

Traffic Violation Analysis

Traffic Violations in USA

Group - 8

Harshavardhan Kuruvella - 01753482

Sai Dheeraj Gollu - 017520503

Abhinav Sriharsha Anumanchi - 017514900

Introduction

This project focuses on **Traffic Violations Data Analysis** and aims to explore traffic violation patterns through data-driven insights. Using a combination of statistical and machine learning techniques, the goal is to classify and cluster traffic violations, uncover correlations with various factors (e.g., vehicle condition, time of day), and visualize geographical trends.

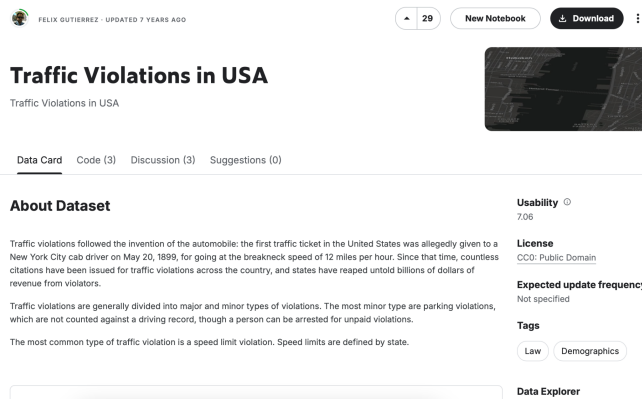
Key Objectives:

Classification: Predict the **Violation Type** based on various factors like time, day, vehicle condition, and location.

Clustering: Group traffic violations into clusters based on geographical location to identify high-risk areas.

Geospatial Analysis: Analyze traffic violations and contributing factors based on geographical data.

Feature Exploration: Analyze relationships between violation types and other features like gender, race, and vehicle age.



Data Description

- **Dataset Name:** Traffic Violations
- **Total Records:** 1,018,634
- **Total Features:** 35 columns
- **Size:** 272 MB

Missing Data:

- Columns with Missing Values:
 - **Latitude, Longitude:** ~84,814 missing values.
 - **Vehicle Year:** ~6,340 missing values.
 - **Color:** ~13,558 missing values.
 - **Geolocation:** ~84,627 missing values.

Key Features:

Feature	Description
Date of Stop	Date when the violation occurred
Time of Stop	Time when the violation occurred
Violation Type	Type of violation (e.g., Citation, Warning, SERO)
Vehicle Condition	Categorized as New, Old, Very Old, or Antique
Race	Race of the driver
Gender	Gender of the driver
Latitude, Longitude	Geographical location of the violation
Arrest Type	How the violation was handled (e.g., Marked Patrol, Laser)
Geolocation	Combined Latitude and Longitude in text form

Literature Survey

Traffic Violations and Machine Learning: Studies have used **Random Forest**, **Decision Trees**, and **SVMs** to predict traffic violations based on various factors.

Geospatial Analysis: **GIS** and **K-Means clustering** help identify high-risk areas for traffic violations and accidents.

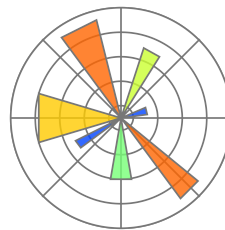
Vehicle Condition and Violations: Research shows that **vehicle age** influences the likelihood of accidents and violations.

Time-based Analysis: **Time of day** and **day of the week** are key factors in predicting traffic violations, aiding in efficient law enforcement.



Technologies Used

- Python:** Primary programming language for data processing and model building.
- Pandas:** Used for data manipulation, cleaning, and transformation.
- Scikit-learn:** Machine learning library used for classification models (Random Forest, Decision Tree, Logistic Regression).
- Seaborn & Matplotlib:** Visualization libraries for data exploration and result presentation.
- Geopandas:** Used for geospatial data analysis and mapping of traffic violations based on geographic locations.
- KMeans Clustering:** To group traffic violations into clusters based on geographical locations.
- Label Encoding:** Converts categorical variables (e.g., Race, Gender) into numerical format for model training.



