Fabric Softener Data Analysis

Prashant Bhowmik

ISM6137: STATISTICAL DATA MINING |

Contents

Introduction

The SKU is an identification number of a particular product and tells about different attributes of that product. These attributes could include, but are not limited to, manufacturer, description, material, size, color, features, and formula. An SKU is not the same as a product model number from a manufacturer, although the model number could form all or part of the SKU. The SKU is established by the merchant and is used for tracking inventory.

Most Consumer choice models in marketing make use of 'Brand' as the fundamental unit of analysis, however many of the decisions made by the consumers, manufacturers, and retailers occur at the level of stock keeping unit. Hence it is important to consider the SKU attributes to better understand consumer trends.

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

Build a consumer choice model among SKU's, using "Fabric Softener" dataset.

Overview of Dataset

The data for fabric softener was distributed in 5 files:

1) D1PUR.DAT: This file contains the household purchase history data. It contains household ID for purchase and trip information. The trip info variable is in AAABBBCCC format where
   AAA = IRI week
   BBB = store#
   CCC = SKU# purchased.
2) MERCH.DAT: This file contains store environment information. The data is divided in 5 fields i.e. SKU, Store ID, IRIweek, price paid and merchandizing info.
   The format of the merchandising variable is AAABCD where
   AAA = regular price
   B = ignore
   C = display
   D = feature
3) ARSP.DAT: Contains the average regular selling price of each SKU in each store. It contains SKU, store ID, and ARSP- average regular selling price of each SKU.
4) BRSINFO.DAT: Contains the attribute information for each SKU. It contains SKU, SKU description Brand, Size, formula and other coded info for SKU attributes.
5) IRIweek.xls: This file contains the week of purchase recorded as IRI week. The measure used for IRI where week 1 corresponds to the week ending on 09/09/79. The file contains a mapping with week number and Week ending date.

Data Cleaning & Preparation

1) The data provided in the .dat files were encoded so we had to clean and prepare the data for our purpose of running the model. We used the script provided by Prof. Daniel to perform data cleaning. We also created dome dummy variables.

   The script with explanation is provided in **Appendix 1**.

2) Dummy variables created from the MERCH.dat file, where merchandising variable was in format AAABCD:
   price_cut = AAA - price_paid (if the result is < 0, price_cut = 0)
   Display = 1 if C >= 1; 0 otherwise
   Feature = 1 if D >= 1; 0 otherwise

3) 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 had 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))
   Wherever 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

4) Data was further split into training, calibration and forecast weeks.
   - Training week from IRIWeek 592 to IRIWeek 641
   - Calibration week from IRIWeek 642 to IRIWeek 643
   - Forecast weeks from IRIWeek 644 to IRIWeek 669

   To perform the forecasting of brand preference by customer, we removed the BRAND column from final_data_forecast sheet.

Visualization

1) Total number of products of each brand with different sizes. As we can see from the below graph that medium size sells the most, whereas XL sells least.
   Also we can observe that for brand DWN small size products are selling more whereas for brand SNG medium size products sell more.

