**Big Mart Sales Prediction Report**

Hemanth Kumar Ponnambalam

Nov 7, 2018

Table of Contents

[Problem Statement 3](#_Toc529222363)

[Executive Summary 3](#_Toc529222364)

[Data Exploration 3](#_Toc529222365)

[Data Preprocessing & Cleaning 5](#_Toc529222366)

[Variable Types 5](#_Toc529222367)

[Variable Analysis 6](#_Toc529222368)

[Outlier Analysis 27](#_Toc529222369)

[Univariate outlier analysis 27](#_Toc529222370)

[Multivariate outlier analysis 28](#_Toc529222371)

[Modelling-Recoded data set 33](#_Toc529222373)

[Data Partition 33](#_Toc529222374)

[Stepwise Regression 34](#_Toc529222375)

[Neural Network 38](#_Toc529222376)

[Decision Tree 40](#_Toc529222377)

[Logistic Regression 42](#_Toc529222378)

[LDA (Discriminant Analysis) 44](#_Toc529222379)

[Ensemble Models 47](#_Toc529222380)

[Summary Table 53](#_Toc529222381)

[Modelling-No Missing Values 53](#_Toc529222382)

[Data Partition 53](#_Toc529222383)

[Stepwise Regression 55](#_Toc529222384)

[Neural Network 56](#_Toc529222385)

[Decision Tree 58](#_Toc529222386)

[Logistic Regression 60](#_Toc529222387)

[LDA (Discriminant Analysis) 61](#_Toc529222388)

[Ensemble Models 62](#_Toc529222389)

[Summary Table 67](#_Toc529222390)

[Recommendations: 67](#_Toc529222391)

[References 67](#_Toc529222392)

# Problem Statement

BigMart has collected 2013 sales data for 1559 products across 10 stores in different cities. Certain attributes of each product and store have been defined in the dataset. The goal is to build a predictive model and find out the properties of products and stores which impact the sales. Using the model, BigMart will try to make necessary changes in order to increase their sales.

# Executive Summary

The purpose of this report is to suggest changes and explain the outcomes of our model to the decision makers of BigMart, so that they can get necessary insights about the factors that are mainly influencing the sales of their products. In order to predict the volume of sales, we had to determine the importance and impact of the characteristics of the products presented in the database: the item identifier, its type, weight, fat content, visibility in the store and its Maximum retail price (MRP). However, the products alone don’t explain the sales, we had to also take into consideration the contribution of the type of each outlet, its location, size and number of running years.

After exploring the data and pre-processing it by dealing with missing values, outliers and errors (detailed steps for which are documented below), we created two data sets, one with recoding the missing values and the other with no missing values. We named the recoded data set as ‘Data\_Imputed’ and the one with no missing data as ‘Data\_NoMissing’. We also updated the existing data dictionary so as to reflect the changes made to the variables.

We built several models on both the data sets to predict the sales. We ended up choosing Decision tree model of the recoded data set as the better model. It turned out that customers tend to prefer a product with a high MRP because the negotiation margin is bigger and because a high price tends to be associated with a better quality. Also, we concluded that the outlet type and how old it is, influences the volume of sales. However, even though our model could predict the sales, it was unable to explain all the variability due to lack of data.

Having more information would help in building a better model which can explain the variability. Also, the quality of data that BigMart has can be improved. There was missing data and ambiguity in the data dictionary for a few variables. Clearly defined descriptions in the data dictionary and a data set with no missing values will benefit the company manifold. If you still have missing values, then it is recommended to reach out to the source in order to fill them up.

A list of recommendations is given at the end of the report.

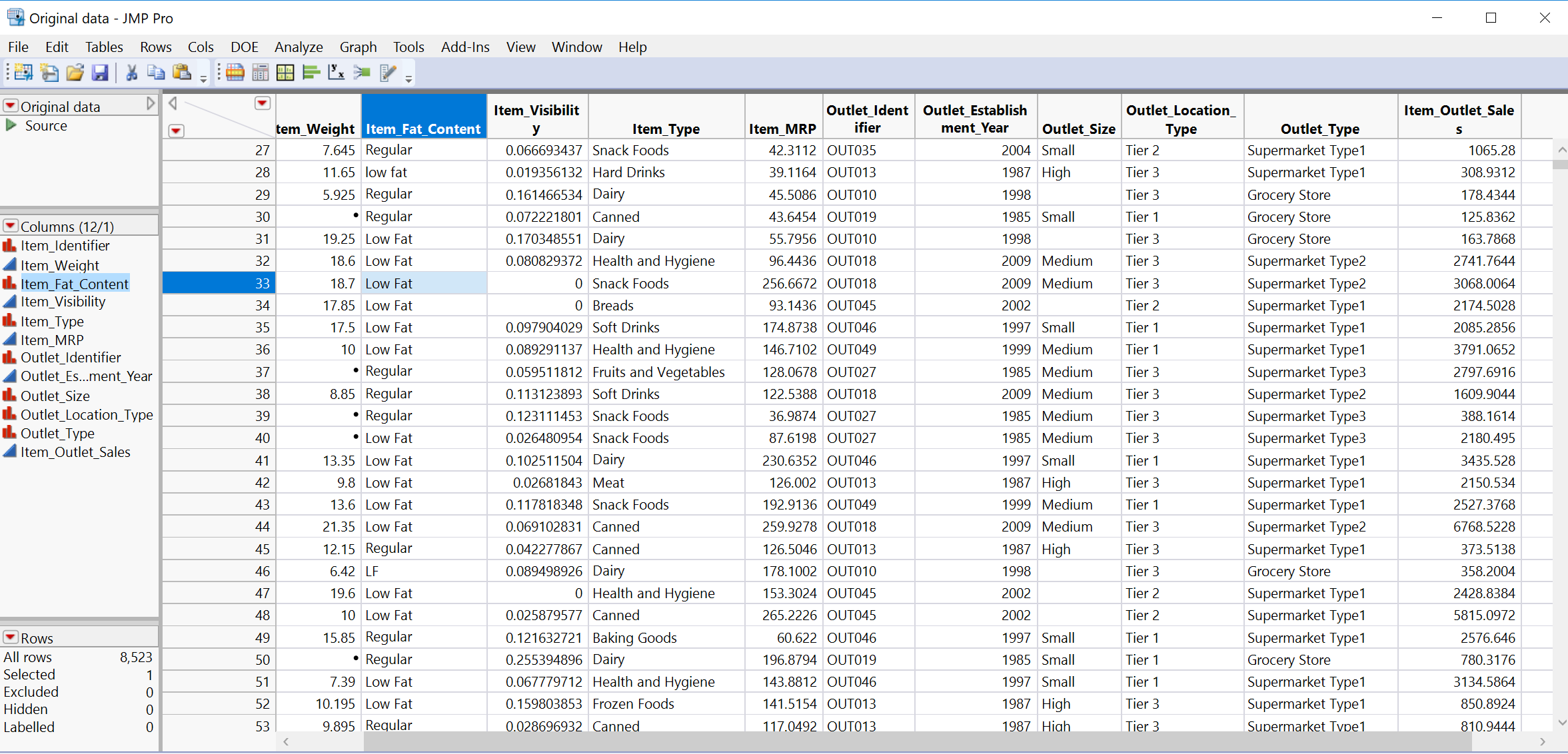
See ----------.

# Data Exploration

Total observations: 8523

Total variables:12

Dependent variable=Item\_Outlet\_Sales



*Data Dictionary:*

|  |  |  |
| --- | --- | --- |
| Variables | Description | Variables type |
| Item\_Identifier | Unique Product ID | Nominal |
| Item\_Weight | Weight of the Product | Continuous |
| Item\_Fat\_Content | Whether the product is low fat or not | Nominal |
| Item\_Visibility | The % of total display area of all products in a store allocated to the product | Continuous |
| Item\_Type | The category to which the product belongs | Nominal |
| Item\_MRP | Maximum Retail Price (list price) of the product | Continuous |
| Outlet\_Identifier | Unique store ID | Nominal |
| Outlet\_Establishment\_Year | The year in which store was established | Nominal |
| Outlet\_Size | The size of the store in terms of ground area covered | Nominal |
| Outlet\_Location\_Type | The type of city in which the store is located | Nominal |
| Outlet\_Type | Whether the outlet is just a grocery store or some sort of supermarket | Nominal |
| Item\_Outlet\_Sales | Sales of the product in the particular store. This is the outcome variable to be predicted. | Continuous |

According to the description stated above, there are two levels of data which is linked to the sales and their inferences that can be made by looking at the problem statement.

* Product properties
* Outlet properties

**For product properties level:**

Item\_Weight：The more weight of one product has, the more expensive.

Item\_Fat\_Content: Low fat are generally used more than others.

Item\_Visibility: Products have more display area will sell better.

Item\_Type: May have relation with weight, fat and Item identifier.

Item\_MRP: The higher the price, the better the sales.

**For Outlet level:**

Outlet\_Establishment\_Year: The longer the outlet established, the more sales per unit.

Outlet\_Size: Stores which are very big in size might have higher sales as they act like one-stop-shops and people would prefer getting everything from one place.

Outlet\_Location\_Type: Downtown and urban stores have better sales related to different location.

Outlet\_Type: May relate to location, size and sales.

# Data Preprocessing & Cleaning

Once the EDA is done, the next step is related to data preprocessing & cleaning where we deal with correcting the variable types, dealing with missing values and outliers.

The steps carried out in this part of data preprocessing are the following:

1) Validate the variables with the data dictionary.

2) Check Column > Recode to observe the individual values, their range and type.

3) Change the variable type according to the above observations (also need to consider inconsistent values).

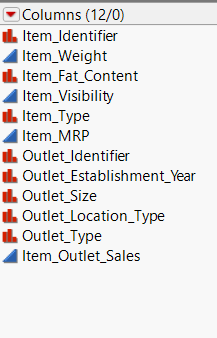
4) Look for missing values.

5) Understand the missing value patterns using Table > Missing Value Pattern

6) Make a decision regarding the missing values (i.e. to delete rows, delete variable, ignore the missing values or impute/recode/transform the variable).

## Variable Types

Variable types are validated against the data dictionary.



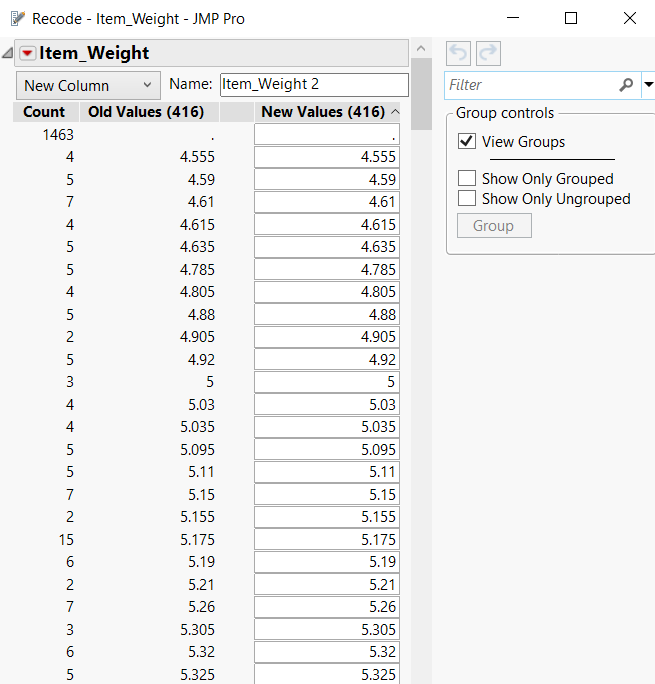
## Variable Analysis

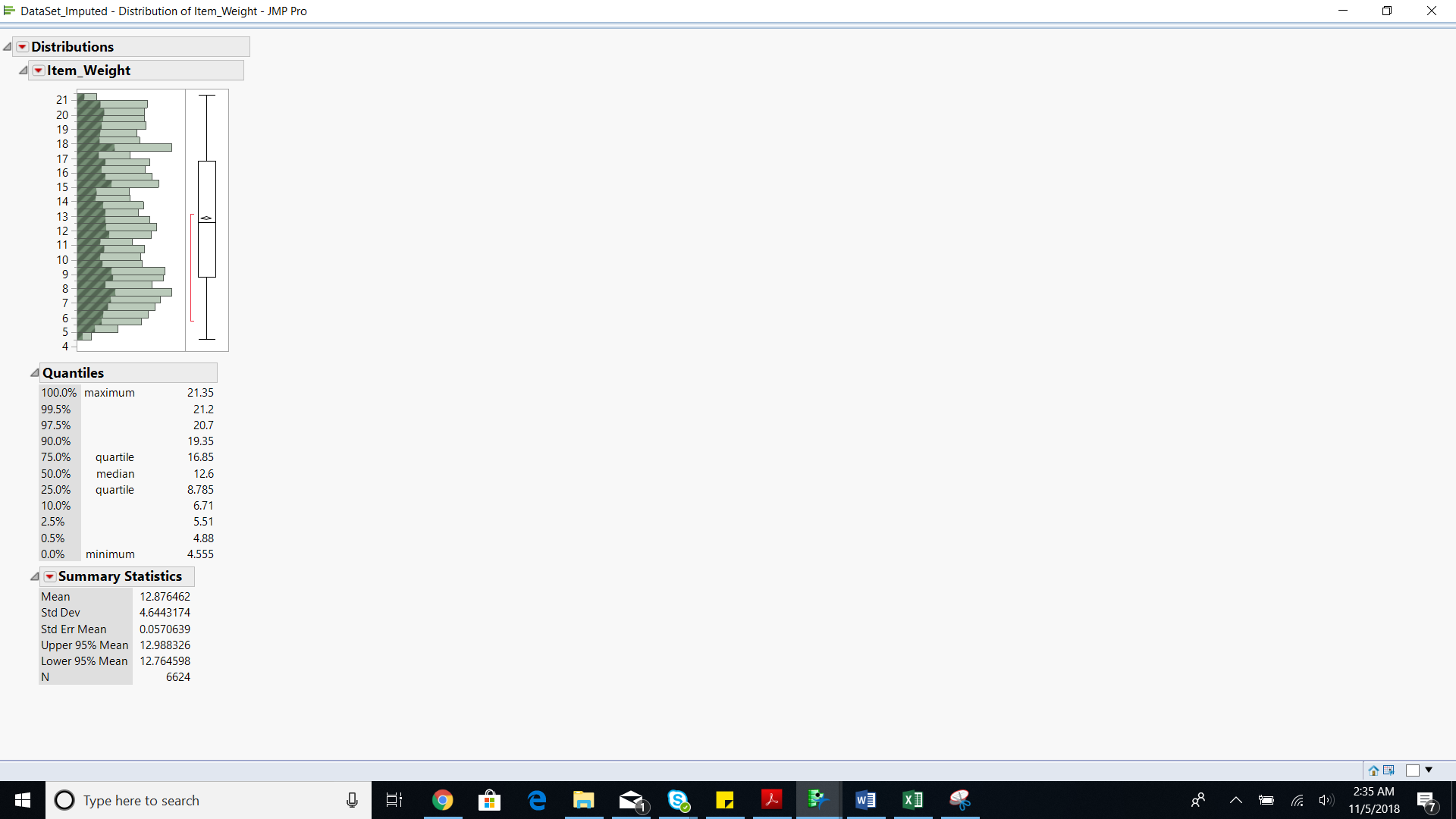
*Item\_Identifier*

Based on the summary table there are 1559 unique product ID values and act as identifiers for every product and doesn’t play any role while building models. It is kept in the dataset for reference purpose. Our model is more concerned about predicting the properties of the products and stores that impact sales.

*Item\_Weight*

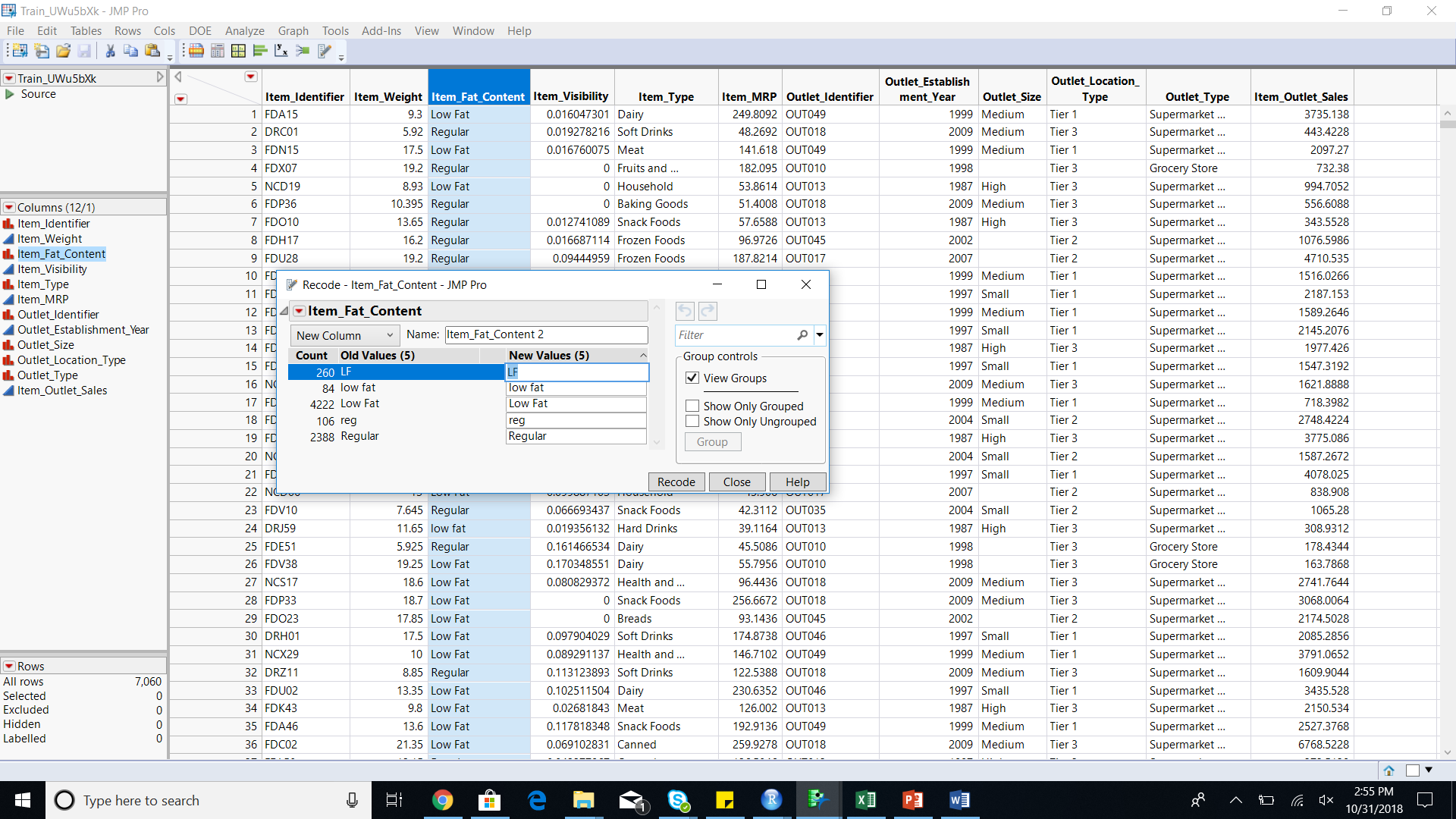
Item Weight has 1463 missing values and has been recoded with mean.

