LEARNING FROM BIG DATA: AN IMPACT ANALYSIS OF A SALES PROMOTION USING SAS

Palma Daawin

Department of Mathematics and Statistics, Miami University
December 9, 2015

Final Term Project for Statistical Programming in SAS

INTRODUCTION

Summary and visual statistics is by far one of the most important tools used in of statistics. The ability to produce, timely accurate and summarized information from large datasets is extremely important and necessary. Exception reporting has become very important concept as there are now large datasets available. Exception reporting enables important and useful information to be picked out from a rather large dataset. Managers need exception information order to make timely decisions. Sales promotions consist of a huge variety of temporary planned promotion tools which aim at generating a preferred response from the consumer (Gilbert and Jackaria, 2002).

A large chain store operator implemented a sales promotion campaign over a two year period on a number of groceries, data has been provided on the selected items, we want to evaluate the impact of the promotion campaign, the most purchased product and how much earned on some of these products, we would study and observe useful trends from the data set. We would answer the following research questions: What was the financial impact of the sales by category, stores or brands?

What is the impact on sales on products displayed or being featured in the circular? What is the impact on units/visit of promotions? How did the above differ by products? By categories? What was the impact of sales promotion by geography? Which brand grossed the most sales? Which products were the highest in demand?

This study seeks to improve on the state of the art of data and provide useful information from the data set from chain store retailer in order to make judgments about the impact of a sales promotion.

In this study, I specially intend to, Read, extract, organize, manipulate, combine and data prepare data from complicated source and provide summary information using graphs and other summary statistics in SAS.

Review and analyze data from large consumer retail chain store and make conclusions to aid in the business's procurement decision based on the dataset to report on the top brands and top grossing products of the promotional period.

DATA DESCRIPTION

The data was obtained from dunnhumby.com, a market analytics company. The data was freely available to the public by the company to help programmers and data scientist in current and future studies The data contained in this file includes selected sales and promotional information from stores over 156 weeks, beginning January 2009 through December 2011. These are the sales metrics of products in 4 categories: mouthwash, pretzels, frozen pizza, and boxed cereal, the top 5 products from each of the top 3 brands in each category. Historical data are provided on 524,922 lines of data on purchases made by customers in 12 variables namely Week End Date, Store Number, UPC, Units, Visits, HHS, Spend, Price, Base Price, Feature, Display, and Temporary Price Reduction.

In a separate excel sheet there is a list of the stores from which the data was extracted with 9 variable names specifying details of each of the 80 stores Store Id, Store Name, Address City Name, Address, State ,Province Code, MSA Code, Seg_Value_Name, Parking_Space_Qty Sales_Area_Size_Num, Avg_Weekly_Baskets.
The third excel sheet contains information about classification of the product purchased UPC, Description, Manufacturer, Category, Sub-Category and Product Size.

METHODOLOGY

This study involved the use of various methodologies in SAS® to enable us answer the research questions. I have provided a summary of some key methods used in analyzing the large sales data of the large chain store operator.

Data Reading, Shaping and Preparation data from complicated source

Data step programming tools in SAS® were used to read the complicated data, combine the data sets and produce summary information for decision making use of conditional statements such as
Three Excel Datasets were read into SAS making use of INFILE and INPUT commands with appropriate variables read as character or numeric where necessary , this included a large excel sheet with over 500,000 transactions , a dataset containing information on the Stores and a third dataset containing information on the various Products of interest.

Merging Datasets

Using the PROC SQL statements (CREATE TABLE, SELECT, FROM, WHERE) , each individual transaction were merged and assigned a Store Characteristics using store number of each transaction .Product characteristics were assigned to each individual transaction, making use of the universal product code. All 500,000 observations were assigned

1) A Store Name in which the transaction took place.
2) The State in which the store is located
3) The name of Manufacturer using the Universal Product Code
4) The Category of the product (i.e. Cold Cereal, Frozen pizza, Bag Snacks, Oral hygiene)
5) And other important variables listed in the data description were assigned to each transactions

SAS Dates

The dates of each individual transaction were converted into SAS Recognized Dates using date formats in SAS. To enable managers make decision on what products to buy there was the need to analyses purchases. To answer the question: What was the weekly or monthly demand of some products by week or month? In this regard, variables for dates were introduced (Transact Week, Transact Month and Transact year). This enabled analysis of information by time (months or years). Converting dates into SAS Recognized dates ensured that transactions were put into a form that will enable a (Weekly or Monthly Sales analysis by product).

