Data 200

Principles and Techniques of Data Science

Data 200 Graduate Project Writeup

Analyzing Weather and Number of Arrests

Gavin Chan (gavin_chan@berkeley.edu, 3033605839)

Andrew Gorin (andrew_gorin@berkeley.edu, 3036271146)

Xinchen Lu (xinchen_lu@berkeley.edu, 3035253599)

December 9, 2022

Abstract

As previous studies have found a positive relationship between daily temperature and crimes, it is likely that the current and future climate change will lead to more crimes and other human conflicts. In the present study, we used the FBI NIBRS Crime Dataset and NOAA Daily Weather Dataset to test if we could build a model to use weather and climate data to accurately predict the number of arrests in different parts of the U.S from 2018 to 2020. Our results indicated that there is a positive correlation between temperature and crimes across the US, and in several states such as Michigan and Texas as well. As we tried to build models to predict the crimes, we found that, very likely, violent crimes are more related with temperature compared with non-violent crimes. Our results confirmed previous findings of the relationship between temperature and crimes, and pointed to the likely consequences of climate change on crimes in the US.

Introduction

Most global warming "doomsday" scenarios invoke social upheaval provoked by climate change. While this seems plausible to many, the precise form that this nebulous "upheaval" may take remains less clear. To date, academics have combed all manner of data looking for possible manifestations of the beginning of this decline. A common thread in many arguments emerges; violent crime, excessive heat, and drought may be linked, and the extent to which this is the case is not yet understood.

In perhaps the most well-known example of this, Kelley (2015) argues that the Syrian Civil War, which began in 2011, was incited by a drought that can be statistically demonstrated to be the result of global climate change. In fact, this drought was the worst in the region in the instrumental record. They suggest that this drought set off migration of farm workers to the capital city of Damascus. These migrants were of a different ethnic background than a majority of the residents already residing in Damascus, provoking racial tensions. Combined with heat, this situation was argued to be a tinder box created by climate change. While some counter that this analysis was flawed (e.g. Meyer, 2018), and others suggest that these sorts of human conflicts are too multifaceted to be quantified with such crude metrics (Selby, 2017), the fact that such vigorous debate exists is the primary motivation for this project.

Understanding the extent to which extreme climate conditions, which are predicted to become more frequent in future decades, affect human behavior will be crucial to mitigating the worst effects of these conditions. While some publications have sought to review how extreme weather conditions affect a wide range of human failings like civil conflict, minority expulsion, political instability, and intergroup civil conflict (Hsiang, 2013), our is project is more focused and aims to understand the relationship between excessive heat and violent crime in the United States. More specifically, we aim to build a model that uses weather data to accurately predict the number of arrests for all reported crime and specific types of crime. Between this report and the associated Jupyter Notebook where we performed our analysis, we intend to showcase the Data Lifecycle and its applications to this research question.

Previous Work

The relationship between weather and crime has been studied within the U.S. since the 19th century, with many studies pointing out that temperature is one main driving factor in this relationship. For instance, Anderson et al. 1997 found a positive correlation between the yearly temperature and serious and deadly assault observed during 1950 to 1995.

Currently, there are two major hypotheses of the relationship: the General Affective Aggression Model (Anderson et al., 1996)—and the routine activities theory (Cohen and Felson 1979). The general affective aggression model predicts that, with increased temperature, physiological heat stress would lead to the more aggressive actions. The routine activities theory, on the other hand, is a model that comprehensively connects the relationship between temperature and crime through the interactions between a motivated offender, a potential victim and the lack of deterrence.

Past studies have demonstrated that, with current climate change, the increased temperature projected could lead to more violent crime. For instance, Hsiang et al., 2013 quantified that each one standard deviation in climate variables would lead to interpersonal violence to increase by 4% and intergroup conflict to increase by 14%. Similarly, Ranson et al., 2014 quantified that, in this century alone (2010-2099), climate change will cause an additional 22,000 murders, 180,000 cases of rape, 1.2 million aggravated assaults in the US.

However, there are a few things that we would like to focus on in the project. As described in the Introduction, we especially want to understand the differences between violent crimes and non-violent crimes and their relationships with temperature, in the context of climate change in the future.

