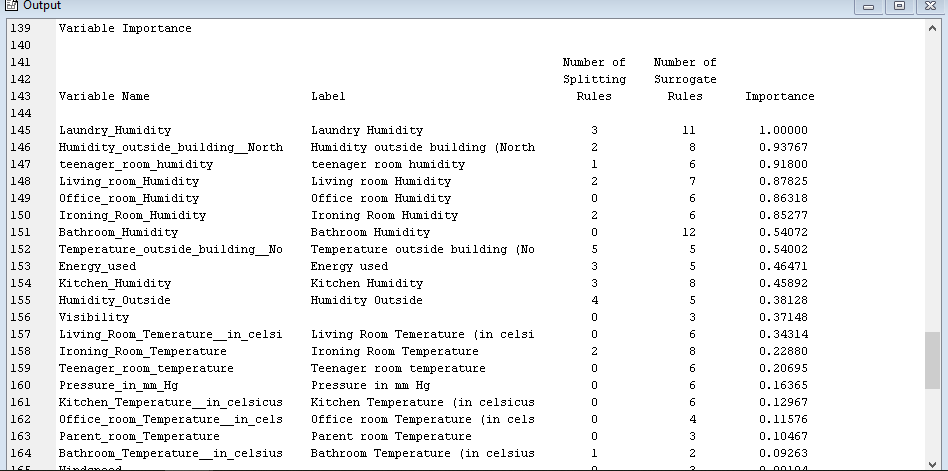
So, we set the maximum clusters to 5 to find out if it could convey any clear picture and as per the analysis it distributed the clusters conveying much more clear picture than the initial plot with the maximum segment size of 357 for segment 4 followed by segment 2 with a frequency of 261 and the segment plot is as shown below.



But lately we decided to perform analysis using all the records using the default methods and automatic cluster numbering. This resulted in the creation of two clusters whose frequencies are in the ratios of 1:2 (i.e., 6858 and 12877) respectively for Clusters 1 and 2. The results window is as shown below.



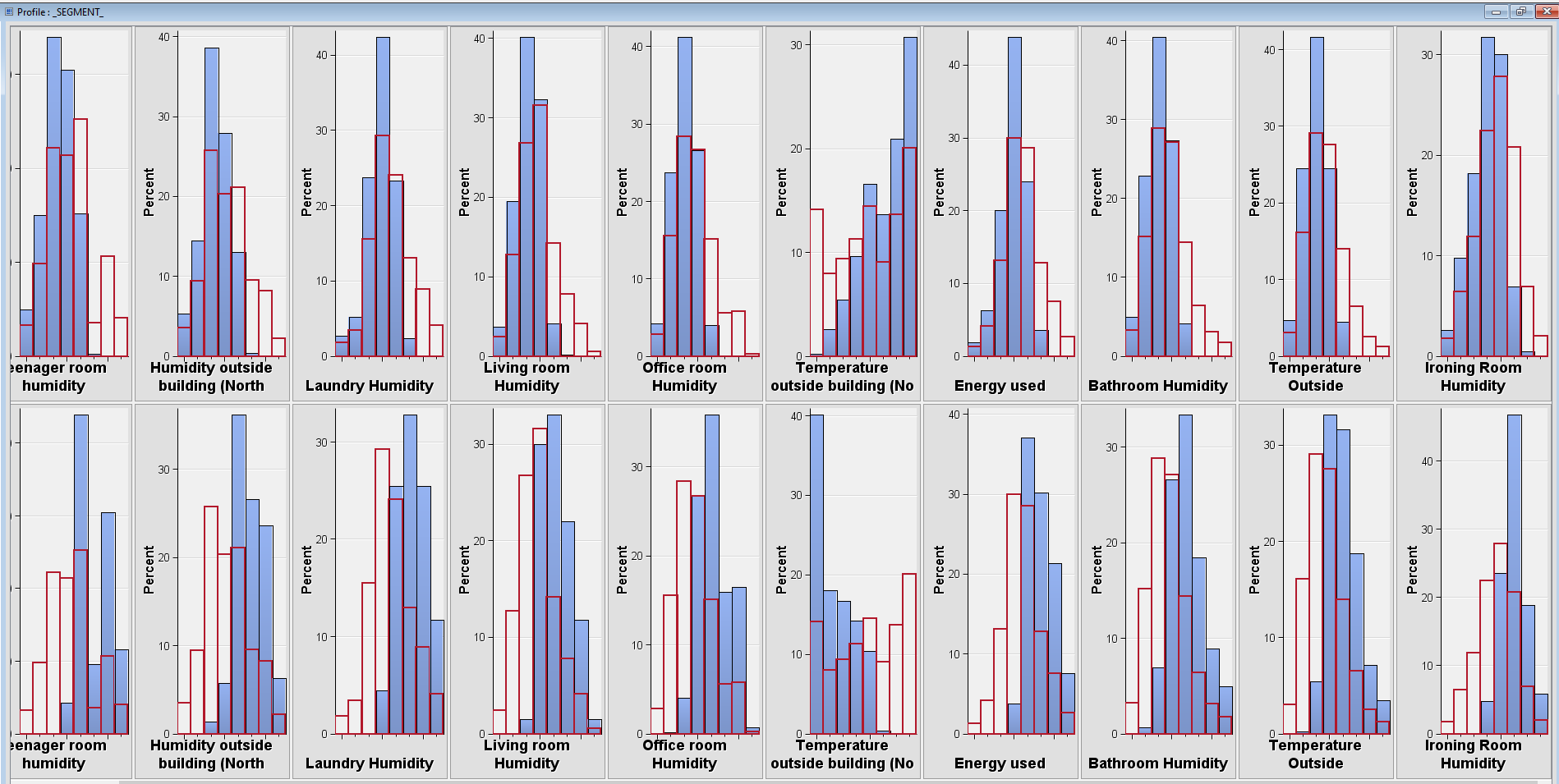
The most important variable used in the splitting is the Laundry Humidity followed by Humidity outside building, teenager room humidity etc. The list of all the important variables used in the splitting and their importance are listed below.



