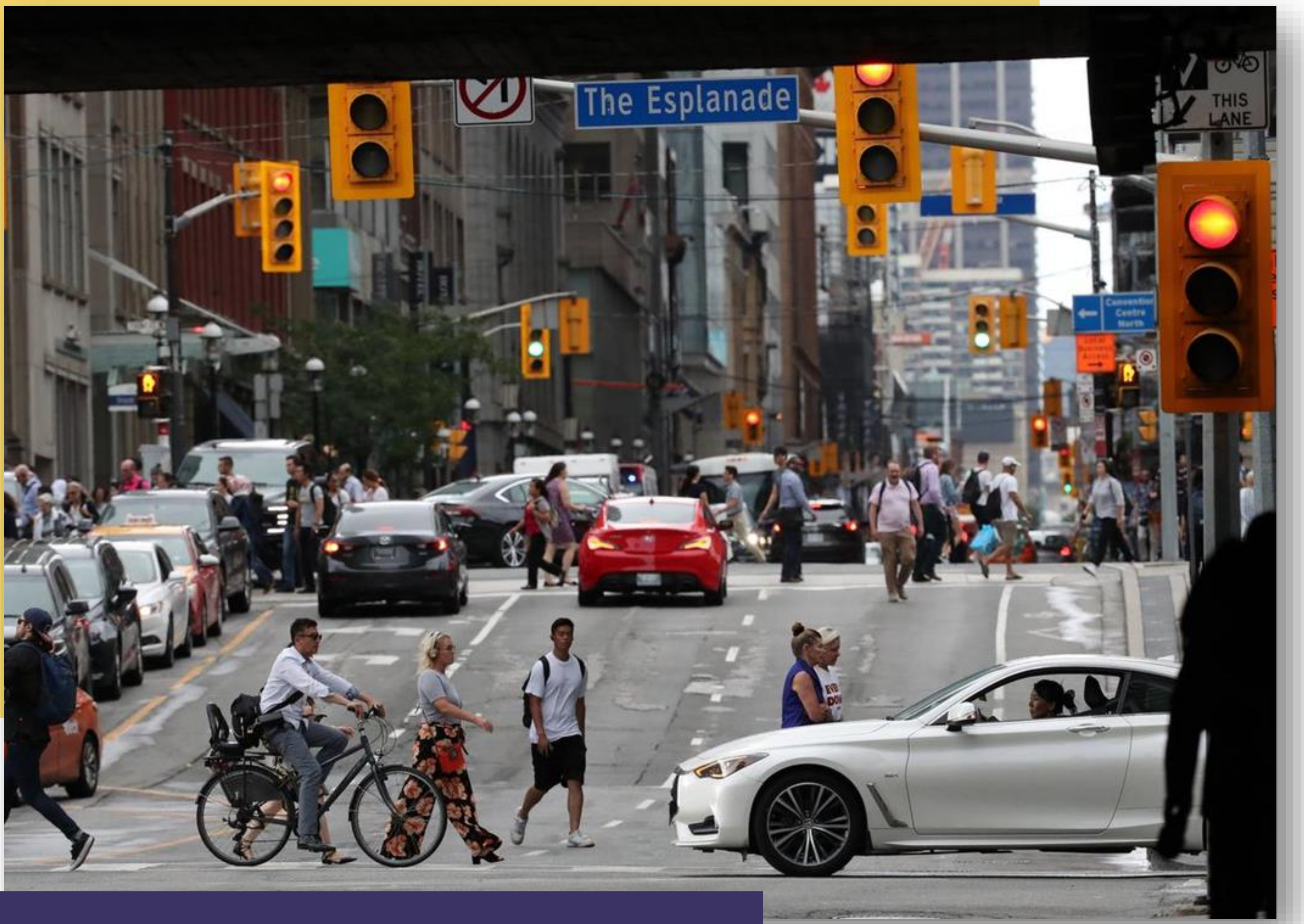


Toronto Traffic Analysis

Foundations of Data Science

Group 1 Bruno de Almeida
Maggie Lau
Fan Ye
Anna Harris
Dennis Norton
Echo Zhang

December 13, 2021

TABLE OF CONTENTS

SUMMARY

Initial Questions

The objective was to analyze vehicle collision data from the City of Toronto's KSI (Killed or Seriously Injured) database to understand the types of collisions that occur and who is involved in those collisions. Some of the initial questions were:

- When do collisions occur?
- Who is involved in collisions?
- How old are the people involved in collisions?
- What type of vehicles are involved in collisions?

Comparison Analysis

Rather than simply analyzing the Toronto collision data in isolation, comparisons were made to Ottawa collision data and Canadian national collision data.

The comparison analysis allowed for potential identification of collision characteristics that are unique to Toronto traffic.

Supplemental Analysis

In addition to analyzing the occurrence of collisions in Toronto, analysis was done on the impact of Red-Light Cameras on collision frequency as well as the availability of public transit in relation to collision locations.

Time Period

The Toronto collisions data covers the period from 2006 to 2020.

2020 Data

2020 included the start of the COVID pandemic which included a lockdown period. During this lockdown period there were fewer people travelling and as a result there was a drop in the number of collisions that occurred. The 2020 data was included in the analysis while noting the impact of the COVID pandemic to the data.

The Ottawa data covers the period of 2013 to 2019 and the national data covers the period from 1999 to 2017. For comparison purposes, the national data prior to 2006 was excluded from the analysis.

Conclusions

1. Toronto has an unusually high frequency of collisions involving pedestrians. The frequency of pedestrian collisions is higher than Ottawa and higher than national data.

> ***"Toronto has an unusually high frequency of collisions involving pedestrians."***

2. Collisions involving a fatality very often involve a pedestrian fatality.
3. A high number of collisions occur in the downtown Toronto core.

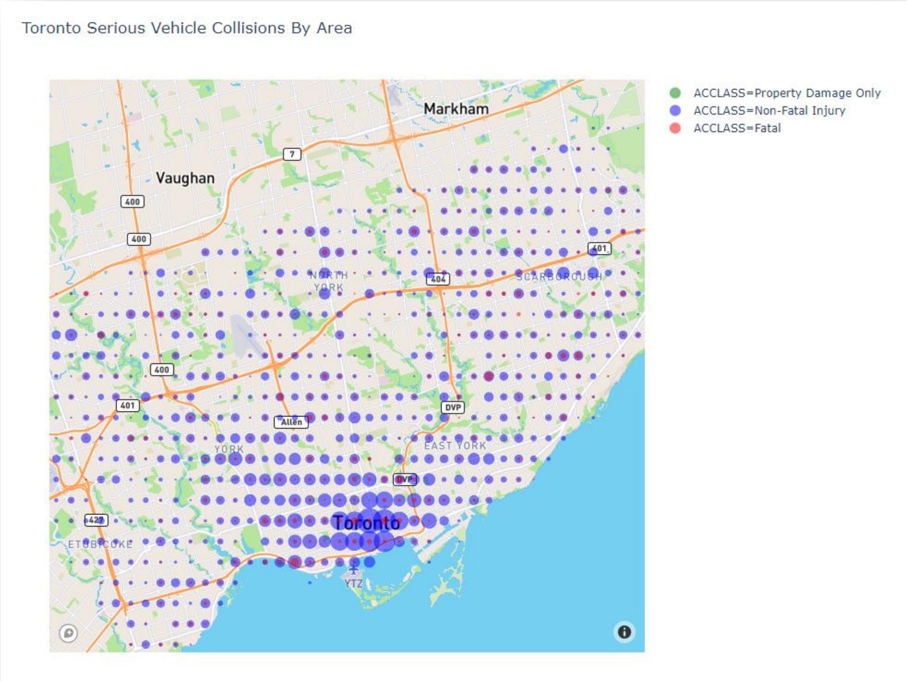


Figure 1: Location of Toronto Serious Vehicle Collisions.

4. Red light cameras do not significantly impact the number of collisions that occur in an area.
5. The likelihood of a Toronto collision causing a fatality could be predicted with over 80% accuracy using the Random Forest Classifier.

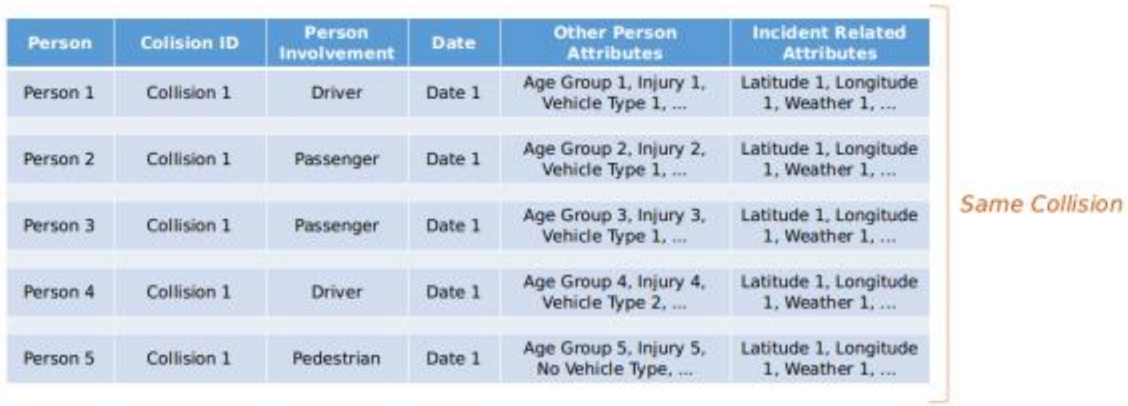
DATA PREPARATION

The primary source of data for this report is the Toronto Police Services Open Data.

Toronto KSI Data

The city of Toronto KSI (Killed or Seriously Injured) Data is a data set provided by the Public Safety Data Portal of Toronto Police Service, and it can be obtained at https://data.torontopolice.on.ca/datasets/ksi/. The file has 16,860 rows and 56 columns.

Each row of the file represents one person involved in each collision. Figure 2 represents how the data is organized for one collision involving five people, being two vehicles with four people in total and one pedestrian:



Person	Colision ID	Person Involvement	Date	Other Person Attributes	Incident Related Attributes
Person 1	Collision 1	Driver	Date 1	Age Group 1, Injury 1, Vehicle Type 1, ...	Latitude 1, Longitude 1, Weather 1, ...
Person 2	Collision 1	Passenger	Date 1	Age Group 2, Injury 2, Vehicle Type 1, ...	Latitude 1, Longitude 1, Weather 1, ...
Person 3	Collision 1	Passenger	Date 1	Age Group 3, Injury 3, Vehicle Type 1, ...	Latitude 1, Longitude 1, Weather 1, ...
Person 4	Collision 1	Driver	Date 1	Age Group 4, Injury 4, Vehicle Type 2, ...	Latitude 1, Longitude 1, Weather 1, ...
Person 5	Collision 1	Pedestrian	Date 1	Age Group 5, Injury 5, No Vehicle Type, ...	Latitude 1, Longitude 1, Weather 1, ...

Same Collision

Figure 2: Structure of Toronto KSI Data.

For each row there are several data fields related to the involved person, such as age group, severity of injury and vehicle type, and also attributes related to the collision, such as date, latitude, longitude, street name and weather.

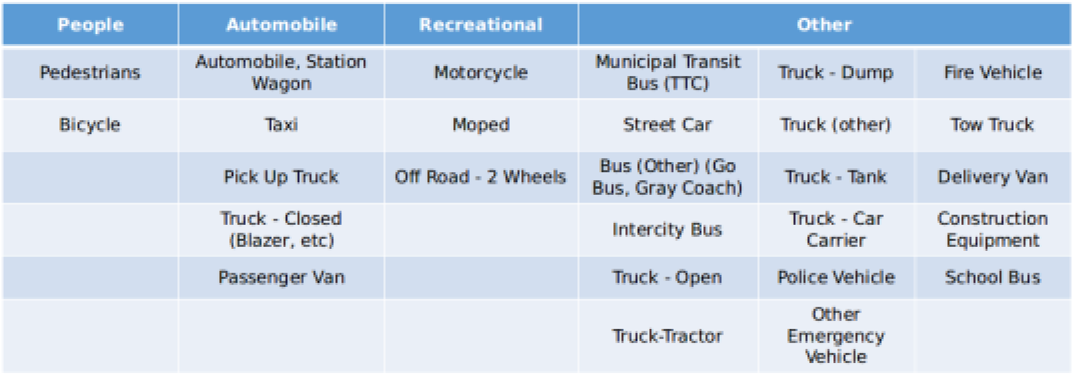
The above example represents only one collision based on the attribute Collision ID (in the original data set this is called ‘ACCNUM’), because the value is the same for all rows. Other common attributes for a collision, such as latitude, longitude, date and weather are also similar among different rows.

In Figure 2 it is possible to conclude that there are two vehicles involved because there are two drivers. On the other hand, there is no way to conclude which car Person 2 and Person 3 are in. This characteristic did not affect the analysis presented here.

There is one attribute representing the collision severity for each person (called ‘INJURY’). This is a categorical attribute varying from ‘None’ to ‘Fatal’. Based on this attribute it is possible to calculate the number of each severity for each collision.

Another important attribute presents the person's involvement in the collision. It can be a ‘Driver’, a ‘Pedestrian’, a ‘Passenger’, a ‘Motorcycle Driver’, and many other options. To support further analysis five categories were created counting the number of ‘Drivers’, ‘Pedestrians’, ‘Passengers’, ‘Cyclists’ and ‘Others’ (such as ‘In-Line Skater’ and ‘Wheelchair’ users) per collision. Additionally, the number of rows with ‘<Null>’ values were also counted.

The vehicle types were also categorized as follow:



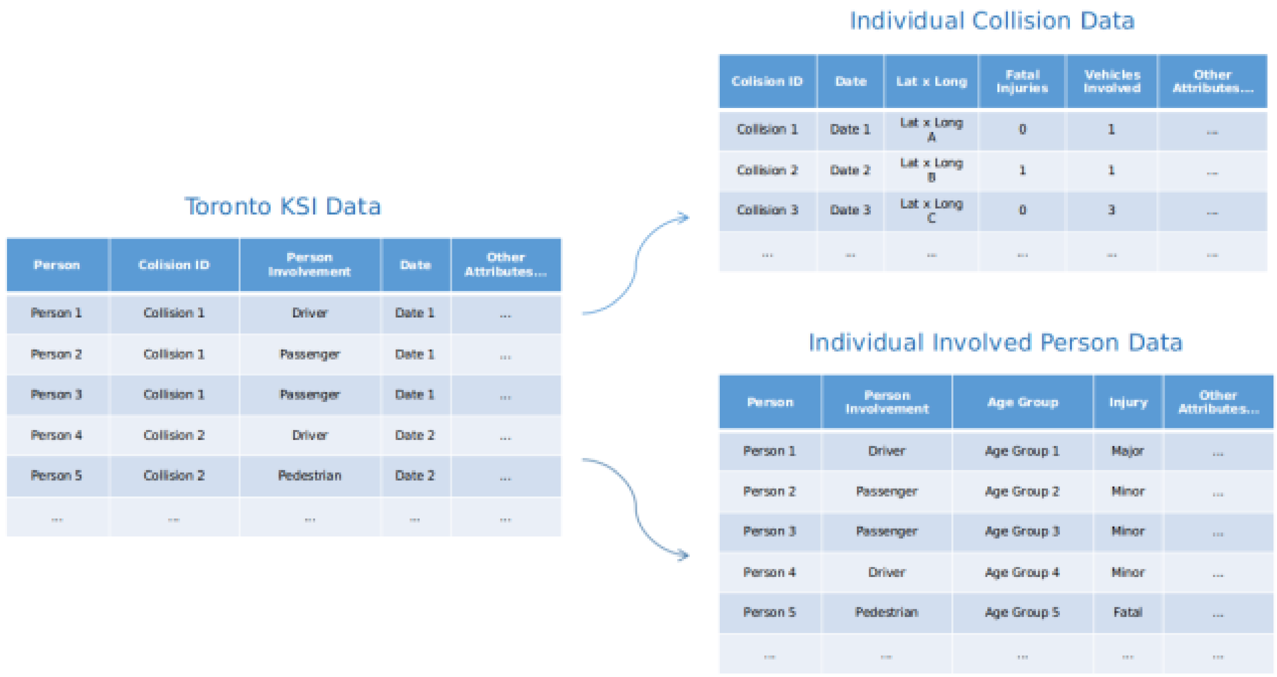
People	Automobile	Recreational	Other		
Pedestrians	Automobile, Station Wagon	Motorcycle	Municipal Transit Bus (TTC)	Truck - Dump	Fire Vehicle
Bicycle	Taxi	Moped	Street Car	Truck (other)	Tow Truck
	Pick Up Truck	Off Road - 2 Wheels	Bus (Other) (Go Bus, Gray Coach)	Truck - Tank	Delivery Van
	Truck - Closed (Blazer, etc)		Intercity Bus	Truck - Car Carrier	Construction Equipment
	Passenger Van		Truck - Open	Police Vehicle	School Bus
			Truck-Tractor	Other Emergency Vehicle	

Figure 3: Vehicle Categories for Toronto KSI Data.

Regarding ages, the original data set provides only age groups for each person involved. They vary in groups of five years from zero to 94 years old. Additionally, this data file does not present the gender of each person.

Finally, the original data set was enriched with additional features, such as time interval, season and month. These attributes will be useful during the analysis stage.

As part of the data clean-up, the data was separated into two different files: in the first one the rows represent a summary of each collision, and in the second rows present the most important information related to each person involved, being similar to the original data set with less attributes. Figure 4 represents this separation:



Toronto KSI Data

Person	Collision ID	Person Involvement	Date	Other Attributes…
Person 1	Collision 1	Driver	Date 1	…
Person 2	Collision 1	Passenger	Date 1	…
Person 3	Collision 1	Passenger	Date 1	…
Person 4	Collision 2	Driver	Date 2	…
Person 5	Collision 2	Pedestrian	Date 2	…
…	…	…	…	…

Individual Collision Data

Collision ID	Date	Lat x Long	Fatal Injuries	Vehicles Involved	Other Attributes…
Collision 1	Date 1	Lat x Long A	0	1	…
Collision 2	Date 2	Lat x Long B	1	1	…
Collision 3	Date 3	Lat x Long C	0	3	…
…	…	…	…	…	…

Individual Involved Person Data

Person	Person Involvement	Age Group	Injury	Other Attributes…
Person 1	Driver	Age Group 1	Major	…
Person 2	Passenger	Age Group 2	Minor	…
Person 3	Passenger	Age Group 3	Minor	…
Person 4	Driver	Age Group 4	Minor	…
Person 5	Pedestrian	Age Group 5	Fatal	…
…	…	…	…	…

Figure 4: Obtained Data Sets from Toronto KSI Data.

City of Ottawa Collision Data

The city of Ottawa organizes their vehicle collision data in separate files for each year from 2013 to 2019. It was necessary to combine the files into one set for ease of analysis. Each year has slight variation in the type of data collected and how it is organized. Cleanup was required to ensure uniformity in column names and data format. The year 2017 in particular formatted the time and date data as epoch time in milliseconds. The fromtimestamp function from the datetime library was used to convert back to readable datetime format.

The files were then combined into one dataframe for analysis. Columns that did not contain meaningful data were removed from the dataframe (i.e. ID numbers). Of the remaining fields, there are 2 columns with significant amounts of missing values:

- ‘TRAFFIC_CONTROL_CONDITION’,
- ‘NUMBER OF PEDESTRIANS’, and
- ‘COLLISION_LOCATION’.

To avoid potential issues later on during analysis, in most cases the missing values were filled with ‘Unknown’ as a dummy value.

However, for the column ‘TRAFFIC_CONTROL_CONDITION’, if the ‘TRAFFIC_CONTROL’ was noted as “10 - No control”, this column was left blank since it would be misleading to fill in as ‘00 - Unknown’, these values were filled as a new variable ‘10 - No control’. Note that not all blanks were due to “10 - No control”, so the remaining null values were filled as “00 - Unknown”.

New variables were also engineered for the Ottawa dataset to use in comparative analysis with the Toronto dataset. These consisted of ‘Season’, ‘Week Day’, ‘Time’.

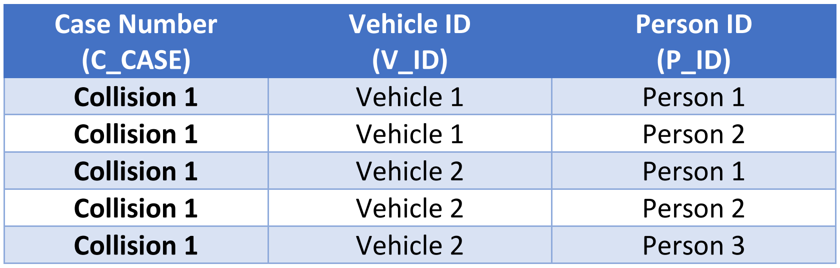
National Collision Data

The national collision dataset is a large file with 6,772,563 rows and 23 columns. There are no null values in the data, however there are values to indicate where data was either not available, not applicable, or of a value not specified as a valid value for the column.

The national collision data is a combination of three types of data;

1. Collision data
2. Vehicle data
3. People data

Figure 5 is an illustration to show how the data was organized. If there was a collision involving two vehicles with two people in the first vehicle and three people in the second vehicle, the dataset would have five rows to represent this collision.



Case Number (C_CASE)	Vehicle ID (V_ID)	Person ID (P_ID)
Collision 1	Vehicle 1	Person 1
Collision 1	Vehicle 1	Person 2
Collision 1	Vehicle 2	Person 1
Collision 1	Vehicle 2	Person 2
Collision 1	Vehicle 2	Person 3

Figure 5: Structure of Toronto KSI Data.

The collision data includes information such as time, date, road conditions, weather, and number of vehicles involved. This information is repeated for each person involved in the collision, in the example above the information would be repeated five times.

Further information about the national collision data can be found in the appendix.

City of Toronto Red Light Camera Data

The red light camera dataset of Toronto has 194 camera activation items with attributes of “camera ID”, “intersection positions”, “activation dates”, “district and police division info”, etc. To explore the correlation of red light cameras and collisions, we only need the location information and the activation time. Hence the red light camera dataset was cleaned to only keep several necessary attributes including “Camera ID”, “Intersection ID”, “Road Name”, “Latitude”, “Longitude” and “Activation time”. The activation time was transformed to datetime datatype for further analysis. Additionally, some cameras were activated twice at the same positions so that the second time of activation was deleted. After clean up, there are 189 intersections in Toronto monitored with redlight cameras.

To analyze the relationship between collisions and red light cameras, the two dataframes were merged. The first work is to filter the KSI collision dataframe to find the collisions that happened at intersections where red light cameras were installed. The latitude and longitude of each collision were compared with that of the cameras and a cutoff distance of 5*10^(-4) for both latitude and longitude was applied to only select the collisions that happened near the intersections where the cameras were installed. Then the filtered dataframe was further adjusted by selecting the collisions that were only marked with “at intersections”, “intersection related” or “<Null>”. 69 collisions without location info were remarked as “intersection related”. Finally, the number of collisions that happened at the intersections with red light cameras is 515.

Given the cameras were activated at different times between 2007 and 2021, only the cameras that were activated before 2020 were used for analysis as the collision dataset is only updated to 2020. Similarly, only the collisions that happened before 2020 were analyzed. Then the numbers of cameras and collisions were reduced to 126 and 381 respectively. Furthermore, the role of cameras was analyzed by comparing the number of collisions that happened one full year before or after the activation of cameras.

