Table of Contents

[Executive Summary 2](#_Toc193029419)

[Overview Of Solutions 3](#_Toc193029420)

[Data Exploration 4](#_Toc193029421)

[Feature Selection 9](#_Toc193029422)

[Data Modeling 10](#_Toc193029423)

[Model Building 11](#_Toc193029424)

# Executive Summary

# Overview Of Solutions

# Data Exploration

The initial data exploration, conducted using a Pandas DataFrame, revealed a dataset containing 18,957 entries across 54 columns. The data types were diverse, encompassing float64 (6 columns), int64 (3 columns), and object (45 columns), with a significant presence of missing values across numerous features. Class counts highlighted a substantial imbalance in accident severity, with a predominance of non-fatal injuries compared to fatalities and property damage.

Statistical assessments revealed insights into the distribution of numeric variables. For LATITUDE, LONGITUDE, x, and y, the close proximity of mean and median values suggested approximately symmetrical distributions. Conversely, TIME and FATAL\_NO exhibited larger disparities, indicating skewed distributions. Variance analysis further elucidated the spread of data: TIME and FATAL\_NO demonstrated high variance, reflecting a wide range of accident times and fatality counts. The large variances for x and y are likely due to coordinate scaling, while the low variances for LATITUDE and LONGITUDE suggested a concentration of accidents within a relatively confined geographical area. The mode analysis identified the most frequent values for each column, highlighting dominant categories and trends.

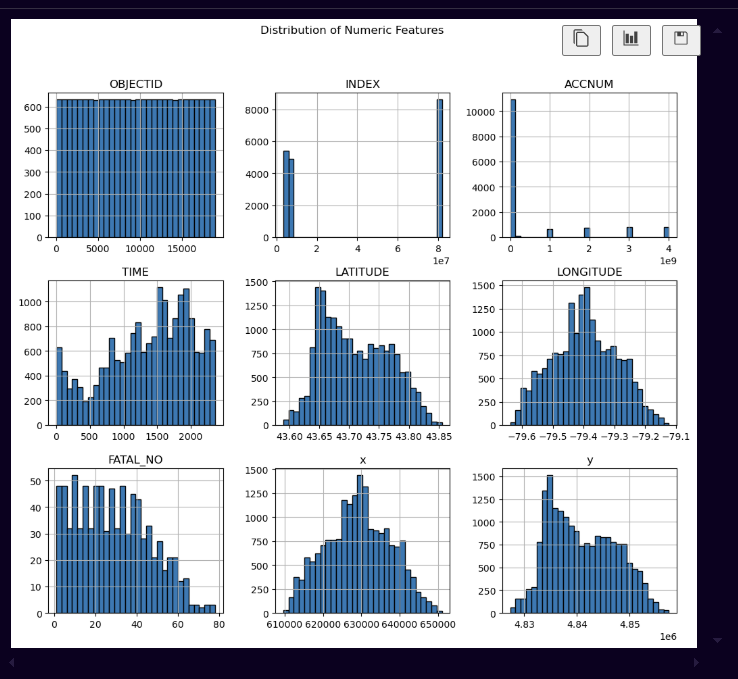
A missing values assessment revealed extensive gaps in the data, with several columns exhibiting severe to moderate levels of missingness. Notably, FATAL\_NO and numerous columns related to accident details and participant characteristics showed extremely high rates of missing data. This underscored the need for robust data cleaning and imputation strategies to ensure accurate and reliable analysis. The correlation heatmap also showed that the numerical columns had some linear relationships, but the object columns, which contained the most missing data, were not able to be analyzed in this way. The visualization of the data was further explored using various types of graphs.

A blue and white barcode

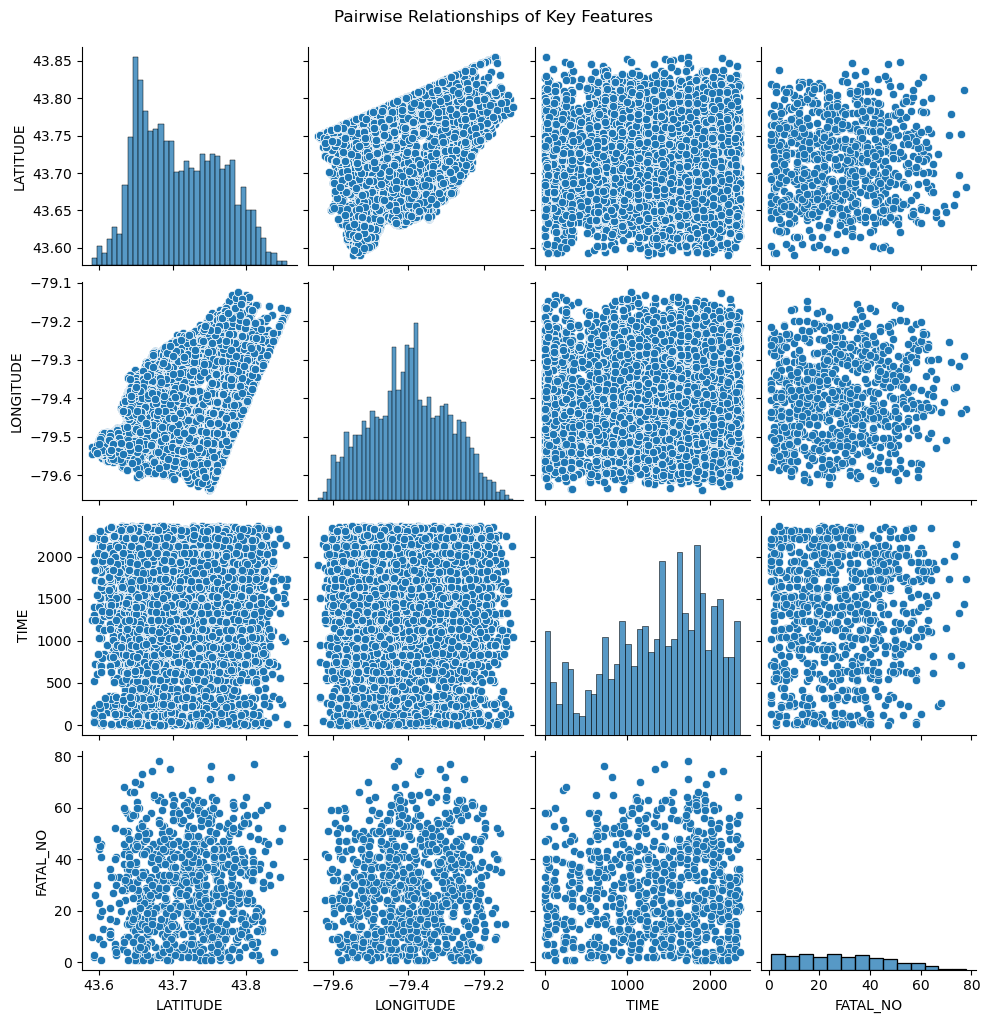
AI-generated content may be incorrect.

