My title*

My subtitle if needed

First author Another author

November 16, 2024

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

1 Introduction

1.1 Overview

This paper analyzes the factors influencing the likelihood of a violent crime versus a non-violent crime, focusing on specific premises and times of day as predictors. Violent crime continues to be a significant societal issue, necessitating evidence-based strategies to improve community safety. Using data from a publicly available dataset on crimes in Toronto, this research investigates how the setting and time correlate with the probability of violent crimes. By exploring these associations, we aim to uncover actionable insights to inform crime prevention policies.

1.2 Estimand

The primary estimand of this study is the probability of a violent crime occurring given specific premises (e.g., residential, commercial, or transit) and the time of day (e.g., early morning, evening). Violent crimes include serious offenses such as assaults and robberies, while non-violent crimes encompass thefts and similar offenses. This research employs a Bayesian logistic regression model to estimate these probabilities, capturing uncertainty while incorporating prior knowledge.

*Code and data are available at: https://github.com/RohanAlexander/starter_folder.

1.3 Results Summary

1.4 Why This Study Matters

Understanding the contextual factors contributing to violent crimes can help city planners and policymakers prioritize safety interventions. This research focuses on premises and times of day where the likelihood of violence is highest, enabling tailored prevention strategies. The insights derived from this study can directly impact resource allocation for policing and urban planning, potentially reducing crime rates and enhancing public safety.

1.5 Paper Structure

The remainder of this paper is structured as follows. In Section 2, we present an overview of the dataset used in this study, along with a detailed description of the variables and the data cleaning process. Section 4 outlines the Bayesian logistic regression model applied in our analysis, including the model setup, assumptions, and justification. In **?@sec-result**, we summarize the results of our analysis and interpret the findings within the context of existing literature on road safety. Finally, Section 6 discusses the implications of these findings, acknowledges the limitations of the study, and suggests directions for future research.

2 Data

2.1 Overview

We use the statistical programming language R (R Core Team 2023) to process, clean, and analyze the dataset for this study. Our data is sourced from the Open Data Toronto portal (Open Data Toronto 2021), specifically from the "Motor Vehicle Collisions Involving Killed or Seriously Injured Persons" dataset, which captures detailed information on traffic incidents in Toronto where individuals were either killed or seriously injured. This dataset, compiled by the Toronto Police Service, includes variables that record the time, location, environmental conditions, and specific parties involved in each collision, providing valuable insights into factors that may increase the risk of severe outcomes.

The detailed information within the dataset enables a rigorous analysis of patterns in collision severity. By examining factors such as driver condition, road surface, and lighting conditions, we aim to identify high-risk scenarios that may require targeted safety interventions. Following principles outlined in "Telling Stories with Data" (Alexander 2023), this study emphasizes transparency, reproducibility, and practical insights in presenting data-driven findings that can inform road safety policies.

3 References for references.bib

3.1 Measurement

The Motor Vehicle Collisions Involving Killed or Seriously Injured Persons dataset from Open Data Toronto provides critical information on traffic incidents involving fatalities or serious injuries in Toronto since 2006. Compiled and maintained by the Toronto Police Service, the dataset captures a variety of incident-specific metrics, including severity, driver condition, road surface condition, and lighting at the time of each collision. Incident severity is categorized as either fatal or serious injury, while environmental conditions, such as road surface (e.g., dry, wet, icy) and lighting (e.g., artificial or natural light), are recorded qualitatively. Each record also includes a timestamp and a location, though the precise geographic coordinates are adjusted to the nearest road intersection to protect the privacy of those involved.

Despite its value, this dataset has limitations that impact the accuracy and applicability of analyses. Most notably, the location of each incident is deliberately offset, meaning that any geographic analysis may not reflect the exact sites of collisions, particularly at the neighborhood or divisional level. Additionally, the dataset is limited to incidents involving significant injuries or fatalities, meaning it does not account for minor or unreported collisions. This focus on severe incidents may lead to an overrepresentation of high-severity events, which could influence interpretations of broader traffic patterns in Toronto. Furthermore, while the dataset aims to provide timely and complete information, the Toronto Police Service does not guarantee its accuracy, completeness, or timeliness, cautioning against direct comparisons with other sources of traffic or crime data. This dataset predominantly consists of categorical variables describing conditions and severity, with minimal quantitative data aside from incident timestamps, making it a qualitative yet powerful resource for studying serious traffic events in Toronto.x

3.2 Data Cleaning

The raw crime data underwent a comprehensive cleaning process to prepare it for analysis, ensuring the dataset was both relevant and consistent for the study's objectives. Initially, only key columns were selected from the raw dataset, including `EVENT_UNIQUE_ID` (Unique Identifier), `REPORT_DATE` (Date of Report), `OCC_HOUR` (Hour of Occurrence), `PREMISES_TYPE` (Type of Premises), and `OFFENCE` (Offense Type). This selection ensured that only the most pertinent information was included for examining factors associated with violent crimes.

Rows with missing or irrelevant data in critical columns, such as `EVENT_UNIQUE_ID`, `PREMISES_TYPE`, `TIME_OF_DAY`, and `VIOLENT_CRIME`, were removed to maintain data integrity and focus on valid observations. Offenses were categorized as either violent or non-violent to create a binary outcome variable (`VIOLENT_CRIME`), where offenses such as "Assault" and "Robbery" were classified as violent (`1`), and crimes like "Theft" or "Break and Enter" were

classified as non-violent (0). Offenses that did not fit into either category were excluded from the dataset.

