Seattle Crime Analysis - 2024 Overview

Team 4 - Yousef, Alysia, Shraddha & Trevor

Executive Summary

This report is an analysis of crime patterns in Seattle for 2024, using public crime records from the Seattle Police Department. Our objective was to identify the most frequent types of crime, where those crimes were located and determine the time and days these incidents occur. It also presents an analysis with the objective of forecasting daily incident counts using regression models. Exploratory data analysis revealed key patterns in crime distribution across offense categories, time of day, day of week, month, and geographical areas. By showcasing these patterns, our goal is to create data-driven strategies for law enforcement resources allocation and community safety planning.

Goal:

Our goal is to identify the most frequent types of crime, identify areas with the highest level of crime, and help plan where to send police and resources.

Description of Dataset:

The dataset was obtained from the Seattle Police Department's public crime records, which document incidents from 2008 to present day. It contains detailed information for each report, including the date and time of the offense, type of crime, location coordinates, neighborhood, precinct, and crime classification categories such as Offense Category, Offense Sub Category, NIBRS Offense Code Description, Crime Against Category, and Offense Parent Group. The

The dataset contained geographical coordinates (Latitude and Longitude), which could be useful for mapping crime hotspots, and possibly giving us an idea as to where to send police officers.

Data Preparation:

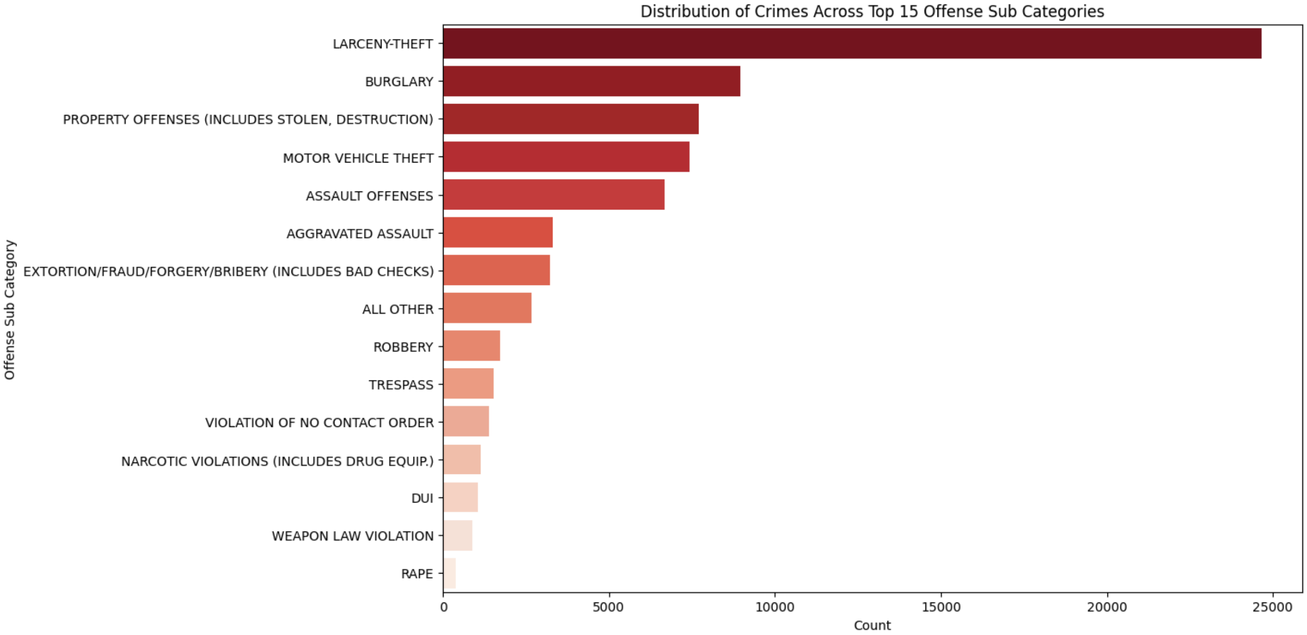
The initial dataset contained over 1.5 million rows, which we planned to analyze, however ran into issues when uploading the file into Google Collab. Therefore, we had to filter the data to the 2024 year to not exceed Collab requirements, and enough to explore crime patterns, identify high-risk areas, and generate actionable insights. Additionally, we had to clean and format our dataset by making the link between report dates and offense dates clear. We had to get rid of columns like Offense Id and Report number which were unique columns in our dataset and were not useful to study any correlations. Lastly, to handle missing and redacted values, we cleaned rows that contained incomplete information to ensure consistency.