City of Toronto Transit Data

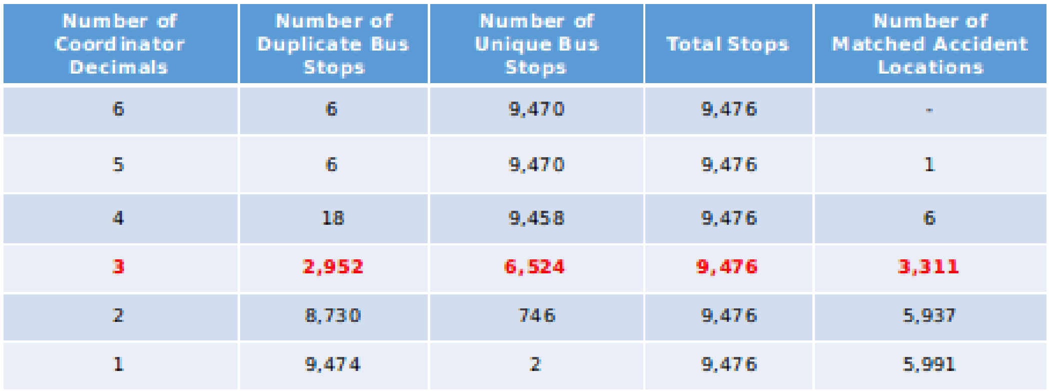
The Toronto Transit Commission (TTC) Routes and Schedules data contains scheduling information (route definitions, stop patterns, stop locations, and schedules). The data is published as a GTFS (General Transit Feed Specification) file format and is publicly available on the Toronto Open data site. The TTC Routes and Schedules zip file contains route definitions, stop patterns, stop locations and schedules. For this report, we will be using the stops.txt file.

There are 12 variables (columns) included in stops.txt file, 50% of them are with null values. After removing those empty or non-relevant variables, we keep the following attributes:

- stop_id: bus stop ids
- stop_code: bus stop codes
- stop_name: bus stop names
- stop_lat: bus stop latitude
- stop_lon: bus stop longitude

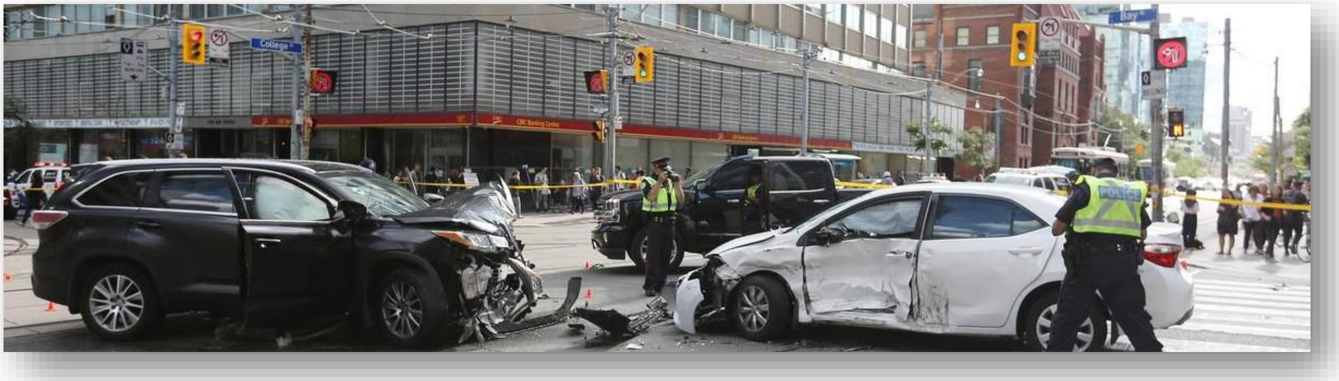
There are 9,476 bus stops and 5,999 collisions in Toronto. The geographical coordinators from the bus stops and from KSI collisions are not exactly matched. We tested the coordinators with decimals from one to six. There are different levels of duplications (as shown in Table 1). Reducing the decimals to three would be the best option for our case with a fare duplication level from bus stops matching with collision locations. To analyze the relationship between collisions and bus stops, we rounded the coordinators to three decimals. Therefore, 55.2% of the collision locations are aligned with one or more than one bus stops.

Table 1: Analysis of the Number of Coordinator Decimals



Number of Coordinator Decimals	Number of Duplicate Bus Stops	Number of Unique Bus Stops	Total Stops	Number of Matched Accident Locations
6	6	9,470	9,476	-
5	6	9,470	9,476	1
4	18	9,458	9,476	6
3	2,952	6,524	9,476	3,311
2	8,730	746	9,476	5,937
1	9,474	2	9,476	5,991

DATA ANALYSIS



Considerations

When comparing the Toronto collision data to the Ottawa data and national data, some factors needed to be taken into account:

Criteria for the data file

- The Toronto data is based on the involvement of a severe injury or fatality
- The national data is for all collisions and includes an indicator for the severity of the injuries.

Time Frame covered

- Each dataset covers a different time frame, but there are overlapping periods in the data where comparisons can be made.
- The Toronto data includes data for 2020 which covers the start of the COVID pandemic. Pandemic related lockdowns dramatically impacted travel patterns and as a result impacted vehicle collision volume.

Fatalities

This section will present an overview of the number of fatalities according to some attributes in the KSI Toronto data file. To carry out this evaluation it was considered the file generated after the cleaning process with Individual Involved Person Data. There were 16,860 records, where 821 people were fatal victims.

The following attributes were individually evaluated:

- Age group: in the original dataset, ages are categorized in groups such as ‘0 to 4’, ‘5 to 9’, ‘10 to 14’, and so on until 94 years old,

- Person involvement: if the person was a driver, a cyclist, a passenger, etc;
- Vehicle type: possible records are ’automobile, station wagon’, bicycle’, ‘motorcycle’, etc,
- Vehicle category: as presented earlier, the vehicles were categorized in four groups – ‘People’, ‘Automobile’, ‘Recreational’ and ‘Other’. Some vehicles could not be categorized based on missing information in the original dataset, being shown here as ‘Not Recorded’,
- Hour of day: the data were grouped in time intervals of one hour,
- Day of Week,
- Month,
- Year, and
- Season.

The analysis considers two scenarios:

1. Considering only the set of fatal victims, how each attribute is distributed among this group.
2. Looking at each attribute, the percentage of fatal victims among all people involved in collisions.

Results

The first observation that was not anticipated is that the number of collisions involving pedestrians is almost as high as vehicle collisions.

The highest number of collisions occur during the afternoon commute home with the peak being during the Friday afternoon commute.

Saturday and Sunday do not demonstrate the spike in collision numbers seen during Monday through Friday and instead have more of a steady rate of collisions.

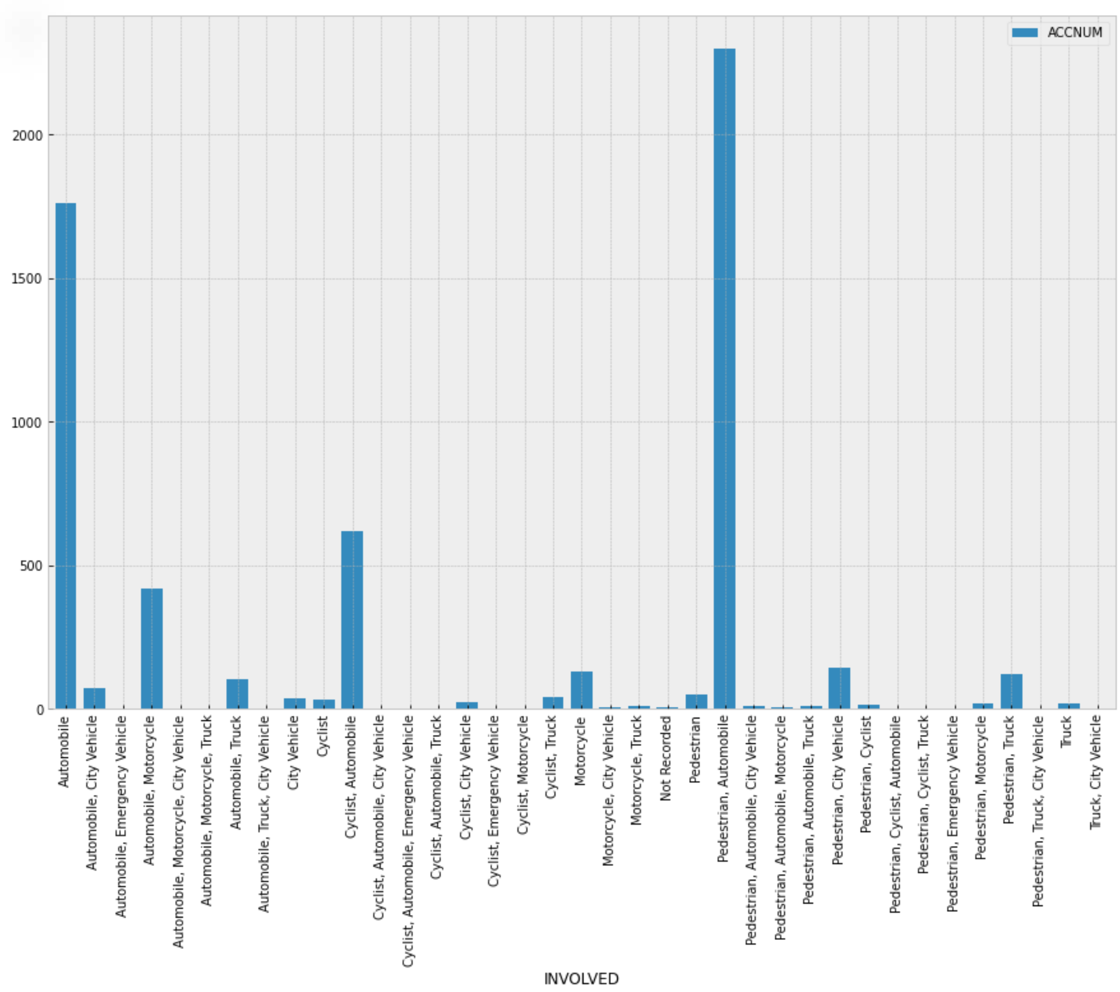


Figure 6: Number of Collisions Per Vehicle and People Involvement.

The following are the main answers obtained during this analysis. Further details can be seen at Appendix F.

Age Group: Among fatalities, 8.0% were between 20 to 24 years old, being the most representative group. However, the distribution of other age groups is quite similar, the exception being children (between zero and 14 years old), where the total number of fatalities is very low in comparison to other groups. This finding could mean that school safety zones are working to make these areas safer for children.

Looking at the percentage of victims for each group age, there is an upward trend as the age increases, where 37.5% of seniors with 90 to 94 years involved in collisions died, in contrast to 4.10% for victims between 20 and 24 years old. This is not unexpected, as the elder are more fragile and susceptible to greater injuries

Person Involvement: This analysis presents how collisions are dangerous for pedestrians who represent 56.3% of all fatalities. In fact, 16.09% of all pedestrians involved in a collision received fatal injuries.

Vehicle Type: The data is unclear, where 40.2% of fatal victims have an unknown vehicle type recorded, followed by 27.0% of 'Other' types. We must remember that this attribute was not treated, being the information obtained from the original data set. Looking at the percentages of fatal victims for each group we see that 12.01% of motorcycle users died, being the worst result, showing how dangerous collisions are for them.

Vehicle Category: this attribute is obtained from the previous one, and again the results are not precise: 40.2% are considered 'Null'. Similarly, 11.73% of victims in 'Null' vehicles died, being the highest value obtained. The second one is the 'Recreational' category, which includes motorcycles, matching with previous results.

Hour of Day: The data are quite scattered. Looking at the total number of fatal victims, 7.3% of the collisions occurred between 18:00 and 18:59, being the worst result. However, there is an interesting result from the perspective of percentages: 9.44% of victims between 05:00 and 05:59 died, a result significantly higher than the next worst result, which is 6.64% of fatal victims between 06:00 and 06:59. To explain this, two possibilities were theorized:

1. At this time traffic is lighter and likely travelling at higher speeds. These higher speeds could account for the higher rate of fatality. More analysis would be required to validate this theory.
2. It could also be related to driver impairment due to alcohol, drugs, or fatigue, from being up late or waking up early. An impaired driver is less to take action to avoid the collision and as a result impact the pedestrian at a higher speed. More analysis would be required to validate this theory.

It was observed that the majority of 27 fatalities on this time interval occurred on Fridays (seven cases), Mondays (six cases) and Tuesdays (six cases), allowing us to discard option 2 for weekends. Besides, 10 of these victims were pedestrians, strengthening thesis 1. Although there were also 10 drivers in the victim groups, any of them were recorded as under alcohol influence. So, for these 10 driver victims, we can conclude that the possibility to drive faster and have a collision increases the chance of a fatality.

Day of Week: In absolute numbers, the worst day is Friday, with 17.4% of fatal victims, followed by Tuesdays with 16.8%. The traffic is usually higher on Fridays because many people decide to travel out of Toronto at the end of the workday. For the rest of weekdays (Tuesday, Wednesday and Thursdays) the results are quite similar and a bit better than Fridays. On the other hand, it is interesting to note that Monday is the safest day, even better than weekends.

Month: September and August concentrate the highest number of fatal victims' occurrences, with 10.8% and 9.7% respectively, meaning that there are more collisions with fatalities in the Summer. However, in January 6.64% of people involved in collsions died, being the worst result. This could be related to the weather conditions. Interestingly, between February and June there is a downward trend, and from July the results start increasing gradually until Winter.

Season: it confirms what was shown previously, when Summer is the worst result in absolute numbers and Winter owns the worst result of fatalities among victims.

Year: from 2006 to 2020, as presented at this dataset, the worst year in number of collisions were 2016, concentrating 9.5% of occurrences, followed by 2018 with 8.0%. 2016 was also the most dangerous year, when 7.75% of people involved in collisions were fatal victims. There is an upward trend between 2011 and 2016, and since then there has been a reduction in the percentage of fatalities.

Finally, it is clear the impact of COVID-19 on 2020 results, where the total number of fatalities was the lowest since 2011, however the percentage of fatalities was still high, similar to 2017-2018 results.

Red Light Cameras: There is no evidence that red light cameras significantly reduce the number of collisions at an intersection.

Toronto Public Transit: Transit options are available in areas of high collisions.

APPENDIX

Appendix A: National Collision Data – Data Preparation

The data contains a collision severity value, at the collision level, to specify if a fatality was involved. The value is a flag and does not indicate the number of fatalities. In the previous example there were five people involved in the collision. If the collision severity indicates that the collision was fatal, it is not known from this information if there were 1, 2, 3, 4, or 5 fatalities.

The vehicle data includes the type of vehicle and year of the vehicle. This information is repeated for each passenger in the vehicle. The people data includes information such as age, gender, severity of injury, and location in the vehicle. As part of the data clean-up, the data was separated into three separate files and the duplicated information was removed.

A full data dictionary (provided with the data) was needed to interpret the values in each of the columns. The values are primarily categorical numeric data. For example, the person's location in the vehicle is a 2-digit number with the first digit representing the row in the vehicle and the second digit representing the seat in the row. The driver is in location 11, the front row passenger is in location 13 (some vehicles have a front middle seat and could have a passenger in location 12).

Another example is the collision configuration which is a 2-digit number representing a description of the collision. The first digit is 0 for a single vehicle collision, is 2 for a two-vehicle collision with both vehicles travelling in the same direction, is 3 for two vehicles traveling in different directions, and is 4 if one of the vehicles was parked. The second digit has varying meanings depending on the first digit.

There were no null values in the data, but there was missing data. In situations where data is not known or not available, one of the following alpha codes was used:

- N - not applicable. For example, the vehicle type is not applicable for a pedestrian.
- Q - other value. For example, if a Zamboni was involved in a collision there is no code for this type of vehicle.
- U - unknown. For example, in the case of a hit and run, some information will not be known. This value is also used in situations involving run-away vehicles. For example, if a vehicle was left in neutral and rolled down a hill until it collided with something, the person's age, gender, location in the vehicle, and injury would all be noted as U.
- X - not provided. This is used when the reporting jurisdiction does not provide this particular data.

For numeric columns, these alpha values were converted to numeric values to allow the column to have an integer data type. For vehicle type and injury severity the alpha values were converted to zero. For other numeric columns the alpha values were converted to -1.

The amount of missing data is very low. The month information is missing only 0.00625% of the time and the hour of the collision is missing only 0.979% of the time. This is very low and will not impact the analysis.

There are some columns with higher percentages of missing data, such as vehicle type which is not specified 4.9% of the time. A missing value in vehicle type is valid in the case of a pedestrian, so this percentage is more than reasonable.

Road configuration and vehicle year are missing 10.7% and 9.9% of the time respectively. These are not used in the comparison to the Toronto data and are not a concern. Also not used in the comparison to the Toronto data is information regarding the safety equipment being used by the individual in the collision (for example, were they wearing a seatbelt). The safety equipment information was missing 21.2% of the time.

The collision ID (known as case ID), vehicle ID, and person ID were all converted to string values by adding a 'C', 'V', and 'P' respectively to the start of each value. This was done to avoid potential issues later if using these values as indices.

New columns were engineered to simplify the analysis:

- Vehicle type:
    - Person - pedestrian or cyclist
    - Recreational - moped, motorcycle, snowmobile
    - Automobile - non-commercial family vehicles such as sedans, pick-up trucks, SUVs, etc.
    - Other - bus, truck, ambulance, etc.
- Collision type:
    - Auto/Person
    - Auto/Auto
    - Auto/Recreational
    - Auto/Other
    - Non-Auto - an collision involving two trucks, or an collision involving only a single motorcycle, etc.
- Summary columns to indicate:
    - Number of vehicles by type in a collision
    - Number of injuries by severity by
- driver,
- auto passenger,
- recreational passenger,
- pedestrian,
- other.

Appendix B: Ottawa Data Analysis

The Ottawa data set is limited in terms of variety and detail of data collected as compared to the Toronto data set. Unfortunately, the dataset is not split into KSI (killed or seriously injured), there is only a classifier for injuries, but it is unknown which of the injuries are serious and which are minor. Consequently, only fatal collisions can be compared across the datasets. The Ottawa data also only spans years 2013-2019. Therefore, the scope of analysis is restricted to the available data common across both cities.

The number of collisions involving pedestrians in Ottawa is displayed in Figure 7. Note, there was no data collected on pedestrians in 2017. By far the majority of collisions in Ottawa do not involve pedestrians. Note that this graph includes all collisions, it is not limited to killed or serious injuries.

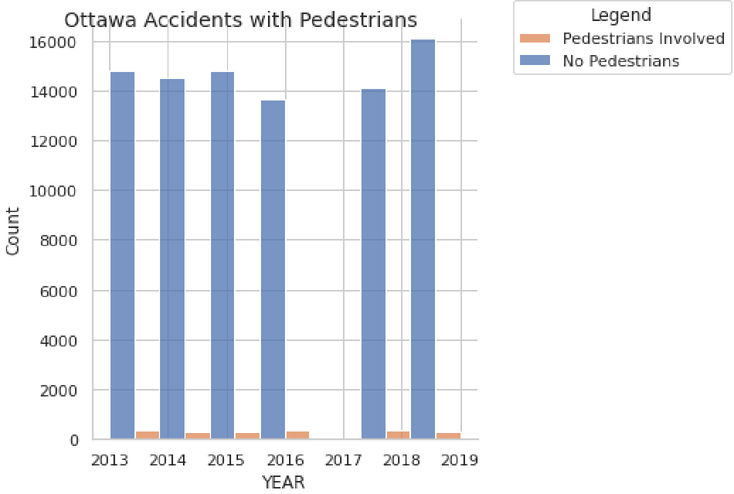


Figure 7: Number of Collisions Involving Pedestrians in Ottawa.

To better compare with the Toronto data, we will filter out the Ottawa dataset for only collisions with fatalities. When we do so, the proportion involving pedestrians increases. So, on the rare occasion a pedestrian collision occurs, it is more likely for it to be a serious collision causing fatalities. When comparing the two cities, there is a much higher proportion of pedestrian involvement in fatal collisions in Toronto than Ottawa.

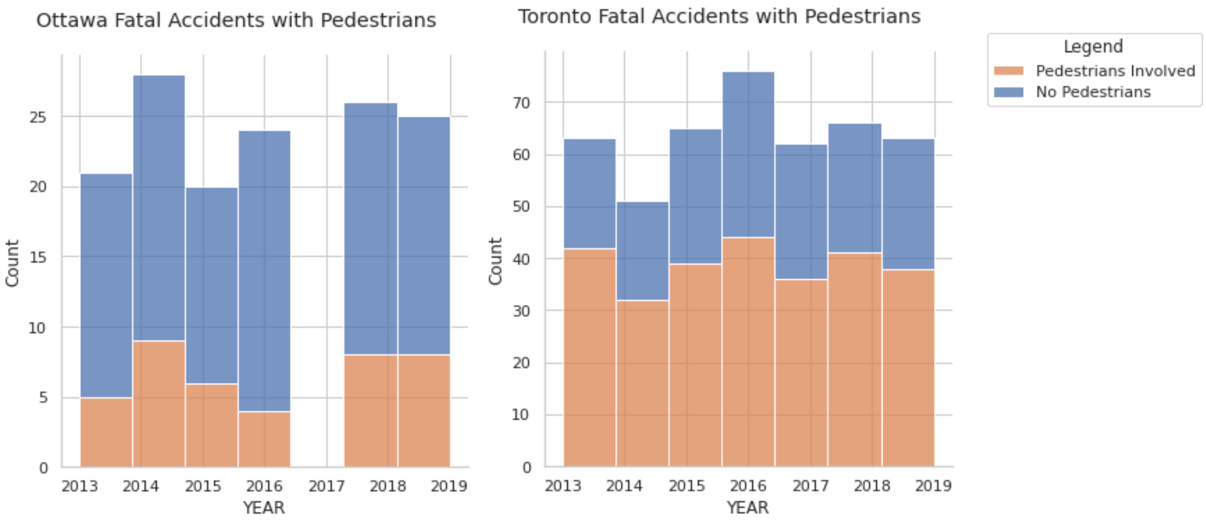


Figure 8: Comparison of Fatal Collisions Involving Pedestrians in Ottawa and Toronto.

The fatal collisions are categorized by season as well to determine which season is the most dangerous. Surprisingly, the most Ottawa fatal collisions occur during the summer, though interestingly Autumn is not far behind and winter is the last. A possible explanation is during the summer months there are more recreational trips with drivers unfamiliar with the area. Another noteworthy observation is Ottawa pedestrian fatalities proportionately are highest in Autumn, perhaps a consequence of the time change and increasing darkness. Initial expectations were that we would see a similar distribution in Toronto, however the data shows that Autumn is actually the season with the most fatalities. Potentially there is more commuter activity that is affected by the Autumn time change.

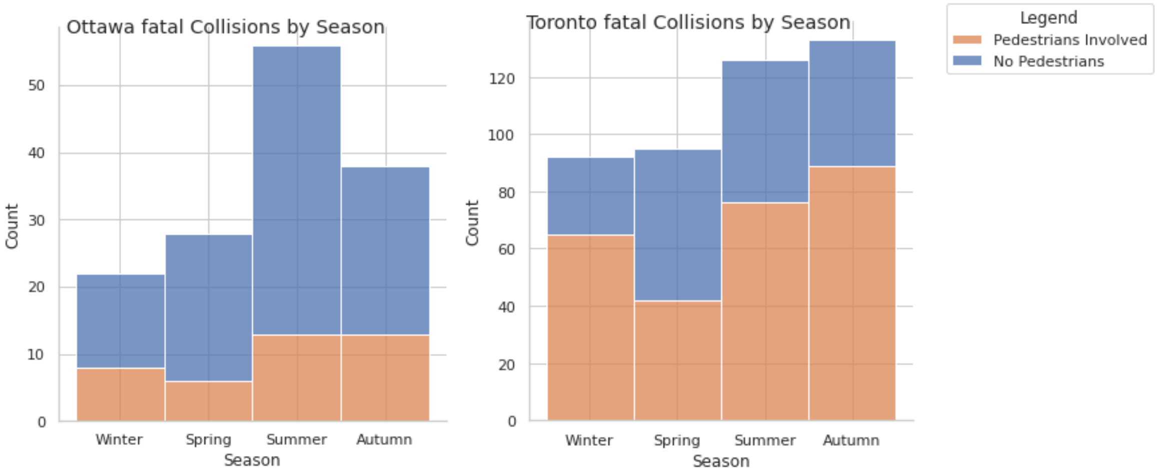


Figure 9: Comparison of Fatal Collisions Involving Pedestrians in Ottawa and Toronto by Season.

