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

Final Report

Gonnade, Prajwal

Nayak, Supreet

Rana, Vijaylakshmi

May 2, 2016

Texas A&M University

Credentials for accessing the data warehouse, reporting and analysis services

SQL Server Authentication

The above credentials can be used to access:

1) Database Engine in the Microsoft SQL Server Management Studio

   Staging area: **601Group1_staging_area**
   Data warehouse area: **1. 601Group1_CategorySales_DW_area**
   2. 601Group1_ProductSales_DW_area

2) Analysis Services of the SQL Server

   Cube deployed for SSAS:
   Question 2 - **601_Group1_Q2**
   Question 5 - **601_Group1_Q5**

3) http://infodata.tamu.edu/ReportServer

   Project folder: **601_Group1**

   The folder contains reports for the five business questions as follows:

infodata.tamu.edu/ReportServer - /601_Group1

[To Parent Directory]

Saturday, April 23, 2016 4:29 PM	46899	Question1
Saturday, April 23, 2016 9:24 PM	42126	Question2
Saturday, April 23, 2016 6:12 PM	22474	Question3
Saturday, April 23, 2016 5:35 PM	21050	Question4

Microsoft SQL Server Reporting Services Version 11.0.5343.0

Table of Contents

A. Introduction

Data Warehouse is a system which serves the purpose of reporting and data analysis. It acts as a central repository of integrated data from multiple sources containing historical as well as current information. The system is used to reach answers to a variety of business questions to support decisions and planning in conjunction with OLAP i.e. Online Analytical Processing technology. As a part of the initial phase of this project, this report contains the outcomes of the requirements gathering process to answer the business questions as a decision support.

In this project, the data warehouse which will be used for academic purpose belongs to Dominick's Finer Foods (DFF) retail store. Dominick's was a retail store chain based out of Chicago area which was closed in 2013 due to poor sales and low performance. This data is part of the partnership between Chicago Booth and Dominick's Finer Foods used to conduct store level research for shelf management and pricing. This vast data contains sales, product, store details of the business done by DFF over the period of nine years. As a part of the data warehouse project, we will be studying this data and try to solve business questions which could have helped DFF to improve its sales, efficiency, and customer base.

1. Challenges faced during project duration

Every project brings forth their set of unique challenges and this project was no different. It had its own set of challenges and tasks which forced us to deeply delve into our thoughts and think innovatively.

Following are some of the challenges we faced during the first phase of the project –

1. As the data was huge and dirty, the team had to spend maximum time and energy on understanding the data and finding logical connections between the disparate parts of this data.
2. A lot of time was spent on creating the Entity-Relationship diagram, as moving forward this was a critical step and any ambiguity at this step could halt the next step of creating business questions.
3. The data was in multiple file format, for example, CSV, TXT, and HTML, which provides a challenge to convert them to Excel format and study.
4. Another challenge we faced was to decide on the priority and effect of the business questions we created. We tried to brainstorm on every business question and find its utility in the modern business world.

After downloading the data, we imported them to Excel files and studied those using Excel features like Charts, Pivot Tables as well as using JMP software as recommended by the professor. We studied the metadata using Dominick's Manual and Codebook document which helped us in designing the business questions and support them with our own analysis.

As advised by the professor, we searched online and on Chicago Booth website and studied the research papers based on this data and acquired the domain understanding necessary for building the business questions. The research papers acted as an important tool for understanding the data and connecting them to current business affairs.

2. Details about the Data

2.1. Understanding of the Data

From sources like research papers, Dominick's Manual and Codebook, and data itself, we understand that Dominick's data contains store-level scanner data which covers a period of around nine years. This sales data corresponds to approximately 3500 UPC's which are divided into 29 different categories and belong to 100 stores around the United States.

The following information is pertaining to the Dominick's data files –

a) Customer Count Files

- The customer count files contain the information of the customer visits and purchase in a store on a daily basis. This file contains the customer counts for every store i.e. in-store traffic.
- This file contains information about coupons redeemed, product sales, data of sale, etc.

b) Store-Specific Demographics

- This data contains the store-specific demographic data. This data originates from the census conducted by the US Government for the Chicago Metropolitan area.
- This data provides the information about the various age groups of customers, their household information, household income, dependent members in the household, whether the customer is unemployed or retired etc.
- This information helps in strategizing on which consumer to target based on age, income, employment status, sex, shopping style etc. and also which products to market for disparate customer types.

c) UPC Files

- This file contains information of the UPC belonging to each category. This information could be Product Name, Size, Description, the number of items in bundled case, UPC number of a product, etc.

d) Movement Files

- The Movement files contain information of the weekly sales data of the products belonging to various categories. This sales data is for each UPC of the category over five years.
- The files contain information like retail price of the product, a number of units sold, gross margin of that UPC product, etc.

- As far as business questions are concerned, these files contain lots if useful information to answer business questions pertaining to holiday sales, weekly sales, demographic-level sales, etc.

e) Week's Decode table

- These files contain information on the codes that symbolizes the week when the sales data is recorded. This data can be found in Dominick's Manual and Codebook document.
- This information is particularly useful to find period-based data to be included in sales reports which could be quarterly, semi-annually etc. and also, for comparison between sales during holidays and normal days.

2.2. Metadata for all the OLTP source files

a) Customer Count Files

- The Customer Count file contains information on store traffic and coupon usage, by store.

b) Store-Specific Demographics

- The demographics file contains detailed information based on the US Government census.

c) UPC Files

- The UPC files contain one record for each UPC in a category (xxx stands for the category acronym). They contain information about product name, size, commodity code, etc. The files are sorted by UPC.

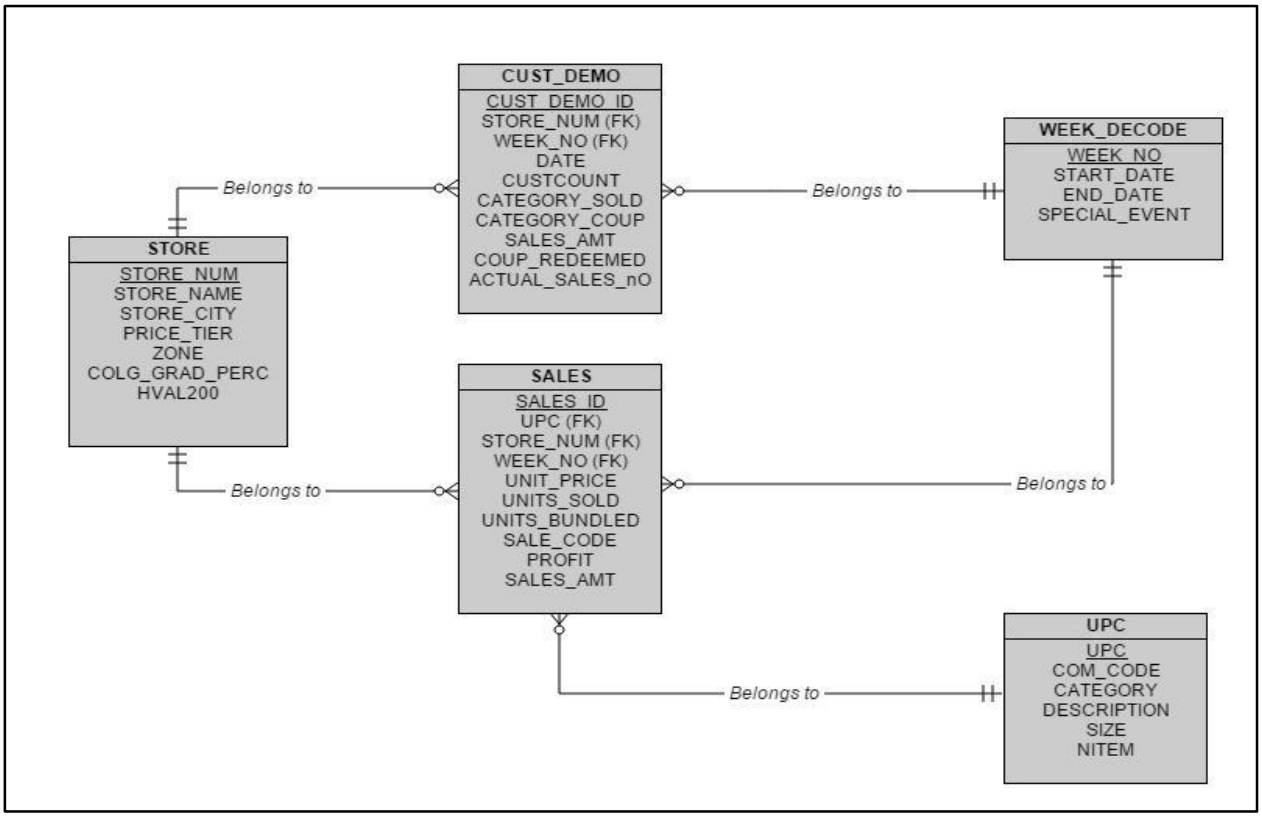
d) Movement Files

- The movement files contain weekly sales data for each UPC in each store for over 5 years. The variables included in these files comprise: price, unit sold, profit margin, deal code, etc. The files are sorted by UPC, store, week.

e) Week's Decode table

- The SAS files contain a week variable that codes the week for which a data point is recorded.

2.3. Entity-Relationship Diagram



3. Domain Understanding

a) Kamakura, Wagner A. and Wooseong Kang, "Chain-wide and Store-level Analysis for Cross-category Management," *Journal of Retailing*, 83 (February 2007): 159-170.

This research paper talks about the effect of price promotions of a particular product across the organization. The author analyzes the chain-level effect of the promotions across the global organization. When a promotion is introduced for a particular product, it is imperative to analyze its cross-category impact. The model proposed by the author analyzes the store-level and cross-category impact of price promotions. It applies the proposed model to Dominick's retail chain and compares them with other approaches to show the benefits of the model over other competing models.

The research paper investigates the Dominick's data with a focus on sales and prices of toothpaste and toothbrush across different brands and stores. The discount by Dominick's in the toothpaste category can have both positive and negative effects on various toothpaste brands like Pepsodent, Aquafresh etc. This paper helps in understanding the effect which sale type of bonus buy, price reduction, the discount has across brands and across different categories.

b) Levy, Daniel, Haipeng (Allan) Chen, Georg Muller, Shantanu Dutta, and Mark Bergen, "Holiday Price Rigidity and Cost of Price Adjustment," Economica 77 (2010): 172-198.

This research paper explains the pricing decisions which retail chain managers have to take during holiday periods. The author argues the increase in the cost of price adjustment during holidays. As during the holidays, people tend to buy more products than non-holiday periods, there are an increase opportunity costs of price adjustments as the tasks like restocking of shelves, customer service, and complaint resolving, logistics like delivery and invoices, take priority.

This paper uses the large scanner price and cost data of Dominic's Finer Foods which is available and concludes that it is less likely that prices change during holiday period than non-holiday periods. Also, as the quality of the products is the same, it's not a factor which reasons the differences between price adjustments between both the periods. From our analysis of the data, we also observed the increase in sales of various products across different categories is more during holidays when compared to non-holiday periods.

c) Pofahl, Geoffrey M. and Timothy J. Richards, "The Valuation of New Products in Attribute Space," American Journal of Agricultural Economics, Vol. 91 (May 2009): 402-415.

The primary objective of the research in this paper is to estimate customer welfare changes associated with the introduction of several food products. New products tend to have higher margins if they contain innovative new features and they provide customers

with a new addition to their choice set. It also helps manufacturers to compete against others' product lines or retailers' private label products. Since new products are costly to develop and bring to market, hence manufacturers and retailers need to understand the incremental value of introducing a new product to make proper investment decisions.

In this research, differences in product attributes for newly introduced bottle juice/juice-drink products were taken into consideration such as prices, quantities, percentage markups, discount information, brand, flavor, and sugar. It was observed that a shopper trying to decide between two brands of a particular juice will likely pay much closer attention to attributes other than flavor, than trying to decide between alternative flavors. Market trends at the time of this research also indicated that bottled juice consumption was on the rise during that period and the nonalcoholic beverage complex was in the midst of a drastic transition in consumer thinking about the role of such drink products in their everyday diet.

d) Slotegraff, Rebecca J. and Keon Pauwels, "The Impact of Brand Equity and Innovation on the Long-term Effectiveness of Promotions," Journal of Marketing Research, Vol. XLV (June 2008): 293-306.

The paper focuses on the permanent and cumulative sales effects from marketing promotions are greater for brands with higher equity and more product introductions whereas brands with low equity gain greater benefits from product introductions. This study has used the Dominick's Finer Foods project to supplement data for this research.

The research shows that promotional efforts are recognized as a potent tool for managing brands with in-store displays, feature advertising and temporary price reductions in a traditional promotional mix. A long-term sales impact may appear in two forms:

- **Permanent** effects which represent a true change in baseline sales
- **Cumulative** effects which summarize the over-time changes

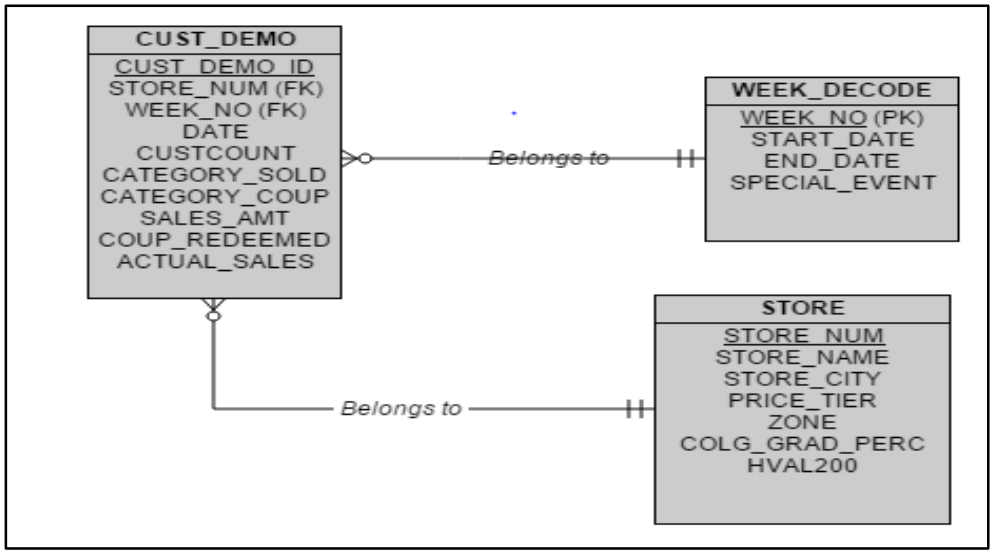
In conclusion, this research established that a brand’s equity and new product introductions play a significant role in the long-term sales elasticity and unit effects from its marketing promotions.

B. Business Questions and their substantiations and explanations

1. Business Questions

1) What is the trend of Beer Sales during Thanksgiving's week for the entire duration?

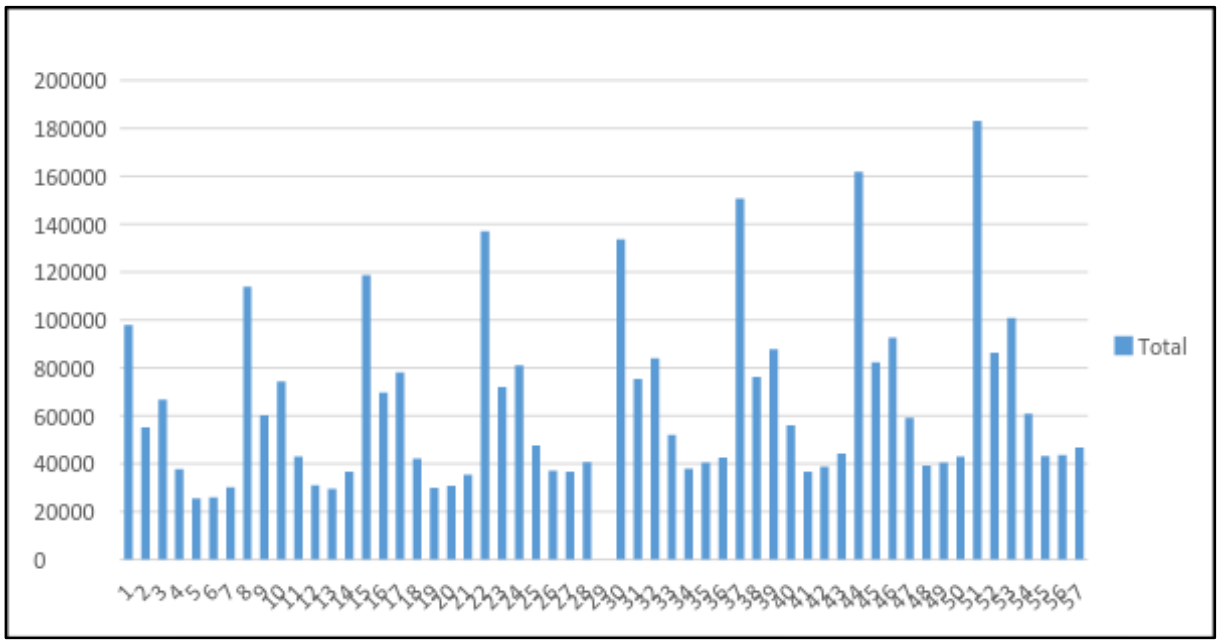
Entity-Relationship Diagram:



Snapshot of Relevant Data:

Row Labels	Average of BEER
11	**504.5967113**
11/22/1989	1018.924167
11/23/1989	574.3886458
11/24/1989	695.5529167
11/25/1989	392.3007292
11/26/1989	266.4779167
11/27/1989	270.7030208
11/28/1989	313.8295833
63	**561.1592785**
11/21/1990	1150.49899
11/22/1990	609.1162626
11/23/1990	750.7169697
11/24/1990	435.1788889
11/25/1990	312.2368687
11/26/1990	298.8643434
11/27/1990	371.5026263

Pivot Chart:

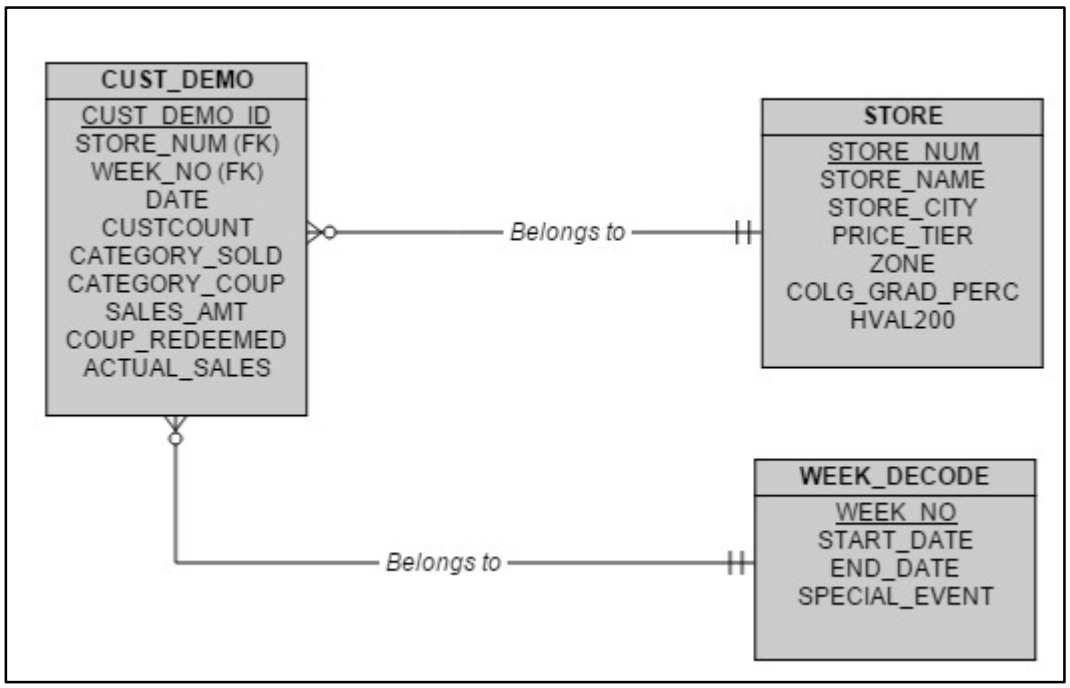


Justification:

A particular product can have increased sales during the holiday period irrespective of providing discounts or bonus coupons. We are trying to capture the trend of one such particular product during a particular holiday period. We realized that the beer sales tend to increase during the thanksgiving. The above graph shows the results of beer sales during the thanksgiving week. Our observations based on the limited data we analyzed concluded that the beer sales tend to be the highest during the initial phase of the week and then tends to decrease continuously as the week progresses. This analysis helps businesses to determine how to handle the holiday weeks in a better way. In our analysis, for example, the business can store the beers in advance before the thanksgiving week starts so as to match the sales. Additionally, increase in prices or promotions sales for products with poor sales can be planned so as to benefit the business.

2) **How are the average price and sales of a particular product changing according to different zones (Fish and Fish Coupon)**

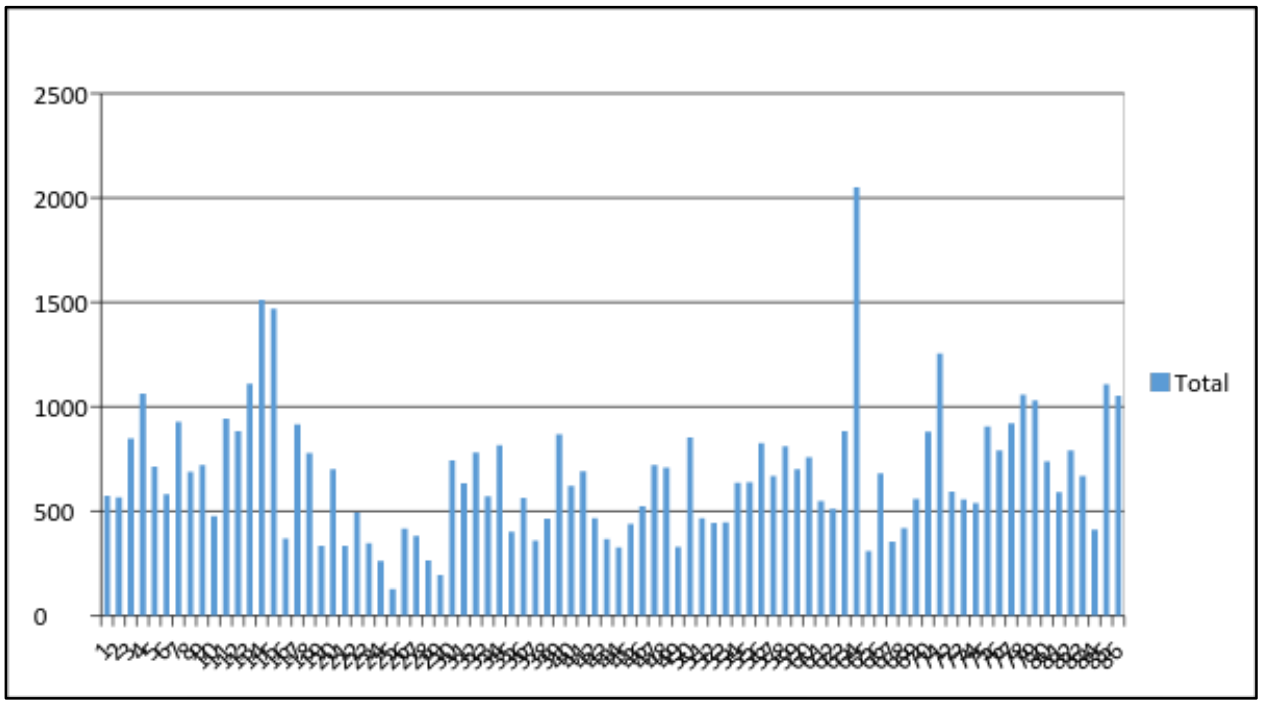
Entity-Relationship Diagram:



Snapshot of Data:

Row Labels	Average of SALES
1	**862.8997954**
2	573.829762
14	566.0905566
32	848.35597
52	1064.182992
62	714.0295339
68	582.0048276
71	928.9794036
72	688.8638244
93	721.5116947
95	476.4069231
111	944.2366532
123	884.0423426
124	1109.852946
130	1512.311731
137	1469.159142

Pivot Chart:



Justification:

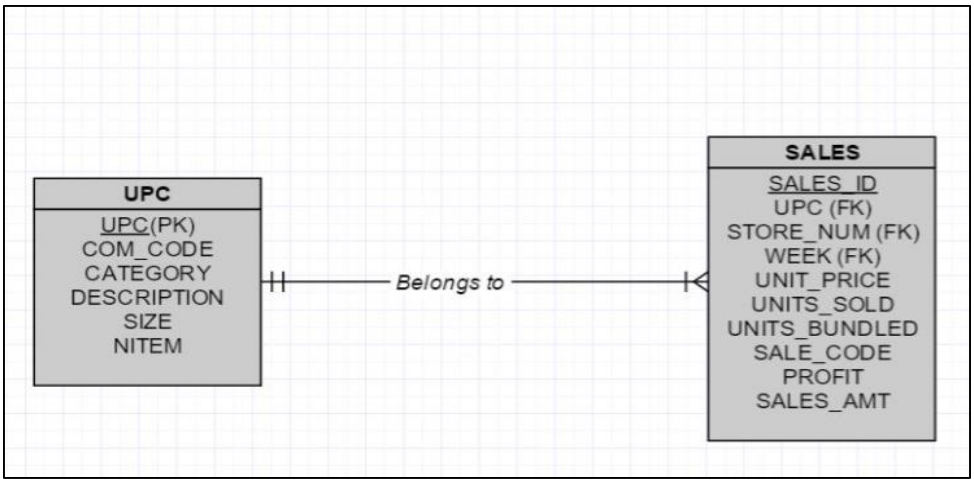
The Dominick data includes store being divided into 16 different zones. Zones help business to distribute their resources in an effective manner. Tracking the trend of product sales according to a zone hence becomes an important factor. Here, we aim to track the highest average sales for a particular product, in our case fish, in a particular zone. The idea behind finding the highest average sales using the data was twofold:

- To compare the sales trend between different zones, and
- To find the trend in a particular zone for that particular product.

Our sample data analysis for fish shows that fish sales trend is highest in zone 12. This would help the business to concentrate on fish sales in zone 12 so as to benefit the business. It also means that effective marketing techniques need to be used to increase the sales in other zones. The graph also shows the stores with highest average sales in each zone which would help the business to decide which store needs to be given emphasis for fish sales based on company sales strategy.

3) Compare the effect of Bonus buy and Price Reduction in Analgesics in different zones

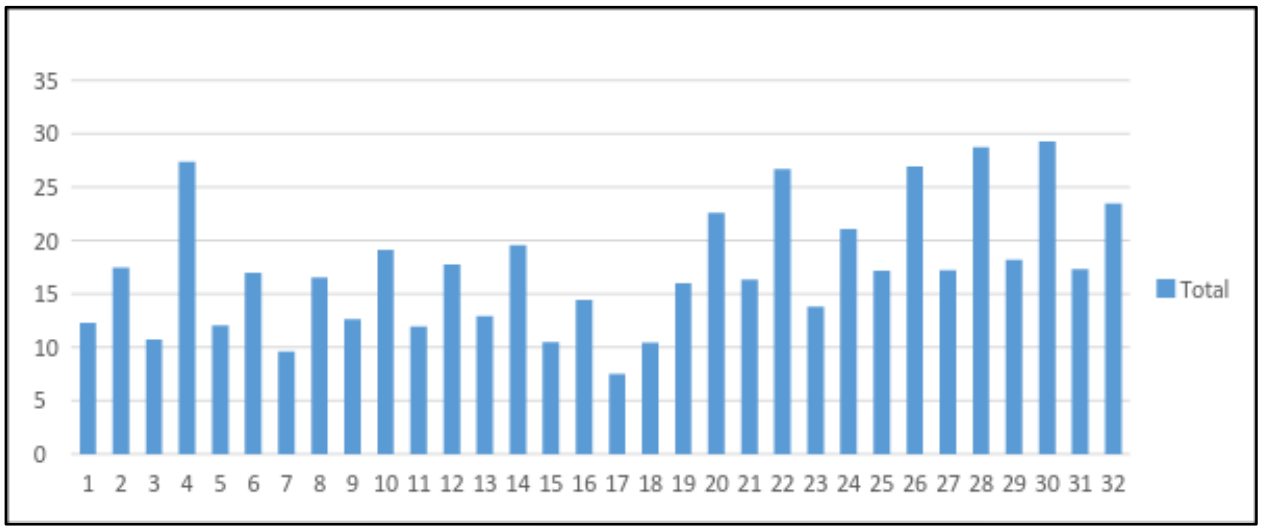
Entity-Relationship Diagram:



Snapshot of Data:

Row Labels	Average of TOTAL SALES
1	**13.72900129**
B	12.30985378
S	17.44449113
2	**16.25669448**
B	10.75234709
S	27.38729131
3	**13.40145062**
B	12.02692308
S	16.97522222
4	**11.75391304**
B	9.584368932
S	16.55956989
5	**14.63545973**
B	12.64932238
S	19.14477855
6	**13.77387755**
B	11.92836275
S	17.77411609
7	**14.87416178**
B	12.92252492
S	19.55498008
8	**11.71569869**
B	10.46370607
S	14.41827586

Pivot Table:

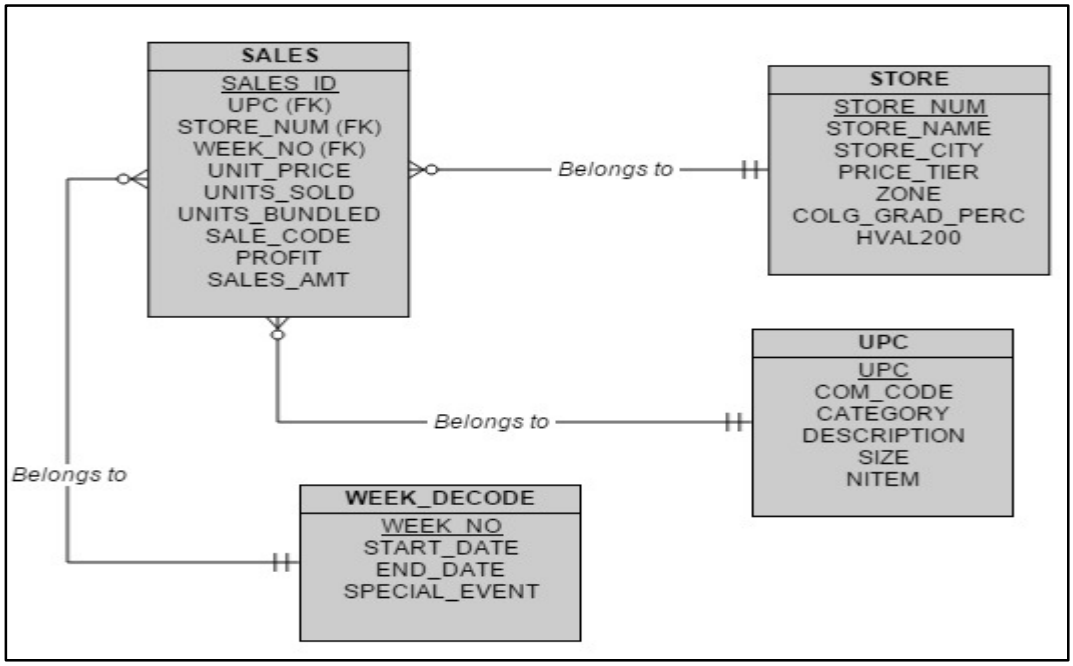


Justification:

The data provides us with the insights of the sales done by the DFF by offering simple price reduction and bonus buy option. The above graph shows the comparison between the two methods of promotion of sales between different stores. The idea here was to find out which method helps to increase the sales of a particular product. In our case, we analyzed the data for the average of total sales for analgesics of the bonus buy option and simple price reduction. We observed that the average sales when a bonus is offered are less than the average sales when there is a simple price reduction. This could help the business in deciding that providing simple price reduction is more beneficial to DFF than providing a bonus buy option. A similar method can also be used for different products so as to compare the method of promotion of sales. The data can also be compared to average sales without promotion so as to get a clear picture of how simple price reduction or bonus buy option improves the sales value.

4) What are the sales of cigarettes in stores with college graduates greater than the average across Chicago?

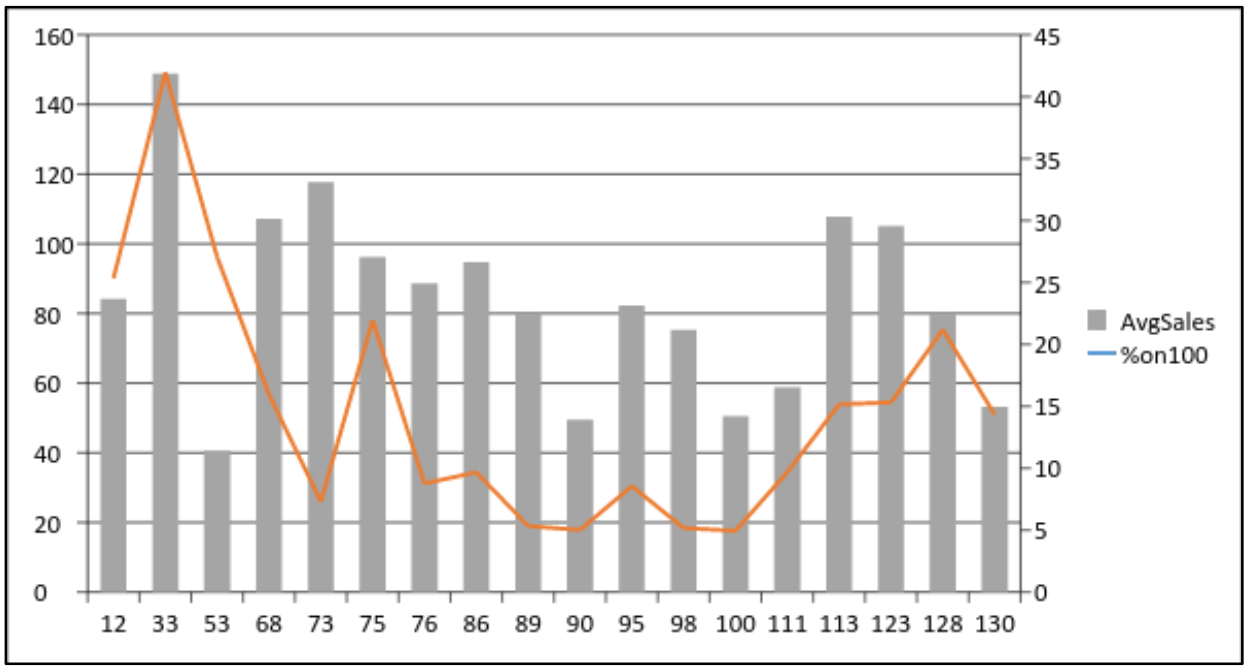
Entity-Relationship Diagram:



Snapshot of Data:

CITY	STORE	EDUC	%on100	GRTNAVG	Sales	AvgSales
CHICAGO	12	0.253413	25.3413	Y	242077.5	84.25
CHICAGO	33	0.419688	41.9688	Y	367613.9	148.83
CHICAGO	53	0.270383	27.03835	Y	96362.23	40.66
CHICAGO	68	0.159722	15.97215	Y	149701.2	107.16
CHICAGO	73	0.073054	7.305396	N	163227.4	117.68
CHICAGO	75	0.219548	21.95485	Y	314958.3	96.25
CHICAGO	76	0.087712	8.771179	N	115851.9	88.7
CHICAGO	86	0.096764	9.676392	N	296817	94.7
CHICAGO	89	0.053349	5.334944	N	112074.1	80.11
CHICAGO	90	0.050193	5.019345	N	89929.54	49.46
CHICAGO	95	0.085642	8.564208	N	485981.9	82.31
CHICAGO	98	0.051703	5.170282	N	326650.6	75.24
CHICAGO	100	0.04955	4.955029	N	233031.3	50.56
CHICAGO	111	0.096929	9.692892	N	4182.86	58.91
CHICAGO	113	0.151592	15.15924	N	231204.9	107.78
CHICAGO	123	0.153191	15.31915	N	297219.3	105.06
CHICAGO	128	0.211897	21.18969	Y	417447.1	80.12
CHICAGO	130	0.143407	14.34068	N	84738.67	53.22

Pivot Table:

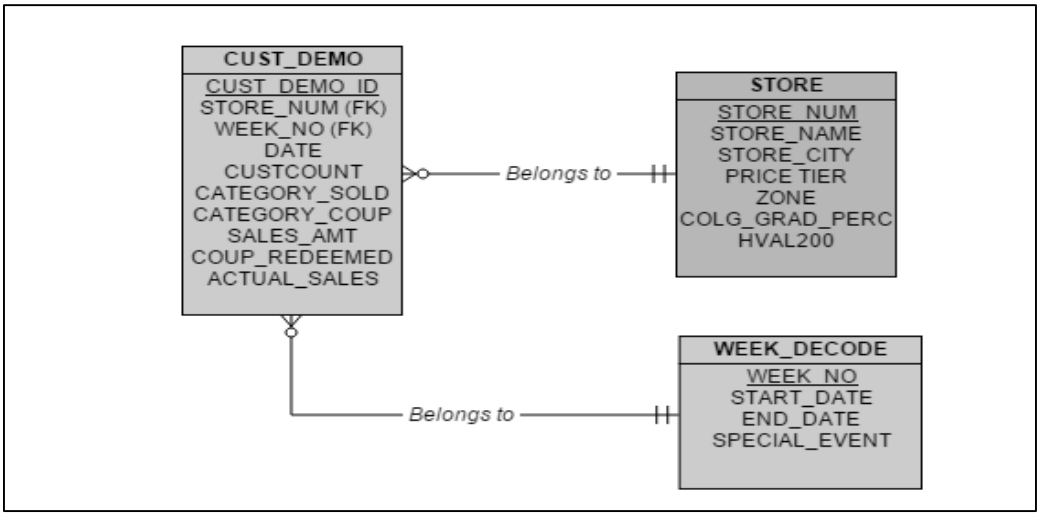


Justification:

The data provides us with demographics of each store with details regarding the percentage of people above or below a certain age level, bifurcation based on family earnings and a variety of other factors. We tried to gauge how the sale of a product in a particular store would be influenced by its demography. In our case, we intended to find sales of cigarettes being influenced by the footfalls of college graduates at a particular store. Analysis of the sample data revealed that higher the number of college graduate visitors in a store, higher is the sale of cigarettes. This would help the business to sell or promote a particular product in a store based on demographics. A similar combination of product and demography, for example, jewelry with working women could be used to benefit DFF in the longer run. Such analysis could also help other stakeholders like manufacturers, employees to promote or demote any particular product in a particular area for future benefits.

