Sam Pennington and Brianna King

Data Mining – Fall 2018

Analysis of Potential Correlation in Atlanta

Weather And Crime

**Introduction**

Atlanta, Georgia is a city of art, business, technology, and mayhem. With the city containing an urban population of 4.5 million and a metropolitan area population of 5.6 million in addition to having the nation’s busiest airport, millions of people travel throughout Atlanta on a daily basis. Unfortunately, some persistently have malicious intent. For the majority of the previous three decades, Atlanta has been ranked as one of the top five highest violent crime cities in the nation and the top city for violent crime in the entire state of Georgia. Curious about why and how this could be happening, we began researching different statistics.

According to a recent article published by the AJC, the Atlanta police department has tracked up to 1,000 different street gangs and over 50,000 individual gang members working within the limits of metropolitan Atlanta. Law enforcement continuously attempts to prevent gang activity and recruitment, but the number of individuals involved in gang activity and recruitment continues to grow at an exponential rate, and law enforcement is simply outnumbered. While this myriad of gangs has certainly contributed to the high rate of crime present in Atlanta throughout recent decades, the FBI strongly cautions against interpreting nation-wide “dangerous city” rankings strictly by crime rate and the presence of gang activity since bare statistics ignores various other important contributing factors such as population density, degree of urbanization, economic conditions, modes of transportation, religious and cultural characteristics, and so on.

According to the NSW Bureau of Crime Statistics and Research, the question of “What factors cause individuals to commit crimes or become involved in crime activity?” is not as simple as it appears. Crime-prone individuals have a variety of contributing factors that may lead them to becoming more involved over time, such as biological factors like genetic propensity, intelligence, truancy, substance abuse, etc. However, the NSW Bureau of Crime Statistics and Research also emphasize that environmental factors such as socialization, family dynamic in regard to parenting, public tolerance of crime, etc. most certainly contributes to crime activity in a given area as well. Therefore, we wondered if the environmental or ecological factor of weather could potentially be a contributing factor as well. Our project will attempt to draw a correlation between crime rates and weather data for the 2017 year in Atlanta. We will evaluate features of data including type of crime committed, type of weather on a given day, temperature, rate of crime committed on a day, and location of the crime committed.

**Prior Studies**

Throughout the previous decades, many studies have been conducted that attempt to correlate the weather on a given date with crime rate in a given area. WeatherOps, a largely successful company that provides organizations with weather decision support, conducted a study to assess the potential correlation between temperature and crime. The industry-leading company published a set of data on their blog website that detailed the statistical data on various types of crime in the Chicago area. They created visual data to display their findings, and the temperatures measured ranged from -2 to 102 degrees Fahrenheit. These graphs were split by crime type which included assault, battery, narcotics, and stalking. The study focused on crimes of opportunity, meaning crimes that occur largely because the chance is there. For example, starting your car to warm up on a cold morning and re-entering your home with the car unlocked would be an ideal opportunity for a break-in, due to the fact that no real effort was required. Unfortunately, this study was only able to draw one conclusion: crimes of opportunity decrease significantly when the temperature is either really hot (>90 degrees) or very cold (<30 degrees). In all other aspects correlating to a given crime and given weather data, no conclusions could be drawn. Crime is shown to steadily rise in the temperature range in between these two values for Assault and Battery while spiking around the 30-degree mark and staying rather steady until the 90-degree point.

According to the Chicago Tribune, the most-read daily newspaper of the Chicago metropolitan area, the seasonal cycle of of Chicago crime peaks during hotter months, but certain crimes are more susceptible to hotter temperatures than other. To back this claim, they conducted a study that focused on different types of crime crime in Chicago from 2012 - 2017. The crimes ranged anywhere from less serious crimes such as theft to incredibly serious crimes such as homicide. The Tribune built a statistical model to quantify the effect of temperatures (or correlations rather) on the different types of crimes available from their dataset. The results showed that all measured crimes had a positive, linear correlation to temperature increases with the exception of homicides. Shootings had the highest rate of increase per degree temperature rise averaging an additional 5 shootings per degree rise, while narcotics had the flattest slope being almost negligible. Homicides were shown to not correlate linearly with temperature and were mistakenly graphed as a flat line.

Because crime is prevalent in most cities throughout the world, we can examine a study conducted by the University of Canterbury that used daily data from 43 districts across New Zealand over the span of a decade in order to investigate the effect of weather on crime. A common myth exists in New Zealand that claims “the Nor’wester wind causes ‘disorderly’ behavior”, and an extension of this study was to empirically investigate this common claim. Property crime and violent crime were the categories analyzed, and temperature and precipitation were used as measurements of weather. They measured the correlation between North-westerly wind and violent crime as well. The General Affect model suggests that higher temperatures facilitate affective aggression, meaning that aggression displayed in hotter temperatures are usually for the sole purpose of harming another individual (Cohn and Rotton 2000a). Contrarily, the Routine Activity theory suggests that criminal activity is more likely to occur provides three key elements are present: a suitable target, a motivated offender, and the absence of capable guardians against the crime as it is occurring (Cohen and Felson). Since better weather (less precipitation and moderate temperatures) are more likely to lead to social outings, the absence of a capable guardian against a crime is more likely given that the weather is nice. In other words, on a beautiful day, a crime (specifically a property crime) is more likely to happen under this theory. The study concluded that both temperature and precipitation have a significant effect on the rate of violent crimes, with higher temperatures and higher levels of precipitation correlating with higher levels of violent crime in a given period, thus aligning with with the General Affect model in terms of temperature. Additionally, the study concluded that only temperature has a significant effect on the number of property crimes recorded, with moderate temperatures indicating higher levels of property time, thus aligning with the Routine Activity theory. The study was inconclusive in regards to the North Westerly winds causing disorderly conduct, however.

Studies have also been done based on the premise of how humidity and rainfall correlate to crime. One such study, cataloged in the British Journal of Criminology, was unable to draw any definitive conclusions about how rain and crime were correlated as the experimented was performed on three separate datasets and drew three separate conclusions. One experiment showed a positive correlation, one showed negative, and one showed negligible correlation. Humidity was also studied with slightly more conclusive results. An experiment performed on multiple different sets of data showed a slightly negative correlation between humidity and crime rate.

From the previous studies mentioned, we can conclude that different experiments will lead to drastically different results. Comparing different experiments that all attempted to find and analyze a potential correlation between the crime rate in a given area and the weather in a given area on a given day simply did not help us in attempting to predict whether or not a correlation would be present. Most interestingly, these potential correlations have never been studied in the Atlanta area, which is widely considered to be one of the most dangerous places in the world based strictly on bare statistics.

**Data Set**

Our overall dataset consists of two subsets of data. The first subset is a comma separated value file detailing crimes that were committed within the Atlanta area. This dataset was found on Kaggle.com[[1]](#footnote-1) and is based off an original data set published by the Atlanta City Police Department. This ACPD data was compiled, cleaned, and republished for our usage. This dataset contains information about the type of crime committed, the location it was committed, the time is was reported, the number of subjects involved, etc.

Our second subset contains data about the weather. This data was measured by NOAA (National Oceanic and Atmospheric Association) from the Fulton County government station. Our data was obtained by submitting a request to the organization for daily weather data for Atlanta on each day in 2017. This information can be requested through their online portal[[2]](#footnote-2). This data includes maximum and minimum temperatures, precipitation information, windspeed, humidity, etc. Measurements were taken approximately every 30 minutes and the data can be averaged together on a day by day basis to do a more generalized analysis.

**Data Cleansing**

The data sets that we used came with a lot of information that we did not deem necessary. As a result, we decided it was best to not waste memory or load time within our Python notebook, and instead deleted the columns of unnecessary data entirely. After going through each of the two spreadsheets, we narrowed down the data to the following columns:

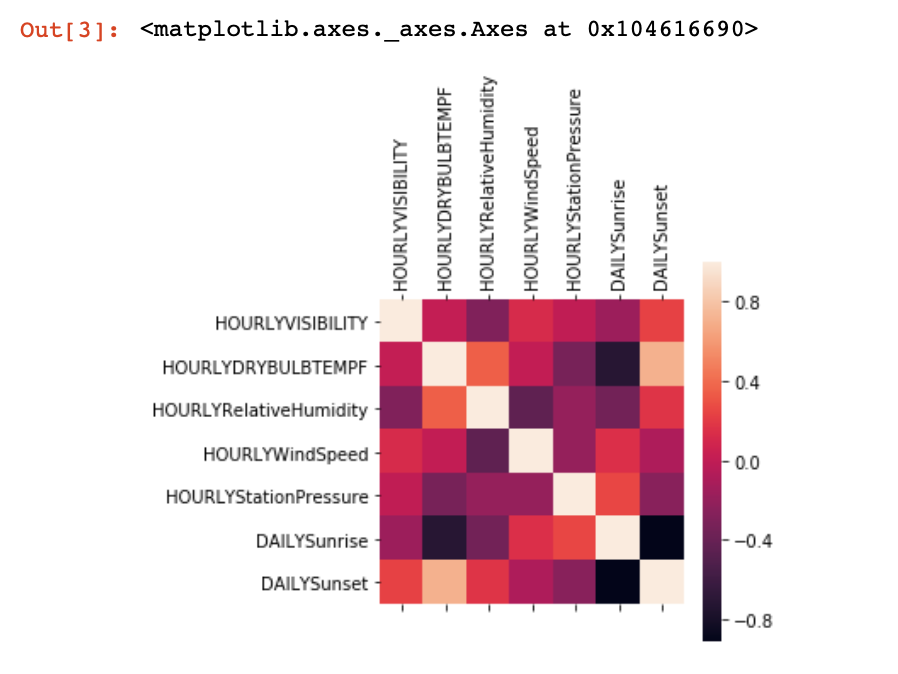
* Crime Data: Date, occurrence time, possible date, possible time, maximum number of victims, the day of the week, the string literal of the type of crime, the neighborhood, and the geographic coordinates where the crime occurred
* Weather Data: Data, visibility, dry bulb temperature in Fahrenheit, relative humidity, wind speed, atmospheric pressure, and the sunrise and sunset times

The weather data needed further cleansing because the data was taken multiple times each day. To overcome this, we averaged the results of each category within each day. These results were then saved in place of the original data, resulting in averaged data for each day of the year.

