# Technical Report Summary: Analysis of Crash Data in New Zealand (2009)

Austin Snyder

2025-01-24

## Introduction

This report investigates crash data from New Zealand in 2009, focusing on incidents involving motorcyclists, pedestrians, and bicyclists. The primary goal is to provide actionable insights for improving road safety and informing policy interventions. Specifically, this analysis addresses two key questions:

- How dangerous is riding a motorcycle as a primary means of transportation?
- Are motorcycles the most common means of transportation involved in crashes on weekends?

To answer these questions, we analyzed crash data categorized by the hour of the day and the day of the week. Using statistical models like Poisson and Negative Binomial regression, we forecast crash counts and identified patterns and interactions among key variables.

### Data Overview

The dataset, sourced from the VGAM package in R, comprises three subsets:

- **Crashmc**: Crash counts involving motorcyclists.
- **Crashp**: Crash counts involving pedestrians.
- **Crashbc**: Crash counts involving bicyclists.

Each subset is structured as a matrix with:

- **Rows**: Hours of the day (24-hour format, 0–23).
- **Columns**: Days of the week (Monday to Sunday).

#### Data Preparation

The data was restructured into a long format to aid in analysis and visualization. A combined dataset was created, including:

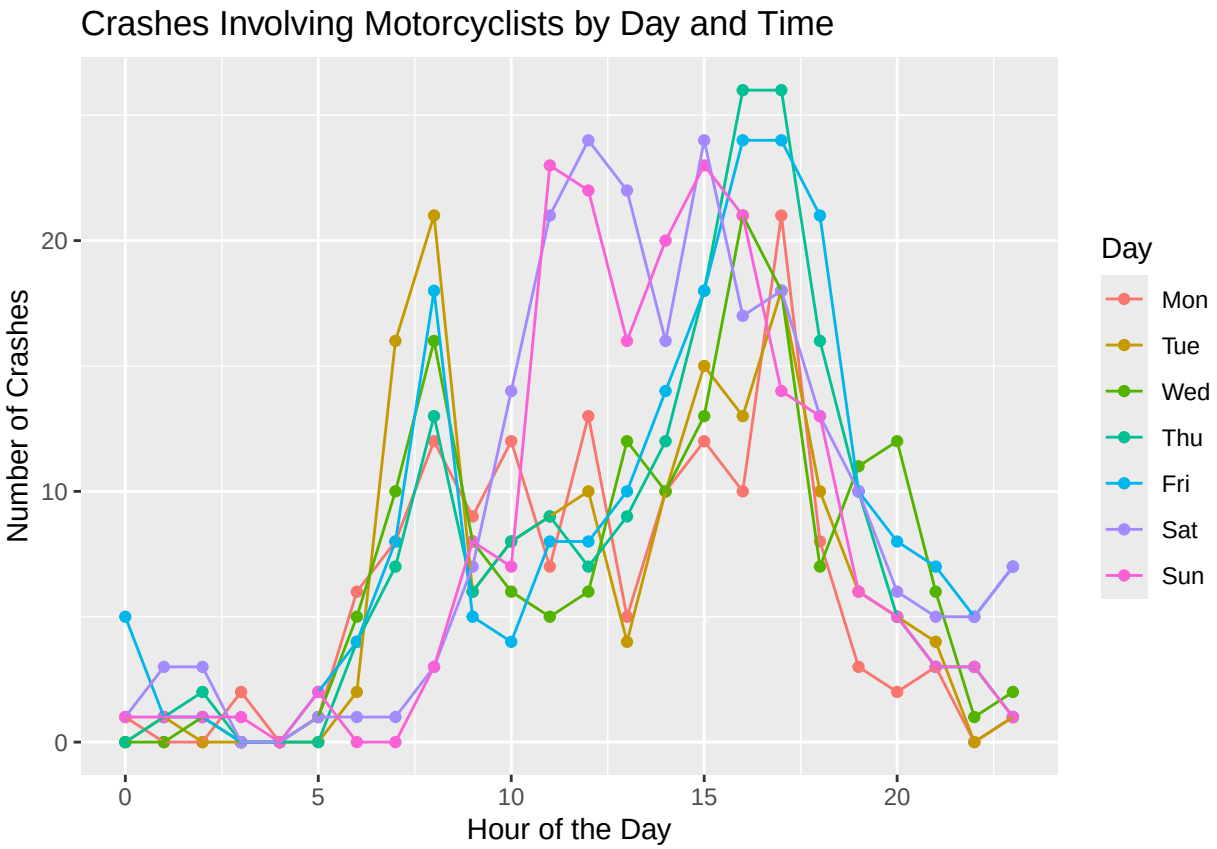
- **Type**: Mode of transportation (motorcyclist, pedestrian, bicyclist).
- **Day**: Day of the week (Monday to Sunday).
- **Hour**: Hour of the day (0–23).
- **Count**: Number of crashes recorded.

