**Predicting Crime Frequency in Prince George’s County, Maryland**

**Introduction**

Located in Maryland just outside of Washington, D.C., Prince George’s County offers comprehensive data on crime incidents within the county, going back to 2017. Using Python along with packages such as requests, Pandas, and Geopandas, this data can be put into a format that allows for easy analysis and visualization. Prediction models such as regression analysis allow for the prediction of the total quantity of crime occurrences in the next month as well as more specific subcategories of crime such as homicide, violent crime, etc.

**Background**

The relationship between crime and the calendar is a long-discussed topic—it is common knowledge that murder rates increase in the summer, resulting in the famous correlation between ice cream sales and murder, oft-used by statistics professors in an attempt to explain the difference between correlation and causation. What other trends can we find when we examine monthly crime data over 5 years’ time? Are certain regions within the county more prone to certain types of crime than others? Using this past November as testing data, the goal of this project is to predict the number of crime incidents in Prince George’s County in December of 2023, with the ability to select for location and subcategory of crime.

To generate predictive data about aggregate crime throughout the county, we turned to a logistic regression model, specifically polynomial regression. From the data we compiled for February of 2017 through November of 2023, we found the best method to be to sort each crime instance into one month bins, displaying the total number of crimes occurring in any given month. This process required extensive effort to clean up and process the initial data, but the result was a scatter plot, showing the total number of instances of crime for each month in the specified period. The subsequent task was to fit this data to a curve. To avoid a polynomial that would overfit or underfit the data, we ran the regression calculation using 80% of the data as training data for every degree of polynomial, from 1 (a line) to 20 to determine which generated the lowest mean absolute error with the remaining 20% of the data, used as testing data. Ultimately, the degree 13 polynomial was found to generate the lowest error. However, (include details of after month of November is completed).

An identical process was done with various types of crime. After separating homicides into their own DataFrame, we performed the same regression analysis. This time, due to the lack of correlation between time and number of monthly homicides, the best fit was a line, and did not reflect the data well.