Data Analytics Capstone Project

Mid-point Documentation

CRIME ANALYSIS

Prepared for: Dong Ye

Prepared by: Group 2

Azim Ibrahim, Isaac Jimoh, Navneet Kaur, Nishaa Muthusamy

Northern Alberta Institute of Technology

DATA 3960 Section OA01

Date of Submission: Tuesday, March 7, 2023

Contents

List Of Figures

List Of Tables

Introduction

Crime analysis is a crucial component of law enforcement, which entails gathering and examining data to find trends and patterns in illegal activity. Data science and machine learning techniques can be used to analyze crime data, find trends, and make predictions that can aid law enforcement organizations in preventing crimes.

Our project aims to develop a system that can analyze and visualize crime data to assist law enforcement organizations in improved decision-making and resource allocation. In this initiative, crime data is collected, cleaned, and analyzed using statistical and machine learning methods.

We have divided our project in six following steps:

1. Collecting and cleaning crime data from various sources
2. Exploring the data and identifying patterns and trends using statistical techniques
3. Visualizing the data.
4. Developing models.
5. Validating the models and testing their accuracy.
6. Communicating the results and recommendations.

Figure 1: Steps in Crime Analysis

Background

Crime Severity Index

The traditional "crime rate" provides information on the number of police-reported incidents that have occurred for a given population. It measures the volume of crime coming to the attention of the police. The rate is simply a count of all criminal incidents reported to and by police divided by the population of interest. Each criminal incident, regardless of the type or seriousness of the offence, counts the same in the rate. For example, one homicide counts the same as one act of mischief.

The Crime Severity Index will, for the first time, enable Canadians to track changes in the severity of police-reported crime from year to year. It does so by considering not only the change in volume of a particular crime, but also the relative seriousness of that crime in comparison to other crimes.

During our initial data exploration of Crime Data Severity Index (from Statistics Canada) data, we noticed the crime severity index for major cities in Canada, Alberta using PowerBi:

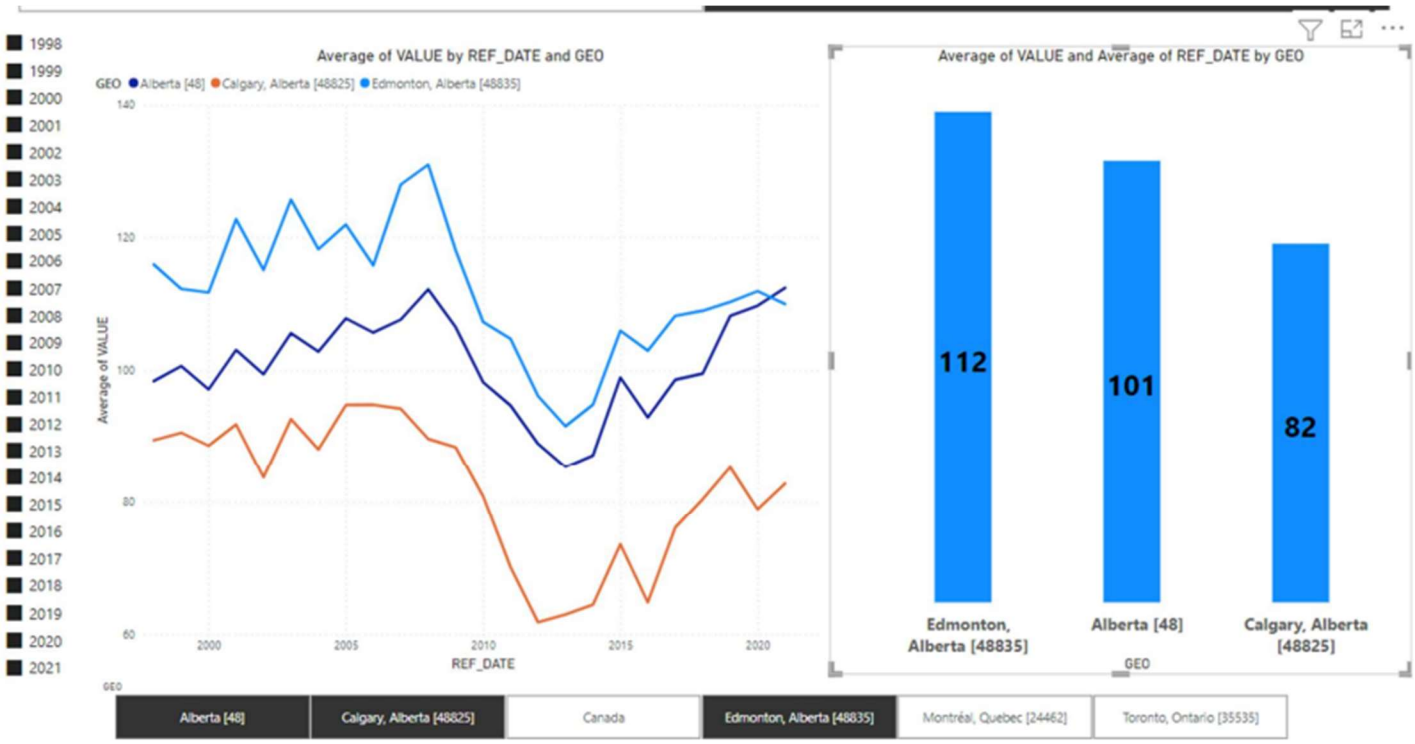

Figure 2:AVG Values of Violent Crime Index against Date.

Violent Crime

Violent crime includes any crime that incorporates force such as **murder**, **assault,** and **rape**, as well as crimes that include the **threat of violence such as robbery**, **harassment, and extortion**. The tracking of violent crime is usually done through indicators such as a violent crime rate or violent crime severity index. The violent crime rate is usually measured simply by counting all violent crimes per 1,000 inhabitants. The violent crime severity index is a measure of violent crime weighted by the severity of the crime. This indicator was developed to provide a clearer picture of serious crimes in Canada that can be hidden in the standard violent crime rate.

Problem Statement: Why does the City of Edmonton consistently have the highest Violent Crime Severity Index in Alberta, even higher than the overall Alberta Index?

Goal: Understanding the crime rate and patterns across the City of Edmonton.

Preliminary Research Questions:

- How are Violent Crimes distributed?
- Overall, between 2009 and 2019, What neighborhoods have the highest sum of occurrences?
- Do those neighborhoods show up as the top neighborhoods for each individual year between 2009 and 2019?
- How do the distinct types of crime fare between these identified neighborhoods?
- Map the neighborhoods via Geopandas and Folium to visualize the concentration of the Violent crimes.
- ML Modeling: Can we predict number of occurrences based on some input features like temperature, income, education status?

Dataset(s)

- Data collection
    1. Criminal Occurrences by Neighborhoods dataset is obtained from Edmonton Police Services website.
    2. Demographics data and Economic indicators datasets are collected from the City of Edmonton open portal.
    3. Weather data is gathered from acis.alberta.ca.
- Datasets merging
    1. Criminal occurrences dataset is merged with Edmonton neighborhoods for latitude and longitude based on Neighborhood names.
    2. Demographics datasets and economic indicators datasets are merged with above merged crime datasets based on neighborhood names.
    3. Weather dataset is merged with crime dataset by using month and year as key.

Challenges

Challenges faced during EDA process,

- Inconsistencies in the neighborhood name in different datasets. Neighborhoods which have few differences in their names are,
    1. CHAPELLE-CHAPELLE AREA
    2. HERITAGE VALLEY TOWN CENTRE-HERITAGE VALLEY TOWN CENTRE
    3. KESWICK-KESWICK AREA
    4. MCCONACHIE-MCCONACHIE AREA
- Edmonton city has renamed some of the old neighborhoods after redevelopment. For example, Edmonton Municipal airport is renamed as Blatchford area after being converted into sustainable building project.
- Edmonton city has developed new neighborhoods which do not have updated crime records. Some of the new neighborhoods are,
    1. THE UPLANDS
    2. DECOTEAU
    3. RIVER'S EDGE
    4. ASTER
    5. EDMONTON SOUTH CENTRAL
- Demographics data are not available for industrial neighborhoods, so, it would be difficult to predict crimes in those neighborhoods.

Why are we concentrating on 2016?

