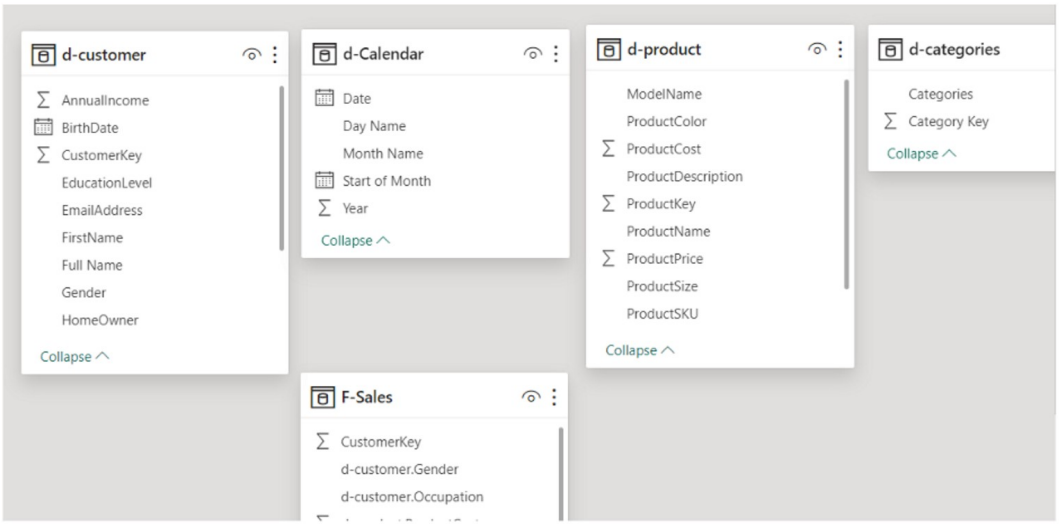
Data Modelling -1

05 September 2024 21:00

Data Modelling - 1

Step 1 : Delete the inbuilt by default relationship being build.

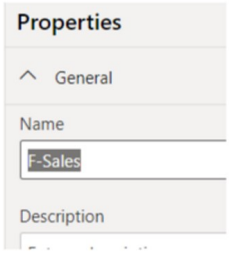
Step 2 : Remove the unnecessary tables , that would not contribute to the final project.

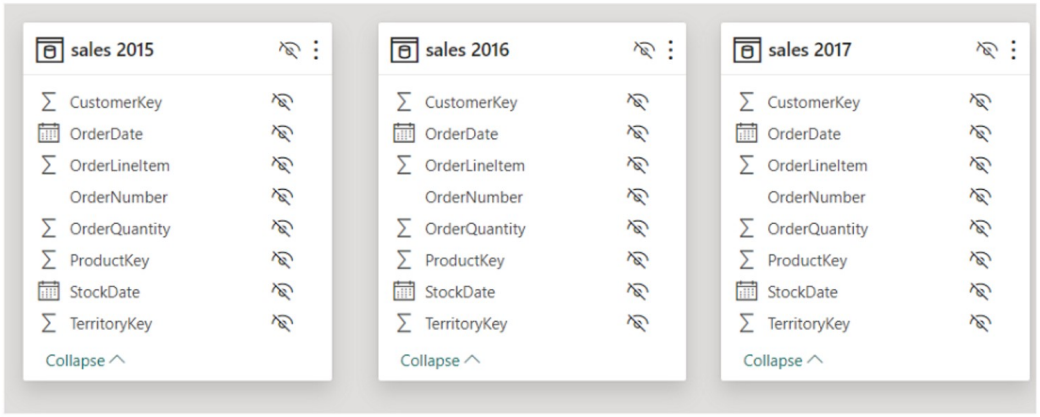


Report View

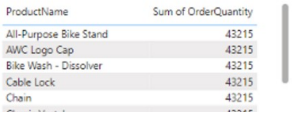
We need to add 2 more tables for our Data Modelling.

- Sub Category [Snow Flake Schema]
- Returns [Fact Table]





Hide this existing Tables. As we only need F-Sales which is appended table from 3 years of span.

