Traffic Collisions Analysis in Maryland State

Stanislav Liashkov
University of Colorado Boulder
Stanislav.Liashkov@colorado.edu

December, 2024

ABSTRACT

Every day, thousands of traffic collisions occur across the United States, many resulting in fatalities or serious injuries. This project focuses on analyzing traffic collision data from Maryland (US) spanning 2015 to 2024 to uncover critical insights into patterns, causes, and risk factors associated with severe collisions. Using advanced data analysis techniques, the project aims to identify the key contributors to serious injuries and significant damage. By doing so, it seeks to inform data-driven policy recommendations, enhance road safety measures, and reduce the overall impact of traffic accidents. The findings could play a crucial role in shaping preventive strategies and improving transportation safety systems.

1. Intro

This study examines traffic collision data from Maryland, spanning 2015 to 2024, sourced from the Automated Crash Reporting System (ACRS). The dataset, comprising approximately 107,000 records, details incidents, drivers, and non-motorists, as reported by local and state police agencies. While comprehensive, the data includes preliminary reports that may be subject to updates and may contain unverified or incomplete information. This dataset is fully open and can be found on data.gov website - incidents dataset link [1]

The analysis focuses on uncovering patterns, causes, and risk factors associated with serious injuries and fatalities. By addressing the inherent limitations of the dataset, the study aims to provide actionable insights into road safety and inform data-driven strategies for reducing traffic-related harm.

2. Related work

Dezman et al. (2016) [2] analyzed traffic collisions in Baltimore from 2009 to 2013, using ARIMA [3] modeling to identify temporal patterns and spatial autocorrelation techniques, such as Moran's I [4] and LISA [5] clustering, to pinpoint crash hotspots. Their study revealed that most collisions occurred in high-density urban areas and at intersections with significant pedestrian activity, with distracted driving emerging as the dominant risk factor. Their findings underscore the importance of addressing modifiable factors, such as road design and driver behavior, to enhance safety. Our geospatial analysis section builds on this work, applying similar methodologies to Maryland's collision data from 2015 to 2024. By reproducing their observations on new data, we aim to validate prior findings and uncover additional insights into statewide traffic collision patterns

3. Proposed work

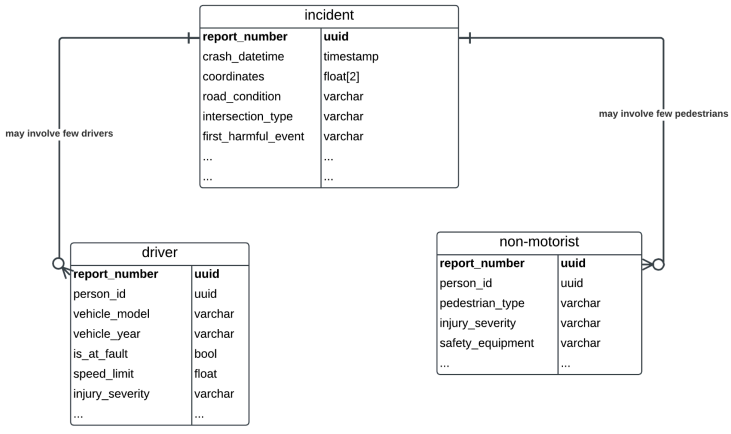
3.1 Data

The dataset consists of three interconnected tables: incidents, drivers, and non-motorists. The incidents table provides general information about each collision, including time, location, road conditions etc. Incidents may involve multiple drivers, detailed in the drivers table, and occasionally non-motorists, such as pedestrians or cyclists, described in the non-motorists table. This structure enables a comprehensive analysis of collisions, considering all parties involved.

Table 1. metadata of tables in Dataset (after normalization)

Table name	Number of rows	Number of columns	Entity
incidents	106k	17	collision
drivers	180k	40	driver
non-motorist	6k	14	pedestrian /cyclist

Figure 1. Entities relationship diagram for dataset



3.2 Goals of Analysis

This analysis is going to have 4 sections and each section focuses on a specific group of related questions that require a certain analytical approach to answer (e.g. geospatial analysis or statistical tests for different groups)

3.2.1 Time-Series Analysis Section

This section examines temporal patterns in collision data, analyzing both minor collisions and those resulting in serious injuries. By exploring the distribution of incidents across time frames such as time of day, day of the week, and season—we identify periods of heightened collision frequency. Additionally, long-term trends in both minor and serious-injury collisions are analyzed to uncover potential temporal risk factors.