To analyze temporal patterns, the OCC_HOUR variable was grouped into broader time categories:

- **Early Morning**: 12 AM to 6 AM
- **Morning**: 6 AM to 12 PM
- **Afternoon**: 12 PM to 6 PM
- **Evening**: 6 PM to 12 AM

Additionally, the PREMISES_TYPE variable, which describes the type of location where the incident occurred, was trimmed to remove excess whitespace and ensure consistent formatting. This step was crucial to prevent unexpected NA values during analysis.

The cleaning process also included filtering for crimes reported during the summer month of July, aligning the dataset with the study's focus on temporal and environmental patterns in criminal activity. After grouping, formatting, and filtering the data, all rows containing NA values in critical columns were dropped to ensure a clean and analyzable dataset.

Finally, the cleaned dataset was saved as a Parquet file for further analysis. This format ensures efficient storage and retrieval while preserving the structure of the data.

3.3 Outcome variable

The primary outcome variable in this study is Violent Crime, which categorizes each incident based on whether the crime is classified as violent or non-violent. Violent crime is coded as a binary variable, where "0" represents non-violent offenses (e.g., theft, break and enter), and "1" denotes violent offenses (e.g., assault, robbery). This classification allows for an assessment of factors contributing to the likelihood of a crime being violent, providing a framework for understanding the underlying conditions associated with violent behavior.

By focusing on crimes involving violence, this outcome variable highlights critical cases of interest for public safety and crime prevention efforts, as violent offenses often have severe consequences for victims and communities. This binary categorization simplifies the statistical analysis, enabling the application of logistic regression models to estimate the probability of violent crimes under varying conditions.

Analyzing violent crime in conjunction with predictor variables such as premises type and time of day helps identify the circumstances under which the risk of violence is heightened. This outcome measure serves as a crucial indicator for assessing and improving community safety

Table 2: Statistics summary of the cleaned Motor Vehicle Collisions dataset

```
 Violent Crime     Premises Type          Time of Day
 0:13502       Apartment  : 8568   Early Morning: 9350
 1:23559       House      : 6402   Morning      : 5951
               Commercial : 7104   Afternoon    : 9258
               Outside    :11235   Evening      :12502
               Educational:  407
               Transit    : 1010
               Other      : 2335
```

policies, as it reflects the most serious consequences of criminal activity and directs attention to factors that could reduce violent incidents.

Table 1: Preview of the cleaned Crime dataset

Violent Crime	Premises Type	Time of Day
1	House	Evening
1	House	Evening
1	House	Evening
1	Apartment	Evening
1	Outside	Early Morning

3.4 Predictor variables

3.5 Predictor Variables

Figure 1 illustrates the distribution of violent versus non-violent crimes across different premises types. The results show that the highest number of crimes, both violent and non-violent, occur in "Outside" locations, followed by "Apartment" premises. Notably, violent crimes significantly outnumber non-violent ones in "Outside" locations, while "Apartments" and "Houses" have a more balanced distribution between the two crime types. Premises categorized as "Educational," "Transit," and "Other" show relatively low numbers of both violent and non-violent crimes, with "Educational" premises having the fewest incidents overall.

Figure 2 highlights the distribution of violent versus non-violent crimes by the time of day. The data reveals that violent crimes are most common during the "Evening," followed by "Afternoon" and "Early Morning." In contrast, non-violent crimes are more evenly distributed throughout the day but still peak during the "Evening." These findings suggest that the time

of day plays a critical role in influencing the likelihood of violent crimes, with evenings being a particularly high-risk period.

By examining these distributions, we gain insights into where and when violent crimes are most likely to occur. Understanding these patterns is essential for developing targeted crime prevention strategies, such as increasing law enforcement presence in high-risk areas during evening hours or implementing community-based interventions in "Outside" locations.

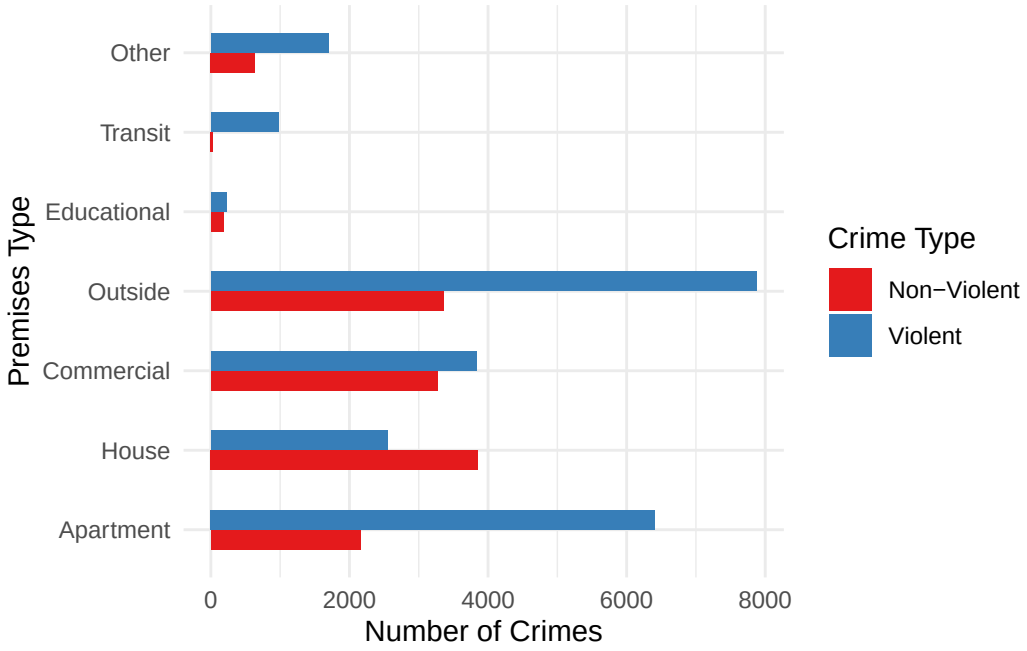


Figure 1: The distribution of violent vs. non-violent crimes by premises type

4 Model

In our analysis, we utilized a Bayesian logistic regression model to examine the relationship between violent crime occurrence and two key predictors: premises type and time of day. By using this model, we aimed to understand how the location and timing of incidents influence the likelihood of a crime being classified as violent. Detailed model diagnostics and background are provided in Appendix B.

