Predictive Analytics Project

Seoul Retail Case

Project Members:
Alekya Kumar
Trupti Jadhav
Anmol Agarwal
Kavin Soni

Under Professor Sanjuktha Das Smith

Plan:

Mr. Choe, the owner of store B is looking for some insight based on the collected data set in the city. Various data (variables) were taken into account for the analysis. He is also interested in how other stores are performing in the city and hence, data for other stores needs to be analyzed to help Mr. Choe make informed decisions.

At multiple instances long weekends could affect the sales and this was aliased as holiday (binary variable). Outlook gives an idea about how the weather and temperature is affecting the sales at the stores.

It has been noticed that 95% of the sales are driven through Japanese tourists. Hence various related data like currency ratio change has been noted throughout the span of 3 years.

Since all the stores are located in the same city distance from major stations has been taken into account.

Decision tree, regression analysis models has been used. We compared different models, picked the best one to interpret the result of given data and visualized appropriately to help Mr.Choe.

There are certain variables which play an important role in prediction. Each variable which is of significance, has been identified and explored to understand the behaviour.

The level of few variables were changed from their original state to more meaningful state.

1) Code - From Interval to Nominal
2) Year - From Interval to Nominal
3) Holiday - From Interval to Binary
4) Month - From Interval to Nominal

Variable Selection :

Certain variables which do not add any value to the data, have been excluded under certain assumptions.

1) Date - It does not provide any meaning when we have variables such as Month and Year.
2) Average Sales per customer - It is just a linear expression of number of customers and total sales. It was rejected on basis of multicollinearity
3) Average Sales per Item - It is just a linear expression of number of items and total sales. It was rejected on basis of multicollinearity.
4) Distance - WHile doing store-wise analysis, the distance was not included for each store except for Store C
5) Store Name - It was rejected for the same reason as (4)
6) Code - It was rejected for all models , but included while performing analysis on Stores A and C regarding the reopening of those.

Data Imputation :

There were two variables with missing values - Japanese Tourists and Outlook.

1) Japanese Tourists - The count of Japanese Tourists per day was same across each store. Japanese tourists just indicates the count of tourists and not the number of customers. Imputation using Tree methodology was used for imputing the missing values.
2) Outlook - It had NA values and spelling errors. The spelling errors were corrected and NA values were removed using Excel and Tree Methodology was used to impute the variable.

Variable Transformation :

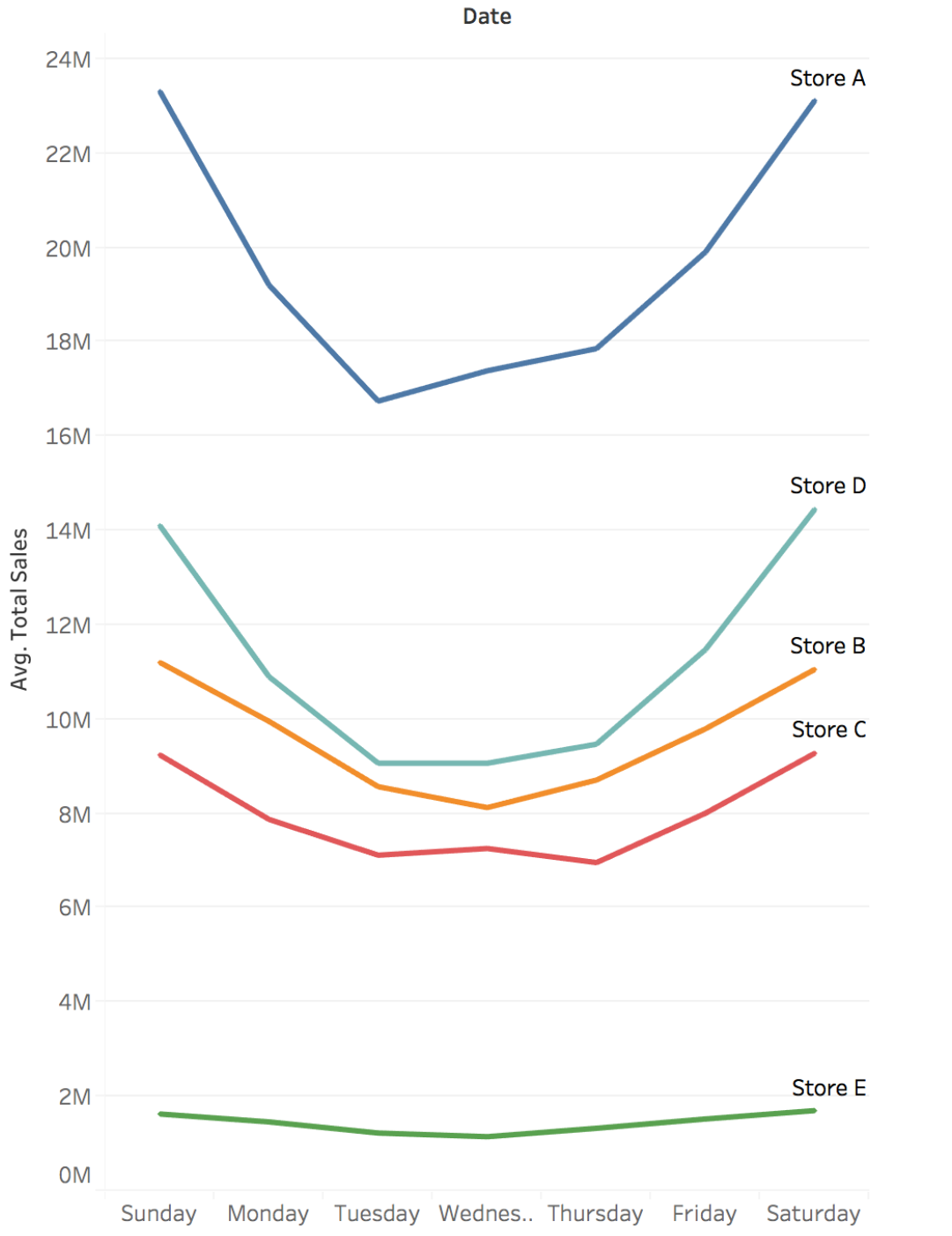
Few variables were highly skewed indicating the presence of outliers. Corrective action needs to be taken in order to get rid of the outliers and produce a good prediction.

1) **Yen-Won Ratio - The variable was negatively skewed, hence Square transformation was used to get rid of the skewness.**
2) **Number of Items - The variable was positively skewed, hence Log Transformation was done to bring it to a normal distribution**
3) **Discount - The variable was positively skewed, hence Log Transformation was done to bring it to a normal distribution**
4) **Imputed Japanese Tourists - The variable was positively skewed, hence Log Transformation was done to bring it to a normal distribution**

Exploratory Data Analysis:

Weekday Analysis for Sales :

Weekend effect

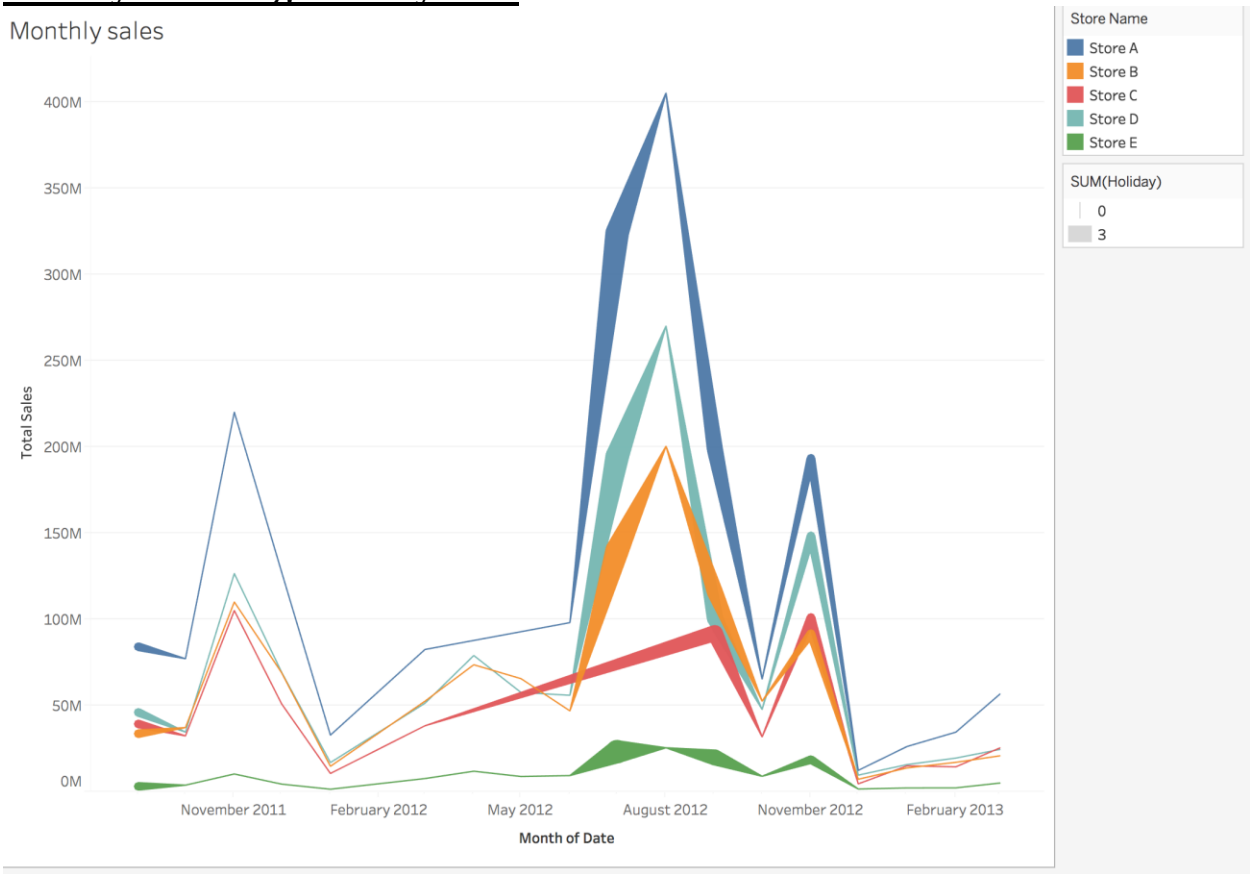


A general trend has been seen that there is greater sale over the weekends for all the stores. In general Store C and E has lower overall sales whereas store A and D has higher sales.

- The sales plunges drastically for store A compared to other stores on Tuesday.
- Store E has the least differences in accordance with the weekdays.

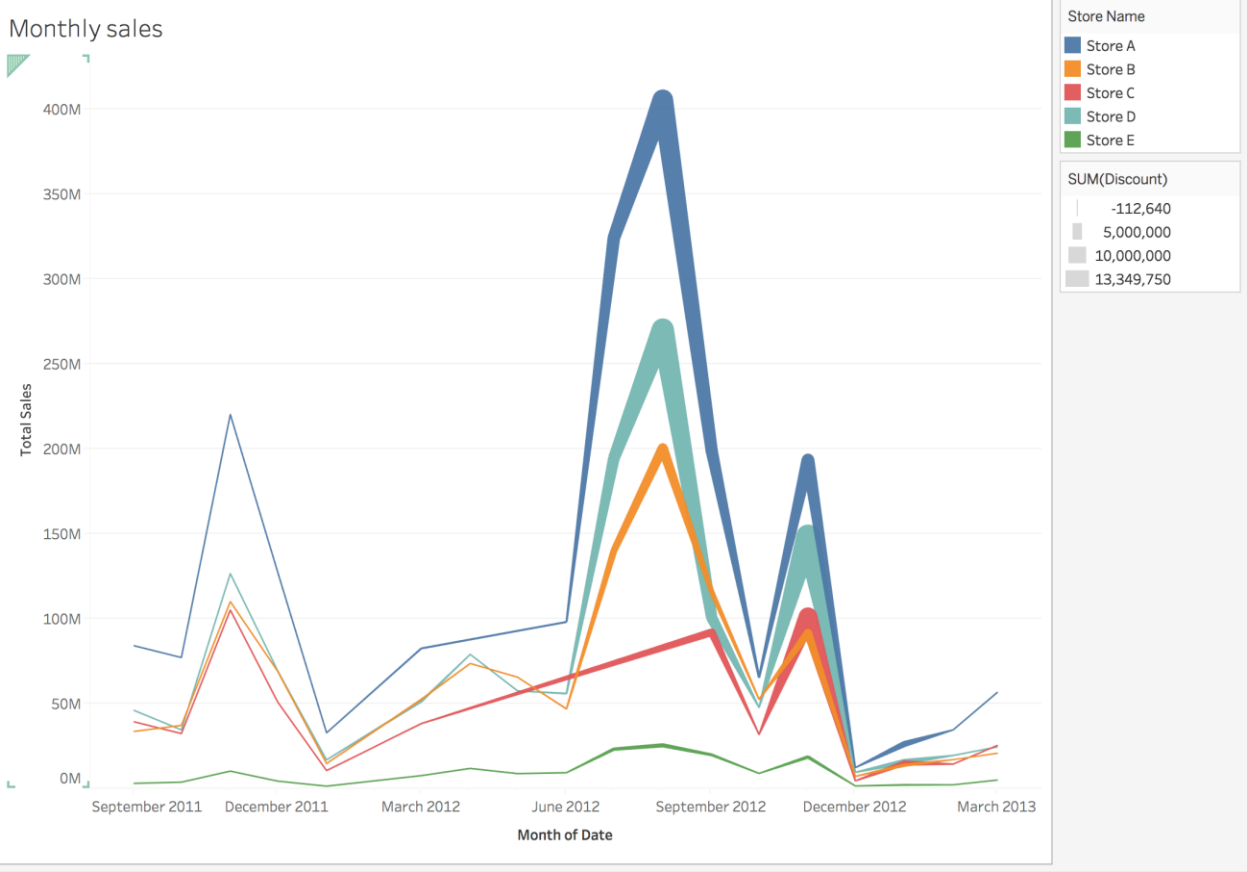
- Store B see the least sales on Wednesdays and hence Mr.Choe must offer more discounts to attract more customers. Also, he must consider what store C and A does since they have little better sales on Wednesdays compared to Tuesdays and Thursdays.

Holidays affecting monthly sales:



We can notice that from June 2012 to December 2012 more holidays occurred and hence there is more sales during that period of time as well

Discounts affecting monthly sales:



The thickness shows that there were more **discounts** during that period of time.

We can notice that from June 2012 to December 2012 more discounts offered were made and hence there is more sales during that period of time as well.

Therefore, it is worth noting that there is positive correlation between discount and holiday.

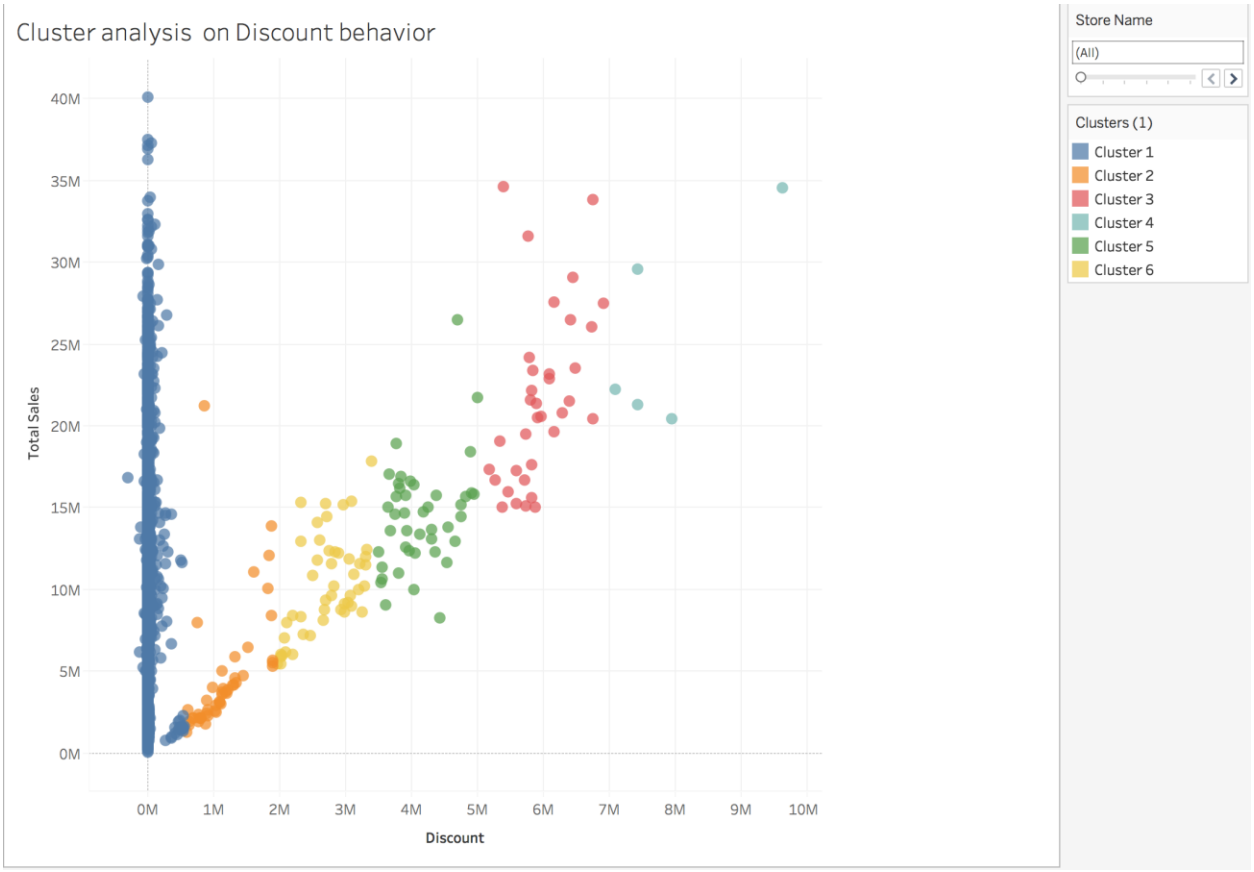
We can also notice that the maximum discounts have been offered by store A & D. Only exception being Store C having more discount during November 2012.

The only exception from the pattern was the store C, which doesn't show very high sales figure.

