STATISTICAL DATA MINING

PREDICTIVE ANALYSIS OF FABRIC SOFTENER BRAND

Journal of Marketing Research Vol XXXIII (November 1996), 442-452 Paper by Fader, Peter S. and Bruce G.S.Hardie on "Modeling Consumer Choice Among SKUs".

FALL 2015 Project

SUBMITTED BY:
Jagpreet Singh Sethi

TABLE OF CONTENTS

INTRODUCTION 3

Problem Statement 3

Overview of Data Set 3

Data Cleaning and Preparation 5

Data Analysis 6

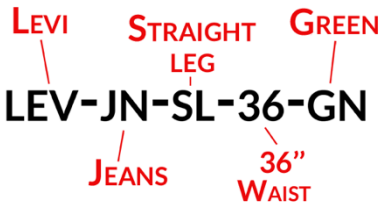
1. Forecast the brand customer has bought based on the SKUs. 6
2. Brand Analysis 7
3. Analyzing dependency of sku on its attribute 9
4. ANalyzing dependency of Price on manufacturer variables 9
5. Analyzing Loyalty of customer towards brand 10
6. To forecast high or low weekly Average sale of the store 11
7. Forecast Sales using moving average (1) model 11
8. Other USeful analysis and visualization 13

Appendix 16

A1. Manipulating with strings and consolidating 16
A2. VGLM Multinomial Logistic Regression Model 17
A3. Forecast the BRAND using Multinorm Function 18
A4. Brand Analysis using Mlogit function 19
A5. Analyzing Dependency of SKu on its attributes 20
A6. Analyzing dependency of Price on its attributes 20
A7. analyzing Loyalty of customers based on Brand 20
A8. To forecast high or low weekly Average sale of the store. 21
A9. Forecast sales using moving average model 22
A10. Other Visualization 24
A11. Ordinal Logistic Regression 25

INTRODUCTION

Stock Keeping Unit (SKU) is a unique number generated based on the characteristics such as Brand, Formula, Size, Form, Model etc. These codes are very useful to track the inventory in the warehouse. The image illustrates a SKU code used by Levi's company. The code LEV-JN-SL-36-GN signifies brand 'Levis' has a product 'Jeans' which has 'Straight Leg' fit with '36inch' waist and is in 'Green' in color.



These SKUs are tracked by stores like Walmart, BestBuy, Target, Staples to track the preferences and demand of the customers. Accordingly, these retail stores can place an order for the products.

Consumer choice model in marketing is often assumed to high correlated with Brand. It was assumed that Brand is the fundamental unit of analysis. However, it has been observed that other features like Size, Form, Color also play a vital role in customers, manufacturers and retailers making decision. Consulting companies like McKinsey and Booz Allen have started making use of these tracked SKUs to bring up better consumer choice model. These model have strong ability to forecast sales for the new product that enter into the market.

Source: *Journal of Marketing Research* Vol XXXIII (November 1996), 442-452 paper by Fader, Peter S. and Bruce G.S. Hardie on "**Modeling Consumer Choice Among SKUs**".

PROBLEM STATEMENT

Develop consumer choice model among SKUs on the Fabric Softener dataset.

OVERVIEW OF DATA SET

Source of the data is from an IRI panel in Philadelphia and cover the period from January 1991 to June 1992. Household who have made at least one purchase in 1991 are included in the data. We have 594 qualified households with a total of 6554 purchases spread over 1.5 years.

- The 4277 purchases from IRIWeek 592 to IRIWeek 641 are used for initialization or training the data.
- The 140 purchases from IRIWeek 642 to IRIWeek 643 are used for calibration of the data.
- The 2137 purchases from IRIWeek 644 to IRIWeek 669 are used for forecasting purposes.

Data was available in separated files. Following is the information about the file:

1. **D1PUR.DAT** – contains household purchase history and has two columns
    a. Column1: HHId
    b. Column2: trip_info. The information stored in trip_info is of the format AAABBBCCC.
        i. AAA stands for IRIWeek
        ii. BBB stands for store#
        iii. SKU# Purchased

2. **MERCH.DAT** – Contains information about store environment and has five columns
    a. Column1: SKU#
    b. Column2: Store#
    c. Column3: IRIWeek
    d. Column4: price_paid
    e. Column5: merchandising. The info is coded in the form of AAABCD
        i. AAA stands for the regular price
        ii. B needs to be ignore
        iii. C is the display
        iv. D is the feature
3. **ARSP.DAT** - Contains the average regular selling price of each SKU in each store and has three fields:
    a. Column1: SKU#
    b. Column2: store#
    c. Column3: ARSP
4. **BRSINFO.DAT** – Contains the attribute information for each SKUs and has 11 fields:
    a. Column1: SKU#
    b. Column2: Description of SKU#
    c. Column3: Brand
    d. Column4: Form
    e. Column5: Formula1
    f. Column6: Formula2
    g. Column7: size
    h. Column8: brand#
    i. Column9: form#
    j. Column10:
    k. Column11:
5. **IRIWeek.xls -** Contains information about the corresponding week. It is properly mapped towards the week.

DATA CLEANING AND PREPARATION

1. The data in the 4 raw DAT files is in the coded form and very difficult to interpret. So our very first task is to make the raw coded data to be in the human readable format with proper column names. We could perform this task with the help of wonderful script provided by Prof. Daniel Zantedeschi.

   The whole script in R with explanation has been given in **Appendix 1**.

2. The names of the Brand, Form, Formula and Size was in the column format with 1s and 0s values in it. To perform the Multinomial Logistic Regression (MLR) on this data, we were supposed to have all the names of the brand in single column. To achieve this task, we used EXCEL quick formula:
   =INDEX(B$1:K$1,MATCH(MAX(B2:K2),B2:K2,0))

   Where ever 1 is found in the row, it will paste corresponding brand name.

   The same task was performed for SIZE, FORMULA and FORM also.

   Sample Preview of the Final Data looks below:

	A	B	C	D	E	F	G	H	I	J	K	L
1	Brand	IRIWeek	HHId	SKU	Form	Formula2	Size	Price	PriceCut	AverageP	Display	Feature
2	PRL	592	9436	103	S	UN	MD	1.29	0	1.182	0	2
3	PRL	631	9571	103	S	UN	MD	0.99	0	1.182	0	2
4	PRL	631	9584	103	S	UN	MD	0.99	0	1.182	0	2
5	PRL	631	9376	103	S	UN	MD	0.99	0	1.182	0	2
6	PRL	621	9451	103	S	UN	MD	1	0	1.182	0	1
7	PRL	622	9595	103	S	UN	MD	1	0	1.182	0	0

3. Data was further split into training, calibration and forecast weeks.
   a. Training week primarily being IRIWeek 592 to IRIWeek 641
   b. Calibration week primarily being IRIWeek 642 to IRIWeek 643
   c. Forecast weeks primarily being IRIWeek 644 to IRIWeek 669

4. To perform the forecasting of brand bought by the customer, we removed the BRAND column from '*final_data_forecast*' sheet. So, that using Multinomial Logistic Regression model, we can predict which brand customer has bought. On a safe side, we did keep '*final_data_forecast*' with brand data as a backup.

DATA ANALYSIS

1. FORECAST THE BRAND CUSTOMER HAS BOUGHT BASED ON THE SKUS.

To forecast the customer choices on the brand, we have used MULTINOMIAL LOGISTIC REGRESSION. We executed the regression model using **three** different functions:

- Mlogit
- VGLM with family=multinomial
- Multinorm function

VGLM function gives more detail that any other function and hence it was used to figure out which independent variables are more significant than others. This was concluded by interpreting the p-value which is listed beside each variable.

PREDICTION:

Using Multinorm function, we developed a normal keeping BRAND as the dependent variable and SKU as the independent variable. Reference brand was kept as DWN.

test_model1 <- multinom(traindata$Brand2 ~ SKU, data = traindata)

Running the prediction model on the validate data set, could help us see how well the model is making prediction. It was observed that model is making prediction with 100% accurancy.

On confirming the same, prediction model was run on forecast dataset and we could actually se what all brand customer would have bought. It is to be noted that BRAND column was deleted from the forecast dataset before performing this operation. Following is the image which shows 0.9981 probability for PRL brand to be bought by customer.

BNC	CLF	FNT	GEN	PRL	SNG	STP	TSN
0	0	8.415606e-135	7.189455e-24	9.981701e-01	0.001829876	1.312423e-97	1.970305e-140
0	0	8.415606e-135	7.189455e-24	9.981701e-01	0.001829876	1.312423e-97	1.970305e-140
0	0	8.415606e-135	7.189455e-24	9.981701e-01	0.001829876	1.312423e-97	1.970305e-140
0	0	8.415606e-135	7.189455e-24	9.981701e-01	0.001829876	1.312423e-97	1.970305e-140
0	0	8.415606e-135	7.189455e-24	9.981701e-01	0.001829876	1.312423e-97	1.970305e-140
0	0	8.415606e-135	7.189455e-24	9.981701e-01	0.001829876	1.312423e-97	1.970305e-140
0	0	8.415606e-135	7.189455e-24	9.981701e-01	0.001829876	1.312423e-97	1.970305e-140
0	0	8.415606e-135	7.189455e-24	9.981701e-01	0.001829876	1.312423e-97	1.970305e-140

CONCLUSION:

- From VGLM model, we could conclude that SKU is a significant independent variable to predict BRAND and IRIWeek and HHId are insignificant variables. **Appendix 2** explains the R-Code performed to reach the conclusion.
- Using Multinorm function, we could predict the BRAND bought by customer with 100% accuracy. **Appendix 3** explain the R-Code behind predicting the BRAND.

