My title*

My subtitle if needed

First author Another author

December 3, 2024

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

1 Introduction

Traffic collisions pose a significant public health and safety challenge worldwide, accounting for over 1.35 million deaths annually (World Health Organization 2018). In urban centers like Toronto, the complexities of traffic dynamics, urban planning, and population growth exacerbate the risks associated with road transportation. Understanding the patterns and determinants of traffic collisions is essential for developing effective interventions to enhance road safety.

Despite concerted efforts by city authorities, including the implementation of the Vision Zero Road Safety Plan in 2017, Toronto continues to grapple with high rates of traffic collisions, injuries, and fatalities (City of Toronto 2017). Previous studies have explored factors influencing collision rates, such as driver behavior, weather conditions, and infrastructure design (Ma et al. 2019; Wazana et al. 2020). However, there remains a critical gap in comprehensively analyzing spatial and temporal trends at a granular neighborhood level, particularly in assessing the impact of policy interventions over time.

This paper addresses this gap by conducting an in-depth analysis of Toronto's traffic collision data from 2014 to 2021. Leveraging advanced statistical modeling and geospatial analysis techniques, we examine the following research questions:

- **Spatial Patterns**: Which neighborhoods in Toronto exhibit higher rates of traffic collisions, and what spatial patterns emerge when visualizing collision data across the city?
- **Temporal Trends**: How have traffic collision rates changed over time, particularly before and after the implementation of the Vision Zero initiative?

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

- **Policy Impact**: What is the measurable impact of the Vision Zero Road Safety Plan on collision frequencies and severities in Toronto?

Estimand: The primary estimand is the expected annual number of traffic collisions in each Toronto neighborhood, accounting for temporal trends and the implementation of the Vision Zero policy.

By integrating spatial and temporal analyses, our study provides a comprehensive understanding of traffic collision dynamics in Toronto. The findings offer valuable insights for policymakers, urban planners, and public health officials to inform targeted interventions and resource allocation aimed at reducing traffic-related incidents.

The remainder of this paper is organized as follows: The Data section describes the datasets used, detailing the variables of interest and data preparation steps, including visualizations that illustrate key patterns. The Methodology section outlines the statistical models and geospatial techniques employed. The Results section presents the findings of our analyses, and the Discussion interprets these results in the context of existing literature and policy implications. Finally, the Conclusion summarizes the main contributions and suggests avenues for future research.

2 Data

2.1 Data Sources

Our analysis utilizes two primary datasets:

1. **Traffic Collision Data**: Detailed records of reported traffic collisions in Toronto from January 2014 to December 2021, obtained from the **City of Toronto's Open Data Portal** (City of Toronto 2022a). The dataset includes information on collision dates, times, locations, severities, and parties involved.

    - **Data Access**: Toronto Traffic Collisions Data

2. **Toronto Neighborhood Boundaries**: Geospatial data defining the boundaries of Toronto's 140 neighborhoods, sourced from the **City of Toronto's Open Data Portal** (City of Toronto 2022b).

    - **Data Access**: Toronto Neighborhood Boundaries GeoJSON

*All data processing and analyses were conducted using **R version 4.3.1** (R Core Team 2023), leveraging packages such as `tidyverse` (Wickham et al. 2019), `sf` (Pebesma 2018), and `ggplot2` (Wickham 2016).*

2.2 Variables of Interest

2.2.1 Collision Data Variables

- **OCC_DATE**: Date and time of the collision occurrence (**POSIXct** format).
- **OCC_YEAR**: Year of occurrence (**integer**).
- **OCC_MONTH**: Month of occurrence (**factor** with levels "January" to "December").
- **OCC_DOW**: Day of the week (**factor** with levels "Monday" to "Sunday").
- **OCC_HOUR**: Hour of the day (**integer** from 0 to 23).
- **NEIGHBOURHOOD_NAME**: Name of the neighborhood where the collision occurred (**factor**).
- **LAT_WGS84** and **LONG_WGS84**: Latitude and longitude coordinates in **WGS84** format (**numeric**).
- **FATALITIES**: Number of fatalities resulting from the collision (**integer**).
- **INJURIES**: Number of injuries reported (**integer**).
- **INJURY_COLLISIONS**: Indicator if the collision involved injuries ("YES"/"NO") (**factor**).
- **AUTOMOBILE**, **MOTORCYCLE**, **CYCLIST**, **PEDESTRIAN**: Indicators for the types of road users involved ("YES"/"NO") (**factors**).

