1. Your problem statement

Identify if a crime involves shooting or not in Boston given the crime type and location, and help law enforcement to identify the situation and act accordingly.

1. Articulation of value

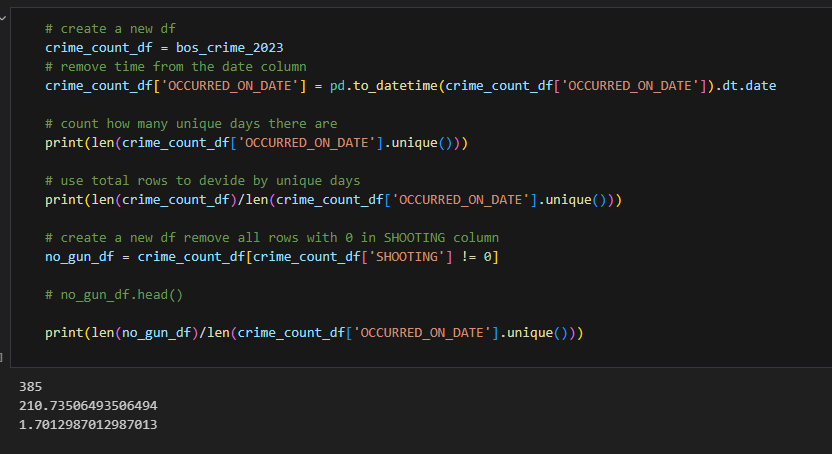
By using this ML model, law enforcements will know how many units they will need to sent to that crime scene and what equipment they need to bring. This will help to reduce the amount of units that will be needed if the model predicted that there will be no shooting involves in the situation. When the model predicts that there will be a gun shooting happened during the crime, they can send more units to have a better control of the situation. This model will help officers feel safer during their field services and could possibly reduce the injury rate and insurance cost for all law enforcement officers.

The data itself (not ML model) can also help produce a heatmap of crime that polices department can use it to develop a better patrol routine and reduce the officers needed for daily patrol. On the other hand, the city planner can also use the heatmap to allocate resources more efficiently and improve public safety.

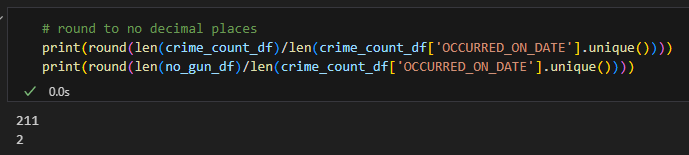
1. Calculation of the potential economic value

According to ziprecruiter, “As of Jan 14, 2024, the average annual pay for an Entry Level Police Officer in Massachusetts is $62,961 a year. Just in case you need a simple salary calculator, that works out to be approximately $30.27 an hour. This is the equivalent of $1,210/week or $5,246/month.”[1]

And according to the policedatainitiative.org, there are 2,139 officers right now hired by Boston.[2]

By using the data(give out by Boston government officially, which I will use for this project):  


After around to 0 digits:



There are average of 211 crimes happened in Boston areas every day in 2023 but only 2 in these have a gun shooting involved.

The predicted saving from this model will be calculated since there will be reducing numbers of polices since less officers will be needed if a crime is determined that it will not involves a gun shooting situation.

By deploying this model, assuming about 25% law enforcement force will not be needed, and this will the city of Boston about 35,000,000 USD a year.

This number is just a estimate, and might be higher since we can not count the insurance fee drop as law enforcements are working with a safer environment as the model is deployed.

1. Project plan: Build a 13-week plan. Identify the steps, identify the weeks, and what you will do in each step (you can look at the syllabus to build this plan)

Week 1: Create a research problem and search for a dataset, create a plan for this problem. (Which is already done in this document)

Week 2-Week3: Data Preprocessing: Clean and prepare the data for analysis. This includes handling missing values, encoding categorical variables, and possibly creating new features ( time of the day, day of the week) Creating new features will mostly done in week 3

Week 4 -5: Data Analysis: Analyze the data to understand patterns and trends in crime incidents. This might involve visualizing crime frequencies by type, district, time. A heatmap as I mentioned in question 2 will most likely to be done in this step.