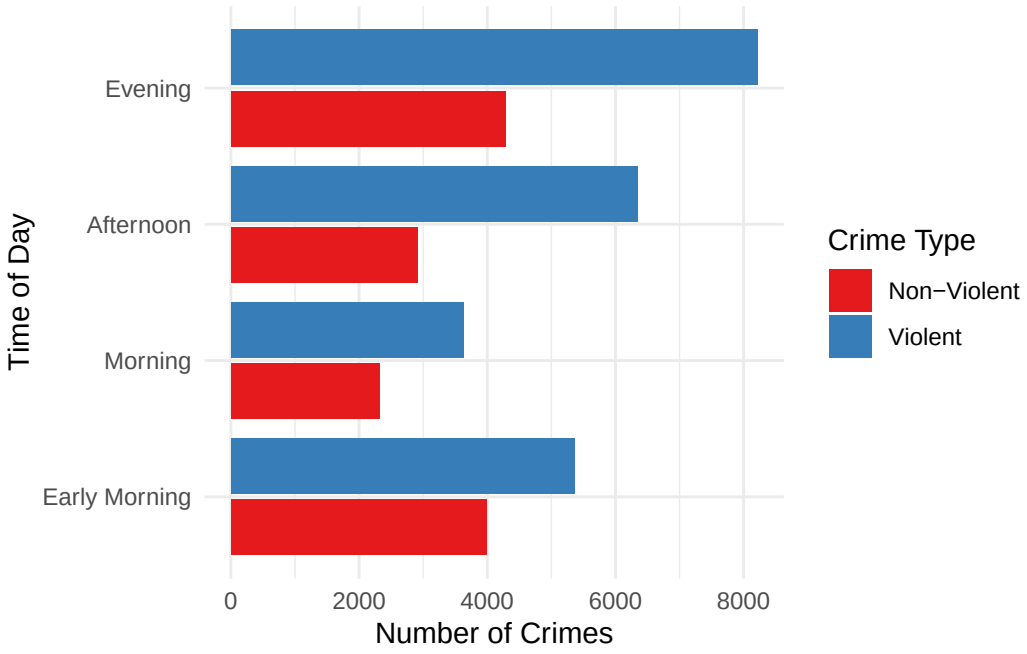


Figure 2: The distribution of violent vs. non-violent crimes by time of day

4.1 Model Set-up

$$y_i|\pi_i \sim \text{Bern}(\pi_i) \tag{1}$$

$$\text{logit}(\pi_i) = \alpha + \beta_1 \times \text{PREMISES\_TYPE}_i + \beta_2 \times \text{TIME\_OF\_DAY}_i \tag{2}$$

$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$

$$\beta_1 \sim \text{Normal}(0, 2.5) \tag{4}$$

$$\beta_2 \sim \text{Normal}(0, 2.5) \tag{5}$$

In this model:

- (y_i) represents the binary outcome variable indicating whether a crime is classified as violent ((1) for violent crimes, (0) for non-violent crimes).
- (_i) is the probability of a violent crime.
- ( ) is the intercept, representing the baseline log-odds of a violent crime when all predictors are at their reference levels.
- (_1) and (_2) are the coefficients associated with the predictor variables:
  - `PREMISES_TYPE`: the type of premises where the crime occurred.
  - `TIME_OF_DAY`: the time of day the crime occurred.

The logistic link function ((logit(_i))) models the log-odds of violent crime as a linear combination of the predictors. The priors for the intercept (( )) and the coefficients ((_1), (_2)) follow a Normal distribution with mean (0) and standard deviation (2.5). These priors regularize the model, constraining the parameters to plausible values based on prior knowledge.

This Bayesian logistic regression model was implemented using the `rstanarm` package, leveraging Markov Chain Monte Carlo (MCMC) sampling for parameter estimation. Further diagnostics and posterior summaries are discussed in the supplementary materials (see Appendix Section B).

4.2 Model Justification

Each predictor in the model represents a specific characteristic of the crime and its context, with hypothesized effects on the likelihood of violent crime. For `PREMISES_TYPE`, it is expected that certain locations, such as "Transit" or "Outside," may have higher odds of violent crime due to increased public interactions or a lack of controlled environments. Conversely, premises like "Educational" or "House" are anticipated to have lower odds of violent crime, reflecting their relatively private or structured settings. Locations categorized as "Other" may include mixed contexts, resulting in varying effects.

For `TIME_OF_DAY`, it is hypothesized that the "Evening" period may exhibit a higher likelihood of violent crimes, as this time is often associated with increased social activities and potential for interpersonal conflict. "Early Morning" may also show elevated odds due to decreased public surveillance and fewer witnesses. In contrast, "Morning" and "Afternoon" are expected to have lower odds, reflecting typical daytime routines with higher levels of oversight and structured activity.

By incorporating these predictors into the Bayesian logistic regression model, this study seeks to identify and quantify the relationships between crime characteristics and the likelihood of violent outcomes, providing actionable insights for targeted interventions.

