Superstore Data Warehouse & Analysis
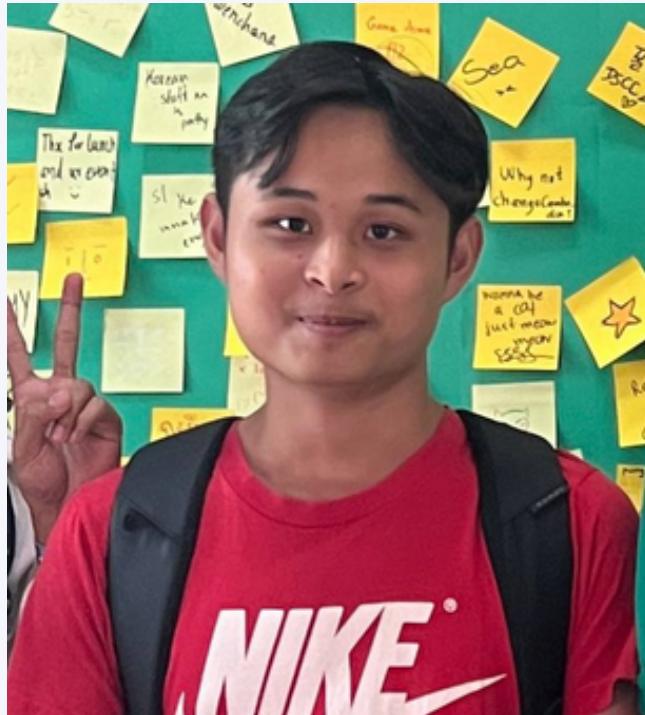
Presentated by Group 17

Lecturer: Mr. Ngin Kimlong

Our Team

Pheth Soriyuon

Phoeun Rajame

Phoeurn Kimhor

Pich Daraphal

Table of Contents

Introduction

In the era of big data, organizations rely on Business Intelligence (BI) to make informed and strategic decisions. BI combines data warehousing, ETL processes, and data visualization to transform raw data into meaningful insights. By organizing and analyzing data effectively, businesses can uncover trends, improve operations, and gain a competitive edge.

Objective

Build a Data Warehouse

Design and create a centralized database to store integrated data from multiple sources in a structured format for analysis.

Perform ETL

Use Pentaho to extract raw data, transform it into a clean and consistent format, and load it into the data warehouse.

Extract Insights from Data

Visualize the data to uncover trends, patterns, and key metrics that help support data-driven decision-making.

Dataset

Raw Dataset

ID	Order ID	Order Date	Ship Date	Ship Mode	Customer	Customer	Segment	Country	City	State	Postal Code	Region	Product ID	Category	Sub-Category	Product Name	Sales
1	CA-2017-1	########	########	Second Cla	CG-12520	Claire Gute	Consumer	United Sta	Henderson	Kentucky	42420	South	FUR-BO-10	Furniture	Bookcases	Bush Some	261.
2	CA-2017-1	########	########	Second Cla	CG-12520	Claire Gute	Consumer	United Sta	Henderson	Kentucky	42420	South	FUR-CH-10	Furniture	Chairs	Hon Delux	731.
3	CA-2017-1	########	16/06/201	Second Cla	DV-13045	Darrin Van	Corporate	United Sta	Los Angele	California	90036	West	OFF-LA-10	Office Sup	Labels	Self-Adhes	14.
4	US-2016-1	########	18/10/201	Standard C	SO-20335	Sean O'Do	Consumer	United Sta	Fort Laude	Florida	33311	South	FUR-TA-10	Furniture	Tables	Bretford C	957.57
5	US-2016-1	########	18/10/201	Standard C	SO-20335	Sean O'Do	Consumer	United Sta	Fort Laude	Florida	33311	South	OFF-ST-10	Office Sup	Storage	Eldon Fold	22.3
6	CA-2015-1	9/6/2015	14/06/201	Standard C	BH-11710	Brosina Ho	Consumer	United Sta	Los Angele	California	90032	West	FUR-FU-10	Furniture	Furnishings	Eldon Expr	48.
7	CA-2015-1	9/6/22	14/06/201	Standard C	BH-11710	Brosina Ho	Consumer	United Sta	Los Angele	California	90032	West	OFF-AR-10	Office Sup	Art	Newell 32	7.
8	CA-2015-1	9/6/200	14/06/201	Standard C	BH-11710	Brosina Ho	Consumer	United Sta	Los Angele	California	90032	West	TEC-PH-10	Technolog	Phones	Mitel 5320	907.1
9	CA-2015-1	9/6/2015	14/06/201	Standard C	BH-11710	Brosina Ho	Consumer	United Sta	Los Angele	California	90032	West	OFF-BI-10	Office Sup	Binders	DXL Angle-	18.5
10	CA-2015-1	9/6/2015	14/14	Standard C	BH-11710	Brosina Ho	Consumer	United Sta	Los Angele	California	90032	West	OFF-AP-10	Office Sup	Appliances	Belkin F5C	114

Dataset after Normalized

dim_customer

Customer	Customer Name	Segment	Postal Code	City	State	Country	
CG-12520	Claire Gute	Consumer	42420	Henderson	Kentucky	United States	
CG-12520	Claire Gute	Consumer	42420	Henderson	Kentucky	United States	
DV-13045	Darrin Van Huff	Corporate	90036	Los Angeles	California	United States	

dim_product

Product ID	Product Name	Category	Sub-Category
FUR-BO-1(	Bush Somerset Collection Bookcase	Furniture	Bookcases
FUR-CH-1(	Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded E	Furniture	Chairs
OFF-LA-10	Self-Adhesive Address Labels for Typewriters by Universal	Office Supplies	Labels
FUR-TA-1(	Bretford CR4500 Series Slim Rectangular Table	Furniture	Tables
OFF-ST-10	Eldon Fold 'N Roll Cart System	Office Supplies	Storage
FUR-FU-1(	Eldon Expressions Wood and Plastic Desk Accessories, Che	Furniture	Furnishings
OFF-AR-10	Newell 322	Office Supplies	Art
TEC-PH-10	Mitel 5320 IP Phone VoIP phone	Technology	Phones
OFF-BI-10(	DXL Angle-View Binders with Locking Rings by Samsill	Office Supplies	Binders
OFF-AP-10	Belkin F5C206VTEL 6 Outlet Surge	Office Supplies	Appliances
FUR-TA-1(	Chromcraft Rectangular Conference Tables	Furniture	Tables
TEC-PH-10	Konftel 250 Conference phone - Charcoal black	Technology	Phones

