DESAUTELS
Faculty of Management
McGill University

# Proof of Concept

## Predicting Collision Risk in Montreal: Identifying High-Risk Conditions and Locations

**Team: The Foundation Lab**

**Multivariate Statistical Analysis**

MGSC-661


**Students:**

Ellie Ha

Rachelle Dong

Zacharie Houle

Muhammad Hydarali

Ibukunoluwa Adeleye

## I. Research Motivation

Between 2012 and 2021, Montreal recorded an average of 21 deaths per year due to vehicle collisions, highlighting the urgent need for smarter prevention strategies. While several studies have analyzed accident data, most take a reactive approach by explaining how severe crashes are once they happen rather than predicting where and when they are likely to occur. Past research, particularly one from Muktar & Fono (2024), also tends to stay broad, identifying general factors such as weather or road conditions without specifying which neighborhoods or specific conditions drive risk. For more detailed comparison, refer to Appendix 1.

Our project, The Foundation Lab, takes a predictive and spatial approach. Instead of studying severity, we model the likelihood of collisions across Montreal, identifying the exact conditions, surfaces, and times that increase risk. We also break the analysis down by neighborhood and postal code, pinpointing red zones and actionable insights that can help prevent collisions before they occur. The value of this approach extends beyond research, as our findings can help the City of Montreal target infrastructure improvements, guide emergency resource planning, and improve safety for drivers, cyclists, and pedestrians alike.

## II. Tentative Hypotheses

Prior to conducting preliminary SAAQ data observations, our hypotheses focus on geospatial and infrastructure variables influencing severe crashes (Class 3: severe injuries; Class 4: fatalities), with plans to control for temporal factors like time of day. These hypotheses are informed by gaps in existing literature, which highlight reactive analyses and broad categorizations (e.g., road conditions, proximity to landmarks) without specific predictive insights.

**Hypothesis 1: Neighborhood-Specific Risks** : The probability of severe crashes is higher in Montreal neighborhoods with high population density, urban services, and major transit activity. Key areas include Downtown (central congestion), High Services and Density Hubs (commercial/residential zones), and Gateways/Major Transit Corridors (high-traffic entry points like highways). Testing this identifies red zones at neighborhoods levels for targeted interventions like signage or traffic calming.

**Hypothesis 2: Cracked Road Conditions** : The probability of severe crashes is higher on roads with visible cracks, reducing traction and stability, especially at varying speeds. This extends literature by specifying crack severity levels, enabling optimized repairs and shifting from reactive to predictive maintenance.

**Hypothesis 3: Flat Curved Road Geometry** : The probability of severe crashes is higher on flat curved roads, where drivers underestimate turns, leading to loss of control. This explores road configuration interactions with behavior, supporting actions like curve warnings, or speed limits.

These hypotheses will be tested via ordinal classification, aiming to reduce annual deaths through policy impacts like infrastructure upgrades and emergency resource allocation.

## III. Dataset Introduction

The dataset, published by Données Québec and originally compiled by the Société de l'assurance automobile du Québec (SAAQ), Québec's road safety authority, records over 200,000 road collisions in Montréal from 2012 to 2021, with updates extending to 2022.

The target variable classifies collision severity into four levels: property damage only (Class 1), minor injury (Class 2), serious injury (Class 3), and fatal collision (Class 4).

Each record includes features that are relevant to the target variable, such as temporal, geospatial, environmental, and infrastructural features, as well as collision type and authorized speed, enabling a comprehensive analysis of both human and environmental determinants of collision severity. For more details on feature variables, refer to Appendix 2.

Before modeling, the dataset underwent correlation and variance analysis to retain only significant predictors. Coordinates were reprojected to match official neighborhood boundaries; categorical variables (e.g., weather, lighting, surface condition) were encoded, and continuous features (e.g., speed limit, time of day) were normalized to ensure model consistency.

Extensive data preprocessing was performed. Eleven columns with over 50% missing values were discarded, and all French column names were translated and standardized into English. To address class imbalance, given the lower frequency of serious and fatal collisions (Classes 3 and 4), robust modeling techniques insensitive to imbalance were employed. Geospatial inconsistencies were resolved by constructing a new "neighbourhood" variable using postal codes and spatial joins, facilitating regional comparisons across areas such as Downtown Montréal and Laval.

This refined dataset enhances the models' ability to identify meaningful patterns in collision severity and contributing factors, while preprocessing ensures data integrity and interpretability.

## IV. Preliminary Analysis

Analysis of the dataset uncovers key patterns in crash severity, with Class 1 (Property Damage) and Class 2 (Minor Injury) incidents dominating, indicating current measures effectively limit escalation to severe outcomes. However, specific high-risk factors emerge, warranting deeper study.

**The "Good Conditions Paradox"**

Surprisingly, severe (Class 3) and fatal (Class 4) crashes predominantly occur under "safe" conditions:

- **Road Geometry:** Straight, flat sections, not complex curves or hills.
- **Road Condition:** Well-maintained surfaces, with degraded conditions nearly absent in severe cases.
- **Environment:** Clear weather and daylight hours.

This suggests risk stems less from obvious hazards and more from exposure, as most driving occurs in these favorable conditions, potentially masking underlying factors.