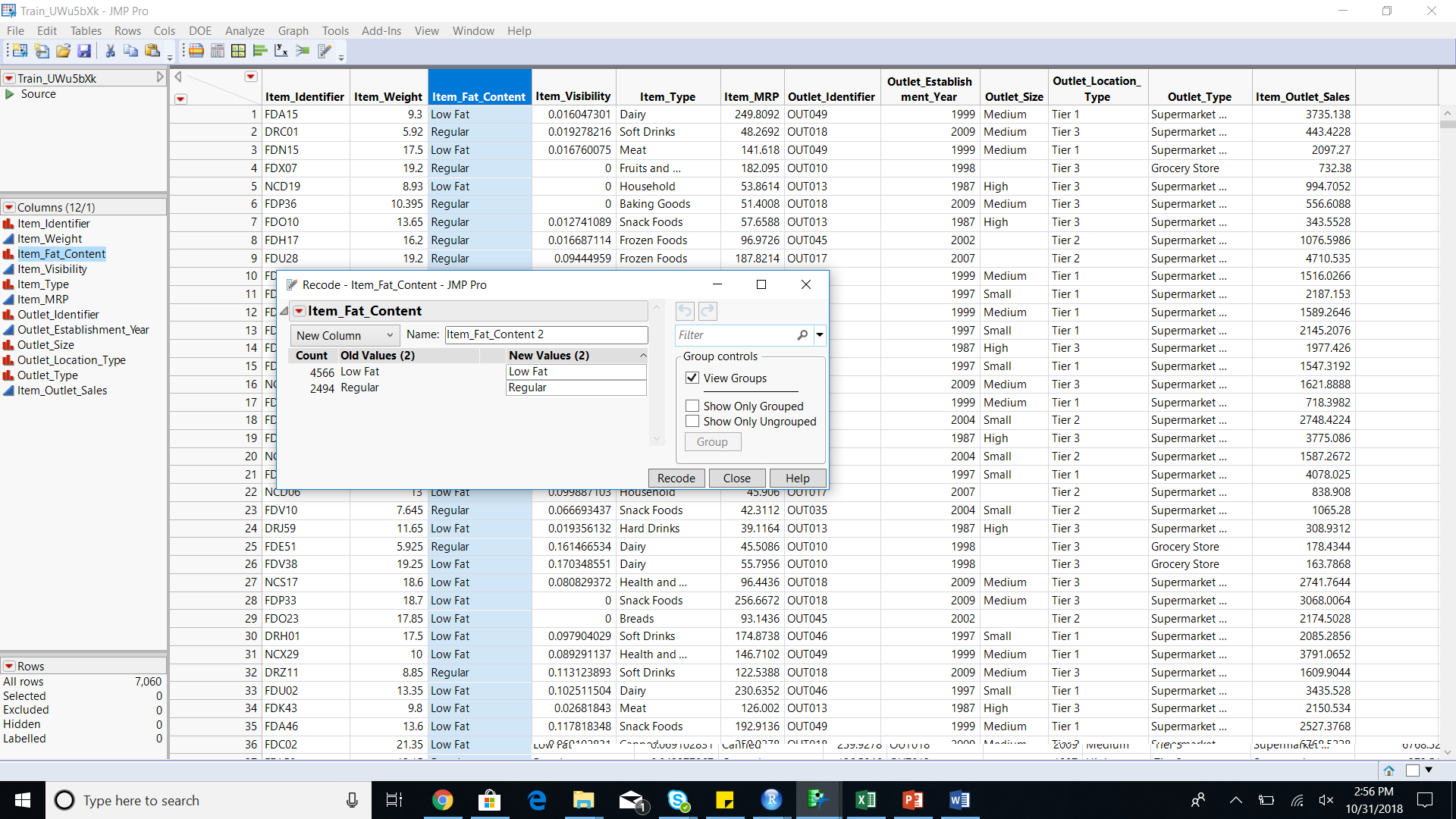




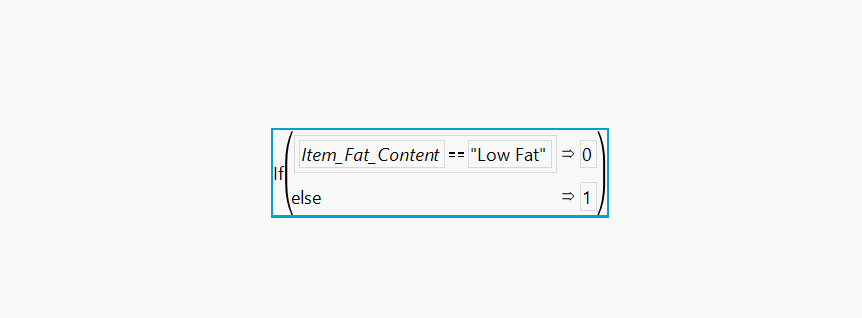
*Item\_Fat\_Content*

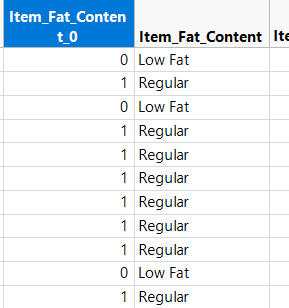
Fixed the incorrect values based on data dictionary. Recoded the values to be Low fat/Regular





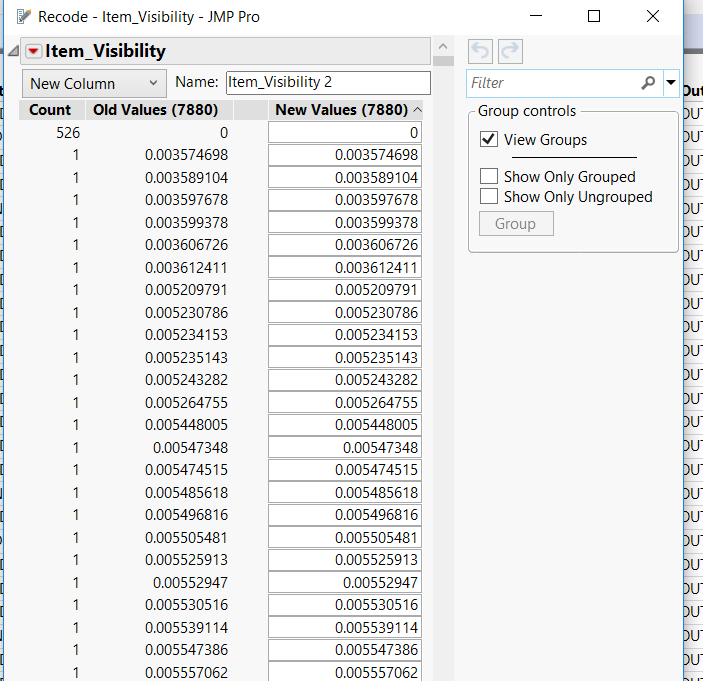
Added new column to convert the item\_fat\_content to continuous variables.

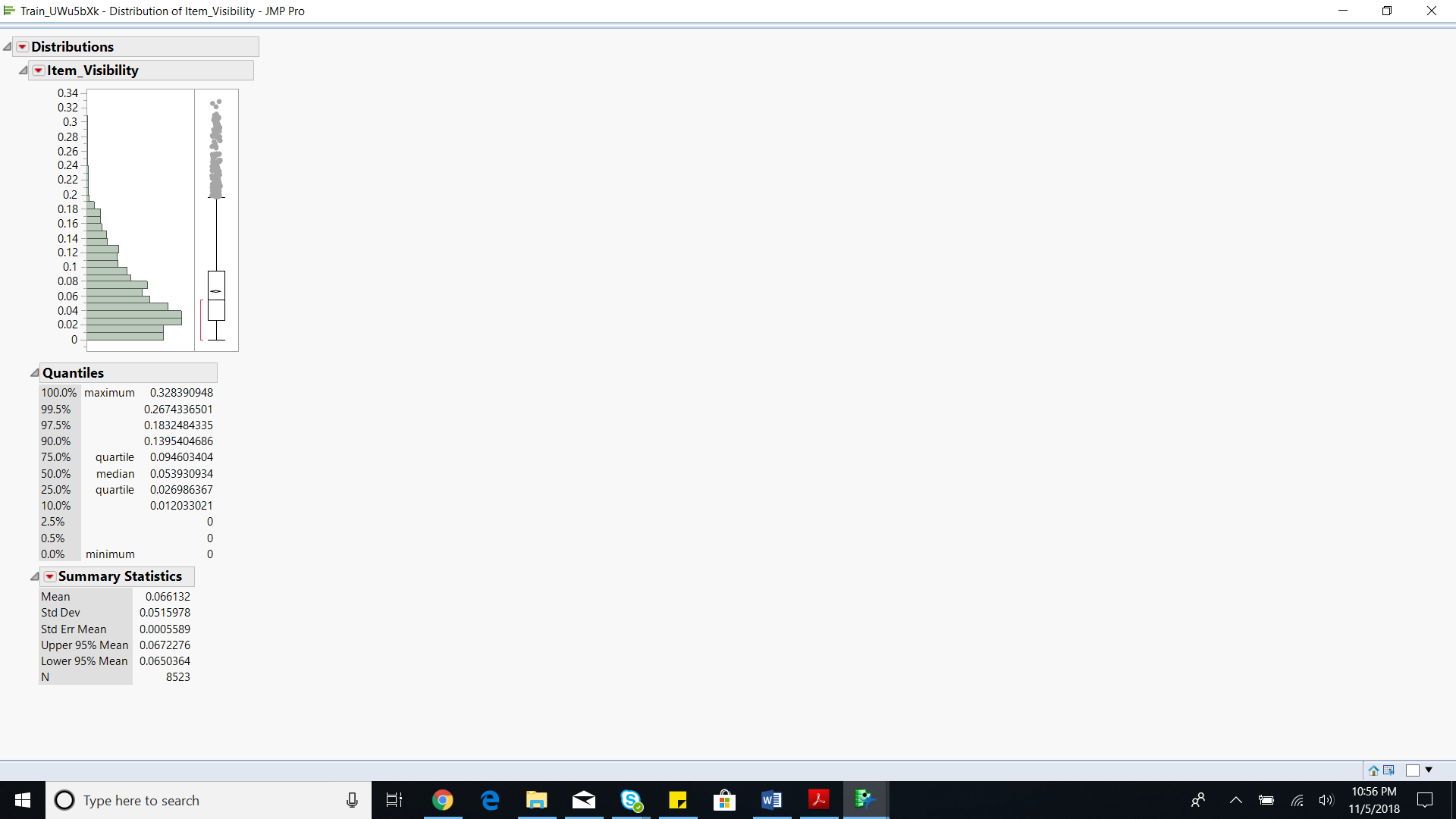


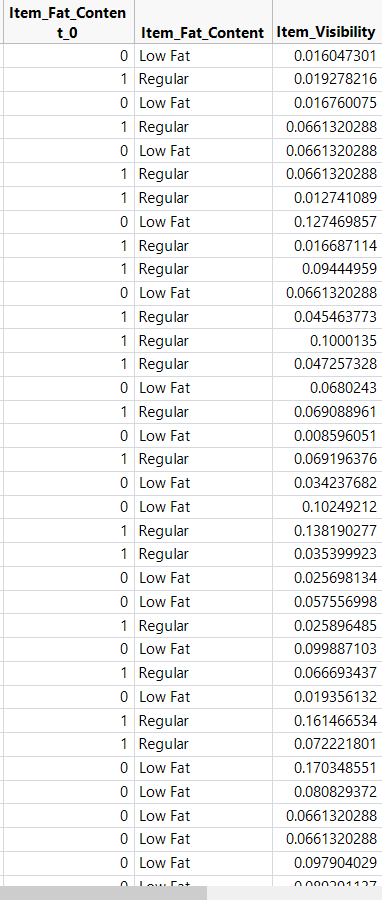


*Item\_Visibility*

Item Visibility has 526 values as 0. We recoded the missing values with the mean.

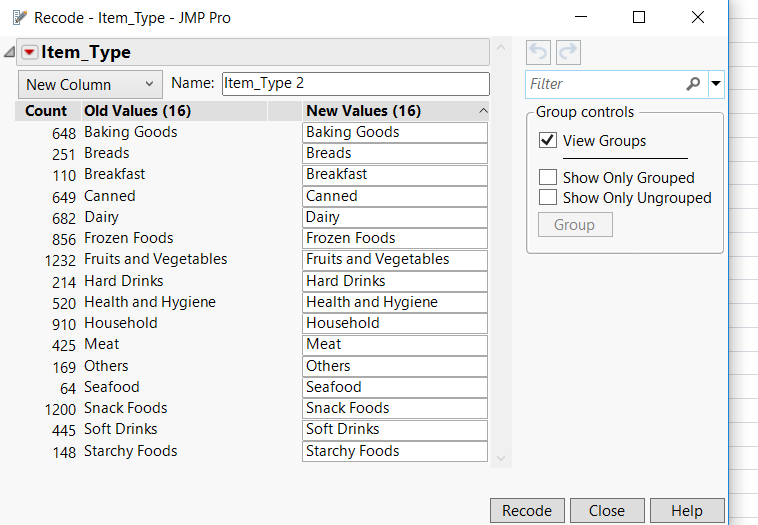


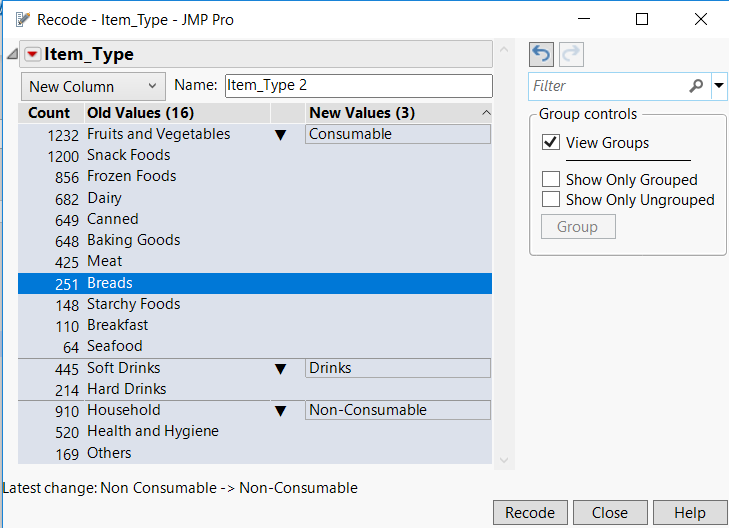


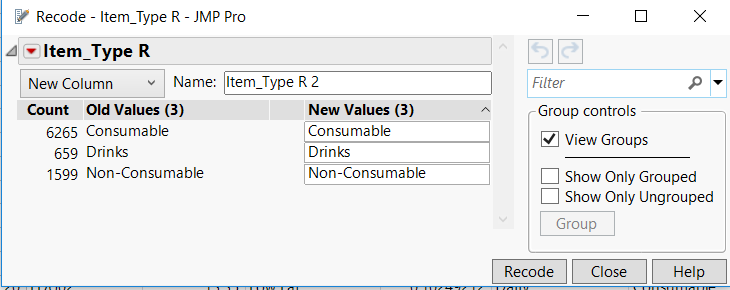


*Item\_Type*

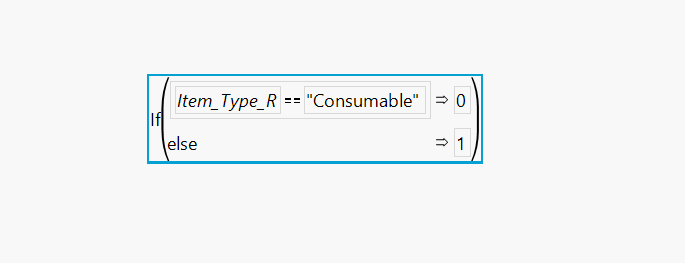
Item\_Type has multiple categories and has been grouped and recoded into consumable, non-consumable and drinks because we are predicting the sales based on categories of products. Also, it would be easier to create dummy variables for 3 categories so as to use them as continuous variables for different models.

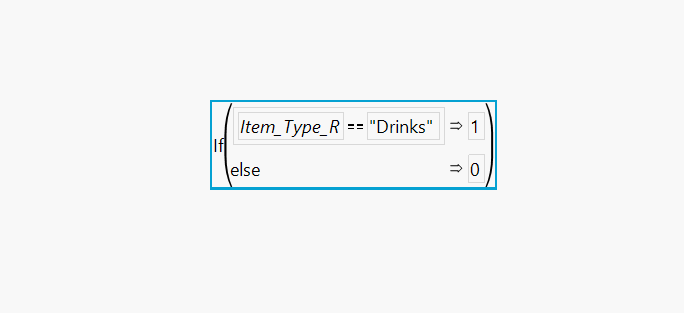


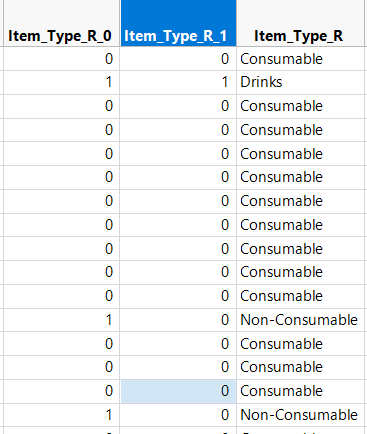




Additional continuous columns have been created that can represent patterns for different type of categories using one-hot encoding method.

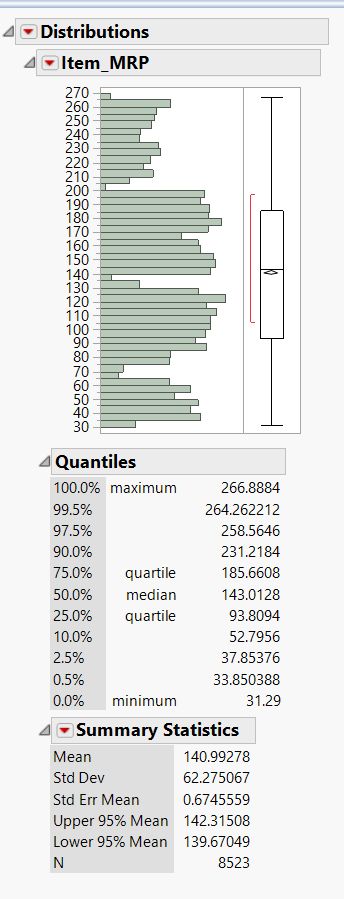






*Item\_MRP*

We did not make any changes to Item\_MRP as it has no missing values and outliers.

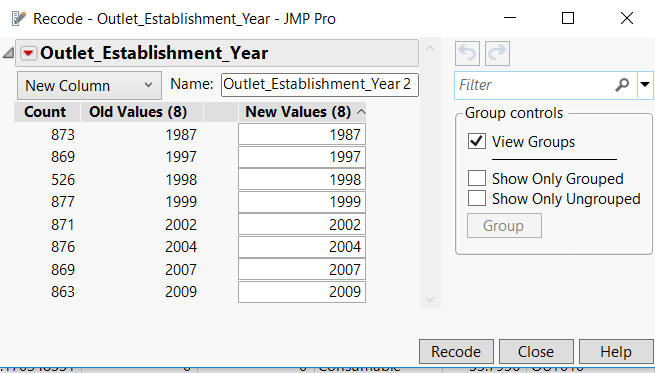


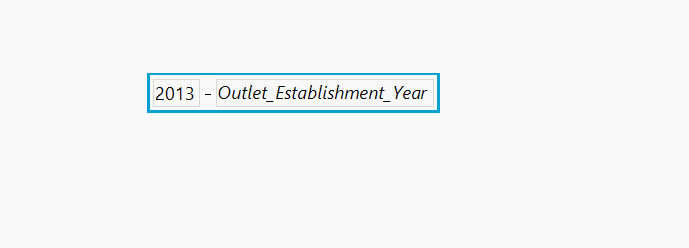
*Outlet\_Identifier*

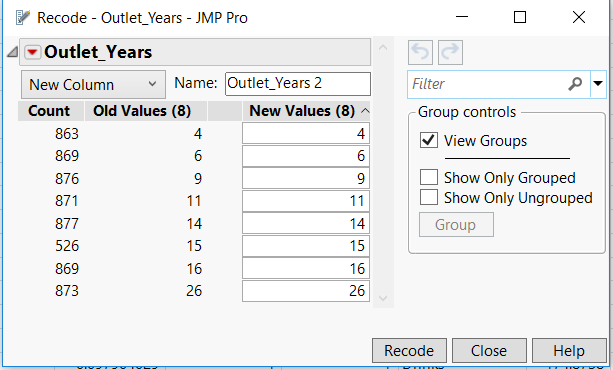
This variable has 10 categories with no missing values. Also, it is not used as a predictor variable. So, we did not make any changes to it.

*Outlet\_Establishment\_Year*

Based on the problem statement, we are supposed to predict Bigmart’s sales as of 2013. So, we can convert Outlet\_Establishment\_Year into a continuous variable by finding the difference between 2013 and Outlet\_establishment\_year in order to get the number of years since establishment.





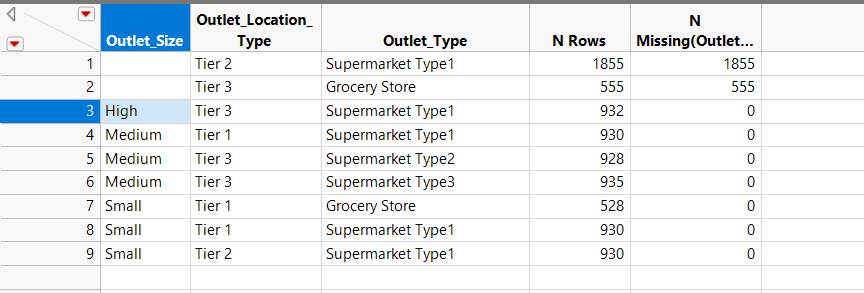


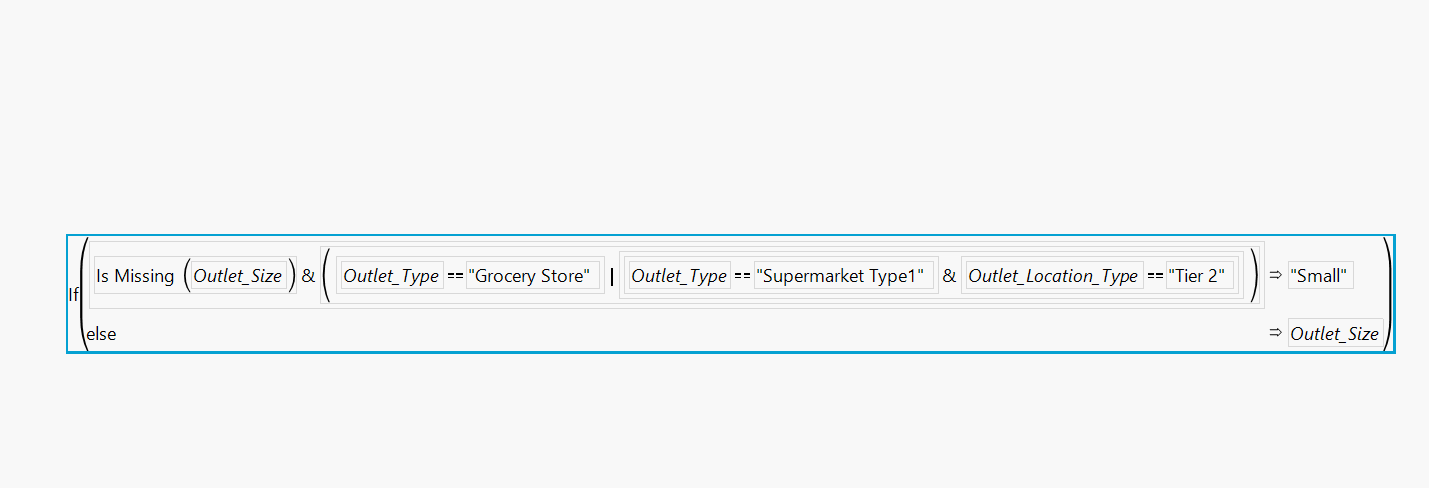
*Outlet\_Size*

Identified missing pattern for Outlet\_size using ‘summary table’.

