My title*

My subtitle if needed

Kevin Roe

December 2, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

Table of contents

*Code and data are available at: https://github.com/Kanghyunroe/traffic_collisions/tree/main.

1 Introduction

Automobile fatalities are a leading cause of death worldwide, posing a public health and safety concern for urban cities. For example, in 2024, thirty people have died on Toronto's roadways so far, which is a 20% increase from last year (insert citation, CBC). However, new headlines, such as CBCs, highlight that Motor Vehicle Collisions (MVC) tend to either be generalized as a summary statistic or typically fatal events are over analyzed to the extent that environmental factors surrounding the crash are ignored (CLARIFY). Moreover, general environmental factors such as the time of the crash or if a bicycle was involved in the accident are important to understand what increases the likelihood of a fatality occurring in an MVC. The use of statistical modeling on increasingly available vehicle collision data presents an opportunity to develop a nuanced understanding of what factors increases the likelihood for a fatality to occur in an accident. This paper uses the Toronto Police Service's Annual Statistical Report from Open Data Toronto to analyze what factors are most responsible in predicting if a fatality occurs in a MVC.

The estimand of interest is the log-adjusted probability of a fatality occuring in a MVC. Specifically, we aim to quanitfy how specific environmental factors increase or decrease the likelihood of a fatality. By applying inferential analysis through Bayesian linear models, we assess not only the magnitude of these effects, but their underlying uncertainties (EDIT).

what was found (FINISH PARAGRAPH)

The paper is not only important from a public health perspective, but the paper also has policy development implications. Road safety and reducing fatal MVCs are a critical agenda

item of any municipal government. The paper informs what factors increase the likelihood of death in an MVC, which informs policymakers' focus for relevant policy design.

The remainder of this paper is structured as follows: Section 2 describes the dataset and methodology and **?@sec-model** exhibits the use of inferential models. **?@sec-results** presents the results of the analysis, detailing the observed relationships between likelihood of death and various circumstantial factors. **?@sec-discussion** discusses the broader implications and limitations of our findings.@sec-appendix presents a detailed idealized methodology to improve data collection, and additional model summary and diagnostic information.

2 Data

2.1 Overview

This dataset, "Police Annual Statistical Report - Traffic Collisions", was published and refreshed on October 21st, 2024, by the Toronto Police Service [insert citation]. The Toronto Police Service publishes various datasets on public safety and crime to inform the public about safety issues (**annual_statistics_report?**). Data on traffic collisions is included in the Toronto Police Service's Annual Statistical Report, which also covers reported crimes, search of persons, firearms, and the Police Service's budget (**annual_statistics_report?**). The data is collected using historical Motor Vehicle Collisions and classifies them into the following categories: * Property Damage (PD) Collisions * Fail to Remain (FTR) Collisions, or commonly known as hit-and-run accidents * Injury Collisions * Fatalities

Following the Municipal Freedom of Information and Protection of Privacy Act, the Toronto Police Service ensures to protect the privacy of individuals involved in the reported crimes when publishing the data. The dataset is updated annually, is open data, and can be used if an attribution statement **?@sec-appendix-attribution** and is properly cited (**tphlicense?**). Each entry in the dataset represents a singular vehicular accident and records all MVCs from 2015.

There is an alternative dataset from the Toronto Police Service called "Motor Vehicle Collisions involving Killed or Seriously Injured Persons" (CITE). Unlike the alternative dataset, this paper's dataset focuses on all collisions, instead of only focusing on ones where someone was either killed or seriously injured. While the alternative dataset has more explanatory variables simply because more data is collected when someone dies or is seriously injured, this paper's aims to generalize if a fatality is more likely to occur based on the general circumstances surrounding a crash, such as the time of day or if property damage occurred. Thus, we ended up not going with the alternative dataset for this paper, but there are variables in the alternative dataset that may motivate future research on this subject. (EDIT TO MAKE MORE CLEAR)

The paper uses the R programming language (R Core Team 2023) to analyze the dataset. The tidyverse package was used to simulate the dataset. Also, the tidyverse (**citetidyverse?**), arrow [CITE] and opendatatoronto (**citeopendatatoronto?**) packages were used to download the Victims of Crime dataset. Then, the tidyverse (**citetidyverse?**) package was used to clean the raw dataset and generate tests. The testthat package [CITE THIS] was used to create tests for our cleaned dataset. Rstanarm [CITE], Arrow [CITE], and bayestestR [CITE] were used to create and test the model. Finally, ggplot2 (**citeggplot2?**), tidyverse (**citetidyverse?**), knitr (**citeknitr?**) and scales (**citescales?**) packages were used to create the tables and graphs to display the results. [edit this paragraph]

