**Introduction**

In this project we use data mining methods to explore relationships between crime rates and weather patterns in the city of Boston. We obtained a dataset containing all reported crime incidents in Boston from mid-2015 to 2018.We began by cleansing our data and exploring each attribute to asses it’s value to our project. We then separated the data by date and zip code, and applied decision tree algorithms using a nominal “crime per day” value as class attribute. While we did not use the decision trees to predict crime rates, they provided insight into the relationship between different weather attributes and crime rates. Overall, we found that there seems to be a correlation between weather temperatures and crime rates. We support this conclusion with graphical analysis of our data.

We used a variety of tools and resources throughout our project. The main tool that we used for data analysis was RapidMiner, which is a data science software that provides machine learning and predictive analytics functionality. We obtained student licenses for RapidMiner and used online tutorials to learn how to use it to fit our needs. Rapidminer also has the ability to assist in data preprocessing and some visualizations, however, we were more comfortable using other tools. We used Excel for initial data exploration and feature selection. For data cleaning and processing we used Python’s Pandas module, which is a library written for data manipulation and analysis. Pandas allowed us to easily select and transform the features we wanted to include in our data, along with clean up missing data, and do in depth data exploration. All of our code was written and presented in Jupyter Notebook, Finally, we used Tableau for visualizations, and to get a better understand our data before modeling.

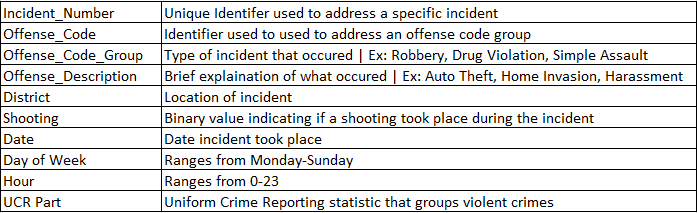
**WHY IS THIS WORK IMPORTANT/INTRODUCTION TO DATASETS?**

Using data mining applications to find relationships between crime rates and weather has the ability to maximize tax payer dollars and minimize the amount of criminal activity in a given community. The reason being is because the information that is discovered by the algorithms can be used to provide data-driven solutions to reduce crime and properly allocate a police force’s resources. For this project, our group decided to home in on crime incidents in the city of Boston, Massachusetts. It was an ideal region to focus on given that Kaggle.com contained a 56.92 MB dataset named “Boston Crime Rates” that consisted of 327,821 rows and 17 fields. Moreover, it also made sense to conduct analysis on the Boston Police Department because they were awarded $416,682,999.94 tax dollars in 2019.

**DATASET EXPLANATIONS:**

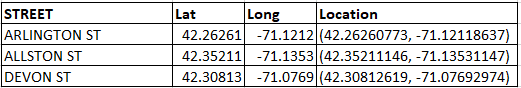
This section of our report will describe the three major datasets used in this data mining project: **Boston Crime Rates** as previously mentioned, the National Oceanic and Atmospheric Administration’s dataset named **Boston Weather** of which was 194 KB in size and contained 1,469 rows and 46 fields, and ZipAtlas’s **Median Household Income in Boston**.

**Boston Crime Rates** was filled with data points, between the years 2015 and 2018, that were essential to the overall mission of this project. Our group used the following relevant fields:



**BOSTON POLICE DISTRICTS:**

There are currently 14 police districts in Boston that correspond to one zip code. However, the dataset did not contain the respective zip codes which made it difficult for us to attach an average income per district, thus we had to manually input them. Nonetheless, after the data was entered, it allowed us to disregard the following fields:



If we know the district and zip code, there is no need to have information regarding the specific streets and geographic coordinates. Our group thought it was best to group crime by district rather than specific locations.

**OFFENSE CODE GROUP & OFFENSE DESCRIPTION:**

The values in the Offense Code Group field provide details regarding the type of crime that occurred, whereas the Offense Description briefly explains what the code group field entails. Below are the different types of crime, there are 66 totals, but only 56 are listed:



This is perhaps one of the most important attributes needed for our project. Not only were we trying to determine a relationship between crime incidents and weather, but also which type of crimes are most likely to occur/ not occur in certain temperatures. For example, in March 2017 the average temperature was 35.18 degrees Fahrenheit and the incident count for Offense Code Group: Liquor Violation and Offense Description: Liquor - Drinking in Public was 17. The incident count then increased by 67.42% to 57 in May when the average temperature was 58.9 degrees Fahrenheit. People tend not drink publicly when the temperature is colder.

**DATE, DAY OF WEEK, HOUR:**

Given that our weather data set only contained one instance per day we decided to use monthly average temperatures to determine the relationship between crime. This wouldn’t have been possible without the date field. However, this made the day of the week and hour insignificant to overall mission of this project. Nonetheless, it was still interesting to see what crimes took place on any given day and hour.