Dataset

dim_ship
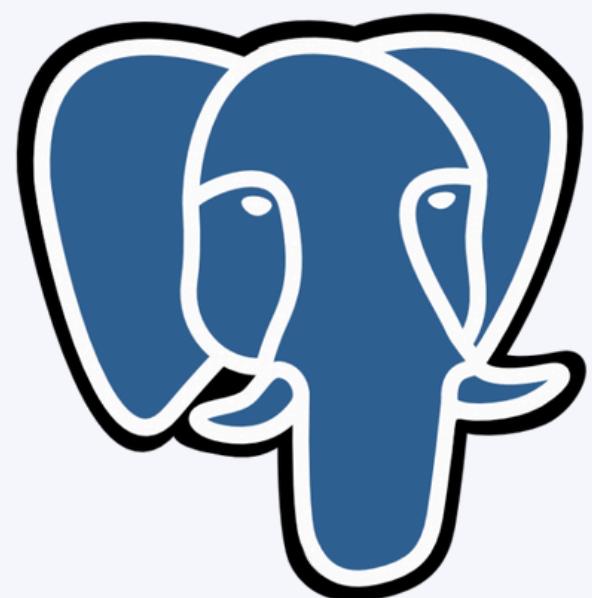
Ship Mode
Second Class
Second Class
Second Class
Standard Class
Standard Class
Standard Class
Standard Class
Standard Class

