Monroe County Car Crash Dataset Analysis

Ayush Das and Blake Gibson

School of Information, The University of Texas at Austin

I 320M: Data Science for Biomedical Health Informatics

Instructor: Dr. Steven Hershman

Date: 04/27/2025

Abstract

Motor vehicle accidents remain a leading cause of injury and mortality, with localized studies offering critical insights for targeted prevention strategies. This study analyzes crash data from Monroe County, Indiana, spanning 2003 to 2015, to identify environmental and behavioral factors most strongly associated with accident risk. Utilizing a combination of descriptive statistics and machine learning, specifically a Random Forest Classifier, patterns across nearly 54,000 incidents were examined. Key variables include collision type, injury severity, temporal factors, location data, and primary crash causes such as failure to yield and speeding. The analysis revealed heightened accident risk during late hours and under specific behavioral conditions. The Random Forest model achieved approximately 80% accuracy in predicting injury and collision outcomes, highlighting its potential for predictive safety interventions. Findings aim to inform local policy, enhance driver education, and support the development of data-driven public safety initiatives, ultimately contributing to a reduction in vehicular accidents and injuries.

Introduction

Motor vehicle accidents are a pervasive public health concern, representing one of the leading causes of injury, death, and economic burden globally. In the United States alone, the National Highway Traffic Safety Administration (NHTSA) reports that millions of accidents occur annually, with significant human and financial costs. Despite ongoing efforts to improve road safety through education, infrastructure enhancements, and enforcement, accidents persist at high rates. Understanding the contributing factors behind traffic accidents at the local

level is crucial for developing targeted interventions that can more effectively mitigate risk and prevent loss of life.

Localized analyses, such as those focused on individual counties or regions, allow researchers and policymakers to identify unique patterns related to environment, behavior, and policy enforcement that may not be apparent in national datasets. Monroe County, Indiana, presents a valuable case study due to its diverse road environments, seasonal weather variations, and a mixture of urban and rural traffic patterns. Investigating crash data in this specific context offers opportunities to uncover risk factors that may be addressable through local action.

Advances in data science, particularly machine learning, provide powerful tools for analyzing large and complex datasets such as traffic crash reports. Machine learning models, such as Random Forest classifiers, enable the identification of subtle patterns and relationships between variables like time of day, collision type, injury severity, and primary crash causes. These insights can guide evidence-based policy making, enhance driver education programs, and inform infrastructure improvements.

However, this field is not without challenges. Issues such as missing or inconsistent data, potential biases in data collection, and limitations in variable granularity can complicate analyses. Moreover, while machine learning models offer strong predictive capabilities, they often operate as "black boxes," providing less interpretability than traditional statistical methods. Care must be taken to validate model results and contextualize findings within the broader social and environmental landscape.

In this paper, we will present the background, methodology, and findings from an analysis of car crash data collected in Monroe County between 2003 and 2015. Our primary goal is to answer the research question, "What are the most dangerous circumstances that significantly increase the likelihood of an accident in Monroe County?" Through data preprocessing, descriptive statistics, and the application of machine learning models, we aim to contribute actionable insights that can help reduce accident rates and improve public safety at the local level.

Dataset Introduction and Variables Overview

Effective analysis of traffic accidents requires access to comprehensive, high-quality data that captures a wide range of incident characteristics. The dataset analyzed in this study was sourced from the Automated Report and Information Exchange System (ARIES), maintained by the Indiana State Police. It includes detailed records of vehicular accidents occurring in Monroe County, Indiana, between 2003 and 2015. Covering a twelve-year span, the dataset provides a robust foundation for identifying long-term trends, seasonal variations, and persistent risk factors associated with motor vehicle accidents. The breadth of the dataset, comprising nearly 54,000 individual incidents, affords the statistical power necessary to detect meaningful patterns that can guide targeted public safety interventions.

The scope of the dataset encompasses a diverse set of variables that reflect both environmental conditions and human behaviors implicated in crash events. These variables are broadly categorized into three domains: temporal details, incident characteristics, and location information. Temporal variables include the year, month, day of the week, and time of day of

each crash, as well as an indicator of whether the incident occurred during a weekend. These factors allow for an examination of how time-based patterns, such as weekday commuting or late-night driving behaviors, correlate with accident frequency and severity.

Incident-specific variables describe the nature of the collision and its resulting impact. They include the type of collision (such as single-vehicle, two-vehicle, multi-vehicle, or pedestrian-related accidents), the injury severity (ranging from no injury to fatal injuries), and the primary factor contributing to the crash (including behaviors like speeding, failure to yield, and following too closely). These measures are vital for identifying specific behaviors and circumstances that pose the greatest risk to public safety.