2. BRAND ANALYSIS

Analysis of various brands was performed using mlogit function.

For the comprehensive R-Script please refer **Appendix A4**

2.1 Most Selling brand – DWN

```
mlogit.model1 <- mlogit(Brand ~ 1 , data=mldata, reflevel="DWN")
summary(mlogit.model1)
```

Interpretation: All intercept coefficients of the brands are *negative* i.e. log odds of preferring other brand over DWN decreases by exponent of coefficient value.

```
Coefficients :
                  Estimate Std. Error t-value  Pr(>|t|)
ARM:(intercept) -2.619879   0.106831 -24.524 < 2.2e-16 ***
BNC:(intercept) -1.557371   0.066716 -23.343 < 2.2e-16 ***
CLF:(intercept) -2.132736   0.085502 -24.944 < 2.2e-16 ***
FNT:(intercept) -1.413780   0.062923 -22.468 < 2.2e-16 ***
GEN:(intercept) -1.879969   0.076490 -24.578 < 2.2e-16 ***
PRL:(intercept) -0.942581   0.052561 -17.933 < 2.2e-16 ***
SNG:(intercept) -0.237578   0.041916  -5.668 1.445e-08 ***
STP:(intercept) -1.479594   0.064622 -22.896 < 2.2e-16 ***
TSN:(intercept) -2.214414   0.088695 -24.967 < 2.2e-16 ***
```

2.2 Worst selling brand – ARM

```
mlogit.model2 <- mlogit(Brand ~ 1 data=mldata, reflevel="ARM")
```

Interpretation: All intercept coefficients of the brands are *positive* i.e. log odds of preferring other brand over ARM increases by exponent of coefficient value.

Out of the following command also verifies above two conclusions:

table(Brand)

ARM	BNC	CLF	DWN	FNT	GEN	PRL	SNG	STP	TSN
97	281	160	1326	322	202	522	1067	297	143

2.3 Least Valuable Brand in terms of Price - CLF

Similarly mode is executed keeping CLF as base model and output is interpreted.

```
mlogit.model4 <- mlogit(Brand ~ 1 | Price, data = mldata, reflevel="CLF")
summary(mlogit.model4)
```

Interpretation: the price coefficient of all other brand price is *positive* in reference to CLF. With every one unit increase in variable of price the log odd of selecting other brands increase over CLF. Hence, people prefer other brands over CLF.

```
ARM:Price      2.37497    0.27191   8.7344 < 2.2e-16 ***
BNC:Price      3.68446    0.25284  14.5725 < 2.2e-16 ***
DWN:Price      3.90079    0.24343  16.0240 < 2.2e-16 ***
FNT:Price      3.04861    0.24803  12.2914 < 2.2e-16 ***
GEN:Price      0.49113    0.28212   1.7409   0.08171 .
PRL:Price      2.08944    0.24119   8.6631 < 2.2e-16 ***
SNG:Price      4.22030    0.24540  17.1975 < 2.2e-16 ***
STP:Price      3.54787    0.25124  14.1214 < 2.2e-16 ***
TSN:Price      2.14811    0.26088   8.2339 2.220e-16 ***
```

2.4 Most Valuable Brand in terms of Price - SNG

Similarly model is executed keeping SNG as reference model and output is interpreted as below:

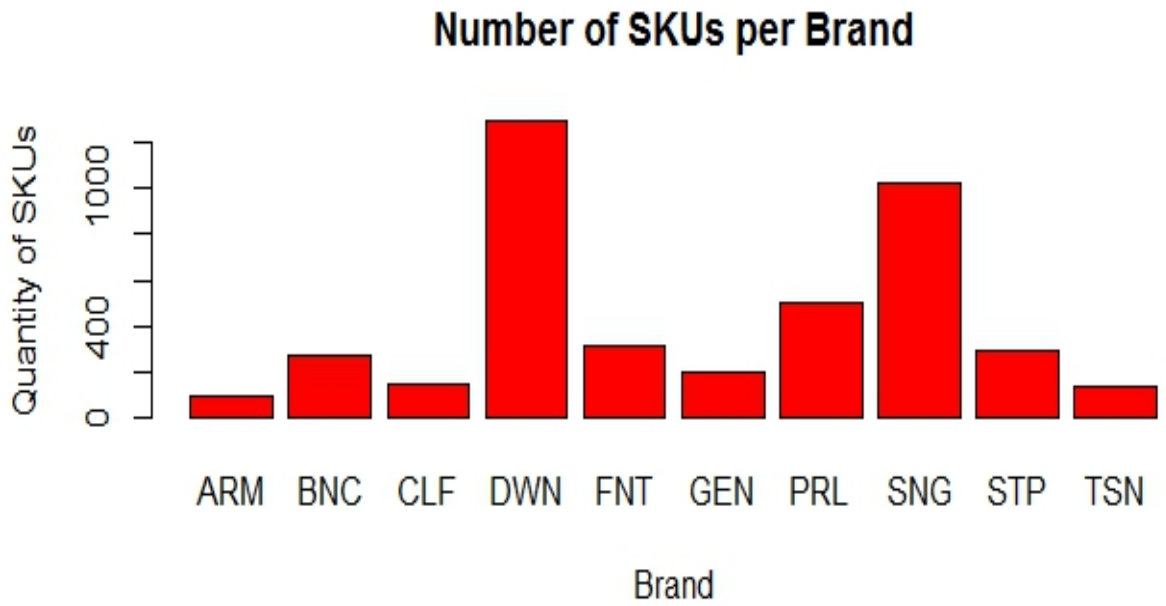
```
mlogit.model3 <- mlogit(Brand ~ 1 | Price, data = mldata, reflevel="SNG")
summary(mlogit.model3)
```

Interpretation: the price coefficient of all other brand price is *negative* in reference to SNG. With every one unit increase in variable of price the log odd of selecting other brands decreases over SNG. Hence, people start preferring SNG.

VISUALIZATION:

We can observe that DWN brand has the most SKUs which signifies it is the brand which is demanded the most. SNG is the next best demanding brand and ARM the least demanding brand.

Accordingly, retailers, manufactures can make decision to put stock in their warehouse.



3. ANALYZING DEPENDENCY OF SKU ON ITS ATTRIBUTE

A linear regression model was executed keeping SKU as the dependent variable and other variables as independent. Starting with Kitchen Sink model, we gradually started removing variables which aren't explaining much information.

We could **conclude** that SKU can be explanation with the help of variance in FORMULA2, FORM, SIZE and BRAND.

We can observe that Multiple R-squared is 99.82% and Adjusted R-squared is also 99.82%. This signifies there is no over-fitting and no interaction occurring between the variables.

Appendix 5 explains the complete R-Script on this operation.

```
Call:
lm(formula = SKU ~ Formula2 + Form + Size + Brand)

Residuals:
    Min      1Q  Median      3Q     Max
-7.6442 -0.6774  0.2842  0.6289  2.8185

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.95713    0.19319  51.542  < 2e-16 ***
Formula2RG   -3.23747    0.06583 -49.182  < 2e-16 ***
Formula2ST   -0.21707    0.20350  -1.067  0.28616
Formula2UN    0.22841    0.15579   1.466  0.14267
FormF         6.75088    0.11438  59.021  < 2e-16 ***
FormL         4.82677    0.18463  26.142  < 2e-16 ***
FormS        10.57346    0.07244 145.956  < 2e-16 ***
SizeMD        0.04457    0.07976   0.559  0.57636
SizeSM       -1.98239    0.10604 -18.694  < 2e-16 ***
SizeXL       -0.45346    0.14436  -3.141  0.00169 **
BrandBNC      6.23414    0.20528  30.370  < 2e-16 ***
BrandCLF     14.69389    0.20059  73.252  < 2e-16 ***
BrandDWN     31.95769    0.17045 187.491  < 2e-16 ***
BrandFNT     55.50017    0.18947 292.923  < 2e-16 ***
BrandGEN     56.99716    0.22571 252.529  < 2e-16 ***
BrandPRL     79.37795    0.17606 450.868  < 2e-16 ***
BrandSNG     99.95162    0.17075 585.367  < 2e-16 ***
BrandSTP    117.65147    0.19140 614.681  < 2e-16 ***
BrandTSN    112.68454    0.19880 566.824  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.49 on 4258 degrees of freedom
Multiple R-squared:  0.9982,    Adjusted R-squared:  0.9982
F-statistic: 1.308e+05 on 18 and 4258 DF,  p-value: < 2.2e-16
```

4. ANALYZING DEPENDENCY OF PRICE ON MANUFACTURER VARIABLES

A linear regression mode was executed keeping PRICE as dependent variable and other variables as independent. We started with Kitchen Sink model gradually removed the one which seem to have caused interaction i.e SKU. We could achieve a model which explains PRICE upto 87.27%. Variables BRAND, FORM, FORMULA2, SIZE, DISPLAY and FEATURE explains the maximum variance in the PRICE variable.

So, we can **conclude** that Price of the product actually depends on ever manufacturing variable.

Appendix A6 has the complete R-Script for this operation.

