My title*

My subtitle if needed

Kevin Roe

December 3, 2024

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

Table of contents

1 Introduction **2**

2 Data **3**
2.1 Overview . . . 3
2.2 Measurement . . . 4
2.3 Outcome variables . . . 5
2.4 Predictor variables . . . 6
2.4.1 Hour . . . 6
2.4.2 Type of Crash . . . 7
2.4.3 Automobile . . . 8
2.4.4 Motorcycle . . . 8
2.4.5 Passenger . . . 8
2.4.6 Pedestrian . . . 11
2.4.7 Police Division . . . 11
2.4.8 Year . . . 12
2.4.9 Interaction Terms . . . 13

3 Model **17**
3.1 Model Set-Up and Justification . . . 17
3.2 Model Assumptions and Limitations . . . 18

4 Results **19**

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

5 Discussion **22**
5.1 Contextualizing Role of Motorcyclists and Pedestrians 22
5.1.1 Motorcyclists: The Most Vulnerable Group 22
5.1.2 Pedestrians 22
5.1.3 Environmental Factors 23
5.2 Role of Protective Factors 23
5.3 Role of High Variation 24
5.4 Implications 25
5.5 Weaknesses and Next Steps 25

Appendix **27**
.1 Exploration of Surveys, Sampling, and Observational Data 27
.1.1 Data Collection Methodology 27
.1.2 Idealized Methodology 28
.1.3 Example Survey 28
.2 Enhancement: Model Card 29
.2.1 Model Overview 29
.2.2 Model Details 29
.2.3 Model Coefficients 29
.2.4 Model Performance Metrics 32
.2.5 Training Time 32
.2.6 Conclusion 33
.3 Data Attribution Statement 34
.4 Model Diagnostics 34

References **36**

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).

Our analysis identifies key predictors of vehicular accident fatalities, emphasizing heightened risks for motorcyclists (coefficient: 17.036) and pedestrians (coefficient: 6.809), especially during rush hours (hours 16 and 17) and late evenings (11:00 PM), where motorcycle collisions show a fatality likelihood of around 66%. Collision type variables, such as injury and property damage collisions, significantly lower fatality risks, while police division and year have minimal impact. However, variability in estimates, particularly for motorcyclist and automobile predictors, highlights uncertainty in their precise effects, underscoring the need for targeted interventions during high-risk times and for vulnerable populations.

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 Section .3 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. In addition, each entry in the alternative dataset represents a killed or injured person, meaning that there exists the possibility for duplicate entries. However, each entry in the selected dataset is a unique Motor Vehicle Collision, making it more suitable for the scope of this paper 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.

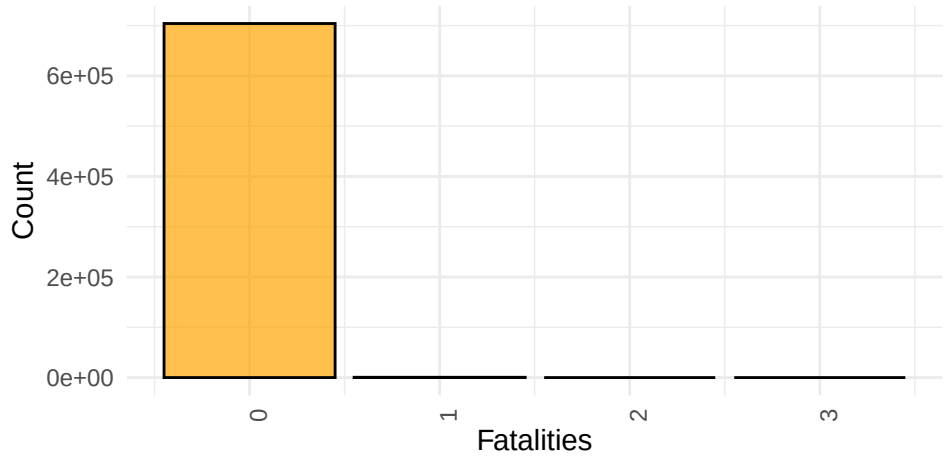

Figure 1: Distribution of Fatalities in the Raw Dataset

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**:

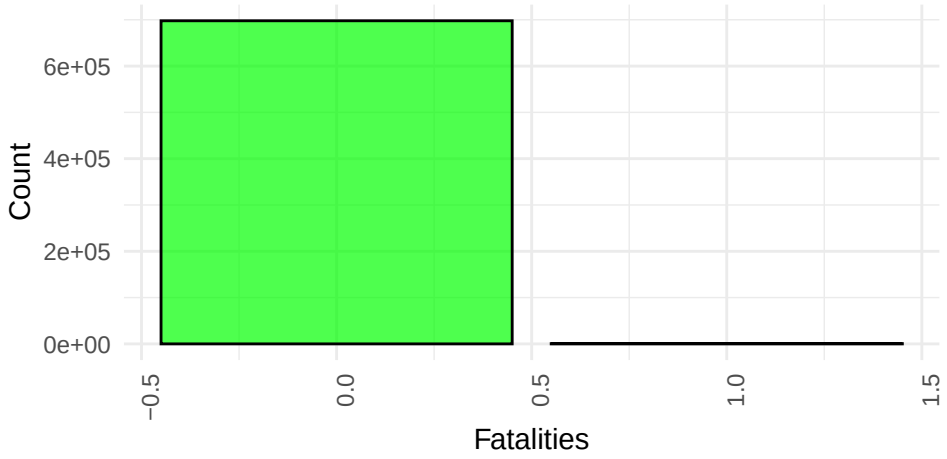

Figure 2: Distribution of Fatalities in the Cleaned Dataset

Table 1: Number of unique high-quality polling organizations

mean	median	min	max	sd	n
0	0	0	1	0.03	698458

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.