Tables and Statistical Summary

One of the research questions was: What was the financial impact of the sales by category, stores or brands? To answer this question there was the need to sort and appropriately combine the data sets. In this regard, I used various techniques in PROC TABULATE with SUM, ALL and REPPCTSUM option to group and

classify the consumer purchases data by category, stores or brands geography or brands. Based on these summaries display summary of information with graphical procedures in SAS to answer the question. Since absolute numbers are sometimes insufficient statistics when making comparisons, I have explored and used Percentages (%) in some cases to show summary information. Also to make tables easy to read and exception circumstances to be identified, I used FORMAT Background options to highlight areas where sales were particularly high or within a range of interest to managers. Other research questions were:
Which brand grossed the most sales? Which products were the highest in demand?
In order to answer these question on which brand grossed the most sales, I used the PROC TABULATE to group the Spend (sales revenue) variable by the Manufacturer, this ensured that summary statistics were produced with appropriate categories to be able to choose which brand grossed the most sales in the 2 year sales promotion period .Due to the successful conversion of transaction dates into SAS dates more information could be given, for example most purchased product by week etc.

Graphical Analysis

Graphs are a great way of telling a story. I used graphs to provide a visual display of the figures, see Appendix. These graphs were modified with additional options in SGPLOT, and formatted in order to provide the best way of representing the facts of the sales promotion exercise.

Trend and Time Series Graph Analysis

Because trend analysis is important in making future predictions, a number of important time series graphs of sales data. Time series plots by week, month and year of the customer total spend on the various Manufacturer Brands were compiled. A major strategy worth noting, which was employed in producing these time series graphs was to use the output from PROC TABULATE. This was necessary because since the data set was large, it was almost impossible to plot each transaction on a graph. There was a need for summary graphs to be produced and this was achieved through the interaction of the ODS output from the PROC Tabulate function together with the SERIES option in SGPLOT. In this regard, the Sum of transactions was categorized in a number of ways to ensure that the time series graphs could be produced.

RESULTS

Figure 1 shows a weekly time series graph of total sales of the observed brands during the promotion period 2009 to 2011. It is worth noting that sales trends changed dramatically in 2010 with shallow sales during the 2^{nd} and 12^{th} week. Aside these weeks where dramatic change was experienced, the overall trend appears to be similar for the 3 years running.

The time series graphs may enable us answer more specific questions about the purchasing decisions of the managers. From Figure 2, Figure 3 and Figure 4 we are able to observe the monthly trend of most purchased brand of products, this will enable a manager predict sales and hence make optimal decisions about which brand of products may be in high demand. Note here that it's not useful to make decisions on the brand Private, as these were several brands of products classified in the original data set which could not be separated. General MI appears to be the top brand in most of the months, followed by Kellogg and Tombstone.

Figure 5 shows a graphical display of combined monthly total sales from this selected brand of products. The tables provide some statistical evidence about observations from the graphical displays. It's worthy to note

that for all the years under review, sales of these brands were evenly distributed throughout the months of the year with an average of about 8%.

Some research questions were: What was the financial impact of the sales by category, stores or brands? Which brand grossed the most sales? Which products were the highest in demand?

From Table 1, total combined sales for these brands of products were $6,294,968.48 in 2009, sales increased in 2010 to $6,910,687.36 but decreased to $6,613,624.40 in 2011.

From Figure 6, we observe that for the 3 years running 2009, 2010, 2011, the single most purchased brand of the store's sales products was General MI, grossing 17% of combined sales for the 3 years, followed by Kellogg grossing 13.65% of total sales for the period under review, with Tombstone (11.92%) and Post Foods (8.42%). See Table 1 for actual figures and percentages.
From Table 2, we can see that for the 3 years running 2009, 2010, 2011, the most purchased category of the store's products was cold cereals grossing combined sales of $10,525,263.39 for the 3 years, followed by frozen pizza, bag snacks and oral hygiene products.

One of the research questions was; what was the impact of sales promotion by geography?
Table 2, shows the total combined sales by State and further broken down into the individual stores for each of the 3 years in Table 3 , results here is not surprising as majority of the stores of the large retail operator are located in the two states Texas and Ohio . The marked difference in the total sales of Indiana and Kentucky was because retailer's operates only a small number of few stores in these states. There were 51 different stores located in various cities. Of these 31 were located in Texas, 16 in Ohio, 3 in Kentucky and only 1 in Indiana. Combined sales from Ohio stores for the 3 years was $9,939,102.92 while Texas made $8,333,459.03. We also note that for the 3 years running, the Cincinnati store grossed the most sales , averaging about 15% of total sales in each of the years. The store located in Houston also contributed the next largest percentage of sales, averagely about 7% per annum. Appropriately combining the datasets has enabled production of information on each location using specified criteria.

CONCLUSION

This study illustrates the benefits of judicious selection of procedures in SAS to produce required information and answer important questions. Simply put we could use SAS methods and procedures to evaluate large data sets and make appropriate conclusions.

