Examining the factors affecting the occurrence of traffic accidents

JSC370 Final Project

RunshiYang

Contents

1 Introduction **2**

1.1 Background 2

1.2 Research Question 2

2 Methods **3**

2.1 Data Collection 3

2.2 Data Cleaning and Data Wrangling 3

3 Results **6**

3.1 Geographic Location (latitude, longitude) 6

3.2 Natural Environment (weather & road conditions) 7

3.3 Drivers (speeding & alcohol) 8

3.3.1 Logistic Regression 8

3.3.2 Decision Tree 9

3.3.3 Bagging, Random Forest 10

4 Conclusions and Summary **11**

1 Introduction

1.1 Background

Traffic accidents are one of the leading causes of injury and death worldwide. Despite significant advances in transportation safety technology, road traffic accidents remain a serious public health concern. In Canada, traffic accidents result in thousands of injuries and fatalities each year, with many of these accidents occurring in the city of Toronto.

Toronto Police Service Open Data Portal offers a Traffic Collisions - Killed or Seriously Injured (KSI) Dataset, containing detailed information on all traffic collision events in Toronto from 2006 to 2021, where at least one person was either killed or seriously injured. This dataset provides comprehensive information that can help stakeholders gain a better understanding of the causes and consequences of traffic accidents in the city.

1.2 Research Question

The purpose of this report is to investigate the factors that contribute to serious traffic accidents in Toronto. The analysis will focus on identifying the contributing factors to incidents resulting in death or serious injury, including geographic location, environmental factors and demographic characteristics of drivers. The findings of this report will help policymakers, transportation planners, and other stakeholders to develop effective policies and interventions aimed at reducing the incidence of serious traffic accidents.

Table 1: cleaned data

ID	YEAR	DATE	ROAD_CLASS	LATITUDE	LONGITUDE	WEATHER	ROAD_CONDITION	INJURY	SPEEDING	ALCOHOL
3490422	2006	2006-10-01 04:00	Major Arterial	43.73215	-79.27859	Clear	Dry	None	FALSE	FALSE
3519703	2006	2006-11-22 05:00	Minor Arterial	43.64055	-79.42799	Clear	Dry	None	FALSE	FALSE
3519704	2006	2006-11-22 05:00	Minor Arterial	43.64055	-79.42799	Clear	Dry	Major	FALSE	FALSE
3522002	2006	2006-11-22 05:00	Major Arterial	43.65355	-79.43299	Clear	Dry	Minimal	FALSE	FALSE
3522003	2006	2006-11-22 05:00	Major Arterial	43.65355	-79.43299	Clear	Dry	None	FALSE	FALSE
3522004	2006	2006-11-22 05:00	Major Arterial	43.65355	-79.43299	Clear	Dry	Major	FALSE	FALSE
3519357	2006	2006-11-23 05:00	Major Arterial	43.74625	-79.56859	Clear	Dry	Major	FALSE	FALSE
3519358	2006	2006-11-23 05:00	Major Arterial	43.74625	-79.56859	Clear	Dry	None	FALSE	FALSE
3490423	2006	2006-10-01 04:00	Major Arterial	43.73215	-79.27859	Clear	Dry	None	FALSE	FALSE
3519359	2006	2006-11-23 05:00	Major Arterial	43.74625	-79.56859	Clear	Dry	Minimal	FALSE	FALSE

2 Methods

2.1 Data Collection

The main data used in this report were obtained from the Toronto Police Service Open Data Portal, which is a reliable source of information on traffic accidents in Toronto. The data are available for download as a CSV file from this website (click to open), and an API is also provided for data retrieval. I chose to use the API since it allows for greater flexibility in downloading data. For example, I am able to specify the range of years and the type of collisions I want to include in the dataset and I can retrieve data in real-time, ensuring that the dataset is up-to-date. However, due to the transfer limitations of the API, only 1000 rows of data could be downloaded at a time. To obtain the complete dataset, the API had to be called 17 times, with the offset value being updated each time to ensure that all rows were retrieved. The acquired dataset contains 16,488 observations and 57 columns, with each line detailing the time, location, and road conditions of the corresponding traffic accident. Each observation in the dataset has a unique index that can be used to identify and track specific accidents.

2.2 Data Cleaning and Data Wrangling

In this study, data cleaning and wrangling are conducted to ensure the accuracy and consistency of the data. First, 11 columns of the KSI dataset are selected based on their potential influence on traffic accidents, including factors such as location, road conditions, and driver behavior. To facilitate analysis and interpretation, columns are renamed with more descriptive names.

