# Toronto Collision Case Study Report

## Introduction:

Traffic collisions represent a significant global public safety concern, contributing to injuries, fatalities, and substantial economic losses every year. According to the World Health Organization (2023), road traffic injuries are the leading cause of death for individuals aged 5–29, with over 1.3 million deaths annually. In Canada, motor vehicle collisions remain a pressing urban challenge, particularly in metropolitan areas such as Toronto where dense traffic, diverse road users, and complex infrastructure create heightened risks (Transport Canada, 2023).

Understanding patterns in collision frequency, severity, and contributing factors is crucial for developing targeted traffic safety policies, optimizing law enforcement deployment, and informing public awareness campaigns. Previous studies have shown that temporal variables such as time of day and day of week, as well as spatial factors like police division boundaries, can significantly influence collision patterns (Chen et al., 2019; Elvik & Bjørnskau, 2019). By applying statistical methods to real-world collision data, cities can proactively implement measures to reduce accidents and improve road safety outcomes

## Dataset and Source

This study uses the **Toronto Traffic Collisions** dataset compiled by Ahapov (n.d.) and publicly available on Kaggle. The dataset originates from the Toronto Police Service’s open data platform and contains detailed, event-level records of reported motor vehicle collisions in the city of Toronto. It provides both spatial and temporal information, enabling a comprehensive analysis of collision patterns and risk factors.

Key variables include:

* **Date and time** of collision
* **Police Division** of occurrence
* **Collision Severity** (Severe / Non-Severe)
* **Number of Injuries**
* **Location Attributes**
* **Vehicle and Road User Involvement**
* **Year, Day of Week, and Hour**
* **Collision Types**

This dataset spans multiple years, allowing for both cross-sectional and longitudinal analyses of traffic collision trends. Prior to analysis, data cleaning steps included removing duplicates, handling missing values, and converting variables to appropriate formats for statistical testing.

## Research Questions

1. **RQ1:** Do daily collision counts differ significantly across Toronto police divisions?
2. **RQ2:** Is there a significant association between collision severity (severe vs. non-severe) and the day of the week?
3. **RQ3:** How do time of day, day of week, and year predict the number of injury collisions?

## Hypotheses

* **H1:** Mean daily collision counts will vary significantly across police divisions, reflecting differences in traffic density, infrastructure, and enforcement levels (Elvik & Bjørnskau, 2019).
* **H2:** Collision severity will be significantly associated with the day of the week, with weekends showing higher proportions of severe collisions due to factors like impaired driving and speeding (Chen et al., 2019).
* **H3:** Time of day, day of week, and year will significantly predict the number of injury collisions, with peak traffic hours contributing to higher counts (Transport Canada, 2023).

## Data Preparation and Exploration

Before conducting statistical analyses, the dataset underwent a series of preprocessing steps to ensure accuracy and consistency:

1. **Data Import and Inspection** – The dataset was imported into R, and variable types were reviewed to ensure correct formats (e.g., dates converted to Date objects, categorical variables converted to factors).
2. **Duplicate Removal** – Duplicate collision records were identified and removed based on unique event IDs (EVENT\_UNIQUE\_ID).
3. **Missing Data Handling** – Missing values in key variables (e.g., division, collision severity, injury counts) were checked. Records missing critical outcome variables were excluded, while non-critical missing values were left intact to preserve sample size.
4. **Variable Recoding** –
   * OCC\_DOW was recorded as an ordered factor to reflect the natural progression of days from Monday to Sunday.
   * DIVISION was standardized for consistency in labels.
   * SEVERITY was derived from a combination of fatality and injury data into a binary “Severe” vs. “Non-Severe” category.
5. **Feature Creation** – Aggregated measures were calculated, such as mean daily collision counts per police division, and injury counts by hour and day of week, to support the research questions.

## Descriptive Statistics

The dataset contains **499,476** reported traffic collisions in Toronto.

* The **meantime** of collisions was 1:30 PM (*M* = 13.5, *SD* = 4.92), with the median at 2:00 PM.
* **Fatalities** were rare, averaging less than one per 1,000 collisions (*M* = 0.00075), with most collisions resulting in zero fatalities.
* **Collision Severity**: The majority of incidents (86.1%) were classified as non-severe (n = 430,284), while 13.9% were severe (n = 69,192).
* **Day-of-Week Trends**: Collisions peaked on Fridays (n = 83,000) and were least frequent on Sundays (n = 49,531).
* **Divisional Distribution**: The NSA division reported the highest collision counts (n = 69,997), followed by D32 (n = 41,363), while D13 reported the fewest (n = 15,884).

## Exploratory Visualizations

To support initial insights, several plots were generated:

* **Bar Plot:** Mean daily collisions by police division, ordered from lowest to highest.
* **Stacked Bar Chart:** Distribution of severe vs. non-severe collisions across days of the week.
* **Heatmap:** Mean daily injury counts by hour of day and day of week.
* **Trend Plot:** Yearly collision trends to examine potential long-term changes.

These exploratory steps provided a strong foundation for selecting appropriate statistical tests to address the research questions.