Managers of the large chain store can use SAS to generate required information that is necessary for decision making of optimal purchasing quantities and frequency. However, to obtain an optimal purchasing strategy, one would have do a number of statistical procedures to make a decision. For example, there was the need to answer the question of whether the reduction in prices affected sales, a Chi-square test of homogeneity could be performed. A Two Sample T- Test of mean total sales units whenever sales prices were reduced or products were displayed in a magazine could be performed to determine if there was really a change in the means sales whenever prices were reduced. This information might be obtained from the literature or from pilot studies. In this study, I have mainly focused on data step, and combining large data sets to produce information that is beneficial to users who want to make quick but accurate decisions.

Since making purchasing decision is inevitable in every business, it is important for managers to develop tools that enable them to make prudent fund allocations required to optimize purchasing decisions. Traditional ways of marketing may yield sub optimal results and in many cases lead to money wasting on certain irrelevant marketing efforts. Great models can be developed and used to understand the effects of promotional

activities and identify the key brands and products that drive the most sales among a group of competing brands and products.

In this study, I have provided some amazing insights to analytics professionals who may be thinking about the best way to analyze large datasets using optimal procedures in SAS .Among several competing brands and products in a large retail store operator we can evaluate weekly trends of customer purchases that will enable key decision makers predict possible quantity of units of particular brand that may be required and bought by their loyal customers. By studying trend of sales of products we would be able to discover very key relationships and hence make optimal purchasing choices. It must be however noted that the data was extract data and not all the data of the retail store operator.

Different results may be achieved if same analysis is performed on the full data set. Conclusions made are based on the data set available for this study.

REFERENCES

[1] Bailer A.J. (2010) *Statistical Programming in SAS*. Cary, NC: SAS Institute Inc.

[2]Clark A, Donald K., Weinmann C. (2014) *Measuring Product-level Promotional Effectiveness using Multiple Linear Regression*

[3] Gilbert, D.C. and Jackaria, N. (2002) "*The efficacy of sales promotions in UK supermarkets: A consumer view" International Journal of Retail & Distribution Management*, Vol. 30, No. 6, pp. 315-322

[4] Huntley S, Middleton W (2012) A Different Point of View with ODS PDF in SAS® 9.3, SAS, Cary, NC, USA

[5] Leonard P M (n.d), SAS Institute Incromotional Analysis and Forecasting for Demand Planning: A Practical Time Series Approach Cary, NC, USA

[6] Persson P.G. (1995) Modeling the Impact of Sales Promotion on Store Profits. *The Economic Research Institute at the Stockholm School of Economics,*

