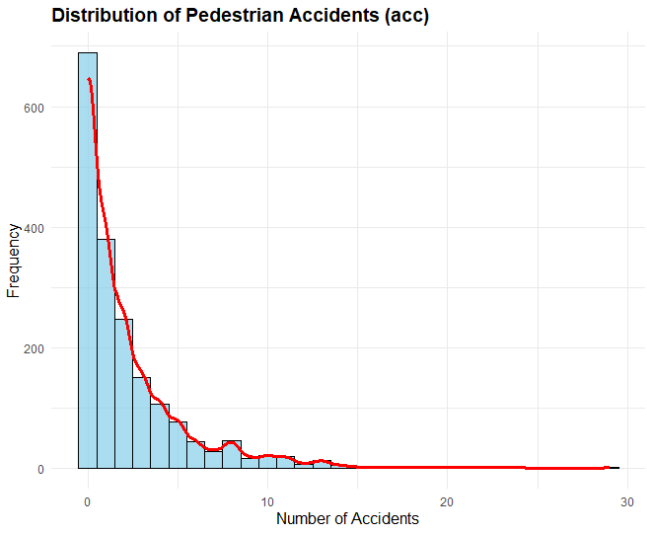
**Advanced Statistical Learning**

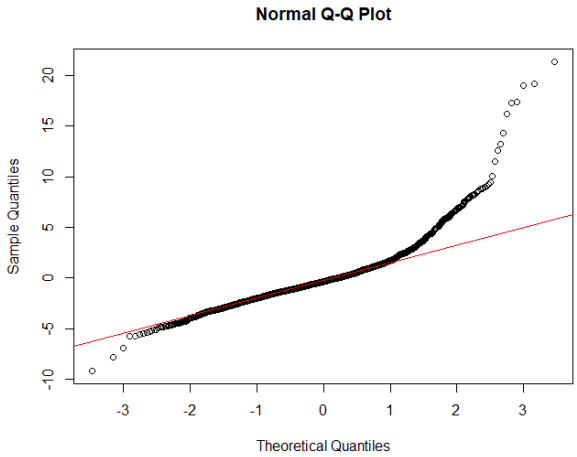
# Descriptive Analysis

## Analysis of distribution of Y (response variable “acc”) :

*Figure 1. Distribution of Pedestrian Accidents*




We first checked the distribution of the response variable acc (number of accidents). It is clearly heavily right skewed and not normal. This may cause some problems for normal linear regression models even though the number of accidents is well above 30. Here a Box- Cox or log transformation might be suitable but we decided to



use normal OLS regression as a baseline test and check the results such as residuals to see if more suitable models work better for this dataset. The results are in Figure 2.

*Figure 2. Q-Q plot*

We see here that the normal Q-Q plot indicates that the residuals from the OLS regression deviate significantly from normality, particularly in the tails, suggesting heavy-tailed residuals and possible outliers. The upper tail shows a strong right-skew, aligning with the earlier histogram of the response variable, while the lower tail also deviates but to a lesser extent. These departures from normality can lead to unreliable standard errors and tests significance in the OLS model. We further did analysis of variance to check overdispersion in image below

