TORONTO POLICE SERVICE

IMPLEMENTATION OF MACHINE LEARNING TECHNIQUES TO UNDERSTAND TENDENCIES IN FATAL COLLISIONS

Agenda

- Client Overview
- Mission Statement & Core Values
- Section 1: Define the Problem
- Section 2 : Methodology
- Section 3: Key Findings & Recommendations
- Section 4: Predictive Modeling & Recommendations
- Conclusion
- Appendix

Company Overview

The Toronto Police Service (TPS) is a municipal police force in Toronto, Ontario, Canada, and the primary agency responsible for providing law enforcement and policing services in Toronto. Working in partnership with communities, the TPS keeps Toronto safe through:

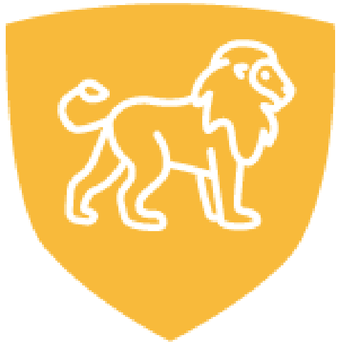
- Community-based crime prevention initiatives
- Enforcement for all applicable laws in the City of Toronto including Provincial Offenses, the Highway Traffic Act and City bylaws
- Maintaining public order to ensure safe and secure communities
- Providing emergency response to major threats and public safety risks.

Mission Statement & Core Values

"We are dedicated to delivering police services, in partnership with our communities, to keep Toronto the best and safest place to be."

Service at our Core

"Have I done all I can do?"



Do the right thing

"Have I lived up to my word and values?"

Connect with Compassion

"Have I treated others as they would like to be treated?"

Reflect and Grow

"What else can I do to improve?"

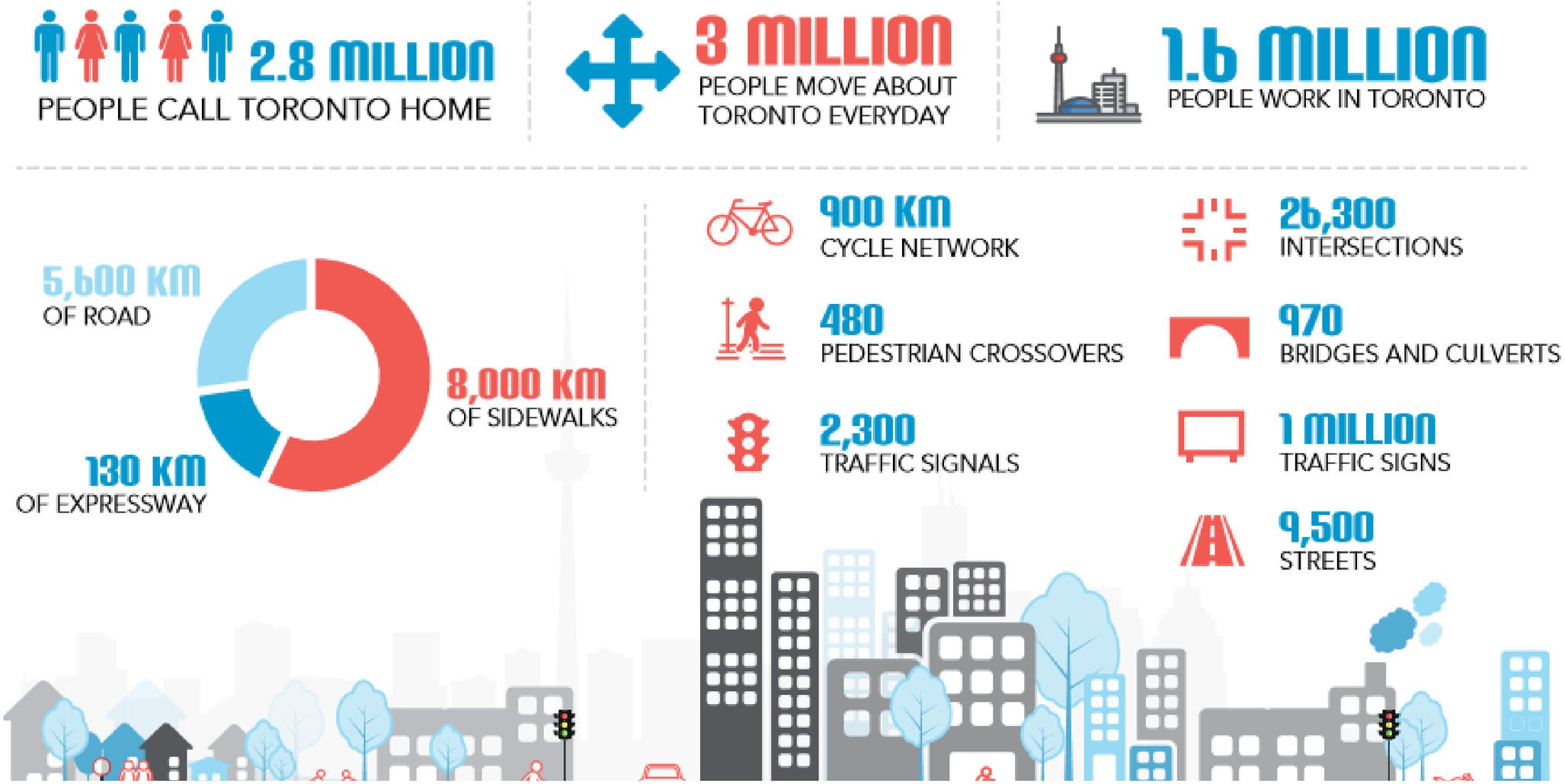
Why do we want to focus on fatal collisions in Toronto?

Section-1

(Defining the Problem)

Support Vision Zero Road Safety Plan

The Vision Zero Road Safety Plan is a comprehensive five year (2017-2021) action plan focused on reducing traffic-related fatalities and serious injuries on Toronto's streets to near zero



Source: Vision Zero: Toronto's Road Safety Plan

Fatality Rate in Toronto v/s North American cities

- Compared to the top 10 North American cities, Toronto's fatality rate is among the lowest.

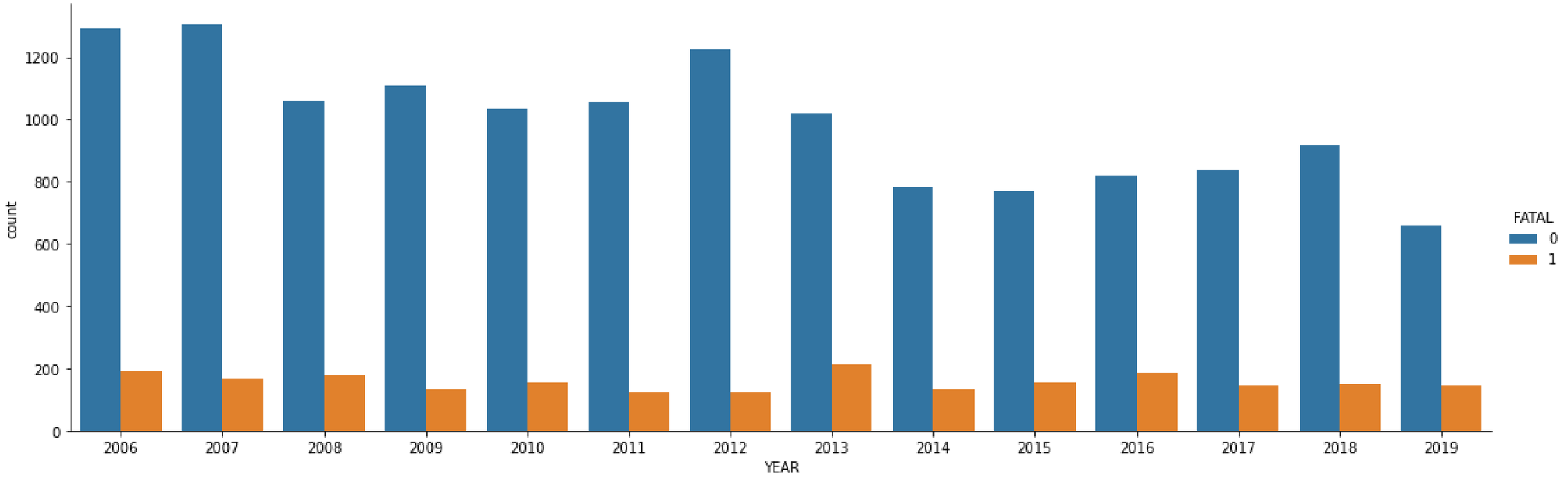
- Continue advancements in the road safety so that as more and more people use the system, their safety can be improved further.

FATALITY COLLISION RATE FOR TEN LARGEST CITIES IN NORTH AMERICA			
Chicago	1.29	Philadelphia	2.44
New York	1.47	Houston	2.68
San Diego	2.32	Dallas	3.2
San Jose	2.17	San Antonio	3.69
Los Angeles	2.44	Phoenix	4.36
Toronto	**1.82**		

Source: National Highway Traffic Safety Administration, Traffic Safety Facts 2014, Rate per 100,000 Population

FATAL V/S Non FATAL Collisions

- Non Fatal Collisions (0's) are higher than fatal (1's) collisions.
- Non-fatal collisons have declined from 2013 onwards
- Fatal collsions have remained nearly constant.
- It is important to analyze fatal and understand ways to reduce them.



Defining the Problem

- **How can Toronto Police Force leverage the patterns observed in the data to bring fatal collisions down to near zero ?**

- **Which variables are major contributors towards fatal collisions in Toronto?**

How can we focus on fatal collisions ?

Section 2
(Methodology)

Data Auditing

We had 2 EXCEL files: Fatal Collsions and Killed & Seriously Injured (KSI) File. The Fatal Collsions is a subset of the KSI file. We decided to use KSI file for our analysis since it had more information.

Metadata

It defines each variable type.

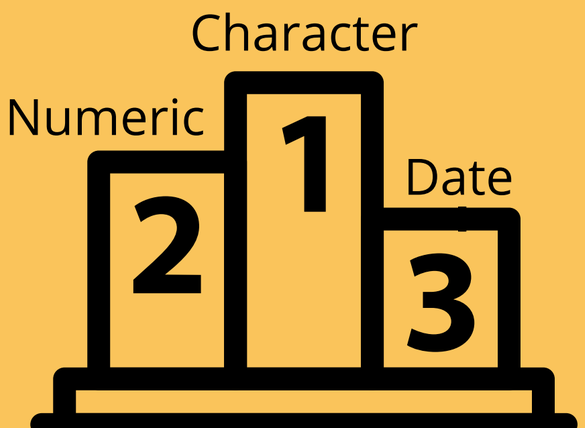
Ex:ACCNUM= Accident Number

Total Variables = 56

Σ

Type of Variables

It tells the type of variables in the dataset. Below is a visual showing type of variables in our dataset.



Quality of Data

Data was assessed from the perspective of its completeness or accuracy. Following things were observed:

- Missing Values
- Redundant Data
- Many categories with empty cells
- Formatting Errors

Frequency Distribution

It shows the number of records for different variables in the datset

PEDESTRIAN	Frequency
Yes	445
NA	336

CYCLIST	Frequency
Yes	37
NA	744

MOTORCYCLE	Frequency
Yes	66
NA	715

AUTOMOBILE	Frequency
Yes	651
NA	130

Key Findings From Data Audit

Unique or Primary ID: We already have unique key in our data. So, we do not need to work on making matchkeys.

Yes/NA VAlues: Some variables have Yes-NA values but we can assume that NAs represent No in this context. We have found this by looking at the other variables.

Unique Values: Some of the variables mostly have unique values such as Street Names. There are only 1 or 2 which are repetitive. We assume these are where traffic and people are more in frequency comapred to other locations.

Observation: Most collisions occurred when the weather was clear and 6 pm and 8 pm are the most common incident hours.

Check Appendix for detailed information

Analytical File

Step1: Source Variable

YEAR: It gives a sense of time and we can count the number of accidents that happened in different years and compare them.

ACCNUM (Accident Number) We use it to find details about each accident

INVTYPE (Involvement Type): It gives us information about the type of commuter whether it was automobile driver, motorcyclist, pedestrian etc. and it can also tell us about the type of vehicle involved.

Step 2: Derived Variable

Month: It can be used to predict or analyze the specific months, seasons, weather conditions which might be contributing towards fatal collision.

Day: It can give an insight on which day the collision rates are high or low.

Minute: It can help identify collisions by minute of the day

Hour: It can be used to identify what time of the day is more prone collisions.

Step 3: Target Variable

FATAL -Target Variable

- Observe the relationships between FATAL and other variables to find the insights.
- The FATAL variable is created by using the information in ACCLASS (Accident Classification) variable.