**UCR\_PART:**

It’s categorization of the crimes committed for the purpose of reporting it to the Federal Bureau of Investigation. The criminal offenses are divided into three major categories: Part I, Part II and Part III offenses. UCR Parts essentially consist of **Offense Code Group** values. This was vital to our analysis given that different crimes may occur depending on the temperature degree.

Part I:

This category of crimes is called Index Crimes, this name is used because the crimes are considered serious under and tend to be reported more reliably than others. It includes violent crimes such as Aggravated assault, forcible rape, murder and robbery; and property crimes such as arson, burglary, larceny-theft, and motor vehicle theft.

Part II:

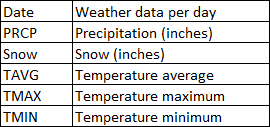
This category of crimes includes crimes such as simple assault, curfew offenses and loitering, embezzlement, forgery, and counterfeiting, disorderly conduct, driving under the influence, drug offenses, fraud, gambling, liquor offenses, offenses against the family, prostitution, public drunkenness, runaways, sex offenses, stolen property, vandalism, vagrancy, and weapon offenses.

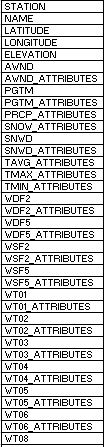
Part III:

This category includes all the other offenses.

**BOSTON WEATHER:**

Once the crime dataset was acquired, we obtained a Boston weather dataset that ranged from 2015-2019. Below are all the fields in the spreadsheet:



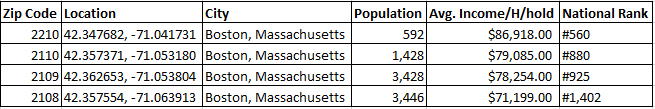


The first problem that we came across was that this spreadsheet only recorded weather data points for Boston, but not for each zip code located within the city. For that reason, we could not discover a relationship between crime rates and weather in specific districts. Instead, we had to assume that each district had the same temperature.

Initially, relevant attributes included PRCP, TAVG, TMAX, TMIN. Knowing how much it rained, the daily average/minimum/maximum temperature, and how much it snowed and rained all played a huge role in this project. Every other attribute was considered useless as they were all dependent on their “parent” attribute. For instance, the latitude and longitude columns were the same throughout because they refer to Boston.

**MEDIAN HOUSEHOLD INCOME IN BOSTON:**

Given that crime is influenced by socioeconomic environments, our group thought it was best to include the average income per household in each district. Below you will see a brief sample of the dataset we worked with:

Immediately, the following fields were removed because they failed to add value/ were redundant: Location, City, and National Rank.

**DATA CLEANING/METHODS:**

The data cleaning and preparation for this project was done using Python’s Pandas library. Pandas allows users to easily load and manipulate csv data into a Data Frame, or a two-dimensional table. Pandas has functions to automatically load csv data into the Data frame, easily select which drop missing values or select certain features and transform data from one format to another. We also used Jupyter Notebook to present and run our data. This allowed us to work in one notebook and see the results of each line of code on our dataset.

We began data cleaning for this project by exploring each attribute in the crime data set and accessing its worth. We looked at the unique values of each attribute, and their distribution in the dataset. This provided a lot of insight into our data, because each row of represented a crime that occurred. Therefore, the value counts of attributes in the dataset shows number of crimes that occurred in each value of the attribute. For example, the value counts of DAY\_OF\_WEEK = ‘Friday’ tells us the number of crimes that occurred on Fridays, within the date range of our data set. This helped us determine what would be a good attribute to keep in our data set and used for later analysis. We also did this for the weather dataset.

After we understood our data, we removed unnecessary attributes, and missing data. Most of our crime data was complete except for a few missing values in Latitude, Longitude and Street. We decided that these attributes are not important to our dataset, as we had zip code and district. We also found that there were many missing values in the “SHOOTING” column of our dataset. We found out that rather than using a True of False value for whether or not a shooting occurred, the dataset contained a “Y” is a shooting was involved in the incident, and a null value if not. We fixed this by mapping all null values to equal “N”. We reduced the data set to only include the following attributes: INCIDENT\_NUMBER', 'OFFENSE\_CODE\_GROUP', 'DISTRICT', SHOOTING,'OCCURRED\_ON\_DATE', 'YEAR', 'MONTH', 'DAY\_OF\_WEEK', 'HOUR', and 'UCR\_PART'. We then repeated the process of exploring and cleaning our data for the Weather dataset. While looking at the weather dataset we found that it contained a large number of attributes with mostly missing values. We later found that this was a similar situation to the shooting attribute in the crime dataset. Weather attributes such as “tornado” would be marked True if they were true that day, and null otherwise. We ended up not using this analysis in our data set because the true values did not seem reasonable, and it did not add value to our analysis.

