Comprehensive Data Analytics Workflow for a Retail Business Documentation

(by Abnerson Ocampo)

1. Project Overview

This project is part of a comprehensive data analytics learning program, designed to apply key concepts in a real-world scenario. The objective is to simulate an end-to-end data analytics process for a hypothetical retail business, using multiple tools to extract insights and support business decision-making.

Understanding customer purchasing behavior, product demand, and sales trends is crucial for the retail industry. However, working with limited and sparse data present unique challenges. In this analysis, we will explore a very limited dataset with only few customers making purchases, and only a handful of products being bought.

Despite the dataset's limitations, we will apply data cleaning, exploratory data analysis, visualization techniques, and statistical concepts to extract meaningful insights.

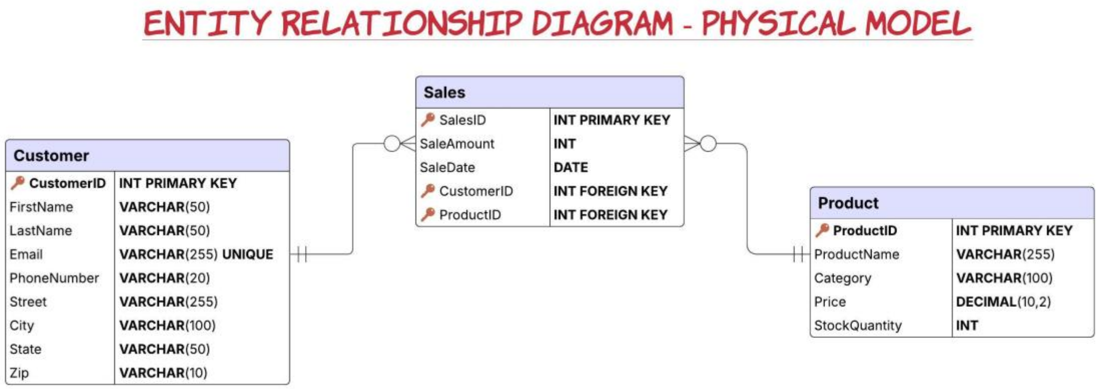
This project will demonstrate how to work with imperfect data. The goal is to extract meaningful insights despite constraints.

Objective: Clean, analyze, and visualize limited sales data to extract meaningful insights.
Tools Used: Python (for data cleaning & transformation), Excel (for pivot tables & dashboard), SQL (for Ad Hoc Customer Behavior Analysis).

2. Data Overview

- **Customers** Table: Contains information about customers.
- **Products** Table: Contains information about the products.
- **Sales** Table: Contains historical sales data.

2.1 Entity Relationship Diagram



ENTITY RELATIONSHIP DIAGRAM - PHYSICAL MODEL

Sales	
SalesID	INT PRIMARY KEY
SaleAmount	INT
SaleDate	DATE
CustomerID	INT FOREIGN KEY
ProductID	INT FOREIGN KEY

Customer	
CustomerID	INT PRIMARY KEY
FirstName	VARCHAR(50)
LastName	VARCHAR(50)
Email	VARCHAR(255) UNIQUE
PhoneNumber	VARCHAR(20)
Street	VARCHAR(255)
City	VARCHAR(100)
State	VARCHAR(50)
Zip	VARCHAR(10)

Product	
ProductID	INT PRIMARY KEY
ProductName	VARCHAR(255)
Category	VARCHAR(100)
Price	DECIMAL(10,2)
StockQuantity	INT

3. Data Cleaning in Python 🐍

Steps Taken:

- Loaded the data using Pandas.
- Cleaned the Customer data
  - Cleaned the Address column(s)
    - Made a copy of the raw data
    - Renamed the unnamed address columns
    - Filled the missing values with blank for consolidation purposes
    - Consolidated the 4 separate address columns
    - Dropped the 4 separate old address columns
    - Split the consolidated address column into 4 columns for deeper analysis
      - Street
      - City
      - State
      - Zip
    - Filled the empty addresses with “Unknown”
    - Dropped the consolidated address column
  - Cleaned the CustomerName column
    - Capitalized every letter of the first word
    - Split into First and Last name
    - Dropped the original column
    - Filled missing values with “Unknown”
  - Cleaned the PhoneNumber column
    - Removed all special characters using regex

