Fabric Softener Analysis

Prepared By –
Sachin Kant Misra
UID#81642960

Submitted to –
Prof. Daniel Zantedeschi
ISM6137: ISDS Dept.
University of South Florida

Contents

Introduction 4

Problem Statement 4

Data Set Overview 4

Data Mining and Cleaning 6

Data Analysis and Modeling Conclusions 7

1). Multinomial Logistic Regression Model using (a) VGLM (b) MLOGIT and (c) MULTINOM functions to Forecast the brand customer has bought based on the SKUs 7

(A). VGLM Function 7

(B). MLOGIT Function 7

(C). MULTINOM Function 8

2). Time Series Model and ARIMA 8

3). Ordinal Logistic Regression Model using OLOGIT function for HIGH/MEDIUM/LOW Spending 9

4). Binomial Logistic Regression Model using LOGIT function with family=binomial 9

5). Loyalty Check for Customers 9

6). Linear Regression Model 10

(A). SKUs to check its dependent variables 10

(B). Price and its dependent variables 10

Visualization Graphs 11

Visualization through Tableau 15

Dashboard 17

Story Line 18

Appendix (R-Script) 20

Appendix 0: Data Cleaning Script: 20

Appendix 1: Mlogit regression 21

Appendix 2: vglm regression 22

Appendix 3: multinorm regression 22

Appendix 4: Ologit regression 24

Appendix 5: Logistic regression 26

Appendix 6: Evaluations of the importance of pricing/sale for IRIWeek 27

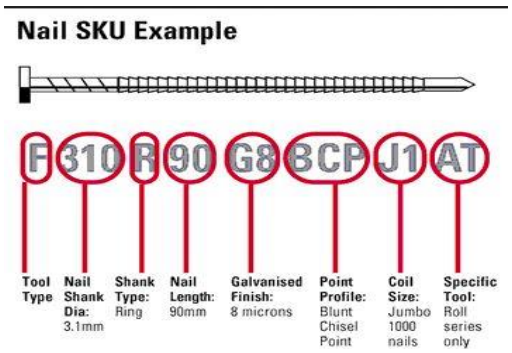
Appendix 7: Time-Series regressions – ARIMA Model 28

Appendix 8: Sale of Product according to Size and according to Formula 30

Appendix 9: Loyalty Check for SKU of the Brand 31

Introduction

The project "Fabric Softener" comprises of different set of Attributes pertaining to Stock Keeping Unit (SKU). The SKU is a unique code that depicts various inventory blocks within a warehouse. SKU characterizes various elements like Brand, Size, Form, Formula, Model, etc. The image here illustrates how SKU signifies various Brands and its attributes.



Retailer Stores like Walmart, BestBuy, and Walgreens keep track of various products through their SKU codes. These are widely used in operation and supply chain management when it comes to a demand for the product by customer.

The motive behind analysis is to create Consumer Choice Model which is strongly dependent upon SKUs and their Brands. Although mostly Brand and Promotions help the prediction for sale but lately it has been observed that feature related to Brand products eg. Size, Color, Form, Formula, Feature, etc. closely follow similar model graph for sale. Many of the fortune 500 companies have started including these attributes in order to prepare their consumer choice models for better forecast sales of their new product being launched into the market.

Problem Statement

To prepare Consumer Choice Model among SKUs and their Brands.

Data Set Overview

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 (4417 purchases-Year 1991, 2137 purchases-Year 1992) purchases over a span of 1.5 years.

CALIBRATION DATASET -

- 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 **Validation** of the data.

FORECAST DATASET -

- The **2137** purchases from IRIWeek 644 to IRIWeek 669 are used for **Forecast** data.

Data was segregated in different files which was cleaned using R-script described later to constitute in a single file named "Finalized_Data.csv". Following is the information about different file and their data:

1. **D1PUR.DAT** – contains household purchase history and has two columns
    a. Column1: HHId – House Hold Identification Number
    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 we use dummies for feature and display:
        DISP = 1 if C >= 1; 0 otherwise
        FEAT = 1 if D >= 1; 0 otherwise
    f. We also separate price into depromoted (regular_price) and price_cut Components:
        i. regular_price = AAA
        ii. price_cut = AAA - price_paid (if the result is < 0, price_cut = 0)
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:Formula2#
    k. Column11:Size#
5. **IRIweek.xls**: This file contains the week of purchase recorded as IRI week. It is a measure of IRIWeek where week 1 corresponds to the week ending on 09/09/79. It maps IRIWeek number with the Week ending date.
6. **Membership Data Panel**: This file contains combined data from IRIWeek.xls and BRSINFO.DAT

Data Mining and Cleaning

R-Script in Appendix 0

Steps to create a final data file that contains all relevant fields from different files given -

Step 1) First part was executed to segregate column values using substring function from D1PUR.DAT file column2 and given appropriate heading as described above.
Step 2) Second part was executed to segregate column values using substring function from MERCH.DAT file based on coding described above point 2. Part (e) and (f).i.e. the for loop segregates Price, Display and Feature attributes for each purchase row in merchandising data file.
Step 3) Third part merges purchase data created from D1PUR.DAT and MERCH.DAT files on column fields IRIWeek, Store and SKU
Step 4) Fourth part reads the combined file Membership panel data created from IRIWeek.xlsx and BRSINFO.DAT
Step 5) Fifth part reads from ARSP.DAT file and merges it with above created attrplusmerch file on SKU and Store Columns.
Step 6) Sixth part adds another column into finaldata frame i.e. PriceCut that contains discount price calculated by subtracting PricePaid with RegPrice column field and writes the finalized_data.csv file for data analysis.
Step 7) 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 'finalized_data.csv' with brand data as a backup.

NOTE: The finalized file created was then modified to contain Brands as categorical field with names in it. This was done by using Excel formula: **"=INDEX(D$1:M$1,MATCH(MAX(D2:M2),D2:M2,0))"** and the data created after this looked like :

AE1 | Brand

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE
1	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	Brand
2	9436	103	592	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	1.29	0	1.182	0	2	PRL
3	9571	103	631	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0.99	0	1.182	0	2	PRL
4	9584	103	631	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0.99	0	1.182	0	2	PRL
5	9376	103	631	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0.99	0	1.182	0	2	PRL
6	9451	103	621	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	1	0	1.182	0	1	PRL
7	9676	103	644	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0.99	0	1.182	0	1	PRL
8	9595	103	622	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	1	0	1.182	0	0	PRL
9	9648	103	622	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	1	0	1.182	0	0	PRL
10	9738	103	665	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0.99	0	1.182	0	0	PRL

Data Analysis and Modeling Conclusions

1). Multinomial Logistic Regression Model using (a) VGLM (b) MLOGIT and (c) MULTINOM functions to Forecast the brand customer has bought based on the SKUs

In this model, we started regression analysis with **Training data** that looked like this:

A1 fx Brand

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	Brand	IRIWeek	HHId	SKU	Form	Formula2	Size	Price	PriceCut	AveragePrice	Display	Feature	B.ARM	B.BNC	B.CLF	B.DWN	B.FNT	B.GEN	B.PRL	B.SNG	B.STP	B.TSN
2	PRL	592	9436	103	S	UN	MD	1.29	0	1.182	0	2	0	0	0	0	0	0	1	0	0	0
3	PRL	631	9571	103	S	UN	MD	0.99	0	1.182	0	2	0	0	0	0	0	0	1	0	0	0
4	PRL	631	9584	103	S	UN	MD	0.99	0	1.182	0	2	0	0	0	0	0	0	1	0	0	0
5	PRL	631	9376	103	S	UN	MD	0.99	0	1.182	0	2	0	0	0	0	0	0	1	0	0	0

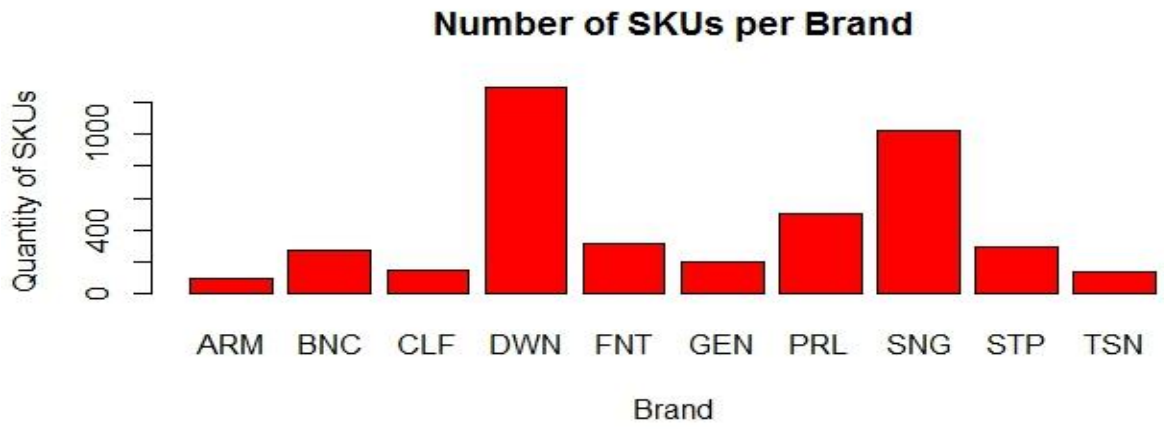
(A). VGLM Function is used to provide significance values of the various predictors through p-value which otherwise aren't provided through other multinomial regression functions. **Check Appendix 1.**

- Based on the summary of the coefficients, we concluded that **Only Brand and SKUs are strong covariates**. Other variables like HHId and IRIWeek don't have much significance in the data variation.
- Price and Brands are highly correlated. Because of this analysis we concluded that we have better chance at creating multinomial regression model on "Brand Vs SKUs" or "Brand vs Price" than other for predictive analysis.

```
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 ***
```

(B). MLOGIT Function is used to predict various Brands and their importance. The model gives a summary of results based on a reference level field as provided in the "reflevel" parameter. Based on the various models executed as shown in **Appendix 2** following are the conclusions:

- **Best selling brand - DWN** - All coefficients of the brands are negative (exp value <1) i.e. Every other brand has lesser value than DWN brand: **(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): **exp(coef(mlogit.model1))**
- **Worst selling brand - ARM** - All coefficients of the brands are positive (exp value >1) i.e. Every other brand has more value than ARM brand: **(mlogit.model2)**
- **SNG is the most expensive brand in terms of Price** since the coefficient of all other brand price is negative (exp value < 1) in reference to SNG: **(mlogit.model3)**
- **CLF is the cheapest brand in terms of Price** since the coefficient of all other brand price is positive (exp value >1) in reference to CLF: (**mlogit.model4**)



(C). MULTINOM Function is used to create a model on **Training Dataset**. It is then validated for the predicted Brand Names on probability values over **Validation Dataset** and then finally predicts the probable Brand names **based on SKUs** for each purchase on **forecast dataset**. **Check Appendix 3.**

- The Brand name "PRL" had 350 purchases in forecast dataset which was accurately predicted with probability value >0.9 by forecast on training data model.
- Similarly, the Brand name "DWN" had 452 purchases in forecast dataset which was accurately predicted with probability value >0.9 by forecast on training data model.
- **MULTINOM model was ~100% accuracy rate with SKUs only as dependent variable and Brand as independent.**

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

2). Time Series Model and ARIMA

In this model, Training dataset is used to create time series model using **function ts** between IRIWeeks and Total_Price (Sale). **Check Appendix 7.**

- The model is further improvised to ARIMA MA(1) model as the data was found stationary using Dickey-Fuller test where **acf function** was found to have partial correlation rejecting Null hypothesis while **pacf function** was insignificant failing to reject Null hypothesis. This confirmed that **the data was stationary and Moving Average (MA) model can be used to predict forecast data.**
- The data model is further verified for **forecast dataset** Total_Price(Sale) value and the resulting graph is found to lie exactly between the predicted graph values for the 95% confidence Interval.