Based on the summary table, two inferences can be made

* Grocery store has Small as Outlet\_size and missing values can be replaced with small.
* Outlet type=Supermarket type 1 and Outlet location type=Tier 2 have Small Outlet Size as a majority. Hence, missing values for these patterns can be replaced with Small.

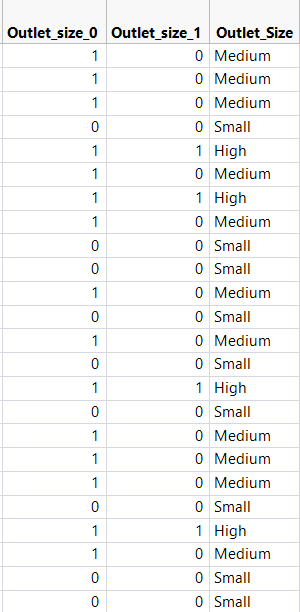




Once this is done, Additional 2 columns have been added with patterns using formulas. These columns can be used in the models where explanatory variables need to be continuous. If the model accepts nominal variables as factors, then Outlet\_Size column can be used.

Below are the patterns:

|  |  |  |
| --- | --- | --- |
| High | 1 | 1 |
| Medium | 1 | 0 |
| Small | 0 | 0 |

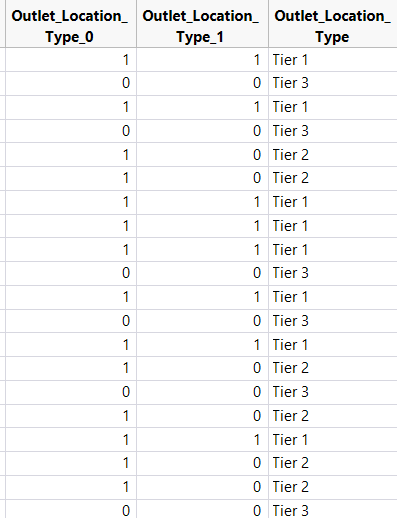
**

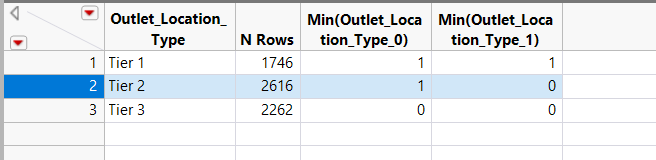
*Outlet\_Location\_Type*

Additional two Continuous Columns: Outlet\_Location\_Type\_0, Outlet\_Location\_Type\_1 are created for the Outlet\_Location\_Type and patterns are assigned using formulas.

Below are the patterns:

|  |  |  |
| --- | --- | --- |
| Tier 1 | 1 | 1 |
| Tier 2 | 1 | 0 |
| Tier 3 | 0 | 0 |



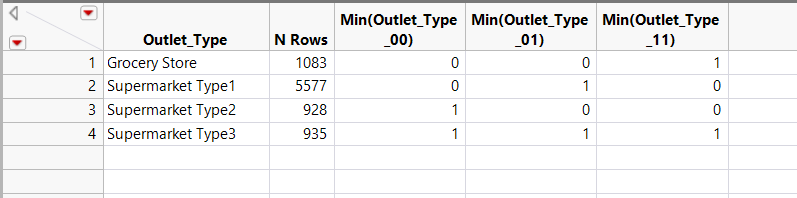


*Outlet\_Type*

Additional 3 Continuous Columns: Outlet\_Type\_00, Outlet\_Type\_01, Outlet\_Type\_11 are created for the Outlet\_Type and patterns are assigned using formulas.

Below are the patterns:

|  |  |  |  |
| --- | --- | --- | --- |
| Grocery Store | 0 | 0 | 1 |
| Supermarket Type1 | 0 | 1 | 0 |
| Supermarket Type2 | 1 | 0 | 0 |
| Supermarket Type3 | 1 | 1 | 1 |



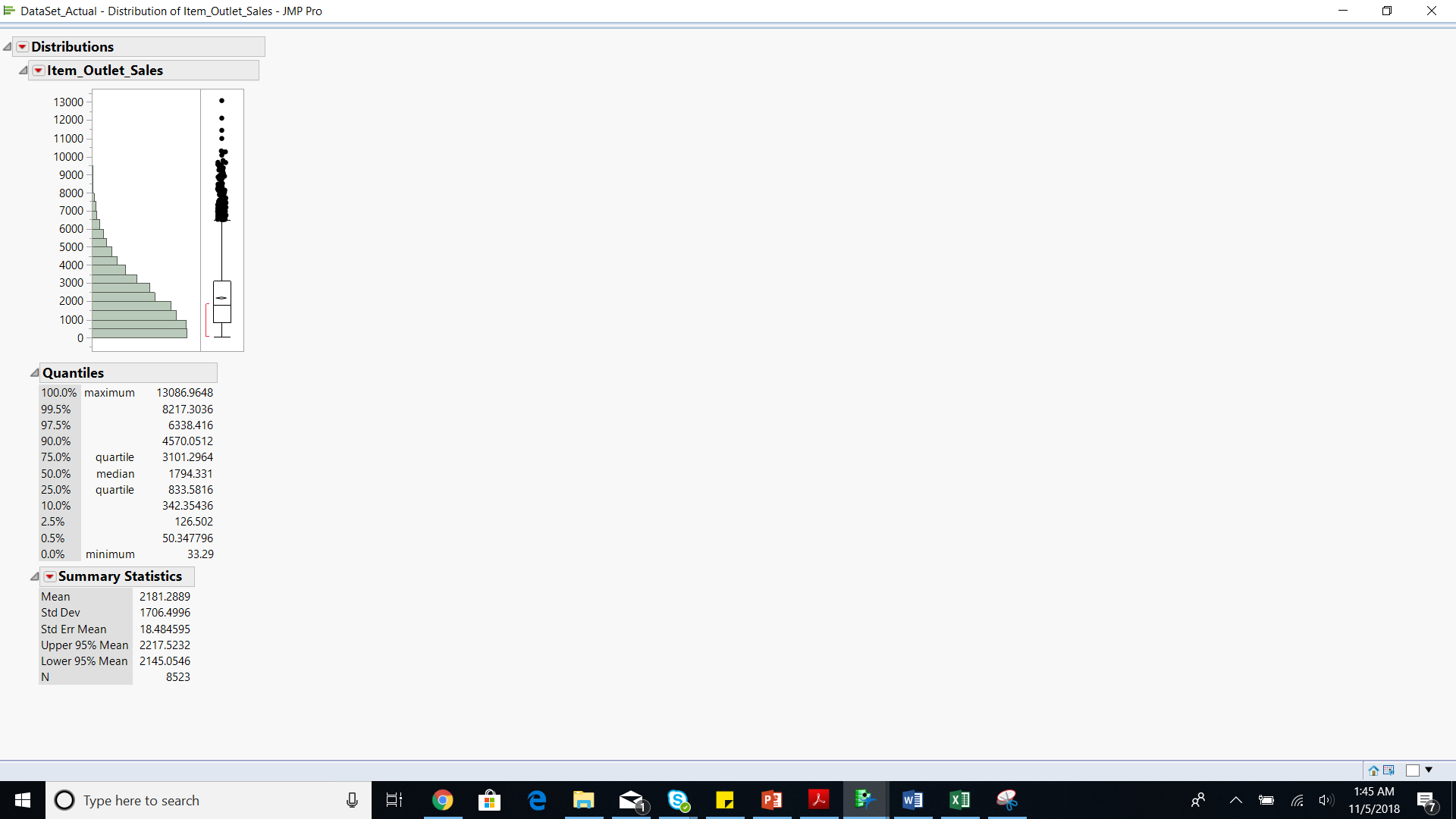


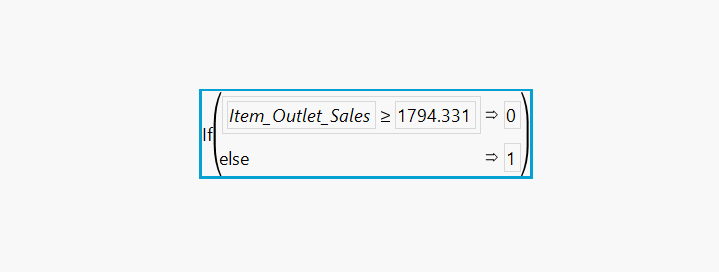
*Item\_Outlet\_Sales*

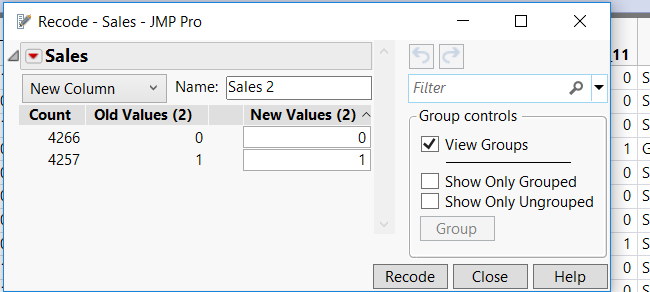
Item\_Outlet\_Sales is our predictor variable.

Additional Column, ‘Sales’ is created with 2 categories (0,1) based on the median value of Item\_Outlet\_Sales.

This is used in building models like Linear Discriminant analysis and Logistic Regression where the Dependent variable needs to be nominal.







## **Outlier Analysis**

### **Univariate outlier analysis:**

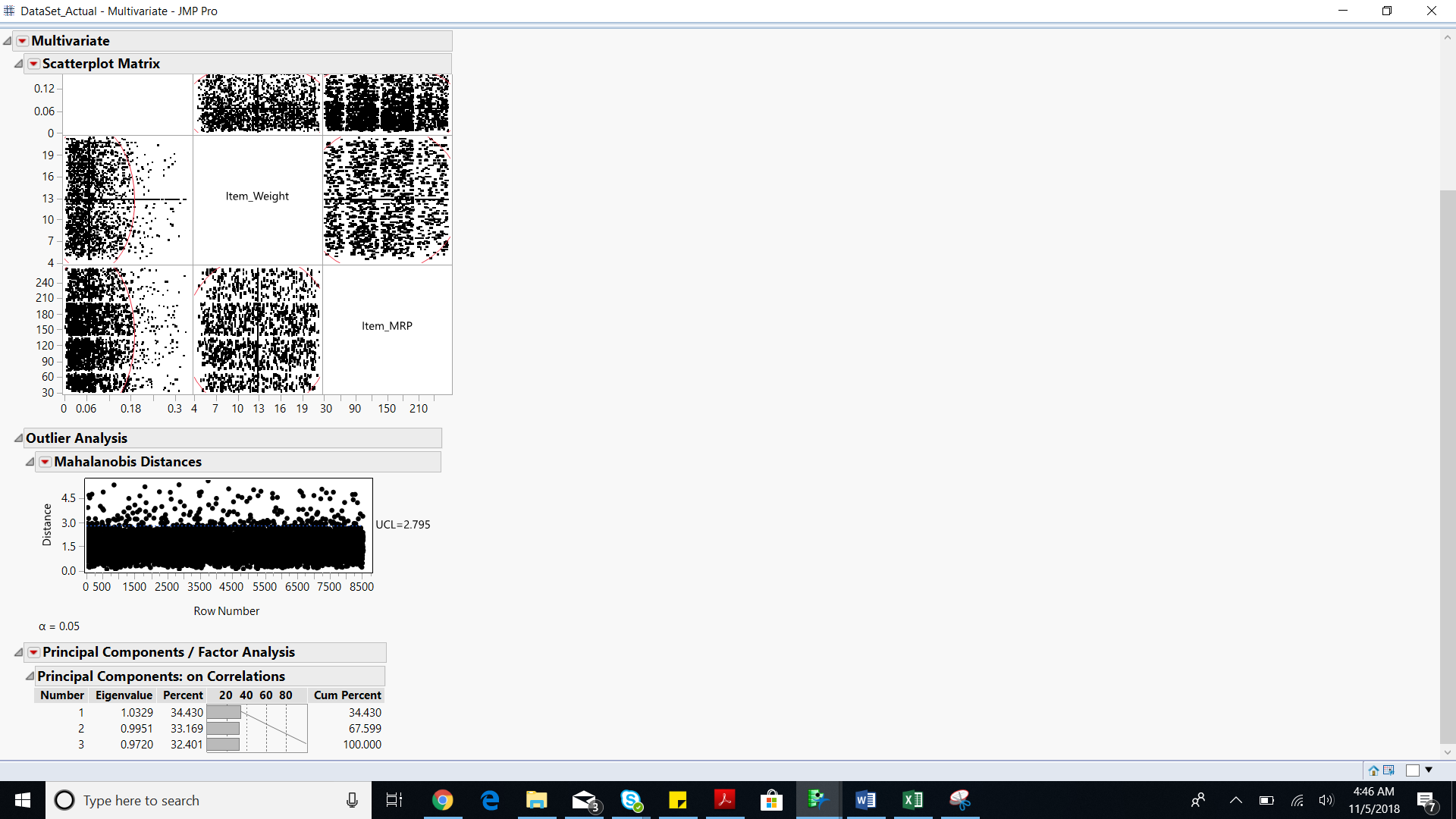
Item\_visibility was found to have outliers that lie outside whiskers of the box plot, which is indicated in the screenshot below. There is an option to transform this variable through Continuous Fit, where all the variables will be transformed to the best fit curve. However, we did not employ this method, since the original data will be modified. Therefore, we did not remove any outliers at this stage as they may be showing variance and thus, will be important for our models.



### **Multivariate outlier analysis**

Multivariate outlier analysis is more robust because it considers the variabilities in the combinations of all the variables as a distance function. This can be show in the three distance functions below.

From our analysis, we observed that there is a lot of variability in the data and there are no clear egregious outliers. Because variability gives us additional information, we may need this at the modelling stage for predictions. Therefore, we did not remove any outliers from our data.

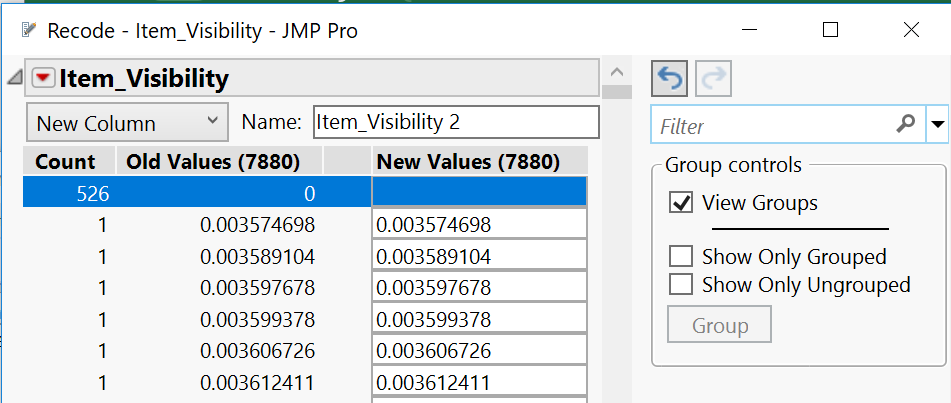


**Modified Data Dictionary:**

|  |  |  |
| --- | --- | --- |
| Columns. | Variables type | Interpretation |
| Item\_Identifier | Nominal | Unique Product ID |
| Item\_Weight | Continuous | Weight of the product |
| Item\_Fat\_Content | Nominal | Low fat/Regular |
| Item\_Fat\_Content\_0 | Continuous | 0=Low fat 1=Regular |
| Item\_Visibility | Continuous | The % of total display area of all products in a store allocated to the product |
| Item\_Type | Nominal | The category to which the product belongs |
| Item\_Type\_R | Nominal | Drinks=Hard Drinks, Soft Drinks Consumable=Dairy, Baking Goods, Breads, Breakfast, Canned, Frozen foods, Fruits and Vegetables, Meat, Seafood, Snack Foods, Starchy Foods Non-Consumable=Health and Hygiene, Household, Others |
| Item\_Type\_R\_0 | Continuous | 0 + 0=Consumable 1 + 1=Drinks 1 + 0=Non-Consumable |
| Item\_Type\_R\_1 | Continuous |
| Item\_MRP | Continuous | Maximum Retail Price of a product |
| Outlet\_Identifier | Nominal | Unique outlet id |
| Outlet\_Establishment\_Year | Nominal | The year in which the store was established |
| Outlet\_Years | Continuous | Number of years since establishment  Outlet\_Years=2013-Outlet\_Establishment\_Year |
| Outlet\_Size | Nominal | Size of the store categorized as High, Medium and low. |
| Outlet\_size\_0 | Continuous | 1 + 1= High 1 + 0= Medium 0 + 0=Small |
| Outlet\_size\_1 | Continuous |
| Outlet\_Location\_Type | Nominal | The type of city in which the store is located |
| Outlet\_Location\_Type\_0 | Continuous | 1 + 1= Tier 1 1 + 0= Tier 2 0 + 0= Tier 3 |
| Outlet\_Location\_Type\_1 | Continuous |
| Outlet\_Type | Nominal | Whether the outlet is just a grocery store or some sort of Supermarket. |
| Outlet\_Type\_00 | Continuous | 0 + 0 + 1 = Grocery Store 0 + 1 + 0 = Supermarket Type1 1 + 0 + 0 = Supermarket Type2 1 + 1 + 1 = Supermarket Type3 |
| Outlet\_Type\_01 | Continuous |
| Outlet\_Type\_11 | Continuous |
| Item\_Outlet\_Sales | Continuous | Sales of the product in the particular store. This is the outcome variable to be predicted. |

**Missing Values**

As we mentioned in data exploration, there are 2 columns have missing data, which are: Item\_Weight and outlet\_size. The 0’s in Item\_Visibility column are considered as missing data.



|  |  |
| --- | --- |
| Columns. | Missing Values |
| Item\_Weight | 1463 |
| Item\_Visibility | 526 |
| Outlet\_Size | 2410 |

We decided to use two different ways to deal with missing data.

1.Delete missing data

2.Recode data based on the mean, median.

Table #

|  |  |  |  |
| --- | --- | --- | --- |
| Columns. | Missing Values | Missing treatment 1 | Missing treatment 2 |
| Item\_Weight | 1463 | Delete | recode using mean |
| Item\_Visibility | 526 | Delete | recode using mean |
| Outlet\_Size | 2410 | Delete | Recode as ‘Small’ |

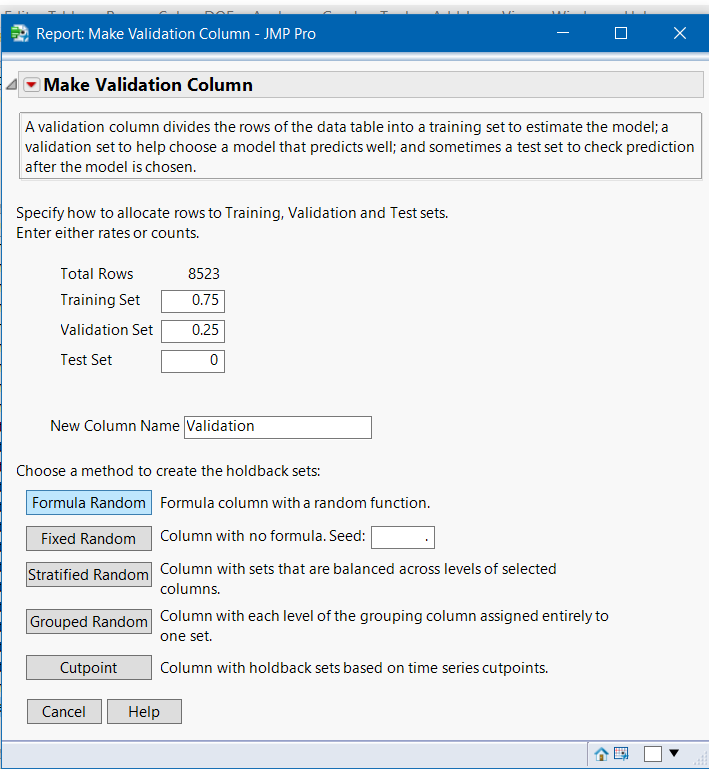
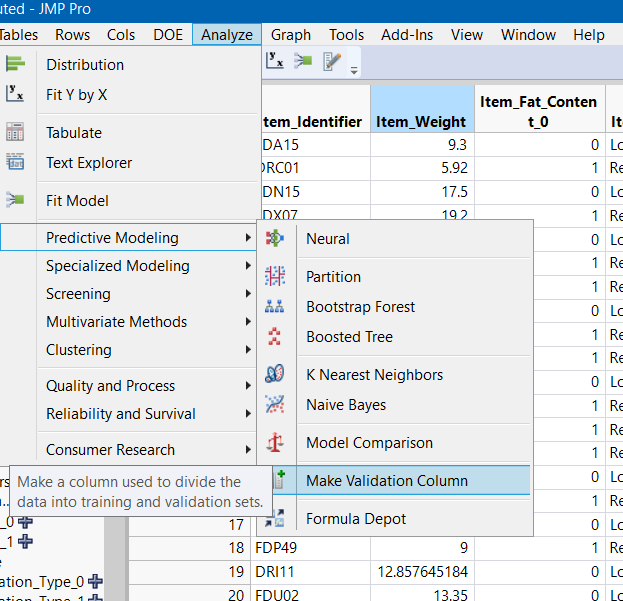
* Hence, we created two data sets, one with no missing values and the other with recoded missing values.

