Analyzing U.S. Crime Data

Technical Documentation

CIE 457

Khaled Elbastawisy 201-800-029

Mohamed Ahmed 201-801-898

Moataz Mohamed 201-801-903

Project Structure

File Hierarchy

Two Jupyter Notebooks were used for the analysis purposes of the project and code execution. The first notebook was used for the data collection and cleaning purposes, while analysis, testing and modeling took place in the second notebook. The first notebook tapped into Google Drive for the storage of the dataset files after data collection and cleaning. Each of the four data sources has a separate drive folder with both the original data files and cleaned versions of the data stored. The state populations dataset is kept in the main folder.

Limitations and Challenges

- The CDE API was especially hard to work with as the website changed in the past few days, and the API endpoints shown on the website lacked the endpoint we were looking for. The server also randomly stopped responding to all requests at times (not a rate limit issue)
- The NIBRS data was missing records for certain offenses in certain years. A good example of this is missing assault offenses in New York for years prior to 2010
- The NCVS had a substantial number of residue, invalid, and unknown records that had to be discarded during analysis and comparisons.
- The lack of consistency for data types of different columns of the Recidivism dataset for the state of Georgia made it difficult to infer deeper knowledge from the records. A good example would be the grouping of some values as “x or more” which created ambiguity and could lead to results that deviate from reality.

General Flow

The flow of the analysis followed the following Structure:

- Data Collection
- Data Cleaning
- EDA
- Answering Questions
- Hypothesis Testing
- Regression Analysis
- Machine Learning Classifier

Detailed Project Stages

Data Collection

The code starts by collecting the datasets off the provided APIs for NCVS, NIBRS, NIJ and State Firearm Laws. Each dataset is retrieved through a specific retrieval function.

For REST based APIs, separate functions are implemented to retrieve the data. These two functions are called “NCVS_data” and “NIJ_data”. The NCVS_data function returns both victimization and population records, while the NIJ_data function returns the Georgia crime records file only. A count of 1000000 was used as an argument in the API requests. As for the NIBRS crime data, An initial request for the state names and abbreviations is made. The list of offenses and offense categories is collected from the CDE API. These lists are then used for a for loop to create a list of URLs that each ask for a state’s records of a certain offense. This list is then passed to the “call_nibrs_api” function which is run simultaneously on 12 threads to make concurrent requests to the API to acquire crime count records for each state and offense. This was to speed up the collection process as the API call process was severely IO bound and would have required hours to execute sequentially. The result of each API call is a list of dictionaries which is appended to the main offense count list, called “offense_count_list.” As for the State Firearms Laws Dataset, A link to the dataset Excel file is used to read the entire dataset directly.

Data Cleaning

1. **NCVS:**

The data was initially filtered by column as only a handful of the existing features were needed for our analysis. The kept features for the victimization were as follows:

- 'Idper' **(ID)**
- 'Year'
- 'Newoff' **(Crime Type)**
- 'Ager' **(Victim Age)**
- 'Race'
- 'Sex'
- 'Hincome1' **(House Income)**
- 'Educatn1' **(Educational level)**
- 'Direl' **(Relation to Offender)**
- 'Notify' **(Crime Reported?)**
- 'Treatment' **(Victim treated?)**
- 'Serious' **(Injury Type)**
- 'Offenderage'
- 'Offendersex'
- ‘Offtracenew' **(Offender Race)**

The kept features for the population records were as follows:

- 'Idper' **(ID)**
- 'Year'
- 'Ager' **(Victim Age)**
- 'Race'
- 'Sex'
- 'Hincome1' **(House Income)**
- 'Educatn1' **(Educational level)**

Most of the used features had categorical values in the form of numerical code. Those values were replaced by more understandable category names that are mentioned in the NCVS codebook.

After the reassignment of categorical values, any records with “Residue” values in the victimization and population datasets were removed since these provided no insight into comparisons we were planning to make. The dataset columns were then renamed to be more easily readable.

2. Recidivism data for the state of Georgia

Dataset columns were renamed accordingly to reflect the intended meaning more accurately. This was done through a manually created “rename_dict”. Null values were investigated afterwards and removed for the “gang_affliated “ column.

