Crash Reporting Analysis: Unveiling Patterns and Predictors of Traffic Incidents

Group 4: *Divya Chenduran Ayyemperumal, Manasa Rao, Sanjana Poojary*

I. SUMMARY

This project focuses on analyzing traffic collision data in Montgomery County to reveal patterns, pinpoint high-risk zones, and determine the primary factors that lead to accidents. Utilizing a dataset from the Maryland State Police's Automated Crash Reporting System (ACRS), this study examines detailed records of traffic collisions on county and municipal roads, contributed by various local law enforcement bodies. The dataset includes specifics like the date, time, and location of crashes, environmental conditions, types of collisions, and details about the vehicles, drivers, and the extent of injuries or fatalities involved.

The objective is to use this data to enhance road safety in Montgomery County by identifying patterns and risk factors, such as peak times for accidents and the impact of weather or driver impairment. The insights gained will guide local authorities and the public in implementing effective safety measures, infrastructure modifications, and policy updates aimed at reducing traffic collisions.

The background of this project centers on addressing the recurring issue of traffic accidents in Montgomery County, seeking to understand their causes and effects through a data-driven approach. Previous related studies have likely examined traffic patterns and safety interventions, providing a foundation for this analysis. The methodology involves statistical analysis of the crash data to identify trends and correlations. This approach will help in understanding the dynamics of traffic incidents and formulating evidence-based recommendations for improving road safety.

In summary, the project aims to provide a comprehensive analysis of traffic collisions in Montgomery County, offering practical solutions to mitigate risks and enhance overall road safety. Through detailed data examination and analytical methods, the project intends to contribute to a safer driving environment in the county.

II. METHODS

Data Pre-Processing, Tidying and Transformation

We perform several key preprocessing tasks on the dataset related to vehicle collisions. There are three main data sources.

Crash_data.csv: This CSV file contains comprehensive data on collision reports. The specific details of each collision, such as report numbers, case numbers, report types, and various attributes of the crash incident, are included.

Incidents_data.csv: This dataset contains detailed records of individual collision incidents. It includes specific details regarding the circumstances of each collision, such as whether it was a hit-and-run, the road conditions, and other incident-specific information.

Non_motorist_data.csv: This file includes data related to collisions involving non-motorists. This encompasses information about pedestrians, cyclists, or any party involved in a collision that was not in a motor vehicle. It also includes data on pedestrian movements, actions, and any traffic signals involved.

These three datasets were loaded separately and then combined into one dataset using join operations on common identifiers, which are Report Number and Local Case Number. The combined dataset allows for a comprehensive analysis that considers multiple aspects of collision incidents, including details on non-motorists and specific incidents.

1. **Data Pre-Processing:**
   In this stage, the environment is prepared by ensuring necessary libraries are installed and loaded, creating a standardized toolbox of functions. The data is then imported into R, with specific placeholders identified as missing values. This sets the stage for a consistent analytical workflow and acknowledges the presence of data imperfections early on.

2. **Data Merging:**
   The ‘Local Case Number’ column from all three datasets is converted to a character type to ensure consistency for joining. These data frames are then merged into one combined data frame using full_join from the dplyr package, with 'Report Number’ and ‘Local Case Number’ as the keys. This creates a large data frame that includes all rows and columns from all datasets, aligning them by the join keys.

3. **Data Tidying:**
   Data tidying involves organizing the data so that it is consistent and easy to work with. This includes converting the text field to lowercase to maintain consistency. By converting all text to lowercase, the data is standardized, which reduces redundancy and confusion that can arise from case-sensitive text.

4. **Data Transformation:**
   Missing values are systematically replaced with standardized placeholders like "unknown" for text or 0 for numbers, ensuring the data can be processed by analytical algorithms without error. The date-time information is parsed into a proper format, and time components are extracted, enriching the dataset with structured temporal information for time-based analysis. Based on the parsed date-time object, new variables are derived that extract specific components of the date-time information.
   This includes:
   Date: Extracting the date part only.
   Hour: Extracting the hour component from the time.
   Weekday: Naming the day of the week.
   Month: Naming the month.
   These derived variables offer new dimensions for analysis such as trend analysis over time or aggregation by the day of the week.

Exploratory Data Analysis (EDA)

We performed some Exploratory Data Analysis to analyze the datasets to summarize their main characteristics, with visual methods. EDA is a way to look at what the data can tell us beyond the formal modeling or hypothesis testing task, for seeing what the data can tell us before making any assumptions.

