A Handbook of Market Mix Model

Stepwise Guide to Complete Modeling Exercise

Compiled By
Soumendu Bhattacharya

Contents

- **Data Transformation / Roll-up**
    - **What is 'Level' of Data**
    - **Data Roll-up Using SQL**
    - **Using SQL and SAS Data Steps for Roll-up**
    - **Using 'Transpose' Function in SAS**
    - **Using Functions in SQL, Effect of Missing Value**
    - **Creation of 'Promo Flags'**
    - **Auditing the Rolled-up Data**
    - **Final Form of Transformed Sales and Promo Data**

- **Capturing Promotion Decay**
    - **What is 'Decay'?**
    - **Calculation of Decay**
    - **Effect of Decay on Promo Sales**
    - **Adding Decay into Promo Data**

- **Eliminating Fluctuation - Seasonality Calculation**
    - **Fluctuations: Seasonal, Cyclical and Random Effects**
    - **Why Smoothening Fluctuation**
    - **Ways to Eliminate Fluctuation: Seasonality Index Creation**
    - **Addding SI Into Analytic Dataset**

- **Modeling Dataset Creation**
    - **What is MODELING Dataset?**
    - **Level of Modeling Dataset**
    - **Contents of a Modeling Dataset**
    - **Auditing the Modeling Dataset**
    - **Sample Modeling Dataset**

Chapter 3: Modeling

- **Fundamentals of Mix Modeling**
- **Why Regression is Important Before Mix Modeling**
- **Regression Procedure: Sample Data, Syntax and Output in SAS**
- **Interpretation of Output: Removal of Multicolinearity**

- **Iterations in Regressions**
- **Final Model Selection Criteria**
- **Proc Mix Procedure: Sample Data, Syntax and Output in SAS**
- **Final Model Creation: Modeling Equation**
- **Validation (in-sample and out-sample)**
- **Deviation in Prediction: MAPE**

Chapter 4: Deliverables

'Predicted Sales' vs. 'Volume-due-to'

- **Calculating Predicted Sales**
- **Calculating Contribution (volume-due-to) from Significant variables**

Return on Investment Calculation

- **What is ROI?**
- **ROI for Different Market Mix Components**
- **Calculation procedure**

Optimum Mix identification

Interpretation and Presentation Basics

Chapter 1: Introduction

- What is Market Mix Model?
- Why Market Mix Model? Advantage Over Other Procedures
- Criterion of Market Mix Model Application
- Major Industries Where Market Mix Model is Important
- Final Output of a Market Mix model

Chapter 1
Introduction

What is Market Mix Model?

Market Mix Modeling is the application of econometrics to identify the volume and profit contribution of each individual marketing activity and external factor.

To comprehend the competitive structure of a market, it is important to understand the short-run and long run effects of the marketing mix on market shares.

This approach provides not just knowledge of the sales returns from each marketing activity, but also allows advice on how these activities can be improved to generate more sales.

Brand plans built on this basis confidently meet objectives whether these are profitability, value share, or volume.

The explanation of past brand performance uses all available past experience to provide an understanding of the overall dynamics of brand and market. Key competitors are identified through sales gains and losses and insight is gained into how consumers purchase the product category.

This includes identifying the segmentation underlying the consumers' purchase decision process.

The range of marketing activities that measurably add to brand sales and can be individually evaluated includes:

- Media - TV, press, outdoor etc.
- Intense marketing - TV + door drops
- Promotions - price, value added etc.
- Point of sale display
- Direct response
- In-store demos
- Price

Why Market Mix Model? Advantage Over Other Procedures

The advantage of Market Mix Models over other predictive/ diagnostic models lies in its power to answer the critical queries like…

- How do you measure the return on investment of each marketing dollar you spend?
- How can you predict the likely return you can expect from future marketing investments?
- How do you know how your sales will be affected by an increase or decrease in your marketing budget?

These models -

- let you look backwards in time to determine ROI for marketing tactics
- help you diagnose the reasons for changes in marketing performance over time.
- let you look forward for forecasting and to create what-if scenarios for response planning

Criterion of Market Mix Model Application

The Market Mix models can be developed for any product / brand if

1. the product/ brand is price and promo elastic. In case the demand of the product is not influenced by the price / promotions (essential commodities), this modeling technique can't be applied
2. the product / brand should have substitutes available in the market. Otherwise we can never judge impact of promo in competitor products/ brands over the products/brands in consideration
3. there are sales data available when 'no promo was on' to calculate baseline
4. detailed data on extent and duration of promo, campaign, TRP and all such factors impacting sales available over a considerable period of time (at least 1 full year to calculate and eliminate seasonality effects)
5. sales and promo data for the brand/ product in consideration are available on daily / weekly basis along with competitors' data

Major Industries Where Market Mix Model is Important

- Consumer Packaged Goods
- Pharmaceuticals
- Financial Services
- Automotive

Final Output of a Market Mix model

The Market Mix Model essentially presents sales of a brand/ product as a multiplicative function of price, seasonality and promotions. However, the major takeaways for the business are ‘volume-due-to’ and ‘ROI’ figures.

The volume-due-to captures the portion of sales that was generated as a result of the particular price change / promotion efforts. Once the volume is converted into monetary form and compared to the cost of such promotions / price changes, we get the Return-On-Investment (ROI) figures for each promotional activity.

This proven analytical approach lets us improve the effectiveness of your marketing through:

- Better allocation of marketing investments. Gain more “bang for the buck” by recommending spending tactics and continuous optimization to maximize profit or volume growth.
- Insight into how to improve your business. Quantify the impact of key volume drivers: price, competition, weather, trade support, marketing, etc.
- Improved business planning
- Confidently deliver forecasts: anticipate the consumer response to marketing activity and marketplace drivers.
- Leverage synergies between different products/brands. Capture any “halo” effects to take advantage of the impact investing in one brand may have on another.

Chapter 2: Starting Point

- **2.1: Different Types of Data Requirement**

    - **Sales/ Transaction Data**
    - **Promotion Data**
    - **Store Location/ Geographic Data**
    - **Brand / Product Indentifier**

- **2.2: Data Mining, Auditing**

    - **Various Forms of Data**
    - **Common Issues Faced During Data Conversion**
    - **Audit Steps on Data Conversion: Test Data/ Code Review / Auditing Through Excel**
    - **Validation of Data Contents and Continuity**
    - **Error Detection at Data, Corrections**

- **2.3: Capping and Missing Value Treatment**

    - **Missing Values vs. Extreme Values**
    - **When Keeping Missing Values are Important**
    - **Univariate Analysis**
    - **Missing Value Treatment**
    - **Difference Between Missing Value Treatment and Capping**
    - **Logic Behind Capping**
    - **Common Capping Practise**
    - **Capping Syntax: Use of Excel for SAS Coding**
    - **Example of Uncapped and Capped Data: Effect on Distribution**
    - **Post Capping Auditing: Checklists**

- **2.4: Data Transformation / Roll-up**

    - **What is 'Level' of Data**
    - **Data Roll-up Using SQL**
    - **Using SQL and SAS Data Steps for Roll-up**
    - **Using 'Transpose' Function in SAS**
    - **Using Functions in SQL, Effect of Missing Value**
    - **Creation of 'Promo Flags'**
    - **Auditing the Rolled-up Data**
    - **Final Form of Transformed Sales and Promo Data**

- **2.5: Capturing Promotion Decay**
    - **What is ‘Decay’?**
    - **Calculation of Decay**
    - **Effect of Decay on Promo Sales**
    - **Adding Decay into Promo Data**

- **2.6: Eliminating Fluctuation - Seasonality Calculation**
    - **Fluctuations: Seasonal, Cyclical and Random Effects**
    - **Why Smoothening Fluctuation**
    - **Ways to Eliminate Fluctuation: Seasonality Index Creation**
    - **Addding SI Into Analytic Dataset**

- **2.7 Modeling Dataset Creation**
    - **What is MODELING Dataset?**
    - **Level of Modeling Dataset**
    - **Contents of a Modeling Dataset**
    - **Auditing the Modeling Dataset**
    - **Sample Modeling Dataset**

Chapter 2
Starting Point

Different Types of Data Requirement

- Sales/ Transaction Data :