By looking at the Mean Statistics table, we can describe about the clusters based on the variable stats displayed in the table.

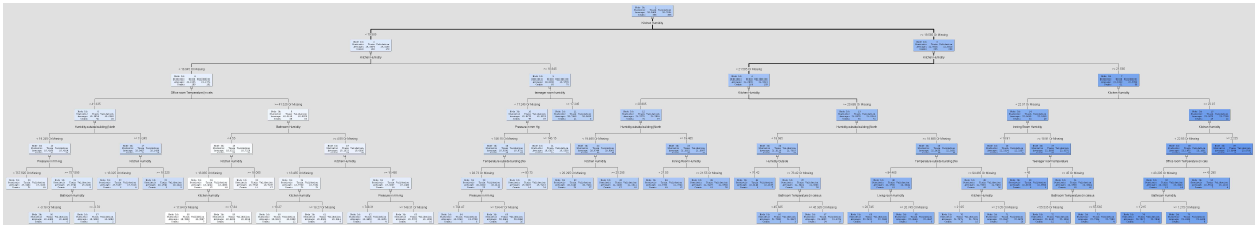
Cluster 1 can be defined as the times where the average appliances used are 108 and lights used are 2 which resulted in the average energy consumption of 23KW. Cluster 2 can be defined as the times of the day where the average appliances used are 91 and lights used are 5 which constitute to the average energy consumption of 20KW. Cluster 1 can be mainly distinguished from cluster 2 in terms of Bathroom Humidity, Temperature outside building and Humidity outside. Cluster 1 can be stated as the times with high humidity percentage in the bathroom. Cluster 2 can be stated as the times with high outside temperatures and humidity percentage.

The other findings that the cluster analysis help us to point out is that two thirds of the total times of all the times the records have been measured, it showed high temperatures and humidity percentages which convey the season to be Summer. There is a relationship that can be pointed at using the values which implies that the Visibility is inversely proportional to the Energy consumed. If there is high visibility it resulted in a lower energy consumption.