Number of collisions are categorized by day of the week (Monday is 0, Sunday is 6). In the city of Ottawa, the highest risk day for fatal collisions in general appears to be Friday for both vehicles and pedestrians. In the city of Toronto, there does not seem to be an outlier day but rather Tuesdays-Fridays all seem similarly risky for fatal collisions involving both vehicles and pedestrians.

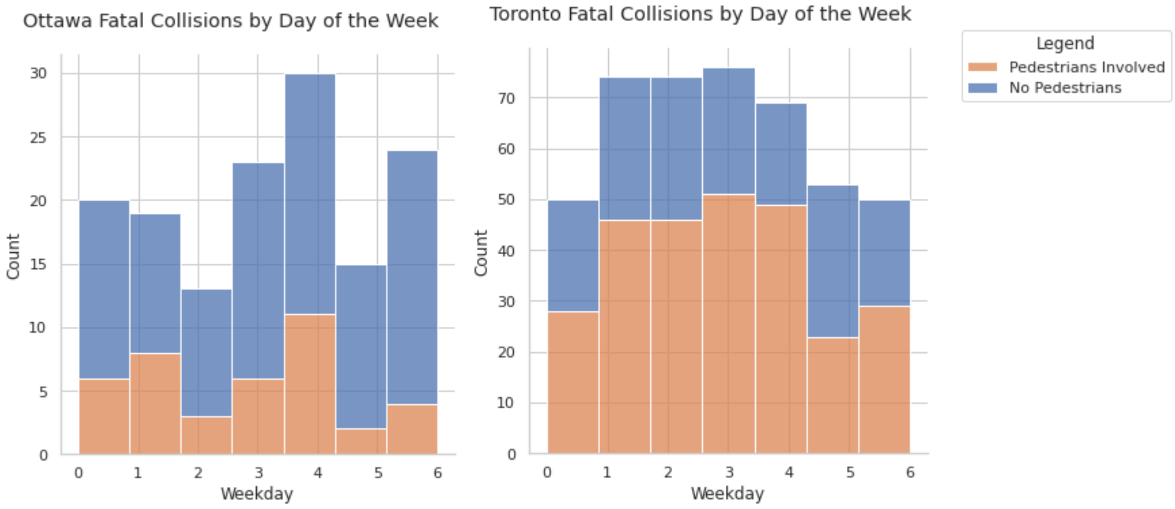


Figure 10: Comparison of Fatal Collisions Involving Pedestrians in Ottawa and Toronto by Day of Week.

Number of fatal collisions are also categorized by time (hour of the day). In Ottawa, fatal collisions appear to be the highest around the noon lunch hour overall, but for pedestrians it is highest at 7am and 5pm. These hours seem to line up with commute times. In Toronto the most fatal collisions occurred at 6pm by far, lining up with the evening rush hour commute. Pedestrian fatalities are high during the evening commute hours as well, with 8pm topping the charts.

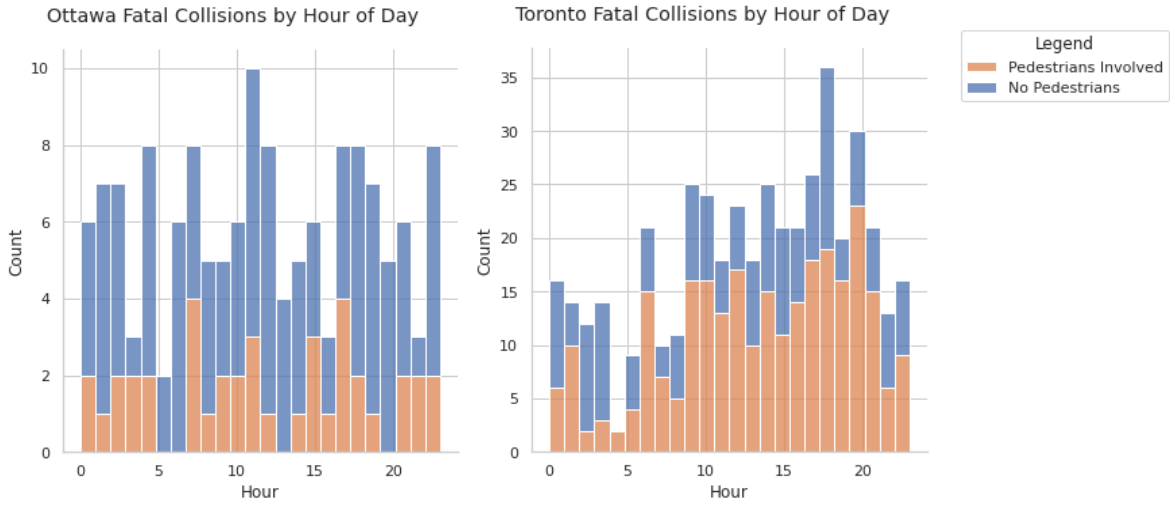


Figure 11: Comparison of Fatal Collisions Involving Pedestrians in Ottawa and Toronto by Hour of Day.

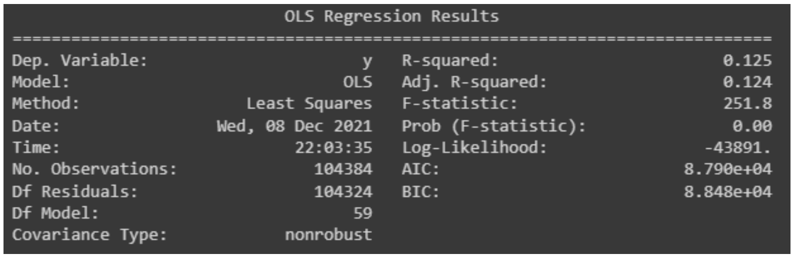
Appendix C: Collision Modelling

Ottawa

An attempt to model collisions was made, the objective was to determine if a model can predict the classification of a collision (Property damage only, injuries, or fatalities) based on the characteristics of the collision.

The features that were initially selected for the model were: ‘ENVIRONMENT', 'LIGHT', 'ROAD_SURFACE', 'TRAFFIC_CONTROL', 'TRAFFIC_CONTROL_CONDITION', 'IMPACT_TYPE', 'NUMBER_OF_PEDESTRIANS', ‘SEASON’ and ‘HOUR’. In order to prepare these features for use in modelling, all categorical features were converted to dummies.

The first model used was the ordinary least squares model. The OLS Regression model resulted in a R-squared value of 0.125 and majority of features with P-values > 0.05 so this model was not a good fit for this dataset.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.125
Model:                            OLS   Adj. R-squared:                  0.124
Method:                 Least Squares   F-statistic:                     251.8
Date:                Wed, 08 Dec 2021   Prob (F-statistic):               0.00
Time:                        22:03:35   Log-Likelihood:                -43891.
No. Observations:              104384   AIC:                         8.790e+04
Df Residuals:                  104324   BIC:                         8.848e+04
Df Model:                          59
Covariance Type:            nonrobust
```

Figure 12: OLS Regression Results.

Another attempt was made to fit the data using the Random Forest Classifiers model using 9 features and n=100 decision trees. This model resulted in a prediction accuracy of 80.8%. The feature importances Pandas function was used to identify which features contributed most to the model. It was found that ‘Hour’ was the top feature by far, followed by ‘NUMBER_OF_PEDESTRIANS’ (boolean value for whether pedestrians were involved in the collision), and impact type.

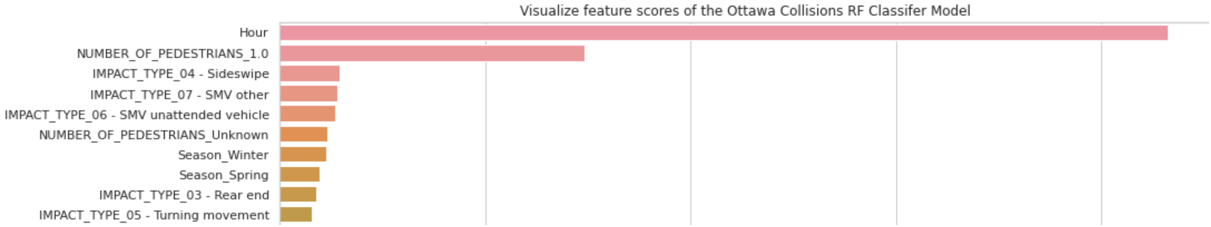
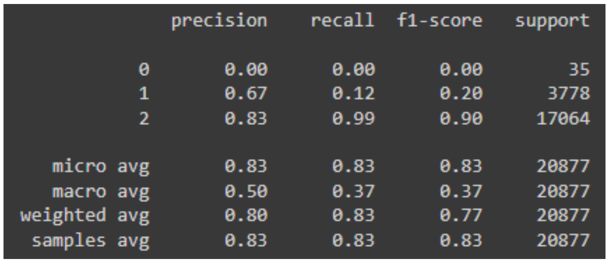


Figure 13: Feature Scores of Random Forest Classifiers (9 Features and 100 Decision Trees) for Ottawa.

To further improve the model, the feature set was narrowed down based on the findings of the feature importance to 5 features: ROAD_SURFACE, IMPACT_TYPE, NUMBER_OF_PEDESTRIANS, Season, and Hour. This marginally improved accuracy to

82.8%. The performance of the model with sklearn's classification_report function results is shown in Figure 14.



```
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        35
           1       0.67      0.12      0.20      3778
           2       0.83      0.99      0.90     17064

   micro avg       0.83      0.83      0.83     20877
   macro avg       0.50      0.37      0.37     20877
weighted avg       0.80      0.83      0.77     20877
 samples avg       0.83      0.83      0.83     20877
```

Figure 14: Random Forest Classifier Results for Ottawa.

In the dataset, the majority of collisions belong to class "2" for Property Damage at 17,064 count, versus class "1" for Injury Only at 3,778 count, and class "0" for Fatalities at only 35. With the limited amount of data points for fatalities, it is not possible for the model to accurately predict this scenario resulting in 0 scores for class 0.

Toronto

The same modelling method using the Random Forest Classifiers model using 8 features and n=100 decision trees was applied to Toronto data. The difference being the Toronto dataset only includes 2 collision classes: Injury and Fatalities. The initial features used are as follows: 'VISIBILITY', 'LIGHT', 'RDSFCOND', 'TRAFFCTL', 'IMPACTYPE', 'INVOLVED', 'Season', 'Hour'. This model resulted in a model accuracy of 81.75% and feature importance as listed in Figure 15.

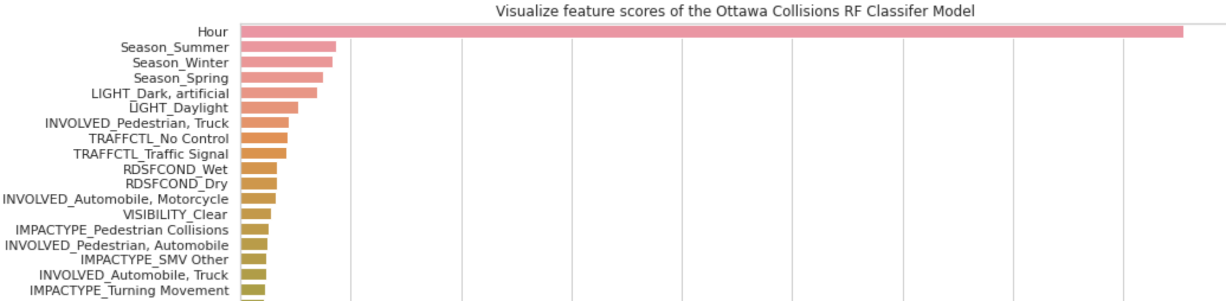
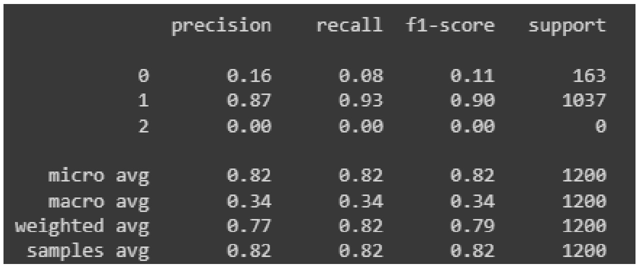


Figure 15: Feature Scores of Random Forest Classifiers (8 Features and 100 Decision Trees) for Toronto.

Interestingly once again 'Hour' is the top feature for both Toronto and Ottawa. Season is the next most important feature for Toronto and also is a top important feature for Ottawa. The performance of the model for Toronto is evaluated with classification_report function and results are shown in Figure 16.



```
              precision    recall  f1-score   support

           0       0.16      0.08      0.11       163
           1       0.87      0.93      0.90      1037
           2       0.00      0.00      0.00         0