After thoroughly cleaning and exploring the datasets, we combined all of our data set into one large data frame. First, we combined the crime dataset and the weather dataset on date. We had to convert the date column in each dataset to a python datetime object, and then manipulate their format so they were both month-day-year format. The crime dataset also included seconds and minutes, which we removed. After converting the data, we combined the dataset on the date column. Once we had this dataset, we used the District by zip code dataset to add Zip codes to our data. The police districts and zip codes lined up, so were able to just map each zip code the appropriate district, using panda’s merge function. Finally, we added Population and average income per zip code to our dataset, once again using panda’s merge function. The columns included in our dataset after the merge were: ( 'INCIDENT\_NUMBER', 'OFFENSE\_CODE', 'OFFENSE\_CODE\_GROUP', 'OFFENSE\_DESCRIPTION', 'DISTRICT', 'SHOOTING', 'DATE', 'YEAR', 'MONTH', 'DAY\_OF\_WEEK', 'HOUR', 'UCR\_PART', 'PRCP', 'TAVG', 'Zip Code', 'Population', 'Avg. Income/H/hold'). This combined dataset was used for analysis and results done in Tableau.

Next, we transformed some of the data attributes into nominal data so that it could be used in our decision tree. We began with the precipitation attribute. The precipitation attribute considered of continuous data that gave the rainfall of that day in inches. Most of the values in this dataset were zero, so we decided it would be best to map all zero values to “no” (as in, no it did not rain), and all non-zero values to “yes”. The new column was named “PRCPnominal” We also transformed the temperature attribute into a nominal value for the decision tree. We binned the temperature into three equal length bins, and labeled them “Low”, “medium”, and “high”. This column became known as WeatherNominal.

The next step in our data cleaning was to create the induvial dataset for each district zip code, based on the number of crimes that occurred each day. To do this, we first sorted the combined dataset by district, so we had all the crimes that in each district. Next, we counted the number of crimes that occurred each day in the district and used each day as a row of data. We then transformed this new attribute, “Number of crimes per day”, into a nominal value so that it could be used as a class attribute. Finally, we converted all of the new datasets for each district into a CSV format, so they could be used in RapidMiner to create a decision tree.

Our analysis was split into two main parts- graphical analysis and the building of the decision tree. Based on our initial exploratory analysis, we decided that a decision tree would be the best way determine patterns between weather and crime data. We began analysis by looking at the relationship between different attributes and crime rates, to get a better idea of how our weather and crime patterns could be become skewed. This also helped us get a better idea of crime in Boston. One of the major discoveries of data analysis was that there is a large difference in number of crimes by district, which also correlated with population and income level. This large difference between the districts’ crime rates is the reason that we split our decision trees up by region. Otherwise, the results would be skewed by these other factors. Another major discovery in our data analysis was trends in crime rates by month and weather each month followed the same graph when plotted (see below). This led us to believe that there is a correlation between temperature and crime rate. At this point, we wanted to see if we could find any associations between weather and crime rate. We initially tried to run a frequent item set associate rule on all of the combined data (every incident was a “itemset”, with attributes like day of the week, hour of the day, weatherNominal, PcrpNominal, etc. However, when we ran these rules we were overwhelmed by the number of frequent item sets found and could not find meaning in the top associations. We attempted to up the minimum support so that we would find less associations but found that we still could not find any meaningful insights. For example, we found that Day = Wednesday and Prcp = Yes were frequent. After attempting to find frequent item sets with no luck, we changed course and decided to create decision trees, using the number of crimes per day as a classifier, and reducing the number of features to just weather attributes. To make up for any outside factors we were possibly introducing by only looking at weather attributes with crime rates, we choose to create decision tree for each district. This was also since we had already seen a correlation between crime and temperature patterns. The decision trees allowed us to gain insight into how weather attributes (temperature and perception) played into deciding crime rate.

RapidMiner allowed us to easily make decision trees after we had separated our data. First, we imported all our data into RapidMiner, and then build process that created the decision trees. We made two process that in total, the first one created the decision tree while the second tested the accuracy of the decision trees. The first process first imported the data, and then selected all of the nominal attributes, set crime rate as the class attribute, and then uses their built-in decision tree algorithm. The second process follows the same step, but instead splits the data into 2/3 test, and 1/3 training sets. After the decision tree was built, we tested it using the training dataset. The output shows the accuracy, recall, precision, and number of TP, FP,TN,FN values.

**DATA EXPLORIATION:**

**Incidents and Average Income per District**

A screenshot of a cell phone

Description automatically generated  
For this graph, we use attributes Count (Incidents), Average income, and District.

This graph gives us insight into whether high rates of crime is co-related to income in any dimension(s). According to this graph, there appears to be some correlation between the two, but not a very strong one. District B2, which has the lowest average income in the city, also has the highest number of criminal incidents out of all the districts. Then, if one looks at district A15 and the number of incidents there, one could think that high crime rate is strongly correlated to low incomes. But districts D4 and A1 put an exception to that rule. The average incomes in those two districts are two of the highest in the City, yet the incidents there are higher than most of the districts. Our group is aware that there may be more variables that could predict crime in certain districts based off income but based off our datasets there is no strong correlation.

