DATA CAPSTONE PROJECT

Part III: Data Assessment and Initial Analysis

Group: Insight Ink Crew

Professor: Zeyad Azem

Date: 1st December, 2023

Submitted by:

Baskaran, Asmita

Coimbatore Jayaraj, Pavithran

Kaur, Harmanjot

Keshav, Pooja

Nalchigar, Niloofar

Pham, Thi

Document Scope

This Report looks at historical patterns of murderous incidents in Toronto using data from the Toronto Police open database. It examines the quantity, location, and timing of incidents, examining data from the preceding five years to identify hotspots across different regions. An investigation of Toronto's police divisions' relative homicide rates will be done to determine whether ones usually have higher or lower rates of homicides. To find any significant differences, the report also compares the frequency of murder cases in Toronto during the preceding five years. By integrating descriptive data and visualizations, we aim to provide insights for informed decision-making and crime prevention measures.

Objectives Overview

The following Data Assessment has been done based on the 'Homicide' dataset from Toronto Police open database and the objectives which were prioritized earlier.

Descriptive

- What are the historical trends in homicide incidents in Toronto in terms of frequency, location, and time of occurrence?
- How many Homicide cases were reported in the last 5 years in Toronto, which regions were hotspots?
- What is the distribution of homicides across different divisions within Toronto, and are there divisions with consistently higher or lower rates?
- Is there any significant difference between Ottawa and Toronto regarding Homicide crime frequency within the last 5 years?

Diagnostic

- To what extent do external factors like economic conditions and major events (like pandemic) or immigration rate impact homicide rates in Toronto?
- Is there any correlation between the type of homicide and the neighborhood (region) it occurs in?
- Are homicides increasing in specific neighborhoods?
- Are there specific socio-economic factors, such as unemployment rates or income inequality, that show a strong correlation with the occurrence of homicides?

Predictive

- How can a predictive model be developed to estimate resource requirements for future homicide incidents in Toronto, incorporating historical data, immigration trends, and other relevant variables?
- What is the likelihood of homicides occurring in specific areas within the next six months?

- How many police resources need to be allocated to be responsive, according to the increasing rate of the population?

Prescriptive

- What specific strategies and interventions can be recommended to law enforcement agencies to enhance response times or proactive initiatives for homicide incidents, taking into account factors like location, time of day, and historical data?
- How can law enforcement agencies be provided with actionable insights to allocate resources effectively in areas where future homicide incidents are predicted to be more likely, using the developed predictive model?
- What sort of educational programs can be implemented that promote the negotiation of disputes and the avoidance of violence within neighborhoods and schools?
- What efforts may be made to get high-risk communities more engaged with community policing to avoid homicides?

Data Assessment

This step is a critical initial step where the quality, structure, and relevance of the available data are thoroughly examined. This phase involves understanding the data sources, assessing data completeness, handling missing values, and identifying potential anomalies or outliers. Also Statistical techniques are employed to gain insights into the distribution and characteristics of the datasets.

Tools and programming languages that have been leveraged for this step are Python and Excel.

1- Overall Description

The basic statistical description for numerical variables of the data set is as below:

```
In [14]: homicides.describe().T
```

Out[14]:

	count	mean	std	min	25%	50%	75%	max
X	1376.0	-8.838600e+06	12506.058839	-8.863555e+06	-8.849713e+06	-8.838515e+06	-8.829406e+06	-8.808499e+06
Y	1376.0	5.420970e+06	7964.930984	5.402687e+06	5.413941e+06	5.420367e+06	5.427392e+06	5.442381e+06
OBJECTID	1376.0	6.885000e+02	397.361297	1.000000e+00	3.447500e+02	6.885000e+02	1.032250e+03	1.376000e+03
OCC_YEAR	1376.0	2.013579e+03	5.860770	2.004000e+03	2.008000e+03	2.014000e+03	2.019000e+03	2.023000e+03
OCC_DAY	1376.0	1.566206e+01	8.786099	1.000000e+00	8.000000e+00	1.600000e+01	2.300000e+01	3.100000e+01
OCC_DOY	1376.0	1.879760e+02	102.861008	1.000000e+00	1.037500e+02	1.950000e+02	2.710000e+02	3.660000e+02
LONG_WGS84	1376.0	-7.939850e+01	0.112344	-7.962267e+01	-7.949833e+01	-7.939773e+01	-7.931590e+01	-7.912809e+01
LAT_WGS84	1376.0	4.371190e+01	0.051714	4.359309e+01	4.366627e+01	4.370801e+01	4.375361e+01	4.385079e+01

As this basic analysis suggests, there are no missing values for the numerical variables.