Although our initial dataset spans a decade from 2009 to 2019 and includes neighborhood crime data, we have chosen to concentrate on 2016 due to the availability of Income data by neighborhood for that year. By examining the demographics of each neighborhood through income data, we aim to gain better understanding of the relationship between socioeconomic status and crime rates. We also recognize that the year 2020 and beyond were impacted by the COVID-19 pandemic and may not be representative of typical crime rates, so we have chosen to limit our analysis to 2019. While we acknowledge that looking at a single year may limit our

analysis, we believe that focusing on 2016 will provide valuable insights into the factors that contribute to crime within each neighborhood.

Preliminary Analysis & Findings

Data exploration: Preliminary Investigation.

2009-2019; Violent Crimes filtered to: Assault, Homicide, Sexual Assaults and Robbery.

The following charts are based on the Dataset that includes the following:

Columns:

NGH_Name: Neighborhood Name (322 Distinct Neighborhoods)

NGH_Number: Neighborhood Number

Latitude: Latitude

Longitude: Longitude

Violation Type: Filtered to [Assault, Homicide, Sexual Assaults and Robbery].

Sum_Occurences: Number of Occurrences based on Violation Type

DT_Year: Year

DT_Month: Month

AVG_Temp: Average Temperature based on Month.

Shape of dataset: 112121 by 9 columns

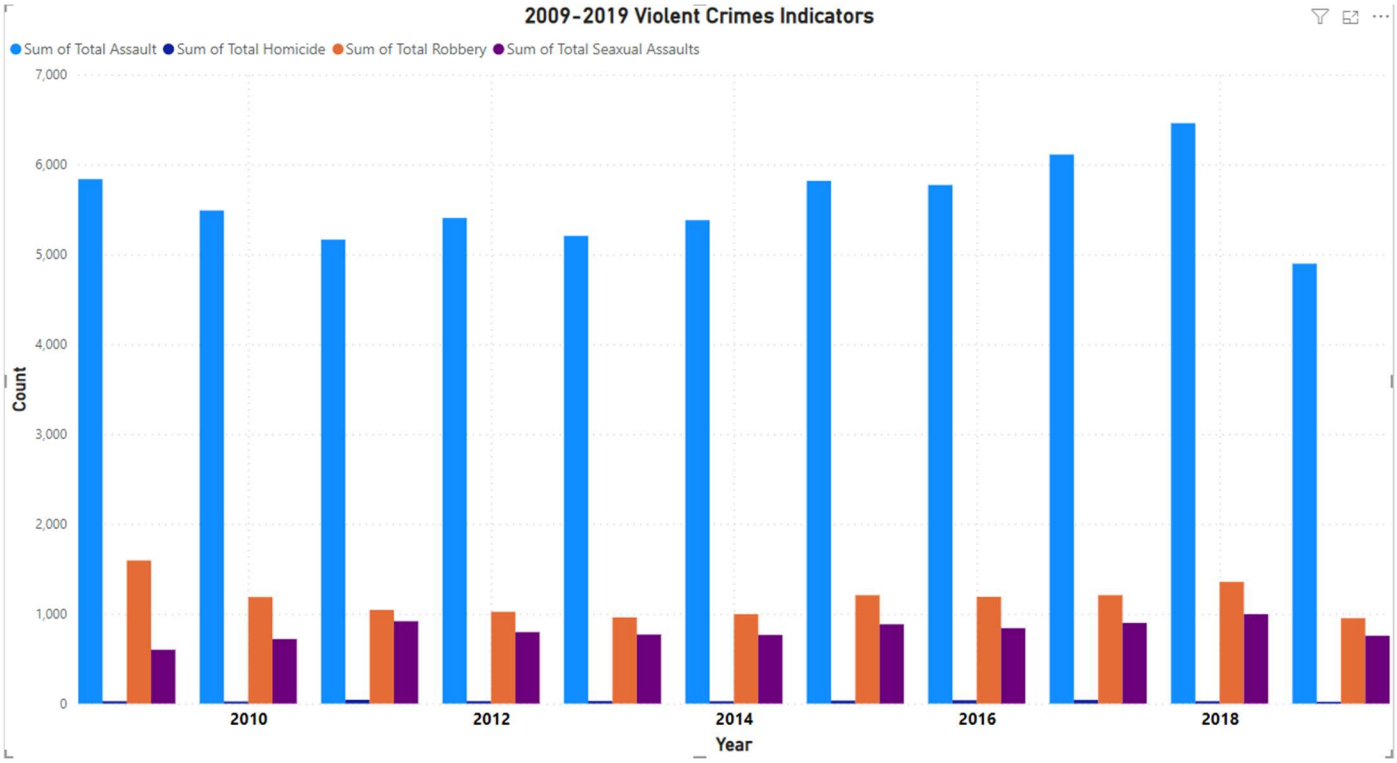

Figure 3: 2009-2019 Violent Crimes count by year.

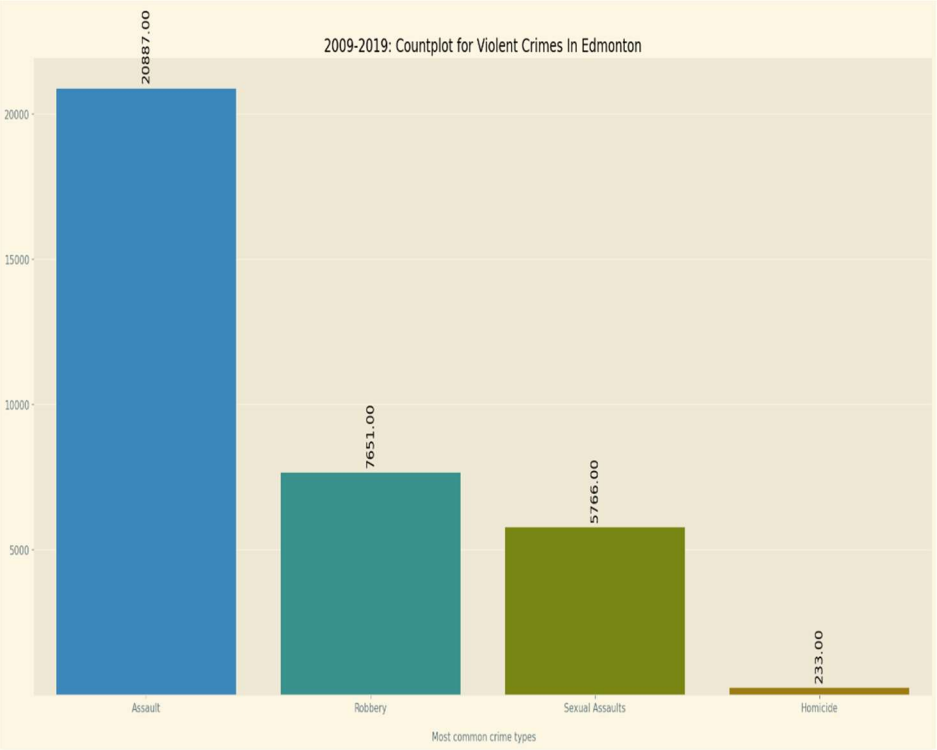

Figure 4:2009-2019 Count plot for Violent Crimes in Edmonton.

After filtering out the data to represent the type of crimes we want to base our investigation on, we were able to determine the count of violent crimes.

Since we live in Edmonton, which is the most northernly large Canadian city, we decided to take alook at how the violent crimes fare over the months between 2009-2019.