**High-Risk Spatio-Temporal Zones**

Spatial analysis identifies hotspots with elevated severe crash rates:

- **Neighborhoods:** Rivière-des-Prairies (H1G: residential/waterfront), Centre-Sud (H2K: downtown nightlife), H1H (eastern industrial-residential), Ville Saint-Laurent (H4R: suburban), and Plateau-Mont-Royal East (H2L: trendy area).
- **Temporal Peaks:** Risk rises during rush hour (e.g., 7-9 AM, 4-6 PM) and nighttime, serving as a control variable.

These findings highlight a paradox linking "safe" infrastructure to severe outcomes, particularly in high-risk zones. Further analysis will explore whether this pattern holds lower-severity cases across all neighborhoods, informing the modeling phase.

## V. Project Feasibility

The project is highly feasible, blending a robust analytical pipeline with strong policy relevance. The workflow will comprise two stages: Evaluation & Validation, ensuring rigor and reliability.

For modeling, linear regression, XGBoost, and LightGBM will assess feature impact on severity, chosen for their ability to handle nonlinear interactions efficiently. Class imbalance (rare Class 3/4 incidents) will be addressed with SMOTE oversampling and weighted losses, with performance to be evaluated via F1 score and Recall to prioritize accurate severe crash (Class 3: serious injury; Class 4: fatal) detection.

Evaluation & Validation will use correlation and variance analyses will to confirm feature contributions and reduce redundancy, followed by comparing model predictions to real world severity distributions. This will validate the model's capture of spatial and infrastructural dynamics, revealing whether well maintained infrastructure alone reduces severity or if context matters.

This scalable framework will bridge data science and policy, guiding resource allocation and planning. Key interventions will include enhanced speed control on straight, high traffic roads and improved lighting in dense/nightlife areas, aiming to reduce severe crashes based on validated patterns. For more details on the workflow stage, refer to Appendix 3.

## REFERENCES

Société de l'assurance automobile du Québec (SAAQ). (2023). Collisions routières – Ville de Montréal [Data set]. Données Québec. https://www.donneesquebec.ca/recherche/dataset/vmtl-collisions-routieres.

Muktar, B., & Fono, V. (2024). Toward Safer Roads: Predicting the Severity of Traffic Accidents in Montreal Using Machine Learning. Electronics, 13(15), 3036. https://doi.org/10.3390/electronics13153036.

## APPENDIX

### Appendix 1 : Literature Review vs. Foundation Lab

| Attributes | Study Findings | Our Project |
|---|---|---|
| Collision Near | "Close to landmark" has 30% higher severity for accident. | Which neighborhood? Is one area riskier than another? We identify where in order for preventive measures to be implemented. |
| Road Condition | Road condition accounts for .05% of accident severity | Which road conditions contribute most? We determine how bad they are so repairs can be prioritized effectively. |

### Appendix 2 : List of Features Based on Category Structures

| Category of variable (Grouping) | Variables | Analytical Purpose |
|---|---|---|
| **1. Structural Factors (Infrastructure & Design)** | **route_category_code, road_aspect_code, roadway_condition_code, road_configuration_code, accident_location_code, accident_position_code, authorized_speed** | **Focus on Inherent Design and Maintenance Risk:** This group assesses the static risk embedded in the urban environment, isolating the impact of permanent design flaws and maintenance failures to demand long-term engineering solutions. Authorized speed is included as a severity multiplier linked to kinetic energy. |
| **2. Collision Event and Context** | **collision_type_code, environment_code** | **Focus on Mechanism and Activity Zones:** These variables pinpoint how the collision occurred (e.g., vehicle-to-pedestrian, vehicle-to-fixed object) and where (e.g., school zone, commercial district). This determines if the severity risk is tied to a specific activity type or a particular collision mechanism. |
| **3. Environmental & Surface Conditions** | **weather_condition_code, surface_condition_code, lighting_condition_code** | **Focus on Dynamic Friction and Visibility:**This set quantifies how environmental dynamics exacerbate severity. It models the influence of visibility (lighting), friction loss (surface condition like ice/slush), |

| | | |
|---|---|---|
| | | and general weather, essential for operational alerts. |
| **4. Spatial Factors (Geography)** | **neighborhood** | **CRITICAL: TO CREATE THE VARIABLE.** This categorical variable allows the model to learn that road conditions are disproportionately severe in specific neighborhoods. |
| **5. Dynamic Factors (Temporal)** | **HR_ACCDN (Hour) and JR_SEMN_ACCDN (Day of the week)** | **Focus on Operational Peaks:** Accounts for time-dependent risk variations (e.g., rush hour congestion, late-night alcohol/fatigue risk, weekend driving patterns) crucial for resource allocation and intervention timing. |

**Appendix 3: Workflow Stage**

| Workflow Stage | Data | Features | Model | Policy Insight |
|---|---|---|---|---|
| **Data Preparation** | Montreal Collision Spatial attributes (10Y) Infrastructural features | Reproject coordinates Encode categorical vars Normalization | Ordinal classification regression model SMOTE, weights XGBoost / LightGBM | Top predictors of high gravity Severity risk across city Maintenance & lighting |
| **Evaluation & Validation** | Correlation and variance analysis | Correlation & Variance analysis | F1-Score | Compare predictions with real severity distribution |