Predicting Pedestrian Fatalities at Traffic Collisions in San Francisco

As a PSA for Traffic Safety

Clarissa Lingas | May 2025

TL ; DR:

TL ; DR:

(Too Long; Didn't Read)

- **Problem:** The public needs an engaging way to become more aware of traffic safety

- **Solution:** Provide the data people need to make better traffic safety decisions in an interactive, fun, widely accessible PSA

- To center awareness around our most vulnerable:
  Develop a ML model that predicts whether a fatal victim of a traffic collision on San Francisco streets is a pedestrian

TL ; DR:

(Too Long; Didn't Read)

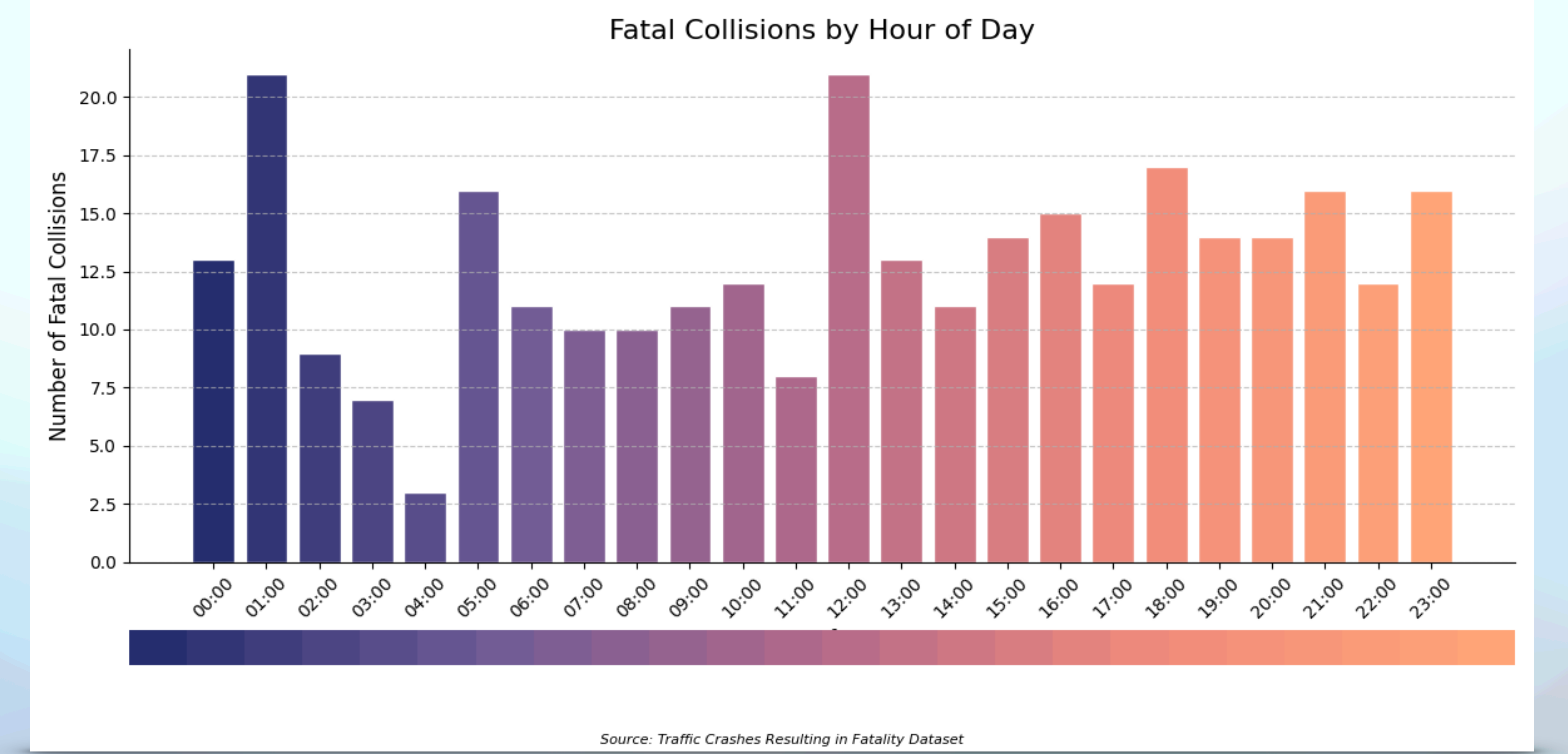
- Model: binary classification model using a Gradient Boosting Classifier

- Metrics:
  Accuracy: 95.52% on the test set
  Precision: 100% (all predicted pedestrian fatalities were correct)
  Recall: 92.11% (the model identified 92.11% of all actual pedestrian fatalities)
  F1 Score: 95.89% (harmonic mean of precision and recall)
  AUC: 99.09% (excellent discriminative ability)
  Cross-Validation Accuracy: 97.59% ± 1.53% (consistent performance across different data splits)

