Sophia Mir
DATS 6103 Data Mining
Professor Amir Jafari
06/24/2022

FINAL PROJECT REPORT

INITIAL PROPOSAL

For the data mining project, I will be using a crime dataset consisting of primary data collected by DC’s Metropolitan Police Department (source: https://crimecards.dc.gov). It consists of over 55,000 observations and the incident data is dated from January 2020 to May 2022. Some of the questions I will address during the EDA are:

1. Is there a relationship between time of day and incidences of crime?
2. Is there a relationship between time of year and incidences of crime?
3. Is there a relationship between time of day and type of crime?
4. Is there a relationship between geographical area and type of offense?
5. Which crimes happen most frequently in each ward?
6. Is there a relationship between property crime and violent crime?
7. What types of crimes happen during the day/evening/midnight?
8. Is there a relationship between the offense type and the method?

For the modeling aspect, I will be creating a classification model to choose between the two categories of crime – property crime and violent crime.

INTRODUCTION

As a research topic criminal activity is ripe with curious minds trying to figure out who is at fault. Particularly in the field of economics, as researchers and policymakers want to know how to curtail crime rates – is it through creating more jobs, by increasing incarceration time or fines, or through developing more community supports? As an applied economist, I was curious of the same mechanism and wanted to explore the data regarding criminal activity in Washington D.C. I wanted to see what I could unearth using the tools we have learned in this class. The dataset comes from the Metropolitan Police Department (MPD) in Washington DC (DC). It primarily consists of data on the types of crimes, the times they occurred, and the places they occurred. Other than that, the dataset has values for police-specific metrics which may or may not turn out to be useful as not all of them as interpretable by laymen. Using that dataset, I will be able to create a classification model to predict whether a crime that occurred was a property crime or a violent crime.

The work on this project will consist of data exploration, preprocessing of the data to clean it and make it usable, exploratory data analysis, modeling, and model evaluation.

DATA LIMITATIONS AND SOURCE

To examine the crime rates in DC I used incident data gathered by the Metropolitan Police Department which I accessed from their website. The data had information on over 55,500 recorded offenses which were reported to, and investigated by, the MPD. The dataset had incidents ranging from January 1st, 2020, through May 31st, 2022. There were a total of 29 variables in the dataset which consisted of mostly qualitative variables. The dataset classified all crimes into two overarching groups - property crimes and violent crimes. These were further divided into nine subcategories. All arson, motor vehicle theft, theft from auto, other theft, and burglary incidents are classified as property crimes. All murder, assault, rape, and robbery incidents are classified as violent crimes.

While it did have truly interesting data, there were a couple of limitations with using this dataset. The first one was that it is not truly representative of crime in the District of Columbia as this crime database system relies on self-report which can lead to underrepresentation of certain crimes and overrepresentation of others. The other limitation was that because this initial dataset consisted of primarily categorical data, it was difficult to get any meaningful summary statistics on it during the exploratory data analysis. As we have already established that I did not believe this dataset to include the entirety of criminal activity occurring in the DC area, I did not assume that the data was normally distributed despite the large number of incidences recorded.

DATA DESCRIPTION AND SMART QUESTIONS

For the SMART questions I explored these ideas:

- Is there a relationship between the time of day and the types of crimes committed?
- Is there a relationship between the geographical area and the types of crimes committed?
- Does a correlation exist between the two broad categories of crimes – property crimes and violent crimes?
- Is there a relationship between the weapon used and the type of crime committed?

While exploring the dataset, I chose to convert all the data types to categorical data. Although there were integer-based values in the dataset like latitudes and longitudes and census tract numbers in the dataset, I chose to convert them to discrete values as I was using them specifically for identification purposes. To see if there were any geographical ties to crime intensity. The 'offensegroup' variable contains the two overarching categories - property or violent crimes and it's the variable I will be as my target variable. The 'OFFENSE' variable is subcategories of crimes, including theft/other, theft f/auto, sex abuse, robbery, motor vehicle theft, homicide, burglary, assault w/dangerous weapon, and arson. 'METHOD' has three values, which are gun, knife, and other. The 'SHIFT' is a time identifier, which divides the whole day into three parts - day (8 AM to 4 PM), evening (4 PM to 12 AM), and midnight (12 AM to 8 AM). The division of time is variable because of shift rotations and difference in seasons but the given times are a good rough approximate. The 'REPORT_DAT' is the time and day the crime was reported to the police department. Similarly, 'YEAR' is the year the crime was reported. 'WARD' is a location identifier to indicate the specific area the crime happened. Washington DC is divided up into eight wards.

