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

Report 4: BI Reports Design and Implementation

Dated: 05.02.2016

Group 4: *Gaurav Diwanji, Pratik Mrinal and Varun Bindra*

Contents

Contents 2
1. Introduction 4
2. Details about the Data 5
   a. Understanding of the Data: 5
   b. Describing Metadata for all the OLTP files: 7
   c. E-R diagram to show the relationship among the major entities on a higher level: 10
3. Domain understanding 11
4. Business Questions 13
Business Questions Explained: 14
   BQ.1 Evaluating the store performance with respect to all the other stores at DFF 14
   BQ2 Identify the most concerned areas of losses for each shop for Dominick's Fine Food (DFF) week-wise. 16
   BQ 3 Generate the report depicting the effect of coupon introduction for increasing the customer count in the store 18
   BQ 4 Identifying the contribution of each product category in the overall sales for DFF 20
   BQ 5 Identifying price elasticity of demand for a particular item, w.r.t. income levels 22
   BQ 6 Identifying the contribution of each store/zone on the overall sales of DFF 24
   BQ 7 How sales of an item of varying packet size is impacted by price change 26
   BQ 8 Forecasting profitability and sales of an item for measured changes in item pricing 28
   BQ 9 Generate the report depicting the sales summary for different types of product for different stores. 30
   BQ 10 Generate the report depicting the effect of coupon introduction for increasing the sales in poverty affected areas. 32
Selected Questions 34
5. Proposed Star schema 35
6. STAR SCHEMA REPRESENTATION 37
7. Mapping Tables 39
8. ETL Development Plan 42
   8.1 Data Quality Concerns with the DFF Data Sets: 42
   8.2 ETL Development Plan: 43

A. Determine target data 43
B. Determine source data 46
C. Mapping tables for staging and data mart loads 46
D. Comprehensive data extraction rules 50
E. Data staging area screen shots 51
F. Data transformation and cleansing rules 55
G. Plan for aggregate tables 59
H. Procedures for all data extractions and loadings 59
I. Mappings definition describing the source to end table for all dimension and fact tables 64
J. SQL Statements used for the ETL operations 65
K. Before After table screenshots 67
9. BI Reporting: 72
BQ2: Identify the most concerned areas of losses for each shop for Dominick's Fine Food (DFF) week-wise. 72
BQ3: Generate the report depicting the effect of coupon introduction for increasing the customer count in the store 76
BQ 4: Identifying the contribution of each product category in the overall sales for DFF 85
BQ 5: Identifying price elasticity of demand for a particular item, w.r.t. income levels 91
BQ 7: How sales of an item of varying packet size is impacted by price change 100
10. Data warehouse infrastructure and front end tools used: 111
11. References 113
12. Work break down 113

1. Introduction

Dominick's Fine Food (DFF) was a grocery store chain in Chicago area and subsidiary of Safeway Inc. With an intent of providing business oriented research for better shelf management and pricing issues, Chicago Booth and Dominick's Fine Foods entered into an agreement between 1989 to 1994, where a set of random researches were conducted in more than 25 different categories throughout all stores in this 100-store chain. With an exhaustive set of data of more than 3500 UPCs obtained from this research, this group project aims at handling few of the very critical business problems for DFF as mentioned below:

a. *Stock out issue:* Stock out of the product has been a very critical issue not for the store chains selling the product, but to the vendors as well. Recent studies have proved that in case of unavailability of the product, the customers are more likely to switch the stores and never come back, giving a direct revenue setback to the retail stores.

b. *Profit margin issue:* To curb the stock out issue, few of the retailers increased their safety stock and over stocked their backups. But this substantially lead to reduced profit for the retail chain and many a times depending upon the market trends and dynamic nature, the piled up product back up were not even used up against shortage cases faced in other scenarios.

c. *Competitive marketing issue:* Adding on the top of the above mentioned concerns for the retail stores is the competitive superiority over their counterparts in the business. Struggling with inner forecasting issue of the product and related profit margins, DFF needs to make sure of making a detailed and realistic projection for every product over every granular timeframe so as to stay above the competition for its own successful existence.

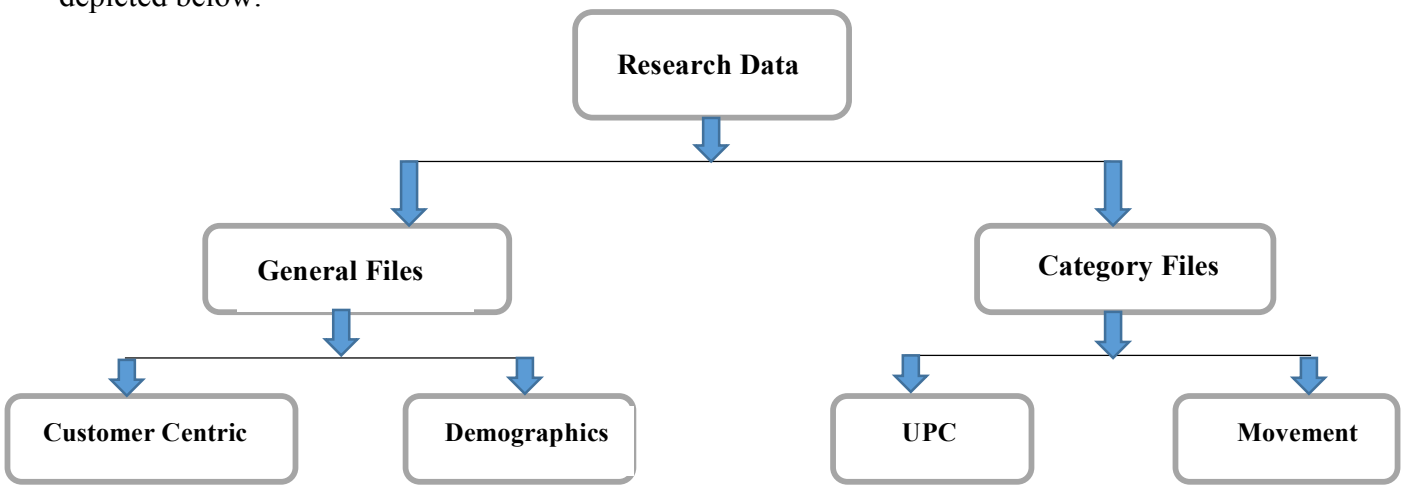
d. *Demographic projection issue*: The success of sales projection relies greatly on the customers, economic health and living habits of its targeted customers. DFF needs to study the intended customer base in different areas, for setting in line expectations towards attracting a large number of customers for its intended growth.

Product sales within a profitable margin is a very critical issue for retail chains like DFF and projection for the same is made harder by the fact that sales of a particular product is affected by a number of factors like seasonal changes, competitor marketing and correct strategy for the sales including price reductions and promotions. The forecasting is dependent on a number of related factors where a delta conditional change can overturn the overall projection directly converting the profits to losses for the retail chain. This group project aims at minutely studying all the artifacts and critical attributes involved in DFF's sales towards devising better realistic business forecasts.

2. Details about the Data

a. Understanding of the Data:

The data for the project has been collected from the James M. Klits Center, University of Chicago Booth School of Business. This data corresponds to approximately 9 years of store level data over more than 3500 UPCs. The collected data has been distributed in a number related files and is present in multiple file formats including csv, html and text. The whole data volume corresponds to 4.76 GB and needs huge manipulations to convert this dirty data to meaningful impacts. We can divide the whole data set into two categories with each further divided into two categories as depicted below:



Few Insights into the data can be summarized as below:

1. The data represents the day to day transactions across the multiple stores of DFF retails store chain and consists of descriptive information about the purchase trends of the consumer.
2. Further the customer data has been broken down into multiple customer categories in order to understand the type of customers DFF is dealing with. This demographic information has been extracted with respect to 1990 population census and is critical for the retail store giving the information over one of key drivers of the business – customers.

3. The provided data is dirty in nature and hence needs to be worked upon to extract the meaningful data out of the same.

4. On a high level, the following major entities have been identified out of the different files:
   a) Store
   b) Customer

c) Coupons
d) Product
e) Category

5. Below is the detailed description for each of the above represented data files:

    a. Customer Centric Files: Information on store traffic and coupon usage. Below are few of the key information extracted from this file:
        i. This files contains information about number of customers visiting the store and purchasing something.
        ii. The file has the mentioned data as per the divided categories of Dairy, Meat, Grocery, Frozen, Beer, etc. as the food category.
        iii. For all the food product category mentioned, the sales done in dollars and coupons redeemed for each of the store for a particular week is also present.

    b. Demographics: Detailed information on the people statistics visiting the stores. Few of the key findings can be summarized as below:
        i. The file gives the category of people visiting each store for a particular week and also mentions the total visits.
        ii. The people visiting have been categorized upon a multiple factors including their age, educational level, household size, working profiles, marital status, etc.
        iii. A majority of the information in this file has been provided in the form of percentages instead of any exact figures.

    c. UPC: Universal Product Code for the products.
    d. Movement files: Weekly sales data giving the information of the product quantities sold and associated profit or loss with each of them.

6. Each of the files contains huge data both in terms of records count and attribute count. These file attributes can be subdivided into a number of entities and hence can be divided into multiple relations while projecting the data into warehouse. Hence, while designing the dimension model, each of these relations should be carefully marked as dimension or fact table, giving rise to a proper absolute star structure.

b. Describing Metadata for all the OLTP files:

This section explains the significance and meaning of each attribute in the data set files.

CCOUNT	
Column	**Description**
STORE	Store Code
DATE	Date of the Observation Week Number
GROCERY	Grocery sales in $
DAIRY	dairy sales in $
FROZEN	frozen sales in $
BOTTLE	bottle sales in $
MVPCLUB	mvp sales in $
GROCCOUP	Grocery coupons redeemed in $
MEAT	meat sales in $
MEATFROZ	frozen meat sales in $
MEATCOUP	Meat coupons redeemed in $
FISH	fish sales in $
FISHCOUP	Fish coupons redeemed in $
PROMO	promotion sales in $
PROMCOUP	Promotion coupons redeemed in $
PRODUCE	produce sales in $
BULK	bulk sales in $
SALADBAR	salad sales in $
PRODCOUP	Produce coupons redeemed in $
BULKCOUP	Bulk coupons redeemed in $
SALCOUP	Salad coupons redeemed in $
FLORAL	floral sales in $
FLORCOUP	Floral coupons redeemed in $
DELI	deli sales in $
DELISELF	deli self sales in $
DELIEXPR	deli express sales in $
CONVFOOD	conventional sales in $
CHEESE	cheese sales in $
DELICOUP	Deli coupons redeemed in $
BAKERY	bakery sales in $
PHARMACY	pharmacy sales in $
PHARCOUP	Pharmacy coupons redeemed in $
GM	GM sales in $

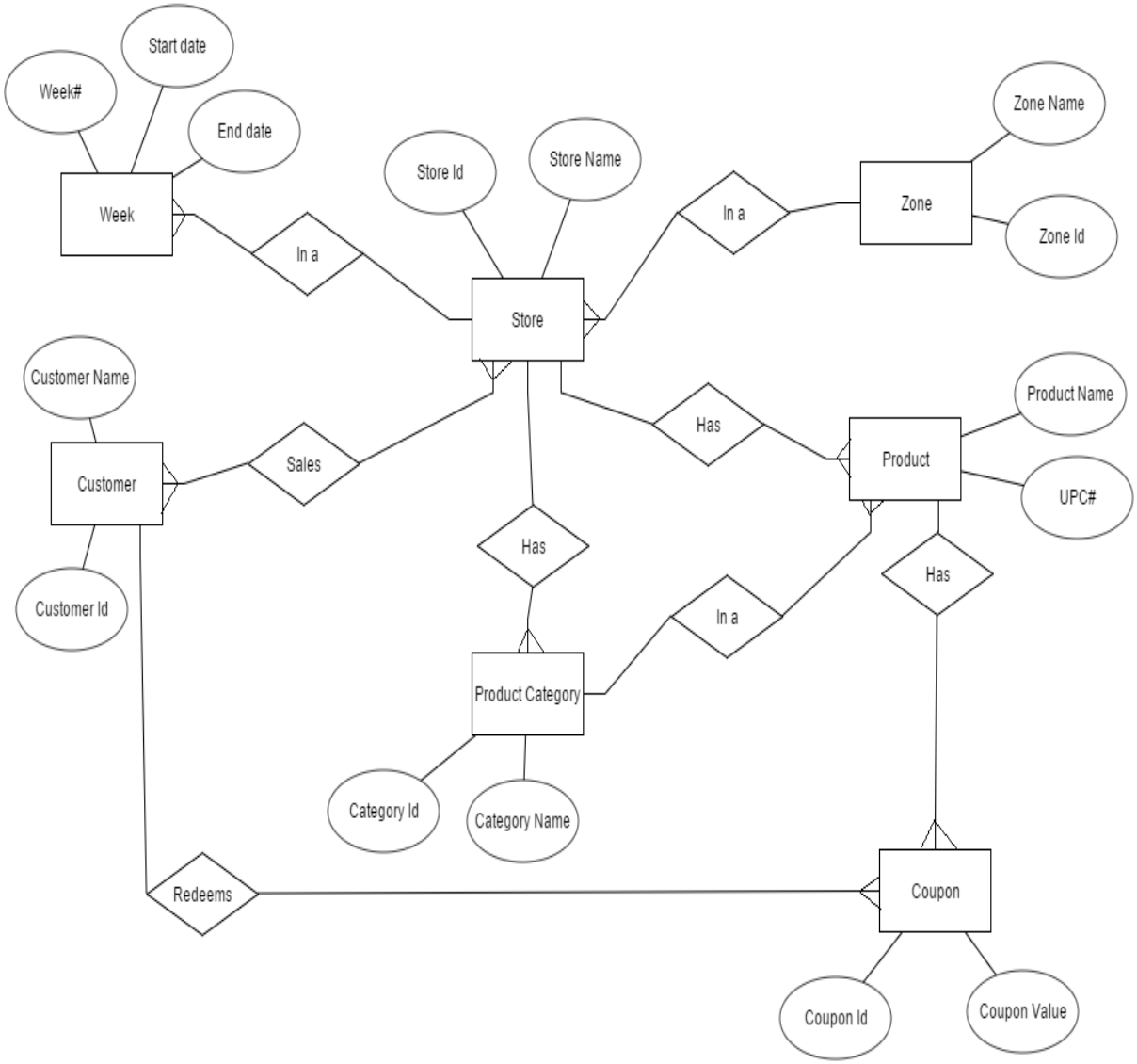
JEWELRY	jewelry sales in $
COSMETIC	cosmetic sales in $
HABA	Health and beauty aids sales in $
GMCOUP	General coupons redeemed in $
CAMERA	camera sales in $
PHOTOFIN	photo sales in $
VIDEO	video sales in $
VIDEOREN	video rental sales in $
VIDCOUP	Video coupons redeemed in $
BEER	beer sales in $
WINE	wine sales in $
SPIRITS	spirit sales in $
MISCSCP	Misc. coupon sales in $
MANCOUP	Manufacturer coupons redeemed in $
CUSTCOUN	number of customers
FTGCHIN	food to go Chinese sales in $
FTGCCOUP	Food to go Chinese coupons redeemed in $
FTGITAL	food to go Italian sales in $
FTGICOUP	Food to go Italian coupons redeemed in $
DAIRCOUP	Dairy coupons redeemed in $
FROZCOUP	Frozen coupons redeemed in $
HABACOUP	Health and beauty aids coupons redeemed in $
PHOTCOUP	Photo coupons redeemed in $
COSMCOUP	Cosmetics coupons redeemed in $
SSDELICP	self-service deli sales in $
BAKCOUP	Bakery coupons redeemed in $
LIQCOUP	Liquor coupons redeemed in $
WEEK	week number

UPC	
Column	**Description**
UPC	UPC number
COM_CODE	Dominick's commodity code
NITEM	Dominick's item code
DESCRIP	Product name
SIZE	Product size
CASE	Number of items in case

Movement	
Column	**Description**
UPC	UPC number
STORE	store number
WEEK	week number
MOVEMENT	number of units sold
PRICE	retail price
QTY	number of items bundled
PROFIT	gross margin
SALE	sale code (B,C,S)
OK	1 for valid, 0 for trash

Demographics	
Column	**Description**
AGE9	Depict population under age 9
AGE60	Depict population over age 60
POVERTY	Percentage of population with income under $15,000
UNEMP	% of Unemployed
HVALMEAN	Mean Household Value (imprecise value)
HVAL200	Percentage Households with Value over $200,000
HVAL150	Percentage Households with Value over $150,000
HSIZEAVG	Average Household Size
INCOME	Log of Median Income
ETHNIC	Percentage Blacks and Hispanics
EDUCATION	Percentage College Graduates
DENSITY	Trading Area in Sq. Miles per Capita
SINHOUSE	Percentage detached houses

c. E-R diagram to show the relationship among the major entities on a higher level:



3. Domain understanding

With our research from a number of resources and papers related to retail domain over cost-quality benefit analysis and related reporting, following below are the major extracts defined:

- Analyzing the data according to sales in different areas. Companies in retail business fixate on an area where they can earn more profit and more business. It is certainly very arduous for the retail company to understand that which particular product is in demand and in which area. So whenever a company launch a product in market, it is not sure that the product will be prosperous in one part of town and not in other. Analyzing the data is the most sizably voluminous quandary for retail business because of its prolific data and numerous optimization quandaries such as optimal prices, discounts, recommendations, and stock levels.
- Another quandary faced by retail business is how to increment sales in penuriousness affected areas. The retail business quandary systematically depends on the penuriousness of communities. There are lots of ways by which retailers endeavor to magnetize customers but all the ways must result in business profit withal not having stores in those particular areas will affect the goodwill of an organization. In order to handle these both quandary simultaneously companies are putting efforts in both ways managerial and technology.
- Astronomically immense retailer companies often introduce schemes of coupons and customer adhesion benefits these are prosperous to some elongate but again the quandary arises that in which area what kind of business is prosperous and what are the authoritative ordinances of people in that particular area. Example sales of non-vegetarian aliment is not being remuneratively lucrative in the area where mostly vegetarian people resides even if you offer any scheme on top of that.
- It is very paramount to examine which store will be prosperous at which location and what type of items will engender revenue. Keeping none authoritatively mandating items in business is an indirect loss to the store.
- Another big challenge for retailers is the business flexibility habituating to an evolving market requires a business to introduce more flexibility into its business processes. For instance, the only way to satiate customer demands for Omni-channel accommodations is for retailers to be flexible in the way that they interact with their customers to provide a consistent experience.
- Customer accommodation is additionally a consequential aspect of running a prosperous business in the field of retailing. Customers are the base of business and to gratify them is the primary motive of the company. Customer accommodation and in-store experience play a massive part in consumer postures to showrooming. Most shoppers only resort to mobile shopping when there is a stock shortage or they receive poor accommodation.

- Retailers wanting to bring in more customers and retain those they already have are tapping into allegiance programs incentives that reward shoppers who make reiterate visits. In this day and age of retail, staunchness is an immensely colossal thing.
- It is no longer adequate for a business to operate utilizing a single channel to market. Now a day's shoppers believed it was paramount to be able to purchase from a retailer utilizing different channels. However, for the forward-cerebrating business, implementation of a system which cumulates all retail channels is a better investment. Kenned as Omni-channel retailing, the conception is to provide a consistent customer experience utilizing a single or integrated multi-channel software platform to power transactions, stock and other internal business processes transparently.
- Retailers have never had more information about their customers, but many are struggling to decipher what it all denotes. In a recent survey by Accenture, 72 percent of 258 North American business bellwethers verbally expressed they orchestrate to spend more on analytics. In many cases that signifies buying software and hiring analysts to make sense of streams of data emanating from mobile contrivances and other electronics.
- Diminutive retailers customarily specialize in a niche area of product. Albeit this can be a vigor when consumers are probing for product expertise, diminutive retailers are dependent on a constrained range of products to make sales. The products offered by boutique shops are customarily not essentialities; consequently, these stores might experience a slump in sales during a time of economic recession.
- Lack of product diversity- Most immensely colossal retailers, which have the goal of engendering a consistent customer experience across all of their stores, will opiate to stock only products with significantly astronomically immense sales. If more incipient products propagate expeditiously, an astronomically immense chain's reluctance to stock them might cost them sales.
- Large retails are usually too big to manage. Albeit astronomically immense retailers benefit from economies of scale, which designates they are able to purchase and sell high amounts of product at a more diminutive per unit cost than independent retailers, companies can expand too expeditiously. Companies can become out of touch with the everyday consumer and fail to provide them with the products they opiate. A chain predicated on the concept of low prices might do well during an economic recession, but suddenly find its sales decline when consumers have more maxima and prefer to spend it on quality or unique goods offered by independents.

4. Business Questions

Priority: High

Case#	Business Question
1	Evaluating the store performance with respect to all the other stores at DFF
2	Identify the most concerned areas of losses for each shop for Dominick's Fine Food (DFF) week-wise.
3	Generate the report depicting the effect of coupon introduction for increasing the customer count in the store.
4	Identifying the contribution of each product category in the overall sales for DFF
5	Identifying price elasticity of demand for a particular item, w.r.t. income levels

Priority: Medium

Case#	Business Question
6	Identifying the contribution of each store/zone on the overall sales of DFF
7	How sales of an item of varying packet size is impacted by price change
8	Forecasting profitability and sales of an item for measured changes in item pricing

Priority: Low

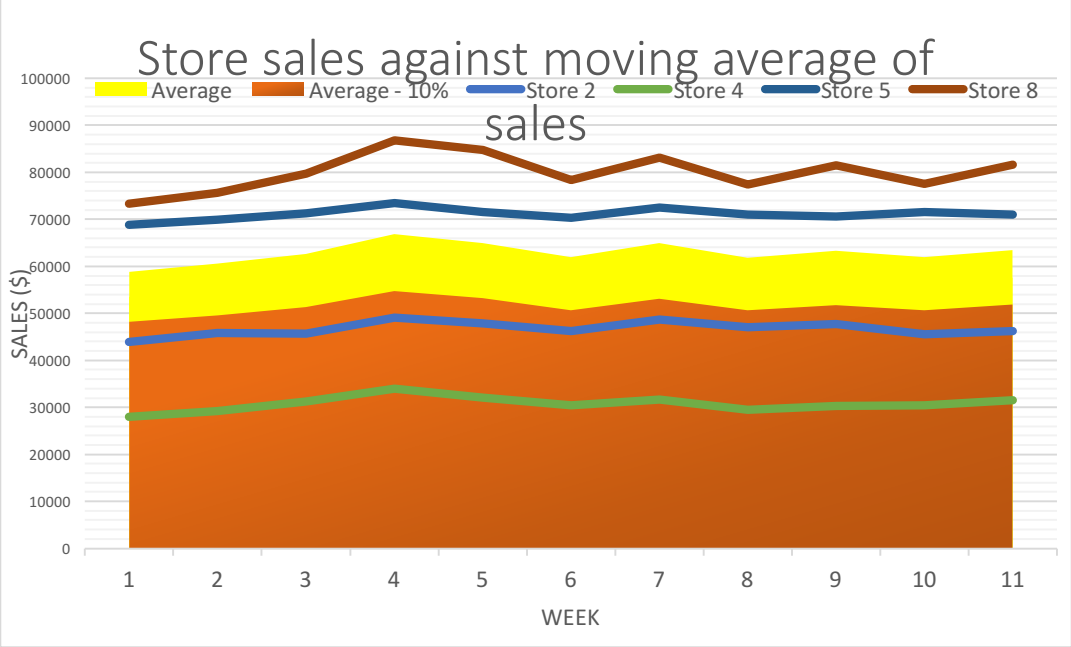
Case#	Business Question
9	Generate the report depicting the sales summary for different types of product for different stores.
10	Generate the report depicting the effect of coupon introduction for increasing the sales in poverty affected areas.

Business Questions Explained:

BQ.1 Evaluating the store performance with respect to all the other stores at DFF

The question here is about comparing the weekly sales of a store to that of the sales for rest of the stores. We can represent the sale for all the stores at DFF by a moving average of all the sales figures and then compare that to the store in question. The question can be further expanded to compare annual sales trends if needed.

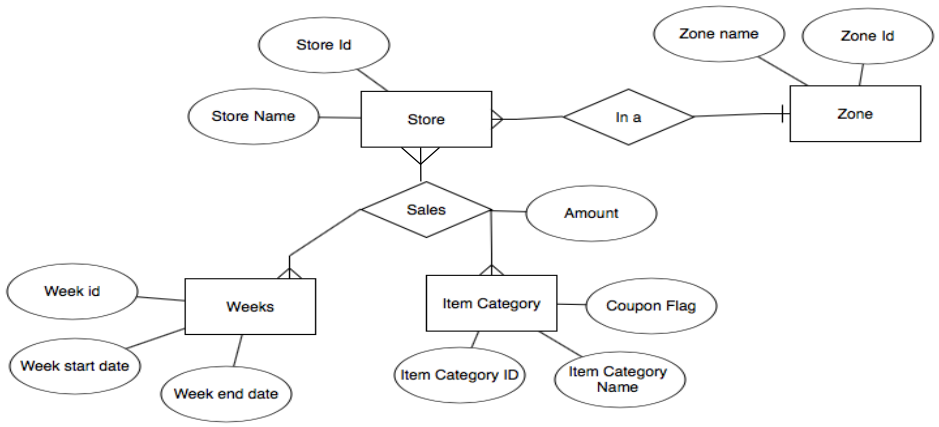
The graph below shows the trend line representation for stores 2, 4,5 ,8 over a ten-week period. A tolerance limit of 10 per cent above and below the average trend is decided and based on that, the graph is separated into 3 regions. The RED region at bottom the area with sales under 'average – 10%'. This represents the lower sales region. The YELLOW region is the region between 'average – 10%' and 'average +10%' sales trend lines. If the trend line of a store falls in this region, then the store has an average performance. The WHITE region at the top of the graph represents the stores with sales above 'average + 10%'. This region can be used to identify the high performing stores.



The graph here indicates that store 2 and store 4 are the lower performing stores. That is, they perform more than 10 per cent lesser than the average performance. The graph doesn't have any trends in the YELLOW region. The WHITE region, which is the higher performance region has the stores 5 and 8.

This information can have many business applications. The results can clearly identify the stores that needed to be worked on. Such stores can be kept under closer supervision and also paired with the higher performance stores of the region for improved learning. The results can be leveraged to target stores for incentives for month on month improvements. Apart from that, promotions and advertising decisions can be greatly influenced by the trends. The results can also be applied in staffing and training decisions for stores.

The ER diagram and the aggregated dataset is quite similar to that in one of the other business questions for identifying the contribution of a store in the overall sales, and hence the same data marts can be used to answer the two question. They are again explained as below:



ER Diagram

Week --> Stores	Average of Sales
17	53518.86
2	43931.3
4	28007.79
5	68859.23
8	73277.14
18	55117.83
2	45765.67
4	29183.43
5	69902.34
8	75619.88
19	56983.84
2	45724.35
4	31297.44
5	71248.15
8	79665.4

Pivot table data derived from CCOUNT dataset

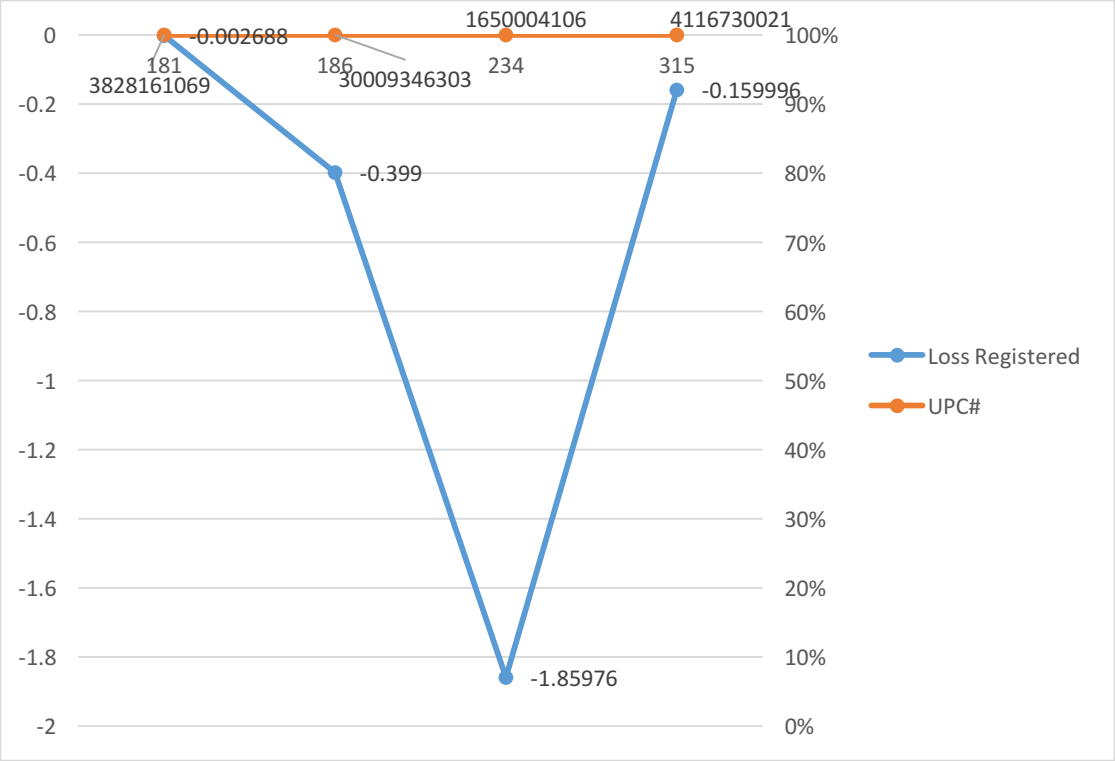
The entity relationship diagram has 4 entities Store entity contains the details for each of the store. It contains the address of the store, demographic details of the store, if they may be needed and other attributes specific to a single store. The zone entity contains details for all the zones. It can contain information like the name of the zone, who is the manager for the zone and certain set of attributes like tax rate etc. that may be common to the entire zone. The week entity has time related information. It has a 'WEEK_ID' and the start date and end date for the week. We may also store information on public holidays that occurred on that day of the week. The item category entity stores information about all the categories of items that are sold. These categories include bakery, beer, bottle, bulk among others as found in the 'CCOUNT' dataset. It also contains information about promotional coupons, which is indicated by the coupon flag. The major relationship in this ERD is the sales relationship which is a many to many relationships among store, weeks and item categories. It has attributes for storing the sales amount and sales dates along with the foreign keys for the entity tables.

BQ2 Identify the most concerned areas of losses for each shop for Dominick's Fine Food (DFF) week-wise.

The major driving force for any business in this world is profit and loss statement. As a retail store chain, DFF also needs to check for the sales volume for each of its store and associated business gains from the same. The business question focuses mainly on loss incurring areas as the business directly come up the summary report of the store week-wise where it registered loss instead of the otherwise profit running trends.

The business objective for this question is to find out the concerning areas of the business with listing down the products upon which DFF is incurring losses. The idea behind to compare this loss with the week is to identify for any related reasons or external factors which led the product to register loss for that particular week. With the data achieved from this report, DFF can actually drilldown to the store authorities providing them with a statistical report about the registered losses and can further brainstorm upon the causes behind it. These brainstorming sessions can actually help the organization understand the factors which are leading it to refrain from the targeted profits. Once analyzed and curbed, these losses can be substantially converted into opportunities of growth for DFF, leading to a direct profit result out of the reporting.

A graphical representation for this report can be show as below:



The above chart clearly depicts the loss registered for each product corresponding to the week. DFF can look up the data to find any external condition due to which these particular product

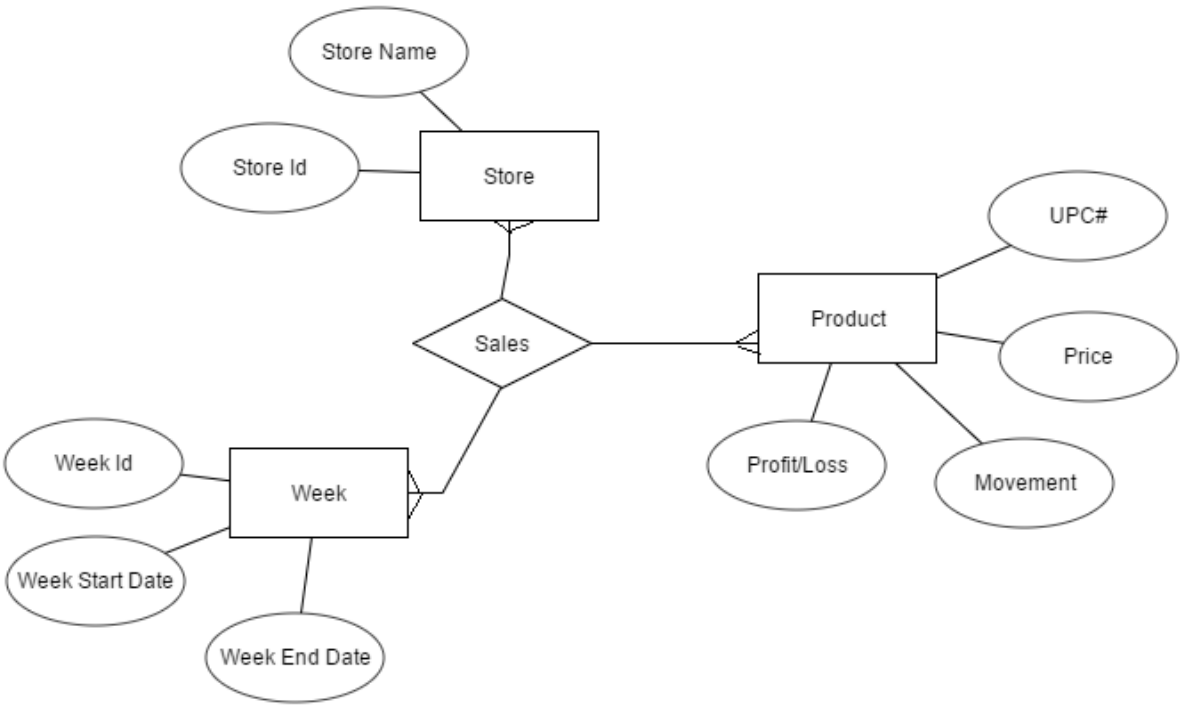
registered loss in their sales for that particular week. Based upon the analysis, the store can be careful in selling the same products under such conditions over the next phase. Below is a sample set of excel data *extracted and extensively manipulated* from 'Movement file' to calculate the losses. Please note that the formula to calculate the sales used is as has been given in the Database manual:

Sales= Price*Move / Qty

Source Data for Chart:

Week Observed	Loss Registered	UPC#
181	-0.002688	3828161069
186	-0.399	30009346303
234	-1.85976	1650004106
315	-0.159996	4116730021

Entity Relationship Diagram:



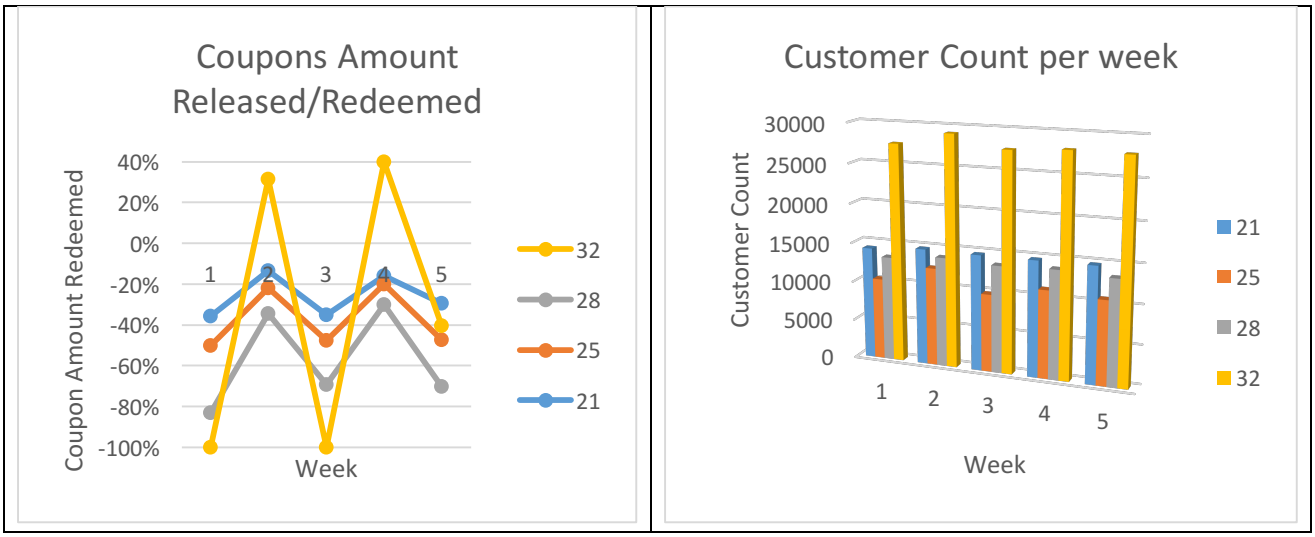
The above diagram represents the entity relationship diagram for this business scenario. The store entity contains the store id and name as important attributes, and is related with week and product entities through 'Sales' relationship. The relationship defines the common linkage between stores selling a product in a particular week. The product has its attributes 'Profit/Loss', UPC#, Price and Movement which is finally utilized in depicting the loss value associated with the store for a particular week.

BQ 3 Generate the report depicting the effect of coupon introduction for increasing the customer count in the store.

DFF has introduced coupon systems for attracting customers and has been consistently working towards being the absolute choice of the customers. The introduction of coupon system is a logical extension of the same policy and this business scenario aims at checking the feasibility of the coupon system for DFF.

The business objective for this question is to find out if the coupons introduction for the store is actually beneficial for the store or not. This business report aims at projecting the worth of coupon system as there is a net value associate with each of the coupon. If the customer purchase is less than the coupon value released by the store, it would pose a threat to the functioning of the store. This question aims at checking the worth of coupon system for DFF so as to further strategize the company's policy over investing more time and energy towards coupon system.

A graphical representation for this report can be show as below:



Please note that the above chart is based on a small sample volume and hence, doesn't represent the extensive trends it will generate once subjected to a higher data set. The data sets used for this projection has been provided below. The coupon amount when has a negative amount, represents that the coupon has been released from the company, but not yet deemed by the customer. A positive value for the same, in turn, represents that the customer has actually redeemed the same. The trend clearly depicts that releasing the coupon to the customers actually increases the customer count in the next week (which is in sync with retail industry's policy that coupons are generally redeemed in the next visit only). The data should present a more exhaustive view when presented with a higher meaningful data sets.

Source Data for Chart:

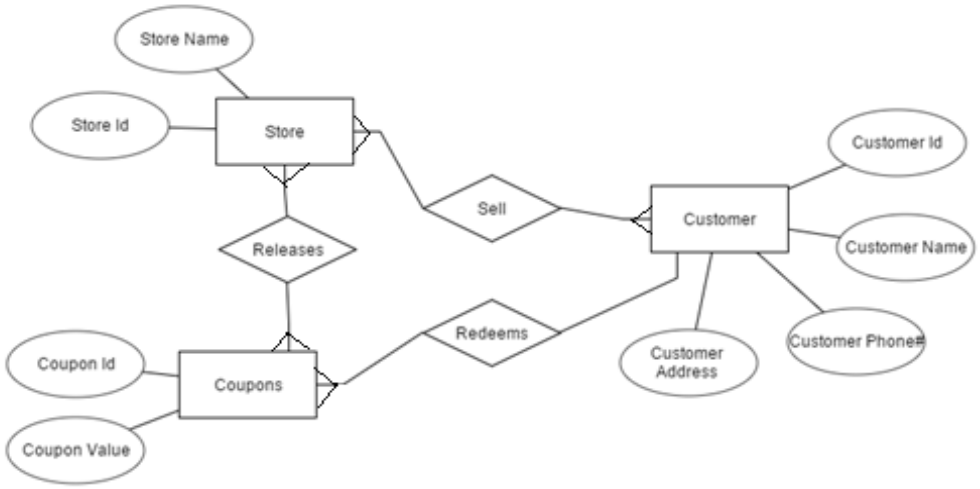
Coupons Redeemed Value per week for each store number:

Week/ Store Number	21	25	28	32
1	-6070.74	-2473.85	-5605.75	-2874.07
2	-648.77	-411.58	-594.97	3166.71
3	-29363.05	-10363.89	-17860.01	-25720.1
4	-974.9	-238.05	-615.42	4241.26
5	-1676.9	-1029.84	-1308.13	1704.03

Customer visited to the different store for each week:

Week/ Store Number	21	25	28	32
1	14229	10356	13270	27568
2	14751	12439	13899	29157
3	14680	9866	13581	27634
4	14707	11178	13810	28023
5	14757	10737	13454	27922

Entity Relationship Diagram:

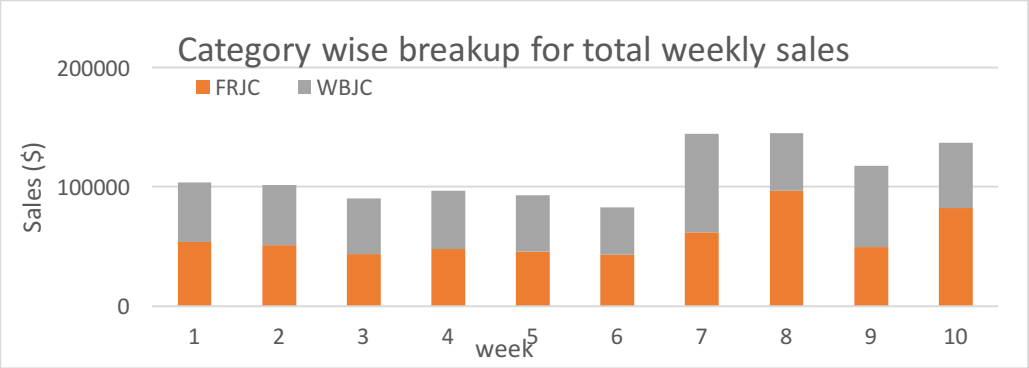


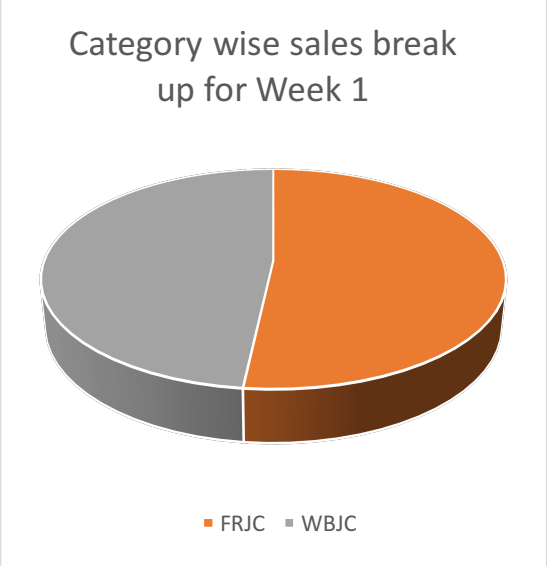
The above diagram represents the entity relationship diagram for this business scenario. The store sells the items to the customers and also releases the coupons. There is another relationship between customer and coupon defined as 'Redeem' which states the redemption option by the customer of the coupon. In a nutshell, there is a triangular relationship between the three critical entities Store, Coupons and Customer for this business case. The coupon value attribute is used to find the total coupon released to the customers. To mention, this relationship uses "CCOUNT" file to facilitate the usage of all the required attributes required for this business case.

BQ 4 Identifying the contribution of each product category in the overall sales for DFF

The products at DFF are divided into 30 categories with daily data of movement available for each item of each category. The question here is to realize the contribution of each product category towards making up the total sales of DFF on a weekly basis. This can also be rolled up to visualize the contribution of each product category to the overall sales for DFF.

