

**ALY6140: Python & Analytics System Technology**

MODULE 5: Capstone Project Draft Report

GROUP MEMBERS:

Ansh Vipul Aya - aya.a@northeastern.ed,

Jie Liu - liu.jie5@northeastern.edu,

Mohammed Abdoul - abdoul.m@northeastern.edu,

Pengxiang Zhang - zhang.pengxi@northeastern.edu

GROUP NUMBER: 5

PROFESSOR NAME: Zhi He

DATE: 10/13/2023

* **Introduction:**
* In the pursuit of enhancing road safety and streamlining accident resolution processes, we embarked on an analytics project using a comprehensive dataset from Montgomery County, Maryland. Basically the dataset is comprised of information about traffic collisions on county and local roads, originates from reliable sources such as the Automated Crash Reporting System (ACRS) of the Maryland State Police, Montgomery County Police, Gaithersburg Police, Rockville Police, and the Maryland-National Capital Park Police. It includes 166,000 rows and 43 columns, this extensive dataset provides a valuable opportunity to gain insights into motor vehicle collisions, a global concern.
* The primary goal of this project is to address the pivotal question of determining fault in these collisions. Accurate fault classification holds the potential to significantly impact various stakeholders, including insurance companies for claims processing, law enforcement for assigning responsibility, and policymakers for road safety improvements. Moreover, it can empower drivers to better understand their driving behavior and identify areas for improvement to contribute to safer roadways. To achieve this, our aim is to create an efficient and reliable system that automates fault determination, ultimately contributing to road safety.
* Considering these objectives, we set out to build four classification models: Logistic Regression, Decision Tree Classification, Support Vector Machine, and Random Forest Classification. These models are tailored to classify driver’s at fault, ultimately providing insights into the responsible party in traffic collisions. In this report, we will present the results of our analysis and discuss the implications of our findings, including recommendations for stakeholders.

**Exploratory Data Analysis:**

After we determine the business objectives, we start working on data processing and exploratory data analysis. We extracted data by using the ‘pd.read\_csv()’ function, and got the shape of the dataframe. Given the total 166128 of rows and 43 columns, it's reasonable for us to assume insights into the factors contributing to those car crashes, such as time, equipment problems, weather conditions, etc.

So we have performed some basic data cleaning operations to get rid of unneeded columns, such as 'Latitude','Agency Name','Route Type', 'Road Name', 'Report Number', 'Longitude', 'Location', 'Person ID', 'Vehicle ID', 'Cross-Street Type', 'Cross-Street Name', 'Municipality', and 'Off-Road Description'. These categorical columns are obviously unrelated to the target variable we defined, so they were dropped.

The column ‘Crash Date/Time’ was trimmed and renamed to ‘Crash Time’, which can help us better understand how time contributes to the crashes and pave the way for us to create time intervals later on.

Handling missing data ensures the accuracy of our analysis. Missing data points were perfectly handled by us. We first counted the total missing values of each column and got the ratios. If any column had over 70% missing values, it was deemed not beneficial for analysis and was dropped. It ensures that we do not rely heavily on sparse data. For columns with under 70% missing values, for instance, ‘Equipment Problems’, we filled null values with the mode. We assumed the most common equipment problem is also the most probable for NaN records. As for some other categorical columns, using mode for replacing the null values is not the best option and it could introduce some bias. So we just dropped the missing values.

After data cleaning, we plotted frequency distributions for several categorical columns:

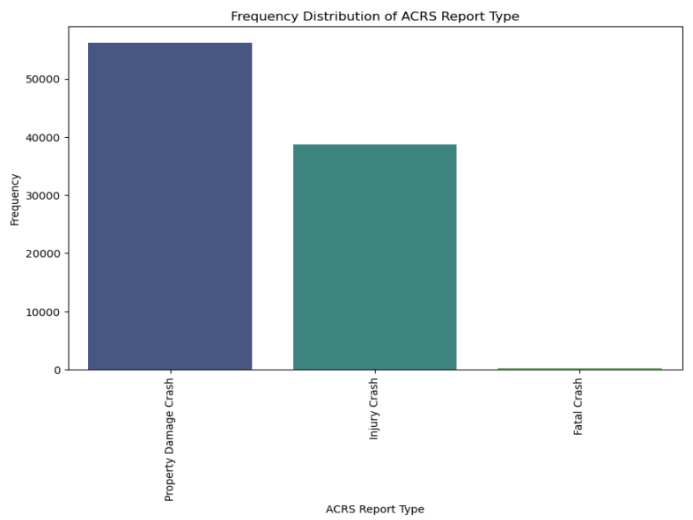


Figure 1. Frequency Distribution of ACRS Report Type

This bar chart represents the frequency distribution of the ACRS (Automated Crash Reporting System) report types. "Property Damage Crash" reports are the most frequent, reaching close to 50,000. It could imply that majority crashes lead to property damage, without any injury to individuals. "Injury Crash" reports are next in frequency, but are notably less than property damage, hovering just above 30,000. "Fatal Crash" reports are minimal in comparison to the other two categories.

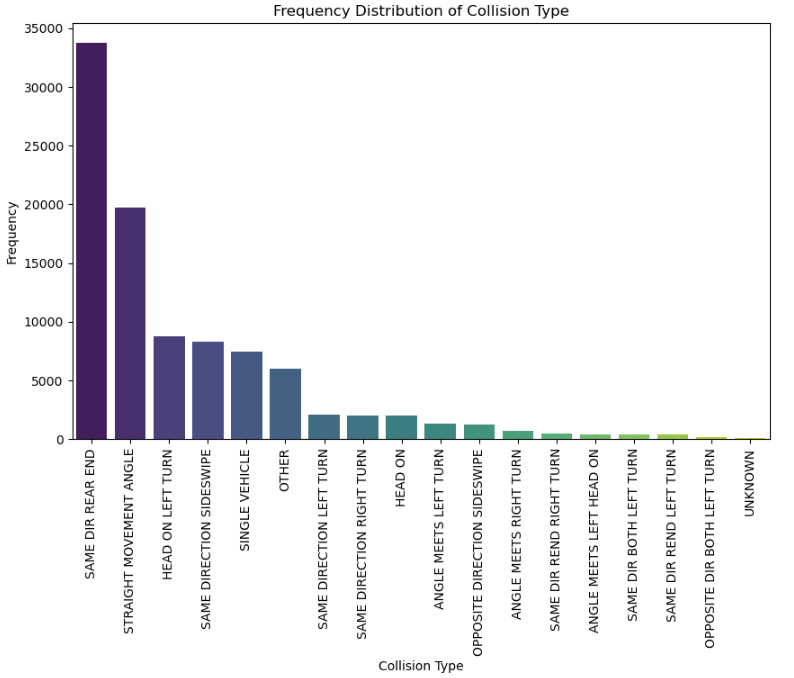


Figure 2. Frequency Distribution of Collision Type

The above bar chart provides insights into the frequency distribution of different collision types. The most common report is the "Property Damage Crash," outnumbering "Injury Crash" and with virtually no reports for "Fatal Crash." As for collision types, "Same Direction Rear-End" collisions dominate, with counts surpassing 35,000, followed by "Straight Movement Angle" nearing 20,000. Other collision types like "Head On Left Turn," "Same Direction Sideswipe," and "Single Vehicle" each hover around 10,000 cases. Less frequent types fall below 5,000, with the least common being "Opposite Direction Both Left Turn" and "Unknown."

