Examining the Effects of Alcohol on Automobile Collision Injuries and Fatalities in the Toronto Area*

Ken Lee

26 April 2021

Abstract

Drinking and driving is a major issue that claims an estimated 1,500 Canadian lives every year, and although fewer people are dying from car automobile accidents, Canada's proportion of deaths involving alcohol was at 34%. Hence, this research paper will explore the effects of alcohol on potential car accident injuries and fatalities. More specifically, it will observe the differences between treated groups (alcohol-related accidents) and control groups (non-alcohol-related accidents) through the use of propensity score matching to control for the many variables (street, age, date, and time) and account for potential biases. All things considered, the insights from this research will not just help us inform the public of the dangers of drunk driving, but also reduce the number of senseless deaths in the process.

Contents

*Code and data are available at: https://github.com/kenllee97/Examining-the-Effects-of-Alcohol-on-Motor-Vehicle-Accident-Casualties-and-Injuries

1 Introduction

Drinking and driving is a major societal issue that has claimed thousands of lives. In the USA, 10,142 people died from drunk driving accidents alone, meaning one person was dying to this problem every 52 minutes ("Overview of motor vehicle crashes in 2019" 2019). Here in Canada, around 1,500 people die every year due to drunk driving, and although motor vehicle crash deaths have been decreasing over the years, alcohol is still significantly linked to these deaths with a proportion of around 34% (S. W. Brown and Robertson 2017). After all, any death involved with this issue is a senseless death, especially when it is not just the drunk drivers who are affected, but the innocent third parties involved in these collisions. Hence, it is important to further understand the potential effects of alcohol on collision injuries and fatalities. These findings would further help inform the general public about the dangers of drunk driving, in addition to informing policies and strategies that limit its injuries and fatalities.

Therefore, this paper will examine the effects alcohol can have on the magnitude of a motor vehicle collision, observing whether the involved individuals were not injured, or had minimal, minor, major, or fatal injuries. The data set is provided by Open Data Toronto, giving us a glimpse of the DUI accidents in the city of Toronto. Of course, this would also involve the 29 control variables, such as road conditions, vision, and street, which would need to be controlled.

This paper will first focus on the experimental design involved. For instance, denoting the tools used and the feature engineering done, such as the regularization and normalization of features that do not meet the logistical regression assumptions involved in the propensity score matching. Most importantly, it will also denote how we will be using propensity score matching and the logistic regression model to compare the differences and control for the many distinct features in order to create internal and external validity on the findings. This will allow us to compare treatment and control groups with the same range of probability of DUI. This type of methodology and model would ultimately not just help control variables (as it takes into account the other variables and attempts to group them together based on the final propensity score), but also reduce selection bias. This isolates variables like age and location, allowing us to compare between groups whose only difference (treatment) is the consumption of alcohol. At last, upon conducting the comparisons, all the differences in effects will be aggregated and analyzed to understand the consequences of alcohol on the type of injury.

2 Experimental Design

For this research paper, R [] and packages such as "Tidyverse" [] and "dplyr" [] were used to analyze and clean the data. Additionally, "knitr" [] was used to create this PDF, while "ggplot2" [] was used to visualize the data.

2.1 Model Description Overview

The experimental design model this paper will be utilizing is propensity score matching. More specifically, it will be using a logistic regression model to find the weights of each of the features in predicting whether the sample would have consumed alcohol before driving. Of course, it is important to note that the injuries feature will not form part of this process, as it is the very metric we will be using to compare between groups at the end, preventing information leakage. All in all, upon discovering the weights, they will be used to calculate the propensity score (probability that they consumed alcohol before driving) of each of the data samples. These scores will then be used to group the samples into similar probabilities and allow us to compare between the control groups (samples that had the same probability of consuming alcohol and did not consume alcohol) with the treatment groups (samples that had the same probability of consuming alcohol and did consume alcohol). Hence, the difference in these comparisons would inform us of the potential effects of drinking while driving, as the propensity score calculated would not just help us counter selection bias, but also control for the many foreign variables by accounting for them in the alcohol consumption probability.

After all, the similarities in propensity score should suggest an affinity in many of the other foreign variables, meaning the only difference is the consumption of alcohol, and hence enhancing the causal validity of the findings.

Nonetheless, it is important to note that to keep this experiment easier to analyze and compare, only some of the features were used in the logistic regression to create the propensity scores. After all, the use of all the potential features would result in an immense diversification of scores which would render the matching useless, as many samples would not have an equal counterpart. The next sections will further explore how the data was cleaned and how specific features were selected to be the main predictors in the logistic regression.