5) **What is the sale of groceries in 4 different price tiers of a store?**

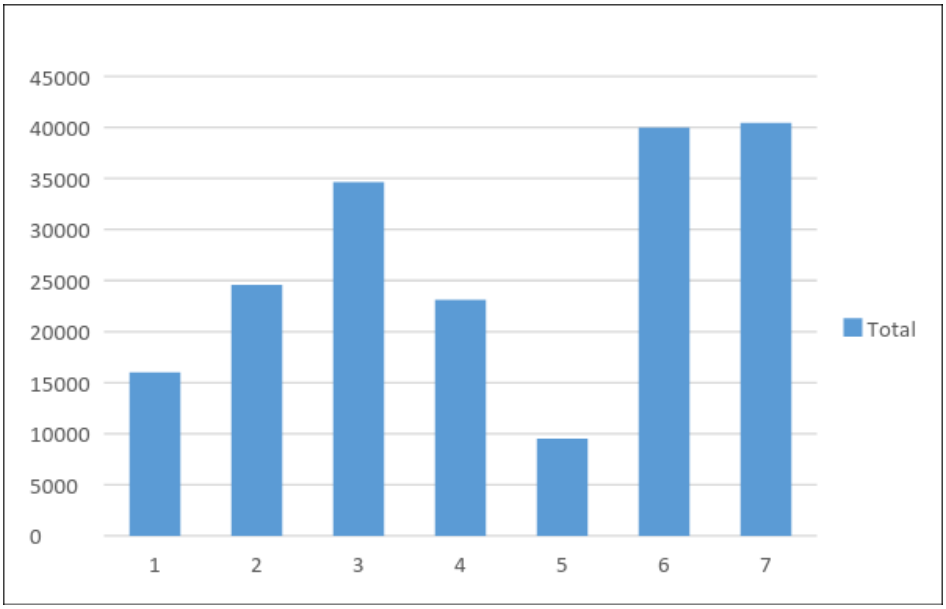
Entity Relationship Diagram:



Snapshot of Relevant Data:

Row Labels	Average of GROCERY
CubFighter	**15991.12**
21	15991.12
High	**31504.31688**
2	24565.264
12	34658.43182
Low	**23115.71692**
18	23115.71692
Medium	**22558.5597**
4	9498.627895
5	40003.6
8	40437.83778
Grand Total	**23956.21171**

Pivot Chart:

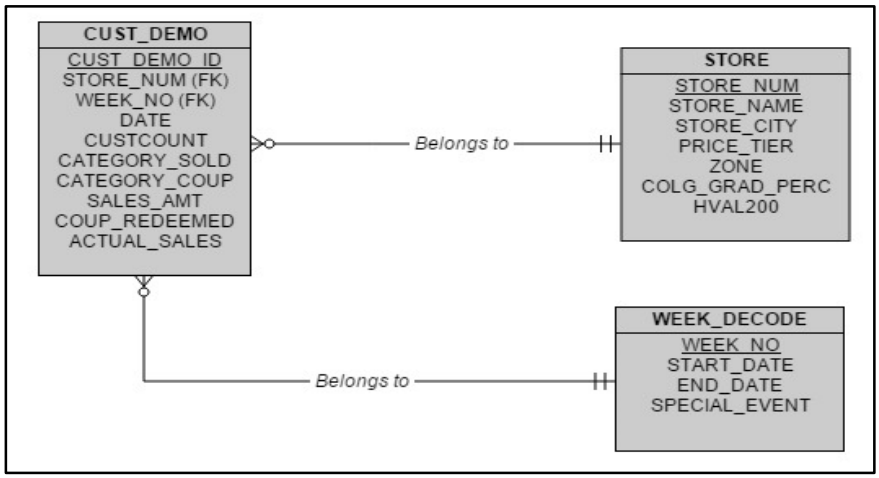


Justification:

According to the given data, the stores have been divided into 4 price tiers which are High, Low, Medium and Cub Fighter. Here, we have taken the average sales of the groceries across different tiers. The graph above shows the trend of grocery sales across these different tiers. This question would help the organization determine the kind of stocks a particular product needs in a particular tier. Additionally, it would also help the business to focus on the sales and marketing of the product in the tier where the sales are high. The data can also be used to find regions where the product sales are low and then determine the cause of low sales. Necessary steps could then be taken to increase the sales in that particular region. We have just taken a particular product to demonstrate the idea. The same method can be implemented on various other products to determine the sales trend in different tiers.

6) **Plot the average profit margin for cigarettes across all the stores. Determine the average of profit for the sales of cigarettes and the stores which are below the average.**

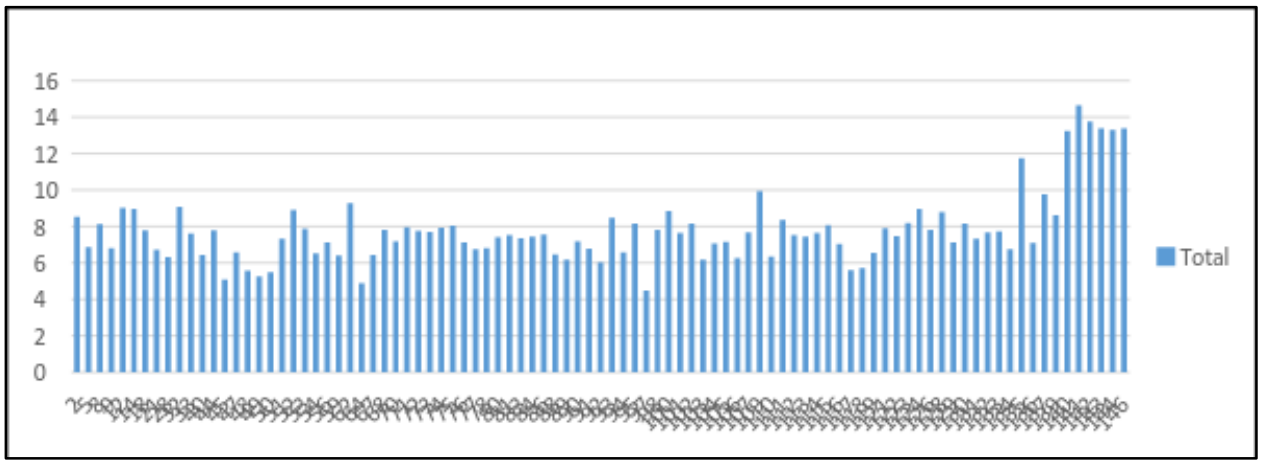
Entity-Relationship Diagram:



Snapshot of Data:

Row Labels	Average of PROFIT
2	8.539907339
5	6.873900968
8	8.138793694
9	6.80818109
12	9.032285297
14	8.969601182
18	7.77645678
21	6.720877402
28	6.322060367
32	9.070253933
33	7.610049429
40	6.444799773
44	7.793772036
45	5.077422832
47	6.571463774
48	5.564828558
49	5.254940924
50	5.475778741
51	7.32772094
52	8.891958938
53	7.863023452
54	6.509922814
56	7.133478583
59	6.416814251
62	9.288948515
64	4.878268604

Pivot Chart:

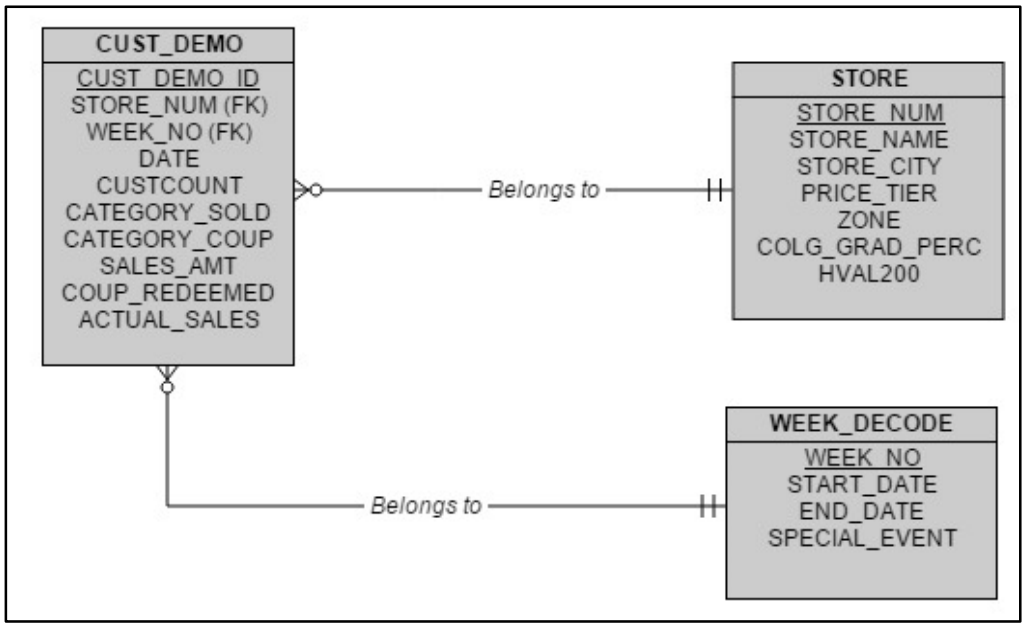


Justification:

The movement data in the given dataset provides us with the profit DFF receives per dollar in a particular product. We planned to analyze the profit and find the highest profit margin of the particular product across the stores. While our question points to the highest profit margin, the graph shows us the trend of profit for the particular product across the stores. The main focus of any business across the world is to secure profit for a particular product they trade in. The above graph shows the profit secured by selling cigarettes across the business line. The average profit secured by store 141 is highest at 14.63. This helps the business to know the store where providing more focus would boost the sales of the product. Additionally, the trend could be used to see the store with sales lower than a particular value e.g.: 7.00 and provide discounts or promotions in such store to boost the sale.

7) **What is the trend of Camera sales from the year 1990 to 1996?**

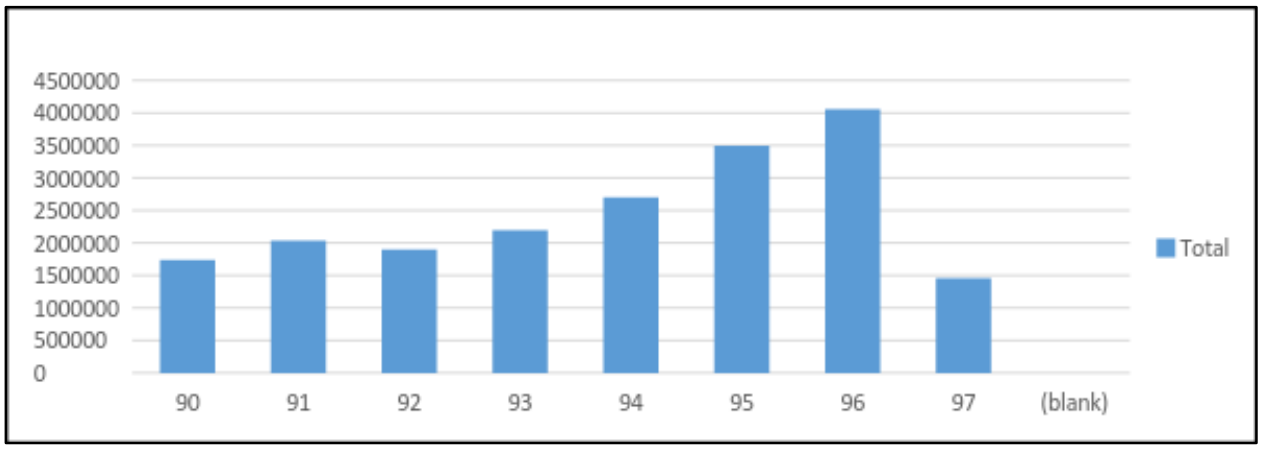
Entity-Relationship Diagram:



Snapshot of Data:

Row Labels	Sum of CAMERA
90	1742050.45
91	2037994.35
92	1902048.53
93	2196794.62
94	2702151.08
95	3499491.37
96	4057030.77
97	1453730.75
(blank)	
Grand Total	**19591291.92**

Pivot Chart:

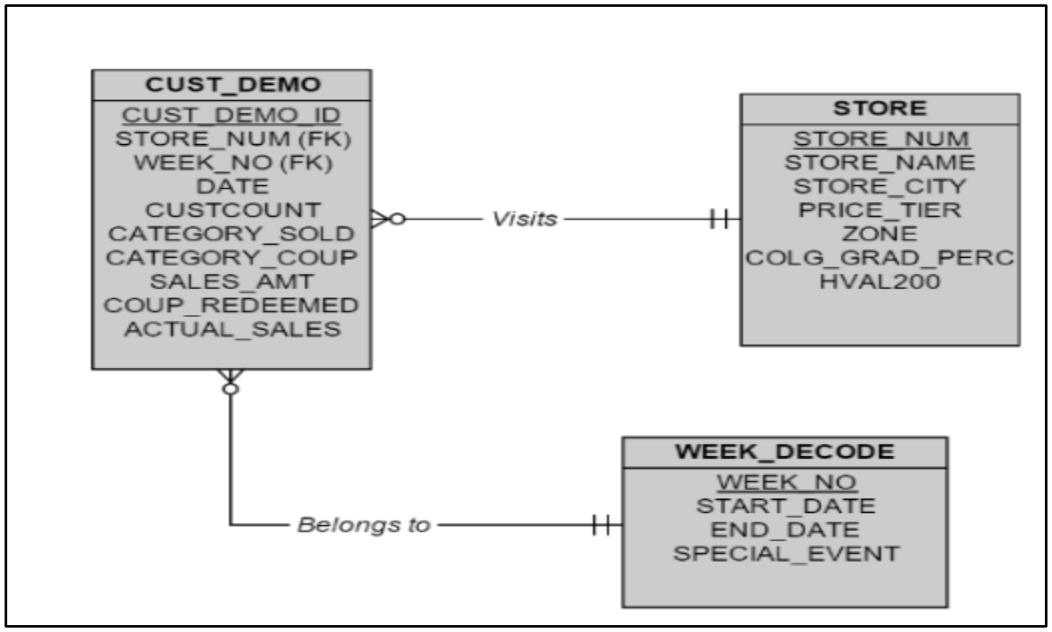


Justification:

The data provides us with sales figure for Camera along with the timestamp i.e. sales of a camera on a particular day. We used this data to plot the sales of a camera on a yearly basis from 1990 to 1996 to showcase the trend of sales. This data analysis would help the business to determine whether a particular product needs to be in store or needs to be removed from the stores. Additionally, the trend would also help manufacturers gauge changes required in the product according to changing times. In our case, Camera sales are increasing as the year proceeds from 1990 to 1996. However, from 1996 to 1997, there is a drastic drop in the sales of a camera. This change could further be analyzed by the businesses and the manufacturers so as to improve the sales in the subsequent years. A similar approach could be used for all the products and data could be analyzed to help the business run more efficiently.

8) During which month were the Meat sales the highest and lowest during the last 3 years?

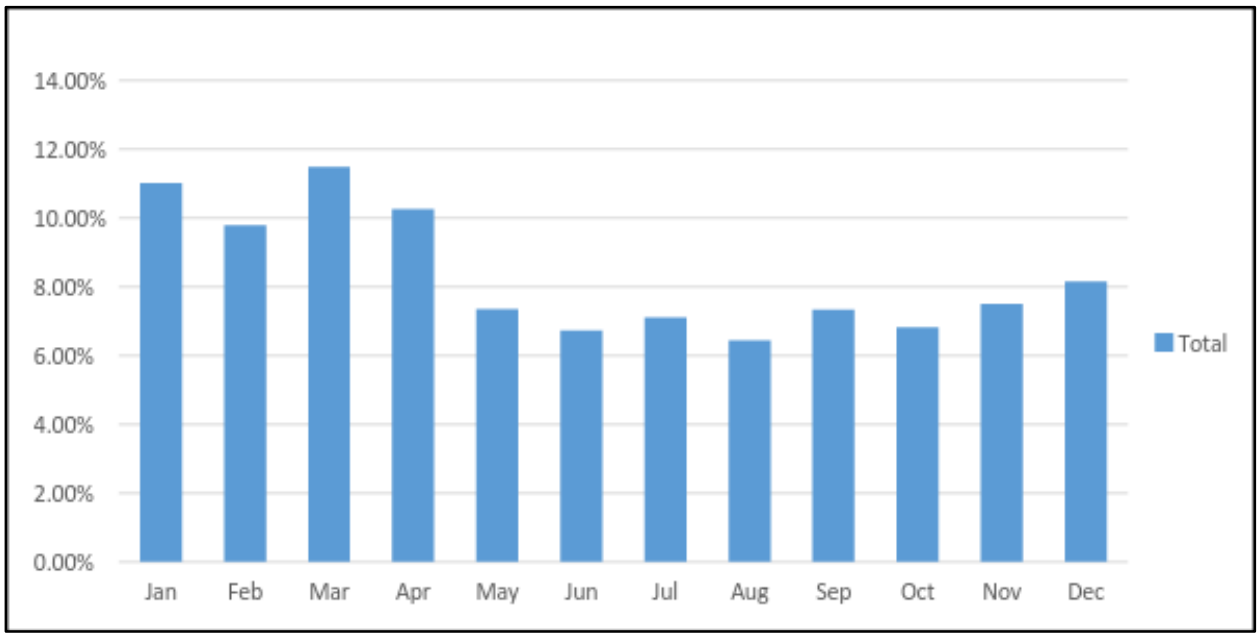
Entity-Relationship Diagram:



Snapshot of Data:

Row Labels	Sum of MEAT
Jan	11.01%
Feb	9.79%
Mar	11.48%
Apr	10.26%
May	7.36%
Jun	6.73%
Jul	7.11%
Aug	6.45%
Sep	7.33%
Oct	6.82%
Nov	7.51%
Dec	8.15%
Grand Total	**100.00%**

Pivot Chart:

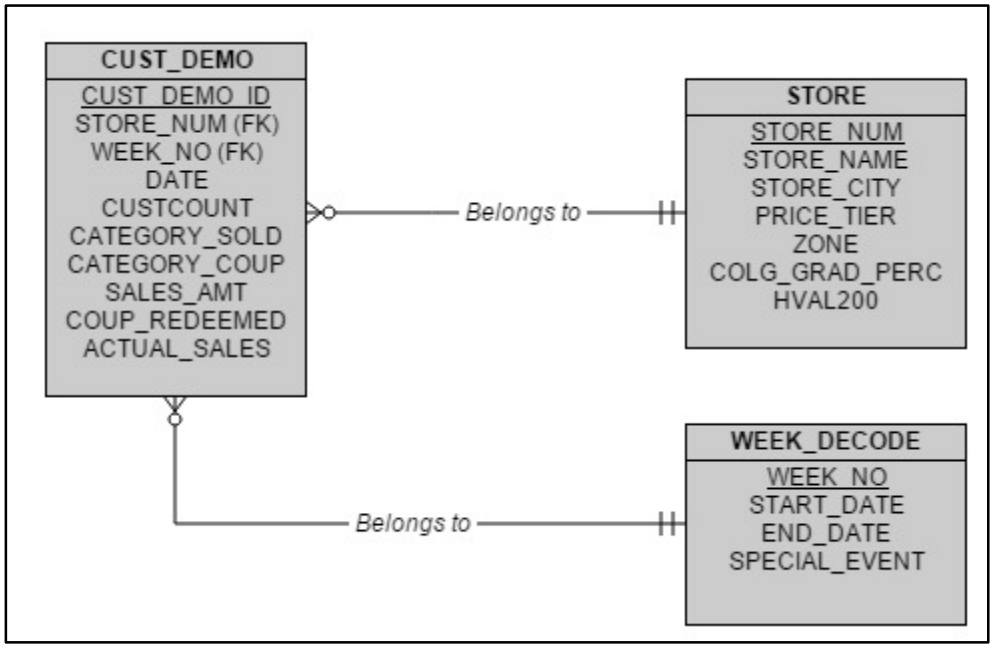


Justification:

For any retail business, it's important to know the yearly trend of sales for a particular product. The store could use this data in a lot of ways. The month with the highest sales could be used to bring a positive price change, improve storage capacity for increased demands and provide promotion for related products. Additionally, the data would provide the month with the lowest sales or the period where the sales are slow. The business could use effective promotion methods to boost the sales of the product such as simple price reduction or bonus buy. We have taken "Meat" as our product and observed the sales trend for meat during last three years. Our sample data shows that meat sales are up during the first four months with March being the month with highest sales. The sales tend to decrease from May to September with August being the month with lowest sales. This trend can be used by the business to improve the sale of meat by using effective sales and marketing strategies

9) **How does "Households with Value over $200,000" greater than the average value compares with those less than the average value in terms of meat sales.**

Entity-Relationship Diagram:



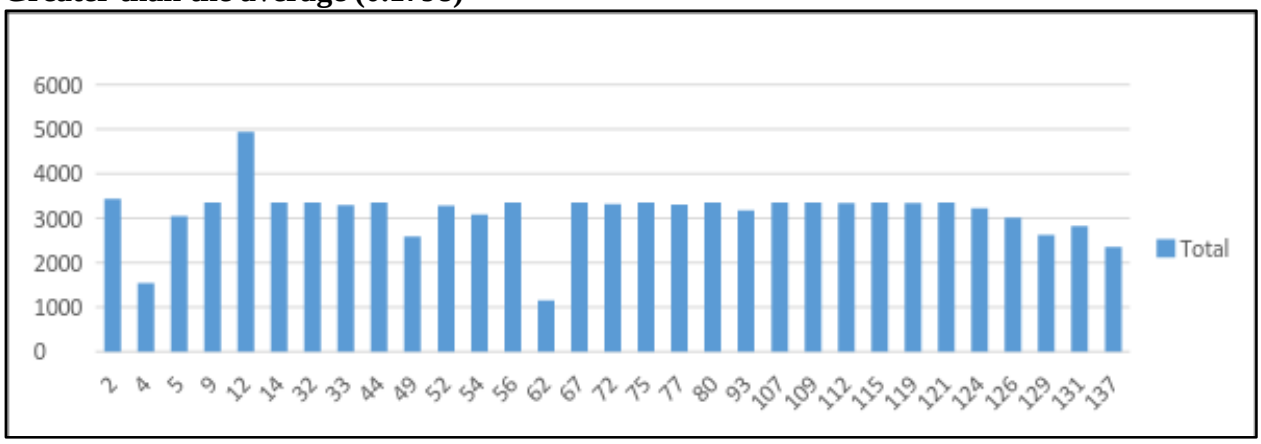
Snapshot of Data:

Row Labels	Sum of Meat
2	3438
4	1537
5	3049
9	3355
12	4941
14	3352
32	3354
33	3293
44	3350
49	2589
52	3276
54	3087
56	3350
62	1157
67	3348
72	3319
75	3351
77	3303
80	3355
93	3181

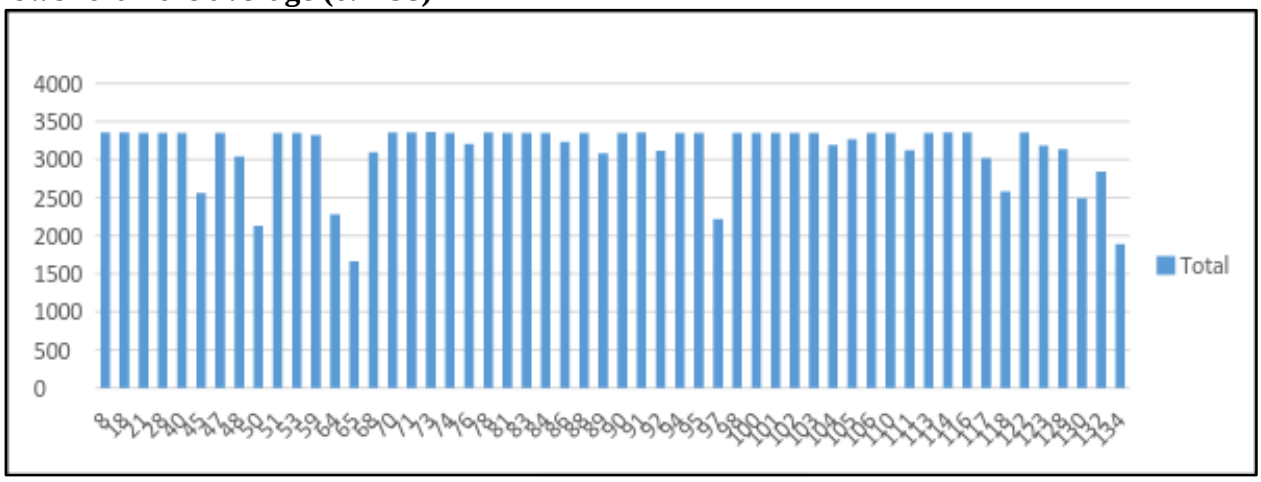
Row Labels	Sum of Meat
8	3354
18	3354
21	3349
28	3348
40	3351
45	2560
47	3350
48	3042
50	2128
51	3346
53	3348
59	3324
64	2280
65	1661
68	3093
70	3353
71	3355
73	3360
74	3350
76	3206
78	3354
81	3346

Pivot Chart:

Meat sales for "%household with value greater than $200,000" having percentage Greater than the average (0.1798)



Meat sales for "%household with value greater than $200,000" having percentage lower than the average (0.1798)

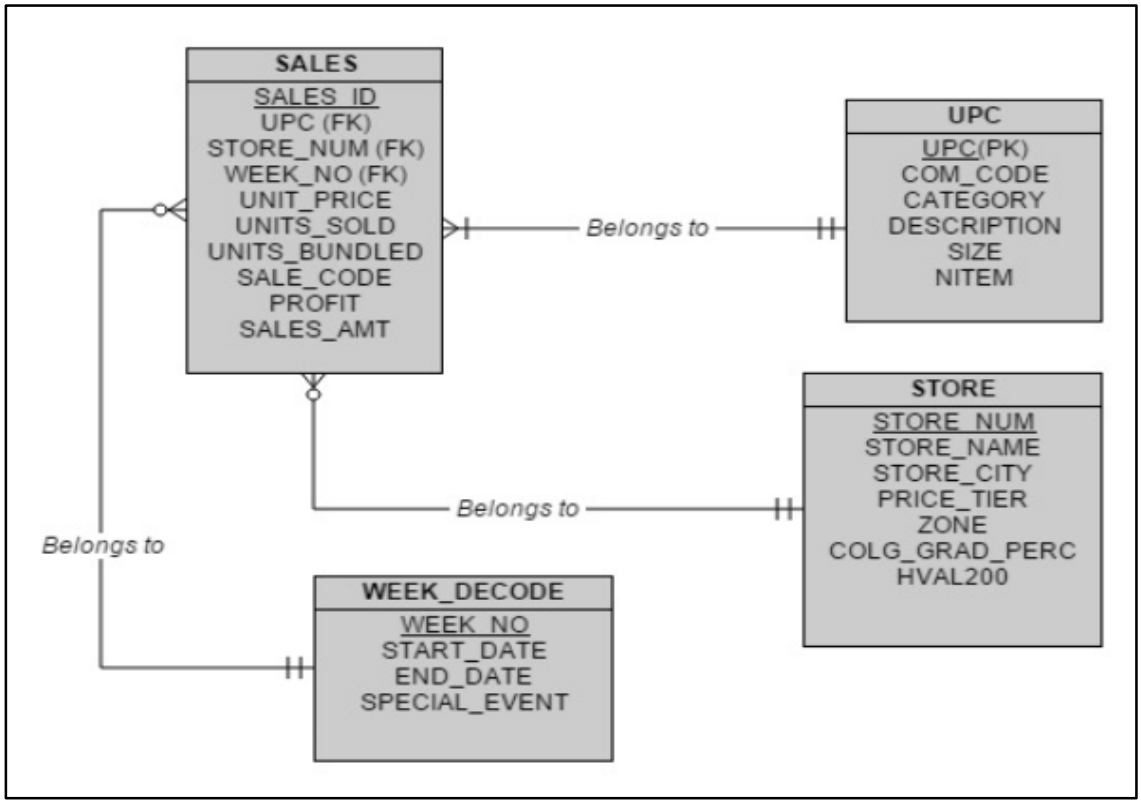


Justification:

The data provides us with demographics data which provides footfall percentage of a particular category of customers in a particular store based on the metrics generated from the census of the area. Our idea here was to compare data for the sales of meat in according to the criteria of "%household with the value greater than $200,000". We first found the average of the percentage of customers in the category and then plotted the graph for the stores above the average value and for the stores below the average value as showed in the graphs above. We were able to observe that the stores which were above the average value reported more sales of meat than those which were lower than the average. This helps us to identify that in areas with percentage household having the value greater than $200,000; we have greater consumption of meat. This data analysis could help DFF to leverage meat sales in these stores. This method could also be used with other combinations of demographics and product to help a business run efficiently.

10) How many Soaps and Frozen Entrée were drop-shipped vs warehoused? (0 or 1 at the last digit of the item code) according to each individual items.

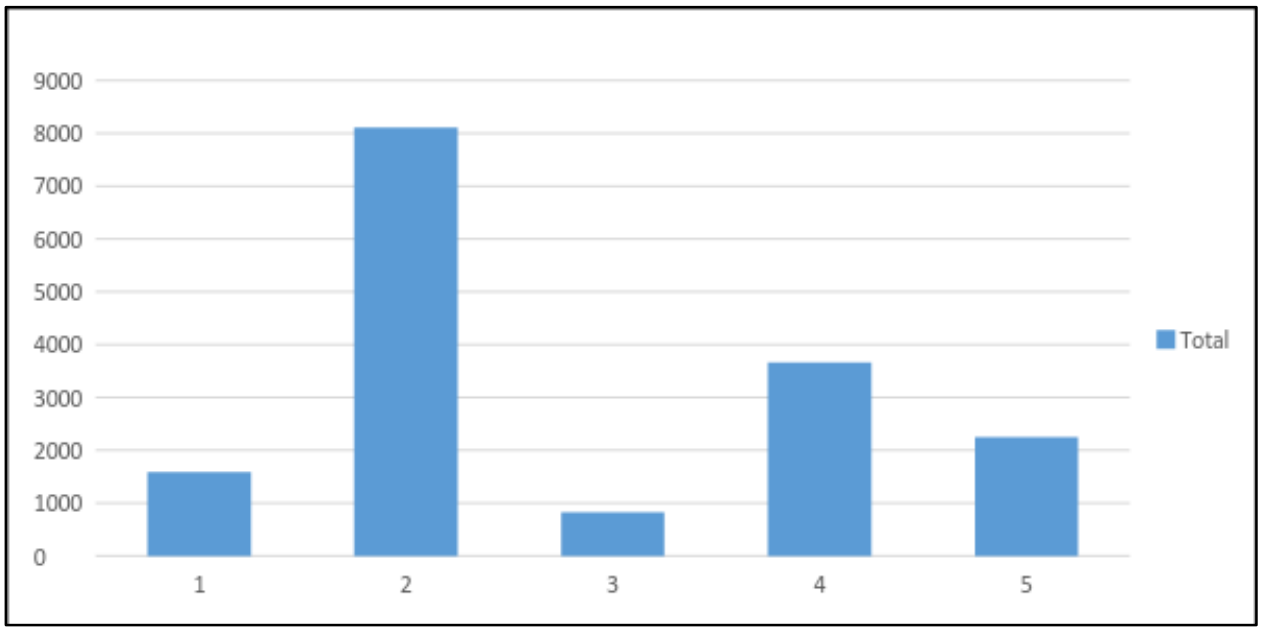
Entity-Relationship Diagram:



Snapshot of Data:

Row Labels	Average of CASE
105	**11.4204947**
0	11.67647059
1	11.371669
658	**24.95555556**
0	37.59090909
1	23.19620253
659	**14.31847134**
1	14.31847134
Grand Total	**13.85834739**

Pivot Chart:



Justification:

The data provides us with details regarding the products that were drop-shipped vs. those that were warehoused. We decided to analyze the data for three products with different characteristics, for example, perishable products vs non-perishable products. In our case, we took a perishable item to be Frozen Entrée and a non-perishable item to be Soap. As perishable products i.e. Frozen Entrée, have shorter expiry dates they need to be shipped directly to the store. Using the data we try to verify the same concept. If there is a change in the trend wherein the perishable items are stored and then transferred to the store, it would mean the loss of products due to improper storage. Business would not want to suffer loss due to such cases and hence, this data analysis would help them store the data in a better way. The analysis of sample data shows us that most of the products are warehoused by Dominick's before being sent to store irrespective of the product category. This analyzes would help business to store the products in a better way.

C. Independent Data Marts design using Kimball's approach

1. Data Mart and Dimension Matrix

We have used Dimensional Modeling technique to create STAR schema for designing the data warehouse for Dominic's Finer Foods retail store chain. Below section of this report, provides details for the various Business dimensions and fact tables, created to answer the business questions pertaining to Dominic's Finer Foods. A thorough explanation of how the data mart design corresponds to the business questions.

1.1. Dimensional Modeling

1.1.1. Dimension Tables

A total of four dimension tables have been created viz. Product dimension, Store dimension, Category dimension and Time dimension. Following are in detail descriptions of each dimension table –

- *DimProduct*

The product dimension table is used to store the details of the various products available in the DFF retail store. It contains reference information about the facts or measures seen in the Product Sales fact table. The Product Sales fact table has been explained in detail in the Fact Table section of this report

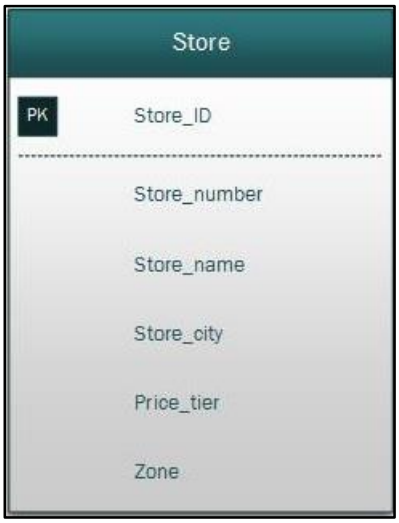
DimProduct

Following is the description of the attributes of the DimProduct dimension table –

Product_ID – A unique identifier of the product dimension table which has been generated as a surrogate key.
Product_name – Name of the product available in the DFF retail store.
UPC_number – The last five digit of the UPC number identify the product, the remaining digits identify the manufacturer.

- *DimStore*

The store dimension table is used to store the details of the various stores in the DFF retail store chain. It contains reference information about the facts or measures seen in the Product Sales fact table and Category Sales fact table. The Category Sales fact table has been explained in detail in the Fact Table section of this report



DimStore

Following is the description of the attributes of the DimStore dimension table –

Store_ID – A unique identifier of the store dimension table which has been generated as a surrogate key.
Store_number – The assigned number to a particular store.
Store_name – Name of a particular store in the DFF retail store chain.
Store_city - The City where the store is located.
Price_tier - The price tier to which a particular zone belongs where the store is located.
Zone – The identifier of the zone to which a store belongs.

- *DimCategory*

The category dimension table is used to store the details of the various categories defined in the DFF retail store chain. It contains reference information about the facts or measures seen in the Category Sales fact table.

DimCategory

Following is the description of the attributes of the DimCategory dimension table –

Category_ID – A unique identifier of the category dimension table which has been generated as a surrogate key
Category_name – Name of a category defined in the DFF retail store chain.

- *DimTime*

The time dimension table is used to stores the time attributes. It contains reference information about the facts or measures seen in the Category Sales fact table.

DimTime

Following is the description of the attributes of the DimTime dimension table –

Time_ID - A unique identifier of the time dimension table which has been generated as a surrogate key
Year – The identifier for the year when sales have occurred
Month - The identifier for the month when sales have occurred
Week_number - The identifier for the week in the year when sales have occurred
Special_event - The identifier to indicate the occurrence of a special event in the year

1.1.2. Fact Tables

- *Product Sales Fact Table*

This table contains the quantitative and aggregated information for the sales of the various products available in the Dominic's Finer Foods retail store chain.

FactProductSales

Fact Table Keys:

Store_ID – A unique identifier for stores which has been generated as a surrogate key and helps in dividing data by store
Product_ID - A unique identifier for products which has been generated as a surrogate key

Fact Table Measures:

Unit_price – Price of a bundle of the product
Quantity – Size of the bundle
Number_of_units_sold – Number of units of a product sold
Profit_per_dollar – Profit made by DFF on the dollar for each sold product
Sale_code – This attribute indicates whether the product was sold on promotion
Product_sales – This is a derived attribute. It indicates the dollar value of the product sales across various dimensions.

*Product_sales = (Unit_price * Number_of_units_sold) / Quantity*

- *Category Sales Fact Table*

  This table contains the quantitative and aggregated information for the sales pertaining to various categories in the Dominic's Finer Foods retail store chain.

FactCategorySales

Fact Table Keys:

Time_ID - A unique identifier for time which helps in dividing data by time dimensions like year, month, week number
Category_ID - A unique identifier for category which has been generated as a surrogate key
Store_ID - A unique identifier for stores which has been generated as a surrogate key and helps in dividing data by store

Fact Table Measures:
Category_sales – This attribute indicates the dollar value of the category sales across various dimensions.

1.1.3. Define Dimension Matrix for data marts

Dimension / Data Mart	DimProduct	DimStore	DimCategory	DimTime
Product Sales	☑	☑		
Category Sales		☑	☑	☑

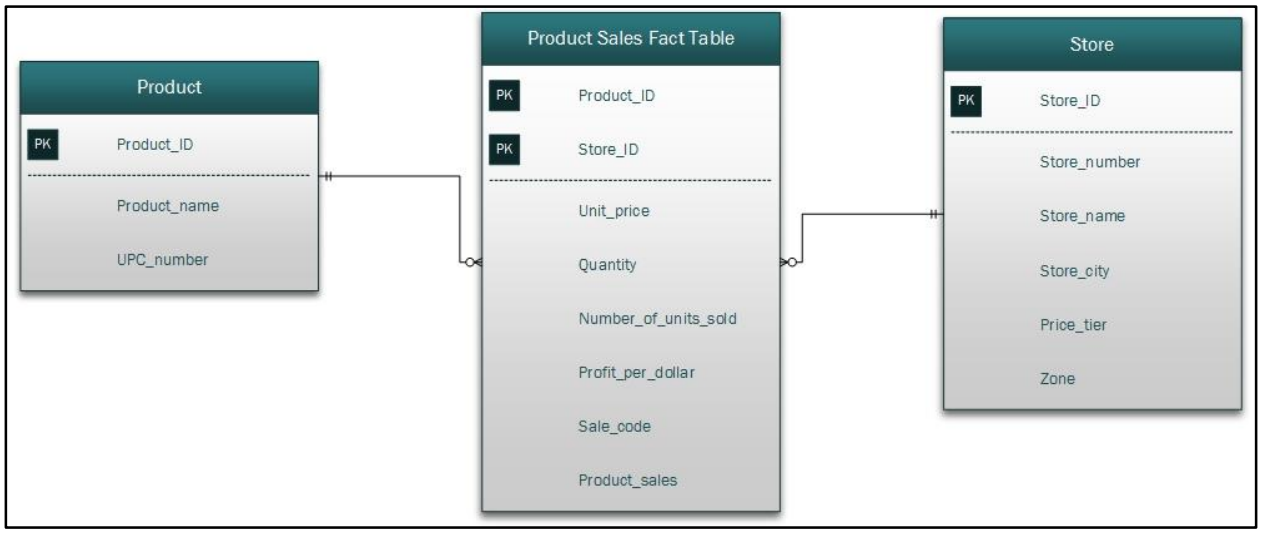
1.1.4. Star Schema

Using the above fact tables and dimension tables, we have integrated them to create two Star Schemas. Those are Product Sales and Category Sales. The data marts created using these tables are sufficient to answer business questions which have been mentioned in Section 3.2.

Following are the data marts created –

Star Schema for Product Sales data mart

Below data mart is the Product Sales data mart. It comprises of Product Sales fact table i.e. FactProductSales and two dimension tables viz. DimProduct and DimStore. The below Star Schema for Product Sales data mart answers business questions 3 and 4 (Refer Section 3.2. for detailed questions and justification)

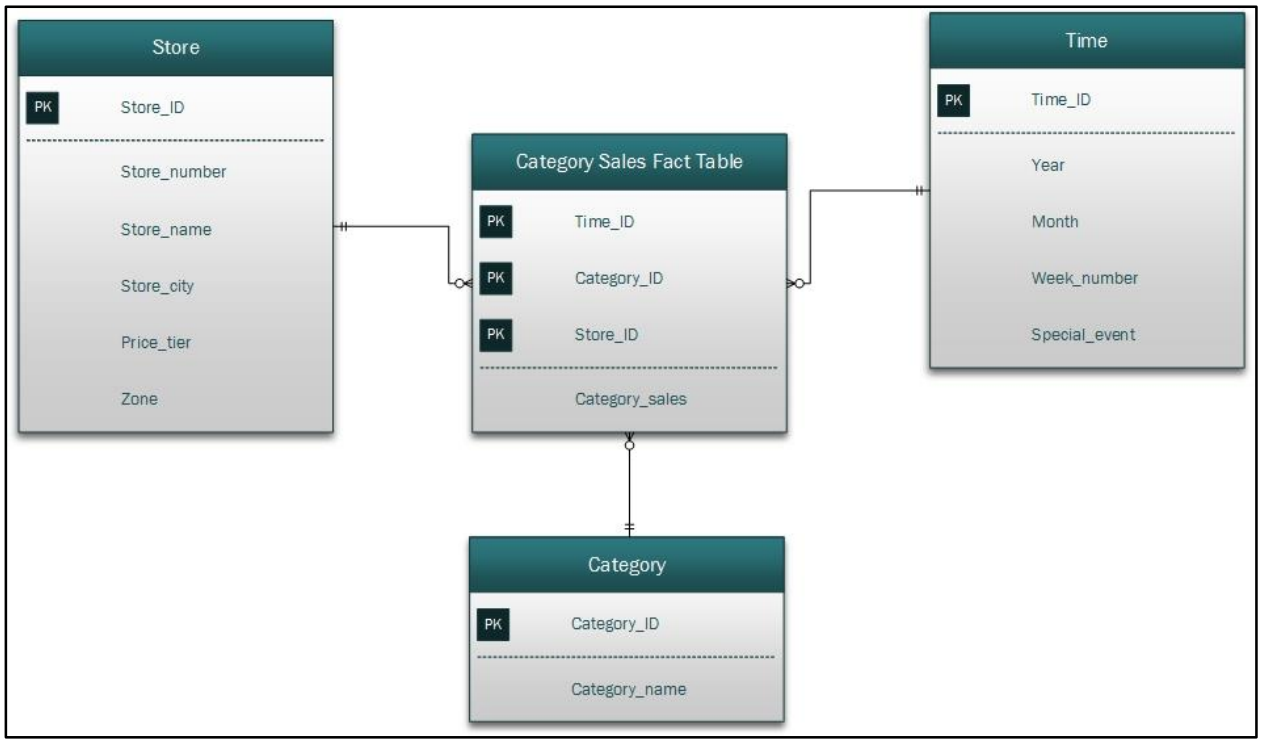


Star Schema for Product Sales data mart

Star Schema for Category Sales data mart

Below data mart is the Product Sales data mart. It comprises of Product Sales fact table i.e. FactCategorySales and three dimension tables viz. DimTime, DimCategory and DimStore.

