A·P·U
ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION

CT045-3-M-ABAV
ASSIGNMENT (PART III)

NAME: AHMAD FUAD BIN KHALIT
STUDENT ID:TP058497
LECTURER: DR. PREETHI SUBRAMANIAM
WEIGHTAGE: 50%

INSTRUCTIONS TO CANDIDATES:

1 **Submit your assignment at the administrative counter**

2 **Students are advised to underpin their answers with the use of references (cited using the Harvard Name System of Referencing)**

3 **Late submission will be awarded zero (0) unless Extenuating Circumstances (EC) are upheld**

4 **Cases of plagiarism will be penalized**

5 **The assignment should be bound in an appropriate style (comb bound or stapled).**

6 **Where the assignment should be submitted in both hardcopy and softcopy, the softcopy of the written assignment and source code (where appropriate) should be on a CD in an envelope / CD cover and attached to the hardcopy.**

Table of Contents

7.0 Data Preparation & Description

7.1 Metadata Explained

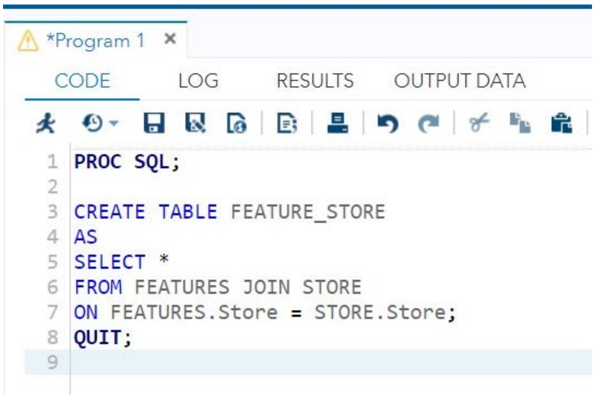
Data source for this project was from https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/data that was published on 2014. It contained Walmart historical sales data for 45 stores in different region from Feb 2010 until Nov 2012 where its cover in 3 saperated dataset files namely "features.csv", "stores.csv" and "train.csv". Below are the detailed for each dataset files:-

1. **features.csv**- these files have information about the internal & external variable that effected Walmart sales. First it contain date which shows in weekly date from 2010-2012. The internal variable is 5 markdown data that related to Walmart promotional data that they are running. The external variable is including average temperature, unemployment rate, fuel price, CPI (Customer price index) and IsHoliday which is in binary format indicate there is special holiday on the week.
2. **Stores.csv**- it contain store, store type and store size. Store describe each of their store in numerical value (1-45), store type describing their type of store in string format (A,B,C) and store size describe the size in square foot in numerical value.
3. **train**.csv- Contain 5 records. Each of variable that contain store, date and IsHoliday already explained as per above. Additional variable is "dept" which is numerical value representing the 99 department on each of the 45 store. Then there is "Weekly Sales" variable which represent the actual sales per week Walmart made during this period.

7.2 Data Preprocessing

For the preprocessing part, we will start first with joining the table using SAS Studio PROC SQL and we will use all the 3 files combine into one files. Below are the step taken to combine all the 3 files into 1 files.

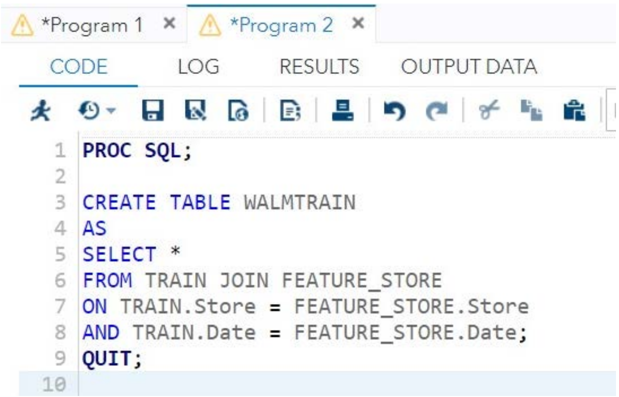
7.2.1 Join Table



```
PROC SQL;

CREATE TABLE FEATURE_STORE
AS
SELECT *
FROM FEATURES JOIN STORE
ON FEATURES.Store = STORE.Store;
QUIT;

```

Figure 7.1: SAS Studio Proc SQL

1. First we create new table name Feature_Store where we combine features.csv and stores.csv by identifying the join condition that used primary key and foreign key relationship.



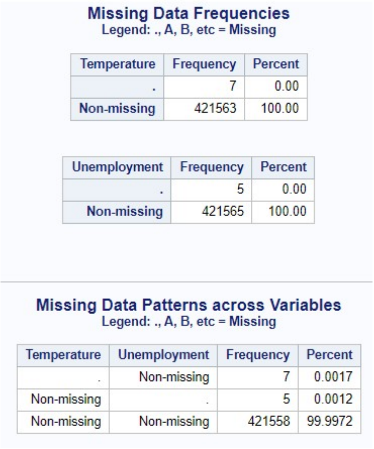
```
PROC SQL;

CREATE TABLE WALMTRAIN
AS
SELECT *
FROM TRAIN JOIN FEATURE_STORE
ON TRAIN.Store = FEATURE_STORE.Store
AND TRAIN.Date = FEATURE_STORE.Date;
QUIT;

```

Figure 7.2: SAS Studio Proc SQL

2. Next step we create our join table name “WALMTRAIN” where we join the newly created table “FEATURE_STORE” and join in with train.csv. then we save the file into our local machine.

7.2.2 Missing Data Imputation

Then we explore our data for missing value. Result was shown as below.



Missing Data Frequencies
Legend: ., A, B, etc = Missing

Temperature	Frequency	Percent
.	7	0.00
Non-missing	421563	100.00

Unemployment	Frequency	Percent
.	5	0.00
Non-missing	421565	100.00

Missing Data Patterns across Variables
Legend: ., A, B, etc = Missing

Temperature	Unemployment	Frequency	Percent
.	Non-missing	7	0.0017
Non-missing	.	5	0.0012
Non-missing	Non-missing	421558	99.9972

Figure 7.3: Missing Data

There is 2 variable that contains missing value. Those variables are Temperature and Unemployment. Temperature contains 7 missing value in the column and Unemployment column contain To handle the missing value we will switch platform and we will impute the missing data using SAS Enterprise Miner. The first step is to import the "WALMTRAIN" file into SAS Enterprise Miner from our local machine by using File Import node in our process flow.

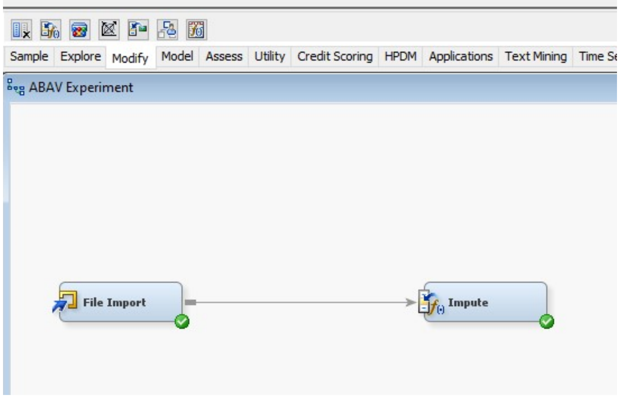
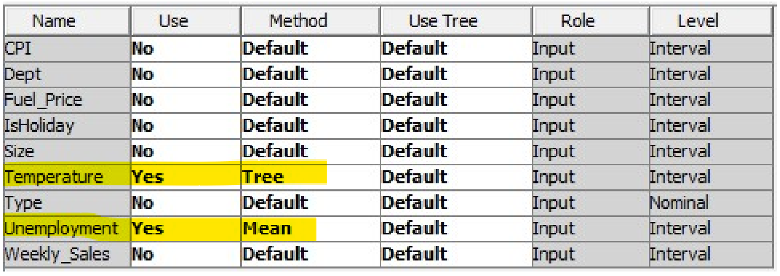


