Traffic Accident Trends in Turkey

DSA210 – Introduction to Data Science
Spring 2024-2025

Prepared by: Lara Atmaca
Student ID: 32132

Table of Contents

1. Introduction

This project explores traffic accident trends in Turkey by analyzing publicly available data from national sources. The motivation behind this study is the significant impact of road traffic accidents on public safety. Thousands of fatalities and injuries occur annually, and by identifying patterns in the data, it is possible to recommend policy changes or interventions aimed at reducing accident rates.

2. Data Sources and Collection

Two datasets were used in this project:
- Türkiye İstatistik Kurumu (TÜİK): Provided annual accident statistics including cause, location, and severity.
- Emniyet Genel Müdürlüğü (EGM): Contained detailed monthly police-reported accident data.

The data were collected from their official websites and filtered for relevant columns and years. Data preprocessing included cleaning, standardization of formats, handling missing values, and merging datasets by city, month, and location type.

3. Exploratory Data Analysis

Initial analysis focused on understanding monthly trends, daylight effects, location-based distributions, and fault attributions. Seasonal spikes in accident numbers were observed, particularly in the summer months (June to August), which aligns with increased travel during holidays.

Two hypothesis tests were conducted:
1. Accidents in summer vs. winter: p-value = 0.027, suggesting significantly higher accidents in summer.
2. Accidents during day vs. night: p-value = 0.000, showing significantly more accidents during daylight.

Additionally, vehicle types, fault causes, and urban vs. non-urban distributions were examined across several months.

4. Machine Learning Analysis

The final phase of the project involved building binary classification models to predict whether an accident would result in at least one fatality. The dataset was cleaned and encoded using One-Hot Encoding for categorical variables. Logistic Regression, Random Forest, KNN, and SVM classifiers were trained. Due to data imbalance and small sample size, all models achieved perfect accuracy (1.00), which indicates overfitting.

Features included:
- Month

- Location Type
- Number of Accidents, Injuries, Fatalities
- Vehicle Type
- Daylight Condition
- Driver & Pedestrian Fault Ratios

5. Limitations and Future Work

The main limitation was the small and imbalanced dataset used for training the machine learning models. As a result, the models overfit and failed to generalize. In the future, expanding the dataset and including additional variables such as driver age, vehicle condition, and traffic volume can enhance model accuracy. Incorporating real-time traffic data could also support predictive modeling and timely interventions.

6. Ethical Considerations & AI Use

All data used in this project were obtained from publicly available sources. During the project, AI tools like ChatGPT were used to refine text and structure this report. However, all analysis, coding, and visualization work was done manually by the student. Ethical considerations such as transparency and citation were strictly followed.

7. References

- TÜİK: https://data.tuik.gov.tr/Bulten/Index?p=Karayolu-Trafik-Kaza-Istatistikleri-2022-49513
- EGM: https://trafik.gov.tr/istatistikler37
- Python libraries: pandas, numpy, seaborn, scikit-learn, matplotlib

