Data Science Assessment Test

Real-Time Traffic Incident Analysis

Prepared by:

HABIMANA Martin

Data Scientist Candidate

Date: December 17, 2024

Analyzing Traffic Incidence —Patterns—Trends— Insights

Contents

1 Dataset Overview

The dataset of real time traffic incident focus on traffic incidents, containing variables that represent different aspects of reported issues, locations, dates, and statuses.

1.1 Key Variables in the datasets:

- **Traffic Report ID:** A unique identifier for each traffic incident.
- **Published Date:** The date when the traffic incident was reported.
- **Issue Reported:** The type/category of traffic incidents (e.g., "Crash Urgent," "Traffic Hazard").
- **Location:** The specific location where the traffic incident occurred.
- **Latitude and Longitude:** Geographical coordinates of where incident occurred.
- **Address:** A textual representation of the incident's location.
- **Status:** The resolution status of the incident (e.g., "Active," "Archived," "Unknown").
- **Status Date:** The date when the incident status was last updated.

I noticed that some variables, such as `Location`, `Latitude`, **`Status`** and `Longitude`, have missing values. This shows potential challenges in geospatial analysis. The `Status` field contains a large number of missing values, which will need to be addressed through imputation.

We can perform temporal analysis since `Published Date` and `Status Date` are given, which will allow for time-based aggregations and trend analysis.

The `Issue Reported` variable likely provides valuable insights into the types of incidents and their frequency, which can help in identifying the most common traffic issues.

1.2 Several analyses can be performed

Key Analyses to Perform:

- Using `Published Date` and `Status Date`, I can explore the number of incidents over time, such as by year, month, or day of the week.
- By analyzing `Latitude` and `Longitude`, it's possible to map incidents and identify high-risk areas for traffic management.
- By analyzing the `Issue Reported` variable, I can determine the most common types of traffic incidents (e.g., accidents, hazards) and identify trends in these categories.
- Analyzing the `Status` and `Status Date` variables will allow for tracking how quickly incidents are resolved and identifying patterns in incident status over time.

2 Tools and Packages Used for Analysis

The analysis was performed using **R**, **Python**, and **Excel** a powerful statistical programming language. The following packages were utilized:

R Packages Utilized

- **readxl**: For reading Excel files into R.
- **writexl**: For exporting cleaned datasets to Excel format.
- **dplyr**: For data manipulation, including filtering, summarizing, and creating new variables.
- **lubridate**: For handling and transforming date-time variables.
- **tidyverse**: For comprehensive data wrangling and visualization.
- **ggplot2**: For creating high-quality visualizations.
- **summarytools**: For generating summary statistics and descriptive reports.

I chose **R** for the data wrangling and analysis because I am most familiar with it and it offers unparalleled capabilities for:

1. Data manipulation and cleaning using efficient libraries like `dplyr`.
2. Visualization through `ggplot2`, which allows for high-quality, customizable plots.
3. Ensuring technical rigor and reproducibility through R's robust ecosystem.

In addition, **Excel** was used for quick data inspection and sharing results in a widely accessible format. This combination of tools ensures an optimal balance between analytical power.

3 Steps for Data Wrangling

The process of data wrangling involved several key steps to prepare the dataset for analysis. Each step is outlined below:

3.1 Inspection of the Dataset

Key Inspection Goals

The dataset was inspected to:

- Understand its structure and variable types.
- Identify the presence of missing or invalid data.
- Ensure that all columns align with the expected data formats.

3.2 Handling Missing Data

Missing Data Strategy

Missing values in critical columns were addressed using appropriate techniques. A summary of missing values across all variables was generated to determine the extent of data loss.

Table 1: **Summary of Missing Values in the Dataset**

S/N	Variable	Missing Count
1	Traffic Report ID	0
2	Published Date	0
3	Issue Reported	0
4	Location	559
5	Latitude	79
6	Longitude	79
7	Address	0
8	Status	1693
9	Status Date	0

Missing values in the **Status** column were imputed with the label *"Unknown"* to ensure completeness.

3.3 Standardizing Date Formats

Date Standardization Process

Date fields, including **Published Date** and **Status Date**, were standardized to ensure consistency and facilitate time-based analysis. Temporal features extracted include:

- **Year** and **Month** for trend identification.

3.4 Feature Engineering

New Features Created

New variables were engineered to enrich the dataset for further analysis:

- **Published Year** and **Published Month**: Extracted from the **Published Date**.
- **Status Year** and **Status Month**: Extracted from the **Status Date**.

3.5 Categorizing Variables

Conversion of Categorical Variables

Categorical variables such as **Issue Reported** and **Status** were converted into factor types to facilitate descriptive analysis and visualization.

