DATA WAREHOUSING

SURYA L RAMESH

The project comprises two main tasks related to data management and analysis.

In Task 1: Data Modelling and ETL, the focus is on various aspects, including table identification, listing relationships, identifying hierarchies, and ensuring the design's future-proofing.

In Task 2: Data Analysis, application of DAX formulae, providing answers to specific questions, utilizing pivot tables and concluding with a discussion and summary.

Table of Contents

Task 1: Data Modelling and ETL

1.1 Table Identification

Table Name	No. of rows	Type	Columns	Justification
Calendar Lookup	729	Dimension	Date	PK
			Day	Calendar table has been created as its essential for time-based analysis. The Day, Month Name, Year and Day are all related to the date, which will be linked to the transaction date.
			Month Name	
			Year	
			Quarter	
Customer	8,842	Dimension	customer_id	PK
			first_name	These columns contain information related to the customer and can be identified using customer_id which is the Primary Key.
			last_name	
			customer_acct_num	
			customer_address	
			birthdate	
			marital_status	
			yearly_income	
			gender	
			total_children	
			num_children_at_home	
			education	
			acct_open_date	
			member_card	
			occupation	
			homeowner	
			customer_postal_code	FK
Product	1,559	Dimension	product_id	PK
			product_name	These columns contain information related to the customer and can be identified using product_id which is the Primary Key.
			product_brand	
			product_sku	
			product_retail_price	
			product_cost	
			product_weight	
			Recyclable	
			low_fat	
Store	24	Dimension	store_id	PK
			store_type	These columns contain information related to the store and can be identified using store_id which is the Primary Key.
			store_name	
			store_street_address	
			store_city	
			store_state	
			store_country	
			store_phone	

			first_opened_date	
			last_remodel_date	
			total_sqft	
			grocery_sqft	
			region_id	FK
Sales Region	23	Dimension	region_id	PK
			sales_region	These columns contain information related to the Sales region and can be identified using region_id which is the Primary Key.
			sales_district	
Postcode Lookup	8,512	Dimension	customer_postal_code	Primary Key
			Customer_city	All related to the postcode and has been added to this separate table to avoid redundancy and can be identified using store_id which is the Primary Key.
			Customer_state_province	
			Customer_country	
Transaction	269,720	Fact	customer_id	PK
			product_id	PK
			store_id	PK
			transaction_date	PK, FK
			quantity	Not related to any other table's primary key and needed for calculations.
			stock_date	Not relates to any other table's primary key and only relevant to the stock purchased i.e. to this table.

1.2 List of Relationships

Table Name	Primary Key	Table	Foreign Key
Date	Date	Transactions	transaction_date
Customer	customer_id	Transactions	customer_id
Product	product_id	Transactions	product_id
Store	store_id	Transactions	store_id
Sales Region	region_id	Store	region_id

1.3 List of Hierarchies

Table Name	Hierarchy Name	Columns included in hierarchy	Justification
Date	**Calendar**	Year Quarter Month Date	All related to the calendar and can be used drill down while using power pivot

Product	**Brand**	product_brand, product_name	All related to the calendar and can be used to slice the information related to brand
Store	**Store_Geography**	store_city store_state store_country	All related to the calendar and can be used to slice the information related to the geography of the store
Sales Region	**Sales**	sales_region sales_district	All related to the calendar and can be used to slice the information
Postcode Lookup	**Geography**	customer_country customer_state_province customer_city customer_postal_code	All related to the calendar and can be used to slice the information

1.4 Future Proofing the design

To allow for future data to be added, the Power Queries are loaded from a Folder instead of a single file. Multiple files associated with each year can be added to this folder.

The query has a filter added to it, to pick up only files with extensions starting with '.xls'. This avoids picking up any document (.doc) or any other type of file (.jpg,.zip etc) in the same folder. However, it will look at '.xls' ,'.xlsx' and any files with macros ('.xlsm')files.

For an accurate date range, which look at future possible dates as well, A 'Calendar Lookup' table has been created which contains all the dates from a min and max range. The min and max ranges where chosen by the following code:

fx = List.Min(Transactions[transaction_date] - renamed as Min Date

fx = List.Max(Transactions[transaction_date]) - renamed as Max Date

And the date range was compiled using the following code:

List.Dates (#"Min Date", Duration. Days(#"Max Date" - #"Min Date") + 1 , #duration(1,0,0,0) - renamed as Date Range

This range of dates was converted into the type 'Date'. This table was added to the current Data model as a connection.

The effectiveness of this , has been tested with spurious data (10 rows) added to a new excel file called 'Coursework_newdata' which was copied to the source folder. It was seen that the data from this file was added to the current file and the range expanded to include the maximum date in the new file as well.