Cluster analysis for discount variable :

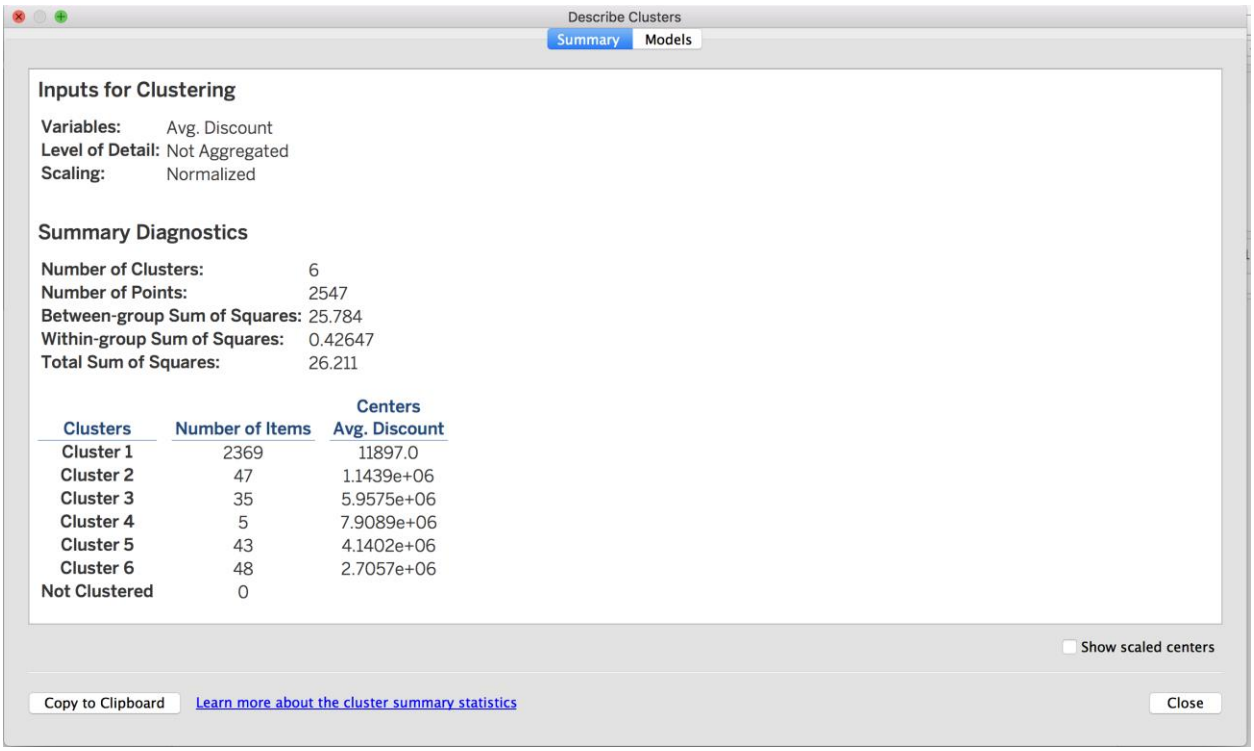
Let's focus more on the discount variable

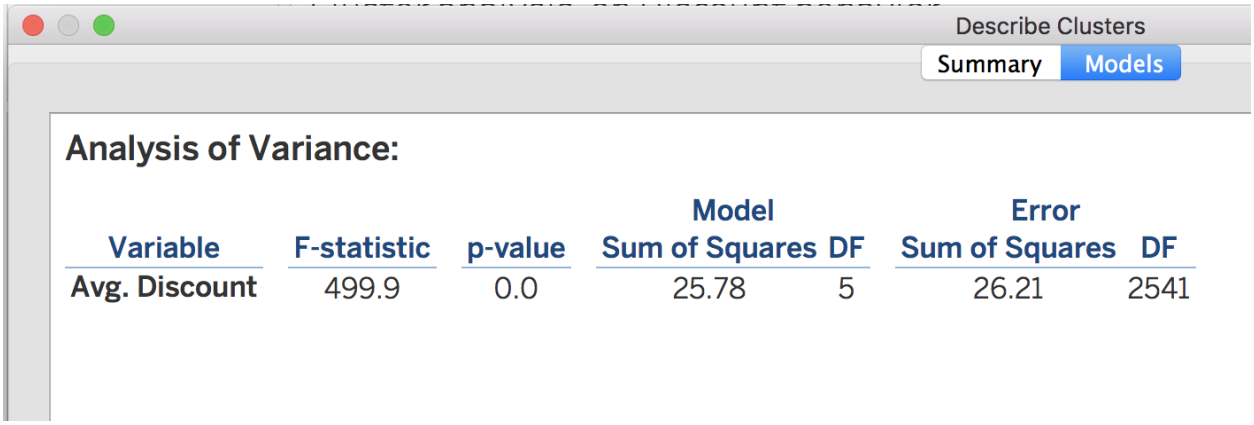
Since there is a relationship between discount and total sales we have done cluster analysis to recognizing the pattern.



We notice that for zero discount or very less discount has no correlation with the sales. This can be seen as the cluster 1.
Any discount offered beyond cluster 1 shows a positive linear relationship with the total sales. Cluster 3 is quite scattered and difficult to draw inference from but cluster 2 & cluster 6 is concentrated and can be labeled easily.

The analysis for the same is as follows:





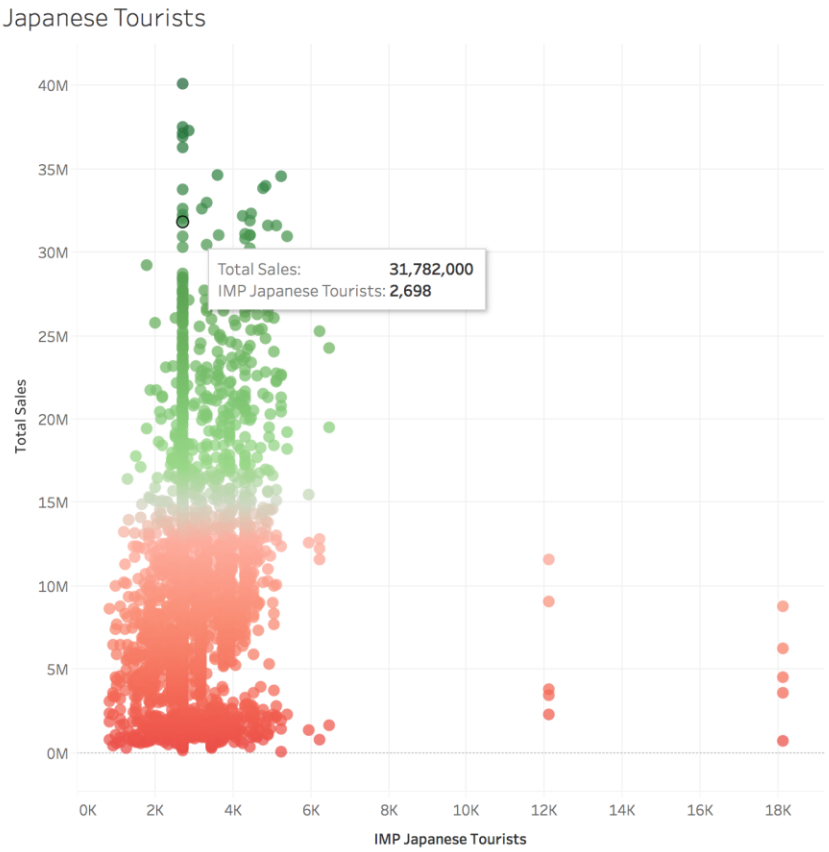
Split decision for decision tree:

For spitting technique in decision trees, we have used F test since it looks across all the branches. For continuous splitting the variance should be less. The variables in certain division should be more like each other. F and variance is used for continuous variables.

Variance – When target variable variance is less, there is greater similarity among them. Hence lower value is desired.

F test – This is not just looking within branch but looking across all branch variation.

Japanese tourists affecting the sales:

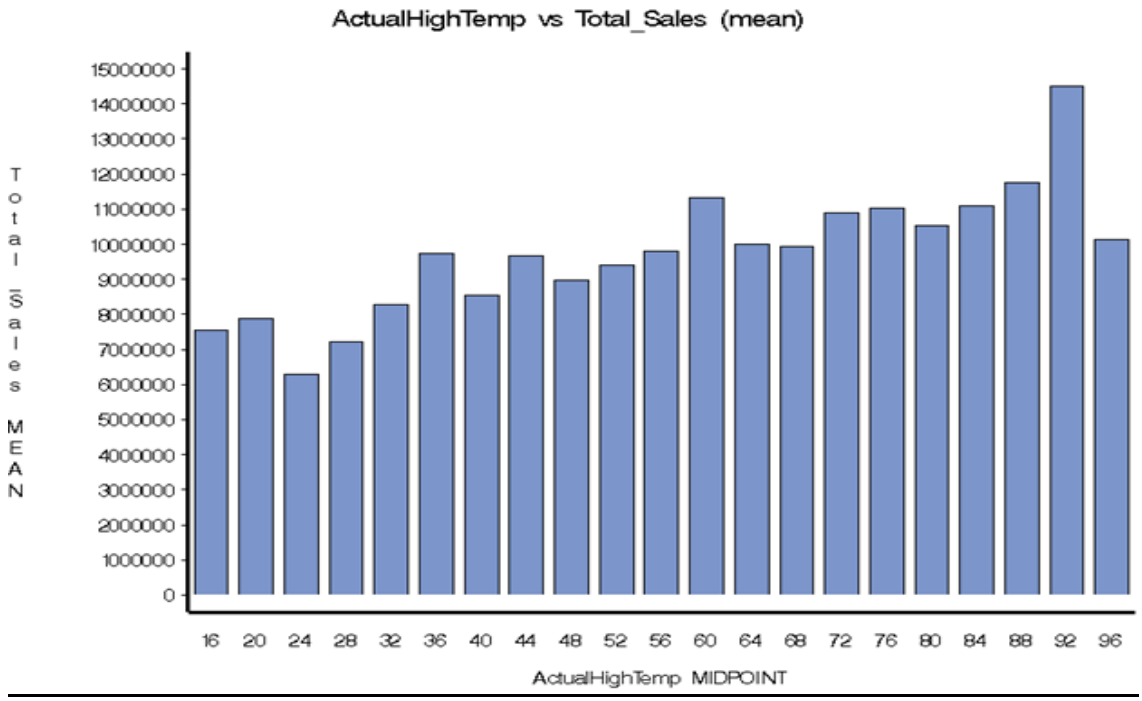


Overall for all outlook we see that maximum sale is seen when tourists are between 2000 to 6000.

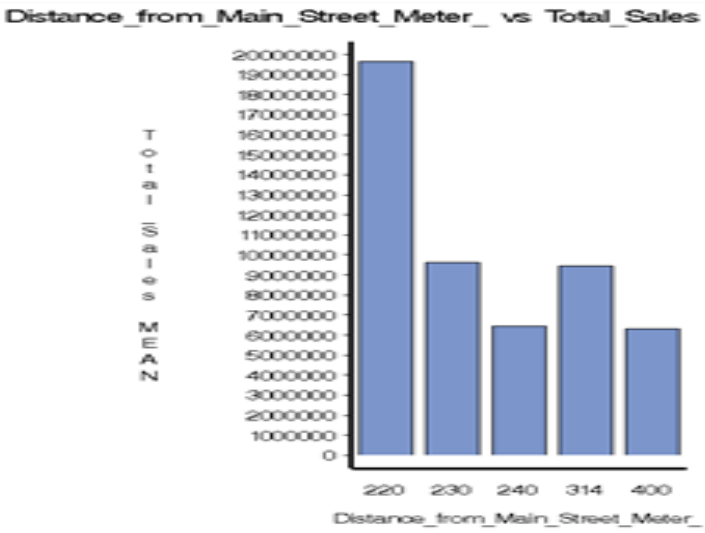
If there are more than 6K tourists, we do not see any improvement in the sales. Hence any efforts to increase the tourism beyond 6K is not worth it.
Also, we can notice that when the number of tourists is less than 2000, there is a drop in the total sales. Hence, efforts should be made to attract more Japanese tourists in the month of April to August to bring them in the desired range (refer to month analysis graph).

Outlook affecting the monthly sales:

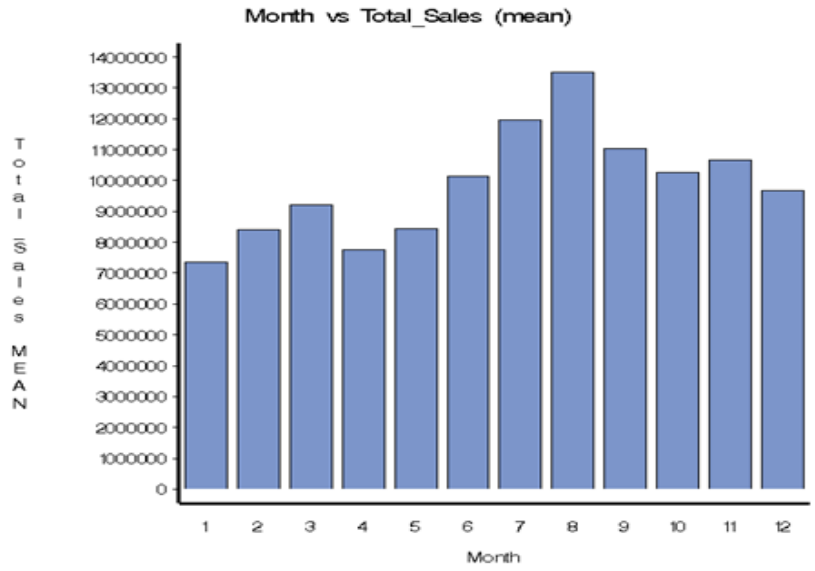
There is a drop in the sales during October 2011 for snowy outlook. Hence, we can infer that there might be a **snow storm** which affected the sales whenever it snowed during this time. This plot shows the relationship between Actual High temperature and the Total sales, the target variable. So, from the graph we can infer that if the temperature is high on a particular day, the sales are also high. This shows a linear relationship between temperature and total sales.



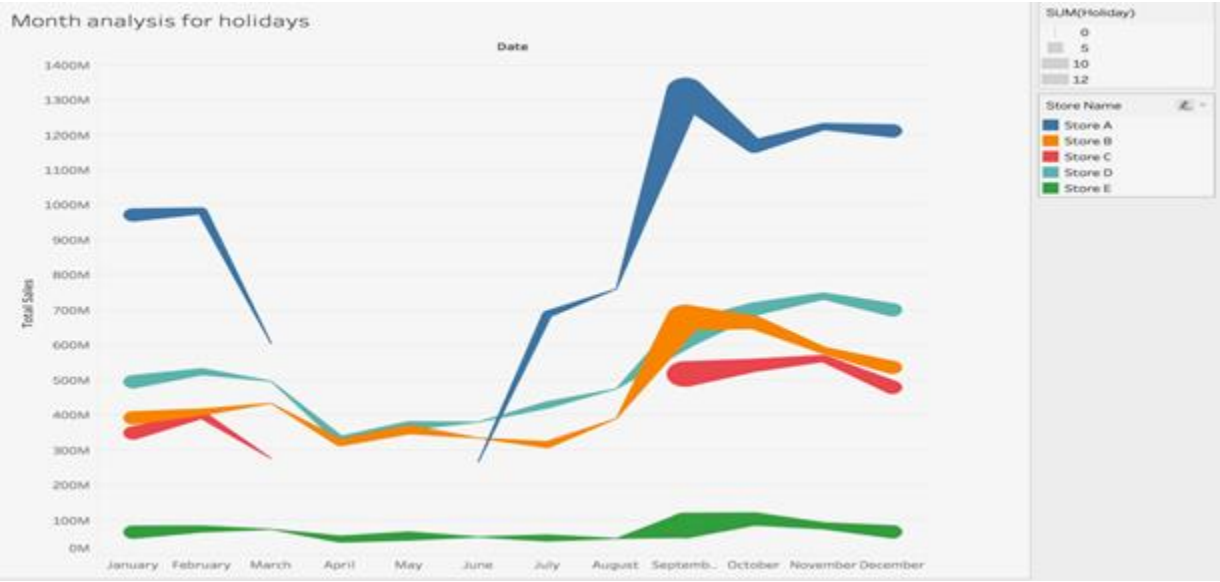
When the distance from Main street increases, the sales also increases accordingly. We see a dip in the value of 240, which tells us that, store C was closed during that time and hence the sales is lower and non proportional to the total sales. If we eliminate that value, it gives an overall negative correlation between Total Sales and Distance from the thoroughfare.



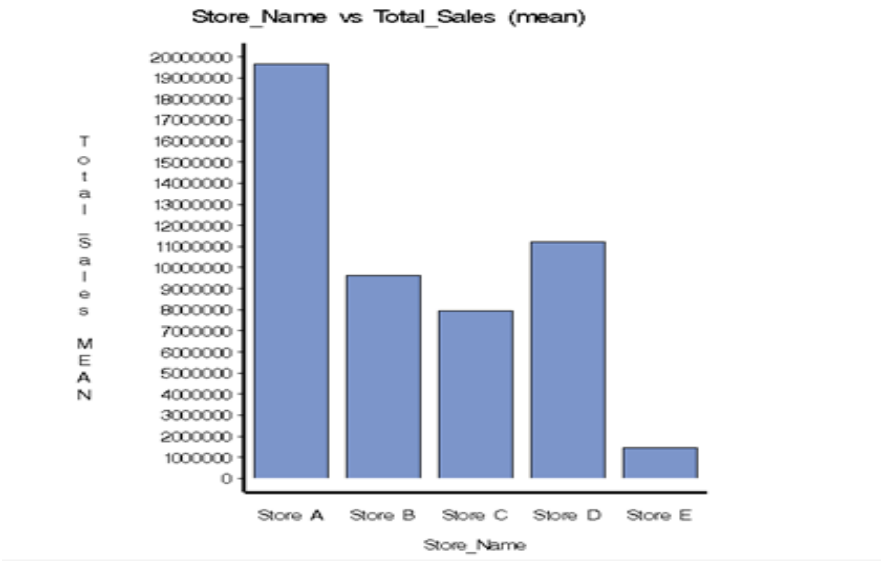
The month-wise average total sales gives an insight about what drives sales in each month. From the graph, it can be inferred that, months August, September and October have high sales. This also implies that the number of holidays during those months was high compared to other months.



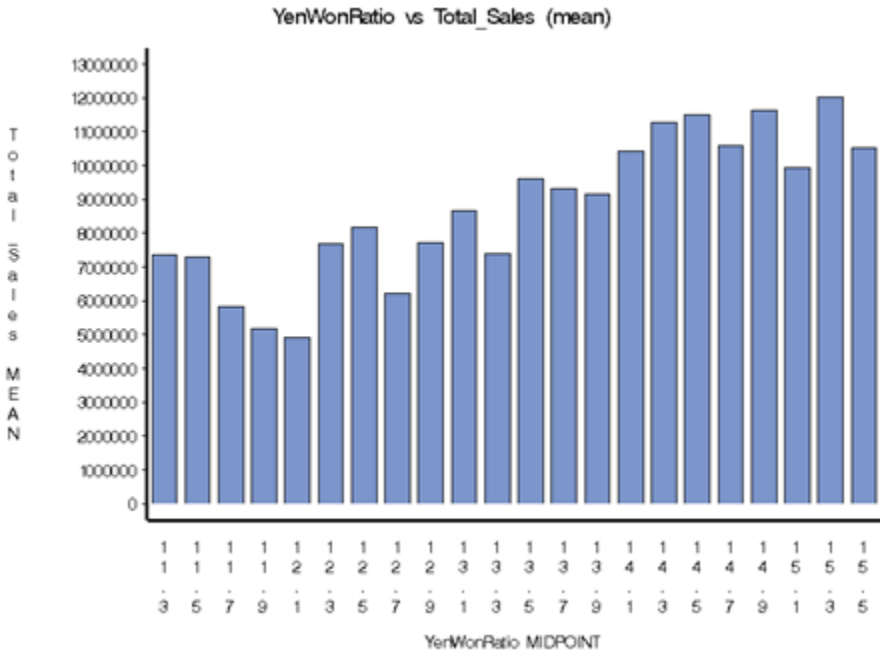
The plot displayed below, shows the store-wise total sales for each month. The thickness resembles more number of holidays.



The store-wise comparison of mean total sales shows that Store A performs much better than any other stores. Beyond the fact that Store A was closed for a short period of time, it seems to outperform other stores.



The Yen-Won ratio is measure of proportionality between Japanese Yen and Korean Won. This is an important measure as this plays a major role in attracting Japanese customers to Seoul. If the ratio is more, then the probabilities of Japanese tourists visiting Seoul is more.

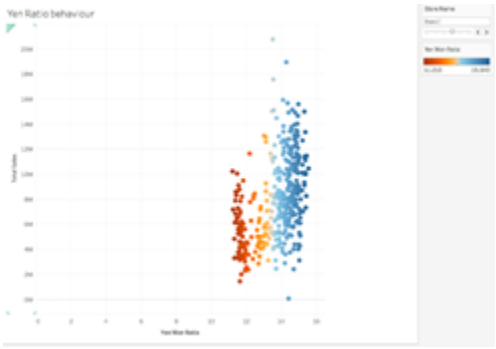


The graphs displayed below, show the relationship of Yen-Won ratio with Total Sales for each store.

