

Design and Implementation of a Data Warehouse for a Retail Store with Store-level Data

Texas A&M University

George, Alpha

Rafi, Parvez

Shetty, Sharvil

Table of Contents

A.	Introduction	4
1.	Potential impacts	4
2.	Details about the Data	5
2.1.	Understanding of the data.....	5
2.2.	Metadata description for all OLTP files.....	6
2.3.	Entity-Relationship diagram.....	10
3.	DOMAIN UNDERSTANDING	11
B.	BUSINESS QUESTIONS AND THEIR SUBSTANTIATIONS AND EXPLANATIONS	13
1.	BUSINESS QUESTIONS	13
1.1.	WHAT IS THE TREND OF BEER SALES DURING CHRISTMAS HOLIDAY SEASON?	13
1.2.	WHICH STORES HAVE MORE POPULARITY AMONG KIDS AND ELDERLY GROUPS?	14
1.3.	WHICH STORES ATTRACT PEOPLE WHO EARN BELOW POVERTY LINE?	16
1.4.	WHAT IS THE EFFECT OF INTRODUCING COUPONS ON TOTAL NUMBER OF CUSTOMER VISITS?	17
1.5.	WHAT IS THE TREND OF A PRODUCT DEMAND IN DIFFERENT PRICE-TIERS?	18
C.	INDEPENDENT DATA MARTS DESIGN USING KIMBALL’S APPROACH.....	20
1.	PROPOSED SCHEMA	20
2.	KIMBALL’S MATRIX FOR DATA MARTS	20
3.	STAR SCHEMA REPRESENTATION	21
4.	MAPPING TABLES	23
5.	Business question justification corresponding to proposed schema	26
D.	DATA INTEGRATION	31
1.	Data Quality Issues with Dominick’s Finer Food Data	31
2.	ETL Development Plan	33
3.	ETL Implementation.....	34
3.1.	Determine the target data.....	34
3.2.	Determine Source data	35

3.3.	Mapping tables for staging and data mart loads.....	36
3.4.	Comprehensive Data Extraction Rules.....	45
3.5.	Screenshots of Data Staging area	47
3.6.	Data Transformation and Cleansing Rules.....	50
3.7.	Plan for Aggregation	52
4.	Data Loading	54
4.1.	Mapping definition describing the source to end table for all dimension and fact tables.....	58
4.2.	SQL statements used for the ETL operations.....	61
4.3.	Staging and Data Mart table screen shots.....	62
E.	BI REPORTING	70
1.	BI Reports Design and Implementation	70
1.1.	Reporting plan.....	70
1.1.1.	Determining all target reports that satisfy business questions	70
1.2.	Mappings from the tables in the data marts to the attributes in the report	72
1.3.	Cube from SSAS and Report from SSRS on top of SSAS for Question 1.....	77
1.4.	SSRS Reporting from Independent Data Mart using for Question 2	86
1.5.	Report generated using Report builder 3.0 for Question 3	92
1.6.	Report generated using Report builder 3.0 for Question 4	97
1.7.	Cubes using SSAS from an independent data mart for Question 5	101
	References	106

A. Introduction

Dominick's was a retail store chain founded in 1918 and headquartered in Illinois. The data being analyzed in this report is from Dominick's Fine Food's (DFF) database through Chicago Booth collected after both formed a partnership for store-level research into shelf management and pricing. The data encompasses 25 product categories sold at DFF's 100 retail chain stores located in the Chicago metropolitan area over the years from 1989 to 1997. The research is of the data and developing a data warehouse on top it, would help DFF systematically analyze the factors that influence product sales. This project report aims at formulating business questions that aim to analyze these factors and help in solving key business problems such as follows.

1. Potential impacts

a. Proper shelf-management for sales growth. The sales increase tactics can broadly be classified under two categories.

- i. *Out-of-store tactics*- they include tactics to attract more consumers or retain existing ones against external competitors.
- ii. *In-store tactics*- it includes tactics to increase sales once consumers are inside the stores.

Shelf-management is an in-store tactic, which focuses on how retailers can boost sales through store-level shelf-management. The analysis of data through UPC scanners makes it *possible to understand* heterogeneity of local area demand. One of the challenges that DFF faces is how it should allocate shelf space to the multitude of products they sell across stores.

b. Effect of consumer demographics on product sales. Another key problem DFF needs to address is how distribution of consumer's age, economic health, household sizes affect the sales across product categories. Answers to these problems would help DFF in targeted marketing.

c. Impact of location of stores on sales. DFF has more than 100 retail stores in Chicago. The store-level research on product sales also aims at answering how positioning of stores translates into revenue for DFF and what makes shoppers choose one store over another.

d. Making pricing decisions based on seasons. Retail stores usually see spikes in sales during vacation seasons such as Thanksgiving and Christmas. It is important for DFF to determine optimal prices of the products that would increase profit margin during these seasons.

- e. **Effect of price promotions.** DFF has been rolling out coupons across product categories for promotions. They need to analyze whether price promotions through strategies like Coupons help any sales growth.

2. Details about the Data

2.1. *Understanding of the data*

As a part of the project, we are planning to do analysis on data collected from James M. Klits Center, University of Chicago Booth School of Business. Total size of this data adds up to 4.76 GB needs lot of changes to convert this dirty data to perform meaningful analysis. The data comprises of 9 years of store level data of more than 3500 UPCs, which were sold through around 100 stores across United States. Most of the Stores are located in Chicago area. We need to analyze the data from 1989-1994 for same number of stores. The complete product line is classified in to 29 different categories.

In total, there are four data files and they can be classified in to two categories namely General files and Category files. These four data files are available in .csv format for the analysis of the sales of Dominick's FF.

a. **Customer Count Files**

- This file contains the information about in-store traffic of all the stores, which were compiled, on a weekly basis from scanners located at each store in DFF.
- This table contains the information about the sale information of each product categories of Beer, Meat, grocery and Dairy etc. store wise on weekly basis.
- This file also contains the separate information of each above product purchased using Coupons at each store in DFF chain on weekly basis.

b. **Store-Specific Demographics**

- This file contains the store wise information of all the customers purchased the products from DFF on weekly and demography basis.
- Information present in this file is obtained by mapping the customer information with the Census information collected by US government for Chicago Metropolitan area.
- Various demography information available in this file include age groups, household income, and number of dependent members, employment status, and retired status of every customer.
- Demographic information is the most important in later stages of the project for Data Warehousing and building different store wise strategies targeting different demography of people. This file has very importance in framing the questions in this phase of the project.

c. UPC Files

- As the name indicates UPC means Unique Product Code. Each Product is mapped to a UPC and other information related to that product.
- Complete mapping of UPC code to product is available in the *List of all UPCs in the category* table. This information can be used for product specific strategies for DFF.

d. Movement Files

- This file contains category wise weekly movement of each product in DFF.
- Information in this file give clear idea about profit margin on each product. This in turn will give idea about strategies need to be adopted by DFF to reduce losses and increase profit. This very important in Business point of view.
- If the Sale quantity is predictable after analyzing weekly sale data, it will give strategic advantage to the inventory department of DFF in preparing for peak and off seasons.

In addition to the above data sources, we have “Weeks Decode Table” which gives the week to date mappings useful for analysis. This is very useful in formulating different business strategies.

2.2. Metadata description for all OLTP files

a. Please find below the description for each attribute in **ccount** file.

Variable	Description	Variable	Description
DATE	Date of the Observation	FTGITAL	Food-to-Go Italian Sales in Dollars
Week	Week Number	GM	General Merchandise Sales in Dollars
Store	Store Code	GMCROUP	General Coupons Redeemed
BAKROUP	Bakery Coupons Redeemed	GROCCROUP	Grocery Coupons Redeemed
BAKERY	Bakery Sales in Dollars	GROCERY	Grocery Sales in Dollars
BEER	Beer Sales in Dollars	HABA	Health and Beauty Aids Sales in Dollars
BOTTLE	Bottle Sales in Dollars	HABACROUP	Health and Beauty Aids Coupons Redeemed

BULK	Bulk Sales in Dollars	JEWELRY	Jewelry Sales in Dollars
BULKCOUP	Bulk Coupons Redeemed	LIQCOUP	Liquor Coupons Redeemed
CAMERA	Camera Sales in Dollars	MANCOUP	Manufacturer Coupons Redeemed
CHEESE	Cheese Sales in Dollars	MEAT	Meat Sales in Dollars
CONVFOOD	Conventional Foods Sales in Dollars	MEATCOUP	Meat Coupons Redeemed
COSMCOUP	Cosmetics Coupons Redeemed	MEATFROZ	Meat-Frozen Sales in Dollars
COSMETIC	Cosmetics Sales in Dollars	MISCSCP	Misc. Coupons Redeemed
CUSTCOUN	Customer Count	MVPCLUB	MVP
DAIRCOUP	Dairy Coupons Redeemed	PHARCOUP	Pharmacy Coupons Redeemed
DAIRY	Dairy Sales in Dollars	PHARMACY	Pharmacy Sales in Dollars
DELI	Deli Sales in Dollars	PHOTCOUP	Photo Coupons Redeemed
DELICOU	Deli Coupons Redeemed	PHOTOFIN	Photo
DELIEXPR	Deli Express Sales in Dollars	PRODCOUP	Produce Coupons Redeemed
DELISELF	Deli Self Service Sales in Dollars	PRODUCE	Produce Sales in Dollars
FISH	Fish Sales in Dollars	PROMCOUP	Promotion Coupons Redeemed
FISHCOUP	Fish Coupons Redeemed	PROMO	Promotion Sales in Dollars
FLORAL	Floral Sales in Dollars	SALADBAR	Salad Bar Sales in Dollars
FLORCOUP	Floral Coupons Redeemed	SALCOUP	Salad Coupons Redeemed
FROZCOUP	Frozen Items Coupons Redeemed	SPIRITS	Spirits Sales in Dollars
FROZEN	Frozen Items Sales	SSDELICP	Self Service Deli Sales in Dollars
FTGCCOUP	Food-to-Go Coupons Redeemed	VIDCOUP	Video Coupons Redeemed
FTGCHIN	Food-to-Go Chinese Sales in	VIDEO	Video Sales in Dollars

	Dollars		
FTGICOUP	Food-to-Go Coupons Redeemed	VIDEOREN	Video Rentals (Dollar Amounts)

b. Please find below the description for each attribute in **Demography** file.

Variable	Description	Variable	Description
age9	% Population under age 9	retired	% of Retired
age60	% Population over age 60	unemp	% of Unemployed
ethnic	% Blacks & Hispanics	wrkch5	% of working women with children under 5
educ	% College Graduates	wrkch17	% of working women with children 6 - 17
nocar	% With No Vehicles	nwrkch5	% of non-working women with children under 5
income	Log of Median Income	nwrkch17	% of non-working women with children 6 - 17
incsigma	Std dev of Income Distribution (Approximated)	wrkch	% of working women with children
hsizeavg	Average Household Size	nwrkch	% of non-working women with children
hsize1	% of households with 1 person	wrkwch	% of working women with children under 5
hsize2	% of households with 2 persons	wrkwnch	% of working women with no children
hsize34	% of households with 3 or 4 persons	telephn	% of households with telephones
hsize567	% of households with 5 or more persons	mortgage	% of households with mortgages
hh3plus	% of households with 3 or more persons	nwhite	% of population that is non-white
hh4plus	% of households with 4 or more	poverty	% of population with income

	persons		under \$15,000
hhsingle	% of households with 1 person	shopcons	% of Constrained Shoppers
hhlarge	% of households with 5 or more persons	shophurr	% of Hurried Shoppers
workwom	% Working Women with full-time jobs	shopavid	% of Avid Shoppers
sinhouse	% Detached Houses	shopstr	% of Shopping Stranges
density	Trading Area in Sq Miles per Capita	shopunft	% of Unfettered Shoppers
hval150	% of Households with Value over \$150,000	shopbird	% of Shopper Birds
hval200	% of Households with Value over \$200,000	shopindx	Ability to Shop (Car and Single Family House)
hvalmean	Mean Household Value (Approximated)	shpindx	Ability to Shop (Car and Single Family House)
single	% of Singles		

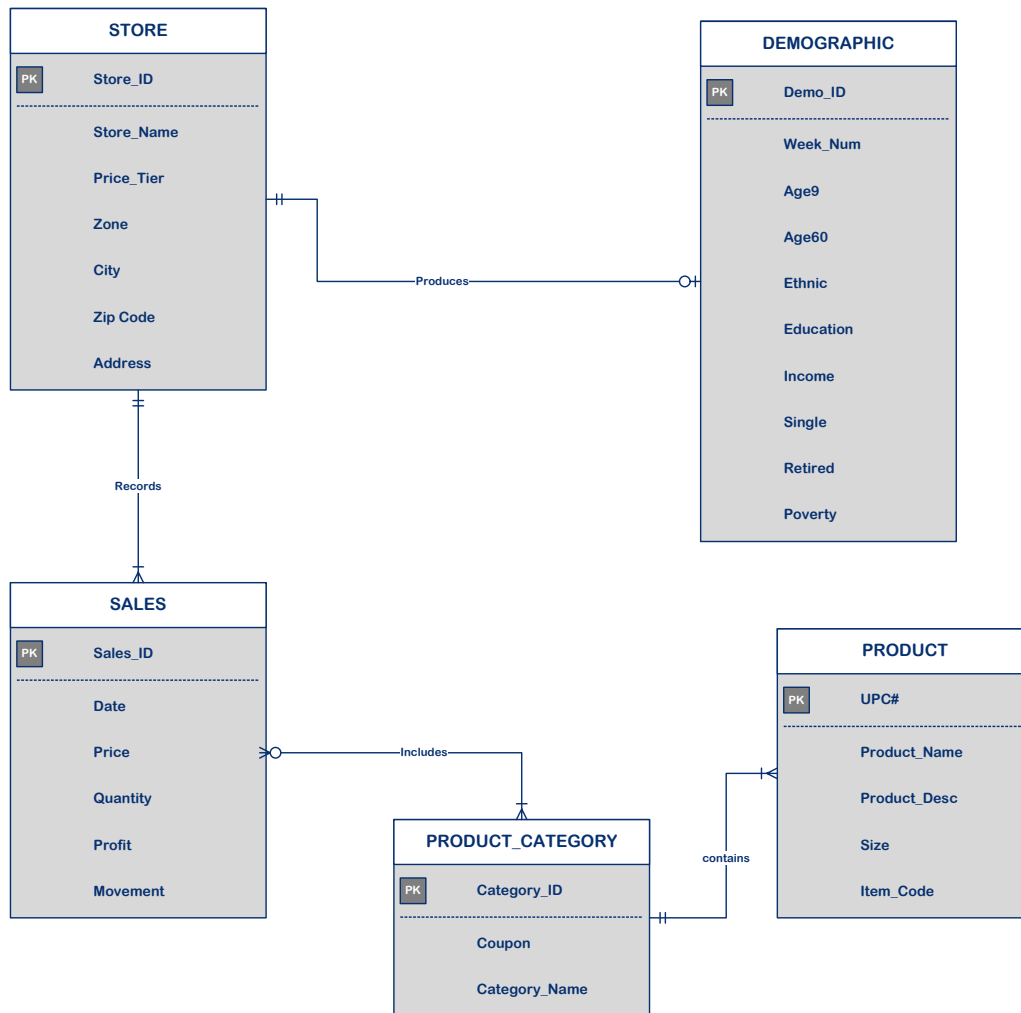
c. Please find below the description for each attribute in **UPC** file

Variable	Description	Variable	Description
upc	UPC Number	descrip	Product Name
com_code	% Population over age 60	size	Product Size
nitem	Dominick's item code	case	Number of items in a case

d. Please find the description for each attribute in **Movement** file

Variable	Description	Variable	Description
upc	UPC Number	qty	Number of item bundled together
store	Store Number	profit	Gross margin
week	Week Number	sale	Sale code (B, C, S)
move	Number of unit sold	ok	1 for valid data, 0 for trash
price	Retail Price		

2.3. Entity-Relationship diagram



3. DOMAIN UNDERSTANDING

We had a pool of research papers to understand how significant and relevant data can be derived from past sales data. Research papers also gave insight about methods by which we can build business strategies on top of the derived data. It was beyond the scope of the project to read and understand all the papers from the pool of research papers. From the complete list, we read three and understood the strategies adopted during the design and implementation of Data Warehousing. Please find our understandings about each paper detailed below.

In their research Nevo, et al [1] summarize the effect of Coupons in the increase or decrease in sales of any retail chain. This paper discusses the scenarios in which the coupons are introduced by a Manufacturers and the effect of it in detail. Paper also discusses that giving coupons is better than spending substantial amount in advertisements. This paper also explains how the price of commodities whose quality is not comparable to its competitors will decrease in long term .This is very important in formulating our strategies in later stages of the project for improving sales based on the meaningful data derived from data warehousing design methodologies.

Nevo, et al also discusses different reasons behind a manufacturer introducing discount/coupon on a particular product. Firstly, it can be for introducing a new product. This will give a larger pool of customers a taste/experience of the new product. Secondly coupons are introduced to boost up sales of a particular product which is lagging behind its customers or if it is being perished in inventory. This is explained under the heading relationships between Coupons and Shelf price. In addition, the paper discusses how the perception of customer about a product can change if the coupons are made available continuously. Effects and correlation of price and coupons in different situations are derived and explained mathematically in the complete paper. This summarizes the complete project.

Sivakumar, et al [2] talk about implications of quality-tier competition by developing a conceptual framework to investigate how optimal pricing decisions can be done the level of brand quality to which it belongs. This paper explained how comparative study could be done on 2 different products about its pricing. It also discusses the implication of the framework introduced and steps involved in the implementation of the framework. This will be very helpful in adopting strategies for pricing the items in DFF considering the competition faced by the product after getting results from the data warehouse we are planning to design in the upcoming stages of the project.

Sivakumar, et al also give insight into Customer decision model explaining what will be the possible decisions of a customer comparing the price and brand category of 2 products competing each other. This idea is very useful in determining what the possible price of each product is, considering the utilities derived from it. The framework described in this paper proposes the empirical distribution of possible prices for a product. It also describes the profit available from each price in the distribution and managerial implication of the same. In total, the paper draws a complete relation between the academic researches in marketing practices of pricing a product.

Lira, et al [3] suggest whether the prices in retail sector decrease when there is a positive demand in the product. This explained in the backdrop of countercyclical relationship of in Chile between cost and increased demand of set of retail products. This paper discusses about various theories analyzing the effect and cause of prices in each theory scenario. Author discusses theories effecting the prices at Producer/Supplier level and Retail level. For example, pro cyclical Price elasticity proposed by Warner Barsky states that Producer price could drop for products experiencing increase in the elasticity of demand. This can happen due to transferring of elasticity of demands to suppliers in periods of generalized shocks in demand. If similar situation arises for a product in DFF for a particular week DFF management transfer the benefit to customers by applying the same principle.

So similarly applying other theories in the paper, equations can be written for each product in DFF on weekly basis considering the changes in sales of the particular product in previous week. Even though applying equations for pricing on each product is not in the purview of the project even in later stages, we can apply these in one or two products if the time permits. The paper also says that above theories stand null and void when the demand of a product is artificially increased by applying coupons. All theories are explained with the variation in price of products in a retail chain at Chile. Results of the study proves that price of set of retail products fall when the demand increases, which is likely to be similar in case of DFF as the scenarios are same in both the cases.

B. BUSINESS QUESTIONS AND THEIR SUBSTANTIATIONS AND EXPLANATIONS

1. BUSINESS QUESTIONS

1.1. WHAT IS THE TREND OF BEER SALES DURING CHRISTMAS HOLIDAY SEASON?

Data Table:

Week Numl	Wine Sales
10	618104.91
11	299353.29
12	271369.49
13	306332.89
14	490994.06
15	687024.94
16	633326.14
17	196033.36
18	224434.69
19	240542.52
20	254735.07
Grand Total	4222251.36

Table.2. Weekly sales trends for beer around Christmas Week

Graphical Analysis:

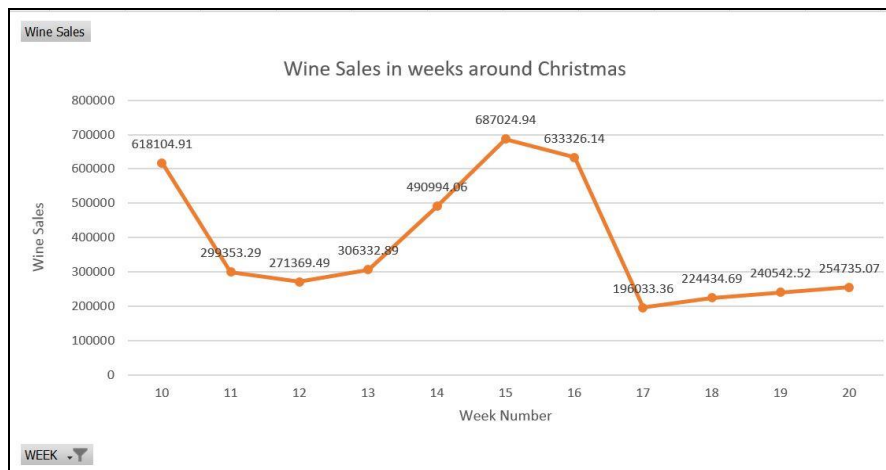


Fig. 2. Weekly sales trends for beer around Christmas Week

Explanation: Some products may see significant boost in sales during festive/vacation seasons. One such product is beer which saw sudden growth from week 13 through week 17 (Christmas through New Year). Such an analysis will help DFF identify all the products which usually see seasonal growth in sales. The supplies of these products would need to be managed effectively

during these seasons. Proper shelf-management has always been one of important concerns for retails businesses. This type of analysis would help in better management of the products that would feature on store shelves.

1.2. WHICH STORES HAVE MORE POPULARITY AMONG KIDS AND ELDERLY GROUPS?

Row Labels	Sum of AGE9	Sum of AGE60
2	11.75%	23%
4	9.51%	26%
5	14.14%	12%
8	12.32%	25%
9	10.35%	27%
12	10.57%	18%
14	12.96%	21%
18	11.01%	27%
21	17.59%	7%
28	12.89%	21%
32	9.91%	25%
33	4.61%	13%
40	13.37%	18%
44	14.49%	19%
45	14.67%	13%
47	14.30%	13%
48	12.18%	10%
49	13.49%	19%
50	12.44%	15%
51	13.25%	18%
52	13.66%	15%
53	12.08%	30%
54	14.79%	9%
56	13.10%	19%
59	17.21%	11%
67	13.35%	22%

Table 3. Store-wise percentage visit of kids(age<9) and elderly people(age>60)

Graphical Analysis:



Fig. 3.A. Store-wise percentage visit of elderly people(age>60)



Fig. 3.B. Store-wise percentage visit of kids(age<9)

Explanation: The graphs above highlight the DFF stores which are most or least popular among kids and elderly group people. Retails companies require such analysis for targeted marketing and sales. For instance, the stores which are popular among kids may be supplied with more of young age merchandize. Similarly, the stores that are popular among elderly section may be supplied with merchandize that sell more to elderly people.

1.3. WHICH STORES ATTRACT PEOPLE WHO EARN BELOW POVERTY LINE?

Data Table:

Row Labels	% population below poverty
2	6%
4	4%
5	3%
8	5%
9	3%
12	17%
14	2%
18	8%
21	2%
28	4%
32	4%
33	9%
40	6%
44	3%
45	2%
47	3%
48	2%
49	3%
50	4%
51	4%
52	2%
53	6%
54	2%
56	3%
59	4%
62	2%
64	4%
65	3%
67	3%
68	10%

Table 4. Store-wise percentage of Poor People (annual salary < \$15,000)

Graphical Analysis:

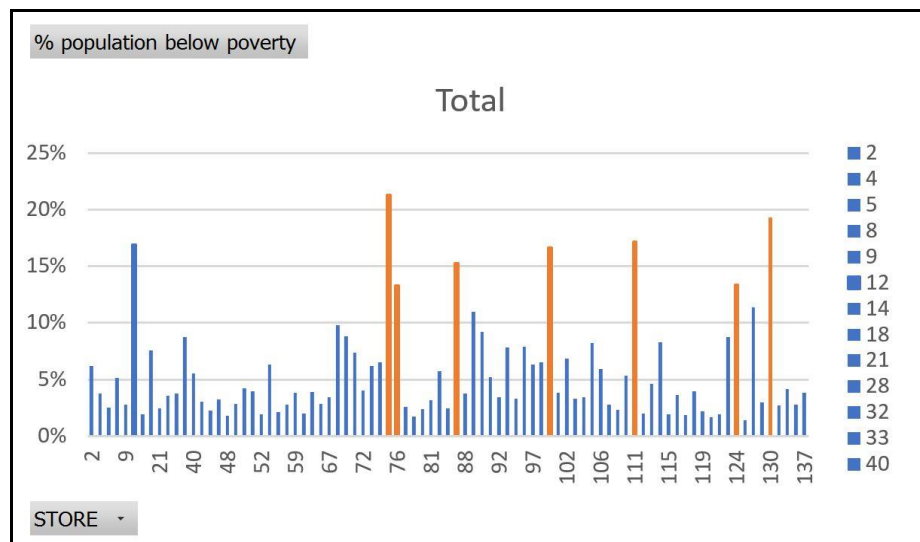


Fig. 4. Store-wise percentage of Poor People (annual salary<\$15,000)

Explanation: The plot above shows percentage of people below poverty line (income less than \$15,000) visiting stores of DFF. We realized that some of the stores were more popular among low income group people. Such an analysis is important for DFF for store-wide distribution of the stock which is usually consumed more by low-income group. For example, low income group usually consumes non-luxury goods more. If we can identify these stores, DFF would be able to make to distribute such goods more effectively.

1.4. WHAT IS THE EFFECT OF INTRODUCING COUPONS ON TOTAL NUMBER OF CUSTOMER VISITS?

Data Table:

Row Labels	Sum of CUSTCOUN
122	101563
123	102886
124	109889
126	69666
128	106820
129	62219
131	91458

Table 5.A. Count of customers per week

Row Labels	Sum of Total Coupons Redeemed
122	41284.44
123	30045.02
124	32632.5
126	27231.97
128	30897.89
129	24595.58
131	9281.81

Table 5.B. Count of Coupons used per week

Graphical Analysis:

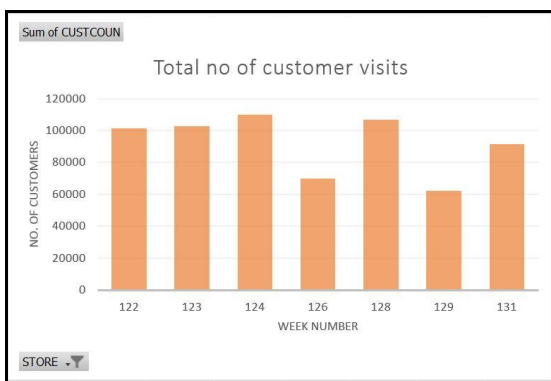


Fig. 5.A. Count of customers per week

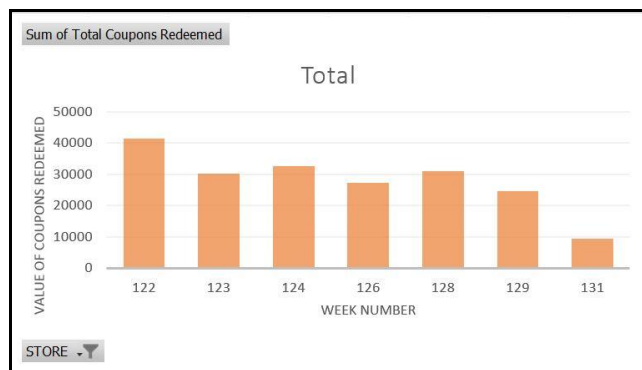


Fig. 5.B. Count of Coupons used per week

Explanation: With data and graph above help analyze the impact of coupons on count of customers who visited DFF stores over a certain period of observation. We assume that number

of coupons that were redeemed is in proportion to the coupons launched. The analysis is needed to understand how customer footfall varied as DFF launched promotional strategies like announcing coupons. We can repeat similar exercise for other promotional strategies and see how each impacts the customer count. This will help identify the effectiveness of such promotional strategies. The analysis will also help to see one promotional strategy fares against other marketing forces in their effect on customer visits.

1.5. WHAT IS THE TREND OF A PRODUCT DEMAND IN DIFFERENT PRICE-TIERS?

Data Table:

Row Labels	8	12	Grand Total
145	343	537	880
146	1286	649	1935
147	1311	609	1920
148	1094	454	1548
149	850	372	1222
150	811	553	1364
151	1034	576	1610
152	953	510	1463
153	944	507	1451
154	898	307	1205
155	1967	902	2869
156	931	486	1417
157	977	467	1444
158	974	514	1488
159	761	402	1163
160	780	458	1238
161	940	86	1026
162	622	371	993
163	691	482	1173
164	622	382	1004
165	635	303	938
166	649	361	1010
167	857	354	1211
168	528	279	807
169	507	314	811

Table. 6. Weekly sales of Cheese in store 8(medium-price tier) store 12(high-price tier)

Graphical Analysis:

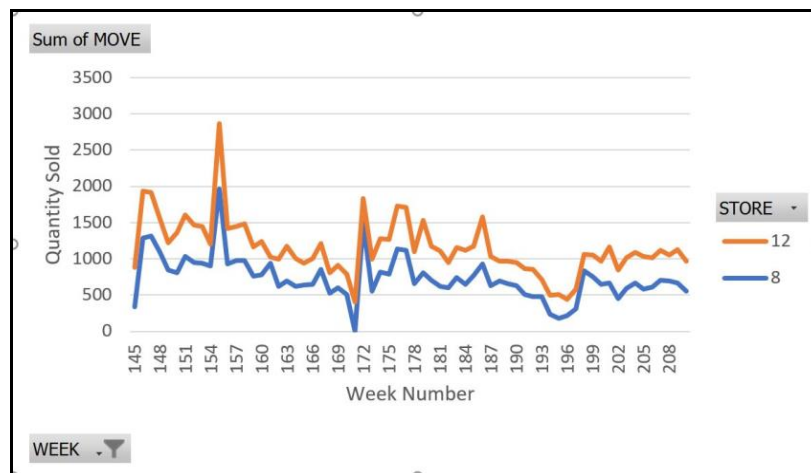


Fig. 6. Weekly sales of Cheese in store 8(medium-price tier) store 12(high-price tier)

Explanation: The plot above compares quantities of Cheese sold in two stores in difference price-tier zones. The store Dominick 12 is in high-price tier zone, while the store Dominick 8 is in low price tier zone. One possible explanation for the trend as above could be high-price zones being in rich neighborhoods where people can afford to buy more. Such an analysis is important in distributing merchandize based on price-tier zones. Out of 16 zones in which different stores are located, they can be divided into 4 price tiers viz. *CubFighter, Low, Medium, High*. In this analysis, we cover only medium-price and high-price tiers. Similar analysis may be expanded to other price tiers which would help in comparative study of the demand trends across these 4 price tiers.

C. INDEPENDENT DATA MARTS DESIGN USING KIMBALL’S APPROACH

1. PROPOSED SCHEMA

The schema being proposed to answer the above business questions is a combination of the five dimensions viz. Store, Product, Time, Demographic, Coupon and two fact tables- SALES_FACT which will hold the data from the MOVEMENT source file and STORE_VISITS_FACT which will hold the data from the CCOUNT source file.

The dimension STORE_DIM dimension contains details about all the DFF stores. This dimension is populated with data from the Dominic’s research project manual, which contains the details about each store under the DFF retail chain. It has an inherent hierarchy in Zone, City and Store number. The PRODUCT_DIM dimension contains the data related to all the products sold at DFF. The UPC files for each product act as the source for this data. It is a collection of data from each of the categories of products. The product category column is populated from the name of the csv file from which the data is being loaded. The TIME_DIM dimension contains the data related to the occurrence of an event. The attribute TIME_ID is an auto incremented surrogate key, acting as the primary key for the table. As stated in the manual, the data contains observations from week 0 to 400, based on this fact, the field WEEK is auto populated. The WEEK acts as the base to calculate MONTH and YEAR as described in the transformation for the time dimension. The COUPON_DIM dimension stores all the types of coupons there are available at the stores. This value is derived from the header column of the CCOUNT.csv file. It is a manual task to separate out the names of the fields, which represent the coupon usage in the dataset. The DEMOGRAPHIC_DIM dimension is derived from the DEMO.csv table. It stores all the demographics related to each store. It contains clearly defined statistics for various demographic factors, which are directly mapped from the source table.

The SALES_FACT table stores the values derived from the movement datasets. Product, Store and Time. The STORE_VISITS_FACT is designed to address business question 2, 3 and 4. This fact table enables us to answer the queries related to introduction of coupons and customer visits. The CUSTOMER_COUNT stores the number of customers who visited the store.

