School of Continuing Studies

3251 – Statistics for Data Sciences

Toronto City – Accident Data Analysis and Building A Prediction Model to Predict Severity of Accident Injury

**Group Project**

**July 23rd, 2019**

**Authors:**

**Group 1**

**Sakshi Sharma, Sridhar Mani, Thangappa Tamilarasan, Vijayalakshmi TA**

# Table of Content

[Table of Content 2](#_Toc14687641)

[1. Introduction 3](#_Toc14687642)

[1.1. Objectives 3](#_Toc14687643)

[2. Data 3](#_Toc14687644)

[2.1. Loading the Data Set 4](#_Toc14687645)

[2.2. Data Analysis 5](#_Toc14687646)

[2.3. Data Transformation 5](#_Toc14687647)

[3. Building A Predictive Model 7](#_Toc14687648)

[3.1. Introduction to the models 7](#_Toc14687649)

[3.1.1. Random Forest Classification 7](#_Toc14687650)

[3.1.2. Logistic Regression Classification 7](#_Toc14687651)

[3.2. Building a prediction Model using Random Forest Classifier method 7](#_Toc14687652)

[3.3. Building a prediction Model using Logistic Regression method 9](#_Toc14687653)

[4. Hypothesis Testing 11](#_Toc14687654)

[4.1. Introduction to Hypothesis Testing 11](#_Toc14687655)

[4.2. Framing the hypothesis: 11](#_Toc14687656)

[4.3. Observations from One Sample T-test: 15](#_Toc14687657)

[4.4. Observations from Two-sample T-test: 15](#_Toc14687658)

[5. Conclusion 15](#_Toc14687659)

[6. References 16](#_Toc14687660)

[7. Appendices 17](#_Toc14687661)

[Appendix A - Headers with the Description in the Original Data Set 17](#_Toc14687662)

# Introduction

Traffic accidents are caused by various factors including but not limited to distracted driving, poor driving conditions, vehicle malfunction, careless driving etc. According to the [World Health Organization](https://en.wikipedia.org/wiki/World_Health_Organization), [road traffic injuries](https://en.wikipedia.org/wiki/Traffic_collision) caused an estimated 1.35 million deaths worldwide in the year 2016.That is, one person is killed every 25 seconds.

Good news is that traffic accidents and critical injury rate in Toronto has been on the decline for the last decade. This is attributed to various reasons such as more awareness among drivers, better infrastructure, higher safety standards in cars etc.

While there is a steady decline in the traffic related accident and resulting serious injury, there is a need for an in-depth analysis of accident data and a predictive model which can predict the criticality of an injury when an accident happens based on various parameters. This key information can be applied in various fields such as hospitals to plan and effectively utilize their resources, insurance companies to optimize insurance rates etc.

## Objectives

Objective of this project is two-fold.

1. Build a prediction model which can predict the severity of an injury when a road related traffic accident happens in Toronto area based on various parameters such as – 1) time of the day 2) neighbourhood 3) road conditions 4) visibility conditions 5) month of the year
2. Based on sample data received, test the alternate hypothesis that severely injured monthly accident rate of last four-year period (2014-2018) is lesser than the severely injured monthly accident rate of previous four-year period (2010 – 2014).

# Data

Source of the data is the Toronto City - killed or severely injured (KSI) accident data set. This data set has information associated with reported accidents in Toronto area between the year 2007 and 2017. The data is available in the form of CSV and the below table provides details of the data attributes.