A simple graphical representation of the desired visualization that might answer this question is as below:

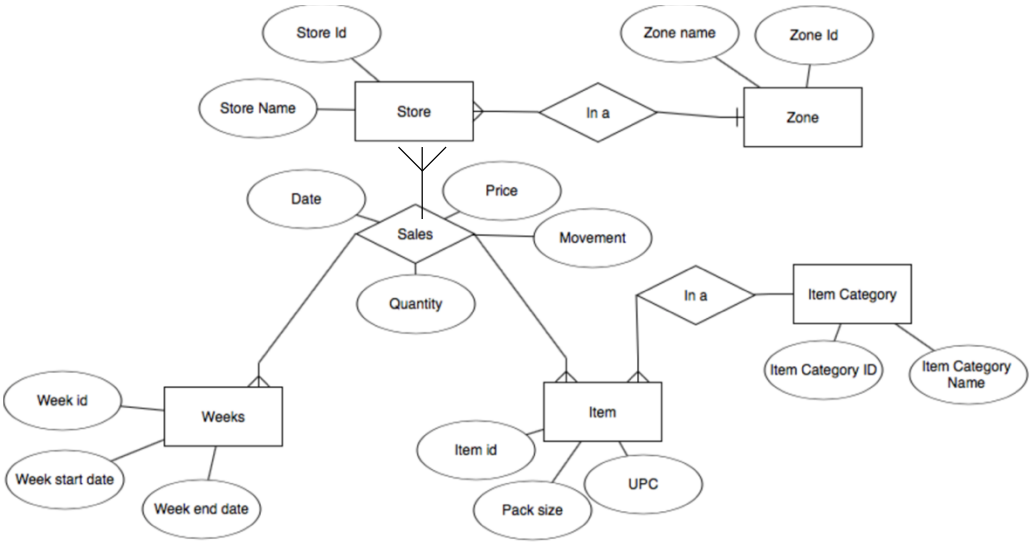




Here, for the sake of simplicity, we have assumed that there are only two categories FRJC Frozen juice and WBJC Bottled juice that are making up the total sales for DFF. The real plot however, will have 30 such categories rather than just one. The bar graph here shows the sales contribution of the two categories over a ten-week period.

The pie chart on the left gives a more accurate idea of the contribution for each category, for week 1.

Such a visualization will have a wide effect on many business decisions of the organization. Higher selling categories and lower selling categories can be identified and can be a major factor in several decisions. Warehouse space can be appropriately allocated to the items. It can affect several inventory level management tasks as well. Higher contributing items can be ordered in higher quantity as they are seen to be moving fast, where as they organization can delay the restocking of lower selling items if the stock goes low and enough storage space is not available. Besides, the information can also help in category promotion and advertising. The ER diagram for this problem can be as below. Most of the data can be populated for the movement dataset, which stores daily sales transactions for each product, as well as UPC data sets, which stores data of all items in a category.



ER Diagram

Sum of Sales	Column Labels	
Week	FRJC	WBJC
17	53734.45	50079.07
18	50802.37	50565.95
19	43659.98	46382.05
20	47854.7	48613.59

Data for graph

The ER diagram here has five entities and three relationships. Store entity holds the details for all the stores there are at DFF. It contains information like STORE_ID, store address, name of store and all the other attributes that are unique for a particular store. It is related to the zone entity by a 'in a' relationship. Here, the zone entity stores details specific to a zone like name of zones, tax rates for zone, name of manager for zones etc. The 'is a' relationship stores information on which store belongs to which zone.

The items entity stores information for each individual item. It has the item code, item size, item name, manufacturer details and other attributes specific to individual items. It is connected by a 'in a' relationship to the category entity. This entity stores details specific to categories like name of category etc. Several items of similar types are bundled in a category.

The 'weeks' entity stores time related information. It stores the 'WEEK_ID', the start date and end date for the week and any special events that might have occurred during the week. The 'sales' relationship is a ternary relationship between items, store and week. It stores the detail specific to each individual day of sale. It contains the sale amount, quantity moved as well as the price for which the item was sold. This relationship contains the numerical data necessary for the calculations for this question.

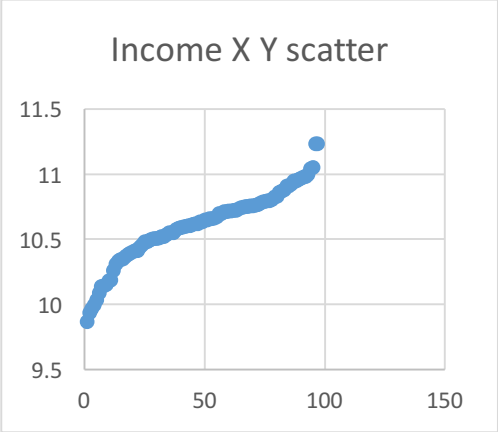
A sample data has been derived above as shown in the table. It is derived from the movement table for FRJC and WBJC by filtering it for weeks 17 to 26, which represent the first weeks of 1991. The sales for each category are aggregated on a weekly basis and then then used as basis to answer the business question. The same technique can be emulated on a larger scale for all the categories, for every week, while answering the actual business question for the project.

BQ 5 Identifying price elasticity of demand for a particular item, w.r.t. income levels

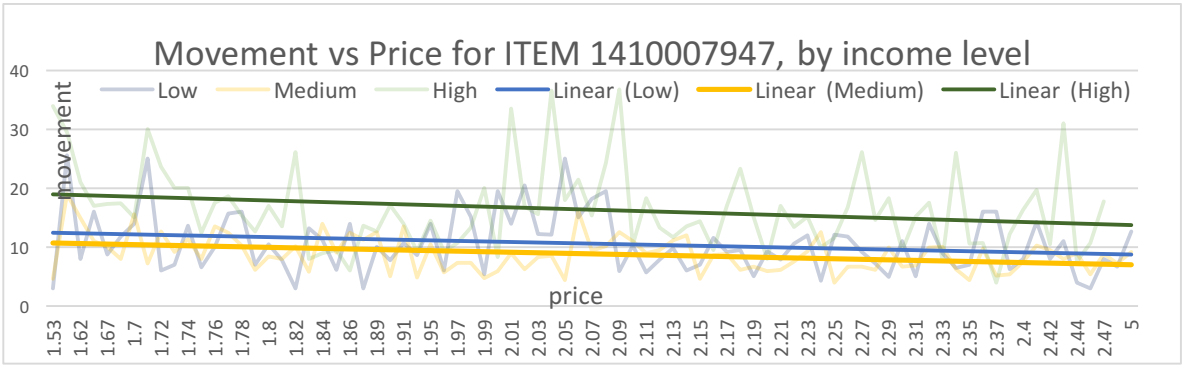
"Price elasticity of demand (PED or E_d) is a measure used in economics to show the responsiveness, or elasticity, of the quantity demanded of a good or service to a change in its price, ceteris paribus. More precisely, it gives the percentage change in quantity demanded in response to a one percent change in price", Wikipedia.com (2016). The question here focuses to visualize the effect of price change on the movement of a given item, with all the other factors constant. The analysis will be more specific and separated on level of income, at the store location where sales occur. We intend to see the difference in E_d for 'low', 'medium' and 'high' income regions for that given item.

Such an analysis will be a helpful tool in predicting the future trends in movement with changing prices. We may identify certain goods that respond well to price change while some may not. The new prices can be arranged accordingly for products. Separating the analysis by income levels will allow DFF to change prices conditionally according to the income parameters of the region. The analysis can help analyze the elasticity for individual items. Luxury items are expected to have lesser elasticity in higher income areas than in lower income areas, that is that their sale is not affected as much by increased prices in higher income areas than that in lower income area stores where the sales is expected to drop. Daily items like food products and toiletry may or may not show such changes. Such an analysis will help DFF make more informed business decisions and help make item promotion and coupon offers more targeted and planned.

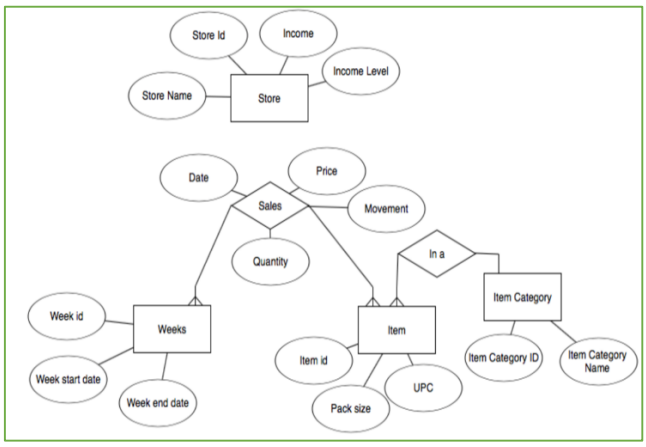
Firstly, to identify and separate the stores into three levels of income: 'low', 'medium' and 'high'. The income for each store, obtained from the store demographic dataset is plotted on a XY scatter and as as visible, there is a stark increase in slope at around 10.4 and 10.8 and we shall assume these to be the separating points as below:
LOW: Income < 10.4
MEDIUM: 10.4 <= Income <= 10.8
HIGH: Income > 10.8



Now, choosing item with UPC code 1410007947 P Farm mint Milano from the 'cookie', as it had the maximum deviation in prices, among the other items of the same category, to make the analysis more visible. The graph can be plotted as:



Here, the graph shows 3 separate trend lines for movement variation for low, medium and high income zones. The trend lines are smoothened out by a liner line as shown in the graph. The slope of this line will be the price elasticity of demand for this item, for the given income groups. Here, as it is evident from the graph, cookies being a fast moving good, shows lesser variation in elasticity for different income groups. But had we chosen a luxury item, the difference in slope would be more evident.



The ER diagram to visualize the world view for this problem is as above. The diagram is in many ways similar to the ER for most other business questions and it focuses around the sales relationship. This relationship stores the values for the price, which can be used to obtain the price variations, movements, which can be used for the Y axis of the graph and other date related parameters. It is worth nothing that a date attribute is not needed in this particular business scenario but it is there nevertheless in case it finds any applications in the future. There are four entities in the diagram. The store entity contains the store specific details. The important attribute in this table is the income demographic attribute which is used to derived the 'income level' attribute as stated previously. The sales data is categorized based on this attribute. Other entities include those for items and item categories. Item category entity is included in case a roll up is needed to find the elasticity at the category level rather than at an individual item level. The week entity stores time related data.
The excel tables needed for plotting the chart is as shown to the right. The data is obtained from the movement table for cookies and an addition column for 'income level' is generated based on the store. This data is categorized by the income level and aggregated for average movement based on the prices. The resulting pivot table generated is as shown in the adjoin tables.

Store Area Income Level	High
Price	Average of Move
1.58	34
1.59	29.5
1.65	21
1.66	17
1.67	17.34920635

Store Area Income Level	Medium
Price	Average of Move
1.53	4.625
1.57	19
1.59	14.95918367
1.6	11
1.61	10

Store Area Income Level	Low
Price	Average of Move
1.53	3
1.59	25.625
1.62	8
1.66	16.07692308

BQ 6 Identifying the contribution of each store/zone on the overall sales of DFF

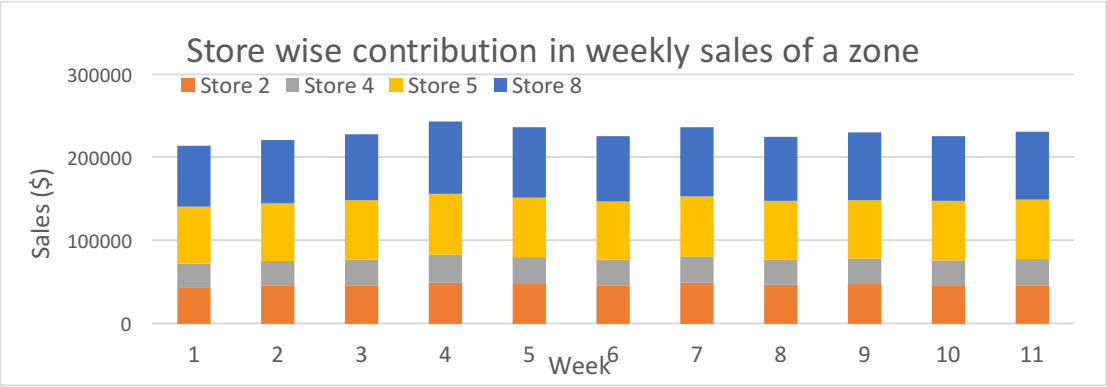
The question focuses on comparing the weekly sales of a particular store against the total sales of that region/zone for DFF. Similarly, it can be rolled up to compare the weekly sales of each zone against the total sales for all regions or alternatively, the sales of a store against overall sales of DFF. The roll up can also be on the time axis and can be used to represent annual sales comparisons as well.

The business implication of such an analysis is that it will help to identify things like stores and zones with lower sales, or the increase in sales of some particular zones at a particular point of time. If a store is found to contribute a lesser to the overall sales, then that zone/store can be individually dealt with after considering all the factors that might affect it. Apart from that, the zonal performance can be a useful tool for giving incentives to the higher performing zones.

This information can also be crucial to staffing decisions for DFF, an outlet which is not making a significant contribution might not need a fully staffed store and human resources can be moved around accordingly.

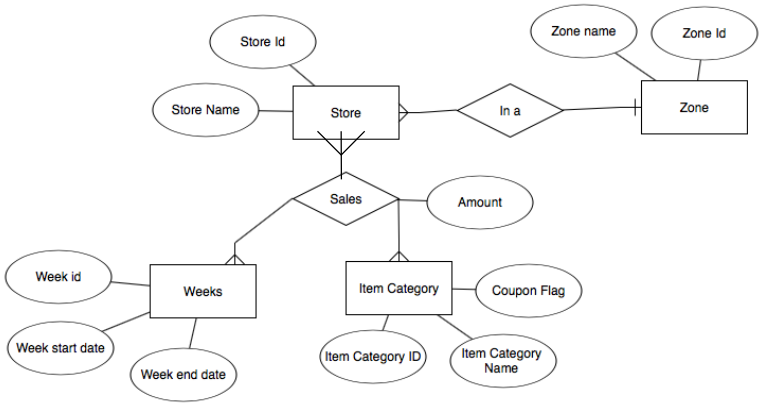
A simple graphical representation which answer this as below. We shall assume that the stores identified as store 2, store 4, store 5, store 8 are in one zone, and the chart shows how each store is contributing to the sales for the zone.

It is apparent here that store 4 isn't performing as well, and efforts should be made to identify the causes.



This chart can be generated from data aggregated as below. This data is derived from the 'ccount' excel sheet by calculating the sales for each week for a given set of stores. Pivot tables and filers have been applied on the 'ccount' dataset to generate the aggregated data.

Besides, the world view of the problem at hand can be represented by as an entity relationship diagram as shown below:



ER Diagram

Week --> Store	Sum of Sales
17	1498528.19
2	307519.08
4	196054.52
5	482014.58
8	512940.01
18	1543299.26
2	320359.72
4	204284
5	489316.39
8	529339.15
19	1595547.44
2	320070.47
4	219082.07
5	498737.07
8	557657.83

Source Data for Chart

The entity relationship diagram above has 4 entities. Store entity contains the details for each of the store. It contains the address of the store, demographic details of the store, if they may be needed and other attributes specific to a single store.
The zone entity contains details for all the zones. It can contain information like the name of the zone, who is the manager for the zone and certain set of attributes like tax rate etc. that may be common to the entire zone.
The week entity has time related information. It has a 'week_id' and the start date and end date for the week. We may also store information on public holidays that occurred on that day of the week.
The item category entity stores information about all the categories of items that are sold. These categories include bakery, beer, bottle, bulk among others as found in the 'ccount' dataset. It also contains information about promotional coupons, which is indicated by the coupon flag.

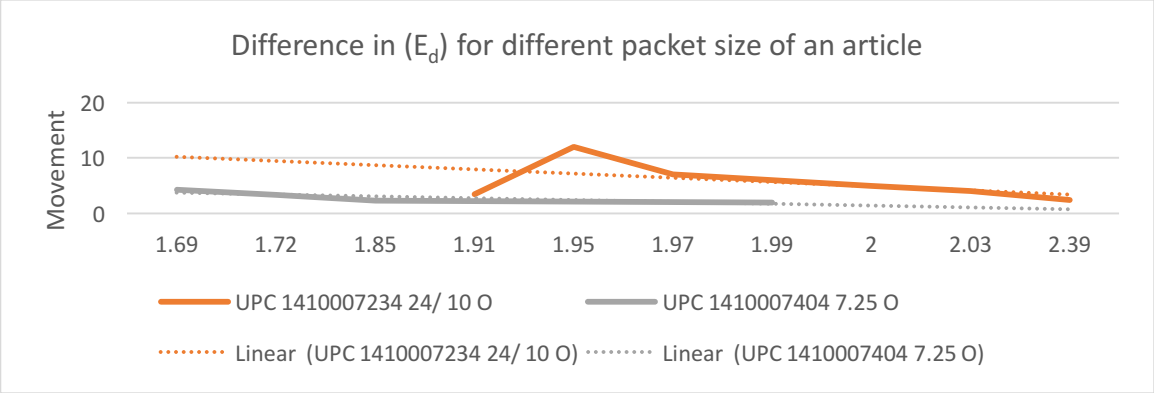
The major relationship in this ERD is the sales relationship which is a many to many relationship among store, weeks and item categories. It has attributes for storing the sales amount and sales dates along with the foreign keys for the entity tables.

BQ 7 How sales of an item of varying packet size is impacted by price change

The question here is to identify the effect of price change on sales for the same item, having different pack sizes. That is to compare the price elasticity of demand among products of varying pack sizes. The price elasticity of demand can be defined as the change in quantity sold per unit change in price for a given item, with all other factors constant. If a graph of sales in units → price is plotted, then the slope of the graph will represent the price elasticity of demand E_d for that product. The E_d for the same item sizes but different pack sizes can be compared as a solution to this business question.

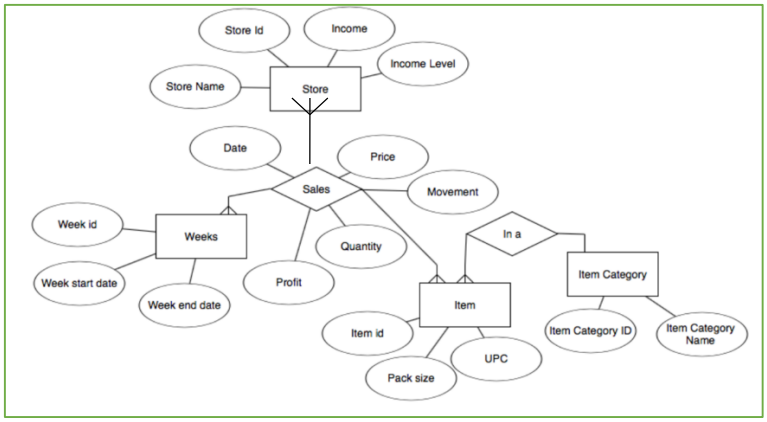
To facilitate this comparison, a graph of average change in movement to the price is plotted for both the pack sizes. A linear smoothing can be applied to the trend line and the slope for different pack sizes can be compared to get an estimate of the effect of price change in sales for each of the sizes.

A sample graph for the product 'almost home oatmeal' for the pack sizes 24/ 10 O (larger) and 7.5 O (smaller) is as below:



As seen in the graph, the trend lines are linearly smoothened and the slope of these lines can be used to estimate the price elasticity of demand. It can be seen that the slope of the orange line which represents the larger pack size is higher than that of the grey line representing the lower pack size. It can be deducted that the larger pack sizes show a more notable decrease in sales on increasing prices than that for the smaller pack sizes. It may also be assumed that the sales shift of the higher pack sizes virtually moves to that of the lower sizes as the price of the item increases.

Several business implications can be derived from this analysis. The sales increase in decrease can be predicted for a particular product of a particular pack size if the price of its other pack sizes are expected to change. Such an analysis can help in stocking and procurement decisions. Also, the knowledge gained can go help the store provide valuable insight for the item manufacturers for regulating the price of their product. The results can also be used for planning promotions for one pack size based on the expected price changes in the other pack sizes.



ER Diagram

The ER diagram for this problem is very similar to the ones for other business questions. The ER diagram is focused around the sales relationship. This relationship stores all the measures necessary for the analysis. It has attributes for storing price, movement, profit on a daily basis. It is a ternary relationship between the entities for stores, items, weeks. The store entity contains details of each store and the unique attributes that are associated with each of the stores. The item entity stores the details for each individual items. The packet size information is stored in this entity and it plays a key role in this business question. Each item is further a part of a category. The week entity has details for the time dimension. The details necessary for answering this business question can be derived from these set of tables.

The pivot table data on the right is used to generate the data for the chart. The data here is primarily obtained from the movement table.

The data is first filtered for the UPC code 1410007234 and the price and movement columns are selected. A pivot table is generated for these two columns with price as the row entity and average of movement as the column entity. The graph for movement → price is then plotted and the trend line is subjected to a linear smoothing. A similar method is carried out for the UPC code 1410007404 and the graph is plotted on the same chart.

The smoothened linear trend lines, when compared with each other form a solution to this business question.

UPC	1410007234

Row Labels	Average of MOVE
1.91	3.495412844
1.95	12
1.97	7
2	5
2.03	4
2.39	2.390011891

UPC	1410007404

Row Labels	Average of MOVE
0	2
1.69	4.253521127
1.72	3.318181818
1.85	2.341463415
1.99	1.983909895

Pivot data for plotting graph

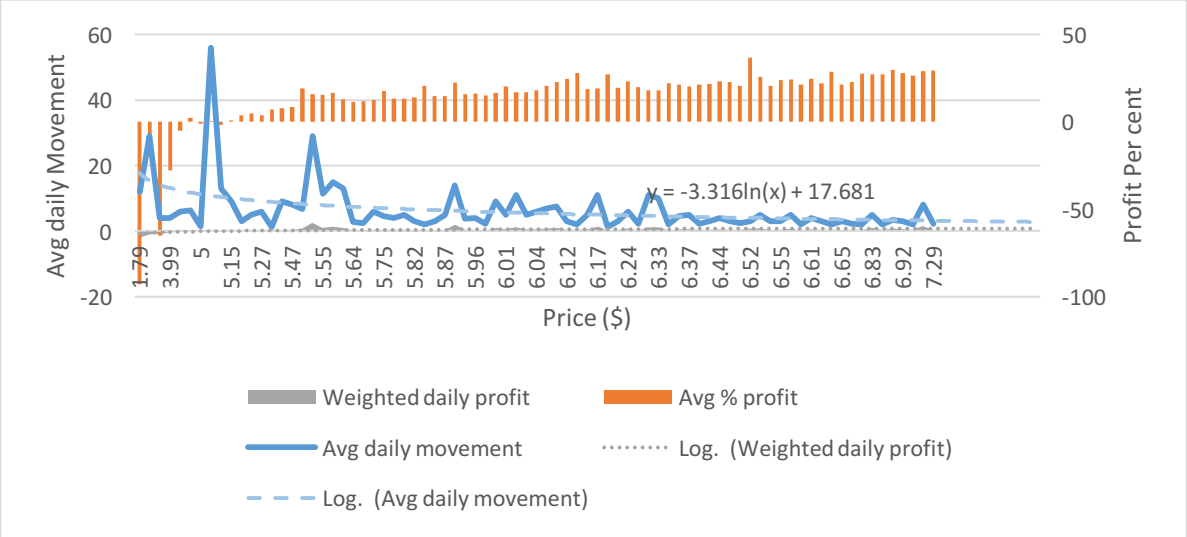
BQ 8 Forecasting profitability and sales of an item for measured changes in item pricing

This question focuses on two things: visualizing the changes in profitability with changing prices as well as the change in movement that follows. The aim is to measure and forecast profitability of an item with changes in prices.

*Average profitability = (price) * (average per cent profitability at the price) * (average movement at that price)*

Hence, the movement dimension is also taken into consideration as movement changes with changes in prices, and as a result affects profitability as well.

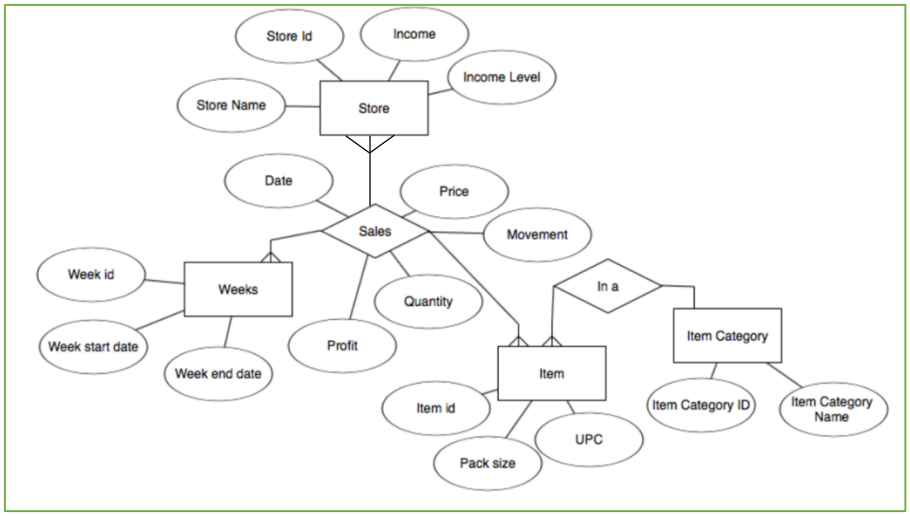
A simple graphical representation of the solution to the problem will help clarify this relationship. The graph is for item 'Caress bath - 6 bar' with UPC code 1111173011.



The graph here shows the effect of change in daily average movement with changes in price on the primary Y axis. The secondary Y axis represents the average percent profit per cent. The prices are shown on the X axis. The profitability is shown on the graph by the grey colored region at the bottom of the graph. We have extrapolated two logarithmic trend lines to predict the future trends in movement and profitability along with subsequent price changes.

From the graph, the decrease in movement with increasing prices can be visualized, along with that, the average profit percent is increasing with increase in prices. As a result, the grey region which represents the profitability (it is small due to a shared axis), is almost constant throughout all the prices. The extrapolated trend lines will help the organization get an idea about the expected behavior in movement and profitability should the prices be further changed.

The business implications for this question include an ability to predict the sales, which in turn can be an important factor while making decisions, like allocation storage space in the inventory. The quantity of the item that the store orders from the supplier can be regulated based on the predicted sales. The analysis will also be helpful to predict the future cash flows which will help the organization forecast more accurate figures in its financial reports.



ER Diagram

The ER diagram for this problem is again very similar to the ones for other business questions. The ER diagram is focused around the sales relationship. This relationship stores all the measures necessary for the analysis. It has attributes for storing price, movement, profit on a daily basis. It is a ternary relationship between the entities for stores, items, weeks. The store entity contains details of each store and the the unique attributes that are associated with each of the stores. The item entity stores the details for each individual items. Each item is further a part of a category. The week entity has details for the time dimension. The details necessary for answering this business question can be derived from these set of tables.

The pivot table data on the right is used to generate the data for the chart. The data here is primarily obtained from the movement table. The data is filtered for item with UPC code 1111173011 and then it is aggregated to get average daily movement and average profit per cent for each of the price values. The column for weighted profitability is calculated by multiplying the price, average profit per cent and the average daily movement. The graph is plotted as shown above. Logarithmic trend lines help in forecasting the trends.

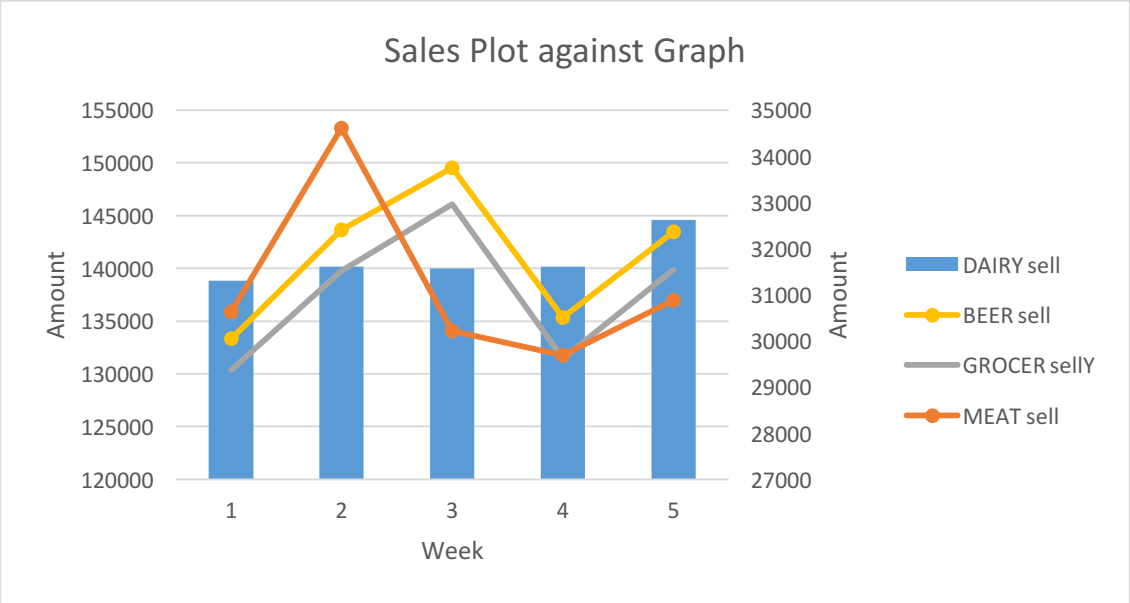
Row Labels	Average of MOVE	Average of PROFIT	Weighted Profit
1.79	12	-92.84	-1.9942032
2.33	29	-14.76	-0.9973332
3.17	4	-65.23	-0.8271164
3.99	4	-27.56666667	-0.439964
4.94	6	-4.93	-0.1461252
4.99	6.36	2.2964	0.072879469
5	1.5	-0.88	-0.0066
5.02	56	0.35	0.098392
5.1	13	-1.69	-0.112047
5.15	9	0.84	0.038934
5.42	9	7.71	0.3760938
5.47	8	8.55	0.374148

Pivot table used for plotting the graph

BQ 9 Generate the report depicting the sales summary for different types of product for different stores.

DFF being a retail store has the sales areas extended over wide breadth of products. The retail store sells a variety of food products and hence calls for one of the most basic reporting about its sales volume in each of the food category.

The business objective for this question is to find out the sales volume for a particular store for each of the food product. Please note that this is one of the most preliminary reporting status considered to be most basic, yet most powerful towards gauging the business standards. Hence, the report has been included as part of the primitive warehousing reporting giving appropriate sales amount for different product category. A graphical representation for this report can be show as below:



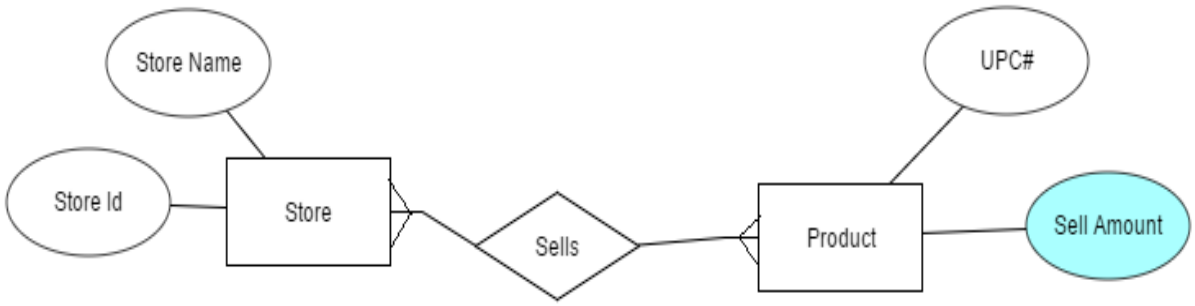
Please note that the above chart is based on a small sample volume and hence, doesn't represent the extensive trends it will generate once subjected to a higher data set. The data sets used for this projection has been provided below. The coupon amount when has a negative amount, represents that the coupon has been released from the company, but not yet deemed by the customer. A positive value for the same, in turn, represents that the customer has actually redeemed the same. The trend clearly depicts that releasing the coupon to the customers actually increases the customer count in the next week (which is in sync with retail industry's policy that coupons are generally redeemed in the next visit only). The data should present a more exhaustive view when presented with a higher meaningful data sets.

Source Data for Chart:

Generalized Report for product wise sell for each week for each store (#14 in example):

Weeks	DAIRY sell	MEAT sell	GROCER sell	BEER sell
1	31301.29	30636.29	130381.57	2966.65
2	31611.93	34609.15	139737.23	3907.73
3	31560.1	30214.52	146081.27	3460.3
4	31601.45	29693.1	131474.34	3880.44
5	32622.91	30877.03	139889.21	3587.19

Entity Relationship Diagram:



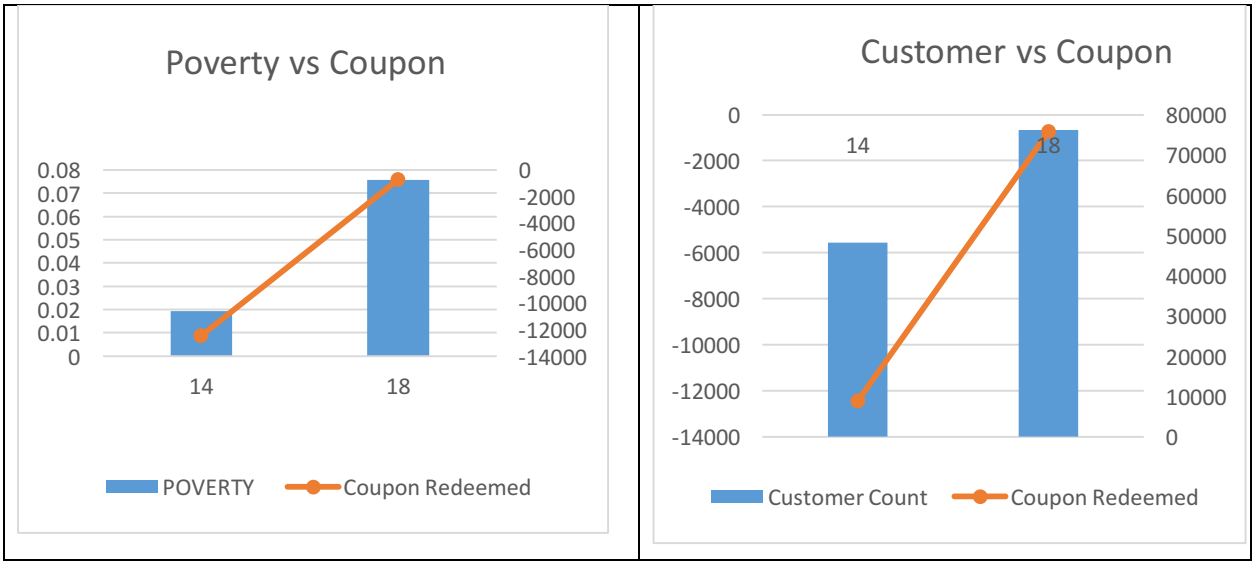
The above diagram represents the entity relationship diagram for this business scenario between store and product amount sold. The store has its main attributes store id and name while the product has its UPC number and sell amount, which is of vital importance for this case. This sell amount attribute it used to measure the sell value for each product for the mentioned week. To mention, this relationship uses “CCOUNT” file to facilitate the usage of all the required attributes required for this business case.

BQ 10 Generate the report depicting the effect of coupon introduction for increasing the sales in poverty affected areas.

DFF, like many other retail chains, has introduced coupons with certain profitability associated with it, with a target to increase its customer base. As per the report "*The Value of Competitive Information in Forecasting FMCG Retail Product Sales and the Variable Selection Problem*", it has been observed that once a customer gets associated with a particular store, he/she becomes more inclined to visit the same store as the last time. And coupons form a major strategic tool to force the customer to change the store preference.

The business objective for this question is to find out if the coupons introduction for the store is actually beneficial for the store or not and counters the poverty scenarios in the nearby areas. As is there that every coupon includes a certain small monetary disadvantage for the store and hence to ensure a profit, the number of customers visiting and actually purchasing should be accordingly increasing. If not the case, it will clearly mark the case of prospective loss for the store and hence the organization. To add to that, the question finds out if the introduction of coupons is beneficial for the organization in the poverty affected areas and increases the customer count or not. This business question addresses the comparison and effectiveness of coupon introduction by the organization in the poverty hit areas.

A graphical representation for this report can be show as below:



Please note that the above chart is based on a small sample volume and hence, doesn't represent the extensive trends it will generate once subjected to a higher data set. The data sets used for this projection has been provided below. The trend clearly depicts whenever the poverty is increasing, the coupon redemption at the store is also increasing and with the redemption increasing, the customer count at the store is also increasing. Hence, for the small data set it is clearly depicting

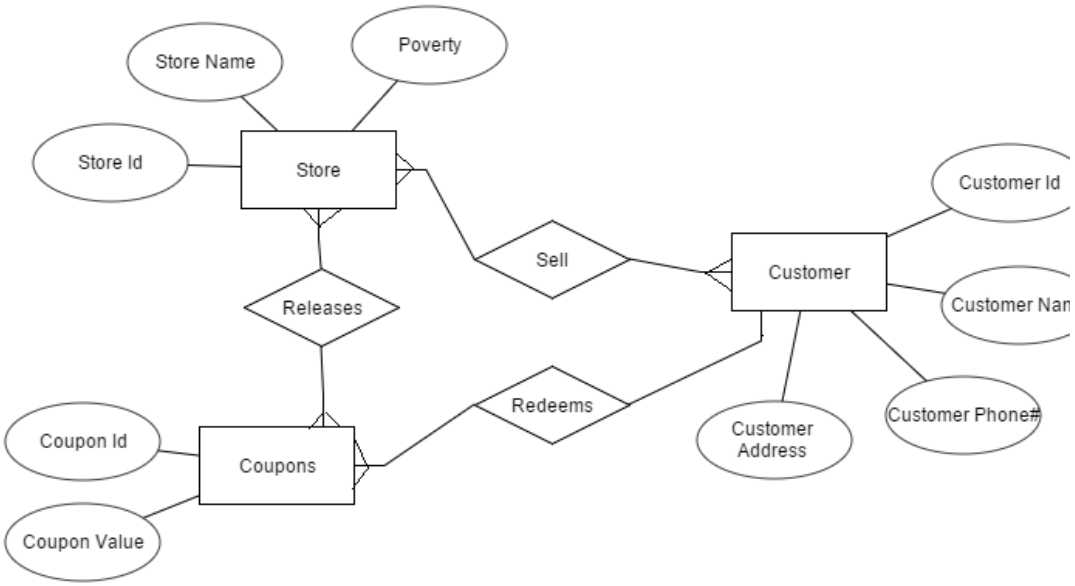
that despite of highly poverty affected areas, the system of couponing has been successful for the organization to increase the customers visiting the stores.

Source Data for Chart:

Stores/Attributes	14	18
Poverty Index	0.019350496	0.075727381
Coupon Redeemed	-12454.84	-741.42

Stores/Attributes	Customer Count	Coupon Redeemed
14	48204	-12454.84
18	76165	-741.42

Entity Relationship Diagram:



The above diagram represents the entity relationship diagram for this business scenario. The store sells the items to the customers and also releases the coupons. There is another relationship between customer and coupon defined as 'Redeem' which states the redemption option by the customer of the coupon. In a nutshell, there is a triangular relationship between the three critical entities Store, Coupons and Customer for this business case. The coupon value attribute is used to find the total coupon released to the customers and poverty attribute depicts the store wise poverty status for the demographics. To mention, this relationship uses "CCOUNT" and "DEMOGRAPHICS" file together to facilitate the usage of all the required attributes required for this business case.

Selected Questions

Below are the business questions which were selected and data marts are planned for them in the subsequent sections.

1.) Identify the most concerned areas of losses for each shop for Dominick's Fine Food (DFF) week-wise.
2.) How sales of an item of varying packet size is impacted by price change?
3.) Identifying price elasticity of demand for a particular item, w.r.t. income levels.
4.) Identifying the contribution of each product category in the overall sales for DFF.
5.) Generate the report depicting the effect of coupon introduction for increasing the customer count in the store.

5. Proposed Star schema

The schema proposed to answer these business questions is a combination of five dimensions: Store, Product, Time, Coupon, Price and two fact tables: one to hold data from the MOVEMENT source files called SALES_FACT and the other to score coupon usage data from CCOUNT.csv called STORESALE_FACT.

The Kimball matrix can be visualized as below:

Business Process	**Dimensions**					
	Date	Product	Store	Customer	Promotions	Packet size
Sales	Y	Y	Y		Y	Y
Coupon	Y		Y		Y	
Areas of losses	Y	Y	Y		Y	
Price elasticity		Y	Y			
Contribution of each product category	Y		Y			
Effect of coupon	Y		Y		Y	

It can be used to derive the schema as below.

The star schemas for the proposed model is an in Figure 1 and Figure 2.

The *time* dimension stores the data related to time. It has three attributes, the TIME_ID is an auto populated surrogate key which is the primary key for the table. According to the Dominick's Manual, the weeks run from 0 to 400, hence they are auto populated in the WEEK attribute. The value for YEAR attribute is calculated from the WEEK value using transformation T2 described later in the document.

The *store* dimension stores detail about each of the stores which is under DFF. This dimension mostly derives its value from the DEMO.csv which holds the demographic data for each of the stores. It has values for zone → city→ store number, in a hierarchical manner. Besides, it also

contains data for the income level of the locality of the store, which is calculated as per transformation T1 as described later. The income level attribute is used in business question 3.

The *product* dimension contains data regarding all the products that are sold in DFF. The source of data for this dimension table are the UPC files for each product category. Data from all the files corresponding to each of the categories flows into this dimension table. The category column is populated from the name of the csv file which is being used to populate the data. For example, for all the data derived from the UPC*FEC*.csv file will have '*FEC*' as the value of the category field. The rest of the fields are populated as described in the mapping tables which follow. The PACKAGE_SIZE and CASE columns are useful in business question 2. The hierarchy of attributes in this dimension is CATEGORY → PRODUCT_DESC → PACKAGE_SIZE, CASE, and UPC.

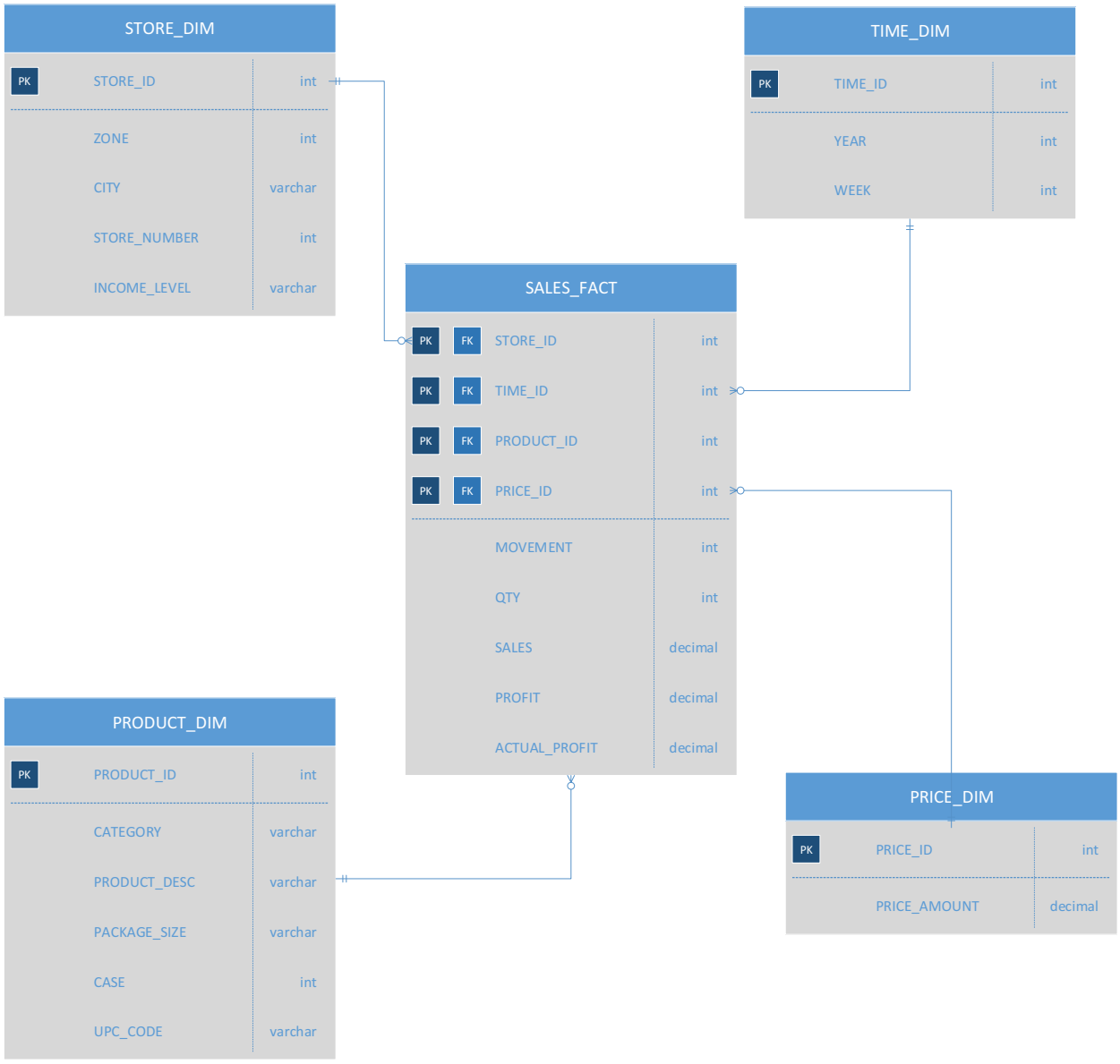
The *coupon* dimension stores the names of all the types of coupons there are available at the stores. This value is derived from the header column of the CCOUNT.csv file. It is a manual task to separate out the names of the field which represent the coupon usage in the dataset. Example of values in this dimension will include PRODCOUP, PROMCOUP, MEATCOUP, etc.