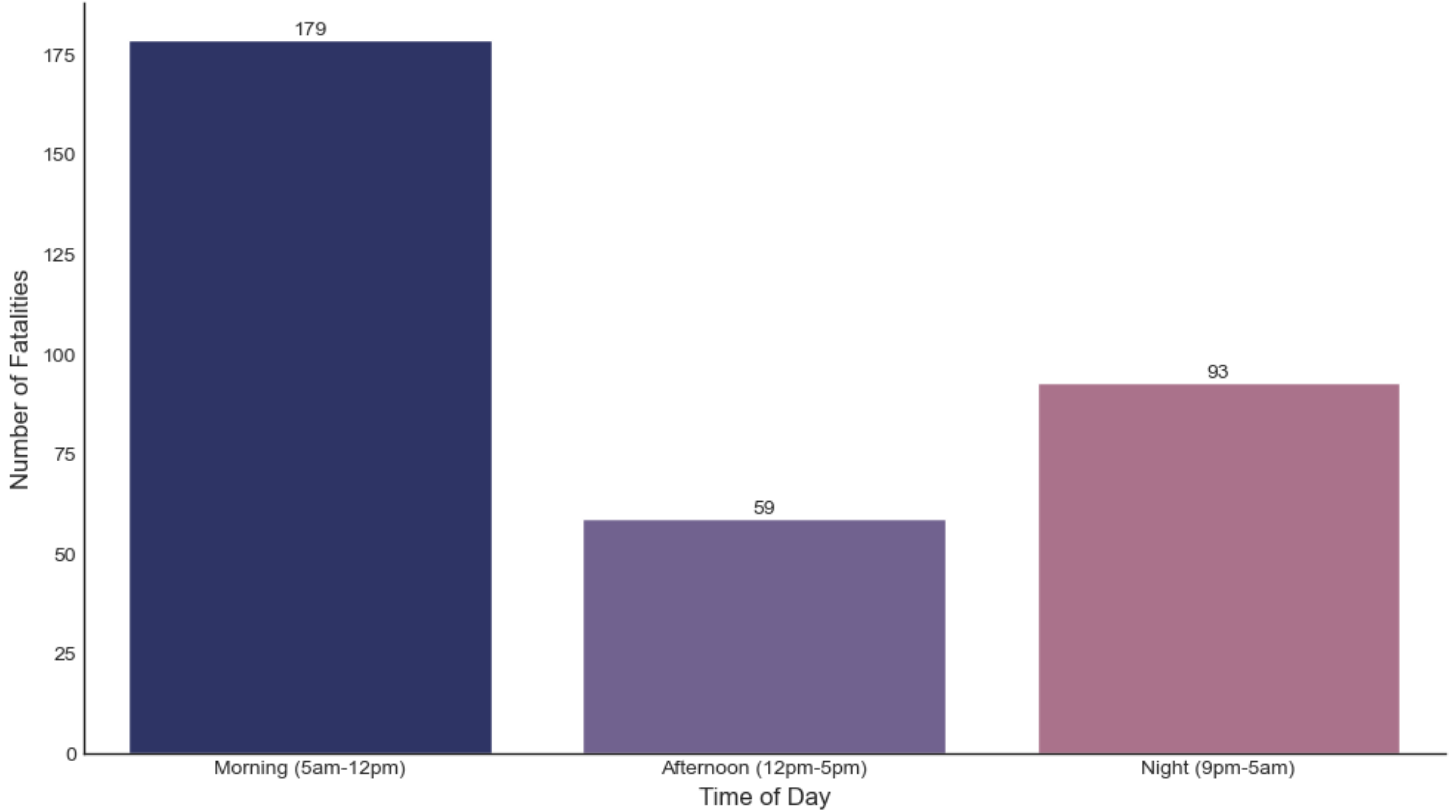
Findings from Data

Findings Summary

From Data Analysis



Source: Traffic Crashes Resulting in Fatality Dataset

