**A black and white logo

Description automatically generated A logo of a globe and a graduation cap

Description automatically generated**

**Accident Analysis**

**Prepared by:**

**Amir Mansour**

**Eslam Abd Elkader**

**Mohammed Ashraf**

**Supervised by:**

**Dr. Amal Mahmoud**

**About Dataset**

**Data source:**

httpshttps://www.kaggle.com/datasets/jaspreetkhokhar/victoria-road-crash-data-2012-2023

**Number of rows:** 152445

**Number of columns:** 16

**Type of data:** Accidents

**Description:**

This dataset contains detailed information on road crashes that occurred from 2012 to 2023. Each record includes data on the accident number, date, time, type, and a descriptive code of the accident. Additional details such as the day of the week, light conditions, road geometry, severity of the crash, speed zone, sensors, and road management authority are also provided.

The dataset can be used for a variety of purposes including traffic safety analysis, public health research, urban planning, and policymaking. By understanding the patterns and causes of road accidents, stakeholders can implement more effective measures to reduce the frequency and severity of crashes, ultimately saving lives and improving road safety.

**The explanation of every column:**

* ACCIDENT\_ID: Unique identifier for each recorded accident.
* ACCIDENT\_DATE: The date when the accident occurred.
* ACCIDENT\_TIME: The specific time of the day when the accident happened.
* Day\_time: Categorized periods of the day (e.g., Morning, Night, Afternoon) to understand the time distribution of accidents.
* DAY\_WEEK\_DESC: Day of the week to observe the frequency of accidents on specific days.
* SEVERITY: The seriousness of the accident, which ranges from "Moderate" to "Severe injury".
* SPEED\_ZONE: Speed limit of the area where the accident occurred, measured in km/h.
* light\_condition\_desc: Description of the lighting conditions during the time of the accident (e.g., "Daylight AM", "Dark Street lights off").
* ROAD\_GEOMETRY\_DESC: Describes the road structure at the accident location (e.g., "T intersection", "Cross intersection").
* Sensors: Whether sensors were used in the accident analysis, or the location had any sensors monitoring road conditions.
* ACCIDENT\_TYPE\_DESC: A description of the accident type (e.g., "Vehicle overturned", "Collision with a fixed object").
* DCA\_DESC: Detailed Crash Analysis description that explains the nature of the collision (e.g., "OUT OF CONTROL", "OFF LEFT BEND INTO OBJECT/PARKED VEHICLE").

**Analysis Objectives**

**The primary goals of this analysis are:**

1. Understand Accident:

- Frequency of accidents over time, including daily, monthly, and yearly.

- Distribution of accidents by time of day and day of the week.

2. Severity of Accidents:

- Distribution of accidents by severity Relationship between severity and speed zones, light conditions.

3. Impact of Environmental Conditions:

- Impact of light conditions on accident frequency and severity.

- Analyze if certain weather conditions correlate with specific types of accidents.

4. Road Safety Insights:

- Relationship between road geometry and accident type.

- Distribution of accidents across different road types.

- Analysis of how intersections contribute to accidents.

5. Accident Causes:

- Investigate the types of accidents based on the Speed.

- Detailed analysis of the most common types of accidents to provide insights on high-risk scenarios.

**Data Preprocessing Steps (using excel, R, python in every step)**

**1.Data exploring:**

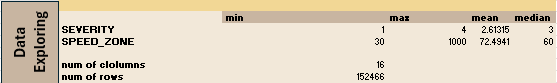
-seeing the data have a quick check for the columns and the rows.

-doing the statics to see the mean, median, mode, 1st quartile ,2nd quartile, checking if there a good correlation between columns.

-checking the amount of missing data.

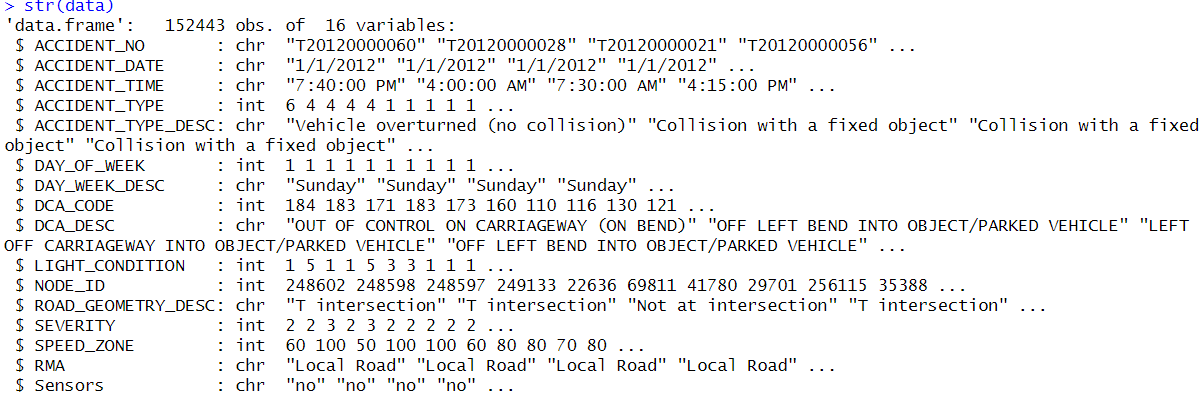
-checking the datatype of each column.