2. KIMBALL’S MATRIX FOR DATA MARTS

Data mart	Dimension				
	STORE_DIM	TIME_DIM	PRODUCT_DIM	DEMOGRAPHIC_DIM	COUPON_DIM
Sales	X	X	X		
Store visits	X	X		X	X

3. STAR SCHEMA REPRESENTATION

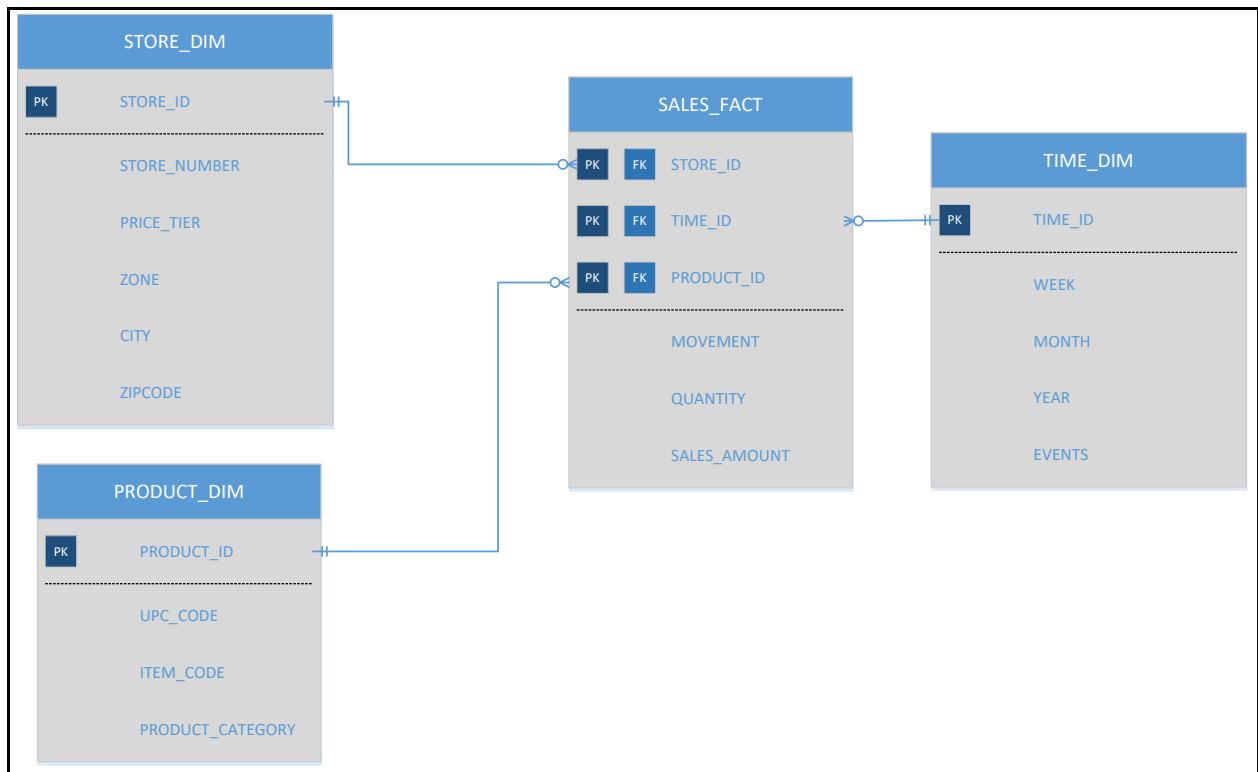


Fig: Star schema for Sales data mart

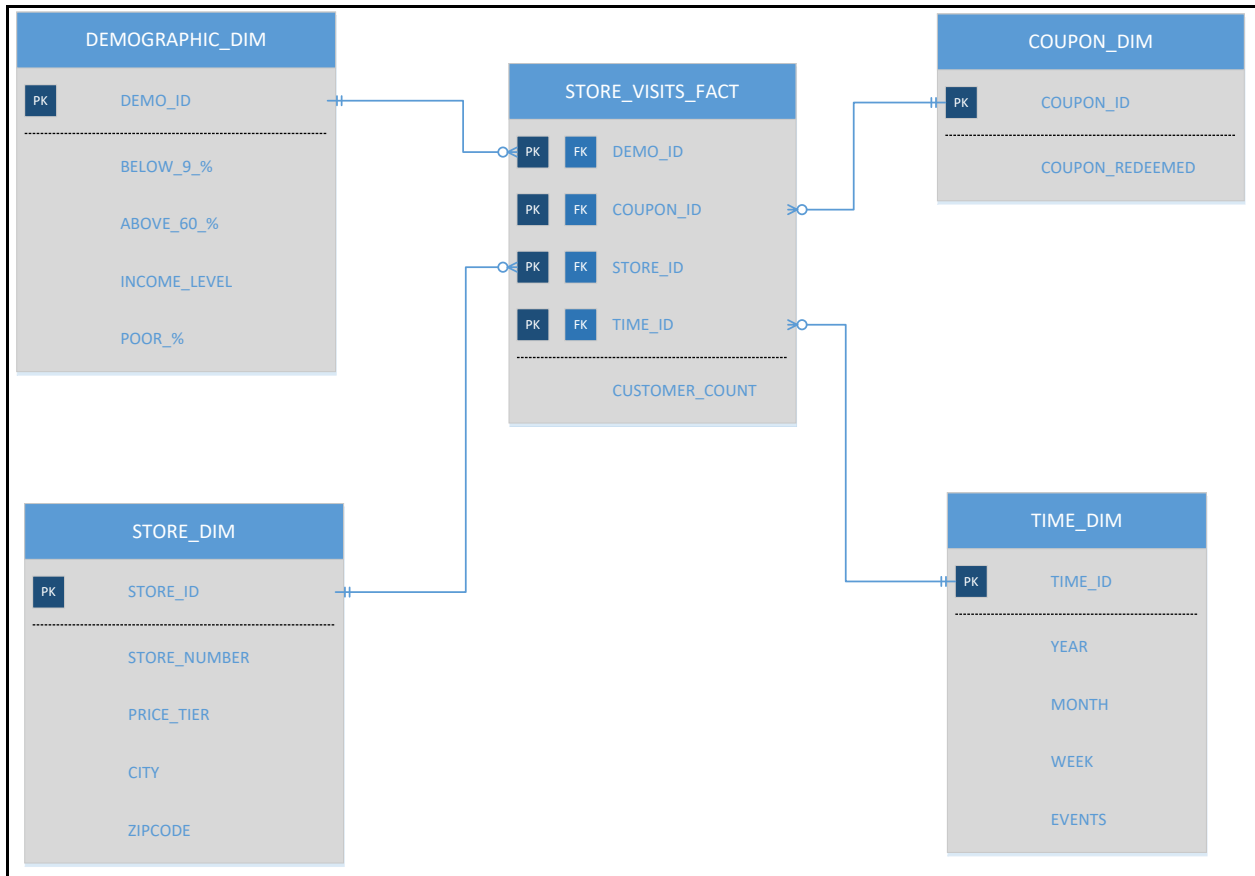


Fig: Star schema for Store visits data mart

4. MAPPING TABLES

Dimension: Product						
DW Target Table	DW Target Column	Target Datatype	Source System/Table	Source Column	Transformation/Business Rule	Error Handling Rules
PRODUCT_DIM	PRODUCT_ID	int		surrogate key		
	UPC_CODE	varchar	UPCXXX.csv	upc		
	PRODUCT_CATEGORY	varchar	UPCXXX.csv		XXX in source file name corresponds to the category code. e.g. If source csv file is named 'upcana', PRODUCT_CATEGORY will be 'ana'	
	ITEM_CODE	varchar	UPCXXX.csv	nitem		

Dimension: Demographic						
DW Target Table	DW Target Column	Target Datatype	Source System/Table	Source Column	Transformation/Business Rule	Error Handling Rules
DEMOGRAPHIC_DIM	DEMO_ID	int		surrogate key		
	INCOME_LEVEL	varchar	DEMO.csv	income	Income in source column is log of median income (M.I.), which ranges from 9.87 to 11.24. $e^{9.87} = \$19,341$ and $e^{11.24} = \$76,114$. We define 3 levels for INCOME_LEVEL: 1. <i>Low</i> : <10.3 (M.I. less than \$30,000) 2. <i>Medium</i> : 10.3 to 11.0(M.I. between \$60,000) 3. <i>High</i> : >11.0 (M.I. greater than \$60,000)	If income field is null in source table, update INCOME_LEVEL with NA
	POOR_%	float	DEMO.csv	poverty		
	BELOW_9%	float	DEMO.csv	age9		
	ABOVE_60%	float	DEMO.csv	age60		

Dimension: Coupon						
DW Target Table	DW Target Column	Target Datatype	Source System/Table	Source Column	Transformation/Business Rule	Error Handling Rules
COUPON_DIM	COUPON_ID	int		surrogate key		
	COUPON_REDEEMED	int	ccount		Sum up coupons redeemed across all product categories. i.e. COUPON_REDEEMED = BAKCOUP+BULKCOUP+COSMCOUP+FISHCOUP+...	for any missing coupons redeemed value across a product category, assume the value to be 0

Dimension: Store						
DW Target Table	DW Target Column	Target Datatype	Source System/Table	Source Column	Transformation/Business Rule	Error Handling Rules
STORE_DIM	STORE_ID	int		Surrogate key		
	STORE_NUMBER	int	Dominick's Stores	Store		
	PRICE_TIER	varchar	Dominick's Stores	Price Tier		If a row doesn't have a value for the price tier, set the target as 'NA'
	ZONE	int	Dominick's Stores	Zone		If a row doesn't have a value for the price tier, set the target as 'NA'
	CITY	varchar	Dominick's Stores	City		
	ZIPCODE	int	Dominick's Stores	Zip Code		

Dimension: Time						
DW Target Table	DW Target Column	Target Datatype	Source System/Table	Source Column	Transformation/Business Rule	Error Handling Rules
TIME_DIM	TIME_ID	int	Surrogate key			
	WEEK	int	Week's Decode Table	Week #	Populated from 1 to 400	
	MONTH	int	Week's Decode Table	Start	The start column is of format MM/DD/YY. Split it to get MM	
	YEAR	int	Week's Decode Table	Start	The start column is of format MM/DD/YY. Split it to get YY	
	EVENT	float	Week's Decode Table	Special Events		

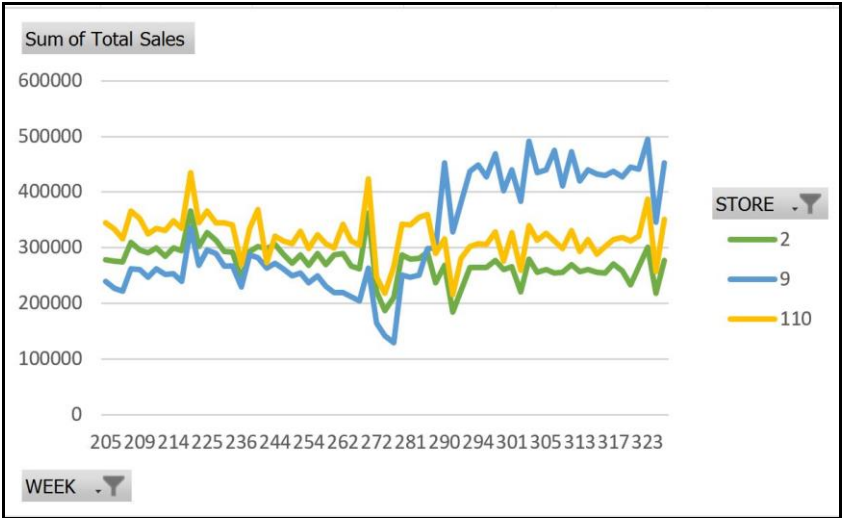
Fact: Store Visits						
DW Target Table	DW Target Column	Target Datatype	Source System/Table	Source Column	Transformation/Business Rule	Error Handling Rules
STORE_VISITS_FACT	DEMO_ID	int	DEMOGRAPHIC_DIM	DEMO_ID	Foreign key corresponding to primary key DEMO_ID of DEMOGRAPHIC_DIM dimension	
	COUPON_ID	int	COUPON_DIM	DEMO_ID	Foreign key corresponding to primary key COUPON_ID of COUPON_DIM dimension	
	STORE_ID	int	STORE_DIM	STORE_ID	Foreign key corresponding to primary key STORE_ID of STORE_DIM dimension	
	TIME_ID	int	TIME_DIM	TIME_ID	Foreign key corresponding to primary key TIME_ID of TIME_DIM dimension	
	CUSTOMER_COUNT	int	ccount	CUSTOMER_COUNT		

Fact: Sales						
DW Target Table	DW Target Column	Target Datatype	Source System/Table	Source Column	Transformation/Business Rule	Error Handling Rules
SALES_FACT	STORE_ID	int	STORE_DIM	STORE_ID	Foreign key corresponding to primary key STORE_ID of STORE_DIM dimension	
	TIME_ID	int	TIME_DIM	TIME_ID	Foreign key corresponding to primary key TIME_ID of TIME_DIM dimension	
	PRODUCT_ID	int	PRODUCT_DIM	PRODUCT_ID	Foreign key corresponding to primary key PRODUCT_ID of PRODUCT_DIM dimension	
	MOVEMENT	int	movement files	move		
	QUANTITY	int	movement files	qty		
	SALES_AMOUNT	float	movement files			SALES_AMOUNT = Price * Move / Qty for a given row

5. Business question justification corresponding to proposed schema

Please find the explanation of how we can answer all the five business question using the above two schemas below.

BQ 1: What is the trend of beer sales during Christmas holiday season?



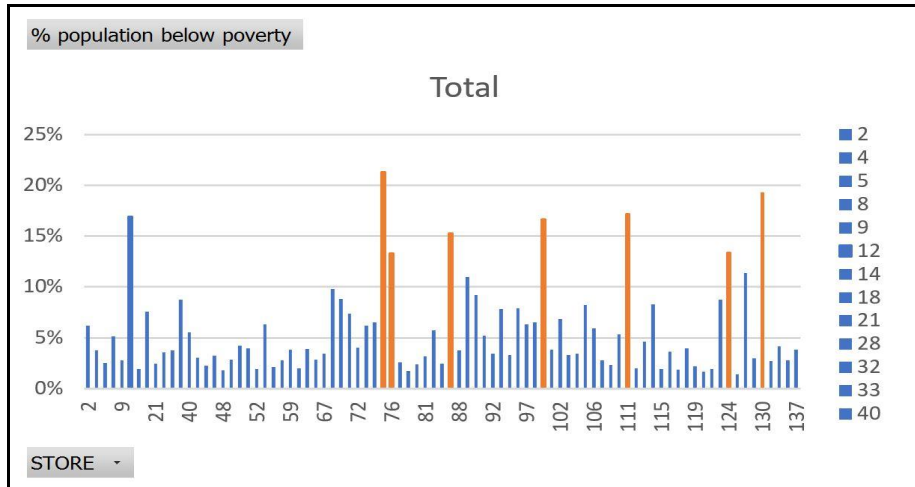
This business question is addressed by analyzing the Beer sales during the Christmas weeks as shown in the above graph. Same analysis can also be done with the Sales data mart shown above. Here the product under consideration is Beer, information related to that can be obtained from PRODUCT_DIM dimension table and Week related information could be taken from WEEK attribute loaded in TIME_DIM dimension table. Beer Sales related information corresponding to each week during Christmas week can be mapped from SALES_FACT fact table in Sales data mart.

BQ 2: Which stores have more popularity among kids and elderly groups?



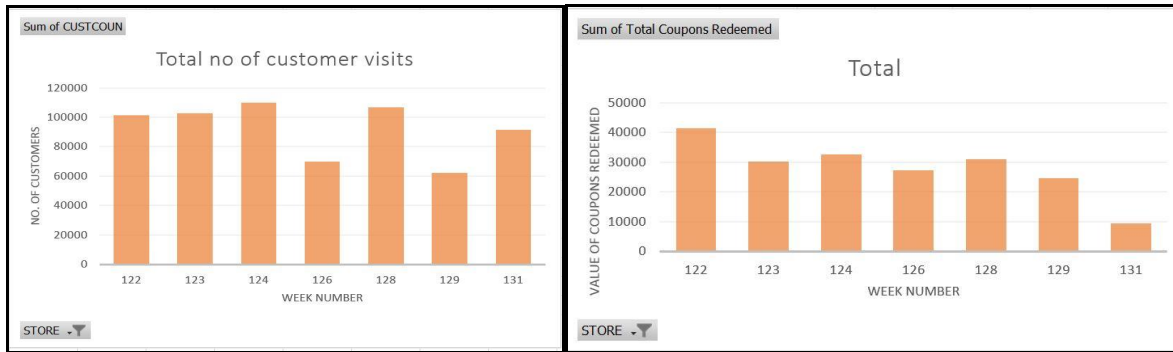
Store visits data mart can be used for analyzing this business question. Number of people below age 9 (BELOW_9_ %) and above 60 (ABOVE_60_ %) are attributes in DEMOGRAPHIC_DIM dimension table and Store Number plotted the above graph can be plotted from attribute STORE_NUMBER from STORE_DIM dimension table. Comparing the two values, the number of store visits, CUSTOMER_COUNT is obtained from STORE_VISITS_FACT fact table. Above graph will help us analyze just the popularity among people above age 60. For answering this business question completely, popularity of stores among kids and elderly group will be analyzed separately in similar way.

BQ 3: Which stores attract people who earn below poverty line?



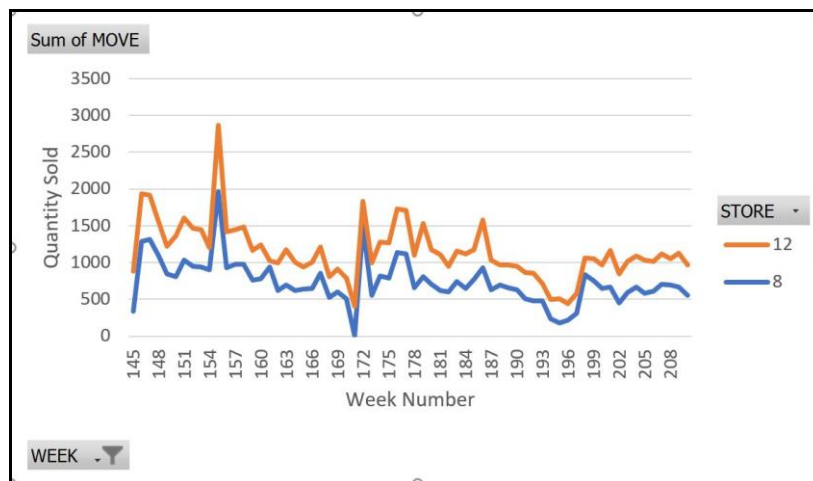
This business question can be analyzed similar to Business Question 2. Same Store visits data mart can be used for analysis. IS_POOR attribute is available in DEMOGRAPHIC_DIM table and store number STORE_NUMBER is obtained from STORE_DIM table. For each corresponding values of STORE_NUM and IS_POOR we can get the CUSTOMER_COUNT from STORE_VISITS_FACT table given in the schema. This data can be used to analyze store attraction among people below poverty line.

BQ 4: What is the effect of introducing coupons on total number of customer visits?



This is comparatively complex business question to analyze even though it is easy to understand. For analyzing the question we need to plot and compare the Total number of customer visits and Total number of coupons redeemed on a weekly basis. Comparing the trend in both the graph's we can reach at a conclusion of effect of coupons on customer visits. So here the Store visits data mart can be used again for analysis. CUSTOMER_COUNT attribute from STORE_VISITS_FACT fact table can be plotted against the WEEK attribute loaded in TIME_DIM dimension table from the schema. Similarly, for plotting the second graph we need calculate the value derived from the summation of coupon sales from different product category, which is given in CCount table. This summed up value is stored in COUPON_DIM dimension table as COUPON_REDEEMED attribute. COUPON_REDEEMED attribute from COUPON_DIM dimension table can be used for plotting the second graph. So now with the two graph's plotted from the data loaded to dimension tables and fact tables we can easily find the effect of introducing coupons in total number of customer visits.

BQ 5: What is the trend of a product demand in different price-tiers?



This business question can be answered by plotting the weekly sales data of each Shop tiers (A, B, C). For this Business question, we can use Sales data mart. In Sales data mart the PRICE_TIER information is loaded in STORE_DIM dimension table and WEEK related data is stored in TIME_DIM dimension table. All the sales related information for all the shop price tiers can be collected from the SALES_AMOUNT attribute in SALES_FACT fact table. This information is plotted for each shop tiers and the distribution of sales in each shop tier can easily be analyzed from the plot.

D. DATA INTEGRATION

1. Data Quality Issues with Dominick's Finer Food Data

The dataset provided by DFF has been highly useful to answer our business questions but the data quality had to be improved in order to make the data useful. We discovered following issues with the data while developing our data warehouse.

Group	Quality	Issues Considered	Data Quality Problem encountered with DFF
Relation to the other data	Referential Integrity	Do tables in DFF datasets have table/file references intact?	Referential integrity in between tables/files was mostly missing, but for some of the tables we linked them using common columns. To link ccount and movement tables we used store numbers. To link UPCXXX files with sales data in ccount, we had to extract product categories(XXX) from files names. Many product categories e.g. Beer which we extracted from UPC files had corresponding movement/ccount data missing.
	Cardinality	Is the structure of relationships among entities and attributes maintained consistently?	The relationships between entities and attributes were highly inconsistent e.g. UPC numbers of products in Movement data and the UPC files were inconsistent for some of the cases.
Structure of fields	Format	Do values follow consistent formatting standards?	The UPCXXX.csv files names were consistent and helped us find list of product categories. Date field in ccount table even though was not in an easily extractable format, but the format throughout the file was consistent i.e. XXXXXX.
	Standard	Are data elements consistently defined and understood?	Store number followed standard integer format which helped linked tables like ccount and demographic. The format of date field in ccount and that in start, end dates in week table were dissimilar. The date in ccount was

			formatted as YYYYMMDD, while in week table is given as MM/DD/YY.
	Consistent	Do values represent same meaning across systems and files?	Most of the attributes e.g. Store Number, Week Number had consistent meaning across the tables. The attributes were clearly explained in the data manual of DFF.
Content within data values	Complete	Is all necessary data present?	Many data values were missing that we needed to answer our business questions with. E.g. Many date values were missing from ccount. Also, UPC data contained more product categories than we had sales figures for.
	Accurate	Does the data accurately represent reality?	Most of the data accurately represented data DFF's business. However, demographic table had demographics in percentage terms instead of absolute number of customers. The count of customers as an additional data column would have provided better understanding.
	Valid	Do data values fall within acceptable ranges defined by the business?	Dominic stores had Store numbers starting from 2, while many ccount records had store values as 0 and 1. We had excluded these records for business question where we analyzed store-wise sales trends.
	Fit for purpose	Is the information valuable to the business? Does the data convey information that can intelligently be consumed by the business?	DFF contains an exhaustive amount of data collected over a long period of more than 5 years. Such data can be used to provide a lot of insights. The data is highly valuable to the business as it can help in understanding marketing trends, demand-supply relations and thus would help in formulating more effective sales strategies.

2. ETL Development Plan

Based upon the previous reports developed, ETL Development Plan is designed to outline the roadmap for the data load process into the data warehouse.

The proposed plan is presented as:

- Determine the target data
- Determine the source data
- Mapping tables for staging and data mart loads
- Comprehensive data extraction rules
- Data staging area and screen shots
- Data transformation and cleansing rules
- Plan for aggregate tables
- Procedures for data extraction and loading
 - ETL for dimension tables
 - ETL for fact tables
- The implementation is shown as:
 - Mapping definition describing the source to end table for all dimension and fact tables
 - SQL statements used for the ETL operations
 - Before and after table screen shots

3. ETL Implementation

3.1. Determine the target data

Our proposed dimensional model contains 5 dimension tables and 2 fact tables. Please find the definitions for each of them at Data Warehouse area below.

Dimensional Tables

Dimension: Product		
DW Target Table	DW Target Column	Target Datatype
601-Group11-DW-Area.PRODUCT_DIM	PRODUCT_ID	int
	UPC_CODE	varchar
	PRODUCT_CATEGORY	varchar
	ITEM_CODE	varchar

Dimension: Demographic		
DW Target Table	DW Target Column	Target Datatype
601-Group11-DW-Area.DEMOGRAPHIC_DIM	DEMO_ID	int
	INCOME_LEVEL	varchar
	POOR_%	float
	BELOW_9_%	float
	ABOVE_60_%	float

Dimension: Coupon		
DW Target Table	DW Target Column	Target Datatype
601-Group11-DW-Area.COUPON_DIM	COUPON_ID	int
	COUPON_REDEEMED	int

Dimension: Store		
DW Target Table	DW Target Column	Target Datatype
601-Group11-DW-Area.STORE_DIM	STORE_ID	int
	STORE_NUMBER	int
	PRICE_TIER	varchar
	ZONE	int

Dimension: Time		
DW Target Table	DW Target Column	Target Datatype
601-Group11-DW-Area.TIME_DIM	TIME_ID	int
	WEEK	int
	MONTH	int
	YEAR	int
	EVENT	varchar

Fact Tables

Fact: Store Visits		
DW Target Table	DW Target Column	Target Datatype
601-Group11-DW-Area.STORE_VISITS_FACT	DEMO_ID	int
	COUPON_ID	int
	STORE_ID	int
	TIME_ID	int
	CUSTOMER_COUNT	int

Fact: Sales		
DW Target Table	DW Target Column	Target Datatype
601-Group11-DW-Area. SALES_FACT	STORE_ID	int
	TIME_ID	int
	PRODUCT_ID	int
	MOVEMENT	int
	QUANTITY	int
	SALES_AMOUNT	float

3.2. *Determine Source data*

Source data to the above designed schema are from CCOUNT.csv, DEMO.csv, MOVEMENT and UPC file that are available with the Dominick's FF data.

3.3. Mapping tables for staging and data mart loads.

Please find the mapping tables for staging and data mart loads below.

a) Mapping of data marts and their sources.

Dimension: Product						
DW Target Table	DW Target Column	Target Data type	Source System/ Table	Source Column	Transformation/Business Rule	Error Handling Rules
601-Group 11-DW-Area. PRODUCT_DIM	PRODUCT_ID	int		surrogate key		
	UPC_CODE	varchar	UPCXXX.csv	upc		
	PRODUCT_CATEGORY	varchar	UPCXXX.csv		XXX in source file name corresponds to the category code. e.g. If source csv file is named 'upcana', PRODUCT_CATEGORY will be 'ana'	
	ITEM_CODE	varchar	UPCXXX.csv	nitem		

Dimension: Demographic						
DW Target Table	DW Target Column	Target Data type	Source System/ Table	Source Column	Transformation/Business Rule	Error Handling Rules
601-Group 11-DW-Area. DEMOGRAPHIC_DIM	DEMO_ID	int		surrogate key		
	INCOME_LEVEL	varchar	DEMO.csv	income	Income in source column is log of median income (M.I.), which ranges from 9.87 to 11.24. $e^{9.87} = \$19,341$ and $e^{11.24} = \$76,114$. We define 3 levels for INCOME_LEVEL: 1. <i>Low</i> : <10.3 (M.I. less than \$30,000) 2. <i>Medium</i> : 10.3 to 11.0 (M.I. between \$60,000) 3. <i>High</i> : >11.0 (M.I. greater than \$60,000)	If income field is null in source table, update INCOME_LEVEL with NA
	POOR_%	float	DEMO.csv	poverty		
	BELOW_9_%	float	DEMO.csv	age9		
	ABOVE_60_%	float	DEMO.csv	age60		

Dimension: Coupon						
DW Target Table	DW Target Column	Target Data type	Source System/ Table	Source Column	Transformation/Business Rule	Error Handling Rules
601-Group 11-DW-Area. COUPON_DIM	COUPON_ID	int		surrogate key		
	COUPON_REDEEMED	int	ccount		Sum up coupons redeemed across all product categories. i.e. COUPON_REDEEMED = BAKCOUP+BULKCOUP+COSMCOUP+FISHCOUP+...	for any missing coupons redeemed value across a product category, assume the value to be 0

Dimension: Store						
DW Target Table	DW Target Column	Target Data type	Source System/ Table	Source Column	Transformation/Business Rule	Error Handling Rules
601-Group 11-DW-Area.S STORE_DIM	STORE_ID	int		Surrogate key		
	STORE_NUMBER	int	Dominic k's Stores	Store		
	PRICE_TIER	varchar	Dominic k's Stores	Price Tier		
	ZONE	int	Dominic k's Stores	Zone		If a row doesn't have a value for the price tier, set the target as '0'
	CITY	varchar	Dominic k's Stores	City		
	ZIPCODE	int	Dominic k's Stores	Zip Code		

Dimension: Time						
DW Target Table	DW Target Column	Target Data type	Source System/ Table	Source Column	Transformation/Business Rule	Error Handling Rules
601-Group 11-DW-Area. TIME_DIM	TIME_ID	int	Surrogate key			
	WEEK	int	Week's Decode Table	Week #	Populated from 1 to 400	
	MONTH	int	Week's Decode Table	Start	The start column is of format MM/DD/YY. Split it to get MM	
	YEAR	int	Week's Decode Table	Start	The start column is of format MM/DD/YY. Split it to get YY	
	EVENT	varchar				

Fact: Store Visits						
DW Target Table	DW Target Column	Target Data type	Source System/ Table	Source Column	Transformation/Business Rule	Error Handling Rules
601-Group 11-DW-Area. STORE_VISITS_FACT	DEMO_ID	int	DEMOGRAPHIC_DIM	DEMO_ID	Foreign key corresponding to primary key DEMO_ID of DEMOGRAPHIC_DIM dimension	
	COUPON_ID	int	COUPON_DIM	DEMO_ID	Foreign key corresponding to primary key COUPON_ID of COUPON_DIM dimension	
	STORE_ID	int	STORE_DIM	STORE_ID	Foreign key corresponding to primary key STORE_ID of STORE_DIM dimension	
	TIME_ID	int	TIME_DIM	TIME_ID	Foreign key corresponding to primary key TIME_ID of TIME_DIM dimension	
	CUSTOMER_COUNT	int	ccount	CUSTOMER_COUNT		

Fact: Sales						
DW Target Table	DW Target Column	Target Data type	Source System/ Table	Source Column	Transformation/Business Rule	Error Handling Rules
601-Group 11-DW-Area.SALES_FACT	STORE_ID	int	STORE_DIM	STORE_ID	Foreign key corresponding to primary key STORE_ID of STORE_DIM dimension	
	TIME_ID	int	TIME_DIM	TIME_ID	Foreign key corresponding to primary key TIME_ID of TIME_DIM dimension	
	PRODUCT_ID	int	PRODUCT_DIM	PRODUCT_ID	Foreign key corresponding to primary key PRODUCT_ID of PRODUCT_DIM dimension	
	MOVEMENT	int	UPCXXX.csv	move		
	QUANTITY	int	UPCXXX.csv	qty		
	SALES_AMOUNT	float	UPCXXX.csv		SALES_AMOUNT = Price * Move / Qty for a given row	

b) Mapping of data from .csv data sources to staging area

Staging Table: dbo.STORE					
File used to load	File Column	Data type	Staging Table	Staging Column	Staging Column Data type
Domini ck's Stores	Store	varchar	601-Group11-Staging-Area.STORE	STORE_NUMBER	int
	Price Tier	varchar		PRICE_TIER	int
	Zone	varchar		ZONE	varchar
	City	varchar		CITY	varchar
	Zip Code	varchar		ZIP_CODE	varchar

Staging Table: dbo.TIME					
File used to load	File Column	Data type	Staging Table	Staging Column	Staging Column Data type
Weeks Decode Table	Week	varchar	601-Group11-Staging-Area.TIME	WEEK	int
	-derived column-	varchar		MONTH	int
	-derived column-	varchar		YEAR	int
	Event	varchar		EVENT	varchar

Staging Table: dbo.PRODUCT					
File used to load	File Column	Data type	Staging Table	Staging Column	Staging Column Data type
UPCXXX.csv	UPC Code	varchar	601-Group11-Staging-Area.PRODUCT	UPC_CODE	int
	Item Code	varchar		ITEM_CODE	varchar
	Product Category	varchar		PRODUCT_CATEGORY	varchar

Staging Table: dbo.COUPON					
File used to load	File Column	Data type	Staging Table	Staging Column	Staging Column Data type
CCOUNT.csv	Coupons Redeemed	varchar	601-Group11-Staging-Area.COUPON	COUPON_REDEEMED	int

Staging Table: dbo.DEMOGRAPHIC					
File used to load	File Column	Data type	Staging Table	Staging Column	Staging Column Data type
Demo.csv	Age9	varchar	601-Group11-Staging-Area.DEMOGRAPHIC	BELOW_9_%	float
	Age60	varchar		ABOVE_60_%	float
	income	varchar		INCOME_LEVEL	varchar
	poverty	varchar		POOR_%	float

Staging Table: dbo.Movement					
File used to load	File Column	Data type	Staging Table	Staging Column	Staging Column Data type
Movements Table	No of Units Sold	varchar	601-Group11-Staging-Area.MOVEMENT	MOVEMENT	int
	Price	varchar		UNIT_PRICE	float
	Quantity	varchar		QUANTITY	int
	-derived column-	varchar		SALES_IN_DOLLARS	float
	Week	varchar		WEEK	int
	Store Number	varchar		STORE_NUMBER	int
	UPC	varchar		UPC_CODE	int

Staging Table: dbo.CCOUNT					
File used to load	File Column	Data type	Staging Table	Staging Column	Staging Column Data type
CCount table	Date	varchar	601-Group11-Staging-Area.CCOUNT	DATE	varchar
	Week	varchar		WEEK	int
	Cust Count	varchar		CUST_COUNT	int
	Store	varchar		STORE_NUMBER	int

c) Mapping of data from staging area to data marts.