[7] Data file *(2015)* Retrieved from http://www.dunnhumby.com/sourcefiles

APPENDIX

LEARNING FROM BIG DATA: AN IMPACT ANALYSIS OF A SALES PROMOTION USING SAS

Appendix of SAS Codes, Figures and Tables

```
/*******************************************************************************
       Nov-Dec 2015

       PROJECT TITLE : LEARNING FROM BIG DATA: AN IMPACT ANALYSIS OF A SALES PROMOTION USING SAS

       Author : Palma Daawin

       Purpose: This program is written as part of a study that seeks to improve on the state of
       the art of data and provide useful information from the data set from chain store retailer in
       order to make judgments about the impact of a sales promotion.
       In this program  I specially intend to, Read, extract, organize, manipulate, combine
       and data prepare data from complicated source and provide summary information using graphs and
       other summary statistics in SAS.

*******************************************************************************************************/
data saleshistory;
       infile "C:\Users\Palma\Desktop\SAS Term Project\CustomerPurchasesProject\salesdata.csv"
       firstobs=3
       dsd; /* delimiter-separated data - comma is default delimiter */
       input transact_date date9. filenum store_num upc units visits hhs spend price base_price feature display
         tpr_only;
       transact_month = month(transact_date); *Create a variable that shows  the Month of each observation;
       transact_week= week(transact_date);
       transact_year = year(transact_date);  *Create a variable that shows  the Year  of each obeservation;

run;

/*** Reading dataset containing product descritption**/
 data productdescript;
       infile "C:\Users\Palma\Desktop\SAS Term Project\CustomerPurchasesProject\productdescription.csv"
       firstobs=3 /* data start on line 2, not line 1 */
       dsd;
       input  upc2 description $ manufacturer $ category $ sub_category $ product_size $;
              manufacturer = propcase(manufacturer);
              category = propcase(category);
run;

/*** Reading dataset containing stores descritption**/
data storedescript;
       infile "C:\Users\Palma\Desktop\SAS Term Project\CustomerPurchasesProject\storesdescription.csv"
       firstobs=3 /* data start on line 2, not line 1 */
       dsd;
       input store_num1 store_name $ city $ state $ msa_code $ seg_value_name $;
run;

/* Merging and Assigning the saleshistory and productdescript data sets by a variable(Universal Product Code) */
ods graphics off;
ods html close;
ods noresults;
proc sql;
              create table salesproclass as
              select *
              from saleshistory,productdescript
              where saleshistory.upc = productdescript.upc2;
              select
transact_date,transact_week,transact_month,transact_year,upc,store_num,units,visits,hhs,spend,price
                           ,base_price,feature,display,tpr_only,description,manufacturer,category
              from  saleshistory,productdescript;
quit;

ods output close;
ods results; * turning back results on ;
ods html path="%sysfunc(getoption(work))";
ods graphics on;

/**  Merging and Assigning the salesproclass and storedescript data sets by a variable (Store Number ) */
ods graphics off;
ods html close;
ods noresults;
proc sql;
              create table salesclassified as
```

```
                select *
                from salesproclass,storedescript
                where salesproclass.store_num = storedescript.store_num1;
                select transact_date,transact_week,transact_month,transact_year,upc,units,visits,hhs,spend,price,
                              base_price,feature,display,tpr_only,description,manufacturer,category
                        ,store_num,store_num1,store_name,city,state
                from  salesproclass,storedescript;
ods output close;
ods results; * turning back results on ;
ods html path="%sysfunc(getoption(work))";
ods graphics on;



/*** Producing tables that summaries multiple sales transactions*/
ods rtf file="C:\Users\Palma\Desktop\SAS Term Project\CustomerPurchasesProject\SalesByWeek";
ods output table=SalesbyWeek; *storing output into another table;
title "Table of summary sales data grouped by Week" ;
proc tabulate data= salesclassified ;
       class transact_week ; * classisy by TimeCategory;
       by transact_year;
       var  spend units; * Use these variables to create table of Statistical summaries;
       table transact_week,(spend )*(reppctsum='Percentage %' sum='Sales Revenue'*f=dollar11.2);
       label transact_year="Sales Year" transact_week="Transaction Week";

run;
ods output close;
ods rtf close;


/**** Producing a time series graph by week and grouped by year ****/
ods html style=htmlbluecml;
title 'Time Series Graph of Total Sales by Week for 3 Years';
proc sgplot data=SalesbyWeek;
       series x=transact_week y=spend_Sum / markers lineattrs=(pattern=solid) group=transact_year;
       *datalabel=transact_week;
       yaxis grid;
       xaxis grid;
       yaxis label="Total sales on products($)";
       xaxis label="Week of Sale";
run;


/*** Producing tables that summaries multiple sales transactions by Manufacturer ,by month and by year****/
ods output table=SBManufMonth;
ods rtf file="C:\Users\Palma\Desktop\SAS Term Project\CustomerPurchasesProject\SalesManuMon.rtf";
title "Table of Monthly summary Sales data grouped by Manufacturer" ;
proc tabulate data= salesclassified;
       class Manufacturer transact_month; * classisy by TimeCategory;
       by transact_year;
       var  spend units; *Use these variables to create table of Statistical summaries;