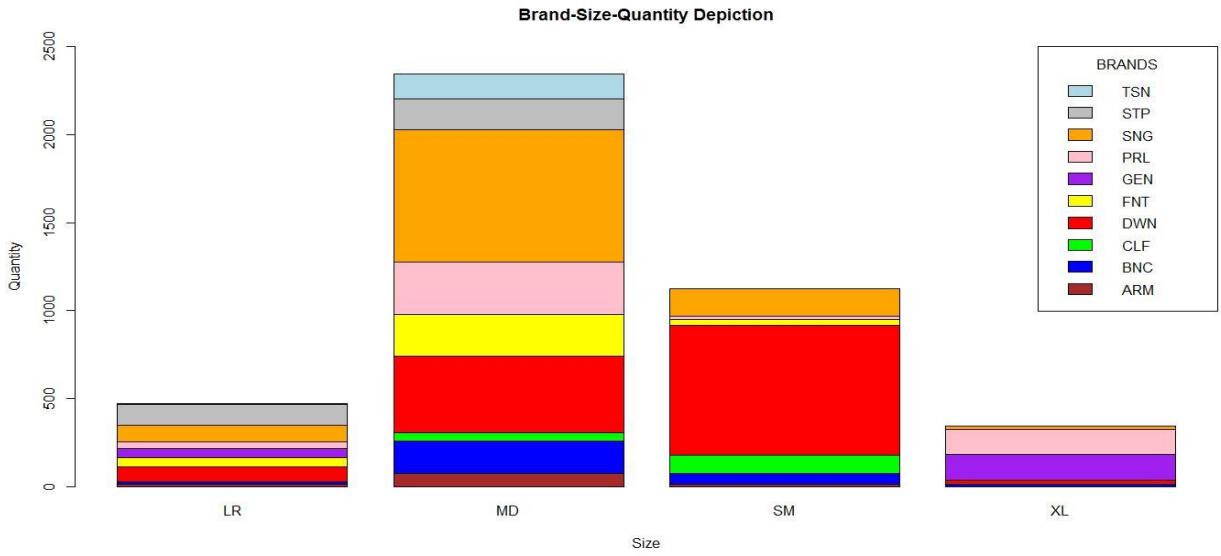
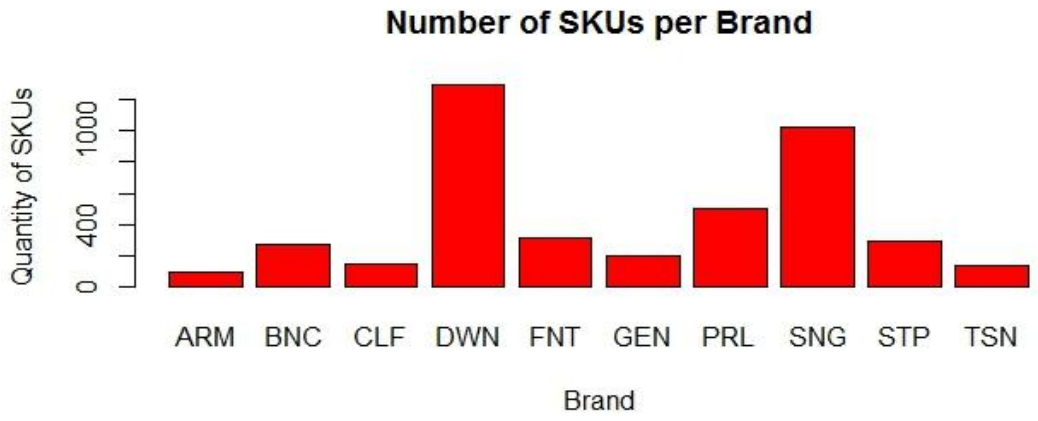


Fig. 1

2) The below figure shows the total number of SKU per brand. This can be interpreted as the number of items sold by a brand. We can observe that DWN and SNG are the most selling brands, whereas ARM is the least selling brand.



The below table shows the number of items sold under each brand.

```
> table(Brand)
Brand
 ARM  BNC  CLF  DWN  FNT  GEN  PRL  SNG  STP  TSN
  97  281  160 1326  322  202  522 1067  297  143
>
```

3) The below line graph explains the number of customer buying fabric softener per week. We can observe that week number 607 and 635 has high number of customer visits. Looking at the IRIweek data we find that these weeks were in the month of April and November respectively.

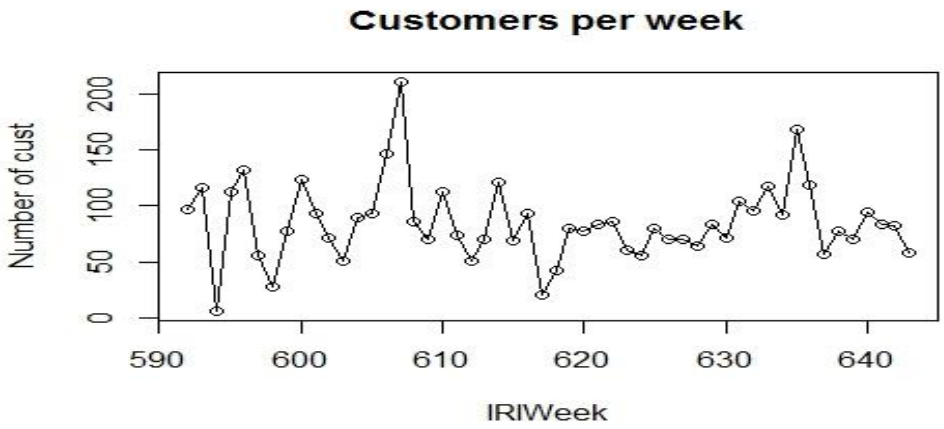


Fig 3.0

4) The Fig 4.0 is the bar plot for the formula used by different Fabric softner Brands. The X-axis represents the Formula and Y-axix the the product quantities using that formula. We can observe that maxumum number of brands use the formula RG. From the data It can also be interpreted as more customers prefer products with formula RG.