2.2 Data Cleaning

When cleaning the data, rows with null values as well as values such as "unknown" and "other" were removed. These variables were removed because they can significantly impact the analysis as null and unknown values are just missing (and speculating and making them up would be harmful to the validity) while "other" variables are too broad for the experiment to control (harming its validity as well). Additionally, features with too many null values (more than 30%) were also removed from the data set. These included variables such as: Offset (Distance and direction of the Collision), Latitude, Longitude, Collision Location, Direction of Travel, Vehicle Maneuver, Driver Action, Driver Condition, Pedestrian Type, Pedestrian Condition, Cyclist Type, Cyclist Action, and Cyclist Condition.

All in all, upon cleaning the data set, another issue at hand was the mere size of the data set, with 38 features left. To simplify the process of the experiment, features that were very similar to each other, and hence redundant, were also removed. For example, features such as index and id were removed because they had no effect in the experiment, while variables such as year were removed because it was already represented in the date features. Nevertheless, the redundancy reduction only managed to bring down the number of features to 31, which is why the feature selection in the next step was key.

2.3 Feature Selection

One of the first steps when picking the right predictors for the logistic regression was through the use of literature reviews. For instance, research papers like the "Driving under the influence of alcohol: frequency, reasons, perceived risk, and punishment" [] concluded that some of the main reasons people consume alcohol before driving are because there were no other means of transportation and the fact that the drinking was associated with their meals. Based on this, we can deduct that features such as streets would have an impact on alcohol consumptions as it would denote the availability of alternative means of transportation as well as the location of restaurants and bars. Additionally, the hour would also play a role as individuals may have certain meals and drinks at a certain time, while some modes of transportations may not be available in the later hours of the day. Similar to the hours, dates would have similar effects as establishments and modes of transportation may not be available on certain dates, while some dates (and hours) may encourage more individuals to go out and drink. At last, due to the topic of alcohol, age may also play an immense role as it would determine their access to said alcohol, as well as their capabilities of going out for meals and drinks.

In addition to literature review, a correlation matrix leveraging Cramer's V coefficient for all the categorical variable combinations was used. This matrix was constructed with the help of AntoniosK [] using the "lsr" [] and "corrplot" [] packages. Cramer's V coefficient was used because it allows us to analyze the correlation between categorical values. This would not just help us identify variables that are most correlated with whether the sample consumed alcohol, but would also help us understand the nature of the variables before being used for the logistic regression. After all, one of the only main assumptions of logistic regression is that the predictors are independent of each other. Hence, with this matrix, we would be able to pick the most significant features while also reducing the inclusion of confounding and mediator variables within the data set.

In figure 1, we can visualize said matrix where red signifies a strong correlation while the transparency denotes the lack of correlation. We can see in the matrix that Street 1 and Street 2 were significantly

