

# **LA Crime (Nine Neighborhoods) vs. Unemployment Report**

## **Introduction**

Crime in the United States of America has been a plague that has decimated families, communities, and even the progress of America becoming a more advanced country. Since its inception in 1776, the American Government has taken a reactive approach to crimes occurring within its coastlines, instead of a proactive approach, which would entail researching ways to prevent crimes from happening before they are committed. The reasons for crimes being committed in America, as well as around the world, are numerous and extensive. For this research project, we (Reilly Tabares and Willie Matthys) decided to conduct an analytic deep dive into the effects that Unemployment Rates in communities have on their respective crime rate/severity of crimes. Does higher unemployment lead to more severe crimes being committed? Does higher unemployment lead to more crimes being committed? These questions and many more will be answered throughout this report. If substantial evidence comes out of these questions, such as higher unemployment being correlated with a higher crime rate, steps can be taken to install nationwide programs to get citizens employed. For this report, we decided to use a sample size of the United States with nine neighborhoods from the City of Los Angeles, due to its population and diversity.

## **Data**

We retrieved the data below from two websites on crime in Los Angeles. One is a government dataset on crime in the Los Angeles area from 2020-2024. The other is a Los Angeles county dataset on unemployment levels in neighborhoods in the Los Angeles area.

## **Unemployment Data Source:**

“Unemployment (Census Tract).” *County of Los Angeles Open Data*, Los Angeles County, 19 Mar. 2025,

data.lacounty.gov/datasets/lacounty::unemployment-census-tract/explore?location=33.694994%2C-118.298767%2C6.48&showTable=true.

Link: [Unemployment Data](#)

### **Crime Data Source:**

“Arrest Data from 2020 to Present - Catalog.” *Data.Gov*, GSA’s Technology Transformation Services, 3 May 2025, catalog.data.gov/dataset/arrest-data-from-2020-to-present.

Link: [Arrest Log Data](#)

### *2.1 ~ L.A Crime*

For the specific crimes being committed in the areas, we collected arrest log data from Data.gov on the city of Los Angeles.

The website had information on the dataset, including where it came from (LAPD Open Data on Arrests). It also contained a convenient CSV file we downloaded and put into a pandas dataframe named "la\_crime". The next step was to clean the data to prepare for a merge with the other dataset containing the unemployment level by community. We first cleaned the data by stripping it of columns of irrelevance, such as certain arrest codes, latitude/longitude, and where the perpetrator was booked. Since the data was from 2020-2024, we only included 2023 dates because the second dataset on unemployment was based on 2023. The rest of the cleaning can be seen in the notebook (wmatthys\_rtabares\_DRAFT.ipynb). From this cleaning, we saved the dataframe to a new pandas dataframe named la\_crime2023.

### *2.2 ~ L.A Unemployment*

We retrieved the data on unemployment level from a website called data.lacounty.gov, which covered the unemployment rate by 'area' in the City of Los Angeles.

Although the dataset had an option to download the file, we felt it would be in our best interest to web scrape it so we could replicate our method on any website/dataset in the future. The scrape was relatively straightforward, although we had to implement a scroll function to load all 2,495 rows before we scraped and gathered the data. This function can be seen in the "wmatthys\_rtabares\_DRAFT" notebook under the "Second Dataset Scraped" header. We then loaded this data into Pandas with the name "unemployment\_df".

After we loaded the data into a Pandas dataframe post-scrape, it was time to prepare this data for merging with our L. A Crime dataset. We began by dropping columns of irrelevance, such as "Shape Area" and "Shape Length". Our only real columns of interest were "Unemployment Rate" and "Area", yet we changed the name of the area column to "community" to align the columns for a merge. Another big portion of this cleaning was editing the data of the now "community" column. The area names were more or less the same, but some had area definitions such as 'City of' or 'Unincorporated'. So, we eliminated every word besides the matching area name itself, and used a ".unique()" function to analyze the communities of both datasets to make sure they would align for merge. We also edited column types to match up for the cleaning and cleared whitespace out, but that was the extent of the cleaning of the dataset.