A boxplot was used to identify outliers in each numerical column. Most columns exhibited a compressed range of values, except for INDEX and ACCNUM, which displayed several outliers above and below their respective distributions. This visualization suggested the need for data transformation or normalization to improve clarity.A graph with black dots and green lines

AI-generated content may be incorrect.



The histogram of TIME indicated that accidents occur fairly evenly throughout the day. LONGITUDE and LATITUDE histograms confirmed that a big chunk of accidents are clustered at a specific geographical area.The FATAL\_NO histogram showed a right-skewed distribution, indicating a lower frequency of high fatality accidents. The x and y histograms also displayed bell-shaped distributions, mirroring LONGITUDE and LATITUDE and suggesting they are highly, if not directly correlated.



The pairplot revealed a strong negative linear relationship between LATITUDE and LONGITUDE, confirming a clear geographical trend in accident locations. Both coordinates exhibited unimodal, approximately normal distributions, suggesting clustering around a central point. However, TIME and FATAL\_NO showed scattered patterns with no discernible linear relationships when paired with any of the other variables. This implied that accident times and fatality counts are not strongly linearly correlated with location or with each other in this dataset.

A screenshot of a graph

AI-generated content may be incorrect.

The correlation heatmap revealed relationships between numerical variables in the accident data. Perfect correlations (1.00) between LATITUDE/y and LONGITUDE/x indicated these are identical coordinates presented differently. Strong positive correlations were observed between OBJECTID and INDEX (0.88), and INDEX and ACCNUM (0.78), suggesting these are related identifiers. A moderate positive correlation (0.42) existed between LATITUDE and LONGITUDE, reflecting their geographical association. However, TIME and FATAL\_NO showed very weak correlations with all other variables, implying limited linear relationships. This suggested that these factors may be influenced by non-linear relationships or other variables not captured in this heatmap. Additionally, the heatmap only reflected linear relationships between numerical data, and did not show relationships between non-numerical data.A graph of a number of blue rectangular objects

AI-generated content may be incorrect.

The class distribution chart highlighted the outcomes of accidents, clearly showing a predominance of non-fatal injuries compared to fatalities and property damage incidents. This class imbalance necessitated the use of SMOTENC during data modeling to ensure a balanced dataset for training the model.

# Feature Selection

For feature selection, we removed redundant and irrelevant columns to refine the dataset for subsequent analysis in Part 2. Specifically, OBJECTID, INDEX, and ACCNUM, which served as identifiers, were dropped. Additionally, x and y were removed due to their perfect correlation with LONGITUDE and LATITUDE, as evidenced by the correlation heatmap, indicating they represent the same data. This had to be done to avoid collinearity. DIVISION was dropped as well as it represents the police division under whose jurisdiction the accident occurred and is not relevant in predicting the outcome.

More features will likely be dropped in the process of training the models in deliverable 2 as a result of analyzing the models’ feature importance.

# Data Modeling

This section outlines the data modeling process to prepare the data for model training.

* **Initial Feature Reduction:** Redundant columns (OBJECTID, INDEX, ACCNUM, x, y, DIVISION) were dropped as a preliminary form of feature selection, with a plan for more advanced feature selection later.
* **Target Definition & Data Cleaning:** The target variable (ACCLASS) was defined, and rows with missing target values were removed.
* **Data Splitting & Feature Identification:** The data was split into features (X) and the target (y), and categorical and numerical features were identified.
* **Imputation:** The missing values were filled using mean and most frequent category for numerical and categorical data respectively. This had to be done before resampling as SMOTENC doesn’t work with missing values.
* **Resampling:** SMOTENC was used to balance the class distribution in the dataset. We chose it because the dataset consists of mostly categorical data, which SMOTENC handles well unlike SMOTE or ADASYN.
* **Scaling:** Numerical data was scaled using StandardScaler to make all features follow the same value range and not overpower each other.
* **Encoding:** One hot encoder was used to convert categorical data into numerical.

# Model Building

This section details the model building process, focusing on training and evaluation.

* **Pipeline Creation:** A machine learning pipeline was constructed to streamline the workflow, encapsulating imputation, resampling (SMOTENC), encoding/scaling, and classification.
* **Model Selection:** Logistic Regression was chosen as the classification algorithm for initial pipeline testing. More models will be tried in the next deliverable.
* **Results:** The model performed relatively well with accuracy scores of 98 on training data and 94 on testing data, suggesting that the preprocessing steps were successful.