Below is a preview of the restructured data:

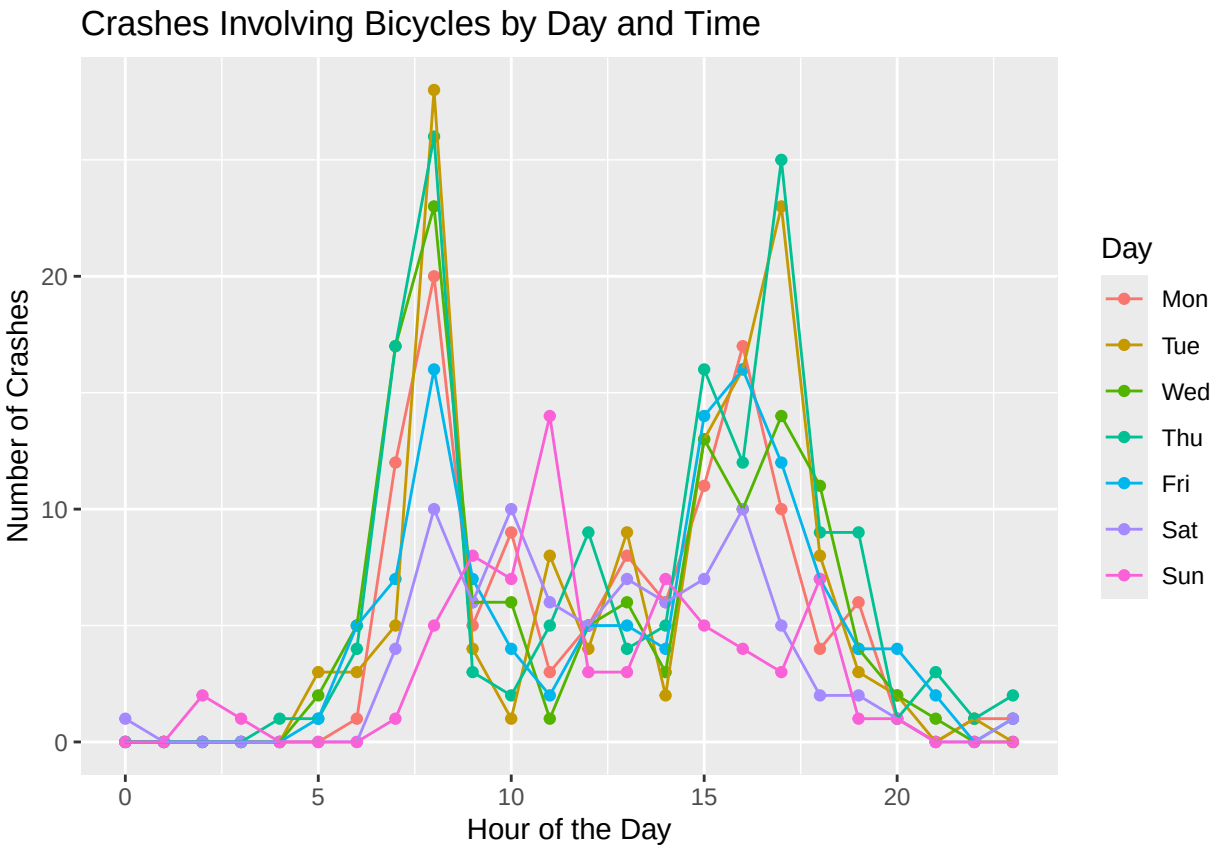
Table 1: Preview of Combined Dataset

| Count | Day | Hour | Type |
|---|---|---|---|
| 1 | Mon | 0 | Motorcyclist |
| 0 | Mon | 1 | Motorcyclist |
| 0 | Mon | 2 | Motorcyclist |
| 2 | Mon | 3 | Motorcyclist |
| 0 | Mon | 4 | Motorcyclist |
| 1 | Mon | 5 | Motorcyclist |