Spatial variables provide additional depth by recording the reported location and GPS coordinates of each crash. Geographic analysis of these data points allows for the identification of high-risk zones within Monroe County, highlighting areas that may benefit from infrastructural improvements, policy changes, or increased enforcement efforts.

Together, the structure and breadth of the dataset enable a multidimensional exploration of traffic accident dynamics. The integration of temporal, incident-specific, and spatial variables forms the foundation for the descriptive and predictive analyses conducted in this study, aimed at uncovering actionable insights into the conditions most strongly associated with vehicular crashes in Monroe County.

Methods

The analysis began with a thorough preprocessing stage to ensure the quality and consistency of the dataset. The original dataset contained approximately 54,000 records of vehicular crashes in Monroe County, Indiana, collected between 2003 and 2015. During initial cleaning, 64 duplicate entries were identified and removed. Furthermore, over 1,300 records were found to have missing critical values and were subsequently excluded from the dataset. After these steps, 52,520 unique and complete crash reports remained for analysis.

```
df.duplicated().sum()
```

```
64
```

```
df = df.drop_duplicates()
```

```
df.isnull().sum()
```

```
Year                   0
Month                  0
Day                    0
Weekend?              68
Hour                 225
Collision Type         6
Injury Type            0
Primary Factor      1119
Reported_Location     35
Latitude              30
Longitude             30
dtype: int64
```

```
df.dropna(inplace=True)
df.shape
```

```
(52520, 11)
```

Figure 1. Jupyter Notebook analysis of missing and duplicate data points from dataset *df* and subsequent preprocessing.

Exploratory data analysis was conducted to uncover patterns relating to the temporal distribution of accidents, the types of collisions, the severity of injuries, and the primary causes of crashes. Trends were examined across various dimensions, including time of day, day of the week, month, and year. In addition, accident characteristics such as the number of vehicles involved, the nature of injuries, and the reported primary causes were analyzed to understand dominant patterns and potential risk factors.

Subsequently, a machine learning model was implemented using a Random Forest Classifier. Random Forests are ensemble learning methods that aggregate the predictions of multiple decision trees to improve predictive accuracy and control overfitting (Breiman, 2001). In Scikit-learn, individual decision trees are binary by construction, splitting the dataset at each node based on feature values. Random Forests address this by introducing two layers of randomness: bagging, where each tree is trained on a different random sample of the data with replacement, and feature randomness, where only a random subset of features is considered at each split. This combination reduces overfitting and improves model generalizability.

Two separate Random Forest models were trained. The first used "Injury Type" as the target variable, while the second used "Collision Type." Each model was constructed with 100 decision trees, as performance improvements beyond 100 trees were observed to be marginal while computational costs scaled linearly. Model performance was evaluated using accuracy, precision, and recall metrics, with special attention paid to the influence of class imbalances on the outcomes.

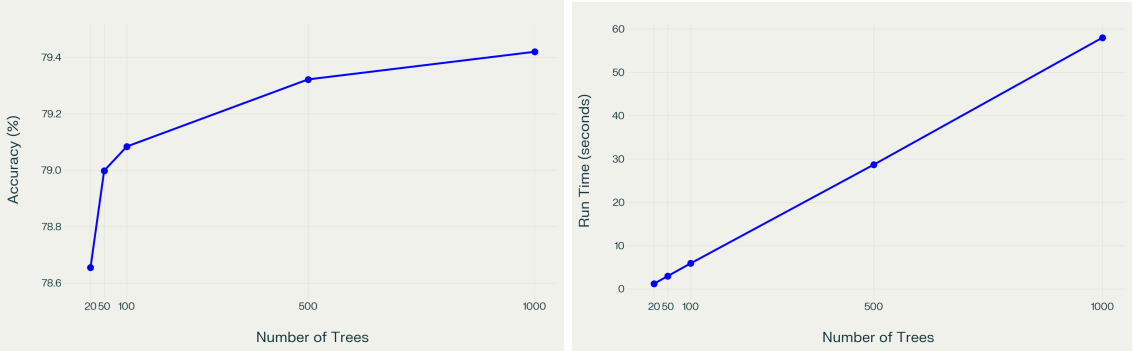


Figure 2. Side-by-side comparison of the accuracy of the model and the run-time showing the computational costs are not worth it past 100 trees.

Analysis

The exploratory data analysis revealed several notable findings. Analysis of temporal trends indicated that the highest number of accidents occurred at approximately 5:00 PM, suggesting a strong association with afternoon commute traffic. Fridays exhibited the highest accident rates, followed by Thursdays, while October recorded the greatest number of accidents across the calendar year. Year-to-year variation was relatively minimal; however, a slight peak was observed in 2008.