Within the crime data set, we used a code creator to assign a code to each of the unique string objects within each data feature. This was done in order to perform our classification models and tests. There were also some blank entries within certain rows and columns. To smooth out the data, we decided to take the average of all other values in the column and fill in the missing values with that new value. This is a much safer way to handle missing data rather than filling in “0” values. It also maximizes the amount of data points in the set by not deleting all rows with a single missing value.

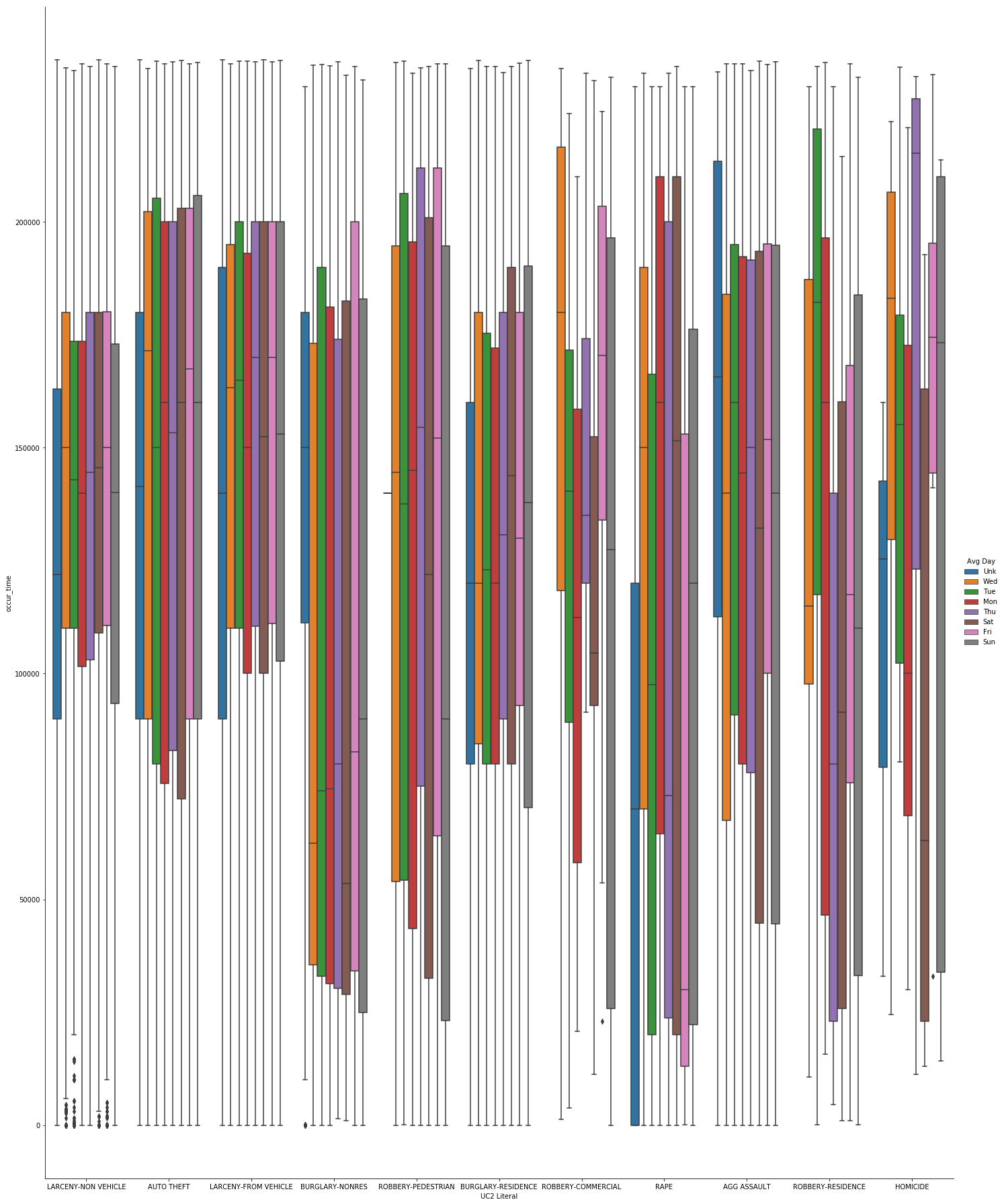
**Exploratory Analysis**

Before beginning to run analysis on a combined weather and crime data set, we decided to run exploratory analysis on each set individually. The aim of this process was to understand our data better in order to better understand the results of our analysis later on. We first decided to draw a correlation matrix based on the numeric attributes of the weather data.

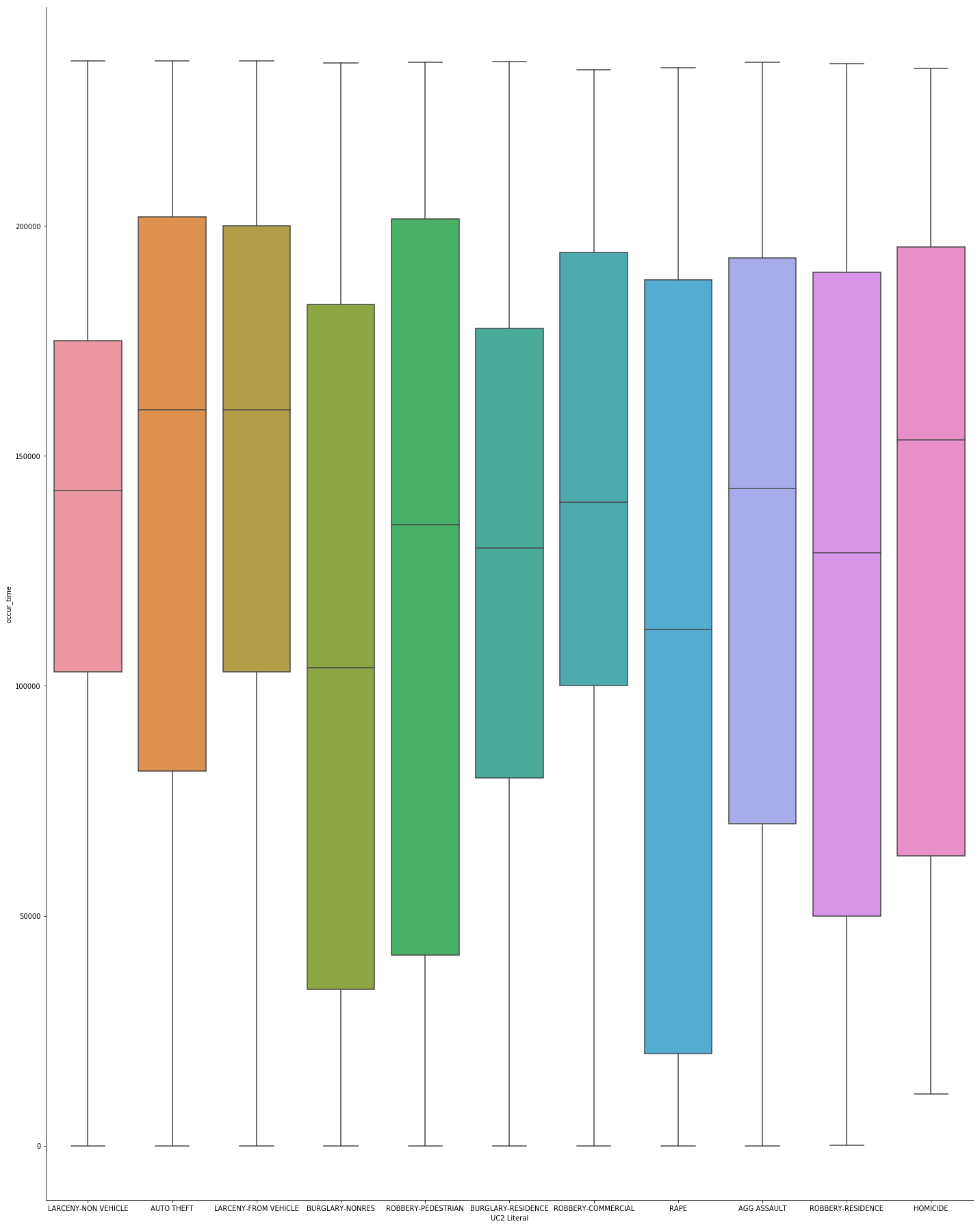


From this correlation matrix heat map, we were able to see that there are strong negative correlation ties between the sunrise and sunset. This would of course be because as days get longer, the sunrise is earlier while the sunset is later, and as the days get shorter, the sunrise gets later while the sunset is earlier. This shows inverse correlation which matches with our matrix. There are also strong positive correlation ties between the relative humidity and the hourly dry bulb temperature. This is to be expected as the temperature generally rises in the summer and cools in the winter. Based on the climate type of the state of Georgia, you can also expect wet, rainy summers, and cool, dry winters. This would seem to match the matrix as a hot summer would be positively correlated to a wet summer, while a cool winter would correlate to a dry winter. One final thing we can see from this matrix is that the dry bulb temperature is positively correlated to the sunset time and negatively correlated to the sunrise time. This would seem to be because as the sunrise gets earlier, and is therefore a smaller numeric value, and the sunset gets later, and is therefore a larger numeric value, the days are longer and the sun has more time to heat the Earth. This would of course result in a higher temperature and would likely be seen in the summer. The result of this matrix is that we should be able to conclude that if one feature in the crime set seems to predict a feature in the weather set or vice versa, then that same crime feature should be able to predict/be predicted by a strongly correlated feature to the corresponding weather feature.