Check Appendix for detailed information

Sample of KSI Analytical File

Source	Source	Derived	Derived	Derived	Derived	Derived	Source	Source	Derived	Source	Derived	Source	Derived	Source	Source	Source	Source	Source	Source
ACCNUM	YEAR	MONTH	DAY	HOUR	MINUTE	WEEKDAY	LATITUDE	LONGITUDE	WardName	Ward Number	Neighbourhood	Hood_ID	Division	District	STREET1	STREET2	OFFSET	ROAD_CLASS	LOCCOOR
893184	2006	1	1	2	36	6	43.6996	-79.318797	Beaches-Eas	19	Woodbine-Lums	60	54	Toronto a	WOODBIN	O CONNOR DR		Major Arterial	Intersecti
893184	2006	1	1	2	36	6	43.6996	-79.318797	Beaches-Eas	19	Woodbine-Lums	60	54	Toronto a	WOODBIN	O CONNOR DR		Major Arterial	Intersecti
893184	2006	1	1	2	36	6	43.6996	-79.318797	Beaches-Eas	19	Woodbine-Lums	60	54	Toronto a	WOODBIN	O CONNOR DR		Major Arterial	Intersecti
893184	2006	1	1	2	36	6	43.6996	-79.318797	Beaches-Eas	19	Woodbine-Lums	60	54	Toronto a	WOODBIN	O CONNOR DR		Major Arterial	Intersecti
893184	2006	1	1	2	36	6	43.6996	-79.318797	Beaches-Eas	19	Woodbine-Lums	60	54	Toronto a	WOODBIN	O CONNOR DR		Major Arterial	Intersecti

Source	Source	Source	Source	Source	Source	Source	Source	Source	Source		Source	Source	Source	Source	Source	Source	Source	Source	Source
ACCLOC	TRAFFCTL	VISIBILITY	LIGHT	RDSFCONI	ACCLASS	IMPACTYP	INVTYPE	INVAGE	INJURY	FATAL_NO	INITDIR	VEHTYPE	MANOEUV	DRIVACT	DRIVCONI	PEDTYPE	PEDACT	PEDCOND	CYCLISTYP
Intersection	No Contro	Clear	Dark	Wet	Non-Fatal	Approachi	Passenger	50 to 54	Major										
Intersection	No Contro	Clear	Dark	Wet	Non-Fatal	Approachi	Passenger	15 to 19	Minor										
Intersection	No Contro	Clear	Dark	Wet	Non-Fatal	Approachi	Driver	55 to 59	Minor		North	Automobi	Going Ahe	Driving Pr	Normal				
Intersection	No Contro	Clear	Dark	Wet	Non-Fatal	Approachi	Passenger	20 to 24	Minor										
Intersection	No Contro	Clear	Dark	Wet	Non-Fatal	Approachi	Passenger	15 to 19	Minor										

Source	Source	Derived	Derived	Derived	Derived	Derived	Derived	Derived	Derived	Derived	Derived	Derived	Derived	Derived	Target
CYCACT	CYCCOND	PEDESTRI	CYCLIST	AUTOMOE	MOTORCY	TRUCK	TRSN_CITY	EMERG_VI	PASSENGE	SPEEDING	AG_DRIV	REDLIGHT	ALCOHOL	DISABILITY	FATAL
		0	0	1	0	0	0	0	1	1	1	0	1	0	
		0	0	1	0	0	0	0	1	1	1	0	1	0	
		0	0	1	0	0	0	0	1	1	1	0	1	0	
		0	0	1	0	0	0	0	1	1	1	0	1	0	
		0	0	1	0	0	0	0	1	1	1	0	1	0	

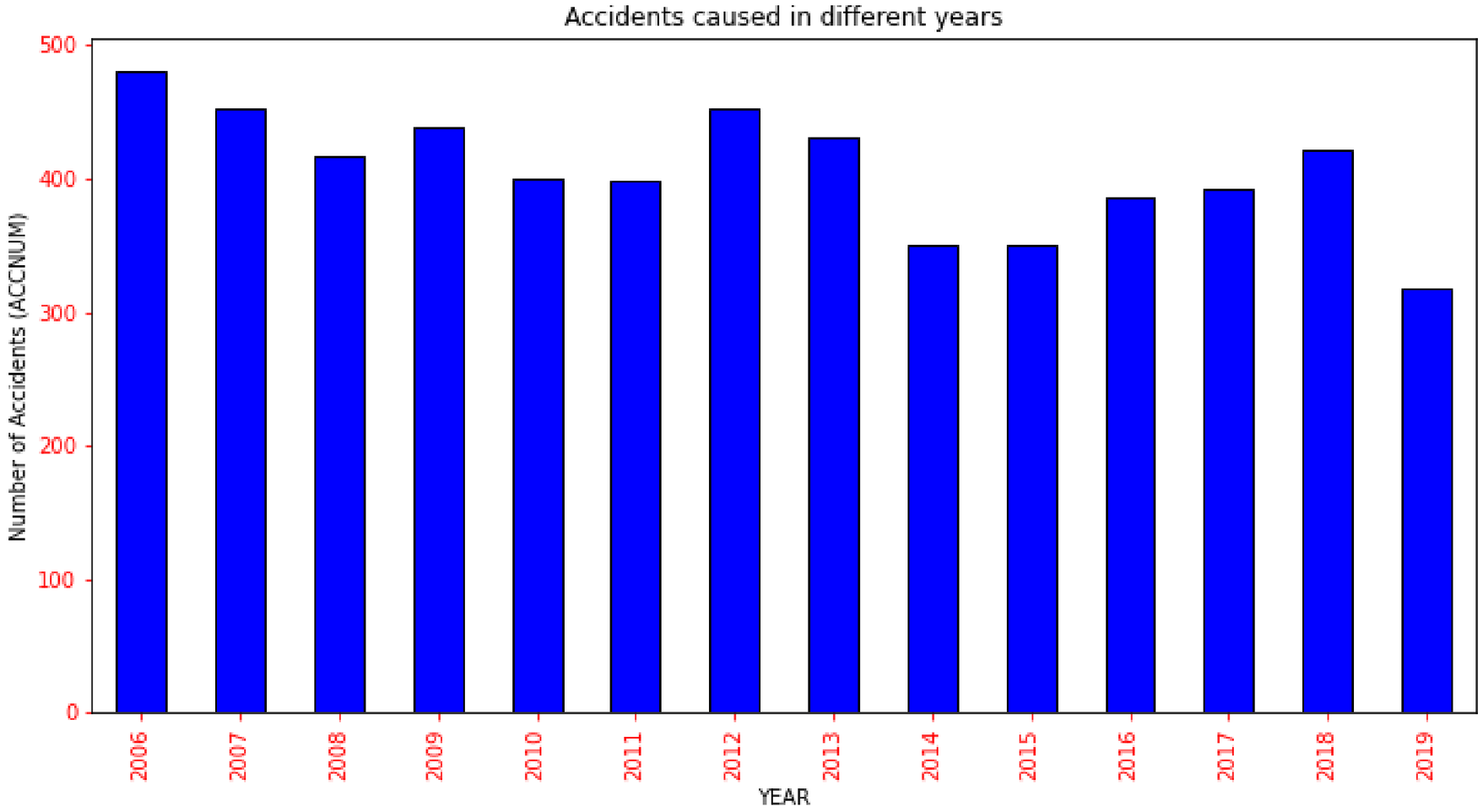
Tendencies In Fatal Collisions

Section 3
(Key Findings & Recommendations)

Collisions By Year

- Year with highest number of collisions: 2006
- Overall, the number of collisions have slightly declined after 2012.

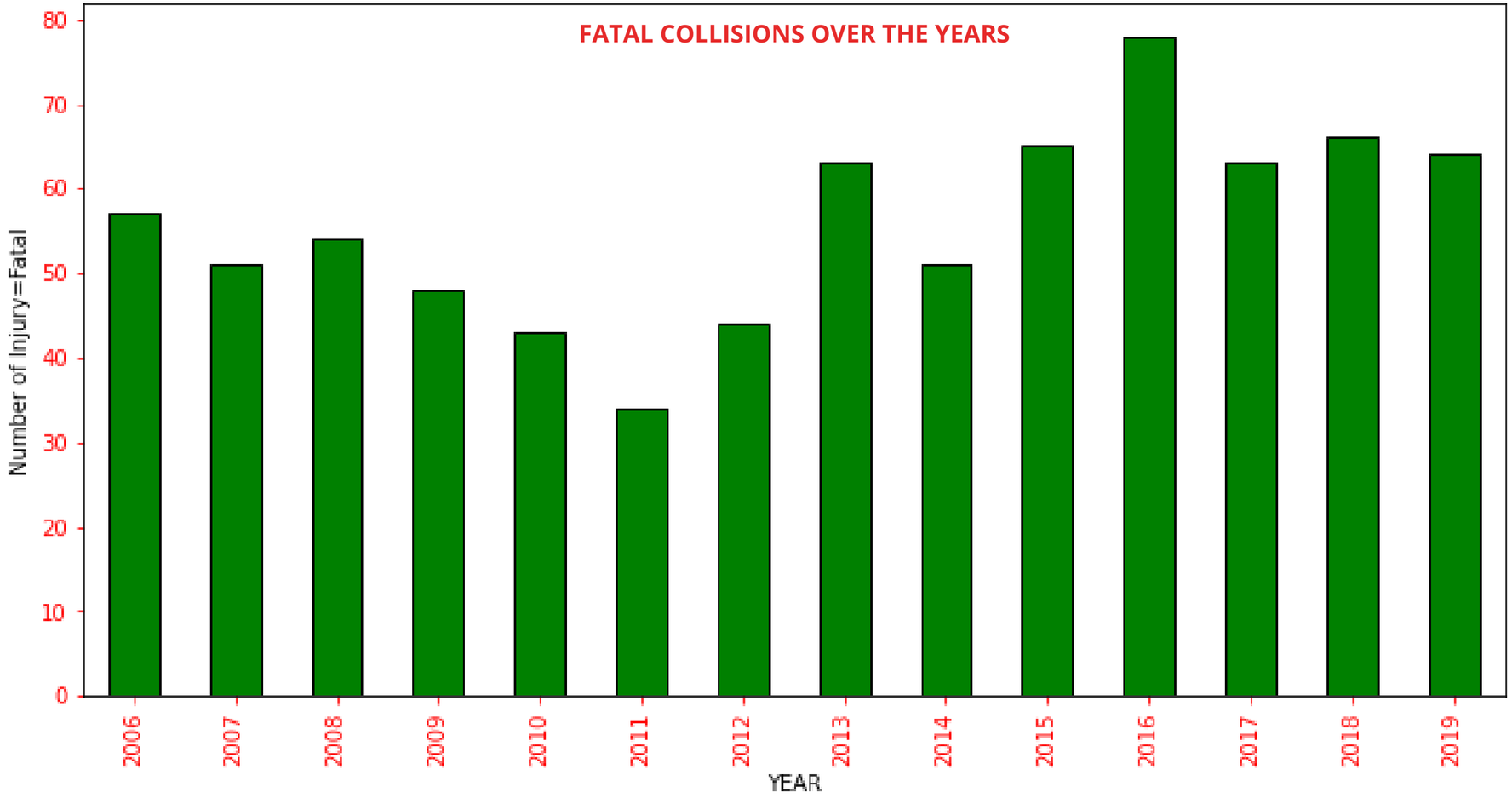
Sharp decline in collisions after 2012 is an indication that strategies adopted to reduce collisions are working well



Number of Fatal Collisions

- Year 2016 recorded the highest number of deaths due the collisions
- Fatalities declined from 2017-2019

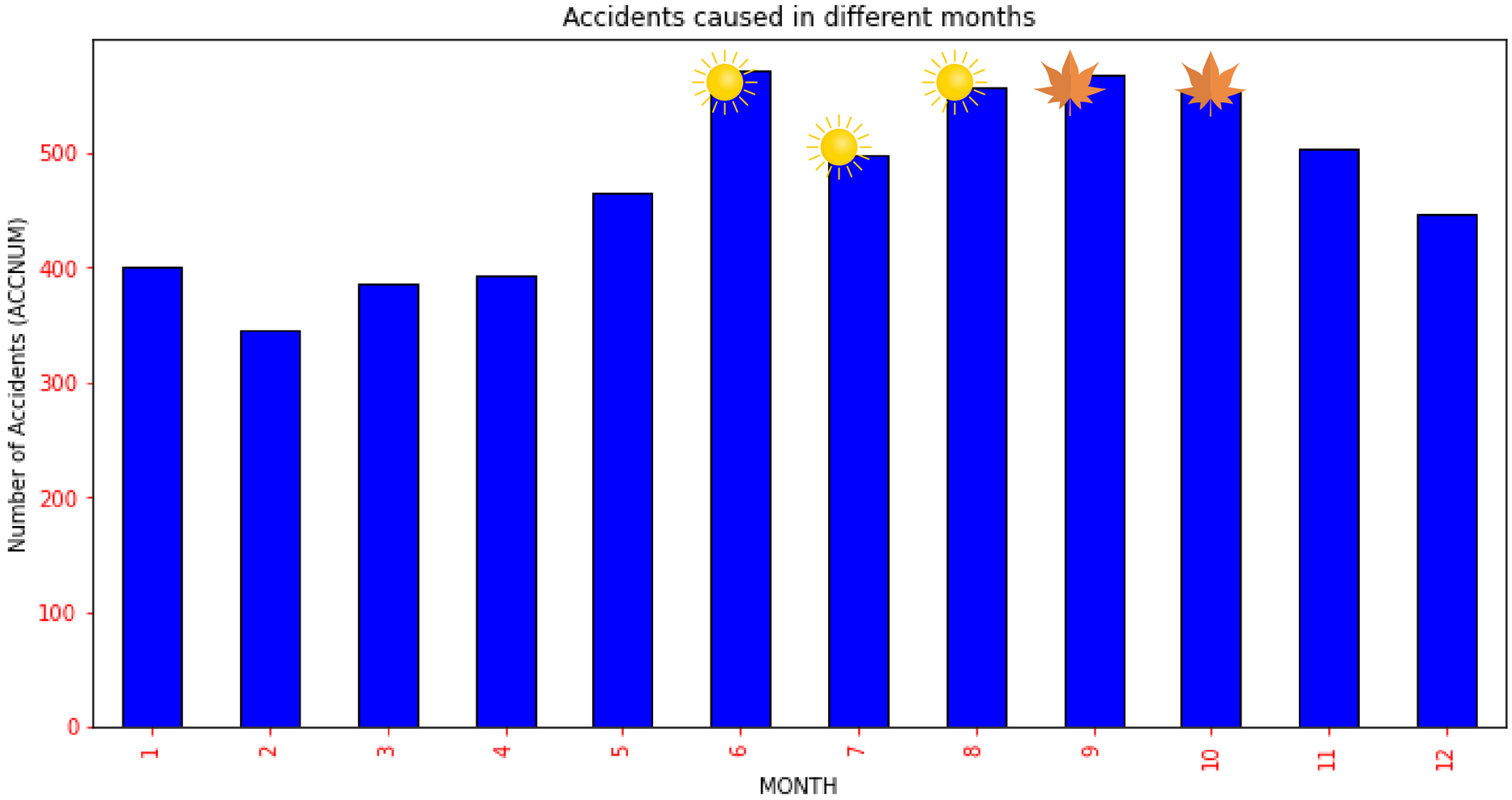
Decline in fatalities in 2017 shows that Vision Zero Safety Plan is working so Toronto Police should continue to enforce it