The *price* dimension table has been created after much deliberation. It stores the unique price values for each of the product prices for each of the product. These prices are populated from each of the movement tables. The purpose of this dimension is to help solve business questions 2 and 3, which involve calculating the price elasticity of demand. The average of movement is to be calculated for a given price to plot a average (movement) → price graph for any particular item. This dimension eases the task of aggregating the movement for each price.

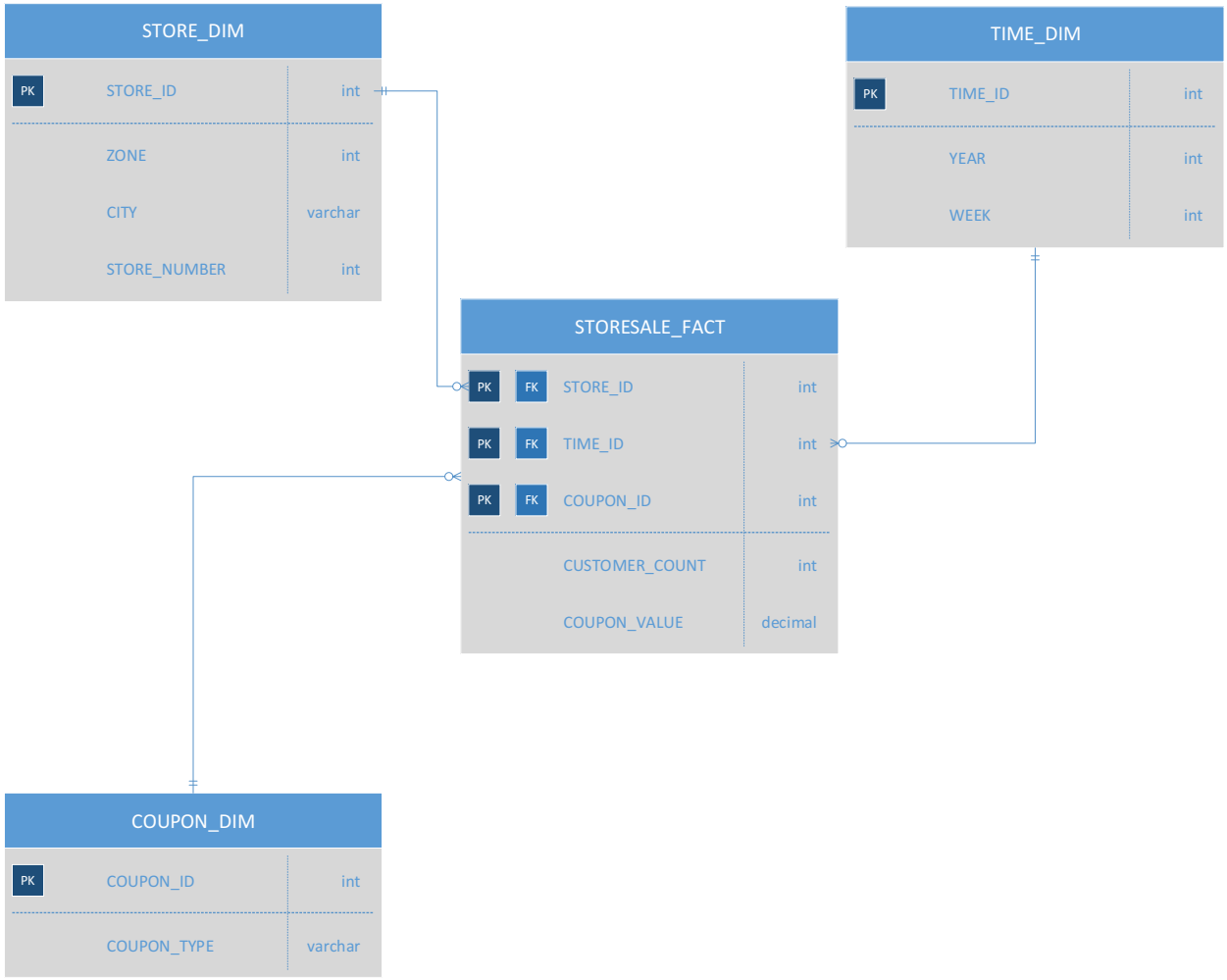
The SALES_FACT is the fact table which stores values derived from the movement datasets. Product, time, store and price dimensions are linked to this fact table. Note that the coupon dimension is not linked to this table. It stores the values for movement (number of items sold), qty (size of bundle), sale ($ amount of sales), profit (per cent of profit earned), actual_profit ($ amount of profit earned). The fields for sales and actual_profit are calculated as in transformations T3 and T4 which are described later. This fact table is used to address business questions 1, 2, 3 and 4.

The STORESALES_FACT is the fact table which is designed to address business question 5. It gets it data from the CCOUNT.csv. For each row of the dataset, there are multiple records populated in the fact table: one record for each of the couple type. Note that the data in CCOUNT is on a daily grain, it must be aggregated to weekly grain and the data for each of the coupon will be populated in the fact table for a given tuple of (week, store). The CUSTOMER_COUNT stores the number of customers who visited the store. The value of CUSTOMER_COUNT will be repeated for each set of (week, store), irrespective of the coupon_id.

6. STAR SCHEMA REPRESENTATION



Schema 1


STORE_DIM
PK
STORE_ID
int
ZONE
int
CITY
varchar
STORE_NUMBER
int
TIME_DIM
PK
TIME_ID
int
YEAR
int
WEEK
int
STORESALE_FACT
PK
FK
STORE_ID
int
PK
FK
TIME_ID
int
PK
FK
COUPON_ID
int
CUSTOMER_COUNT
int
COUPON_VALUE
decimal
COUPON_DIM
PK
COUPON_ID
int
COUPON_TYPE
varchar

Schema2

7. Mapping Tables

MAPPING(SOURCE, DIMENSION and FACT TABLE RELATIONSHIP) For DATA MART IMPLEMENTATION

Dimension: Store

Target Table	Target Column	Target Datatype	Source System/ Table	Source Column	Transformation/Business rule	Error Handling Rules
DWGRP4.STORE_DIM	STORE_ID	int	surrogate key		This is the Surrogate Key of the dimension and is not inserted or updated by the ETL process directly.	
DWGRP4.STORE_DIM	ZONE	int	DEMO.csv	ZONE		
DWGRP4.STORE_DIM	CITY	varchar	DEMO.csv	CITY		
DWGRP4.STORE_DIM	STORE_NUMBER	int	DEMO.csv	STORE		
DWGRP4.STORE_DIM	INCOME_LEVEL	varchar	DEMO.csv	INCOME	T1: Conversion of Log of Median income to income band, High, Medium, Low	If a row doesn't have income column of it is null, then set the target column as 'NA'

Dimension: Time

Target Table	Target Column	Target Datatype	Source System/ Table	Source Column	Transformation/Business rule	Error Handling Rules
DWGRP4.TIME_DIM	TIME_ID	int	surrogate key		This is the Surrogate Key of the dimension and is not inserted or updated by the ETL process directly.	
DWGRP4.TIME_DIM	YEAR	int			T2: Calculate YEAR from WEEK column	
DWGRP4.TIME_DIM	WEEK	int			Populate from 1 through 400	

Dimension: Product

Target Table	Target Column	Target Datatype	Source System/ Table	Source Column	Transformation/Business rule	Error Handling Rules
DWGRP4.PRODUCT_DIM	PRODUCT_ID	int	surrogate key			
DWGRP4.PRODUCT_DIM	CATEGORY	int	UPCXXX.csv	ZONE	Where XXX is the name of the category.	
DWGRP4.PRODUCT_DIM	PRODUCT_DESC	varchar	UPCXXX.csv	CITY	Where XXX is the name of the category.	
DWGRP4.PRODUCT_DIM	CASE	int	UPCXXX.csv	CASE	Where XXX is the name of the category.	
DWGRP4.PRODUCT_DIM	PACKAGE_SIZE	int	UPCXXX.csv	STORE	Where XXX is the name of the category.	
DWGRP4.PRODUCT_DIM	UPC_CODE	varchar	UPCXXX.csv	UPC	Where XXX is the name of the category.	

Dimension: Price						
Target Table	Target Column	Target Datatype	Source System/ Table	Source Column	Transformation/Business rule	Error Handling Rules
DWGRP4.PRICE_DIM	PRICE_ID	int	surrogate key		This is the Surrogate Key of the dimension and is not inserted or updated by the ETL process directly.	
DWGRP4.PRICE_DIM	PRIME_AMOUNT	decimal	DONE-XXXX.csv	PRICE	XXXX are the names of the categories. Source are the .csv files for all movement categories. Insert uniqyue values for price only. Don not insert values arready present in the table.	

Dimension: Coupon						
Target Table	Target Column	Target Datatype	Source System/ Table	Source Column	Transformation/Business rule	Error Handling Rules
DWGRP4.COUPON_DIM	COUPON_ID	int	surrogate key			
DWGRP4.COUPON_DIM	COUPON_TYPE	int	CCOUNT.csv		The header column for each COUPON TYPE will be used to populate this column based on the ETL transformation	

Fact: StoreSale						
Target Table	Target Column	Target Datatype	Source System/ Table	Source Column	Transformation/Business rule	Error Handling Rules
DWGRP4.STORESALE_FA	STORE_ID	int	DWGRP4.STORE_D	STORE_ID	Foreign key from the dimension table corresponding to STORE column of CCOUNT.csv	
DWGRP4.STORESALE_FA	TIME_ID	int	DWGRP4.TIME_DI	TIME_ID	Foreign key from the dimension table corresponding to WEEK column of CCOUNT.csv	
DWGRP4.STORESALE_FA	COUPON_ID	int	DWGRP4.COUPON	COUPON_ID	Foreign key from the dimension table corresponding to STORE column of CCOUNT.csv	
DWGRP4.STORESALE_FA	CUSTOMER_COUN	int	CCOUNT.csv	CUSTCOUN	Sum of CUSTCOUNT over the given week	
DWGRP4.STORESALE_FA	COUPON_VALUE	decimal	CCOUNT.csv	XXXXX	Depending on the Coupon_ID, Coupon_Value will be aggregated for the given week and stored for each type of coupon , like VIDCOUP, CUSTCOUP etc.	

Note: For every row of the Ccount.csv, there will be 22 rows in STORESALE_FACT table corresponding to the sale of each type of coupon for the store

Fact: Sales Fact Table						
Target Table	Target Column	Target Datatype	Source System/ Table	Source Column	Transformation/Business rule	Error Handling Rules
DWGRP4.SALES_FACT	STORE_ID	int	DWGRP4.STORE_DIM	STORE_ID	Foreign key from Dimension table corresponding to STORE of DONE_XXXX.csv	
DWGRP4.SALES_FACT	TIME_ID	int	DWGRP4.TIME_DIM	TIME_ID	Foreign key from Dimension table corresponding to WEEK of DONE-XXXX.csv	
DWGRP4.SALES_FACT	PRODUCT_ID	int	DWGRP4.PRODUCT_DIM	PRODUCT	Foreign key from Dimension table corresponding to UPC of DONE-XXXX.csv	
DWGRP4.SALES_FACT	PRICE_ID	int	DWGRP4.PRICE_DIM	PRICE_ID	Foreign key from Dimension table corresponding to PRICE of DONE-XXXX.csv	
DWGRP4.SALES_FACT	MOVEMENT	int	DONE-XXXX.csv	MOVE		
DWGRP4.SALES_FACT	QUANTITY	int	DONE-XXXX.csv	QTY		
DWGRP4.SALES_FACT	SALES	decimal	DONE-XXXX.csv	PRICE, MO	T3: Calculate Sales=(Price*Movement)/Quantity for given row in DONE-XXXX.csv	
DWGRP4.SALES_FACT	PROFIT	decimal	DONE-XXXX.csv	PROFIT		
DWGRP4.SALES_FACT	ACTUAL_PROFIT	decimal	DONE-XXXX.csv	PRICE, MO	T4: Calculate amount of Profit = Profit * ((Price*Movement)/Quantity) for given row in DONE-XXXX.csv	

8. ETL Development Plan

8.1 Data Quality Concerns with the DFF Data Sets:

Data quality refers to the conformity of the data sets with the intended usage for the system. The data present in a warehouse must conform to the field validation and should represent a comprehensive meaningful set of data for the integrated system as a whole. DFF data set was a raw dirty data set with multiple quality concerns which needed to be addressed while loading the data through ETL processes. Below mentioned are few of the data quality concerns that we addressed for this data set:

Group	Quality	Concern
Content within the data values	Valid	Negative values for week were present for multiple records in CCOUNT sheet.
	Complete	Approx. 24 rows in DEMO sheet had a null STORE name but with a valid zip value and sales volume for different Coupon types.
	Complete	Missing STORE CLUSTER values for multiple store numbers in DEMO sheet.
Structure of the fields	Format	Date standard for the data were ambiguous and sometimes filled with redundant data set.
	Standard	Redundant data like '.' present for fields like WEEK meant to have null values in DEMO sheet.
	Standard	The size field in UPC-XXX files have different standards like OZ in some rows while CT in other rows while none for some rows.
	Accurate	No measurement standard available for the sales of different food products like MEAT, MILK, BAKERY, PHARMACY. All the data have a common unit hence making the standard unit of measurement unclear.
Relation to the other data	Cardinality	The movement data is breakdown into multiple number of files as per the category and no product information or movement information present anywhere in the data sets.

All of these data quality issues were worked upon while loading the data into the staging environment using ETL processes. Also, there were few major calculations done based on the data present in the excel sheets to extract the required data not directly present in the source data set. Few of them include the product category extraction based on the filename or Actual Profit calculation based on the data present in the source, which was later used to address few of the critical business questions in the reporting. The final data extracted into the staging table followed the few characteristics conforming the qualities of high data quality including domain integrity, consistency and completeness discarding the rows with meaningless data or missing values for the data.

Below is the count of the quality data sets loaded into our Staging environment after quality implementation over the data values

Table Name	Cleaned Data Rows Count
[dbo].[CCOUNT]	327051
[dbo].[MOVEMENT]	85034835
[dbo].[UPC]	18117

8.2 ETL Development Plan:

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

The proposed plan is presented as

A. Determine target data
B. Determine source data
C. Mapping tables for staging and data mart loads
D. Comprehensive data extraction rules
E. Data staging area and screen shots
F. Data transformation and cleansing rules
G. Plan for aggregate tables
H. Procedures for data extraction and loading:
    i. ETL for dimension tables
    ii. ETL for fact tables

The implementation is shown as:

I. Mappings definition describing the source to end table for all dimension and fact tables
J. SQL Statements used for the ETL operations
K. Before and after table screen shots

A. Determine target data

The dimensional model has identified 5 dimensional tables and 2 fact tables.

There as as below:

Dimension: Time		
Target Table	**Target Column**	**Target Datatype**
DWGRP4.TIME_DIM	TIME_ID	int
DWGRP4.TIME_DIM	YEAR	int
DWGRP4.TIME_DIM	WEEK	int

Fact: Sales Fact Table		
Target Table	**Target Column**	**Target Datatype**
DWGRP4.SALES_FACT	STORE_ID	int
DWGRP4.SALES_FACT	TIME_ID	int
DWGRP4.SALES_FACT	PRODUCT_ID	int
DWGRP4.SALES_FACT	PRICE_ID	int
DWGRP4.SALES_FACT	MOVEMENT	int
DWGRP4.SALES_FACT	QUANTITY	int
DWGRP4.SALES_FACT	SALES	decimal
DWGRP4.SALES_FACT	PROFIT	decimal
DWGRP4.SALES_FACT	ACTUAL_PROFIT	decimal

Dimension: Product		
Target Table	**Target Column**	**Target Datatype**
DWGRP4.PRODUCT_DIM	PRODUCT_ID	int
DWGRP4.PRODUCT_DIM	CATEGORY	int
DWGRP4.PRODUCT_DIM	PRODUCT_DESC	varchar
DWGRP4.PRODUCT_DIM	CASE	int
DWGRP4.PRODUCT_DIM	PACKAGE_SIZE	int
DWGRP4.PRODUCT_DIM	UPC_CODE	varchar

Dimension: Price		
Target Table	**Target Column**	**Target Datatype**
DWGRP4.PRICE_DIM	PRICE_ID	int
DWGRP4.PRICE_DIM	PRIME_AMOUNT	decimal

Dimension: Coupon		
Target Table	**Target Column**	**Target Datatype**
DWGRP4.COUPON_DIM	COUPON_ID	int
DWGRP4.COUPON_DIM	COUPON_TYPE	int

Fact: StoreSale		
Target Table	**Target Column**	**Target Datatype**
DWGRP4.STORESALE_FACT	STORE_ID	int
DWGRP4.STORESALE_FACT	TIME_ID	int
DWGRP4.STORESALE_FACT	COUPON_ID	int
DWGRP4.STORESALE_FACT	CUSTOMER_COUNT	int
DWGRP4.STORESALE_FACT	COUPON_VALUE	decimal

Fact: Sales Fact Table		
Target Table	**Target Column**	**Target Datatype**
DWGRP4.SALES_FACT	STORE_ID	int
DWGRP4.SALES_FACT	TIME_ID	int
DWGRP4.SALES_FACT	PRODUCT_ID	int
DWGRP4.SALES_FACT	PRICE_ID	int
DWGRP4.SALES_FACT	MOVEMENT	int
DWGRP4.SALES_FACT	QUANTITY	int
DWGRP4.SALES_FACT	SALES	decimal
DWGRP4.SALES_FACT	PROFIT	decimal
DWGRP4.SALES_FACT	ACTUAL_PROFIT	decimal

B. Determine source data

The source data for the data mart are the .csv files. CCOUNT, DEMO and all the files for MOVEMENT and UPC for all the categories act as the source of data for the data mart.

C. Mapping tables for staging and data mart loads

A detail description for the same is provided under as per the report guidelines:

i. All the target data needed in the data marts and corresponding data sources:

MAPPING(SOURCE, DIMENSION and FACT TABLE RELATIONSHIP) For DATA MART IMPLEMENTATION

Dimension: Store							
Target Table	**Target Column**	**Target**	**Staging System/**	**Staging Column**	**File Used to Load**	**Transformation/Business rule**	**Error Handling Rules**
DWGRP4.STORE_DIM	STORE_ID	int	surrogate key			This is the Surrogate Key of the dimension and is not inserted or	
DWGRP4.STORE_DIM	ZONE	int	DEMO	ZONE	DEMO.csv		
DWGRP4.STORE_DIM	CITY	varchar	DEMO	CITY	DEMO.csv		
DWGRP4.STORE_DIM	STORE_NUMBER	int	DEMO	STORE	DEMO.csv		
DWGRP4.STORE_DIM	INCOME_LEVEL	varchar	DEMO	INCOME	DEMO.csv	T1: Conversion of Log of Median income to income band, High, Medium, Low	If a row doesn't have income column of it is null, then set the target column as 'NA'

Dimension: Time							
Target Table	**Target Column**	**Target**	**Staging System/**	**Staging Column**		**Transformation/Business rule**	**Error Handling Rules**
DWGRP4.TIME_DIM	TIME_ID	int	surrogate key			This is the Surrogate Key of the dimension and is not inserted or	
DWGRP4.TIME_DIM	YEAR	int	CCOUNT	Derived Column	CCOUNT.csv	T2: Calculate YEAR from WEEK column	
DWGRP4.TIME_DIM	WEEK	int	CCOUNT	WEEK	CCOUNT.csv	Populate from 1 through 400	

Dimension: Time							
Target Table	**Target Column**	**Target**	**Staging System/**	**Staging Column**		**Transformation/Business rule**	**Error Handling Rules**
DWGRP4.TIME_DIM	TIME_ID	int	surrogate key			This is the Surrogate Key of the dimension and is not inserted or	
DWGRP4.TIME_DIM	YEAR	int	CCOUNT	Derived Column	CCOUNT.csv	T2: Calculate YEAR from WEEK column	
DWGRP4.TIME_DIM	WEEK	int	CCOUNT	WEEK	CCOUNT.csv	Populate from 1 through 400	

Dimension: Product							
Target Table	**Target Column**	**Target**	**Staging System/**	**Staging Column**		**Transformation/Business rule**	**Error Handling Rules**
DWGRP4.PRODUCT_DIM	PRODUCT_ID	int	surrogate key				
DWGRP4.PRODUCT_DIM	CATEGORY	int	UPC	ZONE	UPC.csv	Where XXX is the name of the category.	
DWGRP4.PRODUCT_DIM	PRODUCT_DESC	varchar	UPC	CITY	UPC.csv	Where XXX is the name of the category.	
DWGRP4.PRODUCT_DIM	CASE	int	UPC	CASE	UPC.csv	Where XXX is the name of the category.	
DWGRP4.PRODUCT_DIM	PACKAGE_SIZE	int	UPC	STORE	UPC.csv	Where XXX is the name of the category.	
DWGRP4.PRODUCT_DIM	UPC_CODE	varchar	UPC	UPC	UPC.csv	Where XXX is the name of the category.	

Dimension: Price							
Target Table	**Target Column**	**Target**	**Staging System/**	**Staging Column**		**Transformation/Business rule**	**Error Handling Rules**
DWGRP4.PRICE_DIM	PRICE_ID	int	surrogate key			This is the Surrogate Key of the dimension and is not inserted or	
DWGRP4.PRICE_DIM	PRIME_AMOUNT	decimal	DONE	PRICE	DONE.csv	XXXX are the names of the categories. Source are the .csv files for all movement categories. Insert uniqyue values for price only. Don not insert	

Dimension: Coupon							
Target Table	**Target Column**	**Target**	**Staging System/**	**Staging Column**		**Transformation/Business rule**	**Error Handling Rules**
DWGRP4.COUPON_DIM	COUPON_ID	int	surrogate key				
DWGRP4.COUPON_DIM	COUPON_TYPE	int	CCOUNT		CCOUNT.csv	The header column for each COUPON TYPE will be used to populate this column based on the ETL	

Fact: StoreSale							
Target Table	**Target Column**	**Target**	**Staging System/**	**Staging Column**		**Transformation/Business rule**	**Error Handling Rules**
DWGRP4.STORESALE_FACT	STORE_ID	int	DWGRP4.STORE_D	STORE_ID		Foreign key from the dimension table corresponding to STORE column of	
DWGRP4.STORESALE_FACT	TIME_ID	int	DWGRP4.TIME_DIM	TIME_ID		Foreign key from the dimension table corresponding to WEEK column of	
DWGRP4.STORESALE_FACT	COUPON_ID	int	DWGRP4.COUPON	COUPON_ID		Foreign key from the dimension table corresponding to STORE column of	
DWGRP4.STORESALE_FACT	CUSTOMER_COU	int	CCOUNT	CUSTCOUN	CCOUNT.csv	Sum of CUSTCOUNT over the given week	
DWGRP4.STORESALE_FACT	COUPON_VALUE	decimal	CCOUNT	XXXXX	CCOUNT.csv	Depending on the Coupon_ID, Coupon_Value will be aggregated for the given week and stored for each	

Fact: Sales Fact Table							
Target Table	Target Column	Target	Staging System/	Staging Column		Transformation/Business rule	Error Handling Rules
DWGRP4.SALES_FACT	STORE_ID	int	DWGRP4.STORE_D	STORE_ID		Foreign key from Dimension table corresponding to STORE of	
DWGRP4.SALES_FACT	TIME_ID	int	DWGRP4.TIME_DIN	TIME_ID		Foreign key from Dimension table corresponding to WEEK of DONE-	
DWGRP4.SALES_FACT	PRODUCT_ID	int	DWGRP4.PRODUCT	PRODUCT_ID		Foreign key from Dimension table corresponding to UPC of DONE-	
DWGRP4.SALES_FACT	PRICE_ID	int	DWGRP4.PRICE_DI	PRICE_ID		Foreign key from Dimension table corresponding to PRICE of DONE-	
DWGRP4.SALES_FACT	MOVEMENT	int	DONE	MOVE	DONE.csv		
DWGRP4.SALES_FACT	QUANTITY	int	DONE	QTY	DONE.csv		
DWGRP4.SALES_FACT	SALES	decimal	DONE	PRICE, MOVE, QTY	DONE.csv	T3: Calculate Sales=(Price*Movement)/Quantity for	
DWGRP4.SALES_FACT	PROFIT	decimal	DONE	PROFIT	DONE.csv		
DWGRP4.SALES_FACT	ACTUAL_PROFIT	decimal	DONE	PRICE, MOVE, QTY	DONE.csv	T4: Calculate amount of Profit = Profit * ((Price*Movement)/Quantity) for given row in DONE-XXXX.csv	

ii.) Data mappings for data elements from sources in Excel to staging:

SOURCES IN EXCEL TO STAGING TABLES MAPPING

Staging Table: dbo.STORE					
File Used to Load	File Column	Datatype	Staging Table	Staging Column	Staging Column Datatype
DEMO.csv	ZONE	varchar	DWGRP4_STAGING_AREA.STORE	ZONE	int
DEMO.csv	CITY	varchar	DWGRP4_STAGING_AREA.STORE	CITY	varchar
DEMO.csv	STORE	varchar	DWGRP4_STAGING_AREA.STORE	STORE_NUMBER	int
DEMO.csv	INCOME	varchar	DWGRP4_STAGING_AREA.STORE	INCOME_LEVEL	varchar

Staging Table:dbo.Time					
File Used to Load	File Column	Datatype	Staging Table	Staging Column	Staging Column Datatype
CCOUNT.csv	WEEK	varchar	DWGRP4_STAGING_AREA.TIME	WEEK	int
CCOUNT.csv	WEEK	varchar	DWGRP4_STAGING_AREA.TIME	YEAR	int

Staging Table: dbo.Product					
File Used to Load	File Column	Datatype	Staging System/ Table	Staging Column	Staging Column Datatype
UPC.csv	ZONE	varchar	DWGRP4_STAGING_AREA.UPC	CATEGORY	int
UPC.csv	CITY	varchar	DWGRP4_STAGING_AREA.UPC	PRODUCT_DESC	varchar
UPC.csv	CASE	varchar	DWGRP4_STAGING_AREA.UPC	CASE	int
UPC.csv	STORE	varchar	DWGRP4_STAGING_AREA.UPC	PACKAGE_SIZE	int
UPC.csv	UPC	varchar	DWGRP4_STAGING_AREA.UPC	UPC_CODE	varchar

Staging Table:dbo.Price					
File Used to Load	File Column	Datatype	Staging System/ Table	Staging Column	Staging Column Datatype
DONE.csv	PRICE	varchar	DWGRP4_STAGING_AREA.PRICE	PRIME_AMOUNT	decimal

Staging Table: dbo.Coupon					
File Used to Load	File Column	Datatype	Staging System/ Table	Staging Column	Staging Column Datatype
CCOUNT.csv	COUPON_TYPE	varchar	DWGRP4_STAGING_AREA.COUPON	COUPON_TYPE	int

Staging Table: dbo.Movement					
File Used to Load	File Column	Datatype	Staging System/ Table	Staging Column	Staging Column Datatype
UPC.csv	COUPON_TYPE	varchar	DWGRP4_STAGING_AREA.Movement	STORE	int
UPC.csv	UPC	varchar	DWGRP4_STAGING_AREA.Movement	UPC	bigint
UPC.csv	WEEK	varchar	DWGRP4_STAGING_AREA.Movement	WEEK	int
UPC.csv	MOVE	varchar	DWGRP4_STAGING_AREA.Movement	MOVE	int
UPC.csv	QTY	varchar	DWGRP4_STAGING_AREA.Movement	QTY	int
UPC.csv	PRICE	varchar	DWGRP4_STAGING_AREA.Movement	PRICE	int
UPC.csv	SALE	varchar	DWGRP4_STAGING_AREA.Movement	SALE	int
UPC.csv	PROFIT	varchar	DWGRP4_STAGING_AREA.Movement	PROFIT	int
UPC.csv	OK	varchar	DWGRP4_STAGING_AREA.Movement	OK	int

iii.) Data mapping from staging to data marts (include all transformations):

STAGING To WAREHOUSE TABLES MAPPING

Staging Table: dbo.STORE_DIM					
Staging Table	Staging Column	Staging Column Datatype	Production Table	Production Column	Production Column Datatype
DWGRP4_STAGING_AREA.STORE	ZONE	int	DWGRP4_DW_AREA.STORE_DIM	ZONE	int
DWGRP4_STAGING_AREA.STORE	CITY	varchar	DWGRP4_DW_AREA.STORE_DIM	CITY	varchar
DWGRP4_STAGING_AREA.STORE	STORE_NUMBER	int	DWGRP4_DW_AREA.STORE_DIM	STORE_NUMBER	int
DWGRP4_STAGING_AREA.STORE	INCOME_LEVEL	varchar	DWGRP4_DW_AREA.STORE_DIM	INCOME_LEVEL	varchar

Dimension: dbo.TIME_DIM					
Staging Table	Staging Column	Staging Column Datatype	Production Table	Production Column	Production Column Datatype
DWGRP4_STAGING_AREA.SheetTIME	WEEK	int	DWGRP4_DW_AREA.TIME_DIM	WEEK	int
DWGRP4_STAGING_AREA.SheetTIME	YEAR	int	DWGRP4_DW_AREA.TIME_DIM	Calculated Field	int

Dimension: dbo.PRODUCT_DIM					
Staging Table	Staging Column	Staging Column Datatype	Production Table	Production Column	Production Column Datatype
DWGRP4_STAGING_AREA.Movement	CATEGORY	int	DWGRP4_DW_AREA.PRODUCT_DIM	CATEGORY	int
DWGRP4_STAGING_AREA.Movement	PRODUCT_DESC	varchar	DWGRP4_DW_AREA.PRODUCT_DIM	PRODUCT_DESC	varchar
DWGRP4_STAGING_AREA.Movement	CASE	int	DWGRP4_DW_AREA.PRODUCT_DIM	CASE	int
DWGRP4_STAGING_AREA.Movement	PACKAGE_SIZE	int	DWGRP4_DW_AREA.PRODUCT_DIM	PACKAGE_SIZE	int
DWGRP4_STAGING_AREA.Movement	UPC_CODE	varchar	DWGRP4_DW_AREA.PRODUCT_DIM	UPC_CODE	varchar

Dimension: dbo.PRICE_DIM					
Staging Table	Staging Column	Staging Column Datatype	Production Table	Production Column	Production Column Datatype
DWGRP4_STAGING_AREA.PRICE	PRIME_AMOUNT	decimal	DWGRP4_DW_AREA.PRICE_DIM	PRIME_AMOUNT	decimal

Dimension: dbo.COUPON_DIM					
Staging Table	Staging Column	Staging Column Datatype	Production Table	Production Column	Production Column Datatype
DWGRP4_STAGING_AREA.COUPON	COUPON_TYPE	int	DWGRP4_DW_AREA.COUPON_DIM	COUPON_TYPE	int

D. Comprehensive data extraction rules

The raw dirty data provided in the csv files have been extracted to staging tables. The extraction rules used are as follows:

1. CCOUNT data have been extracted to capture the data stored as column values into the row values for the COUPON dimension. Hence, this gives the row by row detail value of each of the coupon type and related discount offered for each of these coupon category. A representation of the same post extraction is as under below:

	COUPON_ID	COUPON_NAME
1	1	GROCCOUP
2	2	MEATCOUP
3	3	FISHCOUP
4	4	PROMCOUP
5	5	PRODCOUP
6	6	BULKCOUP
7	7	SALCOUP
8	8	FLORCOUP
9	9	DELICOUP
10	10	PHARCOUP
11	11	GMCOUP
12	12	VIDCOUP
13	13	MISCSCP
14	14	MANCOUP

2. All the data types are set for the corresponding columns, records with data like '.' In numeric fields are assumed to be junk data and ignored for CCOUNT table.
3. The condition STORE>0 is used for the extraction for taking the valid store records from the CCOUNT table.
4. The DEMO table was extracted to STORE staging table with the condition that STORE number should be integer to dump the dirty values like '.' In the column values.
5. The UPC data is captured to the staging area using a 'For each' container in SSIS. Data types in the staging area are defined as per the mapping tables above.
6. The MOVEMENT data is captured into the staging area using a 'For each' container in SSIS.
7. In the movement file, the records with OK value 1 were considered as part of the staging to production warehouse as only these records represents the affirmed interactions.
8. The Movement data has been captured in the staging table MOVEMENT, such that it also captures the name of the file with the location. This extracted field is then used to load the product category from staging to dimension table using SUBSTR function. It is as represented below:

Results | Messages

	STORE	UPC	WEEK	MOVE	QTY	PRICE	SALE	PROFIT	OK	CATEGORY
1	71	1192662108	295	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
2	101	1192662108	351	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
3	131	1192662108	308	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
4	12	1650001020	27	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
5	14	1650001020	181	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
6	32	1650001020	347	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
7	52	1650001020	312	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
8	73	1650001020	152	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
9	74	1650001020	244	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
10	76	1650001020	326	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
11	80	1650001020	330	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
12	86	1650001020	244	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
13	93	1650001020	369	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
14	98	1650001020	369	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
15	101	1650001020	19	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
16	102	1650001020	151	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
17	103	1650001020	286	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
18	105	1650001020	267	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....

E. Data staging area screen shots

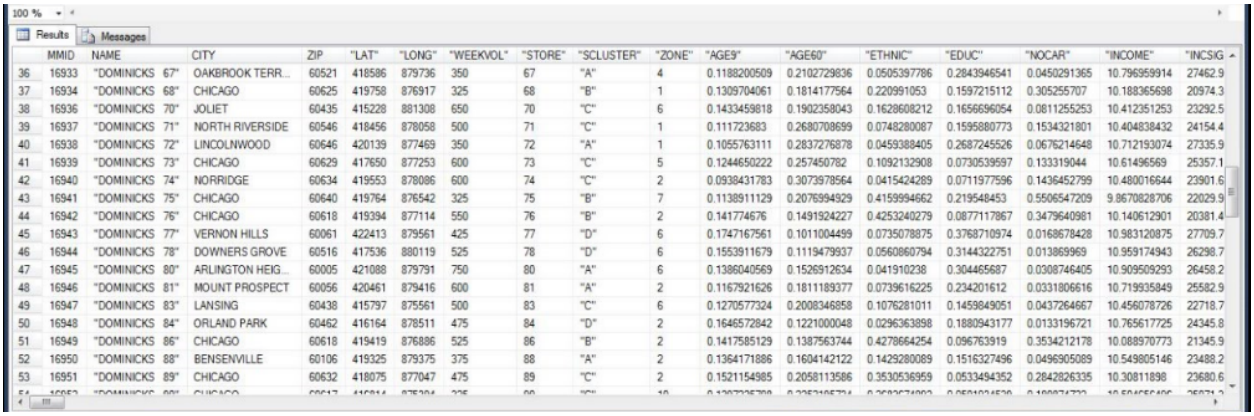
The raw dirty data have been loaded from the provided csv files into the staging data area called as DWGRP4_STAGING_AREA. The subsequent data is then loaded into the production data mart called as DWGRP4_DW_AREA. Both of these schemas have associated number of tables which has been loaded with extracted, cleaned and transformed data as per the rules and logic as described in the earlier sections. PFB the screenshot for the tables in staging area.

Table: DWGRP4_STAGING_AREA.CCOUNT

Results | Messages

	"STORE"	"DATE"	"GROCERY"	"DAIRY"	"FROZEN"	"BOTTLE"	"MVPCLUB"	"GROCCOUP"	"MEAT"	"MEATFROZ"	"MEATCOUP"	"FISH"	"FISHCOUP"	"PROMO"	"PROMCOUP"	"PRODUCE"	"BULK"	"SALADBAR
1	28	"910104"	16319.57	4195	3560.34	3.2	0	-1.8	4677.79	303.44	0	363.98	0	60.95	-25	3959.24	326.16	9.95
2	28	"910105"	18219.25	4476.09	3438.62	5.6	0	-0.75	5310.31	334.15	0	393.63	0	53.94	0	4113.76	307.41	27.19
3	28	"910106"	14153.83	3575.63	2602.33	3.2	0	-5	2894.06	353.77	0	97.69	0	13.99	0	3074.73	234.64	24.8
4	28	"910107"	11046.41	2788.16	2312.86	4	0	0	2401.09	219.91	-3.65	128.55	0	17.98	0	2686.21	225.26	7.4
5	28	"910108"	9906.25	2716.59	2268.01	1.6	0	-3.62	2602.66	258.37	-8.04	195.9	0	21.97	-10	2928.88	248.03	20.82
6	28	"910109"	10080.64	2647.18	2185.49	5.58	0	-2.98	2307.97	194.36	-3.2	120.38	0	0	0	2480.71	189.13	16.23
7	28	"910110"	13542.7	4012.2	2493.59	4.8	0	-480.16	3882.93	381.87	0	246.63	0	21.98	0	2862.36	308.79	19.74
8	28	"910111"	11169.25	3347.88	2344.61	2.4	0	-316.25	3019.83	234.97	0	273.32	0	0	0	2508.91	268.17	15.25
9	28	"910112"	21380.87	5736.87	3908.56	7.2	0	-484.83	6362.18	486.43	0	496.78	0	17.97	0	4916.5	416.65	52.94
10	28	"910113"	13918.15	3945.01	2819.19	2.4	0	-372.31	3003.77	359.73	0	169.18	0	102.9	-19.35	2955.33	339.56	36.11
11	28	"910114"	10877.61	3206.24	2371.67	4.8	0	-331.74	2272.56	249.92	0	167.83	0	64.92	-19.35	2535.79	252.5	16.37
12	28	"910115"	10347.58	3204.39	2261.04	1.6	0	-355.02	2753.37	243.59	0	245.32	0	39.96	-15.48	2632.26	210.45	25.77
13	28	"910116"	9452.35	3013.35	2012.66	3.2	0	-340.54	2224.35	218.81	0	163.2	0	79.92	-24.35	2212.75	200.47	18.21
14	28	"910117"	12180.05	2884.61	2176.39	5.6	0	-193.43	3614.32	319.82	0	210.52	0	80.89	-22.26	2648.73	292.67	28.39
15	28	"910118"	13701.53	3354.95	2688.22	2.4	0	-259.45	4082.49	427.57	0	352.92	0	66.95	-14.84	2982.7	315.45	11.86
16	28	"910119"	19069.48	4662.81	3705.51	11.2	0	-222.25	5707.96	534.68	0	319.52	0	122.86	-29.68	4225.48	338.91	30.32
17	28	"910120"	12715.99	3407.68	2593.86	3.2	0	-115.87	3362.9	313.07	0	99.77	0	84.9	-18.55	3049.17	318.49	25.43
18	28	"910121"	10073.56	2536.38	1877.18	1.6	0	-103.55	2542.18	251.34	0	200.68	0	76.91	-11.71	2230.36	288.57	4.47

Table: DWGRP4_STAGING_AREA.DEMO



	MMID	NAME	CITY	ZIP	"LAT"	"LONG"	"WEEKVOL"	"STORE"	"SCLUSTER"	"ZONE"	"AGE9"	"AGE60"	"ETHNIC"	"EDUC"	"NOCAR"	"INCOME"	"INCSIG
36	16933	"DOMINICKS 67"	OAKBROOK TERR...	60521	418586	879736	350	67	"A"	4	0.1188200509	0.2102729836	0.0505397786	0.2843946541	0.0450291365	10.796959914	27462.9
37	16934	"DOMINICKS 68"	CHICAGO	60625	419758	876917	325	68	"B"	1	0.1309704061	0.1814177564	0.220991053	0.1597215112	0.305255707	10.188365698	20974.3
38	16936	"DOMINICKS 70"	JOLIET	60435	415228	881308	650	70	"C"	6	0.1433459818	0.1902358043	0.1628608212	0.1656696054	0.0811255253	10.412351253	23292.5
39	16937	"DOMINICKS 71"	NORTH RIVERSIDE	60546	418456	878058	500	71	"C"	1	0.111723683	0.2680708699	0.0748280087	0.1595880773	0.1534321801	10.404838432	24154.4
40	16938	"DOMINICKS 72"	LINCOLNWOOD	60646	420139	877469	350	72	"A"	1	0.1055763111	0.2837276878	0.0459388405	0.2687245526	0.0676214648	10.712193074	27335.9
41	16939	"DOMINICKS 73"	CHICAGO	60629	417650	877253	600	73	"C"	5	0.1244650222	0.257450782	0.1092132908	0.0730539597	0.133319044	10.61496569	25357.1
42	16940	"DOMINICKS 74"	NORRIDGE	60634	419553	878086	600	74	"C"	2	0.0938431783	0.3073978564	0.0415424289	0.0711977596	0.1436452799	10.480016644	23901.6
43	16941	"DOMINICKS 75"	CHICAGO	60640	419764	876542	325	75	"B"	7	0.1138911129	0.2076994929	0.4159994662	0.219548453	0.5506547209	9.8670828706	22029.9
44	16942	"DOMINICKS 76"	CHICAGO	60618	419394	877114	550	76	"B"	2	0.141774676	0.1491924227	0.4253240279	0.0877117867	0.3479640981	10.140612901	20381.4
45	16943	"DOMINICKS 77"	VERNON HILLS	60061	422413	879561	425	77	"D"	6	0.1747167561	0.1011004499	0.0735078875	0.3768710974	0.0168678428	10.983120875	27709.7
46	16944	"DOMINICKS 78"	DOWNERS GROVE	60516	417536	880119	525	78	"D"	6	0.1553911679	0.1119479937	0.0560860794	0.3144322751	0.013869969	10.959174943	26298.7
47	16945	"DOMINICKS 80"	ARLINGTON HEIG...	60005	421088	879791	750	80	"A"	6	0.1386040569	0.1526912634	0.041910238	0.304465687	0.0308746405	10.909509293	26458.2
48	16946	"DOMINICKS 81"	MOUNT PROSPECT	60056	420461	879416	600	81	"A"	2	0.1167921626	0.1811189377	0.0739616225	0.234201612	0.0331806616	10.719935849	25582.9
49	16947	"DOMINICKS 83"	LANSING	60438	415797	875561	500	83	"C"	6	0.1270577324	0.2008346858	0.1076281011	0.1459849051	0.0437264667	10.456078726	22718.7
50	16948	"DOMINICKS 84"	ORLAND PARK	60462	416164	878511	475	84	"D"	2	0.1646572842	0.1221000048	0.0296363898	0.1880943177	0.0133196721	10.765617725	24345.8
51	16949	"DOMINICKS 86"	CHICAGO	60618	419419	876886	525	86	"B"	2	0.1417585129	0.1387563744	0.4278664254	0.096763919	0.3534212178	10.088970773	21345.9
52	16950	"DOMINICKS 88"	BENSENVILLE	60106	419325	879375	375	88	"A"	2	0.1364171886	0.1604142122	0.1429280089	0.1516327496	0.0496905089	10.549805146	23488.2
53	16951	"DOMINICKS 89"	CHICAGO	60632	418075	877047	475	89	"C"	2	0.1521154985	0.2058113586	0.3530536959	0.0533494352	0.2842826335	10.30811898	23680.6

Table: DWGRP4_STAGING_AREA.PRODUCT

Results Messages

	COM_CODE	UPC	DESCRIP	SIZE	CASE_SIZE	NITEM	CATEGORY
1	953	1192603016	"CAFFEDRINE CAPLETS 1"	"16 CT"	6	7342431	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
2	953	1192662108	"SLEEPINAL SOFTGEL"	"8 CT"	6	7333311	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
3	953	1650001020	"NERVINE TABS"	"30 CT"	1	8430820	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
4	953	1650001022	"NERVINE SLEEP AID"	"12 CT"	1	8430840	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
5	953	1650004106	"ALKA-SELTZER GOLD"	"20 CT"	1	8430880	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
6	953	1650004108	"ALKA-SELTZER GOLD"	"36 CT"	1	8430900	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
7	953	1650004703	"ALKA MINTS"	"30 CT"	1	8430700	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
8	953	2140649030	"LEGATRIN PM"	"30 CT"	1	8435810	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
9	953	2586600493	"PERCOGESIC A/F ANALG"	"50 CT"	1	8416280	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
10	953	2586610493	"PERCOGESIC A/F ANALG"	"50 CT"	1	8416280	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
11	953	2586610501	"ALEVE TABLETS"	"24 CT"	6	6122441	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
12	953	2586610502	"ALEVE CAPLETS"	"24 CT"	6	6122741	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
13	953	2586610503	"ALEVE TABLETS"	"50 CT"	6	6122451	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
14	953	2586610504	"ALEVE CAPLETS"	"50 CT"	6	6122751	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
15	953	2586610505	"ALEVE TABLETS"	"100 CT"	6	6122461	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
16	953	2586610506	"ALEVE CAPLETS"	"100 CT"	6	6122761	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
17	953	3225259620	"SUNBEAM HEAT WRAP MS"	"1 CT"	1	8402470	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
18	953	3680012732	"TC MOTION SICKNESS T"	"12 CT"	12	6190791	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
19	953	3680012740	"VALUE TIME ASPIRIN"	"250 CT"	12	6108051	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv

Table: DWGRP4_STAGING_AREA.COUPON, derived from CCOUNT , aftter unpivot on header row.