2.2 Measurement

Transforming a real-life Motor Vehicle Collision to an entry in the dataset is a well-documented process by the Toronto Police Service. For insurance purposes, the Toronto Police Service requires drivers to fill out the Motor Vehicle Collision Report for any collisions that occur in Toronto (CITE). Drivers required to fill out a motor vehicle collision report if the combined damage is more than $2000, if someone is injured, if a criminal act such as a DUI occurs, or if a pedestrian is involved in the accident (CITE). These reports are retained for six years by the Toronto Police Service, with the exception of collisions resulting in a fatality, which are retained indefinitely. The form ensures documentation of collision characteristics, location, road condition, and the extent of damages, systematically recording the characteristics of each crash for further criminal investigation and data analysis.

For every collision, basic facts such as the location, time, and date of the collision is recorded through the Motor Vehicle Collision Report. Majority if not all the factors recorded in the dataset are all objective measurements regarding the specific details such as if a motorcycle was involved in the collision or if the collision resulted in property damage. All of these details are recorded in the Motor Vehicle Collision Report for all vehicular collisions and are entered into the dataset. However, while the Motor Vehicle Collision Report logs characteristics such as environmental conditions, alcohol involvement, or fatigue, the data set does not include them due to inconsistent data measurement techniques. Moreover, personal details such as the driver's age are not included to protect the driver's privacy.

2.3 Outcome variables

The main outcome variable records the number of fatalities for each MVC. According to the dataset, a fatal collision occurs when an individual's injuries from a collision results in a fatality within 30 days. Fatal collisions excludes occurences on private property, ones related to suddend eath prior to collision, such as suicide, and where the individuals has died more than 30 days after the collision. However, because we are more interested in predicting the probability that a fatality occurs than the number of fatalities, we transformed the variable that distinguishes collisions between if the collision resulted in any fatalities and those without

fatalities. In the raw dataset, if there were no fatalities, the entry was recorded as NA, but if there were fatalities, then the number of fatalities were recorded. However, we transformed the dataset that all if a fatality occurred then fatalities indicates `1` and if there were no fatalities then the fatalities column records a '0'. The distribution of the raw dataset is shown in Figure 1.

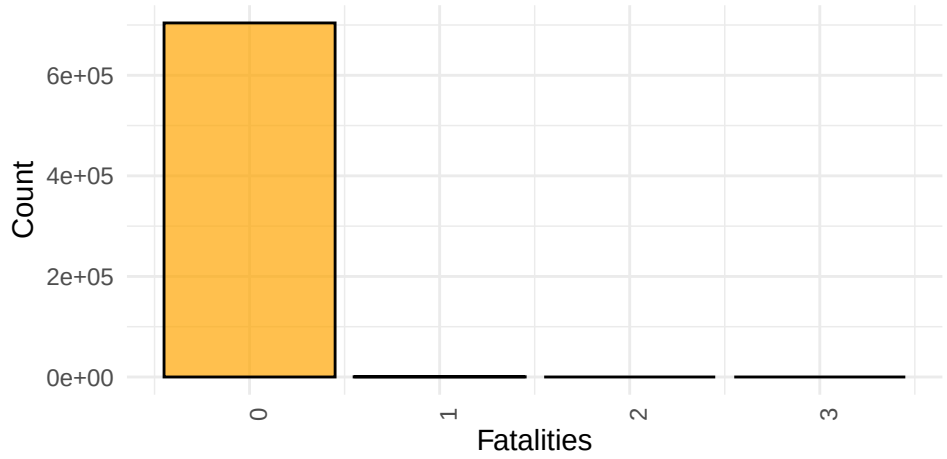

Figure 1: Distribution of Fatalities in the Raw Dataset

However, after the transformation, the outcome variable now takes on binary values and the distribution and summary statistics are shown below in **?@fig-fatalites-cleaned** and **?@tbl-fatalites**:

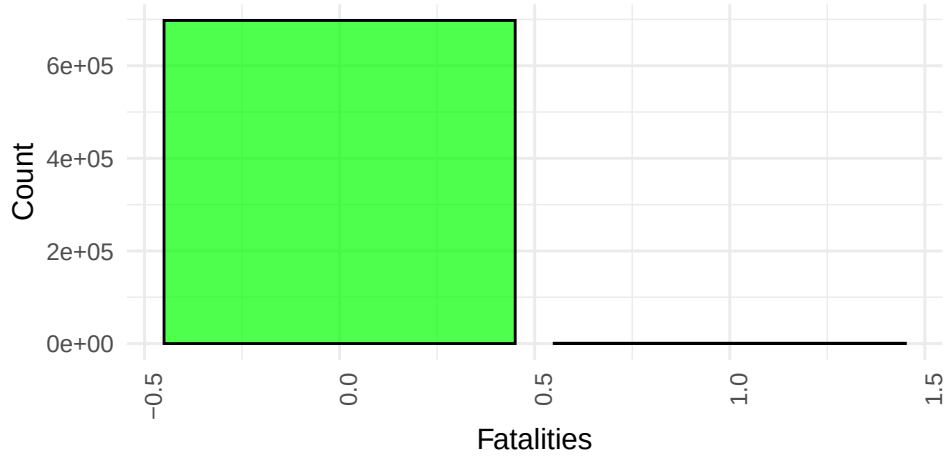

Figure 2: Distribution of Fatalities in the Cleaned Dataset

Table 1: Number of unique high-quality polling organizations

fatalities	Fatalities
0	697852
1	606

2.4 Predictor variables

2.4.1 Hour

The hour variable indicates the time at which the accident occurred using the 24 hour clock, such that a value of 0 represents 12 AM and 23 represents 11 PM. The distribution and summary statistics of the hour variable can be found in Figure 3 and Table 2, respectively. Figure 3 highlights that the majority of accidents happen from 9 AM to 7 PM, which is reasonable as these times include rush hour, where the greatest number of people are driving at the same time for work or school. However, Table 2 also shows that a greater number of fatalities occur at night, which we hypothesize is due to lack of vision or reckless driving.

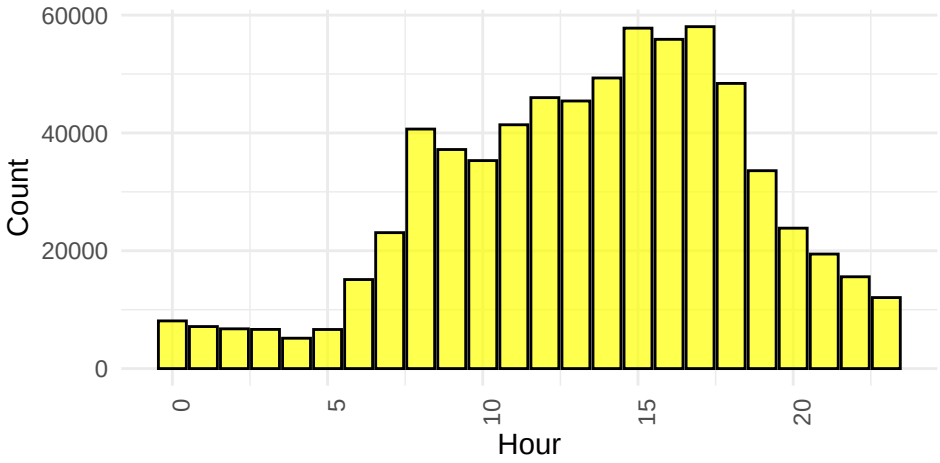

Figure 3: Distribution of the Hour Variable

Table 2: The Summary of the Hour

hour	No Fatalities	Fatality Occurred	Total
0	8067	25	8092
1	7116	15	7131
2	6734	16	6750

3	6630	21	6651
4	5142	11	5153
5	6623	10	6633
6	15083	28	15111
7	23066	10	23076
8	40638	16	40654
9	37153	30	37183
10	35278	27	35305
11	41358	26	41384
12	45955	33	45988
13	45393	25	45418
14	49293	30	49323
15	57762	24	57786
16	55857	27	55884
17	58001	33	58034
18	48357	43	48400
19	33556	34	33590
20	23794	40	23834
21	19414	26	19440
22	15555	29	15584
23	12027	27	12054