One issue identified during the data cleaning process was the date column, which was recorded in Unix time. To make the data more accessible and understandable, the date column is converted to a year/month/day hour:minute format with the help of the `lubridate` package.

Another issue identified during the data cleaning process was the use of `NA` values in the SPEEDING and ALCOHOL columns to indicate that the driver was not speeding or driving under the influence. To ensure consistency and avoid potential confusion, these NA values are replaced with boolean values (`TRUE` and `FALSE`) to indicate whether or not these factors were involved in the collision.

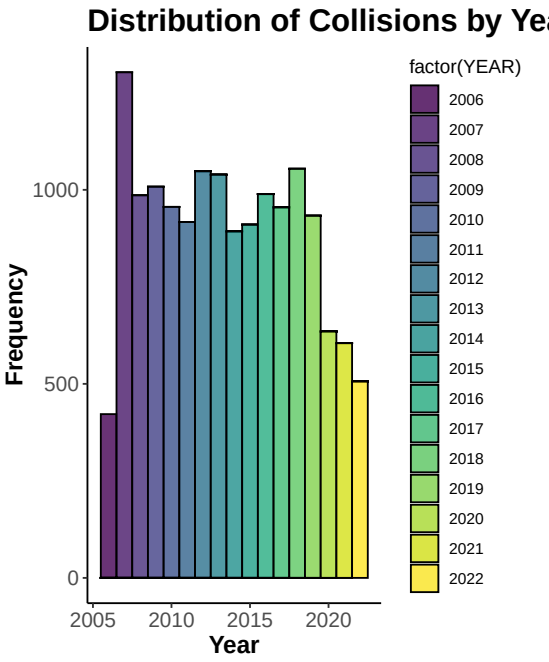
Finally, due to the relatively small number of incomplete observations, it was decided to remove any observation with missing data. The resulting cleaned dataset contains 15,159 observations, all of which have complete data. This ensures that the analysis conducted on this dataset is reliable and accurate. Table 1 shows the first 10 rows of the cleaned data.