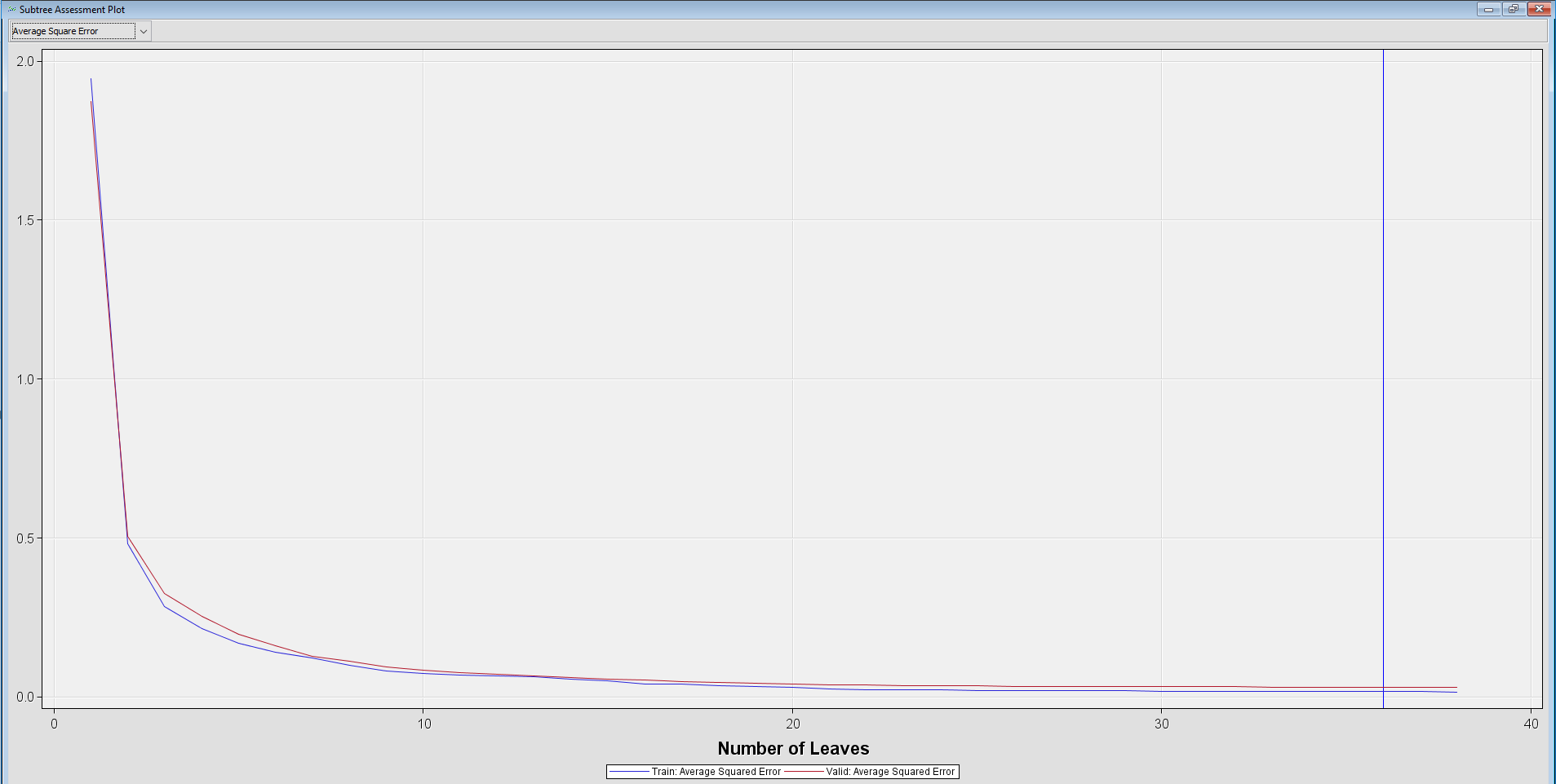
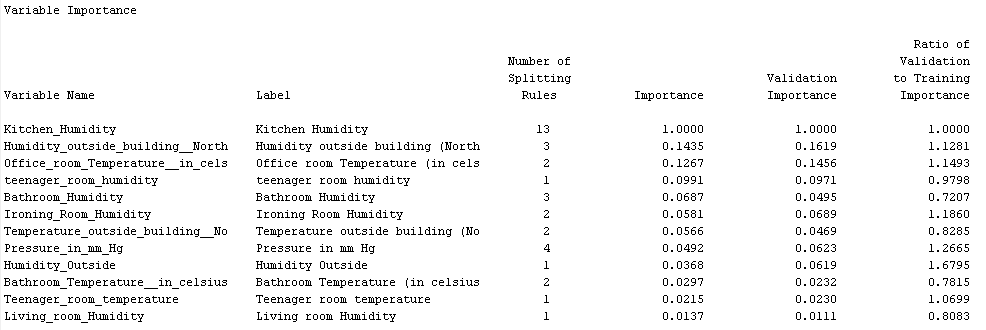


