4741 Final Report: Chicago Crime Scene Investigation

Yoo Jin Bae (yb92), Advika Ravi Kumar (ar732)

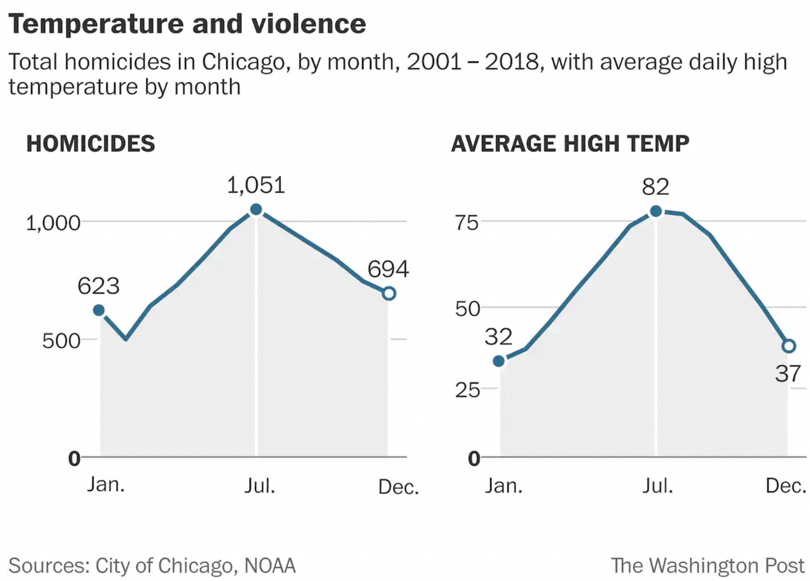
12, May 2023

1 Abstract

This paper focuses on predicting types of crimes based on weather. It has been established by many studies that there exists a linear relationship between temperatures and rates of crimes. Along with such positive existing study, it is especially worthwhile to look at the weather's impact on Chicago's crime rates, since Chicago is known for its variance in temperature throughout the seasons. With data analytics and machine learning techniques, this paper predicts the crimes to be committed with a positive correlation to the weather. Using a combined dataset that has a vast amount of data on crime and weather, it can be seen that there is a clear correlation between the features we used. However, our predictors still had a low accuracy. This study employs Least Squares Regression, Support Vector Machines (SVM) and feature engineering techniques to analyse the relationship between crime types and weather conditions in Chicago.

2 Background and Motivation

Chicago is undeniably linked to two key words: wind and crimes. It ranks 31st for the amount of crime out of approximately 108000 cities in the United States (2023), with a historical precedent that has exhibited an existing relationship between the city and crime (Fieldstadt, 2020). With the recent increase in the homicide rate in 2016 and during COVID-19 years (2020-2022), the 2.7 million citizens of Chicago are under constant fear of crime. In 2020, the city of Chicago has put forth a comprehensive plan to reduce violence in Chicago, proposing that it will implement various measures, such as improving and advancing policing and creating jobs and housing for those affected by violence, to prevent crimes and take care of those that were affected by crimes in the city (City of Chicago, 2020). These measures, however, will take years to implement, and while they may help prevent crimes from happening, they will not guarantee the citizens to feel safe. Moreover a recent study by The Washington Post suggested that higher or rising temperatures may lead to a spike in crime rates as shown in the graph below. We try to replicate these results and apply a prediction algorithm and other data analysis methods in our project.



Diving into the correlation between weather and crime would provide a much more practical solution to relieving the fear in the people of Chicago. Although there is no direct connection between the two that can be deduced on plain sight, there are studies that discovered the existence of a positive association between temperature and violent crimes (Corcoran, 2022). As the information regarding weather is accessible at any time to anyone, if there were to be a strong correlation, it would be a powerful and insightful indicator that could help relieve the worries of the citizens of Chicago.

3 Datasets

There were two datasets used in this project: the Crimes in Chicago dataset and the NOAA Online Weather Data dataset, as there were no pre-existing datasets that include both crime and the weather of the day the crime happened. By merging two datasets that contain daily records, each of crimes and the weather characteristics in Chicago, it would provide us with strong base for our data analysis.

The Crimes in Chicago dataset was acquired from the Chicao Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. It contains 7,755,580 entries of reported crime incidents that occurred in Chicago from January 2001 to March 2023. Some of the features included in the dataset, such as *Date and Primary Type* would be useful in determining the number of crimes and types of crimes.

