Zewail City of Science and Technology

[CIE 457]
4th Year – fall
2022

Course:Statistical Inference and Data Analysis

Technical Report

Code:

CIE 457

Project

Analyzing U.S. Crime Data

Submitted To: Dr. Mahmoud Abdelaziz,

Eng. Asmaa Ismail, Eng. Anhar Abdelmotaleb

Name:	Mohamed Helmy
ID:	201900859
Name:	Youssef Mahmoud Mohamed
ID:	201901093
Name:	Hossam Ashraf
ID:	201901898

Hierarchy:

1.Data Collection & cleaning

2.Exploratory Analysis

3.Answering Questions

4.Hypothesis Testing

5.Regression Analysis

6.Bonus Task

1.1 Data Collection:

That was the first step in the project getting data for multiple datasets with APIs was done for various datasets like

- **The national crime victimization survey (NCVS) data:**
    - **Population dataset:**
        - Dataset collected using a survey in various years and quarters (1993–2021) having the respondent personal data
    - **Victimization dataset**
        - Dataset collected using a survey in various years and quarters (1993–2021) having respondents detailed data for crimes they have faced
- **NIBRS Reported offense count data**
    - **Offense count dataset**
        - This dataset has many states with various crime types and count committed in each state for different years
- **Recidivism data for the state of Georgia [2013-2015]**
    - **Recidivism dataset**
        - This dataset has information about people who got in prison with more details about their crime, respond during the duty time, and many other information

No need for API for the following dataset

Firearm laws per state

- This dataset has the applied firearms laws in each state

Firearm Codebook

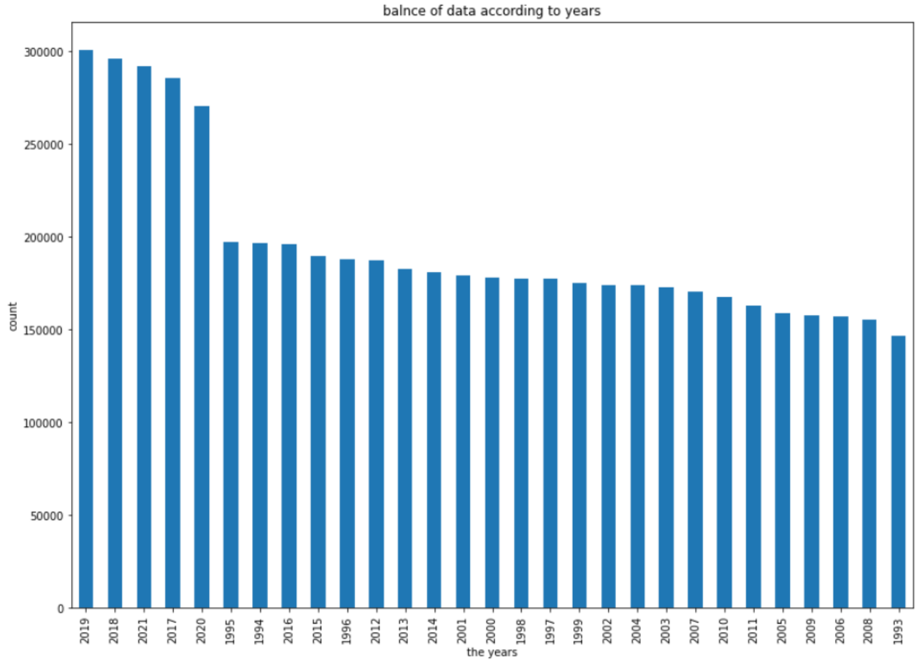
- This dataset is constructed using the Firearm codebook to be able to refer each law to its category and subcategory

1.2 Data Cleaning:

- Codebook Well Reading
- Changing numerical values to categorical variables
- Changing column names to be representable
- Drop unwanted data
- Dealing with null values

Population dataset:

- The first noticed point was that the amount of responses per year wasn’t balanced:



- Changing numeric values to categorical was done for the coming columns:
  - **#ager --> age ranges**

    ```
    specified
    #[1 2 3 4 5 6] -- > ['12-17','18-24','25-34','35-49','50-64','65 or older']
    ```
  - **#gender**

    ```
    #[1 ,2] --> [Male , Female]
    ```
  - **#Hispanic**

    ```
    #specify the Respondent Hispanic origin with its categorical values
    # [1 2 88] --> [hispanic non hispanic residue]
    ```
  - **# Respondent race**

    ```
    # [1 2 3 4 5] -->['White','Black','American Indian/Alaska Native','Asian/Native Hawaiian/Other Pacific Islander','More than one race']
    ```
  - **#Total income of all members of the household for the 12 months preceding the interview. Categories available from 1993-2021**

    ```
    #[ 1  2  3  4  5  6  7 88] -- > ['Less than $7,500','$7,500 to $14,999','$15,000 to $24,999','$25,000 to $34,999','$35,000 to $49,999','$50,000 to $74,999','$75,000 or more','Unknown']
    ```
  - **#Respondentmarital status**

#[1 2 3 4 5 88] --> ['Never married','Married','Widowed','Divorced','Separated','Residue']

- **#Region of respondent residence. The states have been divided into four groups or census regions,starting 1995 Q3.**

  # [-1 1 2 3 4] --> ['Invalid until 1995 Q3','Northeast','Midwest','South','West']

- **#The size range for the place in which the housing unit is located, starting 1995 Q3**

  #[-1 0 1 2 3 4 5] -->['Invalid until 1995 Q3','Not a place','Under 100,000','100,000-249,999','250,000-499,999','500,000-999,999','1 million or more']

- **#Classification of respondent residence based on the Office of Management and Budget definition of metropolitan statistical areas (MSAs)**

  # [1 2 3] --> ['Principal city within MSA','Not part of principal city within MSA','Outside MSA']

- **# level of education**

  # [1 2 3 4 5 88] --> ['No schooling','Grade school','Middle school','High school','College','Residue']

- **# Respondent level of education, starting 2003 Q1**

  # [-1 1 2 3 4 5 6 7 8 98] -->['Invalid until 2003 Q1','No schooling','Grade school','Middle

school','Some high school','High school graduate','Some college and associate degree','Bachelor’s degree','Advanced degree','Residue']

- **#Imputed income categories, starting 2017 Q**
  # [-1 1 2 3 4 5] --> ['Invalid until 2017 Q1','Less than $25,000','$25,000 to $49,999','$50,000 to $99,999','$100,000 to $199,999','$200,000 or more']
- **# Location of household based on BJS geography definitions, starting 2020 Q1**
  # [-1 1 2 3] --> ['Invalid until 2020 Q1','Urban','Suburban','Rural']
- **# Respondent veteran status, starting in 2017**
  # [-2 -1 0 1 8 9] -->['Invalid until 2017 Q1','Under age 18','Not a veteran','Veteran','Residue','Out of universe']
- **# Respondent citizenship status, starting 2017 Q1**
  # [-1 1 2 3 8 9] --> ['Invalid until 2017 Q1','Born U.S. citizen','Naturalized citizen','Non-U.S. Citizen','Residue','Out of universe']

- Dropping the unwanted features:
    - #Drop the year as it is preserved in the column of year and quarter
    - #Drop educatn1 as educatn2 has wider cases including educatn1 cases in addition educ2 is starting from 2003 which neglect all the previous years
    - #Drop Race/ethnicity as it is a mix between Hispanic origin and race columns
    - #Drop hincome2 as it is starting from 2017 and neglect all other --> when get it is unique values it doesn't add any info
- Changing column names to be representable:
    - **{'msa':'metropolitan_statistical_areas', 'idper':'id' , 'yearq':'year_quarter' , 'ager':'age_range','hincome1':'house_hold_income','educatn1':'education_level'}**