5 Results

Our results are summarized in Table 3. The findings align with our expectations, providing insights into how premises type and time of day influence the likelihood of violent crimes. To avoid multicollinearity, the model excludes one level from each categorical predictor as the reference group: "Apartment" for `PREMISES_TYPE` and "Early Morning" for `TIME_OF_DAY`. The intercept represents the estimated log-odds of violent crime occurring when all other predictors are held constant at their reference levels. In this case, the estimated log-odds of a violent crime occurring at an "Apartment" during "Early Morning" is 0.881.

The type of premises strongly influences the likelihood of violent crimes. For instance, crimes occurring in "Transit" premises are far more likely to be violent compared to the reference

Table 3: Explanatory model Injury Severity Prediction (n = 1000)

	Violent Crime
(Intercept)	0.881
	(0.030)
PREMISES_TYPECommercial	−0.925
	(0.034)
PREMISES_TYPEEducational	−0.957
	(0.104)
PREMISES_TYPEHouse	−1.519
	(0.034)
PREMISES_TYPEOther	−0.138
	(0.053)
PREMISES_TYPEOutside	−0.267
	(0.032)
PREMISES_TYPETransit	2.363
	(0.190)
TIME_OF_DAYMorning	0.074
	(0.035)
TIME_OF_DAYAfternoon	0.413
	(0.032)
TIME_OF_DAYEvening	0.325
	(0.029)
Num.Obs.	37 061
R2	0.090
Log.Lik.	−22 567.707
ELPD	−22 577.6
ELPD s.e.	72.5
LOOIC	45 155.2
LOOIC s.e.	145.0
WAIC	45 155.1
RMSE	0.46

group. The estimated coefficient for `PREMISES_TYPETransit` is 2.363, indicating a substantial increase in the log-odds of violent crime in this context. On the other hand, crimes occurring in "Commercial" and "Educational" premises are less likely to be violent, with coefficients of -0.925 and -0.957, respectively. Similarly, crimes at "House" premises exhibit the lowest likelihood of violence, as reflected by a coefficient of -1.519. These results highlight that location plays a critical role in determining whether a crime is violent or non-violent.

Time of day also significantly impacts the likelihood of violent crimes. Compared to the reference level of "Early Morning," the coefficients for `TIME_OF_DAYAfternoon` and `TIME_OF_DAYEvening` are 0.413 and 0.325, respectively, suggesting an elevated likelihood of violent crimes during these times. The "Morning" time of day exhibits a negligible increase in violent crime likelihood, as shown by a coefficient of 0.074.

Figure 7 (see Section B.3) shows the range of coefficient estimates for our model within the 90% probability interval. However, because the credibility intervals for several predictors are relatively narrow, particularly for premises types like "Transit," it is challenging to observe the trends clearly. To address this, Figure 8 was created with the x-axis limited to a range of -5 to 5. Combining Figures 7 and 8, we observe statistical significance for several predictors, including "Transit," "Afternoon," and "Evening," as their credibility intervals do not cross zero.

The coefficients are reported in log-odds, where a positive value indicates an increased likelihood of violent crime, and a negative value reflects a decreased likelihood. The results demonstrate clear statistical significance for key predictors such as "Transit," "Afternoon," and "Evening," emphasizing the importance of both location and time of day in shaping violent crime outcomes.

6 Discussion

6.1 Relationship between Premises Type and Violent Crime

The analysis highlights that premises type significantly influences the likelihood of violent crimes. Crimes occurring in "Transit" locations are strongly associated with higher odds of being classified as violent compared to the reference group ("Apartment"). This finding aligns with expectations, as transit environments often involve high volumes of people and limited surveillance, which may increase the potential for interpersonal conflicts. Similarly, "Outside" locations exhibit a higher likelihood of violent crimes, likely due to their open and less controlled nature, which may facilitate confrontations or opportunistic violence. In contrast, premises such as "Commercial," "Educational," and "House" are associated with lower odds of violent crimes. These findings suggest that environments with greater structural or social control (e.g., schools, private residences) are less conducive to violent activities. These insights

underscore the importance of targeted safety measures in high-risk locations, such as enhancing surveillance in transit areas or increasing community policing efforts in outdoor public spaces.

6.2 Relationship between Time of Day and Violent Crime

Time of day also plays a critical role in determining the likelihood of violent crimes. The model results indicate that "Afternoon" and "Evening" periods are associated with elevated odds of violent crimes compared to the reference period ("Early Morning"). This pattern likely reflects increased social interactions and activities during these times, which may lead to heightened opportunities for conflicts or criminal acts. In contrast, "Morning" hours exhibit a negligible increase in violent crime likelihood, possibly reflecting a time of day characterized by structured routines and reduced social friction. These findings highlight the need for time-sensitive interventions, such as increased police presence during evening hours or community-based programs to address the underlying causes of violent behavior during high-risk times.

6.3 Implications for Policy and Prevention

The results provide actionable insights for crime prevention and public safety policies. For example, transit and outdoor areas, identified as high-risk premises, could benefit from increased surveillance, improved lighting, and targeted community engagement programs. Similarly, allocating law enforcement resources more strategically to focus on evening and afternoon periods could help deter violent crimes during these high-risk times. The findings also support urban planning efforts that enhance safety features in public spaces, such as the installation of security cameras and the promotion of neighborhood watch programs. By addressing the specific premises and times associated with violent crimes, policymakers can implement evidence-based strategies to reduce violence and improve community safety.

6.4 Limitations

6.4.1 Data Constraints

One limitation of this study is the exclusion of crimes that could not be definitively classified as violent or non-violent. While necessary to ensure data consistency, this filtering process may omit incidents that could provide additional context for understanding violent crime dynamics. Future research should explore ways to include these ambiguous cases to improve the comprehensiveness of the analysis.