Figure 7.4: Impute Node

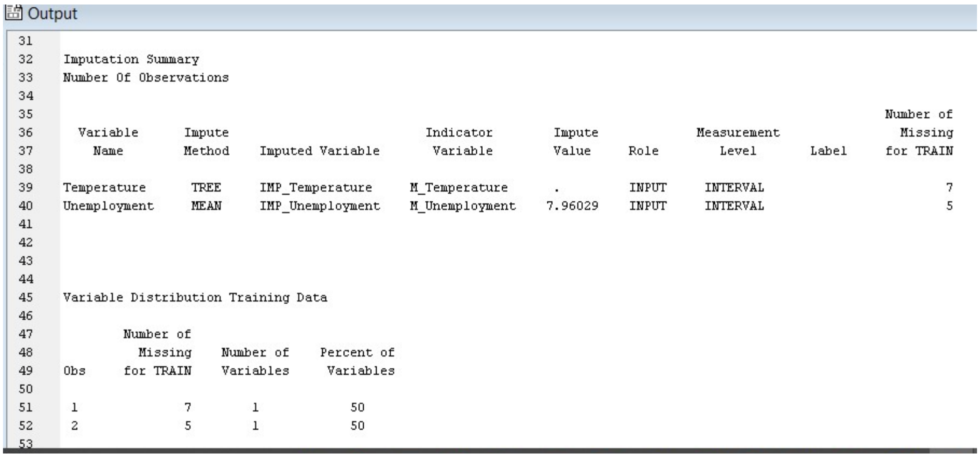
Then after place file import node to our diagram, we connect the file import node to impute node and start to set the instruction on it.



Name	Use	Method	Use Tree	Role	Level
CPI	No	Default	Default	Input	Interval
Dept	No	Default	Default	Input	Interval
Fuel_Price	No	Default	Default	Input	Interval
IsHoliday	No	Default	Default	Input	Interval
Size	No	Default	Default	Input	Interval
Temperature	Yes	Tree	Default	Input	Interval
Type	No	Default	Default	Input	Nominal
Unemployment	Yes	Mean	Default	Input	Interval
Weekly_Sales	No	Default	Default	Input	Interval

Figure 7.5 Impute Node Instruction

We will use 2 method to impute our missing data. The justification of using these 2 methods are since the missing value is not affecting too much on our data, we will use single imputation instead of multiple imputation as our missing value method. Unemployment variable will use Mean imputation where it will used the mean value in the unemployment column and substitute the missing data with mean value of overall value in unemployment column. Where for Temperature variable, we will let the software to use estimated value by predicting the value in the variable and use it to impute the missing value by using tree method. After the node imputed the missing value, it will create new column and created new variable name “IMP_Temperature” and “IMP_Unemployment” in the dataset. The result of these imputation are as below: -



```

Imputation Summary
Number Of Observations


                                                                                                             Number of
  Variable      Impute                          Indicator        Impute             Measurement                Missing
    Name        Method    Imputed Variable       Variable        Value     Role        Level        Label    for TRAIN

Temperature      TREE     IMP_Temperature     M_Temperature       .        INPUT     INTERVAL                        7
Unemployment     MEAN     IMP_Unemployment    M_Unemployment    7.96029    INPUT     INTERVAL                        5




Variable Distribution Training Data

        Number of
          Missing    Number of    Percent of
Obs     for TRAIN    Variables     Variables

 1              7        1             50
 2              5        1             50
```

Figure 7.6: Result

7.2.3 Data Replacement

Another problem we found in the dataset is there is some of the variable that is in “Type” variable are in string value. This will cause a problem when we want to do prediction in later stage. Our target is to change all the string value into numeric value so that the machine will understand better and will create more accurate result. To tackle this problem, we will use replacement node and connect it to file import node in our process flow.

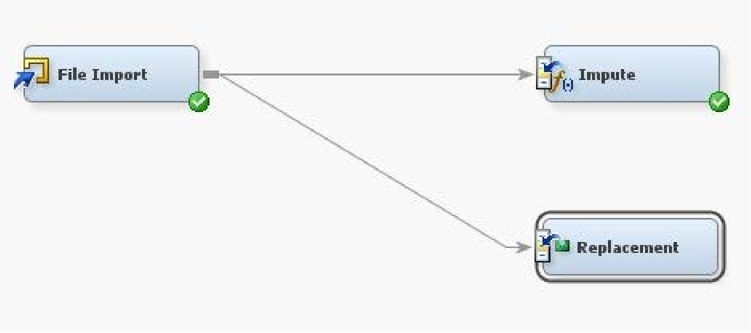
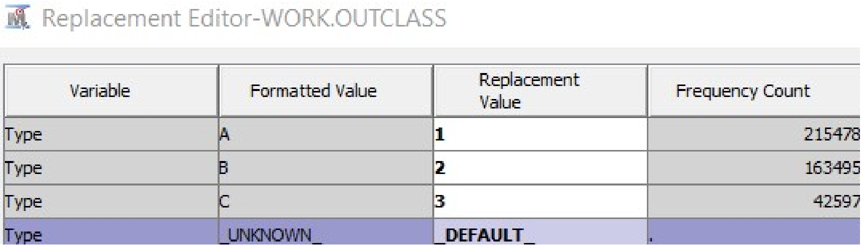


Figure 7.7: Replacement Node Diagram

After connecting the node, then we set the instruction which variable that we want to change.


Replacement Editor-WORK.OUTCLASS

Variable	Formatted Value	Replacement Value	Frequency Count
Type	A	1	215478
Type	B	2	163495
Type	C	3	42597
Type	_UNKNOWN_	_DEFAULT_	.

Figure 7.8 Replacement Editor

As the table above, we will replace value A = 1, B = 2 and C = 3 and since we have no unknown value we will leave it as default. Then we will run the node and the result as follow.

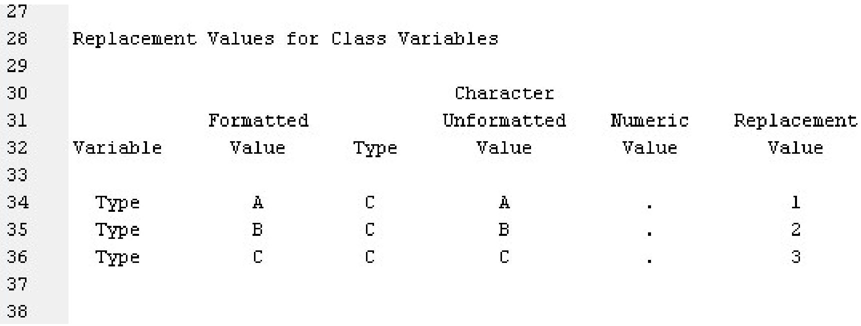

```
Replacement Values for Class Variables


                                          Character
                Formatted               Unformatted      Numeric     Replacement
Variable          Value         Type       Value          Value         Value

  Type              A             C          A              .             1
  Type              B             C          B              .             2
  Type              C             C          C              .             3


```

Figure 7.9: Result

After the datasets is merged, cleansed and transformed then we can be sure that our data is enrich and at the highest quality in order for it to be feed into our models. This also will make sure the processing will be smooth without error in our modelling part as the saying said, “garbage in garbage out”.

8.0 Modeling

8.1 Cluster Analysis

Cluster analysis is a type of unsupervised learning and an exploratory analysis that try to understand the structures within the data. Its was used to find group of similar objects and there is a lot of application of using cluster. There are 3 main cluster analysis algorithm that usually used. There is hierarchical clustering where it is treating each object as separate cluster then its identify and merge two closest cluster together until all the cluster merged together. Another algorithm is k-mean clustering where it will let the user to specify k number of cluster themselves. On the initial phase, the observation will be allocated randomly to cluster. Then the cluster mean are computed and the object started to be allocated to the nearest neighbor until the cluster did not change. And the final algorithm is latent class analysis where it was similar to k-mean clustering except it can be used on both numeric and non- numeric data.

In order to understand the effect of internal and external factor variable (internal factor: type of store, store square foot) (external: CPI, Fuel Price etc) to our dataset, we will run cluster analysis to profile it for us to better understand the behavior of our dataset and help us to make better inferences on it later.