Chicago's daily weather dataset extracted from the NOAA Online Weather Data was obtained from the National Centers for Environmental Information website. It contains 8160 entries of Chicago's daily weather data from January 2001 to May 2023. It contains various weather-related features, from the *average temperature* and *precipitation* to *wind speeds*, that were measured daily from the West Chicago Dupage Airport.

3.1 Data Cleaning

To clean our data in preparation for further analysis, we first explored the two datasets and searched for any missing values. We checked the percentage of missing values in each columns, and dropped the ones that had more than half of its column as null. Then, we looked at the rows to eliminate rows with more null values than existing values. We also improved the average temperature feature by regenerating the average between the provided maximum temperature and minimum temperature features in the weather dataset, as approximately 70% of the average temperature column was null, but the two components that the average temperature is calculated from had no null values. Features that yielded no significant information, such as the *STATION*, *LATITIUDE*, and *LONGITUDE* features from the weather dataset that had the same value for the entire column, the weather type attribute features that consisted of random characters, and features regarding the criminal information pertaining to each crime listed in the crime dataset, were removed. After that, the datatypes were looked at. After that, having discovered that the date attribute for both datasets were not of type datetime, it was converted to datetime, as the crime dataset and the weather dataset have to be merged using the date feature. While examining the dataset, we wanted to compare two ways of creating a merged dataset, so the first merged data had the weather attributes and the number of crimes that happened in each day, and the second merged dataset focused on the crime types that happened each day, alone with the daily weather data pertaining to that day.

3.2 Exploratory Data Analysis

To explore the two datasets we had, we first explored the correlation of pairs of all features in each of the two merged datasets using the heatmap:

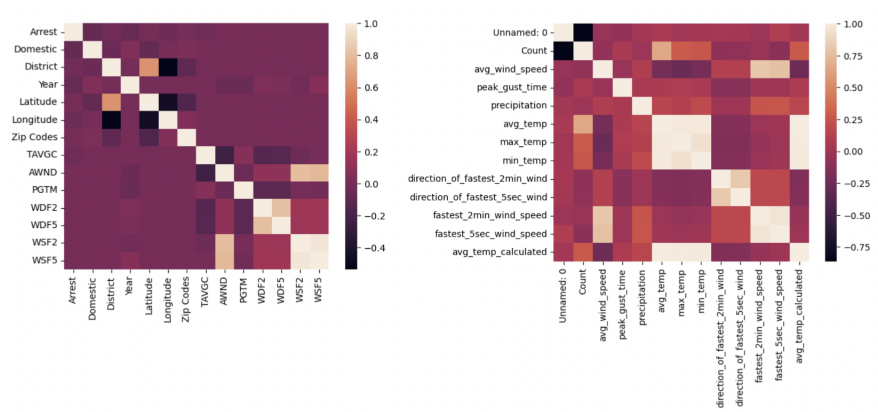


Figure 1: Heatmaps

The heatmap on the left of figure 1 is the merged data with the focus on the feature types and the heatmap on the right of figure 2 is the merged data with the focus on the number of crimes. The correlations can be easily identified on the heatmap with colours; the lighter the colour, the higher the correlation between the two features. We can easily see that there is no significant correlation available on the left heatmap, which makes sense, as the features are not numerical values. As for the heatmap on the right, although there are no significantly light patches in correlation to *Count,* we can see that there are some relatively lighter patches of colour.

The next exploratory data analysis was done using a pairplot:

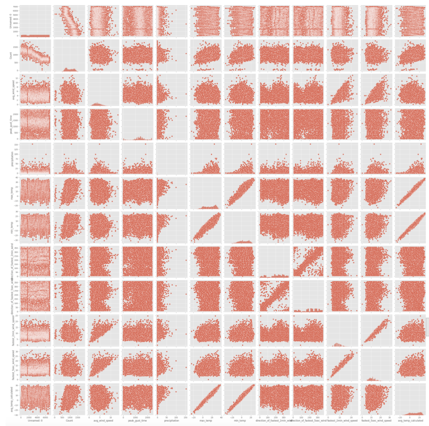


Figure 2: Pairplot

For the pairplot (figure 2), only the merged data focusing on the daily crime count was used, as through the heatmap, we realised that the data visualization for the merged data that focuses on the crime types would not yield significant information about the dataset. In figure 2, we can see that the results align with what we discovered with figure 1; when looking at the *Count* feature, we can see that there is a slightly linear relationship being portrayed with the daily average temperature feature.