When doing exploratory analysis, we were mostly interested in the average time of day and average day of the week that each type of crime was occurring. Pictured below are two boxplots that are broken out by crime type. The first plot shows both time based on day of the week while the second shows an average time regardless of the day of the week.

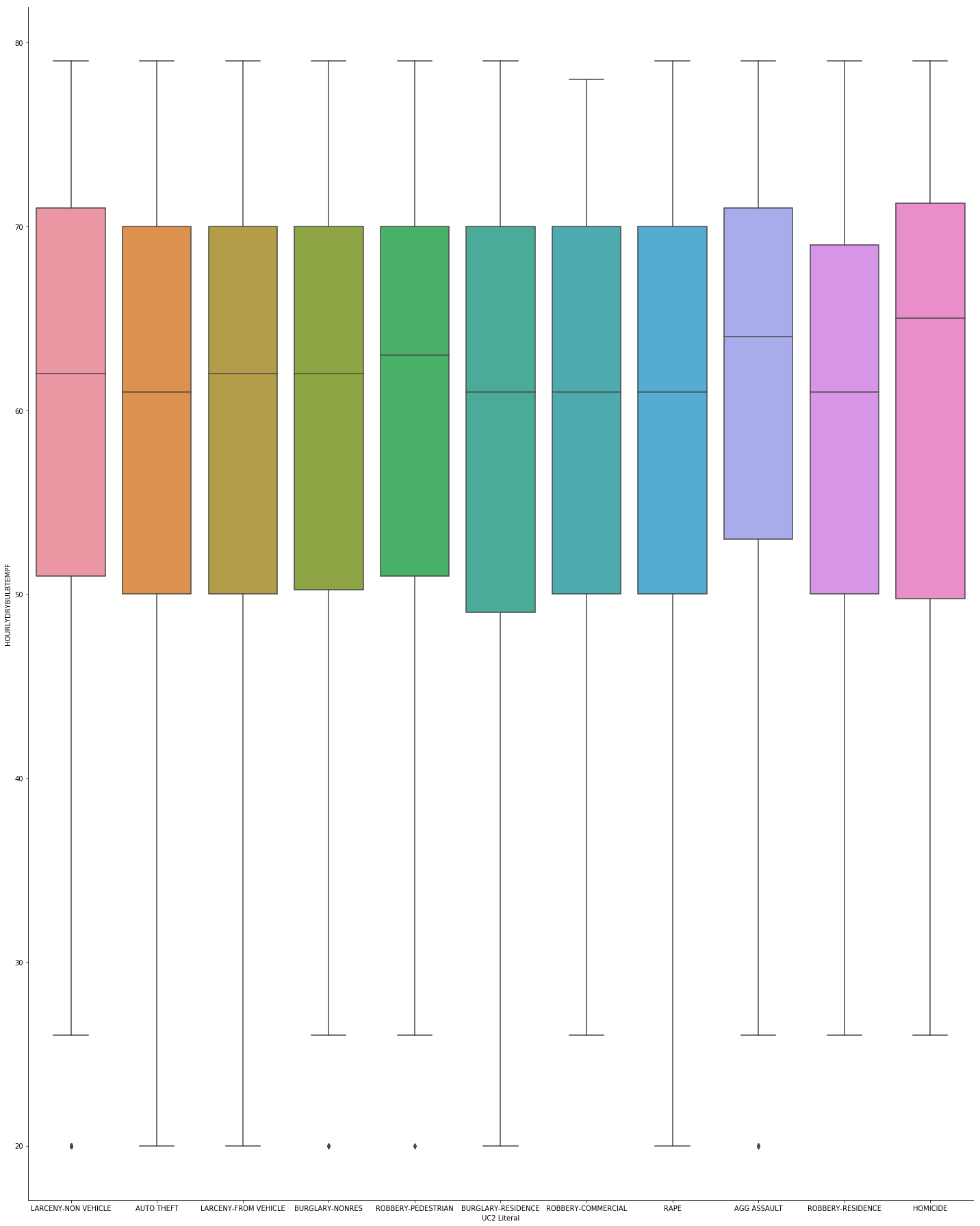


This graph shows a few interesting things. First, it does not seem to matter for most crimes what day it is, as most incidents of the crime will occur around the same time each day. Second, this graph shows that there is a lot of disparity between when a homicide is likely to occur depending on the day of the week. It appears that people are more likely to commit homicide in the evening on Friday, while being more likely to commit homicide in the early morning hours on Saturday. One speculative reason for this is that people are out and about in the beginning hours of the weekend as they have completed their work-week tasks and therefore there are more potential people to commit this crime. This however does not appear to be a trend with any of the other tracked crime types.

****

Compared to the other box plot, this graph shows us significantly less interesting information. One possible thing of note is that most larceny cases are observed between lunch and dinner. Auto theft incidents as well as aggravated assault incidents are also mostly concentrated during the workday hours.

After combining our data into a larger Crime and Weather data set, we did exploratory analysis on the temperature in relation to the type of crime. The results shown below indicate that crimes tend to happen very infrequently in extremer temperature settings. Most crimes in this data set were committed when the temperature was in the range of fifty to eighty degrees Fahrenheit.

****

**Results**

We ran various machine learning algorithms on a few different pairs of attributes. We decided to use a Decision Tree algorithm, a K Nearest Neighbors algorithm, and a Random Forest algorithm.