Methodology

Data Collection & Preprocessing:The dataset was imported, cleaned, and checked for missing values or duplicates.Relevant columns like **Latitude**, **Longitude**, and **Geolocation** were processed to extract useful geographical data.

Feature Engineering:Created new features like **Vehicle Condition** based on vehicle year.Extracted **Day**, **Month**, and **Year** from the **Date of Stop** for time-based analysis.Encoded categorical variables (e.g., **Race**, **Gender**, **Violation Type**) using **Label Encoding**.

Exploratory Data Analysis (EDA):Visualized distributions of violations based on factors like **Day of the Week**, **Hour of Stop**, **Vehicle Condition**, etc.Identified patterns using **count plots**, **heatmaps**, and **scatter plots**.

Vehicle Condition	Year Range	Description	Proportion
New	2015 and later	Modern vehicles with latest features	15%
Old	2000 - 2014	Well-maintained older vehicles	70%
Very Old	1950 - 1999	Aged vehicles prone to more issues	12%
Antique	Before 1950	Classic or collectible vehicles	3%

Gender	Label Encoding
Male	1
Female	0

Methodology Cont..

Clustering

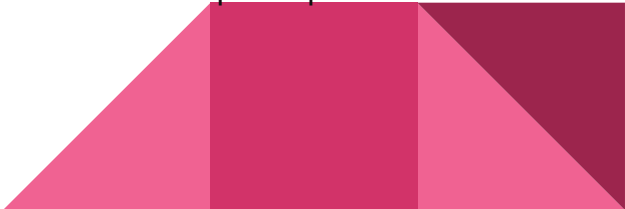
Applied **KMeans clustering** to group traffic violation data based on geographical locations (Latitude & Longitude).

Model Training:

Split data into training and test sets. Trained multiple classifiers (Logistic Regression, Decision Tree, and Random Forest) on the preprocessed dataset.

Model Evaluation:

Evaluated model performance using metrics like **accuracy**, **precision**, **recall**, and **F1-score**. Compared performance across different classifiers.