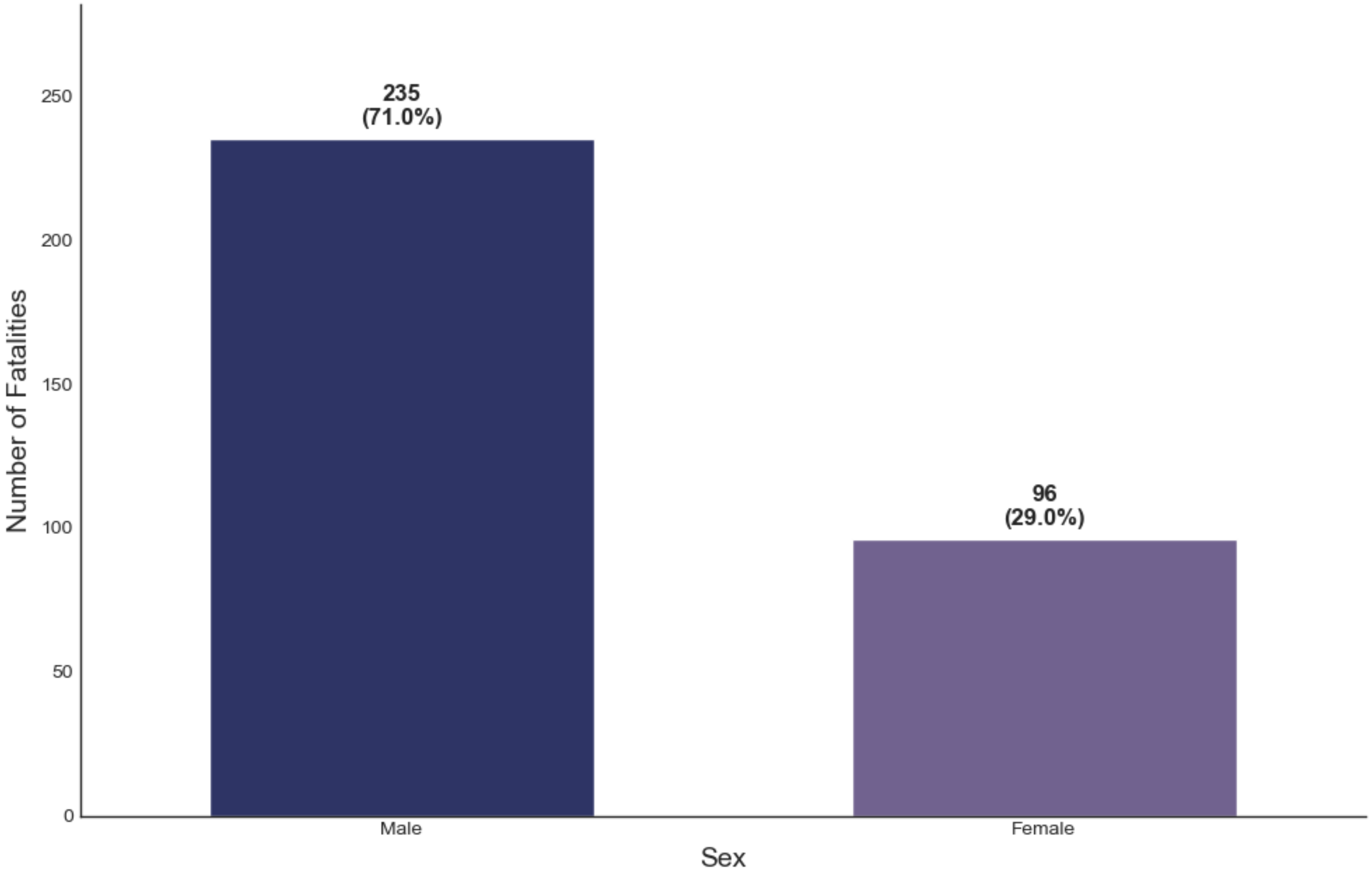


Fatalities by Time of Day, 2014-2025

Source: cleaned_fatalities.csv

Male victims account for 2.4x more traffic fatalities than female