**Crime Rate**

A screenshot of a cell phone

Description automatically generated

We use the attributes Count (Incident), Population and District for this graph.

The reason why we chose this graph is because it tells us about the crime rate in each district, which we tried to associate with the high population or the median income in that district. This way, we could also figure out how much of a factor income is in the crime. But, as seen in the previous graph, the rule that high income is associated to low crimes was put to rest.

Drawing conclusion from the graph, clearly district B2 has the crime rate out of all the districts. And B3, A7 and A15 have one of the lowest crime rates in Boston.

**Incidents vs Month**

A flock of white map

Description automatically generated

We just use two attributes, Count(incidents) and Date.

We used this graph to understand whether there’s a trend in the incident occurrence based on the what time of year it is. The finding from this data could be related to the season, temperature, or in some extreme and unlikely cases even the local culture.

From this chart, we can observe that the crime sees a dip during the months of December, January and February; and peaking in the months of July, August, and September. Now, we could draw a conclusion from this, stating that there is a correlation between months in a year and the crime incidents. We, in our project, further concluded that temperature is what most closely relates to these incidents.

**Temperature and Crime Vs Months (split by UCR part catogerisation)**

A close up of a map

Description automatically generated

A close up of a map

Description automatically generated

For all these graphs, we use attributes Temperature, Count (Incident) and Date. We can observe a strong correlation between the Temperature and the Incident Count. The number of incidents increase when the temperature rises.

We also split the graphs by the UCR categorization. We did that to get a better picture, and to eliminate the fact that, one type of crime, being the majority, might affect the other types. To support this theory, we also computed the R-squared values for each category. R-squared (R2) for Part I, Part II, Part III are 0.48, 0.31, 0.18, respectively. These, considering real world independent sources of data, are pretty high.

**Top 20 Offenses by Types**

A close up of a logo

Description automatically generated

This graph just uses two attributes, the Offense Code Group and Count (Incident). It’s a very straight-forward graph, it displays the top 20 offenses in the city of Boston, regardless of their categories. We can observe that Motor Vehicle Accident is the most common incident.