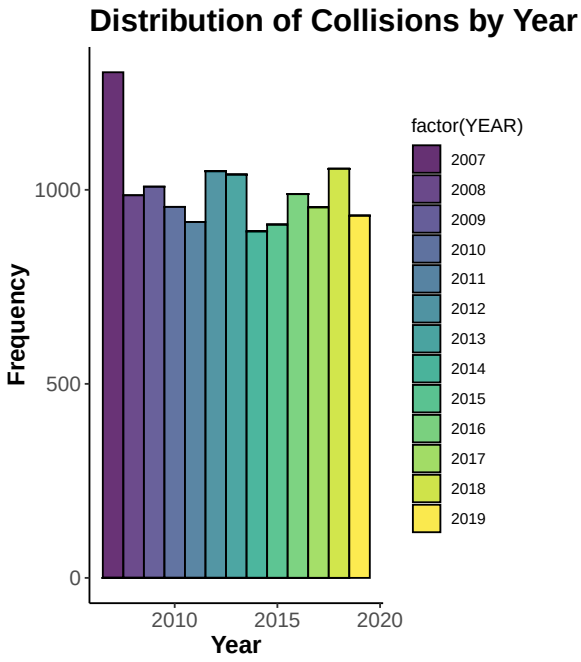
Table 2: Summary Statistics

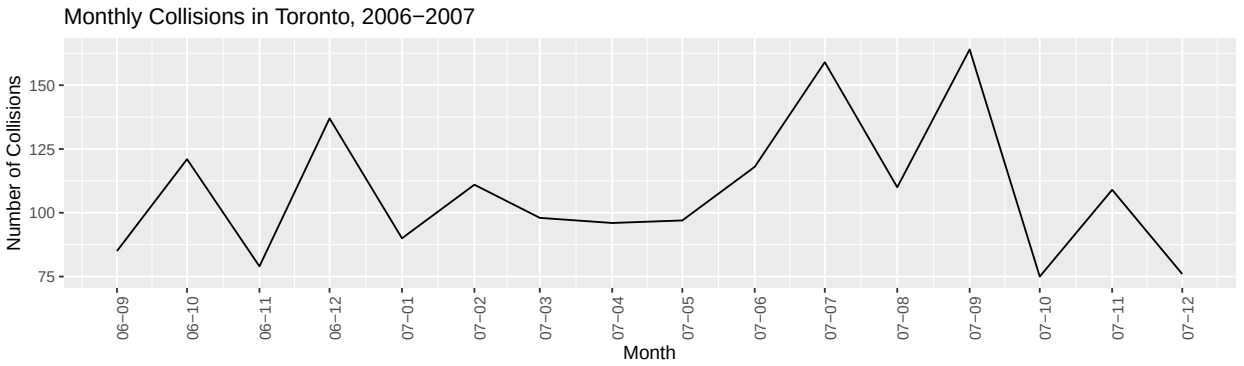
ID	YEAR	LATITUDE	LONGITUDE
Min. : 3470108	Min. :2006	Min. :43.59	Min. :-79.64
1st Qu.: 6105226	1st Qu.:2010	1st Qu.:43.66	1st Qu.:-79.47
Median : 8000017	Median :2013	Median :43.70	Median :-79.40
Mean :43133801	Mean :2014	Mean :43.71	Mean :-79.40
3rd Qu.:80917314	3rd Qu.:2017	3rd Qu.:43.76	3rd Qu.:-79.32
Max. :81705465	Max. :2022	Max. :43.86	Max. :-79.12

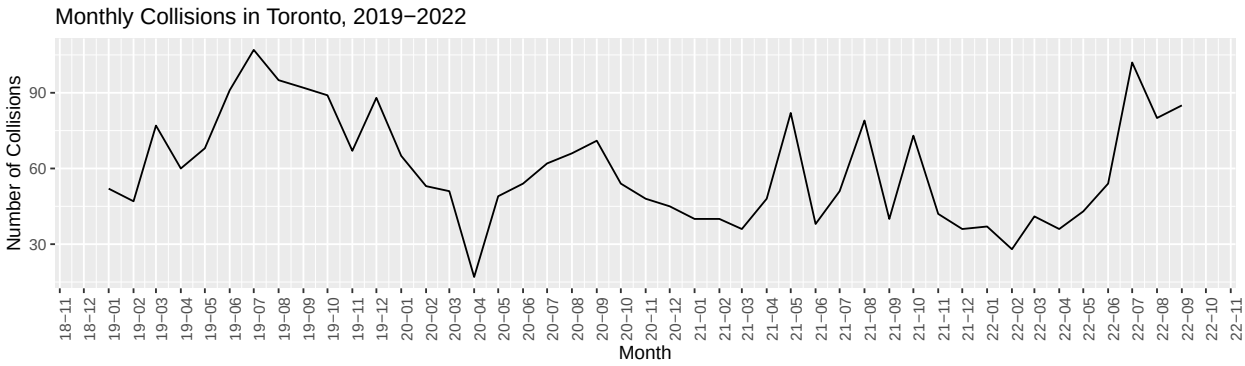
Then I move on to analyzing the data to identify any anomalies or potential errors. One of the first things I did was to create a summary of the numerical variables in the dataset (see Table 2). I do not find anything unusual, which gives me confidence that the data is clean and accurate.

Moving on, I create a bar plot to show the distribution of collisions over the years. From the plot, I observe that the number of KSIs per year decreases as the year increases, except for 2006 where there was a lower number of KSIs compared to the other years. And the number of traffic accidents decreases rapidly from 2020 to 2022. So I plotted the number of traffic accidents from 2006 to 2007 and 2019 to 2022 in the chart below. It turns out that the data for 2006 and 2022 are not complete, the dataset only records data after September 2006 and before October 2022. And I suspect that the data for 2020 and 2021 are not representative since they are affected by Covid-19. Since time of the year is potentially associated with traffic accidents, so I decided to use only the data between 2007 and 2019 for the analysis.









