CRIME FORECASTING AND RESOURCE ALLOCATION: A Predictive Model for Effective Policing in the City of Los Angeles

DNSC6315: Machine Learning II

Final Project

May 10, 2023

Sophie Bukrinsky, Aditya Jaju, Jasvinder Singh Kalha, Estella Mei, Pavneet Singh

CRIME FORECASTING AND RESOURCE ALLOCATION: A Predictive Model for Effective Policing in Los Angeles County

Introduction & Problem Statement

High crime rates in the United States have far-reaching implications, affecting individuals' well-being, community trust, and the allocation of public resources (DOJ, n.d.; Wike et al., 2008). The social and economic costs associated with policing, prosecution, and incarceration further strain taxpayer funds, diverting them from important sectors like education and healthcare (Collins, 2020). In Los Angeles in 2022, the crime rate jumped 11.6%, and violent crimes are the highest in a decade (Kahn, 2023). Addressing crime in Los Angeles is paramount to law enforcement agencies and policymakers. This paper presents a data analysis project focused on predicting crime counts and identifying prevalent crime types by zip code and police station. By utilizing diverse predictive models, we aim to uncover valuable insights regarding the distribution and nature of crimes in different zip codes and their respective police stations. Based on these insights, we will propose recommendations for optimizing police allocation and budgeting in the city of Los Angeles[1].

Objective & Assumptions

We will base our recommendations on the proposed city budget and police allocation for 2023-2024. The budget outlines a $3.18B allocation to the Los Angeles Police Department (LAPD), comprising almost 25% of the total budget. Using the Los Angeles public budget data, we have determined that LAPD recruiting costs, including advertising, equipment, uniforms, salary, and bonuses, among other costs, comprise $693.2M. We have also determined that the total cost for specialized equipment, task forces, vehicles, and uniforms for specific types of crime comprises $318M. We want to use this budgetary information to allocate the police budget efficiently using our model.

[1] Due to time and scope constraints, the analysis and recommendations are only done on the City of Los Angeles, and does not include New York City as a comparison or within the present data analysis. However, the model and analysis may be adapted for other major cities.

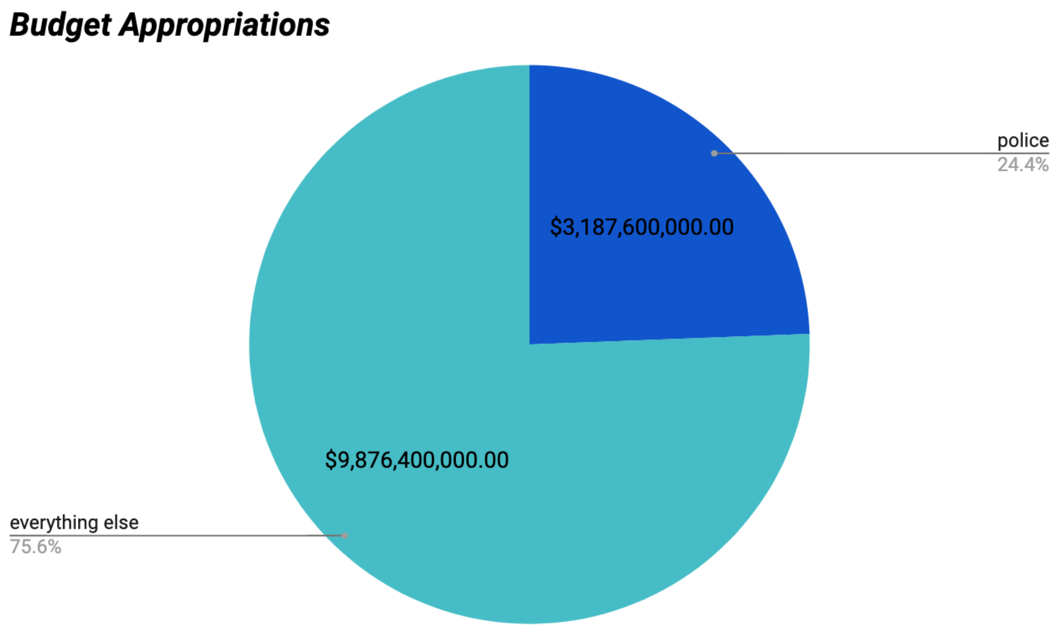


Figure 1. Source: City of Los Angeles

Police Allocation is across 21 police stations, all dedicated to a particular zip code-cluster, with 7070 total officers deployed (City of Los Angeles, 2023).

The data analysis and modeling will allow the team to predict the number of crimes per zip code, and thus per police station, and predict the number of crimes for next year for police localities. Through analysis and modeling, we can reallocate the 7070 police officers across the 21 stations based on the number of crimes our model reveals for the following year to understand recruitment needs. Due to our objective involving complex city budget appropriations and allocation of officers, it was necessary to make the following assumptions:

1. The modeling and analysis are contained to the City of Los Angeles.
2. Modeling and analysis is based on data from 2020-2022.
3. Population gender data is based on US Census information for Los Angeles, estimating 51% of the population is female. This is assumed to be true for all years 2020-2023.

4. We assume that 7070 officers can be distributed as needed across all 21 police stations as a base case. We will build on the current officers allocated and use our model to offer recruitment needs.
5. We assume that police recruitment will increase as crimes are projected to increase. If our model predicts a decrease in police recruitment for a certain station for 2024, it will stay constant with the current 2023 allocation.
6. We assume that the budget for recruitment, including training, salaries, bonuses, uniforms, and any other requirements, is about 22% of the police budget.
7. We assume that 10% of the budget goes to specialized task forces, equipment, and technology relevant to specific crimes.
8. We assume each zip code is consistent with the assigned police station.
9. We assume the given police budget can be distributed to the 21 police stations according to what each police station may require, based on officers recruited and prevalence of crime type, based on the weight of importance deemed necessary.
10. We assume profitability is irrelevant for this project, as it is based on government spending and allocated budget and will only affect budget planning for the following year.