The below Star Schema for Category Sales data mart answers business questions 1, 2 and 5 (Refer Section 3.2. for detailed questions and justification)



Star Schema for Category Sales data mart

2. Design Feedback

2.1. Mapping Table

2.1.1. Mapping tables for Product Sales Data Mart

DW Dimension Table	DW Dimension Attribute	Source Table	Source Table Attribute	Mapping Functions
Product		UPC table		
	Product_ID (Row Number)			
	Product_name		Description	
	UPC_number		UPC	

DW Dimension Table	DW Dimension Attribute	Source Table	Source Table Attribute	Mapping Functions
Store		Dominick's Store and Store Specific Demographics		
	Store_ID (Row Number)			
	Store_number		Store	
	Store_name		Name	
	Store_city		City	
	Price_tier		Price Tier	

	Zone		Zone	

DW Dimension Table	DW Dimension Attribute	Source Table	Source Table Attribute	Mapping Functions
Product Sales Fact Table		Movement table		
	Product_ID (Row Number)			Primary key of the Product dimension
	Store_ID			Primary key of the Store dimension
	Unit_price		Price	
	Quantity		qty	
	Number_of_units_sold		move	
	Profit_per_dollar		profit	
	Sale_code		sale	
	Product_sales			Derived from the formula Sales = (price*move)/qty

2.1.2. Mapping tables for Category Sales Data Mart

DW Dimension Table	DW Dimension Attribute	Source Table	Source Table Attribute	Mapping Functions
Store		Dominick's Store and Store Specific Demographics		
	Store_ID (Row Number)			
	Store_number		Store	
	Store_name		Name	
	Store_city		City	
	Price_tier		Price Tier	
	Zone		Zone	

DW Dimension Table	DW Dimension Attribute	Source Table	Source Table Attribute	Mapping Functions
Time		Week Decode Table		
	Time_ID (Row Number)			
	Year		Start End	To obtain Year, we can use start (date) and end (date) from Week Decode Table.
	Month			To calculate Year, we can use start (date) and end (date) from Week Decode Table.
	Week_number		Week #	
	Special Events		Special Events	

DW Dimension Table	DW Dimension Attribute	Source Table	Source Table Attribute	Mapping Functions
Category		CCount		
	Category_ID (Row Number)			
	Category_name			Various category attributes like Fish, grocery, beer, etc from ccount.

DW Dimension Table	DW Dimension Attribute	Source Table	Source Table Attribute	Mapping Functions
Category Sales Fact Table		Ccount		
	Time_ID			Primary key of the Time dimension
	Category_ID			Primary key of the Category dimension
	Store_ID			Primary key of the Store dimension
	Category_Sales			Sales of categories like Fish, grocery, beer, etc from ccount.

2.2. Justification of Business questions corresponding to data marts

1. **What is the trend of Beer Sales during Thanksgiving's week for the entire duration?**

   This business question is answered by Category Sales data mart. The Week_number and Special_event attributes are used from DimTime dimension to divide the data by Thanksgiving period and the measure Category_sales from the FactCategorySales fact table is used to find the trend of Beer Sales across the years (DimTime dimension)

2. **How are the average price and sales of a particular product changing according to different zones (Fish and Fish Coupon)**

   This business question is answered by Category Sales data mart. The Zone and Store number attributes are used from DimStore dimension to divide the data by Zones and then by Store and the measure Category_sales from the FactCategorySales fact table is used to find the average sales of Fish and Fish Coupon Sales across the zones.

3. **Compare the effect of Bonus buy and Price Reduction in Analgesics in different zones**

   This business question is answered by Product Sales data mart. The Zone attribute is used from DimStore dimension to divide the data by Zones. Then Sale_code attribute and the derived measure Product_sales from the FactProductSales fact table is used to compare the total zonal sales of Analgesics for Bonus Buy and Price Reduction promotions.

4. **Plot the average profit margin for cigarettes across all the stores. Determine the average of profit for the sales of cigarettes and the stores which are below the average.**

   This business question is answered by Product Sales data mart. The Store_number attribute is used from DimStore dimension to divide the data by stores. Then measure Profit_per_dollar from the FactProductSales fact table is used to plot and compare sale of Cigarettes across the stores.

5. **What is the trend of Camera sales from the year 1990 to 1996?**

   This business question is answered by Category Sales data mart. The Year attribute is used from DimTime dimension to divide the data by year. Then the measure Category_sales from the FactCategorySales fact table is used to find the trend of Camera sales across the years.

D. Data Integration

1. Data Quality issues in the DFF data sets

Group	Quality	Issues Considered	Data Quality Problem in Dominic Finer Foods
Relation to Data	Referential Integrity	Do records exist where expected? Do they contain unnecessary or inactive data? Are reference files/tables complete?	The data given for dominic finer foods was not entirely of good quality. Records did not exist where they were expected. There were a lot of null values in all the excel sheet. Also, referential integrity was difficult to figure out as there were values that were null even in keys which could be assumed to be primary key. e.g:- Store number was repeated and had inconsistent values.
	Cardinality	Is the structure of relationships amongst entities and attributes maintained consistently?	The structure of relationships amongst entities and attributes is highly inconsistent.
Structure of fields	Format	Do values follow consistent formatting standards?	Values did not follow consistent formatting standards. E.g.: Date attribute had a format as XXXXXX instead of XX/XX/XX
	Standard	Are data elements consistently defined and understood?	Though data elements were explained well in DFF catalog, looking at the data it was difficult to understand what it represented. E.g.: Coupons had decimal values, so shall a user assume it is coupon's dollar value or it is dirty data.
	Consistent	Do values represent same meaning across systems and files?	Data was explained well in the catalog and hence the data had same meanings across the systems. However, amongst files there was an overlap with the data values which created misrepresentation of data.
Content within Data Values	Complete	Is all necessary data present?	After formulating the business question, it was found that a lot of data was not present. For eg: Sales had blank values in UPCIG file in the time span for which analysis was done.
	Accurate	Does the data accurately represent reality or verifiable resource?	Data was for Dominic Finer Food and hence the data represents reality. However, accuracy was questionable as data had null values where value was expected, negative values for Sales data and decimal value for coupon.
	Valid	Do data values fall within acceptable ranges defined by the business?	Data values did not fall within acceptable ranges defined by the business. Store number have out of range values and hence have invalid values.
	Fit for purpose	Is the information valuable to the business	The information contains a lot of sales data and customer data and hence is valuable to the business.

The above table discusses the Data Quality of Dominic Finer Foods. For working on the ETL implementation, we need to have a good understanding of data. Dummy values, Absence of data values, unofficial use of fields, Cryptic values, Violation of business rules, Non-unique identifiers, Inconsistent values, and incorrect values are the problems faced by users during data analysis. Hence, problems are required to be classified and then appropriate steps in ETL will take care that data generated for BI purpose is accurate.

2. ETL Plan

"ETL is short for extract, transform, load, three database functions that are combined into one tool to pull data out of one database and place it into another database.

- Extract is the process of reading data from a database.
- Transform is the process of converting the extracted data from its previous form into the form it needs to be in so that it can be placed into another database. Transformation occurs by using rules or lookup tables or by combining the data with other data.
- Load is the process of writing the data into the target database.

ETL is used to migrate data from one database to another, to form data marts and data warehouses and also to convert databases from one format or type to another."[1]

Following are steps involved in the ETL plan for data warehouse implementation-

2.1. Determining all the target data needed in the data warehouse

DATA SOURCE	DATA EXTRACTION	DATA TRANSFORMATION	DATA IN DATA WAREHOUSE
UPC	UPCANA, UPCCIG	UPCANA, UPCCIG	DimProduct
MOVEMENT	DONE-WANA, DONE-CIG	FinalProductSource	FactProductSales
STORE	Store	Store	DimStore
WEEK	Week_decode	FinalWeekDecode	DimTime
CCOUNT	CCOUNT	cleanCCOUNT, Total	DimCategory, FactCategorySales

2.2. Determining all the data sources

The data source for the data warehouse is the Dominick's database provided by the University of Chicago Booth School Of Business. The files in this data source that are required according to our business questions are listed as below:

DATA	SOURCE FILES
UPC	UPC<*product_acroymn*>
MOVEMENT	W<*product_acroymn*>
STORE	STORE (created)
WEEK	WEEK_DECODE (created)
CUSTOMER COUNT	CCOUNT

2.3. Preparing data mappings for data elements from sources in CSV to staging and then data mapping from staging to data warehouse (include all transformations)

SOURCE FILE NAME	SOURCE FILE ATTRIBUTES	STAGING AREA TABLE NAME	STAGING AREA TABLE ATTRIBUTES
Week_decode.csv	week number	Week_decode	Week #
	start date		Start
	end date		End
	special events		Special Events
Store.csv	store number	Store	Store
	city		City
	price tier		Price Tier
	zone		Zone
	zip code		Zip Code
	address		Address
CCount.csv	store number	cleanCCOUNT	Store
	fish		Fish
	fishCoupon		Fishcoup
	camera		Camera
	beer		Beer
	week		Week
	date		
upc<*product_acroymn*>.csv		Created Dimproduct using UPC Analgesics and UPC Cigarettes directly	
	UPC		
	comm_code		
	descrip		
	nitem		
	size		
	case		
w<*product_acroymn*>.csv	store number	finalProductSource	Store
	UPC		UPC_number
	week number		
	move		Number_of_units_sold
	quantity		Quantity
	price		Unit_price
	sale		Sale_code
	profit		Profit_per_dollar

			Classification
			Product_sales
	OK		

STAGING AREA TABLE NAME	STAGING AREA ATTRIBUTE	DATAWAREHOUSE TABLE	DATAWAREHOUSE TABLE ATTRIBUTES	MAPPING FUNCTION
Week_decode		dimTime	Time_ID	surrogate key
	Week #		Week_number	
	Start		Month	MONTH(start_date)
	End		Year	YEAR(start_date)
	Special Events		Special_event	
Store		dimStore	Store_ID	surrogate key
			Store_name	
	Store		Store_number	
	City		Store_city	
	Price Tier		Price_tier	
	Zone		Zone	
	Zip Code			
	Address			
Product_UPC		dimProduct	Product_ID	surrogate key
	UPC		UPC_number	
	Com_Code			
	Descrip		Product_name	
	NItem			
	Size			
	Case			
			Classification	extracted from the file name
Category_Staging_move		dimCategory	Category_ID	
	Extracted from CCOUNT column names		Category_Name	
finalProductSource		factProductSales	Fact_ProductSales_ID	surrogate key
			Product_ID	surrogate key
	Store		Store_ID	surrogate key
	Unit_price		Unit_price	
	Quantity		Quantity	
	Number_of_units_sold		Number_of_units_sold	
	Profit_per_dollar		Profit_per_dollar	

	Sale_code		Sale_code	
	Product_sales		Product_sales	Product_sales = (Unit_price * Number_of_units_ sold) / Quantity
	UPC_number			
	Classification			
Total	Sales	factCategorySales	Category_Sales	Weekly Sales for a particular category
	Week			
	Store			
	Category			
			Store_ID	Surrogate key
			Week_ID	Surrogate key
			Category_ID	Surrogate key

2.4. Establishing comprehensive data extraction rules

Data extraction is the process of retrieving data out of data sources for further data processing. The process of data extraction from the source files is the initial step involved in the ETL process of designing and implementing a data warehouse. Data extraction has been performed with extreme caution to help build an efficient data warehouse.

The extraction of data involves combining all the data sources to a single format i.e tables in Microsoft SQL Server Studio, that can be later used for transformation and loading.

The data from the sources are in the Comma Separated Value (.csv) formats.

The data for Week and Store details was extracted from the data manual for Dominick Finer Foods. Comma Separated Value (.csv) files were created for such tabular data.

The data from the source files has been loaded into the data staging area '601Group1_staging_area'.

2.5. Determining data transformation and cleansing rules

Data transformation and cleaning process follows the Data Extraction process. The input to this process is the extracted data that is present in the staging area. Transformation and cleaning rules have been applied to this data to create a data that is clean and consistent throughout. The clean data can then be loaded into the DW area so as to create the data marts and get the appropriate results.

The general transformation and cleaning rules that are applied to the data are as follows:

i. Removal of NULL values

Null values that were existing in the data and that were created while extracting data into tables will be deleted.

ii. Removal of dirty data

Attributes that are not part of the answers to the business questions will be removed. For instance, attributes other than Store, Beer, Fish, Fish Coupons, Camera and Week in the Customer Count files have been eliminated. Also, the blank records, rows with just a '.' (Dot) and other non-meaningful values such as unnecessary negatives will be removed. The records where the value for the sales of various departments are '.' will be removed. Negative values in Fish Coupon will be updated to 0.

iii. Data conversion

The data was extracted from the source files are stored as attributes of type varchar (that is, as string) in the staging area. These have been converted into their respective data types such as int, float and date. Date type field was stored as varchar. The date has been split using functionalities in SSIS and then stored as Year and Month attributes.

iv. Creation of surrogate keys

Surrogate keys have been created for all the dimension tables and fact tables before loading the data in the data warehouse.

v. Derived attributes

Derived attributes exist in two dimensions, described as below:

- Year and month in the Time dimension, obtained from the functions YEAR (date) and MONTH (date) respectively.
- The 'Product_sales' measure in the fact table 'factCategorySales' is obtained from the function
  'Product_sales = (Unit_price * Number_of_units_sold) / Quantity'.
- The 'Category_sales' measure in the fact table 'factCategorySales' is obtained by aggregating sales for a particular week by grouping the sales based on week number. Hence, the measure stores category sales for a particular week.

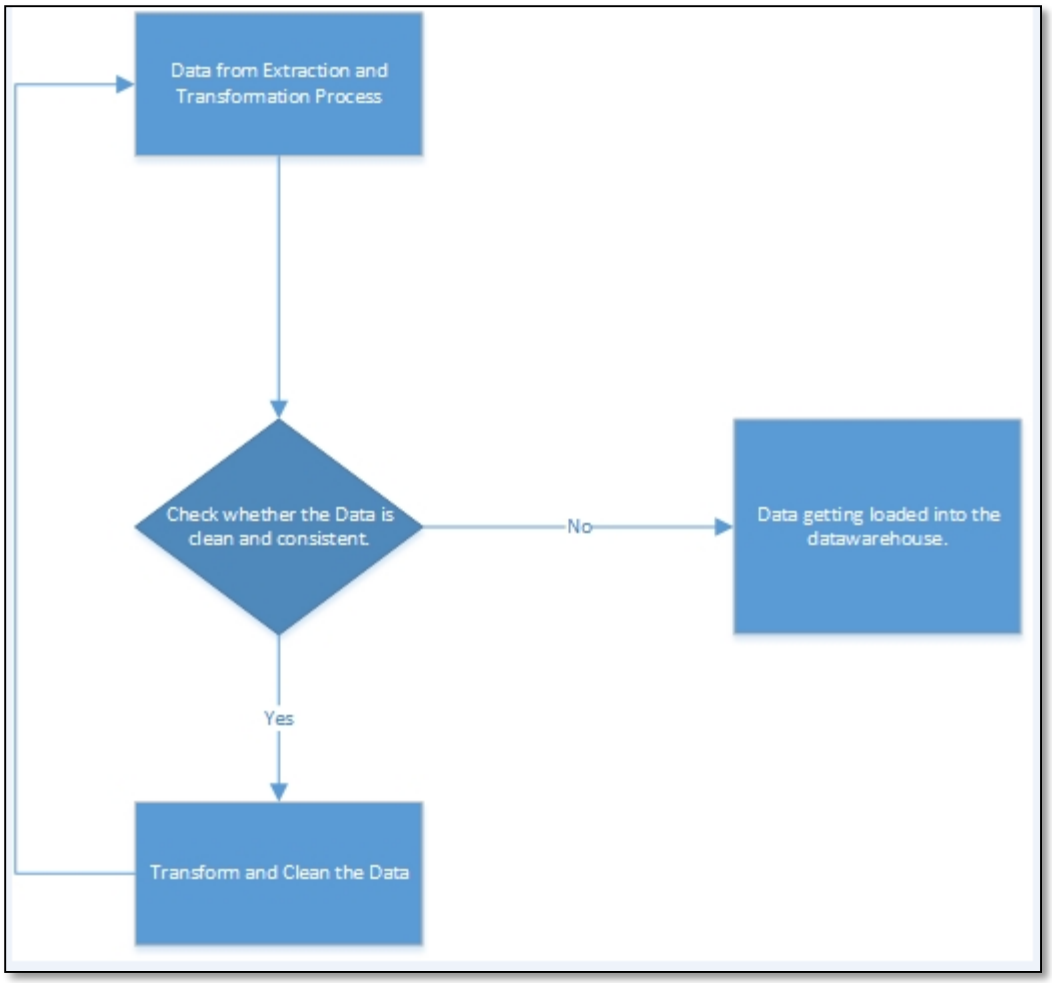
SSIS Functions

The SSIS functions that aided during the transformation and cleaning process include:

1) **LOOKUP**: It joins additional columns to the data flow by looking up values in a table.
2) **DATA CONVERSION**: It converts data from one data type to another.
3) **DERIVED COLUMNS**: It creates new column values by applying expressions to input columns.
4) **AGGREGATE**: It aggregates data with functions such as count and sum.
5) **UNPIVOT**: It makes an un-normalized dataset into a more normalized version by expanding values from multiple columns in a single record into multiple records with the same values in a single column.

vi. Procedure for data loading

The basic idea for the procedure of data loading used for implementing the data warehouse is described using the flowchart below:



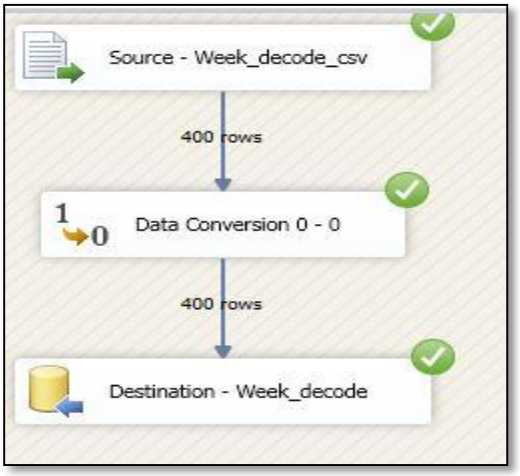
3. ETL implementation

The ETL process for the data warehouse was implemented using the SQL Server Management Studio and SSIS (Microsoft Visual studio – SQL Server Data Tools). The steps involved in this process are discussed as below, along with the SSIS tool screen shots explaining the entire process step-by-step:

3.1. Extraction and Transformation of Source data into Dimensions and Fact Tables

3.1.1. Time Dimension Creation

Extracting Week_decode source file into the staging area



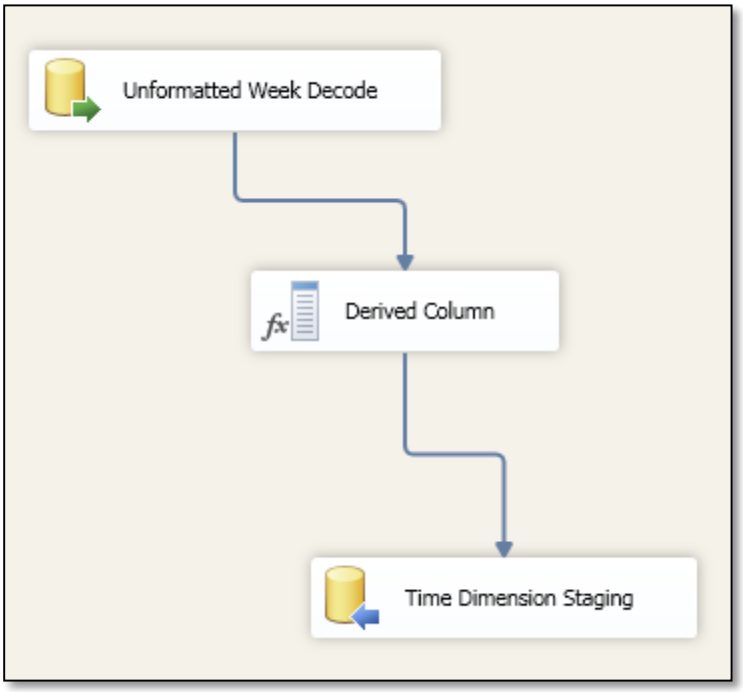
Snapshot of Week Decode table in the Staging Area

```
SELECT * FROM Week_decode
```

Results | Messages

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

Transforming Week_decode to Time Dimension



Snapshot of Time Dimension table

```
SELECT * FROM dimTime
```

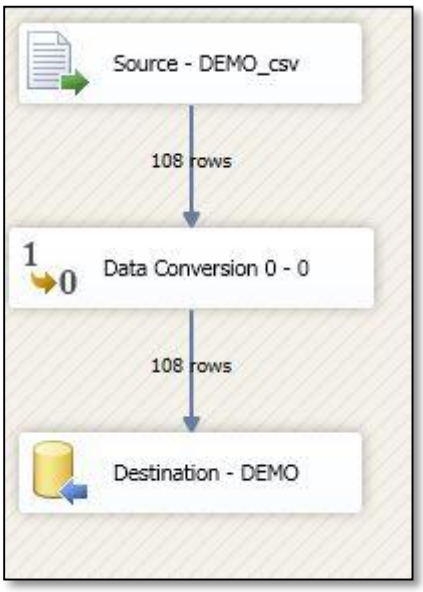
100 %

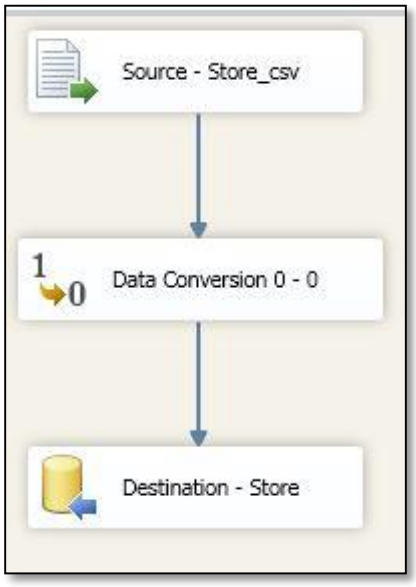
Results | Messages

	Time_ID	Year	Month	Week_number	Special_event
1	1	1989	9	1	
2	2	1989	9	2	
3	3	1989	9	3	
4	4	1989	10	4	
5	5	1989	10	5	
6	6	1989	10	6	
7	7	1989	10	7	Halloween
8	8	1989	11	8	
9	9	1989	11	9	
10	10	1989	11	10	
11	11	1989	11	11	Thanksgiving
12	12	1989	11	12	
13	13	1989	12	13	
14	14	1989	12	14	

3.1.2. Store Dimension Creation

Extracting the Store and Demo source file into the staging area





Snapshot of Demo Table

```
SELECT * FROM DEMO
```

100 %

Results | Messages

	"MMID"	"NAME"	"CITY"	"ZIP"	"LAT"	"
1	.	""	""	.	.	.
2	16892	"DOMINICKS 2"	"RIVER FOREST"	60305	419081	8
3	16893	"DOMINICKS 4"	"PARK RIDGE"	60068	420392	8
4	16894	"DOMINICKS 5"	"PALATINE"	60067	421203	8
5	16895	"DOMINICKS 8"	"OAK LAWN"	60453	417331	8
6	16896	"DOMINICKS 9"	"MORTON GROVE"	60053	420411	8
7	16898	"DOMINICKS 12"	"CHICAGO"	60660	419928	8
8	16899	"DOMINICKS 14"	"GLENVIEW"	60025	420733	8
9	16901	"DOMINICKS 18"	"RIVER GROVE"	60171	419364	8

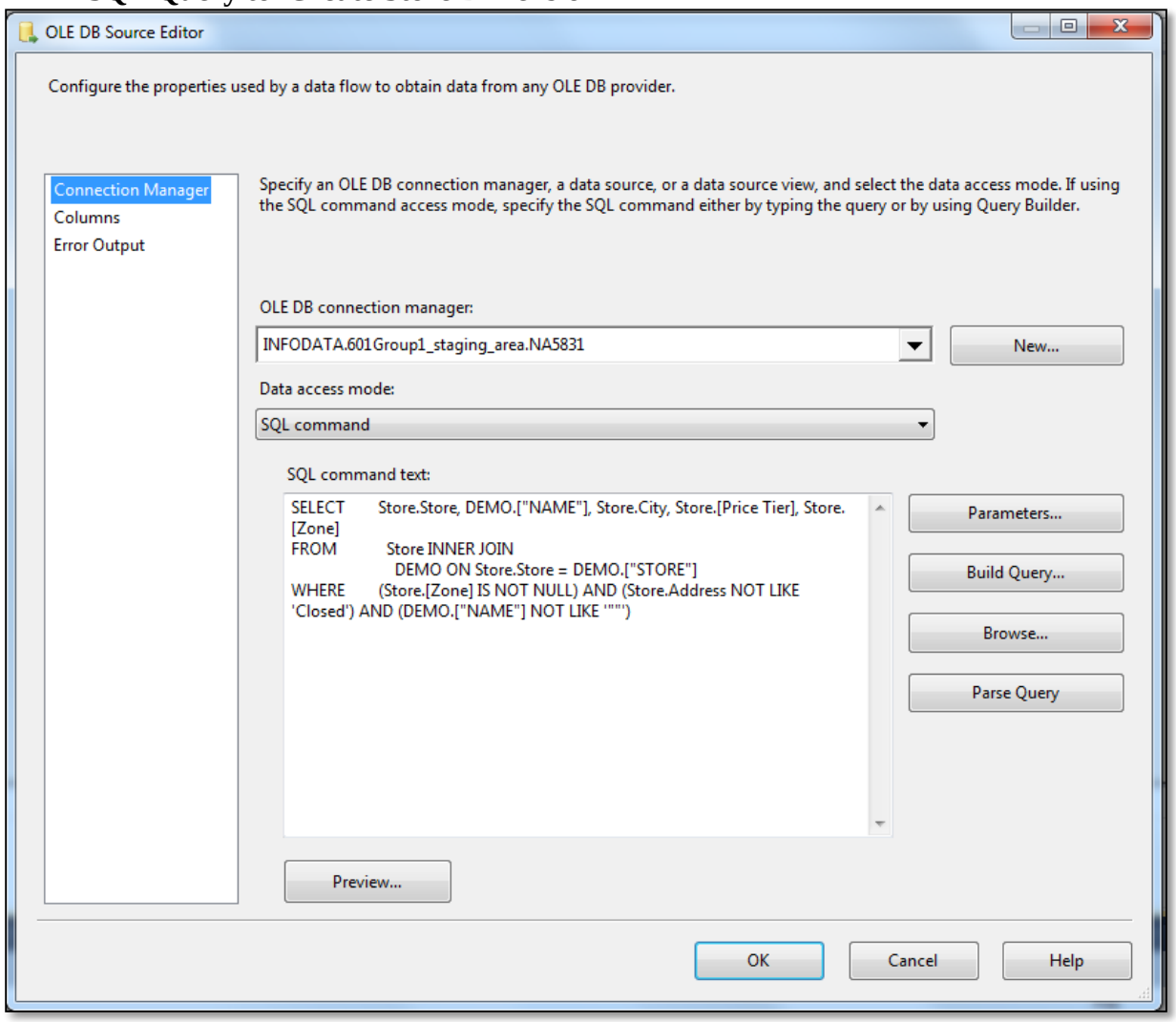
Snapshot of Store Table

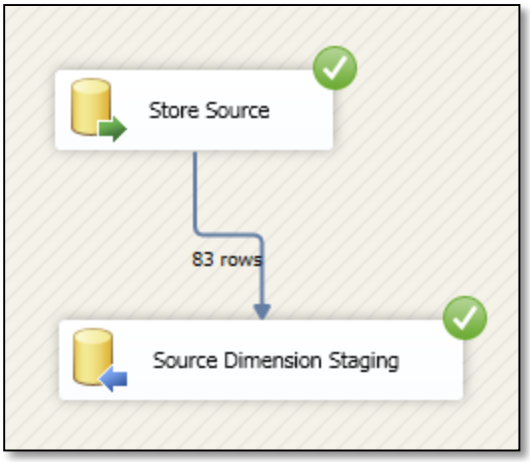
Results | Messages

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

Transformation from Store Table to Store Dimension

SQL Query to Create Store Dimension





Snapshot of Store Dimension

```
SELECT * FROM dimStore
```

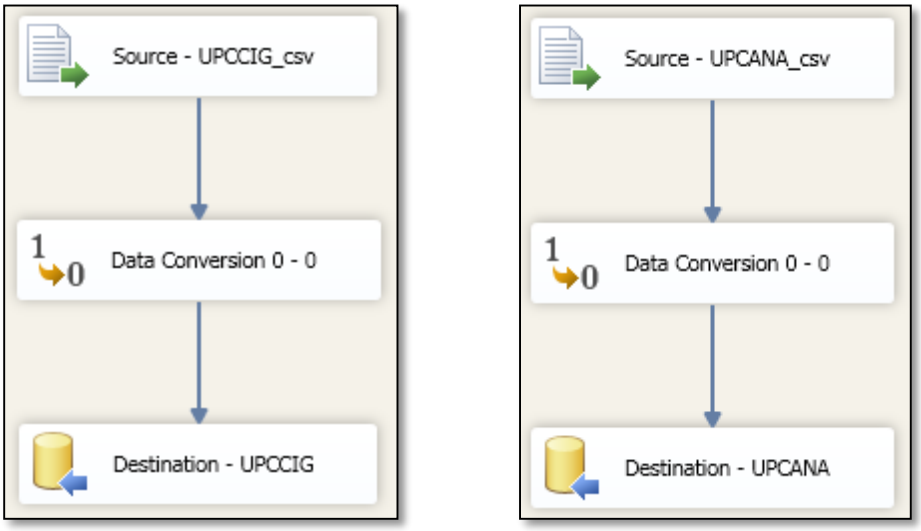
100 %

Results | Messages

	Store_ID	Store_number	Store_name	Store_city	Price_tier	Zone
1	1	2	DOMINICKS 2	River Forest	High	1
2	2	5	DOMINICKS 5	Palatine	Medium	2
3	3	8	DOMINICKS 8	Oak Lawn	Low	5
4	4	9	DOMINICKS 9	Morton Grove	Medium	2
5	5	12	DOMINICKS 12	Chicago	High	7
6	6	14	DOMINICKS 14	Glenview	High	1

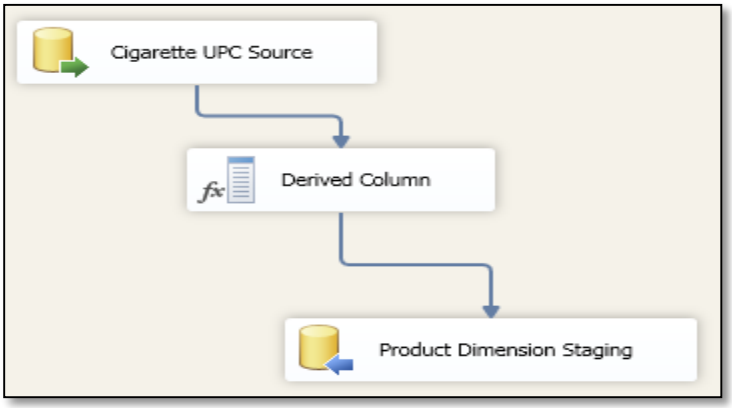
3.1.3. Product Dimension Creation

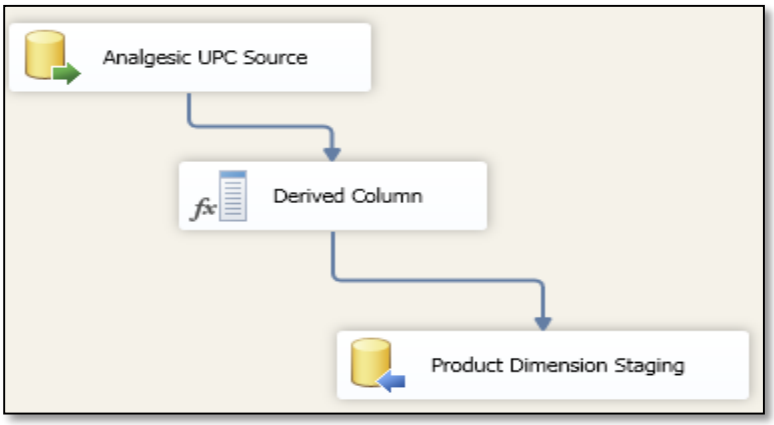
Extracting UPC source file into the staging area



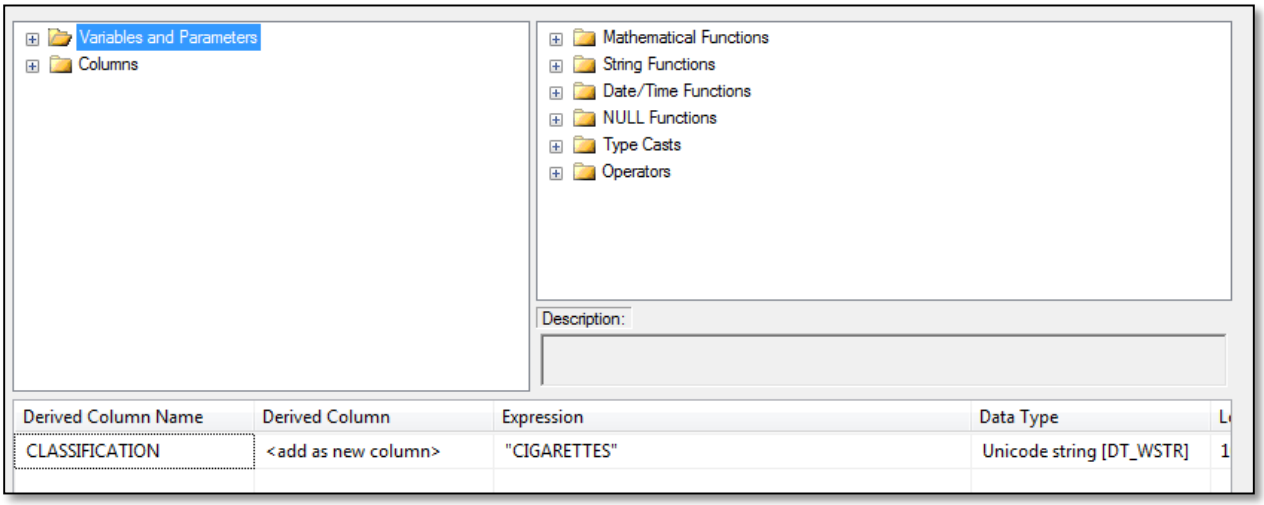
Using only Cigarette and Analgesic since those are the target files

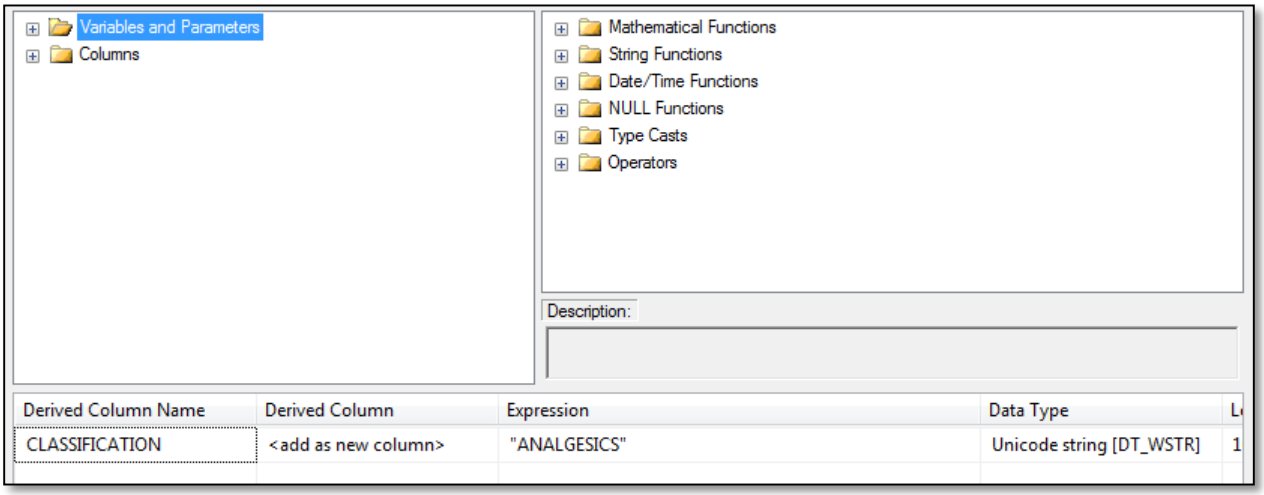
Transforming Source Data into Product Dimension



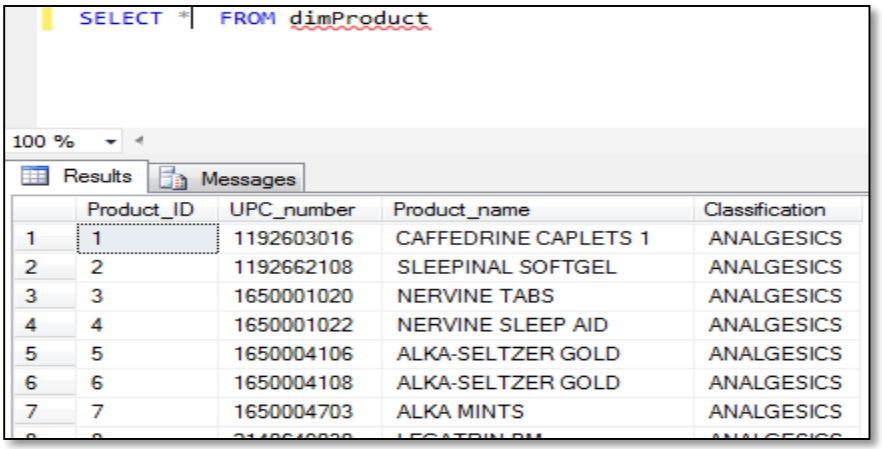


Derived Column for Product Dimension in Analgesic and Cigarette:





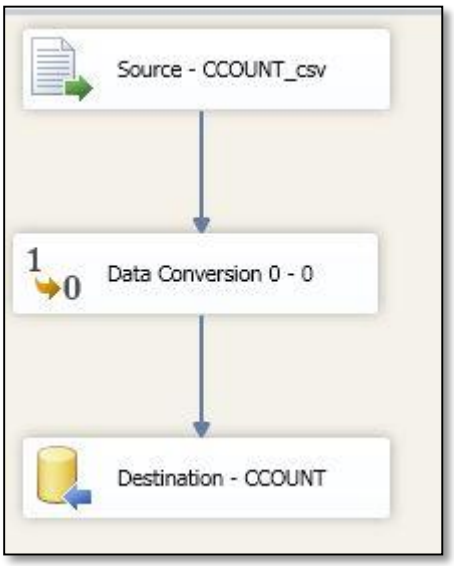
Snapshot of Data in Product Dimension



	Product_ID	UPC_number	Product_name	Classification
1	1	1192603016	CAFFEDRINE CAPLETS 1	ANALGESICS
2	2	1192662108	SLEEPINAL SOFTGEL	ANALGESICS
3	3	1650001020	NERVINE TABS	ANALGESICS
4	4	1650001022	NERVINE SLEEP AID	ANALGESICS
5	5	1650004106	ALKA-SELTZER GOLD	ANALGESICS
6	6	1650004108	ALKA-SELTZER GOLD	ANALGESICS
7	7	1650004703	ALKA MINTS	ANALGESICS