DW Table: dbo.STORE_DIM					
Staging Table	Staging Column	Staging Column Data type	Production Table	Production Column	Production Column Data type
601-Group11-Staging-Area.STORE_DIM	STORE_NUMBER	int	601-Group11-DW-Area.STORE_DIM	STORE_NUMBER	int
	PRICE_TIER	int		PRICE_TIER	int
	ZONE	varchar		ZONE	varchar
	CITY	varchar		CITY	varchar
	ZIP_CODE	varchar		ZIP_CODE	varchar

DW Table: dbo.PRODUCT_DIM					
Staging Table	Staging Column	Staging Column Data type	Production Table	Production Column	Production Column Data type
601-Group11-Staging-Area.PRODUCT_DIM	UPC_CODE	int	601-Group11-DW-Area.PRODUCT_DIM	UPC_CODE	int
	ITEM_CODE	varchar		ITEM_CODE	varchar
	PRODUCT_CATEGORY	varchar		PRODUCT_CATEGORY	varchar

DW Table: dbo.TIME_DIM					
Staging Table	Staging Column	Staging Column Data type	Production Table	Production Column	Production Column Data type
601-Group11-Staging-Area.TIME_DIM	WEEK	int	601-Group11-DW-Area.TIME_DIM	WEEK	int
	MONTH	int		MONTH	int
	YEAR	int		YEAR	int
	EVENT	varchar		EVENT	varchar

DW Table: dbo.COUPON_DIM					
Staging Table	Staging Column	Staging Column Data type	Production Table	Production Column	Production Column Data type
601-Group11-Staging-Area.COUPON_DIM	COUPON_REDEEMED	int	601-Group11-DW-Area.COUPON_DIM	COUPON_REDEEMED	int

DW Table: dbo.DEMOGRAPHIC_DIM					
Staging Table	Staging Column	Staging Column Data type	Production Table	Production Column	Production Column Data type
601-Group11-Staging-Area.DEMOGRAPHIC_DIM	BELOW_9_%	float	601-Group11-DW-Area.DEMOGRAPHIC_DIM	BELOW_9_%	float
	ABOVE_60_%	float		ABOVE_60_%	float
	INCOME_LEVEL	vvarchar		INCOME_LEVEL	vvarchar
	POOR_%	float		POOR_%	float

DW Table: dbo.STORE_VISITS_FACT					
Staging Table	Staging Column	Staging Column Data type	Production Table	Production Column	Production Column Data type
601-Group11-Staging-Area.CCOUNT	CUST_COUNT	int	601-Group11-DW-Area.STORE_VISITS_FACT	CUSTOMER_COUNT	int

DW Table: dbo.SALES_FACT					
Staging Table	Staging Column	Staging Column Data type	Production Table	Production Column	Production Column Data type
601-Group11-Staging-Area.MOVEMENT	MOVEMENT	int	601-Group11-DW-Area.SALES_FACT	MOVEMENT	int
	QUANTITY	int		QUANTITY	int
	SALES_IN_DOLLARS	float		SALES_AMOUNT	float

3.4. Comprehensive Data Extraction Rules

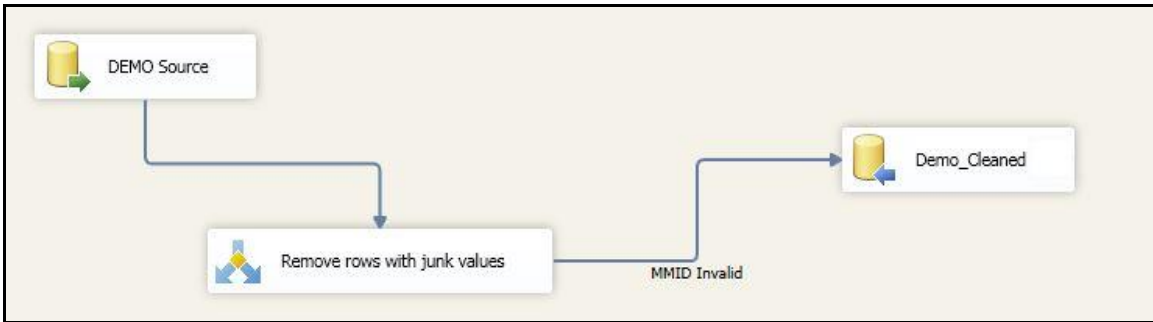
Data has been extracted from the provided CSV files and data manual of DFF. The extraction rules used in the report are as follows.

- The ccount.csv has been used as a source to extract data for COUPON dimension. The coupon redeemed columns across various product categories will be used to COUPON_DIM. So, in order to handle missing value in coupon redeemed columns, we replaced all the missing coupon redeemed values with 0.
- We truncated data with invalid store numbers from the CCOUNT table.
- Additionally, week number will be used to analyze chronological trends, so we removed the junk rows with missing week numbers during extraction process.

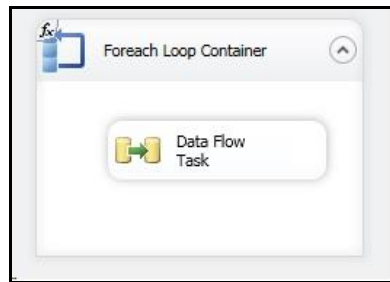
○ After extraction with above rules from ccount.csv, CCOUNT table looks as follows.

	"STORE"	"DATE"	"GROCERY"	"DAIRY"	"FROZEN"	"BOTTLE"	"MVPCLUB"	"GROCCOUP"	"MEAT"	"MEATFROZ"	"MEATCOUP"	"FISH"	"FISHCOUP"	"PROMO"	"PROMCOUP"
1	2	"911120"	13589.67	3219.69	2009.7	-1.6	0	-331.91	4003.91	144.15	0	236.96	0	94.61	-70.61
2	2	"911121"	16080.63	3918.15	2620.38	5.6	0	-513.63	3427.66	223.76	0	561.35	0	71.68	-62.36
3	2	"911122"	19230.15	4349.91	2966.6	-6.1	0	-865.61	5158.77	281.99	0	773.93	-0.5	246.76	-143.12
4	2	"911123"	28219.27	6672.33	3990.66	-8.8	0	-462	8528.5	312.7	0	708.19	0	215.9	-139.98
5	2	"911124"	23909.59	5988.28	3243.23	-1.6	0	-559.52	5107.03	206.44	-7.95	266.29	0	69.39	-47.24
6	2	"911125"	22263.35	5278.44	2999.54	-3.2	0	-557.07	6393.42	235.26	-23.5	360.6	0	76.64	-62.88
7	2	"911126"	26198.63	6789.09	3600.39	1.08	0	-683.4	7188.39	186.57	-24.15	621.07	0	223.49	-182.44
8	2	"911127"	26790.42	6857.89	3597.71	-12.5	0	-545	6027.14	225.57	0	594.37	0	85.38	-53.52
9	2	"911128"	9225.17	2536.86	1225.6	-5.6	0	-165.08	938.99	28.33	0	49.4	0	22.21	-17.28
10	2	"911129"	11466.18	3184.84	1608.58	-1.6	0	-276.31	2052.6	130.84	0	397.4	0	160.36	-87.64
11	2	"911130"	16771.28	4700.58	2452.09	-5.6	0	-394.17	3957.97	162.36	0	532.59	0	93.72	-59.1
12	2	"911201"	16696.58	4727.7	2740.97	-0.8	0	-383.85	3398.71	242.09	0	275.72	0	56.59	-40.4
13	2	"911202"	13046.17	3797.87	2034.78	-0.9	0	-341.43	2827.07	133.83	0	310.9	0	51.45	-36.96
14	2	"911203"	12696.78	3615.3	1970.01	-1.6	0	-339.51	2776.26	180.48	0	215.73	0	46.59	-37.68
15	2	"911204"	12698.58	3626.22	2023.77	0	0	-394.74	2902.57	248.05	0	203.68	0	204.83	-58.8
16	2	"911205"	16929.33	3893.71	2358.99	-7.2	0	-320.13	4799.79	230.92	-5.07	739.47	0	69.93	-2
17	2	"911206"	17686.21	4239.03	2344.4	-6.4	0	-315.55	5199.99	514.19	-104.78	834.66	0	72.23	-57.4
18	2	"911207"	24604.47	5603.38	3621.84	-5	0	-351.54	6874.29	596.36	-74.36	871.58	0	54.2	-8
19	2	"911208"	20661.49	4928.91	3106.43	-4.8	0	-317.38	4226.09	469.7	-109.85	554.47	0	27.2	-29.28
20	2	"911209"	15615.58	3743.13	2054.86	-8.8	0	-281.19	3941.14	413.64	-113.23	298.78	0	12.54	-10.56
21	2	"911210"	14216.87	3412.08	1901.03	-9.8	0	-350.27	2916.21	389.87	-113.23	441.43	0	3.98	0
22	2	"911211"	14566.16	3535.95	2008.92	-4.8	0	-320.19	3843.32	315.78	-52.39	316.7	0	32.78	-11.56
23	2	"911212"	14662.9	3937.46	2140.04	-1.6	0	-423.11	2804.78	278.13	-9.37	764.65	-0.5	50.16	-50.24
24	2	"911213"	17459.12	5036.03	2307.78	-3.2	0	-355.65	4031.88	297.03	0	1090...	0	32.51	-31.56
25	2	"911214"	23016.11	6342.51	3085.35	-11.1	0	-340.21	5770.42	244.23	-8.45	707.02	0	143.9	-100.2
26	2	"911215"	22025.65	5722.32	2792.16	-2.4	0	-284	4819.36	168.27	0	680.67	0	6.27	-5.28
27	2	"911216"	15886.74	4384.49	2087.83	-0.8	0	-296.75	3308.57	197.59	0	592.05	0	50.19	-22.12
28	2	"911217"	14056.23	3934.63	1760.59	-0.8	0	-267.32	4196.14	200.55	0	603.16	0	46.06	-47.12
29	2	"911218"	15175.27	4345.76	1815.3	0	0	-263.5	2951.68	179.46	0	1019...	0	46.43	-26.4
30	2	"911219"	18618.14	5242.56	2286.44	-3.2	0	-125.32	4605.39	186.38	0	1773...	0	125.08	-68.64
31	2	"911220"	18115.91	5195.08	2367.9	-6.4	0	-67.38	5328.8	162.8	0	1621	0	45.93	-1

- The datatype of store IDs, Week and product category coupon redeemed values have been changed from varchar to int during extraction.
- The DEMO table is being extracted from demo.csv. The rows with non-numeric Store IDs have been cleaned during extraction process.
- While extracting data for demo table, junk data with missing MMIDs, which uniquely identify each demo row, have been cleaned.

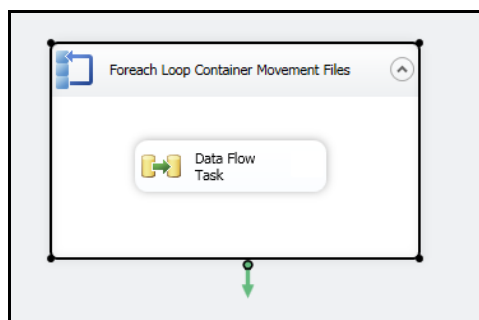


- The columns below_9_%, above_60_% and poor_% have been converted from varchar to numeric.
- The data in UPC table is being extracted using Foreach Container for all UPCXXX.csv files and then combined into final table.



- In the cleaned UPC table we are mapping each row with product category. E.g. for data extracted from the file UPCANA.csv, product category is "ANA". We are using following extraction expression to extract product category from file name.

$$SUBSTRING([User::fileName], LEN([User::fileName]) - FINDSTRING(REVERSE([User::fileName]), "\\", 1) + 2, LEN(RIGHT([User::fileName], FINDSTRING(REVERSE([User::fileName]), "\\", 1) - 1)) - FINDSTRING(REVERSE([User::fileName]), ".", 1))$$
- The data in Movement table is being extracted using Foreach Container for all movement csv files and then combined into final Movement table.



3.5. Screenshots of Data Staging area

The extracted data from the data sources have been stored in the 601-Group11-Staging-Area database. The data from staging area will further be transformed for data marts. The screenshots of different tables in staging area are as follows:

Table: 601-Group11-Staging-Area.CCOUNT

	"STORE"	"DATE"	"GROCERY"	"DAIRY"	"FROZEN"	"BOTTLE"	"MVPLUB"	"GROCCOUP"	"MEAT"	"MEATFROZ"	"MEATCOUP"	"FISH"	"FISHCOUP"	"PROMO"	"PROMCOUP"
1	306	"970211"	31749.19	6904.73	3993.62	0	369.22	-192.4	4921.7	979.5	0	1406.74	0	0	0
2	306	"970212"	33587.58	6967.44	4538.88	0	534.24	-177.39	4203.09	1065.95	0	1701.12	0	0	0
3	306	"970213"	51591.17	9962.01	7048.39	0	528.22	-1696.26	7110.86	1149.31	0	2767.16	0	0	0
4	306	"970214"	45771	9227.58	6678.89	0	542.55	-1494.38	7002.24	1318.15	0	2374.67	0	0	0
5	306	"970215"	68103.74	12989.25	8966.76	0	790.4	-2014.24	11917.79	1940.48	0	2767.42	0	0	0
6	306	"970216"	75410.64	15418.59	10917.55	0	1732.37	-2696.15	12885.57	2935.84	0	4098.71	0	0	0
7	306	"970217"	65010.98	13687.95	9259.68	0	1365.23	-2341	11090.31	2670.01	0	2660.23	0	0	0
8	306	"970218"	59641.5	11930.69	8415.12	0	1338.67	-2198.09	10116.05	2659.87	0	3738.48	0	0	0
9	306	"970306"	43826.32	8444.2	7842.39	0	764.3	-512.4	7161.58	1166.17	0	2296.88	0	0	0
10	306	"970307"	47775.46	9331.62	8489.4	0	846.32	-526	8535.64	1269.64	0	2570.76	0	0	0
11	306	"970308"	63403.35	12921.26	11149.12	0	1084.19	-796.79	11785.44	1822.12	0	2847.06	0	0	0
12	306	"970309"	55686.88	11884.88	8966	0	944.41	-638	9265.04	1940.37	0	2058.52	0	0	0
13	306	"970310"	33379.59	7016.26	5747.78	0	576.77	-354	5294.02	864.86	0	1553.23	0	0	0
14	306	"970311"	33838.78	6694.84	5760.53	0	810.79	-390.2	5274.93	727.56	0	1539.03	0	0	0
15	306	"970312"	29651.13	6416.58	5333.05	0	482.84	-309.9	4881.25	761.06	0	998.06	0	0	0
16	306	"970313"	52941.41	10911.11	8845.37	0	706.62	-891	11073.05	1371.05	-746.92	1845.39	0	0	0
17	306	"970314"	52261.67	11661.04	8614.81	0	872.18	-816.1	11247.1	1256.03	-661.61	2222.44	0	0	0
18	306	"970315"	68154.87	14825.15	10504.6	0	1226.57	-996.3	13827.03	1578.79	-641.4	2317.33	0	0	0
19	306	"970316"	80640.15	16594.36	12205.51	0	1382.4	-3179.3	16798.58	2622.38	-678.37	3472.35	0	0	0
20	306	"970317"	52198.82	11474.69	8147.36	0	950.22	-2202.25	10940.91	1900.67	-555.44	2580.92	0	0	0
21	306	"970318"	45321.49	9447.68	6544.88	0	908.24	-1706.3	8185.25	1340.9	-404.78	2425.97	0	0	0

Table: 601-Group11-Staging-Area.DEMO

	"MMID"	"NAME"	"CITY"	"ZIP"	"LAT"	"LONG"	"WEEKVOL"	"STORE"	"SCLUSTER"	"ZONE"	"AGE9"	"AGE60"
1	16892	"DOMINICKS 2"	"RIVER FOREST"	60305	419081	878131	350	2	"C"	1	0.117508576	0.232864734
2	16893	"DOMINICKS 4"	"PARK RIDGE"	60068	420392	878425	300	4	"A"	2	0.0950895057	0.26202989
3	16894	"DOMINICKS 5"	"PALATINE"	60067	421203	880431	550	5	"D"	2	0.1414334827	0.1173680317
4	16895	"DOMINICKS 8"	"OAK LAWN"	60453	417331	877436	600	8	"C"	5	0.123155416	0.2523940345
5	16896	"DOMINICKS 9"	"MORTON GROVE"	60053	420411	877994	450	9	"A"	2	0.1035030974	0.2691190176
6	16898	"DOMINICKS 12"	"CHICAGO"	60660	419928	876592	450	12	"B"	7	0.1056967397	0.178341405
7	16899	"DOMINICKS 14"	"GLENVIEW"	60025	420733	877994	400	14	"A"	1	0.129589372	0.2139492754
8	16901	"DOMINICKS 18"	"RIVER GROVE"	60171	419364	878331	600	18	"A"	5	0.1100949839	0.2723133684
9	16903	"DOMINICKS 21"	"HANOVER PARK"	60103	420058	881411	500	21	"D"	6	0.1759263459	0.0668964579
10	16905	"DOMINICKS 28"	"MOUNT PROSPECT"	60056	420686	879208	275	28	"A"	2	0.1288795371	0.2133087849
11	16906	"DOMINICKS 32"	"PARK RIDGE"	60068	419872	878378	575	32	"C"	1	0.0990606319	0.2549530316
12	16907	"DOMINICKS 33"	"CHICAGO"	60657	419386	876447	300	33	"B"	7	0.0460709172	0.1341699655
13	16909	"DOMINICKS 40"	"BRIDGEVIEW"	60455	417317	877969	500	40	"D"	6	0.1336846485	0.1818518005
14	16912	"DOMINICKS 44"	"WESTERN SPRINGS"	60558	418033	878903	325	44	"A"	2	0.1448834853	0.1909827761
15	16913	"DOMINICKS 45"	"WHEELING"	60090	421403	879300	300	45	"D"	2	0.1467187625	0.1288573479

Table: 601-Group11-Staging-Area.UPC

	UPC	Store_Number	Week	Category_Name	Unit_price	Quantity	Move	Profit_per_dollar	Sales_in_dollars	Profit_in_dollars
1	38137007220	126	319	Toothbrushes	2.390000	1	2	41.840000	4.780000	1.999952
2	38137007220	126	320	Toothbrushes	0.000000	1	0	0.000000	0.000000	0.000000
3	38137007220	126	321	Toothbrushes	2.390000	1	2	41.840000	4.780000	1.999952
4	38137007220	126	322	Toothbrushes	0.000000	1	0	0.000000	0.000000	0.000000
5	38137007220	126	323	Toothbrushes	2.390000	1	1	41.840000	2.390000	0.999976
6	38137007220	126	324	Toothbrushes	2.390000	1	1	41.840000	2.390000	0.999976
7	38137007220	126	325	Toothbrushes	2.390000	1	1	41.840000	2.390000	0.999976
8	38137007220	126	326	Toothbrushes	2.390000	1	3	41.840000	7.170000	2.999928
9	38137007220	126	327	Toothbrushes	2.390000	1	1	41.840000	2.390000	0.999976
10	38137007220	126	328	Toothbrushes	0.000000	1	0	0.000000	0.000000	0.000000
11	38137007220	126	329	Toothbrushes	0.000000	1	0	0.000000	0.000000	0.000000
12	38137007220	126	330	Toothbrushes	0.000000	1	0	0.000000	0.000000	0.000000
13	38137007220	126	331	Toothbrushes	0.000000	1	0	0.000000	0.000000	0.000000
14	38137007220	126	332	Toothbrushes	0.000000	1	0	0.000000	0.000000	0.000000
15	38137007220	126	333	Toothbrushes	0.000000	1	0	0.000000	0.000000	0.000000
16	38137007220	126	334	Toothbrushes	1.490000	1	2	16.770000	2.980000	0.499746
17	38137007220	126	335	Toothbrushes	1.490000	1	4	16.770000	5.960000	0.999492

Table: 601-Group11-Staging-Area.STORES

	Store	City	Price Tier	Zo...	Zip Code	Address
1	2	River Forest	High	1	60305	7501 W. North Ave.
2	4	Park Ridge	Medium	2	60068	Closed
3	5	Palatine	Medium	2	60067	223 Northwest HWY.
4	8	Oak Lawn	Low	5	60435	8700 S. Cicero Ave.
5	9	Morton Grove	Medium	2	60053	6931 Dempster
6	12	Chicago	High	7	60660	6009 N. Broadway Ave.
7	14	Glenview	High	1	60025	1020 Waukegan Rd.
8	18	River Grove	Low	5	60171	8355 W. Belmont Ave.
9	19	Glen Ellyn		0	60137	Closed
10	21	Hanover Park	CubFighter	6	60103	1440 Irving Park Rd.
11	25	Chicago		0	60639	Closed
12	28	Mt. Prospect	Medium	2	60054	1145-55 Mt Prospect Pz.

Table: 601-Group11-Staging-Area.TIME

	Week #	Start	End	Special Events
1	1	1989-09-14 00:00:00.000	1989-09-20 00:00:00.000	
2	2	1989-09-21 00:00:00.000	1989-09-27 00:00:00.000	
3	3	1989-09-28 00:00:00.000	1989-10-04 00:00:00.000	
4	4	1989-10-05 00:00:00.000	1989-10-11 00:00:00.000	
5	5	1989-10-12 00:00:00.000	1989-10-18 00:00:00.000	
6	6	1989-10-19 00:00:00.000	1989-10-25 00:00:00.000	
7	7	1989-10-26 00:00:00.000	1989-11-01 00:00:00.000	Halloween
8	8	1989-11-02 00:00:00.000	1989-11-08 00:00:00.000	
9	9	1989-11-09 00:00:00.000	1989-11-15 00:00:00.000	
10	10	1989-11-16 00:00:00.000	1989-11-22 00:00:00.000	
11	11	1989-11-23 00:00:00.000	1989-11-29 00:00:00.000	Thanksgiving
12	12	1989-11-30 00:00:00.000	1989-12-06 00:00:00.000	
13	13	1989-12-07 00:00:00.000	1989-12-13 00:00:00.000	
14	14	1989-12-14 00:00:00.000	1989-12-20 00:00:00.000	
15	15	1989-12-21 00:00:00.000	1989-12-27 00:00:00.000	Christmas
16	16	1989-12-28 00:00:00.000	1990-01-03 00:00:00.000	New-Year
17	17	1990-01-04 00:00:00.000	1990-01-10 00:00:00.000	
18	18	1990-01-11 00:00:00.000	1990-01-17 00:00:00.000	

3.6. Data Transformation and Cleansing Rules

a) Transformation of INCOME in CCOUNT to INCOME_LEVEL

Income column on CCOUNT tables stores log of median income (M.I.), which ranges from 9.87 to 11.24. $e^{9.87} = \$19,341$ and $e^{11.24} = \$76,114$. We define 3 levels for INCOME_LEVEL:

Low: <10.3 (M.I. less than \$30,000)

Medium: 10.3 to 11.0 (M.I. between \$60,000)

High: >11.0 (M.I. greater than \$60,000)

We have used following expression in SSIS to generate the income levels from income values in the DEMO table.

INCOME_LEVEL = (INCOME > 11) ? "High" : (INCOME < 10.3 ? "Low" : "Medium")

b) Calculation of Sales Amount value from Price, Movement and Quality columns

The sales amount for each row in Movement table in staging area has been calculated using following transformation.

*Sales_in_dollars = (Unit_Price*Move/Quantity)*

c) Deriving Product_Category from UPCXXX.csv filepath

We derived Product_Category from path of each UPCXXX.csv file. For example, if full path a upc file is C:/folder/UPCANA.csv, we extracted the filename i.e. UPCANA.csv and then we further extracted product category as "ANA". The expression used in derived column SSIS component is as follows:

*SUBSTRING(@[User::fileName],LEN(@[User::fileName]) -
FINDSTRING(REVERSE(@[User::fileName]),"\",1) +
2,LEN(RIGHT(@[User::fileName],FINDSTRING(REVERSE(@[User::fileName]),"\",1) - 1)) -
FINDSTRING(REVERSE(@[User::fileName]),".",1))*

d) Transformation of Time data into Week, Month and Year details

We used Week's Decode Table for this transformation.

	Week #	Start	End	Special Events
1	1	1989-09-14 00:00:00.000	1989-09-20 00:00:00.000	
2	2	1989-09-21 00:00:00.000	1989-09-27 00:00:00.000	
3	3	1989-09-28 00:00:00.000	1989-10-04 00:00:00.000	
4	4	1989-10-05 00:00:00.000	1989-10-11 00:00:00.000	
5	5	1989-10-12 00:00:00.000	1989-10-18 00:00:00.000	
6	6	1989-10-19 00:00:00.000	1989-10-25 00:00:00.000	
7	7	1989-10-26 00:00:00.000	1989-11-01 00:00:00.000	Halloween
8	8	1989-11-02 00:00:00.000	1989-11-08 00:00:00.000	
9	9	1989-11-09 00:00:00.000	1989-11-15 00:00:00.000	
10	10	1989-11-16 00:00:00.000	1989-11-22 00:00:00.000	
11	11	1989-11-23 00:00:00.000	1989-11-29 00:00:00.000	Thanksgiving
12	12	1989-11-30 00:00:00.000	1989-12-06 00:00:00.000	
13	13	1989-12-07 00:00:00.000	1989-12-13 00:00:00.000	
14	14	1989-12-14 00:00:00.000	1989-12-20 00:00:00.000	
15	15	1989-12-21 00:00:00.000	1989-12-27 00:00:00.000	Christmas

For this transformation, we converted start date for each week from varchar to datetime datatype and then derived Month and Year value from the data using following SSIS expressions.

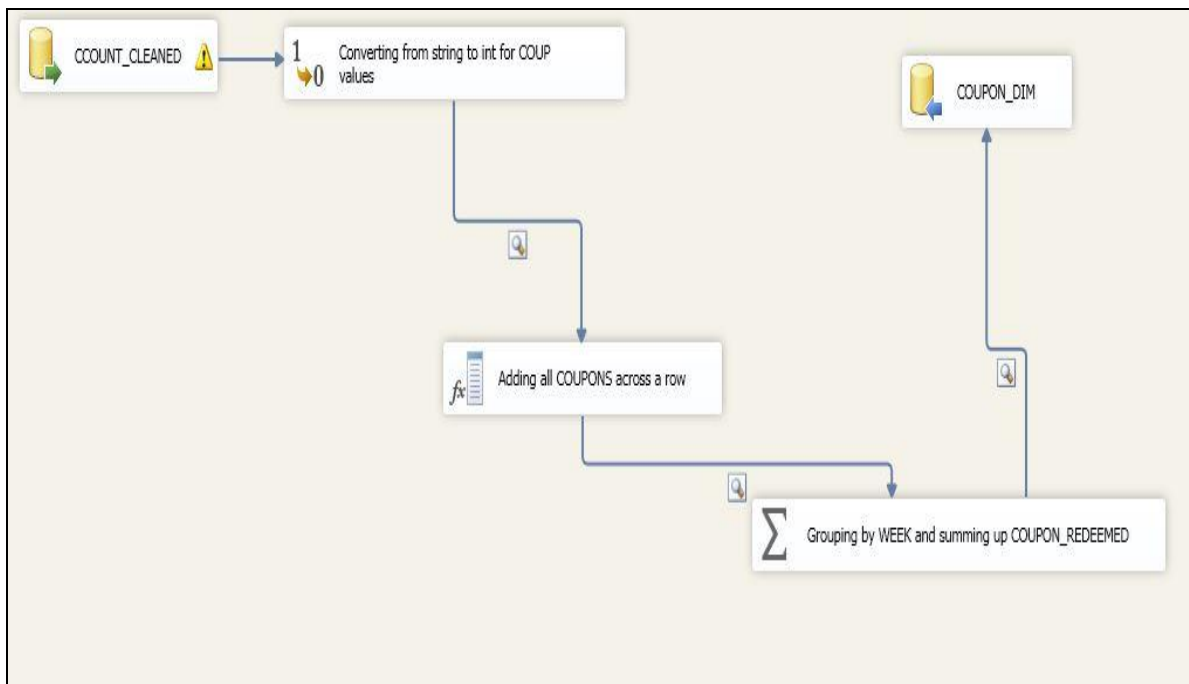
YEAR = YEAR([Datatype convert for Week].Start)
MONTH = MONTH([Datatype convert for Week].Start)

3.7. Plan for Aggregation

Aggregation is done in Coupon dimension table, Sales fact table and Store Visits fact table for answering the business questions and to fix the granularity. Please find the explanation for each aggregation below.

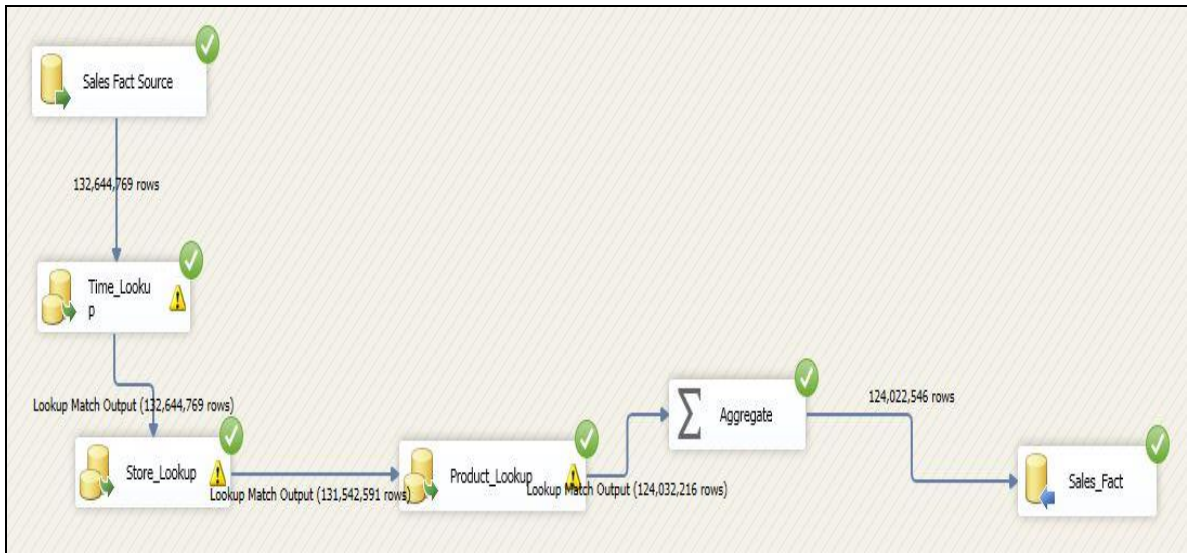
a) Aggregation in Coupon dimension table.

Coupon redeemed data is loaded from CCount table and the data is aggregated on weekly basis to keep the time granularity of the Coupon redeemed sales at Weekly level. As shown in figure below Coupon redeemed attribute in Coupon dimension table is calculated by summing up the Coupon redeemed value on a weekly basis. Hence the surrogate keys generated for Coupon dimension in in one-to-one mapping with week number.



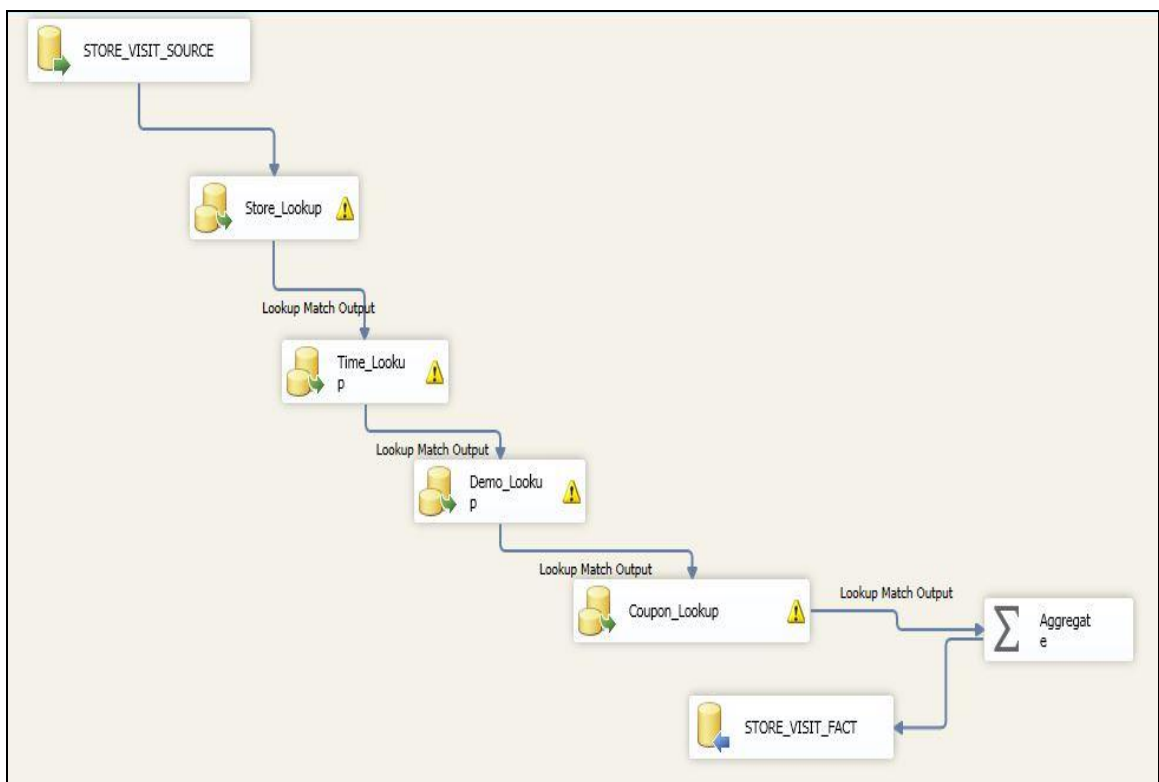
b) Aggregation in SALES_FACT table.