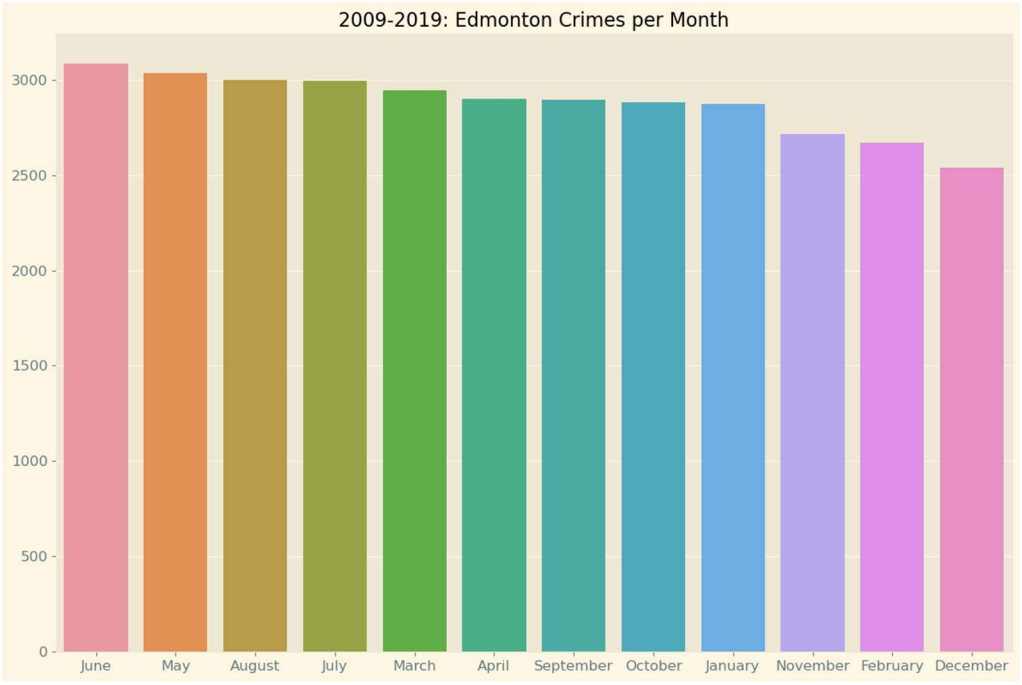

Figure 5: 2009-2019 Countplot for Violent Crimes by Month

As we can see from the chart all months have counts of above 2500. But between May and August we seem to be crossing the 3000 counts for each month.

It seems the spring and summer months are very popular with the criminals in Edmonton.

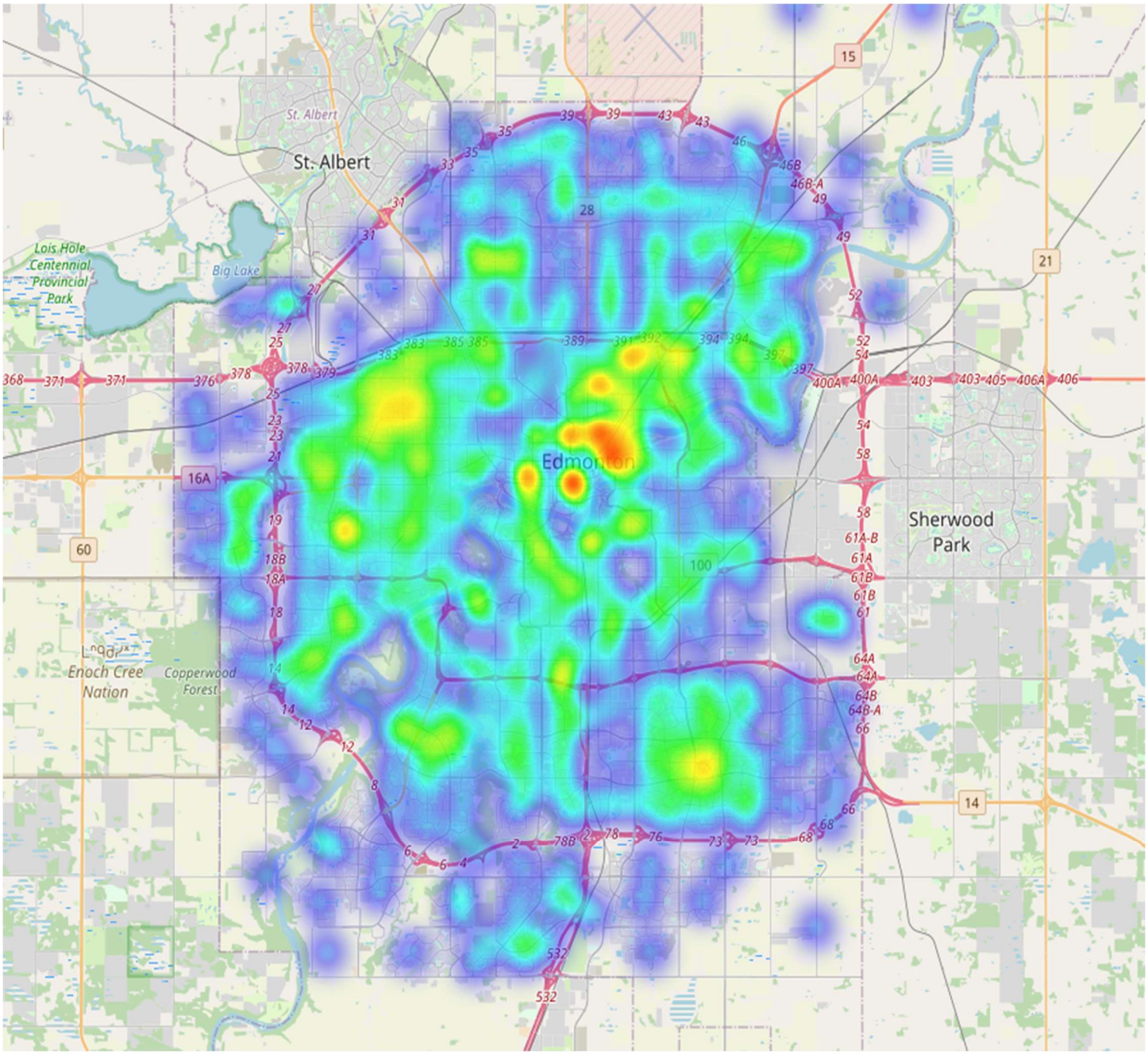

Figure 6: 2009-2019 Folium Heatmap

To visualize the special distribution of crime in the city, a Folium heat map was created to display the frequency of each neighborhood's appearance in the crime count data between 2009 and 2019. The intensity of the heat map reflects the relative number of crimes in each Edmonton neighborhood, with darker colors indicating higher crime rates.

In addition to the Folium heat map, Geopandas choropleth map was also created to visualize the special distribution of crime in Edmonton. The map displays the frequency of each neighborhood's appearance in the crime count data between 2009 and 2019.

