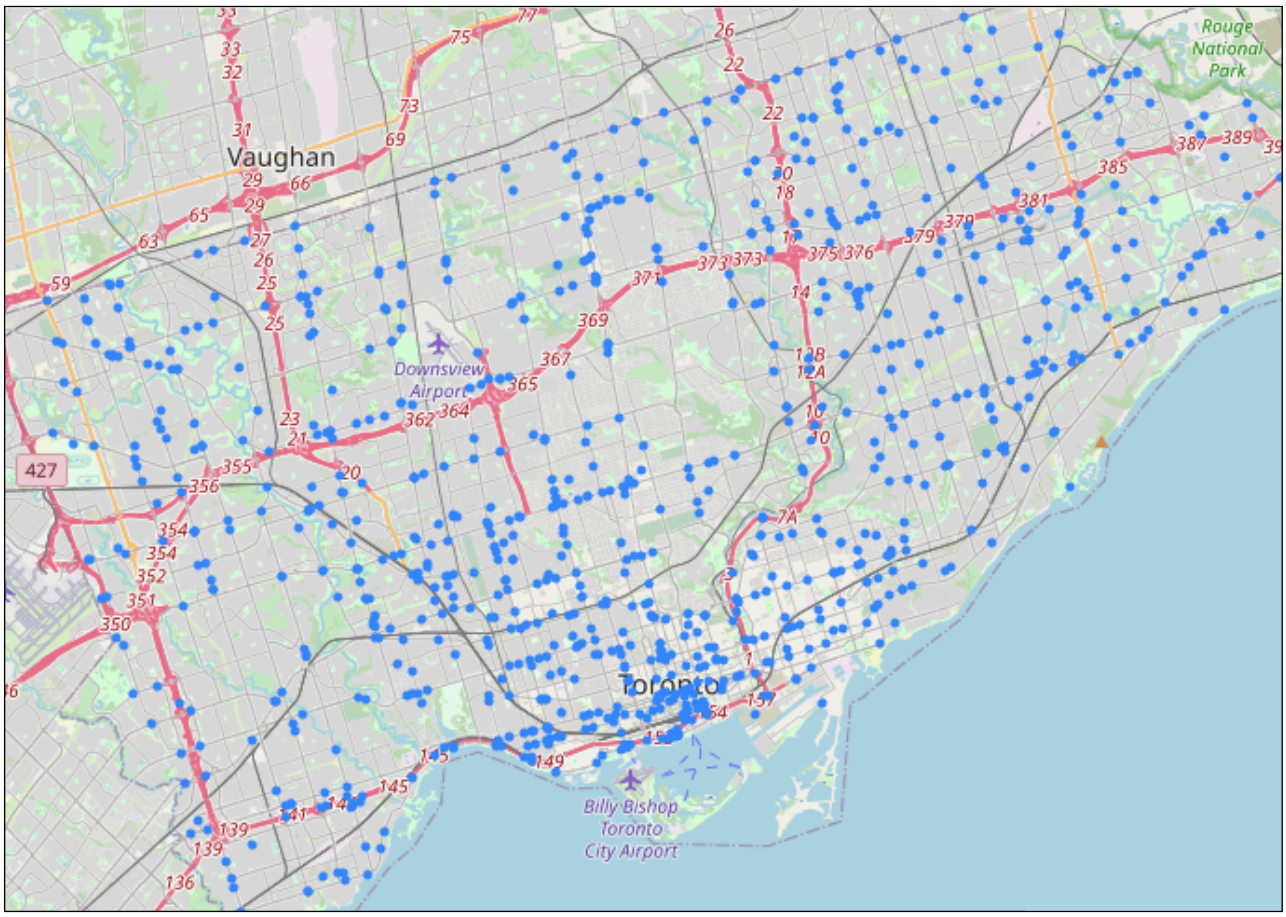
Predicting fatality of a serious accidents in TORONTO

Killed or Seriously Injured Dataset from Toronto Police (2006-2019)



By *Minh Ngoc Pham*

Capstone Project #1

Milestone Report

Introduction

Toronto's commute has often been considered one of the worst in the world. The city's public transport system faces various criticisms as it has a much smaller subway network system compared to cities with similar size. Most of people in Toronto commute by cars as a result leading to frequent traffics and potential accidents. Toronto Police compiled data from 2006-2019 on Killed or Seriously Injured accidents in the city of Toronto over this time period. This project aims to identify the potential variables leading to a serious/fatal accident in Toronto.