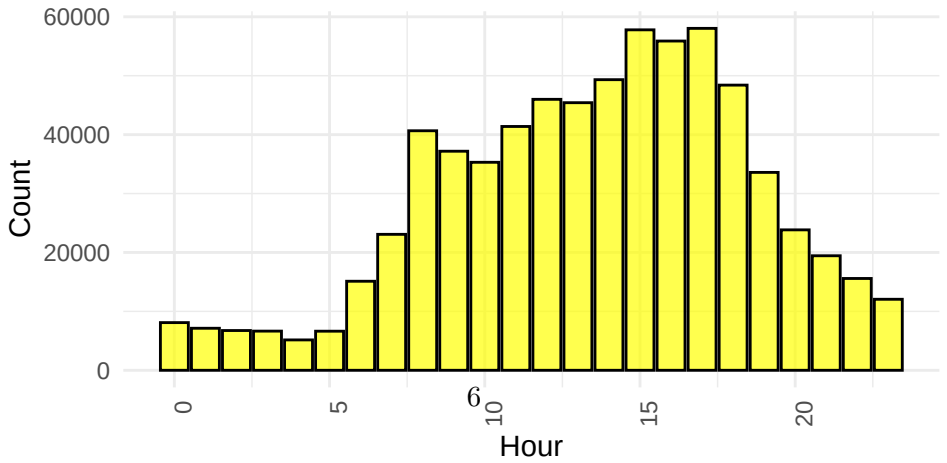

Figure 3: Distribution of the Hour Variable

Table 2: The Summary of the Hour

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

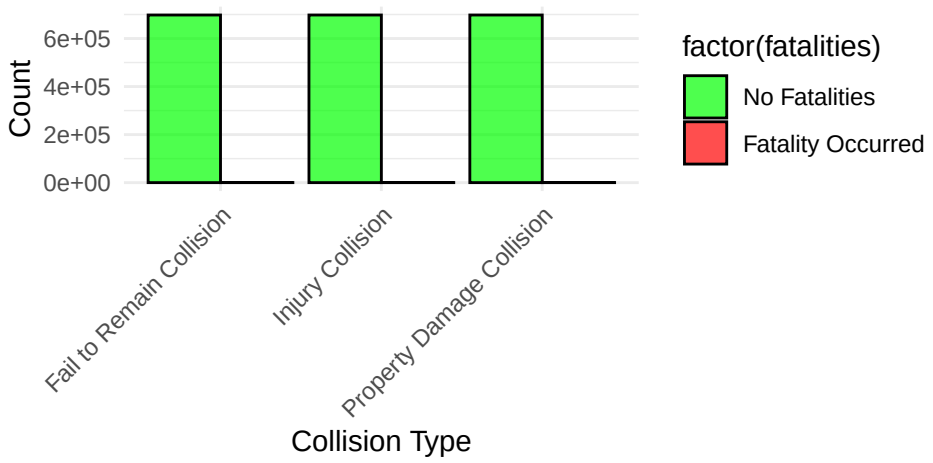

Figure 4: Breakdown of Each Collision Type by Fatalities

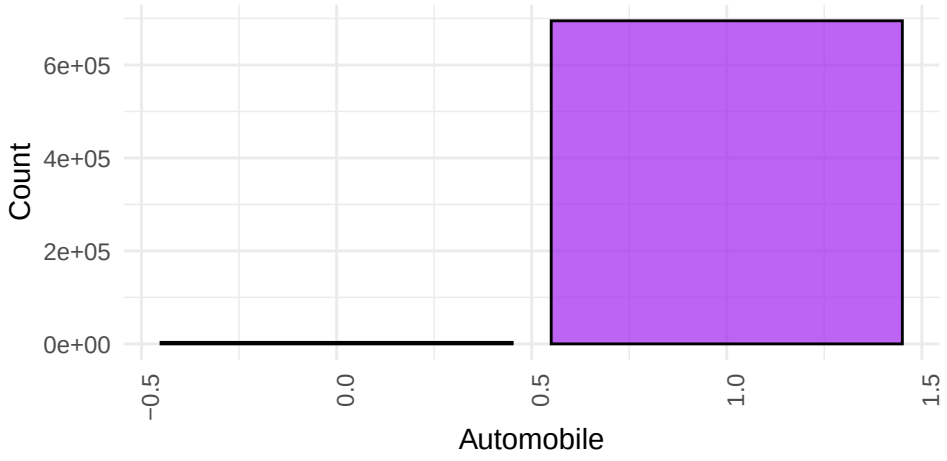

Figure 5: Distribution of the Automobile Variable

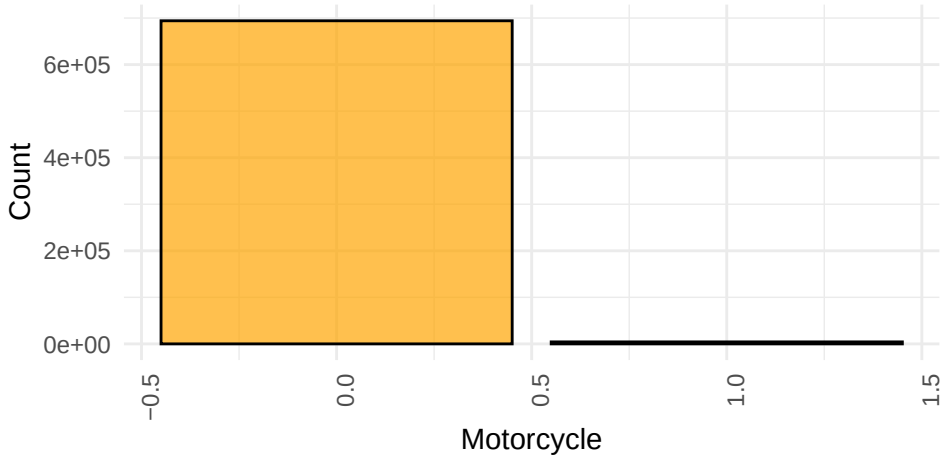

Figure 6: Distribution of the Automobile Variable

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.

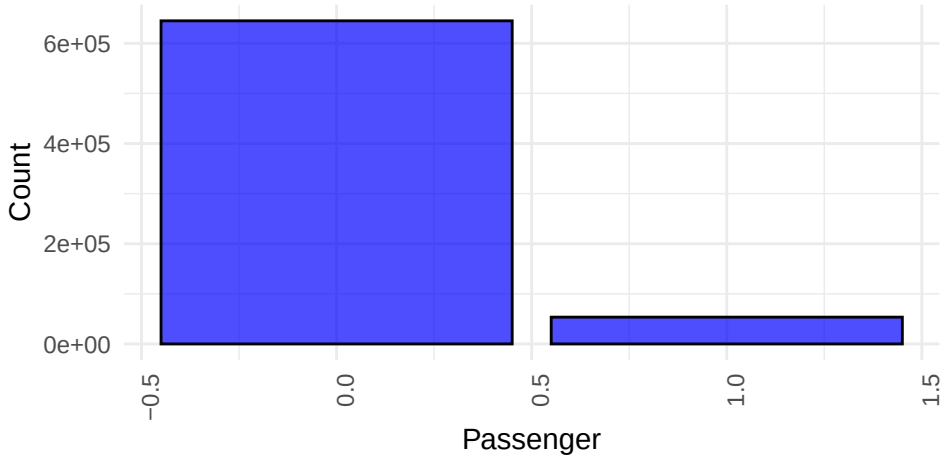

Figure 7: Distribution of the Passenger Variable

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.

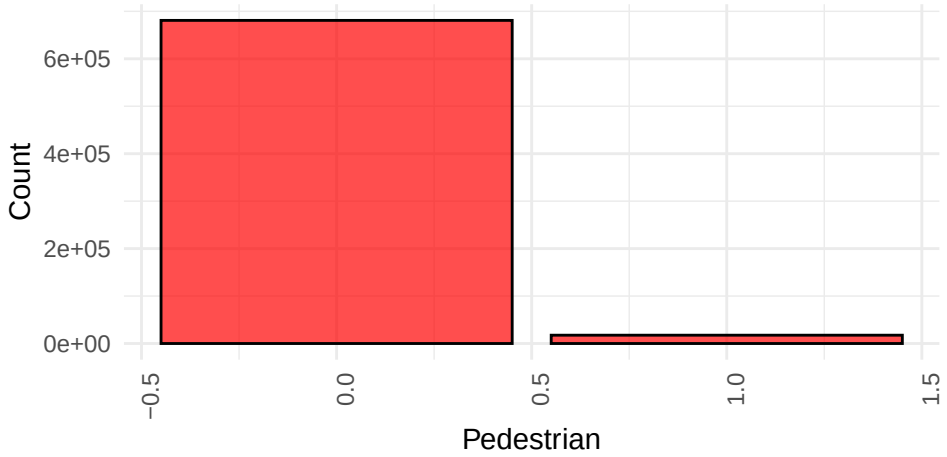

Figure 8: Distribution of the Pedestrian Variable

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 Interaction Terms

Due to the compounded effects of the time of hour on motorcyclists and pedestrians in a motor vehicle collision, we controlled for this effect using an interaction term:

2.4.9.1 Hour and Motorcycle

Figure 11 and Table 10 shows the visual representation and summary of the interaction term, respectively. Figure 11 shows that a significant number of deaths come in the evening from 2 PM to 11 PM.

Table 10: Summary of Fatalities by Hour for Motorcyclists

hour	Motorcyclist Fatalities	Total Motorcyclists
0	3	61
1	2	40
2	2	37
3	2	43
4	2	25
5	0	21
6	1	82
7	1	113
8	1	191
9	0	212
10	3	178
11	0	208
12	1	251
13	5	258
14	4	304
15	4	319
16	2	355
17	4	377
18	8	323
19	4	233
20	5	199
21	4	182

22	12	160
23	5	99

2.4.9.2 Hour and Pedestrian

Figure 12 and Table 11 shows the distribution and summary of the interaction variable, respectively. Figure 12, though showing high variability, shows that a lot more pedestrian deaths also happen in the evening.

