DSC423 – Data Analysis and Regression

Milestone 05 – Project Proposal

By:

Group Dogwood Rose Doppelgangers

Jesse Franks, Pete Gillespie, Reid Case, Robert Cordrey, Salik Hussaini, M. Omar Khan

Crime is a complex and widespread phenomenon that has a constant negative effect on communities throughout history. While the definition of crime shifts throughout time, and across borders, the effect remains - a reduction in quality of life. Tactics for addressing criminal activity can be controversial, expensive, and occasionally seem ineffective. In the US, it is typical for law enforcement to attend the “scene of a crime” after the fact. In this scenario, the damage has already been done. Proactive policing is difficult and often fraught with unfair, and illegal profiling and bias. Using unbiased data and methods to predict crime, or areas, times and seasons when crimes are more likely to occur could assist in allocating law enforcement resources, and may give insight to other social and physical factors that lead to criminal behaviors. We have chosen Boston, Massachuesetts as a source of crime data and paired that with historic weather data to perform analyses that may uncover relationships within this area of concern. Boston is a large metropolitan area with a long and vibrant history tied to the formation of the United States. It has a strong economy and diverse community. FIndings may have a strong impact on quality of life and provide insight that can be spread to other communities.

According to boston.gov, from 2015 to 2019, 393,185 crimes were reported within the Boston Metropolitan Area. Out of this total figure, 49,330 crimes were violent crimes. Based on this data, violent crimes therefore accounted for 12% of all crimes. Annually, for the year of 2017, 93,278 crimes had been reported, and 91,423 crimes were subsequently reported in 2018. Currently, from January 1 through September 25, 2019, 65,519 crimes have been reported. For the year 2019, $400,000,000 was budgeted for the Boston Metropolitan Police Department. As can be seen, the fiscal impact of criminal activity is anything but negligible, and as such, further analysis to better deploy the city’s resources is needed.

Using the data available from boston.gov the purpose of this analysis will be to better predict criminal activity with the goal of providing more efficient and efficacious use of the budget. The analysis will be used specifically to understand what variables affect the incidence and type of criminal activity within the city limits. We will predict whether these variables are effective in making predictions on when and where criminal activity is likely to occur, thereby allowing police departments to more effectively use their resources in combating criminal activity. Furthermore, we will endeavor to make the analysis relatively simple to enhance the general applicability of the model to other regions.

As noted earlier, to help explore criminal activity in the city of Boston, we will use datasets primarily from the Analyze Boston government website available on boston.gov. The crime report dataset is a collection of reports provided by the Boston Police Department ("BPD") documenting the initial details surrounding an incident to which BPD officers respond. This is from a nascent crime incident reporting system, which includes a reduced set of fields focused on capturing the type of incident, as well as where and when it occurred. Other data found from the Analyze Boston website include size and population of the 12 Boston police districts, as well as the GPS coordinates of the police stations. We then calculated the distance from each crime to the responding police station. Lastly, we used a weather dataset from the CRAN package called stationaRy. This dataset recorded the hourly temperature at Logan International Airport. We calculated the average daily temperature and then joined it into our main dataset because, as we will show, we believe daily temperature may be a good variable in determining criminal activity.

Using our Boston crime data frame we hope to answer questions like: do certain crimes occur more readily in certain areas, and if so, what factors are significant in determining whether a specific type of crime will occur? Does temperature affect crime? How does the distance from a police station affect the crime rate. Does the population or day of the week affect the type and location of the crime? Does crime vary with any predictability over some temporal scale. We think answering these questions will help us understand where crime happens and make recommendations to police about where they should concentrate their force. Ultimately, our goal is to be able to provide a thorough analysis that points to a cheaper and more efficient solution when it comes to combating criminal behavior. We intend to achieve this by looking at a myriad of different explanatory variables, and building regression models to determine if any indicators exist when predicting criminal behavior.

Average Daily Temperature vs Incidence of Crime

Based on our observations of the data thus far, we would like to see what relationship exists between the average daily temperature and the incidence of crimes. It is perhaps reasonable to deduce that the criminal activity is limited to when the weather allows for it. That is when the weather is most comfortable for such activity to occur. Generally speaking, it is assumed that on colder days criminal activity, especially that activity requiring an outdoor element, will be significantly reduced. Likewise, on warmer days, we anticipate that criminal activity will be greatly increased. However, it is certainly possible that this is a foregone conclusion, not to mention that certainly there is an upper limit to how warm the climate can be until we see a drop off due to unbearable heat. We will work on creating a model that is designed to specifically identify the relationship on weather and the incidence of criminal activity. This model will project the occurrence of crime given the average daily temperature, and determine if this variable is an adequate predictor of criminal activity. Once the data has been analyzed, it can be leveraged to ensure that the resources of the BPD are being appropriately deployed during periods where the likelihood of criminal activity is highest. This analysis will be conducted by M. Omar Khan.