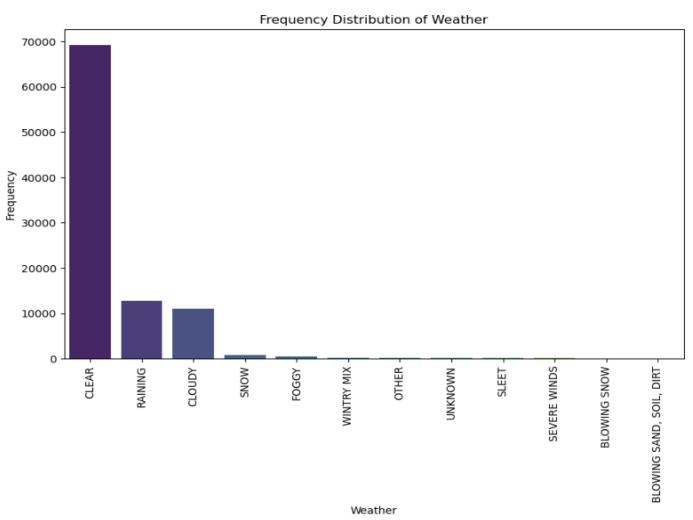


Figure 3. Frequency Distribution of Weather conditions

"Clear" weather overwhelmingly dominates with over 60,000 occurrences. Following that, "Raining" and "Cloudy" conditions have a significant presence, each hovering around or just below 20,000 occurrences. The frequency drops sharply for other conditions like "Snow" and "Foggy," which are below 10,000. Several other weather conditions such as "Wintry Mix," "Other," and "Unknown" register minimal counts, while conditions like "Sleet," "Severe Winds," "Blowing Snow," and "Soil Drift" are almost negligible in frequency.

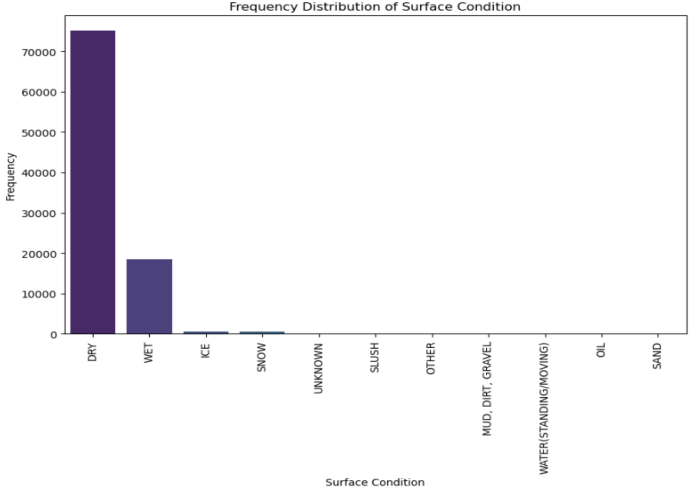


Figure 4. Frequency Distribution of Surface conditions

The "Dry" condition is the most prevalent, with frequencies surpassing 60,000. This is followed by "Wet" conditions, which have a significant but considerably lesser count, roughly around 20,000. Other conditions such as "Ice", "Snow" and "Unknown" have minimal frequencies in comparison. The remaining surface conditions are nearly negligible in their occurrences.

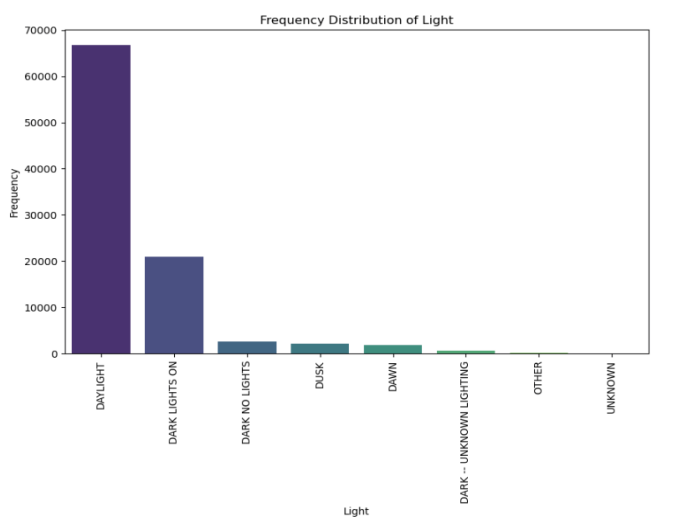


Figure 5. Frequency Distribution of Lights

"Daylight" is overwhelmingly the most common, with frequencies nearing 70,000. The second most frequent condition is "Dark with Lights On," approximating 20,000. Conditions like "Dark with No Lights", "Dusk", and "Dawn" have considerably lower counts, each below 10,000. "Dark with Unknown Lighting", "Other" and "Unknown" have minimal representations, making them relatively rare in this dataset.

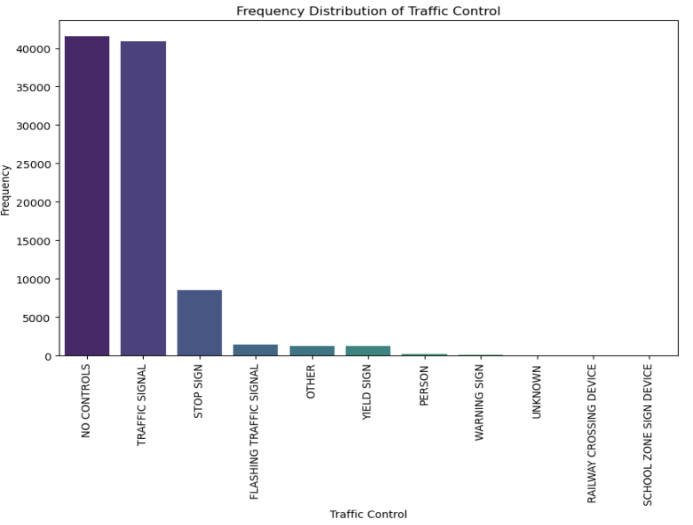


Figure 6. Frequency Distribution of Traffic Controls

"No Controls" and "Traffic Signal" are the most prevalent, with their frequencies approaching or exceeding 40,000. "Stop Sign" is the next most common, but significantly less so, with a count around 10,000. "Flashing Traffic Signal" and "Other" are lesser in frequency, below 5,000. The remaining categories, which have minimal frequencies, indicate they are rare occurrences in this dataset.