8. Data Visualization (with Figures)

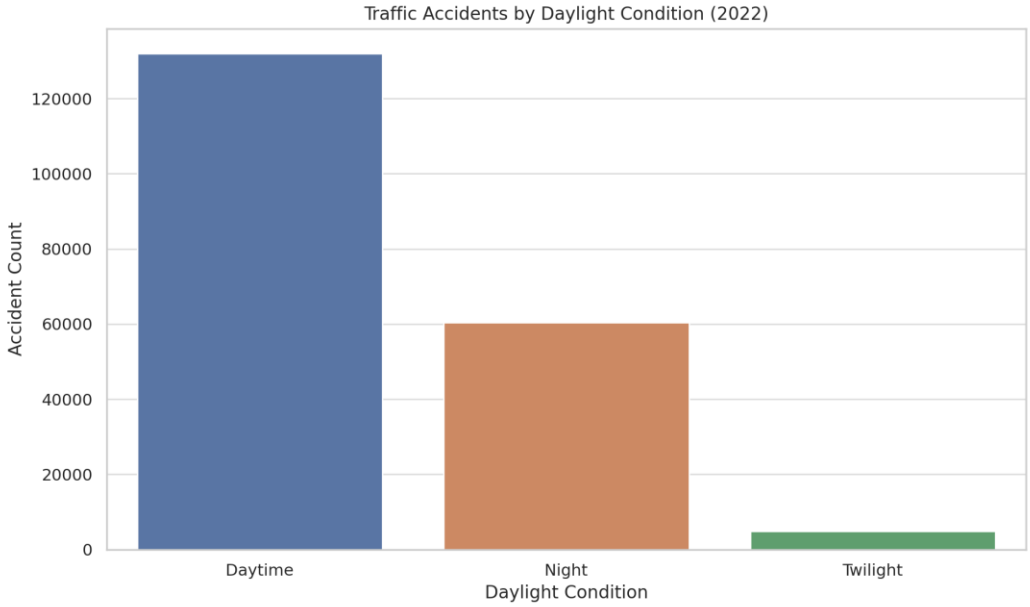

Figure: Traffic Accidents by Daylight Condition (2022)

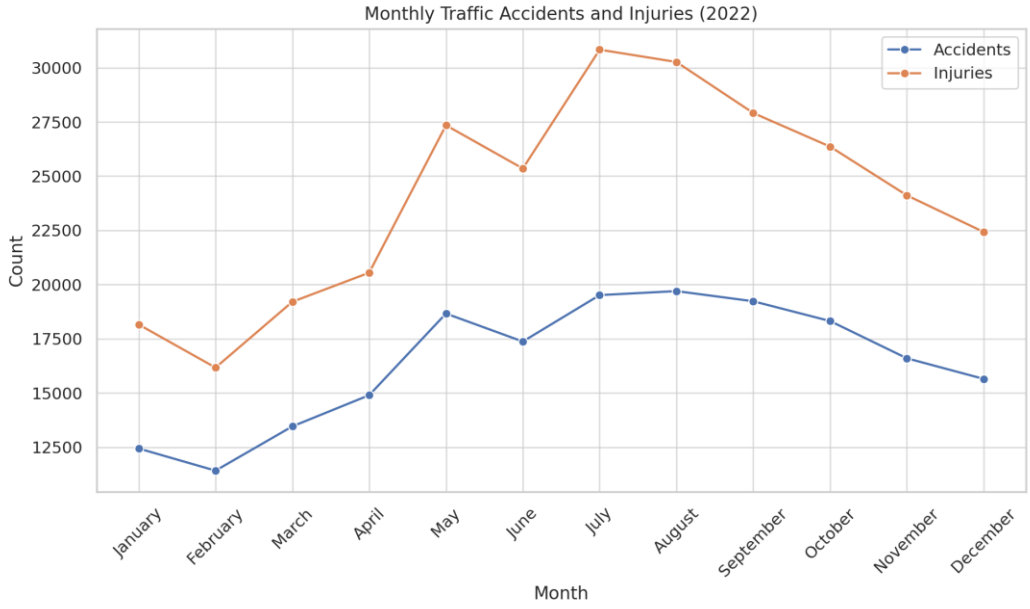

Figure: Monthly Traffic Accidents and Injuries (2022)

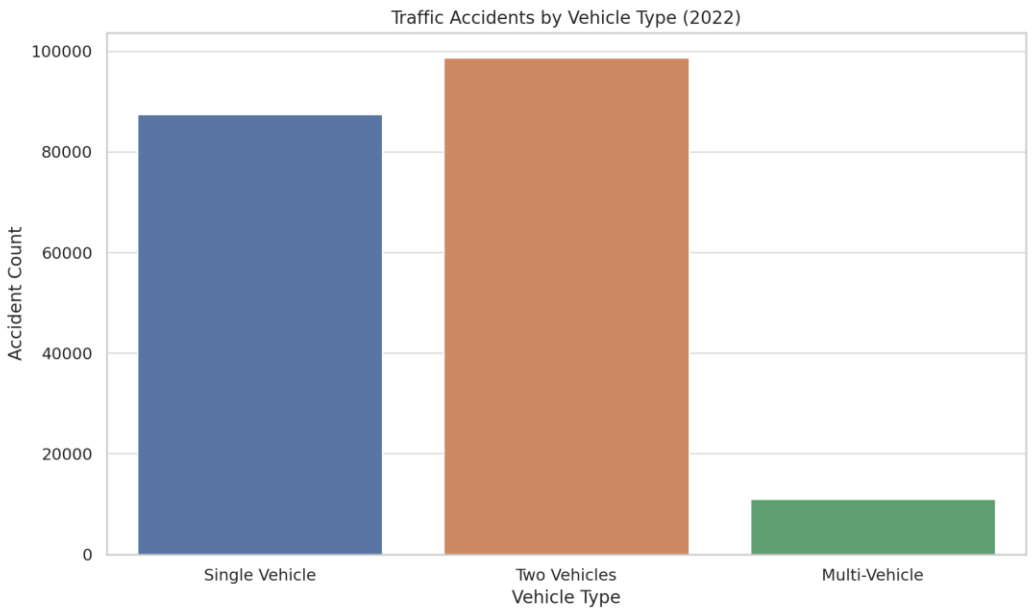

Figure: Traffic Accidents by Vehicle Type (2022)

