**Analysis Report**

**1. Introduction**

This report analyzes a large dataset of traffic accidents, which includes nearly 19,000 entries and 54 different details for each accident. It covers information like the location, time, road conditions, types of vehicles involved, and the outcomes (such as deaths or injuries). The goal is to find patterns, identify key factors that lead to fatal accidents, and prepare the data for predicting future accidents.

**2. Dataset Overview**

The dataset has **18,957 rows** and **54 columns**.

**Data Types:**

* **Numeric:** nine columns (like OBJECTID, LATITUDE, LONGITUDE, TIME, FATAL\_NO)
* **Categorical:** forty-five columns (like DATE, STREET1, ROAD\_CLASS, ACCLASS, INJURY)

**Missing Values:**

* Some columns have a lot of missing data (e.g., EMERG\_VEH is missing 99.74%, DISABILITY is missing 97.40%).
* Important columns like ACCNUM (missing 26.01%) and INJURY (missing 46.93%) were dealt with during preprocessing.

**3. Data Cleaning and Preprocessing:**

* **Duplicate Rows:** No duplicates were found.
* **Missing Values:** Columns with a lot of missing data (like EMERG\_VEH and DISABILITY) were either removed or filled in. Numeric columns like ACCNUM were filled in using suitable methods.
* **Feature Selection:** Key features chosen for modeling include TIME, LATITUDE, LONGITUDE, x, and y. These were picked based on their importance and relevance to predicting fatal accidents.
* **Class Imbalance:** The dataset had many more non-fatal accidents (14,469) compared to fatal ones (696). To balance this, SMOTE was used, resulting in 28,938 rows with an equal number of non-fatal and fatal accidents (14,469 each).

**4.1 Exploratory Data Analysis (EDA):**

**LATITUDE and LONGITUDE:**These features cluster together, showing they are limited to specific geographic areas (which makes sense since they represent physical locations).

**TIME:**Scatter plots for TIME against other features do not show a clear pattern, suggesting TIME likely has a complex relationship with other variables.

**INDEX:** INDEX shows distinct clusters, indicating it might represent categories or ranks rather than continuous data.

**OBJECTID:** OBJECTID appears randomly scattered against other features, typical for a unique identifier rather than a meaningful numeric variable.

The observations about LATITUDE and LONGITUDE being correlated and TIME showing no linear relationship match what is visible in the plot.

A screenshot of a graph

AI-generated content may be incorrect.

**4.2 Boxplots of Key Numeric Features:**

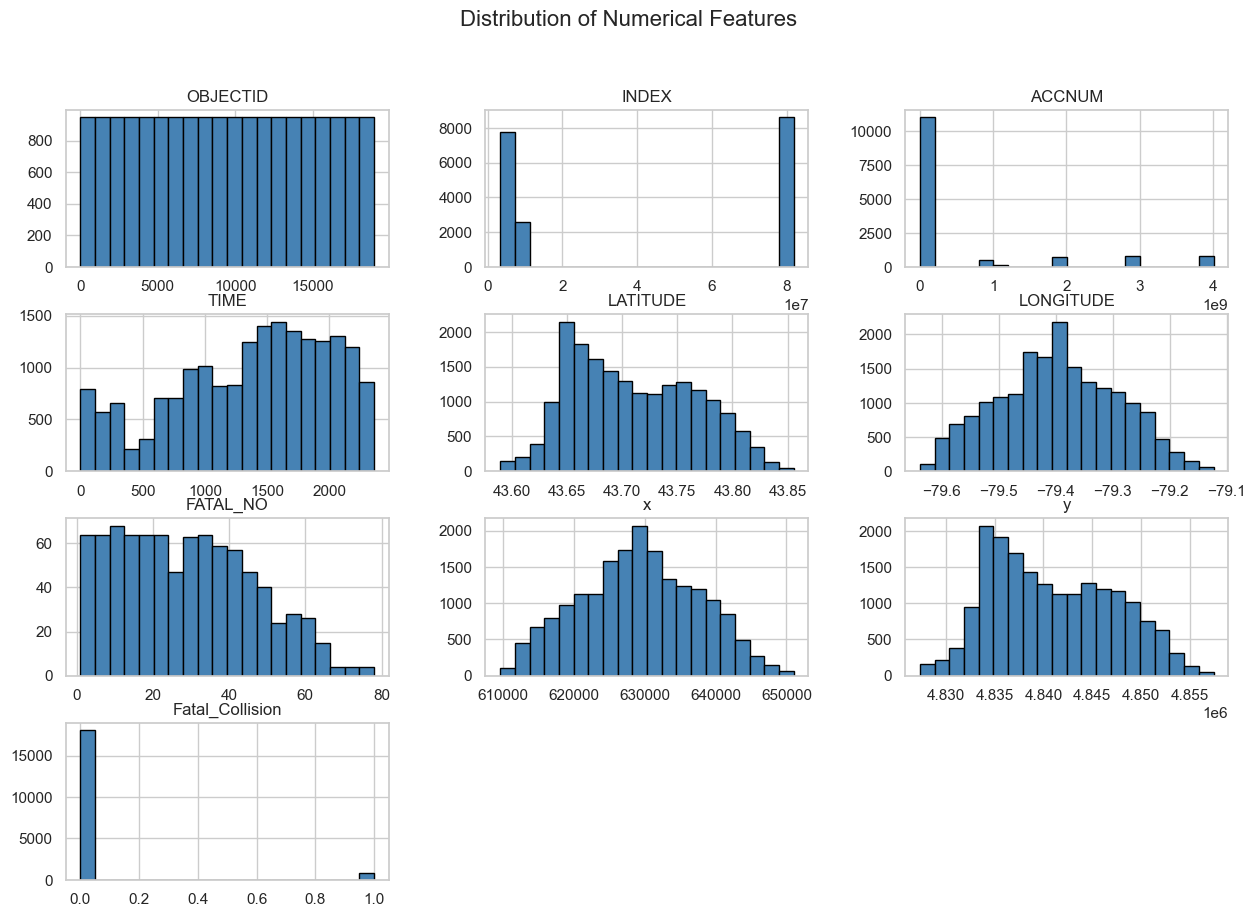
1. **TIME** – TIME has a wide range, with the interquartile range (IQR) roughly between **1000 and 2000** hours, indicating that most collisions happen between mid-morning and early evening.
2. **FATAL\_NO** – The values for FATAL\_NO are tightly clustered near zero, with a few minor variations, suggesting that most collisions are not fatal.
3. **Fatal\_Collision** – Fatal\_Collision is highly skewed, with most values close to zero and one visible outlier, indicating that fatal collisions are relatively rare.

