مدينة زويل للعلوم والتكنولوچيا
Zewail City of Science and Technology

Final Project - Technical Report

CIE 457

Prepared For

Dr. Mahmoud Abdelaziz

Supervised by

Eng. Anhar Hassan - Eng. Asmaa Mahmoud

Prepared By

Ahmed Mahmoud	201800683
Ahmed Elghamry	201801254
Ibrahim Hamada	201800739
Hazem Tarek	201800283

Table of content :

Part 1: Data Collection and Cleaning

1. The National Crime Victimization Survey (NCVS) Data

Dataset Description

The primary source of data on criminal victimization in the country is the National Crime Victimization Survey (NCVS). Each year, statistics are collected from a sample of approximately 240,000 people in approximately 150,000 households, which is nationally representative. The NCVS gathers data on household property crimes (such as burglary/trespassing, motor vehicle theft, and other types of theft), as well as nonfatal personal crimes (such as rape or sexual assault, robbery, aggravated and simple assault, and personal larceny). Respondents to the survey disclose personal data about themselves, including their age, sex, race, and ethnicity (e.g., Hispanic origin), marital status, level of education, and income, as well as if they have ever been victims of victimization.
The NCVS Select - Personal Population section of the survey includes respondents' demographic data. regardless of whether they acknowledged being a victim of personal crime.
Personal criminal victimisations are included in NCVS Select - Personal Victimization. Rape and sexual assault, robbery, simple and aggravated assault, and personal theft/larceny (purse snatching/pocket picking) are all examples of personal crimes.

Dataset Collection

Data is obtained from their website directly, the user does not need an api key to download data.

Data Fields Categories

For the personal population data

- Person ID: Unique person identifier   (idper)
- Year and quarter:Year and quarter data was collected (yearq)
- Year:Year data was collected (year)
- Age:Respondent age on the last day of the month before the interview (ager)
- Sex:Respondent sex (sex)
- Hispanic origin:Respondent Hispanic origin (hispanic)
- Race:Respondent race(race)
- Race/ethnicity:Respondent race and Hispanic origin(race_ethnicity)

- Annual household income: Total income of all members of the household for the 12 months preceding the interview. Categories available from 19932021 with imputed data, starting 2015 Q1.(hincome1)
- Annual household income- imputed: Imputed income categories, starting 2017 Q1 (hincome2)
- Marital:statusRespondent marital status (marital)
- Population size:The size range for the place in which the housing unit is located, starting 1995 Q3(popsize)
- Region:Region of respondent residence. The states have been divided into four groups or census regions, starting 1995 Q3. (region)
- Household MSA: Classification of respondent residence based on the Office of Management and Budget definition of metropolitan statistical areas (MSAs) (msa)
- Household locale:Location of household based on BJS geography definitions, starting 2020 Q1(locality)
- Education level:Respondent level of education (educatn1)
- Education level:Respondent level of education, starting 2003 Q1 (educatn2)
- Veteran status:Respondent veteran status, starting in 2017(veteran)
- Citizenship status:Respondent citizenship status, starting 2017 Q1(citizen)
- Person population weight :Population weight to use on person population data(wgtpercy)

For the personal Victimization data

- Person ID: Unique person identifier (idper)
- Year and quarter:Year and quarter data was collected (yearq)
- Year:Year data was collected (year)
- Age:Respondent age on the last day of the month before the interview (ager)
- Sex:Respondent sex (sex)
- Hispanic origin:Respondent Hispanic origin (hispanic)
- Race:Respondent race(race)
- Race/ethnicity:Respondent race and Hispanic origin(race_ethnicity)
- Annual household income: Total income of all members of the household for the 12 months preceding the interview. Categories available from 19932021 with imputed data, starting 2015 Q1.(hincome1)
- Annual household income- imputed: Imputed income categories, starting 2017 Q1 (hincome2)
- Marital:statusRespondent marital status (marital)
- Population size:The size range for the place in which the housing unit is located, starting 1995 Q3(popsize)
- Region:Region of respondent residence. The states have been divided into four groups or census regions, starting 1995 Q3. (region)
- Household MSA: Classification of respondent residence based on the Office of Management and Budget definition of metropolitan statistical areas (MSAs) (msa)