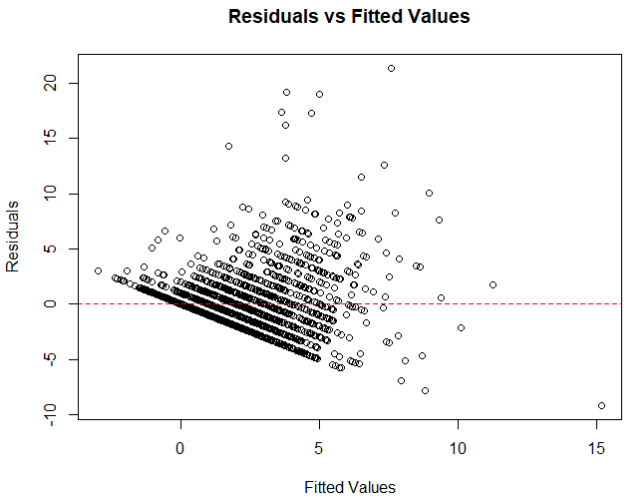



*Figure 3. Q-Q plot*

We see here a clear fan-shaped pattern, where residuals spread out more as fitted values increase, indicating heteroskedasticity (non-constant variance). This violates our OLS assumption, so here the model may not be adequately capturing variability across different levels of the response variable. Additionally, the presence of structured lines in the lower residual range suggests the response variable is **discrete** (likely count-based), reinforcing that here a Poisson or Negative Binomial model would be more appropriate for our case. We also did the studentized Breusch-Pagan test and it returned a p-value less than 2.2e-16 reinforcing that there is heteroscedasticity. We then ran a Poisson model and compared the Pearson residuals to the degrees of freedom (dispersion <- sum (residuals (poisson_model, type = "pearson") ^2) / df.residual(poisson_model)), we got a dispersion parameter of 2.74 ( way over 1) meaning there exists overdispersion and the negative binomial may be more suited here than the poisson model here. We also tested further models as we saw they better fit the data, we will explain them more in detail in the models section of the report.

# Data Pre-processing

Data contains encoding issue in char variables related with French spelling of the streets and borough. The first step is to correct non alphabetic characters.

We decided to remove street names (English and French) as they don't give much info as a characteristic to predict accidents and will make model more complex without added value.

There is one "NA" value in variables ln_distdt, log transformed distance from downtown to the intersection, occurred because the corresponding distance from downtown in the variable distdt, original scale distance, is equal to 0. So, the "NA" value is replaced by 0 as well.

The variable date_ contains 13 missing values. The missing values do not show any pattern and could be classified as missing completely at random. We will replace the NA values with median of the date_.

The data set contains groups of variables that represent the same information in the original scale (pi, fi, fli, fri, fti, cli, cri, cti) compared to their log-transformed (ln_pi, ln_fi, ln_fli, ln_fri, ln_fti, ln_cli, ln_cri, ln_cti). Looking at the distribution of original and log-transformed variables we notice that log-transformed variables would be better for the regression model as it has a more normal distribution , reduces the impact of extreme values and has similar scale with other covariates. To avoid redundancy, we removed the group of variables in original scale and used log-transformed variables as predictors for our model.

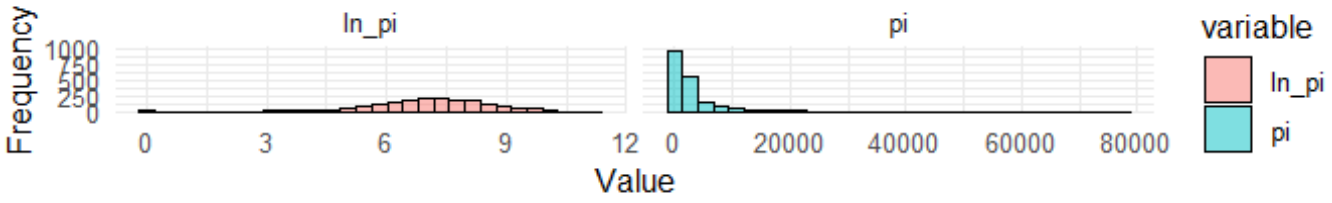



*Figure 4 Distribution of variable in original scale and log-transformed*

However, variable distdt which represents the distance to downtown in meters is preferably to be kept in its original scale rather than log-transformed to preserve the natural spatial distribution of intersections, its higher density closer to downtown. Converting the distance from meters to kilometers reduces the grid scale from 25000 to 25 and improves interpretability of model coefficients with similar scales.