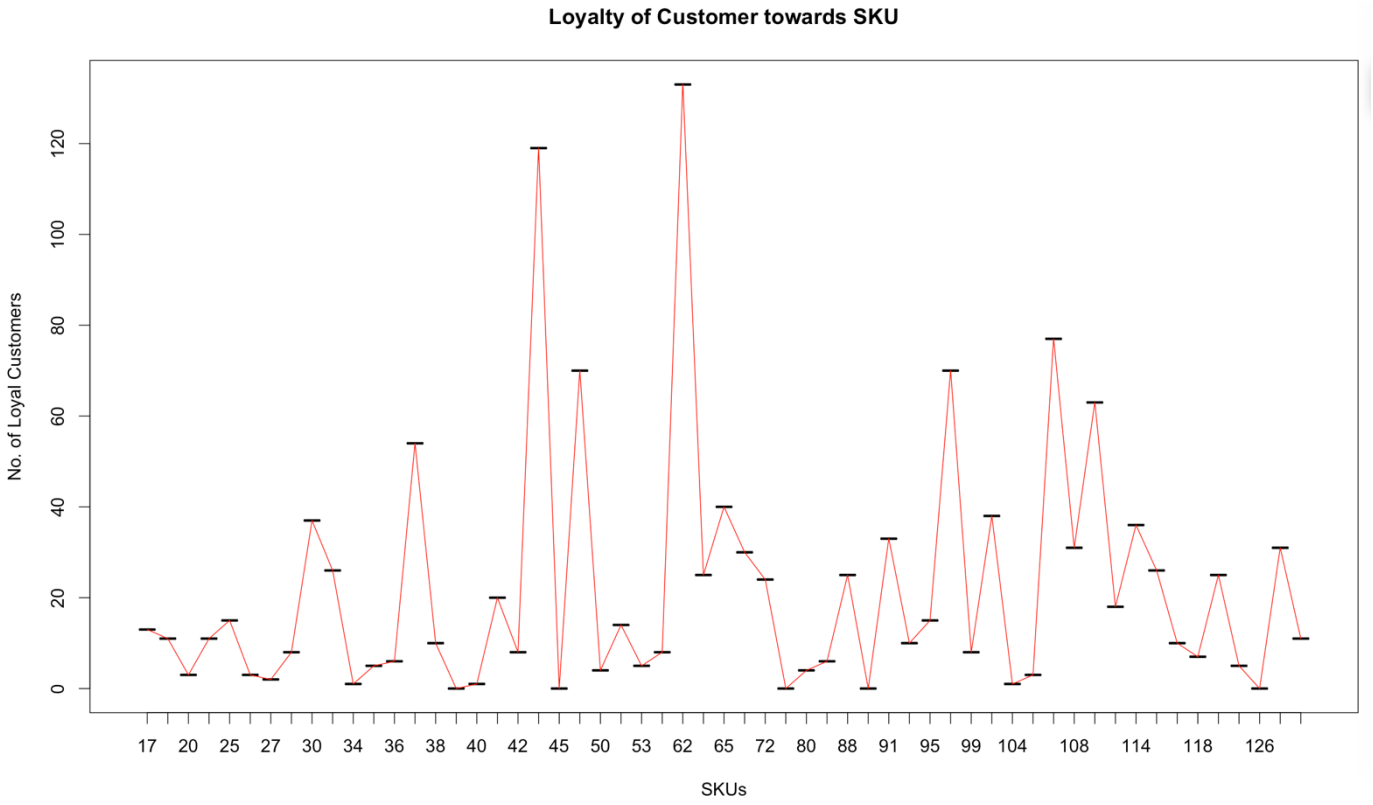
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.5079726  0.0431249  81.345  < 2e-16 ***
BrandBNC     0.8127659  0.0473410  17.168  < 2e-16 ***
BrandCLF     0.0682104  0.0447137   1.525  0.12721
BrandDWN     0.8746513  0.0382207  22.884  < 2e-16 ***
BrandFNT     0.0097991  0.0422594   0.232  0.81664
BrandGEN    -2.5249913  0.0502932 -50.205  < 2e-16 ***
BrandPRL    -0.8453370  0.0391840 -21.574  < 2e-16 ***
BrandSNG     0.8184133  0.0384133  21.305  < 2e-16 ***
BrandSTP    -0.0943348  0.0427081  -2.209  0.02724 *
BrandTSN    -0.1710963  0.0442108  -3.870  0.00011 ***
SizeMD      -1.1283469  0.0177923 -63.418  < 2e-16 ***
SizeSM      -2.3686340  0.0242290 -97.760  < 2e-16 ***
SizeXL       0.4890927  0.0321921  15.193  < 2e-16 ***
FormF        0.6714009  0.0256430  26.183  < 2e-16 ***
FormL        0.0001127  0.0411332   0.003  0.99781
FormS       -0.3053889  0.0164347 -18.582  < 2e-16 ***
Formula2RG  -0.0463677  0.0146507  -3.165  0.00156 **
Formula2ST  -0.0041562  0.0458812  -0.091  0.92783
Formula2UN   0.3739864  0.0346559  10.791  < 2e-16 ***
Display      0.0265755  0.0041180   6.454 1.21e-10 ***
Feature     -0.0213907  0.0073612  -2.906  0.00368 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3313 on 4256 degrees of freedom
Multiple R-squared:  0.8727,    Adjusted R-squared:  0.8721
F-statistic:  1458 on 20 and 4256 DF,  p-value: < 2.2e-16
```

5. ANALYZING LOYALTY OF CUSTOMER TOWARDS BRAND

In the forecast dataset, a new column is created which has 0 and 1 on the basis whether customer/HHId has opted for same SKU in the past i.e. test dataset. If it does then the customer is loyal else it is not.

To perform this activity, we performed few operations in MS Excel using VLOOKUP formula and then did analysis. Following plot between SKUs and No. of Loyal Customers illustrates that maximum number of customer are loyal towards two SKUs

1. **SKU 62** (Brand FNT, Form B, Formula2 RG, Size MD)
2. **SKU 44** (Brand DWN, Form F, Formula2 RG, Size SM)



Appendix A7 contains the complete R-Script on how loyalty is evaluated.

6. TO FORECAST HIGH OR LOW WEEKLY AVERAGE SALE OF THE STORE

To perform this operation, we used Binary Logistic Regression. Weekly average of the Total_price was calculated on a weekly basis. Keeping a threshold of 2.6, the average sale of the store was divided between high and low. These high and low were saved in a new column called SPENDING

Binary Logistic Regression with SPENDING as the response variable and AVERAGEPRICE, PRICECUT were used in the predictor variable to perform the forecasting on the forecast dataset.

The probabilities were calculated and any probability more than 0.5 was considered favorable.

We could observe that 1193 low spending were rated low and 281 high spending were rated high.

The model was able to give us **70% accuracy** which is decent considering less training data available.

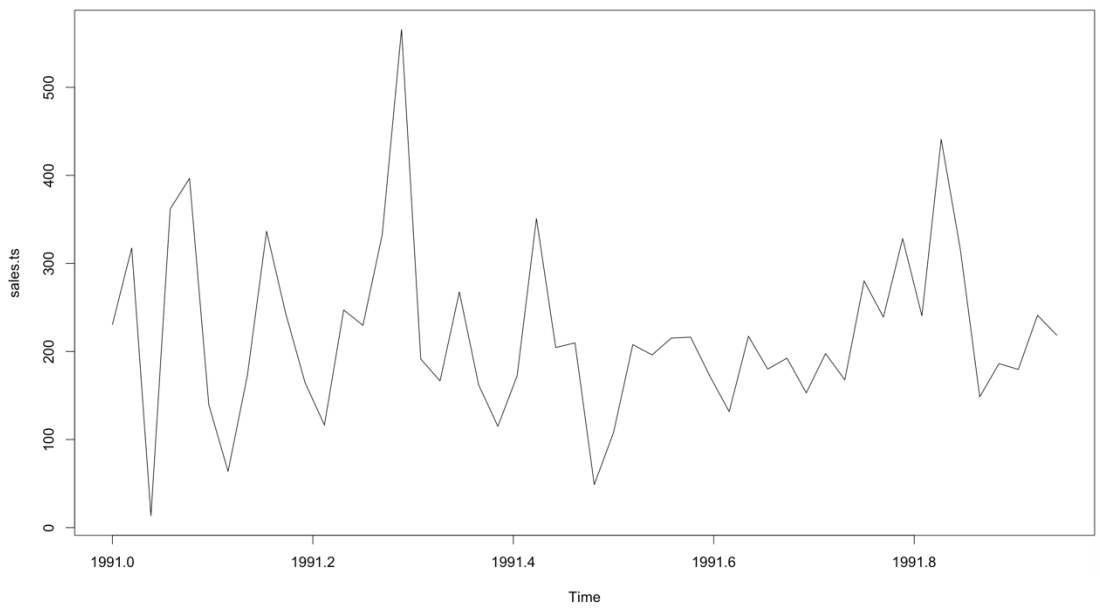
```
> table(foredata$Spending,as.numeric(pred.spending))

     0   1
0 1193 364
1  299 281
> table(foredata$Spending)