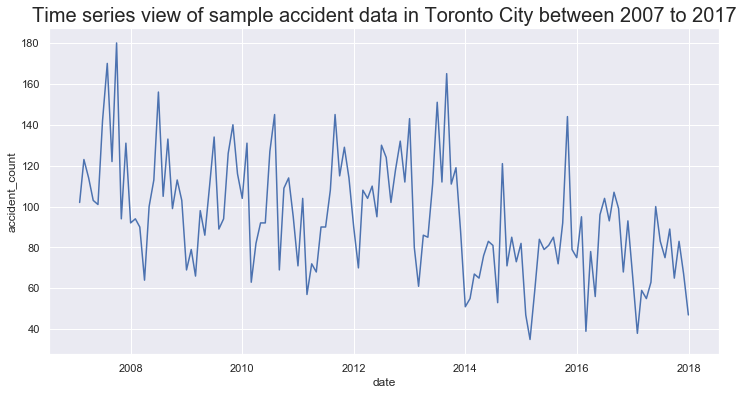
****

Figure 2(a) - shows the Time series view of sample accident data in Toronto City between 2007 to 2017

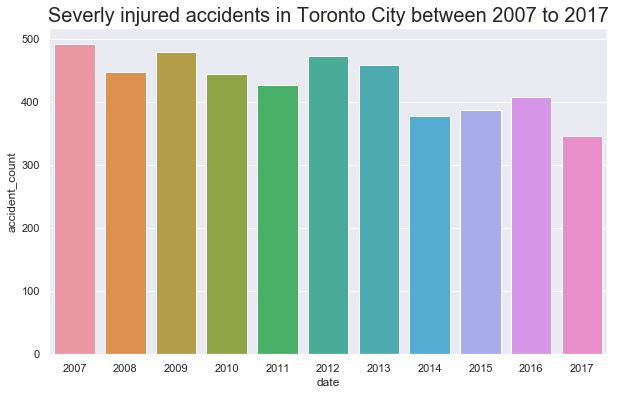


Figure 2(b) shows the Severely injured accidents in Toronto City between 2007 to 2017

## Loading the Data Set

Data was loaded into Python – Jupyter notebook for analysis and model building. Data from CSV was loaded into a Pandas DataFrame object.

## Data Analysis

The loaded data was analysed for completeness and integrity. There were 12557 entries in the CSV data set which was loaded into the DataFrame. A careful analysis of all the columns were performed and the columns which are not deemed critical for analysis and predictive modelling were removed. Following are the list of columns which were removed from the dataframe as they had no significance to data analysis or modelling.

**columns\_to\_drop**=['x','y','index\_','acclass','accnum', 'street1',

'street2', 'offset','latitude', 'longitude',

'loccoord', 'accloc', 'traffctl', 'impactype', 'invtype', 'invage', 'fatal\_no',

'initdir', 'vehtype', 'manoeuver', 'drivact', 'drivcond', 'pedtype',

'pedact', 'pedcond', 'cyclistype', 'cycact', 'cyccond', 'pedestrian',

'cyclist', 'automobile', 'motorcycle', 'truck', 'trsn\_city\_veh',

'emerg\_veh', 'passenger', 'speeding', 'ag\_driv', 'redlight', 'alcohol',

'disability', 'division', 'ward\_name', 'hood\_id',

'hood\_name', 'fid']

## Data Transformation

As the majority of the data attributes are categorical, the data attributes are transformed into numerical values for ease of analysis, consistency and building a prediction model. New data attributes were engineered using available data and some of the existing data attributes were transformed. Below table outlines the list of all new data attributes engineering and existing attributes that were transformed for model building purposes.