Data Collection & Preparation

We collected crime data in Los Angeles and cleaned and preprocessed it for modeling using 8 datasets:

- City of Los Angeles Crime Data from 2020 to Present
- Los Angeles Parks and Recreation Information
- Neighborhood Empowerment/Youth Development Locations
- City of Los Angeles Active Business
- City of Los Angeles Populations by Zip Code 2020-2023
- Police Bureaus, Stations, and Communities Serviced
- City of Los Angeles Schools
- City of Los Angeles Median Household Income, by Zip Code

The data was sourced from open public sources like Los Angeles Open Data, Los Angeles Almanac, and Los Angeles city websites through web scraping. We analyzed police stations and their served communities by scraping LAPD websites and mapping them to zip codes. Crime locations were also mapped to zip codes using longitude and latitude coordinates extracted with the geopandas algorithm. The crime dataset included converting crime occurrence month to a numerical format (1-12). Zip codes were extracted from the datasets using the existing data or from the "address" column using RegEx in Python. Population data was built using US census information and Los Angeles Almanac data, incorporating base populations for 2020 and estimated populations for 2021-2023 based on city trends. The gender distribution assumption for the City of Los Angeles was 51% female and 49% male for 2020-2023, derived from US census data. Missing values were handled by replacing them with column averages. The school's data included an additional column indicating the school category based on the maximum grade taught. The datasets were then aggregated by zip code, and counts were obtained. Figure 2 illustrates an example of grouping park data by zip code and park type.

Zip	Park Type	Count
90002	Gardens	1
90002	Parks	1
90002	Public Computer	2
90002	Recreation Cent	2
90002	Senior Centers	1
90003	Outdoor Fitness	3
90003	Parks	4
90003	Public Computer	1
90003	Recreation Cent	2
90003	Senior Centers	1

Figure 2. Grouping Parks Data by Type and Zip.

Irrelevant columns were dropped from the final concatenated crime dataset. Figure 3 showcases a sample of the final concatenated dataset used for modeling, which consisted of 234,012 rows and 76 features.

	Month	Year	AREA	AREA NAME	Vict Sex	Crm Cd	Crm Cd Desc	ZIP_CODE	POPULATION	POP_SQMI	...	2022	2022 Males	2022 Females	2021	2021 Males	2021 Females	2020
0	1	2020	1	Central	F	110	CRIMINAL HOMICIDE	90013	14779	21418.84	...	13,052	6,395	6,657	13,022	6,381	6,641	13,009
1	1	2020	1	Central	F	121	RAPE, FORCIBLE	90012	39271	10671.47	...	37,391	18,322	19,069	37,305	18,279	19,026	37,268
2	1	2020	1	Central	F	121	RAPE, FORCIBLE	90013	14779	21418.84	...	13,052	6,395	6,657	13,022	6,381	6,641	13,009
3	1	2020	1	Central	F	121	RAPE, FORCIBLE	90014	10365	37017.86	...	9,156	4,486	4,670	9,135	4,476	4,659	9,126
4	1	2020	1	Central	F	121	RAPE, FORCIBLE	90071	17	130.77	...	153	75	78	152	74	78	152

Figure 3. Final Concatenated Dataset Used for Modeling.

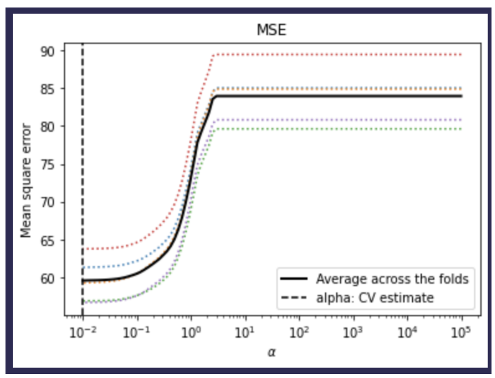
Approach & Methodology,

In our approach to addressing the issue of police allocation and budget resources in the City of Los Angeles, we utilized regression models, classification models, and clustering techniques to predict crime counts per zip code, identify crime types, and understand the underlying factors influencing crime rates. For regression, we employed Lasso regression to identify relevant variables for predicting crime counts and a linear regression model to establish a comprehensible relationship between variables and crime counts per zip code. We also used a random forest regressor because of its predictive accuracy and ability to handle large datasets. To classify crime types, we utilized a random forest classifier known for its enhanced predictive accuracy and ability to mitigate overfitting. Additionally, we employed clustering analysis based on zip codes to gain insights into spatial variations in crime patterns. Our focus revolved around predicting crime counts and classifying crime types for each police station while also conducting supplementary analyses on feature importance, prevalent crime types, and future crime count predictions.

Regression Model (LASSO):

The Lasso regression algorithm to predict crime counts based on various demographic and economic factors. We decided to use the LASSO model as it is a type of linear regression that introduces an additional regularization term to the loss function, which helps to avoid overfitting the model. First, the dataset was split randomly into training and testing using the train_test_split() function from the sklearn library. The Lasso regression model was then trained using the training data, and its coefficients were plotted against the regularization parameter alpha, and the graph plotted for the same for better understanding. Next, utilizing the lassoCV function, we found the best α for your prediction and used this alpha to train the model again and get the coefficients of each of the features printed to understand which features are relevant in predicting the number of crimes and which features are irrelevant. The results of the Lasso regression model show that the regularization parameter alpha was found to be 0.01 for the best MSE, and the coefficients for the model were printed. The coefficients indicated that the feature with the highest positive effect on the crime count was "Population," while the feature with the highest negative effect on the crime count was "Elementary School." Once we got these coefficients, we found the top 10 features that are important in predicting the number of crimes. We also printed the bottom 10 features that are unimportant for better understanding. Here is a snippet of the top_10 features in the data frame and the MSE graph showing the various α values for the lasso cv as follows:

Top 10 important features:	
ZIP_CODE	0.600608
Swimming Pools - Summer	0.699606
Free Wi-Fi Hot Spots	0.811056
HIGH SCHOOL	0.838764
Senior Centers	0.879927
POP_SQMI	0.997814
2023	1.415200
CATEGORY_Violent Crimes	1.480008
Estimated Income Median	1.992603
POPULATION	3.085682



Figures 4 & 5. Top 10 features in the Lasso model and the MSE graph

The MSE calculated for the model was 58.509. This value indicates the average squared difference between the predicted and actual values of the test set. A smaller MSE value indicates a better performance of the model. Additionally, it is interesting to note that the population density of a zip code positively affects the crime count, which means that a higher population density in a zip code is associated with higher crime rates. On the other hand, the Elementary schools in a zip code had the highest negative effect on the crime count, which means that higher Elementary schools in a block group are associated with lower crime rates.