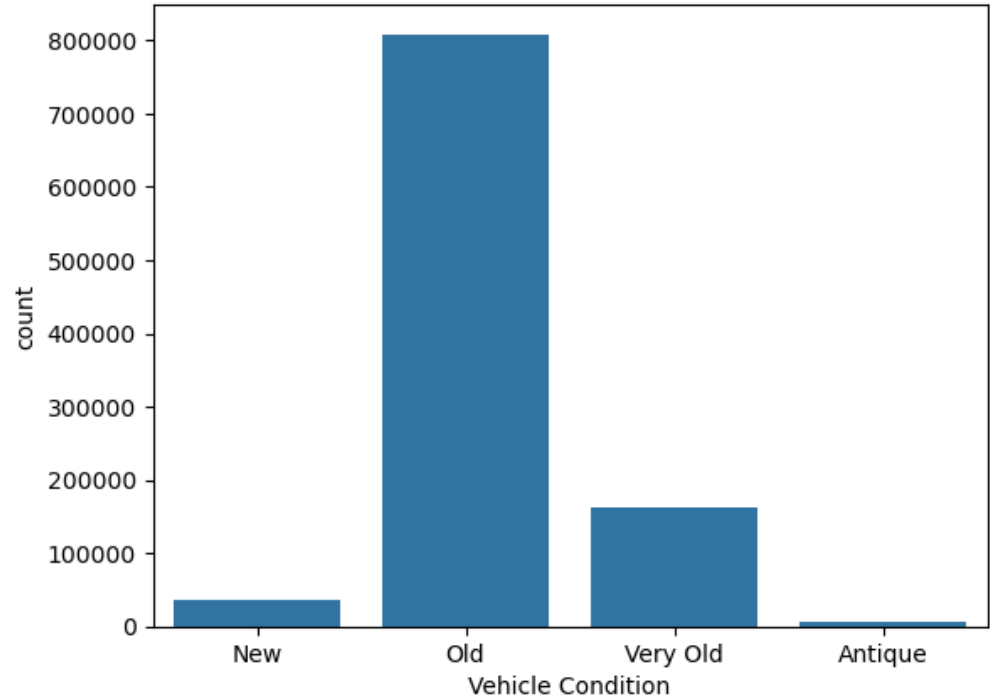
Results

Arrest Type

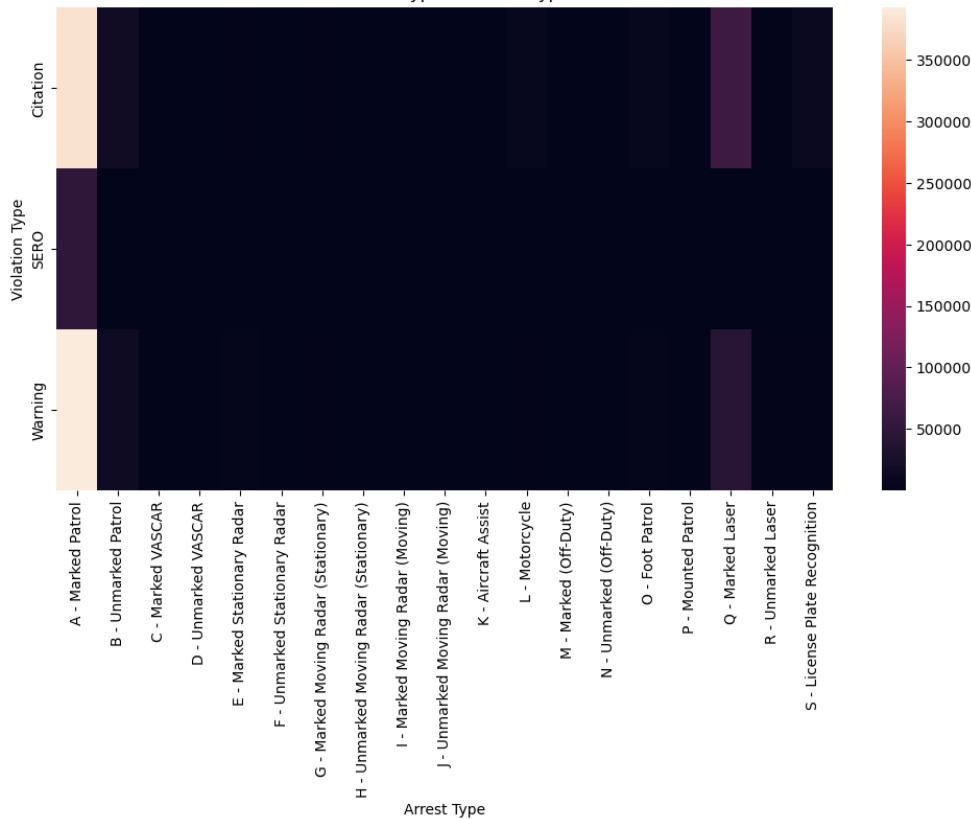
A – Marked Patrol	825129
Q – Marked Laser	104953
B – Unmarked Patrol	32378
S – License Plate Recognition	12530
O – Foot Patrol	10894
L – Motorcycle	9955
E – Marked Stationary Radar	6354
R – Unmarked Laser	4903
G – Marked Moving Radar (Stationary)	3592
M – Marked (Off-Duty)	1563
I – Marked Moving Radar (Moving)	1405
F – Unmarked Stationary Radar	663
H – Unmarked Moving Radar (Stationary)	461
C – Marked VASCAR	379
D – Unmarked VASCAR	226
J – Unmarked Moving Radar (Moving)	225
P – Mounted Patrol	210
N – Unmarked (Off-Duty)	151
K – Aircraft Assist	44