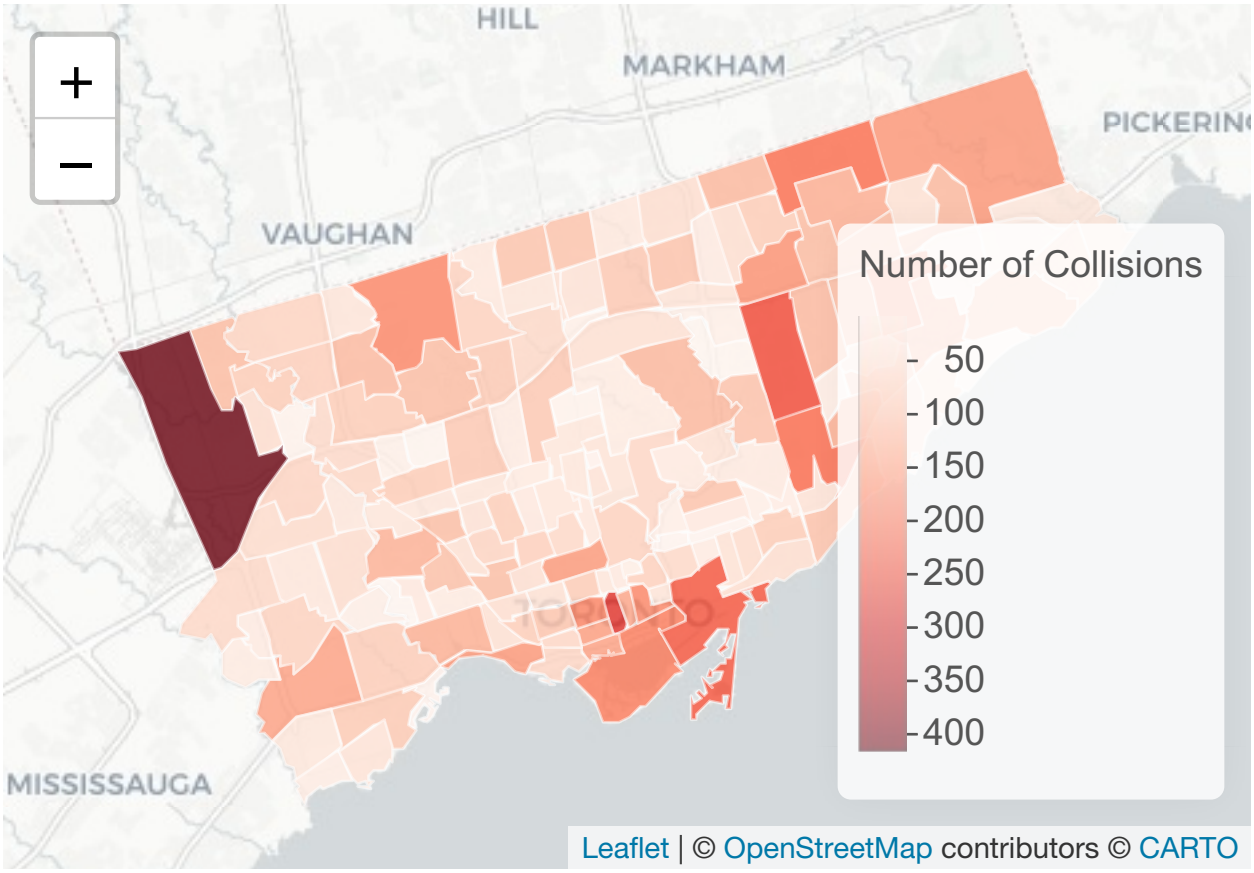
Overall, the analysis gives me confidence that the data is now clean and accurate. By identifying and addressing any potential anomalies or biases, I am able to ensure the validity and reliability of the analysis. The dataset contains a total of 11,312 observations, providing us with sufficient data to draw inferences about the factors that contribute to the incidence of severe traffic accidents in Toronto.

3 Results

I investigate the factors affecting traffic accidents from three perspectives: geographic location, natural environment, and drivers. We examine the relationship between the frequency of accidents and the latitude, longitude, and road type of the accident-prone areas. Furthermore, we explore the correlation between accidents and natural factors such as season and road conditions. Lastly, we delve into the role of speeding and alcohol as critical factors in causing traffic accidents.

3.1 Geographic Location (latitude, longitude)

To investigate the effect of geographic location on traffic accidents, I grouped all traffic accidents that occurred between 2007 and 2021 by neighborhood using information on the latitude and longitude of traffic accidents and the Toronto neighborhood delineated by Statistics Canada census tracts, and color-coded the number of traffic accidents that occurred in each neighborhood on the map. From the interactive map in my website (click to open), we can see that West Humber-Clairville is the most crash-prone neighborhood, with a total of 415 serious crashes from 2007 to 2021. Followed by the Yonge-Bay Corridor and South Riverdale in downtown, with a total of 200-300 crashes. Wexford/Maryvale and Milliken have also experienced a relatively high number of traffic accidents (around 250).

