Independent Study – IMT 600

Implementation of Data Warehouse and Business Intelligence for

Dominick's Finer Foods Dataset

- By Vaibhav Walvekar

1. Introduction

a. What is a data warehouse?

Data warehouse is a logically centralized repository used for Data Management. It is populated from internal and external data source which are combined using bus architecture.

The data warehouse consists of data which is integrated, transformed and optimized for reporting and periodic integration. Thus it also forms an integral part of Business Intelligence.

The general decision making hierarchy involves Strategic, Tactical and Operational decision making with Strategic being at the top of the pyramid, then Tactical being the middle layer and Operational decision making being the last and the base layer. The general databases designs could support operational decision making but Tactical and Strategic required development of Data warehousing. The database designs lacked integration of varied systems, summary of data and also had performance limitations.

Characteristic of Data warehousing differentiating from Databases:

	Characteristics	Operational Database	Data Warehouse
1	Data Handled	Current	Historical
2	Detail Level	Individual	Individual and Summary
3	Orientation	Process	Subject
4	Records per request	Few	Thousands
5	Normalization Level	Mostly Normalized	Relaxed Normalization
6	Update level	Highly Volatile	Less Volatile
7	Data Model	Current	Relational (Star Schema) and Multi-dimensional (Cubes)

Challenges in Data warehousing:

1. Coordination across different organizational units in making a same level granular dataset.
2. Uncertain data quality in data sources.

Architecture Choices

1. Top Down: This is used for large project scope where in higher integration levels are required and it is an Enterprise data warehouse which is logically centralized.

2. Bottom Up: This is used for small project scope where in lower integration levels are required and it has independent data marts which are logically decentralized.

b. What is Business Intelligence (BI)?

Business Intelligence is the next step after development of the Data Warehouse. It is an over-arching term which encapsulates several activities like data mining, cube processing (OLAP and OLTP), querying results and reporting graphical findings.

Business Intelligence is also about delivering relevant and reliable information to the right people at the right time with the goal of achieving better decisions faster.

It is a way to finding what is needed by business by not relying on other systems but simply generating our very own BI system.

Advantages:

1. Discover inefficient Business processes and hidden patterns.
2. Identify areas of strength and weakness.
3. Discover new opportunities.

c. How does Data warehouse and BI help companies?

As Data warehouse and BI enables business executives to understand intricacies and granularity of data, it helps them in taking an informed decision. Companies use Data warehouse and BI to showcase any new opportunities of improvements which can be visually shown to higher management and hence lesser convincing is required.

These two systems combine together to give an integrated and iterative access to the stored data which enables in constant churn of knowledge about the different departments of the company itself.

BI projects can help in Information Management. Most of the firms lack capabilities to efficiently govern possessed data, thus Bi and data warehouse can help in clearing the one version of truth found through day to day operational processing of data.

Key Performance Indicators (KPIs) can help companies measure the metrics which are important for their business to flourish and take decisions in that direction.

These two systems can combine together to provide an insight in to the future growth, needs and deliverables and help companies plan ahead of time for expected upturns or downturns.

2. Understanding Data

a. What is the data about?

I would like to acknowledge, **James M. Kilts Center, University of Chicago Booth School of Business**, for making this data publicly available. I have used this dataset and applied, Data warehousing and Business Intelligence concepts to develop an end to end Business solution model for Retail Sales Company. This is just an Academic Project.

The data is historical, which was collected when Chicago Booth and Dominick's Finer Foods entered into partnership for store level research into shelf management and pricing from year 1989 to 1994. The byproduct of this research was the generation of large amount of data about Customer Count, Products sold, UPC categories, Store demographics, Store Level Scanner data etc.

It is very complex dataset because of the sheer amount of data, the difficulty in establishing logical relationship between data and the domain knowledge requirements of retail industry. But it is also a very unique dataset as it provides information on retail margins and many other specific details.

b. Components of Data

The Dominick's dataset manual and codebook provide below details and below files:

i. Customer Count Files:

This files contains information of the customers visiting and purchasing products at a particular DFF store. It basically provides data about store traffic along with information about coupons redeemed, product sales, data of sale, etc.

The data is present in sas7 zipped and .csv format.

ii. Store-Specific Demographics:

This data contains the store-specific demographic data. It provides information on percentage of age groups of customers, their household information, household income, percent of college graduates, whether the customer is unemployed or retired, ethnicity of the customer etc.

The data has been collected by US government during census. The file format for this is sas7 zipped.

iii. UPC Files:

UPC stands for Universal Product Code. The UPC files contain one record for each UPC in a category. They contain information about product name, size, commodity code, etc. The files are sorted by UPC.

The UPC files contain a description of each UPC in a category. The files are named upcxxx, where xxx is the three-letter acronym for the category.

iv. Movement Files:

The movement files contain weekly sales data for each UPC in each store for over 5 years. The files contain information like retail price of the product, units sold, deal code, profit margin of that UPC product, etc.

The files are sorted by UPC, store, week.

v. Week's Decode table

This file records the week when the particular sale of a product was recorded. It also has information about special events during that week. It is in sas format.

3. Data Cleaning

This was most important and difficult step in this project. As the data was very complex, varied in size and was available in different formats, it required proper usage of the technologies to get to the clean dataset.

a. Narrowing scope

The dataset had information about all the products sold by Dominick's, which was more than 100 products. Thus considering the realistic timeline for this project, I decided to come up with specific Business questions related to sale of specific products which could then be extended to other products. The questions have been designed keeping in mind the business domain of Dominick's Finer Foods.

b. File specific cleaning

i. Customer Count Files:

The data is present in sas7 zipped and .csv format. This file contained 327046 rows of information, thus basic filtering and eyeballing on some of the columns suggested about the faulty information in the file. I removed rows which had "." and "negative" values in the WEEK column. This deletion of rows almost cleaned the whole dataset apart from some columns having "." which were also deleted. The datatype of WEEK column was correctly converted to "int" from "nvarchar". Further data about closed Stores was deleted as there was no information to relate them to correct stores. Based on the narrow scope taken for this project, many of the unwanted columns were also deleted from the Customer Count File.

ii. Store-Specific Demographics:

The file format for this is sas7 zipped. To read the file from this format, I have used R language and further exploration and cleaning has also been done in R. This file consisted of only 108 rows as it provided static information about the store demographics. The cleaning of this data involved deleting rows for which MMID was blank, as there was no MMID, those stores were considered as closed. The column Price_Tier was added to understand if the produce is from "High", "Medium" or "Low" category, which in the initial dataset has been indicated in three different columns.

iii. UPC Files and Movement Files:

From all the individual UPC files, I have considered only files for products, "Analgesics", "Soft Drinks" and "Cereals". These files were in very good condition and just required an addition of column to specify the product name. Same is the case for Movement Files. Some of the columns with NULL values were replaced by blanks.

iv. Week's Decode table

This file contained about 400 rows and was very clean at source. I only replaced NULLS from the Special Events column by blanks.

4. Business Questions

After reading a couple of published papers about this dataset, I was able to design the below business questions. The insights from the papers were about understanding the trends of sale during festive season or understanding the chain-level effect of promotions. Thus below questions probe around analysis of sale of products keeping in mind the concepts learnt from these papers.

a. What is the trend of Wine Sales during Thanksgiving's week?
Idea - As some of the product sales change drastically during festivals, I am trying to capture one such scenario for sales of Wine during Thanksgiving's week.

Expectation - The expected result of such analysis is that we can establish trends in sales.