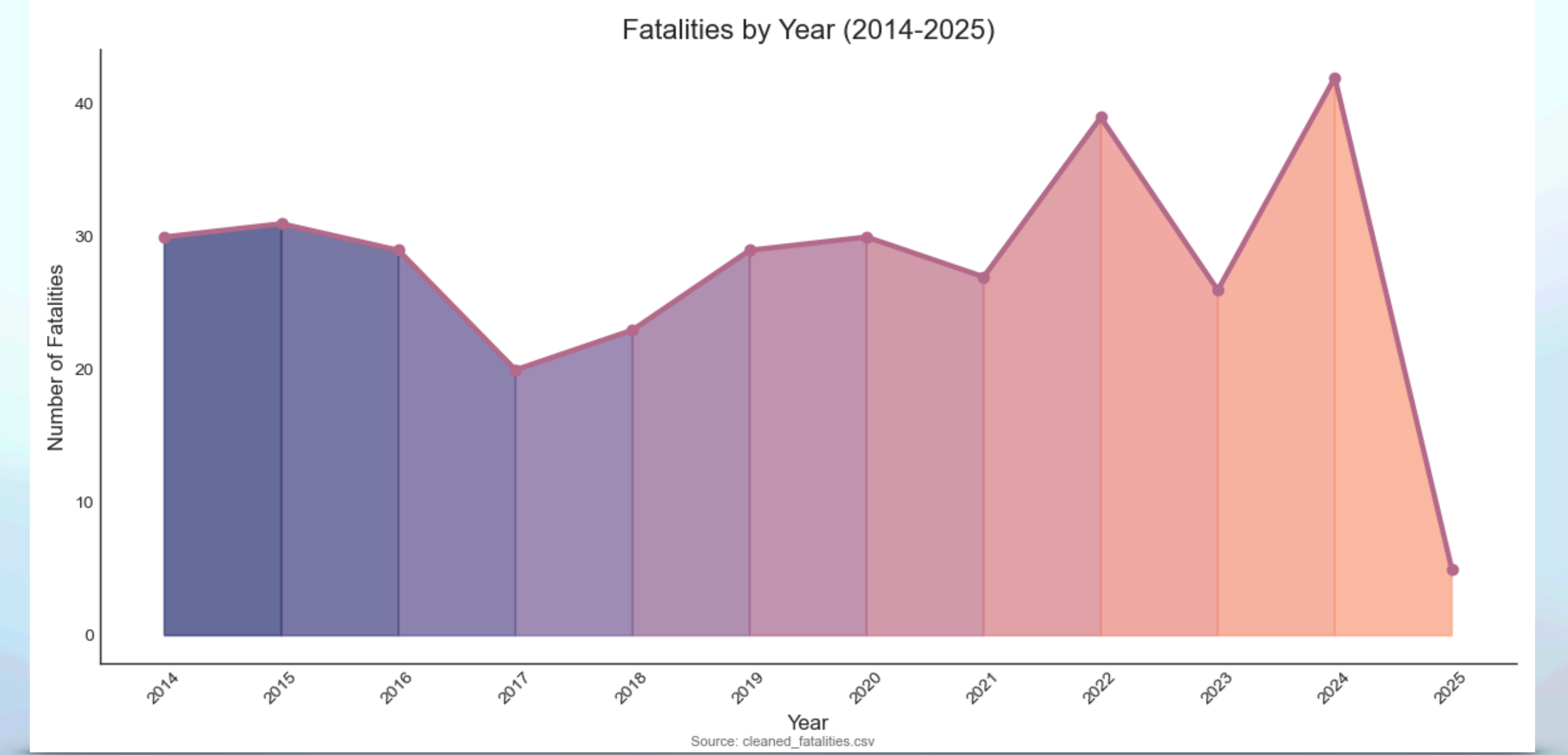
Distribution of Fatalities by Sex



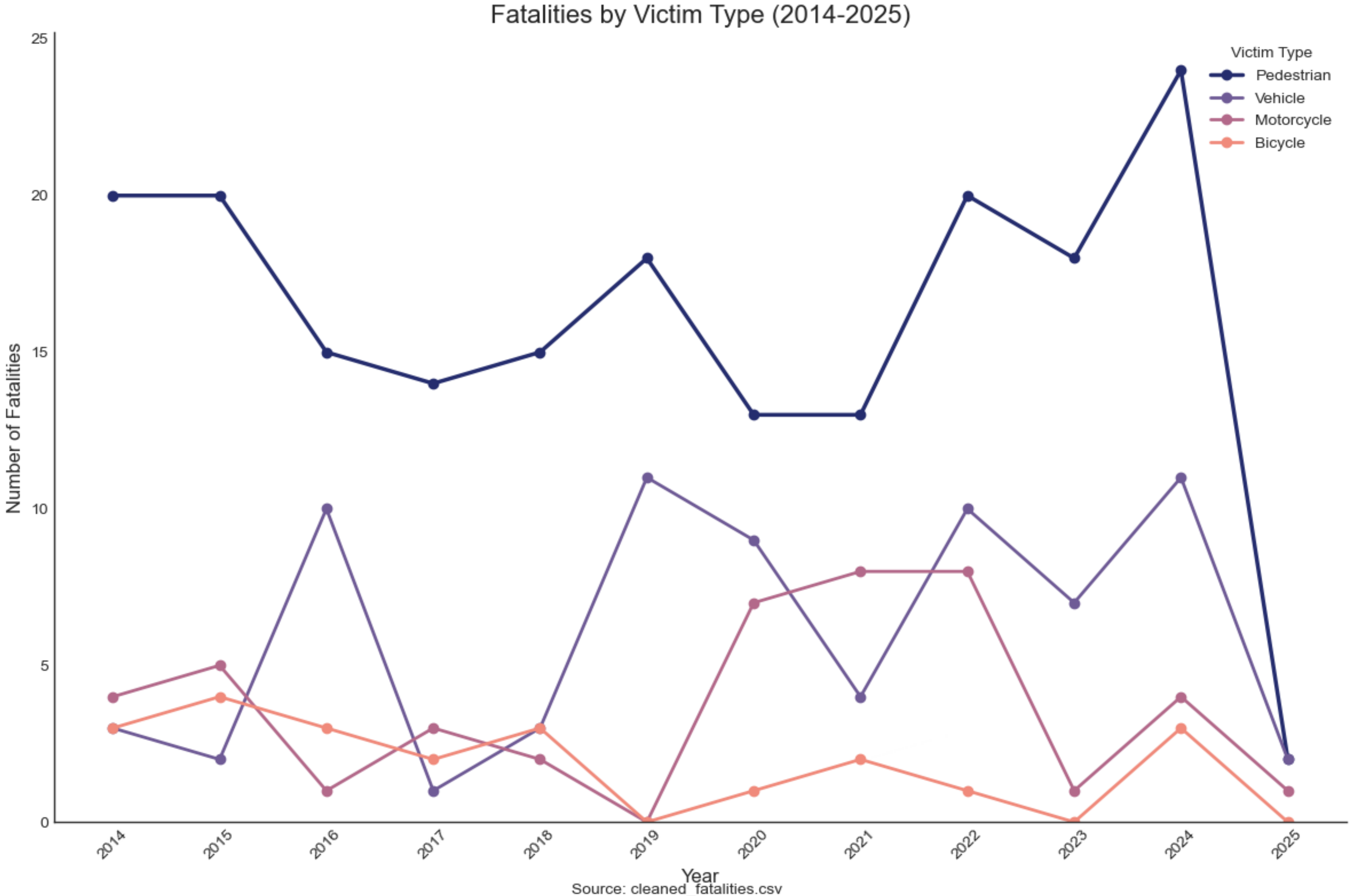
Source: cleaned_fatalities.csv