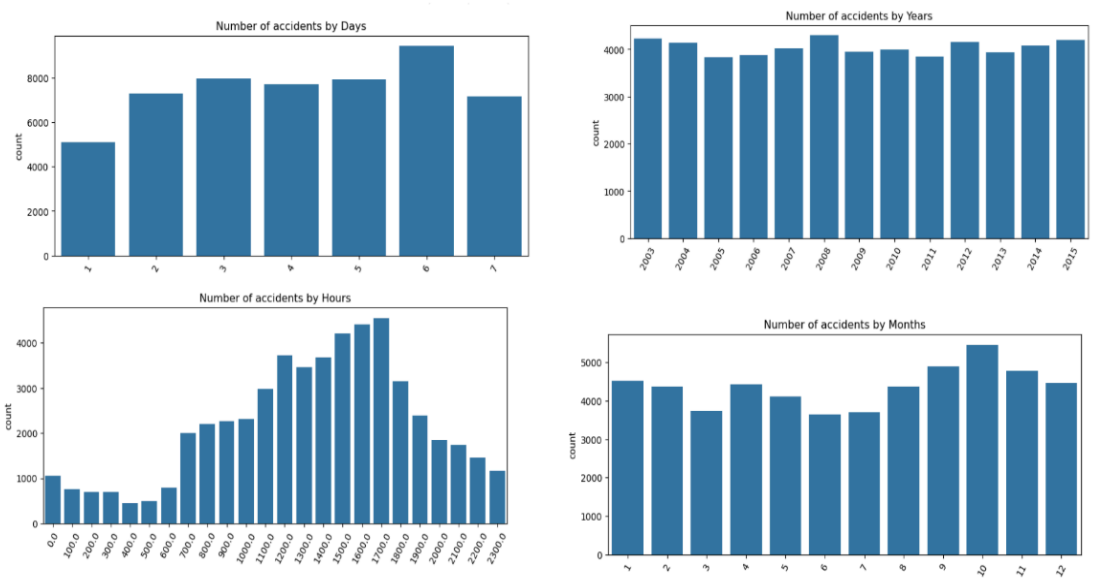


Figure 3. Plots displaying the number of accidents by [timeframe].

When considering the distribution of crashes across the week, it was found that 24.51% of accidents occurred during the weekend, which is lower than the expected 28.5% based on the proportion of weekend days, potentially indicating reduced vehicular activity on weekends.

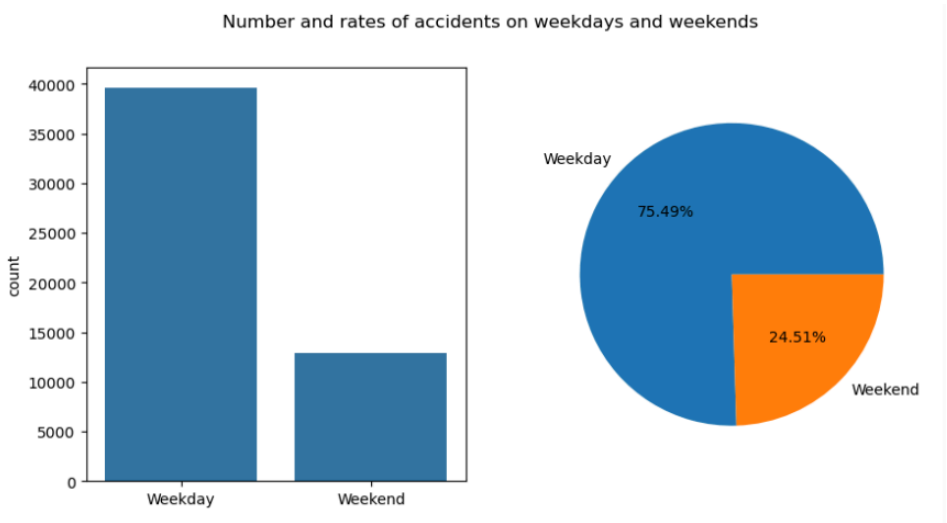


Figure 4. Bar graph and pi chart showing rate of accidents on weekdays vs weekends.

An examination of collision types showed that two-vehicle accidents constituted the vast majority of incidents, followed by single-vehicle crashes and then accidents involving three or more vehicles. Other types of collisions, such as pedestrian or bicycle-related accidents, collectively accounted for only around 6% of the total dataset.