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Transformation Logic** | **Values** |
| date (existing) | Date field was trimmed down to remove the timestamp. Also, the date attribute was converted into pandas datetime object | Before: 2011-08-04T04:00:00.000Z  After:  2011-08-04 |
| accident\_month (new) | Based on the date field, month data was extracted | NA |
| critically\_injured (new) | In the existing data set, following are the different values of injury types – None, Minimal, Fatal, Major & Minor. Based on this a new field crically\_injured was derived. | Any accident record with Fatal or Major injury was marked as critically\_injured =1 while rest of the accident rows were marked as not\_critically\_injured = 0 |
| time\_period (new) | Based on the “hour” when the accident occurred, the data was classified with time\_period as 1, 2 and 3. | Classification criteria are as follows:  Hours 00 to 06 is classified as 1 (Night), Hours 07 to 10 and 16 to 19 are classified as 2 (Rush hour), and the rest were classified as 3 (Non Rush hour) |
| road\_type (new) | Based on the road\_class ('Expressway’, Collectors', Major Arterial', 'Minor Arterial','Local', 'Major Arterial Ramp', 'Expressway Ramp', 'Laneway') where accident occurred, road\_type was classified as 1 or 0 where 1 is related to any accidents in expressway and 0 is for other road categories | 1 – Expressway  0 – Other roads |
| poor\_driving\_conditions (new) | Based on the visibility data associated with the original data set, poor\_driving\_conditions were populated based on the following criteria.  1 – Good  2 – Moderate  3 – Poor | 1 – Good  2 – Moderate  3 – Poor |
| district\_code | Based on the area where the accident occurred, a district code value was assigned  (records with no district value assigned was removed from the data set) | 1 – Toronto East York  2 – Scarborough  3 – Etobicoke York  4 – North York  0 – No district |
| season\_code | Based on the month the accident occurred, a season code value was assigned | 1 – Winter – (Dec to March)  2 – Spring – (April to June)  3 – Summer – (July to Sep)  4 – Fall – (Oct to Nov) |

# Building A Predictive Model

Following algorithms were used to build a predictive model to predict the seriousness / criticality of injury based on various parameters.

* Logistic Regression
* Random Forest Classifier

Listed below are the steps followed in building a predictive model:

* Loading the data set
* Analysing the data
* Transforming the data
* Building the model
* Testing & tuning the model
* Accuracy score analysis

# Introduction to the models

## Random Forest Classification

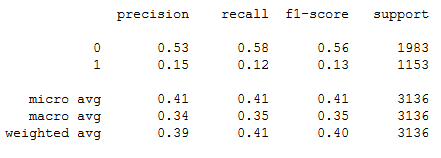
Random forests or random decision forests are an ensemble learning methods for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

## Logistic Regression Classification

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1".

## Building a prediction Model using Random Forest Classifier method

The accident data set is split into two groups - train and test. 25% of the data is randomly selected as test data set while remaining 75% data is selected as training data set. The training data is then fitted under Random Forest Classifier model. Once fitted, the test data set is ran against the fitted model and below results indicate, this model resulted in a weighted average prediction of 39%.



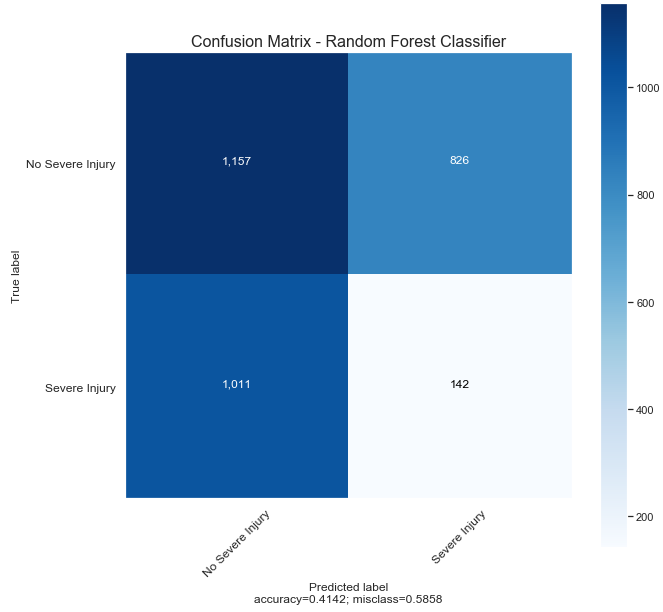


Figure 3-2(a)

As the prediction accuracy is only 39%, the data is further analysed to find the best parameter (n\_estimator) which would give a higher prediction accuracy. A test was conducted for various values of n\_estimator [1, 2, 4, 8, 16, 32, 64, 100, 200] and accuracy score was measured.