3.2.2 Collision Circumstances Analysis Section

This section is dedicated to analysis of various factors affecting the traffic incident such as road condition, weather, driver distraction reasons, substance abuse, intersection type, vehicle body type etc. This section aims to uncover both the most common and most dangerous combination of factors and draw conclusions on how to increase your safety level by avoiding common risk factors.

3.2.3 Geolocations of incidents Analysis Section

This section examines the geographical distribution of collisions, focusing on variations in serious-injury collision rates across Maryland. Spatial patterns are analyzed to identify high-risk areas and determine the most dangerous locations.

3.3 Data processing method

Handling multiple related entities introduces complexities during data processing and aggregation, often requiring extensive data manipulation code to extract the desired insights. To streamline this process and simplify data grouping and joining, DuckDB [6], an embeddable OLAP [7] analytical database, is utilized as a temporary storage and query engine. DuckDB offers a powerful and flexible interface through SQL [8], enabling efficient and intuitive data processing for the analysis.

3.4 Analysis methodology and hypothesis testing

Identifying key risk factors and recommendations to mitigate these risks imply that we need to compare different groups that differ by some factor (e.g. groups of collisions with different road conditions or different weather) to identify whether the groups are different with respect to serious-injury rate of collisions.

Methods of comparison:

3.4.1 Data visualization with confidence interval

Visualization techniques, such as bar plots, line graphs, and box plots, are employed to represent differences in collision rates visually. Confidence intervals are included to highlight the variability and reliability of the observed rates, enabling a clearer understanding of whether differences between groups are statistically significant.This method offers greater informativeness by providing visual representations through plots, which facilitate intuitive understanding of the data. However, caution must be exercised when interpreting conclusions based solely on visualizations

3.4.2 Hypotheses testing

This method requires setup of hypotheses to test and relies on specific statistical tests for group comparison. This work makes use of 2 statistical tests: ANOVA [9] and Kruskal–Wallis [10] test.

While ANOVA assumes homogeneity of variances and normality of the target variable, the Kruskal-Wallis test, being a non-parametric method, does not rely on these specific assumptions, making it suitable for analyzing data that deviate from these conditions.

Data visualization is first used to examine distributions and group differences. If visualizations are inconclusive regarding significance, statistical tests are applied. This involves defining hypotheses, checking ANOVA assumptions, and selecting the appropriate test, either ANOVA or Kruskal-Wallis, based on the data.

4. Evaluation

This section will describe how we evaluate this research and what we mean by “success”. Will be updated later

5. Limitations

This section will be updated later.

6. Discussion

This section will be updated later.

7. Conclusion

This section will be updated later.

8. REFERENCES

[1] Montgomery County, MD - Maryland State Incidents Dataset from 2015 to 2024

https://data.montgomerycountymd.gov/Public-Safety/Crash-Reporting-Incidents-Data/bhju-22kf/about_data.

[2] Z. Dezman et al. , 2016, Hotspots and causes of motor vehicle crashes in Baltimore, Maryland: A geospatial analysis of five years of police crash and census data

(University of Maryland School of Medicine, Baltimore, Maryland, USA; State University of Maringá, Maringá, Brazil; Duke University, Durham, North Carolina, USA)

https://pmc.ncbi.nlm.nih.gov/articles/PMC5572144/#R23

[3] Ratnadip Adhikari, R. K. Agrawal, An Introductory Study on Time Series Modeling and Forecasting

https://arxiv.org/pdf/1302.6613

[4] Wikipedia, Moran’s I

https://en.wikipedia.org/wiki/Moran's_Ic

[5] Luc Anselin, 2020, Local Spatial Autocorrelation (LISA)

https://geodacenter.github.io/workbook/6a_local_auto/lab6a.html

[6] Mark Raasveldt, Hannes Mühleisen, 2019. DuckDB: an Embeddable Analytical Database (CWI, Amsterdam)

https://ir.cwi.nl/pub/28800/28800.pdf

[7] Wikipedia - Online Analytical Processing (OLAP)
https://en.wikipedia.org/wiki/Online_analytical_processing

[8] Wikipedia - Structured Query Language (SQL)
https://en.wikipedia.org/wiki/SQL

[9] Tae Kyun Kim, 2017. Understanding one-way ANOVA using conceptual figures (Department of Anesthesia and Pain Medicine, Pusan National University Yangsan Hospital and School of Medicine)
https://pmc.ncbi.nlm.nih.gov/articles/PMC5296382/pdf/kjae-70-22.pdf

[10] Elif F. Acar, Lei Sun, 2012. A Generalized Kruskal-Wallis Test Incorporating Group Uncertainty with Application to Genetic Association Studies (Department of Statistics, University of Toronto)
https://arxiv.org/pdf/1205.0534