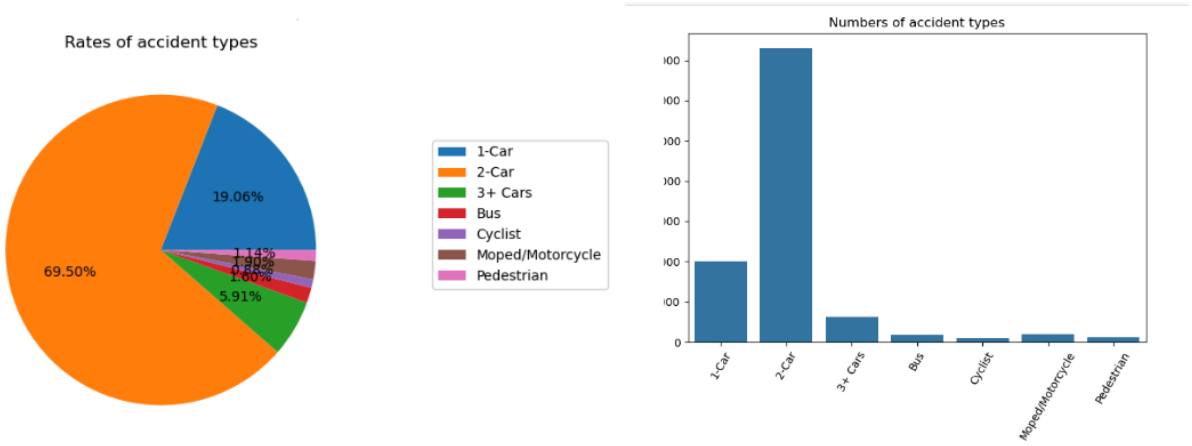


Figure 5. Dataset analysis of accident types, with pi-chart showing percentage of the accident type and bar graph showing dataset counts of the accident type.

Regarding injury severity, the majority of recorded outcomes (76%) indicated no injury or an unknown injury status, while non-incapacitating injuries represented approximately 21% of cases. Serious injuries and fatalities were comparatively rare, suggesting that although accidents were frequent, severe injuries were not the norm.

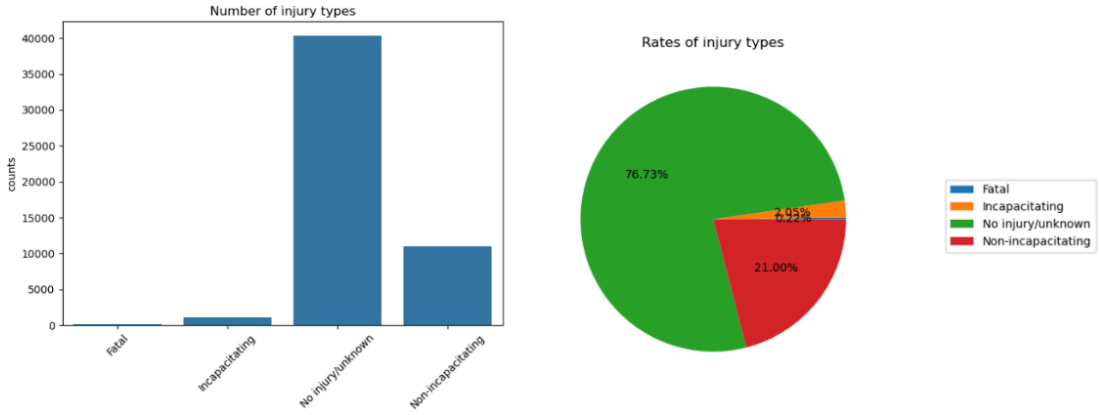


Figure 6. Dataset analysis of accident types, with pi-chart showing percentage of the injury type and bar graph showing dataset counts of the injury type.

Primary crash causes were also investigated. "Failure to Yield Right of Way" emerged as the most common cause, with over 10,000 incidents, followed by "Following Too Closely" and "Other (Driver) - Explain in Narrative." Notably, "Ran Off Road Right" was a frequent occurrence, while there was no equivalent "Ran Off Road Left" category recorded, possibly reflecting the influence of right-side driving norms and visual dominance among drivers.

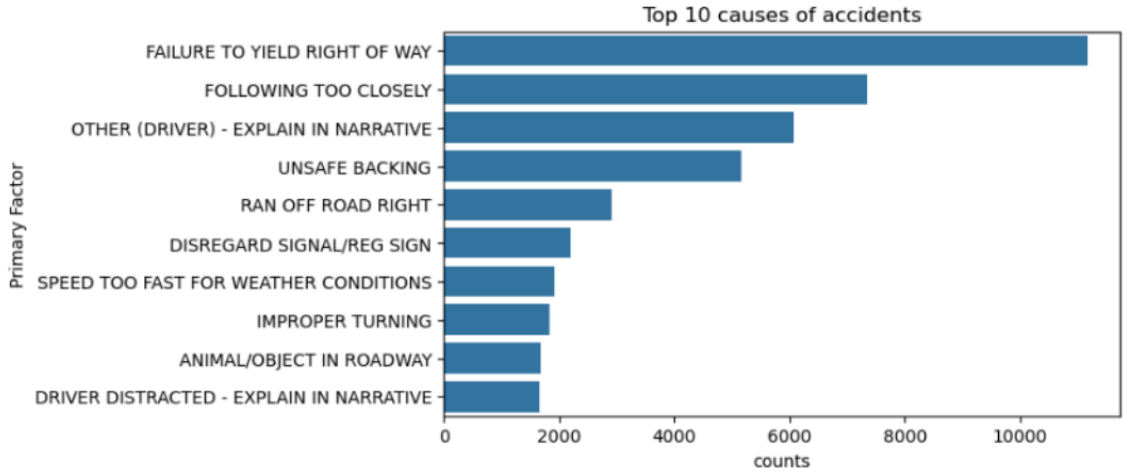


Figure 7. Plot depicting the counts of the accident causes in Monroe County.