# Modelling-Recoded data set.

## **Data Partition**

Creating a validation column:

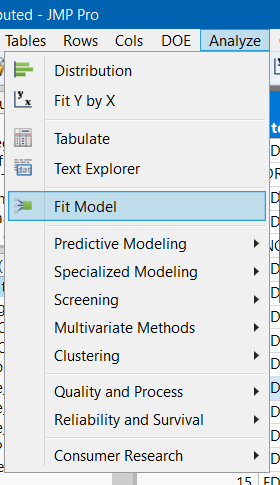
Go to Analyze>Predictive modeling>Make Validation column

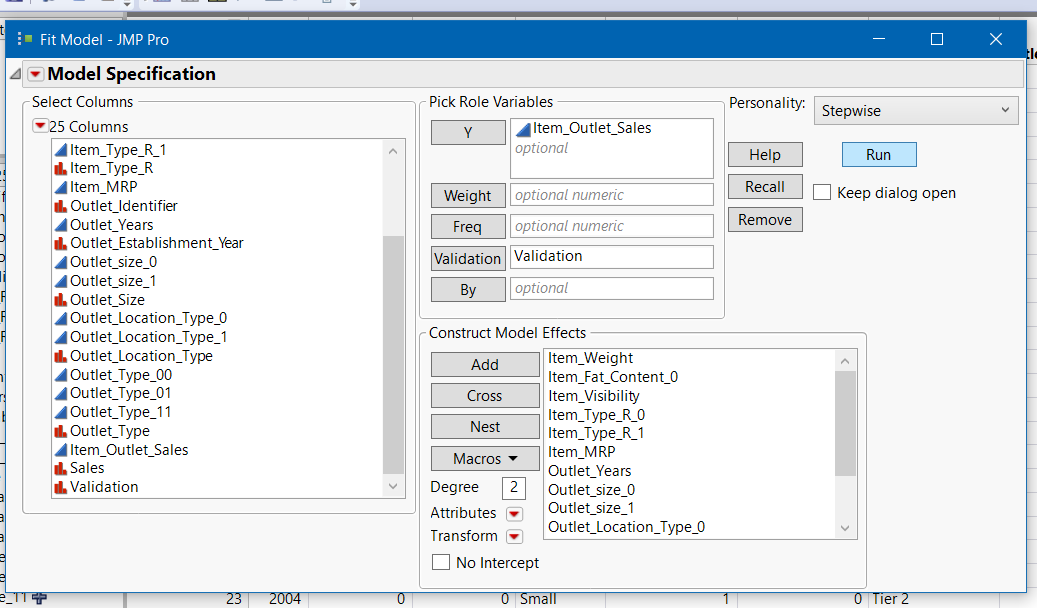
****

* Once the new column is created, we start building models on the data set.

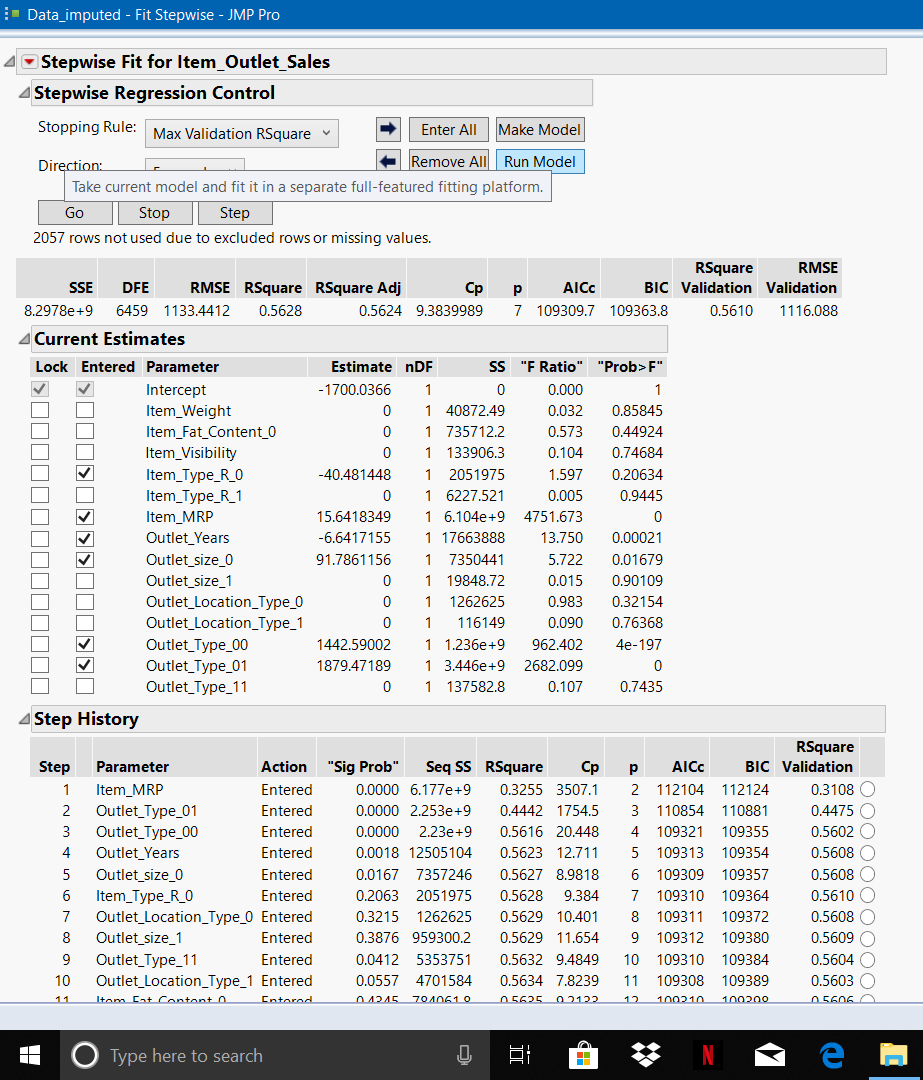
## **Stepwise Regression:**

Go to **Analyze>Fit Model**



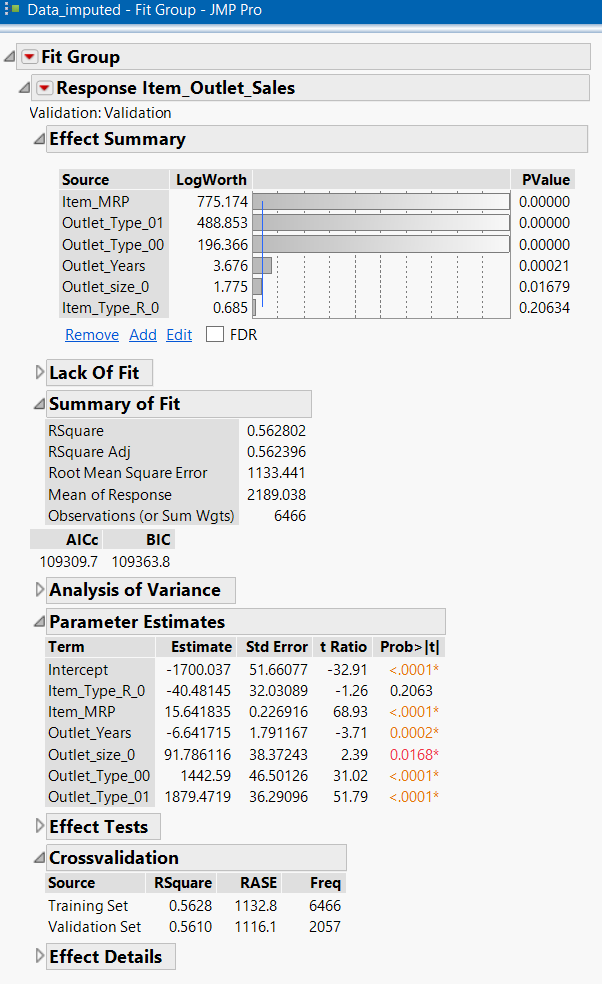


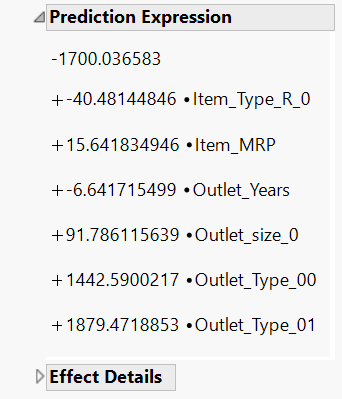
Click **GO,** then **RUN**

****

Then we get the below result,

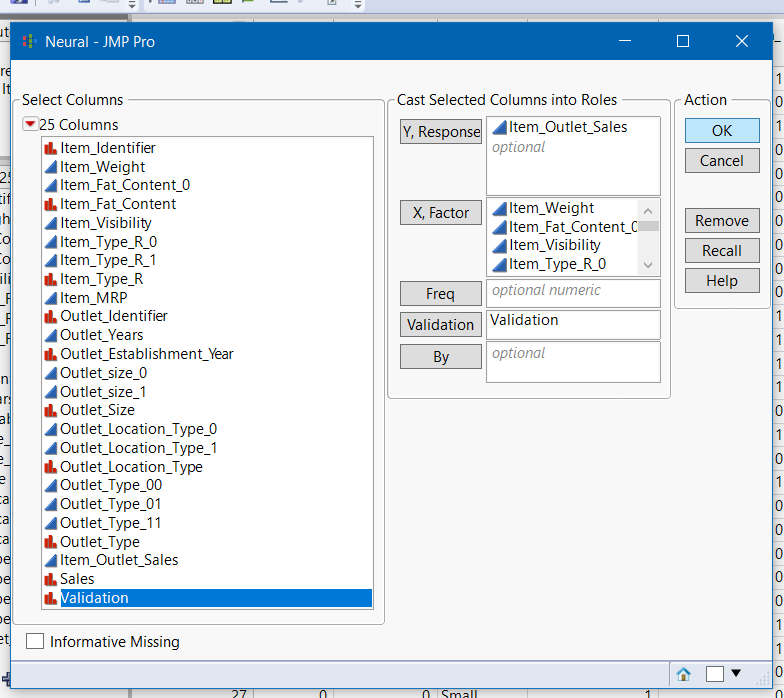
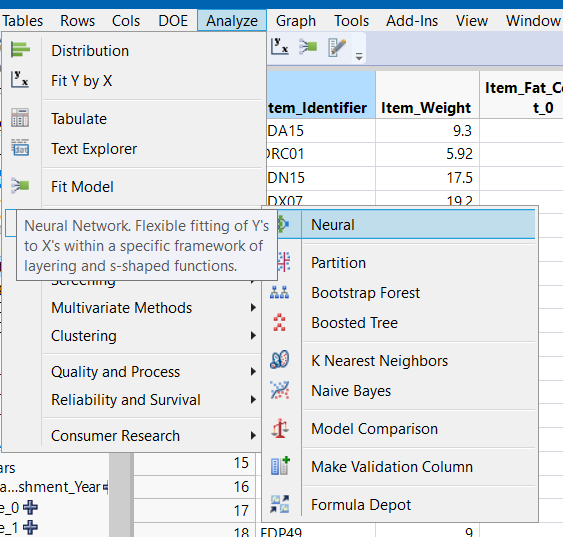
We **saved the script to data set.**



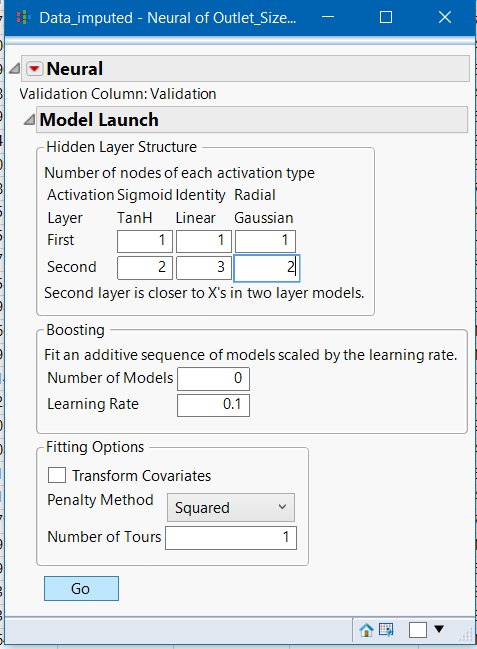


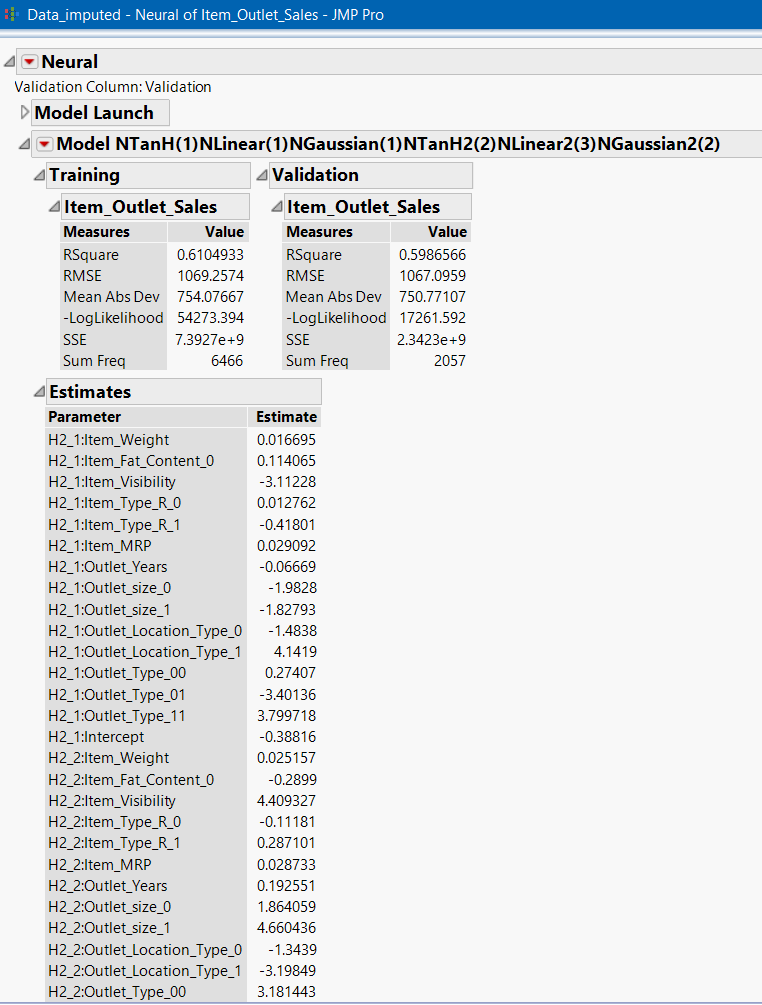
## **Neural Net Model:**

Go to **Analyze>Predictive Modeling>Neural**



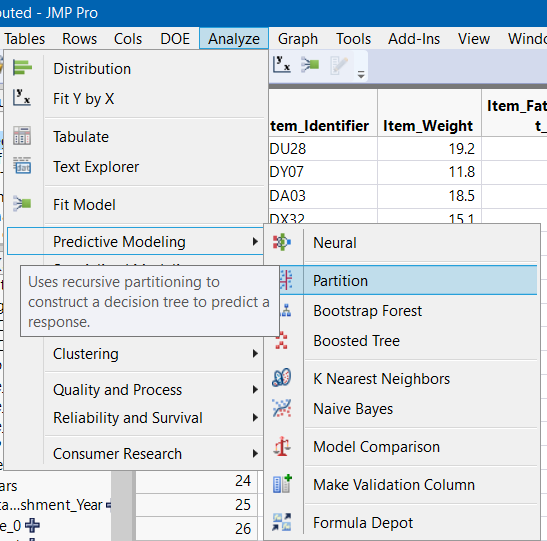
We tried multiple combinations of functions, but the below combination gave us the optimum result. We **saved the script to data set.**

