Retail Data Warehouse - Data Quality Report

1. COMPLETENESS

COMPLETENESS

The warehouse consists of three core tables:
- STORES (dimension): 45 rows, 3 columns
- FEATURES (time and regional features): 8190 rows, 12 columns
- SALES (fact table): 421570 rows, 5 columns

Missing values were analyzed for each table.

For STORES, no missing values were found in any column.

Top missing columns in FEATURES:
- MarkDown2: 5269.0 missing (64.33%)
- MarkDown4: 4726.0 missing (57.7%)
- MarkDown3: 4577.0 missing (55.89%)
- MarkDown1: 4158.0 missing (50.77%)
- MarkDown5: 4140.0 missing (50.55%)

For SALES, no missing values were found in any column.

In the warehousing context, the high level of missingness in the MarkDown fields is critical: these columns are closely tied to promotions and holiday events. When MarkDown values are missing, analytical models and OLAP reports cannot reliably attribute changes in Weekly_Sales to price or promotion changes. This reduces the completeness and business value of the retail data warehouse.

2. CONSISTENCY

CONSISTENCY

Key consistency checks focused on uniqueness of primary keys and referential integrity.

1. Dimension key uniqueness:
   - Stores table: 45 unique Store IDs across 45 rows.

   No duplicate Store IDs were found in the STORES dimension.

2. Fact and feature grain:
- In FEATURES, the logical grain is (Store, Date).
- Number of (Store, Date) combinations with duplicates: 0.
- In SALES, the logical grain is (Store, Dept, Date).
- Number of (Store, Dept, Date) combinations with duplicates: 0.

These results indicate that the warehouse grain is enforced consistently and there are no duplicate fact or feature records at the expected grain. This supports consistent aggregation across OLAP cubes.

3. Referential integrity (Store key):

- Stores in FEATURES but not in STORES: 0
- Stores in SALES but not in STORES: 0

All store IDs used in SALES and FEATURES exist in the STORES dimension, indicating good referential integrity.

3. VALIDITY

VALIDITY

Several checks were used to evaluate validity of the data:

1. Date validity and temporal coverage:

- FEATURES dates range from 2010-02-05 00:00:00 to 2013-07-26 00:00:00
- SALES dates range from 2010-02-05 00:00:00 to 2012-10-26 00:00:00

The difference in date ranges indicates that FEATURES contains observations beyond the SALES period. For a retail warehouse, this means that features exist for weeks where no corresponding Weekly_Sales facts are available, which can limit time-series analyses and create misleading comparisons if not handled properly.

2. Negative Weekly_Sales values:

- Number of rows with Weekly_Sales < 0: 1285

Negative sales values were detected; these may represent returns or data entry errors and should be reviewed for validity.

3. Numeric ranges in FEATURES (potential outliers):

- Temperature: min=-7.290, max=101.950, mean=59.356
- Fuel_Price: min=2.472, max=4.468, mean=3.406
- CPI: min=126.064, max=228.976, mean=172.461
- Unemployment: min=3.684, max=14.313, mean=7.827
- MarkDown1: min=-2781.450, max=103184.980, mean=7032.372
- MarkDown2: min=-265.760, max=104519.540, mean=3384.177
- MarkDown3: min=-179.260, max=149483.310, mean=1760.100
- MarkDown4: min=0.220, max=67474.850, mean=3292.936
- MarkDown5: min=-185.170, max=771448.100, mean=4132.216

Extreme values in these fields may indicate outliers or sensor errors (e.g., unrealistic temperatures or CPI values). In the warehousing context, such values should be flagged and possibly winsorized or corrected during the Transform step of ETL.

4. ACCURACY (WAREHOUSING CONTEXT)

ACCURACY (IN WAREHOUSING CONTEXT)

Accuracy refers to how closely the stored values reflect real-world business events.

In this retail warehouse:

- The strong referential integrity on the Store dimension and the lack of negative sales values (in most cases) suggest that many recorded transactions accurately represent store-level activity.

- However, the missing MarkDown fields reduce the ability to accurately explain changes in Weekly_Sales. When promotion data is absent, analysts may misattribute demand spikes to other factors such as seasonality or store size.

- Differences in temporal coverage between FEATURES and SALES (FEATURES continuing beyond the final SALES date) can generate inaccurate conclusions if analysts assume both tables cover the same period. For example, rolling averages or year-over-year trends may silently mix periods with and without complete sales data.

To improve accuracy, ETL pipelines should:

1. Enforce aligned time windows across fact and feature tables.
2. Capture promotion/markdown data more reliably, especially around key retail holidays.

3. Introduce validation rules and business logic at load time (e.g., rejecting impossible values for Temperature, CPI, or Fuel_Price).