3). Ordinal Logistic Regression Model using OLOGIT function for HIGH/MEDIUM/LOW Spending

In this model **(Check Appendix 4)**, we have filtered the datasets to find average_price (Sale) for every IRIWeek and then categorized the purchase transactions 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 forecast category of the **future transactions on forecast data** (IRIWeeks 644-669).

4). Binomial Logistic Regression Model using LOGIT function with family=binomial

In this model**(Check Appendix 5)**, we did similar filteration as mentioned above for Ordinal Logistic model with only two category of data i..e HIGH/LOW. This model creates the Total_Price(sale) category for the ordinal values using threshold of $2.6 per transaction during IRIWeeks(592-641) on training dataset. Based on the same model, I tried to predict the forecasted value on forecast data(IRIWeeks 644-669) where I got prediction with an accuracy of 68.97.

```
> 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
```

Data View:

SKU	Form	Formula2	Size	Price	PriceCut	AveragePrice	Display	Feature	B.ARM	B.BNC	B.CLF	B.DWN	B.FNT	B.GEN	B.PRL	B.SNG	B.STP	B.TSN	avg_price_value	Spending
103	S	UN	MD	1.29	0.0	1.182	0	2	0	0	0	0	0	0	1	0	0	0	2.377629	0
103	S	UN	MD	0.99	0.0	1.182	0	2	0	0	0	0	0	0	1	0	0	0	2.694808	1
103	S	UN	MD	0.99	0.0	1.182	0	2	0	0	0	0	0	0	1	0	0	0	2.694808	1
103	S	UN	MD	0.99	0.0	1.182	0	2	0	0	0	0	0	0	1	0	0	0	2.694808	1
103	S	UN	MD	1.00	0.0	1.182	0	1	0	0	0	0	0	0	1	0	0	0	2.562024	0
103	S	UN	MD	1.00	0.0	1.182	0	0	0	0	0	0	0	0	1	0	0	0	2.514884	0
103	S	UN	MD	1.00	0.0	1.182	0	0	0	0	0	0	0	0	1	0	0	0	2.514884	0

5). Loyalty Check for Customers –

The customers (HHId) who bought same SKUs both in Calibration Dataset and in Forecast Dataset were considered Loyal via excel formulae as mentioned in **Appendix 9**.The visualization of the same through barplot is done in the below section.

```
> 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")
```

Data View:

ature	ARM	BNC	CLF	DWN	FNT	GEN	PRL	SNG	STP	TSN	B	F	L	S	LT	RG	ST	UN	LR	MD	SM	XL	Loyalty
0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	1
0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	1
0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	1
0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0

6). Linear Regression Model

(A). SKUs to check its dependent variables

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)

```
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
```

Conclusion: 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 variables.

(B). Price and its dependent variables

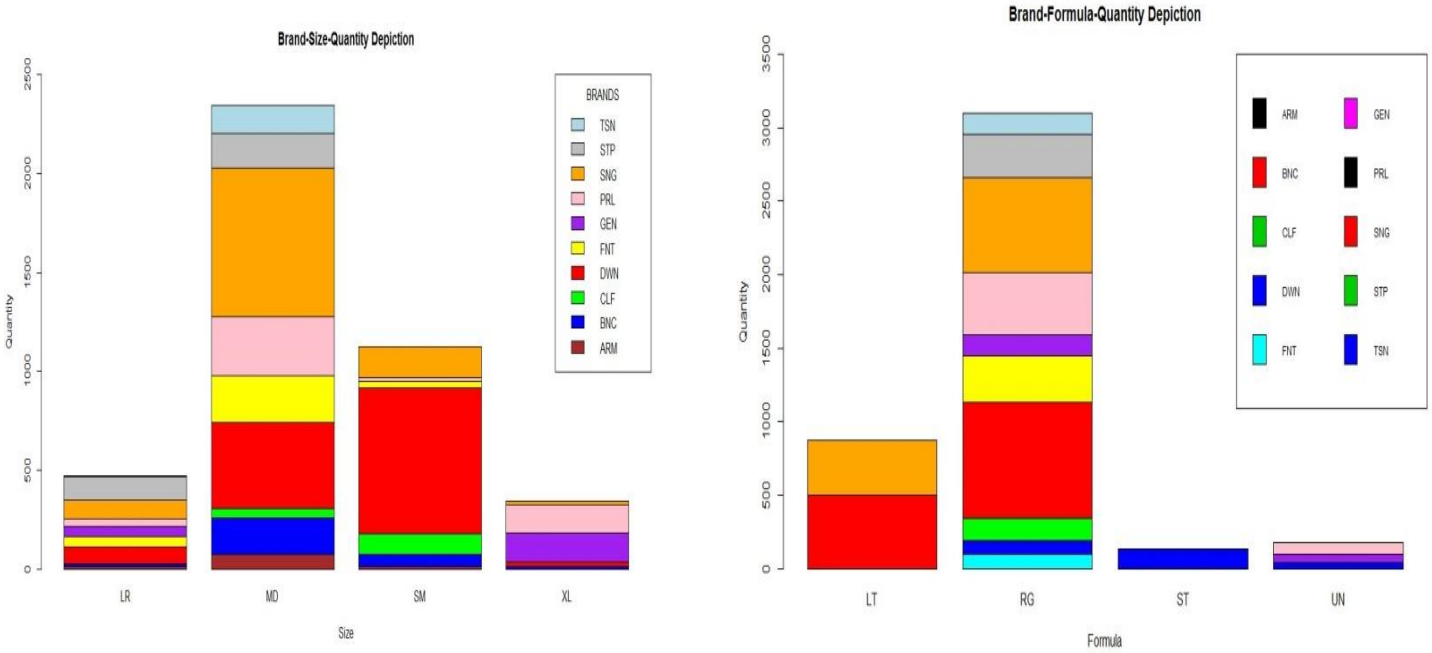
A Linear model with Price as dependent and all other variable signifies that variance of the price is explained by all the environment variables. Every manufactures variable will effect the price of the product.

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

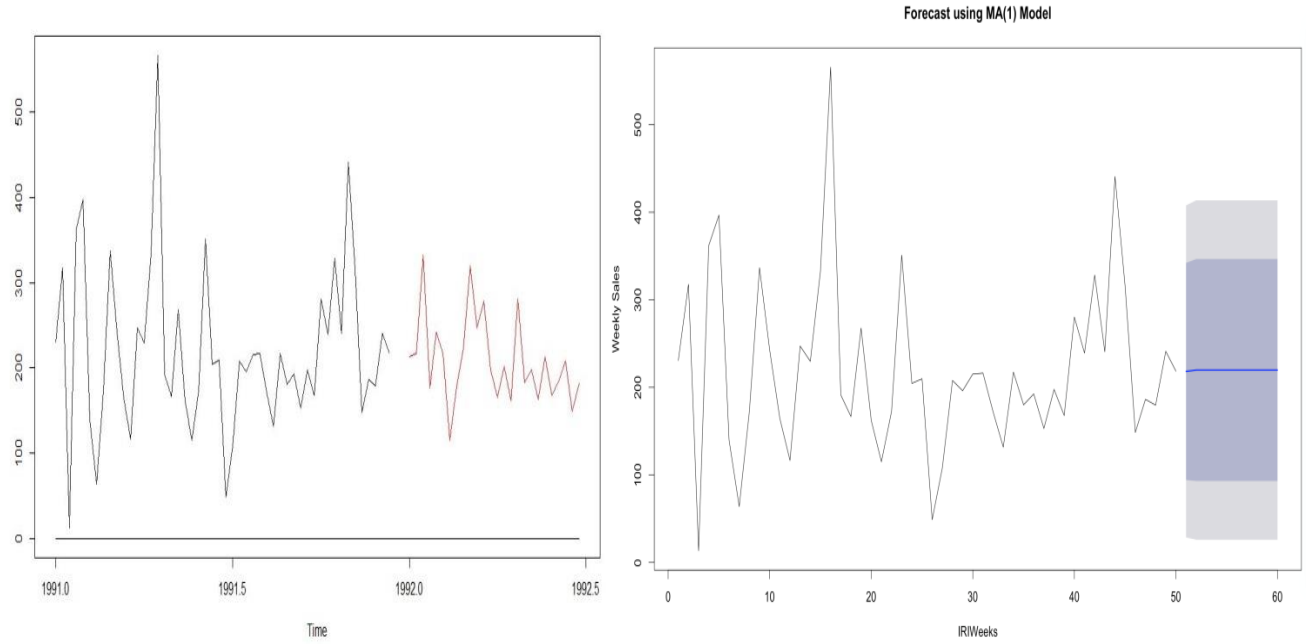
```
Coefficients :
                Estimate Std. Error  t-value  Pr(>|t|)
ARM:(intercept)  2.304801   0.338648   6.8059 1.004e-11 ***
BNC:(intercept)  0.244667   0.267975   0.9130  0.361232
CLF:(intercept)  6.433230   0.359627  17.8886 < 2.2e-16 ***
DWN:(intercept)  1.190421   0.170868   6.9669 3.240e-12 ***
FNT:(intercept)  2.030474   0.238912   8.4988 < 2.2e-16 ***
GEN:(intercept)  6.057266   0.304743  19.8766 < 2.2e-16 ***
PRL:(intercept)  4.531898   0.210891  21.4893 < 2.2e-16 ***
STP:(intercept)  0.694888   0.256923   2.7047  0.006837 **
TSN:(intercept)  3.150613   0.290243  10.8551 < 2.2e-16 ***
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 ***
---
```

Visualization Graphs

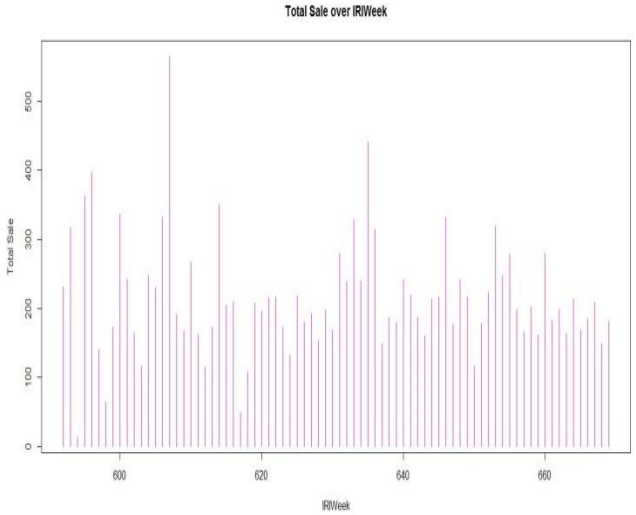
1) **Graph 1: Quantity of Different Sizes for various Brands being sold (Appendix 8):**
   Graph 2: Quantity of Different Formulas for various Brands being sold (Appendix 8):
   Interpretation: Size 'MD' is maximum sold and for Brand SNG from Graph 1. Similarly, Formula 'RG' for Brand 'BNG' is maximum sold.