### *2.3 ~ The Merge of la\_crime2023 and unemployment\_df*

The merge for both of these datasets was a little messy. When we first began the initial merge, we were returned with zero data in our dataframe, or we would get the same community with one million instances. To bypass this error, we created a new dataframe from the unemployed\_df dataframe with a group by function. We grouped the dataframe by its community, and then for each community, we had the aggregate function of its mean unemployment rate, followed by having the supervisor district equal to one. This

returned us the unemployment rate for every community (see Figure 0). The supervisor district constraint was so multiple districts within the same community were not repeated. This gave us a clean merge once we grouped unemployment\_df with these constraints.

*Figure 0.*

	Community	Percent Unemployed
0	central	6.322222
1	harbor	5.233333
2	hollywood	6.651522
3	mission	3.950000
4	n hollywood	6.035897
5	norwalk	3.438095
6	pacific	2.440000
7	van nuys	7.477273
8	west la	1.600000

With unemployed\_grouped (group by unemployed\_df) and la\_crime2023, we merged the two on an inner join, and did it on the column 'Community'. Our merged dataset (la\_unemployed) returned 32,696 rows and 11 columns. We then replaced 'nan' with 'pd.na' to switch it to a Pandas recognizable NA value. We then dropped these NA values and trimmed our dataset to 29,517 rows and changed the name to la\_cleaned. This was the extent of our cleaning with the two datasets/the merge. Although we likely did not need all the columns in la\_cleaned, we kept them to open the door to future analysis questions that may come to mind due to their relevance of the information they provide on the crime. The final dataset is described in a data dictionary (Figure 1) below.

*Figure 1.*

Column	Type	Source	Description
--------	------	--------	-------------

Report ID	Numeric	la_crime2023	Unique crime ID
Arrest Date	Date	la_crime2023	Date of crime
Area ID	Numeric	la_crime2023	Unique ID of Community
Community	Text	both	Neighborhood crime was committed in
Age	Numeric	la_crime2023	Age of perpetrator
Sex Code	Text	la_crime2023	Sex of perpetrator
Charge Group Description	Text	la_crime2023	Group of Charge (homicide, burglary, etc..)
Charge Description	Text	la_crime2023	Specifics of charge (Ex. Robbery → Carjacking)
Address	Text	la_crime2023	Street address of crime committed
Percent Unemployed	Numeric	unemployed_grouped	Percent of population in the community that is unemployed
Supervisor District	Numeric	unemployed_grouped	District that the community is in.

## **Analysis**

### *3.1 ~ Unemployment Rate and Severity of Crime being committed*

***Null Hypothesis: There is NO correlation between unemployment rate and the severity of crimes being committed in a given neighborhood.***

A question that ties directly into the main premise of our research (Unemployment Rate and Crime) is "Does higher unemployment lead to more severe

crimes being committed?". We wanted to explore this question because if people's lives are in danger and there appears to be a correlation between the two, action must be taken immediately.

Before diving into the specific analytic techniques we used to answer the research question, we first need to define the constraints used to shape our solution. First, we conducted a sentiment analysis on the 'Charge Group Description' using VaderSentimentAnalysis. We decided on 'Charge Group Description' instead of 'Charge Description' because of the immediate indicators of violence that pop out (Weapon, Assault, Homicide), as well as the lack of filler words that could affect the sentiment analysis. In other words, the Charge Group Description column is more straightforward and to the point, especially for a text analysis, than the Charge Description column. Moving forward, we divided our text sentiment analysis into analyzing three communities. The three communities selected were chosen because of their diversity in unemployment across the data. West LA was chosen because of its 1.6% unemployment rate (low), as was Mission with an unemployment rate of 3.95% (medium), and finally, we selected Van Nuys with the highest unemployment percentage of 7.47% (high).