- Household locale:Location of household based on BJS geography definitions, starting 2020 Q1(locality)
- Education level:Respondent level of education (educatn1)
- Education level:Respondent level of education, starting 2003 Q1 (educatn2)
- Veteran status:Respondent veteran status, starting in 2017(veteran)
- Citizenship status:Respondent citizenship status, starting 2017 Q1(citizen)
- Person population weight :Population weight to use on person population data(wgtpercy) Person ID: Unique person identifier (idper)
- Year and quarter:Year and quarter data was collected (yearq)
- Year:Year data was collected (year)
- Age:Respondent age on the last day of the month before the interview (ager)
- Sex:Respondent sex (sex)
- Hispanic origin:Respondent Hispanic origin (hispanic)
- Race:Respondent race(race)
- Race/ethnicity:Respondent race and Hispanic origin(race_ethnicity)
- Annual household income: Total income of all members of the household for the 12 months preceding the interview. Categories available from 19932021 with imputed data, starting 2015 Q1.(hincome1)
- Annual household income- imputed: Imputed income categories, starting 2017 Q1 (hincome2)
- Marital:statusRespondent marital status (marital)
- Population size:The size range for the place in which the housing unit is located, starting 1995 Q3(popsize)
- Region:Region of respondent residence. The states have been divided into four groups or census regions, starting 1995 Q3. (region)
- Household MSA: Classification of respondent residence based on the Office of Management and Budget definition of metropolitan statistical areas (MSAs) (msa)

- Household locale:Location of household based on BJS geography definitions, starting 2020 Q1(locality)
- Education level:Respondent level of education (educatn1)
- Education level:Respondent level of education, starting 2003 Q1 (educatn2)
- Veteran status:Respondent veteran status, starting in 2017(veteran)
- Citizenship status:Respondent citizenship status, starting 2017 Q1(citizen)

- Aggregate type of crime: Aggregate type of crime; violent crime includes all rape, sexual assault, robbery, assault (newcrime)
- Type of crimeType of crime(newoff)
- aggravated assault (seriousviolent)
- Reporting to police Specifies whether the crime was reported to police (notify)
- Victim servicesSpecifies whether victims received any help or advice from victim service agencies (vicservices)
- Location of crimeSpecifies where the victimization occurred(locationr)
- Victim-offender relationship Classification of respondent relationship to the offender for crimes involving direct contact (direl)

- Presence of weapon Specifies whether a weapon was present during the victimization (weapon)
- Weapon categoryType of weapon(weapcat)
- InjurySpecifies whether bodily hurt or damage was sustained by a victim as the result of criminal victimization (injury)
- Type of injurySpecifies the type of bodily hurt or damage sustained by respondent (serious)
- Medical treatment for physical injuries Specifies whether respondent received medical treatment for injuries from crime (treatment)
- Offender age:Offender age (offenderage)
- Offender sex:Offender sex (offendersex)
- Offender race/ ethnicity Offender race/ ethnicity, starting 2012 Q1(offtrace)
- Victimization weight Annual victimization weight (wgtviccy)
- SeriesSpecifies whether incident is a series crime (series)
- Series adjusted victimization weight Series adjusted victimization weight (newwgt)

Data Cleaning

Cleaning is done by changing non descriptive column names to descriptive ones, and by changing all numeric values in a category variable to a category(string).

2. NIBRS Reported Offense Count Data

Dataset Description

A more recent style, NIBRS, offers an incident-based perspective on crime. It contains facts about each crime, such as the time of day the event happened, the demographics of the perpetrators and victims, the relationships that are known to exist between the offenders and victims, and several additional information regarding where and how crime is committed. Personally identifying information (PII) about the perpetrators or victims is not included in either format. The FBI intends to switch to the NIBRS format for all crime reporting by 2021, even though many agencies still send SRS data.