Table 11: Summary of Fatalities by Hour for Motorcyclists and Pedestrians

hour	Pedestrian Fatalities	Total Pedestrians
0	12	259
1	9	186
2	3	156
3	6	96
4	6	61
5	5	146
6	22	522
7	8	622
8	9	1012
9	20	867
10	16	784
11	19	869
12	22	902
13	10	943
14	15	1013
15	14	1224
16	15	1069
17	22	1360
18	20	1422
19	22	1179
20	29	924
21	15	857
22	11	623
23	12	458

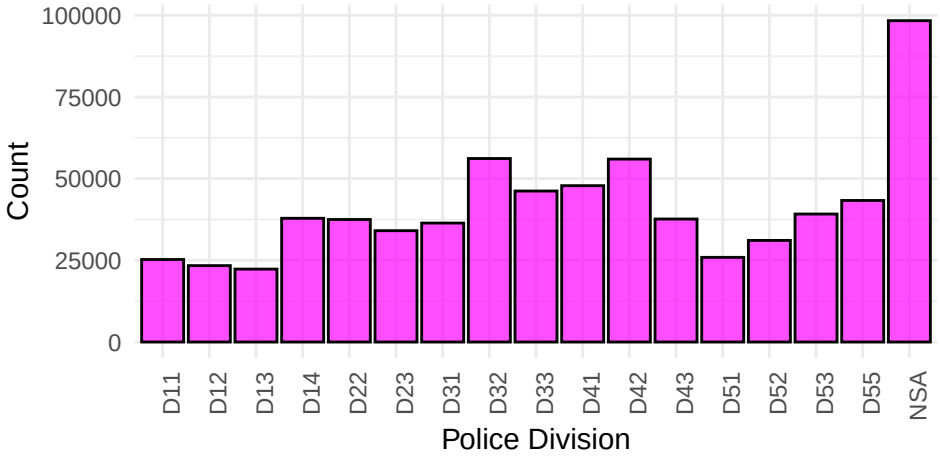

Figure 9: Distribution of the Police Division

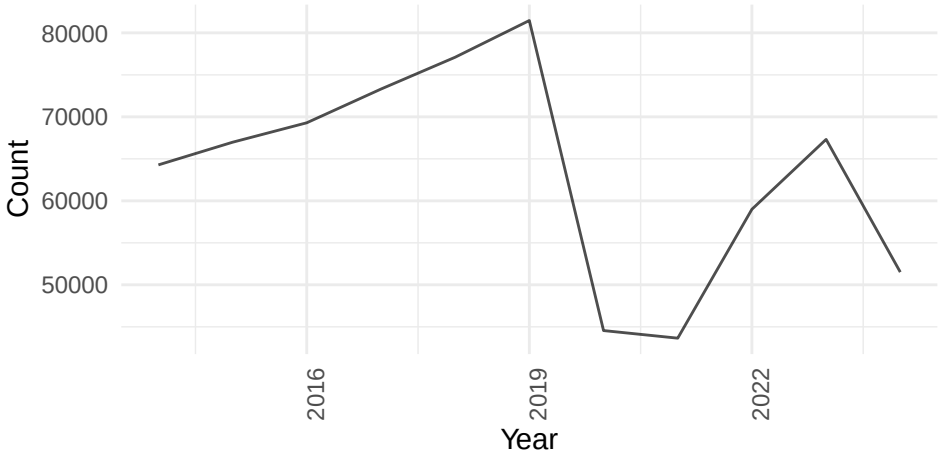

Figure 10: Number of MVCs per Year

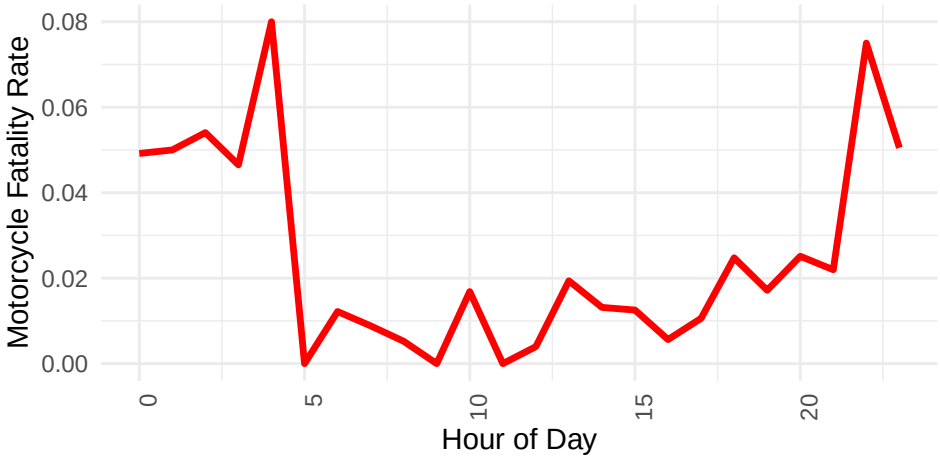

Figure 11: Fatality Likelihood by Hour for Motorcyclists

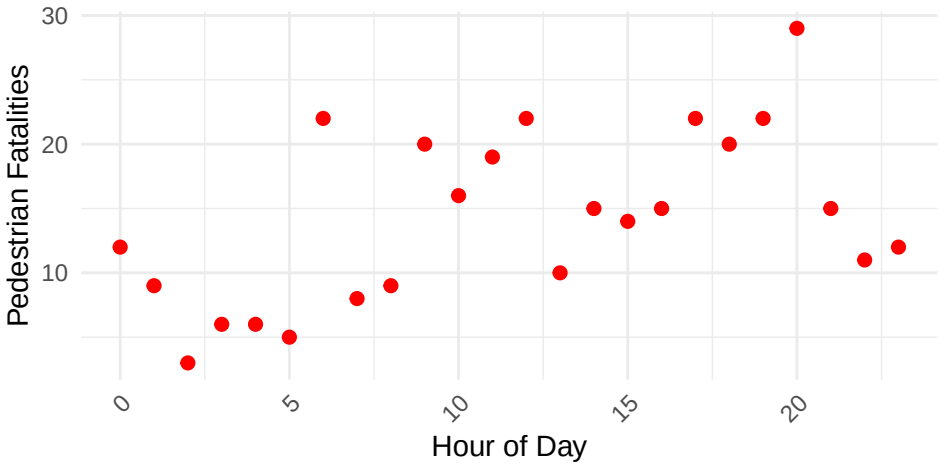

Figure 12: Fatality Likelihood by Hour for Pedestrians

3 Model

3.1 Model Set-Up and Justification

For our analysis, we employ a Bayesian Logistic Linear Model to forecast the likelihood of a fatality in a car crash. This approach allows us to capture known variations between years, police divisions, and the characteristics of a crash, such as the involvement of a motorcycle or if someone was injured.

The first step in the process involved selecting a reliable dataset for model development. We utilized high-quality vehicle collision data gathered by the Toronto Police Service. We first excluded all cases in our predictor variables that had null values. In addition, we transformed fatalities into a binary response variable. Moreover, due to the size of the dataset, we conducted a stratified random sampling with equal allocation across the levels of the fatalities variable to ensure each level of the fatalities variable are adequately represented in the model to help the model learn meaningful patterns for that class. Let Y_i represent the fatality outcome for observation i, where $Y_i = 1$ indicates a fatality occurred and $Y_i = 0$ indicates no fatality. The probability of fatality is modeled through a Bernoulli distribution:

$$Y_i = \begin{cases} 0 & \text{if no fatality occurred (original classification: N/A)} \\ 1 & \text{if any fatality occurred (original classifications: 1, 2, or 3 fatalities)} \end{cases}$$

$$S_i \sim \text{Bernoulli}(p_i) \tag{1}$$

$$
\begin{aligned}
\text{logit}(p_i) = \beta_0 &+ \beta_1 x_{\text{hour}_i} + \beta_2 x_{\text{InjuryCollision}_i} + \beta_3 x_{\text{FTRCollision}_i} + \beta_4 x_{\text{PDCollision}_i} \\
&+ \beta_5 x_{\text{automobile}_i} + \beta_6 x_{\text{motorcycle}_i} + \beta_7 x_{\text{passenger}_i} + \beta_8 x_{\text{pedestrian}_i} \\
&+ \beta_9 x_{\text{PoliceDivision}_i} + \beta_{10} x_{\text{year}_{im}} \\
&+ \beta_{11} x_{\text{hour}_i} \cdot x_{\text{motorcycle}_i} + \beta_{12} x_{\text{hour}_i} \cdot x_{\text{pedestrian}_i} + \epsilon_i
\end{aligned} \tag{2}
$$