Results Messages

	COUPON_ID	COUPON_NAME
1	1	GROCCOUP
2	2	MEATCOUP
3	3	FISHCOUP
4	4	PROMCOUP
5	5	PRODCOUP
6	6	BULKCOUP
7	7	SALCOUP
8	8	FLORCOUP
9	9	DELICOUP
10	10	PHARCOUP
11	11	GMCOUP
12	12	VIDCOUP
13	13	MISCSCP
14	14	MANCOUP
15	15	CUSTCOUN
16	16	FTGCCOUP
17	17	FTGICOUP
18	18	DAIRCOUP
19	19	FROZCOUP

Query executed successfully.

Table: DWGRP4_STAGING_AREA.MOVEMENT

Results Messages

	STORE	UPC	WEEK	MOVE	QTY	PRICE	SALE	PROFIT	OK	CATEGORY
1	71	1192662108	295	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
2	101	1192662108	351	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
3	131	1192662108	308	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
4	12	1650001020	27	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
5	14	1650001020	181	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
6	32	1650001020	347	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
7	52	1650001020	312	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
8	73	1650001020	152	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
9	74	1650001020	244	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
10	76	1650001020	326	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
11	80	1650001020	330	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
12	86	1650001020	244	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
13	93	1650001020	369	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
14	98	1650001020	369	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
15	101	1650001020	19	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
16	102	1650001020	151	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
17	103	1650001020	286	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
18	105	1650001020	267	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....
19	107	1650001020	346	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA....

Table: DWGRP4_STAGING_AREA.STORE

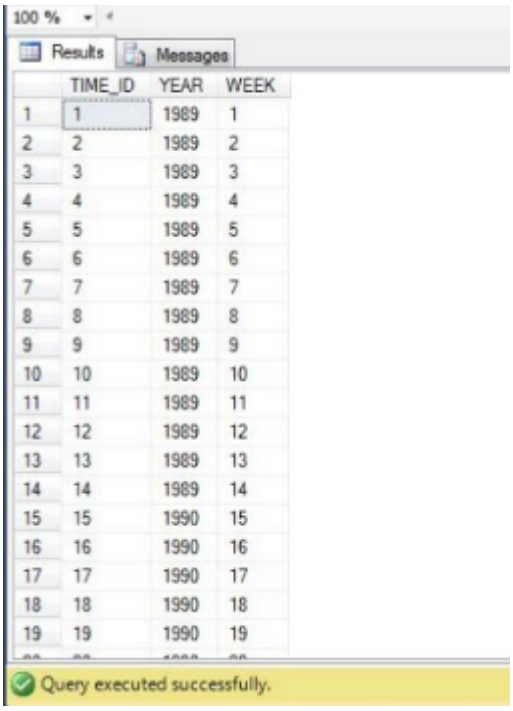
100 %

Results Messages

	STORE_ID	ZONE	CITY	INCOME	STORE_NUMBER	INCOMELEVEL
1	1	1	RIVER FOREST	10.553205	2	MEDIUM
2	2	2	PARK RIDGE	10.646971	4	MEDIUM
3	3	2	PALATINE	10.922370	5	HIGH
4	4	5	OAK LAWN	10.597009	8	MEDIUM
5	5	2	MORTON GROVE	10.787151	9	MEDIUM
6	6	7	CHICAGO	9.996659	12	LOW
7	7	1	GLENVIEW	11.043929	14	HIGH
8	8	5	RIVER GROVE	10.391975	18	LOW
9	9	6	HANOVER PARK	10.716193	21	MEDIUM
10	10	2	MOUNT PROSPECT	10.798534	28	MEDIUM
11	11	1	PARK RIDGE	10.674475	32	MEDIUM
12	12	7	CHICAGO	10.345927	33	LOW
13	13	6	BRIDGEVIEW	10.550250	40	MEDIUM
14	14	2	WESTERN SPRINGS	10.869158	44	HIGH
15	15	2	WHEELING	10.745377	45	MEDIUM
16	16	2	ADDISON	10.635326	47	MEDIUM
17	17	2	SCHAUMBURG	10.756028	48	MEDIUM
18	18	2	DOWNERS GROVE	10.806753	49	HIGH
19	19	2	HICKORY HILLS	10.589307	50	MEDIUM

Query executed successfully.

Table: DWGRP4_STAGING_AREA.SHEET_TIME, Auto generated table with a transformation for derivation of YEAR.



100 %

Results | Messages

	TIME_ID	YEAR	WEEK
1	1	1989	1
2	2	1989	2
3	3	1989	3
4	4	1989	4
5	5	1989	5
6	6	1989	6
7	7	1989	7
8	8	1989	8
9	9	1989	9
10	10	1989	10
11	11	1989	11
12	12	1989	12
13	13	1989	13
14	14	1989	14
15	15	1990	15
16	16	1990	16
17	17	1990	17
18	18	1990	18
19	19	1990	19

Query executed successfully.

F. Data transformation and cleansing rules

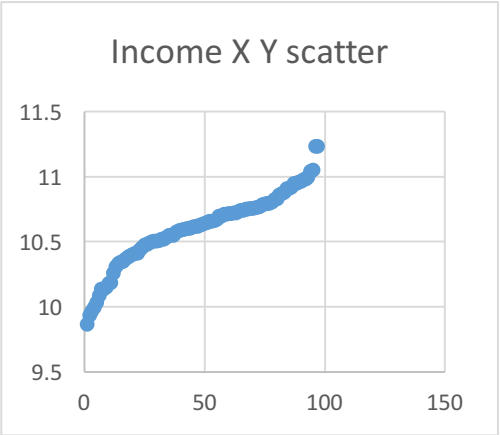
Most of the cleansing tasks were carried out during the extraction phase itself. But there are still certain cleansing tasks while are carried out:

1. CCOUNT.csv data was loaded into the staging table with the condition of removal of insignificant store id in the sheet.
2. A second filter was used on the CCOUNT for removing the erroneous week data. Hence, the junk values like '.' and negative integers were removed from the CCOUNT.
3. In the movement file, the records with OK value 1 were considered as part of the staging to production warehouse as only these records represents the affirmed interactions.
4. The PRODUCT_DESCRIPTION column was filtered to remove any special characters or symbols.
5. Data in PACK_SIZE of UPC_STAGE was filtered to include only numeric values.

Then subsequent transformations were carried out on the stage table data via SSIS packages (described in next section), before loading them in the production data marts. The transformations are as below:

T1: Conversion of Log of Median income to income band, High, Medium, Low

To identify and separate the stores into three levels of income: 'low', 'medium' and 'high'. The income for each store, obtained from the store demographic dataset is plotted on a XY scatter and as visible, there is a stark increase in slope at around 10.4 and 10.8 and we shall assume these to be the separating points as below:
LOW: Income < 10.4
MEDIUM: 10.4 <= Income <= 10.8
HIGH: Income > 10.8



T2: Calculate YEAR from WEEK column

The decoding for WEEK column is given in the Dominick's manual, page 21. A screenshot of the data is as below:

Part 8: Week's Decode Table

The SAS files contain a week variable that codes the week for which a data point is recorded. The table below converts calendar weeks into database weeks.

You can use your browser's Find function to jump to a specific week.

Week #	Start	End	Special Events
1	09/14/89	09/20/89	
2	09/21/89	09/27/89	
3	09/28/89	10/04/89	
4	10/05/89	10/11/89	
5	10/12/89	10/18/89	
6	10/19/89	10/25/89	
7	10/26/89	11/01/89	Halloween
8	11/02/89	11/08/89	
9	11/09/89	11/15/89	
10	11/16/89	11/22/89	
11	11/23/89	11/29/89	Thanksgiving

Here we may use the rough calculation that if WEEK <= 15, then YEAR = 1989

If WEEK > 15 then,

YEAR = 1990 + INTEGER_PART_OF [(WEEK - 15) / 52]

Here, INTERGET_PART_OF is a function which will return the value before the decimal point in case of a decimal number, irrespective of value after the decimal point.
For example,
INTEGER_PART_OF [0.9] = 0
INTEGER_PART_OF [3.1] = 3
INTEGER_PART_OF [1.5] = 1

Here, if WEEK is say, 197
YEAR = 1990 + INTEGER_PART_OF [(197 – 15) / 52]
= 1990 + INTEGER_PART_OF [3.5]
= 1993

T3: Calculate Sales = (Price*Movement)/Quantity for given row in DONE-XXXX.csv

As stated, the value for sale is calculated as (Price*Movement)/Quantity from the row level data available in the DONE-XXXX.csv i.e. the movement data.

T4: Calculate amount of Profit = Profit * ((Price*Movement)/Quantity) for given row in DONE-XXXX.csv

As stated, the value for profit amount is calculated as Profit * ((Price*Movement)/Quantity) from the row level data available in the DONE-XXXX.csv i.e. the movement data.

T5: Derive CATEGORY from file path in staging area.
Use the derived column with function SUBSTRING (37,3) on the file path column to extract the name of the product category.
Example: C:\Users\Desktop\.....\UPCANA.csv becomes ANA.

T6: Weekly aggregation for CCOUNT data
Aggregate CCOUNT_STAGE data over the columns WEEK, STORE, summing up the values for coupons for all categories and customer count.

T7: Un-pivot for CCOUNT_STAGE
The CCOUNT_STAGE data is un-pivot so that the columns representing coupon codes are transformed into subsequent rows.

T8: STORE_DIM lookup

Lookup used to get STORE_ID from store number. Used while populating fact tables.

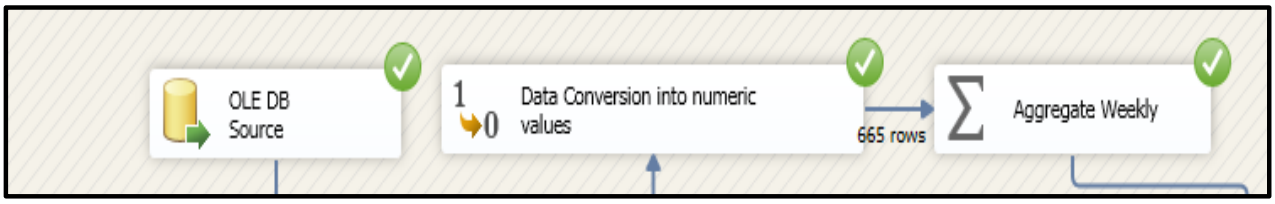
T9: PRODUCT_DIM lookup
Lookup used to get PRODUCT_ID from UPC. Used while populating fact tables.

T10: PRICE_DIM lookup
Lookup used to get PRICE_ID from price. Used while populating fact tables.

G. Plan for aggregate tables

The granularity of the data has been decided to be fixed at a weekly level. There are two main sources of data available to us for loading the measurements: CCOUNT data and MOVEMENT data. The movement data is already at a weekly level and hence no transformation is needed for this data.

The CCOUNT data is at daily granularity and hence it must be aggregated to a weekly level. It is done in an intermediate step in the SSIS dataflow while loading the STORESALES_FACT table. No table is to be created individually. This process is as per the below image:



H. Procedures for all data extractions and loadings

The data extraction tasks were carried out as shown previously. Data was extracted from the source .csv files and it was subsequently loaded in the STAGE tables. Stage tables were created for CCOUNT, MOVEMENT, UPC and DEMO to store wise sales, weekly movement, product data and store demographic data respectively.

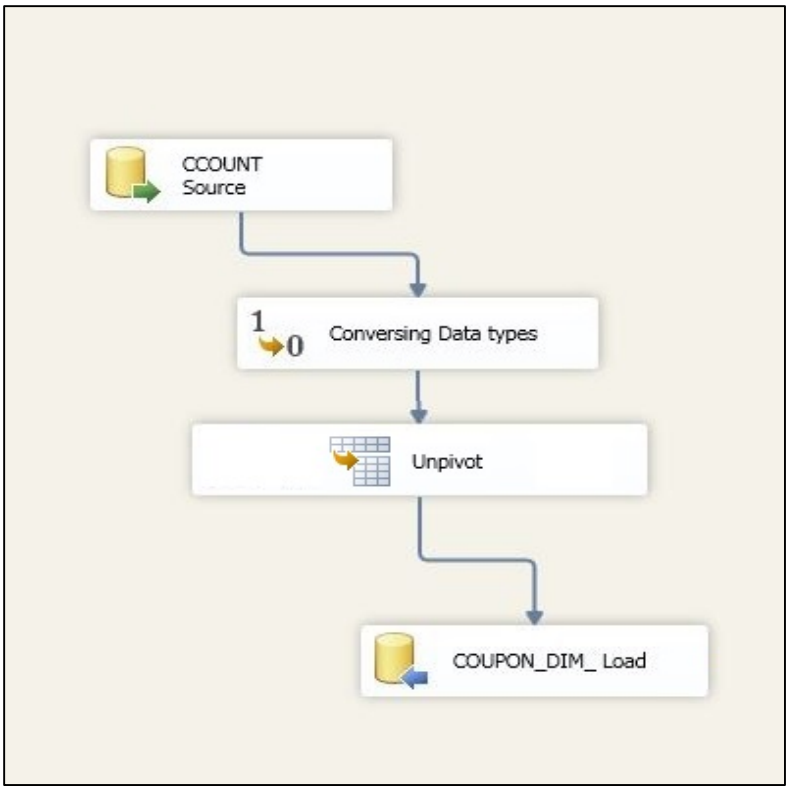
These stage tables were then subjected to rigorous cleansing as described previously.

Then the stage tables underwent several transformations, described in the above section to finally generate the fact and dimension tables for the production data mart. These ETL processes are described as below:

i.) ETL for dimension tables

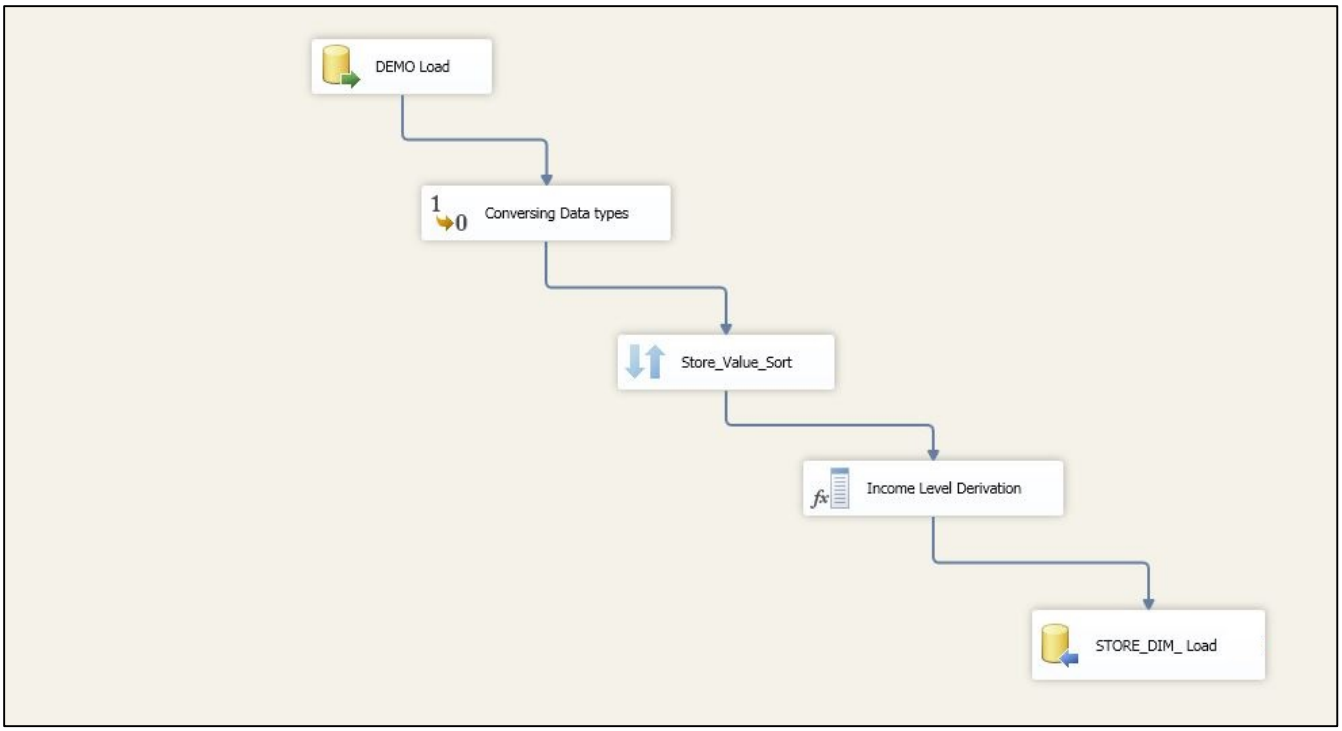
1. **COUPON_DIM**

   This dimension table is loaded from the header row of CCOUNT stage table. The data is transformed into appropriate data type first and then coupon values are derived from them. The data is un-pivot to convert rows to columns. These values are then stored in the COUPON_DIM table.



2. STORE_DIM

This table is derived from the DEMO stage data. Data conversions is carried out first into appropriate data types. Data is then sorted, and the INCOME_LEVEL column is derived from the INCOME column according to transformation T1. The resulting data is then stored in the table STORE_DIM.

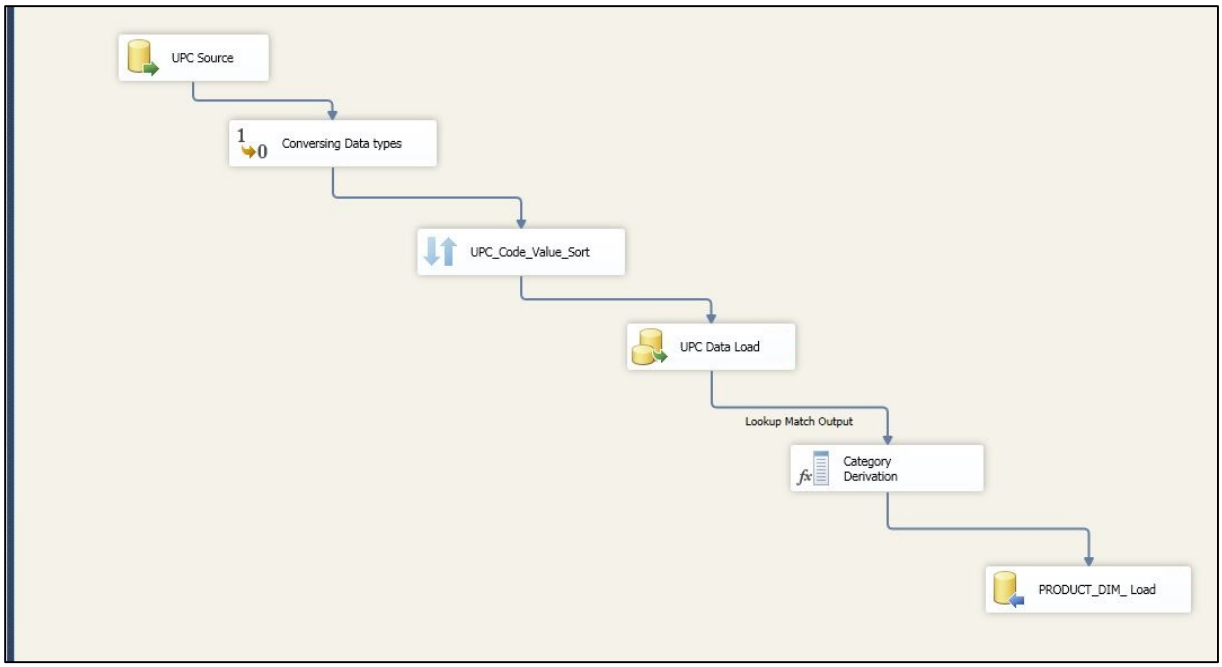


3. **TIME_DIM**

This table is not loaded via any ETL action. The week column is populated with numbers 1 through 400. Then the YEAR column is derived using transformation T2 described previously.

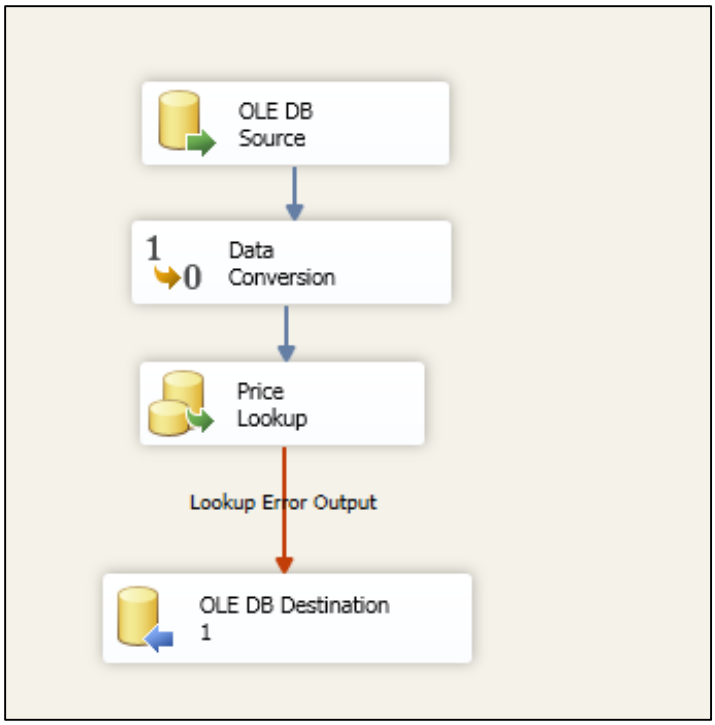
4. **PRODUCT_DIM LOAD**

The product dimension is loaded from the UPC stage table. Data is converted into appropriate data type first of all. It is then sorted on UPC_CODE and checked if the converted values exist in the UPC stage. The category is then derived from the file_name column and the resulting data is stored in the



5. PRICE_DIM

The price dimension table is loaded from the movement stage data. The price amount is first checked if it already exists in the PRICE_DIM table via a look up. If the value does not exist, the look up returns an error. This error record is then stored in the PRICE_DIM table.

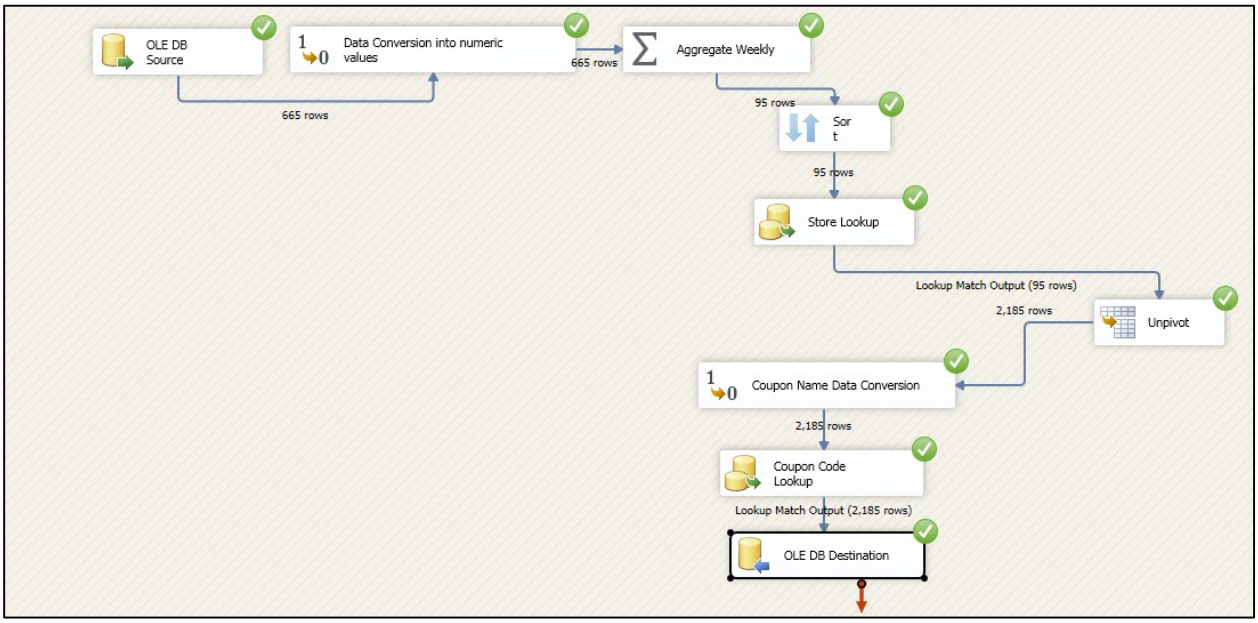


ii.) ETL for fact tables:

The two fact tables ETL data load from staging to Data warehouse tables has been represented below:

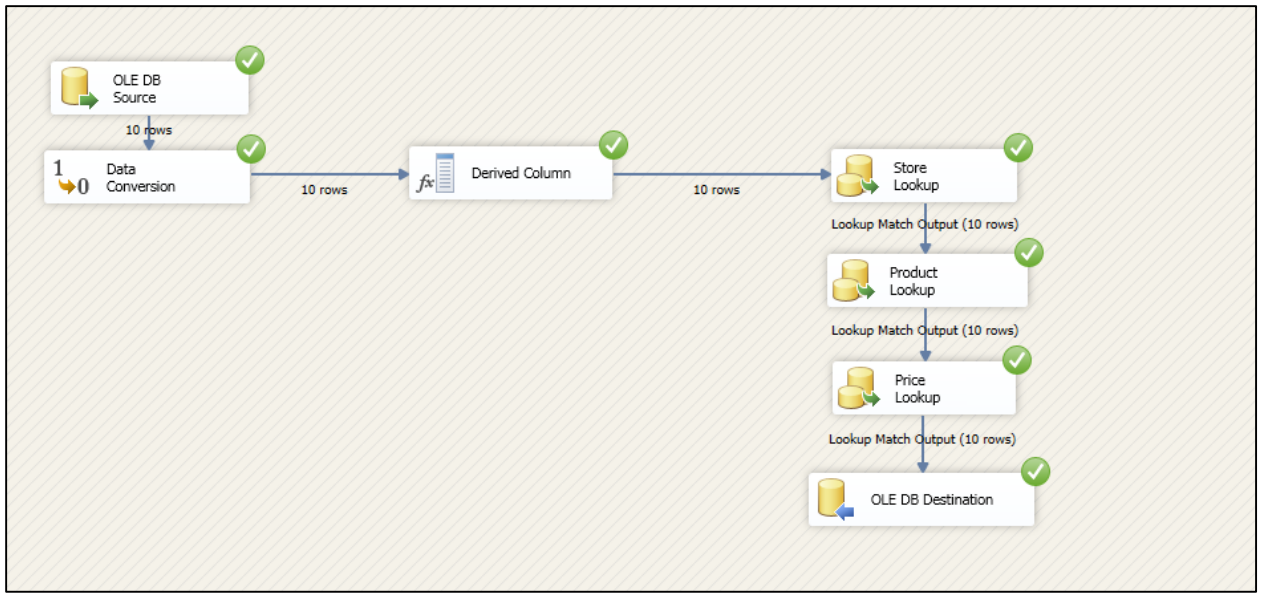
1. STORESALE_FACT

The data flow tasks for populating the storesale_fact are as below. It is populated from the ccount stage. Data is first aggregated into weekly data from daily data. It is then un-pivot. Data is sorted and lookup transformations are used to get store_id, coupon_id from store number and coupon type respectively. This data is then loaded in the fact table.



2. SALES_FACT

The sales_fact table is generated from the movement table. The values for SALES and ACTUAL profit columns are derived as shown in the transformations. Look up transformations are then used to get store_id , product_id, price_id from the store number, UPC name, product price and the resulting data is then stored in the SALES_FACT table.



Implementation:

I. Mappings definition describing the source to end table for all dimension and fact tables

END TO END TABLES MAPPING IMPLEMENTATION

Note: Surrogate Keys have not been displayed here for any of the tables as it doesn't hold any dependenc on any source file/table and needs to be generated automatically in a sequential manner

STORE_DIMENSION LOAD MAPPING FROM FILE TO _DIM TABLE in PRODUCTION WAREHOUSE

File Used to Load	File Column	Datatype	Staging Table	Staging Column	Staging Column Datatype	Production Table	Production Column	Production Column Datatype
DEMO.csv	ZONE	varchar	DWGRP4_STAGING_AREA.STORE	ZONE	int	DWGRP4_DW_AREA.STORE_DIM	ZONE	int
DEMO.csv	CITY	varchar	DWGRP4_STAGING_AREA.STORE	CITY	varchar	DWGRP4_DW_AREA.STORE_DIM	CITY	varchar
DEMO.csv	STORE	varchar	DWGRP4_STAGING_AREA.STORE	STORE_NUMBER	int	DWGRP4_DW_AREA.STORE_DIM	STORE_NUMBER	int
DEMO.csv	INCOME	varchar	DWGRP4_STAGING_AREA.STORE	INCOME_LEVEL	varchar	DWGRP4_DW_AREA.STORE_DIM	INCOME_LEVEL	varchar

TIME_DIMENSION LOAD MAPPING FROM FILE TO _DIM TABLE in PRODUCTION WAREHOUSE

File Used to Load	File Column	Datatype	Staging Table	Staging Column	Staging Column Datatype	Production Table	Production Column	Production Column Datatype
CCOUNT.csv	WEEK	varchar	DWGRP4_STAGING_AREA.SheetTIME	WEEK	int	DWGRP4_DW_AREA.TIME_DIM	WEEK	int
CCOUNT.csv	WEEK	varchar	DWGRP4_STAGING_AREA.SheetTIME	YEAR	int	DWGRP4_DW_AREA.TIME_DIM	Calculated Field	int

PRODUCT_DIMENSION LOAD MAPPING FROM FILE TO _DIM TABLE in PRODUCTION WAREHOUSE

File Used to Load	File Column	Datatype	Staging Table	Staging Column	Staging Column Datatype	Production Table	Production Column	Production Column Datatype
UPC.csv	ZONE	varchar	DWGRP4_STAGING_AREA.Movement	CATEGORY	int	DWGRP4_DW_AREA.PRODUCT_DIM	CATEGORY	int
UPC.csv	CITY	varchar	DWGRP4_STAGING_AREA.Movement	PRODUCT_DESC	varchar	DWGRP4_DW_AREA.PRODUCT_DIM	PRODUCT_DESC	varchar
UPC.csv	CASE	varchar	DWGRP4_STAGING_AREA.Movement	CASE	int	DWGRP4_DW_AREA.PRODUCT_DIM	CASE	int
UPC.csv	STORE	varchar	DWGRP4_STAGING_AREA.Movement	PACKAGE_SIZE	int	DWGRP4_DW_AREA.PRODUCT_DIM	PACKAGE_SIZE	int
UPC.csv	UPC	varchar	DWGRP4_STAGING_AREA.Movement	UPC_CODE	varchar	DWGRP4_DW_AREA.PRODUCT_DIM	UPC_CODE	varchar

PRICE_DIMENSION LOAD MAPPING FROM FILE TO _DIM TABLE in PRODUCTION WAREHOUSE

File Used to Load	File Column	Datatype	Staging Table	Staging Column	Staging Column Datatype	Production Table	Production Column	Production Column Datatype
DONE.csv	PRICE	varchar	DWGRP4_STAGING_AREA.PRICE	PRIME_AMOUNT	decimal	DWGRP4_DW_AREA.PRICE_DIM	PRIME_AMOUNT	decimal

COUPON_DIMENSION LOAD MAPPING FROM FILE TO _DIM TABLE in PRODUCTION WAREHOUSE								
File Used to Load	File Column	Datatype	Staging Table	Staging Column	Staging Column Datatype	Production Table	Production Column	Production Column Datatype
CCOUNT.csv	COUPON_TYPE	varchar	DWGRP4_STAGING_AREA.COUPON	COUPON_TYPE	int	DWGRP4_DW_AREA.COUPON_DIM	COUPON_TYPE	int

Fact: dbo.STORESALE_FACT								
File Used to Load	File Column	Datatype	Staging Table	Staging Column	Staging Column Datatype	Production Table	Production Column	Production Column Datatype
Keys to be generated automatically in the sequential pattern			DWGRP4_STAGING_AREA.STORE	STORE_ID	int	DWGRP4_DW_AREA.STORESALE_FAC	STORE	int
			DWGRP4_STAGING_AREA.TIME	TIME_ID	int	DWGRP4_DW_AREA.STORESALE_FAC	UPC	int
			DWGRP4_STAGING_AREA.COUPON	COUPON_ID	int	DWGRP4_DW_AREA.STORESALE_FAC	WEEK	int
CCOUNT.csv	CUSTCOUN	varchar	CCOUNT	CUSTCOUN	int	DWGRP4_DW_AREA.STORESALE_FAC	MOVE	int
CCOUNT.csv	Coupon Values across the different products, CALCULATED FIELD	varchar	CCOUNT	Calculated Value	decimal	DWGRP4_DW_AREA.STORESALE_FAC	QTY	decimal

Fact: dbo.MOVEMENT_FACT								
File Used to Load	File Column	Datatype	Staging Table	Staging Column	Staging Column Datatype	Production Table	Production Column	Production Column Datatype
UPC.csv	MOVE	varchar	UPC	MOVE	int	DWGRP4_DW_AREA.MOVEMENT_FA	MOVEMENT	int
UPC.csv	QTY	varchar	UPC	QTY	int	DWGRP4_DW_AREA.MOVEMENT_FA	QUANTITY	int
UPC.csv	PRICE, MOVE, QTY	varchar	UPC	PRICE, MOVE, QTY	int	DWGRP4_DW_AREA.MOVEMENT_FA	SALES	int
UPC.csv	PROFIT	varchar	UPC	PROFIT	decimal	DWGRP4_DW_AREA.MOVEMENT_FA	PROFIT	decimal
UPC.csv	PRICE, MOVE, QTY, PROFIT	varchar	UPC	PRICE, MOVE, QTY, PROFIT	decimal	DWGRP4_DW_AREA.MOVEMENT_FA	ACTUAL_PROFIT	Calculated Field

J. SQL Statements used for the ETL operations

The below are the SQL scripts used to create the destination tables.

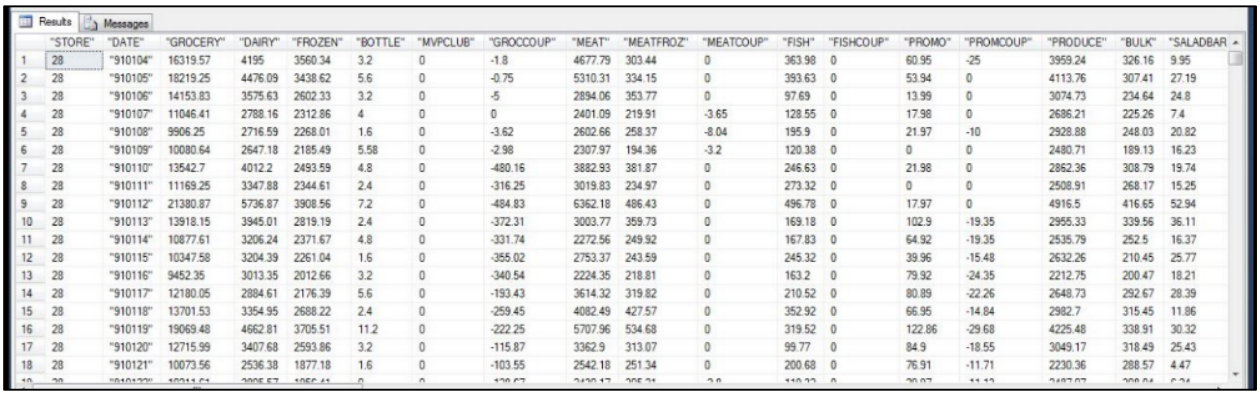
```
CREATE TABLE [COUPON_DIM](
	[COUPON_ID] [int] IDENTITY(1,1) NOT NULL,
	[COUPON_NAME] [char](10) NULL
) ;


CREATE TABLE [PRICE_DIM](
	[PRICE_ID] [int] IDENTITY(1,1) NOT NULL,
	[PRICE_AMOUNT] [decimal](10, 2) NULL
);


CREATE TABLE [PRODUCT_DIM](
	[PRODUCT_ID] [int] IDENTITY(1,1) NOT NULL,
	[CATEGORY] [varchar](100) NULL,
	[PRODUCT_DESC] [varchar](100) NULL,
	[UPC_CODE] [bigint] NULL,
```

```
	[CASE_SIZE] [int] NULL,
	[PACKAGE_SIZE] [varchar](100) NULL
) ;

CREATE TABLE [TIME_DIM](
	[TIME_ID] [int] IDENTITY(1,1) NOT NULL,
	[YEAR] [int] NULL,
	[WEEK] [int] NULL
);

CREATE TABLE [PRICE_DIM](
	[PRICE_ID] [int] IDENTITY(1,1) NOT NULL,
	[PRICE_AMOUNT] [decimal](10, 2) NULL);

CREATE TABLE [STORESALE_FACT](
	[STORE_ID] [int] NULL,
	[TIME_ID] [int] NULL,
	[COUPON_ID] [int] NULL,
	[CUSTOMER_COUNT] [int] NULL,
	[COUPON_VALUE] [decimal](10, 2) NULL
) ;

CREATE TABLE [SALES_FACT](
	[STORE_ID] [int] NULL,
	[TIME_ID] [int] NULL,
	[PRODUCT_ID] [int] NULL,
	[PRICE_ID] [int] NULL,
	[MOVEMENT] [int] NULL,
	[QUANTITY] [int] NULL,
	[SALES] [decimal](10, 2) NULL,
	[PROFIT] [decimal](10, 2) NULL,
	[ACTUAL_PROFIT] [decimal](10, 2) NULL) ;
```

K. Before After table screenshots

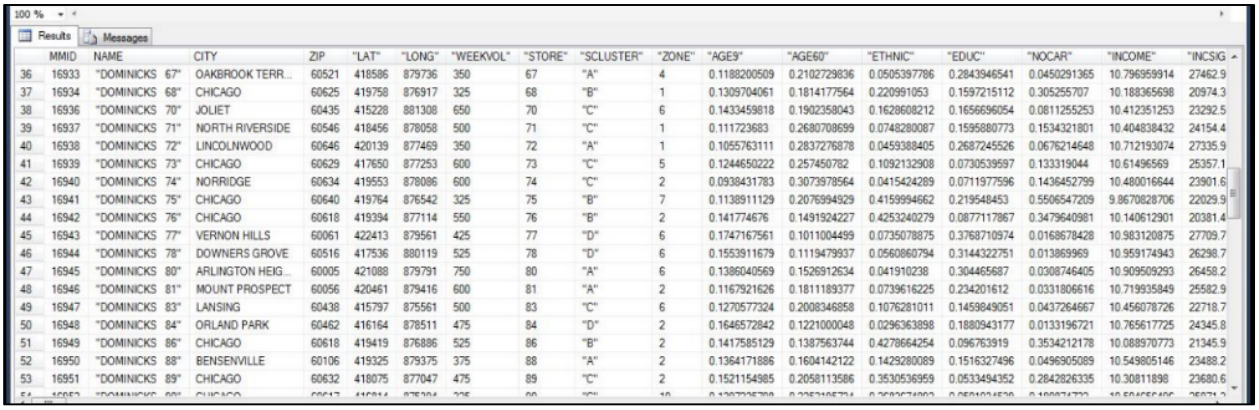
BEFORE TABLES (Stage tables)

Table: DWGRP4_STAGING_AREA.CCOUNT:



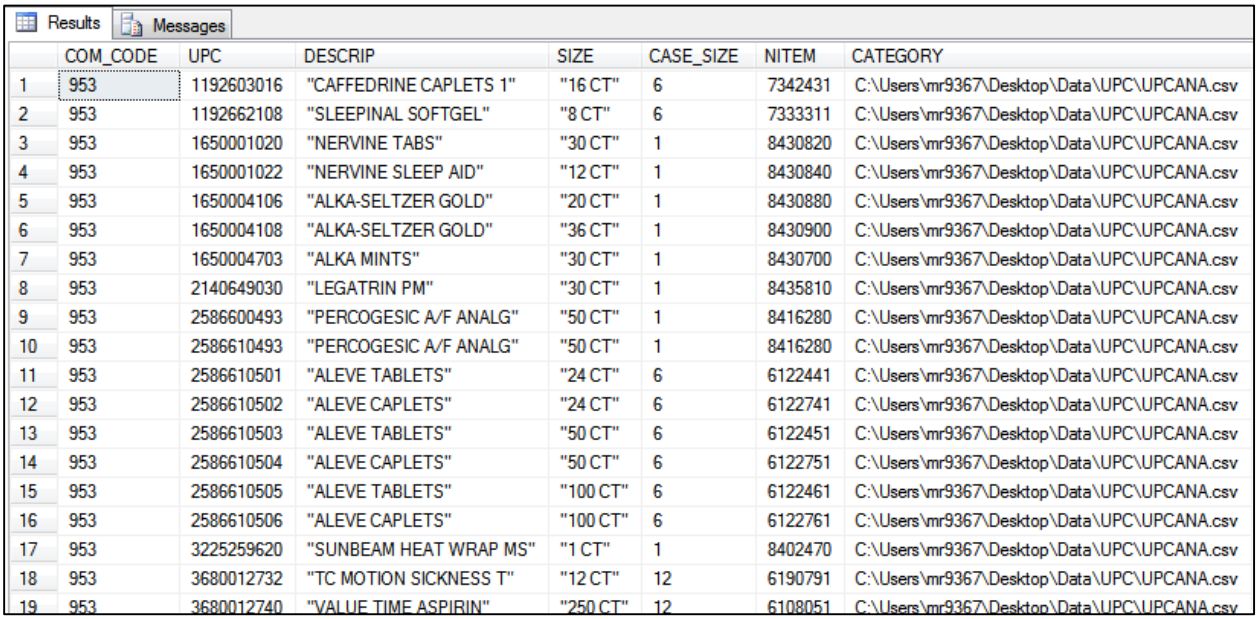
	"STORE"	"DATE"	"GROCERY"	"DAIRY"	"FROZEN"	"BOTTLE"	"MVPCLUB"	"GROCCOUP"	"MEAT"	"MEATFROZ"	"MEATCOUP"	"FISH"	"FISHCOUP"	"PROMO"	"PROMCOUP"	"PRODUCE"	"BULK"	"SALADBAR
1	28	"910104"	16319.57	4195	3560.34	3.2	0	-1.8	4677.79	303.44	0	363.98	0	60.95	-25	3959.24	326.16	9.95
2	28	"910105"	18219.25	4476.09	3438.62	5.6	0	-0.75	5310.31	334.15	0	393.63	0	53.94	0	4113.76	307.41	27.19
3	28	"910106"	14153.83	3575.63	2602.33	3.2	0	-5	2894.06	353.77	0	97.69	0	13.99	0	3074.73	234.64	24.8
4	28	"910107"	11046.41	2788.16	2312.86	4	0	0	2401.09	219.91	-3.65	128.55	0	17.98	0	2686.21	225.26	7.4
5	28	"910108"	9906.25	2716.59	2268.01	1.6	0	-3.62	2602.66	258.37	-8.04	195.9	0	21.97	-10	2928.88	248.03	20.82
6	28	"910109"	10080.64	2647.18	2185.49	5.58	0	-2.98	2307.97	194.36	-3.2	120.38	0	0	0	2480.71	189.13	16.23
7	28	"910110"	13542.7	4012.2	2493.59	4.8	0	-480.16	3882.93	381.87	0	246.63	0	21.98	0	2862.36	308.79	19.74
8	28	"910111"	11169.25	3347.88	2344.61	2.4	0	-316.25	3019.83	234.97	0	273.32	0	0	0	2508.91	268.17	15.25
9	28	"910112"	21380.87	5736.87	3908.56	7.2	0	-484.83	6362.18	486.43	0	496.78	0	17.97	0	4916.5	416.65	52.94
10	28	"910113"	13918.15	3945.01	2819.19	2.4	0	-372.31	3003.77	359.73	0	169.18	0	102.9	-19.35	2955.33	339.56	36.11
11	28	"910114"	10877.61	3206.24	2371.67	4.8	0	-331.74	2272.56	249.92	0	167.83	0	64.92	-19.35	2535.79	252.5	16.37
12	28	"910115"	10347.58	3204.39	2261.04	1.6	0	-355.02	2753.37	243.59	0	245.32	0	39.96	-15.48	2632.26	210.45	25.77
13	28	"910116"	9452.35	3013.35	2012.66	3.2	0	-340.54	2224.35	218.81	0	163.2	0	79.92	-24.35	2212.75	200.47	18.21
14	28	"910117"	12180.05	2884.61	2176.39	5.6	0	-193.43	3614.32	319.82	0	210.52	0	80.89	-22.26	2648.73	292.67	28.39
15	28	"910118"	13701.53	3354.95	2688.22	2.4	0	-259.45	4082.49	427.57	0	352.92	0	66.95	-14.84	2982.7	315.45	11.86
16	28	"910119"	19069.48	4662.81	3705.51	11.2	0	-222.25	5707.96	534.68	0	319.52	0	122.86	-29.68	4225.48	338.91	30.32
17	28	"910120"	12715.99	3407.68	2593.86	3.2	0	-115.87	3362.9	313.07	0	99.77	0	84.9	-18.55	3049.17	318.49	25.43
18	28	"910121"	10073.56	2536.38	1877.18	1.6	0	-103.55	2542.18	251.34	0	200.68	0	76.91	-11.71	2230.36	288.57	4.47

Table: DWGRP4_STAGING_AREA.DEMO:



	MMID	NAME	CITY	ZIP	"LAT"	"LONG"	"WEEKVOL"	"STORE"	"SCLUSTER"	"ZONE"	"AGE9"	"AGE60"	"ETHNIC"	"EDUC"	"NOCAR"	"INCOME"	"INCSIG
36	16933	"DOMINICKS 67"	OAKBROOK TERR...	60521	418586	879736	350	67	"A"	4	0.1188200509	0.2102729836	0.0505397786	0.2843946541	0.0450291365	10.796959914	27462.9
37	16934	"DOMINICKS 68"	CHICAGO	60625	419758	876917	325	68	"B"	1	0.1309704061	0.1814177564	0.220991053	0.1597215112	0.305255707	10.188365698	20974.3
38	16936	"DOMINICKS 70"	JOLIET	60435	415228	881308	650	70	"C"	6	0.1433459818	0.1902358043	0.1628608212	0.1656696054	0.0811255253	10.412351253	23292.5
39	16937	"DOMINICKS 71"	NORTH RIVERSIDE	60546	418456	878058	500	71	"C"	1	0.111723683	0.2680708699	0.0748280087	0.1595880773	0.1534321801	10.404838432	24154.4
40	16938	"DOMINICKS 72"	LINCOLNWOOD	60646	420139	877469	350	72	"A"	1	0.1055763111	0.2837276878	0.0459388405	0.2687245526	0.0676214648	10.712193074	27335.9
41	16939	"DOMINICKS 73"	CHICAGO	60629	417650	877253	600	73	"C"	5	0.1244650222	0.257450782	0.1092132908	0.0730539597	0.133319044	10.61496569	25357.1
42	16940	"DOMINICKS 74"	NORRIDGE	60634	419553	878086	600	74	"C"	2	0.0938431783	0.3073978564	0.0415424289	0.0711977596	0.1436452799	10.480016644	23901.6
43	16941	"DOMINICKS 75"	CHICAGO	60640	419764	876542	325	75	"B"	7	0.1138911129	0.2076994929	0.4159994662	0.219548453	0.5506547209	9.8670828706	22029.9
44	16942	"DOMINICKS 76"	CHICAGO	60618	419394	877114	550	76	"B"	2	0.141774676	0.1491924227	0.4253240279	0.0877117867	0.3479640981	10.140612901	20381.4
45	16943	"DOMINICKS 77"	VERNON HILLS	60061	422413	879561	425	77	"D"	6	0.1747167561	0.1011004499	0.0735078875	0.3768710974	0.0168678428	10.983120875	27709.7
46	16944	"DOMINICKS 78"	DOWNERS GROVE	60516	417536	880119	525	78	"D"	6	0.1553911679	0.1119479937	0.0560860794	0.3144322751	0.013869969	10.959174943	26298.7
47	16945	"DOMINICKS 80"	ARLINGTON HEIG...	60005	421088	879791	750	80	"A"	6	0.1386040569	0.1526912634	0.041910238	0.304465687	0.0308746405	10.909509293	26458.2
48	16946	"DOMINICKS 81"	MOUNT PROSPECT	60056	420461	879416	600	81	"A"	2	0.1167921626	0.1811189377	0.0739616225	0.234201612	0.0331806616	10.719935849	25582.9
49	16947	"DOMINICKS 83"	LANSING	60438	415797	875561	500	83	"C"	6	0.1270577324	0.2008346858	0.1076281011	0.1459849051	0.0437264667	10.456078726	22718.7
50	16948	"DOMINICKS 84"	ORLAND PARK	60462	416164	878511	475	84	"D"	2	0.1646572842	0.1221000048	0.0296363898	0.1880943177	0.0133196721	10.765617725	24345.8
51	16949	"DOMINICKS 86"	CHICAGO	60618	419419	876886	525	86	"B"	2	0.1417585129	0.1387563744	0.4278664254	0.096763919	0.3534212178	10.088970773	21345.9
52	16950	"DOMINICKS 88"	BENSENVILLE	60106	419325	879375	375	88	"A"	2	0.1364171886	0.1604142122	0.1429280089	0.1516327496	0.0496905089	10.549805146	23488.2
53	16951	"DOMINICKS 89"	CHICAGO	60632	418075	877047	475	89	"C"	2	0.1521154985	0.2058113586	0.3530536959	0.0533494352	0.2842826335	10.30811898	23680.6

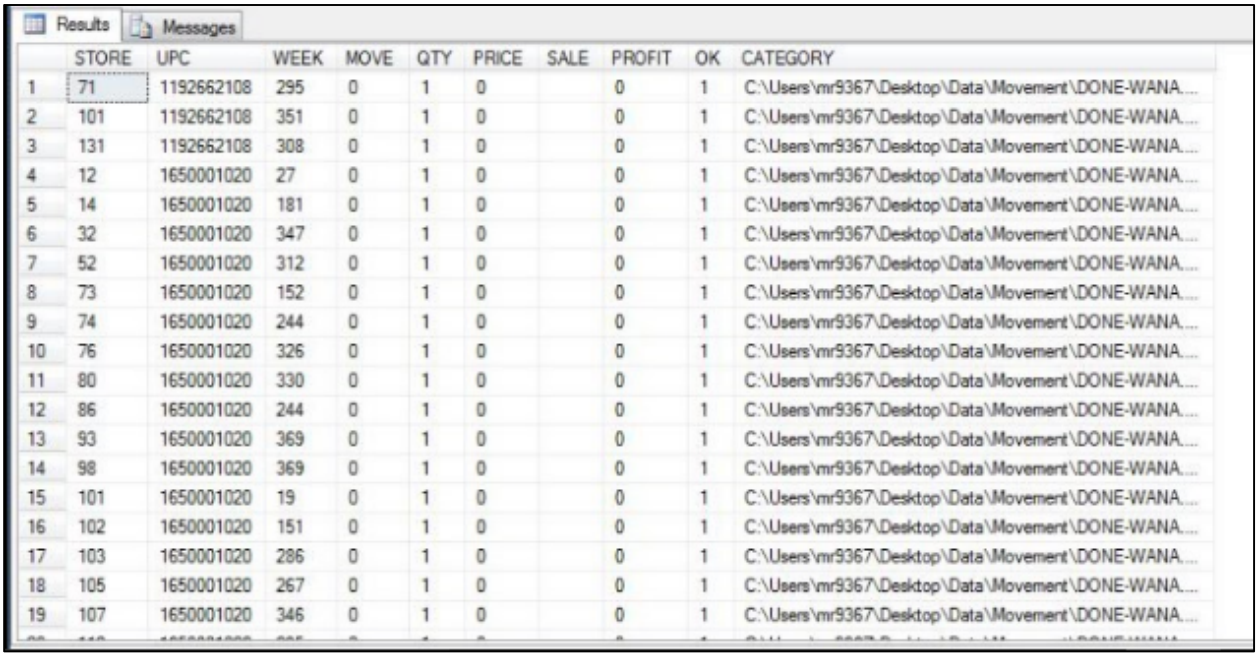
Table: DWGRP4_STAGING_AREA.UPC:



Results | Messages

	COM_CODE	UPC	DESCRIP	SIZE	CASE_SIZE	NITEM	CATEGORY
1	953	1192603016	"CAFFEDRINE CAPLETS 1"	"16 CT"	6	7342431	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
2	953	1192662108	"SLEEPINAL SOFTGEL"	"8 CT"	6	7333311	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
3	953	1650001020	"NERVINE TABS"	"30 CT"	1	8430820	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
4	953	1650001022	"NERVINE SLEEP AID"	"12 CT"	1	8430840	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
5	953	1650004106	"ALKA-SELTZER GOLD"	"20 CT"	1	8430880	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
6	953	1650004108	"ALKA-SELTZER GOLD"	"36 CT"	1	8430900	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
7	953	1650004703	"ALKA MINTS"	"30 CT"	1	8430700	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
8	953	2140649030	"LEGATRIN PM"	"30 CT"	1	8435810	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
9	953	2586600493	"PERCOGESIC A/F ANALG"	"50 CT"	1	8416280	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
10	953	2586610493	"PERCOGESIC A/F ANALG"	"50 CT"	1	8416280	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
11	953	2586610501	"ALEVE TABLETS"	"24 CT"	6	6122441	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
12	953	2586610502	"ALEVE CAPLETS"	"24 CT"	6	6122741	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
13	953	2586610503	"ALEVE TABLETS"	"50 CT"	6	6122451	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
14	953	2586610504	"ALEVE CAPLETS"	"50 CT"	6	6122751	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
15	953	2586610505	"ALEVE TABLETS"	"100 CT"	6	6122461	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
16	953	2586610506	"ALEVE CAPLETS"	"100 CT"	6	6122761	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
17	953	3225259620	"SUNBEAM HEAT WRAP MS"	"1 CT"	1	8402470	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
18	953	3680012732	"TC MOTION SICKNESS T"	"12 CT"	12	6190791	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv
19	953	3680012740	"VALUE TIME ASPIRIN"	"250 CT"	12	6108051	C:\Users\mr9367\Desktop\Data\UPC\UPCANA.csv

Table: DWGRP4_STAGING_AREA.MOVEMENT:



Results | Messages

	STORE	UPC	WEEK	MOVE	QTY	PRICE	SALE	PROFIT	OK	CATEGORY
1	71	1192662108	295	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA...
2	101	1192662108	351	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA...
3	131	1192662108	308	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA...
4	12	1650001020	27	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA...
5	14	1650001020	181	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA...
6	32	1650001020	347	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA...
7	52	1650001020	312	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA...
8	73	1650001020	152	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA...
9	74	1650001020	244	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA...
10	76	1650001020	326	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA...
11	80	1650001020	330	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA...
12	86	1650001020	244	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA...
13	93	1650001020	369	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA...
14	98	1650001020	369	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA...
15	101	1650001020	19	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA...
16	102	1650001020	151	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA...
17	103	1650001020	286	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA...
18	105	1650001020	267	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA...
19	107	1650001020	346	0	1	0		0	1	C:\Users\mr9367\Desktop\Data\Movement\DONE-WANA...

AFTER TABLES (Data Mart dimension and fact tables in PROD)

Table: STORE_DIM:

	STORE_ID	ZONE	CITY	INCOME	STORE_NUMBER	INCOMELEVEL
1	1	1	RIVER FOREST	10.553205	2	MEDIUM
2	2	2	PARK RIDGE	10.646971	4	MEDIUM
3	3	2	PALATINE	10.922370	5	HIGH
4	4	5	OAK LAWN	10.597009	8	MEDIUM
5	5	2	MORTON GROVE	10.787151	9	MEDIUM
6	6	7	CHICAGO	9.996659	12	LOW
7	7	1	GLENVIEW	11.043929	14	HIGH
8	8	5	RIVER GROVE	10.391975	18	LOW
9	9	6	HANOVER PARK	10.716193	21	MEDIUM
10	10	2	MOUNT PROSPECT	10.798534	28	MEDIUM
11	11	1	PARK RIDGE	10.674475	32	MEDIUM
12	12	7	CHICAGO	10.345927	33	LOW
13	13	6	BRIDGEVIEW	10.550250	40	MEDIUM
14	14	2	WESTERN SPRINGS	10.869158	44	HIGH
15	15	2	WHEELING	10.745377	45	MEDIUM
16	16	2	ADDISON	10.635326	47	MEDIUM
17	17	2	SCHAUMBURG	10.756028	48	MEDIUM
18	18	2	DOWNERS GROVE	10.806753	49	HIGH
19	19	2	HICKORY HILLS	10.589307	50	MEDIUM

Table: PRODUCT_DIM:

	PRODUCT_ID	CATEGORY	PRODUCT_DESC	UPC_CODE	CASE_SIZE	PACKAGE_SIZE
1	1	ANA	CAFFEDRINECAPLETS1	1192603016	6	16 CT
2	2	ANA	SLEEPINALSOFTGEL	1192662108	6	8 CT
3	3	ANA	NERVINETABS	1650001020	1	30 CT
4	4	ANA	NERVINESLEEPAID	1650001022	1	12 CT
5	5	ANA	ALKASELTZERGOLD	1650004106	1	20 CT
6	6	ANA	ALKASELTZERGOLD	1650004108	1	36 CT
7	7	ANA	ALKAMINTS	1650004703	1	30 CT
8	8	ANA	LEGATRINPM	2140649030	1	30 CT
9	9	ANA	PERCOGESICAFANALG	2586600493	1	50 CT
10	10	ANA	PERCOGESICAFANALG	2586610493	1	50 CT
11	11	ANA	ALEVETABLETS	2586610501	6	24 CT
12	12	ANA	ALEVECAPLETS	2586610502	6	24 CT
13	13	ANA	ALEVETABLETS	2586610503	6	50 CT
14	14	ANA	ALEVECAPLETS	2586610504	6	50 CT
15	15	ANA	ALEVETABLETS	2586610505	6	100 CT
16	16	ANA	ALEVECAPLETS	2586610506	6	100 CT
17	17	ANA	SUNBEAMHEATWRAPMS	3225259620	1	1 CT
18	18	ANA	TCMOTIONSICKNESST	3680012732	12	12 CT
19	19	ANA	VALUETIMEASPIRIN	3680012740	12	250 CT

Table: PRICE_DIM:

0 %

Results | Messages

	PRICE_ID	PRICE_AMOUNT
	11	1.09
	12	1.10
	13	1.45
	14	2.00
	15	2.11
	16	2.45
	17	2.70
	18	2.75
	19	3.40

Table: COUPON_DIM:

	COUPON_ID	COUPON_NAME
1	1	GROCCOUP
2	2	MEATCOUP
3	3	FISHCOUP
4	4	PROMCOUP
5	5	PRODCOUP
6	6	BULKCOUP
7	7	SALCOUP
8	8	FLORCOUP
9	9	DELICOUP
10	10	PHARCOUP
11	11	GMCOUP
12	12	VIDCOUP
13	13	MISCSCP
14	14	MANCOUP
15	15	CUSTCOUN
16	16	FTGCCOUP
17	17	FTGICOUP
18	18	DAIRCOUP
19	19	FROZCOUP

Table: TIME_DIM:

	TIME_ID	YEAR	WEEK
1	1	1989	1
2	2	1989	2
3	3	1989	3
4	4	1989	4
5	5	1989	5
6	6	1989	6
7	7	1989	7
8	8	1989	8
9	9	1989	9
10	10	1989	10
11	11	1989	11
12	12	1989	12
13	13	1989	13
14	14	1989	14
15	15	1990	15
16	16	1990	16
17	17	1990	17
18	18	1990	18
19	19	1990	19

Table: STORESALES_FACT:

	STORE_ID	TIME_ID	COUPON_ID	CUSTOMER_COUNT	COUPON_VALUE
1	1	1	24	13870	0.00
2	1	1	6	13870	0.00
3	1	1	22	13870	0.00
4	1	1	18	13870	0.00
5	1	1	9	13870	-259.00
6	1	1	3	13870	0.00
7	1	1	8	13870	-3.00
8	1	1	19	13870	0.00
9	1	1	16	13870	0.00
10	1	1	17	13870	0.00
11	1	1	11	13870	-176.00
12	1	1	1	13870	-3268.00
13	1	1	20	13870	0.00
14	1	1	25	13870	0.00
15	1	1	14	13870	0.00
16	1	1	2	13870	0.00
17	1	1	13	13870	-80.00
18	1	1	10	13870	0.00
19	1	1	21	13870	0.00

Table: SALES_FACT:

	STORE_ID	TIME_ID	PRODUCT_ID	PRICE_ID	MOVEMENT	QUANTITY	SALES	PROFIT	ACTUAL_PROFIT
1	6	274	642	11	0	1	0.00	0.00	0.00
2	6	275	642	12	1	1	5.00	8.18	48.99
3	6	276	642	11	0	1	0.00	0.00	0.00
4	6	277	642	11	0	1	0.00	0.00	0.00
5	6	278	642	11	0	1	0.00	0.00	0.00
6	8	248	642	11	0	1	0.00	0.00	0.00
7	8	249	642	11	0	1	0.00	0.00	0.00
8	8	250	642	11	0	1	0.00	0.00	0.00
9	8	251	642	12	2	1	11.00	8.18	97.99
10	8	252	642	11	0	1	0.00	0.00	0.00
11	8	253	642	11	0	1	0.00	0.00	0.00
12	8	254	642	11	0	1	0.00	0.00	0.00
13	8	255	642	11	0	1	0.00	0.00	0.00
14	8	256	642	11	0	1	0.00	0.00	0.00
15	8	257	642	11	0	1	0.00	0.00	0.00
16	8	258	642	11	0	1	0.00	0.00	0.00
17	8	259	642	11	0	1	0.00	0.00	0.00
18	8	260	642	11	0	1	0.00	0.00	0.00
19	8	261	642	11	0	1	0.00	0.00	0.00

9. BI Reporting:

BQ2: Identify the most concerned areas of losses for each shop for Dominick's Fine Food (DFF) week-wise.

The business question aims at identifying the loss areas for the store so as to help management towards a better decision making system. It would show the product wise loss areas with respect to the stores and hence a proper decision could be taken to implement strategies for a particular product promotion in the store area.

a. Target report to satisfy business question:

The report includes a comparison of the product wise loss areas for the stores directly showing the concerned areas with a dip of profit percentage. The target report is plotted with Store data i.e. STORE_NUMBER on x axis and LOSS_PERCENTAGE on y-axis with a generic plot over the UPC_NUMBER. The direct plot of UPC_NUMBER variation over the LOSS_PERCENTAGE values in accordance to the STORE_NUMBER represents the relationship for the loss areas for each of these products.

b. Report attributes and mapping from data mart:

The product wise loss percentage is calculated using the Sales formula as given in the Dominick's web site:

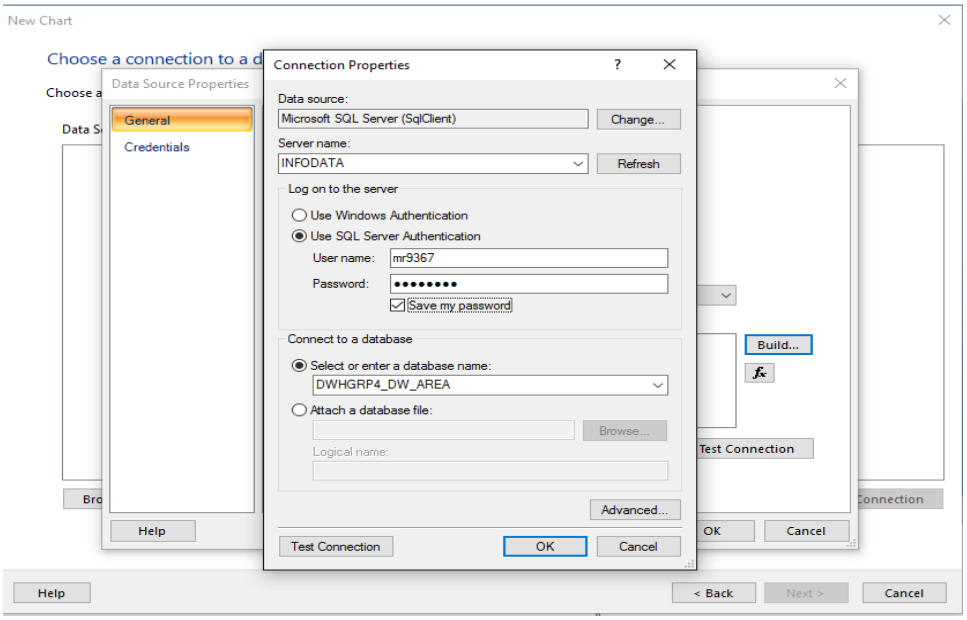
Sales= Price*Move / Qty

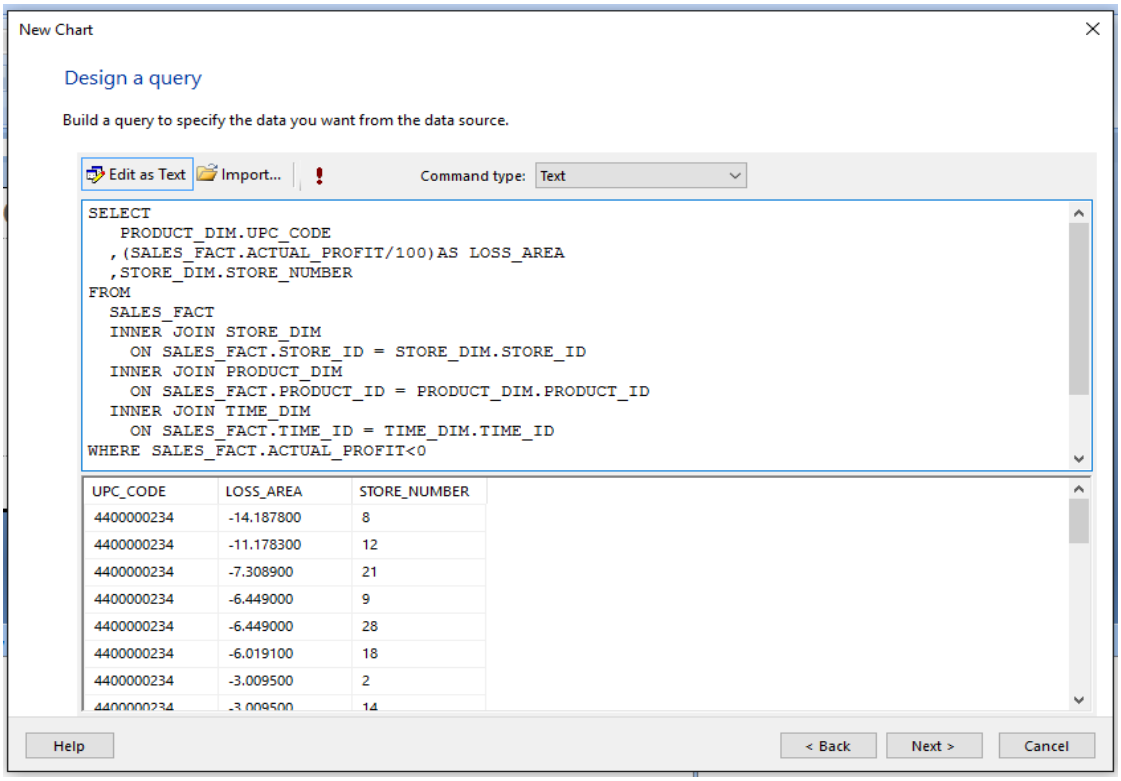
The query takes the ACTUAL_PROFIT volume and divides it by 100 to calculate the profit percentage because the column ACTUAL_PROFIT is calculated using the values in columns SALES and PROFIT. The column PROFIT simply represents the profit percentage. Hence, in order to calculate an accurate profit value, the ACTUAL_PROFIT amount should be divided by 100 for getting the actual profit percentage.

REPORT ATTRIBUTES		SOURCE DATAMART		
Attribute Name	Attribute Description	Table Name	Column Name	Summarization (if any)
Week	The week field	TIME_DIM	WEEK	
Product Item	Each individual product item for which the loss has occured	PRODUCT_DIM	UPC#	
Loss Value	Calculated using the loss percentage and sales for the product	SALES_FACT	ACTUAL_PROFIT	The data with a negative value represents the loss occurrence for the entity.

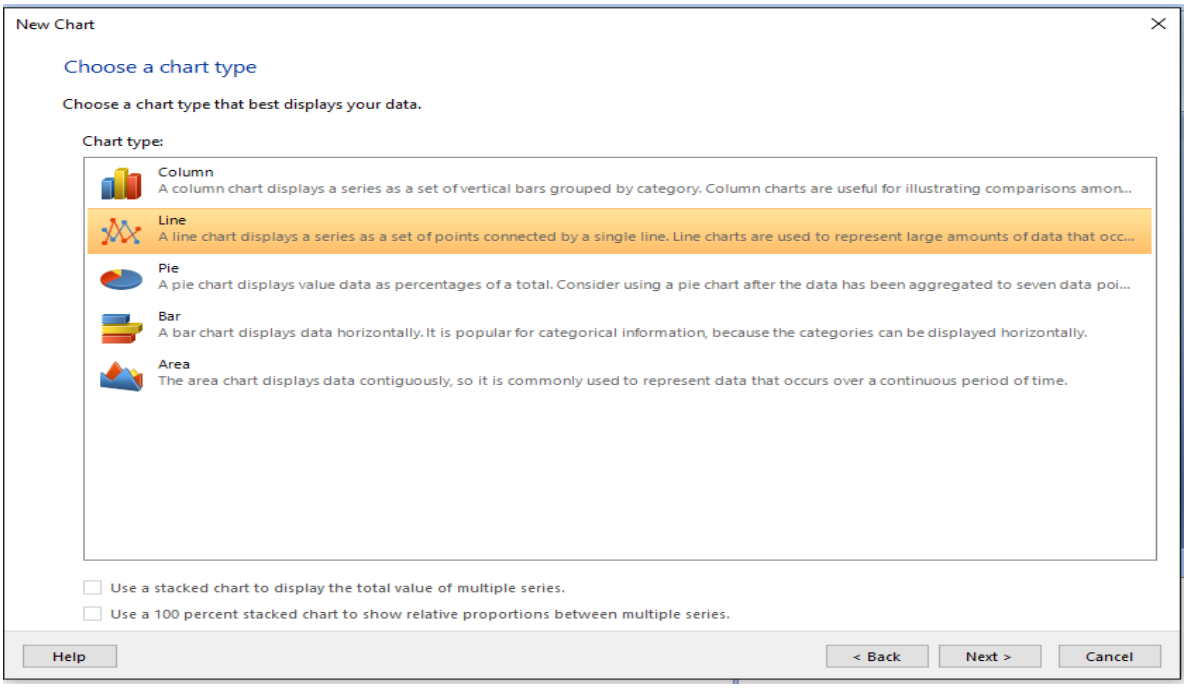
c. Report Implementation using Report Builder 3.0:

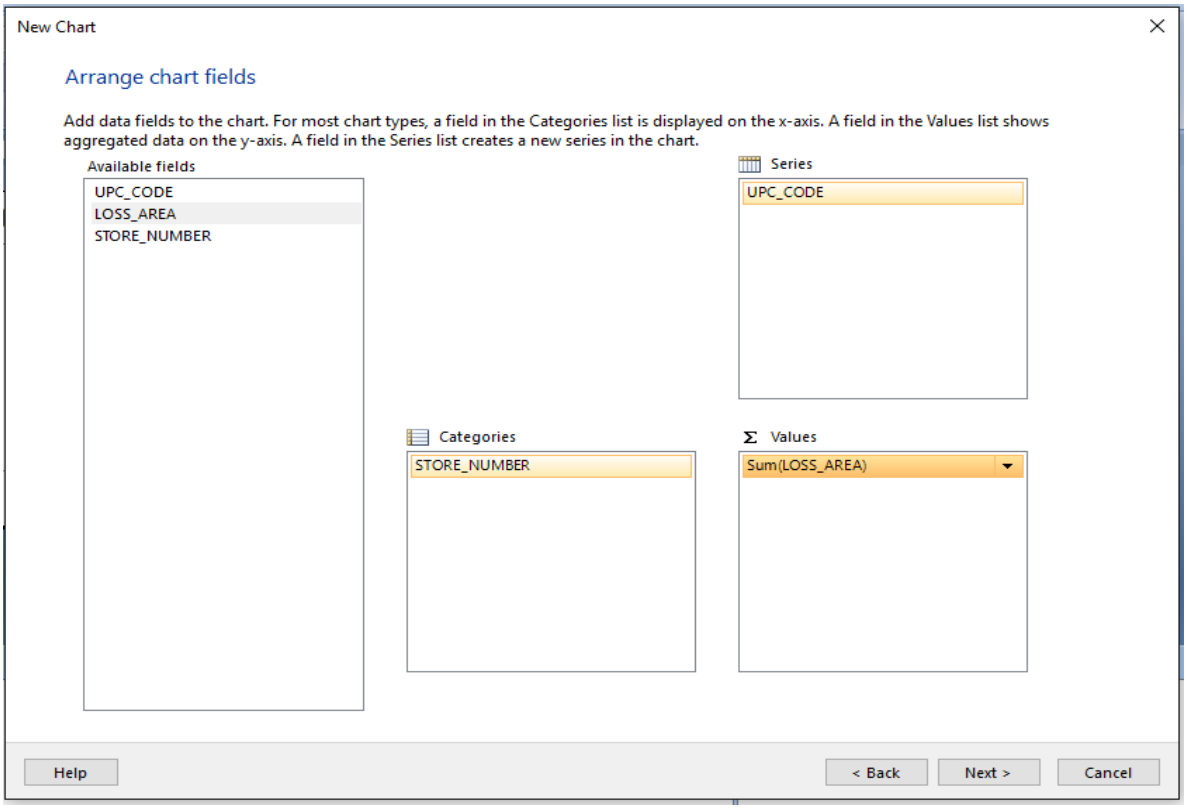
The mapping builds up related to report build up can be represented using the below screenshots:





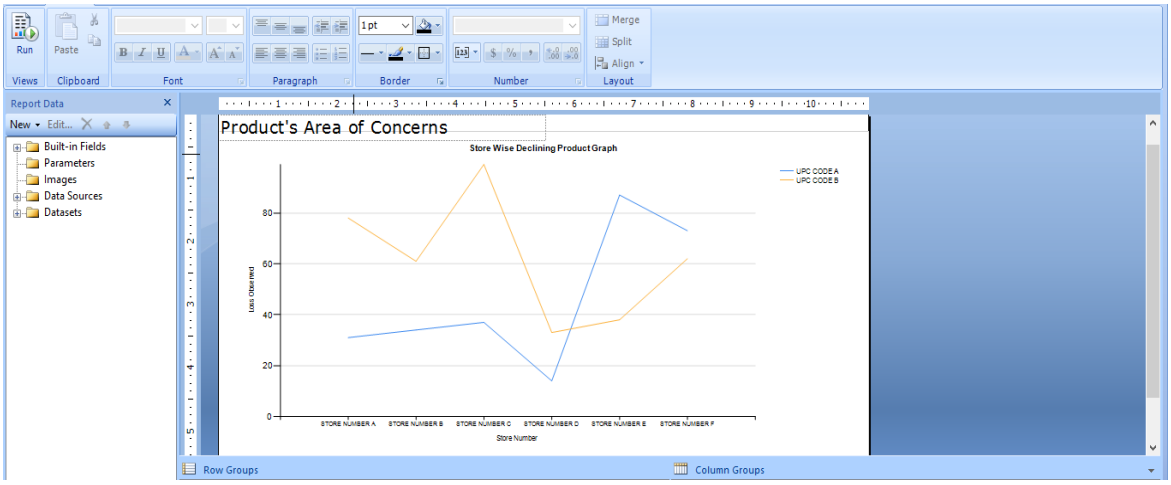
Please remember to apply the condition of ACTUAL_PROFIT<0 so as to obtain only the areas with a loss percentage indicated by a negative amount.



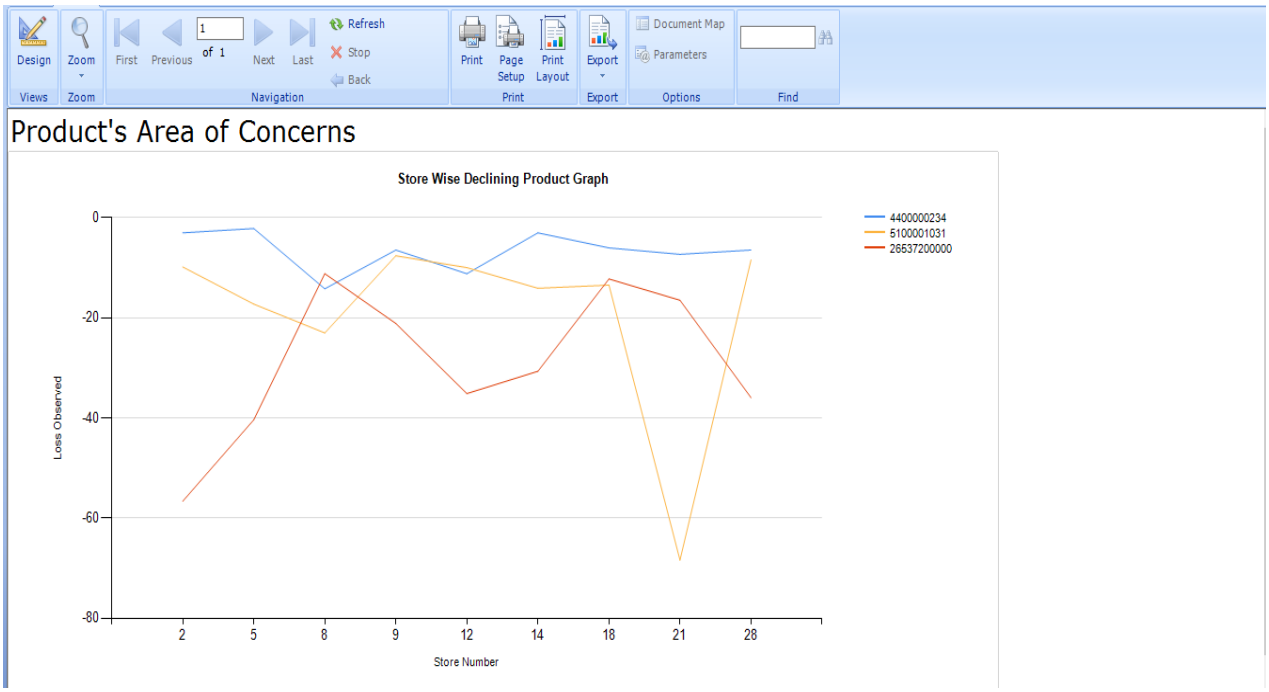


d. Report graph Template:

The report template formed is as under:



e. Reports Output:



This graph represents the Store wise trend for the product depicted by the UPC#. As can be seen from the graph, the UPC# 5100001031 has a steep decrease for the store number 21 after going fairly well with other store area. Hence, the business can implement strategies for increasing the product sales in the area of store 21. Similarly, the store 2 had a very low volume for the product 2653720000 which is again an area of concern that business can focus on for this particular store.

In a nutshell, this business questions represents the drop in the product sales for each of the store where the store can focus on improving the sales strategies for that particular product.

BQ3: Generate the report depicting the effect of coupon introduction for increasing the customer count in the store.

The business question aims at comparing the coupon value against the customer number visiting the store. This report fetches a direct report to check whether the increase in release of coupon cards to the customers actually attracts them to the stores or not. Based upon the output of this report, the business can work on formulizing the strategies to give coupon cards or other sales discounts to the customer.

A. Target report to satisfy business question:

The target report for this business case is split into two different comparisons:

i. Week wise Customer visit to the Store.
ii. Coupon sales by the store with ongoing weeks.

Sample Source Data for Chart:

Coupons Redeemed Value per week for each store number:

Week/ Store Number	21	25	28	32
1	-6070.74	-2473.85	-5605.75	-2874.07
2	-648.77	-411.58	-594.97	3166.71
3	-29363.05	-10363.89	-17860.01	-25720.1
4	-974.9	-238.05	-615.42	4241.26
5	-1676.9	-1029.84	-1308.13	1704.03

Customer visited to the different store for each week:

Week/ Store Number	21	25	28	32
1	14229	10356	13270	27568
2	14751	12439	13899	29157
3	14680	9866	13581	27634
4	14707	11178	13810	28023
5	14757	10737	13454	27922

The two reports obtained out of this business question can be used in parallel to compare if the coupon sales increase is affecting the influx of the customer to the store. The dimensions DIM_COUPON, DIM_STORE and SALES_FACT have been used to map this relationship.

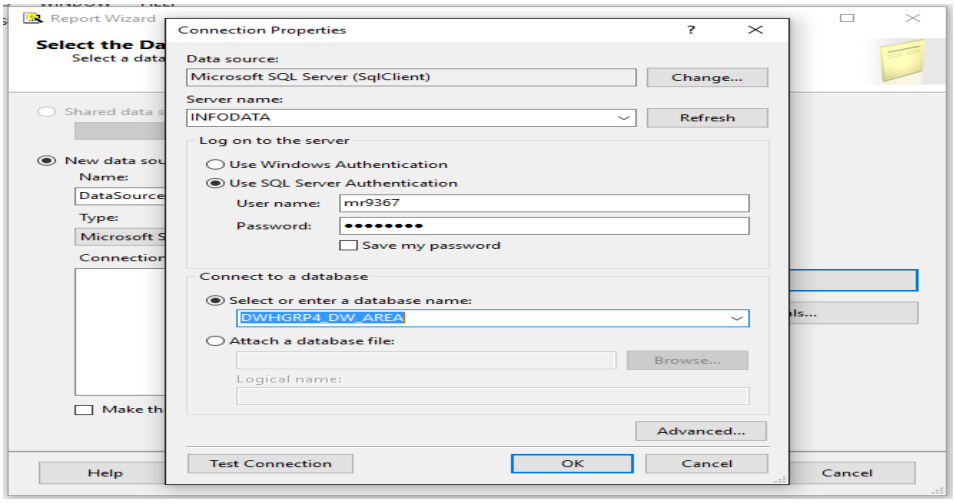
b. Report attributes and mapping from data mart:

To map the relationship between coupon sales and customer count, we used existing dimension tables and fact tables with SSIS server to build the report depicting this relationship.