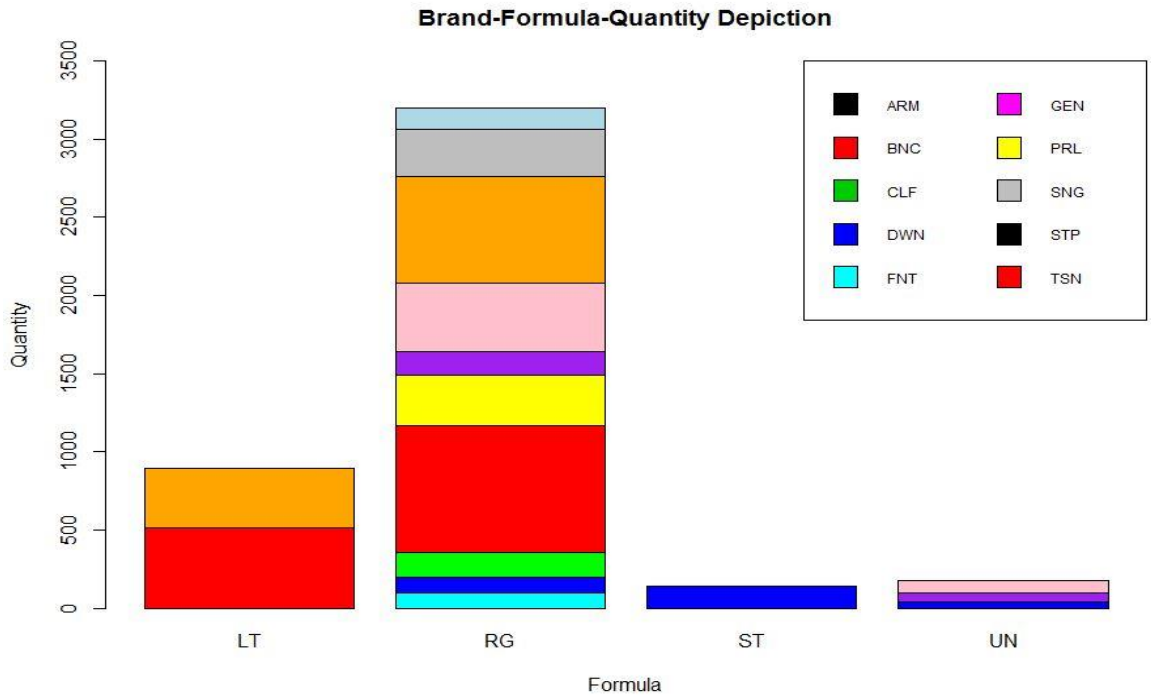


Fig 4.0

5) The below graph shows the weekly sales of the fabric softener. We can observe that week 16 has the maximum sales i.e. $550 approx. Also the average weekly sale hover around $200.

