LAW & WEATHER

SPECIAL DATA UNIT

IST 718 Final Project

Juri Boyar
Audrey Crockett
Kelly Hwang
Michelle Mak
Brandon Niskala

Introduction

According to an article by NPR in December 2018, there is a growing shortage of recruits in the United States police force. To combat this issue, police departments across the country are experimenting with AI technology to help dispatch officers more efficiently. In the same vein, this is an investigation into the possibility of using weather forecasts to help predict crime. Using San Diego, CA, a city which has consistently warm weather, as the control city and New York, NY as the variant, this analysis will explore whether any patterns can be found in the amount of crime or the type of crime that occurs during various weather scenarios throughout the year.

Data Acquisition

The team acquired climate and weather data for San Diego and New York City from the National Oceanic and Atmospheric Administration (NOAA) via the agency's REST API. The data set included daily summaries for the years 2000 through 2018. Because NOAA's API only allowed a maximum query return of 1000 data points, which equated to approximately nine months, a loop was created to accommodate the multiple requests.

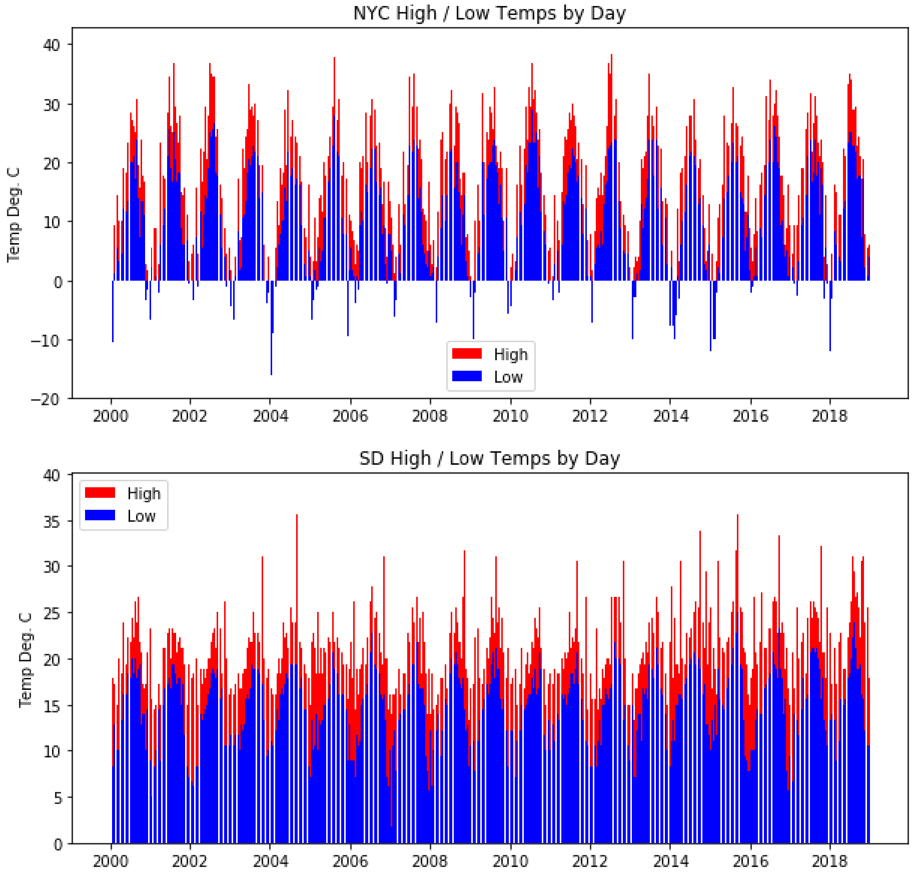


Figure 1: Low and high temperatures for New York City and San Diego. San Diego experienced higher and more consistent temperatures then New York.

To compare with the climatic data, the team acquired datasets for crime in both cities from each city’s respective government data hub. Each dataset documented the time of day each crime occured, a categorical description of the crime, as well as the longitude and latitude of where the crime took place. The San Diego dataset was limited to the years 2007 through 2013, so the New York comparisons for weather and crime were shortened to this timeline for consistency.