Victimization dataset:

The dataset has common features from the population dataset like the victim's demographics, the columns were changed from numeric to categorical just like that mentioned in the population dataset

Additional columns:

- **# offense type**
  # [1 2 3 4 5] --> ['Rape/sexual assault','Robbery','Aggravated assault','Simple assault','Personal theft/larceny']
- **# reported the crime to the police**
  # [1 2 3 8] --> ['Yes','No','Do not know','Residue']

- **# Victim services**
  # [1 2 3 8] --> ['Yes','No','Do not know','Residue']
- **# Location of crime**
  # [1 2 3 4 5] --> ['At or near victim’s home','At or near friend’s,neighbor’s, or relative’s home','Commercial place,parking lot, other public area','School','Other location']
- **# offender and victim relationship**
  # [1 2 3 4 5 6] --> ['Intimates','Other relatives','Well known/casual acquaintance','Strangers','Do not know relationship','Do not know number of offenders']
- **# Offender weapon category**
  # [0 1 2 3 4 5] --> ['No weapon','Firearm','Knife','Other type weapon','Type weapon unknown','Do not know if offender had weapon']
- **# injury type**
  # [1 2 3 4 88] --> ['No injury','Serious injury','Minor injury','Rape w/o other injuries','Residue']
- **# Medical treatment**
  # [0 2 3 88] --> ['Not injured','Not treated','Treated at scene,home, medical office,or other location','Don’t know','Residue']
- **# Offender age**
  # [1 2 3 4 5 88] --> ['11 or younger','12-17','18-29','30 or

older','Multiple offenders of various ages','Residue']

- **# Offender gender**
  # [1 2 3 4 88] --> ['Male','Female','Both male and female offenders','Unknown','Residue']
- **# Offender race**
  # [-1 1 2 3 4 5 6 7 10 11] --> ['Invalid until 2012 Q1','Non-Hispanic white','Non-Hispanic black','Non-Hispanic','American Indian/Alaska Native','Non-Hispanic Asian/Native Hawaiian/Other Pacific Islander', 'Non-Hispanic more than one race','Hispanic','Unknown race/ethnicity''Mixed race group of offenders','Unknown number of offenders']
- **# Specifies whether incident is a series crime**
  # [1 2] --> ['Not a series crime','Series crime']

- Dropping the unwanted features:
  - #Drop the year as it is preserved in the column of year and quarter
  - #Drop educatn2 as educatn1 is valid in all the years
  - #Drop Race/ethnicity as it is a mix between Hispanic origin and race columns
  - #Drop hincome2 as hincome1 is more representative to the data and have more valid rows
  - #Drop newcrime, serious violent as they can be concluded from the crime type 'newoff' column

  -
    ```
    #Drop presence of weapon 'weapon' column as it can be concluded from the Weapon category 'webcat' column
    ```

  -
    ```
    #Drop presence of Injury 'injury' column as it can be concluded from the Type of injury 'serious' column
    ```
- Changing column names to be representable:

  -
    ```
    {'msa':'metropolitan_statistical_areas', 'idper':'id' , 'yearq':'year_quarter' , 'ager':'age_range','hincome1':'house_hold_income','educatn1':'education_level', 'direl':'victim_offender_relation','newoff':'offense_type','locationr':'crime_location','weapcat':'weapon_category','serious':'injury_type', 'wgtviccy':'victmization_weight','newwgt':'series_adjusted_victmization_weigt'}, inplace = True)
    ```

Recidivism dataset:

- Dropping the unwanted features:

  -
    ```
    #Instead of dropping the features we chose the features of interest to be the following
    ```

  -
    ```
    interest_features = ['gender','recidivism_arrest_year1','recidivism_arrest_year2','recidivism_arrest_year3','dependents','education_level','prison_years','prison_offense','recidivism_within_3y
    ```