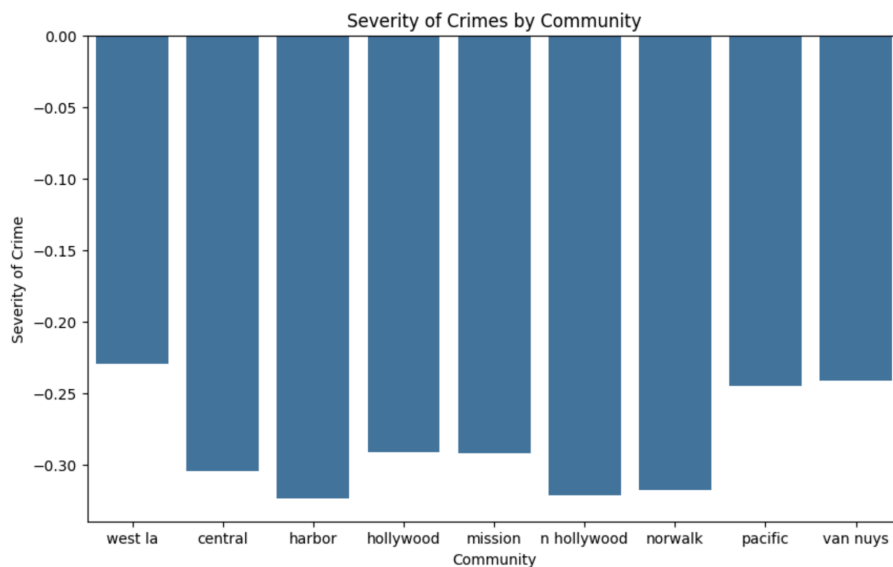
To see the results, we created a new column named 'Sentiment' and compiled the sentiment scores using a lambda and an apply function. We then took the average of the sentiment scores on individual crimes in their respective communities, and the results for the three communities are as follows:

West L.A (1.6% Unemployed) → -0.23 Vader Sentiment Analysis Score  
Mission (3.95% Unemployed) → -0.29 Vader Sentiment Analysis Score  
Van Nuys (7.47% Unemployed) → -0.24 Vader Sentiment Analysis Score

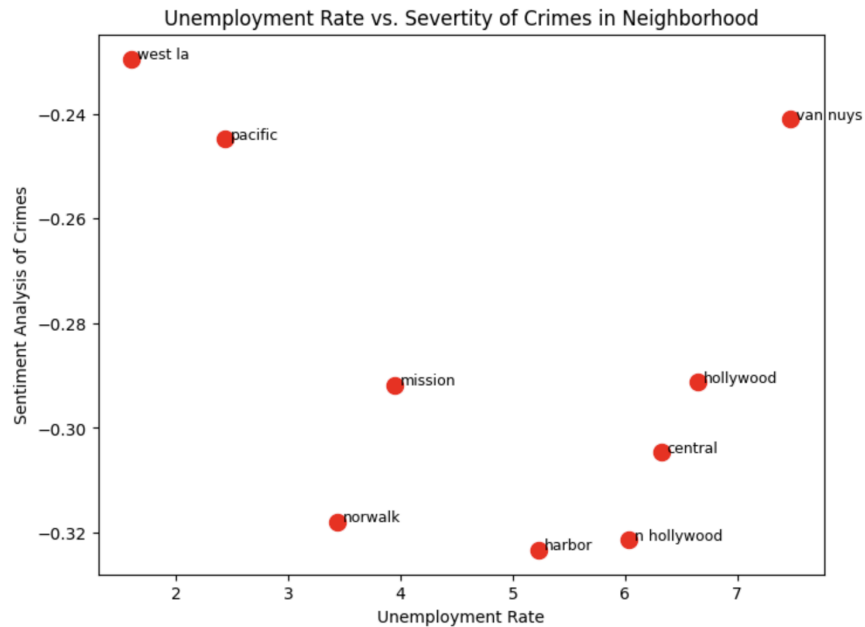
After seeing West L.A and Mission's scores, we were hopeful there would be a continued negative trend as the unemployment percentage went up, yet as seen above, it seemed not to matter for Van Nuys (the highest percentage of unemployed). Because the Sentiment trend was beginning to be promising before Van Nuys, we decided to test

it on every community and show the results in a graph for general trend viewing purposes (Figure 2). After creating the barplot seen in Figure 2, it was still a little uncertain to see a trend across the nine neighborhoods, so we decided to go with a scatter plot to heighten our trend view. From the scatter plot in Figure 3, the trend was not promising. We concluded these graphs of sentiment analysis of the crimes committed within each of the nine communities and their unemployment rate, with a correlation of unemployment rate being related to the 'violentness' of the crimes being committed. That correlation came out to be -0.344.