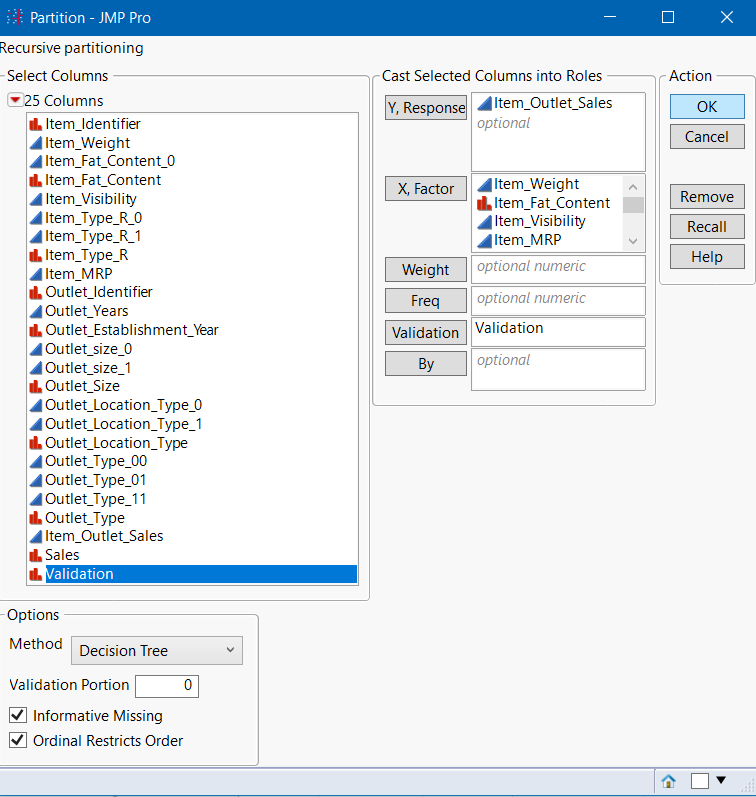




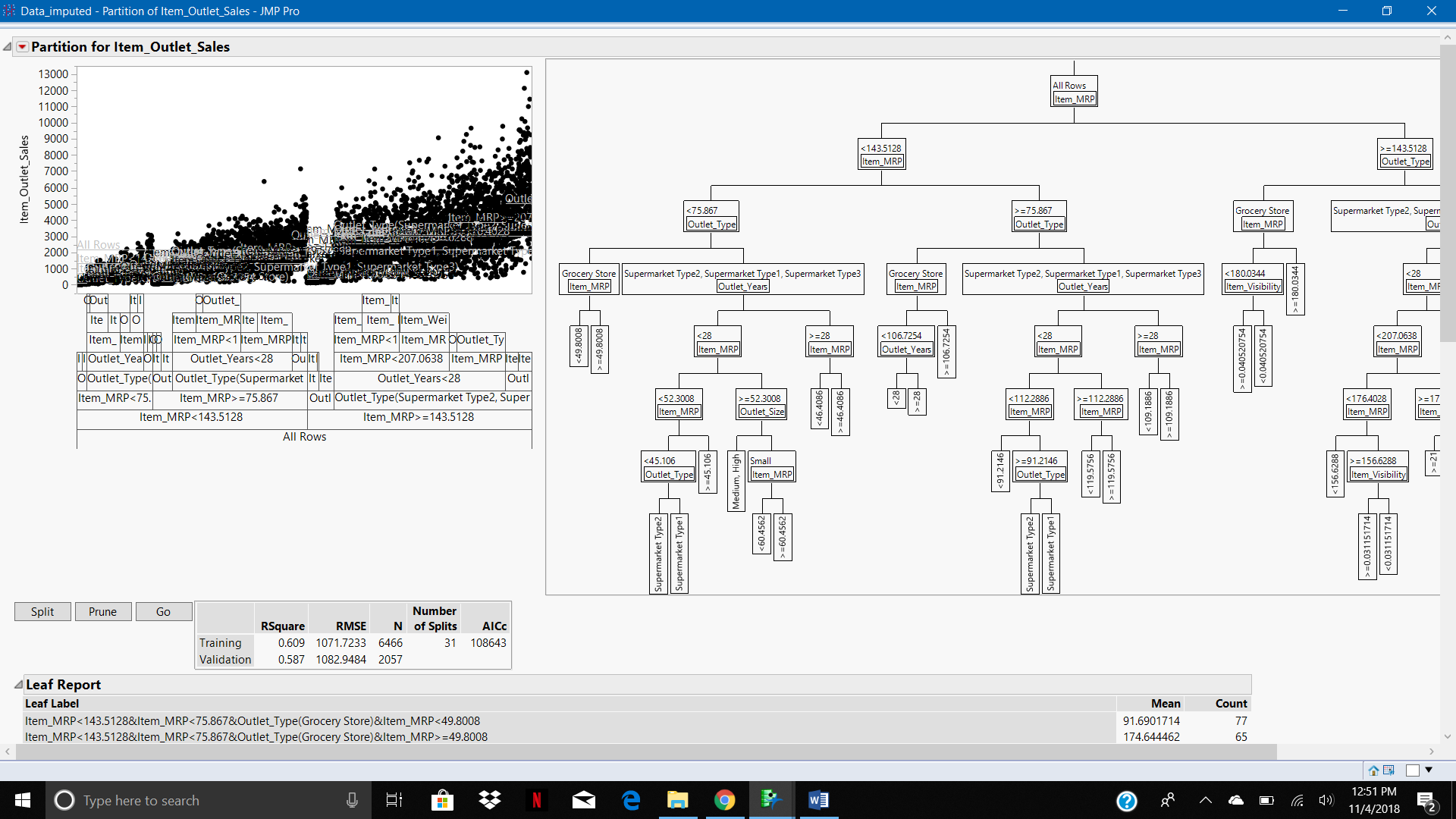
## **Decision Tree Model:**

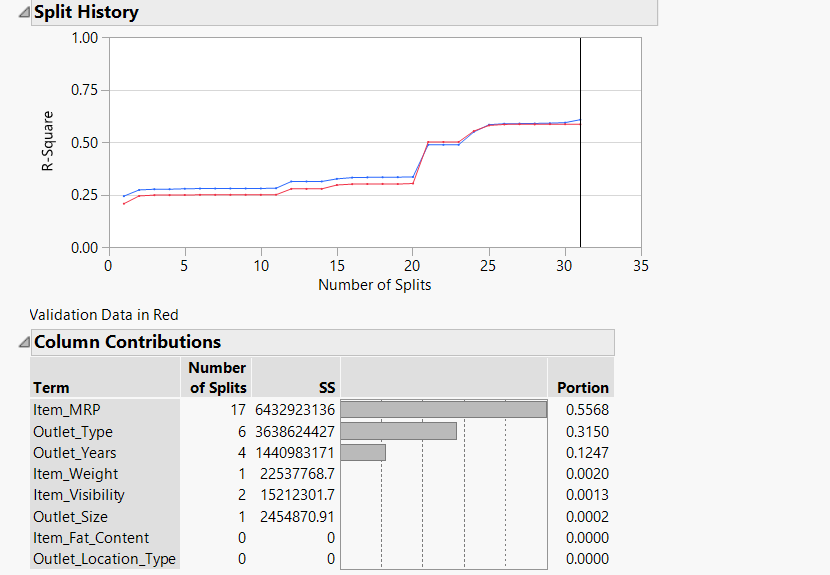
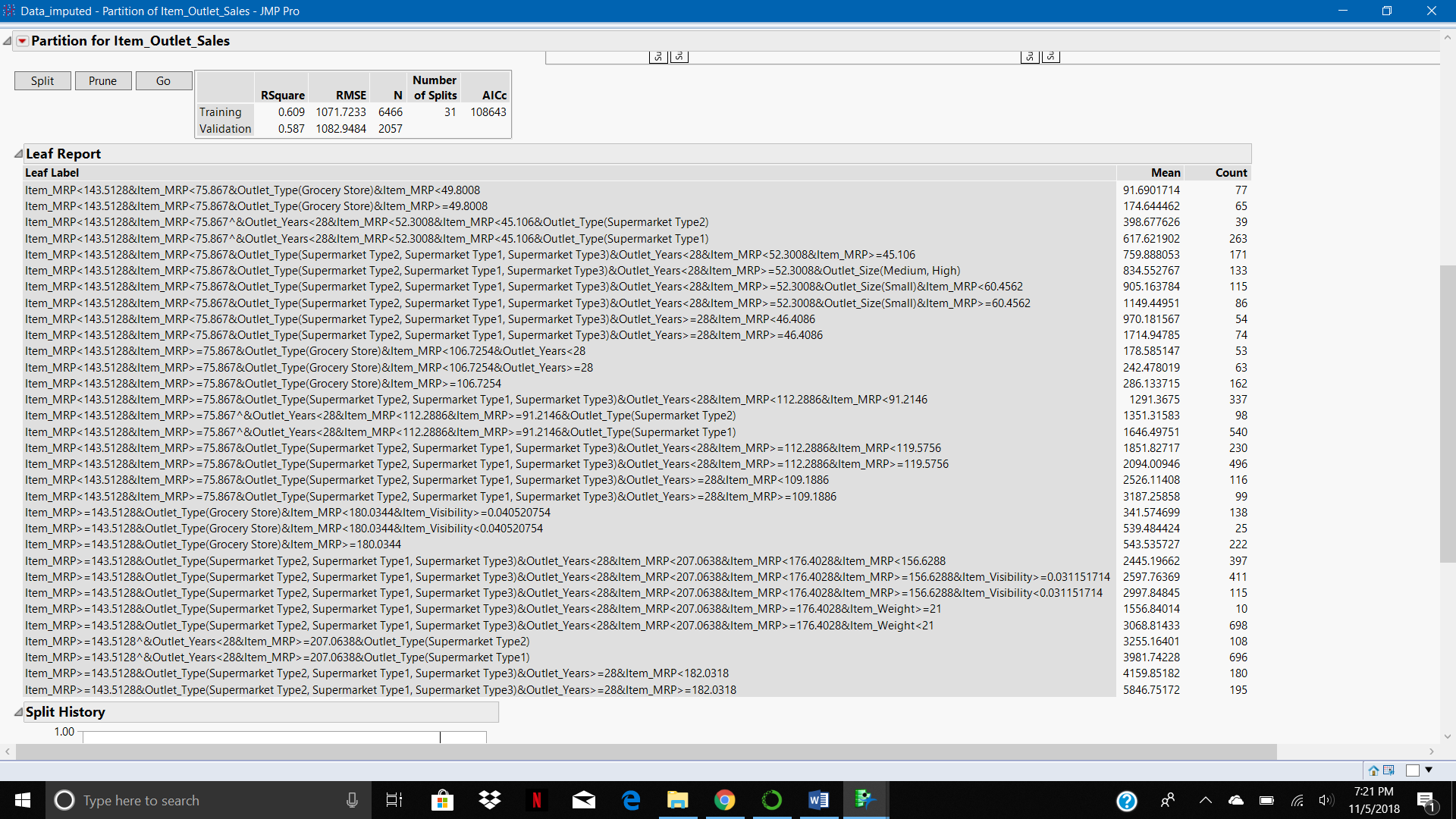
Go to **Analyze>Predictive Modeling>Partition**





Click **GO,** then we get the below results. We **saved the script to data set.**

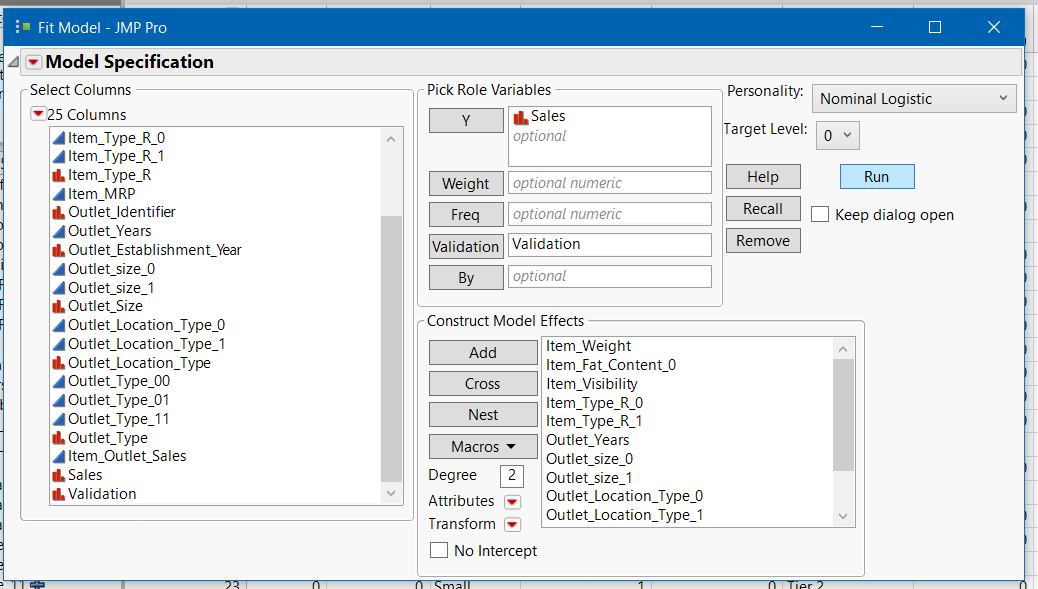


****

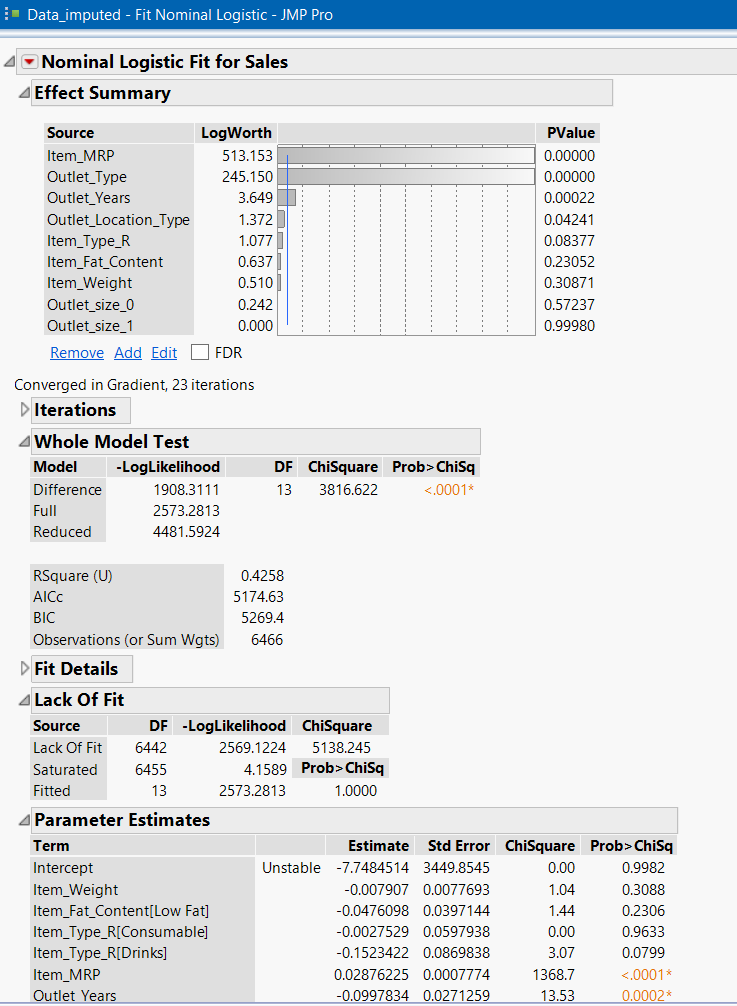
* Though Decision trees and neural net can be applied to categorical target variables, we used the original continuous target variable because it made sense to use the original data than the created dummy variable in order to get better predictions.

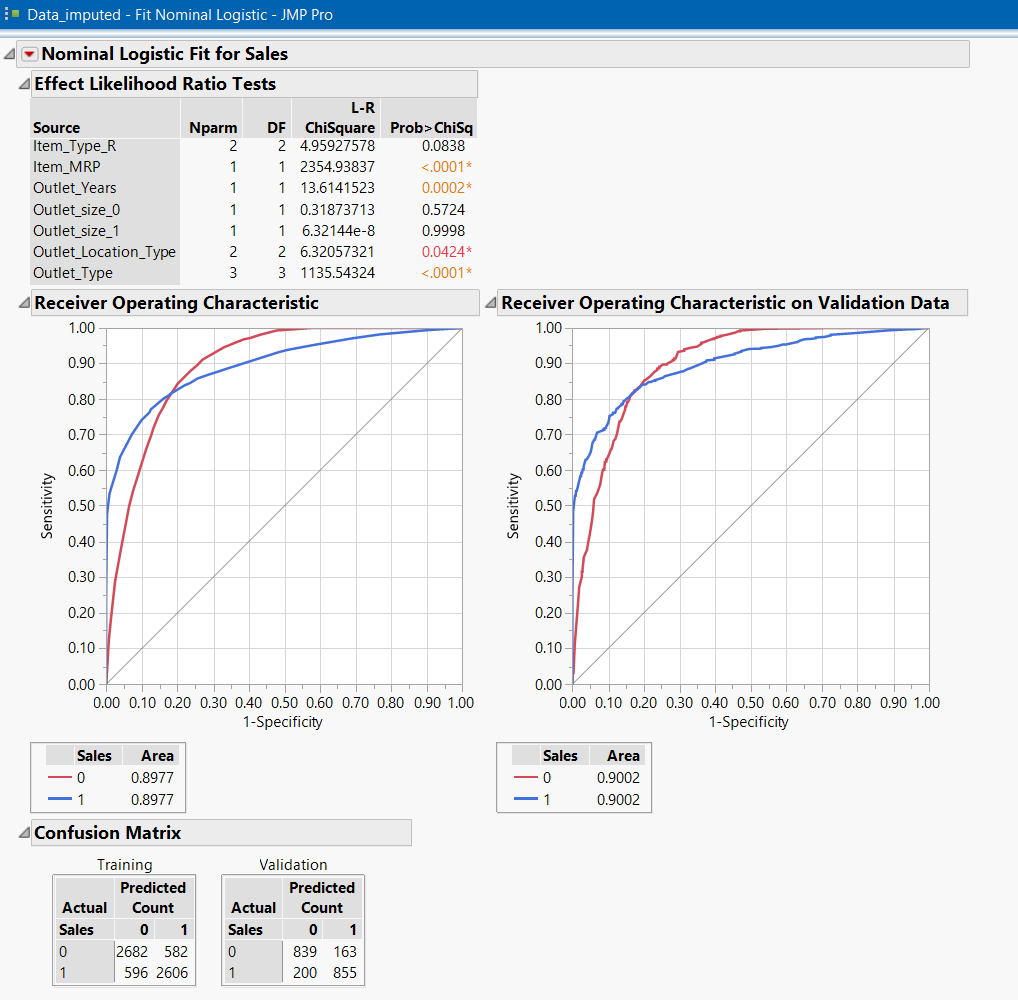
## **Logistic Regression:**

Go to **Analyze>Fit Model.** Here, we used the created categorical variable as the target variable.



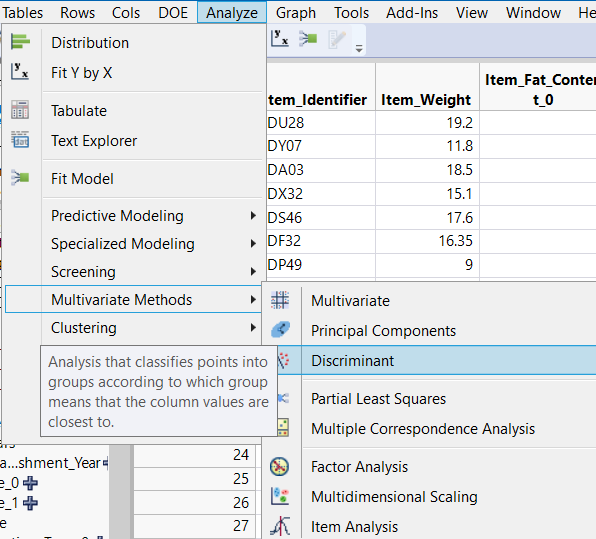
These are the Predictions we got, we **saved the script to data set.**

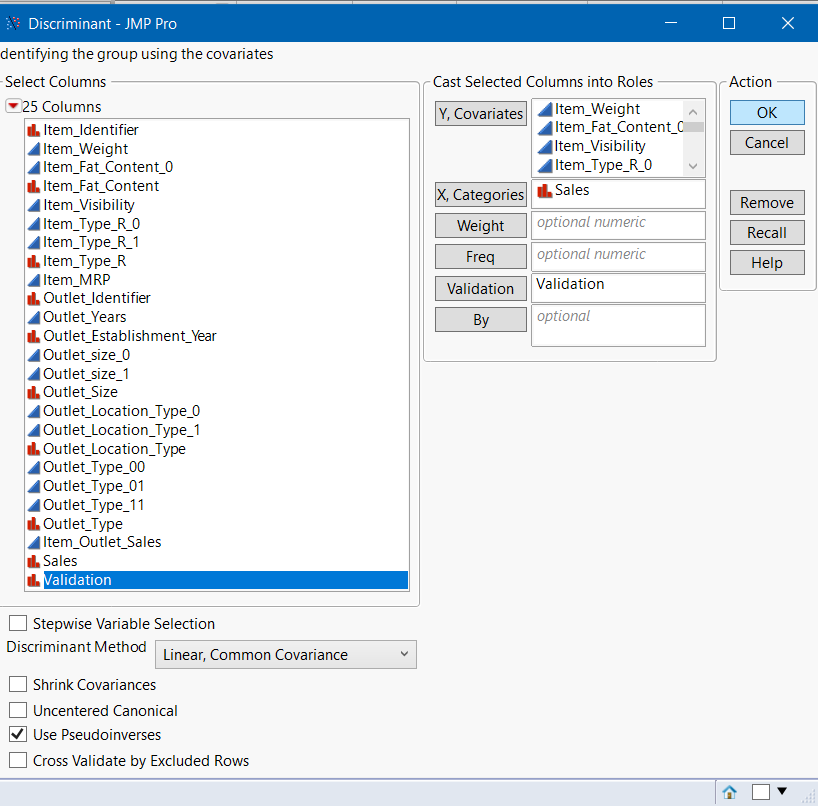




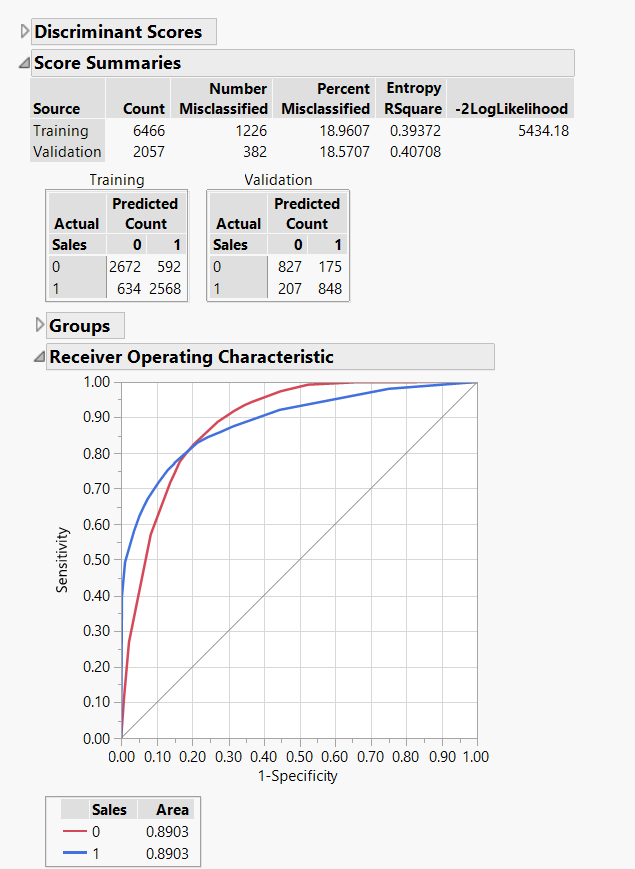
## **LDA (Discriminant Analysis):**

Go to **Analyze>Multivariate Methods>Discriminant.** Target variable is categorical.