2.4.2 Type of Crash

Beyond identifying fatalities, the Motor Vehicle Collision Reports notes if the crash was one of three types: an injury collision, a fail to remain collision, and a property damage collision. A personal injury collision occur when an individual involved in a MVC suffers personal injuries. Fail to remain collisions are recorded when an individual involved in a collision fail to stop and provide their information at the scene of a collision. Property damage collisions occur when an individual has been damaged in a collision or the value of damages is less than $\2000 for all parties. The distribution and summary statistics of these three variables can be found in Figure 4 and Table 3. The results in Table 3 show that crashes that classify under these three categories usually do not lead to death.

Table 3: Breakdown of MVCs Into the Different Categories, Broken Down by Fatalities

collision_type	No Fatalities	Fatality Occurred	Total
Fail to Remain Collision	112404	0	112404
Injury Collision	94326	2	94328
Not Applicable	0	604	604

Property Damage Collision	491122	0	491122
Total	697852	606	698458

2.4.3 Automobile

The automobile variable is a indicator variable to show if a collision involved a person in an automobile. In the raw dataset, the variable was labeled as Yes, No, None or N/R (Not Recorded). We labelled Not Recorded as NA because there is no reliable way of characterizing the variable. Further, we labeled No and None as 0 and Yes as 1, where 1 represents that an automobile was involved and 0 represents that an automobile was not, such as a crash between two motorcycles. We also employed this method for the following variables: motorcycle, passenger, and pedestrian.

Figure 5 and Table 4 shows that 588 of 608 deaths happened when an automobile was involved, which is not surprising given majority of vehicles on the road are cars.

Table 4: Summary of the Automobile Variable

automobile	No Fatalities	Fatality Occurred	Total
0	3337	18	3355
1	694515	588	695103

2.4.4 Motorcycle

The motorcycle variable is another indicator variable to show that whether the collision involved a person in a motorcycle. 1 represents that a motorcycle was involved and 0 represents that a motorcycle was not involved in the crash. Figure 6 and Table 5 shows the distribution and summary of the motorcycle variable, respectively. Table 5 shows that only 75 vehicular deaths involved a motorcycle.

Table 5: Summary of the Automobile Variable

motorcycle	No Fatalities	Fatality Occurred	Total
0	693656	531	694187
1	4196	75	4271

2.4.5 Passenger

The passenger variable is an indicator variable that highlights if the collision involved a passenger in a motor vehicle. 1 represents there was a passenger and 0 shows that there was not a

