Crash Reporting Analysis: Unveiling Patterns and Predictors of Traffic Incidents

Group 4 – Divya Chenduran Ayyemperumal

Manasa Rao

Sanjana Poojary

Agenda

- Introduction
- Methods
  - Data Tidying and Transformation
  - Exploratory Data Analysis
  - Data Modeling
- Results
- Conclusion
- Future Work
- References

Introduction

- Conduct a comprehensive analysis of traffic collision data in Montgomery County.
- Primary goal is to enhance road safety by identifying collision patterns and the primary risk factors contributing to accidents.
- Data-driven methodology to statistically analyze the crash data.
- Employs three key datasets to enhance the breadth and depth of our safety analysis.

Methods

- Data Tidying and Transformation
- Exploratory Data Analysis (EDA)
- Data Modeling

Data Tidying and Transformation

Data Tidying:

- Organize data to make it consistent and easy to manipulate.
- Converting all text fields to lowercase to avoid redundancy and inconsistencies.

Data Transformation

- Handling missing values.
- Date-Time parsing.
- Extraction of Time Components: Derive new variables like Date, Hour, Weekday, and Month.

Exploratory Data Analysis

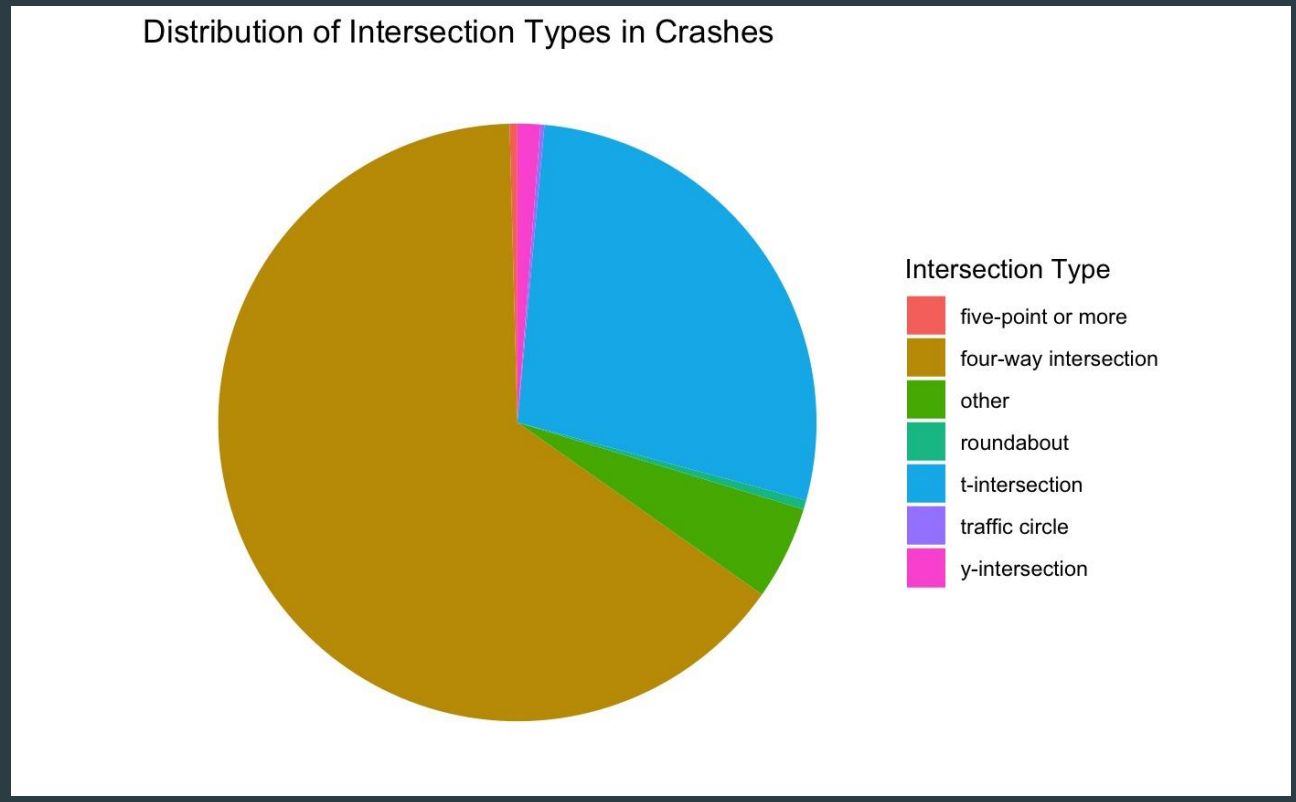


Figure 1: Distribution of Intersection types in Crashes

- Four-way intersections are the most common sites for crashes.
- Other common intersection types also show significant occurrences in crashes.
- Implications for traffic safety.

Exploratory Data Analysis

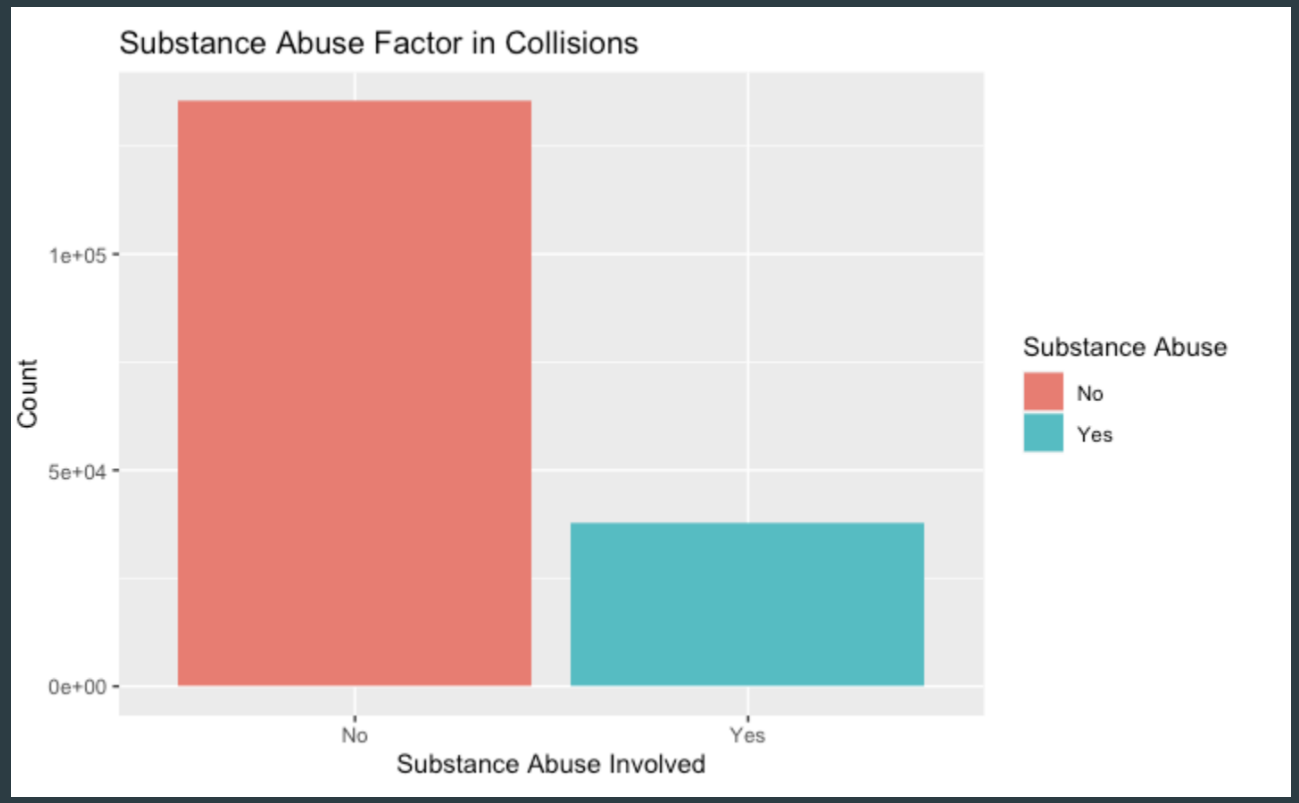


Figure 2: Substance Abuse Involvement

- Far fewer collisions involving substance abuse.
- Potential underreporting.