fact_sales

| Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Product ID	Sales
1	CA-2017-152156	8/11/2017	11/11/2017	Second Class	CG-12520	FUR-BO-10001798	261.96
2	CA-2017-152156	8/11/2017	11/11/2017	Second Class	CG-12520	FUR-CH-10000454	731.94
3	CA-2017-138688	12/6/2017	16/06/2017	Second Class	DV-13045	OFF-LA-1000240	14.62
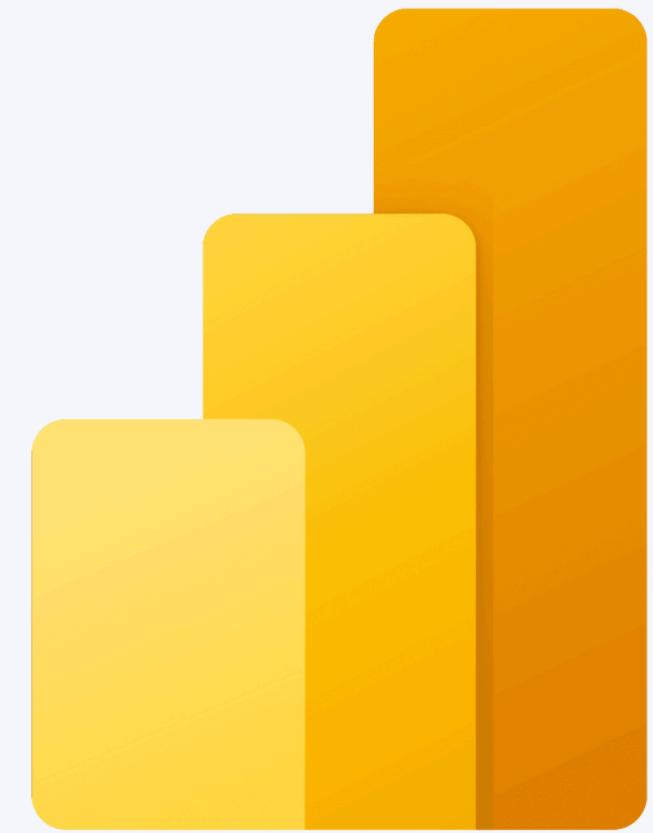
Tools Used
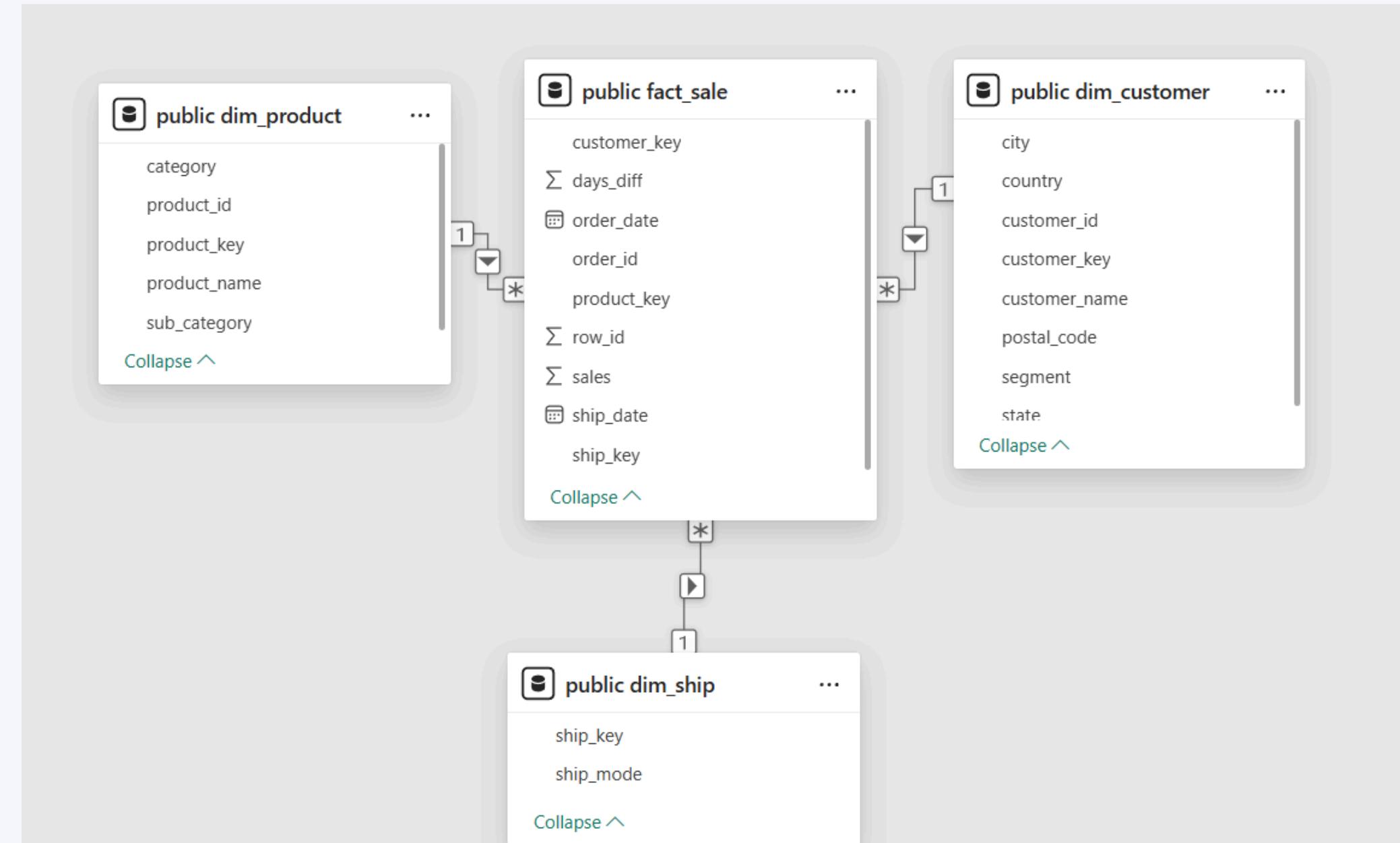
Data Warehouse Design
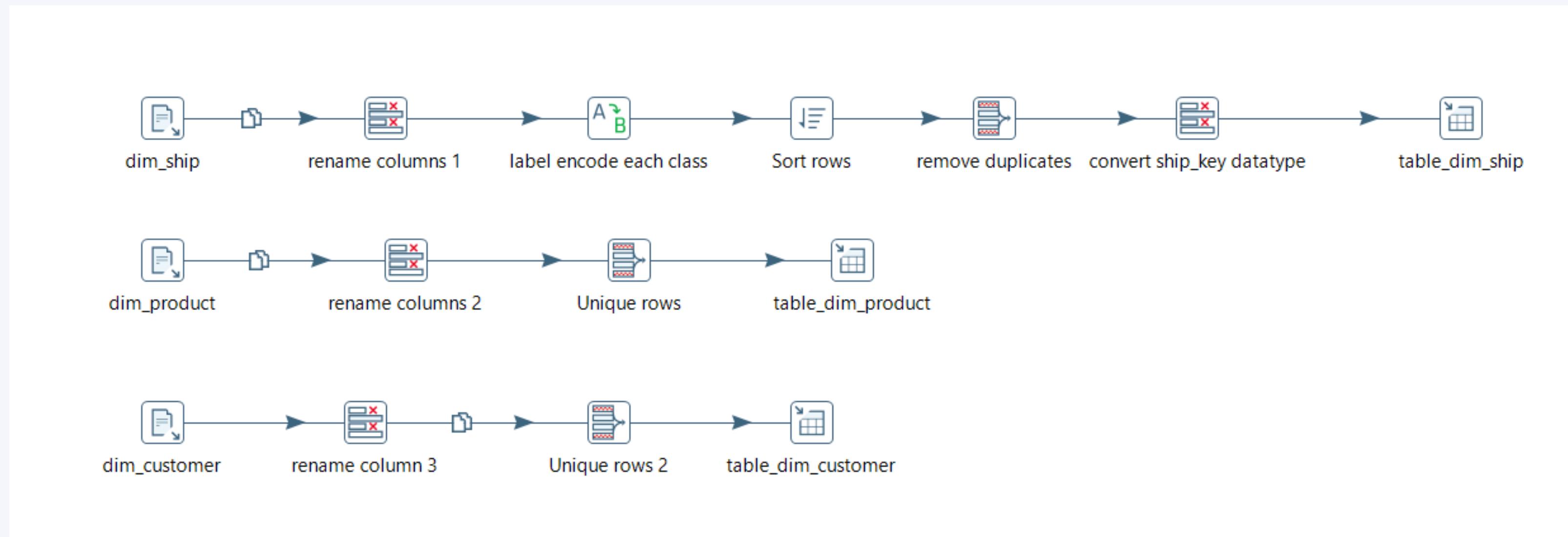
Data Warehouse Creation

```sql
▷ Run | ⧉ Select | ⊡ Ask AI
CREATE TABLE public.dim_ship (
  ship_key          BIGSERIAL,
  ship_mode         VARCHAR(50)
);


▷ Run | ⧉ Select | ⊡ Ask AI
CREATE TABLE public.fact_sale (
  row_id BIGINT PRIMARY KEY,
  order_id VARCHAR(25),
  order_date DATE,
  ship_date DATE,