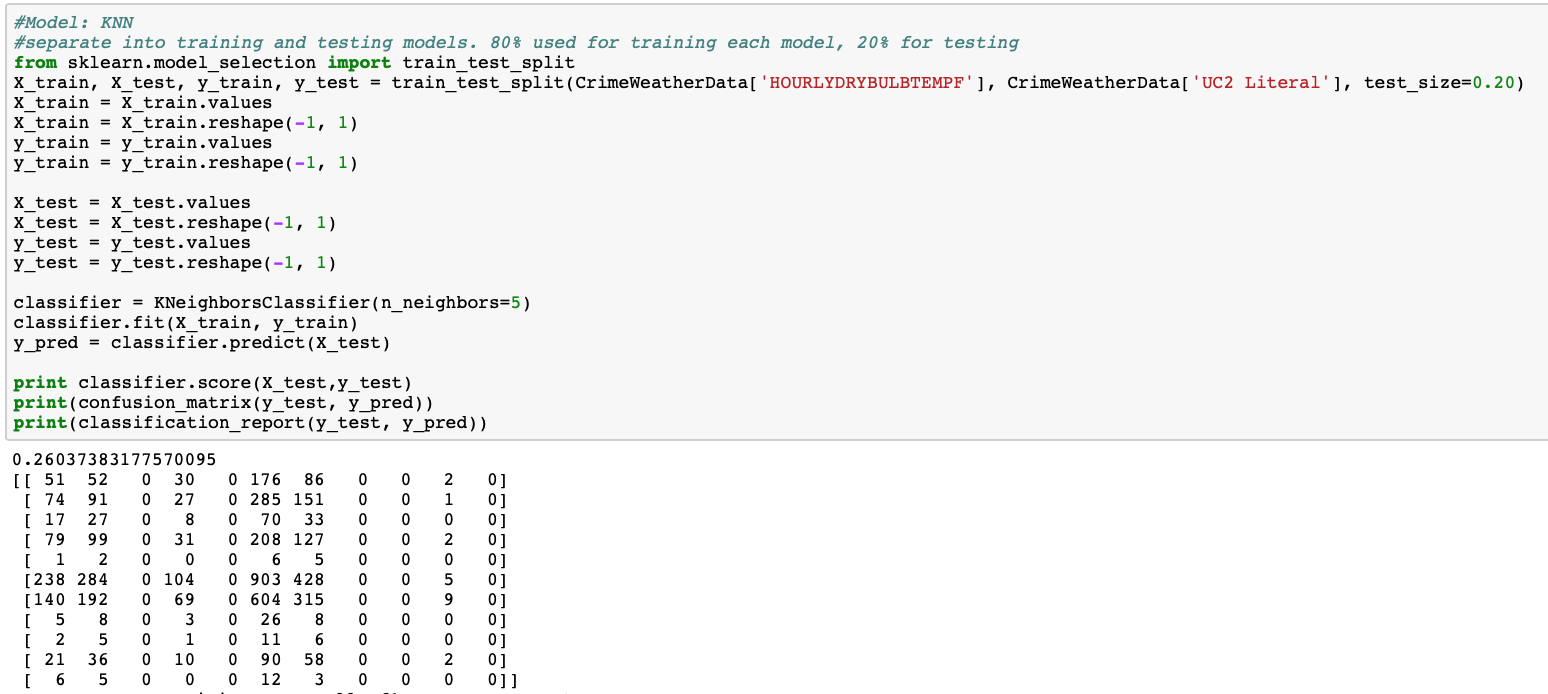
Temperature vs. Crime Type

Figure 5. K-Nearest-Neighbors algorithm

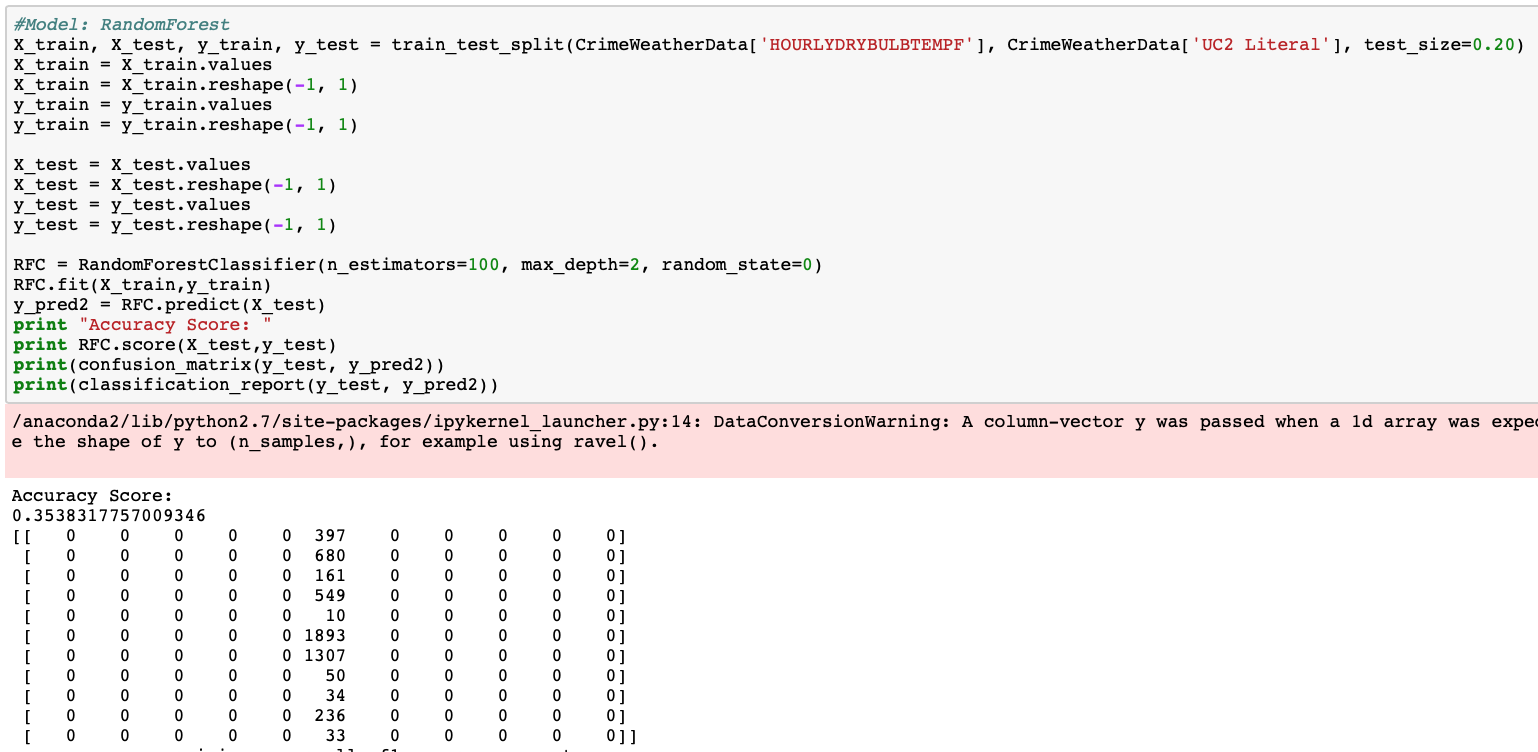


Figure 6. Random Forest algorithm

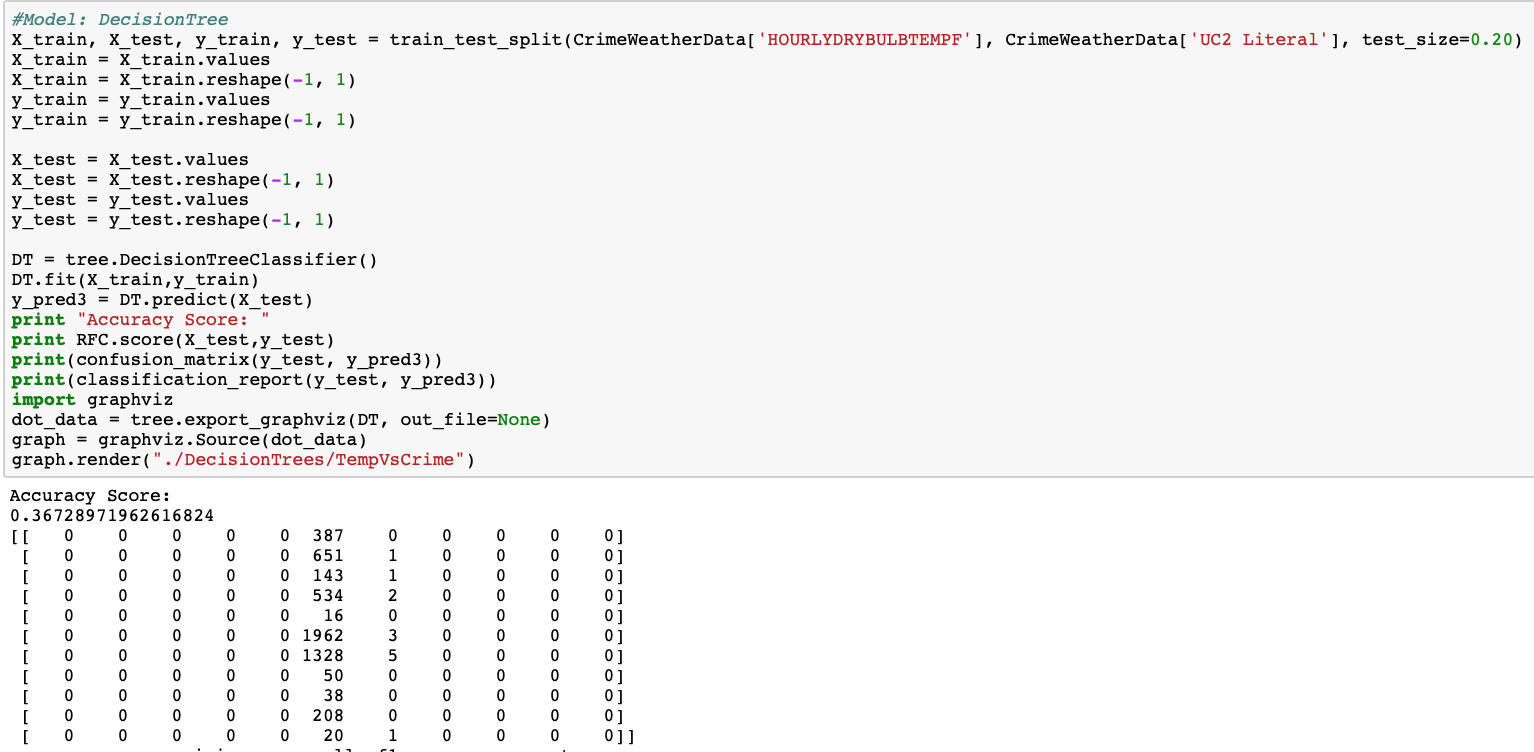


Figure 7. Decision Tree algorithm

As seen in the figures above, temperature does not seem to be a good predictor of crime type. The highest accuracy we were able to achieve after running each of these tests multiple times was 36.7% when using the Decision Tree algorithm. The confusion matrices for both Random Forest and Decision Tree showed decent results but had low accuracy while the confusion matrix for the KNN algorithm was abysmal and this algorithm produced poor accuracy scores as well.