6.4.2 Focus on a Single City

The dataset is specific to crimes reported in Toronto, which limits the generalizability of the findings to other regions or cities. Crime dynamics can vary widely based on local socioeconomic, cultural, and environmental factors. Further studies should replicate this analysis in other urban areas to assess the extent to which these findings apply across different contexts.

6.4.3 Simplification of Time Categories

Grouping time into four broad categories (Early Morning, Morning, Afternoon, and Evening) simplifies analysis but may obscure more granular patterns in violent crime occurrence. Future research could explore finer temporal granularity, such as analyzing crimes on an hourly basis, to capture additional nuances in time-of-day effects.

6.5 Future Steps

To enhance the utility and scope of this research, future studies should consider incorporating additional predictors, such as socioeconomic factors, neighborhood characteristics, or the presence of law enforcement resources. These variables could provide a more comprehensive understanding of the conditions contributing to violent crimes. Additionally, integrating spatial analysis with precise location data could identify specific hotspots for targeted interventions. Policymakers and city planners could leverage these insights to allocate resources more effectively, reduce violent crimes, and promote public safety in Toronto and beyond.

Appendix

A Additional data details

B Model details

B.1 Posterior Predictive Check and Prior-Posterior Comparison

The posterior predictive check in Figure 3 provides an evaluation of how well the Bayesian logistic regression model replicates the observed data. The posterior distribution ((y_{rep})) closely aligns with the observed data ((y)), indicating that the model provides a good fit and captures the essential patterns in the data.

In Figure 4, the posterior distributions of the model parameters are compared against their priors. This visual comparison highlights how the observed data influenced the posterior estimates. For instance, predictors such as `PREMISES_TYPETransit` and `TIME_OF_DAYAfternoon` show a notable shift in their posterior distributions compared to the priors, reflecting strong evidence provided by the data. Conversely, predictors with minimal shifts suggest less data-driven evidence or smaller effects on the model outcome.

