**Data Import and Cleaning:**

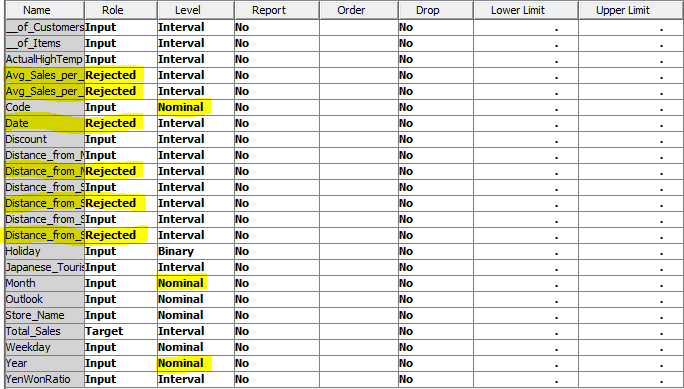
The dataset Korea.csv was imported into SAS Enterprise Miner. In the imported dataset, certain variables that were irrelevant were rejected at the start and not included in future analysis. The variables rejected were

* **Date** – the variable date was rejected as there were smaller units like month, weekday and year which were kept.
* **Distance from Main Street in meters, Distance from Station Y in meters, and Distance from Station X in meters** were rejected and the corresponding distances in feet were kept.
* **Average sales per item and Average sales per customer** are derived variables and hence we rejected them.

I also changed the Level of few variables as the levels were being displayed incorrectly.

* Month – Nominal
* Year – Since the year value is 2011, 2012 or 2013 we changed in to nominal.
* Code – This variable denotes the ownership of the store and hence changed it to nominal.

**The variable Total sales was changed to target variable.**



Once the variables to be used were selected, **data partitioning was performed** to create Training, Validate and test dataset. I used the default 40-30-30 dataset allocation method to create the 3 datasets. Also I used the **stratified method for sampling** and used **Store Name as the stratification variable.**

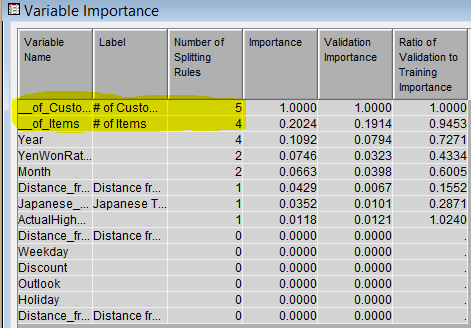
**This will ensure that in the sampled data set, all the stores are represented and the final model can be applied to each store individually.**

Once the data set was partitioned, I created multiple decision tree models and regression models and selected the best model that helps us to best determine the factors which influence the sales in each store individually.

**Decision Tree:**

In the first decision tree I created, I considered all the variables. I also created a 4 way tree using all the variables.

By analyzing the trees, the most important factors influencing sales were Number of customers and number of items sold. The below variables are for 4 way tree which is better than 2 way tree.



These 2 variables are obvious and highly co-related to total sales. A co-relation matrix obtained using statExplore confirmed it.

Correlation Statistics

(Maximum 500 observations printed)