Single-vehicle accidents were most often associated with vehicles running off the road to the right, while two-vehicle accidents were typically due to failures to yield. Accidents involving three or more vehicles commonly stemmed from following too closely, and bus-related accidents often required detailed narrative explanations under the "Other" category, reflecting additional context to understand the collision.

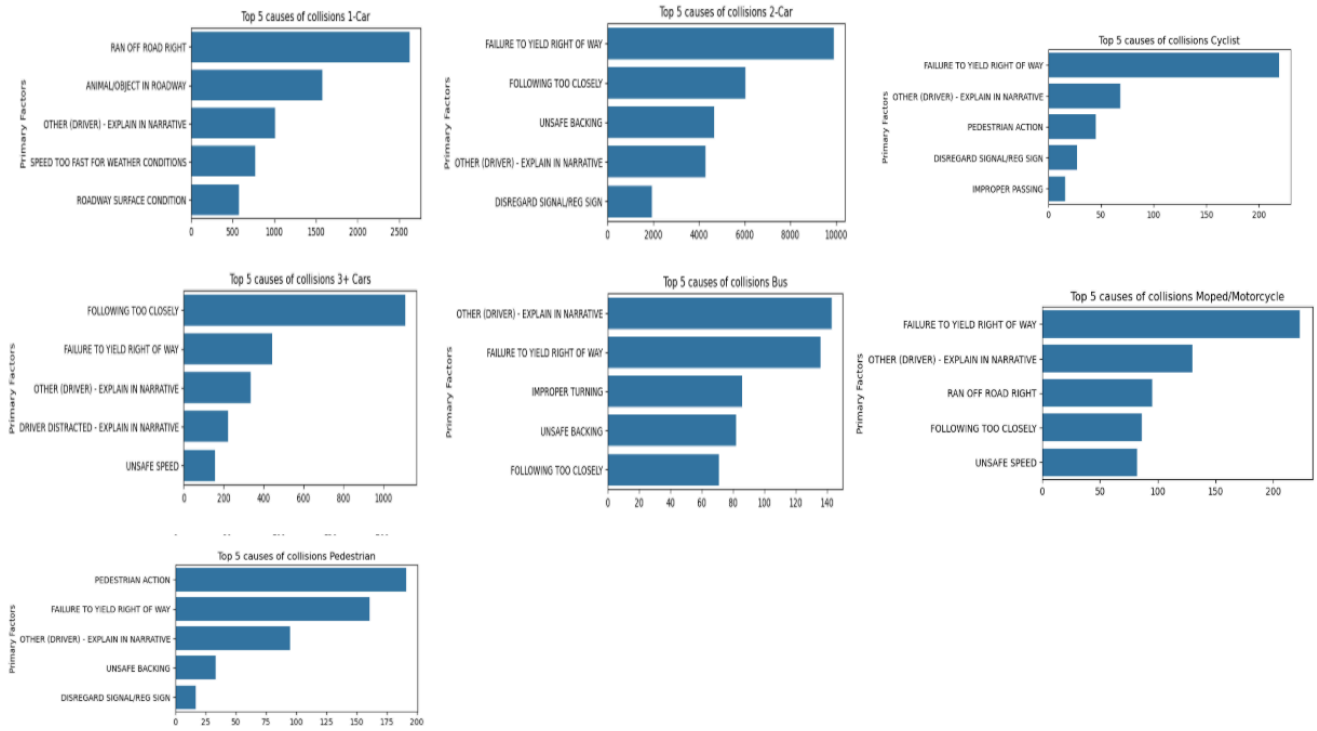


Figure 8. Seven plots depicting the five most common reasons for the collision types.

An additional trend emerged when analyzing collision times sorted by the type of accident. While two-vehicle and multi-vehicle collisions peaked sharply around 5:00 PM, corresponding with the afternoon traffic rush, single-vehicle accidents were distributed more evenly throughout the day. This broader spread of single-vehicle crashes may be attributed to factors less dependent on traffic density, such as distracted driving or loss of vehicle control,

which can occur at any time rather than being tied to commuting patterns.