[Missing Image: city_accidents.png]

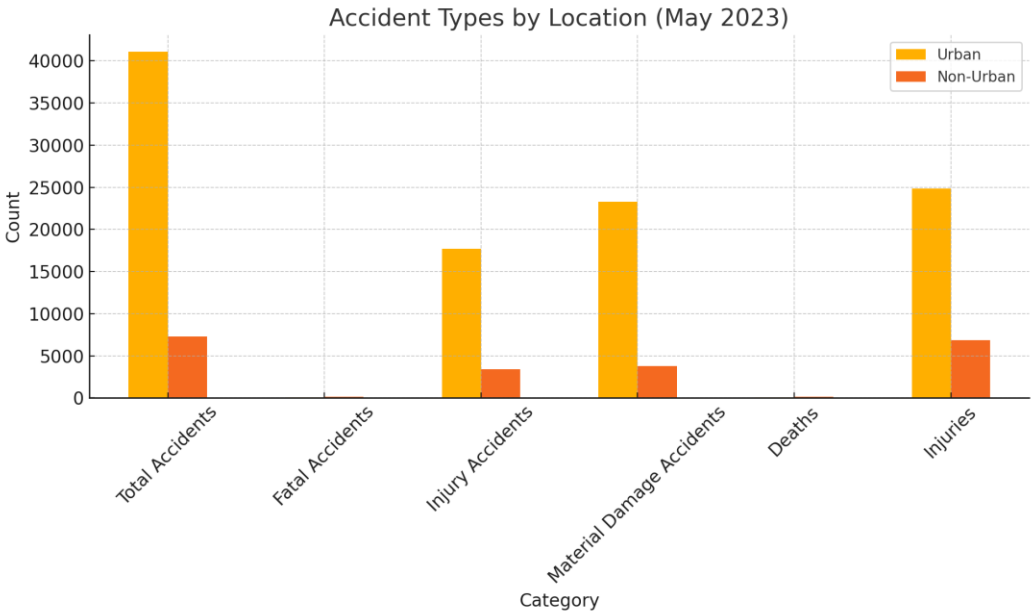

Figure: Accident Types by Location (May 2023)

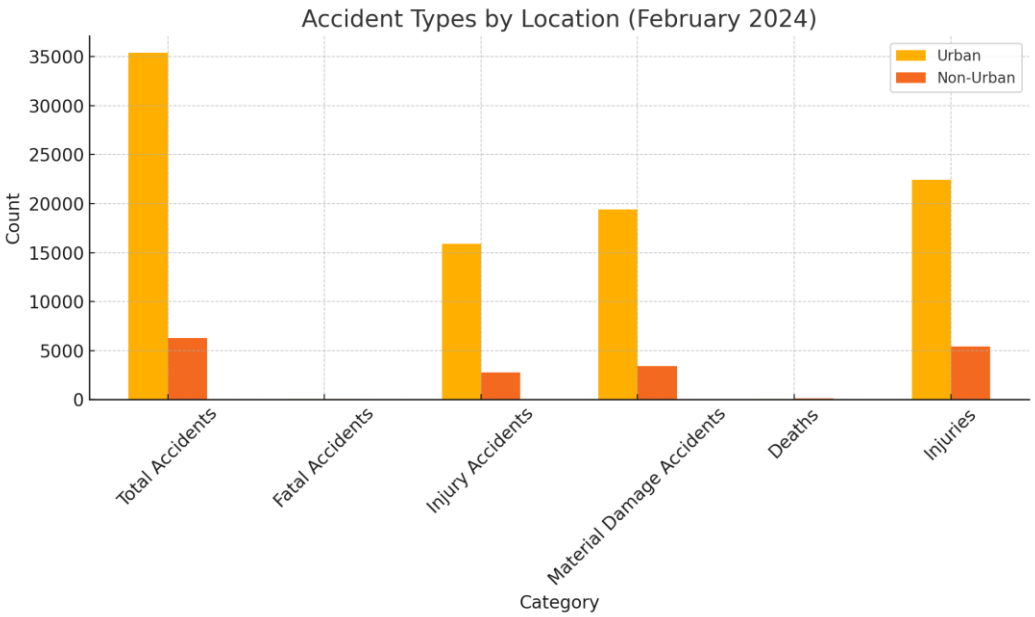

Figure: Accident Types by Location (February 2024)