Outcome – We have a better knowledge of the upcoming surge or drop in sales and are able manage our inventory better during holiday seasons.

b. How are the average sales of a particular product changing according to different zones in past year (Fish)?
Idea – Understand the trend of sales of products (Fish) on zone by zone basis. The Dominick's Finer Foods has been segregated into 16 different zones for better supply chain management, thus analysis of average sales can lead us to understand which zone is doing better and what are the reasons for its better performance and consequent steps can be taken to help the underperforming zones.

Expectation – The positives of over achieving zones can be applied to under achieving zones and we can expect to have better sales in all zones.

Outcome - Implementation of a such report will help grow the business and understand plus and minus points specific to Zones.

c. Compare the effect of Bonus buy and Price Reduction in Analgesics in different zones?
Idea – Dominick's Finer Foods uses Bonus Buy and Price Reduction option to boost their sales. The idea is to analyze which among the two strategies is more successful in boosting the sales. I consider "Analgesics" product to analyze this use case.

Expectation – The expectation of this analysis to establish any one among the two strategies as better. I plan to do this by comparing the sales when each of the strategy was employed.

Outcome – This would be a onetime report for each product and thus we would know the best strategy for that product. As not all products would have same success strategy we could customize or may combine both strategies for other products.

d. What is the sale of cereals in 3 different price tiers of a store for the whole period?
Idea – Dominick's Finer Foods sells its products in three different price tiers namely: High, Medium and Low. The idea is to analyze which price tier works best and how the company can focus on sales and marketing aspects of the products in different price tiers. I consider "Cereals" product to analyze this use case.

Expectation – The expectation of this analysis is to understand sale of products in different price tiers and make specific marketing strategies to improve on sluggish performance in a particular tier.

Outcome – This report again would be specific to products and outcome of this would be to have deeper knowledge of products doing good or bad in specific price tiers, which in turn would help rethink the strategies.

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

Idea – The data provides us with sales figure for Camera along with the timestamp i.e. sales of a camera on a particular day. We can use this data to plot the sales of a camera on a yearly basis from 1990 to 1996 to showcase the trend of sales.

Expectation – The expectation is to understand sales for each individual product and identify products to be discontinued or needing more marketing push.

Outcome – This report is a must for each product, as it will help track performance. The above 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. A similar approach could be used for all the products and data could be analyzed to help the business run more efficiently.

f. During which month were the Meat sales the highest and lowest during the last 3 years?
Idea – As the data available has a time line we can do analysis for understanding trends in a season or for particular months. This is similar to the other question where we look for trends during festivals, but the idea behind this analysis is to check for any strong patterns in the last three years which then can be used to capitalize on coming months or seasons.

Expectation – Establish a seasonal trend or at least make a claim this product doesn't have to do with any season. This will help again long term strategy building and being prepared for expected surge or drop in sales.

Outcome – Preparedness on the inventory front and grabbing the opportunity to up the sales during the time of demand and reduce losses during the time of downturn.

g. What are the sales of pharmacy products in stores with % Population over age 60 greater than the average across Chicago?
Idea – The idea behind this question is to work on the demographic of the location and come up with analysis which helps boost sale of products. As we know that Pharmacy is regular requirement for Old age people, so any trend confirming this proposition would help company make better decisions. These kind of products sales related to demographics can be analyzed in detail.

Expectation – The expectation of this analysis is to establish a pattern for this product and the demographic.

Outcome – On establishing such a pattern, company can make specific schemes or increase or decrease price knowing the consequences.

h. Plot the average profit margin for soft drinks across all the stores. Determine the average of profit for the sales of soft drinks and the stores which are below the average.
Idea – Profit is the main aim of any business, thus understanding the profit margins of products can help the executives know their main source of incomes and thus it would help strategize better going forward.

Expectation –The expectation is to find out stores performing well and stores performing below par for particular products and take appropriate action to increase the average profit margins.

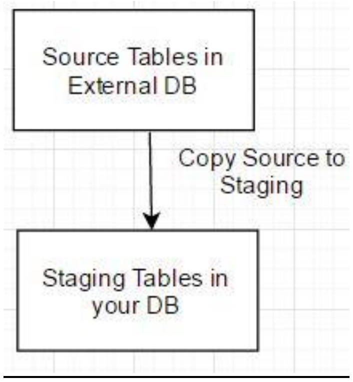
Outcome – This reporting will help gauge the direction from which the profits have been flowing in and thus better strategies will be implemented and more revenue would be generated for the company.

5. ETL Design and Implementation

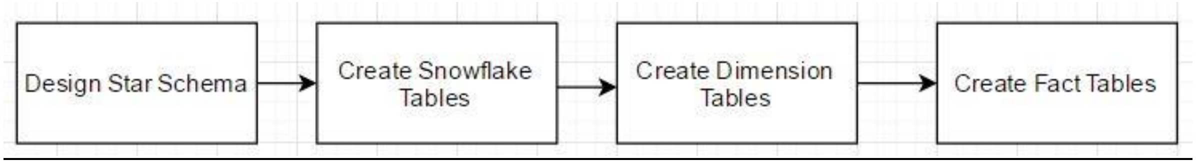
ETL Design is the process of Extract, Transform and Load. Extracting involves collection of data from internal and external sources, transforming phase involves manipulation of data to get it into correct format or at a correct grain and finally loading phase involves storing the data in the database design developed as appropriate for the business.

The general steps followed in developing an ETL are as follows:

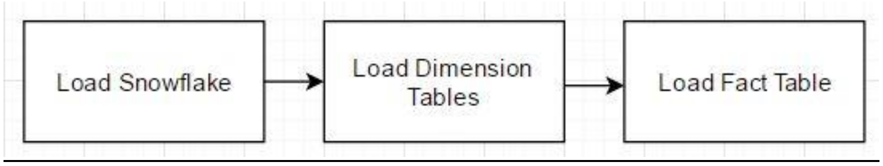
Step 1: Copy Source tables to Staging (Extract)



Step 2: Design Star Schema and Create Tables (Transform and Create)

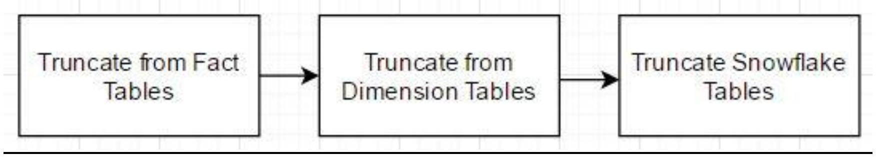


Step 3: Insert data into tables (Load)



We follow above procedure to create new data warehouse, but once the data warehouse is created and we need to refresh the data from the source we follow below steps:

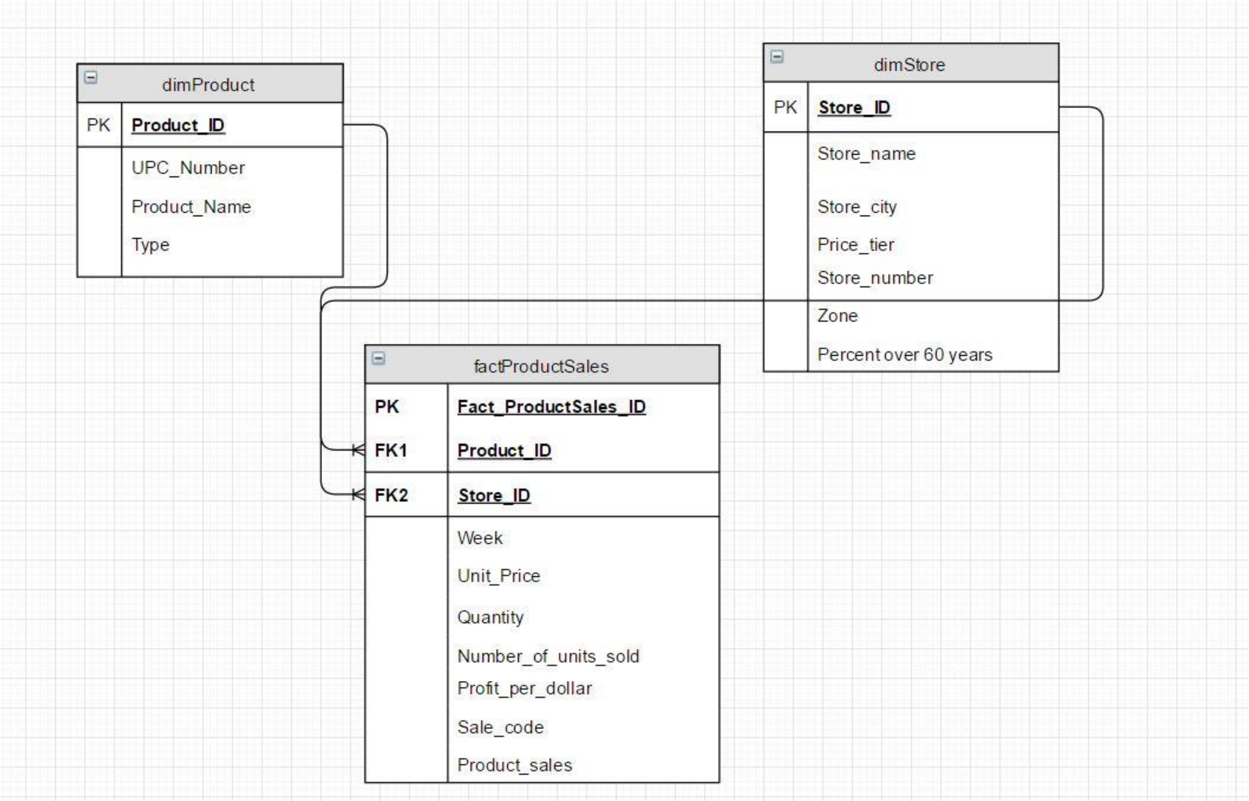
Step 4: Truncate Tables (Before this Source tables have been refreshed to staging tables)


Truncate from Fact Tables
Truncate from Dimension Tables
Truncate Snowflake Tables

After the truncation of tables, we perform step 3 and load the tables with the new data and same steps are followed after each refresh in the source tables.

For the Dominick's Finer Foods, according to the scope considered for this project, below are the final star schema ERD diagrams:

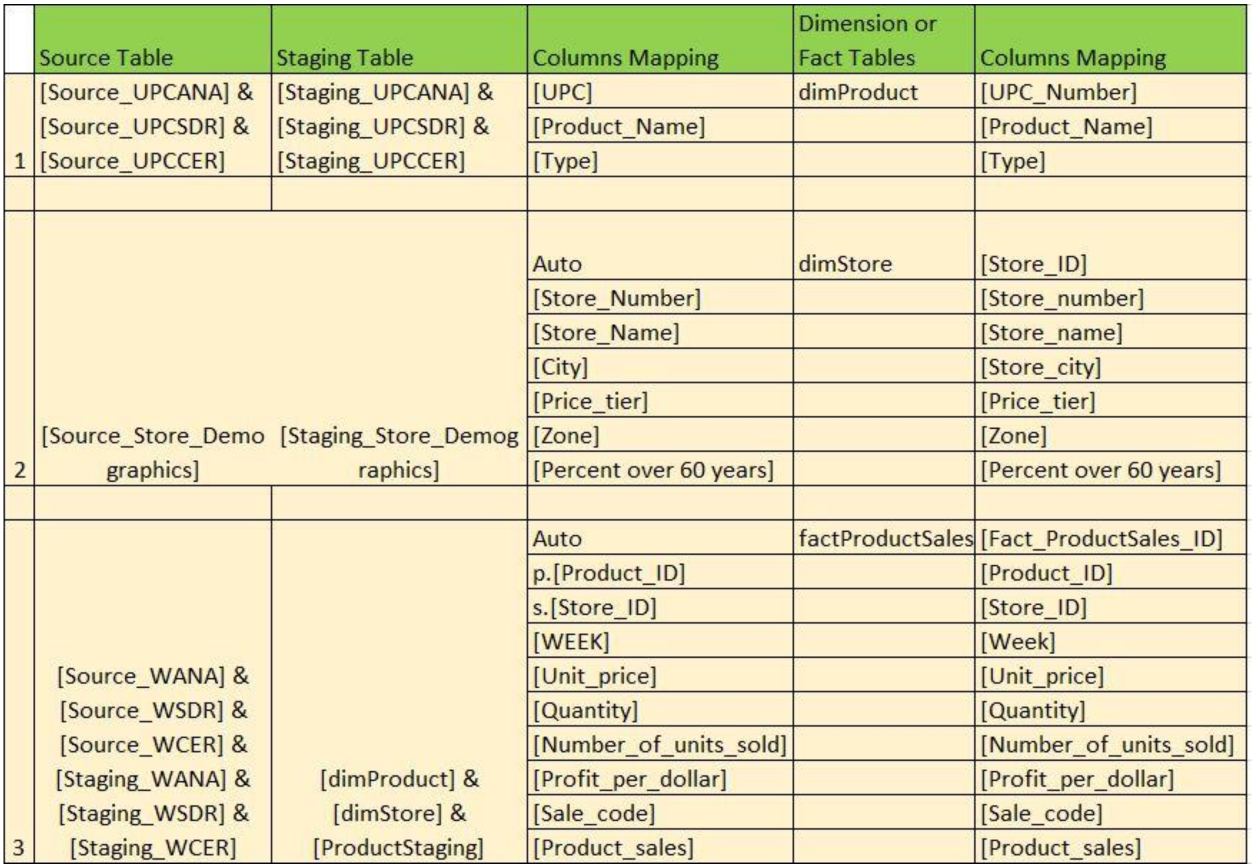
Product Sales ERD -



The above data mart has been developed to answer part of the questions as posed in the above section. The final Product Sales ERD and final Category Sales ERD, have been developed in an iterative manner, as constant design changes were required considering the scale and magnitude of information in the Dominick's database.

