University of Science and Technology
Communications & Information Engineering Program
CIE 457: Statistical inference and data analysis

Final Course Project: Analyzing U.S. Crime Data

In this project we will use the publicly available crime data offered by the Federal Bureau of Investigation (FBI) and the Bureau of Justice Statistics (BJS) to analyze the patterns of crime in the U.S across time, regions, and demographics.

PART 1: Data Collection and Cleaning: [17.5 %]

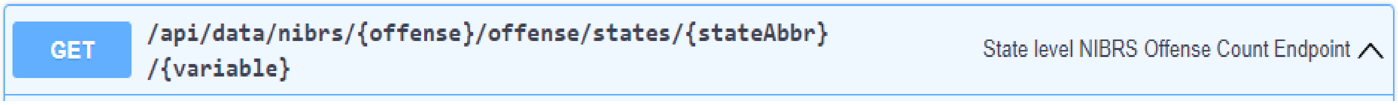
2.1 Data collection: [10%]

The national crime victimization survey (NCVS) data:

- Use the NCVS data API (Available without an API key), to download as much of the *Personal crime victimization* and *Personal population* data as is available for all years (set limit >= 1000000).
- To make sense of the dataset content, you need to carefully read its *codebook*. The codebook contains a detailed description of variables and their meaning.

NIBRS Reported offense count data:

- Use the FBI Crime Data API (Must apply for an API key first), to download the state-level NBIRS offense count data for all available states and offenses.



GET /api/data/nibrs/{offense}/offense/states/{stateAbbr}/{variable}

State level NIBRS Offense Count Endpoint

- You will need to get a list of state abbreviations and names first. You can do this manually or by code with a separate API call to the *Lookups Controller/agencies* server. You should have a list of 54 states.
- Collect the data by repeated requests from the state-level NIBRS count server for all possible states and offenses to get the *COUNTs*.
- Carefully read the offense definitions and categories.

Recidivism data for the state of Georgia [2013-2015]

Recidivism is the tendency of a convicted criminal to reoffend after being released from prison.

- Use the SODA API to download as much of the NIJ's recidivism data as is available (set limit >=1000000).
- Carefully read the data description and codebook.

Firearm laws per state

- Download the Firearm laws per state dataset and codebook.
- Read the codebook to understand the structure of the dataset, and the purport of each referenced law.

2.2 data cleaning: [7.5%]

Do whatever cleaning/restructuring steps you see necessary on the data. However, your final output must fulfill the following:

1- Have descriptive column/variable names.
2- numerically-encoded categories must be replaced by their descriptive string in the analysis outputs and plots.

In the next sections, you get to decide which one or combination of these datasets to use in each analysis task.

PART 2: Exploratory Analysis: [15%]

ALL

2.1 Use the appropriate statistics and plots to investigate the following:

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

2. The average percentage of violent crimes[1] relative to total crime per state over all available years.
3. National homicide rates, as well as total violent crime rates per year over all years.
4. The frequency of non-fatal crime incidents in relation to victim demographics[2].
5. The frequency of non-fatal crime incidents in relation to offender demographics.
6. The relationship between the victim's education level, their gross household income, and their rate of victimization.

All analysis points must be accompanied by your commentary and at least one appropriate visualization.

PART 3: Answering Questions: [12.5%]

Jihad

3.1 Use the appropriate statistics and plots to answer the following questions:

1. Which type of non-fatal crime is the most under-reported? Is there an association between the offender-victim relationship and the likelihood of a crime being reported? (reported: ie, police notified at time of occurrence)
2. Who are the people (the demographic segment) that appear to be most at risk of violent victimization? Who is the least at risk?
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?

[1] **all_violent_crime = ['assault_offenses','homicide_offenses', 'robbery','kidnapping_abduction' ,'sexual_assault']**

[2] **demographics: age, sex, race/ethnicity**

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

All analysis points must be accompanied by your commentary and at least one visualization.

PART 4: Hypothesis Testing: [15%]

All

Claim: “U.S. states that implement stricter firearm control laws, have lower violent crime rates on average”

4.1 Formulate a hypothesis test to assess the validity of this claim given the available data:

- State the test you will use and justify your choice.
- Clearly state the hypotheses.
- Conduct the test and report the result.
- Make a conclusion as to the validity of the claim, assume a significance level of 0.05.

4.2 Come up with your own claim and formulate a hypothesis test to assess it following in the same steps.

PART 5: Regression Analysis: [7.5%]

Mohamed

Use The recidivism in Georgia dataset 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.**

1. Report your model's coefficients and p-values.
2. Which of these variables are good predictors of the variabilities in the target? Which are bad ones?
3. Are any of these predictors correlated with each other? Assess the quality of your model.

PART 6: Bonus Task:[10%]

Mohamed

Train a machine/deep learning classifier to predict the likelihood of recidivism within 3 years of release based on the state of Georgia recidivism records.

PART 7: Documentation:[25%]

ALL

You are required to submit two documents:

1- <u>A business report:</u> containing all the analysis results from parts 2 to 6 along with the associated visualizations and comments. In addition, include an introduction and a conclusion explaining the significance and potential limitations of your findings.

1- <u>A technical documentation:</u> that explains:

- The structure of your project: file hierarchy, general flow, etcetera.
- A description of all functions and their usage.
- The steps you followed for data collection/cleaning, and all subsequent analysis requirements.
- The challenges/limitations/assumptions involved in any step.

PART 8: Presentation:[7.5%]

Mahmoud

For the final presentation you should prepare slides summarizing all your findings with visualizations and brief comments.