Task 2: Data Analysis

2.1 DAX formulae

Table Name	CC/ M #	Name of CC or M	CC or M	Formula
Product	1	Max Retail Price	M	Max Retail Price:=MAX([product_retail_price])
Customer	2	Average Age	M	Average Age:=AVERAGE([Customer_Age])
Store	3	days_since_opening	CC	= DATEDIFF(Store[first_opened_date],TODAY(),DAY)
Store	4	supermarket_size	CC	=SWITCH(TRUE(),Store[total_sqft]>35000,"Large",Store[total_sqft]>25000&&Store[total_sqft]<=35000,"Medium","Small")
Customer	5	Total Customers	M	Total Customers:=COUNTA([customer_id])
Store	6	store_street_number	CC	= LEFT(Store[store_street_address],SEARCH(" ",Store[store_street_address])-1)
Customer	7	age	CC	=DATEDIFF(Customer[birthdate],TODAY(),YEAR)
Customer	8	customer_priority	CC	= IF ((Customer[total_children]>=3) && (Customer[member_card] = "Golden") && (Customer[homeowner] = "Y"),"High-Priority","Normal")
Customer	9	house_number	CC	= LEFT(Customer[customer_address],SEARCH (" ",Customer[customer_address])- 1)
Transactions	10	Weekend	CC	= IF((WEEKDAY(Transactions[transaction_date])= 1)|| (WEEKDAY(Transactions[transaction_date])= 7),"Y","N")
Transactions	11	Low-Fat Quantity	M	Total Low Fat quantity sold:=SUM([LowF_quantityperitem])
Transactions	12	Total Cost	M	Total_CostperItemSold = RELATED('Product'[product_cost]) * [quantity] Total Cost:=SUM([Total_CostperItemSold])
Transactions	13	Total Revenue	M	Total_RevenueperItemSold =RELATED ('Product'[product_retail_price]) * [quantity] Total Revenue:=SUM([Total_RevenueperItemSold])
Transactions	14	Profit	M	Profit:=[Total Revenue] -[Total Cost]
Transactions	15	Product Brand Rank	M	=IF(HASONEVALUE('Product'[product_brand]),RANKX(ALL('Product'[product_brand]),Transactions[Sum of Profit_perItemsold])) Output seen in Excel sheet: Rank
Transactions	16	MTD Profit	M	=CALCULATE(Transactions[Sum of Profit_perItemsold],DATESMTD('Calendar Lookup'[Date]))
Transactions	17	Last Month Profit	M	Last Month Profit:=CALCULATE((Transactions[Profit]),DATEADD ('Calendar Lookup'[Date],-1,MONTH))

Transactions	18	MoM Profit % Change	M	SalesDiff:=[MTD Profit]-[Last Month Profit] MoM Profit % Change:=DIVIDE([SalesDiff],[Last Month Profit]) Format the result to %

2.2 Answers to Questions

Question #	Question	Answer
1	What is the maximum retail price for the "Green Ribbon" product brand?	$ 3.11
2	Which store opened first (store number)?	22
3	How many customers are female?	4386
4	What was the total Low-Fat quantity sold for "High Top" product brand?	10635
5	What is the total cost of Tri-State products sold?	$20,283
6	Which district saw the highest profit?	Los Angeles
7	Which brand is ranked #25?	Bravo

2.3 Pivot Tables

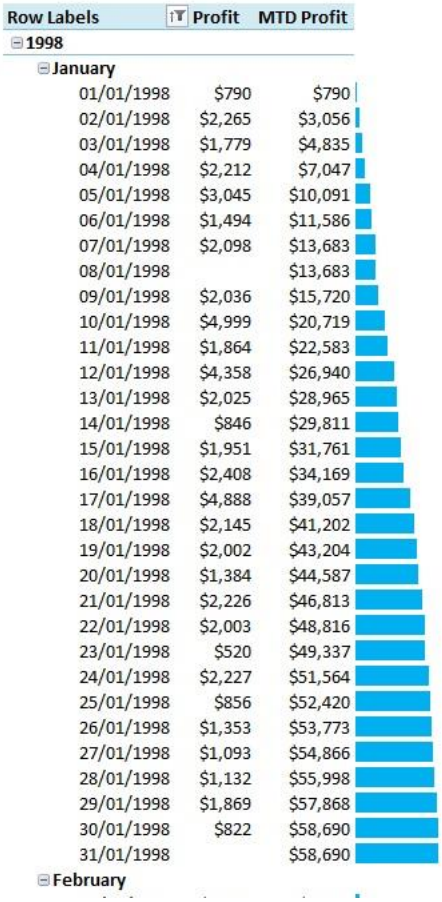
Row Labels	Profit	MTD Profit
⊟1998		
⊞January	$58,690	$58,690
⊞February	$56,451	$56,451
⊞March	$58,612	$58,612
⊞April	$56,505	$56,505
⊞May	$56,918	$56,918
⊞June	$57,938	$57,938
⊞July	$59,016	$59,016
⊞August	$56,462	$56,462
⊞Septembe	$60,480	$60,480
⊞October	$55,067	$55,067
⊞November	$67,872	$67,872
⊞December	$71,682	$71,682

Figure 1: Pivot Table 1: Month View

If the Data bars are applied only for the MTD profit column, in association with the Dates, it will show a gradually increasing bar. This may not be of much use. Hence, it has been chosen to add the Month in this chart as well. Now, it is possible to have two views:

- The daily profits (by expanding the fields) and the MTD profits associated with it.
- Month view by collapsing the fields and been able to see a more relevant Data bars with it.