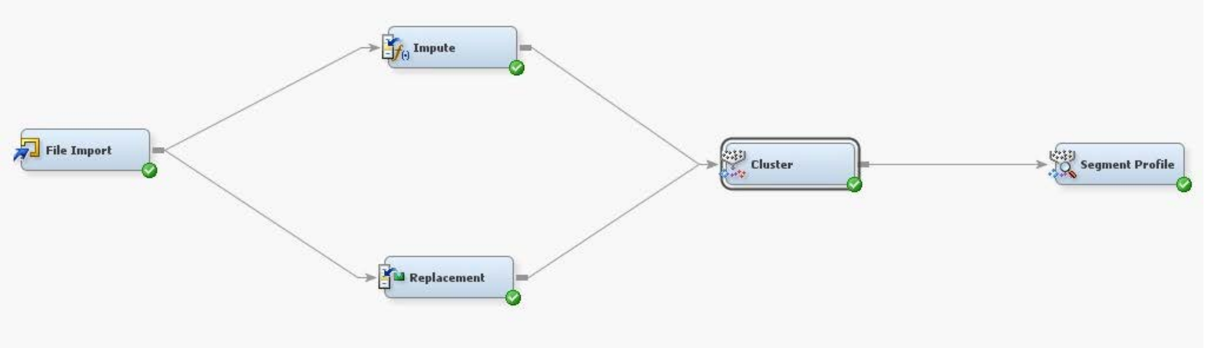


Figure 8.1 Cluster Analysis Process Flow

Above is the process flow diagram where we added cluster node and segment profile node into our process flow. After we done with impute and replacement node, cluster node will further process the data into group and the result will be profile by segment profile node.

On the clustering node, we will use store ID as the segment cluster variable role and we set the clustering method into centroid. We also cluster variable role to ‘segment’ and this resulted into 3 unique cluster as the pie chart shown below

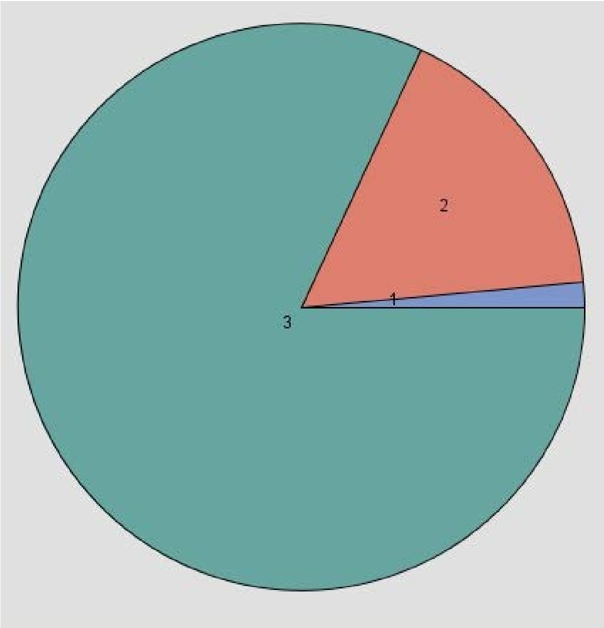
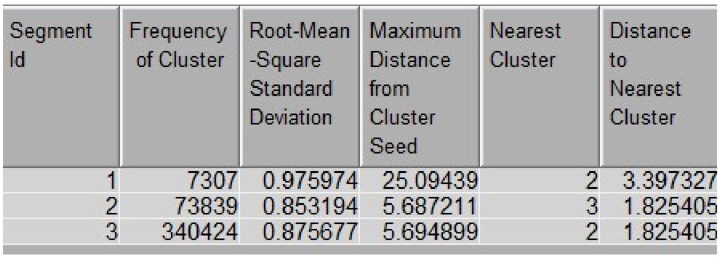


Figure 8.2: Pie Char for Cluster

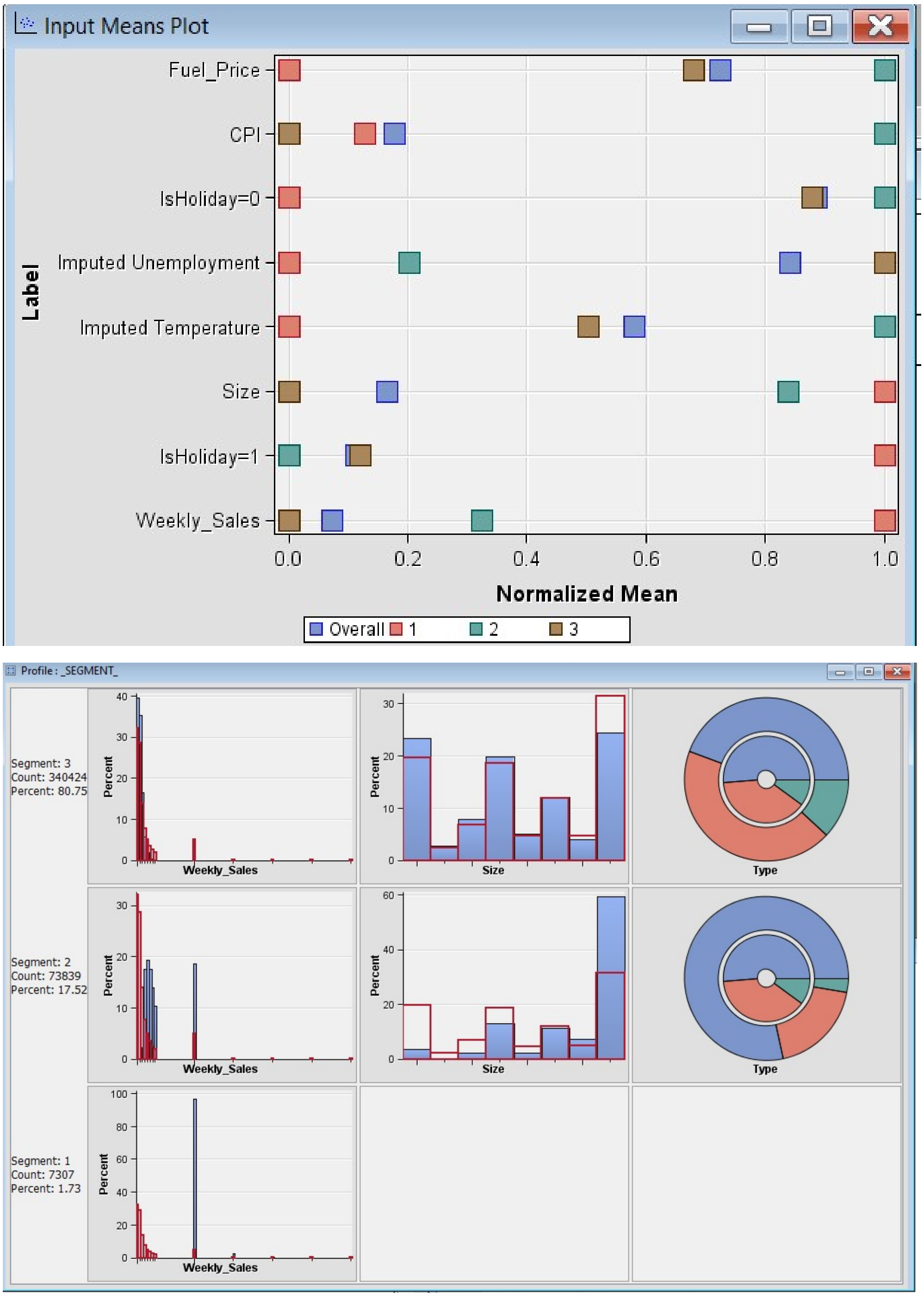
Then we analyze the behavior of the cluster by looking at the statistics. The biggest cluster is cluster number 3 with 340,424 observation fall under this cluster. Second biggest is cluster number 2 with 73,839 observation. The smallest is cluster 1 with only 7307 observation. The maximum distance from cluster seed show intra cluster distance (how far are the distance between cluster). As shown on the table below the distance between cluster is acceptable.