Collisions By Month

- Most collisions happen from June to October which are summer to early fall months

Toronto Police needs to exercise stricter traffic laws during summer and fall months to avoid fatalities



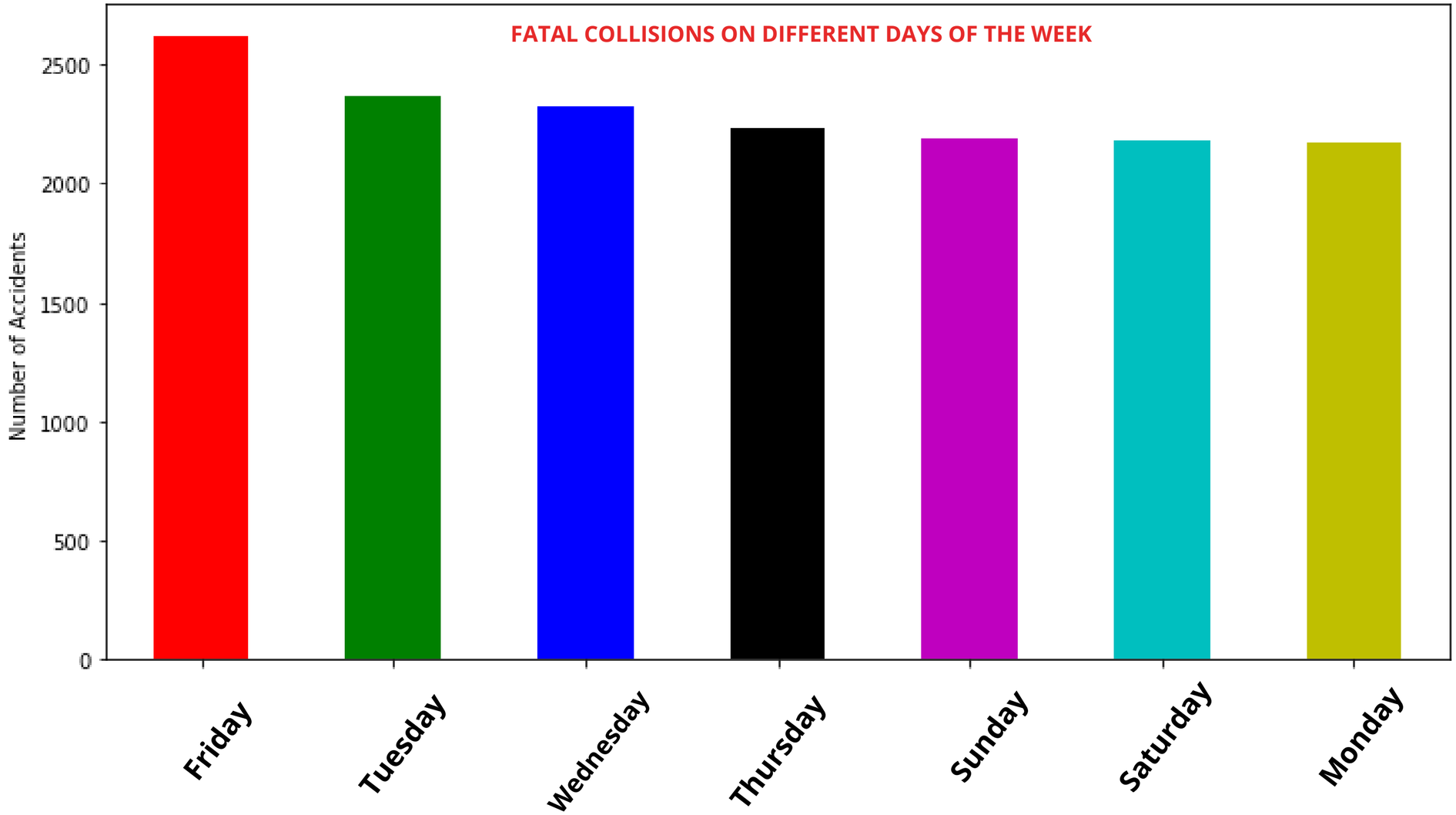
Since, summer and fall months have ideal weather conditions which means that more pedestrians, vehicles on the road and hence, more collisions

Collisions by Day of the Week

- Highest number of collisions:
  Friday

- Lowest number of collisions:
  Monday
  Saturday
  Sunday

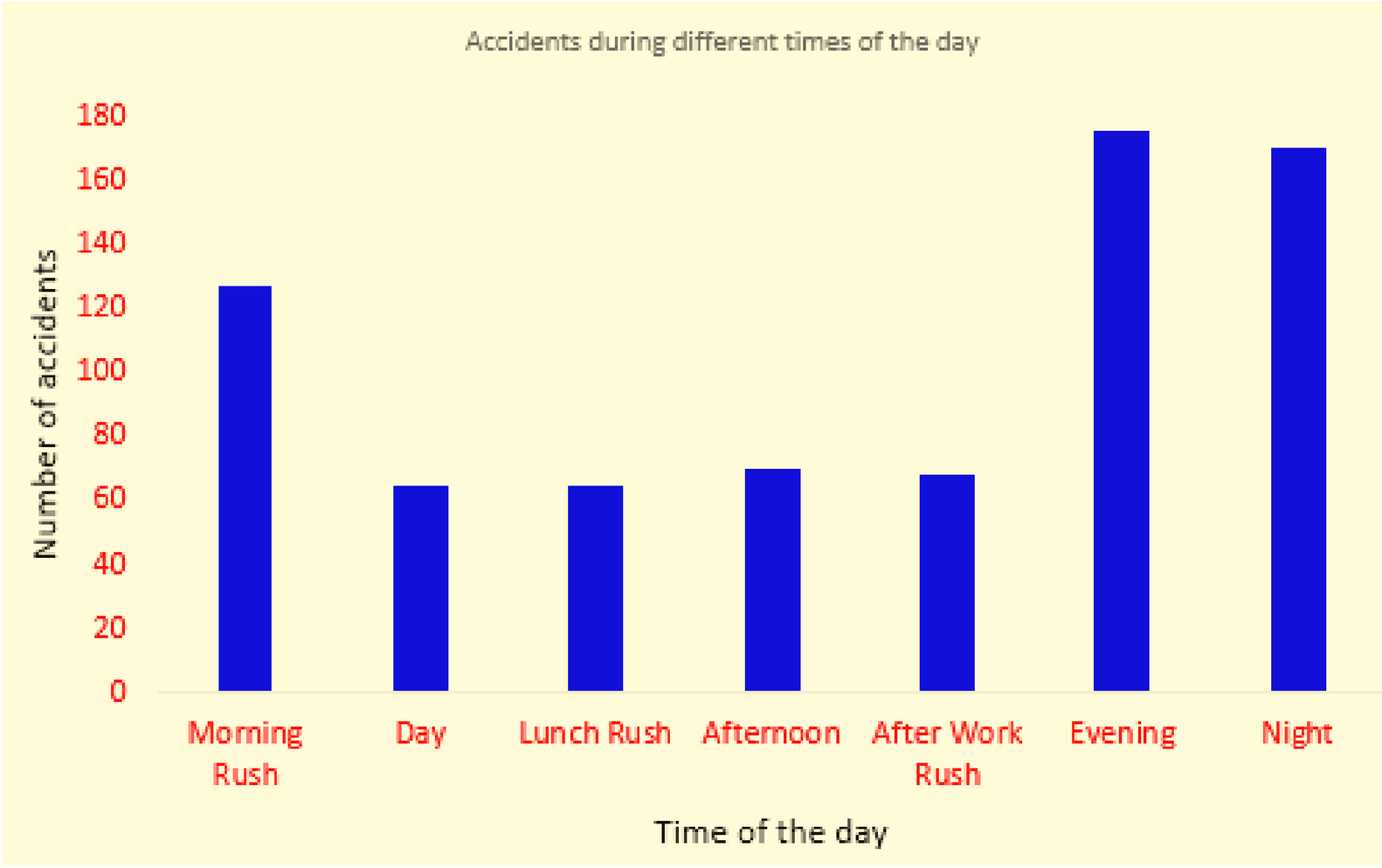
More police patrolling on Friday



Collisions By Time Of The Day

- Most collisions occur during following time of the day:
  Evening
  Night
  Morning Rush Hour

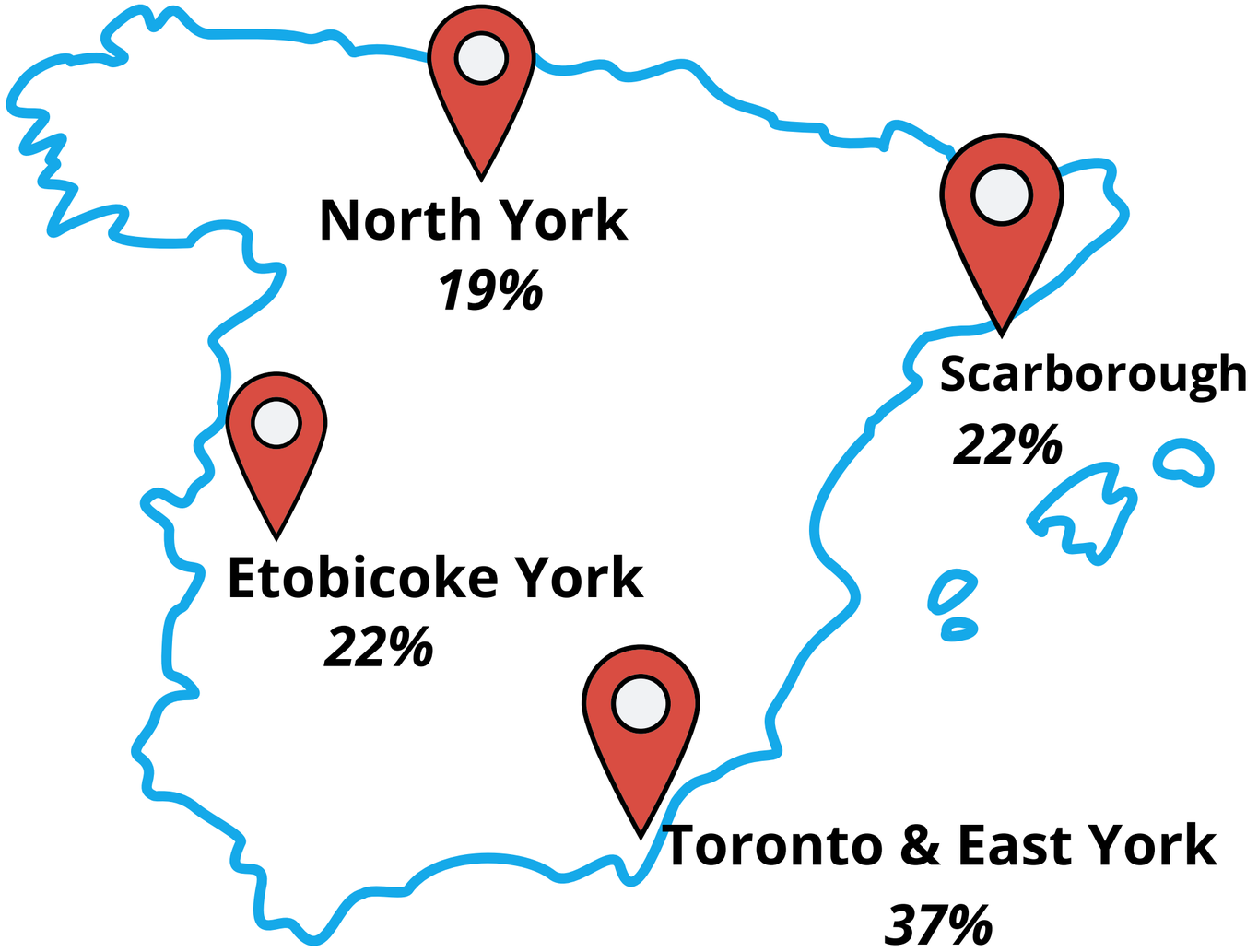
TPS should focus more on evening, night and morning rush hours as they are also busy from work commute perspective



District Wise Fatal Collision

- Most Collisions: Toronto & East York
  Least Collisions: North York
- Observation: Lack of Traffic Controls such as traffic signal, stop or yield sign is common in Etobicoke York, Toronto & East York and Scarborough
- North York is the only exception where there is some form of traffic controls (such as traffic signal)

Local authority can consider installing traffic controls in Toronto & East York, Etobicoke & Scarborough along with more police in these districts