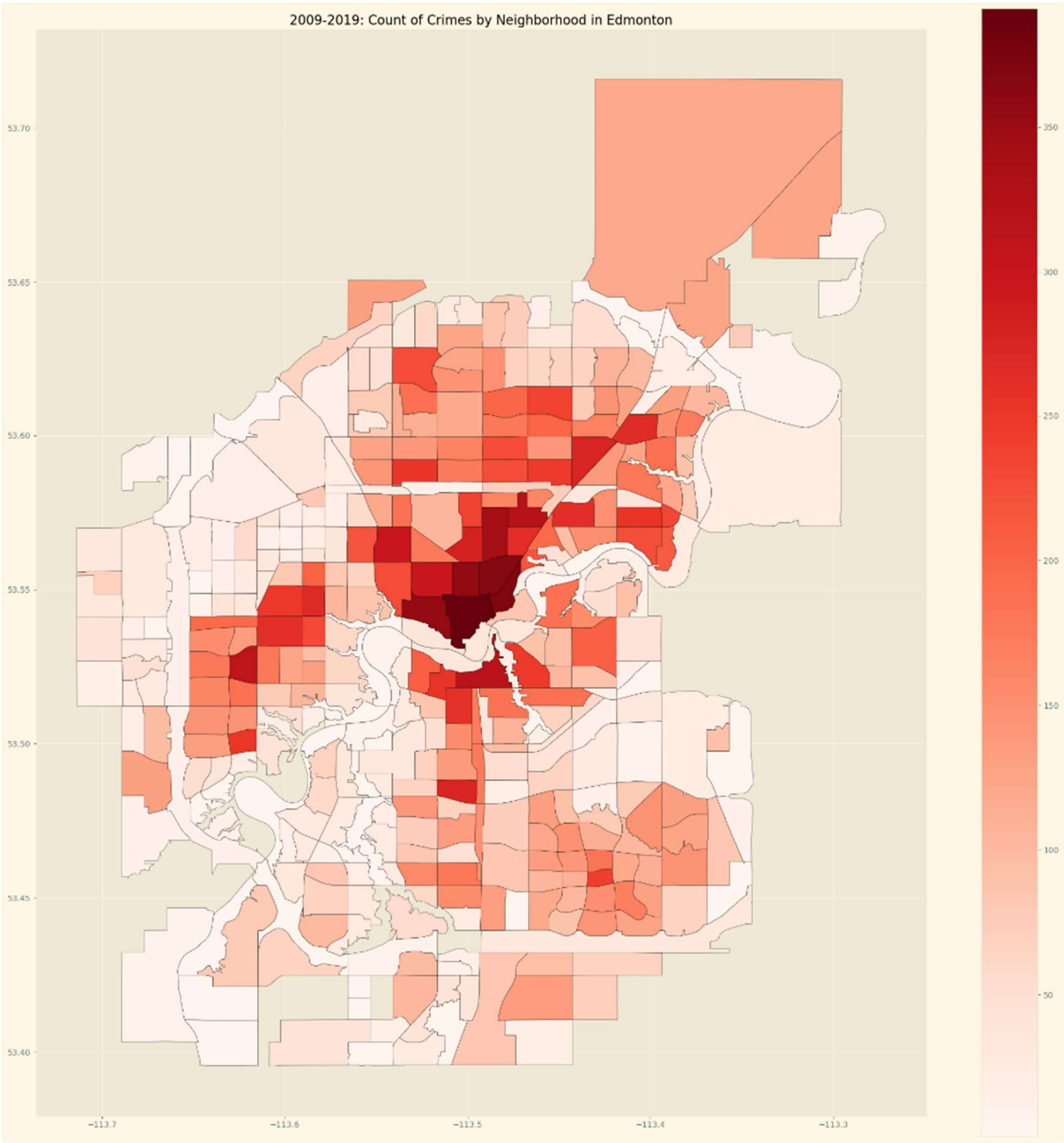

Figure 7: 2009-2019 Geopandas Choropleth Map for Number of Occurrences by Neighborhood.

After examining the concentration of the crime in the Folium heat map and Geopandas map, we identified the neighborhoods that appeared in the darker shades of the map. By plotting a count plot of the 6 top neighborhoods that had the six highest total count of Violent crimes, we found that the neighborhoods that

correspond to the darker areas on the heat map were: Downtown, McCauley, Central McDougall, Boyle Street, Oliver and finally Alberta Avenue.

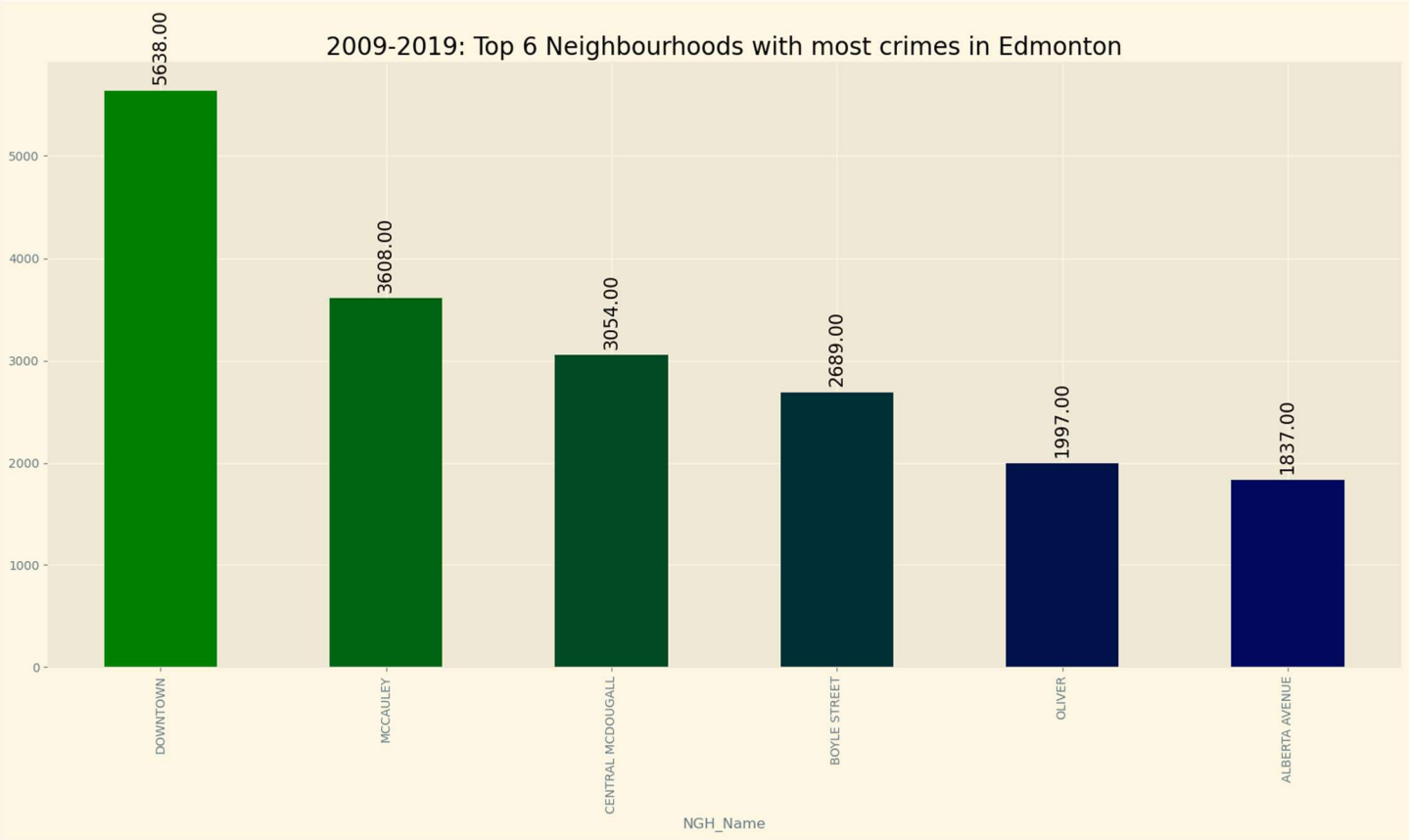

Figure 8: 2009-2019 Top 6 Neighborhoods with highest count of Violent Crimes.

The top 6 Neighborhoods hold 12.37% of the overall crimes between 2009-2019.

	Year	Percentage
0	2009.0	13.120206
1	2010.0	13.870981
2	2011.0	12.448829
3	2012.0	11.676364
4	2013.0	10.379999
5	2014.0	12.453222
6	2015.0	12.554543
7	2016.0	11.227348
8	2017.0	12.188557
9	2018.0	12.822894
10	2019.0	12.901033

Violation_Type	Assault	Homicide	Robbery	Sexual Assaults
Year				
2009	19.020682	15.789474	21.097322	15.524194
2010	19.723054	17.647059	19.434629	15.476190
2011	19.134577	21.212121	16.915423	18.898810
2012	17.425520	29.411765	20.845921	14.637904
2013	16.352941	33.333333	18.629550	14.785374
2014	19.055745	27.272727	17.972832	13.826367
2015	20.034996	24.000000	19.508058	18.648649
2016	18.979520	16.000000	17.631806	14.647887
2017	20.100083	16.000000	18.575064	13.825503
2018	21.341560	11.111111	16.437309	17.370892
2019	20.491493	15.384615	14.400806	15.186246

Figure 9: Violent Crimes Recorded for Top 6 Neighborhoods as a percentage of overall data.

To investigate further, we examined whether the same neighborhoods consistently appeared in the top 6 neighborhoods with the highest crime counts between 2009 and 2019. By analyzing the data and using PANDAS and Seaborn library, we were able to confirm that all six neighborhoods mentioned above consistently appeared at the top of the rankings, indicating their status as high crime areas over the course of several years.