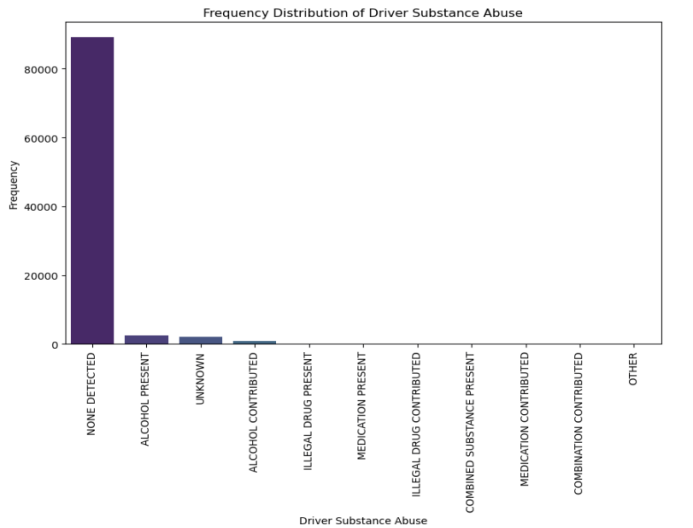


Figure 7. Frequency Distribution of Driver Substance Abuse

The overwhelming majority of occurrences are categorized as "None Detected," which has a frequency close to 80,000. All other categories have such low frequencies that they are not visibly discernible on the chart.This indicates that in the represented data, the vast majority of drivers did not have any detectable substance abuse.

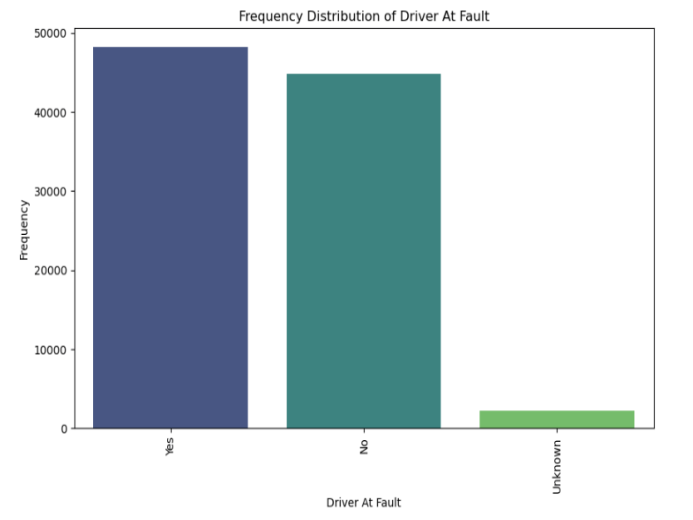


Figure 8. Frequency Distribution of Driver At Fault

The category "Yes" indicates that the driver was at fault. It has a frequency of approximately 50,000. While the category "No" denotes instances where the driver was not at fault and its frequency is slightly below that of the "Yes" category, but still quite substantial. A much smaller number of cases fall under the "Unknown" category, suggesting that in those situations, it's unclear whether or not the driver was at fault.

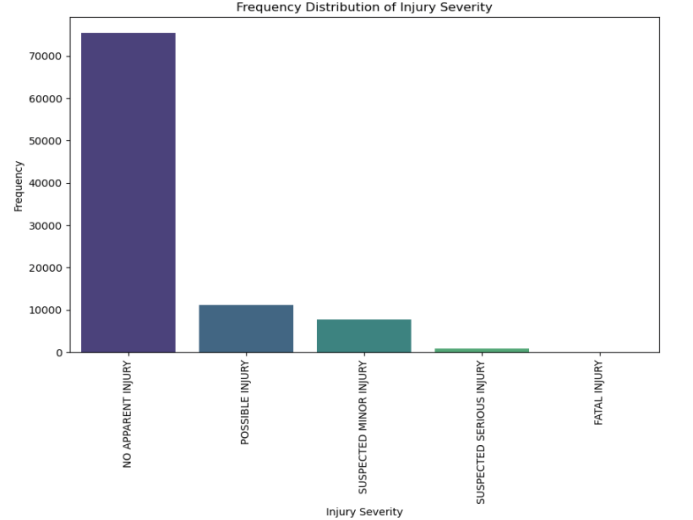


Figure 9. Frequency Distribution of Injury Severity

The majority, approximately 70,000 cases, resulted in no apparent injuries. There's a significant drop with around 20,000 possible injuries and numbers further decline as injury severity increases. Fatal injuries, fortunately, are the least common outcome in this dataset.

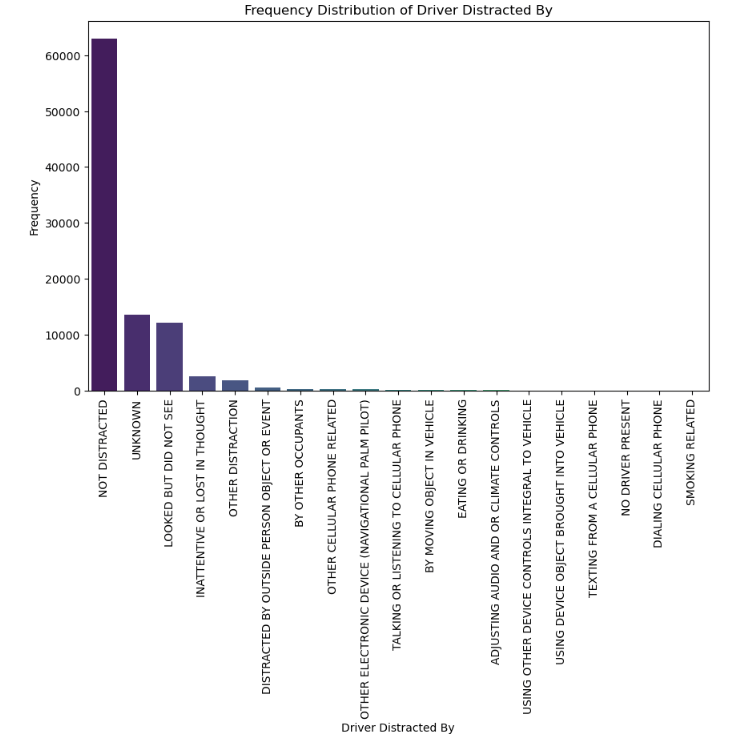


Figure 10. Frequency Distribution of Distractions Experienced By Drivers

A vast majority, nearing 60,000 occurrences, involved drivers who were not distracted. A smaller yet notable proportion experienced unknown distractions or were inattentive. Other notable reasons for distraction include getting lost in thought and distractions by outside persons or events. Electronic device use, like talking or texting on cell phones, constituted fewer cases. Overall, the data highlights that while technology can be a distraction, there are many other causes leading to driver inattention.