Since the focus of this project is for last 5 years, the new dataset has been defined :

```
In [22]: homicides2 = homicides[homicides.OCC_YEAR > 2018]

In [23]: homicides.shape
Out[23]: (1376, 18)

In [24]: homicides2.shape
Out[24]: (359, 18)
```

2- Missing Values

```
In [11]: homicides.isnull().sum()
Out[11]: X                   0
         Y                   0
         OBJECTID            0
         EVENT_UNIQUE_ID     0
         OCC_DATE            0
         OCC_YEAR            0
         OCC_MONTH           0
         OCC_DAY             0
         OCC_DOW             0
         OCC_DOY             0
         DIVISION            0
         HOMICIDE_TYPE       0
         HOOD_158            0
         NEIGHBOURHOOD_158   0
         HOOD_140            0
         NEIGHBOURHOOD_140   0
         LONG_WGS84          0
         LAT_WGS84           0
         dtype: int64
```

3- Duplicates & Outliers

Upon reviewing the data, we have confirmed the absence of duplicates and outliers.

```
duplicates = homicides[homicides.duplicated()]

if duplicates.empty:
    print ("No duplicates found.")
else:
    print ("Duplicates found:")
    print (duplicates)
```

```
No duplicates found.
```

```
import pandas as pd
from scipy import stats
```

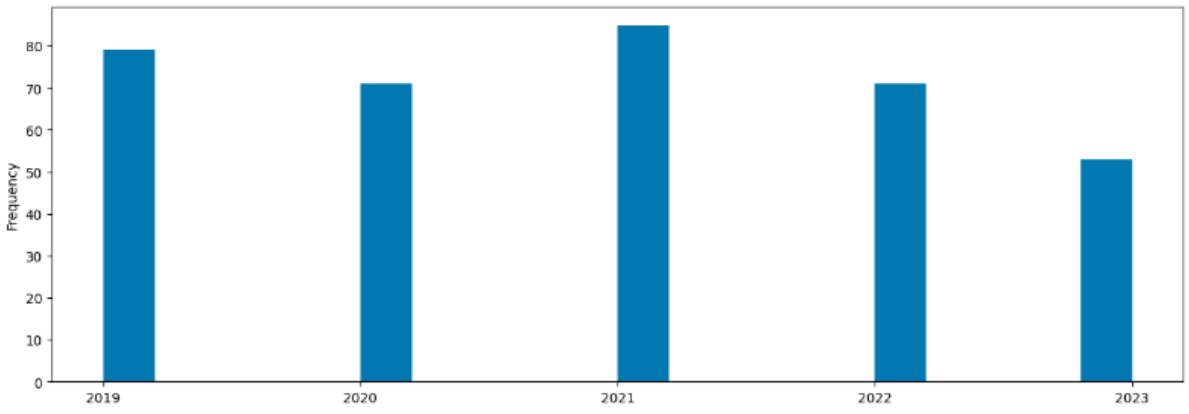
```
numerical_columns = ['X', 'Y','OCC_YEAR', 'LONG_WGS84', 'LAT_WGS84']
Z_scores= stats.zscore(homicides[numerical_columns])

threshold = 3
outlier_indices = (abs(Z_scores) > threshold).any(axis=1)

outliers = homicides[outlier_indices]
print ("Rows with outliers")
print(outliers)
```

```
Rows with outliers
Empty DataFrame
Columns: [X, Y, OBJECTID, EVENT_UNIQUE_ID, OCC_DATE, OCC_YEAR, OCC_MONTH, OCC_DAY, OCC_DOW, OCC_DOY, DIVISION, HOMICIDE_TYP
E, HOOD_158, NEIGHBOURHOOD_158, HOOD_140, NEIGHBOURHOOD_140, LONG_WGS84, LAT_WGS84]
Index: []
```

