Cleaning Log

1. Data Merging

- **Action**: Merged monthly data tables into a single temporary table using `UNION ALL`.
- **Reason**: To consolidate all data into one table for comprehensive analysis.

2. Handling Missing Values

- **Action**: Checked for null values in `start_station_name` and `end_station_name`.
    - `start_station_name`: 905,237 null values
    - `end_station_name`: 956,579 null values
- **Action**: Updated null values in `start_station_name` and `end_station_name` to 'dummy_start_station' and 'dummy_end_station' respectively.
- **Reason**: To avoid data loss and bias, and maintain data integrity.

3. Adding and Populating New Columns

- **Columns Added**:
    - `start_date` (date)
    - `start_time` (time)
    - `end_date` (date)
    - `end_time` (time)
- **Action**: Populated new columns by splitting `started_at` and `ended_at` into `start_date`, `start_time` and `end_date`, `end_time`, respectively.

    - **Calculation**: `SET start_date = CONVERT(DATE, CAST(started_at AS date), 112);`
    - **Calculation**: `SET start_time = CAST(started_at AS time);`
    - **Calculation**: `SET end_date = CONVERT(DATE, CAST(ended_at AS date), 112);`
    - **Calculation**: `SET end_time = CAST(ended_at AS time);`
- **Reason**: For better readability and ease of analysis.

4. Handling Missing `start_date` Values

- **Action**: Updated missing `start_date` values with corresponding `end_date` values.
    - **Calculation**: `SET start_date = end_date WHERE start_date IS NULL;`
- **Reason**: Assumed most rides end on the same day to fill missing `start_date` values logically.

5. Calculating and Filling Missing `start_time` Values

- **Action**: Calculated average ride duration.
    - **Calculation**: `AVG(DATEDIFF(MINUTE, start_time , end_time))` = 9 minutes
- **Action**: Updated missing `start_time` values by subtracting 9 minutes from `end_time`.
    - **Calculation**: `start_time = DATEADD(minute, -9, end_time)`
- **Reason**: To populate missing `start_time` values based on the average ride duration.

6. Deriving New Columns

- **Action**: Added `day_of_week` column.
    - **Calculation**: `day_of_week = DATENAME(DW, start_date)`
- **Reason**: To facilitate better analysis and insights based on the day of the week.

7. DAX Calculations in Power BI

- **Columns Added Using DAX**:
    - `start_month`: Extracted month from `start_date`
        - **Calculation**: `start_month = Date.Month([start_date])`
    - `start_hour`: Extracted hour from `start_time`
        - **Calculation**: `start_time = Time.Hour([start_time])`
    - `same_station`: Determined how many rides start/end at the same station.
        - **Calculation**: `same_station = IF(start_station_name = end_station_name, TRUE, FALSE)`
- **Reason**: To create additional columns for detailed analysis and visualization in Power BI.