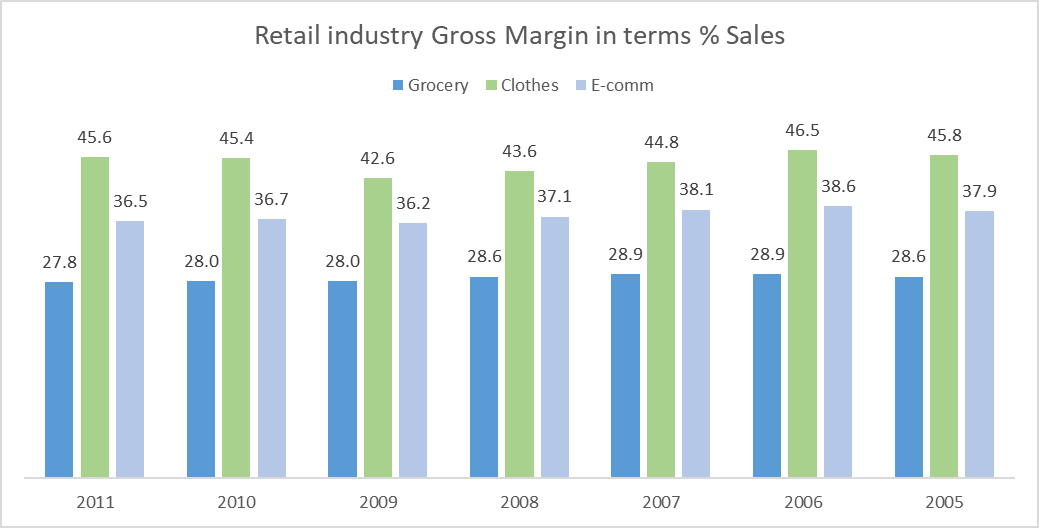
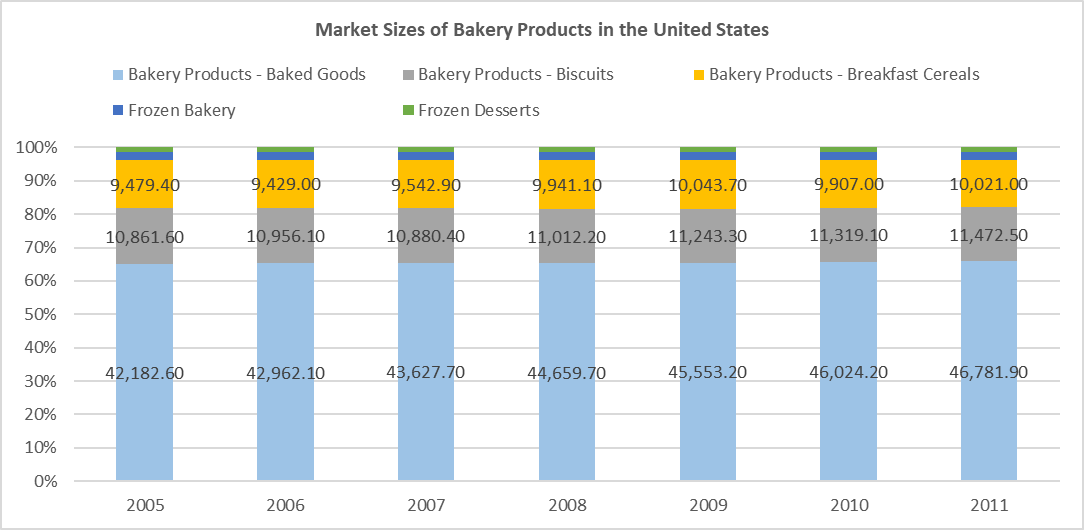
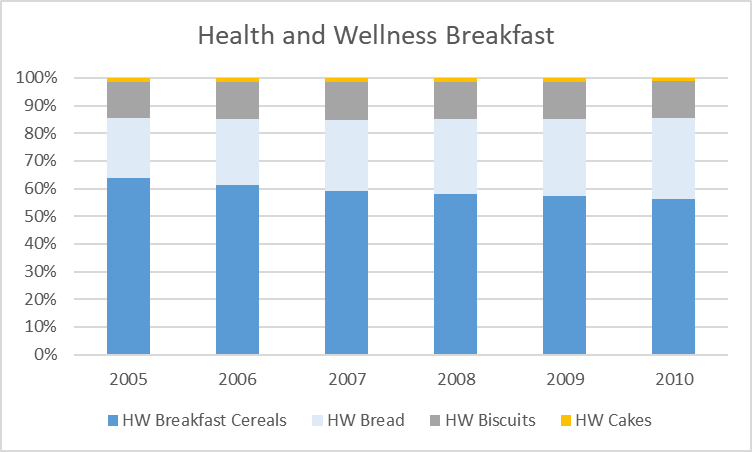
Predicting Sales and Improving Profitability of Retail Chain

# Introduction

This project deals with analysis of retail transaction data of a retail chain located at five states of the United States of America, that includes the sales, promotion information for multiple products and brands with different categories for the past 156 weeks.

