Brand & Month Consistency Check

Overview

This document verifies that the submission template, test data, and training data follow consistent rules for the two prediction scenarios.

Submission Template Structure

The template file (`submission_template.csv`) contains **7,488 rows** for **340 brands**.

Scenario	Start Month	End Month	Brands	Rows per Brand	Total Rows
Scenario 1	0	23	228	24	5,472
Scenario 2	6	23	112	18	2,016
Total	-	-	**340**	-	**7,488**

Key Insight

The template combines **both scenarios** into a single file:

- Some brands need months **0-23** predicted (Scenario 1)
- Other brands need months **6-23** predicted (Scenario 2)

Test Data Availability

Volume Test Data (`df_volume_test.csv`)

Scenario	Brands	Available Months	Max Month	Data Available
Scenario 1	228	-24 to -1	**-1**	Pre-entry only
Scenario 2	112	-24 to 5	**5**	Pre-entry + 6 months post-entry

Verification Results

```
Scenario | Test Max Month | Count | Rule
---------|----------------|-------|-----
   1     |       -1       |  228  | ☑ Only pre-entry data
   2     |        5       |  112  | ☑ Pre-entry + months 0-5
```

☑ TEST DATA matches TEMPLATE exactly!

The test data availability determines which scenario each brand belongs to:

- Brands with data only up to month **-1** → Scenario 1 (predict 0-23)
- Brands with data up to month **5** → Scenario 2 (predict 6-23)

Training Data Structure

Volume Train Data (`df_volume_train.csv`)

Metric	Value
Total Brands	1,953
Min Month (all brands)	-24
Max Month (all brands)	23
Complete Data	☑ Yes

All 1,953 training brands have COMPLETE data from month -24 to 23.

This allows you to:

1. Train models that learn the full erosion pattern
2. Simulate both scenarios during validation
3. Compute `avg_vol` (months -12 to -1) and `bucket` labels

Brand Overlap Analysis

Dataset	Unique Brands
Train	1,953
Test	340
Overlap	**0**

Train and test have completely separate brands!

This means:

- You cannot use test brand history from training
- Models must generalize to unseen brands
- Features should capture general erosion patterns, not brand-specific memorization

---

Scenario Definitions

Scenario 1: "0 Actuals" (Phase 1-a)

> **Situation**: Generic just entered the market. You only have pre-entry sales data.

- **Available data**: Months -24 to -1 (pre-entry)
- **Predict**: Months 0-23 (24 months)

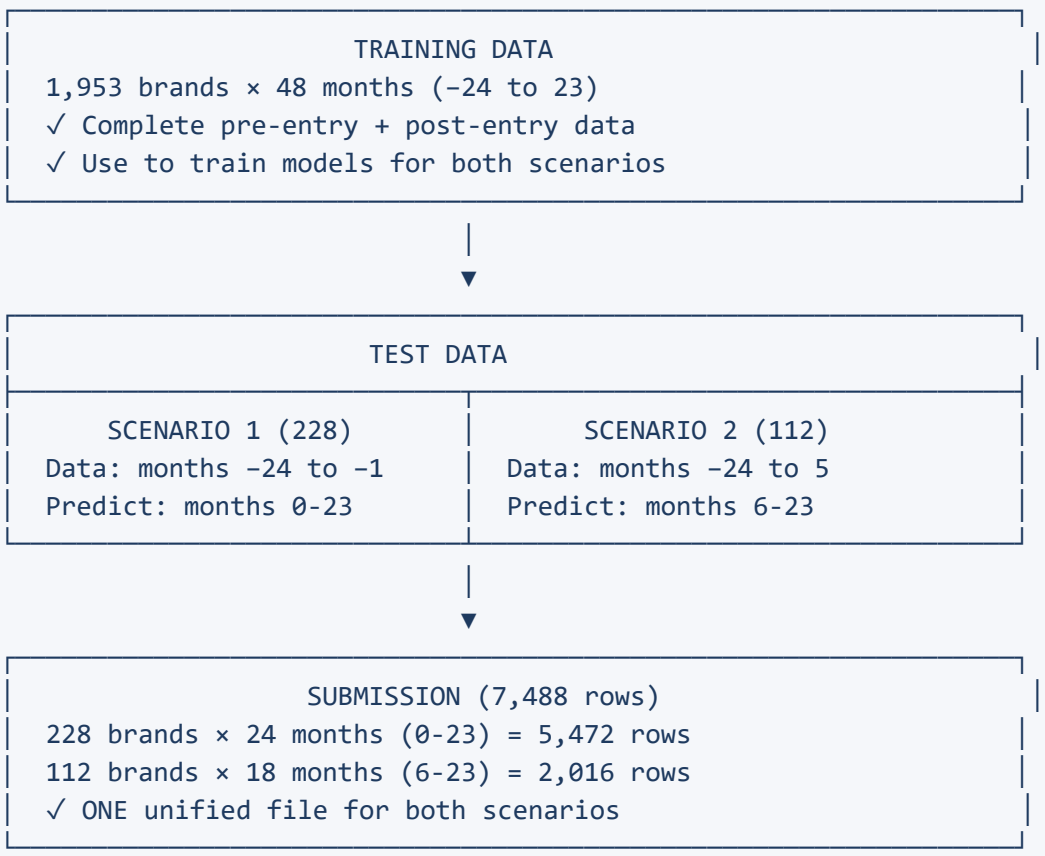
- **Test brands**: 228
- **Metric weights**:
  - 20% monthly errors (months 0-23)
  - 50% cumulative error (months 0-5) ← **Heavily weighted!**
  - 20% cumulative error (months 6-11)
  - 10% cumulative error (months 12-23)

Scenario 2: "6 Actuals" (Phase 1-b)

> **Situation**: Generic entered 6 months ago. You have early post-entry sales.

- **Available data**: Months -24 to 5 (pre-entry + 6 months post-entry)
- **Predict**: Months 6-23 (18 months)
- **Test brands**: 112
- **Metric weights**:
  - 20% monthly errors (months 6-23)
  - 50% cumulative error (months 6-11) ← **Heavily weighted!**
  - 30% cumulative error (months 12-23)

Data Flow Diagram



Validation Rules for Submission

☑ Must Match Template Exactly

1. **Same brands**: All 340 (country, brand_name) pairs
2. **Same months per brand**:
   - Scenario 1 brands: months 0-23
   - Scenario 2 brands: months 6-23
3. **Same row count**: 7,488 total rows
4. **No NaN values** in volume column
5. **Positive volumes** (volume > 0)

❌ Common Mistakes

Mistake	Problem
Separate files for S1 and S2	Should be ONE unified file
All brands with months 0-23	S2 brands should start at month 6
Missing months for some brands	Must match template exactly
Extra rows	Must match template row count

Quick Reference

```
# Check which scenario a brand belongs to
template = pd.read_csv('submissions/submission_template.csv')
brand_start = template.groupby(['country', 'brand_name'])
['months_postgx'].min()

# Scenario 1: start_month == 0
# Scenario 2: start_month == 6
```

```
# Generate correct submission
template = pd.read_csv('submissions/submission_template.csv')
submission = template.copy()
submission['volume'] = your_predictions  # Match row order!
submission.to_csv('submission_final.csv', index=False)
```

Summary

Check	Status
Template has correct structure (S1: 0-23, S2: 6-23)	☑

Check	Status
Test data availability matches template scenarios	☑
Train data has complete 48-month history	☑
Train and test brands are separate (no overlap)	☑
Submission should be ONE unified file	☑

The data is consistent. Your submission must match the template structure exactly!