Data might be skewed

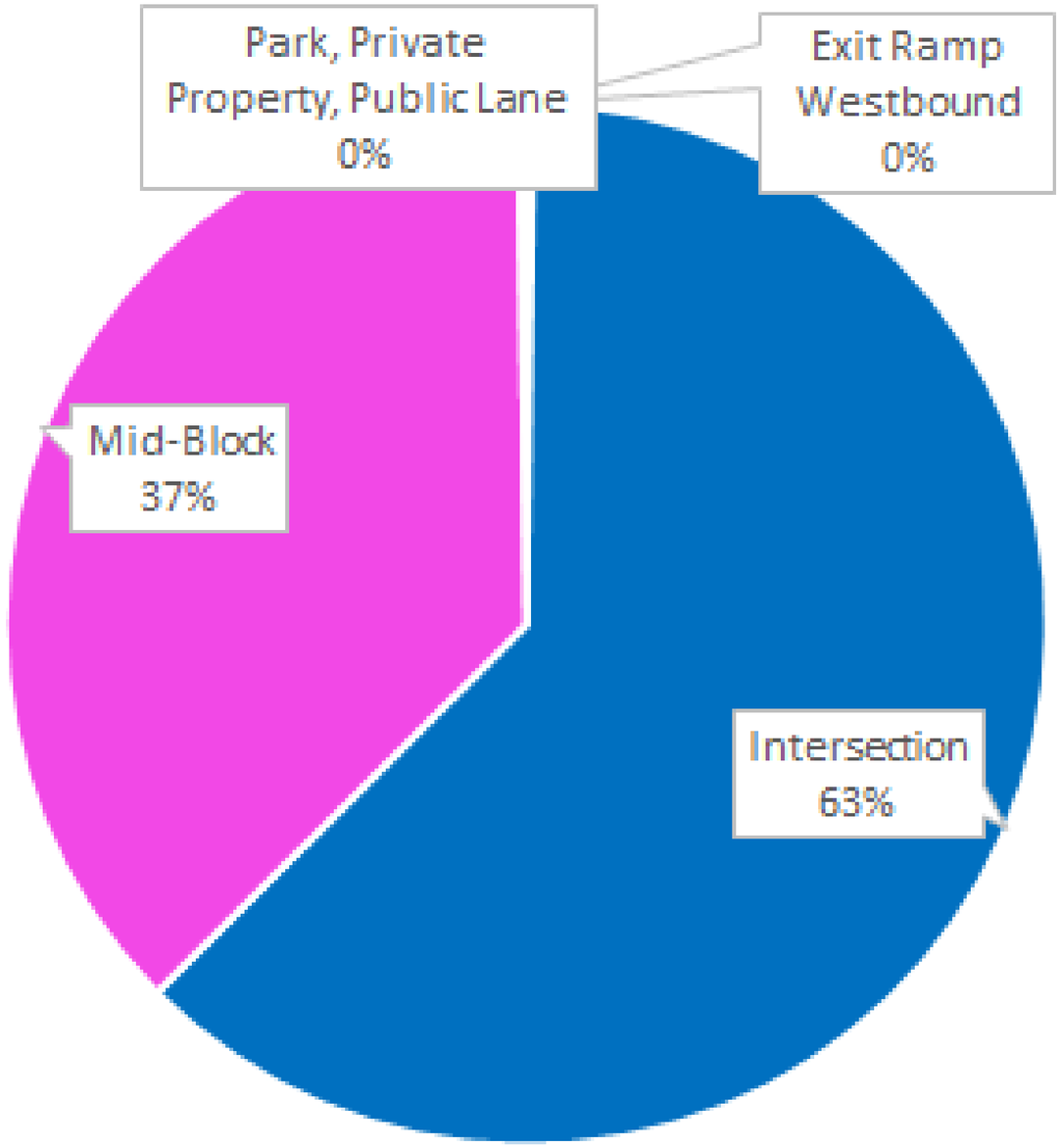
StatCan Data 2016 Population numbers are Scarborough (632,098), Toronto & East York (118,071), North York (869,401), Etobicoke York (365,143)

Location or spot wise Fatal Collisions

Four locations where collisions take place:
(ranked from highest to lowest)

1. Intersection (63%)
2. Mid-Block (37%)
3. Exit Ramp Westbound
4. Park, Private Property, Public Lane

Install cameras or police officers at intersections and mid-blocks to review the cause of fatal collisions at these spots and try to minimize fatality.



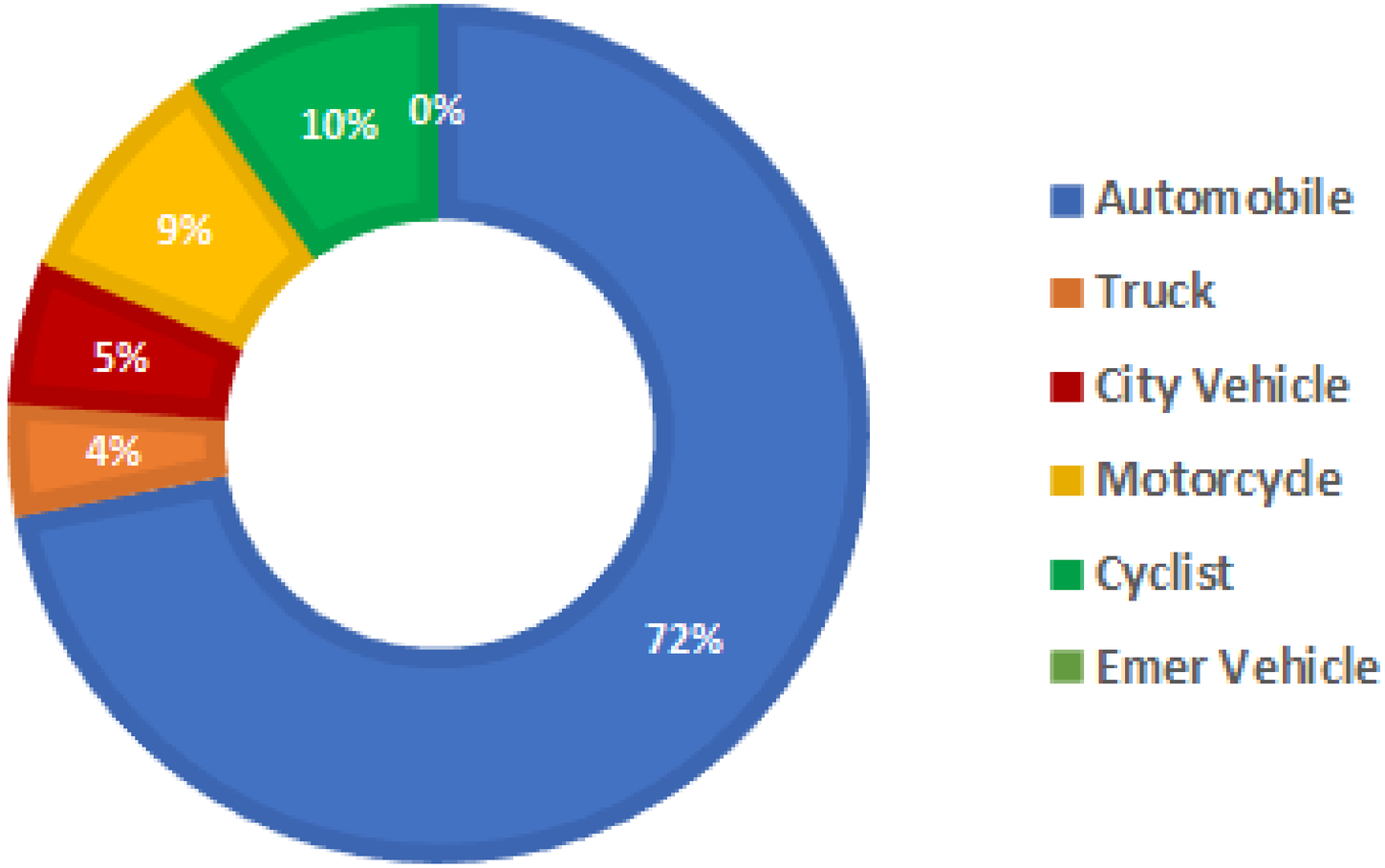
Vehicle Type Involved In Collision

- Most Collision Prone: Automobiles, Cyclist, Motorcycle

- Least Collision Prone: Emergency Vehicles (The reason can be because of less number of emergency vehicles on road)

Make city transit more attractive & convenient as an option for people to commute as it involves only 5% fatalities when compared to automobiles (72%)

Stricter rules for automobiles

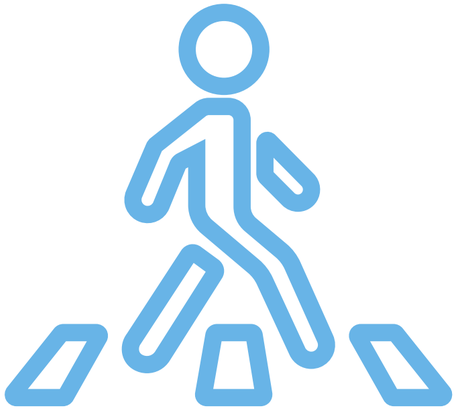


Victims of Collisions

- More Victims:
  Passengers and Pedestrians
  (approx. 43 %each)

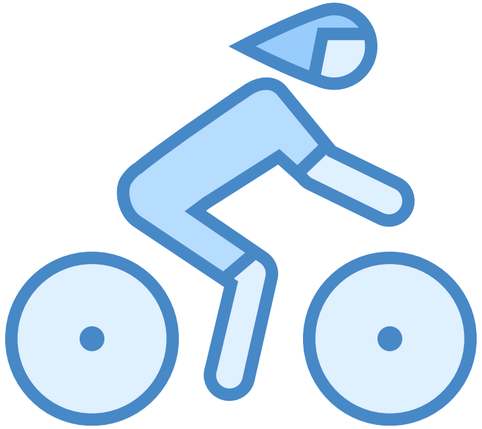
- Less Victims:
  Cyclists (13%)

Passenger- 43.6%



Pedestrian- 43.3%

Toronto Police should focus on programs that can enhance safety for passengers, pedestrians & cyclists.



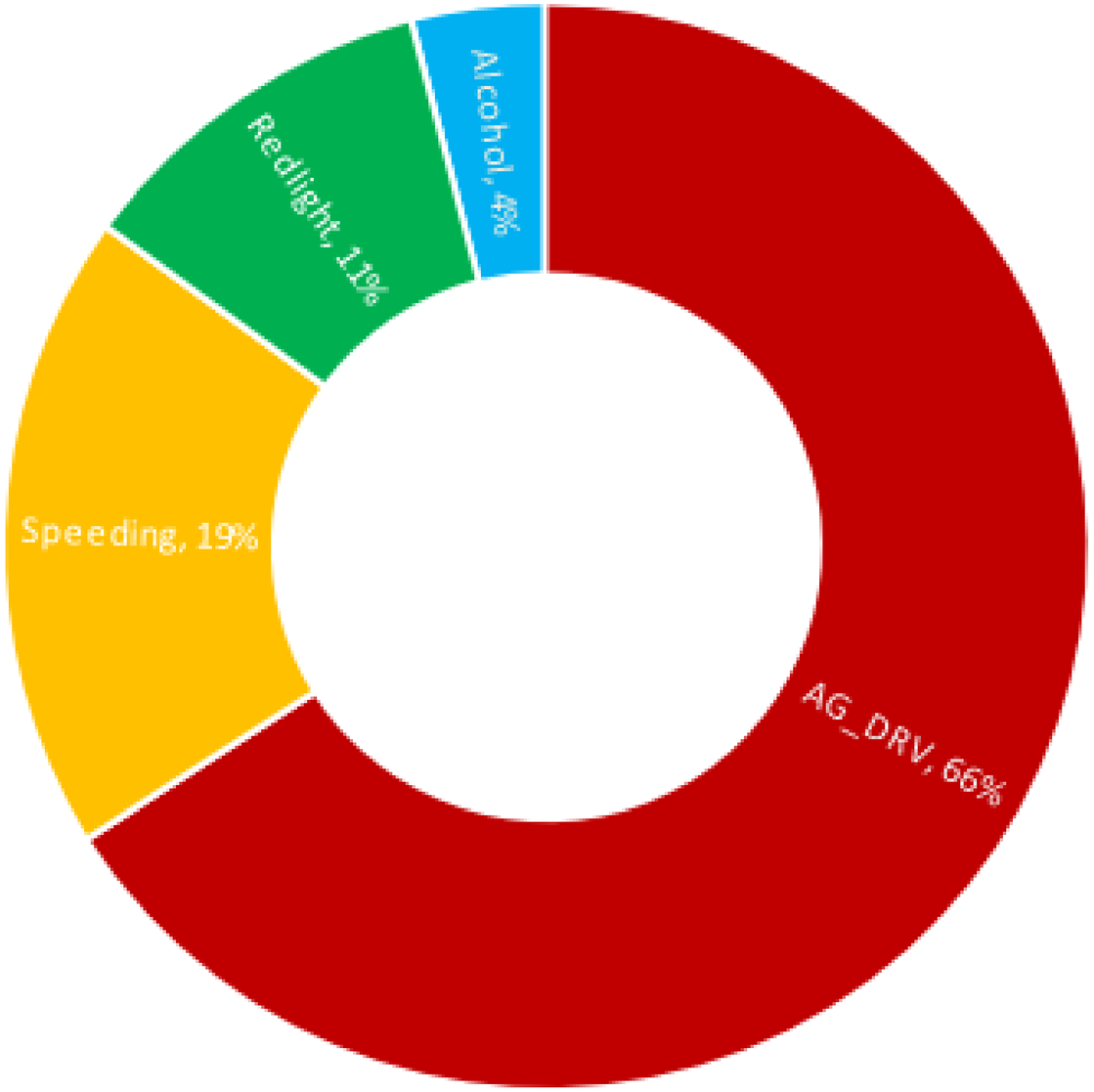
Cyclist- 13.2%

Causes of Fatal Collision

- There are 4 main reasons of fatal collision:
  1. Aggressive & Distracted Driving (66%)
  2. Speeding related collisions (19%)
  3. Red Light related collisions (11%)
  4. Alcohol related collisions (4%)

- Aggressive and Distracted Driving is a major reason resulting in fatal collisions.