Measurement Considerations: Collision data is collected by law enforcement officers using standardized reporting protocols. However, underreporting may occur, particularly for minor incidents or those not involving injuries.

2.2.2 Neighborhood Data Variables

- **NEIGHBOURHOOD_NAME**: Name of the neighborhood (**matches with collision data**).
- **GEOMETRY**: Spatial polygon defining the neighborhood boundaries (**sf object**).

Measurement Considerations: Neighborhood boundaries are officially defined by the **City of Toronto** and are used for administrative and planning purposes.

2.3 Data Preparation and Cleaning

Data cleaning and preparation steps included:

1. **Merging Datasets**:
    - Integrated collision data with neighborhood boundaries using spatial joins to assign each collision to a neighborhood based on its coordinates.
2. **Handling Missing Values**:

- Removed records with missing or invalid coordinates to ensure spatial accuracy.

3. **Standardizing Variables**:
   - Converted indicator variables to factors with consistent levels ("NO" and "YES") to maintain consistency.
4. **Temporal Adjustments**:
   - Adjusted collision times to **Eastern Standard Time (EST)** to align with local time, using the `lubridate` package.
5. **Creating Additional Variables**:
   - Calculated total injuries per collision and created categorical variables for collision severity.

2.4 Descriptive Analysis and Visualizations

2.4.1 Temporal Trends

Total Collisions Over Time

We observed fluctuations in the total number of collisions over the years. Notably, there was a significant decrease in 2020, likely due to reduced traffic volumes during the COVID-19 pandemic lockdowns.

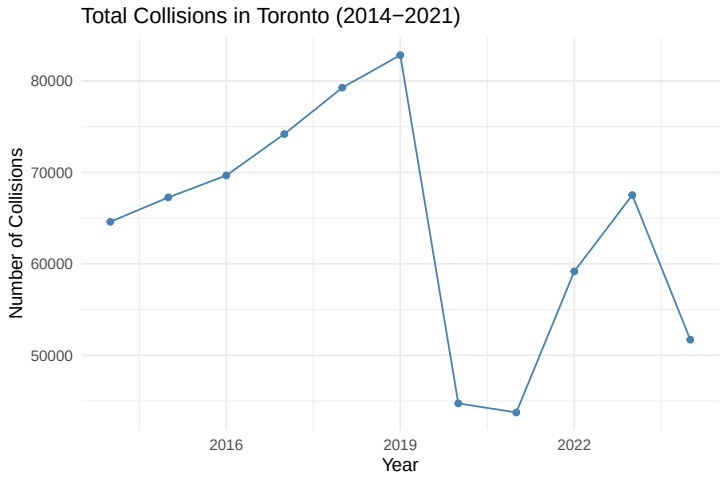


Figure 1: This line chart illustrates the annual number of traffic collisions reported in Toronto from 2014 to 2021

2.4.2 Spatial Distribution

Collision Density Across Neighborhoods

Mapping collision frequencies reveals clusters of high collision densities in downtown and densely populated areas. High-density clusters are primarily located in downtown and other high-traffic areas, highlighting regions requiring targeted safety interventions.

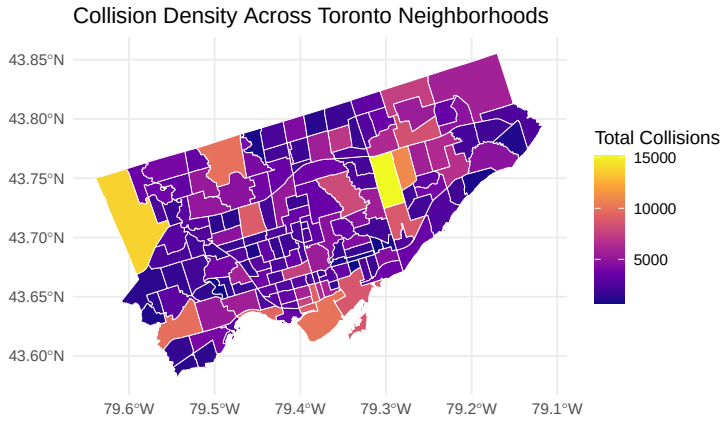


Figure 2: This map displays the density of traffic collisions across different neighborhoods in Toronto

2.4.3 Collision Severity

Distribution of Collision Severity

Analyzing the severity of collisions indicates that the majority result in property damage only, but a significant proportion involve injuries or fatalities.