3.1.4. Category Dimension Creation

Extracting CCOUNT files in the staging area



Snapshot of CCOUNT

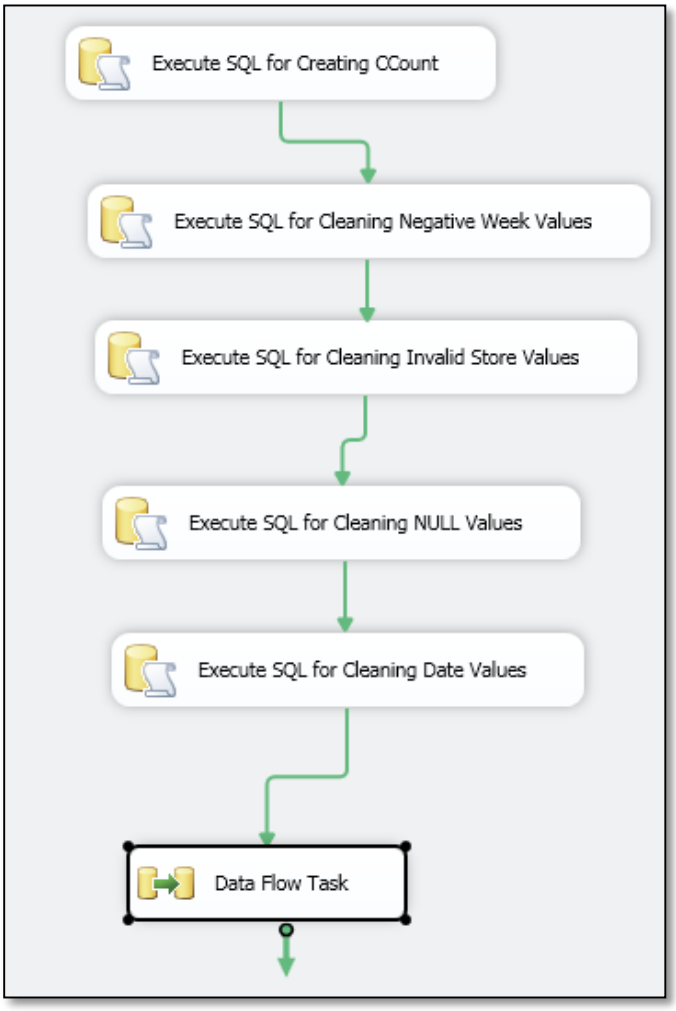
```
select * from CCOUNT;
```

100 %

Results | Messages

	"STORE"	"DATE"	"GROCERY"	"DAIRY"	"FROZEN"	"BOTTLE"	"MVPCLUB"	"GROCCOUP"	"MEAT"	"MEATFROZ"	"MEATCOUP"	"FISH"	"FISHCOUP"	"PROMO"	"PROMCOUP"
1	21	"920210"	12957.59	3515.46	2394.35	-2.4	0	-339.51	2456.35	414.44	0	130.92	0	83.88	-5
2	21	"920211"	10981.58	2843.6	2167.34	-11	0	-85.31	2261.36	376.98	0	248.03	0	43.89	-12
3	21	"920212"	13186.23	3179.05	2398.36	0	0	-123.89	2199.38	323.23	0	156.66	0	62.89	-8
4	21	"920213"	18084.96	4863.55	3565.86	-5.6	0	-438.55	4171.65	585.65	-9.5	597.62	0	158.85	-16

Cleaning CCOUNT files using SQL Queries



SQL Queries Used

SQL for Creating CCount

```
SELECT      ["STORE"], ["DATE"],["FISH"],["FISHCOUP"], ["CAMERA"],
["BEER"],["WEEK"]
INTO        cleanCCOUNT
FROM        CCOUNT
```

SQL for Cleaning Negative week Values

```
DELETE FROM cleanCCOUNT
WHERE       (["WEEK"] < 0)
```

SQL for Cleaning Invalid Store Values:

```
DELETE FROM cleanCCOUNT
WHERE       (["STORE"] NOT IN
(SELECT     Store_number
FROM        dimStore))
```

SQL for Cleaning Null Values

```
DELETE FROM cleanCCOUNT
WHERE       (["WEEK"] IS NULL) OR
(["STORE"] IS NULL) OR
(["DATE"] IS NULL) OR
(["FISH"] IS NULL) OR
(["FISHCOUP"] IS NULL) OR
(["BEER"] IS NULL) OR
(["CAMERA"] IS NULL)
```

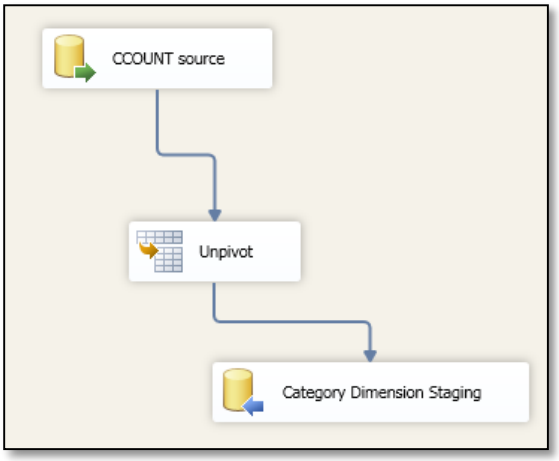
SQL for Cleaning Date Values

```
UPDATE      cleanCCOUNT
SET         ["DATE"] = REPLACE(["DATE"], "'", '')
```

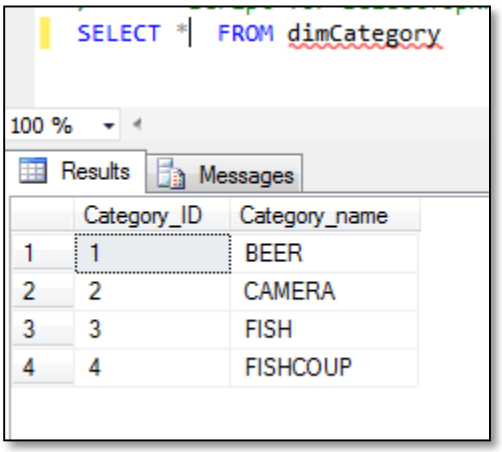
SQL for removing invalid store number

```
DELETE      from cleanCCOUNT
WHERE       ["STORE"] NOT IN (SELECT [Store_number] from dimStore);
```

Transforming CCount to Category Dimension:

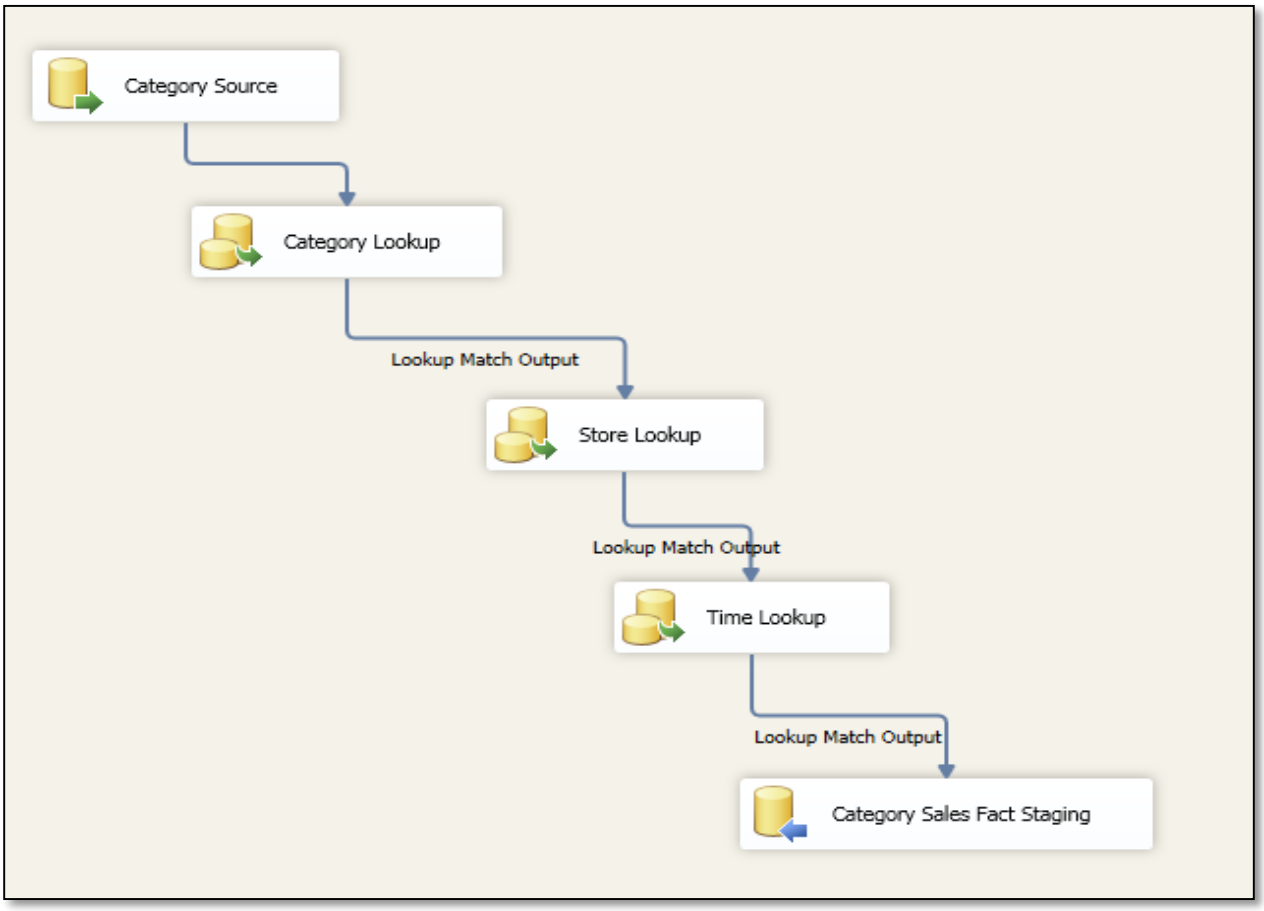


Snapshot of Category Dimension

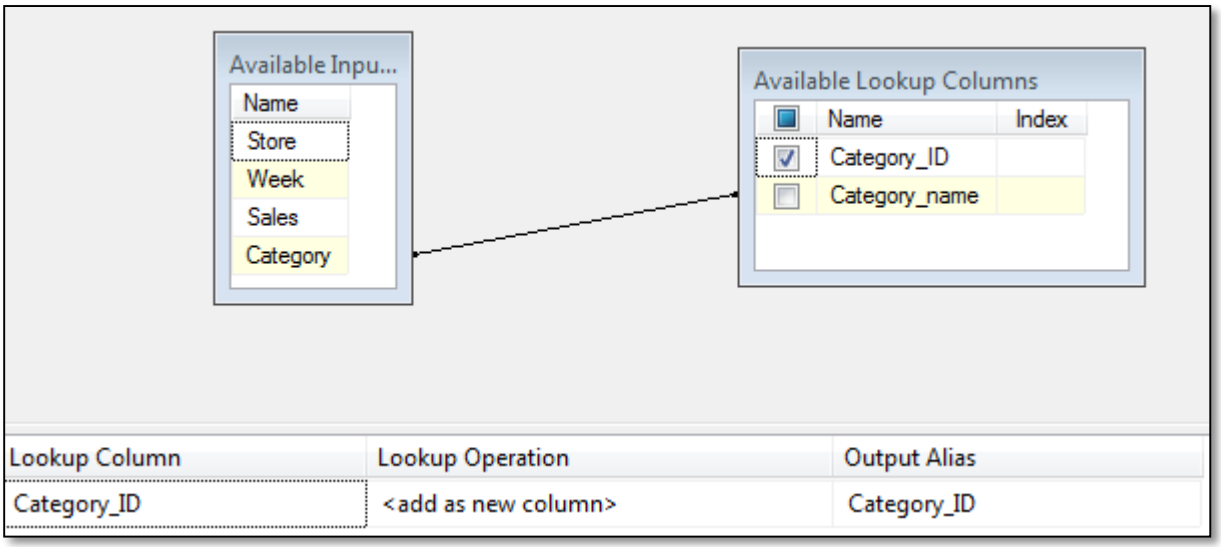


	Category_ID	Category_name
1	1	BEER
2	2	CAMERA
3	3	FISH
4	4	FISHCOUP

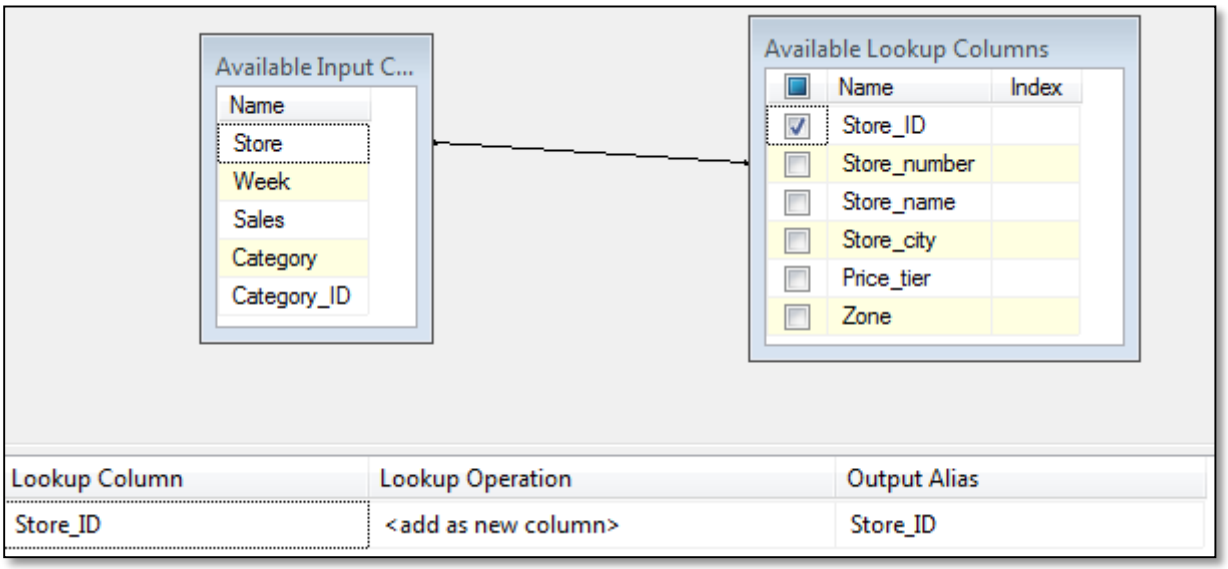
3.1.5. Category Fact Table Creation



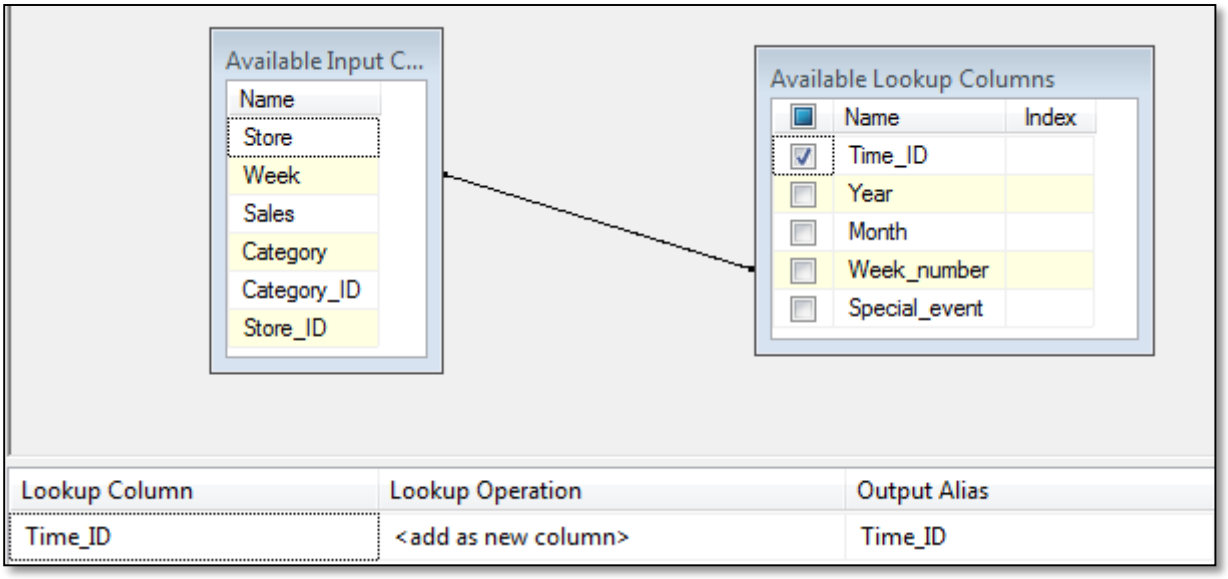
Category Lookup



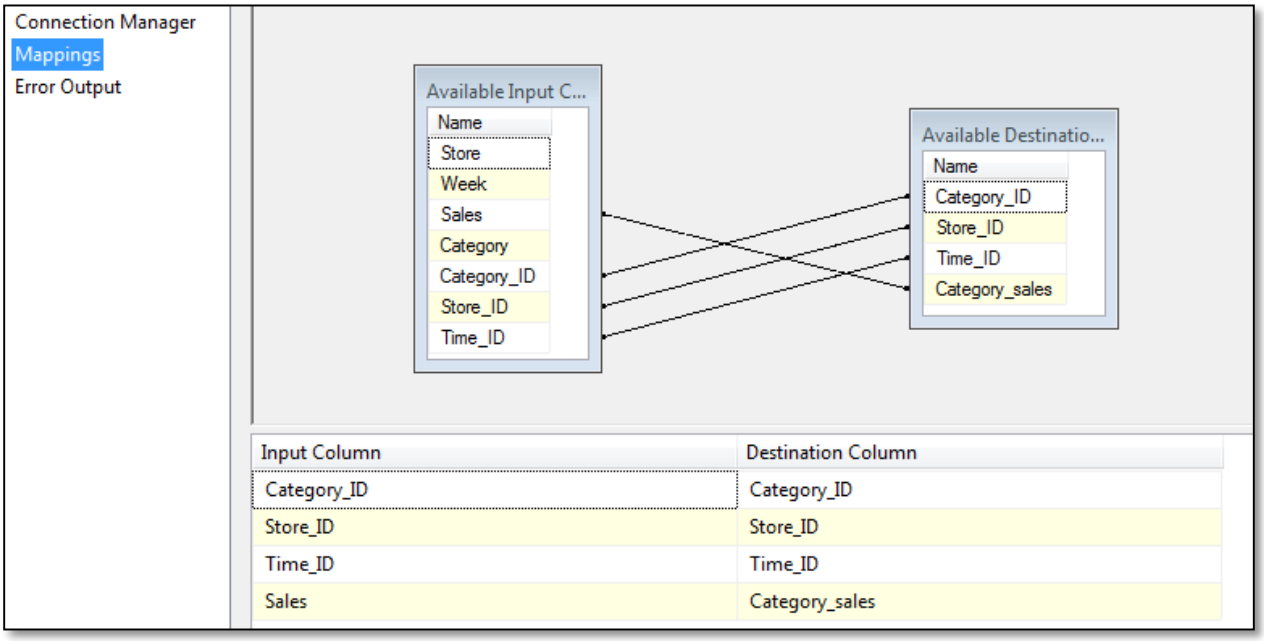
Store Lookup



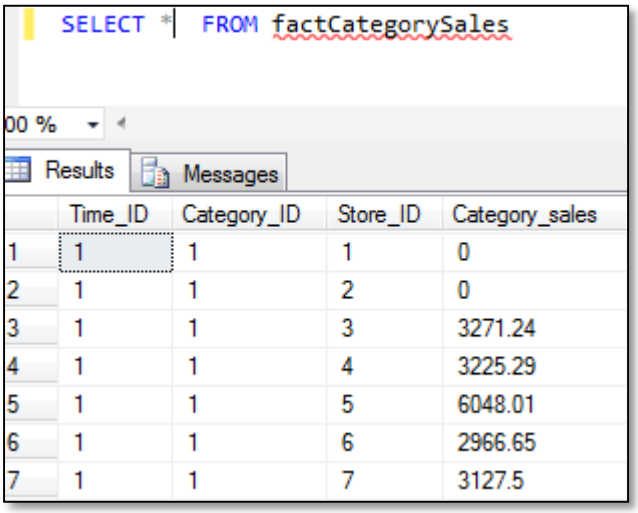
Time Lookup



Final Mapping

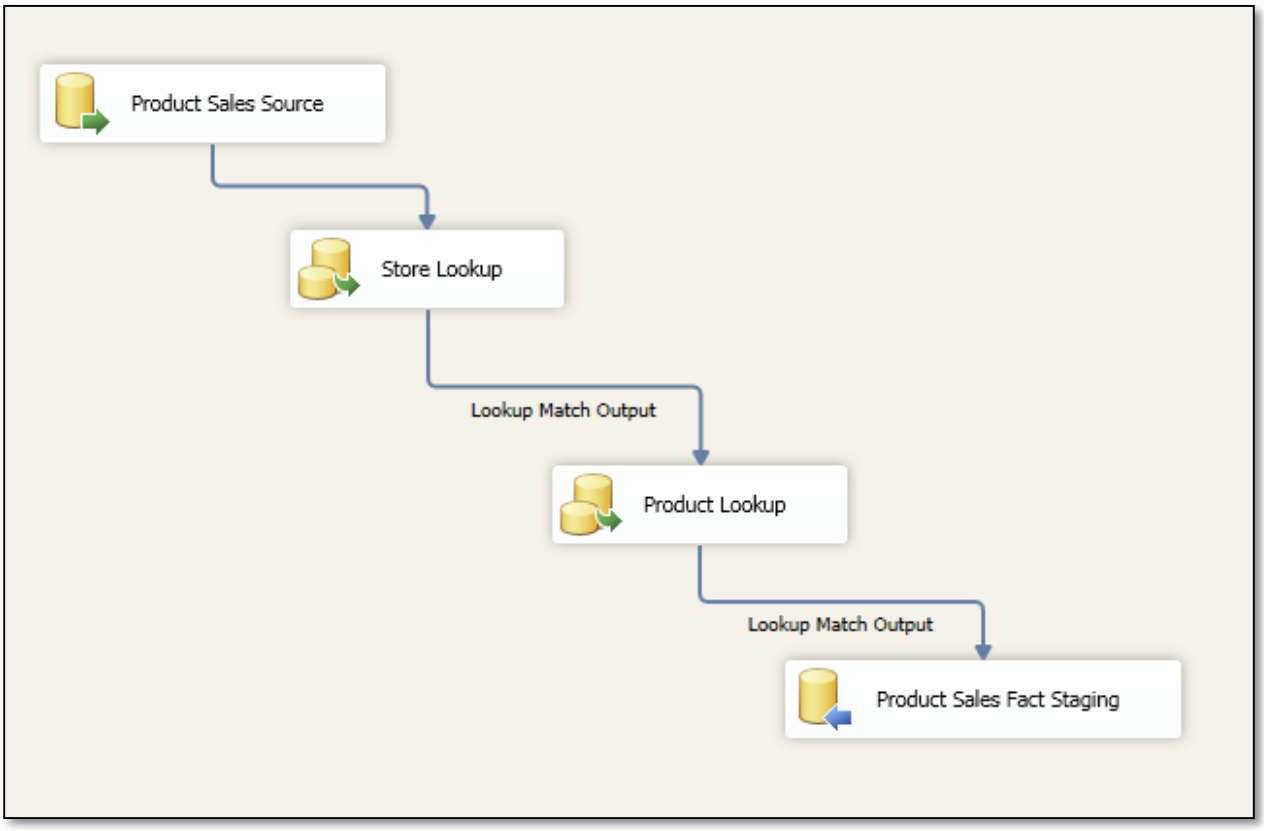


Snapshot of Data in Category Fact table

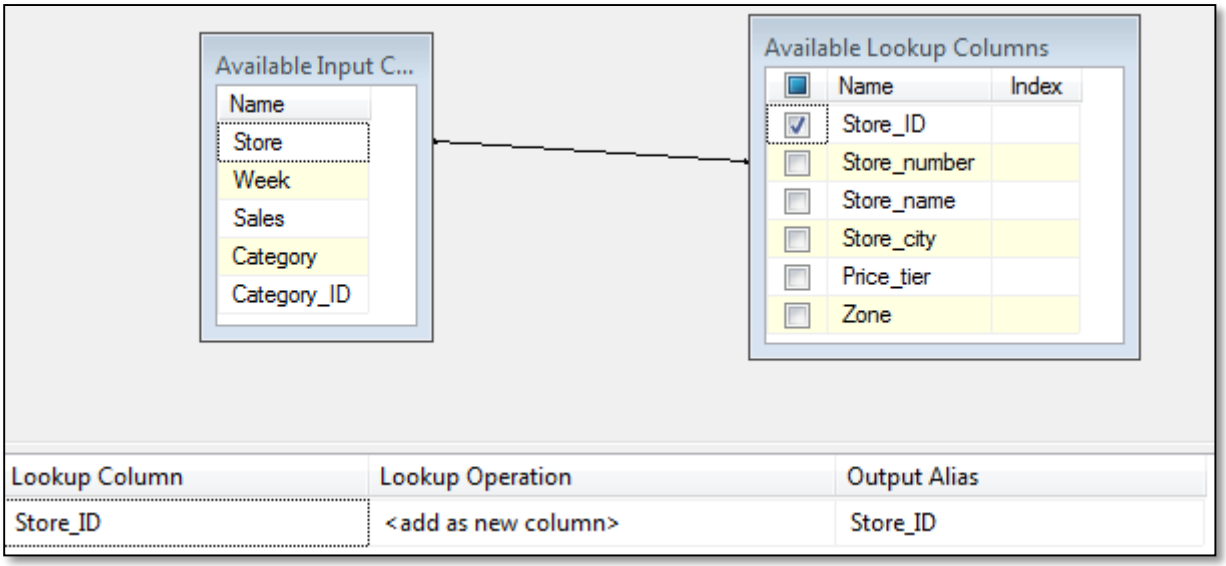


	Time_ID	Category_ID	Store_ID	Category_sales
1	1	1	1	0
2	1	1	2	0
3	1	1	3	3271.24
4	1	1	4	3225.29
5	1	1	5	6048.01
6	1	1	6	2966.65
7	1	1	7	3127.5

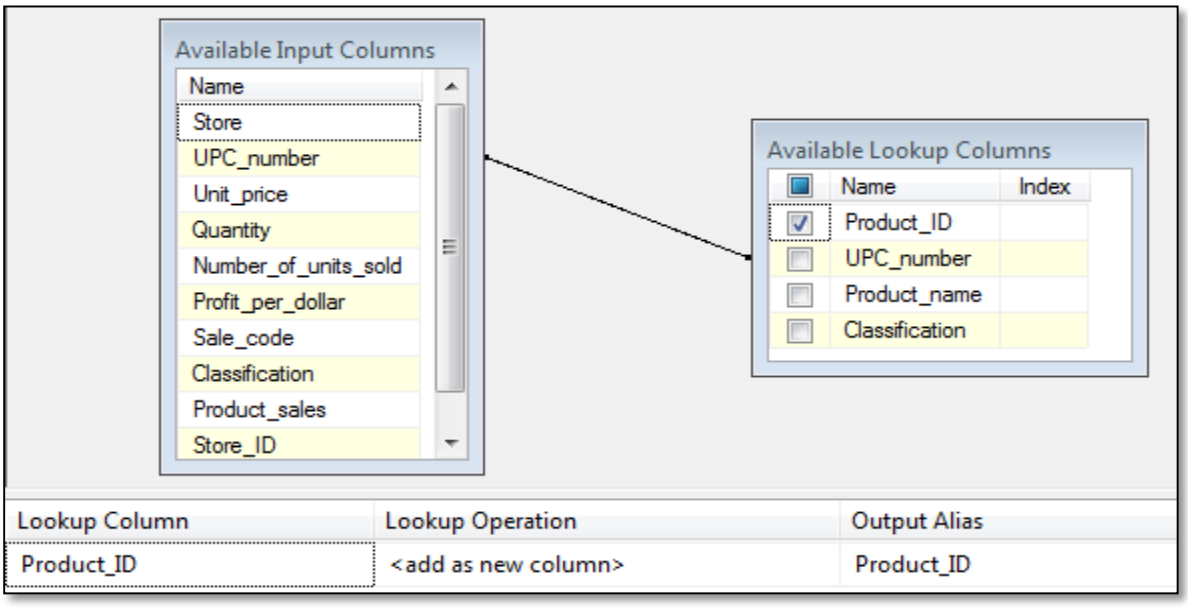
3.1.6. Product Fact Table Creation



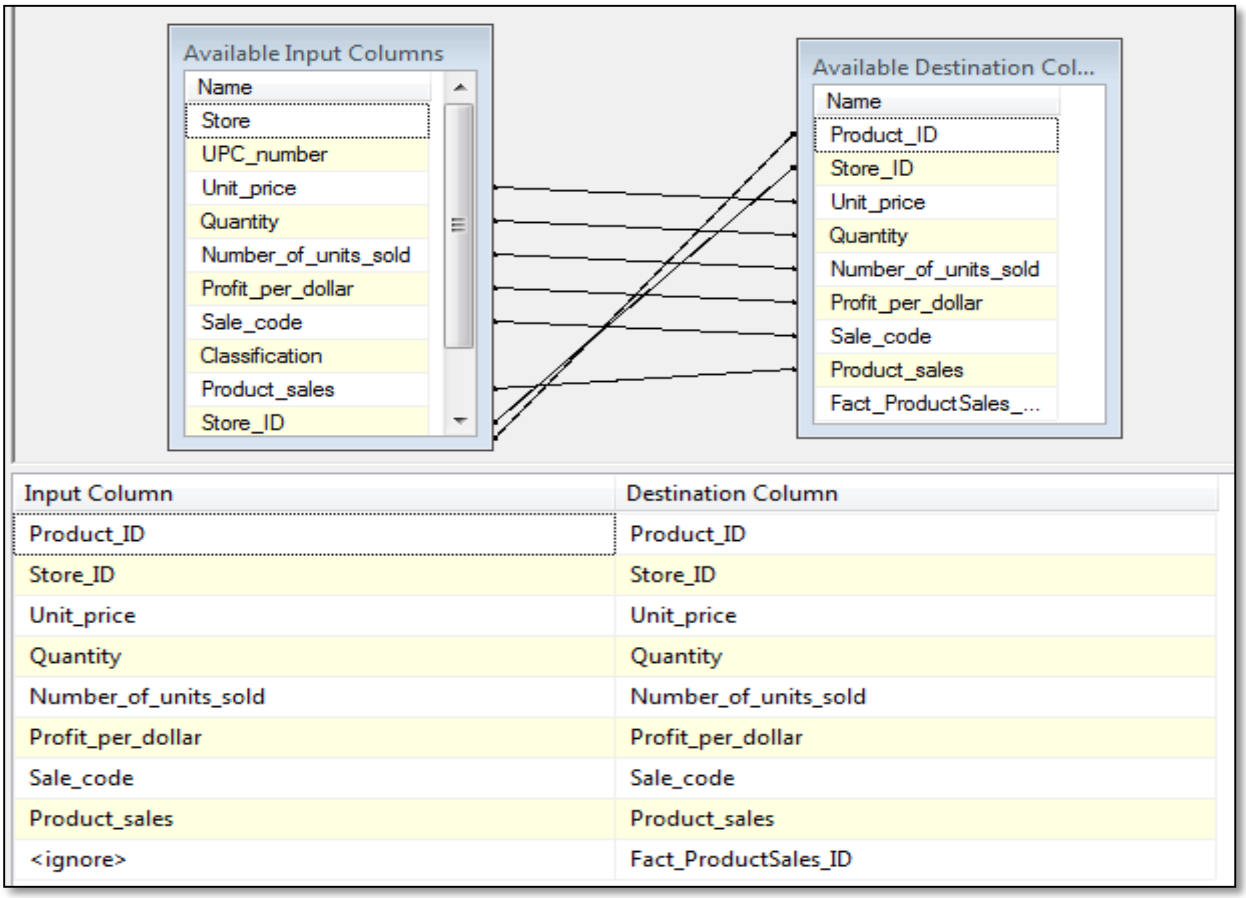
Store lookup



Product Lookup

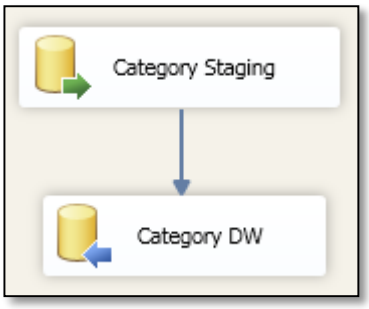


Final Mapping

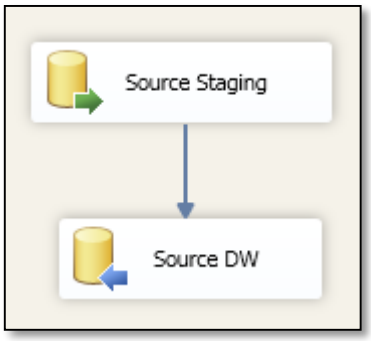


4. Loading the dimension and fact tables from Staging area to DW area

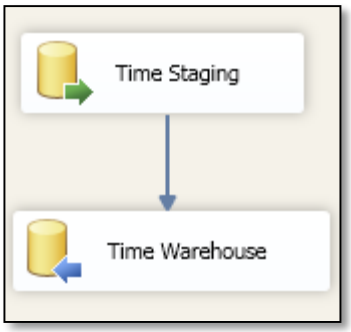
Loading of Category dimension from Staging area to Category Sales and Product Sales Data marts



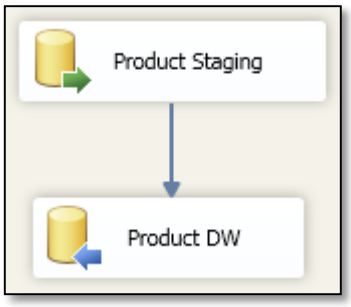
Loading of Store dimension from Staging area to Category Sales and Product Sales Data marts



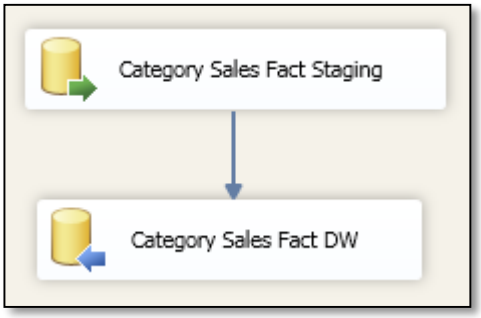
Loading of Time dimension from Staging area to Category Sales and Product Sales Data marts



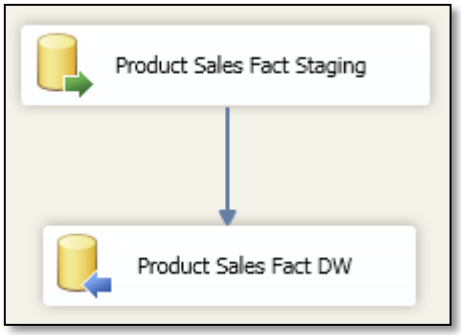
Loading of Product dimension from Staging area to Product Sales Data mart



Loading of Category Sales Fact table from Staging area to Category Sales Data mart



Loading of Product Sales Fact table from Staging area to Product Sales Data mart



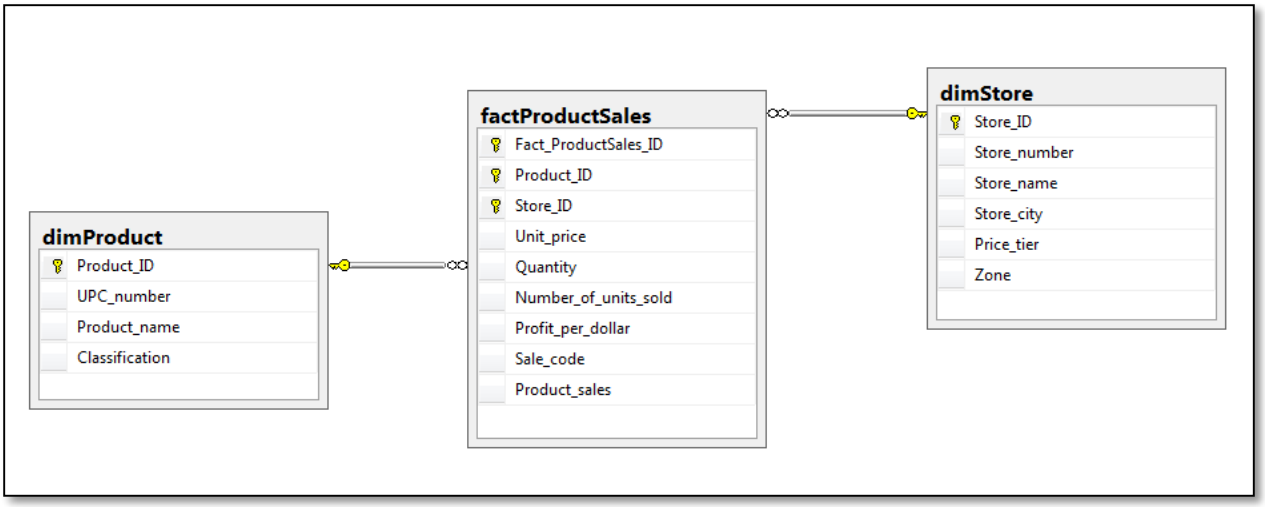


Table structure for the Product Sales Data Mart

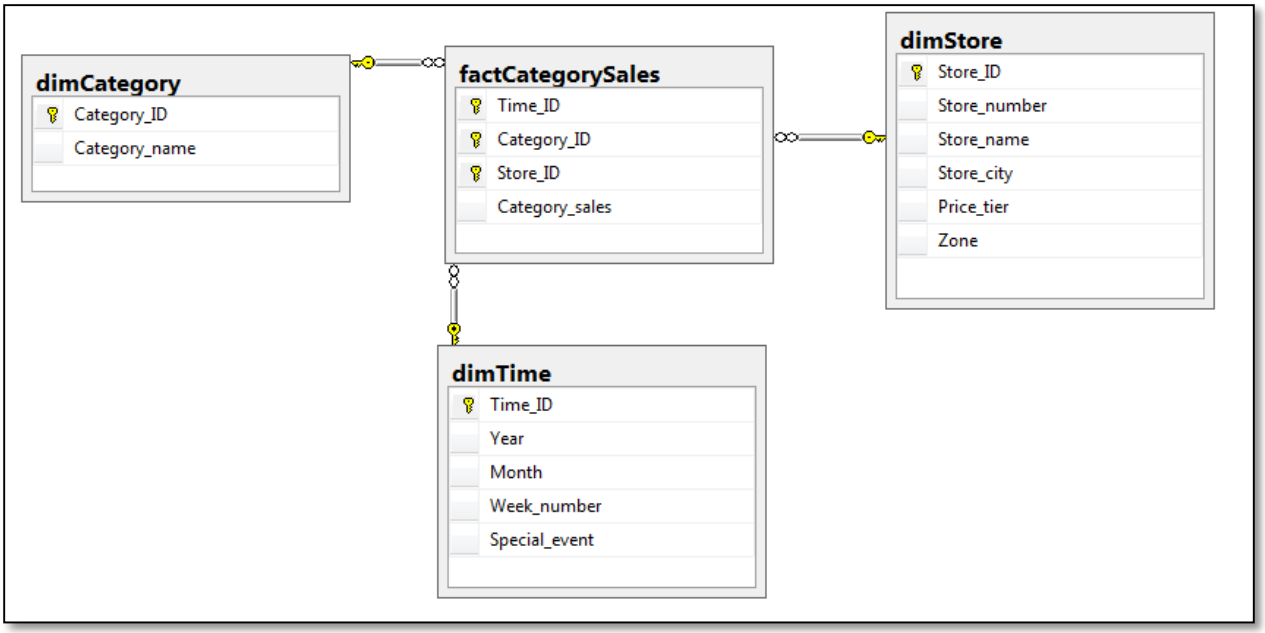
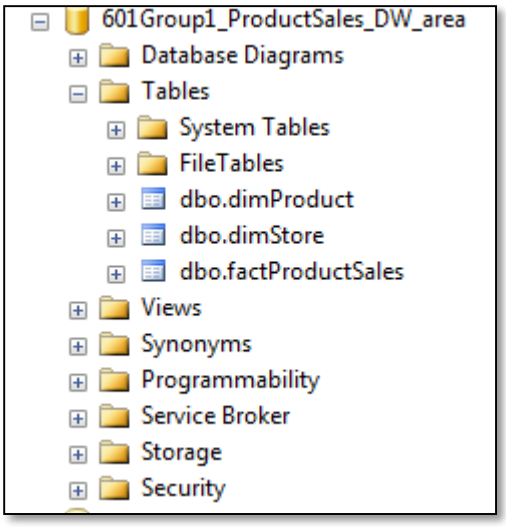