Exploratory Data Analysis

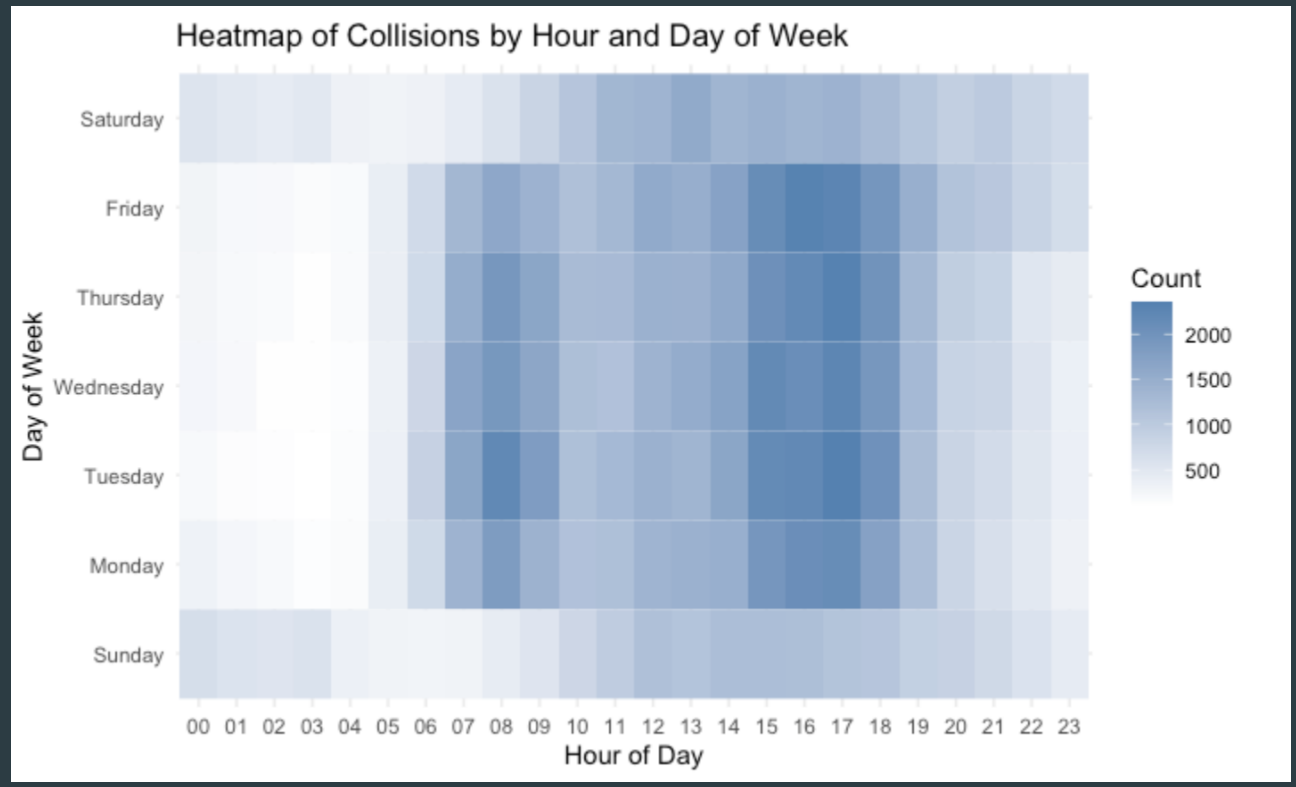


Figure 3: Heatmap of collisions by hour and day of the week.

- Darker shades during morning and evening rush hours - higher collision frequencies.
- Lighter shades during midday and late-night hours have fewer collisions - lower traffic volumes.
- Weekday vs. Weekend Trends.