Regression Model (RandomForestRegressor):

The RandomForestRegressor model is used to predict the number of crimes in different areas based on the given features. In our analysis, we performed data preprocessing and cleaning by identifying categorical and numerical columns, converting numerical columns to float when necessary, and handling missing values by dropping rows. For data transformation, we utilized a column transformer to handle categorical variables with OneHotEncoder and performed feature scaling on numerical columns using StandardScaler. The preprocessor was then fitted and applied to the dataset. In the model training and evaluation phase, we split the dataset into training and testing sets, trained a RandomForestRegressor model with 100 estimators, and evaluated its performance using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R2). Figure 6 shows the performance metrics.

Mean Absolute Error:	0.02616279069767441
Mean Squared Error:	0.025805813953488358
R-squared:	0.9793531155344007

Figure 6. Performance Metrics of Random Forest Regressor

Classification Model (RandomForestClassifier):

Our methodology encompassed several essential steps. We began by preprocessing and cleaning the data, ensuring its integrity and consistency. This involved handling categorical and numerical columns and addressing missing values. Subsequently, we transformed the data using a column transformer, which effectively managed categorical variables and conducted feature scaling for numerical columns. This prepared the dataset for modeling. Next, we trained a RandomForestClassifier model with 100 estimators, enabling it to learn patterns and associations between the input features and crime types. To assess the model's performance, we employed accuracy as a metric to measure overall prediction correctness. Additionally, a comprehensive classification report was generated, providing valuable insights into precision, recall, F1-score, and support for each crime type class. This detailed evaluation offered a deeper understanding of the model's predictive capabilities. By utilizing the RandomForestClassifier model, we aimed to predict crime types based on diverse input features, facilitating a better understanding of criminal patterns and formulating effective crime prevention strategies. Figure 7 shows the performance metrics.

Accuracy	0.9186046511627907

Figure 7. Performance Metric of Random Forest Classifier.

We conducted a supplementary analysis to identify the most influential factors by calculating importance from the trained RandomForestClassifier model. We then predicted prevalent crime types by grouping the dataset by the community and using the model for predictions. We created a function to generate a dataset with specified time periods and ZIP codes for future crime counts and used the trained RandomForestRegressor model for predictions. This allowed us to anticipate crime trends and provide valuable insights. Our approach aimed to inform decision-making and targeted crime prevention efforts by uncovering important features, predicting crime types, and forecasting future crime counts. Our project successfully implemented regression and classification models to predict crime counts and classify crime types, respectively. The performance metrics indicate the effectiveness of our models. The supplementary analyses provided additional insights into feature importance, prevalent crime types, and future crime counts prediction. This information can benefit law enforcement and local authorities to allocate resources better and implement preventive measures.

Clustering Model Methodology

The analysis involved four main steps: data preprocessing, feature selection, clustering, and interpretation. For preprocessing, the goal was to ensure that the data was in a format that could be used for clustering analysis. It involved reverting the data to longitude and latitude coordinates, transforming it to a numeric format, and removing missing values. The second step was to select the most relevant features for the clustering analysis. We selected two key features: crime count and estimated median income. These features will provide insights into the relationship between crime and income in LA. The third step involved clustering the data using the K-means clustering and K-median clustering algorithms. We used the elbow method to find the optimal K to determine the optimal number of clusters. We also visualized the results using scatterplots to gain insights into the data distribution. The final step was to interpret the results of the clustering analysis. The optimal number of clusters was found at three. The K-median clustering is more robust to the outliers, and the graph can be interpreted easily. Cluster 1 has a high crime count with a low median Income. Cluster 2 has a moderate crime count and a moderate estimated median income. A low crime count and a high estimated median income characterize cluster 3. Our analysis also revealed a strong negative correlation between crime count and estimated median income. This suggests that neighborhoods with lower incomes tend to have higher crime rates. Figure 8 provides a heatmap of crime in the City of Los Angeles.

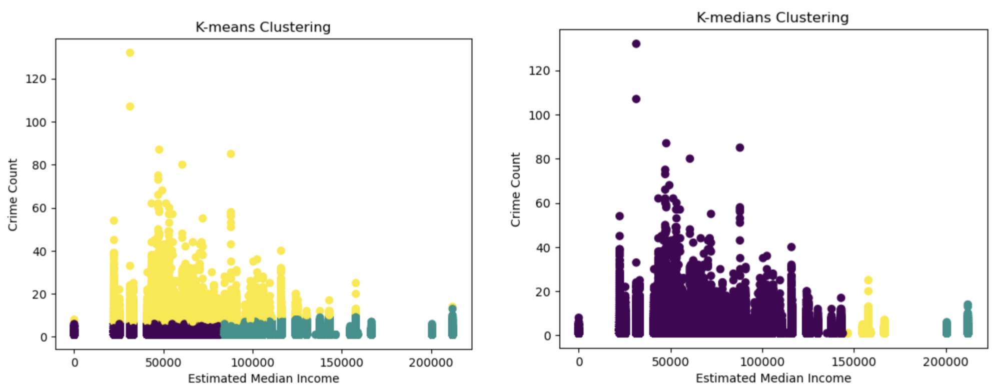


Figure 8. Clusters of crime rates in the City of Los Angeles

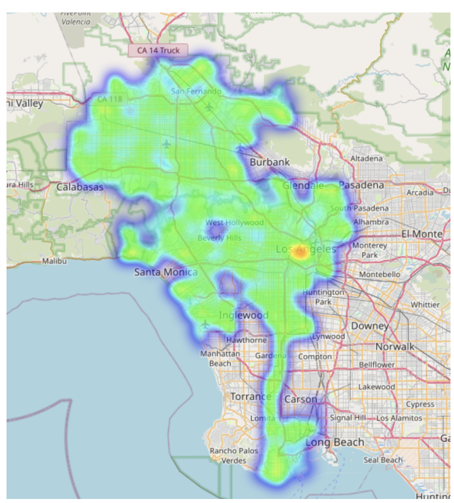


Figure 9. Heatmap of Crime in the City of Los Angeles

Insights and Findings

Several exciting insights and findings can be drawn from each model that ultimately affects the recommendations for police allocation and budget appropriations.