Figure 1 shows the view with all fields collapsed and Figure 2 shows the partial view (For Jan) with all Dates. Figure 3 shows the second Pivot table with the Monthly profit, LMP and Mom Profit change %.



Row Labels	Profit	MTD Profit
1998		
January		
01/01/1998	$790	$790
02/01/1998	$2,265	$3,056
03/01/1998	$1,779	$4,835
04/01/1998	$2,212	$7,047
05/01/1998	$3,045	$10,091
06/01/1998	$1,494	$11,586
07/01/1998	$2,098	$13,683
08/01/1998		$13,683
09/01/1998	$2,036	$15,720
10/01/1998	$4,999	$20,719
11/01/1998	$1,864	$22,583
12/01/1998	$4,358	$26,940
13/01/1998	$2,025	$28,965
14/01/1998	$846	$29,811
15/01/1998	$1,951	$31,761
16/01/1998	$2,408	$34,169
17/01/1998	$4,888	$39,057
18/01/1998	$2,145	$41,202
19/01/1998	$2,002	$43,204
20/01/1998	$1,384	$44,587
21/01/1998	$2,226	$46,813
22/01/1998	$2,003	$48,816
23/01/1998	$520	$49,337
24/01/1998	$2,227	$51,564
25/01/1998	$856	$52,420
26/01/1998	$1,353	$53,773
27/01/1998	$1,093	$54,866
28/01/1998	$1,132	$55,998
29/01/1998	$1,869	$57,868
30/01/1998	$822	$58,690
31/01/1998		$58,690
February		

Figure 2:Pivot Table 1- Partial Daily View

Row Labels	Profit	Last Month Profit	MoM Profit % Change
1998			
January	$58,690	$33,998	72.63%
February	$56,451	$58,690	-3.81%
March	$58,612	$56,451	3.83%
April	$56,505	$58,612	-3.60%
May	$56,918	$56,505	0.73%
June	$57,938	$56,918	1.79%
July	$59,016	$57,938	1.86%
August	$56,462	$59,016	-4.33%
September	$60,480	$56,462	7.12%
October	$55,067	$60,480	-8.95%
November	$67,872	$55,067	23.25%
December	$71,682	$67,872	5.61%

Figure 3: Pivot Table 2

2.4 Discussion & Conclusion

Cost: This is a very cost-effective solution for businesses who do not want to invest in additional licenses for a dedicated database solution. Everything is done through Excel, which is a well-known software with additional functionalities built in.

Amount of data: The use of PowerPivot allows to circumvent the maximum limit of rows that is an issue in '.xls' and '.xlsx' files. As it is possible to load data from multiple files into our data model

Scalability: As everything is done in-memory, it may not be suited for large amounts of data. It is a bit slow during the stage where data is been refreshed. It does function well enough for this superstore which has 24 branches. But will need to be reconsidered if they are expecting to expand a lot more or dramatically increase their sales.

Rigid Format of input files: This design will work on the premises that a folder has been created for the source files to be saved into. Any accidental files in this folder (not of type Excel) will be ignored. It also must be noted that the users will need to add additional data (for future) in the similar format.

Data Clean-up: This solution also assumes that the data that is contained in the Excel sheet is pretty accurate, if you required to clean up the data in any way, a different solution may need to be looked at.

Postcode Lookup: In this solution, the postcode has been separated into a different table, which eliminates data redundancies in the customer table, as multiple customers can have the same postcode.

Removal of Irrelevant Columns: Ideally, it should be confirmed with the client if the stock_date (in transaction table) is of any future use in analysis. For Data models in Excel, the degree of compression that can be achieved, depends primarily on the number of unique values in each column [REF 2]. Hence if a column is not needed, it needs to be removed to minimise memory usage. If it is needed and additional info regarding stock date will be provided in future, then a separate table for stock info can be created.

Calendar Lookup: A calendar has been created from the min and max range of dates in the transaction table and will be modified according to the data, when it is refreshed. Having a separate calendar table allows the users to analyse data based on various calendar periods (day, month, quarter, year) and identify sales anomalies and take right course of action to remediate it. However, the calendar can be made more efficient by only pulling in dates that have transactions in it.

The solution also hides certain columns from the Client tools (Day in Calendar, Primary keys in Fact Table) . The primary keys in the Fact table are hidden to avoid multiple options been given to the user. The day is hidden as it tabulates the results by all the days in the year starting with the same number and not by date as needed by the requirements. The day column has be left in the calendar table, in case future analysis by particular day in all the months (for example: beginning of the month or after pay day) is needed. But if it is confirmed by the client as not needed, it needs to be removed.

Overall, the solution has met all the requirements given by the client. Further modifications as discussed above can be made, based on more information from the client. If the scale of Business is hoped to increase dramatically in the next few years, then a different database solution will need to be looked at. Till then, this solution is recommended to be used their business Intelligence solution.