  sales NUMERIC(12, 4),
  days_diff INT,
  customer_key BIGINT NOT NULL,
  product_key BIGINT NOT NULL,
  ship_key BIGINT NOT NULL
);
```
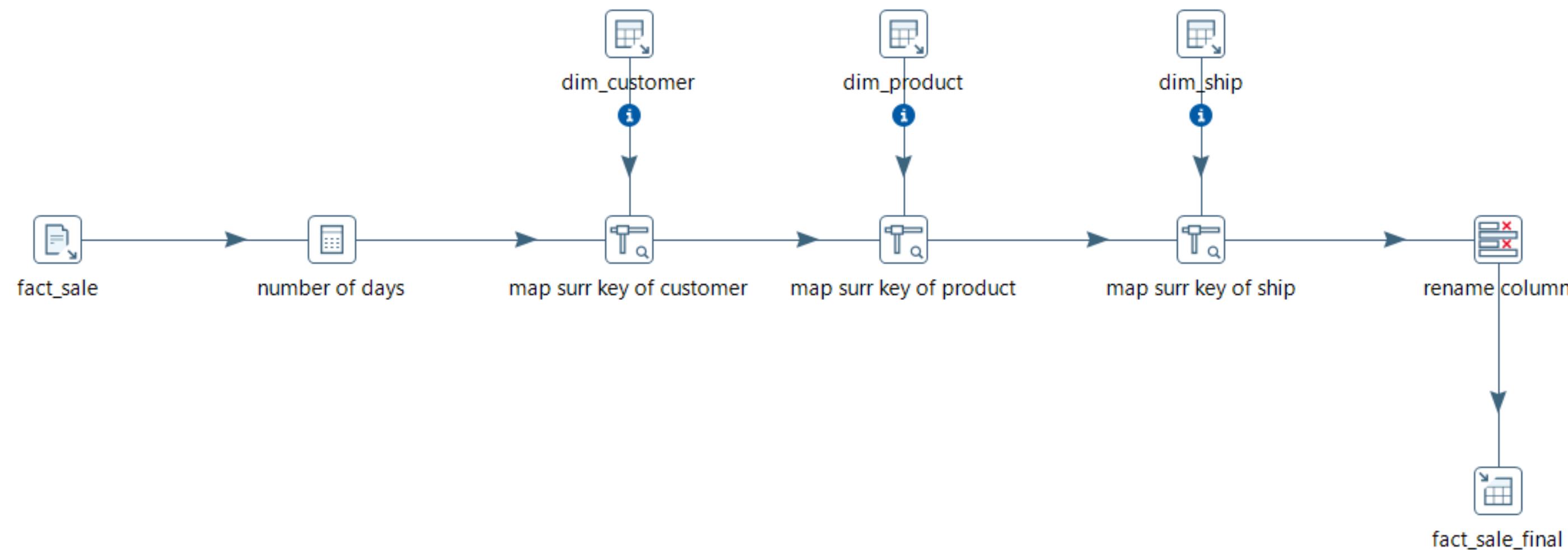
```sql
CREATE TABLE public.dim_product (
  product_key       BIGSERIAL PRIMARY KEY,
  product_id        VARCHAR(30),
  product_name      VARCHAR(150),
  category          VARCHAR(50),
  sub_category      VARCHAR(50)
);


▷ Run | ⧉ Select | ⊡ Ask AI
CREATE TABLE public.dim_customer (
  customer_key      BIGSERIAL PRIMARY KEY,
  customer_id       VARCHAR(20),
  customer_name     VARCHAR(100),
  segment           VARCHAR(50),
  postal_code       NUMERIC(10, 1),
  city              VARCHAR(100),
  state             VARCHAR(100),
  country           VARCHAR(100)
);
```
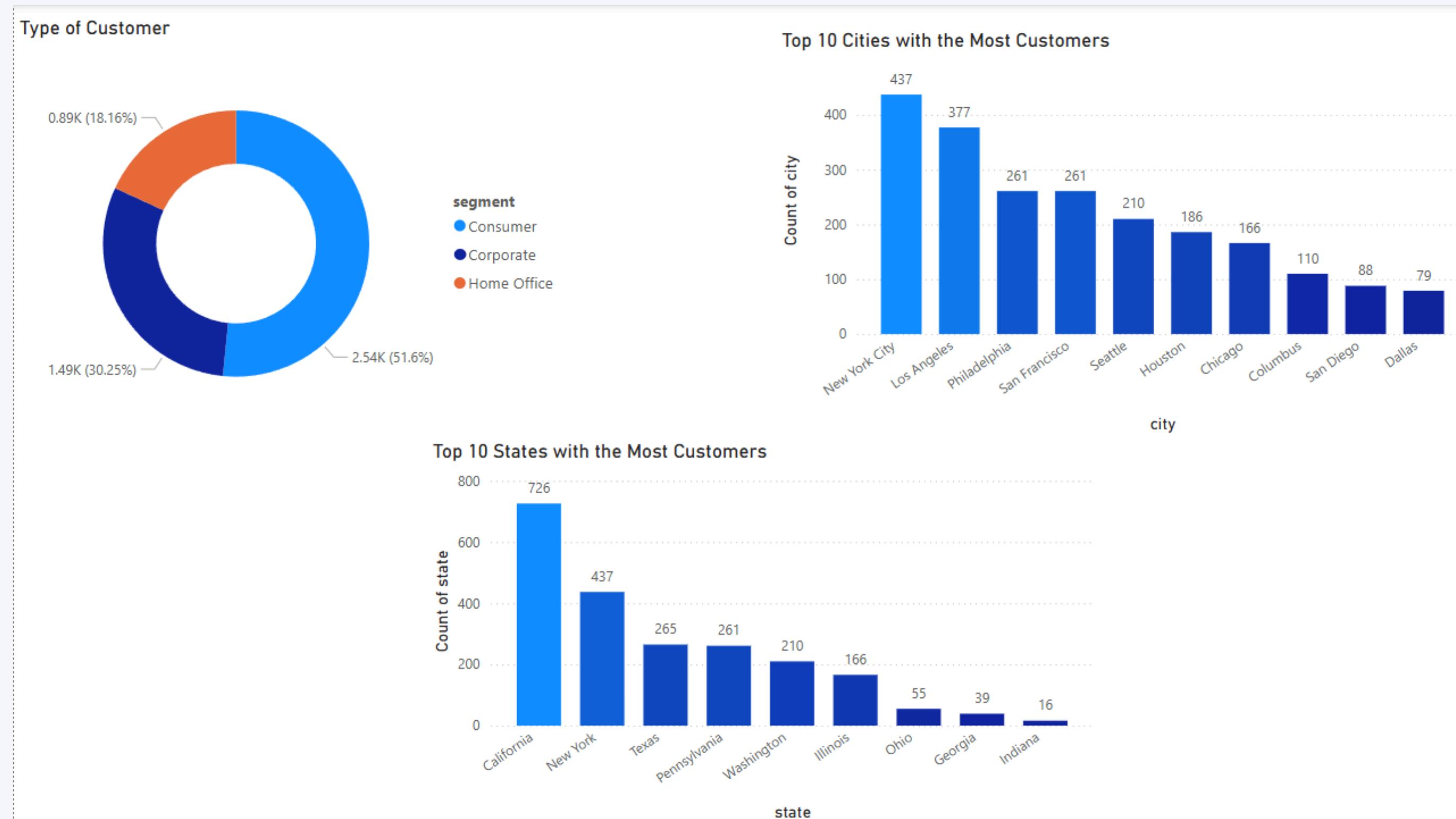
ETL of dim table