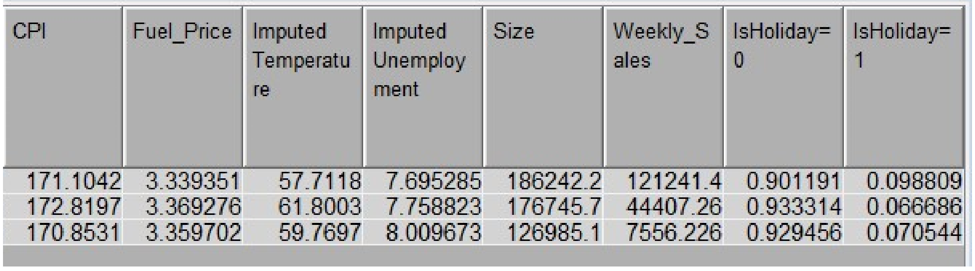


Segment Id	Frequency of Cluster	Root-Mean -Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster
1	7307	0.975974	25.09439	2	3.397327
2	73839	0.853194	5.687211	3	1.825405
3	340424	0.875677	5.694899	2	1.825405

Figure 8.3: Distance between Cluster

Result Analysis



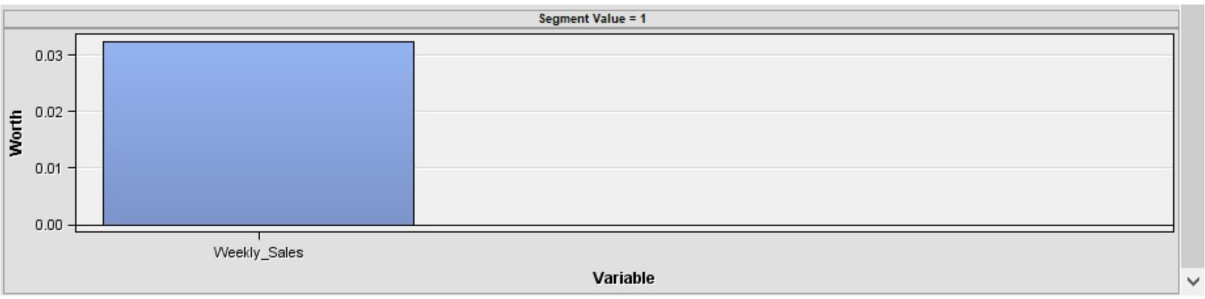


CPI	Fuel_Price	Imputed Temperature	Imputed Unemployment	Size	Weekly_Sales	IsHoliday=0	IsHoliday=1
171.1042	3.339351	57.7118	7.695285	186242.2	121241.4	0.901191	0.098809
172.8197	3.369276	61.8003	7.758823	176745.7	44407.26	0.933314	0.066686
170.8531	3.359702	59.7697	8.009673	126985.1	7556.226	0.929456	0.070544

Figure 8.4: Input Mean Plot & Mean Variable On Each Cluster Result

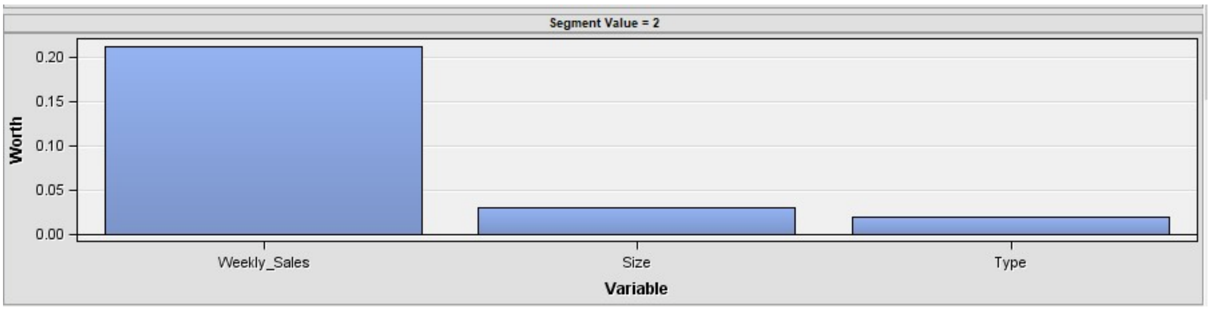
Based on Input means plot, we can make inference based on the cluster behavior

Cluster 1 :- Cluster 1 are mostly has the highest weekly sales despite only belong to small observation of 7307 which is 1.73% of total observation. This cluster has the lowest unemployment rate & temperature which is the reason why it generate so much sales as most of consumer in this area are employed and has high spending power. From this information, Walmart can target this cluster for more promotion or sales credit to customer that belong here.



For variable worth (maximum logworth) in this segment, the only worthy variable for this sagment is weekly sales that has more descriptive power than any other variable.

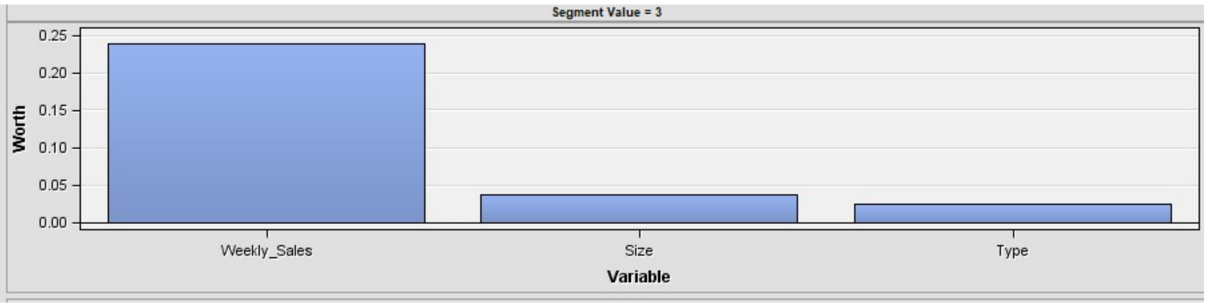
Cluster 2: - Store in cluster 2 that contain 73,879 observation (17.52%) has the highest CPI, FuelPrice and Temperature more than average compared to other cluster. This store is not effected by any holiday season. Based on the information that this cluster has high Fuel Price, meaning that consumer from this cluster are living in rural area as rural area commonly has higher fuel price due to distance from source of supply resulting higher prices on Fuel Price.



The store in segment 2 has 3 variable that worthy which is weekly sales, size and type.

Cluster 3: - Cluster 3 has is a biggest cluster 340,424 or 80.75% observation but also has the lowest weekly sales among these cluster with average as per Input Mean Plot. This can be explained by store in this cluster has the highest unemployment rate above average. We can see the correlation between CPI and Unemployment rate that all indicate consumer in this cluster has lower purchasing power thus low sales in this cluster. We can also assume that spending behavior of consumer on this cluster like to save their money and only spend it during holiday as we can see the IsHoliday variable do has some effect on this cluster.

Due to the fact it has low sales, sales strategy on this segment might not has the same impact as segment 1 and different approach to tackle this problem. Low Price should be main approach for this segment due the fact people has low purchasing power & high employment rate.

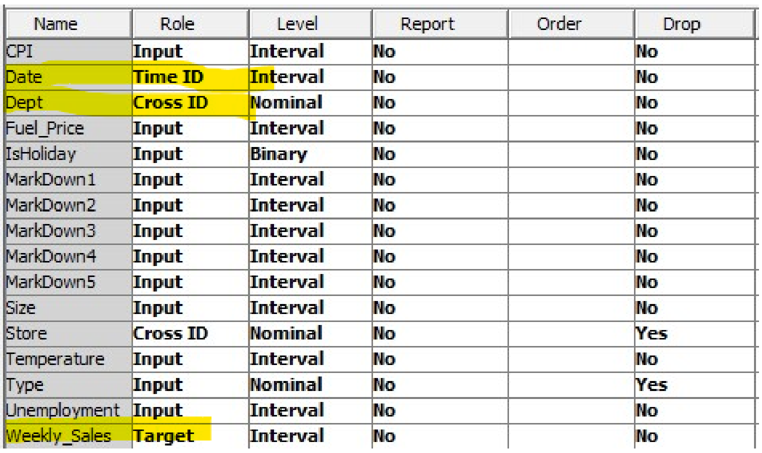


The worthiest variable for segment 3 is similar with segment 2 but the different is only with weekly sales is higher than weekly sales in segment 2. Another variable that important is size and type. This show that this these 3 variable especially weekly sales has more significant impact for this segment.

8.2 Time Series Exponential Smoothing

Setting Up Variable Role

For the second model, we will use time series exponential smoothing to predict the sales prediction from November 2012 until November 2013. The first step we taken is to reupload the file and make new process flow. We will change the role of the variable as per picture below.