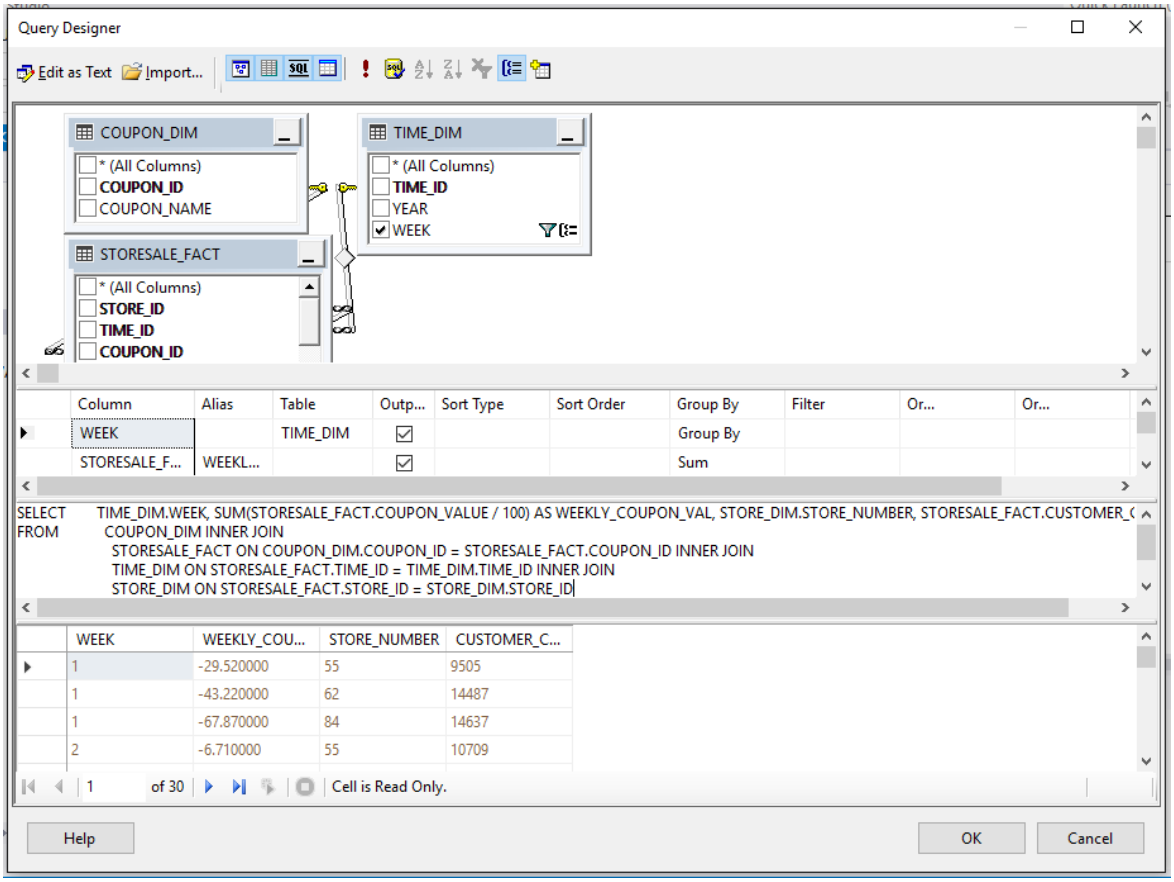
REPORT ATTRIBUTES		SOURCE DATAMART		
Attribute Name	Attribute Description	Table Name	Column Name	Summarization (if any)
Week	The week field	TIME_DIM	WEEK	
Store Number	Store Id for the different stores in DFF	STORE_DIM	STORE_NUMBER	
Coupon Value	The value for the coupon amount released by the store to the customers	STORESALES_FACT	COUPON_VALUE	Calculated using the coupon value of each type of food products introduced by the store
Customer Count	Customers visiting the store over the week	STORESALES_FACT	CUSTOMER_COUNT	The number of the customer visiting the store

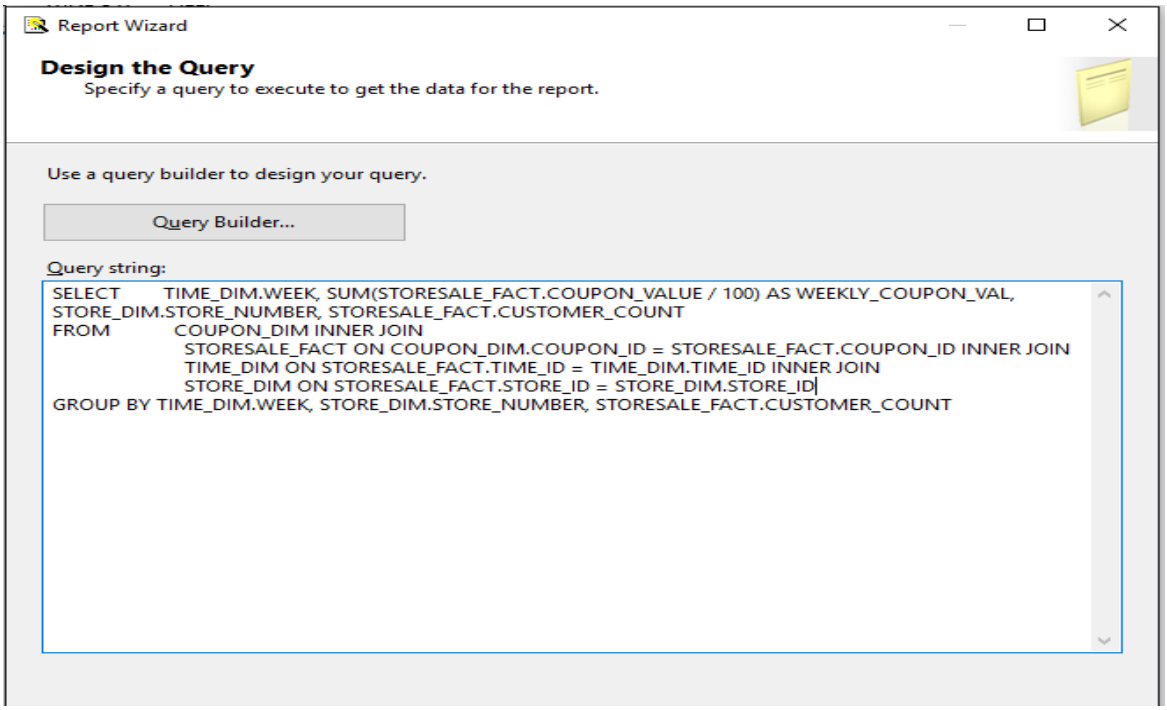
c. Report Implementation - SSRS:

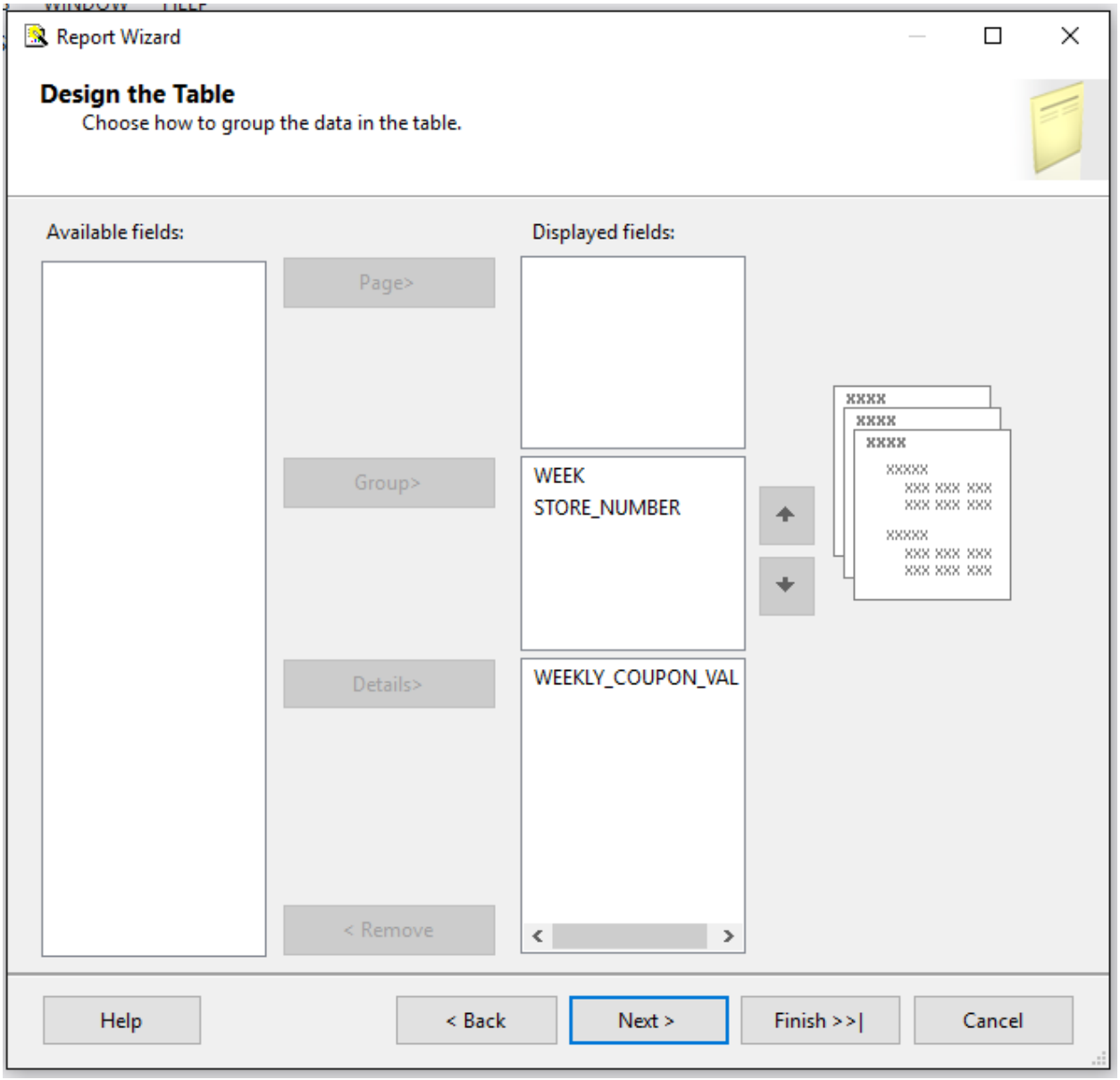
A series of steps for the can be represented as below:

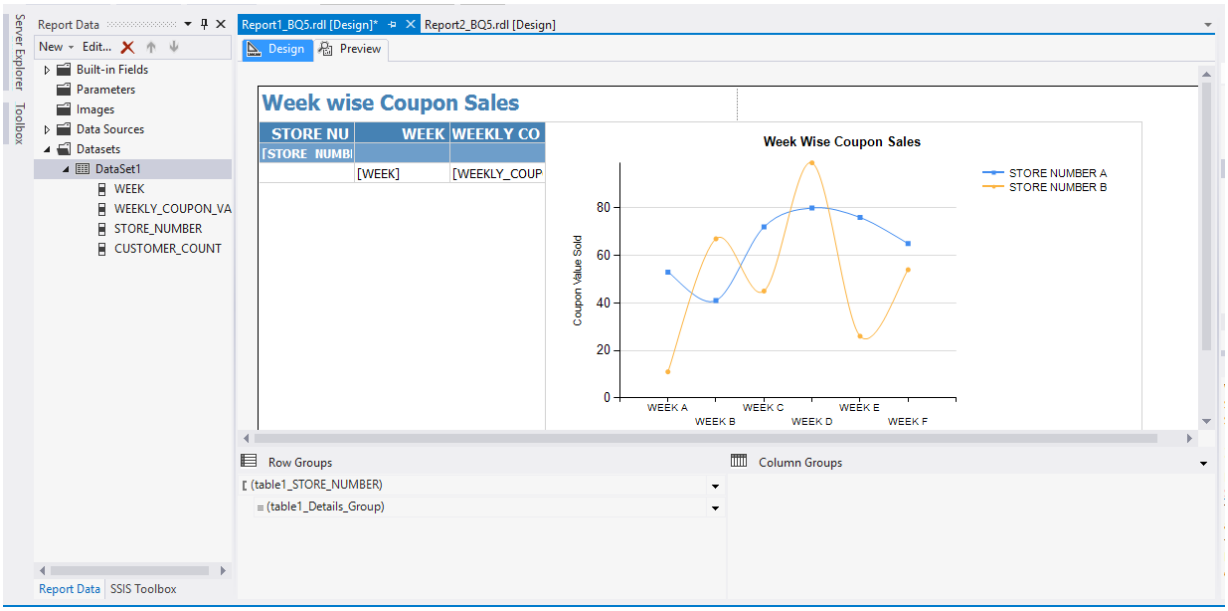


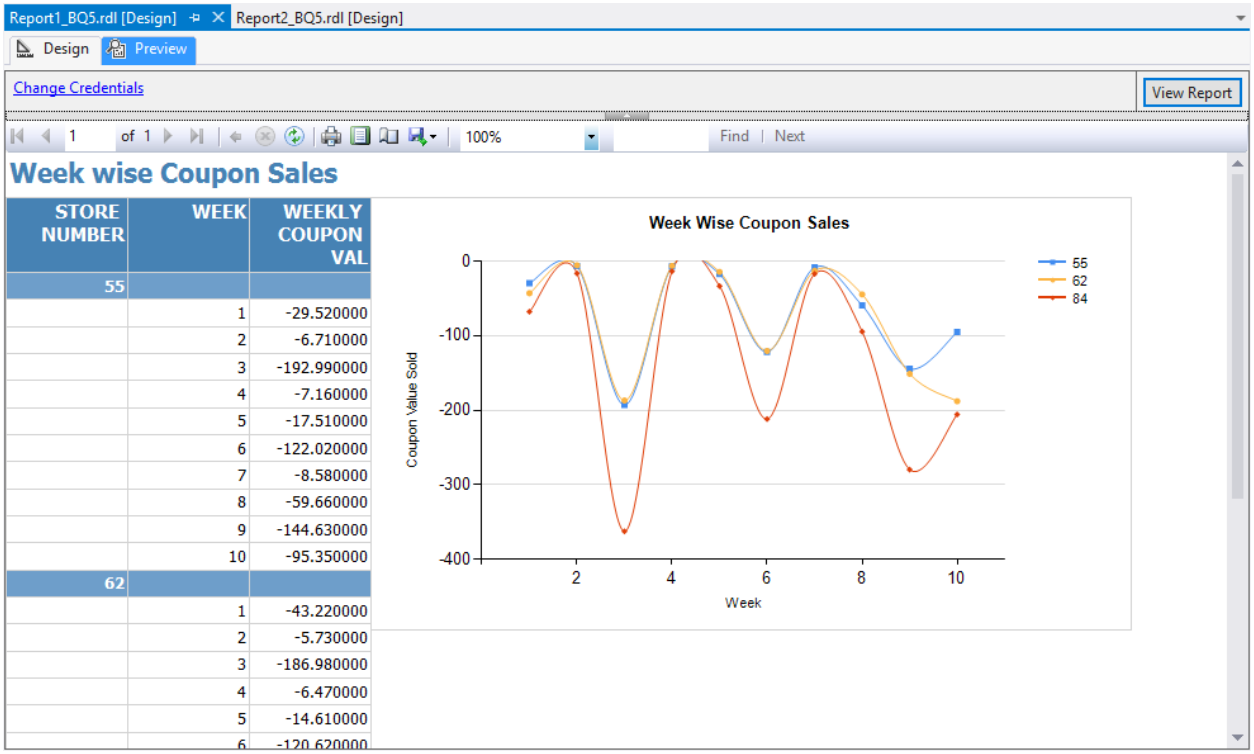
Query Designer depicting the queries used:





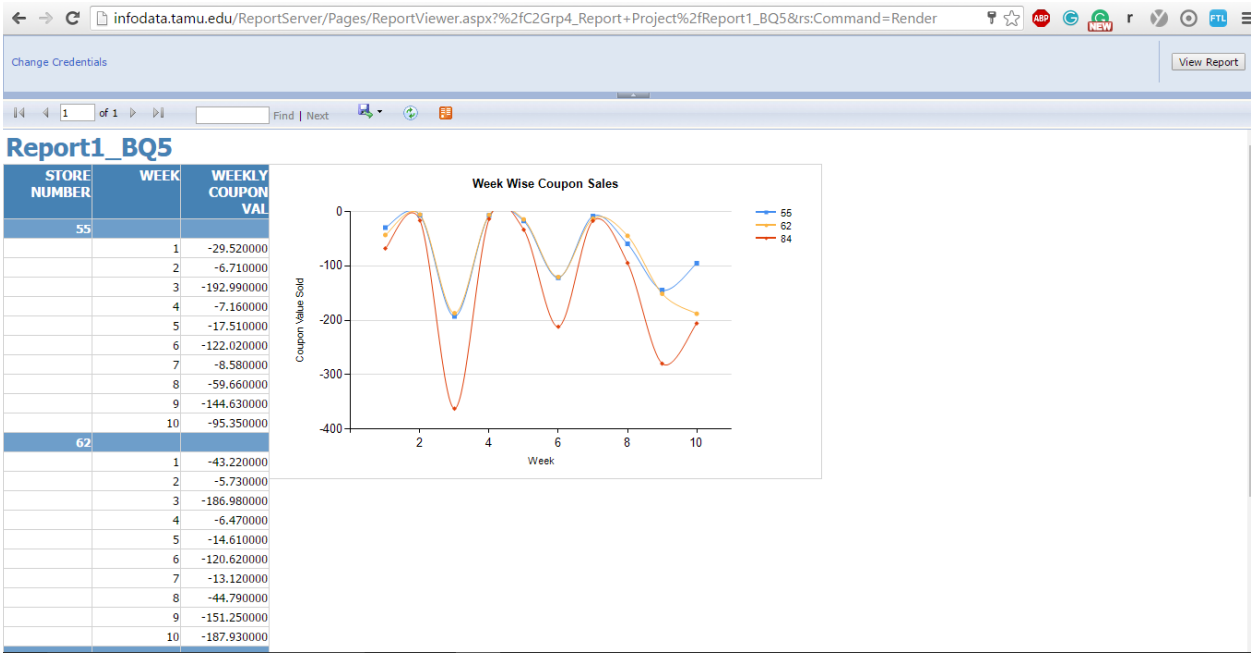

Report Wizard
Design the Table
Choose how to group the data in the table.
Available fields:
Displayed fields:
Page>
Group>
Details>
< Remove
WEEK
STORE_NUMBER
WEEKLY_COUPON_VAL
Help
< Back
Next >
Finish >>|
Cancel





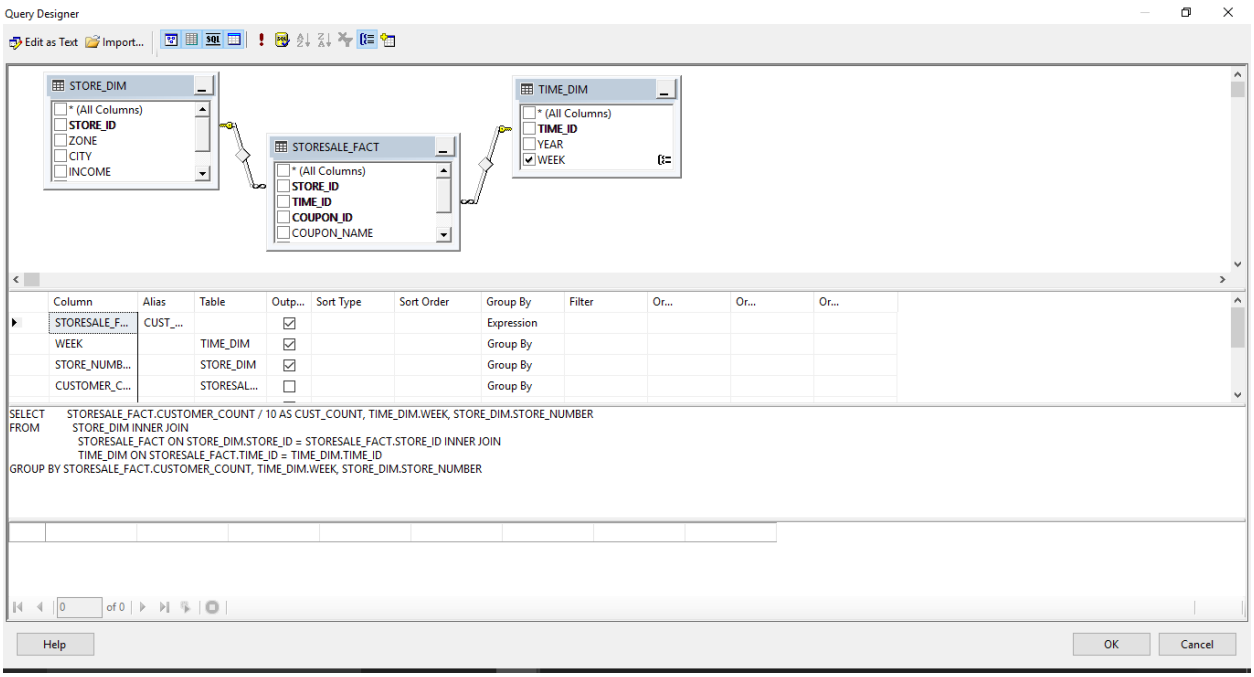
STORE NUMBER	WEEK	WEEKLY COUPON VAL
55		
	1	-29.520000
	2	-6.710000
	3	-192.990000
	4	-7.160000
	5	-17.510000
	6	-122.020000
	7	-8.580000
	8	-59.660000
	9	-144.630000
	10	-95.350000
62		
	1	-43.220000
	2	-5.730000
	3	-186.980000
	4	-6.470000
	5	-14.610000
	6	-120.620000

d. Report Deployed on the Server:



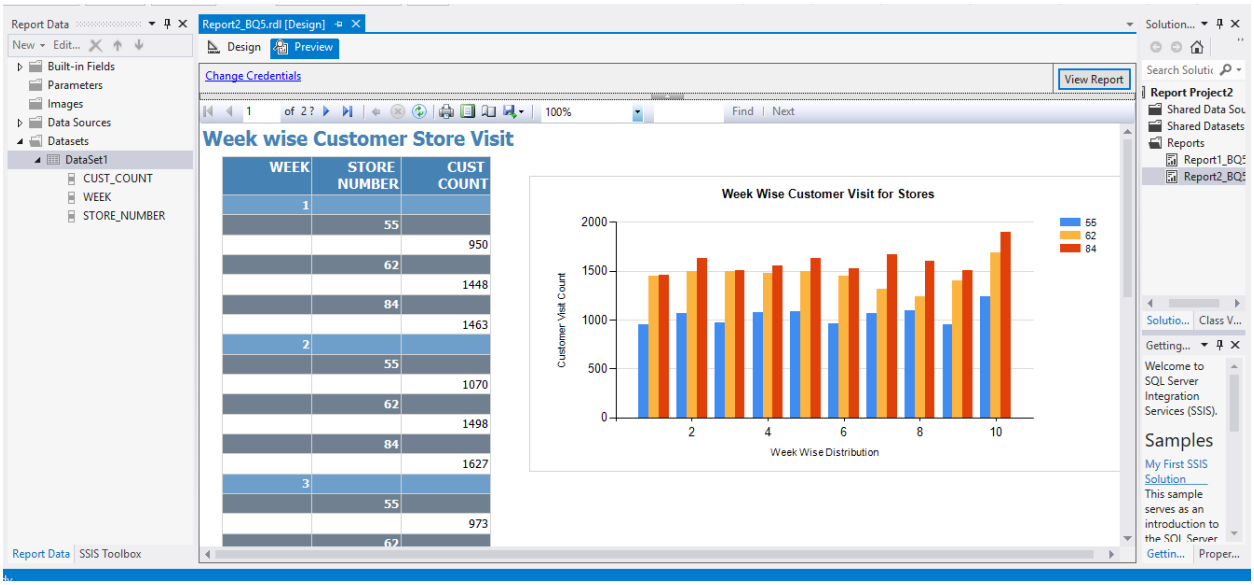
STORE NUMBER	WEEK	WEEKLY COUPON VAL
55		
	1	-29.520000
	2	-6.710000
	3	-192.990000
	4	-7.160000
	5	-17.510000
	6	-122.020000
	7	-8.580000
	8	-59.660000
	9	-144.630000
	10	-95.350000
62		
	1	-43.220000
	2	-5.730000
	3	-186.980000
	4	-6.470000
	5	-14.610000
	6	-120.620000
	7	-13.120000
	8	-44.790000
	9	-151.250000
	10	-187.930000

Similarly, the second report for depicting the relationship between Customer count visiting the store over the week time, is depicted into the second graph relation. Below are few of the implementation screenshots for the same:

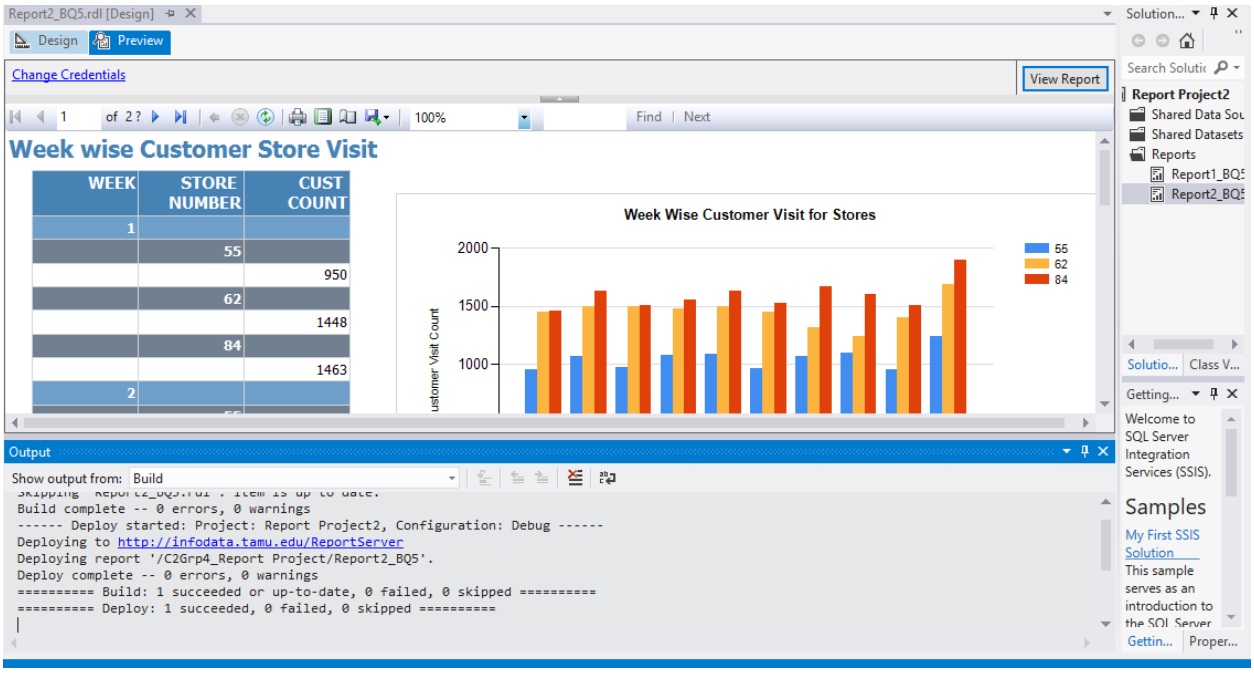


e. Report Template:

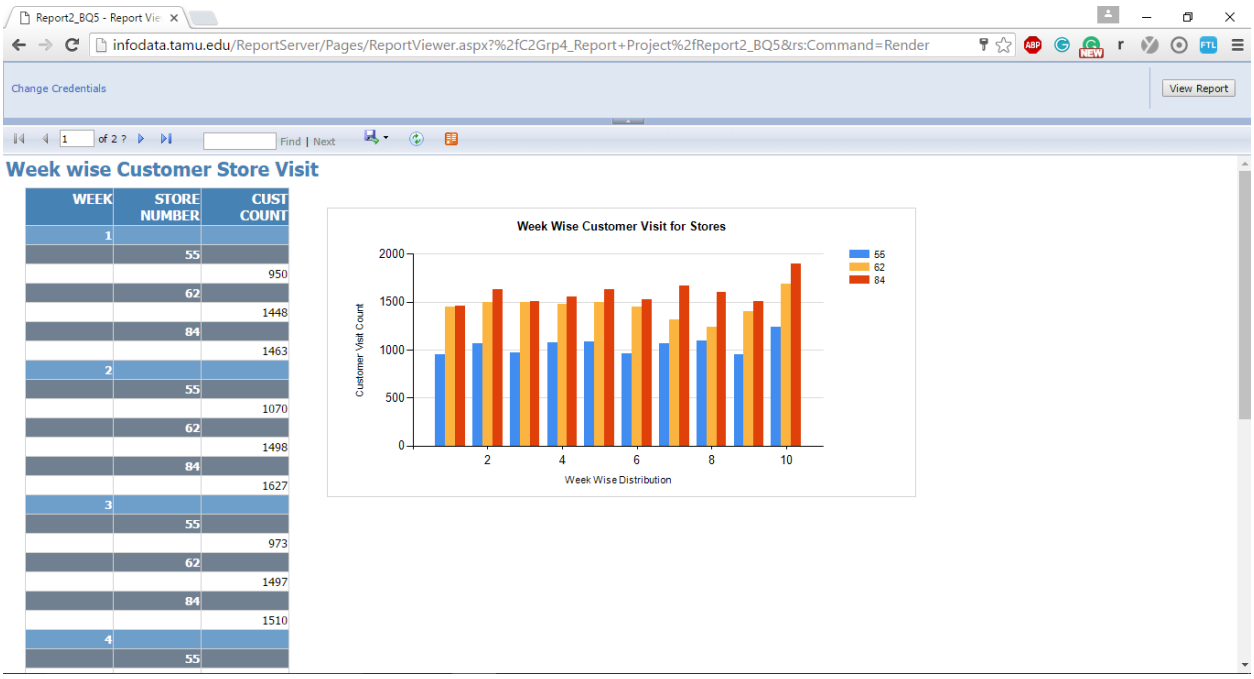
f. Reports Output:



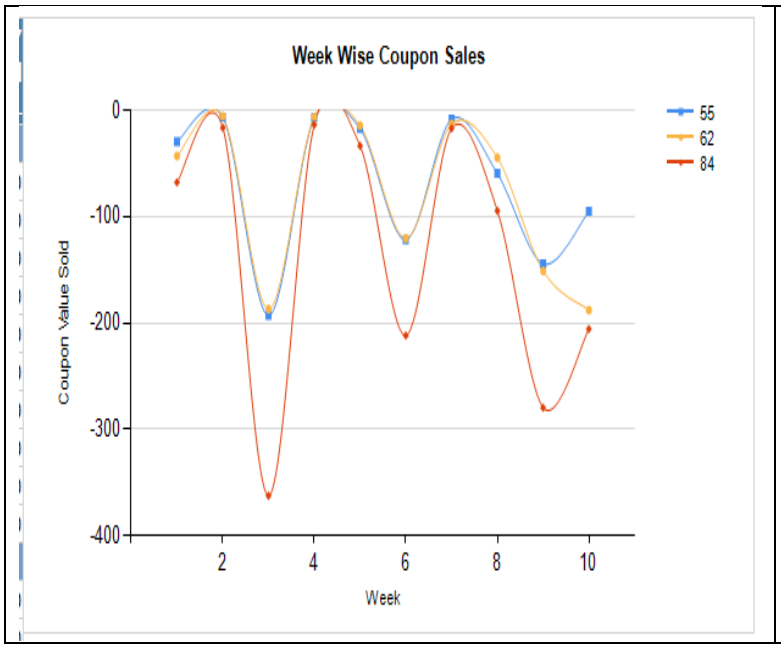
Report Deployment:

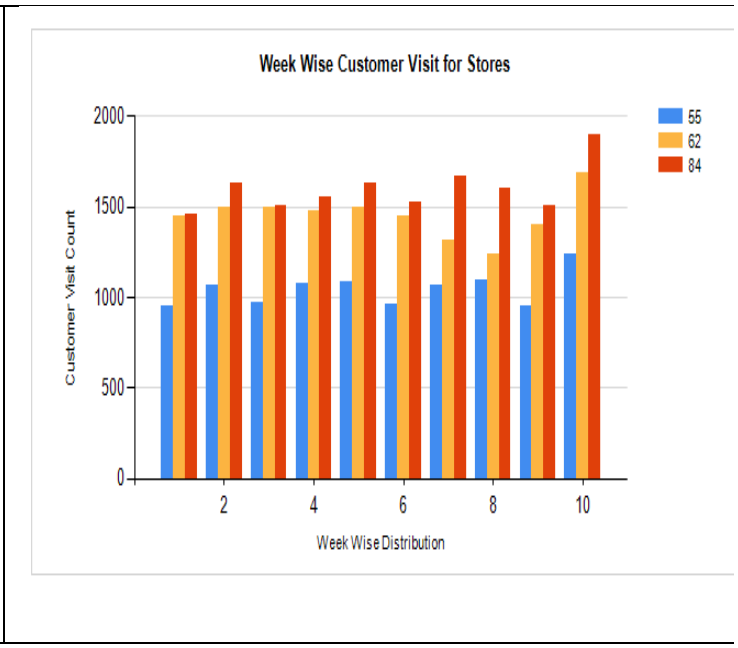


Server Report Output:

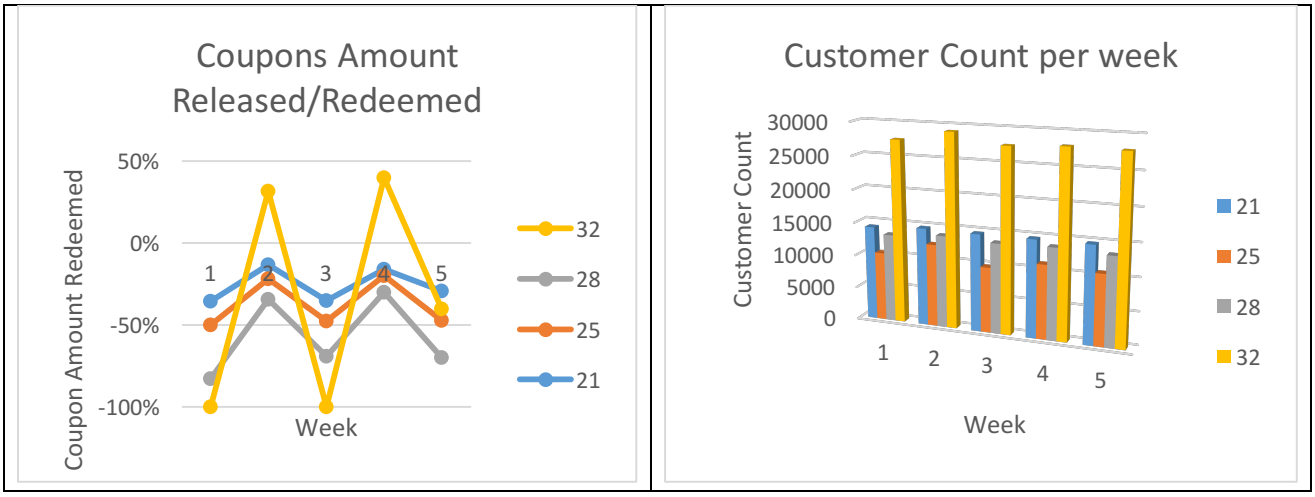


Comparing both the reports extracted as part of this business question, the following observation can be deduced by the business:





Hence, the obtained graph stands exactly in accordance to the proposed pivot table graph as under:



As can be observed in the above comparison charts, the coupon release value in store no 84 (depicted in red) decreased and hence decreased the number of customers in the second chart. Similarly, when in the 10th week, the number of coupon released by the store increase, the customer visits to the store also increased, affirming the fact that customer count increases linearly with respect to increase in coupon introduction. The same pattern can be deduced for other stores as well. This coupon-customer relationship will help the business to come up with further strategic coupon market that can attract the customer to the store.

BQ 4: Identifying the contribution of each product category in the overall sales for DFF

As previously stated, the objective here is to realize the contribution of each product category towards making up the total sales of DFF on a weekly or annual basis. This can also be rolled up to visualize the contribution of each product category to the overall sales for DFF.

a.) <u>Target report to satisfy business question</u>

Determining the target report to fulfil this business question, we will need a report which will aggregate sales of product into product categories, on an annual or weekly level. Such a report will enable us to successfully visualize the contribution of each product category to the annual sales at DFF.

The sample data (for just two categories here) can be as:

Sum of Sales	Categories	
Year	FRJC	WBJC
17	53734.45	50079.07
18	50802.37	50565.95
19	43659.98	46382.05
20	47854.7	48613.59

Data for graph

The graphical representation for the report can be visualized as following:

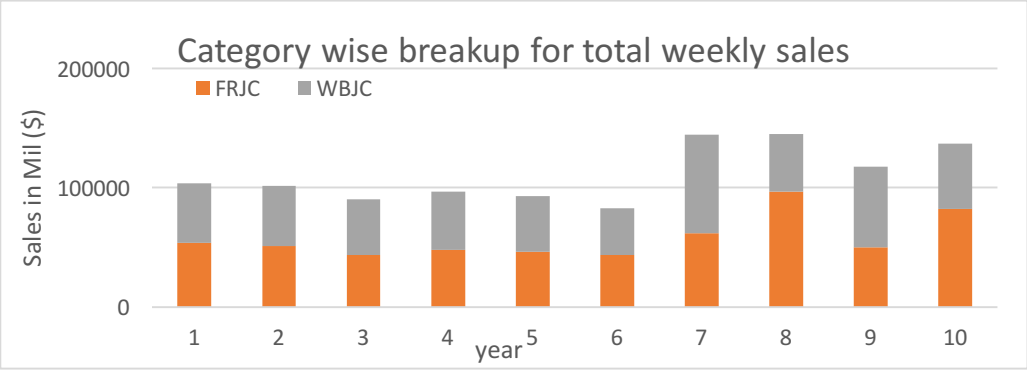


Figure: Target graph for BQ 4

b.) Report attributes & mapping from data mart

The attributes needed in the report are as below:

- Year of Sales
- Product categories
- Sum of sales for each product category, for each year

Looking at the nature of data needed, the dimension which will be used in this report will be: PROD_DIM, TIME_DIM. The product dimension will be used to extract the product category from the PRODUCY Key (PRODUCT_ID column in the data mart), and the time dimension will be used to derive the week from the time key (TIME_ID column in TIME_DIM). The fact tables used will be SALES_FACT and the measure we plan to use is SALES.

The mapping for each of the report parameters from the data mart tables is as below:

REPORT ATTRIBUTES		**SOURCE DATAMART**		
Attribute Name	Attribute Description	Table Name	Column Name	Summarization (if any)
Year	The year field	TIME_DIM	YEAR	
Product Category	Each individual product category	PRODUCT_DIM	CATEGORY	
Sum of Sales	Sum of sales, aggregated for each category, for each year	SALES_FACT	SALES	Summarized at product category level, at an annual level

Table: Report attribute mapping BQ 4

c.) Description of report template

The report will be presented in a format as described below:
There will be a column for year and for each year, we shall have the total sales for all the respective categories. The categories will be described by their category short name. The report will of course, have all the product categories as well as columns for the sub totals.

They will be presented in the following manner in the report:

SALES	**Product Category**					
Year	**ANA**	**BER**	**BJC**	**CER**	**CHE**	**Total**
1989	99	99	99	99	99	**495**
1990	99	99	99	99	99	**495**
Total	198	198	198	198	198	**990**

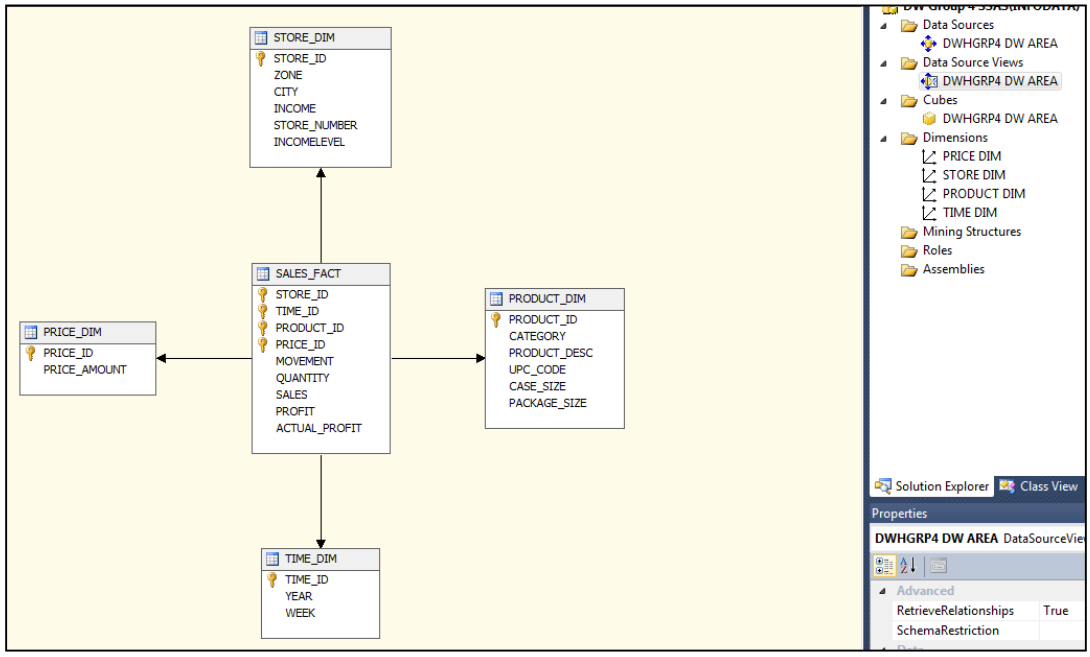
Table: Sample report attributes

Note: No Report designer screen shot has been attached here as the report is designed using a SSAS Cube (without SSRS).