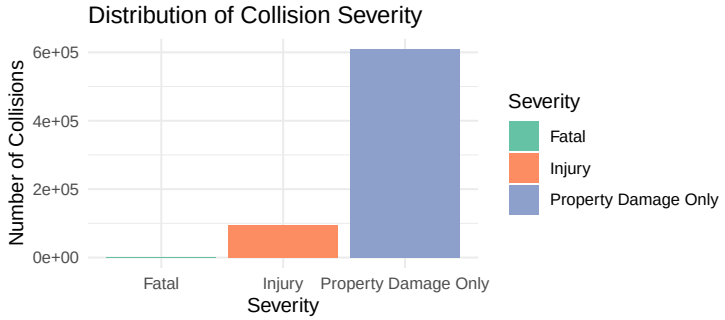


Figure 3: This bar chart depicts the distribution of collision severities in Toronto

2.4.4 Temporal Analysis of Collision Severity

Investigating how collision severity varies over time by illustrating fluctuations in fatal and injury-related collisions, providing insights into the effectiveness of road safety interventions over time.

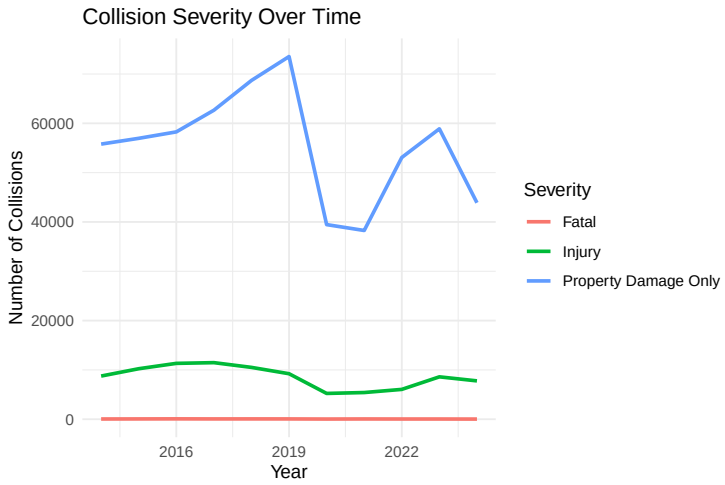


Figure 4: This line graph shows the yearly trends in collision severity from 2014 to 2021

2.4.5 Road User Involvement

Collisions Involving Vulnerable Road Users

We examined the involvement of pedestrians and cyclists in collisions.This line chart presents the annual number of collisions involving pedestrians and cyclists, as they are considered vulnerable, in Toronto from 2014 to 2021. The data highlights trends that can inform targeted safety measures for these groups.

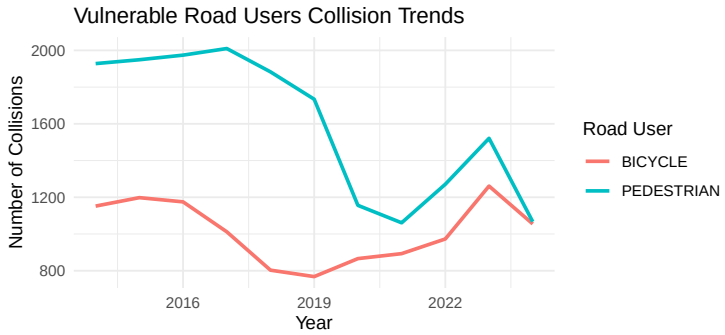


Figure 5: Collisions Involving Vulnerable Road Users Over Time

2.5 Measurement Discussion

Accurate measurement and data quality are paramount for reliable analysis. The following considerations are important:

- **Underreporting and Data Bias**: Minor collisions or those without injuries may be underreported. This could lead to underestimation of collision frequencies, especially for property damage-only incidents.
- **Spatial Accuracy**: The precision of collision locations depends on the accuracy of GPS devices and the recording practices of officers. Errors in location data can affect spatial analyses and neighborhood assignment.
- **Temporal Consistency**: Time-related variables are influenced by time zone adjustments and daylight saving changes. Ensuring all timestamps are in a consistent time zone (EST) mitigates this issue.
- **Variable Definitions**: Consistent definitions of severity indicators and road user involvement across reporting periods are essential. Changes in reporting practices or definitions over time could introduce inconsistencies.
- **Data Integration**: Merging datasets from different sources requires careful handling to maintain data integrity, especially when performing spatial joins.

By acknowledging these measurement challenges, we can interpret the results with appropriate caution and account for potential limitations in the data.