**US Retail industry and its growth during 2005-2011**

 The retail industry in united states is highly competitive esp in food and beverage (grocery) stores command comparatively low gross margin with reference to clothes and e-commerce industry, with reference to Annual Retail Trade Survey1(Fig 1). To survive in such competitive environment, the brick and mortar retail must be innovative and competitive to stay in the market. Hence the management of this retail chain wish to classify the stores, value of each products they are delivering and thus enhance the margin to the retailer.

The dataset generated out of transaction/sales (spends by the customer) done over the year that highly depends on the customer preference of retail stores over the competition. The entire dataset filtered to products that used for prior and during breakfast. The figure 2, market size of bakery product in the United States of the America shows that bakery products (65%)2 dominate the consumption during breakfast. Also, it reinforces that consumption pattern of breakfast is not changed for the past during 2005-2011.But the drive of health and wellness among the consumers added another segment i.e. Health and wellness breakfast made an internal change in the subcategory. The health and wellness breakfast comprises cereals, bread, biscuits and cakes, the study shows that cereals losing their placing with breads. The breakfast shares for cereals decreased from 62 % to 57%, evident from fig 3.

In the prevailing condition, the retailer wish to understand the consumption/spending pattern of their consumers in five different states with past three-year data and preferably store the consumers choice of the product in the shelves.

The objective of this study to predict the sales for the future for different products, test hypothesis to address the relation between price and discount, price thresholds and price elasticity of the products. The study should estimate the impact on sales by changing the price gaps between items, promotions, displays and feature. Also, address price cushion on increase in margin to retailer that would increase the sales. At the end, considerable effort should be spent to identify the products that would enhance the sales and margin to the retailer.

# Dataset Description

The dataset comprises transaction data at different stores located at five different states of the united states of America. It contains the UPC (unique product code wise), price, baseline and customer spend pattern on weekly basis for past 156 weeks.

The dataset segregated in terms of product wise transaction with reference to UPC (unique product code in date time line, details of the store in which the transaction made, details of the product with details on packaging, brand and manufacturer. There were few categorical variables such as product, category, sub category and Store id. There were many continuous variables such as Spend (Sale), sales areas and price.

## Factors



# Methodology

## Exploratory Descriptive Analysis

* Inferences
* Outliers and treatment
* Category wise
* Area of interest

## Inferences from Dataset

A screenshot of a social media post

Description generated with very high confidenceThe sales of the retail chain were as shown in the figure. The retail chain commands average sale of 9600K /year There is no predominant seasonality as seen in the chart. But this need to be ascertained with suitable hypothesis.

In line with US population choice, cereals are most preferred food for breakfast in these five states too. This is evident from the product preference chart and sub category.

A screenshot of a social media post

Description generated with very high confidenceA screenshot of a social media post

Description generated with very high confidenceA screenshot of a social media post

Description generated with very high confidenceThe brand penetration is highly evident from the chart shown below, the leaving the private lables and general misc brands, the sales of the retail chain is more than 80% contributed by major brands such as Kellogs,Quarker and Tombstone. This shows that population in the five states were highly concerned about the brands they consume. Apart A picture containing screenshot

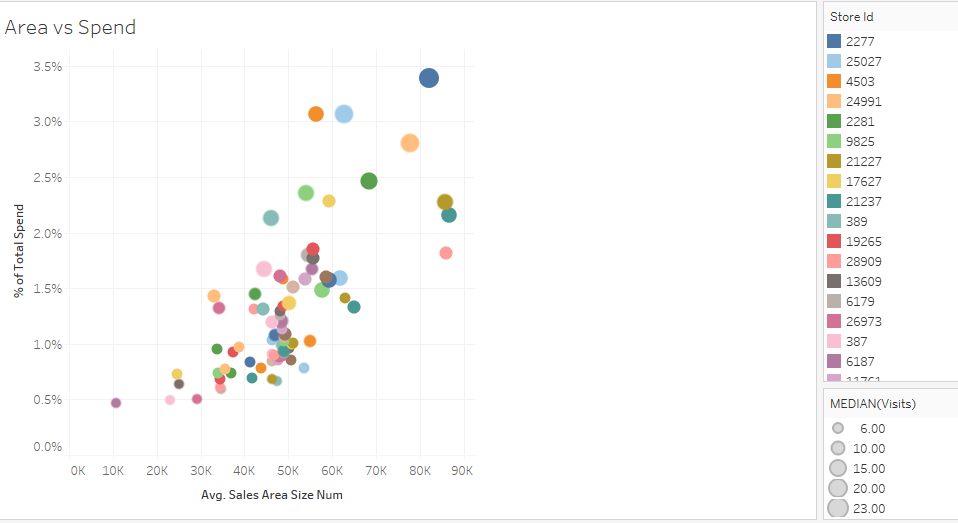
Description generated with very high confidencefrom major brands, while chosing the cold cereals, snacks and frozen pizza, consumers doesnot seem to worry about the brand and select their choice in the private labels. This augments by the picture shown below.