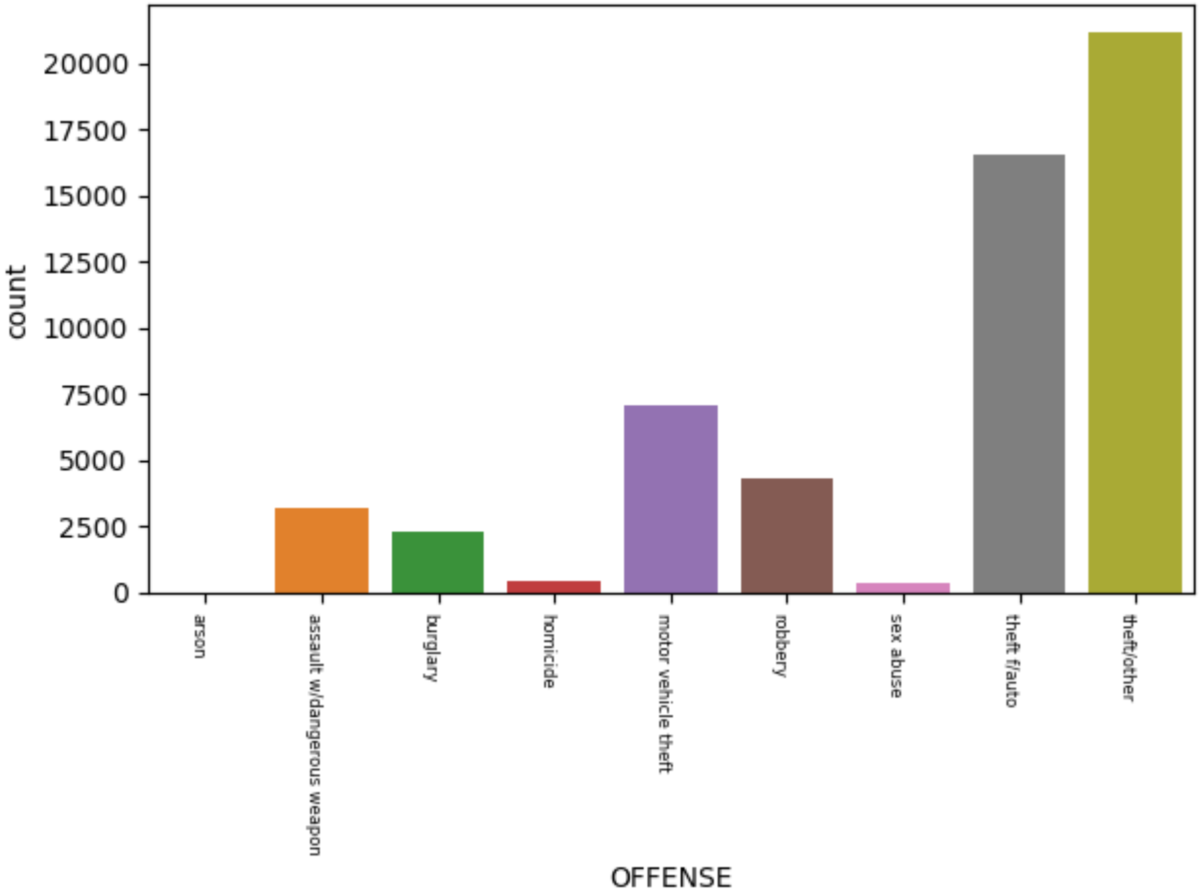
PREPROCESSING

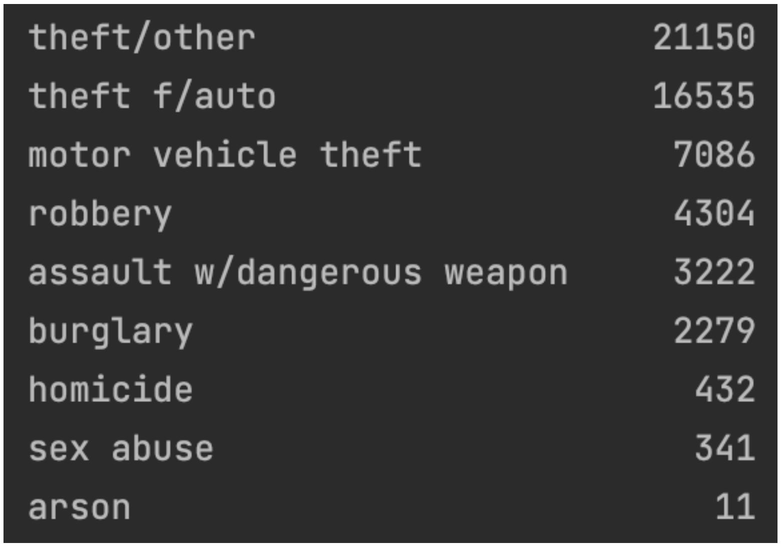
After importing the data, I looked at its structure and checked for missing values. I also looked at the column names and values. I found a column with over 45000 missing values, and two features with repetitive data. I decided to drop those along with three others which had data that was uninterpretable ('ucr-rank') and the time that the report was filed and when it was closed. UCR Rank stands for uniform crime reporting rank, but I couldn't find out what that metric was based on which is why I decided to drop that variable. The variables for the start date and end date of the reports didn't affect where or when the crimes were happening and spoke more to the bureaucracy and efficiency of the police system. As that didn't pertain to the model I was trying to create, I dropped those features.

After the initial look at the data and removing features, I decided to change all the data types to categorical data. The dataset initially included float, integer and string data but it wasn't on a continuous scale and doing any kind of arithmetic on it wouldn't yield meaningful results. After changing the data to categorical, I decided to impute missing values using SimpleImputer and setting the hyperparameter 'strategy' to 'most-frequent'. Then I decided to explore the values within the variables. I tried to use the describe function but with that many features it didn't give me the whole picture which is why I decided to use unique on the individual columns. Doing that led to learning the number of unique values within each feature such as the 8 wards, 7 districts, 46 neighborhoods that DC was divided into. I also discovered the different subcategories of crime, the different methods specified in a police report, and the years the data spanned across.

I wasn't too sure how valuable the 'REPORT_DAT' feature would be, but I thought it would be interesting to see if there was a certain month in which crime was higher. In retrospect, it would have also been interesting to look at days of the week to see if more criminal incidents happened during the week or over the weekend. To extract the month from 'REPORT_DAT', I used the string split feature in python along with the lambda function to get the numeric value for each month those incidents took place and created a new feature named 'MONTH'.