Random Forest Regression Models and Classification Model

Our analysis utilizing the Random Forest Regressor and Random Forest Classifier models have yielded significant insights and findings regarding crime patterns. We have discovered the most prevalent crime types in various cities within our dataset, such as "BURGLARY FROM VEHICLE" in Los Angeles (California State Univ Northridge) and "THEFT PLAIN - PETTY ($950 & UNDER)" in Los Angeles (University of Southern California). Additionally, we observed a correlation between population density and crime rates, with densely populated areas like Los Angeles (University of California Los Angeles) experiencing higher instances of crime. Our analysis also indicated a potential relationship between certain recreational facilities (e.g., parks, public computer centers, and rental facilities) and crime rates, warranting further investigation. Moreover, our models highlighted a potential link between median income and crime rates, with lower-income areas like zip code 90013 exhibiting higher crime rates. Lastly, our predictions for 2024 indicate that crime rates may continue to fluctuate in the future, emphasizing the importance of ongoing monitoring and the development of targeted crime prevention strategies.

Clustering Model

The heat map clearly showed a concentration of crime in certain areas, particularly in the **downtown regions**. And the clusters have proved that the crime rates are higher in areas with a median income between $25,000 to $125,000. These findings suggest that these areas may be more vulnerable to crime and could benefit from targeted policing efforts or other interventions. Overall, the clustering visualizes the socioeconomic patterns of crime and identifies areas of particular concern.

Model Application and Managerial Decision

We will apply the model by first understanding how to allocate police accordingly. We will start with 21 police stations and have 7070 police officers to deploy based on the analysis.

We used our trained regression models and dummy variables to build out crime count prediction per zip code for 2024 and mapped the zip codes to the relevant police stations. We then got the total crime count for each police station to understand the total number of crimes predicted for 2024. We then divided the total number of crimes predicted by the 7070 police officers to understand the ratio of police officers to crimes. We then used this ratio with the predicted number of crimes to allocate police officers. Because the total crime count and crime rate will be **increasing in 2024,** we determined that the **difference in police allocation** is the most relevant metric for the upcoming budget. Additionally, if the model returned the output of **fewer police officers,** we have determined that it is necessary to keep the previously allocated police officers because of the increase in crime in total. Figure 8 shows a table describing the 2024 police allocation logic. We have identified that LAPD needs to recruit 1791 more police officers for 2024.

Police Station Name	Year	TOTAL CRIME COUNT	Officers Allocated Present	2024 OFFICERS REQUIRED	REQUIRED RECRUITMENT
77th Street Community Police Station	2024	23938	407	685	278
Central Community Police Station	2024	19200	410	550	140
Devonshire Community Police Station	2024	10543	265	302	37

Foothill Community Police Station	2024	2905	277	83	0
Harbor Community Police Station	2024	4479	280	128	0
Hollenbeck Community Police Station	2024	5011	311	143	0
Hollywood Community Police Station	2024	18401	356	527	171
Mission Community Police Station	2024	11053	302	316	14
N Hollywood Community Police Station	2024	8100	271	232	0
Newton Street Community Police Station	2024	9211	327	264	0
Northeast Community Police Station	2024	13820	299	396	97

Olympic Community Police Station	2024	17040	255	488	233
Pacific Community Police Station	2024	687	294	20	0
Rampart Community Police Station	2024	18509	282	530	248
Southeast Community Police Station	2024	15462	377	443	66
Southwest Community Police Station	2024	15330	383	439	56
Topanga Community Police Station	2024	3593	253	103	0
Van Nuys Community Police Station	2024	15520	251	444	193
West LA Community Police Station	2024	10620	229	304	75

West Valley Community Police Station	2024	10543	252	302	50
Wilshire Community Police Station	2024	13039	240	373	133

Figure 10. Required Recruitment for LAPD for 2024.

We also have a total police budget of $3.18B. We determined that 10% of the total police budget should be allocated to curb the most prevalent crime under the jurisdiction of each police station. We need $693M and the ratio of new recruits for the abovementioned recruitment. The rest of the budget will be allocated as per usual operations. The budget allocated to each police station to form its special task force is **proportional to the crime count of the most prevalent crime category.** Figure 11 provides the weights assigned to each type of crime. This is then multiplied by the budget and **fraction of officers required for recruitment.**

Recruitment	Ratio of recruited/total	Most Prevalent Crime Type	Counts	% Weight of Most Prevalent Crime	Budget Allocated for Special Crime Task Force	Budget for Recruitment	Total Budget
278	16%	Violent Crimes	550	9%	$27,340,941.07	$107,598,883.31	$134,939,824.37
140	8%	Property Crimes	402	6%	$19,983,742.38	$54,186,488.00	$74,170,230.37
37	2%	Violent Crimes	301	5%	$14,962,951.38	$14,320,714.68	$29,283,666.07
0	0%	Theft Crimes	63	1%	$3,131,780.52	$0.00	$3,131,780.52
0	0%	Violent Crimes	112	2%	$5,567,609.82	$0.00	$5,567,609.82
0	0%	Property Crimes	135	2%	$6,710,958.26	$0.00	$6,710,958.26
171	10%	Property Crimes	525	8%	$26,098,171.02	$66,184,924.62	$92,283,095.64
14	1%	Theft Crimes	258	4%	$12,825,386.90	$5,418,648.80	$18,244,035.70
0	0%	Property Crimes	223	3%	$11,085,508.83	$0.00	$11,085,508.83
0	0%	Property Crimes	204	3%	$10,141,003.60	$0.00	$10,141,003.60
97	5%	Property Crimes	325	5%	$16,156,010.63	$37,543,495.25	$53,699,505.88
233	13%	Property Crimes	451	7%	$22,419,571.67	$90,181,797.88	$112,601,369.55
0	0%	Theft Crimes	53	1%	$2,634,672.50	$0.00	$2,634,672.50
248	14%	Property Crimes	512	8%	$25,451,930.59	$95,987,493.02	$121,439,423.61
66	4%	Violent Crimes	384	6%	$19,088,947.94	$25,545,058.63	$44,634,006.57
56	3%	Property Crimes	372	6%	$18,492,418.32	$21,674,595.20	$40,167,013.52
0	0%	Violent Crimes	113	2%	$5,617,320.62	$0.00	$5,617,320.62
193	11%	Property Crimes	448	7%	$22,270,439.27	$74,699,944.17	$96,970,383.43
75	4%	Theft Crimes	332	5%	$16,503,986.24	$29,028,475.71	$45,532,461.96
50	3%	Violent Crimes	301	5%	$14,962,951.38	$19,352,317.14	$34,315,268.52
133	7%	Property Crimes	333	5%	$16,553,697.05	$51,477,163.60	$68,030,860.64
1791	**100%**		**6397**	**100%**	**$318,000,000.00**	**$693,200,000.00**	**$1,011,200,000.00**