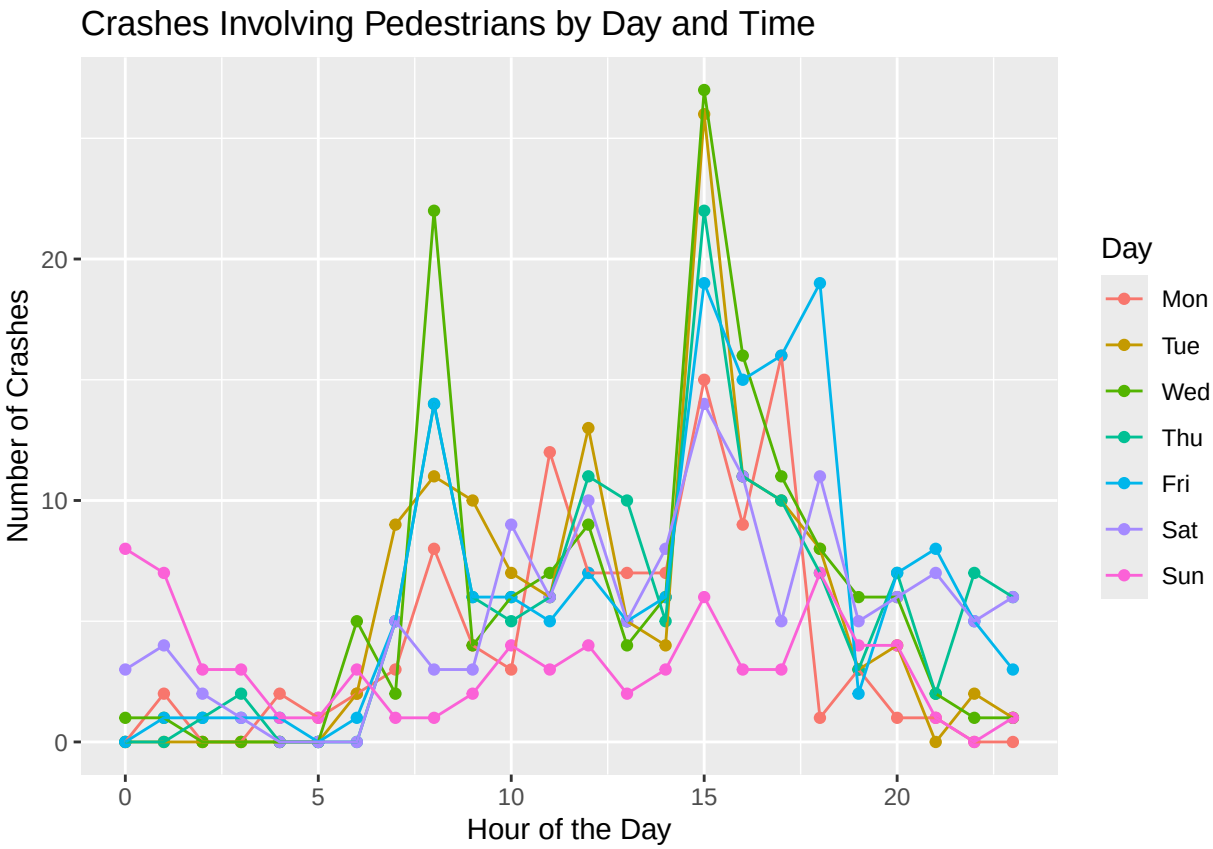
The restructured dataset enables detailed analysis of crash patterns by time, day, and type of transportation.




Here we see the patterns of crashes are relatively similar for weekdays, there are some noticeable differences between weekdays and weekends, specifically between the hours of 5 - 15.




Similar to the Motorcyclist data, the Bicycle data shows a similar pattern for weekdays, with some noticeable differences between weekdays and weekends.




The Pedestrian data shows a similar pattern to the Motorcyclist and Bicycle data, you can see weekend instances are higher in the first couple hours of the day.

# Methods

## Statistical Models

### Poisson Regression

Poisson regression is often used to model count data. This model assumes that the mean and variance of the response variable are equal. For our dataset, crash counts are the response variable, while predictors include:

- **Type**: Mode of transportation (motorcyclist, pedestrian, bicyclist).
- **DayGroup**: Weekdays (Monday–Friday) vs. weekends (Saturday–Sunday).
- **TimeCategory**: Divided into Early Morning (12–6 AM), Morning (6 AM–12 PM), Afternoon (12–6 PM), and Evening (6 PM–12 AM).