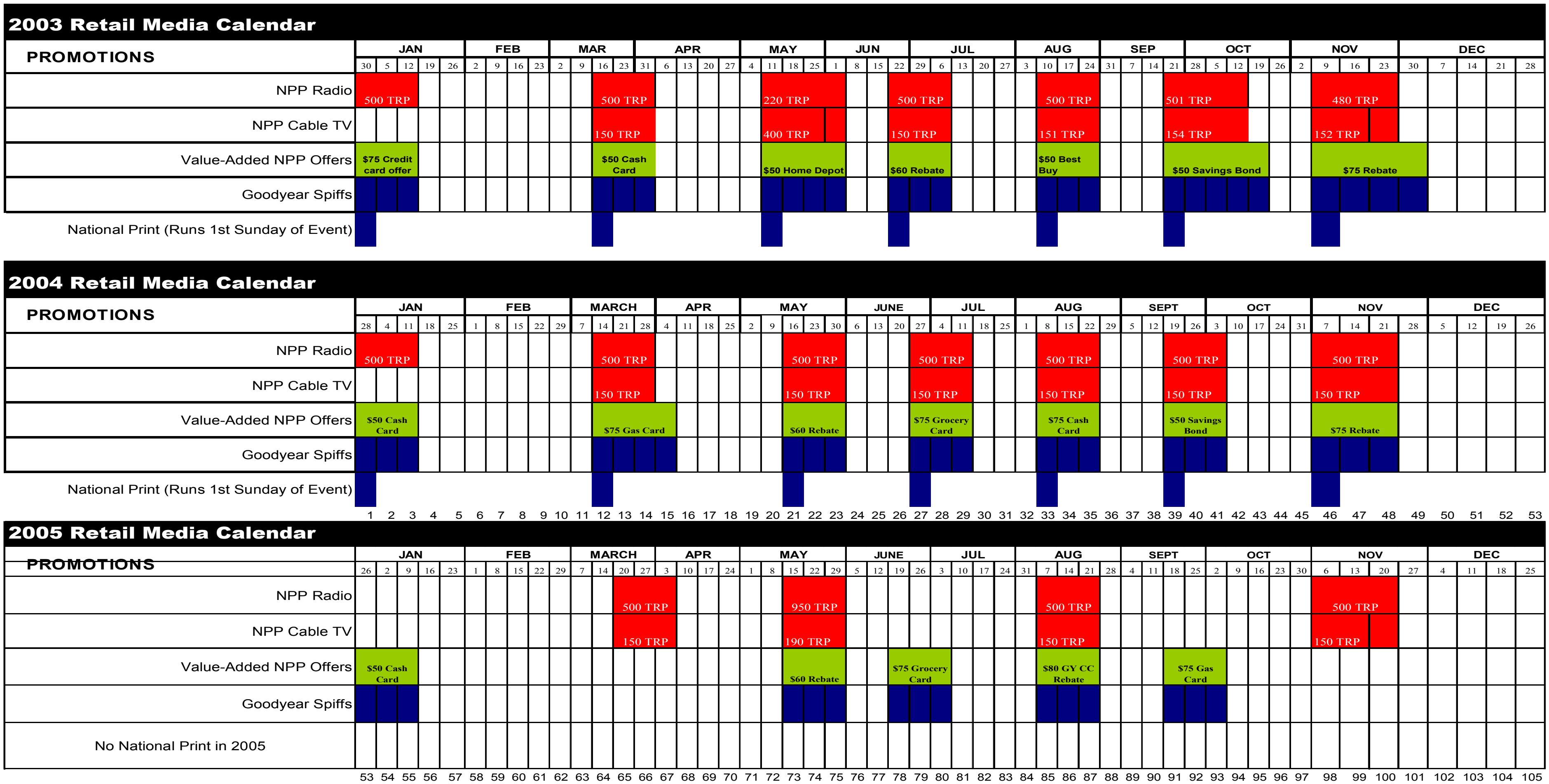
Market mix modeling needs time series data on sales volume as well as sales revenue. In most of the cases the data needs to be at daily / weekly level for all the stores separately. The level of granularity in transaction data is essentially based on the time and duration of promotions. If all promotions run from say Sunday to Saturday, then summarized weekly data (from Sunday to Saturday) can be used. However, if the promotions don't follow any such fixed routine, then use of daily data is preferred.

The data is normally available at invoice level. We need to roll it up for different brands / products for each store for each date separately. However, if the client supplies the summarized data in the desired level, the roll-up exercise becomes redundant

TypeCode	Region	District	Nonsig	OpenStatus	WeekNumber	InvoiceDate	ProductNumber	ClassOfBusiness	ItemDesc	Invoices	NumberUnits	TotalItemPrice	DiscountAmt	NumberTires	TotalTirePrice	TotalServicePrice	UniqueID
J	1103	0722	901270	Y	1	1/1/2003	041263000	V		1	1	3	0	0	3		2527
J	1103	0722	901270	Y	1	1/1/2003	044263000	V		1	1	10	0	0	10		2528
J	1103	0722	901270	Y	1	1/1/2003	071000000	V		1	1	2.5	0	2.5	0		2529
J	1103	0722	901270	Y	1	1/1/2003	347000105	V		1	1	128	1	128	0		2530
J	1103	0722	901260	Y	1	1/2/2003	040000000	2		2	2	0	0	0	0		46583
J	1103	0722	901260	Y	1	1/2/2003	040101000	2		1	4	50	0	0	50		46584
J	1103	0722	901260	Y	1	1/2/2003	040265000	2		8	8	90	0	0	90		46585
J	1103	0722	901260	Y	1	1/2/2003	040265000	M		1	1	15	0	0	15		46586

- Promotion Data:

Typically the promotion calendar is used as the input of market mix modeling exercise. The dates when these promotions run are flagged suitably. Different kinds of promos are flagged separately. Ideally there should not be overlap of 2 promotions on a particular day; else the effects from these promotions can't be calculated separately.



2003 Retail Stores Advertising Calendar

HF&A

PROMOTIONS	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
	30 5 12 19 26	2 9 16 23	2 9 16 23	31 6 13 20 27	4 11 18 25	1 8 15 22	29 6 13 20 27	3 10 17 24	31 7 14 21	28 5 12 19 26	2 9 16 23	30 7 14 21 28
National Brand Advertising Wings		2/16 - 3/15		4/20 - 5/17					8/31 - 9/27			
Save More Credit Card		Save More $20 $40 off			Save More $20 $40 off			Save More $20 $40 off			Save More $20 $40 off	
Value-Added NPP Offers	$75 Credit card offer		$50 Cash Card		Home Depot/ Best Buy rebate	$60,$40,$20 Eaglewrangler aqua		$50 best buy	$50 savings		$50 Cash Card	
NPP Support W/ Direct Mail												
Local Saturation Direct Mail												
Gemini — Seasonal Packages			Spring Car Care March 1- May 11		Car care Vac Days May 19 - July 26				Fall Car Care Sept 1 - Nov 22			
				Gemini Brand Campaign					Gemini Brand Campaign			
Website												
VIP Responder Program												
Database Marketing — Service Reminder												
1st-Time Customer												
Lapsed Customer												
New Movers												
nds and Family/Peel a Deal						Peel a Deal						
Thank You Card (Hallmark)				In Home 12/30, 1/6, 2/10 2/17 or 2/10 2/17 3/2 3/16 exp 3/31 0r 4/19						500 per store drops Sept 1		
SPIKE DAYS SALES	16 20	13 17	27 31	26 30	14 17	26 30	24 28	14 18	26 30	16 20	13 17	11 15
Hallmark Birthday Card												

Drop Date Continuity

- Store Location/ Geographic Data

The Geographic data typically captures the city / zip / state location etc. mapping. This data is required to summarize the finding at different geographic levels.

Type Store	Terr	Manager	Street Address	City	ST	Zip Code	County	Phone Nbr	Fax Nbr	DMR #	DMR Name	DMR E-Mail
ASC	0128	PIECHNIK, JEFFREY	801 NEW LOUDON RD #2	LATHAM	NY	12110	ALBANY	518-7854151	-	3168	KEN WETZONIS	KENWETZONIS@GOODYEAR.COM
ASC	0131	OBRIEN, THOMAS	46 48 WOLF RD	ALBANY	NY	12205	ALBANY	518-4599122	518-4599122	3168	KEN WETZONIS	KENWETZONIS@GOODYEAR.COM
ASC	0132	RAGOSTA, CHARLES	3713 STATE ST	SCHENECTADY	NY	12304	SCHENECTADY	518-3748342	518-3746070	3168	KEN WETZONIS	KENWETZONIS@GOODYEAR.COM
ASC	0142	FERGUSON, STANLEY	RT 44 PLAZA SHOPPING	POUGHKEEPSIE	NY	12603	DUTCHESS	845-4858430	914-4853294	3182	DICK JOHNSTON	RICHARD_T_JOHNSTON@GOODYEAR.COM
ASC	0220	WADE, FREDERIC	6034 BALT. NAT. PIKE	CATONSVILLE	MD	21228	BALTIMORE	410-8699200	410-8699274	3193	KARL KAMPLAIN	KARL.KAMPLAIN@GOODYEAR.COM
ASC	0222	CALLOW, NICOLA	8667 PHILADELPHIA RD	BALTIMORE	MD	21237	BALTIMORE	410-7804441	410-7804696	3193	KARL KAMPLAIN	KARL.KAMPLAIN@GOODYEAR.COM
ASC	0223	WILLIAMS, ROBERT	3156 BLADENSBURG RD	WASHINGTON	DC	20018	DISTRICT OF COLUMBIA	202-5263885	202-2690708	3193	KARL KAMPLAIN	KARL.KAMPLAIN@GOODYEAR.COM
ASC	0225	REYNOLDS JR, DONALD	1400 EASTERN BLVD	ESSEX	MD	21221	BALTIMORE	410-6878212	410-6879271	3193	KARL KAMPLAIN	KARL.KAMPLAIN@GOODYEAR.COM
ASC	0226	MUNRO, MILDRED	2212 BELAIR ROAD	FALLSTON	MD	21047	HARFORD	410-6387180	410-6387906	3193	KARL KAMPLAIN	KARL.KAMPLAIN@GOODYEAR.COM

- Brand / Product Indentifier

In most cases the products and brands are identified by some identifier (codes). Based on the requirement / project scope, the market mix modeling is carried out at different brand / product/ overall level. The complete mapping among these helps in rolling up the data at suitable category level.

prodcdT	Prodcd	Prod-cd	DD	Desc	Brand	Reduced Brand	Size	reduced size	PBU	Market Area	Market Group	Product Group	Product Line	HIERARCHY	RimD	UTQG
178462286	178462286	178-462-286		260/80R20 00 106A8 SUP TRAC RADDTL	GOODYEAR	GOODYEAR	260/80R20	260-8020	Farm	Farm - MA	Radial Rear Farm	R1W-Radial	Super Traction Rd	20200840247002000	20	
178042324	178042324	178-042-324		385/85R34 00 141G DYNA TOR RAD TL	GOODYEAR	GOODYEAR	385/85R34MPT	385-8534MT	Farm	Farm - MA	Radial Rear Farm	R1-Radial	Dyna Torque Radial	20200840246001800	34	
178473324	178473324	178-473-324		320/85R34 00 132D DYNA TOR RAD TL	GOODYEAR	GOODYEAR	320/85R34	320-8534	Farm	Farm - MA	Radial Rear Farm	R1-Radial	Dyna Torque Radial	20200840246001800	34	
178469762	178469762	178-469-762		380/90R54 00 170A8B DT800 TL	GOODYEAR	GOODYEAR	380/90R54	380-9054	Farm	Farm - MA	Radial Rear Farm	R1W-Radial	DT800 Radial	20200840247002000	54	
178468286	178468286	178-468-286		520/85R46 00 169A8B SUP TRAC RAD TL	GOODYEAR	GOODYEAR	520/85R46	520-8546	Farm	Farm - MA	Radial Rear Farm	R1W-Radial	Super Traction Rd	20200840247002000	46	
178466762	178466762	178-466-762		320/90R54 00 149A8B DT800 TL	GOODYEAR	GOODYEAR	320/90R54	320-9054	Farm	Farm - MA	Radial Rear Farm	R1W-Radial	DT800 Radial	20200840247002000	54	
178465763	178465763	178-465-763		260/70R20 00 113A8 DT810 TL	GOODYEAR	GOODYEAR	260/70R20	260-7020	Farm	Farm - MA	Radial Rear Farm	R1W-Radial	DT810 Radial	20200840247002000	20	

Data Mining / Auditing:

- Various Forms of Data

The raw data normally are available in the following formats

1. .dat file (flat file with fixed locations)
2. .csv files (comma separated value)
3. Excel file
4. SPSS file
5. SAS dataset
6. Access database (.mdb)

We need to convert these input files into SAS environment by suitable technique.