A graph with a blue rectangle and a blue rectangle

AI-generated content may be incorrect.

**Distributions of Numerical Features**

1. **TIME** – TIME shows a multimodal pattern, with notable peaks during **morning and evening rush hours**.
2. **LATITUDE and LONGITUDE** – Both LATITUDE and LONGITUDE follow roughly normal distributions, indicating that most collisions are concentrated around specific geographic areas.
3. **FATAL\_NO** – The distribution of FATAL\_NO is right-skewed, with most values clustered near zero, indicating that the majority of collisions are non-fatal.
4. **Fatal Collision** – The distribution shows that most values are zero, with very few instances of fatal collisions, suggesting that fatal collisions are rare.



**4.4. Missing Values Heatmap**

**Visualization:** A heatmap was created to visualize missing values across all columns.

* Columns like **EMERG\_VEH, DISABILITY, and CYCCOND** have a high percentage of missing values.
* Other columns, such as **VEHTYPE, PEDCOND, and CYCACT**, also contain moderate missing data.
* **ACCNUM** appears to have minimal missing values, contrary to earlier assumptions.
* The missing values were carefully addressed during preprocessing.

A yellow and purple lines

AI-generated content may be incorrect.

**4.5. Correlation Heatmap**

* LATITUDE & y and LONGITUDE & x are almost the same, so one set could be removed.
* OBJECTID and INDEX are very similar (0.88 correlation), meaning one might not be needed.
* FATAL\_NO (fatalities) doesn’t strongly relate to other numbers, so factors like road conditions or driver behavior might play a bigger role

A screenshot of a computer

AI-generated content may be incorrect.

**4.6. Distribution of Fatal vs. Non-Fatal Collisions**

* The dataset is highly imbalanced, with non-fatal collisions far outnumbering fatal ones.
* This imbalance may impact model performance, requiring techniques like SMOTE to balance the classes if needed.

A graph of a number of injuries

AI-generated content may be incorrect.

**Key Findings:**

**Missing Data:**

* A lot of the dataset has missing values, especially in columns about injury severity (INJURY), fatalities (FATAL\_NO), alcohol involvement (ALCOHOL), speeding (SPEEDING), and others like PEDESTRIAN and CYCLIST.
* These missing values should be handled by filling them in, removing them, or using other methods, depending on the analysis or machine learning model used.

**Categorical Variables:**

* Many categorical variables have few unique values (like ROAD\_CLASS, DISTRICT, TRAFFCTL), which can be encoded for modeling.
* Some categories in columns like INJURY, FATAL\_NO, and VEHICLE type can strongly indicate accident severity and vehicle involvement.

**Geospatial Data:**

* The dataset includes LATITUDE and LONGITUDE data, useful for geospatial analysis to find accident hotspots or patterns in specific areas.

**Traffic Conditions:**

* Columns like TRAFFCTL (Traffic Control) and RDSFCOND (Road Surface Condition) are important for understanding accident causes. For example, accidents on "Wet" or "Slush" roads are more likely to be severe.

**Injury Analysis:**

* The INJURY column has many missing entries, which could bias the analysis unless properly handled.

**Conclusion:**

This analysis reveals important patterns in the traffic accident data, offering valuable insights into accident severity, contributing factors, and potential predictive features. Key findings include:

* **Class Imbalance:** There are many more non-fatal accidents than fatal ones, so techniques like SMOTE are needed to balance the dataset.
* **Missing Data:** It's crucial to handle missing values, especially in important columns, for accurate modeling.
* **Geospatial Insights:** Geographic data (LATITUDE and LONGITUDE) can be used to identify accident hotspots.
* **Traffic Conditions:** Road surface condition and traffic control are key factors in understanding accident severity.