Toronto Police should prioritize controlling the aggressive and distracted driving



Summary of Recommendations

1. **Crucial Months:** Toronto Police needs to enforce stricter traffic rules during summer and fall months.

2. **Crucial Day:** More police force on Fridays.

3. **Time of the Day:** Morning rush hours,night and evenings have more people commuting because of work so more patrolling during these hours.

4. **Districts:** Local authority should focus on installing traffic controls in Etobicoke, Scarborough and Toronto & East York as most collisions took place at spots where there was no form of control. Also, stricter law enforcement is required in these districts.

Summary of Recommendations (contd.)

5. **Location:** Install cameras or traffic police at intersections and mid-block to analyze cause of collision at these spots and reduce fatalities

6. **Vehicle Type:** Make city transit more attractive & convenient as an option for people to commute as it involves only 5% fatalities when compared to automobiles (72%). Also,strict rules for automobiles to reduce fatalities

7. **Victims:** Toronto Police should focus on programs that can enhance safety for passengers and pedestrians & cyclists.

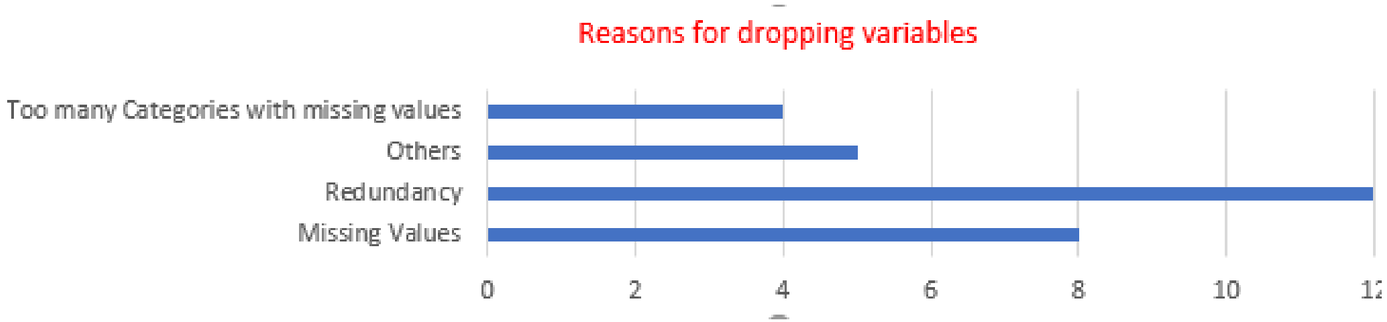
8. **Habit:** Toronto Police should prioritize controlling the aggressive and distracted driving

Predictive Modeling

Which variables have the most contribution in fatal collisions?

Section 4
(Modeling & Recommendations)

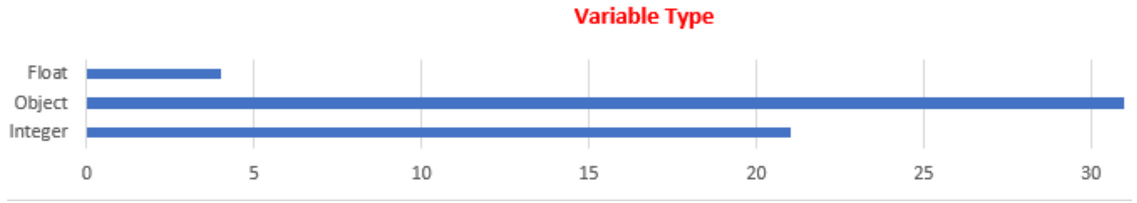
Initial Steps in Model Building



4 **Drop Variables:**
>80% Missing Values
Redundant Variables
Too many categories

Object → **Category**

3 **Convert variable 'object' type into 'Category'**



2 **Explore variable types**

1 **Upload data set in Python**

Short listed Variables for Model Building

52 Variables have been chosen for fiting a Logistic Regression Model

```
'YEAR', 'MONTH', 'DAY', 'HOUR', 'MINUTE', 'WEEKDAY', 'LATITUDE',
'LONGITUDE', 'Hood_ID', 'PEDESTRIAN', 'CYCLIST', 'AUTOMOBILE',
'MOTORCYCLE', 'TRUCK', 'TRSN_CITY_VEH', 'EMERG_VEH', 'PASSENGER',
'SPEEDING', 'AG_DRIV', 'REDLIGHT', 'ALCOHOL', 'DISABILITY', 'FATAL',
'VISIBILITY_Clear', 'VISIBILITY_Drifting Snow',
'VISIBILITY_Fog, Mist, Smoke, Dust', 'VISIBILITY_Freezing Rain',
'VISIBILITY_Other', 'VISIBILITY_Rain', 'VISIBILITY_Snow',
'VISIBILITY_Strong wind', 'RDSFCOND_Dry', 'RDSFCOND_Ice',
'RDSFCOND_Loose Sand or Gravel', 'RDSFCOND_Loose Snow',
'RDSFCOND_Other', 'RDSFCOND_Packed Snow', 'RDSFCOND_Slush',
'RDSFCOND_Spilled liquid', 'RDSFCOND_Wet', 'LIGHT_Dark',
'LIGHT_Dark, artificial', 'LIGHT_Dawn', 'LIGHT_Dawn, artificial',
'LIGHT_Daylight', 'LIGHT_Daylight, artificial', 'LIGHT_Dusk',
'LIGHT_Dusk, artificial', 'LIGHT_Other', 'DISTRICT_Etobicoke York',
'DISTRICT_North York', 'DISTRICT_Scarborough',
'DISTRICT_Toronto and East York'],
```

Forward Stepwise Regression Results

Forward Stepwise Regression gives us 29 variables responsible for fatal collisions

Important Variables are:

Truck
Pedestrian
Speeding
Toronto & East York District
Transit City Vehicle

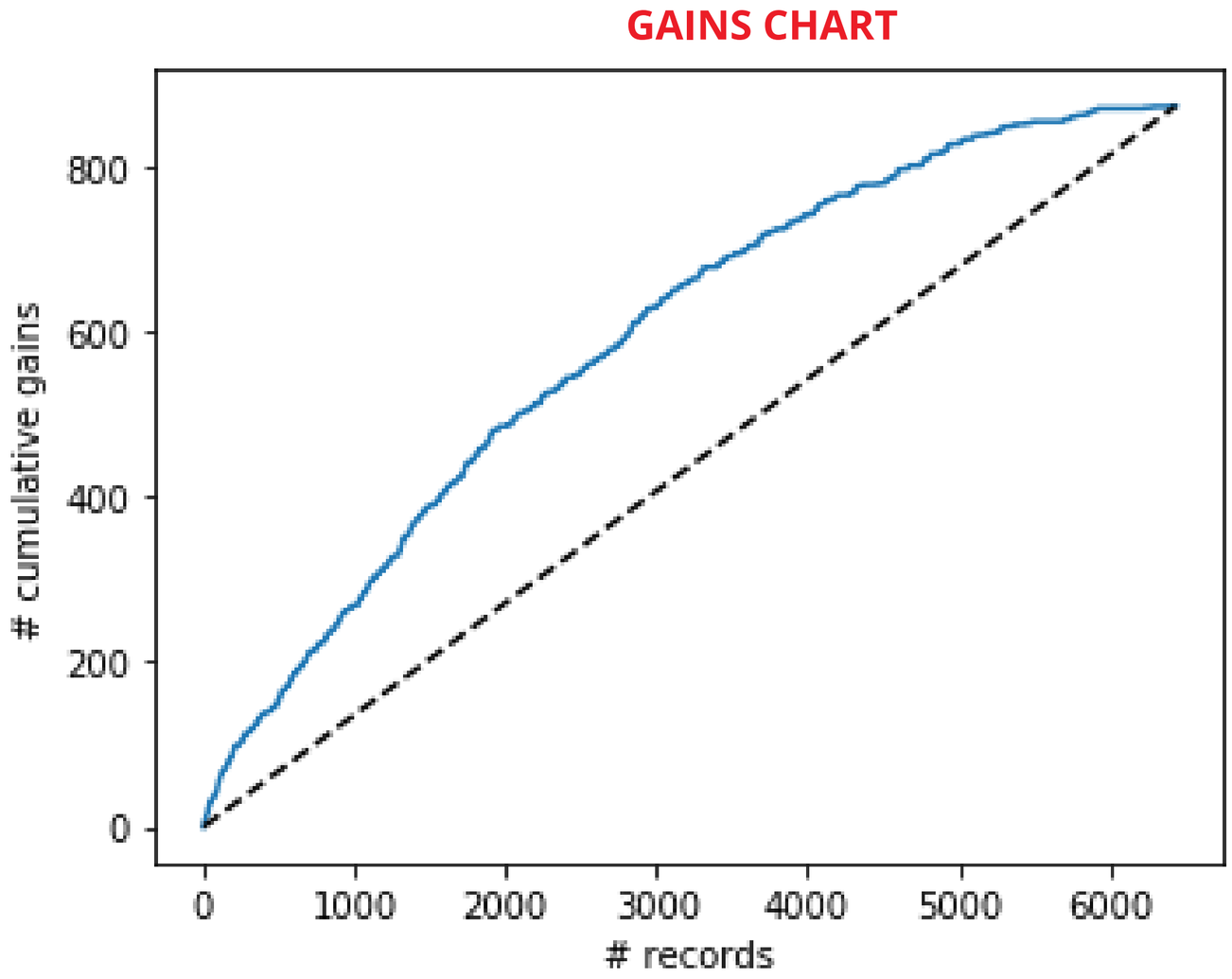
```
Start: score=6864.06, constant
Step: score=6739.91, add TRUCK
Step: score=6620.41, add PEDESTRIAN
Step: score=6500.53, add SPEEDING
Step: score=6432.52, add DISTRICT_Toronto and East York
Step: score=6375.27, add TRSN_CITY_VEH
Step: score=6329.21, add VISIBILITY_Other
Step: score=6298.14, add LIGHT_Dark
Step: score=6273.03, add AG_DRIV
Step: score=6253.09, add YEAR
Step: score=6242.67, add DISTRICT_Scarborough
Step: score=6233.13, add HOUR
Step: score=6223.60, add RDSFCOND_Ice
Step: score=6214.98, add RDSFCOND_Other
Step: score=6208.44, add REDLIGHT
Step: score=6204.23, add LIGHT_Dusk
Step: score=6199.84, add MOTORCYCLE
Step: score=6194.72, add PASSENGER
Step: score=6192.44, add VISIBILITY_Snow
Step: score=6190.33, add MINUTE
Step: score=6188.73, add DISABILITY
Step: score=6187.33, add VISIBILITY_Fog, Mist, Smoke, Dust
Step: score=6186.51, add LIGHT_Other
Step: score=6186.22, add LONGITUDE
Step: score=6182.72, add DISTRICT_Etobicoke York
Step: score=6182.62, add VISIBILITY_Rain
Step: score=6174.92, add VISIBILITY_Clear
Step: score=6166.62, add VISIBILITY_Freezing Rain
Step: score=6150.66, add VISIBILITY_Drifting Snow
Step: score=6130.17, add VISIBILITY_Strong wind
Step: score=6130.17, unchanged None
```

Results of Logistic Regression Model- Gains Chart

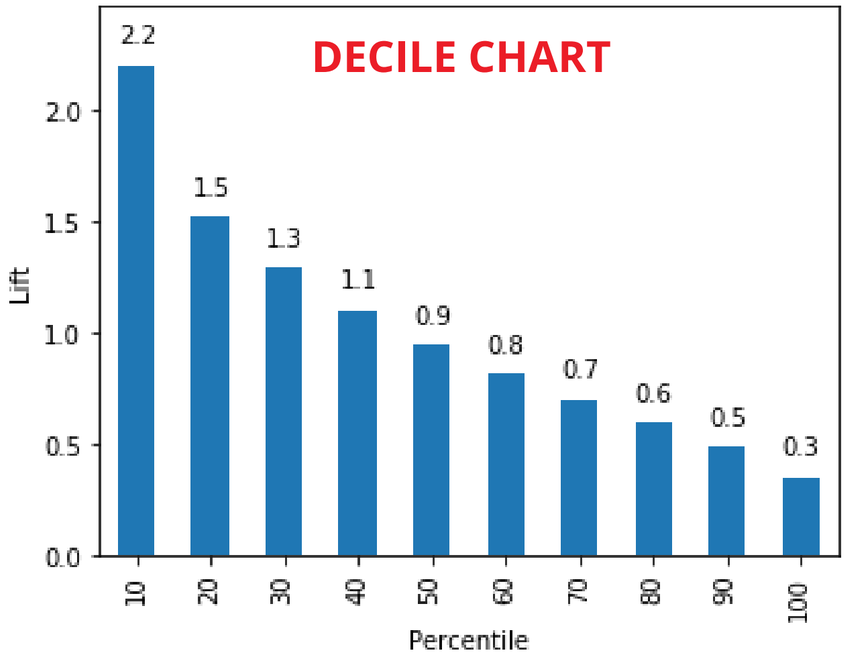
- For any given number of records (the x-axis value), The Gains Chart represents the expected number of fatal collisions we would predict if we did not have a model but simply selected cases at random.

- The dotted line provides a benchmark against which we can see performance of the model.