The Dataset

The dataset for this project was published publicly by the Toronto Police Service.

The dataset has 14,457 rows including detailed description of 5,690 accidents over the period of 14 years (2006-2019). There are more rows than the number of accidents because each accident is recorded in the number of rows equivalent to the number of people involved (passengers, drivers, cyclists, pedestrians and etc.).

The columns of interest include: 'ACCNUM' -accident code, 'ROAD_CLASS', 'LOCCOORD' - where the accident is located, 'TRAFFCTL' - whether there is traffic signal, 'VISIBILITY', 'LIGHT', 'RDSFCOND' - road conditions, 'ACCLASS' - whether it was fatal or non-fatal, 'IMPACTYPE' - nature of impact, 'INVTYPE' - who is involved, 'INVAGE' - age of involved person, 'INJURY' - nature of injury, 'VEHTYPE' - vehicle, 'MANOEUVER', 'DRIVACT' - driver's action, 'DRIVCOND', 'PEDTYPE', 'PEDACT', 'PEDCOND', 'CYCLISTYPE', 'CYCACT', 'CYCCOND', 'PEDESTRIAN', 'CYCLIST', 'AUTOMOBILE', 'MOTORCYCLE', 'TRUCK', 'EMERG_VEH', 'PASSENGER', 'SPEEDING', 'AG_DRIV' - aggressive driving, 'REDLIGHT', 'ALCOHOL', 'DISABILITY'

All of the columns are categorical. The independent variable of interest is ACCLASS - which tells us whether the accident is fatal or not and INJURY - which tells us how injured is the person.

Source: https://data.torontopolice.on.ca/datasets/ksi/data?geometry=-80.371%2C43.550%2C-78.393%2C43.897&page=2

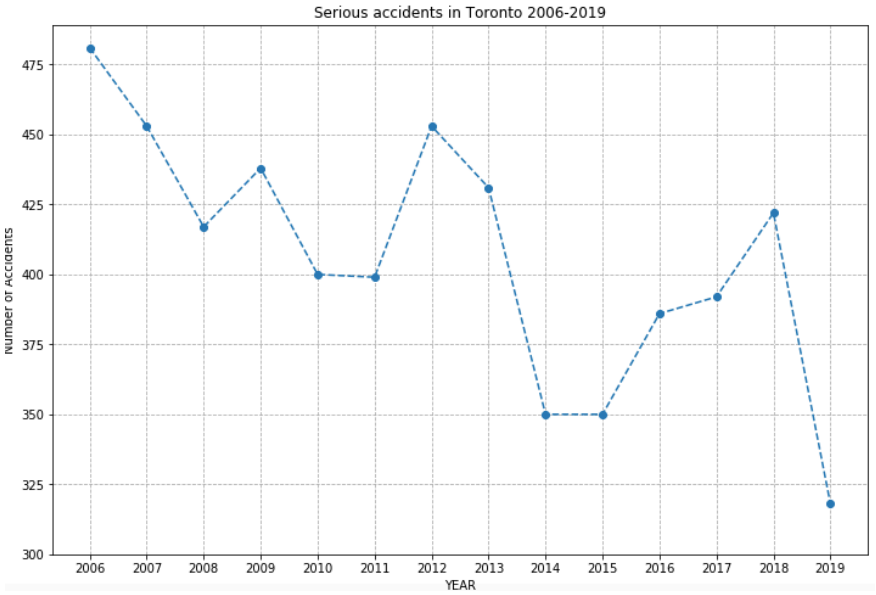
I. Data Wrangling:

- Combine DATE and TIME columns together to create a date time column
- Replace all the empty columns with NaN
- Dropping columns that are not of interest: 'OFFSET', 'Division','ACCLOC', 'Hood_ID', 'FATAL_NO', 'Index_', 'YEAR', 'DATE', 'TIME', 'HOUR', 'X', 'Y', 'WardNum'
- Cleaning and filling District/Ward/Neighbourhood columns

- Replacing all NaN values with 'No' (for columns with Yes/No values) and 'Unknown/Other' (for other categorical columns).