Description of Data

This project draws its data from two data sources: the Daily Summary from the Global Historical Climate Network Dataset (listed as Topic 2: Dataset A (Climate and the Environment - General Measurements and Statistics) on the graduate project website) and the NIBRS Crime Dataset.

The Federal Bureau of Investigation's National Incident-Based Reporting System (FBI NIBRS) is the nation's most complete crime and arrests database and is the successor of the FBI Summary Reporting System (SRS), which simply tracked summary statistics about crime and arrests in each state. This older system, the SRS, dates back to 1927, and was still used by many states until 2021. Both of these systems rely on states to individually and voluntarily report their arrest records or summaries of arrest records to the federal government. States are allowed, but not required, to mandate that local law enforcement participate in this data collection. It's important to consider that, if significant numbers of police departments do not participate, our interpretations of the data could be incorrect, as departments who are willing to participate may be different than those who are not.

Importantly, in 1998, states were encouraged to participate in the first iteration of NIBRS. Come 2015, the Federal Bureau of Justice Statistics announced that the SRS would be phased out completely by 2021; all crime statistics would need to be submitted in this newer form. Despite the significant notice from the Bureau of Justice Statistics, only about 66% of the nation's population was actually accounted for in this database by the year 2021 (Figure 1), when the

SRS was retired. To combat the potential bias-inducing effects of these reporting issues, we focused our analyses on states where the percentage of local agencies reporting was at least 98%. It's rather interesting to observe that California, New York, and Florida are the states with the least amount of crime statistics reported; Los Angeles County and New York County both did not participate in these reports. Hopefully these issues will be ameliorated in future iterations as the country becomes more attuned to the intricacies of this reporting system.

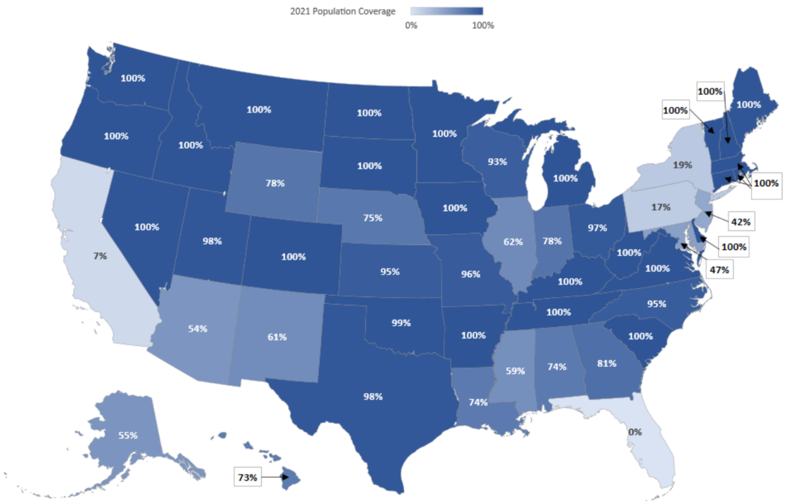

Figure 1: Coverage of Reporting by States to NIBRS System

The weather data used for this project is a bit more straightforward. Our weather data comes from the Global Historical Climatology Network (GHCN), which is an "integrated database of daily climate summaries from land surface stations across the globe" (NCEI, 2022). To ensure that the data are reliable and accurate, the data are routinely subjected to a suite of quality assurance reviews. While this database spans the entire globe, we focus on using temperature records in the United States.

The course project provided us with a year worth of data from this source, but in order to better understand longer-term temperature averages, and how the temperature on any given day compared to those, we decided it would be best to download additional years' worth of data. To this end, we explored this dataset with data from years 2010-2021.

Exploratory Data Analysis / Interesting Findings of Individual Datasets

The GHCN Weather Dataset only includes latitudes and longitudes of locations. Therefore, we had to use the reverse_geocoder package to map these locations to cities, states, and countries. After doing so, we can plot the weather data over time. For example, we can graph the average daily temperature of Texas from the beginning of 2010 to the end of 2021, as shown in Figure 2. As expected, we see that daily temperatures rise and fall seasonally. We can average such data over all of the years to graph the average temperature with respect to the day of the year, also shown in Figure 2.