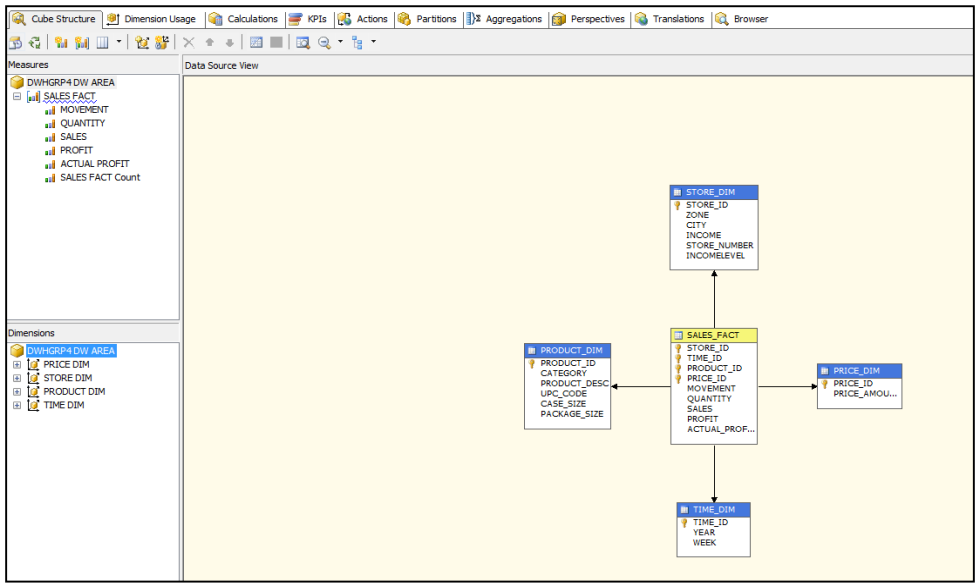
d.) **Report development – Using SSAS Cubes**

The report requires a very high amount of aggregation. The data must be aggregated at each category level as well as weekly data must be converted to yearly data. Direct querying from the data mart will be faster than queries from the transactional system. But it will still be very inefficient, considering the large amount of data at hand. Hence, we can build multidimensional cubes to satisfy this business question. The steps and the screen shots in the development of the cube are as below:

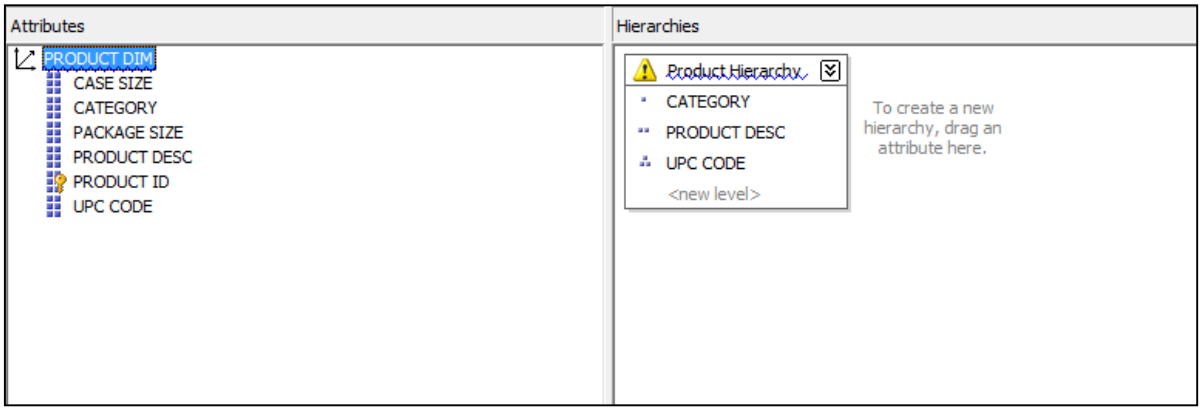
Data view:



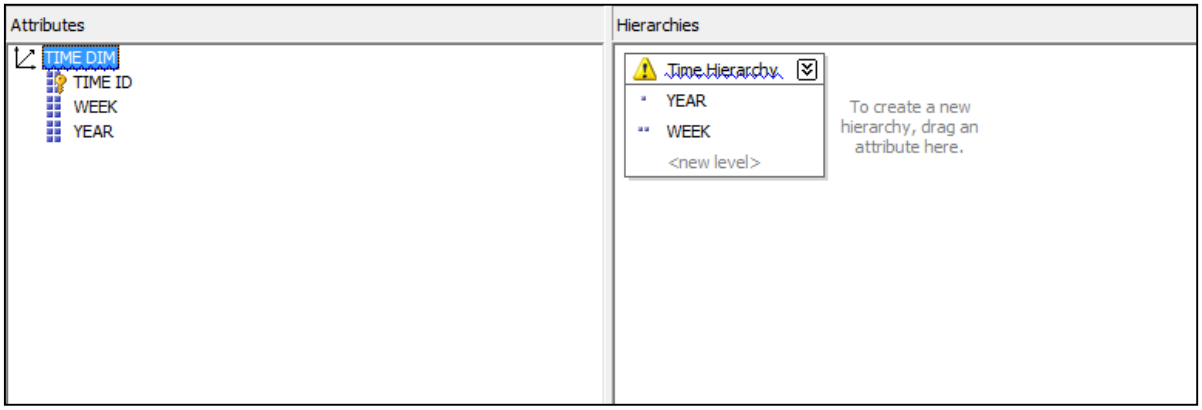
Cube:



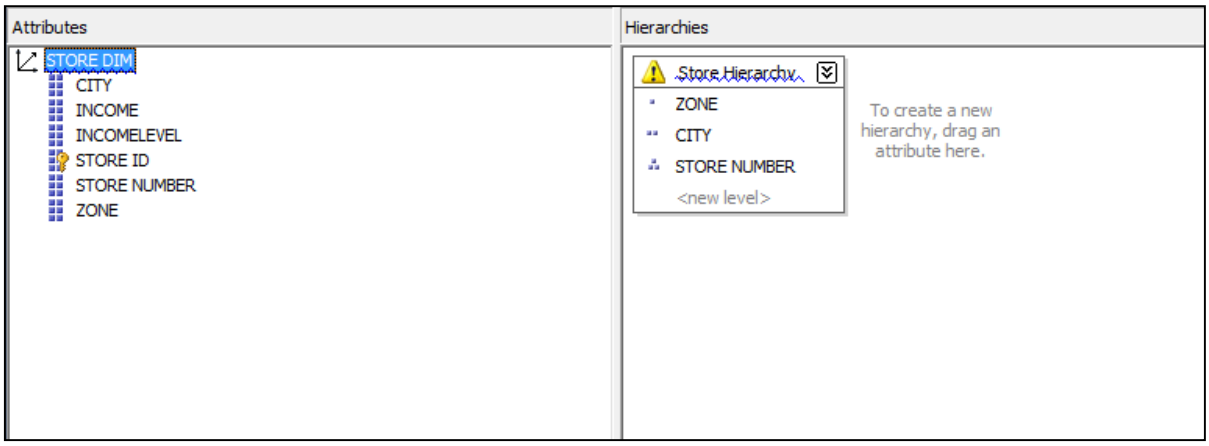
Dimension Hierarchies:



PRODUCT_DIM hierarchy



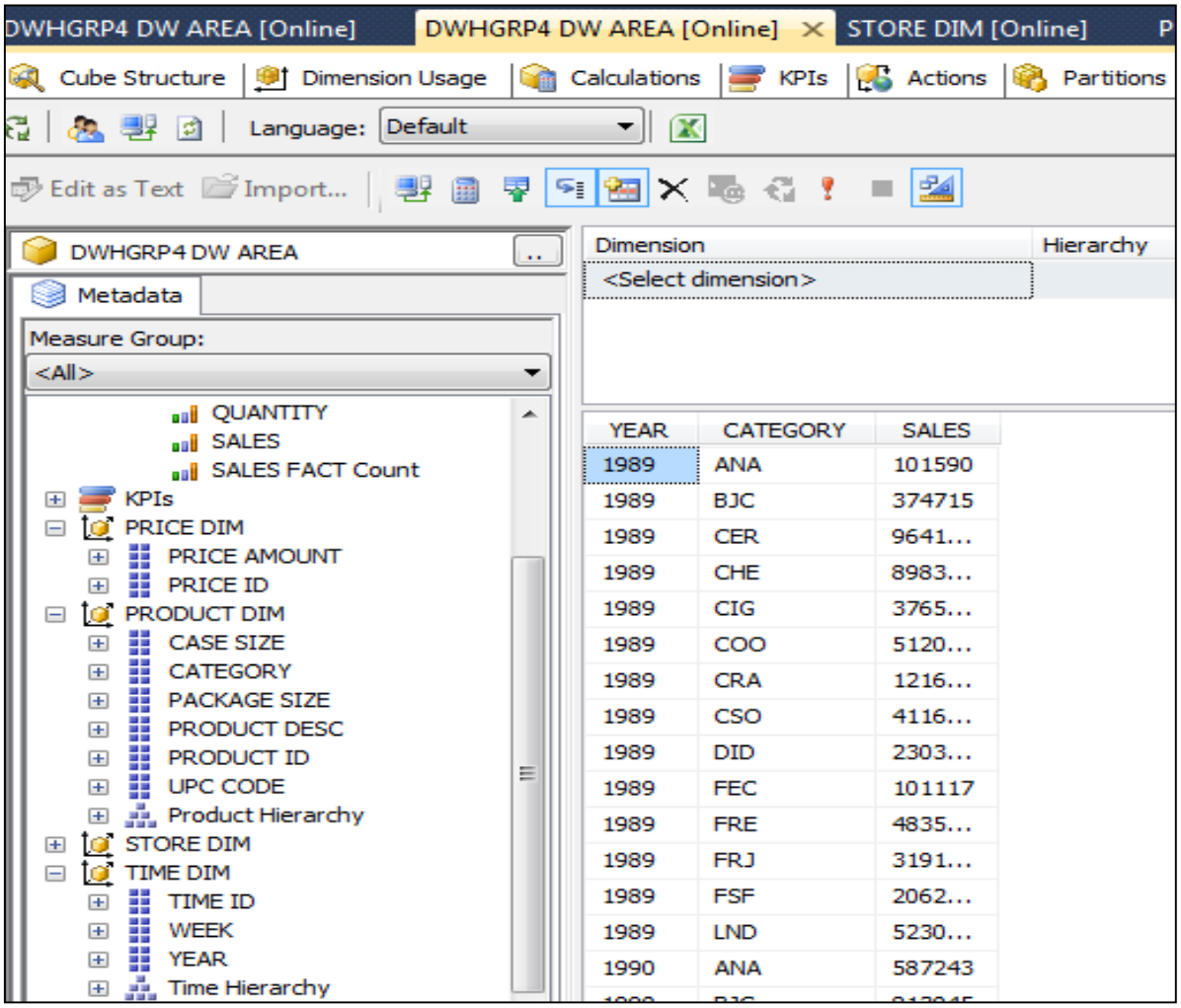
TIME_DIM hierarchy



STORE_DIM hierarchy

Cube browser:

No, we select the appropriate attributes from the CUBE browser, either from SSAS, or SQL server management studio and generate the desired report output. It is as shown below:



This data will serve as our report to answer the business question. We can export this report to MS Excel by clicking the little Excel icon on the top and a graph can be plot accordingly.

Excel gives an option to create a pivot from the newly imported data, and thus, we can create a pilot table and generate a graph from it as shown in the next section.

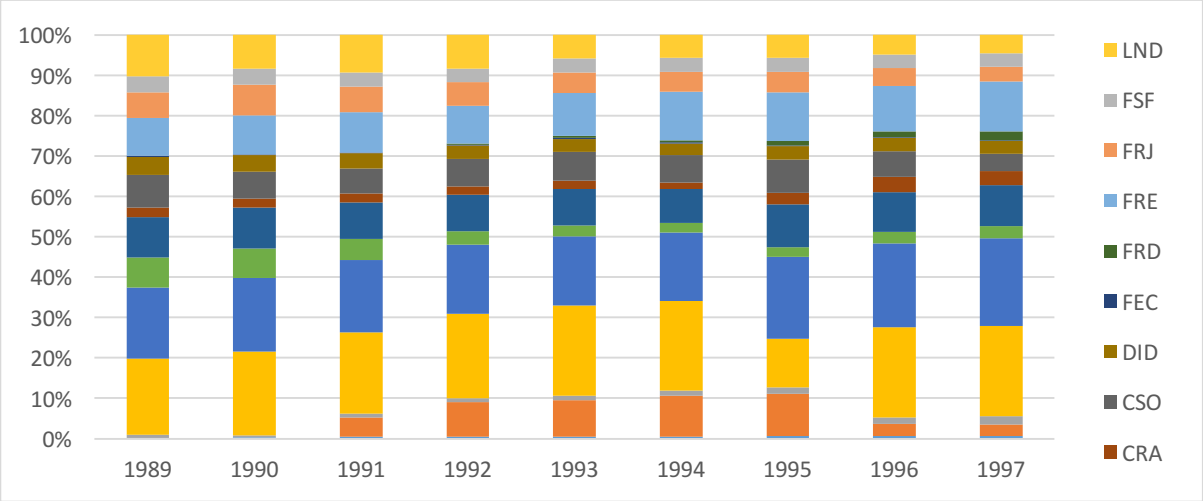
e.) **Report output**

The output from the newly formed pivot table in Excel is as below:

SALES Row Labels	Column Labels ANA	BJC	CER	CHE	CIG	COO
1989	101590	374715	9641320	8983080	3765848	5120251
1990	587243	813045	34848838	30618320	12420679	16994725
1991	784675	1614623	36664742	32796380	9662965	16399809
1992	863240	1932938	37751868	31124516	5965065	16427134
1993	865456	2037850	38069142	29037608	4508213	15376902
1994	773376	2050527	35816082	27424123	3780083	13416962
1995	840387	2289552	17675914	29753867	3400397	15618937
1996	860117	2236439	33088736	30719293	4148401	14470892
1997	384525	1277008	13944263	13621950	1807095	6386736
Grand Total	**6060609**	**14626697**	**257500905**	**234079137**	**49458746**	**120212348**

Note: The data is truncated for readability purpose.

The graph can be visualized from this data as below:



Thus, this report successfully answers the business question effectively shows the contribution of each product category in the annual sales of DFF. We can easily see which categories is contribution how much towards the sales, and can then make informed business decisions whether to push a product category further, as per the sales objectives of the organization.

BQ 5: Identifying price elasticity of demand for a particular item, w.r.t. income levels

"Price elasticity of demand (PED or E_d) is a measure used in economics to show the responsiveness, or elasticity, of the quantity demanded of a good or service to a change in its price, ceteris paribus. More precisely, it gives the percentage change in quantity demanded in response to a one percent change in price", Wikidepia.com. As previously stated, the objective here is to find the different values of movement for all the various values of the price, for a particular item and plot a graph for it. Here however, to study the effect with respect to income levels, we shall plot three separate graphs for LOW, HIGH and MEDIUM income levels, and study the difference in the pattern of the graph for each of them.

A. Target report to satisfy business question

Determining the target report to fulfil this business question, we will need a report which will show the average movement of particular product for all the possible values of prices. Also, the data is to be separated based on income level, found in the store dimension. We will need three different sets of price and average movement for the three income levels.

The sample data (for just two categories here) can be as (for a fixed UPC code):

Store Area Income Level	High
Price	Average of Move
1.58	34
1.59	29.5
1.65	21
1.66	17
1.67	17.34920635

Store Area Income Level	Medium
Price	Average of Move
1.53	4.625
1.57	19
1.59	14.95918367
1.6	11
1.61	10

Store Area Income Level	Low
Price	Average of Move
1.53	3
1.59	25.625
1.62	8
1.66	16.07692308

The graphical representation for the report can be visualized as following:

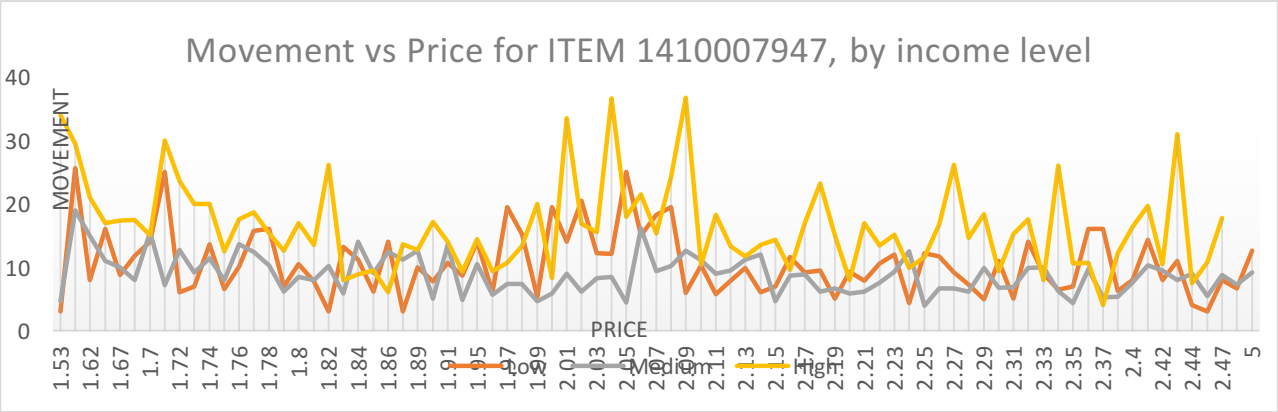


Figure: Target graph for BQ5

b. Report attributes & mapping from data mart

The attributes needed in the report are as below:

- Product UPC code
- Income Level
- Price
- Average movement, for given price, income level and UPC code.

Looking at the nature of data needed, the dimension which will be used in this report will be: PROD_DIM, STORE_DIM. The product dimension will be used to extract the product category from the PRODUCY Key (PRODUCT_ID column in the data mart) based on the UPC code. It will act as a filter for the data. Similarly, the STORE_KEY from STORE_DIM will act as a filter on the INCOME_LEVEL column. The fact tables used will be SALES_FACT and the measure we plan to use is MOVEMENT.

The mapping for each of the report parameters from the data mart tables is as below:

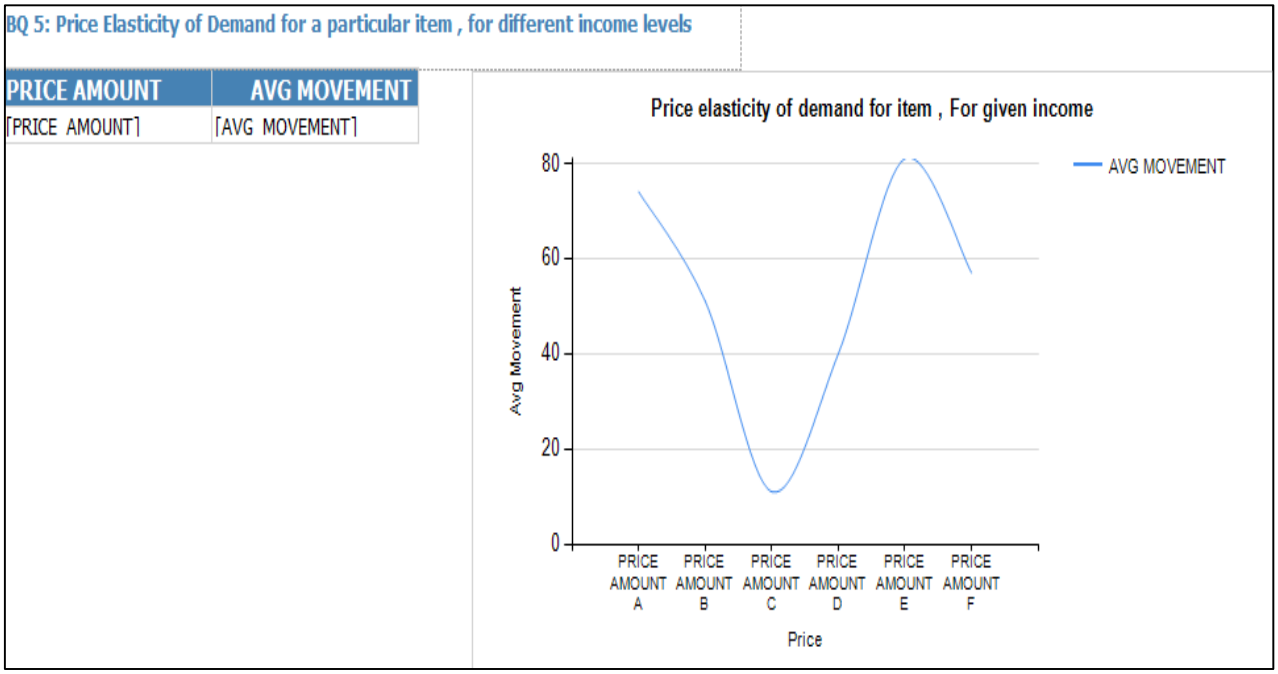
REPORT ATTRIBUTES		**SOURCE DATAMART**		
Attribute Name	Attribute Description	Table Name	Column Name	Summarization (if any)
UPC Code (Parameter)	The UPC code of the product selected by user	PRODUCT_DIM	UPC_CODE	
Income Level (Parameter)	The income level for which the report is to be viewed, selected by the user.	STORE_DIM	INCOME_LEVEL	
Price	The price values for which the product was sold	SALES_FACT	PRICE	
Average Movement	Average movement which will be used as Y-axis of graph.	SALES_FACT	MOVEMENT	Average of movement calculated for given product, income level and price.

Table: Report attribute mapping BQ 5

c. Description of report template

The report will be presented in a format as described below:
The report will have two parameters: UPC code, and income level (LOW, MEDIUM and HIGH). Once the user selects the appropriate query parameters, then the output will be displayed as per the below template. There will be a table with columns for price amount (in $) and one for the average movement. There will also be an accompanying graph for the plot.

They will be presented in the following manner in the report:

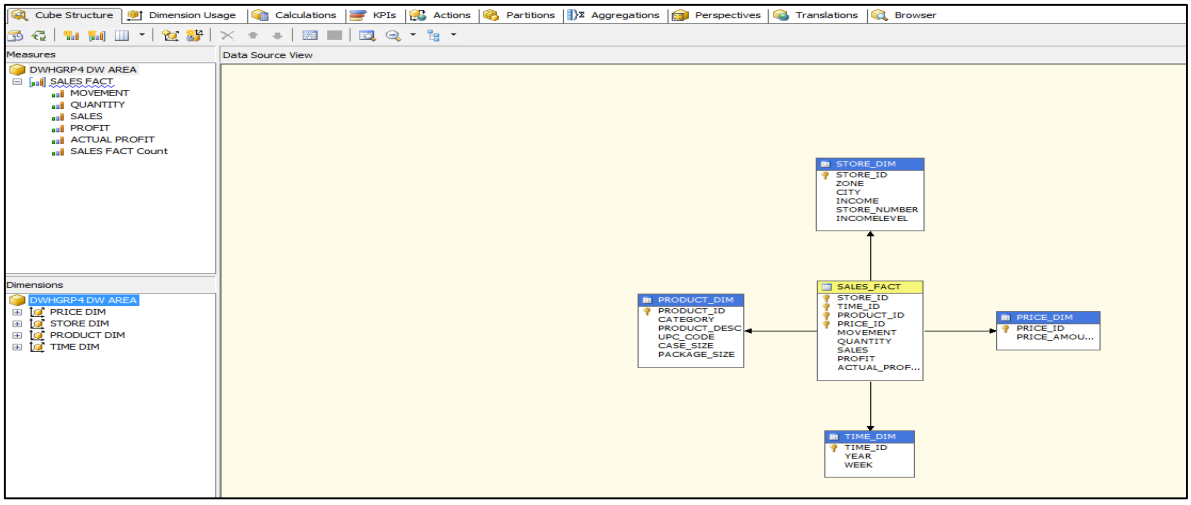


Report Template BQ 5

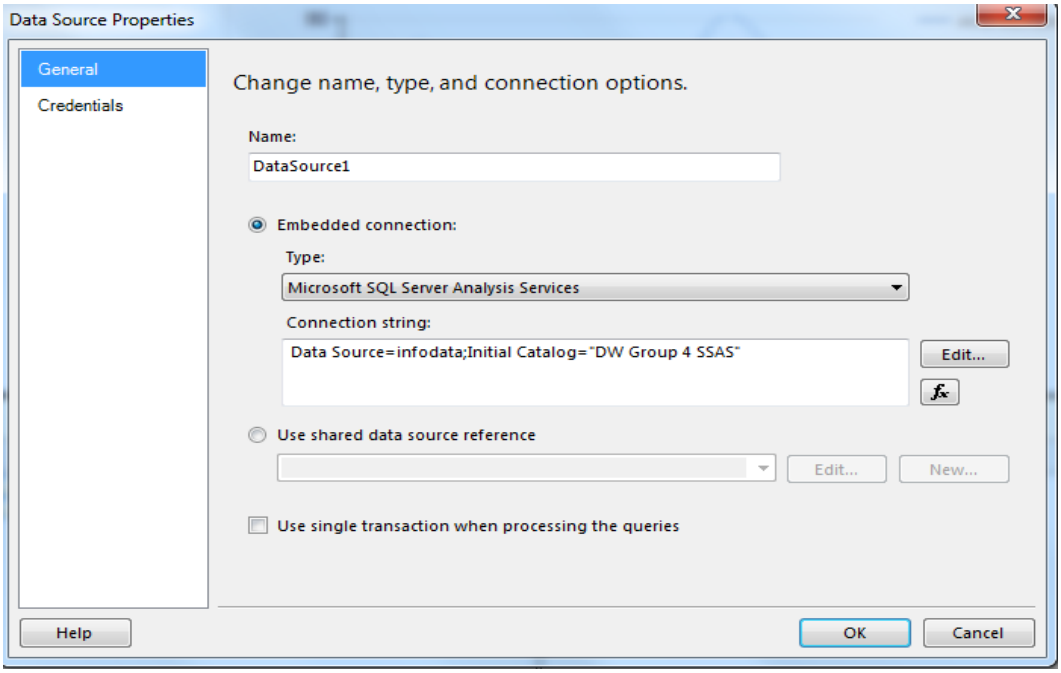
d. **Report development – Using SSRS based on SSAS Cube**

The report requires a high amount of aggregation. So this report was built on on top of the cube which was developed in the previous business question. This cube is used as a data source in SSRS and the report is developed in SSRS. Some screenshots of the development process are as below:

SSAS Cube:



SSRS Report data source:



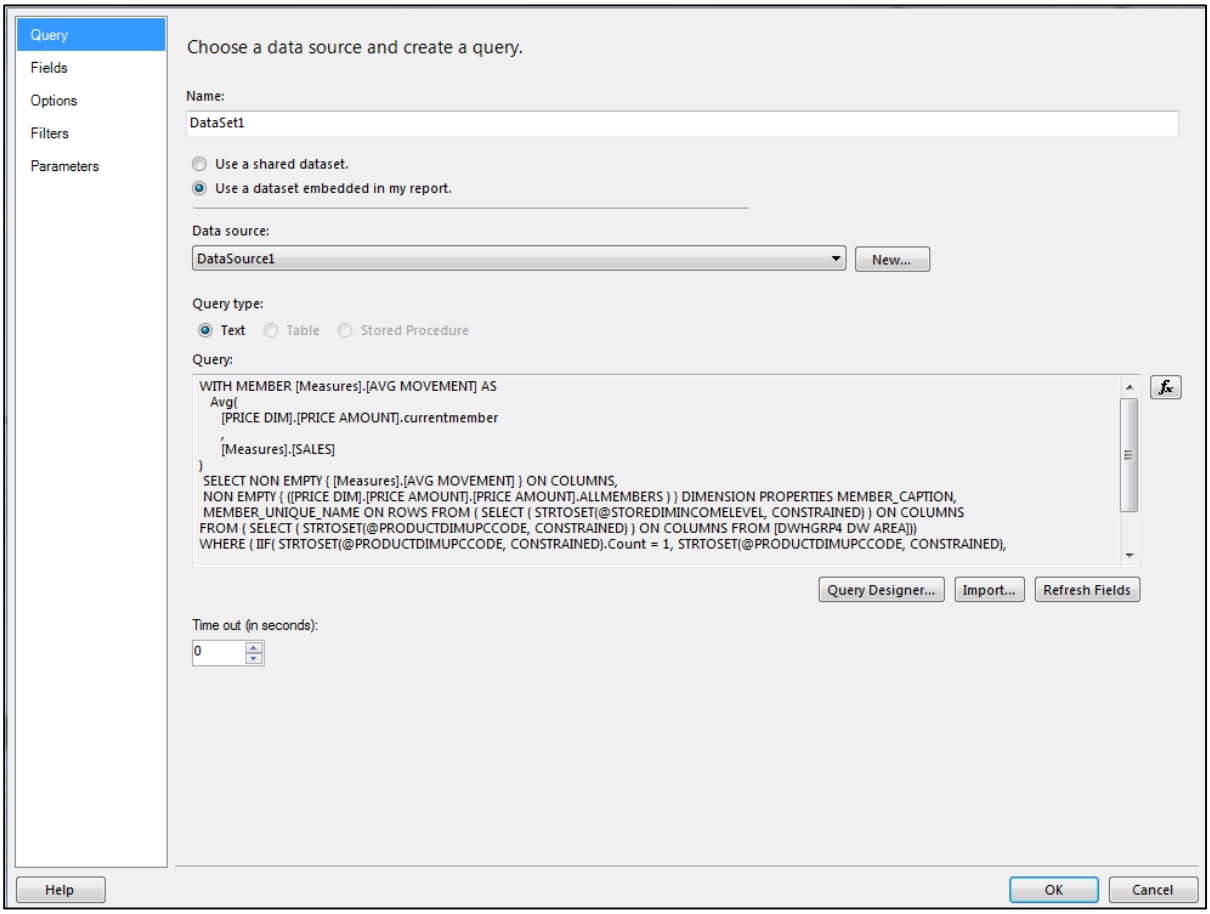
Here, as seen, we have used the SSAS cube as the data source.

Data CUBE query:

As seen, a cube query has been written to answer the business question. Note that the query builder tool could not be used in this case, as the requirement has some peculiar calculations for calculation the average. The MDX query used for this purpose is:

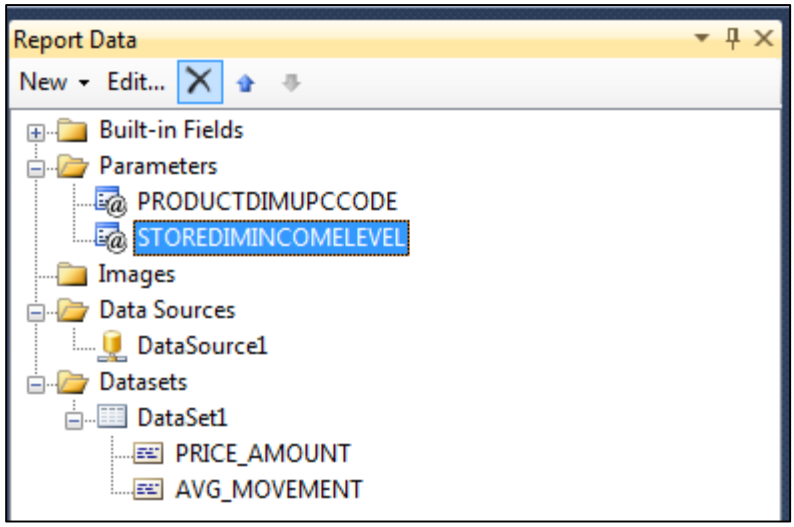
```
WITH MEMBER [MEASURES].[AVG MOVEMENT] AS
   AVG(
      [PRICE DIM].[PRICE AMOUNT].CURRENTMEMBER
      ,
      [MEASURES].[SALES]
)
 SELECT NON EMPTY { [MEASURES].[AVG MOVEMENT] } ON COLUMNS,
 NON EMPTY { ([PRICE DIM].[PRICE AMOUNT].[PRICE AMOUNT].ALLMEMBERS ) }
DIMENSION PROPERTIES MEMBER_CAPTION,
 MEMBER_UNIQUE_NAME ON ROWS FROM ( SELECT (
STRTOSET(@STOREDIMINCOMELEVEL, CONSTRAINED) ) ON COLUMNS
FROM ( SELECT ( STRTOSET(@PRODUCTDIMUPCCODE, CONSTRAINED) ) ON COLUMNS
FROM [DWHGRP4 DW AREA]))
WHERE ( IIF( STRTOSET(@PRODUCTDIMUPCCODE, CONSTRAINED).COUNT = 1,
STRTOSET(@PRODUCTDIMUPCCODE, CONSTRAINED),
 [PRODUCT DIM].[UPC CODE].CURRENTMEMBER ), IIF(
STRTOSET(@STOREDIMINCOMELEVEL, CONSTRAINED).COUNT = 1,
```

```
 STRTOSET(@STOREDIMINCOMELEVEL, CONSTRAINED), [STORE
DIM].[INCOMELEVEL].CURRENTMEMBER ) ) CELL PROPERTIES VALUE,
 BACK_COLOR, FORE_COLOR, FORMATTED_VALUE, FORMAT_STRING, FONT_NAME,
FONT_SIZE, FONT_FLAGS
```

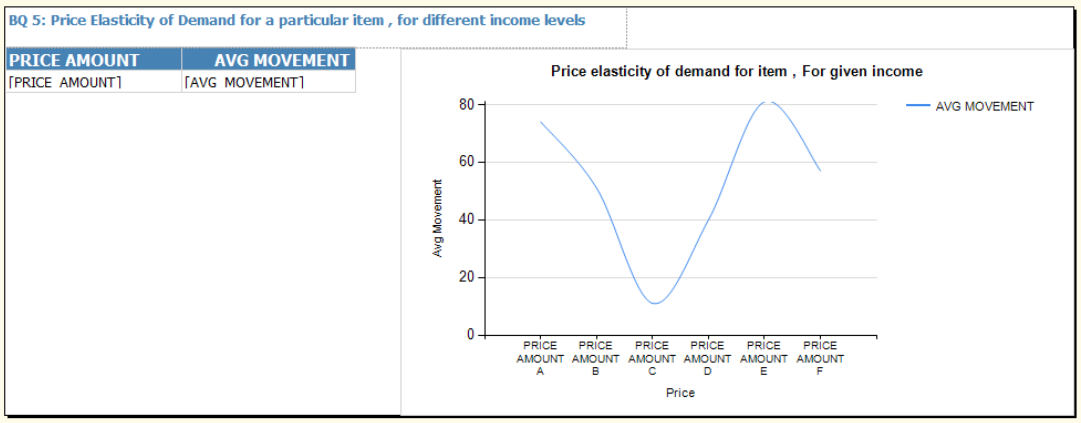


Report Parameters:

We have defined the two report parameters, for UPC code and Income level as shown below:

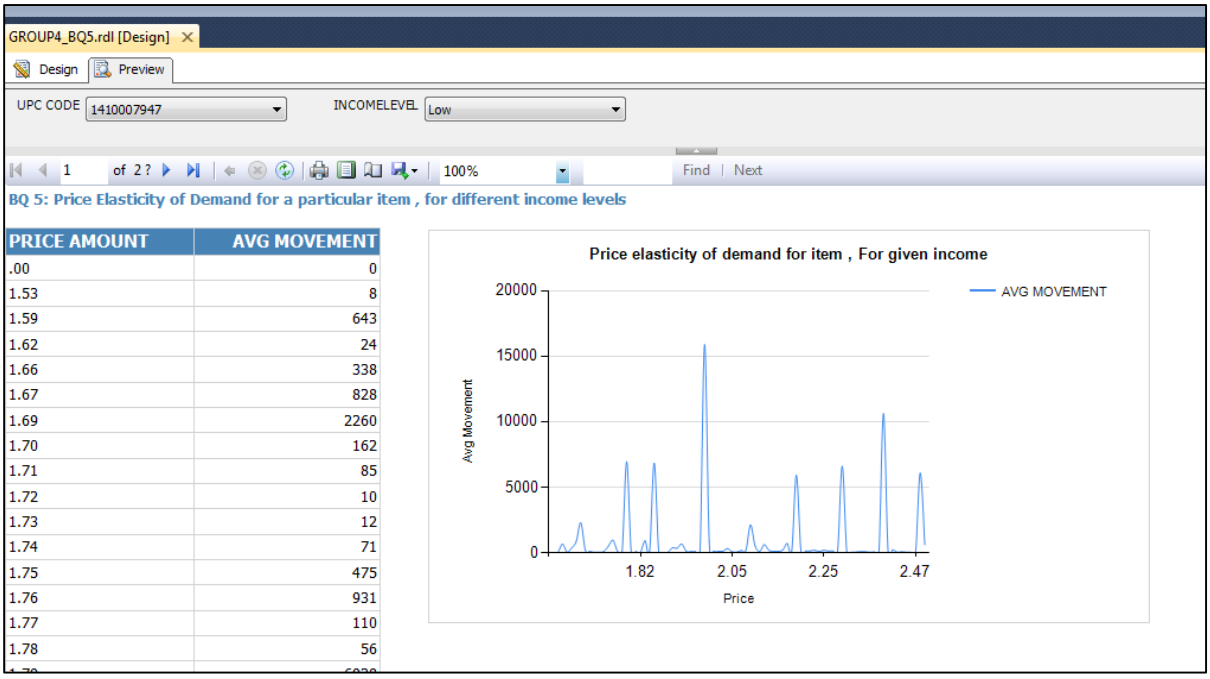


Report Design:

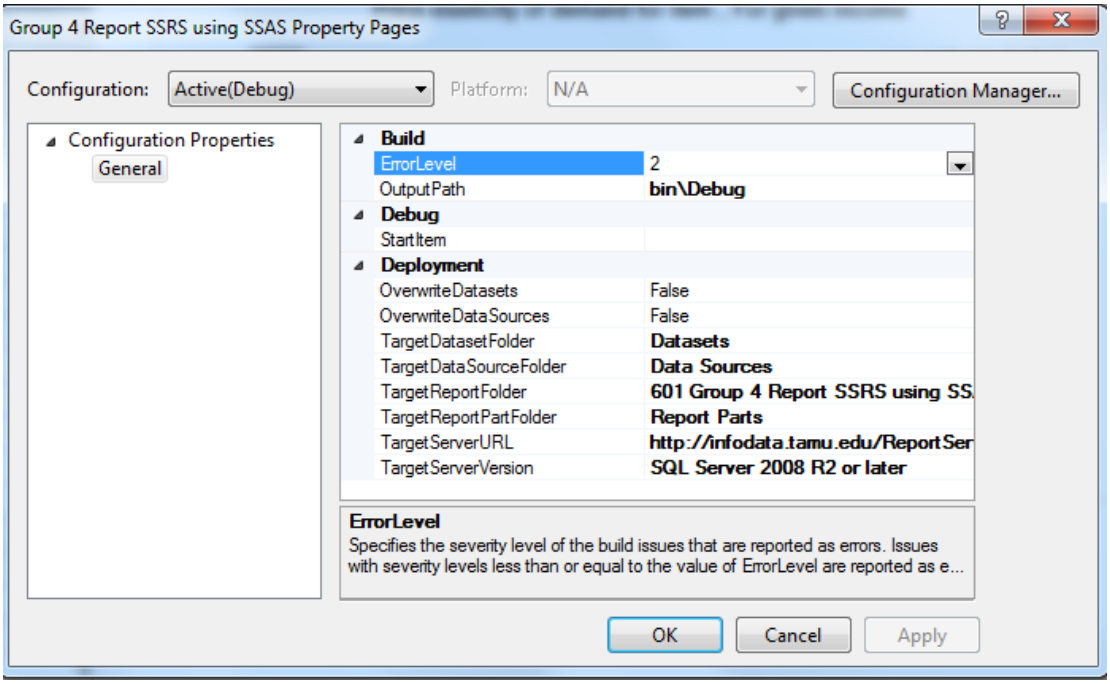


Report Preview:

A sample preview for the report can be seen below.



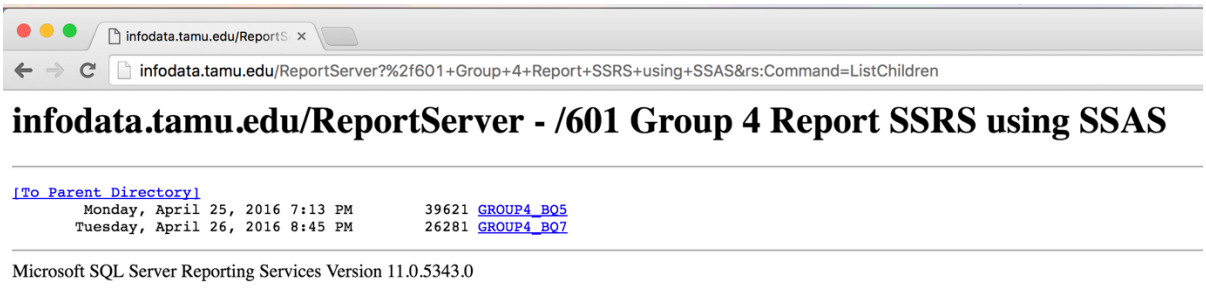
Report Deployment parameters:



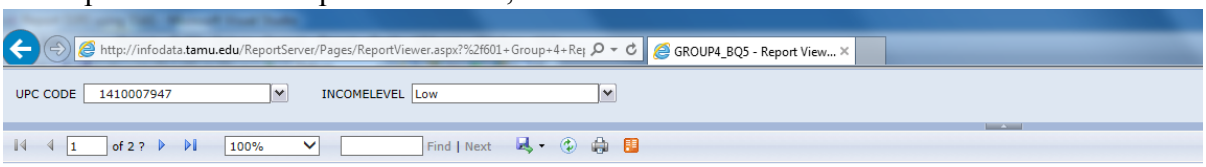
e. **Report output**

The report, once deployed on the server can be run directly.

The server screen for the reports is as below:

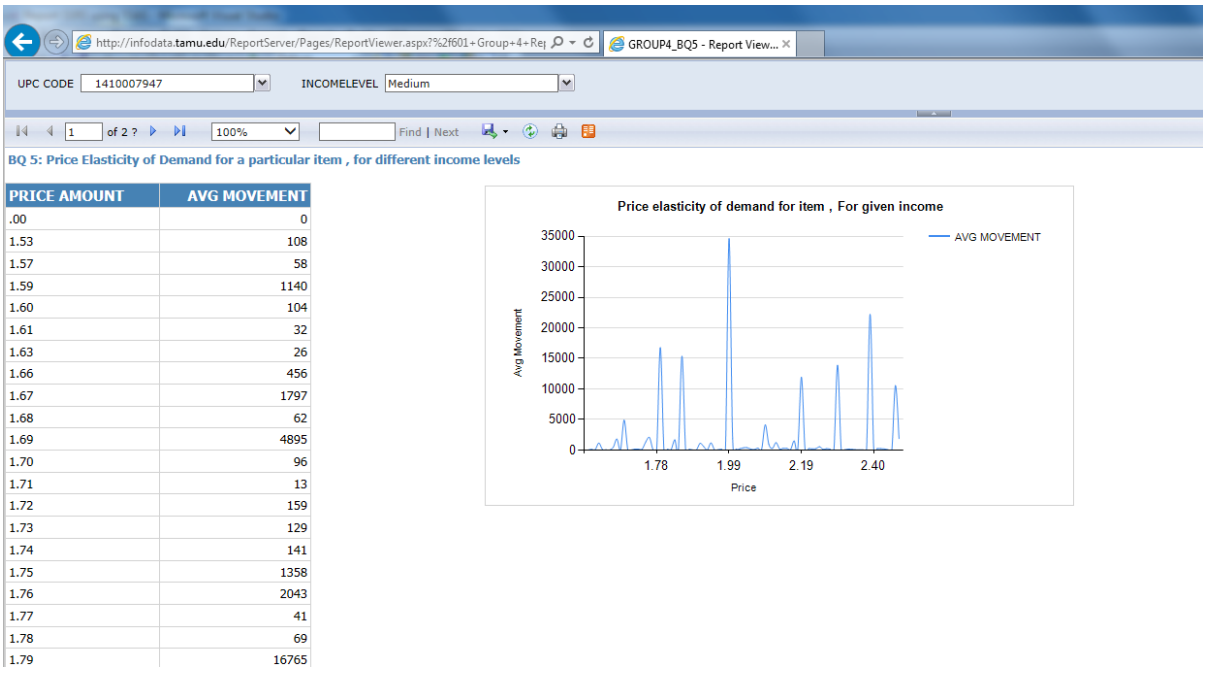


The report is run for one particular item, for all three income levels. The screen shots are as below.

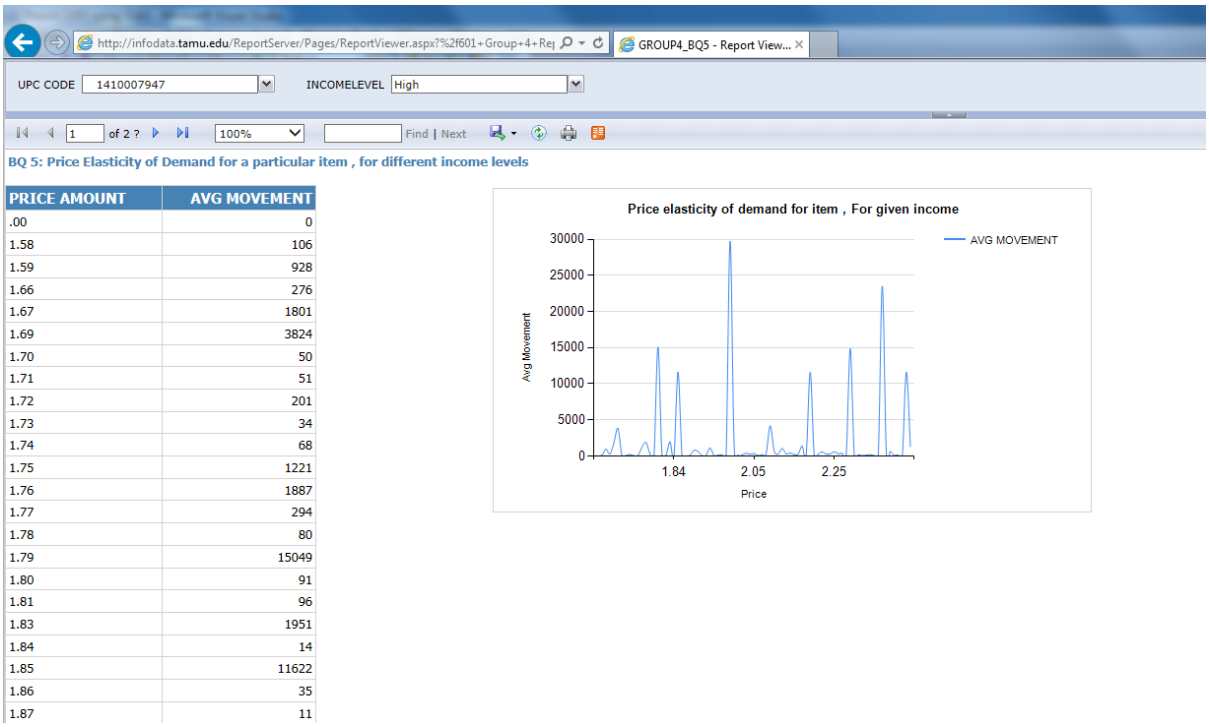


BQ 5: Price Elasticity of Demand for a particular item , for different income levels

PRICE AMOUNT	AVG MOVEMENT
.00	0
1.53	8
1.59	643
1.62	24
1.66	338
1.67	828
1.69	2260
1.70	162
1.71	85
1.72	10
1.73	12
1.74	71
1.75	475
1.76	931
1.77	110
1.78	56
1.79	6928
1.80	37
1.81	68
1.82	5
1.83	895



PRICE AMOUNT	AVG MOVEMENT
.00	0
1.53	108
1.57	58
1.59	1140
1.60	104
1.61	32
1.63	26
1.66	456
1.67	1797
1.68	62
1.69	4895
1.70	96
1.71	13
1.72	159
1.73	129
1.74	141
1.75	1358
1.76	2043
1.77	41
1.78	69
1.79	16765



PRICE AMOUNT	AVG MOVEMENT
.00	0
1.58	106
1.59	928
1.66	276
1.67	1801
1.69	3824
1.70	50
1.71	51
1.72	201
1.73	34
1.74	68
1.75	1221
1.76	1887
1.77	294
1.78	80
1.79	15049
1.80	91
1.81	96
1.83	1951
1.84	14
1.85	11622
1.86	35
1.87	11

Thus, this report successfully answers the business question and efficiently shows the trends in price elasticity of demand for a product of choice, for the three different levels on income. This information can be very useful for making several important business decisions like offers and discounts, product promotion etc.

BQ 7: How sales of an item of varying packet size is impacted by price change

The question here is to identify the effect of price change on sales for the same item, having different pack sizes. That is to compare the price elasticity of demand among products of varying pack sizes. The price elasticity of demand can be defined as the change in quantity sold per unit change in price for a given item, with all other factors constant. To facilitate this comparison, a graph of average change in movement to the price is plotted for both the pack sizes.

A. Target report to satisfy business question

Determining the target report to fulfil this business question, we will need a report which will show the average movement of particular product for all the possible values of prices. Also, the data is to be separated based on different pack sizes of the same article, found in the product dimension. The product will be selected based on the product description. We will need three different sets of price and average movement for the all the available pack sizes.