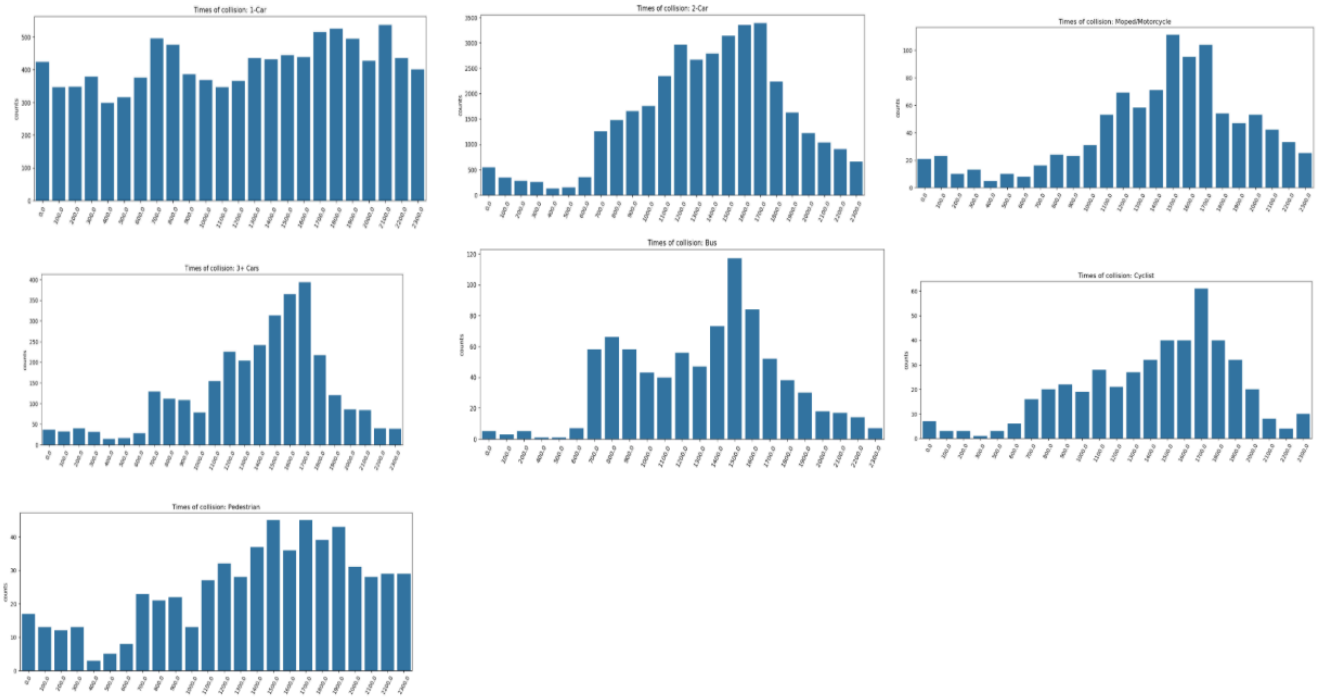


Figure 9. This plot depicts the collision times for each of the collision types.

Spatial analysis of collision types across Monroe County revealed that most locations exhibited a dominance of two-vehicle accidents. However, specific intersections, such as SR37 and Vernal Pike, demonstrated a notably high proportion (15.31%) of accidents involving three or more vehicles, indicating heavy congestion. Furthermore, locations like N Walnut Street and E 10th Street showed elevated proportions of single-vehicle accidents (17–18%), potentially suggesting issues related to road design or surface conditions.