ETL of Fact_sale
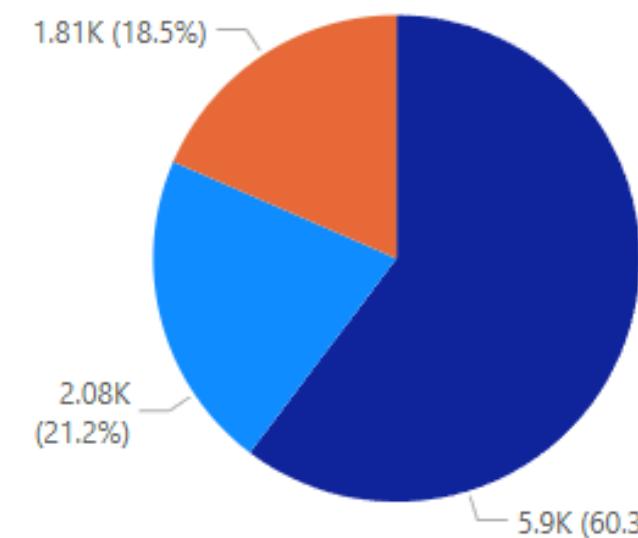
# Visualization of Customer Table
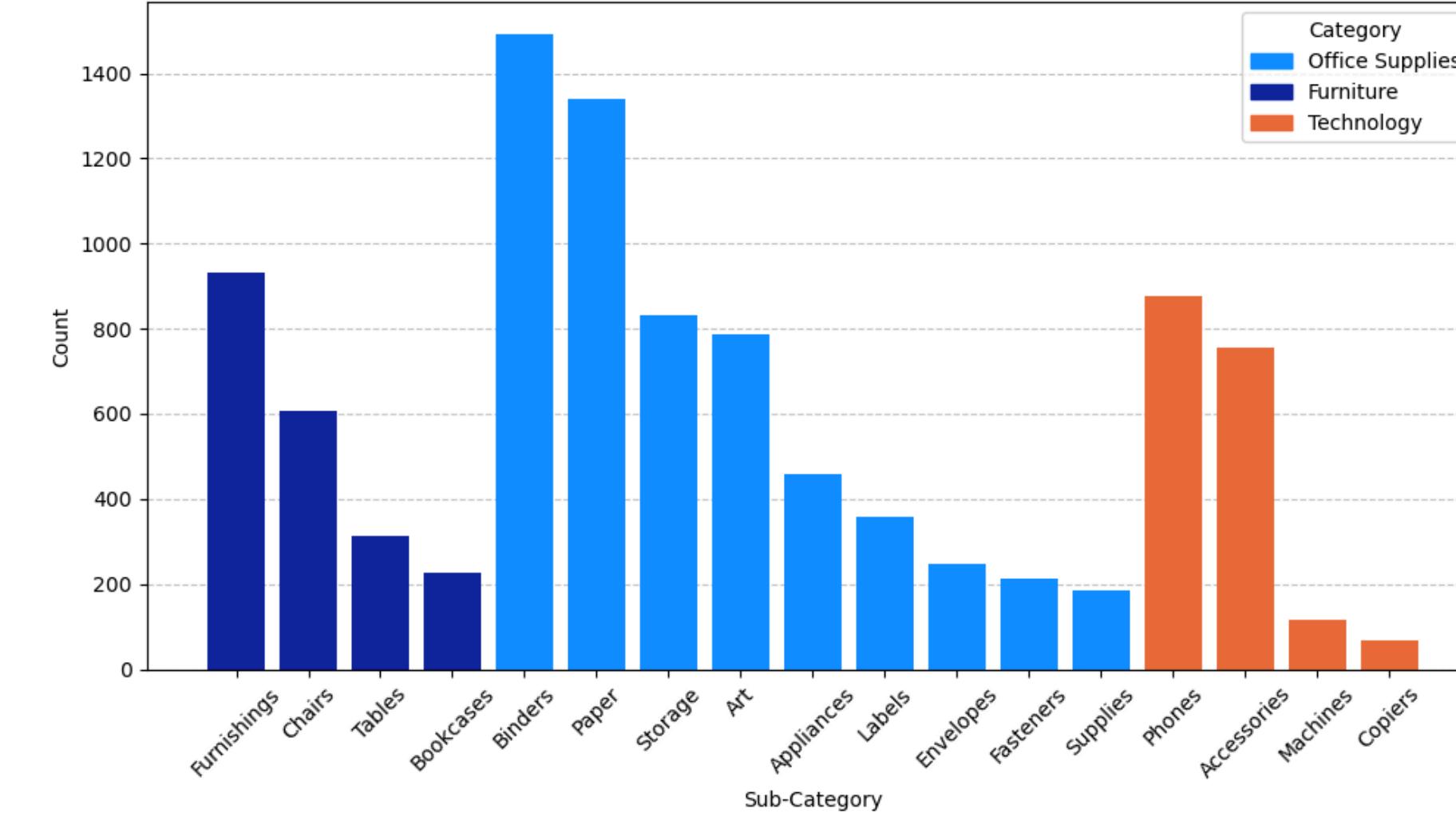