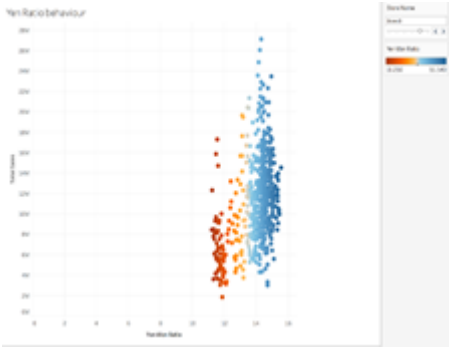
Store A :

Store B :

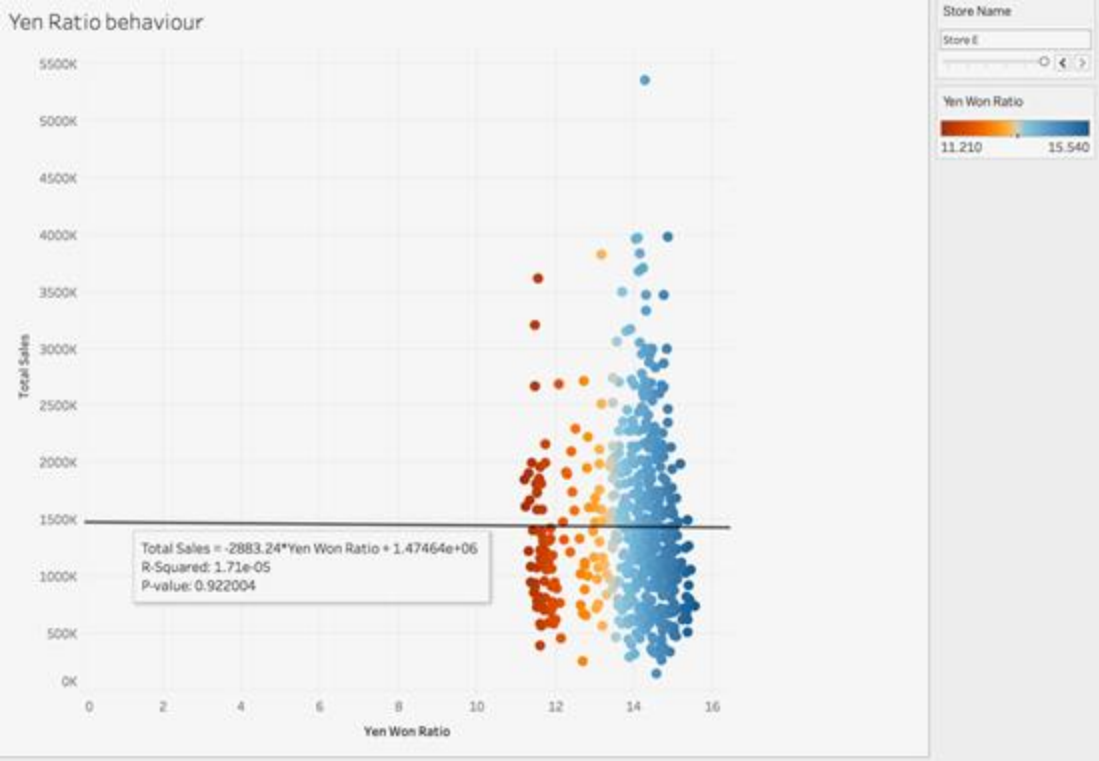
Store C:



Store D:



Store E :



The plots for Stores A through D, show a good correlation between Yen-Won ratio and the total sales. Store E, though it shows a good correlation, the trend line in store E is flat and gives an inference that is not aligned to other stores.

Store-wise Analysis:

MODEL SELECTION:

Models were built taking into consideration two aspects for all the stores.

- ❑ **Non Inclusion of Number of Customers and Number of Items**
- ❑ **Inclusion of Number of Customers and Number of Items:**

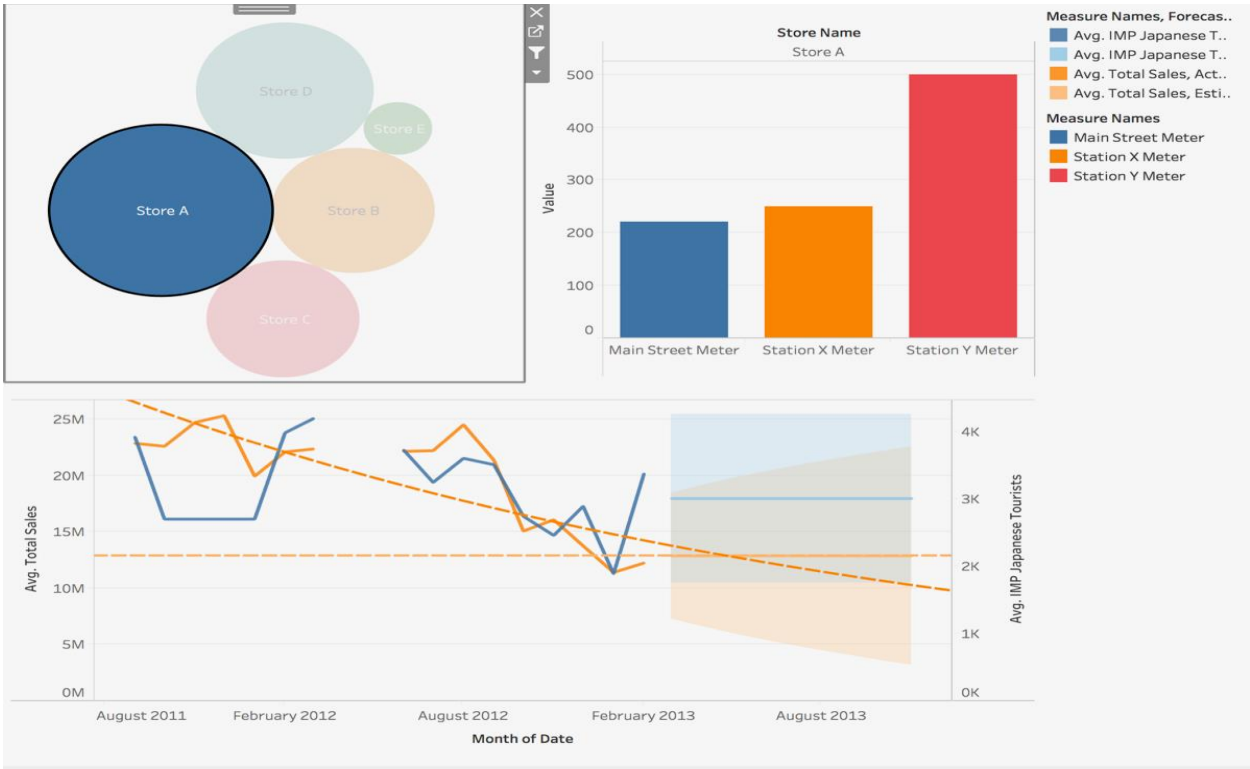
  These two variables have a very strong correlation with the target variable.

DATA PARTITION:

The raw data is divided into five mutually exclusive sets based on the Store name. These are stored in separate excel and imported into SAS Enterprise Miner.

STORE A

The general trend of total sales of Store A and Japanese visitors. The distance of the store A from different locations is highlighted too. The dotted line shows the trend for the total sales. the colored area shows the forecasting to predict the future sales.



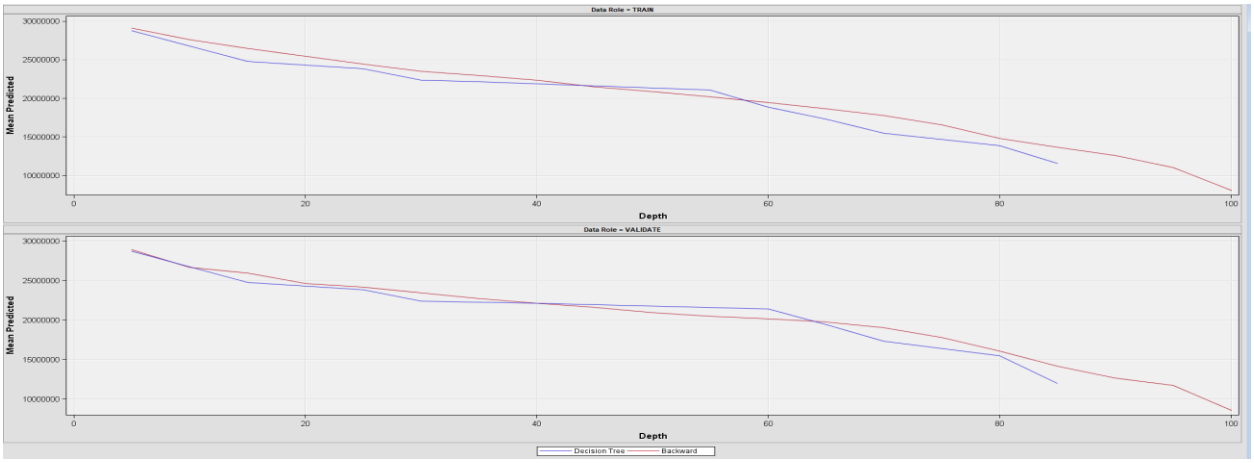
MODEL SELECTION:

Part A: Non Inclusion of Number of Customers and Number of Items

This model is when the store hasn't relocated. The best model for our inferences will be Backward regression model, as selected by the tool.

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Average Error Function	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom
Y	MdlComp...	Reg20	Backward	Total Sal...	Total Sales	1.555E13	10000.87	1.517E13	1.517E13	305	23
	MdlComp3	Tree	Decision ...	Total Sal...	Total Sales	1.91E13	.	1.58E13	.	.	.

Selected Model

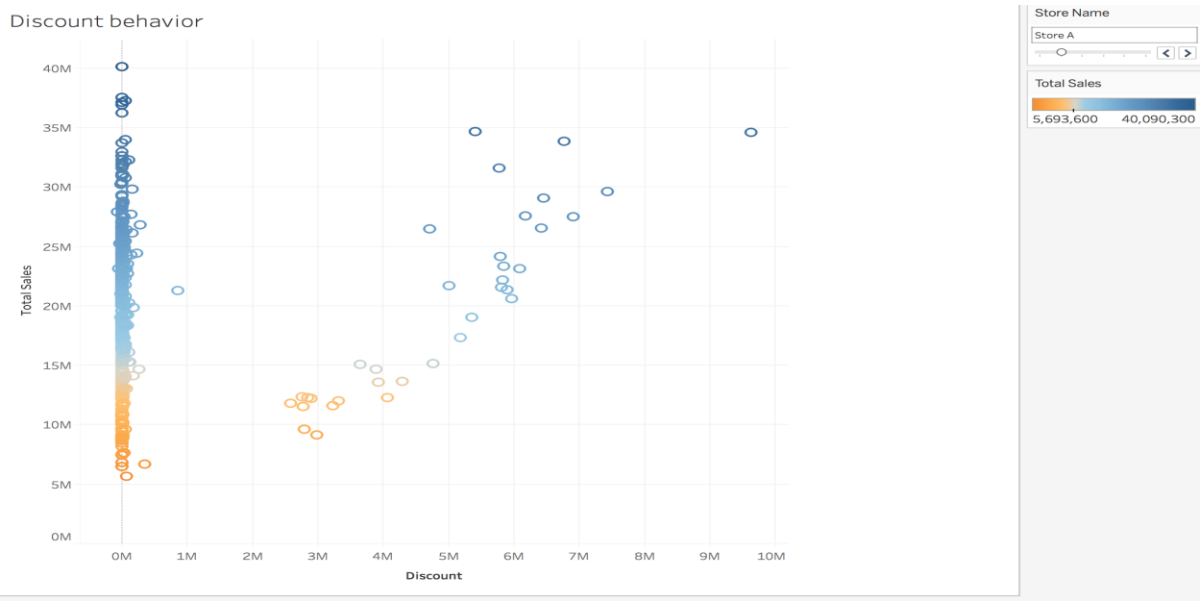


The inferences that can be drawn are:

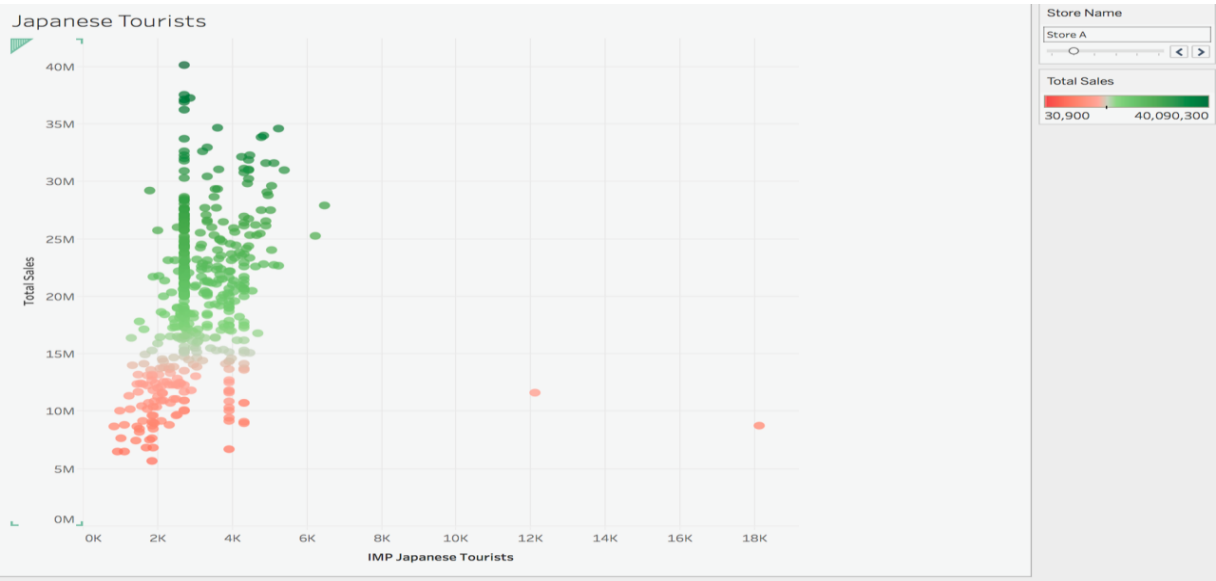
Intercept		1	-4.339E7	8840909	-4.91	<.0001
ActualHighTemp		1	-15442.4	31944.0	-0.48	0.6291
Holiday	0	1	-248466	415170	-0.60	0.5500
LOG_Discount		1	285026	189797	1.50	0.1342
LOG_IMP_Japanese_Tourists		1	4168866	872692	4.78	<.0001
Month	1	1	-1534714	1163402	-1.32	0.1881
Month	2	1	-220025	1096732	-0.20	0.8411
Month	3	1	2090610	1071464	1.95	0.0520
Month	6	1	750592	1597557	0.47	0.6388
Month	7	1	1400062	1107331	1.26	0.2071
Month	8	1	3725574	1259220	2.96	0.0033
Month	9	1	-1051090	844658	-1.24	0.2143
Month	10	1	-4408648	681353	-6.47	<.0001
Month	11	1	-315069	840466	-0.37	0.7080
SQR_YenWonRatio		1	130597	22258.2	5.87	<.0001
Weekday	Friday	1	-228959	553899	-0.41	0.6796
Weekday	Monday	1	842876	588201	1.43	0.1529
Weekday	Saturday	1	3070109	560962	5.47	<.0001
Weekday	Sunday	1	3541102	553682	6.40	<.0001
Weekday	Thursday	1	-1683485	565488	-2.98	0.0031
Weekday	Tuesday	1	-2914708	573385	-5.08	<.0001
Year	2011	1	1720673	1182883	1.45	0.1468
Year	2012	1	-1402884	463966	-3.02	0.0027

There's strong impact of LOG_Japanese_Tourists, Discounts, Month of March, July, August, SQR_YenWonRatio, sales on Monday, Saturday, Sunday and year 2011 on the Total_Sales. The following Tableau diagrams also support the inferences.

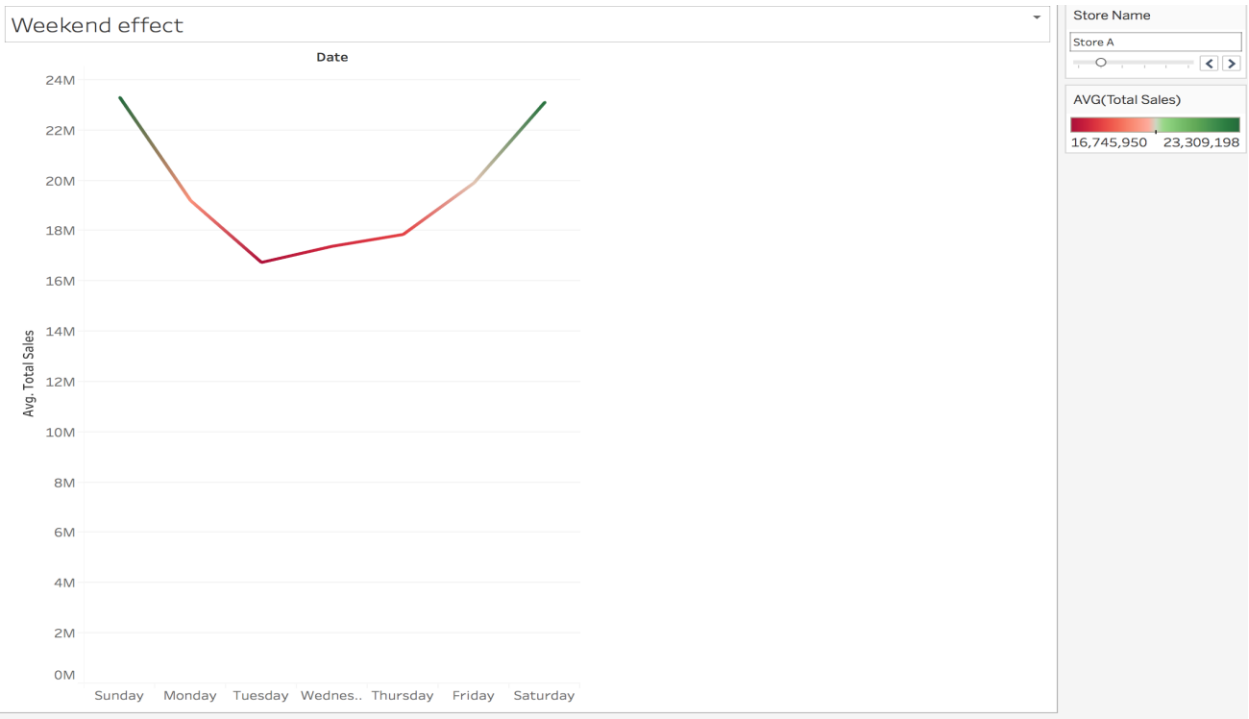
More the Japanese tourists, more are the Total_sales. Similarly, with the discounts.



We can see that discount has a linear relation with the total sale except when there is no discount given.



Japanese tourists has the best range from 2000 to 6000



Store A has least sale on tuesday and and wednesday which needs some attention.

- The sales are high from Saturday to Monday. The manager should make sure that he gives more discounts on items on these days as more discounts too, correspond to high Total_Sales.
- The manager should keep a check on YenWonRatio rates, especially from Saturday to Monday as high YenWonRatio corresponds to higher sales.

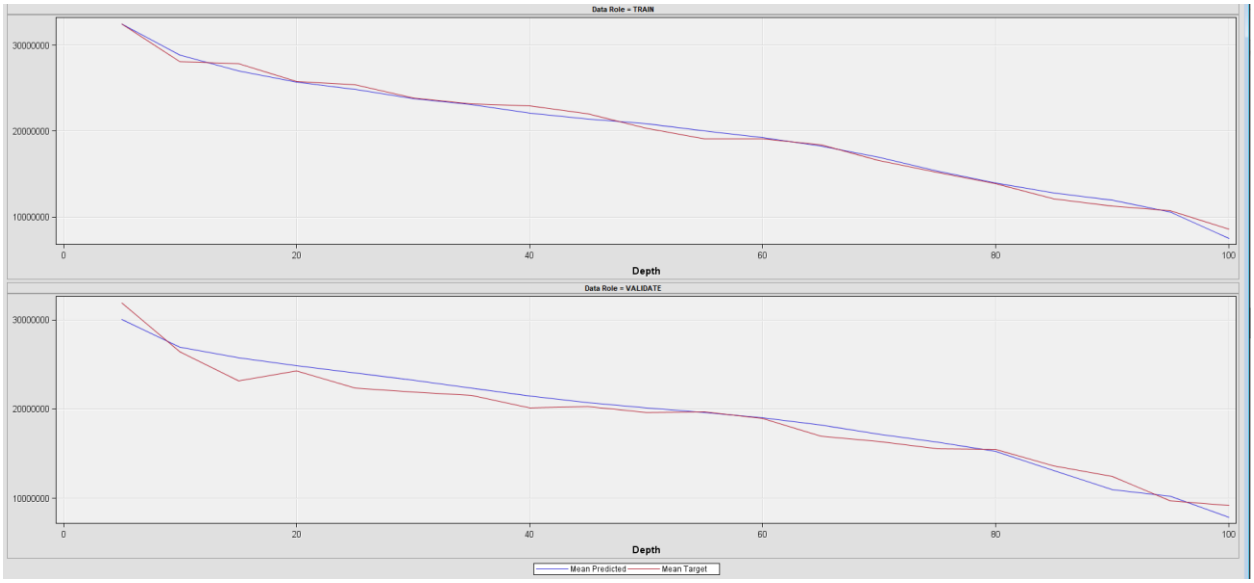
Part B: Inclusion of Number of Customers and Number of Item variable.