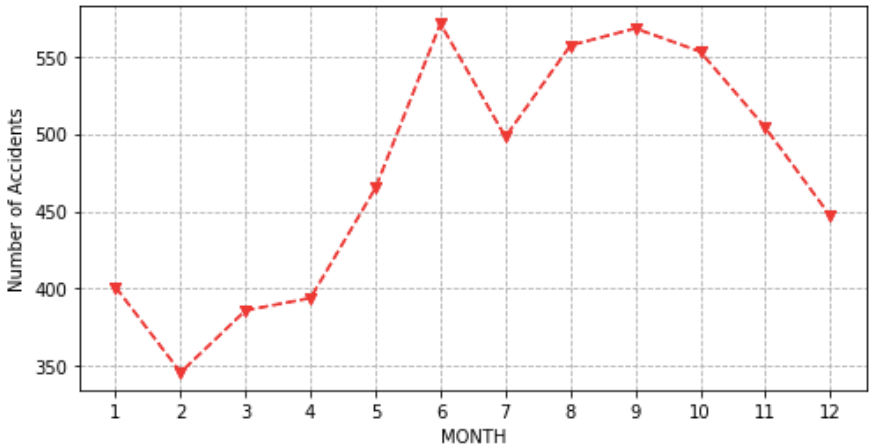
II. Findings from exploratory analysis:

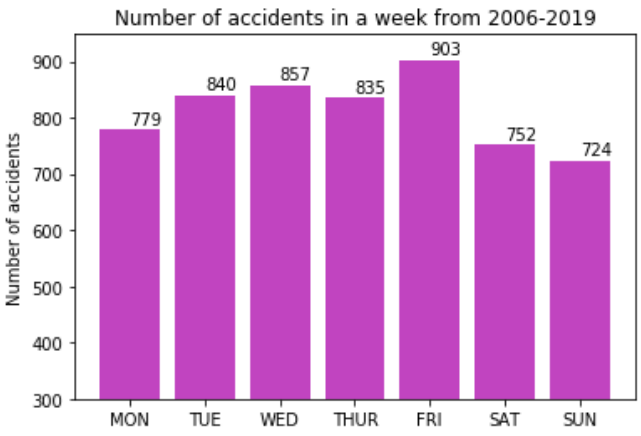
Looking at the trend over time:



OVER THE YEAR: From the graph above, we can see there is a general decline in the number of serious accidents happening in Toronto in the past 7 years (from 2013 to 2019) compared to from 2006 to 2013. The biggest drops are from the period from 2013 and 2014 and the period from 2018 to 2019 where we can see a big decline in the number of serious accidents.

OVER THE MONTH: We can see from the graph that there is a very low number of traffic accidents happening from December until April in Toronto, which are also winter months in Toronto where the temperature is lowest with a lot of snow days.

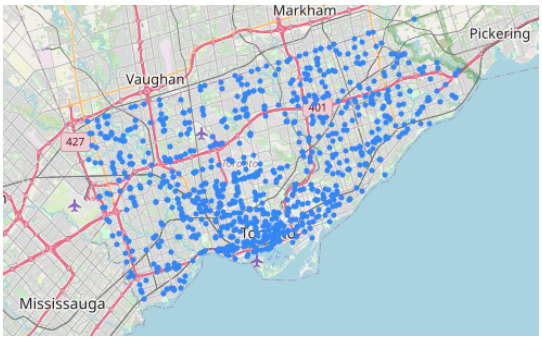




This seems to correspond to the fact that people stay in more during winter, hence there are less number of accidents.

OVER THE WEEK: From the bar chart above, we can see that there is a higher number of serious accidents on Friday. Assuming that it's the last day of the week so people tend to go out more at night, we can have a look at the data of those accidents happening on Friday only and see what time it usually occurred.