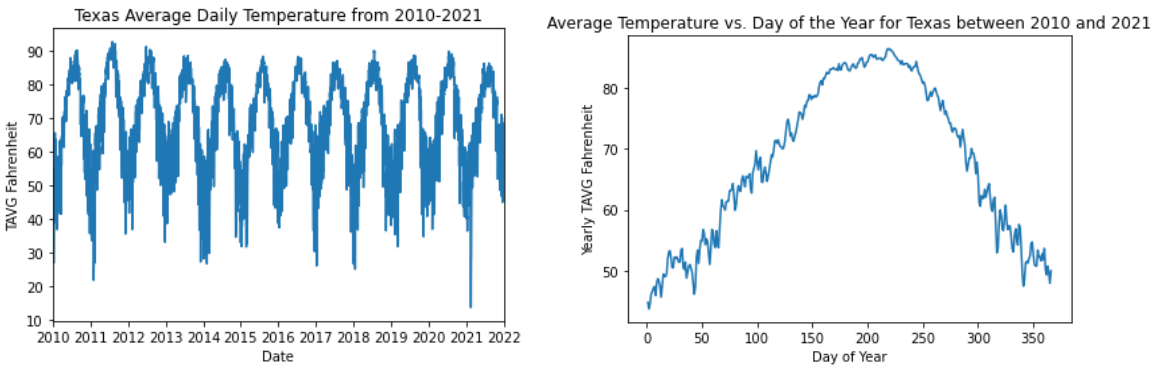

Figure 2: Texas Average Daily Temperature from 2010-2021 & Average Daily Temperature vs. Day of Year.

For the NIBRS Crime Dataset, first we note that each year's data is stored in large text files, each text file being multiple gigabytes. Each arrest is stored in a line of the text file, with associated information encoded in the column positions of each line. This required significant work in parsing, and the associated help file is also stored as an image. Therefore, manual conversions were required to change the text file into a Pandas dataframe, and the code can be found in the Jupyter notebook.

However, after parsing such data, there are some interesting statistics we would like to point out. For example, between 2018-2020, there were 12108897 unique incidents that had reported arrests. For one particular incident (encoded 'H3-F3FLGRVSC'), there were 152 arrests. Additionally, Table 1 shows the top 5 most common offenses between 2018 and 2020.

Table 1: Top 5 Offense between 2018-2020

Offense	**Number of Occurrences**
Simple Assault	2835704
Destruction/Damage/Vandalism of Property	2568366
All Other Larceny	2434626
Drug/Narcotic Violations	2080385
Theft from Motor Vehicle	1672188

We can then graph the number of arrests over time, which is shown in Figure 3. There appears to be some periodically changing number of arrests over time (similar to how temperature changes over the course of the year) leading us to believe that there is some relationship between temperature and number of arrests. We can further compare between a violent and nonviolent crime to see if there is seasonality in both types of offenses. Specifically, we chose 'False Pretenses/Swindle/Confidence Game' (non-violent) and 'Aggravated Assault' (violent), and graph the results, as seen in Figure 3.

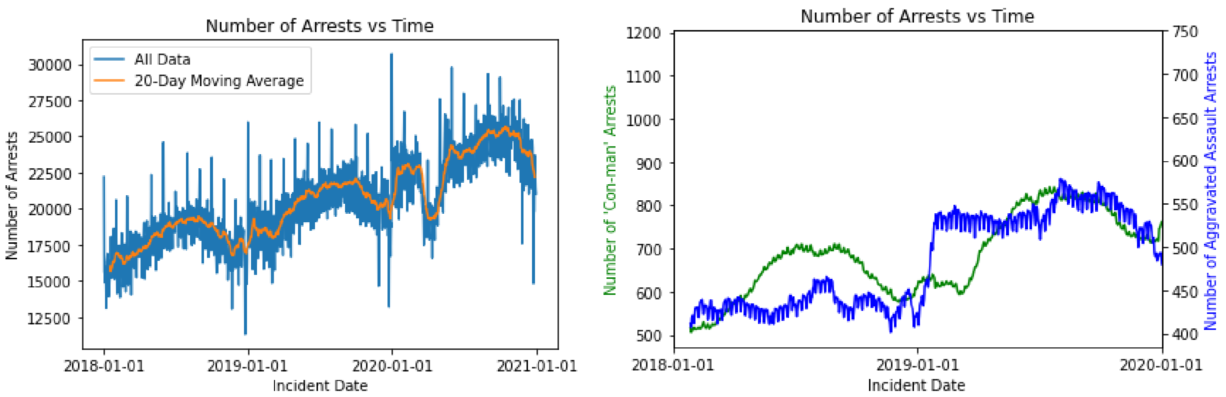

Figure 3: Number of Arrests over Time (Total and Violent vs Nonviolent Crime) Between 2018-2019

There does appear to be some seasonality in both of the data, with a large jump in aggravated assault arrests. This could come from a number of different factors, such as a number of reporting agencies increasing.