Visualization of Product Table
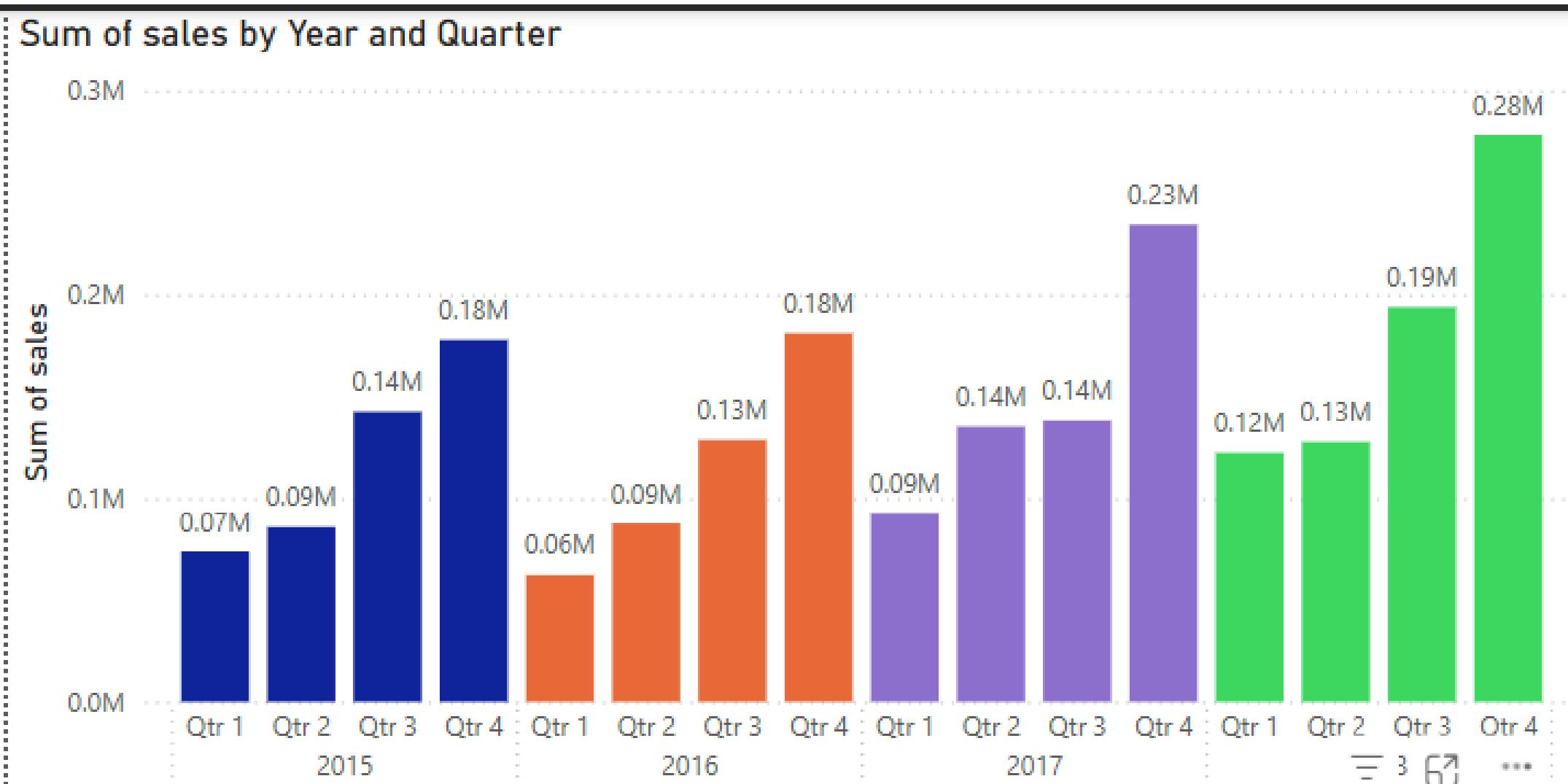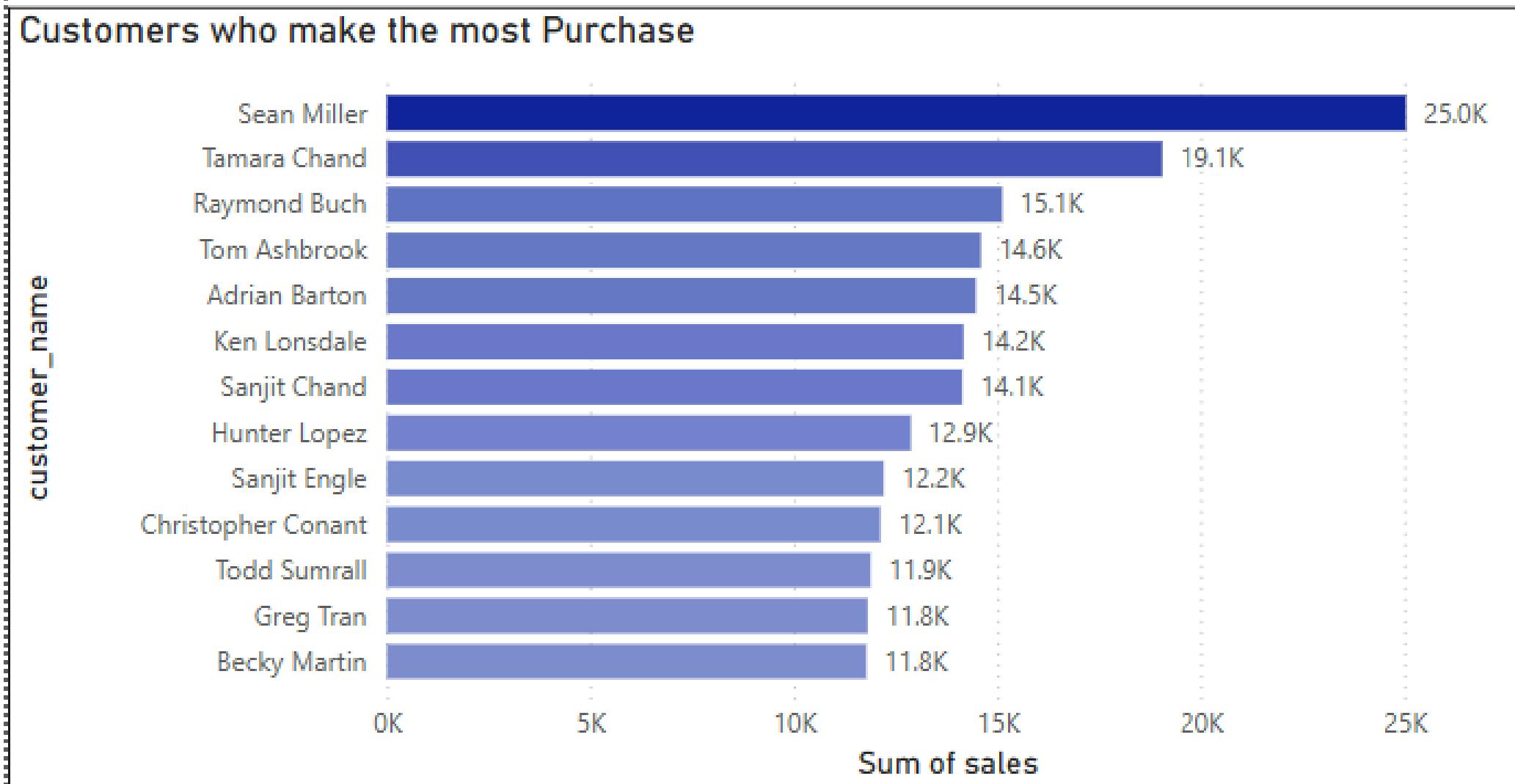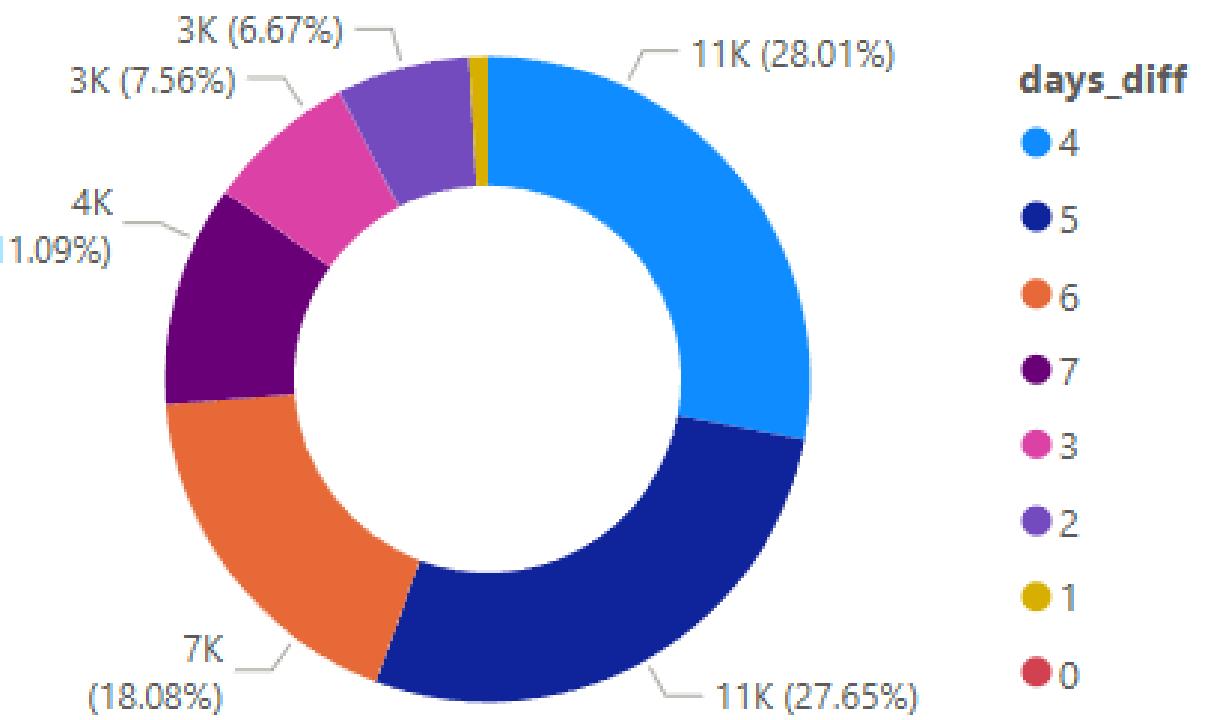