LOCATION

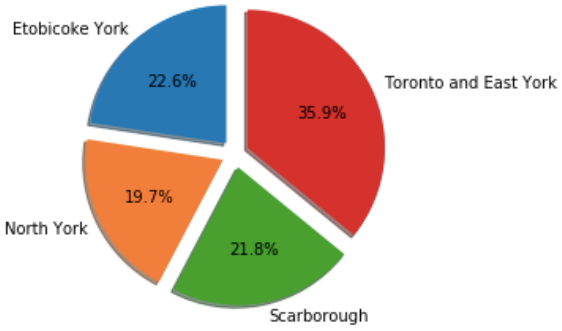


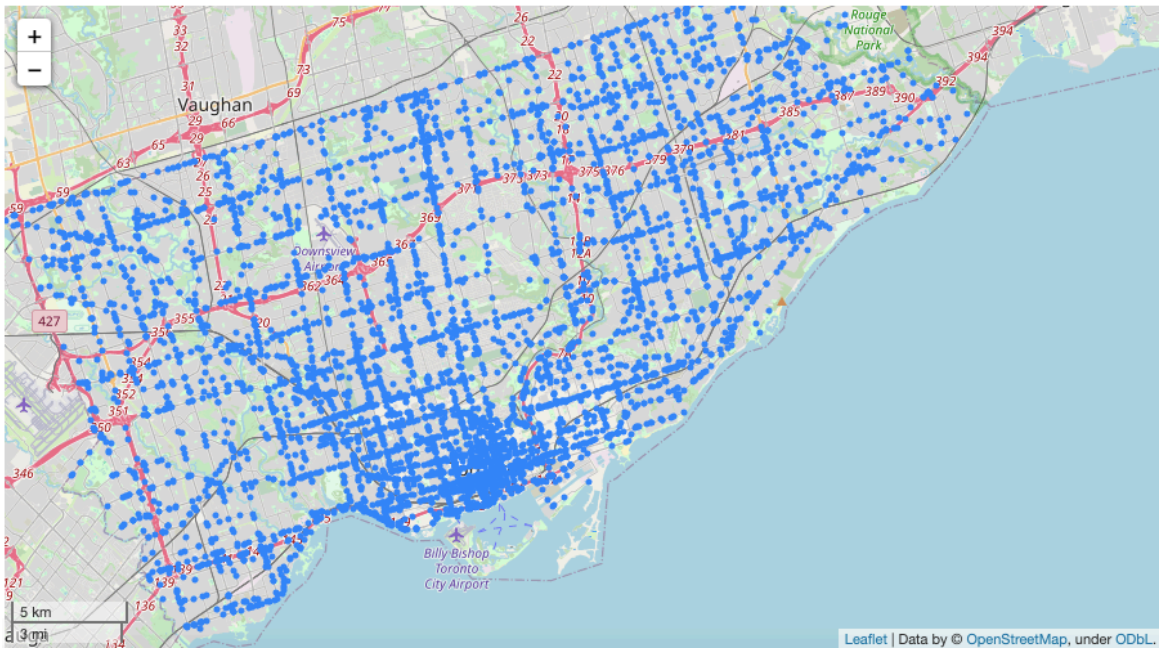
Having a closer look at the accidents happening on Friday (based on coordinates), we notice a large number of accidents happening in the downtown area on Friday, which corresponds to party night for many people.

By **Neighbourhood**:

The highest number of accidents happening in Toronto and East York seems to coincide with the neighbourhood of Waterfront Communities-The Island (77) as this neighbourhood is located in this district.

Neighbourhood	ACCNUM
Waterfront Communities-The Island (77)	205
West Humber-Clairville (1)	166
Bay Street Corridor (76)	134
Rouge (131)	129
Woburn (137)	116





The map above shows that our finding is quite accurate in terms of the location of the accidents, a lot of them occurred in the waterfront area of Toronto (strong cluster of accidents over the year).

ROAD_CLASS	Total accidents	Fatal	Rate of Fatal Accidents
Collector	354	49.0	0.138418
Expressway	6	1.0	0.166667
Laneway	4	3.0	0.750000
Local	289	44.0	0.152249
Major Arterial	3981	538.0	0.135142
Major Arterial Ramp	1	0.0	0.000000
Minor Arterial	931	107.0	0.114930
Unknown/Other	124	24.0	0.193548

Looking at the ***road class*** of the accident location:
Without looking at the rate of Fatal accidents for Unknown and Laneway, we can see that accidents happening on Expressway tend to be more fatal (0.166666) than major arterial (0.135) even though there are more accident happening on major arterial roads. A possible explanation for this could be the fact that people tend to drive faster on Expressway making it more fatal when collision occurs.