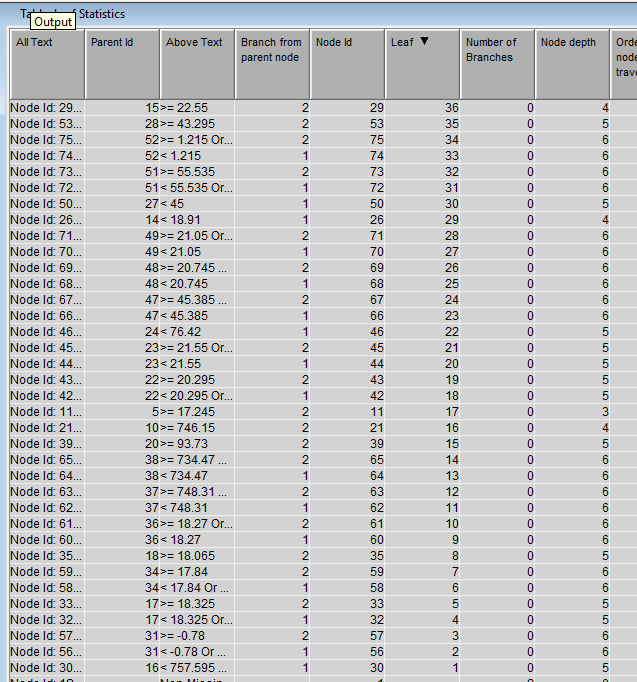
**Decision Trees**

We wanted to test our dataset by building the decision trees, so we initially added the entire dataset and added a Data Partition node prior to building Decision Trees. In the Data Partition node, we set the Training and Validation values as 50, 50 respectively. After executing the decision trees node, the results show that the tree created has 62 leaves and it has thrown a great challenge in analyzing the results with 20K records. The first initial split was made using the Laundry Humidity for the values <22.615 and for the values >= 22.615. The next important variable used for the split was Living Room Humidity for the values <20.63 and > = 20.63. Since the tree is so huge with 62 number of leaves, we felt it very difficult to analyze the tree, so we decided to work building the decision trees by allowing only 1000 rows using the option in the File Import node. So, the resulting tree created using the 1000 records is as shown below.

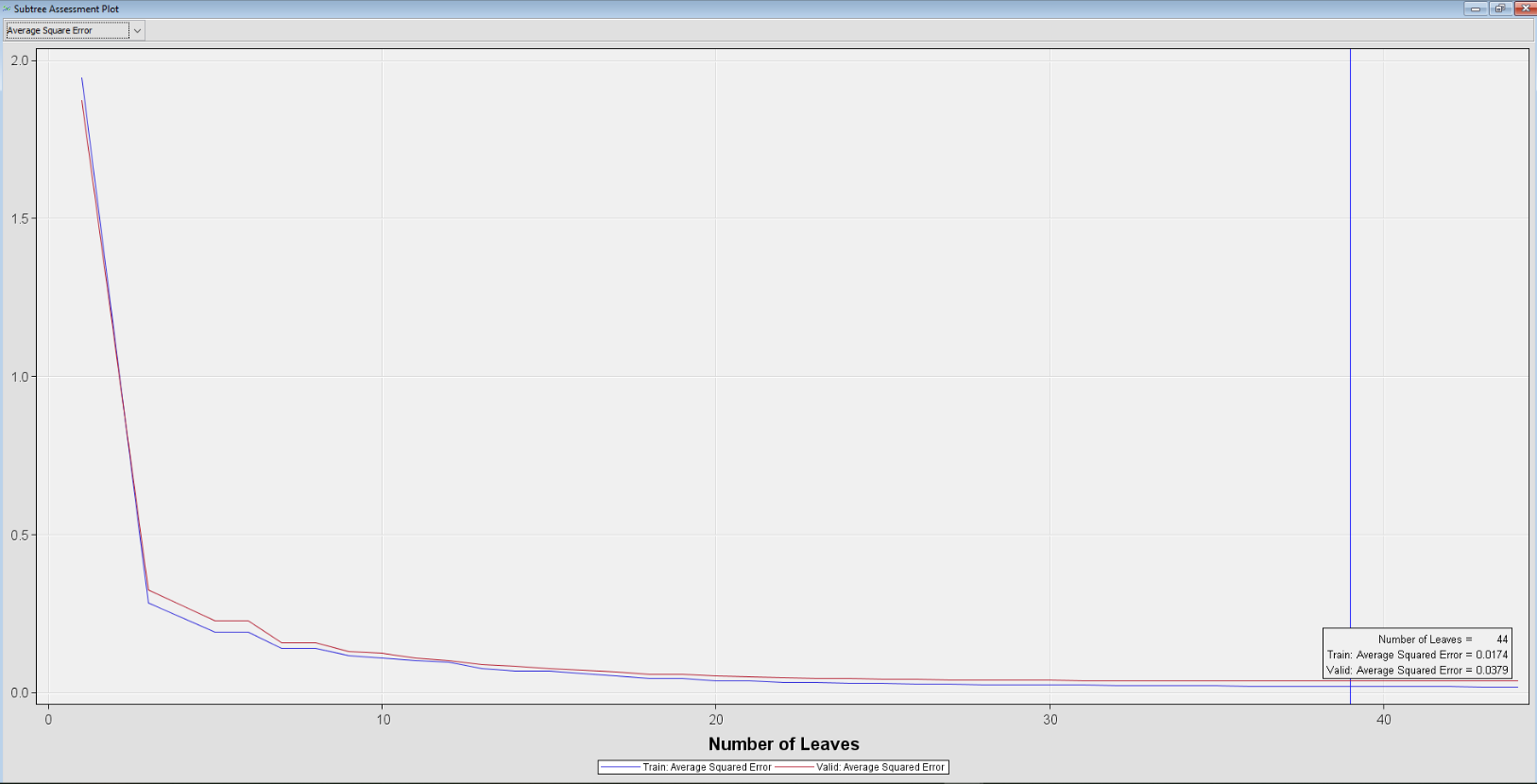
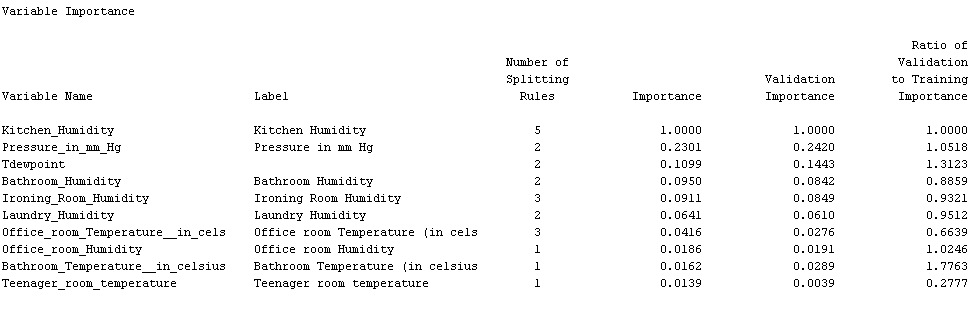