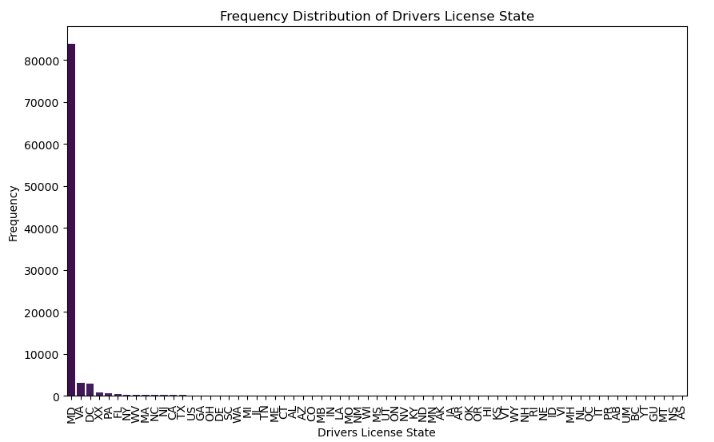


Figure 11. Frequency Distribution of Drivers License States

It's evident that the state MD dominates with nearly 80,000 occurrences, which makes sense because the dataset is from Maryland State Police. The remaining states have a significantly lower frequency, almost negligible in comparison.

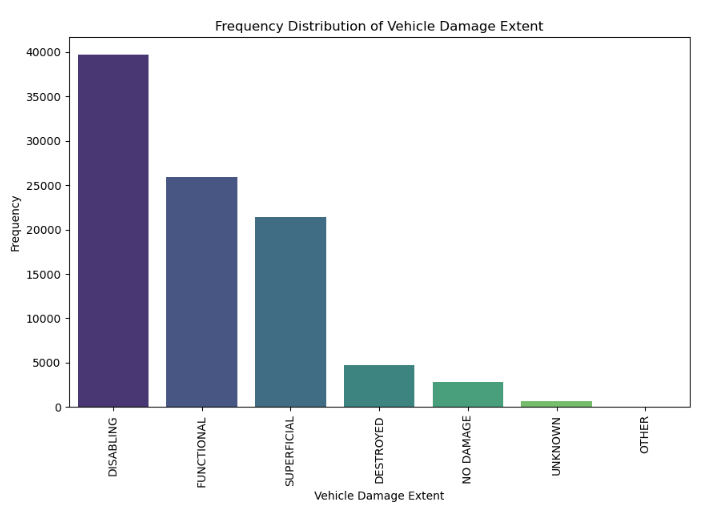


Figure 12. Frequency Distribution of Vehicle Damage Extent

The most prevalent damage type is "Disabling" followed by "Functional" and "Superficial". Categories like "Destroyed" and "No Damage" have considerably fewer occurrences. The "Unknown" and "Other" categories also have minimal frequencies.

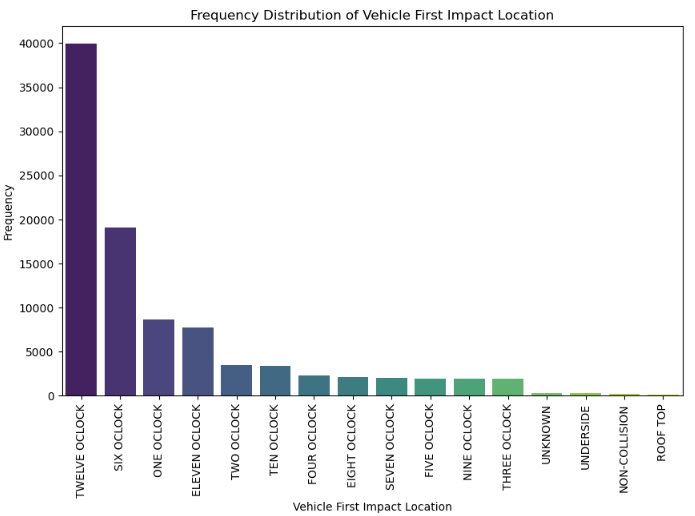


Figure 13. Frequency Distribution of Vehicle First Impact Location

The highest frequency of impact occurs at the "Twelve O'Clock" position, indicating frontal collisions. The subsequent highest frequencies are found at the "Six O'Clock" and "One O'Clock" positions. Other locations, such as sides and rear (Three O'Clock, Nine O'Clock, etc.), register lower frequencies.This suggests that head-on collisions or frontal impacts are predominant in this dataset.

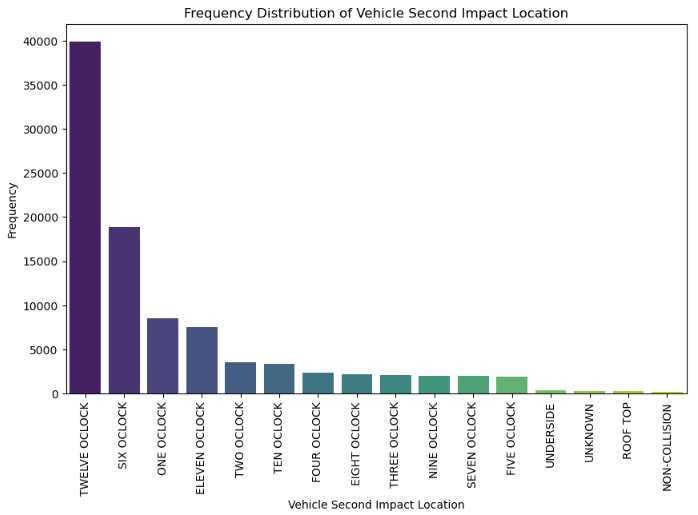


Figure 14. Frequency Distribution of Vehicle Second Impact Location

The data from the second impact largely mirrors the trends observed in the first impact. However, the prominence of frontal and rear collisions even in the second impact underscores the potential severity and complexities of multi-car accidents or scenarios where vehicles are pushed into subsequent collisions after an initial impact.