Name: count, dtype: int64

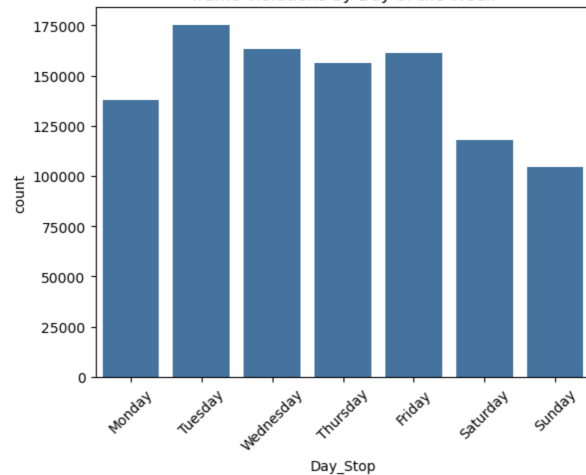
Distribution of Vehicle Conditions



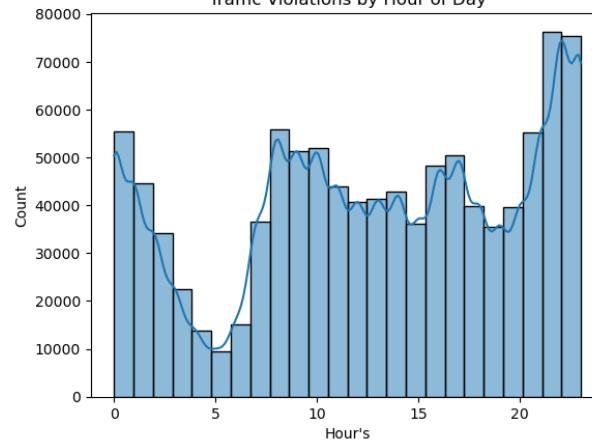
Violation Type vs Arrest Type

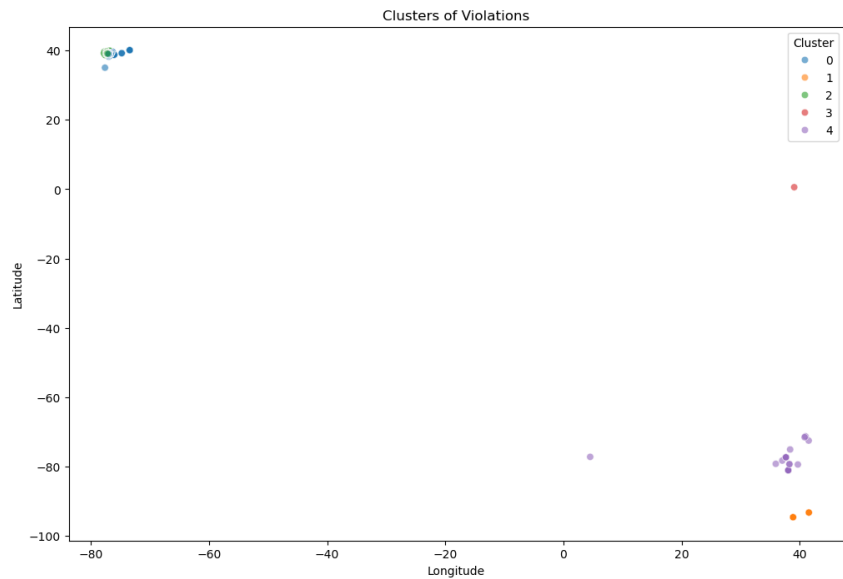
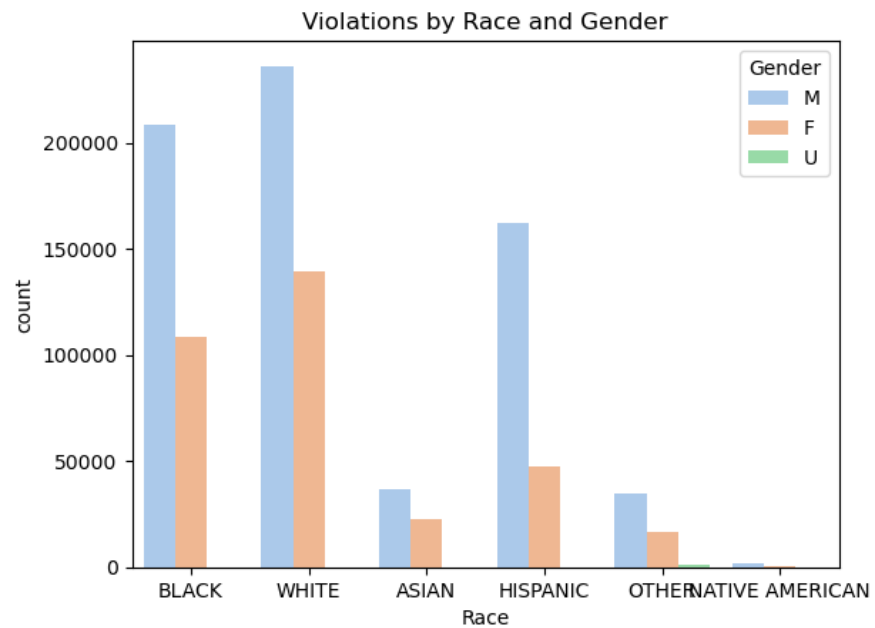