Analysis Objective:

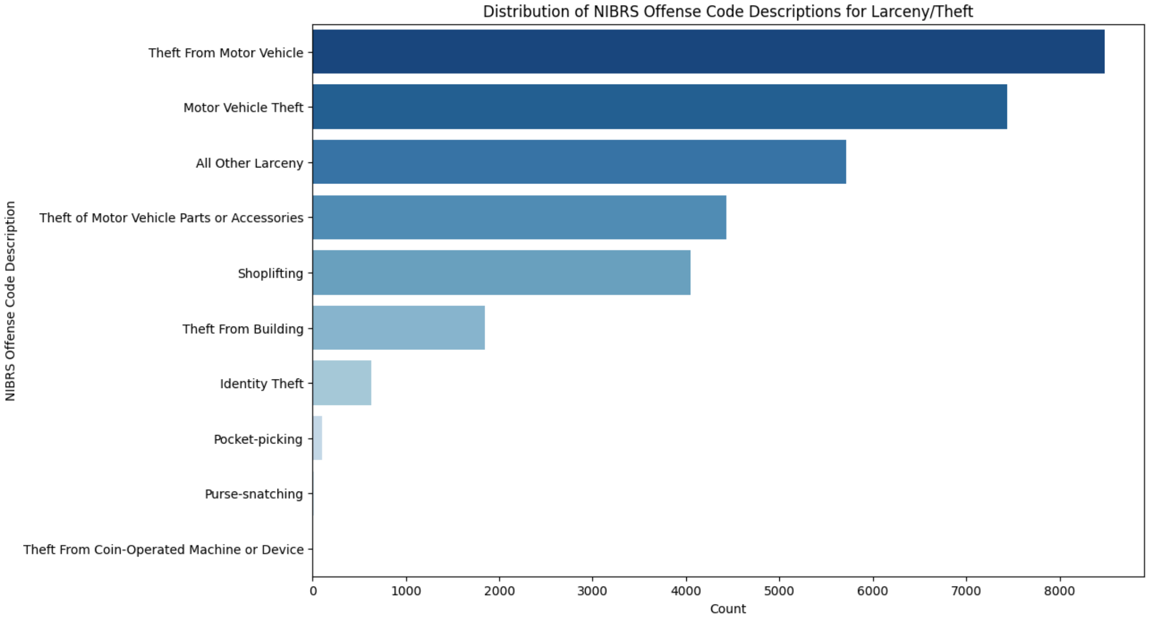
We have three main objectives for our analysis: First, to identify the **most frequent types of crime**. Second, to identify the areas with **the highest levels of crime**. Last, to determine **the times and days** when crime is most prevalent. With these objectives in mind, we can help plan where to send police and resources, create benefits to the public, and suggest recommendations to meet specific community needs.

Visualizations & Insights:

Distribution of Crimes by Top 15 Offense Categories: The chart shows the distribution of crimes based on the top 15 crime subcategories, with Larceny/Theft dominating at nearly 25,000 cases, followed by Burglary, Property Offenses, Motor Vehicle Theft, and Assault. Less frequent crimes include DUI, Weapon Law Violation, and Rape. Darker red shades indicate a higher count. This breakdown can help prioritize law enforcement focus and resource allocation toward the most prevalent offenses to reduce overall crime rates.