Data Role=TRAIN Type=PEARSON Target=Total\_Sales

Input Correlation

\_\_of\_Customers 0.94430

\_\_of\_Items 0.91990

Avg\_Sales\_per\_Customer 0.47721

Distance\_from\_Station\_X\_Feet\_ 0.45734

Avg\_Sales\_per\_Item 0.43559

YenWonRatio 0.25625

Discount 0.22979

Japanese\_Tourists 0.20475

ActualHighTemp 0.16447

Code 0.16275

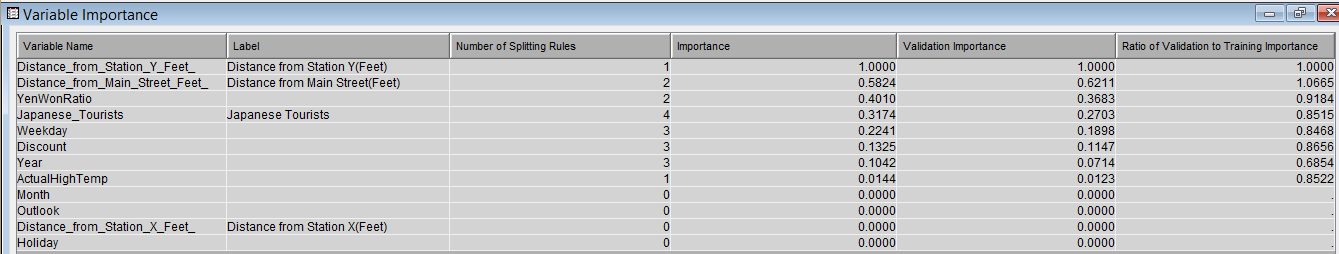
Year -0.23680

Distance\_from\_Station\_Y\_Feet\_ -0.31882

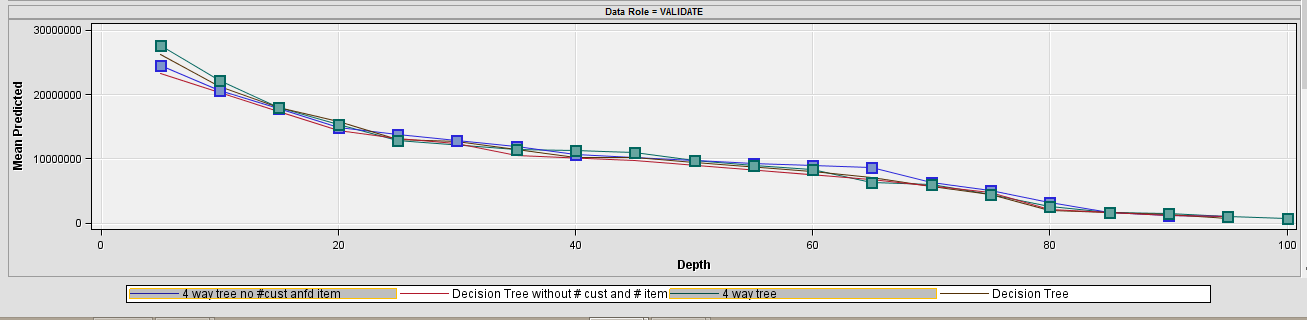
Distance\_from\_Main\_Street\_Feet\_ -0.47465

It shows that both Number of customers and number of items are highly co-related to total sales.

So, I again created two trees: 2 way decision tree and 4 way decision tree by not considering the variable number of customer and number of items. In this case we got different variables which are not easily visible. Below is the important variable for the 4 way tree.



Once all the decision trees we constructed, I compared the 4 decision trees.

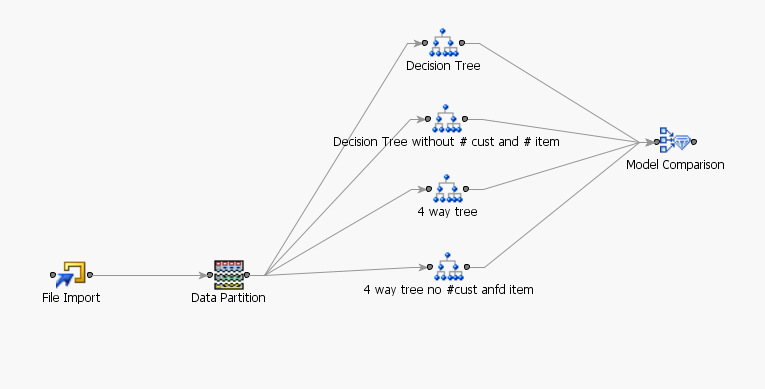


I concluded that the 2 way decision tree are performing poorly as compared to the 4 way decision trees.

Amongst the 4 way decision tree, the decision tree with number of customer and number of item (Green) is performing better for the first 20 percentile. This is expected as the top influencing factors are number of customer and number of items. While the decision tree without number of customer and number of item (Blue) is performing slightly better between the 20 and 60 percentile mark.

Thus the decision tree without number of customer and number of item will better determine factors which are important but not easily visible.

**So I decided to go with the decision tree that did not consider number of customers and number of items.**

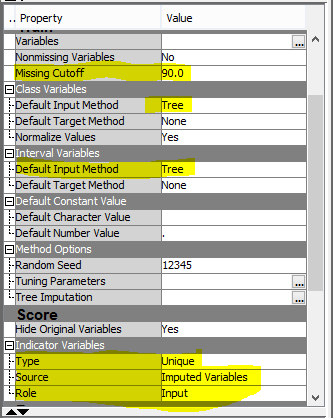


**Regression:**

In case of regression, before starting with the modelling, I need to manage the missing values. There are many ways to go about it and I decided to impute the missing values.