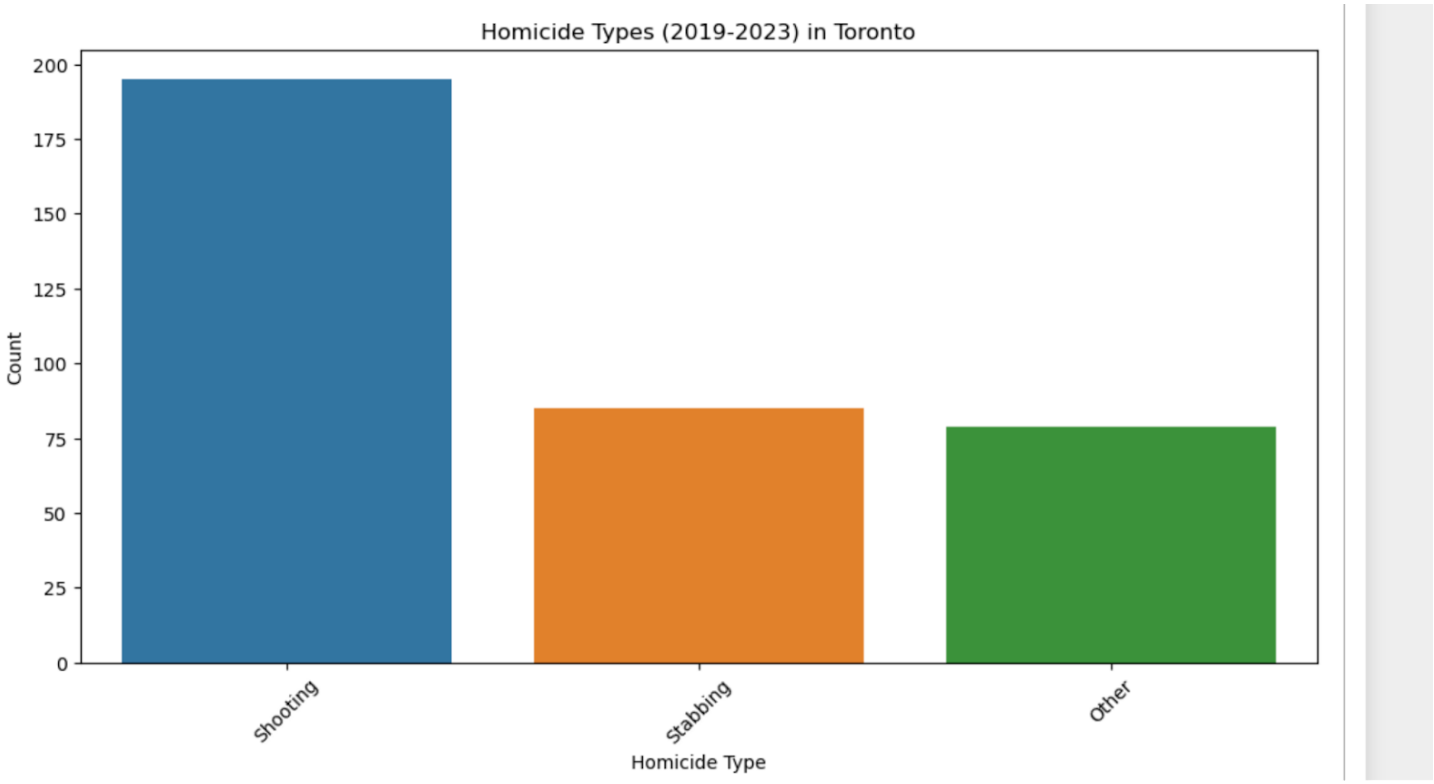
4- Initial Descriptive Analysis

Frequency of the incidents in last 5 years :