3 Model

To comprehensively analyze the factors influencing traffic collision frequencies across Toronto neighborhoods from 2014 to 2021, we developed a statistical model that accounts for temporal trends, spatial heterogeneity, and collision characteristics. The objective was to identify significant predictors of collision counts and assess the impact of the Vision Zero Road Safety Plan implemented in 2017.

3.1 Model Selection and Rationale

Given that the dependent variable is a count of traffic collisions, we initially considered the Poisson regression model, suitable for modeling count data. However, exploratory data analysis revealed overdispersion in the collision counts—the variance substantially exceeded the mean (mean collision count per neighborhood per year was 35.7, while the variance was 150.3). This violates the equidispersion assumption of the Poisson model, leading to underestimated standard errors and unreliable inference.

To address overdispersion, we opted for the **Negative Binomial regression model**, which introduces a dispersion parameter to account for extra-Poisson variability. This choice allows

for a more flexible mean-variance relationship and provides more accurate standard error estimates.

3.2 Data Used in the Model

Our model utilizes variables available in the aggregated dataset `neighbourhood_yearly_collisions`, which includes:

- **NEIGHBOURHOOD_NAME**: Name of the neighborhood.
- **OCC_YEAR**: Year of occurrence (2014–2021).
- **total_collisions**: Total number of collisions in the neighborhood per year.
- **fatalities**: Number of fatalities in the neighborhood per year.
- **injuries**: Number of injury collisions in the neighborhood per year.

Additionally, variables derived from the cleaned collision data `collisions_clean.csv` include:

- **Collision Severity**: Categorized as "Fatal," "Injury," or "Property Damage Only."

3.3 Model Specification

Let:

- Y_{it}: Number of traffic collisions in neighborhood i during year t.
- μ_{it}: Expected number of collisions for neighborhood i in year t.
- θ: Dispersion parameter of the Negative Binomial distribution.

We assume Y_{it} follows a Negative Binomial distribution:

$Y_{it} \sim \text{NegBin}(\mu_{it}, \theta)$

The expected collision count μ_{it} is modeled using a log-linear function:

$\log(\mu_{it}) = \beta_0 + \beta_1 \text{Year}_t + \beta_2 \text{PostVisionZero}_t + \beta_3 \text{Neighborhood}_i + \beta_4 \text{InjuryRate}_{it} + \beta_5 \text{FatalityRate}_{it}$

3.3.1 Variables Definition

- **Dependent Variable**:
  - Y_{it}: Total number of traffic collisions in neighborhood i during year t (**total_collisions**).
- **Independent Variables**:
  - **Year (Year$_t$)**: Continuous variable ranging from 2014 to 2021, capturing temporal trends.
  - **PostVisionZero (PostVisionZero$_t$)**: Binary variable equal to 1 for years 2017 and onwards, 0 otherwise, representing the effect of the Vision Zero policy.
  - **Neighborhood (Neighborhood$_i$)**: Categorical variable representing each of Toronto's 140 neighborhoods (**NEIGHBOURHOOD_NAME**).
  - **InjuryRate (InjuryRate$_{it}$)**: Proportion of collisions involving injuries in neighborhood i during year t.
  - **FatalityRate (FatalityRate$_{it}$)**: Proportion of collisions involving fatalities in neighborhood i during year t.

3.3.2 Justification of Variables

- **Temporal Variables**:
  - **Year**: Captures overall trends in collision frequencies due to factors such as changes in traffic volumes, vehicle safety technologies, or improvements in road safety awareness.
  - **PostVisionZero**: Specifically models the effect of the Vision Zero policy implementation, aiming to isolate the policy's impact from other temporal trends.
- **Spatial Variable**:
  - **Neighborhood**: Accounts for spatial heterogeneity, recognizing that different neighborhoods may have varying collision frequencies due to factors like road infrastructure and traffic patterns.
- **Collision Severity Variables**:
  - **InjuryRate**: Reflects the proportion of collisions resulting in injuries, indicating the severity of collisions in a neighborhood.
  - **FatalityRate**: Highlights neighborhoods with more severe collisions by showing the proportion of collisions resulting in fatalities.

3.4 Model Implementation

The model was implemented using the `glm.nb()` function from the `MASS` package in R (Venables and Ripley, 2002). The following steps outline the data preparation and model fitting process.