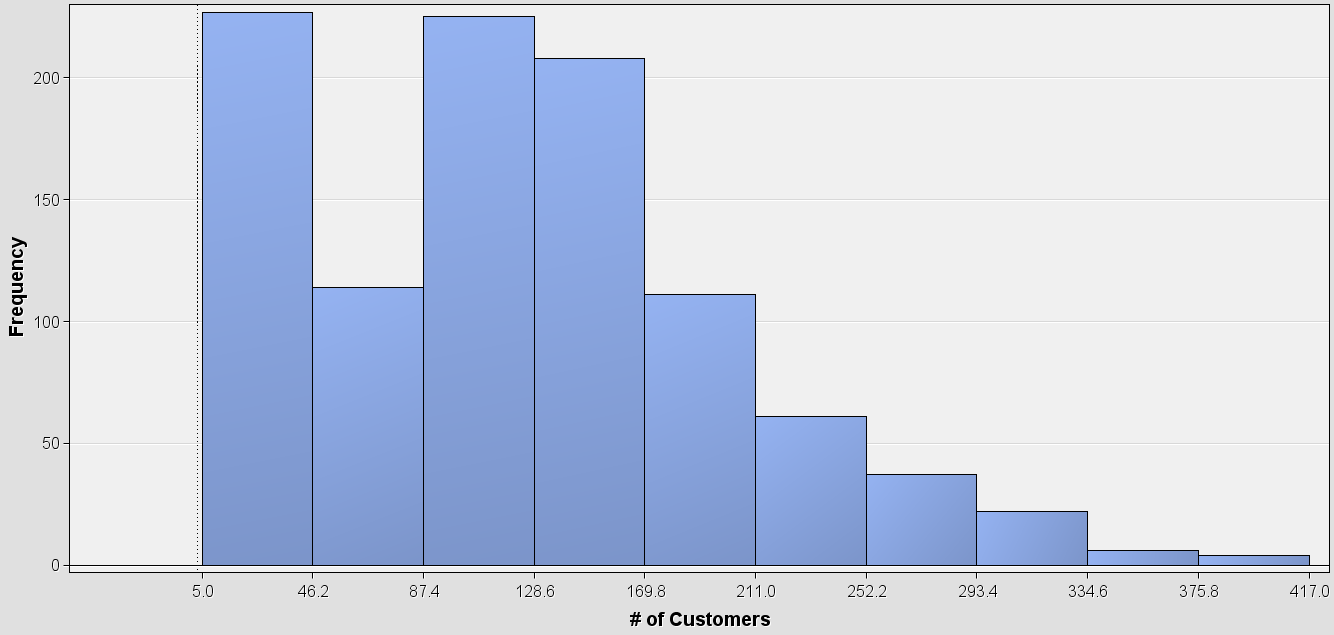
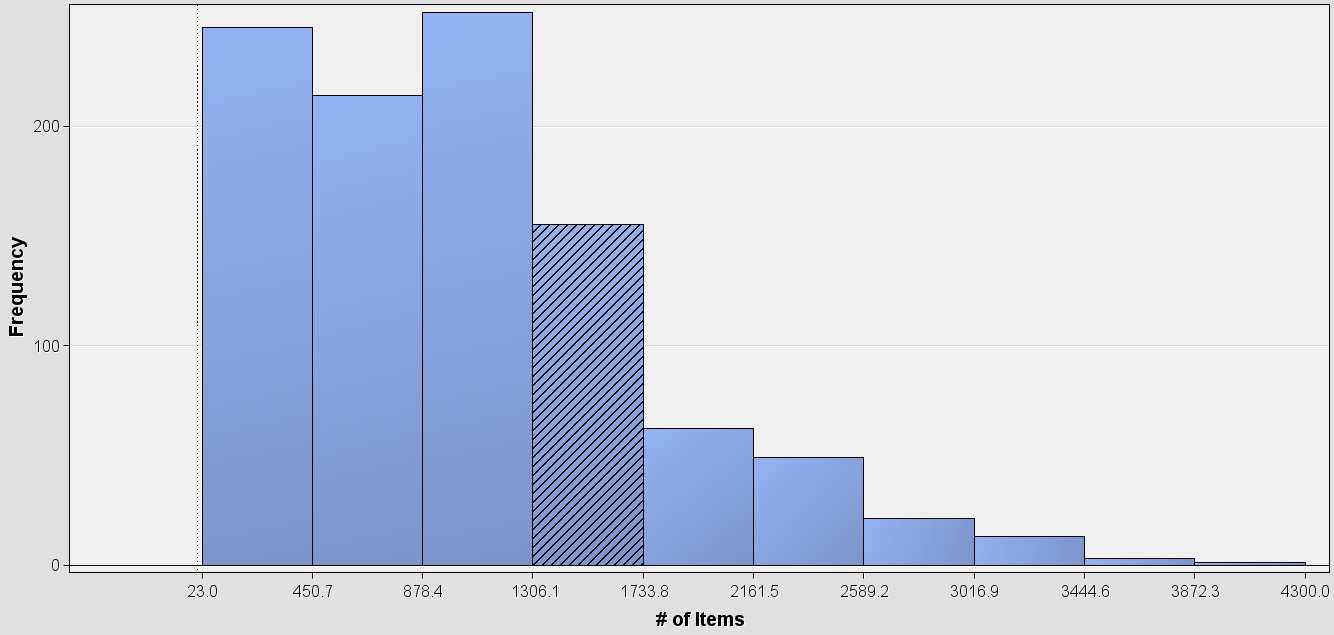
In this data, the Japanese Tourist data has a lot of missing values and need to be imputed. Since the number of tourist visiting will vary depending on holidays or weekends, I decided to go with tree as the imputation method and changed the missing value cut off to 90%.