The sample data (for just two categories here) can be as (for a fixed fixed product description: 'almost home oatmeal'):

UPC	1410007234
Pack size	24 / 10 O
Row Labels	Average of MOVE
1.91	3.495412844
1.95	12
1.97	7
2	5
2.03	4

UPC	1410007404
Pack Size	7.5 O
Row Labels	Average of MOVE
0	2
1.69	4.253521127
1.72	3.318181818
1.85	2.341463415
1.99	1.983909895

The graphical representation for the report can be visualized as following:

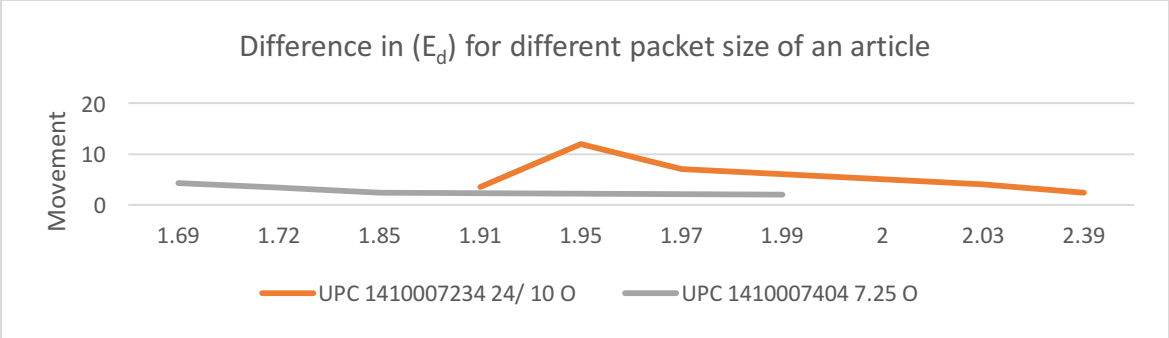


Figure: Target graph for BQ7

b. Report attributes & mapping from data mart

The attributes needed in the report are as below:

- Product UPC code
- Product Description
- Pack Size
- Price
- Average movement, for given price, product description and pack size.

Looking at the nature of data needed, the dimension which will be used in this report will be: PROD_DIM. The product dimension will be used to extract the product category from the PRODUCY Key (PRODUCT_ID column in the data mart) based on the selected product description, pack size and the UPC code. The dimension will be used to populate the list of values available in the report parameters selection list as well. It will act as a filter for the data. The fact tables used will be SALES_FACT and the measure we plan to use is MOVEMENT.

The mapping for each of the report parameters from the data mart tables is as below:

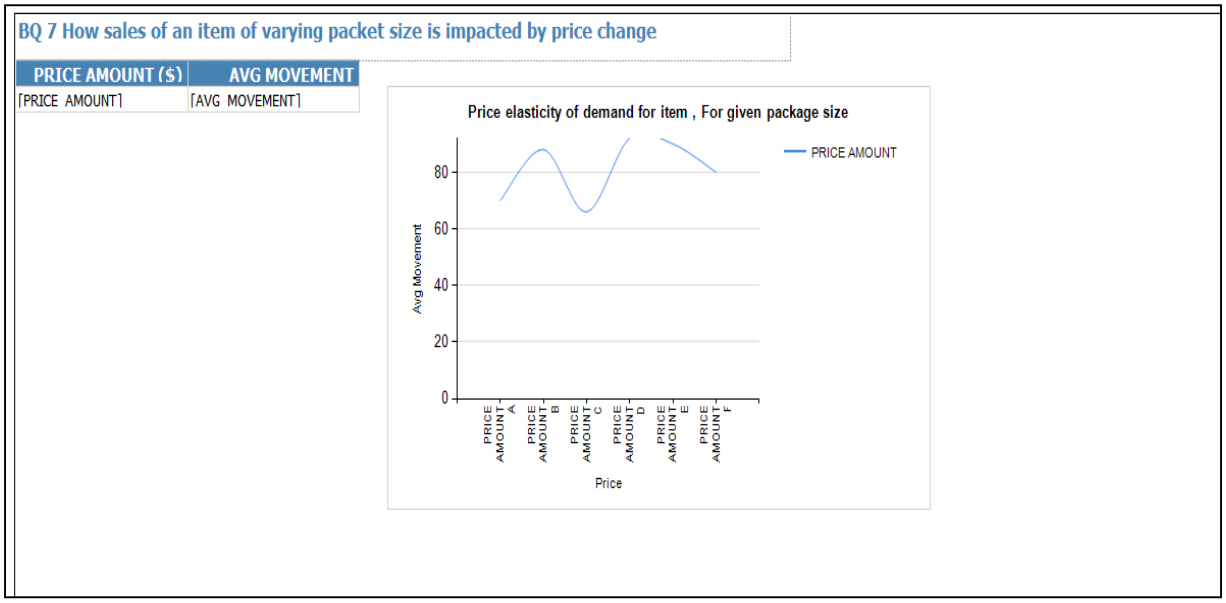
REPORT ATTRIBUTES		**SOURCE DATAMART**		
Attribute Name	Attribute Description	Table Name	Column Name	Summarization (if any)
Product Description	The name of the product to be studied.	PRODUCT_DIM	PRODUCT_DESC	
Pack Size (Parameter)	The pack size for which the data is to be filtered, selected by the user.	PRODUCT_DIM	PACK_SIZE	
UPC Code (Parameter)	The UPC code of the product selected by user, and the pack size , needed in case of multiple products having same description and pack size.	PRODUCT_DIM	UPC_CODE	
Price	The price values for which the product was sold	SALES_FACT	PRICE	
Average Movement	Average movement which will be used as Y-axis of graph.	SALES_FACT	MOVEMENT	Average of movement calculated for given product description, pack size, UPC code and price.

Table: Report attribute mapping BQ 5

c. Description of report template

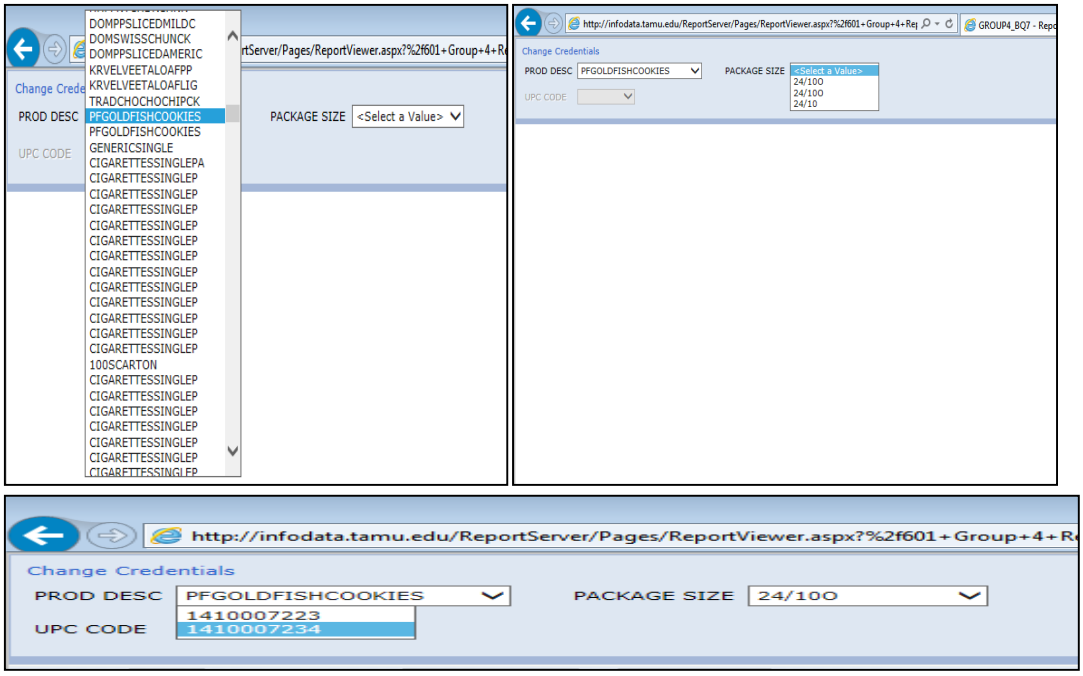
The report will be presented in a format as described below:
The report will have three parameters: Product Description, Pack Size and UPC code. Once the user selects the appropriate query parameters, then the output will be displayed as per the below template. There will be a table with columns for price amount (in $) and one for the average movement. There will also be an accompanying graph for the plot.

They will be presented in the following manner in the report:



Report Template BQ7

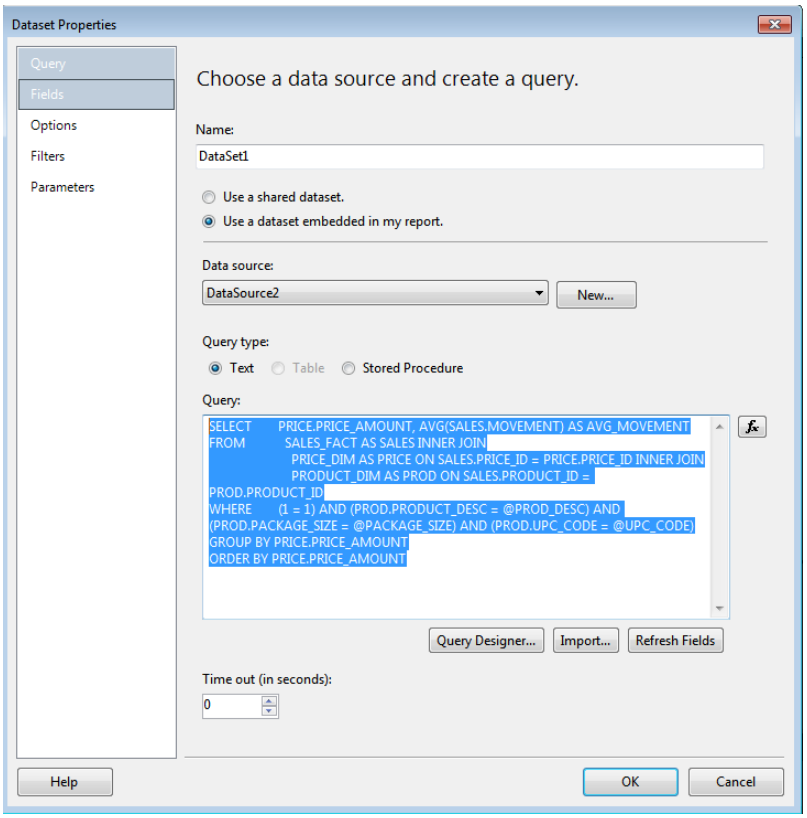
Besides, the parameter screen will also need to be designed. Once the user selects the product description, the selection box for the pack sized must be updated to show all the pack sized of that product. Once, the product description and pack size are selected, then the user needs to select the appropriate UPC code, in case of multiple items with same description and pack size, as the last parameter for the query. It is as shown below:



d. **Report development – Using SSRS**

The report will be developed using SSRS. We shall use SQL queries for the report, as well as specifying the list of available values for the report parameters. The screen shots for the report development process are as below:

SSRS Report data source:



Here, as seen, we have used the SQL server 2012 as the data source.

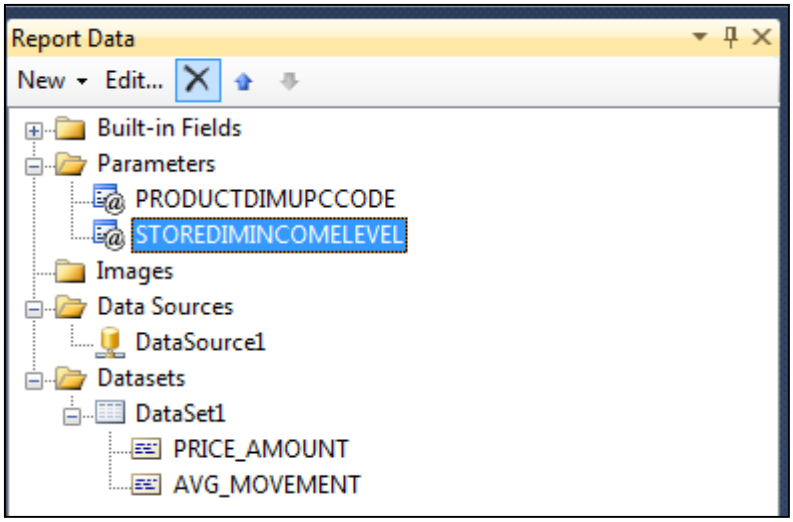
SQL query:

The query that has been used to generate the report is as below:

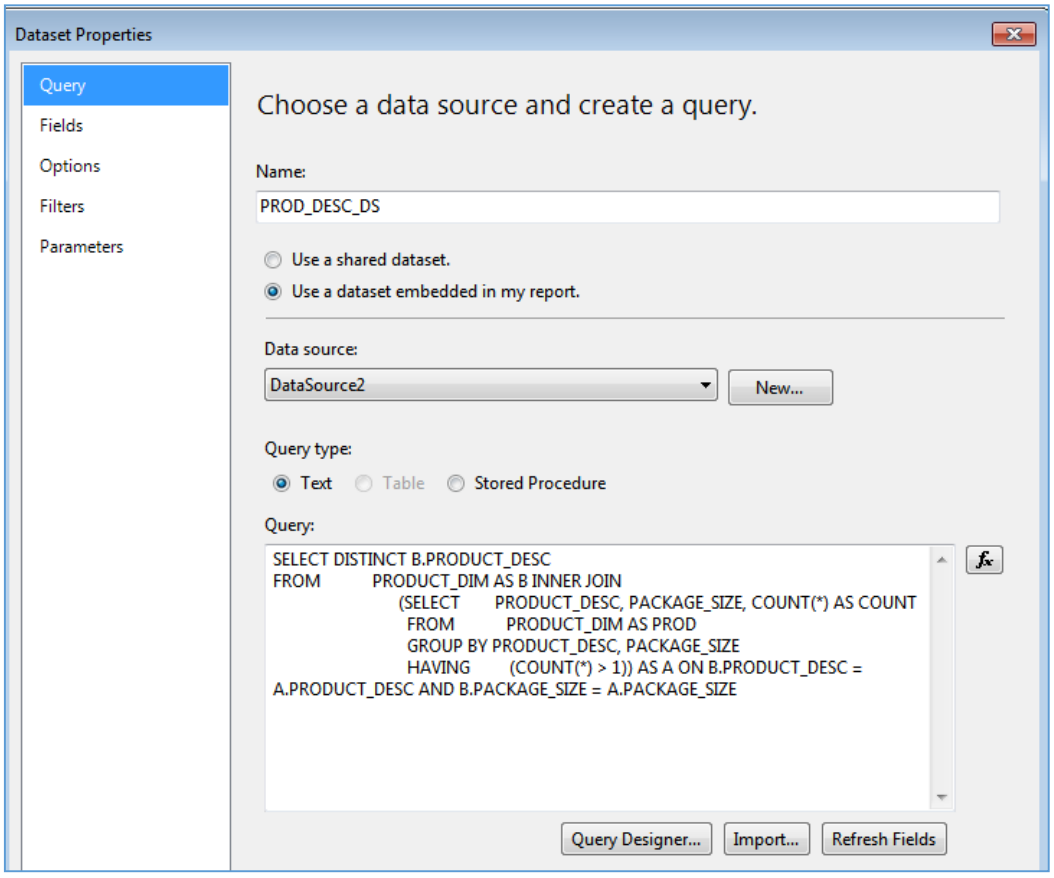
```
SELECT PRICE.PRICE_AMOUNT, AVG(SALES.MOVEMENT) AS AVG_MOVEMENT
FROM
SALES_FACT AS SALES INNER JOIN
PRICE_DIM AS PRICE ON SALES.PRICE_ID = PRICE.PRICE_ID INNER JOIN
PRODUCT_DIM AS PROD ON SALES.PRODUCT_ID = PROD.PRODUCT_ID
WHERE (1 = 1)
AND (PROD.PRODUCT_DESC = @PROD_DESC)
AND (PROD.PACKAGE_SIZE = @PACKAGE_SIZE)
AND (PROD.UPC_CODE = @UPC_CODE)
GROUP BY PRICE.PRICE_AMOUNT
ORDER BY PRICE.PRICE_AMOUNT
```

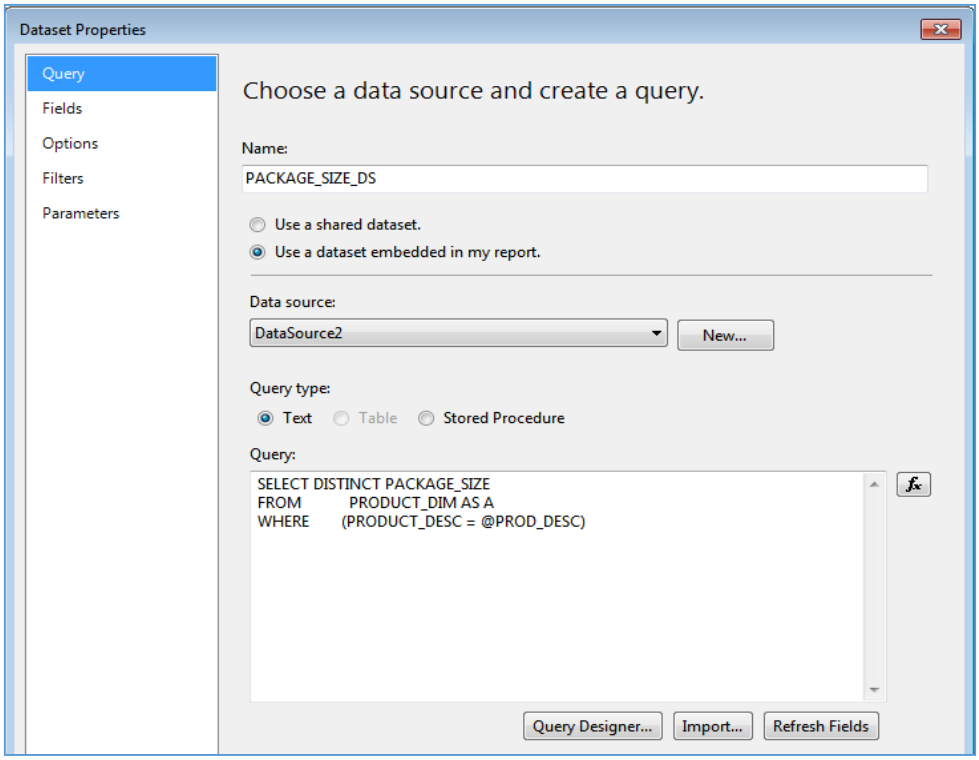
Report Parameters:

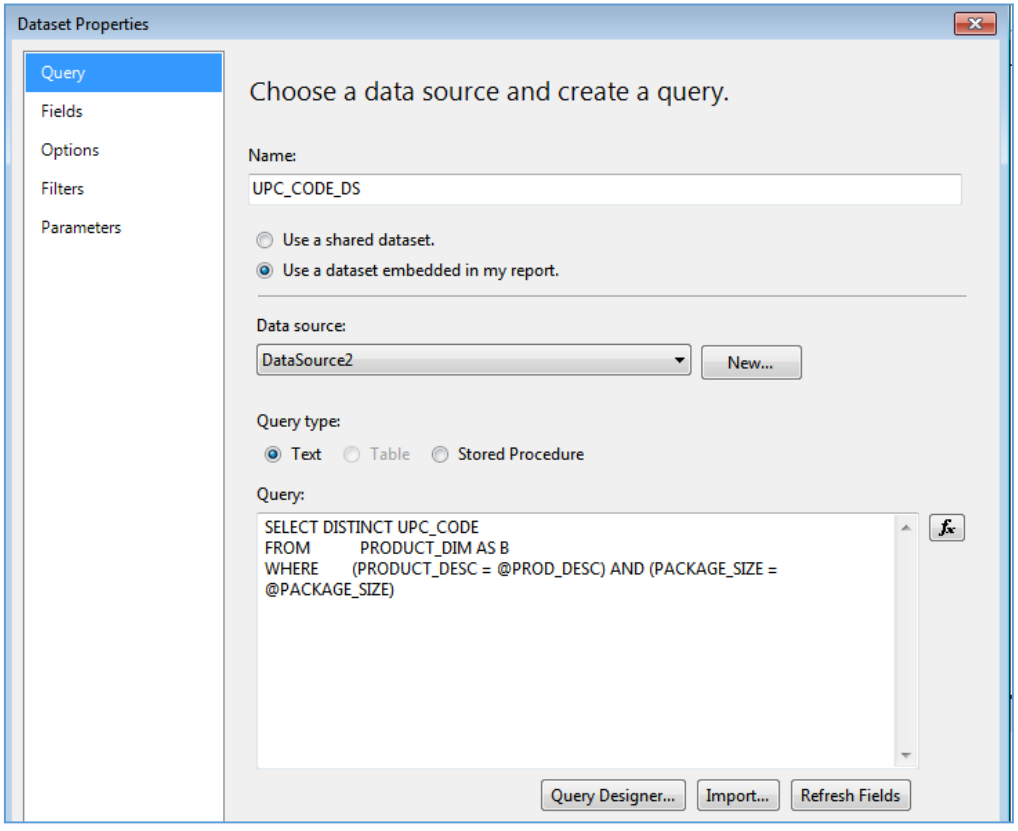
We have defined the three report parameters, the product description, the pack size and the UPC code as the report parameters. Once, the user selects the product description, the text box for the pack sizes is populated. Once the product description and pack sizes are populated, then the list box for UPC codes is generated. The implementation is as shown below.



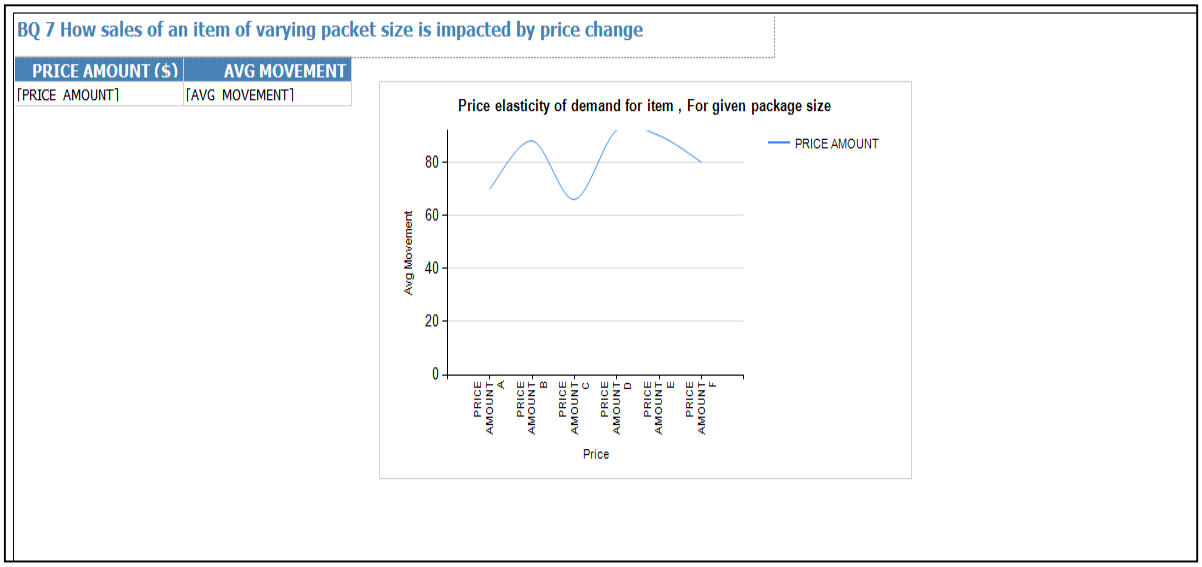
The data sets used for the parameters are as below.



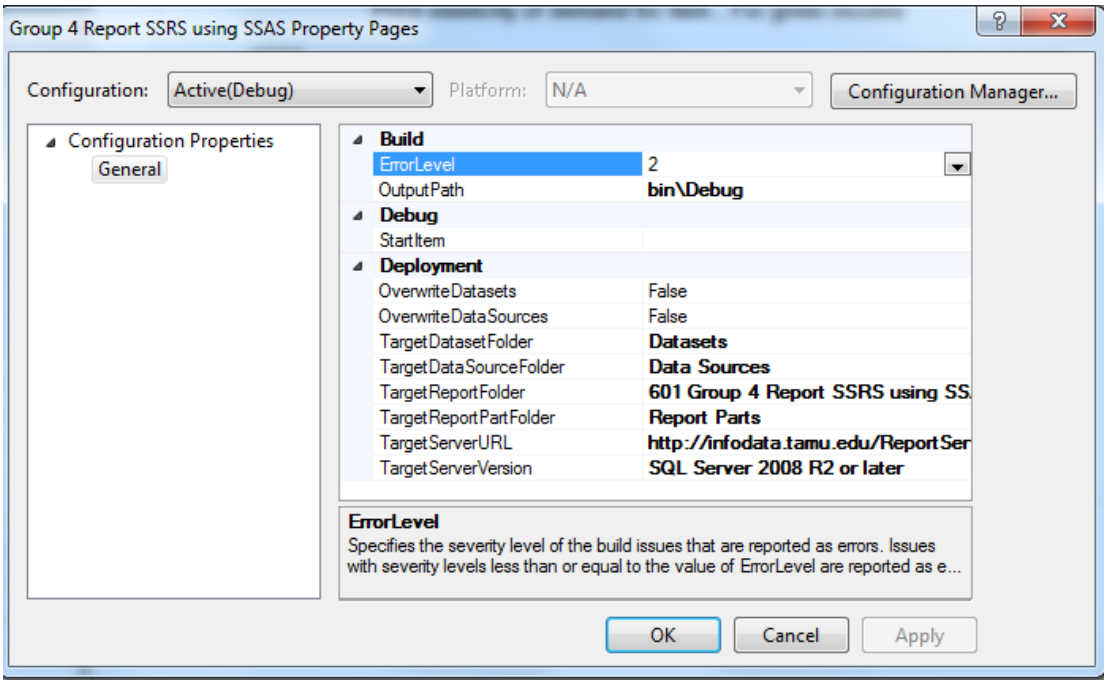

Dataset Properties
Query
Fields
Options
Filters
Parameters
Choose a data source and create a query.
Name:
PACKAGE_SIZE_DS
Use a shared dataset.
Use a dataset embedded in my report.
Data source:
DataSource2
New...
Query type:
Text
Table
Stored Procedure
Query:
SELECT DISTINCT PACKAGE_SIZE
FROM PRODUCT_DIM AS A
WHERE (PRODUCT_DESC = @PROD_DESC)
Query Designer...
Import...
Refresh Fields


Dataset Properties
Query
Fields
Options
Filters
Parameters
Choose a data source and create a query.
Name:
UPC_CODE_DS
Use a shared dataset.
Use a dataset embedded in my report.
Data source:
DataSource2
New...
Query type:
Text
Table
Stored Procedure
Query:
SELECT DISTINCT UPC_CODE
FROM PRODUCT_DIM AS B
WHERE (PRODUCT_DESC = @PROD_DESC) AND (PACKAGE_SIZE =
@PACKAGE_SIZE)
Query Designer...
Import...
Refresh Fields

Report Design:



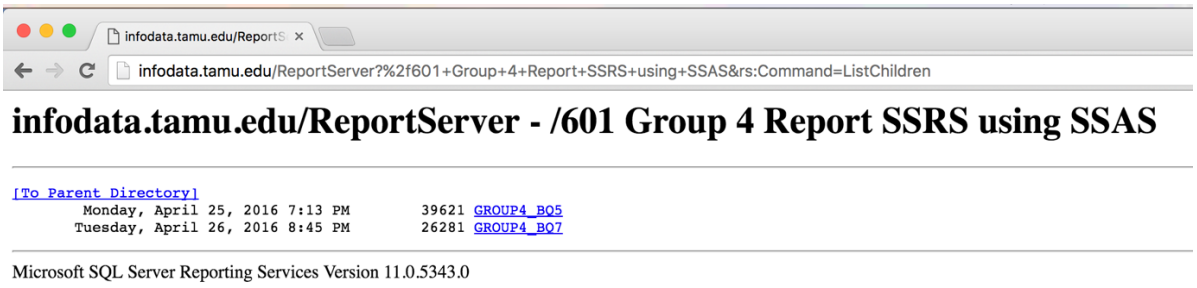
Report Deployment parameters:



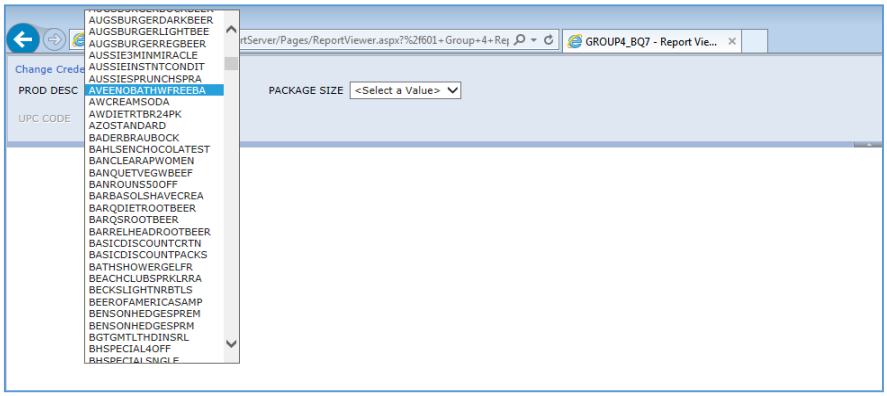
e. **Report output**

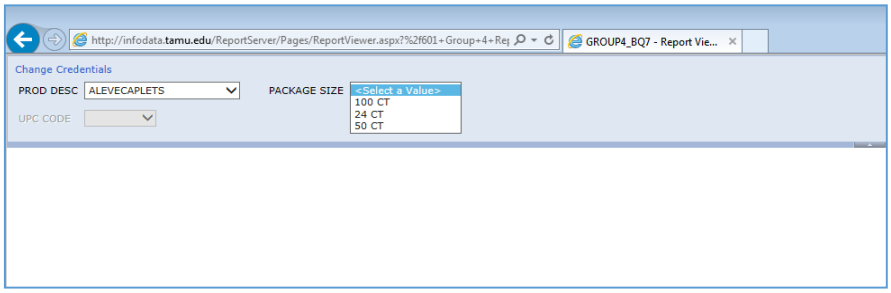
The report, once deployed on the server can be run directly.

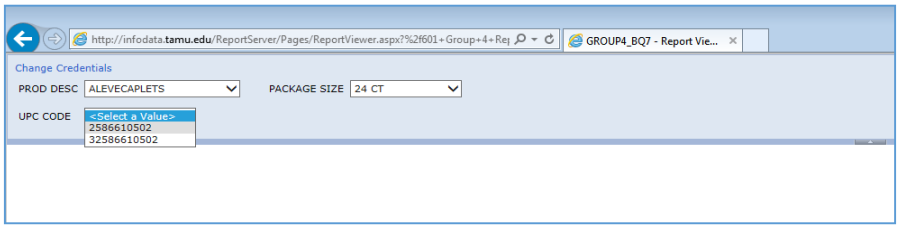
The server screen for the reports is as below:



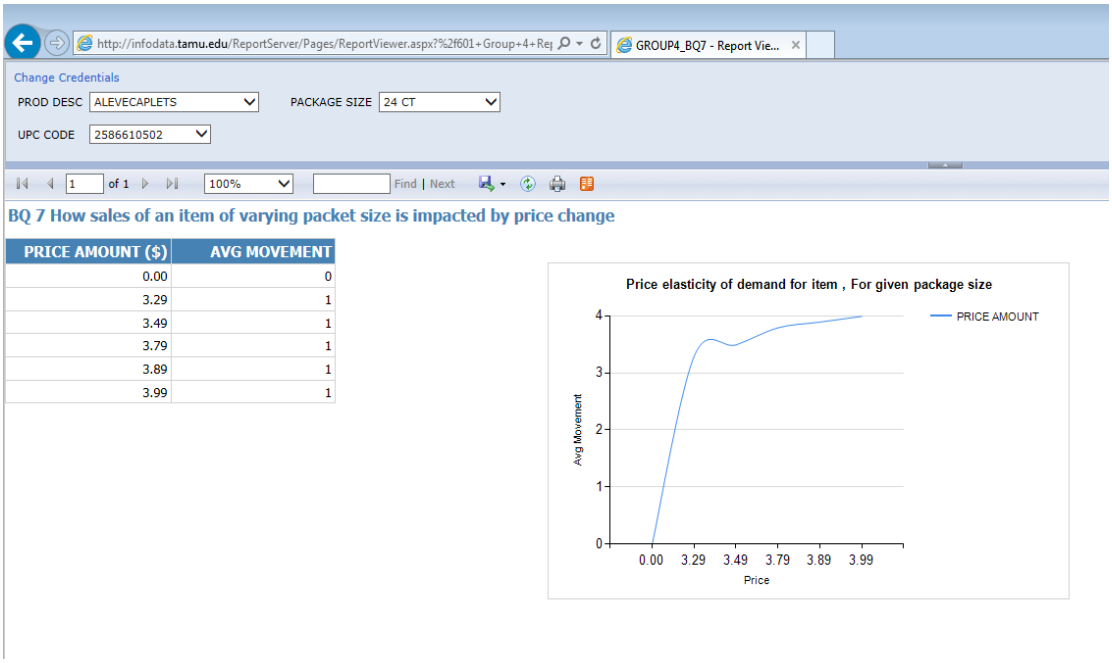
The parameter selection screen for the report is as below. The subsequent parameter box becomes active once the previous box is completed.

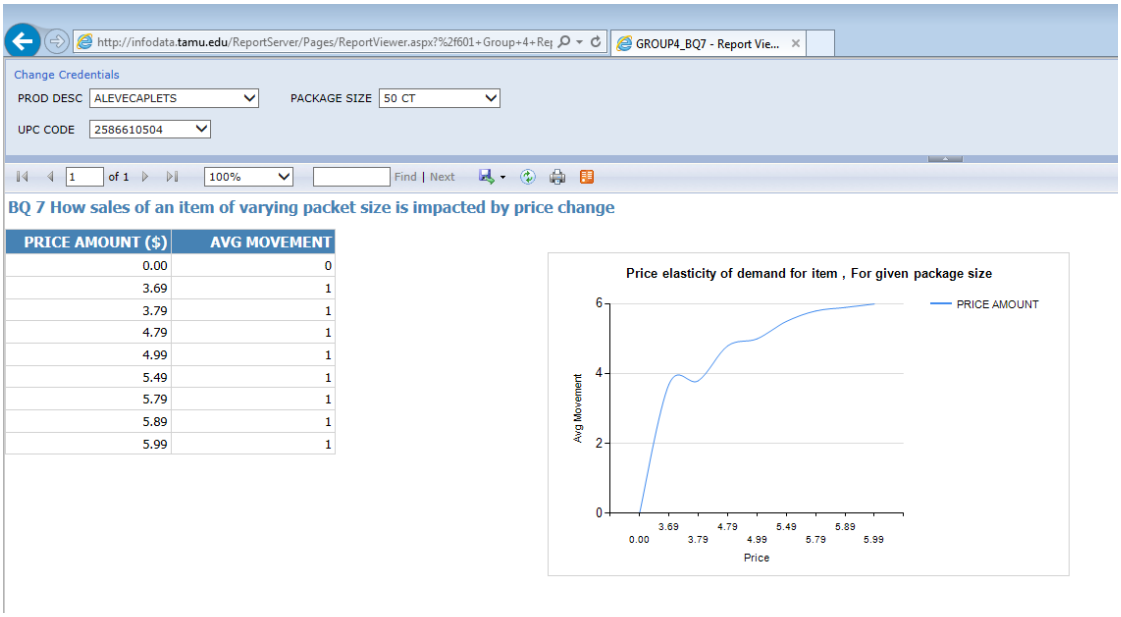


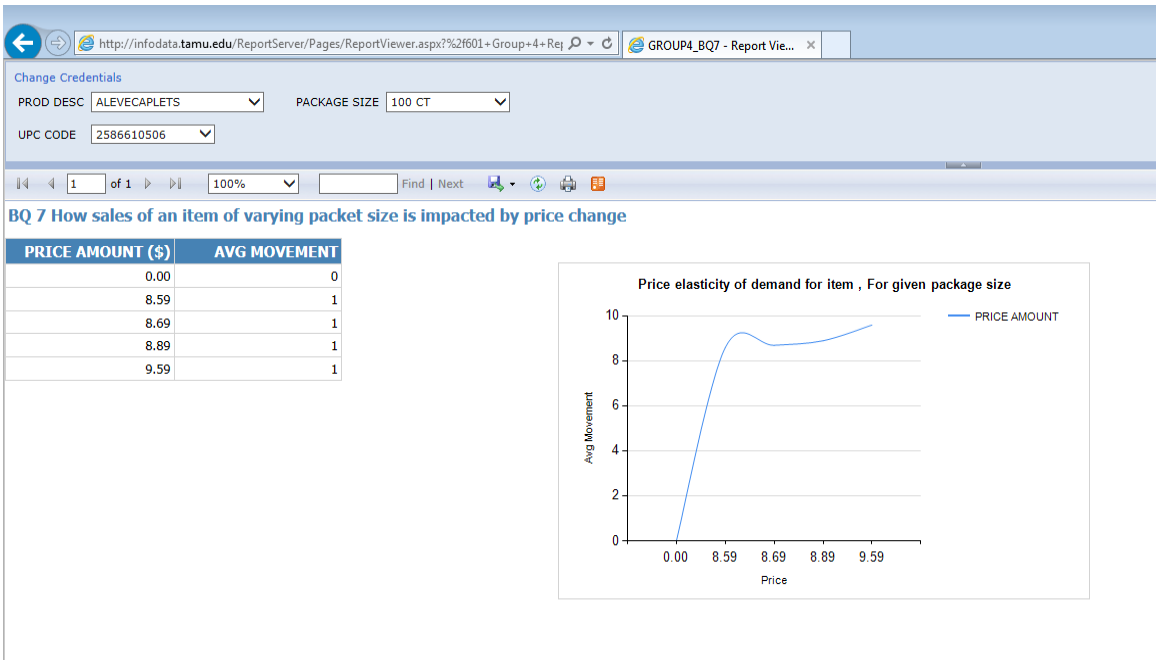




The output of the report for several pack sizes is as below:



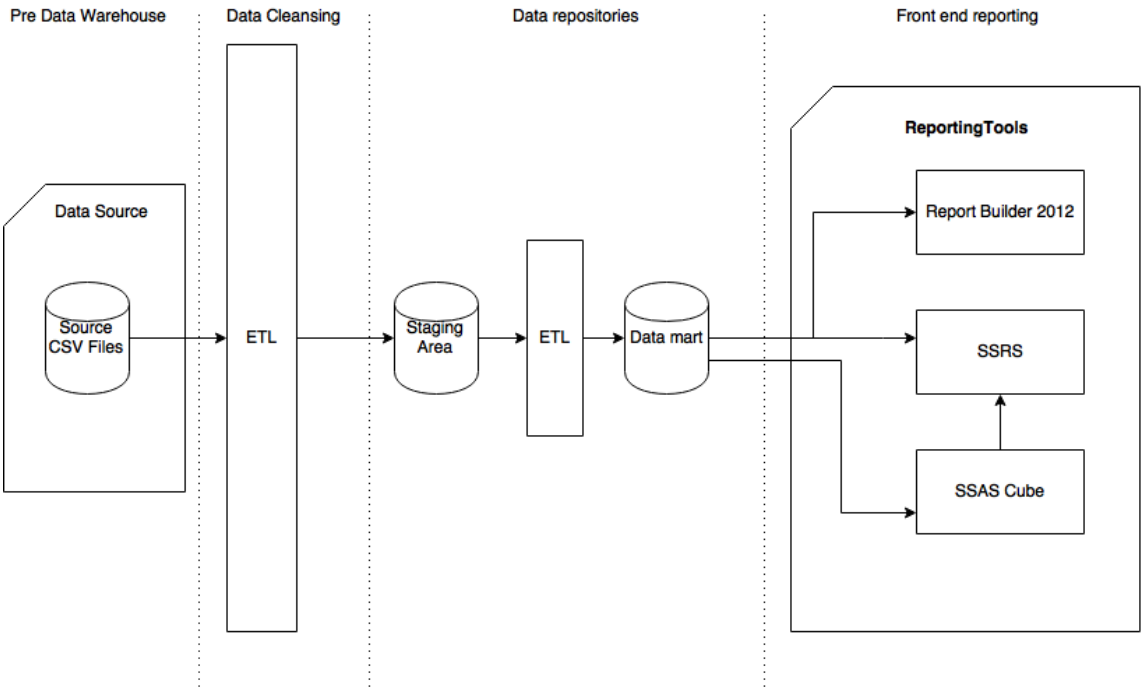




Thus, this report successfully answers the business question and efficiently shows the trends in price elasticity of demand for a product of choice, for the different pack sizes available. This information can be very useful for making several important business decisions like offers and discounts, product promotion etc. And we may also know the quantitative effect the packet sizes have on the overall price elasticity and how it will vary for multiple items.

10. Data warehouse infrastructure and front end tools used:

The project was planned and developed based on the infrastructure as shown below.



Data warehouse infrastructure

The source for the data are the .csv, which were loaded into the data warehouse stage tables after the necessary extract, transform and load operations. Once this data was in the staging areas, rigorous cleansing and transformation tasks were carried out on the data, as described in the previous sections of the report, Finally, the data was loaded in the final data mart areas.

Once the data was in the production data marts, several front end tools were used to address the business questions. They were SQL Server Reporting Services (SSRS), SQL Server Analysis Services (SSAS), to build a multi-dimensional Cube around the data mart. A cube was used for standalone reports as well as in conjunction with SSRS. Report Builder 2012 was used as well. These tools were successful in answering the business questions that were posed at the start of the project. The list business questions answered by each of the tools is as below:

Number	Business Question	Tool used
BQ2	Identify the most concerned areas of losses for each shop for Dominick's Fine Food (DFF) week-wise.	Report Builder 2012
BQ 3	Generate the report depicting the effect of coupon introduction for increasing the customer count in the store.	SSRS
BQ 4	Identifying the contribution of each product category in the overall sales for DFF	SSAS Cube
BQ 5	Identifying price elasticity of demand for a particular item, w.r.t. income levels	SSAS Cube with SSRS
BQ 7	How sales of an item of varying packet size is impacted by price change	SSRS

11. References

1. Dubé, Jean-Pierre and Sachin Gupta, "Cross-Brand Pass-Through in Supermarket Pricing," *Marketing Science*
2. Pofahl, Geoffrey M. and Timothy J. Richards, "The Valuation of New Products in Attribute Space," *American Journal of Agricultural Economics*
3. Slotegraff, Rebecca J. and Keon Pauwels, "The Impact of Brand Equity and Innovation on the Long-term Effectiveness of Promotions," *Journal of Marketing Research*
4. Levy, Daniel, Haipeng (Allan) Chen, Georg Muller, Shantanu Dutta, and Mark Bergen, "Holiday Price Rigidity and Cost of Price Adjustment,"
5. Song, Inseong and Pradeep K. Chintagunta, "Measuring Cross-Category Price Effects with Aggregate Store Data," Management Science

12. Work break down

This sections includes work break down structure for analysis and brainstorming over the business case translations for the delivery. Below is the detailed description of the works involved:

Task No	Task Defined	Task Owner	Time Taken
1.	Data mart analysis	Gaurav, Pratik	2 days
2.	Loading stage data	Gaurav and Pratik	5 days
3.	Data cleansing	Gaurav, Pratik	1 days
4.	Making transformation rules	Pratik	1 day
5.	SSIS package for transform, load	Gaurav, Pratik	2 days
6.	SQL and tables creation	Gaurav, Varun	0.5 days
7.	Report builder 3.0 Implementation	Pratik	2 days
8.	SSRS Implementation	Gaurav	2 days
9.	SSAS Implementation	Gaurav, Pratik	3 days
10.	Power Point presentation	Varun	1 day
11.	Documentation	Pratik, Gaurav, Varun	2 day