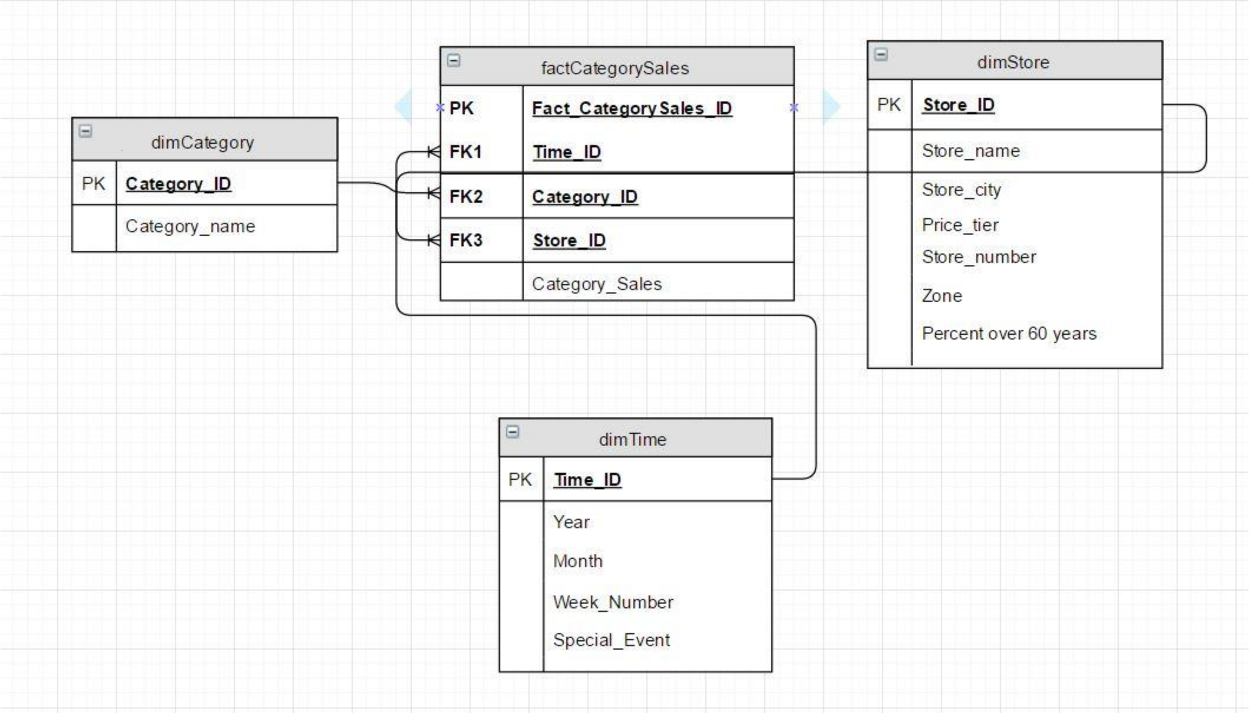
According to my analysis the above ERD will help us answer questions c, d and h.

The source and staging tables used for creating above dimension and fact tables along with their column mappings are as given below:



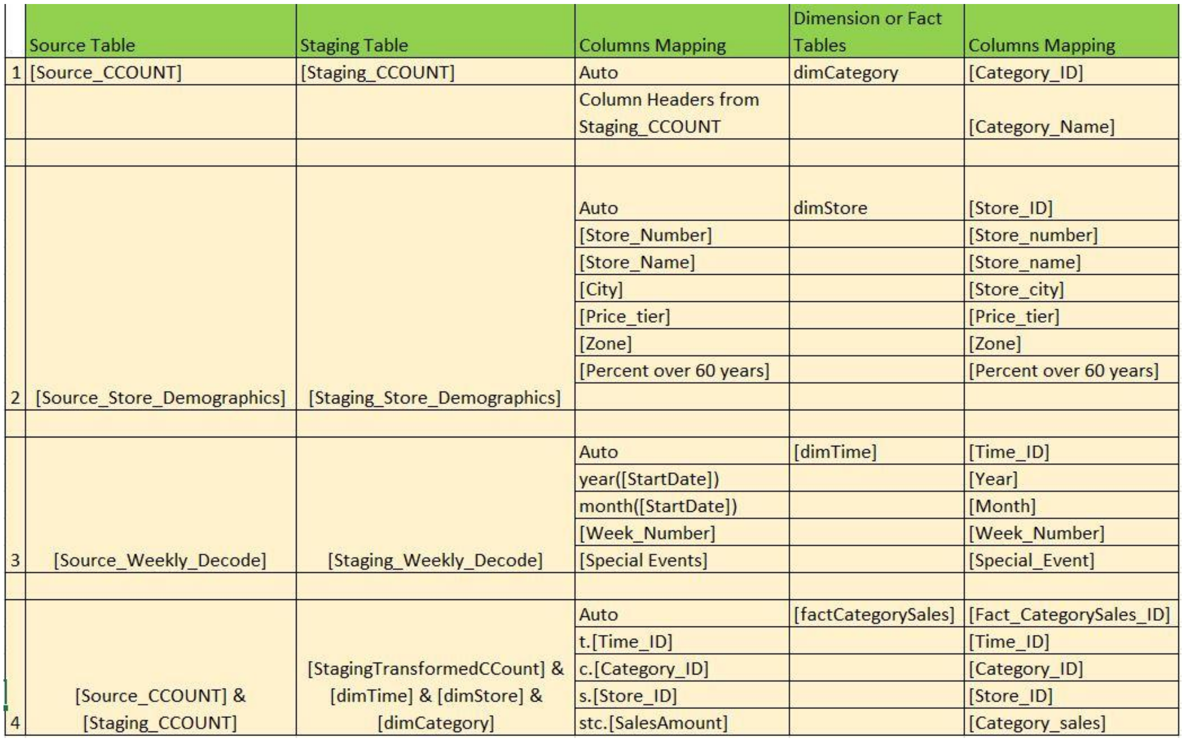
	Source Table	Staging Table	Columns Mapping	Dimension or Fact Tables	Columns Mapping
1	[Source_UPCANA] & [Source_UPCSDR] & [Source_UPCCER]	[Staging_UPCANA] & [Staging_UPCSDR] & [Staging_UPCCER]	[UPC]	dimProduct	[UPC_Number]
			[Product_Name]		[Product_Name]
			[Type]		[Type]
2	[Source_Store_Demographics]	[Staging_Store_Demographics]	Auto	dimStore	[Store_ID]
			[Store_Number]		[Store_number]
			[Store_Name]		[Store_name]
			[City]		[Store_city]
			[Price_tier]		[Price_tier]
			[Zone]		[Zone]
			[Percent over 60 years]		[Percent over 60 years]
3	[Source_WANA] & [Source_WSDR] & [Source_WCER] & [Staging_WANA] & [Staging_WSDR] & [Staging_WCER]	[dimProduct] & [dimStore] & [ProductStaging]	Auto	factProductSales	[Fact_ProductSales_ID]
			p.[Product_ID]		[Product_ID]
			s.[Store_ID]		[Store_ID]
			[WEEK]		[Week]
			[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]

Category Sales Data Mart –



According to my analysis the above ERD will help us answer questions a, b, e, f and g.

The source and staging tables used for creating above dimension and fact tables along with their column mappings are as given below:

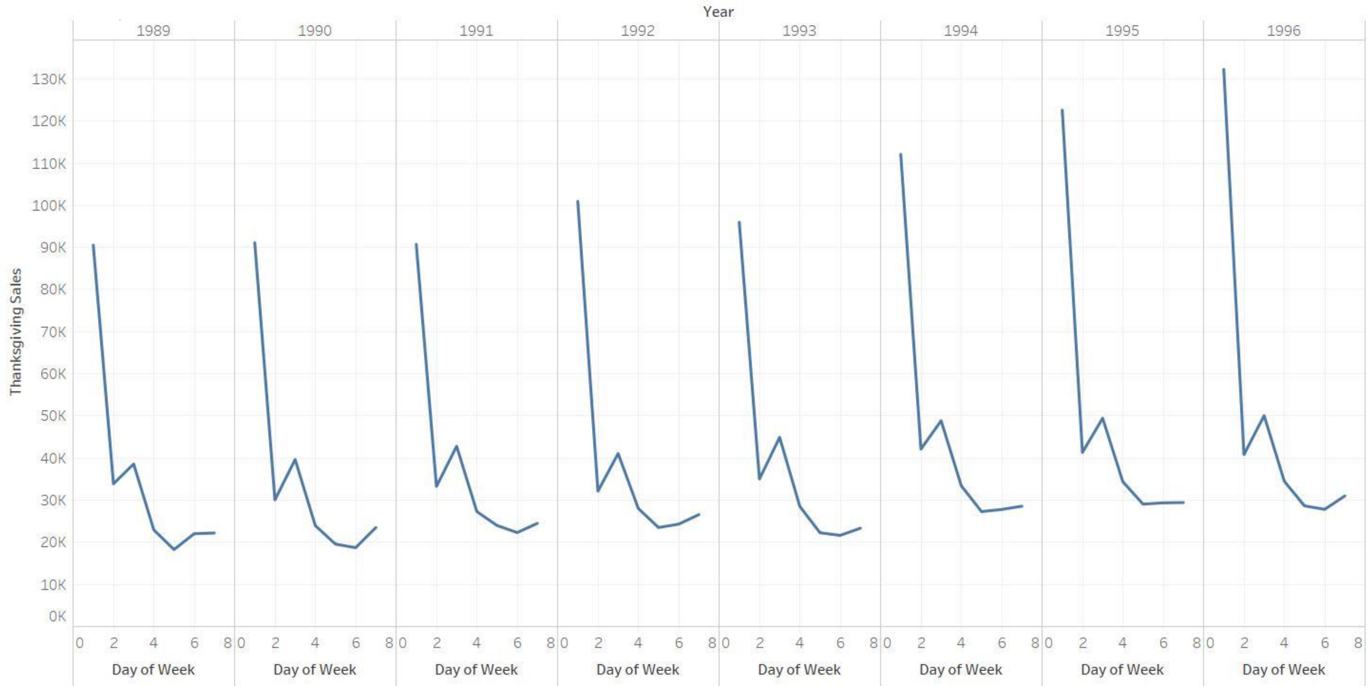


	Source Table	Staging Table	Columns Mapping	Dimension or Fact Tables	Columns Mapping
1	[Source_CCOUNT]	[Staging_CCOUNT]	Auto	dimCategory	[Category_ID]
			Column Headers from Staging_CCOUNT		[Category_Name]
2	[Source_Store_Demographics]	[Staging_Store_Demographics]	Auto	dimStore	[Store_ID]
			[Store_Number]		[Store_number]
			[Store_Name]		[Store_name]
			[City]		[Store_city]
			[Price_tier]		[Price_tier]
			[Zone]		[Zone]
			[Percent over 60 years]		[Percent over 60 years]
3	[Source_Weekly_Decode]	[Staging_Weekly_Decode]	Auto	[dimTime]	[Time_ID]
			year([StartDate])		[Year]
			month([StartDate])		[Month]
			[Week_Number]		[Week_Number]
			[Special Events]		[Special_Event]
4	[Source_CCOUNT] & [Staging_CCOUNT]	[StagingTransformedCCount] & [dimTime] & [dimStore] & [dimCategory]	Auto	[factCategorySales]	[Fact_CategorySales_ID]
			t.[Time_ID]		[Time_ID]
			c.[Category_ID]		[Category_ID]
			s.[Store_ID]		[Store_ID]
			stc.[SalesAmount]		[Category_sales]

6. BI Reporting

a. What is the trend of Wine Sales during Thanksgiving's week?

Dominick's Finer Foods - Wine Sales During Thanksgiving Week across years



The trend of sum of Thanksgiving Sales for Day of Week broken down by Year.

Conclusion

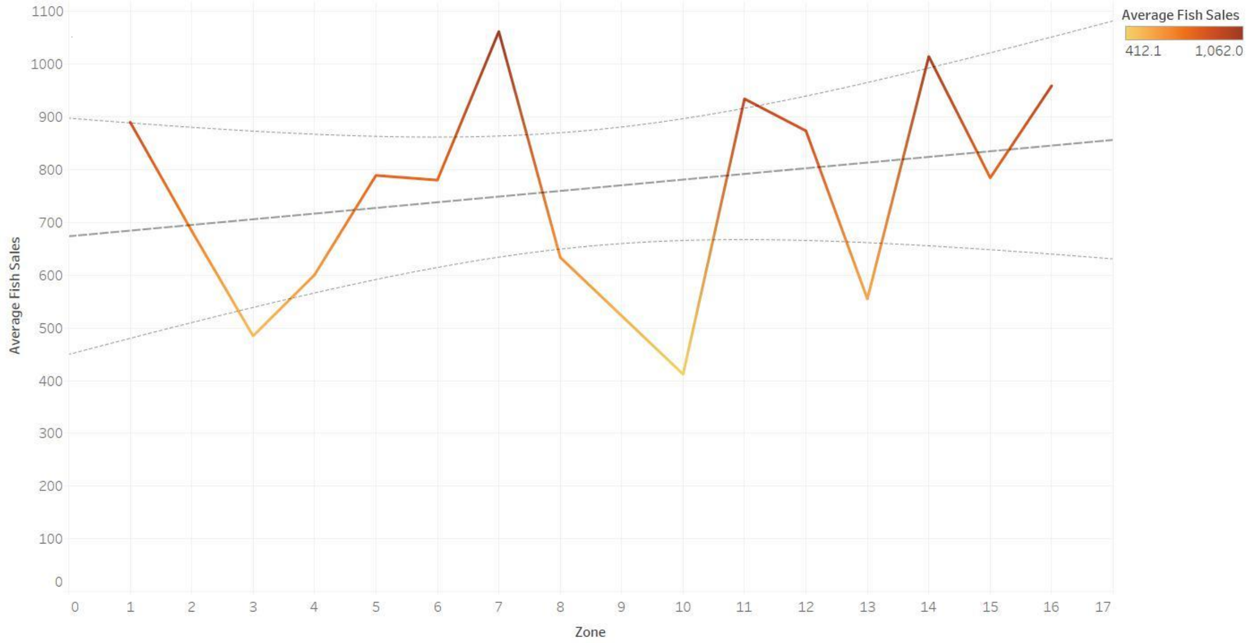
1. The Wine sales are higher during the initial part of the Thanksgiving week as people tend to shop in advance for next weekend festivities.
2. Similar trend is seen across all the years.
3. The wine sales during Thanksgiving have been increasing through the years.

Action on part of Dominick's Stores

1. Stock Wine in stores a week before Thanksgiving.

b. How are the average sales of a particular product changing according to different zones in past year (Fish)?

<Dominick's Finer Foods DB - Avg. Fish Sales across 16 Zones>



The trend of sum of Average Fish Sales for Zone. Color shows sum of Average Fish Sales.

<u>Action on part of Dominick's Stores</u>

1. Six out of the 16 zone are performing below average in Fish sales, thus concentrate on these zones to improve their Business.
2. Three zones are very near to average sales, keep an eye on these Zones in coming quarters.

Below we see the Model used to generate the Trend Line for Fish Sales:

<u>Trend Lines Model</u>

A linear trend model is computed for sum of Average Fish Sales given Zone.