   0    1
1557  580
> accuracy_rate <- (1193+281)/(1193+364+299+281)
> accuracy_rate
[1] 0.689752
```

Appendix A8 explains the detailed R-Script behind this forecasting method.

We tried to perform the same activity using three ordered levels – High , Medium, Low using **ORDINAL LOGISTIC REGRESSION**. Unfortunately, we couldn't achieve much accuracy on the predicted results. **Appendix A11** contains the details code and the explanation on the Ordinal Logistic Regression efforts.

7. FORECAST SALES USING MOVING AVERAGE (1) MODEL

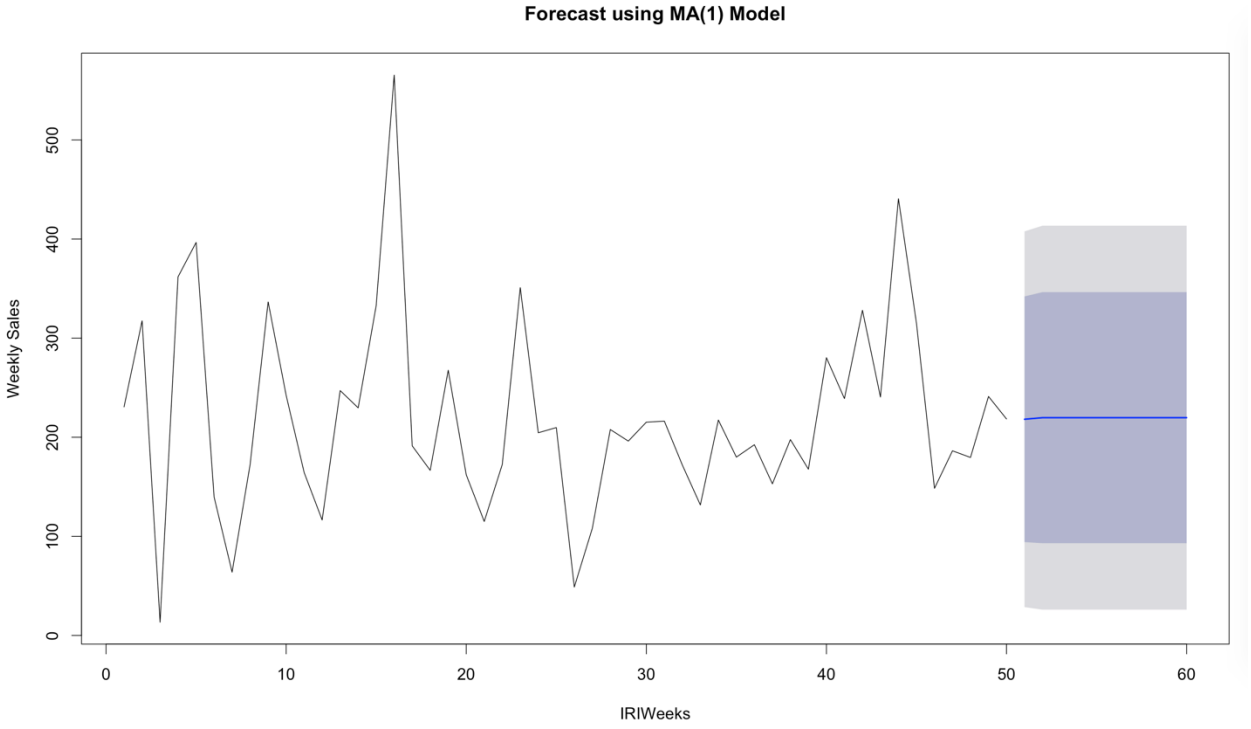
A new column containing the sum of total_price on a weekly basis was created in the test dataset. Using unique and sorted entries time series plot was created. This gives us a following time series graph with sales on x-axis and IRI Week on the y-axis



Using Dickey-Fuller test we could see that p-value is very small for the null hypothesis of time series being non-stationary, and hence we could reject it and concluded that time series graph over 1 year is Stationary.

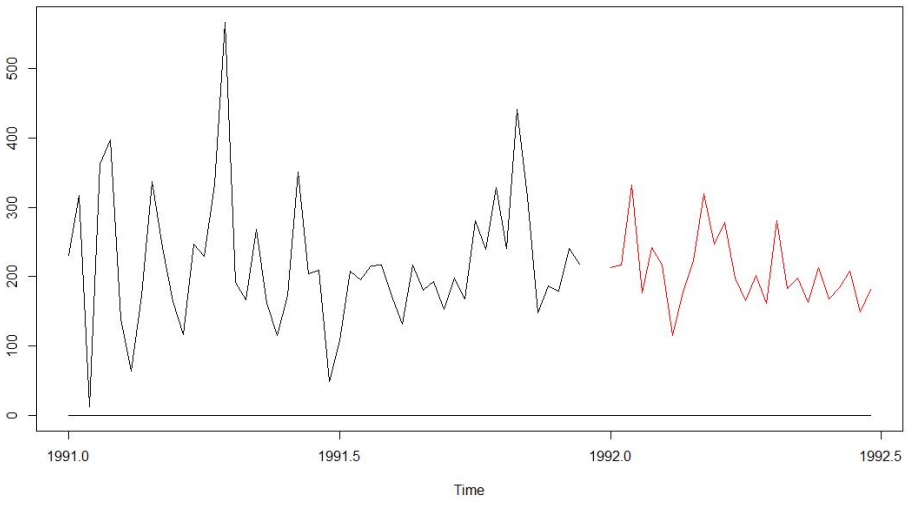
Autocorrelation function (ACF) gave us a single significant point and partial auto-correlation function (PACF) has nothing significant. This help us conclude that Moving Average (1) needs to be executed.

Using Arima function we gave c(0,0,1) as a parameter and we could achieve a Moving Average (1) model, which was used for forecast on the forecast dataset.



Above image gives us the 80% and 95% confidence interval over which our future sales would exist. This prediction was done for next 10 weeks by giving a parameter to forecast.arima.

The graph in BLACK is the sales from the training data and graph in RED is the sales from the forecast data set. We compared the actual forecasted sales data and the range over which MA model forecasted it.



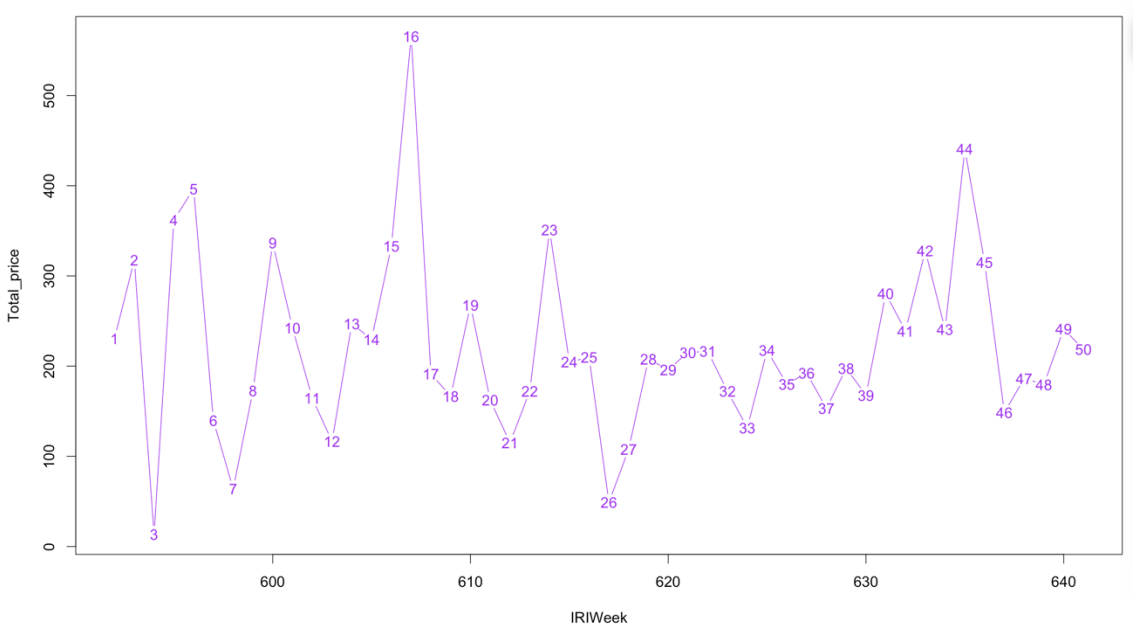
Conclusion: It was inside that range and our prediction of sales was accurate.