In equation 2, each β represents a coefficient determined through regression analysis. The variables chosen for the model are hour, the different types of collisions, automobile, motorcycle, passenger, pedestrian, year, and police division. Each predictor variable was selected based on its significance in vehicle fatality prediction analysis. Temporal variables such as hour and year are introduced as fixed effects to account for biases across time without introducing unnecessary complexity. Similarly, location introduces biases as some neighborhoods may be more susceptible to vehicle accidents than others, prompting the inclusion of police department as a fixed effect and a proxy for location. Furthermore, depending on the time of day, motorcyclists and pedestrians are more susceptible to fatal vehicle accidents than during the day. Thus, β_{11} and β_{12} represents the coefficient of an interaction term between motorcycle and pedestrian with the categorical variable hour, respectively, to account for potential variations in risk patterns associated with different times of the day. These interaction terms allow the model to capture the increased vulnerability of motorcyclists and pedestrians during specific hours, such as nighttime or early morning, when visibility and traffic conditions may differ significantly, thereby improving the precision and interpretability of the predictions. In addition, any variables, especially the various interaction terms, will be omitted if they display high multicolinearity or insignificance. Finally, ϵ_i is the Gaussian-distributed error term, accounting for residual variation in the model.

To enhance the model, Bayesian priors were applied, introducing regularization and incorporating plausible ranges grounded. For the coefficient priors β, a normal distribution with a mean of 0 and a scale of 2.5 (autoscaled) was chosen to provide flexibility while mitigating overfitting. Similarly, the intercept uses a normal prior with a mean of 0 and a scale of 2.5 to stabilize model estimates. For the error term (sigma), an exponential prior with a rate of 1 was selected to constrain the residuals, aligning with Gaussian assumptions.

The model was implemented in R (R Core Team 2023) using the `rstanarm` package, which offers an accessible interface for Bayesian generalized linear models (GLMs), allowing specification of priors and customization of model parameters.

3.2 Model Assumptions and Limitations

The model assumes independence of observations, meaning that each observation's outcome is not influenced by others. However, this assumption may be partially violated in the context of vehicle accidents, as incidents occurring within close temporal or spatial proximity may

share common influencing factors, such as weather, road conditions, or traffic patterns. While the inclusion of temporal and spatial variables like hour and police division aims to mitigate such dependencies, it may not fully account for potential clustering effects. As a result, the model could underestimate or overestimate the significance of some predictors, impacting the robustness of its conclusions.

Another important limitation is the potential for omitted variable bias and unmeasured confounding factors. For example, the model does not include variables like road infrastructure quality, driver impairment (e.g., alcohol or drug use), or real-time weather conditions, all of which could influence accident outcomes. Additionally, selection bias is a concern, as the dataset may disproportionately represent severe accidents reported for insurance claims while omitting less severe incidents. This could skew the model's predictions toward more severe outcomes and limit its generalizability. Addressing these limitations would require incorporating additional data sources and applying techniques to account for potential dependencies and missing information. Additionally, the model relies on reported data, which may include potential selection bias, as minor accidents are less likely to be documented.

4 Results

To assess model reliability, we examined several key metrics. Convergence metrics, such as Rhat values, were very close to 1 for all parameters, indicating strong convergence. Additionally, the effective sample size n_{eff} was high across all parameters, suggesting low autocorrelation and contributing to model stability. See more details of our model diagnostics here: (INSERT REFERENCE)

Table 12 presents the estimated coefficients for the predictors in our GLM model. These coefficients fit into the GLM equation, 2, allowing us to interpret the impact of each predictor on the likelihood of fatality in a vehicular accident. Most notably, motorcyclist (17.036) and pedestrian (6.809) collisions exhibit substantial positive associations with fatality occurrence, which is in line with expectations. Moreover, motorcyclist accidents during rush hour (hour 16 and 17) have the highest coefficients at 40.4 and 26.8, respectively, and during the late evening, such as at 11:00 PM (39.365), indicating that a collision involving a motorcycle during rush hour or late evening is around 66% likely to result in a fatality. We also find similar results with pedestrians as well. The type of collision variables all have a significant negative effect on the likelihood of death, as well.However, police division and year don't seem to have a significant impact on the likelihood of death.

Figure 13 represents the model coefficients, with error bars indicating the confidence interval for each estimate. While most variables are relatively in line, it seems that the interaction terms all have wide confidence intervals, suggesting high variability in its estimates.

Table 12: Summary for Motor Vehicle Collision Fatality Prediction Model

	(1)
(Intercept)	9.705
hour1	1.442
hour2	1.920
hour3	2.908
hour4	0.793
hour5	0.382
hour6	0.966
hour7	−1.673
hour8	−0.511
hour9	−0.581
hour10	0.032
hour11	0.727
hour12	0.855
hour13	0.178
hour14	0.332
hour15	−1.227
hour16	−0.551
hour17	−0.354
hour18	1.238
hour19	0.460
hour20	0.474
hour21	1.237
hour22	0.631
hour23	1.572
injury_collision	−34.794
fail_to_remain_collision	−29.912
property_damage_collision	−30.004
automobile	2.648
motorcycle	17.036
passenger	3.624
bicycle	3.809
pedestrian	6.832
police_divisionD12	0.846
police_divisionD13	0.075
police_divisionD14	−2.806
police_divisionD22	1.797
police_divisionD23	1.193
police_divisionD31	0.619
police_divisionD32	1.007
police_divisionD33	0.055
police_divisionD41	0.989
police_divisionD42	1.140
police_divisionD43	0.582
police_divisionD51	0.915
police_divisionD52	1.465
police_divisionD53	1.389
police_divisionD55	0.389
police_divisionNSA	−4.040
year2015	−0.500
year2016	0.329