3. NIBRS Reported offense count

The collected data spanned a large set of offenses, most of which were close in nature. After studying the data codebook we decided to group offenses based on their offense category. For instance, all these offenses (aggravated-assault, simple-assault , intimidation) fall under the category of assault offenses and so on for the other offense categories. We added the count of all single offenses and assigned the resulting count to the bigger offense category. For each unique combination of (year, state, offense) there was an offense count. The columns were rearranged for higher clarity and saved for later analysis.

4. Firearm laws per state

The firearm dataset contains almost 150 laws for gun regulations. This would have made analysis extremely difficult. Therefore the laws were grouped into 14 law categories that are as follows:

- Dealer_regulations
- Buyer_regulations
- Prohibitions_for_highrisk_gun_possession
- Background_checks
- Ammunition_regulations
- Possession_regulations
- Concealed_carry_permitting
- Assault_weapons_and_large_capacity_magazines
- Child_access_prevention
- Gun_trafficking
- Domestic_violence
- Preemption
- Immunity
- Stand_your_ground

The laws applied within each category were then summed to give an idea of how strict or loose a state’s gun regulation is.

Exploratory Analysis

1- National criminal offense rates per year across all available years for the top five most frequent offense categories.

The NIBRS and state population data is first loaded. The state population data is then transformed. The transformed state population data contains columns for state, year, and population.

Each record represents a specific state's population in a specific year. To calculate the yearly national criminal offense rates, we needed to divide the number of offenses for each category in each state in a given year by the state population in that year. This would give us the offense rate per 100,000 population for each category in each year for each state.

Then for each year we used the mean offense rate per 100,000 population over the offense rates of different states. This gave us the national offense rate for each category and year

The reason for following this analysis instead of summing the offense counts for all the states and dividing by the country's population is due to the fact that some criminal offense counts are not reported for some states in some years. So dividing by the whole population would lead to misleading rates.

To identify the top five most frequent offense categories, we used the offense rates of each category and got their mean over the years. We then ranked the offense categories based on the mean offense rates and selected the top five. A line plot was then used to visualize the trends over time.

2- The average percentage of violent crimes relative to total crime per state over all available years.

First, we calculated the total number of crimes for each state for each year by summing up the offense counts for all crime categories. Next, we calculated the total number of violent crimes for each state for each year by summing up the offense counts for all violent crime categories. Then, for each state, we calculated the average percentage of violent crimes relative to total crime by dividing the total number of violent crimes by the total number of crimes and multiplying by 100.

To visualize the results, we created a bar plot showing the average percentage of violent crimes per state. This allowed us to easily compare the average percentage of violent crimes among the different states.

3- National homicide rates, as well as total violent crime rates per year over all years.

To get the national homicide offense rates, we will use the data frame constructed earlier for the yearly national offense rates and select the homicide offenses only. To get the violent crime rates, we will use the same dataframe and get the average of the violent offenses for each year which would give us the violent crime rate per year. A line plot was used to show national homicide rates per 10000 and 100000 populations, and a scatter plot was used to compare the homicide and violent crime offense rates to see if there is a correlation between them.

4- The frequency of non-fatal crime incidents in relation to victim demographics.

To get the frequency of non-fatal crime incidents per demographic, the victimization NCVS dataset was used, where the data set was grouped by sex, race, and age group respectively, and the count occurrence for each crime type was then divided by the entire victimized count for each category. The data was displayed using stacked bar chart and pie chart plots, where the stacked bar chart was used to show the proportion of each crime to the total crime count of each demographic parameter (i.e. the percentage of total female victims that were subjected to larceny), and the pie charts were used to show the proportion of each demographic parameter to the total count of each crime type (i.e. the percentage of all larceny records where the victims were of African American race).

5- The frequency of non-fatal crime incidents in relation to offender demographics.

To get the frequency of non-fatal crime incidents per offender demographic, the victimization NCVS dataset was used, where the data set was grouped by offender sex, offender race, and offender age group respectively, and the count occurrence for each crime type was then divided by the entire victimized count for each category. The data was displayed using stacked bar chart and pie chart plots, where the stacked bar chart was used to show the proportion of each crime to the total crime count of each demographic parameter (i.e. the percentage of total female offenders that committed larceny), and the pie charts were used to show the proportion of each demographic parameter to the total count of each crime type (i.e. the percentage of all larceny records where the offenders were of African American race).

6- The relationship between the victim’s education level, their gross household income, and their rate of victimization.