These steps ensured that the dataset was thoroughly cleaned, standardized, and structured, making it ready for meaningful analysis and visualization.

4 Findings and Visualizations

4.1 Traffic Incidents by Year

Key Observation

2018 stands out as the year with the highest number of traffic incidents (**70,070**), followed by a decrease in 2019 1.
Note:

- The data for **2022** only includes reports from January and February.
- The data for **2017** is limited to the last four months (September–December).

These limitations explain the smaller number of traffic incidents recorded in 2017 and 2022.

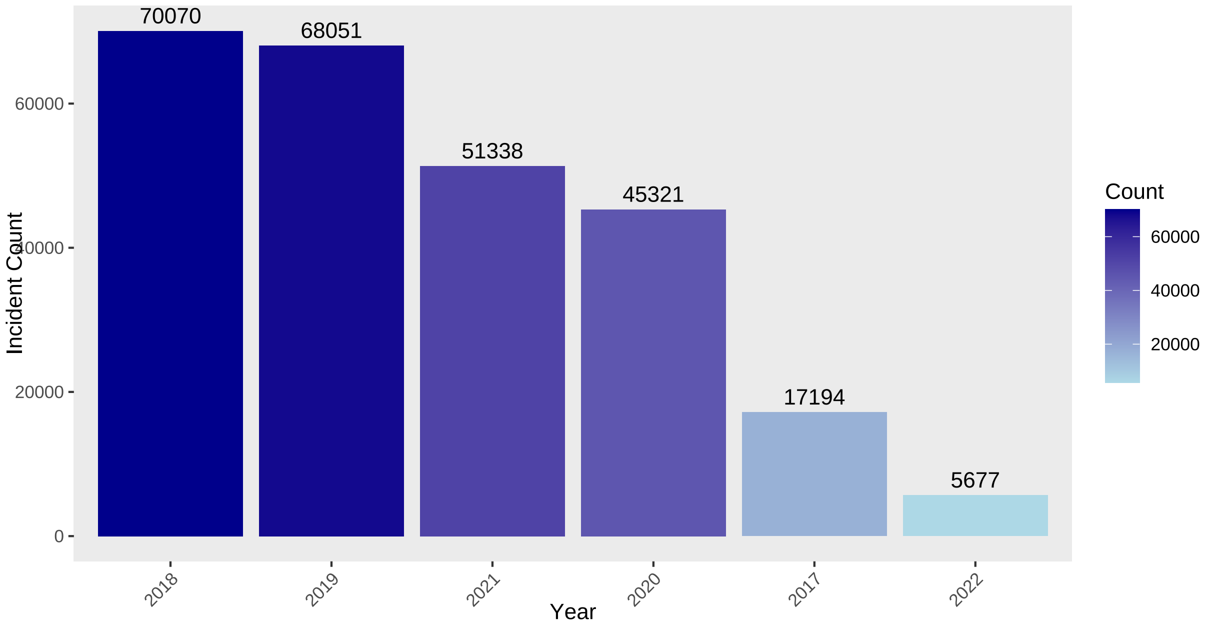

Figure 1: **Traffic Incidents by Year**

To gain a more comprehensive understanding of incident trends, it would be beneficial to have complete data for all months in both 2017 and 2022.

4.2 Traffic Incidents by Month

The analysis of traffic incidents by month reveals seasonal patterns, with noticeable fluctuations in traffic incident counts across months and years. The trends, however, do not follow a consistent trajectory, as some years exhibit increases while others show decreases (Figure: 2).