Exploratory Data Analysis

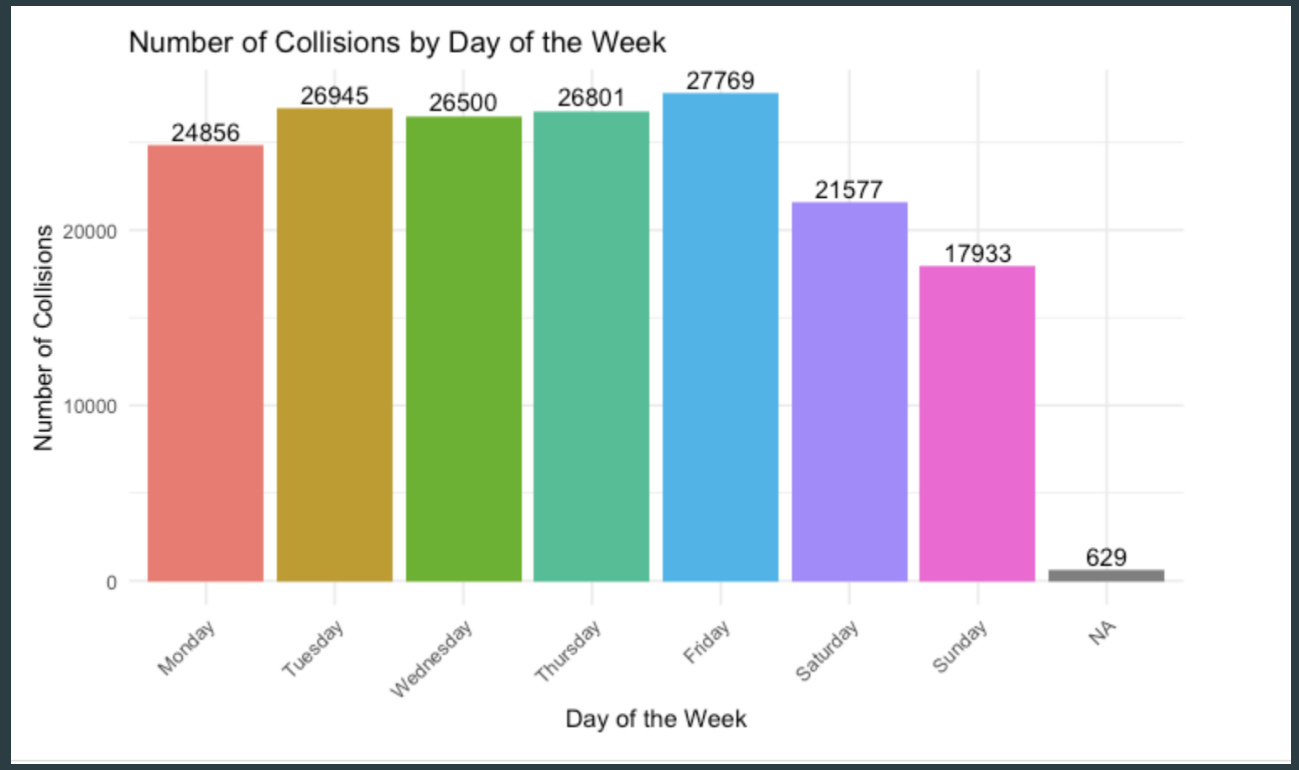


Figure 4: Number of collisions according to the days of the week.

- Friday shows the highest number of collisions, likely due to increased traffic.
- The low number of collisions on Sunday suggests less traffic due to reduced commuting.

Data Modeling - ARIMA

- Implemented ARIMA - an AutoRegressive Integrated Moving Average model.
- Popular statistical method for time series forecasting that uses past data points to predict future ones.
- Leverages the inherent time series structure of traffic collision data.
- Conducted a thorough evaluation of the 30-day forecast's accuracy to ascertain the predictive reliability of the model.

Data Modeling – Regression Analysis

- Implemented Multiple Linear Regression.
- Forecasts by utilizing the relationship between multiple independent and the dependent variables (injury severity).
- Recognizes and leverages the inherent structure of traffic collision data to predict injury severity.
- Performed factor encoding, predictor (p-values), the overall fit of the model (R-squared) and added interaction and polynomial terms to capture more complex relationships between variables.
- Conducted a thorough evaluation of the model's predictive performance by comparing the actual and predicted values of injury severity.