   micro avg       0.82      0.82      0.82      1200
   macro avg       0.34      0.34      0.34      1200
weighted avg       0.77      0.82      0.79      1200
 samples avg       0.82      0.82      0.82      1200
```

Figure 16: Random Forest Classifier Results for Toronto.

The Toronto data set has 1037 scenarios under class ‘1’ for Injuries and 163 scenarios under class’0’ for Fatalities. There are still not enough example scenarios under Fatalities for this model to achieve high scores for precision and recall.

Appendix D: Toronto Red Light Camera Data Analysis

Figure 17 is a map describing the locations of red light cameras and the collisions that only happened at camera monitored intersections. There are 189 intersections in Toronto equipped with red light cameras from 2007 to 2021. Only 168 intersections were recorded with collisions from 2006 to 2020.

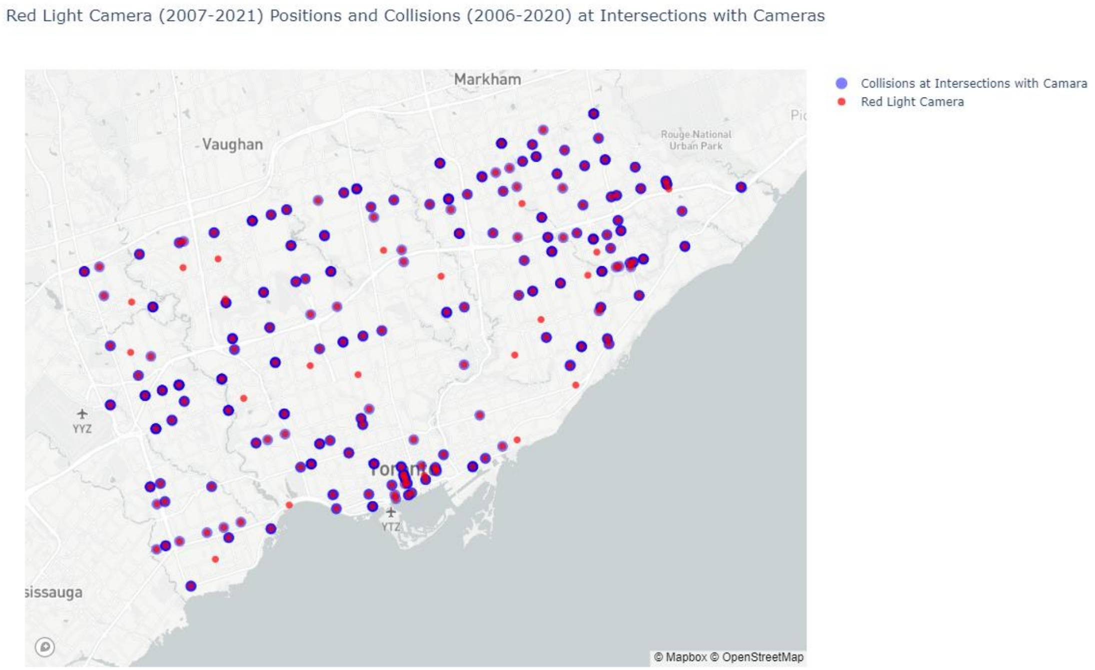

Red Light Camera (2007-2021) Positions and Collisions (2006-2020) at Intersections with Cameras
● Collisions at Intersections with Camara
● Red Light Camera
Markham
Vaughan
Pic
Rouge National Urban Park
YYZ
Toronto
YTZ
sissauga
© Mapbox © OpenStreetMap

Figure 17: Location of Red Light Cameras and Collisions at Camera Monitored Intersections.

As the cameras were installed at different times, it is reasonable to only compare the numbers of collisions that happened one full year before and after the activation of cameras. Then only 52 collisions at 37 intersections equipped with cameras were considered. Figure 18 shows that 29 collisions happened before the activation of cameras and 23 collisions occurred after the activation of cameras. Among 37 cameras installed at different intersections, 19 intersections only have collisions in one year before the activation of cameras, 6 intersections have collisions in one year both before and after the activation of cameras and 12 intersections only have collisions that happened in one year after the activation of cameras. It suggests that camera surveillance can help reduce the collisions at intersections, but it is far from impressive.

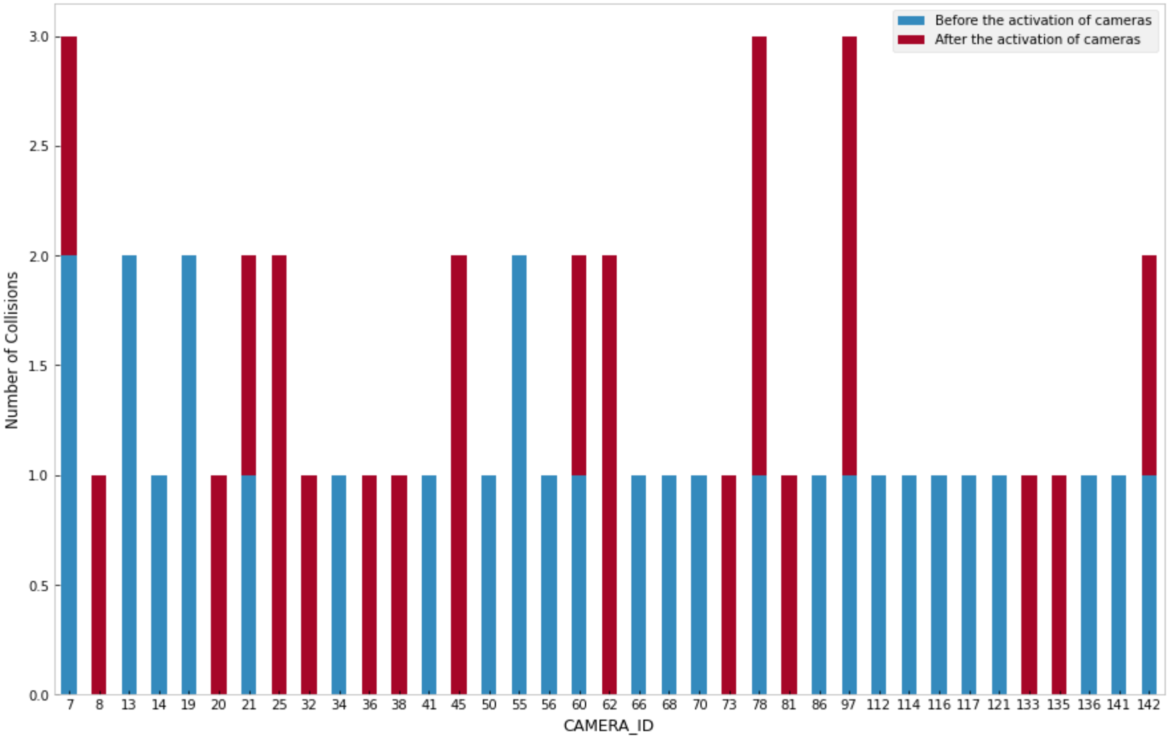


Figure 18: Collisions in One Year Before and After Cameras Activation.

The locations of redlight cameras were also examined. Figure 19 displays the camera coverage at intersections. It can be found in Fig. 19(a) that all the intersections with total collisions of greater than 6 from 2006 and 2020 have been monitored by cameras, indicating a good coverage of redlight cameras at most high-risk intersections. However, the camera coverage rates of lower than 50% at intersections that have total collisions of 3-6 are still not high enough. When looking at Fig. 19(b) of the intersections covered by cameras, more than 50% of cameras were installed at intersections with collisions of less than 3. Among them, about 10% of cameras were installed at intersections with no collisions. This suggests the locations of cameras should be further optimized. Fig. 19 (c) shows the collisions at intersections from 2006 to 2020 and the locations of redlight cameras. The size of blue circle is proportional to the collision numbers.

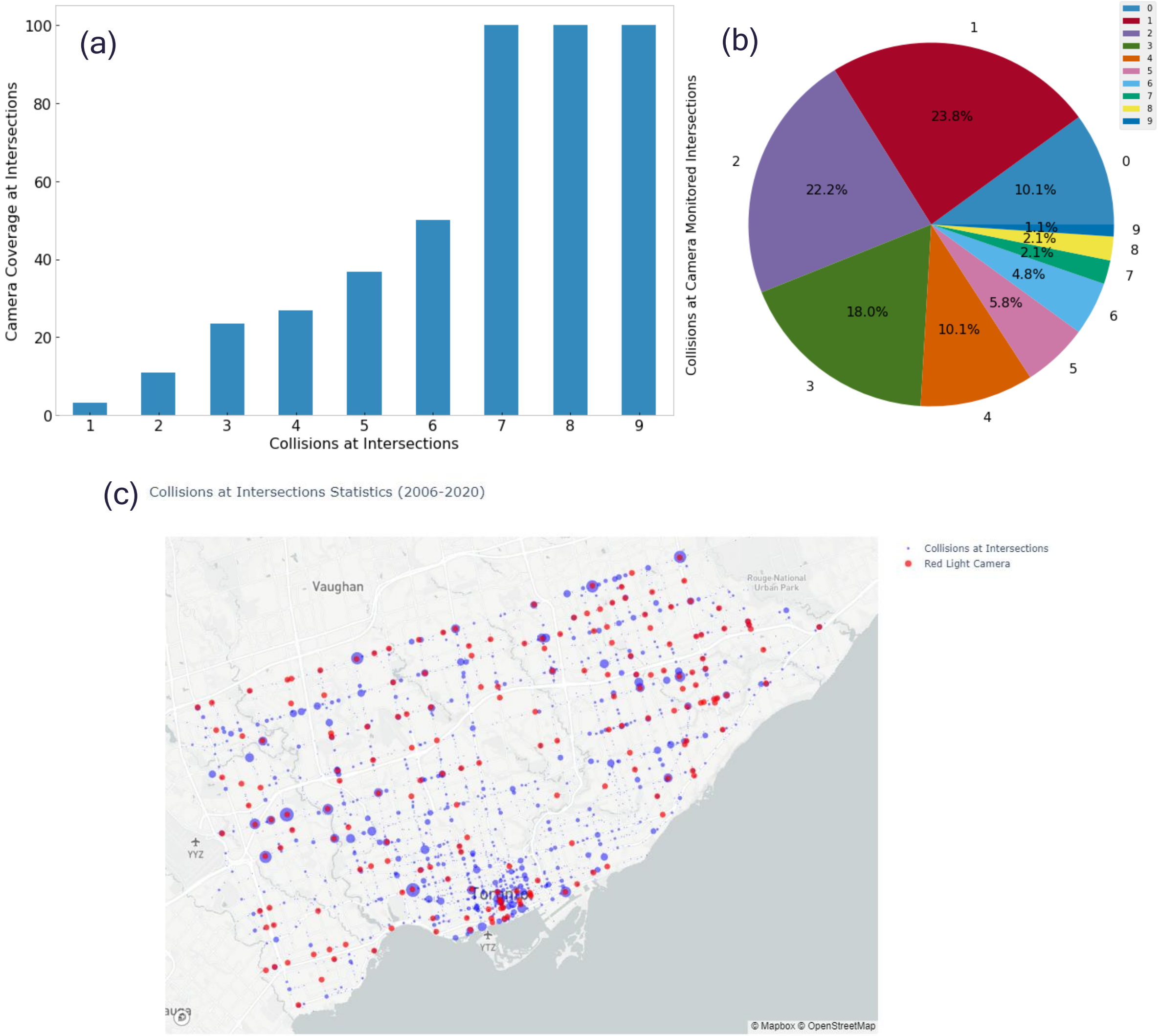


Figure 19: (a) Camera Coverage Rate at intersections with different collisions; (b) Collision Distribution at Camera Monitored Intersections; (c) Map of Collisions at Intersections.

Appendix E: Toronto Transit Data Analysis

Based on the cleaned up coordinators containing only 3 decimals for both latitude and longitude, the transit map is shown in Figure 20. 55.2% of the collision locations are aligned with one or more than one bus stops. With those 55.2% (or 3,311) collision locations, there are 409 (12.4%) fatal collisions and 2,902 (87.6%) non-fatal collisions. Moreover, there are 1,394 (42% of total collision locations which can be aligned with bus stops) collision locations with only 1 bus stop nearby, 1,232 with 2 bus stops nearby. Collision locations with 3-5 bus stops are accounted for 20% (or 673).