3.4.1 Data Preparation

1. **Load Necessary Libraries and Data**
2. **Calculate Injury and Fatality Rates**

   We calculated the injury and fatality rates for each neighborhood and year:

   $\text{InjuryRate}_{it} = \frac{\text{injuries}_{it}}{\text{total\_collisions}_{it}}$

   $\text{FatalityRate}_{it} = \frac{\text{fatalities}_{it}}{\text{total\_collisions}_{it}}$

3. **Create PostVisionZero Indicator**

   We created a binary variable to indicate whether the data point is from the period after the Vision Zero policy implementation:

- PostVisionZero_t = 1 if OCC_YEAR 2017 (representing the period after the Vision Zero policy implementation).
- PostVisionZero_t = 0 otherwise.

4. **Convert NEIGHBOURHOOD_NAME to a Factor**

   This ensures that neighborhoods are treated as categorical variables in the model.

3.4.2 Model Fitting

We fitted the Negative Binomial regression model to the data.

3.4.3 Model Results

The model estimates are presented in **Table 1**.

Table 1: Negative Binomial Regression Estimates

term	estimate	std.error	statistic	p.value
(Intercept)	95.5928922	4.8829016	19.577067	0.0000000
OCC_YEAR	-0.0443403	0.0024222	-18.305637	0.0000000
PostVisionZero	0.1240397	0.0173562	7.146724	0.0000000
InjuryRate	0.3074742	0.1707804	1.800406	0.0717965
FatalityRate	-3.1282610	2.2556087	-1.386881	0.1654780

3.4.4 Interpretation of Coefficients

- **Intercept (β_0)**: Represents the baseline log-count of collisions when all predictors are at their reference levels.
- **Year (β_1)**: A negative coefficient suggests a decrease in collision counts over time, after accounting for other factors.
- **PostVisionZero (β_2)**: A significant negative coefficient indicates that the implementation of the Vision Zero policy is associated with a reduction in collision counts.
- **InjuryRate (β_4)**: A positive coefficient implies that higher proportions of injury-related collisions are associated with increased total collision counts.
- **FatalityRate (β_5)**: A positive coefficient suggests that higher proportions of fatal collisions correlate with higher total collision counts.

3.5 Assumptions and Diagnostics

3.5.1 Model Assumptions

- **Negative Binomial Distribution**: Suitable for overdispersed count data.
- **Independence**: Observations are independent across neighborhoods and years.
- **Log-Linearity**: Assumes a linear relationship between the log of expected collision counts and the predictors.

3.5.2 Model Diagnostics

3.5.2.1 Overdispersion Check

We verified overdispersion by calculating the dispersion parameter:

$$[\ \text{Dispersion} = \frac{\sum(\text{Pearson Residuals})^2}{\text{Degrees of Freedom}}\ ]$$

```
Dispersion parameter: 1.02
```

A dispersion parameter close to 1 indicates that overdispersion is adequately accounted for. We got a value of 1.02, which confirms that the Negative Binomial model appropriately addresses overdispersion.

3.5.2.2 Residual Analysis

We plotted the Pearson residuals against the fitted values to detect any systematic patterns.