- Common Issues Faced During Data Conversion

While converting the raw input files into SAS datasets following issues may arise

1. **Data truncation**: If we import the CSV files directly into SAS, by default the format of the variables in the first row is accepted as the format for all the records for those variables. Due to this, if the first row has some character variable of length 5, the next record imported for the same variable also gets a length of 5 characters. If the actual record has more characters then only the first 5 characters appear
2. **Missing values**: Missing values generate if we don't have any observation in a variable. Such missing values, if not treated properly, can result in computational errors.
3. **Columns misplacement**: This is a typical problem when excel files are imported into SAS directly that has some values missing in some variable(s). The missing cells sometimes get replaced by the next variable's value if available. This leads to creation of erroneous data for analysis.
4. **Issues due to special character / spaces / 'enter's**: SAS considers space as separators. If we don't specify any length for the input variables, the character values in the variable having space within are placed in two different columns.

We need to audit the data carefully to ensure the conversion is correct.

- Audit Steps on Data Conversion:
  - Using PERL script for reading the variable format and length before importing into SAS

    Programme (PERL Script) developed by Mu Sigma team is available internally.

  - Test Data/ Code Review / Auditing Through Excel
  -

    While converting data from different forms into SAS intially consider part of it than full and generate the required code. Specifying of variables names, their formats (numeric or character) and their sizes are to be done in required form so that no data is missed, or misread.

- Validation of Data Contents and Continuity

  Verify whether the number of records read into SAS are equal to the raw data record nos.

  Verify if they are any duplicates in the dataset.

  Verify whether there are any missing values in the imported dataset.

  Verify whether the data is populated in correct format while merging

- Error Detection at Data, Corrections

If no. of records read into sas are not equal to the actual number,check the raw dataset and code used while importing. Duplicates are verified by using ‘**nodupkey’** option in SAS and if any duplicate record is found then they are deleted them from the dataset.

In case of missing values records are either ‘treated’ or they are removed from the dataset based on the particular situation.

Capping and Missing Value Treatment

- Missing Values vs. Extreme Values

Missing values are generated when there is absence of one or more entries in the variable column. The variable can be numeric / character. In SAS environment the missing values in character variable are denoted as a blank entry (‘’). The numeric missing values are presented by a dot (.) and is interpreted in SAS an infinitely negative number (- ∞).

Extreme values are those which don’t fit into the natural distribution of the variable. For example if we have ‘date of birth’ as one variable and we have calculated age as of today and suppose in few cases the year of birth is wrongly printed in 15^{th} century, the age will have some huge values which are wrong. Similarly while dealing with sales data, there can be sales amount in millions for 1-2 transactions that affect the entire analysis. Though the entry may be right, it doesn’t capture any normal behavior of the sample. Such values are called as extreme values.

- Univariate Analysis

Univariate analysis is done to check the distribution and central tendencies of the variables and to see the # missing values / presence of extreme values. Post univariate analysis we do the necessary capping / missing value treatment.

- Missing /Extreme Value Treatment

Missing values generate lot of computational errors if not addressed properly. For example, if we want to take ratio between two variables and the variables in denominator have missing value, then division by missing values generate unexpected results.

In some cases* we replace missing value by zero to avoid any mathematical computational issues. However based on the nature of the variable (like age) we might use some representative number (mean / median) to treat the missing values.

For extreme values we use the method called ‘capping’ (explained later below).

- *When Keeping Missing Values are Important

At times we need to KEEP the missing values in its original form. This is mostly done when we need to calculate average / standard deviation etc using SAS syntax. In such cases the missing values are by default not considered by SAS. But if we replace the missing values by zero, such values WILL BE considered by SAS as valid entries and the average calculated will be lesser than actual.

- ## Difference Between Missing Value Treatment and Capping

Missing Value Treatment and Capping are at times used interchangeably but they are quite different. They are two different approaches to tackle two different kinds of issues.

Detection of missing value is fairly easy. Missing values are mostly replaced by zero / mean / median. Detection of all extreme values involve lot more effort and quite a number of iterations. We might need to look at the univariate distribution and identify the extreme values. After treating them properly, we might need to recheck the univariate distribution again and redo the same exercise to rule-out presence of any further extreme values.

- ## Logic Behind Capping

If we don't 'cap' the extreme variables properly, following errors may occur

a) huge impact on the basic statistic like mean / standard deviation that can affect the entire diagnosis / insight gained from the data
b) increase the residual errors
c) rejection of a good model on fitment ground

- ## Common Capping Practise

In case of extreme values, to ensure the normality of the distribution (in large sample cases the distribution becomes normal / quasi-normal) we need to 'treat' them properly. The extreme values can be negative / positive. Based on the nature of the variable we need to suitably decide whether we need to 'bin' it at some point. We need to look at the distribution of the variable and decide the lowest / highest values we can accept for that variable that will not impact the normalcy of the distribution. A combination of max/min/mean etc can be used to cap a variable effectively

- ## Capping Syntax: Use of Excel for SAS Coding

The following code generates the capping code that can be used in SAS directly (the max and min to be decided after the univariate test and removing outliers)

	Col A	Col B	Col C	Col D	Col E	Col F
Row#	Variable	Format	Min	Max	Excel Formula	Capping Code to be used in SAS
3	asfin33	Num	0	379	CONCATENATE(A3," ="," MIN","(","MAX","(",A3,",",C3,")",",",D3,")",";")	asfin33 = MIN(MAX(asfin33,0),379);
4	avgcram12	Num	0	3,780	CONCATENATE(A4," ="," MIN","(","MAX","(",A4,",",C4,")",",",D4,")",";")	avgcram12 = MIN(MAX(avgcram12,0),3780);
5	avgcram3	Num	0	3,942	CONCATENATE(A5," ="," MIN","(","MAX","(",A5,",",C5,")",",",D5,")",";")	avgcram3 = MIN(MAX(avgcram3,0),3942);
6	avgcram6	Num	0	2,736	CONCATENATE(A6," ="," MIN","(","MAX","(",A6,",",C6,")",",",D6,")",";")	avgcram6 = MIN(MAX(avgcram6,0),2736);
7	avgcram9	Num	0	3,828	CONCATENATE(A7," ="," MIN","(","MAX","(",A7,",",C7,")",",",D7,")",";")	avgcram9 = MIN(MAX(avgcram9,0),3828);

- Example of Uncapped and Capped Data: Effect on Distribution

Distribution of Uncapped variable

```
                    The SAS System         13:17 Tuesday, July 25, 2006  10

                    The UNIVARIATE Procedure
                 Variable:  wtdPrice_Associate

                             Moments

N                      133508    Sum Weights             133508
Mean               45.8075877    Sum Observations    6115679.42
Std Deviation      11.6398839    Variance            135.486896
Skewness           1.06068607    Kurtosis            2.55571652
Uncorrected SS      298232971    Corrected SS        18088449.1
Coeff Variation    25.4103838    Std Error Mean      0.03185628


                   Basic Statistical Measures

         Location                    Variability

     Mean     45.80759     Std Deviation           11.63988
     Median   44.41566     Variance               135.48690
     Mode     35.00000     Range                  169.32500
                           Interquartile Range     13.57583


                   Tests for Location: Mu0=0

        Test           -Statistic-    -----p Value------

        Student's t    t  1437.945    Pr > |t|    <.0001
        Sign           M     66754    Pr >= |M|   <.0001
        Signed Rank    S  4.4561E9    Pr >= |S|   <.0001


                    Quantiles (Definition 5)

                    Quantile       Estimate

                    100% Max       172.9500
                    99%             83.0000
                    95%             67.0000
                    90%             60.3333
                    75% Q3          51.5758
                    50% Median      44.4157
                    25% Q1          38.0000
                    10%             32.8333
                    5%              29.8889
                    1%              24.8571
                    0% Min           3.6250
```

```
                    The UNIVARIATE Procedure
                 Variable:  wtdPrice_Associate

                      Extreme Observations

             ------Lowest-----        -----Highest----

               Value      Obs          Value      Obs

             3.62500   103175         130.00    30915
             5.94636   129428         130.00    42105
             6.00000    57406         130.00    67722
             6.00000    43519         133.00    28231
             7.42000    50695         172.95    48686


                Histogram                          #             Boxplot
 175+*                                             1                *
    .
    .
    .
    .*                                             4                *
    .*                                             6                *
    .*                                            42                *
    .*                                           148                *
    .*                                           475                0
    .**                                         1134                0
    .***                                        3111                0
    .*********                                  9136                |
    .*************************                 25620             +-----+
    .************************************************ 50661      *--+--*
    .***********************************       36386             +-----+
    .*******                                    6531                |
    .*                                           244                0
   5+*                                             9                0
     ----+----+----+----+----+----+----+----+----+---
     * may represent up to 1056 counts
```