The data set was analyzed by including Number of Customers and Number of Items in the model. Along with those variables, predictors such as Discount, Holiday, Weekday , Year , Month, Yen-Won ratio, Weekday also play an important role.

Thus including these two variables(Number of Items and Number of Customers) prove the assumption that they are correlated to the target variable.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	-4.556E7	6138409	-7.42	<.0001
Holiday	0	1	-398670	256619	-1.55	0.1213
LOG_Discount		1	-540487	124155	-4.35	<.0001
LOG___of_Items		1	5543586	842182	6.58	<.0001
Month	1	1	-1784574	514048	-3.47	0.0006
Month	2	1	-3255008	513268	-6.34	<.0001
Month	3	1	-2275384	634366	-3.59	0.0004
Month	6	1	-574316	848023	-0.68	0.4988
Month	7	1	1734236	508341	3.41	0.0007
Month	8	1	1868771	572799	3.26	0.0012
Month	9	1	358407	420618	0.85	0.3948
Month	10	1	-381860	455765	-0.84	0.4028
Month	11	1	2001082	498544	4.01	<.0001
SQR_YenWonRatio		1	85733.9	13965.6	6.14	<.0001
Weekday	Friday	1	-25370.2	345071	-0.07	0.9414
Weekday	Monday	1	516769	366554	1.41	0.1596
Weekday	Saturday	1	1427727	356430	4.01	<.0001
Weekday	Sunday	1	1351757	360896	3.75	0.0002
Weekday	Thursday	1	-783719	354532	-2.21	0.0278
Weekday	Tuesday	1	-1607747	361777	-4.44	<.0001
Year	2011	1	-4398586	787480	-5.59	<.0001
Year	2012	1	-510029	286249	-1.78	0.0758
__of_Customers		1	52852.4	5239.8	10.09	<.0001



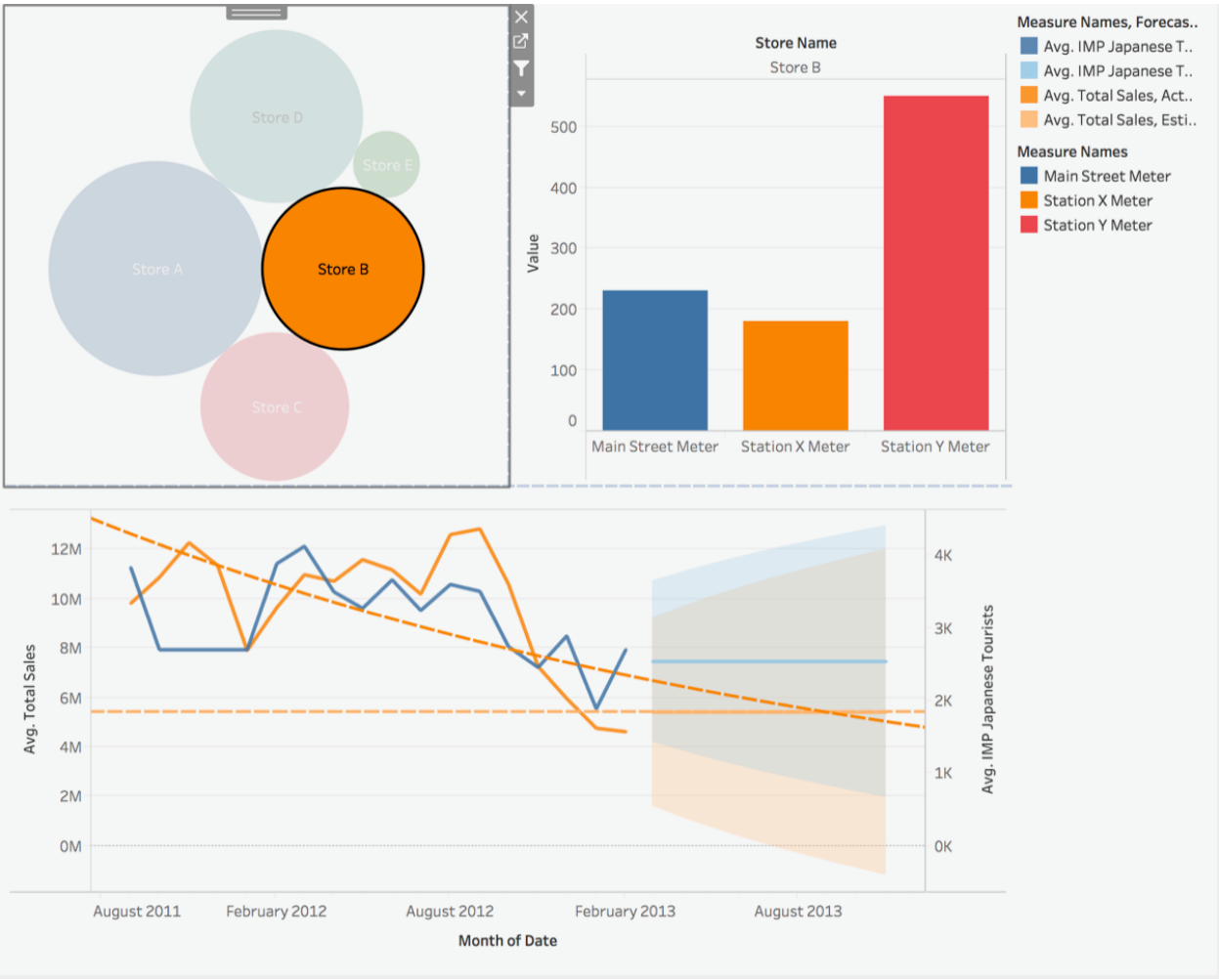
Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Sum of Frequencies	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error
Y	Reg5	Reg5	Backward	Total Sal...	Total Sal...	6.041E12	328	8753998	1.929E15	5.881E12	2424989
	Tree	Tree	Decision ...	Total Sal...	Total Sal...	1.25E13	328	9701789	2.934E15	8.944E12	2990690

We observe that here also Backward Regression model plays well. The factors picked up by the model are the same for Part A. So here the two variables are not acting differently and the result is the same. Thus the bifurcation does not affect in this case.

STORE B:

The following shows the general trend of sales at store B along with the distances from various locations.

The dotted line shows the trend for the total sales. The colored area shows the forecasting to predict the future sales. The predicting values according to the trend line shows that there will be a decrease in the total sales and hence, Mr. Choe must be advised to make chances in the business model.

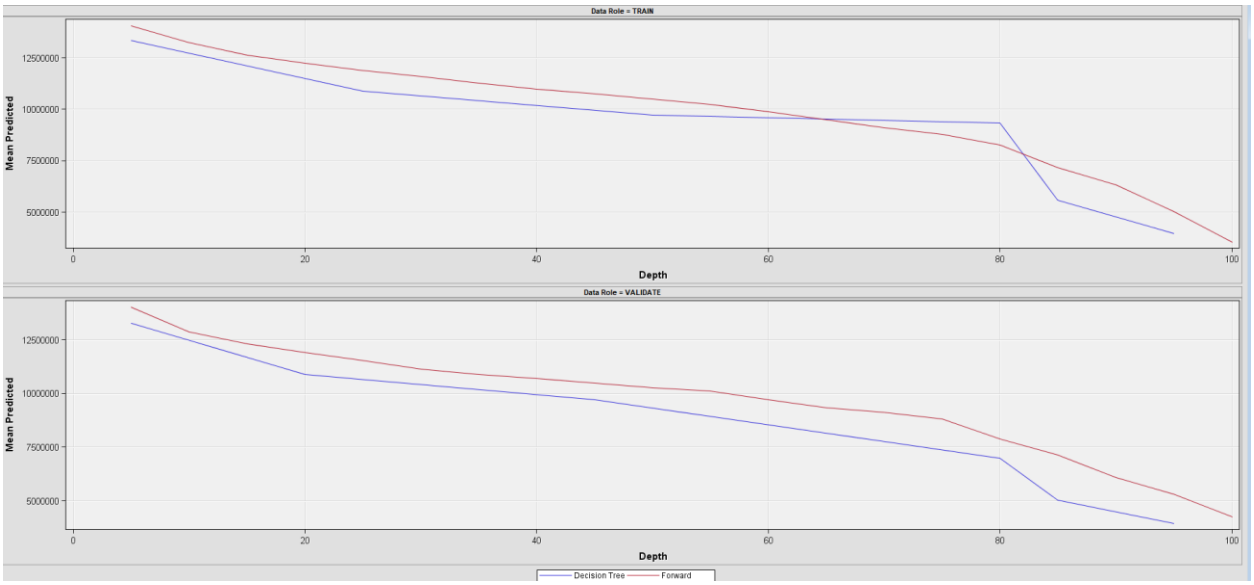


MODEL SELECTION:

Part A: Non Inclusion of Number of Customers and Number of Items

The final model that we'll be using for our inferences will be Forward Regression model as selected by the software.

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Sum of Frequencies	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error
Y	MdlComp...	Reg22	Forward	Total Sal...	Total Sales	8.841E12	394	16108250	3.426E15	8.696E12	2948851
	MdlComp2	Tree11	Decision ...	Total Sal...	Total Sales	9.906E12	394	16899771	3.248E15	8.244E12	2871229

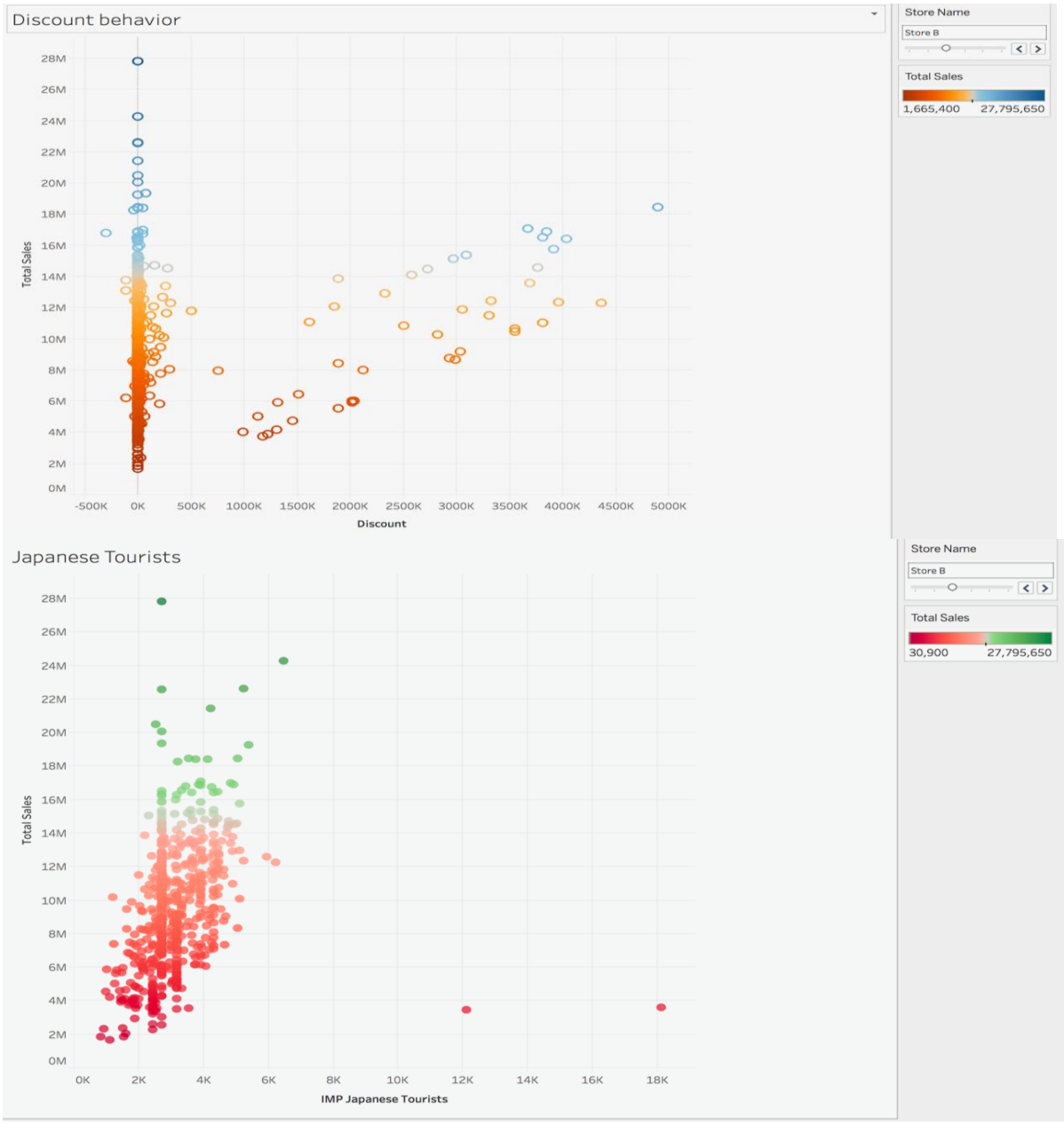


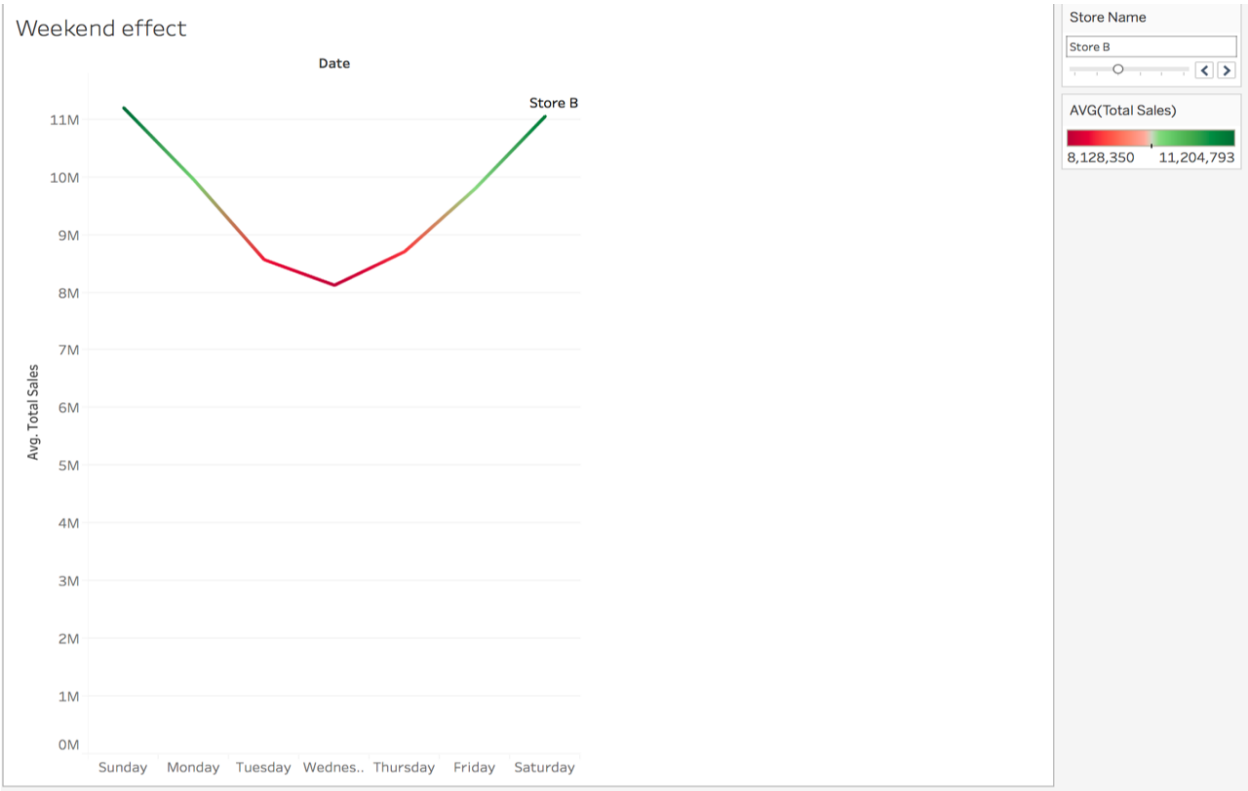
The inferences that can be drawn from the prediction models are:

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	-2.793E7	5397334	-5.17	<.0001
LOG_IMP_Japanese_Tourists		1	3092632	685129	4.51	<.0001
Month	1	1	-1610279	522357	-3.08	0.0022
Month	2	1	-1679290	525266	-3.20	0.0015
Month	3	1	546582	553510	0.99	0.3240
Month	4	1	581147	572396	1.02	0.3106
Month	5	1	794203	658664	1.21	0.2287
Month	6	1	-527227	595320	-0.89	0.3764
Month	7	1	-662583	570743	-1.16	0.2464
Month	8	1	1785049	556122	3.21	0.0014
Month	9	1	12056.7	439842	0.03	0.9781
Month	10	1	337413	473617	0.71	0.4767
Month	11	1	390591	464407	0.84	0.4009
SQR_YenWonRatio		1	58870.4	6269.4	9.39	<.0001
Weekday	Friday	1	301952	365124	0.83	0.4088
Weekday	Monday	1	538265	382133	1.41	0.1598
Weekday	Saturday	1	1014559	361559	2.81	0.0053
Weekday	Sunday	1	1359877	357508	3.80	0.0002
Weekday	Thursday	1	-907175	354051	-2.56	0.0108
Weekday	Tuesday	1	-958804	369843	-2.59	0.0099

The Total Sales are effected by Japanese Tourists, YenWonRatio, Month of April, May, August, weekday Friday, Monday, Saturday and Sunday.

Following can be observed even with the Tableau diagrams:


Discount behavior
Total Sales
28M
26M
24M
22M
20M
18M
16M
14M
12M
10M
8M
6M
4M
2M
0M
-500K
0K
500K
1000K
1500K
2000K
2500K
3000K
3500K
4000K
4500K
5000K
Discount
Store Name
Store B
Total Sales
1,665,400
27,795,650
Japanese Tourists
Total Sales
28M
26M
24M
22M
20M
18M
16M
14M
12M
10M
8M
6M
4M
2M
0M
0K
2K
4K
6K
8K
10K
12K
14K
16K
18K
IMP Japanese Tourists
Store Name
Store B
Total Sales
30,900
27,795,650



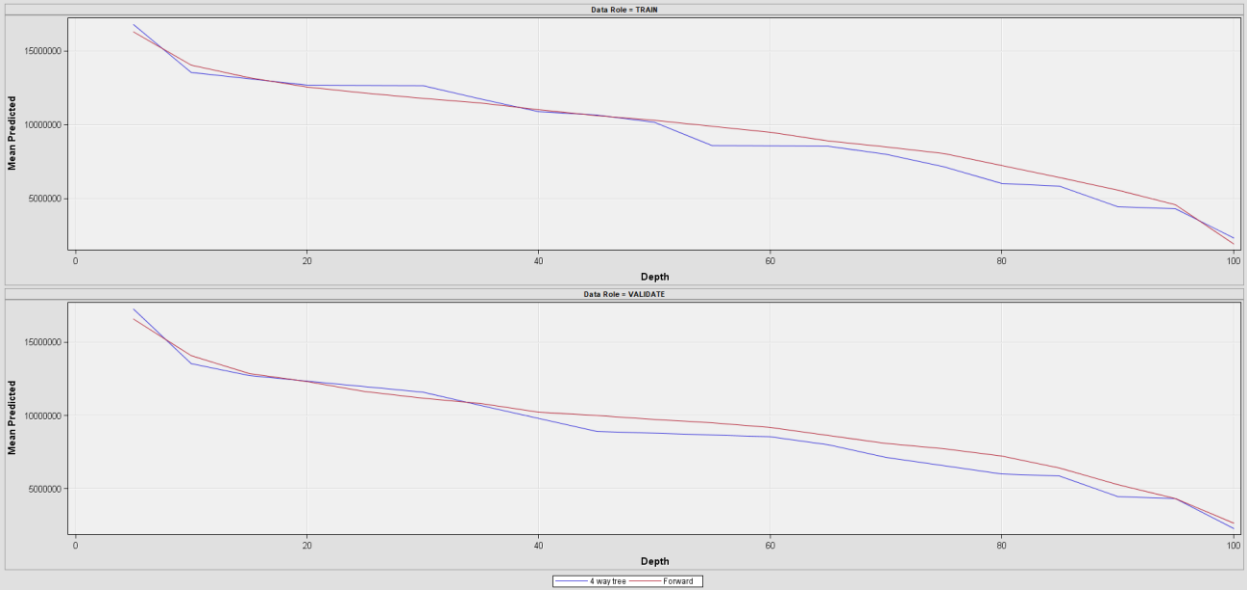
Based on the above diagrams:

- The sales are higher in the months of April, May and August. Thus, the manager should make sure that he keeps a special check on the YenWonRatio rates for the days of these months.
- Since the YenWonRatio rates are strongly affecting the total sales, the manager should be observing its values, especially on the days from Friday to Monday.
- The japanese tourists highly affect the total sales. Thus, the manager should keep a check on the number of tourists visiting Seoul and accordingly, keep more items on sale which are of specific interest to Japanese people.