A close up of a map

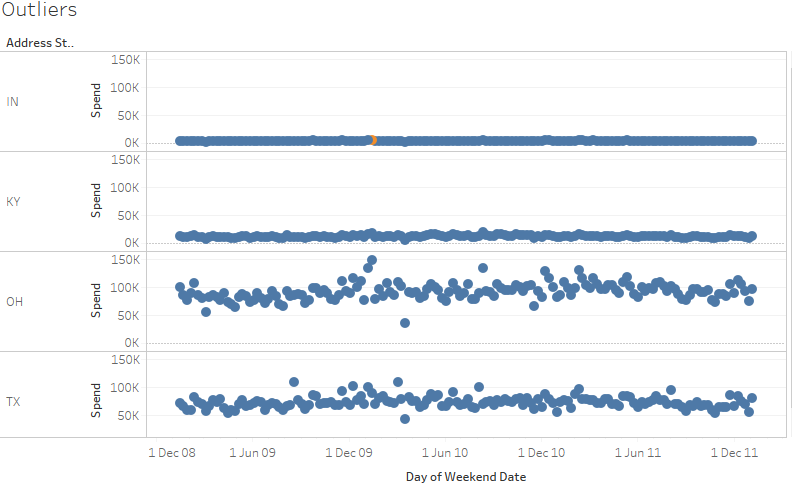
Description generated with high confidenceA screenshot of a cell phone

Description generated with very high confidenceBased on retailer choice, the top 10 retailers command better share in the sales of the retail chain. The sale of the retail chain primarily dependant on visits made by each customer . if the retailer has median visits more than 9 times , then the sales contributed would be 50% more than median sales. The customer visit also inherent function of sales area of the shop and average weekly basket size. There exist a linear relation between the number of visits and sales generated by the customer and this needs to be studied with suitable hypothesis. Those are evident from the above figures.

Another inference, if the store size or sale area greater than 70k Sq ft, it attracts the customers in to spend more and sales in those stores in the magnitude of 2X to 3X greater than average sale. There possible reasons could be lot of space for promotional mechanism such as instore display . This needs to be studied in detail with suitable hypothesis.

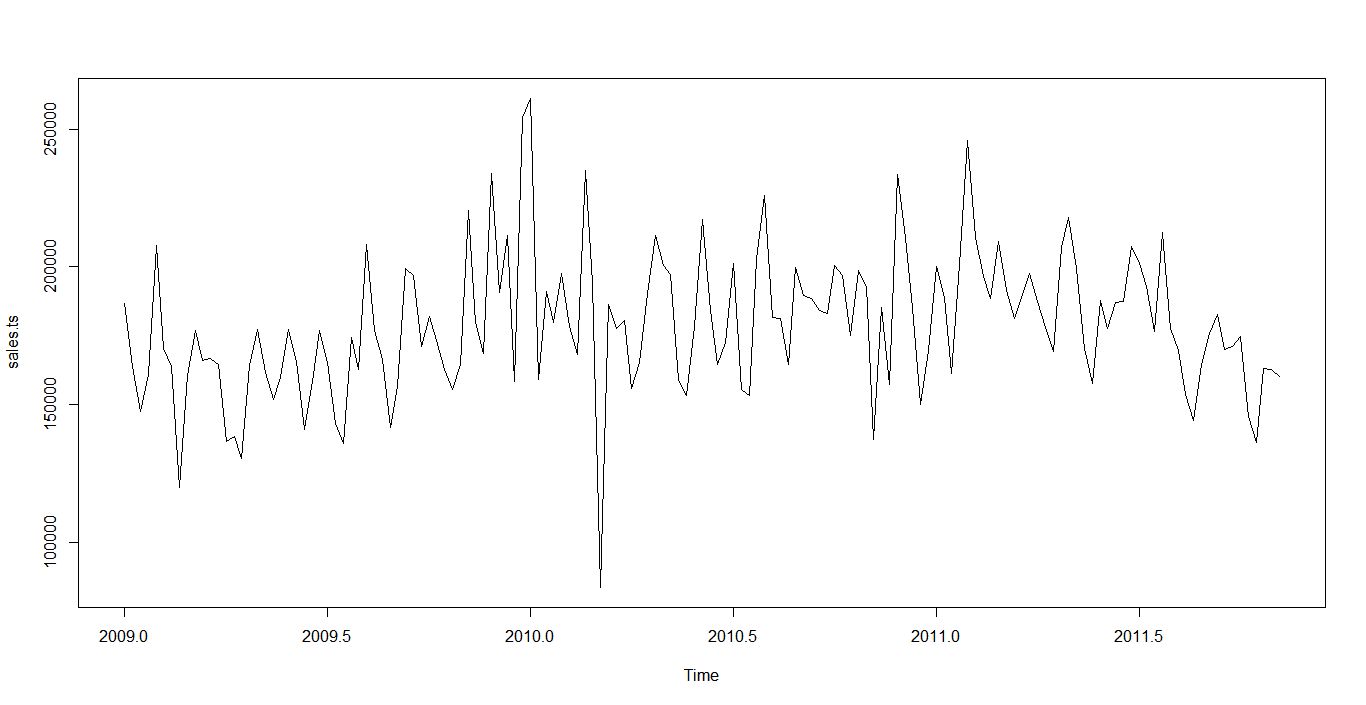


Out of 16 Manufactures, top 4 contributes to 65% of the overall sales and out of 7 subcategories, top 3 contributes to 68% of the overall sales.



We have found only 1 outlier (above 95% quantile or below 5% quantile).