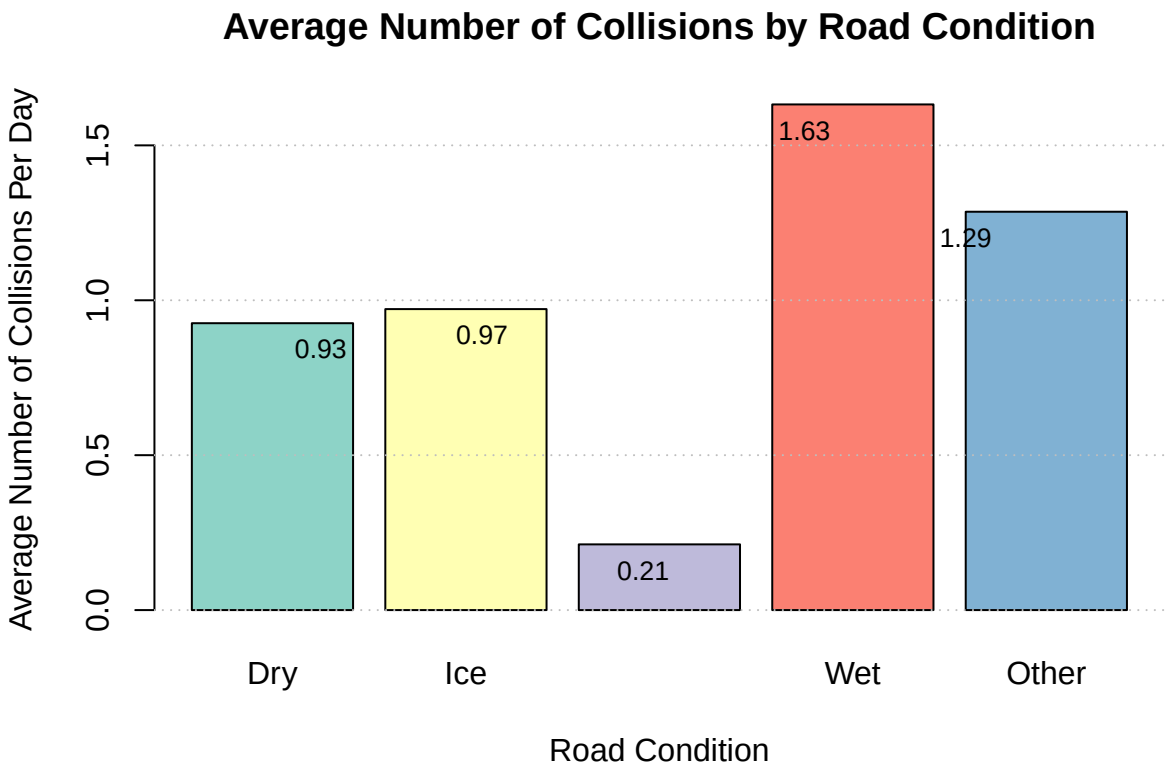


3.2 Natural Environment (weather & road conditions)

To analyze the likelihood of traffic accidents based on different road conditions, it is not appropriate to simply tally the number of accidents per road condition. This is because road conditions like wet, icy, or with loose sand or gravel are infrequent occurrences due to less frequent rain, snow, or dust storms. Hence, to accurately determine the likelihood of serious traffic accidents per day on a specific road condition, I divide the total number of accidents by the number of days in which that particular road condition may arise.