```
ears','race', 'age_at_release',
'gang_affiliated'
,'prior_conviction_episodes','prior_convict
ion_episodes_1','prior_conviction_episodes_
2','prior_conviction_episodes_3','prior_con
viction_episodes_4','prior_conviction_episo
des_5','prior_conviction_episodes_6','prior
_conviction_episodes_7','supervision_risk_s
core_first']
```

- Filling the nulls in the `'gang_affiliated'` and `'prison_offense'` columns using `'ffill'` method as the rows are large numbers about 3000 rows
  And dropped the nulls in `supervision_risk_score_first` as it is small sample size

2 Exploratory Analysis

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

```
#offenses count dataset is used
#get the highest 5 unique offense categories
#in 5 sub plot:
#counts bar plot for each category rate in all years
```

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

```
#offenses count dataset is used
```

```
#classify crimes to violent and non-violent crimes
#Using group by method:
  ● #get the counts of violent crimes per state
  ● #get the counts of total crimes per state
#Plotting:
  • #show the percentage of the violent crime in each
    state
  • # compare between violent and nonviolent crimes to
    avoid the first graph bias
```

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

```
# group by year and offense name and counts values in each
year

# plot the rate of violent crimes per year over  all years

# clarify the homicide_offenses
# plotting rate of homicide per year over all years after
using group by method
```

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

```
#calculating the frequency of crimes for each victim
demographic in the victimization dataset
#get the total number of each sub-category for demographics
in the population dataset
#normalize the calculated frequencies
#plot 4 subplots for the relationship between non-fatal
crimes and victim’s demographics
```

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

```
#calculating the frequency of crimes for each offender
demographic in the victimization dataset
```

```
#Didn’t perform normalization as the offender’s
demographics has categories more than that in the
population dataset
#plot 3 subplots for the relationship between non-fatal
crimes and offender’s demographics
```

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

```
#calculating the frequency of crimes for each offender
demographic in the victimization dataset
#Didn’t perform normalization as the offender’s
demographics has categories more than that in the
population dataset
#plot 3 subplots for the relationship between non-fatal
crimes and offender’s demographics
```

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? (reported: ie, police notified at time of occurrence)**

```
#a)Which type of non-fatal crime is the most
under-reported?
#get nonfatal crimes + not reported + group by crime
```

Personal theft is found to be the most under-reported crime

```
#b)get reported crimes from victimization dataset and group
by the victim-offender relationship counting number of
reported crimes:
```

Strangers came into the first place, while well known got the second with a minor slight between them, so it is not a good association between them

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

```
#Grouping data by each demographic while counting the crimes.
```

```
#Normalizing the data
```

```
#Plotting the data
```

After normalization Most risk is found with:

- **Younger people (age)**
- **American/indian (race)**
- **Hispanic category is not affecting the victimization risk**
- **Gender is not affecting the victimization 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?**

```
#The percentage of treatment was first obtained for every demographic and plotted, it was observed that the percentage of treatment in injury cases is nearly equal for all sub-categories in each demographic.
```