Methods

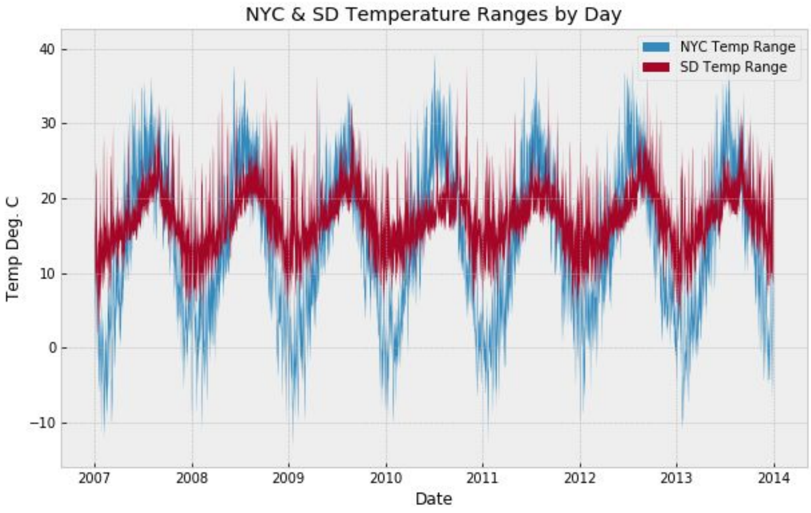
Preprocessing & Cleaning

The data from NOAA is downloaded in JSON format and converted into dictionaries using Python’s JSON library. New York data was missing only one temperature value while San Diego data had a large amount. A loop is used to find and replace all missing values with a null. The values of each dataset were converted to whole numbers and saved out to a CSV to use for analysis.

Crime data is read into the notebook as a CSV file. The data columns of each dataset are stripped of whitespace and uniform column names were added. San Diego had less crime data from the years 2008 - 2013, while New York crime went as far back as the 1940’s to present. The New York City crime data was sampled to apply to the same date range as the San Diego dataset.

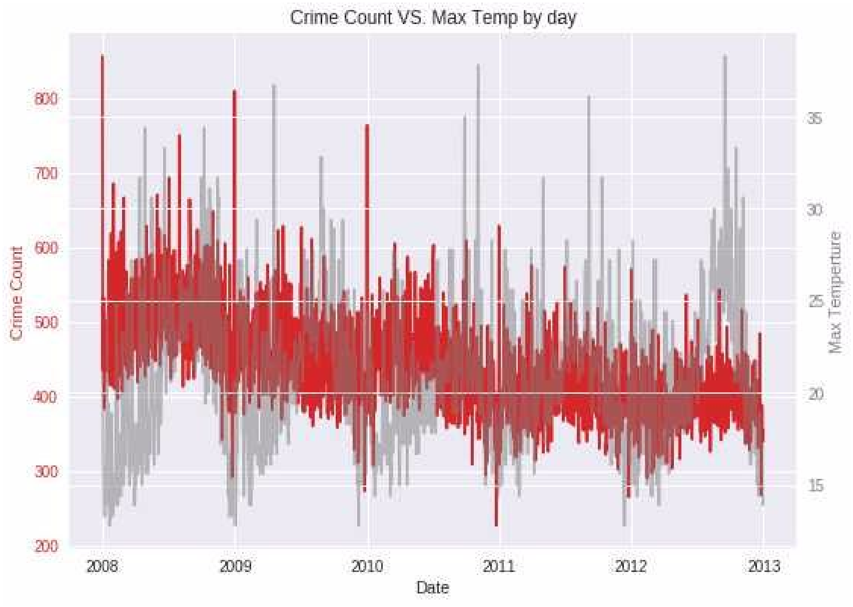
Exploratory Analysis

Before beginning the time series modeling, which would demonstrate whether weather patterns could help predict crime, a few preliminary questions are explored in the initial analysis. First, the seasonal differences between the two cities is established. Figure 1 below displays that while the seasons follow the same patterns in San Diego and New York, the temperatures in San Diego are clearly less divergent than that of New York’s. Since San Diego does, in fact, demonstrate more stable weather, it can be used as a