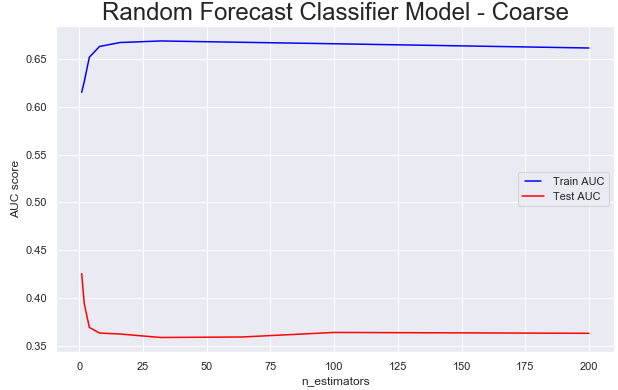


Figure 3-2(b)

As per the above graph, it is evident that the value of estimators cannot be further adjusted to increase the accuracy score.

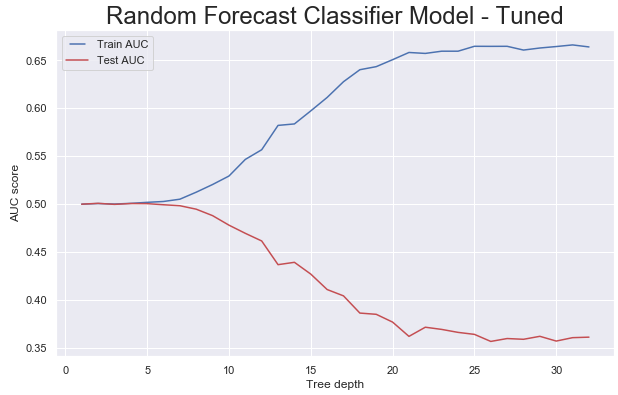


Figure 3-2(c)

In addition to tuning the n\_estimator value, various value of max\_depth parameter was tested to see if it improves accuracy score. But as the above results indicate, changing the max\_depth value doesn’t seem to improve the prediction accuracy significantly.

## Building a prediction Model using Logistic Regression method

The accident data set is split into two groups - train and test. 25% of the data is randomly selected as test data set while remaining 75% data is selected as training data set. The training data is then fitted under Logistic Regression Model. Once fitted, the test data set is running against the fitted model and below results indicate, this model resulted in a weighted average prediction of 38%.

precision recall f1-score support

0 0.62 1.00 0.76 1932

1 0.00 0.00 0.00 1204

micro avg 0.62 0.62 0.62 3136

macro avg 0.31 0.50 0.38 3136

weighted avg 0.38 0.62 0.47 3136

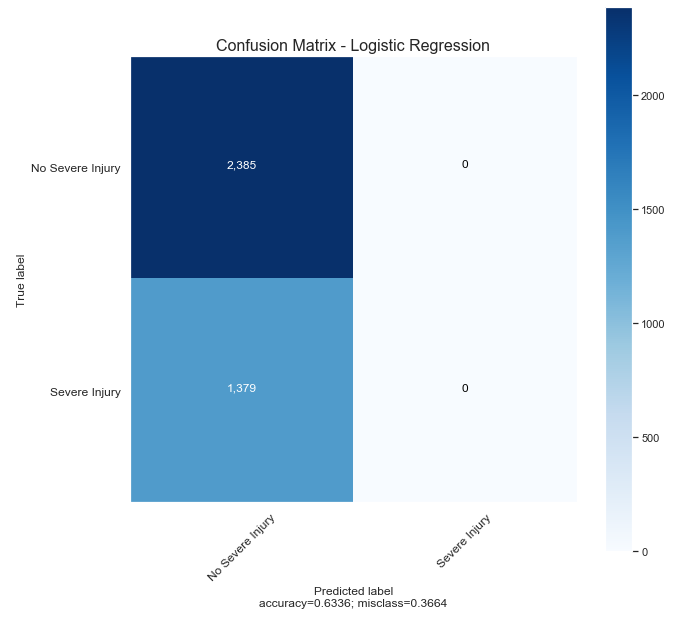


Figure 3-3(a)