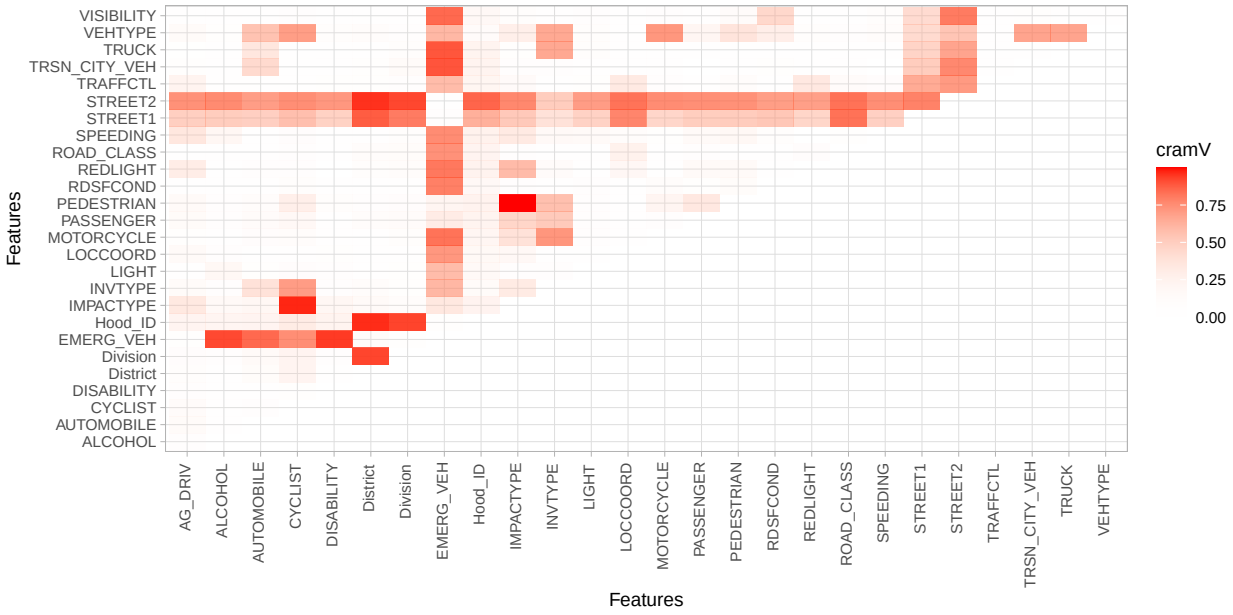

Figure 1: Feature Correlation Matrix Using Cramer's V Coefficient

correlated with many of the other variables, meaning its inclusion in the logistic regression would also take into account the many other potential effects of other correlated variables not included. Additionally, they both also share some correlation with the consumption of alcohol, meaning they would be good predictors. More importantly, they did not correlate with each other. This finding also seems to be backed by the academic review, as the streets (location) would play a significant role in the availability of alternative means of transportation, as well as the available locations for individuals to have their meals and drinks.

2.4 Propensity Score Matching

After considering the academic review as well as the correlation matrix, Date, Hour, Age, Street 1, and Street 2 were selected as the main predictors for the logistic regressions. The reason we have selected a logistic regression model to calculate the propensity score is that we are trying to determine the probability the treatment (alcohol) is involved, which is a dichotomous variable, meaning the sample can either have alcohol or not. Moreover, the logistic regression is easy to implement and can take into account categorical data (streets) and numerical data (dates, hour, age) without needing to consider the distribution of the features.

Of course, this also has its own drawbacks like the fact that each variable needs to be independent of others (no multicollinearity). This makes it hard to pick the right variables as it is complicated to determine the correlation of so many distinct features. To add to this, there may also be other unknown confounding or mediator variables that are at play, which could affect the accuracy and precision of the logistic regression model. There are also the basic assumptions of linearity between dependent and independent variables which could harm the outcomes if said assumption is not met. Hence, although logistic regressions seem to be appropriate to estimate the probability that the samples consume alcohol before driving, there are concerns on whether the data meets its assumptions. For instance, there may be unknown variables at play, we may not have picked an adequate amount of features to create an accurate probability, or some features may not be linearly correlated with the dependent variable.

All in all the logistic regression model is the following:

$$log(\frac{p}{p-1}) = \beta_0 + X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + X_4\beta_4 + X_5\beta_5$$

Where p denotes the probability of the event being treated (alcohol), β_0 represents the reference group, other numbered β's represent the feature coefficients of Date, hour, Age, Street 1, and Street 2, while the numbered X's represent the explanatory variables (values representing said features in the data set inputted to calculate the propensity score)

Upon finding the weights of the predictors using the "broom" [] package, the logistic regression model was then used to calculate all the propensity scores of each of the individual samples. The package "arm" [] was then used to match the treated samples with control samples with the same score, allowing us to compare the differences in injuries.

3 Data

The data we are using for this report comes from the R package "opendatatoronto" []. This package helps us gather data sourced from Toronto's Open Data Portal, the official source for data collected from Toronto's divisions and agencies. The data set used was the "Motor Vehicle Collisions involving Killed or Seriously Injured Persons" [], published under the Open Government License by Toronto Police Services and was last updated August 18, 2020 (data set refreshes annually). Because these are police reports, one of the major strengths of this data set is the abundant and precise information it provides. This data ultimately includes all traffic collisions where individuals were killed, injured or neither from 2006 - 2019. It consists of a sample size of 16,093, with 56 features.

3.1 Data Biases

One of the main potential weaknesses or biases of the data set is that these are all collected from police reports, which means collisions that are not reported or witnessed will not be included. This would create some validity issues as we would not be considering incidents samples that were not recorded. For instance, for an accident, especially involving alcohol, the culprits may flee and not report the incident due to legal consequences. Hence, data may be misrepresented and heavily biased for only reported incidents, which most likely do not involve alcohol. Additionally, there may also be unreported cases where individuals did not consume alcohol and the accident was so minor that it did not merit a report.