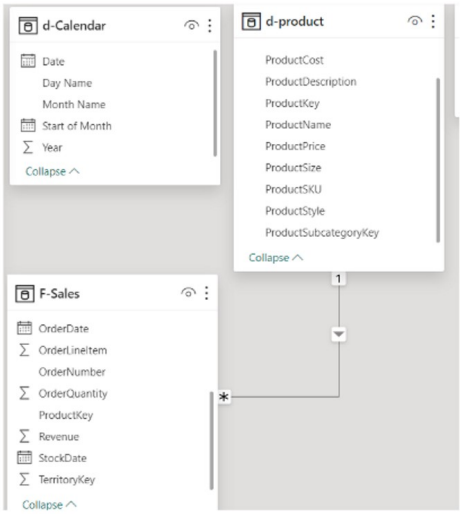


ProductName	Sum of OrderQuantity
All-Purpose Bike Stand	43215
AWC Logo Cap	43215
Bike Wash - Dissolver	43215
Cable Lock	43215
Chain	43215

if you don't have a relationship , it keeps on giving cumulative result

All-Purpose Bike Stand	43215
AWC Logo Cap	43215
Bike Wash - Dissolver	43215
Cable Lock	43215
Chain	43215
Classic Vest, L	43215
Classic Vest, M	43215
Classic Vest, S	43215
Fender Set - Mountain	43215
Front Brakes	43215
Front Derailleur	43215
Full-Finger Gloves, L	43215
Full-Finger Gloves, M	43215
Full-Finger Gloves, S	43215
Half-Finger Gloves, L	43215
Half-Finger Gloves, M	43215
Half-Finger Gloves, S	43215
Headlights - Dual-Beam	43215
Headlights - Weatherproof	43215
Total	**43215**

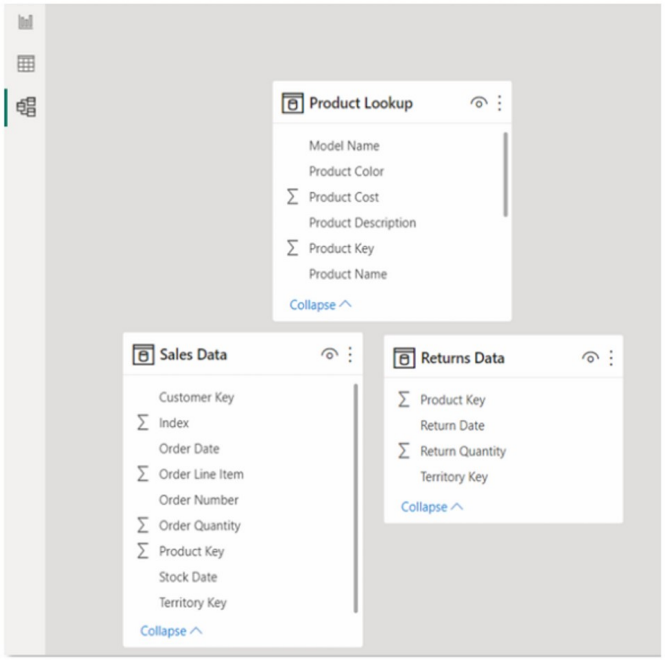
if you don't have a relationship , it keeps on giving cumulative result and shows wrong answer.



ProductName	Sum of OrderQuantity
All-Purpose Bike Stand	131
AWC Logo Cap	2091
Bike Wash - Dissolver	946
Classic Vest, L	84
Classic Vest, M	96
Classic Vest, S	85
Fender Set - Mountain	2585
Half-Finger Gloves, L	438
Half-Finger Gloves, M	460
Half-Finger Gloves, S	456
Hitch Rack - 4-Bike	188
HL Mountain Tire	981
HL Road Tire	374
Hydration Pack - 70 oz.	370
LL Mountain Tire	889
LL Road Tire	696
Long-Sleeve Logo Jersey, L	225
Long-Sleeve Logo Jersey, M	224
Long-Sleeve Logo Jersey, S	221
Total	**43215**

Dimension Table should be above Fact Table to build a Downstream flow having single filter.

WHAT IS A DATA MODEL?

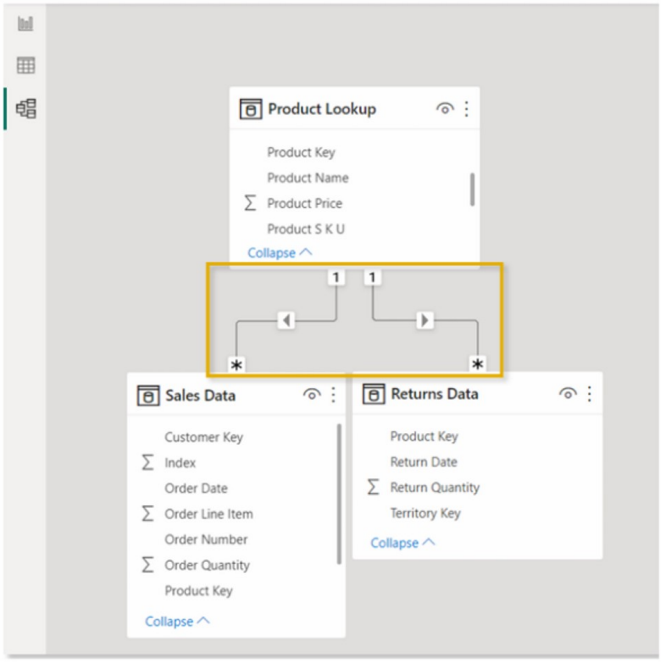


This IS NOT a data model 😞

- This is a collection of independent tables, which share no connections or relationships
- If you tried to visualize Orders and Returns by Product, this is what you'd get