Time vs Crime Type

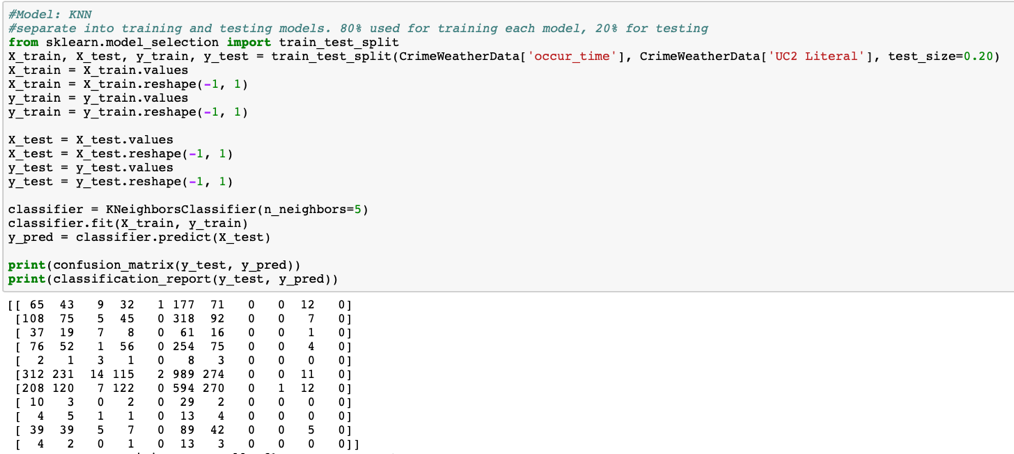


Figure 8. K-Nearest-Neighbors

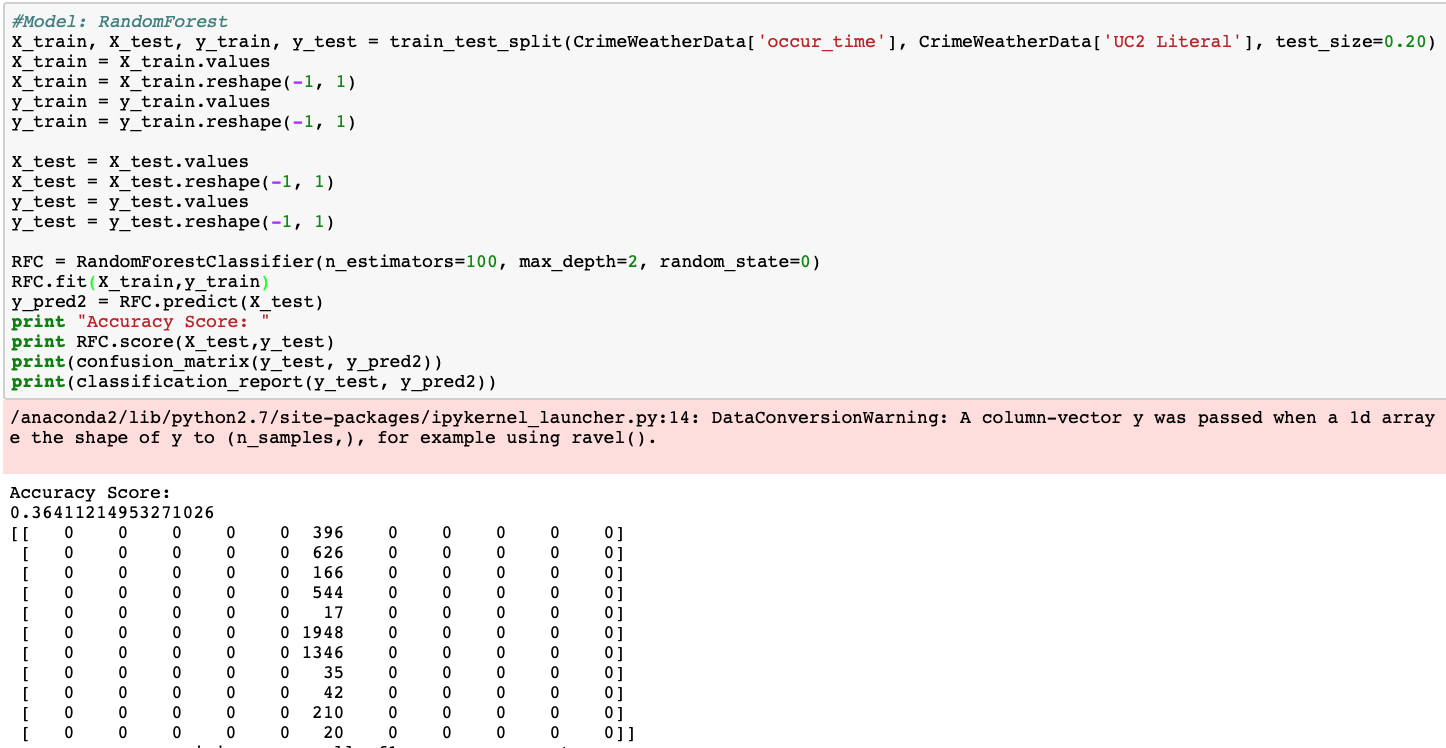


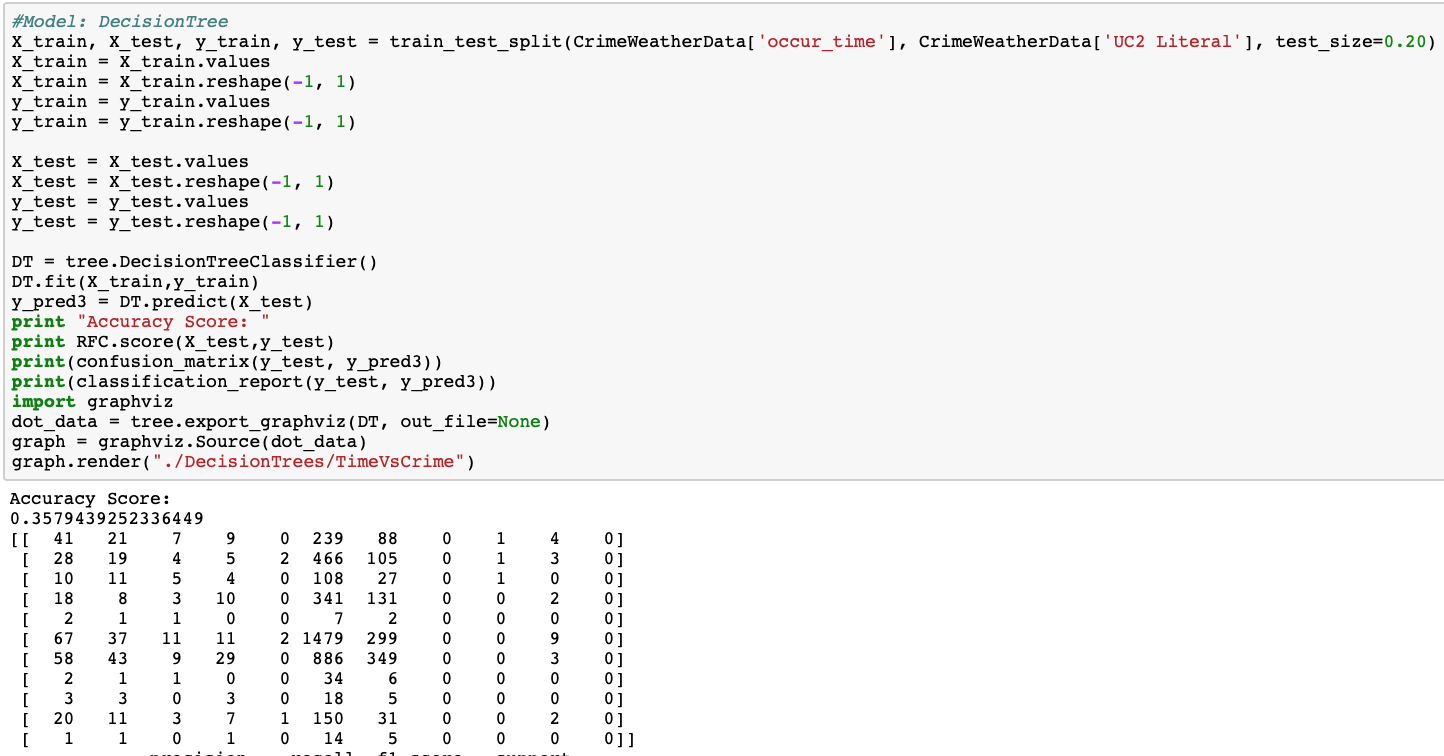
Figure 9. Random Forest 

Figure 10. Decision Trees

As seen in the figures above, time is also not an accurate predictor of crime. While one may think that certain crimes may happen at certain times of the day and could try and verify this proposition by looking at our exploratory analysis results, it turns out that this is not the case. As seen above, the highest accuracy score we achieved was 36% using a Random Forest algorithm. The other two algorithms generated lower accuracy scores with the KNN algorithm producing a low enough score that we chose to omit it from the findings all together.