Figure 11. Budget Allocation for new initiatives

References

https://cops.usdoj.gov/RIC/Publications/cops-w0669-pub.pdf

https://www.pewresearch.org/global/2008/04/15/where-trust-is-high-crime-and-corruption-are-low/

https://www.vox.com/the-highlight/21430892/defund-the-police-funding-abolish-george-floyd-breonna-taylor-daniel-prude

https://scholars.unh.edu/cgi/viewcontent.cgi?article=1676&context=honors

https://www.journals.uchicago.edu/doi/full/10.1086/715713#:~:text=The%20estimated%20annual%20cost%20of,%243.92%20trillion%20net%20of%20transfers.

https://norton.house.gov/media-center/press-releases/norton-introduces-bill-to-create-local-task-forces-on-policing-0

https://catalog.data.gov/dataset/nypd-arrests-data-historic

https://catalog.data.gov/dataset/crime-data-from-2020-to-present

https://data.lacity.org/

https://opendata.cityofnewyork.us

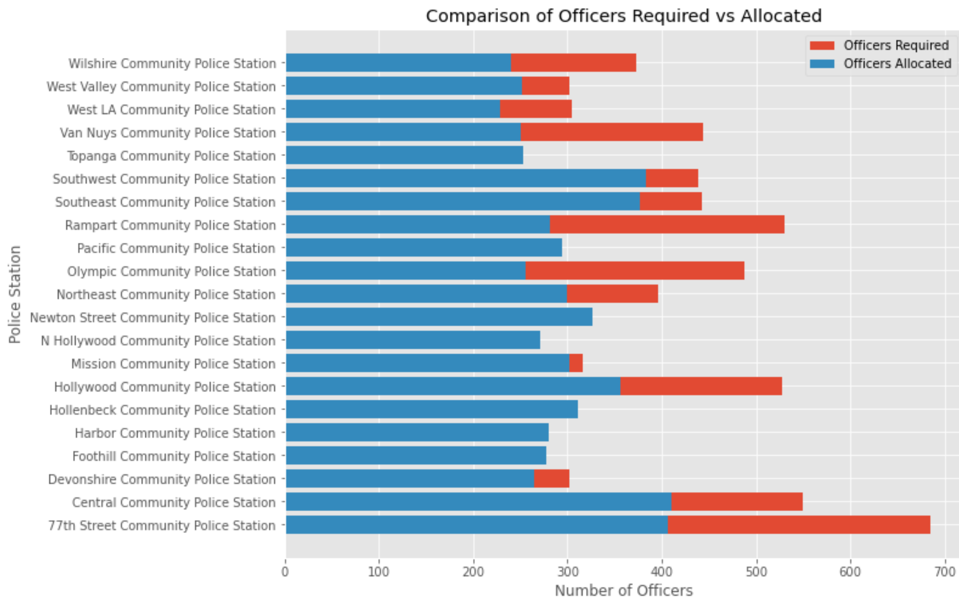
https://xtown.la/2023/01/23/crime-rate-los-angeles-2022/

https://cao.lacity.org/budget23-24/2023-24Proposed_Budget.pdf

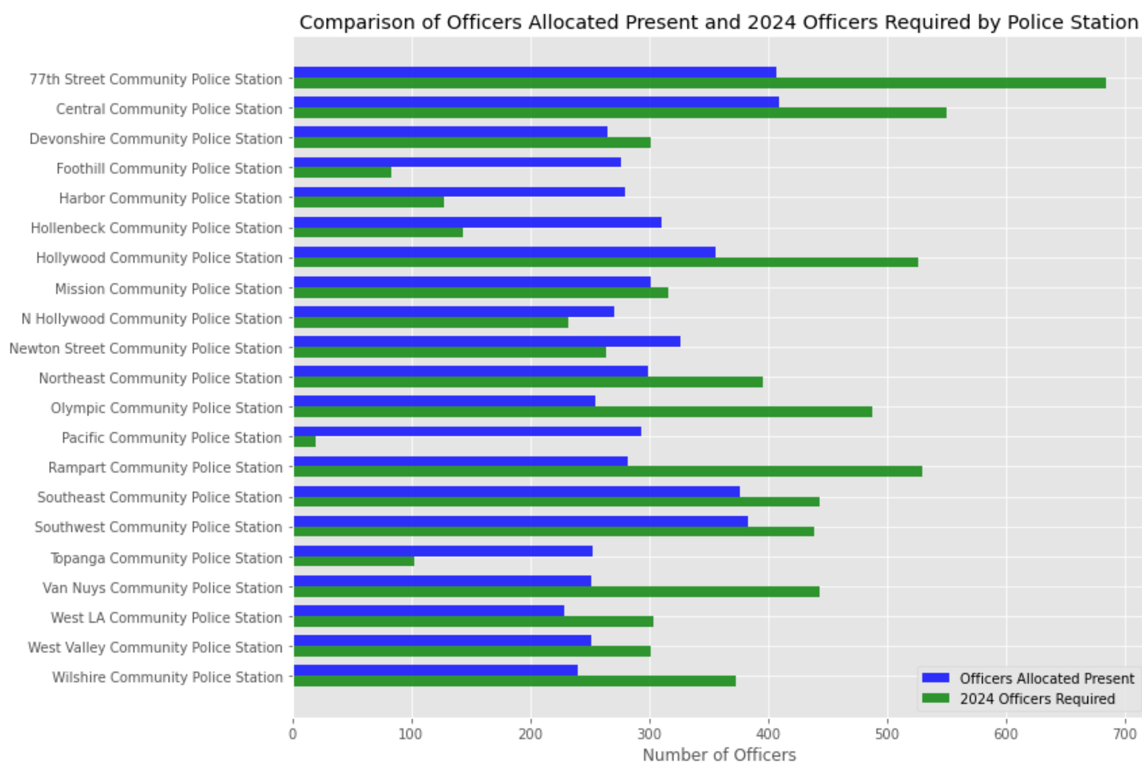
https://www.laalmanac.com/crime/cr70a.php