*Figure 2.*



*Figure 3.*



Due to its weak correlation, we failed to find any strong indication that the unemployment level affects the heinousness of the crime. To further conclude our findings, comparing the Unemployment Rate and Severity of Crime returned a p-value of 0.369. This leads us to fail to reject the null hypothesis that there is no correlation between unemployment and the severity of the crime. In simpler terms, there is no statistically significant evidence that the unemployment level in a given neighborhood in L.A affects the heinousness of the crimes being committed.

*3.2 ~ Does the Unemployment Level have an effect on the amount of crimes being committed in a neighborhood?*

***Null Hypothesis: There is NO correlation between unemployment level and the number of crimes committed.***

The next topic of research that we were led to was whether the unemployment level affected the number of crimes being committed. This research question was pretty straightforward, considering we found no statistical significance that the unemployment level had an effect on the heinousness of crimes. This led us to tackle the quantitative



approach (number of crimes) since the qualitative approach (heinousness) was taken care of.

We first started to answer this research question by creating a dataframe named 'la\_city\_crime\_summary' to compile all of the information we would need to answer the question of interest. We compiled this dataframe with a groupby function, and our three columns were "Community, Number of Crimes Committed, Percent Unemployed". Each community had its respective number of total crimes committed in the area, as well as its percentage of the population that was unemployed. We chose a scatter plot because of our interest in correlation, as well as the desire to see different neighborhoods more clearly from the rest (see Figure 4). In addition to Figure 4, we felt we needed to make a unique graph that would show the areas of the highest unemployment with the highest number of crimes. In Figure 5, you see a heat map, which shows the Community, its group of unemployment, and the color of its number of crimes (darker being higher). From the heat map, we can see trends that the darkest colors also seem to be the most unemployed. Due to both Figures 4 and 5 showing promising signs of correlation, we went on with the statistical significance tests.

After analyzing the scatterplot and heat map visually, it was clear that a positive trend in unemployment level and level of crime was apparent. So, we ran a correlation test that returned a value of 0.535. Although this correlation is not strong by any means, it is a moderate correlation that the higher the unemployment, the higher the number of crimes in the area. Keeping this moderate correlation in mind, we decided to conduct a 'p-value' test on the statistical significance of unemployment level leading to more crimes committed. The result was a p-value of 0.13. This led us to fail to reject the null hypothesis, meaning that there is no strong statistical evidence that the unemployment rate leads to more crimes being committed.

Although there was no statistical evidence of the amount of crimes being linked to the unemployment rate, we concluded that we need to keep an eye on the number of crimes as the unemployment rate fluctuates because of the moderate and positive correlation between the two, as seen in Figures 4 and 5.