       table  transact_month*Manufacturer,(spend units)*(reppctsum sum);
     label transact_year="Sales Year" transact_week="Transaction Week" transact_month="Transaction Month";
run;



/****  Producing a time series graph Manufacturer by month by year***/
title 'Time Series Graph of Total Sales per Manufacturer by Month';
ods html style=htmlbluecml;
proc sgplot data=SBManufMonth ;
       series x=transact_month y=spend_Sum / lineattrs=(pattern=solid) markers group=manufacturer;
       by transact_year;
       xaxis grid;
       yaxis grid;
       format spend_Sum dollar8.0;
       yaxis label="Total sales on products($)";
       xaxis label="Month of Transaction";
       label transact_year="Sales Year";
run;




/*** Producing tables that summaries multiple total sales transactions by month and by year****/
```

```
ods output table=SalesbyMonth;
title "Table of summary sales data grouped by Month" ;
proc tabulate data= salesclassified ;
        class transact_month ; * classisy by TimeCategory;
        by transact_year; * create different tables of Statistical summaries by year;
        var  spend ; * Use these variables to create table of Statistical summaries;
        table  transact_month,(spend )*(reppctsum='Percentage %' sum='Sales Revenue'*f=dollar11.2);
        label transact_year="Sales Year" transact_month="Transaction Month";
run;
ods output close;

/****  Producing a time series graph of total sales by month by year***/
ods html style=htmlbluecml;
title 'Time Series Graph of Consumer Spending';
proc sgplot data=SalesbyMonth ;
        series x=transact_month y=spend_Sum / lineattrs=(pattern=solid) markers group=transact_year
        datalabel=transact_month;
        *format transact_month monname.;
        xaxis grid;
        yaxis grid;
        format spend_Sum dollar8.0;
        yaxis label="Total sales on products($)";
        xaxis label="Month of Transaction";
run;
ods html close;


/****Formatting colors in table by this criteria*/
proc format;
   value expfmt low-<100000='white'
                     100000-<150000='blue'
                     150000-high='yellow';
run;

/*** Producing tables that summaries total sales transactions by city and state****/
ods rtf file="C:\Users\Palma\Desktop\SAS Term Project\CustomerPurchasesProject\SalesCityState.rtf";
ods html style=htmlbluecml;
title "Table of total sales over 3 years classified by City and State" ;
proc tabulate data= salesclassified  format=8.2 style=[backgroundcolor=expfmt.];
        class state city transact_year ; * classisy by City;
        var  spend; * Use these variables to create table of Statistical summaries;
        table (state*city all='Total'), (spend)*(reppctsum='Percentage %' sum='Sales Revenue'*f=dollar14.2)
                *(transact_year all='Total');
        label transact_year="Sales Year" sum="Sales Revenue" city="Store Location" spend="Sales" state="State";
run;
ods html close;
ods rtf close;



/**Produces data required for sgplot  step below*/
ods output table=SBManuf;
title "Table of summary Sales data grouped by Manufacturer" ;
proc tabulate data= salesclassified;
        class Manufacturer; * classisy by TimeCategory;
        by transact_year;
        var  spend units; *Use these variables to create table of Statistical summaries;
        table  Manufacturer,(spend units)*(reppctsum='Percentage %' sum);
        /*format sum'Sales Revenue'*f=dollar10.2*/
run;
ods html close;
ods rtf close;



ods html close;
/****Formatting colors in table by this criteria*/
proc format;
   value expfmt low-<500000='white'
                     500000-<1000000='blue'
                     1000000-high='yellow';
run;
```

```
ods rtf file="C:\Users\Palma\Desktop\SAS Term Project\CustomerPurchasesProject\SaleManYrFin.rtf";
ods output table=SBManufTotal;
title "Table of summary Sales over 3 years grouped by Manufacturer" ;
proc tabulate data= salesclassified format=8.2 style=[backgroundcolor=expfmt.];
        class Manufacturer transact_year ; * classisy by TimeCategory;
        var  spend ; *Use these variables to create table of Statistical summaries;
        table  (Manufacturer all='Total'),(spend )*(reppctsum='Percentage %' sum='Sales Revenue'*f=dollar14.2)
               *(transact_year all='Total');
        label transact_year="Sales Year" transact_week="Transaction Week" transact_month="Transaction Month"
                 spend="Sales";
run;


ods html style=htmlbluecml;
title "Graph of  summary Sales  classified by Manufacturer for the 3 years" ;
proc sgplot data=SBManufTotal ;
        hbar manufacturer /response=spend_Sum dataskin=gloss group=transact_year;
        format spend_Sum dollar8.0;
        yaxis grid;
        xaxis grid;
        xaxis label="Total sales on products($)";
        yaxis label="Manufacturer";
run;


/*** Producing tables that summaries total sales  by category , year  and state****/
ods rtf file="C:\Users\Palma\Desktop\SAS Term Project\CustomerPurchasesProject\SalesCatYrState.rtf";
ods html style=htmlbluecml;
title "Table of total sales over 3 years classified by Category ,State and Year" ;
proc tabulate data= salesclassified ;
        class state transact_year category ; * classisy by City;
        var  spend; * Use these variables to create table of Statistical summaries;
        table (category all='Total' state all='Total'), (spend)*(reppctsum='Percentage %' sum='Sales
               Revenue'*f=dollar14.2) *(transact_year all='Total');
        label transact_year="Sales Year" sum="Sales Revenue" city="Store Location" spend="Sales" state="State";
run;
ods html close;
ods rtf close;
```