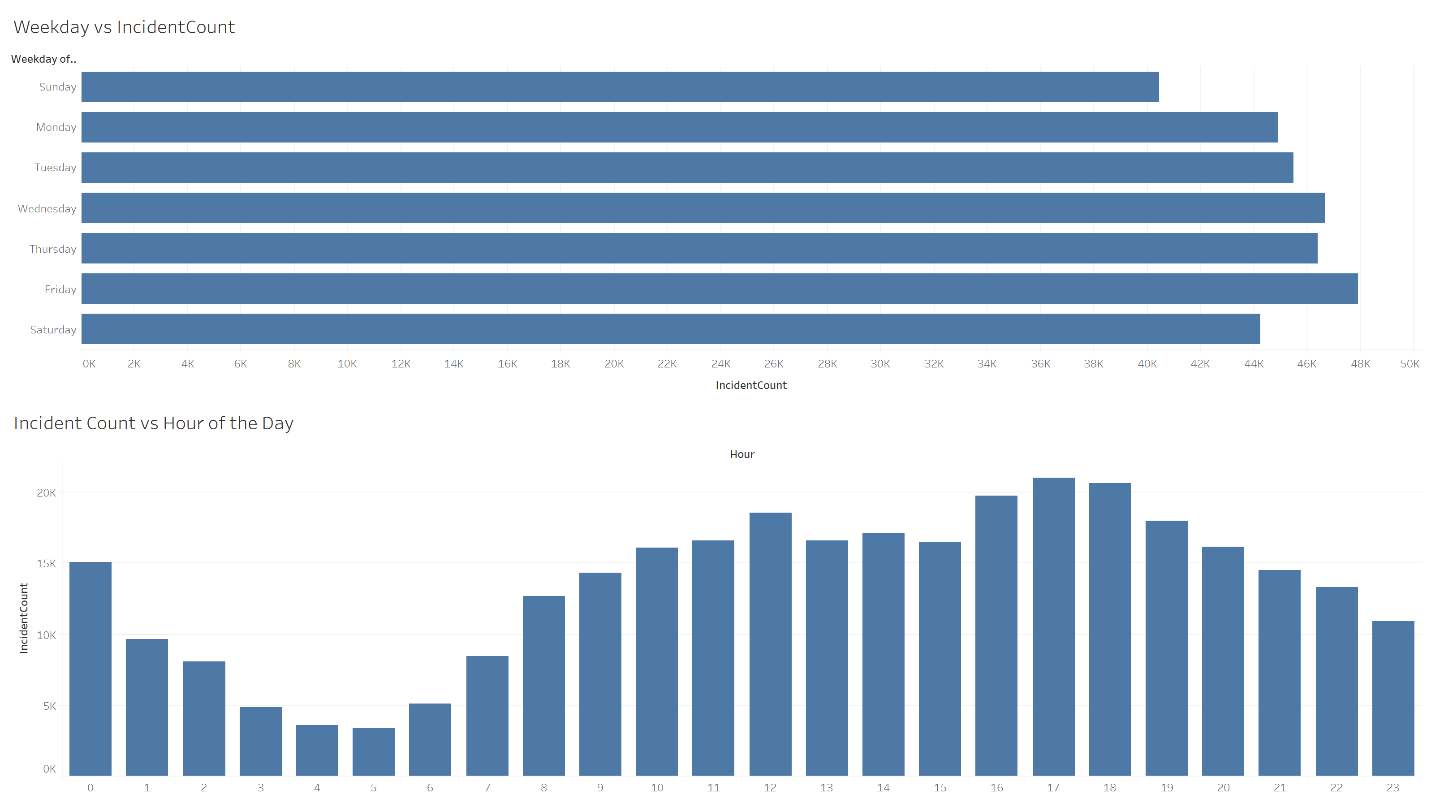
**Shooting**

A screenshot of a cell phone

Description automatically generated

This graph tells us about the number of shootings, categorized by the Offense Code Groups. This tells us which type of incidents had the greatest number of shootings.

**Number of Incidents vs Day of the Week and Hour of the Day**

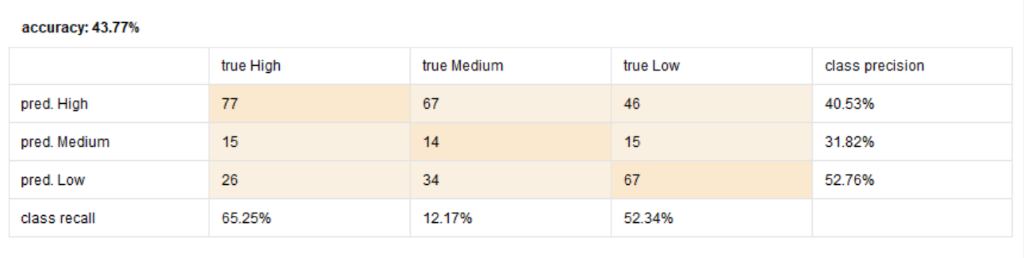
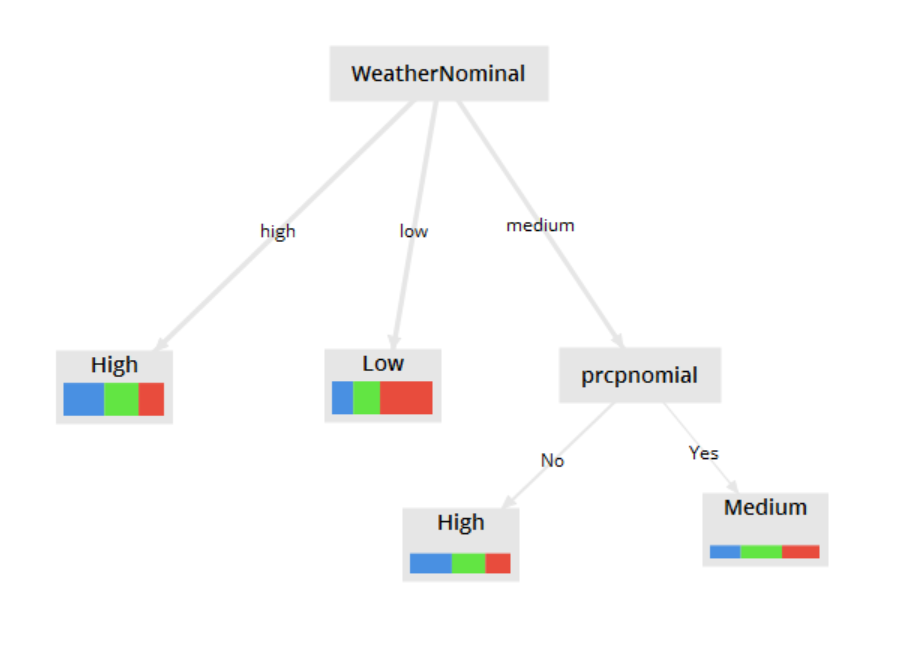


Although the Day of the Week and Hour of the Day did not play a role in our final project, it was still interesting to see when crimes were more prevalent on a specific time and day.

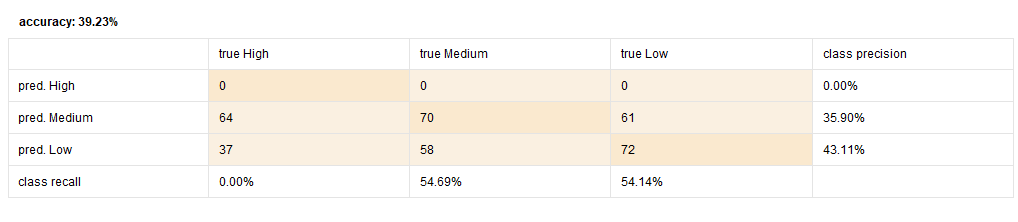
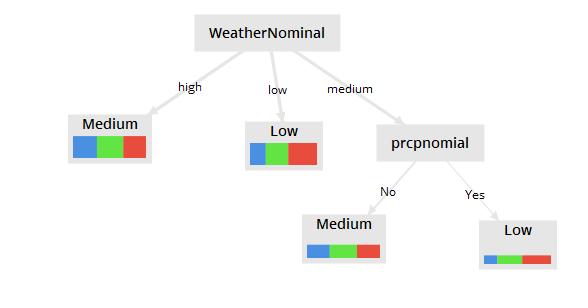
**RESULTS:**