Appendix:

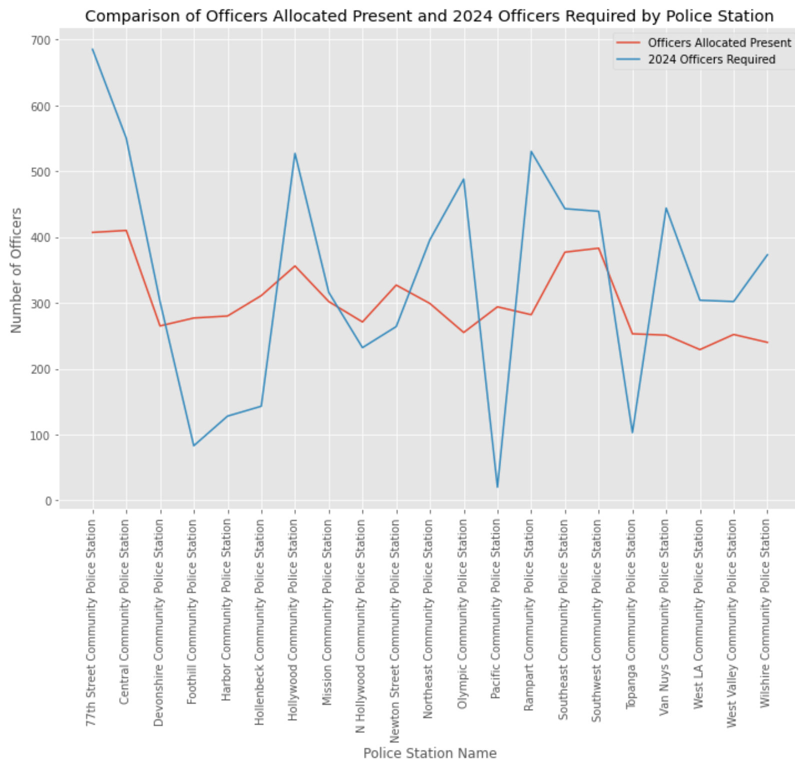
Comparison of Officers required vs Allocated



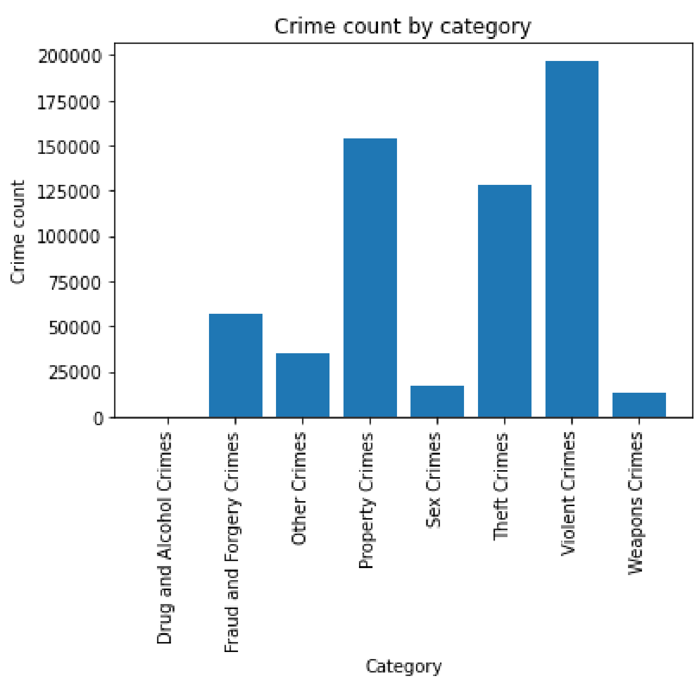
Comparison of Officers allocated at present vs 2024 required officers



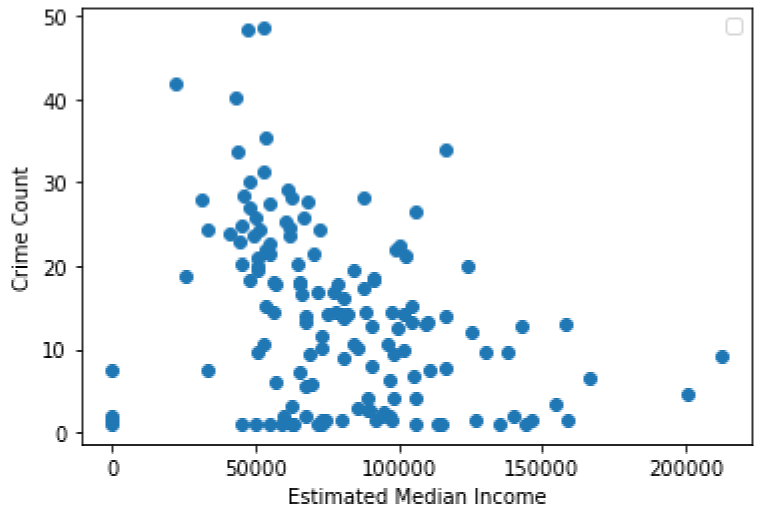
Line graph of the comparison between Police officer allotment and requirement