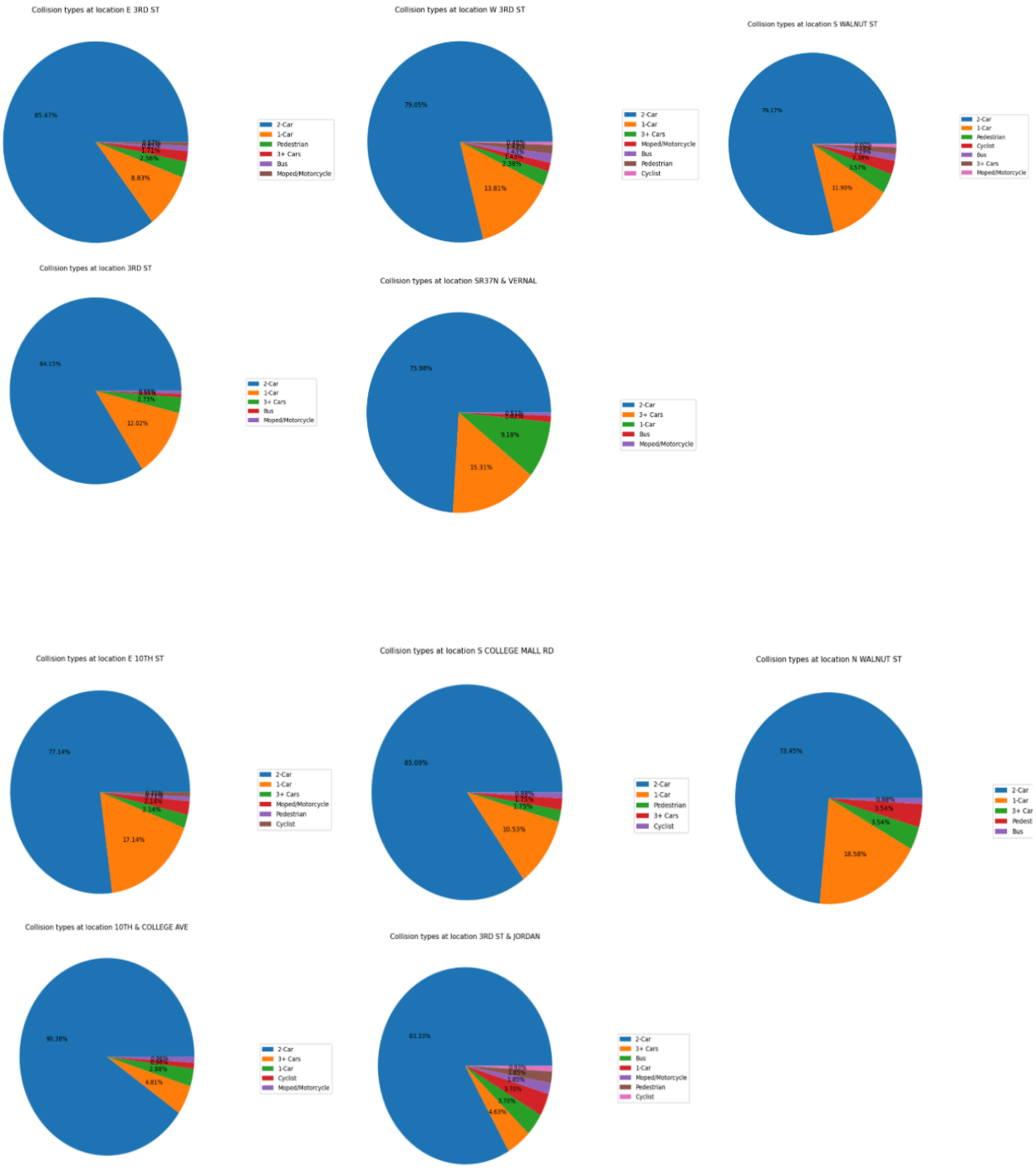
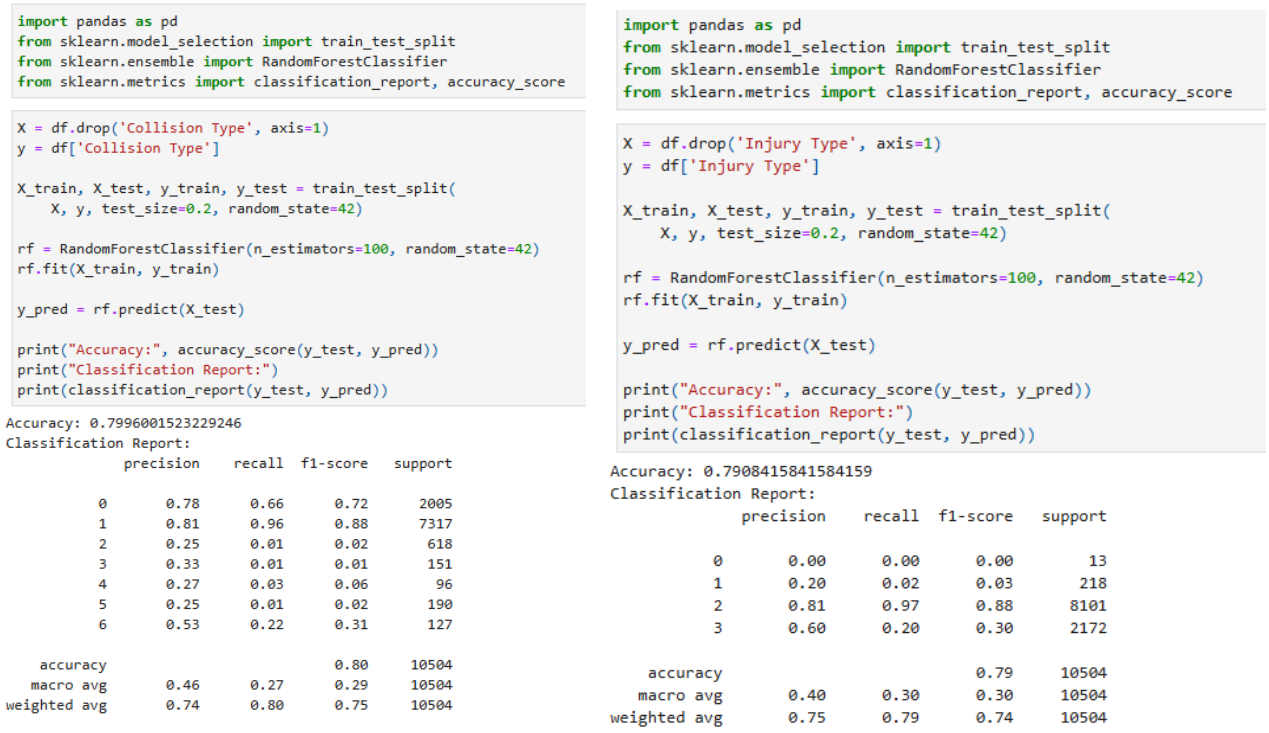