Table Structure for the Category Sales Data Mart

4.1. Snapshot of the Product Sales Data Warehouse area



The data that exists in the dimension tables of the 601Group1_ProductSales_DW_area

```
SELECT * FROM dimProduct;
```

100 %

Results | Messages

	Product_ID	UPC_number	Product_name	Classification
1	1	1192603016	CAFFEDRINE CAPLETS 1	ANALGESICS
2	2	1192662108	SLEEPINAL SOFTGEL	ANALGESICS
3	3	1650001020	NERVINE TABS	ANALGESICS
4	4	1650001022	NERVINE SLEEP AID	ANALGESICS
5	5	1650004106	ALKA-SELTZER GOLD	ANALGESICS
6	6	1650004108	ALKA-SELTZER GOLD	ANALGESICS
7	7	1650004703	ALKA MINTS	ANALGESICS
8	8	2140649030	LEGATRIN PM	ANALGESICS

```
SELECT * FROM dimStore;
```

100 %

Results | Messages

	Store_ID	Store_number	Store_name	Store_city	Price_tier	Zone
1	1	2	DOMINICKS 2	River Forest	High	1
2	2	5	DOMINICKS 5	Palatine	Medium	2
3	3	8	DOMINICKS 8	Oak Lawn	Low	5
4	4	9	DOMINICKS 9	Morton Grove	Medium	2
5	5	12	DOMINICKS 12	Chicago	High	7
6	6	14	DOMINICKS 14	Glenview	High	1

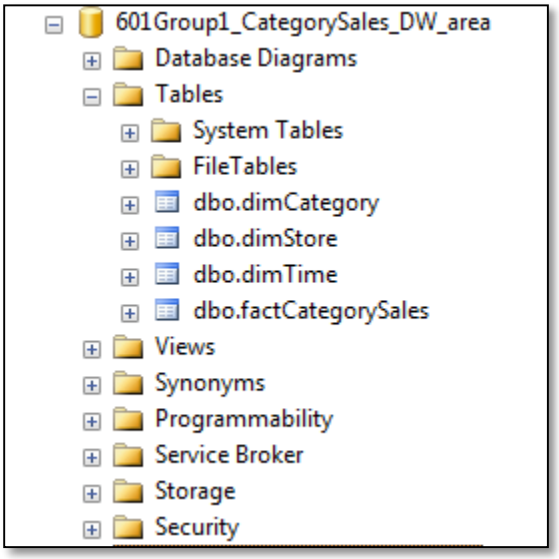
```
SELECT * FROM factProductSales;
```

100 %

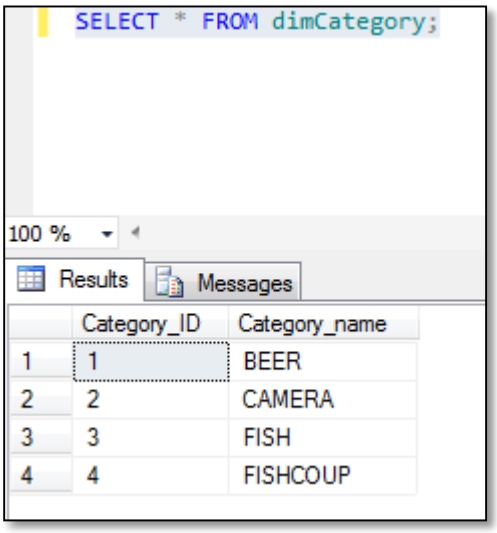
Results | Messages

	Fact_ProductSales_ID	Product_ID	Store_ID	Unit_price	Quantity	Number_of_units_sold	Profit_per_dollar	Sale_code	Product_sales
1	1	142	19	2.49	1	2	34.13		4.98
2	2	142	19	0	1	0	0		0
3	3	142	19	0	1	0	0		0
4	4	142	19	0	1	0	0		0
5	5	142	19	0	1	0	0		0
6	6	142	19	0	1	0	0		0
7	7	142	19	0	1	0	0		0

4.2. Snapshot of the Category Sales Data Warehouse area



The data that exists in the dimension tables of the 601Group1_CategorySales_DW_area



	Category_ID	Category_name
1	1	BEER
2	2	CAMERA
3	3	FISH
4	4	FISHCOUP

```
SELECT * FROM dimStore;
```

100 %

Results | Messages

	Store_ID	Store_number	Store_name	Store_city	Price_tier	Zone
1	1	2	DOMINICKS 2	River Forest	High	1
2	2	5	DOMINICKS 5	Palatine	Medium	2
3	3	8	DOMINICKS 8	Oak Lawn	Low	5
4	4	9	DOMINICKS 9	Morton Grove	Medium	2
5	5	12	DOMINICKS 12	Chicago	High	7
6	6	14	DOMINICKS 14	Glenview	High	1

```
SELECT * FROM dimTime;
```

100 %

Results | Messages

	Time_ID	Year	Month	Week_number	Special_event
1	1	1989	9	1	
2	2	1989	9	2	
3	3	1989	9	3	
4	4	1989	10	4	
5	5	1989	10	5	
6	6	1989	10	6	
7	7	1989	10	7	Halloween

```
SELECT * FROM factCategorySales;
```

100 %

Results | Messages

	Time_ID	Category_ID	Store_ID	Category_sales
1	1	1	1	0
2	1	1	2	0
3	1	1	3	3271.24
4	1	1	4	3225.29
5	1	1	5	6048.01
6	1	1	6	2966.65
7	1	1	7	3127.5
8	1	1	8	2655.25
9	1	1	9	0
10	1	1	10	3656.02

4.3. SQL statements to create Staging Area and Data Warehouse

Staging Database

CREATE DATABASE [601Group1_staging_area]

CCount

```
CREATE TABLE [dbo].[cleanCCOUNT](
["STORE"] [int] NULL,
["DATE"] [varchar](50) NULL,
["FISH"] [float] NULL,
["FISHCOUP"] [float] NULL,
["CAMERA"] [float] NULL,
["BEER"] [float] NULL,
[WEEK] [int] NULL
)
```

Category Dimension

```
CREATE TABLE [dbo].[dimCategory](
[Category_ID] [int] IDENTITY(1,1) NOT NULL,
[Category_name] [nvarchar](255) NULL,
CONSTRAINT [PK_dimCategory] PRIMARY KEY CLUSTERED
(
[Category_ID] ASC
)
```

UPC Source Files

```
CREATE TABLE [dbo].[UPCANA](
["COM_CODE"] [varchar](50) NULL,
["UPC"] [bigint] NULL,
["DESCRIP"] [varchar](50) NULL,
["SIZE"] [varchar](50) NULL,
["CASE"] [varchar](50) NULL,
["NITEM"] [varchar](50) NULL
)

CREATE TABLE [dbo].[UPCCIG](
["COM_CODE"] [varchar](50) NULL,
["UPC"] [bigint] NULL,
["DESCRIP"] [varchar](50) NULL,
["SIZE"] [varchar](50) NULL,
["CASE"] [varchar](50) NULL,
["NITEM"] [varchar](50) NULL
)
```

Product Dimension

```
CREATE TABLE [dbo].[dimProduct](
[Product_ID] [int] IDENTITY(1,1) NOT NULL,
[UPC_number] [bigint] NULL,
[Product_name] [varchar](50) NULL,
[Classification] [nvarchar](20) NULL,
CONSTRAINT [PK_dimProduct] PRIMARY KEY CLUSTERED
```

```
(
[Product_ID] ASC
)
```

Store Source

```
CREATE TABLE [dbo].[Store](
[Store] [int] NULL,
[City] [varchar](50) NULL,
[Price Tier] [varchar](50) NULL,
[Zone] [int] NULL,
[Zip Code] [int] NULL,
[Address] [varchar](50) NULL
)
```

Store Demographics Source

Store Dimension

```
CREATE TABLE [dbo].[dimStore](
[Store_ID] [int] IDENTITY(1,1) NOT NULL,
[Store_number] [int] NULL,
[Store_name] [nvarchar](50) NULL,
[Store_city] [varchar](50) NULL,
[Price_tier] [varchar](50) NULL,
[Zone] [int] NULL,
CONSTRAINT [PK_dimStore] PRIMARY KEY CLUSTERED
(
[Store_ID] ASC
)
```

Week Decode Source

```
CREATE TABLE [dbo].[finalWeekDecode](
[Week #] [int] NULL,
[Start] [date] NULL,
[End] [date] NULL,
[Special Events] [varchar](50) NULL
)
```

Time Dimension

```
CREATE TABLE [dbo].[dimTime](
[Time_ID] [int] IDENTITY(1,1) NOT NULL,
[Year] [int] NULL,
[Month] [int] NULL,
[Week_number] [int] NULL,
[Special_event] [varchar](50) NULL,
CONSTRAINT [PK_dimTime] PRIMARY KEY CLUSTERED
(
[Time_ID] ASC
)
```

Category Sales Fact table Source

```
CREATE TABLE [dbo].[CategorySales](
[Store] [int] NULL,
```

[Week] [int] NULL,
[Date] [date] NULL,
[Sales] [float] NULL,
[Category] [nvarchar](8) NULL
)

CREATE TABLE [dbo].[Total](
[Sales] [float] NULL,
[Week] [int] NULL,
[Category] [nvarchar](8) NULL,
[Store] [int] NULL)

Category Sales Fact table

CREATE TABLE [dbo].[factCategorySales](
[Time_ID] [int] NOT NULL,
[Category_ID] [int] NOT NULL,
[Store_ID] [int] NOT NULL,
[Category_sales] [float] NULL,
PRIMARY KEY CLUSTERED
(
[Time_ID] ASC,
[Category_ID] ASC,
[Store_ID] ASC
)

Product Movement Source files

CREATE TABLE [dbo].[DONE-WANA](
["STORE"] [int] NULL,
["UPC"] [bigint] NULL,
["WEEK"] [int] NULL,
["MOVE"] [int] NULL,
["QTY"] [int] NULL,
["PRICE"] [float] NULL,
["SALE"] [varchar](50) NULL,
["PROFIT"] [float] NULL,
["OK"] [int] NULL
)

CREATE TABLE [dbo].[Done-WCIG](
["STORE"] [int] NULL,
["UPC"] [bigint] NULL,
["WEEK"] [int] NULL,
["MOVE"] [int] NULL,
["QTY"] [int] NULL,
["PRICE"] [float] NULL,
["SALE"] [varchar](50) NULL,
["PROFIT"] [float] NULL,
["OK"] [int] NULL
)
CREATE TABLE [dbo].[cleanWANA](
["STORE"] [int] NULL,

["UPC"] [bigint] NULL,
["WEEK"] [int] NULL,
["MOVE"] [int] NULL,
["QTY"] [int] NULL,
["PRICE"] [float] NULL,
["SALE"] [varchar](50) NULL,
["PROFIT"] [float] NULL,
["OK"] [int] NULL
)
CREATE TABLE [dbo].[cleanWCIG](
["STORE"] [int] NULL,
["UPC"] [bigint] NULL,
["WEEK"] [int] NULL,
["MOVE"] [int] NULL,
["QTY"] [int] NULL,
["PRICE"] [float] NULL,
["SALE"] [varchar](50) NULL,
["PROFIT"] [float] NULL,
["OK"] [int] NULL
)

Product Sales Fact table source

CREATE TABLE [dbo].[finalProductSource](
[Store] [int] NULL,
[UPC_number] [bigint] NULL,
[Unit_price] [float] NULL,
[Quantity] [int] NULL,
[Number_of_units_sold] [int] NULL,
[Profit_per_dollar] [float] NULL,
[Sale_code] [varchar](50) NULL,
[Classification] [nvarchar](10) NULL,
[Product_sales] [float] NULL
)

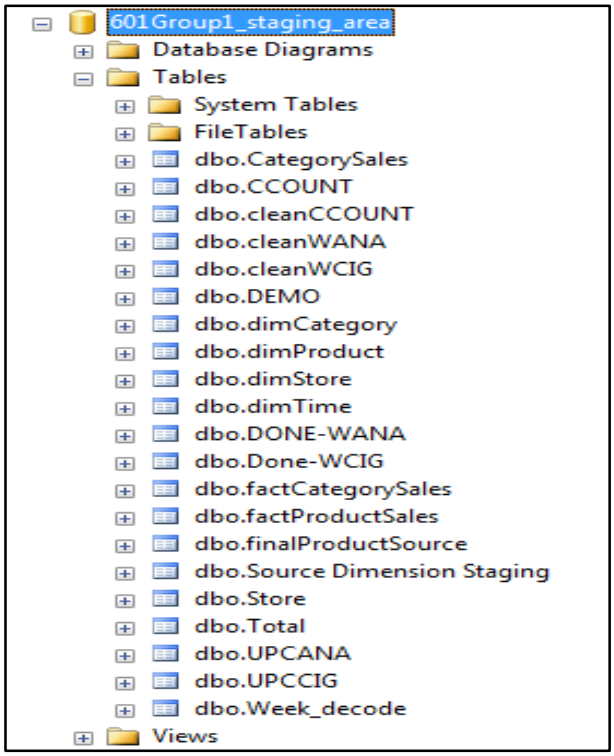
Product Sales Fact table

CREATE TABLE [dbo].[factProductSales](
[Fact_ProductSales_ID] [int] IDENTITY(1,1) NOT NULL,
[Product_ID] [int] NOT NULL,
[Store_ID] [int] NOT NULL,
[Unit_price] [float] NULL,
[Quantity] [int] NULL,
[Number_of_units_sold] [int] NULL,
[Profit_per_dollar] [float] NULL,
[Sale_code] [varchar](50) NULL,
[Product_sales] [float] NULL,
PRIMARY KEY CLUSTERED
([Fact_ProductSales_ID] ASC,
[Product_ID] ASC,
[Store_ID] ASC
)

Category Sales Data Mart
CREATE DATABASE [601Group1_CategorySales_DW_area]

Product Sales Data Mart
CREATE DATABASE [601Group1_ProductSales_DW_area]

4.4. Removal of temporary tables that exists in the data staging area



The tables existing before data transformation process which contained dirty data were removed. The above snapshot lists the tables which were retained in the staging area after the transformation process and required for the loading process.

E. BI reporting (use SSRS, SSAS and Report Builder 2012)

1. Reporting plan

> "Business intelligence (BI) is a technology-driven process for analyzing data and presenting actionable information to help corporate executives, business managers and other end users make more informed business decisions. BI encompasses a variety of tools, applications and methodologies that enable organizations to collect data from internal systems and external sources, prepare it for analysis, develop and run queries against the data, and create reports, dashboards and data visualizations to make the analytical results available to corporate decision makers as well as operational workers."[2]

We are using SSRS, SSAS, SSAS+SSRS and ReportBuilder3.0 in our report building with the following breakdown for each of the reports:

Reporting Tool	Question Number
SSRS	Question 1
SSAS	Question 5
SSAS + SSRS	Question 2
ReportBuilder3.0	Question 3 & 4

1.1. Determine all target reports that satisfy business questions.

Question1: What is the trend of Beer Sales during Thanksgiving's week for the entire duration?

Report generated from SSRS alone

To solve this business question, we used the category name as BEER from the Category dimension, week number in ascending order and special event as "Thanksgiving" from the Time dimension and mapped them with the Category Sales fact table to get the Category sales. The sales from the fact table will be the trend of Beer sales during the Thanksgiving week for the entire duration. We have used SQL query and SSRS to generate the report for this business question. We have made use of two chart types to give a better visualization to answer this question. One is a bar chart and the other is the pie chart.

Question 2: How are the average price and sales of a particular product changing according to different zones (Fish and Fish Coupon)?

Report generated from SSRS on top of SSAS

In order to answer this business question, we created a cube in the Analytics Services of the SQL Server data tools. We used the Category Sales Data mart with Category, Time and Store dimensions along with the Category Sales fact table. We then created a SQL query to using the attributes Zone and Store number from the Store dimension, Category name from the Category dimension and Category Sales from the Category Sales fact table. After successful deployment of the cube on server, we used SSRS on top of SSAS cube to generate a report. We used a bar chart to visualize the average sales of the Fish and Fish coupon changing according to different zones.

Question 3: Compare the effect of Bonus buy and Price Reduction in Analgesics in different zones.

Report generated from Report Builder 3.0

We used independent data mart – Product Sales to answer this business question. Bonus buy and Price Reduction are the sale code come from the fact table are stored in the form of codes such as 'B' – Bonus Buy and 'S' – Price Reduction. We used the Store dimension, Product dimension and Product Sales fact table from this data mart to build our report. We have grouped the Average of Product sales from Product Sales fact table and the Zones from Store dimension by Sale code. The attributes such as Product ID are used to create inner joins in the SQL

query. We used the bar chart to visualize the effect of Bonus Buy and Price Reduction.

Question 4: Plot the average profit margin for cigarettes across all the stores. Determine the average of profit for the sales of cigarettes and the stores which are below the average.

Report generated from Report Builder 3.0

We used independent data mart – Product Sales to answer this business question. The profit margin can be determined from the Profit per dollar attribute that gives the amount of profit a product received on per dollar sale. We used the Store dimension, Product dimension and Product Sales fact table from this data mart to build our report. We have grouped the Average of Profit per dollar by Store number. The attributes such as Product ID are used to create inner joins in the SQL query. We have used bar chart to plot and visualize the average profit margin for cigarettes across all the stores.

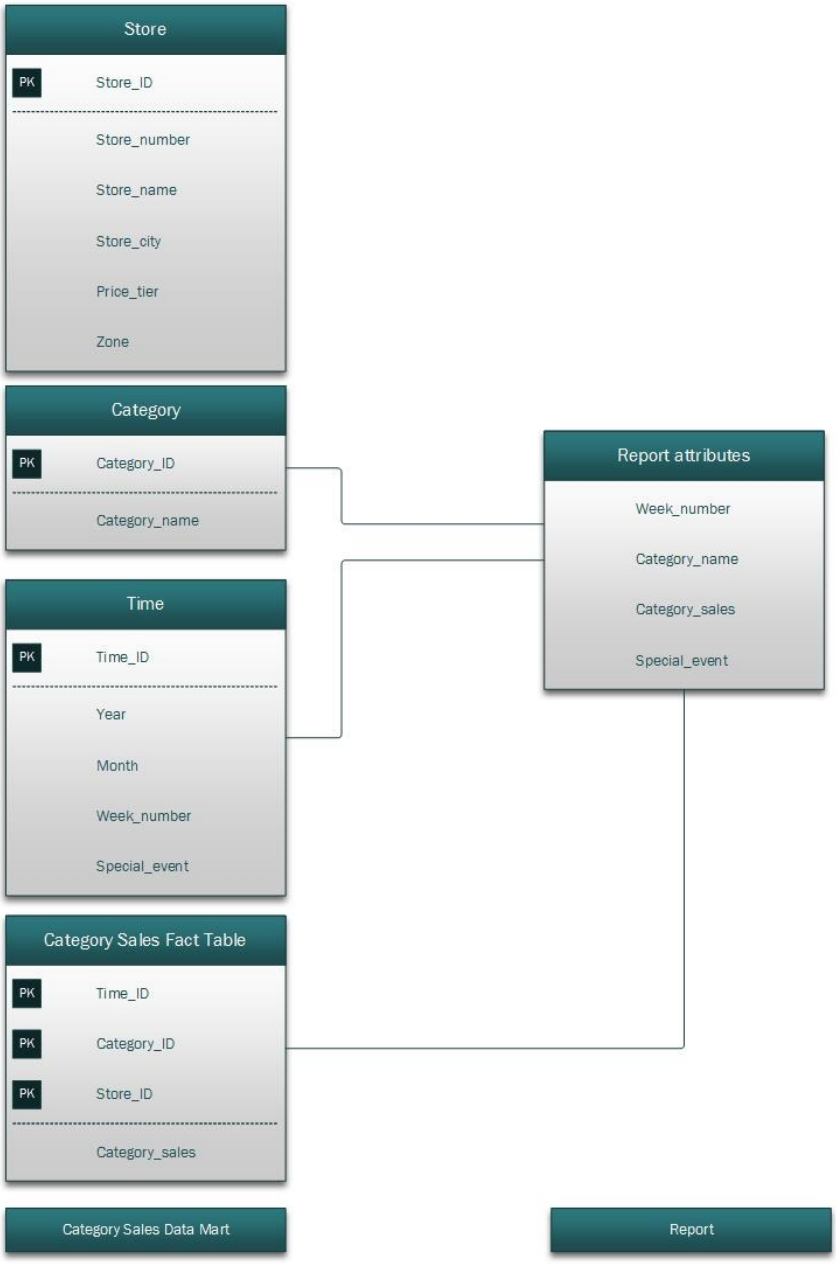
Question 5: What is the trend of Camera sales from the year 1990 to 1996?

Analysis of the cube created from SSAS only

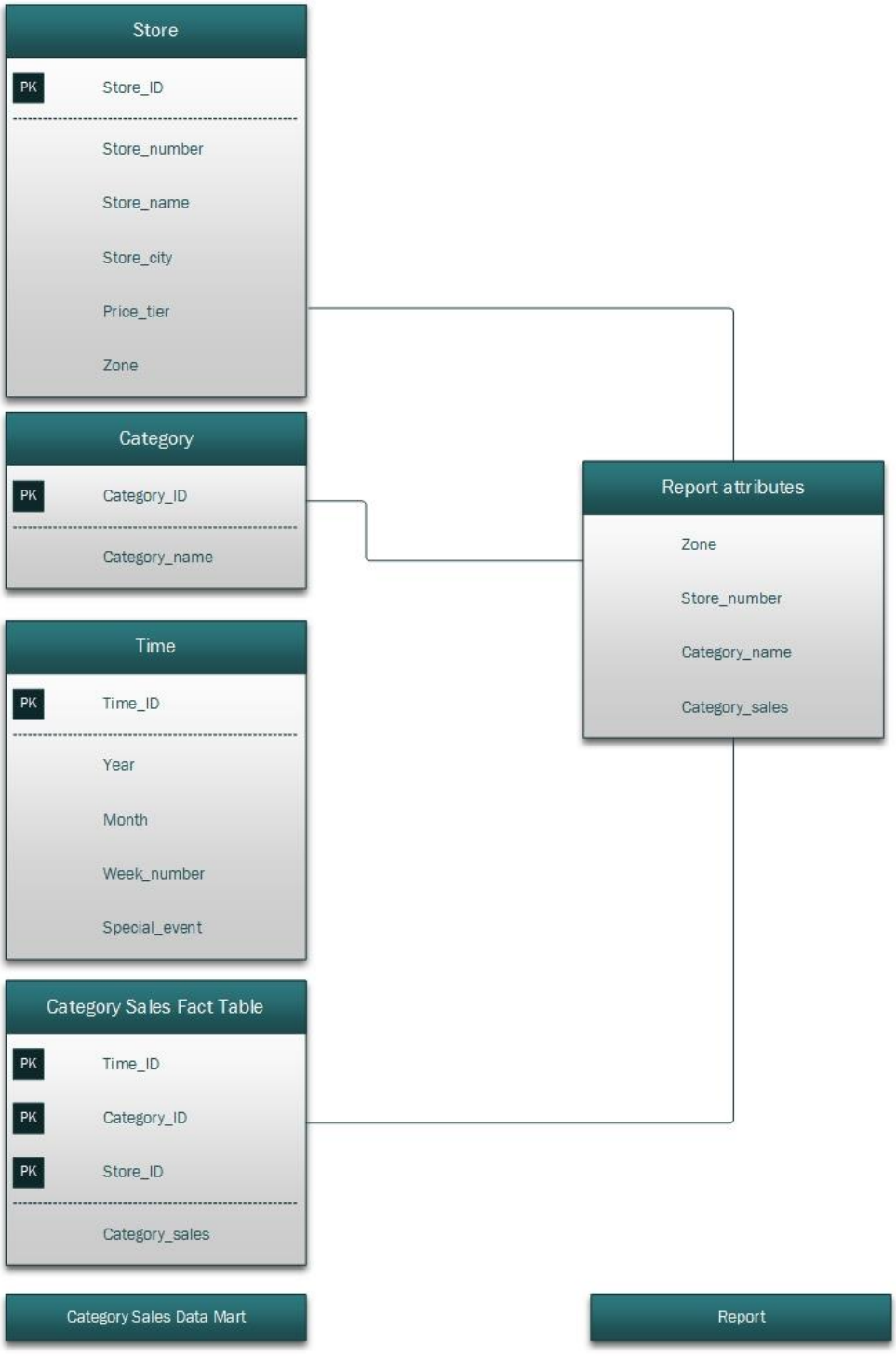
We used SSAS only to answer this business question. We deployed a cube which consists of the Category sales fact table, Store dimension, Category dimension and Time dimension tables. The cube groups the Category sales by the Category name "Camera" and the SQL query gives the results of the sales of camera trending over the period of 1990 to 1996. In order to analyze the results, we have chosen Pivot charts with bar graph for the visualization of the trend.

1.2. Mapping s from the tables in the data marts to the attributes in the report

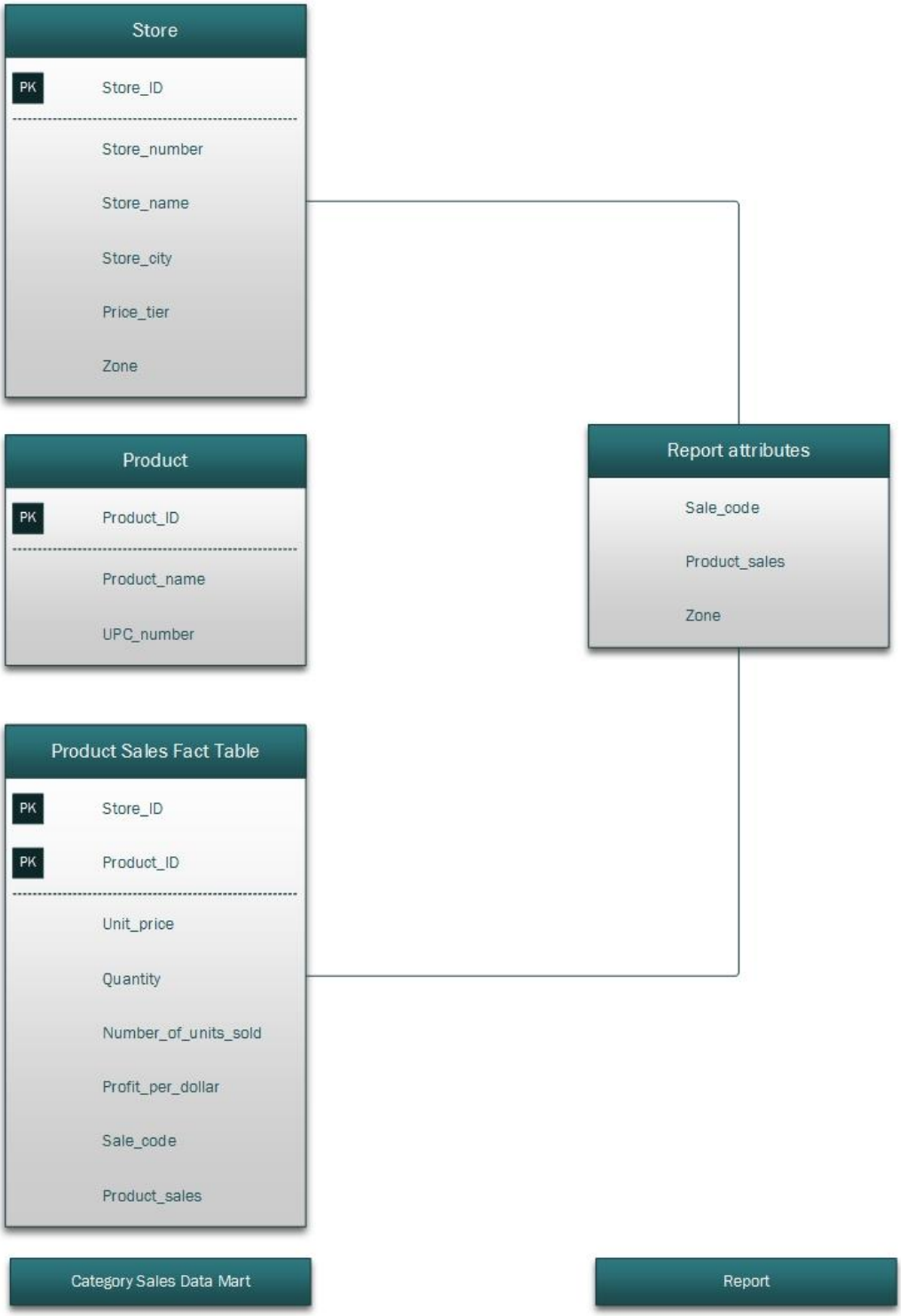
Question1: What is the trend of Beer Sales during Thanksgiving's week for the entire duration?



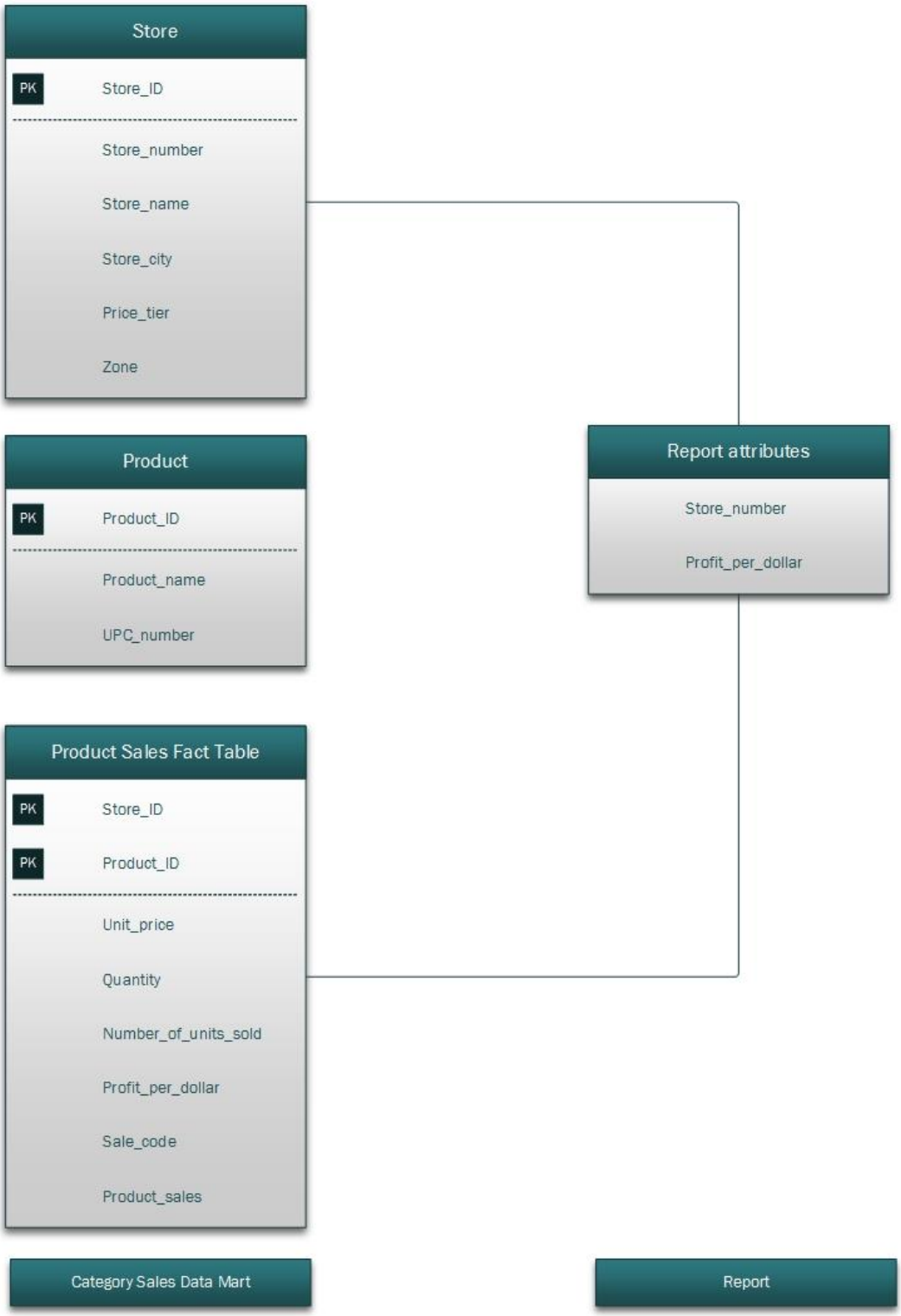
Question 2: How are the average price and sales of a particular product changing according to different zones (Fish and Fish Coupon)?



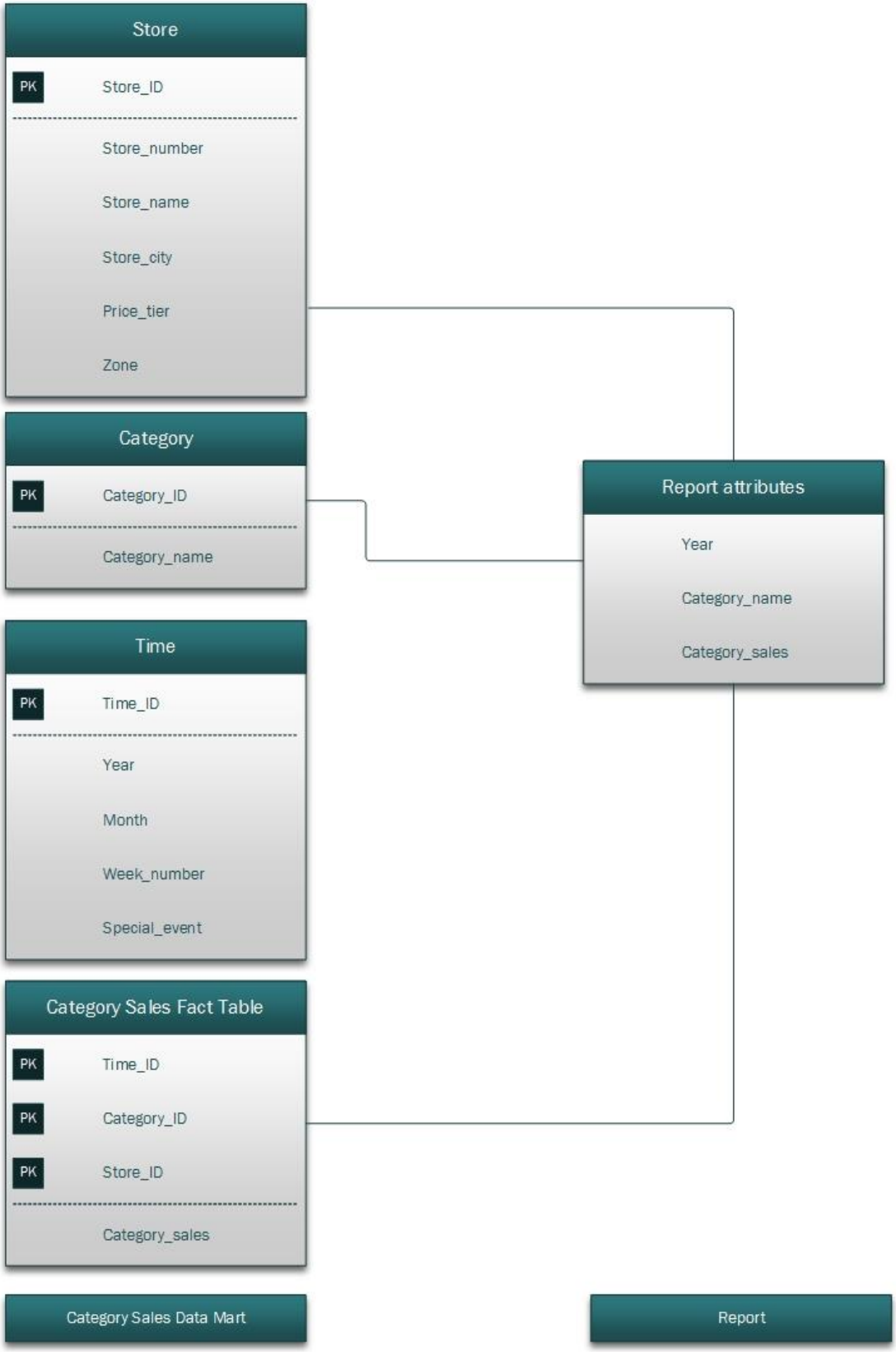
Question 3: Compare the effect of Bonus buy and Price Reduction in Analgesics in different zones.



Question 4: Plot the average profit margin for cigarettes across all the stores. Determine the average of profit for the sales of cigarettes and the stores which are below the average.



Question 5: What is the trend of Camera sales from the year 1990 to 1996?

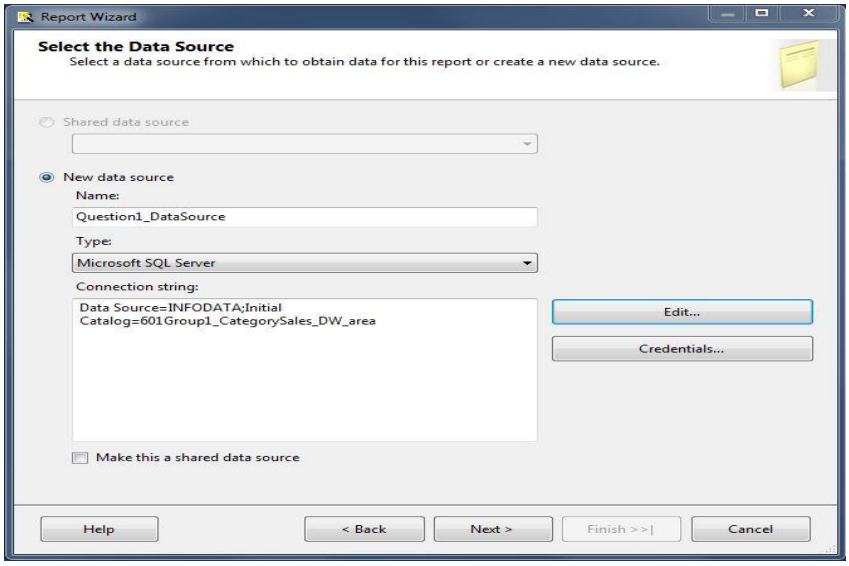


2. Report Building from Individual Data Mart using SSRS for Question 1

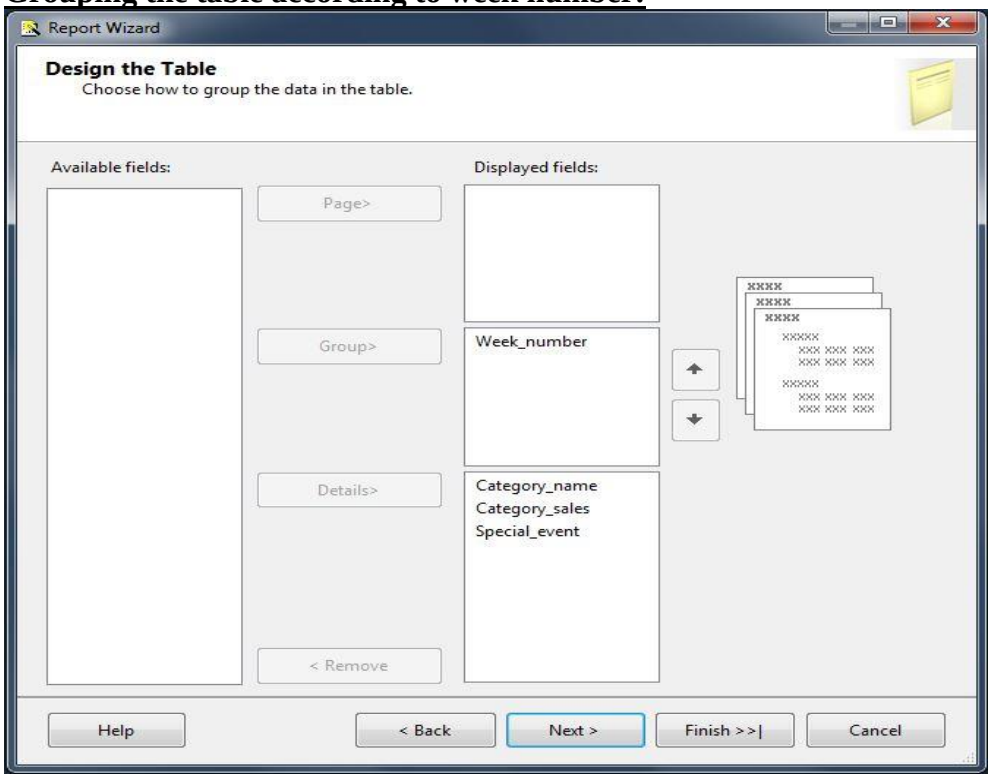
Based on the reporting plan discussed in the report earlier, we are using report building from individual data mart using SSRS approach to create visualizations for question 1.