Part B:Inclusion of Number of Customers and Number of Item.

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Average Error Function	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom
Y	Reg7	Reg7	Forward	Total Sal...	Total Sal...	4.001E12	11413.16	3.508E12	3.508E12	378	16
	Tree5	Tree5	4 way tree	Total Sal...	Total Sal...	5.026E12	.	2.576E12	.	.	.

On building the model, we found that Forward Regression model was performing better compared to others.



The inferences that can be drawn from the prediction models are:

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	-2.132E7	2386544	-8.94	<.0001
LOG___of_Items		1	3676381	413566	8.89	<.0001
Month	1	1	-176255	331602	-0.53	0.5954
Month	2	1	-1244752	329296	-3.78	0.0002
Month	3	1	-1979028	352397	-5.62	<.0001
Month	4	1	-497681	378783	-1.31	0.1897
Month	5	1	1110013	428947	2.59	0.0100
Month	6	1	88814.9	384899	0.23	0.8176
Month	7	1	668892	376787	1.78	0.0767
Month	8	1	1150944	368082	3.13	0.0019
Month	9	1	-1036872	309889	-3.35	0.0009
Month	10	1	869861	318118	2.73	0.0065
Month	11	1	466120	313569	1.49	0.1380
Year	2011	1	-1363909	283703	-4.81	<.0001
Year	2012	1	665996	161210	4.13	<.0001
__of_Customers		1	46487.9	5010.4	9.28	<.0001

The factors affecting the target are Number of items, Month, Year and Number of Customers. Many of the other factors from Part A are not coming into picture here.

- With a unit increase in log of number of items, the total sales seems to have increased by 3676381
- Similarly, With a unit increase in number of customers, the total sales seems to have increased by 46488
- While the sales seems to have reduced by 1363909 for the year 2011.

- Overall sales is poor(loss) in months like February, March, April and September.

STORE C

The general trend observed for store C is:

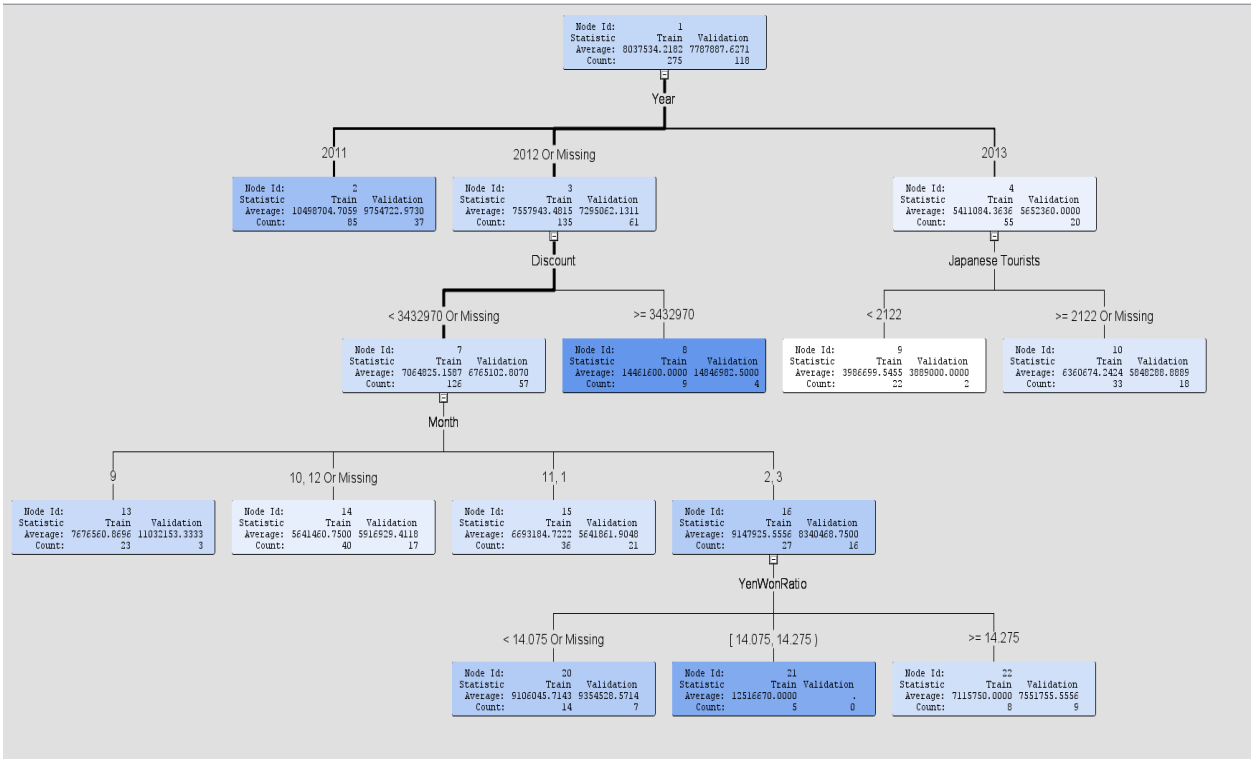
The dotted line shows the trend for the total sales. The colored area shows the forecasting to predict the future sales.

MODEL SELECTION:

Part A: Non Inclusion of Number of Customers and Number of Items

The best model that we get is 4 way decision tree according to the tool.

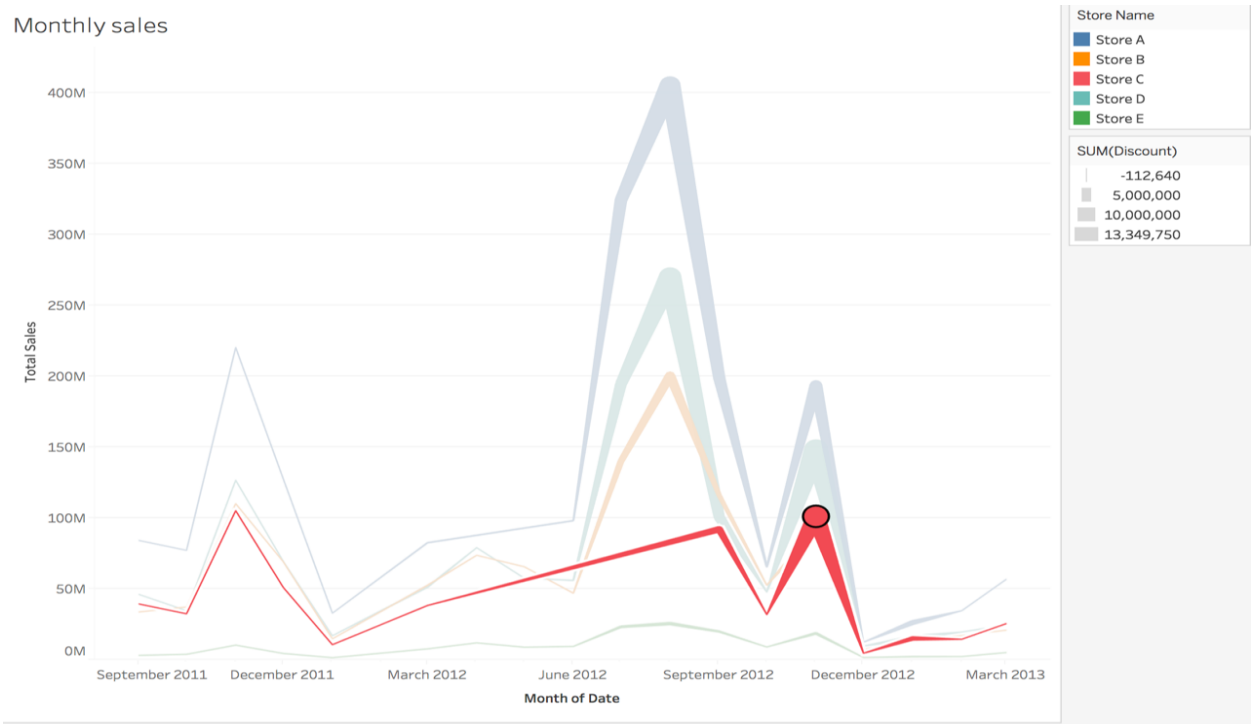
Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Sum of Frequencies	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error
Y	MdlComp5	Tree14	4 way tree	Total Sal...	Total Sales	4.871E12	275	10879015	1.303E15	4.738E12	2176605
Selected Model	p...	Reg3	Backward	Total Sal...	Total Sales	5.347E12	275	8789178	1.365E15	4.964E12	2227997

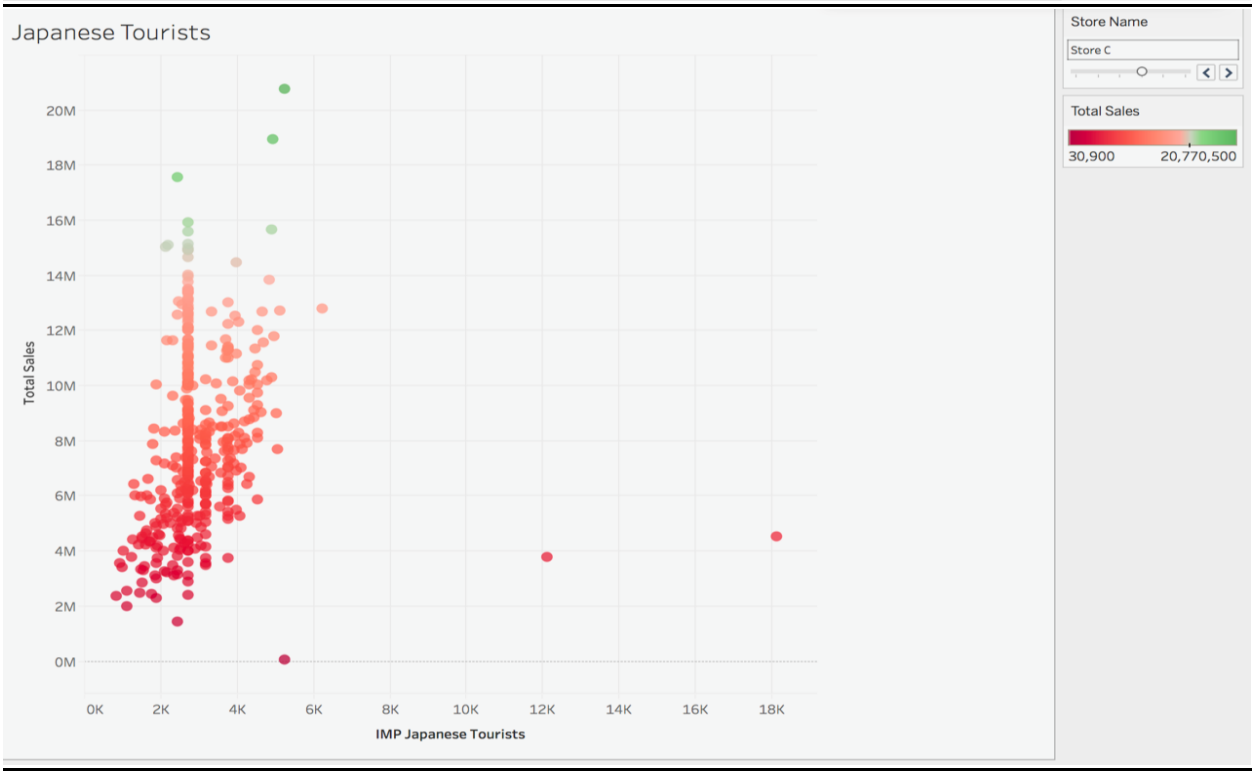


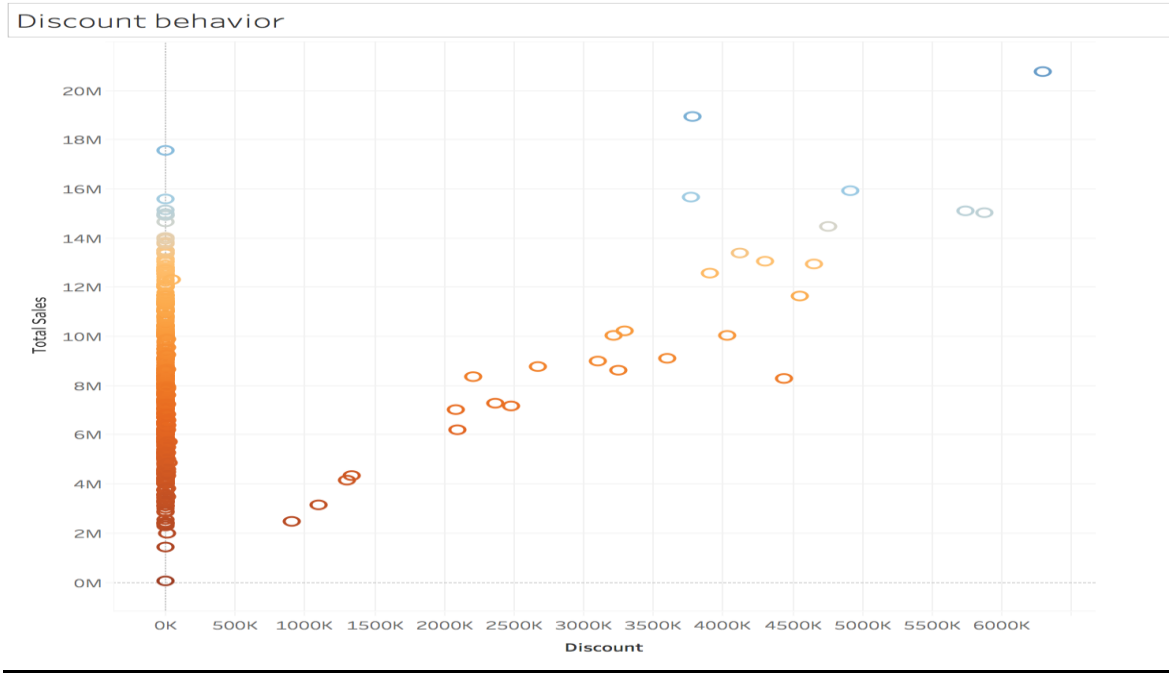
According to the model given, following are the inferences that can be concluded (Tableau diagrams included):

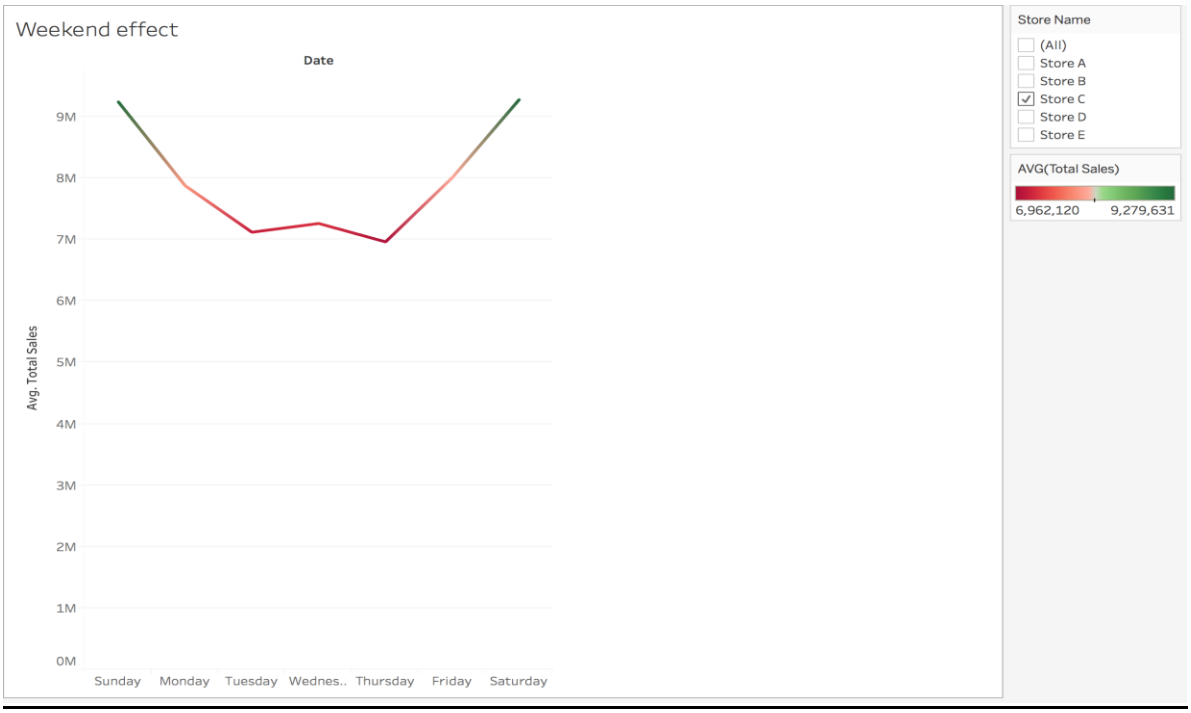
The sales for the store are high in 2011 as compared to 2012. (The peak being in 2011). The thickness represents the discounts offered by the store. It can be observed that, in 2012 the sales soar because a lot of extended holidays fall in 2012, attracting more customers. Thus, even more discounts were offered in this period.

```
if YenWonRatio < 14.075 or MISSING
AND Year IS ONE OF: 2012 or MISSING
AND Month IS ONE OF: 2, 3
AND Discount < 3432970 or MISSING
then
 Tree Node Identifier   = 20
 Number of Observations = 14
 Predicted: Total_Sales = 9106045.7143
```


Monthly sales
Store Name
Store A
Store B
Store C
Store D
Store E
SUM(Discount)
-112,640
5,000,000
10,000,000
13,349,750
400M
350M
300M
250M
200M
150M
100M
50M
0M
Total Sales
September 2011
December 2011
March 2012
June 2012
September 2012
December 2012
March 2013
Month of Date


Japanese Tourists
Store Name
Store C
Total Sales
30,900
20,770,500
20M
18M
16M
14M
12M
10M
8M
6M
4M
2M
0M
Total Sales
0K
2K
4K
6K
8K
10K
12K
14K
16K
18K
IMP Japanese Tourists



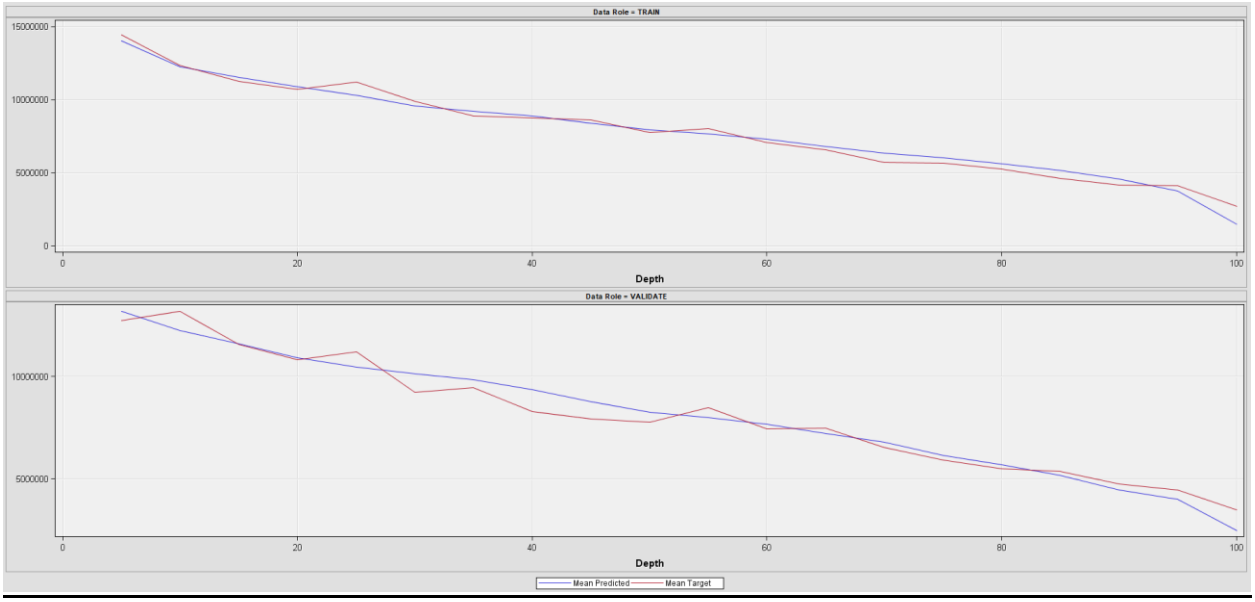


Since there's a strong correlation with YenWonRatio, the manager should keep a check on the rates. The sales peak in the months (September to October) when there are more Japanese extended holidays, so, along with the YenWonRatio, a track of when the extended holidays are falling should be taken care of.