The UNIVARIATE Procedure
Variable: wtdPrice_Associate

```
                     Normal Probability Plot
175+                                                  *
   |
   |
   |
   |                                                  *
   |                                                  *
   |                                                  *
   |                                                  *
   |                                                  *
   |                                               ****
   |                                           *****+++
   |                                      ******++
   |                             +*********
   |                    ***********
   |         ************
   |**********++
   |*++
  5+*
    +----+----+----+----+----+----+----+----+----+----+
        -2        -1         0        +1        +2
```

Post removing the outlier, this is how the distribution (univariate) would look like

The UNIVARIATE Procedure
Variable: wtdPrice_Associate

Moments

N	131303	Sum Weights	131303
Mean	45.4858692	Sum Observations	5972431.08
Std Deviation	10.5134949	Variance	110.533576
Skewness	0.66681492	Kurtosis	0.46042063
Uncorrected SS	286174499	Corrected SS	14513279.6
Coeff Variation	23.1137607	Std Error Mean	0.02901415

Basic Statistical Measures

Location		Variability	
Mean	45.48587	Std Deviation	10.51349
Median	44.35672	Variance	110.53358
Mode	35.00000	Range	58.96986
		Interquartile Range	13.25083

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	1567.714	Pr > |t|	<.0001
Sign	M	65651.5	Pr >= |M|	<.0001
Signed Rank	S	4.3102E9	Pr >= |S|	<.0001

Quantiles (Definition 5)

Quantile	Estimate
100% Max	82.9990
99%	75.8000
95%	65.3000
90%	59.6360
75% Q3	51.3300
50% Median	44.3567
25% Q1	38.0792
10%	33.0000
5%	30.1500
1%	26.0000
0% Min	24.0291

The UNIVARIATE Procedure
Variable: wtdPrice_Associate

Extreme Observations

------Lowest-----		-----Highest----	
Value	Obs	Value	Obs
24.0291	133479	82.950	62951
24.0357	43743	82.990	55242
24.0625	74544	82.990	58800
24.0769	10288	82.990	62416
24.0909	129247	82.999	56992

```
                        Histogram                               #             Boxplot
82.5+*                                                         439               0
    .***                                                      1144               0
    .****                                                     1967               0
    .*******                                                  3378               |
    .***********                                              5758               |
    .******************                                       9686               |
52.5+*****************************                           15934            +-----+
    .********************************************            24169            |  +  |
    .************************************************        26492            *-----*
    .******************************************              23152            +-----+
    .************************                                13234               |
    .**********                                               5393               |
22.5+**                                                        557               |
     ----+----+----+----+----+----+----+----+----+---
     * may represent up to 552 counts
```

```
                    Normal Probability Plot
82.5+                                                 *
    |                                               ***
    |                                           *****++
    |                                        ****++++
    |                                     ****++
    |                                  ****
52.5+                             ******
    |                        +*****
    |                   *******
    |             *******
    |       *******+
    |*********++
22.5+*+++++
     +----+----+----+----+----+----+----+----+----+----+
         -2        -1         0        +1        +2
```

- Post Capping Auditing: Checklists
    - Run the univariate code again and look for any discontinuity
    - For each variable count the cases where the variable is missing (=.)

Ideally, there will be no missing variable left and the count will be zero. If its otherwise, revisit the exercise.

Data Transformation / Roll-up

- What is ‘Level’ of Data

The level of a data indicates the level of granularity at which the data is unique. For example, if we have daily sales data from a retail chain which presents the date of sales, items bought / returned, name (or some identifier) of the buyer etc, the data can be unique at the product level (that is the lowest level of granularity captured at the data). An example of such a dataset is presented below. As we can see, the combination of customer id, date, invoice number, product number is unique for each record.

Customer Id	Sales date	Invoice number	Product code	Price
AAAAA	1-Jan-06	1234	AB123	235
AAAAA	1-Jan-06	1234	AS234	234
AAAAA	11-Jan-06	1125	AS346	124
AAAAA	11-Jan-06	1126	FG563	345
AAAAA	21-Jan-06	1149	HJK45	76
AAAAA	21-Jan-06	1912	FJK78	2
AAAAA	29-Jan-06	2421	AB123	235
BBBBB	8-Feb-06	1599	JK534	32
BBBBB	8-Feb-06	1599	HJK45	76
BBBBB	8-Feb-06	1599	DF566	56
BBBBB	18-Feb-06	1491	SEF45	36
BBBBB	8-Mar-06	1514	GF732	234

In this case, as can be seen, there are repetitions of first 3 variables. However, the lowest level of granularity is captured at product level; and the combination of Customer id, Sales date, invoice number and Product code is unique for each row. Hence we say the level of data is at ‘product level’.

- **Data Roll-up Using SQL**

Using SQL, we can summarize (roll-up) the data at the desired level (customer. Sales date, invoice number, product code etc.). This can be done at combination level as well (like total daily sales by each customer, total sales for each product in a given period etc.). We need to understand the time and other dimensions we want to capture and baased on that we need to decide the level at which we want to roll-up the data.

- **Using SQL / SAS Data Steps for Roll-up**

SQL is always preferred for data roll-up due to the options available for level of data to be rolled up (“group by” function). This flexibility is not available in SAS. Below is one example.

Raw Data

week	customer_number	sales
1	1234	456
1	2345	46
1	2312	245
2	2345	564
2	4566	458
2	2334	122
3	1234	212
3	7645	89
3	5655	123
4	4232	124
4	1234	466
4	3424	566

SQL Code for rolling up the data at weekly level

```
Proc sql;
Create table weekly_summary as select
week,
      count(distinct customer_number) as unique_customer,
      count(case when sales>0 then customer_number end) as
total_transaction,
      sum(sales) as weekly_tot_sales

from example
      group by week;
      quit;
```

Output from SQL

week	unique_customer	total_transaction	weekly_tot_sales
1	3	3	747
2	3	3	1144
3	3	3	424
4	3	3	1156

- **Using 'Transpose' Function in SAS**

  At times we need to transpose the data to increase the level of the data while summerizing some variables. For example, suppose the data is available to us at store, date & brand level (the data is unique at store + date + brand combination level). We need to data to be summarized at date level while capturing the summarized form of brand level info at different columns. Here we need to use PROC TRANSPOSE in SAS. Below is one example.

Suppose the raw data is available to us as follows

Type Code	Region	District	Nonsig	Invoice Date	mega_sub_brand	wtd_avg
ASC	1211	3168	900321	1/2/2003	Assurance	67.67
ASC	1211	3168	900321	1/2/2003	Foreign	95
ASC	1211	3168	900324	1/2/2003	Eagle	94
ASC	1211	3168	900324	1/2/2003	GY_other	44.18
ASC	1211	3168	900324	1/3/2003	Wrangler	68.67
ASC	1211	3168	900325	1/2/2003	Assurance	80
ASC	1211	3168	900325	1/2/2003	Eagle	106
ASC	1211	3168	900325	1/2/2003	Wrangler	132
ASC	1211	3168	900351	1/6/2003	Assurance	204
ASC	1211	3168	900351	1/6/2003	Dunlop	59
ASC	1211	3168	900354	1/3/2003	Assurance	65.6
ASC	1211	3168	900354	1/3/2003	Foreign	53
ASC	1211	3168	900354	1/3/2003	Republic	41

Suppose we need the data to be unique at store + date level in the following format

Type Code	Region	District	Nonsig	Invoice Date	wtdPrice_ Assurance	wtdPrice_ Foreign	wtdPrice_ Eagle
ASC	1211	3168	900321	1/2/2003	xx	xx	xx
ASC	1211	3168	900324	1/2/2003	xx	xx	xx
ASC	1211	3168	900324	1/3/2003	xx	xx	xx
ASC	1211	3168	900325	1/2/2003	xx	xx	xx

Here is the SAS code for transposing the variables

```
proc transpose data=base out=base_tranposed prefix=wtdPrice_;
      by TypeCode Region District Nonsig InvoiceDate;
    id mega_sub_brand;
    var wtd_avg;
      run;
```

And here is the output…

TypeCode	Region	District	Nonsig	InvoiceDate	wtdPrice_Assurance	wtdPrice_Foreign	wtdPrice_Eagle	wtdPrice_GY_other	wtdPrice_Wrangler	wtdPrice_Dunlop
ASC	1211	3168	900321	1/2/2003	67.67	95	.	.	.	.
ASC	1211	3168	900324	1/2/2003	.	.	94	44.18	.	.
ASC	1211	3168	900324	1/3/2003	.	.	.	.	68.67	.
ASC	1211	3168	900325	1/2/2003	80	.	106	.	132	.
ASC	1211	3168	900351	1/6/2003	204	.	.	.	.	59
ASC	1211	3168	900354	1/3/2003	65.6	53	.	.	.	.
ASC	1211	3168	900354	1/6/2003	.	61	.	.	.	91.29
ASC	1211	3168	900361	1/3/2003	.	.	87	.	.	.
ASC	1211	3184	901023	10/3/2005	.	.	.	.	.	139.95
ASC	1211	3184	901047	9/26/2005	.	.	.	.	.	104.36

- **Using Functions in SQL, Effect of Missing Value**

  In SAS / SQL, the missing values for any variable are not considered for any computation. For example, if there are 3 values like 10, 5 and missing (=.) and we try to calculate the average of that variable, the result will be 7.5 {=(10+5)/2}.

  If we blindly convert those missing values to zero then it will be considered as a valid number (= 0) and will be considered in the computations. Considering the same example presented above, the average now will be calculated as (10 + 5+ 0)/3 = 5.