*Figure 4.*

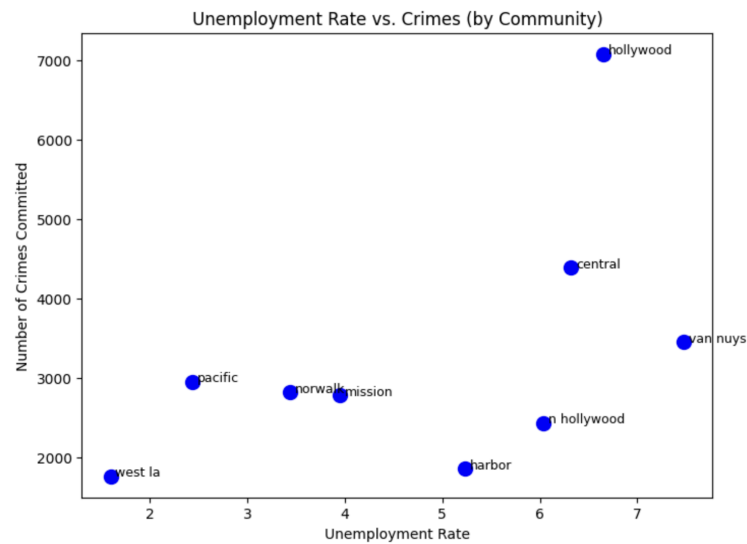
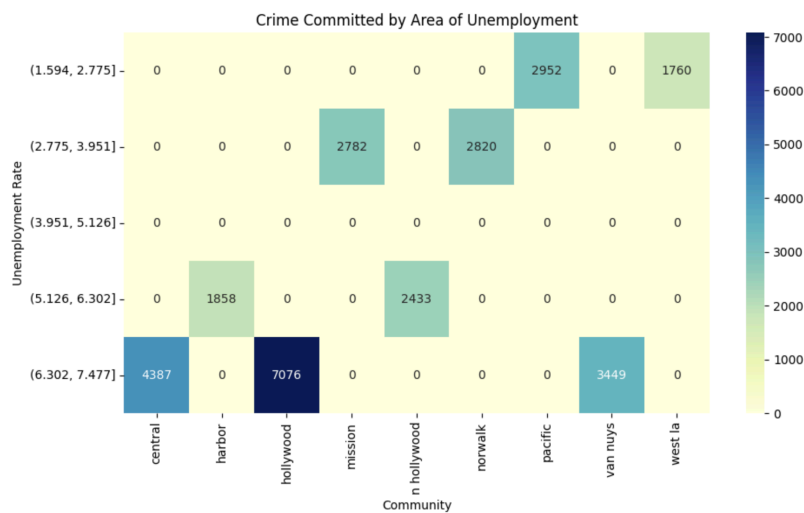


Figure 5.



### 3.3 Research Question 3: Can we predict if a crime is violent depending on factors such as age, sex, unemployment rate, and community?

Before building the solution to this problem, we first need to lay down a concrete foundation on what we are basing our target variable on. Our target variable is whether a crime is violent or not. To determine if a crime is violent, we decided to create a

dictionary of words that have violent undertones. An example of some of those words are Assault, Weapon, Robbery, Homicide, Injury, and many more. With words of violence, we created a column in `la_unemployed` dataframe named `'is_violent'`. Basically, if any row in the 'Charge Description' column has any of those words of violence in it, there will be a 1 as the value in the corresponding row of the `'is_violent'` column. This is for machine learning to be possible so it can numerically predict if a crime is violent or not with a 1 being yes, and a 0 being no.

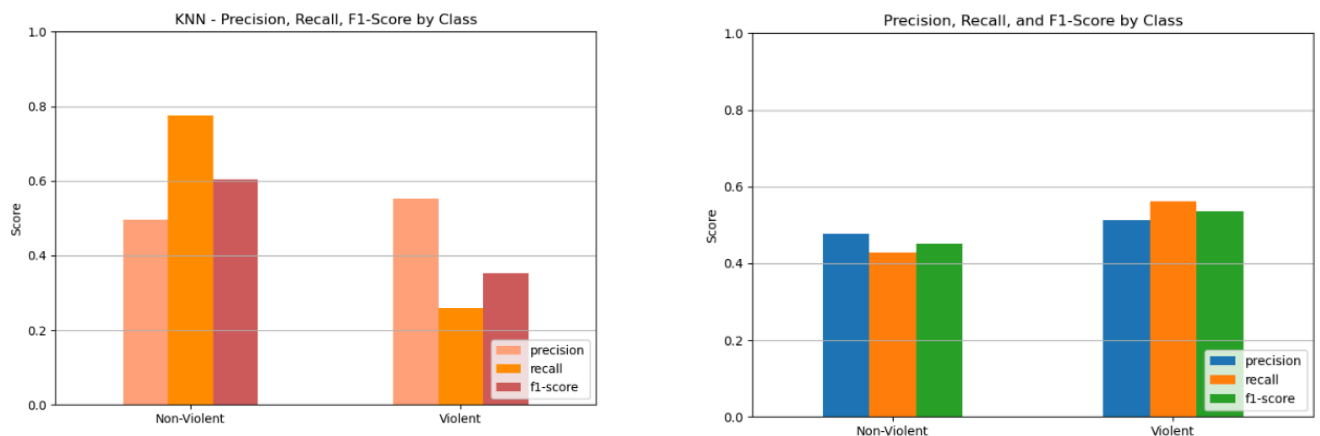
We first started to answer this research question by looking at words in the `la_cleaned` dataframe under the 'Charge Description' column'. We filtered our dataframe down to only include the columns 'Percent Unemployed' and 'Violent Crimes'. We compiled about 15 words that are often associated with violent acts, and put them under the `violent_crimes` list. From there we created a new column in the dataframe, categorizing each crime as violent (1) or non-violent (0). The number of non violent crimes was much higher than the violent crimes, so we balanced the dataset for testing purposes. Next, we split the balanced dataset into training and testing sets, applied feature scaling, and ran a binary logistic regression model.