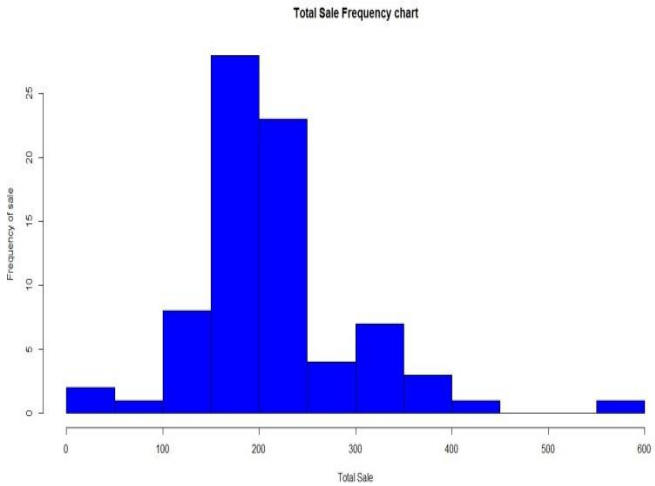


2) **Graph on the left depicts Forecast value (Red) and Training value (Black) for Time series that lies within 95% C.I as predicted in the grey section of the graph on the right.(Appendix 4)**

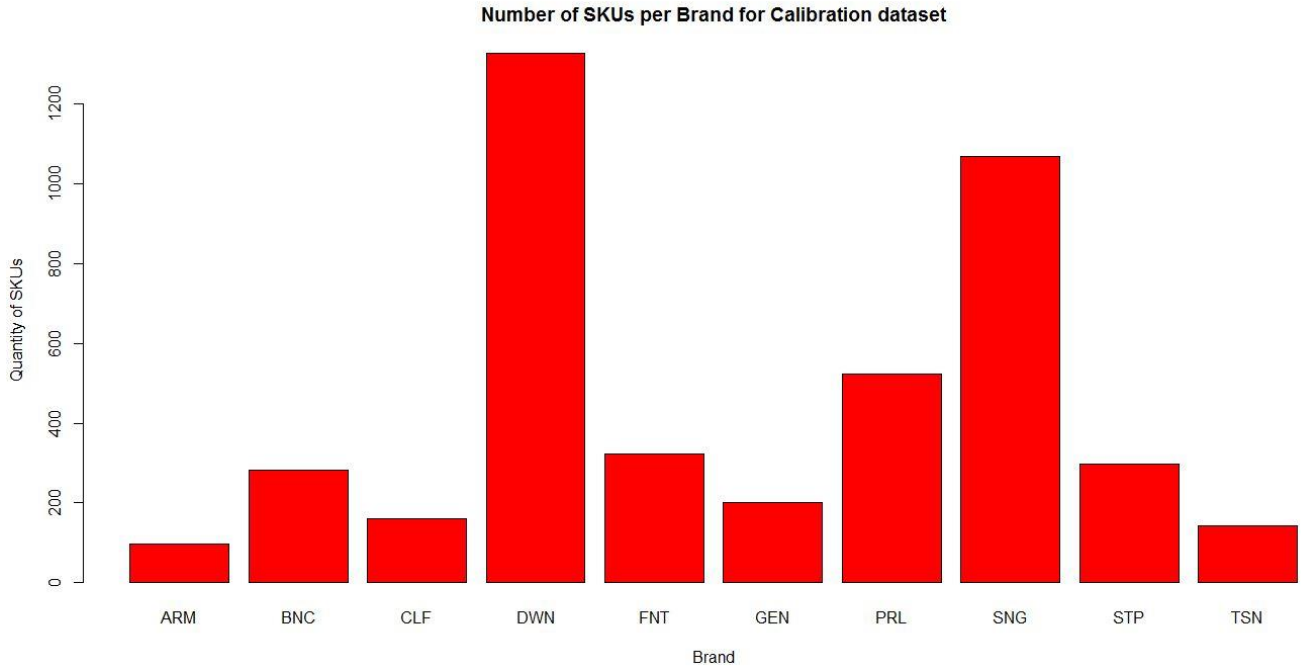


3) **Graph depicting Total Sale over Unique IRIWeeks (Appendix 6) :**
Interpretation: Most of the IRIWeeks end up having sale of around $180-$220 products.

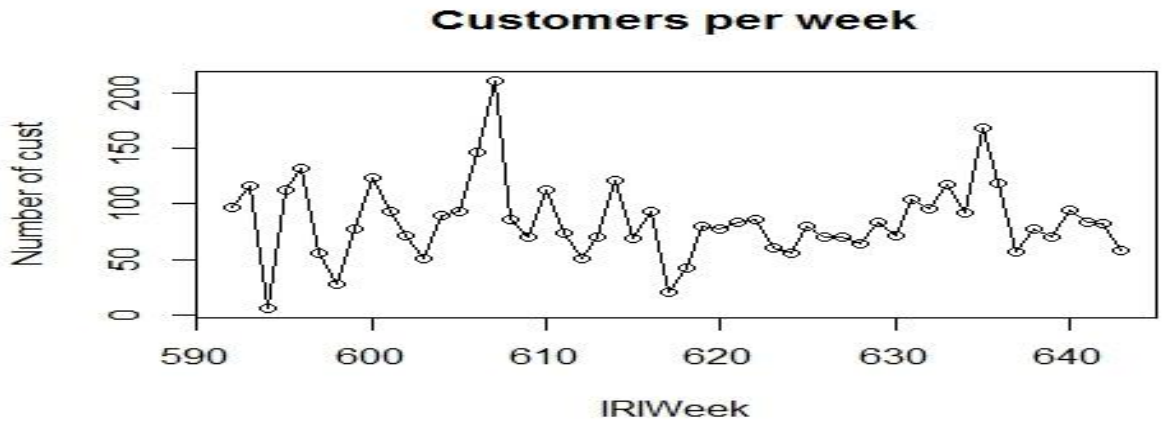




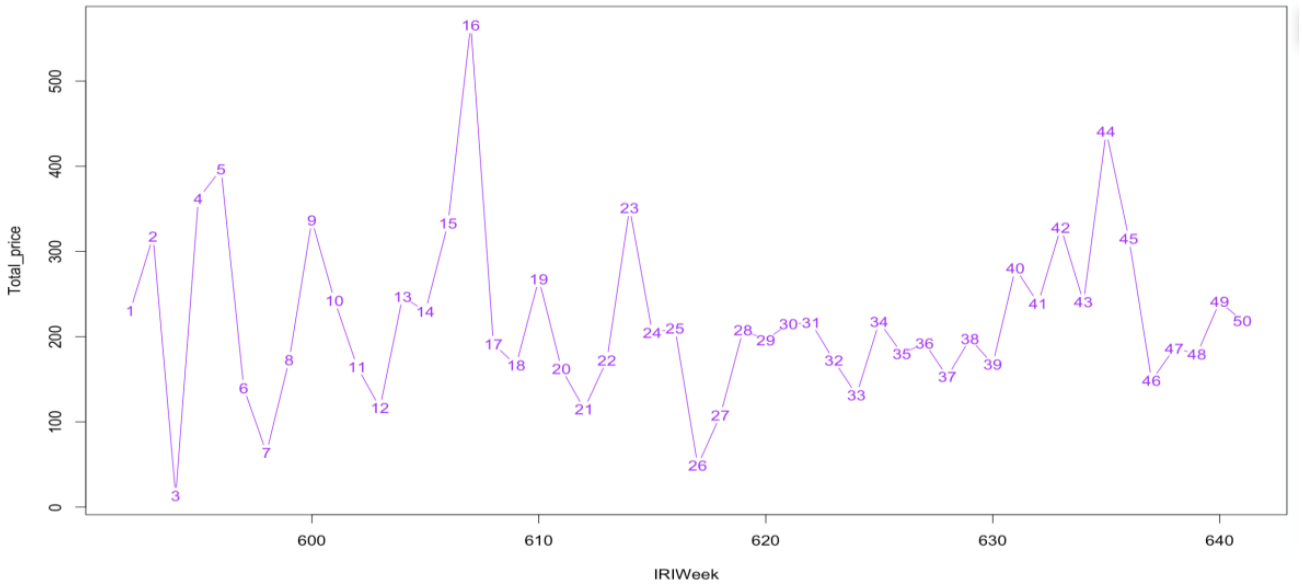
4) **Graph showcasing the number of SKUs for different Brands from Calibration dataset (Appendix 1):**
Interpretation: Clearly, DWN has max SKUs being sold followed by SNG.



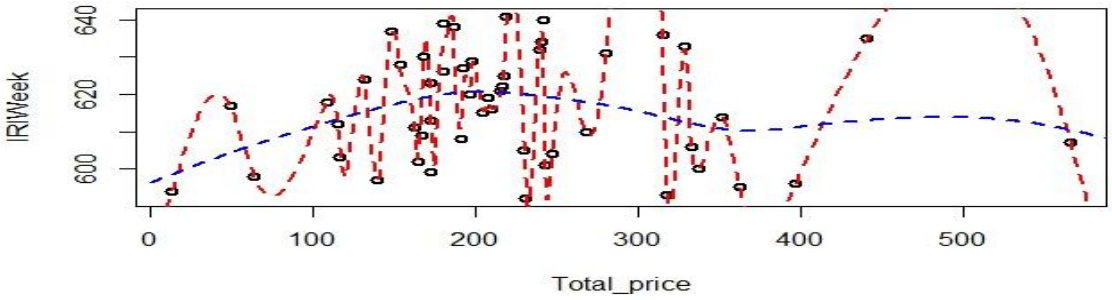
5) **Graph showcasing the number of customers visiting the store per week. This helps in identifying the weeks where there is a huge rush within store for shopping:**
Interpretation: For weeks 607 (14-Apr'91) and 635 (3-Nov'91), there is huge rush of customers reason may be this is the time the store offers discounted sales on different products.



