Minh Ngoc Pham
Data Science Career Track
25 August 2020

Capstone Project I:
Exploratory Data Analysis - Inferential Statistics

I. Project Question:

The dataset in question is the KSI (Killed or Seriously Injured) data from Toronto Police Portal. The most important question we will be asking is:
Can you tell whether a person involved in an accident is fatal or not based on different features of that specific accident?
The variables in questions are:

```
dataanalysis.columns
```

```
Index(['ACCNUM', 'ROAD_CLASS', 'LOCCOORD', 'TRAFFCTL', 'VISIBILITY', 'LIGHT',
       'RDSFCOND', 'ACCLASS', 'IMPACTYPE', 'INVTYPE', 'INVAGE', 'INJURY',
       'INITDIR', 'VEHTYPE', 'MANOEUVER', 'DRIVACT', 'DRIVCOND', 'PEDTYPE',
       'PEDACT', 'PEDCOND', 'CYCLISTYPE', 'CYCACT', 'CYCCOND', 'PEDESTRIAN',
       'CYCLIST', 'AUTOMOBILE', 'MOTORCYCLE', 'TRUCK', 'EMERG_VEH',
       'PASSENGER', 'SPEEDING', 'AG_DRIV', 'REDLIGHT', 'ALCOHOL',
       'DISABILITY'],
```

All these variables are categorical variables and **INJURY** is the dependent variable that we will be predicting.

II. 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

INJURY	Fatal	Non-Fatal
INVTYPE		
Cyclist	37.0	632.0
Cyclist Passenger	0.0	2.0
Driver	745.0	25768.0
Driver - Not Hit	0.0	1.0
In-Line Skater	0.0	4.0
Moped Driver	0.0	26.0
Motorcycle Driver	85.0	467.0
Motorcycle Passenger	1.0	38.0
Passenger	87.0	1681.0
Pedestrian	441.0	2251.0
Truck Driver	3.0	49.0
Witness	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
```

person is does not necessarily determined whether their injury will be fatal or not.

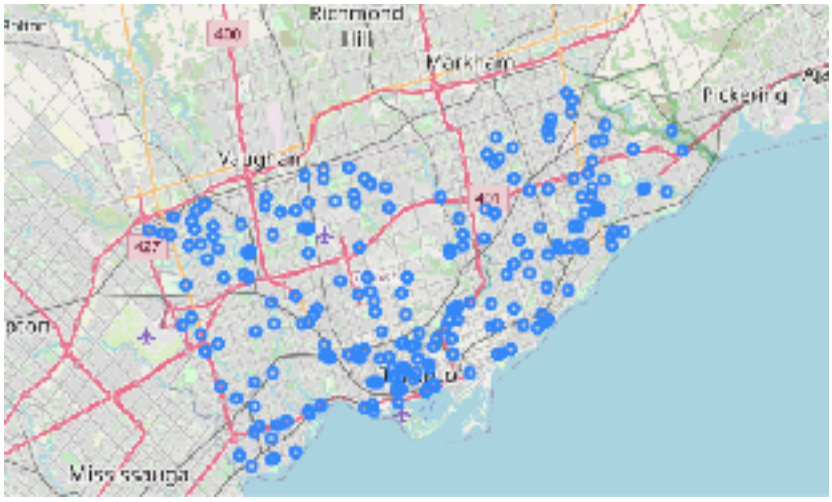
- Is there a relationship between Neighbourhood and whether alcohol is involved?

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.