Victims vs Crime Type

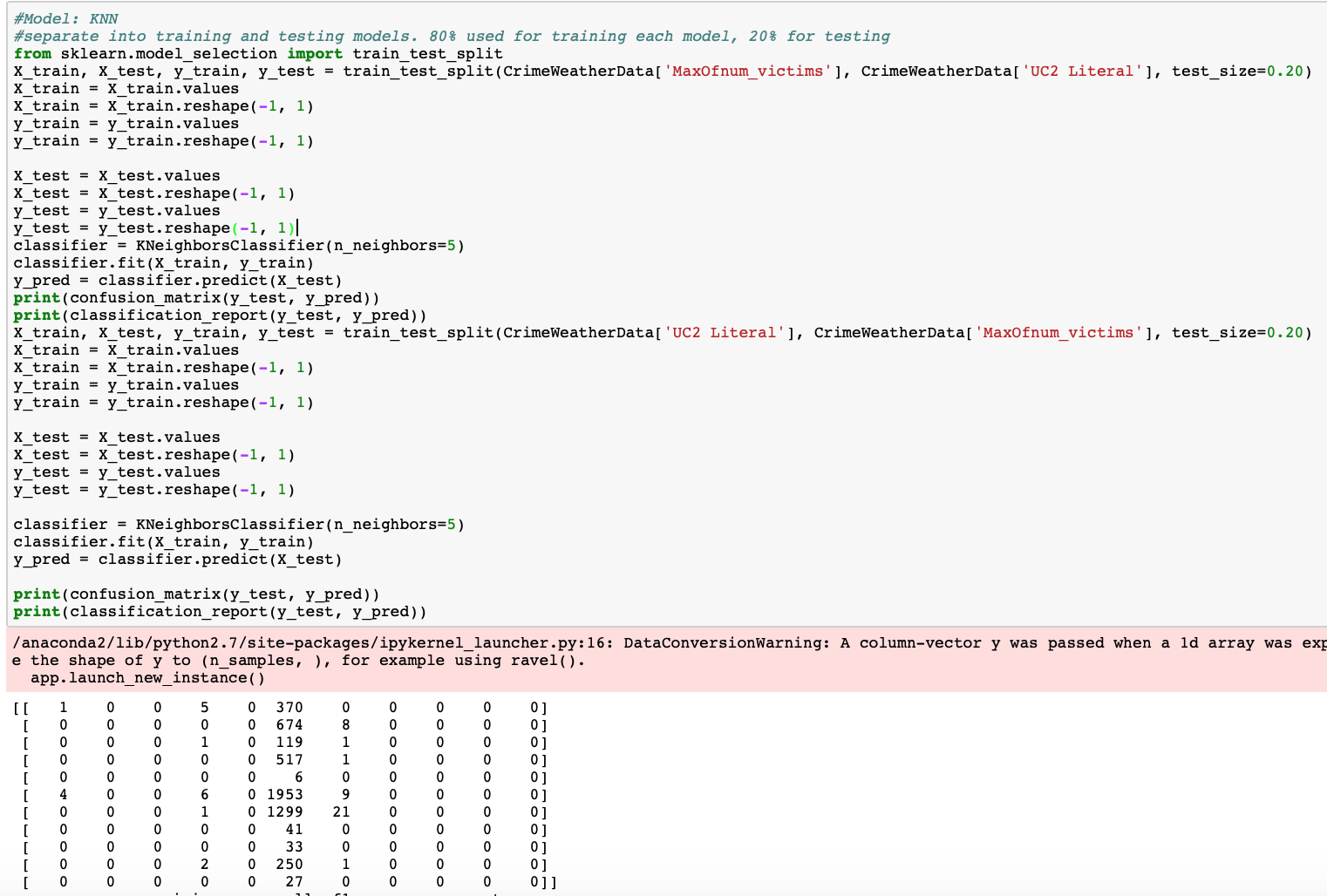


Figure 11. K-Nearest-Neighbors

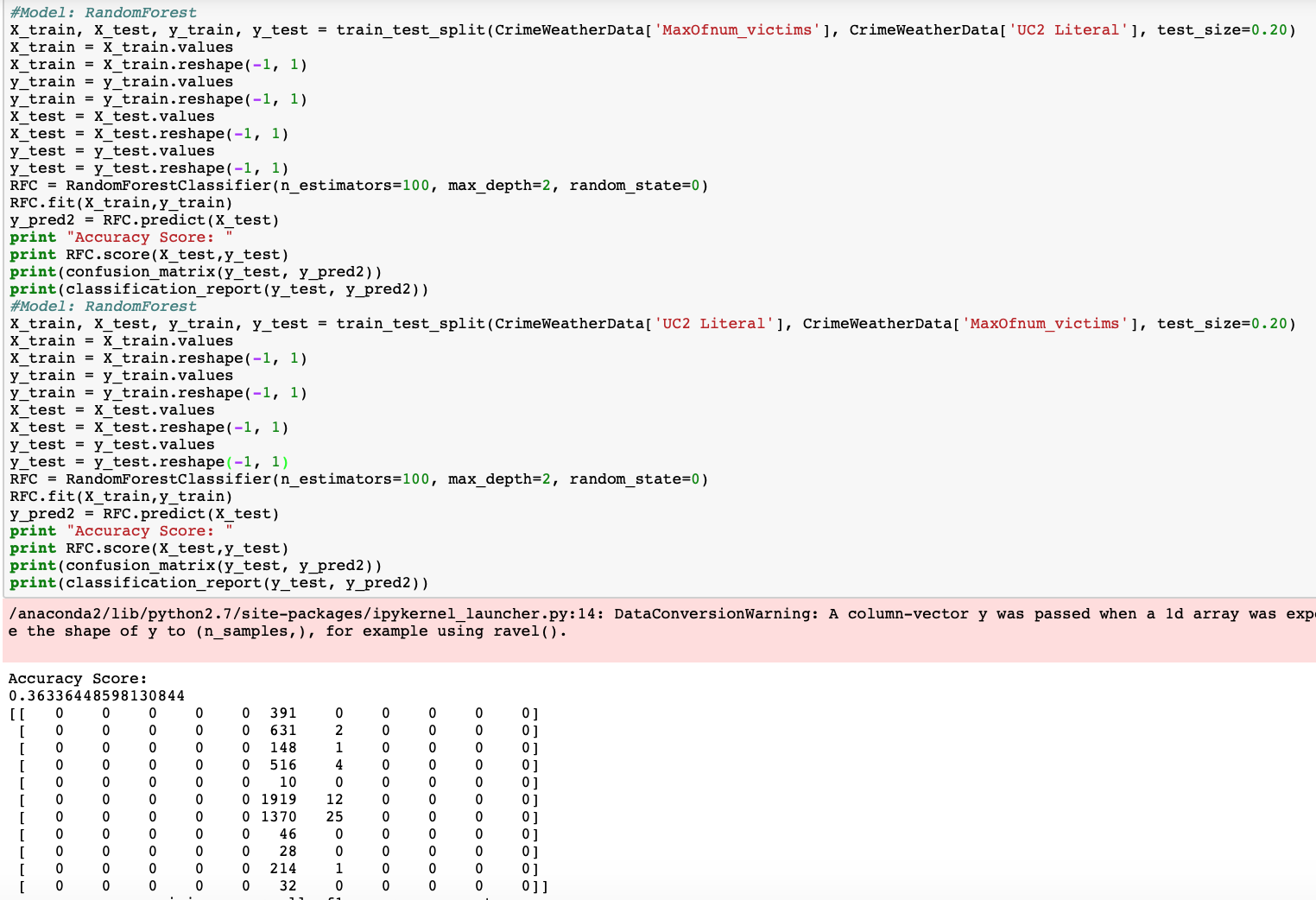


Figure 12. Random Forest

Crime Type vs Victims

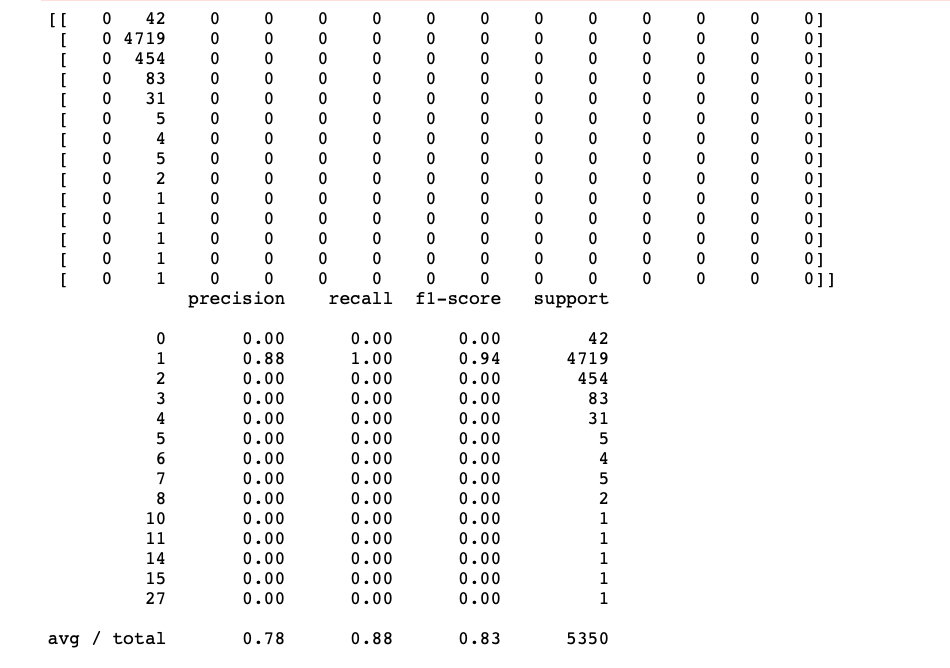


Figure 13. KNN

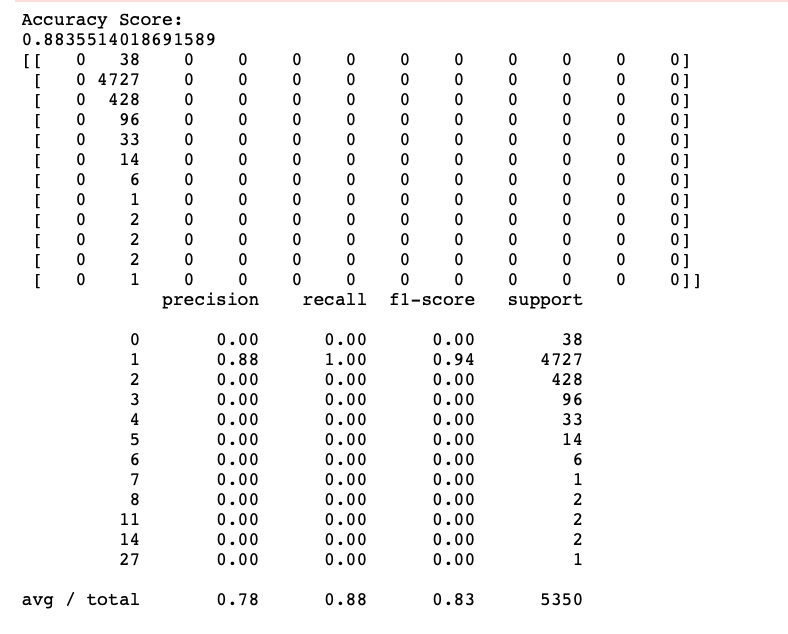


Figure 14: Random Forest

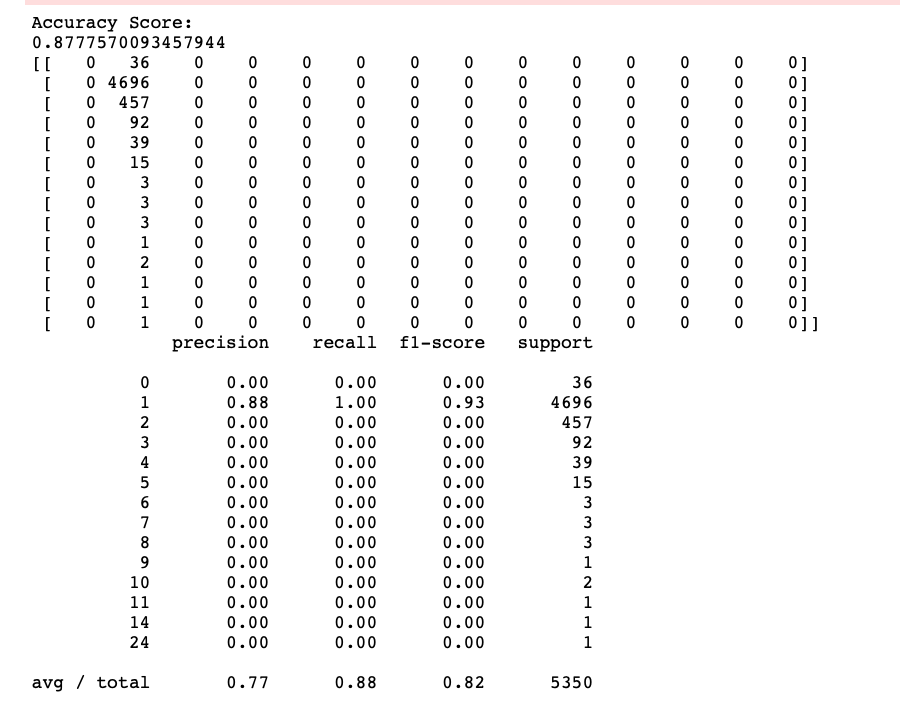


Figure 15. Decision Tree

Our most accurate prediction pair came in the form of Crime Type vs Victim Count. As can be seen in the scores and confusion matrices above, each of the three machine learning algorithms produces a result that has a recall of either 77 or 78%. This should be considered high enough recall accuracy that we can conclude that crime type is a relatively good predictor of the number of victims.

