Final Report

Group 11: Caroline Simmons, Jeannette Jiang, Jenny Jang, Sejal Patrikar

Objectives

Our primary goal is to find the best way to convey Baltimore's crime data in a way that is helpful to the Baltimore Police Department. Our secondary goals were:

- To predict what type of crime occured (violent or nonviolent)
- To predict how many crimes will occur in a given day
- To create visualizations that present interesting, informative findings

Motivation

Predicting and analyzing crime is a field that has been around for a long time, and for good reason. Finding the motivations and situations that lead to a crime occuring could lead to a safer, more secure world. The "routine activity approach" is a theory which emphasizes the circumstances in which a crime occurs, rather than the traits of the perpetrator, to analyze crime trends (Cohen & Felson, 1979). There is evidence that property crimes are driven by pleasant weather, which is consistent with this approach (Hipp & Curran, 2004). Property crimes are nonviolent crimes which involve the theft or destruction of someone else's property.

There is also empirical evidence that crime rates increase in hotter years, and that crime is more prevalent in hotter parts of the year. Furthermore, the effect of temperature is stronger on violent crimes than it is on nonviolent crimes (Anderson, 1987). Conversely, there is a strong positive effect of unemployment on property crimes, while the evidence is much weaker for violent crimes (Raphael & Ebmer, 2001).

For this reason crimes in the dataset were divided into two broad categories: violent and nonviolent. By predicting which category a single crime falls into, this provides insight into what factors are the most influential on that type of crime occurring.

Predicting the amount of daily crime can be useful for many reasons, the most important of which is staffing. By taking into account the weather, current unemployment rate, or last week's numbers, the Baltimore Police Department might be able to use their officer's time more effectively by staffing only the officers they need. It may also reveal which factors are the most important.

By creating interesting data visualizations it is easier to convey complicated trends and patterns in a straightforward fashion. The apprehensibility of the graphics is especially true considering the primary audience for these findings, the BPD.

Data Cleaning

After the Baltimore PD data set was loaded in, there were a few adjustments made. First, we split the variable CrimeDate, into its respective month, day, and year. Then, we decided to work with only the years 2015-2019. The raw dataset included data from the 1900's to 2020's, but through initial glance, the only documented offense up until 2015 was rape. This could be due to the fact that the site from which we retrieved the data did not initiate until 2015. Additionally, we cut off the dataset at the year 2019 to include the full cycle of crime, and not imbalance it since the number of crimes committed varies greatly from month to month.

Then, weather data from the National Center for Environmental Information was added. This included daily average temperature, precipitation, snow, and snow depth. Next Baltimore's monthly unemployment rates from the Maryland Department of Labor was added.

Next, the original dataset included multiple types of the same offenses (such as common assault and aggravated assault), which we decided to consolidate into a single offense. This was done for larceny, assault, robbery, and larceny.

Furthermore, we coded in "1" for violent crimes and "0" for nonviolent crimes. To do this, we used the Description variable and determined a Violent crime as any offense with intent to harm and a Nonviolent crime as a property crime. Violent Crimes include shootings, homicides, assaults, and rapes. Nonviolent crimes included robbery, larceny, burglary, and arson.

Utilizing the variable, CrimeTime, we created a separate Hour variable, which allowed us to create a part of day for each offense.

Logistic Regression

Overview

We chose to use logistic regression to try to determine what type of crime (violent or non-violent) would occur based on our predictors. It is most helpful in understanding which of our predictors have the most influence in a violent or non-violent crime occurring. This information would be important to know so that Baltimore PD would know where to allocate its resources. It would also help with staffing placement in order to allow for a more secure and safe city.

Assumptions

In order to conduct logistic regression, there are a set of assumptions that must be met by our data. The first step was to check whether or not these assumptions were met.

- The assumption of independence is met because every entry corresponds to a separate crime committed by an individual.

- The assumption that the response variable is binary was also met. In order to meet this assumption we grouped different types of crimes into violent and non-violent using our own discretion, so that our response variable (type of crime) was binary.
- The assumption of linearity for the explanatory variables was also met. In order to check for this, scatter plots for each of the explanatory variables against the logit model were created. They indicated that every predictor was linear, except for the predictor Hour. See the scatter plots for the non-linear variable Hour and the linear variable Month in Figure A below for comparison. The red line depicts the linear regression line if the predictor follows a linear trend and the blue line follows the predictor as is. If the blue and red lines are relatively lined up, we considered the predictor to meet the assumption of linearity.