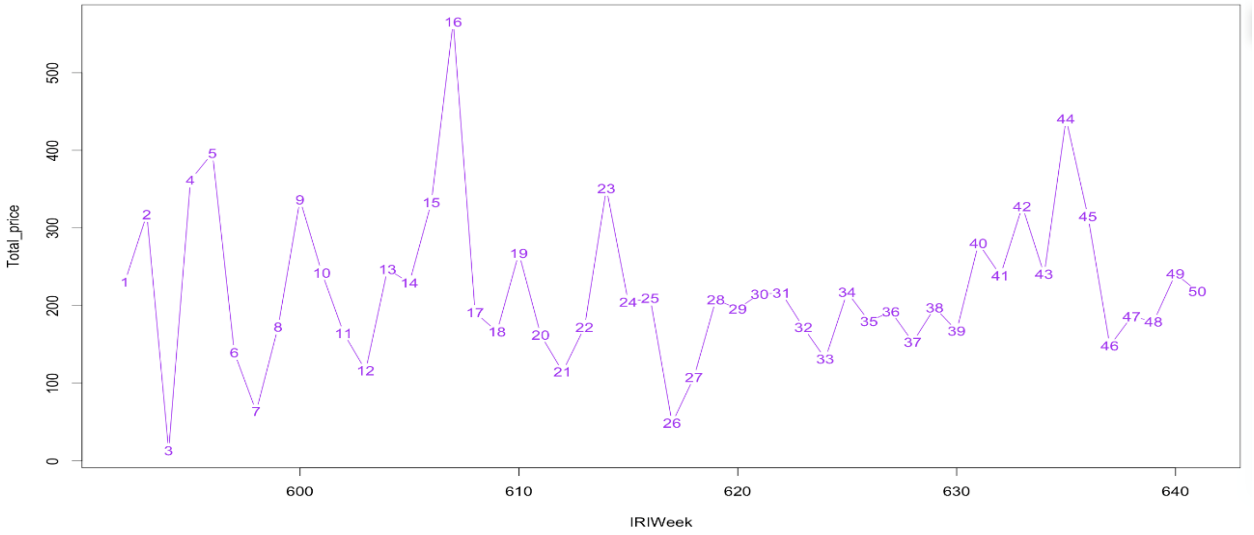


Fig 5.0

Data Analysis

Brand Analysis

We performed the brand analysis using multinomial logistic regression as we had categorical data for the brand. For R-Script refer **Appendix 2**

Most Selling brand – DWN

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.

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

```
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 ***
```

Worst selling brand – 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.

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

```
Coefficients :
                Estimate Std. Error t-value  Pr(>|t|)
BNC:(intercept)  1.06251    0.11964  8.8805 < 2.2e-16 ***
CLF:(intercept)  0.48714    0.13105  3.7172 0.0002014 ***
DWN:(intercept)  2.61988    0.10683 24.5236 < 2.2e-16 ***
FNT:(intercept)  1.20610    0.11757 10.2584 < 2.2e-16 ***
GEN:(intercept)  0.73991    0.12536  5.9024 3.582e-09 ***
PRL:(intercept)  1.67730    0.11237 14.9269 < 2.2e-16 ***
SNG:(intercept)  2.38230    0.10780 22.0995 < 2.2e-16 ***
STP:(intercept)  1.14028    0.11849  9.6235 < 2.2e-16 ***
TSN:(intercept)  0.40547    0.13316  3.0450 0.0023265 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Cheapest Brand – CLF

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.

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

Most Expensive Brand - SNG

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.

```
ARM:Price      -1.845334   0.149587 -12.3362 < 2.2e-16 ***
BNC:Price      -0.535846   0.091453  -5.8592 4.650e-09 ***
CLF:Price      -4.220301   0.245402 -17.1975 < 2.2e-16 ***
DWN:Price      -0.319514   0.055363  -5.7712 7.869e-09 ***
FNT:Price      -1.171688   0.088692 -13.2108 < 2.2e-16 ***
GEN:Price      -3.729167   0.186628 -19.9818 < 2.2e-16 ***
PRL:Price      -2.130862   0.086866 -24.5303 < 2.2e-16 ***
STP:Price      -0.672428   0.089257  -7.5336 4.929e-14 ***
TSN:Price      -2.072195   0.131301 -15.7820 < 2.2e-16 ***
```

Forecasting customer preferences based on the SKU.

To forecast customer preference of brands for the Fabric Softener, we build multinomial Logistic regression models. There were 3 functions available in R to perform this task i.e. Mlogit, VGLM with family=multinomial and Multinorm function.

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

Models:

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

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% accuracy.

To confirming the above model, we ran the prediction model on forecast dataset and could identify which 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

Results:

- From VGLM model, we could conclude that SKU is a significant independent variable to predict BRAND and IRIWeek and HHId are insignificant variables. **Appendix 7** explains the R-Code performed to reach the conclusion.
- Using Multinorm function, we could predict the BRAND bought by customer with 100% accuracy. For R code refer **Appendix 3**.

Analysing dependency on SKU attributes

```
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
```

A linear regression model was build keeping SKU as the dependent variable and other variables as independent. Starting with Kitchen Sink model, we gradually started removing variables which weren’t explaining much about the attributes.

We concluded that SKU can be explaned 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 4 explains the complete R-Script on this operation.

Analyzing dependency of Price on manufacturer variables