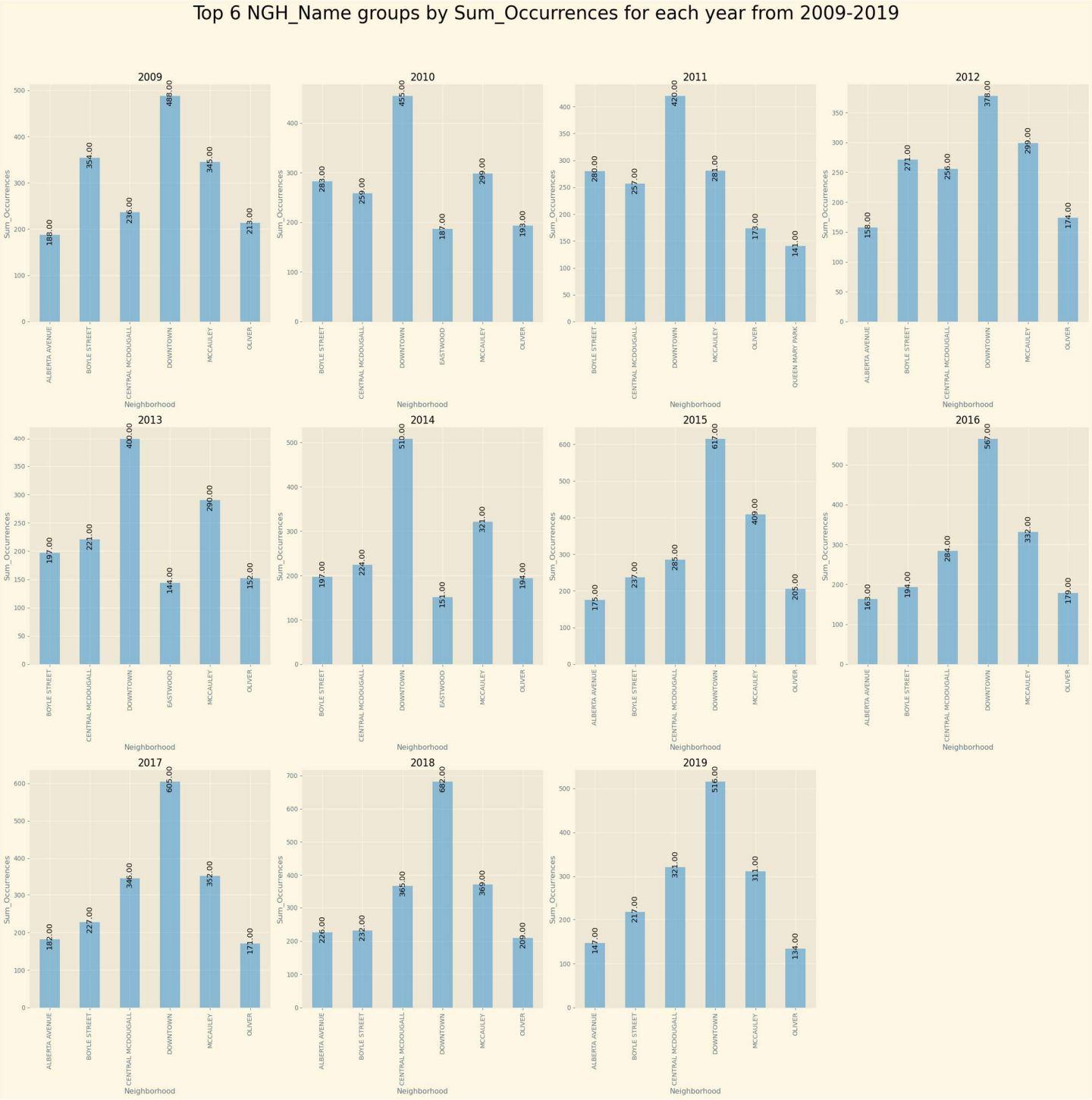

Figure 10: 2009-2019 Top 6 Neighborhoods with highest count of Violent Crimes for all years individually.

To gain a more detailed understanding of the characteristics of the high crime neighborhoods, individual visualizations for each of the top 6 neighborhoods over the specified time were created. The following pandas bar plot allowed us to analyze the patterns of violent crimes within each neighborhood between 2009 and 2019.

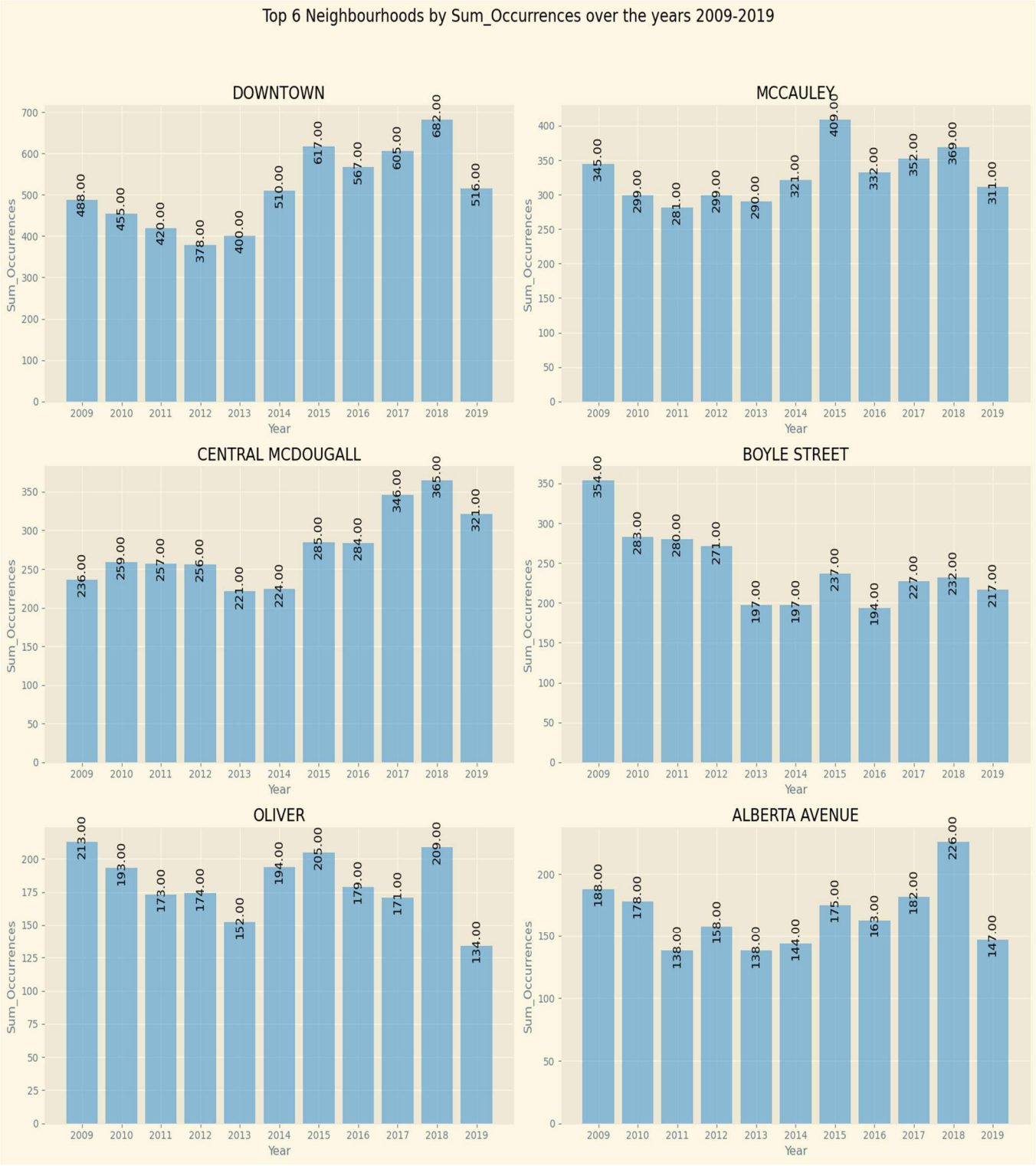

Figure 11: 2009-2019 Top 6 Neighborhoods and how they fared between 2009 and 2019.

To further investigate the patterns of violent crimes within the top 6 neighborhoods, we created Countplot for each type of violent crime. The violent crimes filtered to Assault, Homicide, Sexual Assault and Robbery. These Countplot, which were generated using seaborn, allowed is to analyze each type of violent crime within the neighborhoods.

