DATASET ASSESSMENT DOCUMENT (DAD)

1. Dataset Overview

Dataset Name: DataCo Global Supply Chain Dataset

Description:
This dataset contains detailed information about supply chain operations, sales transactions, customer demographics, product details, logistics metrics, and delivery performance. It is suitable for analysis in domains such as supply chain efficiency, sales performance, customer insights, and operational analytics.

File Format: CSV

Number of Rows: 180,519
Number of Columns: 53

---

2. Source Information

Source Type: Provided directly by the user (external dataset)

Data Origin: Ecommerce/Supply Chain operational data generated from transactional systems.

---

3. Data Structure Overview

The dataset contains the following major categories of information:

- **Customer Information** (City, Country, Segment, Gender)

- **Order Information** (Order ID, Order Date, Delivery Status)

- **Shipping & Logistics** (Shipping Mode, Real vs Scheduled Shipping Days)

- **Product Information** (Product Name, Price, Category, Department)

- **Performance Metrics** (Sales per customer, Benefit per order)

---

4. Data Quality Assessment

4.1 Completeness Check

- Check for missing values across critical fields:

    - Customer City

    - Order ID

    - Product Name

    - Shipping Date

o   Benefit per order

4.2 Consistency Check

- Verify consistent data formats:

    o   Dates (Order Date, Shipping Date)

    o   Numeric fields (Sales, Price, Benefit)

- Check for inconsistent category names or duplicated category IDs.

4.3 Accuracy Check

- Validate whether real shipping days align logically with scheduled days.

- Check for unrealistic values (negative prices, impossible dates).

---

5. Data Issues Identified

Potential Data Quality Issues:

- Presence of **null values** in customer or shipping columns.

- **Incorrect data types** (e.g., date stored as string).

- **Duplicate records** in orders.

- Some columns may contain improper formatting.

---

6. Data Cleaning Plan

Steps for Cleaning:

1. **Handle Null Values**

    o   Replace, remove, or impute depending on field importance.

2. **Correct Data Types**

    o   Convert string dates to proper date formats.

    o   Convert numeric fields stored as text.

3. **Remove Duplicates**

    o   Identify duplicate Order IDs or transaction lines.

4. **Rename Fields**

    o   Use user-friendly, standardized field names.

5. **Pivot/Reshape Data** (if needed)

    o   Convert wide-format data to long-format for analysis.

6. **Create New Fields**

   o Calculated metrics such as delivery delay, profit margin, etc.

---

7. Preparedness for Analysis

After cleaning, the dataset will support the following analyses:

- Supply chain performance and delivery efficiency

- Product and sales performance

- Customer segmentation and behaviour patterns

- Trend and forecasting visuals

The dataset is considered **ready for Tableau visualization** after the cleaning steps are completed.

---

8. Notes & Recommendations

- Ensure date fields are consistent for time-series charts.

- Validate category names for better filtering.

- Consider grouping products/departments for improved dashboard readability.

Document Prepared By: *MUHAMMED SANAN P*