Although more work needs to be done, we overall found that there is a correlation between weather and crimes rates. Based on the graph above, we can see that weather temperatures and crime rates follow the same pattern, indicating that temperature and crime are related. These results were to be mirrored in our decision trees. We found that high temperature was mapped to high crimes rate in most cases. Temperature was always a part a decided in our decision tree while, precipitation was only used occasionally, and typically under temperature = low, or medium. We also found that our decision trees did not have great accuracy. However, we are not using the decision tree to predict weather, but only to find associates between temperature. Each decision tree and their respective accuracies are shown below. The weatherNominal Category represent High, medium, or low temperature, while the deciding factor at the bottom of the tree is high, medium or low crime rate.

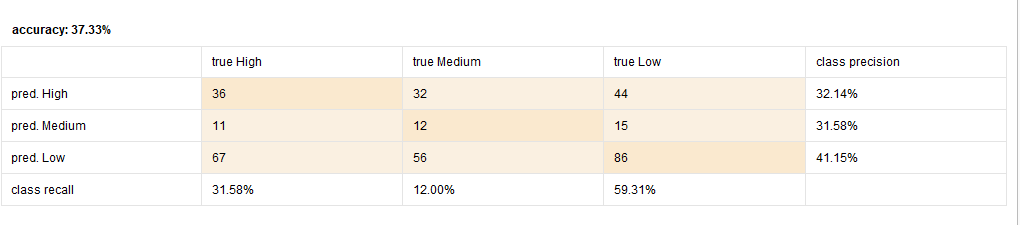
**District A1:**

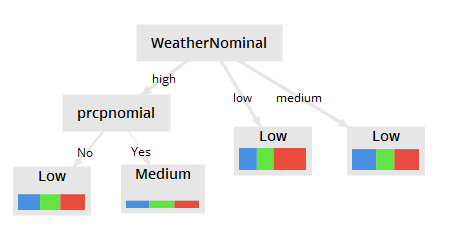


**District A7:**

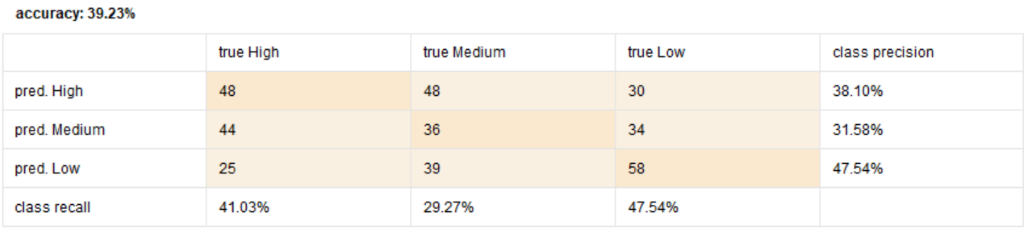


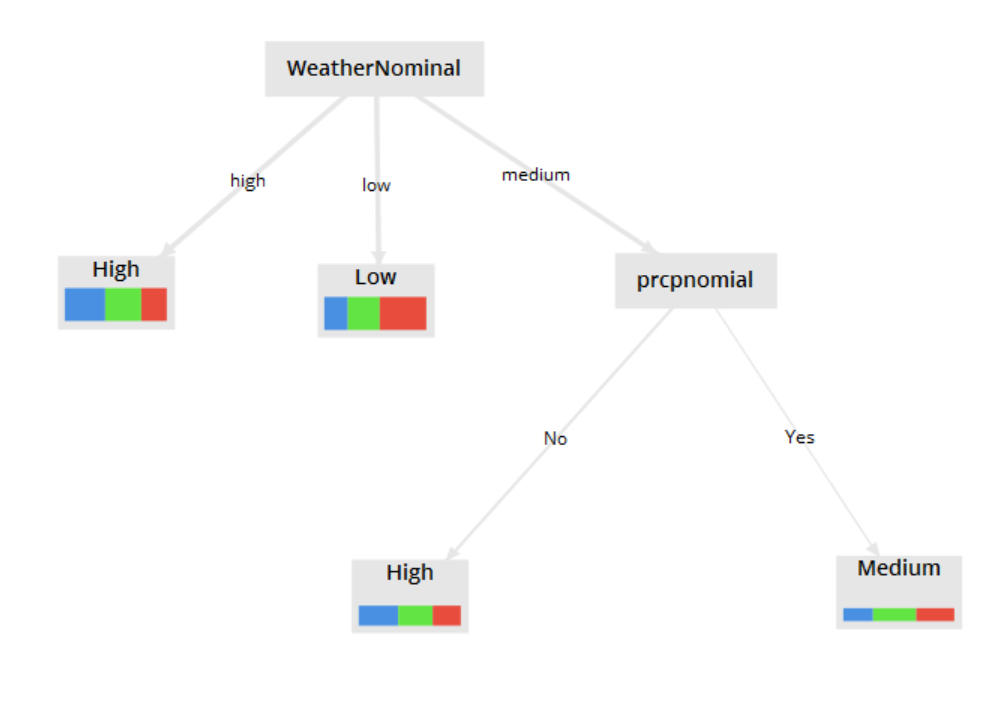
A15:



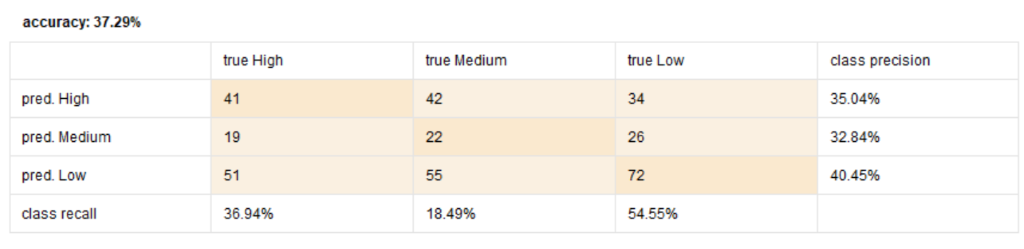
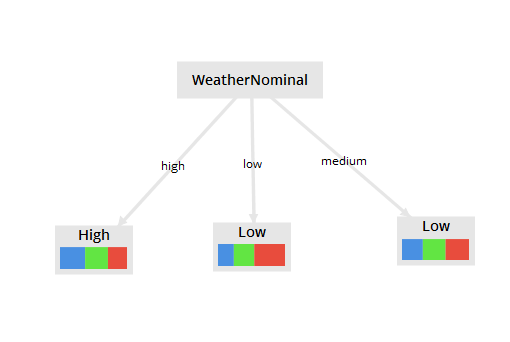


**District B2:**

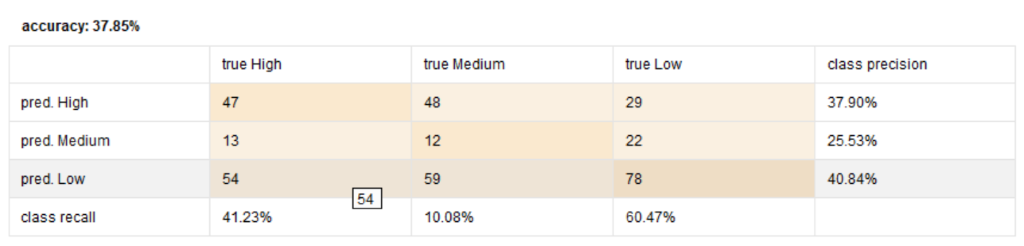
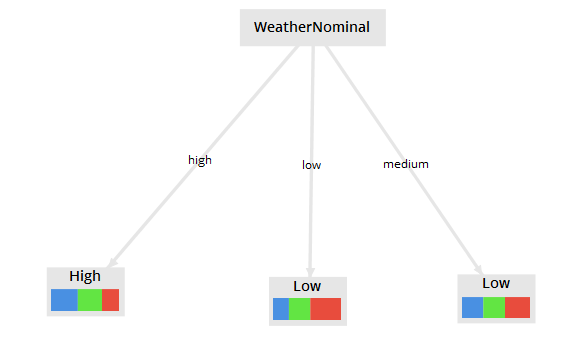




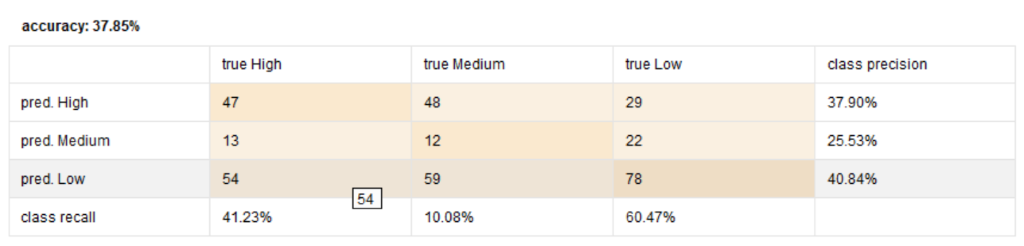
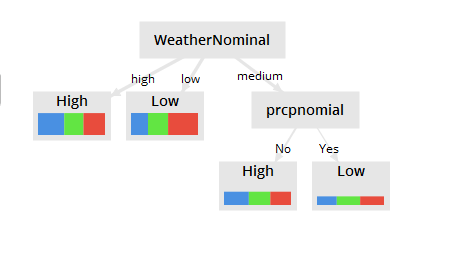
**District B3:**