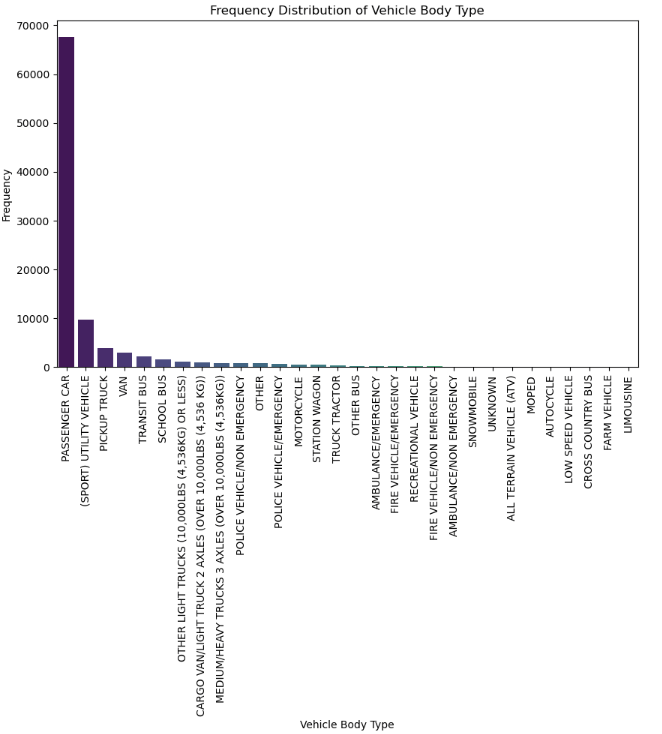


Figure 15. Frequency Distribution of Vehicle Body Type

The roadways are dominated by passenger cars, with a significant presence of utility vehicles and transport buses. While heavy trucks and emergency vehicles aren't as common, their roles in transportation and emergency response, respectively, make them crucial components of road traffic. The wide range of other vehicle types showcases the diversity of vehicles utilized for various purposes.

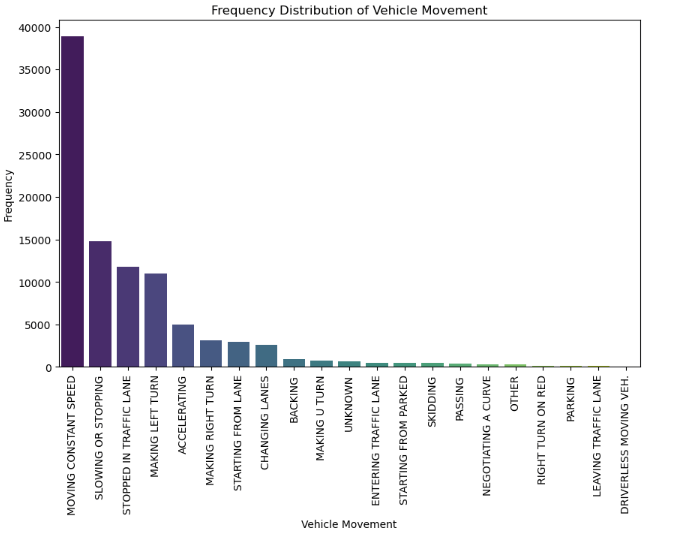


Figure 16. Frequency Distribution of Vehicle Movement

The majority of vehicles maintain a steady pace, as indicated by the high frequency of "Moving at Constant Speed". This is closely followed by vehicles "Slowing or Stopping" and those "Stopped in Traffic Lane", likely due to traffic signals or congestion. Turns are also common, with both "Making Right Turn" and "Making Left Turn" being frequently observed, hinting at numerous intersections or directional changes. Additionally, "Changing Lanes" and "Accelerating" are notable movements, signifying active road maneuvering and speed adjustments. Meanwhile, movements like "Starting in Traffic Lane" and "Backing" reflect temporary halts and reverse maneuvers, respectively.

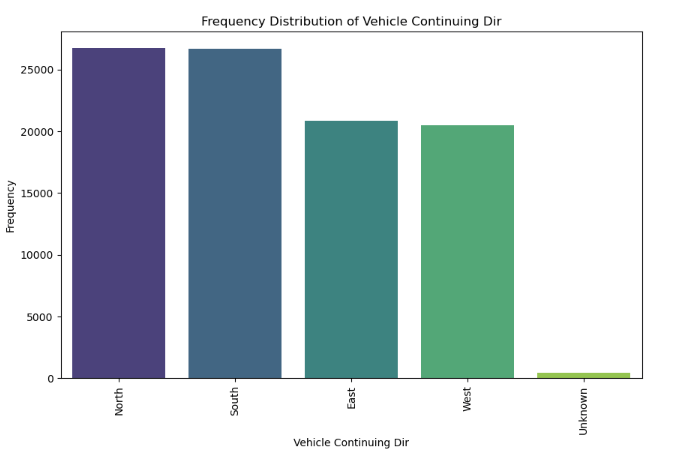


Figure 17. Frequency Distribution of Vehicle Continuing Direction

It's evident that the North direction is the most predominant, closely trailed by the South. The East and West directions showcase nearly identical frequencies, indicating a balanced vehicular movement between these two cardinal directions. A negligible number of entries fall under the "Unknown" category.

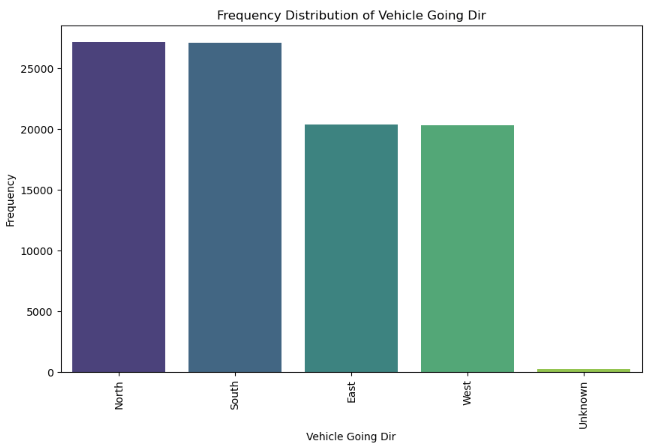


Figure 18. Frequency Distribution of Vehicle Going Direction

The vehicle going direction chart is almost the same as the previous chart, both showing the North direction appears as the leading direction, with the South being very close in frequency. The East and West directions exhibit similar frequency counts.

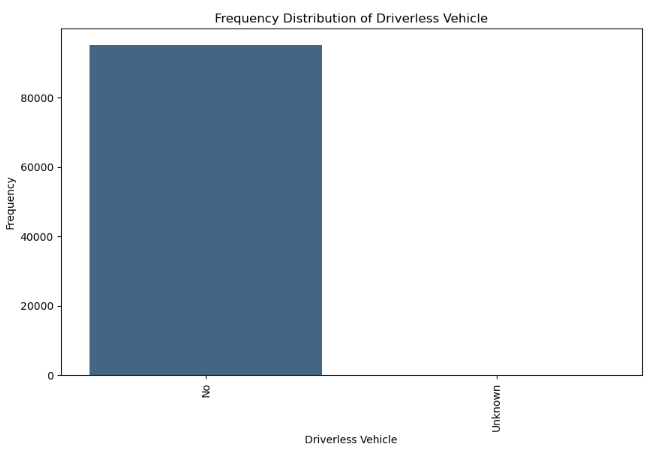


Figure 19. Frequency Distribution of Driverless Vehicle

From the chart, it's clear that a large number of vehicles are categorized as "No" for being driverless, while a significantly smaller number fall under the "Unknown" category. This indicates that driverless technology, as per this dataset, has yet to permeate the majority of vehicles or that the information about many vehicles' driverless status remains undetermined.

