READ.ME FILE

The aim of this project is to take a critical approach towards the use of crime data for the geospatial prediction of crime as a means to aid law enforcement. The approach this project takes focuses on the location of bias through data exploration and data mining, as well as predictive modelling, including both logistic regression and 3 classification models in order to examine which attributes contribute the most to predicting our dependent variables, the crimes of murders and robberies.

This project begins with a preliminary statistical analysis: first, our needed libraries are imported. The main libraries this project draws from will be Pandas and Numpy. Throughout this repository, it will also draw from sklearn, matplotlib for more specific functions. The preliminary analysis is meant to provide an opportunity to familiarize oneself with the dataset in general, its organization, column names, data types, missing values, and its shape. The missing values will be prepared here for further processing later on. In this section, there will be an examination of our summary statistics, to further understand the nature of each attribute, such as observing the mean, standard deviation and ranges of each to get an early sense of each distribution. This section will conclude with sub-setting and aggregation. The dataset will be aggregated by the attribute STATE, with an examination of the number of rows represented by each state, and a general summary statistics of how the attributes behave under this aggregation.

In our initial code and results, we will begin to deal with the missing values discovered in the preliminary statistics. A threshold of 0.5 has been set: any attributes that are missing more than 50% of its values will are dropped. I have also chosen my target variables based on missing values: because of the risk of introducing bias with imputation of any target variables, I have chosen the variables that have the least amount to no missing values.

Then, we will begin to address the remainder of the missing values in the dataset, by column. First, we will address the values in columns that fell significantly under the 0.5 threshold, columns missing 1-20 values in total: these are imputed with the MODE of the column. We then move to the missing values that are beneath the threshold but are still significant. The approach taken for these values is to aggregate the data by the attribute, STATE, as there is a higher likelihood of more geographical and social similarities. The ratio of missing values to the length of rows represented by each state will be assessed. If the missing values reach or exceed a threshold of 0.75, the state will be dropped. Otherwise, they will be imputed with the mean of the column of the specific state in which these missing values are located.

Once all of the missing values are dealt with, we move onto exploring correlation. A correlation matrix will be run for the entire dataset to check for highly correlated columns. Then, we create 2 separate datasets for our target variables: one set for MURDERS (murdPerPop) and another set for ROBBERIES (robbbPerPop). For each dataset, we will run a correlation matrix to examine the correlations between the independent variables and the dependent variable. Then, another correlation will be run, but for the entire dataset, without specifying the target variable. All columns with a correlation greater than 0.8 will be dropped. The remaining attributes will be examined. Skew values will be checked, as well as the distributions of each by looking at histograms per each attribute.

Based on the histogram and resulting skew value, the next step is to normalize the attributes. Attributes with a normal distribution will be normalized with a 0-1 range. Attributes that have a non-normal distribution will be transformed into a categorical variable. The levels are established according to quartiles in order to balance them.

Next, both the MURDER DATASET and ROBBERIES dataset will be prepared for both the classification and regression models. The target variables will be transformed to a categorical variable for classification, and a normalized range of 0-1 for regression. The datasets will be ready to use in order to run the models. The datasets will be divided into train and test splits. The first train/test split will be using hold-out cross validation: the simple and common 80/20 split. The models will be run again, but using stratified k-fold cross validation to address any imbalances and ensures equal representation in all folds. The models that will be used are logistic regression and 3 classification models: KNN, Decision Tree and Random Forest. A confusion matrix will be created for the output of each model, and the classification models results will be compared based on the evaluation metrics: Accuracy, Precision, and Recall.