Week 2 Reports: Explore Data + data cleaning

Target Variable:

The target variable will be a new column after feature engineering called Severe Crimes. This column will be a be classifier which only contained value of 1 or 0. I choose this as target variable because this is a 200 vs 2 binary problem (200 + no shooting, 2 with shooting) if I only choose shooting as target variable, I believe there might be a problem that makes the model “believes” that there will be no shooting for all crimes.

Here is the rule of which crime will be considering as Severe Crimes:

1. Involves gun shooting
2. ASSAULT - SIMPLE/AGGRAVATED: Assault involves intentionally causing physical harm to another person. Aggravated assault typically involves a weapon or results in serious injury, making it more severe.
3. MURDER, NON-NEGLIGENT MANSLAUGHTER: Murder involves the intentional killing of another person. Non-negligent manslaughter involves unintentional killing but with a disregard for human life. Both are severe due to the loss of life and moral implications.
4. ARSON: Arson involves intentionally setting fire to property, causing destruction and potentially endangering lives.
5. BREAKING AND ENTERING (B&E) MOTOR VEHICLE: Breaking into a vehicle to commit theft or other crimes is severe because it involves property violation and potential financial loss.
6. KIDNAPPING/CUSTODIAL KIDNAPPING/ABDUCTION: Kidnapping involves unlawfully taking and detaining a person, often for ransom or other illegal purposes. It's severe due to the violation of personal freedom and potential harm to the victim.
7. MANSLAUGHTER - VEHICLE - NEGLIGENCE: This involves causing death through negligent operation of a vehicle, such as reckless or drunk driving. It's severe because it results in loss of life due to irresponsible behavior.

Predictors

1. DISTRICT: The geographical area where the crime occurred. Some districts may have higher rates of shootings due to various socio-economic factors.
2. REPORTING_AREA: Similar to DISTRICT, this could provide more granular geographical information.
3. DAY_OF_WEEK: The day of the week might influence the likelihood of a shooting occurring, as certain days might have different activity patterns.
4. HOUR: The time of day when the incident occurred. Certain times might have higher risks of severe crimes, for example late night.
5. STREET: The specific street location might be relevant if certain areas are known to have higher shooting incidents.

Exploration of the dataset:

1. INCIDENT_NUMBER: A unique number assigned to each crime incident.
2. OFFENSE_CODE: A code that represents the type of offense committed.
3. OFFENSE_DESCRIPTION: A detailed description of the offense.
4. DISTRICT: The police district where the incident occurred.
5. REPORTING_AREA: A specific area within the district where the crime was reported.
6. SHOOTING: A binary indicator (usually 0 or 1) showing whether the incident involved a shooting.
7. OCCURRED_ON_DATE: The date and time when the incident occurred.
8. YEAR: The year when the incident occurred.
9. MONTH: The month when the incident occurred.
10. DAY_OF_WEEK: The day of the week when the incident occurred.
11. HOUR: The hour of the day when the incident occurred.
12. STREET: The street where the incident occurred.
13. Lat: Latitude coordinate for the location of the incident.

14. Long: Longitude coordinate for the location of the incident.
15. Location: A combined field of latitude and longitude coordinates.

Data Types & Data Sizes

```
# show datatype of all columns
print(raw_data.dtypes)

# show dataframe size
print(raw_data.shape)
```

✓ 0.0s

```
_id                    int64
INCIDENT_NUMBER       object
OFFENSE_CODE           int64
OFFENSE_CODE_GROUP   float64
OFFENSE_DESCRIPTION   object
DISTRICT              object
REPORTING_AREA        object
SHOOTING               int64
OCCURRED_ON_DATE      object
YEAR                   int64
MONTH                  int64
DAY_OF_WEEK           object
HOUR                   int64
UCR_PART             float64
STREET                object
Lat                  float64
Long                 float64
Location              object
dtype: object
(81133, 18)
```

This dataset has 81133 rows and 18 columns, provided by Boston Police Department (BPD) to document the initial details surrounding an incident to which BPD officers respond. This is a dataset containing records from the new crime incident report system, which includes a reduced set of fields focused on capturing the type of incident as well as when and where it occurred. Records in the new system begin in June of 2015. However, this dataset I will be using only contains data from beginning of 2023 to now.