Findings Summary

From Data Analysis



Source: cleaned_fatalities.csv

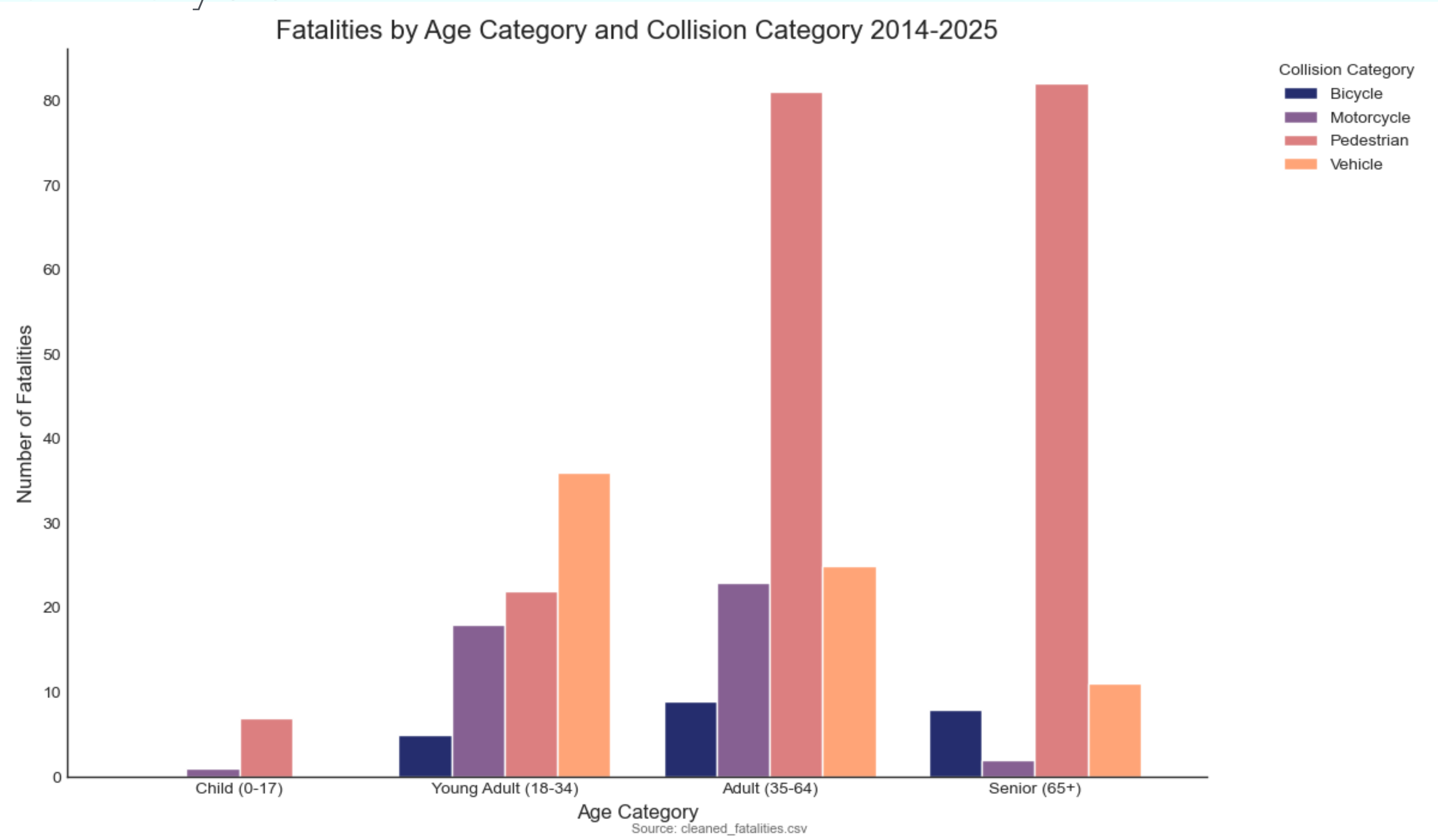


Fatalities by Victim Type (2014-2025)