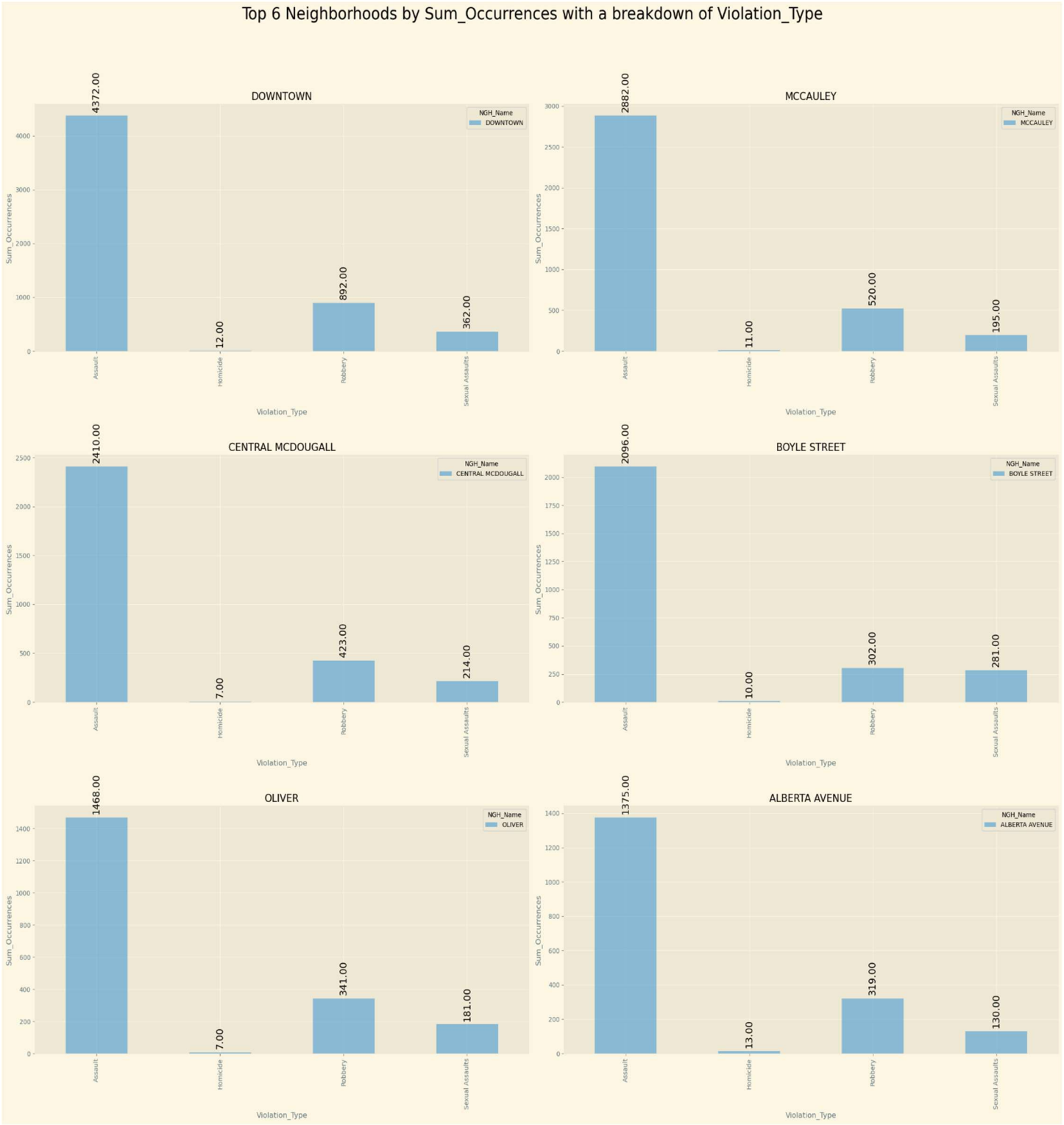

Figure 12: 2009-2019 Breakdown of Violation Types for each of the Top 6 neighborhoods.

To gain more meaning and understanding of the distributions of Violent crimes within the top 6 neighborhoods, individual Kernel Density plots for each type of filtered violent crime were created. These plots, which were generated using the seaborn library, allowed us to analyze the shape and density of each distribution within the top 6 neighborhoods. By examining the KDE plots for each neighborhood, differences, and similarities in the distribution of the crimes across the high crime areas could be identified.

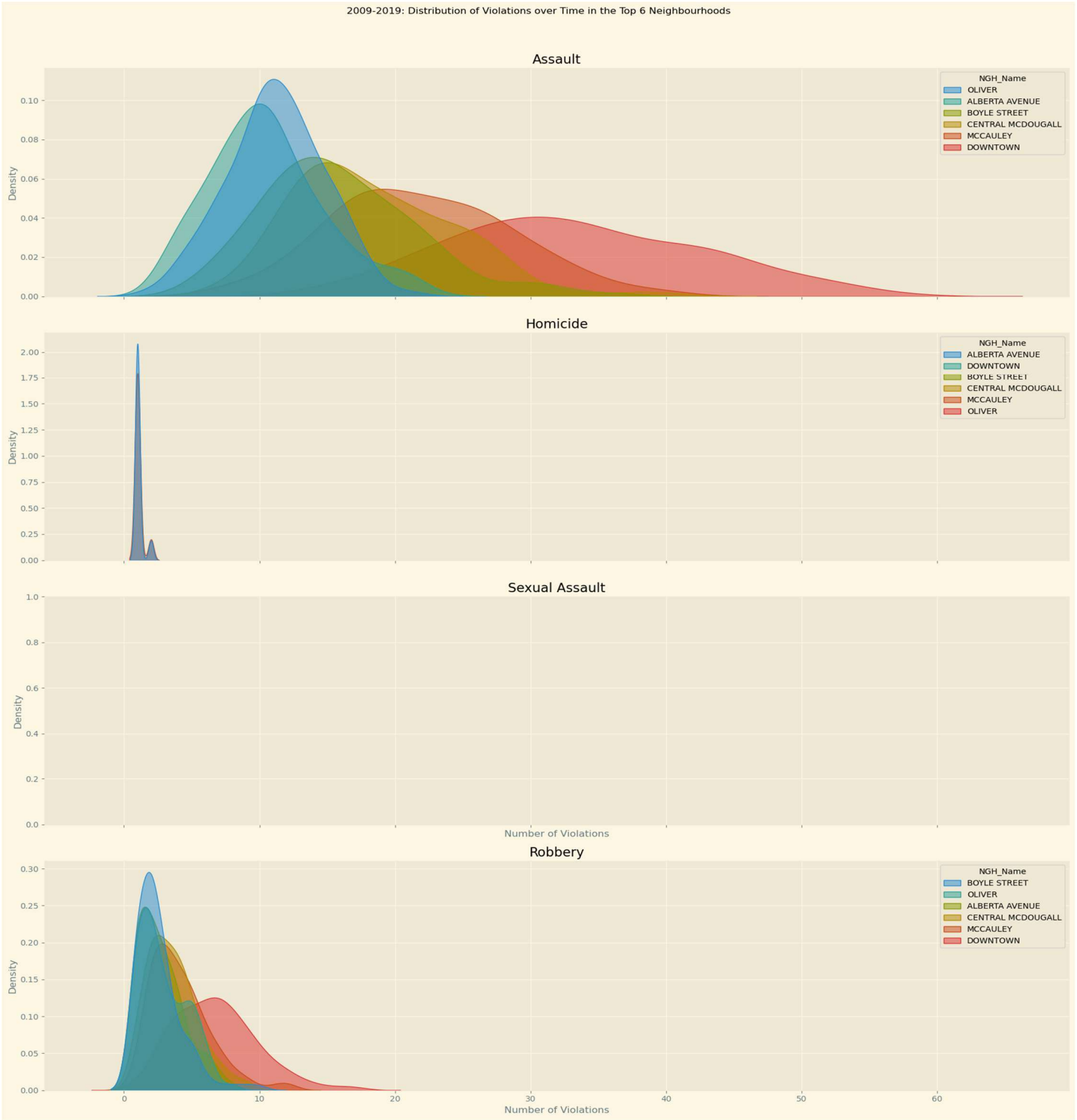

Figure 13: 2009-2019 KDE plots for each Violent Crimes based on the Top 6 Neighborhoods.

Modelling

Several models were explored to predict the number of occurrences based on various input features.

- Target – Number of occurrences
- Input features- Neighborhood Names, latitude and longitude, Type of crime, Occurrence Month, Air temperature, Education level of residents, Income Level of residents, Employment Status of Residents, Permanent Residents number
- The following models were trained on dataset and the models' performance was evaluated using the coefficient of determination (R^2) on the test set.