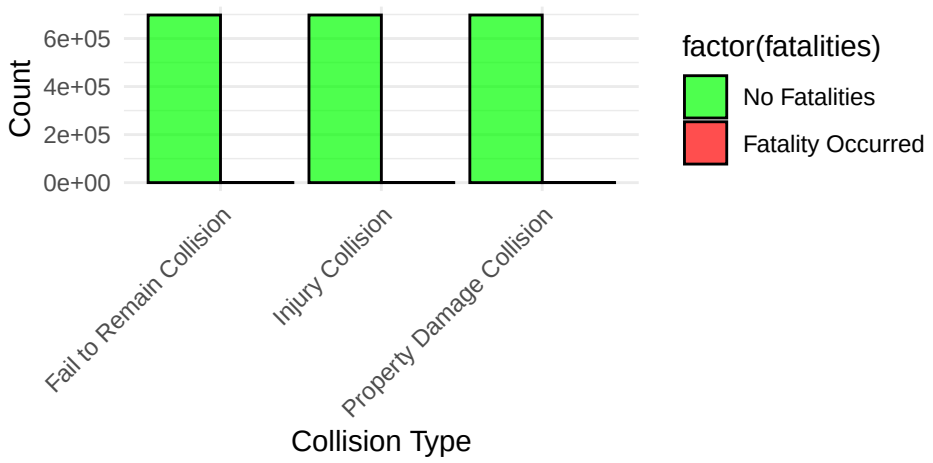

Figure 4: Breakdown of Each Collision Type by Fatalities

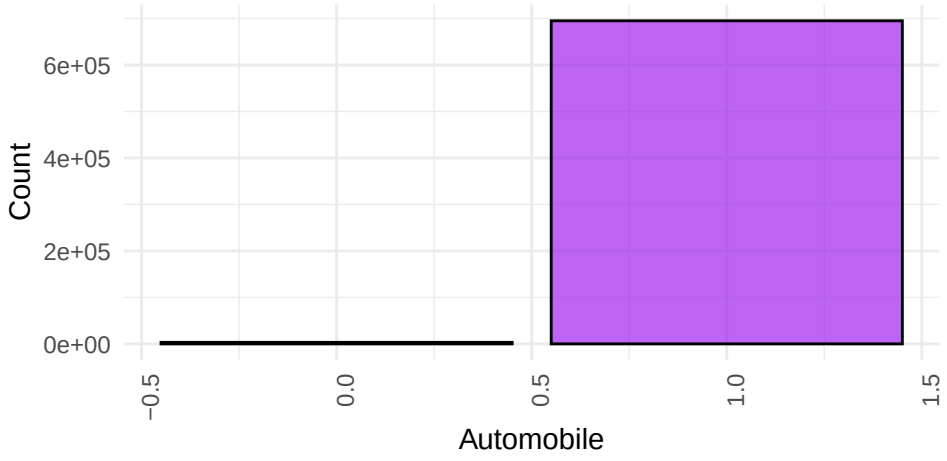

Figure 5: Distribution of the Automobile Variable

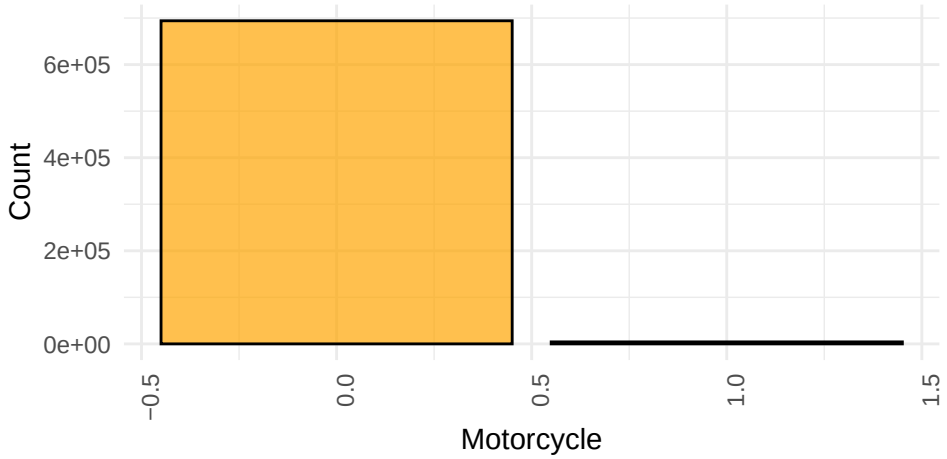

Figure 6: Distribution of the Automobile Variable

passenger involved. Figure 7 and Table 6 shows the distribution and summary of the passenger variable, respectively. Table 6 shows that 177 vehicular deaths involved a passenger.

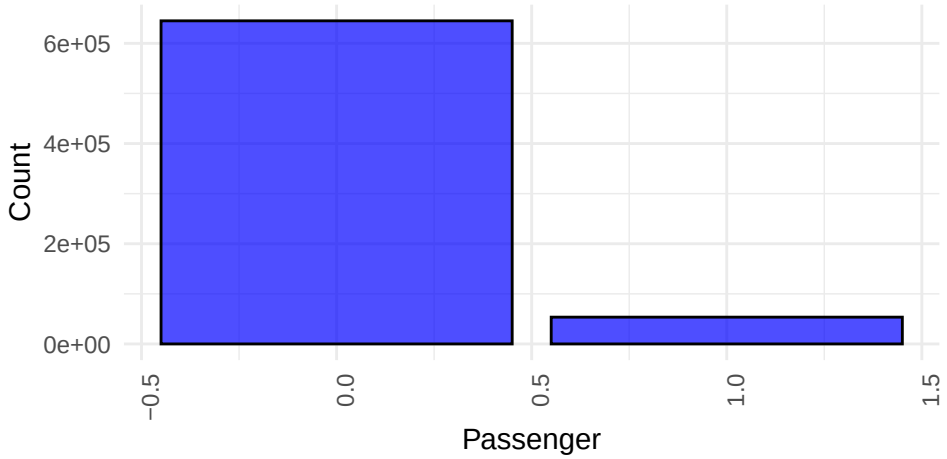

Figure 7: Distribution of the Passenger Variable

Table 6: Summary of the Passenger Variable

passenger	No Fatalities	Fatality Occurred	Total
0	644546	429	644975
1	53306	177	53483

2.4.6 Pedestrian

The pedestrian variable is an indicator variable that highlights if the collision involved a pedestrian. 1 represents a pedestrian was involved and 0 shows that there was no pedestrian. Figure 8 and Table 7 shows the distribution and summary of the pedestrian variable, respectively. Table 7 highlights that of 606 deaths, 342 deaths involved pedestrians, which is a significant percentage.

