Data Quality Report

1. Data Cleaning Methodology

The dataset used for this project was pulled from KAGGLE . It contains information on 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. Its features allow viewing orders from multiple dimensions: from order status, price, payment, and freight performance to customer location, product attributes, and finally reviews written by customers.

Missing Values Handling:

- Checked missing values for each dataset and calculated missing percentages.
- Missing categorical values were replaced with 'Unknown' where appropriate.
- Timestamp columns were filled using logical estimations based on median values for similar categories.
- Missing numerical values were imputed using the median strategy.

Duplicate Removal:

- Identified and removed duplicate rows based on unique identifiers (e.g., customer_id, order_id, seller_id).
- Applied different deduplication strategies for each dataset to maintain data integrity.

Outlier Detection & Handling:

- Used the **Interquartile Range (IQR) method** to detect numerical outliers.
- Identified and removed extreme outliers in key variables like order values and review scores.

Data Type Standardization:

- Converted timestamps to appropriate `datetime` format.
- Standardized categorical columns (e.g., city and state names) to title and uppercase formats to avoid inconsistencies.

Normalization & Export:

- Dropped irrelevant columns that were not needed for analysis.
- Ensured cleaned datasets were stored in **Parquet format** for efficiency.

2. Feature Engineering

Feature Extraction:

- **Customer-Level Aggregations:**

    - Total number of lifetime orders per customer.
    - Average order value.
    - Days since last order.
- **Payment Features:**
    - Average number of installments per customer.
    - Identified high spenders based on the top 10% spending threshold.
- **Customer Experience Metrics:**
    - Average review score per customer.
    - Total number of reviews given.

Churn Prediction Features:

- Defined churn threshold as 6 months of inactivity.
- Labeled customers as "churned" or "active" based on last order date.

Feature Scaling & Balancing:

- Applied **StandardScaler** to normalize continuous features.
- Used **Synthetic Minority Oversampling (resampling)** to balance churned and non-churned customer data for model training.

This report summarizes the data preparation process, ensuring data integrity, consistency, and improved feature relevance for predictive modeling.