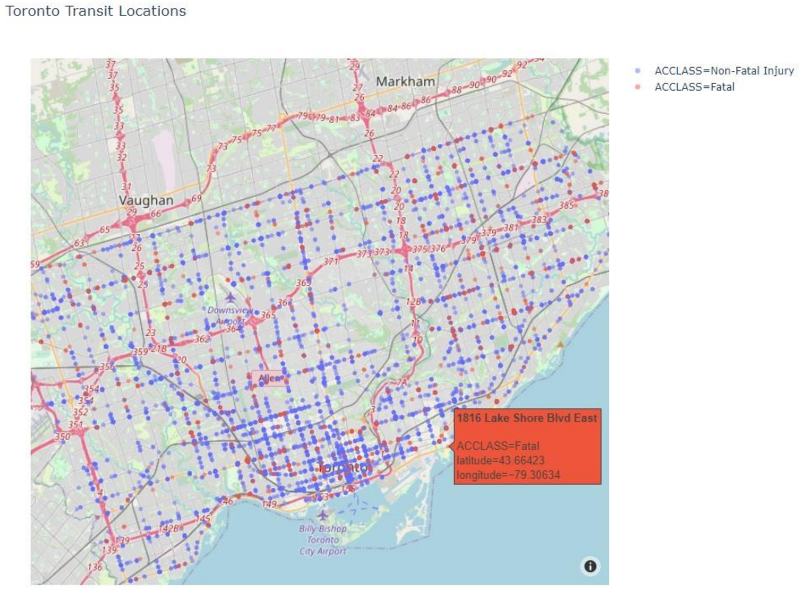

Toronto Transit Locations
● ACCLASS=Non-Fatal Injury
● ACCLASS=Fatal
Markham
Vaughan
Downsview
Toronto
Billy Bishop Toronto City Airport
1816 Lake Shore Blvd East
ACCLASS=Fatal
latitude=43.66423
longitude=−79.30634

Figure 20: Bus Stops Aligning with Past Fatal Collision Locations.

Appendix F: Toronto Fatal Victims

Age

Regarding fatal victims, there were only two people of unknown age. These records were removed from this evaluation.

The pie chart in Figure 21 presents how the fatal victims are distributed over age groups. The majority of fatal victims were between 20 to 24 years old, although the distribution of other age groups are quite similar.

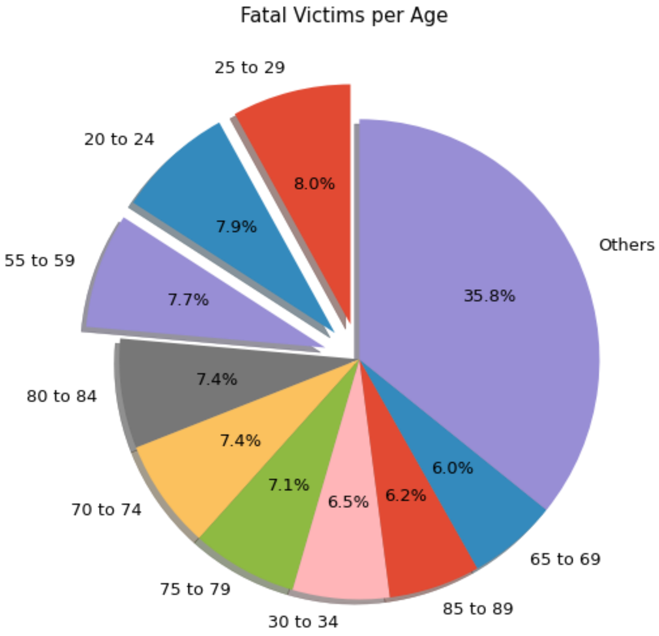


Figure 21: Distribution of Fatal Victims Among Age Groups.

The total number of fatal victims among each age group is displayed in Figure 22. Here, it is also possible to see the percentage of fatal victims in each age group. For example, for those involved in collisions and between 90 and 94 years old, the chances to become a fatal victim is 37.5%. On the other hand, this probability is lower for younger people. For instance, the chances of those between 20 and 24 years old (the most representative group in absolute number analysis) reduces to 4.10%. This is expected, as the elder are more fragile and susceptible to greater injuries.

Special attention must be paid for elementary age victims' results (between zero and 14 years old), where the total number of fatal victims is really lower than other groups. This finding could mean that school safety zones are working, turning these places safer.

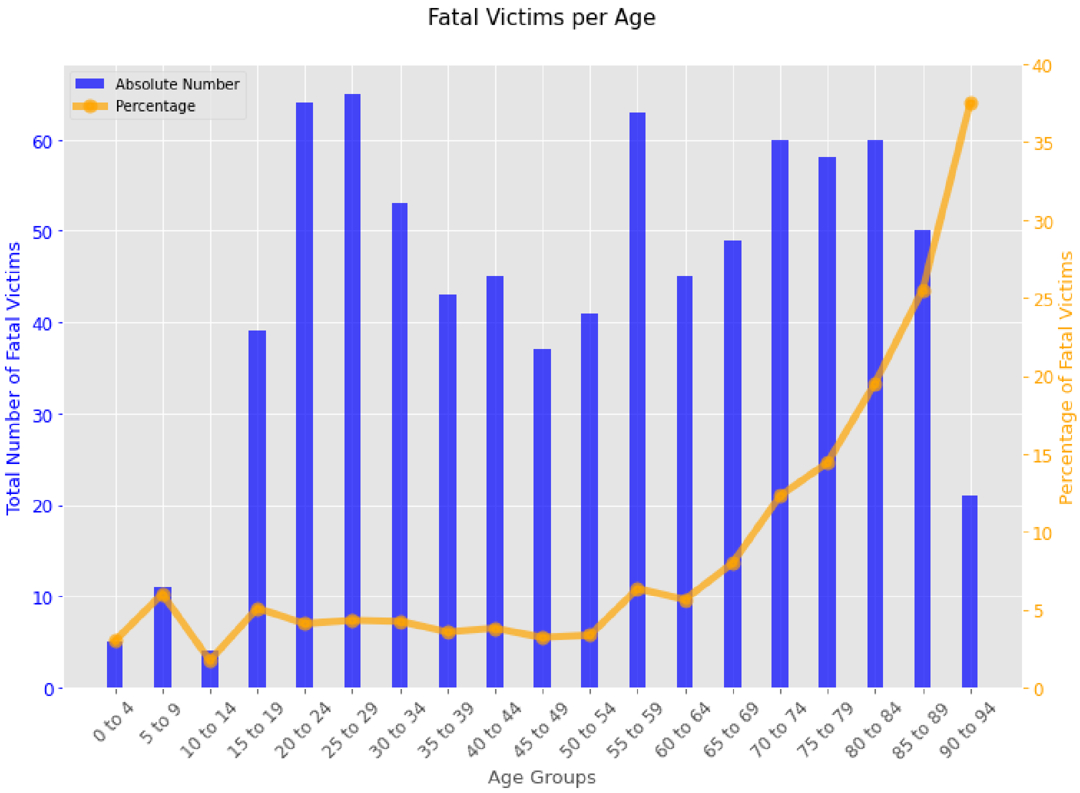


Figure 22: Percentage of Fatal Victims Among all People Involved per Age Group.

Person Involvement

There is no fatal victim in the original dataset without an appropriate involvement recorded. The majority of fatal victims were pedestrians, representing 56.3% of all occurences, followed by drivers with 18.3% This shows how vulnerable a pedestrian is when affected by a vehicle collision.

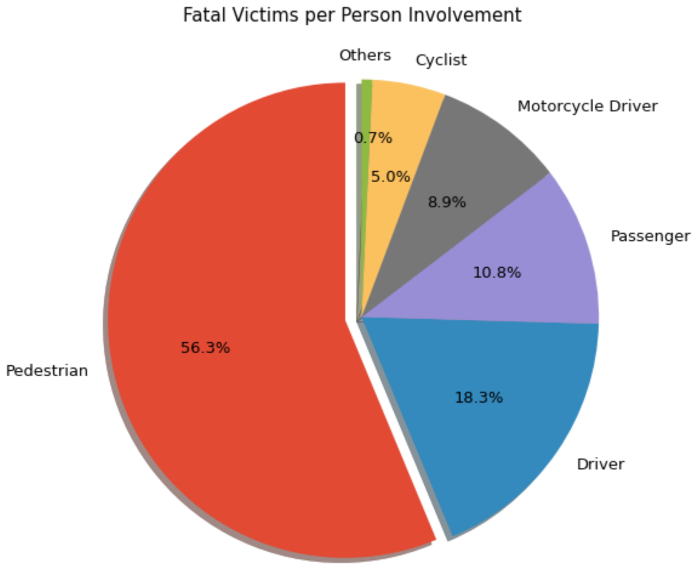


Figure 23: Distribution of Fatal Victims Among People Involvement.

The above statement is confirmed when we look at Figure 24, which displays that 16.09% of pedestrians involved in collisions died.

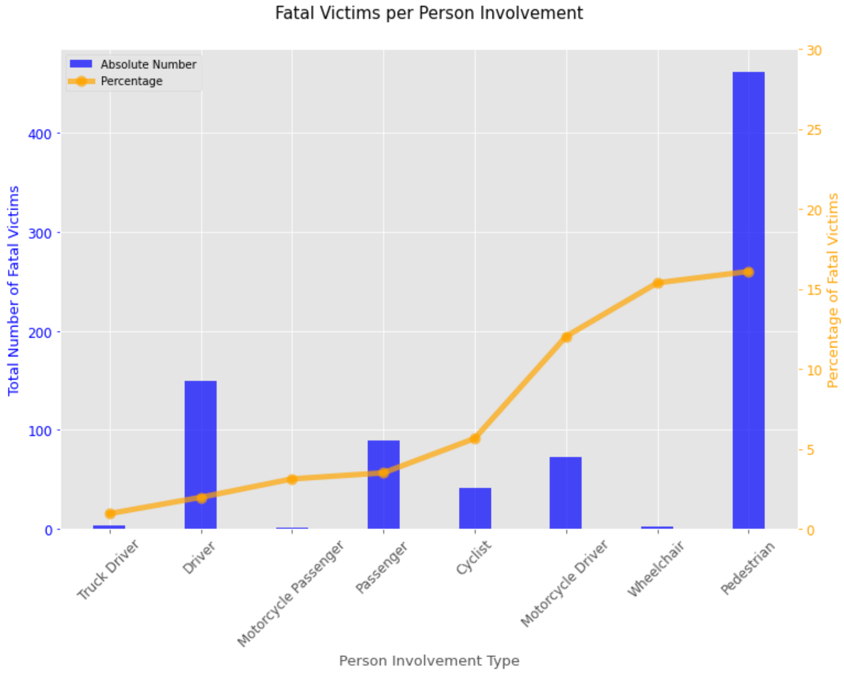


Figure 24: Percentage of Fatal Victims Among all People Involved per Person Involvement.

Vehicle Type

At this section we will evaluate the vehicle type recorded at the original data set, which means that no cleaning was carried out for this information. Looking at this attribute, we can see that 40.2% of fatal victims have an unknown vehicle type recorded (330 of 821 records), followed by the category ‘Other’ with 27.0%. This high number of not detailed information makes the task of evaluating the absolute occurrences less precise.

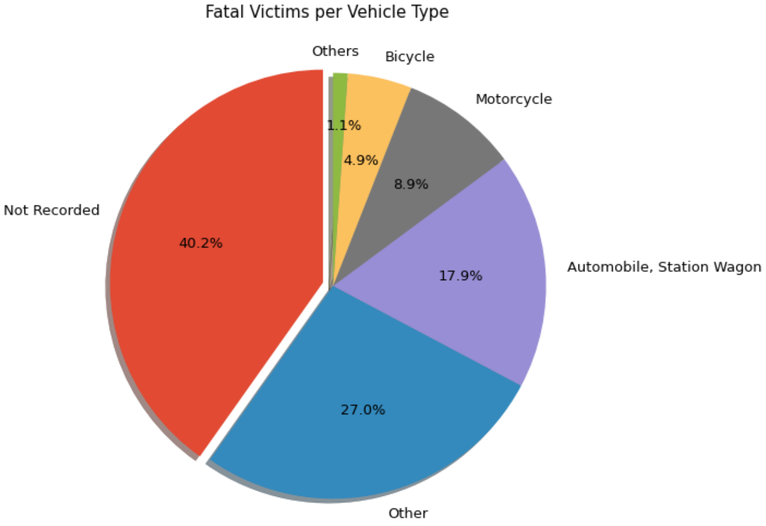


Figure 25: Distribution of Fatal Victims Among Vehicle Types.

However, Figure 26 displays how dangerous a vehicle collision is for those in motorcycles: fatal victims are 12.01% of people involved, being the highest value. On the other hand, again the data is not precise: there is also a high percentage of fatal victims in vehicles not categorized, being 11.73% not recorded and 4.68% for other types of vehicles.

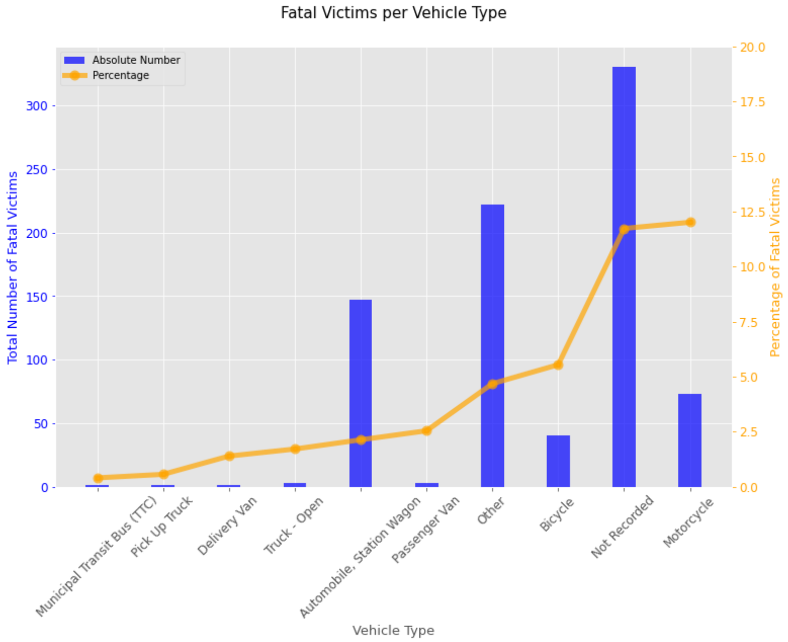


Figure 26: Percentage of Fatal Victims Among all People Involved per Vehicle Type.

Vehicle Category

Vehicle categories are derived from vehicle types, and again the obtained results are not precise: 40.2% are considered ‘Null’. The group ‘Other’, which represents some vehicle types defined earlier, is also significant, with 27.6%. Following that, automobiles have the third highest value, with 18.4% of the occurrences.