Name	Role	Level	Report	Order	Drop
CPI	Input	Interval	No		No
Date	Time ID	Interval	No		No
Dept	Cross ID	Nominal	No		No
Fuel_Price	Input	Interval	No		No
IsHoliday	Input	Binary	No		No
MarkDown1	Input	Interval	No		No
MarkDown2	Input	Interval	No		No
MarkDown3	Input	Interval	No		No
MarkDown4	Input	Interval	No		No
MarkDown5	Input	Interval	No		No
Size	Input	Interval	No		No
Store	Cross ID	Nominal	No		Yes
Temperature	Input	Interval	No		No
Type	Input	Nominal	No		Yes
Unemployment	Input	Interval	No		No
Weekly_Sales	Target	Interval	No		No

Figure 8.6: Variable Role Setting

We change variable ‘Weekly_Sales’ into target role as we want to predict the sales for 1 year. Then we specify variable ‘Date’ with Time ID role as its is the requirement when running a time series node. And we dropped variable Type as it is in nominal value. Then we make sure all the level are set correctly especially for variable Dept & Store where by default SAS automatically set it as interval but its actually is nominal as it represent the count of store in numeric format.

Another crucial step into setting up variable role is the cross id. Since we want to predict the each 99 department sales for Walmart, the cross id will be set up on variable ‘Dept’. Other interval variable will be set as Input variable.

The next step after setting up variable role is to use the time series node. But since we are uploading the dataset back, we need to use impute nodes back to handle the missing value. After that we’ll use the TS Data Preparation node to explore our data in depth then finally make the sales prediction for each department using TS Exponential Smoothing node. Below are the process diagram.

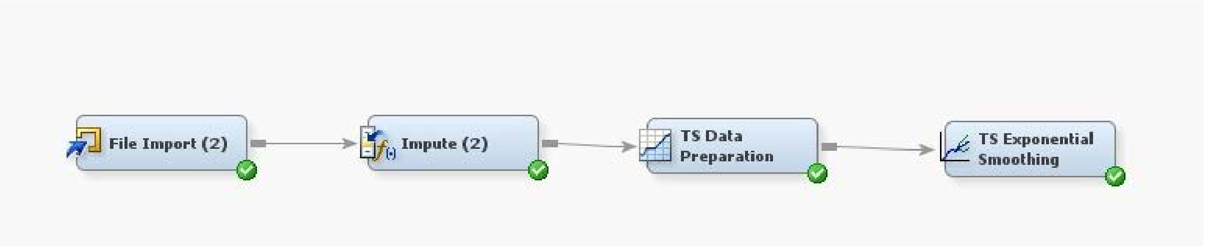


Figure 8.7: Time Series Process Flow Diagram

TS Data Preparation

Before we run TS Data Preparation node, we need to make sure all data are interval (except for Dept variable which have cross id role) in order for it to run without error. Another important parameter to set before running is to manually specify time interval for this data as the SAS unable to figure the time interval by itself. We set the interval as in week due to the dataset time is in weekly format. There are several finding that we find interesting when we run this node and we will discuss it below.

Finding & Interpretation

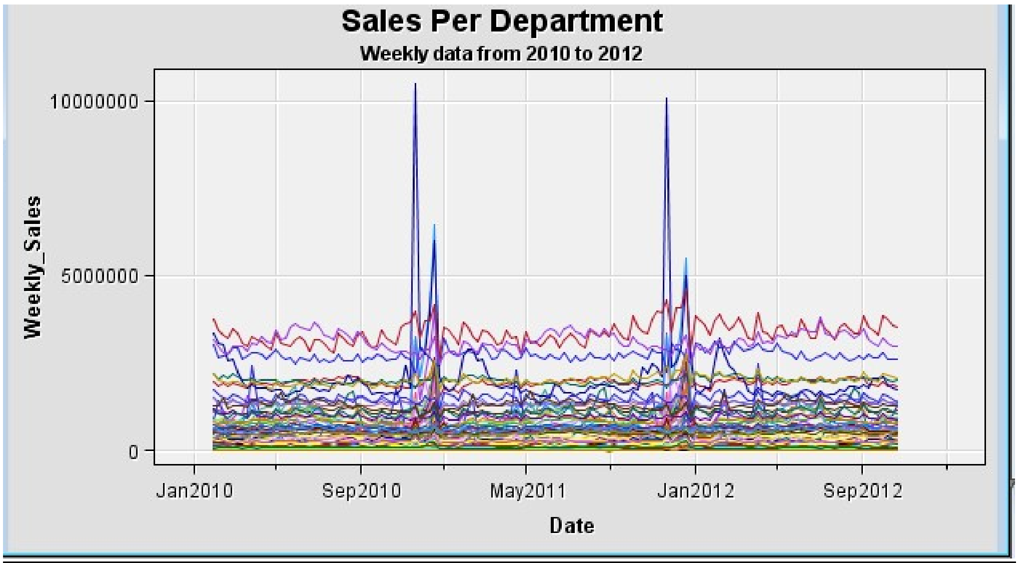


Figure 8.8 Sales Per Department Trend

On above line graph, each color line are represent the 99 department of Walmart. With this line graph we can see which department has the highest contribution on Walmart weekly sales and which department sales is affected by holiday season clearly represented on this graph. We will zoom in to the specific department to make inference on this finding.

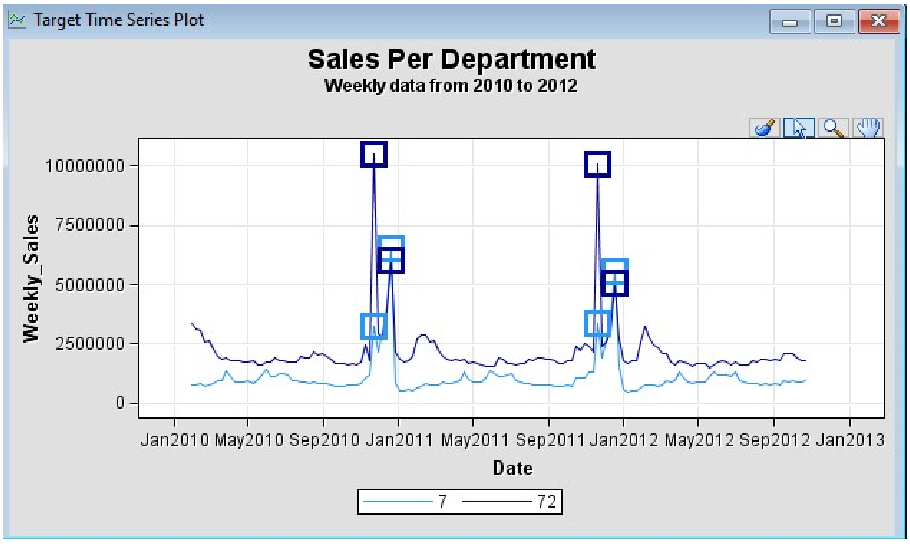

Target Time Series Plot
Sales Per Department
Weekly data from 2010 to 2012
Weekly_Sales
10000000
7500000
5000000
2500000
0
Jan2010 May2010 Sep2010 Jan2011 May2011 Sep2011 Jan2012 May2012 Sep2012 Jan2013
Date
7
72

Figure 8.9 Sales Trend for Dept 7 & 72

Based on above line graph, we focused on 2 department namely department 7 and department 72 which has similar pattern. Both shown sudden increase on Nov and Dec on each year (Thankgiving & Christmas Holiday) so we can conclude that both department are having holiday related product such toys or electronic as this two items are commonly purchase during holiday by consumer as a gifts. We can make inference that both department should be in focus for any Walmart strategy during holiday period.