I selected to create a unique variable for imputed values to see the effect of missing values.

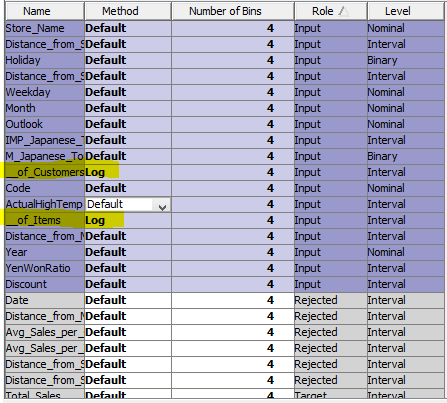


After having dealt with the missing values, I now need to deal with the Skewness of the data. For that I need to perform transformation.

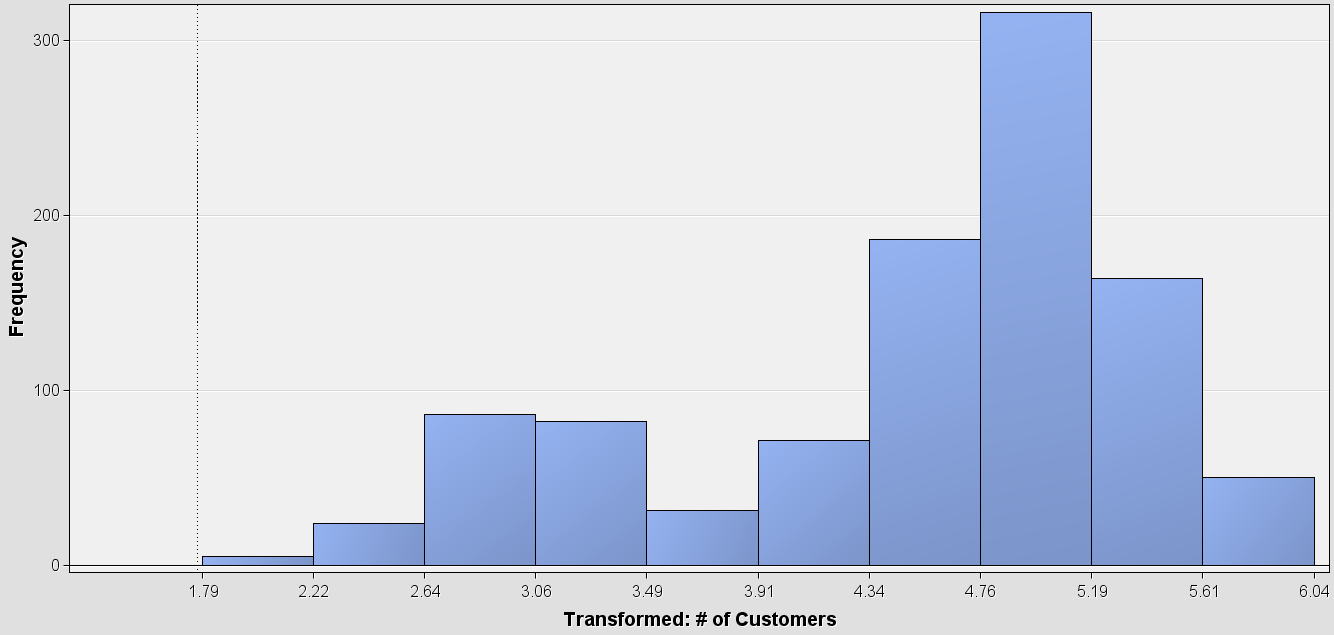
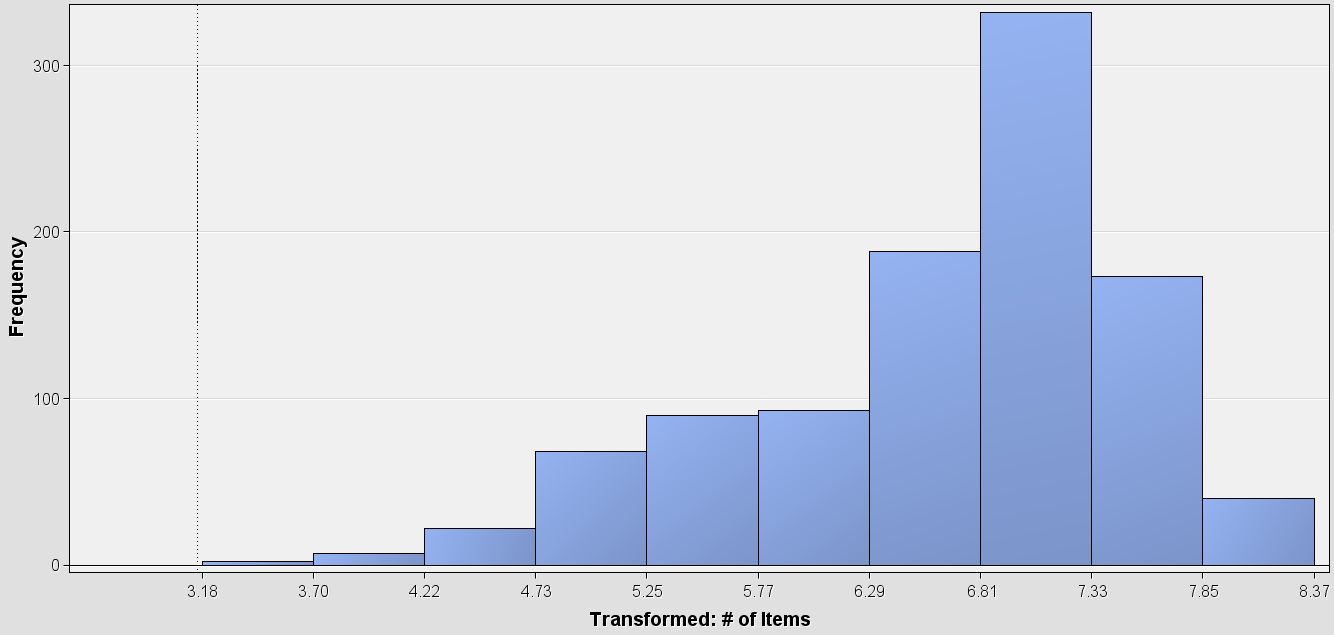
After exploring the variables, only variables number of customer and number of items seemed to be skewed.

To remove Skewness I performed log transformation on these two variables.



**Post transformation:**



Like decision tree, I again created two different models. One with the variables number of customer and number of items and one without those variables. For both these models I used the stepwise regression as it best measures the impact.