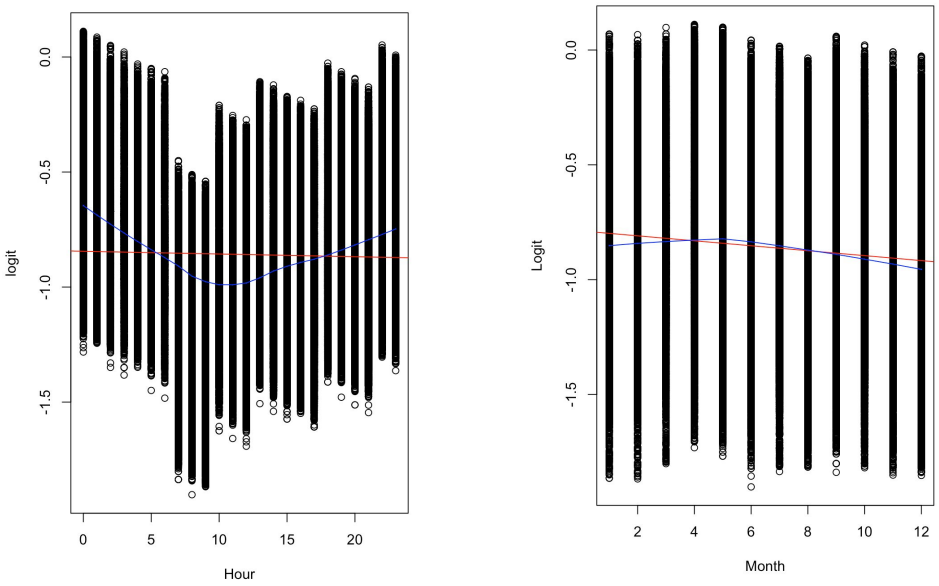

Figure A

- The assumption of no multicollinearity was also met by all predictors except for the predictor Hour with a VIF value above 5. All other predictor VIF values were around 1.
- Since the predictor, Hour, did not meet the assumption of linearity and lack of multicollinearity, we chose to remove it from the model.

Logistic Regression Model 1

Building the Model

- We initially formed a logistic regression model including all of our predictors except for Hour to determine which were insignificant. The output showed that all were significant except for the predictors Snow and Snow Depth.
- We then split the data 80:20 into training and testing data. We ran logistic regression to form a model based on the training data set, not including Hour, Snow, and Snow Depth.

Then we used the model to predict values based on our testing data. Our cutoff probability for a positive value or a Violent Crime occuring, was 50%. Using the probabilities output, we decided that if a probability was greater than or equal to 50%, we would consider that as a "1" (Violent Crime Occured) and a probability less than 50% would be considered a "0" (Non-Violent Crime Occured).

Results

- We then used the testing data to determine how accurate our model was. The confusion matrix output showed an accuracy rate of 70%, which is relatively good, but could definitely be improved.
- The sensitivity rate, the rate that the model detects true positives, was extremely low at around 0.1%, meaning our Type II error rate, predicting a violent crime will not occur, when in fact it did occur, is extremely high at 99.9%. Type II error rate is calculated by subtracting sensitivity from 1. (Newberg, 2005) This extremely high value is not ideal, especially when trying to predict the occurence of violent crime.
- The specificity rate, the rate that the model detects true negatives, was high at around 99.93%, meaning our Type I error rate, predicting a violent crime will occur, when in fact it did not, is low at 0.07%. Type I error rate is calculated by subtracting specificity from 1. (Newberg, 2005)
- We then formed an ROC curve with an AUC of 0.58, which isn't ideal but still indicates the model does better than predicting at random. This curve is depicted below in Figure B.

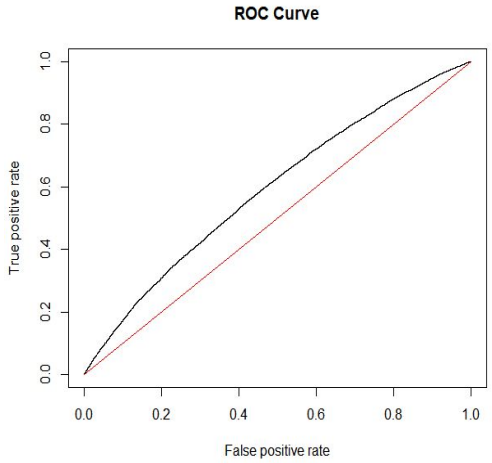

Figure B

- Overall Analysis of Model:
    - Based on the results outputted through the confusion matrix, we wanted to increase the sensitivity rate, which would then decrease the type II error rate. We want to attempt to decrease this value because we believe it will be more harmful if the model inaccurately predicts no violent crime when in fact there is a violent

crime. In other words, we would prefer for a violent crime to be predicted and not occur (type I error) over a violent crime to not be predicted for and a violent crime to occur (type II error). Since the results indicate that our model would not be able to predict well enough, we wanted to see what we could potentially do to improve upon this current model.

Future Improvements

- Lowering our threshold of a positive output to increase sensitivity rate thus decreasing type II error since we want to be as accurate as possible when it comes to predicting if a violent crime will occur or not.
- Re-evaluating the inclusion of our current predictors through stepwise regression

Logistic Regression Model 2: Improving the Model

Building the Model

- We continued with our logistic regression model evaluation, but this time used an automated variable selection process to further tune the variables within our model. After running stepwise regression, the variable, Day, was removed on top of the variables removed from the first model.
- Following the process of our first model, we then split the data 80:20 into training and testing data and ran logistic regression on the training data.
- We then predicted values using testing data, which output probabilities that a violent crime would occur. To improve from our first model, we lowered the threshold of a positive result, or a "1", to be greater than or equal to 35% instead of 50% from the initial model to be more conservative.