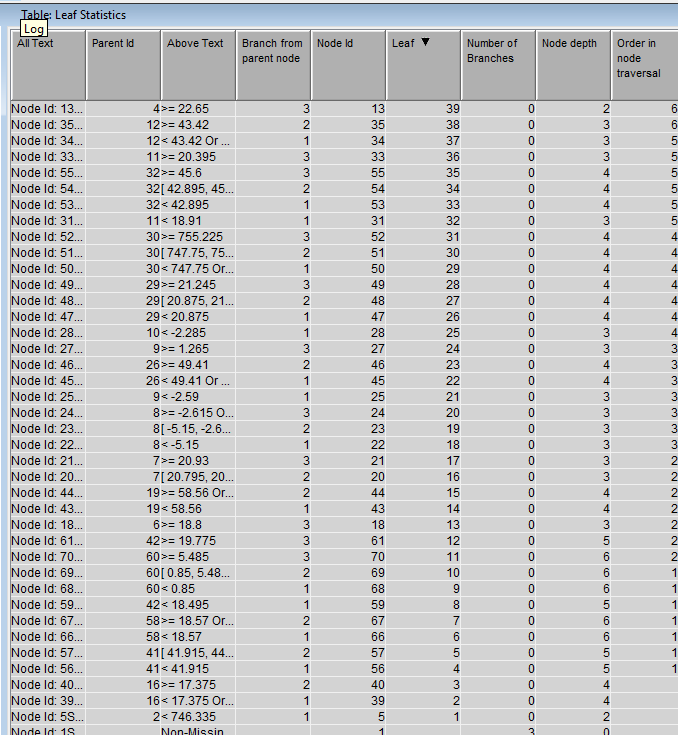
The number of leaves in this resulting tree are 36. And the main variable used in splitting is the Kitchen Humidity for the values <19.585 and >= 19.585, followed by Humidity outside the building and the list is shown below along with their importance.

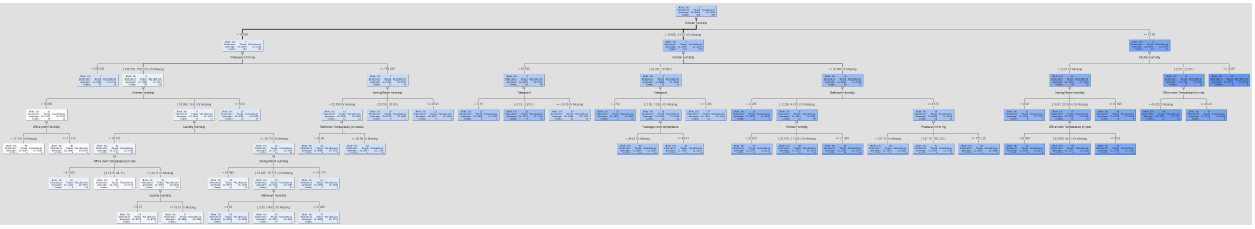


The same tree when implemented using the 3-way branching generates the trees with 44 number of leaves and the main split is made using the variable Kitchen Humidity for the values < 19.585, >=21.55 and for the values between 19.585 and 21.55. The next important variable used for the splitting is Pressure followed by Dewpoint and the other variables used in the split are listed in the pic below.