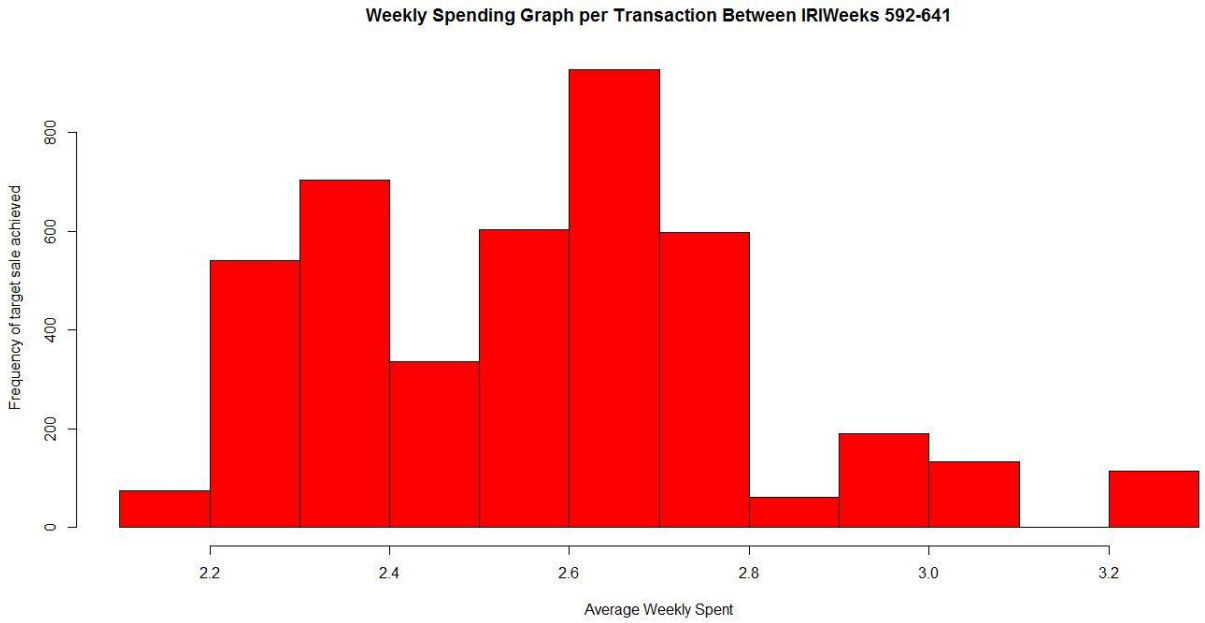
6) **Graph below shows the Total_price or Sale of the store for corresponding IRIWeeks. We observe that the majority of the sales occur in the range of $180-$220 per week Appendix 6:**



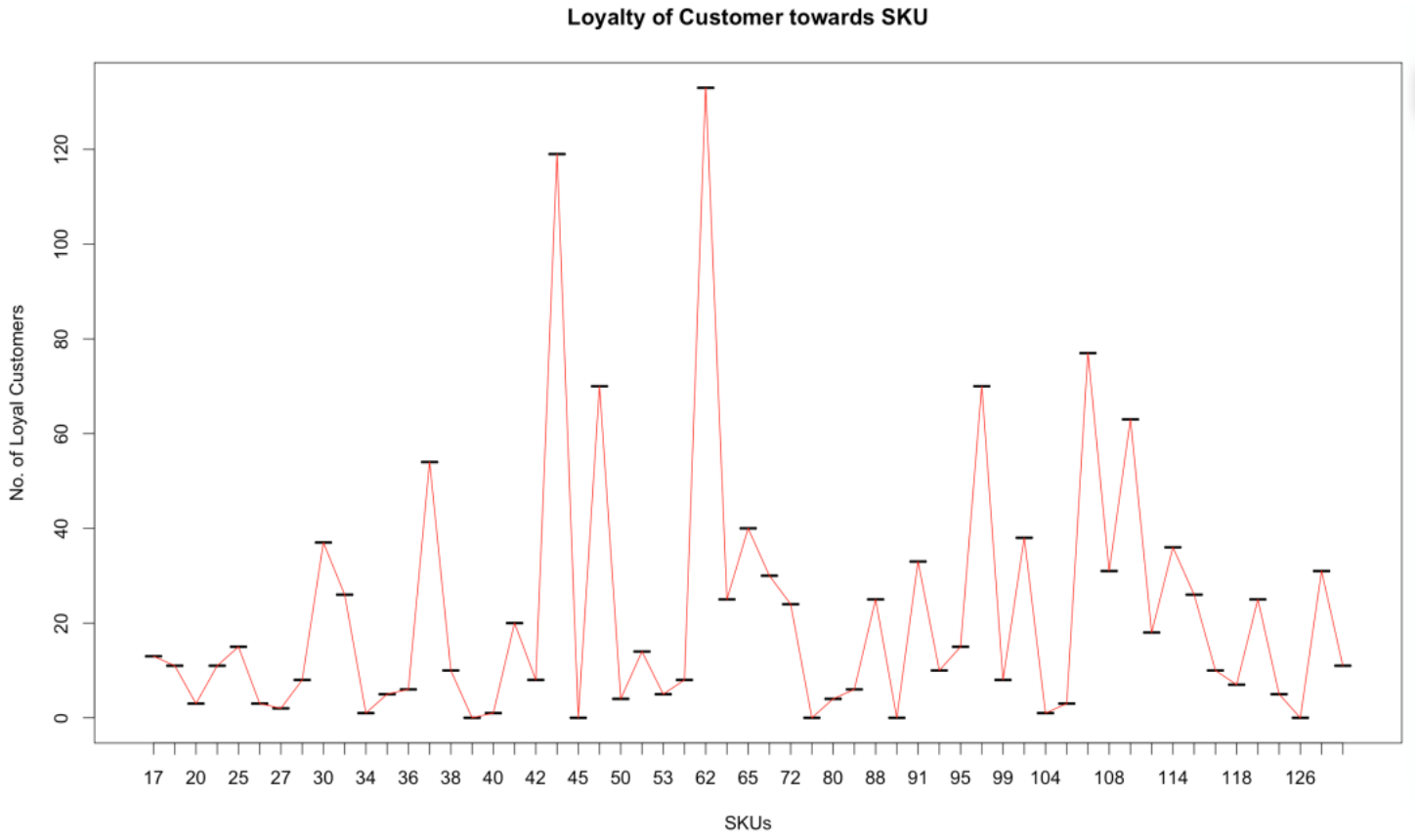
7) **Graph below illustrate the Spline curves(Red) and the best fit curve (Blue) for the weekly sales in Time series (Appendix 7):**



8) **Graph below provides view on average sale per transaction (granular level detail). This helps the business get to know how much target sale per transaction they have achieved Appendix 4:**

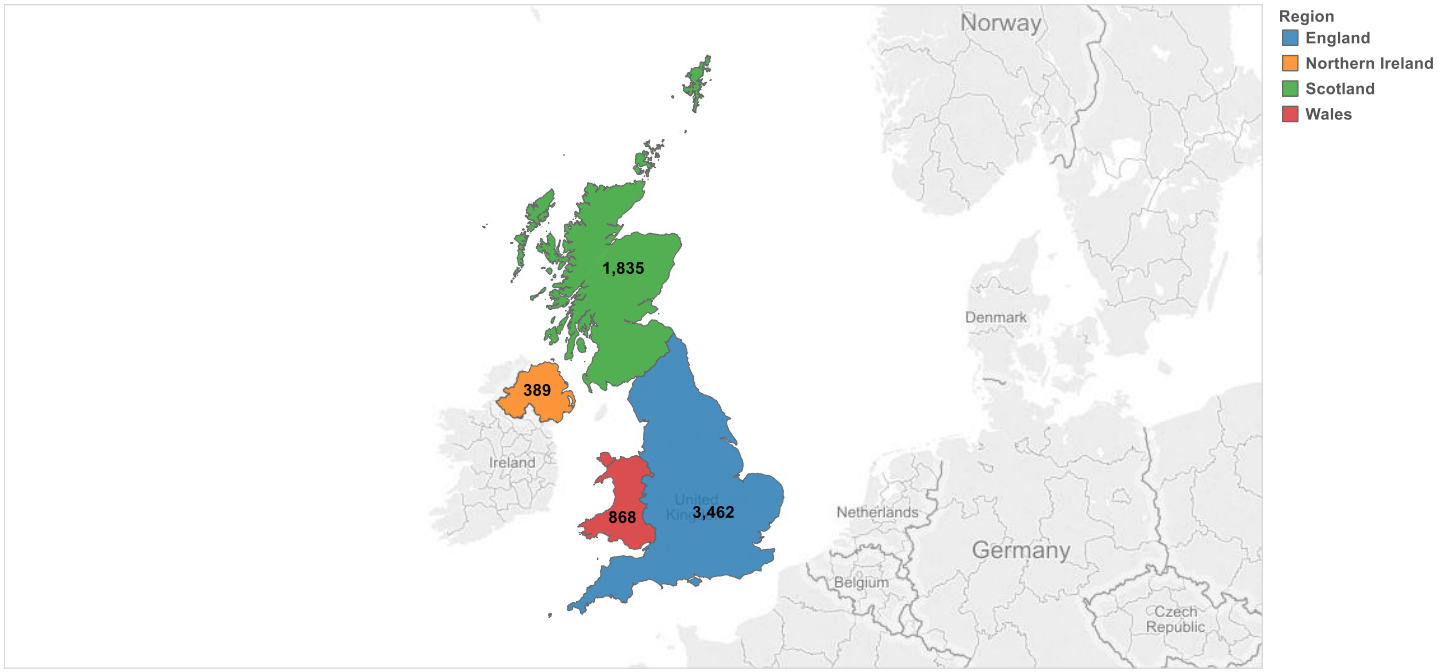


1). Loyalty Check Graph Below depicts the clearly that most number of loyal customers are for SKU 62 and second most loyal are for SKU 44 – Check Appendix 9



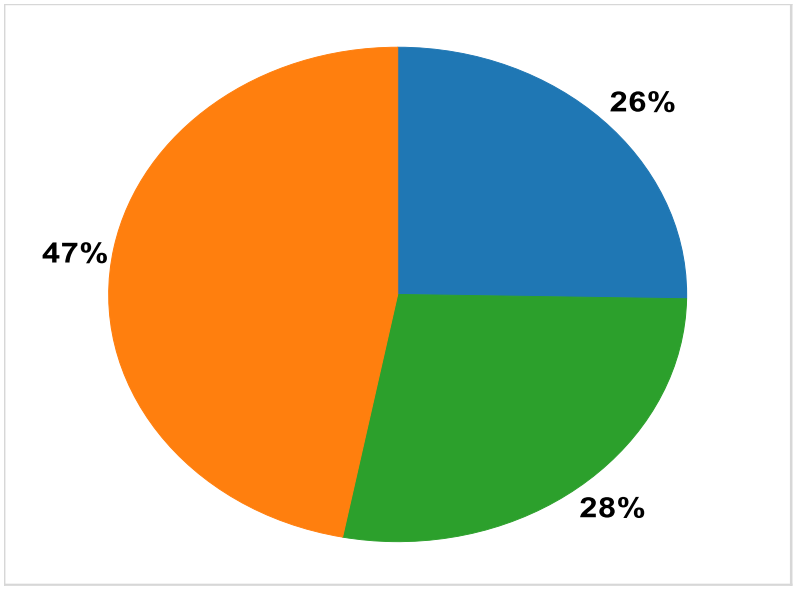
Visualization through Tableau

Map



Map based on Longitude (generated) and Latitude (generated). Color shows details about Region. The marks are labeled by sum of Number of Records.