Results

- To evaluate our model, we output a confusion matrix with an accuracy rate of 65%,
- The sensitivity rate, the rate that the model detects true positives, increased to around 31%, meaning our type II error rate, predicting a violent crime will not occur, when in fact it did occur, decreased to 69%.
- The specificity rate, the rate that the model detects true negatives, decreased to around 79% , meaning our type I error rate, predicting a violent crime will occur, when in fact it did not, increased to 20%.
- The ROC Curve shows that our model is not the most accurate, but it performs better than if someone were to randomly guess if a violent crime would occur or not. The AUC, or area under the curve, value is ~0.592, which means our model is not the best, but there is a slight improvement from the first model's AUC. The ROC curve is depicted below in Figure C below.

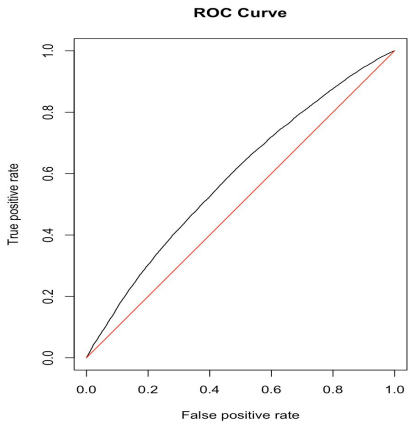

Figure C

- Overall Analysis of Model 2:
    - Based on the results, decreasing the probability threshold resulted in a more balanced model in terms of sensitivity and specificity rates. By increasing sensitivity and decreasing specificity, the type II error rate decreased from 99.9% to 69%. Type I error rate increased from 0.07% to 20%, but we could afford to be more conservative on this value, especially when it comes to predicting violent crime. The accuracy rate did decrease 5%, which is not ideal. The AUC is about the same at around .592.

Future Improvements For Logistic Regression

- With our current time constraints, we understand that modeling human behavior with our predictors is a difficult task. If we were to have more time, we would try to search for more datasets that include more variables regarding human behavior that could have a correlation with crime (literacy rates throughout Baltimore, socioeconomic levels etc.)
- We would also run more iterations of the model to try to further balance sensitivity, specificity and accuracy.

Comparison of Model 1 and Model 2

- The results from our two logistic regression models did not show that much of a difference, especially when looking at the AUC values of 0.58 vs. 0.59, however, the biggest difference would be with the accuracy rate and sensitivity rate.
- Between the two models, our accuracy rate decreased from 70% to 65% and our sensitivity rate increased significantly from 0.1% to 30%. The increase in sensitivity resulted in a decrease of type II error, which means the second model predicted true positives at a higher rate and false negatives at a lower rate than the first model. This value, however, could still be improved by potentially modifying the probability threshold once again.

- Type II error, or the probability that the model predicts no violent crime when in fact there is one, led to an increase in type I error, which is preferred over a high type II error rate.
- Overall, we recognize that both models are not the most ideal, but the second model is a better fit for our interest in predicting if a violent crime will occur or not even if our accuracy rate is lower than our first model.

In the end, the most significant variables used to predict if a violent crime would occur or not in Baltimore for logistic regression are month, district, precipitation, average temperature , unemployment, and part of day.

Decision Tree

Overview

We chose to use decision tree analysis methods to create models that predict daily total crime and significant variables. Decision tree is a one of the popular methods in machine learning analysis that provides a good visualization of the result. This would be beneficial because the decision tree almost mimics the way humans make decisions, and may provide a good way to prepare for upcoming crimes.

Assumptions

- There is no particular assumption that needs to be met to be able to use this method.

Model Building

Initial Tree Model

- We splitted the dataset into half to create a train dataset to build a model and a test dataset to test our model.
- We first created a very basic tree initially. See Figure D below.

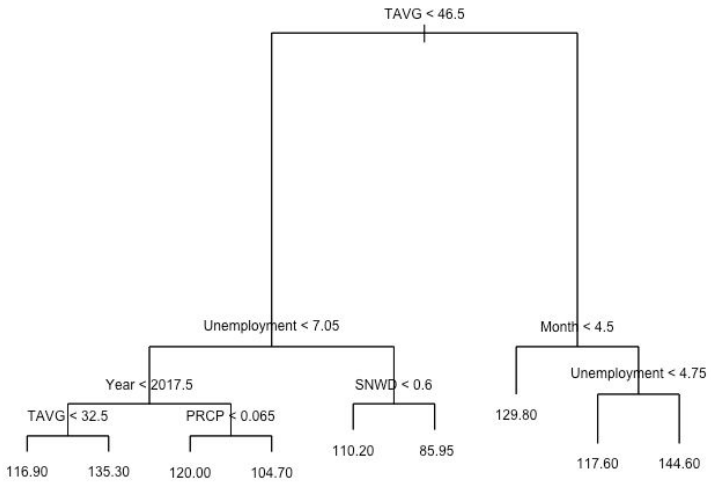

Figure D

Here we can see that the temperature average is the most important variable. For a day with temperature less than 46.5 F, the next most significant variable was unemployment, followed by year and snow depth. For a day with temperature more than 46.5 F, then Month was the next significant variable, followed by unemployment.

- The test MSE was 22.93837.