When combining the two datasets, we can then plot the average daily temperature versus the number of arrests per day. This is shown in Figure 4. When plotting average temperature with respect to a particular state and or crime, we generally see relationships. That is, as the temperature increases, the number of crimes also generally increases. This motivates the use of a linear model for prediction between these two variables.

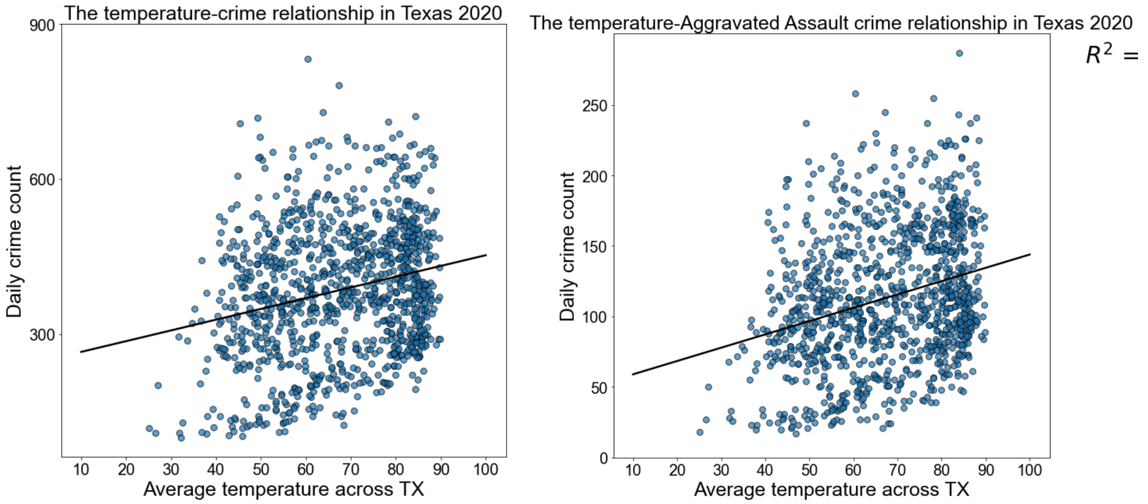

Figure 4: Average Daily Temperature vs. Crime Relationship (All and Aggravated Assault) for Texas in 2020

Description of Methods / Methodology

Before investigating whether we can build a model to accurately predict the number of arrests using climate data, we first discuss *causal inference*. Causal inference is when one attempts to infer whether one variable causes another. In other words, by modifying one variable and observing another, causal inference can be used to relate them. However, it is important to note that causal inference has its drawbacks. One needs to make assumptions in order to draw conclusions about these relationships, and those assumptions may not be correct. In the case of our study, we are making the implicit assumption that temperature is a primary driver of the number of arrests in a given location. This is obviously incorrect, as many other variables such as population density, geographic location, age, local policing priorities, and demographic makeup of the local populus may affect this relationship as well. Despite these clear limitations, we will still attempt to describe a relationship between daily temperature and number of arrests.

Modeling Temperature vs. Number of Arrests

To model our data, we chose to use ordinary least squares (OLS) as our model. As discussed in our Exploratory Data Analysis Section, there is a linear relationship between daily average temperature (℉) and the number of arrests across the entire United States and per state. This demonstration is made despite incomplete reporting in the NIBRS dataset. Therefore, by building a linear model with temperature as our feature and number of arrests as our prediction, we produced a linear model of $Y = \theta_0 + X\theta_1$. By splitting the data into training and test data and then training our model appropriately, we were able to plot the relationship between our test data point and the model predictions, as shown in Figure 5.

