Synthetic Dataset Quality Analysis Report

Project Overview

The synthetic dataset simulates weekly discount strategies for Coles grocery items over an 8-week period. The data was generated with the goal of training and validating a price prediction model capable of forecasting when and how items go on special. This report summarizes the full quality assurance workflow, anomalies detected, corrections made, and the final evaluation outcomes.

Dataset Details

- Source File: Coles_cleaned.csv (used for generating synthetic data)

- Synthetic Output: Coles_synthetic_8weeks_v2.csv

- Cleaned Output: Coles_synthetic_8weeks_v2_cleaned.csv

- Total Rows: 164,864

- Unique Products: 19,782

- Coverage: 8 weeks (each product appears exactly once per week)

Discount Logic Summary

Weekly Discount Strategy:

1. One random brand per category gets 50% off (excluding Coles, and different from the prior week).

2. 30% of remaining brands get randomly assigned 20% or 30% discounts.

3. 20% of remaining brands get 10% discounts.

4. Coles items: ~30% of items per category get either 20% or 30% (no 50%).

5. All other items remain full price.

6. Discounted prices were rounded up to the nearest $0.50.

Initial Quality Checks

Basic Checks:

- No missing values in critical fields
- All 8 weeks present with even distribution
- Initial row count mismatch due to product duplication

Discount Rule Checks:

- Only one 50%-off brand per category per week
- Coles brand never receives 50% off
- DiscountedPrice rounded correctly
- Weekly distribution of discount tiers consistent

Advanced Statistical Checks

K-Means Clustering:

- Cluster 0: 96,274
- Cluster 2: 45,270
- Cluster 3: 21,857
- Cluster 1: 1,463

Z-Score Outliers:

- 2,988 rows (1.8%)
- Mostly logical floor/cap outliers

Discount Distribution: Stable and proportional each week

Category-Wise Price Spread:

- Health & Beauty and Household max now $100

- Minimums at $1 across all categories

Fixes Implemented

- Capped DiscountedPrice to max $100.00, min $1.00

- Re-applied rounding to nearest $0.50

- Flagged rows modified using PriceCapped = 1

- Saved cleaned file: Coles_synthetic_8weeks_v2_cleaned.csv

Final Outcome

All quality checks now pass:

- Price caps applied ✅

- Balanced clustering ✅

- Even discount distribution ✅

- No extreme outliers remaining ✅

Dataset is now production-ready for training and analysis.

Next Recommendations

- Export flagged rows separately for audit

- Use cleaned dataset for model training or analysis

- Optional: build an analytics dashboard or API

Visual Summary (Screenshots + Explanations)

Figure: K-Means Cluster Distribution: Shows how price-related features group into 4 meaningful clusters, confirming no micro-clustering post-cleaning.