# Hypothesis Testing

## Introduction to Hypothesis Testing

A statistical hypothesis, sometimes called confirmatory data analysis, is a [hypothesis](https://en.wikipedia.org/wiki/Hypothesis) that is testable on the basis of [observing](https://en.wikipedia.org/wiki/Observable_variable) a process that is [modeled](https://en.wikipedia.org/wiki/Statistical_model) via a set of [random variables](https://en.wikipedia.org/wiki/Random_variable). A statistical hypothesis test is a method of [statistical inference](https://en.wikipedia.org/wiki/Statistical_inference). Commonly, two statistical data sets are compared, or a data set obtained by sampling is compared against a synthetic data set from an idealized model. A hypothesis is proposed for the statistical relationship between the two data sets, and this is compared as an [alternative](https://en.wikipedia.org/wiki/Alternative_hypothesis) to an idealized null hypothesis that proposes no relationship between two data sets. The comparison is deemed [statistically significant](https://en.wikipedia.org/wiki/Statistically_significant) if the relationship between the data sets would be an unlikely realization of the [null hypothesis](https://en.wikipedia.org/wiki/Null_hypothesis) according to a threshold probability—the significance level. Hypothesis tests are used when determining what outcomes of a study would lead to a rejection of the null hypothesis for a pre-specified level of significance.

## Framing the hypothesis:

Number of severely injured road accidents in Toronto City is on the decline. To provide this, a hypothesis test is conducted. Null and Alternate Hypothesis are as follows:

1. **Null Hypothesis:** There is no difference between monthly average of critically injured road accidents in the recent 4-year period (2014-2018) and previous 4-year period (2010-2014)

HO: Mean (mu) = 36

1. **Alternate Hypothesis:** monthly average of critically injured road accidents in Toronto City has declined during the 4-year period (2014-2018) when compared with previous 4-year period (2010-2014)

HA: Mean (mu) < 36

In order to test the above hypothesis, student t-test method was chosen as the data set meets the following criteria to run a student t-test:

* Data sets are independent observation
* Data sets are near normal or normal distribution
* Observations are > 30
* There are no outliers

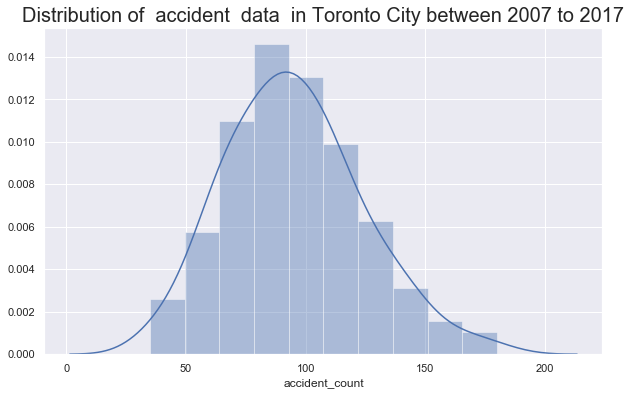
****

Figure 4-2(a) **Distribution plot of monthly severely injured accident rate between 2007 to 2017**

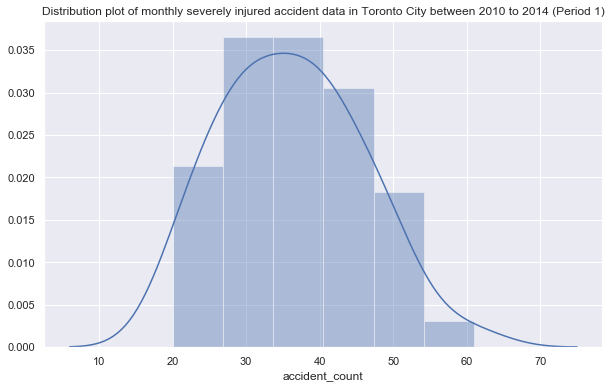
****

Figure 4-2(b) Distribution **plot of monthly severely injured accident data in Toronto City between 2010 to 2014 (Period 1)**