Source data for Sales fact table is Movement table .lookup transformation is performed with cleaned Movement table from data staging area on Time, Store and Product dimension tables on the matching of attributes Week Number, Store Number and UPC Code with respective tables. Aggregation of the MOVEMENT, QUANTITY and SALES_AMOUNT in the resultant table is aggregated on the basis of Store_ID, Time_ID and Product_ID. Store_ID, Time_ID and Product_ID are mapped to Store Number, Week Number and UPC_Code respectively. So the aggregation ensures the granularity in the above levels for each dimensions. We are concerned mainly about the Time granularity for drilling down and rolling up in later phases of the project and the time granularity will be on Week level. This can be rolled up to Month and Year.



c) Aggregation in STORE_VISITS_FACT table.

Source data for the STORE_VISITS_FACT table is the CCount data. Lookup transformation was performed with cleaned CCount table from the staging area on STORE_DIM, TIME_DIM, DEMOGRAPHIC_DIM and COUPON_DIM tables in the same order as shown in the below figure. Attributes matched for lookup in STORE_DIM, TIME_DIM, DEMOGRAPHIC_DIM and COUPON_DIM tables are STORE_ID, TIME_ID, DEMO_ID and COUPON_ID respectively. Time granularity for the CUSTOMER_COUNT in the STORE_VISITS_FACT is WEEK. This can be rolled up to Month and Year level.



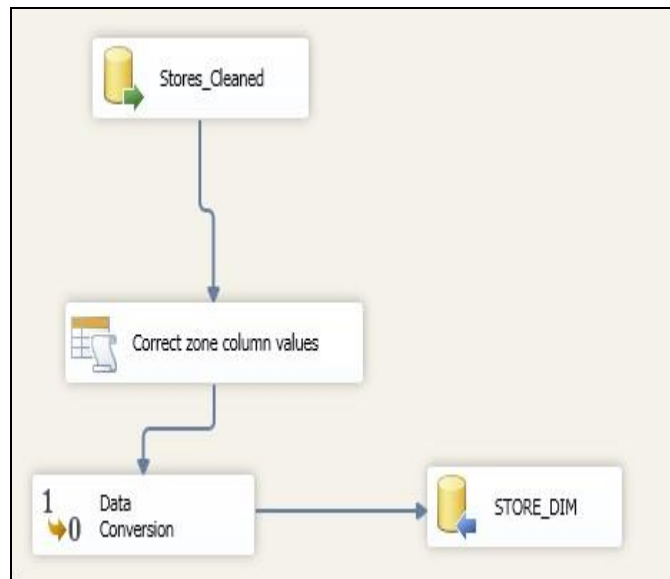
4. Data Loading

Main source of data for this project is main from 4 .csv files- CCount, Demographics, UPC and Movement. This data was directly loaded into data staging area of the warehouse directly and data cleaning operations were performed as explained in this report. This cleaned data was further transformed using different ETL processes as explained below for the 5 final dimension(STORE_DIM, PRODUCT_DIM, TIME_DIM, COUPON_DIM, DEMOGRAPHICS_DIM) tables and 2(SALES_FACT, STORE_VISITS_FACT) fact tables.

a) ETL for DIMENSION tables

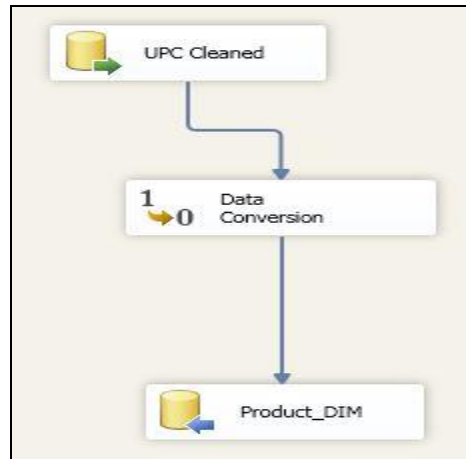
STORE_DIM

Source data for the STORE_DIM is Dominick's Stores data table. This data was cleaned and ETL was performed on it as shown in the below diagram. Few Column's had no Zone value in Dominick Stores table and then data conversion was performed on it from String to Integer. This transformed data was stored in STORE_DIM table.



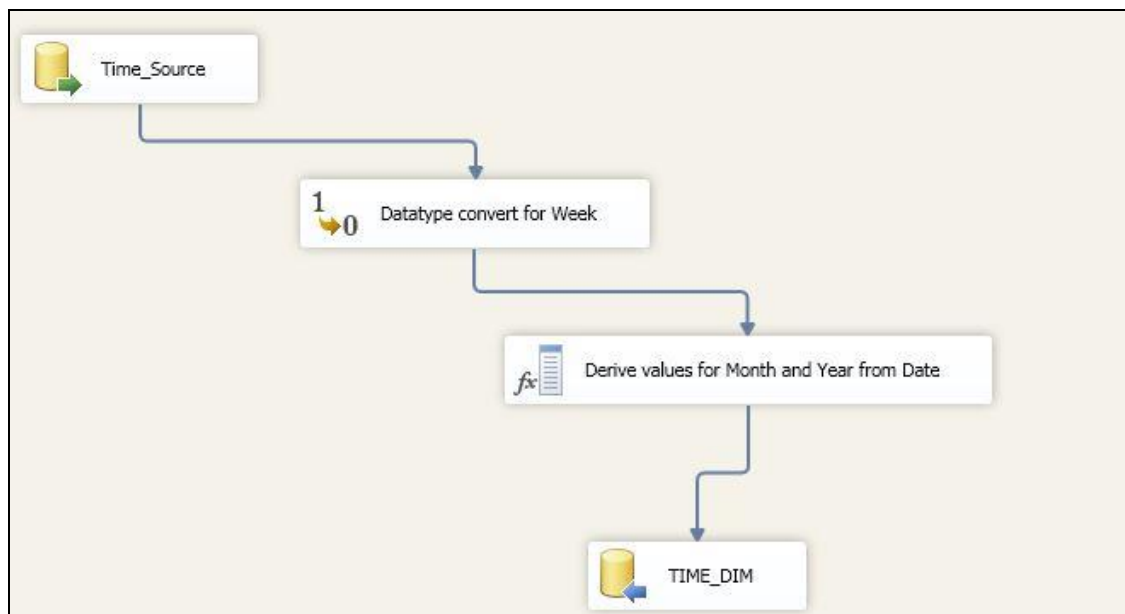
PRODUCT_DIM

Product Dimension tables took source data from UPC tables in Dominick's data. Each product has its own UPC tables and all tables were merged together. This table was cleaned as explained previously. On the cleaned UPC data, data conversion was performed for ITEM_Code to convert from Varchar to Integer as shown below.



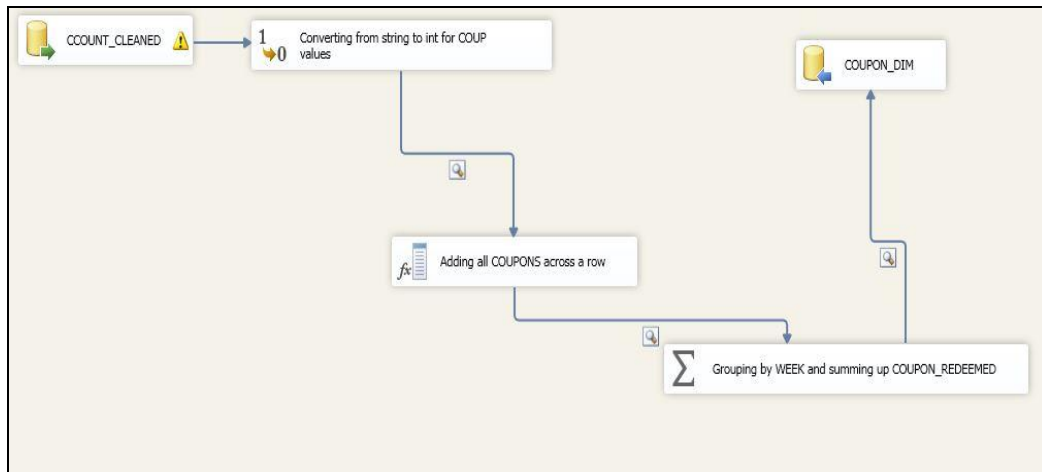
TIME_DIM

Weeks Decode table from Dominick's manual acts as the source table for the TIME_DIM table. Initially the data type for Week is Varchar and it is converted to Integer. Thereafter the Month and Year attributes in TIME_DIM was derived from the date as per the Transformation described previously. This final transformed table is stored in TIME_DIM.



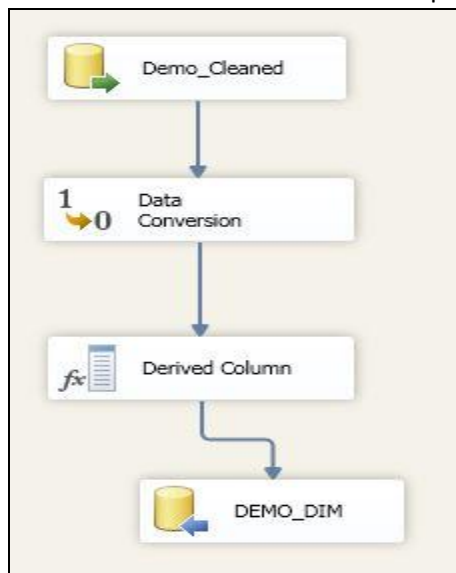
COUPON_DIM

Coupon redeemed values for the COUPON_DIM was sourced from CCount table. ETL is performed on the cleaned CCount table in data staging area. First of all data conversion from varchar to Integer was performed on Coupon redeemed values. Thereafter is performed on the Coupon redeemed values on weekly basis to get the weekly coupon redeemed values every week. This transformed table is kept COUPON_DIM table.



DEMOGRAPHICS_DIM

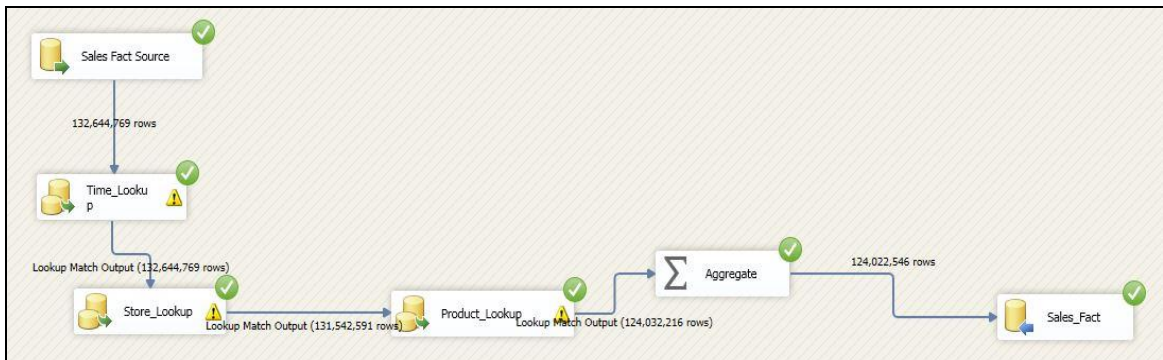
Demographics table from Dominick's is the source table for DEMOGRAPHICS_DIM. The cleaned data is kept in data staging area and ETL transformation performed on the same. Firstly the values BELOW_9_% and ABOVE_60_% undergoes data type conversion from varchar to float. There after derived column INCOME_LEVEL is set as per the range of log of median income. Business rule for the same is explained in the mapping table. This will be the second stage of ETL transformation and the resultant table after this step is stored in DEMO_DIM.



b) ETL for FACT tables

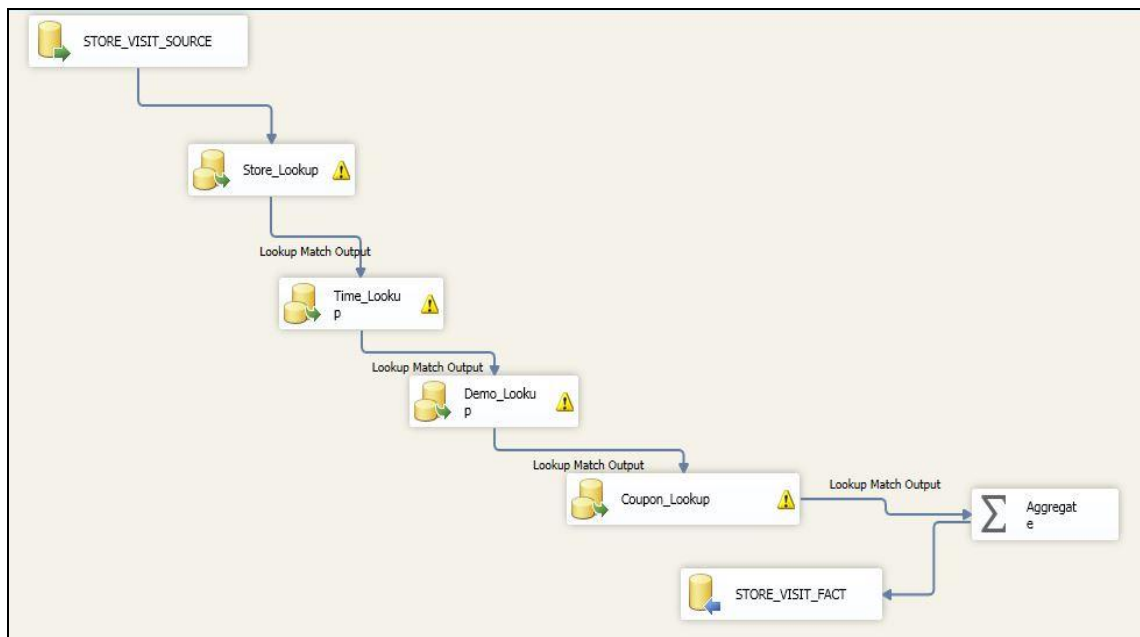
SALES FACT

Movement table is the source data for Sales fact table. SALES_AMOUNT in Movement table is the value derived from Price, Quantity and Movement values in Movement table. Cleaning was performed on this table and kept in Staging area. Lookup transformation was performed on this table this table as source on TIME_DIM, STORE_DIM, PRODUCT_DIM and aggregation was performed as explained previously before loading the data in SALES_FACT table.



STORE VISITS FACT

Source data for STORE_VISITS_FACT table is CCount table. Cleaned CCount data in staging area undergoes Lookup transformation with STORE_DIM, TIME_DIM, DEMO_DIM and COUPON_DIM. Finally aggregation is performed as explained previously and final data is loaded in STORE_VISITS_FACT table.



4.1. Mapping definition describing the source to end table for all dimension and fact tables

End to end mapping of elements from source to destination data marts

Dimension: Product								
Source File	Column Name	Datatype	Staging Table	Staging Column	Staging Column Datatype	Production Table	Production Column	Production Column Datatype
UPC.csv	UPC_CODE	varchar	UPC-Cleaned	UPC	varchar	PRODUCT_DIM	UPC_CODE	int
	PRODUCT_CATEGORY	varchar		PROD_CATEGORY	varchar		PRODUCT_CATEGORY	varchar
	ITEM_CODE	varchar		NITEM	varchar		ITEM_CODE	int

Dimension: Store								
Source File	Column Name	Datatype	Staging Table	Staging Column	Staging Column Datatype	Production Table	Production Column	Production Column Datatype
Demo.csv	STORE	varchar	Stores-Cleaned	STORE	int	STORE_DIM	STORE_NUMBER	int
	PRICE TIER	varchar		PRICE TIER	varchar		PRICE_TIER	varchar
	ZONE	varchar		ZONE	varchar		ZONE	int
	CITY	varchar		CITY	varchar		CITY	varchar
	ZIPCODE	varchar		ZIPCODE	varchar		ZIPCODE	varchar

Dimension: Time								
Source File	Column Name	Datatype	Staging Table	Staging Column	Staging Column Datatype	Production Table	Production Column	Production Column Datatype
Dominick's Stores	Week #	varchar	Time-Cleaned	Week #	int	TIME_DIM	WEEK	int
	Start	varchar		Start	Datetime		MONTH	Calculated Field
	End	varchar		End	Datetime		YEAR	Calculated Field
	Special Events	varchar		Special Events	varchar		SPECIAL_EVENTS	varchar

Dimension: Demographic								
Source File	Column Name	Datatype	Staging Table	Staging Column	Staging Column Datatype	Production Table	Production Column	Production Column Datatype
Demo.csv	Age9	varchar	DEMO-Cleaned	BELOW_9_%	varchar	DEMOGRAPHIC_DIM	BELOW_9_%	decimal
	age60	varchar		ABOVE_60_%	varchar		ABOVE_60_%	decimal
	poverty	varchar		POOR_%	varchar		POOR_%	decimal
	income	varchar		INCOME_LEVEL	decimal		INCOME_LEVEL	Calculated Value
	Store	varchar		STORE_NUM	int		STORE_NUM	int

Dimension: Coupon								
Source File	Column Name	Datatype	Staging Table	Staging Column	Staging Column Datatype	Production Table	Production Column	Production Column Datatype
CCOUNT.csv	Coupon Redeemed	Coupon value across the different products- Calculated field	CCOUNT-Cleaned	COUPON_REDEEMED	decimal	COUPON_DIM	COUPON_REDEEMED	varchar

Fact Table: Sales								
Source File	Column Name	Datatype	Staging Table	Staging Column	Staging Column Datatype	Production Table	Production Column	Production Column Datatype
Keys are auto incremented values			PRODUCT_DIM	PRODUCT_ID	int	SALES_FACT	PRODUCT_ID	int
			STORE_DIM	STORE_ID	int		STORE_ID	int
			TIME_DIM	TIME_ID	int		TIME_ID	int
UPC.csv	move	vvarchar	UPC-Cleaned	movement	int		MOVEMENT	int
UPC.csv	qty	vvarchar	UPC-Cleaned	quantity	int		QUANTITY	decimal
UPC.csv	Sales	Calculated value	UPC-Cleaned	Sales amount	decimal		SALES_AMOUNT	decimal

Fact Table: Store visits								
Source File	Column Name	Datatype	Staging Table	Staging Column	Staging Column Datatype	Production Table	Production Column	Production Column Datatype
Keys are auto incremented values			DEMOGRAPHIC_DIM	DEMO_ID	int	STORE_VISITS_FACT	DEMO_ID	int
			COUPON_DIM	COUPON_ID	int		COUPON_ID	int
			STORE_DIM	STORE_ID	int		STORE_ID	int
			TIME_DIM	TIME_ID	int		TIME_ID	int
CCOUNT.csv	CUSTOMER_COUNT	vvarchar	CCOUNT-Cleaned	Customer Count	int		CUSTOMER_COUNT	int

4.2. SQL statements used for the ETL operations

Below are the SQL scripts used to create the destination tables:

```
CREATE TABLE [dbo].[COUPON_DIM](
    [COUPON_ID] [int] NOT NULL,
    [COUPON_REDEEMED] [real] NULL,
    PRIMARY KEY [COUPON_ID]
)
CREATE TABLE [dbo].[DEMOGRAPHIC_DIM](
    [DEMO_ID] [int] IDENTITY(1,1) NOT NULL,
    [BELOW_9_%] [numeric](10, 2) NULL,
    [ABOVE_60_%] [numeric](10, 2) NULL,
    [POOR_%] [numeric](10, 2) NULL,
    [INCOME_LEVEL] [varchar](50) NULL,
    [STORE_NUM] [int] NOT NULL,
    PRIMARY KEY [DEMO_ID]
)
CREATE TABLE [dbo].[PRODUCT_DIM](
    [PRODUCT_ID] [int] IDENTITY(1,1) NOT NULL,
    [UPC_CODE] [varchar](50) NULL,
    [PRODUCT_CATEGORY] [varchar](50) NULL,
    [ITEM_CODE] [varchar](50) NULL,
    PRIMARY KEY [PRODUCT_ID]
)
CREATE TABLE [dbo].[STORE_DIM](
    [STORE_ID] [int] IDENTITY(1,1) NOT NULL,
    [STORE_NUMBER] [int] NULL,
    [PRICE_TIER] [varchar](50) NULL,
    [ZONE] [int] NULL,
    [CITY] [varchar](50) NULL,
    [ZIPCODE] [varchar](50) NULL,
    PRIMARY KEY [STORE_ID]
)
CREATE TABLE [dbo].[TIME_DIM](
    [TIME_ID] [int] IDENTITY(1,1) NOT NULL,
    [WEEK] [int] NULL,
    [MONTH] [int] NULL,
    [YEAR] [int] NULL,
    [EVENTS] [varchar](50) NULL,
    PRIMARY KEY [TIME_ID]
)
CREATE TABLE [STORE_VISITS_FACT] (
    DEMO_ID int FOREIGN KEY references DEMOGRAPHIC_DIM(DEMO_ID),
    COUPON_ID int FOREIGN KEY references COUPON_DIM(COUPON_ID),
    STORE_ID int FOREIGN KEY references STORE_DIM(STORE_ID),
    TIME_ID int FOREIGN KEY references TIME_DIM(TIME_ID),
    COSTOMER_COUNT int
    PRIMARY KEY (DEMO_ID, COUPON_ID, STORE_ID, TIME_ID)
)
```

```

CREATE TABLE [SALES_FACT] (
    PRODUCT_ID int FOREIGN KEY references PRODUCT_DIM(PRODUCT_ID),
    STORE_ID int FOREIGN KEY references STORE_DIM(STORE_ID),
    TIME_ID int FOREIGN KEY references TIME_DIM(TIME_ID),
    MOVEMENT int,
    QUANTITY int,
    SALES_AMOUNT numeric,
    PRIMARY KEY (PRODUCT_ID,STORE_ID,TIME_ID)
)

```

4.3. Staging and Data Mart table screen shots

Staging tables:

601-Group11-Staging-Area.CCOUNT-Cleaned

	"STORE"	"DATE"	"GROCERY"	"DAIRY"	"FROZEN"	"BOTTLE"	"MVPCLUB"	"GROCCOUP"	"MEAT"	"MEATFROZ"	"MEATCOUP"	"FISH"	"FISHCOUP"	"PROMO"	"PROMCOUP"
1	306	"970211"	31749.19	6904.73	3993.62	0	369.22	-192.4	4921.7	979.5	0	1406.74	0	0	0
2	306	"970212"	33587.58	6967.44	4538.88	0	534.24	-177.39	4203.09	1065.95	0	1701.12	0	0	0
3	306	"970213"	51591.17	9962.01	7048.39	0	528.22	-1696.26	7110.86	1149.31	0	2767.16	0	0	0
4	306	"970214"	45771	9227.58	6678.89	0	542.55	-1494.38	7002.24	1318.15	0	2374.67	0	0	0
5	306	"970215"	68103.74	12989.25	8966.76	0	790.4	-2014.24	11917.79	1940.48	0	2767.42	0	0	0
6	306	"970216"	75410.64	15418.59	10917.55	0	1732.37	-2696.15	12885.57	2935.84	0	4098.71	0	0	0
7	306	"970217"	65010.98	13687.95	9259.68	0	1365.23	-2341	11090.31	2670.01	0	2660.23	0	0	0
8	306	"970218"	59641.5	11930.69	8415.12	0	1338.67	-2198.09	10116.05	2659.87	0	3738.48	0	0	0
9	306	"970306"	43826.32	8444.2	7842.39	0	764.3	-512.4	7161.58	1166.17	0	2296.88	0	0	0
10	306	"970307"	47775.46	9331.62	8489.4	0	846.32	-526	8535.64	1269.64	0	2570.76	0	0	0
11	306	"970308"	63403.35	12921.26	11149.12	0	1084.19	-796.79	11785.44	1822.12	0	2847.06	0	0	0
12	306	"970309"	55686.88	11884.88	8966	0	944.41	-638	9265.04	1940.37	0	2058.52	0	0	0
13	306	"970310"	33379.59	7016.26	5747.78	0	576.77	-354	5294.02	864.86	0	1553.23	0	0	0
14	306	"970311"	33838.78	6694.84	5760.53	0	810.79	-390.2	5274.93	727.56	0	1539.03	0	0	0
15	306	"970312"	29651.13	6416.58	5333.05	0	482.84	-309.9	4881.25	761.06	0	998.06	0	0	0
16	306	"970313"	52941.41	10911.11	8845.37	0	706.62	-891	11073.05	1371.05	-746.92	1845.39	0	0	0
17	306	"970314"	52261.67	11661.04	8614.81	0	872.18	-816.1	11247.1	1256.03	-661.61	2222.44	0	0	0

601-Group11-Staging-Area.DEMO-Cleaned

	"MMID"	"NAME"	"CITY"	"ZIP"	"LAT"	"LONG"	"WEEKVOL"	"STORE"	"SCLUSTER"	"ZONE"	"AGE9"	"AGE60"	"ETHNIC"	"EDUC"	"Y"
1	16892	"DOMINICKS 2"	"RIVER FOREST"	60305	419081	878131	350	2	"C"	1	0.117508576	0.232864734	0.1142799489	0.2489349342	0
2	16893	"DOMINICKS 4"	"PARK RIDGE"	60068	420392	878425	300	4	"A"	2	0.0950895057	0.26202989	0.0621612744	0.2207894147	0
3	16894	"DOMINICKS 5"	"PALATINE"	60067	421203	880431	550	5	"D"	2	0.1414334827	0.1173680317	0.0538752774	0.3212257298	0
4	16895	"DOMINICKS 8"	"OAK LAWN"	60453	417331	877436	600	8	"C"	5	0.123155416	0.2523940345	0.0352433281	0.0951732743	0
5	16896	"DOMINICKS 9"	"MORTON GROVE"	60053	420411	877994	450	9	"A"	2	0.1035030974	0.2691190176	0.0326188257	0.2221723183	0
6	16898	"DOMINICKS 12"	"CHICAGO"	60660	419928	876592	450	12	"B"	7	0.1056967397	0.178341405	0.3806979879	0.2534129693	0
7	16899	"DOMINICKS 14"	"GLENVIEW"	60025	420733	877994	400	14	"A"	1	0.129589372	0.2139492754	0.034178744	0.3482930237	0
8	16901	"DOMINICKS 18"	"RIVER GROVE"	60171	419364	878331	600	18	"A"	5	0.1100949839	0.2723133684	0.0744171442	0.0722464558	0
9	16903	"DOMINICKS 21"	"HANOVER PARK"	60103	420058	881411	500	21	"D"	6	0.1759263459	0.0668964579	0.1050387773	0.1775034504	0
10	16905	"DOMINICKS 28"	"MOUNT PROSPECT"	60056	420686	879208	275	28	"A"	2	0.1288795371	0.2133087849	0.0559354726	0.233162564	0
11	16906	"DOMINICKS 32"	"PARK RIDGE"	60068	419872	878378	575	32	"C"	1	0.0990606319	0.2549530316	0.0319385141	0.1982598608	0
12	16907	"DOMINICKS 33"	"CHICAGO"	60657	419386	876447	300	33	"B"	7	0.0460709172	0.1341699655	0.1301271793	0.4196880043	0
13	16909	"DOMINICKS 40"	"BRIDGEVIEW"	60455	417317	877969	500	40	"D"	6	0.1336846485	0.1818518005	0.0440530671	0.0721286047	0
14	16912	"DOMINICKS 44"	"WESTERN SPRINGS"	60558	418033	878903	325	44	"A"	2	0.1448834853	0.1909827761	0.0376320741	0.3297383876	0
15	16913	"DOMINICKS 45"	"WHEELING"	60090	421403	879300	300	45	"D"	2	0.1467187625	0.1288573479	0.0872338729	0.2801501642	0
16	16915	"DOMINICKS 47"	"ADDISON"	60101	419364	880022	350	47	"D"	2	0.1429616817	0.125798297	0.1206758914	0.1405986509	0
17	16916	"DOMINICKS 48"	"SCHAUMBURG"	60193	420503	880775	325	48	"D"	2	0.1217670803	0.0979219614	0.0994917028	0.3032603841	0
18	16917	"DOMINICKS 49"	"DOWNERS GROVE"	60515	418111	879869	275	49	"A"	2	0.1348777349	0.1874731875	0.0383526384	0.3199499687	0
19	16918	"DOMINICKS 50"	"HICKORY HILLS"	60457	417169	878347	275	50	"A"	2	0.1244204019	0.1533573759	0.0709256397	0.1287642783	0
20	16919	"DOMINICKS 51"	"PALOS HEIGHTS"	60463	416594	877775	400	51	"D"	3	0.132472108	0.1761597181	0.0254257193	0.1719167112	0
21	16920	"DOMINICKS 52"	"NORTHBROOK"	60062	421364	878825	450	52	"A"	1	0.13660619	0.1522411953	0.0848986126	0.3729272959	0
22	16921	"DOMINICKS 53"	"CHICAGO"	60662	420039	877069	300	53	"C"	7	0.1208391392	0.3002786809	0.065722248	0.2703834998	0
23	16922	"DOMINICKS 54"	"NAPERVILLE"	60540	417975	881225	375	54	"D"	2	0.1479145854	0.0902222777	0.0466408591	0.4211256441	0
24	16924	"DOMINICKS 56"	"COUNTRYSIDE"	60525	417683	878881	375	56	"A"	2	0.1310138278	0.192888549	0.0413560186	0.2375507193	0

601-Group11-Staging-Area.MOVEMENT-Cleaned

	UPC	Store_Number	Week	Category_Name	Unit_price	Quantity	Move	Profit_per_dollar	Sales_in_dollars	Profit_in_dollars
1	38137007220	126	319	Toothbrushes	2.390000	1	2	41.840000	4.780000	1.999952
2	38137007220	126	320	Toothbrushes	0.000000	1	0	0.000000	0.000000	0.000000
3	38137007220	126	321	Toothbrushes	2.390000	1	2	41.840000	4.780000	1.999952
4	38137007220	126	322	Toothbrushes	0.000000	1	0	0.000000	0.000000	0.000000
5	38137007220	126	323	Toothbrushes	2.390000	1	1	41.840000	2.390000	0.999976
6	38137007220	126	324	Toothbrushes	2.390000	1	1	41.840000	2.390000	0.999976
7	38137007220	126	325	Toothbrushes	2.390000	1	1	41.840000	2.390000	0.999976
8	38137007220	126	326	Toothbrushes	2.390000	1	3	41.840000	7.170000	2.999928
9	38137007220	126	327	Toothbrushes	2.390000	1	1	41.840000	2.390000	0.999976
10	38137007220	126	328	Toothbrushes	0.000000	1	0	0.000000	0.000000	0.000000
11	38137007220	126	329	Toothbrushes	0.000000	1	0	0.000000	0.000000	0.000000
12	38137007220	126	330	Toothbrushes	0.000000	1	0	0.000000	0.000000	0.000000
13	38137007220	126	331	Toothbrushes	0.000000	1	0	0.000000	0.000000	0.000000
14	38137007220	126	332	Toothbrushes	0.000000	1	0	0.000000	0.000000	0.000000
15	38137007220	126	333	Toothbrushes	0.000000	1	0	0.000000	0.000000	0.000000
16	38137007220	126	334	Toothbrushes	1.490000	1	2	16.770000	2.980000	0.499746
17	38137007220	126	335	Toothbrushes	1.490000	1	4	16.770000	5.960000	0.999492
18	38137007220	126	336	Toothbrushes	0.000000	1	0	0.000000	0.000000	0.000000
19	38137007220	126	337	Toothbrushes	2.390000	1	5	41.840000	11.950000	4.999880

601-Group11-Staging-Area.STORES-Cleaned

	Store	City	Price Tier	Zo...	Zip Code	Address
1	2	River Forest	High	1	60305	7501 W. North Ave.
2	4	Park Ridge	Medium	2	60068	Closed
3	5	Palatine	Medium	2	60067	223 Northwest HWY.
4	8	Oak Lawn	Low	5	60435	8700 S. Cicero Ave.
5	9	Morton Grove	Medium	2	60053	6931 Dempster
6	12	Chicago	High	7	60660	6009 N. Broadway Ave.
7	14	Glenview	High	1	60025	1020 Waukegan Rd.
8	18	River Grove	Low	5	60171	8355 W. Belmont Ave.
9	19	Glen Ellyn		0	60137	Closed
10	21	Hanover Park	CubFighter	6	60103	1440 Irving Park Rd.
11	25	Chicago		0	60639	Closed
12	28	Mt. Prospect	Medium	2	60054	1145-55 Mt Prospect Pz.
13	32	Park Ridge	High	1	60068	1900 S. Cumberland Ave.
14	33	Chicago	High	7	60657	3012 N. Broadway Ave.
15	39	Waukegan		0	60085	Closed
16	40	Bridgeview	CubFighter	6	60455	8825 S. Harlem Ave.....
17	44	Western Spring	Medium	2	60558	14 Garden Market St.
18	45	Wheeling	Medium	2	60090	550 W. Dundee Rd.
19	46	Carol Stream	Low	5	60187	Closed
20	47	Addison	Medium	2	60101	545 W. Lake St.
21	48	Schaumburg	Medium	2	60193	20 E. Golf Rd.
22	49	Downers Grove	Medium	2	60515	120 E. Ogden Ave.