Week 6,7,8: Feature selection and Model development, The reason why I Put these two steps together is I believe these two steps are using same think process and I do not want to cut them so I will be thinking straight down these two steps. During these two steps, I will need to identify the most relevant features and train the model to predict if there will be shooting or not.

Week 9,10,11,12: Evaluation, Debugging, and use the model's outputs to provide actionable insights and recommendations for crime prevention and resource allocation. (it also gives me time to start over if something is wrong and I cannot fix it).

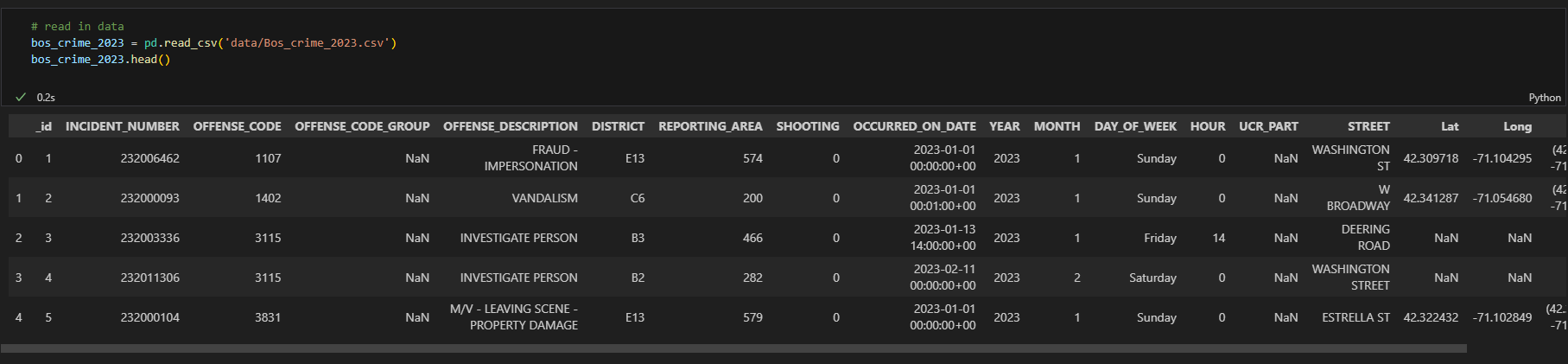
Week 13: prepare for submission.

1. Discuss the dataset: describe the dataset you've found, where you've found it from, and how will this dataset solve the problem statement?

The dataset I found is from Boston.gov ‘s analyze Boston project, the link is

<https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system/resource/b973d8cb-eeb2-4e7e-99da-c92938efc9c0>

It has 18 columns and 81,133 rows including all crimes happened in Boston from 2023-1-1 to today.



Here is the screenshot of the data frame’s head

It contains:

1. Incident Number: A unique identifier for each crime incident.
2. Offense Code: Numerical code representing the type of crime.
3. Offense Code Group and Description: Categorization and description of the crime.
4. District: Police district where the incident occurred.
5. Reporting Area: A more specific area within the district.
6. Shooting: Indicator if the incident involved a shooting.
7. Occurred On Date: Date and time when the incident occurred.
8. Year and Month: Year and month of the incident.
9. Street: Street name where the incident occurred.
10. Latitude and Longitude: Geographical coordinates of the incident.

And other columns that might be used after I finished feature engineering.

1. Identify the type of modeling that will solve the problem you identified: supervised or unsupervised. If supervised, is this a classification or regression problem? If it is classification, is it binary or multi class? If it is unsupervised what type of unsupervised learning is it?

It will be a supervised binary classification problem. The only thing I am worry about is that this is a 200 vs 2 binary problem (200 + no shooting, 2 with shooting), I believe there might be a overfitting problems that makes the model “believes” that there will be no shooting for all crimes.