**Data Review and Preparation for Machine Learning Project**

**Introduction and Business Understanding:** The Chicago Police Department utilizes their Citizen Law Enforcement Analysis and Reporting (CLEAR) system to track crime statistics. Along with geographical information, the CLEAR system also houses the outcome of whether the crime resulted in an arrest and/or whether it was a domestic dispute. We will utilize 2016 data to create a Machine Learning (ML) model. This narrative explains how we prepared for building the ML model by framing the problem, along with steps taken to explore and wrangle the data. Additionally, this narrative provides our initial assumptions that we will check against the results of our ML model.

**Frame the Problem*:*** The data within the CLEAR system can be utilized to answer several business objectives. For example, a visualization of the data could show which districts have the highest crime rate (Figure 1) or how many arrests each beat made (Figure 2).The main business objective our model will answer is predicting the likelihood of an arrest based on various factors, such as type, location, and time of day.

**Data Understanding and Preparation:** CLEAR reported a total of 265,462 crimes in 2016. The original data set contained 23 columns. Several steps were taken to prepare the dataset for modeling, including removing unnecessary columns, updating data types, and addressing missing values.

***Unnecessary Columns:*** The “Year” column was removed from our dataset, as all records contained a value of 2016 within the column. Additionally, several other columns were removed since they would not be beneficial to our model nor help us answer our business objectives ("Unnamed: 0","Case Number", "IUCR", "X Coordinate", "Y Coordinate", "Updated On", "Block", and "FBI Code").

***Data Types:*** The default data types for "District", "Ward", and "Community Area" were float. However, after a review of the data, it was determined that an integer data type is more appropriate and was updated within the dataset. The default data type for the “Date” column was object. A review of the data shows that this column includes a date and timestamp, therefore this field was updated to a datetime datatype.

***Missing Values:*** 14,189 records were missing several values within the “Latitude”, “Longitude” and “Location”columns and 783 records were missing a “Location Description”. These records were dropped from our dataset, as they made up a small portion of the population (5.55%) and there was no easy way to decide what their values should be. After dropping these records, the number of rows within the dataset was reduced to 250,732.

***Binning:*** Counts and distribution were explored for location description. Ultimately, since location description is a categorical variable, it didn’t make sense to bin based on like counts the way you would with a time series for example. The categories were evaluated for commonality to preserve the informative nature of the measure. 127 location descriptions were binned to 10 categories based on description commonality. This will provide a consolidated view of the data while preserving a sufficient level of insight into the locations where crimes are committed. This same approach was also completed with the “Primary Type” variable, reducing the number of values from 31 to 13.

To make the data easier to analyze and reveal clearer patterns, we separated the hour and day from the date variable and created three separate bins. For the hour of the crime, we first examined how crime frequency changed throughout the day. Based on that distribution, we created five time blocks: 1–8, 9–11, 12–16, 17–20, and 21–0 (Figure 5). These bins were chosen to reflect natural breaks in daily activity and differences in crime rates during those periods. For example, crimes committed in the late afternoon and early evening showed noticeably higher frequencies, while early morning hours had fewer incidents.

We also binned the day of the week into just two categories: Weekday (Monday through Friday) and Weekend (Saturday and Sunday). Since social behavior often changes on weekends, we wanted to see if this shift correlated with differences in crime types or arrest rates. Lastly, we grouped each crime into a season (Spring, Summer, Fall, or Winter) based on the date it occurred. This seasonal binning was meant to explore whether changes in weather, daylight, or school schedules might influence crime patterns. These simplifications helped reduce noise in the data while still preserving meaningful distinctions for our model.

The final part of the binning phase included additional data cleaning. Three distinct non-criminal labels were consolidated into a single category for clarity. From there, we also included a new way to identify arrests between violent and non-violent crimes to better understand where these efforts are typically concentrated. These feature engineering steps lay the foundation for enhanced interpretation and potential for stronger model performance moving forward.

***Test/Training:*** To prepare for modeling, the data was split into test and training sets. Due to the large volume of data points in the set, we were able to allocate 25% of the dataset to a testing set. This is still within standard practice. However, we were able to use a percentage towards the higher end of the standard range due to the previously mentioned high volume of data. The random\_state is set to 42 to ensure that through iterations of the data and coding the distribution of data between test and training sets remains the same. This provides us with an apples-to-apples perspective for comparison of results throughout the iterations of our model development.