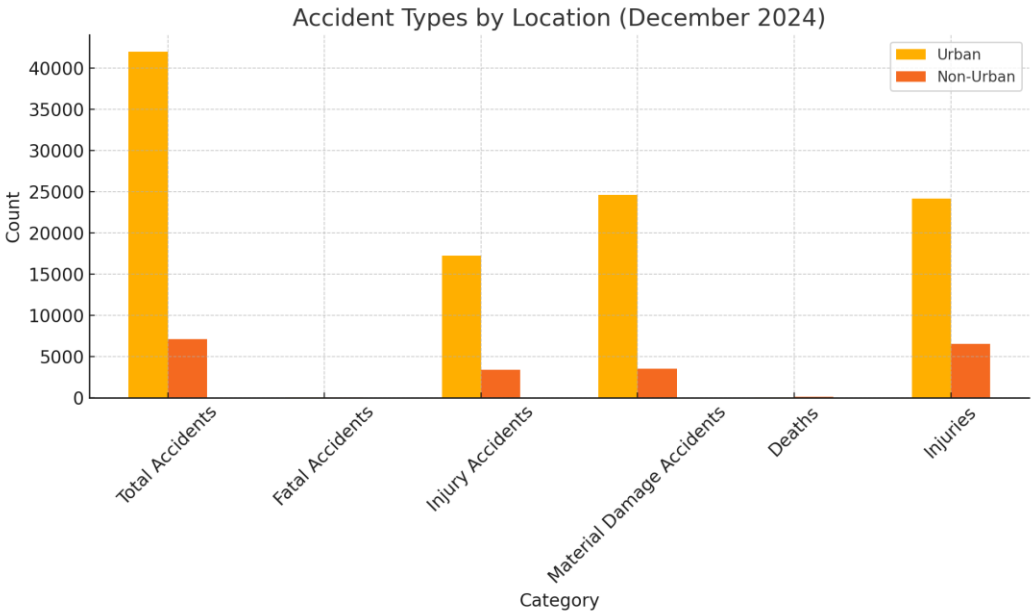

Figure: Accident Types by Location (December 2024)