**Important factors with the 2 variables:**

Distance\_from\_Main\_Street\_Feet\_

Distance\_from\_Station\_X\_Feet\_

Distance\_from\_Station\_Y\_Feet\_

LOG\_\_\_of\_Customers

LOG\_\_\_of\_Items

Month

Year

YenWonRatio

**Important Factor without the 2 variables:**

Discount

Distance\_from\_Main\_Street\_Feet\_

Distance\_from\_Station\_Y\_Feet\_

IMP\_Japanese\_Tourists

M\_Japanese\_Tourists

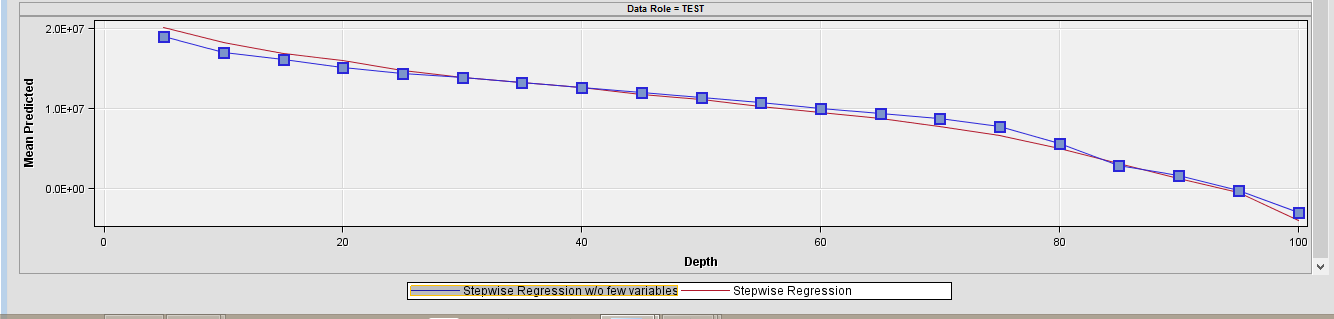
Month

Weekday

Year

YenWonRatio

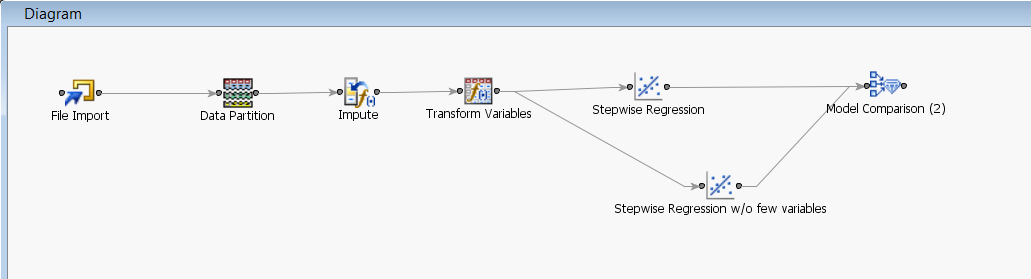
I compared the 2 regression model.



Amongst the 2 regression models, the model with number of customer and number of item (Red) is performing better for the first 20 percentile. This is expected as the top influencing factors are number of customer and number of items. While the regression model without number of customer and number of item (Blue) is performing slightly better between the 20 and 60 percentile mark.

Thus the regression model without number of customer and number of item better determines factors which are important but not easily visible.

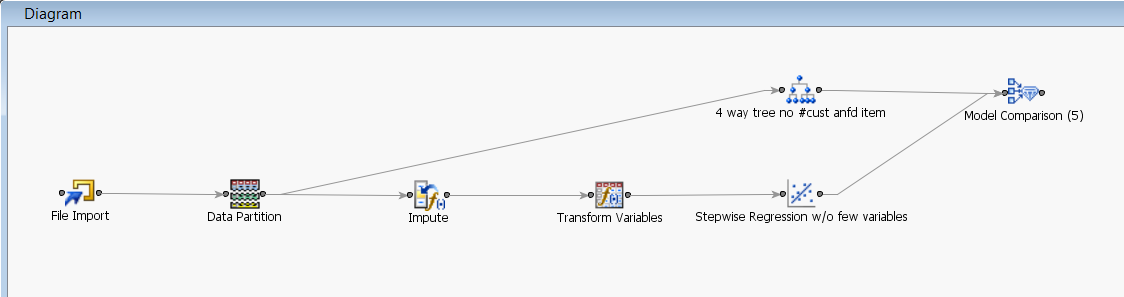
**So I decided to go with the regression model that did not consider number of customers and number of items.**



Once I had selected the decision tree and regression model, I compare these two models.