Traffic Violations by Day of the Week

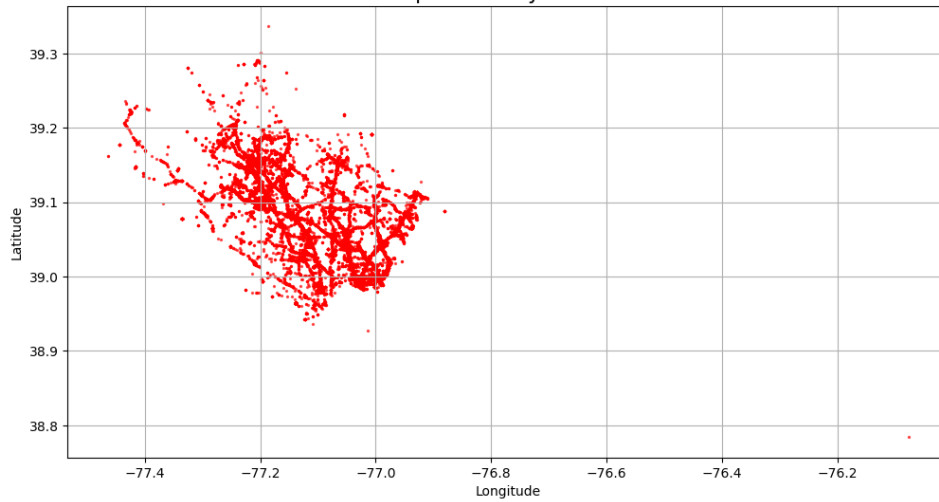


Traffic Violations by Hour of Day

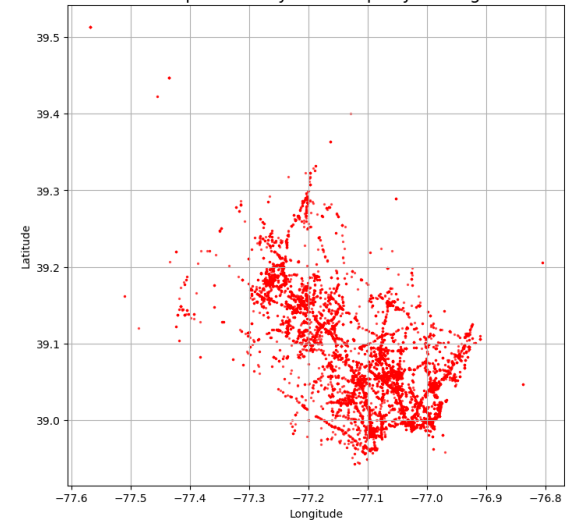




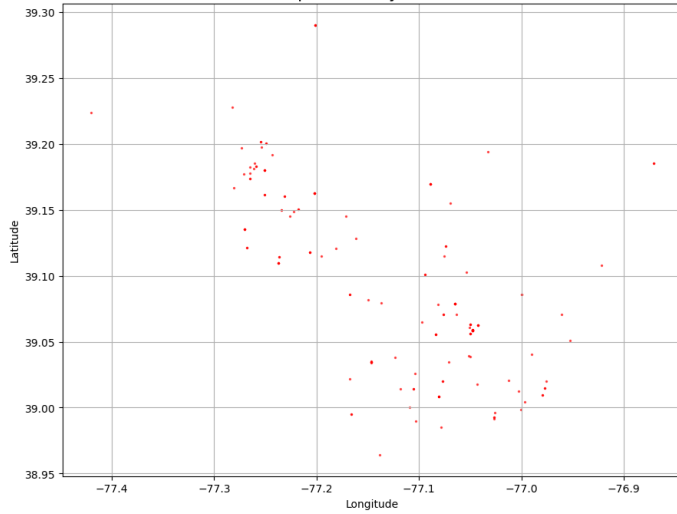
Geospatial Analysis of Belts



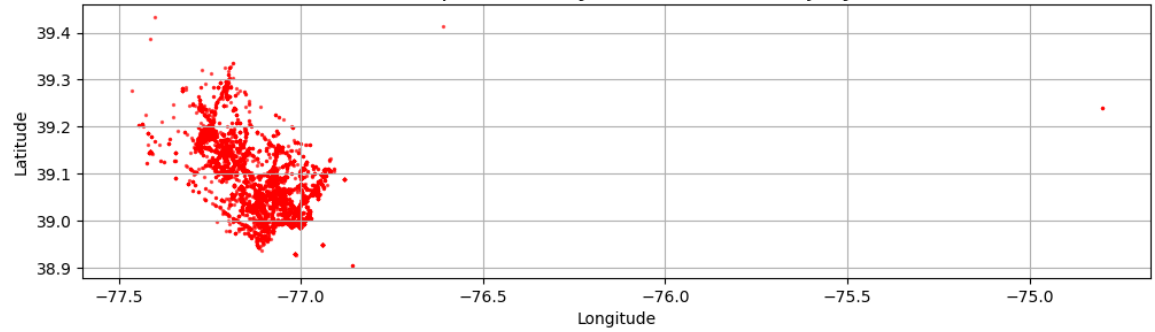
