# Analysis of Crime in Seattle Since 2008

Maria Berova, Archit Ganapule, Aileen Kuang, Arjun Naik, Maitreyi Parakh

February 19, 2023

## Introduction

In recent years, crime in Seattle has increased; this raises the question: what types of criminal offenses have increased, and what types have decreased? To address this, we created a visualization of Seattle crime data from 2008 and 2022.

Furthermore, an increase in crime has a significant impact on community members and the Seattle Police Department (SPD). Thus, we also create a **machine-learning model** that predicts crime (for the next three months, though this is scalable) based on the type of offense in order to assist SPD with their work.

Overall, we aimed to both provide evidence of and explanations for specific patterns in the data, as well as be able to predict future crime occurrences on specific dates in relation to specific crimes. We worked on both data analysis & visualization and machine learning.

## *(DATA VISUALIZATION)* Rise and Fall of Criminal Offenses

In this section, we explain and analyze our visualizations of the rise and fall of offenses based on the crime subcategory, the offense parent group, and offense description.

### Methods

We used the provided SPD data and ran two programs to find the totals of the Crimes Against, Offense Parent Groups, and Offenses columns. We removed extraneous columns and used PowerBI to create graphs displaying trends in the Crimes Against category. Furthermore, we used Google Sheets to analyze specific Offense Parent Groups and Offenses, aiming to uncover trends in our graphs. Our Google Sheets visualization focuses on 2008, 2013, 2018, and 2022 in order to hone our analysis of specific trends over these years.

### Graphs

Through these methods, we created various charts in PowerBI and Google Sheets. Our data visualization in PowerBI depicts crime in Seattle from 2008 to 2023 by the crime against category; two examples are given in Fig. 1 & 2, and the link to the PowerBI report is here. The link to the models made in Google Sheets are given here.

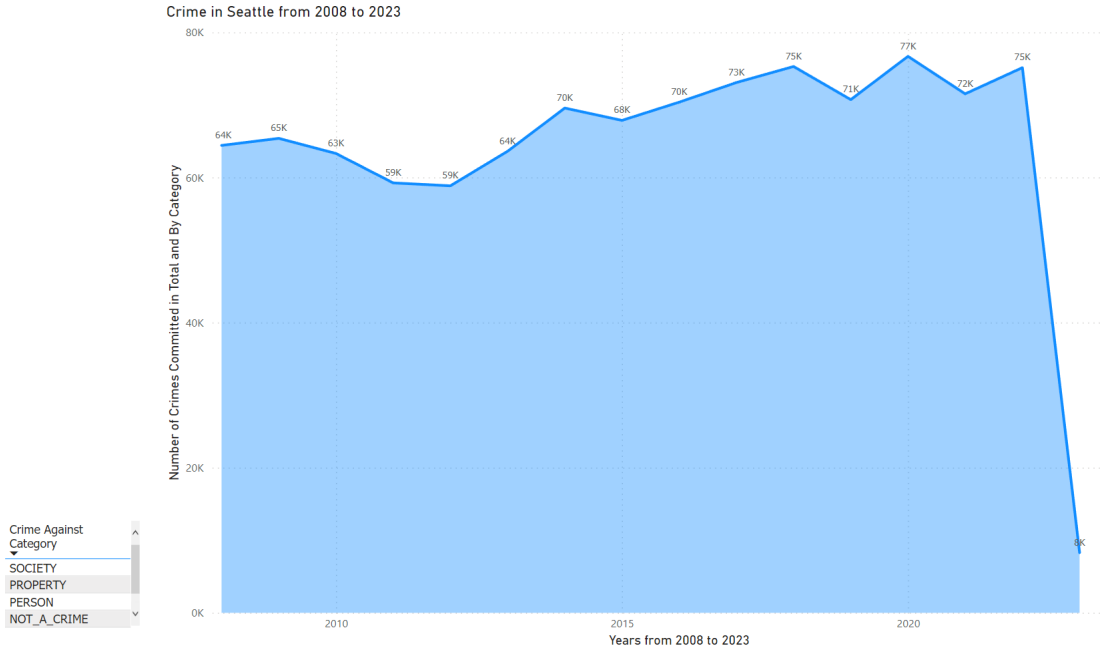


Figure 1. Crime (in general) in Seattle from 2008-2023.