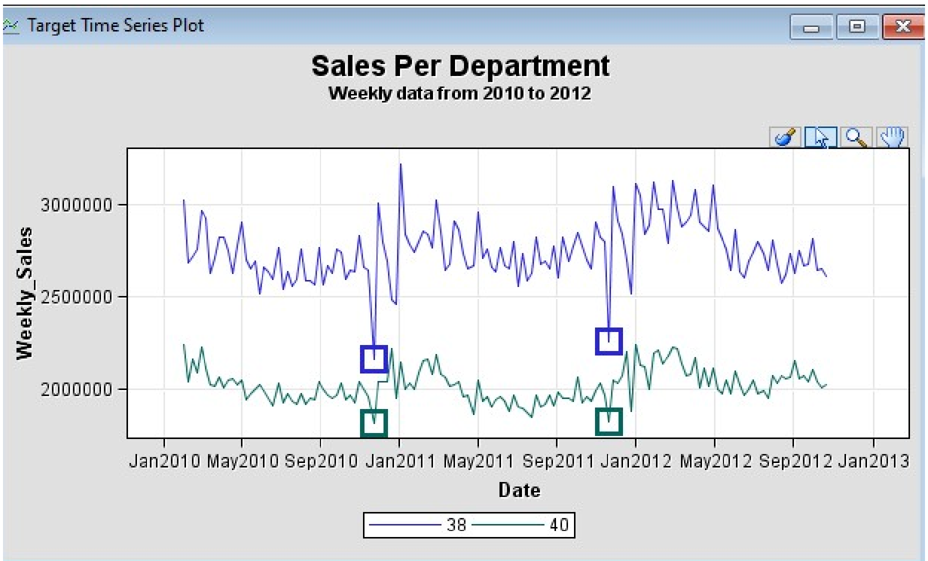

Target Time Series Plot
Sales Per Department
Weekly data from 2010 to 2012
Weekly_Sales
3000000
2500000
2000000
Jan2010 May2010 Sep2010 Jan2011 May2011 Sep2011 Jan2012 May2012 Sep2012 Jan2013
Date
38
40

Figure 8.10: Sales Trend for Department 38 & 40

On the contrary, we also found out that department has inverse reaction during holiday period. On Thanksgiving holiday period, both department 38 & 40 dropped sales sharply which we can conclude make this department is not popular during this particular holiday period. This is useful information to Walmart management to look at as they can reallocate their personnel in this department into more demanding department for example.

Sales Forecasting

For Sales Forecasting, we will provide Walmart with full 2013 forecasting for them to strategize their sales. We will used SAS Enterprise Miner TS Exponential Smoothing node to make this prediction.

Forecasting Model

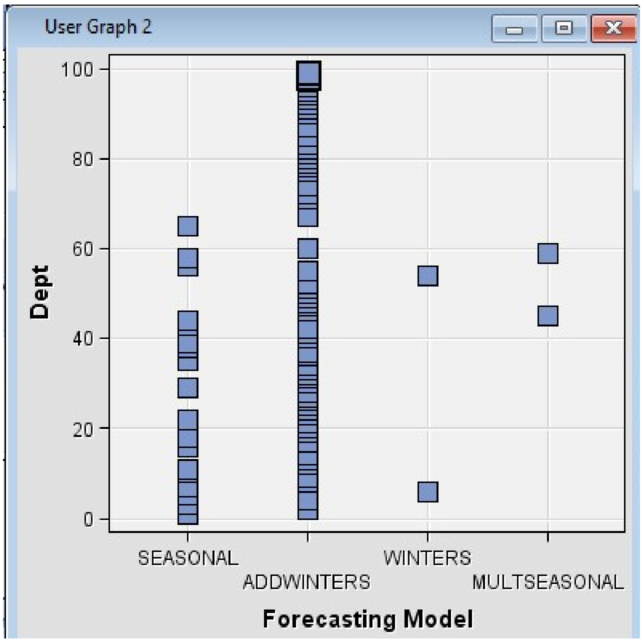


Figure 8.11: Commonly used Forecasting Model

For each department, different model was use to forecast the model. Since in the parameter we has specify to use the best model among smoothing forecasting modelto used, the model with least error will be selected. On the graph above, Winter additive method (ADDWINTERS) & additive seasonal(SEASONAL) model proved to be the best fit model among the department.

Performance Benchmark

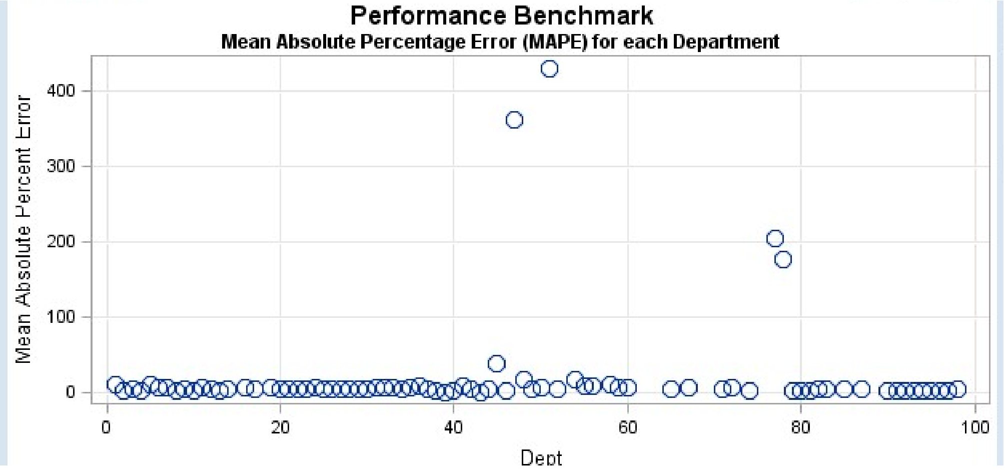


Figure 8.12: MAPE for each Department

First step is to make sure all the parameter & instruction was set correctly before we run the nodes. After we run the node, we access the performance of our forecasting by using fit statistic model. We will use Mean Absolute Percentage Error to evaluate our forecasting performance as it’s the most commonly used performance measurement when using time series forecast.

From the fit statistic model, we can see that majority of our Dept is very close to zero which is good indicator of performance. Only 4 department has high MAPE error dept 47, 51, 77 & 78 but we can exclude this 4 department from our inference.

Inferences on Result

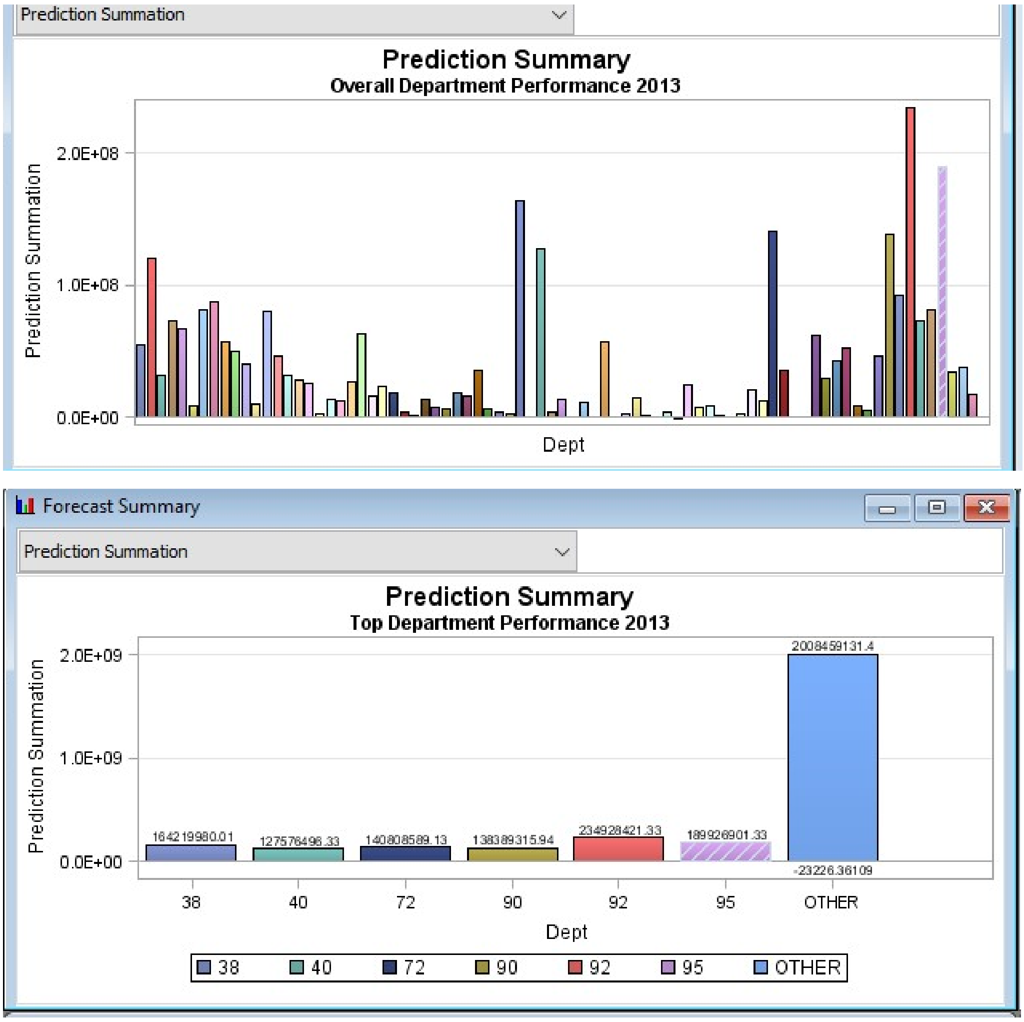


Figure 8.13: Prediction Summery (Overall & Top 6 Dept)

From the prediction summary, we can see that there is 6 department was predicted to perform better in sales compared from others. Top of the department was department 92 followed by department 95 and department 38.