These are the predictions we got, we **saved the script to data table.**



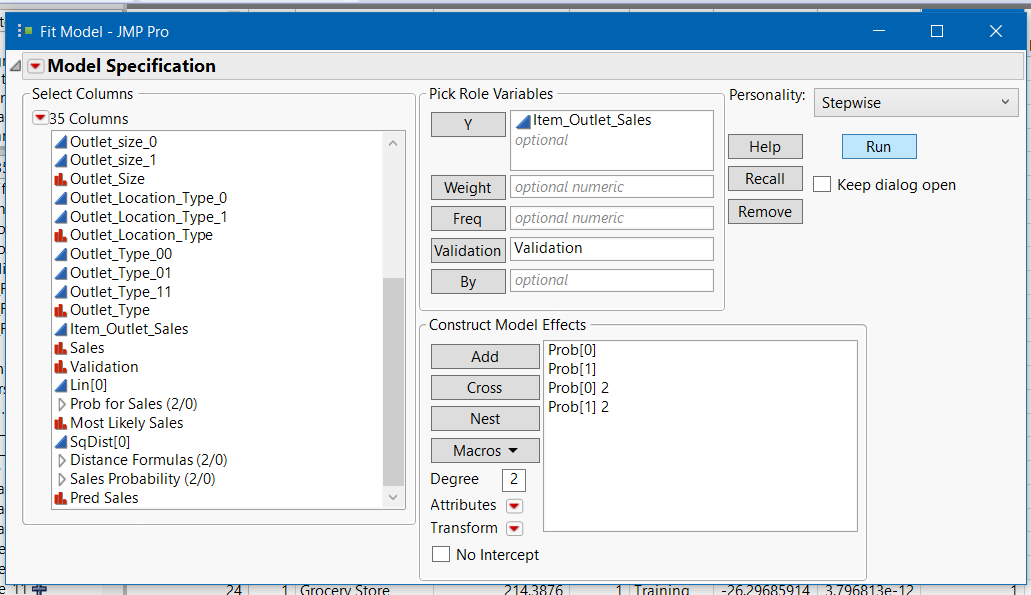
* **We saved the prediction formulas for all the models.**

## **Building Ensemble Models**

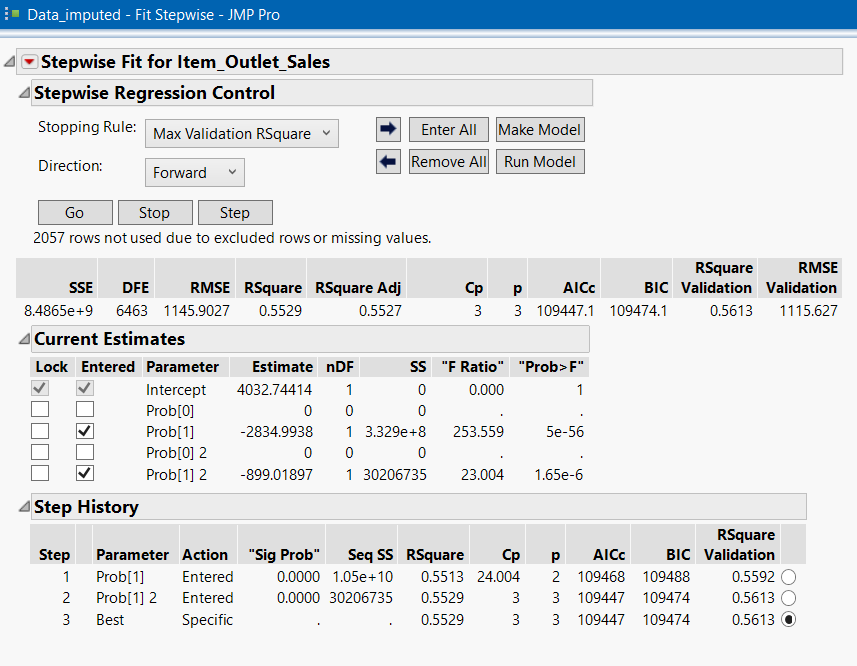
The probabilities of Logistic Regression and LDA are fed as predictor variables for different ensemble models.

### **Ensemble Stepwise Regression:**

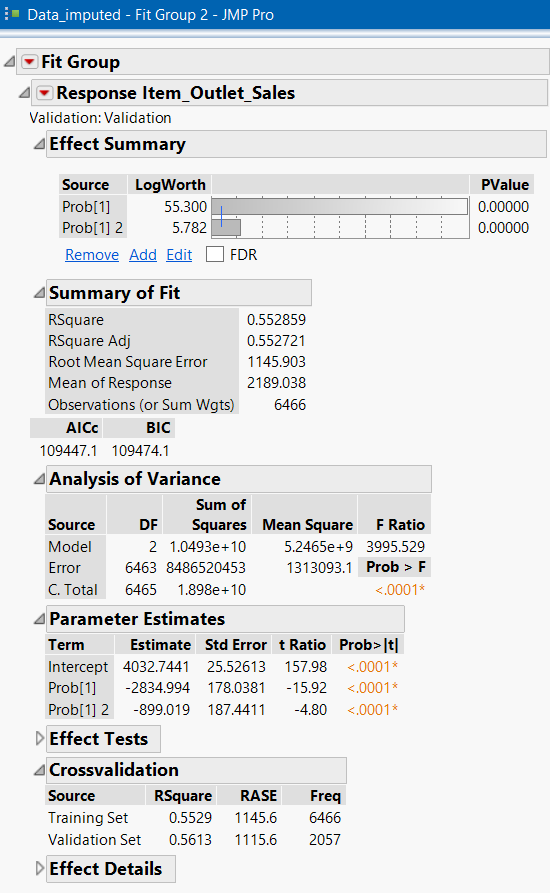
We followed the following steps,



Click **GO** then **Run model**

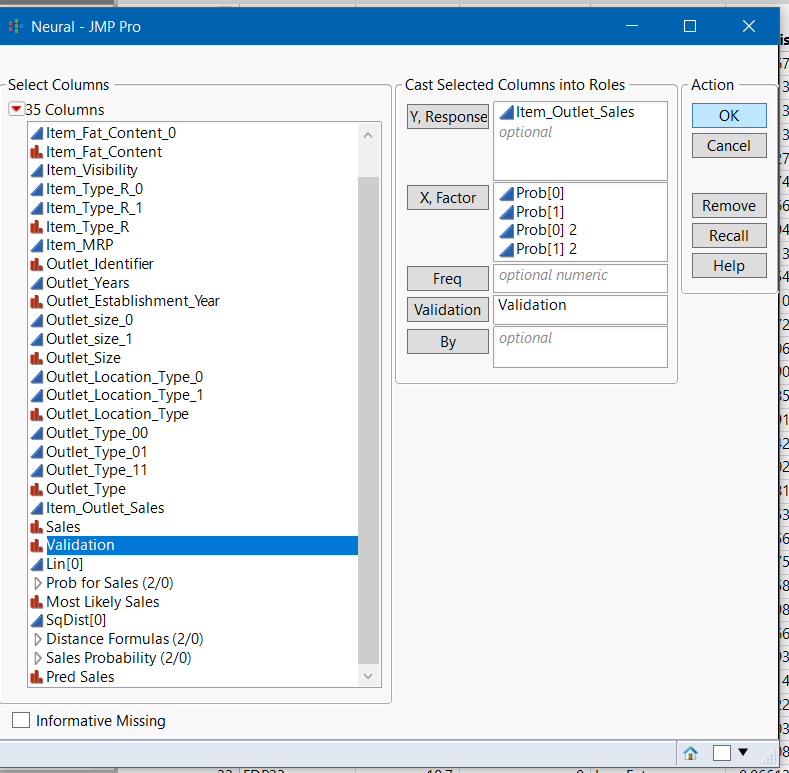


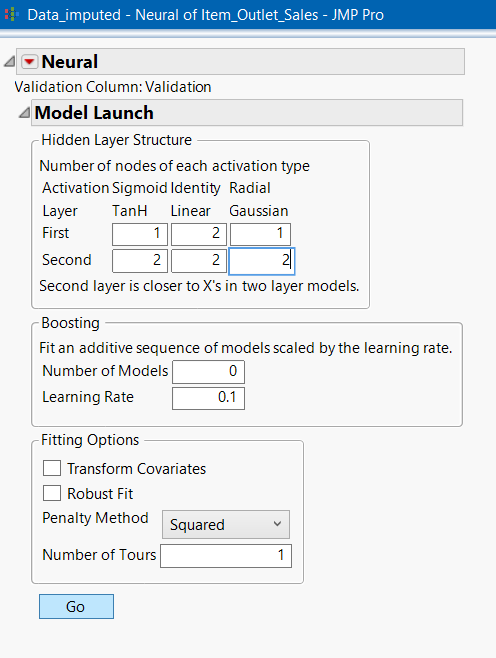
We got the below results,



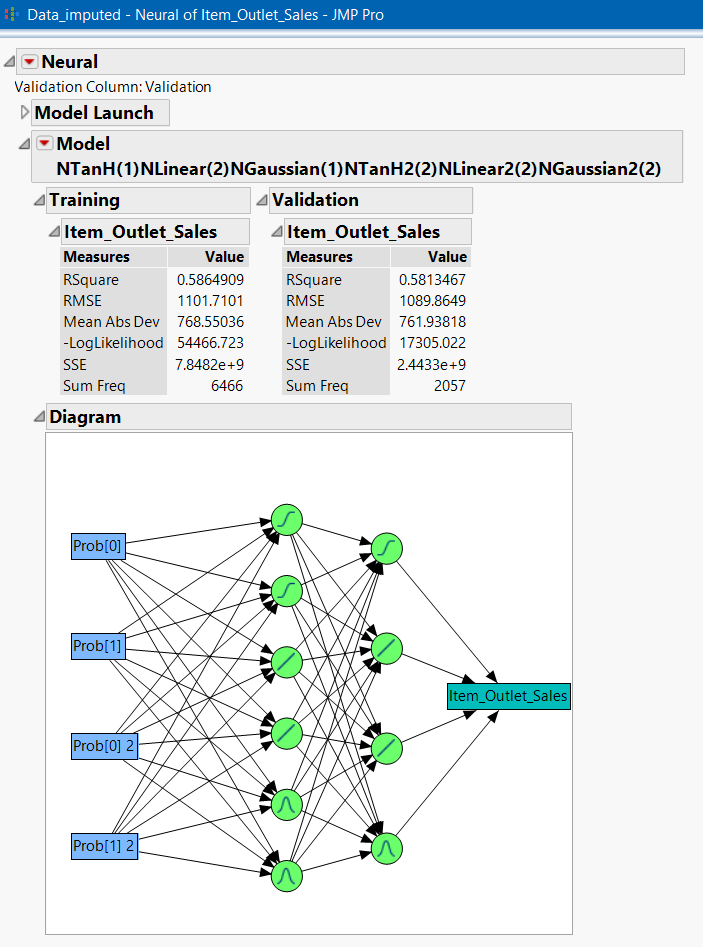
### **Ensemble Neural Net:**

We followed the following steps,



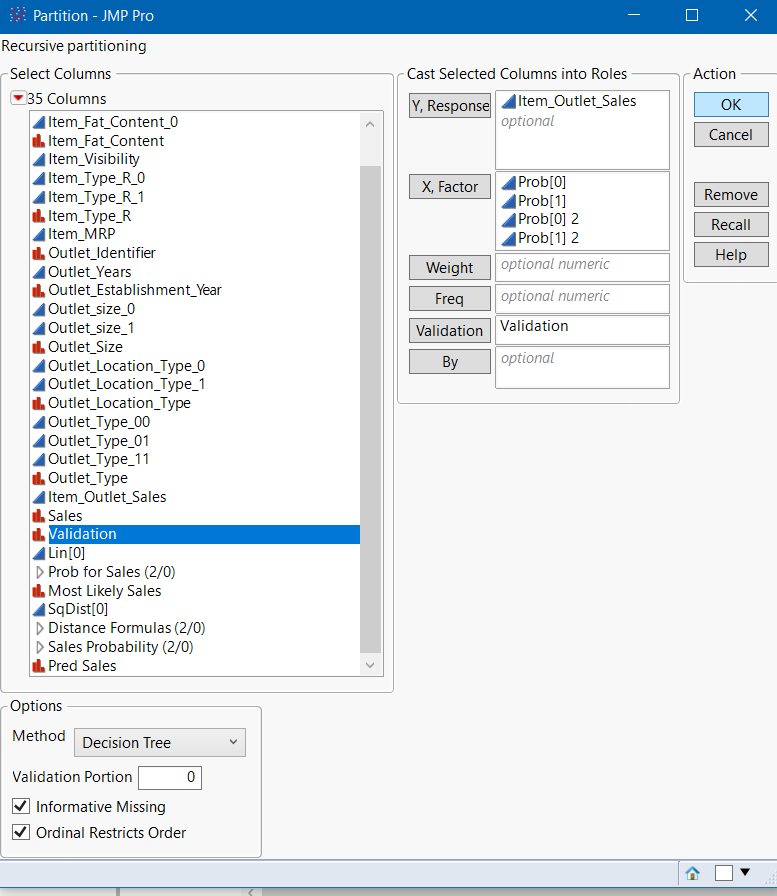


These are the results we got,

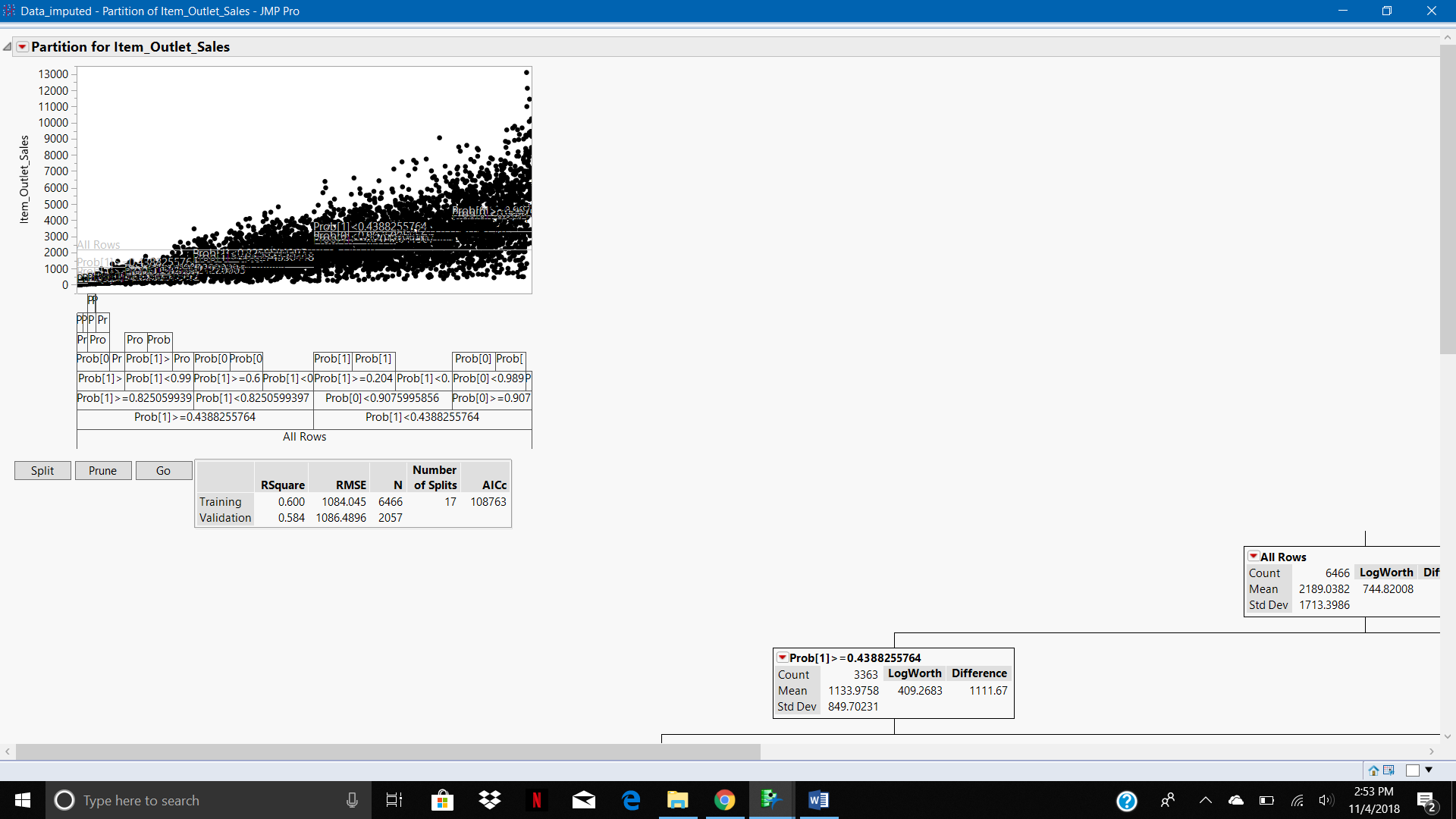


### **Ensemble Decision Tree:**

We took the following steps,



This is the result we got, after clicking **GO.**



### **Summary Table of Imputed data set:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model Type** | **R-Square** | **RMSE** | **AICs** | **BICs** |  |
| Stepwise Reg | 0.562 | 1133.44 | 109309.7 | 109363.8 |  |
| Logistic reg | 0.425 |  | 5147.63 | 5269.4 |  |
| Decision tree | 0.587 | 1082.94 | 108643 |  |  |
| LDA | 0.40 |  |  |  |  |
| Neural Network | 0.5986 | 1067.09 |  |  |  |
| Ensemble DT | 0.584 | 1086.48 | 108763 |  |  |
| Ensemble Reg | 0.5613 | 1145.9 | 109447.1 | 109474.1 |  |
| Ensemble Neural | 0.581 | 1089.86 |  |  |  |

The best model for this data set is Decision Tree because it performs better than all the models except the neural net model. Neural nets can’t be explained to the clients, but decision trees can be. Therefore, we are going with the Decision tree model.

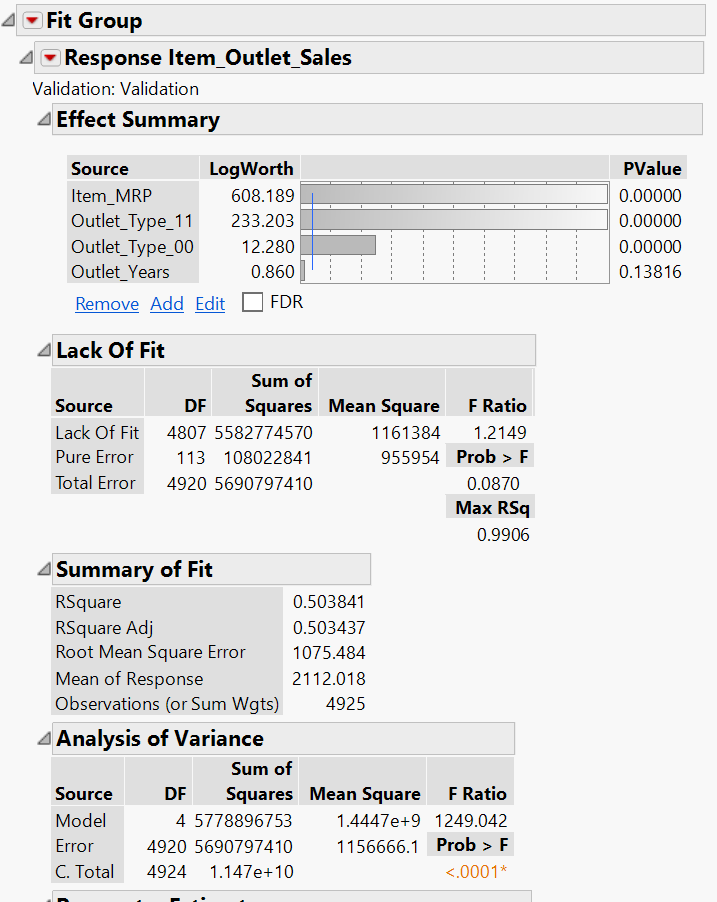
# Modelling-No Missing Values

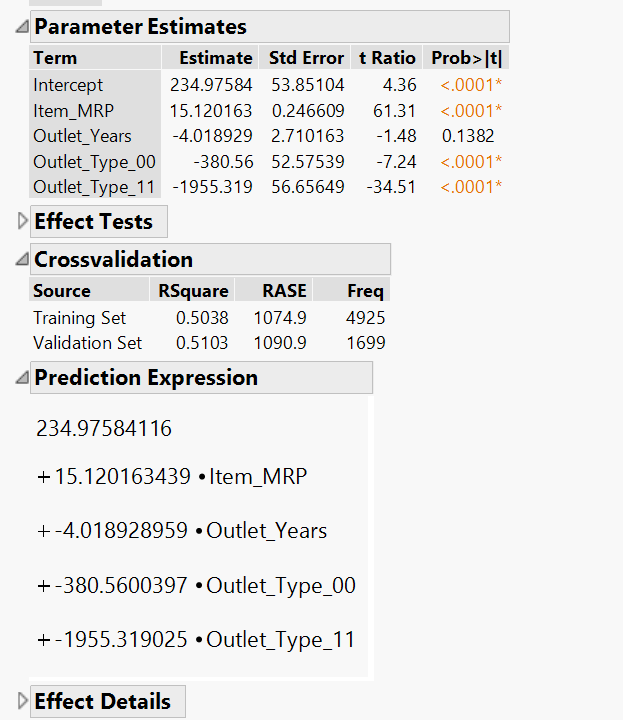
We built the same models as above using the data set with deleted missing values.

The same steps were followed for this data set as well.