  This shows how critical it is to handle the missing values with utmost care. They should be capped as zero ONLY WHEN we are sure the missing value means zero. Else at times we will prefer to have the missing value un-treated to avoid such computational errors.

- **Creation of 'Promo Flags'**

  The promotions are ideally captured as binary flags (1 / 0) at monthly / weekly / daily level. The level is mostly decided by promo implementaion level at the clients' end. If we have GRP data, then that info can also be added in the same data as a separate variable, ***AFTER DECAY ADJUSTMENT*** (See page 29 for details). A typical way to capture the promo at weekly level is as fillows.

Store	Year_Week	Year	week ending date	Promo 1	Promo 2	Promo 3	Promo 4	GRP Promo 1
900321	200400	2004	1/3/2004	1	0	0	1	70.0
900321	200401	2004	1/10/2004	0	0	0	0	91.0
900321	200402	2004	1/17/2004	0	0	1	0	27.3
900321	200403	2004	1/24/2004	1	0	0	1	113.2
900321	200404	2004	1/31/2004	0	1	0	0	34.0
900324	200410	2004	3/13/2004	0	0	0	0	10.2
900324	200411	2004	3/20/2004	0	0	0	0	3.1
900324	200412	2004	3/27/2004	1	0	0	1	105.9
900324	200413	2004	4/3/2004	1	0	0	1	171.8
900324	200414	2004	4/10/2004	1	0	1	1	191.5
900324	200415	2004	4/17/2004	0	0	1	0	57.5
900324	200416	2004	4/24/2004	0	1	0	0	17.2
900324	200417	2004	5/1/2004	0	1	0	0	5.2

  The promo data is available from the client either in form of promo calendar / execution files. All type of such input files are needed to be converted into the format presented above.

- **Auditing the Rolled-up Data**

  Same steps to be followed as discussed in "Common Issues Faced During Data Conversion" in "Data Mining / Auditing" section in page # 16.

- **Final Form of Transformed Sales and Promo Data**

  Here is a basic structure as to how the sales / promo / pricing and seasonality are to be clubbed together to create the analytical dataset.

Store	Year_Week	Year	week ending date	Sales for Brand 1	Sales for Brand 2	Promo 1	Promo 2	Weighted Price for Brand 1	Weighted Price for Brand 2	Seasonality Index
900321	200400	2004	1/3/2004	11	23	0	1	15.2	9.1	1.02
900321	200401	2004	1/10/2004	13	24	0	0	15	9.2	1.03
900321	200402	2004	1/17/2004	12	21	1	0	15	9.25	0.99
900321	200403	2004	1/24/2004	15	26	0	1	15.2	9.2	0.97
900321	200404	2004	1/31/2004	11	31	0	0	15.3	9.2	1.05
900324	200410	2004	3/13/2004	10	26	0	0	15.3	9.2	1.06
900324	200411	2004	3/20/2004	9	22	0	0	15.3	9.2	1.02
900324	200412	2004	3/27/2004	11	25	0	1	15.3	9.15	1
900324	200413	2004	4/3/2004	16	24	0	1	15.2	9.15	0.98
900324	200414	2004	4/10/2004	16	21	1	1	15.2	9.15	0.97
900324	200415	2004	4/17/2004	15	23	1	0	15.2	9.15	0.96
900324	200416	2004	4/24/2004	12	29	0	0	15.2	9.18	0.98
900324	200417	2004	5/1/2004	10	32	0	0	15.3	9.2	0.97

- **Capturing Promotion Decay**

What is 'Decay'?

In case of advertising in a media or through some promotional activity the customer reacts to the information and make a note of it in mind. But the inforamtion starts losing from his memory over time. The process of calculating the memory from the customers thought is known as Decay.

Calculation of Decay

Before we discuss about how we calculate the 'decay', lets understand the concept of decay in bit details with some example.
Suppose in a week, a particular Ad campaign has received 100 GRPs. In absence of any decay we can assume the effect of all those 100 decays will be realized in the same week itself. However, in practice, a 'part' of the effect of those campaign (GRPs) are realized in the same week and the remaining parts are realized over a period of time (like a GP series). Suppose we assume a decay rate of 60% (Adstock=0.4), we relize the effect of 60 GRPs in the first week. Again 60% of the remaining 40 GRPs (= 24 GRPs) in the second week and so on. However, if we run the advertisement in the second week also and get another 100 GRPs, then in the second week the total GRPs realized will become 60 (from the fresh Ad) + 24 (effect from the previous week) = 84.

The table below presents a hypothetical example and calculation of the decay calculation at weekly level.

Week	GRPs	Adstock by alpha									
		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	100	100	90.0	80.0	70.0	60.0	50.0	40.0	30.0	20.0	10.0
2	100	100	99.0	96.0	91.0	84.0	75.0	64.0	51.0	36.0	19.0
3	100	100	99.9	99.2	97.3	93.6	87.5	78.4	65.7	48.8	27.1
4		0	10.0	19.8	29.2	37.4	43.8	47.0	46.0	39.0	24.4
5		0	1.0	4.0	8.8	15.0	21.9	28.2	32.2	31.2	22.0
6		0	0.1	0.8	2.6	6.0	10.9	16.9	22.5	25.0	19.8
7		0	0.0	0.2	0.8	2.4	5.5	10.2	15.8	20.0	17.8
8	70	70	63.0	56.0	49.2	43.0	37.7	34.1	32.0	30.0	23.0
9	80	80	78.3	75.2	70.8	65.2	58.9	52.5	46.4	40.0	28.7
10	70	70	70.8	71.0	70.2	68.1	64.4	59.5	53.5	46.0	32.8
11		0	7.1	14.2	21.1	27.2	32.2	35.7	37.5	36.8	29.5
12		0	0.7	2.8	6.3	10.9	16.1	21.4	26.2	29.4	26.6
13		0	0.1	0.6	1.9	4.4	8.1	12.8	18.4	23.5	23.9
Total		520	520.0	519.9	519.2	517.1	511.9	500.7	477.2	425.8	304.6

Effect of Decay on Promo Sales
Introduction of 'Decay' helps to capture the delayed effect of the promotions on sales.

Adding Decay into Promo Data
Please refer to page # 27 for example

2.6: Eliminating Fluctuation - Seasonality Calculation

- **Fluctuations: Seasonal, Cyclical and Random Effects**

The sales of any product normally varies with time. Different macro / micro economic factors, product life cycle related factors, short-run and long-run factors and different marketing initiavives can be atributed to these fluctuations.

The fluctuations in sales can be devided into two broad categories, Trend and Seasonality.

"Trend" present somewhat stable movement that are generated due to macro-economic factors like population / market growth etc. Also obsolation of some products (like tape recorder) lead to slow but steady decline in the demand for the same in the market.

The latter may have a formally similar nature (e.g., a plateau followed by a period of exponential growth); however, it repeats itself in systematic intervals over time. Those two general classes of time series components may coexist in real-life data. For example, sales of a company can rapidly grow over years but they still follow consistent seasonal patterns (e.g., about 25% of yearly sales each year are made in December, while only 4% in August).

"Seasonality" can be decomposed into 2 sub-categories, seasonal fluctuation (like sales of AC is higher in summer and lower in winter) and cyclical fluctuation (repeats in a shorter duration - like sales of cookies is high in the first week of every month as compared to other weeks).

Random effects are guided by factors we can't control.

- **Why Smoothening Fluctuation**

We need to smoothen the fluctuation to understand what the expected base sales would be at that period had there been no exogenous factors other than seasonality affects the sales. Unless we benchmark this, we might fall in trap of giving credit of some spike in sales to a particular promo or undermine the performance of a promo when the sales is not so high. The actual fact might be that in absence of the promo, the sales would have been further down in that lean season!

- **Ways to Eliminate Fluctuation: Seasonality Index Creation**

There are different ways to calculate seasonality index. Most common is to calculate the moving average of sales and calculate the ratio between the averaghe and the actual sales. However, in SAS, the PROC TIMESERIES procedure can calculate the seasonality index automatically based on the time series data on sales.

- **Addding SI Into Analytic Dataset** (please refer to "Final Form of Transformed Sales and Promo Data")

2.7 Modeling Dataset Creation

- **What is MODELING Dataset?**

A modeling dataset, as the name suggests is the dataset which has

- the dependent and independent variables properly defined and populated
- data rolled up at appropriate level
- all variables properly capped / binned / treated for missing values
- seasonality index populated
- All variables (excluding the binary promo flags) are log-transformed (natural base)
- All variables (including promo flags) as mean centered at store / other geographic level (to be decided based on the level of the data)

- **Level of Modeling Dataset**

The level of a dataset has 2 dimensions – time and geography. In terms of time dimension, modeling dataset can be daily / weekly / monthly level. The level is mostly decided by promo implementaion policy / capability at the clients' end. The geographical level of the data can be at store / location / district / state/ county etc level. Again, this is also decided by the clients' promo implementation capabilities. If customized promo can be implemented at store level, then the geographical level of the data must be at store level. However, in a market mix exercise, its always a standard practise to have the data at store level that can be suitably rolled-up to appropriate level for implementation of the suggestions.

- **Contents of a Modeling Dataset**

1. Store and other geographical location identifier
2. Month / week number / date
3. Sales (dependent variable) for different brands (log transformed and mean centered)
4. Promo flags / GRPs / quantities for different variables (mean centered)
5. Price variables (log transformed and mean centered)
6. Seasonality index (log transformed and mean centered)

- **Auditing the Modeling Dataset**

Same steps to be followed as discussed in "Common Issues Faced During Data Conversion" in "Data Mining / Auditing" section in page # 16.

- **Sample Modeling Dataset**

The sample data presented in "Final Form of Transformed Sales and Promo Data" presents a glimps of the actual modeling dataset. For a more detailed structure of the modeling dataset, refer to Mu Sigma's shared drive where some real life data are saved.

