CS105 Final Report

Written by Molly Shopper, Amalia Safer, Raphael Baysa
mshop@bu.edu, asafer@bu.edu, raph737@bu.edu
Full marked down readme and all the information can be found in:
https://github.com/Raphib737/Crimes

Introduction: In this project, we analyzed Boston crime data. Our goal was to predict at what times of day and what days of the week crimes were more likely. We also saw the most frequent kind of crimes from our data. We then used data mining techniques to predict what types of crimes were more likely to occur at certain times and on certain days. This information could be useful to police forces, who would be able to better assign resources. For example, if crimes with guns were more common at a certain time, police could be sure to wear bulletproof vests at that time. Days with more crime could have more police officers on duty.

The data did show that some days and times had higher crime than other days or times. Unfortunately, it was difficult to successfully predict which types of crimes were occurring.

Dataset description: Our project used the Boston crime data from https://data.cityofboston.gov/Public-Safety/Crime-Incident-Reports/7cdf-6fgx, which contains location, type of weapon, type of crime, day of the week, time of day, among other information. We created our database as follows: (Inc_Type_Desc varchar(255), Time varchar(255), WeaponType varchar(255), Shooting varchar(255), DayWeek varchar(255)), so we only kept the type of crime (and only kept actual crimes, as there were many types of incidents in the original database that were not relevant), the time the crime was committed, the type of weapon used for the crime, whether or not there was a shooting, and the day of the week it was committed.

Data preparation: We first created a python program that cleaned the data from the original database (cleancsv.py); it made a new database that only had the rows and columns specified above, and also got rid of the date and turned the time into 24 hour time in the DayWeek attribute. Lastly, to make mining with weka easier, we shortened time to just the hour (getting rid of minutes and seconds (floor rounding)).

Data analysis: We performed SQL queries on the database (see sqlcommands.py) to obtain basic statistics from the database, such as the number of crimes committed during the day vs at night and the number of crimes per day of the week. We also created 3 visual graphs that displayed the crimes per day, crimes per hour and types of crimes committed with the data. We also used sql commands in main.py to analyze the data so you can look at it in main.py.

Number of crimes per day of the week, in descending order:
Friday 17919
Wednesday 16457
Thursday 16434

Tuesday 16406
Monday 16301
Saturday 16186
Sunday 14592

Number of crimes committed on friday before noon vs after:
5903 | 12016

After using several very complex data mining algorithms on our training set and getting accuracy rates of under 25%, we decided to use 1R to create our model - a very simple classification algorithm. We chose 1R because we were worried about the complex algorithms overfitting the model to the training set and the accuracy rate was very similar. 1R only uses one attribute to classify, so the chance of overfitting is much lower.

The goal of our model was to predict what type of crime was occurring based on time of day and day of the week. When we ran 1R, time of day was a much better predictor of type of crime than day of the week. If day of the week was the decision attribute, no matter what the day of the week was, the model predicted that the type of crime was 'OTHER LARCENY' - the most common type of crime. When time of day was the decision attribute, the model was more divisive:
TIME:

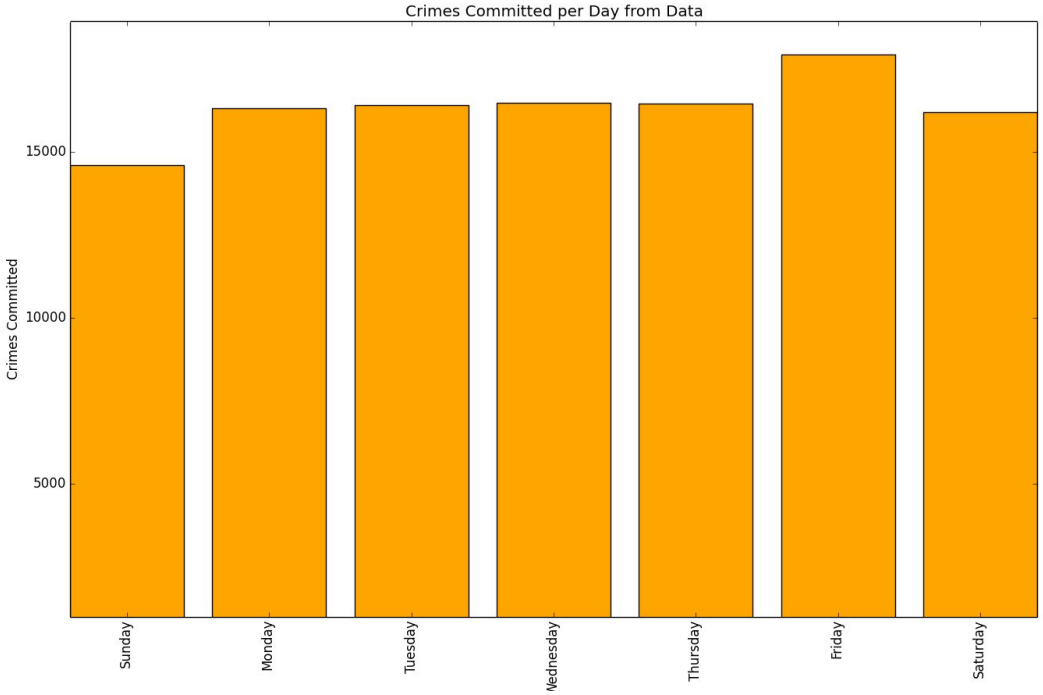
```
< 2.5    -> SIMPLE ASSAULT
< 6.5    -> VANDALISM
< 7.5    -> DRUG CHARGES
< 20.5   -> OTHER LARCENY
>= 20.5 -> VANDALISM
```

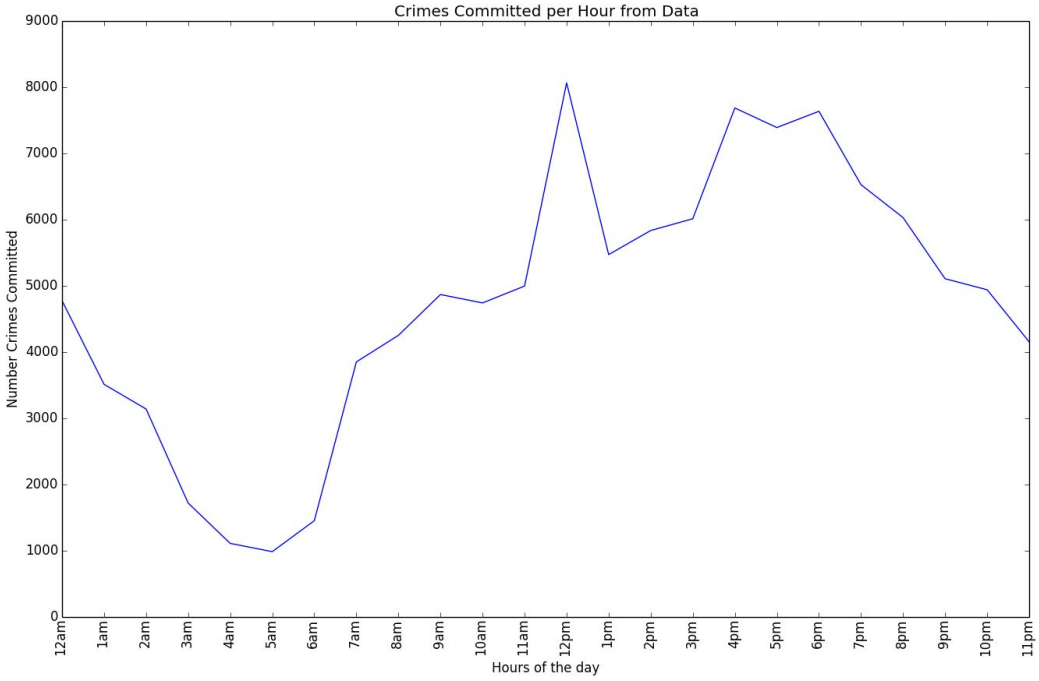
The accuracy of this second model was only 24%.