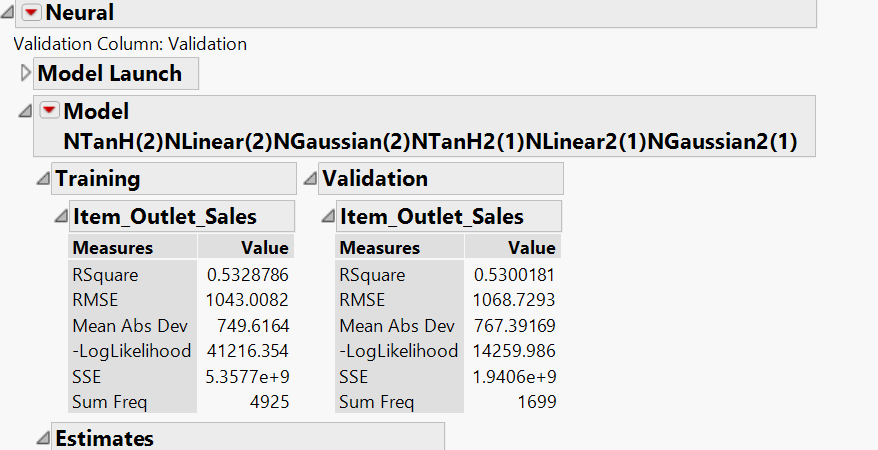
**Stepwise Regression:**

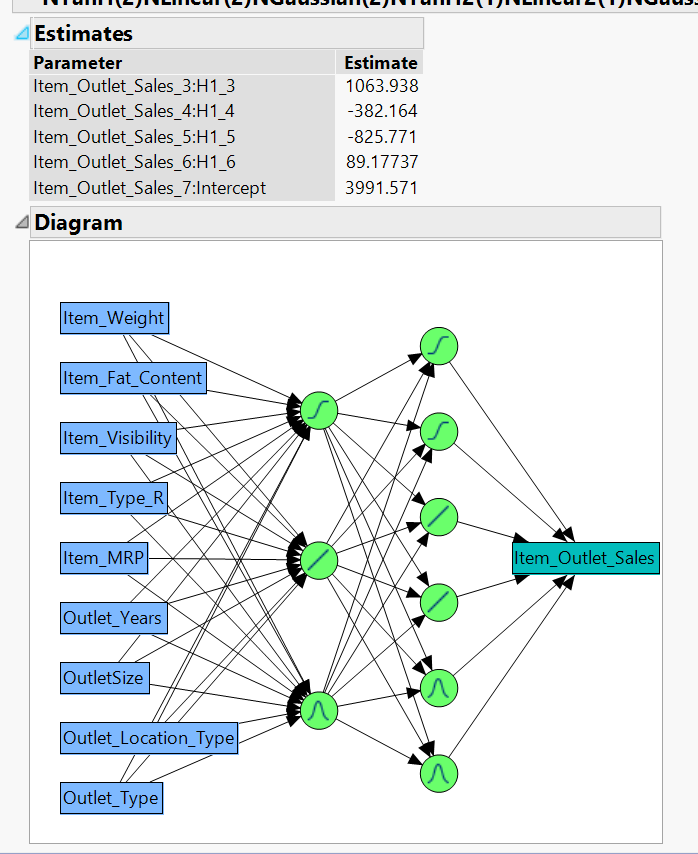


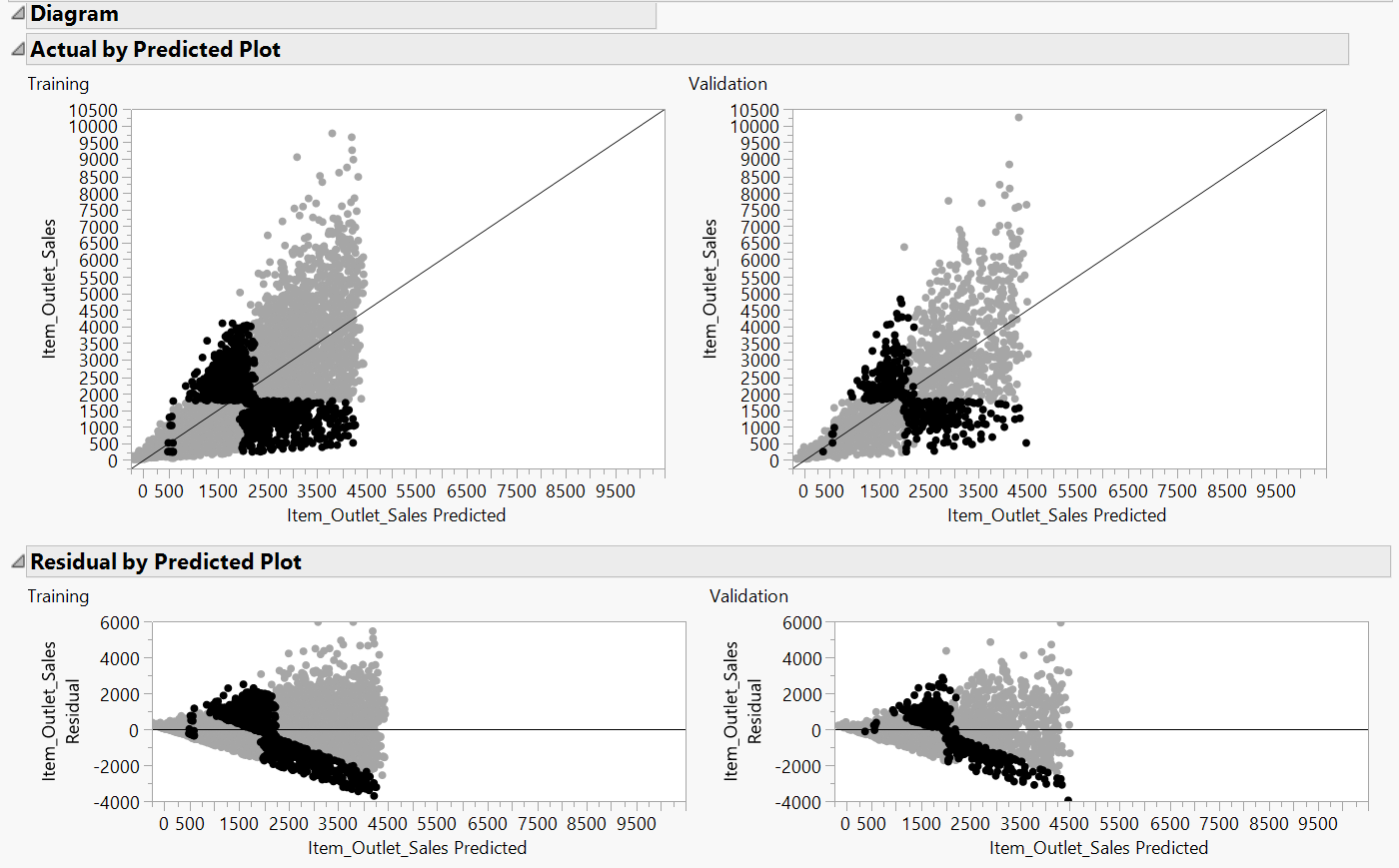




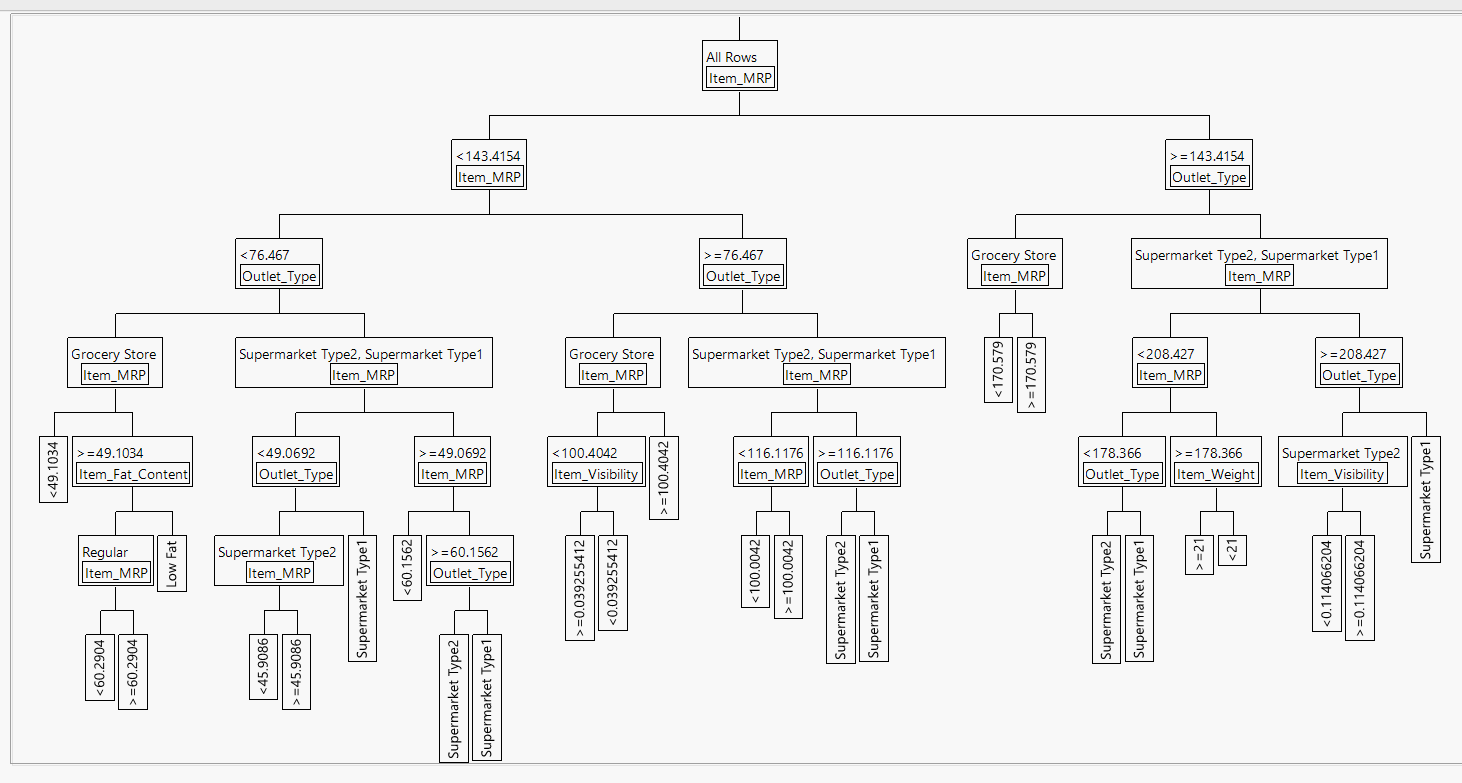
## **Neural Net Model:**

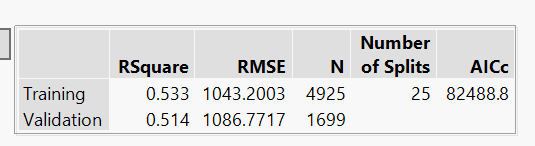


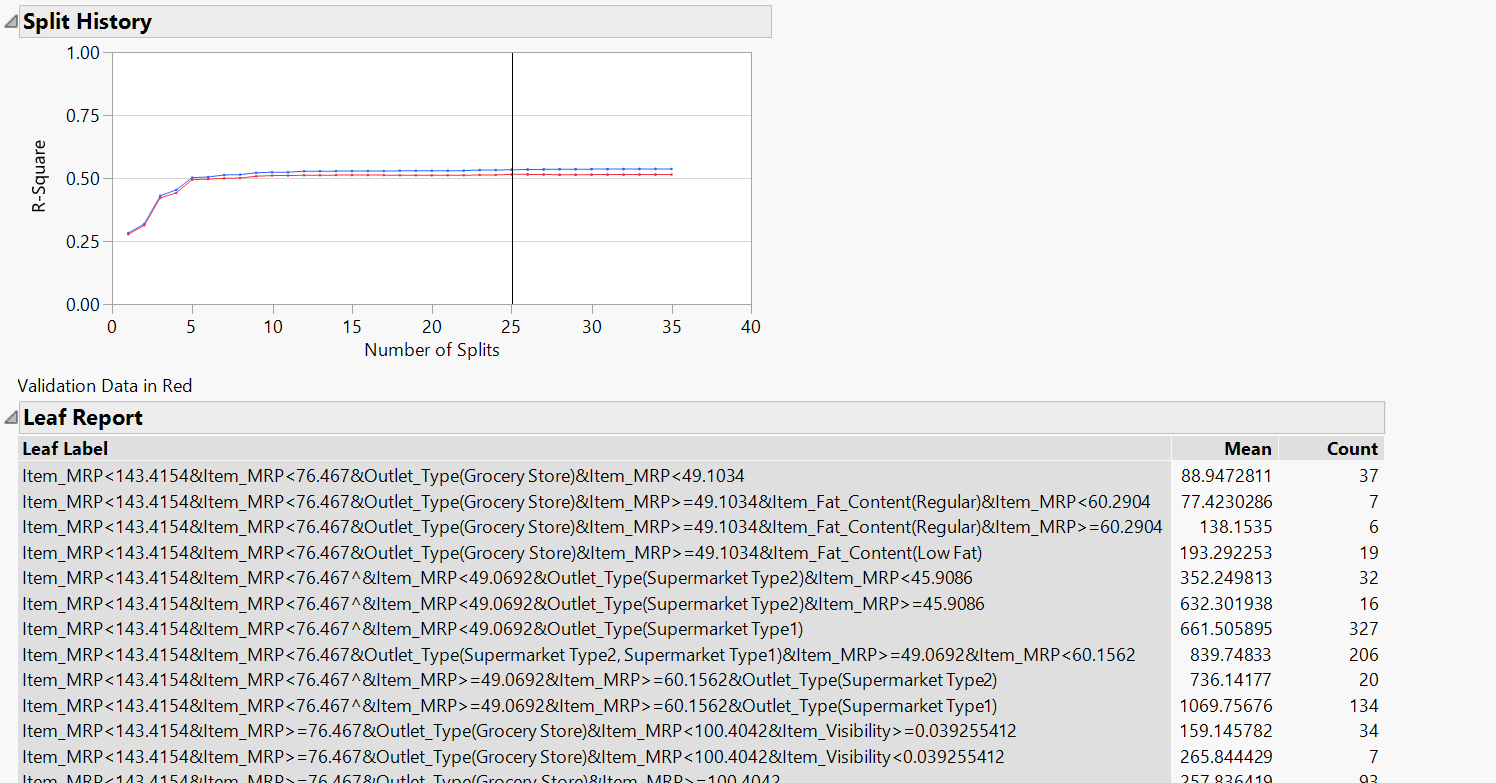




## **Decision Tree:**

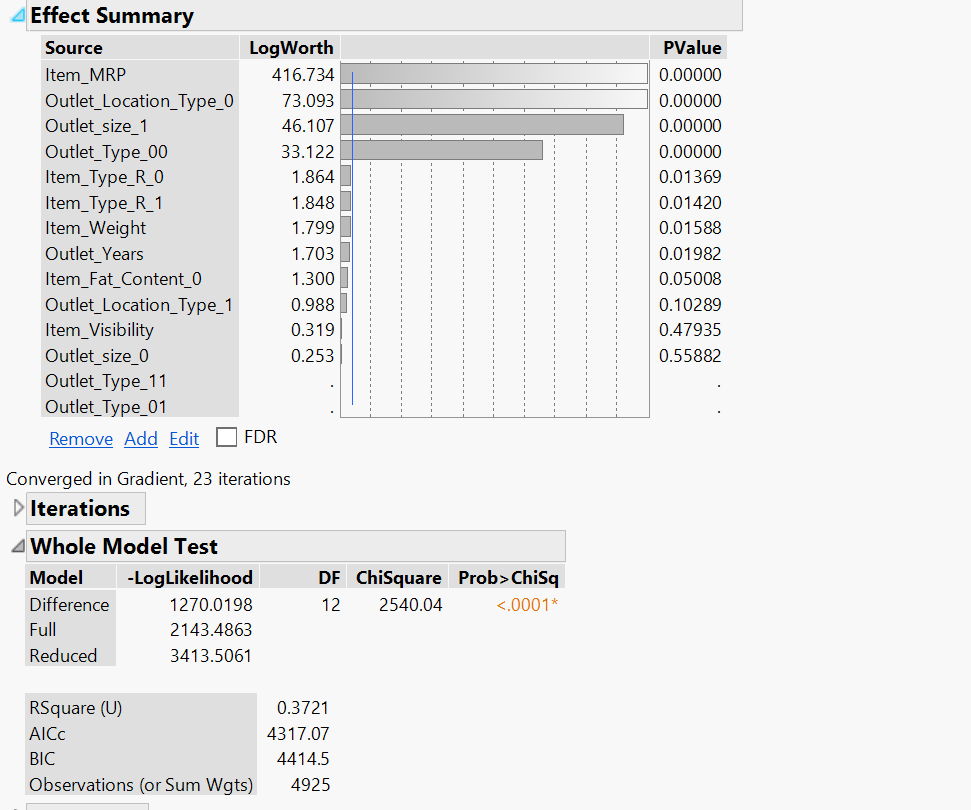


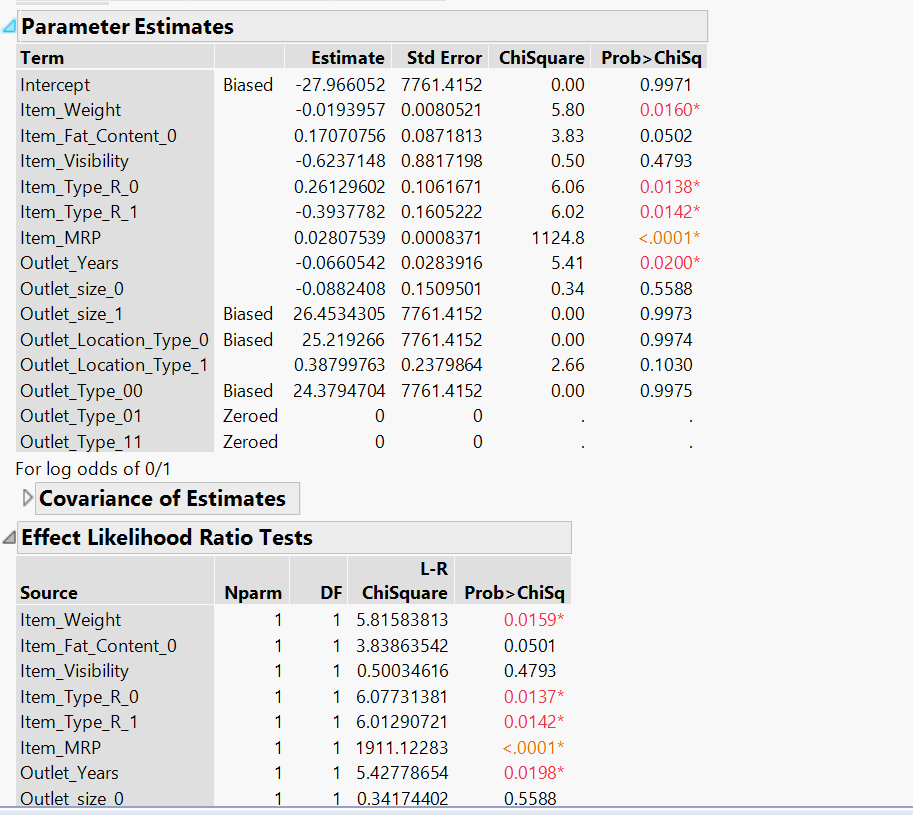






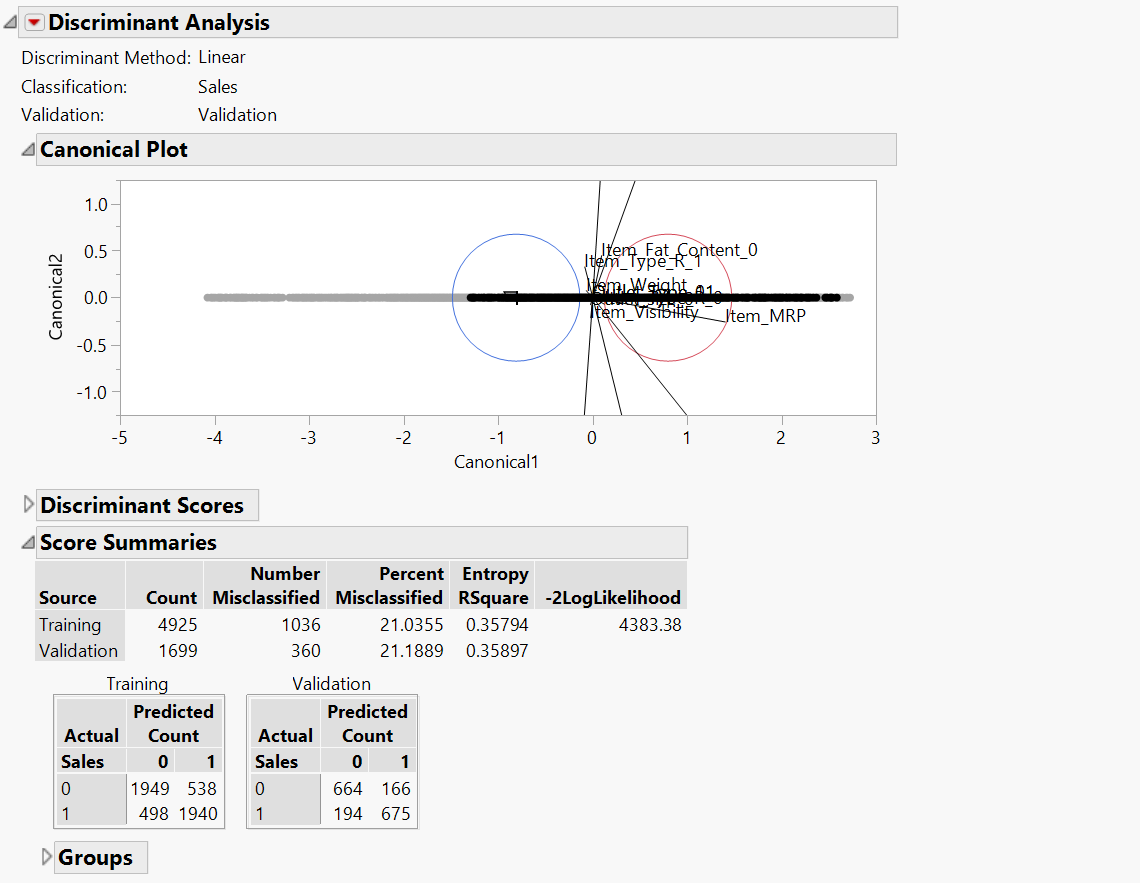
## **Logistic Regression:**





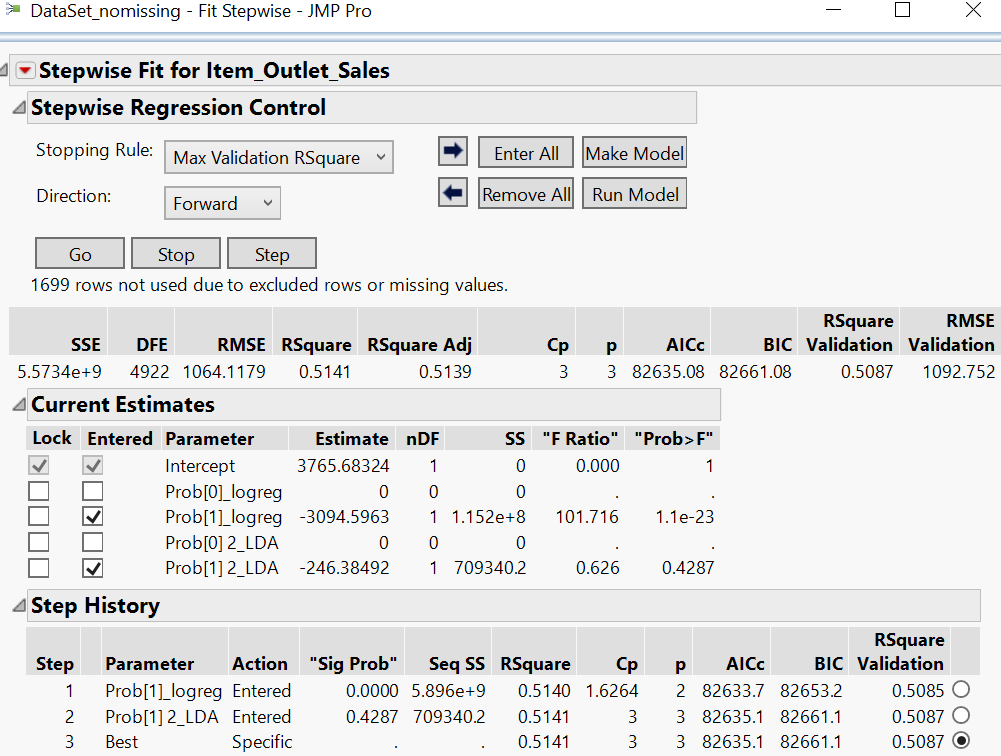


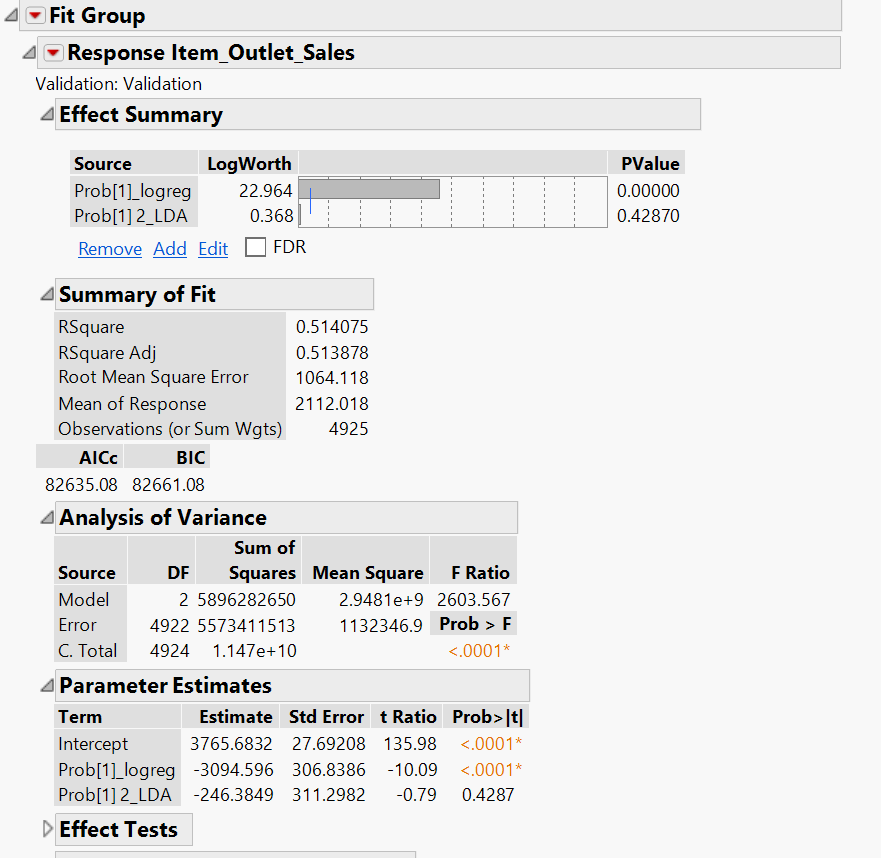
## **LDA (Discriminant Analysis):**



## **Building Ensemble Models**

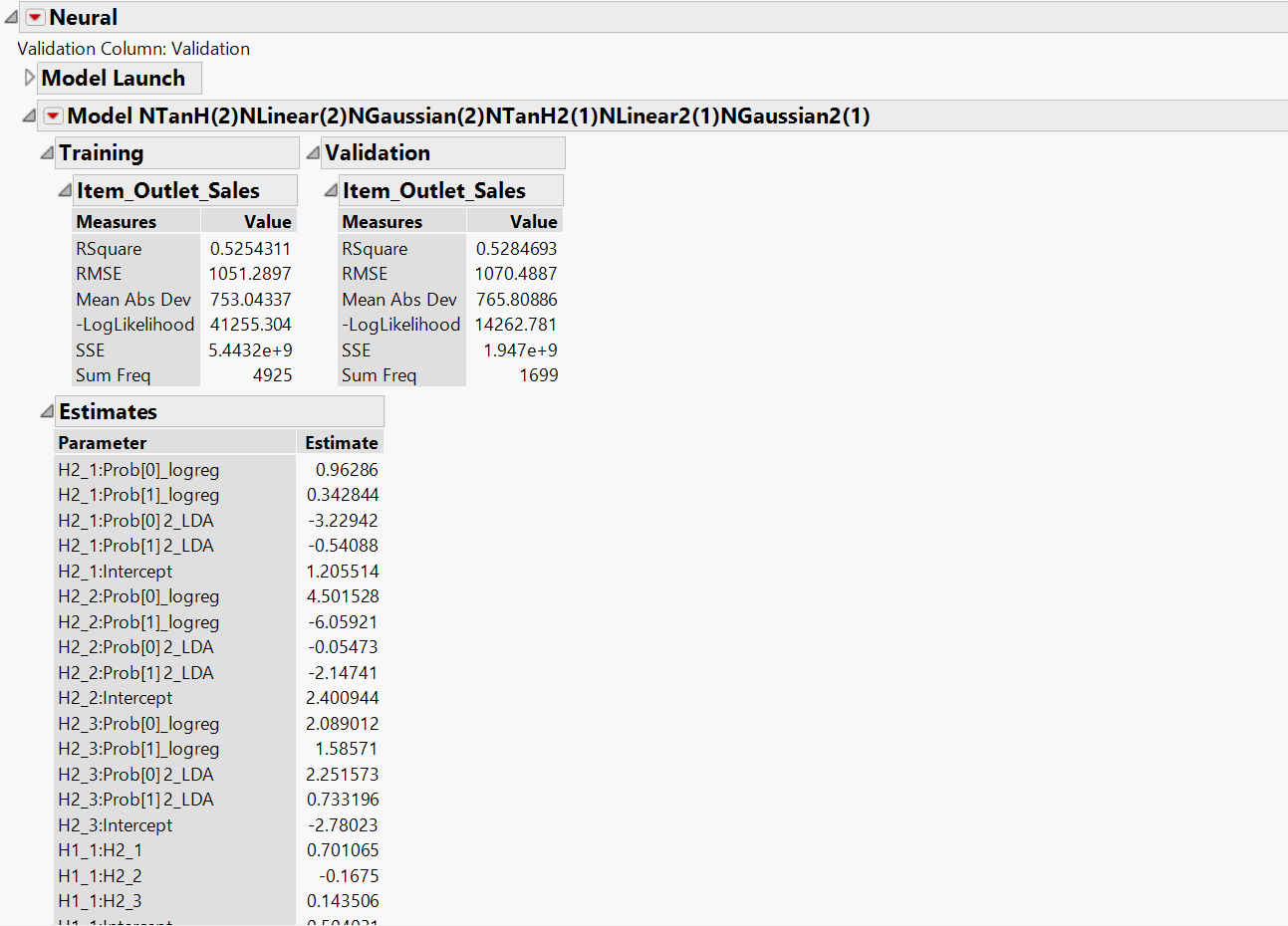
### **Ensemble Stepwise Regression Model:**

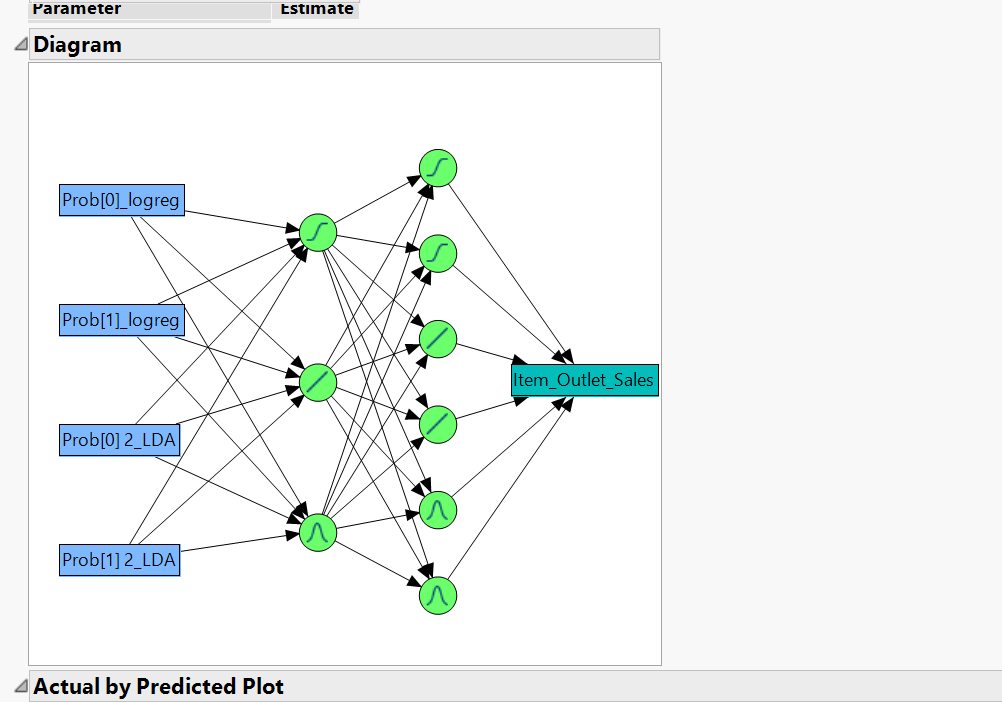


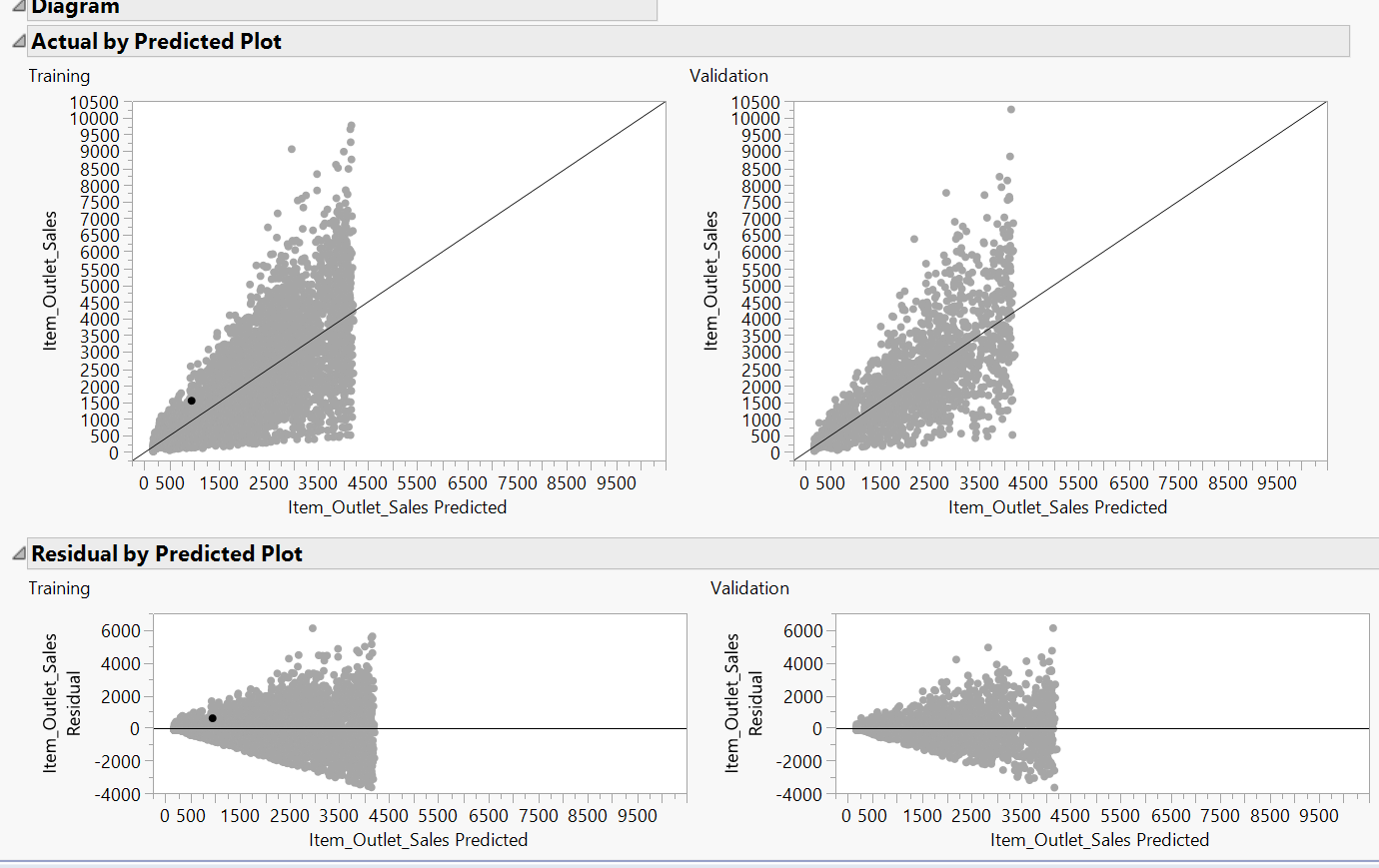




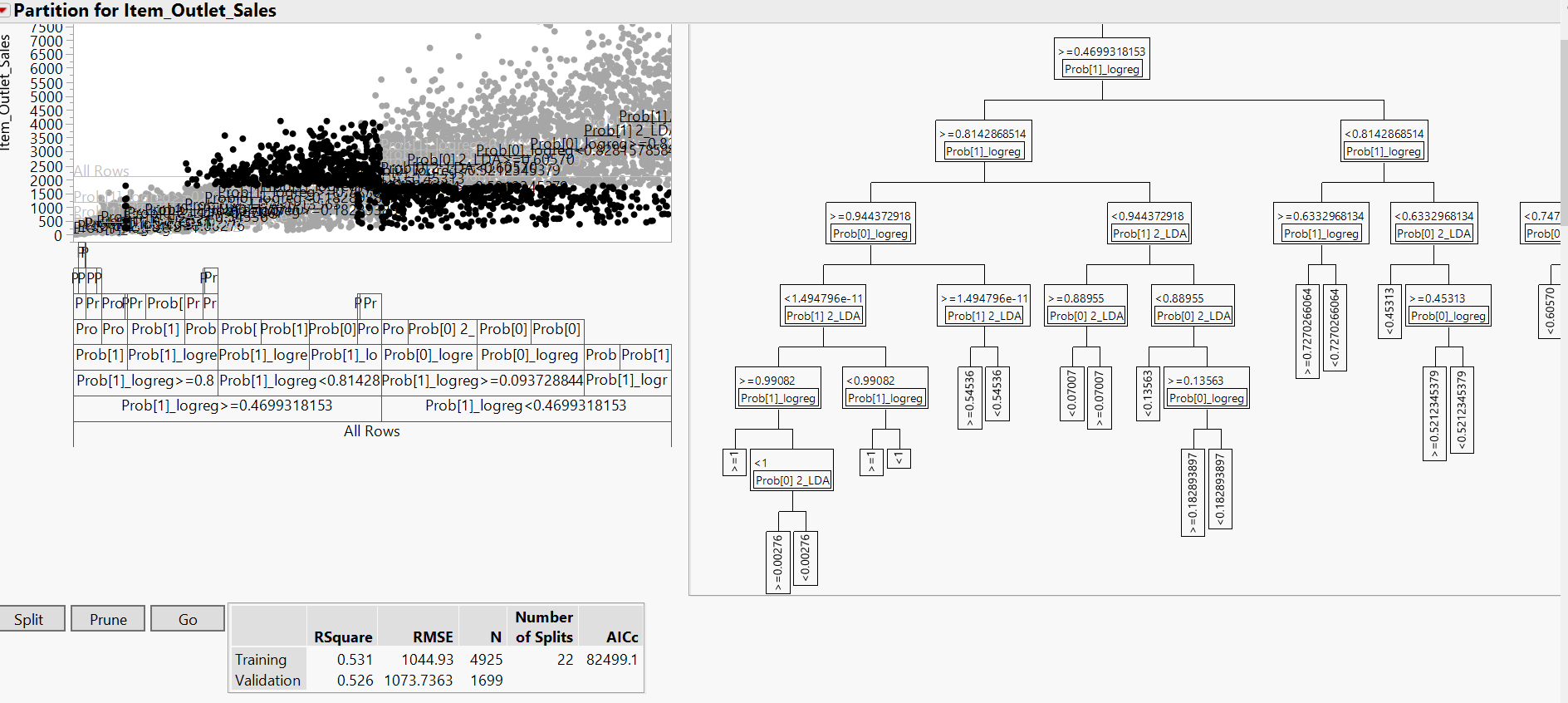
### **Ensemble Neural Network:**



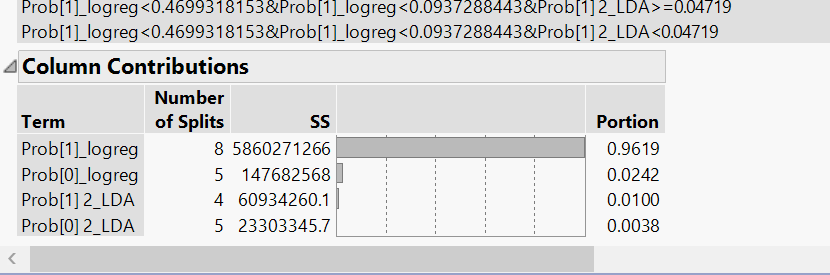




### **Ensemble Decision Tree:**







### **Summary Table of No missing values data set:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model Type** | **R-Square** | **RMSE** | **AICs** | **BICs** |  |
| Stepwise Reg | 0.503841 | 1075.484 | 82741.74 | 82780.74 |  |
| Logistic reg | 0.3721 |  | 4317.07 | 4414.5 |  |
| Decision tree | 0.514 | 1086.717 | 82488.8 |  |  |
| LDA | 0.35897 |  |  |  |  |
| Neural Network | 0.5300181 | 1068.7293 |  |  |  |
| Ensemble DT | 0.526 | 1073.7363 | 82499.1 |  |  |
| Ensemble Reg | 0.514075 | 1064.118 | 82635.08 | 82661.08 |  |
| Ensemble Neural | 05284693 | 1070.4887 |  |  |  |