Dataset Collection

The data generated by FBI Crime Data API requires state abbreviation, offense, and the key. After generating a key for the FBI Crime Data API.

Data Fields Categories

1) Offense:

All offenses are defined in this link and categorized in this link. We have found that the offense categories are representative of all offenses. Hence, the used categories are: -

- 'animal-cruelty'
- 'arson'
- 'assault-offenses'
- 'bribery'
- 'burglary-breaking-and-entering'
- 'counterfeiting-forgery'
- 'destruction-damage-vandalism-of-property'
- 'drugs-narcotic-offenses'
- 'embezzlement'
- 'extortion-blackmail'
- 'fraud-offenses'
- 'gambling-offenses'
- 'homicide-offenses'
- 'human-trafficking-offenses'
- 'kidnapping-abduction'
- 'larceny-theft-offenses'
- 'motor-vehicle-theft'
- 'pornography-obscence-material'
- 'prostitution-offenses'
- 'robbery'
- 'sex-offenses'
- 'stolen-property-offenses'
- 'weapon-law-violation'

2) State: All state abbreviations were extracted from the states API.

3) Year: Years of records from 1991 to 2021.

4) Count: Frequency of each offense category for each state and year.

Data Cleaning

The dataset contains 36309 records representing the offense categories at each state across all years from 1991-2021. Each one has a count of offenses that occurred at that time and state.

- There were no nulls.
- All state abbreviations were mapped to full state names..

3. Recidivism Data for the State of Georgia [2013-2015]

Dataset Description

The Georgia Department of Community Supervision and the Georgia Crime Information Center have provided data on individuals who were released from prison in Georgia and placed under the supervision of the Department of Community Supervision between 2013 and 2015. The data includes demographic information, details about their prison and parole cases, prior supervision history, and information about their supervision activities such as employment, program attendance, and any violations. The Georgia Crime Information Center also provided data on the individuals' criminal histories, including arrests and convictions prior to their prison time, and information on any new arrests within three years of the start of their parole supervision. This data includes all charges from an arrest episode, with the exception of domestic violence and gun charges which are counted across all episodes.

Data Fields Categories

1) Unique Person:

- ID (Unique Person ID)

2) Supervision Case Information:

- Gender (M=Male/F=Female)
- Race (Black or White),
- Age_at_Release (Age Group at Time of Prison Release (18-22, 23-27, 28-32 33-37, 38-42, 43-47, 48+)
- Residence_PUMA* (Residence US Census Bureau PUMA Group* at Prison Release)
- Gang_Affiliated (Verified by Investigation as Gang Affiliated)
- Supervision_Risk_Score_First (First Parole Supervision Risk Assessment Score) (1-10, where 1=lowest risk)
- Supervision_Level_First (First Parole Supervision Level Assignment) (Standard, High, Specialized)

3) Prison Case Information:

- Education_Level (Education Grade Level at Prison Entry) (< high school, High School diploma, at least some college)
- Dependents (# Dependents at Prison Entry) (0, 1, 2, 3+)
- Prison_Offense (Primary Prison Conviction Offense Group) (Violent/Sex, Violent/Non-Sex, Property, Drug, Other)
- Prison_Years (Years in Prison Prior to Parole Release) (< 1, 1-2, 2-3, 3+)

4) Prior Georgia Criminal History

- Prior_Arrest_Episodes_Felony (# Prior GCIC Arrests with Most Serious Charge=Felony)
- Prior_Arrest_Episodes_Misdemeanor (# Prior GCIC Arrests with Most Serious Charge=Misdemeanor)
- Prior_Arrest_Episodes_Violent (# Prior GCIC Arrests with Most Serious Charge=Violent)
- Prior_Arrest_Episodes_Property (# Prior GCIC Arrests with Most Serious Charge=Property)
- Prior_Arrest_Episodes_Drug (# Prior GCIC Arrests with Most Serious Charge=Drug)
- Prior_Arrest_Episodes_PPViolationCharges (# Prior GCIC Arrests with Probation/Parole Violation Charges)