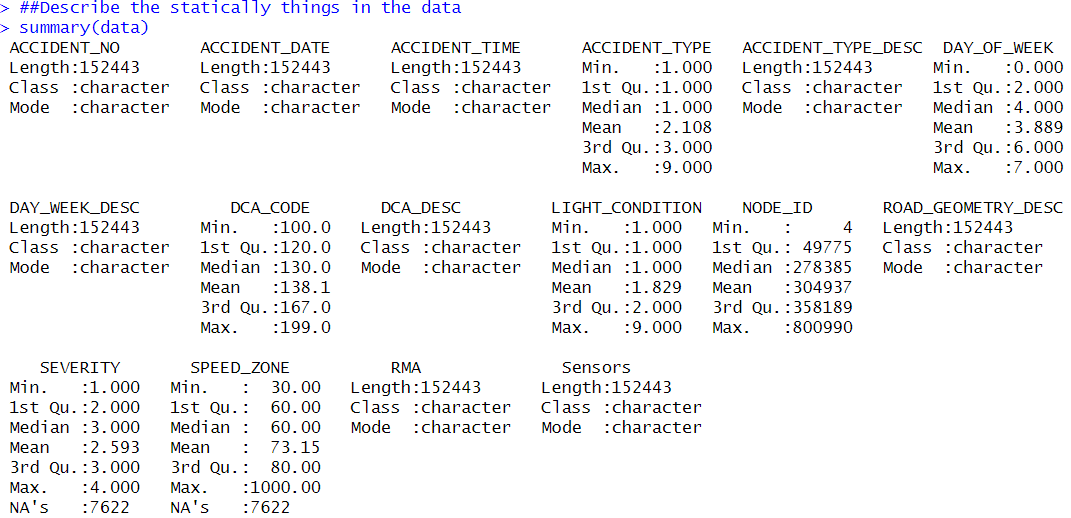
-**Excel:**



\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

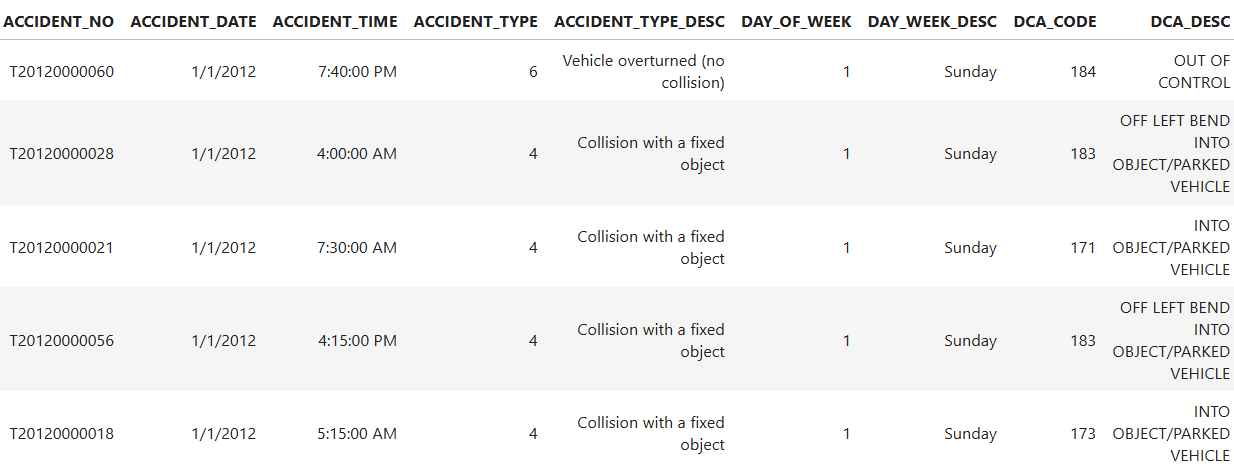
**- R:**

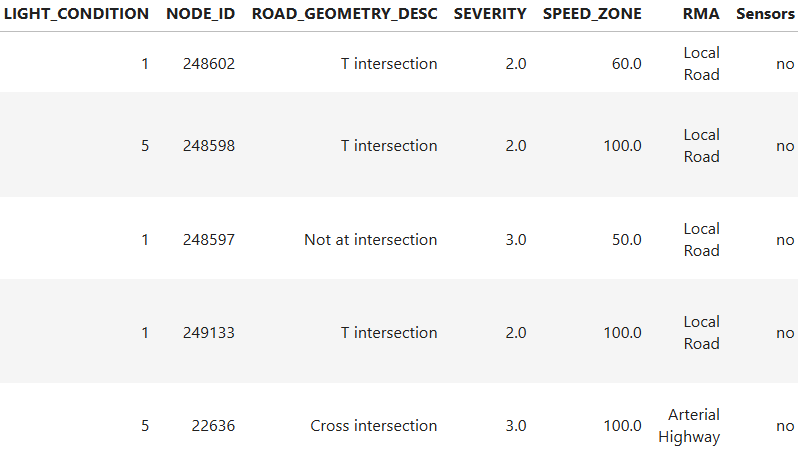
****

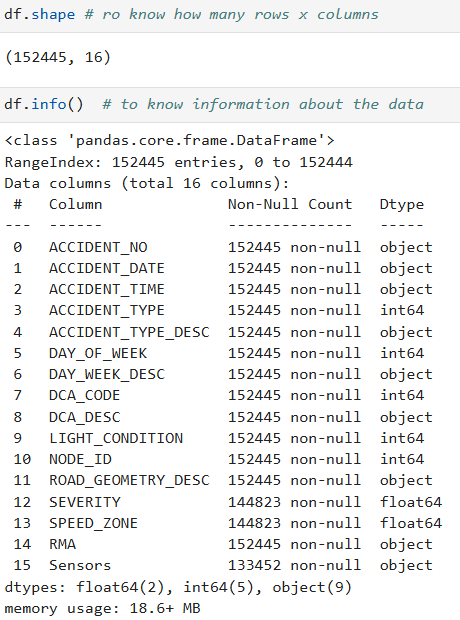
****

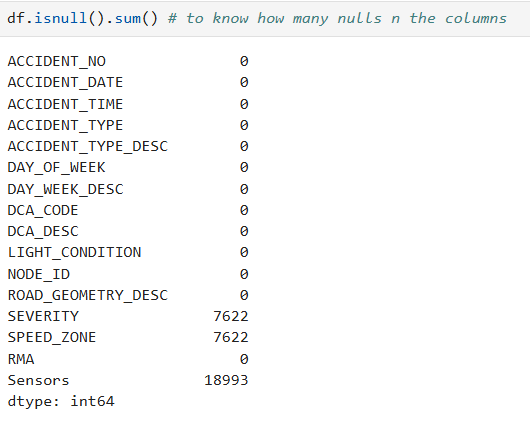
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**- Python:**