From the above table, it can be inferred that the neural net and ensemble decision tree models perform the best. But because of the complexity of explaining the model, we are selecting the decision tree as the better model.

# Recommendations:

## **For the quality of data:**

1. According to the data dictionary, ‘Item\_Identifier’ is a unique product id, but it actually repeats quite a few times. The same id is used for different ‘Outlet\_Establishment\_Year’ which creates confusion. Clear explanation of the variable is suggested.
2. ‘Item\_Visibility’ has many values as ‘0’. According to the data dictionary it is the percentage of visibility of the product, but it is logically impossible for a product to have sales if it is not visible at all. Clear explanation of the scale and basis of calculation is required.
3. There were missing values in ‘Item\_Weight’ and ‘Outlet\_size’. Reaching out to the source to collect this information is suggested.
4. An updated data dictionary with clear explanation of variables and a data set with no missing values and errors would help in building better models which can address business problems more efficiently.

## **For the Model:**

# Decision Tree from the imputed data set seems to be performing better than the decision tree from the no-missing values data set. Therefore, we suggest using the Decision tree model from the Imputed data set.

# ‘Item\_MRP’, ‘Outlet\_Type’ and the number of years since establishment of the store give most of the information about the sales. So, in the future, we suggest considering these three properties for optimal sales.

# A quick scan of the leaf report screenshot explains the trends of sales. [Decision Tree Model:](#_Decision_Tree_Model:)

# References