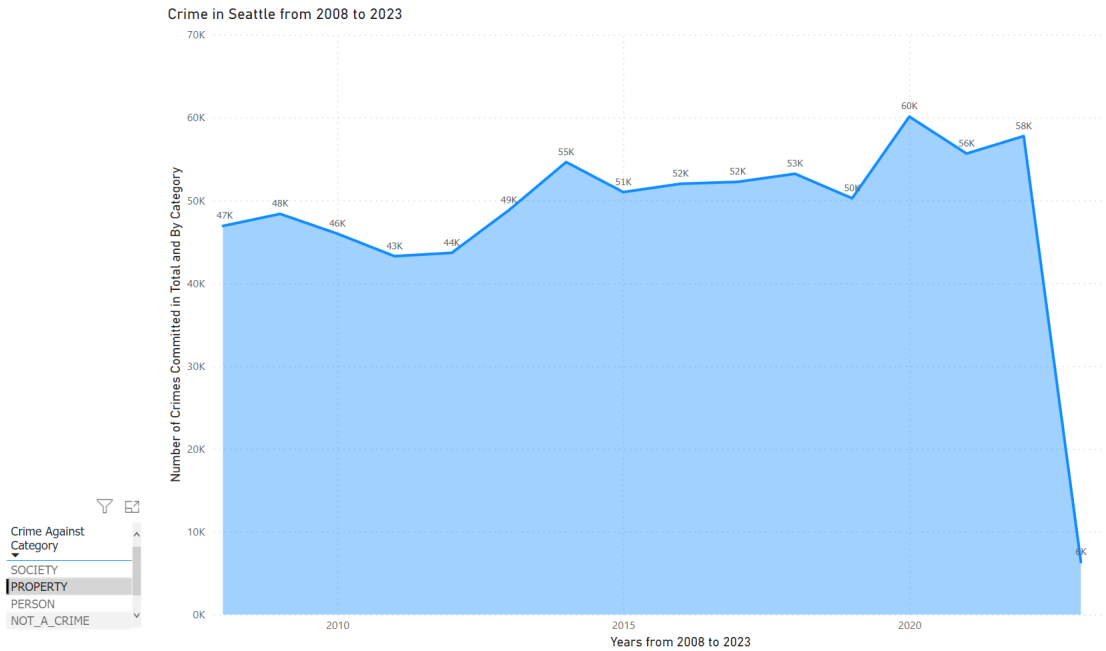


Figure 2. Crime against property in Seattle from 2008-2023.

# *(MACHINE LEARNING)* URDAD (Utilizing Regression for Date Aggregate Datasets)

Although this model is not perfect, we developed an original neural-network-like approach to model complex data. We organized the data by "layers" of time, which each contributed their own transforms to a basic prediction. These "layers" resemble a classic neural network layer approach.

This section documents the design of URDAD (Utilizing Regression for Date Aggregate Datasets). The purpose of our implementation was to make a machine learning model based off of extensive past Seattle crime data to predict future crime rates at a given date.

URDAD is structured around "layers" of date specifics.

- The **top layer** is the year. A polynomial regression is used to predict the total amount of a specific crime committed in a given future year.
- The **second layer** is the month. We calculated the percent of the specified crime in each month that made up the aggregate crime count in each year, averaged each month's percent values to come up with one average contribution (given as a fraction).
- The **bottom layer** is the day. We calculated the percent that each day made up in the aggregate crime (of that type) for each month, then averaged each day's percent values to come up with one average contribution per month (given as a fraction).

The outline for calculating the crimes committed on a given date (year, month, day) is given as follows:

1. Use the top layer to predict the total number of crime committed that year. For sake of convenience, call this number $N$.
2. Obtain the given month's percent contribution from the second layer (call this fraction $M$). $N \times M$ will then be the amount of crime predicted to occur in the given month.
3. Obtain the given day's percent contribution to the month from the bottom layer (call this fraction $D$). $N \times M \times D$ will then be the amount of crime predicted to occur on the given day.

## Manipulating Layers of Time Data

### Motivation

We decided to model trends over the years, months, and days separately from each other, then combine these trends to make an accurate prediction. This allows us to reflect specific trends of crime over months, and over days of the month, in our predictions. (For example, there could be more robberies on Black Friday, which would be reflected in how we model days in a month). It also prevents overfitting data by making general models for different layers.