Distance from Station vs Time-of-Day, Day-of-Week, Population.

The variable that represents the distance from the nearest police station should, intuitively, have some relationship with crime rates. Considering the patterns identified in the previous exploratory analysis, there may be some effect from "seasonality" present in those rates as well. Potential strain on law enforcement resources may be implicitly derived from this analysis when paired with population densities. We will work to build a model that projects the incidence of crimes in terms of "Distance-to-Station". We will also compare this with separate variables such as population, the time of day a crime occurred, and the day-of-week a crime occurred to see if it may be possible to identify potential shortage or surplus in law enforcement based on the station coverage as populations shift, and as the day and week progresses. This will allow for a better understanding of areas that may have lower law enforcement coverage. By doing so, we will be able to identify gaps in geographical coverage and provide greater insight on where resources should be deployed. This analysis will be conducted by Reid Case.

Date vs Crime

We will be reviewing and analyzing the relationship between date and crime. Specifically we will review to see if there is a relationship between crime occurrence and a specific day of the week, or in a given month. Per the analysis from the data examined thus far, it is clear that the incidence of crime is localized to specific days of the week. Based on this insight, it will be hugely beneficial to design a model that indicates what impact the day of the week has on the incidence of criminal activity. This will go a long way in assisting the BPD in deciding on how to deploy resources such as policemen on the street on a given day, or in a given month. Based on the results of the regression model, if there exists a clear relationship, we will be able to indicate what impact that relationship has on predicting criminal behavior. This analysis will be done by Salik Hussaini.

Location of Drug Violations, Auto Thefts, Auto Recovery vs Median Income

We believe auto theft, stolen auto recovery, and drug violations are 'single point' crimes that are spatially clustered and frequently occur in the same areas. Based on this we can hypothesize that a multiple regression model can identify areas that these crimes are most likely to occur based on latitudinal and longitudinal data. We will use median income as an explanatory variable to see if the incident of crime is localized to those areas with lower median incomes. One such example would be that auto theft may occur more predominantly in areas of higher median income, but auto theft recovery would occur in areas of lower median income. That is to say that cars are stolen from high median income neighborhoods and recovered in lower income areas. Our hope is that this information will provide justification for a larger police presence, and a change in municipal services or infrastructure in those areas where criminal activity is more likely to occur. This analysis will be completed by Jesse Franks.

Crime vs Location

We will be reviewing the correlation between the occurrence of a crime and the location of the infraction. It is likely that pockets of crime will be found by location and may be extremely dependent on the type of offense committed. While it can be nearly impossible to predict when a crime will occur, it is possible to predict the likelihood of a specific type of crime occurring within a defined area. Within each Police Precinct, there are Police Substations that cater to a more specific, defined area. By having a better understanding of where crimes occur and the frequency of occurrence, the police will be better equipped to handle situations as they arise. If location is a good predictor of criminal activity, then police departments can more effectively target those areas directly, thereby reducing the total overall costs associated with combating crime. This analysis will be completed by Robert Cordrey.

Time of Day vs Type of Crime

Crime in Boston tends to occur most frequently during the week in the afternoon hours. We believe that certain categories of criminal activity may be more prevalent at different times of the day and occur more frequently in certain police districts as opposed to others. Therefore, we will split the crimes by the total number of districts and be able to compare the models against each other based on the nature of criminal activity, and the nature of the districts (i.e. size, population, etc.) We anticipate that certain criminal activity is localized to very specific times of day, and that they may be categorized based on violent or non-violent crime. Using this information, the BPD would be able to greatly benefit in ensuring the most adequate protection for its resources is available at those times of day that have a more prevalent instance of violent crime. This analysis will be completed by Pete Gillespie.

In summary, combating crime is one of the largest cost drivers for any metropolitan area. By being able to more accurately predict the criminal activity, when it is likely to happen, where it is likely to occur, and how frequently, we endeavor to provide a model that can allow cities fight crime in a more cost effective and efficient manner. There is perhaps no greater detriment to human collective progress than criminal activity. The goal of this model will be to compare changes in our response variables and see how much that change can be explained with our explanatory variables. Once we have come to a conclusion, we will hope that our model is able to predict criminal behavior more accurately, or at least provide a list of factors that froment an environment conducive to criminal activity. Armed with this information, cities will be able to better combat crime in a more cost effective manner.