ProductName	OrderQuantity	ReturnQuantity
All-Purpose Bike Stand	84,174	1,828
AWC Logo Cap	84,174	1,828
Bike Wash - Dissolver	84,174	1,828
Cable Lock	84,174	1,828
Chain	84,174	1,828
Classic Vest, L	84,174	1,828
Classic Vest, M	84,174	1,828

Cable Lock	84,174	1,828
Chain	84,174	1,828
Classic Vest, L	84,174	1,828
Classic Vest, M	84,174	1,828
Classic Vest, S	84,174	1,828
Fender Set - Mountain	84,174	1,828
Total	**84,174**	**1,828**



This IS a data model! 😊

- The tables are connected via relationships, based on a common field (Product Key)
- Now Sales and Returns data can be filtered using fields from the Product Lookup table!

ProductName	OrderQuantity	ReturnQuantity
All-Purpose Bike Stand	234	8
AWC Logo Cap	4,151	46
Bike Wash - Dissolver	1,706	25
Classic Vest, L	182	4
Classic Vest, M	182	7
Classic Vest, S	157	8
Fender Set - Mountain	3,960	54
Half-Finger Gloves, L	840	18
Half-Finger Gloves, M	918	16
Total	**84,174**	**1,828**

DATABASE NORMALIZATION

Normalization is the process of organizing the tables and columns in a relational database to reduce redundancy and preserve data integrity. It's commonly used to:

- Eliminate redundant data to decrease table sizes and improve processing speed & efficiency.
- Minimize errors and anomalies from data modifications (inserting,

processing speed & efficiency.

- Minimize errors and anomalies from data modifications (inserting, updating or deleting records).

- Simplify queries and structure the database for meaningful analysis.

In a normalized database, each table should serve a distinct and specific purpose (i.e. product information, transaction records, customer attributes, store details, etc.)

date	product_id	quantity	product_brand	product_name	product_sku	product_weight
1/1/1997	869	5	Nationeel	Nationeel Grape Fruit Roll	52382137179	17
1/7/1997	869	2	Nationeel	Nationeel Grape Fruit Roll	52382137179	17
1/3/1997	1	4	Washington	Washington Berry Juice	90748583674	8.39
1/1/1997	1472	3	Fort West	Fort West Fudge Cookies	37276054024	8.28
1/6/1997	1472	2	Fort West	Fort West Fudge Cookies	37276054024	8.28
1/5/1997	2	4	Washington	Washington Mango Drink	96516502499	7.42
1/1/1997	76	4	Red Spade	Red Spade Sliced Chicken	62054644227	18.1
1/1/1997	76	2	Red Spade	Red Spade Sliced Chicken	62054644227	18.1
1/5/1997	3	2	Washington	Washington Strawberry Drink	58427771925	13.1
1/7/1997	3	2	Washington	Washington Strawberry Drink	58427771925	13.1
1/1/1997	320	3	Excellent	Excellent Cranberry Juice	36570182442	16.4

→ Models that aren't normalized contain redundant, duplicate data. In this case, all of the product-specific fields could be stored in a separate table containing a unique record for each product id

→ This may not seem critical now, but minor inefficiencies can become major problems at scale!

FACT & DIMENSION TABLES

Data models generally contain two types of tables: fact ("data") tables, and dimension ("lookup") tables:

- Fact tables contain numerical values or metrics used for summarization (sales, orders, transactions, pageviews, etc.)