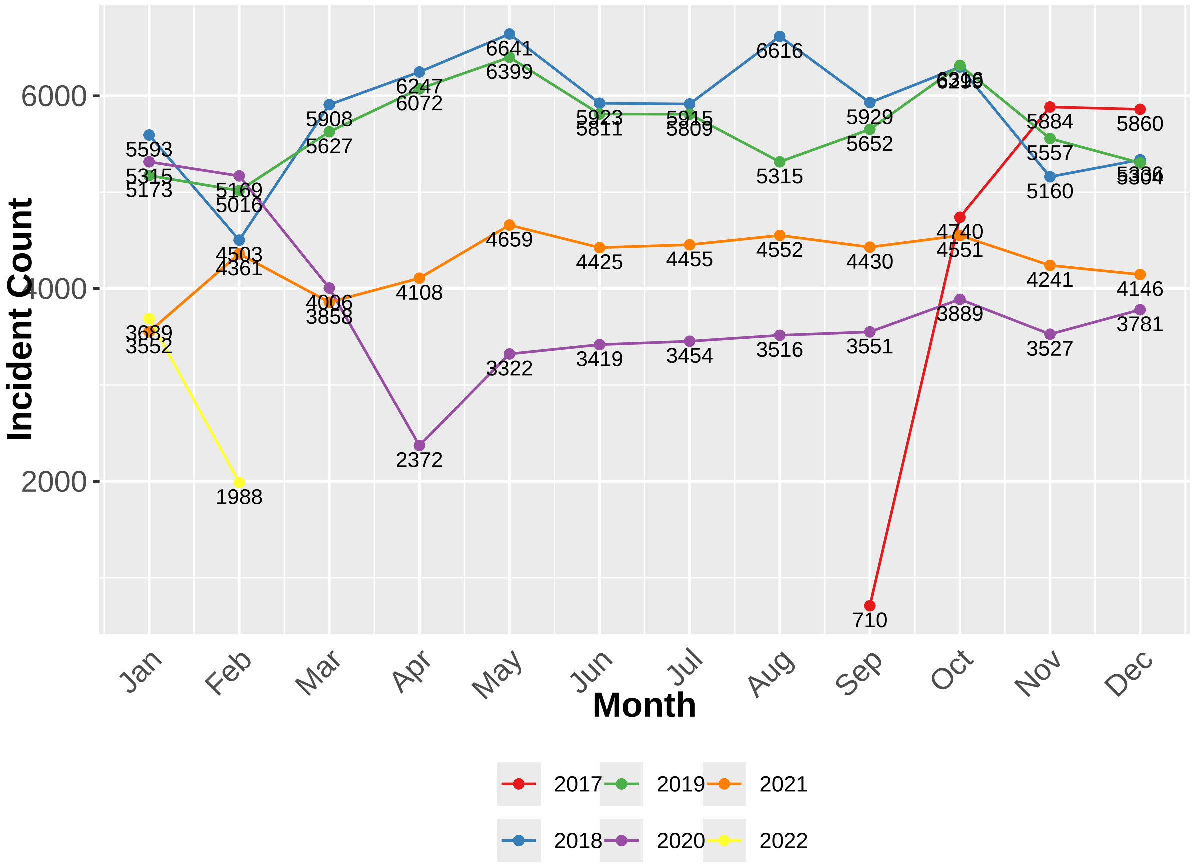

Figure 2: **Traffic Incidents by Month and Year**

Key Insights

- Across all years, there is a sharp **increase in incidents from April to May**, followed by a significant decrease between May and July.
- In **2017**, traffic incidents spiked dramatically:
  - September: **710 incidents**.
  - October: **4,740 incidents**.
  - November: **5,884 incidents**.
  - December: **5,860 incidents**.
- In **2022**, there was a steep decline from **3,689 incidents** in January to **1,988 incidents** in February.

The heatmap provides a clear visual representation of incident frequency over time, highlighting both seasonal trends and anomalies, such as the sharp increases and declines noted above.

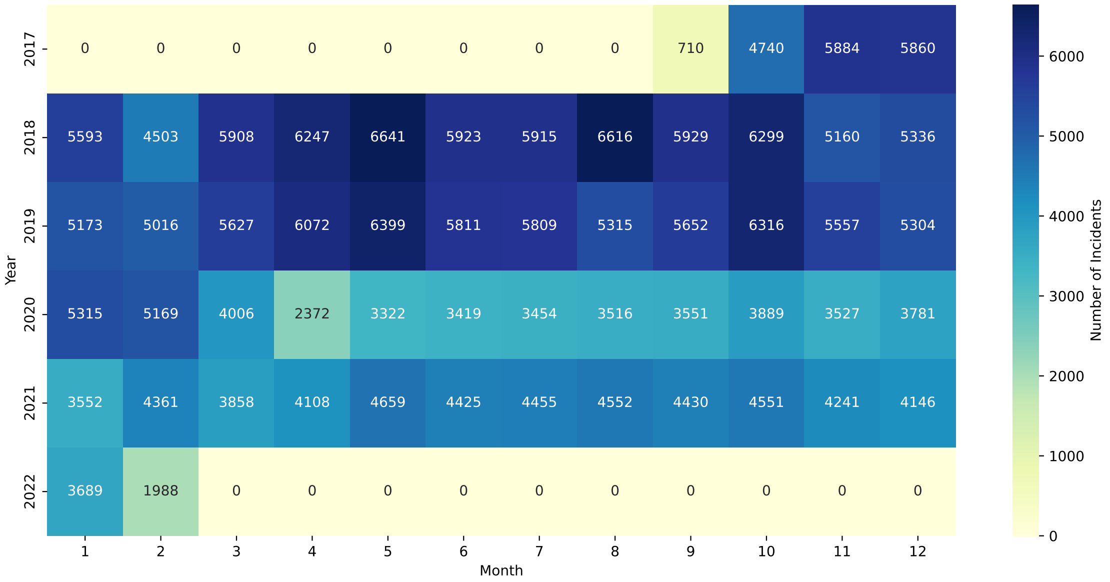

Figure 3: **Heatmap of Traffic Incidents by Year and Month**

4.3 Issue Type with Highest Traffic Incidents

The most frequent types of traffic incidents include **Crash Urgent** and **Traffic Hazard**, followed closely by **Crash Service**, **Collision**, and **Debris**. These categories account for the majority of reported traffic issues.