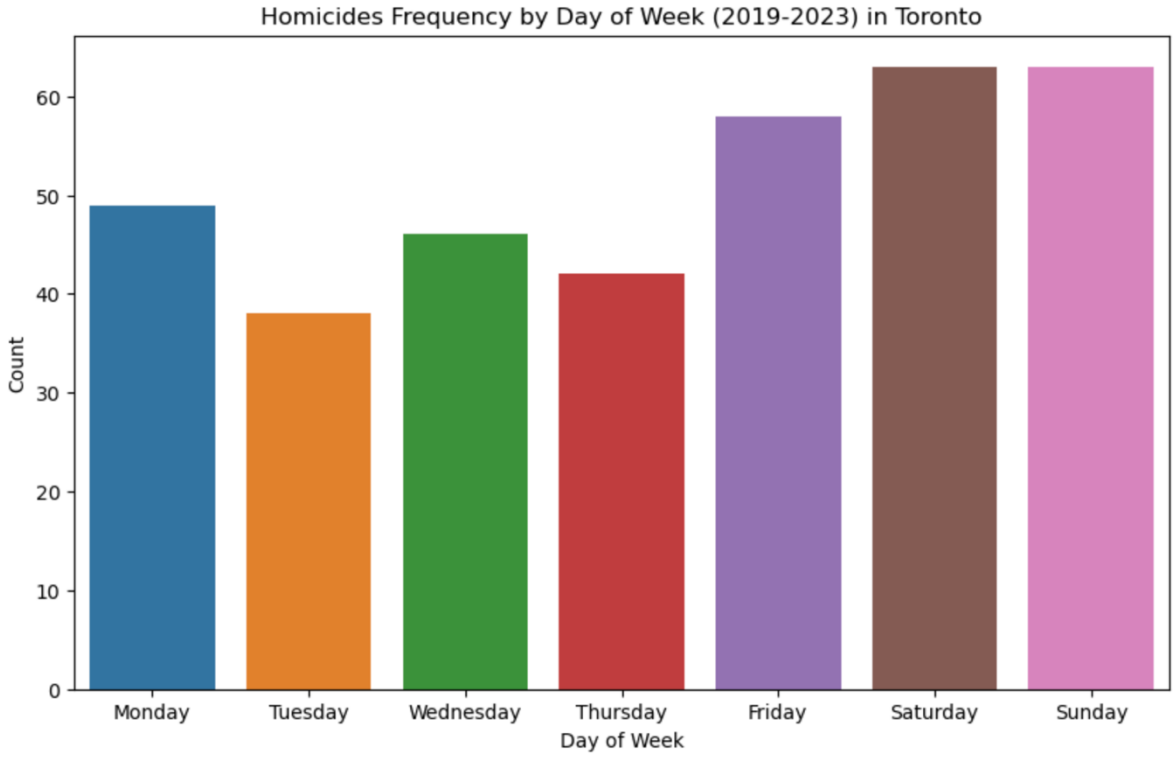
```
In [31]: import matplotlib.pyplot as plt

In [33]: homicides2["OCC_YEAR"].plot.hist(bins=20, figsize=(15, 5))
         plt.xticks(range(min(homicides2["OCC_YEAR"]), max(homicides2["OCC_YEAR"])+1))
         plt.show()
```



Initially, the analysis reveals a substantial count of approximately 200 cases involving shootings, accompanied by nearly 80 incidents categorized as stabbings. Additionally, roughly 75 cases fall under various other reported circumstances during the comprehensive study of homicides spanning from 2019 to 2023.



The data on homicides in Toronto from 2019 to 2023, categorized by the day of the week, reveals interesting patterns. Sundays and Saturdays exhibit the highest frequency, each accounting for 63 cases. Fridays follow closely with 58 reported incidents, while Mondays have 49 cases. Wednesdays, Thursdays, and Tuesdays show decreasing frequencies, with 46, 42, and 38 cases, respectively. This information sheds light on the distribution of homicides throughout the week, providing valuable insights into potential temporal patterns or trends in crime occurrences.