4 Data Analysis

After cleaning and running the exploratory data analysis on the dataset, three machine learning techniques were applied to the dataset in order to achieve our goal of determining a significant correlation between crimes in Chicago and the weather.

4.1 Ordinary Least Squares Regression

For our first model, we chose the ordinary least squares, or OLS, regression, as it was discovered during the exploratory data analysis that there is a possible relationship between the crime count and the daily average temperature. Running an OLS regression will allow us to see whether the crime count is dependent on the temperature.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Count   R-squared:                       0.100
Model:                            OLS   Adj. R-squared:                  0.099
Method:                 Least Squares   F-statistic:                     749.5
Date:                Sat, 13 May 2023   Prob (F-statistic):         1.30e-156
Time:                        02:20:04   Log-Likelihood:                -46862.
No. Observations:                6779   AIC:                         9.373e+04
Df Residuals:                    6777   BIC:                         9.374e+04
Df Model:                           1
Covariance Type:            nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                   911.9330      3.915    232.943      0.000     904.259     919.607
avg_temp_calculated       7.2839      0.266     27.378      0.000       6.762       7.805
==============================================================================
Omnibus:                      591.393   Durbin-Watson:                   0.183
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              187.470
Skew:                           0.062   Prob(JB):                     1.96e-41
Kurtosis:                       2.195   Cond. No.                         19.5
==============================================================================
```

Figure 3: Ordinary Least Squares Regression Summary

The model's summary in figure 3 shows that the r-squared value is 0.100, meaning that the daily average temperature explains approximately 10% of the variability of the crime count. The p-value is displayed to be 0.000, indicating that the relationship between the crime count and the daily average temperature is statistically significant, as low p-values, specifically those that are less than $\alpha = 0.05$ show the likelihood of the predictor being a meaningful addition.

A visualisation based on the model was generated to verify the OLS regression results:

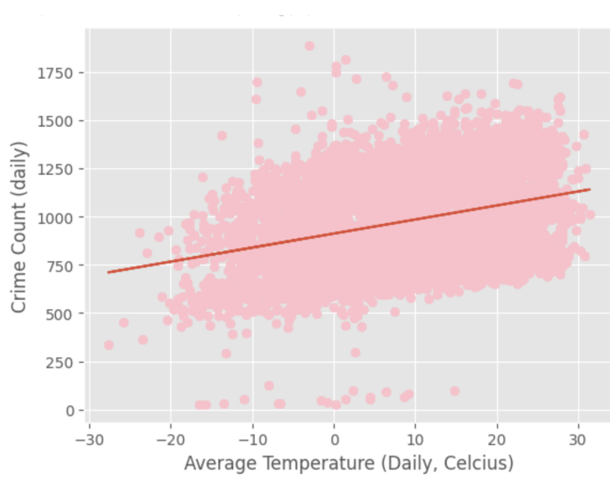


Figure 4: Ordinary Least Squares Regression Plot

The scatterplot (figure 4) displays a fitted regression line over the pairs of data points from the dataset displayed in pink. The scatterplot (figure 4) confirm that the fitted regression line properly encapsulates the relationship between the daily average temperature and the crime count. The fitted regression line highlights the positive linear relationship between the two features, bringing forth the conclusion that higher temperature leads to more crime.

4.2 Feature Engineering

In order to prepare the data to be analyzed and to apply machine learning and other techniques to it, we are using feature engineering techniques to clean and get the data in the right format to be used. This should enhance the performance of our models and increase their accuracy, especially since we plan on using SVM to analyze the data.

We started by assessing the missing values, since we are working with crime data, a lot of it can be hidden due to privacy reasons. We decided to exclude any of these rows when analyzing our data. On the other hand, for the weather data, we impute missing values using statistical measures such as mean since this data is more reliable and can be used to generalize weather patterns.

One-hot encoding and many-hot encoding were other methods that were used to identify categorical data in the datasets such as crime type. This was then used in our SVM model. This was done in order to avoid bias and then we performed scaling methods as well to ensure this.

Overall, feature engineering played a crucial role in the data preparation phase, increasing the quality and usability of the data for analysis. It enabled us to extract valuable information, reduce bias, and ensure that the SVM model could leverage the dataset's characteristics to deliver meaningful insights into crime analysis based on weather conditions.

4.3 Support Vector Machine