Part B: Inclusion of Number of Customers and Number of Item.

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Sum of Frequencies	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error
Y	Reg10	Reg10	Forward	Total Sal...	Total Sal...	2.721E12	275	8130583	5.041E14	1.833E12	1353867
	Tree7	Tree7	Decision ...	Total Sal...	Total Sal...	3.108E12	275	5364673	4.834E14	1.758E12	1325867

The Forward Regression model was selected by the software. The variables selected are number of Items , Month , YenWonRatio, Year , number of Customers.



The train and validate data do not go hand in hand in this case.

The inferences that can be drawn from the prediction models are:

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	-1.948E7	2250251	-8.66	<.0001
LOG___of_Items		1	2685280	306050	8.77	<.0001
Month	1	1	-705590	287254	-2.46	0.0147
Month	2	1	-751189	252337	-2.98	0.0032
Month	3	1	-445419	297956	-1.49	0.1361
Month	9	1	197983	222980	0.89	0.3754
Month	10	1	314648	224131	1.40	0.1615
Month	11	1	1033435	243836	4.24	<.0001
SQR_YenWonRatio		1	17355.2	7457.5	2.33	0.0207
Year	2011	1	-528086	408283	-1.29	0.1970
Year	2012	1	-48973.0	158761	-0.31	0.7580
__of_Customers		1	51026.5	5085.6	10.03	<.0001

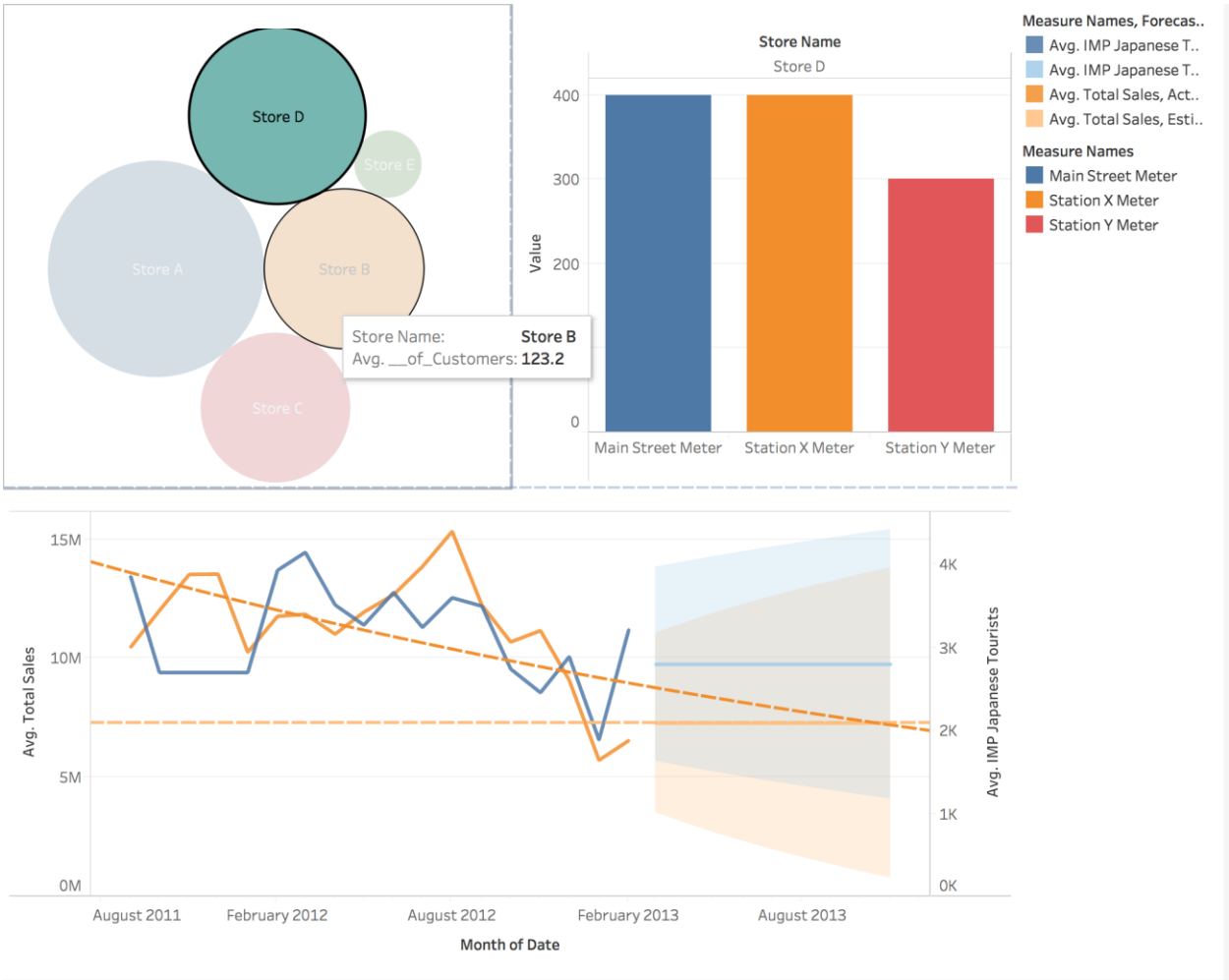
The factors playing role in this model are not the same for Part A.

- With a unit increase in log of number of items, the total sales seems to have increased by 2685280
- Similarly, With a unit increase in number of customers, the total sales seems to have increased by 51026
- While the sales seems to have reduced by 528086 for the year 2011 and by 48973 for year 2012.
- Overall sales is poor(loss) in months like January, February, March.
- With a unit increase in square of YenWonRatio, the total sales seems to have increased by 17355.

STORE D

The general trend seen:

The dotted line shows the trend for the total sales. The colored area shows the forecasting to predict the future sales.



MODEL SELECTION:

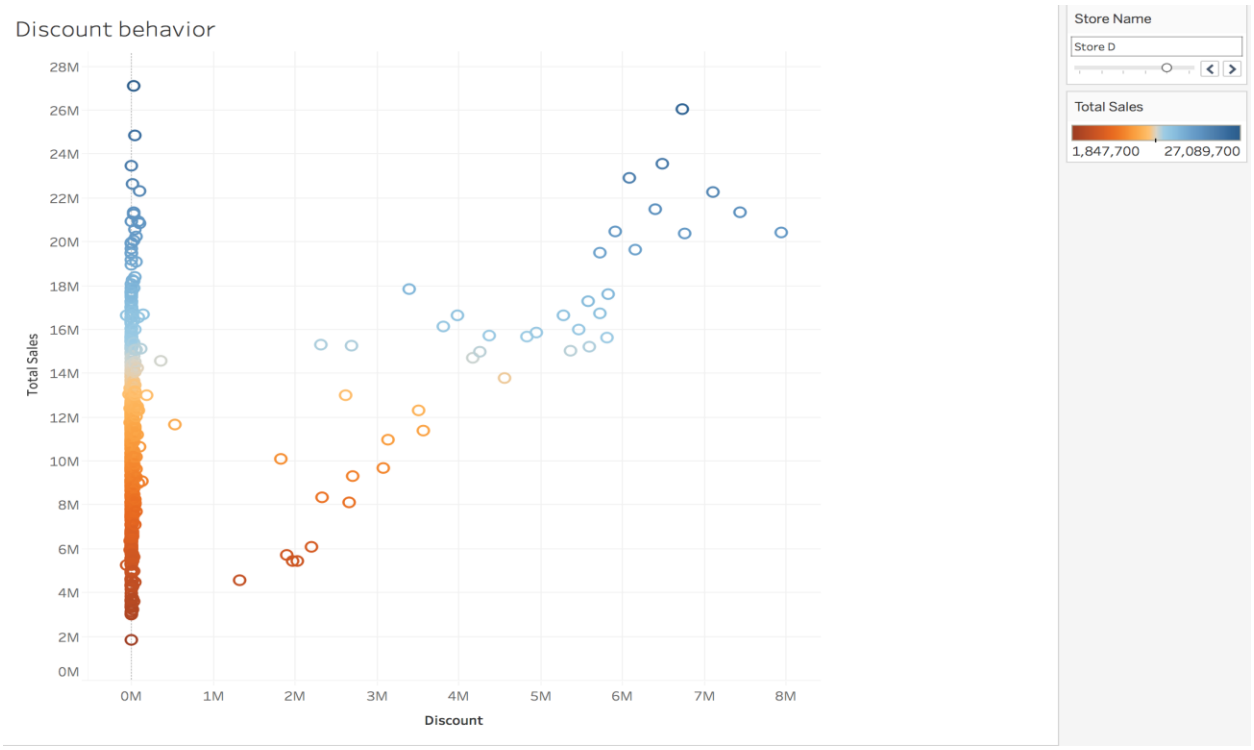
Part A: Non Inclusion of Number of Customers and Number of Items

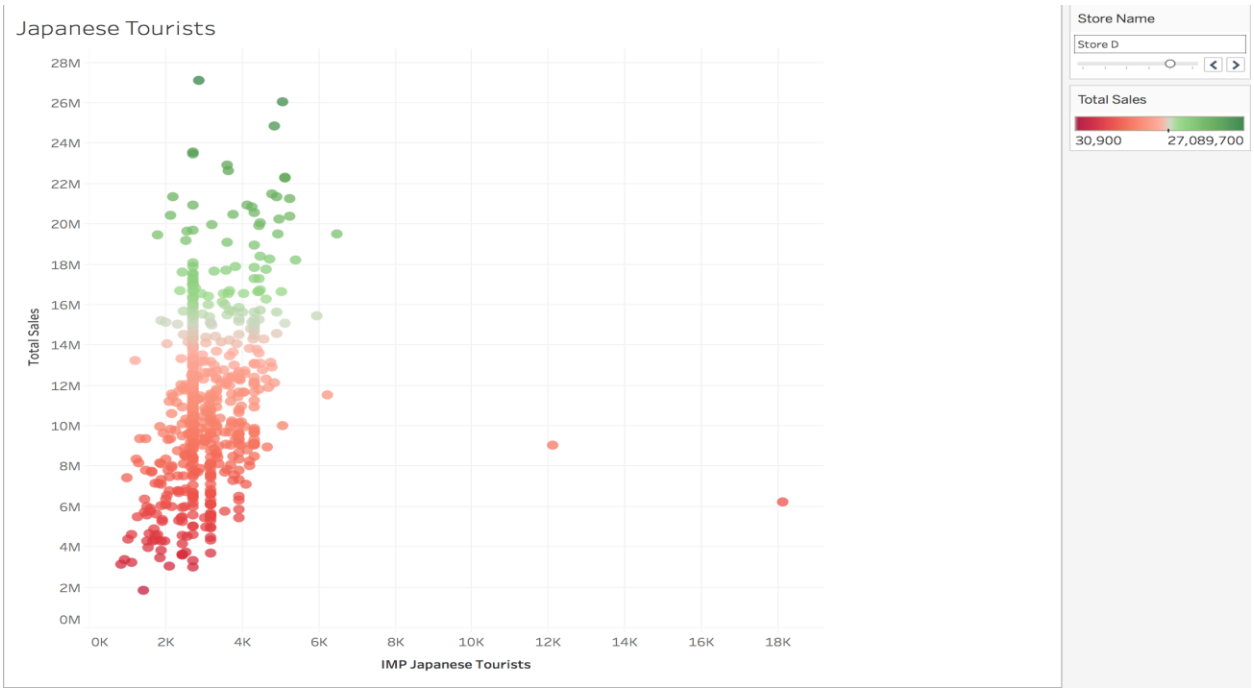
Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Sum of Frequencies	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error
Y	MdlComp...	Reg6	Backward	Total Sal...	Total Sales	8.806E12	391	11479181	3.303E15	8.448E12	2906480
	MdlComp7	Tree2	Decision ...	Total Sal...	Total Sales	9.052E12	391	12780502	3.037E15	7.767E12	2787017

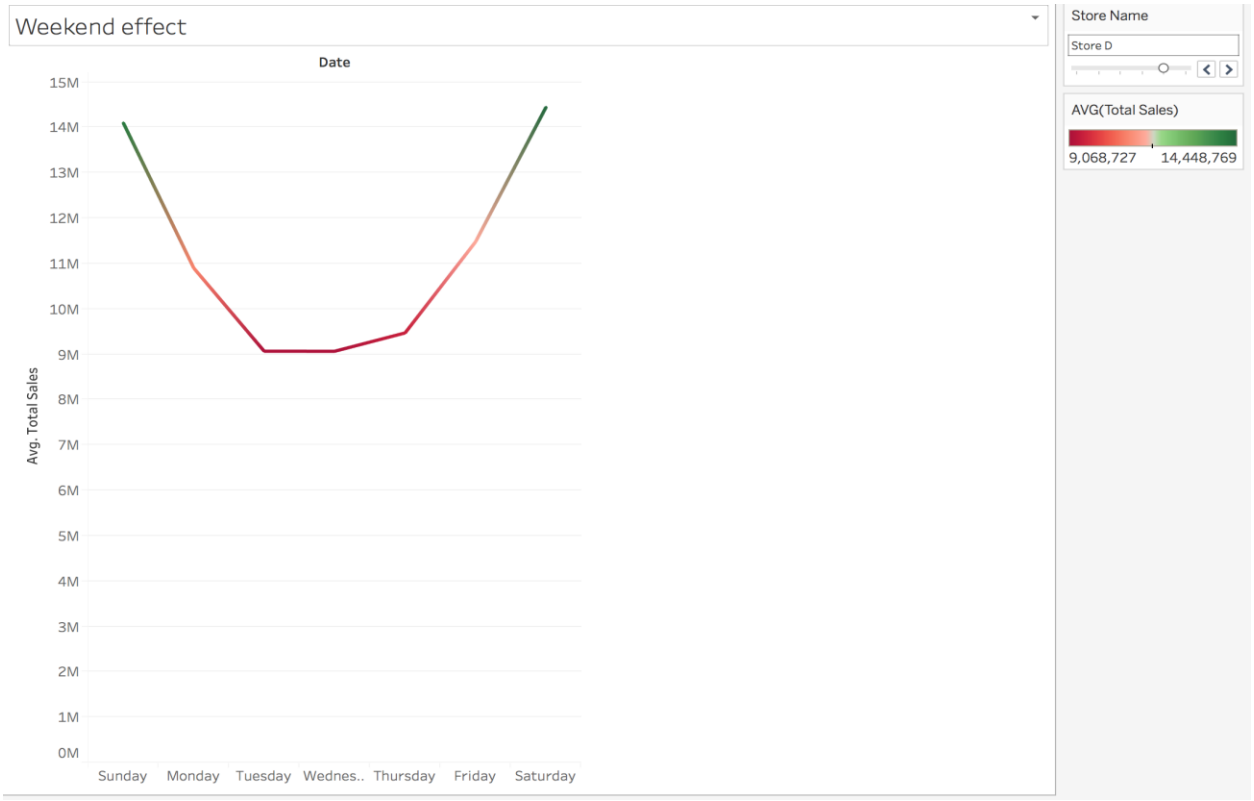
The inferences that can be drawn from the prediction models are:

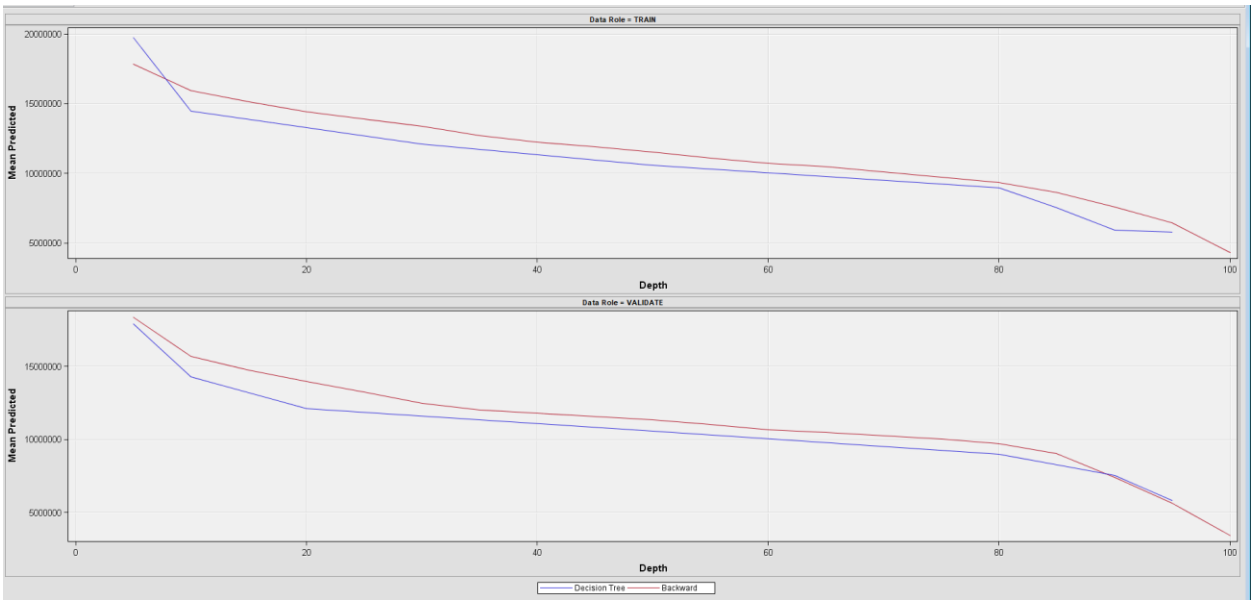
Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	-2882920	936442	-3.08	0.0022
ActualHighTemp		1	6896.1	1834.1	3.76	0.0002
Distance_from_Main_Street_Meter_		0	0	.	.	.
Distance_from_Station_X_Meter_		0	0	.	.	.
Distance_from_Station_Y_Meter_		0	0	.	.	.
Holiday	0	1	-92571.6	58193.8	-1.59	0.1125
LOG_Discount		1	54779.8	7500.1	7.30	<.0001
LOG_IMP_Japanese_Tourists		1	624657	134091	4.66	<.0001
SQR_YenWonRatio		1	-4953.4	1449.0	-3.42	0.0007
Weekday	Friday	1	101396	79088.1	1.28	0.2006
Weekday	Monday	1	16576.8	82347.2	0.20	0.8406
Weekday	Saturday	1	49387.9	78583.2	0.63	0.5301
Weekday	Sunday	1	99442.6	78438.1	1.27	0.2056
Weekday	Thursday	1	-19475.8	77300.6	-0.25	0.8012
Weekday	Tuesday	1	-129884	82615.4	-1.57	0.1167