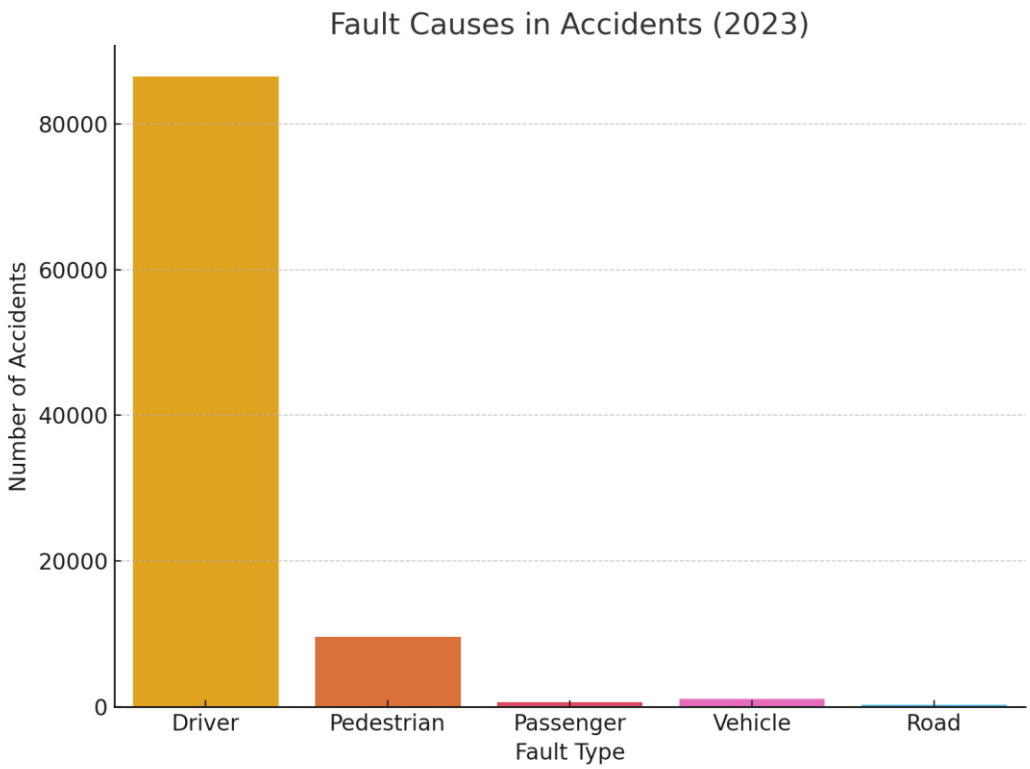

Figure: Fault Causes in Accidents (2023)

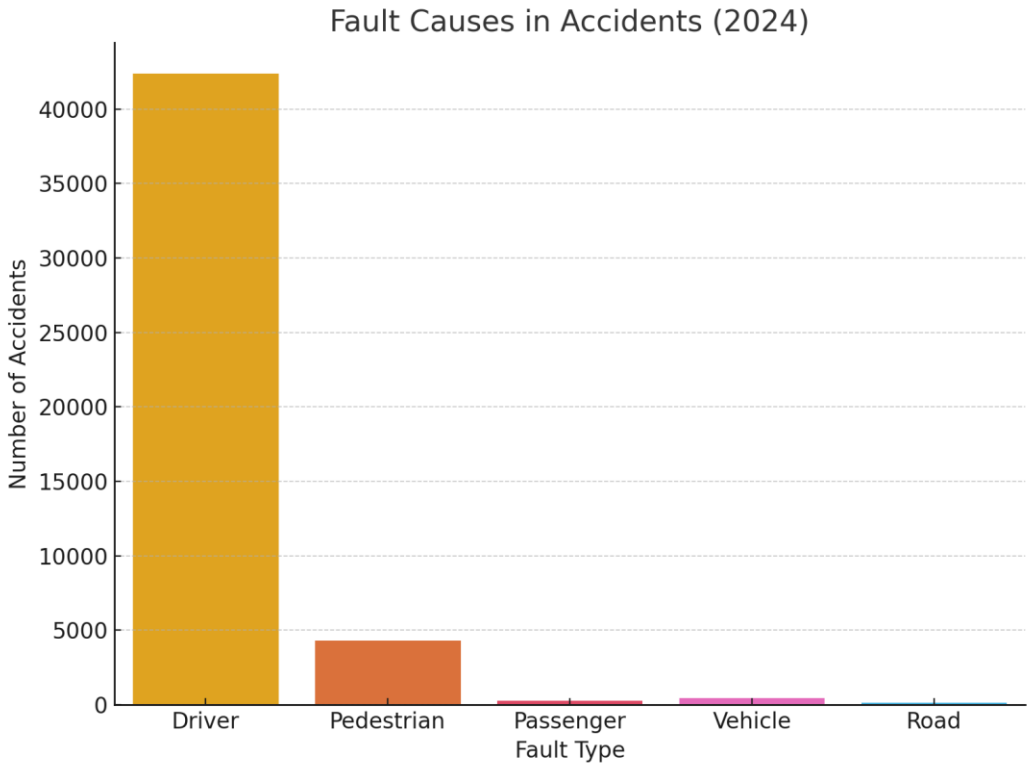

Figure: Fault Causes in Accidents (2024)