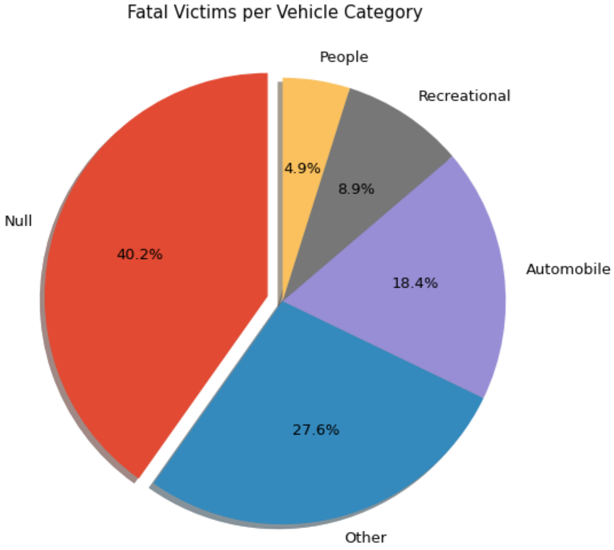


Figure 27: Distribution of Fatal Victims Among Vehicle Categories.

This imprecise data is reflected in Figure 28, with a high number of fatal victims in ‘Null’ vehicles. Here, we must pay attention to the result for ‘Recreational’ vehicles (which includes motorcycles): there are fatal victims in 11.57% of them. This matches to the result shown in Vehicle Type analysis.

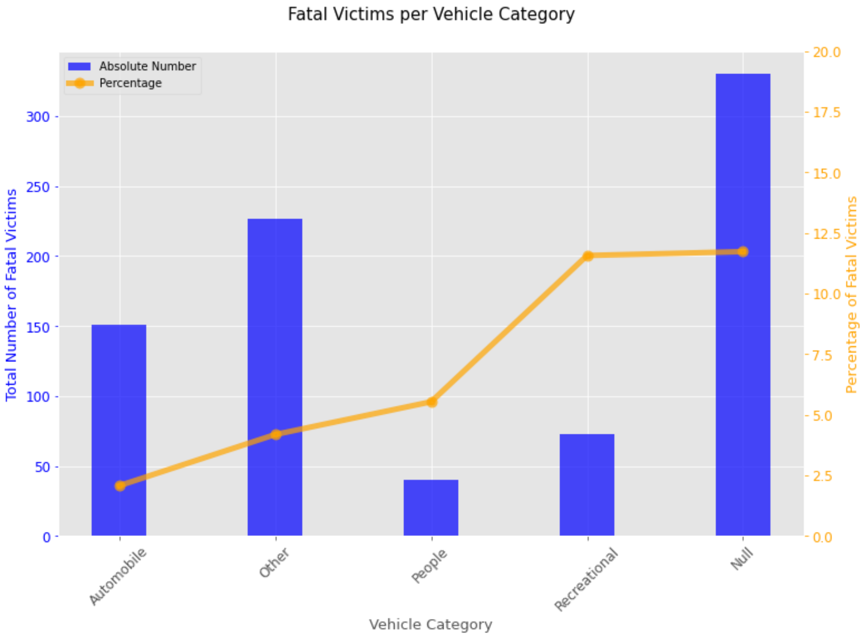


Figure 28: Percentage of Fatal Victims Among all People Involved per Vehicle Category.

Hour of Day

Regarding hours of day, the data are quite scattered. The time interval with the greater number of fatal collisions is between 18:00 and 18:59, with 7.3%. This is a time with greater traffic, and then it is expected that the number of collisions increases.

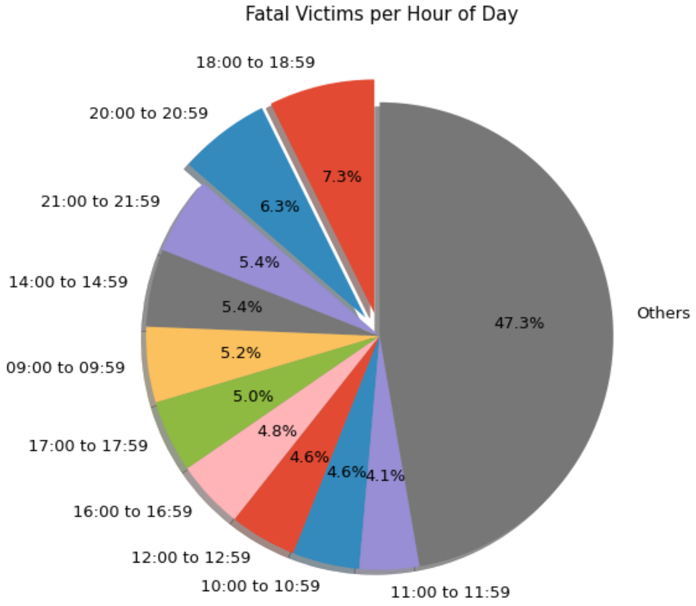


Figure 29: Distribution of Fatal Victims Among Hours of Day.

Although, looking at the relative analysis in Figure 30, it is interesting to note that 9.44% of collisions between 05:00 and 05:59 culminated in fatal victims, being the worst result found and significantly higher than the next worst result, which is 6.64% of fatal victims between 06:00 and 06:59. For the rest of the day the results oscillate around 3.3 to 6.6%.

To explain this behavior, two possibilities were investigated:
(1) At this time there are already more vehicles running on the streets than earlier, but without traffic, letting them run faster. On the other hand, there are also many early jogging people, and a collision in higher speeds would easily culminate in a fatality;
(2) It could have drivers under alcohol influence driving after night parties, especially at weekends.

It was observed that the majority of 27 fatalities on this time interval occurred on Fridays (7 cases), Mondays (6 cases) and Tuesdays (6 cases), allowing us to discard option 2 for weekends. Besides, 10 of these victims were pedestrians, strengthening thesis 1. Although there were also 10 drivers in the victim groups, any of them were recorded as under alcohol influence. So, for these 10 drivers victims, we can conclude that the possibility to run faster and have a collision increases the chance of a fatality.

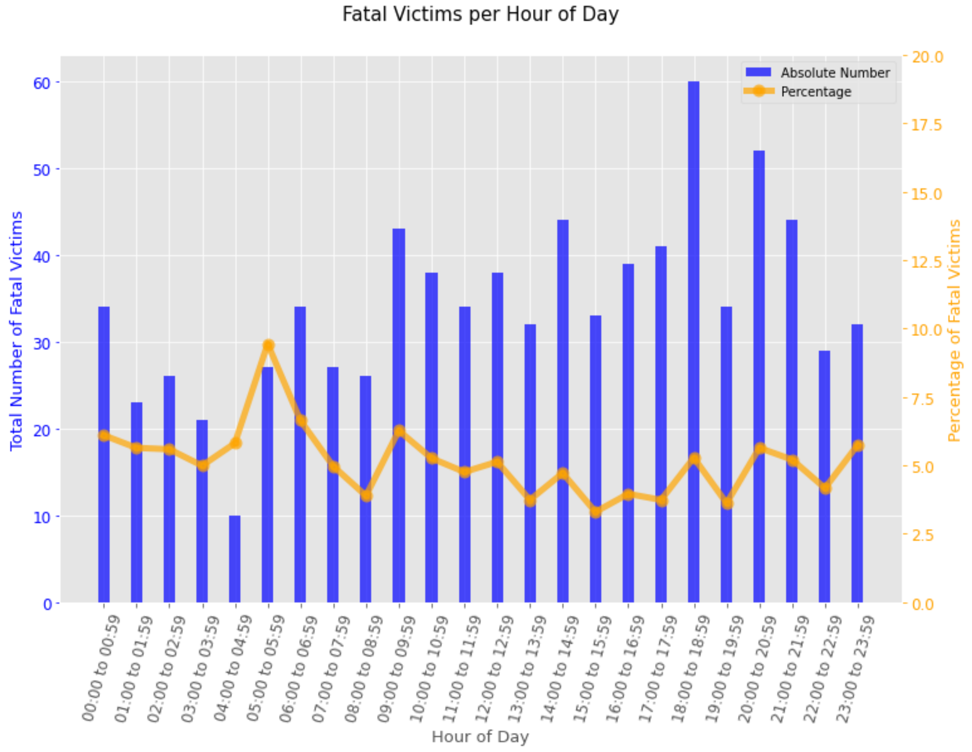


Figure 30: Percentage of Fatal Victims Among all People Involved per Hour of Day.

Day of Week

In absolute numbers, the worst day is Friday, with 17.4%, followed by Tuesdays with 16.8%. The traffic is usually higher on Fridays because many people decide to travel out of Toronto in the end of the workday. For the rest of weekdays (Tuesday, Wednesday and Thursdays) the results are quite similar and a bit better than Fridays.

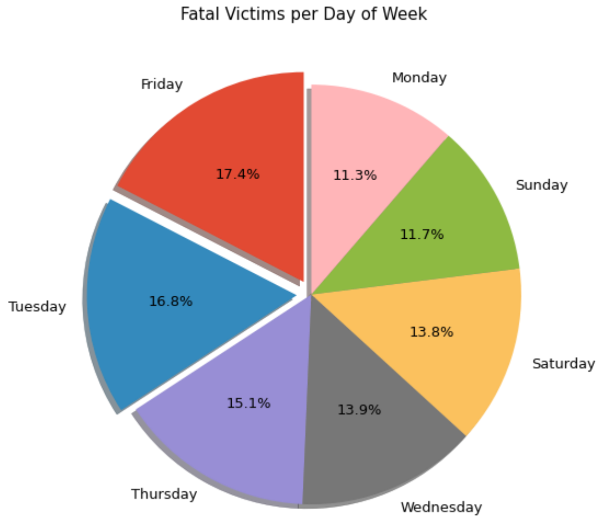


Figure 31: Distribution of Fatal Victims Among Day of Week.

Figure 32 presents that the percentage of collisions causing fatal victims over all days is quite similar. Interestingly, looking at both plots we can conclude that Mondays are the safer day to drive.

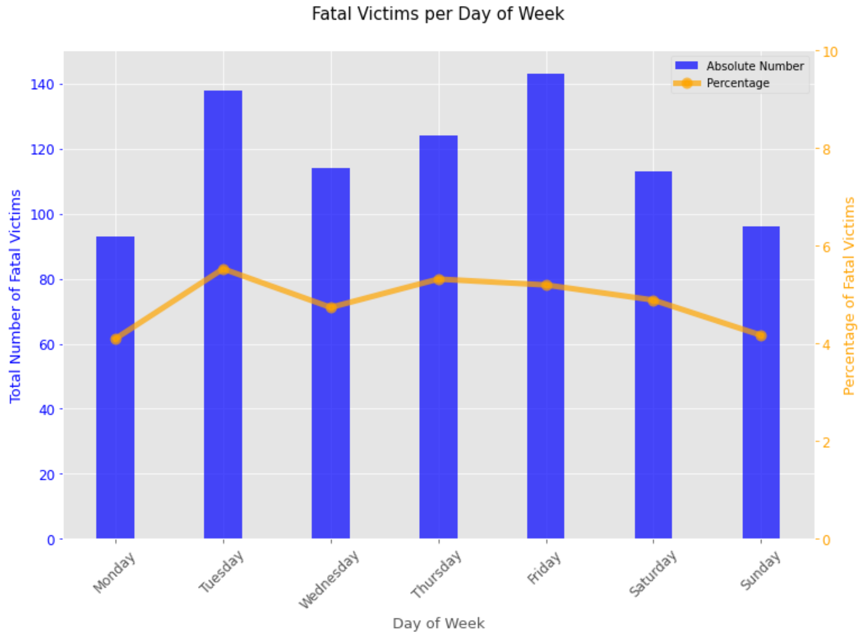


Figure 32: Percentage of Fatal Victims Among all People Involved per Day of Week.

Month

The results are quite similar among months. September and August concentrate the highest number of fatal victims occurences, with 10.8% and 9.7% respectively. We conclude that in Summer there are more collisions (this will be confirmed at Season Analysis). Really close to these results we see January, and this could be related to the weather conditions in this month.

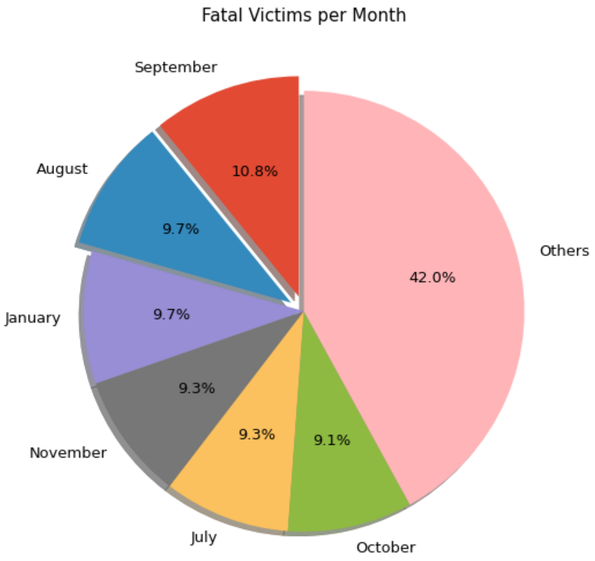


Figure 33: Distribution of Fatal Victims Among Months.