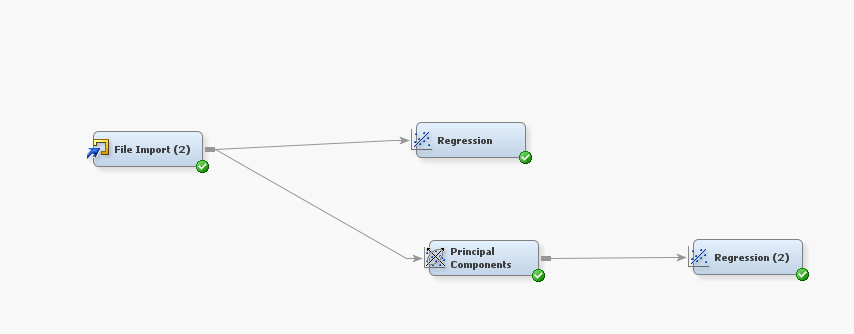
The generated tree is as shown below

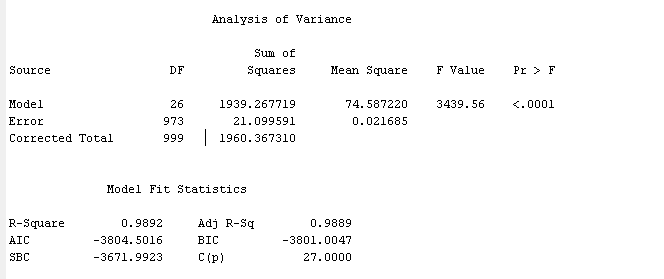


**Principle Component Analysis:**

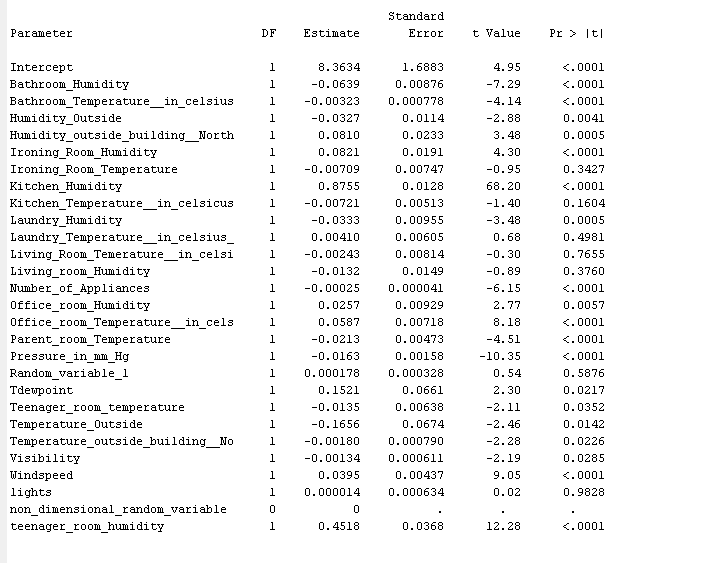
When all the input variables are taken in to consideration among them there are few input variables which don’t contribute much for our regression analysis. We can find such variables from the t value statistics and their associated probabilities (Pr > |t|).

From the variance analysis the model is significant as the Pr > F value is < 0.0001 and the R-square value is 0.989

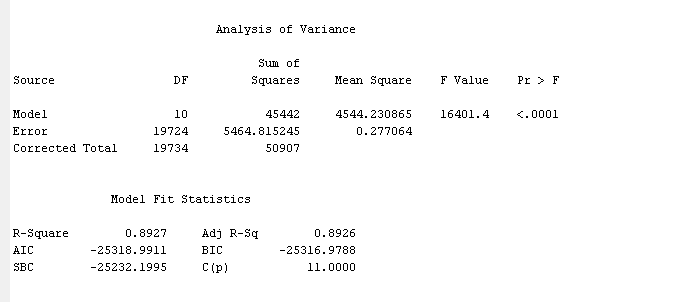




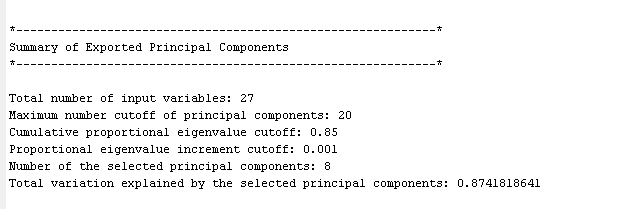
Based on the t-value statistics and associated probabilities parameters Bathroom Humidity, Bathroom Temperature, Ironing room humidity, Kitchen Humidity, Number of Appliances, office room temperature, parent room temperature, pressure and wind speed are significant.



