PriceOptima – Dataset Validation Report (Milestone 1)

Prepared by: *Kareemunnisa*
Project: *PriceOptima – Dynamic Pricing System*
Milestone: **Requirements & Data Preparation**

Introduction

The objective of Milestone 1 is to prepare and validate the datasets required for developing the PriceOptima Dynamic Pricing System.
The following datasets were collected, cleaned, and validated using Python and Pandas scripts:

1. **Sales Dataset**
2. **Inventory Dataset**
3. **Competitor Pricing Dataset**
4. **Product Master Dataset**

Each dataset was checked for mandatory fields, data consistency, quality, and usability for downstream pricing analysis and KPI calculations.

Step 1 – Field Validation

We validated whether each dataset contained all required fields.
A Python script was used to load the datasets and confirm the presence of required columns:

Script Used (Step 1)

```
def check_fields(df, required_fields):

    missing = [field for field in required_fields if field not in
df.columns]

    if missing:

        return f"Missing: {missing}"

    return "All required fields present.
```

This script allowed us to inspect raw field names and map/rename them to the required schema.

1.1 Sales Dataset – Field Validation

Required Fields:

- Date
- Product ID
- Units Sold
- Price
- Revenue

Actions performed:

- Loaded the cleaned `sales_dataset.csv`
- Verified the presence of renamed fields:
  - `Date`
  - `Product ID`
  - `Units Sold`
  - `Price`
  - `Revenue`

Result: *All required fields are present and correctly formatted.*

1.2 Inventory Dataset – Field Validation

Required Fields:

- Product ID
- Stock Level
- Restock Date
- Warehouse/Store ID

Actions performed:

- Extracted from `retail_store_inventory.csv`
- Generated Restock Date using `Date + 7 days` logic

Result: *All required fields exist in the final dataset.*

1.3 Competitor Pricing Dataset – Field Validation

Required Fields:

- Product ID
- Competitor Price
- Competitor Name

Actions performed:

- Loaded Flipkart dataset
- Applied logic:
  - `Competitor Price = discounted_price if available else retail_price`
  - `Competitor Name = brand`

Result: *All fields are valid.*

1.4 Product Master Dataset – Field Validation

Required Fields:

- Product Name
- Category
- Cost Price

Actions performed:

- Derived from inventory dataset using unique Product IDs
- Mapped Category from dataset
- Used Product ID as Product Name (dataset had no descriptive name field)

Result: *All required fields are present.*

Step 2 – Data Quality Checks

A Python script was used for systematic quality checks:

Script Used (Step 2)

```
for name, df in datasets.items():
    print(f"\n--- {name} ---")
    print("\nMissing Values:\n", df.isnull().sum())
    print("\nData Types:\n", df.dtypes)
    print("\nDuplicate Rows:", df.duplicated().sum())
    print("\nBasic Statistics:\n", df.describe(include='all'))
```

This script validated datatypes, missing values, duplicates, and extreme values.

2.1 Missing Values Check

- Sales Dataset → No missing values in required columns

- Inventory Dataset → No missing values in required fields
- Competitor Pricing Dataset → Some missing brand values (acceptable; not required)
- Product Master Dataset → No missing values after transformation

Result: *All mandatory fields are complete.*

2.2 Duplicate Check

- Sales Data → No unexpected duplicates
- Product Master → Deduplicated to **21 unique products**
- Inventory → Multiple rows per product (valid, since they represent daily stock snapshots)

Result: *Dataset duplication follows expected patterns.*

2.3 Outlier Check (Basic)

- Price values fall in a realistic range
- Stock levels are reasonable
- Competitor prices are logical
- No negative or zero pricing

Result: *No critical outliers detected.*

Step 3 – Validation Summary

The datasets meet the requirements for:

- Dynamic pricing feature engineering
- Revenue and margin KPI measurement
- Inventory turnover calculations
- Competitor benchmarking
- Product cost and profitability modeling

Each dataset was successfully cleaned, validated, and transformed to match the required schema for Milestone 1.

Conclusion

The dataset preparation phase is complete.
All four datasets are validated, consistent, and aligned with the PriceOptima system requirements.