Report on Data Engineering Pretest

Task Overview

The primary objective was to process and analyze the provided datasets, ensuring data quality and preparing them for analytical purposes. The tasks included:

1. **Data Extraction**:
    - Loaded datasets (items, promotion, sales, and supermarkets) from CSV files.
    - Integrated these datasets into a PostgreSQL database for streamlined analysis.
2. **Data Cleaning**:
    - Verified datasets for missing values and duplicate rows—none were found.
    - Ensured primary keys were unique across all datasets.
    - Addressed duplicate entries in secondary keys, which were acceptable based on context.
3. **Data Transformation**:
    - Added unique IDs to the promotion and sales datasets for enhanced tracking.
    - Proposed the addition of a datetime column for the sales dataset but deferred due to a lack of year information.
4. **Schema Redesign**:
    - Developed a new database schema to incorporate additional primary keys while maintaining relational integrity.

Data Cleaning & Transformation

Steps Undertaken:

1. **Initial Quality Checks**:
    - Validated primary keys for uniqueness and confirmed no missing values.
2. **Duplicate Removal**:
    - Removed redundant rows to ensure data consistency.
3. **Key Enhancements**:

    - Introduced unique IDs for the promotion and sales datasets to simplify relational integration and improve traceability.
4. **Potential Enhancements**:
    - Suggested combining time and day columns in the sales dataset to create a datetime field. However, this remains unimplemented due to missing year information.

Business Insights

A. Branch-Level Sales Patterns

Analyzing supermarket transaction data revealed **branch-based sales patterns**, **regional trends**, and **performance benchmarks**:

1. **Top-Performing Branches**:
    - Example: **Supermarket 71 in Province 1** recorded the highest sales and units sold (sales: **12,111.45**, units: **8,234**).
    - **Actionable Insight**: Expand high-performing branches or replicate their strategies across underperforming locations.
2. **Underperforming Branches**:
    - Example: **Supermarket 13 in Province 1** had significantly lower performance (sales: **687.06**, units: **553**).
    - **Actionable Insight**: Investigate contributing factors like location, marketing efforts, or competitor presence and optimize strategies.
3. **Regional Trends**:
    - Provinces with higher sales indicate better performance.
    - **Actionable Insight**: Reallocate inventory or staff resources based on demand to improve overall efficiency.

B. Promotion Effectiveness

Evaluated the **impact of features and displays on promotions** by analyzing aggregated metrics:

1. **Consistent Display Utilization**:

    - For **feature = 1**, both provinces achieved 100% promotion effectiveness (`Effectiveness = 1.0`).
    - **Actionable Insight**: Maintain coordination between displays and promotional efforts.
2. **Display Count Variation**:
    - **Province 1**: 207,193 displays.
    - **Province 2**: 144,179 displays.
    - **Actionable Insight**: Explore reasons behind the disparity (e.g., market size, resources, or demand) and adjust marketing budgets.
3. **Regional Strategies**:
    - High display counts in Province 1 may reflect greater demand or higher investment in promotional activities.
    - **Actionable Insight**: Use these insights to guide resource allocation for optimized results.

Challenges Faced

1. **Temporal Data Limitations**:
    - Absence of year data in the sales dataset restricted advanced temporal analyses.
2. **Integration Complexity**:
    - Ensuring compatibility between datasets with newly introduced primary keys required careful schema design.
3. **Lack of Common Columns for Joins**:
    - The datasets did not include a common column for join operations, significantly limiting the scope of machine learning and business analysis. This prevented the ability to combine datasets effectively and derive richer insights.

Conclusion

The data engineering tasks ensured clean, reliable datasets ready for further analysis. Key improvements, such as unique IDs and schema redesigns, significantly enhanced usability and scalability. However, incorporating a common column for joins and more comprehensive temporal data in future datasets would unlock greater analytical potential.