For categorical variables, categories with fewer than 30 observations were combined to avoid misrepresentation due to low sample sizes. As a result, we introduced new variables with suffix _comb for categorical variables with combined categories:

commertial_comb - category 5 represents 5 and more entrances/exits to commercial properties.

of_exclusi_comb – category 3 is for 3 and more exclusive left turn lanes;

total_lane_comb – category 2 is for 2 and less and category 8 – 8 and more number of lanes;

number_of_comb – category 2 –is for 2 and less number of approaches at the intersection;

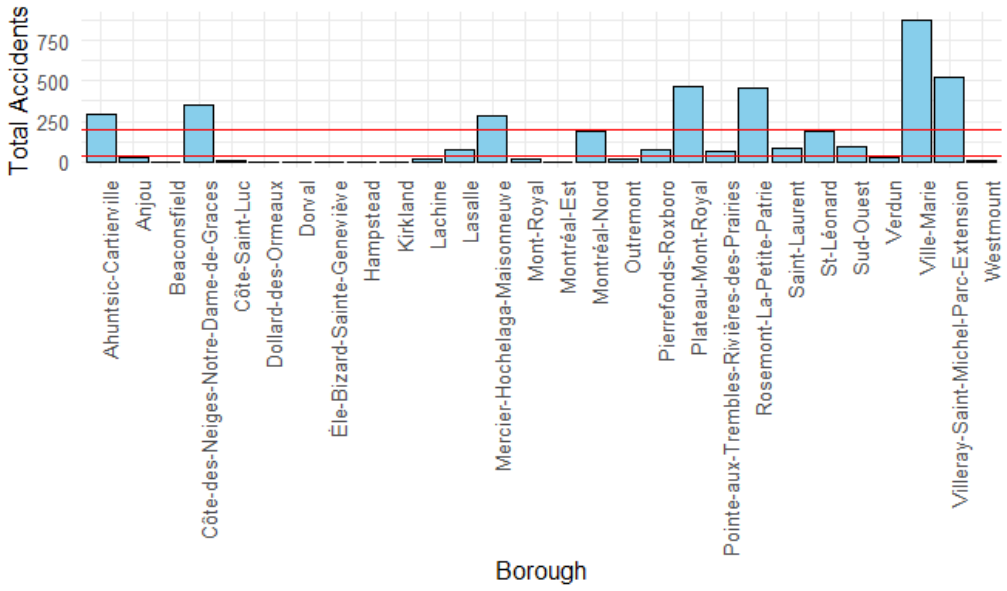



*Figure 6 Total Accidents by Borough*

Variable borough ( in image above) has 28 unique categorical variables with some having very few counts. We decided to introduce a new variable borrow_comb where we group boroughs by number of accidents with categories less than 40 accidents as "low", more than 200 as "high", and between 40 and 200 as "medium". These thresholds are based on natural breaks in the data distribution, ensuring clear differentiation between categories, and provide a meaningful classification for the model.

After verifying the proportion of binary variables, we discovered that variable all_red_an has proportion 99.35% of values 0 to 0.65% of values of 1, which means that this column is not representative and contains majority of identical values and should be removed from the dataset because it is not variable enough.

We eliminated the redundant variables north_ped, east_ped, south_ped, and west_ped since their sum is perfectly equal to the value pi, thus introducing multicolinearity into the model which is not a good thing for inference. Similarly, the combined total of north_veh, east_veh, south_veh, and west_veh exhibits perfect collinearity with fi. Additionally, the sum of fli, fri, and fti is identically collinear with fi, further necessitating their removal to avoid multicollinearity issues.

# Model Analysis

## Linear Model

We begin with an Ordinary Least Squares (OLS) regression as our benchmark model using all predictors. Despite the non-normality observed in the response variable, we decided to try this approach first to assess how well linear models could capture the relationship between predictors and the response variable. The model performance was evaluated using AIC and BIC criteria, we used the AIC and BIC as benchmark to see if further models we theoretically judged to choose, were actually better performing in practice. The results are given in the last section of this report when comparing all models. Actually, normal OLS was one of the worst performing, validating our assumptions

Given the count nature of our response variable (number of accidents) we needed to explore more appropriate models.

## Poisson model

As discussed earlier, considering that the response variable is a count non-negative data we implement a Poisson model, which is more appropriate for this type of data. However, a key assumption of Poisson regression is similar mean and variance. We calculated the dispersion parameter. The calculation yielded dispersion parameter of 2.747349, which is greater than 1. That indicates clear overdispersion in our data. This confirms that Poisson model underestimates the variance in the accidents, potentially leading to incorrect standard errors and misleading significance tests.