Humidity vs Crime Type

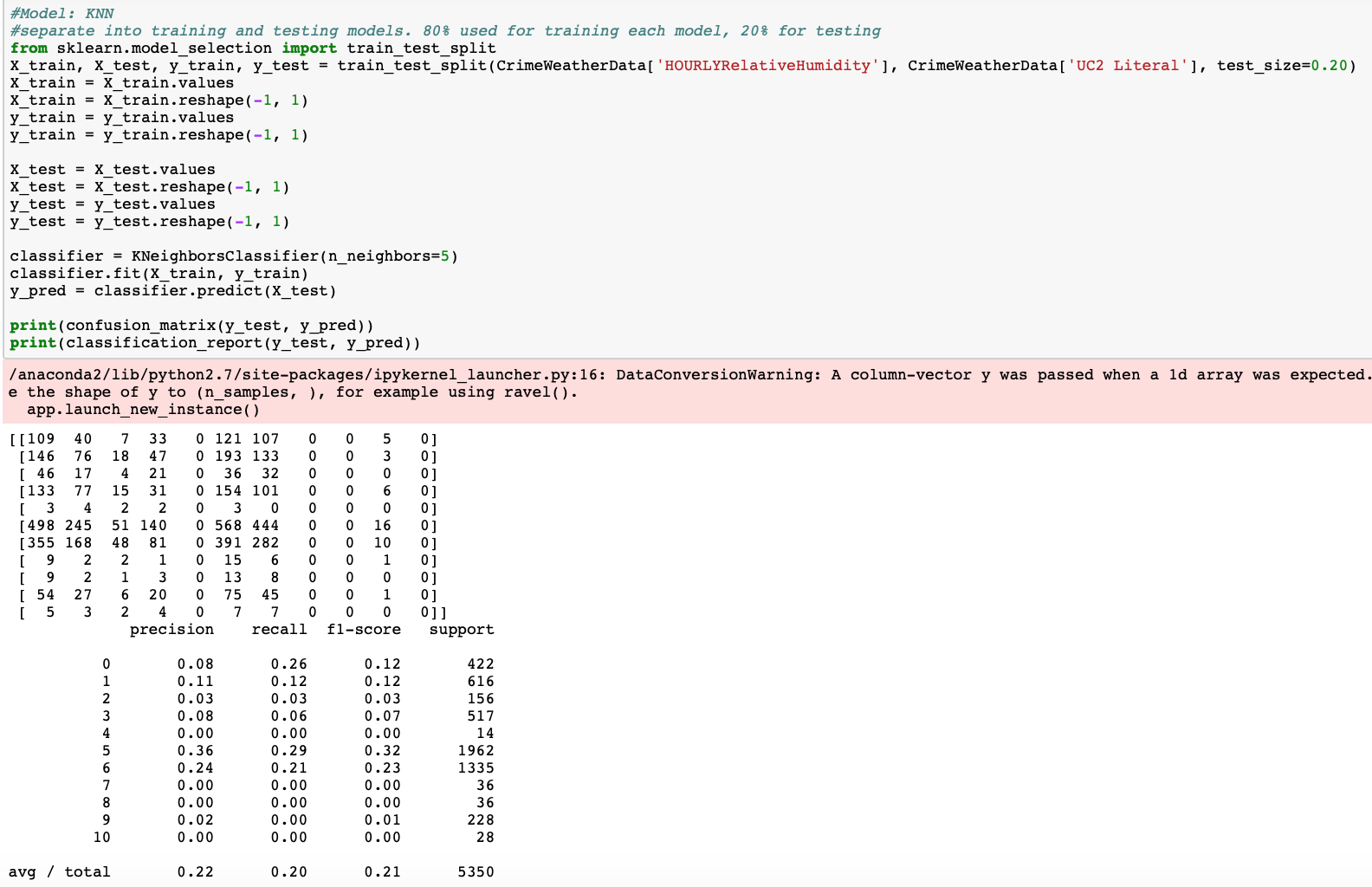


Figure 16. K-Nearest-Neighbors

As seen above, humidity is a poor indicator of crime type based on the KNN algorithm. Both the precision score and the support value are low when compared to the overall set of data.

Overall, through the included figures, one can see that when comparing temperature and crime, time and crime, and humidity and crime, it is very unreliable trying to predict crime with these attributes. However, you get higher accuracy when trying to predict number of victims based on the type of crime.