To explain why department 92 & 95 achieve top sales, we will look back at graph 8.8 (Sales Department Trend Line Graph), both department has high steady sales throughout the year. This show that consumer behavior toward this department does not change and demand for product on this department is always high no matter the season. Essential goods item (i.e, Grocery, household item) are example of consumer product that has high demand all year long. Essential goods demand does not effected by any sort of promotion as it has low elasticity of demand

Department 38, 40 & 72 showing sales demand affected by holiday as shown on the line graph 8.8. Huge spike during holiday are showing that consumer demand is the factor that drive the sales of this department and its is non-essential goods. These department should be the focus by Walmart management to boost their sales as non-essential good has high elasticity of demand.

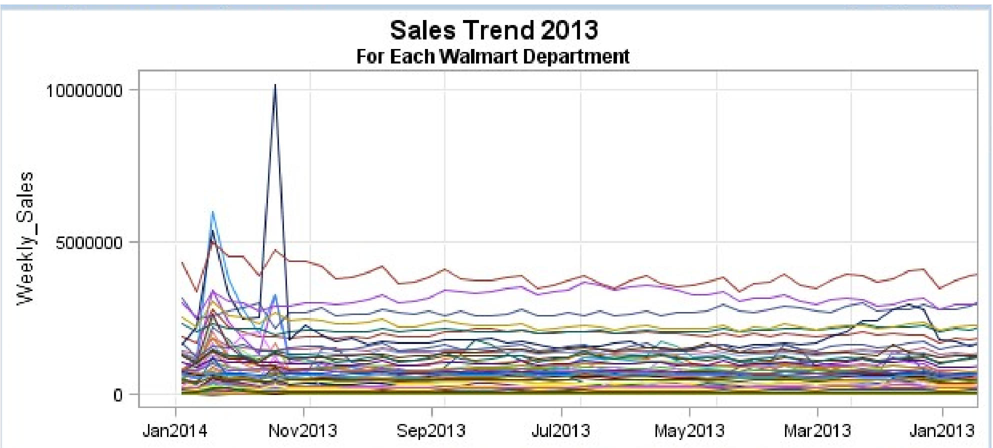


Figure 8.14: Predicted Sales Trend 2013 for Each Department

For the sales trend, this graph is very useful for Walmart management to analyze the effect of holiday. We can see from these forecasting model, spike happen during thankgiving & Christmas holiday period. There are also some department that has spike during February (Superbowl Period) and September (Labour Day).

February (Super Bowl Season)

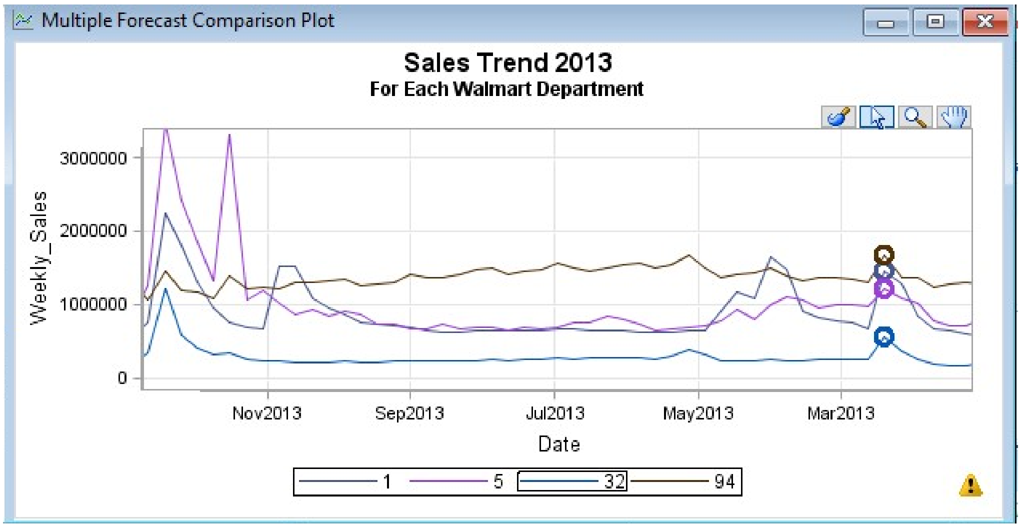


Figure 8.15 Super Bowl Trend

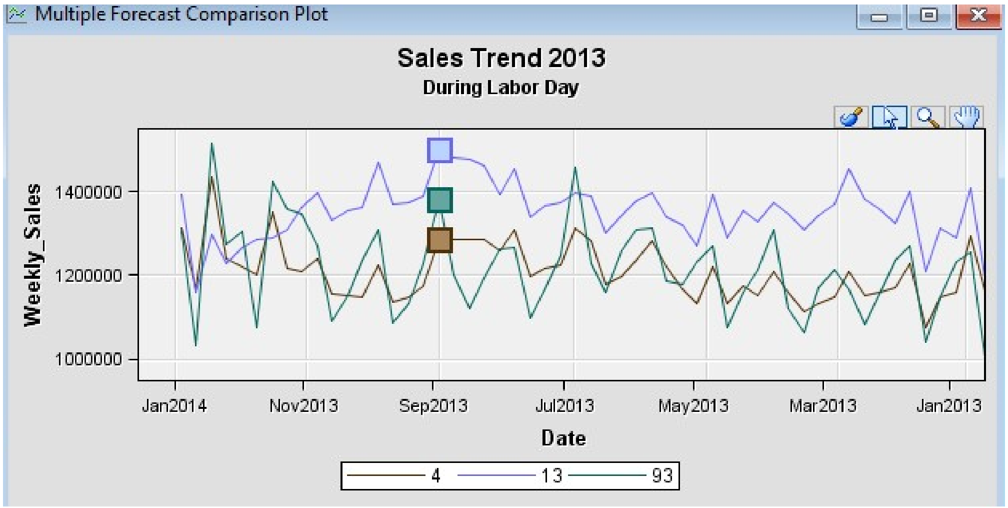
8.15 Super Bowl

Super Bowl Sunday is often considered as unofficial holiday in America. Its a day where National Footbal League (NFL) annual championship game. For Superbowl season, most of consumer predominantly male, will make plan to watch the games with their friend. Sport social gathering is key feature on this holiday.

Consumer behavior during this period are forecasted that there are 4 department that will spike during 10 February 2013 which is department 1, 5, 32 and 94. Promotion campaign related to popular product during this period should be focusing on male related product such as tobacco & beer, sport product (i.e. sport apparel, caps & football gear) and food product (Chicken wings, beef rib, snack food) which we believe is related to these 4 department.

Strategy during this period is to focus more gander based male-related item due to the fact that majority of Super Bowl fan are male. Sport theme should be used in this promotion for decoration or advertisement such as hiring NFL player to be ambassador to Walmart promotion.