Job Classification



% of Total Number of Records. Color shows details about Job Classification. Size shows sum of Number of Records. The marks are labeled by %

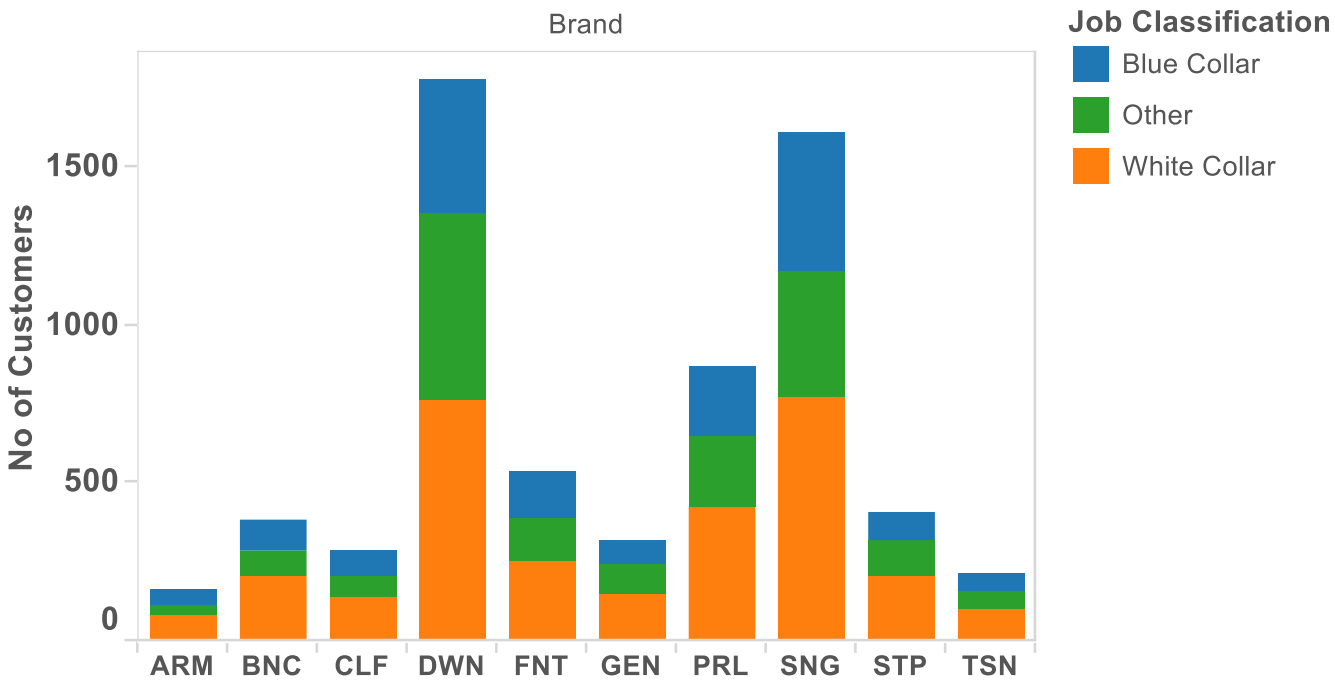
Number of Records

6,554

Job Classification

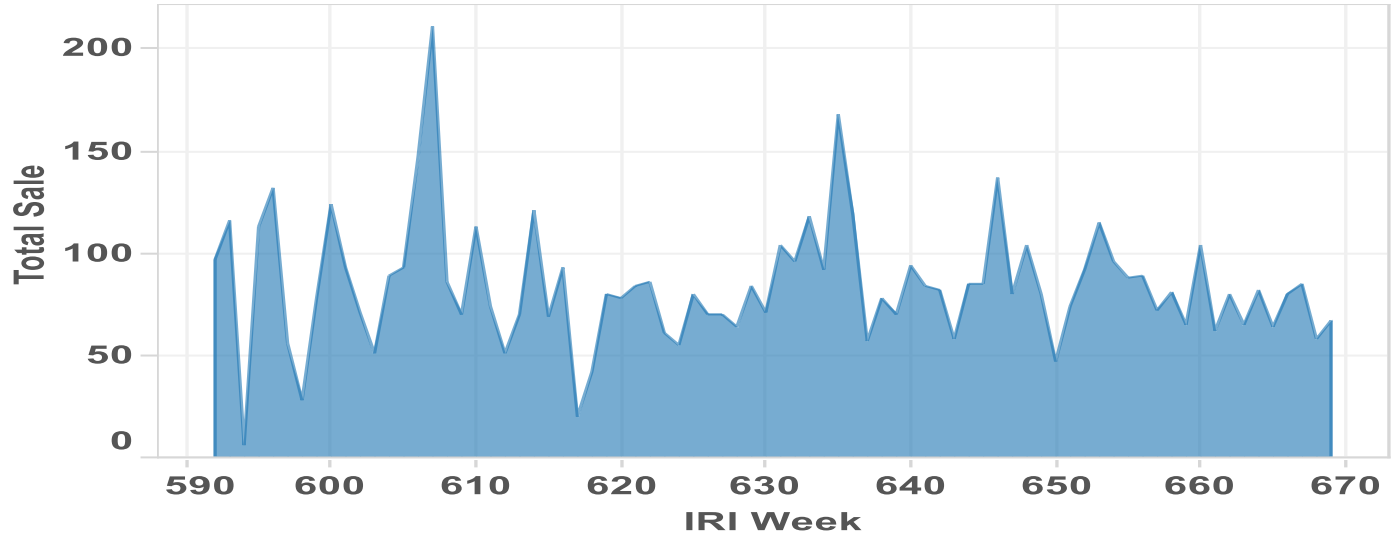
- Blue Collar
- Other
- White Collar

Brand Value



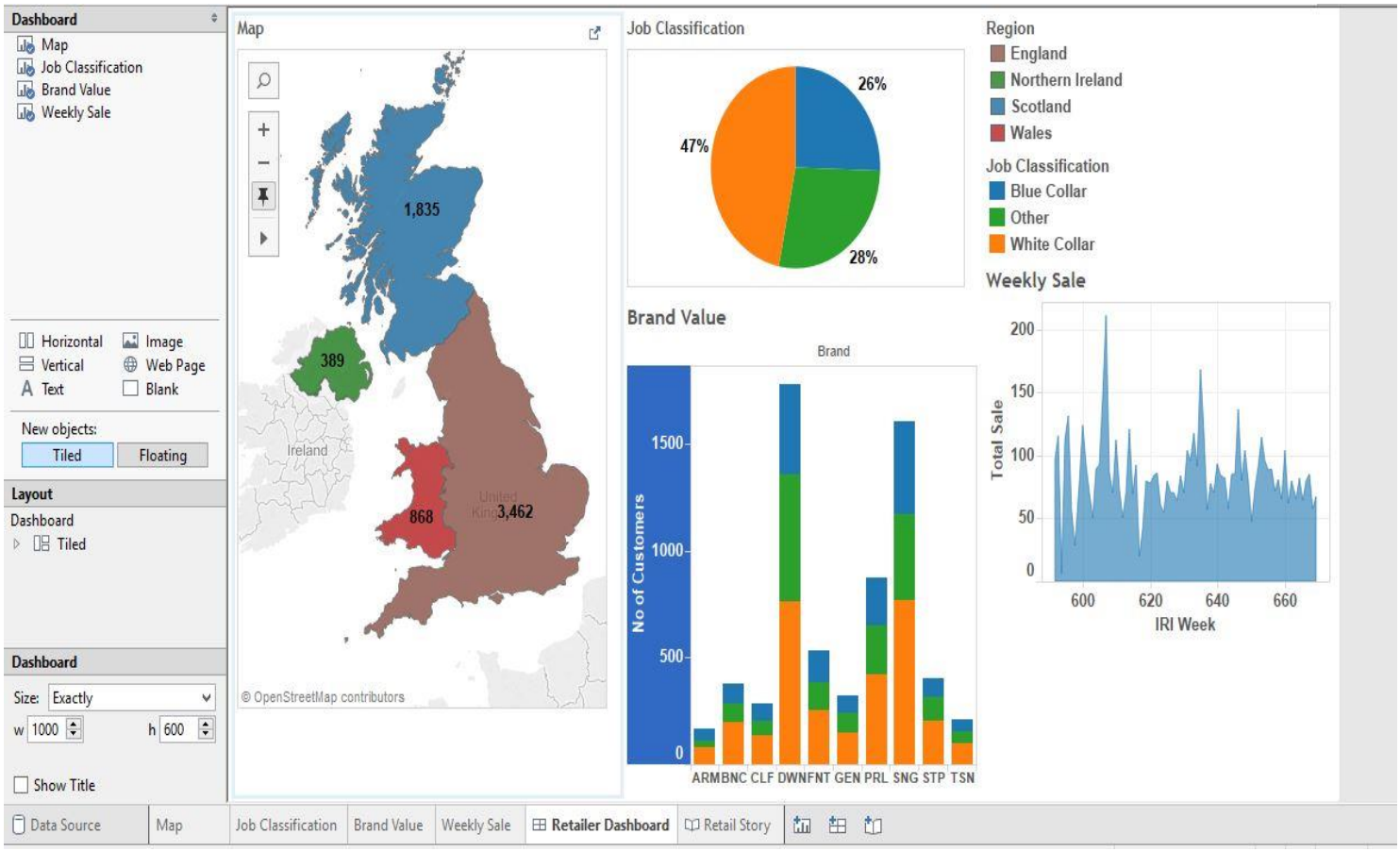
Sum of Number of Records for each Brand. Color shows details about Job Classification. The data is filtered on Region and Action (Region). The Region filter keeps England, Northern Ireland, Scotland and Wales. The Action (Region) filter keeps 4 members.