# Sales – Time Series Analysis

The transaction data for the sales happened during the period 14012009 to 14012012 has been collected. The transaction, the spend (sales) is aggregated sum on a weekly basis .the weekwise sales data for the retail chain collated and shown as in the figure attached.

Check for stationarity:

A stationary time series is one whose statistical properties such as mean, variance, autocorrelation, etc. are all constant over time. The ADF (Augmented Dickey-Fuller Test) study on the sales data shows that the alternate hypothesis needs to accepted (p<0.05 – Significant at 5%), hence it support the conclusion that the data is stationary.

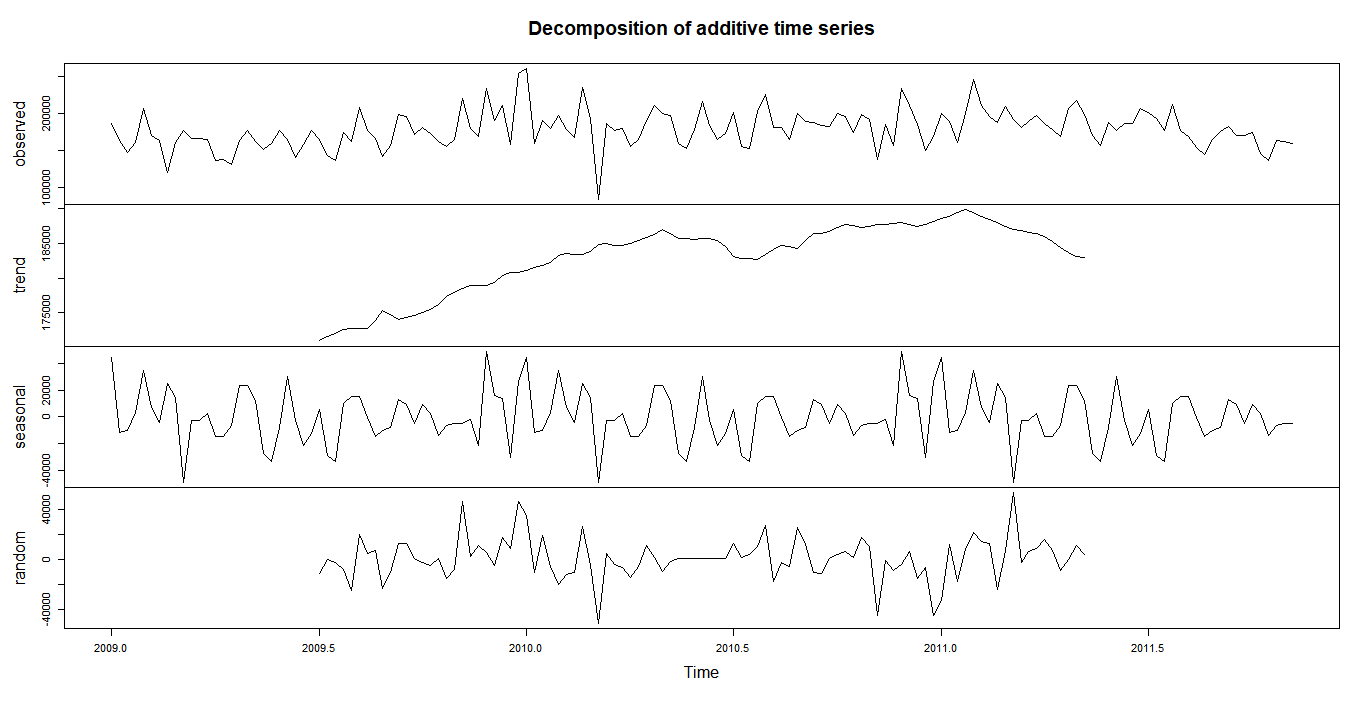
adf.test(sales.ts/scale(sales.ts-sales.comp$seasonal), alternative = "stationary")

Augmented Dickey-Fuller Test

data: sales.ts/scale(sales.ts - sales.comp$seasonal)