**Decision Tree Results**

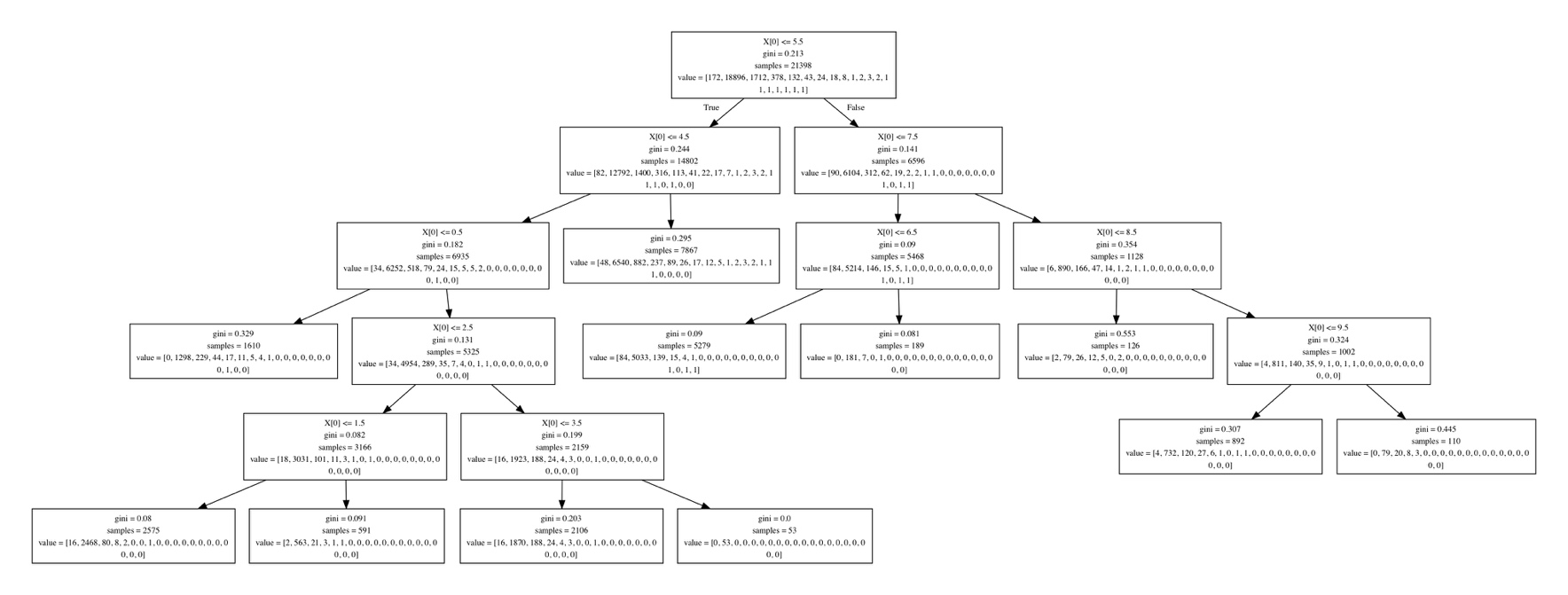


Figure 14. Decision Tree split on Crime Type

The root node at X[0] is the numeric representation of the type of crime. You can see that the tree splits along axes representing which type of crime was committed. These numbers are fairly arbitrary as they are just codes representing string literal values. In other words, “1” could represent rape or homicide. It is worth noting that once the numeric value is assigned to a crime type, that numeric value continues to represent the crime type throughout the entire analysis process. This tree was used to predict the number of victims based on the crime type.

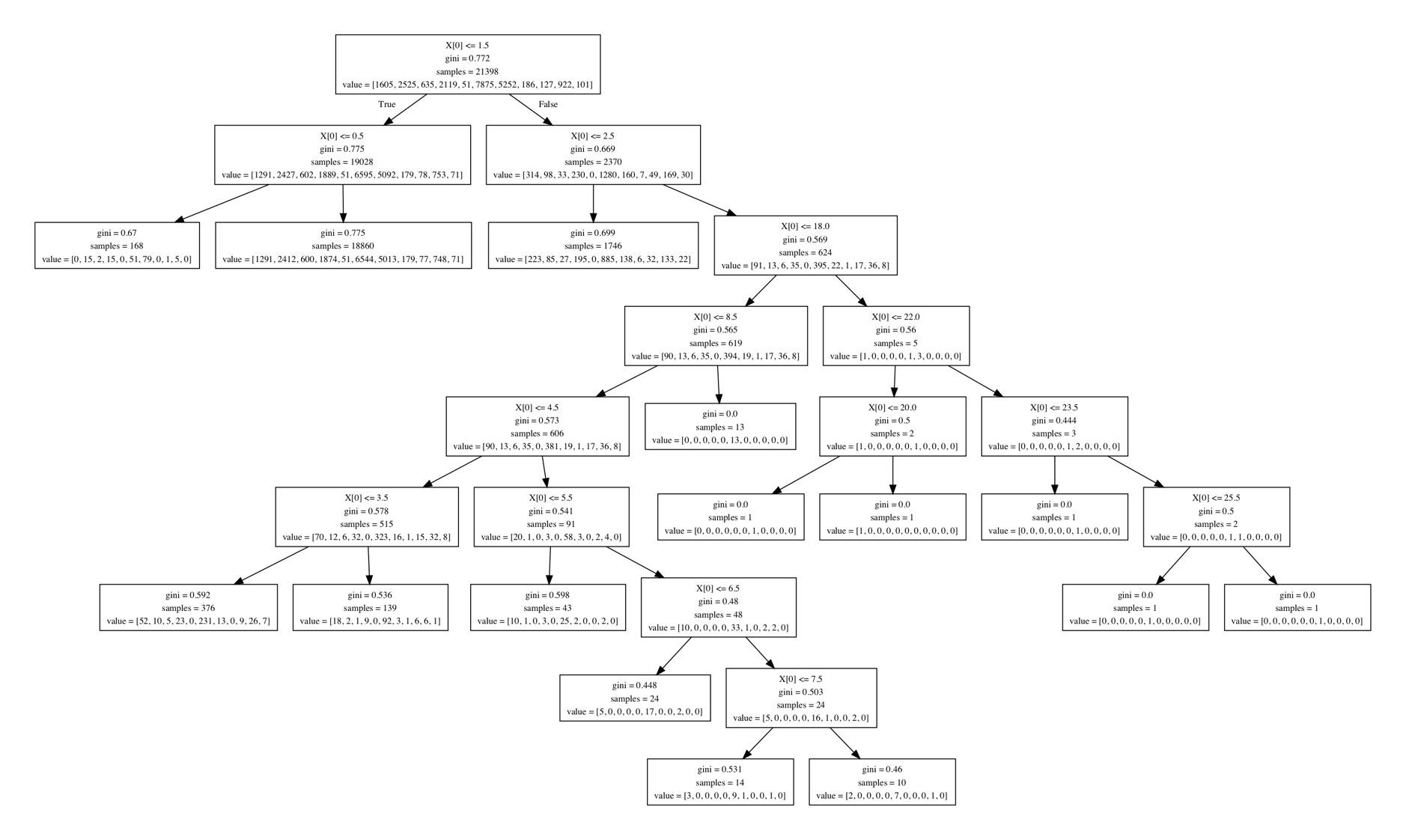


Figure 15. Decision tree of //what.

The root node at X[0] represents the number of victims. The tree splits at each level based on a certain victim number with the subtrees to the right containing instances with a higher victim count and subtrees on the left containing instances with a lower victim count.

Additionally, we created other decision trees in order to visualize other attribute prediction pairs. However, these trees were far too large to show clearly within a document.

**Conclusion**

Depending on the data available, we can find undeniable correlations or simply none at all. Therefore, in order to have the most accurate results possible, we searched for as much data on crime and weather in Atlanta as possible. However, despite our best efforts, we could only salvage data from the entire year of 2017 rather than a sequence of years prior, due to the fact that some data was kept private, was too old, or was simply negligible. There are many measures that could have potentially improved our study, but we used the data attainable in the best way possible.

In order to search for potential correlations, we ran numerous tests between different attributes including temperature and crime rate/type, time of day and crime rate/type, humidity and crime rate/type, and victim count and crime rate/type. Both contrary and predictable compared to other studies prior to ours, our tests were largely inconclusive. At the end of our study, we did conclude that while weather patterns do not seem to lead to a conclusion about type of crime or victim count, we can draw a prediction of how many victims are part of each crime, but conversely, we cannot determine what type of crime occurred based on the amount of victims recorded.

Sources

<http://worldpopulationreview.com/us-cities/atlanta-population/> - Atlanta’s population

<https://web.archive.org/web/20101109075330/http://www.ajc.com/opinion/how-to-create-a-711260.html> - ranking

<https://ucr.fbi.gov/ucr-statistics-their-proper-use> - FBI cautioning

<https://www.bocsar.nsw.gov.au/Documents/CJB/cjb54.pdf> - Bureau of statistics and research

<https://www.ajc.com/news/crime--law/georgia-gang-situation-getting-worse-investigators-say/hf4OxArZ9YL8UaUQ9yP6jJ/>

<https://blog.weatherops.com/how-does-the-weather-affect-crime-rates>

<https://www.chicagotribune.com/news/data/ct-crime-heat-analysis-htmlstory.html>

<https://www.nzae.org.nz/wp-content/uploads/2011/08/Horrocks_and_Menclova__The_Effects_of_Weather_on_Crime.pdf>

<https://www.researchgate.net/publication/232515288_Assault_as_a_function_of_time_and_temperature_A_moderator-variable_time-series_analysis>

<https://www.jstor.org/stable/2094589?seq=1#page_scan_tab_contents>

<https://www.nzae.org.nz/wp-content/uploads/2011/08/Horrocks_and_Menclova__The_Effects_of_Weather_on_Crime.pdf>

1. https://www.kaggle.com/priscillapun/crime-in-atlanta-2017?fbclid=IwAR2P\_xMijBLjw4UtxWwqKuoFtARy1WqCG5GqU\_n47iLV1SmLVAjGomVV9YQ#COBRA-YTD2017.csv [↑](#footnote-ref-1)
2. https://www.ncdc.noaa.gov/cdo-web/datatools/lcd [↑](#footnote-ref-2)