Sales Analysis Project Documentation

Project Objective

To identify key factors influencing sales performance based on historical transactional data, with the goal of uncovering actionable insights to increase sales and improve business strategy.

Dataset Specifications

- Original Dataset: 9800 rows × 18 columns
- Post-Cleaning Dataset: 9792 rows × 14 columns
- Target Variable: price
- Initial Structure: 3 numerical columns, 15 categorical columns
- Assumption: Each row represents a unique transaction for a specific product.

Data Quality Assessment

Completeness

- postal code column had 11 missing values (~0.11%), all from Burlington, Vermont — filled using verified postal code from the web.

Consistency

- Same transactions (by product ID and order ID) had inconsistent price entries → grouped and summed price accordingly.
- Product name inconsistencies were observed for same product IDs → retained only product id and dropped product name.

Uniqueness

- No duplicate records post-cleaning.

Accuracy

- price column has valid and reasonable values throughout.

Data Types

- Converted:
  - order date and ship date to proper datetime format
  - postal code to text format (preserving leading zeros)

Data Cleaning Summary

- Removed unnecessary row id, product name, customer name, and country columns.
- Changed all column names to lowercase for consistency.
- Grouped by product id and order id to sum prices where necessary.
- Rounded sales column to 2 decimal places for clarity.
- Created separate tables: order_details, customer_details, product_details.
- Structured and exported all cleaned tables to Excel for use in Power BI.
- Uploaded final dataset and Power BI dashboard to GitHub.

Power BI Dashboard Overview

The dashboard includes:

- Key metrics: Total Sales, Average Sales, Order Count
- Time-based analysis: Monthly and yearly trends
- Customer segmentation and product performance insights
- City and state-wise sales distribution
- Filtered views by segment, category, and region