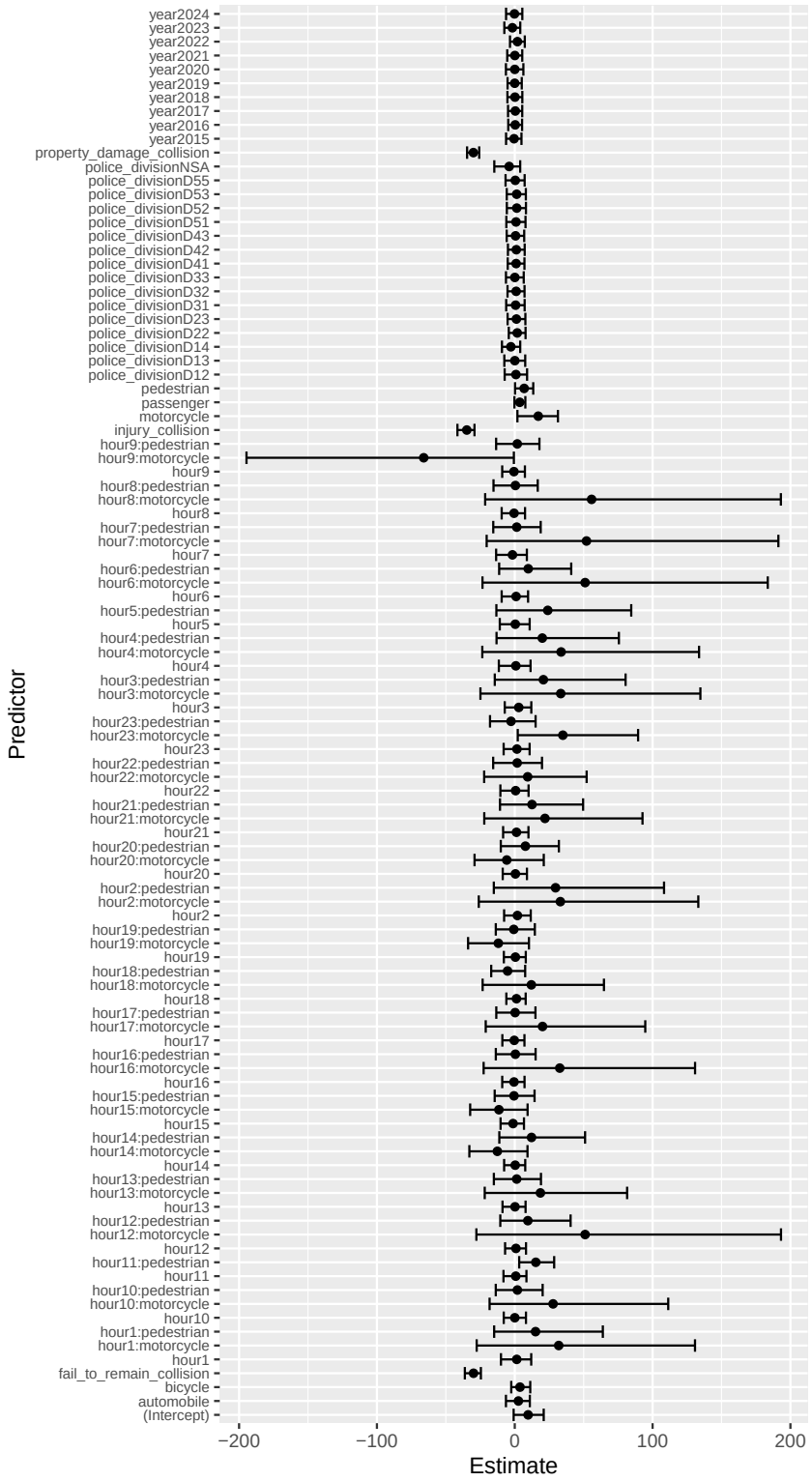

Figure 13: Coefficient Estimates for Predictors

5 Discussion

5.1 Contextualizing Role of Motorcyclists and Pedestrians

The results of this analysis both validate and challenge several common assumptions about motor vehicle collision (MVC) fatalities. Specifically, the high risk faced by motorcyclists, pedestrians, and individuals involved in accidents during rush hours or late evenings aligns with long-held perceptions about these vulnerable groups. However, the study also introduces a nuanced understanding of these factors, suggesting deeper complexities that could inform public policy.

5.1.1 Motorcyclists: The Most Vulnerable Group

In Table 12, the coefficient of 17.036 for motorcyclists indicates that they are the most at-risk group in terms of MVC fatalities. This result strongly aligns with the widely recognized vulnerability of motorcyclists in traffic accidents. Motorcycles, by their very design, offer little protection compared to other vehicles. In the event of a crash, motorcyclists are more likely to suffer severe injuries due to the lack of structural safeguards such as airbags, seat belts, or the protective frame that cars provide. Additionally, their smaller size makes them less visible to other drivers, increasing the likelihood of collisions. This heightened risk is compounded by the fact that motorcycles are often involved in high-speed crashes, particularly on highways or in urban environments with high traffic volumes. The coefficient underscores the need for targeted interventions, such as improved rider safety training, better infrastructure (e.g., dedicated motorcycle lanes), and stricter regulations on helmet use.

5.1.2 Pedestrians

Similarly, shown in Table 12, the coefficient of 6.809 for pedestrians reinforces the widely acknowledged dangers pedestrians face in urban settings. Pedestrian safety has long been a critical area of concern for urban planners and policymakers, and this result highlights the need for continued focus on this issue. Pedestrians are inherently vulnerable in any accident, and the lack of physical barriers between them and vehicles significantly increases their chances of injury or death in a collision. Moreover, in many urban areas, crosswalks, sidewalks, and pedestrian signals are either inadequate or poorly maintained, which contributes to pedestrian risk. This finding calls for heightened investment in urban infrastructure designed to protect pedestrians, including the expansion of safe crossing points, better traffic signaling, and measures to reduce vehicle speeds in pedestrian-heavy areas. It also suggests that policies addressing urban sprawl and public transportation options may help reduce pedestrian fatalities by limiting the need to walk on high-traffic streets.

5.1.3 Environmental Factors

The analysis also highlights the significant role that environmental factors—specifically, the time of day—play in MVC fatalities. During rush hours (particularly between hours 16 and 17) and late evenings (around 11:00 PM), the likelihood of a fatality, especially in motorcycle crashes, increases substantially. The 66% fatality likelihood for motorcycle crashes during these times suggests that several environmental and behavioral factors may be at play. During rush hours, the density of traffic can lead to higher stress levels among drivers, potentially contributing to reckless driving, lane changes, or a lack of attention to motorcyclists. Additionally, the rush-hour period coincides with fatigue, both for drivers who may have been on the road for extended periods and for motorcyclists who may be navigating more complex, congested traffic conditions.

Late evening crashes, on the other hand, are often linked to impaired driving and reduced visibility. Fatigue and alcohol impairment increase as the night progresses, leading to a higher likelihood of accidents that involve poor decision-making or delayed reactions. In the case of motorcycles, the combination of low visibility, decreased driver attention, and the increased risks of driving under the influence could explain why fatalities are particularly high during late-night hours. The study's findings align with common assumptions about these time-based risks, but they also reinforce the importance of implementing targeted interventions, such as stricter enforcement of impaired driving laws and improved lighting and signage to enhance visibility for motorcyclists during late-night hours.

5.2 Role of Protective Factors

The variables "injury_collision" and "property damage collisions" play a crucial role in lowering the likelihood of fatalities in motor vehicle collisions (MVCs). These outcomes suggest that, in contrast to fatal accidents, certain collisions are characterized by lower severity, different dynamics, and an overall reduced risk to life. By exploring why these variables lower fatality risks, we can better understand the factors that contribute to non-fatal collisions and how interventions might be designed to replicate these protective conditions across all collisions.

Injury collisions often result in less severe damage to both the vehicle and the individuals involved compared to fatal accidents. The presence of injuries in these cases indicates that, while there may be harm to the people involved, the force of impact and the nature of the collision may not have been severe enough to result in death. Injury collisions may involve safety features, such as airbags, seat belts, or vehicle crumple zones, that significantly mitigate the impact on the occupants, reducing the risk of fatal outcomes. Furthermore, when injuries occur, they often prompt quicker medical responses, either by emergency services or through hospitals' ability to treat injuries promptly. This quick medical intervention can prevent fatalities by addressing life-threatening injuries before they become fatal. Analyzing injury collisions can therefore offer valuable insights into what makes these accidents less deadly and

how certain safety features, such as improved emergency response times, advanced vehicle safety systems, and better road infrastructure, can prevent fatalities.

Property damage collisions are another protective factor that lowers the likelihood of fatalities. These types of collisions, by definition, involve no physical harm to the individuals involved, but they still represent significant risks of damage to the vehicles and the surrounding environment. The lower fatality risk in property damage collisions may be attributed to the collision dynamics being less intense or violent compared to crashes that result in fatalities. For example, these types of crashes may occur at lower speeds, with more gradual deceleration, which is less likely to cause life-threatening injuries. Additionally, property damage collisions may be associated with specific accident scenarios where vehicles are involved in minor impacts, such as fender-benders, or where the nature of the crash—such as a collision involving parked cars—does not subject the occupants to the full force of a high-speed crash.