Question 1: What is the trend of Beer Sales during Thanksgiving's week for the entire duration?

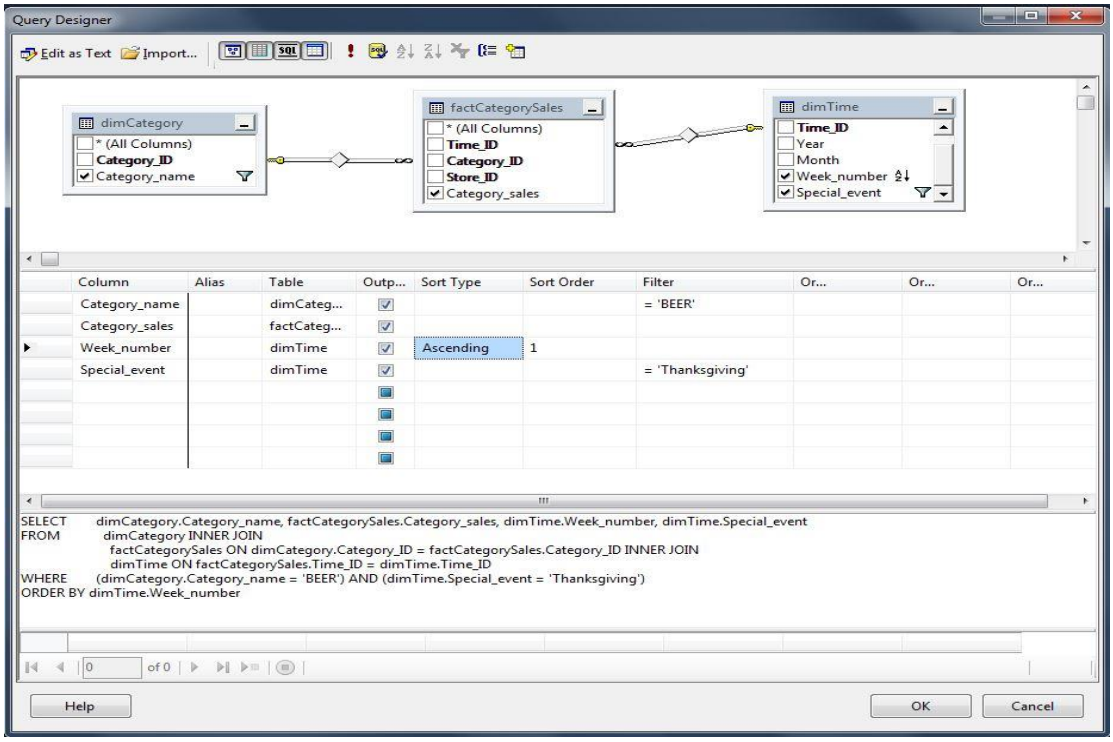
Data Source Creation:

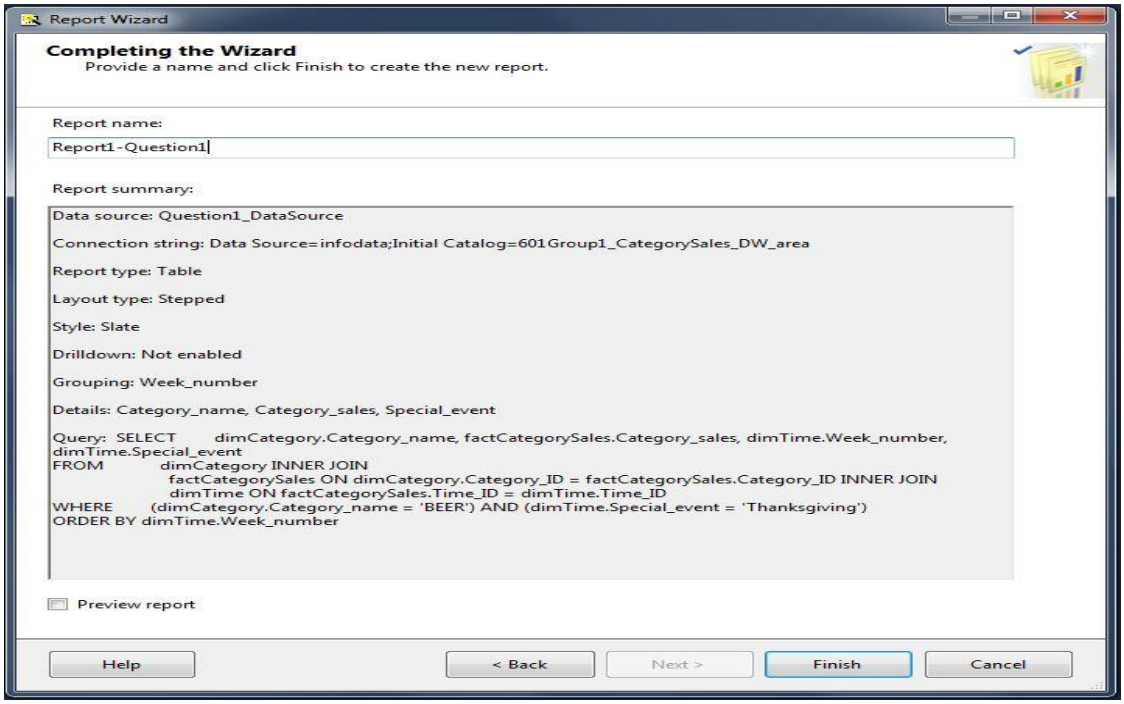


Grouping the table according to week number:



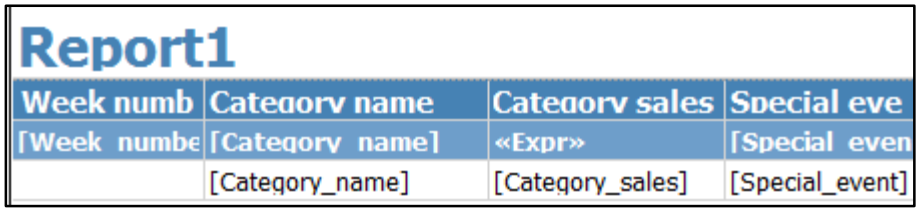
Query Designer:



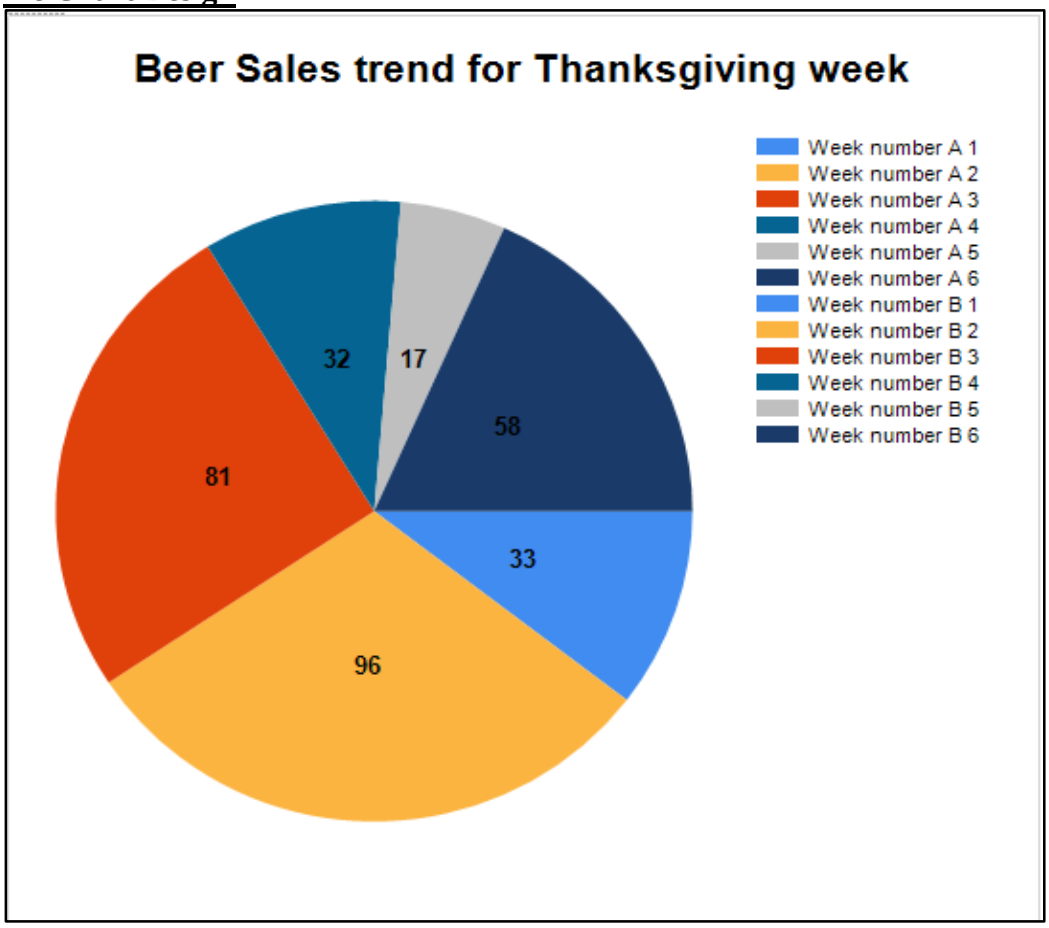


Final Chart Creation

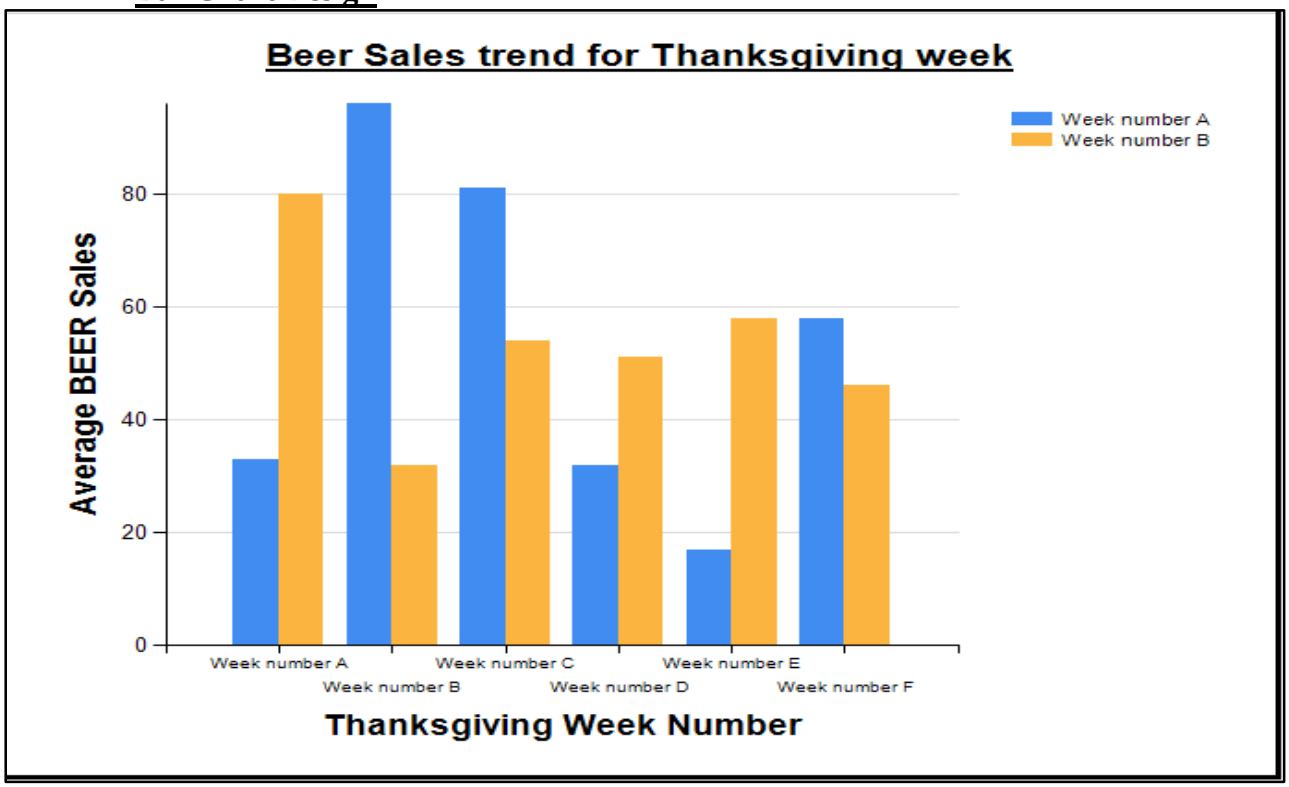
Report Design



Pie Chart Design

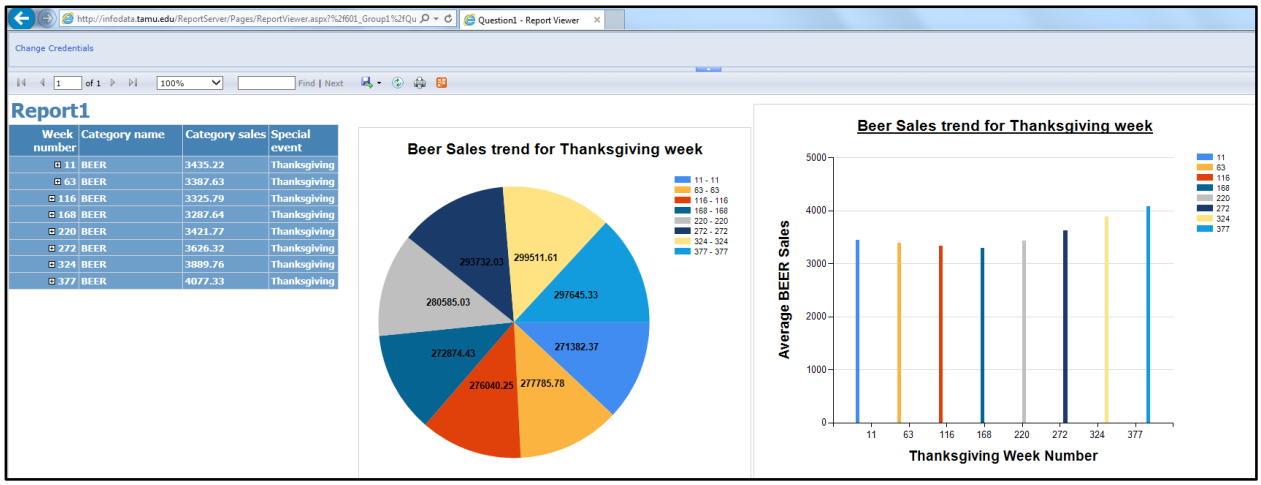


Bar Chart Design

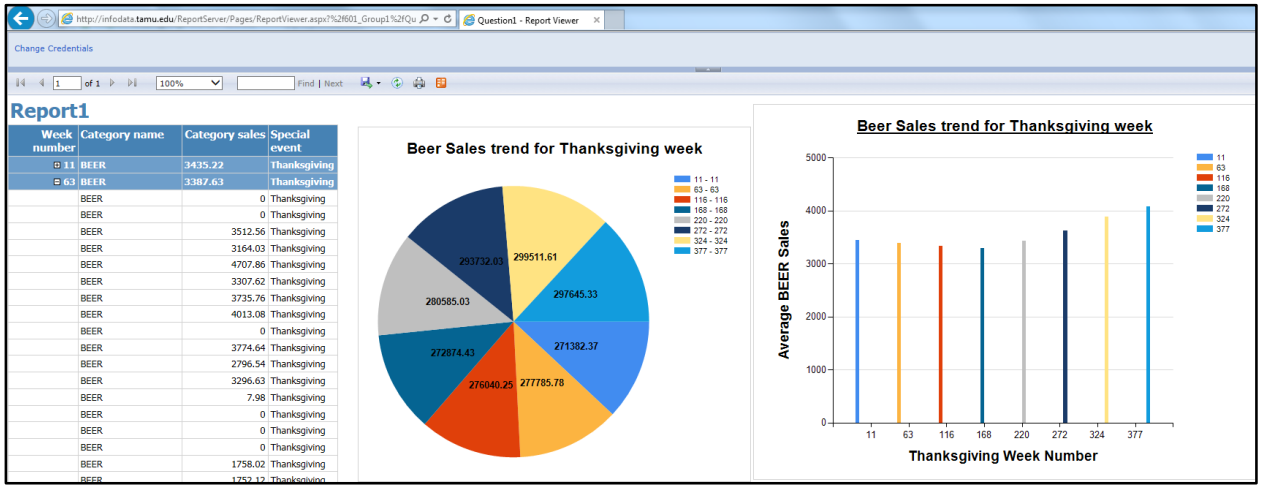


Deployment of Report in infodata.tamu.edu

Report deployed before Drilldown:



Report deployed after Drilldown:



Conclusion:

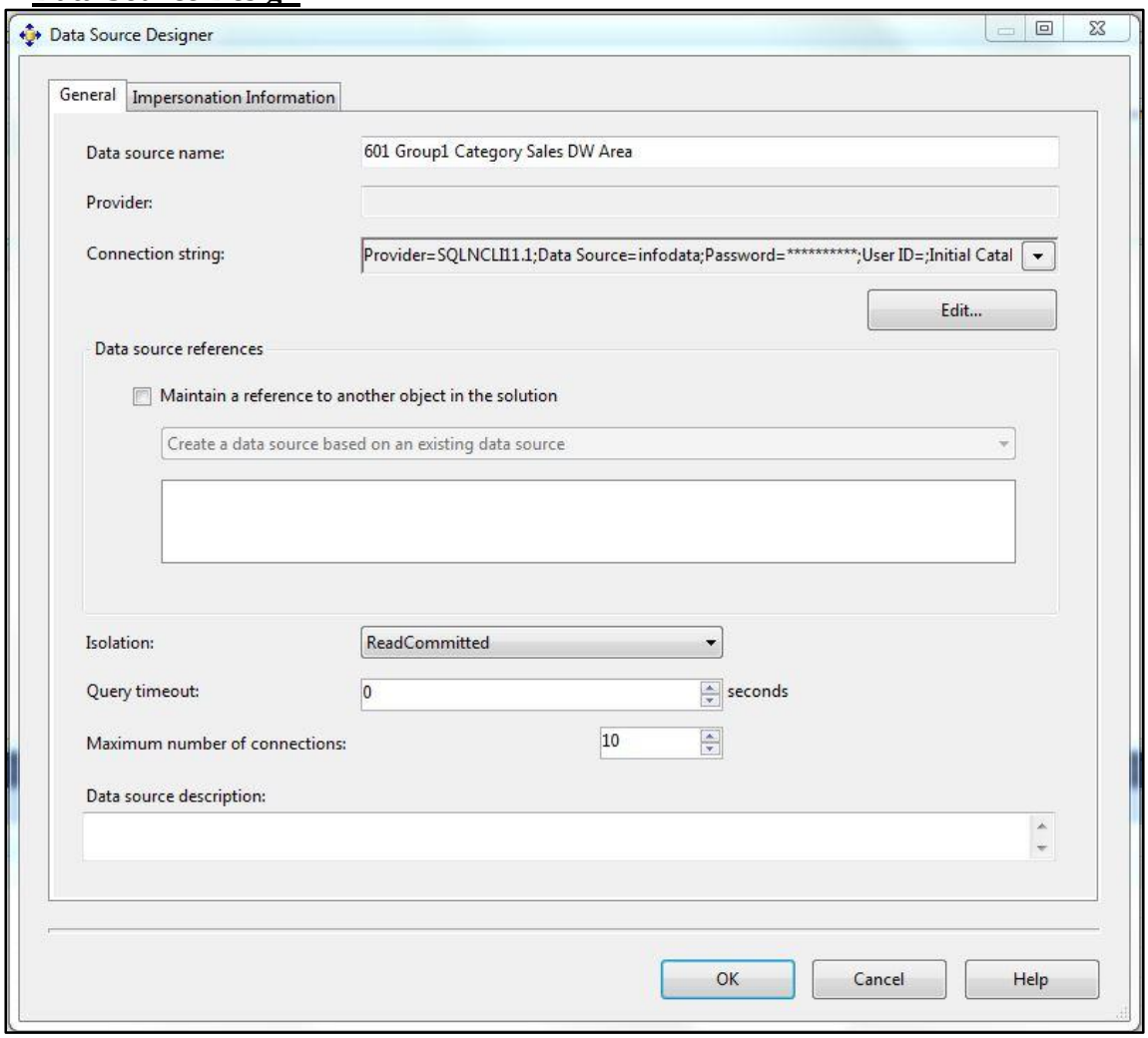
Based on the graphs that we created using SSRS, we can conclude that the sales of beer during the thanksgiving week see an increasing trend. This graph could then be used to compare with the sales trend of beer during other weeks and then an appropriate sales strategy during the thanksgiving week could be used by the stakeholders to improve business. Dominic Finer Foods can then use this data to also use various promotions with other products that do not have much sales to help the business overall. A similar approach could be used on other products during different holiday season to improve business. Hence, data warehousing along with business intelligence can be used to effectively answer business questions.

3. Cube from SSAS and Report from SSRS on top of SSAS for Question 2

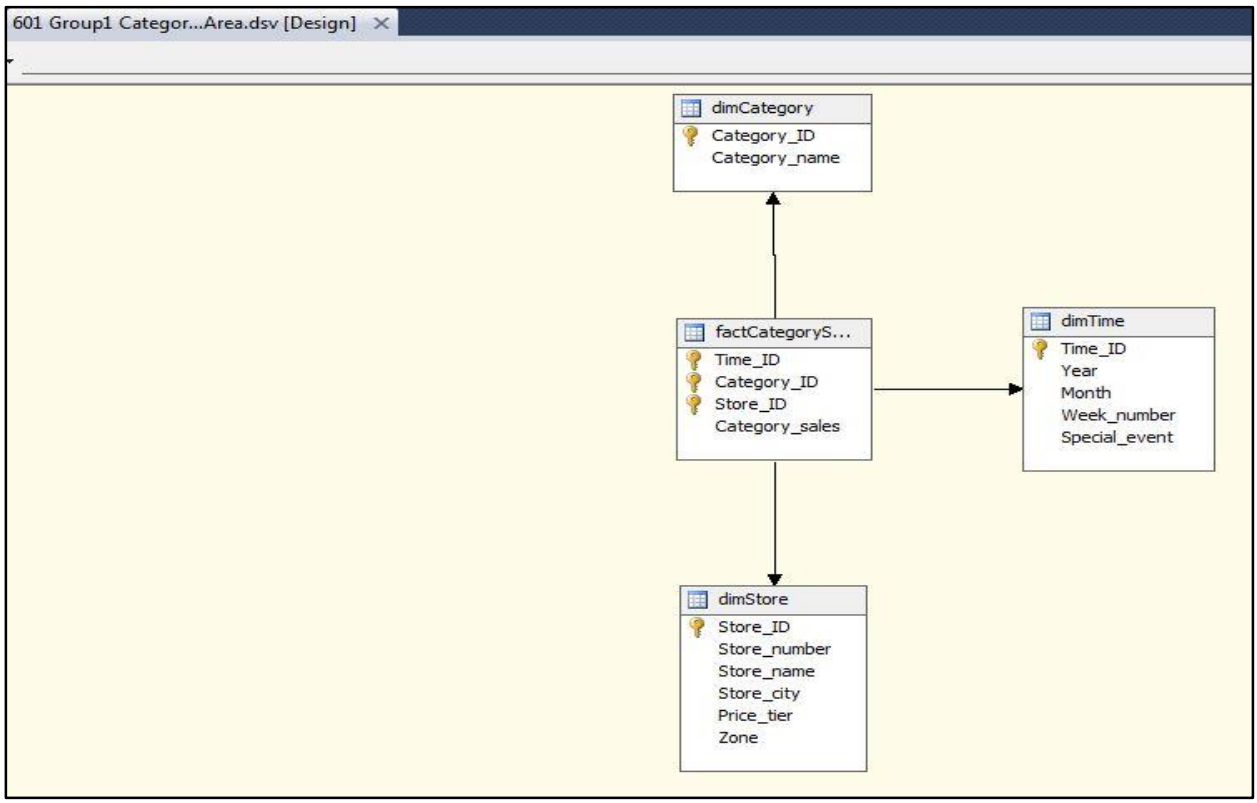
Based on the reporting plan discussed in the report earlier, we are building cube from SSAS and then using SSRS to create reports for question 2.

Question 2: How are the average price and sales of a particular product changing according to different zones (Fish and Fish Coupon)?

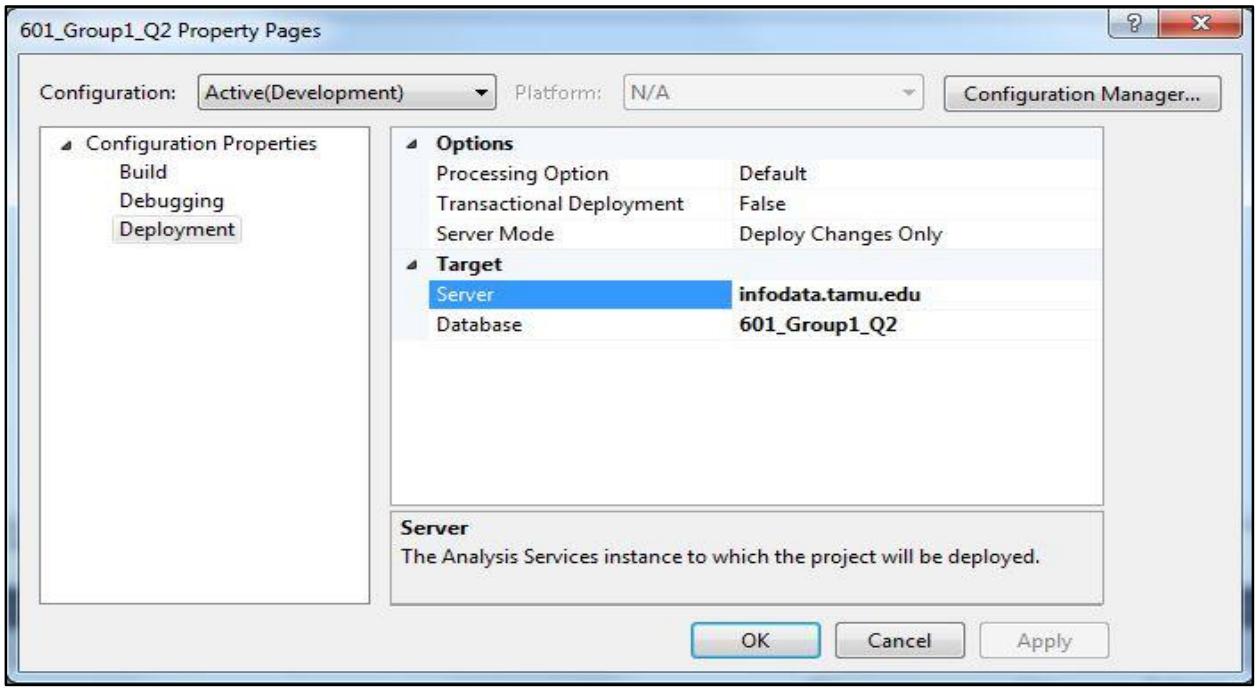
Data Source Design



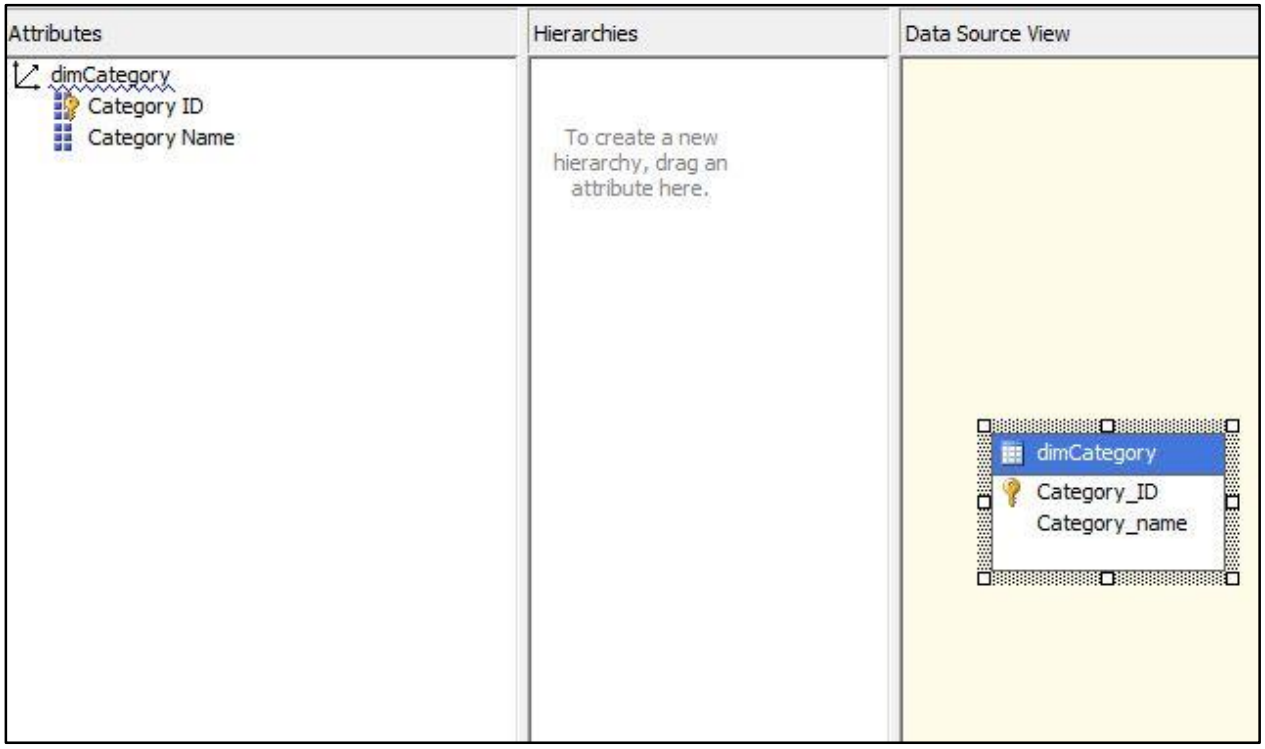
Cube Creation:



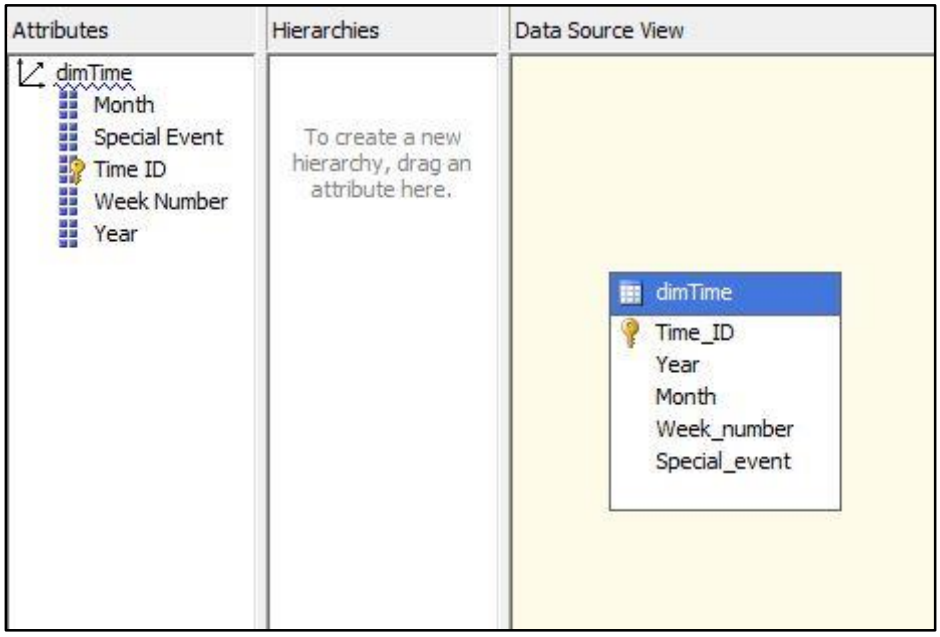
Multidimensional database property:



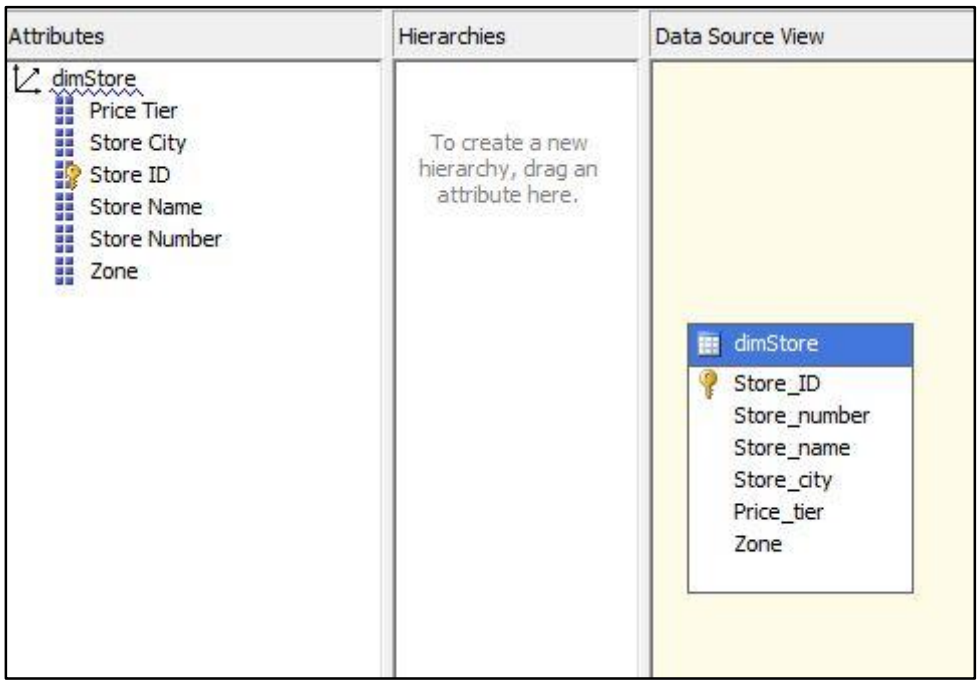
Category Dimension



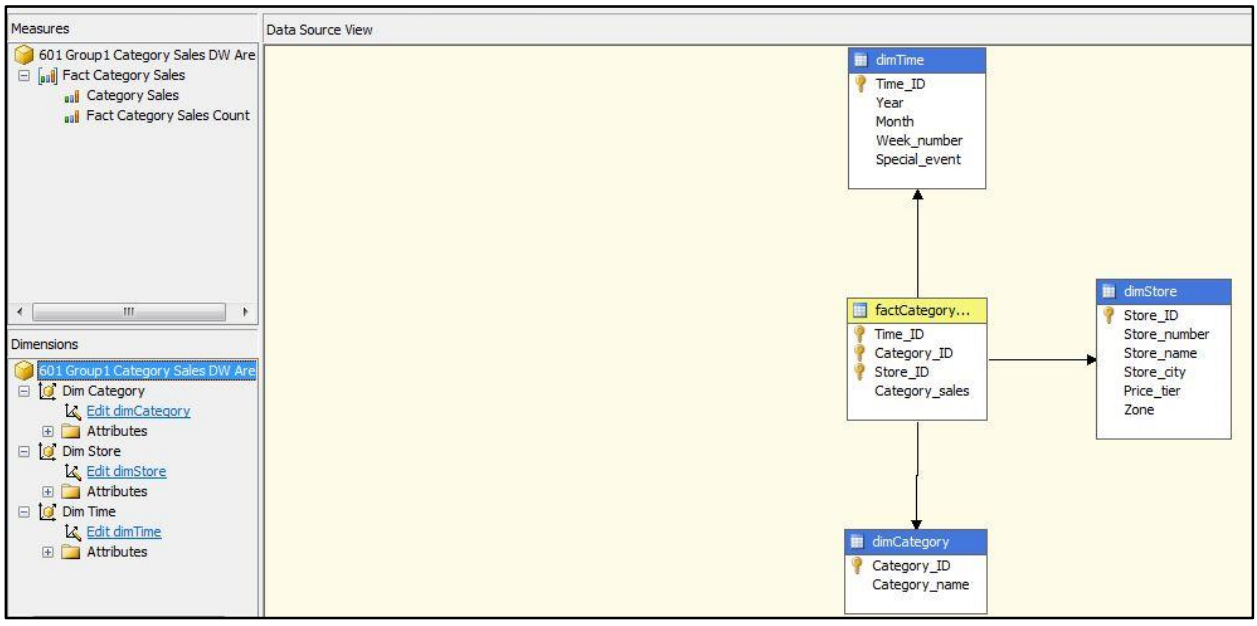
Time Dimension



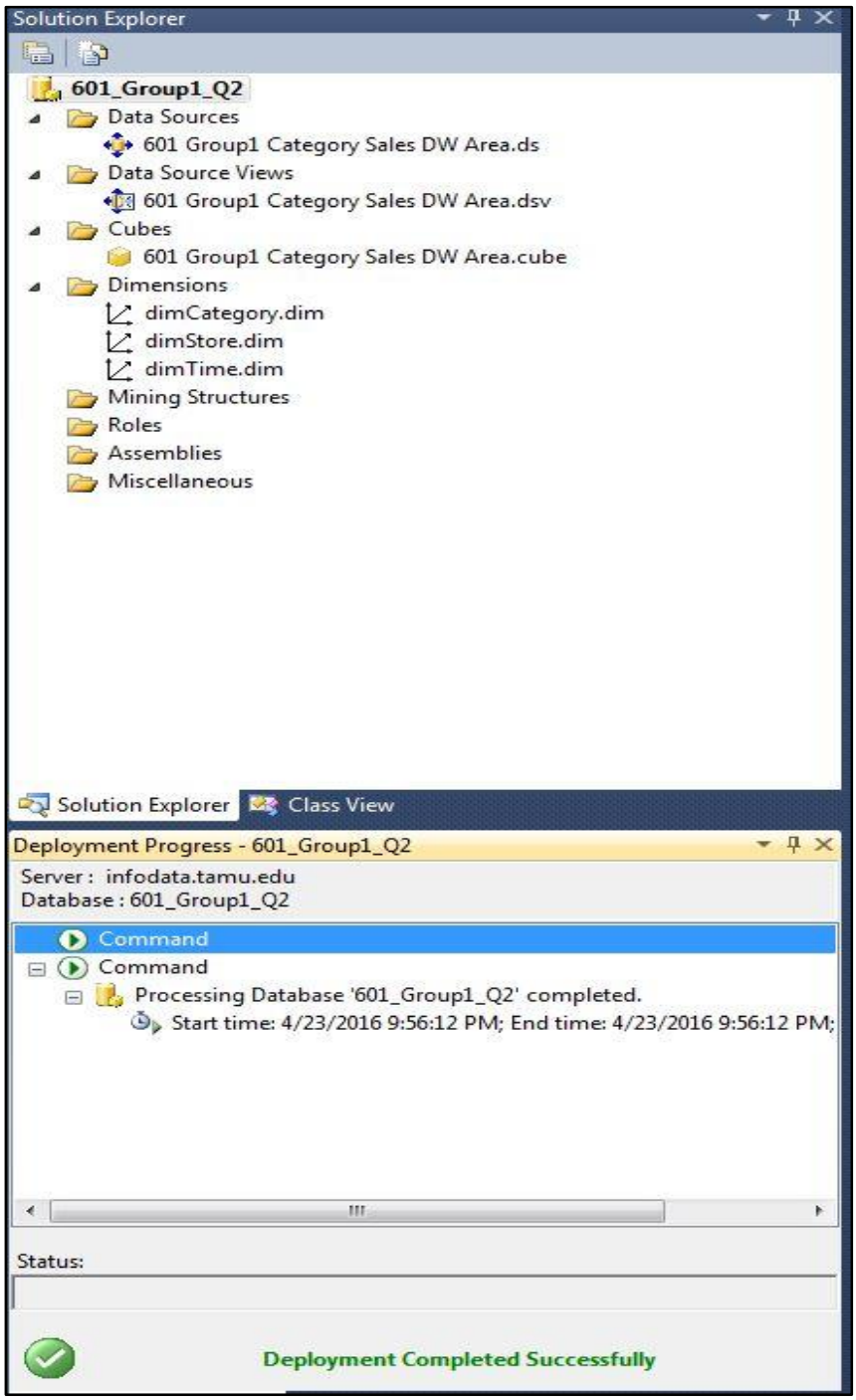
Store Dimension



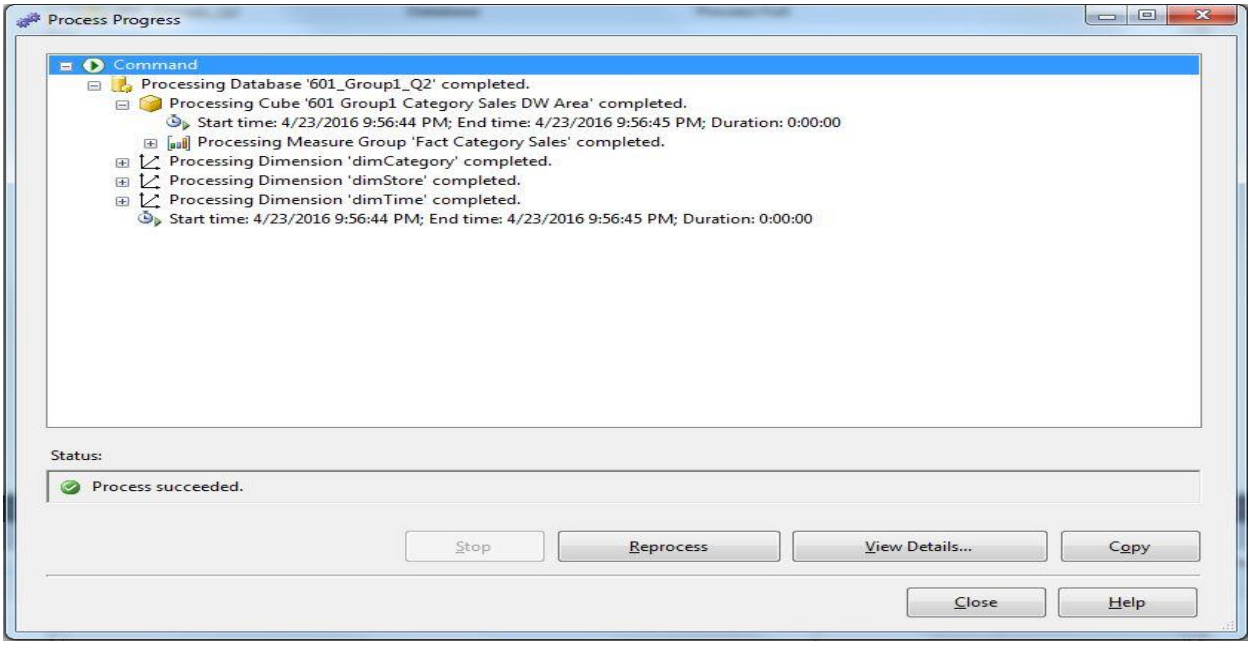
Cube Structure



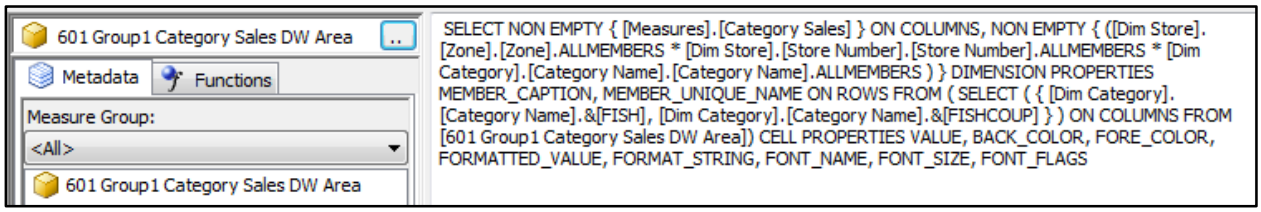
Successful Deployment



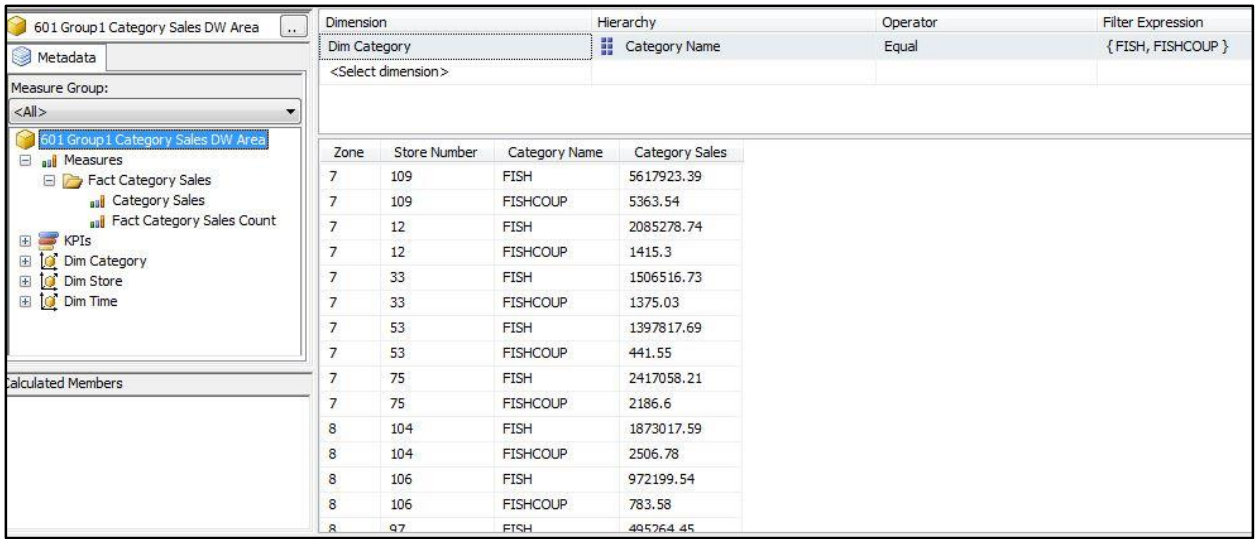
Successful Process



Cube Browser – Query



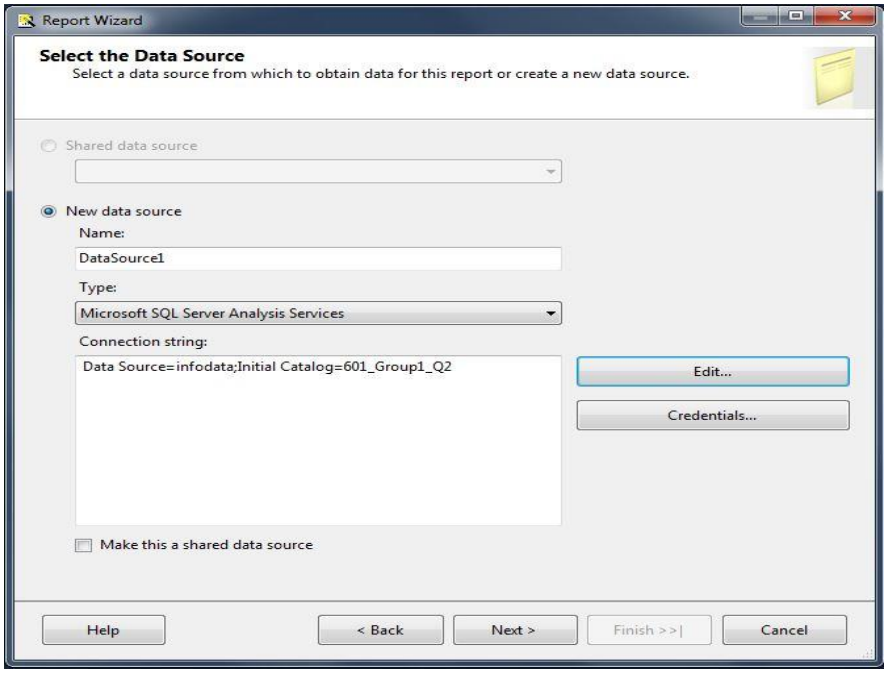
Cube Browser



4. SSRS on top of SSAS for Question 2

After successful creation of cube using SSAS, we can now use SSRS to create the reports for users. SSAS cube will be used as the data source, hence creating the report using the SSRS on top of SSAS for Question 2.

SSRS Data Source



SSRS Query Design

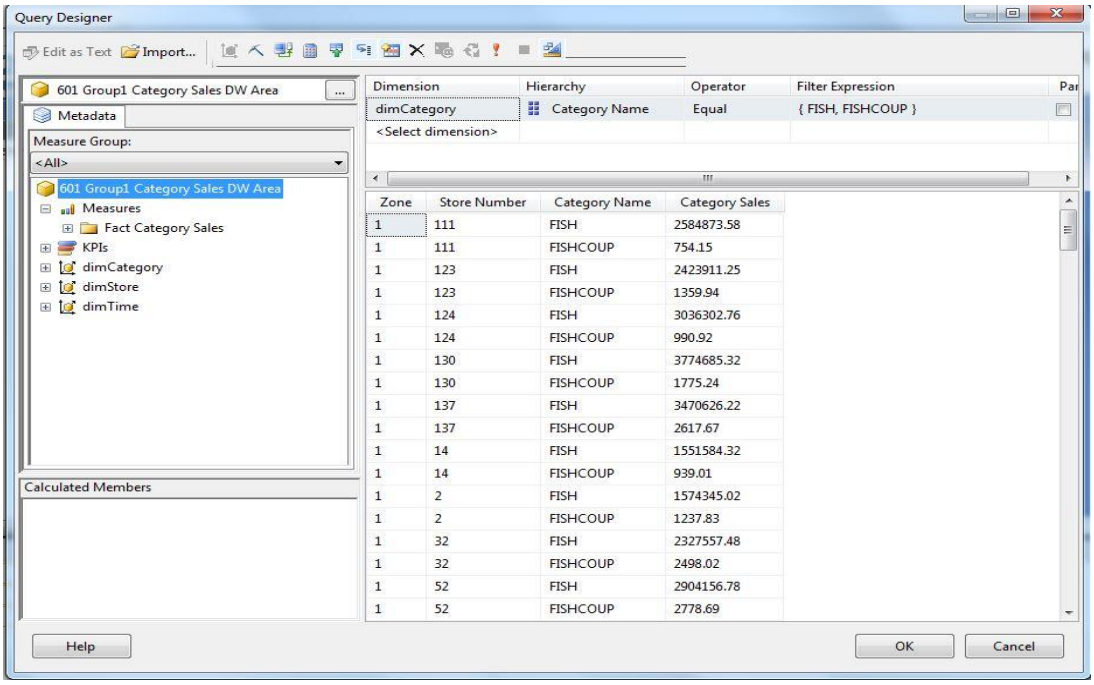


Table Design

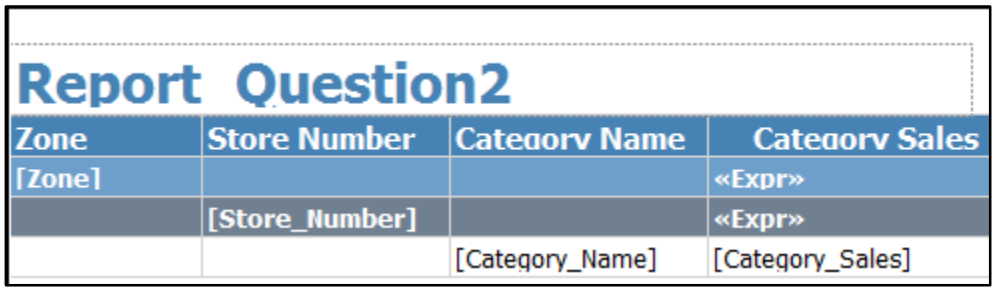
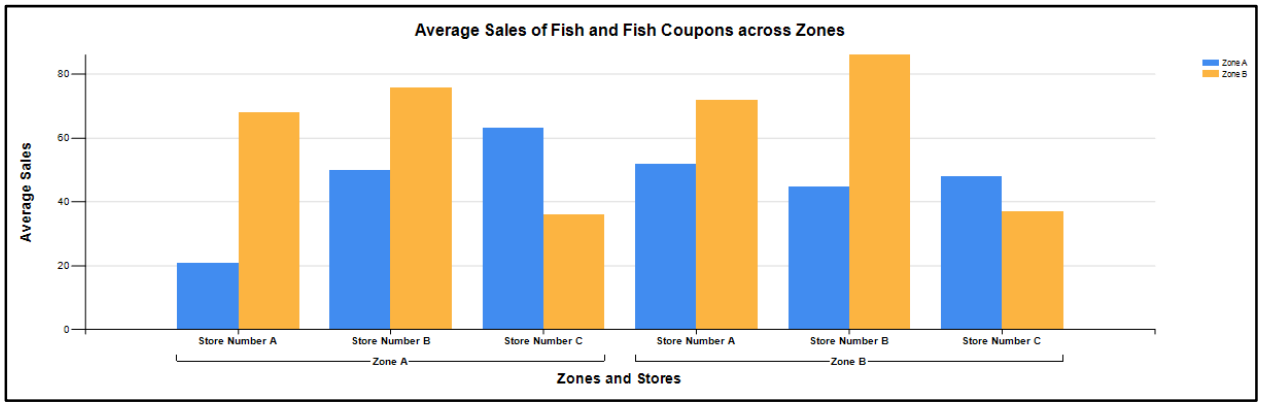
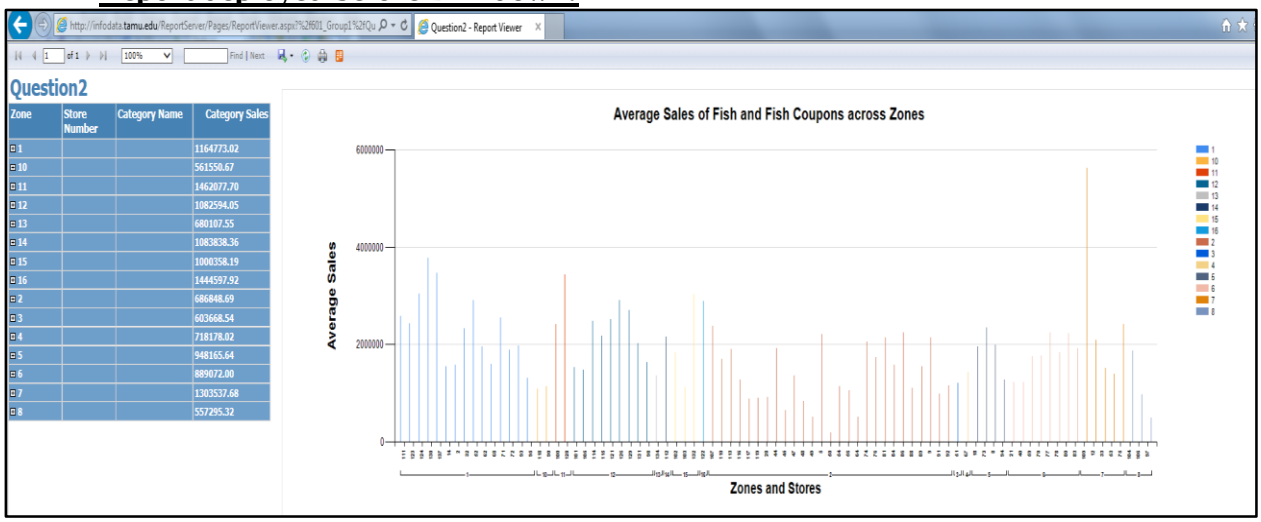


Chart Design



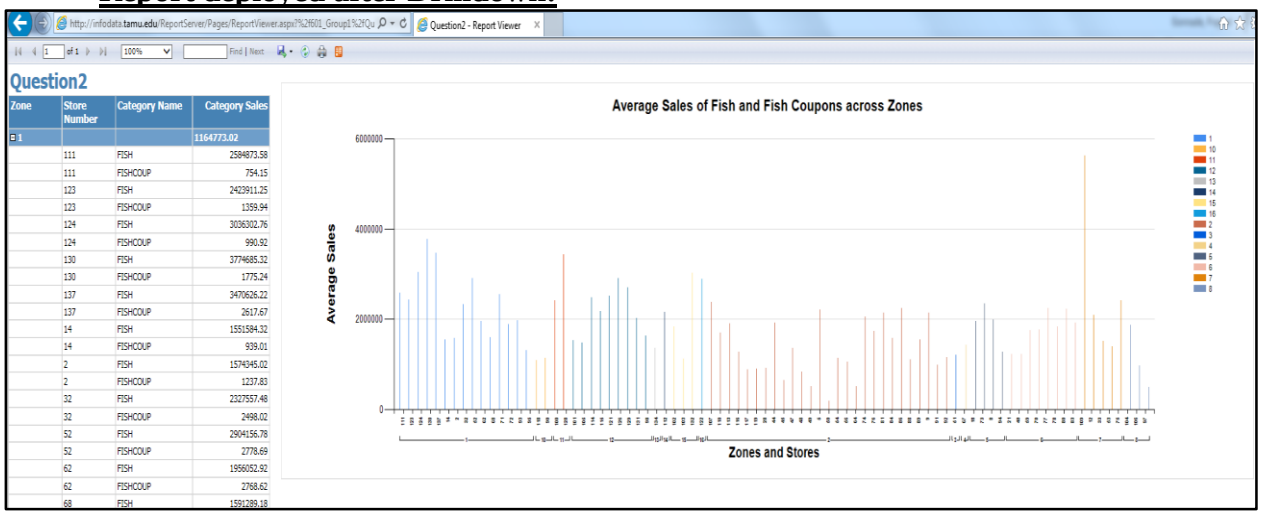
Deployment of Report on infodata.tamu.edu

Report deployed before Drilldown:



Zone	Store Number	Category Name	Category Sales
1			1164773.02
10			561550.67
11			1462077.70
12			1082594.05
13			680107.55
14			1083838.36
15			1000358.19
16			1444597.92
2			686848.69
3			603668.54
4			718178.02
5			948165.64
6			889072.00
7			1303537.68
8			557295.32

Report deployed after Drilldown:



Zone	Store Number	Category Name	Category Sales
1			1164773.02
	111	FISH	2584873.58
	111	FISHCOUP	754.15
	123	FISH	2423911.25
	123	FISHCOUP	1359.94
	124	FISH	3036302.76
	124	FISHCOUP	990.92
	130	FISH	3774685.32
	130	FISHCOUP	1775.24
	137	FISH	3470626.22
	137	FISHCOUP	2617.67
	14	FISH	1551584.32
	14	FISHCOUP	939.01
	2	FISH	1574345.02
	2	FISHCOUP	1237.83
	32	FISH	2327557.48
	32	FISHCOUP	2498.02
	52	FISH	2904156.78
	52	FISHCOUP	2778.69
	62	FISH	1956052.92
	62	FISHCOUP	2768.62
	68	FISH	1591289.18

Conclusion:

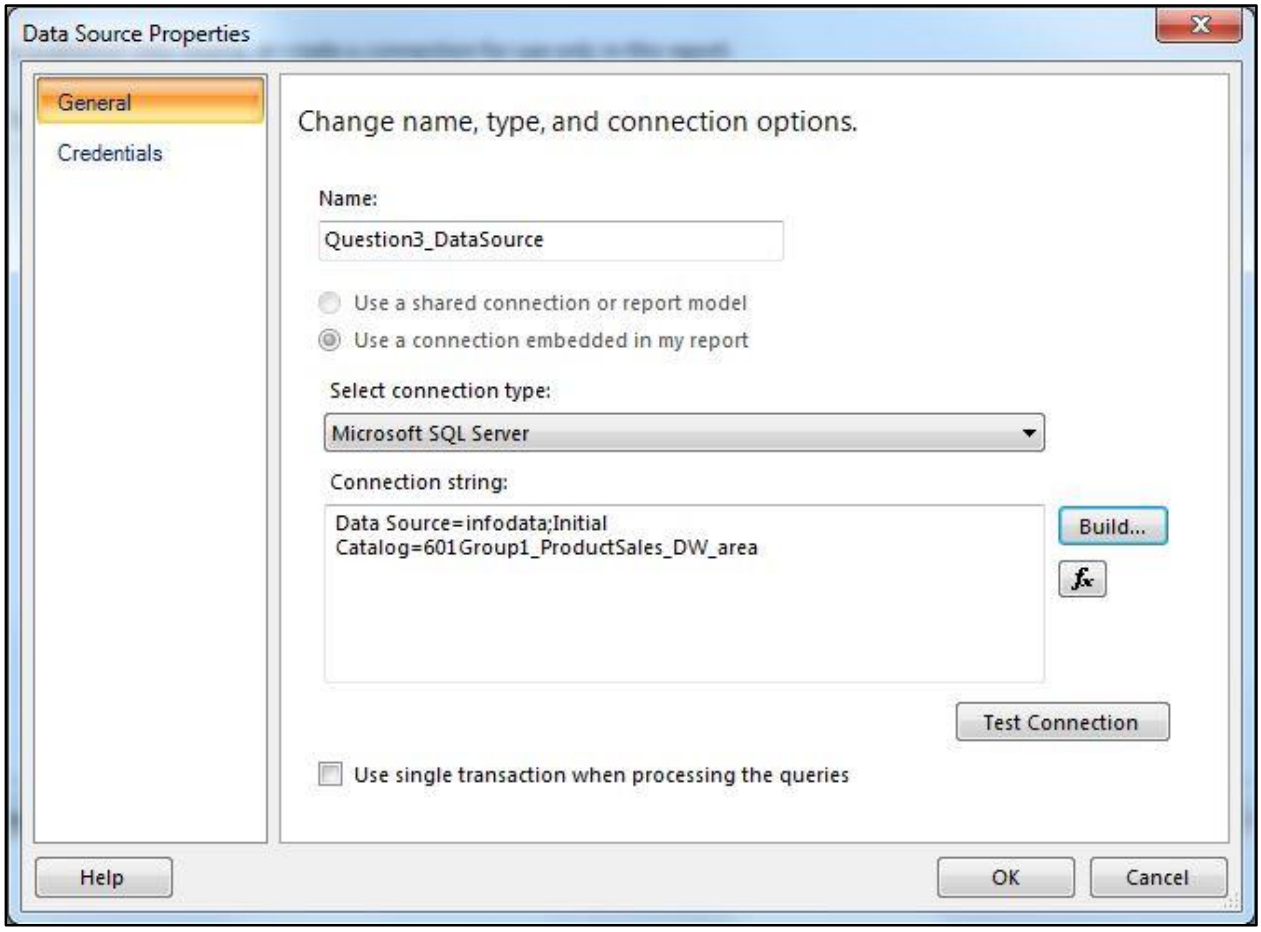
Based on the graphs that we created using SSRS on top of SSAS, we can conclude average sales of fish and fish coupons are differing in different zones. Dominic Finer Foods can then use this graph to compare sales across zones and create strategies to improve sales in zones with less sales and sustain sales in areas with more sales. Also, sales trend across zones would help the business to store the product in a better way so as to help business gain profits. A similar approach could be used on all the products across the zones so as to help the business gain profits, improve business and provide promotions to the customer. SSRS on top of SSAS helps to answer specific business question that a stakeholder of the business would have and then provides visualization accordingly to answer the business question.

5. Reports using ReportBuilder3.0 for Question 3

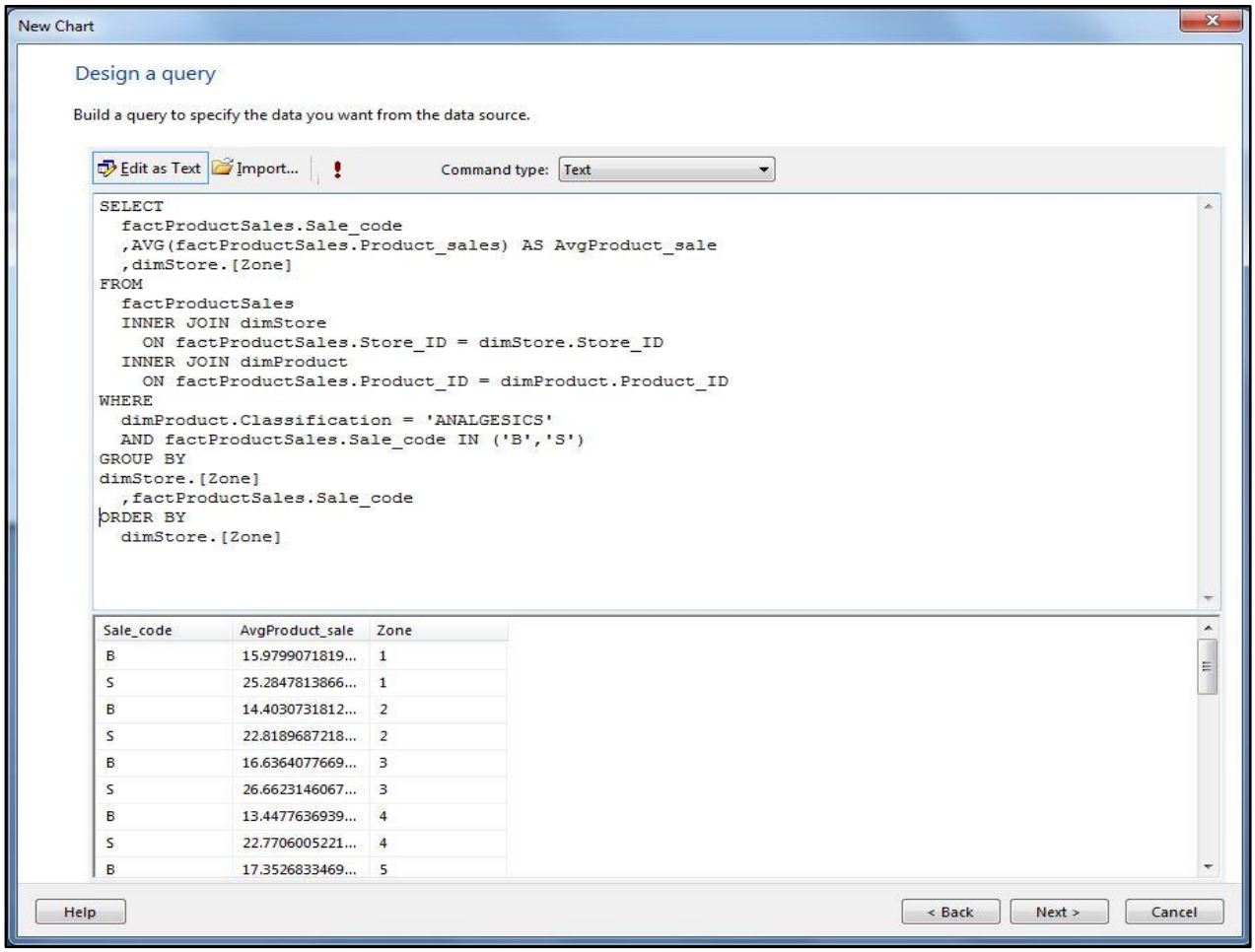
Based on the reporting plan discussed in the report earlier, we are creating reports for Question 3 using ReportBuilder3.0.

Question 3: Compare the effect of Bonus buy and Price Reduction in Analgesics in different zones.

Data Source Design:



Query Design



Arrange Chart fields

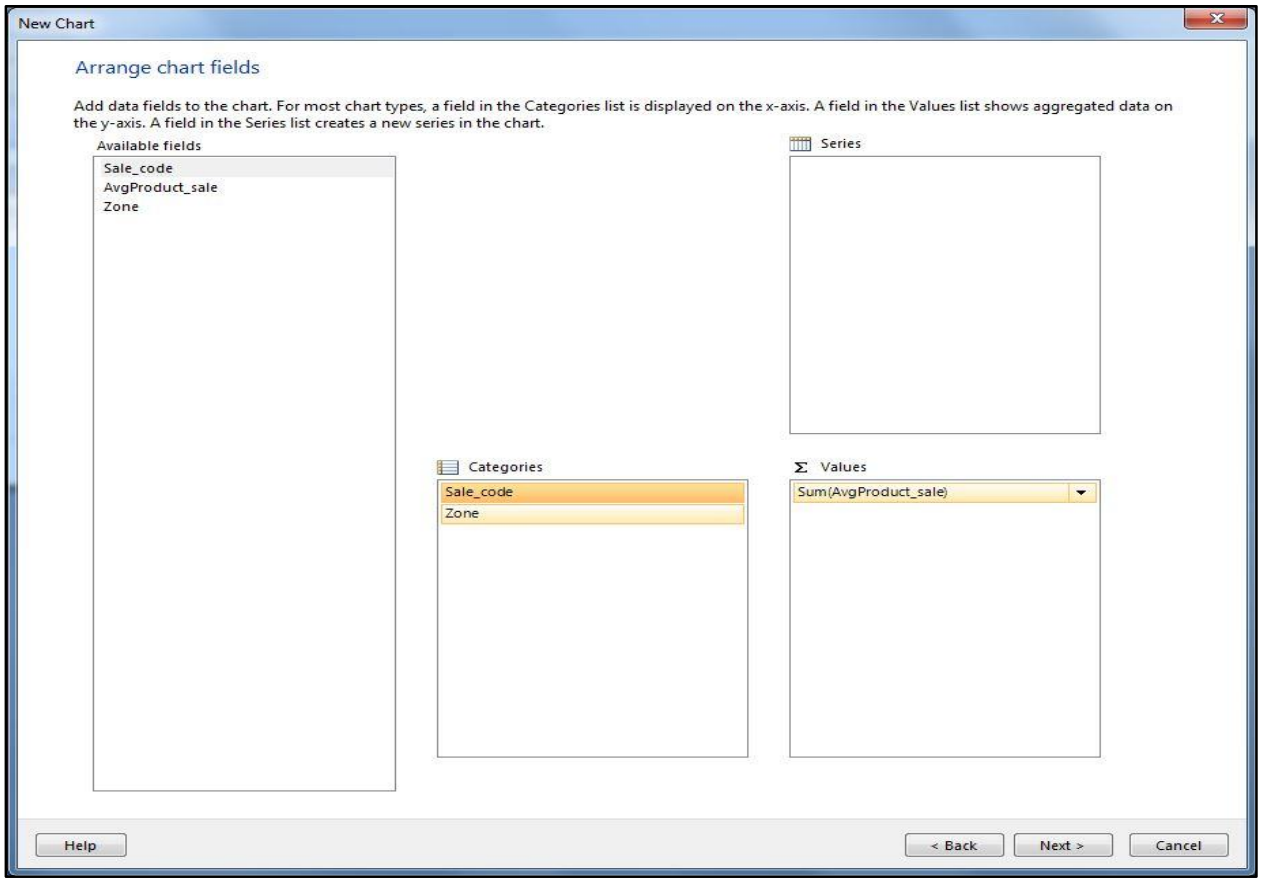
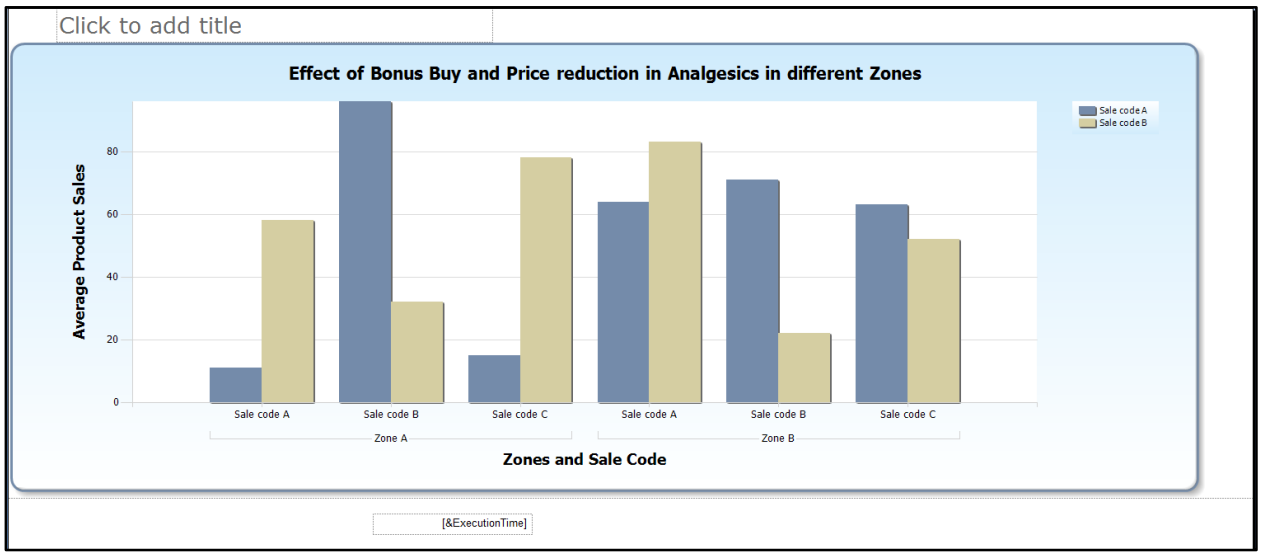
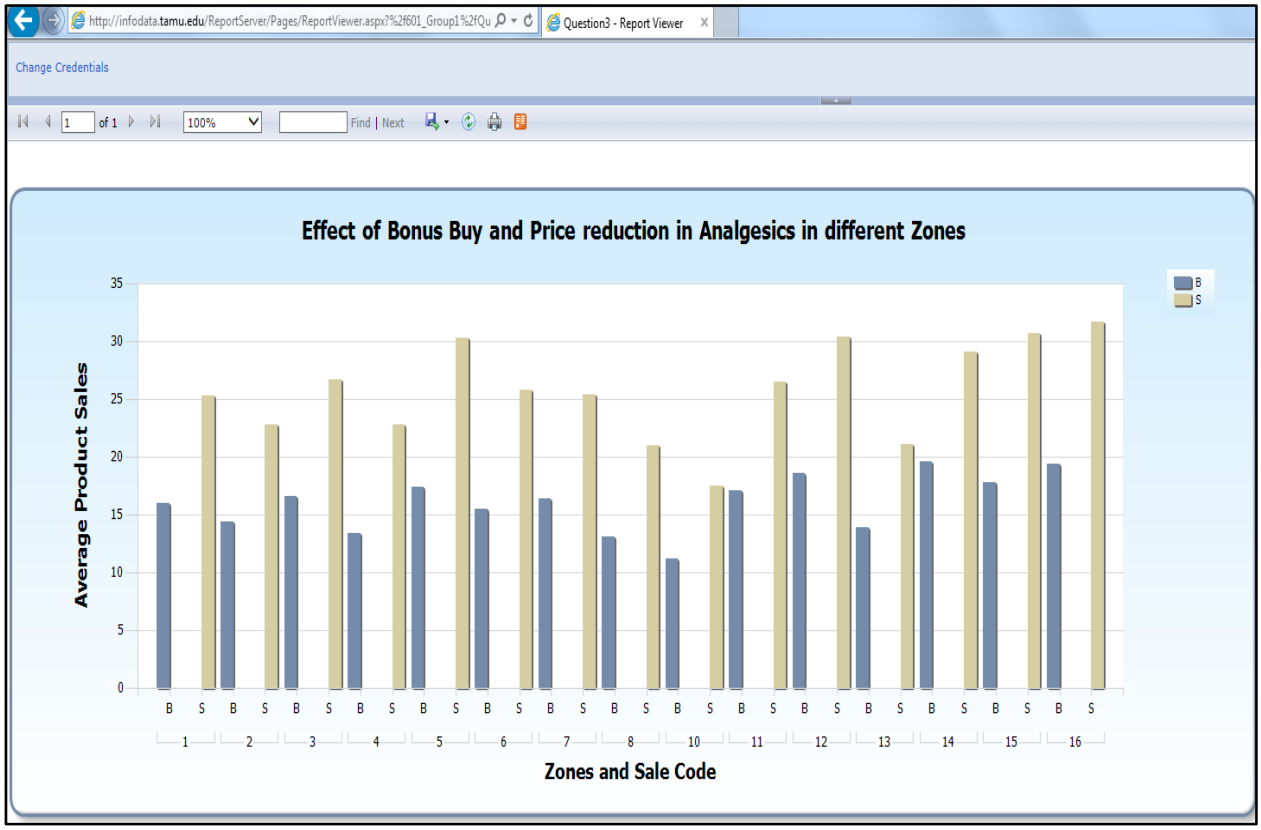


Chart Design



Report deployed on infodata.tamu.edu

Report Deployed



Conclusion:

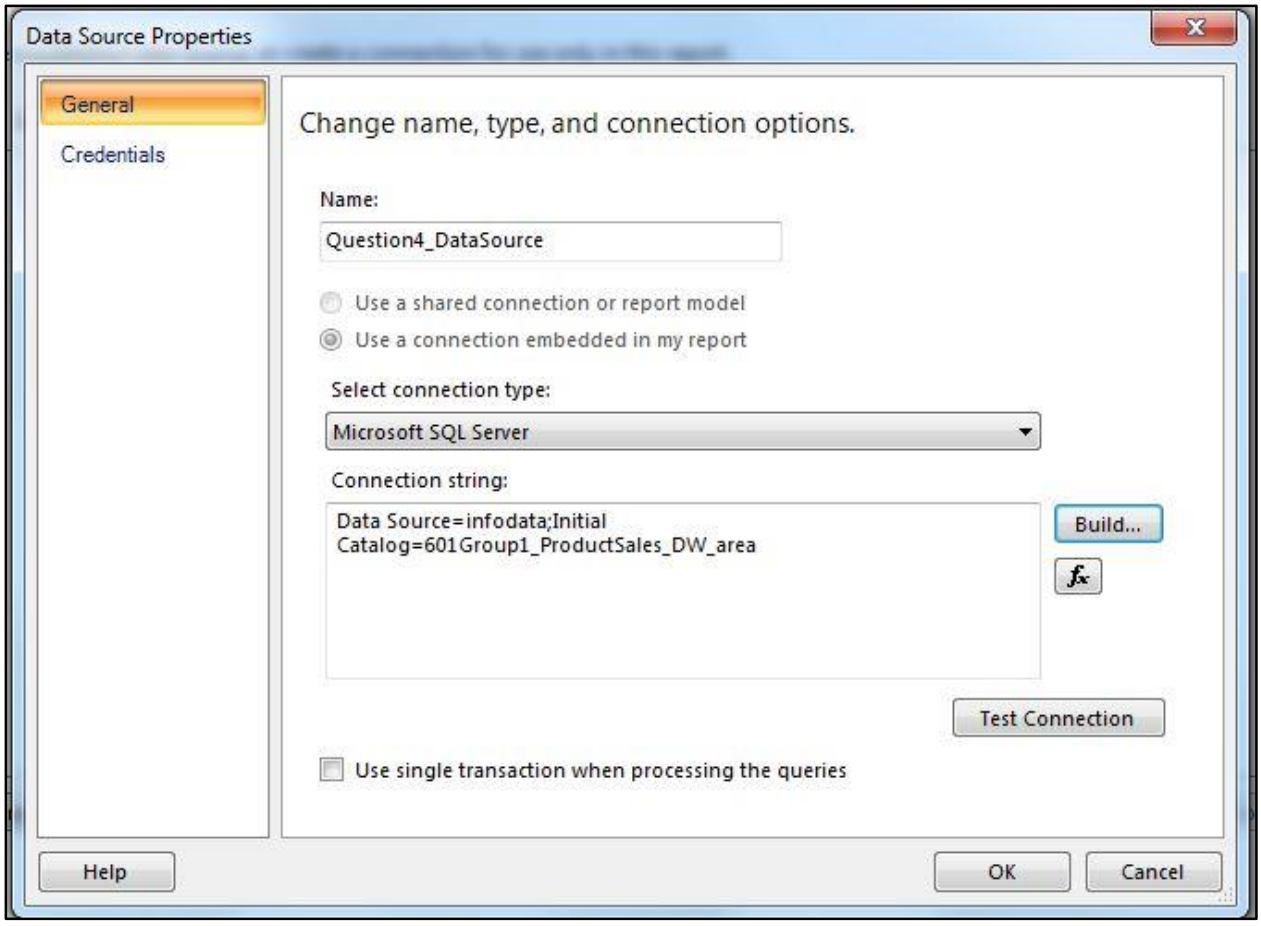
We answered the business questions "Compare the effect of Bonus buy and Price Reduction in Analgesics in different zones." using the ReportBuilder3.0. As we can see in the graph, price reduction helps improve sales more than the bonus buy in almost all the zones. Hence, we can focus more on ways to reduce the price so as to help Dominick Finer Food with its sales. We can use the same method for different products and different promotions so as to help improve the strategies that the business would focus upon. Additionally, promotions which do not yield results for the organization could be scrapped or replaced and new promotions could be brought in using the sales data given above. Zone manager could also use this to improve sales in their zones. Thus reportbuilder3.0 was used effectively to create reports to help stakeholders answer business questions using effective visualizations.