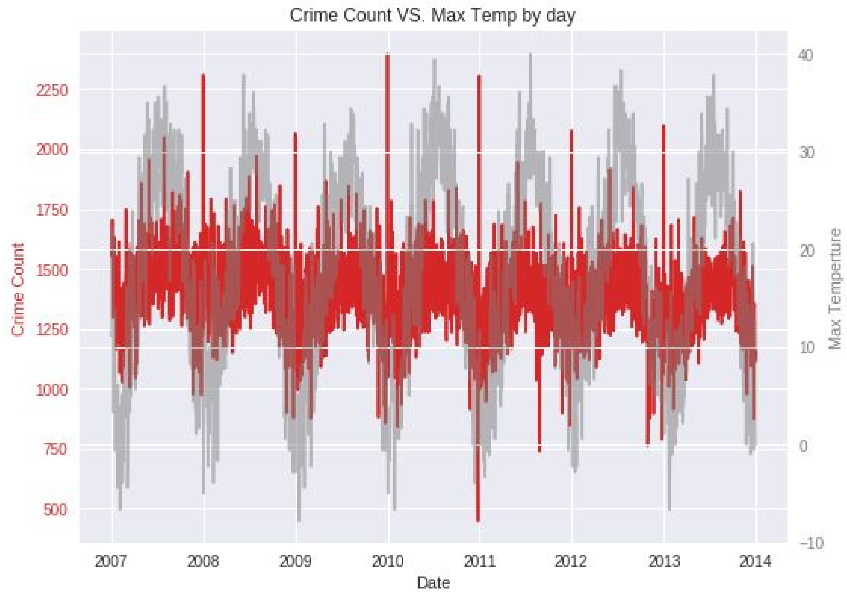


“control” of sorts in determining the possible relationship that exists between weather and crime. Now that seasonality is established, we look at whether there are any patterns in when the crimes occur. The following charts show the weather of the two cities (in gray) overlaid with the number of crimes committed (in red) throughout the time sample.

San Diego

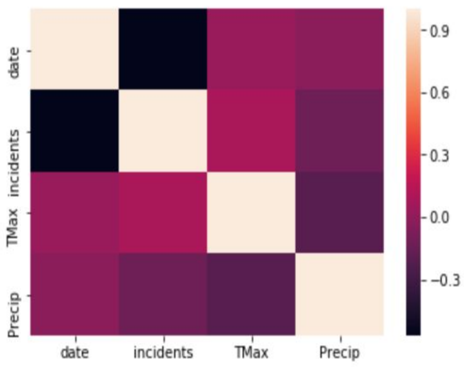


New York

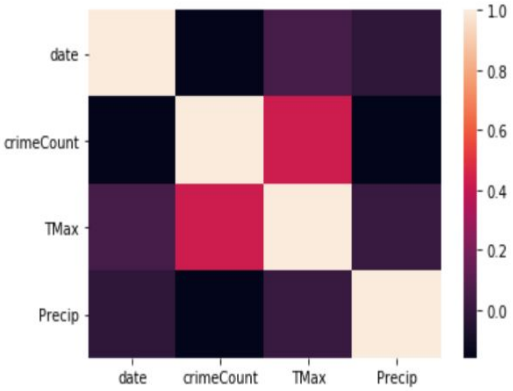


While the patterns are not exact, crime does seem to follow some sort of seasonality year-over-year. To further investigate the correlation between weather and crime, we explore the heatmaps created for the two cities.