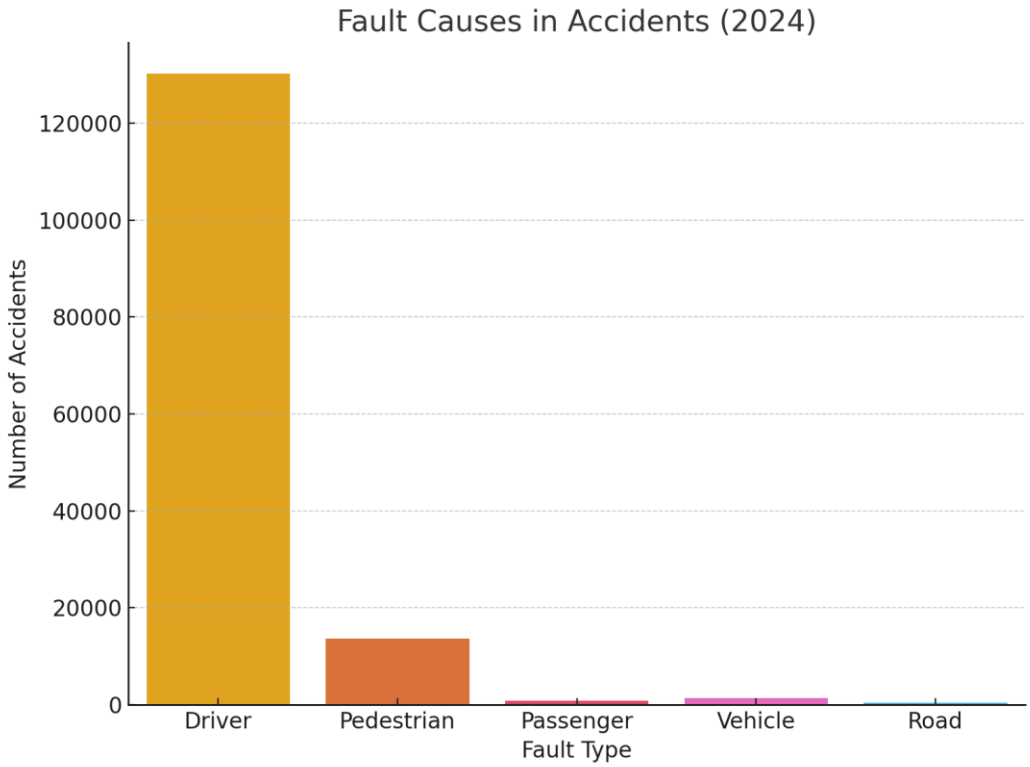

Figure: Fault Causes in Accidents (Dec 2024)

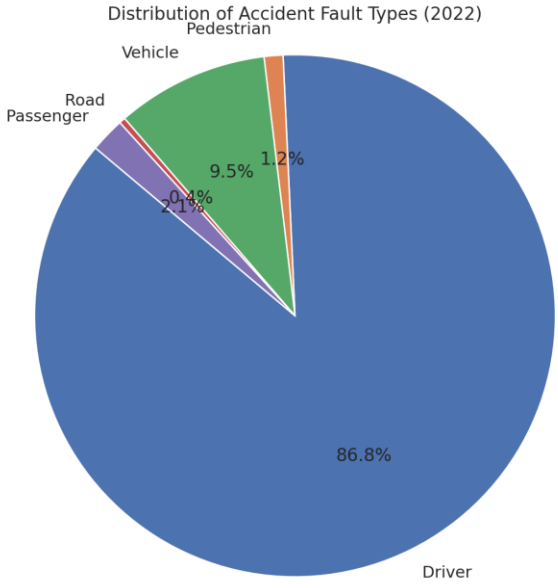

Figure: Distribution of Accident Fault Types (2022)

