Anexo 2.4 — Informe de EDA (ComparaYa)

*Versión 0.2 · Fecha: 2025-09-17 · Autores: Equipo ComparaYa*

# 1. Objetivo

Resumir hallazgos del análisis exploratorio de datos (EDA) usando muestras 2025 de datasets Consumidor y Mayorista, para respaldar dashboards, comparador y reglas de normalización.

# 2. Dataset: Consumidor

Filas: 231035 · Columnas: 15 · Fecha: 2024-12-30 → 2025-09-01 (freq: diaria) · Precio: 'Precio promedio'

## 2.1 Nulos por Columna (Top 10)

|  |  |
| --- | --- |
| Columna | % Nulos |
| Anio | 0.0% |
| Mes | 0.0% |
| Semana | 0.0% |
| Fecha inicio | 0.0% |
| Fecha termino | 0.0% |
| ID region | 0.0% |
| Region | 0.0% |
| Sector | 0.0% |
| Tipo de punto monitoreo | 0.0% |
| Grupo | 0.0% |

## 2.2 Duplicados (estimación rápida)

Registros duplicados exactos: 0

## 2.3 Cobertura y Cardinalidades

Productos: 286 · Regiones: 9 · Unidades: 19

## 2.4 Unidades más frecuentes (muestra)

|  |  |
| --- | --- |
| Unidad | Conteo |
| $/kilo | 162348 |
| $/unidad | 39897 |
| $/envase 1 kilo | 5858 |
| $/bandeja 12 unidades | 3616 |
| $/bandeja 30 unidades | 2543 |
| $/Caja de 1 Litro | 2444 |
| $/botella 900 ml | 2441 |
| $/bolsa 800 grs | 2414 |
| $/pan de 250 gramos | 1221 |
| $/envase 400 gramos | 1219 |
| $/caja 180 unidades | 1216 |
| $/pote 500 gramos | 1215 |
| $/litro | 1210 |
| $/bolsa 1 kilo | 1203 |
| $/kilo (en envase de 1 kilo) | 748 |

## 2.5 Diagnóstico de Separadores en 'producto'

Coincidencias (muestra): {" ' | ' ": 0, " '|' ": 152}. Sugerencia: normalizar al separador canónico '|'.

## 2.6 Outliers (IQR) en Precio

Umbrales IQR: lo=-8524.29, hi=17807.14. % de outliers estimado: 1.47%.

## 2.7 Recomendaciones

• Normalizar separadores en 'producto' (regex '\s\*\|\s\*' → '|') y tokenizar a base/calidad/presentación.

• Construir tabla de equivalencias de unidades (ej.: 'caja 18 kg' → factor 18 a kilo).

• Preferir columna de 'promedio ponderado' cuando exista; etiquetar precio\_tipo y flag\_ponderado.

• Publicar granularidad (semanal/mensual) según frecuencia estimada.

# 2. Dataset: Mayorista

Filas: 140003 · Columnas: 14 · Fecha: 2025-01-02 → 2025-09-05 (freq: diaria) · Precio: 'Precio promedio ponderado'

## 2.1 Nulos por Columna (Top 10)

|  |  |
| --- | --- |
| Columna | % Nulos |
| Fecha | 0.0% |
| ID region | 0.0% |
| Region | 0.0% |
| Mercado | 0.0% |
| Subsector | 0.0% |
| Producto | 0.0% |
| Variedad / Tipo | 0.0% |
| Calidad | 0.0% |
| Unidad de comercializacion | 0.0% |
| Origen | 0.0% |

## 2.2 Duplicados (estimación rápida)

Registros duplicados exactos: 0

## 2.3 Cobertura y Cardinalidades

Productos: 81 · Regiones: 9 · Mercados: 12 · Unidades: 145

## 2.4 Unidades más frecuentes (muestra)

|  |  |
| --- | --- |
| Unidad | Conteo |
| $/unidad | 12546 |
| $/bandeja 18 kilos | 10734 |
| $/saco 25 kilos | 6016 |
| $/bandeja 10 kilos | 5677 |
| $/kilo (en caja de 17 kilos) | 5677 |
| $/malla 18 kilos | 5629 |
| $/caja 20 kilos | 5206 |
| $/bandeja 15 kilos granel | 5195 |
| $/bins (400 kilos) | 5115 |
| $/caja 18 kilos | 4953 |
| $/caja 15 kilos | 4169 |
| $/bandeja 18 kilos granel | 3204 |
| $/caja 15 unidades | 2983 |
| $/bins (450 kilos) | 2898 |
| $/malla 25 kilos | 2498 |

## 2.5 Diagnóstico de Separadores en 'producto'

Coincidencias (muestra): {" ' | ' ": 0, " '|' ": 0}. Sugerencia: normalizar al separador canónico '|'.

## 2.6 Outliers (IQR) en Precio

Umbrales IQR: lo=-17250.00, hi=40750.00. % de outliers estimado: 6.84%.

## 2.7 Recomendaciones

• Normalizar separadores en 'producto' (regex '\s\*\|\s\*' → '|') y tokenizar a base/calidad/presentación.

• Construir tabla de equivalencias de unidades (ej.: 'caja 18 kg' → factor 18 a kilo).

• Preferir columna de 'promedio ponderado' cuando exista; etiquetar precio\_tipo y flag\_ponderado.

• Publicar granularidad (semanal/mensual) según frecuencia estimada.