Chapter 3: Modeling

- **Fundamentals of Mix Modeling**
- **Why Regression is Important Before Mix Modeling**
- **Regression Procedure: Sample Data, Syntax and Output in SAS**
- **Interpretation of Output: Removal of Multicolinearity**
- **Iterations in Regressions**
- **Final Model Selection Criteria**
- **Proc Mix Procedure: Sample Data, Syntax and Output in SAS**
- **Final Model Creation: Modeling Equation**
- **Validation (in-sample and out-sample)**
- **Deviation in Prediction: MAPE**

- **Fundamentals of Mix Modeling**

The form of a mixed model is similar to any other linear model, such as a regression model…

- … however, all individual models are estimated at once, rather than as separate independent models

- Each causal factor is considered either "fixed" or "random"

– Fixed effects are coefficients which are constant across subjects
– Random effects are coefficients which vary for each subject — generally expressed as a constant fixed effect across subjects, plus a random deviation for each subject

- This approach alleviates the difficulties found in estimating individual models

– For subjects where causal values do not change, the estimated coefficient is the constant fixed effect
– Multicollinearity is reduced because all data is used to create estimates, not just those for the specific subject

- **Why Regression is Important Before Mix Modeling**

The PROC MIX procedure can't detect the multicolinear variables. So we use a similar linear regression model to detect and remoce the multicolinearity. Once we have the non-colinear variables, we use them as explanatory variables in a PROC MIX procedure.

- **Regression Procedure: Sample Data, Syntax and Output in SAS**

The regression procedure establishes linear relationship between a dependant variable and a set of independent variables. PROC REG procedure in SAS performs the analysis. As a bi-product, the model output throws the VIF, T-value, P-value and R^2 (showing overall fitment of the model). The process requires multiple iteration. After the first run, we identify and drop the variable with highest VIF and rerun the model with leftover variables. The process continues till we end up with few 'significant' variables (P value<= 0.05) with negligible multicolinearity (VIF <=2). Also the analyst needs to apply his judgment in selecting the final list of variables based on their applicability in real life situation. For example, in a model the 'income' of the individual appears significant. However, as per US rules, no organization is allowed to discriminate the customers in terms of income / gender / age etc. It might be wise to drop that variable from the set of selected explanatory variable and see the impact and do some more iterations to get a good model without such variables.

A sample data for regression modeling is as follows

Store	yrwk	Other_Dir	Advo_Tab_4	GRP_Cbl	ln_sls	ln_Eagle	ln_Fortera	ln_Wrangler
900128	200400	0	1	0	-1.51928	0.42776	-0.00722	-0.03379
900128	200401	0	1	0	-0.3666	-0.15813	-0.00722	-0.03379
900128	200402	1	0	0	-0.50768	-0.53187	0.03852	-0.07375
900128	200403	1	0	0	-0.24299	-0.04363	-0.00722	-0.02581
900128	200404	0	0	0	-0.5702	0.17957	-0.00722	-0.03866
900128	200405	0	0	0	-0.39327	0.31543	-0.00722	-0.31491
900128	200406	0	0	0	-0.17554	0.21046	-0.00722	0.01544
900128	200407	0	0	0	-0.63689	-0.31887	-0.00722	-0.03379
900128	200408	0	0	0	-0.5702	0.00655	-0.00722	0.13617
900128	200409	0	0	0	-0.22	-0.17475	-0.00722	0.05194
900128	200410	0	0	0	-1.05975	0.04942	-0.00722	-0.03379
900165	200400	0	0	0	-0.70835	0.04394	-0.00722	-0.03379
900165	200401	0	0	0	-0.63689	0.05307	-0.00722	0.0821
900165	200402	0	1	0	-0.09216	0.05307	-0.04323	-0.03379
900165	200403	0	1	0	-0.74609	0.57833	-0.00722	-0.18375
900165	200404	1	0	0	-0.07236	0.11903	-0.00722	-0.03379
900165	200405	1	0	0	-0.22	-0.16747	-0.00722	0.00056
900165	200406	0	0	0	-0.19752	0.05986	-0.00722	-0.00476
900165	200407	0	0	0	-0.42067	-0.12562	-0.00722	0.15887
900165	200408	0	0	0	-0.50768	-0.15821	-0.00722	-0.00337
900165	200409	0	0	0	-0.15404	-0.28597	-0.00722	0.12627
900165	200410	0	0	0	-0.24299	0.03736	-0.00722	0.00767
900256	200400	0	0	0	-0.13299	1.13024	-0.00722	-0.09177
900256	200401	0	0	0	-0.3666	-0.0691	-0.00722	0.27666
900256	200402	0	1	0	-0.42067	-0.14066	-0.00722	0.17757
900256	200403	0	1	0	-0.39327	0.23058	-0.00722	0.08295
900256	200404	0	0	0	0.12295	0.11476	-0.00722	-0.02156
900256	200405	0	0	0	-0.63689	0.03316	-0.00722	-0.03379
900256	200406	0	0	0	-0.63689	-0.2385	-0.00722	-0.03379
900256	200407	1	0	0	-0.05294	-0.24077	-0.00722	0.11725
900256	200408	1	0	0	-0.19752	0.2124	-0.00722	-0.03379
900256	200409	0	0	0	0.28627	0.02442	0.15714	-0.06604
900256	200410	0	0	0	0.47315	0.0578	-0.00722	-0.04232

In this dataset, the dependent variable is ln_sales. Here is the proc reg code to be used in SAS

```
ods listing select FitStatistics ParameterEstimates anova;
proc reg data = temp;
           model ln_sls = ln_Eagle ln_Fortera ln_Wrangler Advo_Tab_4 GRP_Cbl
Other_Dir/ vif tol;
            ods output ParameterEstimates = VIF;
      run;quit;
quit;
```

The output in SAS needs to be interpreted carefully to remove multicolinearity.

○ Interpretation of Output: Removal of Multicolinearity

The SAS output looks like below.
The R^2 (encircled in red) shows the overall model fitment.

The VIF (encircled in blue) shows the effect of multicolinearity. In this sample, all the variables have acceptable level of VIF. However, during the initial iteration we might get high VIFs for some variables. We need to drop them one after another in each iteration and try building the model with the remaining variables.

The estimates (beta coefficients, encircled in dark yellow) are to be used for developing the linear equation in a normal regression model building case. However, at this stage we are not focusing on the same. As a part of the market mix model development, we will use PROC REG as a tool to remove multicolinearity and select the 'good' variables.

The SAS System 15:30 Tuesday, October 10, 2006 19

The REG Procedure
Model: MODEL1
Dependent Variable: ln_sls

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	37.20908	6.20151	55.55	<.0001
Error	60664	6772.13445	0.11163		
Corrected Total	60670	6809.34353			

Root MSE	0.33412	R-Square	0.654
Dependent Mean	-3.4068E-16	Adj R-Sq	0.555
Coeff Var	-9.8074E16		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	9.80541E-17	0.00136	0.00	1.0000	.	0
ln_Eagle	1	-0.05605	0.00597	-9.39	<.0001	0.99569	1.00433
ln_Fortera	1	-0.09671	0.01236	-7.82	<.0001	0.99488	1.00514
ln_Wrangler	1	-0.04400	0.00783	-5.62	<.0001	0.99403	1.00601
Advo_Tab_4	1	0.05144	0.00640	8.03	<.0001	0.99655	1.00346
GRP_Cbl	1	0.00056412	0.00027812	2.03	0.0425	0.99960	1.00040
Other_Dir	1	0.09777	0.01307	7.48	<.0001	0.99976	1.00024

- **Iterations in Regressions**

As have already been discussed earlier. We need to start with all the independent variables while building a regression model. However, based on the multicolinearity and P value we need to drop the insignificant variables one after another and rerun the model till we get the final list of variables with acceptable level of VIF / P value.

- **Final Model Selection Criteria**

The final model must fulfill the following criteria

1. All the variables have acceptable P value (normally less than 0.05)
2. The VIF are within acceptable range (<=2)
3. R^2 is acceptable (higher the R^2, better is the model)
4. The variables selected makes business sence
5. Total number of variables not more than 10. Too many variables selection willmake the model unstable in future.

- **Proc Mix Procedure: Sample Data, Syntax and Output in SAS**

Sample data for PROC MIX is same as what presented for PROC REG.

The SAS syntax for mixed modeling is as follows…

```
proc mixed data =temp  SCORING=5 DFBW IC METHOD=REML;
            class nonsig;
            model ln_sls =
ln_Eagle
ln_Fortera
ln_Wrangler
Advo_Tab_4
GRP_Cbl
Other_Dir
/ solution;
            ods output solutionf = FXD_ln_tot_sales;
run;quit;
```

The SAS output looks like the one presented in the next page.
We need to focus on the estimates and the P value of the variables. Final model selection criteria remains same as discussed in the regression model building procedure.