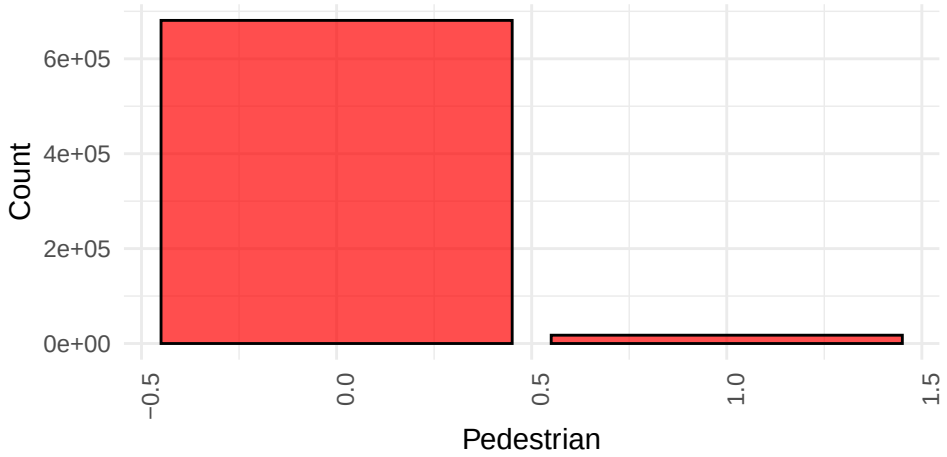

Figure 8: Distribution of the Pedestrian Variable

Table 7: Summary of the Pedestrian Variable

pedestrian	No Fatalities	Fatality Occurred	Total
0	680640	264	680904
1	17212	342	17554

2.4.7 Police Division

The Police Division variable represents the police division where the collision occured. The paper plans to include the police division variable as a general proxy for location and account if certain areas are more susceptible to crashes than others. Figure 9 and Table 8 shows the distirbution and breakdown of MVCs among police departments.Based on the Table 8, there seems to be no discernible pattern but D41 and D42 have the highest number of vehicular fatalities at 63 and 66, respectively.

Table 8: Summary of the Police Divison Variable

police_division	No Fatalities	Fatality Occurred	Total
D11	25230	24	25254
D12	23366	17	23383
D13	22295	28	22323
D14	37837	30	37867
D22	37436	55	37491
D23	34032	39	34071
D31	36370	39	36409
D32	56106	49	56155
D33	46169	39	46208
D41	47789	63	47852
D42	55916	66	55982
D43	37598	42	37640
D51	25877	35	25912
D52	31076	15	31091
D53	39128	34	39162
D55	43287	30	43317
NSA	98340	1	98341

2.4.8 Year

The year variable shows the number of MVCs per year. Figure 10 and Table 9 shows the number of MVCs over time. Looking at Figure 10 and Table 9 there is a noticeable dip in MVCs in 2020 and 2021 due to COVID-19 but MVC levels have not hit their 2019 peaks most likely due to people in Toronto not driving as much as before. However, idea of fewer drivers is a hypothesis and needs further research.

Table 9: Summary of Number of MVCs by Year

year	No Fatalities	Fatality Occurred	Total
2014	64228	51	64279
2015	66915	65	66980
2016	69219	76	69295
2017	73245	62	73307
2018	77034	66	77100
2019	81410	63	81473
2020	44511	40	44551
2021	43589	58	43647
2022	58950	48	58998

2023	67270	45	67315
2024	51481	32	51513

2.4.9 Relevant Interaction Terms

Why you have interaction between hour and all these things

3 Model

The study aims to predict the likelihood of fatality in a car crash. The model was implemeneted using the rstatarm package in R, leveraging its framework for Baysian regression modelling.