```
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
```

A linear regression mode build 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 manufacturing variable.

Refer **Appendix 5** for R-Script.

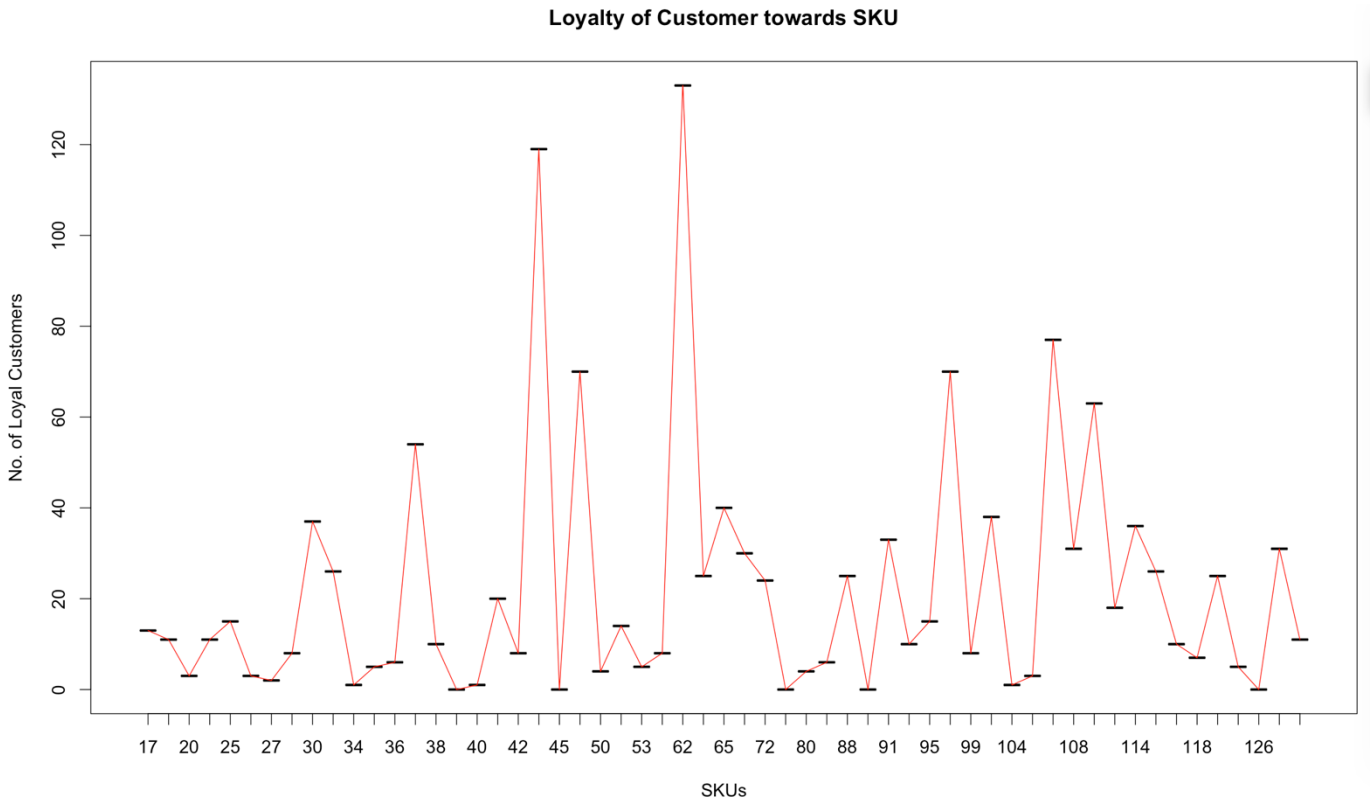
Analyzing Loyalty of customer towards brand

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

Data Preparation: To perform this activity, we performed few operations in MS Excel using VLOOKUP formula and then did analysis.

Analysis: Below line plot explains the number of loyal customers for each SKU. We can observe that 2 SKU's have the maximum number of loyal customers.

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



Appendix 6 contains the complete R-Script.

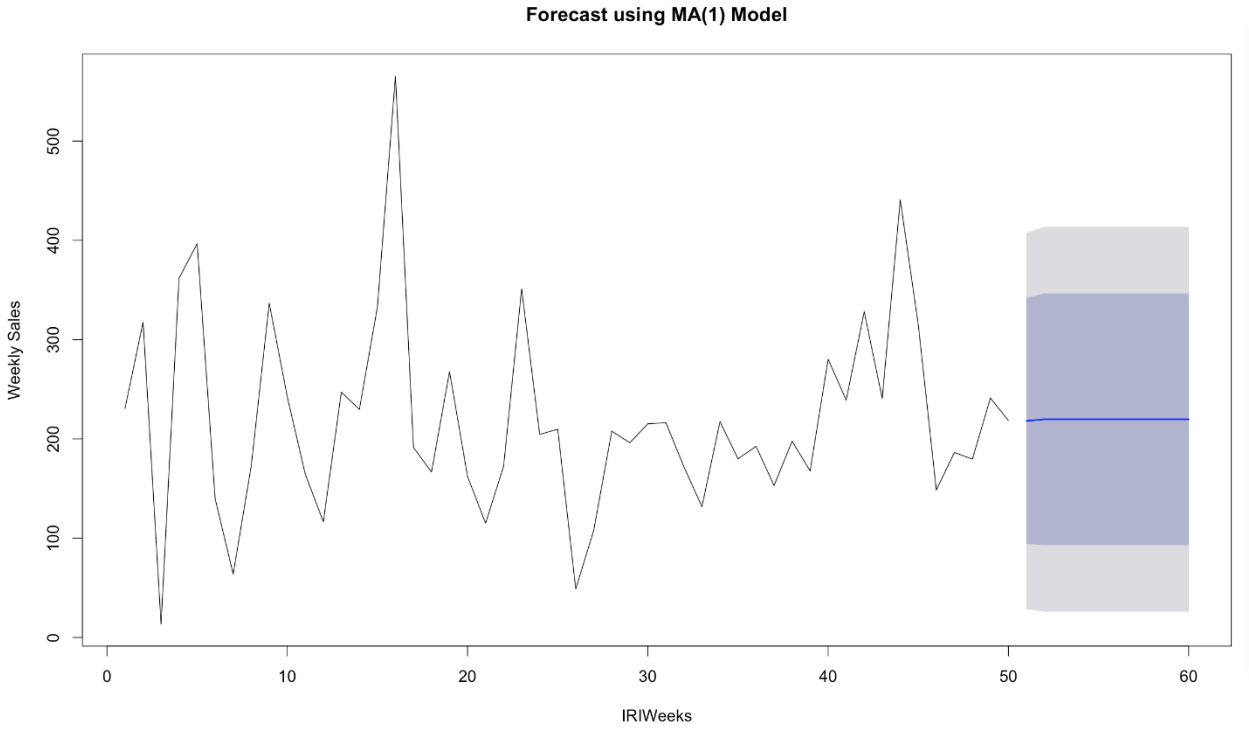
Forecast Sales using moving average (1) model

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.

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**.