- Prior_Arrest_Episodes_DomesticViolenceCharges (Any Prior GCIC Arrests with Domestic Violence Charges)
- Prior_Arrest_Episodes_GunCharges (Any Prior GCIC Arrests with Gun Charges)
- Prior_Conviction_Episodes_Felony (# Prior GCIC Felony Convictions with Most Serious Charge=Felony)
- Prior_Conviction_Episodes_Misdemeanor (# Prior GCIC Convictions with Most Serious Charge=Misdemeanor)
- Prior_Conviction_Episodes_Violent (Any Prior GCIC Convictions with Most Serious Charge=Violent)
- Prior_Conviction_Episodes_Property (# Prior GCIC Convictions with Most Serious Charge=Property)
- Prior_Conviction_Episodes_Drug (# Prior GCIC Convictions with Most Serious Charge=Drug)
- Prior_Conviction_Episodes_PPViolationCharges (Any Prior GCIC Convictions with Probation/Parole Violation Charges)
- Prior_Conviction_Episodes_DomesticViolenceCharges (Any Prior GCIC Convictions with Domestic Violence Charges)
- Prior_Conviction_Episodes_GunCharges (Any Prior GCIC Convictions with Gun Charges)

5) Prior Georgia Community Supervision History:

- Prior_Revocations_Parole (Any Prior Parole Revocations)
- Prior_Revocations_Probation (Any Prior Probation Revocations)

6) Georgia Board of Pardons and Paroles Conditions of Supervision:

- Condition_MH_SA (Parole Release Condition = Mental Health or Substance Abuse Programming)
- Condition_Cog_Ed (Parole Release Condition = Cognitive Skills or Education Programming)
- Condition_Other (Parole Release Condition = No Victim Contact or Electronic Monitoring or Restitution or Sex Offender Registration/Program)

7) Supervision Activities:

- Violations_ElectronicMonitoring (Any Violation for Electronic Monitoring)
- Violations_InstructionsNotFollowed (Any Violation for Not Following Instructions)
- Violations_FailToReport (Any Violation for Failure to Report)
- Violations_MoveWithoutPermission (Any Violation for Moving Without Permission)
- Delinquency_Reports (# Parole Delinquency Reports)
- Program_Attendances (# Program Attendances)
- Program_UnexcusedAbsences (# Program Unexcused Absences)
- Residence_Changes (# Residence Changes/Moves (new zip codes) During Parole)
- Avg_Days_per_DrugTest (Average Days on Parole Between Drug Tests)
- DrugTests_THC_Positive (% Drug Tests Positive for THC/Marijuana)
- DrugTests_Cocaine_Positive (% Drug Tests Positive for Cocaine)
- DrugTests_Meth_Positive (% Drug Tests Positive for Methamphetamine)
- DrugTests_Other_Positive (% Drug Tests Positive for Other Drug)
- Percent_Days_Employed (% Days Employed While on Parole)
- Jobs_Per_Year (Jobs Per Year While on Parole)
- Employment_Exempt (Employment is Not Required (Exempted))

8) Recidivism Measures:

- Recidivism_Within_3years (New Felony/Mis Crime Arrest within 3 Years of Supervision Start)
- Recidivism_Arrest_Year1 (Recidivism Arrest Occurred in Year 1)
- Recidivism_Arrest_Year2 (Recidivism Arrest Occurred in Year 2)
- Recidivism_Arrest_Year3 (Recidivism Arrest Occurred in Year 3)

Data Collection and Cleaning:

- The Dataset was downloaded from the link.
- It contains 25835 samples, each of them has 54 features.
- The null values were checked and they are shown below:

Feature	Null Counts
Gang_Affiliated	3167
Supervision_Risk_Score_First	475
Supervision_Level_First	1720
Prison_Offense	3277
Avg_Days_per_DrugTest	6103
DrugTests_THC_Positive	5172
DrugTests_Cocaine_Positive	5172
DrugTests_Meth_Positive	5172
DrugTests_Other_Positive	5172
Percent_Days_Employed	462
Jobs_Per_Year	808

- Categorical terms that have null values were filled with the mode value.
- Numerical terms that have null values were filled with the mean value.