### Prepping the Data

We separated the data into three "layers" of dataframes as follows:

Layer 1: Years. There was one dataframe structured as below:

| | Offense |
|---|---|
| Year 1 | # Occurrences |
| ... | ... |
| Year $n$ | # Occurrences |

Layer 2: Months. There was one dataframe structured as below:

| | | Offense |
|---|---|---|
| Month 1 | Year 1 | % Occurrences |
| | ... | ... |
| | Year $n$ | % Occurrences |
| | ... | ... |
| Month 12 | Year 1 | % Occurrences |
| | ... | ... |
| | Year $n$ | % Occurrences |

Layer 3: Days. There were 12 dataframes structured as below, with one dataframe per month:

| | | Offense |
|---|---|---|
| Day 1 | Year 1 | % Occurrences |
| | ... | ... |
| | Year $n$ | % Occurrences |
| | ... | ... |
| Day $m$ | Year 1 | % Occurrences |
| | ... | ... |
| | Year $n$ | % Occurrences |

Then, we averaged the % Occurrences values over all the years for layers 2 and 3. Layers 2 and 3 ended up looking as follows:

Layer 2: Months.

| | Offense |
|---|---|
| Month 1 | Average % Occurrences |
| ... | ... |
| Month 12 | Average % Occurrences |

Layer 3: Days.

| | Offense |
|---|---|
| Day 1 | Average % Occurrences |
| ... | ... |
| Day $m$ | Average % Occurrences |

This gives us a reliable way of tracking each month and day of the month's relative contribution to a year's total crime count.

**Layer Implementation**

The "top layer", or the year, is crucial to making future predictions. We fit a polynomial model to past crime occurrence data by gradient descent. [After subsequent train-test splits and refinements, we determined that an n-degree polynomial was best suited for our model.]

The "second layer", or the month, is essentially a model of each month's contribution to the total crime count of a year. We first calculated the percent of each year's crimes committed in a given month for every month. Then, we averaged the percents we obtained over all the years sampled. Overall, this yielded the **average percentage of a year's crimes committed in each given month**.

The "bottom layer," or the day, is essentially a model of each day's contribution to the total crime count of a month. We first calculated the percent of each month's crimes committed on a given day for every day of every month, over all of the years. Then, we averaged the percents we obtained over all the years sampled. Overall, this yielded the **average percentage of a month's crimes committed on a given day**.

## Polynomial Regression Implementation

We used a polynomial regression model to predict the total number of crimes in future years. The main challenge with this model was choosing the optimized degree of the polynomial to implement. To tackle this, we split our data into train and test subsets. We iterate through degrees of the polynomial, from one to ten. On each iteration, we create a new polynomial model based on the train subset. Then we calculate the MSE (mean-squared error) for the model compared to the train subset, and the MSE compared to the test subset. Finally, we take the difference between these two MSEs. Throughout the iterations, we keep track of the minimal MSE difference, continuously updating it, along with the related polynomial model. After the iterations, we are left with the optimized polynomial model for the data.

## Scalability

An important consideration when designing our model was scalability. We created separate trends for days, months, and years in order to avoid overfitting for specific dates. The top layer uses a carefully considered model to make future predictions, so it is viable for many years in the future. Thus, it can be extended beyond the required "3 months into the future" prediction from the Datathon prompt.

Our implementation and parsing of the data uses optimized functions that mostly operate on linear time, or O(a*n). On extremely large datasets, it conducts calculations within seconds. Additionally, our layered method for predicting the frequency of a certain crime or crime category on a specific day makes the overall process faster - regression is only calculated for the year data which both highlights larger trends and allows the rest of the computation to be done almost instantly (based on pre-determined average values for monthly and daily percents).

## Datathon-Specific Implementation

An important consideration when designing our model was scalability. We created separate trends for days, months, and years in order to avoid overfitting for specific dates. The top layer uses a carefully considered model to make future predictions, so it is viable for many years in the future. Thus, it can be extended beyond the required "3 months into the future" prediction from the Datathon prompt.

Our implementation and parsing of the data uses optimized functions that mostly operate on linear time, or O(a*n). On extremely large datasets, it conducts calculations within seconds. Additionally, our layered method for predicting the frequency of a certain crime or crime category on a specific day makes the overall process faster - regression is only calculated for the year data which both highlights larger trends and allows the rest of the computation to be done almost instantly (based on pre-determined average values for monthly and daily percents).

## Instructions for Usage

To use the implemented model, use the *predictCrime*() method in the linearregression notebook in the "notebooks" folder in our github repo. The method takes the name of the crime (either of a parent category or specific offense), snf the year, month, and day. The final input is a string: set it to "po" if you want statistics on specific crime names, and to "poopg" if you want statistics on parent crime categories.