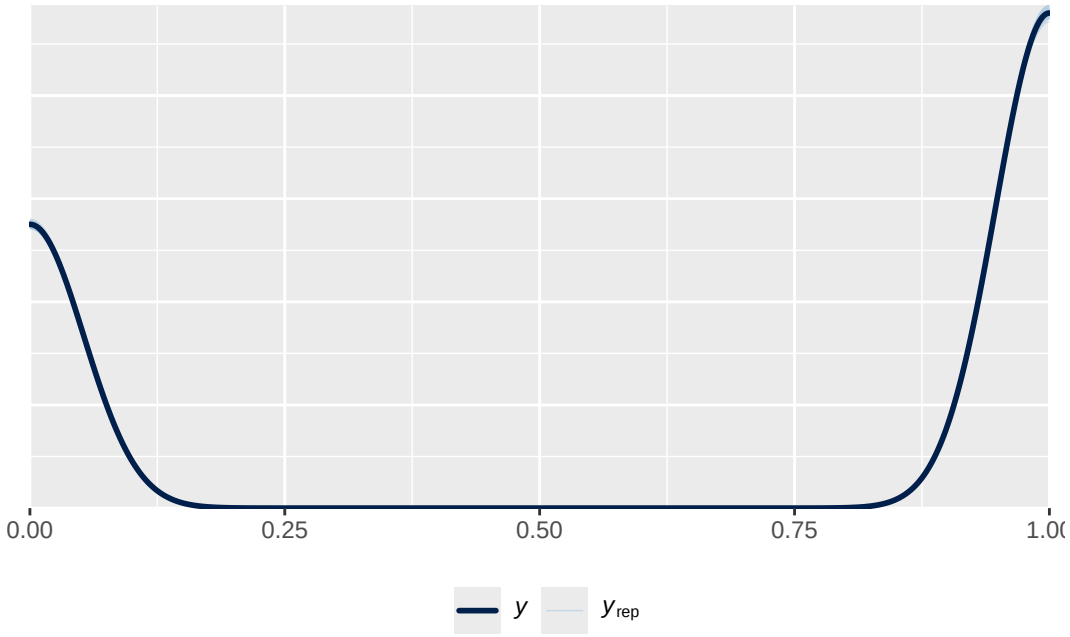


Figure 3: Posterior distribution for logistic regression model

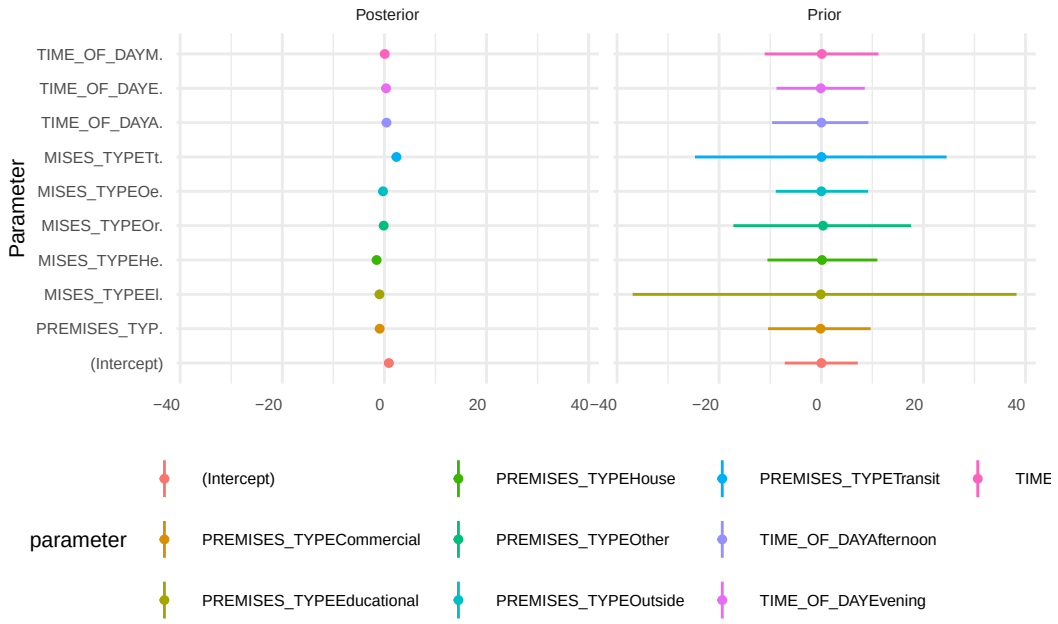


Figure 4: Comparing the posterior with the prior

B.2 Diagnostics

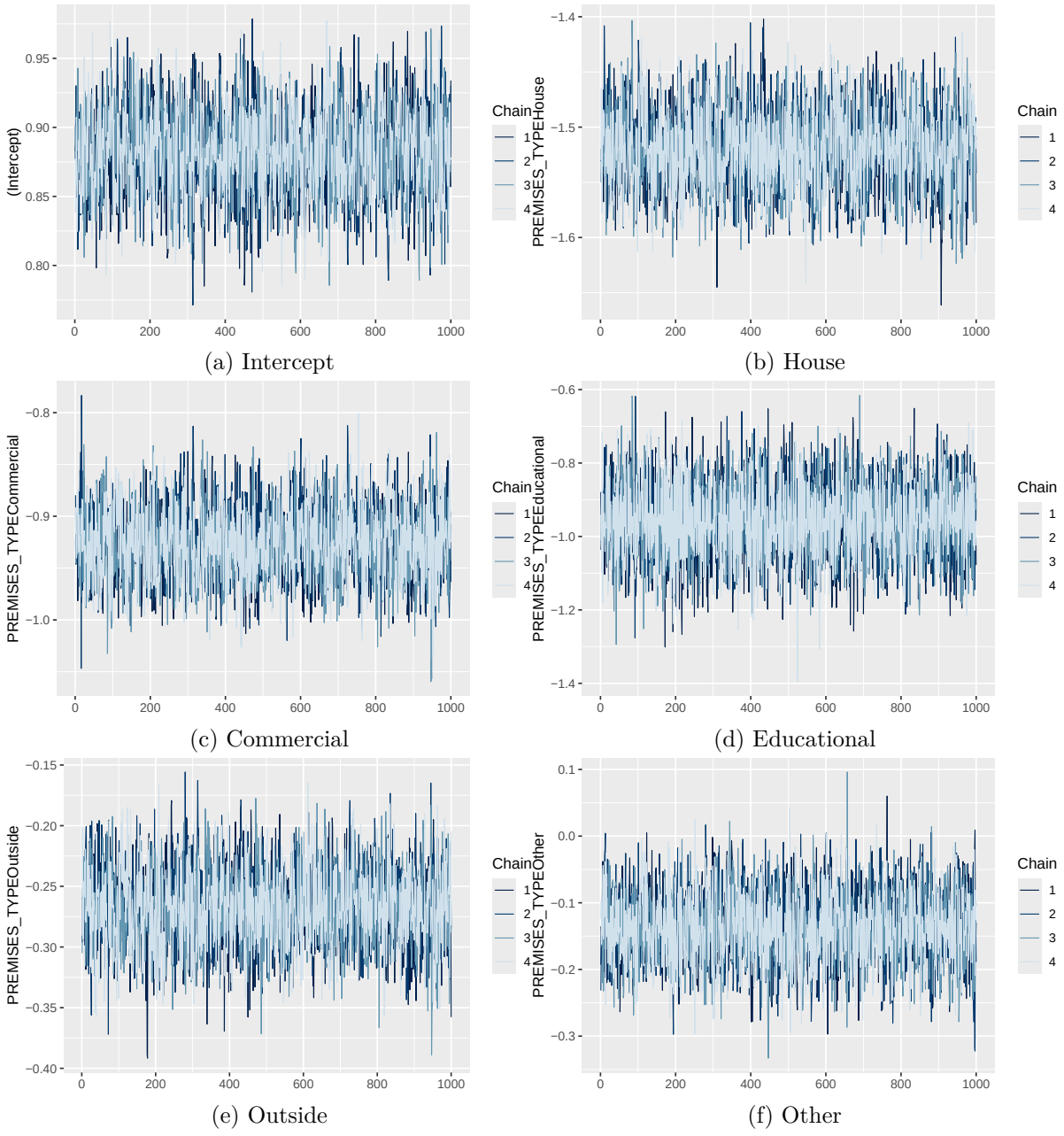
B.2.1 Markov Chain Monte Carlo Convergence Check

To assess the convergence of the Markov Chain Monte Carlo (MCMC) sampling for the Bayesian logistic regression model, trace plots and Rhat plots were analyzed. These diagnostics are crucial to ensure that the chains mix well and converge to a stable distribution, providing reliable parameter estimates.

B.2.1.1 Trace Plots

Figure 5 and Figure 6 display the trace plots for the model's intercept, premises type predictors, and time of day predictors. Each plot shows the sampled parameter values across iterations for all four chains.

- **Premises Type Predictors**: As seen in Figure 5, the chains for predictors like `House`, `Commercial`, `Educational`, and `Outside` fluctuate around stable mean values, indicating good mixing and convergence.
- **Time of Day Predictors**: Figure 6 illustrates the trace plots for `Morning`, `Afternoon`, and `Evening` time categories, along with `Transit Premises`. The chains for these predictors also show no significant trends or drift, confirming convergence.