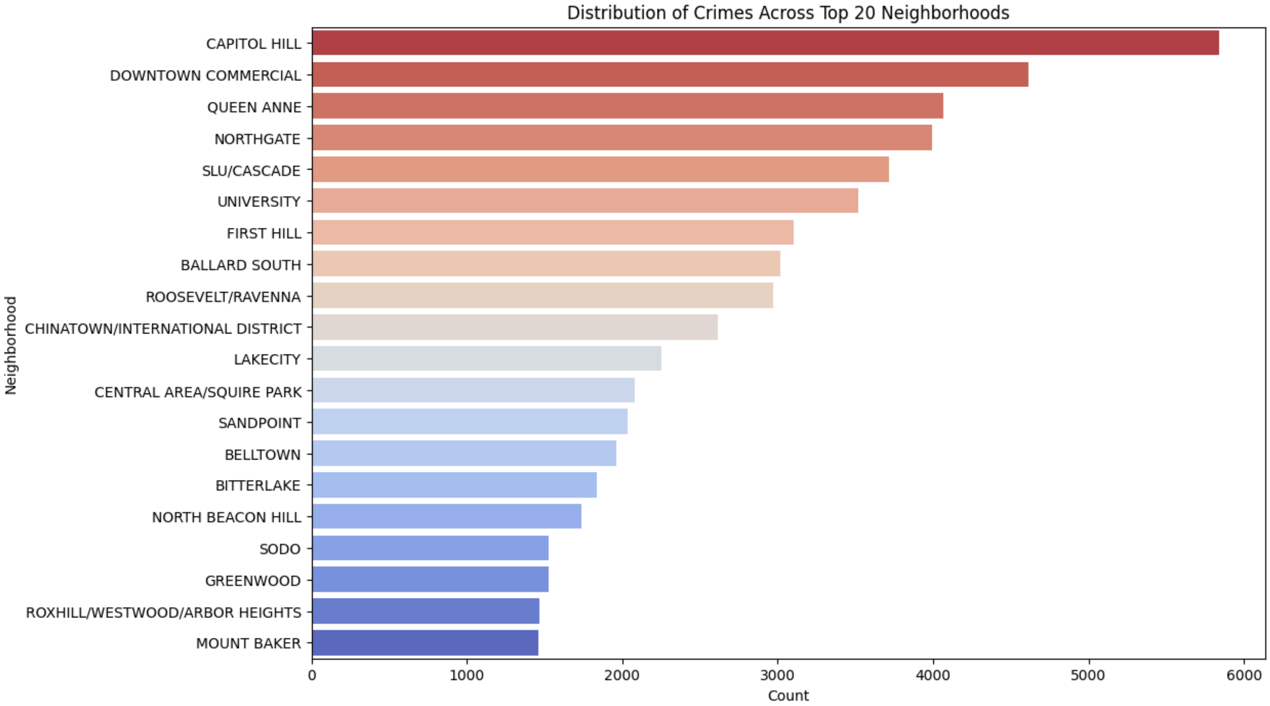


Distribution of Offenses for Larceny Theft: We wanted to take a closer look at the most prominent types of theft that occurred in 2024, and this visualization indicates Theft from Motor Vehicle as most common, followed by Motor Vehicle Theft, Theft of Vehicle Parts, and Shoplifting. Less frequent types include Theft from Building, and rarer crimes like Identity Theft, Pocket-picking, and Purse-snatching. There is a category labeled “All Other Larceny" which serves as a placeholder for all other thefts that do not fit into the above categories.

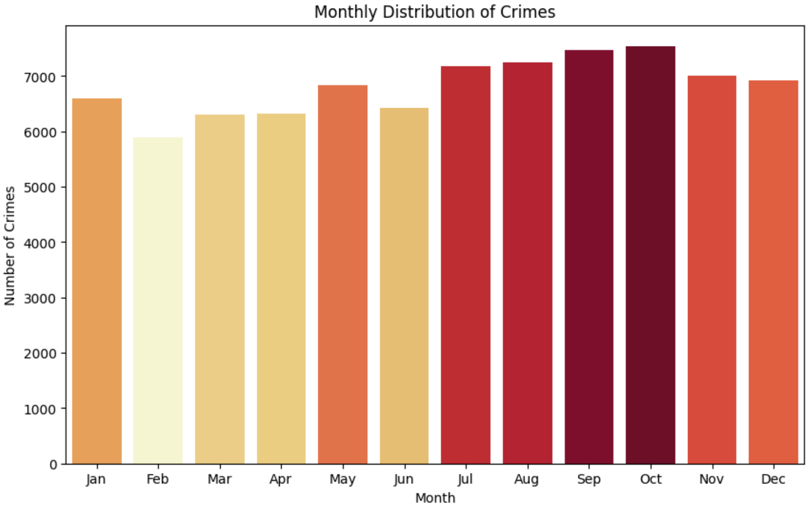


Distribution of Crimes by Neighborhood: To help in answering the question, “Where should we allocate resources”, we examined the crime distribution by the top 20 neighborhoods with crime rates. This graph shows Capitol Hill with the highest incident count, nearing 6,000 cases, followed by Downtown Commercial, Queen Anne, and Northgate. Other high-activity areas include South Lake Union/Cascade, University District, and First Hill. Neighborhoods like Mount Baker, Roxhill/Westwood/Arbor Heights, and Greenwood report the lowest counts among the top 20, with a gradual decline in incidents moving down the list. The color gradient shifts from dark red for higher-crime areas to dark blue for lower-crime ones. This data can guide the strategic reallocation of law enforcement and community safety resources to neighborhoods with the highest crime levels, maximizing prevention and response efforts.