Figure 34 presents that January is the most dangerous month: 6.64% of people involved in collisions died. This can also be related to the weather conditions, because there is a trend to reduce this percentage during Spring. However, during the Summer the results become worse again, increasing gradually until the Winter.

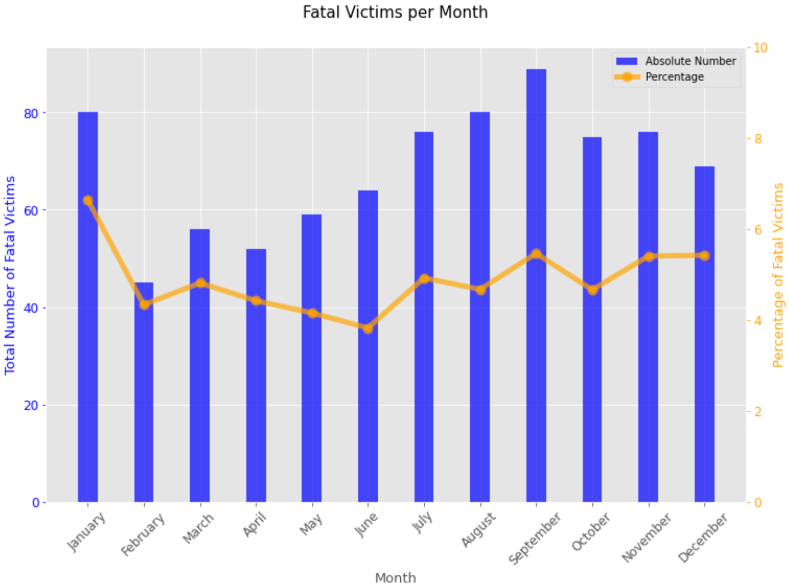


Figure 34: Percentage of Fatal Victims Among all People Involved per Month.

Season

Figure 35 confirms what was shown before: in absolute numbers, the worst season is Summer.

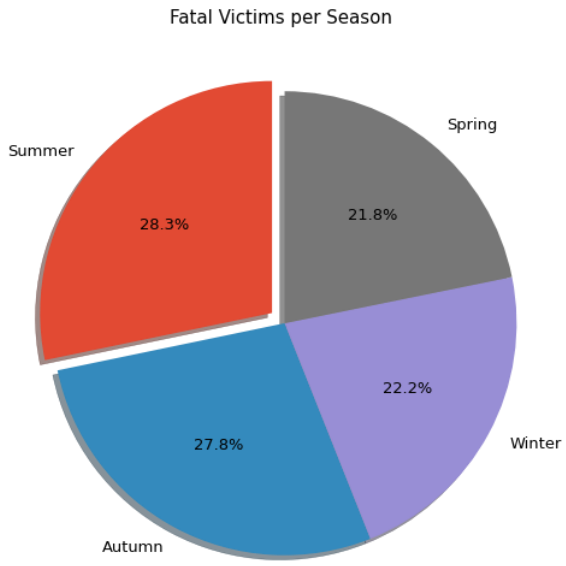


Figure 35: Distribution of Fatal Victims Among Seasons.

However, looking at the percentage of fatal victims in Figure 36, Winter is worse, with 5.25%, followed by Autumn with 5.22%.

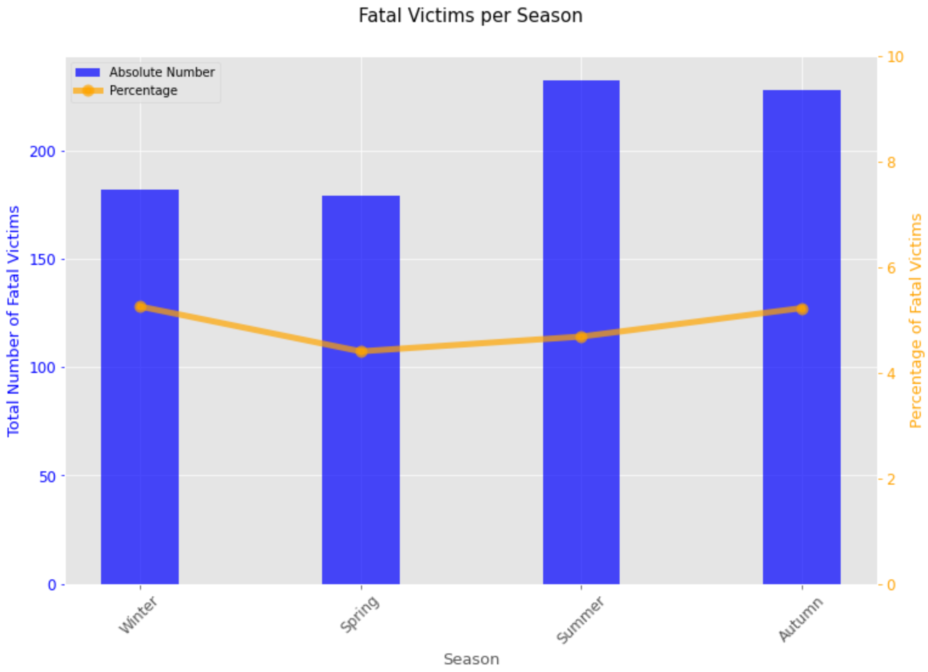


Figure 36: Percentage of Fatal Victims Among all People Involved per Season.

Year

From 2006 to 2020, as presented at this dataset, the worst year in number of collisions were 2016, concentrating 9.5% of occurrences, followed by 2018 with 8.0%.

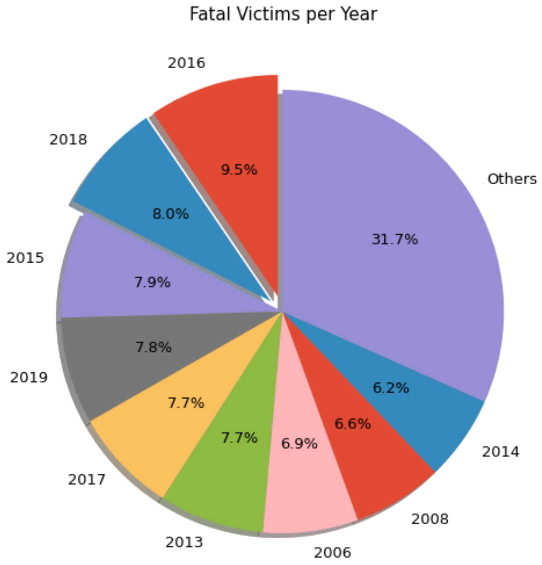


Figure 37: Distribution of Fatal Victims Among Years.

2016 was also the most dangerous year, when 7.75% of people involved in collisions were fatal victims. We can clearly see an upward trend between 2011 and 2016 in Figure 38, and since then there has been a reduction in the percentage of fatalities. Here, it is clear the impact of COVID-19 on 2020 results, where the total number of fatalities was the lowest since 2011, however the percentage of fatalities was still high, similar to 2017-2018 results.

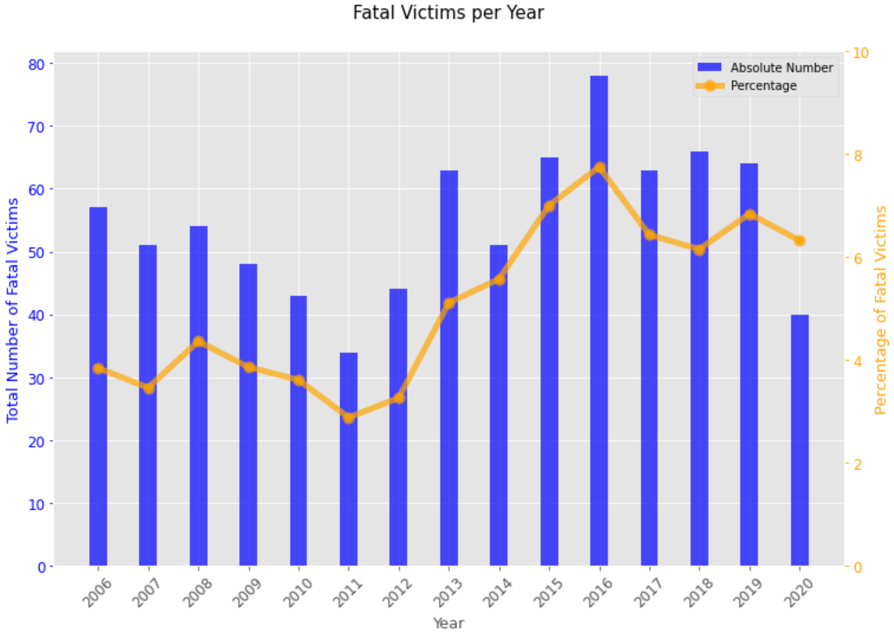


Figure 38: Percentage of Fatal Victims Among all People Involved per Year.

Appendix G: Toronto - Additional Plots

Pedestrians

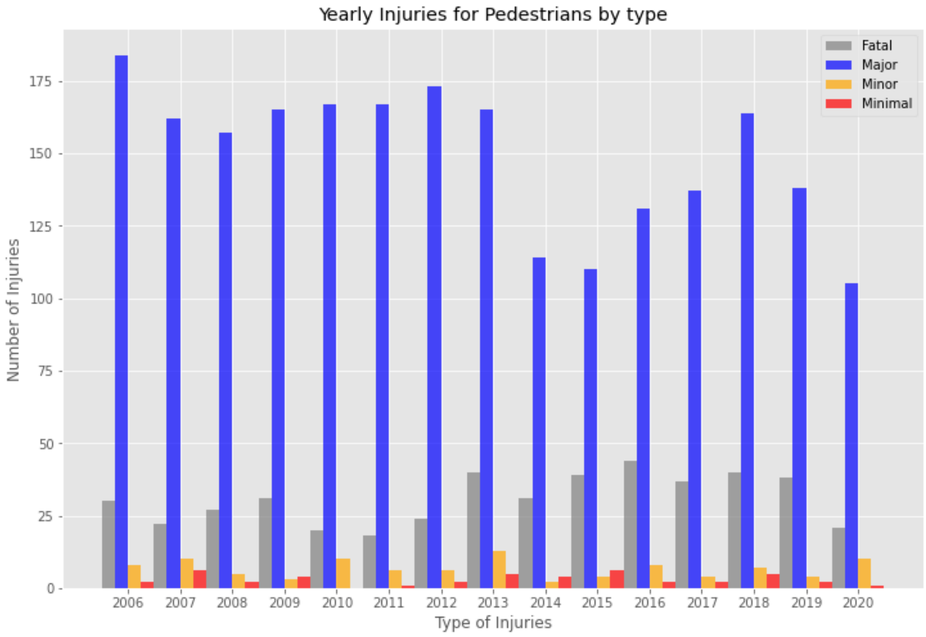


Figure 39: Yearly Injuries for Pedestrians by Type.

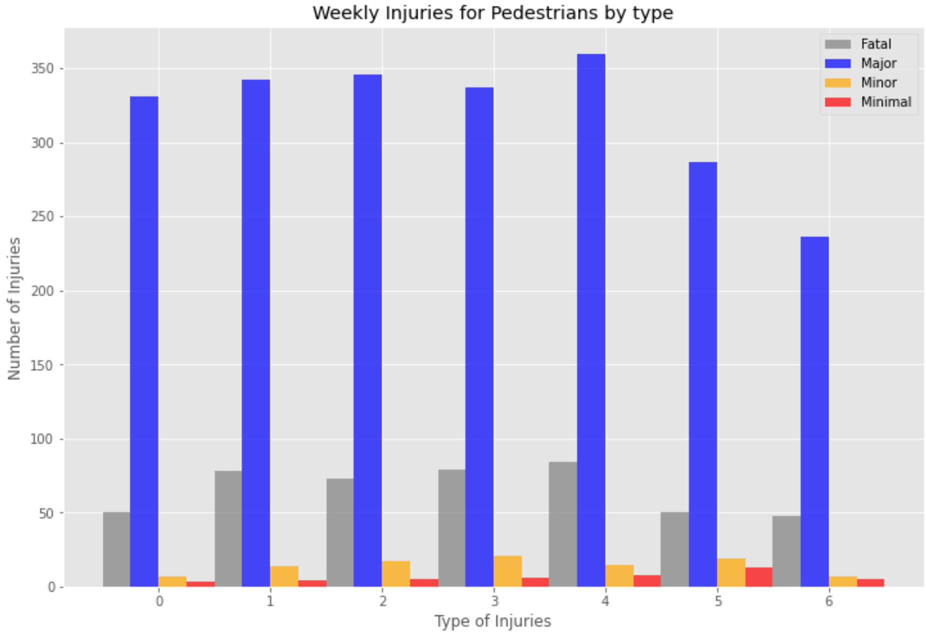


Figure 40: Weekly Injuries for Pedestrians by Type.

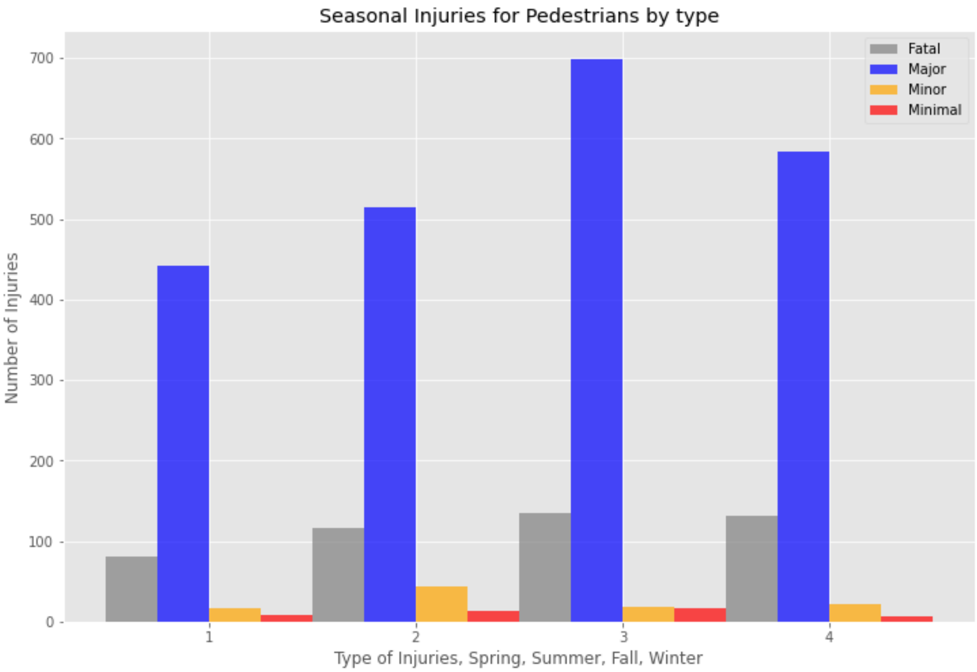


Figure 41: Seasonal Injuries for Pedestrians by Type.

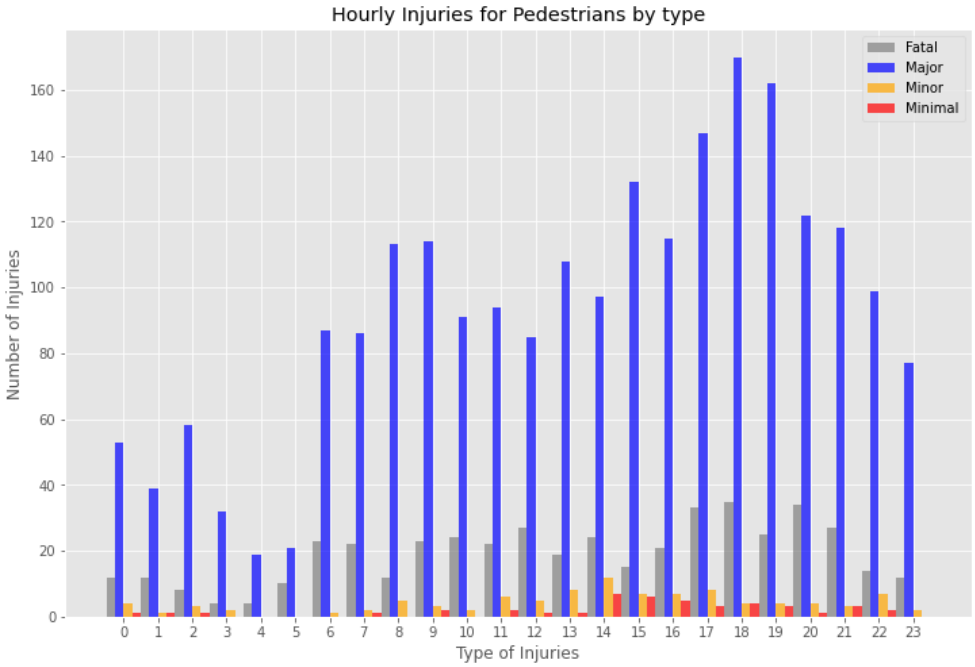


Figure 42: Hourly Injuries for Pedestrians by Type.

Personal Vehicles

Personal vehicles and injuries by Year, weekday, season, and hour. Personal vehicles include the vehicle types: (Automobile, Passenger Van, Truck - Closed (Blazer, etc),Pick up Truck, Taxi)

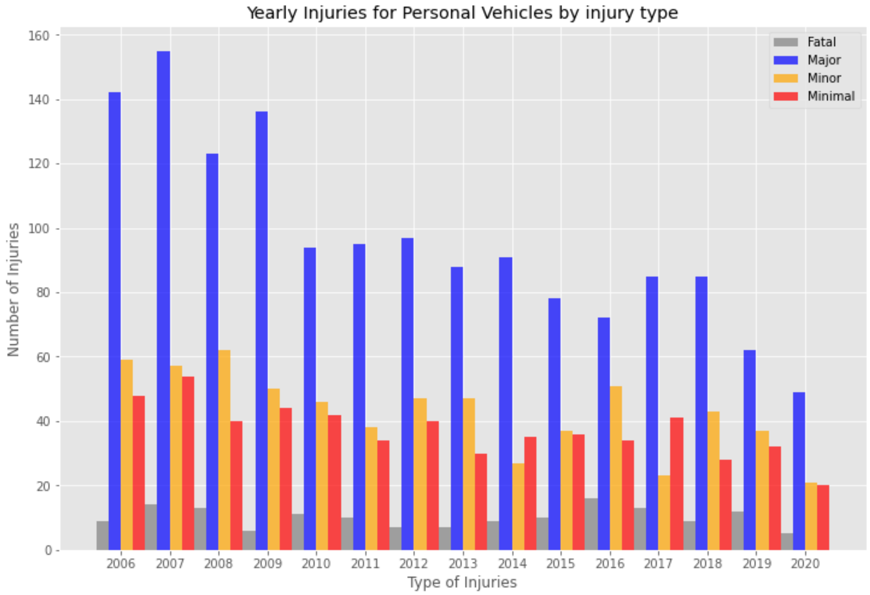


Figure 43: Yearly Injuries for Personal Vehicles by Injury Type.

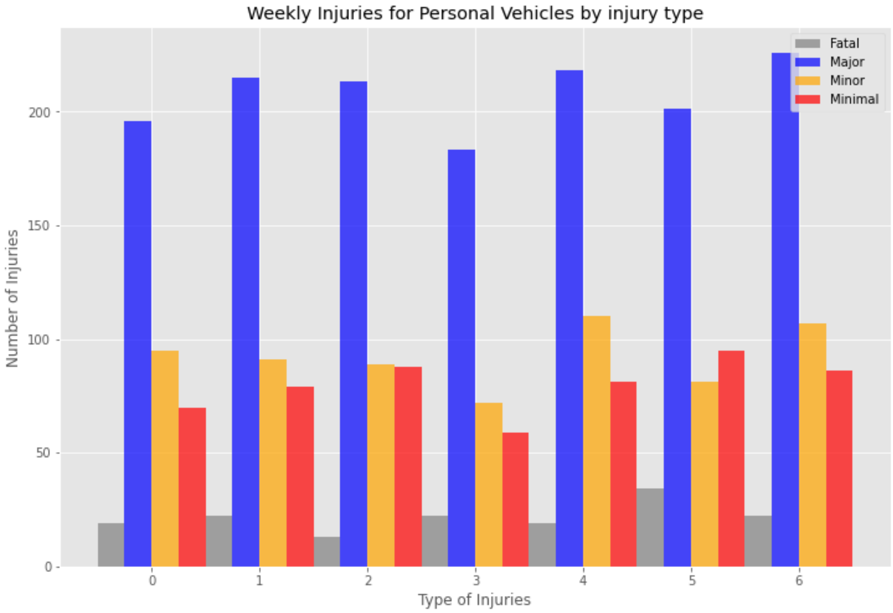


Figure 44: Weekly Injuries for Personal Vehicles by Injury Type.

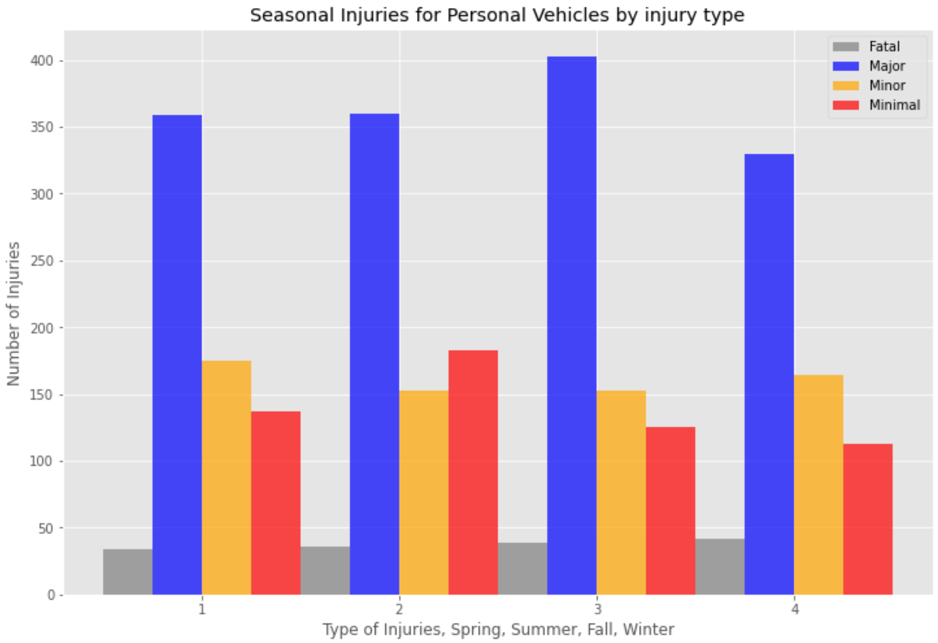


Figure 45: Seasonal Injuries for Personal Vehicles by Injury Type.

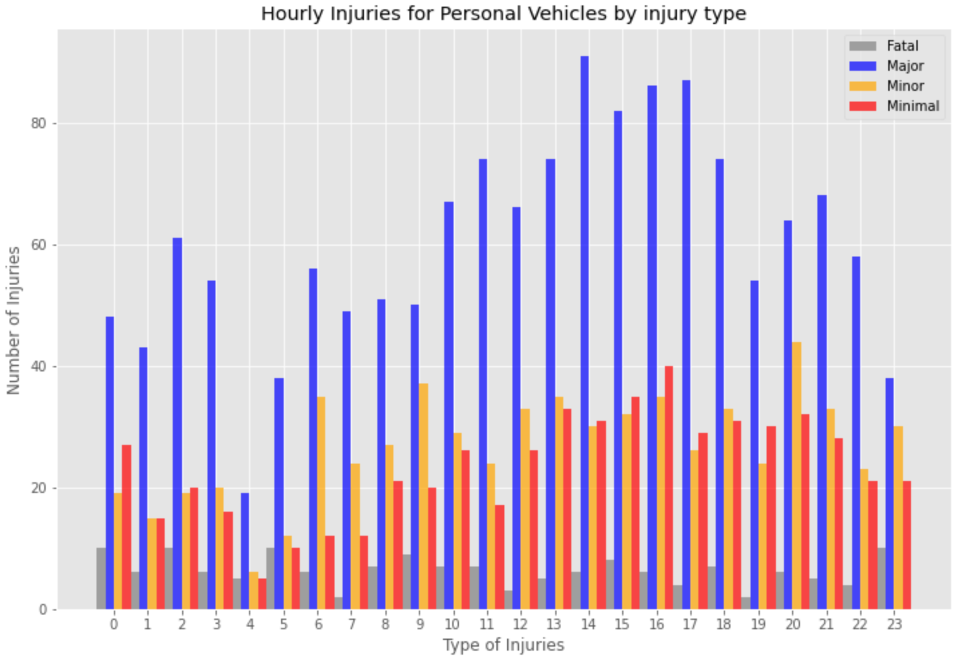


Figure 46: Hourly Seasonal Injuries for Personal Vehicles by Injury Type.

Motorcycles

Motorcycle vehicles injuries by year, weekday, season, and hourly. Motorcycle includes the following vehicle types: Motorcycle, Moped, Off Road - 2 Wheels

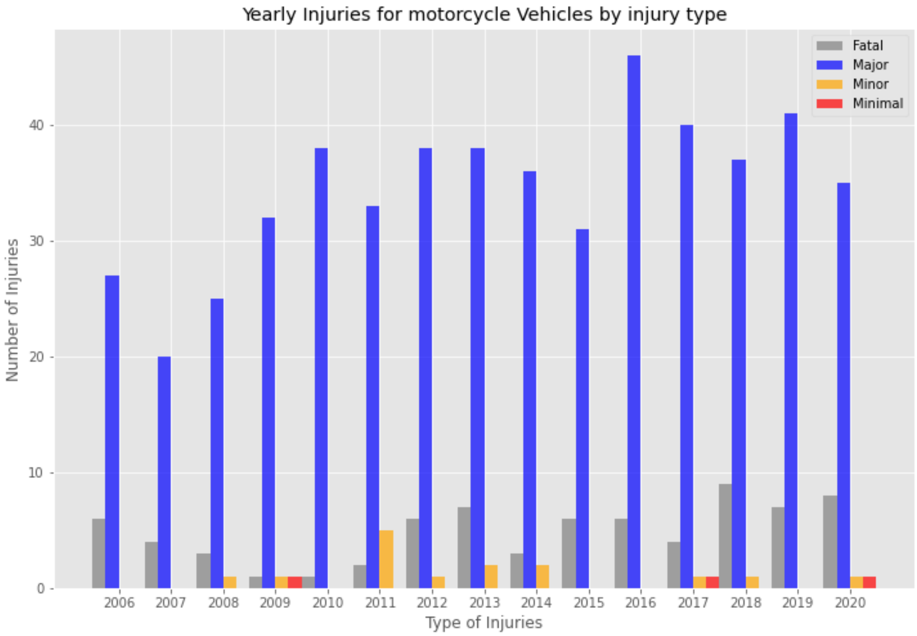


Figure 47: Yearly Injuries for Motorcycle Vehicles by Injury Type.

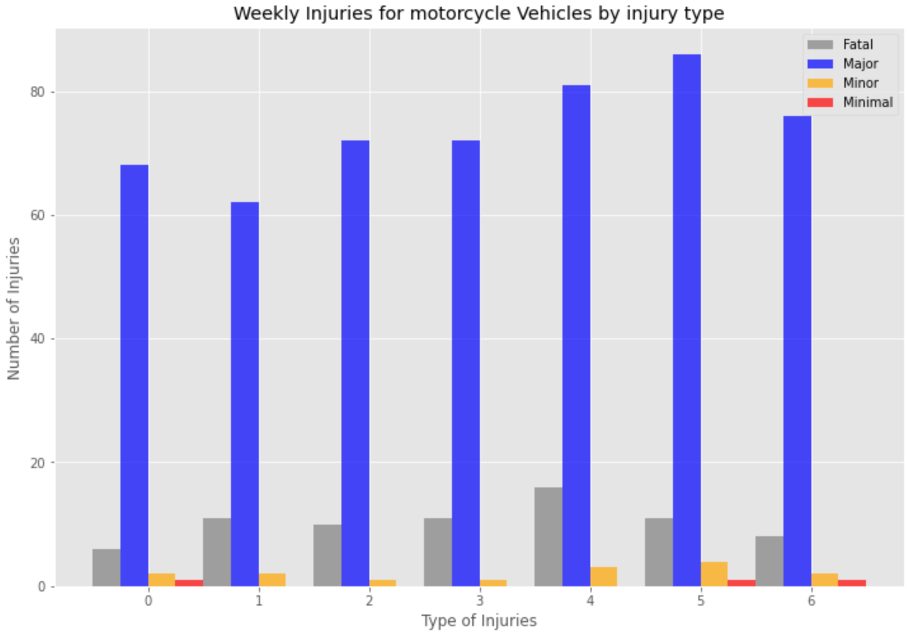


Figure 48: Weekly Injuries for Motorcycle Vehicles by Injury Type.

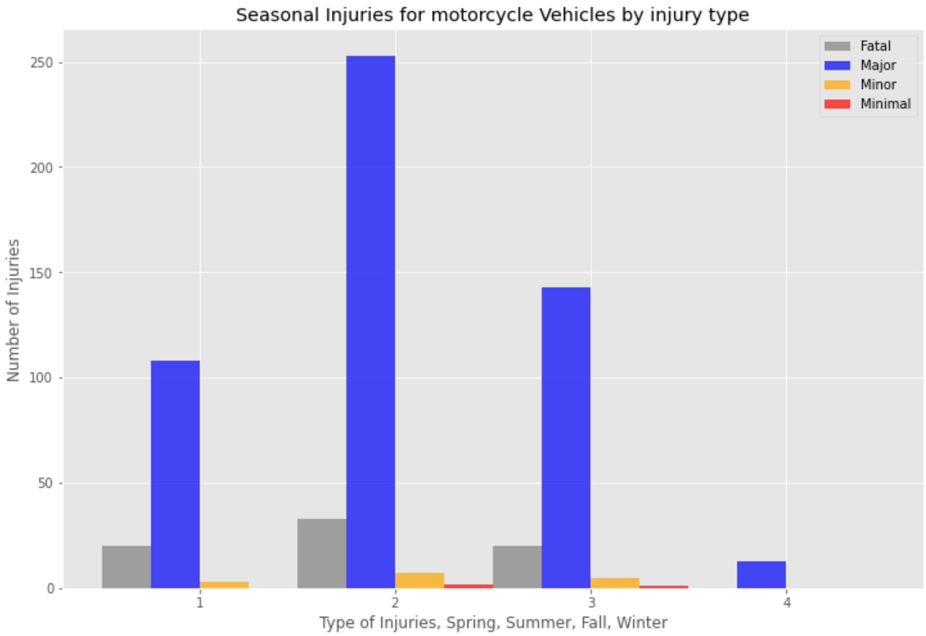


Figure 49: Seasonal Injuries for Motorcycle Vehicles by Injury Type.

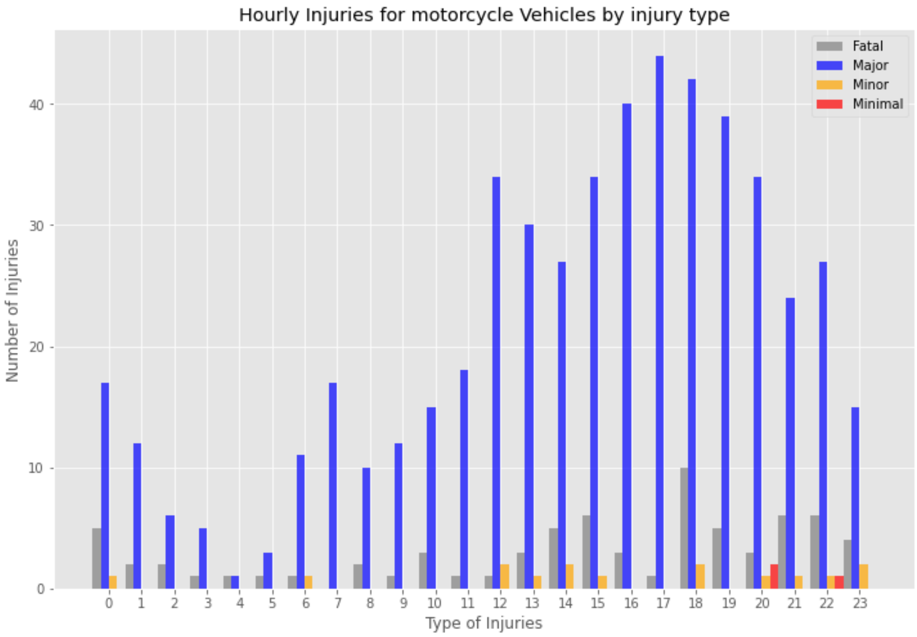


Figure 50: Hourly Injuries for Motorcycle Vehicles by Injury Type.

Cyclists

Bicycle injuries by Year, weekday, season, and hourly.

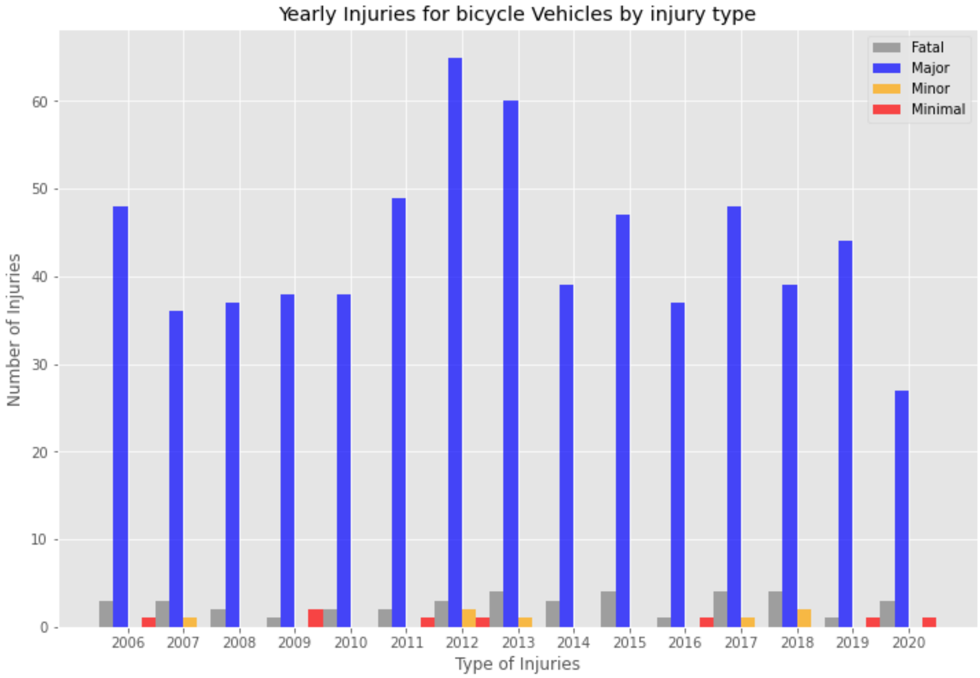


Figure 51: Yearly Injuries for Bicycle Vehicles by Injury Type.

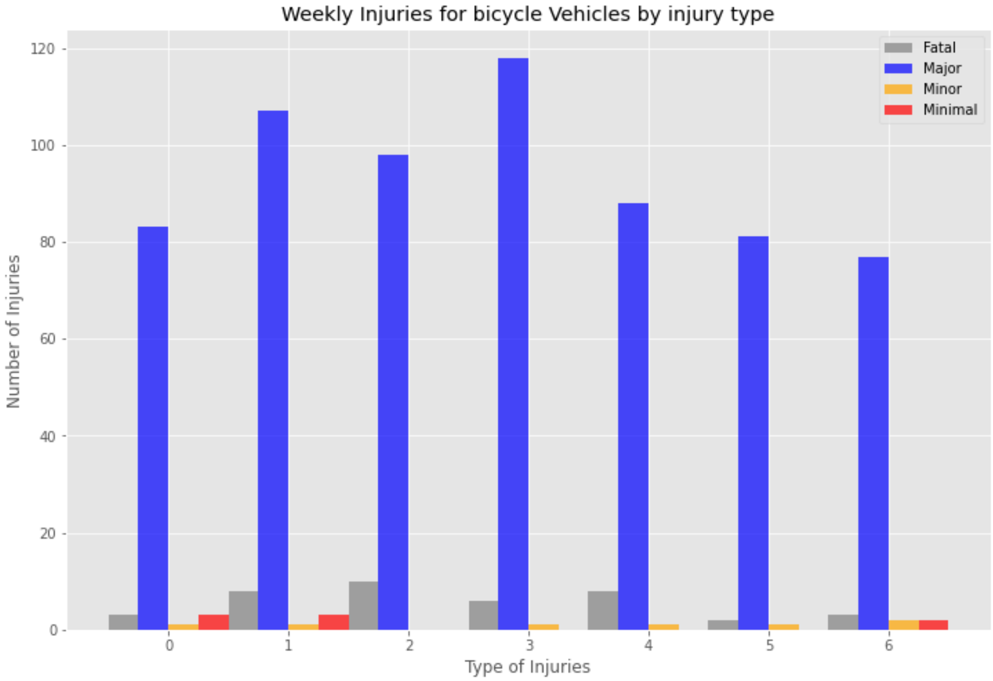


Figure 52: Weekly Injuries for Bicycle Vehicles by Injury Type.

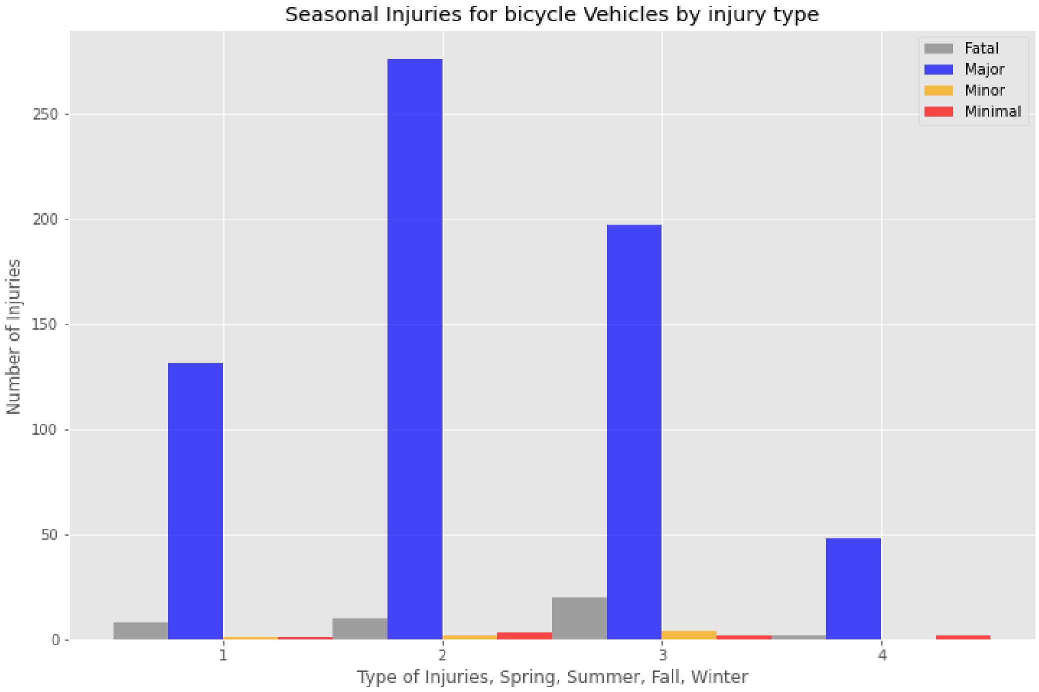


Figure 53: Seasonal Injuries for Bicycle Vehicles by Injury Type.

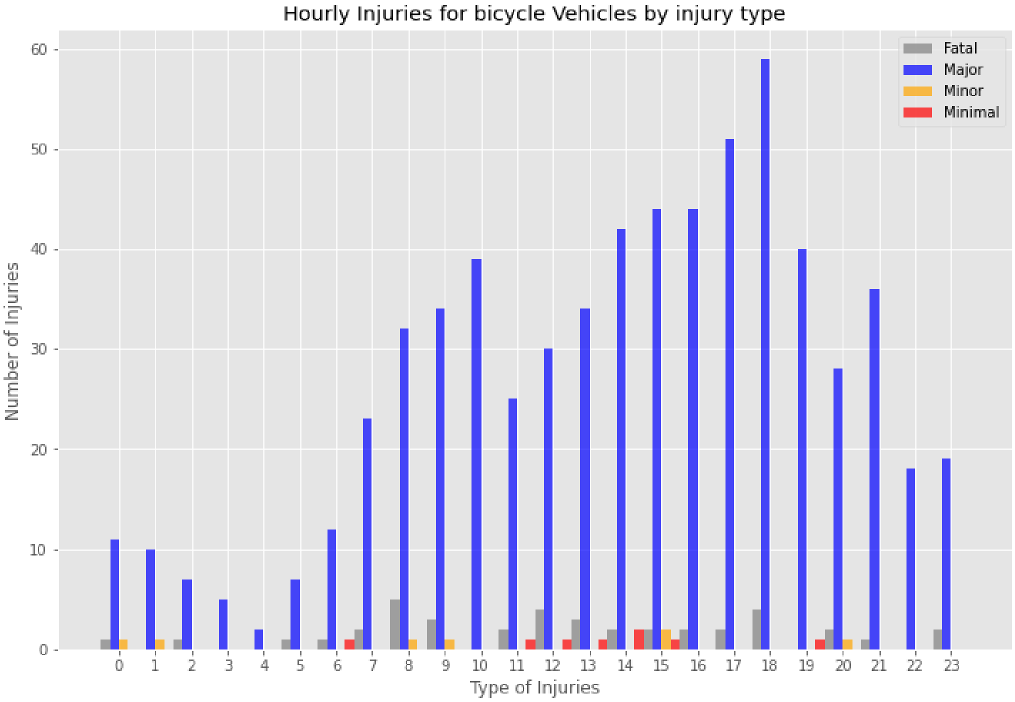


Figure 54: Hourly Injuries for Bicycle Vehicles by Injury Type.

Public Transit

Public Transit injuries by year, weekday, and season. Public Transit includes: Municipal Transit Bus (TTC), Street Car, Bus (Other) (Go Bus, Gray Coach),Intercity Bus, School Bus