****

****

****

****

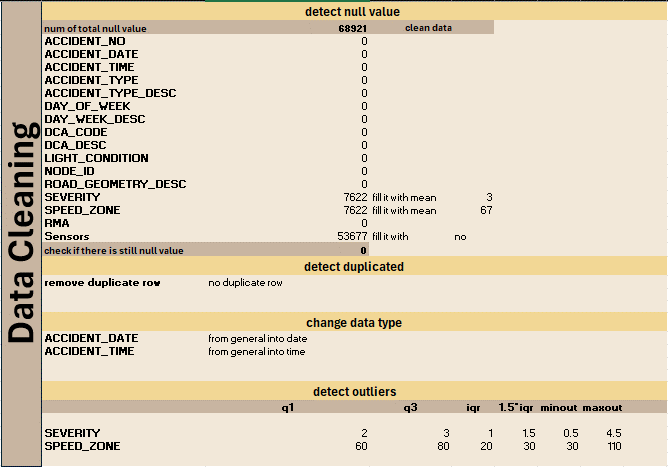
**A screenshot of a graph

Description automatically generated**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2. Data Cleaning:**

**- Excel**



-Missing Values: in Sensors, severity, and speed zone and

fill severity and speed\_zone by mean.

fill sensor by “no”.

- Datatype: change data type of columns

accident\_data from object into date.

accident\_time from object into time.

- Dupplicated there is no duplicated in this dataset .

- Outliers: get Q1 , Q3 and IQR to detect outliers and get min outliers and max outliers and check if there is value greater than man outlier return it into max outlier and if there is value less than min outlier return it into min outlier

- Adding columns: to clear some columns we added

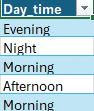
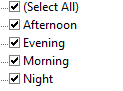
light\_condition\_desc .

Severity\_desc .

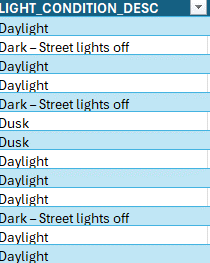
Day\_time .

Data Transformation:

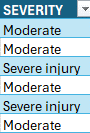
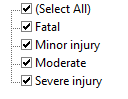
-Categorization: Convert time-based variables like `ACCIDENT\_TIME` into specific periods (Morning, Afternoon, Evening, Night).

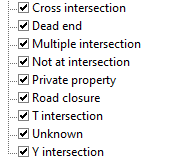
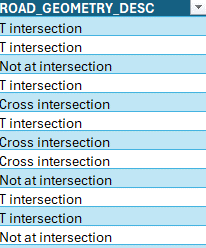
-Categorization: Convert variables like `light\_condition ` into specific variables (Dark Street lights on, Dusk, Dawn, Daylight AM, Dark Street lights off, Dark No streetlights, Overcast).



-Categorization: Convert variables like `severity ` into specific variables Minor, Moderate, Severe, Fatal).

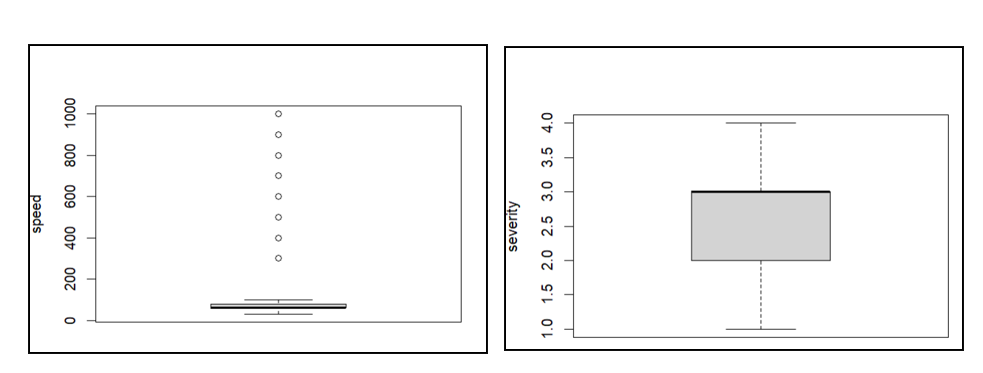
-Categorization: Convert variables like `geometry ` into specific variables (T intersection, not at intersection, Cross intersection, Multiple intersection, Unknown, Y intersection, Dead end, Road closure, Private property).

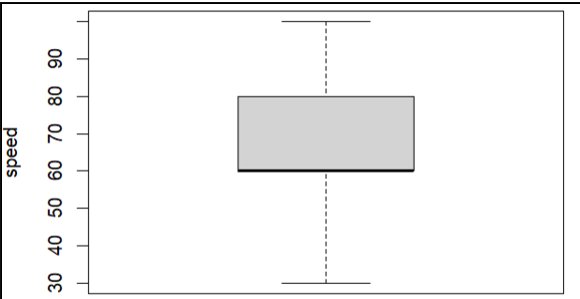
 

- Delete unneeded column like : Dca\_DESC .

