## **Road Traffic Accidents**

Data Curation  
  
Name : Hamza Saleh

## Data Sourcing

* **Name**: Road Traffic Accident Dataset of Addis Ababa City
* Description Dataset of road traffic accidents in Addis Ababa from 2017 to 2020, collected for a Master's research project. Includes accident causes, locations, time, environmental factors
* **Dataset Details:** 12,316 Rows & 32 Columns
* **Size:** 4,300KB (4.3MB)
* **Source**: [Kaggle Dataset - Road Traffic Accidents](https://www.kaggle.com/datasets/saurabhshahane/road-traffic-accidents?select=RTA+Dataset.csv)
* Dataset Creator: Tarikwa Tesfa Bedane

## Data Profiling

* Apply Excel filter to the dataset.
* Examine each attribute's unique value for inconsistency.
* Some statistical summaries:

**Day of Week**:

* **Most Common**: Friday (17% of entries).
* Other days like Thursday (15%) and other unspecified days (68%).

**Age Band of Driver**:

* **Most Common**: 18-30 (35% of entries).
* The other categories are 31-50 (33%) and others (32%).

**Educational Level of Driver**:

* **Most Common**: Junior high school (62%).
* Other education levels include Elementary school (18%) and Other (21%).

**Cause of Accident**:

* **Most Common**: No distancing (18%).
* Other causes include Changing lane to the right (15%).

Notable Features :

**Missing Data**:

* Various columns have missing values, including:
  + **Educational Level** (6% missing).
  + **Owner of Vehicle** (4% missing).
  + **Service Year of Vehicle** (32% missing).
  + **Defect of Vehicle** (36% missing).
  + **Area of Accident Occurred** (2% missing).
  + **Driving Experience** (7% missing).
  + **Type of Vehicle** (8% missing).

**Null/NA Values**:

* Several columns, such as **Casualty Class** (36% NA), **Work of Casualty** (26% null), and **Fitness of Casualty** (21% null), have significant portions of null values.

**Unique Values Discrepancy**:

* Columns like **Age Band of Casualty** and **Cause of Accident** contain some "Other" categories with over 60% of the data, which may suggest classification issues or unknown categories.

## Data Wrangling

### **Data Wrangling and Cleaning Steps**

**Educational Level:**

Imputed missing values with "Junior high school" (most common category) to maintain consistency.

**Rows affected**: 741.

**Vehicle Driver Relation:**

Removed rows where the Vehicle\_driver\_relation was missing or unclear.

**Rows deleted**: 579.

**Type of Vehicle:**

Replaced missing or "Unknown" values with the most common vehicle type, "Automobile," to ensure consistency in the dataset.

**Rows affected**: 950.

**Owner of Vehicle:**

Replaced missing data in the Owner\_of\_vehicle column with the most common value, "Owner," to standardize the entries.

**Rows affected**: 394.

**Service Year of Vehicle:**

Filled missing values in the Service\_year\_of\_vehicle column with "Unknown," which helps retain row completeness without introducing biased values.

**Rows affected**: 3928.

**Defect of Vehicle:**

Removed rows where the Defect\_of\_vehicle column had no diversity or specific values, as these entries didn't add value to the analysis.

**Rows deleted**: Entire row removed for lack of information.

**Area Accident Occurred:**

Deleted rows with missing values (Null) in the Area\_Accident\_Occurred column to ensure complete and reliable data.

**Rows deleted**: 204.

--

There are more wrangling steps i will make it during the project

**Data Table Schema**:

|  |  |  |
| --- | --- | --- |
| FEILD | TYPE | DESCRIPTION |
| Time | DATETIME | The time the accident occurred. Format: YYYY-MM-DD HH:MM:SS. |
| Day\_of\_week | STRING | The day of the week the accident occurred (e.g., 'Monday', 'Tuesday', etc.). |
| Age\_band\_of\_driver | STRING | Age category of the driver (e.g., '18-24', '25-34', '35-44', etc.). |
| Sex\_of\_driver | STRING | Gender of the driver (e.g., 'Male', 'Female'). |
| Educational\_level | STRING | Educational background of the driver (e.g., 'Junior high school', 'Graduate', 'PhD', etc.). |
| Vehicle\_driver\_relation | STRING | Relationship of the driver to the vehicle owner (e.g., 'Owner', 'Friend', 'Family', etc.). |
| Driving\_experience | INTEGER | Number of years the driver has been driving. |
| Type\_of\_vehicle | STRING | Type of vehicle involved in the accident (e.g., 'Car', 'Motorcycle', 'Truck', etc.). |
| Owner\_of\_vehicle | STRING | Ownership status of the vehicle (e.g., 'Owner', 'Non-owner'). |
| Service\_year\_of\_vehicle | STRING | Year the vehicle started service (e.g., '2000', '2015', 'Unknown'). |
| Area\_accident\_occured | STRING | Area where the accident occurred (e.g., 'Urban', 'Rural', 'Highway', etc.). |
| Lanes\_or\_Medians | STRING | Information about lanes or medians involved (e.g., 'Single Lane', 'Multiple Lanes', 'Median', etc.). |
| Road\_allignment | STRING | Type of road alignment (e.g., 'Straight', 'Curved', 'Bent', etc.). |
| Types\_of\_Junction | STRING | Type of junction (e.g., 'T-junction', 'Roundabout', 'Crossroad', etc.). |
| Road\_surface\_type | STRING | Type of road surface (e.g., 'Asphalt', 'Gravel', 'Concrete', etc.). |
| Road\_surface\_conditions | STRING | Conditions of the road surface at the time of the accident (e.g., 'Dry', 'Wet', 'Icy', etc.). |
| Light\_conditions | STRING | Light conditions at the time of the accident (e.g., 'Daylight', 'Dark', 'Streetlights', etc.). |
| Weather\_conditions | STRING | Weather conditions at the time of the accident (e.g., 'Clear', 'Rain', 'Fog', 'Snow', etc.). |
| Type\_of\_collision | STRING | Type of collision (e.g., 'Head-on', 'Rear-end', 'Side-impact', etc.). |
| Number\_of\_vehicles\_involved | INTEGER | Number of vehicles involved in the accident. |