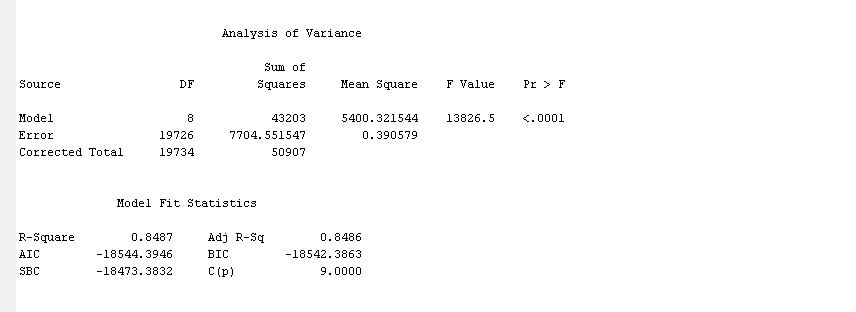
After rejecting all the non-significant parameters from the data set and running the regression analysis once again we get the R-square value as 0.89 and this model is significant as Pr > F values is <0.0001. This implies that all the significant variables can explain about 89% of total variance.



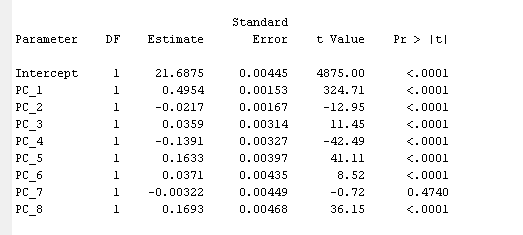
After setting the cumulative frequency to 0.85 and checking the principle component we get a total of 8 principle components which alone explains 87 % of total variation.



This model is significant as the Pr > F value is < 0.0001

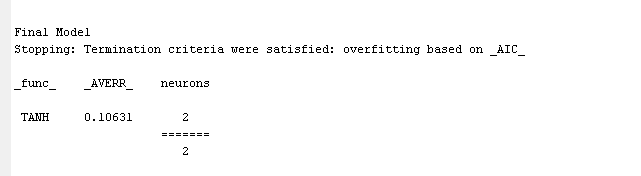


From this model we can see that the principle component 7 (PC 7) is insignificant as Pr > |t| is greater than 0.05 value.

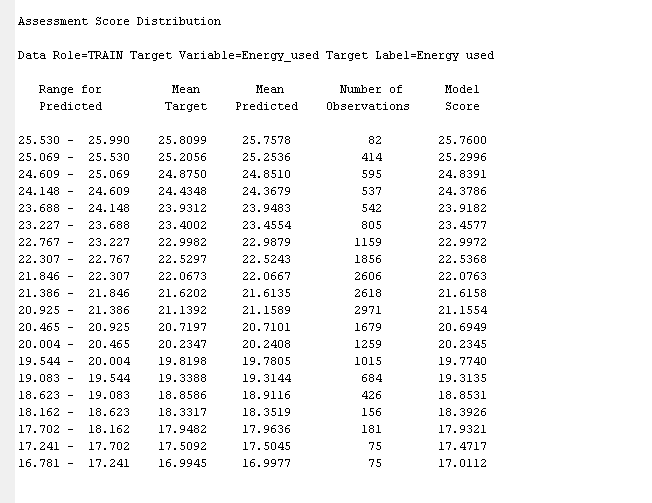


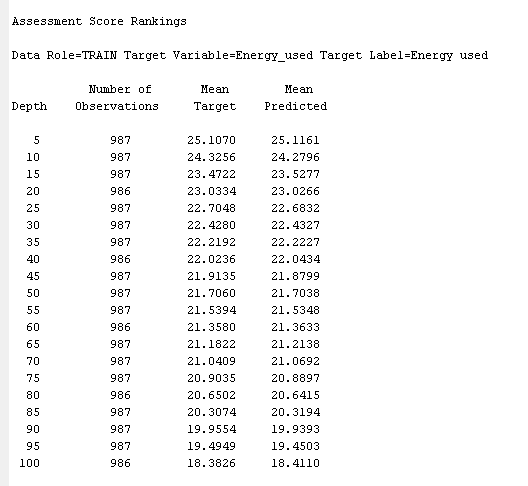
**Neural Network Analysis:**

We used Auto neuralNetwork to analyze our data set and developed a neural network model. We selected different room temperatures, number of appliances, number of lights, humidity levels of different rooms, outside temperature, humidity, pressure, dew value, visibility as the input parameters for the neural networks. For these input parameters we got a neural network model with 2 Neurons.