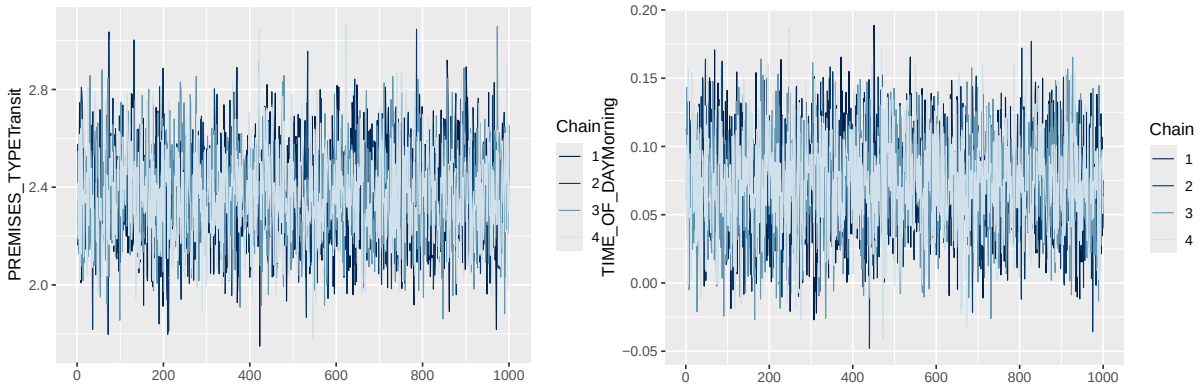


(a) Intercept (b) House

(c) Commercial (d) Educational

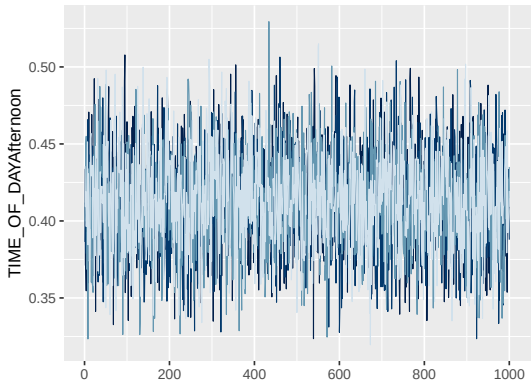
(e) Outside (f) Other

Figure 5: Trace plots of intercept and premises type predictors

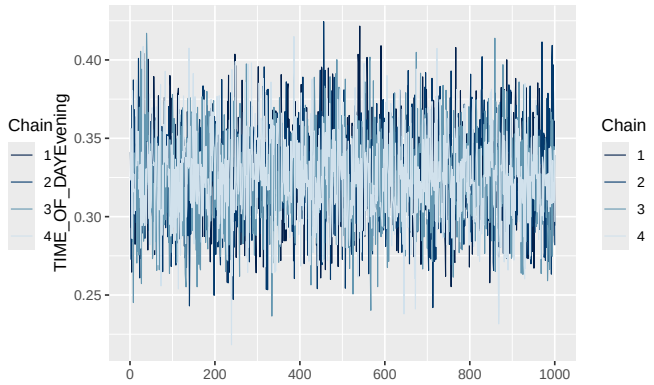


(a) Transit Premises

(b) Morning



(c) Afternoon



(d) Evening

Figure 6: Trace plots of time of day predictors

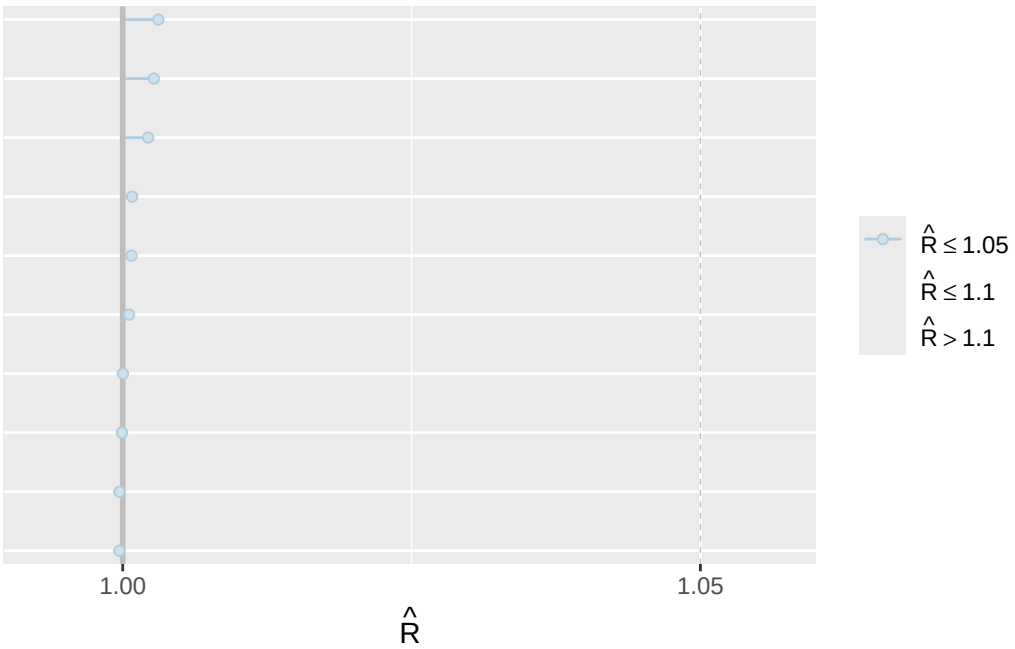


Figure 7: Rhat plot for MCMC convergence

B.3 90% Credibility Interval

Figures Figure 8 and Figure 9 visualize the 90% credible intervals for the predictors in the Bayesian logistic regression model. These plots provide insights into the uncertainty surrounding the estimated coefficients for each predictor.