To investigate the relationship between victimization rate, house income and educational level, the victimization and population NCVS datasets were used, where the data set was grouped using house income and educational level as a multi-index based table. The same grouping was performed on the population dataset to get the counts for each income-education combination. The count occurrence for each house-income-educational level combination was then divided by the entire population count for each category to get the percentage of population that was victimized for each category. The total victimization rate values were then altered using a log2 function to accentuate the variability between values as they were initially all extremely small. Since we have three variables to visualize, a heat map was used to show the victimization rate of each group. A pie chart showing the victimization rates of the top three groups for each crime type was also used to gain further insight.

Answering Questions

1. **Question 1:**

Which type of non-fatal crime is the most under-reported?

The NCVS victimization dataset was used for this question. The data for the counts of different crime types was grouped using the "is_reported" feature. This left us with a dataframe of three rows (reporting status) and five columns (crime type) The 'unknown' status values were afterwards discarded since they do not add any information to the current comparison. The data was visualized using stacked bar-chart and pie chart graphs. The bar chart showed the counts of crime occurrences for each reporting status. The pie chart showed the percentage of reported and unreported occurrences of each crime.

Is there an association between the offender-victim relationship and the likelihood of a crime being reported?

To investigate the relationship between offender-victim relationship and likelihood of crime reporting, the victimization NCVS dataset was used, where the data set was grouped using reporting status and victim-offender relationship as a multi-index based table. The count occurrence for each reporting-relation combination was then divided by the entire population count for each reporting status to get the rate of reporting for each group. Since we have three variables to visualize, a heat map was used to show the reporting rate of each group.

2. **Question 2:**

Who are the people (the demographic segment) that appear to be most at risk of violent victimization? Who is the least at risk?

The NCVS victimization and population datasets were used for this question. The dataset was grouped by sex, race, and age group respectively, and the victim counts were subtracted from the total population to get the non-victim counts for each demographic category (i.e. female victims and non-victims). The counts for each category of victims and non victims were then visualized using a pie chart. The pie charts were then used collectively to decide the most and least at risk demographic.

3. **Question 3:**

Of all victims of non-fatal crimes who suffer an injury, which demographic is the most likely to receive medical attention at the scene? Which is the least likely?

The NCVS victimization dataset was used for this question. The dataset was grouped by sex, race, and age group respectively, and the treated victim counts were subtracted from the total victim count to get the non- treated victim counts for each demographic category (i.e. female treated victims and non-treated victims). The counts for each category of treated and non-treated victims were then visualized using a pie chart. The pie charts were then used collectively to decide the most and least treated demographic.

4. **Question 4:**

Which class of crimes is associated with the highest rate of same-offense-recidivism; i.e. prison re-entry for the same offense within 3 years of release?

The Recidivism data for the state of Georgia was used for this question. Data was grouped using the prison offense categorical column and both it and the “recidivism within 3 years column” were used to create a bar plot of the data.

5. Question 5

Are prisoners who are younger at the time of release more or less likely to reoffend than those who are older?

The Recidivism data for the state of Georgia was used for this question. Data was grouped using the age at release categorical column and both it and the “recidivism within 3 years column” were used to create a bar plot of the data.

Hypothesis Testing

Hypothesis Testing 1.0

We first tested the Claim that *"U.S. states that implement stricter firearm control laws, have lower violent crime rates on average"* using two-tailed t-test and linear regression analysis test.

I. Two-tailed t-test:

To assess the validity of the claim that "U.S. states that implement stricter firearm control laws have lower violent crime rates on average," we used a two-tailed dependent samples t-test. This test is appropriate because it compares the mean violent crime rates of states that have stricter firearm control laws to the mean violent crime rates of states that have less strict firearm control laws, and we have repeated measures (violent crime rates) for each state over the past 30 years.

The hypotheses for this test would be:

- Null hypothesis (H0): There is no difference in the mean violent crime rates of states that have stricter firearm control laws versus those that have less strict laws.
- Alternative hypothesis (Ha): There is a difference in the mean violent crime rates of states with stricter firearm control laws and those with less strict laws.

To conduct the test, we needed to first classify the states into two groups: those with stricter firearm control laws and those with less strict laws. We did this by calculating the mean number of gun control laws applied in each state across all years in the dataset, and then dividing the states into two groups based on whether they have above or below the median number of laws applied.

After that, we calculated the mean violent crime rate for each group of states. Finally, we conducted the t-test by using a SciPy statistical software package.