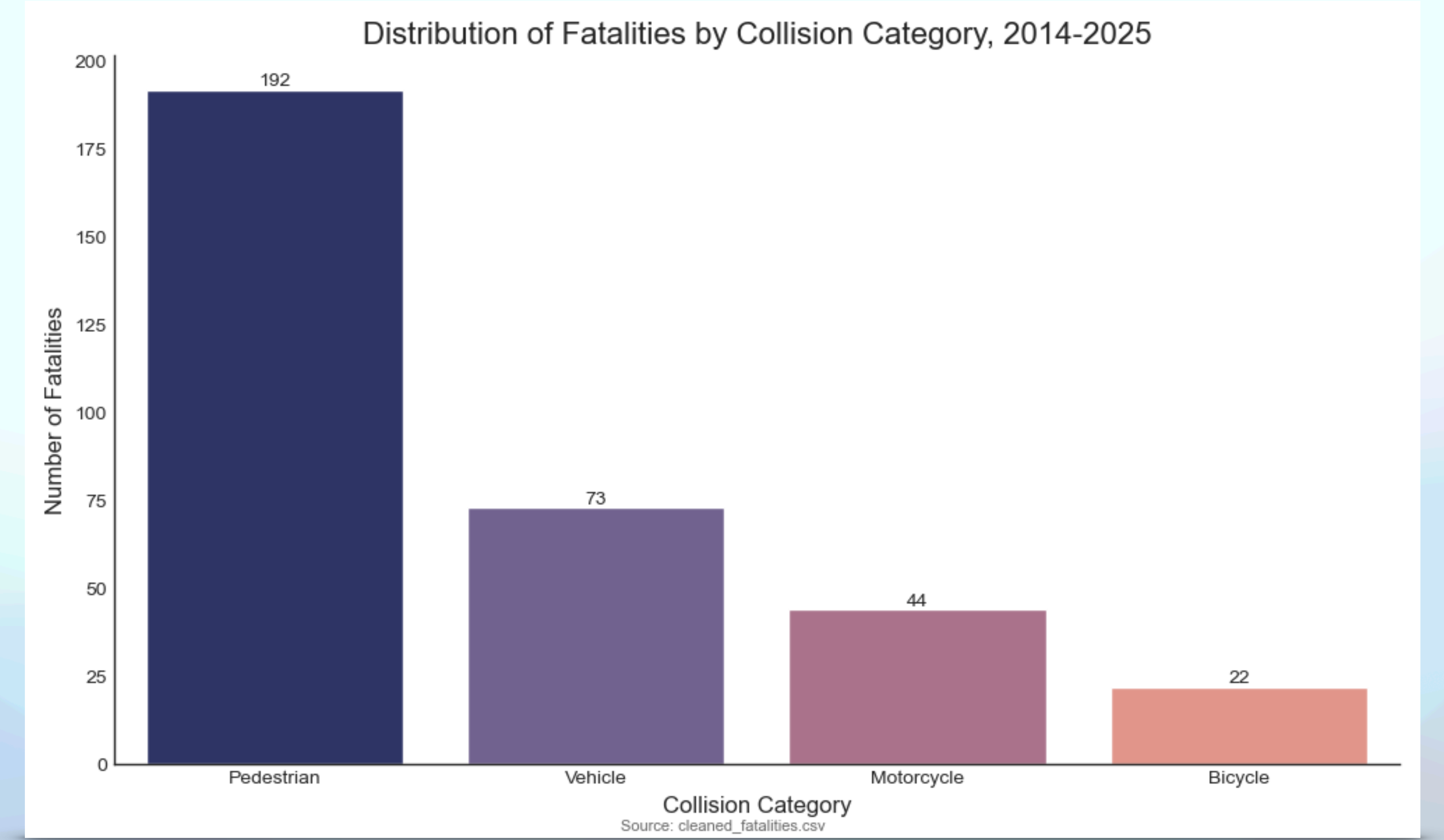
Source: cleaned_fatalities.csv

Findings Summary

From Data Analysis



Source: cleaned_fatalities.csv



Distribution of Fatalities by Collision Category, 2014-2025

Source: cleaned_fatalities.csv

Binary Classification Model

Binary Classification Model

Using a Gradient Boosting Classifier

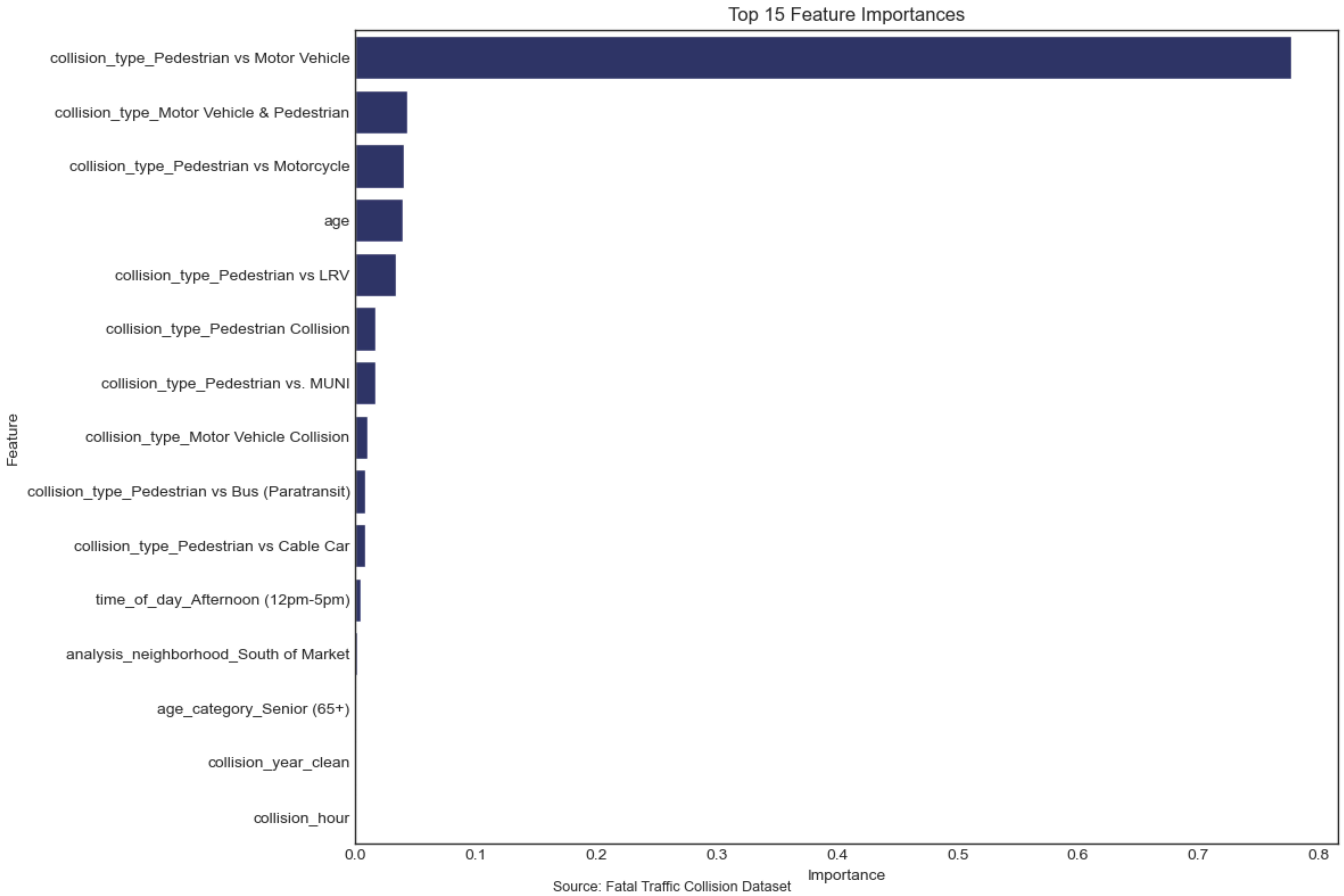
- **Purpose:** To predict whether a fatal victim of a traffic collision in San Francisco is a pedestrian, our most common fatality type in this city

- Part of an engaging interactive PSA to help SFers become more aware of traffic safety

Binary Classification Model

Using a Gradient Boosting Classifier

- **Features:**
  - Collision type: "Pedestrian vs Motor Vehicle" was by far the most important feature (77.7% importance)
  - Collision type: "Motor Vehicle & Pedestrian" (4.3% importance)
  - Collision type: "Pedestrian vs Motorcycle" (4.1% importance)
  - Age of the victim (4.0% importance)
  - Collision type: "Pedestrian vs LRV" (3.4% importance)



Source: Fatal Traffic Collision Dataset

Binary Classification Model

Using a Gradient Boosting Classifier

The model uses a combination of:

- **Spatial features:** neighborhood
- **Temporal features:** time of day
- **Categorical features:** age category, collision type, sex (M or F)

Binary Classification Model

Using a Gradient Boosting Classifier

For Preprocessing:

- **StandardScaler for numerical features:** Normalizes numerical data to improve model performance
- **OneHotEncoder for categorical features:** Transforms categorical variables into a format suitable for ML algorithms
- **ColumnTransformer:** Combines these preprocessing steps into a unified pipeline