One method we use to analyze this data is a Support Vector Machine (SVM). SVM is a supervised learning algorithm used for classification or regression tasks. We try to find a hyperplane that best separates the input data into different classes using an SVM. We chose to use SVM here because of its ability to handle high-dimensional data and since it can capture non-linear relationships between features and the target variable using kernel functions. SVMs also have several hyperparameters that can be tuned to optimize their performance. We chose to work with the regularization parameter (C) which controls the trade-off between maximizing the margin and minimizing the classification error. A larger value of C leads to a narrower margin and fewer misclassifications which helps balance the bias-variance trade-off and prevent over or underfitting.

We begin the process with our cleaned dataset and extracting the relevant information. By selecting features such as 'CrimeType', 'Temperature', 'Precipitation', 'Humidity', and 'WindSpeed' we are able to get the data ready for the SVM by splitting the data into features X and target variable Y. After processing this data we use a train test split of 70 percent of the data and the remaining 30 percent of the data. We also perform feature scaling so that it helps ensure that all features contribute equally to the model's training process.

After completing this process we are able to make predictions on the testing set and evaluate the model. The results we achieved in this process were that the accuracy is 0.2035. This indicates poor model performance and suggests that the model is not able to effectively learn and generalize patterns from the data.

5 Conclusion

Through our project we intended to get a better understanding of the correlation between crime types and weather patterns. Our findings were intended to be used by the government to decrease crime rates around the Chicago area.

5.1 Weapons of Math Destruction

Through exploratory data analysis, although not the most distinct, there was a clear observed correlation between weather and crimes committed. Since the predictions we make using our model has a low accuracy, however, using our predictive model could seriously impact the safety of the citizens. The model has the potential to become a weapon of math destruction, posing a serious threat to citizens in the case that resources are redistributed unevenly and may even achieve the opposite by creating a spike in crime rates which defeats the purpose.

5.2 Fairness

Fairness aims to ensure that predictions and decisions made by models do not result in unjust or biased outcomes that disproportionately impact certain groups or communities. Our model tries to account for this through the use of feature engineering. Feature scaling in SVM specifically ensures that the predictions are not trained on biased data. Our method does employ techniques to avoid biases and ensure fairness. Nevertheless, our model's performance suggests that the model is not able to effectively learn and generalize patterns from the data.

5.3 Future Improvements

In terms of accuracy, we do have a lot of scope for improvement. A low R-squared value from the ordinary least squares regression method indicates that the regression model is not able to explain a significant portion of the variability in the dependent variable. The SVM also gave us an accuracy score of 0.2 which shows that there is great room for improvement. One way to do this may be to find other factors that we think might affect crime rates and use that as well for features of predicting the target variable. There are also other methods that we could consider using instead for cleaning the data and making prediction. One particular method we could consider is to apply regularization techniques, such as L1 or L2 , to penalize complex models and prevent overfitting. This can improve generalization to unseen data, and avoid the inclusion of unnecessary or noisy features. This is one of the problems that was faced by the Moreover, a correlation may not necessarily imply causation so we should also consider rethinking and researching our topic in greater depths.

6 Citations

City of Chicago. (2020). A Comprehensive Plan to Reduce Violence in Chicago. Chicago, Illinois.

Corcoran, J., Zahnow, R. (2022). Weather and crime: A systematic review of the empirical literature. Crime Science, 11, Article16. https://doi.org/10.1186/s40163-022-00179-8

Fieldstadt, E. (2020, November 9). The most dangerous cities in America, ranked. CBS News. Retrieved March 19, 2023, from https://www.cbsnews.com/pictures/the-most-dangerous-cities-in-america/2/

Max, H. (2023, February 21). CHICAGO CRIME SPIKES IN 2022, BUT FIRST DROP IN MURDER SINCE PANDEMIC. Illinois Policy. Retrieved March 19, 2023, from https://www.illinoispolicy.org/chicago-crime-spikes-in-2022-but-first-drop-in-murder-since-pandemic/

The Washington Post. Two new studies warn that a hotter world will be a more violent one. (2019, July 16). https://www.washingtonpost.com/business/2019/07/16/two-new-studies-warn-that-hotter-world-will-be-more-violent-one/

United States Cities Database. simplemaps. (2023, January 31). https://simplemaps.com/data/us-cities: :text=Up