Dickey-Fuller = -4.7267, Lag order = 5, p-value = 0.01

alternative hypothesis: stationary



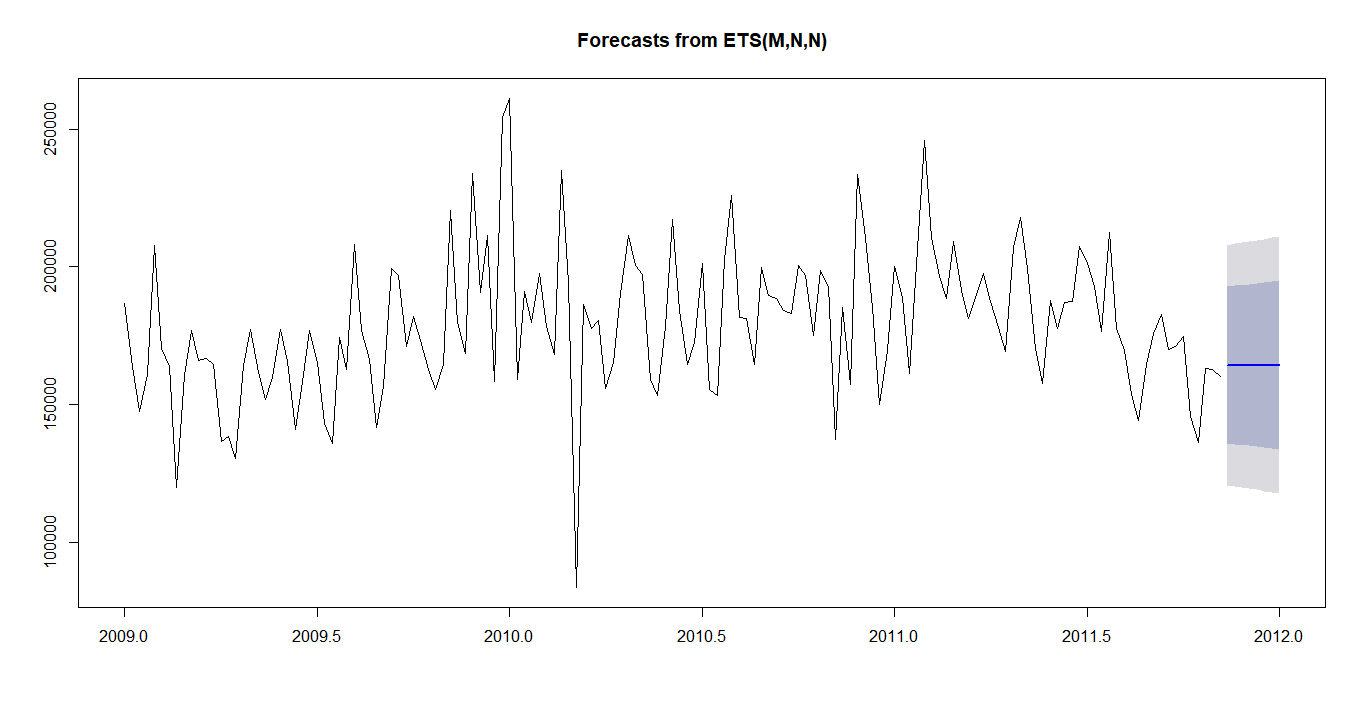
## Forecasting the sales

Forecasting the retail chain sales, is a crucial part in getting the things for the successful operations management. There are several forecasting model available and best models need to be selected based on minimal MAPE. To name few critical

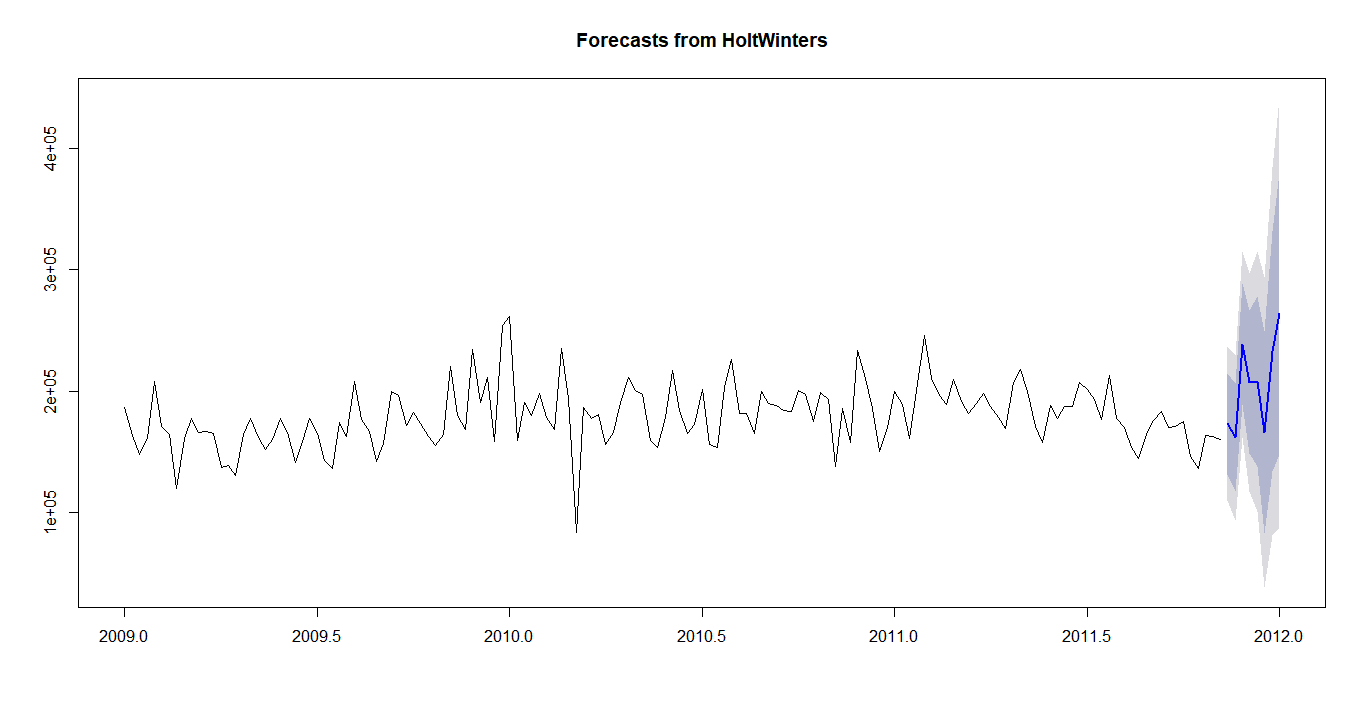
1. Exponential Time Smoothing
   1. Holt Winters
2. ARIMA model
3. TBATS
4. NNETAR

Two subsets were formed out of the time series data of sales, First 145 weeks were considered for the training set and last 8 weeks were considered to be test set. Based on the MAPE of each model and stability of the model in predicting the next weeks, the best model would be selected.