Japanese Tourists
Total Sales
28M
26M
24M
22M
20M
18M
16M
14M
12M
10M
8M
6M
4M
2M
0M
0K
2K
4K
6K
8K
10K
12K
14K
16K
18K
IMP Japanese Tourists
Store Name
Store D
Total Sales
30,900
27,089,700


Weekend effect
Date
Avg. Total Sales
15M
14M
13M
12M
11M
10M
9M
8M
7M
6M
5M
4M
3M
2M
1M
0M
Sunday
Monday
Tuesday
Wednes..
Thursday
Friday
Saturday
Store Name
Store D
AVG(Total Sales)
9,068,727
14,448,769



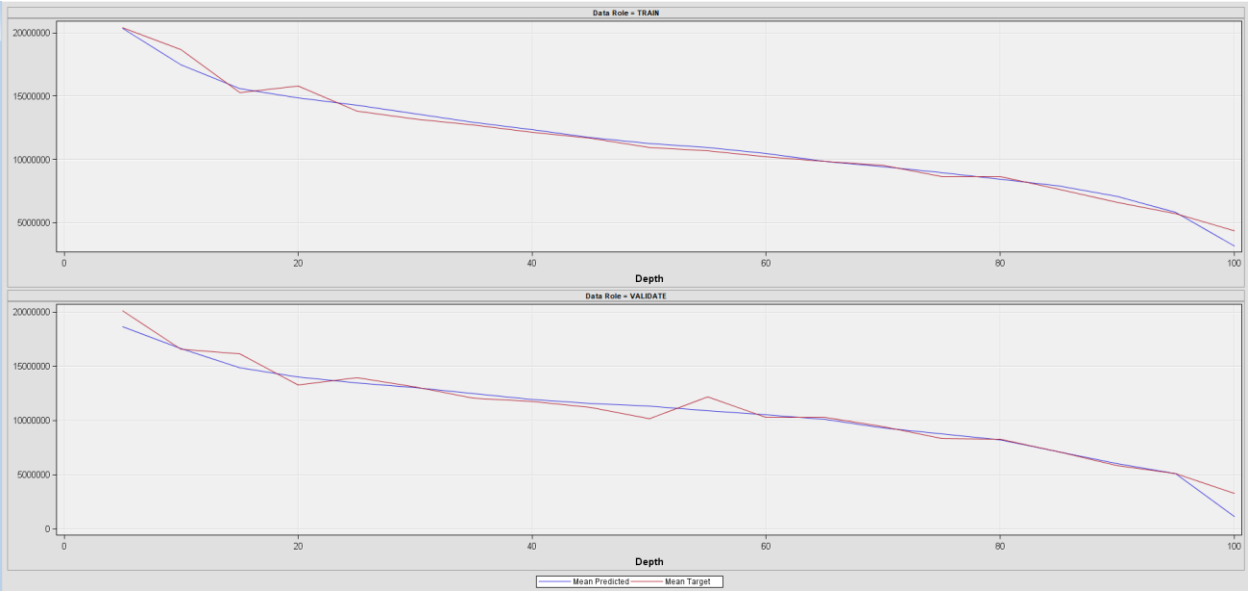
Inference drawn are:

- The sales are higher on friday, saturday and sunday. Thus, the manager should make sure that the discounts offered are more on these days.
- YenWonRatio effects strongly. Thus, apart from being kept a track on these days, they should also be kept a track of regularly. The days when the rate's high, discounts can be offered to attract customers.
- Extended holidays with warmer temperatures attract more customers. Thus, the manager should ensure that on days like these, proper discounts are given.
- More Japanese customers, more the sales. Special offers for Japanese products should

Part B: **Inclusion of Number of Customers and Number of Item.**

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Sum of Frequencies	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error
Y	Reg13	Reg13	Forward	Total Sal...	Total Sal...	2.72E12	391	7391241	1.111E15	2.841E12	1685541
	Tree11	Tree11	4 way tree	Total Sal...	Total Sal...	3.663E12	391	6632622	1.304E15	3.336E12	1826386

Model selected here is Forward Regression.



The features that contribute for model building in this case are Discount , Number of Items Month, YenWonRatio, Year, number of Customers.

The inferences that can be drawn from the prediction models are:

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	-2.577E7	3067215	-8.40	<.0001
LOG_Discount		1	-429865	85964.8	-5.00	<.0001
LOG___of_Items		1	3765902	461951	8.15	<.0001
Month	1	1	-1080431	344138	-3.14	0.0018
Month	2	1	-1744668	311257	-5.61	<.0001
Month	3	1	-1605684	309243	-5.19	<.0001
Month	4	1	205215	390205	0.53	0.5993
Month	5	1	-12463.7	364077	-0.03	0.9727
Month	6	1	-304088	381388	-0.80	0.4258
Month	7	1	1265207	350599	3.61	0.0003
Month	8	1	712520	343730	2.07	0.0389
Month	9	1	-743892	298450	-2.49	0.0131
Month	10	1	129372	293914	0.44	0.6601
Month	11	1	1447551	326350	4.44	<.0001
SQR_YenWonRatio		1	30812.9	9114.5	3.38	0.0008
Year	2011	1	-2645703	503798	-5.25	<.0001
Year	2012	1	244587	186408	1.31	0.1903
__of_Customers		1	59231.7	4360.2	13.58	<.0001

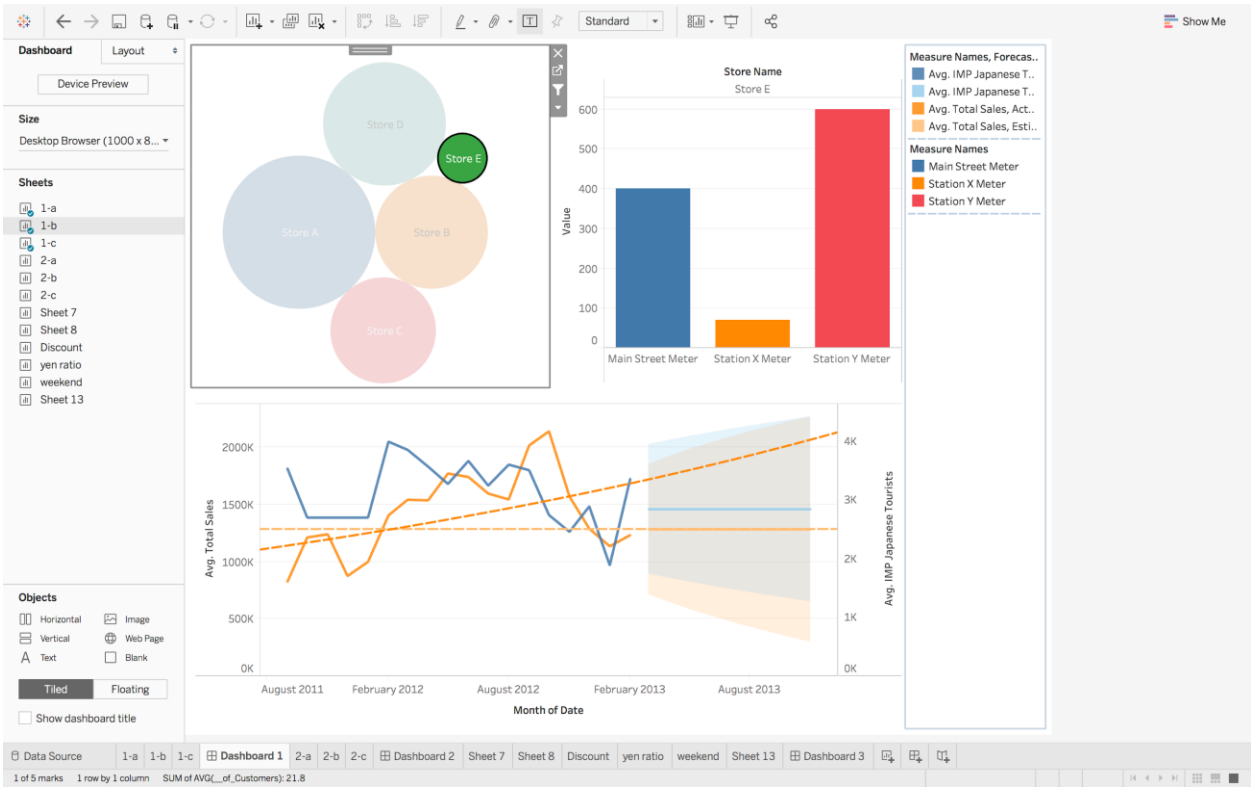
The factors playing role in this model are not the same for Part A.

- With a unit increase in log of number of items, the total sales seems to have increased by 3765902
- Similarly, With a unit increase in number of customers, the total sales seems to have increased by 59231
- While the sales seems to have reduced by 2645703 for the year 2011 and increased by 244587 for year 2012.
- Overall sales is poor(loss) in months like January, February, March.
- With a unit increase in square of YenWonRatio, the total sales seems to have increased by 30812.
- With a unit increase in log of discount, the total sales seems to have reduced by 429865

STORE E

The general trend is shown as follows:

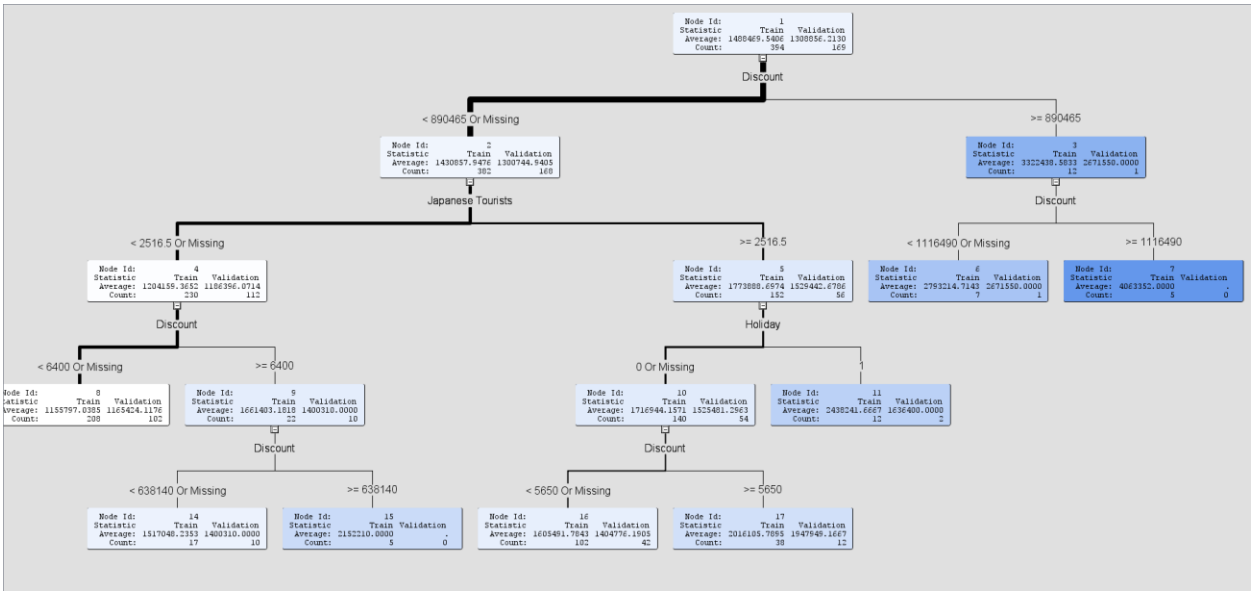
The dotted line shows the trend for the total sales. The colored area shows the forecasting to predict the future sales. Store E has a good trend line hinting there is a likeliness of increment in the total sales in the near future.



MODEL SELECTION:

Part A: Non Inclusion of Number of Customers and Number of Items

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Sum of Frequencies	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error
Y	MdlComp8	Tree4	Decision ...	Total Sal...	Total Sales	4.096E11	394	1946394	1.222E14	3.1E11	556818.3
	MdlComp...	Reg10	Stepwise	Total Sal...	Total Sales	4.456E11	394	2749598	1.523E14	3.866E11	621789.6



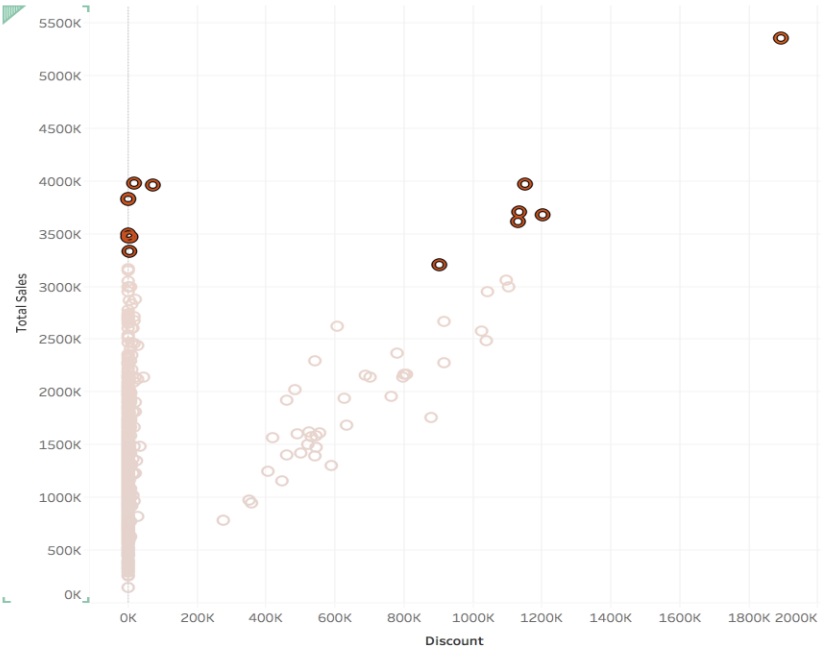

```
*-------------------------------------------------------------*
if Japanese Tourists >= 2516.5
AND Holiday IS ONE OF: 1
AND Discount < 890465 or MISSING
then
 Tree Node Identifier   = 11
 Number of Observations = 12
 Predicted: Total_Sales = 2438241.6667

*-------------------------------------------------------------*
```

The inference drawn:

- The sales are high when there's an extended holiday. The manager should keep a track of all the extended holidays.
- The discounts offered on extended holidays should be kept more by the manager as high discounts correspond to higher sales.
- Higher japanese tourists in the store lead to higher sales. Thus, special offers on Japanese items should be kept to ensure more tourists are attracted by the manager's store.

Discount behavior



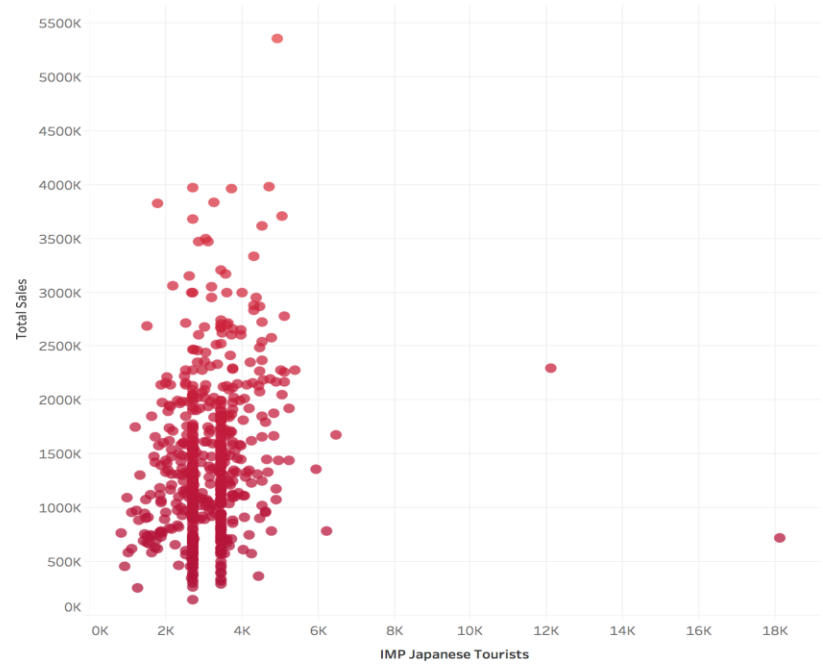
Store Name

Store E

Total Sales

148,500 14,410,200

Japanese Tourists

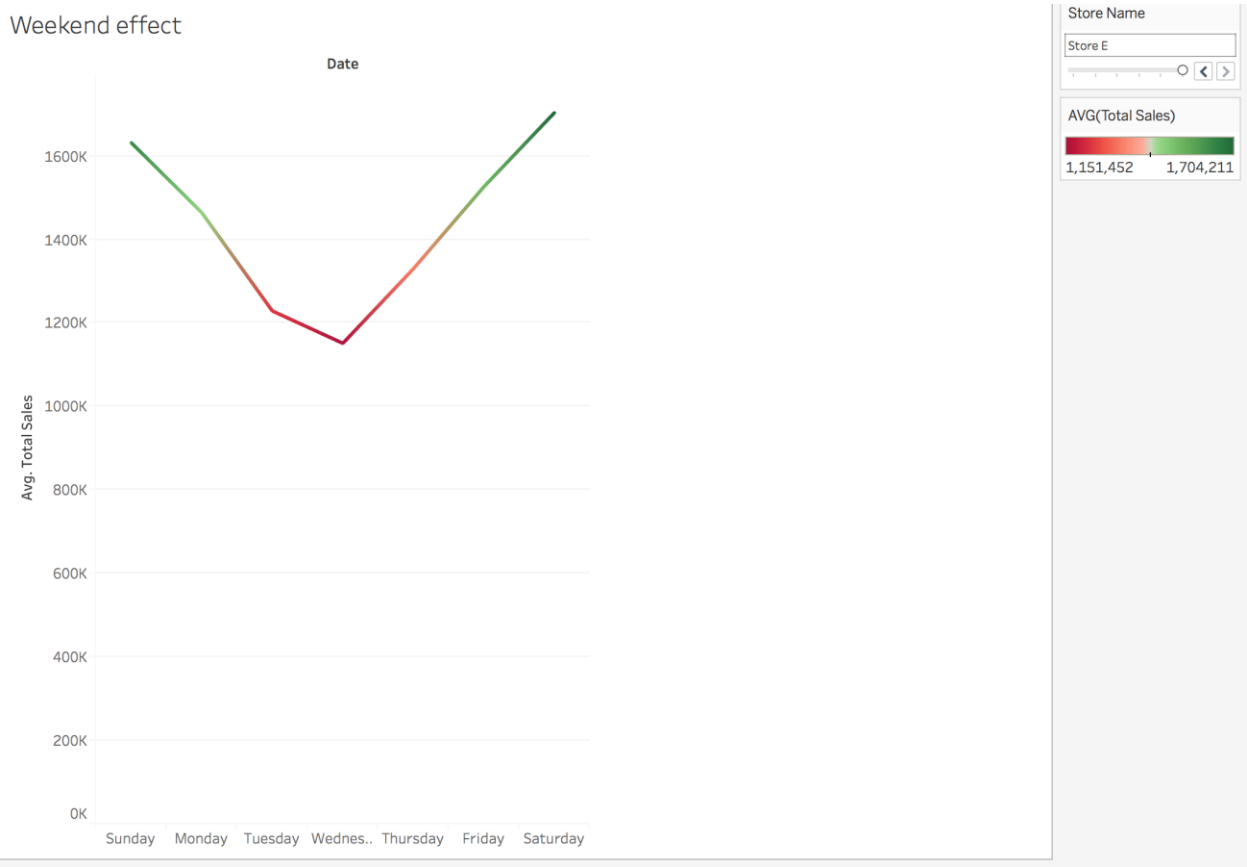


Store Name

Store E

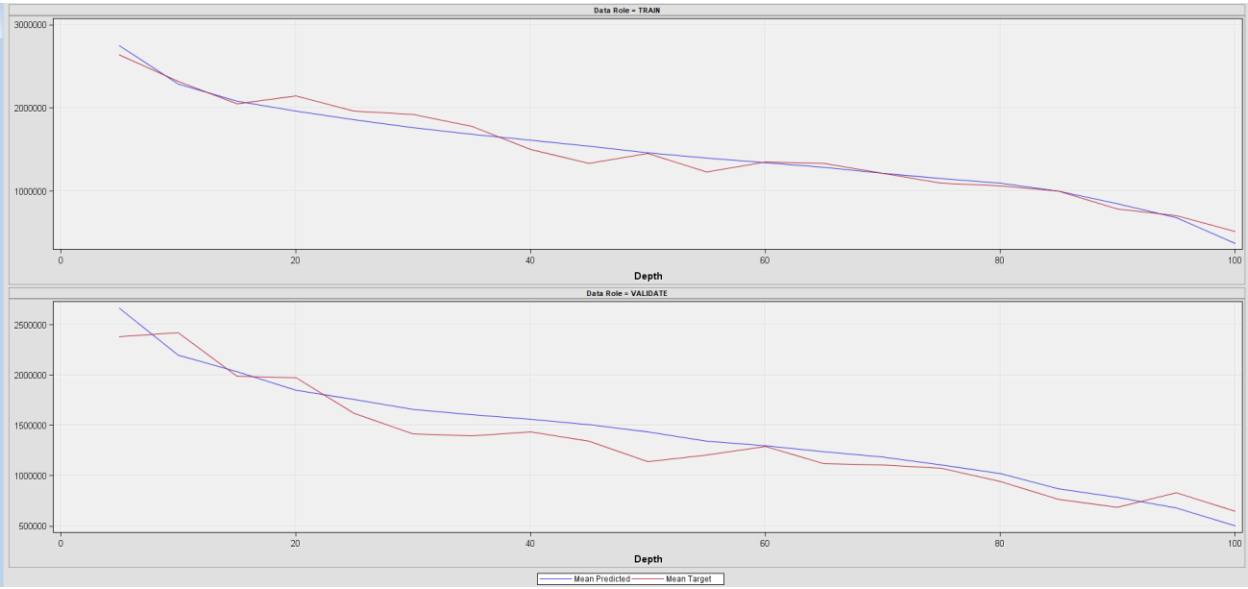
Total Sales

30,900 15,058,100



Part B: Inclusion of Number of Customers and Number of Item.