Looking at the location with ***high frequency of accidents*** (those intersections with more than 5 accidents over the years):

We can see from the table here that the almost half (205 out of 451) of the location with high frequency of accidents are with pedestrian collisions.

IMPACTYPE	ACCNUM
Angle	41
Approaching	10
Cyclist Collisions	28
Other	7
Pedestrian Collisions	205
Rear End	32
SMV Other	37
SMV Unattended Vehicle	1
Sideswipe	7
Turning Movement	82

INVTYPE	INJURY Fatal	Non-Fatal
Cyclist	37.0	632.0
Cyclist Passenger	0.0	2.0
Driver	145.0	2578.0
Driver - Not Hit	0.0	1.0
In-Line Skater	0.0	4.0
Moped Driver	0.0	26.0
Motorcycle Driver	65.0	467.0
Motorcycle Passenger	1.0	28.0
Passenger	87.0	1581.0
Pedestrian	441.0	2251.0
Truck Driver	3.0	49.0
Wheelchair	2.0	11.0

```
fatal=compare1['Fatal']
non_fatal=compare1['Non-Fatal']
table = pd.crosstab(fatal, non_fatal)
from scipy.stats import chi2_contingency
chi2, p, dof, expected = chi2_contingency(table.values)
print (chi2, p)
```

```
95.99999999999999 0.262528152279769
```

III.Chi-square test for two variables

Most of the variables used in the analysis are categorical variables so we will use Chi-square to determine if there is a relationship between two variables.
We will carry out chi-square test for a couple of variable-pairs to determine if there is some correlation between them (they

are not independent if p value from the chi-square test is low)

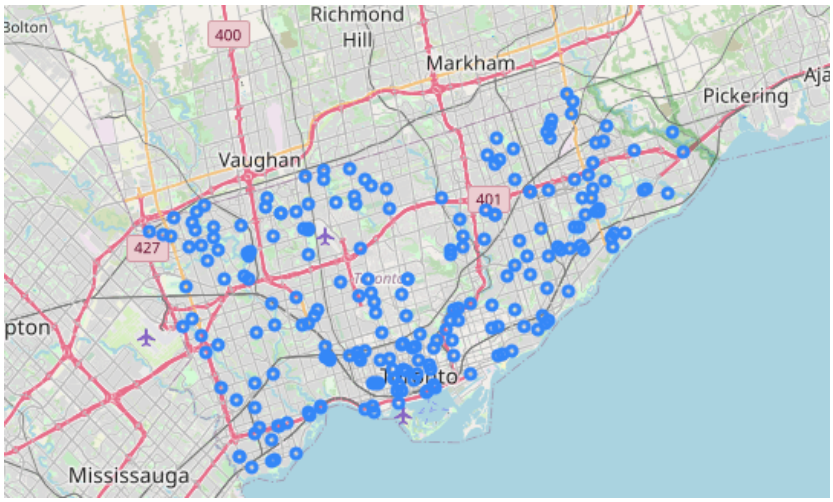
- INVOLVED PERSON AND INJURY: low P-value, that means who the involved person is does not necessarily determined whether their injury will be fatal or not.
- Is there a relationship between Neighbourhood and whether alcohol is involved?

ALCOHOL	No	Yes
Neighbourhood		
Agincourt North (129)	145.0	5.0
Agincourt South-Malvern West (128)	142.0	0.0
Alderwood (20)	57.0	7.0
Annex (95)	185.0	9.0
Banbury-Don Mills (42)	143.0	19.0
...	...	...
Wychwood (94)	92.0	0.0
Yonge-Eglinton (100)	45.0	4.0
Yonge-St.Clair (97)	46.0	0.0
York University Heights (27)	205.0	9.0
Yorkdale-Glen Park (31)	102.0	3.0

140 rows × 2 columns

```
not_drunk=compare3['No']
drunk=compare3['Yes']
table3 = pd.crosstab(not_drunk, drunk)
from scipy.stats import chi2_contingency
chi2_3, p_3, dof_3, expected_3 = chi2_contingency(table3.values)
print (chi2_3, p_3)
```

```
2328.8207547169814 9.225194703874935e-05
```