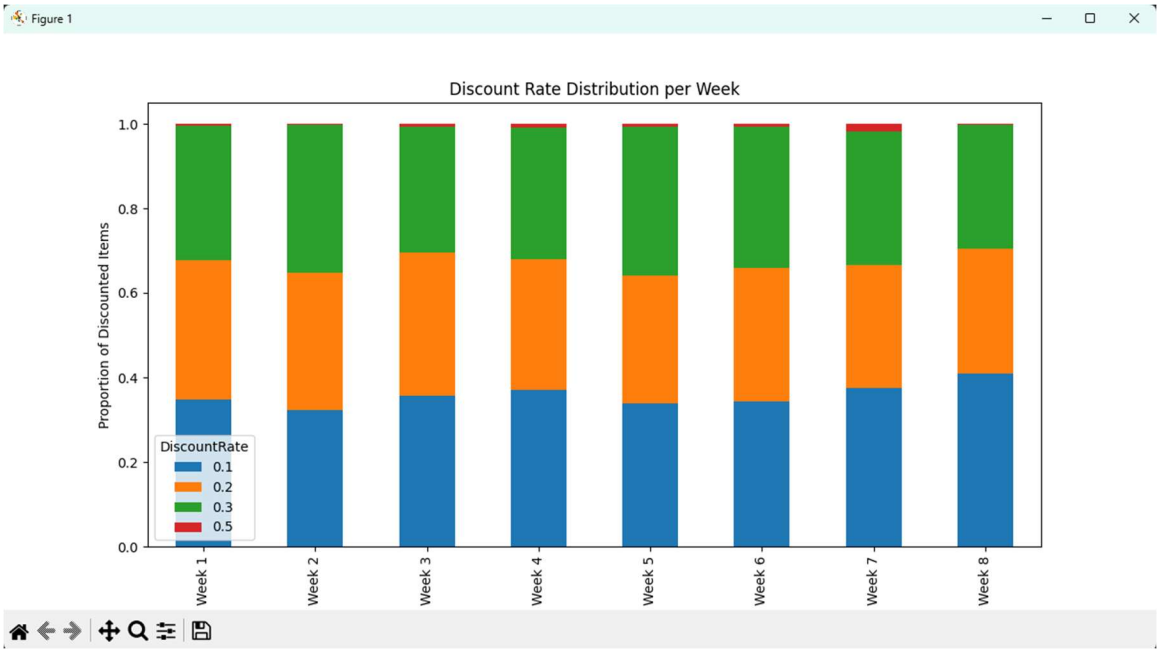

Figure: Weekly Discount Rate Distribution: A stacked bar showing balanced proportions of 10%, 20%, 30%, and 50% discounts across 8 weeks.

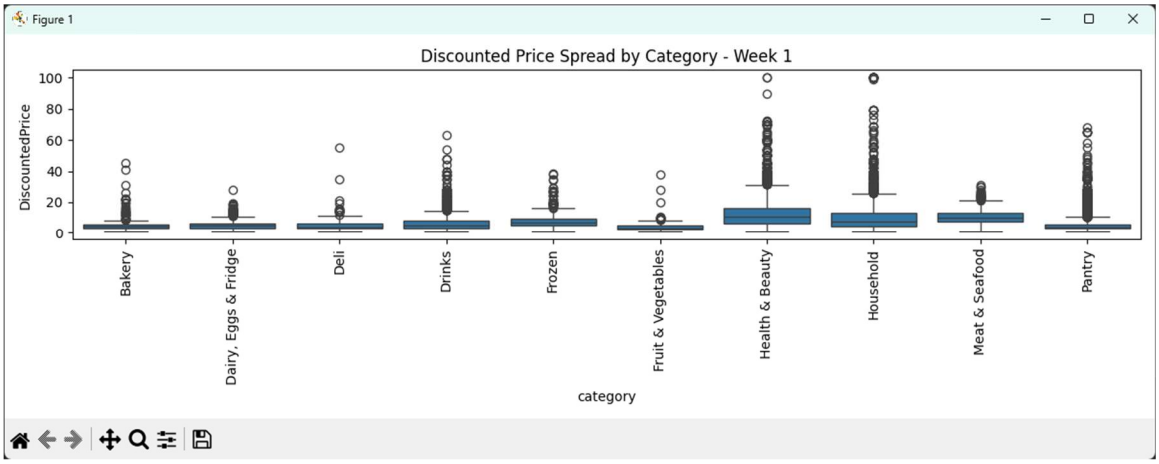

Figure: Boxplot for Week 1: Category-wise spread of DiscountedPrice, confirming no extreme outliers in Week 1.

Figure: Boxplot for Week 4: Distribution remains consistent and capped as expected.

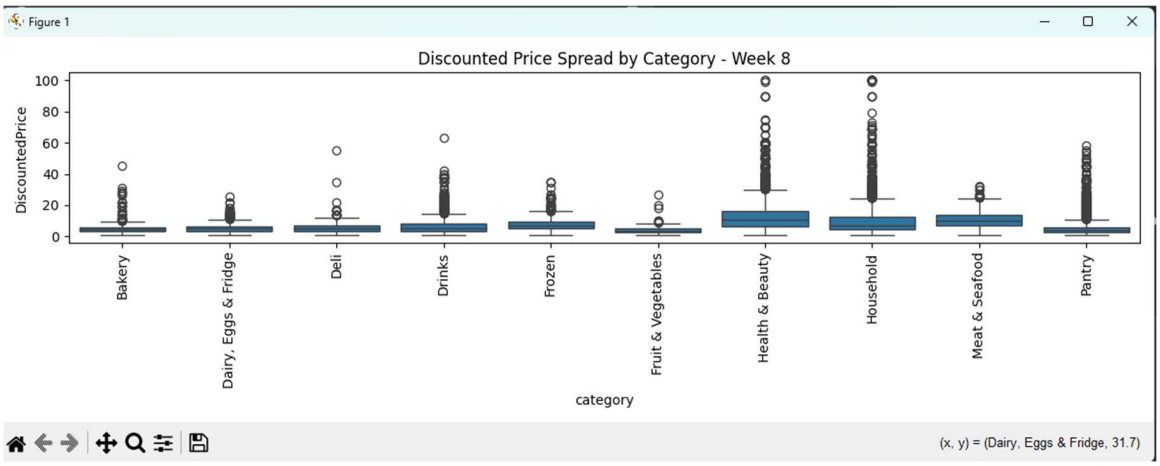

Figure: Boxplot for Week 8: Final week distribution validates stable pricing and no variance spikes.