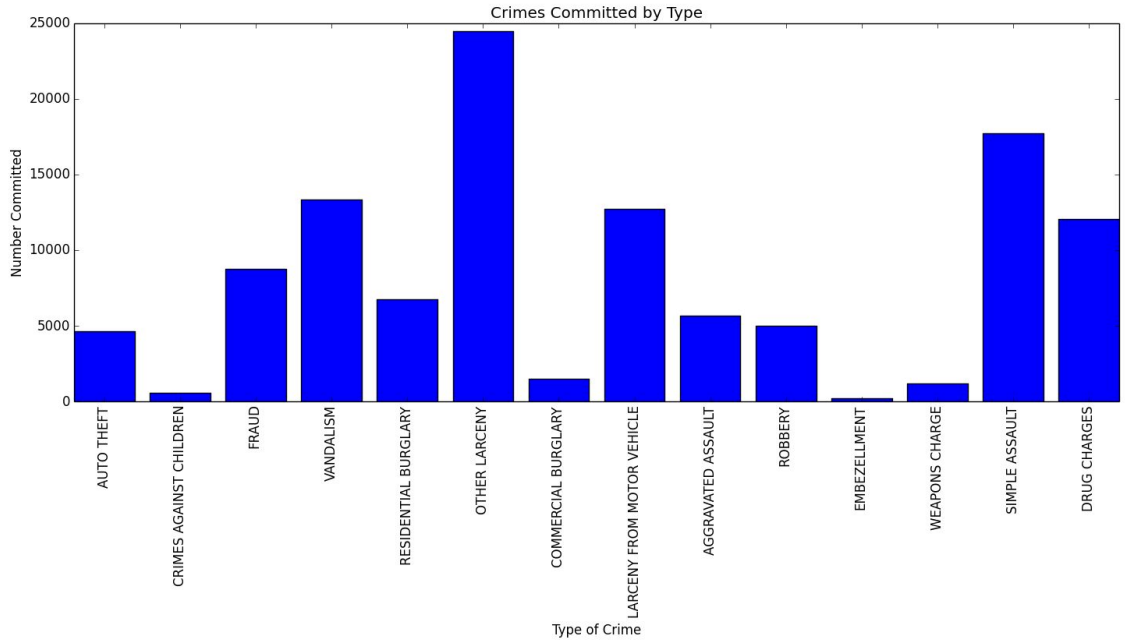
Results: When the model was used on the test data, the accuracy was 23%. This value is very similar to the accuracy of the model when it was used on the training set, which means the model generalizes relatively well. Of course, the accuracy is still incredibly low, so the model will usually be wrong anyway. This tells us that time of day is probably not the best indicator for what type of crime is occurring. The results from the SQL queries show that more crimes happen later in the day, but it's difficult to predict which ones.

Earlier in the report, we mentioned that time of day is a better predictor of type of crime than day of week. If time of day turns out to be a bad predictor of type of crime, day of week is even worse. This tells us that the day of the week may not be at all related to what types of crimes occur.

We created some graphics to show how many crimes are happening at different times of day and different days of the week, as well as a graph to show which crimes happen more.







Conclusions: Unfortunately, the data mining model's accuracy rate was low, showing that it is difficult to predict what types of crimes are occurring on any given day or time. However, the sql commands and graphics show that some times and days have more crimes than others. More crimes (of all types) occur later in the day. Also, Fridays have slightly more crime than other days of the week, and Sundays have less. Police forces can use this information to best utilize their staff - more officers should be out on Fridays than on Sundays. As far as the type of crime occurring, they'll just have to be ready for anything!

Appendix: cleancsv.py, sqlcommands.py, crime.db, crime.csv, main.py(contains all the code from other files detailed and documented well). Other Directories:

All the information can be found in https://github.com/Raphib737/Crimes