6. Reports using ReportBuilder3.0 for Question 4

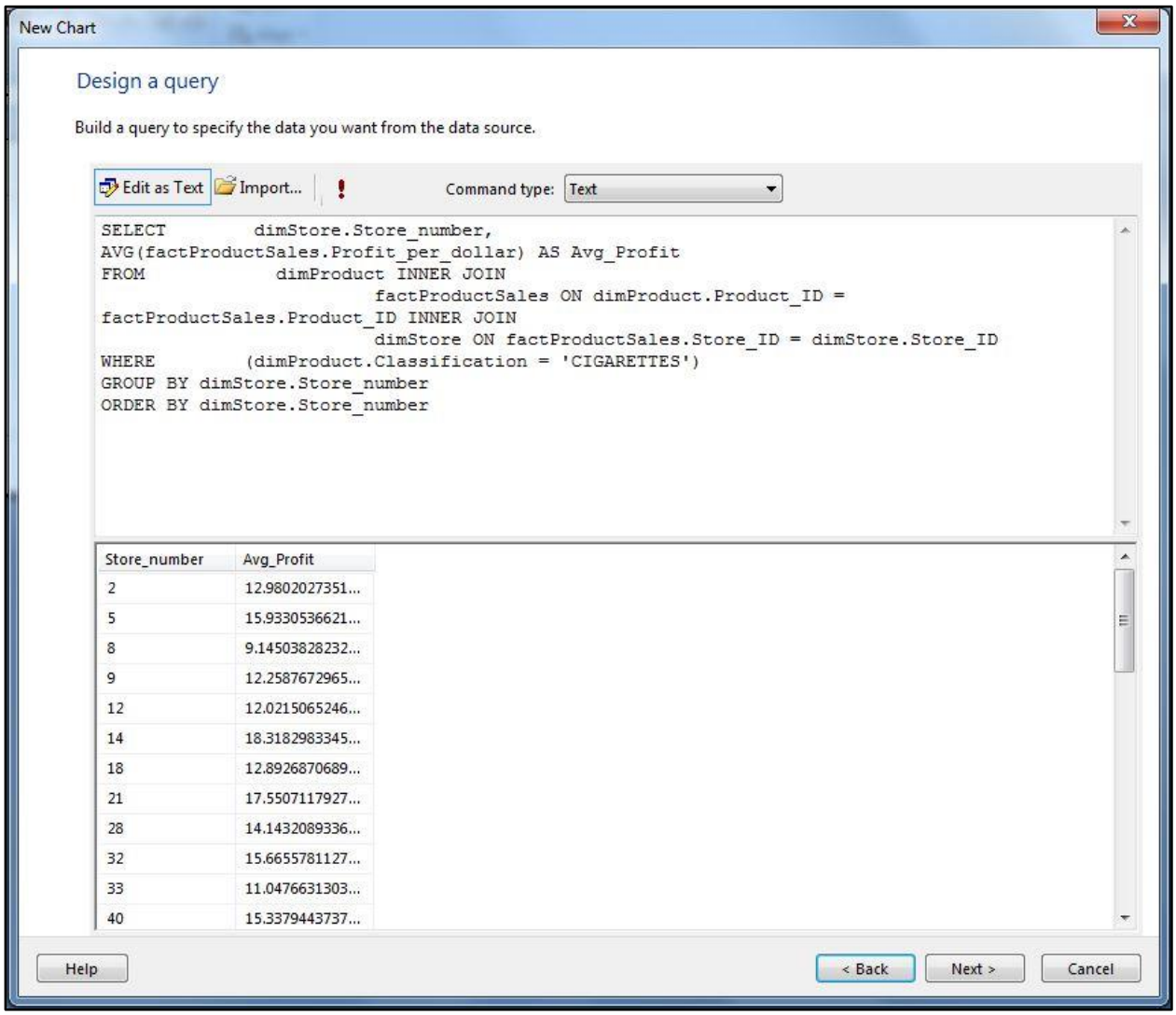
Based on the reporting plan discussed in the report earlier, we are creating reports for Question 4 using ReportBuilder3.0.

Question 4: Plot the average profit margin for cigarettes across all the stores. Determine the average of profit for the sales of cigarettes and the stores which are below the average.

Data Source Design



Query Design



Arrange Chart fields

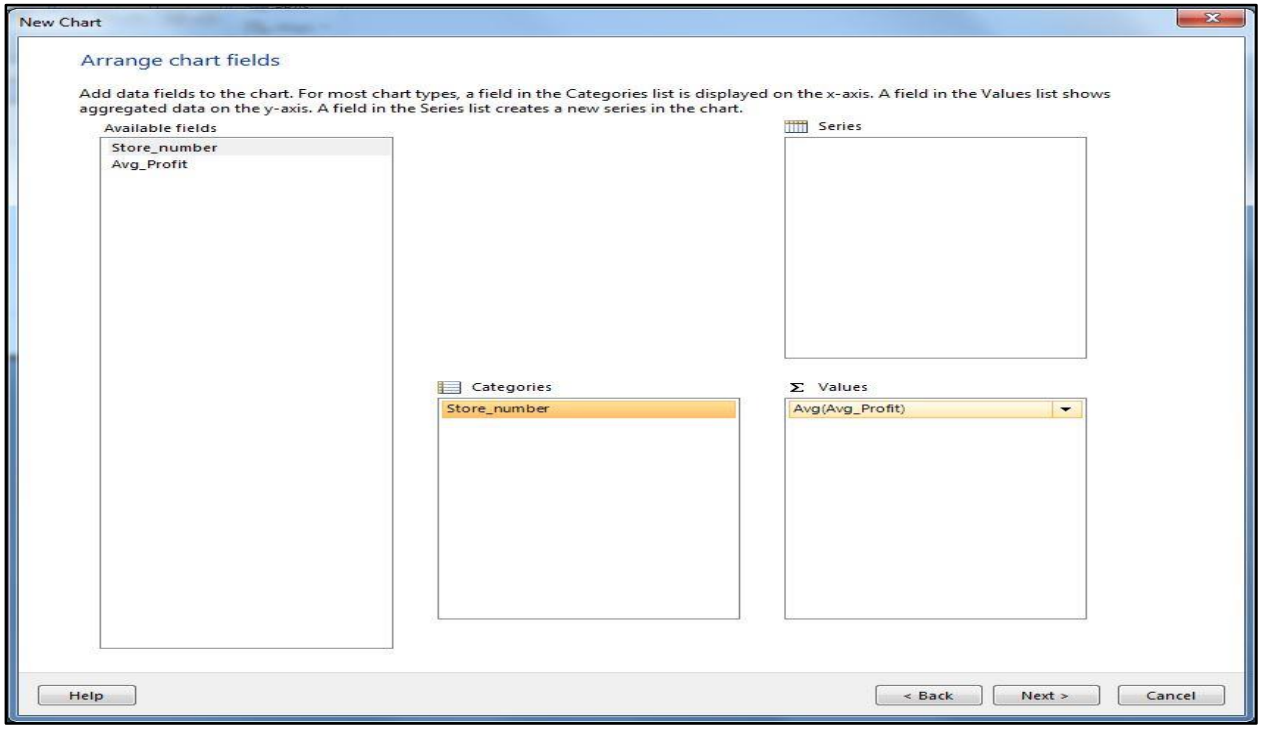
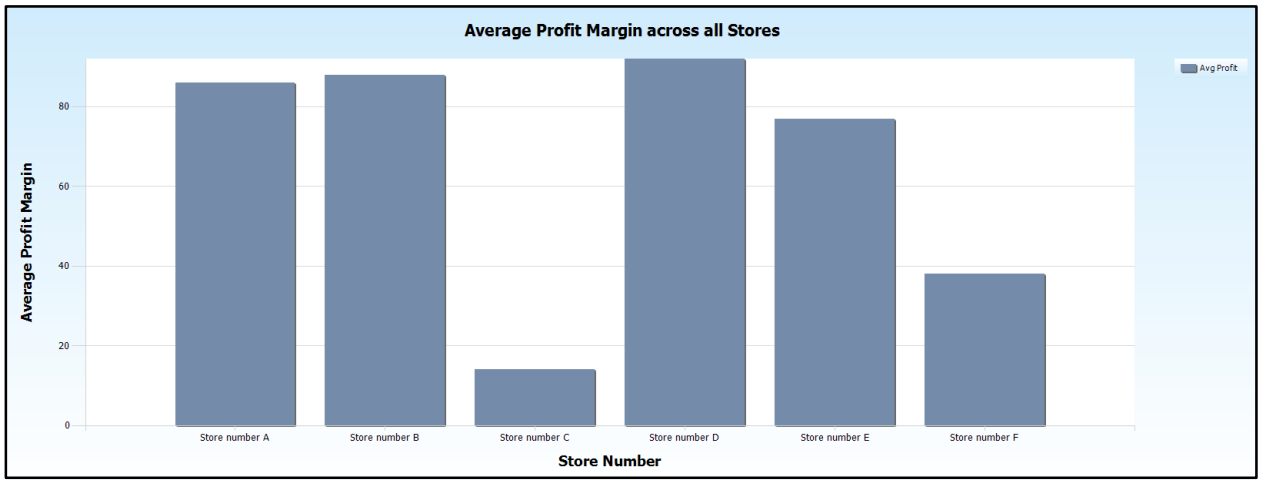
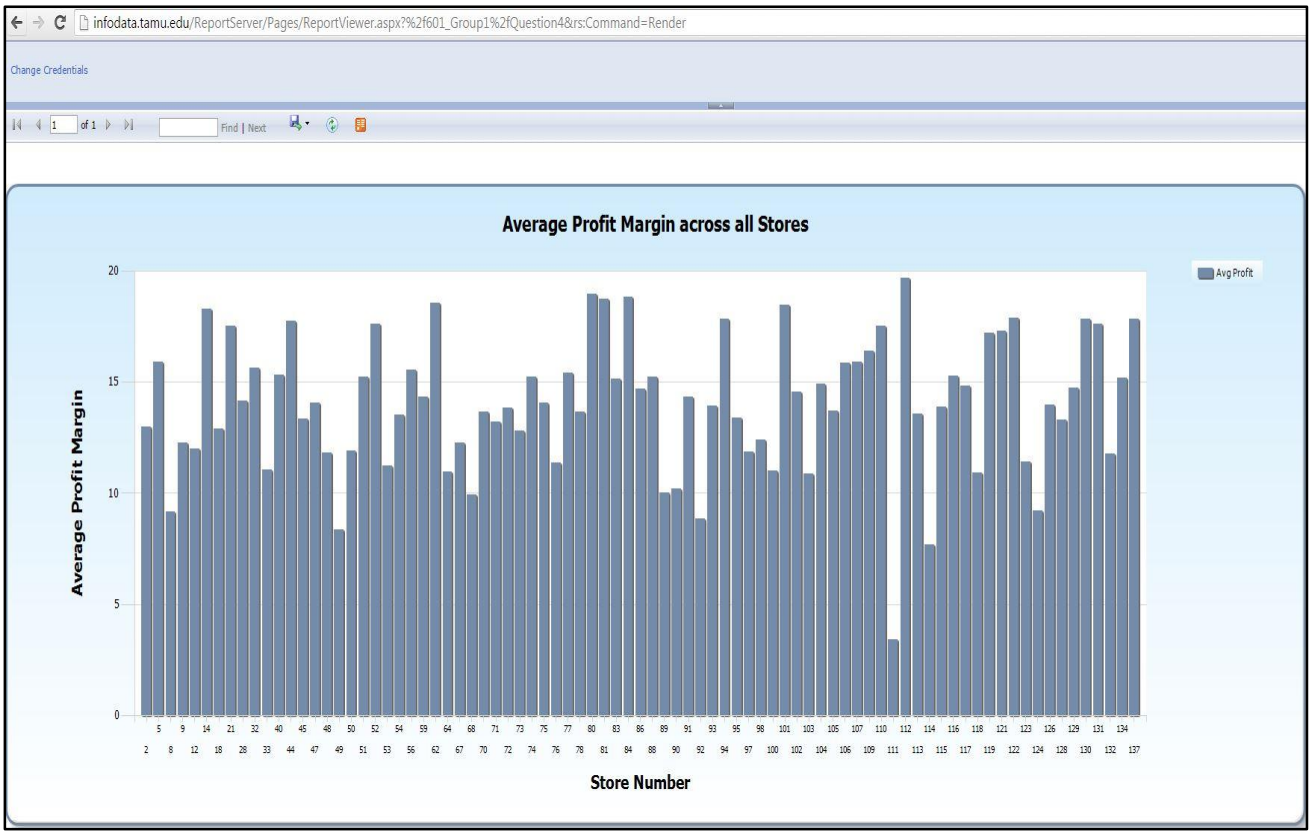


Chart Design



Report deployed on infodata.tamu.edu

Report Deployed



Conclusion:

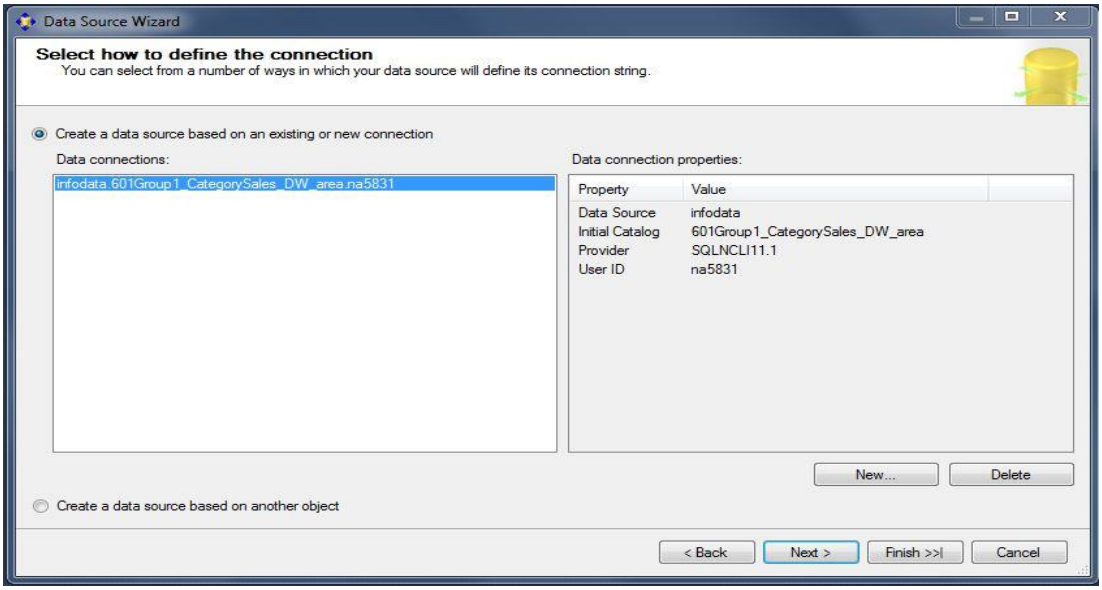
Using report builder, we created visualization to answer business question 4. We were trying to get the average sales of cigarettes across all the stores. This would help us to find stores which have lower than average sales and help the business to devise strategy to improve business in that area. Similar approach used on other products would boost the sales for that particular product. We can also observe that store 111 has the lowest of all sales and store 112 has the highest sales. The business can then focus on all these stats to create promotions and improve sales. The managers and supervisors can also use this data to compare the sales amongst other stores and devise strategies similar to stores with high profitability to achieve better results for Dominic Finer Foods.

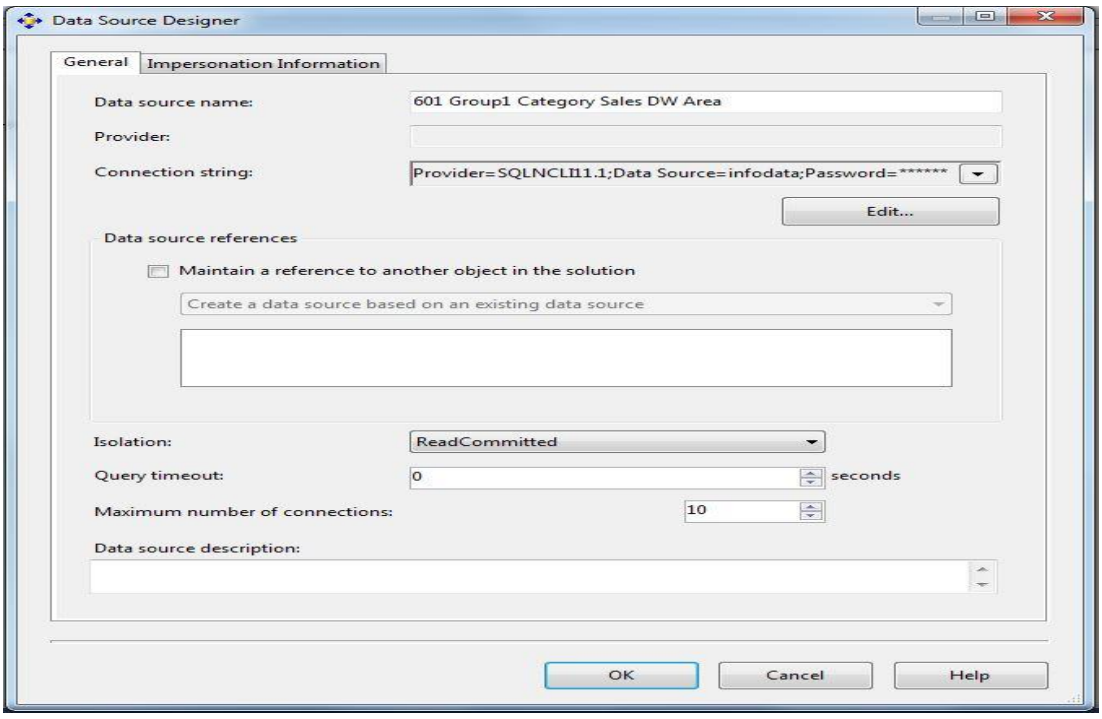
7. Cubes from SSAS for Question 5

Based on the reporting plan discussed in the report earlier, we are building cube from SSAS and then using pivot charts in Excel to create reports for question 5.

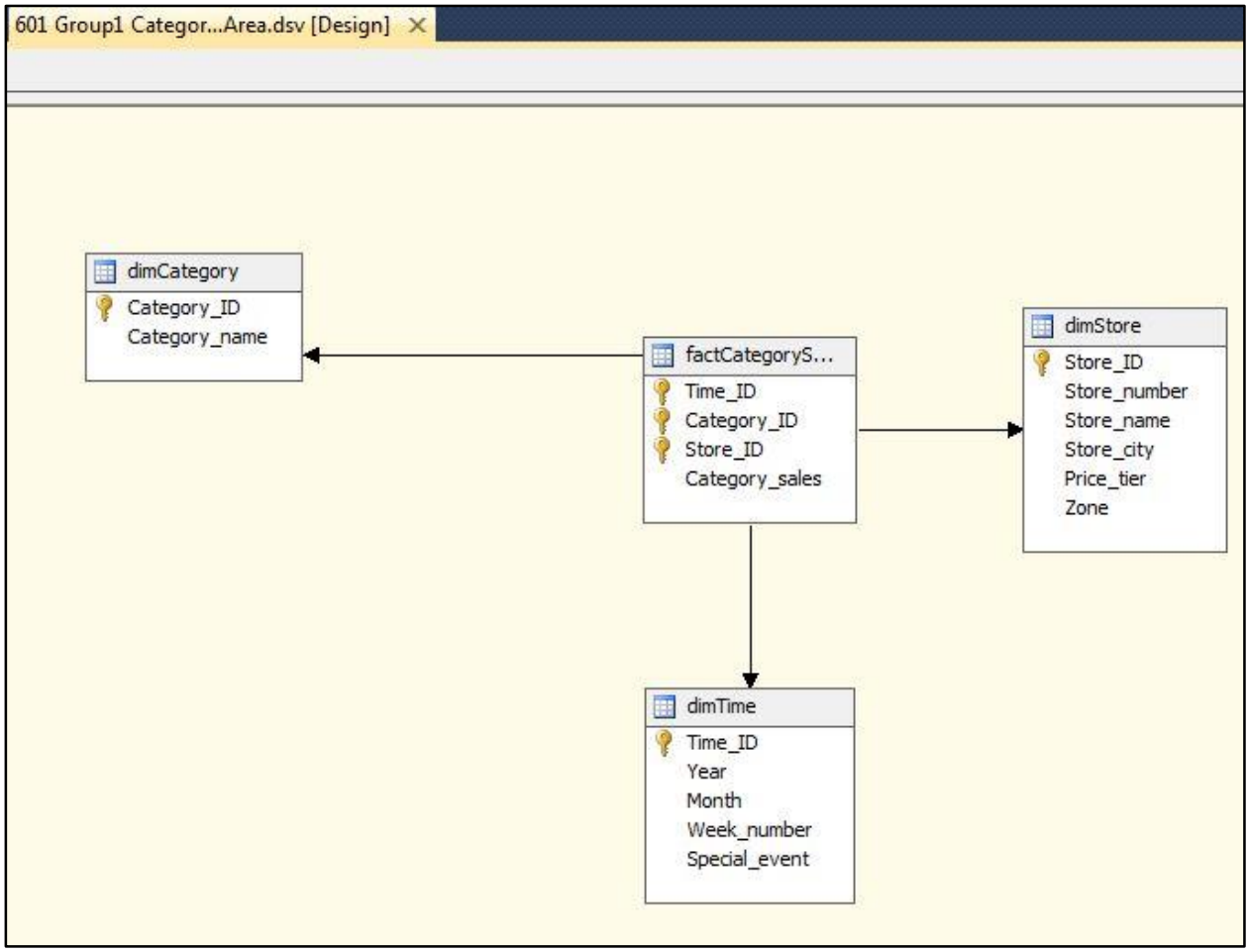
Question 5: What is the trend of Camera sales from the year 1990 to 1996?

Data Source Creation





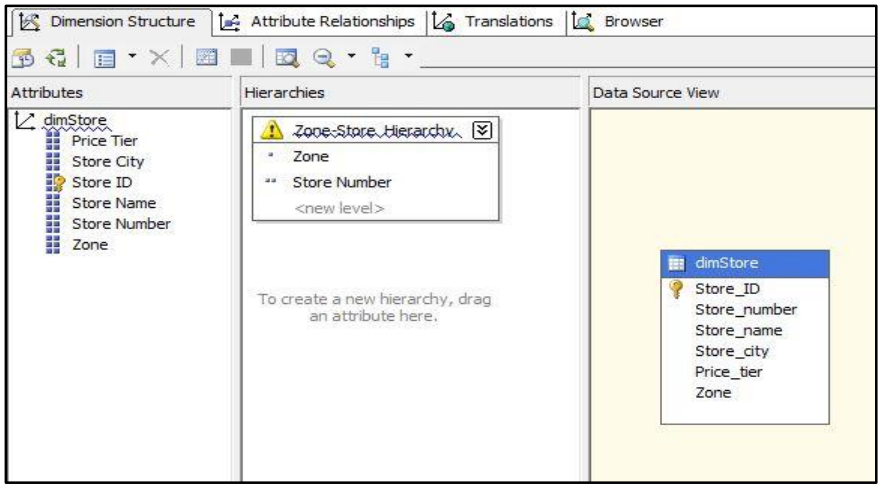
Data Source View



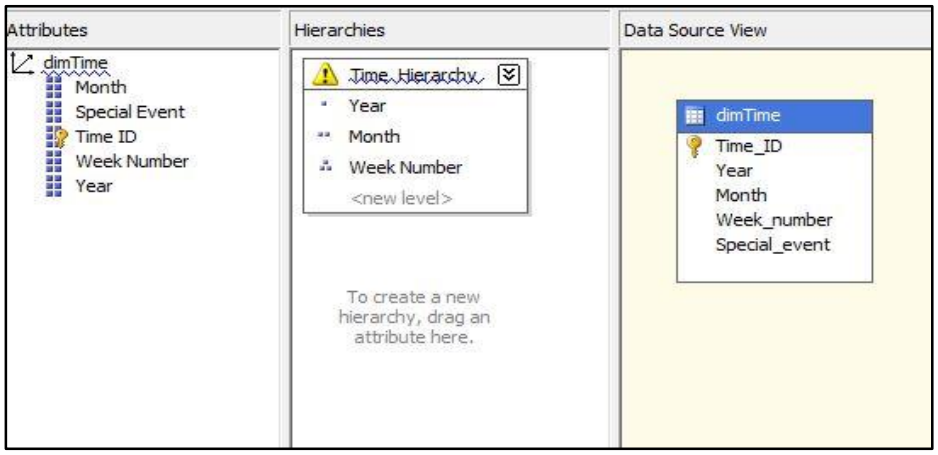
Dimensions

Category Dimension

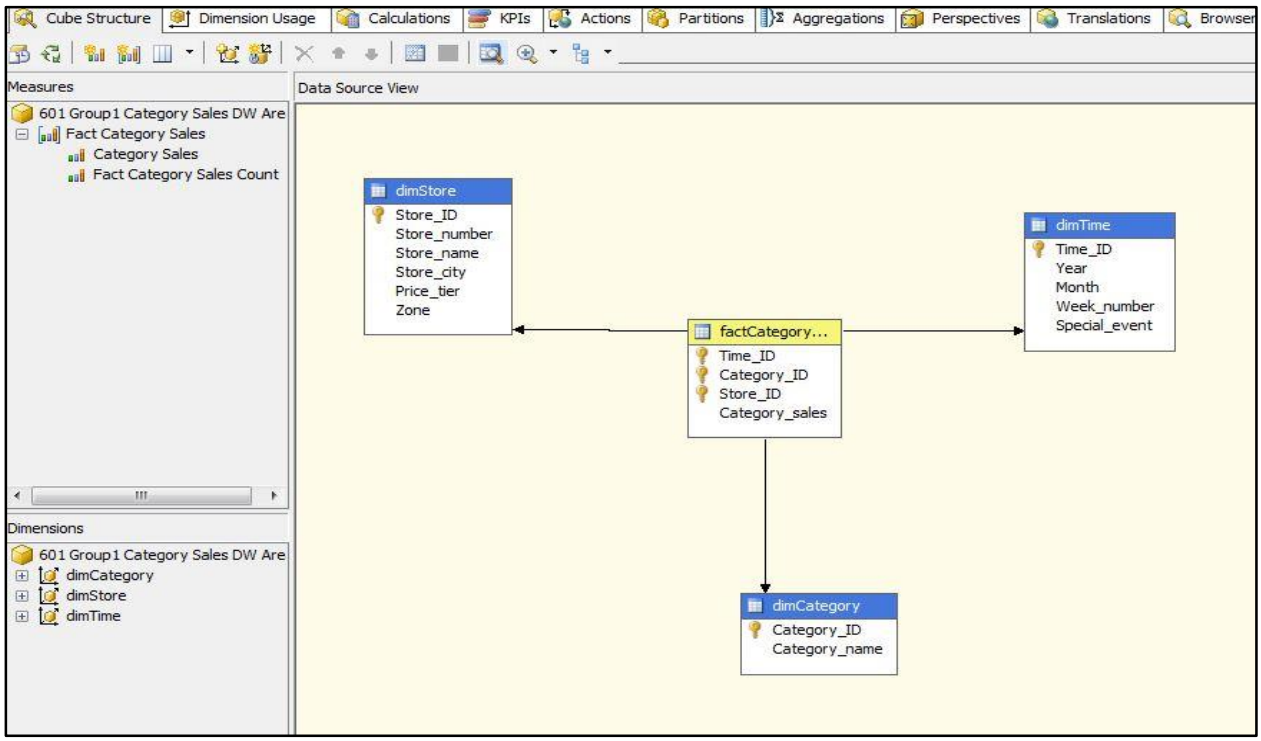
Store Dimension



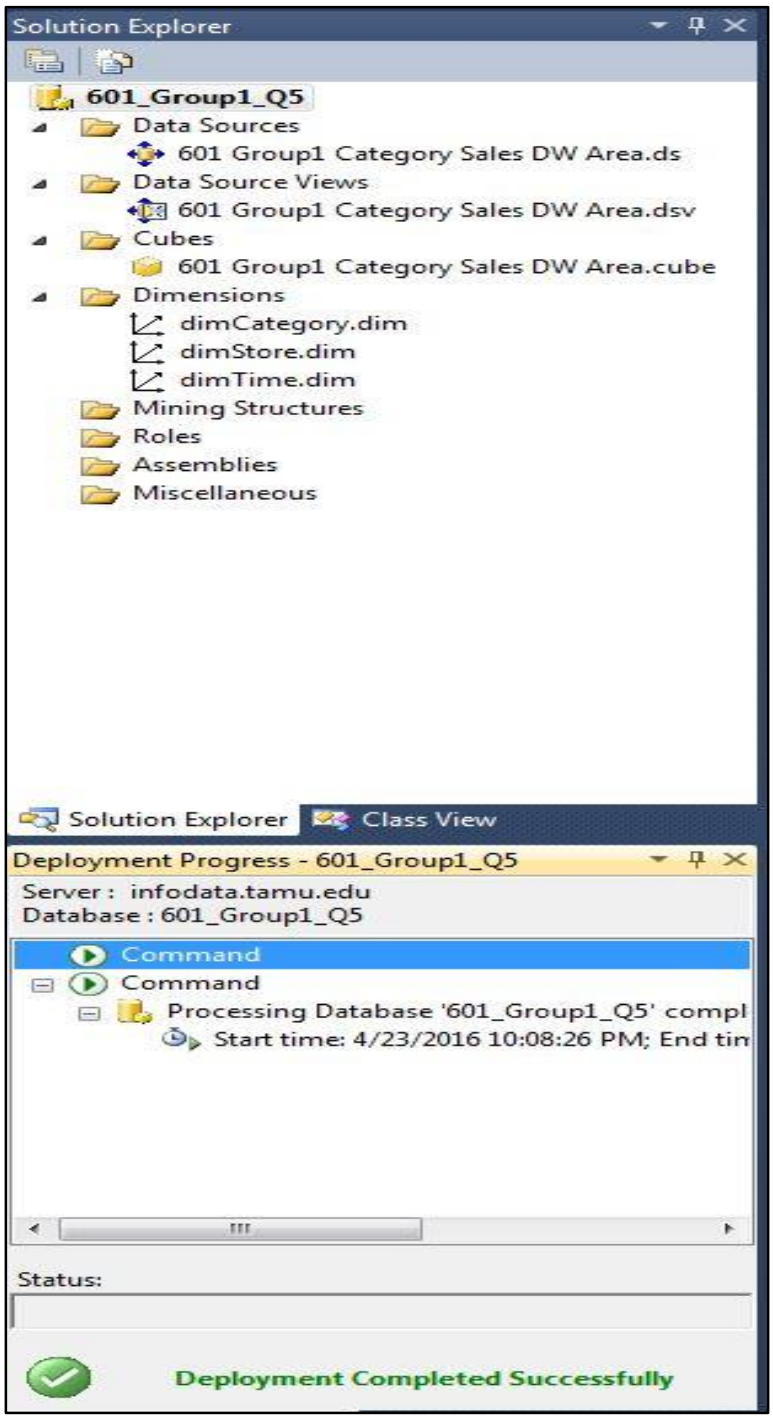
Time Dimension



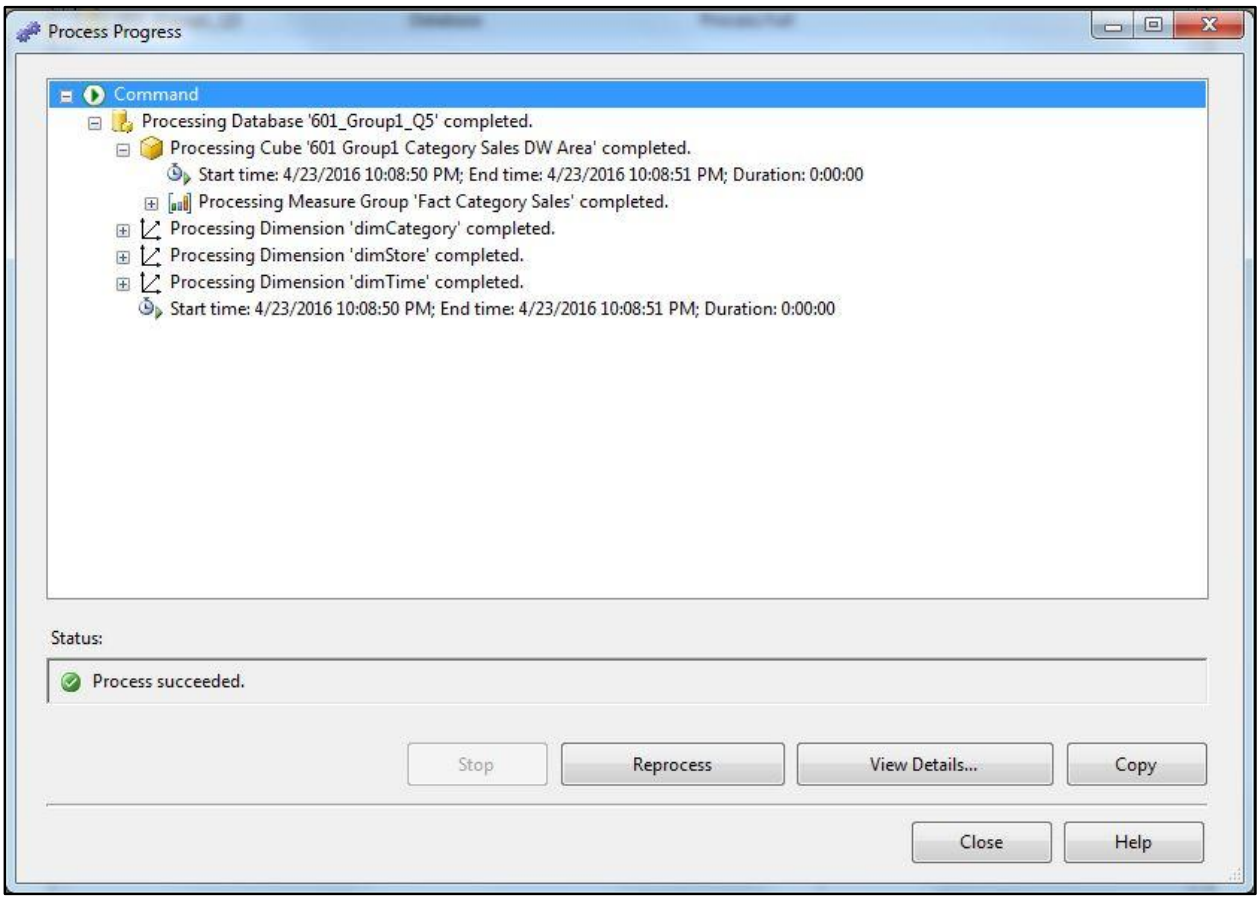
Cube Structure



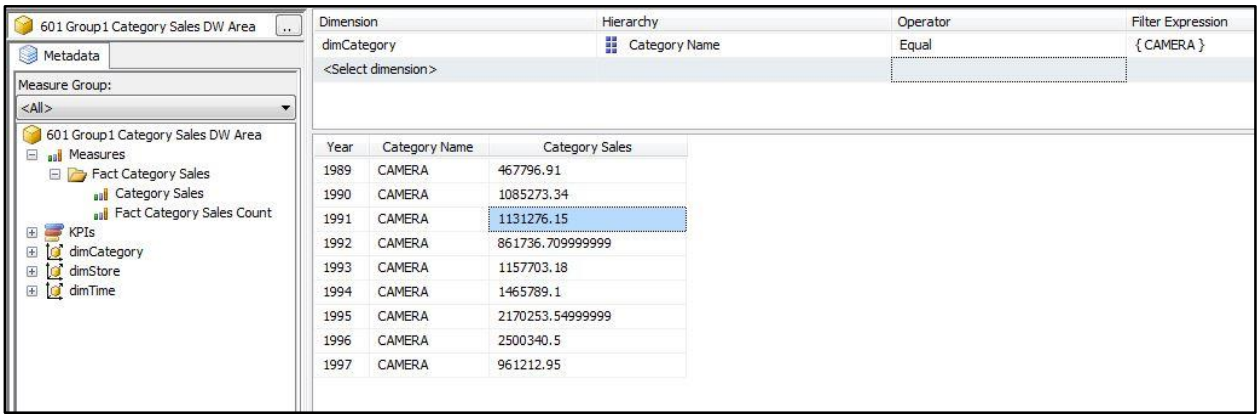
Successful Deployment



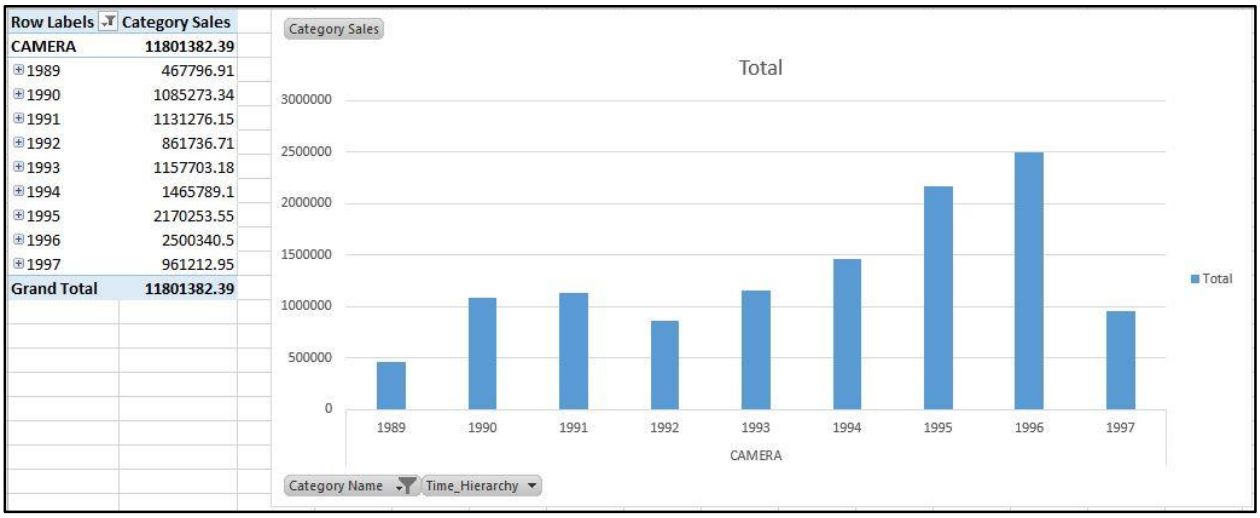
Successful Process



Cube Browser



Analysis using Pivot Chart



Row Labels	Category Sales
CAMERA	11801382.39
1989	467796.91
1990	1085273.34
1991	1131276.15
1992	861736.71
1993	1157703.18
1994	1465789.1
1995	2170253.55
1996	2500340.5
1997	961212.95
Grand Total	11801382.39

Conclusion:

We used SSAS and Excel to answer the fifth and final question of our report "What is the trend of Camera sales from the year 1990 to 1996?" Analysis of Camera sales would help us to determine how well camera is doing as a product. It also helps us to determine whether camera as a product should continue on shelf or be discontinued so as to include other products that would be beneficial to the organization. Similar analysis could be carried out for different products and the results could help Dominic Finer Foods decide on whether a particular product should be on the shelf or not. Thus the business question was answered using only SSAS and reports when then created using Pivot Charts in Excel.

F. References:

1) http://www.webopedia.com/TERM/E/ETL.html
2) http://searchdatamanagement.techtarget.com/definition/business-intelligence

G. Work Breakdown

Team Member Details	Tasks Performed	Duration for each tasks	Signature
Vijaylakshmi Rana [SQL Server Consultant]	1. Arrange meetings and assign tasks 2. Plan report tasks 3. Implement Report Builder	2 day 4 days 4 days	
Prajwal Gonnade [Developer]	1. Plan the project tasks 2. Devise strategy and step by step procedure to implement SSAS 3. Collate and format final Report	1 day 3 days 4 days	
Supreet Nayak [Business Consultant]	1. Collect all work and combine into a project 2. Perform SSRS and deployment procedures 3. Review and check Final document	2 days 4 days 4 days	