Frequency of Crimes by Month: Additionally, we wanted to find out when crimes are primarily occurring, to help police plan safety initiatives and increase operations during high-crime time-periods. This chart shows the monthly distribution of crimes, with the lowest counts in

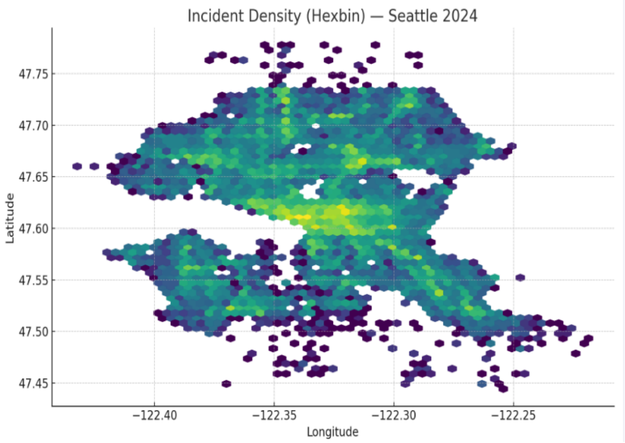


February and March, and the highest in September and October, both exceeding 7,500 incidents. Contrary to popular belief, most crimes don't occur during the summer months, this data shows a different trend, with crime peaking in early fall.



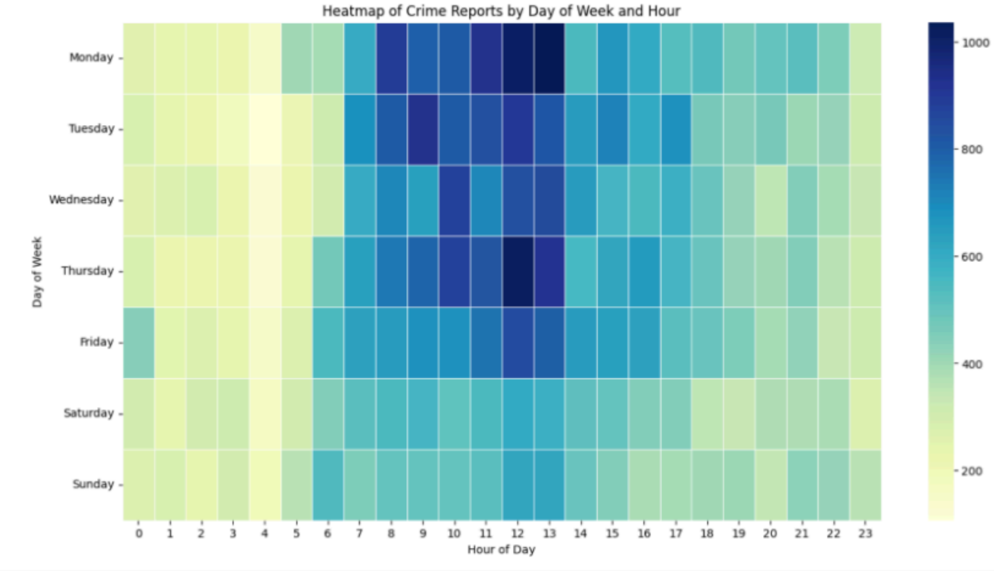
Crime Density

This visual shows the distribution of crime incidents across Seattle, with color representing the density of reports. The highest concentration is shown in yellow, and is clustered in Capitol Hill, Downtown, and University District. Surrounding neighborhoods show moderate density, and further areas show fewer incidents. This visual highlights hotspots where law enforcement and community safety initiatives should be prioritized to reduce the greatest concentration of this crime.



Predictive Modelling:

To better understand when crimes are likely going to occur, we used a heatmap that shows the distribution of these reported incidents by hour and day of week. This shows that the crime reports peak during late mornings and early afternoons on weekdays with Monday and Thursday around noon showing the highest crime. Reports also show that crime is low between midnight and 5 a.m. One thing that was surprising to us was that we see fewer incidents on weekends which suggests that crime activity is tied to weekday patterns such as school and work hours.

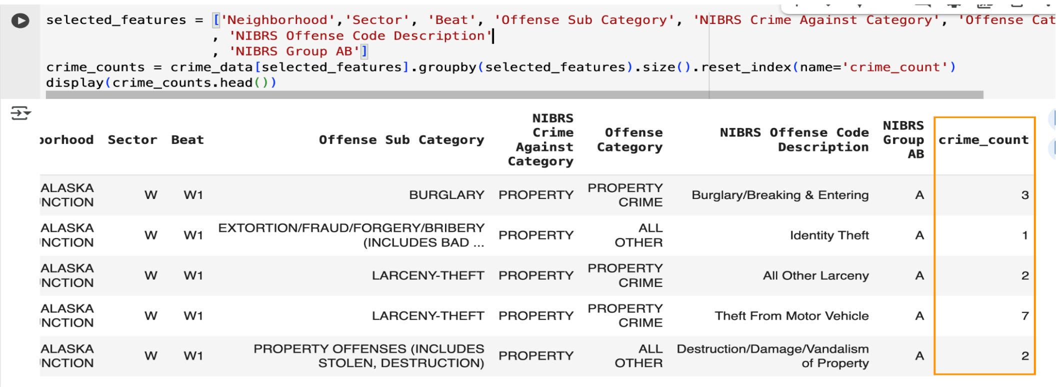


Regression Models

Linear Regression Model - Crime Count Prediction based on areas and categories as features:

When we tried to build linear regression models on our dataset, we found it slightly tricky. As we had learned for building a linear regression model we need to have a continuous numeric value, which was missing in our dataset. As part of data preparation we need to derive the quantitative value. Our main objective is to predict the number of crimes in an area using historical crime data and relevant features. By analyzing the data we discovered the features related to area and location are related to predicting the number of crimes. Following are the selected features: 'Neighborhood', 'Sector', 'Beat', 'Offense Sub Category', 'NIBRS Crime Against Category', 'Offense Category', 'NIBRS Offense Code Description' , 'NIBRS Group AB'

To prepare the quantifying data we had to aggregate crime data by relevant features (area & time).



```
selected_features = ['Neighborhood','Sector', 'Beat', 'Offense Sub Category', 'NIBRS Crime Against Category', 'Offense Cat
                    , 'NIBRS Offense Code Description'
                    , 'NIBRS Group AB']
crime_counts = crime_data[selected_features].groupby(selected_features).size().reset_index(name='crime_count')
display(crime_counts.head())
```

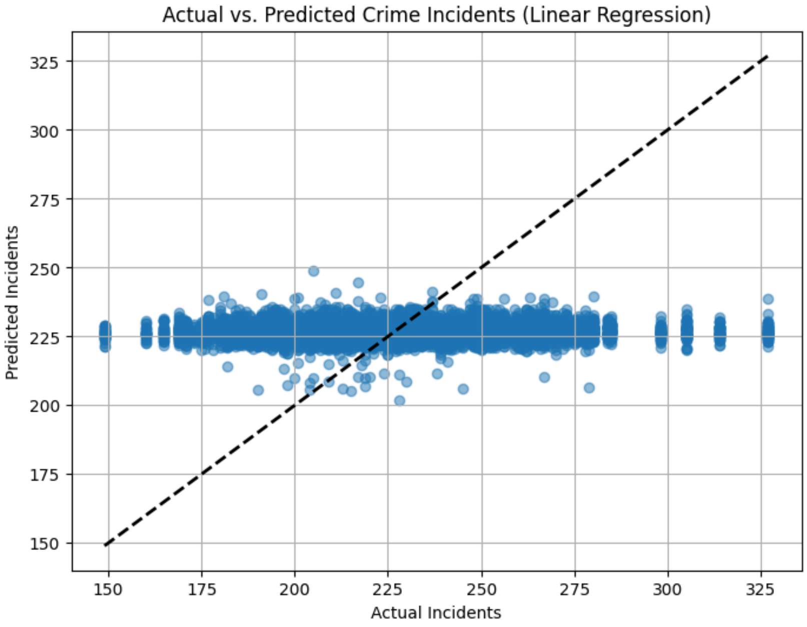
orhood	Sector	Beat	Offense Sub Category	NIBRS Crime Against Category	Offense Category	NIBRS Offense Code Description	NIBRS Group AB	crime_count
ALASKA NCTION	W	W1	BURGLARY	PROPERTY	PROPERTY CRIME	Burglary/Breaking & Entering	A	3
ALASKA NCTION	W	W1	EXTORTION/FRAUD/FORGERY/BRIBERY (INCLUDES BAD ...	PROPERTY	ALL OTHER	Identity Theft	A	1
ALASKA NCTION	W	W1	LARCENY-THEFT	PROPERTY	PROPERTY CRIME	All Other Larceny	A	2
ALASKA NCTION	W	W1	LARCENY-THEFT	PROPERTY	PROPERTY CRIME	Theft From Motor Vehicle	A	7
ALASKA NCTION	W	W1	PROPERTY OFFENSES (INCLUDES STOLEN, DESTRUCTION)	PROPERTY	ALL OTHER	Destruction/Damage/Vandalism of Property	A	2

Then we converted categorical variables to numerical format (0,1) via one-hot encoding.

After that the data was split into training and testing sets. (80% train and 20% test), divided data randomly at the value 42. The model was trained using Linear Regression model on training data and predicted value i.e. number of crimes for the test dataset. Further, we evaluated model performance by calculating Root Mean Squared Error (RMSE). RMSE is the difference between the actual and predicted values.

RMSE: 27.96

MAE: 22.20



The scatter plot of actual versus predicted incidents visually shows the poor performance, with predicted values clustered around a relatively narrow range regardless of the actual incident count.

The current Linear Regression model using only the specified categorical features is not effective in predicting crime counts. Future steps should involve exploring additional features or feature engineering techniques to improve predictive performance. Then, we further explored forecasting by narrowing down our features and using feature engineering techniques.