Information Criteria

Neg2LogLike	Parms	AIC	AICC	HQIC	BIC	CAIC
39210.8	1	39212.8	39212.8	39215.6	39221.8	39222.8

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	9.81E-17	0.001356	61E3	0.00	1.0000
ln_Eagle	-0.05605	0.005967	61E3	-9.39	<.0001
ln_Fortera	-0.09671	0.01236	61E3	-7.82	<.0001
ln_Wrangler	-0.04400	0.007826	61E3	-5.62	<.0001
Advo_Tab_4	0.05144	0.006404	61E3	8.03	<.0001
GRP_Cbl	0.000564	0.000278	61E3	2.03	0.0425
Other_Dir	0.09777	0.01307	61E3	7.48	<.0001

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
ln_Eagle	1	61E3	88.23	<.0001
ln_Fortera	1	61E3	61.19	<.0001
ln_Wrangler	1	61E3	31.60	<.0001
Advo_Tab_4	1	61E3	64.52	<.0001
GRP_Cbl	1	61E3	4.11	0.0425
Other_Dir	1	61E3	56.00	<.0001

Obs	Effect	Estimate	StdErr	DF	tValue	Probt
1	Intercept	9.81E-17	0.001356	61E3	0.00	1.0000
2	ln_Eagle	-0.05605	0.005967	61E3	-9.39	<.0001
3	ln_Fortera	-0.09671	0.01236	61E3	-7.82	<.0001
4	ln_Wrangler	-0.04400	0.007826	61E3	-5.62	<.0001
5	Advo_Tab_4	0.05144	0.006404	61E3	8.03	<.0001
6	GRP_Cbl	0.000564	0.000278	61E3	2.03	0.0425
7	Other_Dir	0.09777	0.01307	61E3	7.48	<.0001

- ### **Final Model Creation: Modeling Equation**

Writing the modeling equation out of a PROC MIX output is the most critical part of the whole exercise. Before we discuss that step, lets refresh the contents of the modeling dataset.

1. Store and other geographical location identifier
2. Month / week number / date
3. Sales (dependent variable) for different brands (log transformed and mean centered)
4. Promo flags / quantities for different variables (mean centered)
5. Price variables (log transformed and mean centered)
6. Seasonality index (log transformed and mean centered)

Clearly, in a mixed model we try to build the relationship between the log transformed and mean centered dependent variable and the set of independent variables (all are mean centered, some variables like price / seasonality etc. are log transformed too).

The mean centering is done to ensure zero intercept. However, when we convert the modeling output into a mathematical equation, we need to bring the effect of the mean back to the equation. This can be justified by the fact that mean centered price variables essentially present the deviation of the actual price from the average and the model captures the effect of those changes in total sales. However, the baseline sales will depend in the 'mean' price only. Any change in the price will result in sales fluctuating from that base. Clearly when we are trying to predict sales as a function of price variable, there are 2 in-built relations we are trying to explain. One, the relation between mean price and baseline sales. And second, the effect of price changes on sales. This explains why we need to bring the 'mean' value of each variables back into the final equation.

Before we do this, we need to go back to original analytical dataset (variables are not mean centered) and calculate the mean of each of the significant variables using PROC MEANS procedure in SAS. *PLEASE NOTE, THE LEVEL OF CALCULATION THE MEAN HAS TO BE SAME AS THE GEOGRAPHICAL LEVEL OF THE DATASET.*

Here is the syntax for proc means procedure

```
proc means data = RAW_DATA ;
class store;
var
ln_sls ln_Eagle ln_Fortera ln_Wrangler Advo_Tab_4 GRP_Cbl Other_Dir ;
output out = as1 mean = mln_sls mln_Eagle mln_Fortera mln_Wrangler
mAdvo_Tab_4 mGRP_Cbl mOther_Dir;
run;
```

Now append these means to the original dataset

```
proc sort data = raw_data out= meancnt;
by nonsig;run;

proc sort data = as1;by nonsig;run;
```

```
data origmeancnt;
 merge as1(in=a) meancnt(in=b);
 by nonsig;
 if a and b;
run;
```

Now select only the required variables and write the modeling equation as follows.

```
data origmeancnt1 (keep =
ln_sls
ln_sls
ln_Eagle
ln_Fortera
ln_Wrangler
Advo_Tab_4
GRP_Cbl
Other_Dir
mln_sls
mln_Eagle
mln_Fortera
mln_Wrangler
mAdvo_Tab_4
mGRP_Cbl
mOther_Dir
pred
tot_sales
res
yrwk
nonsig
);
set origmeancnt;

pred =
1.011725065*exp(
-0.05605*(ln_Eagle-mln_Eagle)
-0.09671*(ln_Fortera-mln_Fortera)
-0.04400*(ln_Wrangler-mln_Wrangler)+
 0.05144*(Advo_Tab_4-mAdvo_Tab_4)+
0.000564*(GRP_Cbl-mGRP_Cbl)+
 0.09777*(Other_Dir-mOther_Dir)+
+mln_sls);
res = pred - tot_sales;
run;
quit;
```

- **Validation (in-sample)**

Market mix model are typically validated ‘in-sample’. The actual sales and the predicted sales are compared within the analytical dataset using some basic SAS procedure as follows.

```
proc summary data = origmeancnt1 nmiss nway; class yrwk ;
var pred tot_sales; output out = test10 sum = ; run; quit;
```

Here is how the output looks like

Year-Week	Predicted	Actual
200401	41,089	39,916
200402	41,310	41,249
200403	41,631	39,503
200404	40,358	42,194
200405	40,362	41,028
200406	41,083	43,135
200407	42,736	42,692
200408	43,656	43,550
200409	41,415	41,752
200410	39,365	40,471
200411	41,184	43,850
200412	44,597	44,526
200413	46,121	47,384
200414	41,349	43,083
200415	41,268	41,763
200416	40,727	40,703
200417	42,451	44,327

- **Deviation in Prediction: MAPE**

MAPE is the measure of goodness of fit of a Market Mix model. Its calculated at the daily / weekly / monthly level (based on the level of the modeling dataset). MAPE (Mean Abolute Percentage Error) is defined as the ratio between the 'absolute deviation between predicted and actual sales' and actual sales. A good model must have less than 10% overall MAPE. Here is how the MAPE is calculated.

Year-Week	Predicted	Actual	Residual =abs(pred – actual)	MAPE = residual/ actual
200401	41,089	39,916	1,173	2.9%
200402	41,310	41,249	61	0.1%
200403	41,631	39,503	2,128	5.4%
200404	40,358	42,194	1,836	4.4%
200405	40,362	41,028	666	1.6%
200406	41,083	43,135	2,052	4.8%
200407	42,736	42,692	44	0.1%
200408	43,656	43,550	106	0.2%
200409	41,415	41,752	337	0.8%
200410	39,365	40,471	1,106	2.7%
200411	41,184	43,850	2,666	6.1%
200412	44,597	44,526	71	0.2%
200413	46,121	47,384	1,263	2.7%
200414	41,349	43,083	1,734	4.0%
200415	41,268	41,763	495	1.2%
200416	40,727	40,703	24	0.1%
200417	42,451	44,327	1,876	4.2%
200418	40,324	42,683	2,359	5.5%
Total	751,025	763,809	19,997	2.6%

Chapter 4: Deliverables

‘Predicted Sales’ vs. ‘Volume-due-to’

- Calculating Predicted Sales
- Calculating Contribution (volume-due-to) from Significant variables

Return on Investment Calculation

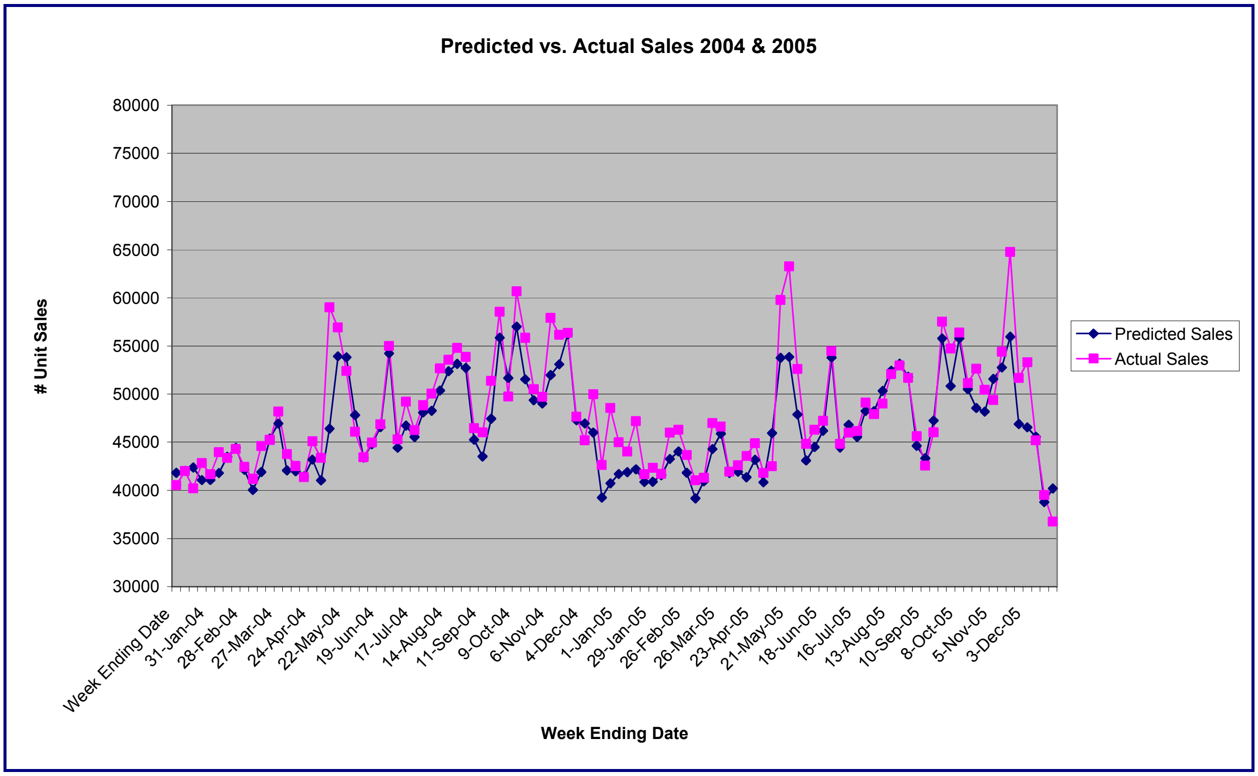
- What is ROI?
- ROI for Different Market Mix Components
- Calculation procedure

Optimum Mix identification

Interpretation and Presentation Basics

○ Calculating Predicted Sales

The calculation of the predicted sales has already been discussed in details. However, in terms of 'deliverable', the best way to present the comparison diagrammatically as below.