- Tree Pruning
  - Pruning the tree finds a smaller tree with a smaller variance that performs better on test data, but since even after pruning the size of the tree is the same as the unpruned tree. So disregard this part.
- Although this initial tree gives us some good idea on what affects daily total crime, it is not very reliable. It is vulnerable to a small change in data, and less accurate then most of the other approaches. However, using methods like bagging, random forest and boosting that aggregates a lot of decision trees, our model can become substantially more reliable.
  - Bagging/Random Forest
    - Bagging and random forests use the bootstrap to enable us to reduce the variance of the tree by taking the "average" prediction across many trees. The only difference between bagging and random forests is the number of predictors that are candidates for splitting during each split in the tree: in bagging, all predictors are candidates, whereas in random forests, a subset of predictors are candidates.

After Bagging

- For bagging, we used all predictors to be considered for bootstrap.
- Generated 2000 trees in total.
- % Var explained: 28.98
  - % Var explain is a measure of how well out-of-bag predictions explain the target variance of the training set. Unexplained variance would be due to true random behavior or lack of fit.

- Even though it seems low, we think that this is expected since it is hard to predict people's behavior with little information.

- Because it is hard to interpret a large number of trees, we would obtain an overall summary of the importance of each predictor using the RSS.

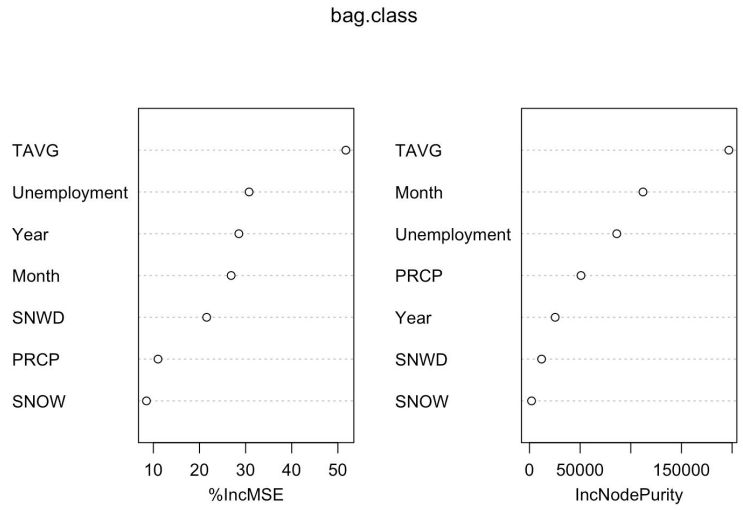

Figure E

- The first measure plot is based on the decrease in accuracy in predictions on the out-of-bag samples when a given predictor is excluded from the model, averaged over all trees. See Figure E above.
- The second measure plot is the decrease in node impurity from splitting on the predictor, averaged over all trees. Since the response is quantitative, the decrease in the residual sum of squares is used.
- Based on these two graphs, temperature average is the most important predictor, followed by unemployment, month, and year.
- Test MSE after bagging: 18.83332

After Random Forest:

- Usually, when using a random forest method to build a regression tree, we would use the p/3. In this case it would be 2.33, so we tried both 2 and 3. Using 2 gave a better test MSE, so we ended up going with 2.
- Also generated 2000 trees.
- With random forest, the % Var explained increased to 33.77%.

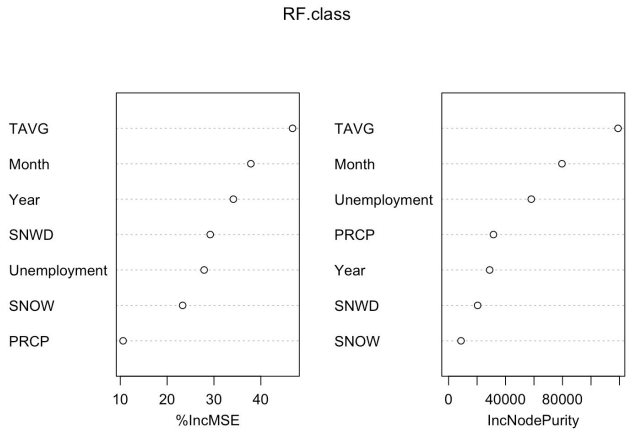

Figure F

- Figure F is the same plot as Figure E, but with random forest.
- The test MSE after random forest: 17.33683.

Boosting

- Boosting is another way to improve the predictive performance of tree-based methods on test data, by building trees sequentially where a tree is built given the information provided by previous trees.
- We also generated 2000 trees.

	var	rel.inf
TAVG	TAVG	29.5512063
Unemployment	Unemployment	29.1941317
Month	Month	17.9533018
PRCP	PRCP	13.4715205
Year	Year	7.2931761
SNWD	SNWD	2.2106371
SNOW	SNOW	0.3260265

Figure G

- Below is the plotted version.

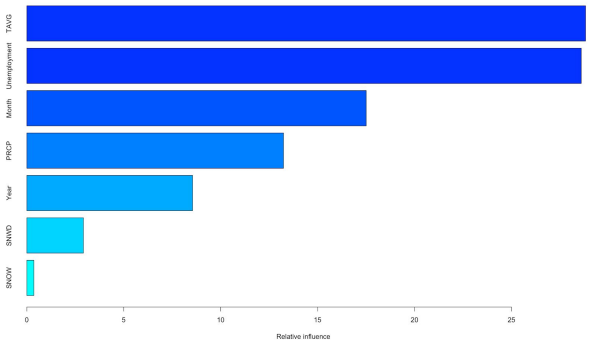

Figure H

- The test MSE after boosting is 16.98436.