Interaction terms were included to capture combined effects of predictors:

$$\text{Count}_i \sim \text{Type}_i * \text{DayGroup}_i * \text{TimeCategory}_i$$

### Addressing Overdispersion

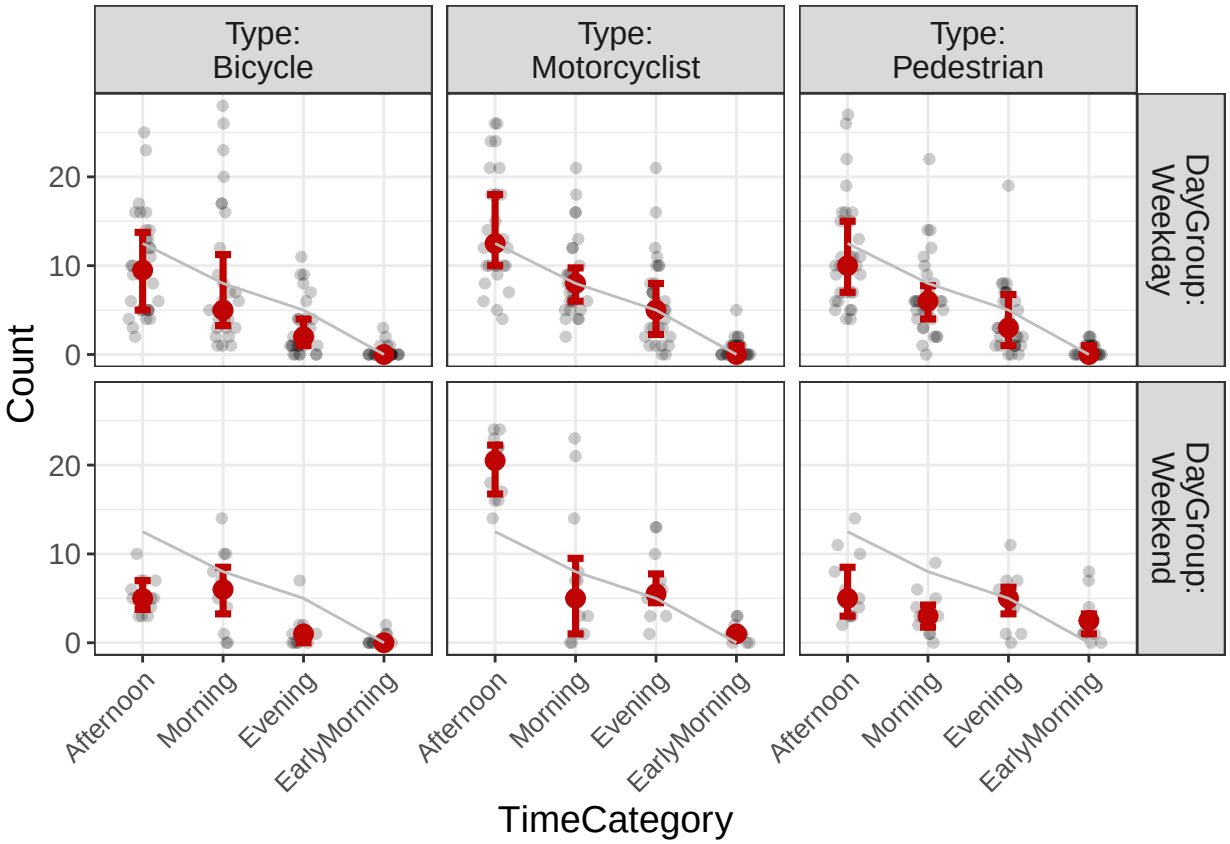
Initial diagnostics revealed overdispersion in the Poisson model, indicated by a residual deviance significantly higher than the degrees of freedom. Overdispersion occurs when the variance exceeds the mean, violating Poisson assumptions. To address this, we fitted a **Negative Binomial regression model**, which introduces a dispersion parameter to better handle variability in the data.

### Model Evaluation

Models were compared using the following metrics:

- **Akaike Information Criterion (AIC)**: Measures model fit, with lower values indicating better fit.
- **Bayesian Information Criterion (BIC)**: Similar to AIC but penalizes model complexity more strongly.
- **Mean Squared Prediction Error (MSPE)**: Evaluates predictive performance using k-fold cross-validation.