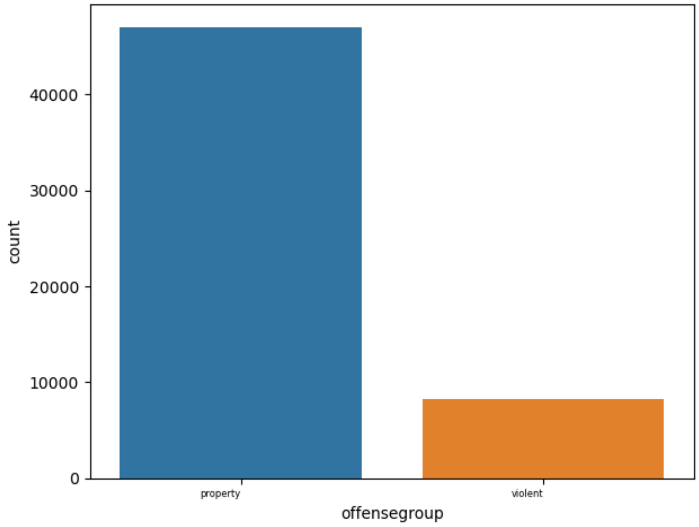
At this point I felt confident that my data was at a good level of completeness, so I proceeded to visualizing it to get an idea of the relationships which existed in the data, if at all. These are the visualizations that I created in python using seaborn, pandas, and matplotlib:

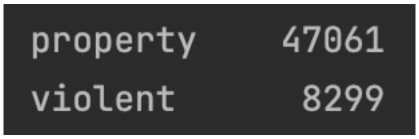




```
theft/other                    21150
theft f/auto                   16535
motor vehicle theft             7086
robbery                         4304
assault w/dangerous weapon      3222
burglary                        2279
homicide                         432
sex abuse                        341
arson                             11
```

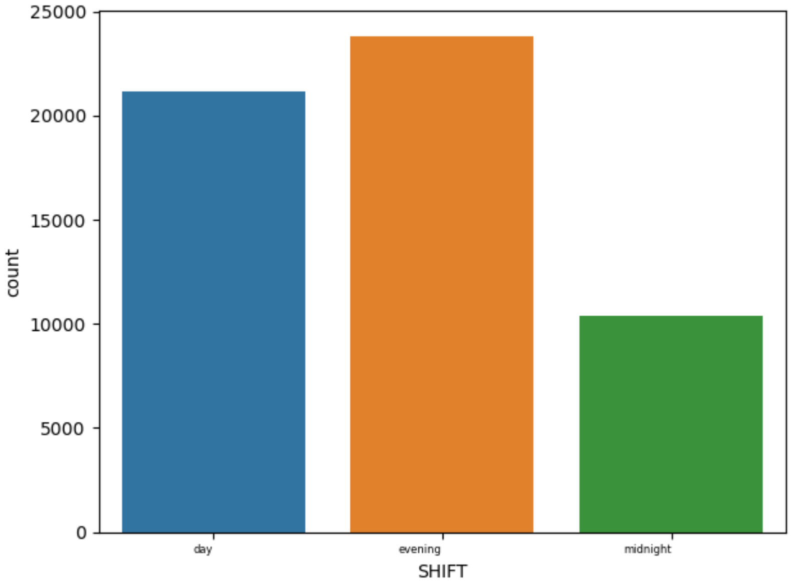
A frequency distribution of the different criminal offenses. It shows that among the different crimes occurring in DC, theft is the most frequent.

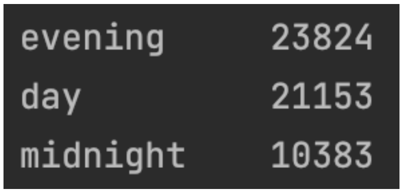




```
property    47061
violent      8299
```

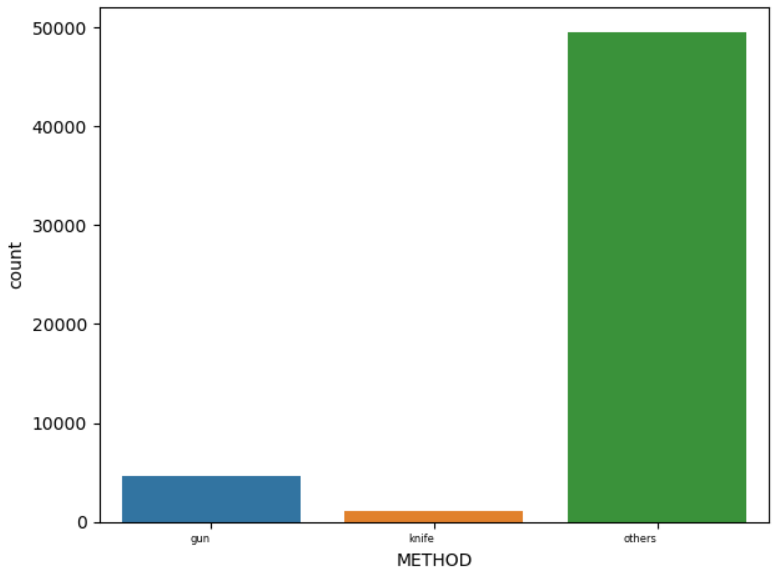
This plot shows the discrepancy between the types of crimes. Most crimes that occurred in DC in the last couple of years are classified as property crimes. I later learned that I should have accounted for the difference in my modeling and balanced the two outcomes to obtain a model with greater accuracy of prediction.