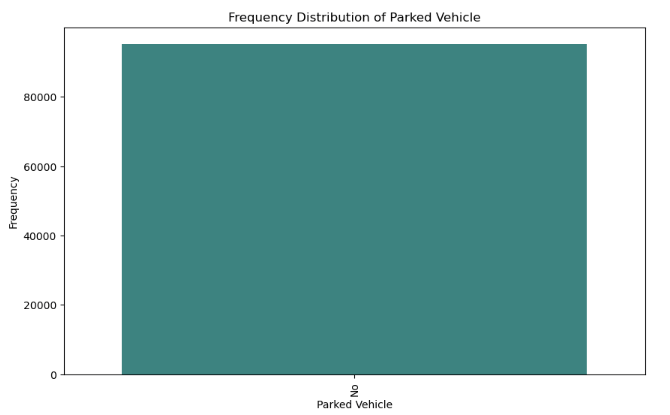


Figure 20. Frequency Distribution of Parked Vehicle

From the chart above, we could tell that all of the vehicles in the dataset are labeled as "No" for being parked.

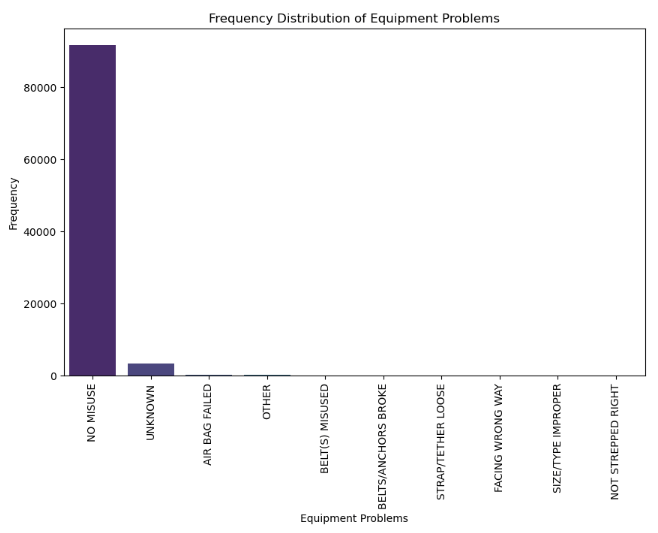


Figure 21. Frequency Distribution of Equipment Problem

A substantial majority of entries fall under the "No Misuse" category, suggesting that most vehicles in this dataset did not exhibit any equipment problems. A smaller segment shows equipment issues labeled as "Unknown". Various other specific equipment problems such as "Air Bag Failed", "Belt Misused", "Seat Anchors Broke", and others are represented but with considerably lower frequencies. The minimal representation of these specific problems emphasizes the overall good condition or proper usage of equipment in the majority of vehicles within this dataset.