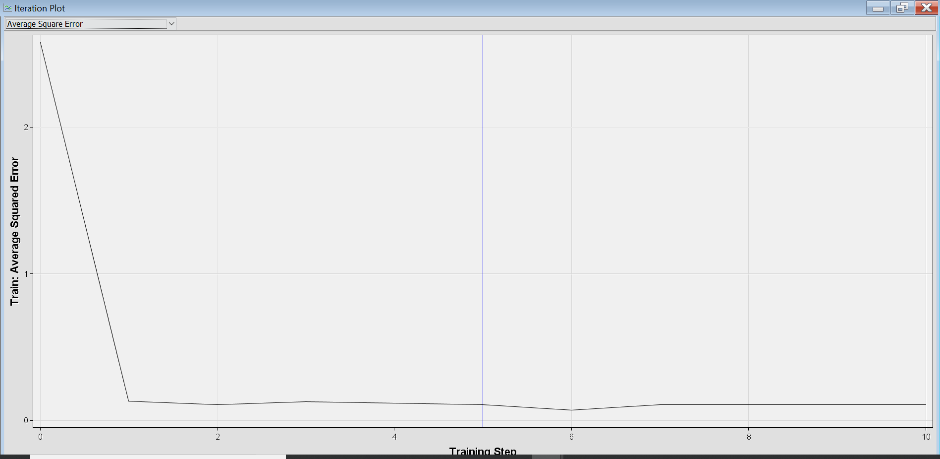


The frequency range and dept for the target variable Energy used are as shown below:

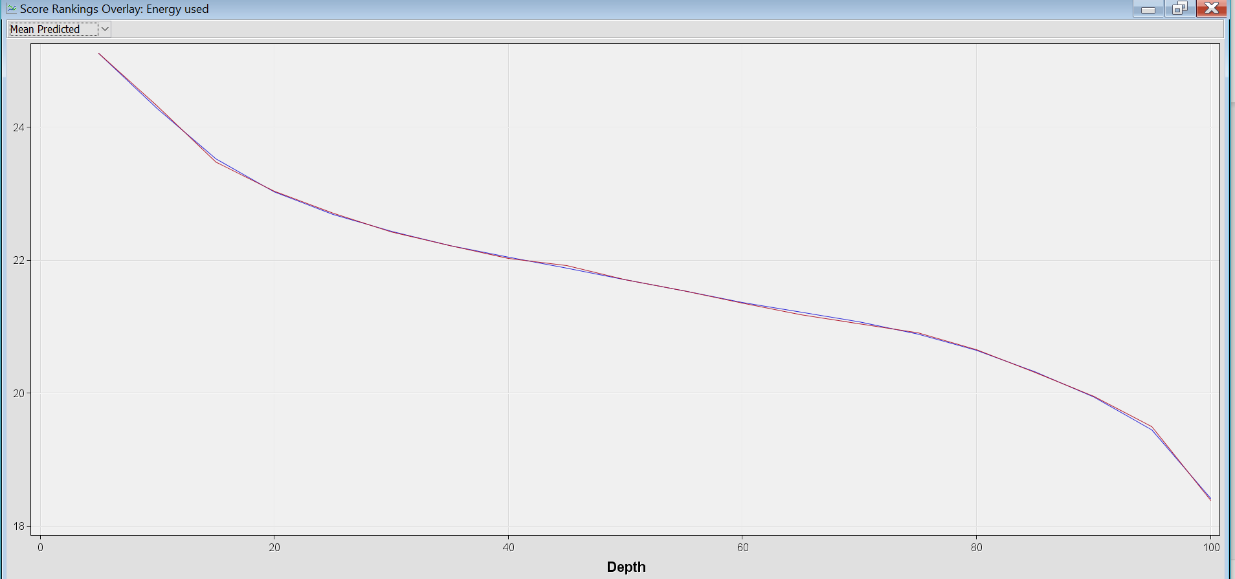




We obtained Average Square Error iteration plot from the results of the Auto neural networks which indicates the estimated process requires 10 iterations. After the fifth iteration the average square error slightly decreases for the data set.



The mean predicted, and mean target graph of the data is almost similar.

**Findings:**

We initially found that the data had no unusual data pattern or missing values but some of the columns had many decimal values which made us to prepare data using Excel to round them off so that the Enterprise Miner does not raise an Error. While conducting the Regression Model, we found the model to be significant but there were some insignificant variables that affect the R2 value, so we had to reject the variables like Visibility, Random Variable 1 and Number of Appliances and rerun the model. We performed the Regression models using features like the Intercept Suppression and using Interactive Binning. On conducting various analyses such as Linear Regression, Cluster Analysis, Decision trees and Predictive Component Analysis we found that the Temperature, Humidity and Pressure contribute to the major part of the predictive analysis of the target variable. The other findings that the cluster analysis help us to point out is that two thirds of the total times of all the times the records have been measured, it showed high temperatures and humidity percentages which convey the season to be Summer. There is a relationship that can be pointed at using the values which implies that the Visibility is inversely proportional to the Energy consumed. If there is high visibility it resulted in a lower energy consumption.

**Conclusions**

From our analysis of different techniques on the selected data set we conclude that the energy consumption depends not only on the number of appliances and lights but also on various factors such as surrounding and room temperatures, humidity levels, pressure, visibility and windspeed.