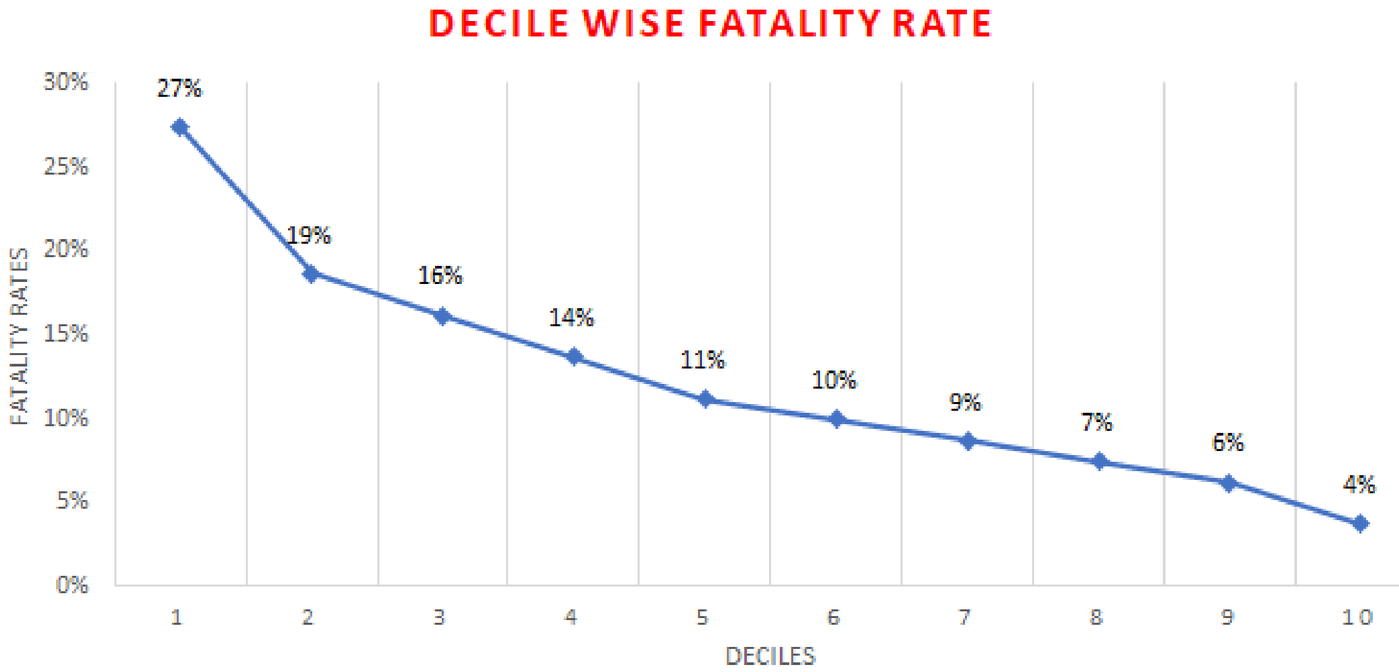
- For 3000 records, our model predicts close to 600 fatal collisions whereas randomly selected cases would have predicted less than 400 fatal collisions.



Results of Logistic Regression Model- Decile Chart



- The bars show the factor by which our model outperforms a random assignment of 0's and 1's.
- For the 1st decile or top 10%, we have close to 650 records and having no model gives us 80 fatal records whereas our model gives us 176 (2.2 times) fatal records.
- For 20%, we have 1300 records and in the absence of model, we have close to 200 fatal collisions whereas our model gives us little less than 400 fatal collisions.



- The graph shows the fatality rates for each decile starting from 1 to 10
- Fatality Rate highest for 1st decile- 27%
- Fatality Rate lowest for 10th decile- 4%

Interpreting the Confusion Matrix

Validation Set

Confusion Matrix (Accuracy 0.8329)

Actual	Prediction 0	Prediction 1
0	5203	362
1	714	159

Training Set

Confusion Matrix (Accuracy 0.8387

Actual	Prediction 0	Prediction 1
0	7832	488
1	1069	266

Conclusion:

Accuracy of 83% in both cases proves that the Logistic Regression model can be relied upon for predictive analysis

EDA's for Key Variables

Truck_involved	number_of_records	fatal_rate
1	990	30%
0	15103	13%

Out of 990 Truck Related Collisions, close to 300 are fatal, which is 30%

Pedestrian_involved	number_of_records	fatal_rate
1	6484	18%
0	9609	11%

Out of 6484 Pedestrian Related Collisions, close to 1170 are fatal, which is 18%

SPEEDING_involved	number_of_records	fatal_rate
1	2157	22%
0	13936	12%

Out of 2157 Speeding Related Collisions, 475 are fatal, which is 22%

City_Transport_Vehicle	number_of_records	fatal_rate
1	982	22%
0	15111	13%

Out of 982 Transit Related Collisions, 216 are fatal, which is 22%

District	number_of_records	fatal_rate
Toronto East York	5632	16%
Scarborough	3599	11%

Out of 5632 collisions in Toronto & East York District, 901 are fatal, which is 16%

Out of 3599 collisions in Scarborough District, 396 are fatal, which is 11%

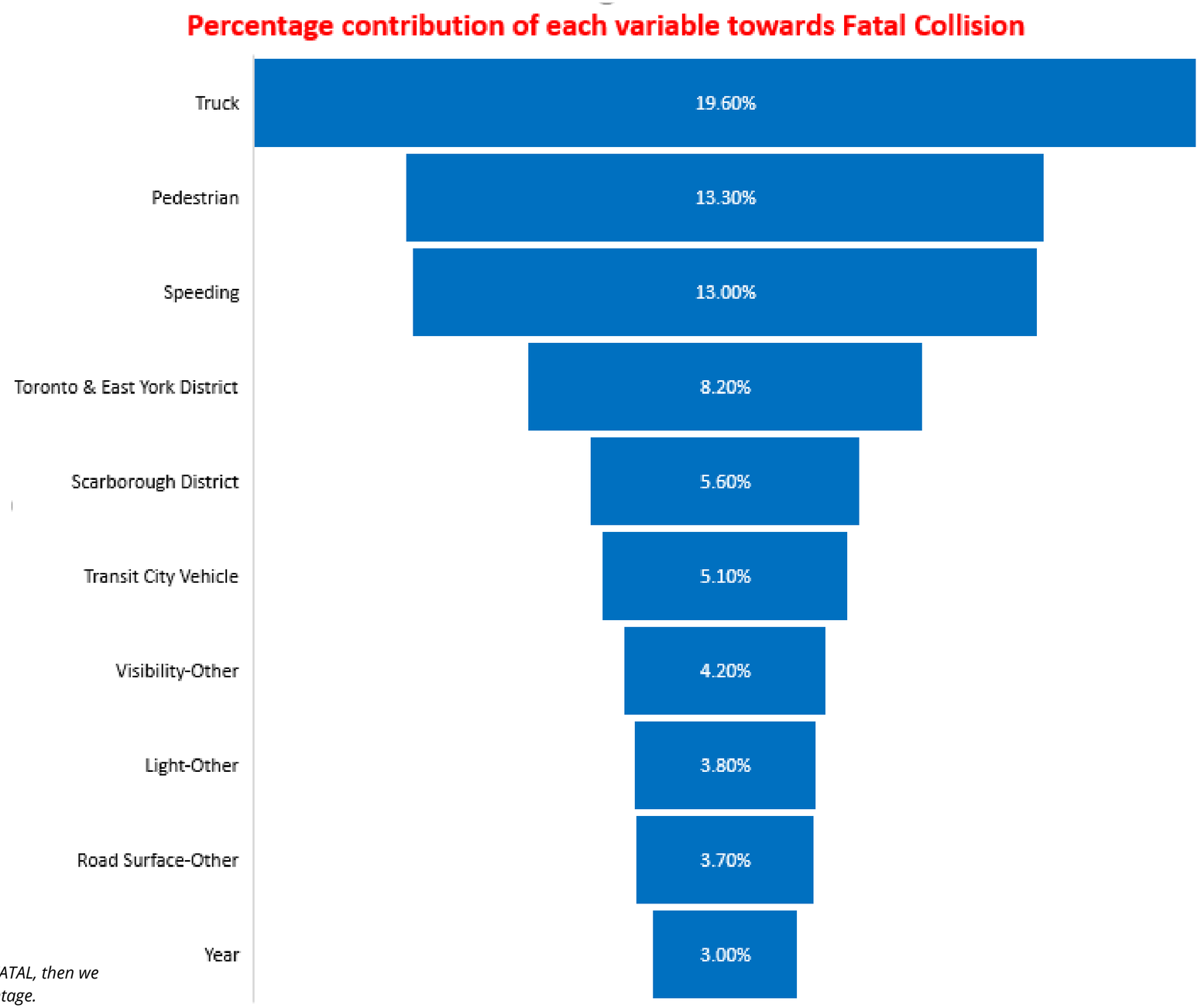
Key Contributors towards Fatal Collisions

Top 10 Variables:

Truck (+)
Pedestrian (+)
Speeding (+)
Toronto & East York District (-)
Scarborough District (+)
Transit City Vehicle (+)
VIsibility-Other (+)
Light-Other (+)
Road Surface-Other (+)
Year (+)

+ : Positive correlation between Fatal (target) and predictors

- : Negative correlation



Note: We found the partial R Squares by finding the squares of correlation of each variable to our target FATAL, then we divided partial R squares by adjusted R sqaure to get the contribution and then converting them to percentage.

Recommendations based on Predictive Modeling

1. **Vehicle Involved:** Stricter rules to be enforced for Trucks and Transit City Vehicles to reduce fatalities related to them

2. District: More police force can be deployed in Toronto & East York and Scarborough to minimize fatal collisions.

3. Pedestrian: Safety for pedestrians can be enhanced by fines on violation of rules such as illegal crossing

4. Speeding: Toronto Police should practice zero tolerance policy for vehicles caught exceeding the speed limits.

5. Other Conditions: Toronto Police Force in collaboration with concerned authorities should work on reducing road surface, light conditions and visibility related fatalities.

Conclusion

Vision Zero Program aimed at bringing the number of fatalities down to near zero in Toronto is surely challenging but possible.

The recommendations based on patterns in the data such as deploying more police force during summer and fall months, on Fridays, evening and morning rush hour, with focus on intersections and mid-blocks will help in reducing the number of fatalities in Toronto.

Toronto Police Force can leverage the variables from the model and look for strategies to reduce Truck, Pedestrian, Speeding, Toronto & East York District etc. related fatalities.

The key findings and recommendations from patterns and predictive modeling pave way for making all road users in Toronto safer and establish the city as one of the safest for any form of commute.

THANK YOU

Questions?

Prepared By: Batuhan, Esha, Hitesh, Phu, Reshma

Appendix

Metadata

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

Data Auditing-Methodology

Q1. Data Load or sample extract of each file you are working with

Answer – We are working on Toronto Police fatal collision data which is provided in CSV format. Moreover, we loaded data into python without any issue because the given data is nicely formatted which takes care of missing values itself. We did not need to work with that perspective.

We extracted the 100 random values from data into excel file with proper formatting (attached excel sheet).

Q2. Meta data or data diagnostics report for each file or tablet.

Answer – The data diagnosis report tells us about the variables how many **number of records** they consist of, **Data Field Format, Number of unique values, and Number of missing values** for each variable. In total, we have 13 Numeric variables, 40 Character variables, and 1 Date variable. (The data diagnostics and meta data report are attached in excel file).

Q3. Frequency distribution reports for each field on each file.

Answer - We run a frequency analysis on categorical variables to find out insights within them.

Q4. Summary of key findings and insights from data audit and next steps in terms of creating the analytical file.

Answer – We already have unique key in our data. So, we do not need to work on making matchkeys. In addition to this, some of our variables have Yes-NA values but we can assume that NAs represent **No** in this context. We have found this by looking at the other variables to figure out what is the pattern for missing values. Some of the variables mostly have unique values such as Street Names. There are only 1 or 2 values that catch the eyes in terms of frequency. We assume these are where traffic and people are located more than other locations. In the data, it looks like most of the collisions occurred when the weather was clear and 6 pm and 8 pm are the most common incident hours.

Provide examples of 5 source variables that might be useful in your analysis and why they might be useful. If there are no source variables or less than 5 source variables, provide some rationale as to why this exists.

1. **YEAR**: It gives a sense of time and we can count the number of accidents that happened in different years and compare them.
2. **ACCNUM**: Is the accident number. We use it to identify the accidents.
3. **IMPACTYPE:** It can give information about the impact types such as pedestrian collisions, SMV other, approaching, cyclist movement.
4. **INVTYPE**: It gives us information about the type of commuter whether it was automobile driver, motorcyclist, pedestrian etc. and it can also tell us about the type of vehicle involved.
5. **LATITUDE** and **LONGITUDE:** It can help identify the location and create a map of accidents.

Indicate your target variables and their relevancy in your analysis

Because the objectives of our project are deriving insights and patterns among variables, building predictive analytics models and mapping hot-spot for fatal collisions in Toronto. Therefore, we think the FATAL variable is the target variable.