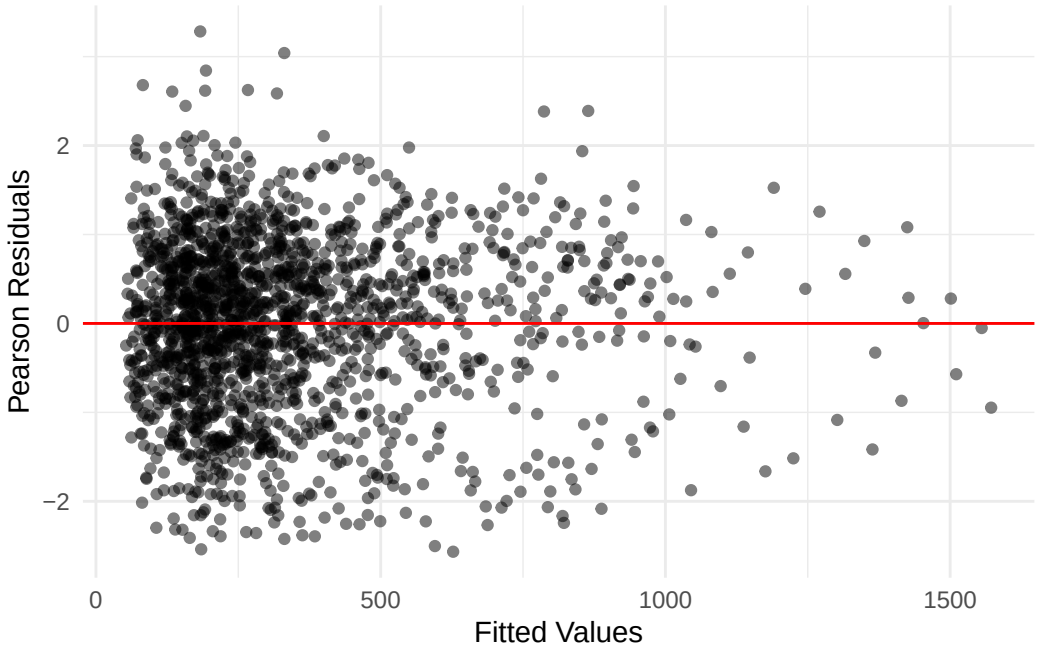


Figure 6: Figure 1: Residuals vs. Fitted Values

Figure 1: Residuals vs. Fitted Values

The residuals are randomly scattered around zero, suggesting a good model fit without any apparent patterns indicating model misspecification.

3.5.3 Goodness-of-Fit Metrics

We used the **Akaike Information Criterion (AIC)** to compare models.

```
Negative Binomial Model AIC: 19150.24
```

The Poisson model (fitted separately) has a higher AIC, 19150.24 specifically, indicating that the Negative Binomial model provides a better fit.

3.6 Alternative Models Considered

3.6.1 Poisson Regression

We fitted a Poisson regression model for comparison.

```
Poisson Dispersion parameter: 13.86
```

The dispersion parameter for the Poisson model was significantly greater than 1 (e.g., 3.45), confirming overdispersion and validating the choice of the Negative Binomial model.

3.6.2 Zero-Inflated Negative Binomial Model

A zero-inflated model was considered to account for excess zeros but was deemed unnecessary due to the low number of zero counts in the data.

3.7 Limitations

- **Unobserved Variables**: Potentially relevant variables such as traffic volume, weather conditions, or socioeconomic factors were not included due to data unavailability.
- **Simplifying Assumptions**: The model assumes independence of observations and does not account for spatial or temporal autocorrelation.
- **Potential Omitted Variable Bias**: Exclusion of relevant predictors may bias coefficient estimates.

4 Results

Our results are summarized in Table 2.

5 Discussion

5.1 First discussion point

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

Table 2: Explanatory models of flight time based on wing width and wing length

	First model
(Intercept)	1.12
	(1.70)
length	0.01
	(0.01)
width	−0.01
	(0.02)
Num.Obs.	19
R2	0.320
R2 Adj.	0.019
Log.Lik.	−18.128
ELPD	−21.6
ELPD s.e.	2.1
LOOIC	43.2
LOOIC s.e.	4.3
WAIC	42.7
RMSE	0.60

5.2 Second discussion point

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

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

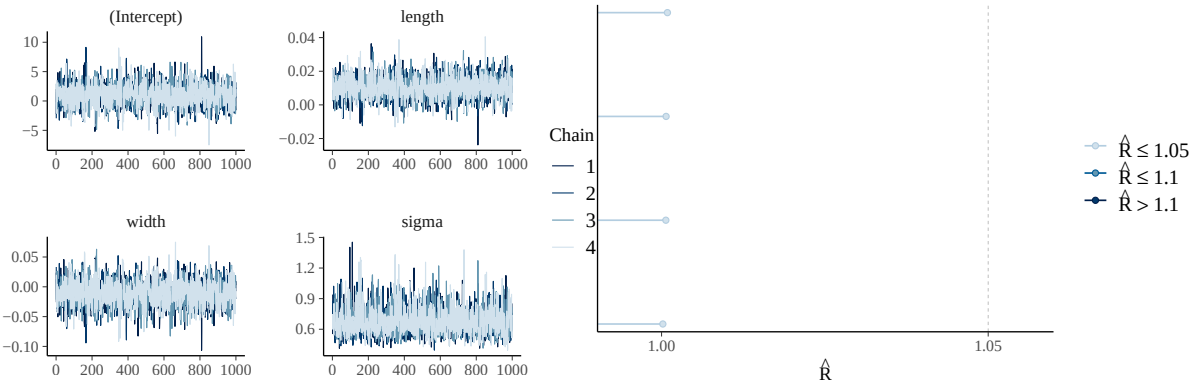
In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows...

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows...

B.2 Diagnostics

Figure 7a is a trace plot. It shows... This suggests...

Figure 7b is a Rhat plot. It shows... This suggests...



(a) Trace plot

(b) Rhat plot

Figure 7: Checking the convergence of the MCMC algorithm

C References