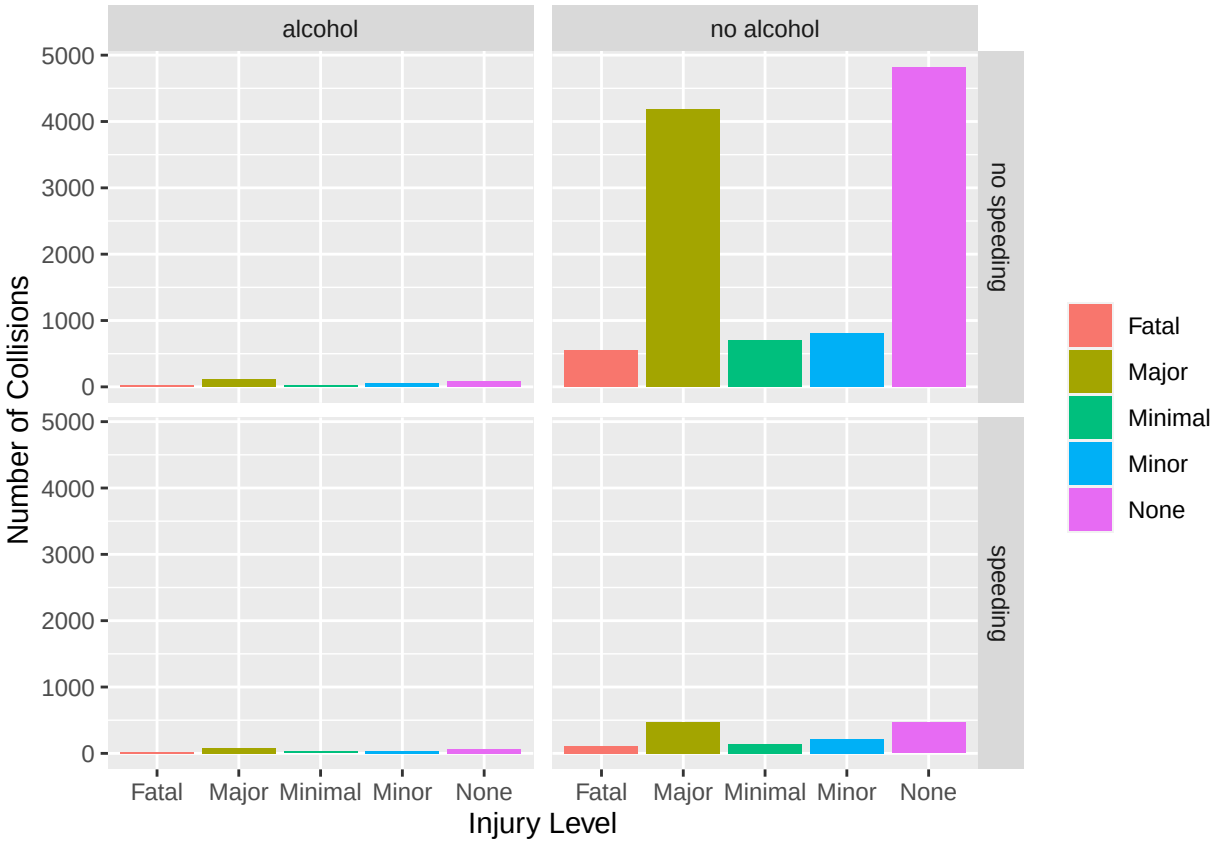
The resulting average number of accidents per day under a specific road condition is obtained. The bar plot below shows that, on average, 1.62 collisions occur on wet roads, 0.93 collisions occur on icy roads, and merely 0.92 collisions occur on dry roads.

Nonetheless, this approach is subject to bias since fewer vehicles are likely to travel during inclement weather conditions, resulting in an underestimation of the average number of traffic accidents occurring on all road conditions other than dry roads.



3.3 Drivers (speeding & alcohol)

I tallied the occurrences of collisions for each year based on the following categories: collisions where the driver was not under the influence of alcohol (DUI) and not speeding, collisions where the driver was DUI but not speeding, collisions where the driver was not DUI but was speeding, and collisions where the driver was both DUI and speeding. Additionally, I recorded the distribution of injuries for each of these categories. My analysis shows that in the case of DUI, speeding and non-speeding are almost equally likely, while most accidents occur without speeding in the case of no DUI. These findings suggest that DUI is a critical factor leading to speeding and highlight the need for strict enforcement of DUI laws and better education on the dangers of drunk driving.



I also notice that for speeding and alcohol-induced traffic accidents, injuries occur more frequently. So I test this hypothesis using logistic regression, decision tree and random forest:

3.3.1 Logistic Regression

From Table 1, we can see that the logistic regression analysis conducted indicates that the predictors, SPEEDING and ALCOHOL, have a significant impact on the likelihood of a traffic collision resulting in injury. The coefficient for SPEEDING is $\hat{\beta}_1 = 0.406$, indicating that the odds of injury are approximately $e^{\hat{\beta}_1} = 1.5$ times higher when speeding is a factor in the collision event. Similarly, the coefficient for ALCOHOL is $\hat{\beta}_2 = 0.465$, which suggests that the odds of injury are approximately $e^{\hat{\beta}_2} = 1.6$ times higher when alcohol is involved. The intercept coefficient, $\hat{\beta}_0$, is 0.262. This represents the log odds of injury when neither SPEEDING nor ALCOHOL are present in the collision event. The p-value for all coefficients is 0, indicating that they are statistically significant predictors of injury.