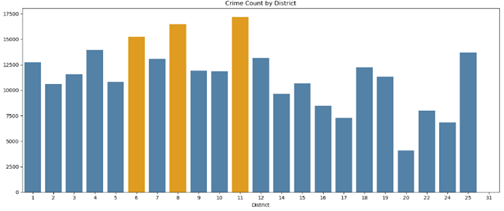
**Initial Assumptions:** Before building our model, we made several assumptions based on what we already know about crime and a first look at the data. We expected violent crimes to lead to more arrests because they are more serious and get more police attention. We also thought drug-related crimes would have a high number of arrests since police often target those specifically. Crimes that happen in the summer or during the afternoon were assumed to result in more arrests, since there are usually more people around and more police activity at those times. We believed that school grounds, especially public schools, would have more violent crimes, even if the total number of crimes there is low. Finally, we expected most crimes to happen in public places connected to infrastructure, as well as in private homes, and that these areas would also have a higher proportion of violent crimes compared to non-violent crimes. These ideas helped guide how we prepared the data and what we looked for in our model.

**Conclusion:** 2016 data collected from the Chicago Police Department’s CLEAR reporting system will be utilized to build a ML model. The objective of the ML model will be to predict the likelihood of arrest based on various factors. Various steps were taken to evaluate and wrangle the data to prepare if for modeling. Additionally, initial assumptions were made that will guide us when evaluating the outcome of our ML model.

**References:** Crimes in Chicago Data, <https://www.kaggle.com/datasets/currie32/crimes-in-chicago?select=Chicago_Crimes_2012_to_2017.csv>;

Machine Learning Project Checklist Roadmap

**Figure 1: Number of crimes by district**

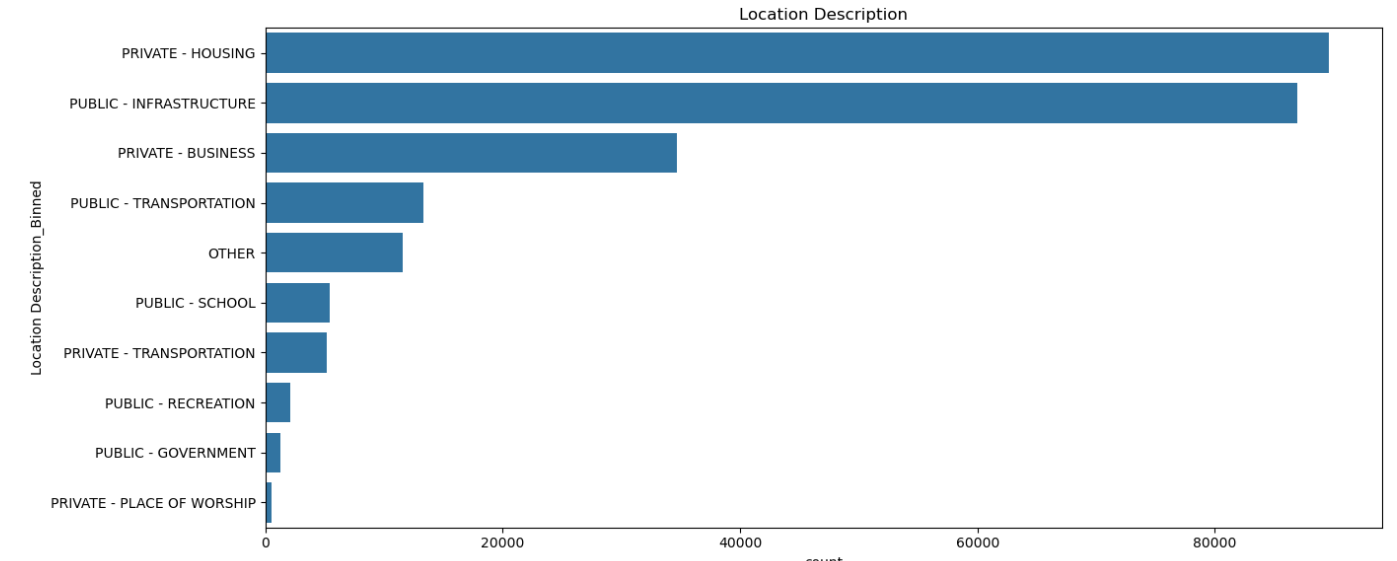
****

**Figure 2: Number of arrests by beat**

**A graph of blue bars

AI-generated content may be incorrect.**

**Figure 3: Number of Crimes by Location Description**

****

**Figure 4: Violent and Non-Violent Crimes by Location Description**

**A graph with orange and blue bars

AI-generated content may be incorrect.**

**Figure 5: Crime Counts by Hours**

**A graph of blue bars

AI-generated content may be incorrect.**