601-Group11-Staging-Area.TIME-Cleaned

	Week #	Start	End	Special Events
1	1	1989-09-14 00:00:00.000	1989-09-20 00:00:00.000	
2	2	1989-09-21 00:00:00.000	1989-09-27 00:00:00.000	
3	3	1989-09-28 00:00:00.000	1989-10-04 00:00:00.000	
4	4	1989-10-05 00:00:00.000	1989-10-11 00:00:00.000	
5	5	1989-10-12 00:00:00.000	1989-10-18 00:00:00.000	
6	6	1989-10-19 00:00:00.000	1989-10-25 00:00:00.000	
7	7	1989-10-26 00:00:00.000	1989-11-01 00:00:00.000	Halloween
8	8	1989-11-02 00:00:00.000	1989-11-08 00:00:00.000	
9	9	1989-11-09 00:00:00.000	1989-11-15 00:00:00.000	
10	10	1989-11-16 00:00:00.000	1989-11-22 00:00:00.000	
11	11	1989-11-23 00:00:00.000	1989-11-29 00:00:00.000	Thanksgiving
12	12	1989-11-30 00:00:00.000	1989-12-06 00:00:00.000	
13	13	1989-12-07 00:00:00.000	1989-12-13 00:00:00.000	
14	14	1989-12-14 00:00:00.000	1989-12-20 00:00:00.000	
15	15	1989-12-21 00:00:00.000	1989-12-27 00:00:00.000	Christmas
16	16	1989-12-28 00:00:00.000	1990-01-03 00:00:00.000	New-Year
17	17	1990-01-04 00:00:00.000	1990-01-10 00:00:00.000	

601-Group11-Staging-Area.UPC-Cleaned

	COM_CODE	UPC	DESCRIP	SIZE	CASE	NITEM	PROD_CATEGORY
1	953	1192603016	CAFFEDRINE CAPLETS 1	16 CT	6	7342431	ANA
2	953	1192662108	SLEEPINAL SOFTGEL	8 CT	6	7333311	ANA
3	953	1650001020	NERVINE TABS	30 CT	1	8430820	ANA
4	953	1650001022	NERVINE SLEEP AID	12 CT	1	8430840	ANA
5	953	1650004106	ALKA-SELTZER GOLD	20 CT	1	8430880	ANA
6	953	1650004108	ALKA-SELTZER GOLD	36 CT	1	8430900	ANA
7	953	1650004703	ALKA MINTS	30 CT	1	8430700	ANA
8	953	2140649030	LEGATRIN PM	30 CT	1	8435810	ANA
9	953	2586600493	PERCOGESIC A/F ANALG	50 CT	1	8416280	ANA
10	953	2586610493	PERCOGESIC A/F ANALG	50 CT	1	8416280	ANA
11	953	2586610501	ALEVE TABLETS	24 CT	6	6122441	ANA
12	953	2586610502	ALEVE CAPLETS	24 CT	6	6122741	ANA
13	953	2586610503	ALEVE TABLETS	50 CT	6	6122451	ANA
14	953	2586610504	ALEVE CAPLETS	50 CT	6	6122751	ANA
15	953	2586610505	ALEVE TABLETS	100 CT	6	6122461	ANA
16	953	2586610506	ALEVE CAPLETS	100 CT	6	6122761	ANA
17	953	3225259620	SUNBEAM HEAT WRAP MS	1 CT	1	8402470	ANA
18	953	3680012732	TC MOTION SICKNESS T	12 CT	12	6190791	ANA
19	953	3680012740	VALUE TIME ASPIRIN	250 CT	12	6108051	ANA
20	953	3680012742	VALUE TIME ACETA	100 CT	12	6108031	ANA
21	953	3680012888	TC IBUROFEN TABLETS\$	100 CT	6	6190091	ANA
22	953	3680012890	TC X/STR PAIN RLF TA	30 CT	12	6191211	ANA

Data Mart tables:

601-Group11-DW-Area.PRODUCT_DIM

	PRODUCT_ID	UPC_CODE	PRODUCT_CATEGORY	ITEM_CODE
1	1	1192603016	ANA	7342431
2	2	1192662108	ANA	7333311
3	3	1650001020	ANA	8430820
4	4	1650001022	ANA	8430840
5	5	1650004106	ANA	8430880
6	6	1650004108	ANA	8430900
7	7	1650004703	ANA	8430700
8	8	2140649030	ANA	8435810
9	9	2586600493	ANA	8416280
10	10	2586610493	ANA	8416280
11	11	2586610501	ANA	6122441
12	12	2586610502	ANA	6122741
13	13	2586610503	ANA	6122451
14	14	2586610504	ANA	6122751
15	15	2586610505	ANA	6122461
16	16	2586610506	ANA	6122761
17	17	3225259620	ANA	8402470
18	18	3680012732	ANA	6190791

601-Group11-DW-Area.TIME_DIM

	TIME_ID	WEEK	MONTH	YEAR	EVENTS
1	1	1	9	1989	
2	2	2	9	1989	
3	3	3	9	1989	
4	4	4	10	1989	
5	5	5	10	1989	
6	6	6	10	1989	
7	7	7	10	1989	Halloween
8	8	8	11	1989	
9	9	9	11	1989	
10	10	10	11	1989	
11	11	11	11	1989	Thanksgiving
12	12	12	11	1989	
13	13	13	12	1989	
14	14	14	12	1989	
15	15	15	12	1989	Christmas
16	16	16	12	1989	New-Year
17	17	17	1	1990	
18	18	18	1	1990	
19	19	19	1	1990	
20	20	20	1	1990	
21	21	21	2	1990	

601-Group11-DW-Area.STORE_DIM

	STORE_ID	STORE_NUMBER	PRICE_TIER	ZONE	CITY	ZIPCODE
1	1	2	High	1	River Forest	60305
2	2	4	Medium	2	Park Ridge	60068
3	3	5	Medium	2	Palatine	60067
4	4	8	Low	5	Oak Lawn	60435
5	5	9	Medium	2	Morton Grove	60053
6	6	12	High	7	Chicago	60660
7	7	14	High	1	Glenview	60025
8	8	18	Low	5	River Grove	60171
9	9	19		0	Glen Ellyn	60137
10	10	21	CubFighter	6	Hanover Park	60103
11	11	25		0	Chicago	60639
12	12	28	Medium	2	Mt. Prospect	60054
13	13	32	High	1	Park Ridge	60068
14	14	33	High	7	Chicago	60657
15	15	39		0	Waukegan	60085
16	16	40	CubFighter	6	Bridgeview	60455
17	17	44	Medium	2	Western Spring	60558
18	18	45	Medium	2	Wheeling	60090
19	19	46	Low	5	Carol Stream	60187
20	20	47	Medium	2	Addison	60101
21	21	48	Medium	2	Addison	60101

601-Group11-DW-Area.DEMOGRAPHIC_DIM

	DEMO_ID	BELOW_9_%	ABOVE_60_%	POOR_%	INCOME_LEVEL	STORE_NUM
1	1	0.11	0.23	0.06	Medium	2
2	2	0.09	0.26	0.03	Medium	4
3	3	0.14	0.11	0.02	Medium	5
4	4	0.12	0.25	0.05	Medium	8
5	5	0.10	0.26	0.02	Medium	9
6	6	0.10	0.17	0.16	Low	12
7	7	0.12	0.21	0.01	High	14
8	8	0.11	0.27	0.07	Medium	18
9	9	0.17	0.06	0.02	Medium	21
10	10	0.12	0.21	0.03	Medium	28
11	11	0.09	0.25	0.03	Medium	32
12	12	0.04	0.13	0.08	Medium	33
13	13	0.13	0.18	0.05	Medium	40
14	14	0.14	0.19	0.03	Medium	44
15	15	0.14	0.12	0.02	Medium	45
16	16	0.14	0.12	0.03	Medium	47
17	17	0.12	0.09	0.01	Medium	48
18	18	0.13	0.18	0.02	Medium	49
19	19	0.12	0.15	0.04	Medium	50
20	20	0.13	0.17	0.03	Medium	51
21	21	0.13	0.15	0.01	High	52
22	22	0.12	0.30	0.06	Medium	53
23	23	0.14	0.09	0.02	Medium	54
24	24	0.13	0.19	0.02	Medium	56
25	25	0.17	0.11	0.02	Medium	58

601-Group11-DW-Area.COUPON_DIM

	COUPON_ID	COUPON_REDEEMED
1	0	450100.7
2	1	880910.3
3	2	318295.8
4	3	2773639
5	4	383151.4
6	5	435256.2
7	6	1687740
8	7	498353.6
9	8	926468.2
10	9	2193234
11	10	2039579
12	11	523196.5
13	12	2418025
14	13	319055.9
15	14	372650.8
16	15	314056.7

601-Group11-DW-Area.STORE_VISITS_FACT

	DEMO_ID	COUPON_ID	STORE_ID	TIME_ID	COSTOMER_COUNT
1	1	1	1	1	13870
2	1	2	1	2	15539
3	1	3	1	3	13987
4	1	4	1	4	14810
5	1	5	1	5	14415
6	1	6	1	6	13812
7	1	7	1	7	14574
8	1	8	1	8	14649
9	1	9	1	9	13551
10	1	10	1	10	17016
11	1	11	1	11	12731
12	1	12	1	12	13706
13	1	13	1	13	13890
14	1	14	1	14	15250
15	1	15	1	15	13977
16	1	16	1	16	13778
17	1	17	1	17	13690
18	1	18	1	18	13473
19	1	19	1	19	13464

601-Group11-DW-Area.SALES_FACT

	PRODUCT_ID	STORE_ID	TIME_ID	MOVEMENT	QUANTITY	SALES_AMOUNT
1	41	78	327	0	1	0
2	41	78	328	1	1	5
3	41	78	329	1	1	5
4	41	78	330	1	1	5
5	41	78	331	1	1	5
6	41	78	332	0	1	0
7	41	78	333	1	1	5
8	41	78	334	1	1	5
9	41	78	335	1	1	5
10	41	78	336	1	1	5
11	41	78	337	0	1	0
12	41	78	338	0	1	0
13	41	78	339	1	1	5
14	41	78	340	2	1	11
15	41	78	341	1	1	5
16	41	78	342	0	1	0
17	41	78	343	0	1	0
18	41	78	344	3	1	17
19	41	78	345	0	1	0
20	41	78	346	3	1	17
21	41	78	347	0	1	0
22	41	78	348	0	1	0

E. BI REPORTING

1. BI Reports Design and Implementation

1.1. Reporting plan

For our reporting and analysis, we have used SSRS, SSAS, ReportBuilder 3.0, SSRS on top of SSAS. Following is our reporting strategy.

Business Question 1	SSAS + SSRS
Business Question 2	SSRS
Business Question 3	Report Builder 3.0
Business Question 4	Report Builder 3.0
Business Question 5	SSAS

1.1.1. Determining all target reports that satisfy business questions

Question1: What is the trend of Beer sales during Christmas holiday season?

Report generated from SSRS on top of SSAS

We decided to create a cube using SALES_FACT fact table and added the attributes of Week, Product Category, Sales Amount and Event from TIME_DIM and PRODUCT_DIM dimension tables. We decided to use this plan because SALES_FACT has the largest size in the entire data warehouse and a cube created from it could be used to answer business Question 5 as well. Creating a cube added flexibility to our analysis and then allowed report easily using SSRS on top of the cube we created. We used line graph for data visualization since it was required to understand the trend.

Question 2: Which stores have more popularity among kids and elderly groups?

Report generated from SSRS alone

To solve this question, we used STORE_VISITS_FACT, DEMOGRAPHIC_DIM and STORE_DIM tables to generate the reports. Since we needed only two separate list of popular stores, one representing popularity among kids and another among elderly people, we decided to go for SSRS alone for this business question. We used STORE_DIM for STORE_NUMBER attributes. The ABOVE_60_% and

BELOW_9_% attributes have been extracted from DEMOGRAPHIC_DIM table. Since, we were interested in only top 10 results, we order by demographic percentages in descending order and picked top 10 store numbers.

Question 3: Which stores attract people who earn below poverty line?

Report generated using Report Builder 3.0

This business question required generating report with the list of stores that are popular among poor income group and we decided to use Report Builder 3.0. The report is based on number of customer visits in the stores, so we used STORE_VISITS_FACT. This fact table provided us stores numbers which could be linked with the attributes POOR_% available in DEMOGRAPHIC_DIM table. DEMO_ID key has been used to create inner join between the table. We were interested in most popular stores among poor income group, so order the results by POOR_% in descending order.

Question 4: What is the effect of introducing coupons on total number of customer visits?

Report generated from Report Builder 3.0

For this business question, we used independent data mart – Store Visits to answer this business question. We required a report that represented coupons redeemed weekly trend against week numbers and decided to use Report Builder 3.0. The COUPON_ID attribute has been used to join COUPON_DIM and STORE_VISITS_FACT to generate COUPON_REDEEMED attribute. The TIME_ID attribute has been used to join TIME_DIM and STORE_VISITS_FACT to generate WEEK attribute of the report. We aggregated COUPON_REDEEMED for each week to generate the trends and used bar graph for data visualization. We created another bar chart by aggregating customer visit counts per week. Comparing the trends of coupons redeemed per week and customer visits per week, we could analyze correlation between coupons redeemed and customer visits

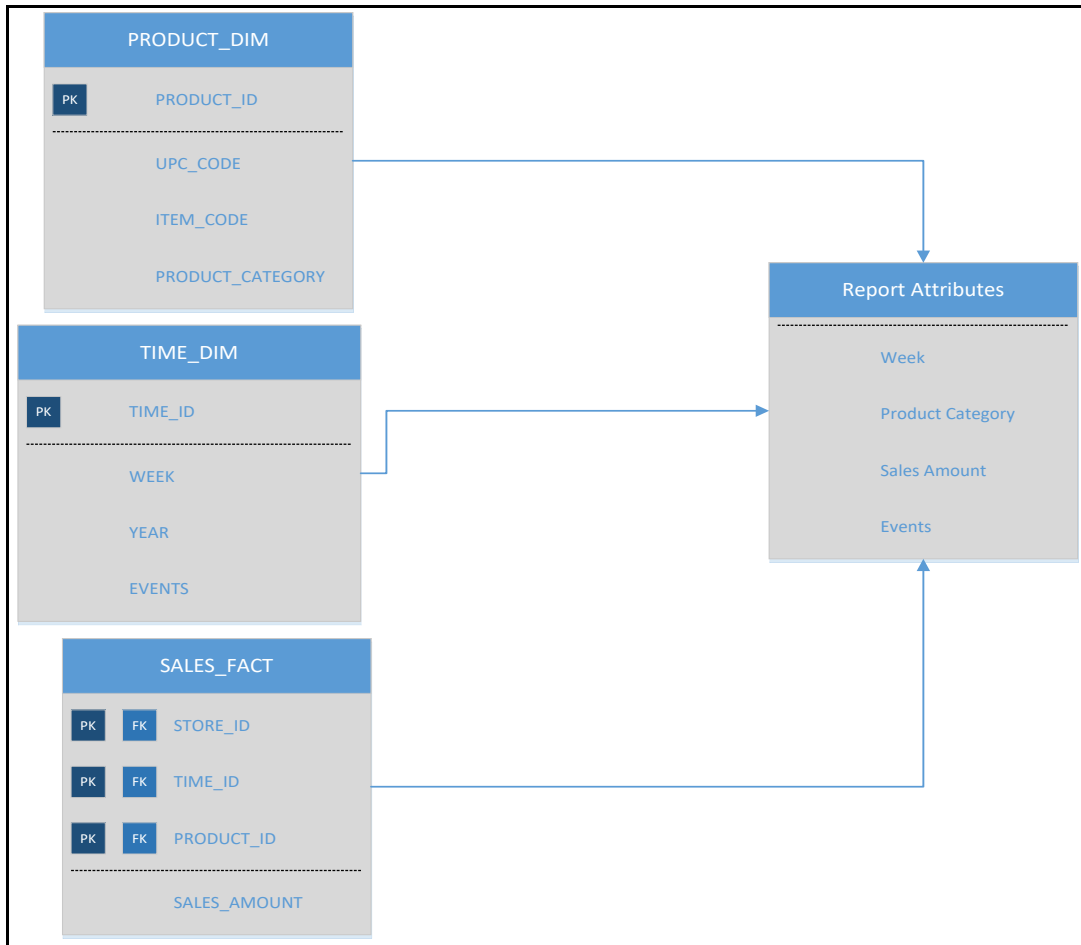
Question 5: What is the trend of a product demand in different price-tiers?

Analysis of the cube created from SSAS alone

We analyze this business question using SSAS as we needed to use SALES_FACT fact table and creating a cube using this large table added flexibility to our analysis. This cube further used STORE_DIM and TIME_DIM for dimensions. The STORE_DIM provided attribute for price tiers (High, Low, Medium) and TIME_DIM provided week number attribute. We generated data using inner joins between the 3 tables and aggregated on weekly basis. We finally analyzed the cube using pivot chart.

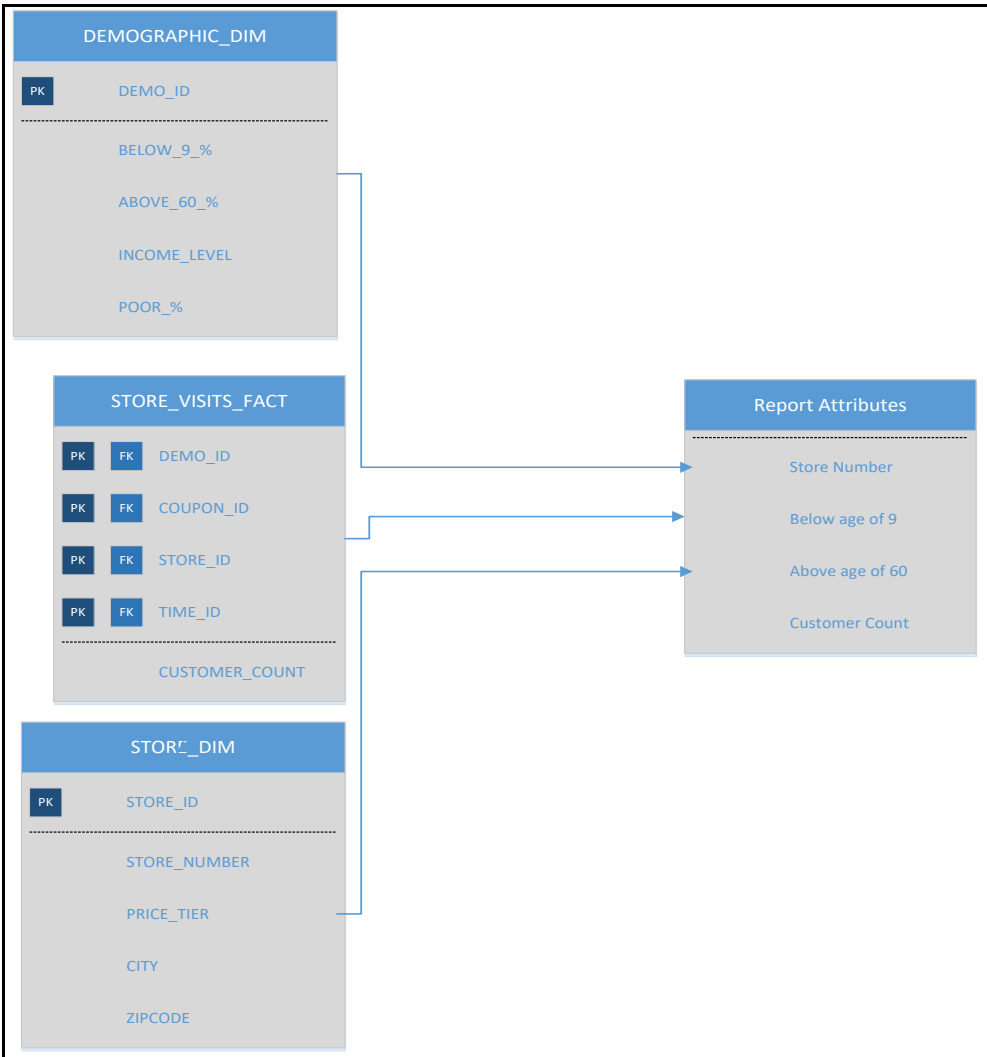
1.2. Mappings from the tables in the data marts to the attributes in the report

Question1: What is the trend of beer sales during Christmas holiday season?



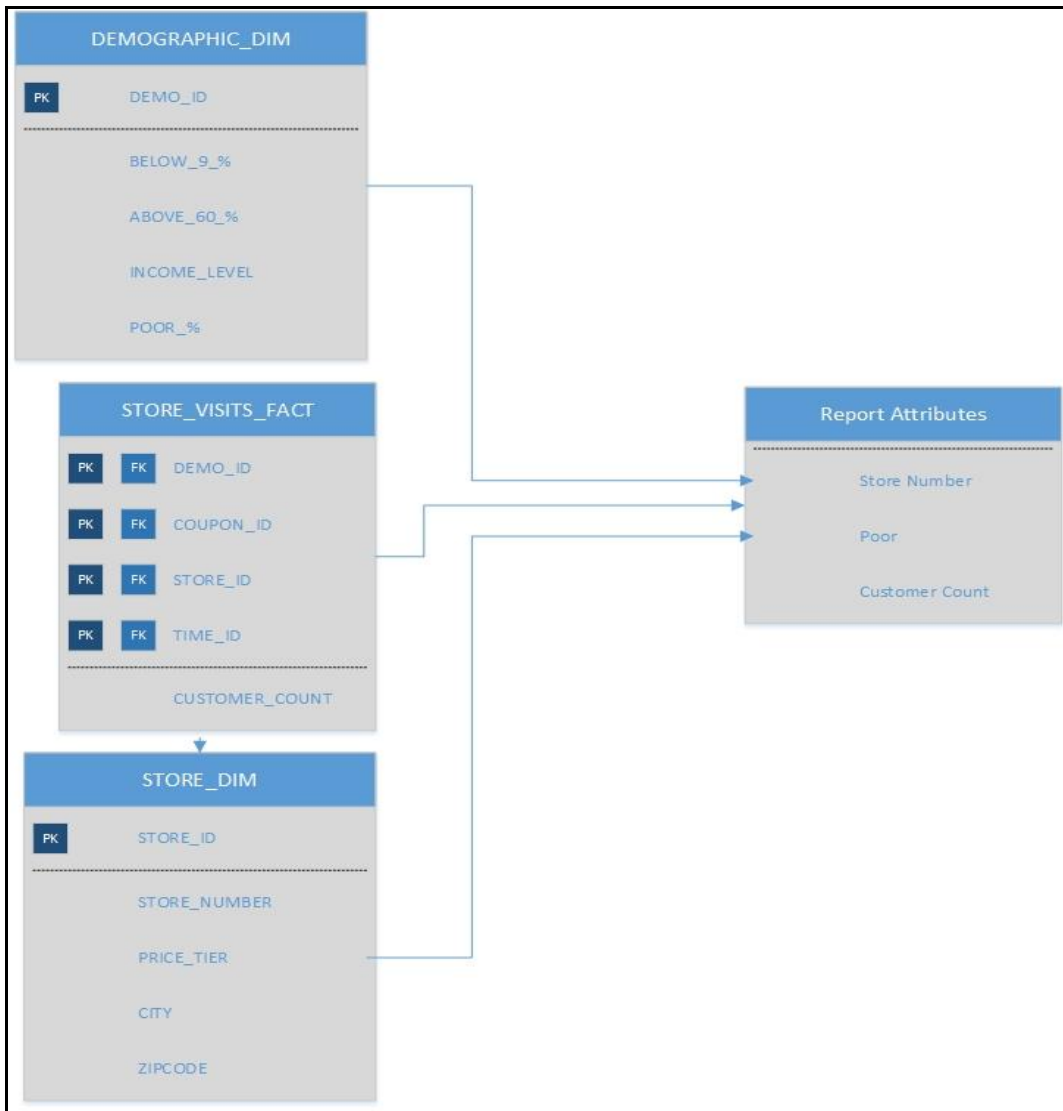
Attribute name	Dimension / Fact Table	Groupings/ Filters	Report Attribute
PRODUCT_CATEOGRY	PRODUCT_DIM		Week Number
WEEK	TIME_DIM		Product Cateogry
EVENTS	TIME_DIM	Filter on Christmas	Events
SALES_AMOUNT	SALES_FACT		Sales Amount

Question2: Which stores have more popularity among kids and elderly groups?



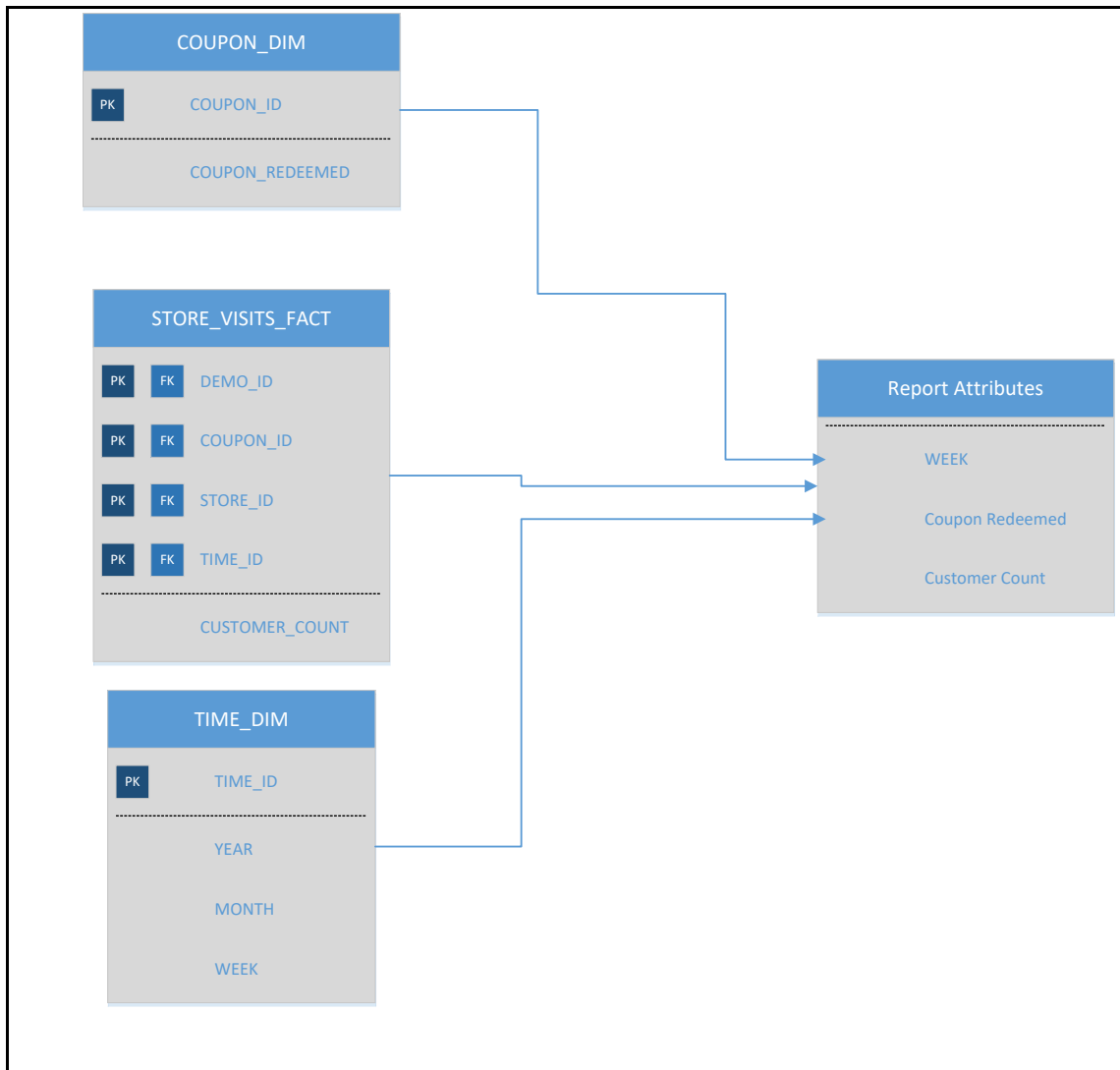
Attribute name	Dimension / Fact Table	Groupings/ Filters	Report Attribute
BELOW_9_%	DEMOGRAPHIC_DIM		Below age of 9
ABOCE_60_%	DEMOGRAPHIC_DIM		Above age of 60
STORE_NUMBER	STORE_DIM	TOP 10 Stores	Store Number
CUSTOMER_COUNT	STORE_VISITS_FACT		Customer count

Question3: Which stores attract people who earn below poverty line?



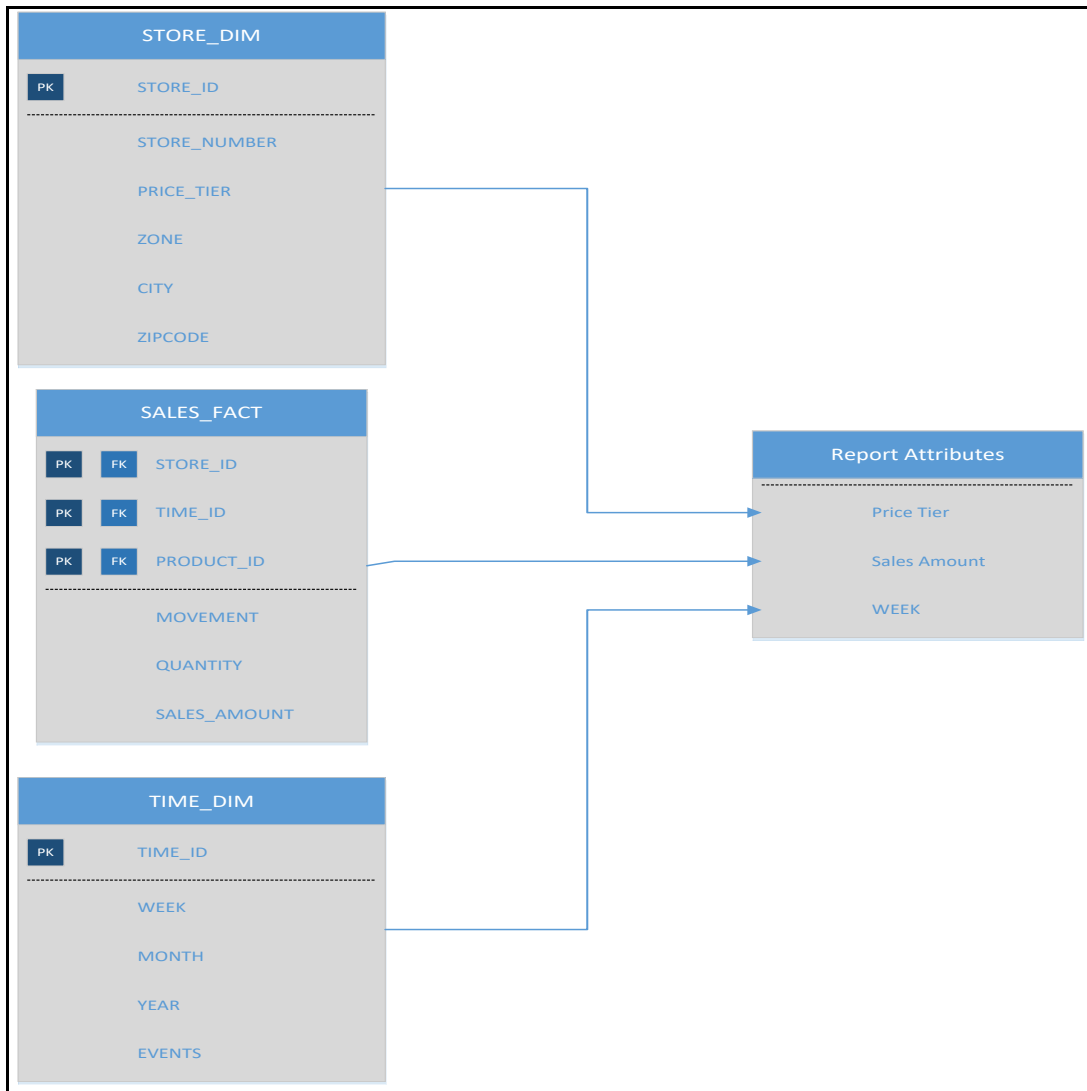
Attribute name	Dimension / Fact Table	Groupings/ Filters	Report Attribute
STORE_NUMBER	STORE_DIM	TOP 10 Stores	Store Number
POOR_%	DEMOGRAPHIC_DIM		Product Category
CUSTOMER_COUNT	STORE_VISITS_FACT		Customer count

Question4: What is the effect of introducing coupons on total number of customer visits?



Attribute name	Dimension / Fact Table	Report Attribute
WEEK	TIME_DIM	Week
COUPON_REDEEMED	COUPON_DIM	Coupons redeemed
CUSTOMER_COUNT	STORE_VISITS_FACT	Customer count

Question5: What is the trend of a product demand in different price-tiers?