Figure 10. The plots showing the collision types at 10 different locations around Monroe County.

Following the exploratory analysis, the Random Forest Classifier models were applied. When using "Injury Type" as the target, the model achieved an overall accuracy of approximately 79%. However, performance was heavily influenced by the dominance of the "No Injury/Unknown" class, which resulted in a precision score of 0.81 and a recall score of 0.97 for that category, while performance on other injury types was considerably lower. Similarly, when "Collision Type" was used as the target, the model achieved an accuracy of approximately 79.9%. In this case, high precision (0.81) and recall (0.96) were observed for two-vehicle accidents, and moderately high precision (0.78) and lower recall (0.66) for single-vehicle accidents. The models struggled to accurately predict minority classes, reflecting the underlying class imbalance.



```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, accuracy_score

X = df.drop('Collision Type', axis=1)
y = df['Collision Type']

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42)

rf = RandomForestClassifier(n_estimators=100, random_state=42)
rf.fit(X_train, y_train)

y_pred = rf.predict(X_test)

print("Accuracy:", accuracy_score(y_test, y_pred))
print("Classification Report:")
print(classification_report(y_test, y_pred))
```

```
Accuracy: 0.7996001523229246
Classification Report:
              precision    recall  f1-score   support

           0       0.78      0.66      0.72      2005
           1       0.81      0.96      0.88      7317
           2       0.25      0.01      0.02       618
           3       0.33      0.01      0.01       151
           4       0.27      0.03      0.06        96
           5       0.25      0.01      0.02       190
           6       0.53      0.22      0.31       127

    accuracy                           0.80     10504
   macro avg       0.46      0.27      0.29     10504
weighted avg       0.74      0.80      0.75     10504
```

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, accuracy_score

X = df.drop('Injury Type', axis=1)
y = df['Injury Type']

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42)

rf = RandomForestClassifier(n_estimators=100, random_state=42)
rf.fit(X_train, y_train)

y_pred = rf.predict(X_test)

print("Accuracy:", accuracy_score(y_test, y_pred))
print("Classification Report:")
print(classification_report(y_test, y_pred))
```

```
Accuracy: 0.7908415841584159
Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        13
           1       0.20      0.02      0.03       218
           2       0.81      0.97      0.88      8101
           3       0.60      0.20      0.30      2172

    accuracy                           0.79     10504
   macro avg       0.40      0.30      0.30     10504
weighted avg       0.75      0.79      0.74     10504
```

Figure 11. The model results on the two “target” variables, Collision Type and Injury Type.

Conclusion

This study offers a comprehensive analysis of vehicular accidents in Monroe County, Indiana, over a twelve-year period. The findings highlight that the afternoon commute, particularly around 5:00 PM and on Fridays, represents a period of heightened accident risk. Behavioral factors such as failure to yield right of way and following too closely were the leading contributors to accidents, suggesting that targeted educational and enforcement efforts could substantially improve public safety outcomes.

Spatial analyses identified specific intersections and corridors with elevated accident rates, providing actionable insights for potential infrastructure improvements. Although the Random Forest models demonstrated reasonably high overall accuracy, their performance was limited by the imbalanced nature of the dataset. The models were adept at predicting the majority classes but performed poorly on less frequent but critical classes such as severe injuries and multi-vehicle collisions.

Future research should consider employing resampling techniques to balance the dataset or exploring alternative machine learning methods designed to better handle class imbalance. Incorporating additional contextual features, such as weather conditions or traffic density, may also improve predictive power. The insights derived from this study can help create data-driven strategies to reduce accident rates and enhance roadway safety in Monroe County.

References

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

https://doi.org/10.1023/A:1010933404324

R, J. D. (2024, January 8). Car crash dataset. Kaggle.
https://www.kaggle.com/datasets/jacksondivakarr/car-crash-dataset