**Severely injured accident data in Toronto City between 2010 to 2014 (Period 1)**

|  |  |
| --- | --- |
| Parameter | Value |
| Count | 48.00 |
| Mean | 36.14 |
| Std | 9.54 |
| Min | 20.00 |
| 25% | 29.00 |
| 50% | 36.00 |
| 75% | 42.00 |
| Max | 61.00 |

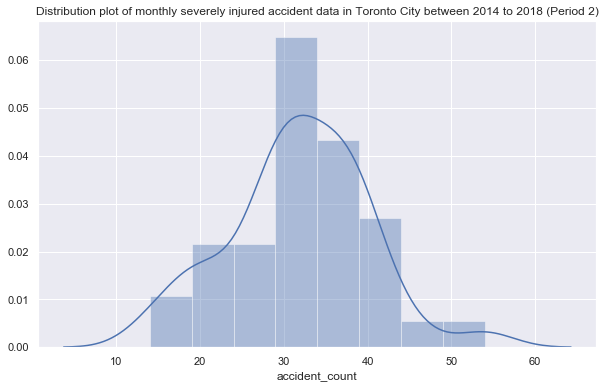


Figure 4-2(c) Distribution **plot of monthly severely injured accident data in Toronto City between 2014 to 2018 (Period 2)**

**Severely injured accident data in Toronto City between 2010 to 2014 (Period 2)**

|  |  |
| --- | --- |
| Parameter | Value |
| Count | 37.00 |
| Mean | 31.78 |
| Std | 8.31 |
| Min | 14.00 |
| 25% | 28.00 |
| 50% | 31.00 |
| 75% | 37.00 |
| Max | 54.00 |

## Observations from One Sample T-test:

After running a one-sample t-test using stats package in python, following are the results obtained:

*Ttest\_1sampResult(statistic=-3.19149, pvalue=0.00293)*

Above results indicate that the probability of observing mean value of 36 by chance is very less than the significance level of 0.05. Hence Null Hypothesis is rejected and Alternate Hypothesis is accepted. Therefore, the number of accidents per month have decreased since the 2010-2014 period.

## Observations from Two-sample T-test:

After running a two-sample t-test using stats package in python, following are the results obtained:

*Ttest\_indResult(statistic=2.207, pvalue=0.03007*)

Above results indicate that the probability of observing two samples having an identical average by chance is very less than the significance level of 0.05. Hence Null Hypothesis is rejected and Alternate Hypothesis is accepted. Therefore, the number of accidents per month have decreased since the 2010-2014 period.

# Conclusion

Data set having information on Killed or Seriously injured accidents in Toronto area between 2007 to 2017 was analysed with two key objectives:

1. Build a model to predict the severity of an accident in the event an accident occurs in Toronto city based on various parameters
2. Test the alternate hypothesis that monthly rate of accidents resulting in serious injury has declined in the last 4-year period (2014-2018) when compared with its previous 4-year period (2010-2014).

Prediction model was built using Random Forest Classifier method and Logistic Regression method. While Random Forest Classifier approach resulted in 39% accuracy, Logistic Regression approach resulted in only 38% accuracy. Also, tuning the model did not result in significant increase in accuracy scores. As both the models have only < 50% accuracy score, prediction models cannot be treated final and used in Production for real use. This suggests that there are more opportunities to broaden the data set by adding more data attributes and/or adapt different approach to improve the accuracy score of the prediction model.

Alternate hypothesis was tested using both one sampled and two-sample t-test and the results indicate that the null hypothesis which states there is no difference between severely injured accident rate between two periods under study (2010-2014 & 2014-2018) can be rejected and alternate hypothesis which states that there is a decline in severely injured accident rate from 2010-14 period to 2014-18 can be safely accepted.