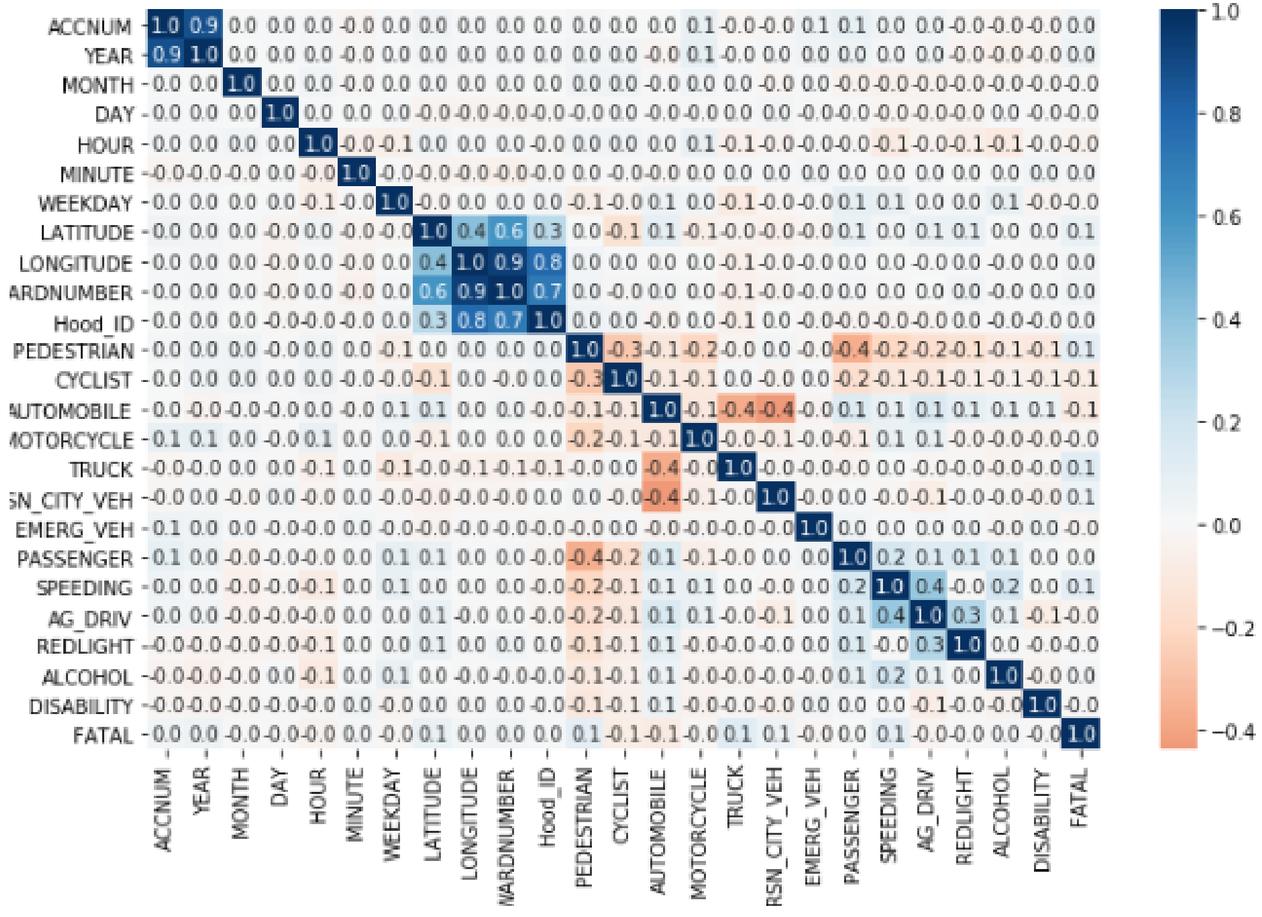
Provide the logic in how you will create the target variables

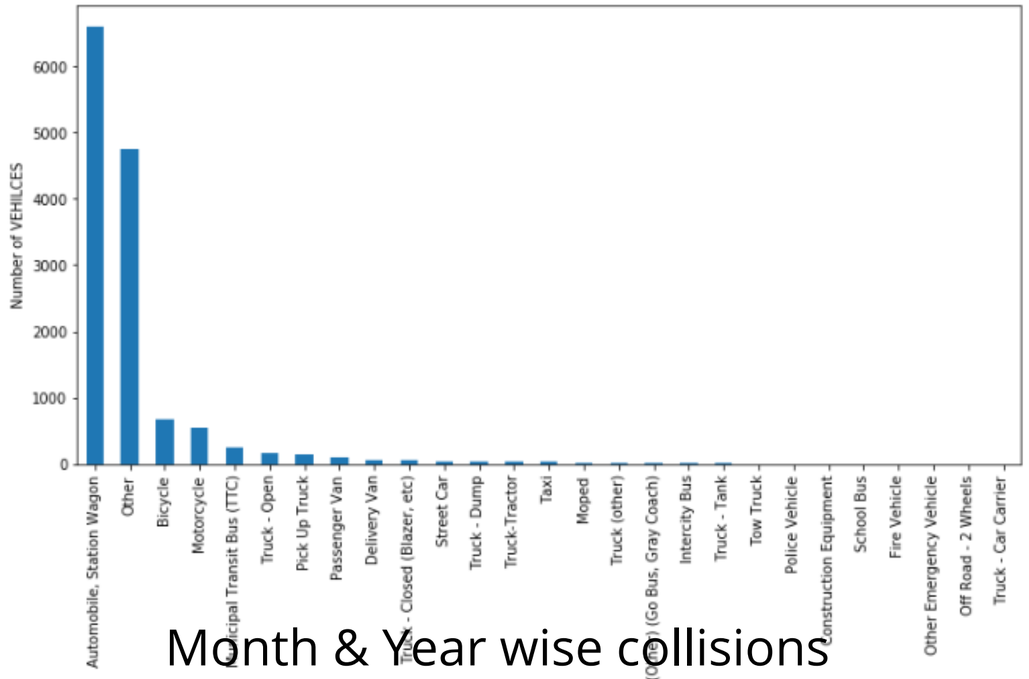
The FATAL variable is created by using the information in ACCLASS variables. If the accidents are classified as fatal, it will be 1 in FATAL columns and the rest are 0. Because the idea for creating FATAL is to differentiate fatal class collisions and non-fatal class collisions. So we can compare and develop a model to predict fatal collisions.

Provide examples of 15 derived variables that might be useful in your analysis and why they might be useful

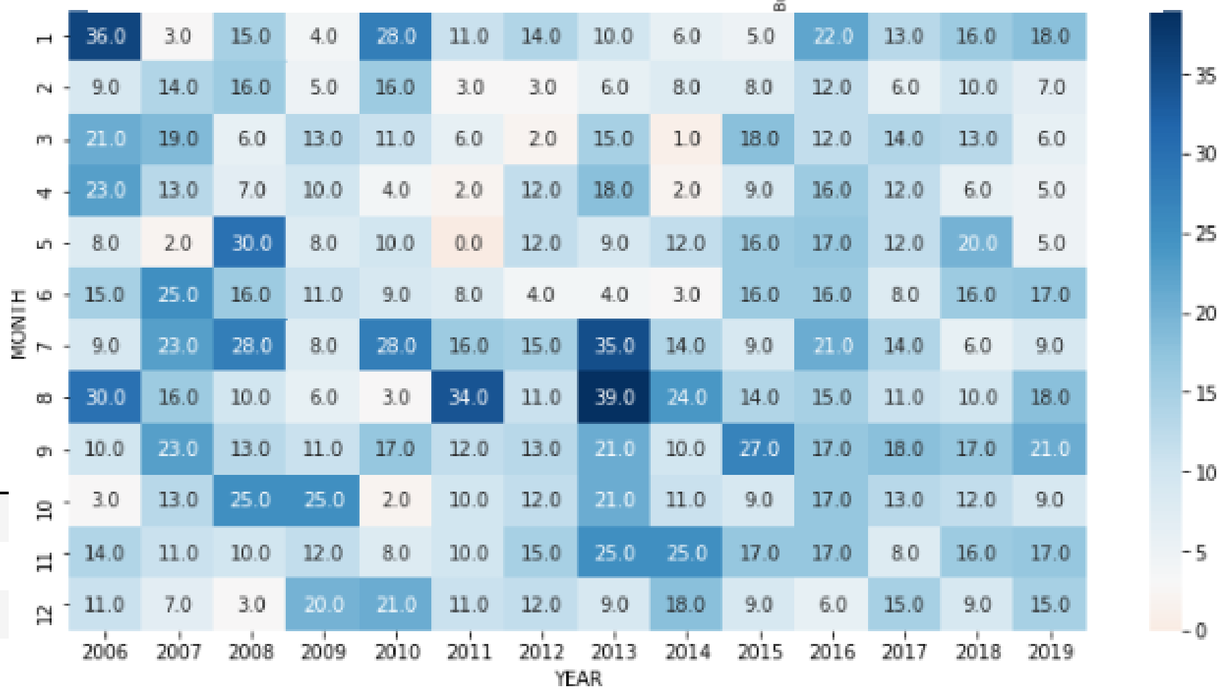
1. **MONTH:** It gives more details about which month the accident happen and we can compare accidents happening in different months of the same year or different years. This way we can predict or analyze the specific seasons, weather conditions which might be contributing in the collision rate.
2. **DAY:** It gives more details about which day of the week the accidents happen and the same as the month. We can compare between them in different months, years. Similarly like month, here we can have an inslight on which day the rates are higher. For example: Weekends.
3. **MINUTE:** It gives more details about when the accidents happen than HOUR.
4. **HOUR:** It gives more details about the time when the accidents occurred than DAY. By doing this identifying what time of the day its more prone to happen.
5. **WEEKDAY:** It gives more insights when the collisions happened during the week such as Monday, Friday or Weekend. Depending on the days of the week we can compare and find the patterns.
6. **WARDNAME:** There are 25 wards in Toronto since 14/08/2018. This variable can give information about which wards accidents happened we can compare and find the patterns between wards based on type, rate, frequency… of accidents.
7. **DIVISION:** It is the branch ID of Toronto police service for a designated area. We can understand the relationship between fatal rate, injury type under each division and how far the accidents from a particular division office.
8. **FATAL:** This variable can help identify which accident is fatal and non-fatal. So we can find out the relationship between variables and find the patterns.
9. **DISABILITY:** It can help identify if it is a medical or physical disability-related collision or not.
10. **REDLIGHT:** It can help identify if it is a red light related collision or not.
11. **ALCOHOL:** It can help identify if it is an alcohol-related collision or not.
12. **SPEEDING:** It can help identify if it is a speeding-related collision or not.
13. **AG_DRIV:** It can help identify if it is an aggressive and distracted driving condition or not.
14. **EMERG_VEH:** It can help to gain insights between the emergency vehicles involved and the fatal rate.
15. **TRSN_CITY_VEH:** It can help identify if transit or city vehicles are involved in collision or not.

Correlation Matrix of shortlisted variables





Month & Year wise collisions



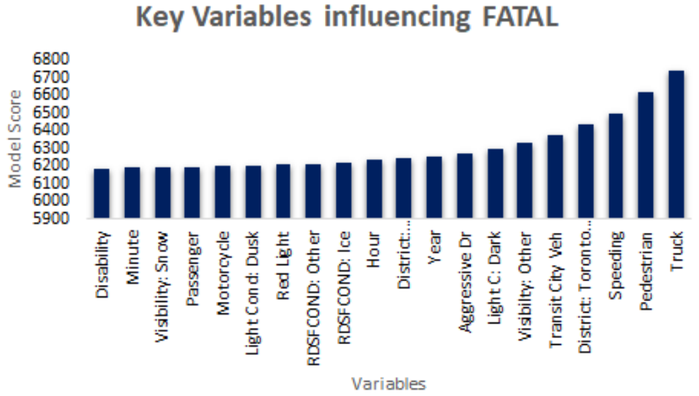
District wise collisions

YEAR	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
DISTRICT														
Etobicoke York	55	38	43	30	32	36	30	33	28	33	48	27	38	33
North York	34	29	57	28	16	35	33	35	40	39	36	24	41	46
Scarborough	71	59	32	34	43	22	38	71	32	44	51	52	45	50
Toronto and East York	29	43	47	41	66	30	24	73	34	41	53	41	27	16

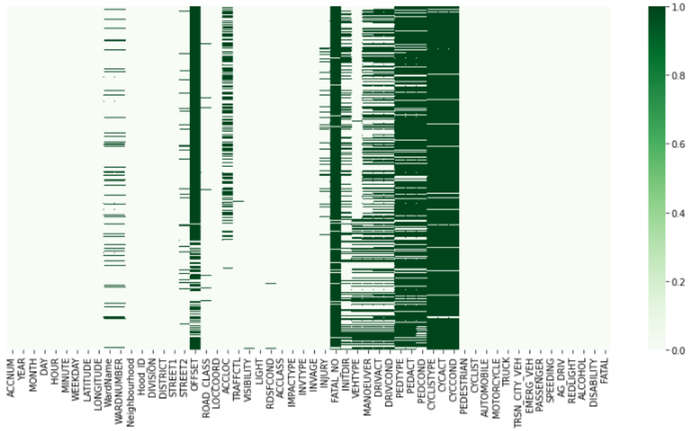
Stepwise Regression- Best Variables

```
# Finding partial R-squared for each varaible with respec
r=corr1**2
r.sort_values(ascending=False)
```

```
FATAL                                   1.000000e+00
TRUCK                                   1.485830e-02
PEDESTRIAN                              1.005900e-02
SPEEDING                                9.832765e-03
DISTRICT_Toronto and East York          6.186839e-03
DISTRICT_Scarborough                    4.238701e-03
TRSN_CITY_VEH                           3.852806e-03
VISIBILITY_Other                        3.180678e-03
LIGHT_Dark                              2.901754e-03
RDSFCOND_Other                          2.791418e-03
YEAR                                    2.266051e-03
HOUR                                    1.155188e-03
LIGHT_Other                             8.832393e-04
AG_DRIV                                 7.465469e-04
LONGITUDE                               6.274704e-04
MINUTE                                  5.511500e-04
RDSFCOND_Ice                            4.641257e-04
VISIBILITY_Drifting Snow                1.879675e-04
VISIBILITY_Freezing Rain                1.861443e-04
VISIBILITY_Freezing Rain                1.861443e-04
VISIBILITY_Snow                         1.510901e-04
MOTORCYCLE                              1.417051e-04
LIGHT_Dusk                              1.074063e-04
VISIBILITY_Strong wind                  7.909010e-05
VISIBILITY_Rain                         7.472220e-05
VISIBILITY_Clear                        4.069752e-05
VISIBILITY_Clear                        4.069752e-05
DISABILITY                              2.697492e-05
VISIBILITY_Fog, Mist, Smoke, Dust       1.109588e-05
PASSENGER                               5.502588e-06
REDLIGHT                                2.261324e-07
DISTRICT_Etobicoke York                 1.245596e-09
Name: FATAL, dtype: float64
```