Appendix A9 contains the detailed R-Script with explanation for forecasting sales using MA Model.

8. OTHER USEFUL ANALYSIS AND VISUALIZATION

1. Sales per IRIWeek.

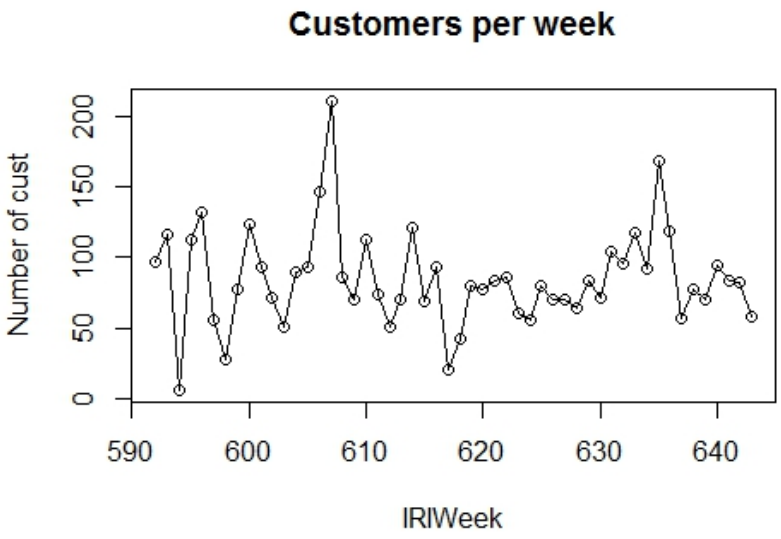
Interpretation: IRIWeek 16 clocked highest sales i.e 14 Apr 1991 to 21 Apr 1991. May be because of some festival or beginning of Summer season people are in a mood to go for shopping and spend more.



2. Customers per IRIWeek.

Through the graph looks similar to Sales per IRIWeek, but it is not. There is slight variation specially between 600 to 610 weeks. However, it is easy to predict that with increase in customer, store should expect more sales.

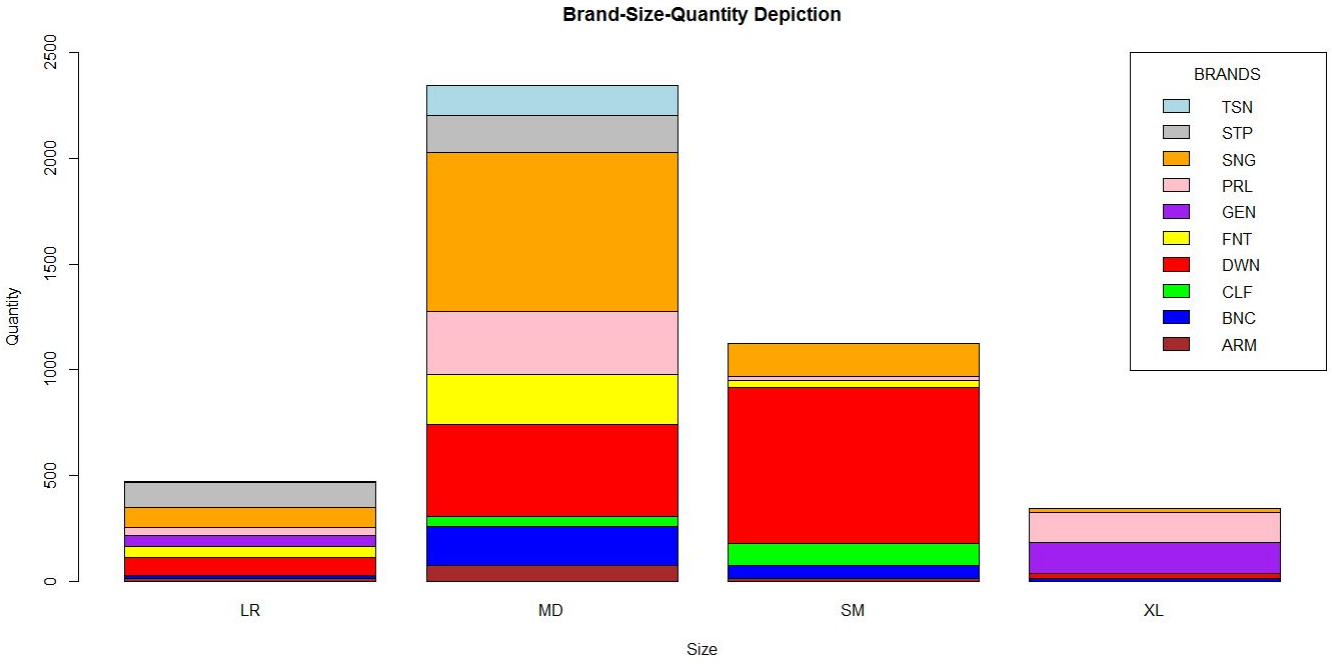
There might be cases that even few heavy customers do the required monthly sales target of the store, but from the data we can say all the customer do similar price shopping with not so drastic variance.



3. Quantity vs Size.

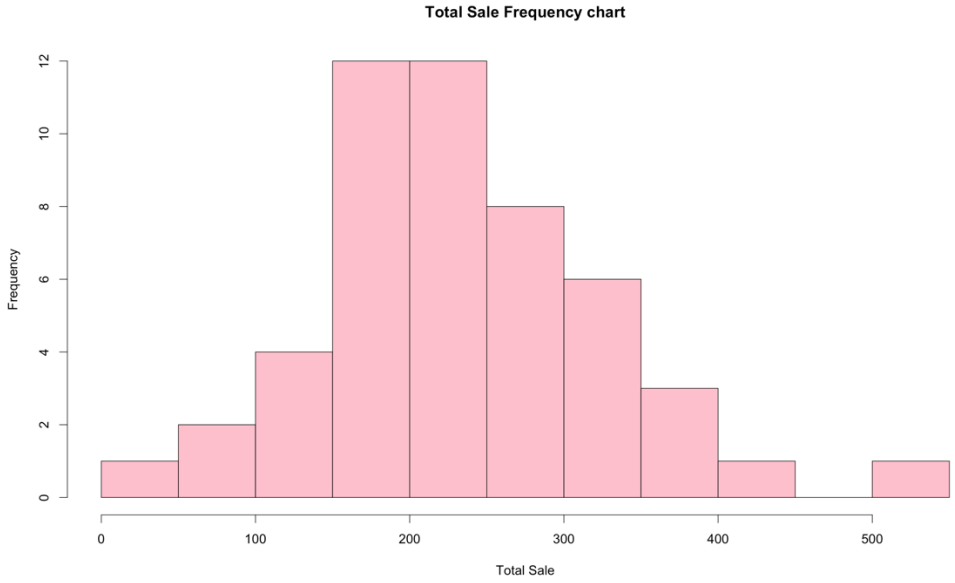
From the graph we can see that Medium size is the most sold size and XL being the least sold size. People generally like to keep themselves fit and generally fall in a category of Medium Size. Cases of Obesity are rare and hence the requirement of XL size is very less.

Medium size of Brand SNG is being sold the most and brand DWN sells Small size the most.



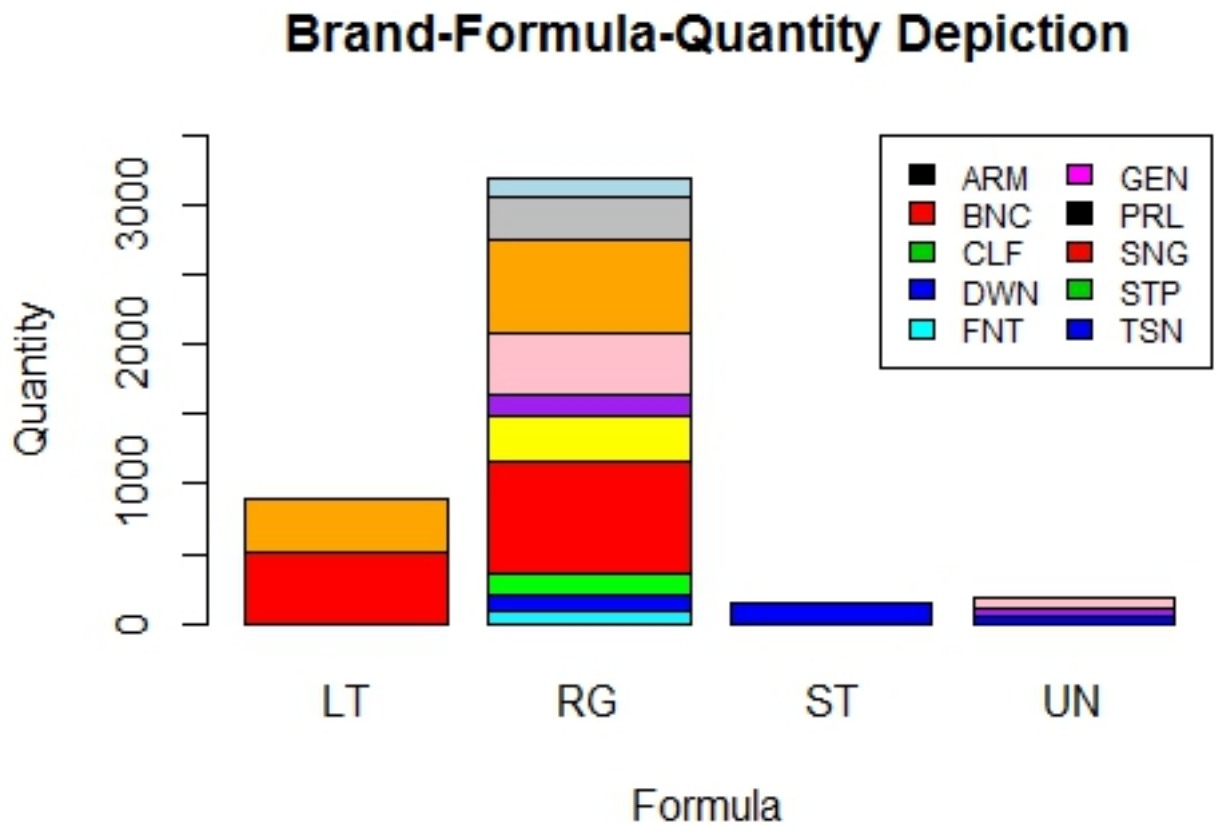
4. Sales Frequency.

From the graph we could see that for most part of the year, store clocks sales between 150 and 250. For very few week, sales are extreme.



5. Formula Frequency

From the graph we can see that people highly prefer brand which has RG as a Formula. Manufacture should buld more product with RG as a formula this will actually give them more turn-over in sales.



APPENDIX

A1. MANIPULATING WITH STRINGS AND CONSOLIDATING 4 DIFFERENT RAW FILES – D1PUR.DAT, MERCH.DAT, ARSP.DAT, BRSINFO.DAT INTO 1 FILE – FINALIZED DATA.CSV