Model Comparison

- The best model came out to be the one after boosting, with a test MSE of 16.98436, the lowest out of all three.
- Looking at all analysis results, it can be confirmed that average temperature of the day is the most important variable in predicting total number of crimes. The next important variable turns out to be unemployment, following with Month and precipitation.

Linear Regression

Overview

We chose to use linear regression to create a basic model to answer how many crimes will occur in a day based on our predictors. This model can be used as a reference for other models that we create since linear regression is not the best model to rely on in this case.

Analysis

Assumptions

- In order to conduct linear regression, there are four main assumptions that need to be met.
- First assumption is linearity. This has been checked by making a scatter plot for each explanatory variable with the response variable. Variables 'Snow' and 'SnowDepth' were removed because they did not meet this assumption.
- Second assumption is homoscedasticity. This assumption was checked by creating a residual plot, and the plot showed homoscedasticity.
- Third assumption is that observations are independent of each other. This has been met since we are assuming that each crime is independent from one another.
- The last assumption is normality. This was confirmed by creating a QQ plot.

Model Building

- We used the function lm to create a model with the chosen variables.
- The significant variables were 'Month', 'Temperature', and 'Precipitation'.
- The Adjusted R-Squared was 0.3041 which is not high and indicates that our model does not do a good job in predicting the outcome, but this is expected as it is hard to predict people's behavior with only a few variables.
- The final model:
    - Total Crime = 1506.4027 + (-0.7054)*Year + (1.3848)*Month + (0.0863)*Day + (0.5819)*Temperature + (-6.8273)*Precipitation + (1.0690)*Unemployment

Time Series

Overview

In addition to the linear regression model, we also wanted to create a time series model to predict daily total crime based on past values. This would be beneficial to Baltimore PD because the results could give an indication as to how they should staff for the next week.

Assumptions

- One major assumption that must be met in order to form a time series model is that the data must be stationary.
    - The data had to be detrended because it initially had a strong seasonal pattern, as well as a nonlinear trend.
- Once the model parameters have been chosen, if the model fits well all higher order coefficients will be significant and the model's diagnostic plots must follow certain patterns:
    - The residual plot will have no clear pattern, indicating constant variance.
    - The ACF plot of the residuals should be 0 after lag 0.
    - The QQ plot values should fall on the diagonal line, indicating normality
    - The Ljung-Box p-values should be insignificant, rejecting the alternative hypothesis that the fitted time series is autocorrelated.

SARIMA Models

Model Building

- Notice the seasonal and nonseasonal trend in Figure I, which must be removed.

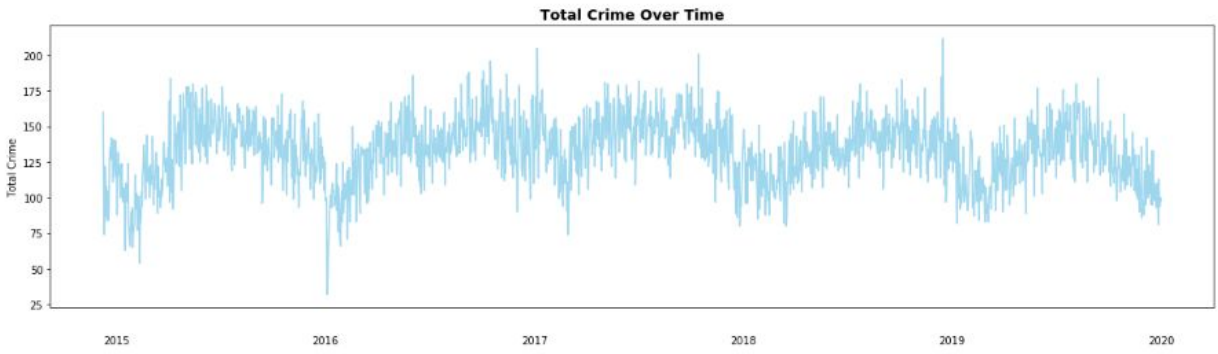

Figure I

- Examine the periodogram in Figure J, which indicates that the data should be smoothed. A Daniel kernel with L=15 was used to determine the period. On the smoothed periodogram, you can see spikes at around 0.15 and 0.29, which indicates the frequency is 0.15 and the period is 6.67, so m=7.

