**[TITLE]**

**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.

***Business Objectives:*** 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:***

***Test/Training:***

* 25% of the data will be used as test – standard practice; 62,683 records
* Random\_state at 42 to ensure that the results don’t change

**Initial Assumptions:**

* Violent crimes have likelihood of higher arrest rate
* Summer has a likelihood of higher arrest
* Crimes in the afternoon have a higher likelihood of arrest
* Crimes related to narcotics have the highest likelihood of arrest

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

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

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