Model formula:	(Zone + intercept)
Number of modeled observations:	15
Number of filtered observations:	0
Model degrees of freedom:	2
Residual degrees of freedom (DF):	13
SSE (sum squared error):	504257
MSE (mean squared error):	38789
R-Squared:	0.0722317
Standard error:	196.949

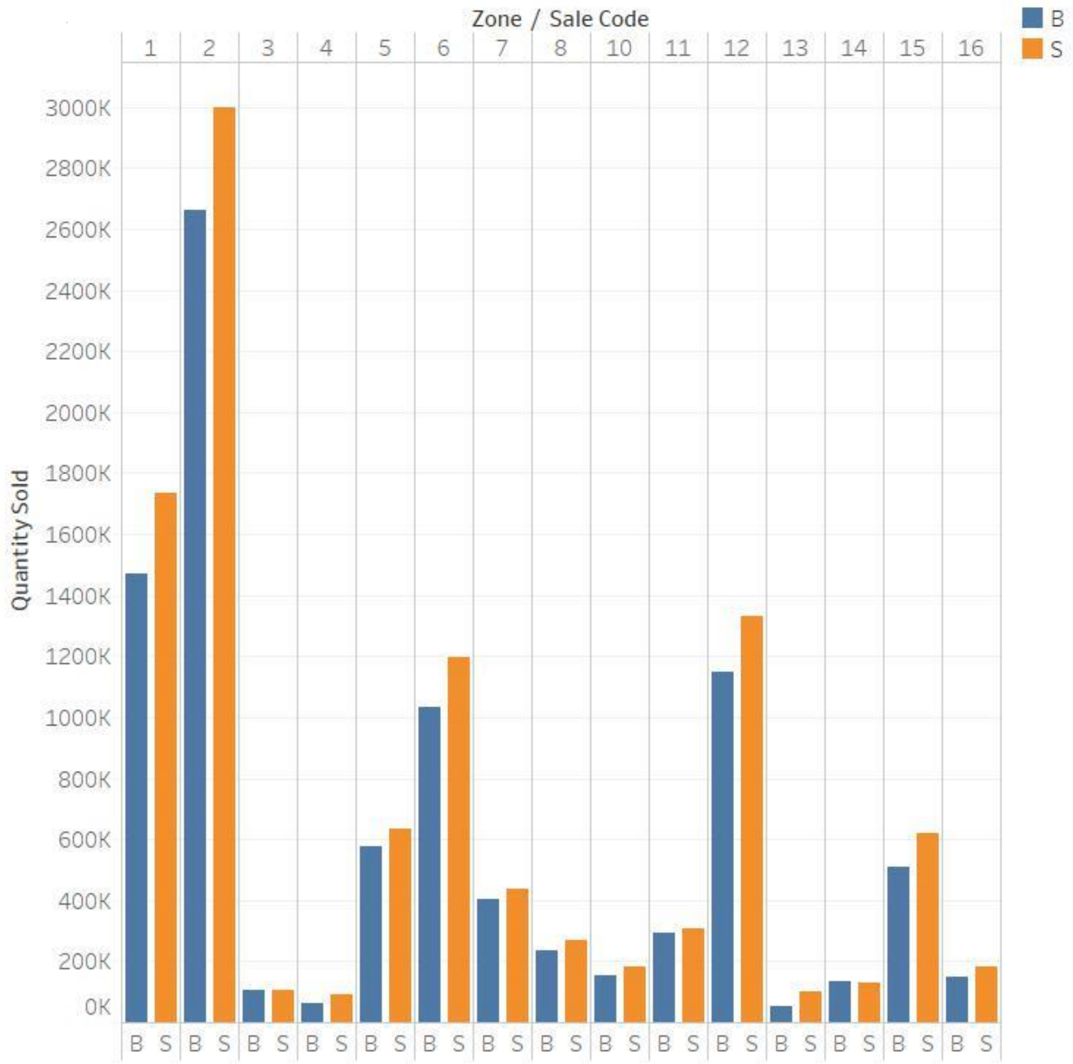
p-value (significance): 0.332755

Individual trend lines:

Panes		**Line**		**Coefficients**				
Row	**Column**	**p-value**	**DF**	**Term**	**Value**	**StdErr**	**t-value**	**p-value**
Average Fish Sales	Zone	0.332755	13	Zone	10.7498	10.6853	1.00604	0.332755
				intercept	672.83	103.781	6.48317	< 0.0001

c. Compare the effect of Bonus buy and Price Reduction in Analgesics in different zones?

Dominick's Finer Foods - Sale Code Analysis for Analgesics Product on Zone basis



Sum of Quantity Sold for each Sale Code broken down by Zone. Color shows details about Sale Code.

B indicates – Bonus Buy, S indicates – Simple Price Reduction

<u>Conclusion</u>

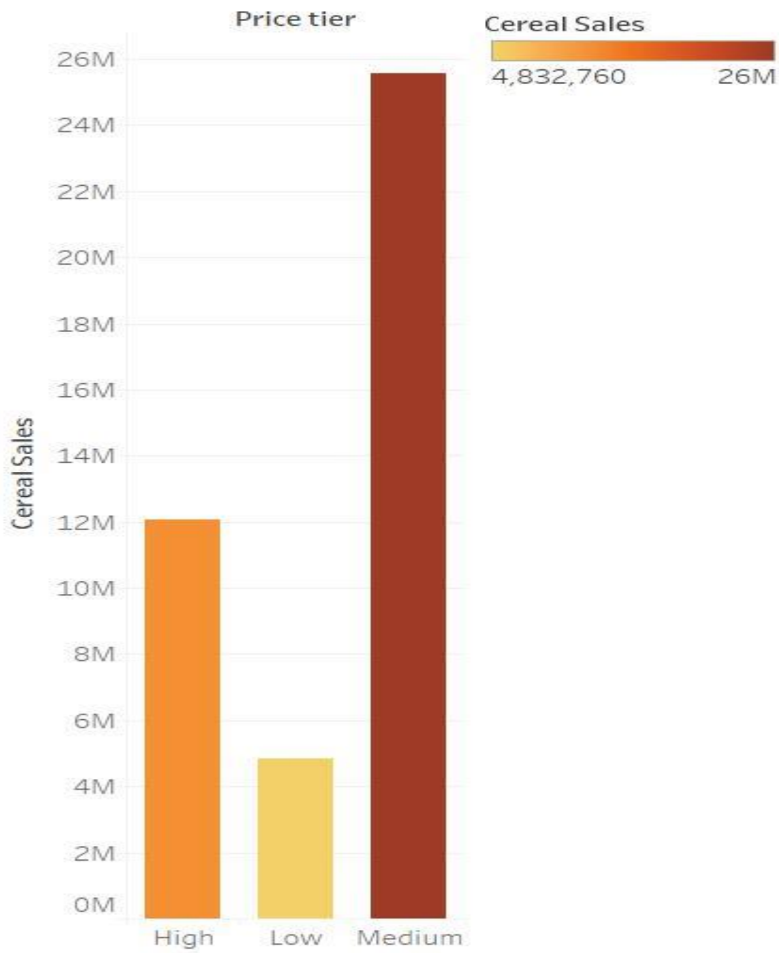
1. From the above graphic we see that in each of the zones, Simple Price Reduction leads to better sale than Bonus buy.
2. Only 2-3 zones have comparable sales for both Bonus Buy and Simple Price Reduction.

<u>Action on part of Dominick's Stores</u>

1. Try to strategize sales in providing more price reduction than bonus buy as that helps more sales.

d. What is the sale of cereals in 3 different price tiers of a store for the whole period?

Dominick's Finer Foods - Cereals sale in different Price Tiers



Sum of Cereal Sales for each Price tier. Color shows sum of Cereal Sales.