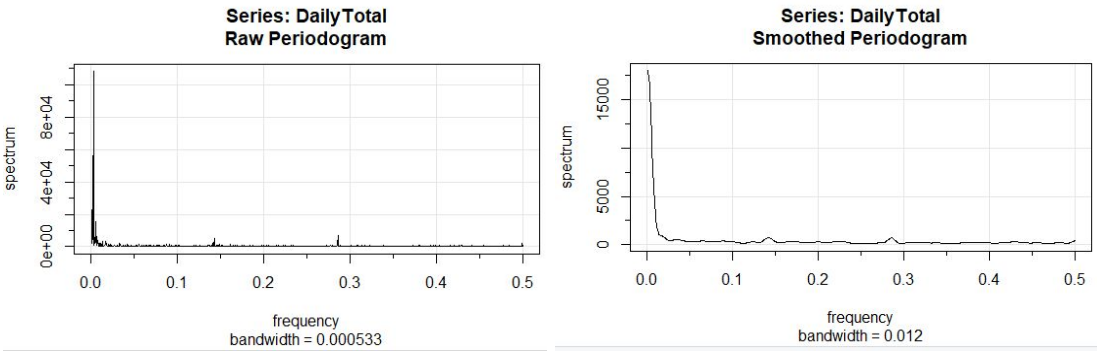

Figure J

- Examine the time series plot in Figure K with the last 100 observations so that the trends can clearly be observed. To remove the trends, the first difference was taken, in addition to the seasonal difference with a period of 7.

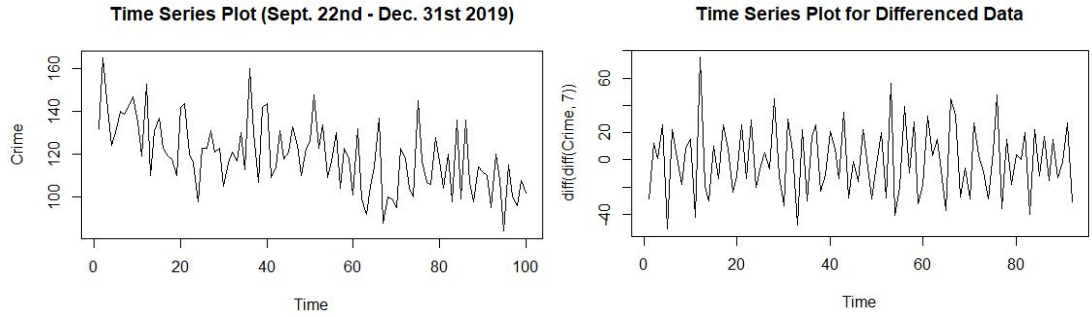

Figure K

- From there, several possible parameter values for a SARIMA model were chosen from the ACF and PACF plots in Figure L below.
    - For the nonseasonal AR component, note that the PACF is significant for the few first lags. This indicates that p=1,2, 3, or 4.
    - For the nonseasonal MA component, note that there is a decay present in the first few lags of the PACF plot, and that the ACF is significant at lag 1 and 3. Lag 3 might be a false positive, or it might be significant so q=1 or 3.
    - For the seasonal AR component, note that the ACF has a decay at lag 7, and the PACF is significant at several multiples of 7. P=1,2, 3, 4, or 5.
    - For the seasonal MA component, note that there is a decay present after each multiple of 7 in the PACF plot, and that the ACF is significant at lag 7 so Q=1.

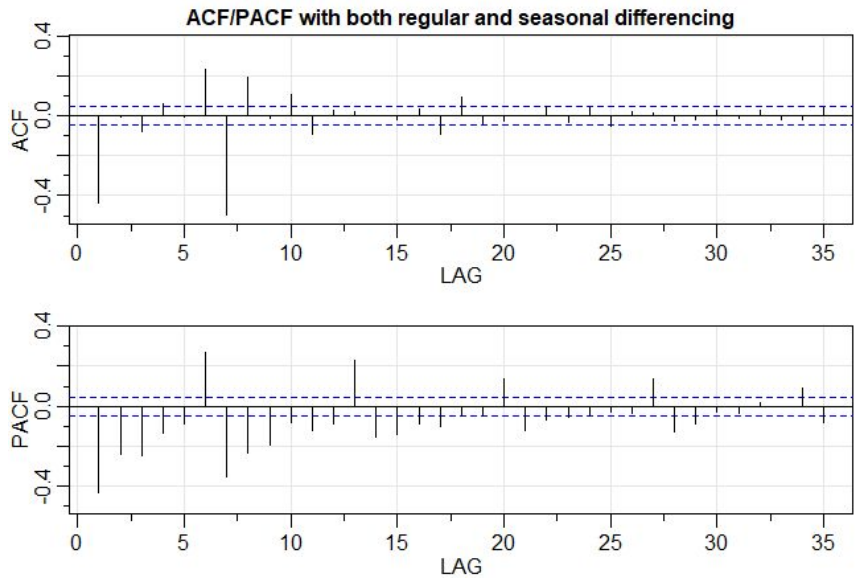

Figure L

- Of the many models tried, only two models had significant higher order coefficients as well as sufficient diagnostic plots. See Figure M below. Both the SARIMA(1,1,1)x(0,1,1)$_7$ model and the SARIMA(0,1,3)x(0,1,1)$_7$ model satisfied these requirements:
    - All higher order coefficients were significant
    - The residuals look constant from the residual plot, and the data appears normal from the QQ plot.
    - The p-values are mostly insignificant for the ACF of the residuals.
    - Most or all p-values for the Ljung box statistic fall above the significance line, except for one value on the SARIMA(1,1,1)x(0,1,1)$_7$ model , which is sufficient.