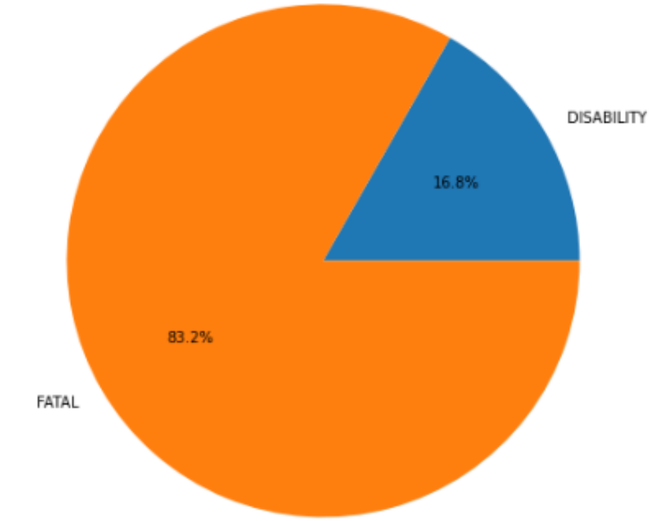

```
In [10]: fig, ax = plt.subplots(figsize=(15,7))
         #heatmap to visualize features with most missing values
         sns.heatmap(KSI_CLEAN.isnull(), yticklabels=False,cmap='Greens')
Out[10]: <matplotlib.axes._subplots.AxesSubplot at 0x29883117608>
```



```
#correlation of each variable with FATAL
corr1 = new_df.corr()['FATAL']
corr1
```

```
TRUCK                                    0.121895
YEAR                                     0.047603
PEDESTRIAN                               0.100295
SPEEDING                                 0.099160
DISTRICT_Toronto and East York          -0.078656
TRSN_CITY_VEH                            0.062071
VISIBILITY_Other                         0.056397
LIGHT_Dark                               0.053868
AG_DRIV                                 -0.027323
DISTRICT_Scarborough                     0.065105
HOUR                                    -0.033988
RDSFCOND_Ice                            -0.021544
RDSFCOND_Other                           0.052834
REDLIGHT                                -0.000476
LIGHT_Dusk                               0.010364
MOTORCYCLE                              -0.011904
PASSENGER                                0.002346
VISIBILITY_Snow                         -0.012292
MINUTE                                   0.023477
DISABILITY                              -0.005194
VISIBILITY_Fog, Mist, Smoke, Dust        0.003331
LIGHT_Other                              0.029719
LONGITUDE                                0.025049
DISTRICT_Etobicoke York                 -0.000035
VISIBILITY_Rain                         -0.008644
VISIBILITY_Clear                        -0.006379
VISIBILITY_Freezing Rain                -0.013643
VISIBILITY_Clear                        -0.006379
VISIBILITY_Freezing Rain                -0.013643
VISIBILITY_Drifting Snow                -0.013710
VISIBILITY_Strong wind                  -0.008893
FATAL                                    1.000000
```

	coef	std err	z	P>|z|	[0.025	0.975]
TRUCK	0.1785	0.014	12.434	0.000	0.150	0.207
YEAR	0.0026	0.001	3.005	0.003	0.001	0.004
PEDESTRIAN	0.1042	0.008	13.130	0.000	0.089	0.120
SPEEDING	0.1237	0.011	11.026	0.000	0.102	0.146
DISTRICT_Toronto and East York	-0.0582	0.010	-6.085	0.000	-0.077	-0.039
TRSN_CITY_VEH	0.1093	0.014	7.711	0.000	0.082	0.137
VISIBILITY_Other	-0.5590	0.106	-5.251	0.000	-0.768	-0.350
LIGHT_Dark	0.0598	0.009	6.672	0.000	0.042	0.077
AG_DRIV	-0.0459	0.008	-5.870	0.000	-0.061	-0.031
DISTRICT_Scarborough	0.0204	0.011	1.859	0.063	-0.001	0.042
HOUR	-0.0020	0.001	-3.559	0.000	-0.003	-0.001
RDSFCOND_Ice	-0.1156	0.052	-2.245	0.025	-0.217	-0.015
RDSFCOND_Other	0.1265	0.044	2.846	0.004	0.039	0.214
REDLIGHT	0.0370	0.013	2.849	0.004	0.012	0.062
LIGHT_Dusk	0.0730	0.029	2.507	0.012	0.016	0.130
MOTORCYCLE	0.0401	0.013	2.994	0.003	0.014	0.066
PASSENGER	0.0224	0.008	2.897	0.004	0.007	0.038
VISIBILITY_Snow	-0.7792	0.096	-8.145	0.000	-0.967	-0.592
MINUTE	0.0004	0.000	1.839	0.066	-2.35e-05	0.001
DISABILITY	0.0408	0.022	1.874	0.061	-0.002	0.083
VISIBILITY_Fog, Mist, Smoke, Dust	-0.5944	0.118	-5.050	0.000	-0.825	-0.364
LIGHT_Other	0.2941	0.172	1.706	0.088	-0.044	0.632
LONGITUDE	0.0563	0.022	2.529	0.011	0.013	0.100
DISTRICT_Etobicoke York	0.0010	0.011	0.095	0.924	-0.020	0.023
VISIBILITY_Rain	-0.7434	0.093	-7.969	0.000	-0.926	-0.561
VISIBILITY_Clear	-0.3642	0.046	-7.850	0.000	-0.455	-0.273
VISIBILITY_Freezing Rain	-0.3782	0.058	-6.536	0.000	-0.492	-0.265
VISIBILITY_Clear	-0.3642	0.046	-7.850	0.000	-0.455	-0.273
VISIBILITY_Freezing Rain	-0.3782	0.058	-6.536	0.000	-0.492	-0.265
VISIBILITY_Drifting Snow	-0.8192	0.142	-5.777	0.000	-1.097	-0.541
VISIBILITY_Strong wind	-0.8656	0.178	-4.857	0.000	-1.215	-0.516

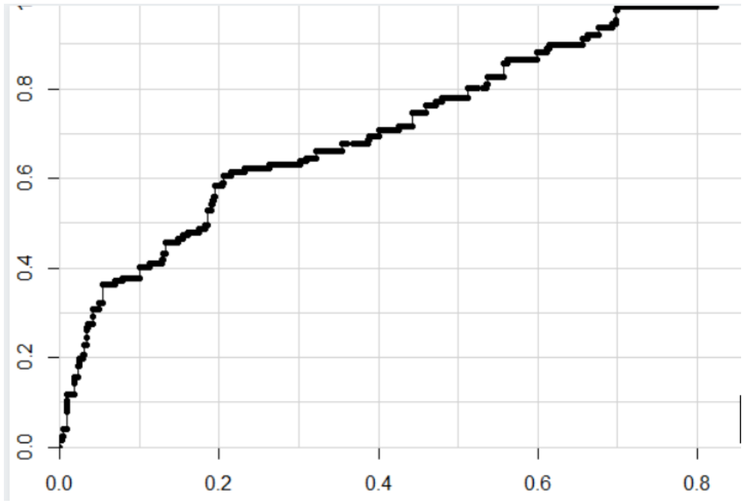



```
# R2 from linear regression
reg = LinearRegression()
reg.fit(train_X2, train_y2)



# print performance measures
regressionSummary(train_y2, reg.predict(train_X2))

pred_y2 = reg.predict(train_X2)

print('adjusted r2 : ', adjusted_r2_score(train_y2, pred_y2, reg))
r2=adjusted_r2_score(train_y2, pred_y2, reg)
```

```
Regression statistics

            Mean Error (ME) : 0.0000
Root Mean Squared Error (RMSE) : 0.3313
     Mean Absolute Error (MAE) : 0.2218
adjusted r2 :  0.07577691251059449
```

```
#Getting contribution of all the variables to the model.
contribution=r/r2
contribution.sort_values(ascending=False)
contribution_df=contribution.sort_values(ascending=False)
contribution_df.to_excel (r'contribution.xlsx', index = True, header=True
```

- Roc curve helped us to find the appropriate threshold . Each point in the graph responds to certain true positive rate and false positive rate.
- **Aim:** Keep false positive rate as low as possible to avoid misclassifying observations that are non fatal as fatal.
- **Result:** Chosen threshold 0.25 even though it did not give a good true positive rate but our fatal predictions match

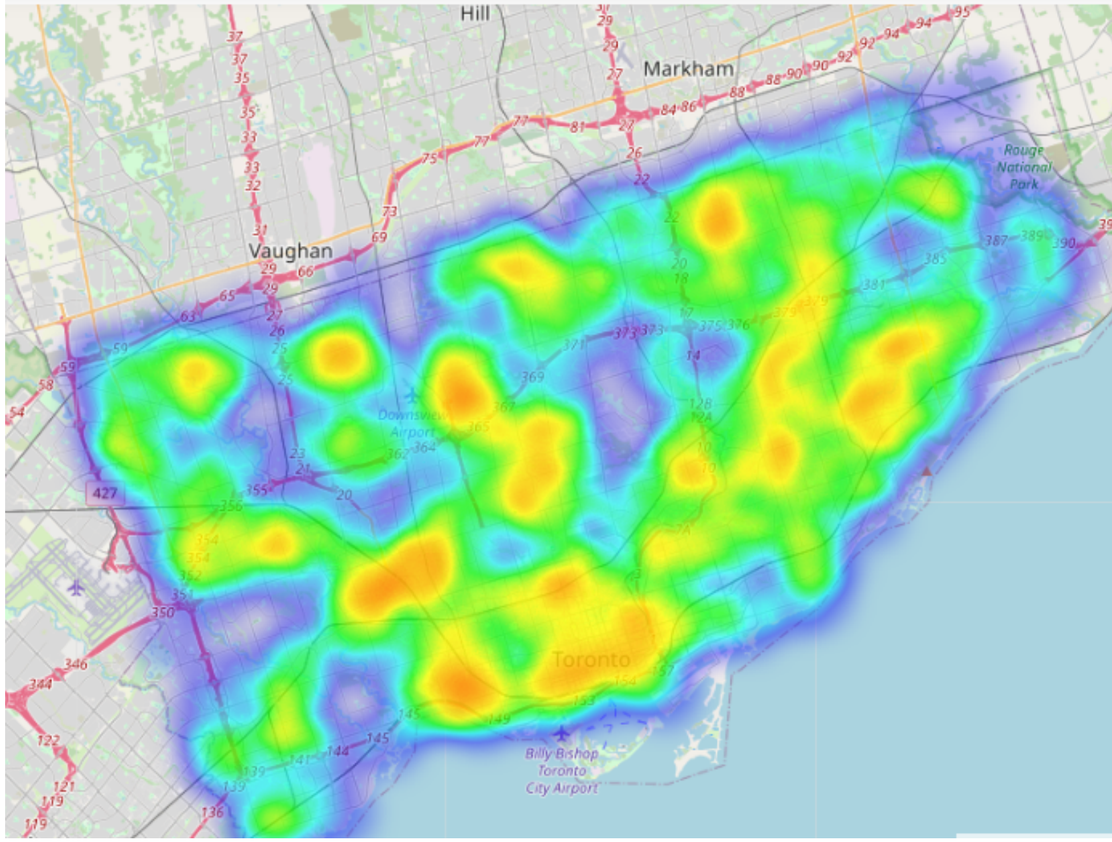


Variable Contribution to fatal

```
TRUCK                                  19.607950
PEDESTRIAN                             13.274493
SPEEDING                               12.975937
DISTRICT_Toronto and East York          8.164544
DISTRICT_Scarborough                    5.593658
TRSN_CITY_VEH                           5.084406
VISIBILITY_Other                        4.197423
LIGHT_Dark                              3.829338
RDSFCOND_Other                          3.683731
YEAR                                    2.990424
HOUR                                    1.524459
LIGHT_Other                             1.165578
AG_DRIV                                 0.985190
LONGITUDE                               0.828050
MINUTE                                  0.727332
RDSFCOND_Ice                            0.612490
VISIBILITY_Drifting Snow                0.248054
VISIBILITY_Freezing Rain                0.245648
VISIBILITY_Freezing Rain                0.245648
VISIBILITY_Snow                         0.199388
MOTORCYCLE                              0.187003
LIGHT_Dusk                              0.141740
VISIBILITY_Strong wind                  0.104372
VISIBILITY_Rain                         0.098608
VISIBILITY_Clear                        0.053707
VISIBILITY_Clear                        0.053707
DISABILITY                              0.035598
VISIBILITY_Fog, Mist, Smoke, Dust       0.014643
PASSENGER                               0.007262
REDLIGHT                                0.000298
DISTRICT_Etobicoke York                 0.000002
Name: FATAL, dtype: float64
```

Decile wise fatality rate

Deciles	Fatality Rates
1	27%
2	18%
3	16%
4	14%
5	11%
6	10%
7	9%
8	7%
9	6%
10	4%
	122%



Cause / Neighbourhood	SPEEDING	AG_DRIV	TRUCK	AUTOMOBILE	TRSN_CITY_VEH	ALCOHOL	REDLIGHT
Waterfront Communities-The Island (77)	292	56	42	501	45	25	78
West Humber-Clairville (1)	278	99	86	467	24	33	50
Rouge (131)	231	78	10	347	19	16	47
Bay Street Corridor (76)	199	22	22	326	40	2	65
Wexford/Maryvale (119)	198	60	23	328	14	19	34
Woburn (137)	198	56	16	350	5	32	57
Islington-City Centre West (14)	144	45	25	255	9	14	24
South Riverdale (70)	141	23	30	265	26	4	21
York University Heights (27)	137	26	28	220	5	9	29
Banbury-Don Mills (42)	134	30	14	173	0	21	11