## Negative Binomial Models

Based on our dispersion analysis, we implemented Negative Binomial regression models. These models provided a better fit than both linear and Poisson regression, as they properly accommodated the count nature and overdispersion of our response variable. Since our goal is to identify the most dangerous intersections, we include in our analysis models with the offset term of pedestrian and vehicle intensity. This approach helps us make rate-based evaluations of accidents per unit of exposure rather than focusing only on accident counts.

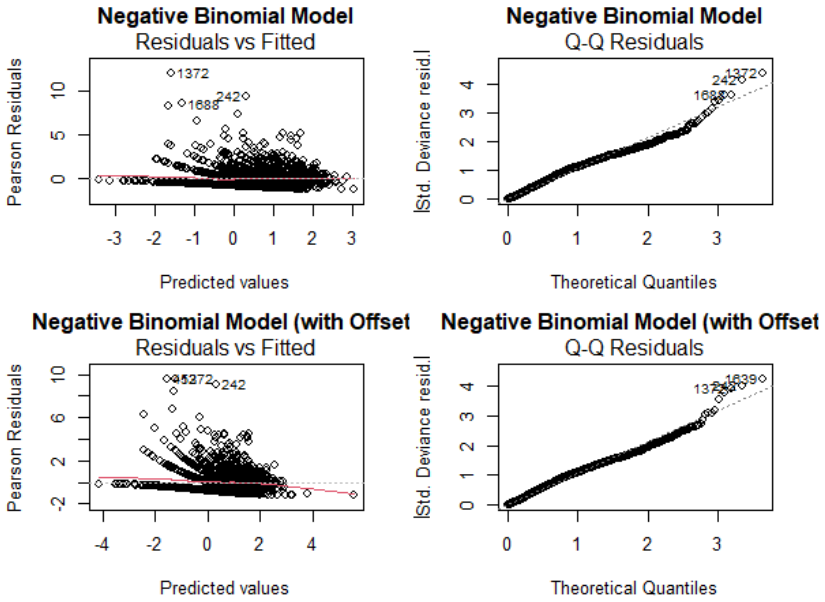
| Model | AIC | BIC | #coef |
|---|---|---|---|
| LM Full | 8971.096 | 9125.949 | 27 |
| Poisson Full | 7309.568 | 7458.890 | 27 |
| NB Full | 6580.733 | 6735.586 | 27 |
| NB with Offset | 6630.220 | 6774.012 | 25 |

*Table 1 Comparison of AIC BIC values across models*

Comparing the diagnostic plots for the two Negative Binomial models (standard and offset):

> Both models have similar Q-Q plots showing slightly better aligning with normality with NB with offset and both are slightly better than OLS.

*Figure 8 Comparing the diagnostic plots for the two Negative Binomial models (standard and offset)*




## Variable Selection Methods

To follow parsimony principle and address potential multicollinearity, we used several variable selection techniques.

## Stepwise Selection with AIC and BIC

The BIC criterion, which penalizes model complexity more heavily, produced a more parsimonious model with fewer predictors than the AIC criterion.

We attempted to use stepwise selection with our Negative Binomial models, but these algorithms did not converge, suggesting that model might be too big and complex for Negative Binomial, or the structure of our spatial data might not very suitable for NB model.

# Regularization Methods

Given the convergence issues with stepwise selection for Negative Binomial models, we explored regularization methods for variable selection.

## Lasso Regression.

The method performs variable selection by shrinking certain coefficients to zero. We implemented both the minimum lambda model and the more conservative "one standard error" rule. (as a note: we used LASSO with default code (and R assumes it is a  Gaussian distribution) as it cannot take a family " negative binomial" as an input . We ran NB model with glmnet.nb .