SAN DIEGO Crime-Climate Correlation

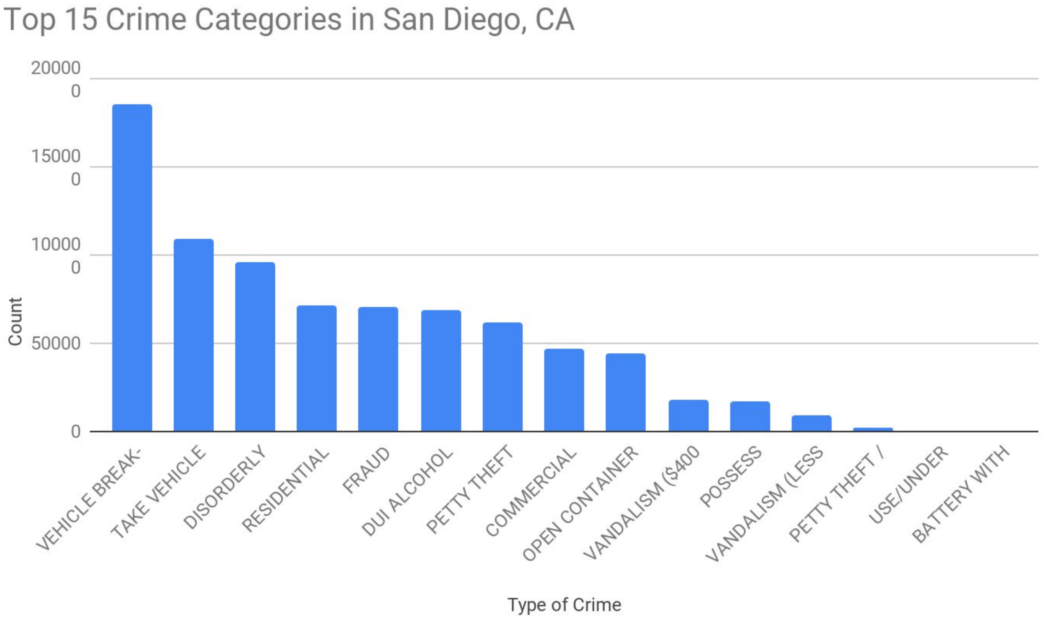


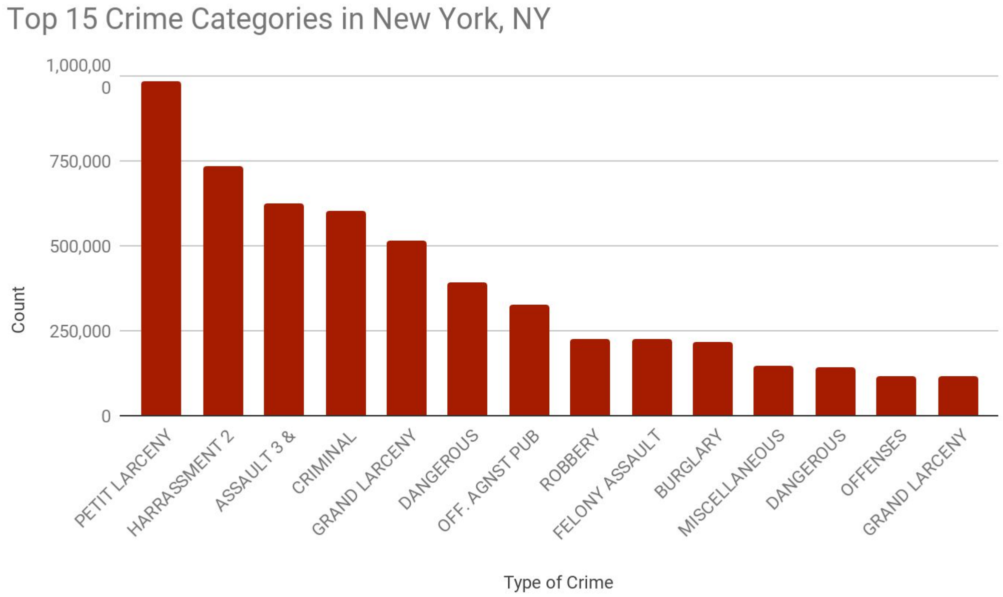
NEW YORK Crime-Climate Correlation



At first glance, it is clear that the correlation of temperature, precipitation, and incidents between the two cities have some differences. Using San Diego as the baseline, the variables of high temperatures (TMax) and incidents have a slight positive correlation while precipitation and incidents have a definite negative correlation. On the other hand, New York's high temperatures and incidents (crimeCount) have a slightly stronger positive correlation than that of San Diego's and a noticeably stronger negative correlation between precipitation and incidents. While the patterns in both cities reflect the same trend, it is clear that weather in New York, which has the more temperamental, has stronger correlations with crime incidents.

A third question explored is 'What types of crime are the most popular in the two cities?' The top 15 types of crime are plotted in the bar graphs below to answer this question.





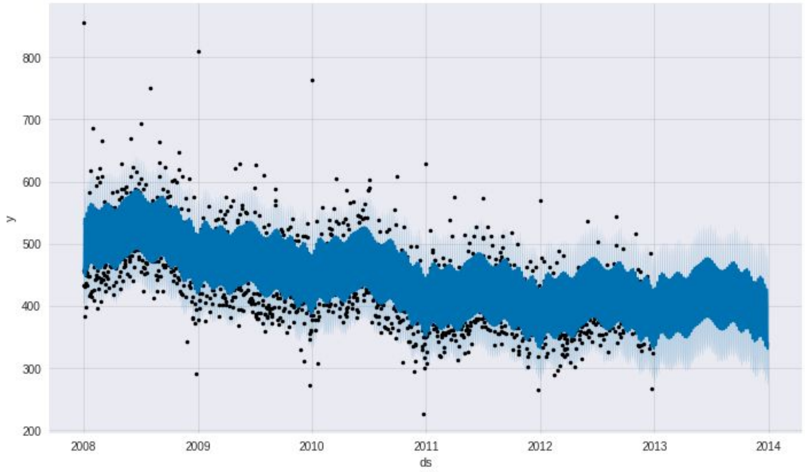
It looks as if San Diego and New York have two very different crimescapes. While San Diego seems to have more vehicular and residential crime, New York crime is a bit more serious. In addition, the amount of crime committed in New York is significantly greater than that committed in San Diego. This is largely due to population differences. This initial investigation will not control for population differences, but future experiments should certainly take note of this important distinction.

