TAKE-HOME ASSIGNMENT: ANALYTICS ENGINEER

Mayah Bosworth

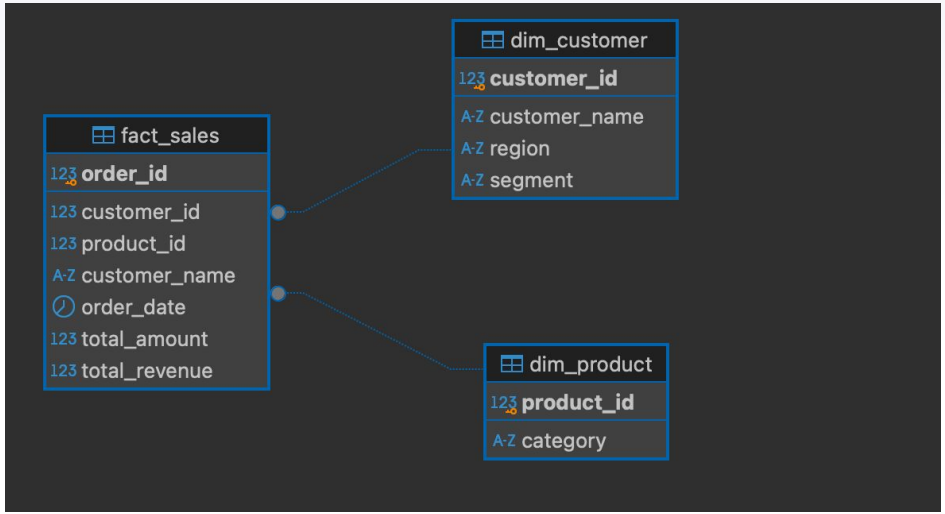
Overview

- Clean raw data
  - Correct data types and format
  - Remove irrelevant columns
  - Remove NULL values
  - Use logic to substitute missing information
- Create dimension and fact tables
  - Implement PKs and FKs to link tables
- Output sales_performance table optimized for tableau reporting

Data Cleaning

customers.csv	• NULL values for customer_id • missing customer_name for some rows • inconsistent region strings • irrelevant columns	• filter out rows with a NULL customer_id using CAST and regex • standardize region strings using LOWER • removed irrelevant columns during table creation
orders.csv	• utf-16le file encoding • NULL values for order_id • invalid dates • irrelevant columns. • some customer_id values did not exist in the customers table	• use pandas to handle the utf-16le encoding • filter out rows where order_id was NULL • correct invalid dates (leap year) • removed irrelevant columns during table creation • added "Unknown" entries to dim_customer for missing customer information
products.csv	• utf-16le file encoding • NULL values for product_id • irrelevant columns	• use pandas to handle the utf-16le encoding • filter out rows where product_id was NULL • removed irrelevant columns during table creation

Schema Diagram



This schema represents a star schema design, where the `fact_sales` table serves as the central fact table. It connects to two dimension tables, `dim_customer` and `dim_product` via foreign key relationships.

Snapshot of sales_performance table:

	order_id	customer_name	customer_id	region	segment	product_id	category	order_month	total_revenue
1	1	Amanda Vang	13	west	Corporate	8	Electronics	1	486.3
2	2	Unknown	999	Unknown	Unknown	12	Furniture	4	26.41
3	3	Heather Pace	20	east	Retail	999	Clothing	4	382.1
4	4	Mary Austin	27	east	Small Business	6	Electronics	2	279.09
5	5	Sherry Jones	15	south	Small Business	8	Electronics	7	338.14
6	6	Mary Austin	27	east	Small Business	12	Furniture	3	405.81
7	7	Cynthia Lee	25	north	Retail	15	Electronics	3	285.68
8	8	Steven Price	5	east	Retail	9	Furniture	5	498.79
9	9	Emily Warren	10	north	Retail	4	Clothing	1	362.35
10	10	Kyle Gonzalez	24	west	Small Business	15	Electronics	12	470.01

Tableau use case examples:

category	total_revenue_generated
Clothing	994,035.37
Electronics	759,462.74
Furniture	610,085.05
Accessories	144,707.84

Determine top performing product categories

order_month	monthly_revenue
1	339,378.35
2	307,354.45
3	182,656.8
4	156,074.4
5	173,217.6
6	204,444
7	221,965.2
8	190,518.3
9	124,531.2
10	156,184.2
11	148,483.8
12	303,482.7

Track monthly revenue trends

segment	total_revenue_generated
Small Business	1,000,028.47
Retail	865,962.36
Corporate	510,932.57
Unknown	131,367.6

Understand which customer segments generate the most revenue