Regression Models for Citywide Daily Crime Incident Forecasting

Objective: The aim of this analysis was to forecast daily crime incident counts for the city of Seattle using historical data from 2024. Two regression approaches were explored: Linear Regression and Random Forest Regression. These both were applied and compared to identify which method offers higher predictive accuracy and better captures the patterns in crime occurrences.

Data Preparation

Aggregation: Aggregate raw crime incident data to daily incident counts for the entire city.

Date Range Completion: A complete date range for 2024 was generated to ensure all days were represented, with missing days filled with a count of zero incidents.

Feature Engineering:

- Trend: Day index from the start of the dataset to capture long-term changes.
- Day of Week: One-hot encoded variables for each weekday to model weekly variations.
- Annual Seasonality (sin_y, cos_y): Sine and cosine transformations of the day-of-year to capture recurring annual cycles, such as seasonal crime fluctuations.

Data Split: The prepared dataset was split into 80% train and 20% test dataset.

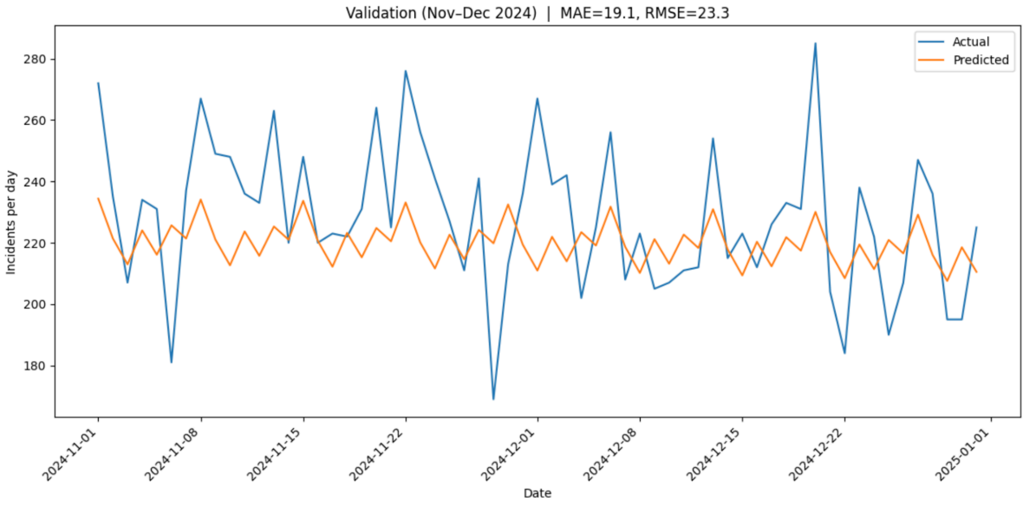
The data was split into: Random State: 42, Training Set: January 1 - October 31, 2024, Validation Set: November 1 - December 31, 2024. Both models were trained and evaluated using the same training/validation split for a fair comparison.

Model 1: Linear Regression

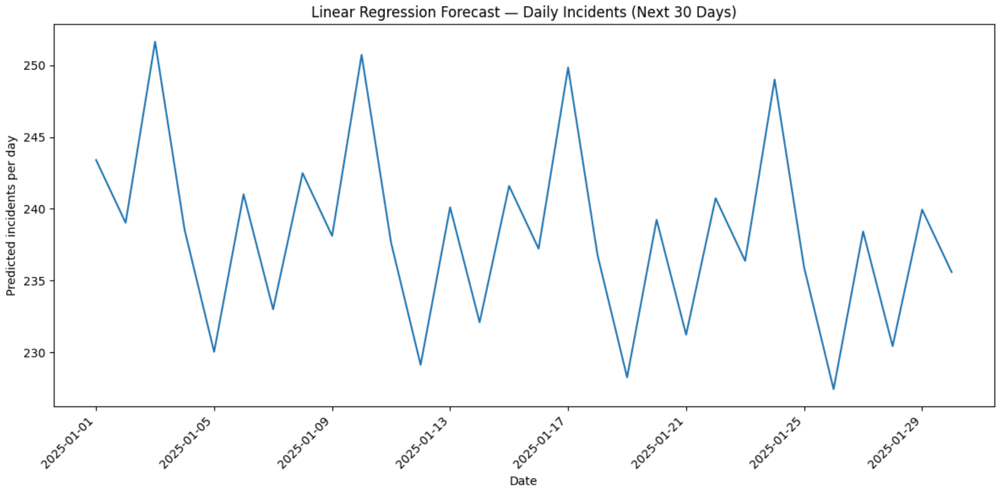
A standard Linear Regression model was trained on the engineered features from the training set. This model assumed a linear relationship between predictors and the daily crime counts.

