Scope of Data Preparation & Analysis

This document describes the expected scope of data preparation, cleaning, and structuring required for the EXL Round 2 Mini Project. Students must follow these guidelines to transform the raw datasets into clean, analysis-ready tables suitable for SQL queries and business insights.

1. Data Arrangement & Standardization

- Standardize formatting for all columns (consistent casing and trimming).
- Handle extra spaces, inconsistent values, and unwanted characters.
- Convert all numeric-like fields (amount, price, quantity) into numeric formats.
- Normalize and standardize date formats into a single valid datetime format.
- Validate and correct foreign keys (user_id, product_id, order_id, sku).

2. Deduplication & Null Handling

- Handle duplicate rows from all datasets.
- Handle NULL values using appropriate logic (drop, fix, or impute).
- Ensure no NULLs exist in required join keys.
- Validate data integrity to ensure joins work correctly.

3. Structuring Data for Analysis

- Build a Star Schema with fact_orders, dim_users, dim_products, and dim_date.
- Ensure each fact record links properly with all dimension tables.
- Generate clean, transformed datasets ready for SQL analysis.

4. Query- Ready Data Preparation

The cleaned data must be structured to support analytical queries such as:

- Get order details placed in Mumbai in the month of December.
- Get percentage of total sales contributed by each city.
- Calculate average order value per user.
- Identify repeat customers.
- Generate monthly revenue trends.
- Identify highest-selling products or categories.