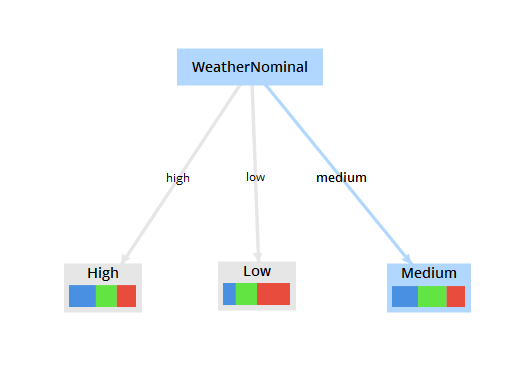
**District C6:**

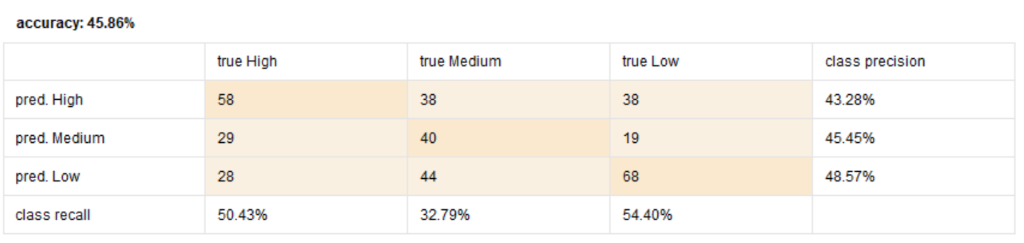


**District C11:**



**District D4:**





From these graphs we can see that high weather maps to high crime rate in six out of 8 of the districts. While low temperature is mapped to low crimes rates in seven out of the eight trees. The “middle” weather tends to get a little tricky, but this is expected as it is not an extreme. While we realize that this evidence is not conclusive to saying weather and crime rate are directly related, we believe that it provides reasonably clause, and more work should be done in this research.

**CONCLUSION:**

There appears to be strong correlations between the following in Boston, MA: Temperature and Number of Crime Incidents, High Crime with High Average Temperature. Moreover, 6 / 8 of the decision trees supported the fact that High Temperatures were to High Crime. In addition, 8 / 8 of the decision trees demonstrated that Low Temperatures yielded Low Crime. It was wonderful to have found a strong relationship between weather and crime rates, however, the next step is to provide suggestions for the Boston Police Department. In the future, we would want to gather pay roll data for every employee in each district. Ideally, it will have the following fields: Date of shift, type of shift (early morning, swing, graveyard), and district worked in. We could then use this information to figure out if BPD is allocating their human capital resources in the most optimal way possible by considering the day of the incident, hour range based off shift, and district. For example, in our crime dataset we noticed that most crimes were committed on Friday’s at 7 PM. We also discovered that in March there were 17 counts of drinking alcohol in public, with a low temperature of 35.18 degrees Fahrenheit.

BPD has Enforcement Division that implements alcohol law enforcement strategies. If we had the dataset containing the shifts work by employees, we could see how many liquor enforcement officers are working in during the month of March on Friday’s at 7 PM. If the amount seems to be too high based off the incident count in the past, the BPD could lay off some officers. Nonetheless, this is just one way of expanding on this project.

**APPENDIX**

Stephanie Bankes

* Introduction
* Data Cleaning and preparation
* Created decision trees in RapidMiner

Daniel Ramirez Jr.

* Why is this work important/introduction to datasets?
* Dataset Explanations:
* Conclusion

Prakhar Saxena

* Dataset Explanations (UCR Parts)
* Data Exploration
* Utilized Tableau Software to create all data visualizations)

**REFERENCES**

AnkurJain. (2018, October 04). Crimes in Boston. Retrieved May 26, 2019, from <https://www.kaggle.com/ankkur13/boston-crime-dataNational> Institute of Justice. (n.d.). 

Police Districts. Retrieved from <https://bpdnews.com/districts>. Boston Police Department. (n.d).

Sources of Crime Data: Uniform Crime Reports and the National Incident-Based Reporting System. Retrieved from <https://www.nij.gov/topics/crime/pages/ucr-nibrs.aspxU.S>. Federal Government,. (n.d.).

About the Uniform Crime Reporting Program. Retrieved May 26, 2019, from <https://www.bjs.gov/ucrdata/abouttheucr.cfmWalden> University. (2019, February 14).

Why National Crime Statistics Are Important. Retrieved from <https://www.waldenu.edu/online-bachelors-programs/bs-in-criminal-justice/resource/why-national-crime-statistics-are-important>

Household Income in Boston, MA by Zip Code. Retrieved from <http://www.zipatlas.com/us/ma/boston/zip-code-comparison/median-household-income.htm>. ZipAtlas.com Development Team. (n.d.).