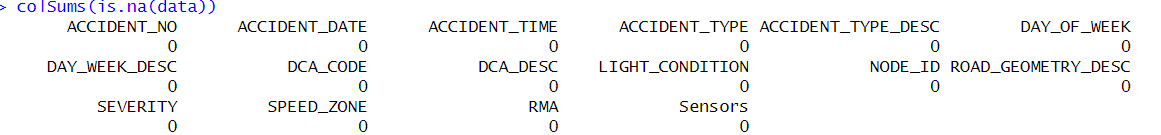
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**R:**





-Outliers: check the outliers and fixing it from speed zone, severity and check it using boxplot in R in speed.



-Missing Values: Filling any missing in columns such as Sensors, severity, and speed zone using mean in speed and median in severity and filling sensors with yes.

A black and blue text

Description automatically generated

-Datatypes: Ensure correct datatypes for each column (Accident\_Date, Accident\_Time).



-Remove Duplicates: Check for and eliminate duplicate entries based on ACCIDENT\_ID but there wasn’t any duplicated data

A screenshot of a computer code

Description automatically generated

A red and yellow squares with white text

Description automatically generated

-check the correlation between all the columns.

A computer code with text

Description automatically generated

-Delete unneeded column Dca\_DESC.

**-Data Transformation:**

A computer code with numbers and letters

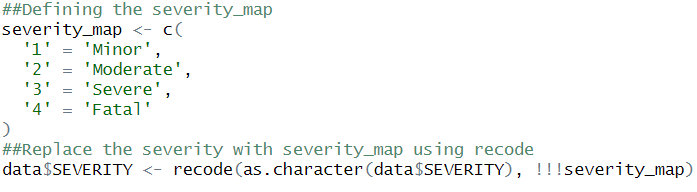
Description automatically generated with medium confidence

-Categorization: Convert time-based variables like `ACCIDENT\_TIME` into specific periods (Morning, Afternoon, Evening, Night).

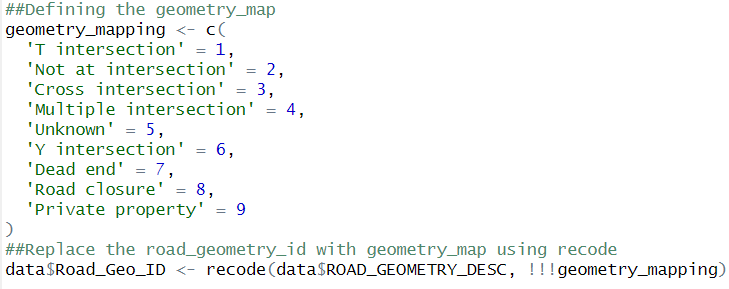
A computer screen shot of a code

Description automatically generated

-Categorization: Convert variables like `light\_condition ` into specific variables (Dark Street lights on, Dusk, Dawn, Daylight AM, Dark Street lights off, Dark No streetlights, Overcast).



-Categorization: Convert variables like `severity ` into specific variables Minor, Moderate, Severe, Fatal).



-Categorization: Convert variables like `geometry ` into specific variables (T intersection, not at intersection, Cross intersection, Multiple intersection, Unknown, Y intersection, Dead end, Road closure, Private property).

A white background with black and white text

Description automatically generated

-Categorization: Convert variables like `Rma ` into specific variables (Local Road, Arterial Highway, Freeway, Arterial Other, Non-Arterial).

**Python:**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

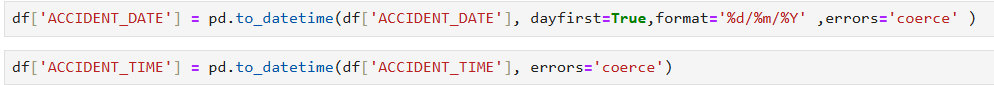
Description automatically generated**

-Outliers: check the outliers and fixing it from speed zone, severity and check it.

A screenshot of a computer program

Description automatically generated

-Missing Values: Filling any missing in columns such as Sensors, severity, and speed zone using simple impute and fillna in sensors.



-Datatypes: Ensure correct datatypes for each column (Accident\_Date, Accident\_Time).

A close-up of a computer screen

Description automatically generated

-Remove Duplicates: Check for and eliminate duplicate entries based on ACCIDENT\_ID but there wasn’t any duplicated data

A screenshot of a computer code

Description automatically generated

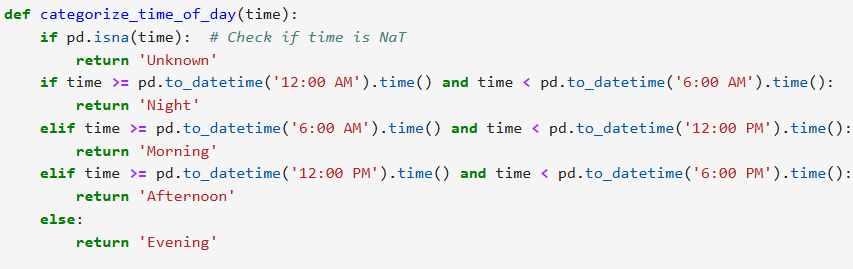
-check the correlation between all the columns.

A screenshot of a computer code

Description automatically generated

-Delete unneeded column Dca\_DESC.

**-Data Transformation:**

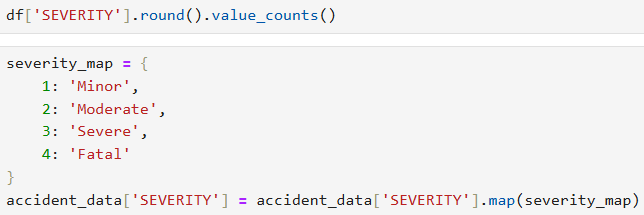


-Categorization: Convert time-based variables like `ACCIDENT\_TIME` into specific periods (Morning, Afternoon, Evening, Night).

A screenshot of a computer

Description automatically generated

-Categorization: Convert variables like `light\_condition ` into specific variables (Dark Street lights on, Dusk, Dawn, Daylight AM, Dark Street lights off, Dark No streetlights, Overcast).