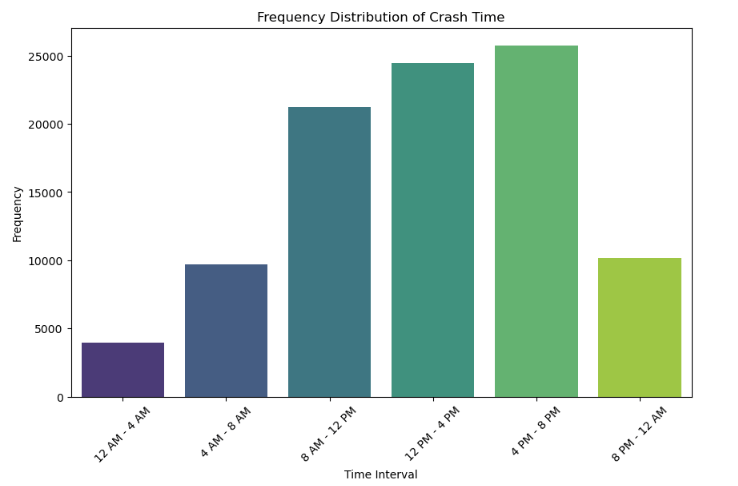
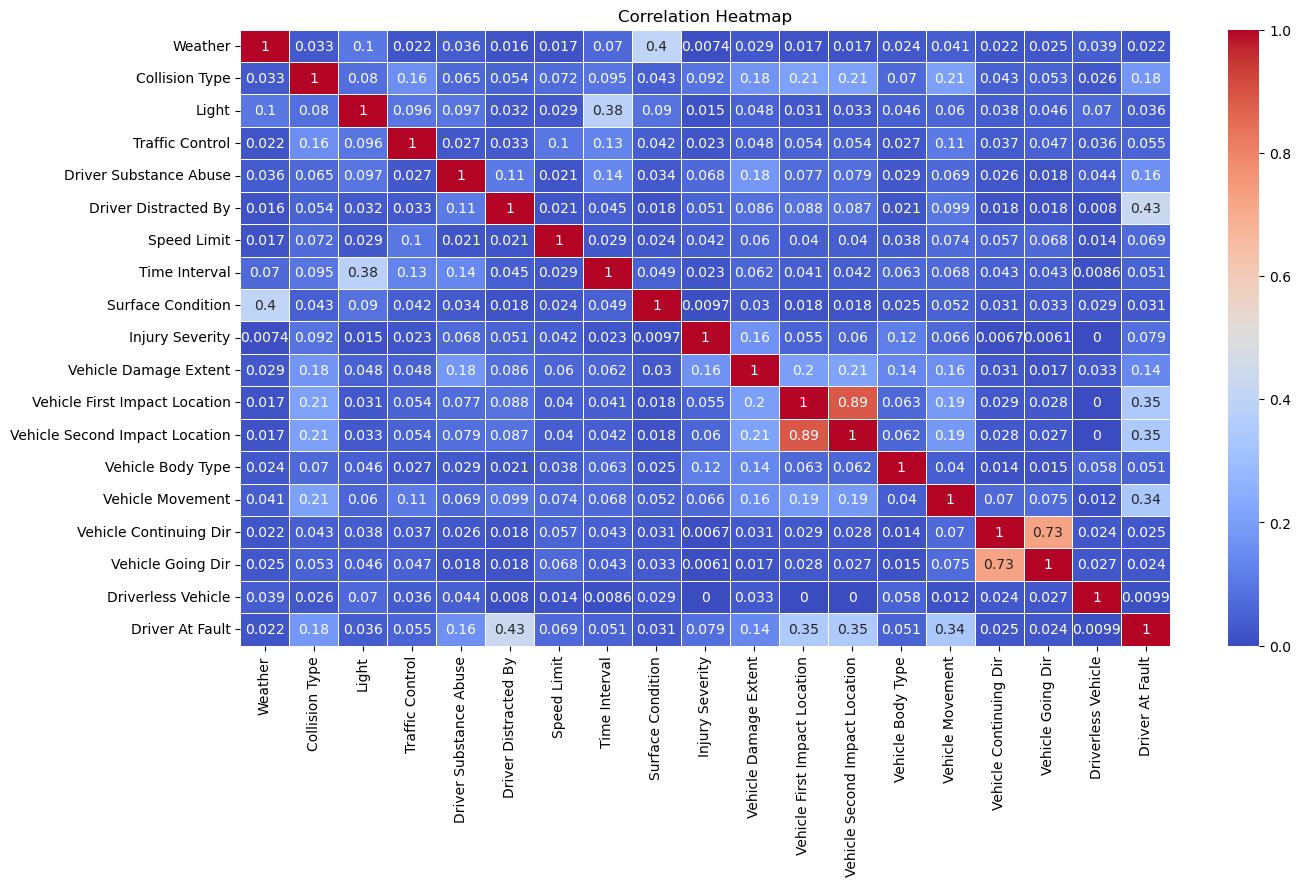
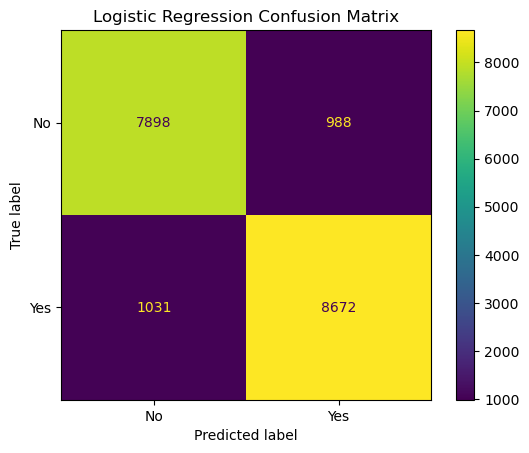
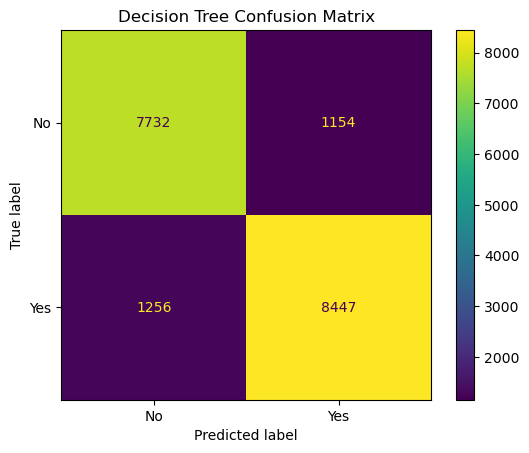
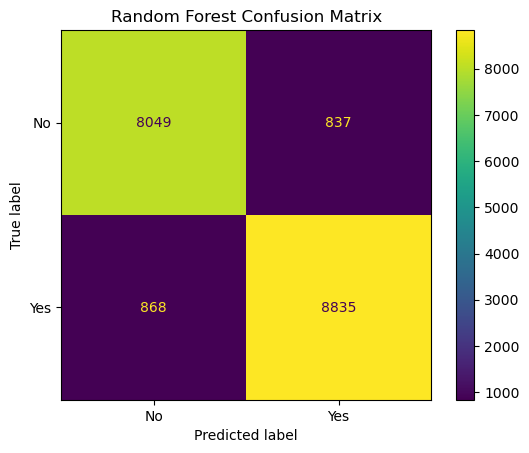
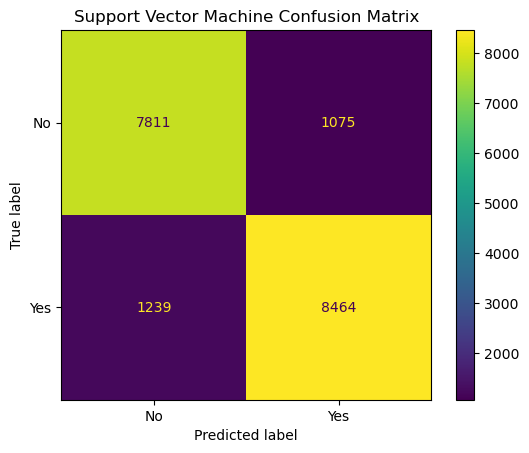


Figure 22. Frequency Distribution of Crash Time

We defined ‘time\_bins’, which are hour intervals that we categorized crash times into. As the graph shows, we set the time interval for every 4 hours. The lowest frequency of crashes occurs between midnight and 4 AM. A rise is noted from 4 AM to 8 AM, and it peaks from 8 AM to 12 PM. The afternoon hours from 12 PM to 4 PM and early evening from 4 PM to 8 PM have similar high frequencies, marking them as time intervals with the most crashes. The frequency then decreases notably between 8 PM and midnight. From this data, it can be inferred that the highest incidence of crashes occurs during daylight hours, particularly from late morning to early evening, while nighttime sees a considerable reduction in crash occurrences.

* **Predictive Analysis:**  
  Among the vast 43 attributes, we meticulously picked 12, each exhibiting a correlation value greater than 0.05 with driver fault (using Cramer’s V correlation analysis), crafting a focused lens through which we view this complex puzzle of classifying the fault of the driver in these collisions. This endeavor is not just about attributing blame, but about creating ripple effects: redefining insurance paradigms, empowering law enforcement with data-backed insights, and enabling drivers to introspect and evolve. All of the details of the features used below are confirmed by data visualizations including frequency distributions of categorical variables from the exploratory data analysis section and also the heatmap below generated with Cramer’s V correlation analysis which includes all variables after cleaning the dataset.
* 
* · Vehicle First Impact Location: A major chunk of the first impact happens at the front of the vehicle. This again hints towards rear-end collisions and emphasizes the significance of maintaining a safe following distance.
* · Vehicle Second Impact Location: While the first impact is telling, the second impact provides insights into the subsequent course of events post-collision. A substantial number of vehicles don't have a second impact location, but for those that do, it's predominantly the side, suggesting a swerve post initial collision.

· Vehicle Body Type: Sedans and SUVs account for the majority of vehicles involved in these collisions. This could be a reflection of their popularity or indicate specific vulnerabilities associated with these vehicle types.