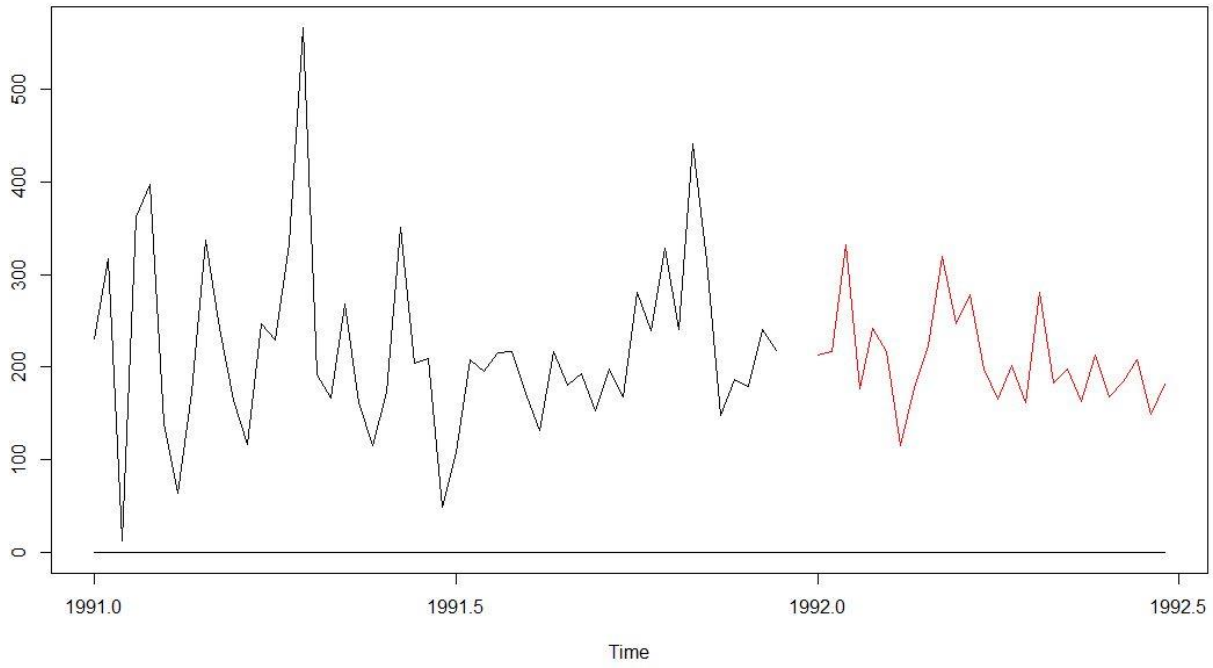
Autocorrelation function (ACF) gave us a single significant point and partial auto-correlation function (PACF) has nothing significant. This help us conclude that **Moving Average** (1) needs to be executed.

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. The X-axis is marked on half yearly basis.



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

Appendix 11 contains the detailed R-Script with explanation for forecasting sales using MA Model.

Conclusion

We could conclude by saying that although BRAND was the most important parameter in our model for predicting customer behavior, but the SKU attributes could not be simply ignored as they give more details of customer behavior when included in the model. From the Fabric Softener data we could say that BRAND, SIZE, FORMULA and FORM were significant attributes during modeling.

Appendixes

Appendix 1:

```
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")
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"))
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")
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","Feat
ure")]
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","Featu
re")]
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)
```

Appendix 2:

```
#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))
```

Appendix 3:

#using multinorm function - Forecast the BRAND preferred 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.
```

```
> 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
```

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

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

```
> 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
```

Appendix 4:

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.

Appendix 5:

#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)

Appendix 6:

analyzing Loyalty of customers based on Brand

#Loyalty Check Script and Number of SKUs per Brand

1) 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.
2) 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.
3) I also copied and pasted all SKU column from Test dataset into Column 'AL' of this combined sheet.
4) 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.
5) 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")
```

Appendix 7:

#VGLM Multinomial Logistic Regression Model – To check for the significance of various predictor variables.

```
install.packages("VGAM")
library(VGAM)
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+HHId, 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.

Appendix 8:

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----------

mydata_SKU<-as.data.frame.matrix(table(mydata$Brand,mydata$SKU))
barplot(apply(mydata_SKU,1,sum),xlab="Brand",ylab="Quantity of SKUs", main = "Number of SKUs per Brand", 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)
```

4. **Number of SKUs Per Brand for the calibration dataset**

```
calibdata_SKU<-as.data.frame.matrix(table(calibdata$Brand,calibdata$SKU))
barplot(apply(calibdata_SKU,1,sum),xlab="Brand",ylab="Quantity of SKUs", main = "Number of SKUs per Brand for Calibration dataset", col="red")
```

Appendix 9:

Logistic regression – to predict the sale(HIGH/LOW) for the ordinal values using threshold of $2.6 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) with an accuracy of 68.97.

```
class(SKU)
#######logistic model data#######
#for loop for checking logistic model on weekly spent
for (i in 1:nrow(finaldata)){
  if(finaldata$avg_price_value[i] <= 2.6){
    finaldata$Spending[i] <- "0" }
  else if(finaldata$avg_price_value[i] > 2.6){
    finaldata$Spending[i] <- "1"
  }


SKU <- as.factor(SKU)
mod1 <- glm(Spending ~ AveragePrice+PriceCut,family = binomial);
summary(mod1);
exp(0.91384)
#For every unit increase in PriceCut,The odds of High Spending increase by
exp(0.91384)=2.493881.


test.data <- read.csv("Spam-Test.csv");
pred.prob <- predict.glm(mod1,finaldata,type="response");
summary(pred.prob)
```

```
cut.off <- 0.5;
pred.spending <- (pred.prob > cut.off);
table(pred.spending);
#tablewise classification
table(finaldata$Spending,as.numeric(pred.spending))
table(finaldata$Spending)
(1193+281)/(1193+364+299+281)

finaldata <- cbind(finaldata,predict.glm(mod1,finaldata,type="response"))
names(finaldata)
##################################################
```

Appendix 10:

Linear Regression Model – evaluations of the importance of pricing and promotions

```
detach(finaldata)
setwd("E:\\MBA\\GMAT\\SKM_MS-MIS_Docs\\USF\\MIS\\SDM\\Final Project")
finaldata<-read.csv("Final_Data_Calibration_Training.csv")
finaldata<-read.csv("finalized_data_Latest.csv")
dim(finaldata)
names(finaldata);
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$Total_price[j] <- sum(finaldata[which(IRIWeek == i),c("Price")])
    }
  }}