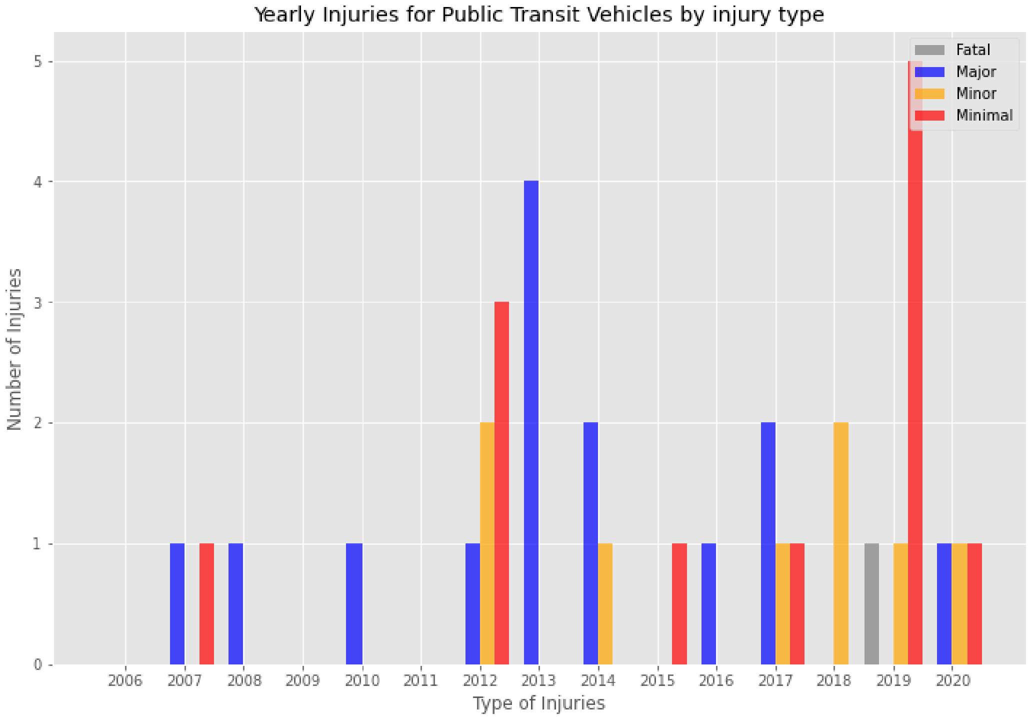


Figure 55:Yearly Injuries for Public Transit Vehicles by Injury Type.

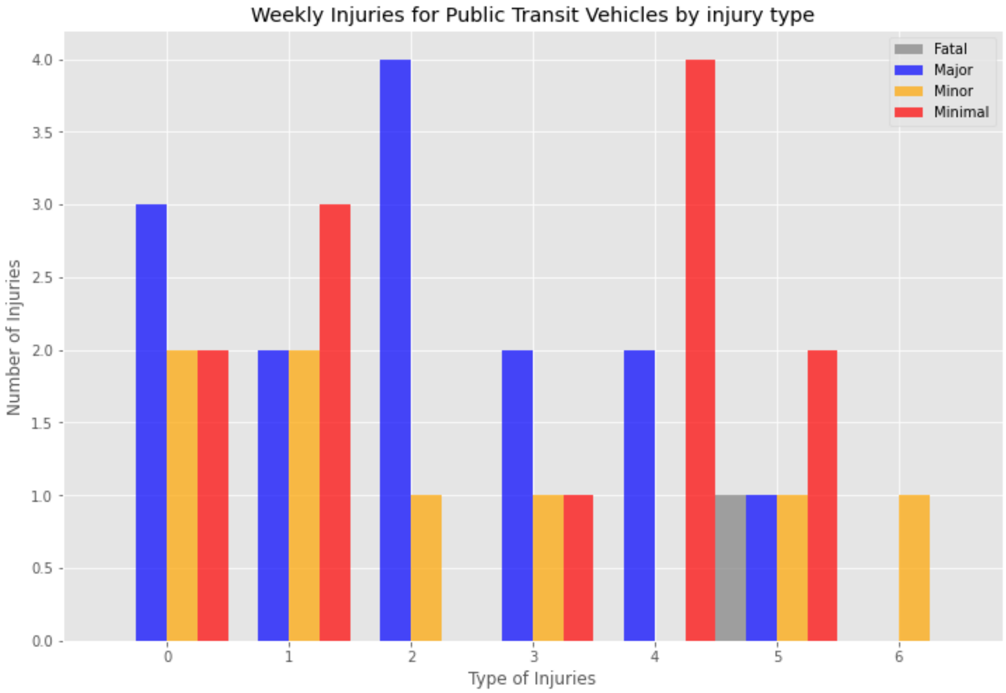


Figure 56: Weekly Injuries for Public Transit Vehicles by Injury Type.

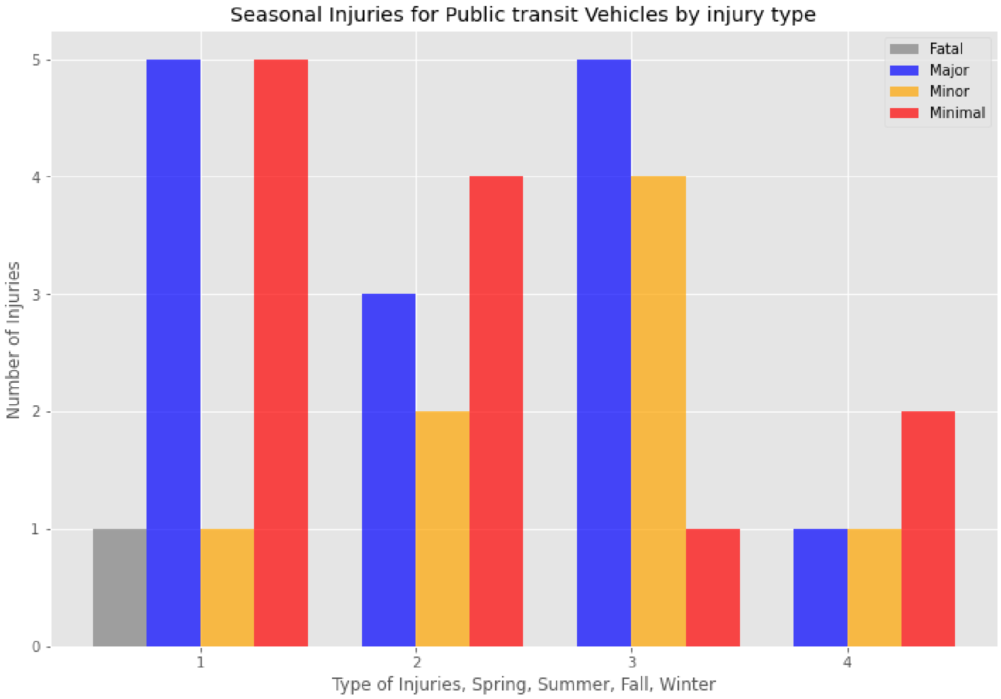


Figure 57: Seasonal Injuries for Public Transit Vehicles by Injury Type.

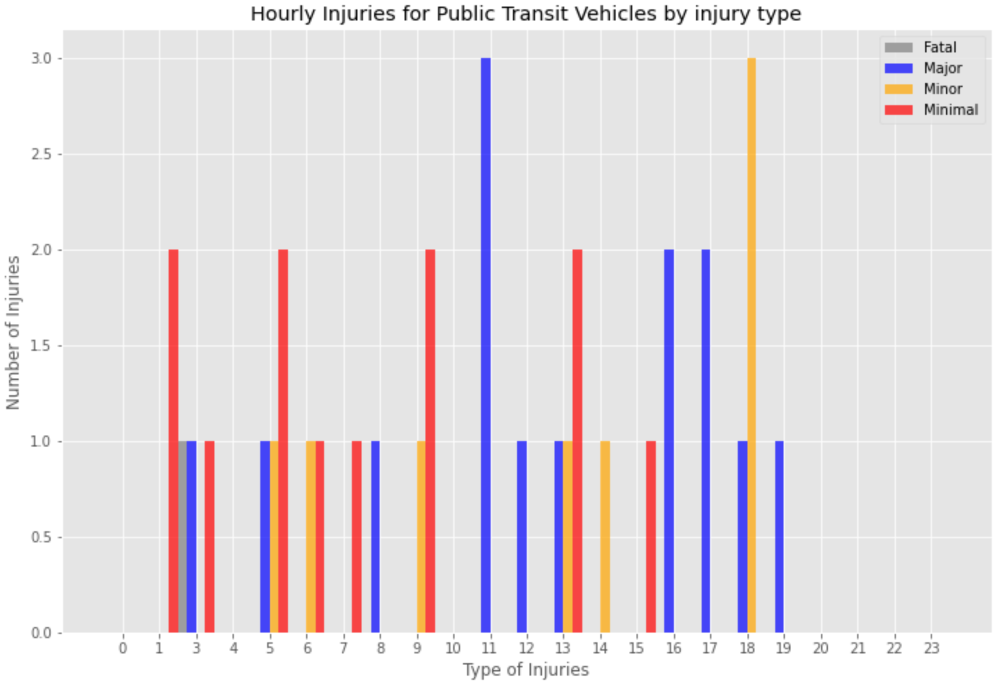


Figure 58: Hourly Injuries for Public Transit Vehicles by Injury Type.

Emergency Vehicles

Emergency Vehicles by year, weekday, season, and hourly. Emergency Vehicles includes: (Police Vehicle, Other Emergency Vehicle, Fire Vehicle)

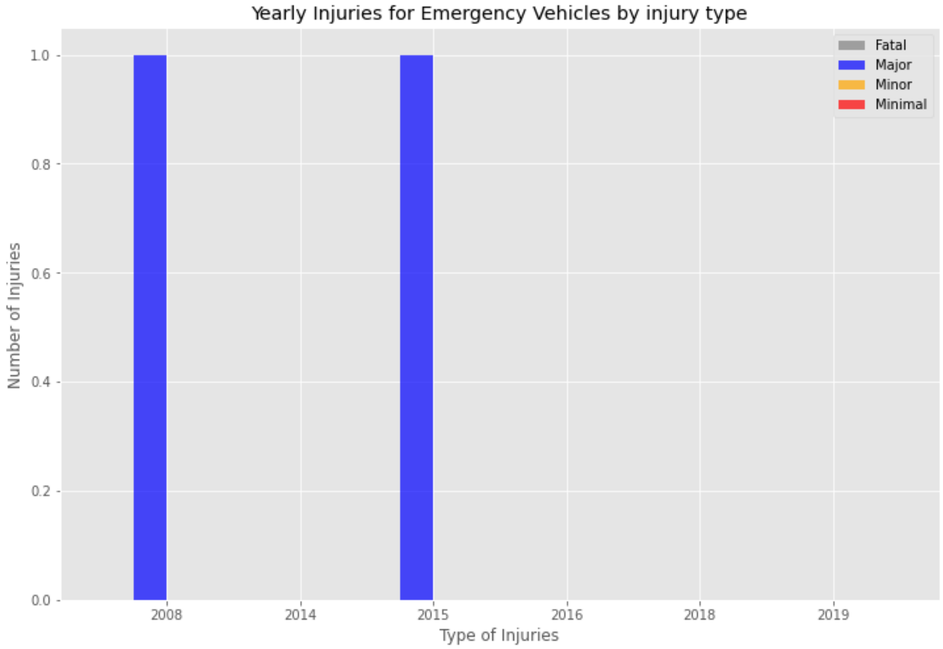


Figure 59: Yearly Injuries for Emergency Vehicles by Injury Type.

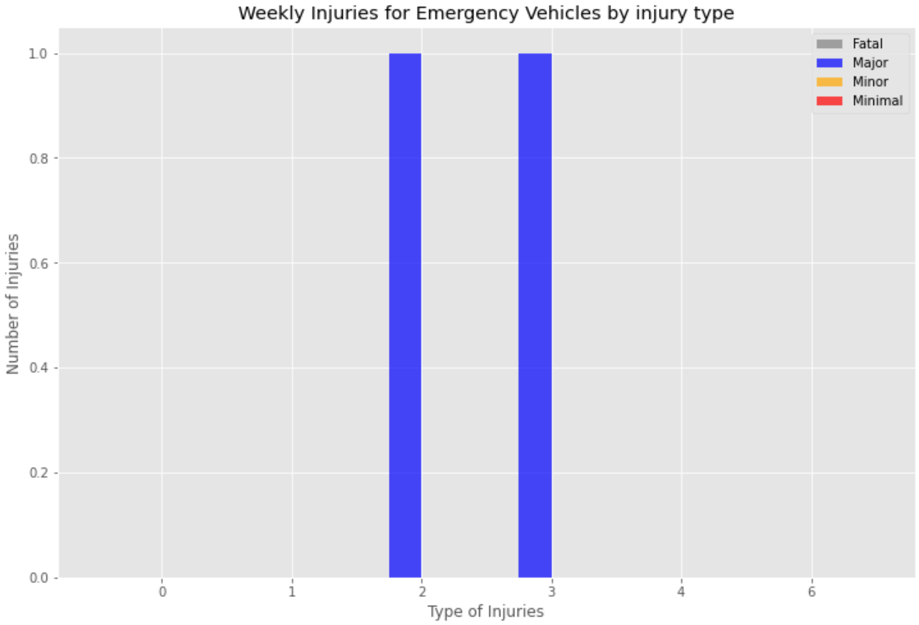


Figure 60: Weekly Injuries for Emergency Vehicles by Injury Type.

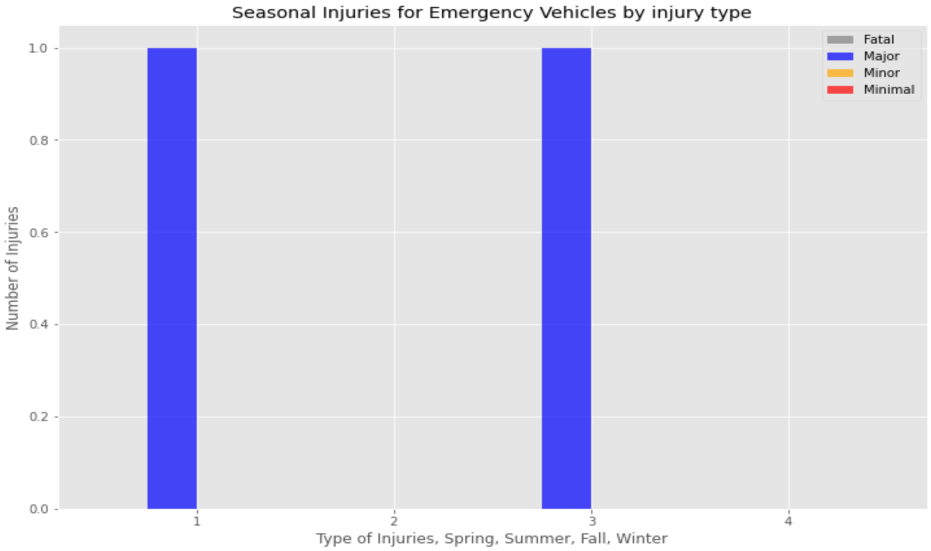


Figure 61: Seasonal Injuries for Emergency Vehicles by Injury Type.

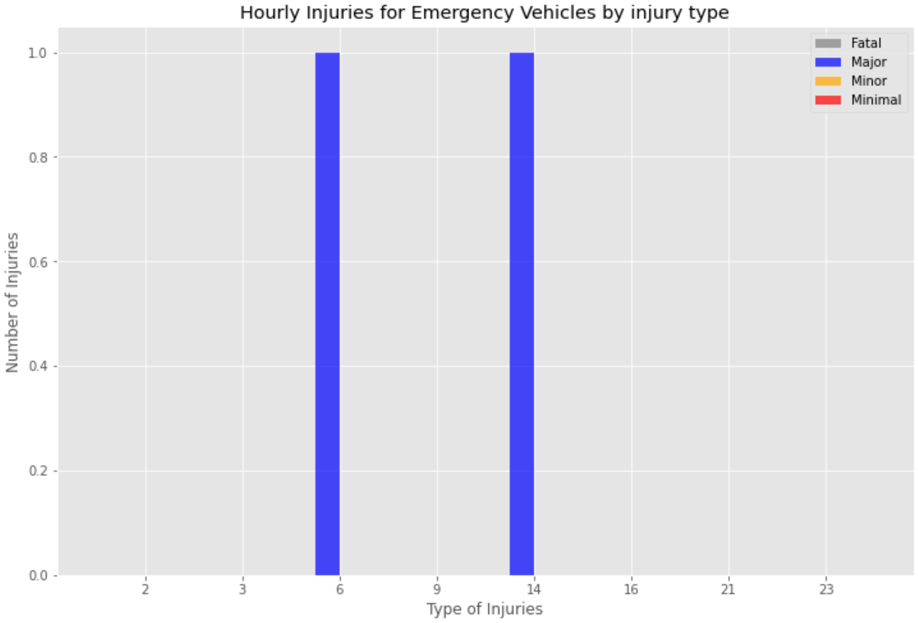


Figure 62: Hourly Injuries for Emergency Vehicles by Injury Type.

Large Trucks

Large Trucks injuries by year, weekday, season, and hourly. Large Trucks includes: (Truck - Open, Truck-Tractor, Truck - Dump, Construction, Delivery Van, Equipment, Truck (other), Truck - Tank, Tow Truck, Truck - Car Carrier)

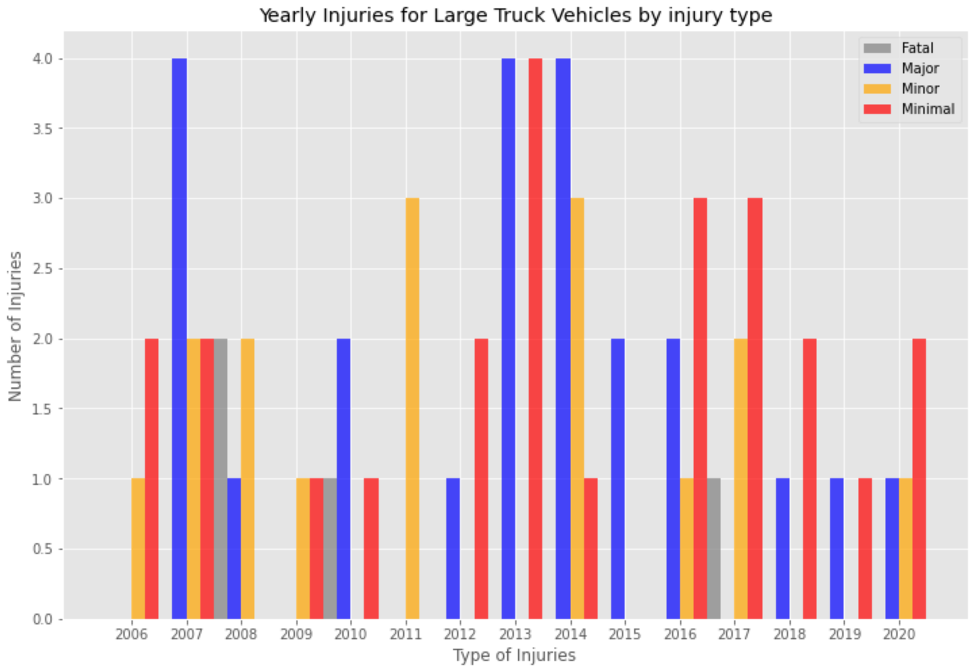


Figure 63: Yearly Injuries for Large Truck Vehicles by Injury Type.

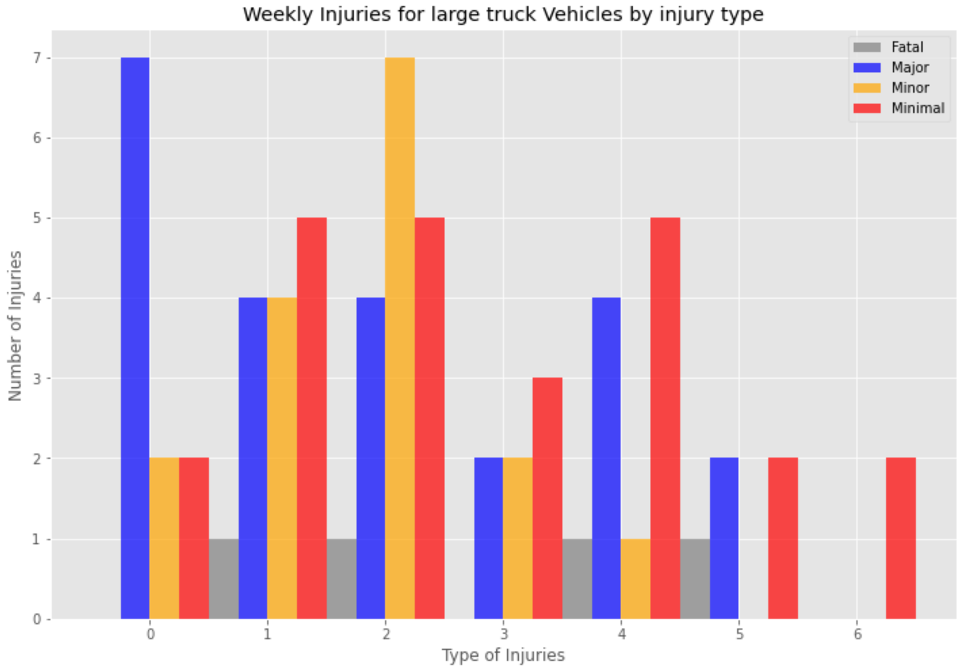


Figure 64: Weekly Injuries for Large Truck Vehicles by Injury Type.

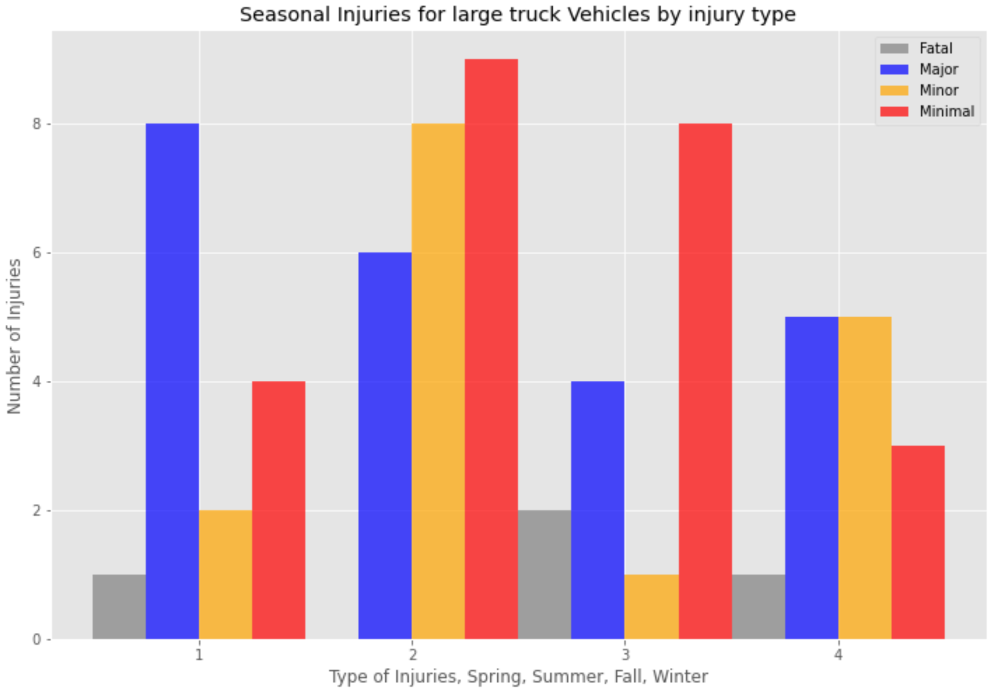


Figure 65:Seasonal Injuries for Large Truck Vehicles by Injury Type.

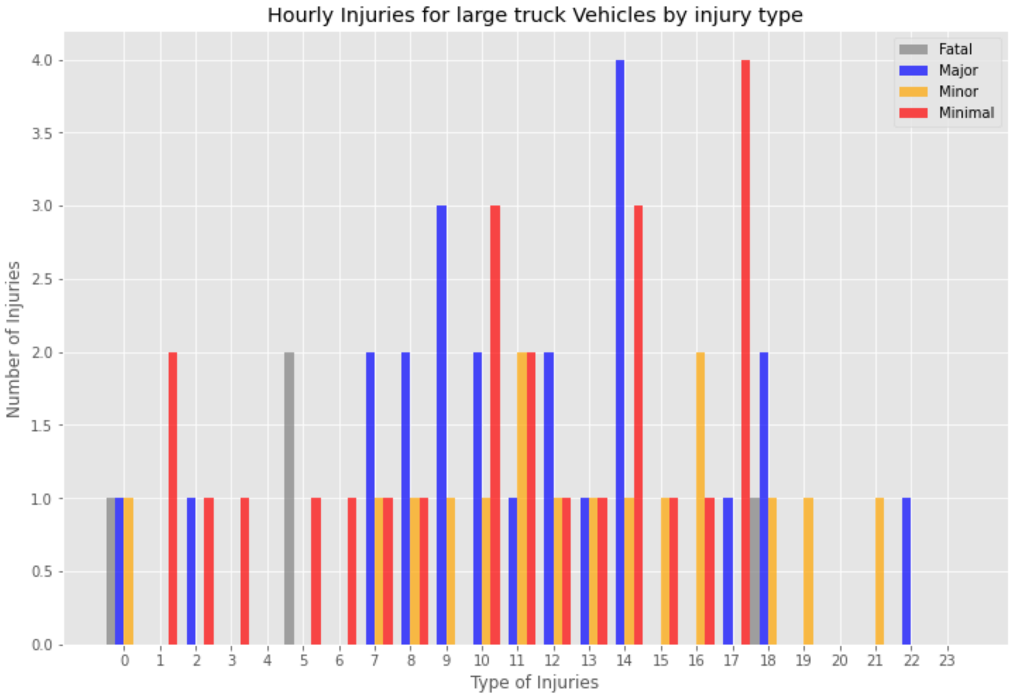


Figure 66: Hourly Injuries for Large Truck Vehicles by Injury Type.

Appendix H: Ottawa x Toronto Comparison - Additional Plots

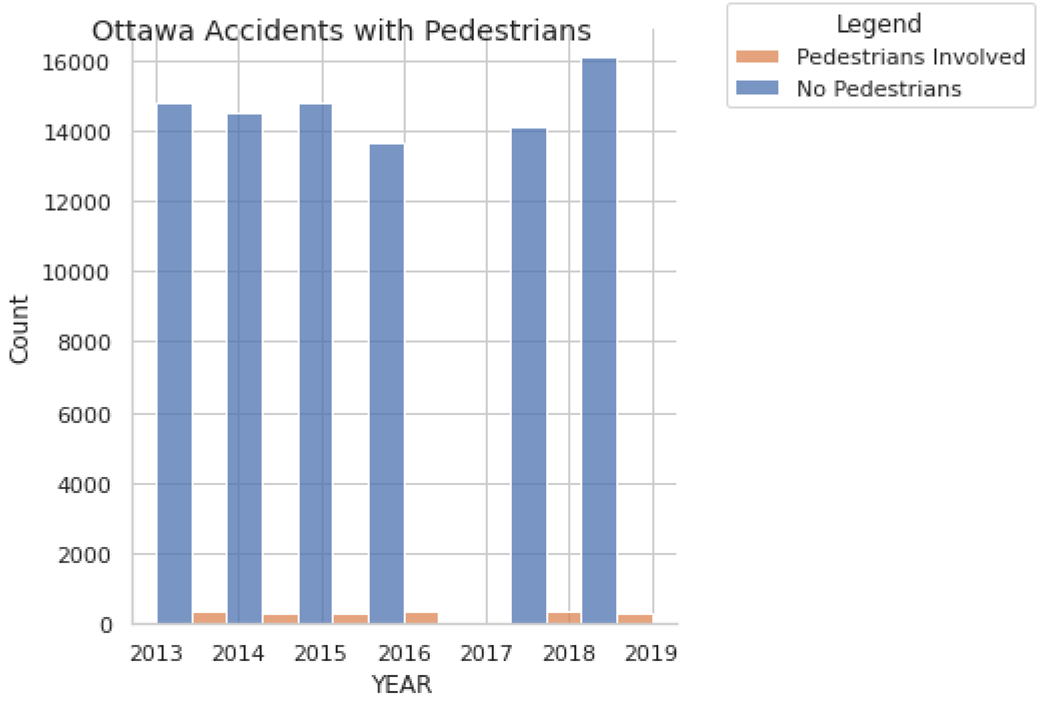


Figure 67: Ottawa Collisions with Pedestrians.

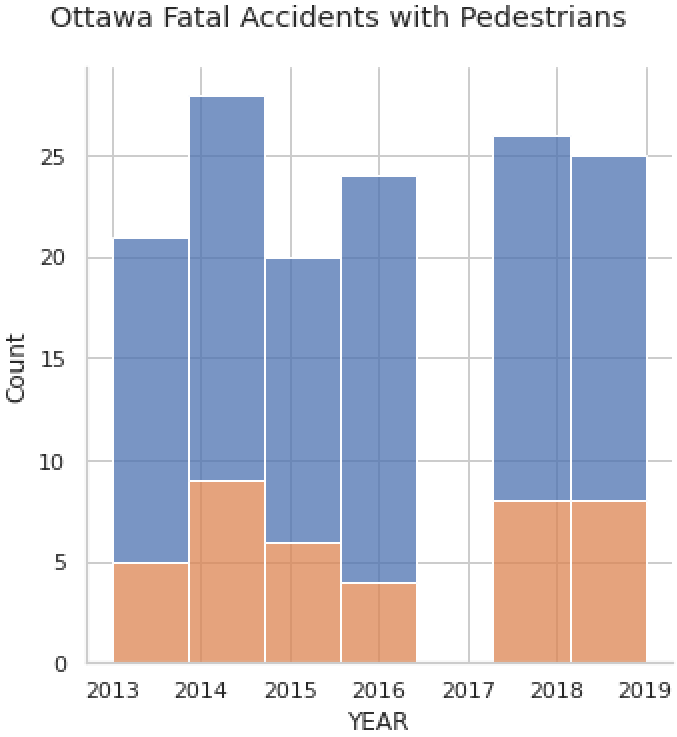


Figure 68: Ottawa Fatal Collisions with Pedestrians.

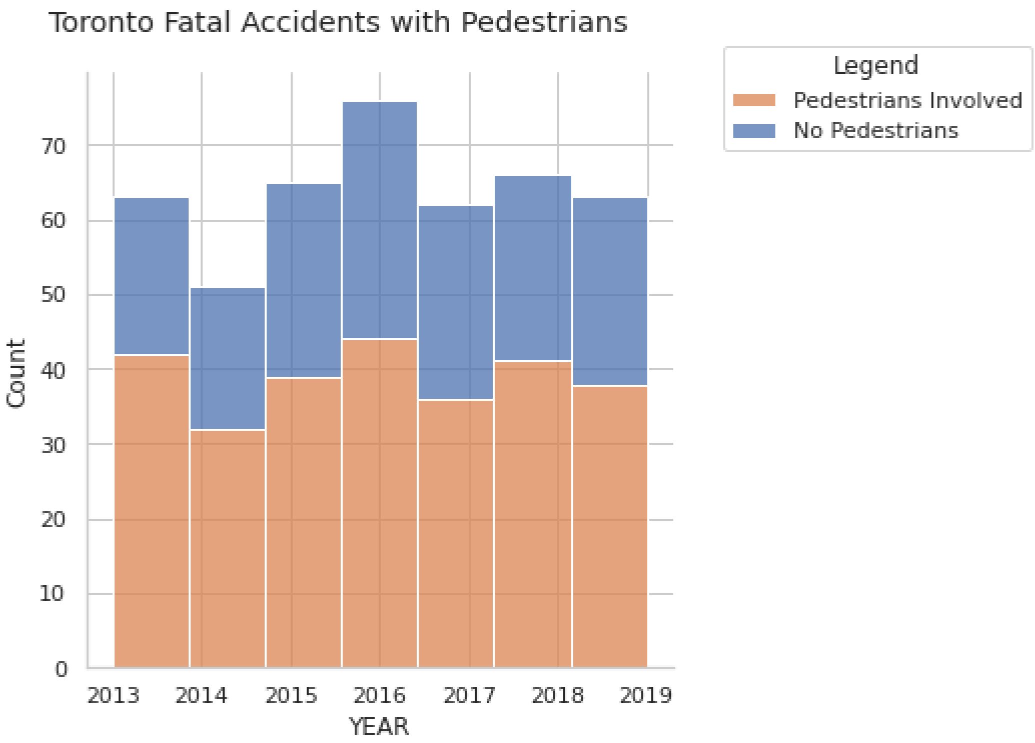


Figure 69: Toronto Collisions with Pedestrians.

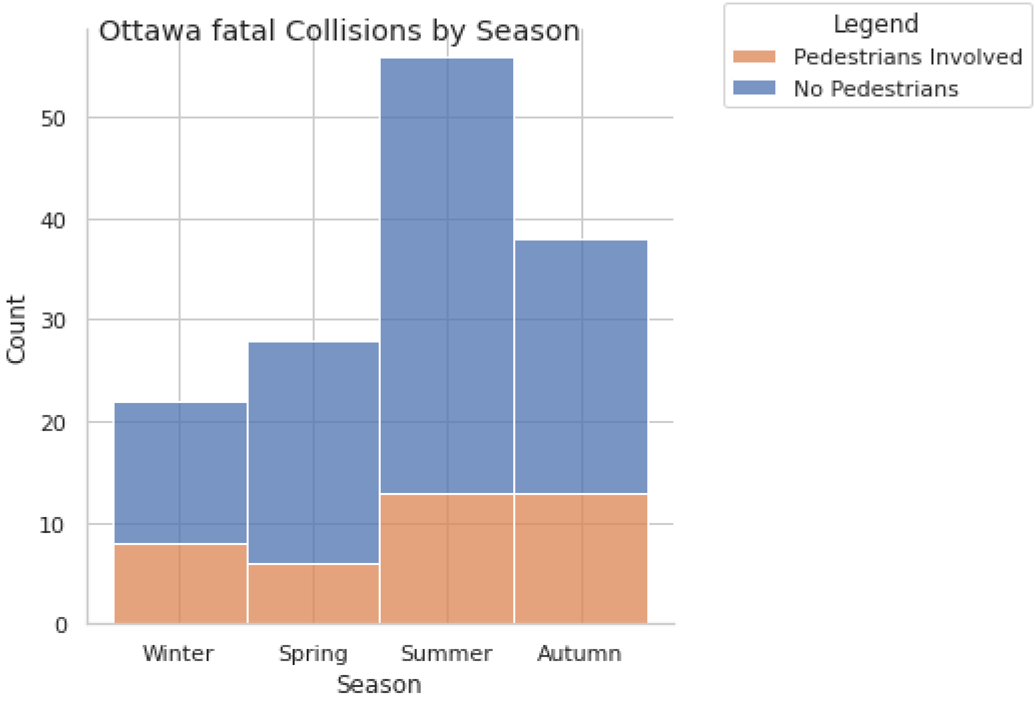


Figure 70: Ottawa Fatal Collisions by Season.

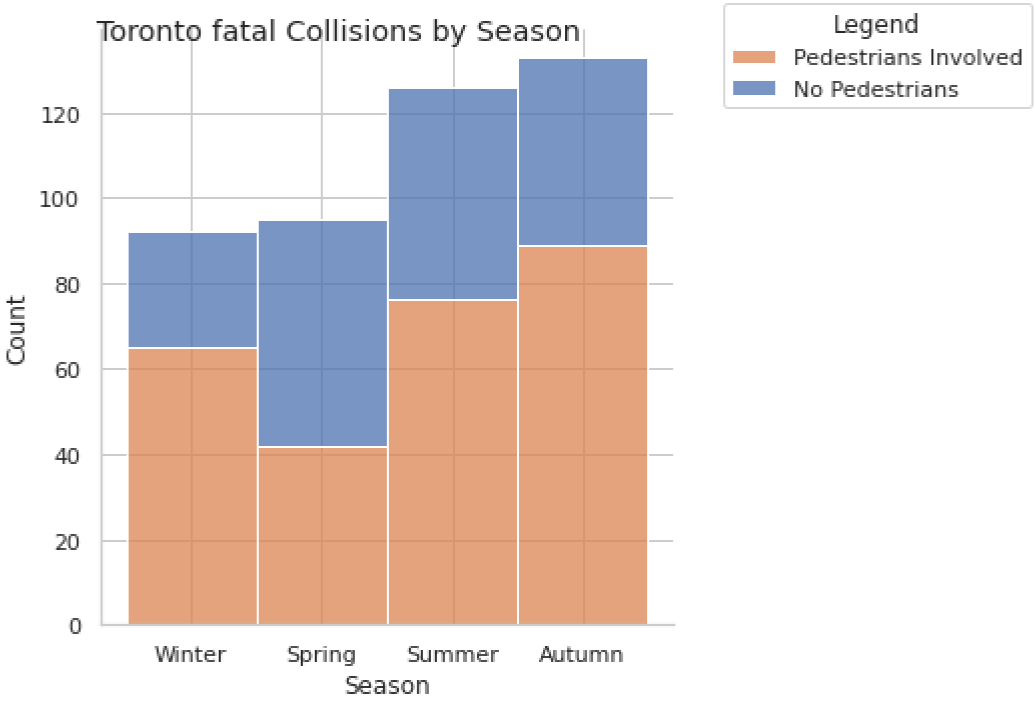


Figure 71: Toronto Fatal Collisions by Season.

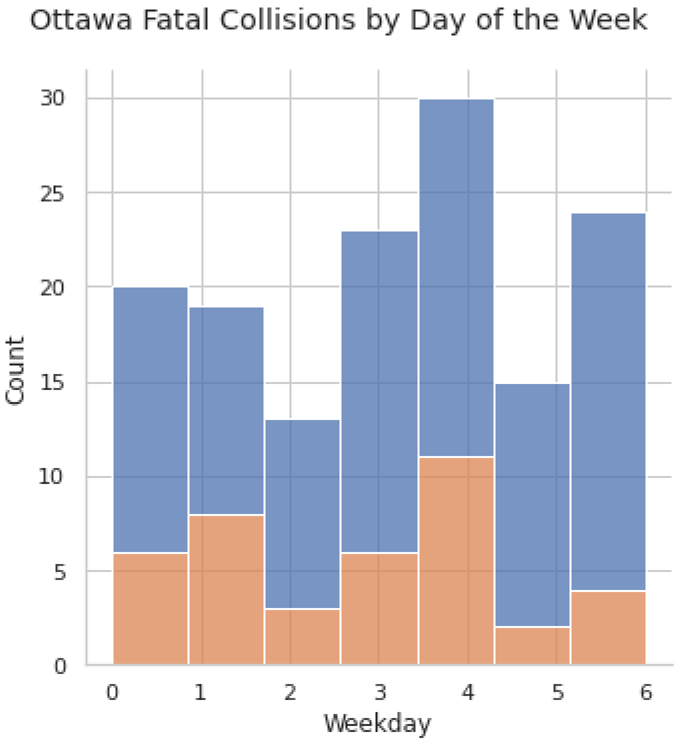


Figure 72: Ottawa Fatal Collisions by Day of the Week.

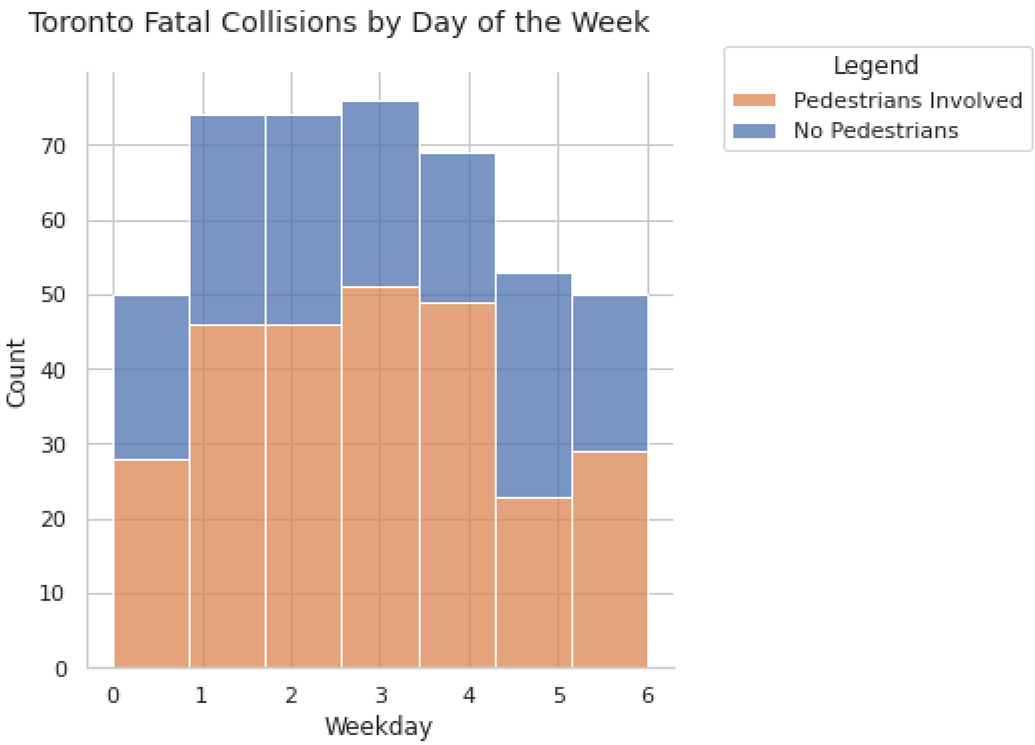


Figure 73: Toronto Fatal Collisions by Day of the Week.

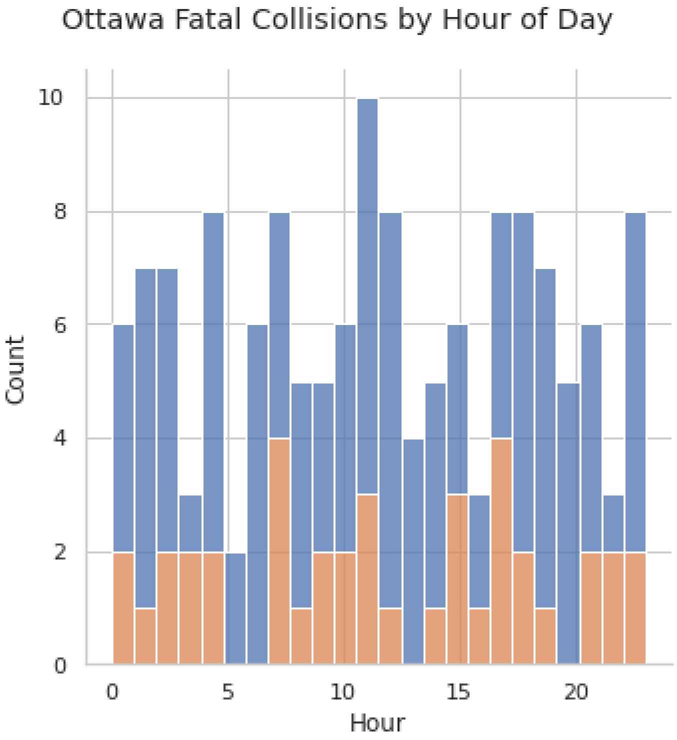


Figure 74: Ottawa Fatal Collisions by Hour of Day.

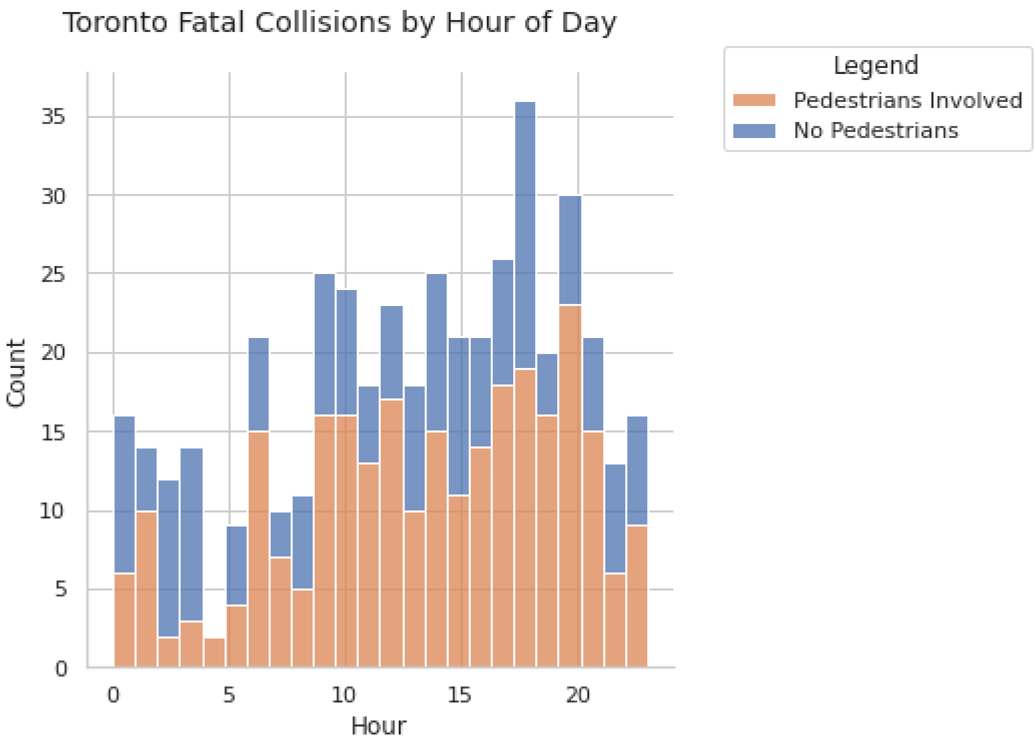


Figure 75: Toronto Fatal Collisions by Hour of Day.

References

"KSI," *Public Safety Data Portal*. [Online]. Available: https://data.torontopolice.on.ca/datasets/ksi/. [Accessed: Nov-2021].

"Open Ottawa collision data," *Open Ottawa*. [Online]. Available: https://open.ottawa.ca/search?q=fatality. [Accessed: Nov-2021].

"National Collision Database," *Open Government Portal*. [Online]. Available: https://open.canada.ca/data/en/dataset/1eb9eba7-71d1-4b30-9fb1-30cbdab7e63a. [Accessed: Nov-2021].

"TTC Routes and Schedules," *City of Toronto Open Data Portal*. [Online]. Available: https://open.toronto.ca/dataset/ttc-routes-and-schedules/. [Accessed: Nov-2021].

"Toronto Red Light Cameras," *City of Toronto Open Data Portal*. [Online]. Available: https://open.toronto.ca/dataset/red-light-cameras/. [Accessed: Nov-2021].

T. S. Dept., "Road Safety Action Plan," *City of Ottawa*, 21-Jul-2021. [Online]. Available: https://ottawa.ca/en/parking-roads-and-travel/road-safety/road-safety-action-plan. [Accessed: 08-Dec-2021].

Ottawa 311, rep., 2013. ,*City of Ottawa*, 2013. [Online]. Available: https://documents.ottawa.ca/en/file/6783/download?token=Ro5t9-aT. [Accessed: 09-Dec-2021]