○ Calculating Contribution

One of the important deliverables from a Market Mix model is calculating the contribution of the significant factors to total sales.

As we have already discussed, the total unit sales has two components, baseline + fluctuations. Baseline is dependent on the mean level of the significant variables, while any changes from the mean level results in 'fluctuations' in sales.

While calculating contribution we don't touch the 'mean' part of the independent variables. We remove the mean-centered of the variable from the equation for predicting sales. This removal impacts the predicted sales. The difference between the actual predicted sales and the predicted sales after removing a mean centered variable from the equation is the contribution from that particular variable.

Below is a sample SAS syntax for the same (note, for calculating the contribution from each variable, we omit the log transformed form of that variable from the equation).

SAS Syntax:

```
data contribution (keep =
ln_sls ln_sls ln_Eagle ln_Fortera ln_Wrangler Advo_Tab_4 GRP_Cbl Other_Dir
mln_sls mln_Eagle mln_Fortera mln_Wrangler mAdvo_Tab_4 mGRP_Cbl mOther_Dir
pred tot_sales res yrwk  nonsig pred_ln_Eagle pred_ln_Fortera
pred_ln_Wrangler pred_Advo_Tab_4 pred_GRP_Cbl pred_Other_Dir Contr_ln_Eagle
Contr_ln_Fortera Contr_ln_Wrangler Contr_Advo_Tab_4 Contr_GRP_Cbl
Contr_Other_Dir);
set origmeancnt;

pred =
1.011725065*exp(
-0.05605*(ln_Eagle-mln_Eagle) -0.09671*(ln_Fortera-mln_Fortera) -
0.04400*(ln_Wrangler-mln_Wrangler)+ 0.05144*(Advo_Tab_4-mAdvo_Tab_4)+
0.000564*(GRP_Cbl-mGRP_Cbl)+ 0.09777*(Other_Dir-mOther_Dir)+ +mln_sls);
res = pred - tot_cars_week;

pred_ln_Eagle=
1.011725065*exp(
-0.05605*( -mln_Eagle)-0.09671*(ln_Fortera-mln_Fortera)-0.04400*(ln_Wrangler-
mln_Wrangler)+ 0.05144*(Advo_Tab_4-mAdvo_Tab_4)+0.000564*(GRP_Cbl-mGRP_Cbl)+
 0.09777*(Other_Dir-mOther_Dir)++mln_sls);
Contr_ln_Eagle=pred - pred_ln_Eagle;

pred_ln_Fortera =
1.011725065*exp(
-0.05605*(ln_Eagle -mln_Eagle)-0.09671*( -mln_Fortera)-0.04400*(ln_Wrangler-
mln_Wrangler)+ 0.05144*(Advo_Tab_4-mAdvo_Tab_4)+0.000564*(GRP_Cbl-mGRP_Cbl)+
 0.09777*(Other_Dir-mOther_Dir)++mln_sls);
Contr_ln_Fortera =pred - pred_ln_Fortera;

run;
```

After this, we need to summarize the data to get the contribution from each variable like below

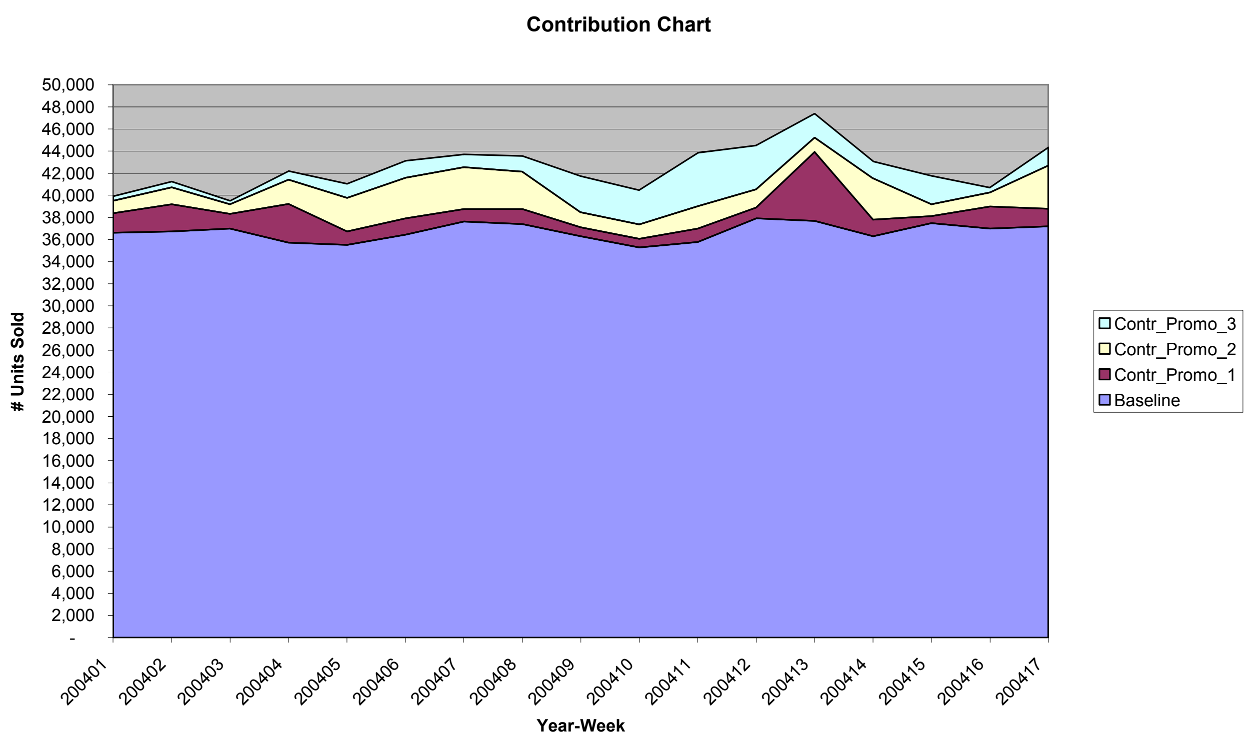
```
proc summary data = contribution;
class yrwk;
var
res
pred
tot_sales
Contr_ln_Eagle
Contr_ln_Fortera

;
output out = test9
sum=
;
run;quit;
```

The output Looks like:

Year-Week	Predicted	Actual	Baseline	Contr_Promo_1	Contr_Promo_2	Contr_Promo_3
200401	41,089	39,916	36,601	1,790	1,127	398
200402	41,310	41,249	36,738	2,436	1,534	541
200403	41,631	39,503	36,977	1,364	859	303
200404	40,358	42,194	35,731	3,490	2,197	776
200405	40,362	41,028	35,516	1,213	3,031	1,268
200406	41,083	43,135	36,447	1,471	3,678	1,538
200407	42,736	42,692	37,634	1,113	3,782	1,163
200408	43,656	43,550	37,402	1,353	3,382	1,414
200409	41,415	41,752	36,283	820	1,367	3,282
200410	39,365	40,471	35,289	777	1,296	3,109
200411	41,184	43,850	35,780	1,210	2,017	4,842
200412	44,597	44,526	37,930	989	1,649	3,958
200413	46,121	47,384	37,703	6,228	1,292	2,162
200414	41,349	43,083	36,308	1,491	3,726	1,558
200415	41,268	41,763	37,494	640	1,067	2,562
200416	40,727	40,703	36,991	2,004	1,262	445
200417	42,451	44,327	37,213	1,565	3,912	1,636

The same output is presented graphically (area curve) in the following way



Return on Investment Calculation

○ What is ROI?

Return on Investment (ROI) is the ratio between the incremental profit earned from a promotion and the cost of the promo.

Incremental profit is again calculated as the 'volume-due-to' for a particular promo for a brand multiplied by the profit margin for that brand. If the promo is applicable for more than one brand then total incremental profit from all the brands divided by the cost of the promo gives ROI for that particular Promo in the referred time window.

○ ROI for Different Market Mix Components

For different market mix components (promo) the same procedure described avove needs to be repeated for each promo separately.

○ Calculation procedure

ROI for a promo (X) is calculated as

$$\{\sum_{i=1}^{n}(VDT\_i * profitability\_i) / \text{Promo\_Cost\_X}\}$$

Where VDT_i = volume-due-to for brand 'i' and profitability_i = profitability for brand 'i'.

Optimum Mix identification

All the promos are rank ordered in terms of their ROI and volume-due-to (contribution).
It can so happen that one partivular promo generated more contribution than others but the total profit from the same is lower due to low profit margin for the brand. Alternatively, some promo might be active in a smaller market and the cost for the same is also quite low though the same promo generated high ROI.

Business needs to take a decision (as per their marketing strategy) based on the contribution and ROI as to how to allocate the fund among different promotions.

As an extension to the market-mix modeling, we can present some 'what-if'scenario using some basic simulation technique showing how the sales will be impacted if spend and coverage / durationof some promos changed. The final selection of optimum mix are done by the Client.

Interpretation and Presentation Basics

Market mix models and the outputs are generally presented in powerpoint files. Also at times the actual SAS codes / dataset and logs are demanded by the Client. Though there is no stringent way to present the model and the findings / recommendations to the client, below is a brief checklist that can be referred while presenting the model.

- Background, Scope and Objective
- Project Approach
- Summary of the Model
- Prediction vs. Actual (graph)
- Bivariate analysis to explain the mismatch between Predicted vs. Actual in some periods, if any
- Contribution (area graph)
- Bivariate analysis to support the contribution (time series analysis)
- Cross section (like region / state / district) level analysis to explain the uniformity / deviations in contribution from different promos across diff locations at same point of time, if any.
- ROI calculation
- Rank order of the promos in term od ROI and Contribution
- What-if analysis (if client wants)