- Dimension tables contain descriptive attributes used for filtering or grouping (products, customers, dates, stores, etc.)

date	product_id	quantity
1/1/1997	869	5
1/1/1997	1472	3

date	product_id	quantity
1/1/1997	869	5
1/1/1997	1472	3
1/1/1997	76	4
1/1/1997	320	3
1/1/1997	4	4
1/1/1997	952	4
1/1/1997	1222	4
1/1/1997	517	4
1/1/1997	1359	4
1/1/1997	357	4
1/1/1997	1426	5
1/1/1997	190	4
1/1/1997	367	4
1/1/1997	250	5
1/1/1997	600	4
1/1/1997	702	5

This Fact table contains quantity values, along with date and product_id fields

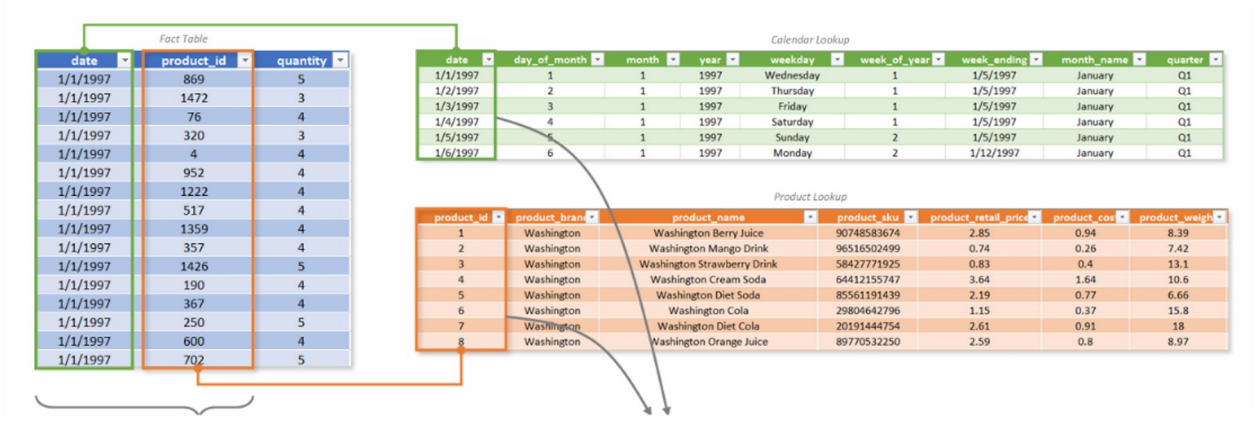
date	day_of_month	month	year	weekday	week_of_year	week_ending	month_name	quarter
1/1/1997	1	1	1997	Wednesday	1	1/5/1997	January	Q1
1/2/1997	2	1	1997	Thursday	1	1/5/1997	January	Q1
1/3/1997	3	1	1997	Friday	1	1/5/1997	January	Q1
1/4/1997	4	1	1997	Saturday	1	1/5/1997	January	Q1
1/5/1997	5	1	1997	Sunday	2	1/5/1997	January	Q1
1/6/1997	6	1	1997	Monday	2	1/12/1997	January	Q1

This Calendar Lookup table contains attributes about each date (month, year, quarter, etc.)

product_id	product_brand	product_name	product_sku	product_retail_price	product_cost	product_weight
1	Washington	Washington Berry Juice	90748583674	2.85	0.94	8.39
2	Washington	Washington Mango Drink	96516502499	0.74	0.26	7.42
3	Washington	Washington Strawberry Drink	58427771925	0.83	0.4	13.1
4	Washington	Washington Cream Soda	64412155747	3.64	1.64	10.6
5	Washington	Washington Diet Soda	85561191439	2.19	0.77	6.66
6	Washington	Washington Cola	29804642796	1.15	0.37	15.8
7	Washington	Washington Diet Cola	20191444754	2.61	0.91	18
8	Washington	Washington Orange Juice	89770532250	2.59	0.8	8.97

This Product Lookup table contains attributes about each product_id (brand, SKU, price, etc.)

PRIMARY & FOREIGN KEYS



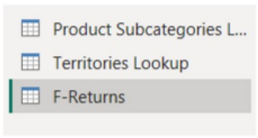
These are foreign keys (FK)

They contain multiple instances of each value, and relate to primary keys in dimension tables

These are primary keys (PK)

They uniquely identify each row of the table, and relate to foreign keys in fact tables

don't forget to add 3 new tables using Transformation :
1. Sub Category Lookup
2. Territory Lookup
3. Returns [Fact Table]



https://docs.google.com/spreadsheets/d/1VY9rWyWCBxbjM7fGb2_xrcNSrnphr4WjgGT2NeZlVRA/edit?usp=sharing

https://docs.google.com/spreadsheets/d/1KPuyGY-SZ5dc8M46B0jOQFYfX-Yuni_tfELSSAh8SRY/edit?usp=sharing

https://docs.google.com/spreadsheets/d/1KTA2trZaV5LeffduNf7yrlsZC1GHUhzr9H5Tlnk-qUc/edit?usp=sharing