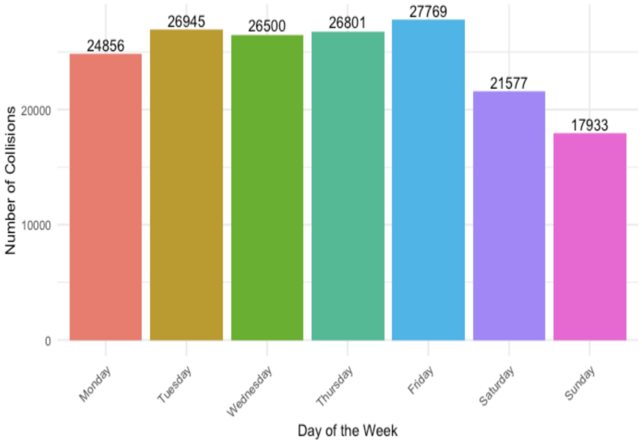

Figure 1: Variation in collision numbers on different days of the week

This bar chart provides a visual representation of the number of collisions according to the days of the week. The x-axis represents the days of the week from Monday to Sunday. The y-axis indicates the number of collisions. Each bar's height corresponds to the total count of collisions that occurred on each day of the week. This visualization reveals preliminary patterns, such as the variation in collision numbers on different days, with the highest on Friday and the lowest on Sunday. The decrease on weekends, especially Sundays, could be due to lower traffic volumes as people tend to stay home or there's less commuting.

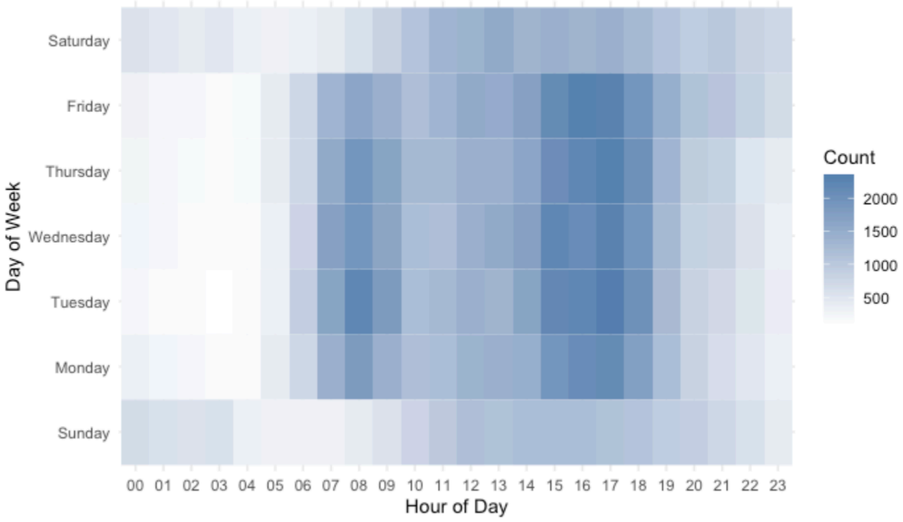

Figure 2: Heatmap of collisions by hour and day of the week

The heatmap represents the number of collisions according to the hour of the day and the day of the week. The x-axis shows the hour of the day, ranging from 00 (midnight) to 23 (11 PM). The y-axis represents the days of the week. The color intensity within the heatmap indicates the count of collisions. Darker shades represent a higher number of collisions, while lighter shades indicate fewer collisions. Certain times of day, like morning and evening rush hours, have darker shades, indicating higher collision frequencies during those times. Midday and late-night hours show lighter shades, suggesting fewer collisions. The weekdays show different patterns compared to the weekend, potentially with the workweek showing more consistent collision occurrences.