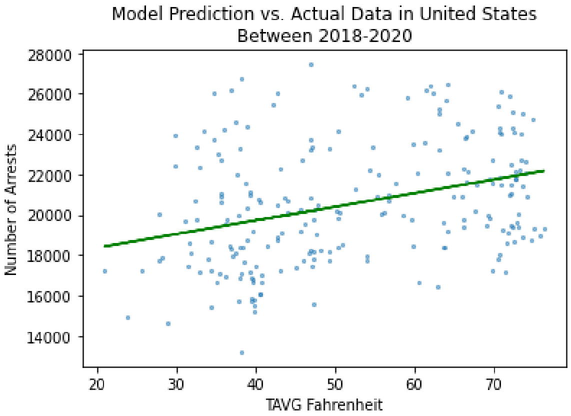

Figure 5: Model Prediction vs. Test Data in United States Between 2018-2020

When plotting the Residuals in Figure 6 (both the Actual vs. Predicted and Residual Plot), we see that our data is scattered randomly, as desired.

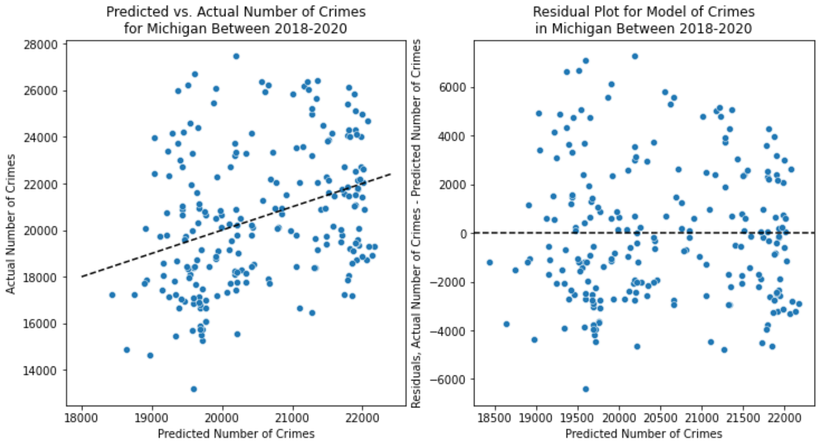

Figure 6: Residual Plots for United States Between 2018-2020

However, as the temperature across the United States spans a large range of values. We were concerned that different populations do not experience a constant temperature the same way. For example, the local population in Phoenix, AZ is likely to be experienced at dealing with 100 degree temperatures, while the population of Ann Arbor MI may not.

To address this concern, we chose to focus on a subset of states which had both a large range in yearly temperature conditions, as well as a 100% NIBRS reporting rate. Michigan's weather changes significantly between the summer (hot and humid) and the winter (cold and dry), and the state also has a 100% NIBRS reporting rate, which makes it an ideal candidate for study. Plotting both the temperature vs. number of arrests and the model's output and test data, as shown in Figure 7, we can again see a linear relationship between the two variables.

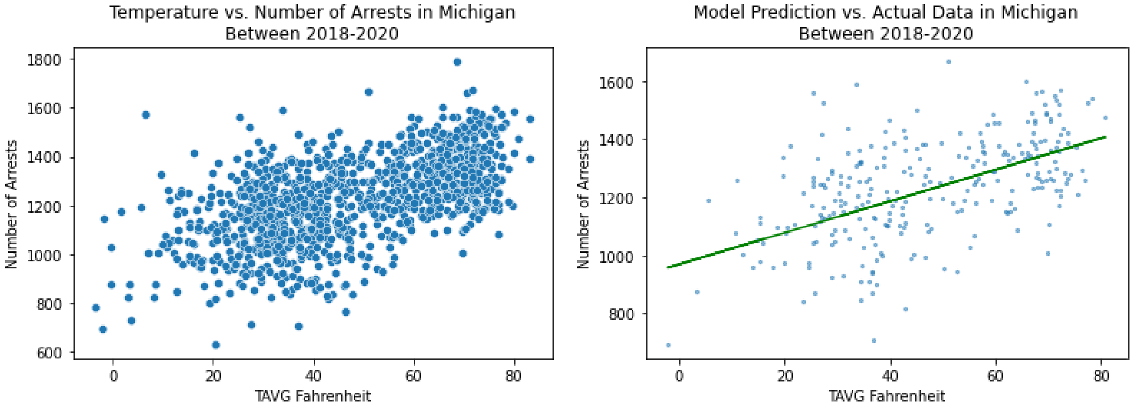

Figure 7: Temperature vs. Number of Arrests and Model Prediction for Michigan Between 2018-2022