Weekly Sale

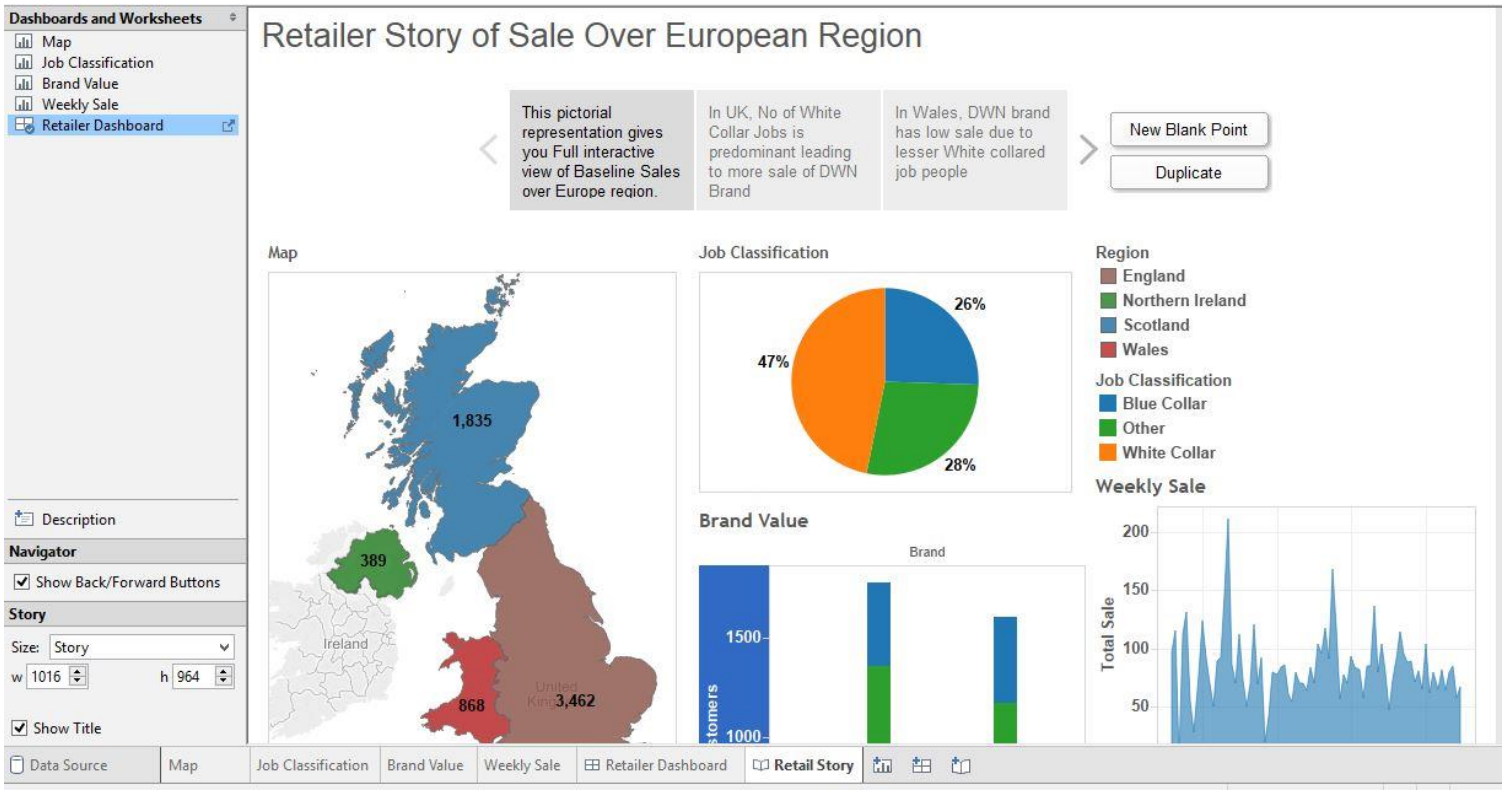


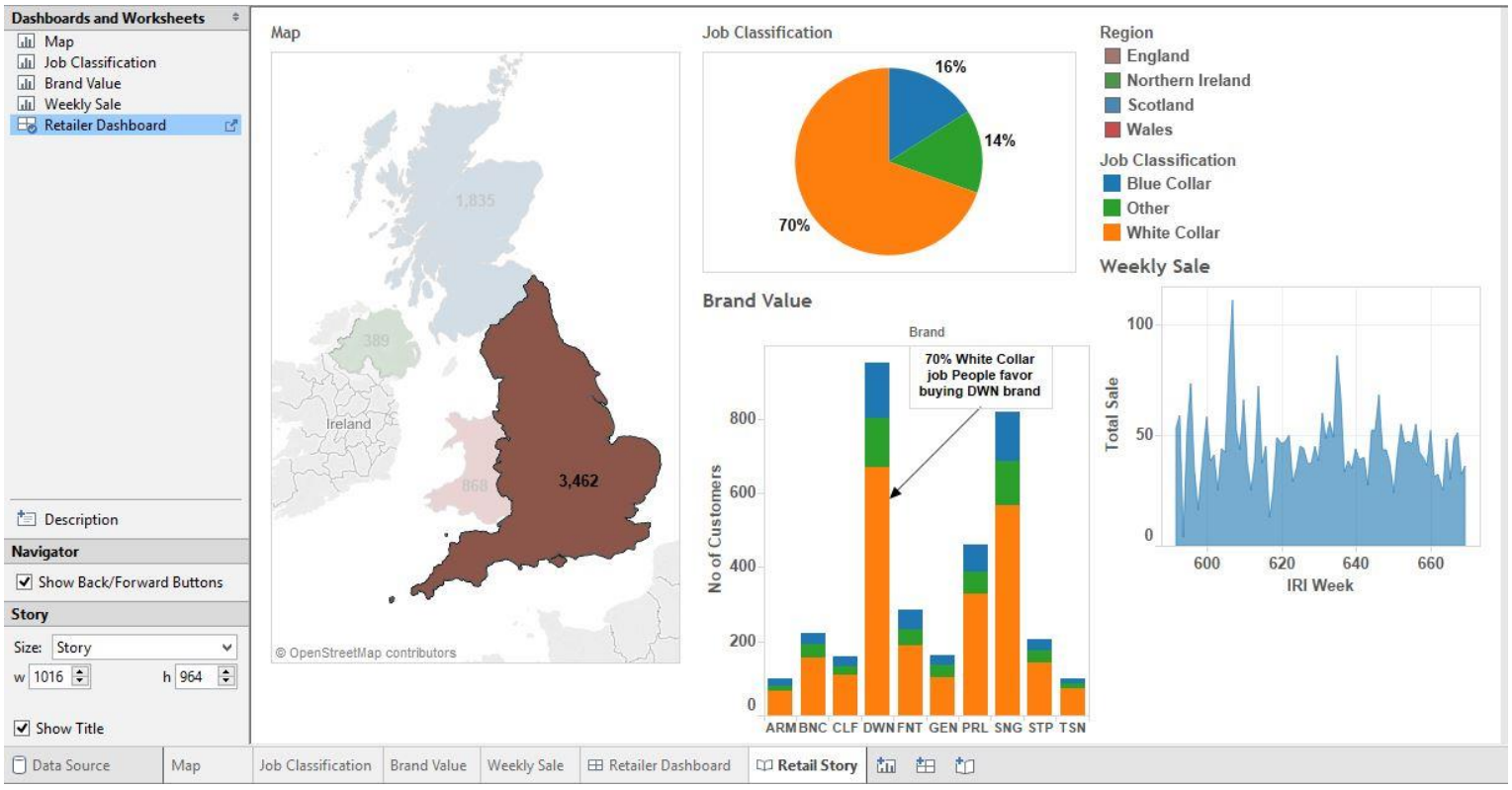
The plot of sum of Number of Records for IRI Week. The data is filtered on Action (Region) and Action (Brand,Job Classification). The Action (Region) filter keeps 4 members. The Action (Brand,Job Classification) filter keeps 30 members.

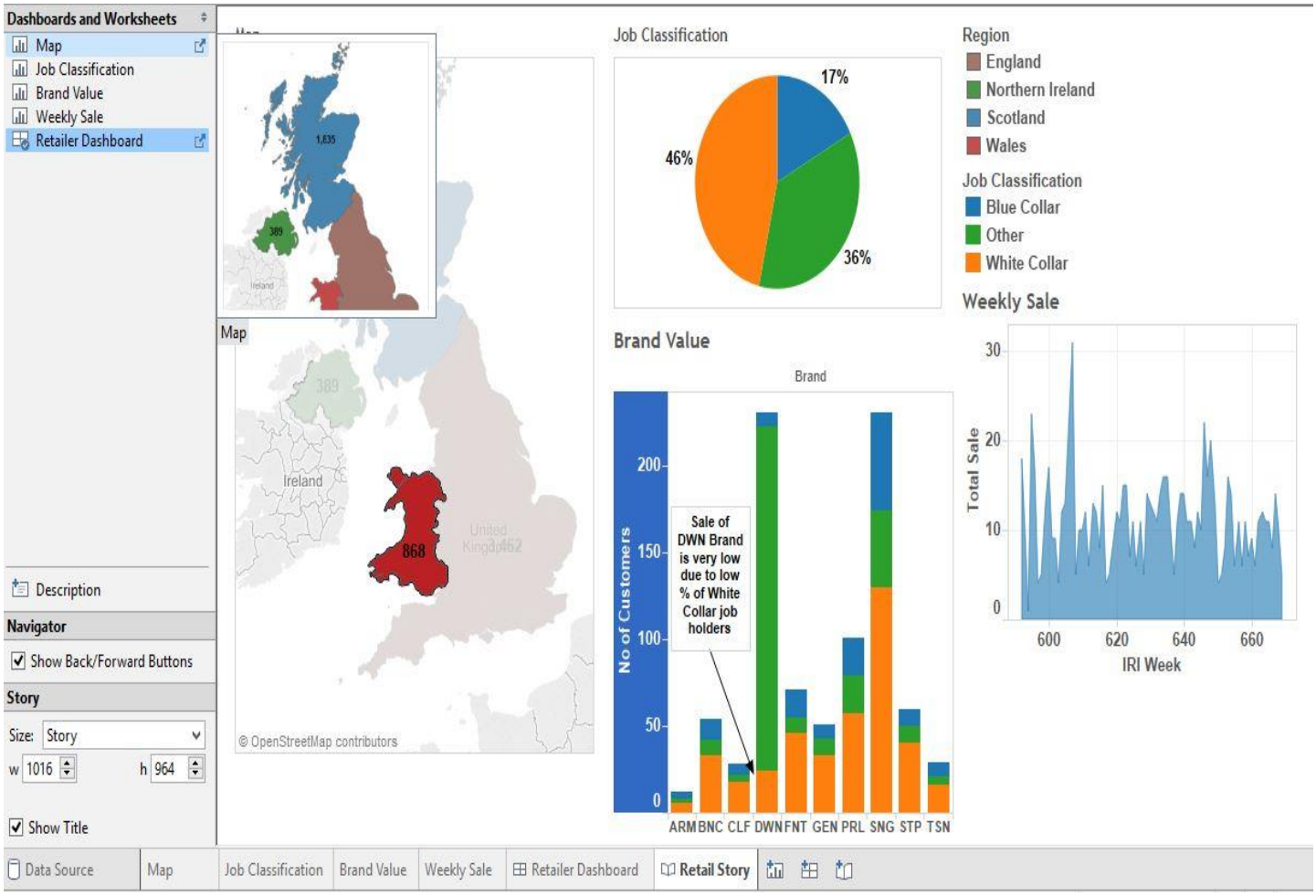
Dashboard


Dashboard
Map
Job Classification
Brand Value
Weekly Sale
Horizontal
Vertical
Text
Image
Web Page
Blank
New objects:
Tiled
Floating
Layout
Dashboard
Tiled
Dashboard
Size: Exactly
w 1000
h 600
Show Title
Map
1,835
389
868
3,462
Ireland
© OpenStreetMap contributors
Job Classification
26%
47%
28%
Region
England
Northern Ireland
Scotland
Wales
Job Classification
Blue Collar
Other
White Collar
Brand Value
Brand
No of Customers
1500
1000
500
0
ARM BNC CLF DWN FNT GEN PRL SNG STP TSN
Weekly Sale
Total Sale
200
150
100
50
0
600
620
640
660
IRI Week
Data Source
Map
Job Classification
Brand Value
Weekly Sale
Retailer Dashboard
Retail Story