Various Models without optimization		
Model Name	Training scores(r2)	Testing scores(r2)
Linear Regression	0.508	0.479
Elastic Net	0.149	0.149
Random Forest Regressor	0.974	0.826
SGD	0.445	0.426
Extra tree	1.0	0.822
XGBoost	0.953	0.827
Gradient Boosting	0.874	0.828

Table 1:Evaluation of various models using r2 scores.

Model optimization

To further improve the performance of the model, hyperparameter tuning was performed using grid search. The hyperparameters of the model were optimized by evaluating different combinations of hyperparameters on our training dataset. The GridsearchCV function from the Scikit-Learn library was used to search for the optimal combination of hyperparameters. 5-fold cross-validation was used during the grid search to ensure the generalization performance of models. The performance of the tuned models was done using the evaluation metrics, the mean squared error (MSE) and the coefficient of determination (R^2) on the test set.

Model	Tuned Hyperparameters	Optimal Hyperparameters	Training scores(r2)	Testing Scores(r2)	Testing Scores(RMSE)
Random Forest Regressor	• 'bootstrap': [True, False] • 'max_features': ['sqrt']	• 'bootstrap': False • 'max_features': 'sqrt' • min_samples_leaf':2 • n_estimators': 500	0.814	0.706	1.7975

	• 'min_samples_leaf': [2,3,4] • 'n_estimators': [250,500]				
Gradient Boosting Regressor	• 'max_features': ['sqrt', 'log2'] • "n_estimators":[500] • "max_depth": [3,4,5] "learning_rate":[0.01,0.05,0.2] • "alpha" :[0.1,0.5]	• 'alpha': 0.1 • 'learning_rate': 0.05 • 'max_depth': 5 • 'max_features': 'sqrt' • 'n_estimators': 500	0.94	0.784	1.5413
XG Boost Regressor	• learning_rate: [0.05, 0.10,0.20, 0.30] • max_depth: [3, 4, 5] • min_child_weight: [1, 3, 5], • gamma: [0.1, 0.2] • colsample_bytree:[0.3, 0.5, 0.7]	• 'colsample_bytree': 0.7 • 'gamma': 0.2 • 'learning_rate': 0.1 • 'max_depth': 5 • 'min_child_weight': 5	0.893	0.83	1.346

Table 2:Models with hyperparameters tuning.

Feature selection-Random Forest

Feature selection was performed using the Random Forest regression model (Embedded Method). Random Forest is a powerful algorithm for feature selection because it can evaluate the importance of each feature in the dataset.

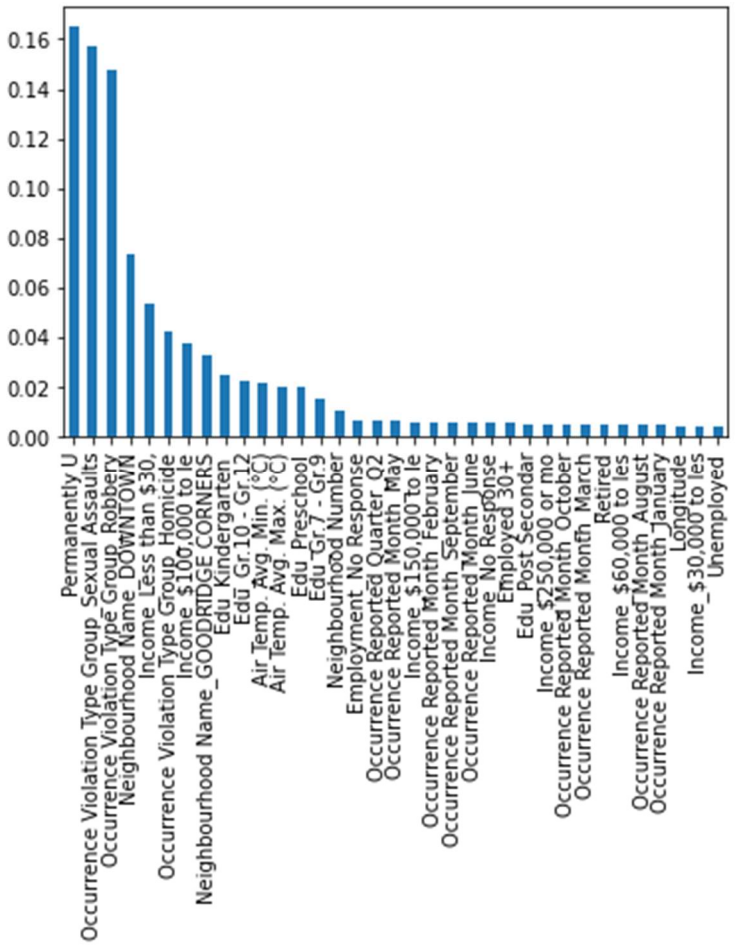

Figure 14:Feature Selection using Random Forest Regressor

Modelling Using Transformed Target Regressor

Target was positively skewed, so Transformed Target Regressor was used.

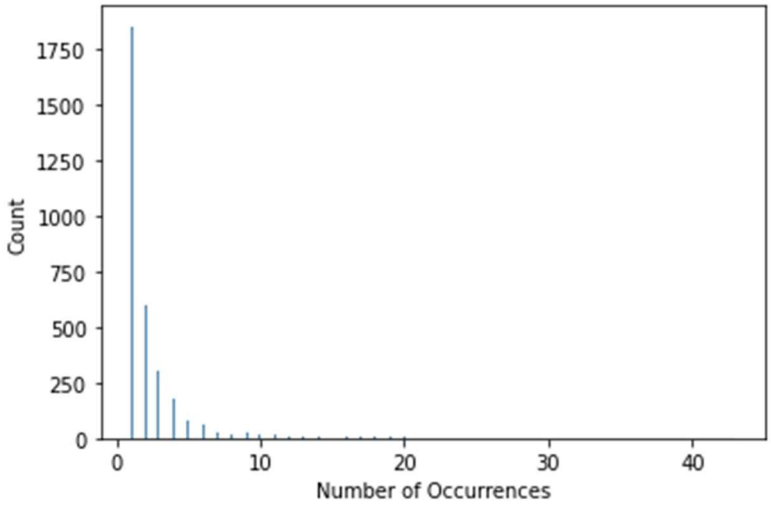

Figure 15:Distribution of Target variable

Various Models with Target Transformed Regressor		
Model Name	Training scores(r2)	Testing scores(r2)
Linear Regression	0.508	0.479
Elastic Net	0.149	0.149
Random Forest Regressor	0.974	0.826
SGD	0.445	0.426
Extra tree	1.0	0.822
XGBoost	0.953	0.827
Gradient Boosting	0.874	0.830

Table 3:Evaluation of models with Transformed Target Regressor

Model	Tuned Hyperparameters	Optimal Hyperparameters	Training scores(r2)	Testing Scores(r2)	Testing Scores(RMSE)
Random Forest Regressor	• 'bootstrap': [True, False] • 'max_features': ['sqrt'] • 'min_samples_leaf': [2,3,4] • 'n_estimators': [250,500]	'model__bootstrap': False 'model__max_features': 'sqrt' 'model__min_samples_leaf': 2 'model__n_estimators': 250}	0.805	0.699	1.818
Gradient Boosting Regressor	• 'max_features': ['sqrt', 'log2'] • "n_estimators":[500] • "max_depth":[3,4,5] "learning_rate":[0.01,0.05,0.2] • "alpha" :[0.1,0.5]	'model__colsample_bytree': 0.7 'model__gamma': 0.2 'model__learning_rate': 0.2 'model__max_depth': 5 'model__min_child_weight': 1	0.882	0.835	1.818
XG Boost Regressor	• learning_rate: [0.05, 0.10,0.20, 0.30] • max_depth: [3, 4, 5] • min_child_weight: [1, 3, 5], • gamma: [0.1, 0.2]	'model__alpha': 0.1 'model__learning_rate': 0.05 'model__max_depth': 4 'model__max_features': 'sqrt' 'model__n_estimators': 500	0.893	0.835	1.818

	• colsample_bytree:[0.3, 0.5, 0.7]				

Table 4:Hyperparameters tuning of models with Transformed Target Regressor

Feature Selection- Gradient Boosting

It was observed that using RFECV accuracy dropped to 0.166, resulting in only 3 optimal features.