* · Vehicle Movement: Going straight prior to the collision has the highest count. This emphasizes the importance of lane discipline, signaling, and awareness, even when not making turns or overtaking.
* · Collision Type: Rear-end collisions dominate the dataset. This type of collision often signifies distracted driving, sudden braking, or aggressive tailgating behaviors.
* · Traffic Control: A noteworthy number of collisions happen in areas where there's a traffic signal. Whether it's malfunctioning signals, improperly timed lights, or drivers running red lights, this indicates a key area for intervention.
* · Driver Substance Abuse: A concerning number of drivers were under the influence of drugs or alcohol. This dangerous behavior is a reminder of the perils of impaired driving and the importance of stringent DUI checks.
* · Driver Distracted By: Distractions, especially mobile phones, emerge as significant contributors. This underscores the life-threatening risks of multitasking while driving.
* · Speed Limit: Collisions predominantly occur in zones with speed limits between 30-50 mph. This range is typical for urban and suburban areas and indicates zones that may benefit from additional traffic calming measures.
* · Time Interval: The late afternoon to early evening time frame, typically corresponding to rush hours, sees a spike in collisions. This could be attributed to increased traffic, driver fatigue, or even factors like sun glare.
* · Injury Severity: While many collisions result in non-incapacitating injuries, the toll of even minor injuries can be significant in terms of trauma, recovery time, and associated costs.
* · Vehicle Damage Extent: Many vehicles involved in these collisions sustain substantial damage, emphasizing the financial, safety, and even environmental repercussions of these accidents.
* **Models Employed & Their Rationale:**
* Our central task was to determine if a driver was at fault. We used an 80/20 test/train split for all the models. The confusion matrices provided a granular view of each model's capabilities in this regard:
* · True Positives (TP): These represent the instances where drivers were rightly predicted to be at fault. A high TP rate across models underscores their effectiveness.
* · True Negatives (TN): This metric captures the cases where drivers were correctly identified as not being at fault. High TN values reaffirm our confidence in the models.
* · False Positives (FP) & False Negatives (FN): These are the cases where models erred. FP indicates drivers wrongly classified as being at fault, while FN represents the opposite. These metrics emphasize the challenges inherent in collision data and remind us that while our models are robust, they aren't infallible.
* Here are details about models, their accuracy and why they were chosen:
* · Logistic Regression: This model excels in binary outcomes, making it apt for our 'at-fault' or 'not-at-fault' classification. With an accuracy of 89.14%, it provided a foundational lens, establishing baselines and offering preliminary insights. The matrix for logistic regression showcased a high number of true positives and true negatives, suggesting it adeptly distinguishes between fault categories. However, like any model, it's not without its false positives and false negatives, hinting at potential areas of improvement or the complexities inherent in some collision scenarios.
* 
* · Decision Tree Classification: Hierarchical by design, this model effectively broke down the significance hierarchy of features. At 87.04% accuracy, it not only predicted outcomes but also visually conveyed the decision-making process, offering interpretability. The decision tree, while offering a structured view of feature importance, had a slightly lower accuracy. Its matrix revealed instances where it might have misclassified certain cases, possibly due to overfitting or not capturing certain nuances.
* 
* · Random Forest Classification: An ensemble of decision trees, this model thrives in capturing complex data patterns. Its superiority was evident with a 90.83% accuracy. By leveraging multiple decision trees, it reduced overfitting, ensuring robust and generalized predictions. Random Forest emerged as the most accurate model. Being an ensemble method, it aggregates results from multiple trees, providing a more balanced and nuanced classification. The matrix underscored its proficiency in accurately predicting both fault categories, making it a promising tool for our task.
* 
* · Support Vector Machine (SVM): Perfect for high-dimensional datasets, especially post one-hot encoding, SVM exhibited an accuracy of 87.55%. It worked by finding the optimal hyperplane that best divides the dataset into classes, making it suitable for our rich feature set. Its matrix, however, showed certain misclassifications, possibly hinting at the challenges posed by the non-linear boundaries of our data or the need for hyperparameter tuning.
* 
* While each model offers unique insights, together they present a holistic picture. Logistic Regression provides foundational understandings; Decision Trees visually break down decisions, highlighting key determinants; Random Forests, with their ensemble approach, capture nuances and intricate patterns; and SVM, with its capacity to handle high-dimensional data, ensures that even subtle, less apparent patterns are not missed. By juxtaposing the insights from confusion matrices with our task, it's evident that while we can predict driver fault with high confidence, there's room for refinement. Whether it's through feature engineering, hyperparameter tuning, or exploring other algorithms, these matrices serve as a roadmap for continuous model enhancement. In summary, the collective insights from these models solidify our main goal/question for the project: specific features, when synergized, can predict driver fault with remarkable accuracy.

**Interpretation:**

To address our research question of fault determination in traffic collisions, we performed an extensive analysis on the dataset. Our approach involved correlational analysis, aimed at identifying the most influential variables impacting the target variable, which is driver’s faults. Specifically, we selected columns with a correlation value below 0.05, as these variables exhibited the most significant impact on fault determination. This methodological choice was crucial in ensuring our models would be robust and more accurate in their predictions.

The results of our analysis highlighted the Random Forest Classification model as the standout performer, demonstrating the highest accuracy among the models considered. This choice of model is pivotal in ensuring that fault determination is not only automated but also reliable and efficient.  
**Conclusions:**

In conclusion, our analytics project on traffic collision data from Montgomery County, Maryland, has yielded valuable insights and conclusions. We successfully addressed the central question of fault determination, which has far-reaching implications for various stakeholders in the context of road safety. Our primary goals have been met, and we are poised to make the following recommendations based on our findings:

* Adoption of Random Forest Classification: We recommend the adoption of the Random Forest Classification model as the preferred choice for automating fault determination. This model demonstrated the highest accuracy and robustness in our analysis, making it an efficient tool for stakeholders involved in accident resolution.
* Enhanced Road Safety Measures: Our findings highlight the potential for policy interventions in areas with a higher incidence of at-fault collisions. Law enforcement and policymakers can use this information to implement targeted safety measures, improving road safety.
* Improved Driver Awareness: By providing insights into fault determination, this system can be leveraged to enhance driver awareness of their own driving behavior. This information can help drivers identify areas for improvement and reduce accidents.

Overall, this project has not only addressed the pressing issue of fault determination but also laid the foundation for an automated system that can revolutionize accident resolution and contribute to safer roadways in Montgomery County and beyond.

References

Montgomery County, M. (2023, September 22). *Crash reporting - drivers data: Open Data Portal*. Crash Reporting - Drivers Data | Open Data Portal. <https://data.montgomerycountymd.gov/Public-Safety/Crash-Reporting-Drivers-Data/mmzv-x632>.

Publisher data.montgomerycountymd.gov. (2023, September 22). Crash reporting - drivers data. Catalog. <https://catalog.data.gov/dataset/crash-reporting-drivers-data>.