Story Line


Dashboards and Worksheets
Map
Job Classification
Brand Value
Weekly Sale
Retailer Dashboard
Retailer Story of Sale Over European Region
This pictorial representation gives you Full interactive view of Baseline Sales over Europe region.
In UK, No of White Collar Jobs is predominant leading to more sale of DWN Brand
In Wales, DWN brand has low sale due to lesser White collared job people
New Blank Point
Duplicate
Map
1,835
389
868
3,462
Ireland
Job Classification
26%
47%
28%
Region
England
Northern Ireland
Scotland
Wales
Job Classification
Blue Collar
Other
White Collar
Weekly Sale
Total Sale
Brand Value
Brand
1500
1000
Description
Navigator
Show Back/Forward Buttons
Story
Size: Story
w 1016
h 964
Show Title
Data Source
Map
Job Classification
Brand Value
Weekly Sale
Retailer Dashboard
Retail Story


Dashboards and Worksheets
Map
Job Classification
Brand Value
Weekly Sale
Retailer Dashboard
Map
1,835
389
Ireland
868
3,462
© OpenStreetMap contributors
Job Classification
16%
14%
70%
Region
England
Northern Ireland
Scotland
Wales
Job Classification
Blue Collar
Other
White Collar
Weekly Sale
Total Sale
IRI Week
Brand Value
Brand
70% White Collar job People favor buying DWN brand
No of Customers
ARM BNC CLF DWN FNT GEN PRL SNG STP TSN
Description
Navigator
Show Back/Forward Buttons
Story
Size: Story
w 1016
h 964
Show Title
Data Source
Map
Job Classification
Brand Value
Weekly Sale
Retailer Dashboard
Retail Story


Dashboards and Worksheets
Map
Job Classification
Brand Value
Weekly Sale
Retailer Dashboard
Description
Navigator
Show Back/Forward Buttons
Story
Size: Story
w 1016 h 964
Show Title
1,635
389
Ireland
868
Map
© OpenStreetMap contributors
Job Classification
17%
46%
36%
Region
England
Northern Ireland
Scotland
Wales
Job Classification
Blue Collar
Other
White Collar
Brand Value
Brand
No of Customers
Sale of DWN Brand is very low due to low % of White Collar job holders
ARM BNC CLF DWN FNT GEN PRL SNG STP TSN
Weekly Sale
Total Sale
IRI Week
Data Source
Map
Job Classification
Brand Value
Weekly Sale
Retailer Dashboard
Retail Story

Appendix (R-Script)

Appendix 0: Data Cleaning Script:

```
setwd(".....\\USF\\MIS\\SDM\\Final Project")
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")
    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$IRIWeek)
    purplusmerch <- merge(purdata, merchdata, by=c("IRIWeek", "Store","SKU"))
   attrdata<-read.csv("Membership panel Data.csv")
attrdata<-attrdata[,-1]#this removes the first column in the dataframe "attrdata" and shifts all remaining by 1
attrplusmerch <- merge(purplusmerch, attrdata, by=c("SKU")) #merges purplusmerch data created above with attrdata
arspdata<-read.table("ARSP.DAT")
    names(arspdata)<-c("SKU","Store","ARSP")
    finaldata <- merge(attrplusmerch, arspdata, by=c("SKU","Store"))
    finaldata<-
finaldata[,c("HHId","SKU","IRIWeek","ARM","BNC","CLF","DWN","FNT","GEN","PRL","SNG","STP","TSN","B","F","L","
S","LT","RG","ST","UN","LR","MD","SM","XL","PricePaid","RegPrice","ARSP","Display","Feature")]
    finaldata$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) #this is the finalized_data.csv file used for further analysis
```

Appendix 1: Mlogit regression – *# Multinomial logit model coefficients*

#Best selling brand - Every other brand has less significant value than DWN brand
#Worst selling brand - Every other brand has significant value than ARM brand
 # Multinomial logit model coefficients (with different base outcome)
#CLF is the cheapest brand in terms of Price
#SNG is the most expensive brand in terms of Price

```
#Multinomial Logistic Regression Model
    install.packages("VGAM")
    library(VGAM)
    install.packages("mlogit")
    library(mlogit)
    detach(calibdata)
    rm(calibdata)
    #The calibration file been used here contains the data from Final data file created after Data cleaning only for
IRIWeeks 592-641.
    setwd("..USF\\MIS\\SDM\\Final Project")
    calibdata<-read.csv("Final_Data_Calibration.csv")
    dim(calibdata)
    names(calibdata)
    attach(calibdata)
    #Descriptive statistics of Brand Variable. There are 10 Different Brands with corresponding purchase rows in
Calibration Dataset
    table(Brand)
    #Reshaping the data from wide to long format
    calibdata$Brand<-as.factor(calibdata$Brand)
    mldata<-mlogit.data(calibdata, varying=13:22, choice="Brand", shape="wide")
    mldata[1:25,]
    # Multinomial logit model coefficients
    #Best selling brand - DWN - All coefficients of the brands are negative (exp value <1) i.e. Every other brand has lesser
value than STP brand.
    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 corelated because for other predictor values like
HHId,Form,Formula2,Size,etc. the p-value was not significant (>0.05)
    #Worst selling brand - ARM - All coefficients of the brands are positive (exp value >1) i.e. Every other brand has more
value than BNC brand
    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 expensive brand in terms of Price since the coefficient of all other brand price is negative (exp value <
1) in reference to SNG.
  mlogit.model3 <- mlogit(Brand ~ 1 | Price, data = mldata, reflevel="SNG")
  summary(mlogit.model3)
  exp(coef(mlogit.model3))
```

```
#CLF is the cheapest brand in terms of Price since the coefficient of all other brand price is positive (exp value >1) in reference to CLF.
mlogit.model4 <- mlogit(Brand ~ 1 | Price, data = mldata, reflevel="CLF")
summary(mlogit.model4)
exp(coef(mlogit.model4))
#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 2: vglm regression — *to check for the significance of various predictors using regression model p-values.*

```
#vglm Function for multinomial regression modeling.
install.packages("VGAM")
library(VGAM)
rm(list=ls())
setwd("...USF\\MIS\\SDM\\Final Project")
calibdata<-read.csv("Final_Data_Calibration.csv")
dim(calibdata)
names(calibdata)
attach(calibdata)
SKU <- as.factor(SKU)
#Using vglm function model to predict all significant dependent 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=calibdata, family=multinomial)
summary(vglm_mod1)
exp(coefficients(vglm_mod1))
#Only Brand and SKUs are strong covariants. Other variables like HHId and IRIWeek doesn't have much significant in the data variation.
#Using vglm function model to predict if Price attribute is dependent on Brand.
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=calibdata, family=multinomial)
summary(vglm_mod2)
exp(coefficients(vglm_mod2))
#Hence Price and Brands are highly correlated. Because of this analysis we concluded that we have better chance at creating multinomial regression model on "Brand Vs SKUs" or "Brand vs Price" for predictive analysis.
```

Appendix 3: multinorm regression – *we ran multinomial logistic regression model on Brand and SKU(check for +Price) for the calibration training data (IRIWeeks 592-641) and then predicted the brand names for the forecast data(IRIWeeks 644-669).(check for accuracy)*

```
#Multinom Function - Multinomial regression on Brand and SKU Level.
rm(list=ls())
require(foreign)
require(nnet)
detach(calibdata)
detach(traindata)
detach(foredata)
```

```
setwd("E:\\MBA\\GMAT\\SKM_MS-MIS_Docs\\USF\\MIS\\SDM\\Final Project")
traindata<-read.csv("Final_Data_Calibration_Training.csv")
dim(traindata)
names(traindata)
attach(traindata)
#Giving reference level of Brand="DWN" for multinom Model.
traindata$Brand <- relevel(traindata$Brand, ref = "DWN")
#Multinom regression Model
multi_mod_test <- multinom(traindata$Brand ~ SKU+Price, data = traindata)
summary(multi_mod_test)
# Model gives AIC value is 70.08043 for the significant dependent attributes SKU and Price.
multi_mod <- multinom(traindata$Brand ~ SKU, data = traindata)
summary(multi_mod)
# Model is significantly fit as the AIC value is 38.58645(Lower than the value obtained other combination of predictors).Hence we think Only Brand Vs SKUs (excluding Price) will give us better model.
#The below part calculates p-values corresponding to above model as multinom function does not produce p-values in the model.
z <- summary(multi_mod)$coefficients/summary(multi_mod)$standard.errors
z
p <- (1 - pnorm(abs(z), 0, 1)) * 2
p
exp(coef(multi_mod))
head(pp <- fitted(multi_mod))
#Below part performs the validation of the model created using multinom function above using validation file that contains data values from IRIWeek 642-643.
detach(validdata)
setwd("...USF\\MIS\\SDM\\Final Project")
validdata<-read.csv("Final_Data_Calibration_Validation.csv")
dim(validdata)
names(validdata)
attach(validdata)
#Below function is used to predict the probable values of the brand into our validation data set.
predict(multi_mod, newdata = validdata, "probs")
#Here, the values ~0.99 or exactly 1.0 confirms the presence of the corresponding Brand name for the purchase row in the dataset.
#On Cross verification with the dataset, we observed the model is accurately predicting the values. For eg. 75-109 row denoted DWN probability = (0.9,1.0) which is true.
#Similarly, we will try to predict the Brand name for all the purchase rows in the forecast data that contains data values from IRIWeek 643-669.
detach(foredata)
setwd("....USF\\MIS\\SDM\\Final Project")
foredata<-read.csv("Final_Data_Forecast.csv")
dim(foredata)
names(foredata)
attach(foredata)
#predicting values based out of previous model generated
Brandpred_value <- data.frame(predict(multi_mod, newdata = foredata, "probs"))
#adding predicted values into our dataset
Brand_predicat <- cbind(foredata,Brandpred_value)
```