The method will return a tuple: [the predicted number of the specified crime on the specified day, margin of error].

We found that our margin of error was quite large for most runs. This error may have stemmed from rounding: when taking averages of data, or otherwise modifying it to use in our implementation, values tended to get very close to zero. This could result in said values simply being rounded to zero, and thus not being considered in our modeling.

## Further Inquiry

One of the main features we could implement to improve this model is polynomial regression not only for year trends, but for month and day trends. Recall that we used a polynomial regression model to predict the amount of crimes committed in future years. We could do the same for predicting future percentages of crime for each month and each day of each month. However, implementing this would be challenging because we would need a separate polynomial model for each month, and each day of the month – leading to up to 378 additional polynomial models to create and optimize! In this case, a more time-efficient approach to optimizing a polynomial model would be needed.

## Data Visualization Analysis

We find that offense counts within parent groups remain relatively steady. The 2008 model in Google Sheets demonstrates the primary forms of offenses and offense parent groups, where larceny-theft is the most prevalent. As we see in the 2013, 2018, and 2022 models (which are extremely similar to one another, with slight fluctuations that can be explained by significant events occuring during those years), these counts do not change much. This indicates that there was no rise or fall within types of criminal offenses from 2008-2022.

Additionally, in the model showing crime against society from 2008-2023, there is a drop in crimes in 2011. During this year, SPD focused on restricting crime throughout Seattle for whiter, wealthier citizens, which may explain the sudden change in criminal offense counts that targeted those of color [1]. As police began to divert their attention away from enforcing misdemeanors, crime rates increased until their all-time high up until 2014 [2]. SPD's criticism of force led to lower misdemeanor rates, reducing targeted populations' fear of breaking the law. Since then, crime has steadily increased, with a few dips explained by alternate circumstances. The role of punishment in the justice system for certain crimes is a highly hypothesized theory, which may have been present in the fluctuations during these few years [3].

Another interesting fluctuation is present throughout the crimes in 2020. Although there was a slight increase in overall crimes and in crimes against property, crimes against society and people were lower than they had been throughout our data set. Throughout 2020, political tensions were heightened and several remarkable incidents affected criminal offenses. Firstly, the Black Lives Matter movement incited riots and property damage throughout the country and not just within Seattle. Crimes against society dropped, but homicides (especially targeting black men) increased 48.57% since 2019 [4] [5]. Violent crime against homeless people was also further exacerbated due to the heightened rates of eviction and the growth in Seattle's already large, predominantly Black and Hispanic population [6]. Homeless-related events for violent crimes in shootings and shots fired increased by 122% and Seattle residents' primary concerns included homelessness.[7]

All in all, our data visualizations indicate that there was no change in rise or fall in types of crimes and provide avenues of interesting interdisciplinary analysis.

## Conclusion

This project has explored the rise/fall of crime in Seattle through data visualization tools. Through our models, we have found that there was no significant rise or fall within types of criminal offenses from 2008-2022. Moreover, we have identified interesting insights about crime rates (ex. the sudden drop in crimes in 2011) and have provided potential explanations for each insight.

We have also created a machine-learning model in order to predict crime in Seattle for the next three months (and beyond). Through this model, we found that data depending on multiple types of time categories (e.g., years, months, and days) requires a layered approach to recognizing trends. We found "detangling" such data and separating it into separate trends helped get more accurate and less overfitted results. This information can be used in future, similar endeavors.

[1] https://www.capitolhillseattle.com/2011/07/survey-east-precinct-residents-think-spd-misconduct-is-a-problem-crime-getting-worse/

[2] https://www.seattletimes.com/seattle-news/report-cites-plunge-in-spd-enforcement-of-low-level-crime/

[3] https://www.house.mn.gov/hrd/pubs/deterrence.pdf

[4] https://www.seattlepi.com/local/seattlenews/article/2020-crime-Seattle-highest-homicide-rate-15864266.php

[5] https://www.seattle.gov/documents/Departments/Police/Reports/2021_SPD_CRIME_REPORT_FINAL.pdf

[6] https://kingcounty.gov/depts/health/covid-19/data/homeless.aspx

[7] https://www.seattle.gov/documents/Departments/Police/Reports/2021_SPD_CRIME_REPORT_FINAL.pdf

Although we did encounter issues in the specifics of our dataframe traversal, we laid out a solid skeleton and framework for approaching data topics that deal with predictions on specific dates (URDAD). We believe this can be scaled to areas outside of crime rate analysis.