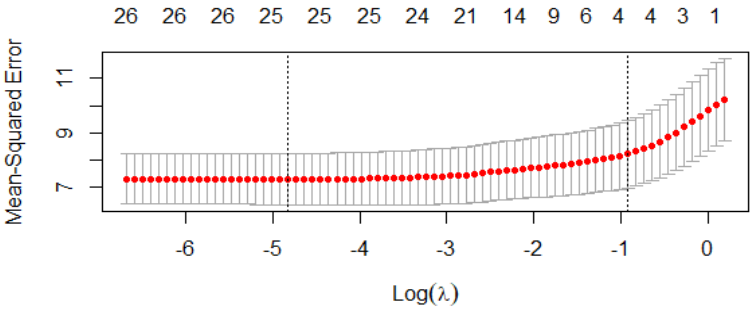



*Figure 9 Lasso Method*

## Ridge Regression

Ridge regression addresses multicollinearity by shrinking coefficients but doesn't reach zero.

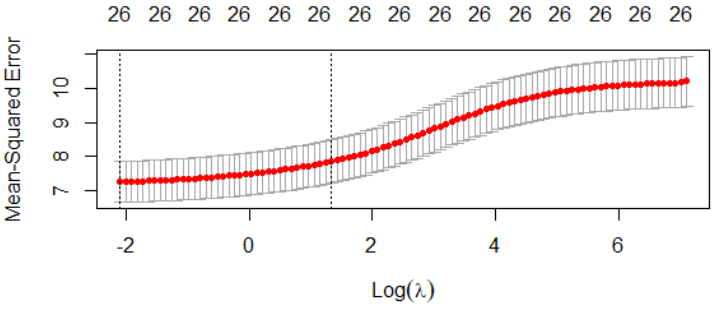



*Figure 10 Ridge Method*

## Relaxed Lasso

The Relaxed Lasso applies variable selection like standard Lasso as a first step, but the second step is to refit the model with the selected during first step variables.

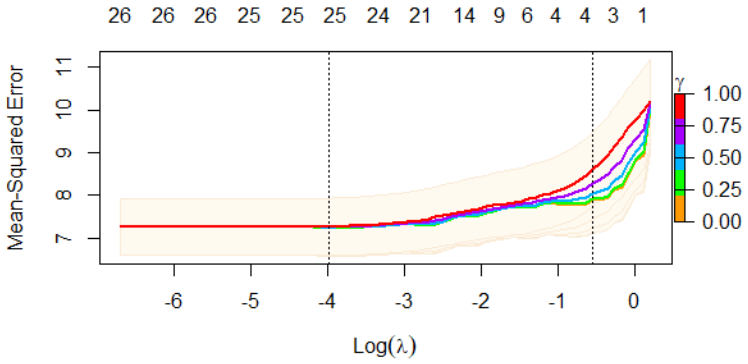



*Figure 11 Relaxed Lasso Method*

*the results are given in last table of report*

## Non-parametric Approach

To capture potential non-linear relationships between covariates and response variable as well as handle potential interactions automatically in our data, we implemented several non-parametric models.

### Random Forest

The method provides flexible non-parametric approach that can automatically capture non-linear relationships and complex interactions between variables. In our implementation we started from running 500 trees and then determined the optimal number of trees 378.

Also, Random Forest can help in variable selection by two criteria MSE and Node Purity. The main inconvenience in RF method is that it's a black box model that doesn't allow any interpretation as of inference, only prediction. (check Figure 12 for variable importance)

### Generalized Additive Models (GAM)

For GAMs, we implemented versions with different levels of complexity to explore goodness interpretability.

*Simplified GAM.* This parsimonious model focused on the most important predictors identified by earlier Lasso+1se approach. Using less predictors made computation more efficient and more interpretable while capturing the most significant non-linear relationships.

