DATA CLEANING LOG (BITÁCORA)
Reto MA2003B

November 26, 2025

Project Information

- Course: Aplicación de métodos multivariados en ciencia de datos
- Dataset Source: CSV files located in `data_raw/`

1. Initial Loading and Inspection

- The datasets `categorias`, `clientes`, `metodos_pago`, `productos`, and `ventas` were imported.
- Structure, summary statistics, and missing values were examined using:
  - `head()`
  - `info()`
  - `describe()`
  - `isna().sum()`
- Result: No missing values were found in any dataset.

2. Duplicate Record Handling

- Duplicate checks were performed on all tables.
- Results:
  - `ventas`: 29 duplicate entries detected and removed

  - All other tables: 0 duplicates
- Outcome: All datasets now contain unique records.

3. Data Type Corrections

- Issues identified:
  - Dates stored as text
  - Monetary values with comma decimal separators
- Corrective actions:
  - Converted `Fecha` and `Fecha_Resgistro` to datetime format
  - Converted `Precio_Unitario` to numeric after replacing commas
- Result: All involved fields now have proper data types.

4. Outlier Detection (IQR Method)

- Variables evaluated:
  - `Precio_Unitario`
  - `Stock`
  - `Cantidad`
- Findings:
  - One price outlier detected in `Precio_Unitario`
  - No outliers detected in `Stock` or `Cantidad`
- Decision: Outlier retained, as it represents a realistic high-value product.

5. Construction of Master Dataset

- Performed a progressive merge of all datasets using appropriate keys.
- Created derived variables:
  - Year (`anio`)

  - Month (`mes`)
  - ISO week number (`semana`)
  - Revenue per transaction (`ingreso = cantidad * precio_unitario`)
- Verified structural consistency after merging.

6. Final Export

- Exported final cleaned dataset in Parquet format:

`data_cleaned/master.parquet`

Final Status Summary

- No missing values remain
- No duplicate records
- Data types corrected
- Outliers analyzed and validated
- Fully merged analytical dataset created
- Ready for EDA, clustering, and predictive modeling