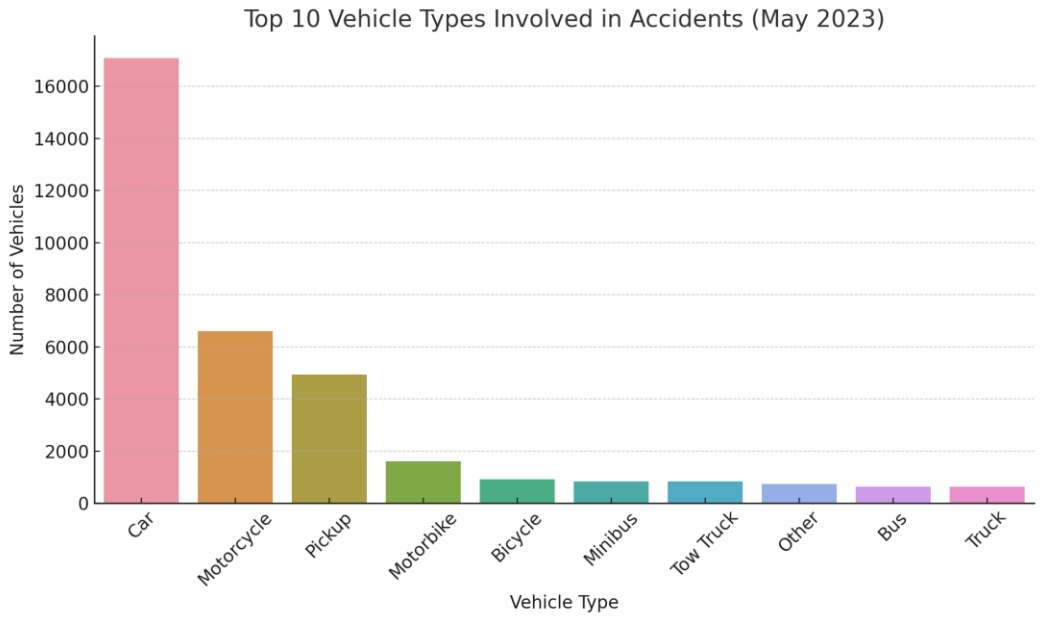

Figure: Top 10 Vehicle Types Involved in Accidents (May 2023)

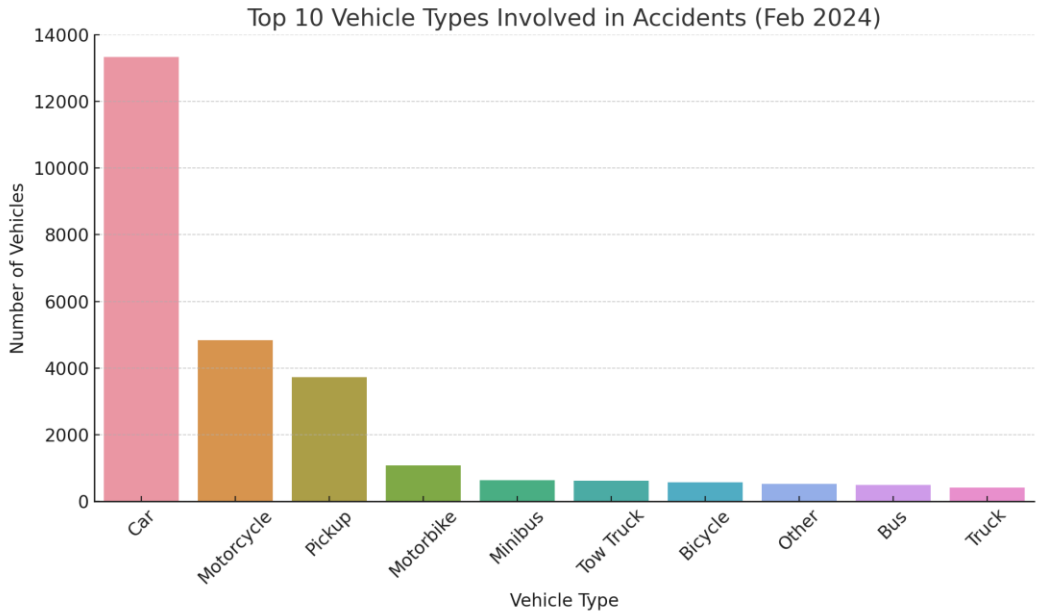

Figure: Top 10 Vehicle Types Involved in Accidents (Feb 2024)