FIGURE 1

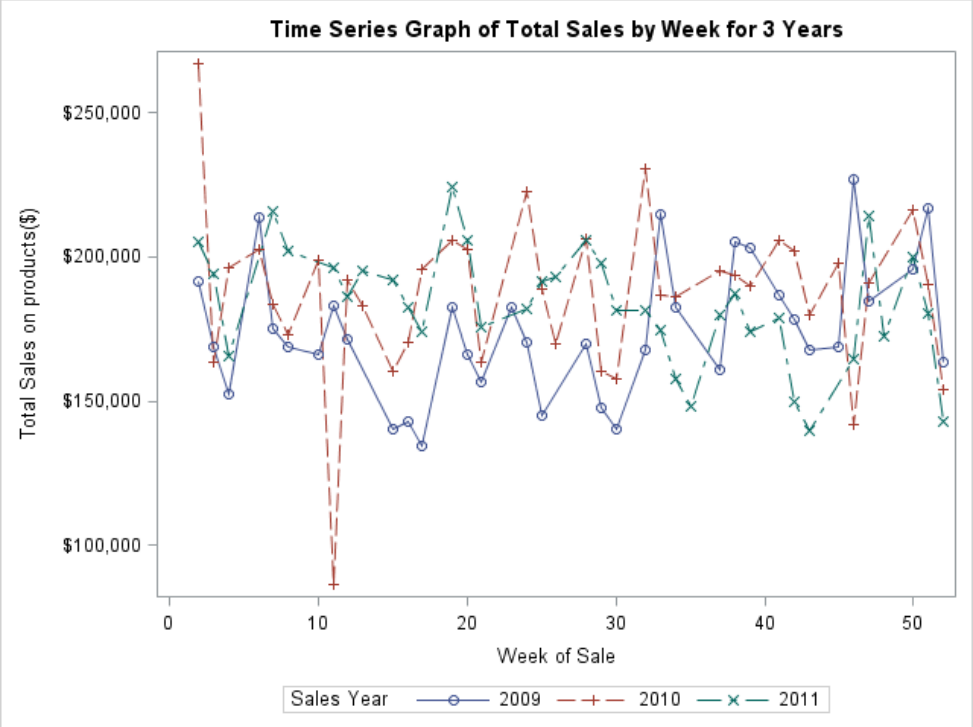


FIGURE 2

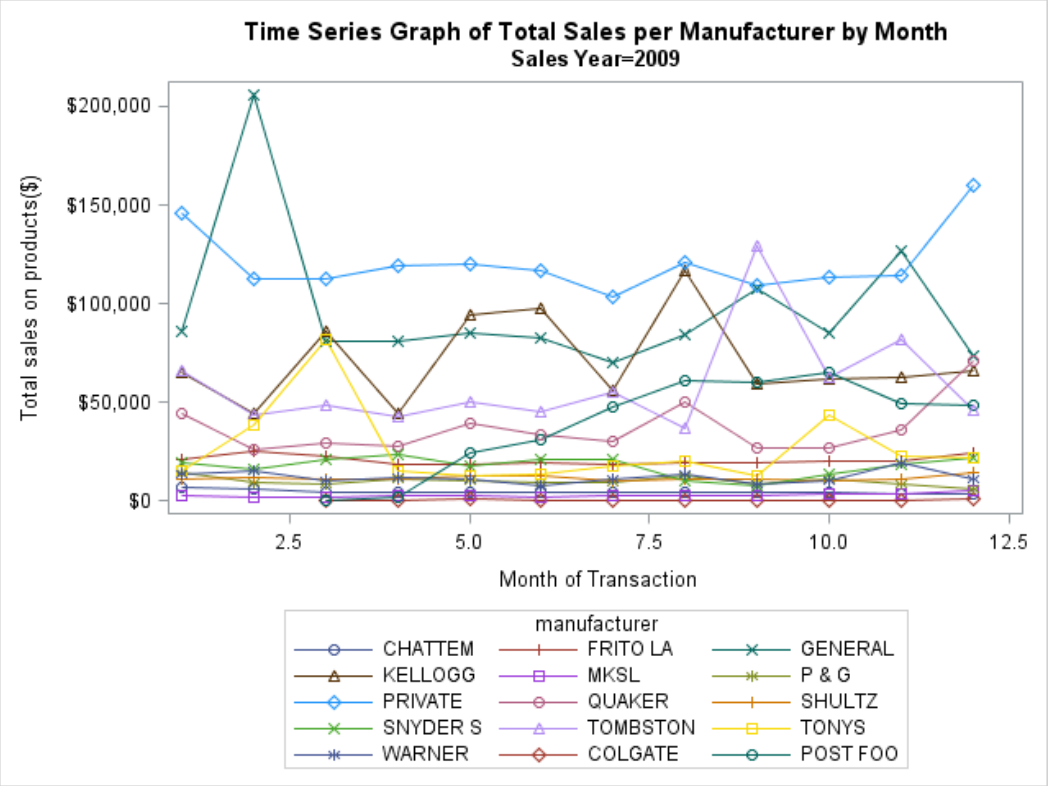


FIGURE 3

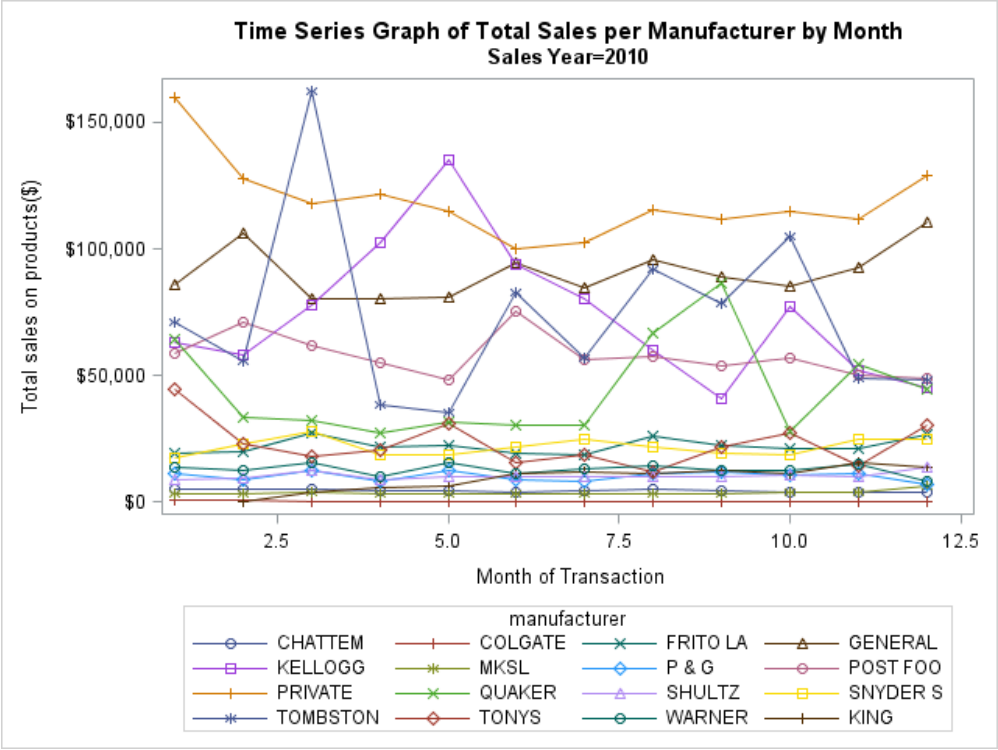


FIGURE 4

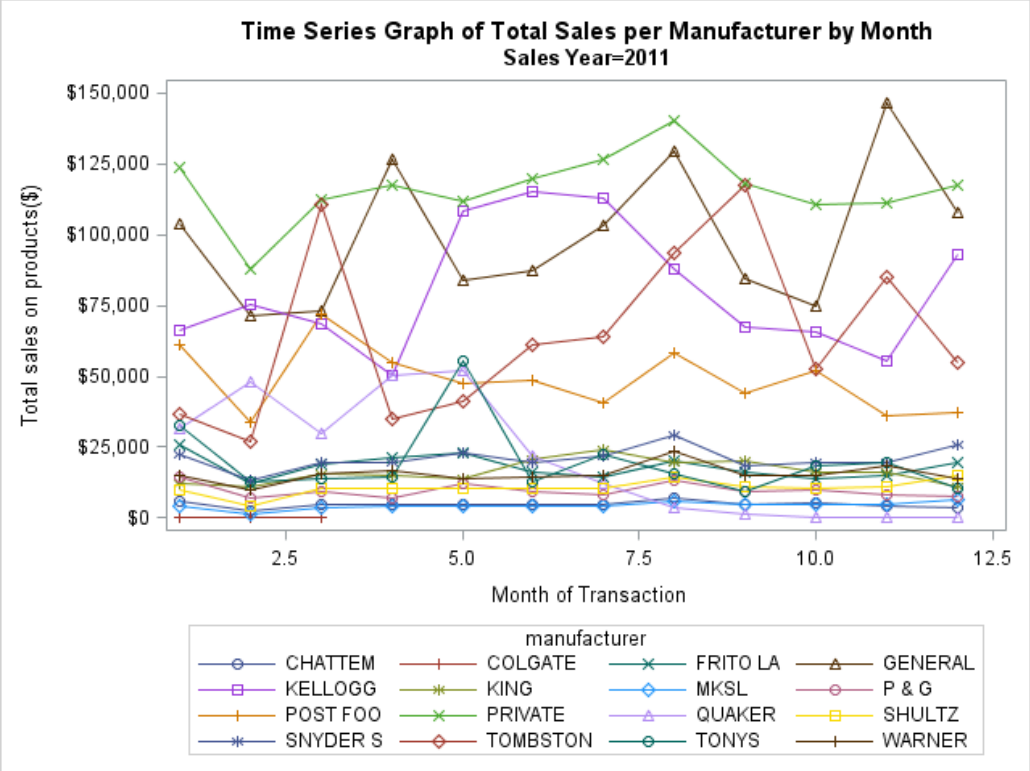