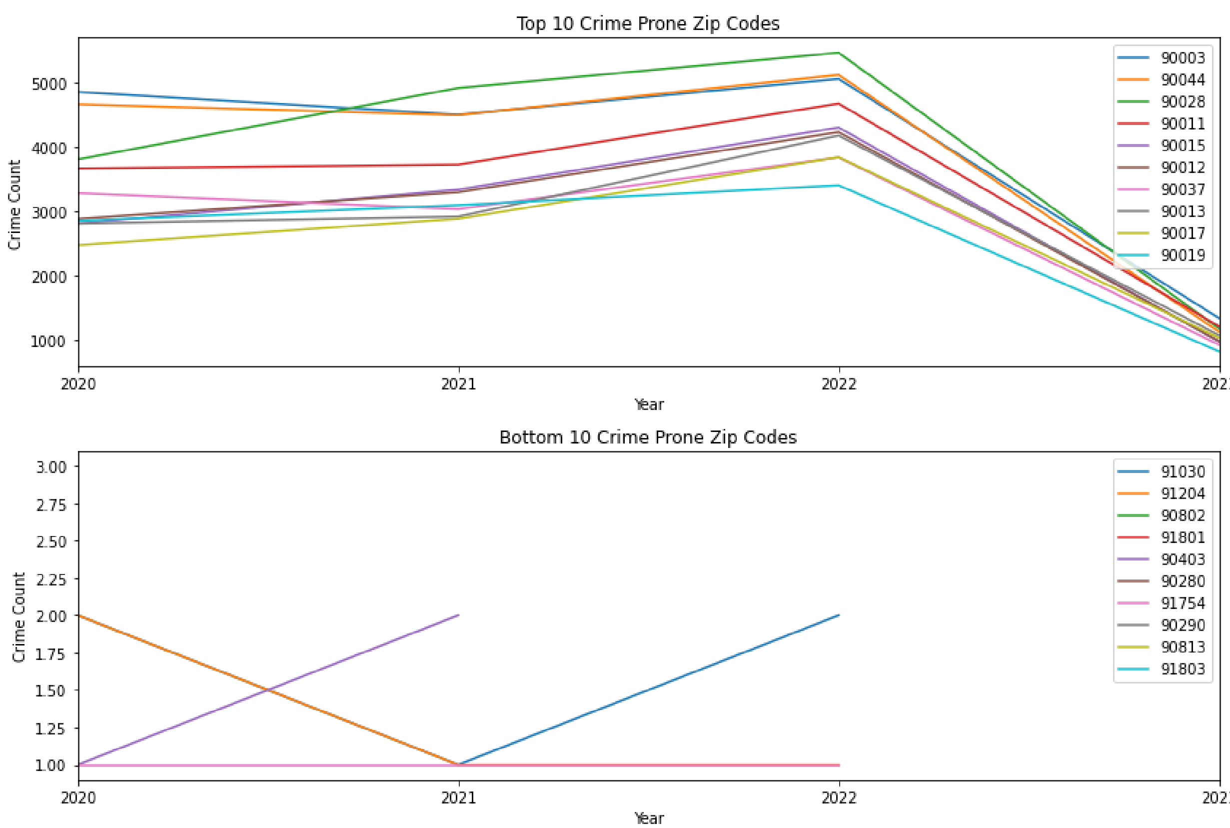
Bar chart of crime type



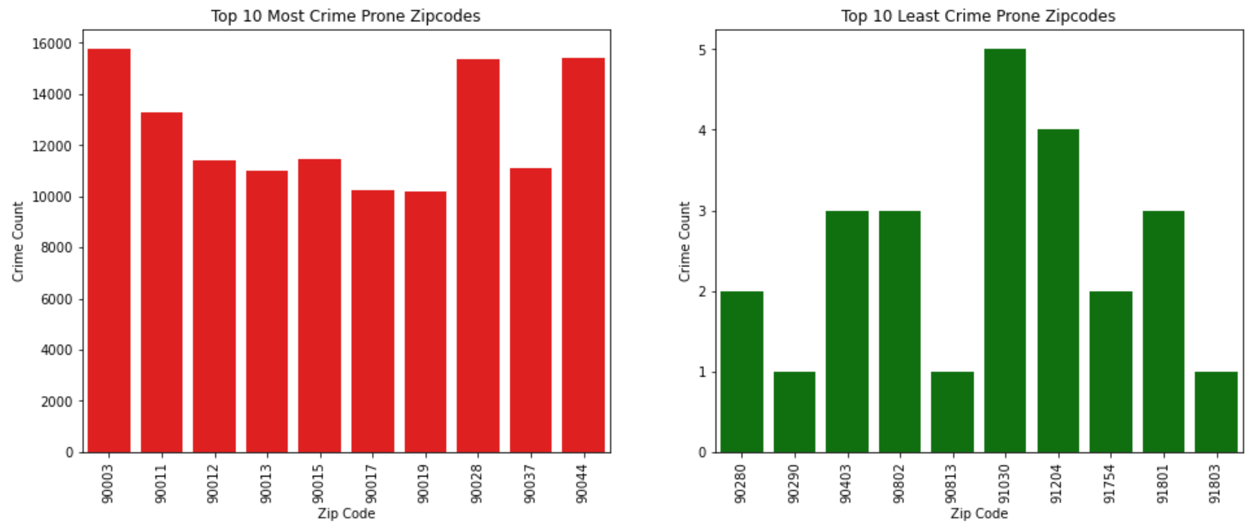
Comparison of Crimes with respect to Estimated Median Income:



Top10 and Bottom 10 zip codes that are crime prone:



Bar chart of top10 and bottom 10 crime prone areas:



<u>Code for Predicting 2024 crime rate and most important crime type as per police station:</u>

import pandas as pd c2024 = pd.read_csv('2024_predictions.csv') c2024.head() pdf = pd.read_csv('PS_Names_zips.csv') pdf.head() Final_pred = c2024.merge(pdf, on='ZIP_CODE') Final_pred.to_csv('Final_pred_2024_merged.csv', index=False) Final_pred

filter the rows for Year 2024

#df_Year = merged_df[merged_df['Year'] == 2024]

group the rows by ZIP_CODE and Year, and sum the CRIME_COUNT for each group

crime_counts_by_stations = Final_pred.groupby(['Police_Station_Name', 'Year'])['CRIME_COUNT'].sum().reset_index()

crime_counts_by_stations = crime_counts_by_stations.rename(columns={'CRIME_COUNT': 'TOTAL_CRIME_COUNT'})

print the resulting Series

print(crime_counts_by_stations) total_crime_counts = Final_pred['CRIME_COUNT'].sum() print('Total Crime Counts in 2024 are:', + total_crime_counts)

Define the total crime count and total number of officers

total_officers = 7070

Calculate the overall crime-to-officer ratio

crime_to_officer_ratio = total_crime_counts / total_officers

Multiply the crime count for each station by the ratio to estimate the number of officers required

crime_counts_by_stations['2024_OFFICERS_REQUIRED'] = (crime_counts_by_stations['TOTAL_CRIME_COUNT'] / crime_to_officer_ratio).round().astype(int) crime_counts_by_stations.head() crime_counts_by_stations ps_present = pd.read_csv('PoliceAllocation.csv') ps_present final_2024_off = pd.merge(crime_counts_by_stations, ps_present, on='Police_Station_Name', how='left')

final_2024_off final_2024_off.to_csv('2024_Final_Prediction_Stationwise.csv', index=False) import matplotlib.pyplot as plt