Modeling

Time series models are created using the FBprophet package, using dates as the ‘ds’ column and the number of incidents as the ‘y’ column for both data sets. Two separate lists are created and combined to create a ‘holidays’ dictionary with which the impact of weather will be measured. All days with temperatures hotter than 25℃ and precipitation greater than 10 cm of rain are considered ‘holidays.’

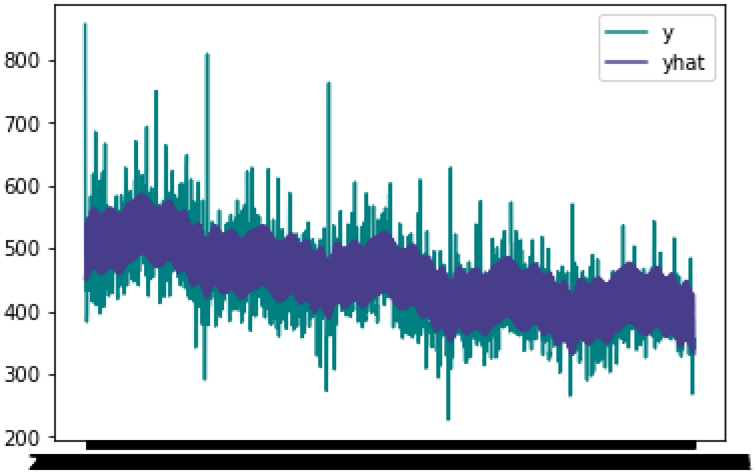
The time series model created by FBprophet for San Diego reflects a slight downward trend of crime activity from the years 2008 to 2014.

San Diego Crime Time Series Model



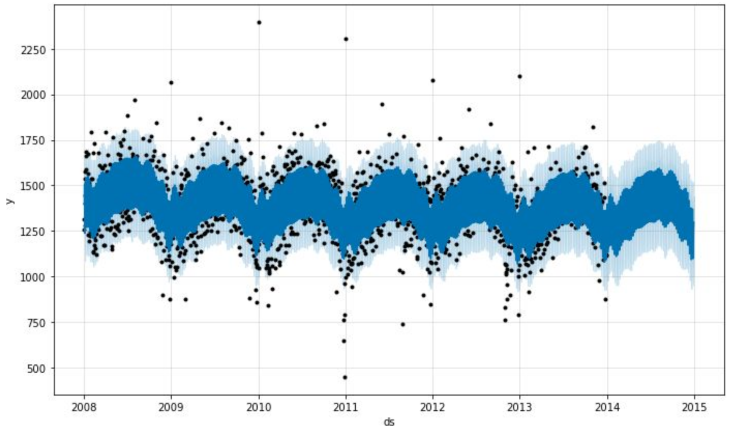
The root mean squared error is 40.35, which is calculated using the ‘mean_squared_error’ function from the sklearn.metrics package. By dividing the average crime count, which is 436.78 for San Diego, we get the error rate of 10.9%.

San Diego RMSE Plot

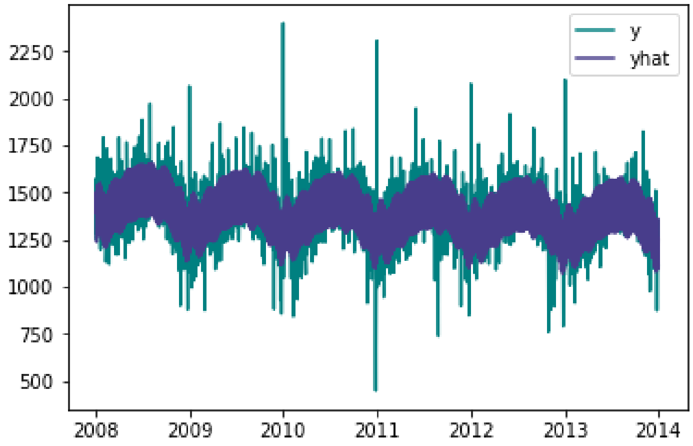


The time series model for New York City shows a more stable amount of crime activity over the years, with an RMSE of 120.37. The average crime count is 1390.20, which results with an error rate of 11.5%.