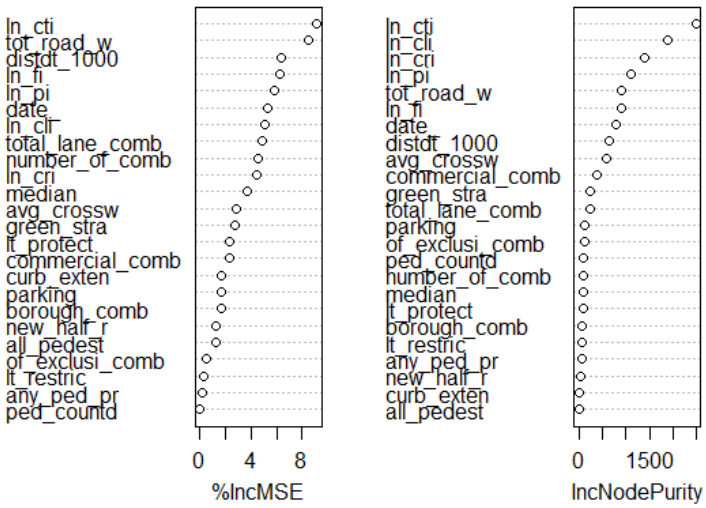



*Figure 12 Variables importance in RF model*

*Full GAM*. The more complex model including all variables with smooth terms for continuous variables. While this expose the risk of overfitting and more computationally costly, it might provide additional details that are missing in less complex model.

*Full GAM with Offset.* The offset version specifically modeled accident rates rather than counts by setting coefficient of vehicle and pedestrian flow to 1. This approach is particularly useful for identifying intersection that were more dangerous than would be expected given their traffic volumes.

Depending on relationship between response variable and covariates, we tried to use several models like factor in R, smoothing splines and LOWESS as we thought it might well suited for spatial data.

## Cross-Validation and Models Evaluation

To properly assess predictive performance, we split our data into training (70%) and testing (30%) sets. We then trained each model on the training data and evaluated performance on the test data using Mean Squared Error (MSE) and Mean Absolute Error (MAE)

| **Model** | **MSE** | **MAE** | **#Coef** |
|---|---|---|---|
| Random Forest | 5.938093 | 1.556478 | 24 |
| RIDGE | 7.394377 | 1.826233 | 26 |
| Relaxed LASSO | 7.440846 | 1.833043 | 25 |
| LASSO | 7.453075 | 1.835215 | 25 |
| LASSO 1se | 7.580342 | 1.849721 | 6 |
| Relaxed LASSO 1se | 7.580342 | 1.849721 | 6 |
| RIDGE 1se | 7.918078 | 1.877127 | 26 |
| GAM | 11.066816 | 1.996392 | 12 |
| GAM Full | 11.171635 | 2.052558 | 39 |
| NB Offset | 11.265502 | 2.097897 | 25 |
| GAM Full Offset | 11.273025 | 2.103641 | 37 |

*Table 2 Model performance metrics on test data*

1. The LASSO model with one standard error rule achieved the best balance between model complexity and predictive accuracy

2. Random Forest demonstrated strong predictive performance, particularly for identifying high-risk intersections

3. GAM models effectively captured non-linear relationships but with higher computational complexity ( the possible reason we got poor MSE for GAMS is due to very high variability because of very complex model in terms of parameters)

| Model | AIC | BIC | #coef |
|---|---|---|---|
| GAM Full | 4590.985 | 4797.913 | 39 |
| GAM Full Offset | 4620.544 | 4817.126 | 37 |
| GAM | 4680.964 | 4748.216 | 12 |
| NB Full | 6580.733 | 6735.586 | 27 |
| NB with Offset | 6630.220 | 6774.012 | 25 |
| Poisson Full | 7309.568 | 7458.890 | 27 |
| LM Full | 8971.096 | 9125.949 | 27 |

*Table 3 Model performance metrics on AIC BIC*

The table shows GAM models outperform all others, with the Full GAM achieving the lowest AIC (4590.985) despite having the most coefficient (39), indicating non-linear relationships are crucial for this data. NB model does better then Poisson, confirming the importance of accounting of overdispersion. Linear models perform worst, confirming its inappropriacy for count data.

**In Conclusion:**

We decided to choose the best model (GAMS) in terms of goodness of fit (AIC and BIC) to recommend the city of Montreal based on the most significance variables affecting either the increasing or lowering of number of accidents. For intersections ranking in terms of dangerousness, we decided to utilize the best predictive model with best MSE (random Forest) to recommend the city on possible adjustments and considerations concerning these intersections.