4. Firearm Laws per State

Dataset Description

The Firearm laws dataset contains all the laws and regulations for possessing Guns and firearms applied in all 50 U.S states throughout the period starting from 1991 until 2020.

Data Collection

The dataset can be downloaded from this link. For a better understanding of the dataset, there is a code book which explains the laws in detail and how they are encoded in the dataset.

Part 2: Explanatory Analysis

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

- It is related to **NCVS and NIBRS** datasets. Hence, we combined both datasets and completed the task using the combined dataset.
- By aggregating the sum of the count of offenses from different states, it was found that the top five most frequent offense categories were [**"larceny-theft-offenses", "assault-offenses","destruction-damage-vandalism-of-property","drugs-narcotic-offenses", and "burglary-breaking-and-entering"]**
- Columns of interest **'Offense', 'Year', and 'Count'** were selected and filtered to contain the top 5 offense categories in frequency.
- This count was normalized to the total number of US citizens in each year, but due to the linear increase, it did not affect the crime rates distributions across all years a lot.

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

- It is related to **NIBRS** datasets.
- Violent crimes were filtered out of the whole dataset on having offense categories of **['assault-offenses','homicide-offenses','robbery','kidnapping-abduction','sex-offenses'].**
- By aggregating the sum of the count of offenses from different states, we had the total count of violent crimes in each state and year.

- Columns of interest **‘State’, ‘Year’, and ‘Count’** were selected and filtered from the whole dataset and from the violent crimes dataset, but some states did not have violent crimes which were not present in the violent crimes dataset. Hence, I added the missing states in each year and put 0 at the **‘Count’**.
- This count was normalized to the total number of crimes in each year. Then, an average of crime rates was taken along all years for each state.

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

- It is related to **NIBRS** datasets.
- Violent crimes were filtered out of the whole dataset on having offense categories of **['assault-offenses','homicide-offenses','robbery','kidnapping-abduction','sex-offenses'].**
- Also, the homicide offenses were filtered out of the whole dataset.
- By aggregating the sum of the count of offenses from different states and adding both rates in a single dataframe, we had the total count of violent crimes and homicide crimes in each year.
- The data was not balanced, so it was normalized by the total number of crimes in each year.
- Columns of interest **‘State’, ‘Year’, and ‘Count’** were selected and filtered from the whole dataset and from the violent crimes dataset, but some states did not have violent crimes which were not present in the violent crimes dataset. Hence, I added the missing states in each year and put 0 at the **‘Count’**.
- This count was normalized to the total number of crimes in each year. Then, an average of crime rates was taken along all years for each state.
- It was found that Minnesota had the highest violent crime rates.

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

- It is related to the NCVS dataset.
- Victims of non-fatal crimes were filtered first.
- Columns of interest ‘**race_ethnicity’,'age','sex',’count’** were selected and filtered then normalized to the total count of victims by aggregating the columns of interest.
- A data frame containing all combinations of the three demographic variables was formed having the non-fatal crime percentage rates .

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

- It is related to the NCVS dataset.
- Offenders of non-fatal crimes were filtered first.
- Columns of interest ‘**race_ethnicity’,'age','sex',’count’** were selected and filtered then normalized to the total count of victims by aggregating the columns of interest.
- A data frame containing all combinations of the three demographic variables was formed having the non-fatal crime percentage rates.

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

- It is related to the NCVS dataset.
- Columns of interest ‘**Respondent_education_level’, 'household_with_imputed_data', ‘counts’** were selected and filtered then normalized to the total population of each combination of the two demographic variables after aggregating the columns of interest.
- A data frame containing all combinations of the two demographic variables was formed having the non-fatal crime percentage rates.

Part 3: Answering 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?