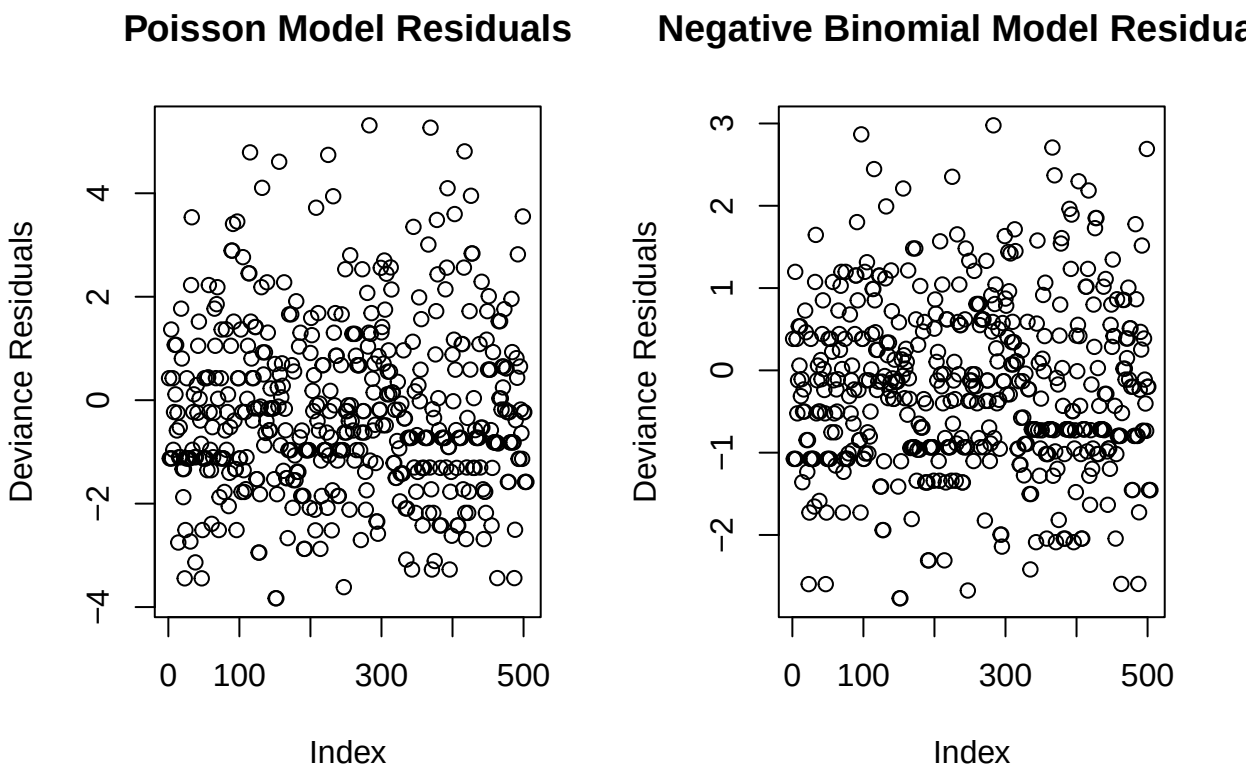
flexplot is used to visualize our restructured data. Flexplot is a package developed by Dr. Dustin Fife that reduces the friction of creating visualizations in R and leverages human strengths while reducing human biases.




We can see that there are some likely interactions between the type, day group, and time category. When modeling our poisson regression model we will include these interactions.

## Residual Diagnostics

To assess model fit, we plotted residuals for both the Poisson and Negative Binomial models:




Residual plots highlight that the Negative Binomial model provides a better fit, as deviations are more evenly distributed.

### Model Comparison

The table above summarizes model performance, with the Negative Binomial model demonstrating lower AIC and BIC values, indicating better fit. While the Poisson model shows a slightly lower MSPE, its inability to handle overdispersion justifies selecting the Negative Binomial model for predictive analysis.

## Results

### Question 1: How Dangerous is Riding a Motorcycle as a Primary Means of Transportation?

The analysis revealed that motorcyclists are at a higher risk of being involved in crashes compared to pedestrians and bicyclists. This risk is especially pronounced during commuting hours (morning and evening on weekdays) and recreational hours (afternoon on weekends). The table below summarizes the predicted crash counts for motorcyclists during these time periods:

Table 2: Predicted Crash Counts for Commuting and Recreational Hours

| Type | DayGroup | TimeCategory | PredictedCrashes |
|---|---|---|---|
| Bicycle | Weekday | Morning | 255 |
| Bicycle | Weekday | Evening | 88 |
| Bicycle | Weekend | Afternoon | 65 |
| Motorcyclist | Weekday | Morning | 260 |

| | | | |
|---|---|---|---|
| Motorcyclist | Weekday | Evening | 178 |
| Motorcyclist | Weekend | Afternoon | 237 |
| Pedestrian | Weekday | Morning | 196 |
| Pedestrian | Weekday | Evening | 124 |
| Pedestrian | Weekend | Afternoon | 74 |

The results indicate that motorcycles consistently exhibit higher predicted crash counts across these periods compared to bicycles and pedestrians.

