# Mobility Investigation for City of Toronto

## Model Prediction for Collisions Numbers at Intersections, Based on Traffic Volume Factors

Christen (Yuchen) Ye

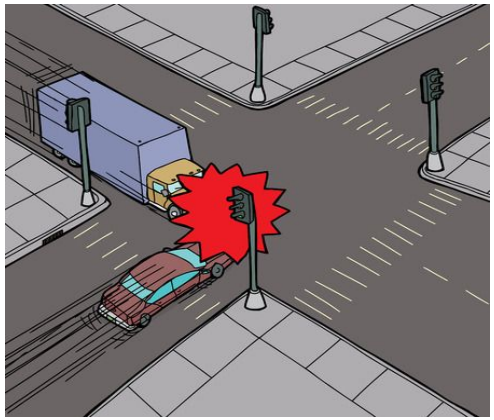
# Introduction & Research Question

Client: City of Toronto

Case Description: analyze available data from 2014 to 2021 so that they can understand the mobility in the metropolitan area.

***Research Question:***

***Which traffic-volume-related factors exert a significant effect on collision number in the intersections?***



We expect that the expected collision number in each intersection will depend linearly on traffic-volume factors. The predictive model includes six variables of interest: total volumes observed at specific intersections of cars, trucks, buses, pedestrians, cyclists, and others.

*https://www.tariolaw.com/determining-fault-in-intersection-car-accidents/

# Data Analysis & Data Visualization

## Pre-Analysis Data Cleaning:

- Traffic Collisions - 4326. csv (**Collision Table**):

1) Split 'geometry' (offset to nearest intersection) and create new columns ‘lng’ and ‘lag’
2) Round ‘lng’ and ‘lat’ to 4 digits then drop missing or 0 value observations
3) Create ‘location id column’, by joining Collision Table with Reference Table

   **Dataset Size after Cleaning**: **188194** unique observed collisions at intersections from 2014 to 2021

- raw-data-2010-2019.csv, raw-data-2020-2029.csv (**Traffic Volume Table**):

1) Concatenate the Traffic Volume table for 2010-2019 and 2020-2029, drop data from 2010 to 2013 and 2022
2) Keep rows that ‘centreline_type’ =2, this corresponds with an intersection
3) Round ‘lng’ and ‘lat’ to 4 digits

   **Dataset Size after Cleaning**: **202508** unique observations for traffic volume at intersections from 2014 to 2021

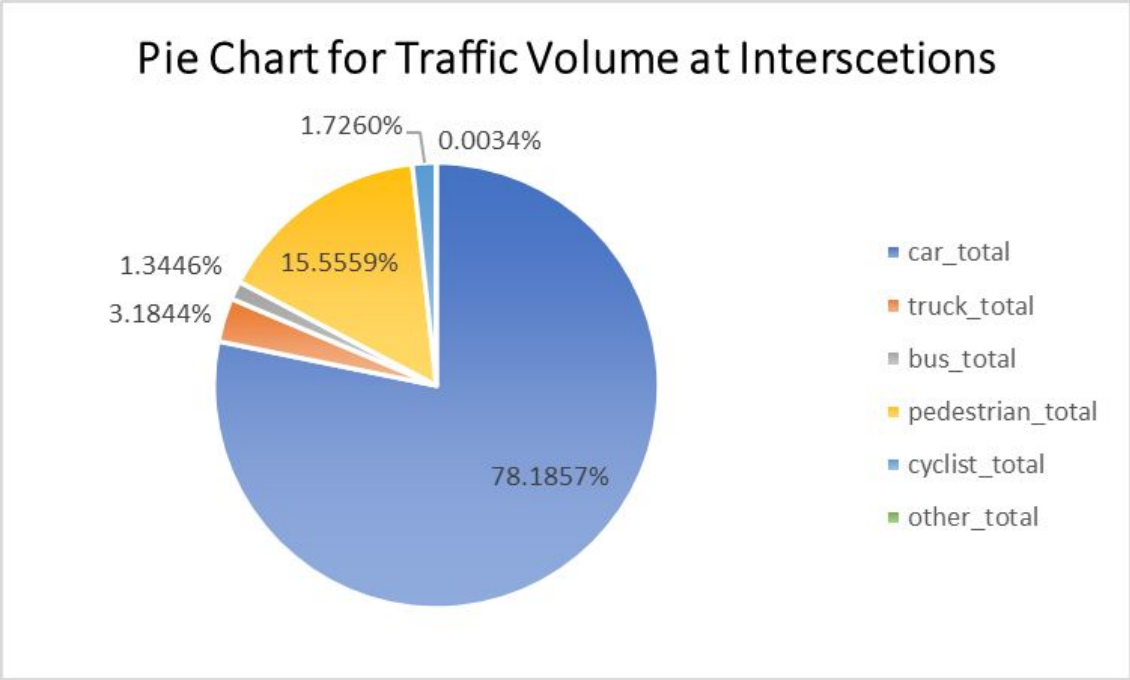
- location.csv (**Reference Table**):