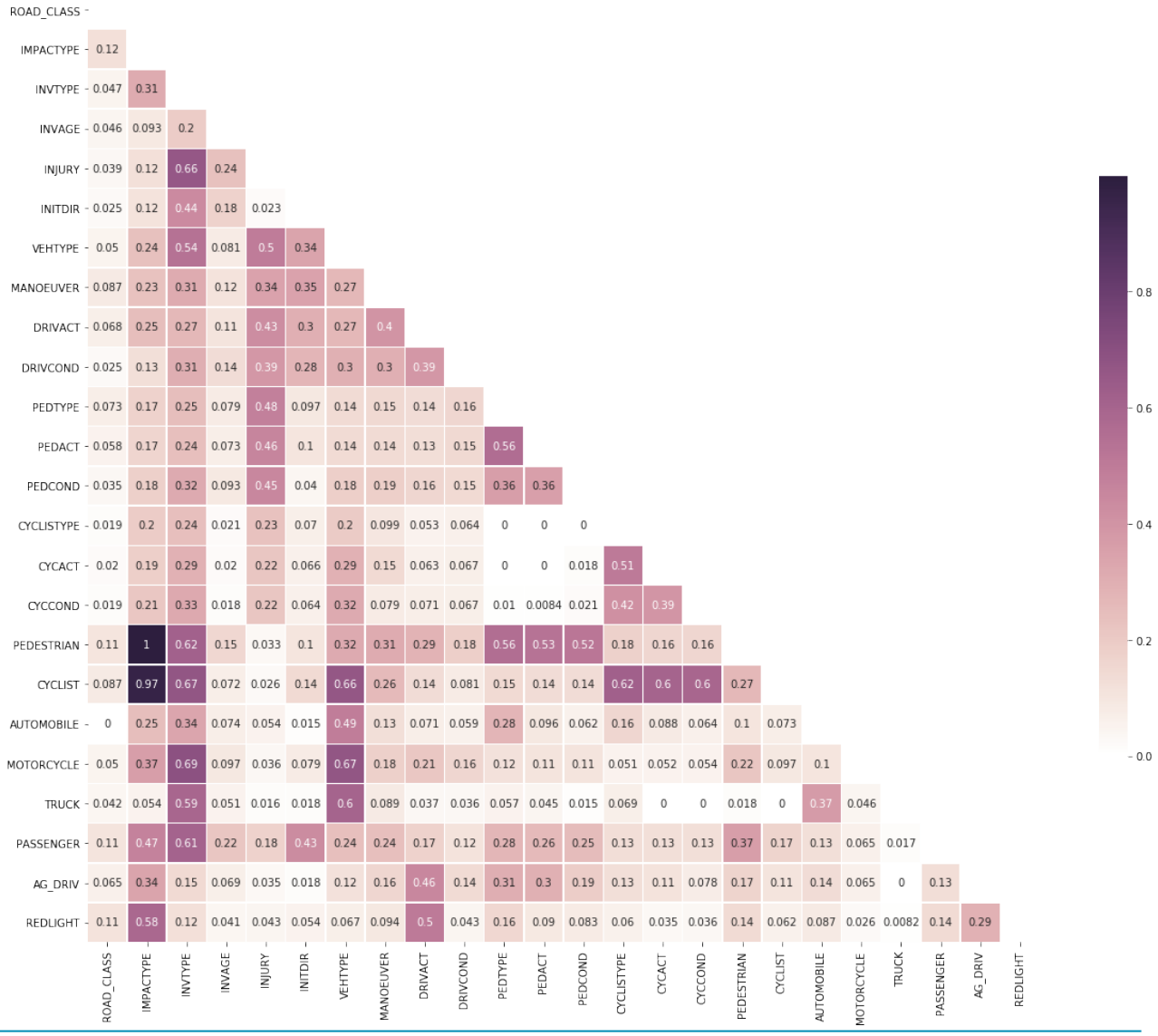
The P-value is actually very low here, suggesting a relationship between location of the accident and whether there is alcohol involved. Further look into the relationship shows that most of the alcohol-involved accidents occur in about 87 out of 140 neighbourhoods.