Table 3: Logistic Regression Summary Table

Coefficients	Estimated value	Std. Error	z value	Pr(>|z|)
(Intercept)	0.262	0.019	13.722	0.000
SPEEDINGTRUE	0.406	0.057	7.128	0.000
ALCOHOLTRUE	0.465	0.100	4.646	0.000

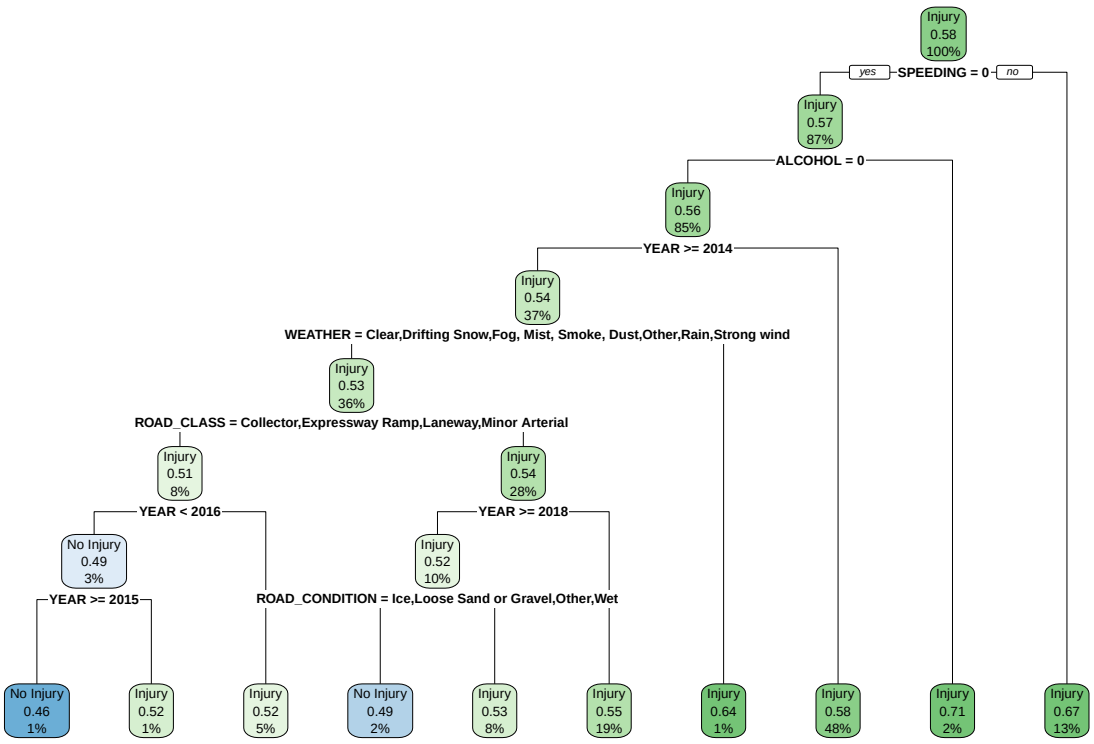
The logistic regression equation for this model is as follows:

$$\text{logit}(\hat{\pi}) = \log \frac{\hat{\pi}}{1-\hat{\pi}} = 0.262 + 0.406(SPEEDING) + 0.465(ALCOHOL),$$

where π represents the probability of injury and SPEEDING and ALCOHOL are indicator variables for whether speeding or alcohol were present in the collision event.

3.3.2 Decision Tree

The binary decision tree model predict injury in traffic accidents based on a set of predictors (YEAR, ROAD_CLASS, WEATHER, ROAD_CONDITION, SPEEDING, and ALCOHOL).



Based on the output from the decision tree model, the most important factors leading to injury in traffic accidents are the variables "SPEEDING" and "ALCOHOL". And the These variables are used as the first two splits in the decision tree, indicating that they have a strong predictive power in determining whether an accident resulted in an injury.