-Categorization: Convert variables like `severity ` into specific variables Minor, Moderate, Severe, Fatal).

A screen shot of a computer code

Description automatically generated

-Categorization: Convert variables like `geometry ` into specific variables (T intersection, not at intersection, Cross intersection, Multiple intersection, Unknown, Y intersection, Dead end, Road closure, Private property).

A screenshot of a computer code

Description automatically generated

-Categorization: Convert variables like `Rma ` into specific variables (Local Road, Arterial Highway, Freeway, Arterial Other, Non-Arterial).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Analysis Approach:**

- Investigate if certain types of roads (Local Roads, Arterial Highways) are more prone to accidents.

- Investigate if certain types of Speed and light condition are more prone to accidents with high severity.

**Key Insights to Extract:**

1.High-Risk Areas: Identification of road types, intersections, and speed zones that are more prone to accidents.

2.Accident Severity Determinants: Factors that contribute the most to severe accidents (high-speed zones combined with poor lighting).

3. Time-Related Trends: Peak accident times (rush hours, late-night hours) and dangerous days of the week (weekends).

4. Impact of Lighting: How poor lighting conditions (streetlights off) contribute to accidents and whether these accidents tend to be more severe.

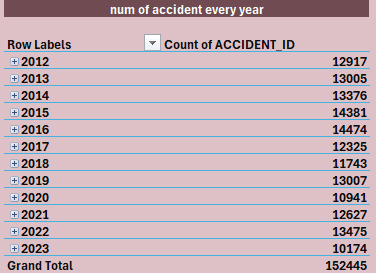
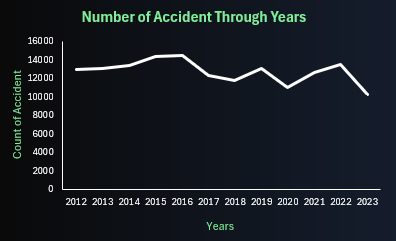
5. Common Accident Types: Most frequent accident causes, such as vehicle

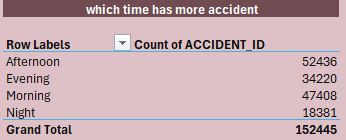
overturns or collisions with fixed objects, and recommendations to prevent these accidents.

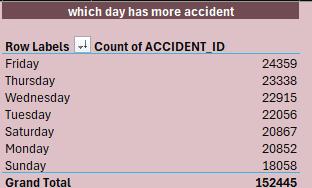
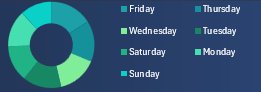
6. Impact of sensors: The number of accidents before sensors and after the beginning of it, number of accidents with sensors and without.

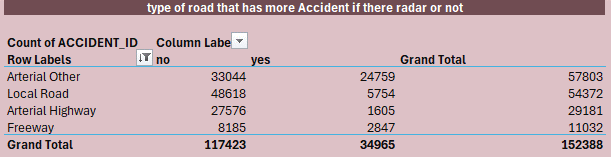
**Data analysis and visualization:**

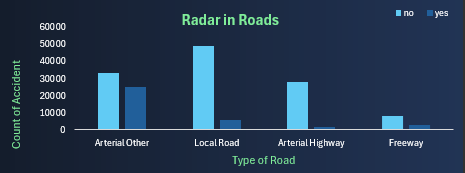
  ****

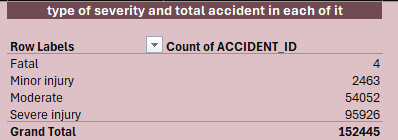
  ****

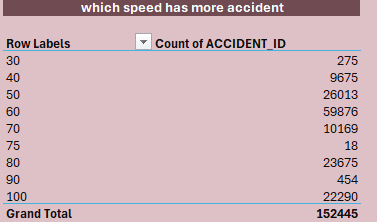
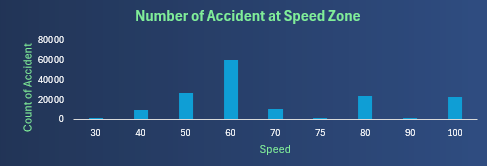
 ****

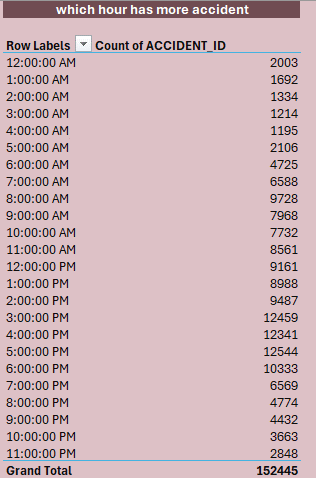
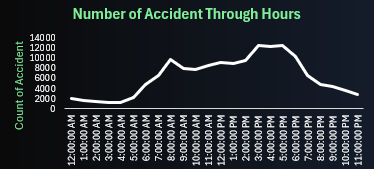
  ****