To evaluate the performance of our models, we used RMSE, as a large RMSE is related to a badly-fitting model while a smaller RMSE is related to a more well-fit model. We can determine underfitting and overfitting by comparing the RMSE for our training data vs. predictions and test data vs. predictions (using the Holdout method). The associated RMSE values can be found in our Jupyter Notebook and in the Summary Section.

We then attempted to improve our models using both feature engineering and regularization. First, we used Day of the Year as a feature for OLS. As shown in Figure 8, there is a (non-linear) relationship between day of the year and number of arrests in Michigan.

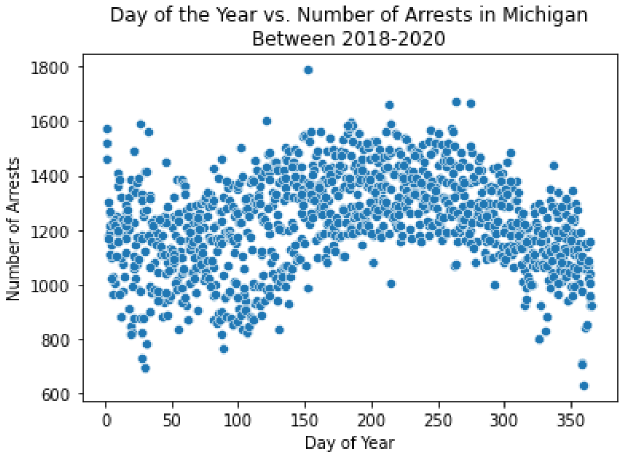

Figure 8: Day of the Year vs. Number of Arrests in Michigan Between 2018-2022

Because we do not want the actual value of the day of the year to affect the model, we one-hot encoded this variable as a feature in our OLS model. However, upon training a model and computing the test data RMSE, we found that this feature significantly overfit the data. It is important to note that, since we are using the day of the year as a feature to the model, each day of the year only has three data points (2018, 2019, and 2020). Therefore, these data are *too* specific, which leads to overfitting. We attempted to use Ridge Regression to penalize such highly weight coefficients, but found this did not decrease our RMSE. A discussion of these results can be found in the Summary Section.

To further combat the overfitting, we engineered a different feature which one-hot encoded the month of the year. We figured that this would give each feature more data, and might improve model fit. As shown in Figure 9, there is still a clear relationship between the month of the year and the Number of Arrests in Michigan.

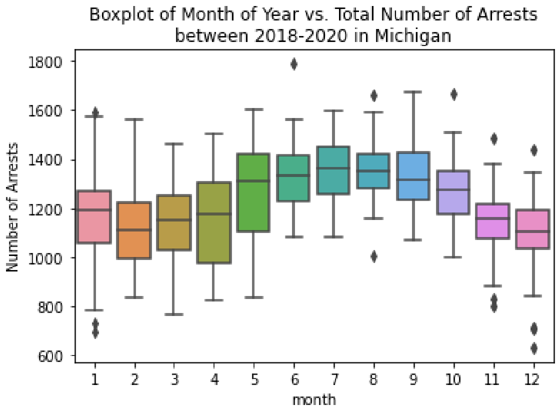

Figure 9: Month of the Year vs. Number of Arrests in Michigan Between 2018-2022

Again, this did not decrease our RMSE, likely due to collinearity between month and daily average temperature.. A further discussion of results can be found in the Summary Section.

Models for Violent vs. Non-violent Crime

To understand how heat affects different types of human behavior, we explored an even more specific question: Does temperature affect violent and nonviolent crime rates differently? In Figure 10, we compare the relationship between temperature and aggravated assault (violent crime) and drug/narcotic violations (non-violent crime) in Michigan.