The goal of our modelling strategy is twofold. Firstly,...

Here we briefly describe the Bayesian analysis model used to investigate... Background details and diagnostics are included in Appendix B.

3.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \alpha + \beta_i + \gamma_i \tag{2}$$
$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\gamma \sim \text{Normal}(0, 2.5) \tag{5}$$
$$\sigma \sim \text{Exponential}(1) \tag{6}$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

what doe seach one

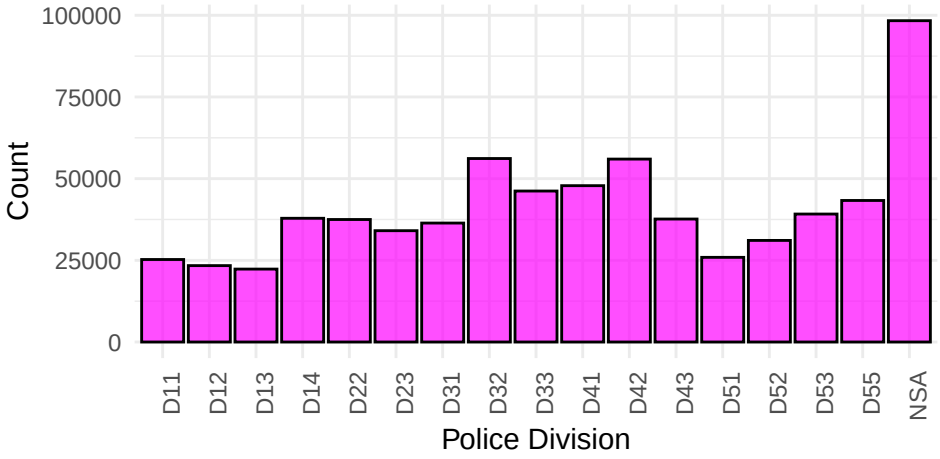

Figure 9: Distribution of the Police Division

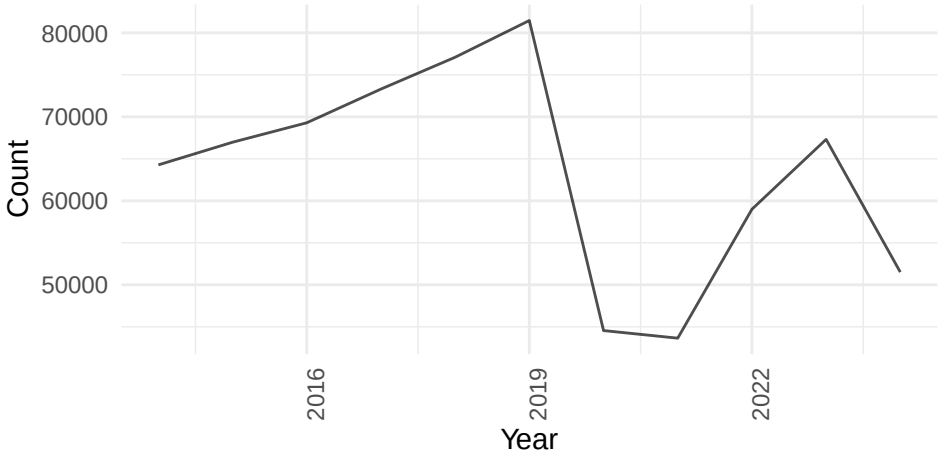

Figure 10: Number of MVCs per Year

Table 10: Explanatory models of flight time based on wing width and wing length

3.2 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular...

We can use maths by including latex between dollar signs, for instance θ.

3.3 Model assumption and limitation

We expect a positive relationship between the size of the wings and time spent aloft. In particular...

We can use maths by including latex between dollar signs, for instance θ.

3.4 Alternative model considerations

We expect a positive relationship between the size of the wings and time spent aloft. In particular...

We can use maths by including latex between dollar signs, for instance θ.

4 Results

Our results are summarized in Table 10.

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

In we implement a posterior predictive check. This shows...

In we compare the posterior with the prior. This shows...

Examining how the model fits, and is affected by, the data

B.2 Diagnostics

is a trace plot. It shows... This suggests...

is a Rhat plot. It shows... This suggests...

Checking the convergence of the MCMC algorithm

References

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.