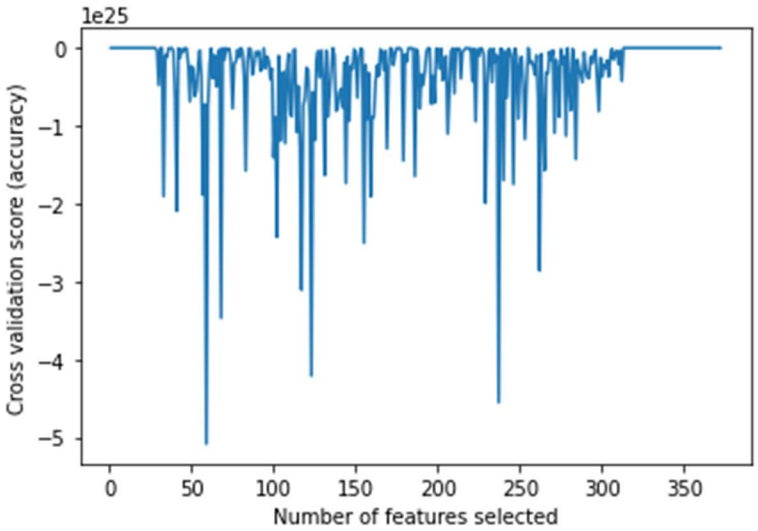

Figure 16:Feature Selection using RFECV

The same pattern was observed for XGBoosting. Testing accuracy dropped to 0.36.

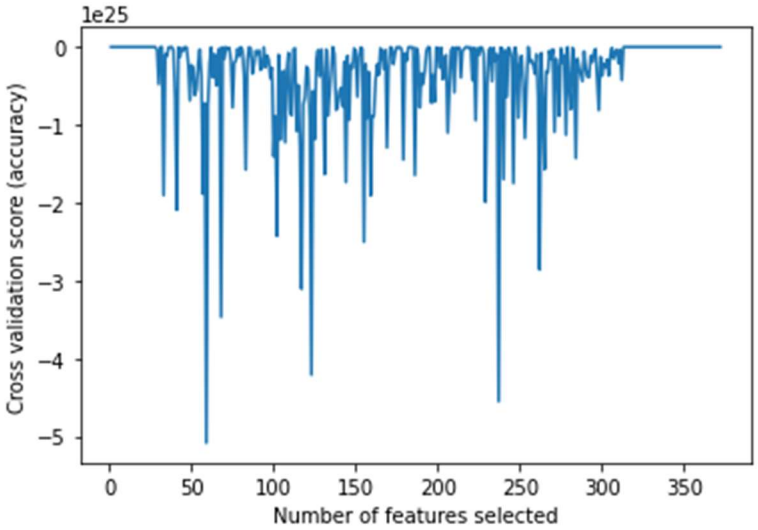

Figure 17: Feature Selection using RFECV(XG Boosting)

As accuracy was dropped with feature selection using RFECV, it was implied that the features that were chosen are not the most informative. RFECV chooses features based on how well they contribute to the performance of the model, but this metric may not always match the specifications of the job at hand. It's conceivable that the features that were eliminated were essential for attaining high accuracy.

Moving Forward

Updated Data: Yearly Analysis of crime trends

- Data collection
    1) Violent and non-violent crime occurrences data from 2009-2019 was obtained from Edmonton police Services website.
    2) Demographics data such as population, age, education level was gathered from alberta.ca open portal.
    3) Economic indicators such as unemployment rate, median household income, GDP was collected from Alberta.ca open portal and statistics Canada.
    4) Law enforcement factors such as number of police officers, EPS Budget was gathered from Edmonton police Services website.
- Data cleaning

  Some of the factors such as unemployment rate, education level census data are taken once every five years only. So, we had to impute the values for other years.
- Datasets joining

  Crime occurrences dataset is merged with population, education level, unemployment rate, median household income, GDP, EPS budget, EPS strength datasets with key being years from 2009 to 2019.

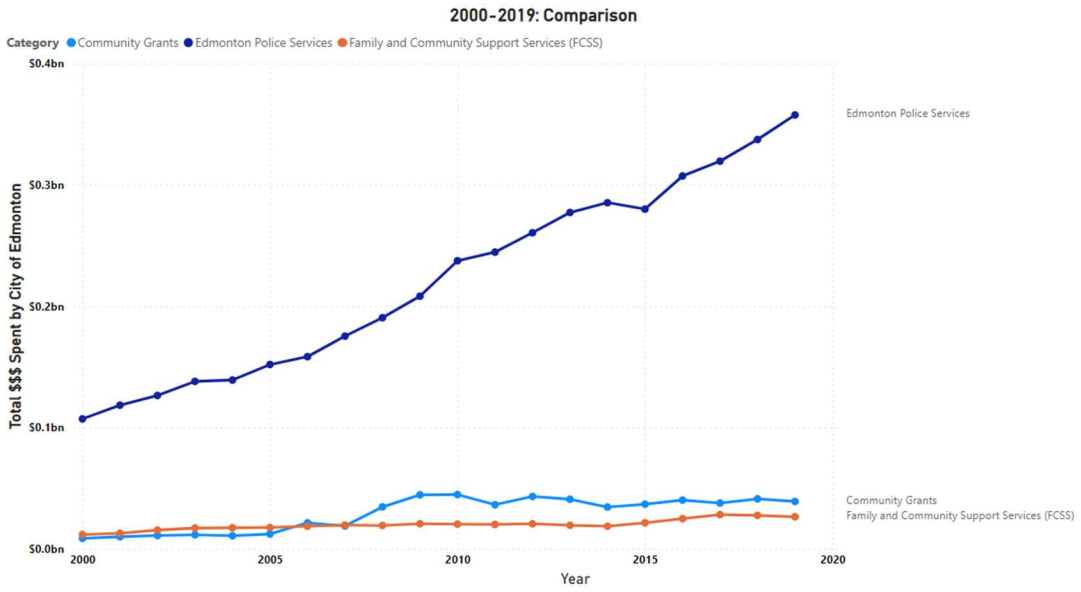

Figure 18: City of Edmonton Allocation

Statistical Analysis: We plan to perform some advanced statistical analysis to gain deeper insights into the relationships between different variables in the dataset. Specifically, we will explore whether there are any significant correlations between the number of occurrences of different types of crimes and other features such as temperature, neighborhood income levels, and education levels. For this we will use data from 2016. Additionally, we would like to investigate whether there are any significant differences in the number of occurrences of different types of crimes across different neighborhoods and years using techniques such as ANOVA and Chi-squared tests. Finally, we will examine whether there are any significant interactions between different variables, such as whether the relationship between temperature and crime varies depending on the neighborhood.

ANOVA and Chi-Squared tests are commonly used in analyzing crime data to determine if there is a significant difference in the mean values of a continuous variable (in our case the number of occurrences pertaining to Violent crimes, Assault, Sexual Assault, Homicide and Robbery) across different levels of a categorical variable (in our case, neighborhood, and type of violation). ANOVA is particularly useful when analyzing the relationship between two or more continuous variables, while Chi-squared is useful in analyzing the relationship between categorical variables.

Our group would focus on yearly analysis of crime trends in Edmonton. Our group would try to study correlation between prime environmental factors like unemployment rate, GDP, Age groups that can influence crime and crime related data to conduct analyses.

ML Modeling: Moving forward, we would like to also explore the use of clustering algorithms to group neighborhoods with similar trends. This could provide insights into the underlying factors that contribute to crime in specific areas. Additionally, time permitting, we could apply time series analysis techniques to better understand the patterns and trends of crime occurrences over time. This could help us identify seasonal and long-term trends, as well as potential factors that contribute to changes in crime rates.

Bringing it all together, by combining these approaches with the statistical analysis (ANOVA, Chi-Squared hypotheses testing), we can gain a more comprehensive understanding of the factors that contribute to crime in the City of Edmonton.

References

1) https://www.statista.com/statistics/436253/violent-crime-severity-index-in-canada-by-province/#:~:text=Violent%20crime%20includes%20any%20crime,or%20violent%20crime%20severity%20index.