DSC423 – Data Analysis and Regression

Milestone 07 – Individual Milestone 1

By: Salik Hussaini

Criminal law stems from the fact of criminalization. To criminalize, or in other words, an act of violation of the law, ought not to be done. However, crimes continue to occur in Boston. Our team has begun to analyze the Boston crime data and would like to propose strategies on how to reduce them. We will begin to explore crimes based upon several factors and create a regression model to help reduce them from occurring. Variables that will help build the linear regression model will be the location, time of year, time of day, nature of the crime, affected city district, and local weather to ultimately predict what and when the crime will occur.

This report will focus specifically on the date and the crime that would occur. Date parameters such as month, day, year, hour, day of the week will be addressed as the independent variable of this model. The dependent variable will be the distance to the station from where the crime occurred. This will tell the users how far a crime will likely occur based upon the date parameters that will be given. Ultimately the regression model that will be created will help the Boston police department in predicting crimes in the future.

From the original data set that was worked with the team, which included 18 variables; the dataset that was included in this report includes 20 variables. The additional 2 variables that were manipulated were the addition of month and year which stemmed from the original date variable. This ultimately begins this report with a building of a model with just the date independent variables. The dependent variable again will be the distance from the station where the crime occurred. The variable selection was made to the fact that this report is selectively focused on the date parameters. To begin, model one was made with independent variables day of the week, hour, month, date, and year, area, and population. These independent variables were used to predict the dependent variable of distance to the station where the crime occurred. The variables that were removed in the model were the reporting area if a shooting was involved, the average daily temperature, the latitude of the crime, the longitude of the crime, station latitude, station longitude, and recoded type of crime. These variables were removed because they didn't apply to the overall purpose of the report which was the date.

After building the first linear model with the proposed independent variables that consisted of the date parameters, one can say the model performed poorly. Looking at the adjusted R squared of the model, it was 5%. This means 5% of the variability of the distance of the station where the crime occurred is explained by our model. Although our model had a low r squared, the mean squared error was quite good, it was 0.3948. With a low MSE, this shows our model is minimizing the variance and shows how the model fits the data without significantly harming the model's predictive ability. Looking at the F value we can say that one of our variables beta's is not equal to zero, meaning we could reject the null hypothesis that they are all equal to zero. Each of the variables has a significant individual T-test except the variables hour and date. In the next model, I will be removing these variables to see if the model's adjusted r squared will be increased.

In the next model I create I will be using the variables day of the week, month, year, area, and population, average daily temperature, 2nd order term for population (population * population), 2nd order term for area (area * area), interaction between population and area, and also interaction between temperature and area. The dependent variable again will be the distance to the station. In this model the adjusted r-squared is 7.5%, this means 7.5% variability of the distance of the station where the crime occurred is explained by my model. The mean squared error for this model was 0.3849. This is also a quite low MSE. The F value significance level is also below .05 which means one of our variables beta's is not equal to zero, meaning we could reject the null hypothesis that they are all equal to zero. Finally looking at the individual T-test for all our independent variables, all but one showed a significant level of below .000 except one which was a significant value of .01. This explains that the beta's for these individual variables is not equal to zero since the p-value is so low. This idea will be useful when we are analyzing our data.

Looking at the two models that were created, the second model seems a bit more significant than the first model. The use of interaction terms and second-order variables helped increase the model adjusted r squared. Nonetheless, looking at the overall betas for the second model, we can see that 6 out of 10 of the variables have a negative sign in front of them. Meaning an increase in, for example, the month and year will cause a decrease in the distance to the station from where the crime occurred. Also an increase in population, the daily temperature will decrease the distance to the station. An increase in a year, of a possible crime date, would cause the distance from the station to be reduced by .005, which is in miles. Ultimately, more vigilance should be required. A possible suggestion for a solution to this issue is installing cameras utilizing computer vision in the vicinities within police stations. This may be a possible solution to decrease overall crimes by analyzing relative human behavior and motion concerning possible crime.

In the next milestone, I would love to increase my adjusted R squared for my model. Possible solutions to increase my adjusted R squared will be looking at the residual plots and looking if I could add any more interaction terms in my model or an additional second-degree variable. Another possible solution will be looking at to see if multicollinearity is happening in my independent variables. I will make an analysis of the Variance inflation factor for my independent variables. However, the concept of time and crime is hard. No one can perfectly predict the future, however, I hope to see small incremental improvements in my model to help prevent future crime.