# References

1. <https://en.wikipedia.org/wiki/List_of_countries_by_traffic-related_death_rate>
2. <https://data.torontopolice.on.ca/datasets/ksi>
3. <https://en.wikipedia.org/wiki/Statistical_hypothesis_testing>
4. <https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/>
5. <https://en.wikipedia.org/wiki/Random_forest>

# Appendices

## Appendix A - Headers with the Description in the Original Data Set

|  |  |  |
| --- | --- | --- |
| **Number** | **Field name** | **Description** |
| 1 | Index | Unique Identifier |
| 2 | ACCNUM | Accident Number |
| 3 | YEAR | Year Accident Occurred |
| 4 | DATE | Date Accident Occurred |
| 5 | TIME | Time Accident Occurred |
| 6 | HOUR | Hour Accident Occurred |
| 7 | STREET1 | Street Accident Occurred |
| 8 | STREET2 | Street Accident Occurred |
| 9 | OFFSET | Distance and direction of the accident |
| 10 | ROAD\_CLASS | Road Classification |
| 11 | District | City District |
| 12 | LATITUDE | Latitude |
| 13 | LONGITUDE | Longitude |
| 14 | LOCCOORD | Location Coordinate |
| 15 | ACCLOC | Accident Location |
| 16 | TRAFFCTL | Traffic Control Type |
| 17 | VISIBILITY | Environment Condition |
| 18 | LIGHT | Light Condition |
| 19 | RDSFCOND | Road Surface Condition |
| 20 | ACCLASS | Classification of Accident |
| 21 | IMPACTYPE | Initial Impact Type |
| 22 | INVTYPE | Involvement Type |
| 23 | INVAGE | Age of Involved Party |
| 24 | INJURY | Severity of Injury |
| 25 | FATAL\_NO | Sequential Number |
| 26 | INITDIR | Initial Direction of Travel |
| 27 | VEHTYPE | Type of Vehicle |
| 28 | MANOEUVER | Vehicle Maneuver |
| 29 | DRIVACT | Apparent Driver Action |
| 30 | DRIVCOND | Driver Condition |
| 31 | PEDTYPE | Pedestrian Crash Type - detail |
| 32 | PEDACT | Pedestrian Action |
| 33 | PEDCOND | Condition of Pedestrian |
| 34 | CYCLISTYPE | Cyclist Crash Type - detail |
| 35 | CYCACT | Cyclist Action |
| 36 | CYCCOND | Cyclist Condition |
| 37 | PEDESTRIAN | Pedestrian Involved In Collision |
| 38 | CYCLIST | Cyclists Involved in Collision |
| 39 | AUTOMOBILE | Driver Involved in Collision |
| 40 | MOTORCYCLE | Motorcyclist Involved in Collision |
| 41 | TRUCK | Truck Driver Involved in Collision |
| 42 | TRSN\_CITY\_VEH | Transit or City Vehicle Involved in Collision |
| 43 | EMERG\_VEH | Emergency Vehicle Involved in Collision |
| 44 | PASSENGER | Passenger Involved in Collision |
| 45 | SPEEDING | Speeding Related Collision |
| 46 | AG\_DRIV | Aggressive and Distracted Driving Collision |
| 47 | REDLIGHT | Red Light Related Collision |
| 48 | ALCOHOL | Alcohol Related Collision |
| 49 | DISABILITY | Medical or Physical Disability Related Collision |
| 50 | Police Division | Police Division |
| 51 | City Ward | City Ward |
| 52 | City Ward ID | City Ward Identifier |
| 53 | Neighbourhood ID | Neighbourhood Identifier |
| 54 | Neighbourhood Name | Neighbourhood Name |
| 55 | FID | Object ID (Unique Identifier) |
| 56 | X | Latitude |
| 57 | Y | Longitude |