#evaluations of the importance of pricing and promotions
attach(finaldata)
names(finaldata)
hist(Total_price)
Lmod1 <- lm(IRIWeek~ Brand*Price+log(SKU)+Form+Total_price+Size*Feature+Display)
summary(Lmod1)

plot(IRIWeek~Total_price)

Lmod2 <- lm(Price~Brand+SKU+Size)
summary(Lmod2)

Lmod3 <- lm(PriceCut~Brand+Size+Display)
```

```
summary(Lmod3)

#evaluations of the importance of the different attributes for each SKU (to this aim, you may have to code the attributes with an appropriate set of dummies).
detach(finaldata)
rm(finaldata)

finaldata<-read.csv("finalized_data_Latest.csv")
dim(finaldata)
names(finaldata);
attach(finaldata)

Lmod4 <- lm(SKU~Formula2+Form+Size+Brand)
summary(Lmod4)
#R2 is 99.81 % means SKU is combination of these features and there is no interaction among them.
```

Appendix 11:

Time-Series–

```
#Time Series between IRIWeek and Total_Price Sales

 ## TIME SERIES CODE

 mydata<-read.csv("Final_Data_Calibration_Training.csv", header = TRUE)
attach(mydata)
for(j in 1:nrow(mydata)){
 for(i in 592:641) {
  if(i %in% IRIWeek[j]){
   mydata$Total_price[j] <- sum(mydata[which(IRIWeek == i),c("Price")])
  }
 }}

writedata <- mydata[,c("IRIWeek","Total_price")]

duplicates = duplicated(IRIWeek,Total_price)
duplicates[1:10]
unique_writedata <- writedata[!duplicated(writedata[c("IRIWeek","Total_price")]),]
sorted_writedata <- unique_writedata[order(unique_writedata$IRIWeek),]

write.csv(sorted_writedata,"test.csv")
ts<-read.csv("test.csv", header=TRUE)
attach(ts)
```

```
Time_Model<- ts(ts, frequency=52)
Time_Model
plot.ts(Time_Model,col="purple")

data <- read.csv("Enrollments.csv");attach(data);names(data);
IRIWeek    <- IRIWeek[2:49];
IRIWeek.l1 <- IRIWeek[1:48];
Total_price    <- Total_price[2:49];
View(ts)
mod3 <- lm(IRIWeek~IRIWeek.l1);
summary(mod3);


#plotting smoothing splines with different smoothing parameters
##note: the parameter "spar" sets the smoothness
par(mfrow=c(1,1))
plot(Total_price, IRIWeek, xlab="Total_price", ylab="IRIWeek",lwd=2);
sm1 <- smooth.spline(Total_price, IRIWeek, spar=0.2);
sm2 <- smooth.spline(Total_price, IRIWeek, spar=1);
x.pred <- seq(0,25000);
sm1.pred <- predict(sm1,x.pred)$y;
sm2.pred <- predict(sm2,x.pred)$y;
lines(x.pred,sm1.pred, lwd=2,lty=2,col="red")
lines(x.pred,sm2.pred, lwd=2,lty=2,col="blue")
-----------------------------------------------------------------------------------------------------
#-------------------------------------------------------------------------
#Time Series - MA model

#Training Dataset
detach(finaldata)
rm(finaldata)
traindata<-read.csv("Final_Data_Calibration_Training.csv")
dim(traindata)
names(traindata);
attach(traindata)

#Below for loop 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)

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),]

install.packages("tseries")
library(tseries)
library(xts)
install.packages("forecast")
library(forecast)
install.packages("TTR")
library("TTR")
library(ggplot2)
sales.ts<-ts(sorted_traindata$Total_price,frequency = 52, start=c(1991,1))
plot.ts(sales.ts)

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

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

plot(acf(sorted_traindata$Total_price), main="ACF for Stationary Data") #One Significant partial correlation is there. Which suggests MA(1) model
plot(pacf(sorted_traindata$Total_price), main="PACF for Stationary Data") # nothing is significant

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

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.

#working Code and file below
detach(foredata)
rm(foredata)

foredata<-read.csv("Final_Data_Forecast_Backup_Sorted.csv")
dim(foredata)
names(foredata);
attach(foredata)

#Below for loop 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),]


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

forecast_ts <- ts(plot.forecast(arimapred1, main = "Forecast using MA(1) Model", xlab="IRIWeeks", ylab="Weekly Sales"))
ts.plot(sales.ts ,newsales.ts, parallel=TRUE, gpars = list(col = c("black","red")))
```