New York Crime Time Series Model

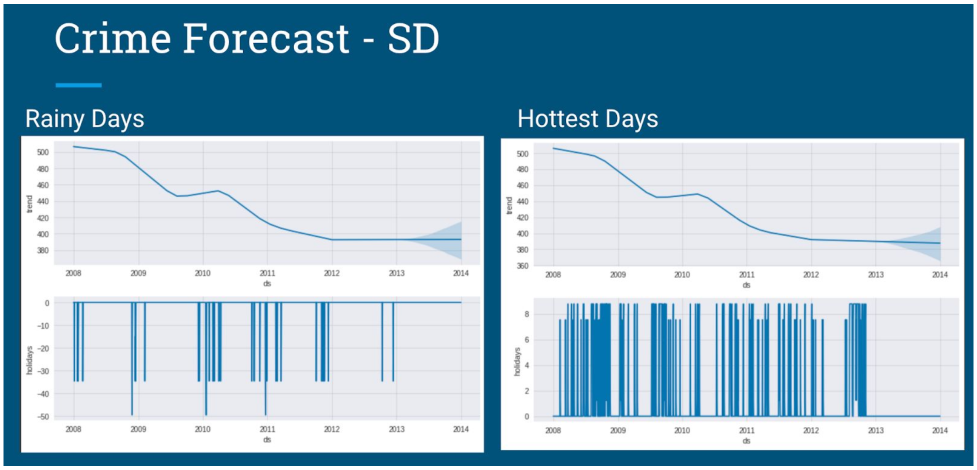


New York RMSE Plot



The error rate of both models is satisfactory as they are both above 85% accuracy. However, using weather 'holidays' to help predict the crime did not seem to be beneficial. The breakdown of the 'holiday' results is achieved through the 'plot_components' function that is built into prophet.

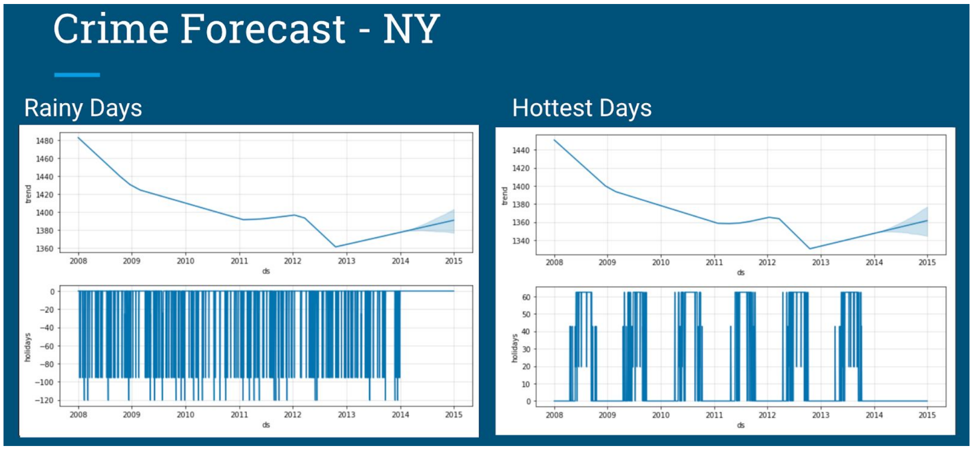
The two visuals below illustrate that rainy days do, in fact, have some sort of negative relationship with crime while high temperatures have a positive relationship with crime. This is in line with the findings in the exploratory analysis. However, the seasonality of the visuals below do not match the seasonal patterns found in the time series, which means that these two variables are not the best for predicting crime.



Crime Forecast - SD

Rainy Days

Hottest Days



Crime Forecast - NY

Rainy Days

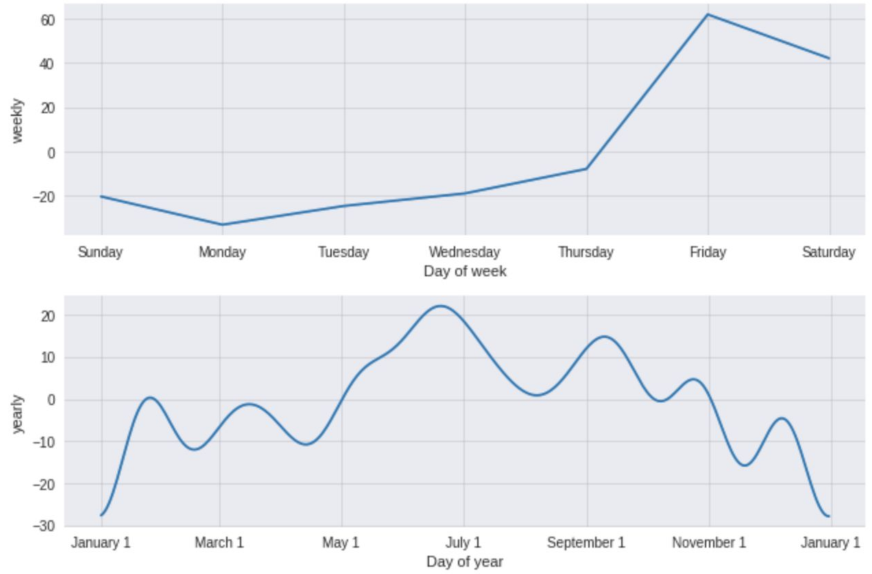
Hottest Days

Results

Despite having seasonality and slight correlation in crime and weather, these rain and temperature do not seem like the best variables to use in predicting crime. However, this analysis did demonstrate that patterns in criminal activity do exist in a way that can help better staff police departments.

