How the Data Flows Phase-by-Phase

🔹 **Phase 1 – Raw Data → Cleaned Data**

Input:

- Raw RetailSmart datasets
  (customers.csv, products.csv, sales.csv, marketing.csv, reviews.csv)

Output (created by learners):

- customers_cleaned.csv
- sales_cleaned.csv
- marketing_cleaned.csv
- products_cleaned.csv

📌 **These cleaned datasets become the official inputs for all future phases.**

🔹 **Phase 2 – Predictive Modeling (Uses Cleaned Data)**

Input:

- Cleaned CSVs from Phase 1

Output:

- model_input.csv
- final_rf_model.pkl
- scaler.pkl

📌 In this phase, learners **merge and integrate** cleaned data to create **model_input.csv**, which is the modeling-ready dataset.

This dataset is then used for training churn prediction models.

🔹 **Phase 3 – Advanced Analytics (Also Uses Cleaned Data)**

Input:

- Cleaned data from Phase 1
- Optional: modeling features from Phase 2 (RFM, aggregated metrics)

Output:

- customers_with_clusters.csv
- cluster_summary.csv
- forecast_results.csv

📌 Clustering and forecasting also rely on the **cleaned datasets** from Phase 1.
Phase 2's dataset can optionally enhance segmentation.

🔹 **Phase 4 – Power BI (Uses ALL Outputs from Previous Phases)**

Input:

- Cleaned datasets (Phase 1)
- Model predictions & feature sets (Phase 2)
- Clustering & forecasting outputs (Phase 3)

Output:

- RetailSmart_Dashboard.pbix
- RetailSmart_Storytelling_Report.docx

📌 This phase is a **consumption phase**, pulling all CSVs + model outputs into Power BI.