Results - ARIMA

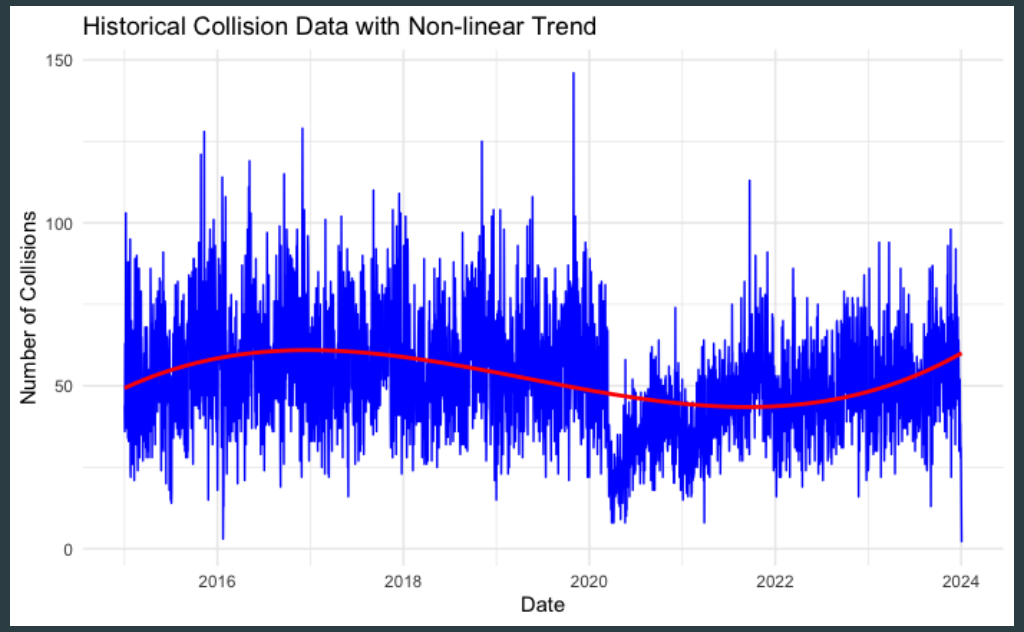


Figure 5: Historical data on traffic collisions

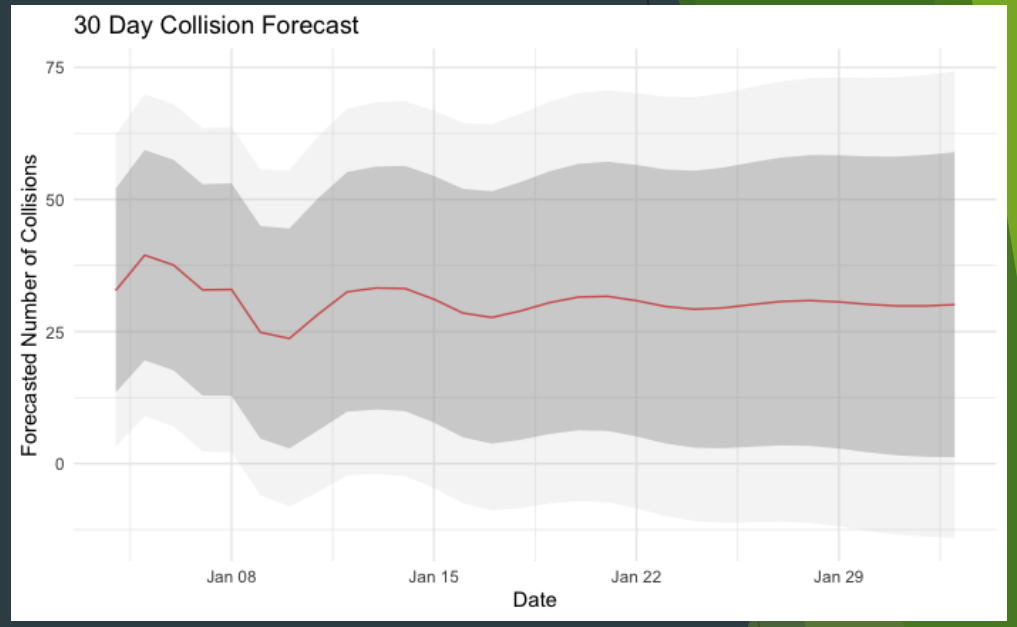
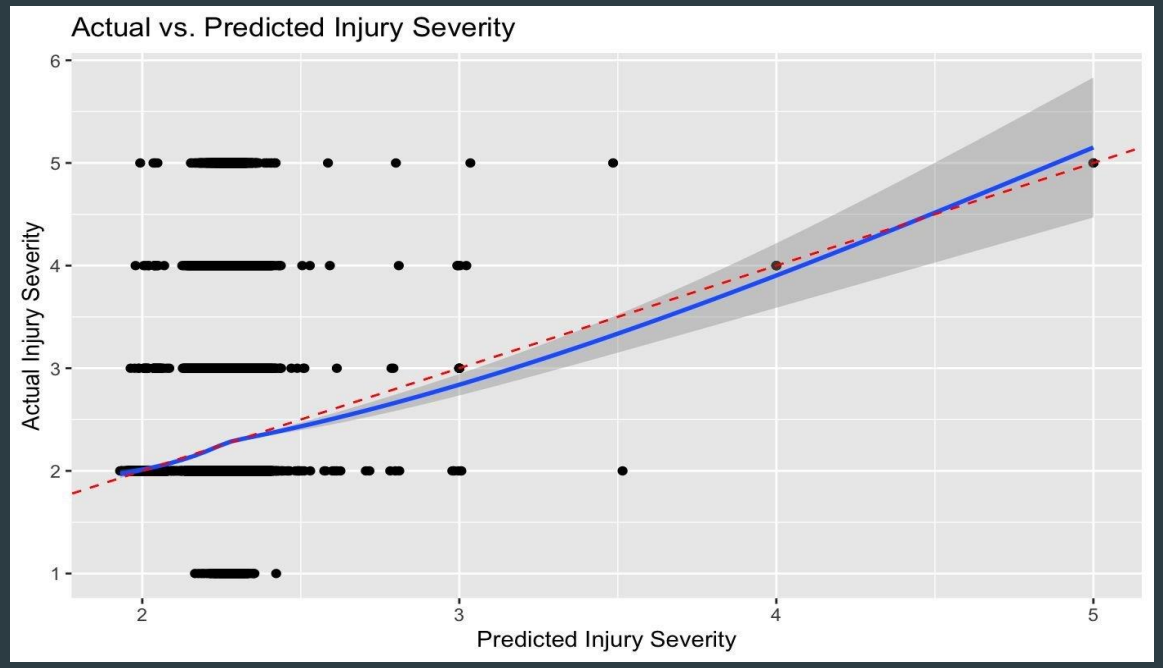


Figure 6: 30 Day Collision Forecast

- The trend line suggests variability- notable decline in the frequency of collisions around the year 2020, followed by a subsequent rise.
- 30-day forecast - relatively steady trend without significant peaks or troughs over the next month.

Results – Regression Analysis




```
Residual standard error: 0.6164 on 172236 degrees of freedom
Multiple R-squared:  0.008309,  Adjusted R-squared:  0.00748
F-statistic: 10.02 on 144 and 172236 DF,  p-value: < 2.2e-16

[1] "Pearson correlation:  0.0911520163034978"
```

Figure 7: Actual vs. Predicted Injury Severity

- The regression model used here is a statistical tool that predicts the severity of injuries in accidents.
- Based on various factors such as weather, surface condition, light, vehicle make, and model.
- Helps in understanding the relationship between these factors and injury severity.
- Indicates a weak positive linear relationship between the actual and predicted values.

Conclusion

- The ARIMA model 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.
- Regression analysis 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.

Future Work

- Advanced modeling techniques.
- Integrate additional datasets to enrich the model inputs.
- Incorporating real-time traffic and weather data into the forecasting model.

References

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