Data Analysis Project Requirements:

Section 1: Business Rules

Customer Data

- Each CustomerID must be unique; duplicates must be resolved by keeping the most recent SignupDate or merging records.
- Email addresses must be valid (contain "@" and a domain “cuzies.org”); null or invalid emails should be flagged.
- State must be standardized to two-letter codes (e.g., "California" → "CA").
- SignupDate must be in YYYY-MM-DD format; invalid or null dates should be set to a default (e.g., "1900-01-01") or excluded.

Order Data

- OrderID must be unique; duplicates should be removed, keeping the first occurrence.
- CustomerID and ProductID must exist in the Customers and Products tables, respectively; invalid IDs should be flagged.
- Quantity must be positive; negative or zero values should be corrected or excluded.
- TotalAmount must equal Quantity * Product.Price; discrepancies should be recalculated or flagged.
- OrderDate must be in YYYY-MM-DD format; missing dates should be set to a default or excluded.

Product Data

- ProductID must be unique; duplicates should be resolved by keeping the record with a valid Price.
- ProductName must not be null; missing names should be set to "Unknown Product".
- Category must be standardized (e.g., "Electronics", "Home Appliance", "Tools").
- Price must be positive; negative prices should be set to a default (e.g., median price) and unknown prices excluded.

Section 2: Naming Conventions

Database Tables

- Use singular nouns: Customer, Order, Product.
- Use PascalCase for table names.

Columns

- Use PascalCase for column names (e.g., CustomerID, OrderDate).
- Prefix columns with the table context where ambiguous (e.g., Customer_City in a joined table).

Staging and Final Tables

- Staging tables (pre-ETL): Prefix with stg_ (e.g., stg_Customer).
- Final tables (post-ETL): Prefix with dim_ for dimension tables, agg_ for aggregated tables and facts_ for facts tables (e.g., dim_Customer).

Database Schema

- Store raw data in bronze schema.
- Store cleaned, standardized, transformed data in silver schema.
- Store analysis-ready data in gold schema.
- Use views for analysis.

Section 3: Data Preparation and Transformation for Sales Analysis

Objective

Prepare and transform the customers, orders, and products datasets into a clean, integrated, and analysis-ready data model. The goal is to support accurate, meaningful business insights through visualizations and reporting.

Specifications

Data Sources
- Input: Three unclean CSV datasets: customers.csv, orders.csv, and products.csv.
- Each dataset contains raw transactional or reference data extracted from separate operational systems.

Data Cleaning
- Handle missing, inconsistent, and duplicate data across all tables.
- Normalize data formats (e.g., date/time formats, phone numbers, address formatting).
- Standardize categorical variables and remove anomalies (e.g., invalid product IDs, out-of-range quantities).

Data Transformation
- Generate a clean and structured model with proper relational joins (e.g., orders linked to customers and products).
- Derive additional fields useful for analysis (e.g., total order value, order frequency per customer).
- Convert raw fields into meaningful categories (e.g., segment customers based on spending behavior).

Data Integration
- Merge the three datasets into a unified, flat reporting table or star schema that supports fast querying.
- Ensure referential integrity across joined datasets.

Scope
- Focus on the most recent and relevant subset of data if the datasets span multiple years or contain historical artifacts.
- Exclude archival or irrelevant records unless necessary for current analysis.

Analytical Readiness
- Ensure the final dataset is suitable for dashboarding tools (e.g., Power BI, Tableau) or statistical analysis in Python/R.
- Include clearly defined column names and data types.

Deliverables
- A final clean dataset (CSV or Excel) ready for analysis.
- A summary report or data dictionary describing:
- - Cleaning steps
- - Field definitions
- - Data quality issues encountered and how they were resolved

BI: Analytics and Reporting (Data Analysis)
Objective: Develop SQL-based analytics to deliver detailed insights into:

- Customer Behavior
- Product Performance
- Sales Trends

These insights empower stakeholders with key business metrics, enabling strategic decision-making.