```
setwd("/Users/jagpreet/Downloads/fabric_softener/")
#File D1PUR.DAT is read and strings are manipulated to assign to proper column name. So that it is in human readable format. Columns HHId, IRIWeek, Store and SKUs are created out of coded data.
purdata<-read.table("D1PUR.DAT")
purdata$IRIWeek<-substring(purdata$V2,1,3)
purdata$Store<-as.numeric(substring(purdata$V2,4,6))
purdata$SKU<-as.numeric(substring(purdata$V2,7,9))
purdata<-purdata[,c("V1","IRIWeek", "Store", "SKU")]
names(purdata)<-c("HHId","IRIWeek","Store","SKU")

#File MERCH.DAT is read and strings are manipulated to assign to proper column name. So that it is in human readable format. Columns SKU, Store, IRIWeek, PricePaid, RegPrice, Display, Feature are created out of coded data.
merchdata<-read.table("MERCH.DAT")
for(i in 1:length(merchdata$V5)) {
if(nchar(merchdata[i,"V5"])<6){
ZeroString<-character()
for(j in 1:(6-nchar(merchdata[i,"V5"]))) {
ZeroString<-paste(ZeroString,0,sep='') }
merchdata[i,"V5"]<-paste(ZeroString,merchdata[i,"V5"],sep='') }}
merchdata$Price<-as.numeric(substring(merchdata$V5,1,3))
merchdata$Display<-as.numeric(substring(merchdata$V5,5,5))
merchdata$Feature<-as.numeric(substring(merchdata$V5,6,6))
merchdata$Price<-merchdata$Price/100
merchdata<-merchdata[,-5]
merchdata<-merchdata[,c("V1","V2","V3","V4","Price","Display","Feature")]
names(merchdata)<-c("SKU","Store","IRIWeek","PricePaid","RegPrice","Display","Feature")
merchdata$IRIWeek<-as.numeric(merchdata$IRIWeek)
purplusmerch <- merge(purdata, merchdata, by=c("IRIWeek", "Store","SKU"))
#After processing the BRSINFO.DAT, Membership Panel Data file is created which is used further for processing.
attrdata<-read.csv("Membership panel Data.csv")
attrdata<-attrdata[,-1]
attrplusmerch <- merge(purplusmerch, attrdata, by=c("SKU"))
arspdata<-read.table("ARSP.DAT")
names(arspdata)<-c("SKU","Store","ARSP")
# All the files are merged here to form a single finalized file
finaldata <- merge(attrplusmerch, arspdata, by=c("SKU","Store"))
finaldata<-
finaldata[,c("HHId","SKU","IRIWeek","ARM","BNC","CLF","DWN","FNT","GEN","PRL","SNG","STP","TSN","B","F","L"
,"S","LT","RG","ST","UN","LR","MD","SM","XL","PricePaid","RegPrice","ARSP","Display","Feature")]
finaldata$PriceCut<-finaldata$RegPrice-finaldata$PricePaid
```

```
finaldata<-
finaldata[,c("HHId","SKU","IRIWeek","ARM","BNC","CLF","DWN","FNT","GEN","PRL","SNG","STP","TSN","B","F","L"
,"S","LT","RG","ST","UN","LR","MD","SM","XL","RegPrice","PriceCut","ARSP","Display","Feature")]
names(finaldata)<-
c("HHId","SKU","IRIWeek","ARM","BNC","CLF","DWN","FNT","GEN","PRL","SNG","STP","TSN","B","F","L","S","LT","
RG","ST","UN","LR","MD","SM","XL","Price","PriceCut","AveragePrice","Display","Feature")
write.csv(finaldata," finalized_data.csv",row.names=FALSE)
```

A2. VGLM MULTINOMIAL LOGISTIC REGRESSION MODEL – TO CHECK FOR THE SIGNIFICANCE OF VARIOUS PREDICTOR VARIABLES.

```
install.packages("VGAM")
library(VGAM)
setwd("/Users/jagpreet/Downloads/fabric_softener/")
traindata<-read.csv("Final_Data_Training.csv")
class(SKU)
SKU <- as.factor(SKU)
```

#Using vglm function model to predict the significant independent variables.

```
vglm_mod1=vglm(cbind(B.ARM,B.BNC,B.CLF,B.FNT,B.GEN,B.PRL,B.SNG,B.STP,B.TSN,B.DWN)~SKU+IRIWeek+H
HId, data=traindata, family=multinomial)
summary(vglm_mod1)
exp(coefficients(vglm_mod1))
```

#Conclusion: Only Brand and SKUs are strong covariant and explain the variance. Other variables like HHId and IRIWeek doesn't have much significant in the data variation. This was concluded by looking at the p-values.

#Using vglm function model to predict if Price attribute is dependent on Brand.

```
SKU:1        -6.408e+00  3.315e-01 -19.329  < 2e-16 ***
SKU:2        -4.519e+00  2.574e-01 -17.559  < 2e-16 ***
SKU:3        -1.988e+00  1.573e-01 -12.644  < 2e-16 ***
SKU:4         1.421e+00  1.628e-01   8.727  < 2e-16 ***
SKU:5         4.366e+00  2.507e-01  17.413  < 2e-16 ***
SKU:6         5.522e+00  2.961e-01  18.650  < 2e-16 ***
SKU:7         8.659e+00  4.063e-01  21.314  < 2e-16 ***
SKU:8         1.034e+01  4.417e-01  23.417  < 2e-16 ***
SKU:9         1.248e+01  4.733e-01  26.371  < 2e-16 ***
IRIWeek:1     7.717e-02  4.071e-02   1.896 0.057998 .
IRIWeek:2    -5.265e-03  2.426e-02  -0.217 0.828224
IRIWeek:3     2.104e-02  1.902e-02   1.107 0.268488
IRIWeek:4    -2.846e-03  5.144e-02  -0.055 0.955879
IRIWeek:5     1.173e-03  5.455e-02   0.021 0.982848
IRIWeek:6    -1.086e-02  6.949e-02  -0.156 0.875777
IRIWeek:7     8.957e-02  7.404e-02   1.210 0.226367
IRIWeek:8     4.695e-02  8.263e-02   0.568 0.569938
IRIWeek:9    -7.570e-03  8.934e-02  -0.085 0.932473
HHId:1       -1.907e-03  2.903e-03  -0.657 0.511101
HHId:2       -8.632e-04  1.805e-03  -0.478 0.632469
HHId:3       -4.939e-04  1.428e-03  -0.346 0.729503
HHId:4        5.303e-04  4.223e-03   0.126 0.900087
HHId:5       -4.905e-05  4.510e-03  -0.011 0.991321
HHId:6       -1.582e-03  5.700e-03  -0.278 0.781289
HHId:7       -1.898e-03  5.920e-03  -0.321 0.748473
HHId:8       -2.147e-03  6.699e-03  -0.320 0.748605
HHId:9       -2.430e-03  7.314e-03  -0.332 0.739731
```

```
vglm_mod2=vglm(cbind(B.ARM,B.BNC,B.CLF,B.FNT,B.GEN,B.PRL,B.SNG,B.STP,B.TSN,B.DWN)~Price,
data=traindata, family=multinomial)
summary(vglm_mod2)
exp(coefficients(vglm_mod2))
```

#Conclusion: P-values could tell us that Price and Brands are highly correlated. From this analysis we can concluded that we have better chance at creating multinomial regression model on "Brand Vs SKUs" or "Brand vs Price" for predictive analysis.

A3. USING MULTINORM FUNCTION - FORECAST THE BRAND BOUGHT BY THE CUSTOMER USING SKUs AS INDEPENDENT VARIABLE

```
require(foreign)
require(nnet)
traindata<-read.csv("Final_Data_Training.csv", header = TRUE)
traindata$Brand2 <- relevel(traindata$Brand, ref = "DWN")
#DWN brand is kept at a reference level
train_model1 <- multinom(traindata$Brand2 ~ SKU, data = traindata)
summary(train_model1)
z <- summary(train_model1)$coefficients/summary(train_model1)$standard.errors
z
p <- (1 - pnorm(abs(z), 0, 1)) * 2
p
# Z-value and P-value is listed by a separate formula as multinorm doesn't explicitely displays these values.
exp(coef(train_model1))
head(pp <- fitted(train_model1))
validate_data<-read.csv("Final_Data_Validation.csv")
head(predict(train_model1, newdata = validate_data, "probs"))
# Gives prediction probabilities on the validate data. From here we can validate the accuracy of the model.
forecast_data<-read.csv("Final_Data_Forecast.csv")
```

```
> head(predict(test_model1, newdata = validate_data, "probs"))
            DWN ARM BNC CLF           FNT           GEN          PRL          SNG           STP           TSN
1 3.390601e-261   0   0   0 8.415606e-135 7.189455e-24 9.981701e-01 0.001829876 1.312423e-97 1.970305e-140
2 1.036738e-269   0   0   0 1.304002e-140 7.722626e-27 7.453192e-03 0.992546808 9.494042e-90 8.060073e-131
3 1.871144e-303   0   0   0 3.062795e-166 6.042730e-44 1.958920e-17 1.000000000 2.269049e-74 3.483506e-110
4 1.871144e-303   0   0   0 3.062795e-166 6.042730e-44 1.958920e-17 1.000000000 2.269049e-74 3.483506e-110
5 1.871144e-303   0   0   0 3.062795e-166 6.042730e-44 1.958920e-17 1.000000000 2.269049e-74 3.483506e-110
6 1.871144e-303   0   0   0 3.062795e-166 6.042730e-44 1.958920e-17 1.000000000 2.269049e-74 3.483506e-110
```

```
Brandpred_value <- data.frame(predict(test_model1, newdata = forecast_data, "probs"))
Brand_predicat <- cbind(forecast_data,Brandpred_value)
View(Brand_predicat)
```