5.3 Role of High Variation

While the results of the analysis provide valuable insights into the factors influencing MVC fatalities, they also highlight significant uncertainties, particularly with some predictors such as the interaction terms. In Figure 13, the variability of the interaction terms underscores the inherent challenges of modeling rare events such as fatalities, where small changes in conditions can lead to disproportionately large impacts. The uncertainty observed in these predictors is not just a statistical artifact but reflects the complexity and stochastic nature of accidents. These results highlight the need for caution when interpreting model outputs and point to the importance of further data collection, refinement, and more sophisticated modeling techniques to enhance the accuracy and reliability of predictions.

The variability in the coefficients for predictors l suggests that there may be additional, unobserved factors influencing the likelihood of fatalities that are not captured by the current model. This uncertainty could stem from gaps in the available data or from the limitations in the types of variables collected, such as the lack of detailed road conditions, driver behavior, or vehicle-specific characteristics (e.g., model, age of the vehicle). To reduce this uncertainty, a more comprehensive data collection strategy is needed, focusing on variables that could provide a deeper understanding of accident dynamics, such as weather conditions, road infrastructure quality, or even the socio-economic status of the areas where accidents occur. By expanding the scope and granularity of the data, it would be possible to refine the model, improving the precision of the estimates and providing clearer insights into the factors that contribute to fatalities.

The variability in the model's predictions also speaks to the stochastic nature of accidents, where the outcome of any given crash is influenced by a multitude of unpredictable factors. MVC fatalities, while significant, are relatively rare compared to non-fatal accidents, making them inherently difficult to predict. In this context, the uncertainty observed in the model's results can be understood as a reflection of the randomness and unpredictability involved in

traffic collisions. Factors such as driver response times, unexpected road hazards, or even the randomness of the sequence of events leading up to a fatal crash are beyond the scope of even the most sophisticated predictive models. In addition, fatalities often result from a confluence of circumstances—such as high speed, poor weather conditions, or the failure of safety mechanisms—that may not be fully captured by existing data or model specifications.

5.4 Implications

The findings from this analysis provide valuable insights for policymakers and law enforcement agencies seeking to improve road safety. One of the most pressing implications is the identification of high-risk periods, particularly during rush hours and late evenings, where the likelihood of fatal collisions is significantly higher. Targeted interventions could focus on enforcing speed limits and impaired driving laws during these critical hours, which may help mitigate the increased risk of accidents. For example, police could increase patrols and sobriety checkpoints, particularly in areas with higher concentrations of motorcycle or pedestrian collisions. Moreover, understanding the time of day when accidents are most likely to result in fatalities allows for more efficient deployment of resources, ensuring that law enforcement is present during times when their intervention could have the greatest impact on reducing fatalities.

In addition to targeted law enforcement, the analysis also points to infrastructure improvements that could reduce fatal outcomes for vulnerable groups, such as pedestrians and motorcyclists. Police and local governments should prioritize the creation of safer urban environments by enhancing lighting in poorly lit areas and developing separated lanes for bicycles and motorcycles. These measures could help reduce the exposure of vulnerable road users to traffic hazards, particularly during high-risk times. Furthermore, predictive insights from statistical modeling could revolutionize real-time road safety strategies. For instance, traffic management systems could use predictive data to adjust traffic light patterns during peak hours, optimizing traffic flow and reducing congestion-related accidents. Additionally, emergency services could use data-driven insights to allocate resources more efficiently, ensuring that areas with higher likelihoods of fatal collisions are prioritized for immediate response. In this way, statistical modeling can provide law enforcement and urban planners with the tools necessary to not only react to but proactively address high-risk situations on the road.

5.5 Weaknesses and Next Steps

The model's limitations stem from both the stochastic nature of traffic accidents and potential biases in the data. Fatalities are rare events, and the randomness inherent in accident outcomes—such as variations in road conditions, driver behavior, or other small factors—makes it difficult for models to predict them with high precision. Additionally, biases in data collection, like inconsistent reporting of certain variables or underreporting of less severe accidents, may lead to seleciton bias towards more severe accidnets in our datasets, affecting

our predictions. For example, in the United States, almost a third of car crashes involving a vulnerable road user go unreported (INSERT CITATION). Though this is a statistics from the United States, the article's findings highlight that the potential biases present in our data. These challenges introduce uncertainties in the model's estimates, particularly for less predictable outcomes like fatalities.

Moreover, the model's reliance on aggregate data limits its ability to account for the nuanced, dynamic nature of real-time road events. The factors influencing a fatality may not be static and can vary significantly based on the specific context of each collision. This variability makes it challenging to create a one-size-fits-all model for MVC fatalities, and the results are likely to change with shifts in environmental, societal, or technological factors. To improve accuracy and robustness, further data collection is essential, particularly to capture rare but impactful events and refine predictors of fatality outcomes. Enhancements in data quality and the incorporation of more granular, real-time data will improve model predictions and provide better insight into the complex dynamics of traffic accidents.

As a next step, future research should focus on collecting more granular, real-time data to capture the dynamic factors influencing MVC fatalities. This could involve leveraging emerging technologies such as smart traffic sensors, vehicle telematics, and data from wearable devices to capture more detailed and context-specific information. Additionally, refining statistical techniques, such as Bayesian hierarchical models or machine learning approaches, can help better account for the inherent randomness of accidents and the uncertainty in predicting rare outcomes. Improved modeling could lead to more accurate, actionable insights for policy development and real-time road safety interventions.

Appendix

.1 Exploration of Surveys, Sampling, and Observational Data

.1.1 Data Collection Methodology

The Toronto Police Service's Motor Vehicle Collision database is collected using a systematic process. When a collision occurs such that where persons are injured or a combined damage is valued at more than $\2000 to vehicles or property, or damage to any private, municipal or highway property, police officers complete a standardized Motor Vehicle Accident Report form. This data is then maintained by the City of Toronto, where the collection process follow a standardized process highlighted in the Motor Vehicle Collision Report manual (CITATION), ensuring reporting consistency. The dataset also defines a collision: the contact resulting from the motion of a motor vehicle or a street car of its load, which produces property damage, injury or death. A standardized definition ensures consistent reporting by the Toronto Police Service and highlights the dataset's clear boundaries.

For every accident that requires one to fill out the Motor Vehicle Accident Report form, the police officer fills out the SR-LD-402 form (CITATION). The process ensures consistency in reporting key details surrounding a motor vehicle accident.However, because police officers only need to fill out this form in specific instances, the data set only includes collisions where police officers were present or serious incidents that required a report. As a result, minor offences are usually left out in the dataset and often go unreported.

The dataset primarily focuses on all types of Motor Vehicle Collisions. The dataset specifically focuses on property damage collisions, fail to remain collisions, injury collisions and fatalities. Fatal Collisions occur when an individual's injuries from a MVC result in a fatality within 30 days. Fatal Collisions excludes occurrences on private property, ones related to sudden death prior to collision, such as suicide, and occurrences where the individual has died more than 30 days after the collision. Personal Injury Collisions occur when an individual involved in a MVC suffers personal injuries. Fail to Remain Collisions occur when an individual involved in a MVC fails to stop and provide their information at the scene of a collision. Property Damage Collisions occur when an individual's property has been damaged in a MVC or the value of damages is less than $2,000 for all involved These defintions, specifically the one for Fatal Collisions, sets a clear boundary for data inclusion, but excludes motor vehicle accidents that occur on private property.

Each entry in the dataset represents a single Motor Vehicle Collision incident. Therefore, multiple offences or victims can be associated with each record. This means that a singular person might have multiple records if they have gotten into multiple Motor Vehicle Collision incidents.

A notable aspect of the dataset is its handling of spatial data. Each collision includes the approximate latitude and longitude coordinates. The location of calls for service have been

deliberately offset to the nearest road intersection node to protect the privacy of parties. Because of the offset of offences location, the coordinates may not reflect the exact count of occurrences.The statistics at the neighborhood level may not reflect the true location of the collisions. The Toronto Police Service acknowledges these limitations and does not guarantee the accuracy, completeness, timeliness of the data, and TPS encourages to not compare it to any source of crime data.