Another drawback would be the fact that the locations of the collisions were offset to the nearest intersection. In other words, these locations, divisions, and neighborhoods may not reflect exactly the location of the incident, making it harder to control for these variables and hence weaken the causality validity of the paper.

At last, factors such as light, visibility, and road surface conditions may also lack accuracy, as it would be quite hard to determine the exact conditions of the collisions. After all, the report is only recorded after the fact, so conditions may have changed between the time of the collisions and the time the police arrive. All in all, the data set is not perfect and does contain weaknesses that could derail the analysis and causal validity of the study. Nonetheless, the data set is still substantive enough to derive insights from.

3.2 Feature Exploration

Although it is not necessary to check the distribution of features for logistic regression, as it does not require features to be normally distributed, exploring these features will help us understand better the type of values we will be processing.

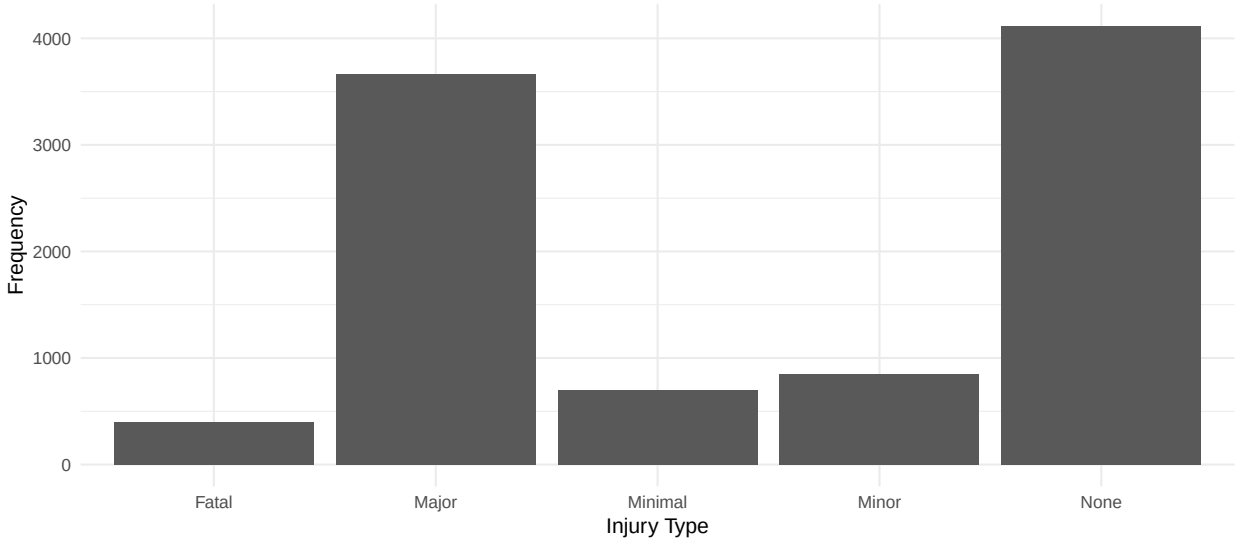

Figure 2: Injury Type Frequency

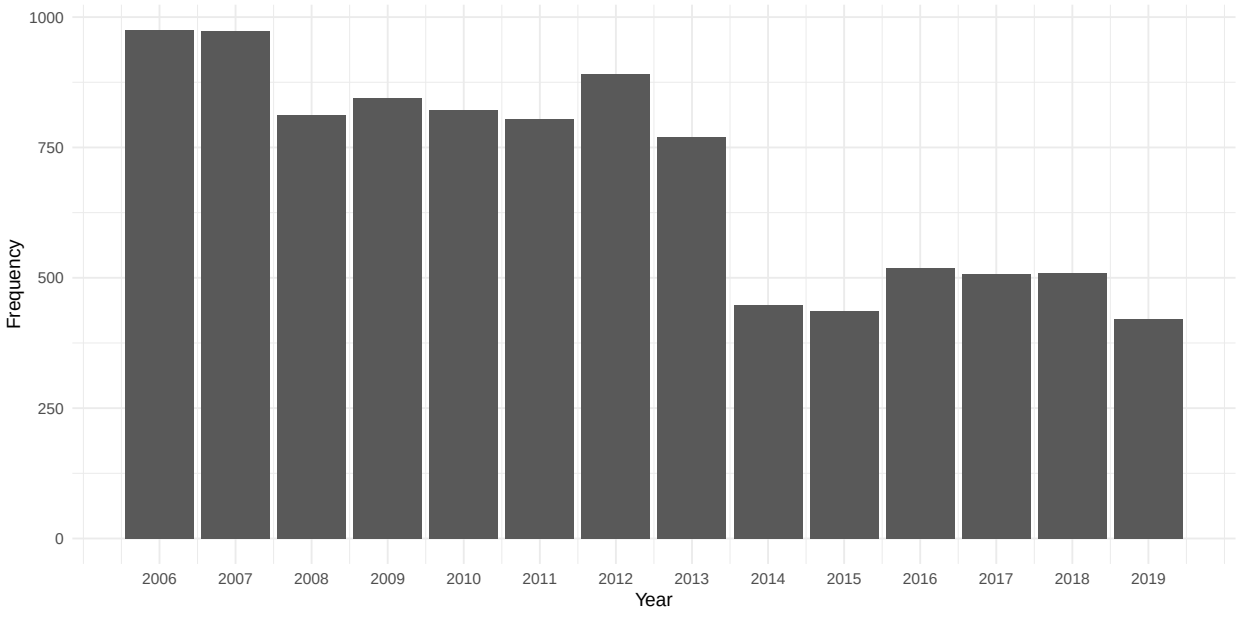

Figure 3: Year Collision Distribution

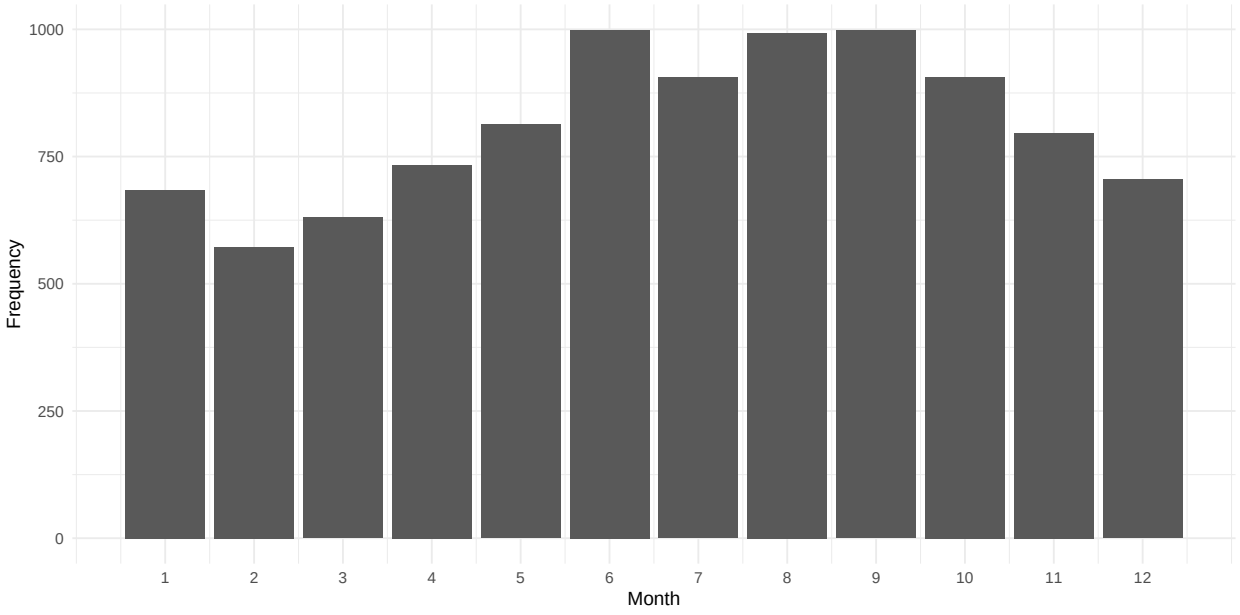

Figure 4: Month Collision Distribution

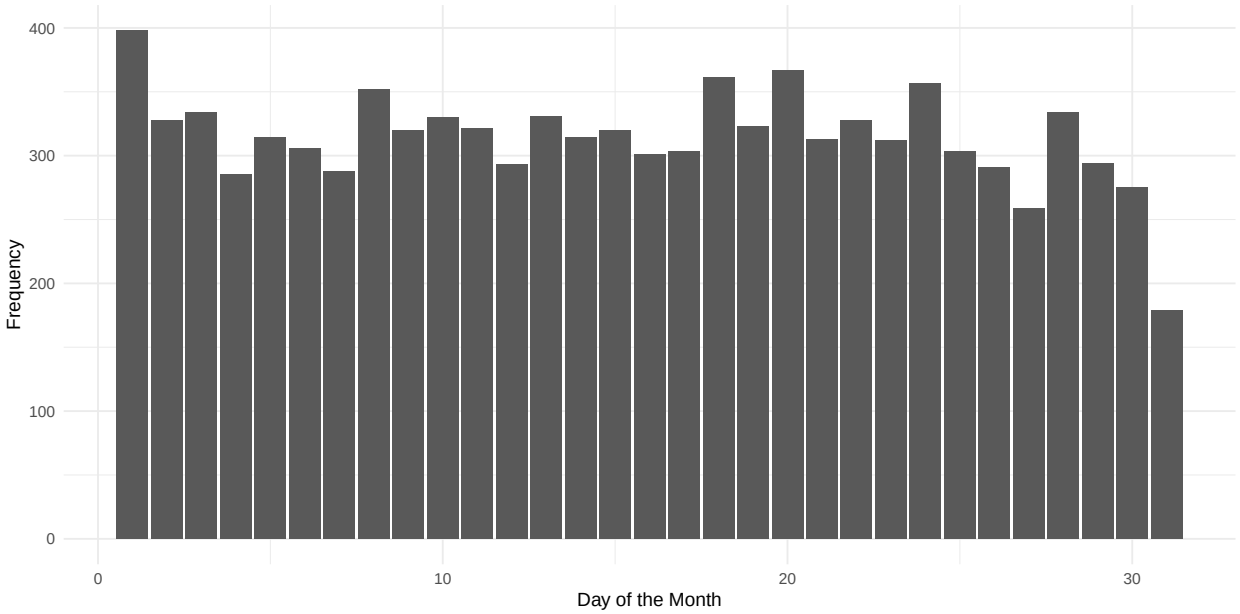

Figure 5: Day of the Month Collision Distribution

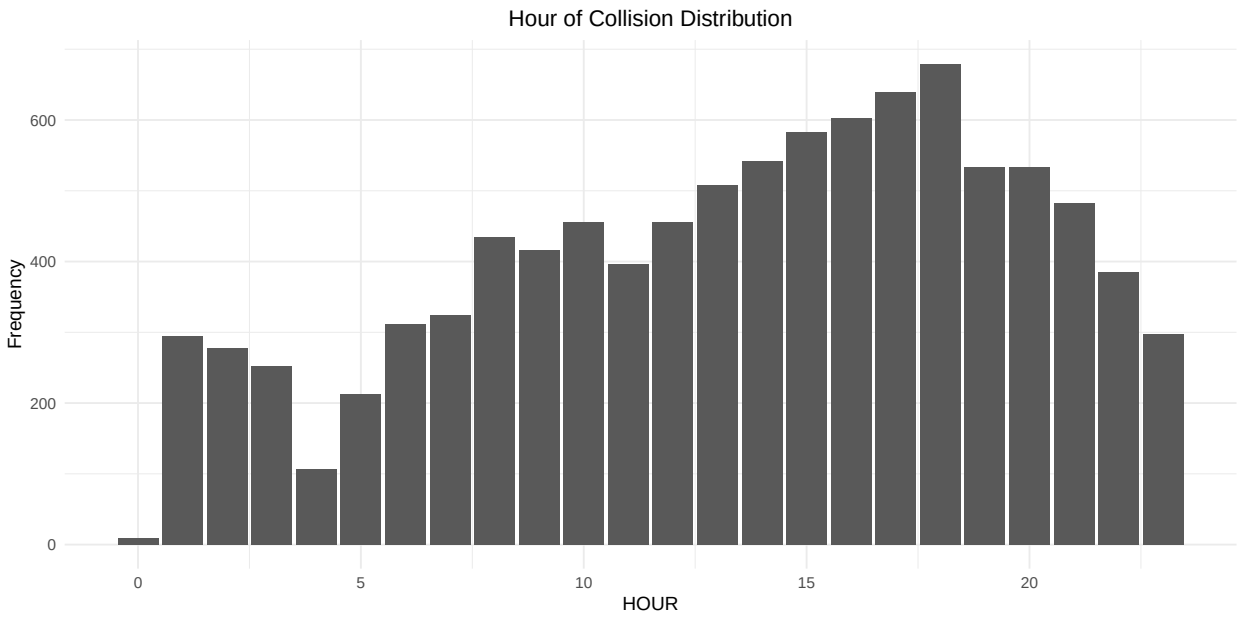

Figure 6: Hour of Collision Distribution

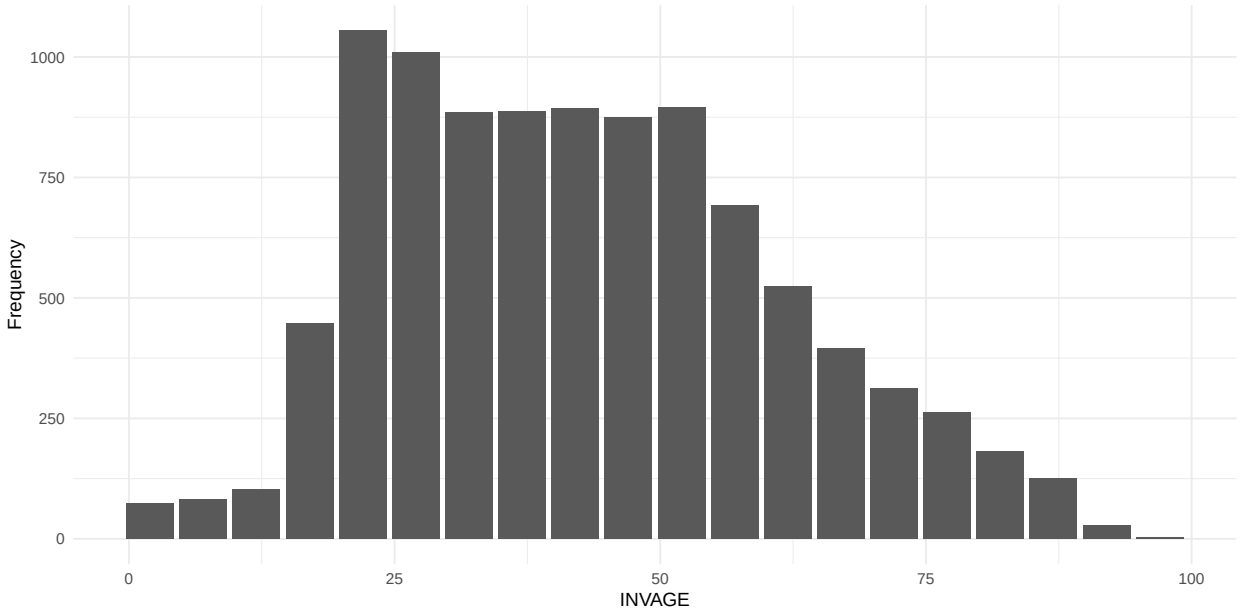

Figure 7: Age Distribution

Table 1: Alcohol Distribution

Alcohol	Frequency
0	9316
1	418

Table 2: Street 1 Distribution

STREET1	Frequency