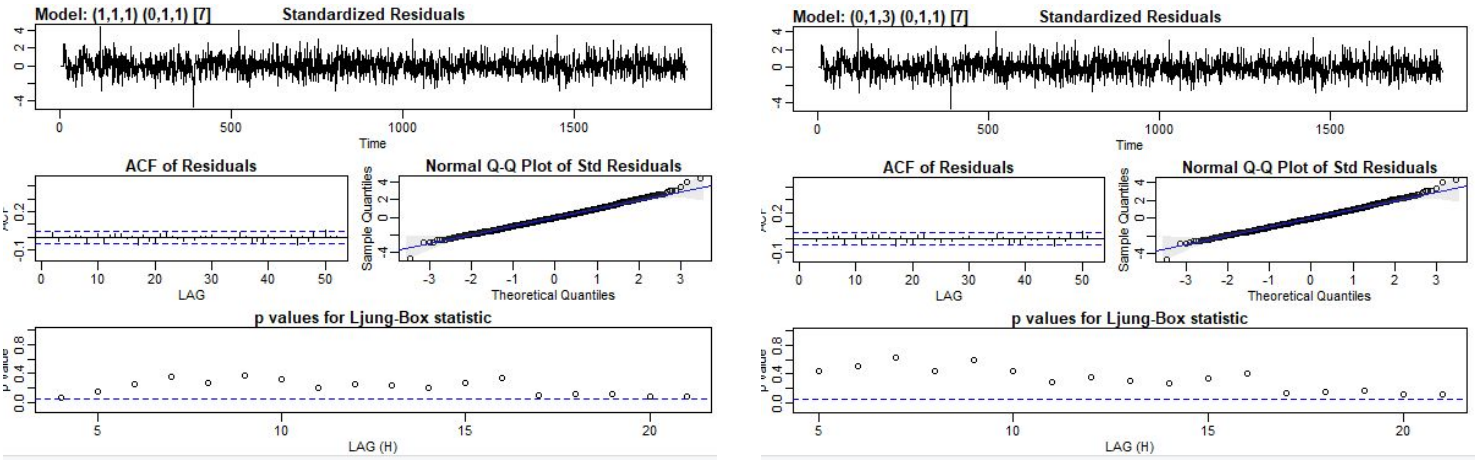

Figure M

- Both models were fitted to the data until Dec. 31st, 2019. Then it was used to predict for 31 days. See Figure N and O below.