Geospatial Analysis of Property Damage



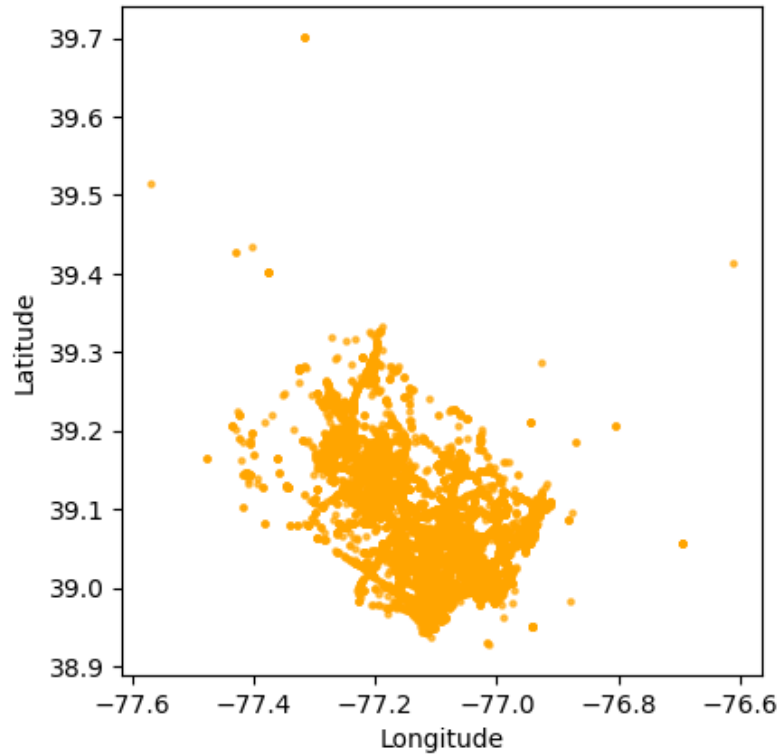
Geospatial Analysis of Fatal



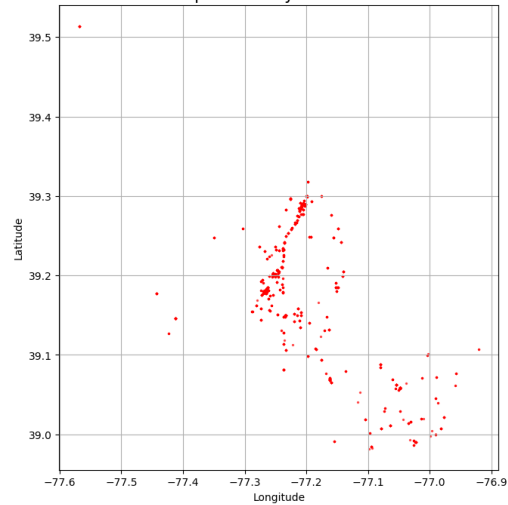
Geospatial Analysis of Personal Injury



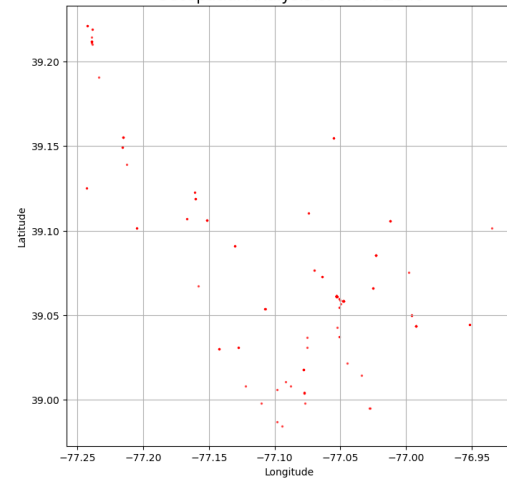
Geospatial Analysis of Contributed Accidents