Sum of sales by Year and Quarter
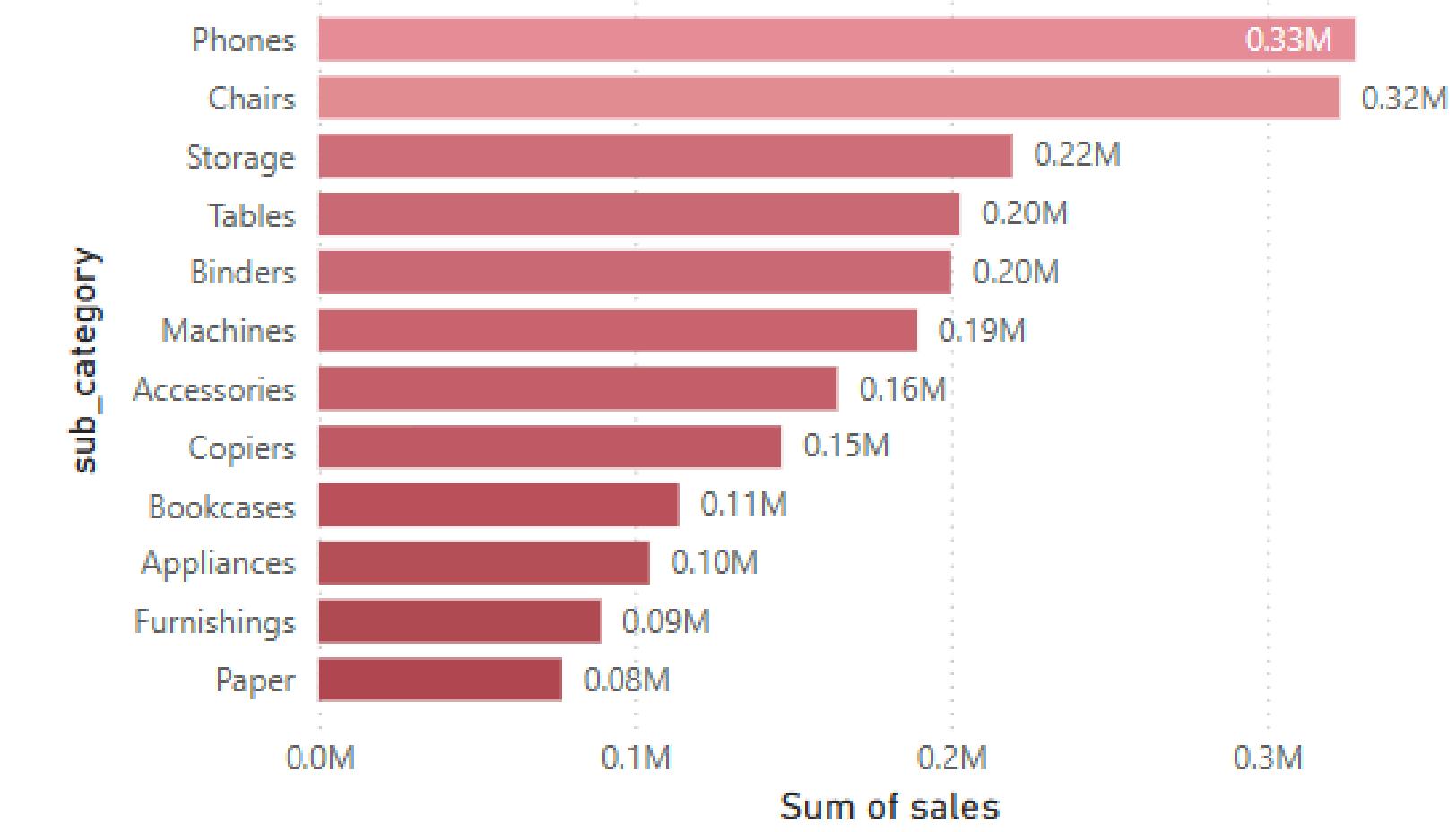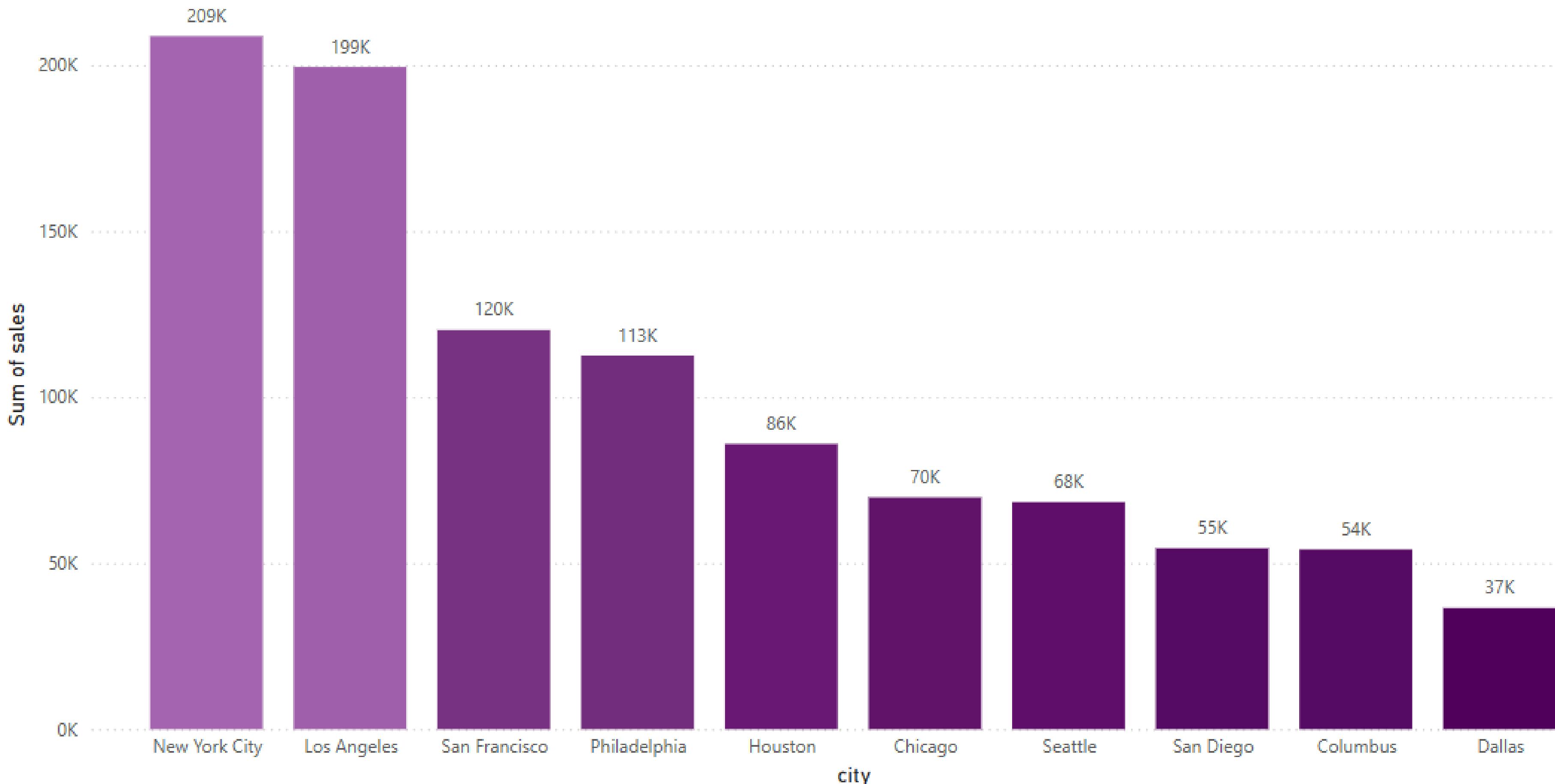
How long It Takes to be Delivered