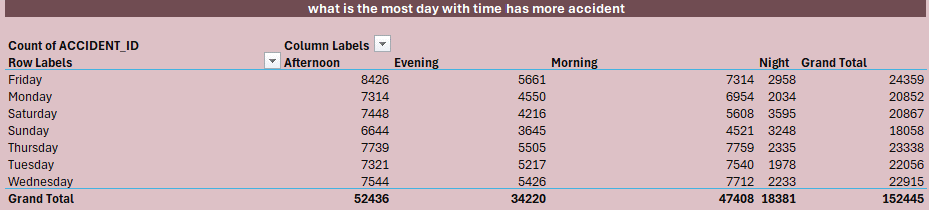


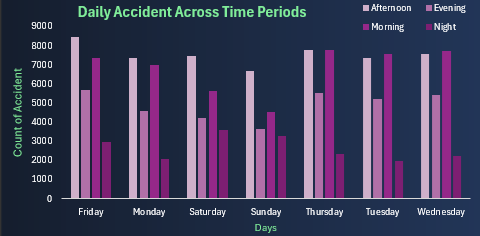
****

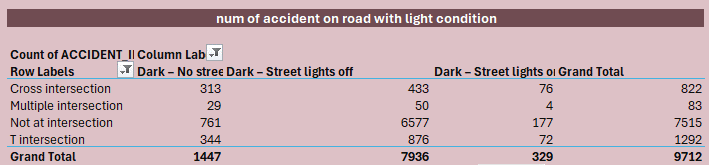
  ****

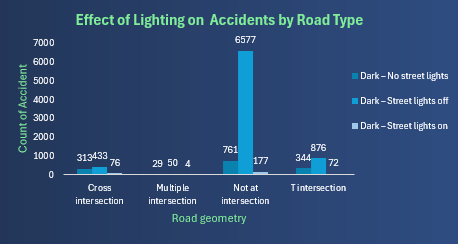
  ****

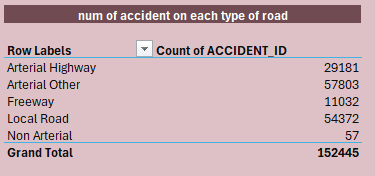
  ****

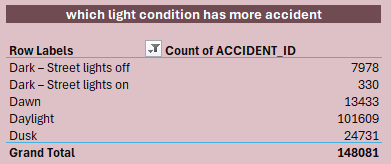


****



****

  ****

**Key Insights to Extract:**

-- The Number of accidents.

-- which year has the highest number of accidents.

-- The number of accidents in each day.

-- The number of accidents at each period of the day.

-- The number of accidents at each speed.

-- The number of accidents at each severity type.

-- The number of accidents at each light condition.

-- which Accident\_Type has the highest number of accidents.

-- which Road geometry has the highest number of accidents.

-- which Rma has the highest number of accidents.

-- If The Number of Accident Decrease after the Rader is available or not.

-- show what is the severity of the accident for each road.

-- show what is the severity of the accident for each speed.

-- show in speed 60 why is the highest number.

-- show why Accident\_Type Collision with vehicle is the highest number.

-- Before 2016, a sensor was not available.

**Deliverables**

Data Visualizations: Charts and pie that summarize key findings using python and R