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Average Error Function	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom
Y	Reg16	Reg16	Forward	Total Sal...	Total Sal...	1.887E11	10354.82	2.515E11	2.515E11	388	6
	Tree14	Tree14	4 way tree	Total Sal...	Total Sal...	2.28E11	.	2.684E11	.	.	.



The inferences that can be drawn from the prediction models are:

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	-1809117	272512	-6.64	<.0001
LOG_Discount		1	21068.9	6334.9	3.33	0.0010
LOG___of_Items		1	497744	62046.0	8.02	<.0001
Year	2011	1	-209673	48935.3	-4.28	<.0001
Year	2012	1	155324	38554.8	4.03	<.0001
__of_Customers		1	23542.7	4204.1	5.60	<.0001

Forward Regression model is the best amongst all. The features selected here are Discount , number of tems, Year, number of Customers..

The factors playing role in this model are not the same for Part A.

- With a unit increase in log of number of items, the total sales seems to have increased by 497744.
- Similarly, With a unit increase in number of customers, the total sales seems to have increased by 23542
- While the sales seems to have reduced by 209673 for the year 2011 and increased by 155324 for year 2012.
- With a unit increase in log of discount, the total sales seems to have increased by 21069.

Impact of the change in ownership at Store A and change in ownership & location at Store C

STORE A

The store was split code wise to understand the factors affecting before closing and after reopening the store..

Store A : Before Closing

Analysis of Maximum Likelihood Estimates

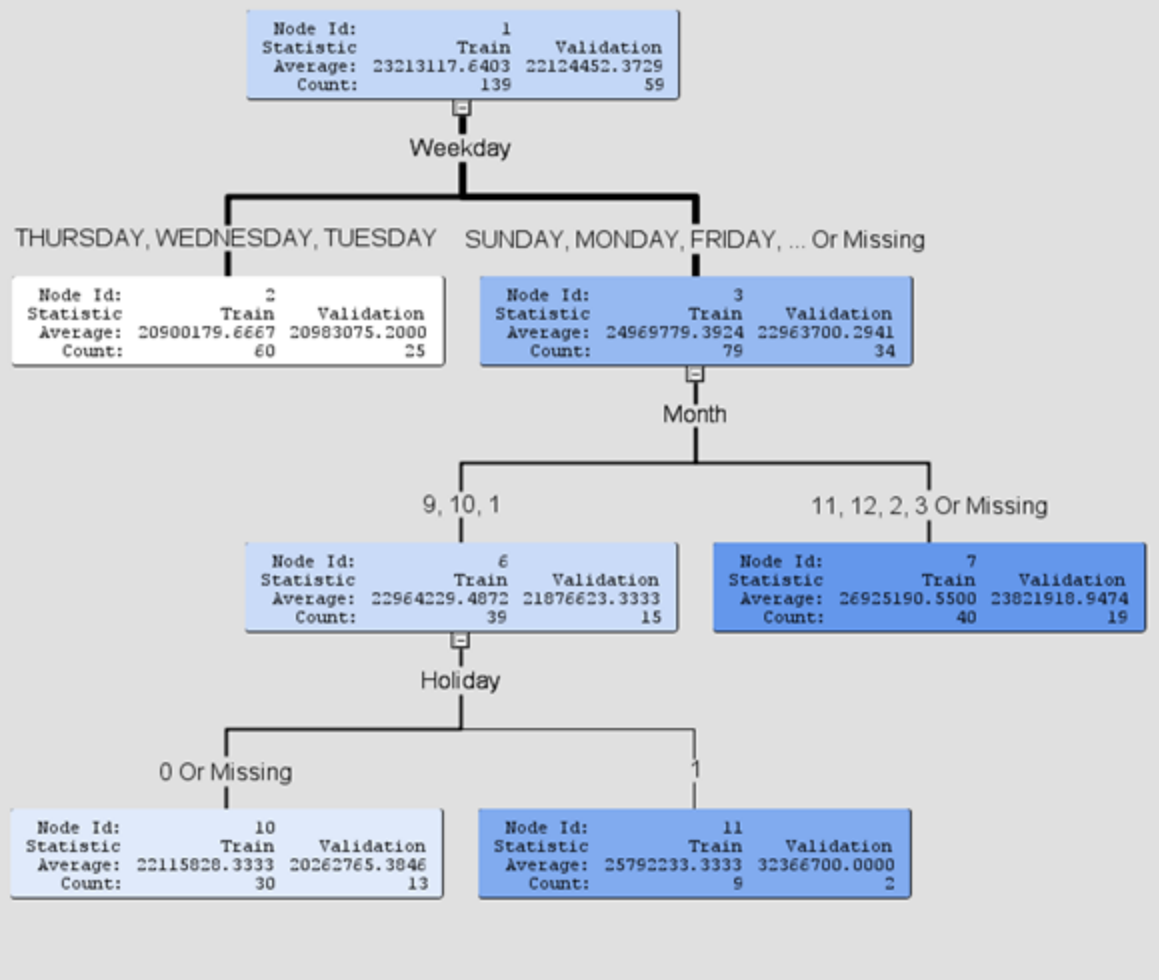
Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	-7.232E7	40612369	-1.78	0.0774
LOG_IMP_Japanese_Tourists		1	7114081	4507018	1.58	0.1170
Month	1	1	-3323943	1130421	-2.94	0.0039
Month	2	1	-531337	1235598	-0.43	0.6679
Month	3	1	2042969	1737984	1.18	0.2421
Month	9	1	-996493	1244143	-0.80	0.4247
Month	10	1	-2782275	1193798	-2.33	0.0214
Month	11	1	2609617	1157903	2.25	0.0260
SQR_YenWonRatio		1	158032	39770.2	3.97	0.0001
Weekday	Friday	1	551204	899372	0.61	0.5411
Weekday	Monday	1	1200030	974734	1.23	0.2206
Weekday	Saturday	1	2720983	899380	3.03	0.0030
Weekday	Sunday	1	2174904	874140	2.49	0.0142
Weekday	Thursday	1	-1280894	874183	-1.47	0.1454
Weekday	Tuesday	1	-3622288	943512	-3.84	0.0002

The Months January, February, March, September, October and November play important roles in maximizing the sales. On the other hand, Days from Monday thru Sunday except Wednesdays, are significant

The factors affecting sales before the closing the store are Japanese Tourists, Month, Yen-Won Ration and Weekday. The regression model performs better than the Decision Tree model

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Sum of Frequencies	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error
Y	Reg20	Reg20	Backward	Total Sal...	Total Sales	1.457E13	139	13085881	2.231E15	1.605E13	4006706
	Tree16	Tree16	Decision ...	Total Sal...	Total Sales	1.811E13	139	13165109	2.731E15	1.965E13	4432810

Let’s explore the variables from the decision tree.



The main variable that the data was split into is Weekday. Days related to Weekend such as Sunday, Monday, Friday and Saturday take more precedence over other days. It can be concluded that Sales happens higher over the weekend than the Weekdays.
Beyond that, Holidays also play a major role. If there is Holiday for Japan, the sales seems to be high.

Store A : After Reopening

The store was reopened on June 19, 2012 under a new ownership.

Regression models and Decision Tree models were used on the data set to understand the factors affecting the sales after new ownership came into picture. Comparing the two models, the decision tree model seems to work better than the Forward regression model.

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Sum of Frequencies	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error
Y	Tree6	Tree6	Decision ...	Total Sal...	Total Sales	1.652E13	190	14331804	2.53E15	1.332E13	3649409
	Reg22	Reg22	Forward	Total Sal...	Total Sales	1.821E13	190	12289406	2.712E15	1.428E13	3778346

Before analyzing the decision tree, let's take a look at the variables the Regression model finds significant.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	17602653	314901	55.90	<.0001
Month	1	1	-5509898	762084	-7.23	<.0001
Month	2	1	-5548125	921230	-6.02	<.0001
Month	3	1	-2304565	1032194	-2.23	0.0268
Month	6	1	5258662	1497527	3.51	0.0006
Month	7	1	5075079	800898	6.34	<.0001
Month	8	1	7908290	822485	9.62	<.0001
Month	9	1	3373255	832902	4.05	<.0001
Month	10	1	-2700585	834937	-3.23	0.0015
Month	11	1	-1603821	894128	-1.79	0.0746
Weekday	Friday	1	198747	668195	0.30	0.7665
Weekday	Monday	1	-405569	687156	-0.59	0.5558
Weekday	Saturday	1	3137997	751241	4.18	<.0001
Weekday	Sunday	1	4837820	779257	6.21	<.0001
Weekday	Thursday	1	-2390592	689517	-3.47	0.0007
Weekday	Tuesday	1	-2484814	676684	-3.67	0.0003

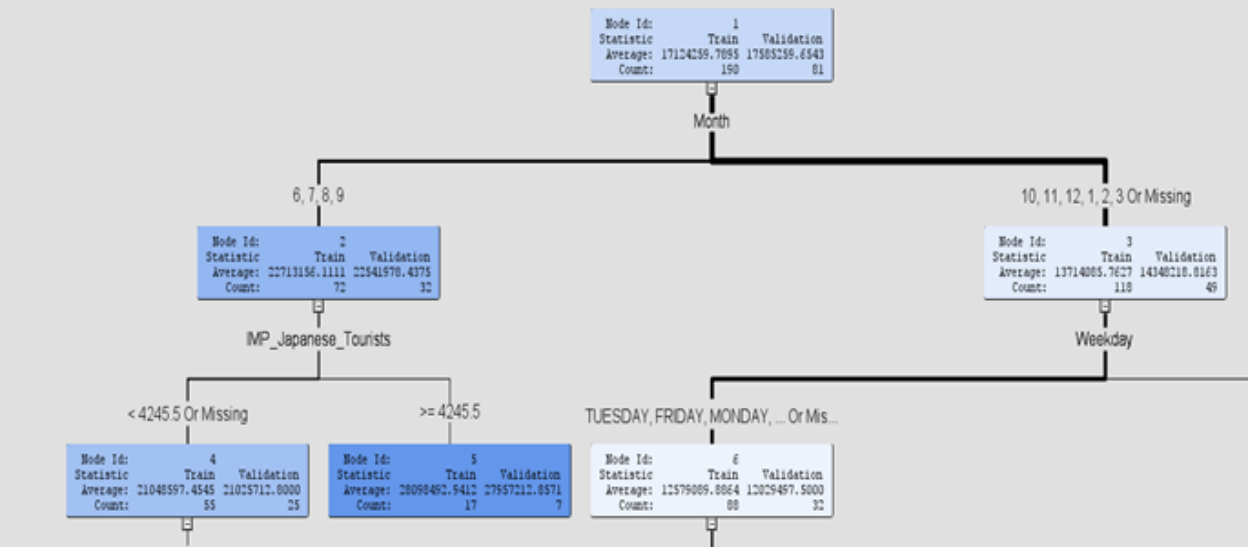
The regression model shows only Weekday and Month are significant, the same as it was for the previous data.

From the decision trees, we can find that variables such as Japanese Tourists, Actual High temperature also taken into consideration other than Weekday and Month.

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
Month	Month	2	1.0000	1.0000	1.0000
IMP_Japanese_Tourists	IMP_Japanese_Tourists	4	0.5491	0.5403	0.9839
Weekday	Weekday	1	0.3438	0.5797	1.6860
ActualHighTemp	ActualHighTemp	1	0.1927	0.4037	2.0950

For months of June, July, August, September, the sales are high and also, another important factor affecting the sales is the Japanese tourists. When the number of Japanese Customers are higher than 4245.5, the sales is better. Therefore, the higher the number of Japanese tourists, the better the sales is.



Thus, we can conclude that, after the new ownership, more number of Japanese tourists started visiting Seoul , thus increasing the sales in the stores. Also, temperature is also playing an important role in attracting people to the store.

STORE C

When the store is reopened

The distance of the Store is highlighted as below: (A decrease in the distance observed)

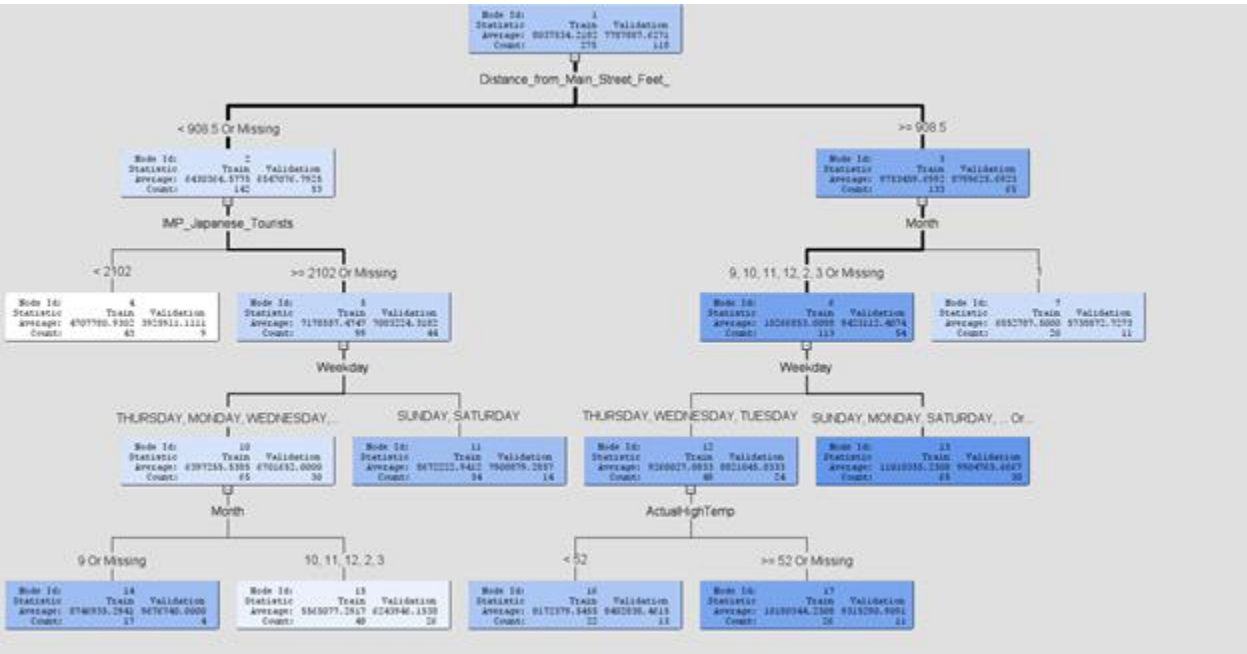
The distance from main street is decreased by around 60 meters.


Store D
Store E
Store A
Store B
Store C
Store Name
Store C
Value
500
400
300
200
100
0
Main Street Meter
Station X Meter
Station Y Meter
Measure Names, Forecas..
Avg. IMP Japanese T..
Avg. IMP Japanese T..
Avg. Total Sales, Act..
Avg. Total Sales, Esti..
Measure Names
Main Street Meter
Station X Meter
Station Y Meter
Avg. Total Sales
12M
10M
8M
6M
4M
2M
0M
Avg. IMP Japanese Tourists
4K
3K
2K
1K
0K
August 2011
February 2012
August 2012
February 2013
August 2013
Month of Date


Store D
Store E
Store A
Store B
Store C
Store Name
Store C
Value
500
400
300
200
100
0
Main Street Meter
Station X Meter
Station Y Meter
Measure Names, Forecas..
Avg. IMP Japanese T..
Avg. IMP Japanese T..
Avg. Total Sales, Act..
Avg. Total Sales, Esti..
Measure Names
Main Street Meter
Station X Meter
Station Y Meter
Avg. Total Sales
12M
10M
8M
6M
4M
2M
0M
Avg. IMP Japanese Tourists
4K
3K
2K
1K
0K
August 2011
February 2012
August 2012
February 2013
August 2013
Month of Date

Closing and Reopening of Store C:

The store was relocated under the new ownership. Since there was a relocation, the distances from the metro stations and the thoroughfare also changed accordingly.



The result from the decision tree shows that when the distance from the main street thoroughfare is smaller, the sales is higher. This indicates that the new ownership and the location change attract more customers to the shop. Other important factors are the Japanese Tourists and Month. The sales is high during the months of February, March, September through December. Also, Japanese tourists as customers are more after the location changed.

And, weekdays play a significant role here. Sunday through Saturday, the sales is high, indicating there might be more Japanese tourists visiting the place.

How other stores affect store B?

Store A has very high total sales. Store D and store C are the biggest competitor for Store B so Mr. Choe should closely monitor these stores. Store D encourages discounts on long weekends as seen from the graph which has been a strategy of store B as well. Store C has been competing very close with the sales over long weekend.

Store D does not perform very well on Tuesdays, Wednesdays and Thursdays. This is a window of opportunity of store B to offer higher discounts and other marketing techniques to draw more customers. Additionally store D is closer from Y station which makes it difficult to compete with store D in that area. Hence Mr. Choe should focus marketing at the main street and station X which serves as an advantage to store B.

Store C has recently moved closer to main street so store B will likely face more competition to advertise on that street. Store B has advantage over at store C at X station. Thus, Mr.Choe should aggressively advertise on X station because this gives them the advantage over both store D and C.

Store A and B has huge advantage on account of Yen ratio. To compete better with store A in this aspect store B can provide better conversion rate discounts to customers.

Although store E has much lower sales, it shows a trend line indicating increasing sales over coming future. Hence Mr. Choe must monitor the activities Store E is doing which can be beneficial (Eg- kind of discount provided). Store E is very close to X station which can give a some competition.

Number of holidays is very high during August to December. Store D takes the highest advantage of the the holidays, hence Mr.Choe must compete more intensely over this period of time and avoid any renovation or construction in the store. January has many holidays but for some reason has lower sales compared to other high holidays months.