The analysis indicates that the five regions, namely Forest Hill North, Bathurst Manor, Yonge-St. Clair, Danforth East York, and Long Branch, each reported only a single case, marking them as the areas with the least occurrences in the dataset.

```
region_occurrences = homicides['NEIGHBOURHOOD_158'].value_counts()

bottom_5_regions = region_occurrences.tail(5)

print("Top 5 regions with the least occurrences:")
print(bottom_5_regions)
```

```
Top 5 regions with the least occurrences:
Forest Hill North     1
Bathurst Manor        1
Yonge-St.Clair        1
Danforth East York    1
Long Branch           1
Name: NEIGHBOURHOOD_158, dtype: int64
```

Upon analyzing the dataset, it's evident that Moss Park and Mount Olive-Silverstone-Jamestown are the most concerning regions, with 40 occurrences each, leading in reported incidents. Following closely behind, Glenfield-Jane Heights and South Riverdale show a slightly lower but still substantial number of incidents at 38 and 26, respectively. Black Creek, with 25 incidents, also exhibits a significant frequency of reported events. These top 5 regions collectively account for a considerable portion of the reported incidents, signaling a concentrated area of concern for further detailed examination and targeted intervention strategies.

```
hotspot_regions = homicides['NEIGHBOURHOOD_158'].value_counts().head(5)

print("Top 5 regions with the most occurrences:")
print(hotspot_regions)
```

```
Top 5 regions with the most occurrences:
Moss Park                          40
Mount Olive-Silverstone-Jamestown  40
Glenfield-Jane Heights             38
South Riverdale                    26
Black Creek                        25
Name: NEIGHBOURHOOD_158, dtype: int64
```

Below is the detailed information for these five hotspot areas:

Region: Moss Park

DIVISION	OCC_DATE	HOMICIDE_TYPE	NEIGHBOURHOOD_158
D51	2004/04/01 05:00:00+00	Other	Moss Park
D51	2004/06/01 04:00:00+00	Other	Moss Park
D51	2004/06/18 04:00:00+00	Stabbing	Moss Park
D51	2004/07/24 04:00:00+00	Shooting	Moss Park
D51	2004/08/30 04:00:00+00	Other	Moss Park

Region: Mount Olive-Silverstone-Jamestown

DIVISION	OCC_DATE	HOMICIDE_TYPE	NEIGHBOURHOOD_158
D23	2004/02/21 05:00:00+00	Other	Mount Olive-Silverstone-Jamestown
D23	2004/05/01 04:00:00+00	Shooting	Mount Olive-Silverstone-Jamestown
D23	2005/02/12 05:00:00+00	Shooting	Mount Olive-Silverstone-Jamestown
D23	2005/08/03 04:00:00+00	Shooting	Mount Olive-Silverstone-Jamestown
D23	2005/10/22 04:00:00+00	Shooting	Mount Olive-Silverstone-Jamestown

Region: Glenfield-Jane Heights

DIVISION	OCC_DATE	HOMICIDE_TYPE	NEIGHBOURHOOD_158
D31	2005/05/05 04:00:00+00	Stabbing	Glenfield-Jane Heights
D31	2005/05/27 04:00:00+00	Stabbing	Glenfield-Jane Heights
D31	2005/08/30 04:00:00+00	Shooting	Glenfield-Jane Heights
D31	2005/09/13 04:00:00+00	Shooting	Glenfield-Jane Heights
D31	2005/12/23 05:00:00+00	Shooting	Glenfield-Jane Heights

Region: South Riverdale

DIVISION	OCC_DATE	HOMICIDE_TYPE	NEIGHBOURHOOD_158
D55	2004/10/08 04:00:00+00	Other	South Riverdale
D55	2008/01/17 05:00:00+00	Shooting	South Riverdale
D55	2008/03/18 04:00:00+00	Shooting	South Riverdale
D55	2008/10/25 04:00:00+00	Shooting	South Riverdale
D55	2009/04/17 04:00:00+00	Stabbing	South Riverdale

Region: Black Creek

DIVISION	OCC_DATE	HOMICIDE_TYPE	NEIGHBOURHOOD_158
D31	2004/05/15 04:00:00+00	Shooting	Black Creek
D31	2004/07/01 04:00:00+00	Shooting	Black Creek
D31	2005/01/26 05:00:00+00	Other	Black Creek
D31	2005/06/24 04:00:00+00	Shooting	Black Creek
D31	2005/09/10 04:00:00+00	Shooting	Black Creek