Data Cleaning Documentation – Cyclistic Case Study

Tool used: Microsoft Excel (Power Query)

Purpose: To clean and prepare the 2024 Cyclistic bike trip data for analysis in order to answer the business question: "How do casual riders and members use Cyclistic bikes differently?"

# Cleaning Steps

## 1. Import and Combine Data

- Used Power Query to import all 12 monthly .xlsx files stored in a single folder.  
- Selected Data > Get Data > From Folder to load all files at once.  
- Chose Combine & Transform to stack all datasets into one table.  
- Verified that all files had consistent column headers and formats.

## 2. Removed Unnecessary Columns

- Deleted any columns not relevant to the analysis (e.g., ride\_id, start\_station\_name, end\_station\_name) to reduce clutter and focus on key fields.

## 3. Standardized Data Types

- Ensured started\_at and ended\_at were recognized as Date-Time type.  
- Confirmed member\_casual was categorized as Text.

## 4. Created New Calculated Columns

- ride\_length: Subtracted started\_at from ended\_at to get ride length.  
- Formula: [ended\_at] - [started\_at] (data type: Duration)  
- ride\_length\_minutes: Converted ride\_length to total minutes.  
- Formula: Duration.TotalMinutes([ended\_at] - [started\_at])  
- day\_of\_week: Extracted the name of the weekday from started\_at.  
- Formula: Date.DayOfWeekName([started\_at])

- Group all columns using ‘’Group by’’ to count the ride length for each group.

## 5. Filtered Invalid Data

- Removed rows where:  
- ride\_length was negative (indicating data entry errors or cancelled rides).  
- ride\_length\_minutes was unusually long (e.g., over 24 hours).  
- Ensured no blank values in key columns like started\_at, ended\_at, member\_casual.

## 6. Loaded Cleaned Data

- Clicked Close & Load to return the cleaned data to Excel for analysis.