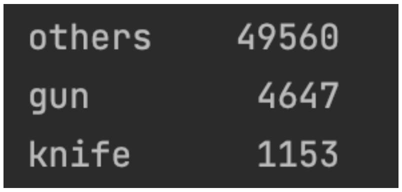




```
evening     23824
day         21153
midnight    10383
```

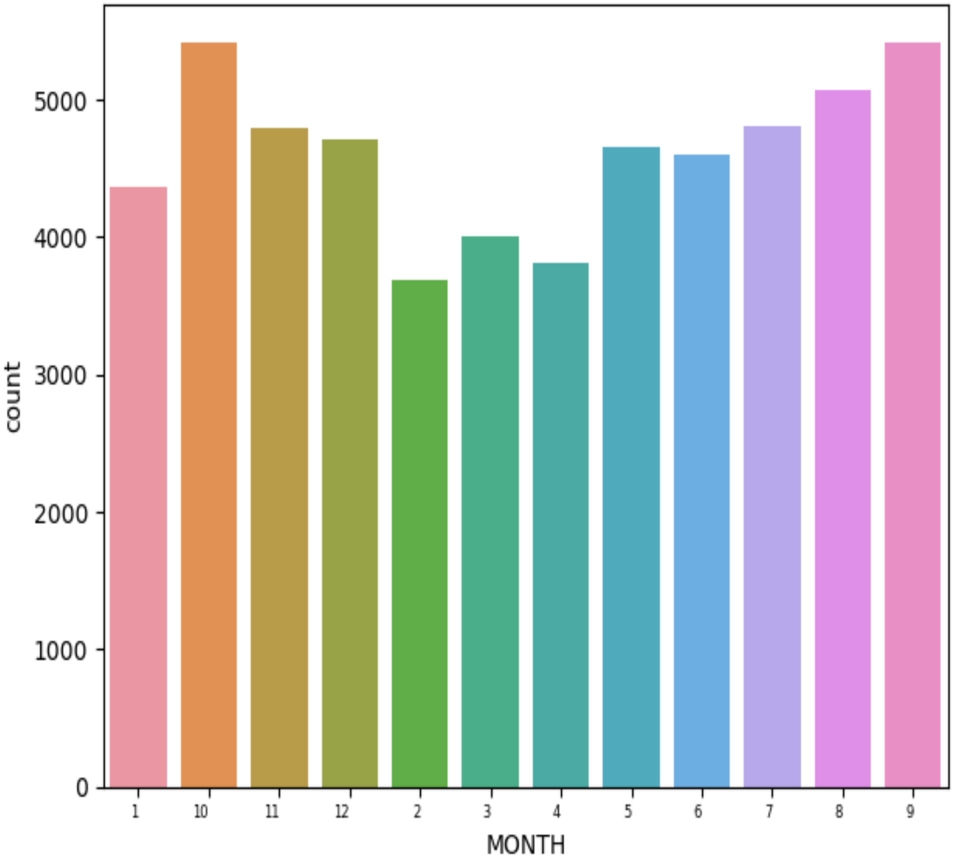
It's not too surprising to see that most crimes occur in the evening, roughly between 4 pm and 12 am. What I was surprised by was that the second most frequent division was during the day as popular media depicts crime as something that happens after dark.

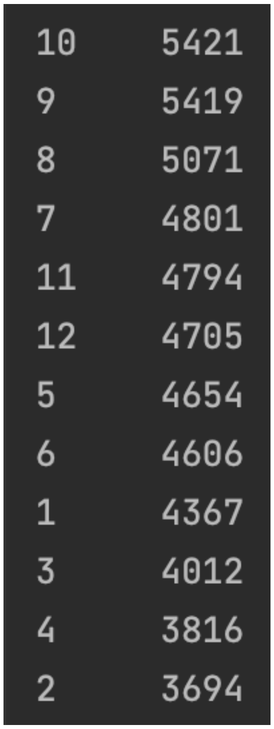




```
others    49560
gun        4647
knife      1153
```

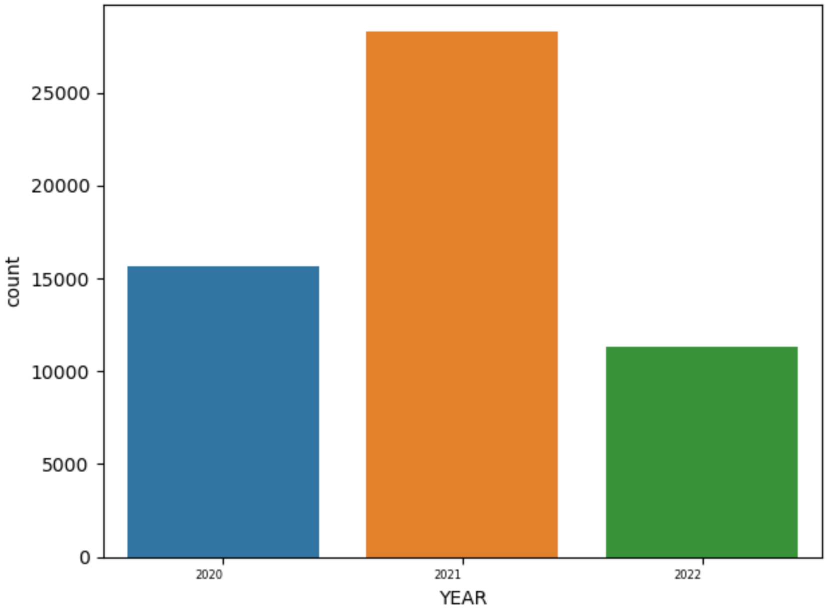
I thought this feature would be more informative than it actually was. I had been hoping that more of the crimes would be classified as commonly having a weapon involved in order to create a relationship between the method and the offense. However, outside of the sphere of wanting specific data, I think it’s a good thing that there isn’t a lot of gun or knife use in crimes committed in DC. This feature also highlights a shortcoming of the MPD’s data collection system. There are too many crimes reported with the use of ‘other’ as a method.