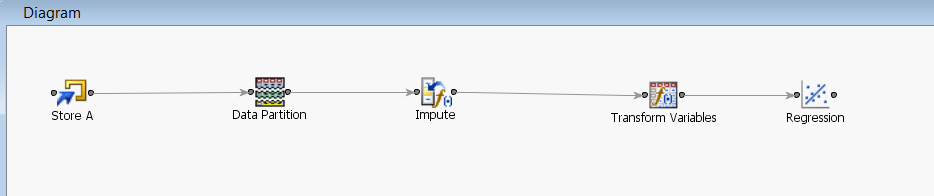


Based on this comparison, I decided to with the regression model (red) as it performs betters overall from 20 percentile.



I took this regression model and applied it to the data from each store individually to come up with factor influencing sales for each store.

Since I had selected samples from each store, I could apply the same model to all the stores individually.



**Answer to the Required Questions:**

**1. The factors that affect sales in each store and the importance of such factors.**

**Store A**

The important factors that influence sales are:

**Number of Japanese tourists**

The number of tourists is directly proportional to the total sales.

**Month of the year**

The Time the year influences sales. During the winter the sales are considerable low as compared to spring and summer.

**Weekday**

Friday, Saturday, Sunday and Monday are the days when the store does most of its sales.

**Year**

The year 2011 was better for sales as compared to year 2012. But in 2012, the store had a change of ownership which impacted the total sales.

**YenWonRatio**

The YenWonRatio is directly proportional to the total sales.

**Store B**

The important factors that influence sales are:

**Actual High Temperature**

Higher temperature corresponds to higher total sales. So, sunny days have more sales than snowy or cold days. But outlook is not a factor which is interesting to see.

**Japanese Tourists**

The number of tourists is directly proportional to the total sales.

**YenWonRatio**

The YenWonRatio is directly proportional to the total sales.

**Weekdays**

Friday, Saturday, Sunday and Monday are the days when the store does most of its sales.

**Store C**

The important factors that influence sales are:

**Code**

Change in ownership results in a negative impact on the total sales.

**Discount**

If the discount provided is more, the total sales tend to increase.

**Month**

The months of September, October and November tend to have more sales as compared to January, February and March.

**Store D**

The important factorsthat influence sales are:

**Japanese Tourists**

The number of Japanese tourist is directly proportional to total sales.

**YenWonRatio**

The YenWonRatio is also directly proportional to the total sales.

**Weekdays**

Friday, Saturday, Sunday and Monday are the days when the store does most of its sales.

**Month**

November, December, January and August are the months when the sales are better compared to other months.

**Discount**

If the store gives a discount, then the total sales increase.

**Actual High temperature**

Higher temperature corresponds to higher total sales. So, sunny days have more sales than snowy or cold days. But outlook is not a factor which is interesting to see.

**Store E:**

The important factors influencing the sales are

**Japanese tourists**

The number of Japanese tourist is directly proportional to total sales.

**Discount**

If the store gives a discount, then the total sales increase.

**Month**

September, October, November and December are the months when sales are better than the other months.

**Year**

The sales were better in 2012 as compared to 2011. Thus sales have been increasing.

**2. How these factors differ across stores, in particular how these factors compare against the key factors for Store B.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Factors | Store A | Store B | Store C | Store D | Store E |
| Japanese Tourist | 4330560 | 3579125 |  | 2063.4 | 163.5 |
| Month | Yes |  | Yes | Yes | Yes |
| Weekday | Yes | Yes |  | Yes |  |
| Year | Yes |  |  |  | Yes |
| YenWonRatio | 3351377 | 1502104 |  | 1565783 |  |
| Actual High temp |  | 27032.2 |  | 53450.3 |  |
| Code |  |  | Yes |  |  |
| Discount |  |  | Yes | Yes | Yes |

Table with influential factors for each store.

For Store B:

* Japanese Tourists has a low positive affect on the total sales in comparison with Store A while tourist have more impact as compared to store D and store E.
* YenWonRatio has a medium positive affect on the total sales in comparison with other stores.
* ActualHighTemp has a positive affect on the total sales but not as much as Store D.
* In case of weekdays all stores are having the same trends.