```
#To check for the ACCURACY(Lets keep the cut off probability as 0.9)
cut.off <- 0.9
pred.brand <- (Brandpred_value > cut.off)
#No of purchase rows for Brand "PRL" with probability higher than 0.9 are 350 which is true if we see from the file directly.
table(Brand_predicat$PRL)
table(forecast_data$Brand)
#Similarly for Brand "DWN" are 452 which is also true.
table(Brand_predicat$DWN)
```

```
> table(Brand_predicat$PRL)

   0    1
1787  350
> table(forecast_data$Brand)

ARM BNC CLF DWN FNT GEN PRL SNG STP TSN
 68  99 123 452 211 118 350 539 108  69
> #Similarly for Brand "DWN" are 452 which is also true.
> table(Brand_predicat$DWN)

   0    1
1685  452
```

```
#Hence our multinom logistic regression model is highly accurate.
```

A4. BRAND ANALYSIS USING MLOGIT FUNCTION

```
#Multinomial Logistic Regression Model using mlogit
install.packages("mlogit")
library(mlogit)

#The training file been used here contains the data from Final data file created after Data cleaning only for IRIWeeks 592-641.
traindata<-read.csv("Final_Data_Training.csv")
attach(traindata)

#Descriptive statistics of Brand Variable. There are 10 Different Brands with corresponding purchase rows in Training Dataset
table(Brand)

#Reshaping the data from wide to long format
traindata$Brand<-as.factor(traindata$Brand)
mldata<-mlogit.data(traindata, varying=13:22, choice="Brand", shape="wide")
mldata[1:25,]

# Multinomial logit model coefficients
#MOST SELLING BRAND - DWN - All intercept coefficients of the brands are negative i.e. log odds of preferring other brand over DWN decreases by exponent of coefficient value.
mlogit.model1 <- mlogit(Brand ~ 1,  data=mldata, reflevel="DWN")
summary(mlogit.model1)
exp(coef(mlogit.model1))
#Brand and IRIWeek are the only two attributes that are highly correlated because for other predictor values like HHId,Form,Formula2,Size,etc. the p-value was not significant (>0.05)

#WORST SELLING BRAND - ARM - All intercept coefficients of the brands are positive i.e. log odds of preferring other brand over ARM increases by exponent of coefficient value.
mlogit.model2 <- mlogit(Brand ~ 1,  data=mldata, reflevel="ARM")
summary(mlogit.model2)
exp(coef(mlogit.model2))

# Multinomial logit model coefficients (with different base outcome)
#SNG is the most valuable brand in terms of Price since the price coefficient of all other brand price is negative in reference to SNG. With every one unit increase in variable of price the log odd of selecting other brands decreases over SNG. Hence, people start preferring SNG.
mlogit.model3 <- mlogit(Brand ~ 1 | Price, data = mldata, reflevel="SNG")
summary(mlogit.model3)
exp(coef(mlogit.model3))

#CLF is the least valuable brand in terms of Price since the price coefficient of all other brand price is positive in reference to CLF. With every one unit increase in variable of price the log odd of selecting other brands increase over SNG. Hence, people prefer other brands over SNG
mlogit.model4 <- mlogit(Brand ~ 1 | Price, data = mldata, reflevel="CLF")
summary(mlogit.model4)
exp(coef(mlogit.model4))
```

A5. ANALYZING DEPENDENCY OF SKU ON ITS ATTRIBUTES

 # A linear model with SKU as dependent and other variables FORMULA2, FORM, SIZE and BRAND as independent.

```
Lmod1 <- lm(SKU~Formula2 + Form + Size + Brand )
summary(Lmod1)
```

#The model explains 99.81% of the variance in the SKU by those variables. Also the adjusted R2 was exactly same, which signifies there is no interaction and no over-fitting among the independent variable.

A6. ANALYZING DEPENDENCY OF PRICE ON ITS ATTRIBUTES

 # A Linear model with Price as dependent and all other variable signifies that variance of the price is explained by all the environment variables. Every manufactures variable will effect the price of the product.

```
Lmod2 <- lm(Price~Brand+Size+Form+Formula2+Display+Feature)
summary(Lmod2)
```

A7. ANALYZING LOYALTY OF CUSTOMERS BASED ON BRAND

#Loyalty Check Script and Number of SKUs per Brand

#The below file contains additional Column "Loyalty" that has "0" or "1" value in case the Customer/HHId opted for same SKU in the Test dataset.
#To create a Column "Loyalty" I have used Excel formula. I combined both HHId and SKU data into one column using "=C2&" "&D2" for both files i.e. Testing having 4277 rows(into column 'AK') and Forecast having 2137 rows(into column 'AM') into single excel sheet.
#I also copied and pasted all SKU column from Test dataset into Column 'AL' of this combined sheet.
#After this, I have used excel formula using "=VLOOKUP(AM2,AK2:AL4418,2,FALSE) to get all corresponding SKU values for that HHId Match from calibration dataset and put it in the forecast dataset in Col 'AN'. Once i have SKUs from Forecast(col 'D2') and SKUs from Calibration(col 'AN') side by side on every HHId(col'C2'), simply used formula '=IF(D2=AN,1,0)' i.e. if SKU from forecast is same as SKU from Calibration put '1' denoting the customer/HHId is loyal and so on.
#We then sort the file based on HHId and then on SKUs.

```
    forecastdata<-read.csv("Final_Data_Forecast.csv")

    plot(Loyalty~SKU)

    loyal<-as.data.frame(table(SKU, Loyalty))
    loyal
```

#To plot graph for loyal customer, we have to take count from 58-114 row of the above table "loyal"

```
    plot(loyal$SKU[58:114],loyal$Freq[58:114], main="Loyalty of Customer towards SKU", xlab="SKUs", ylab="No. of Loyal Customers")
    lines(loyal$SKU[58:114],loyal$Freq[58:114], type="l", col="red")
```

A8. TO FORECAST HIGH OR LOW WEEKLY AVERAGE SALE OF THE STORE.

```
# Logistic regression model to predict the various purchase transaction category between HIGH(>$2.6/purchase transaction) or LOW(<=$2.6/purchase transaction)

traindata<-read.csv("Final_Data_Training.csv")

#Below for loop to find mean/avg price/sale in a particular IRIWeek for Training Dataset
for (j in 1:nrow(traindata)){
  for(i in 592:641) {
    if(i %in% IRIWeek[j]){
      traindata$avg_price_value[j] <- mean(traindata[which(IRIWeek == i),c("Price")])}}}

attach(traindata)

#Adding a column for HIGH(0)/LOW(1) based on avg_price_value of the transaction over an IRIWeek.
for (i in 1:nrow(traindata)){
  if(traindata$avg_price_value[i] <= 2.6){
    traindata$Spending[i] <- "0" }
  else if(traindata$avg_price_value[i] > 2.6){
    traindata$Spending[i] <- "1"
  }}

View(traindata)
attach(traindata)

Spending <- as.factor(Spending)
SKU <- as.factor(SKU)
#Creating a Logistic Regression Model for Spending Category (HIGH/LOW) on AveragePrice and PriceCut(Promotions) over IRIWeek.
glm_mod1 <- glm(Spending ~ AveragePrice+PriceCut,family = binomial);
summary(glm_mod1)
#AIC value is 5586.9 which is most decent in comparison to various cobination of attributes provided in the dataset.
exp(0.91384)
#For every unit increase in PriceCut, the odds of High Spending increase by exp(0.91384)=2.493881.

#Predicting futuristic category of forecast transactions
foredata<-read.csv("Final_Data_Forecast.csv")
attach(foredata)
#Below for loop to find mean/avg price/sale in a particular IRIWeek for Forecast Dataset
for (j in 1:nrow(foredata)){
  for(i in 644:669) {
    if(i %in% IRIWeek[j]){
      foredata$avg_price_value[j] <- mean(foredata[which(IRIWeek == i),c("Price")])}}}

attach(foredata)

#Adding a column for HIGH(0)/LOW(1) based on avg_price_value of the transaction over an IRIWeek.
for (i in 1:nrow(foredata)){
  if(foredata$avg_price_value[i] <= 2.6){
```

```
    foredata$Spending[i] <- "0" }
  else if(foredata$avg_price_value[i] > 2.6){
    foredata$Spending[i] <- "1" }}

attach(foredata)
pred.prob <- predict.glm(glm_mod1,foredata,type="response");
summary(pred.prob)
cut.off <- 0.5;
pred.spending <- (pred.prob > cut.off);
table(pred.spending);
#tablewise classification
table(foredata$Spending,as.numeric(pred.spending))
table(foredata$Spending)
accuracy_rate <- (1193+281)/(1193+364+299+281)
#below variable gives us the accuracy rate for the prediction which 68.975 ~70%
accuracy_rate
```

A9. FORECAST SALES USING MOVING AVERAGE MODEL

```
#Time Series between IRIWeek and Total_Price Sales
    #Time Series - MA model

    #Training Dataset
    install.packages("tseries")
    library(tseries)
    install.packages("forecast")
    library(forecast)
    install.packages("TTR")
    library("TTR")
    traindata<-read.csv("Final_Data_Training.csv")
    attach(traindata)

    #Below for loop to find sum of sale price in a particular IRIWeek for Training Dataset
    for (j in 1:nrow(traindata)){
      for(i in 592:641) {
        if(i %in% IRIWeek[j]){
          traindata$Total_price[j] <- sum(traindata[which(IRIWeek == i),c("Price")])
        }
      }}

    attach(traindata)

    #Filtering the training dataset to have unique sorted rows on IRIWeek and Total_price
    duplicates = duplicated(IRIWeek,Total_price)
    duplicates[1:10]
    unique_traindata <- traindata[!duplicated(traindata[c("IRIWeek","Total_price")]),]
    sorted_traindata <- unique_traindata[order(unique_traindata$IRIWeek),]

    #plotting a time series model on weekly basis starting from year 1991 for training dataset.
    sales.ts<-ts(sorted_traindata$Total_price,frequency = 52, start=c(1991,1))
    plot.ts(sales.ts)
```

```
#The above graph gives the view of total sale throughout the year 1991 for data values  present in training dataset.