Attribute name	Dimension / Fact Table	Groupings/ Filters	Report Attribute
PRICE_TIER	PRODUCT_DIM	Filters on High, Low, Medium	Price Tier
SALES_AMOUNT	SALES_FACT		Sales Amount
WEEK	TIME_DIM		WEEK

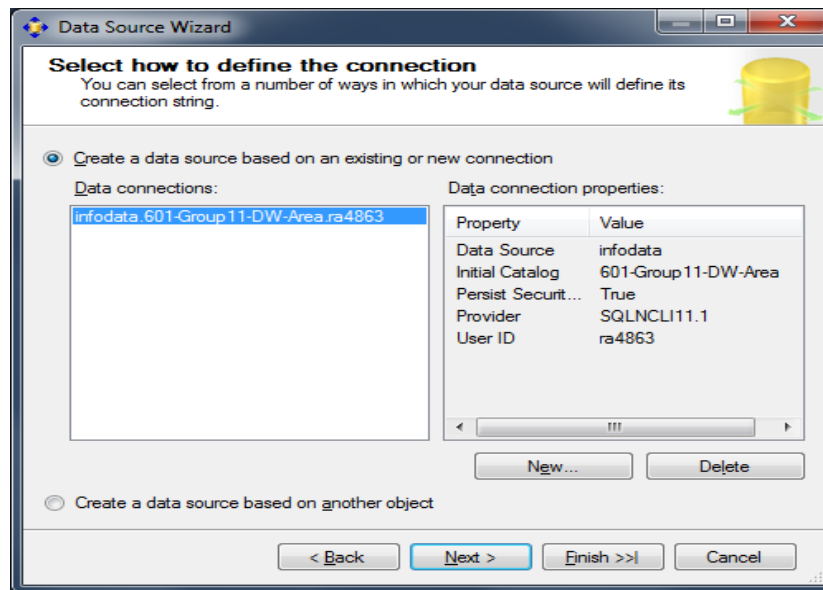
1.3. Cube from SSAS and Report from SSRS on top of SSAS for Question 1

As per the reporting plan described in previous part of report we are building the cube in SSAS and SSRS will be used on top of SSAS cube to generate report for question 1.

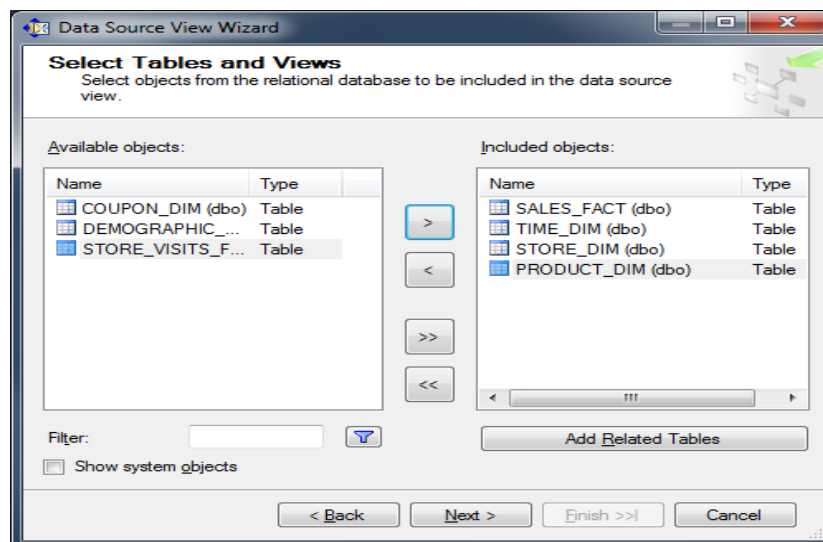
Question 1: What is the trend of Beer sales during Christmas holiday season?

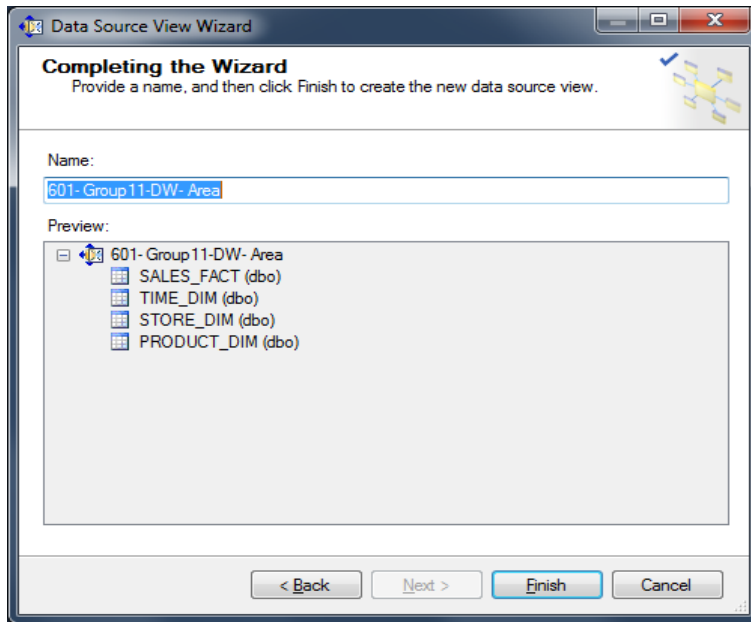
Designing Cube Using SSAS

Selecting Data Source for SSAS

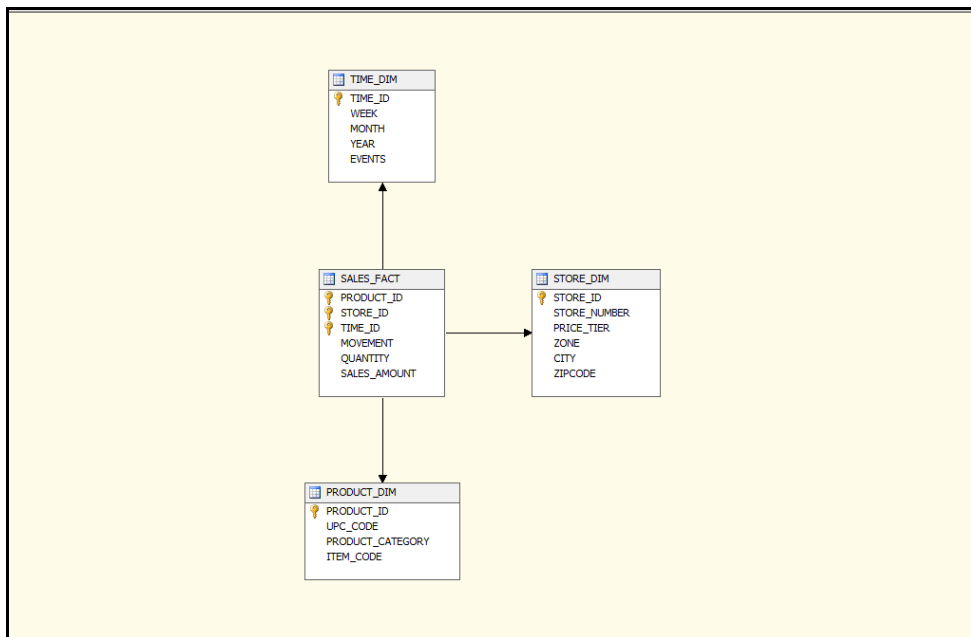


Designing Data Source Views

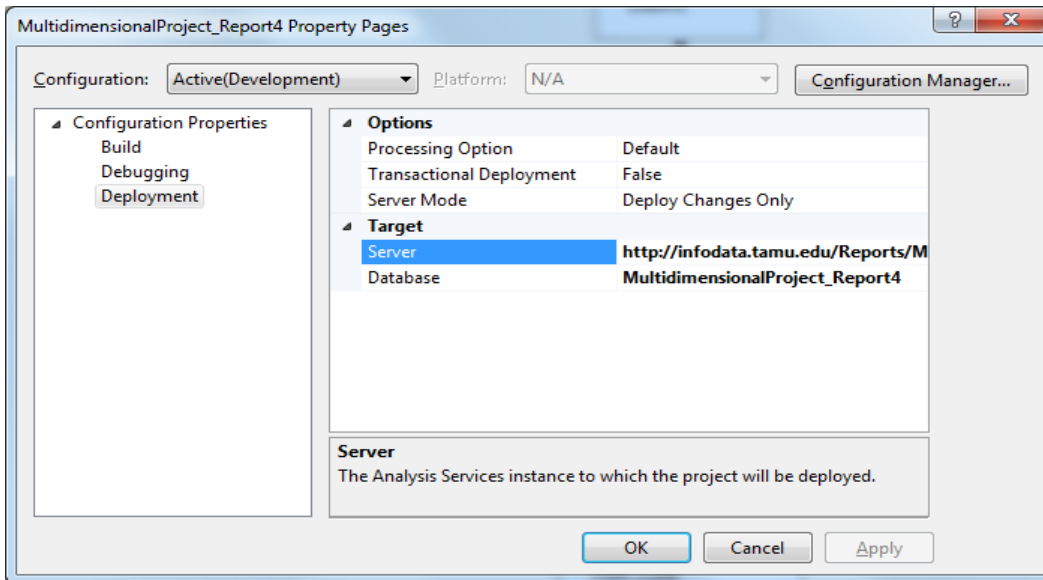




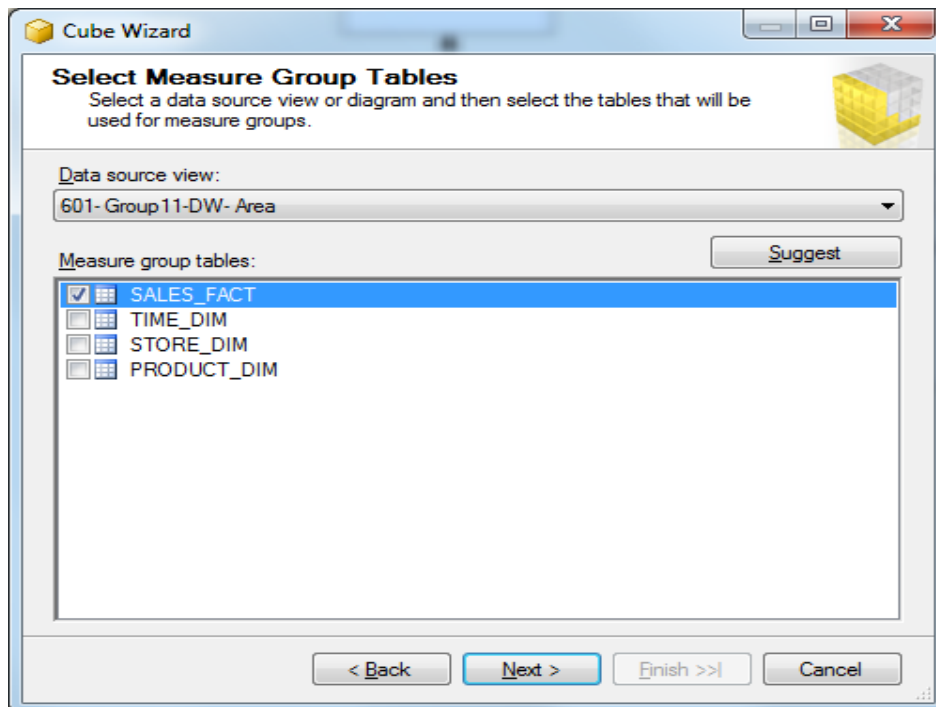
The data source view looks as follows.

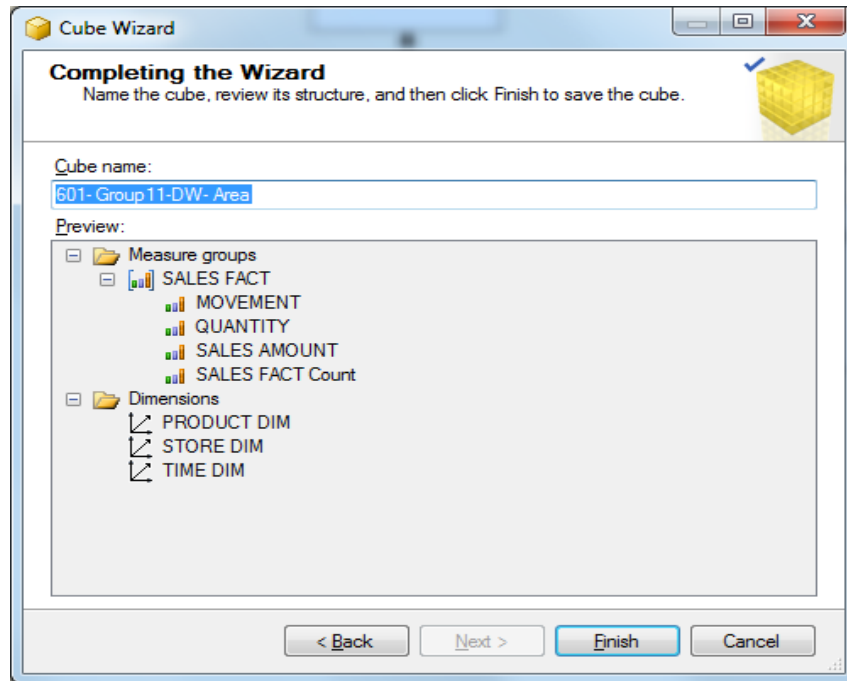


Configuring SSAS Multidimensional Database

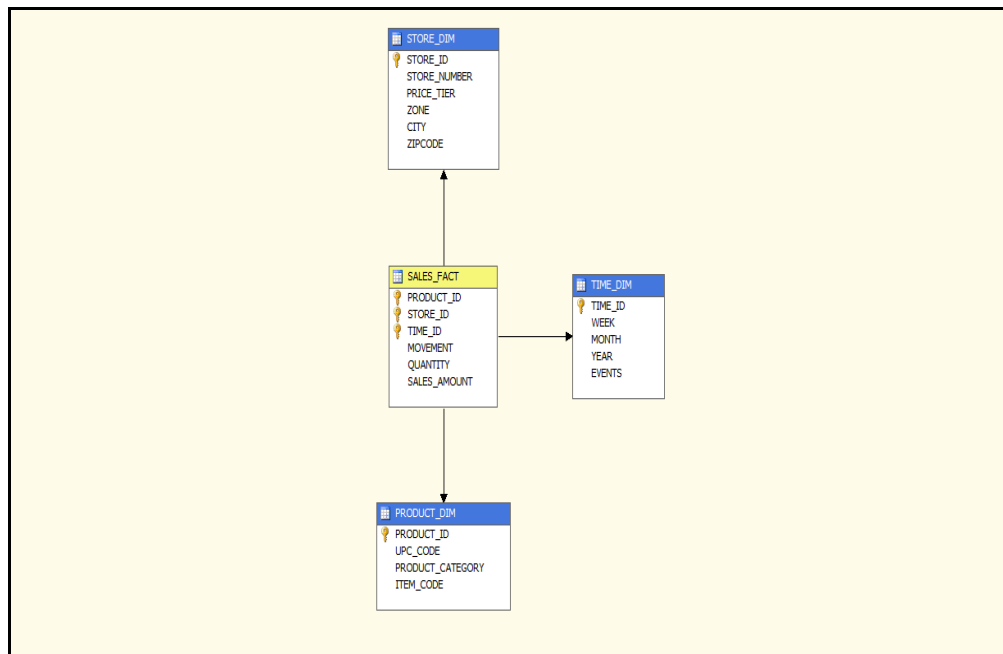


Creating Cube from Data Source View

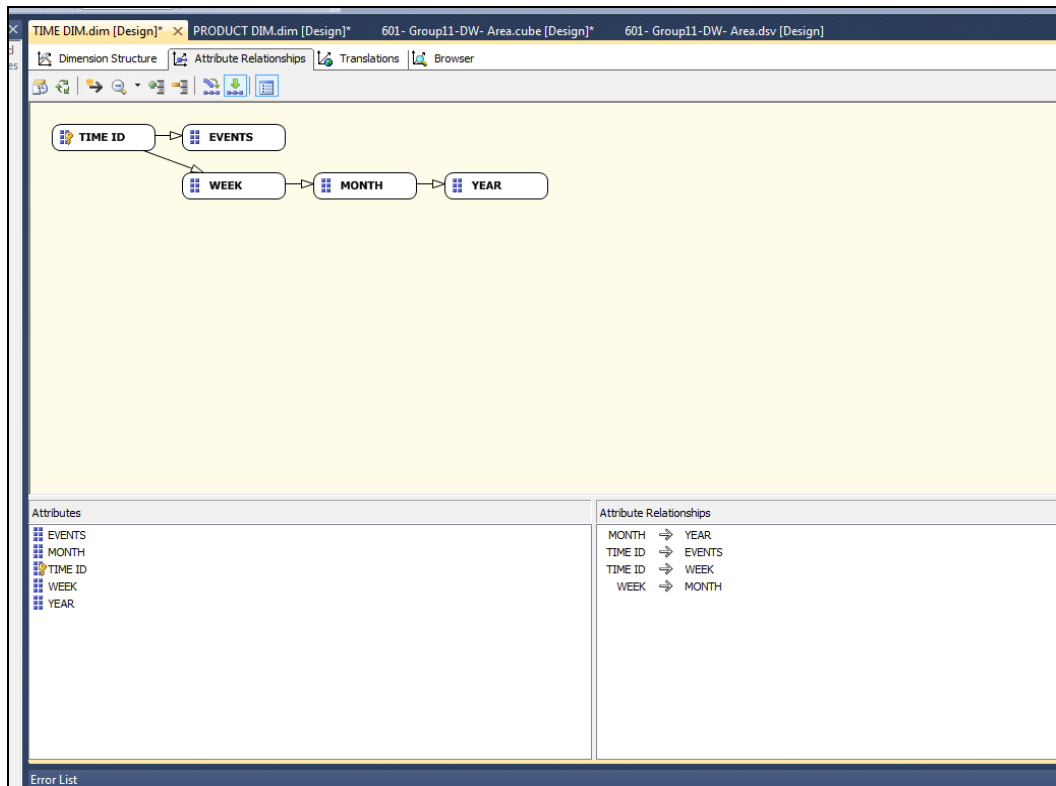




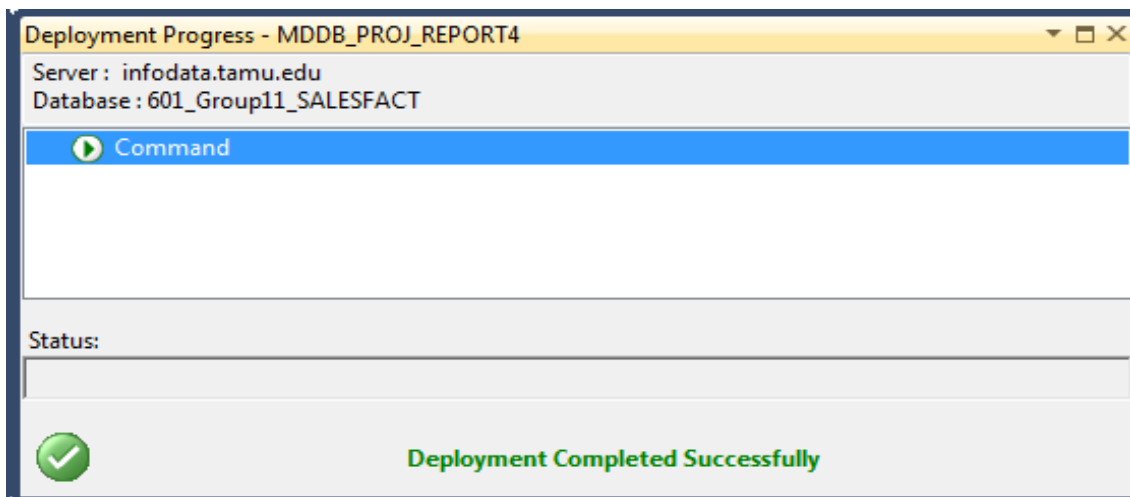
The cube appears as follows.

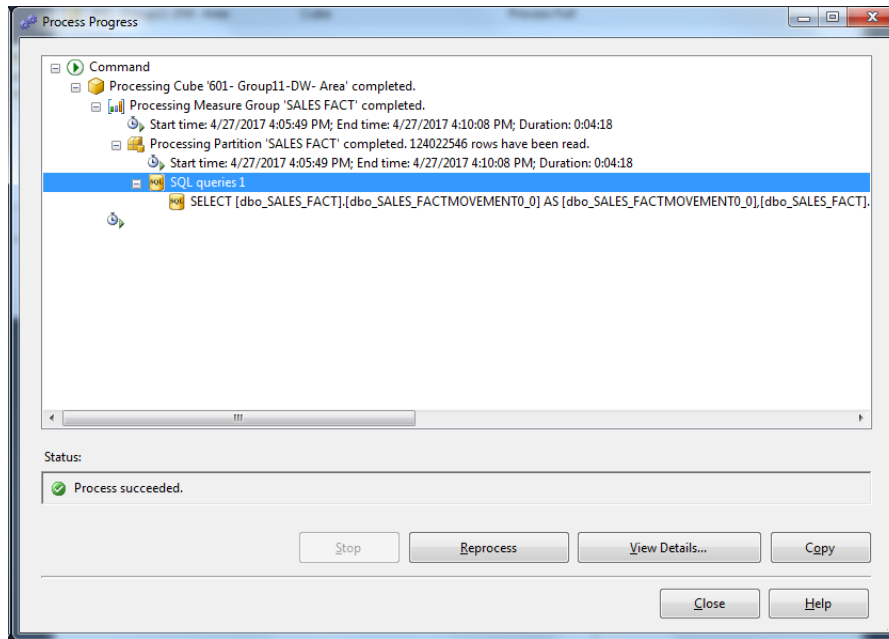


Defining hierarchies for Time Dimension

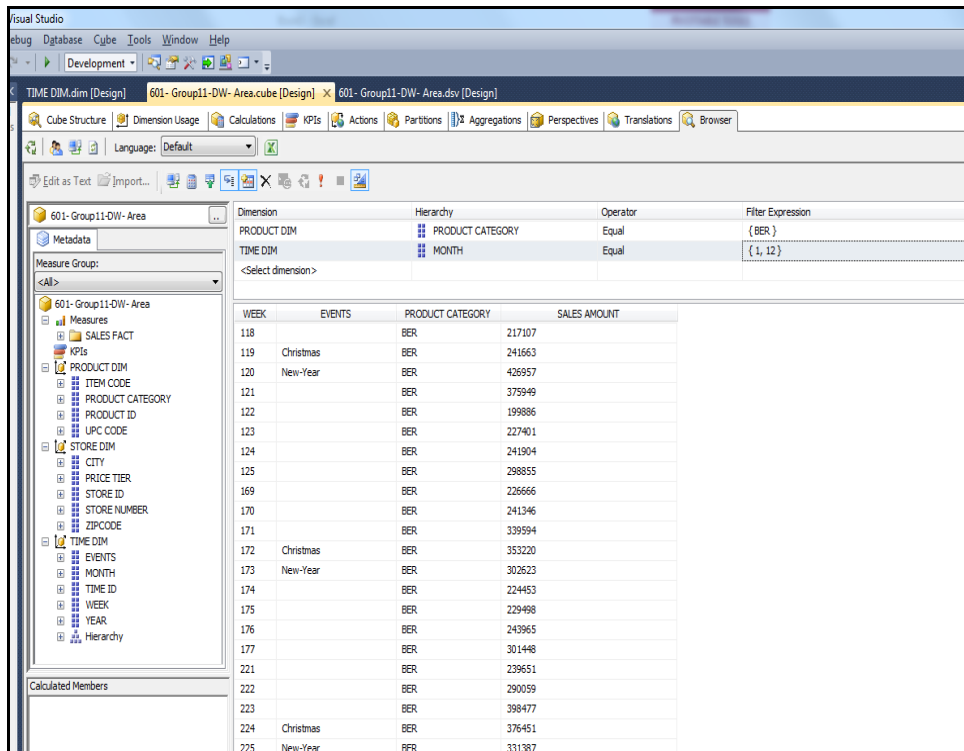


Deploying SSAS Cube



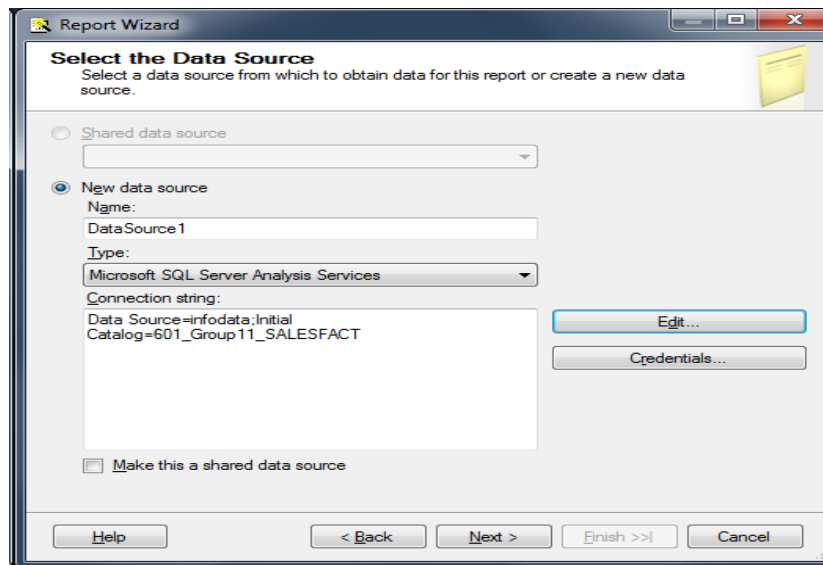


Browsing the Deployed Cube

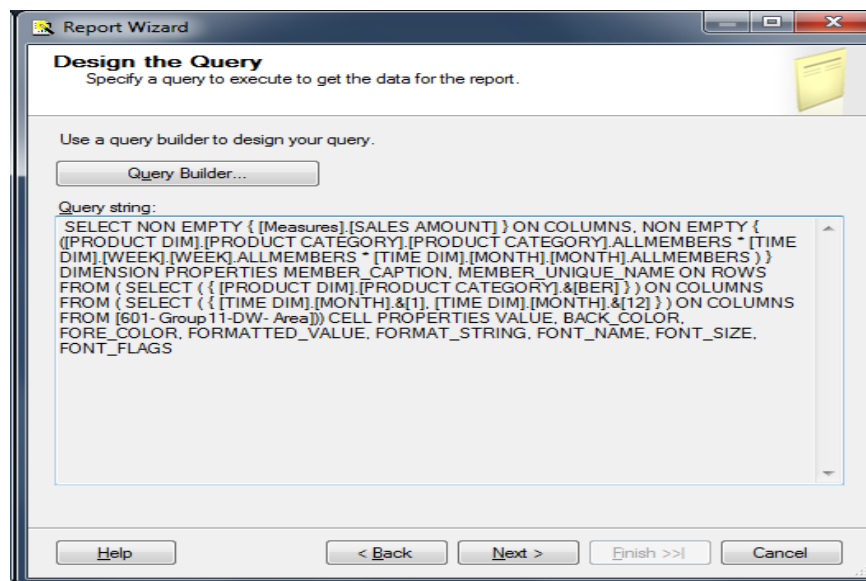


SSRS Reporting on Top of SSAS for Question1.

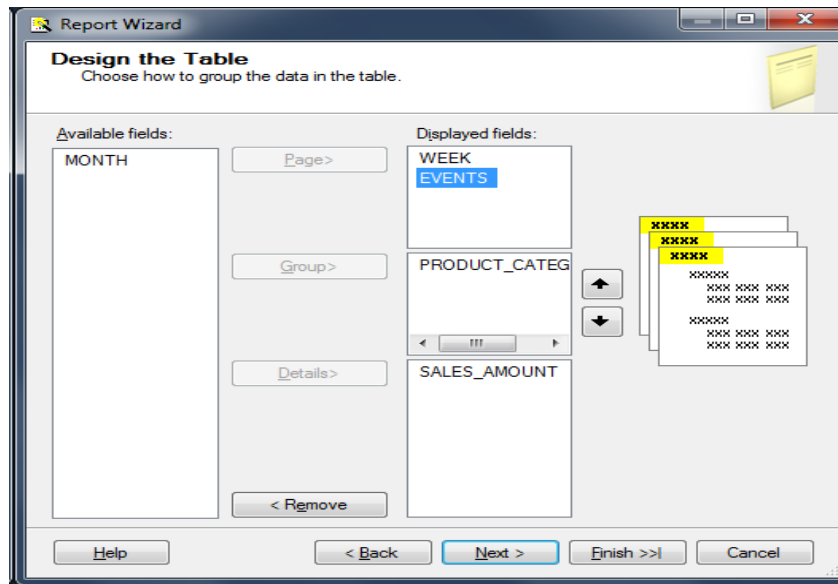
Selecting data source as Cube deployed in previous step



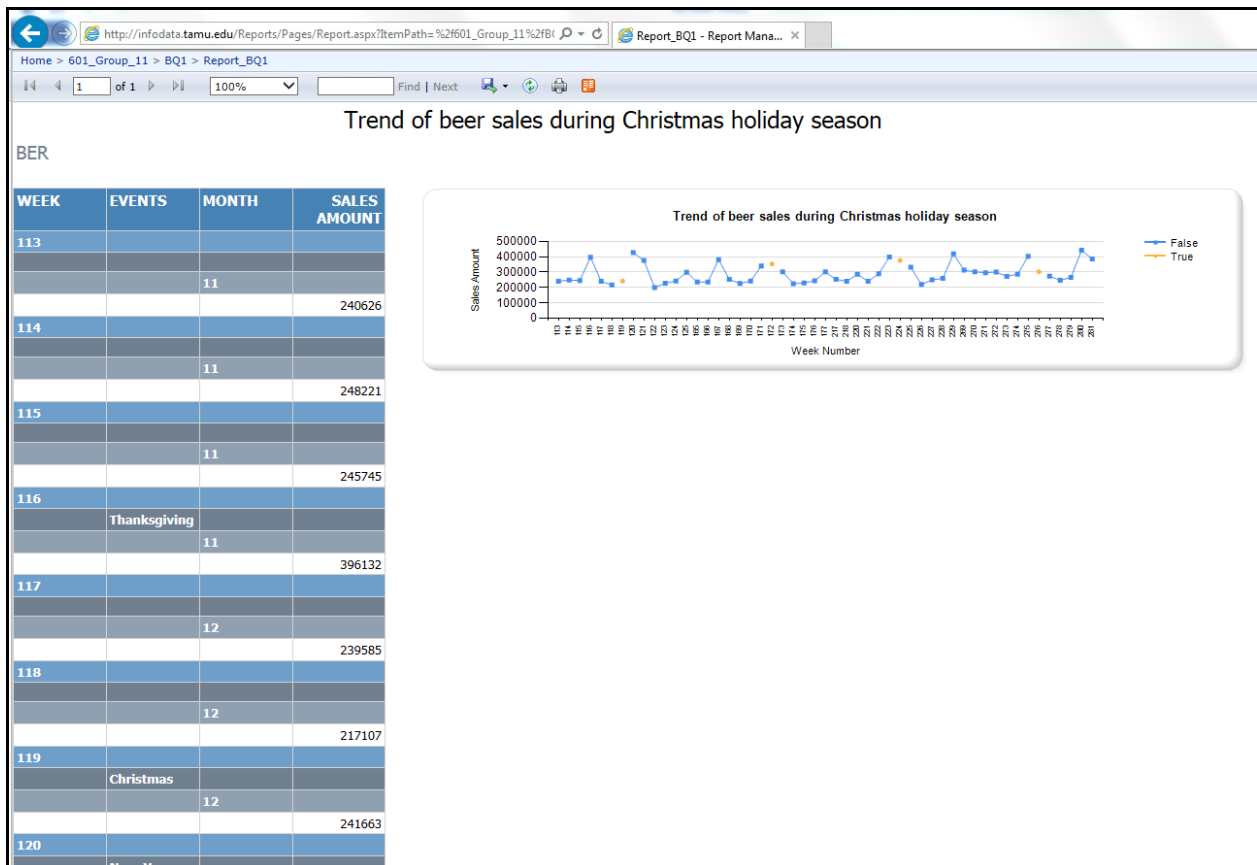
Designing query for Report generation



Designing table for report



Final Report deployed on the server



Conclusion

The above report plots the weekly sales of Beer in the month of November, December and January. From the graph, we get a clear idea about the trend of beer sales during the Christmas-New year holiday period. Also we can conclude that the Beer sales across the shops in Dominic Finer Food have several peaks in the months of November, December and January, even though the peak is not strictly during Christmas and New Year Weekends.

How the report assists the management:

This will give a clear direction to inventory team at DFF to store more Beer during the Christmas-New Year period. A similar approach can be used by Business Intelligence team at DFF for predicting the sales of different products during different festival season well ahead.