After analyzing the binary logistic regression test, the results of the test were not especially strong. The confusion matrix (Figure 6) showed the model had difficulty correctly classifying violent crimes, and the precision scores (Figure 7) indicated moderate reliability at best. The precision for predicting violent crimes was around 0.49, meaning that just under half of the crimes labeled as violent by the model were actually violent in reality. The model's confusion matrix further demonstrated an even mix of true and false predictions, which supports the idea that using unemployment alone as a predictor does not provide enough separation between classes.

To validate these findings, we ran a second test using a K-Nearest Neighbors (KNN). KNN is useful in this context because it classifies new data points based on how similar they are to others nearby in the dataset. After scaling the data and running the model, the results were similar to those of the logistic regression (Figure 6.). Neither model demonstrated strong predictive performance using unemployment as the sole feature.

From this analysis, we concluded that although unemployment might have some association with violent crime, it is not a strong predictor by itself. The low precision scores and balanced confusion matrices suggest that other variables would likely need to be included for more accurate classification.

*Figure 6 (left) & Figure 7 (right)*



## **Conclusion**

To close the books on our crime project, we analyzed how unemployment percentages in nine different neighborhoods in Los Angeles impact the crime in those neighborhoods. In our analysis of how unemployment rates affect crime rates, we created the following questions for a deeper understanding of the unemployment rates relationship to crime:

**Question 1 ~** Does the unemployment level in a given neighborhood impact the heinousness of the crimes in the neighborhood?

There is no statistical evidence that higher unemployment leads to more heinous crimes being committed. Our statistical testing also returned a correlation coefficient of  $-0.344$  (negative because violence is portrayed with a negative). So, to conclude this research question: There is a weak correlation between unemployment levels relating to

the heinousness of crimes in certain areas, but it is not statistically significant enough to further research.

**Question 2 ~** Does the unemployment level in a given neighborhood impact the crimes committed in the specified neighborhood?

Through the visual optics of the scatter plot created to show the relationship between unemployment and quantity of crimes, as well as the correlation coefficient of 0.52, we have concluded that the two are moderately correlated. Even with a p-value of .13, statistics tells us not to reject the claim that there is no correlation between the two. Yet, with a topic as visceral and serious as crimes being committed, a moderate correlation is too hard to ignore. Due to this moderately strong correlation with only nine neighborhoods, we feel as if the unemployment rate does affect the number of crimes being committed.

**Question 3 ~** *Can the unemployment level in a neighborhood help us predict if violent crime is more likely to be committed?*

After the logistic regression and KNearestNeighbor tests, we have come to the conclusion that unemployment level in a neighborhood is not a good predictor of whether a given crime will be violent or not. Even with taking multiple different sample sizes, the results were not significant enough to make a clear indication that unemployment level can predict violent crime.

Looking back on the results of our project on unemployment vs. crime, it is hard to ignore the limitations of the data that could have produced more extreme results, or even stronger trends. For example, we only had data on nine communities in Los Angeles. For future research to be more compelling, we need a larger quantity of data on metro communities regarding the unemployment level. Looking forward, if we want to become proactive on crime in America, we need more centralized data on unemployment for samples as small as Los Angeles communities, so we can have more confidence in our results and more clarity for implementation.