## Statistical Analysis

**RQ1: Do mean daily collision counts differ across Toronto Police Divisions?**

**Test and Model Used**  
Initially, a one-way ANOVA was applied to compare mean daily collision counts across 17 police divisions. Assumptions were checked using Shapiro–Wilk’s test for normality and Levene’s test for homogeneity of variances. Due to significant violations of both assumptions, a non-parametric Kruskal–Wallis’s rank-sum test was performed. Post-hoc pairwise comparisons were conducted using Dunn’s test with Holm correction.

**Results**  
The ANOVA indicated a significant difference in mean daily collision counts, *F*(16, 46,269) = 1562, *p* < .001. Assumption checks revealed non-normal residuals (*W* = 0.888, *p* < .001) and heterogeneity of variances, *F*(16, 46,269) = 710.91, *p* < .001. The Kruskal–Walli’s test confirmed significant differences, χ²(16) = 13,513, *p* < .001. Dunn’s post-hoc tests identified numerous significant differences, including D13 reporting significantly fewer collisions than D14 (*p* < .001).

**Interpretation**  
Collision frequency is not uniform across Toronto’s police divisions. Higher daily counts in certain divisions may be linked to higher traffic volumes, denser urban infrastructure, or differences in enforcement and reporting practices.

**RQ2: Is collision severity associated with the day of the week?**

**Test and Model Used**  
A chi-square test of independence was used to examine the relationship between collision severity (Severe vs. Non-Severe) and day of the week. Effect size was measured using Cramér’s *V*.

**Results**  
The association between severity and day of the week was statistically significant, χ²(6, N = 499,476) = 41.82, *p* < .001, but the effect size was negligible (Cramér’s *V* = 0.009).

**Interpretation**  
While the results suggest a relationship between collision severity and day of the week, the practical significance is minimal. Severe collisions occur at relatively similar proportions throughout the week.

**RQ3: Do time of day, day of week, and year predict the number of injuries per collision?**

**Test and Model Used**  
A negative binomial regression model was applied to predict injury counts based on hour of occurrence, day of the week, and year. This model was chosen to account for the overdispersion observed in count data.

**Results**  
The model was significant, 2 × log-likelihood = –156,275.58, with a dispersion parameter of 1.02, indicating no substantial overdispersion. Several early morning hours (1:00 AM, 2:00 AM, 3:00 AM) showed significantly fewer injuries compared to the midnight reference category (*p* < .01).

**Interpretation**  
Time of day, day of week, and year all significantly influence injury counts in collisions. Lower injury rates during early morning hours may be related to reduced traffic density, while higher counts during peak hours may reflect increased exposure and risk.

## Conclusion

This analysis of Toronto’s traffic collision data revealed notable differences in collision frequency across police divisions, a statistically significant yet negligible association between collision severity and day of the week, and strong predictive effects of time, day, and year on injury counts. Specifically, certain divisions consistently reported higher daily collision counts, reflecting differences in traffic density, road network design, and local driving patterns. Injury counts were influenced by temporal factors, with lower counts during early morning hours and elevated counts during high-traffic periods.

## Limitations

1. **Reporting bias** – The dataset includes only reported collisions, which may omit minor or unreported incidents.
2. **Missing contextual factors** – External variables such as weather, road surface conditions, and seasonal effects were not incorporated into the analysis.
3. **Spatial granularity** – Some location data were aggregated to division-level, limiting the ability to detect finer spatial patterns.
4. **Static time frame** – The analysis does not account for potential long-term changes in traffic policy, road design, or population growth.

## Suggestions for Future Research

1. **Integrate additional datasets** – Combine collision data with weather, traffic volume, and road infrastructure information for richer insights.
2. **Conduct longitudinal analysis** – Track changes over multiple years to evaluate the effects of policy interventions or infrastructure upgrades.
3. **Explore advanced modeling techniques** – Use spatial-temporal models to capture both location-based and time-based collision trends more accurately.
4. **Assess behavioral factors** – Examine driver demographics, traffic enforcement activity, and public awareness campaigns to better understand collision causes.

**References**

Ahapov, A. (n.d.). *Toronto Traffic Collisions Dataset*. Kaggle. Retrieved from <https://www.kaggle.com/datasets/andriyahapov/traffic-collisions-toronto>

Chen, H., Zhang, Y., & Wang, C. (2019). Spatiotemporal analysis of traffic crashes and contributing factors in urban areas. *Accident Analysis & Prevention, 123*, 312–321. https://doi.org/10.1016/j.aap.2018.11.012

Elvik, R., & Bjørnskau, T. (2019). Safety-in-numbers: A systematic review and meta-analysis of evidence. *Safety Science, 117*, 39–49. https://doi.org/10.1016/j.ssci.2019.03.003

Transport Canada. (2023). *Canadian Motor Vehicle Traffic Collision Statistics*. Government of Canada. Retrieved from https://tc.canada.ca/

World Health Organization. (2023). *Global status report on road safety 2023*. Retrieved from <https://www.who.int/publications/i/item/9789240077610>