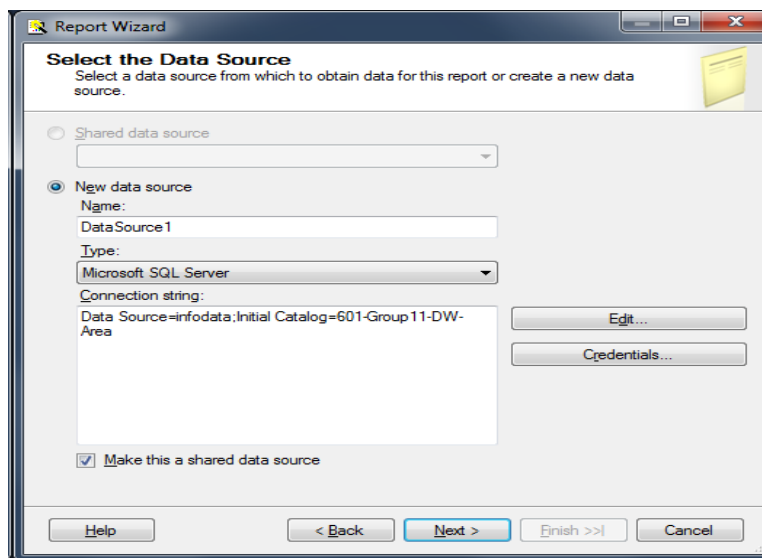
1.4. SSRS Reporting from Independent Data Mart using for Question 2

As per the reporting plan described in previous parts we are generating a report on SSRS reporting tool for answering the question number 2.

Question 2: Which stores have more popularity among kids and elderly groups?

Designing SSRS Report for Popularity among Kids

Selecting Data Source for SSRS



The screenshot shows the 'Report Wizard' dialog box, specifically the 'Select the Data Source' step. The dialog has a title bar with 'Report Wizard' and standard window controls. The main area is titled 'Select the Data Source' with a subtitle 'Select a data source from which to obtain data for this report or create a new data source.' There are two radio buttons: 'Shared data source' (unselected) and 'New data source' (selected). Under 'New data source', there are fields for 'Name:' (containing 'DataSource1'), 'Type:' (a dropdown menu set to 'Microsoft SQL Server'), and 'Connection string:' (containing 'Data Source=infodata;Initial Catalog=601-Group11-DW-Area'). To the right of the connection string field are two buttons: 'Edit...' and 'Credentials...'. At the bottom left of the main area is a checked checkbox labeled 'Make this a shared data source'. At the bottom of the dialog are five buttons: 'Help', '< Back', 'Next >', 'Finish >>', and 'Cancel'.

Designing the query for SSRS report

Column	Alias	Table	Outp...	Sort Type	Sort Order	Group By	Filter	Or...	Or...	Or...
STORE_ID		STORE_VIS...	<input checked="" type="checkbox"/>			Group By				
[BELOW_9_%]		DEMOGRA...	<input checked="" type="checkbox"/>	Descending	1	Group By				

```

SELECT TOP (10) STORE_VISITS_FACT.STORE_ID, DEMOGRAPHIC_DIM.[BELOW_9_%]
FROM STORE_VISITS_FACT INNER JOIN
    DEMOGRAPHIC_DIM ON STORE_VISITS_FACT.DEMO_ID = DEMOGRAPHIC_DIM.DEMO_ID INNER JOIN
    STORE_DIM ON STORE_VISITS_FACT.STORE_ID = STORE_DIM.STORE_ID
GROUP BY STORE_VISITS_FACT.STORE_ID, DEMOGRAPHIC_DIM.[BELOW_9_%]
ORDER BY DEMOGRAPHIC_DIM.[BELOW_9_%] DESC
    
```

STORE_ID	BELOW_9_%
34	0.19
93	0.19
68	0.18
65	0.18
77	0.18
62	0.18
72	0.17
10	0.17
88	0.17
45	0.17

Designing the table for SSRS Report

Design the Table
Choose how to group the data in the table.

Available fields: [Empty]

Displayed fields: STORE_ID, BELOW_9_

Buttons: Page>, Group>, Details>, < Remove

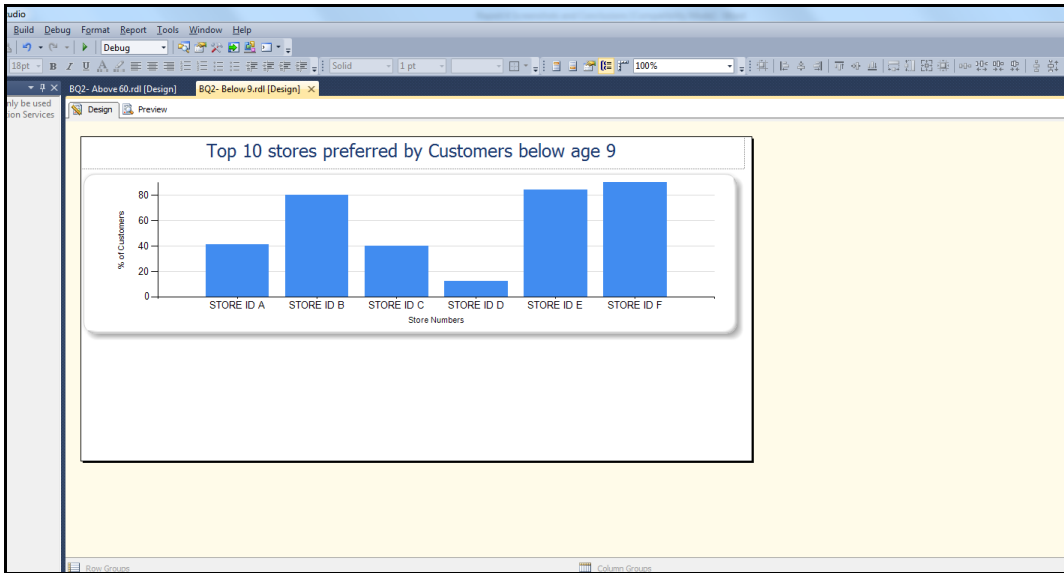
Preview:

```

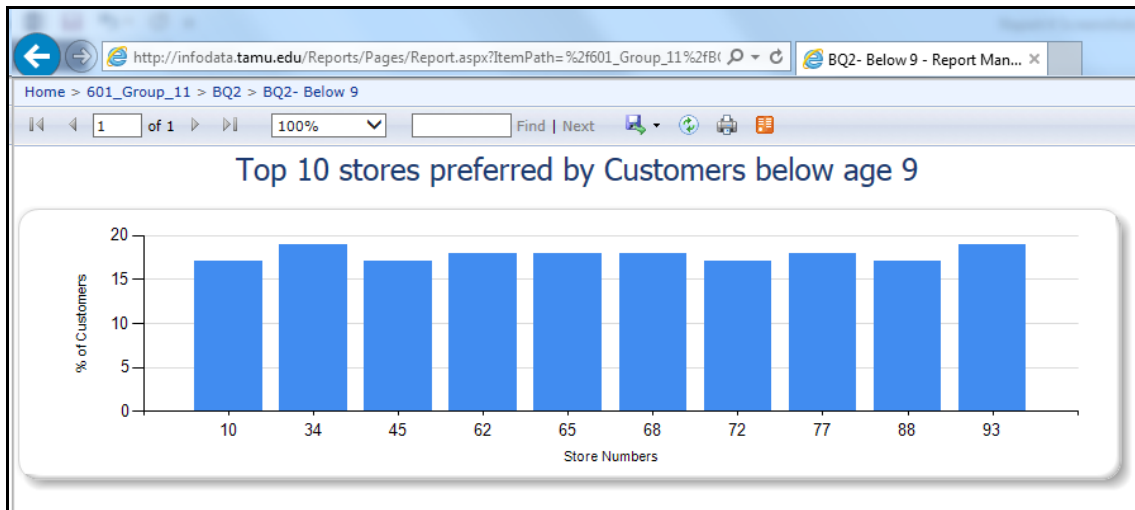
XXXXX
XXXXX
XXXXX
XXXX XXXX XXX
XXXX XXXX XXX
XXXXX
XXXX XXXX XXX
XXXX XXXX XXX
    
```

Buttons: Help, < Back, Next >, Finish >>!, Cancel

Plotted Bar Graph for store popularity among kids

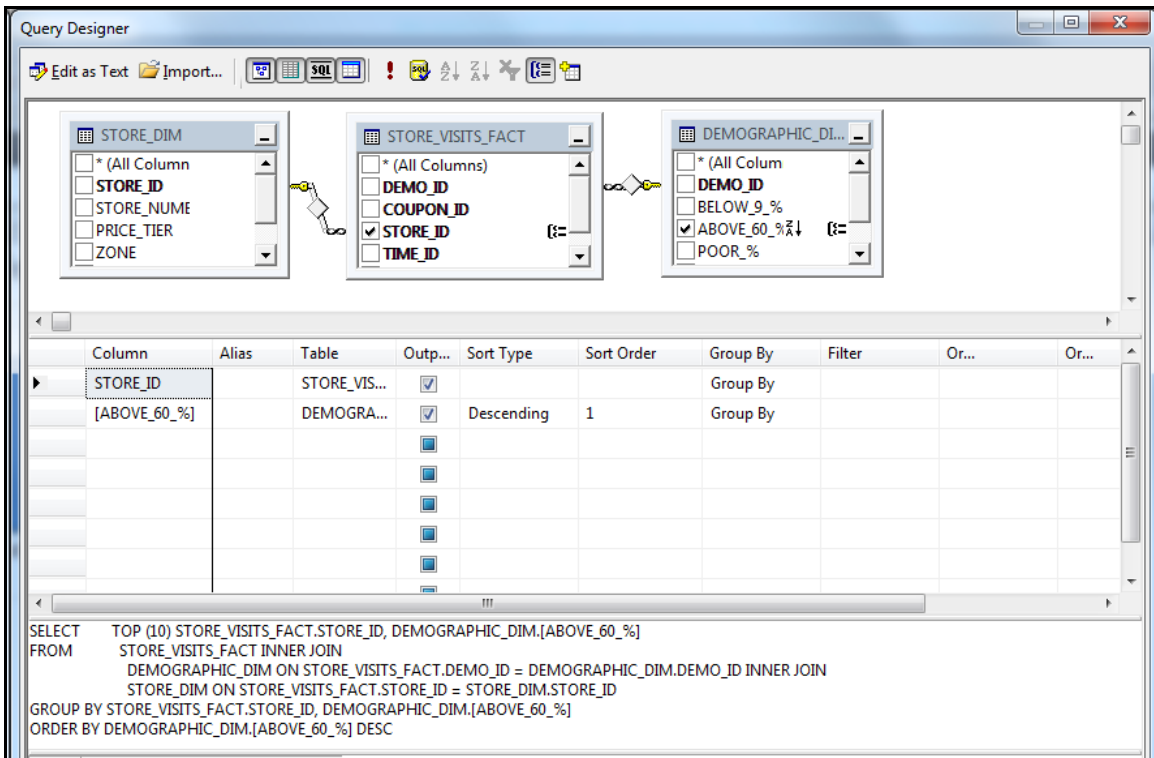


Browsing Report Deployed on the server



Designing SSRS Report for Popularity Among Elderly Customers

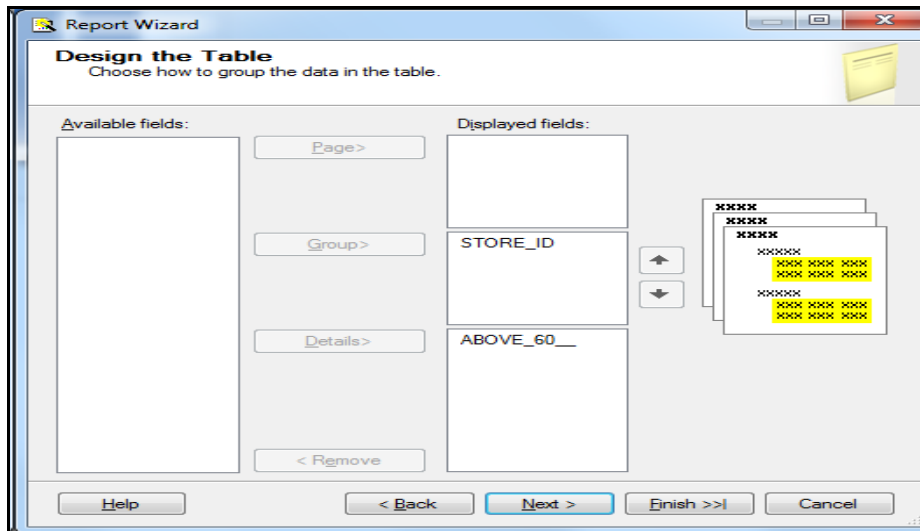
Designing Query for plotting Store popularity among elderly group



Column	Alias	Table	Outp...	Sort Type	Sort Order	Group By	Filter	Or...	Or...
STORE_ID		STORE_VIS...	<input checked="" type="checkbox"/>			Group By			
[ABOVE_60_%]		DEMOGRA...	<input checked="" type="checkbox"/>	Descending	1	Group By			

```
SELECT TOP (10) STORE_VISITS_FACT.STORE_ID, DEMOGRAPHIC_DIM.[ABOVE_60_%]
FROM STORE_VISITS_FACT INNER JOIN
    DEMOGRAPHIC_DIM ON STORE_VISITS_FACT.DEMO_ID = DEMOGRAPHIC_DIM.DEMO_ID INNER JOIN
    STORE_DIM ON STORE_VISITS_FACT.STORE_ID = STORE_DIM.STORE_ID
GROUP BY STORE_VISITS_FACT.STORE_ID, DEMOGRAPHIC_DIM.[ABOVE_60_%]
ORDER BY DEMOGRAPHIC_DIM.[ABOVE_60_%] DESC
```

Designing Table for Report



Design the Table
Choose how to group the data in the table.

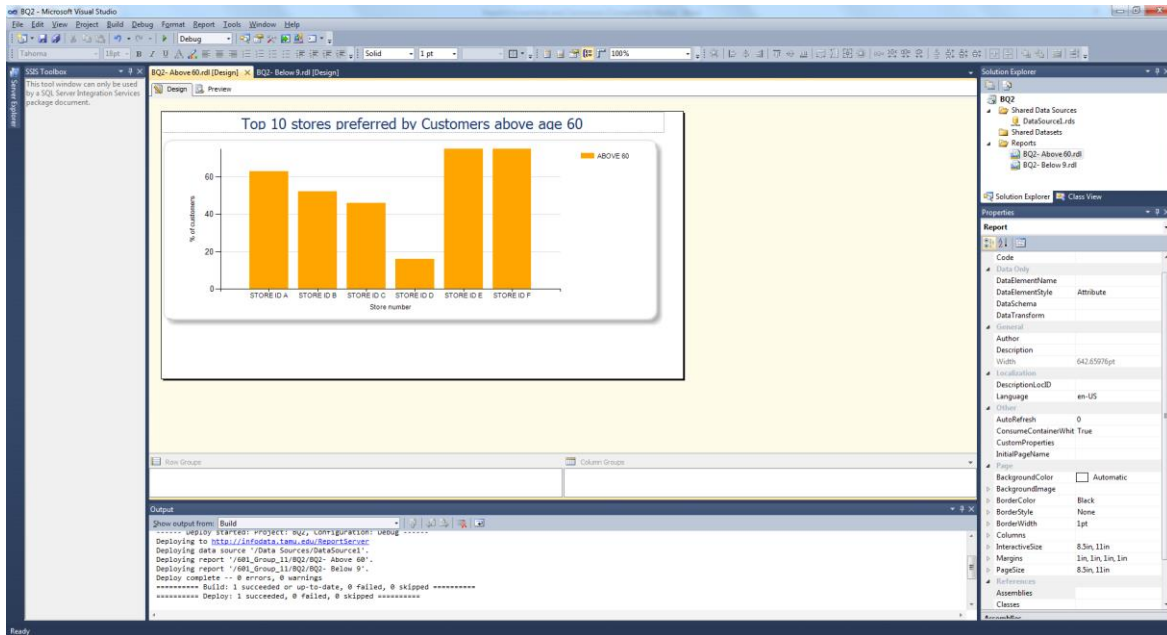
Available fields: [Empty]

Displayed fields: STORE_ID, ABOVE_60_

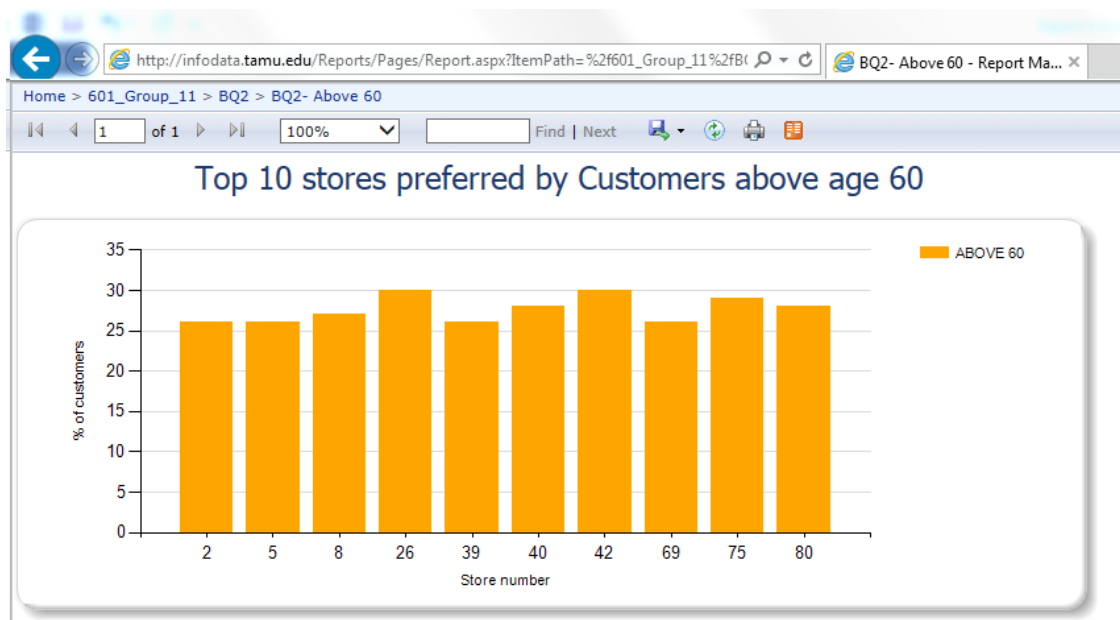
Buttons: Page>, Group>, Details>, < Remove, Help, < Back, Next >, Finish >>, Cancel

Preview: A table with columns for STORE_ID and ABOVE_60_.

Screenshot of report in the design view



Browsing Deployed Report



Conclusion

We have generated a SSRS report from an independent data mart for answering popularity of different stores in DFF among Kids and elderly groups. For kids and elderly groups, the report plots top 10 popular stores in each case considering percentage of age group visited among the total number of store visits. From this analysis we get the stores popular among the kids and elderly group directly. This is customizable to get top n number of stores among both groups. Report and chart generated with SSRS gives a clear idea of stores popular among kids and elderly group along with margin by which they are popular from the height difference of each plot in the graph.

How the report assists the management:

With popular stores list, now it is up to the upper management to understand what these stores are doing that is attracting these 2 groups and to spread the favorable conditions in this stores to other stores across DFF. Similar demographic analysis can be done by analytics team at DFF and strategies can be built on that for boosting the overall sales of DFF.

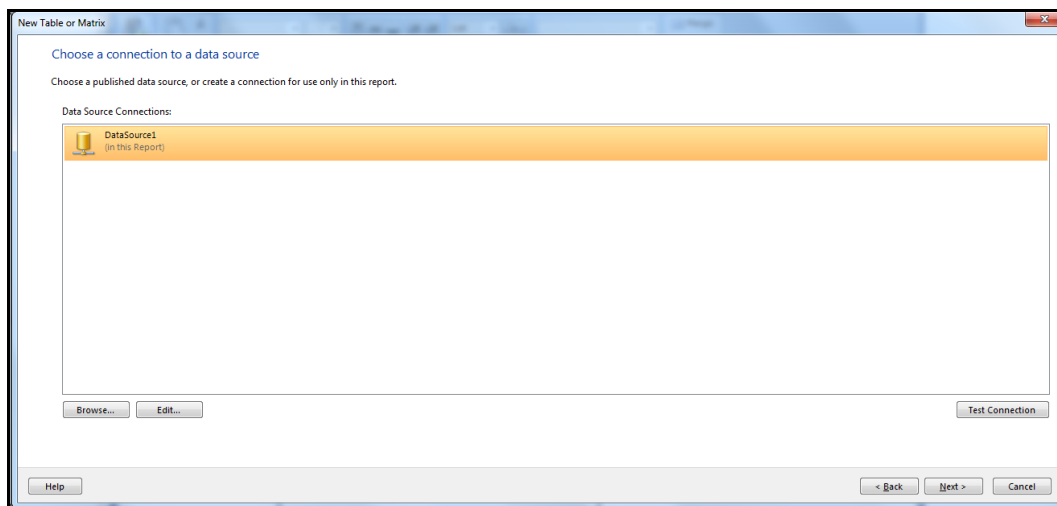
1.5. Report generated using Report builder 3.0 for Question 3

As per the reporting plan described in previous parts we are generating a report using Report Builder 3.0 for answering the question number 3.

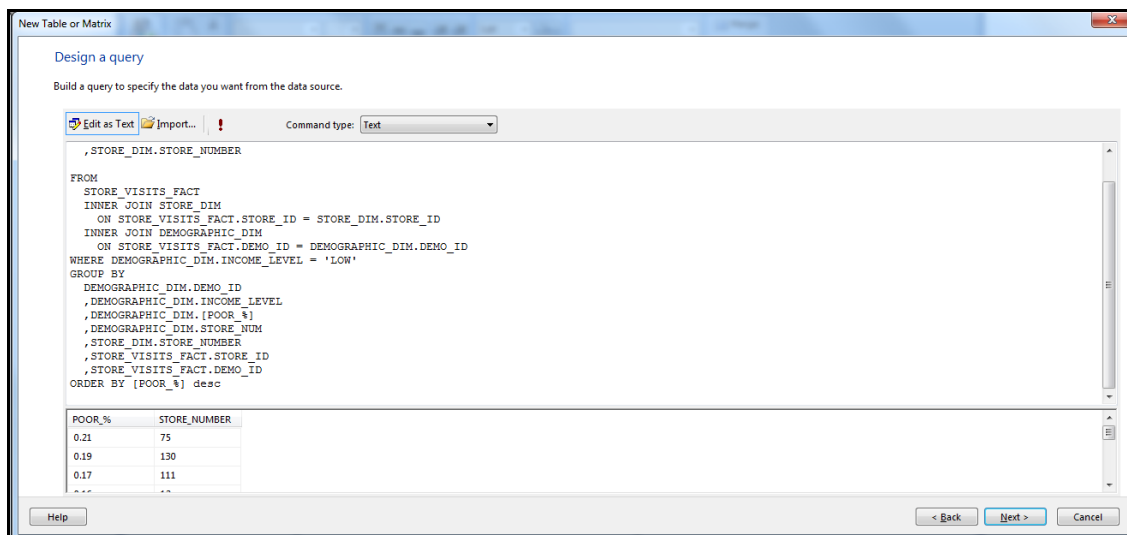
Designing Report Using Report Builder

Question 3: Which stores attract people who earn below poverty line?

Selecting data source



Designing query for report builder



Creating Table for Report

New Table or Matrix

Arrange fields

Arrange fields to group data in rows, columns, or both, and choose values to display. Data expands across the page in column groups and down the page in row groups. Use functions such as Sum, Avg, and Count on the fields in the Values box.

Available fields

- POOR_
- STORE_NUMBER

Column groups

Row groups

- STORE_NUMBER

Values

- Sum(POOR_)

Help < Back Next > Cancel

New Table or Matrix

Choose the layout

If you choose to show subtotals and grand totals, you can place them above or below the group. Stepped reports show hierarchical structure with indented groups in the same column.

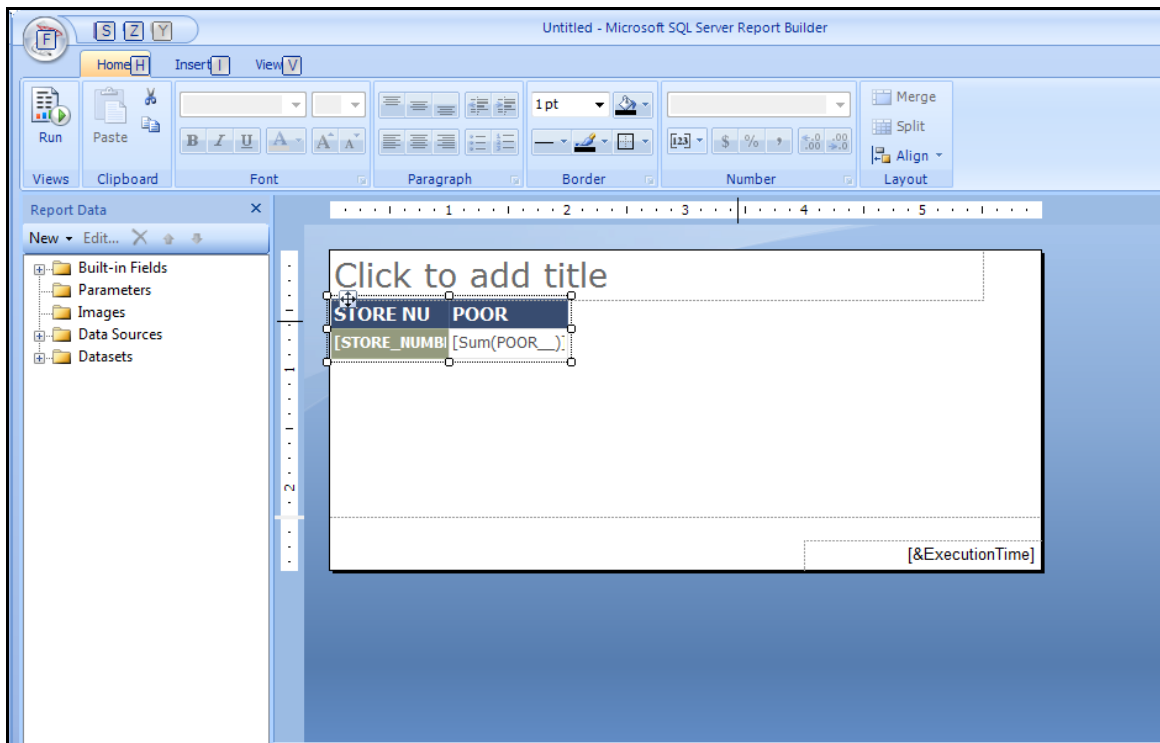
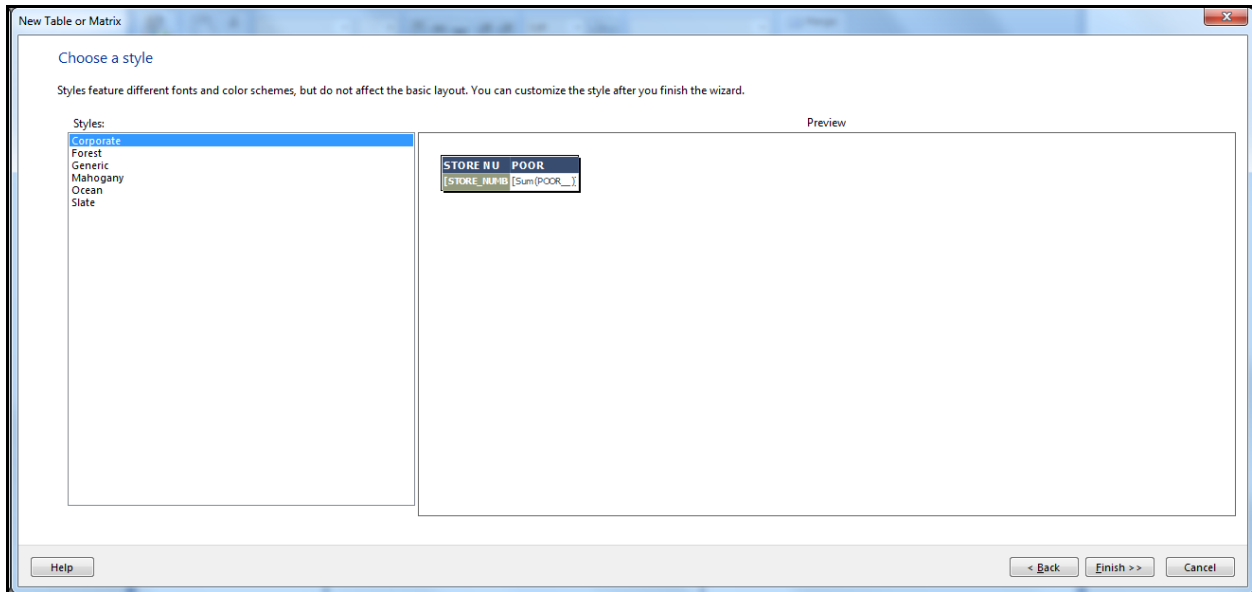
Options:

- Show subtotals and grand totals
 - Blocked, subtotal below
 - Blocked, subtotal above
 - Stepped, subtotal above
- Expand/collapse groups

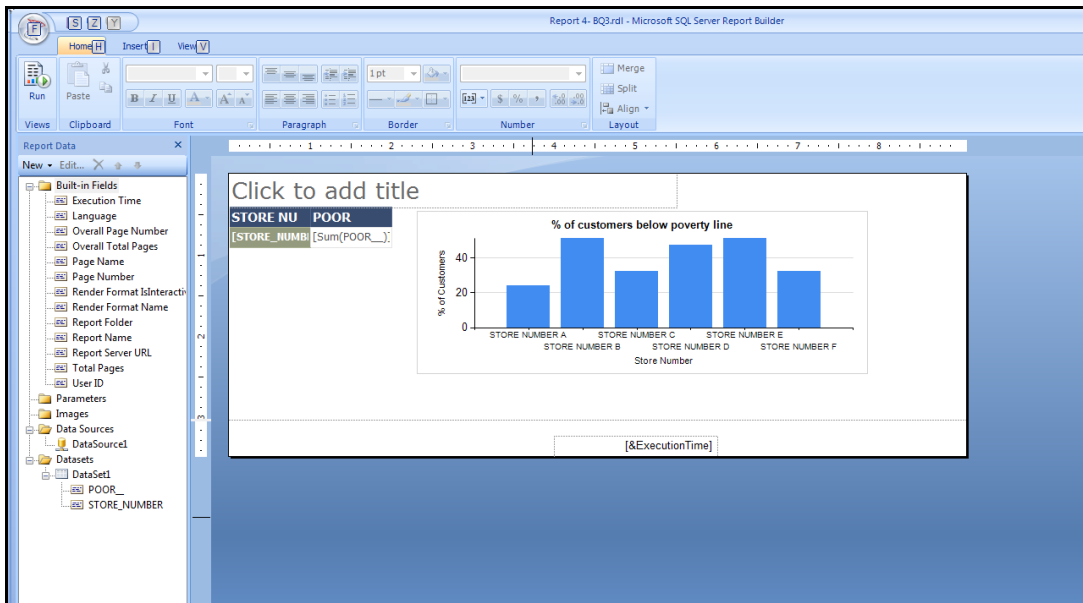
Preview

STORE NU	POOR
STORE_NUMBER	Sum(POOR_)
Total	Sum(POOR_)

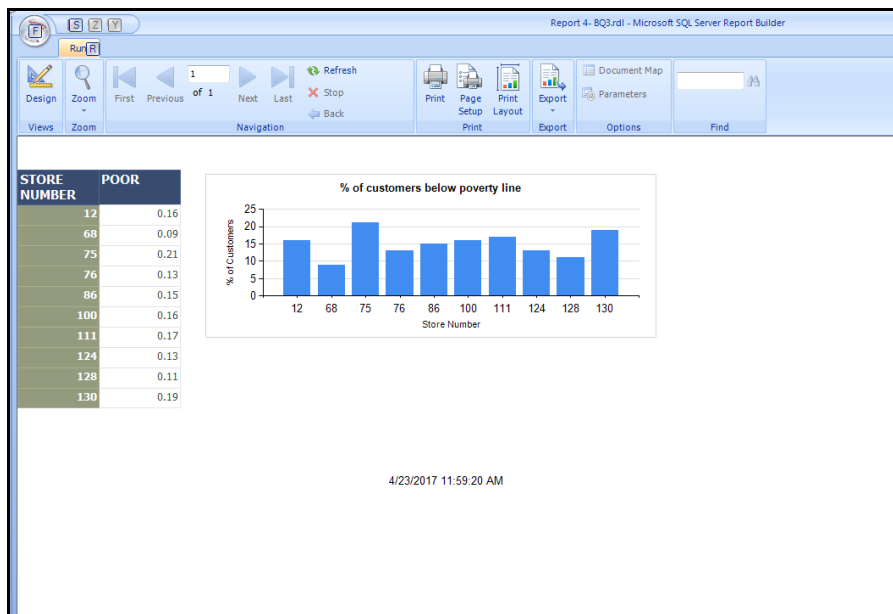
Help < Back Next > Cancel



Browsing Generated Graph



Snapshot of the report deployed on the server



Conclusion

We have generated a report using Report Builder 3.0 for answering popularity of different stores in DFF among section of people who earn below poverty line. The report plots Top 10 stores which are popular among people below the poverty line as a percentage of total store visits. From this analysis we got the Store numbers popular among this section of people. This we can customize to get top n number of stores among people below poverty line. Report and chart generated with Report Builder 3.0 gave a clear idea of stores popular among people who earn below poverty line along with margin by which they are popular from the height difference of each plot in the graph.