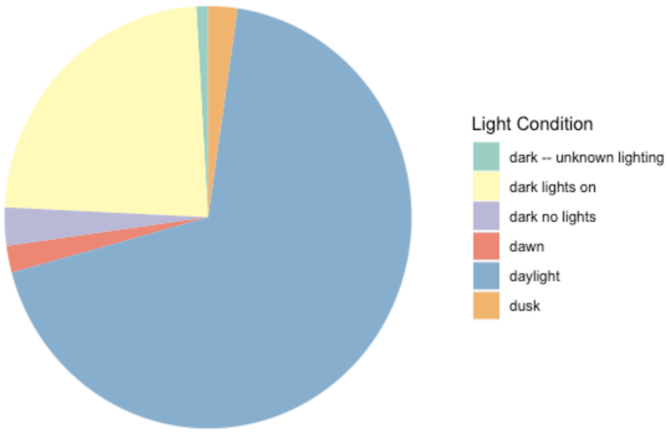

Figure 3: Proportion of accidents by light conditions

The pie chart shows the proportion of accidents by light conditions. Each slice of the pie represents a different light condition under which the accidents occurred. The size of each slice is proportional to the number of accidents that occurred in that particular light condition. From the chart, it seems that a significant majority of accidents occur in daylight, indicated by the largest slice in blue, this is followed by dark with lights on. This could reflect the overall higher volume of traffic during daylight hours. Other light conditions such as dawn, dusk, and dark make up a smaller portion of the total accidents.

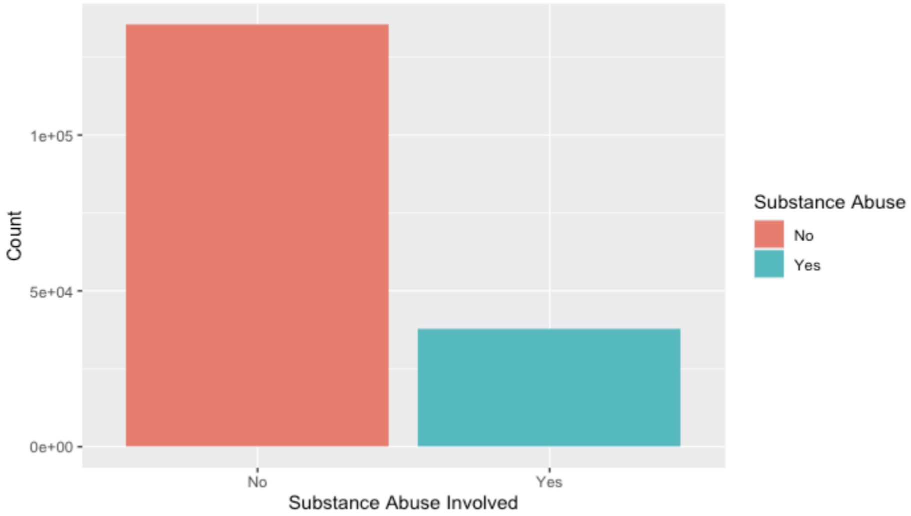

Figure 4: Substance abuse factor in collisions

This bar chart visualizes the role of substance abuse in collisions. The x axis displays whether substance abuse was reported to be involved in the collisions. The y-axis represents the count of collisions. The scale of the y-axis is logarithmic, as indicated by the "1e+05" notation, which allows for a more manageable display of wide-ranging values. The height of the bar for 'No' is significantly taller than the one for 'Yes,' suggesting that there were many more collisions where substance abuse was not reported as a factor. It's also worth noting that the actual number of incidents involving substance abuse may be underreported.

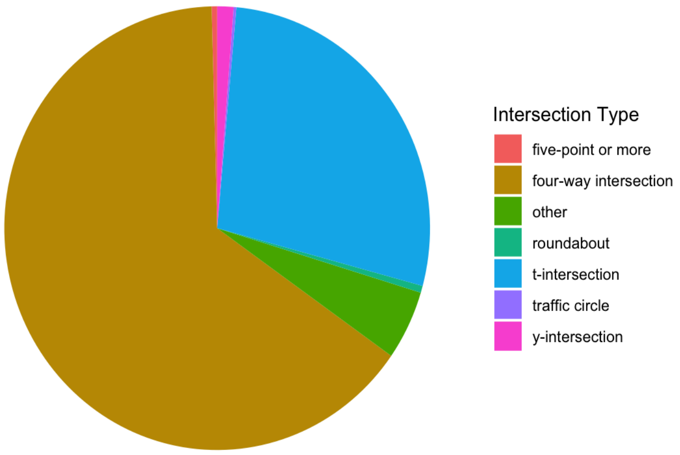

Figure 5: Intersection Type

The chart is divided into segments, each representing a type of intersection involved in crashes. The largest segment, colored in brown, represents **“four-way”** intersections. This suggests that these types of intersections have the highest occurrence in crashes. Another most common type of intersection involved in crashes is “t-intersection”. Traffic authorities could focus on these areas for potential safety improvements. This chart could be useful for traffic authorities to understand which types of intersections are most prone to crashes and could potentially benefit from safety improvements.