Results: Mean Absolute Error (MAE): 19.1

Root Mean Squared Error (RMSE): 23.3



The model captured overall trends, weekly patterns, and annual seasonality but could not predict sudden spikes or drops in daily incidents. The actual vs. predicted plot for the validation period showed good alignment with the general pattern, though with some deviations during high-crime days.



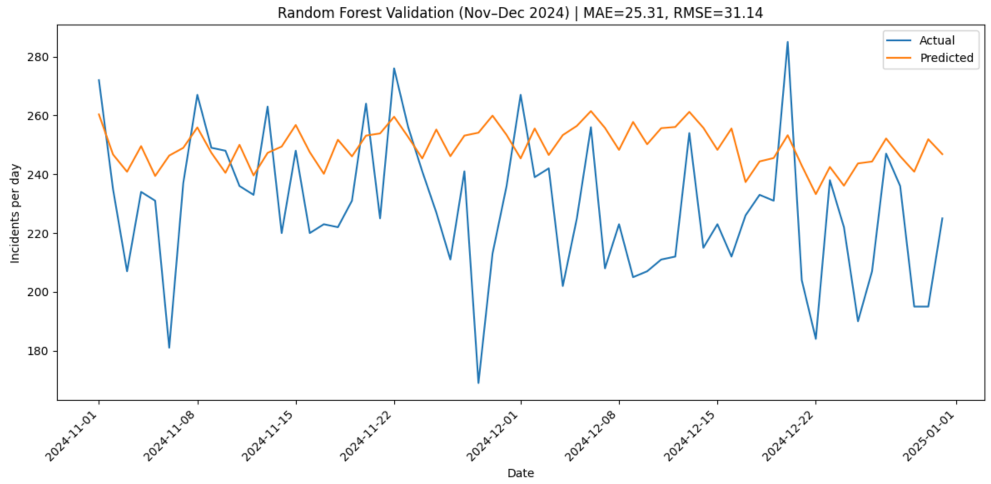
Forecast: A 30-day forecast for January 1 - January 30, 2025, was generated using the trained model. The forecast plot reflected the learned trend, weekly effects, and seasonal variations.

Model 2: Random Forest Regression

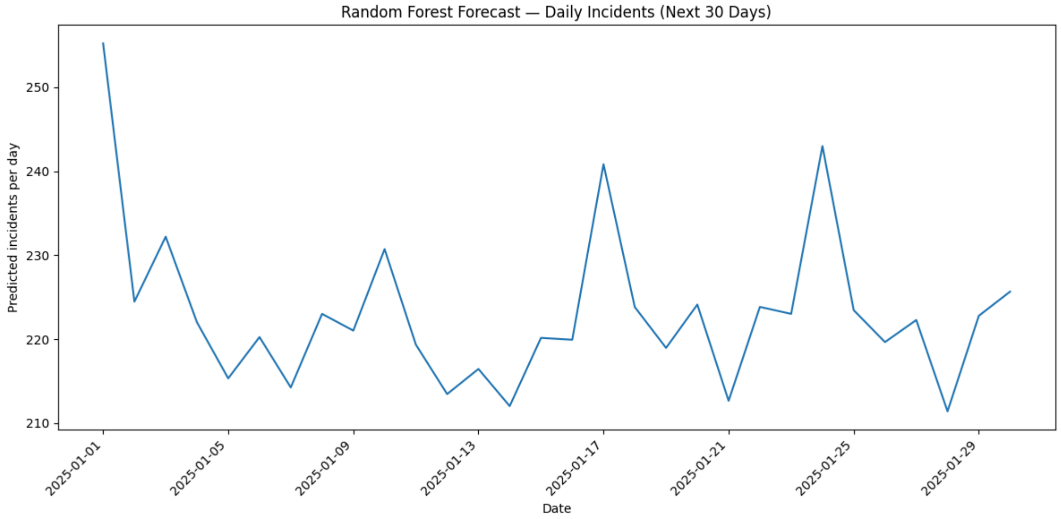
A Random Forest Regressor model was trained using the same features and training data. The model was configured with 100 estimators and a fixed random. Unlike Linear Regression, Random Forest is known to capture complex non-linear relationships.

Results: Mean Absolute Error (MAE): 25.31

Root Mean Squared Error (RMSE): 31.41



The model more accurately showed fluctuations in daily crime incidents during the validation period. The validation plot showed close alignment between predicted and actual values, even for irregular patterns.



Forecast: A 30-day forecast for January 1 - January 30, 2025, was generated using the trained Random Forest model. The forecast captured both regular seasonal patterns and potential short-term variations.