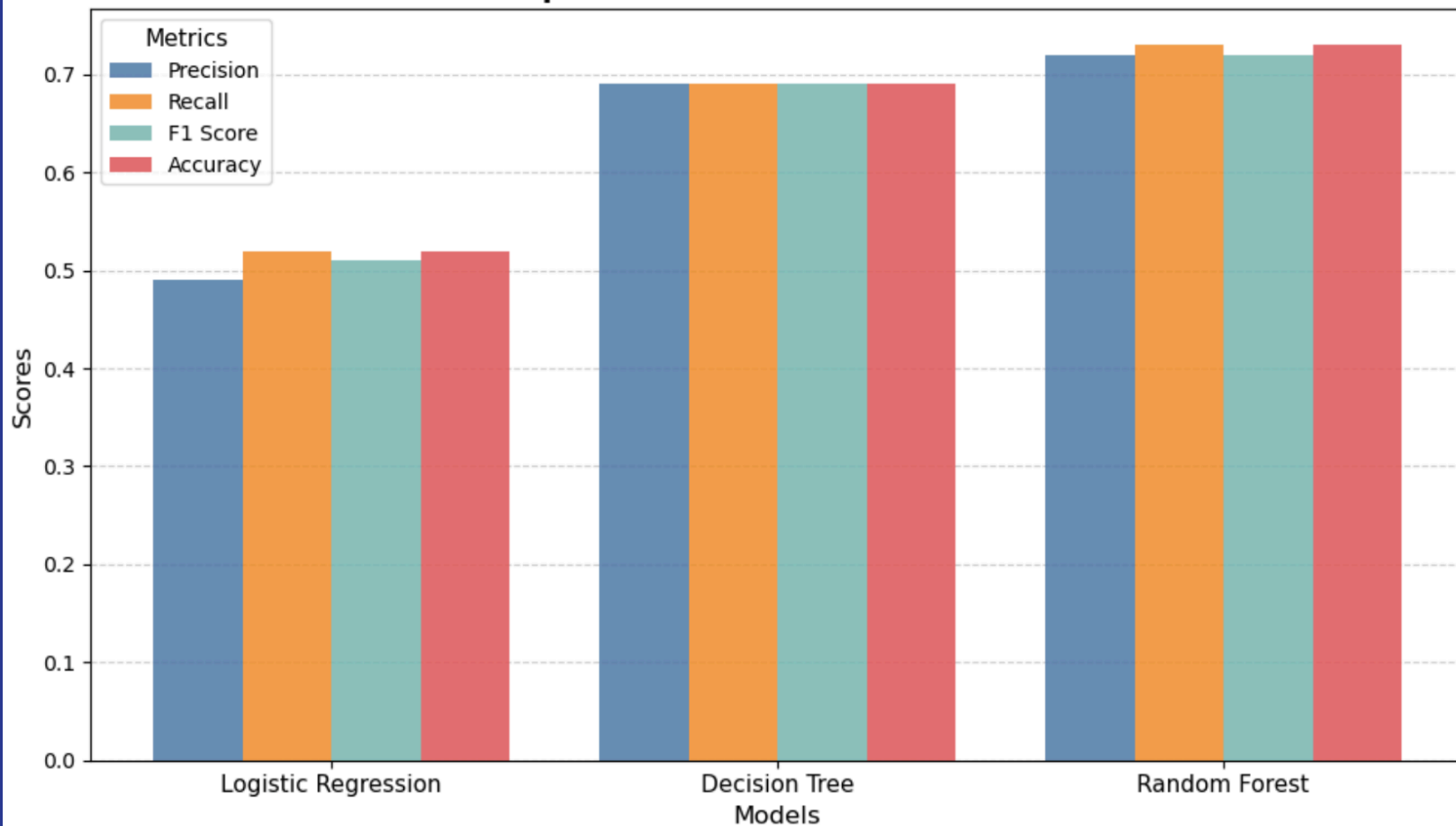
Geospatial Analysis of Alcohol



Geospatial Analysis of Work Zone



Comparison of Metrics Across Models



Conclusion

1. **Data Insights:** The analysis uncovered patterns in traffic violations influenced by vehicle condition, time of day, day of the week, race, and gender.
2. **Geospatial Analysis:** Clustering revealed high-risk geographical hotspots, enabling targeted enforcement strategies.
3. **Model Performance:**
 - Random Forest:** Achieved the highest accuracy (73%) due to its ability to handle complex and non-linear relationships.
 - Decision Tree:** Performed well with 69% accuracy by effectively capturing non-linear patterns, but overfits the data
 - Logistic Regression:** Achieved the lowest accuracy (52%) as it assumes linear relationships, which are not suitable for the dataset's complexity.
4. **Feature Engineering:** Created new features like Vehicle Condition and time-based attributes, which significantly improved model predictions.
5. **Impact:** Insights can assist in prioritizing resources, reducing violations in high-risk areas, and forming a foundation for predictive road safety measures.



Future Scope

Real-time Data Integration: Future work can focus on integrating **real-time traffic violation data** for dynamic predictions and responses.

Advanced Model Development: Explore more advanced models like **XGBoost** or **Neural Networks** for improved prediction accuracy.

Geospatial Optimization: Utilize **advanced geospatial techniques** like clustering with different algorithms (e.g., DBSCAN) to better understand violation hotspots.

Actionable Insights: Develop an **interactive dashboard** for law enforcement agencies to visualize violation trends and hotspots for quick decision-making.



Thank You