FIGURE 5

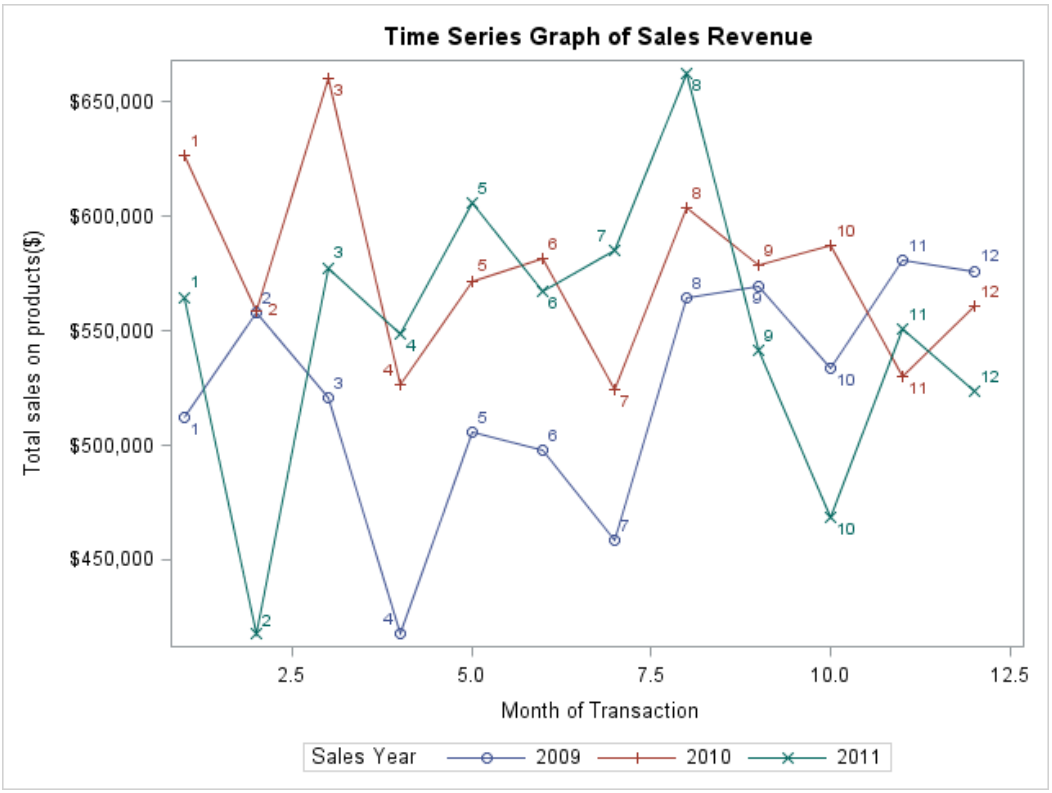
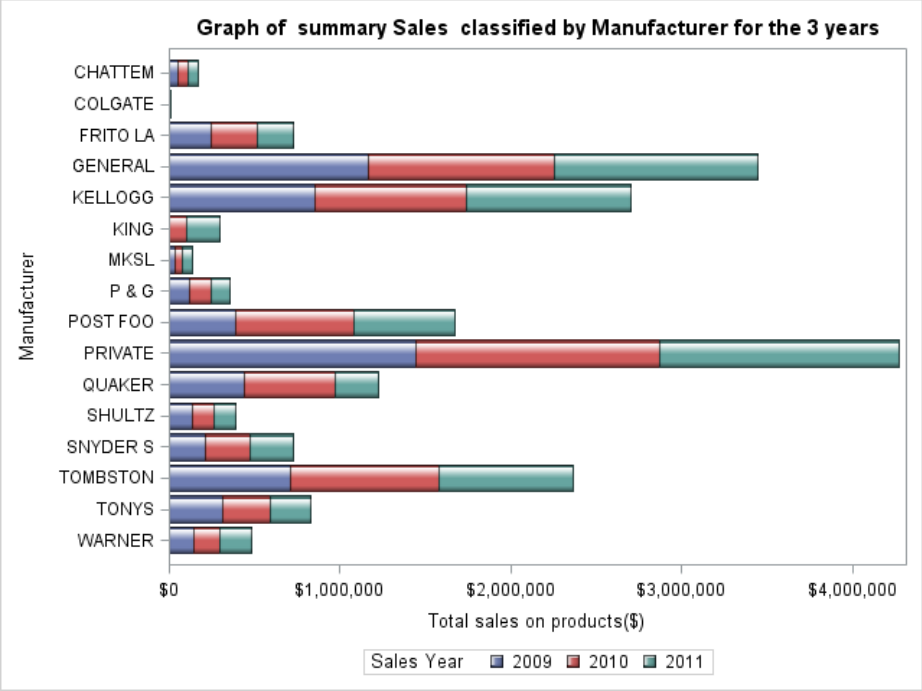


FIGURE 6


Graph of summary Sales classified by Manufacturer for the 3 years
Manufacturer
CHATTEM
COLGATE
FRITO LA
GENERAL
KELLOGG
KING
MKSL
P & G
POST FOO
PRIVATE
QUAKER
SHULTZ
SNYDER S
TOMBSTON
TONYS
WARNER
$0
$1,000,000
$2,000,000
$3,000,000
$4,000,000
Total sales on products($)
Sales Year
2009
2010
2011