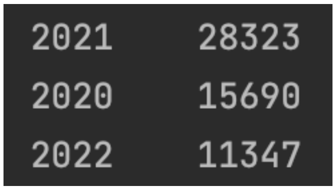




```
10    5421
9     5419
8     5071
7     4801
11    4794
12    4705
5     4654
6     4606
1     4367
3     4012
4     3816
2     3694
```

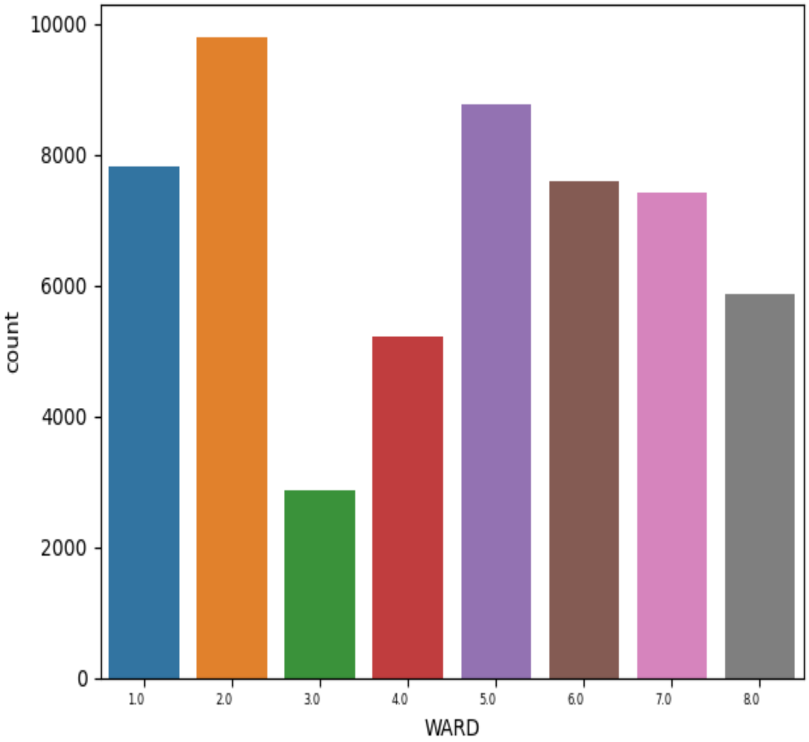
This shows that many criminal incidents in DC take place during the latter half of the year, specifically around September and October.

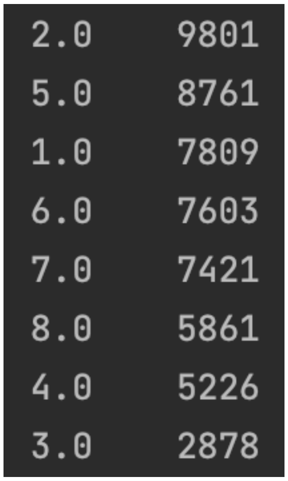




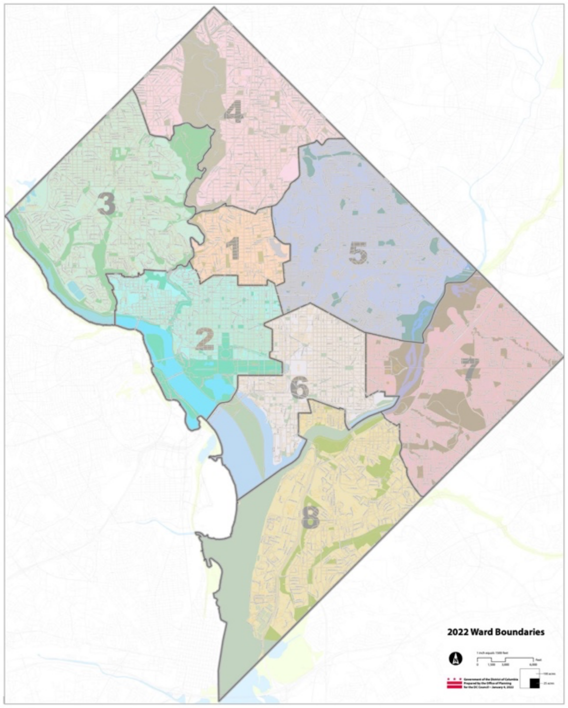
```
2021      28323
2020      15690
2022      11347
```

The incidents in 2020 were much lower than 2021 – almost half. Extrapolating from the data, I think 2022 numbers at the end of the year will be similar to 2021's. 2020 was an anomaly because of the COVID-19 pandemic and stay-at-home orders.



