Data Cleaning

Objective:

The data cleaning file focuses on cleaning and preparing the traffic dataset for further analysis. The goal is to remove redundant columns, correct inconsistencies in the data, and ensure the dataset is structured for efficient processing in subsequent analysis.

Dataset Overview:

The dataset includes information about:

- **Road Characteristics**: road_name, location, suburb, speed_limit, and direction.
- **Vehicle Classification**: Multiple columns detailing counts of different vehicle types (e.g., vehicle_class_1 to vehicle_class_13).
- **Traffic Metrics**: average_speed, 85th_percentile_speed, maximum_speed, and total vehicle counts.
- **Geospatial Information**: Latitude, Longitude, Geo Shape x, and Geo Shape y.

Data Cleaning Steps:

1. **Data Loading**:
    - The dataset was loaded from a CSV file and inspected for missing values, data types, and overall structure.
    - A summary of the dataset was generated using the .info() method, showing that the dataset has **63,120 rows** and multiple features of varying data types (integers, floats, and objects).
2. **Column Selection**:
    - Irrelevant columns such as seg_descr, poly_area, and unnecessary identifiers like gisid and seg_part were identified for removal, as they do not contribute to traffic or vehicle analysis.
    - **Geospatial features** such as Geo Shape x and Geo Shape y were retained for potential spatial analysis.
3. **Vehicle Classification**:
    - Vehicle classifications from vehicle_class_1 to vehicle_class_13, as well as motorcycle and bike, were reviewed to identify heavy vs. light vehicle trends.
    - Any potential inconsistencies or missing data related to vehicle classes were corrected.
4. **Speed Metrics**:
    - Columns such as average_speed, 85th_percentile_speed, and maximum_speed were kept for further analysis, as they provide insight into speeding trends on various road segments.

- Speed limits were validated against the corresponding average speeds to flag potential inconsistencies.

5. **Handling Missing Values**:
    - Missing values were reviewed across the dataset, particularly in critical columns like vehicle counts and speed metrics. Columns with excessive missing values were marked for either imputation or removal in future steps.
6. **Final Dataset**:
    - The cleaned dataset is now ready for further exploratory data analysis (EDA), with unnecessary fields removed and relevant traffic and vehicle data retained.

Key Insights:

- **Redundant Data**: Columns such as seg_descr, gisid, and seg_part were identified as redundant and removed, as they did not provide actionable insights for the traffic analysis.
- **Vehicle Classification**: The dataset includes detailed breakdowns of different vehicle classes, which will be valuable for understanding traffic composition on different road segments.
- **Speed Metrics**: Speed-related columns were retained for analyzing speeding incidents and determining which road segments are most prone to speeding violations.