YONGE ST	218
BATHURST ST	185
DUNDAS ST W	170
EGLINTON AVE E	165
FINCH AVE W	165
DUFFERIN ST	162
EGLINTON AVE W	158
BLOOR ST W	154
STEELES AVE E	153
LAWRENCE AVE E	150

3.2.1 Alcohol

From Table 1, we can see that the majority of the collisions did not involve alcohol, and only around 418 of them consumed alcohol. Hence, when conducting the propensity score matching, we should have 418 pairs, with a total sample of 836.

3.2.2 Injury

From Figure 2, we can see that the majority of the injuries were either none or major, which goes to show the volatility of injury types when it comes to collisions. It is for this reason that people have to take car collisions seriously, as they can result in major injuries.

Table 3: Street 2 Distribution

STREET2	Frequency
BATHURST ST	81
YONGE ST	73
LAWRENCE AVE E	72
FINCH AVE E	69
SHEPPARD AVE E	66
DUNDAS ST W	60
EGLINTON AVE W	59
EGLINTON AVE E	57
ISLINGTON AVE	55
BIRCHMOUNT RD	53
WESTON RD	53

3.2.3 Street 1

From Table 2, we can see the top 10 streets 1 where collisions occurred. At the top, we have Yonge street, which makes sense, as it is the longest road in the world [], and hence its probability of having collisions in it is larger. Nevertheless, the collisions are largely dispersed across 1,140 unique streets.

3.2.4 Street 2

From Table 3, we can see the top 10 street 2 where collisions occurred. Like the previous feature, collisions were largely dispersed across 2,432 unique streets.

3.2.5 Date (Year)

From Figure 3, we can see the distribution of collisions across the years from 2006 to 2019. In fact, it also goes to support our previous research, which is the fact that car collisions have been decreasing over the years, as most of the collisions happen before 2013.