How the report assists the management:

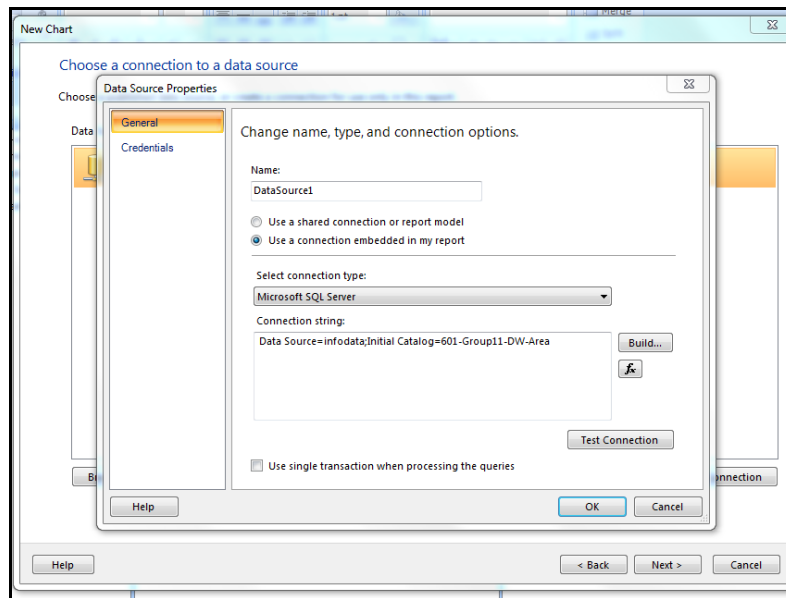
As we got the popular stores list, now it is up to the management to prepare the store increase the popularity among people below poverty line after analyzing the factors associated with this. This question is similar to the question number we answered previously and more such demography analysis can be done in similar way.

1.6. Report generated using Report builder 3.0 for Question 4

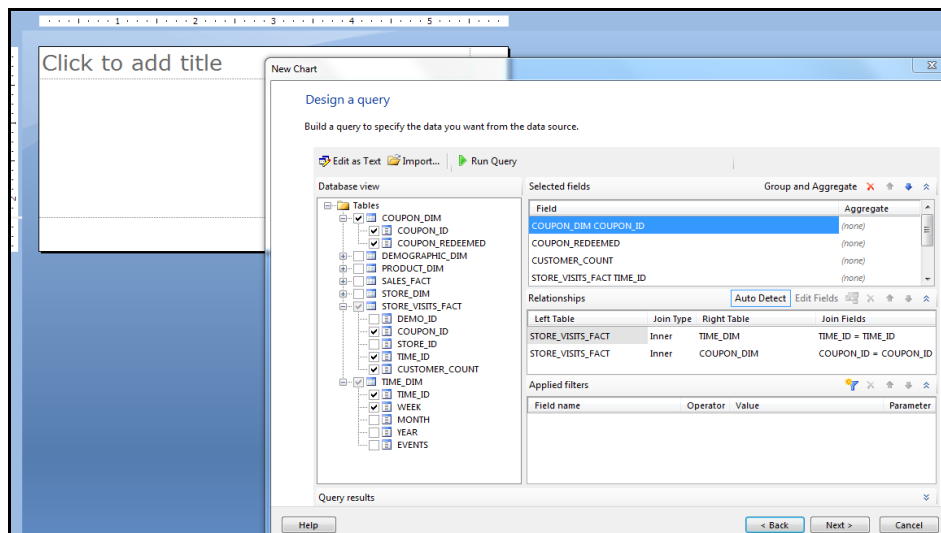
As per the reporting plan described in previous parts we are generating a report using Report Builder 3.0 for answering the question number 4.

Question 4: What is the effect of introducing coupons on total number of customer visits?

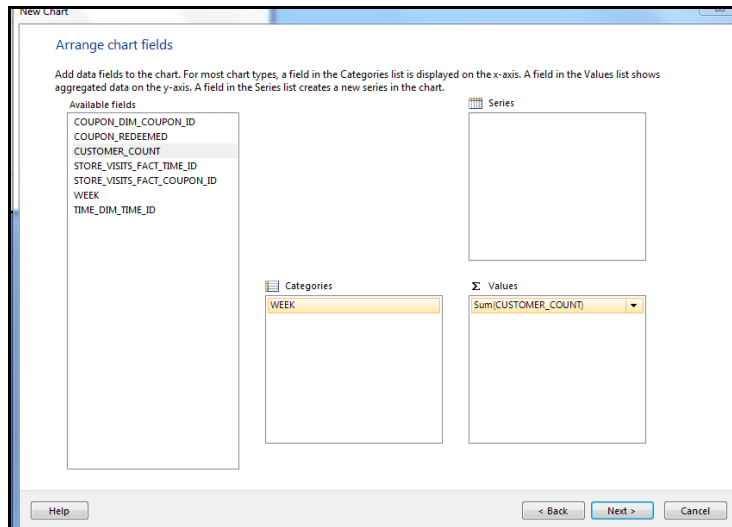
Selecting data source for Report Builder 3.0



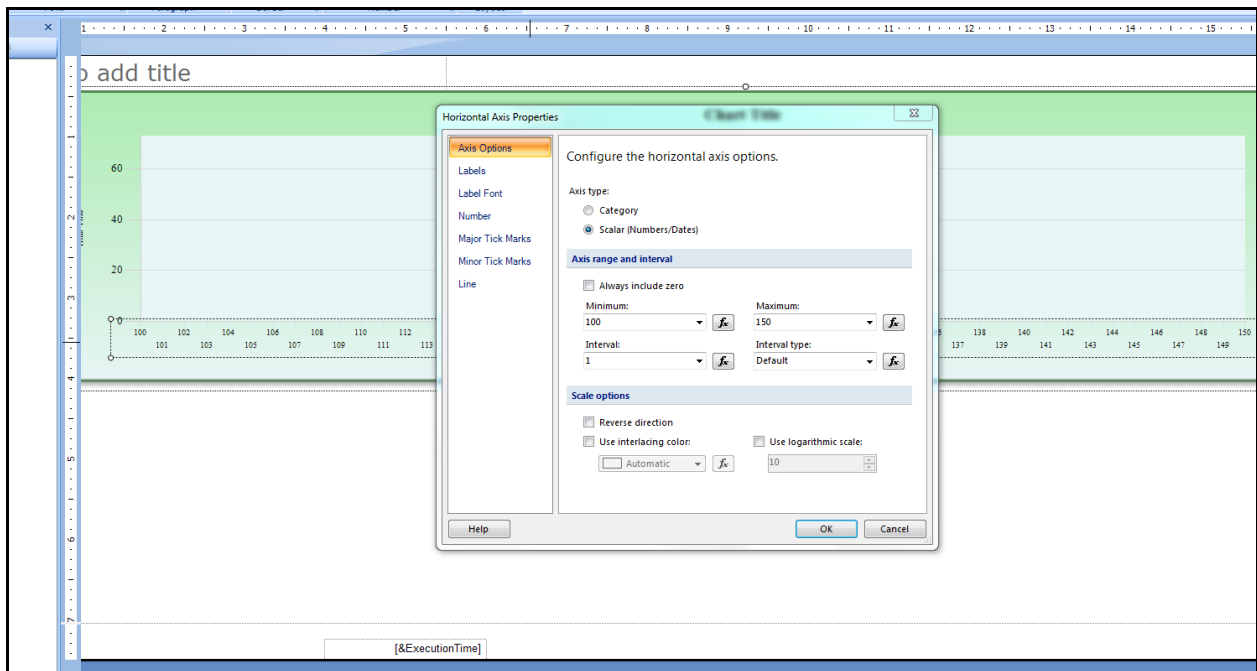
Designing query to pull the data required for report



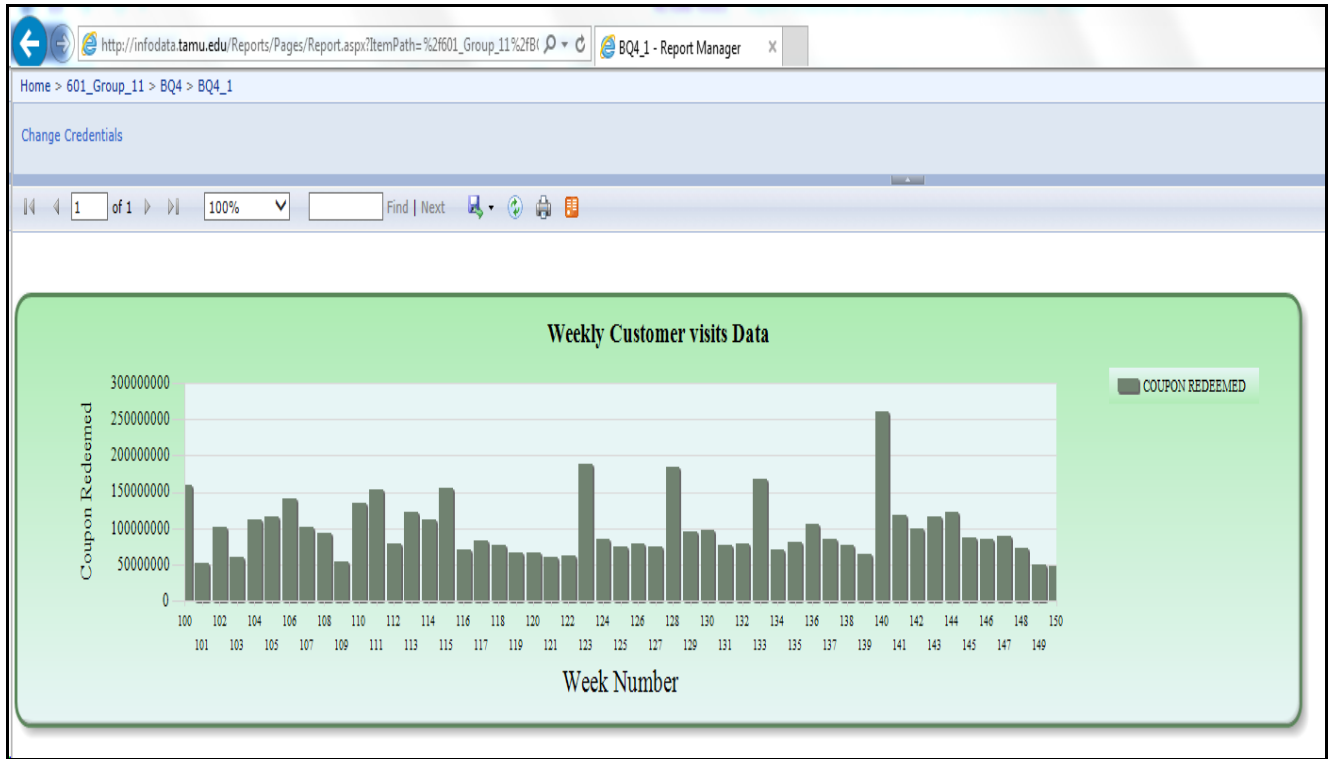
Assigning fields for the report table



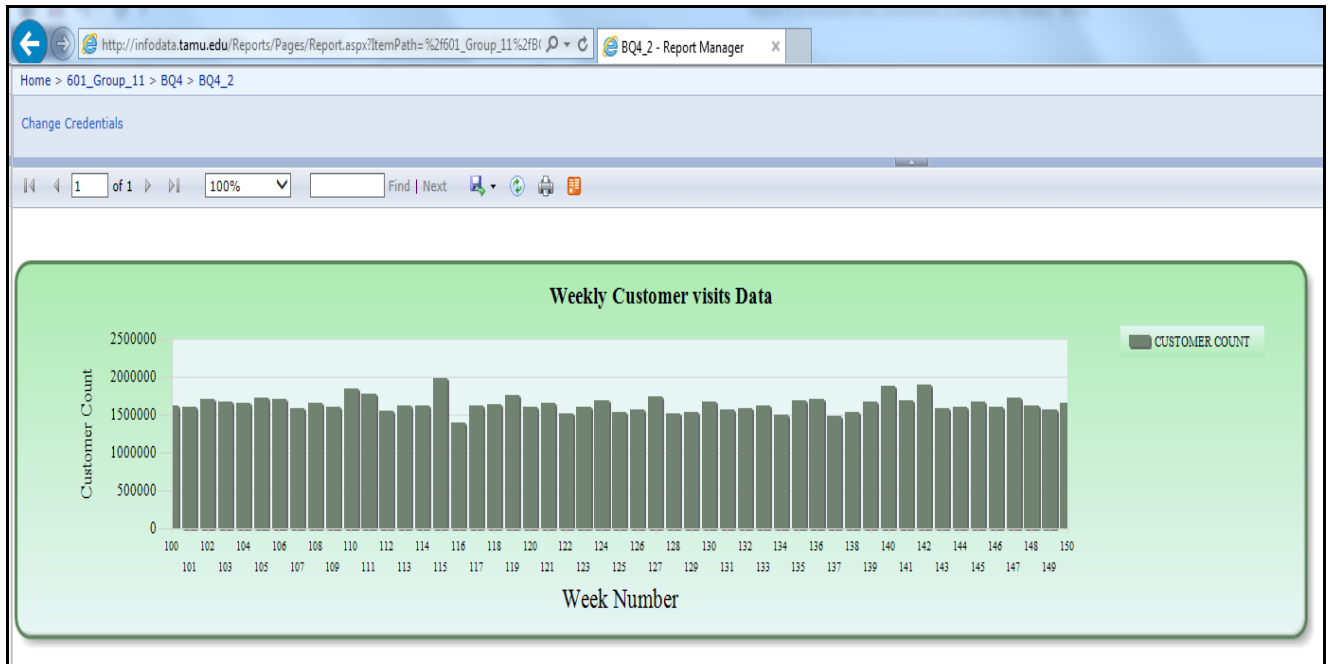
Configuring the horizontal axis interval in the report



Coupon redeemed values plotted on weekly basis



Customer Visits data plotted for the same weeks for which Coupon redeemed is plotted



Conclusion

There is a general opinion that introducing coupons increases the customer flow. Hence in this question we expected a correlation between the Coupons redeemed and number of customer visits as Coupon issued is considered as kind of freebie to the customers. From the graph we could conclude that number of customer visits is almost constant among the weeks under consideration here, whereas the coupon redeemed have its own peaks and lows for the weeks under consideration. From this it is evident that Coupon redeemed have no relation with the number of customer visits.

How the report assists the management:

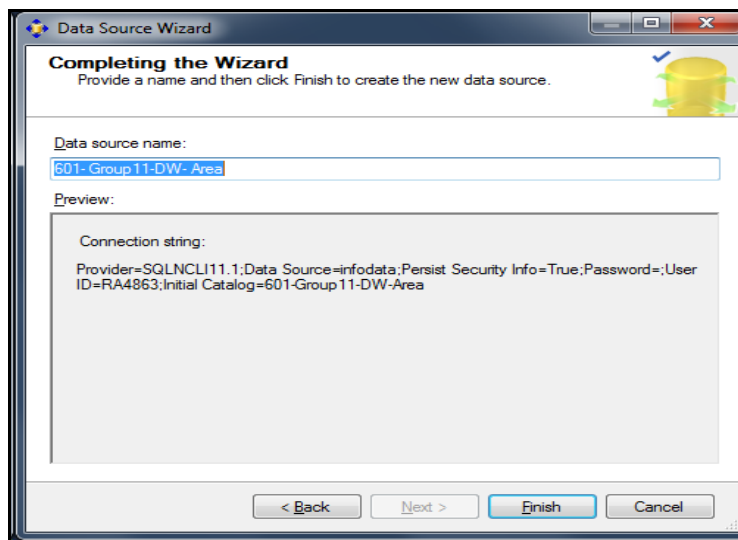
As per the result Coupon redeemed might have only decreased the total sales due to discounts and has no direct effect in increase in Customer counts. So the management can look into decreasing the coupons disbursed as the number of customer visits remains unaffected.

1.7. Cubes using SSAS from an independent data mart for Question 5

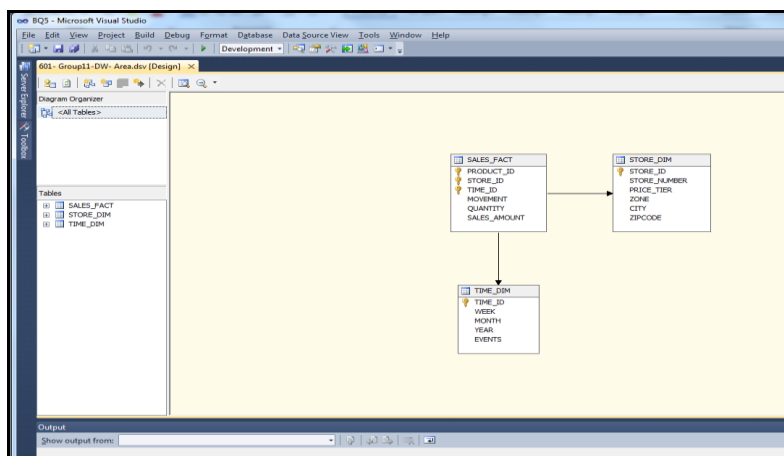
As per the reporting plan described in previous parts we are generating a report SSAS using an independent data mart for answering the question number 5.

Question 5: What is the trend of a product demand in different price-tiers?

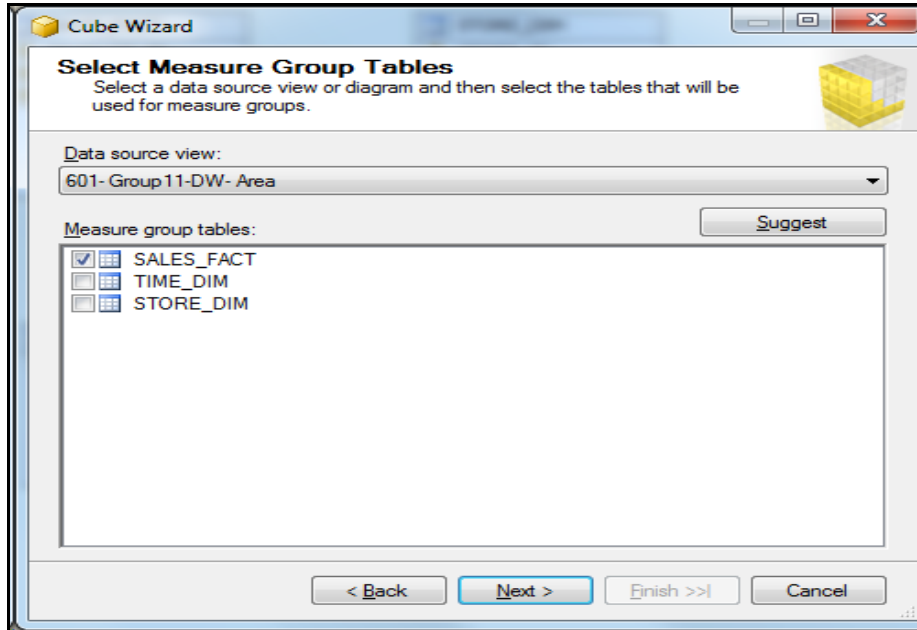
Selecting data source for SSAS



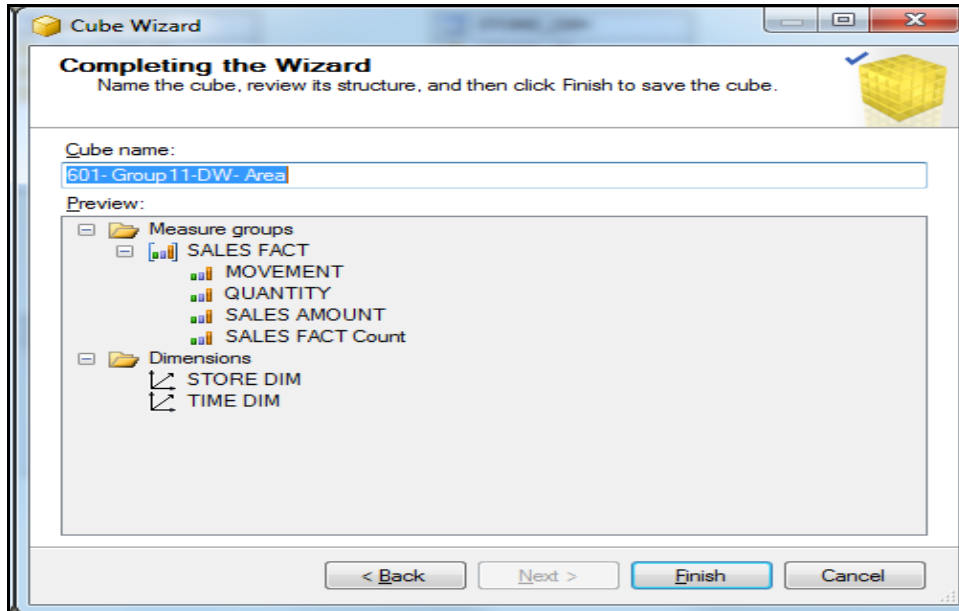
Designing data source view



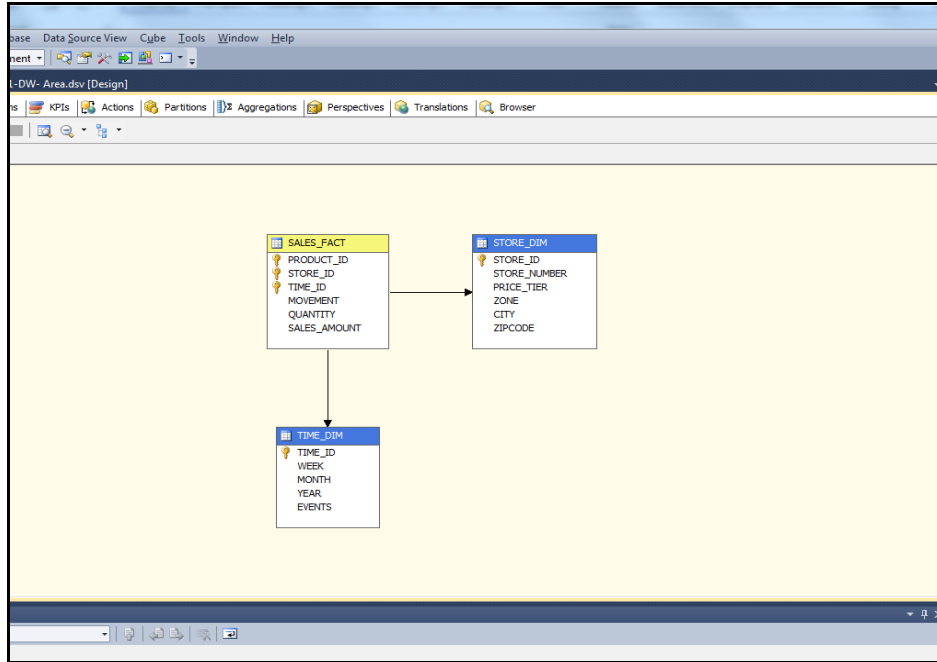
Fact table set as the measures Group table



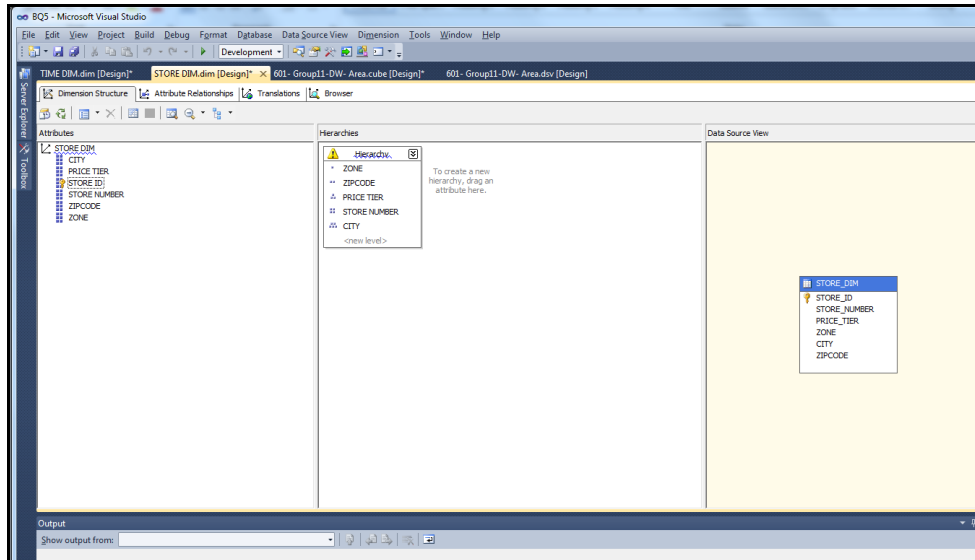
Completing the wizard after defining dimensions



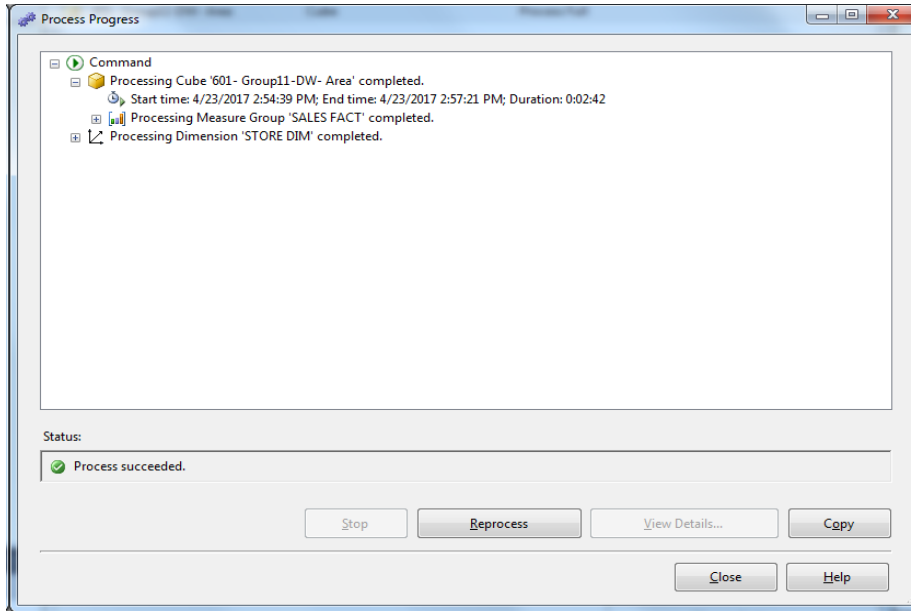
Created Cube



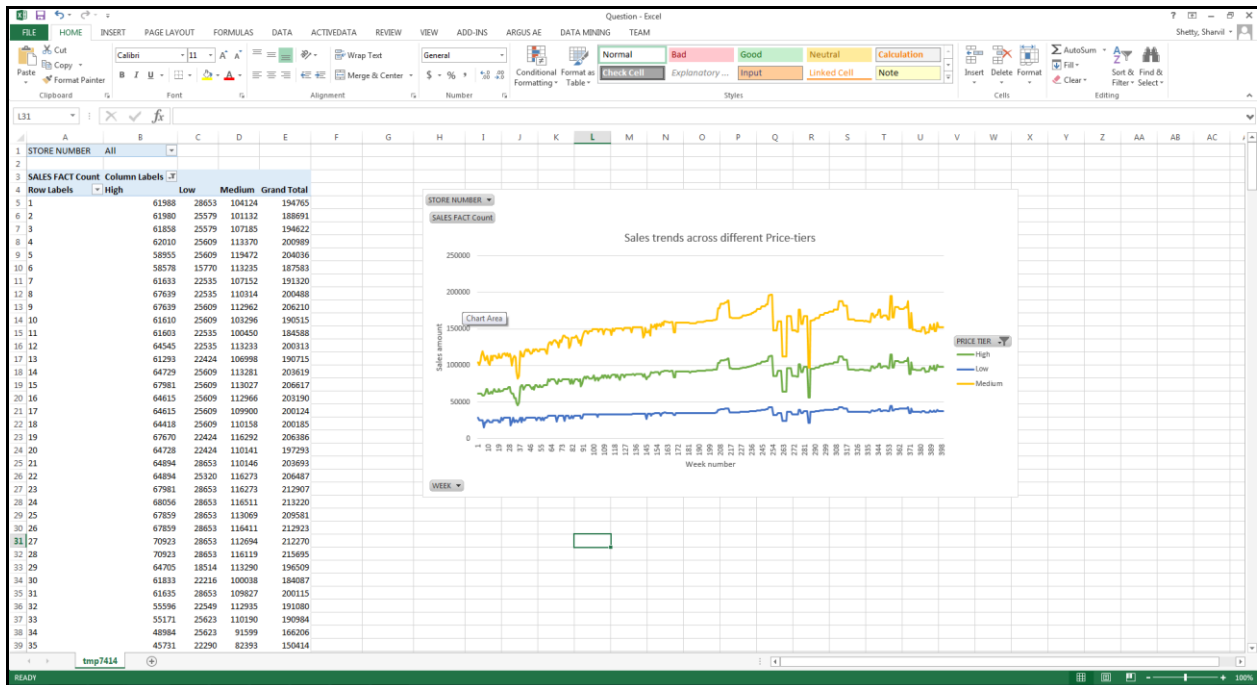
Defining hierarchies for drill down and roll up



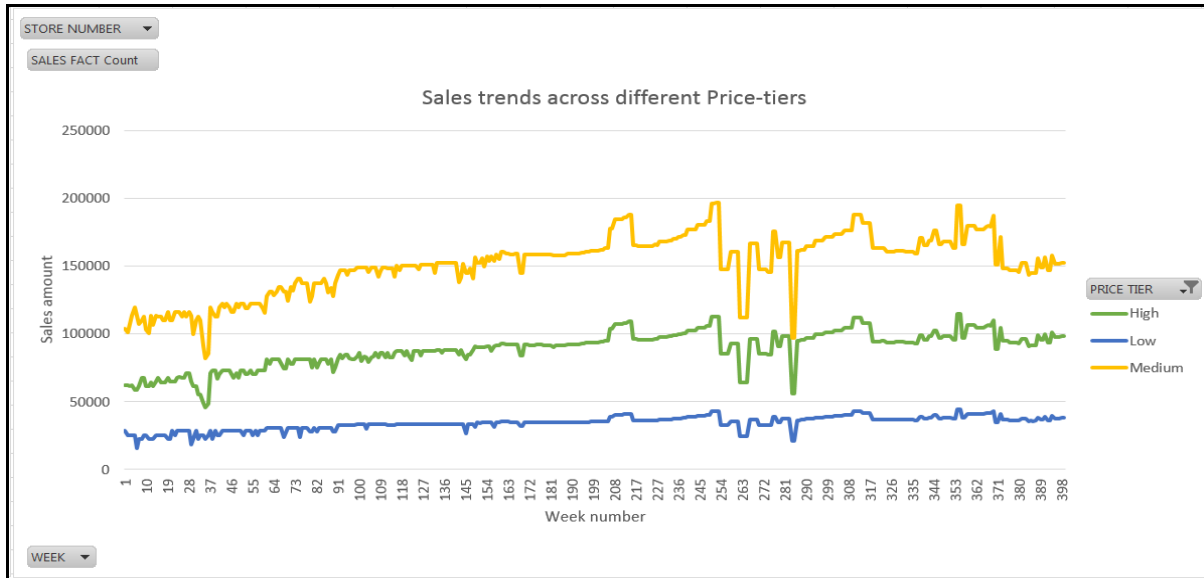
Deploying Cube for analysis



Report generated using Pivot table



Sales chart for different Price Tiers



Conclusion

Sales amount in each price tier is calculated by adding individual store sales amount in each price tier on weekly basis. So the graph was plotted against Sales amount and Week. Surprisingly we got result from analysis had strikingly similar correlation between the different Price tiers High, Low and Medium. Whenever there is a dip or peak in sales amount for a particular week for any price tier the same trend was followed by other price tiers. This was important finding after doing analysis for question 5.

How the report assists the management:

The difference in Sales amount across different price tiers remain constant throughout the analysis duration. Management can dig up further for knowing the reasons for constant difference across varying price tiers.

References

1. Nevo, Aviv and Catherine Wolfram, "Why do Manufacturers Issue Coupons? An Empirical Analysis of Breakfast Cereals," *The RAND Journal of Economics*, 33 (Summer 2002): 319-339.
2. Sivakumar, K., "Quality-Tier Competition and Optimal Pricing," *Journal of Business Research* 33 (1995): 251-260.
3. Lira, Loreto, "Why Do Some Prices in the Retail Sector Drop When Demand Rises? Evidence from the Chilean Case," *Revista ABANTE* 10 (Octubre 2007): 151-168
4. <https://en.wikipedia.org/wiki/Dominick's>
5. <https://research.chicagobooth.edu/marketing/databases/dominicks/docs/1994-shelfmanagement.pdf>
6. Drèze, Xavier; Hoch, Stephen J.; Purk, Mary E. (1994): *Shelf Management and Space Elasticity*
7. <http://research.chicagobooth.edu/marketing/databases/dominicks/papers.aspx>
8. www.chicagobooth.edu/research/kilts/marketing-databases/dominicks/demo
9. https://research.chicagobooth.edu/marketing/databases/dominicks/docs/2002_Why_Do_Manufacturers.pdf
10. <https://research.chicagobooth.edu/marketing/databases/dominicks/docs/1995-Quality-TierCompetition.pdf>
11. <https://research.chicagobooth.edu/marketing/databases/dominicks/docs/2007-WhyDoSomePricesDrops.pdf>