# Descriptive statistics and plotting the data
summary(sorted_traindata$Total_price)

# Dickey-Fuller test for variable
adf.test(sorted_traindata$Total_price, alternative="stationary", k=0)
#p-value is 0.01 i.e. H0 is rejected and hence alternate hypothesis holds true. This means the data is stationary.
adf.test(sorted_traindata$Total_price, alternative="explosive", k=0)
#p-value is 0.99 i.e. H0 failed to reject and hence Null hypothesis holds true. This means the data is not explosive.

plot(acf(sorted_traindata$Total_price), main="ACF for Stationary Data") #One Significant autocorrelation lag is there. Which suggests MA(1) model
plot(pacf(sorted_traindata$Total_price), main="PACF for Stationary Data") # nothing is significant
# ACF has significant autocorrelation lag whereas in PACF nothing is significant. Thus MA(1) is recommended.
arima(sorted_traindata$Total_price, order = c(0,0,1))
arima001 <- arima(sorted_traindata$Total_price, order = c(0,0,1))
arimapred1 <- forecast.Arima(arima001, h=10)
arimapred1

#Below graph forecasts the predictive confidence interval for the futuristic value of weekly sales for next year i.e. 1992 data(forecast dataset).
plot.forecast(arimapred1, main = "Forecast using MA(1) Model", xlab="IRIWeeks", ylab="Weekly Sales")
# Since the data provided is only for 1 year, we could not figure out the seasonality nor could we see the trend and hence we can conclude that this model is cyclic. Due to which it becomes difficult to forecast data with higher accuracy.

#Predicting the value graph for total sales on forecast dataset.
foredata<-read.csv("Final_Data_Forecast.csv")
attach(foredata)

#Below for loop to find sum of sale price in a particular IRIWeek for Forecast Dataset
for (j in 1:nrow(foredata)){
  for(i in 644:669) {
    if(i %in% IRIWeek[j]){
      foredata$Total_price[j] <- sum(foredata[which(IRIWeek == i),c("Price")])
    }
  }}

attach(foredata)

duplicates = duplicated(IRIWeek,Total_price)
duplicates[1:10]
unique_foredata <- foredata[!duplicated(foredata[c("IRIWeek","Total_price")]),]
sorted_foredata <- unique_foredata[order(unique_foredata$IRIWeek),]

#Modeling time-series graph for forecast dataset on weekly basis starting with year 1992
```

```
newsales.ts<-ts(sorted_foredata$Total_price,frequency = 52, start=c(1992,1))
plot.ts(newsales.ts)
```

#The above graph gives the view of total sale throughout first half of the year 1992 for data values present in forecast dataset.

#Now, we will try to forecast the same graph based on predict variable(arimapred1) that was created on training dataset above.

```
forecast_ts <- ts(plot.forecast(arimapred1, main = "Forecast using MA(1) Model", xlab="IRIWeeks", ylab="Weekly Sales"))
```

#Below plot draws both the graphs i.e. predicted graph for 1992 and measured graph for 1991 side by side.

```
ts.plot(sales.ts ,newsales.ts, parallel=TRUE, gpars = list(col = c("black","red")))
```

#We observe that the forecast/predictive graph lies within the confidence interval of 95% as predicted on the training dataset.

A10. OTHER VISUALIZATION

1. Sale of Product according to Size –

```
#----------Sale of product according to size----------
with(finaldata, table(finaldata$Brand,finaldata$Size))
mydata<-read.csv("Final_Data_Calibration_Training.csv", header = TRUE)
mydata_size<-table(mydata$Brand,mydata$Size)
table(mydata_size)
barplot(mydata_size, legend = rownames(mydata_size), pch = c(1,10), ylim=c(0,2500),col = c("brown", "blue", "green", "red", "yellow","purple", "pink", "orange", "grey", "light blue"), xlab = "Size" , ylab = "Quantity", main = "Brand-Size-Quantity Depiction")
args.legend = list(title = "SES", x = "topright", cex = .7)
barplot(mydata_size, legend = rownames(mydata_size), args.legend = list(title = "BRANDS", x = "topright"), ylim=c(0,2500),col = c("brown", "blue", "green", "red", "yellow","purple", "pink", "orange", "grey", "light blue"), xlab = "Size" , ylab = "Quantity", main = "Brand-Size-Quantity Depiction")
```

2. Number of SKUs per Brand –

```
#----------No of SKUs per Brand----------
testdata_SKU<-as.data.frame.matrix(table(testdata$Brand,testdata$SKU))
barplot(apply(testdata_SKU,1,sum),xlab="Brand",ylab="Quantity of SKUs", main = "Number of SKUs per Brand for Calibration dataset", col="red")
```

3. Sale of products according to Formula:

```
table1<-table(Brand,Formula2)
barplot(table1, pch = c(1,10), ylim=c(0,3500),col = c("cyan", "blue", "green", "red", "yellow","purple", "pink", "orange", "grey", "light blue"), xlab = "Formula" , ylab = "Quantity", main = "Brand-Formula-Quantity Depiction")
legend("topright", legend = row.names(table1), fill = 1:6, ncol = 2, cex = 0.75)
```

A11. ORDINAL LOGISTIC REGRESSION

 # To predict the sale(HIGH/MEDIUM/LOW) for the ordinal values using threshold of $2.4, $2.8 per transaction during IRIWeeks(592-641) as the sale target. Based on the same logic, we tried to predict the forecasted value on forecast data(IRIWeeks 644-669).

```
detach(finaldata)
rm(finaldata)
setwd("E:\\MBA\\GMAT\\SKM_MS-MIS_Docs\\USF\\MIS\\SDM\\Final Project")
finaldata<-read.csv("Final_Data_Calibration_Training.csv")
attach(finaldata)

#Below for loop for Training Dataset
for (j in 1:nrow(finaldata)){
 for(i in 592:641) {
  if(i %in% IRIWeek[j]){
   finaldata$avg_price_value[j] <- mean(finaldata[which(IRIWeek == i),c("Price")])}}}}

#Below for loop for Validation Dataset
for (j in 1:nrow(finaldata)){
 for(i in 642:643) {
  if(i %in% IRIWeek[j]){
   finaldata$avg_price_value[j] <- mean(finaldata[which(IRIWeek == i),c("Price")])}}}}

#Below for loop for Forecast Dataset
for (j in 1:nrow(finaldata)){
 for(i in 644:669) {
  if(i %in% IRIWeek[j]){
   finaldata$avg_price_value[j] <- mean(finaldata[which(IRIWeek == i),c("Price")])}}}}

attach(finaldata)
hist(finaldata$avg_price_value, main = "Weekly Spending Graph Between IRIWeeks 592-641", xlab
= "Average Weekly Spent", ylab = "Frequency of target sale", col = "Red", border = "Black")
summary(finaldata$avg_price_value)

for (i in 1:nrow(finaldata)){
 if(finaldata$avg_price_value[i] <= 2.4){
  finaldata$Spending[i] <- "Low" }
 if(finaldata$avg_price_value[i] > 2.4 && finaldata$avg_price_value[i] <= 2.8){
  finaldata$Spending[i] <- "Medium" }
 if(finaldata$avg_price_value[i] > 2.8 ){
  { finaldata$Spending[i] <- "High" }}}
```

```
#Ordinal Logistic Regression
#m <- polr(Spending ~ IRIWeek, data = finaldata, Hess=TRUE)
      #summary(m)

      install.packages("rms")
      library(rms)

Y <- cbind(Spending)
      X <- cbind(IRIWeek,Brand)
      Xvar <- c("IRIWeek","Brand")

      # Descriptive statistics
      summary(Y)
      summary(X)
      table(Y)

      # Ordered logit model coefficients
      ddist<- datadist(Xvar)
      options(datadist='ddist')

      ologit<- lrm(Y ~ X, data = finaldata)
      print(ologit)

      # Ordered logit model odds ratio
      #summary(ologit)

      # Ordered logit predicted probabilities
      # xmeans <- colMeans(X)
      # newdata1 <- data.frame(t(xmeans))
      fitted <- predict(ologit, newdata=finaldata, type="fitted.ind")
      colMeans(fitted)

      finaldata <- cbind(finaldata,predict(ologit, newdata=finaldata, type="fitted.ind"))
      names(finaldata)

      write.csv(finaldata,"Final_Data_Weekly_Forecast_Spending_Predicted.csv")
      ------------------------------------------------------------------------------
       #Alternate for Ordered Logistic model
       library(MASS)
```

```
m <- polr(Spending ~ Brand+IRIWeek+SKU+HHId+Price+PriceCut+AveragePrice, data = finaldata, Hess=TRUE)
summary(m)
predict(m, finaldata, type = "probs")
exp(cbind(OR = coef(m),confint(m)))

rm(newdat)
newdat <- read.csv("Final_Data_Weekly_Spending_Forecast.csv", header = T)
newdat <- read.csv("Final_Data_Weekly_Spending_Validation.csv", header = T)
newdat <- cbind(newdat, predict(m, newdat, type = "probs"))
```