Plotting a bar chart

plt.figure(figsize=(12, 6)) plt.bar(final_2024_off['Police_Station_Name'], final_2024_off['TOTAL_CRIME_COUNT'], label='Total Crime Count') plt.bar(final_2024_off['Police_Station_Name'], final_2024_off['2024_OFFICERS_REQUIRED'], label='Officers Required') plt.bar(final_2024_off['Police_Station_Name'], final_2024_off['Officers_Allocated_Present'], label='Officers Allocated') plt.xticks(rotation=90) plt.xlabel('Police Station') plt.ylabel('Count') plt.title('Crime Count and Officers Comparison') plt.legend() plt.show()

from IPython.display import display

Displaying the dataframe as a table

display(final_2024_off) import matplotlib.pyplot as plt

Set the style of the plot

plt.style.use('ggplot')

Set the x and y values for the chart

police_stations = final_2024_off['Police_Station_Name'] officers_required = final_2024_off['2024_OFFICERS_REQUIRED'] officers_allocated = final_2024_off['Officers_Allocated_Present']

Create a horizontal bar chart

fig, ax = plt.subplots(figsize=(10,8)) ax.barh(police_stations, officers_required, label='Officers Required') ax.barh(police_stations, officers_allocated, label='Officers Allocated') ax.legend()

Set the labels for the chart

ax.set_xlabel('Number of Officers') ax.set_ylabel('Police Station')
ax.set_title('Comparison of Officers Required vs Allocated')

Show the chart

plt.show()

import matplotlib.pyplot as plt import numpy as np

create bar chart showing comparison between Officers_Allocated_Present and 2024_OFFICERS_REQUIRED for each police station

fig, ax = plt.subplots(figsize=(12,8)) index = np.arange(len(final_2024_off)) bar_width = 0.35 opacity = 0.8

rects1 = ax.barh(index, final_2024_off['Officers_Allocated_Present'], bar_width, alpha=opacity, color='b', label='Officers Allocated Present')

rects2 = ax.barh(index + bar_width, final_2024_off['2024_OFFICERS_REQUIRED'], bar_width, alpha=opacity, color='g', label='2024 Officers Required')

ax.set_yticks(index + bar_width / 2)
ax.set_yticklabels(final_2024_off['Police_Station_Name']) ax.invert_yaxis()
ax.set_xlabel('Number of Officers') ax.set_title('Comparison of Officers Allocated Present and 2024 Officers Required by Police Station') ax.legend()

plt.tight_layout() plt.show()

import matplotlib.pyplot as plt

Set figure size

plt.figure(figsize=(12, 8))

Create line chart for Officers_Allocated_Present column

plt.plot(final_2024_off["Police_Station_Name"], final_2024_off["Officers_Allocated_Present"], label="Officers Allocated Present")

Create line chart for 2024_OFFICERS_REQUIRED column

plt.plot(final_2024_off["Police_Station_Name"], final_2024_off["2024_OFFICERS_REQUIRED"], label="2024 Officers Required")

Set title and axis labels

plt.title("Comparison of Officers Allocated Present and 2024 Officers Required by Police Station") plt.xlabel("Police Station Name") plt.ylabel("Number of Officers")

Set legend

plt.legend()

Rotate x-axis labels for better readability

plt.xticks(rotation=90)

Show plot

plt.show()

type2024 = pd.read_csv('2024_predictions.csv') type2024.head() #df2423 = type2024.CRIME_COUNT.sum() #print(df2423) pdf123 = pd.read_csv('PS_Names_zips.csv') pdf123.head() Final_pred123 = type2024.merge(pdf123, on='ZIP_CODE') Final_pred123.head() Final_pred123.to_csv('PS_Names_Added.csv', index=False)

Group the data by police station name and category, and count the number of occurrences

grouped = Final_pred123.groupby(["Police_Station_Name", "CATEGORY"]).size().reset_index(name="Counts")

Find the category that occurs most frequently for each police station

max_category = grouped.groupby("Police_Station_Name").apply(lambda x: x.loc[x["Counts"].idxmax()])

Select the desired columns from the grouped data

output = max_category[["Police_Station_Name", "CATEGORY", "Counts"]]

Reset the index of the 'output' DataFrame

output = output.reset_index(drop=True)

Calculate the total counts of crimes in each police station

total_counts = Final_pred123.groupby("Police_Station_Name").size().reset_index(name="Total_Counts")

Merge the 'output' and 'total_counts' DataFrames on 'Police_Station_Name'

output = output.merge(total_counts, on="Police_Station_Name")

Display the output

print(output)

output['Total_Counts'].sum() output.to_csv('Final_2024_CrimeType_per_Station.csv', index = False)

filter the rows for Year 2024

#df_Year = merged_df[merged_df['Year'] == 2024]

import pandas as pd

set display options to show all columns in the same line

pd.set_option('display.max_columns', None) pd.set_option('display.expand_frame_repr', False)

your code here

group the rows by ZIP_CODE and Year, and find the most frequent CATEGORY value for each group

crime_counts_by_type = Final_pred123.groupby(['Police_Station_Name', 'Year'])['CATEGORY'].apply(lambda x: x.value_counts().idxmax()).reset_index()

add a new column showing the count of each CATEGORY value for each group

crime_counts_by_type['CATEGORY_COUNT'] = Final_pred123.groupby(['Police_Station_Name', 'Year', 'CATEGORY'])['CRIME_COUNT'].sum().reset_index()['CRIME_COUNT']

crime_counts_by_type = crime_counts_by_type.rename(columns={'CATEGORY': 'CATEGORY_TYPE'})

group the data by Police_Station_Name and Year, and sum the CRIME_COUNT for each group crime_counts_by_station = Final_pred123.groupby(['Police_Station_Name', 'Year'])['CRIME_COUNT'].sum().reset_index()

merge the two DataFrames based on Police_Station_Name and Year

result = pd.merge(crime_counts_by_type, crime_counts_by_station, on=['Police_Station_Name', 'Year'])

add a new column showing the total CRIME_COUNT per station

#result['TOTAL_CRIME_COUNT'] = result.groupby('Police_Station_Name')['CRIME_COUNT'].transform('sum')

print the resulting DataFrame

print(result)

total_crime_counts241 = result['CATEGORY_COUNT'].sum() print('Total Crime Counts in 2024 are:', + total_crime_counts241)