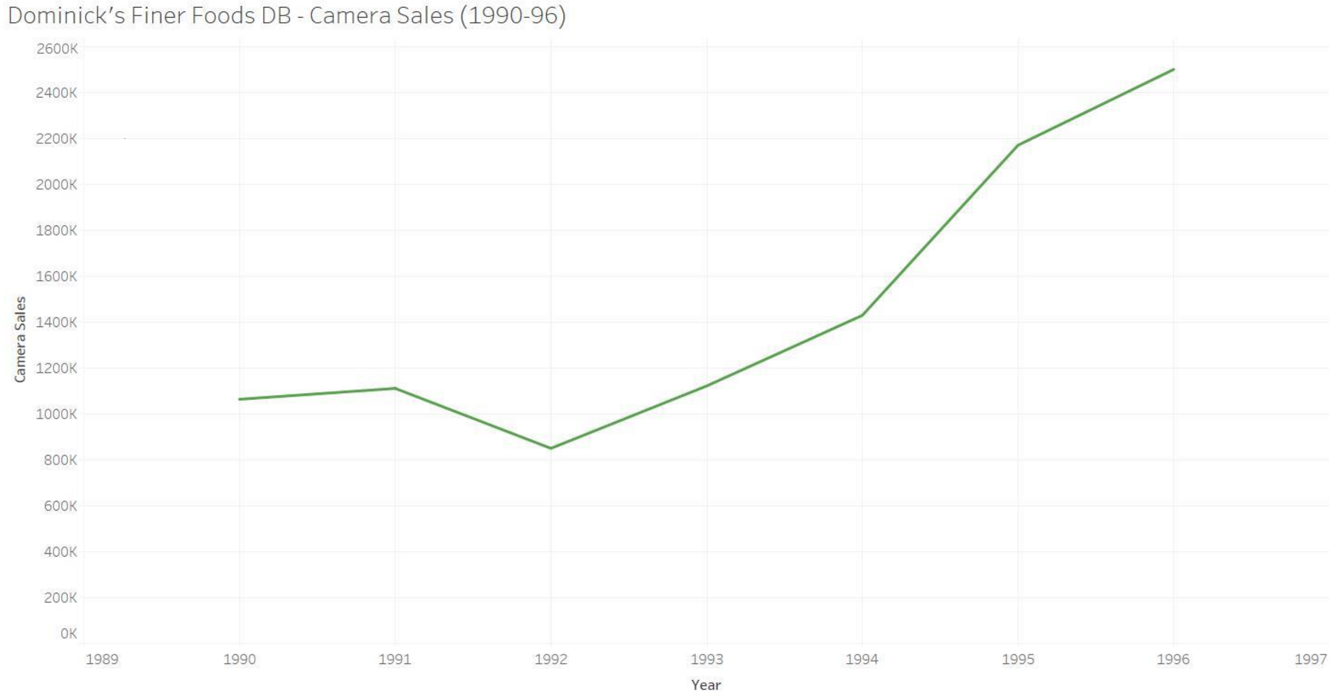
<u>Conclusion</u>

1. For the whole period of over 7-8 years, the Medium tier Cereals have sold more than High and Low tier.
2. Low Tier cereals have sold very less in terms of Medium sales, which indicates there must not be much difference in price between these two tiers.

<u>Action on part of Dominick's Stores</u>

1. Concentrate on bridging a gap between Medium and Low tier cereals to promote sale of goods in each tier or strategize to remove Low tier goods altogether.

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



The trend of sum of Camera Sales for Year.

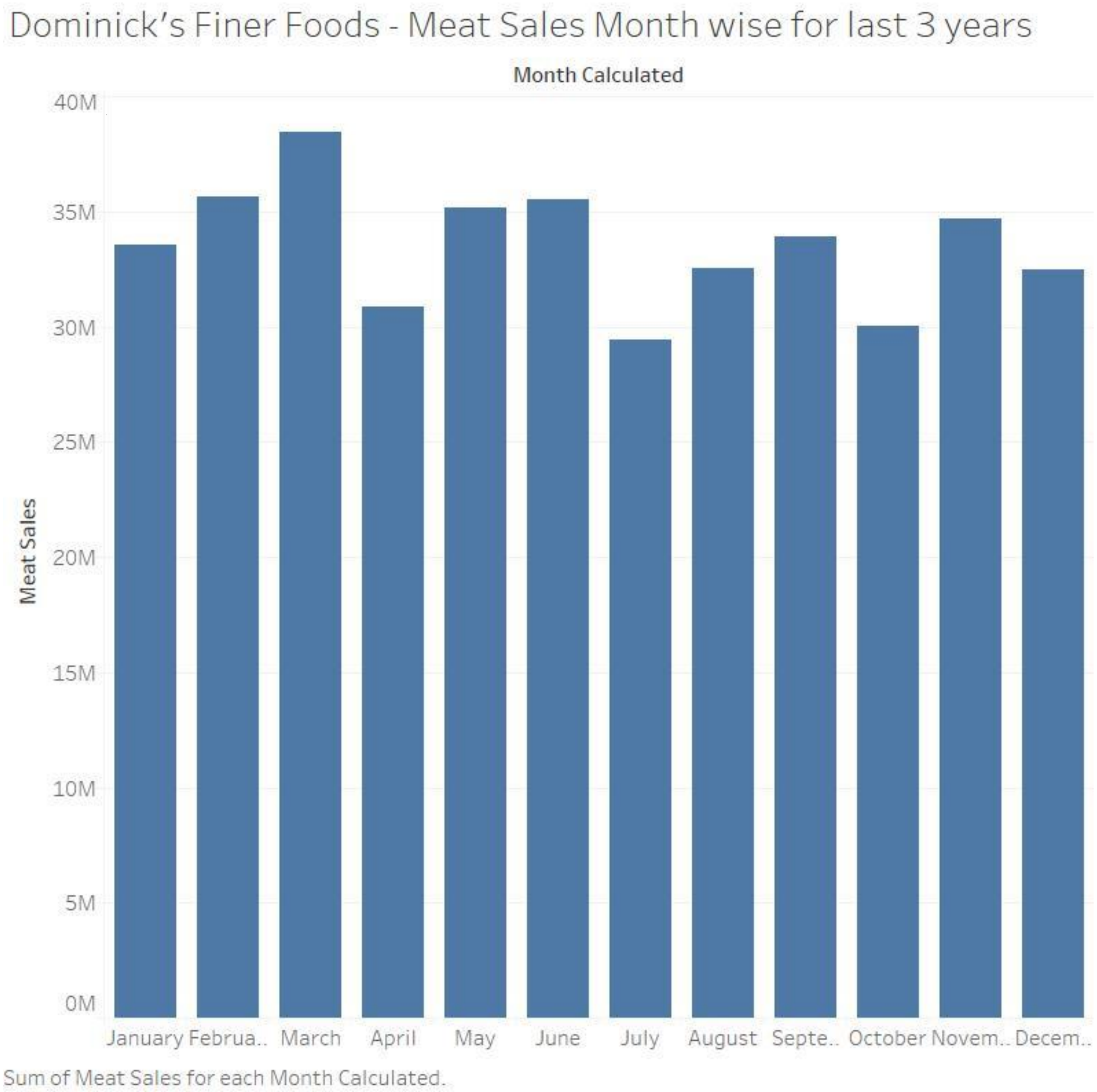
Conclusion

1. Camera sales has reduced from 1990 – 1992, but since has been increasing.
2. 1994 – 95, the increase in slope is maximum.

Action on part of Dominick's Stores

1. Camera Sales has been doing good for last several years, research more on the things going right for this product and try to implement similar strategies in other product lines.
2. Keep innovating as boom is always followed by a bear, keep an eye out for possible downturns.

f. During which month were the Meat sales the highest and lowest during the last 3 years?



