**ROAD TRAFFIC ANALYSIS EXCEL PROJECT**

**2.0. Data Collection**

**Dataset Details**

* **Source:** [Kaggle (Road Traffic Accident Dataset)](https://www.kaggle.com/datasets/saurabhshahane/road-traffic-accidents)
* **Time Period:** 2017–2020
* **Location:** Addis Ababa (Sub-city police department data)
* **Size:** 12,316 entries with 32 features (columns).
* **Data Type:** Tabular data in Excel format.

**3.0. Data Cleaning and Preparation**

Data cleaning is a crucial step to ensure the dataset is consistent, accurate, and ready for analysis. The following tasks were performed to clean and prepare the road traffic accident dataset:

**A. Standardizing the Time Column**

The time format was converted from the 12-hour format (12:00:00 AM to 11:59:59 PM) to descriptive time periods:

* **Morning**: 6:00 AM to 11:59:59 AM
* **Afternoon**: 12:00 PM to 5:59:59 PM
* **Evening**: 6:00 PM to 11:59:59 PM
* **Night**: 12:00 AM to 5:59:59 AM

This transformation helps simplify time-based analysis.

**B. Handling Missing Values**

Missing or blank cells across multiple columns were replaced with "Unknown" to maintain data consistency and avoid errors during analysis:

* **Educational\_Level**: 741 blank cells replaced with "Unknown."
* **Vehicle\_driver\_Relation**: All blanks replaced with "Unknown."
* **Driving\_experience**: All blanks replaced with "Unknown."
* **Type\_of\_vehicle**: All blanks replaced with "Unknown."
* **Owner\_of\_vehicle**: All blanks replaced with "Unknown."
* **Service\_year\_of\_vehicle**: All blanks replaced with "Unknown."
* **Area\_accident\_occured**: All blanks replaced with "Unknown."
* **Lanes\_or\_Medians**: All blanks replaced with "Unknown."
* **Road\_allignment**: All blanks replaced with "Unknown."
* **Types\_of\_Junction**: All blanks replaced with "Unknown."
* **Road\_surface\_type**: All blanks replaced with "Unknown."
* **Type\_of\_collision**: All blanks replaced with "Unknown."
* **Vehicle\_movement**: All blanks replaced with "Unknown."
* **Fitness\_of\_casuality**: All blanks replaced with "Unknown."

**C. Standardizing Text Format**

Inconsistent text values were corrected for uniformity:

* **Driving\_experience**: Corrected all instances of "unknown" to "Unknown."
* **Type\_of\_vehicle**:
  1. Replaced "Lorry(41?100Q)," "Lorry(11?40Q)," and "Long lorry" with "Lorry."
  2. Consolidated "Public (> 45 seats)," "Public (13?45 seats)," and "Public (12 seats)" into "Public Bus."
* **Defect\_of\_vehicle**:
  1. Replaced numeric values with descriptive labels:
     + 5 replaced with "Minor Defect."
     + 7 replaced with "Unclassified."
* **Casualty\_class**: Replaced "na" with "Unknown."
* **Age\_band\_of\_casualty**: Replaced "na" with "Unknown."
* **Sex\_of\_casualty**: Replaced "na" with "Unknown."
* **Fitness\_of\_casuality**: Corrected "NormalNormal" to "Normal."

**4. Handling Duplicates**

The dataset was checked for duplicates, and no duplicates were found.

**5. Reclassifying Numerical Values**

For better interpretation, numeric scales were replaced with meaningful labels:

* **Casualty\_severity**:
  1. 1 replaced with "Minor."
  2. 2 replaced with "Serious."
  3. 3 replaced with "Severe."