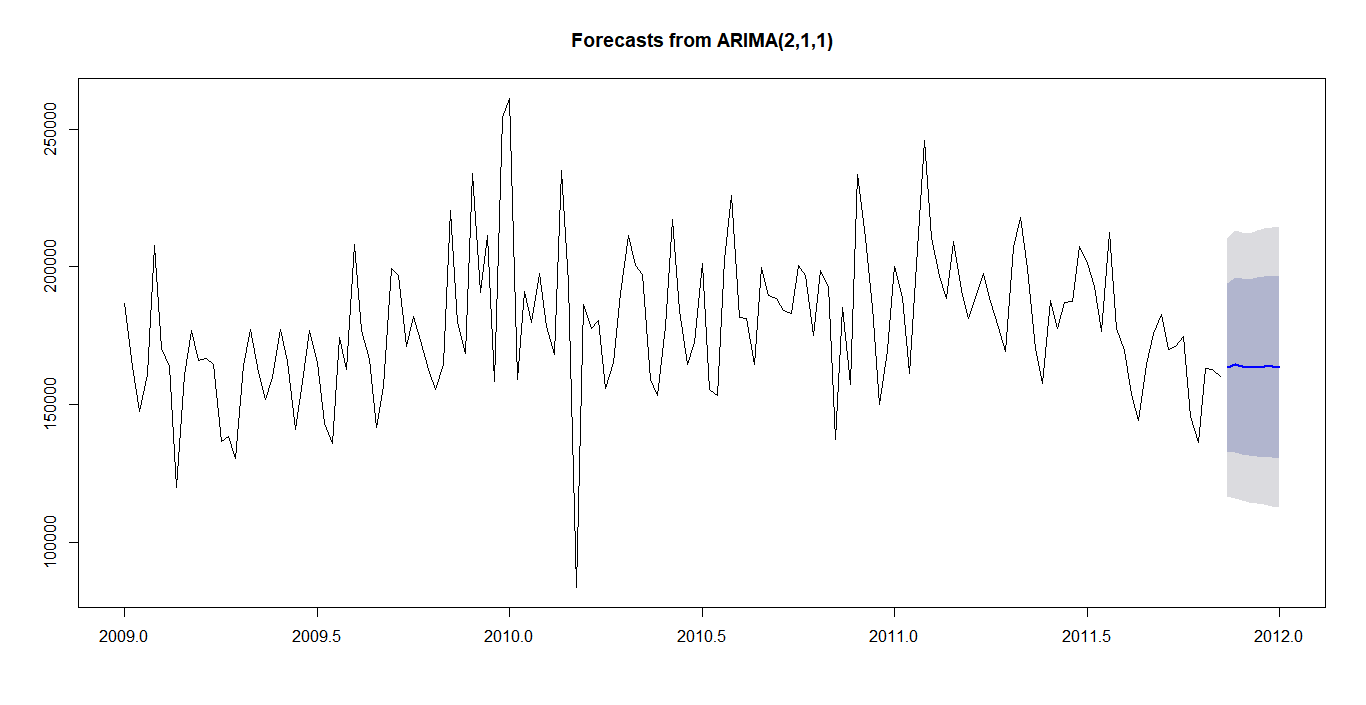
### Exponential Time Smoothing



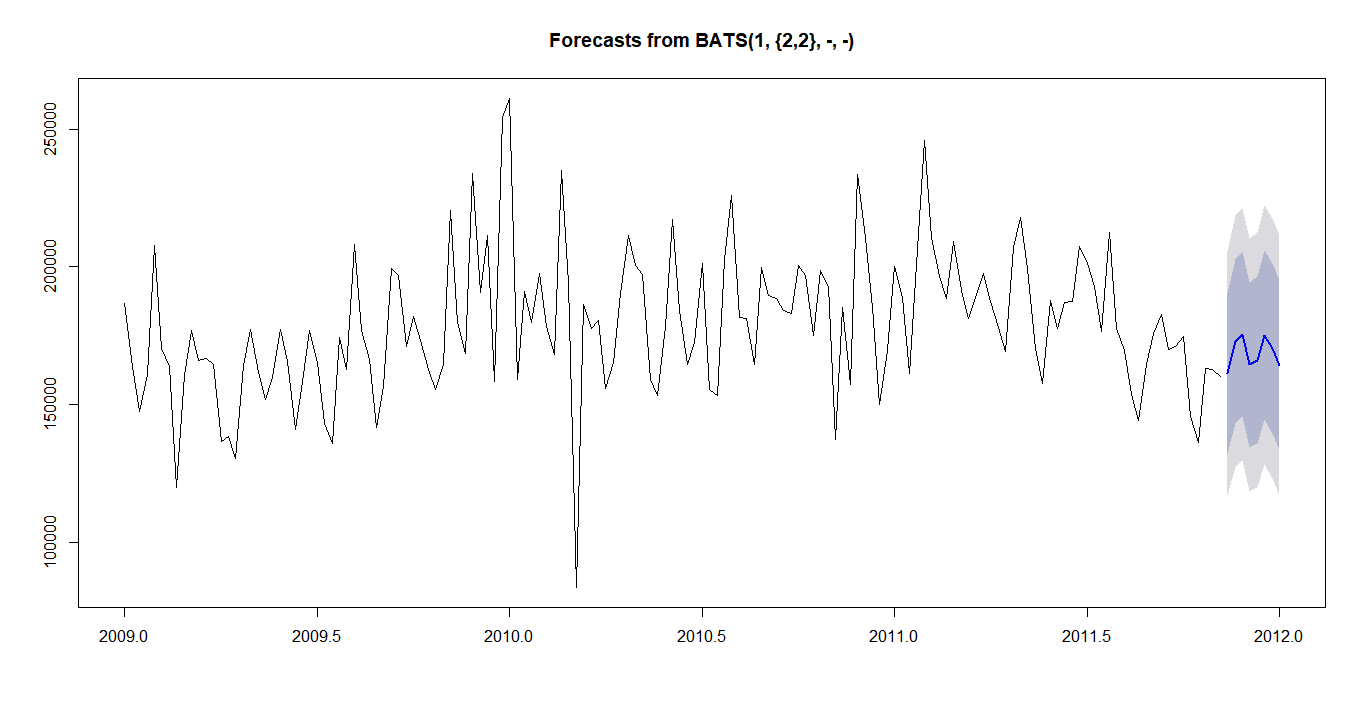
### Holt winters Model



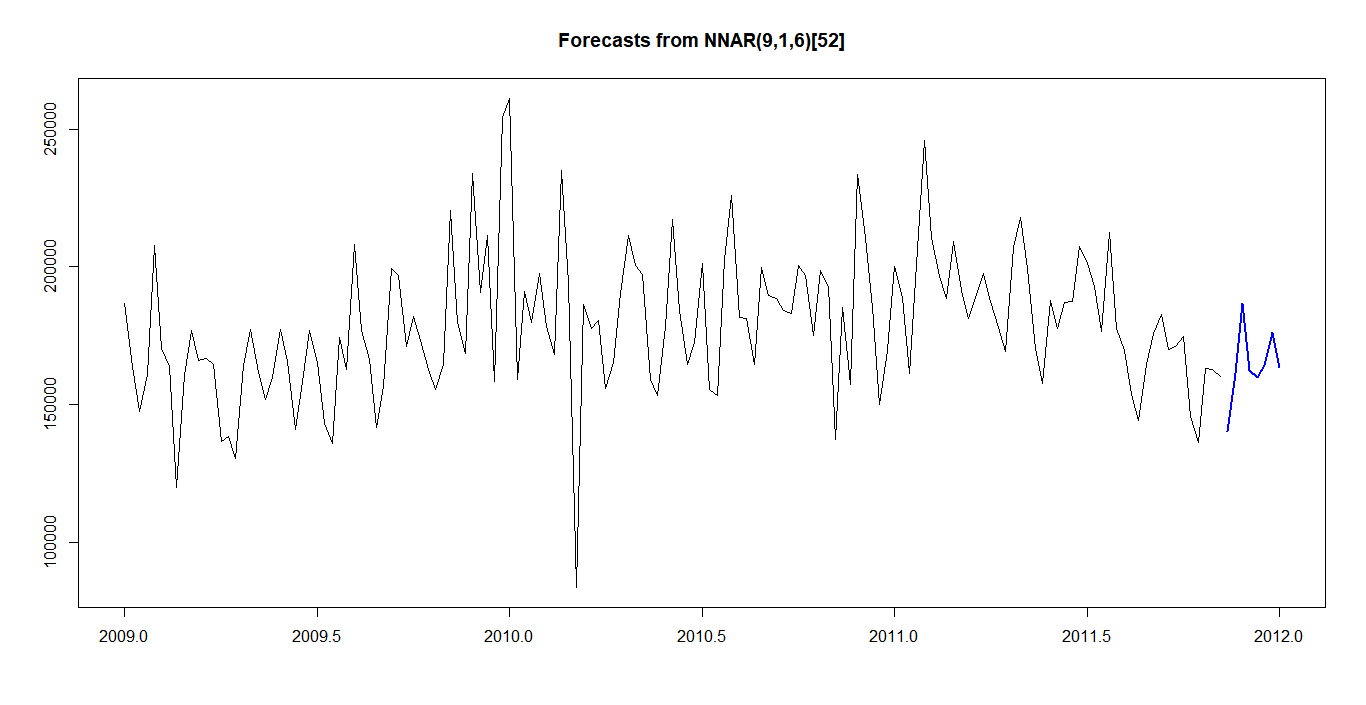
### ARIMA Model



### TBATS Model



### NNETAR Model



MAPE as Measure

Different models are verified with test set with 8 week sales data), the MAPE are listed for different model

|  |  |  |
| --- | --- | --- |
|  |  | MAPE |
| ETS | Training | 10.26999 |
| Test | 10.92934 |
| HW | Training | 11.46126 |
| Test | 32.92964 |
| ARIMA | Training | **9.952777** |
| Test | **10.87441** |
| TBATS | Training | 9.386718 |
| Test | 12.99714 |
| NN | Training | 1.109472 |
| Test | 15.54247 |

The ARIMA model is so stable with Training data as well as Test data, though Neural Net model is good in accuracy with training data set and have very high MAPE with Test data. Hence the NNETAR model could not be considered.

