Sales Data Analytics and ETL Pipeline Project

Objective:

To build an end-to-end ETL pipeline using Python and SQL that cleans raw sales data, stores it in a SQLite database, and visualizes key business metrics using Power BI.

Tools & Technologies:

- Python (Pandas, sqlite3, matplotlib/seaborn)

- SQL (SQLite)

- Power BI

- GitHub

Skills Demonstrated:

- Data Cleaning

- ETL Pipeline Automation

- Database Management

- Data Visualization

- Business Insight Generation

- Version Control (GitHub)

ETL Steps:

Step 1: Extract

- Loaded sales\_data.csv using Pandas.

Step 2: Transform

- Cleaned missing values.

- Converted Profit to numeric.

- Removed duplicates.

- Standardized column names.

Step 3: Load

- Loaded cleaned data into sales\_data.db (SQLite).

- Verified with SQL queries.

Python ETL Script Summary:

import pandas as pd

import sqlite3

# Load raw CSV

df = pd.read\_csv("sales\_data.csv")

# Data Cleaning

df["Profit"] = pd.to\_numeric(df["Profit"], errors="coerce")

df.dropna(inplace=True)

df.drop\_duplicates(inplace=True)

# Load to SQLite

conn = sqlite3.connect("sales\_data.db")

df.to\_sql("sales\_cleaned", conn, if\_exists="replace", index=False)

conn.close()

SQL Query Example:

SELECT Category, SUM(Profit) AS Total\_Profit

FROM sales\_cleaned

GROUP BY Category;

Power BI Dashboard Visuals:

1. Total Profit by Category – Bar Chart

2. Quantity Sold by Sub-Category – Column Chart

GitHub Repository: https://github.com/Oliviamartin33?tab=repositories

Includes: etl\_script.py, sales\_data.db, screenshots of Power BI dashboard, and this Word doc.

Link: https://github.com/Oliviamartin33