   - It is related to **NCVS**dataset.
   - Columns of interest **'reported_to_police', 'type_of_crime'** were selected.
   - Counts columns were added to the dataframe to aggregate the type of crimes.
   - **'Simple assault '** crimes were found to be the least reported.

   - It is related to **NCVS**dataset.
   - Columns of interest **'reported_to_police', 'relation_with_offender'** were selected.
   - Counts columns were added to the dataframe to aggregate the type of crimes.
   - **'Stangers '** relationship were found to be the most reported to the police, while the number of '**well-known'** relationships increased in number when it comes to not reported to the police station

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

   - It is related to **NCVS**dataset.
   - Columns of interest **'age','sex','race_ethnicity' ,'aggregate_type_of_crime'**were selected.
   - **'Aggregate_type_of_crime' set to violent crimes**
   - Counts and rates columns were added to the dataframe to aggregate the type of crimes.

    - Rates are determined by: the counts in the victimization dataset/ the total counts in the population dataset of the same grouping.
    - **'18-24 years females non hispanic with more than one race'** demographic was found to be the with the highest risk to be violently victimized, while **‘65 or older females hispanic’** demographic are at least at risk to be violently victimized.

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?

- It is related to the NCVS dataset.
- Injured victims of non-fatal crimes were filtered firstly.
- Columns of interest **'age','sex','race_ethnicity'** were selected and filtered for victims who had medical attention at the scene **(‘treatment’=='Treated at scene, home, medical office, or other location')** then normalized to the total population of each combination of the the columns of interest after aggregating them.
- A data frame containing all combinations of the columns of interest was formed having the non-fatal crime percentage rates.

4. Which class of crimes is associated with the highest rate of same-offense-recidivism?

- It is related to **Recidivism data for the state of Georgia [2013-2015]** dataset.
- Columns of interest **'Prison_Offense', 'Recidivism_Within_3years'** were selected.
- The null values were filled using the mode value.
- Counts columns were added to the dataframe to aggregate the number of offenses from each type **['Drug', 'Violent/Non-Sex', 'Property', 'Other', 'Violent/Sex']** whose offender reoffended another offense within 3 years.
- **'Property'** crimes were found to be associated with the highest rate of offense-recidivism.

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

- The question is related to **Recidivism data for the state of Georgia [2013-2015]** Dataset.
- Columns of interest **Age_at_Release, 'Recidivism_Within_3years'** were selected.
- The null values were filled using the mode value.
- Counts columns were added to the dataframe to aggregate the number of offenders from each age category **['43-47', '33-37', '48 or older', '38-42', '18-22', '23-27', '28-32']** who reoffended another offense within 3 years.
- **'23-27'** offenders are more likely to reoffend than those who are from age ranges.
- Hence, Prisoners who are younger at the time of release are more likely to reoffend than those who are older.

Part 4: Hypothesis Testing

Technical Approach for The Data Analysis

A. Categorization: in this step, we started categorizing the laws in the dataset according to their primary categories. The categories included the following:
   - Dealer regulations
   - Buyer regulations
   - High-risk guns prohibition
   - Background checks
   - Ammunition regulations
   - Possession regulations
   - Concealed carry permitting
   - Assault weapons and large-capacity magazines
   - Child access prevention
   - Gun trafficking
   - Stand your ground

- Preemption
- Domestic violence

B. Defining strictness: the column strictness represents how strict the state is when it comes to applying the laws. Strictness is calculated by summing over all the laws categories applied by the state.

C. Calculating the average strictness per state through the whole period (1991-2020)

D. Categorizing the states: the states are splitted into two categories, heavily restricted and not-heavily restricted. If the average strictness per state exceeds the mean of the average strictness for all states, the state is considered heavily restricted. Else, the state is considered not_heavily restricted.

Hypothesis Testing

A. Testing Claim 1

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

Steps: the states are divided into two categories, heavily strict states and less strict states and the threshold that decides will be the mean of the average strictness per state across all years.

Null hypothesis H0: there will be no difference in the average crime rate between heavily strict states and the other states.