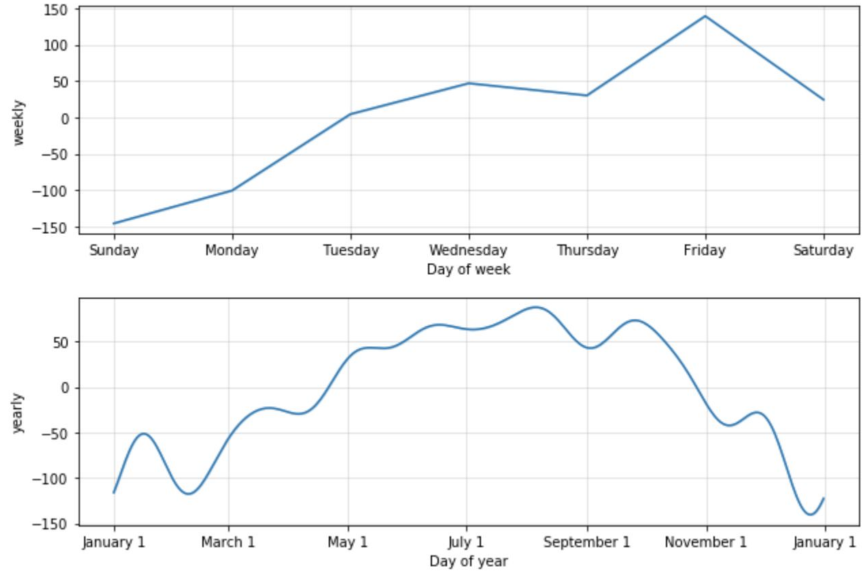
For example, the components breakdown in FBprophet shows that San Diego historically has lower crime on Mondays and higher crimes Friday through Saturday. In addition, summer months like June and July seem to have a higher number of incidents.