II. Linear regression analysis test:

Another appropriate test for this scenario is linear regression analysis, as it can help determine the strength and direction of the relationship between two continuous variables (in this case, the number of gun control laws and the violent crime rate).

The null hypothesis for this test would be that there is no relationship between the number of gun control laws and the violent crime rate, while the alternative hypothesis would be that there is a relationship between the two (either positive or negative, depending on the direction of the association).

We then needed to fit a linear regression model using the number of gun control laws as the predictor variable and the violent crime rate as the outcome variable. We then evaluated the model's performance using metrics such as R-squared and the F-statistic, and determined the p-value for the regression coefficient for the predictor variable.

If the p-value is less than the chosen significance level of 0.05, we can reject the null hypothesis and conclude that there is a significant relationship between the number of gun control laws and the violent crime rate. On the other hand, if the p-value is greater than 0.05, we cannot reject the null hypothesis and cannot conclude that there is a significant relationship between the two variables.

Hypothesis Testing 2.0

We then formulated another claim that *"U.S states with larger population tend to have more crime rate"* using Pearson's correlation test

Pearson's correlation test is appropriate because it can be used to determine the strength and direction of the linear relationship between two continuous variables, in this case the population size and the crime rate.

The hypotheses for this test would be:

- H0 (null hypothesis): There is no relationship between population size and crime rate. The population size and crime rate are independent of each other.
- H1 (alternative hypothesis): There is a relationship between population size and crime rate. The population size and crime rate are related to each other.

To conduct this test, we used the crime offense count dataset coupled with the states' population dataset. We calculated the crime offense rate for each offense category in each state for each year. Then we calculated the mean crime rate in each state for each year (by averaging over all offense categories for every state in each year). Since that we already have the data for all the states' population for each year. We used those two variables to try to establish a relationship between the crime rate and the state's population.

Regression Analysis

In this section we attempted to fit a regression model that predicts the Offender's supervision risk score based on :

- All prior convictions.
- Offender's race.
- Offender's gang affiliation.
- Offender's age at release.

Data preprocessing:

Gang Affiliation Column:

Since that gang_affliated and first_parole_risk_score columns are important for the regression models, we will take the decision to drop the rows with gang_affiliated null values which are a very small percentage of the data.

Race Column:

We encoded the race column using one-hot encoding since the races are either black or white.

Prior Conviction Columns:

There are two types of prior conviction columns: Columns that are coded with the number of prior convictions, the convictions with their possible values are:

- felony_convictions = {0, 1, 2, 3 or more}
- misdemeanor_convictions = {0, 1, 2, 3, 4 or more}
- property_convictions = {0, 1, 2, 3 or more}
- drug_convictions = {0, 1, 2 or more}

We kept the column values as they are, since they represent a logical quantity. We just replaced the last number with its exact number (3 or more >> 3, and so on)

The second type of prior conviction columns has a binary coded values that indicate if the offender had been convicted of that offense before and these are columns are:

- violent_convictions
- parole_violation_convictions
- domestic_violence_convictions
- gun_charges_convictions

For these columns we left the binary codes (true : 1) and (false : 0)

Age at Release Column:

The age_at_release column can take values of the possible category of values: (18-22, 23-27, 28-32, 33-37, 38-42, 43-47, 48 or older). We numerically encoded this column by assigning each age group a number from 1-7 based on its natural order (ordinal encoding).

Fitting and Visualization:

We used a SKlearn based linear regression model to fit the data and the results of the model were then displayed using a simple print statement. The coefficients and p-values of the predictors were analyzed to identify the good/bad predictors. The R-squared measure was used to assess the quality of the model. A correlation matrix with the entire feature space was then used to visualize the correlation between different features of the dataset.

Bonus Task (ML Classification):

In this task we used state of Georgia recidivism records to train a random forest classifier to predict the likelihood of recidivism within 3 years of release.

We built a pipeline for processing the dataset that followed these simple steps:

- Applied one hot encoding for the categorical columns
- Split the dataset into train and test samples using the last column in the dataset (‘training_sample’)
- Dropped redundant columns
- Fitted an imputer transformer on the training dataset with a mean strategy to fill the missing numerical values.The training set only was used, and not on the test set. This is because the test set should be treated as "unseen" data, and the imputation values should not be based on the test set.
- Using SKlearn library, we created a random forest classifier object and fitted it on the training data
- We then used the model to predict the target variable (recidivism within 3 years) and analyzed its accuracy.