        - Applied the (123) 456-7890 format
        - Filled missing values with “Unknown”
    - Cleaned the Email column
        - Assumed that the ideal format is firstlastname@email.com from the data pattern since there is no documentation or stakeholders to refer to
        - Concatenated first name and last name + email.com
        - Filled missing values with “Unknown”
    - Dropped duplicates
    - Dropped unnecessary rows with all missing values
    - Converted the columns to appropriate data type
- Cleaned the Product data
    - Cleaned the ProductName column
        - Made a copy of the raw data
        - Dropped duplicates
        - Filled missing values with “Unknown”
        - Trimmed whitespaces
        - Capitalized first letter of each word
    - Cleaned the Category column
        - Filled missing values with “Unknown”
        - Trimmed whitespaces
        - Capitalized first letter of each word
        - Replaced blanks with “Unknown”
        - Updated missing values using inference since there is no documentation or stakeholders to refer to
    - Cleaned StockQuantity column
        - Filled missing values with 0
        - Converted column to integer data type
    - Cleaned Price column
        - Filtered out rows with missing values since I can’t really guess or impute with mean or median and an “Unknown” placeholder will make the column’s data type an object
        - Made the column of type float
- Cleaned the Sales data
    - Made a copy of the raw data
    - Converted columns to appropriate data type
        - SalesID -> Int64
        - CustomerID -> Int64
        - ProductID -> Int64
        - SaleAmount -> Int64
        - SaleDate -> datetime %Y-%m-%d
    - Handled missing values
        - Dropped rows where SalesID is missing

        - Imputed missing SaleAmount with mean
- Merged the 3 tables (master_data)
- Feature Engineered a new column Revenue = SaleAmount * Price
- Exported all the clean data for further analysis in Excel

4. Analysis & Visualization in Excel 📊

4.1 Pivot Tables:

4.1.1 **Sales Performance by Product** – Total sales per product.

4.1.2 **Revenue by State** – Total revenue by State

4.1.3 **Sales Trend Over Time** – Monthly/weekly analysis of sales patterns.

4.1.4 **Monthly Stock Report** – Stock report of products sold.

4.1.5 **Percentage of Sales by Category**

Dashboard Features:

Filters & Slicers: Allow dynamic filtering by product category, state, and date.

Charts & Visuals: Bar charts, line graph, pie chart, tables (for a deeper and more granular analysis)

Interactivity: Users can explore sales trends easily.

5. Key Insights & Findings 🔍

5.1 **Best-Selling Product**: Laptops had the highest sales.

5.2 **Best-Selling Category:** Electronics is the most dominant category.

5.3 **Sales Patterns:** Sales peaked in May but dropped significantly from July onwards.

5.4 **Highest Revenue State**: Illinois had the highest revenue of all states.

6. Issues

6.1 **Limited Geographic Reach**: Sales transactions were only recorded in **two** states, limiting business growth opportunities

6.2 **Low Customer Engagement**: Only **three** unique customers made purchases, indicating a lack of customer base growth

6.3 **Restricted Product Sales**: Out of the wide range of available inventory, only **three** distinct products were sold, highlighting a potential issue with product demand or visibility, leaving most inventory unsold

6.4 A **decline** occurred in May with a particular sharp drop in September

7. Recommendations & Action Plan

7.1 Expand Market Research

- Conduct market research to identify potential regions for expansion
- Start small, scale fast
- Begin with 2-3 expansion states to test demand
- Focus on e-commerce and digital marketing to minimize upfront costs

7.2 Improve Customer Acquisition & Retention

- Launch a customer loyalty program
- Improve advertising strategies to reach a broader audience
- Use email marketing and personalized offers

7.3 Optimize Product Offerings

- Analyze customer preferences and pricing strategy to promote underperforming products
- Offer bundled deals or discounts on less popular items to increase sales

7.4 Analyze Customer Preferences

- Identify why certain categories remain unsold—are they priced too high, not well marketed, or not relevant to the target audience?
- Conduct customer surveys to understand demand for other product categories
- Implement strategic discounts

7.5 Address what caused the sales decline after May and especially in September

- **Possible causes**
  - Seasonality
  - Pricing
  - Competition
  - Inventory Issues
- **Marketing Boost during low performing months and to expand product sales**
  - Discounts
  - Promotions

  - Targeted Ads

8. Ad Hoc Customer Behavior Analysis in SQL

Customer behavior analysis helps businesses understand spending patterns, purchase frequency, and retention trends. Using SQL, we can generate ad hoc reports to analyze customer data dynamically.

8.1 Key Questions Answered:

- Identified Best-Selling Products
- Identified High-Value Customers
- Identified Inactive Customers Who May Need Re-Engagement
- Found the States with the Highest Revenue
- Identified the Number of Returning Customers per Month