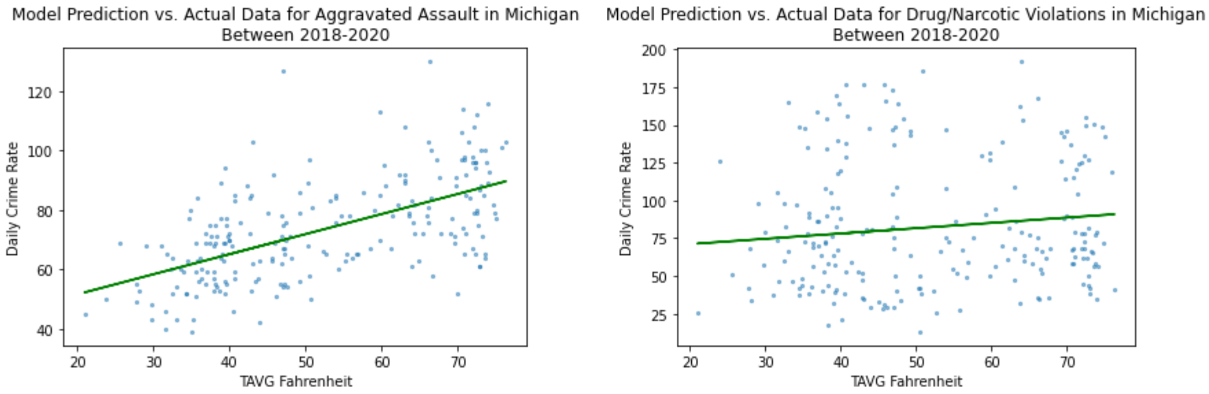

Figure 10: Models for Temperature vs. Violent and Non-violent Crimes

Notably, in the specific case that we explored, violent crime has a greater response to changes in temperature than non-violent crimes.

Summary and Analysis of Results

We summarize our RMSE values in Table 2.

Table 2: Summary of RMSE Values for Various Trained Models for Michigan Between 2018-2022, Test Set = 20%

Model	**Training RMSE**	**Test RMSE**
OLS - Temperature vs. # Arrests	139.12	147.09
OLS - Temperature, OHE Day of Year vs. # Arrests	114.93	147.09
Ridge - Temperature, OHE Day of Year vs. # Arrests	127.96	152.17
OLS - Temperature, OHE Month of Year vs. # Arrests	136.18	144.14

Notably, adding one-hot encoded features (such as day of the year or month of the year) does not decrease our test RMSE. This is likely because of *collinearity*. Because temperature data is correlated with the month or day of the year, temperature data already ‘includes’ the month or day of the year. That is, hotter temperatures typically occur during the summer, and colder temperatures typically occur during the winter. Therefore, adding such features has little to no effect on our models, and strictly using absolute temperature is sufficient as features for our model.

Discussion

In this project, we analyzed the relationship between weather (especially the air temperature) and crime in several parts of the US using the GHCN Weather Dataset and the NRBIS crime data from 2018 to 2020.

Our results indicated that there can be some positive relationship between the daily temperature and the daily crimes. As the results suggested, overall, from 2018 to 2020, we found a positive correlation between daily temperature and the number of crimes each day. At the same time, we also conducted some state-level analysis based on these data. For instance, we found that there is a positive relationship between daily temperature and number of crimes in Michigan from 2018 to 2020. Similar results were found in Texas (not shown in the report). We also tested the use of Ridge regression models and one-hot encoding of the month to capture the relationships.

We were struck by our inability to effectively engineer more features. For example, we attempted to develop a metric that described temperature deviation "away from the average." Initially, we applied a Z-Score (standardizing the data with mean zero and standard deviation one) for a metric of 'x standard deviations' away from the average, however, this yielded no linear relationship between Z-Score and Number of Arrests (Figure 11).