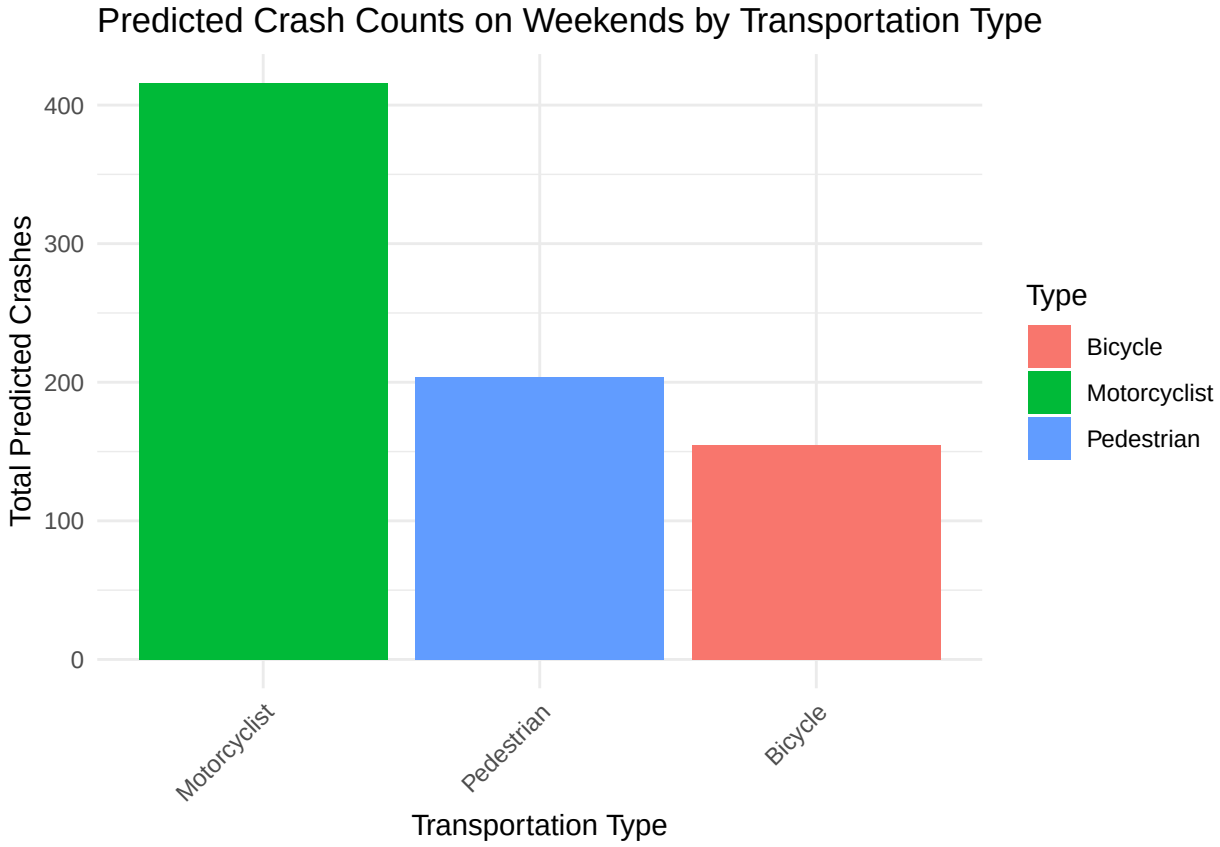
## Question 2: Are Motorcycles the Most Common Means of Transportation Involved in Crashes on Weekends?

Weekend crash data highlights that motorcycles are indeed the most common means of transportation involved in crashes, aside from cars. The total predicted crash counts for weekends by type of transportation are summarized below:

Table 3: Total Predicted Crash Counts for Weekends by Transportation Type

| Type | TotalPredictedCrashes |
|---|---|
| Motorcyclist | 416 |
| Pedestrian | 204 |
| Bicycle | 155 |

The bar plot below provides a visual comparison of the predicted crash counts:



The analysis predicts 416 crashes involving motorcyclists on weekends, reinforcing their higher risk compared to bicyclists and pedestrians.

### Overall Interpretation

Motorcyclists face significantly higher risks of crashes, particularly during recreational and commuting hours. The elevated risk on weekends underscores the need for targeted safety interventions, such as public awareness campaigns or enhanced traffic enforcement during high-risk periods.

### Conclusion

This analysis emphasizes the increased vulnerability of motorcyclists, particularly on weekends. The Negative Binomial model was utilized for predicting crash trends, as it effectively addressed overdispersion in the data. While the findings offer valuable insights, the analysis is limited by the lack of variables such as weather conditions, road characteristics, and demographics.

### Future Work

To build upon this analysis, the following areas should be explored:

- **Improved Data Collection**: Collect additional variables, including traffic flow, weather, road conditions, speed limits, and crosswalk presence, to enhance model accuracy and reliability.
- **Policy Recommendations**: Develop targeted safety measures for motorcyclists, with a focus on high-risk times and days identified in this study.
- **Model Refinements**: Explore dynamic models, such as time series, to forecast crash risks in real-time and assess the impact of interventions over time.