.1.2 Idealized Methodology

The current data collection methodology employed by the Toronto Police Service provides valuable insights into Motor Vehicle Collisions (MVCs), yet several enhancements could make the dataset more comprehensive and actionable. Expanding the criteria for reporting is a critical first step. By including minor collisions that do not meet the $2,000 damage threshold or require police presence, the city can gain a fuller understanding of road safety dynamics. A self-reporting mechanism, such as an online portal or app, would allow drivers and witnesses to log details of these incidents conveniently. Additionally, redefining fatality criteria to include deaths occurring beyond 30 days due to collision-related injuries would align the dataset with public health standards, ensuring fatalities are not undercounted.

Incorporating advanced data collection techniques could significantly enhance data accuracy and granularity. Real-time telemetry from vehicles, collected via partnerships with automakers and insurers, can provide precise information on collision dynamics such as speed, braking patterns, and collision forces. Crowdsourcing data through a city-wide mobile app could also capture unreported incidents and near-misses, enriching the dataset with granular environmental and behavioral details. Integration of traffic camera data, analyzed using AI-powered algorithms, would add an unbiased and continuous data stream to validate collision reports and identify patterns. These methods collectively address gaps in the current system, providing deeper insights into the factors contributing to MVCs.

Expanding the scope of variables collected in collision reports is another avenue for improvement. Adding details about driver behavior (e.g., speeding or distracted driving), the involvement of emerging vehicle technologies (e.g., self-driving systems), and the road environment (e.g., construction zones) can offer richer insights into the conditions that contribute to MVCs. This would help policymakers and researchers better understand the interplay of human, environmental, and technological factors in collision outcomes.

.1.3 Example Survey

To gather more information about motor vehicle collisions, I've attached a proposed survey to better understand the factors that contribute to fatalities and injuries. We aim to replicate something similar to the Motor Vehicle Collision Report and aim to collect data for car crashes in Toronto. All responses will be confidential. I've attached the proposed survey here.

.2 Enhancement: Model Card

.2.1 Model Overview

This model predicts the occurrence of motor vehicle collisions (MVCs) based on various factors, including the time of day (hour), type of vehicle involved (motorcycle, pedestrian, automobile, etc.), and police division. The model is trained using data from 2015 to 2024, with a focus on the interactions between the hour of the day and motorcycle or pedestrian involvement.

.2.2 Model Details

- **Model Type**: Bernoulli regression (logistic regression for binary outcomes)
- **Target Variable**: Binary classification of motor vehicle collisions (occurrence of MVCs)
- **Training Data**: 1606 observations
- **Features**:
  - Hour of the day (`hour1` to `hour23` representing 24 hours in a day)
  - Vehicle type (motorcycle, pedestrian, automobile, etc.)
  - Type of accident (injury_collision, fail_to_remain_collision, and property_damage_collision)
  - Police division (D12, D13, etc.)
  - Year of the accident (2015–2024)

.2.3 Model Coefficients

- **Intercept**: 9.705
- **Hour Coefficients**:
  - `hour1` through `hour23`: Varies with each hour of the day. Positive coefficients for hours between 2 PM to 11 PM suggest higher accident probability in these hours.
  - **Key Hours**:
    - `hour2`: 1.920
    - `hour3`: 2.908
    - `hour6`: 0.966
    - `hour7`: -1.673
    - `hour8`: -0.511
    - `hour21`: 1.237
    - `hour22`: 0.631
    - `hour23`: 1.572
- **Vehicle Type Coefficients**:
  - **Motorcycle**: 17.036, indicating a high likelihood of motorcycle-related collisions.

  - **Pedestrian**: 6.832, showing the significance of pedestrian involvement in accidents.
  - **Bicycle**: 3.809, showing a moderate likelihood of bicycle-related accidents.
  - **Automobile**: 2.648, indicating that automobile-related collisions are also common.

- **Collision Type Coefficients**:
  - **Injury Collision**: -34.794, indicating a strong negative association with injury-related collisions.
  - **Fail to Remain Collision**: -29.912, indicating a strong negative association with collisions where the driver fails to remain.
  - **Property Damage Collision**: -30.004, showing a significant negative impact for property damage-related collisions.

- **Police Division Coefficients**:
  - `police_divisionD12`: 0.846, indicating division-specific effects, where division D12 has a positive association with collisions.
  - `police_divisionD13`: 0.075
  - `police_divisionD14`: -2.806
  - `police_divisionD22`: 1.797
  - `police_divisionD23`: 1.193
  - `police_divisionD31`: 0.619
  - `police_divisionD32`: 1.007
  - `police_divisionD33`: 0.055
  - `police_divisionD41`: 0.989
  - `police_divisionD42`: 1.140
  - `police_divisionD43`: 0.582
  - `police_divisionD51`: 0.915
  - `police_divisionD52`: 1.465
  - `police_divisionD53`: 1.389
  - `police_divisionD55`: 0.389
  - `police_divisionNSA`: -4.040

- **Year Coefficients**:
  - `year2015`: -0.500
  - `year2016`: 0.329
  - `year2017`: 0.412
  - `year2018`: 0.165
  - `year2019`: -0.025
  - `year2020`: 0.021
  - `year2021`: 0.079
  - `year2022`: 1.947, showing a significant increase in collisions during this year.
  - `year2023`: -1.708, showing a decrease in collisions during this year.
  - `year2024`: -0.210

- **Interaction Terms**:
  - **Hour × Motorcycle**:
    - `hour1 × motorcycle`: 31.900
    - `hour2 × motorcycle`: 33.094
    - `hour3 × motorcycle`: 33.377
    - `hour4 × motorcycle`: 33.633
    - `hour6 × motorcycle`: 51.053
    - `hour7 × motorcycle`: 52.055
    - `hour8 × motorcycle`: 55.746
    - `hour9 × motorcycle`: -66.051
    - `hour10 × motorcycle`: 27.802
    - `hour12 × motorcycle`: 51.049
    - `hour13 × motorcycle`: 18.620
    - `hour14 × motorcycle`: -12.621
    - `hour15 × motorcycle`: -11.489
    - `hour16 × motorcycle`: 32.651
    - `hour17 × motorcycle`: 20.150
    - `hour18 × motorcycle`: 12.044
    - `hour19 × motorcycle`: -11.899
    - `hour20 × motorcycle`: -5.807
    - `hour21 × motorcycle`: 21.915
    - `hour22 × motorcycle`: 9.369
    - `hour23 × motorcycle`: 34.897
  - **Hour × Pedestrian**:
    - `hour1 × pedestrian`: 15.095
    - `hour2 × pedestrian`: 29.594
    - `hour3 × pedestrian`: 20.771
    - `hour4 × pedestrian`: 19.986
    - `hour5 × pedestrian`: 23.943
    - `hour6 × pedestrian`: 9.778
    - `hour7 × pedestrian`: 1.434
    - `hour8 × pedestrian`: 0.484
    - `hour9 × pedestrian`: 1.831
    - `hour10 × pedestrian`: 1.895
    - `hour11 × pedestrian`: 15.280
    - `hour12 × pedestrian`: 9.497
    - `hour13 × pedestrian`: 1.397
    - `hour14 × pedestrian`: 12.171
    - `hour15 × pedestrian`: -0.552
    - `hour16 × pedestrian`: 0.453
    - `hour17 × pedestrian`: 0.272
    - `hour18 × pedestrian`: -5.210

* `hour19 × pedestrian`: -0.707
* `hour20 × pedestrian`: 7.720
* `hour21 × pedestrian`: 12.510
* `hour22 × pedestrian`: 1.760
* `hour23 × pedestrian`: -2.690

.2.3.1 Interaction Terms

- **Hour and Motorcycle**:
  - `hour1 × motorcycle`: 31.900, significant interaction in early morning hours, suggesting motorcycle accidents are more likely during certain hours.
  - Other significant hours for motorcycle collisions are `hour6 × motorcycle` (51.053) and `hour9 × motorcycle` (-66.051).
- **Hour and Pedestrian**:
  - `hour1 × pedestrian`: 15.095, significant pedestrian accidents in the early morning hours.
  - Other significant hours include `hour2 × pedestrian` (29.594) and `hour5 × pedestrian` (23.943), highlighting peak pedestrian accident hours.