IV. Machine learning process:

The KSI dataset consists mostly of categorical data, and hence we will use the method of supervised learning to make prediction on whether a certain person will survive or not from an accident (each row represents a single person involved in the accident).
The independent variable in question is 'INJURY', which we will classify as Fatal/Major or Minimal as impact in a person from an accident.

1. Preprocessing:

- Dropping unnecessary columns and engineer the date columns to get hour and day of the week
- Our dataset looks quite balanced in terms of the two classes: Minimal and Fatal/Major with slightly more number of people falling into the Minimal Injury (57.4% versus 42.6%)
- The majority of our data includes categorical variables so we will perform Chi-square to determine independence with independent variables (in this case 'INJURY')
- From Chi-square we can see the following variables: HOUR, District, LOCCOORD, TRAFFCTL, VISIBILITY, LIGHT, RDSFCOND, EMERG_VEH, SPEEDING, ALCOHOL,



DISABILITY, Hood_ID, TRSN_CITY_, WEEKDAY show high P-value and we will drop them from the models. In fact, this seems quite intuitive that speeding/ alcohol makes an accident fatal/major.

- Another interesting way to determine relationship between data is by looking at the association heatmap. Cramer's V is a measure of association between nominal variables. The value is between 0 and 1, with a number closer to 1 denoting a strong association. It is based on the chi squared statistic.

- Based on the heat map, there is a strong association between Pedestrian and impact type and cyclist and impact type. It seems intuitive since PEDESTRIAN and CYCLIST are dummy variables and whether an accident involves them or not will classify the type of impact it is. For this reason, we will also drop IMPACTYPE from the columns.

- ***Converting data into numerical values:***

Machine learning requires that the input and the output variables being numerical values. As a result, the first step in getting the data ready is through making sure all the variables are in numerical form.

For the X value, we will use the method called OneHotEncoder to transform the data into different columns since all the data we have are unordered categorical variables, we do not want our model to run as if the numbered ordering matters. OneHotEncoder will split data in the number of columns corresponding to the number of options for each categorical data.

```
from sklearn.preprocessing import OneHotEncoder, LabelBinarizer
# Transform all the X's categorical values into numerical values
onehotencoder = OneHotEncoder()
X_2 = onehotencoder.fit_transform(X).toarray()
# y= column 'INJURY' is of binary value and hence we will convert it into binary label
lb = LabelBinarizer()
y_2 = lb.fit_transform(y)

X_2
```

```
array([[0., 0., 0., ..., 1., 1., 0.],
       [0., 0., 0., ..., 1., 1., 0.],
       [0., 0., 0., ..., 1., 1., 0.],
       ...,
       [0., 0., 0., ..., 1., 0., 1.],
       [0., 0., 0., ..., 1., 0., 1.],
       [0., 0., 0., ..., 1., 0., 1.]])
```

As we notice by looking at X_2, we can see that all the data now includes dummy variables and X_2, the features, look like a matrix with 0 and 1 values only.

2. Building machine learning model:

Now that we have our data ready, we will build a model to predict whether a person will suffer from a major/fatal injury or a minimal injury impact by running our data above. The classification problem is a binary problem and hence decision tree, which will split data in different groups seem appropriate.

```
from sklearn.model_selection import cross_validate, StratifiedKFold, ShuffleSplit
from sklearn.tree import DecisionTreeClassifier
from sklearn import tree
# Define our data splitting
split = StratifiedKFold(n_splits=5, shuffle=True, random_state=0)
treemodel = DecisionTreeClassifier()
score = cross_validate(treemodel, X_2, y_2.ravel(), cv=split)
print("Test score:                 {}".format(scores["test_score"]))
# Print average across K tests
print("Average test score:       any   {} (+/- {})".format(scores["test_score"].mean(), scores["test_score"].std() * 2)
```

```
Test score:                 [0.81507086 0.81044621 0.81528883 0.82601176 0.81840194]
Average test score:       any   0.8170439196318646 (+/- 0.010306092754609726)
```

The decision tree model gives us a pretty good scoring of approximately 81.7% accuracy of predicting whether a person would suffer a Fatal/Major injury or not from the accident.