San Diego Weekly and Yearly Crime Patterns

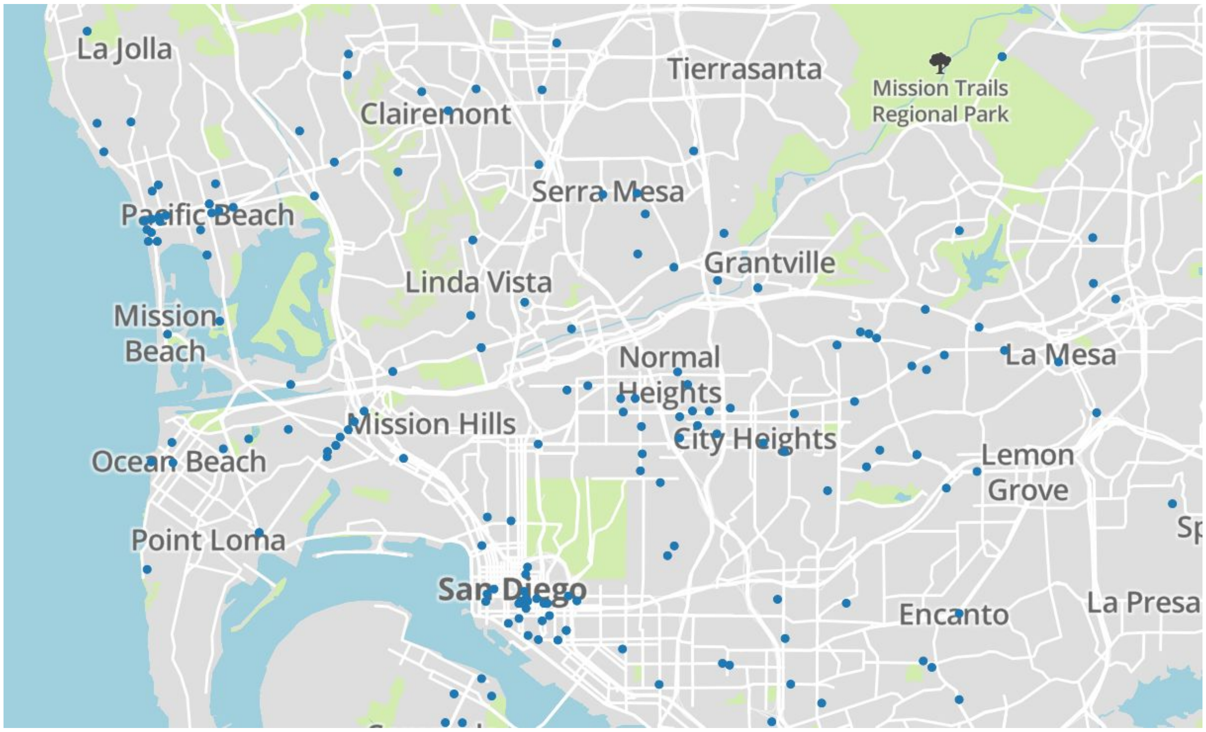


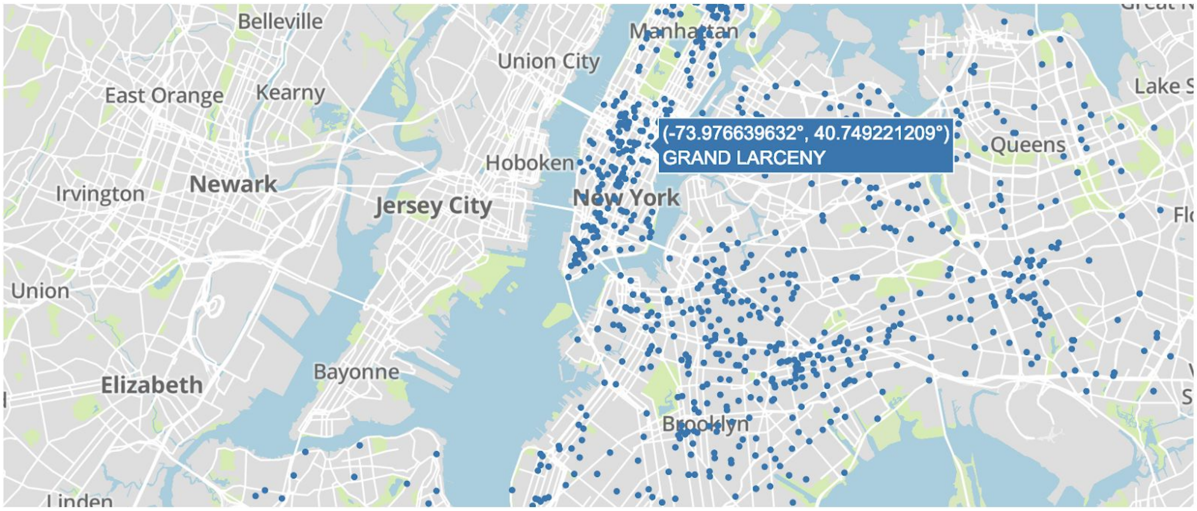
Likewise, New York City records show that criminal activity steadily rises throughout the week, peaking on Fridays. And summer crimes in New York are a bit longer lived, with higher rates June through October.

New York CIty Weekly and Yearly Crime Patterns



Since crime tends to be higher on hot days, a plot of the criminal activity on the two hottest days within the sample time frames are created to see where crimes take place. These interactive maps can be found for San Diego and New York. The maps of these crimes show that densely populated areas have more criminal activity.





Recommendations & Conclusion

The results of this investigation are preliminary and can be improved. Controlling for population, for example, is a huge step in leveling out the differences between San Diego and New York City -- especially because more crimes happen in densely populated areas. Another insightful addition could be to create a severity scale of crime. This analysis has no differentiation between violent crime and misdemeanor. Since only the count of criminal incidents is taken into account, a high number of incidents could skew the need for police support if the majority of those reports are minor infractions. In addition, there is a large difference between the number of crimes reported and the number of crimes that actually take place.

The overall finding of this investigation is that this data can, indeed, be useful to help staff police departments more efficiently when faced with recruitment and budget shortage. For example, staffing the departments less heavily during the beginning of the week, when criminal activity seems to be lower in both San Diego and New York, might open up some budget to increase police presence on Fridays and Saturdays. Additionally, saving up some budget for summer months, when criminal activity is at its peak would be effective as well. This might be accomplished by offering voluntary time off for officers when heavy rain is predicted, as evidence shows that less crime occurs during this type of climate.

Furthermore, educational campaigns can be used to remind the public to lock the doors and windows when leaving for the beach or vacation in the summer, since high temperatures correlate with higher crime rates. Campaigns like this are preventative and can perhaps lessen the potential for thefts.