Figure Figure 8 displays the 90% credible intervals for all predictors without restrictions on the x-axis. This plot shows how the predictors such as `PREMISES_TYPE` and `TIME_OF_DAY` influence the likelihood of violent crimes. For example, `PREMISES_TYPETransit` has a high positive credible interval, indicating a significant increase in the likelihood of violent crime in transit areas. Conversely, `PREMISES_TYPECommercial` and `PREMISES_TYPEEducational` have negative credible intervals, indicating a lower likelihood of violent crimes in these premises types.

In Figure Figure 9, the x-axis has been restricted to a range of -5 to 5 to better visualize predictors with smaller intervals. This refined view helps highlight subtle but meaningful differences between categories like `TIME_OF_DAYAfternoon` and `TIME_OF_DAYEvening`, which show a noticeable increase in the likelihood of violent crimes compared to the reference category (`TIME_OF_DAYEarly Morning`).

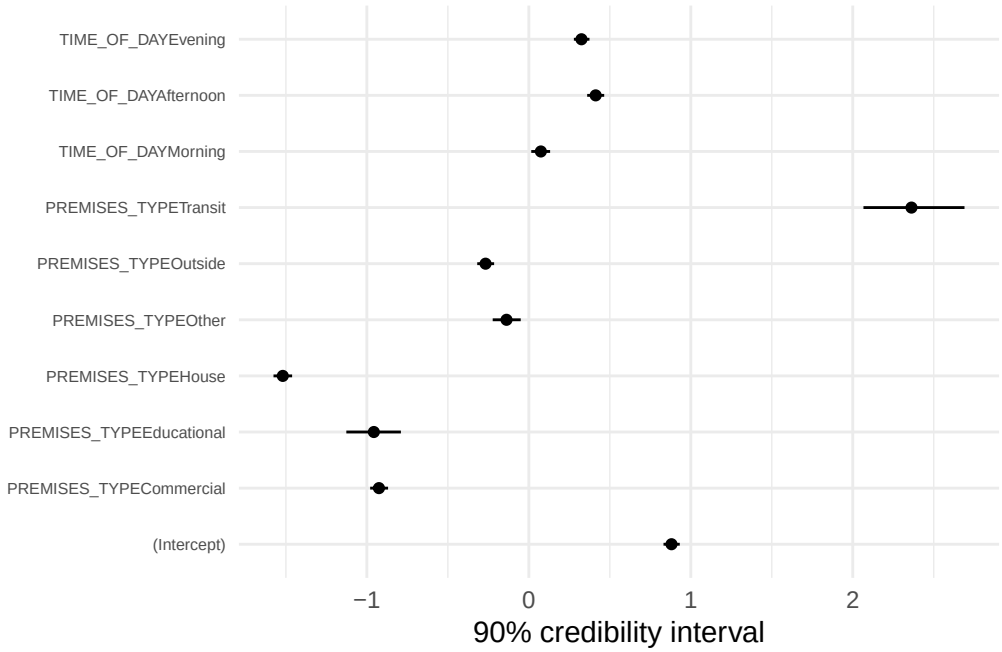


Figure 8: 90% credible intervals for predictors of violent crime likelihood

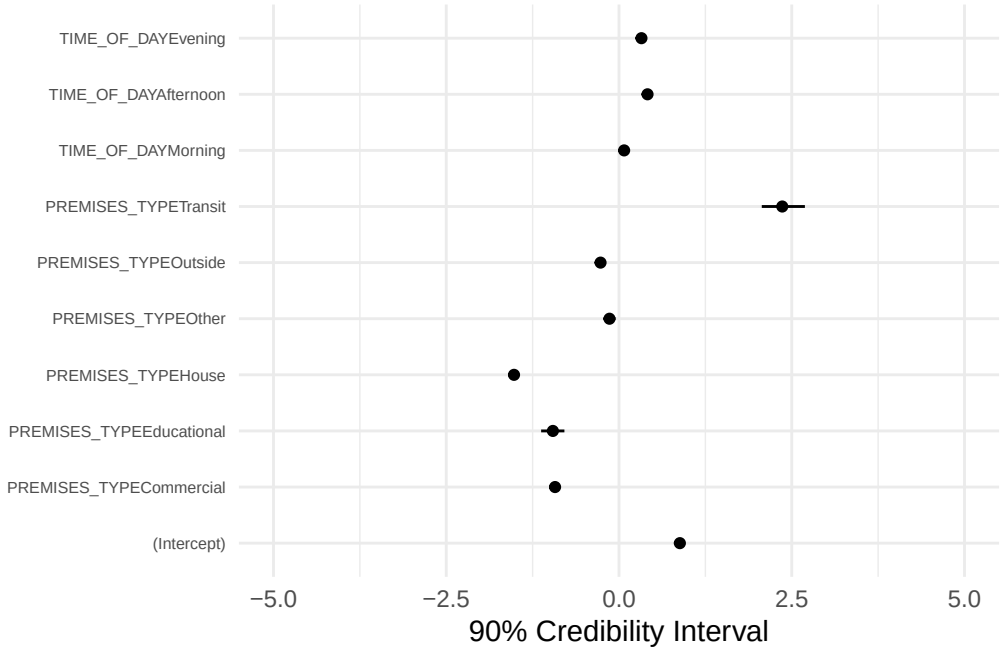


Figure 9: 90% credible intervals for predictors of violent crime likelihood with restricted x-axis limits

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

Open Data Toronto. 2021. "Motor Vehicle Collisions Involving Killed or Seriously Injured Persons." https://open.toronto.ca/dataset/motor-vehicle-collisions-involving-killed-or-seriously-injured-persons/.

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