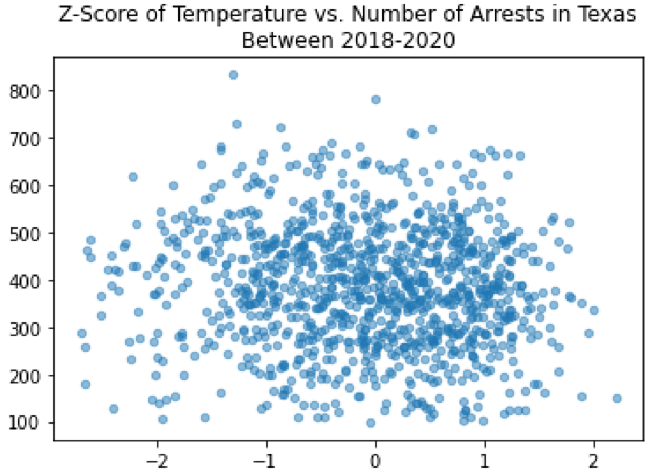

Figure 11: Z-Score of Temperature vs Number of Arrests for Texas Between 2018-2020

Additionally, we tried to create a slightly different metric that we called Temperature Ratio. This metric was the daily temperature divided by the average temperature for that day of the year, however; this feature also yielded no relationship between crime and temperature. We therefore discarded this metric and decided to use absolute temperature as our model feature. If there was a relationship between Z-Score and/or Temperature Ratio, we would likely be able to use the Day/Month of the Year as additional features, as there would likely be less collinearity.

We hypothesize that our lack of ability to engineer more useful features in our model is a result of the fact that we were only able to process about 1 decade of temperature data, which may not have been sufficient to estimate averages by which to calculate deviations.

Overall, our results overall agreed with the previous findings that related crimes to temperature, for instance, Anderson et al. 1997. And it also pointed to the necessity of a consideration of the months of the crimes into the models as the number of crimes is clearly related to the corresponding day of year.

Limitations of Methods Used

We recognize that our method of aggregating data may not be the most accurate. While grouping together multiple years of data, we typically averaged our data of interest, such as temperature across a state and temperature across years. We recognize that there is room for improvement, but given the scope and time frame of the Data 200 course, we opted to use these methods. Additionally, we recognize that our data is (by nature) time series data. Models exist for time series data that may be more accurate, but are not applied here due to the scope of the Data 200 course. Further investigation should be done to apply these models.

Conclusion

Overall, we were able to analyze both the FBI NIBRS Crime Dataset and NOAA Daily Weather Dataset. We explored both sets individually, and then were able to merge the two datasets together for analysis. We determined that, for our datasets, using absolute temperature was the best predictor for determining daily crime rates. Although we initially believed that some metric analogous to a Z-Score would be a better predictor, this was not the case. We suspect that this was an artifact of our data aggregation methods. Additionally, we found that, in specific cases, violent crime may have a more positive relationship with Temperature, while non-violent crime may not. However, further work in each of these topics should be explored for a stronger conclusion of relationships.

As temperature may be correlated to crime rates, this can be extremely beneficial to police forces and security. That is, if a temperature is higher for a particular day, on average, those groups may expect to see more crime.

Future Work

Further work should be completed to investigate if absolute temperature is an accurate measure for determining a relationship between climates and number of arrests. Due to limitations of our computational power, we had difficulty processing more datasets (additional years of NIBRS crime datasets and NOAA weather data), and therefore, further investigations should be made in processing additional data for data accuracy when averaging.

References

Anderson, C.A., Bushman, B.J. and Groom, R.W., 1997. Hot years and serious and deadly assault: empirical tests of the heat hypothesis. Journal of personality and social psychology, 73(6), p.1213.

Anderson, Craig A., and Kathryn B. Anderson. "Violent crime rate studies in philosophical context: A destructive testing approach to heat and southern culture of violence effects." Journal of personality and social psychology 70.4 (1996): 740.

Cohen, L.E. and Felson, M., 2010. Social change and crime rate trends: A routine activity approach (1979). In Classics in environmental criminology (pp. 203-232). Routledge.

Hsiang, S.M., Burke, M. and Miguel, E., 2013. Quantifying the influence of climate on human conflict. Science, 341(6151), p.1235367.

Kelley, C.P. et al. (2015) 'Climate change in the Fertile Crescent and implications of the recent Syrian drought', Proceedings of the National Academy of Sciences, 112(11), pp. 3241–3246.

Meyer, R. (2018) 'Does Climate Change Cause More War?', The Atlantic, 13 February.

Ranson, M., 2014. Crime, weather, and climate change. Journal of environmental economics and management, 67(3), pp.274-302.

Selby, J. et al. (2017) 'Climate change and the Syrian civil war revisited', Political Geography, 60, pp. 232–244.