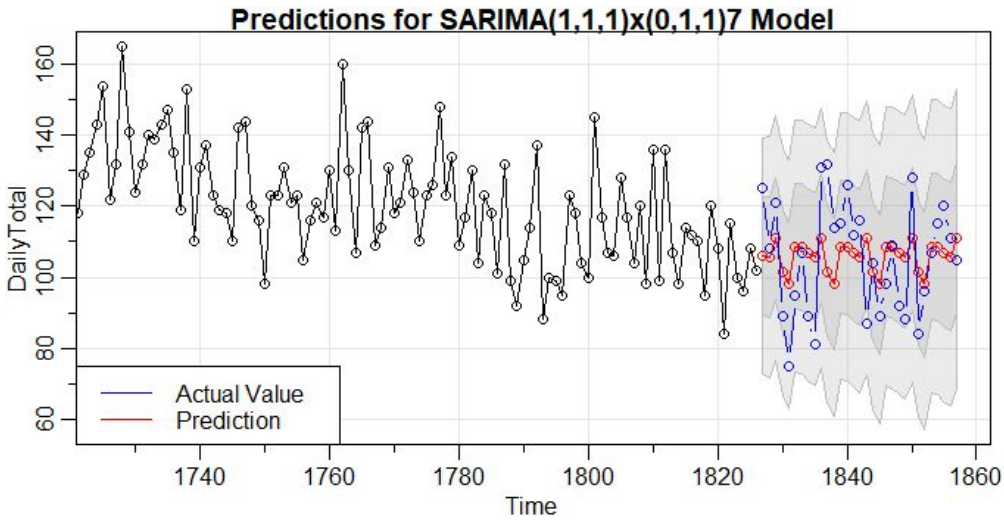

Figure N

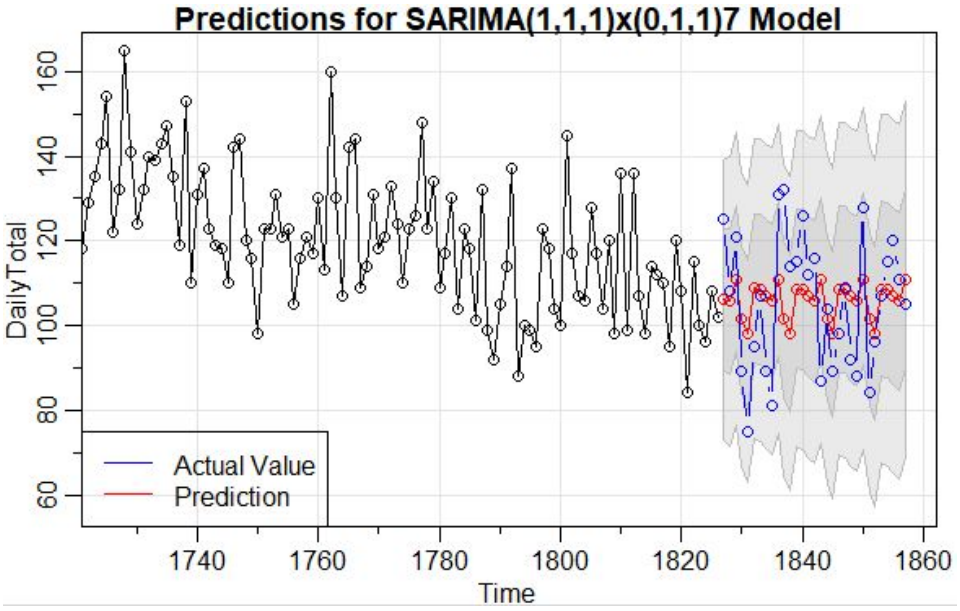

Figure O

Comparing the Models

Figure P contains several of the most common metrics to measure accuracy, as well as three information criterion (AIC, BIC, AICc). The information criteria provide an estimate of how much information is lost by a given model, and are performed on the training data set. The accuracy metrics were performed on the test data, the month of January in 2020. The better values are bolded in the table.

	Model 1 SARIMA(1,1,1)x(0,1,1)$_7$	Model 2 SARIMA(0,1,3)x(0,1,1)$_7$
AIC	8.449309	**8.448679**
AICc	8.449316	**8.448692**
BIC	**8.461382**	8.463771
MAE	**12.21906**	12.22173
RMSE	**14.5193**	14.53334
ME	**0.4678299**	-0.7078554
MPE	**-2.50532**	-2.73881
MAPE	**12.08573**	12.11812

Figure P

- Model 1 has a smaller BIC and a larger AIC and AICc.
    - Since the sample size is large, AIC and AICc are very close. Both select model 2.
    - AIC and BIC and very similar measures, but BIC places a heavier penalty on complex parameterization, which is why model 1 was selected.
- All of the accuracy metrics select model 2.
    - RSME is the key metric here, as it is measured in the units of the original data and penalizes outliers more heavily than MAE.
        - In staffing, the further off the number of officers is the worse the impact is.
    - The MPE is provided because it is easy to interpret - the actual values are on average 2.5% less than forecasted for model 1, and 2.7% less than forecasted for model 2.
    - The MAPE is provided because it is easy to interpret - the absolute value of the difference between the actual values and the forecasts is on average 12.09% for model 1, and 12.12% for model 2.
- Since the AIC, AICc, and RSME select for model 2, this is the better model.

Results

- Both the SARIMA(1,1,1)x$(0,1,1)_7$ model and the SARIMA(0,1,3)x$(0,1,1)_7$ model were sufficient in predicting the total daily crime, although the second model did perform better.
    - Both models had prediction intervals for January that contained the real value.
- This indicates that crime can be modeled with seasonal components - in particular, this tells us that the weekly cycle is important in crime.

Future Improvements for Time Series

- Both models performed reasonably well, however a more accurate model might exist. In particular, lowering the standard error will provide narrower prediction intervals.
- Additional algorithms might yield more insight, such as:
    - Holt Winters Exponential Smoothing is used for univariate time series with trend and seasonal components, which makes it a good fit since crime is a scalar.
    - Seasonal Autoregressive Integrated Moving-Average with Exogenous Regressors (SARIMAX) is an extension of SARIMA modeling, and includes exogenous variables.
- Another possibility is the use of multiple seasonality trends. No viable models with the lag of 365 were found, but it might be that the weekly trend was obscuring the yearly trend. Modeling with the yearly and weekly trend might produce better results.

Future Improvements for All Models

All of the models would benefit from adding more variables that represent human behavior to be able to predict crime more accurately, such as:

- Additional indicators of economic condition:
    - Poverty level
    - Cost of living
    - Consumer price index.
- Information by district:
    - Literacy rates
    - Socioeconomic levels
    - Demographics (ethnic and racial makeup, age composition, gender composition).
- Information by neighborhood:
    - Population density
    - Degree of urbanization
    - Median income
    - Youth concentration

    - Crime reporting practices of the residents
- Information about the offender provided by the BPD such as age or gender
- If the crime occured during a holiday, festival, and/or school vacation period.

Appendix

All of our code can be found on the github repository https://github.com/crsimmons1/capstone

The final deliverable is the Rrubs dashboard https://rpubs.com/jj_99/baltimorepdcrime.

Sources

Anderson, C. A. (1987). Temperature and aggression: Effects on quarterly, yearly, and city rates of violent and nonviolent crime. Journal of personality and social psychology, 52(6), 1161.

Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. American sociological review, 588-608.

Hipp, J. R., Curran, P. J., Bollen, K. A., & Bauer, D. J. (2004). Crimes of opportunity or crimes of emotion? Testing two explanations of seasonal change in crime. Social Forces, 82(4), 1333-1372.

Newberg, L. (2005). Some Useful Statistics Definitions. https://www.cs.rpi.edu/~leen/misc-publications/SomeStatDefs.html

Raphael, S., & Winter-Ebmer, R. (2001). Identifying the effect of unemployment on crime. The Journal of Law and Economics, 44(1), 259-283.