Binary Classification Model

Using a Gradient Boosting Classifier

Model Training and Evaluation

- **Train-test split:** 80/20 split with stratification to maintain class balance

- **Pipeline architecture:** Ensures preprocessing and model training are consistently applied

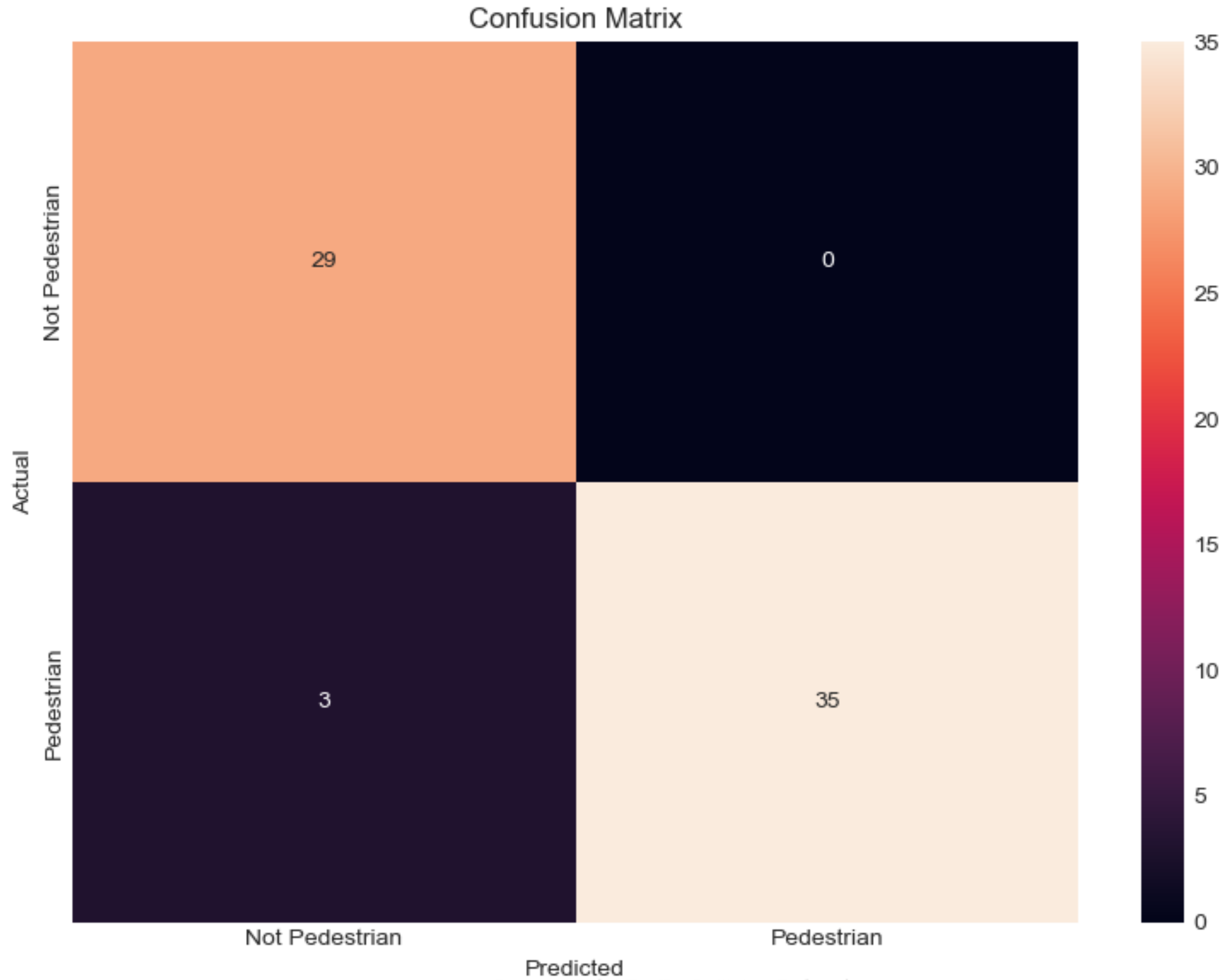
- **Classification metrics:** Uses classification report (precision, recall, F1-score) for model evaluation

- **ROC-AUC score:** Evaluates the model's ability to discriminate between classes

Binary Classification Model

Using a Gradient Boosting Classifier

- **Scores:**
  Accuracy: 95.52% on the test set
  Precision: 100% (all predicted pedestrian fatalities were correct)
  Recall: 92.11% (the model identified 92.11% of all actual pedestrian fatalities)
  F1 Score: 95.89% (harmonic mean of precision and recall)
  AUC: 99.09% (excellent discriminative ability)
  Cross-Validation Accuracy: 97.59% ± 1.53% (consistent performance across different data splits)



Source: Fatal Traffic Collision Dataset