## Apendicies

#### Poisson Regression code and output

```
mod_grouped <- glm(Count ~ Type* DayGroup * TimeCategory, data = combined_data_grouped, family = "poisso
summary(mod_grouped)
```

```
##
## Call:
## glm(formula = Count ~ Type * DayGroup * TimeCategory, family = "poisson",
##     data = combined_data_grouped)
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                         -1.3218     0.3536
## TypeMotorcyclist                                     0.8650     0.4215
## TypePedestrian                                       0.5596     0.4432
## DayGroupWeekend                                      0.2231     0.6124
## TimeCategoryMorning                                  3.4618     0.3591
## TimeCategoryAfternoon                                3.6310     0.3582
## TimeCategoryEvening                                  2.3979     0.3693
```

```
## TypeMotorcyclist:DayGroupWeekend                            0.3878     0.7064
## TypePedestrian:DayGroupWeekend                              1.5506     0.6905
## TypeMotorcyclist:TimeCategoryMorning                       -0.8456     0.4306
## TypePedestrian:TimeCategoryMorning                         -0.8228     0.4533
## TypeMotorcyclist:TimeCategoryAfternoon                     -0.5423     0.4282
## TypePedestrian:TimeCategoryAfternoon                       -0.4411     0.4502
## TypeMotorcyclist:TimeCategoryEvening                       -0.1606     0.4411
## TypePedestrian:TimeCategoryEvening                         -0.2167     0.4646
## DayGroupWeekend:TimeCategoryMorning                        -0.5854     0.6269
## DayGroupWeekend:TimeCategoryAfternoon                      -0.8429     0.6275
## DayGroupWeekend:TimeCategoryEvening                        -1.0761     0.6731
## TypeMotorcyclist:DayGroupWeekend:TimeCategoryMorning       -0.1925     0.7296
## TypePedestrian:DayGroupWeekend:TimeCategoryMorning         -1.8612     0.7245
## TypeMotorcyclist:DayGroupWeekend:TimeCategoryAfternoon      0.5832     0.7241
## TypePedestrian:DayGroupWeekend:TimeCategoryAfternoon       -1.5394     0.7155
## TypeMotorcyclist:DayGroupWeekend:TimeCategoryEvening        0.5435     0.7718
## TypePedestrian:DayGroupWeekend:TimeCategoryEvening         -0.5585     0.7618
##                                                          z value Pr(>|z|)
## (Intercept)                                               -3.738 0.000185 ***
## TypeMotorcyclist                                           2.052 0.040134 *
## TypePedestrian                                             1.263 0.206710
## DayGroupWeekend                                            0.364 0.715565
## TimeCategoryMorning                                        9.641  < 2e-16 ***
## TimeCategoryAfternoon                                     10.137  < 2e-16 ***
## TimeCategoryEvening                                        6.494 8.38e-11 ***
## TypeMotorcyclist:DayGroupWeekend                           0.549 0.583074
## TypePedestrian:DayGroupWeekend                             2.246 0.024720 *
## TypeMotorcyclist:TimeCategoryMorning                      -1.964 0.049552 *
## TypePedestrian:TimeCategoryMorning                        -1.815 0.069496 .
## TypeMotorcyclist:TimeCategoryAfternoon                    -1.267 0.205297
## TypePedestrian:TimeCategoryAfternoon                      -0.980 0.327195
## TypeMotorcyclist:TimeCategoryEvening                      -0.364 0.715905
## TypePedestrian:TimeCategoryEvening                        -0.466 0.640960
## DayGroupWeekend:TimeCategoryMorning                       -0.934 0.350378
## DayGroupWeekend:TimeCategoryAfternoon                     -1.343 0.179156
## DayGroupWeekend:TimeCategoryEvening                       -1.599 0.109856
## TypeMotorcyclist:DayGroupWeekend:TimeCategoryMorning      -0.264 0.791865
## TypePedestrian:DayGroupWeekend:TimeCategoryMorning        -2.569 0.010195 *
## TypeMotorcyclist:DayGroupWeekend:TimeCategoryAfternoon     0.805 0.420566
## TypePedestrian:DayGroupWeekend:TimeCategoryAfternoon      -2.152 0.031423 *
## TypeMotorcyclist:DayGroupWeekend:TimeCategoryEvening       0.704 0.481282
## TypePedestrian:DayGroupWeekend:TimeCategoryEvening        -0.733 0.463457
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3205.2  on 503  degrees of freedom
## Residual deviance: 1333.2  on 480  degrees of freedom
## AIC: 2798.3
##
## Number of Fisher Scoring iterations: 6
```