3.2.6 Date (Month)

From Figure 4, we can see the distribution of collisions across all the months. This graph shows that most of the collisions spike up during the latter half of the year. This insight makes sense as it is also supported by our academic review since the summer and winter months mean there may be more people going out for meals due to celebrations and holidays.

3.2.7 Date (Day)

From Figure 5, we can see the distribution of collisions across the days of the month. However, this graph is not as reliable as not all months have a 31st, which is why the 31st may be lower and the 1st has the highest frequency of collisions. All things considered, there seems to be some uniform distribution.

3.2.8 Hour

From Figure 6, we can see the distribution of collisions across the hours of the day. As expected, the collisions are happening later in the day/evening possibly due to the fact that individuals are going out for their dinner meals, which encourage them to consume alcohol. In other words, the distribution is negatively skewed with a mean of 13.44 and a median of 14.

3.2.9 Age

At last, Figure 7 shows us a normal distribution of the age of individuals involved in the collisions. This makes sense as it is the range of ages that are able to drive and go out for meals (as discussed in our academic review). All in all, it had a mean of 42.78 and a median of 42.

3.3 Results

4 Discussion

4.1 Main Findings

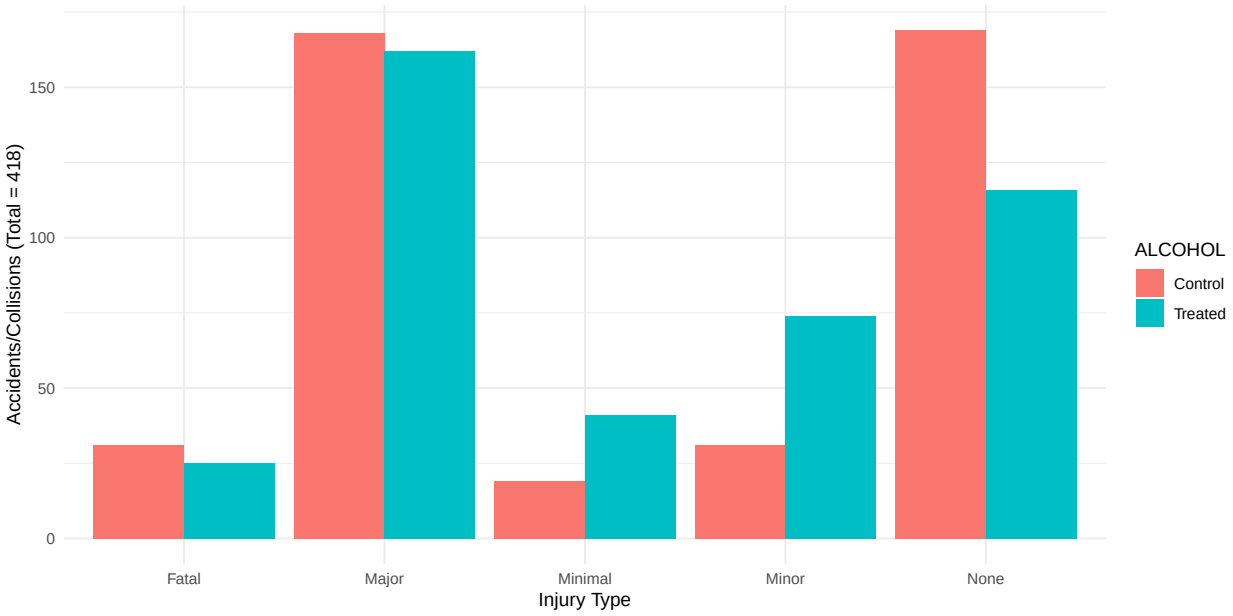

Figure 8: Bar Chart Difference in Injuries Between Treated and Control Groups

Table 4: Percentage Difference in Injuries Between Treated and Control Groups

INJURY	Control	Treated
Fatal	7.42	5.98
Major	40.19	38.76
Minimal	4.55	9.81
Minor	7.42	17.70
None	40.43	27.75

4.2 Limitations

4.3 Future Work

References

"Overview of motor vehicle crashes in 2019." 2019. *National Center for Statistics and Analysis.* https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813060.

S. W. Brown, W. G. M. Vanlaar, and R. D. Robertson. 2017. "The Alcohol and Drug-Crash Problem in Canada 2014 Report." *Canadian Council of Motor Transport Administrators.* https://ccmta.ca/images/publications/pdf/2014_Alcohol_and_Drug_Crash_Problem_Report.pdf.