The test: the test used here is the T-test as we are comparing the means of two different groups.

Test result: the p-value = 0.544, which means that the null hypothesis cannot be rejected. Therefore, we can deduce that American society has violent behavior regardless of laws being strict or not.

B. Testing Claim 2

Claim: Black people are assigned a high risk score compared to white people.

Steps: the criminals from the dataset "Recidivism data for the state of Georgia [2013-2015]"will be divided into two categories; Blacks and Whites.

Null hypothesis H0: there will be no difference in the Supervision Risk Score between Black and White people.

The test: the test used here is the T-test as we are comparing the means of two different populations.

Test result: the p-value = 0.999, which means that the null hypothesis cannot be rejected. Therefore, we cannot say that the blacks are assigned a high risk score compared to the whites.

Part 5: Regression Analysis

Task Description

- Using The recidivism in Georgia dataset to fit a regression model that predicts the Offender’s supervision risk score based on :
    a) All prior convictions.
    b) Offender’s race.
    c) Offender’s gang affiliation.
    d) Offender’s age at release.

Approach

- Columns of interest were selected
- Columns of interest = ['Race', 'Age_at_Release', 'Gang_Affiliated', 'Supervision_Risk_Score_First', 'Prior_Conviction_Episodes_Felony', 'Prior_Conviction_Episodes_Misd', 'Prior_Conviction_Episodes_Viol', 'Prior_Conviction_Episodes_Prop', 'Prior_Conviction_Episodes_Drug',, 'Prior_Conviction_Episodes_PPViolationCharges', 'Prior_Conviction_Episodes_DomesticViolenceCharges', 'Prior_Conviction_Episodes_GunCharges']

- One Hot Encoding was performed on the non-numerical features, where the last column was dropped.
- The encoded data was passed to fit an Ordinary Least Square (OLS) model from the statsmodels library.
- An intercept column was added in order for the model to find $\boldsymbol{\beta}_0$.
- The model's coefficients and p-values were obtained after fitting the model.
- Residuals Analysis were performed in order to find if error terms (residuals) follow a normal distribution with zero mean or not.
- QQ-Plot was used to assess if the residuals follow a normal distribution.
- Residuals-vs-Fits Plot was used to check if the regression function is linear or not.
- Residual-vs-Orders Plot was used to check the dependence and correlation between error terms.
- Good and Bad Predictors were obtained using the p-value. If the p-value < 0.05, the predictor would be considered as good predictors, while variables that have p-value > 0.05 are bad predictors.
- Correlation matrix was used to check the highly correlated features.
- A threshold value (0.5) was set such that variables that had correlation value larger than that threshold value would be considered as highly correlated variables.

Part 6: Bonus Task

Task Description:

- 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.

Approach:

- All columns of the dataset were selected except 5 columns **['ID', 'Recidivism_Arrest_Year1', 'Recidivism_Arrest_Year2', 'Recidivism_Arrest_Year3']**

- The columns that carry information about Recidivism were dropped because **‘Recidivism_Within_3years’** would be considered as the target value and all of these 4 columns are highly correlated since they nearly represent the same information.
- Ordinal Encoding was performed on the non-numerical features in order not to increase the number features as in the case of using One Hot Encoding (Curse of Dimensionality).
- The encoded data was then splitted into training and testing dataset based on the **‘Training_Sample’** column.
- **‘Training_Sample’** column has a value of ‘1’ for the columns that would be used for training, while having a value of ‘0’ for testing rows.
- **‘Recidivism_Within_3years’** column was selected to be the target feature, while all the other columns are considered as input features.
- Logistic Regression, Logistic Regression with Feature Selection: RFE, Random Forest, AdaBoost, and Bagging models were used for classification of **‘Recidivism_Within_3years’**.
- Metrics such as Accuracy, Precision, Recall, and F1-score were reported after predicting the testing dataset by comparing the predicted and true values.
- The model with the highest accuracy was Bagging and its accuracy was: 73.857 %