1) Keep rows that ‘centreline_type’ =2, this corresponds with an intersection
2) Round ‘lng’ and ‘lat’ to 4 digits

   **Dataset Size after Cleaning**: **5570** unique location_id for intersections

# Data Analysis & Data Visualization

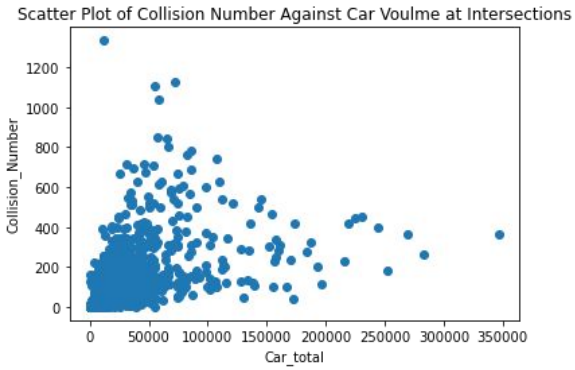
**Exploring the distribution of traffic-volume-related factors:**

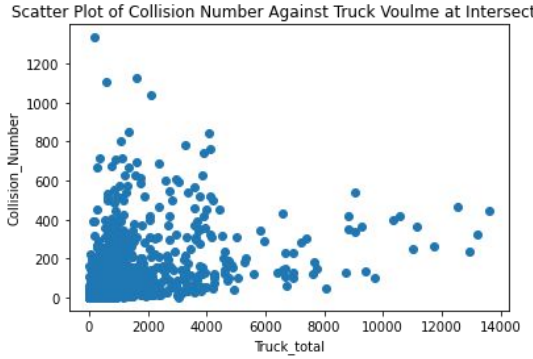


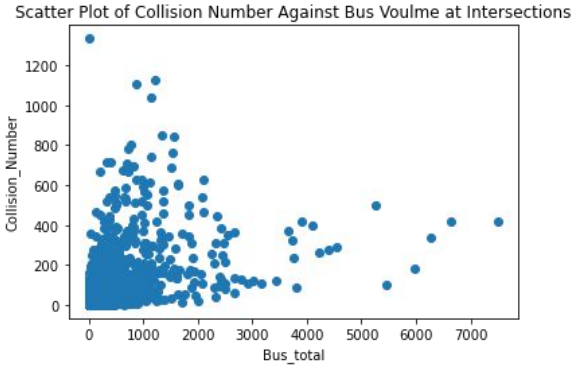

**Conclusion**: Car and Pedestrian are main resources of traffic volume at intersections, taking about 93% in total.Other is the least important resource of  traffic volume at intersections.

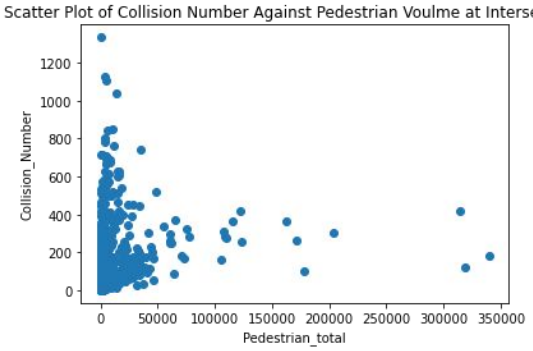
# Data Analysis & Data Visualization

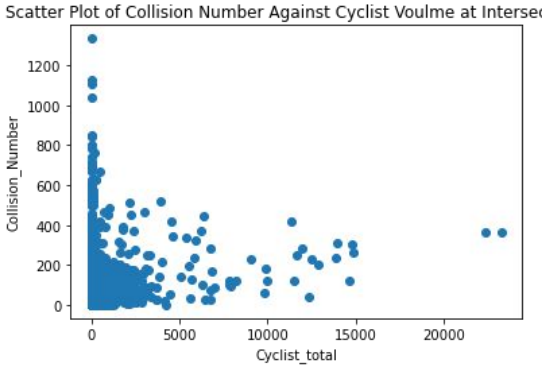
**Exploring if there exists a linear relationship between the number of collisions and traffic volume factors at intersections:**