September (Labor Day)



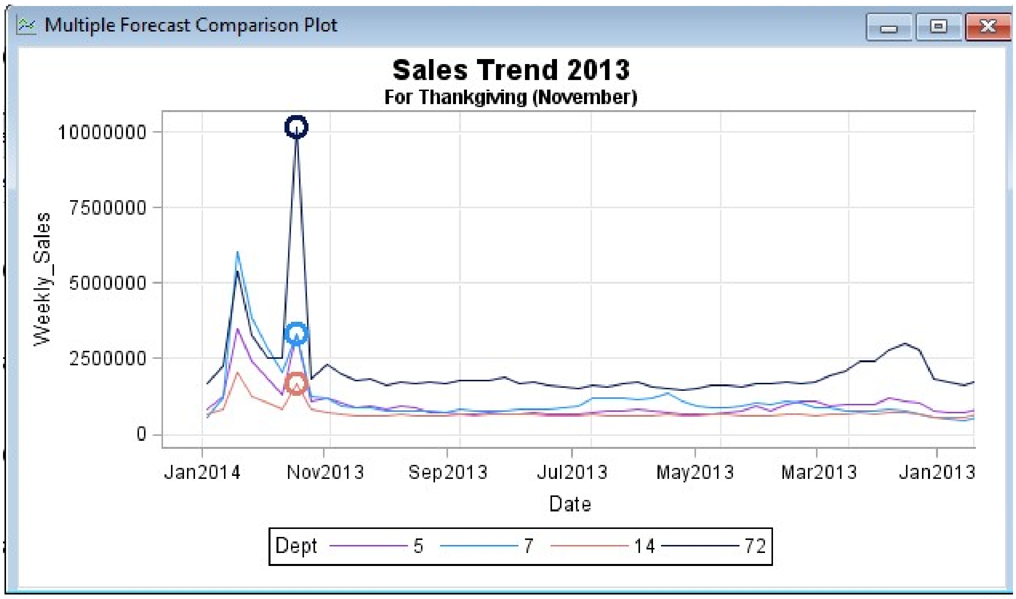
8.16: Labor Day Sales Trend

To understand consumer behavior during Labor day we need to understand the characteristic of this holiday. This holiday is often run a back to school sale where parent and student purchase their school supplies and apparel for upcoming school year.

On the Line graph 8.16, we can see there is 3 department that was predicted to has some impact in this holiday period. The item in this department can be assume as academic book, stationery and school apparel (School related item)judge by the boost of sales in this department.

The sales strategy in this period is School theme. There are 2 type of targeted customer during this season. One is teenager due to the fact that they have the access to money their parent giving them and can decide what to buy for themselves and parent who decide what to buy for their primary school children. For teenager, they more attracted to more fashionable item so strategy is to used local artist to endorse the item. By convincing the customer that Walmart is offering item at lowest price compare to other competitor as most parent want to save their hard earn money by buying only basic stuff to their children.

November (Thanksgiving)

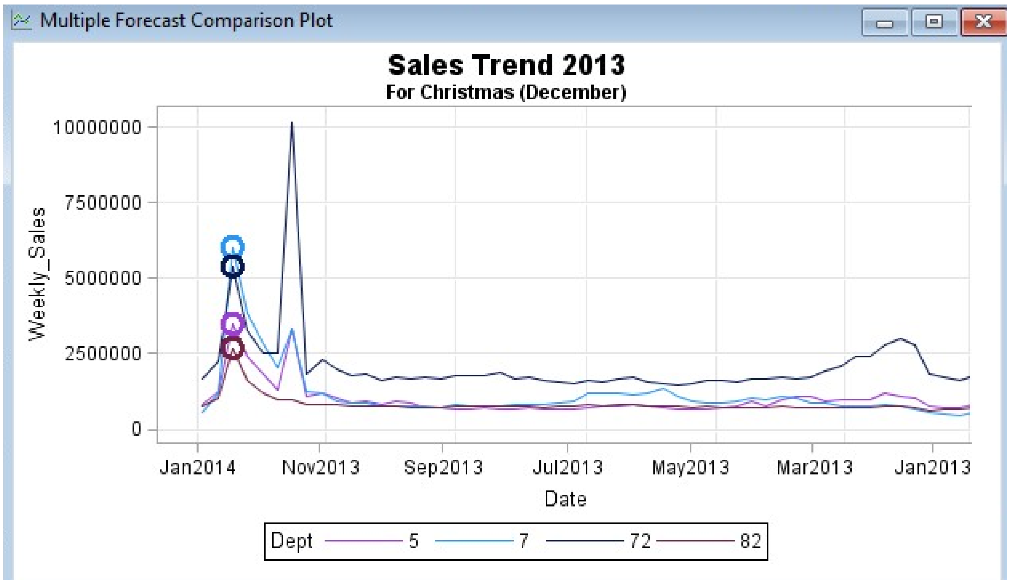


8.17: Thanksgiving Sales Trend

On thanksgiving, most consumer are expected to have gathered with their family or friends. The increase sales predicted to be started on 10 November and will end until end of December. Even though Thanksgiving Day was on 28 November, the forecast was expecting consumer to started to purchase their holiday gift early on preparation should be started on end of October itself. The peak time was in 17 November with department 72 spike sharply on this day follow by another department such as 5, 7 and 14.

Contact with customer via email, brochure, to promote sales on these particular department should start. Stock inventory for department 72 should be increase due to the fact the spike in sales will continue until December.

December (Christmas)



8.18: Christmas Sales Trend

For Christmas period, sales boost was predicted to start during Thanksgiving day on 28 November and peak at 15 December, week before Christmas. Highest department sales coming from department 7 followed by department 72, 5 and 82.

Consumer on Christmas period would more interested on holiday party and gifts ideas and it should be the theme to be focused on. Most popular product are toys and electronic which is popular gifts that parent give to their kids. Spike on this department that has this item are expected so promotion should be focus on this popular item.

Walmart also need to be aware that other retailer also competing for consumer in this lucrative holiday season as consumer will be hopping from shop to shop to find bargains. Its important to capture a deal consumer on their visit to Walmart as fast as they can and this are important thing that should be taken into consideration.

9.0 Recommendation

Based on the sales prediction, we can see exactly which department should be focused by Walmart decision maker for them to strategize and anticipate problem early on. From our observation on Sales Prediction made by our model there are few recommendation that we can propose to Walmart.

- **Perfect time to start year-end holiday promotion**:- from the data we can see the predicted increase sales boost start at 10 November 2013 and stop at 29 December 2013. There was 2 peak time during this period which is on 17 November and 15 December 2013. It would be recommended that the sales period start at 10 November until 29 December.
- **All resources focus to high sales department**:- There is huge demand on department 72 during year end period (Nov~Dec). Increase on stock inventory, sales personnel whether hiring temporary staff during this period or moving personnel from non-popular department to this department is highly recommended.
- **Product Bundling**:- Another recommendation is to bundle together product in high sales department with low sales department. For example product in department 72 can be bundle together with product from lower sales department 40 to push sales into customer during Thanksgiving period.
- **Markdown Strategy:-** There is some month where sales was not dipped below average. To increase sales, its is recommend to use Markdown strategy to boost the sales during this month such as stock clearance sales. On 20 January & 21 April 2013 sales was predicted to decline. This is where Markdown strategy can be use to increase the sales.

10.0 Discussion & Conclusion

First, we want to clarify that there is slight changes on Aim and Objective where before this we try to forecast sales of 45 Walmart Store and forecast which store is most profitable due to some mistake. Instead of forecasting the store, we are forecasting each 99 department on Walmart store. Below are the revised objective:-

- To find out which department the most profitable between all
- To analyze the correlation between variable
- To develop a model that able to forecast the sales of each 99 Walmart Department
- To come out with business-wise recommendation such as appropriate time to launch promotion campaign to maximize profit based on the prediction model

For the first objective which to find which department is most profitable, by using Time series forecasting, we able to pinpoint 6 department that will be most profitable for Walmart.

Second objective is to analyze the correlation between variable, we have use cluster analysis to group together store and find some interesting pattern among variable that useful when used in conjunction with sales prediction and assist them during decision making.

Third objective is completed by using Time Series Forecasting in SAS Enterprise Miner to make sales prediction on 2013 for Walmart. TS Exponential Smoothing node was used to make sales prediction and achieve high accuracy for most of the department.

Fourth objective is business wise recommendation is done by giving 4 recommendation that help Walmart in promotion aspect, resources allocation, marketing strategy and product bundling.

Therefore, we would like to conclude that the analysis on the Walmart Sales Prediction datasets is successful but in future, possible use of datasets that contain more granular information such as product items instead of department only variable will prove to be more useful and easier to make accurate inference.