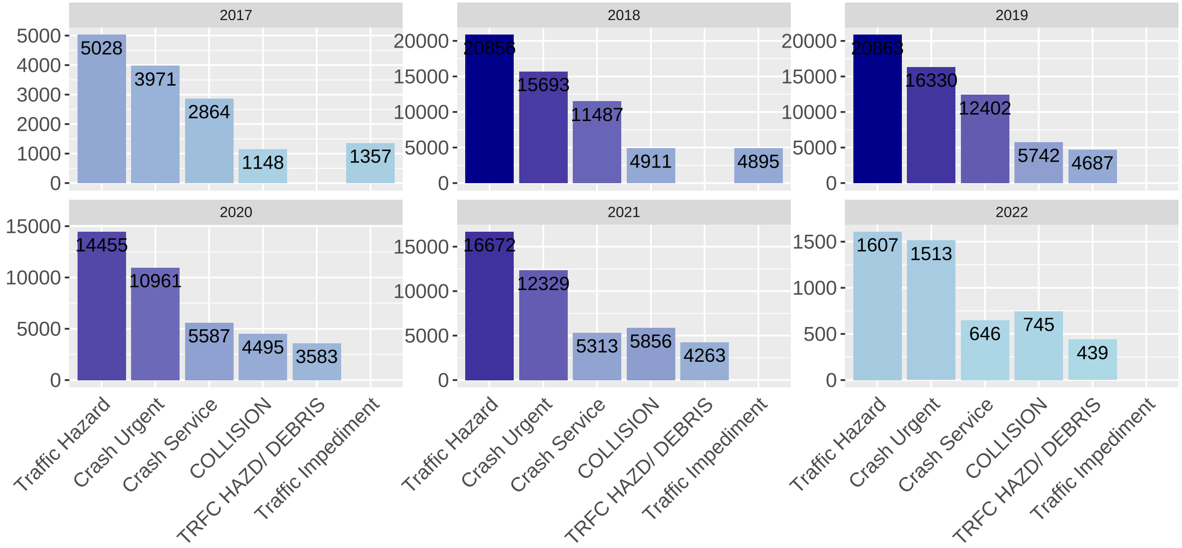

Figure 4: **Top 5 Traffic Incidents by Issue Reported**

Key Insights

- **Crash Urgent** and **Traffic Hazard** are the leading incident types, indicating frequent occurrences of urgent crashes and hazardous road conditions.
- Other notable incident types include:
  - **Crash Service**
  - **Collision**
  - **Debris**

4.4 Incident Status Analysis

The analysis of incident statuses sheds light on the resolution progress of reported traffic incidents. The vast majority of incidents are marked as **Archived**, indicating that they have been resolved. In contrast, a smaller portion remains **Active**, suggesting ongoing or unresolved incidents.

4.5 Conclusions

The analysis of traffic incidents provides valuable insights into temporal trends, the types of issues reported, and potential geographical patterns.

Key Findings

- **General Increase in Traffic Incidents:** There is an overall upward trend in traffic incidents over the years.
- **Seasonal Trends:** Traffic incident counts peak in May and show significant increases during the last three months of the year.
- **Frequent Incident Types:** Certain types of traffic incidents, such as **Crashes** and **Traffic Hazards**, occur more frequently, highlighting critical areas for traffic management.

These findings emphasize the importance of targeted interventions to mitigate frequent traffic issues and address seasonal spikes effectively. Future analysis could focus on geographical mapping to identify high-risk zones and discover spatio - temporal correlation.

4.6 Complementary Datasets for Deepening Insights into traffic incident

To enhance the traffic incident analysis, the following datasets can be used for cross-analysis and to uncover more insights from traffic:

Dataset	Key Variables	Potential Insights
Road Infrastructure Data	Road Type, Speed Limits, Road Conditions	Influence of poor infrastructure and road quality on incidents.
Weather Data	Temperature, Precipitation, Visibility, Weather Condition	Impact of weather (e.g., rain, fog) on incident frequency and severity.
Traffic Volume Data	Vehicle Count, Congestion, Travel Speed	Relationship between traffic volume, peak hours, and incident occurrence.