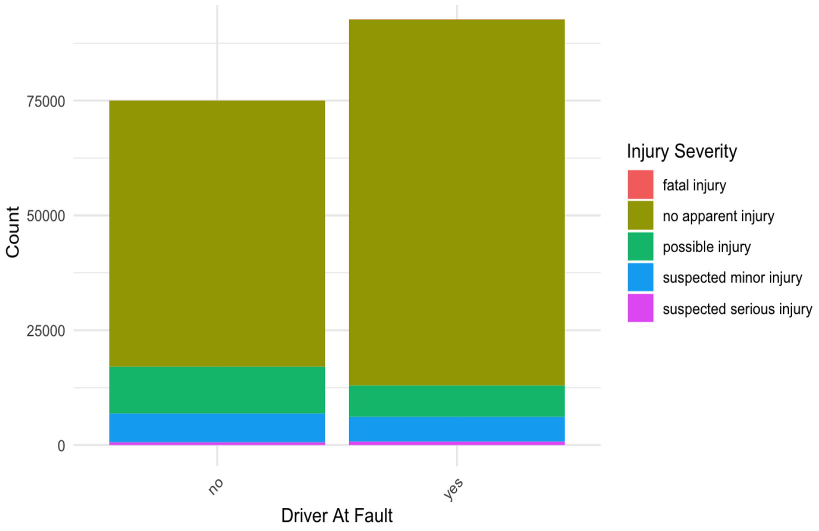

Figure 6: Injury Severity by Driver at Fault

In this graph, the x-axis represents whether the driver is at fault or not, with categories “no” and “yes”. The y-axis represents the count of incidents, ranging from 0 to 75,000. In conclusion, the chart indicates that there are similar counts of injuries regardless of whether the driver is at fault or not. However, there are more cases of no apparent injury when the driver is at fault.

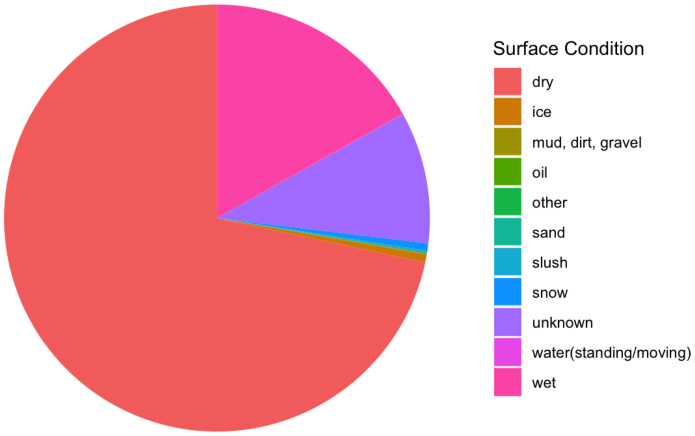

Figure 7: Proportion of Surface conditions

The chart is divided into segments, each representing different types of surface conditions, including dry, ice, mud/dirt/gravel, oil, other, sand, slush, snow, unknown, water(standing/moving), and wet. In conclusion, the chart indicates that a significant majority of the surfaces are dry, with smaller proportions being wet, and even smaller portions representing other conditions like ice, mud, dirt, gravel, etc. This suggests that most driving conditions are dry, but drivers should still be prepared for a variety of surface conditions.

Data Modeling

We have applied ARIMA and Regression analysis for data modeling in our crash reporting project. The choice of these two models was driven by the nature of our data and the specific analytical needs of the project.

ARIMA stands for AutoRegressive Integrated Moving Average. It's a popular statistical method for time series forecasting that uses past data points to predict future ones. For ARIMA, we recognized that traffic collision data inherently follows a time series structure, with potential correlations between observations at different time points. ARIMA is a robust method that accounts for such temporal dependencies. It can help us understand and predict future collision trends based on historical patterns. The model's integration aspect makes it particularly suitable for non-stationary data, which is a common characteristic of collision counts that may fluctuate over time due to various underlying factors.

On the other hand, regression analysis was chosen to identify and quantify the relationship between the number of collisions and other relevant variables, such as weather conditions, traffic volume, and road type. It provides a clear framework to assess the influence of these factors on collision frequency, allowing us to draw actionable insights. Moreover, by including a trend component, regression analysis helps us capture long-term patterns in the data, which complements the more short-term focus of the ARIMA model.

III. RESULTS

Our ARIMA model was fitted to this time series to both understand historical patterns and forecast future collisions for the next 30 days. We visualized the historical data, applying a polynomial regression to reflect non-linear trends, especially a noticeable pattern post-2020.