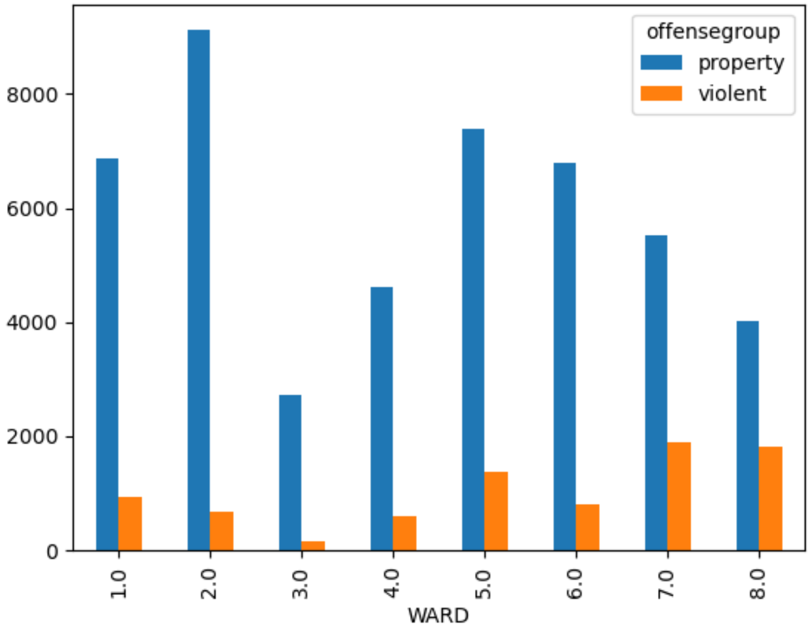


```
2.0     9801
5.0     8761
1.0     7809
6.0     7603
7.0     7421
8.0     5861
4.0     5226
3.0     2878
```

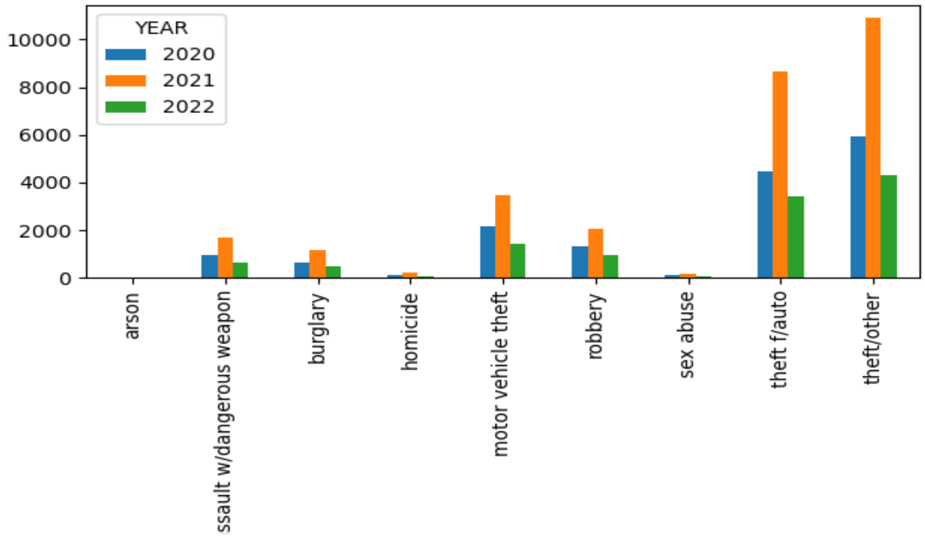


The greatest number of criminal offenses in DC take place in Ward 2, followed by Ward 5 and Ward 1 respectively. The fewest incidences take place in Ward 3. I've added a picture of the 2022 Ward boundaries from the DC Mayoral office for context.

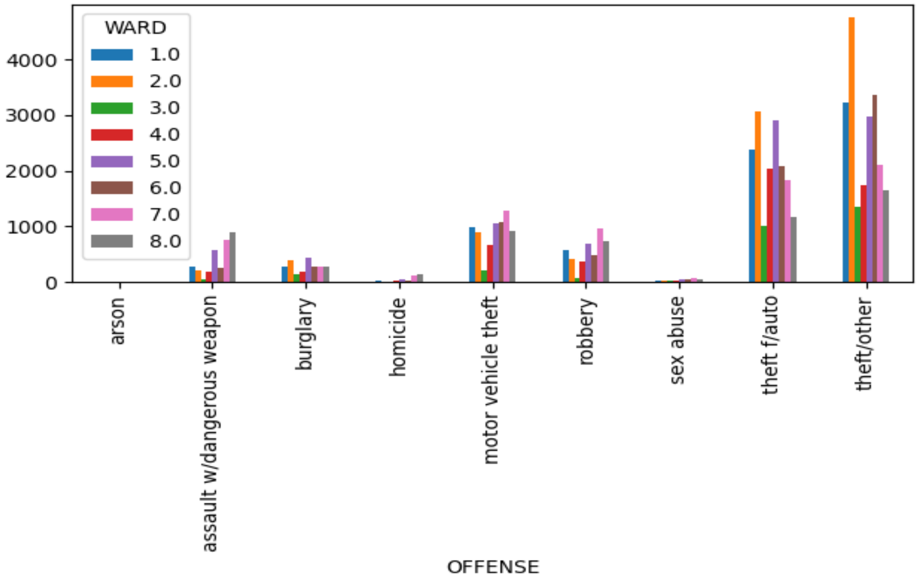
After looking at the features independent of one another, I decided to look at them together to see if there were any relationships to be found.



