Novartis Datathon – Plain-Language Explanation

This document explains **what the Datathon wants**, **what you must do**, and **pharma vocabulary**, based on both provided instruction files.

---

1. What the Datathon Wants (In Plain Language)

The task is to **predict how much a drug's sales will fall after generics enter the market**.
When a drug's patent expires, cheaper copy versions (called **generics**) appear.
Sales of the original brand usually **drop sharply**—this drop is called **generic erosion**.

The Datathon wants you to:

1. Build models that **forecast monthly sales** for the 24 months after generic entry.
2. Make predictions for two situations:
   - **Scenario 1**: Right when generics arrive (0 months of post-generic data).
   - **Scenario 2**: Six months later (you already know 6 months of real post-entry sales).
3. Pay special attention to **high-erosion drugs** (big sales crashes).
4. Explain your work clearly with EDA, preprocessing, modeling rationale, and visuals.
5. Follow the rules for submissions, file formats, and deadlines.

In short:

> Your job = **Predict the sales cliff** after generics arrive, and explain your methods well.

---

2. Step-by-Step Breakdown of What You Should Do

Step 1 – Understand the Business Story

- A drug launches → grows → stabilizes.
- Patent expires → **Loss of Exclusivity (LoE)**.
- Generics enter → **sales of the original drop**.
- Novartis needs accurate forecasts to plan budgets and strategy.

Step 2 – Understand the Provided Data

You get **three datasets**:

1. Volume Dataset

Contains sales history:

- `country`
- `brand_name`
- `month`
- `months_postgx` (month relative to generic entry)
- `volume` (units sold)

You must **predict** the `volume`.

2. Generics Dataset

Shows market competition:

- `n_gxs` = number of generics available that month.

3. Medicine Info Dataset

Static product features:

- therapeutic area
- hospital rate
- package type
- biological or small molecule

Step 3 – Generic Erosion and Buckets

Mean Generic Erosion = average normalized volume during 24 months after generics enter.

Buckets:

- **Bucket 1**: High erosion (0–0.25). Very important.
- **Bucket 2**: Everyone else.

Bucket 1 is **double weighted** during scoring.

Step 4 – Two Forecasting Scenarios

1. **Scenario 1**
    - Predict months **0–23**.
    - No post-entry data provided.
2. **Scenario 2**
    - Predict months **6–23**.
    - Actual months **0–5** provided.

Step 5 – Build Models

You must:

- Handle missing values and inconsistent units.
- Create features (lags, trends, # of generics).
- Explore data visually.
- Build time-series or ML models.
- Explain *why* you chose your approach.

Step 6 – The Metric (How You're Scored)

They use a **custom Prediction Error (PE)** with:

- Monthly errors,
- Period-sum errors (0–5, 6–11, 12–23),
- Normalization by pre-generic volume.

Bucket 1 scores count **twice**.

Lower PE = better.

Step 7 – Competition Flow

- Join Microsoft Teams.
- Get data in the "Files" section.
- Build and refine models.
- Submit CSV predictions (3 per 8 hours).
- Leaderboard ranks based on the public test set.
- Before the deadline, choose **two final submissions**.
- Final scoring uses a private test set.
- Top 10 → Top 5 → finalists present to the jury.

---

3. Pharma & Datathon Vocabulary Explained

Brand / Brand Name – Original drug (e.g., "Diovan").
Generic Drug – Cheaper copy with same medical effect.
Active Ingredient – The actual chemical treating the condition.
Loss of Exclusivity (LoE) – Patent expiration date.
Generic Entry / Month 0 – First month generics appear.
Generic Erosion – Sales drop after generics enter.
Mean Generic Erosion – Average normalized sales months 0–23.
Bucket 1 – High-erosion drugs.
Bucket 2 – Lower erosion.
Therapeutic Area – Disease category.
Biological Drug – Made from living cells.
Small Molecule – Classic chemical drug.
Hospital Rate – Fraction of sales through hospitals.
Volume – Units sold.
Scenario 1 / Scenario 2 – Two forecasting situations.
Prediction Error (PE) – Custom scoring metric.
Public vs Private Test Set – Public for leaderboard; private for final results.

---

4. Summary

They want you to:

- **Predict post-generic sales erosion**,
- **Model high-erosion cases well**,

- **Submit correct CSV forecasts**,
- **Explain your approach clearly**,
- And **present** if you reach the final.