**Negative Binomial Regression code and output**

```
mod_grouped_nb <- glm.nb(Count ~ Type * DayGroup * TimeCategory, data = combined_data_grouped)
summary(mod_grouped_nb)
```

```
##
## Call:
## glm.nb(formula = Count ~ Type * DayGroup * TimeCategory, data = combined_data_grouped,
##     init.theta = 3.244067105, link = log)
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                         -1.3218     0.3678
## TypeMotorcyclist                                     0.8650     0.4452
## TypePedestrian                                       0.5596     0.4658
## DayGroupWeekend                                      0.2231     0.6411
## TimeCategoryMorning                                  3.4618     0.3866
## TimeCategoryAfternoon                                3.6310     0.3858
## TimeCategoryEvening                                  2.3979     0.3961
## TypeMotorcyclist:DayGroupWeekend                     0.3878     0.7556
## TypePedestrian:DayGroupWeekend                       1.5506     0.7407
## TypeMotorcyclist:TimeCategoryMorning                -0.8456     0.4759
## TypePedestrian:TimeCategoryMorning                  -0.8228     0.4965
## TypeMotorcyclist:TimeCategoryAfternoon              -0.5423     0.4738
## TypePedestrian:TimeCategoryAfternoon                -0.4411     0.4937
## TypeMotorcyclist:TimeCategoryEvening                -0.1606     0.4855
## TypePedestrian:TimeCategoryEvening                  -0.2167     0.5069
## DayGroupWeekend:TimeCategoryMorning                 -0.5854     0.6819
## DayGroupWeekend:TimeCategoryAfternoon               -0.8429     0.6824
## DayGroupWeekend:TimeCategoryEvening                 -1.0761     0.7245
## TypeMotorcyclist:DayGroupWeekend:TimeCategoryMorning -0.1925     0.8223
## TypePedestrian:DayGroupWeekend:TimeCategoryMorning  -1.8612     0.8177
## TypeMotorcyclist:DayGroupWeekend:TimeCategoryAfternoon  0.5832     0.8175
## TypePedestrian:DayGroupWeekend:TimeCategoryAfternoon -1.5394     0.8098
## TypeMotorcyclist:DayGroupWeekend:TimeCategoryEvening  0.5435     0.8600
## TypePedestrian:DayGroupWeekend:TimeCategoryEvening  -0.5585     0.8510
##                                                    z value Pr(>|z|)
## (Intercept)                                         -3.594 0.000326 ***
## TypeMotorcyclist                                     1.943 0.052011 .
## TypePedestrian                                       1.201 0.229603
## DayGroupWeekend                                      0.348 0.727778
## TimeCategoryMorning                                  8.954  < 2e-16 ***
## TimeCategoryAfternoon                                9.411  < 2e-16 ***
## TimeCategoryEvening                                  6.053 1.42e-09 ***
## TypeMotorcyclist:DayGroupWeekend                     0.513 0.607836
## TypePedestrian:DayGroupWeekend                       2.093 0.036315 *
## TypeMotorcyclist:TimeCategoryMorning                -1.777 0.075613 .
## TypePedestrian:TimeCategoryMorning                  -1.657 0.097521 .
## TypeMotorcyclist:TimeCategoryAfternoon              -1.145 0.252304
## TypePedestrian:TimeCategoryAfternoon                -0.893 0.371656
## TypeMotorcyclist:TimeCategoryEvening                -0.331 0.740879
## TypePedestrian:TimeCategoryEvening                  -0.427 0.669062
## DayGroupWeekend:TimeCategoryMorning                 -0.859 0.390568
```

```
## DayGroupWeekend:TimeCategoryAfternoon                   -1.235 0.216737
## DayGroupWeekend:TimeCategoryEvening                     -1.485 0.137471
## TypeMotorcyclist:DayGroupWeekend:TimeCategoryMorning    -0.234 0.814876
## TypePedestrian:DayGroupWeekend:TimeCategoryMorning      -2.276 0.022840 *
## TypeMotorcyclist:DayGroupWeekend:TimeCategoryAfternoon   0.713 0.475539
## TypePedestrian:DayGroupWeekend:TimeCategoryAfternoon    -1.901 0.057293 .
## TypeMotorcyclist:DayGroupWeekend:TimeCategoryEvening     0.632 0.527355
## TypePedestrian:DayGroupWeekend:TimeCategoryEvening      -0.656 0.511615
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(3.2441) family taken to be 1)
##
##     Null deviance: 1325.79  on 503  degrees of freedom
## Residual deviance:  548.67  on 480  degrees of freedom
## AIC: 2446.8
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  3.244
##           Std. Err.:  0.370
##
##  2 x log-likelihood:  -2396.801
```