<u>Conclusion</u>

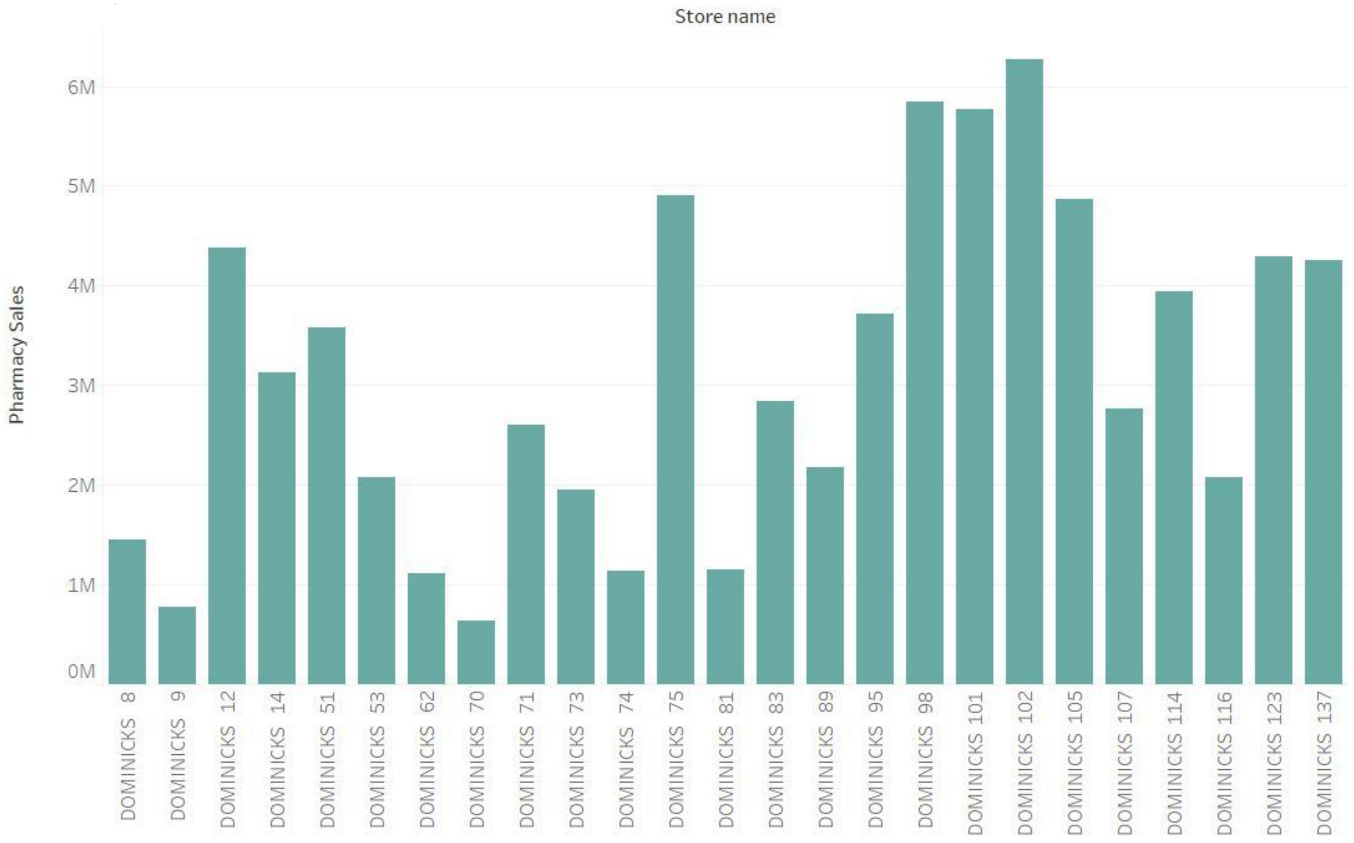
1. Meat sales were highest in month of March.
2. It was lowest during the month of June.

<u>Action on part of Dominick's Stores</u>

1. Understand the reasons for high and low sales, may be seasonal changes and work towards maintaining above average sale throughout the year.

g. What are the sales of pharmacy products in stores with % Population over age 60 greater than the average across Chicago?

Dominick's Finer Foods - Pharmacy Product sales in Stores where Population over 60 years is greater than average in city of Chicago



Sum of Pharmacy Sales for each Store name. The data is filtered on Pharmacy Sales, which ranges from 10000 to 6269608.2.

<u>Conclusion</u>

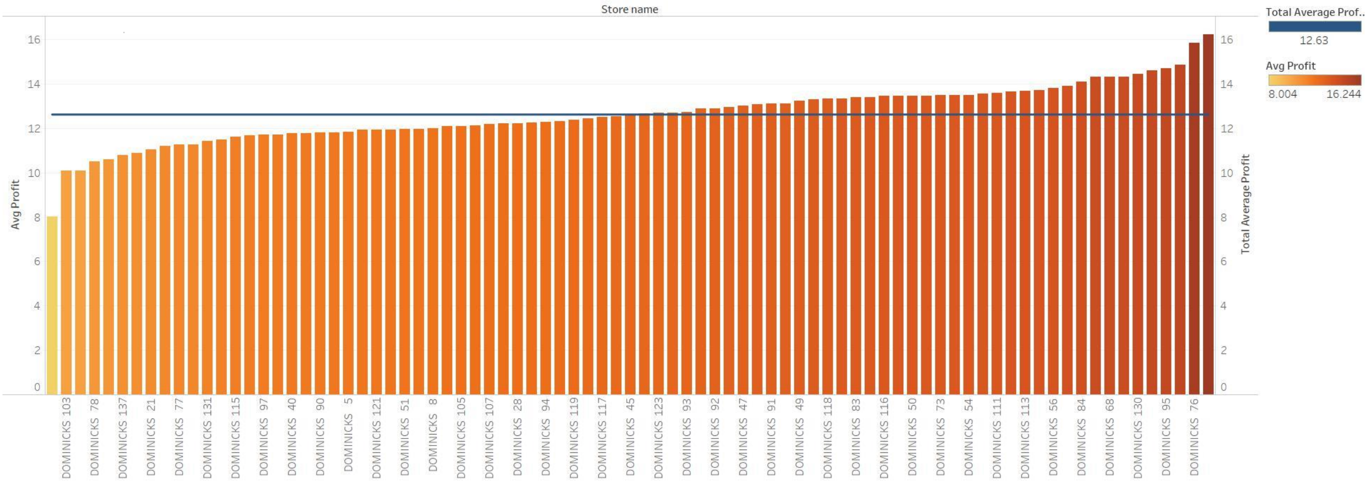
1. DOMINICKS 102 has the highest sale and DOMINICKS 70 has the lowest sale of Pharmacy Products in areas where population over 60 years is greater than average in the city of Chicago.
2. The sales spread is quite large among all stores (considering large scale pharmacy stores only)

<u>Action on part of Dominick's Stores</u>

1. Reasoning for such wide spread in the sales of Pharmacy products where elderly people are concentrated more than average points us to some missing statistic for Stores, thus analysis is required to reach to the ground reality of such differences in sales.

h. Plot the average profit margin for soft drinks across all the stores. Determine the average of profit for the sales of soft drinks and the stores which are below the average.

Dominick's Finer Foods - Avg. Profit of SoftDrinks across Dominick's Stores



The trends of Avg Profit as an attribute and Total Average Profit for Store name. For pane Avg Profit as an attribute: Color shows details about Avg Profit. For pane Total Average Profit: Color shows sum of Total Average Profit.

Conclusion

1. The average of profit margin for soft drinks across all store is 12.63
2. Out of 83 Stores almost 50% of the stores are below the average in terms of profit margin.

Action on part of Dominick's Stores

1. Try to increase profit margin of the stores performing poorly.
2. Gauge the average profit margin of 12.63 with other product's profit margins and check if Soft Drink as a product is a good product for Dominick's Finer Foods to continue to sell.