**Python:**

**A graph with blue lines and dots

Description automatically generated**

A graph of accident per hours

Description automatically generated

A graph of blue rectangular bars

Description automatically generated with medium confidence

A graph of a pie chart

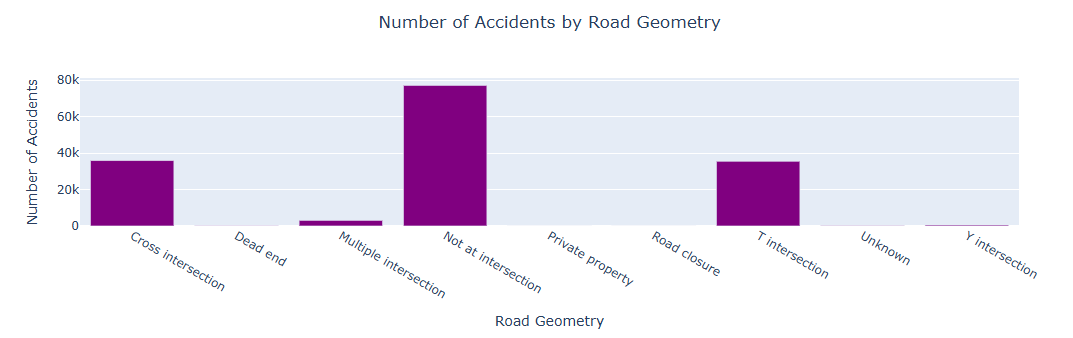
Description automatically generated

A graph of a graph

Description automatically generated with medium confidence

A graph of accident statistics

Description automatically generated with medium confidence



A blue and orange circle with text

Description automatically generated

A graph of different types of road

Description automatically generated

A graph of different colored squares

Description automatically generated

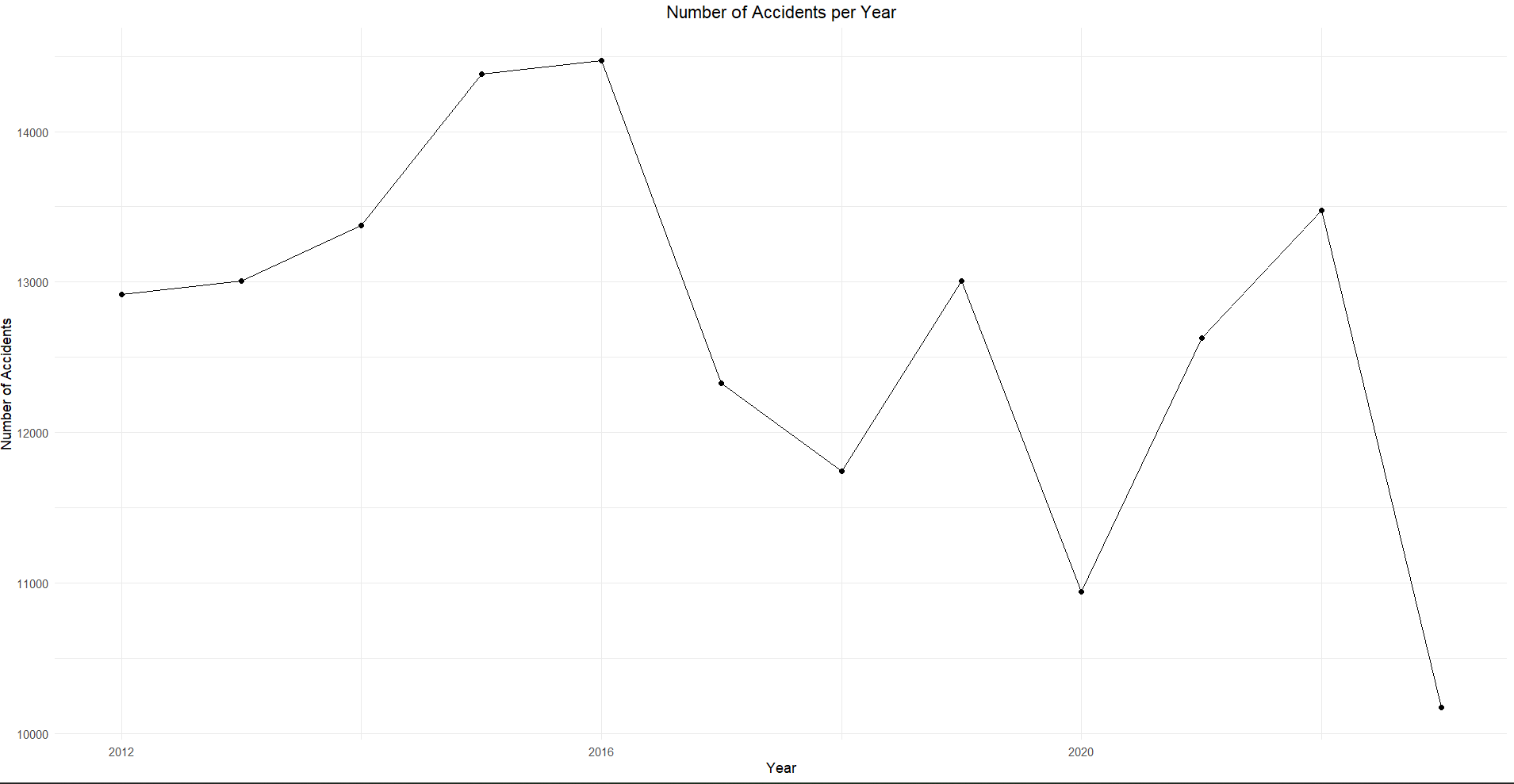
A graph with numbers and text

Description automatically generated with medium confidence

A graph of a graph

Description automatically generated with medium confidence

Using R:



A pie chart with numbers and text

Description automatically generated

A graph of a number of accidents

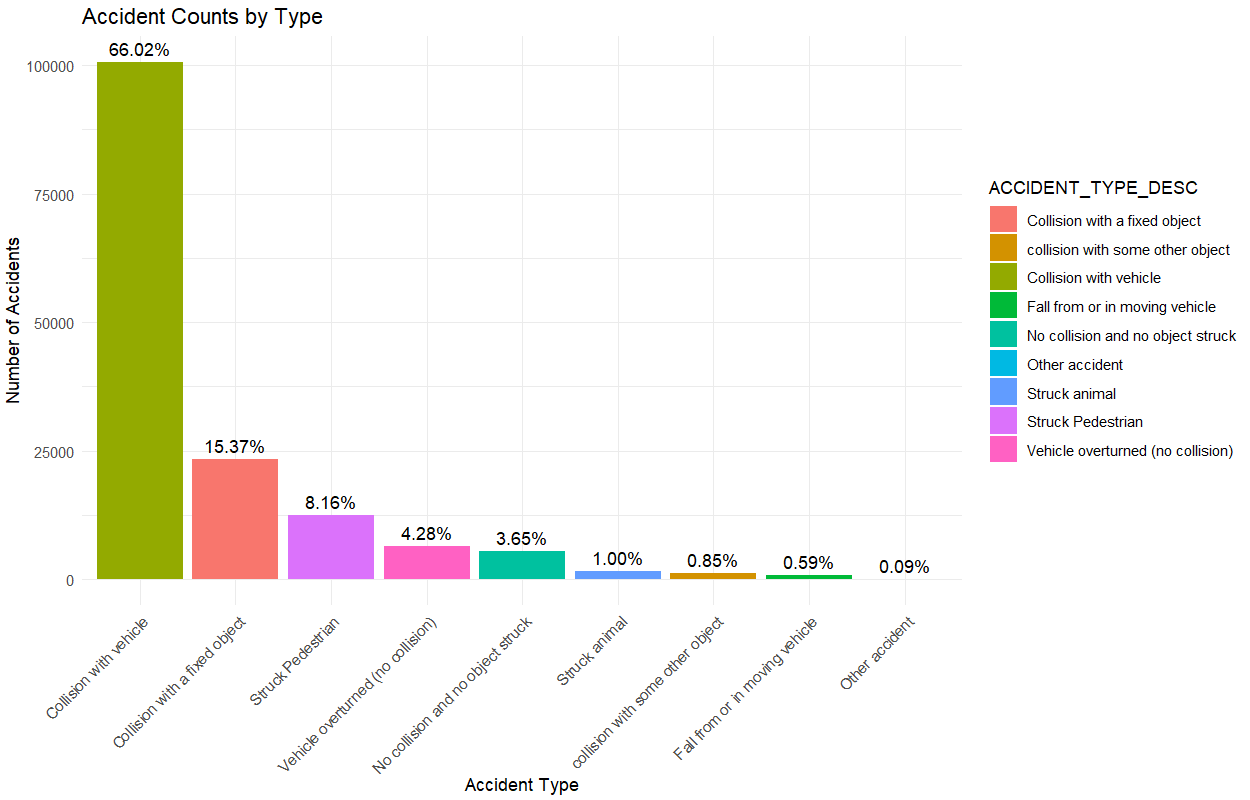
Description automatically generated

A graph of blue squares

Description automatically generated

A graph with different colored squares

Description automatically generated

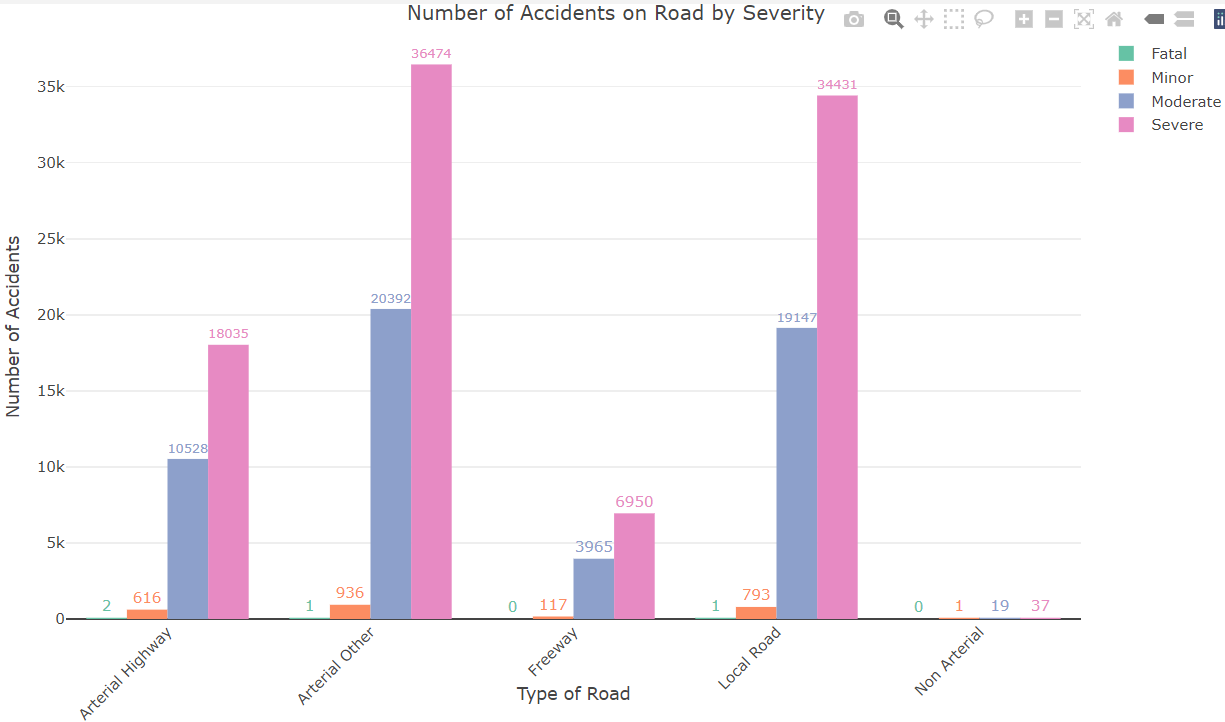


A graph with numbers and a bar

Description automatically generated

A graph with different colored squares

Description automatically generated



A graph of a number of injuries

Description automatically generated

A graph with red squares and white text

Description automatically generated

A graph of light condition

Description automatically generated

**Summary Report:**

**Accident Analysis (2013–2023)**

This accident data analysis provides critical insights into accident trends, accident types, road conditions, and external factors influencing safety, delivering key takeaways that can guide future policy-making, road safety improvements, and resource allocation.

**1. Accident Volume and Severity:**

- A total of 152,445 accidents were recorded from 2013 to 2023, with an average of 12,954 accidents per year. Of these, 100,306 collisions involved vehicles.

- Severe injuries accounted for a significant proportion (95,926), with minor injuries at a much lower count (2,463). Fortunately, fatal accidents remained relatively low (4).

**2. Time of Day & Accident Occurrence:**

- Accidents were most frequent in the afternoon (52,436) and least frequent during nighttime (18,381). This implies that peak traffic hours, likely linked to commuting periods, are higher-risk times.

- The hourly distribution of accidents also reveals a sharp increase from 6 AM to 8 AM, likely aligning with morning rush hours. Another spike occurs around 3 PM to 5 PM, suggesting higher risk during evening commutes.

**3. Road Type & Geometry:**

- Most accidents occurred on local roads (54,372), followed by arterial other roads (57,803). Understanding the road types involved provides opportunities to improve infrastructure and road conditions in these high-risk areas.

- Road geometry also played a role, with intersections, particularly cross intersections (35,900) and T-intersections (35,492), contributing to a large portion of accidents.

**4. Day of the Week Trends:**

- Friday shows the highest number of accidents (24,359), which could be associated with increased travel at the end of the workweek. Accidents are relatively stable across other weekdays, with a slight dip on Sundays (18,058).

**5. Impact of Speed Zones:**

- Accidents were most frequent in the 60 km/h speed zone (52,254), likely representing urban and suburban driving environments. Interestingly, both low-speed zones (40 km/h) and high-speed zones (100 km/h) also reported significant accident volumes, indicating that both urban and highway driving require attention.

**6. Lighting Conditions:**

- Lighting conditions at the time of accidents are crucial for safety improvements. Daylight hours see the highest number of accidents, but dark streets with no lighting also contribute a significant portion, highlighting the importance of well-maintained street lighting.

**7. Accident Type Breakdown:**

- Most accidents were collisions with vehicles (100,648), followed by collisions with parked vehicles (23,430). This suggests that the focus should be on reducing vehicle-to-vehicle and vehicle-to-object collisions through better traffic management and driver awareness campaigns.

**Recommendations**:

- Specific interventions to reduce accidents in high-risk areas (e.g., improving lighting in dark areas, installing speed control mechanisms).

- Road design recommendations (e.g., re-evaluating intersections with high accident rates).

- Time-based measures such as increased patrols or monitoring during high-risk hours.