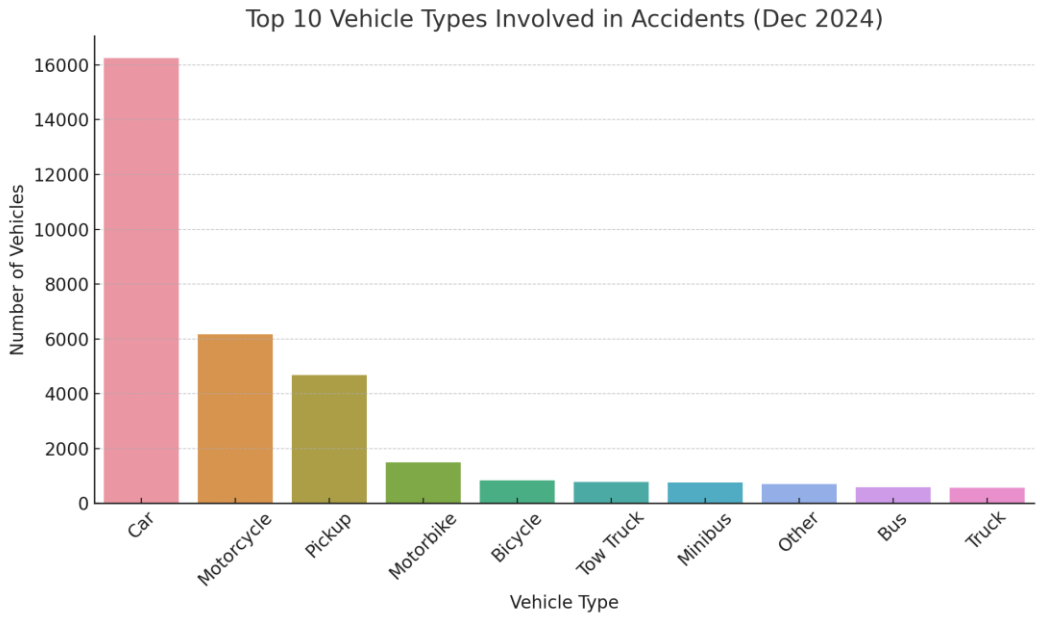

Figure: Top 10 Vehicle Types Involved in Accidents (Dec 2024)

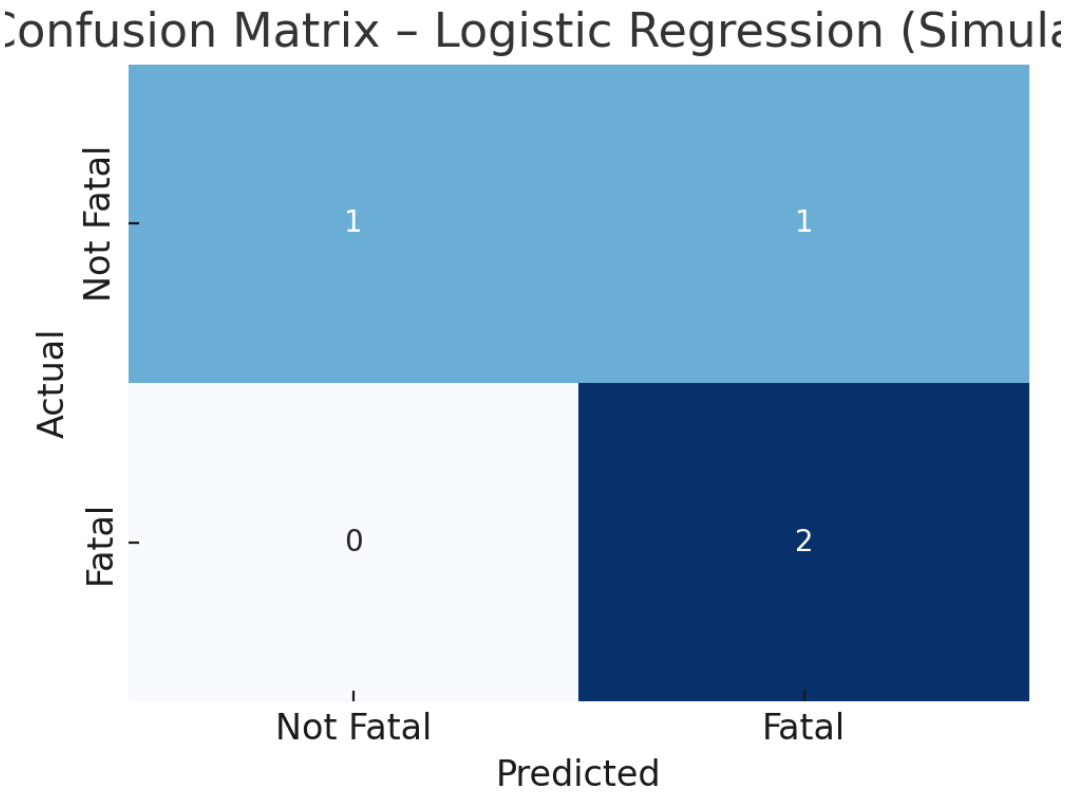

Figure: Confusion Matrix – Logistic Regression (Simulated)

Key Takeaways from Visual Data Analysis

Based on the graphical data, most accidents occur during daylight and in urban areas, with a noticeable rise during the summer months. Passenger cars are the leading vehicle type involved in accidents, while driver-related errors consistently emerge as the dominant cause across all periods. Provinces with higher populations such as Istanbul and Ankara report significantly more traffic incidents. Furthermore, the machine learning confusion matrix reflects a high classification accuracy, though it may be influenced by class imbalance.

Appendix: Machine Learning Performance Metrics

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	1.00	1.00	1.00	1.00
Random Forest	1.00	1.00	1.00	1.00
KNN	1.00	1.00	1.00	1.00
SVM	1.00	1.00	1.00	1.00