```
View(Brand_predicat)
#Here again we observe that the probability values corresponding to each purchase row accurately points to the Brand name. For eg. rows 1-48 give Prob of 0.998 for PRL which is true.
#To check for the accuracy(Lets keep the cut off probability as 0.9)
cut.off <- 0.9
pred.brand <- (Brandpred_value > cut.off)
table(pred.brand)
table(Brand_predicat$PRL) #No of purchase rows for Brand "PRL" with probability higher than 0.9 are 350 which is true if we see from the file directly.
table(Brand)
#Similarly for Brand "DWN" are 452 which is also true.
table(Brand_predicat$DWN)
table(Brand)
#Hence our multinom logistic regression model is highly accurate.
```

Appendix 4: Ologit 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).*

```
#Ologit function for Ordinal Logistic regression model based on HIGH,MEDIUM and LOW ordered values for Sale Price of Purchase(HIGH>2.8,2.8>MEDIUM>2.4,LOW<2.4 per transaction)
#We were not able to fully predict the futuristic values for the Sale price on purchase through this model on all three categories
#Because of which we also ran Logistic Regression model after this model to predict only HIGH or LOW (2 categories: HIGH>2.6 and LOW<2.6 per transaction) which has ~70% accuracy rate.
rm(list=ls())
install.packages("rms")
library(rms)
require(foreign)
require(ggplot2)
require(MASS)
require(Hmisc)
require(reshape2)
setwd("...USF\\MIS\\SDM\\Final Project")
traindata<-read.csv("Final_Data_Calibration_Training.csv")
dim(traindata)
names(traindata)
attach(traindata)
#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)
#To check for the frequency of the target sale achieved per transaction in an IRIWeek for the training data.
hist(avg_price_value, main = "Weekly Spending Graph Between IRIWeeks 592-641", xlab = "Average Weekly Spent", ylab = "Frequency of target sale achieved", col = "Red", border = "Black")
```

```
summary(avg_price_value)
#Adding a column for HIGH/MEDIUM/LOW based on avg_price_value of the transaction over an IRIWeek.
for (i in 1:nrow(traindata)){
  if(traindata$avg_price_value[i] <= 2.4){
    traindata$Spending[i] <- "Low" }
  if(traindata$avg_price_value[i] > 2.4 && traindata$avg_price_value[i] <= 2.8){
    traindata$Spending[i] <- "Medium"
  }
  if(traindata$avg_price_value[i] > 2.8 ){
    { traindata$Spending[i] <- "High" }
  }
}
View(traindata)
attach(traindata)
Y <- cbind(Spending)
X <- cbind(IRIWeek,Brand)
Xvar <- c("IRIWeek","Brand")
# Descriptive statistics
summary(Y)
summary(X)
table(Y)
# Ordered logit model coefficients
ddist<- datadist(Xvar)
options(datadist='ddist')
ologit<- lrm(Y ~ X, data = traindata)
print(ologit)
# Ordered logit predicted probabilities
xmeans <- colMeans(X)
newdata1 <- data.frame(t(xmeans))
#Predicting futuristic category of forecast transactions
setwd("...USF\\MIS\\SDM\\Final Project")
foredata<-read.csv("Final_Data_Forecast.csv")
dim(foredata)
names(foredata)
attach(foredata)
#Below for loop to find mean/avg price/sale in a particular IRIWeek for Forecast Dataset
for (j in 1:nrow(foredata)){
  for(i in 644:669) {
    if(i %in% IRIWeek[j]){
      foredata$avg_price_value[j] <- mean(foredata[which(IRIWeek == i),c("Price")])
    }
  }}
View(foredata)
attach(foredata)
#Adding a column for HIGH/MEDIUM/LOW based on avg_price_value of the transaction over an IRIWeek.
for (i in 1:nrow(foredata)){
  if(foredata$avg_price_value[i] <= 2.4){
    foredata$Spending[i] <- "Low" }
  if(foredata$avg_price_value[i] > 2.4 && foredata$avg_price_value[i] <= 2.8){
```

```
   foredata$Spending[i] <- "Medium"
  }
  if(foredata$avg_price_value[i] > 2.8 ){
   { foredata$Spending[i] <- "High" }
  }
}
#Adding a column to verify the actual category with the predicted category. We could see some discrepency on this and hence we decided to only run logistic regression for HIGH/LOW vaues.
Spending_pred <- predict(ologit, newdata=foredata, type="fitted.ind")
colMeans(Spending_pred)
foredata <- cbind(traindata,Spending_pred)
View(foredata)
write.csv(foredata,"Final_Data_Weekly_Forecast_Spending_Predicted.csv")
```

Appendix 5: 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.*

```
#Logistic regression model to predict the various purchase transaction category between HIGH(>$2.6/purchase transaction) or LOW(<=$2.6/purchase transaction)
rm(list=ls())
install.packages("rms")
library(rms)
require(foreign)
require(ggplot2)
require(MASS)
require(Hmisc)
require(reshape2)
setwd("...USF\\MIS\\SDM\\Final Project")
traindata<-read.csv("Final_Data_Calibration_Training.csv")
dim(traindata)
names(traindata)
attach(traindata)
#Below for loop to find mean/avg price/sale in a particular IRIWeek for Training Dataset
for (j in 1:nrow(traindata)){
 for(i in 592:641) {
  if(i %in% IRIWeek[j]){
   traindata$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 combination 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
setwd("...USF\\MIS\\SDM\\Final Project")
foredata<-read.csv("Final_Data_Forecast.csv")
dim(foredata)
names(foredata)
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$avg_price_value[j] <- mean(foredata[which(IRIWeek == i),c("Price")])
    }
  }}
View(foredata)
attach(foredata)
#Adding a column for HIGH(0)/LOW(1) based on avg_price_value of the transaction over an IRIWeek.
 for (i in 1:nrow(foredata)){
  if(foredata$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
write.csv(foredata,"Final_Data_Weekly_Forecast_Spending_Predicted.csv")
```

Appendix 6: Evaluations of the importance of pricing/sale for IRIWeek

```
#Evaluation of the pricing/sale for IRIWeek
```

```
library(ggplot2)
detach(finaldata)
setwd("....\\USF\\MIS\\SDM\\Final Project")
#Below is Finalized Dataset with "Brand" Column having all categorical values in it.
finaldata<-read.csv("finalized_data_Latest.csv")
dim(finaldata)
names(finaldata);
attach(finaldata)
table(Brand)
#Below loop sums up the price attribute for all IRIWeeks i.e. Total sale per week.
for (j in 1:nrow(finaldata)){
  for(i in 592:669) {
    if(i %in% IRIWeek[j]){
      finaldata$Total_price[j] <- sum(finaldata[which(IRIWeek == i),c("Price")])
    }
  }}
attach(finaldata)
duplicates = duplicated(IRIWeek,Total_price)
duplicates[1:10]
unique_finaldata <- finaldata[!duplicated(finaldata[c("IRIWeek","Total_price")]),]
sorted_finaldata <- unique_finaldata[order(unique_finaldata$IRIWeek),]
attach(sorted_finaldata)
names(finaldata)
SKU <- as.factor(SKU)
#Below histogram shows that majority of the weeks end up having total sale of around $200 in a week for the given set of 10 Brands.
hist(Total_price, main = "Total Sale Frequency chart", ylab = "Frequency of sale", xlab = "Total Sale", col = "Blue")
#Below plot showcase the total sale Per IRIWeek
plot(IRIWeek,Total_price, type="h", ylab="Total Sale",  col="Magenta", main="Total Sale over IRIWeek")
```

Appendix 7: Time-Series regressions – ARIMA Model

```
#Time Series between IRIWeek and Total_Price Sales
 #Time Series - MA model
#Training Dataset
rm(list=ls())
install.packages("tseries")
library(tseries)
library(xts)
install.packages("forecast")
library(forecast)
install.packages("TTR")
library("TTR")
library(ggplot2)
setwd("...\\USF\\MIS\\SDM\\Final Project")
traindata<-read.csv("Final_Data_Calibration_Training.csv")
dim(traindata)
names(traindata);
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$Total_price[j] <- sum(traindata[which(IRIWeek == i),c("Price")])
    }
  }}
attach(traindata)
#Filtering the training dataset to have unique ans sorted rows on IRIWeek and Total_price
duplicates = duplicated(IRIWeek,Total_price)
duplicates[1:10]
unique_traindata <- traindata[!duplicated(traindata[c("IRIWeek","Total_price")]),]
sorted_traindata <- unique_traindata[order(unique_traindata$IRIWeek),]
#plotting 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")
#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$Total_price, alternative="stationary", k=0)
#p-value is 0.01 i.e. H0 is rejected and hence alternate hypothesis holds true. This means the data is stationary and
hence requires MA(Moving Average) model for analysis.
adf.test(sorted_traindata$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 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
#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.
detach(foredata)
rm(foredata)
setwd("......\\USF\\MIS\\SDM\\Final Project")
foredata<-read.csv("Final_Data_Forecast.csv")
dim(foredata)
names(foredata);
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
```

Appendix 8: Sale of Product according to Size and according to Formula

```
rm(list=ls())
setwd("....\\USF\\MIS\\SDM\\Final Project")
mydata<-read.csv("Final_Data_Calibration_Training.csv", header = TRUE)
dim(mydata)
names(mydata)
attach(mydata)
with(mydata, table(mydata$Brand,mydata$Size))
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")
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)
```

Appendix 9: Loyalty Check for SKU of the Brand

```
#Loyalty Check Script and Number of SKUs per Brand
#working Code and file below
rm(list=ls())
 setwd("....\\USF\\MIS\\SDM\\Final Project")
 #The below file contains additional Column "Loyalty" that has "0" or "1" value in case the customer/HHId opted for
same SKU in the Calibration dataset.
 #For this, I've used Excel formula. I've combined both HHId and SKU data into one column using "=C2&" "&D2" for
both files i.e. Calibration having 4418 rows(into column 'AK') and Forecast having 2137 rows(into column 'AM') into
single excel sheet.
 #I've also copied and pasted all SKU column from Calibration dataset into Column 'AL' of this combined sheet.
 #After this, I've 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.
    finaldata<-read.csv("Final_Data_Forecast_Backup_Sorted.csv")
    dim(finaldata)
    names(finaldata);
    attach(finaldata)
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")
```