DATA PREPROCESSING REPORT
Analysis of Marketing Campaign Performance Ending in May 2025

This report presents the process of inspecting, cleaning, and standardizing data to support the analysis of marketing campaign performance that ended in May 2025. The original data was collected from three main tables: campaign information (campaign_data.csv), conversion data (conversion_data_complex.csv), and user region information (user_region.csv). After processing, three corresponding cleaned datasets were created to ensure consistency, completeness, and accuracy for analysis.

1. Overview of the Original Data

- **Campaigns**: Includes 8 campaigns with details on ID, campaign name, budget, start date, and end date.
- **Conversions**: Includes 541 conversion transactions, with columns such as conversion_id, campaign_id, user_id, conversion_date, and revenue_generated.
- **User Region**: Includes 237 unique users, each assigned to one region (East, West, South, North).

2. Data Issues and Handling Methods

a. Duplicate conversion_id

A total of 41 duplicate conversion_ids were detected in the conversion table. These were not exact duplicates but differed in user_id or conversion_date, indicating they were valid transactions mistakenly recorded under the same ID.
Solution: These rows were not removed. Instead, all conversion_ids were regenerated using an automated method to ensure uniqueness while preserving the original data. This approach maintains data integrity throughout the analysis process.

b. Missing revenue data

There were 47 rows missing values in the revenue_generated column.
Solution: Instead of removing these rows, a simple interpolation system was developed to estimate missing revenue values. This system relied on:

- The average performance of each campaign
- Similar user behavior patterns
  Result: After processing, there were no missing revenue values remaining.

c. Date formatting and filtering for May

The start_date, end_date, and conversion_date columns were initially stored in inconsistent text formats.
Solution: All dates were converted to a standard datetime format. Then, campaigns that ended in May 2025 were filtered out, including:

- *Summer Sale*
- *New User Promo*
- *Flash Deal*
- *Member Exclusive*

d. Mapping user and region data
All user_ids in the conversion data matched entries in the user region table.
There were no missing region values. The region data consisted of four clearly defined regions and contained no errors.

3. Summary of Changes After Cleaning

- **Campaigns**: The original dataset contained 8 rows, but only 6 campaigns had actual data. The remaining 2 rows were entirely blank and were removed. Among the 6 valid campaigns, 2 were excluded because their end_dates were after May 2025.
- **Conversions**: From 541 rows, the cleaned dataset included 530 rows after resolving ID issues and filling in missing revenue data.
- **Users**: Remained unchanged at 237 users, with complete region information.

4. Data Readiness for Analysis

After preprocessing, the data meets the following criteria:

- No duplicate or missing critical data
- Standardized date formats enabling time-based analysis
- Strong foreign key relationships between tables (campaign_id, user_id)
- Data is ready for analysis of campaign effectiveness using indicators such as:
    - Total revenue
    - ROAS (Return on Ad Spend)
    - CPL (Cost per Lead)
    - Performance by region

5. Conclusion

The cleaned dataset is fully reliable for analyzing the effectiveness of marketing campaigns that ended in May 2025. By preserving all records and estimating missing revenue data, the analysis remains comprehensive and avoids data loss, allowing for more accurate insights.