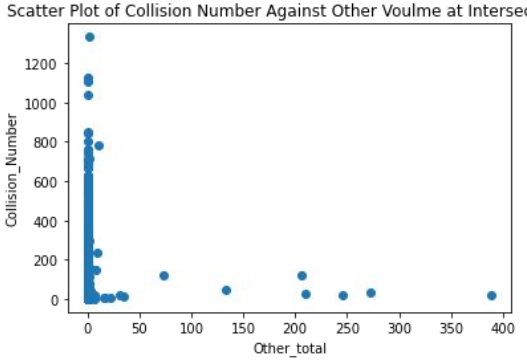


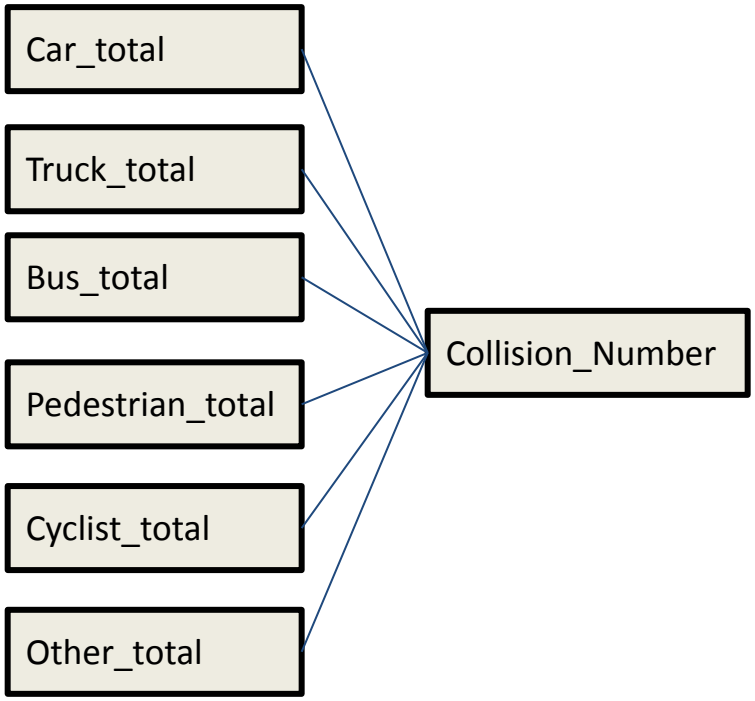

**Conclusion:** There seems to exist a linear relationship between the response variable of collision number and the selected variables, namely car, truck, bus and cyclist volumes at intersections. In addition, we can tell that a log transformation on pedestrian and other volume might be helpful in building a predictive model.

# Predictive Model & Model Adjustment

**Full Model:**

Table 1: Main Results

| | *Dependent variable:* |
|---|---|
| | Collisions |
| car_total | 0.003*** |
| | (0.0002) |
| truck_total | −0.009*** |
| | (0.003) |
| bus_total | 0.008 |
| | (0.007) |
| pedestrian_total | 0.00001 |
| | (0.0002) |
| cyclist_total | −0.020*** |
| | (0.002) |
| other_total | −0.012 |
| | (0.168) |
| Constant | 27.658*** |
| | (3.057) |
| Observations | 1,877 |
| $R^2$ | 0.328 |
| Adjusted $R^2$ | 0.326 |
| Residual Std. Error | 105.003 (df = 1870) |
| F Statistic | 152.338*** (df = 6; 1870) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |



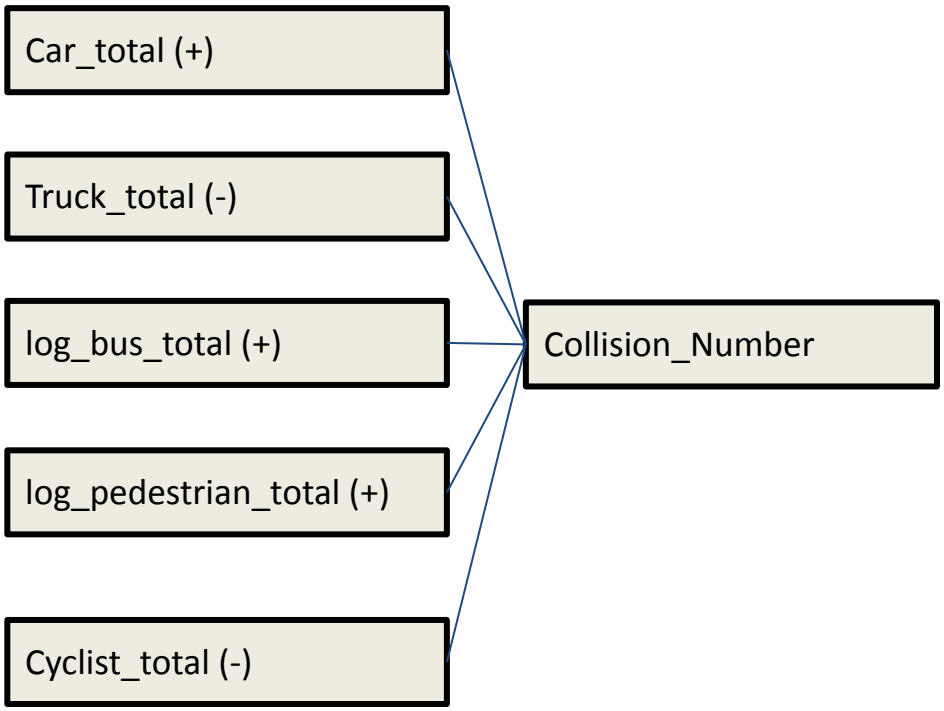
# Predictive Model & Model Adjustment

**Adjusted Model (Drop other_total, take log 10 of pedestrian_total and bus_total):**

*Collision_Number = -114.1 + 0.002644 * car_total - 0.00679 * truck_total + 8.79 * log(bus_total) + 48.02 * log(pedestrian_total) - 0.026 * cyclist_total*

Table 1:

| | *Dependent variable:* |
|---|---|
| | Collision_Number_at_Each_Intersection |
| car_total | 0.003*** |
| | (0.0002) |
| truck_total | −0.007** |
| | (0.003) |
| log_bus_total | 8.790** |
| | (4.010) |
| log_pedestrian_total | 48.018*** |
| | (3.787) |
| cyclist_total | −0.026*** |
| | (0.002) |
| Constant | −114.075*** |
| | (11.286) |
| Observations | 1,877 |
| $R^2$ | 0.390 |
| Adjusted $R^2$ | 0.388 |
| Residual Std. Error | 100.074 (df = 1871) |
| F Statistic | 238.803*** (df = 5; 1871) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |



# Conclusion & Discussion

**Suggestions for Limitations:**

- Include more characteristics for the intersections to avoid omitted variables, like road width
- Assign location_id to the Collisions table, making future data analysis easier
- Add the type of vehicle involved in collisions into the Collisions data, for future conditional analysis
- Split the ‘geometry’ column in Collision table and create new columns ‘lng’ and ‘lag’, increasing readability

| geometry |
| --- |
| {'type': 'Point', 'coordinates': (-79.4149080887603, 43.6504985877039)} |

| lng | lag |
| --- | --- |
| -79.4149 | 43.6505 |

# Appendix & References

## Dataset link:

Traffic Volumes at Intersections for All Modes:
https://open.toronto.ca/dataset/traffic-volumes-at-intersections-for-all-modes/

Police Annual Statistical Report - Traffic Collisions:
https://open.toronto.ca/dataset/police-annual-statistical-report-traffic-collisions/

## Code (Python):

https://drive.google.com/file/d/18SkvNvUtpjBMfInhYxqYLYYHfFq6TT4t/view?usp=sharing

## Code (R):

https://drive.google.com/file/d/1cuWKQdzQOYkd4WpimoAgxQd8AljC1wyz/view?usp=sharing

# Thank you