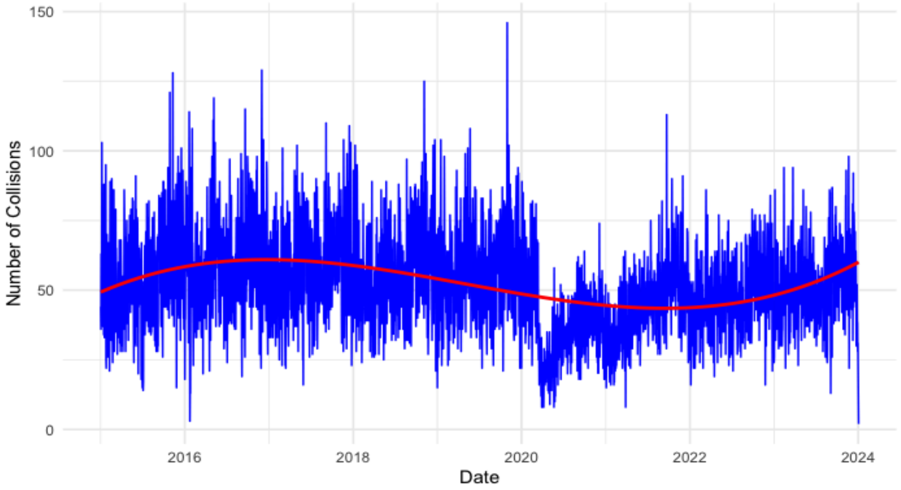

Figure 8: Historical Collision Data with Non-linear trend

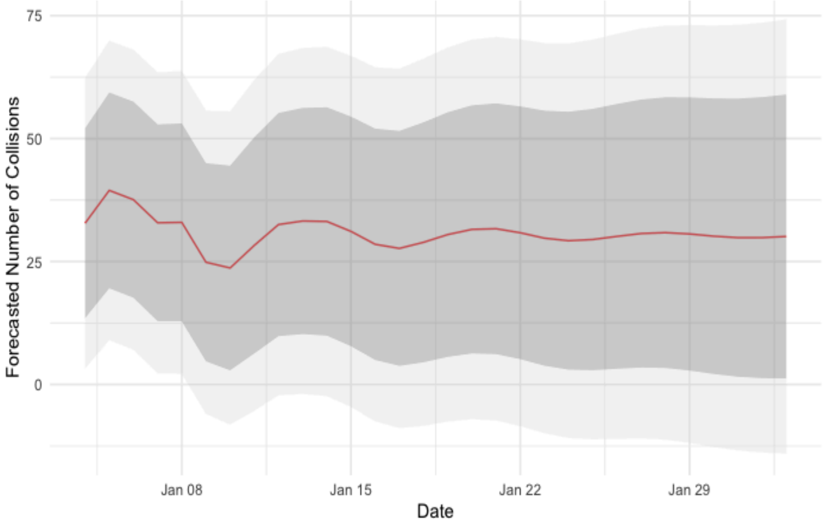

Figure 9: 30 day collision forecast

Regression analysis:

This regression analysis helps in understanding the relationship between the independent variables (weather, surface condition, light, vehicle characteristics) and the dependent variable (injury severity). It allows us to identify which variables have significant effects on injury severity and how changes in these variables impact the outcome. Additionally, the visualization provides a clear comparison between the actual and predicted values, aiding in the assessment of the model's predictive performance.

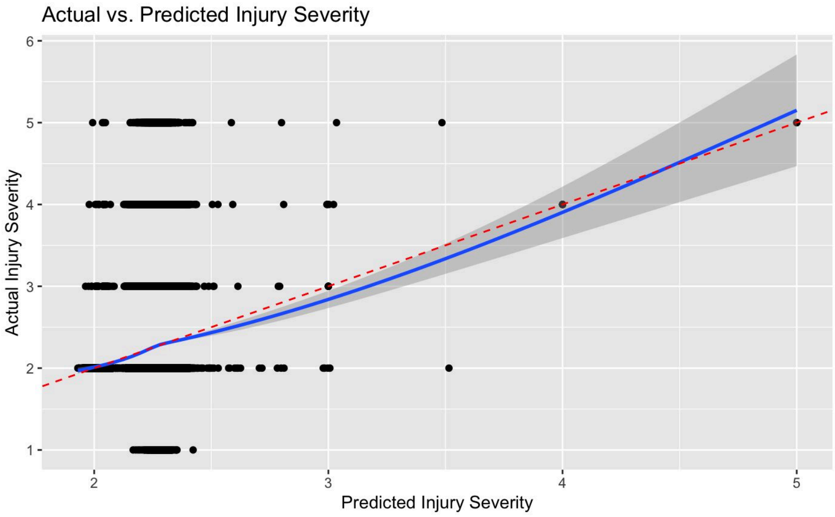

Figure 10: Actual vs. Predicted Injury Severity

The above plot is comparing actual injury severity against predicted injury severity from a regression model. Here's what the elements of the plot indicate:

Points (Black Dots): Each point represents an individual observation from the dataset, where the x-axis value is the injury severity predicted by the regression model, and the y-axis value is the actual recorded injury severity.
Solid Blue Line: This line represents the best-fit line from the regression analysis. Ideally, it should pass through the core of the distribution of the points. It represents the relationship the model has found between the predictors and the actual injury severity.
Dashed Red Line: This is likely the line of perfect prediction, where the predicted values are exactly equal to the actual values. If a point lies on this line, it means the prediction was exactly correct for that observation.
Shaded Area (Grey): This shaded area represents the confidence interval around the best-fit line, showing the range within which we can expect the true best-fit line to fall, given the current data, with a certain level of confidence (often 95%).

If most of the points are close to the solid blue line and it is close to the dashed red line of perfect prediction, it indicates a better fit of the model to the data. If there's a systematic pattern in the residuals (vertical distance between the points and the blue line), this suggests the model could be improved.

Additionally we also performed Pearson correlation which computes the correlation between two variables that provides a precise determination of correlation. Below is the conclusion of the Pearson correlation:

1. **Coefficients:** The coefficients of Vehicle_Make_Num and Vehicle_Model_Num are statistically significant at the 0.001 level. This means that these variables have a significant impact on the Injury_Severity_Num. The positive coefficient for Vehicle_Make_Num suggests that as the numeric encoding of the vehicle make increases, the injury severity tends to increase.
2. **Model Fit:** The Multiple R-squared value is 0.008243, and the Adjusted R-squared is 0.007425. These are measures of how well the model fits the data. However, these values are quite low, suggesting that the model explains only a small portion of the variability in the Injury_Severity_Num.
3. **Pearson Correlation:** The Pearson correlation between the actual and predicted values of injury severity is approximately 0.091. This indicates a weak positive linear relationship between the actual and predicted values.

In conclusion, our model suggests that the make and model of the vehicle have a significant impact on the severity of injuries in accidents. However, the overall fit of the model is not very strong, and it explains only a small portion of the variability in injury severity. The weak correlation between the actual and predicted values also suggests that the model's predictive performance is limited. This analysis can still provide valuable insights for crash analysis. For instance, it suggests that vehicle make and model might be important factors to consider when investigating the severity of injuries in accidents

IV. DISCUSSION

The ARIMA model analysis of historical traffic collision data in Montgomery County reveals a pattern of fluctuation, with a notable decline around 2020 and a rise thereafter. The 30-day forecast indicates a relatively stable trend in collision frequency in the near future, without significant fluctuations.

In contrast, our regression analysis examining the severity of injuries in accidents concludes that the make and model of vehicles are statistically significant predictors. However, the model demonstrates a modest explanatory power, suggesting that other unaccounted-for factors may also play a crucial role. Despite its limitations, this finding is a valuable addition to the understanding of factors influencing injury severity in traffic collisions.

Together, ARIMA and regression analysis offer a comprehensive approach to our data modeling needs. While ARIMA harnesses the chronological order of the data for forecasting, regression analysis deciphers the impact of different predictors, giving us a multi-faceted understanding of the dynamics at play in traffic collision occurrences.

V. STATEMENT OF CONTRIBUTIONS

Divya Chenduran Ayyemperumal: Data cleaning and manipulation and Documentation.
Manasa Rao: Data compilation, Exploratory Data Analysis, Modeling and Documentation.
Sanjana Poojary: Data compilation, Exploratory Data Analysis, Modeling and Documentation.

VI. REFERENCES

[1] https://www.geeksforgeeks.org/time-series-analysis-using-arima-model-in-r-programming/
[2] https://catalog.data.gov/dataset/crash-reporting-drivers-data
[3] https://data.montgomerycountymd.gov/Public-Safety/Crash-Reporting-Incidents-Data/bhju-22kf
[4]https://data.montgomerycountymd.gov/Public-Safety/Crash-Reporting-Non-Motorists-Data/n7fk-dce5