# Next steps

* Develop the mathematical model for profitability
* Identify the influence of promotional activity vis-à-vis sales
* Test the hypothesis highlighted
* Classify stores based on key parameters
* Calculate Price elasticity for every product
* Predict the sales based on UPC wise and aggregate to the high level
* Forecast the sales for quarter

# R Code

#Read data from Work file

path=getwd()

data.work=read.csv("work1.csv")

data.work$timestamp=as.Date(data.work$WEEKEND\_DATE,format="%d-%b-%y")

data.work=data.work[-1]

require(dplyr)

new.df = data.work %>% group\_by(timestamp) %>% mutate(SPEND\_SUM = sum(SPEND))

subs.df=new.df[,25:26]

subs.new=unique(subs.df[,1:2])

library(smooth)

library(forecast)

library(graphics)

library(datasets)

#Time Series of Sales (Spend) of Retail chain

sales.ts=ts(subs.new[,-1],start =c(2009,1),end =c(2011,45),frequency = 52)

sales.ts1=ts(subs.new[,-1],start =c(2011,46),end =c(2012,1),frequency = 52)

plot(sales.ts)

sales.comp=decompose(sales.ts)

plot(sales.comp,title(main="Decomposition of Sales"))

# Differenced Series & Autocorrelation #

d1=diff(sales.ts,lag = 1)

d2=diff(sales.ts, lag=2)

plot(d1)

auto.correl=acf(sales.ts, lag.max=50) # Autocorrelation

pa.correl=pacf(sales.ts, lag.max=10) # Partial autocorrelation

require(tseries)

adf.test(sales.ts/scale(sales.ts-sales.comp$seasonal), alternative = "stationary")

# Anova for Display and Sales

data.work$DISPLAY = as.factor(data.work$DISPLAY)

require(ez)

anova( aov(SPEND ~ DISPLAY, data=data.work) )

Mean\_data.work = data.work %>%

select(DISPLAY,SPEND) %>%

group\_by(DISPLAY) %>%

mutate (M = mean(SPEND, na.rm = T))

unique(Mean\_data.work[,c(1,3)])

#Choosing Forecasting model~ Predict next 8 future values

library(forecast)

library(dplyr)

#Exponential model

sales.ets=ets(sales.ts)

sales.fs\_ets=forecast(sales.ets,h=8)

acc.ets=accuracy(sales.fs\_ets,sales.ts1)

plot(sales.fs\_ets)

acc.ets

#Holt-Winters Model - Tuned with R code

#Tuning model

j=1

MAPE = numeric()

a1 = numeric()

b1 = numeric()

c1 = numeric()

sales.ts.dev = head(sales.ts,145)

sales.ts.hold = head(sales.ts,8)

for(a in seq(from=0.1,to= 0.9,by=0.1)){

for(b in seq(from=0.1,to= 0.9,by=0.1)){

for(c in seq(from=0.1,to= 0.9,by=0.1)){

HW\_Model = HoltWinters(sales.ts.dev, alpha=a, beta=b, gamma=c)

hold\_predict = forecast(HW\_Model,5)

MAPE[j] = sum(abs(hold\_predict$mean[1:5]-sales.ts.hold[1:5])/sales.ts.hold[1:5])/length(sales.ts.hold)

a1[j] = a

b1[j] = b

c1[j] =c

j = j+1

}

}

print(a)

}

temp = data.frame(MAPE = MAPE, alpha = a1, beta = b1, gamma = c1)

subset(temp, MAPE==min(MAPE))

#Holt Winter's Forecast

sales.HW= HoltWinters(sales.ts, alpha=0.2, beta=0.9, gamma=0.5)

sales.fs.HW= forecast(sales.HW, 8)

plot(sales.fs.HW)

acc.HW=accuracy(sales.fs.HW,sales.ts1)

acc.HW

#Arima Model

sales.aa=auto.arima(sales.ts)

sales.fs.aa = forecast(sales.aa, h=8,model=Arima)

acc.aa=accuracy(sales.fs.aa,sales.ts1)

plot(sales.fs.aa)

acc.aa

#TbATS Model

sales.tbats = tbats(sales.ts)

sales.fs.tbats = forecast(sales.tbats, h=8)

acc.tb=accuracy(sales.fs.tbats,sales.ts1)

plot(sales.fs.tbats)

acc.tb

#NNETS Model

sales.nn = nnetar(sales.ts)

sales.fs.nn = forecast(sales.nn, h=8)

acc.nn=accuracy(sales.fs.nn,sales.ts1)

plot(sales.fs.nn)

acc.nn

# References