Metric	**Linear Regression**	**Random Forest Regression**
MAE	19.1	25.31
RMSE	23.3	31.14

We can see the linear regression model has a lower MAE and RMSE, meaning it makes more accurate predictions on the validation data. It proved really good at spotting the main pattern over time, like if the number is generally increasing or decreasing.

So, while the Random Forest is a powerful model in general, our analysis shows that for this dataset and forecast period, the simpler Linear Regression model was more accurate based on the chosen metrics. Random Forest handles complex and irregular patterns better,it looks for lots of little specific rules and combinations of things to make its guess. It can be better at finding tricky, irregular patterns that the Linear Regression might miss.

Here the Linear Regression was stronger at showing overall trends and seasonal patterns.

Further feature engineering or hyperparameter tuning could potentially improve the performance of the Random Forest model, or exploring other time series forecasting models might be beneficial.

The forecasts will be very helpful for the Seattle police department to be more proactive and efficient in their efforts to maintain public safety.

- Knowing when and where crime is likely to occur helps the police department strategically allocate resources like officers and equipment.
- Over time, consistent forecasting can provide insights into long-term crime trends, which can inform broader policy decisions related to public safety, urban planning, and social programs.

Who Benefits?

Everyone and everything benefits from a safer city. Residents experience an enhanced quality of life and have a greater peace of mind. Safe neighborhoods foster a greater sense of community, and social interaction and participation in local activities is increased. Mental and physical health benefits exponentially due to reduced fear of crime and a higher sense of security. Businesses will thrive from this, leading to economic growth. The city will have a greater reputation, which leads to more tourism as well. All in all, a safer city will lead to only positive changes.

Recommendations

We can remove any challenges civilians may face with reporting crimes, so that we can lessen this time difference between report and offense time. Perhaps automation can be implemented to be able to ensure crime is reported as soon as it occurs, and that someone can be sent as soon as possible to mitigate this problem. We can also work on improving police and community relationships. When people trust law enforcement, they are more likely to rely on police, cooperate and participate in investigations, and collaborate on crime reduction strategies.

From our analysis, we see that theft is the highest crime in Seattle. Theft related to a motor vehicle is most common. To prevent that, Seattle should have more secure parking lots with security guards, secure gates and surveillance cameras. We need more police monitoring areas with more crime. From our data, we see the top 3 neighborhoods with highest crime are Capitol Hill, Downtown and Queen Anne. This will help minimize time that police can get to the scene as well. Civilians can also take measures to prevent theft, such as locking doors and windows, using a steering wheel lock, having an audible alarm, removing valuables from cars, and parking in well lit areas. As a city, they can also ensure civilians are well informed with alerts of crime in the area.

To inform management, we have data to prove which crimes to look out for. Quality and well-curated data is crucial to illustrate our findings. Public safety solutions can then be designed to meet specific community needs. We have several graphs that clearly inform exactly what crimes have occurred, which crimes are more prominent, and which neighborhoods need more enforcement. Management will be informed of when crimes usually occur, how to prepare and mitigate crimes, and recommendations will be shared to them. This will also allow them to better allocate their budget and resources accordingly, whether it is to hire more police or installing safer infrastructure. This will drive more conversation as well on how to ensure the city is more informed of crime happening in the area so they can better prepare.

Challenges

Working with over 1.5 million rows of data from 2008-2024 caused Google Collab to crash due to RAM limitations which forced us to only use data from 2024. The dataset also contains both report date/time and offense date, however it is unclear whether the report time is when the crime was reported, when it was logged, or when a police officer was sent onto the site. The dataset also did not include demographic or environmental variables that could provide insight into crime patterns. Without this the root cause remains difficult to understand.

Future Improvements

Because our dataset was so large, we had to limit the dataset to 2024 so that Google Colab will be able to run the data. In the future, we would love to analyze more years of data so we can see trends in the data, and how crime changes throughout the year so the city can be best informed. We would also love to combine more data sets to see if there is more data in terms of exactly what time the crime is committed. We are unsure if the report time indicates that the police officer will be able to reach the scene immediately or if there is a delay for that to happen. It will also be great if we can see data on the number of police officers on site throughout the year, and locations of police officers in the city.

How AI tools assisted in the project

We utilized Gemini to support the creation of bar charts and to offer additional perspective during data interpretation. Additionally, we asked Gemini to provide ideas on what visuals would best represent the data based on our objectives, helping us generate more quantifiable properties for clearer analysis. After collecting the data, we used ChatGPT as a fact-checker to verify our findings were in line with real-life information. All final decisions, interpretations, and conclusions were made by our team to ensure the work reflected our own analysis and judgement.

Data Source:

https://data.seattle.gov/Public-Safety/SPD-Crime-Data-2008-Present/tazs-3rd5/about_data