.2.4 Model Performance Metrics

- **R^2**: 0.992, indicating excellent fit to the data.
- **Log Likelihood**: -0.082
- **Effective Log Predictive Density (ELPD)**: -16.2, with an associated standard error of 2.2.
- **Leave-One-Out Information Criterion (LOOIC)**: 32.4, with a standard error of 4.4.
- **Watanabe-Akaike Information Criterion (WAIC)**: 21.3
- **Root Mean Squared Error (RMSE)**: 0.00, indicating near-perfect predictive accuracy.

.2.5 Training Time

- **Chain 1**:
  - Warm-up: 29.99 seconds
  - Sampling: 26.11 seconds
  - Total: 56.1 seconds
- **Chain 2**:

  - Warm-up: 19.62 seconds
  - Sampling: 17.23 seconds
  - Total: 36.84 seconds

- **Chain 3**:
  - Warm-up: 19.88 seconds
  - Sampling: 19.30 seconds
  - Total: 39.18 seconds

- **Chain 4**:
  - Warm-up: 22.40 seconds
  - Sampling: 19.17 seconds
  - Total: 41.57 seconds

.2.6 Conclusion

The model performs exceptionally well in predicting motor vehicle collisions based on hour, vehicle type, and other factors, with the highest predictive power for motorcycle-related accidents. Key interactions, such as between hour and motorcycle or pedestrian, highlight the importance of specific hours of the day in predicting collision occurrences. This model can be used for more effective traffic safety planning and intervention strategies.

.3 Data Attribution Statement

"This data contains information licensed under the Open Government License - Toronto" (**tphlicense?**).

.4 Model Diagnostics

Figure 14 compares observed data (dark line) with replicated posterior predictions (light blue lines). The close alignment suggests that the model accurately captures the data's central tendency and variability. Figure 15 and Figure 16 show that the sampling algorithm used, the Markov chain Monte Carlo (MCMC) algorithm, did not run into issues as the posterior distribution for the model was created. Using the checks presented by (**citetellingstorieswithdata?**), both graphs do not show anything abnormal since the trace plots in Figure 15 display substantial horizontal fluctuation across chains, indicating good mixing, while the Rhat values in Figure 16 are close to 1 and well below 1.05, further supporting convergence.

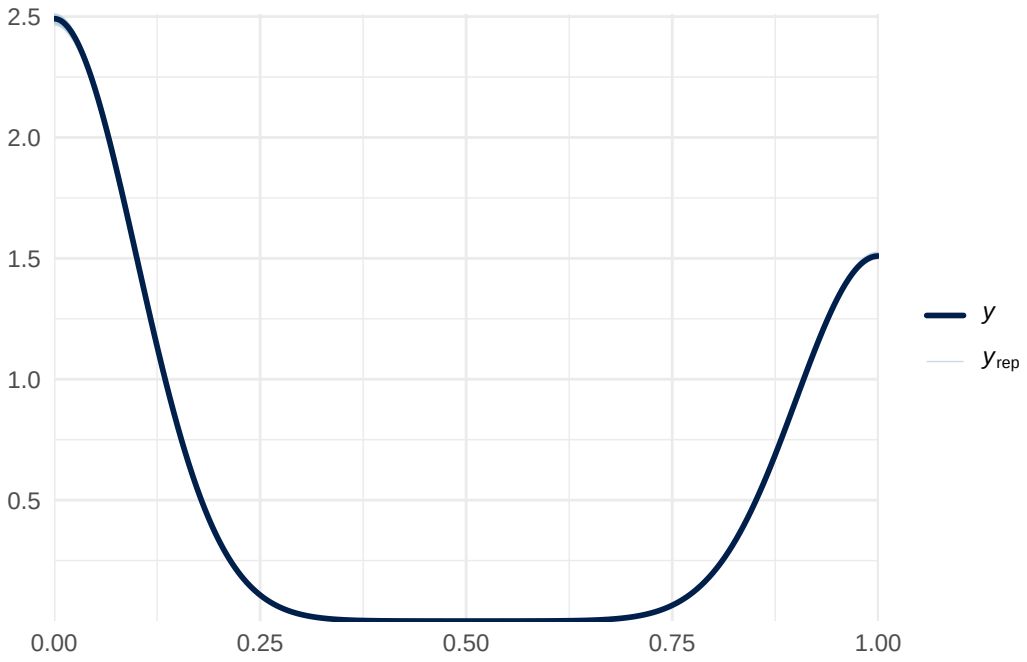

Figure 14: Posterior Predictive Check: Comparison of Observed and Replicated Data

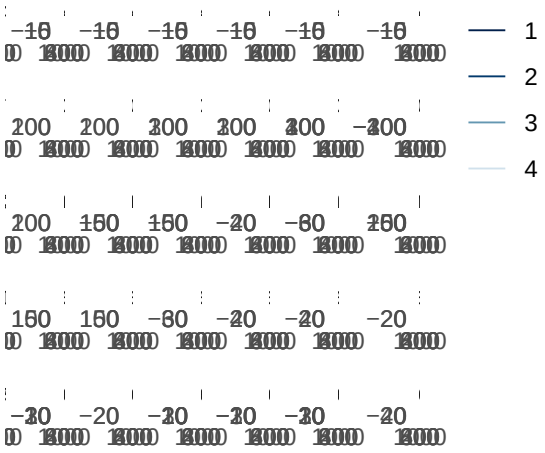

Figure 15: Checking the convergence of the MCMC algorithm - Trace

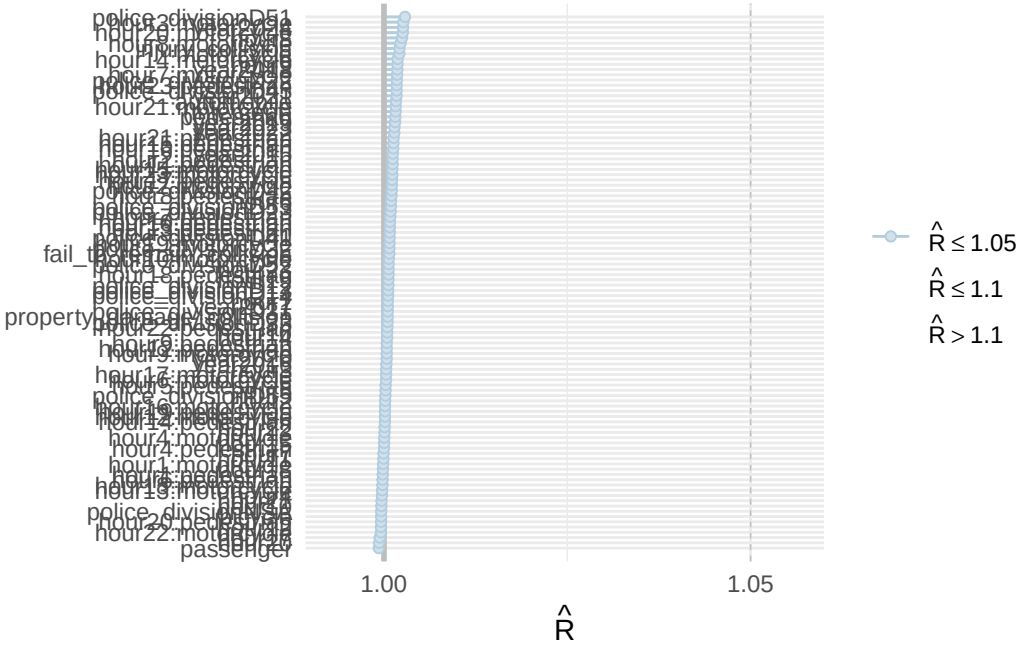

Figure 16: Checking the convergence of the MCMC algorithm - Rhat

References

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