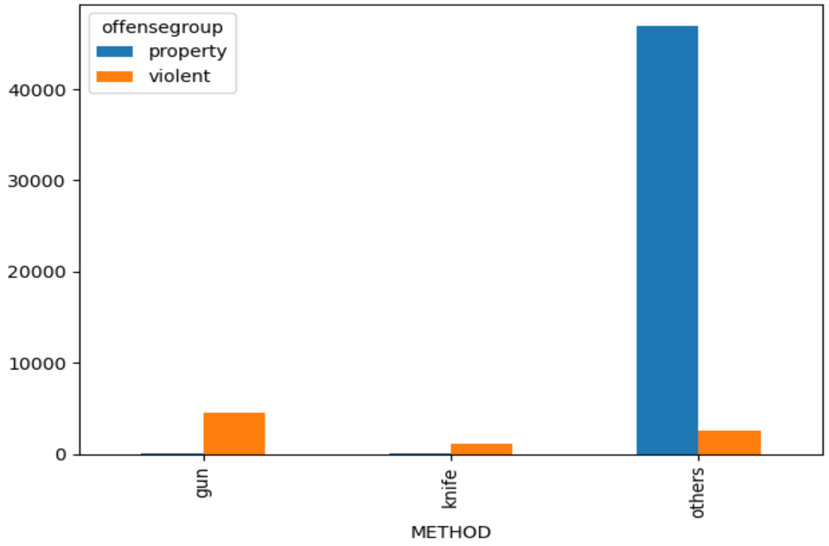
There isn't a particular area in DC that has only one type of crime. Both property crimes and violent crimes can be found across the different wards in DC. The relative scaling is different in each ward, but both types of crimes exist. Which behooves the question of whether there exists a relationship between the types of crimes themselves.



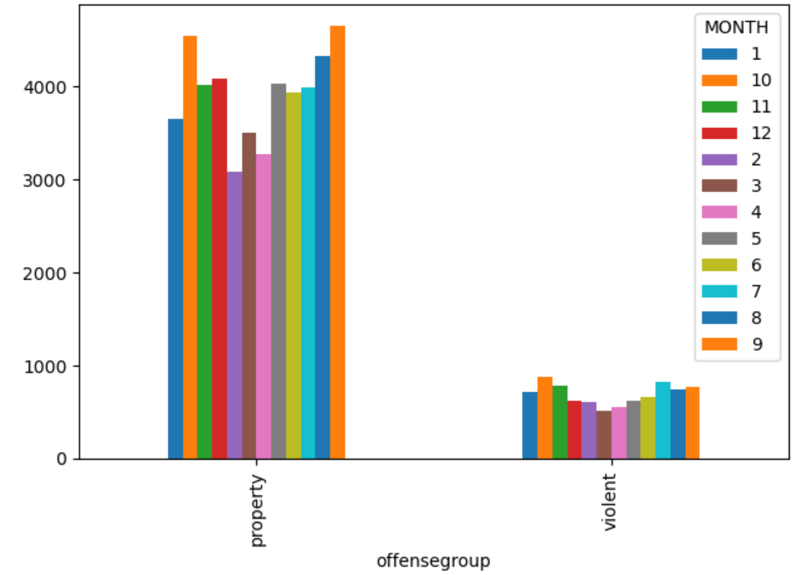
Other than the dip in all crimes in 2020, it seems like the frequency of the types of crimes remains similar across years.



The same can be said for the distribution of crime frequency across the wards.



As I mentioned above, there's not much useful information to be gleaned from the relationship between the method and the type of crime although there's a strong relationship there.

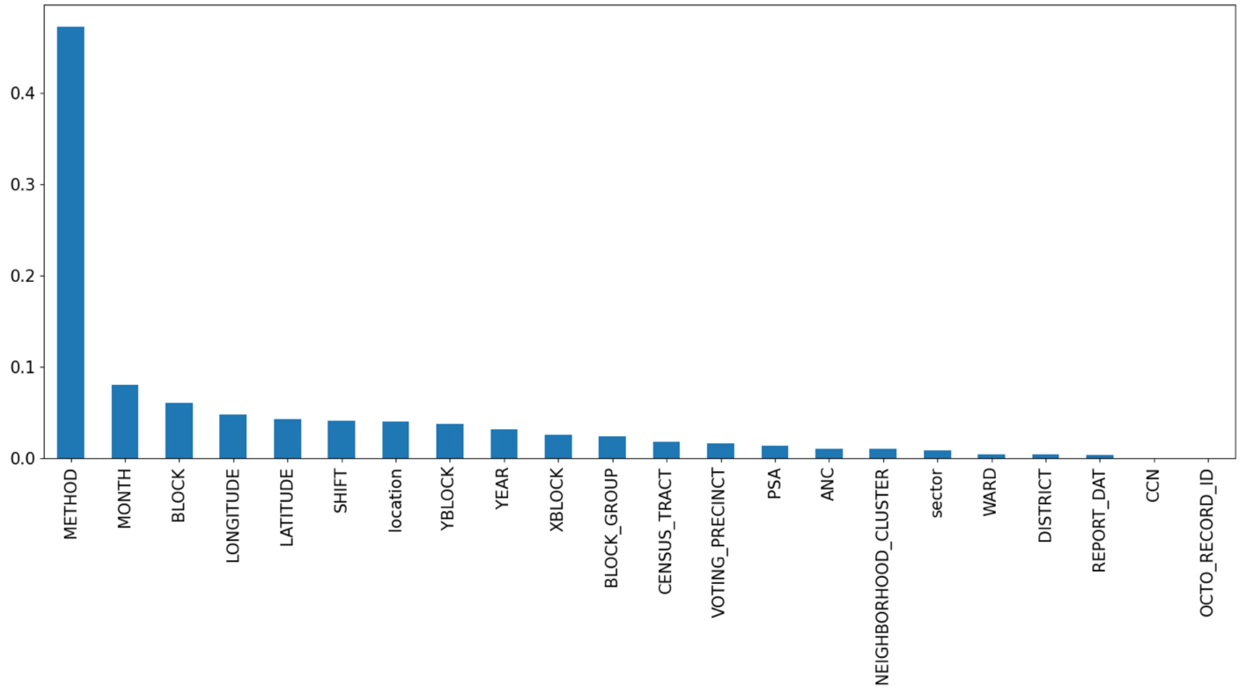


The division of the types of crimes across the different months reflects the division of the crimes as a whole as seen earlier. The rise in property crimes reflect the rise in overall crime as seen in the month feature histogram but the highest frequency for violent crimes is in October and July.