days_diff
- 4
- 5
- 6
- 7
- 3
- 2
- 1
- 0

Pie chart values:
- 11K (28.01%)
- 11K (27.65%)
- 7K (18.08%)
- 4K (11.09%)
- 3K (7.56%)
- 3K (6.67%)

Customers who make the most Purchase

customer_name	Sum of sales
Sean Miller	25.0K
Tamara Chand	19.1K
Raymond Buch	15.1K
Tom Ashbrook	14.6K
Adrian Barton	14.5K
Ken Lonsdale	14.2K
Sanjit Chand	14.1K
Hunter Lopez	12.9K
Sanjit Engle	12.2K
Christopher Conant	12.1K
Todd Sumrall	11.9K
Greg Tran	11.8K
Becky Martin	11.8K

Categories with the Highest Sales

sub_category	Sum of sales
Phones	0.33M
Chairs	0.32M
Storage	0.22M
Tables	0.20M
Binders	0.20M
Machines	0.19M
Accessories	0.16M
Copiers	0.15M
Bookcases	0.11M
Appliances	0.10M
Furnishings	0.09M
Paper	0.08M

Sum of sales by city

city	Sum of sales
New York City	209K
Los Angeles	199K
San Francisco	120K
Philadelphia	113K
Houston	86K
Chicago	70K
Seattle	68K
San Diego	55K
Columbus	54K
Dallas	37K

Conclusion

Working with Pentaho helps in making visualization and working on data better.

This helps bring in developer and data analyst working better on tools.

Thank you

Any question?