The first split on "SPEEDING" suggests that accidents with no speeding are less likely to result in injury compared to accidents with speeding. The second split on "ALCOHOL" suggests that accidents with no

Table 4: Variable Importance for Decision Tree

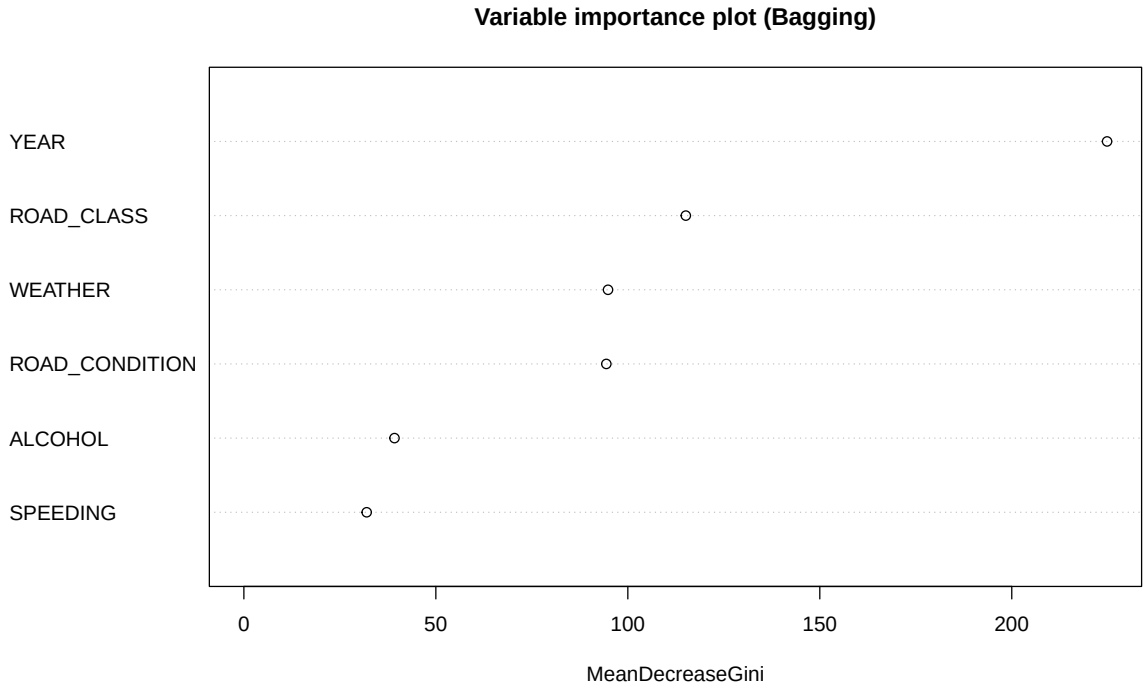
Variable	Importance
SPEEDING	32.782057
YEAR	14.813990
ALCOHOL	12.620716
WEATHER	2.199895
ROAD_CLASS	1.562877
ROAD_CONDITION	1.354774

alcohol involved are less likely to result in injury compared to accidents with alcohol involved. And the importance of these two variables is relatively high compared with the others.

While the other predictor variables (YEAR, ROAD_CLASS, WEATHER, and ROAD_CONDITION) are also included in the decision tree model, they are not as influential as "SPEEDING" and "ALCOHOL" in predicting injury. Therefore, addressing the issues of speeding and alcohol consumption in drivers may have the most significant impact in reducing the number of injuries resulting from traffic accidents.

3.3.3 Bagging, Random Forest

However, when I tried to use bagging to build multiple decision trees on randomly sampled subsets of the data to reduce overfitting and improve the generalization performance of the model, it turns out that SPEEDING and ALCOHOL has relative low importance, which contradict with the previous results. It is possible that the variables ALCOHOL and SPEEDING are strongly correlated with other variables in the dataset, which means that they may not be as important in predicting injury when other variables are also taken into account. When using a single decision tree, the algorithm may give more weight to these variables if they happen to be the best predictor of the outcome in the tree's splits. However, when using bagging, the algorithm is able to average out the importance of each variable across multiple trees, which may result in a smaller relative importance for these variables.



4 Conclusions and Summary

From my analysis, I have found that there are multiple factors contributing to serious traffic accidents in Toronto. The geographic location of accident-prone areas are West Humber-Clairville, Yonge-Bay Corridor, Wexford/Maryvale and Milliken. The natural environment, specifically the wet roads after rain contributes the most to the occurrence of accidents, followed by icy roads after snow. Finally, DUI remains a critical factor leading to speeding and accidents involving both speeding and alcohol are more likely to result in injuries.

The study suggests that safety measures such as increased police presence and traffic enforcement could be implemented in the accident-prone areas and during bad weather, and efforts to raise awareness about the dangers of driving under the influence should be made to reduce the occurrence of such accidents.