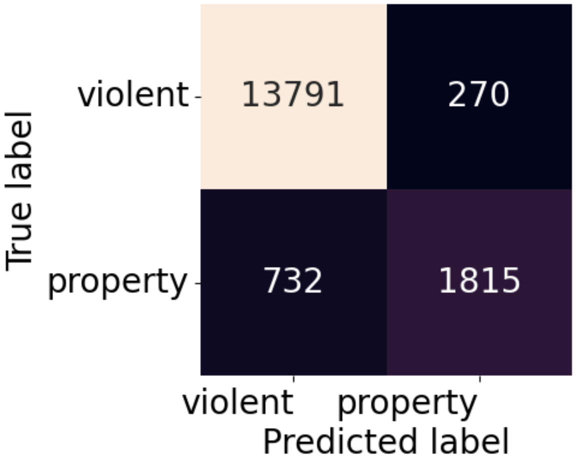
MODELING

Using the information learned above, I decided to use the Random Forest algorithm to create my model. Given that the data was categorical and the split in the frequency of the two types of crimes was so expansive, I thought it would be a fitting model. I also liked the fact that it had bootstrapping aggregation built into the model to get better prediction rates.

The first step for the modeling was splitting the data into training and testing groups. I used the train-test-split function in sklearn and split the data 70/30 with 30% of it being held in reserve for testing the trained model. This meant 38752 observations in the training set and 16608 observations in the testing set.

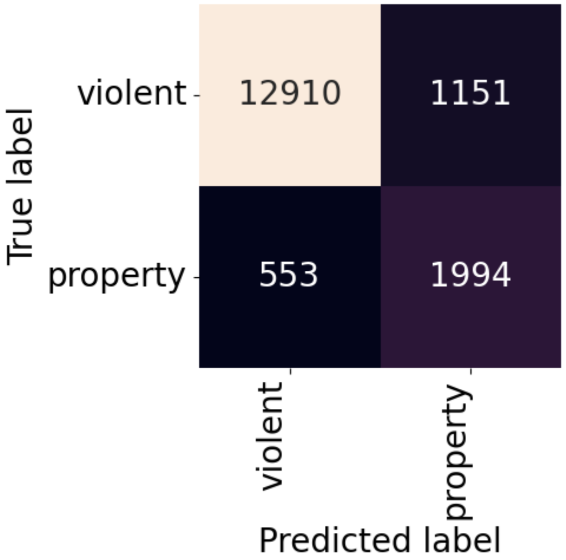


I used the sklearn's functions RandomForestClassifier to train the model and features_importances to rank the features in order of most important to least important. The plot for which you can see above. The features that I thought would be important, like Ward, did not turn out to be important. I trained the data on the Random Forest algorithm with all the features, and with the 11 most important features. Both gave an accuracy rate of around 94% with the model trained with the k features more accurate than the model trained on all features.



This is the confusion matrix that shows the division of false positives and false negatives along with accurately predicted values which resulted from the model trained using the Random Forest algorithm.
It shows that there were 270 observations which were misclassified as property crimes when they were violent crimes and 732 observations which were misclassified as violent crimes when they were actually property crimes.

Given that the criminal offenses didn't happen at the same time and were somewhat independent of one another and other features, I decided to train the data using the Naïve Bayes model as well. I split the data along the same 30/70 lines. With 38752 observations in the training set and 16608 observations in the testing set. Training the model on the Naïve Bayes algorithm led to an accuracy of around 89.7% and the confusion matrix below.



It shows that there were 1151 observations which were misclassified as property crimes when they were violent crimes and 553 observations which were misclassified as violent crimes when they were property crimes.

As the Random Forest model resulted in a higher accuracy score and it used bagging, I decided to use that as the model to train the data. The confusion matrix for the Naïve Bayes algorithm also showed a greater imbalance within the predictions for property and violent crimes. The imbalance between the two was something that I hadn't fixed in my preprocessing but the bagging feature in Random Forest alleviated the issue to some extent.

CONCLUSION

Further Development

Further research on this topic and model building will incorporate the unemployment variable to see how crime and unemployment intersect in Washington DC. Other potentially relevant variables that may help understand incidences of crime in DC and may be of interest and should be obtained for each ward include population, demographics, income distribution, and age distributions.

I would definitely like to learn how to manipulate categorical data to a greater extent. I feel like my current level of python skills didn't do justice to the dataset. I'd like to work on the issues I've mentioned throughout the text and that were pointed out to me during the presentation, like balancing the data. I would also like to create a correlation plot and create better visuals with the data. Training the data on more models would have also given me either a better outcome or more confidence in my decided outcome. These are things I'd like to work on for my next data mining/machine learning project.

Disclosure on Bias

Because crime is self-reported, this database will not include incidences of crime that are not reported or were not humored by law enforcement. Some crimes may be more or less likely to be reported to law enforcement due to societal pressures, stigma, etc. Because of this potential for under representation, this data may inaccurately reflect the true parameters of crime in DC.