```
#At the end, data got grouped by demographics, and treatment counting treatment case for every subcase, then normalized
```

After normalization demographics that most likely to receive medical attention at the scene:

- **Younger people (age)**
- **American/indian (race)**

Gender is not affecting the victimization risk

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?**

   #Using the recidivism dataset, the data was grouped by crime type counting 'recidivism_within_3years', then plotting a barplot.
   Property crime was found to be the most crime that a prisoner leave the prison and re-enter it within 3 years from the release date

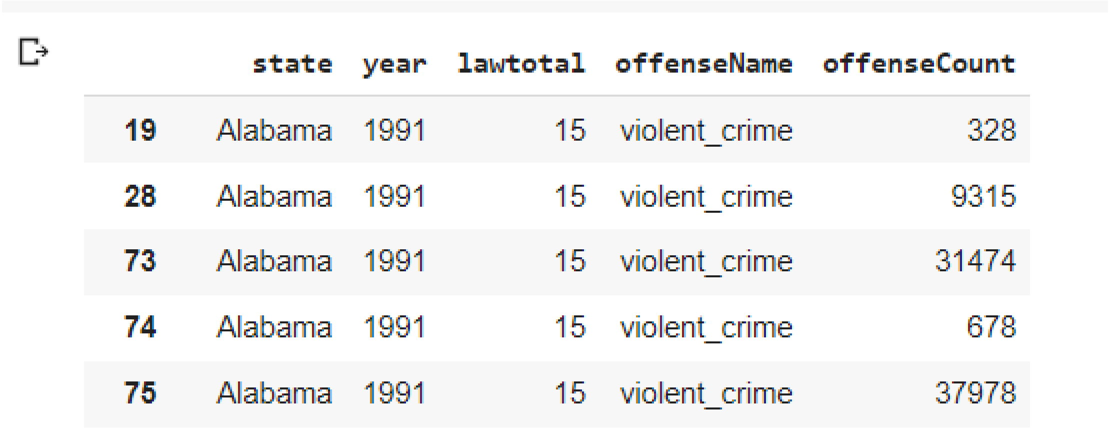
5. **Are prisoners who are younger at the time of release more or less likely to reoffend than those who are older?**

   #Using the recidivism dataset,the data was grouped by prisoner age at release counting 'recidivism_within_3years', then plotting a bar plot after normalization.
   It was found that number of returned prisoners decrease as they get older, so yes; prisoners who are younger at the time of release more likely to reoffend than those who are older

4 Hypothesis Testing

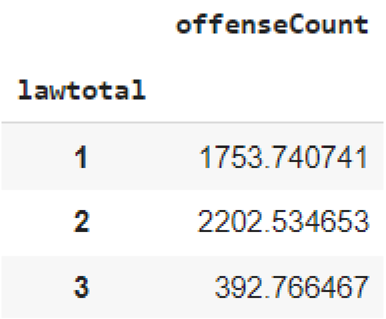
Since the claims we work on are related to (number of laws, violent crime rates, year of each combination of them along with the state) :

1. **Select the two data frames of interest, the one relating each state with number of laws applied in each year, and the other dataframe counting the occurrences of each type of crime in each state per year.**
2. **Merging both as they have in common “State” and “year” columns**
3. **Categorize all crimes into 2 categories only Violent and non-violent then selecting violent only**



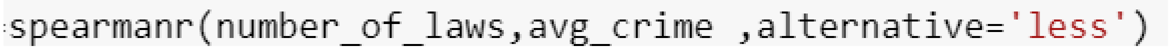
	state	year	lawtotal	offenseName	offenseCount
19	Alabama	1991	15	violent_crime	328
28	Alabama	1991	15	violent_crime	9315
73	Alabama	1991	15	violent_crime	31474
74	Alabama	1991	15	violent_crime	678
75	Alabama	1991	15	violent_crime	37978

4. **<u>For Claim 1</u> studying relation between number of laws and crime rates, drop the state, year, and offenseName columns as they are of no interest.**
5. **Then group by number of applied laws and calculate average of crime rates at each number of laws**



	offenseCount
lawtotal	
1	1753.740741
2	2202.534653
3	392.766467

6. **<u>For the correlation:</u> import spearmans function from scipy.stats then calculating the correlation between the two variables to test the hypothesis of either crime rates remain same of increase (Null-hypothesis) or decrease with number of laws (Alternative hypothesis)**



```
spearmanr(number_of_laws,avg_crime ,alternative='less')
```

7. **Plot a scatter plot between the 2 variables.**
8. **For claim 2: studying crime rates over years, drop the state, lawtotal, and offenseName columns as they are of no interest.**
9. **Then group by year and calculate average of crime rates at each year**
10. **Repeat the same steps as the above claim**

5 Regression Analysis:

```
Offender’s supervision risk score based on :
# - All prior convictions.
# - Offender’s race.
# - Offender’s gang affiliation.
# - Offender’s age at release.
```

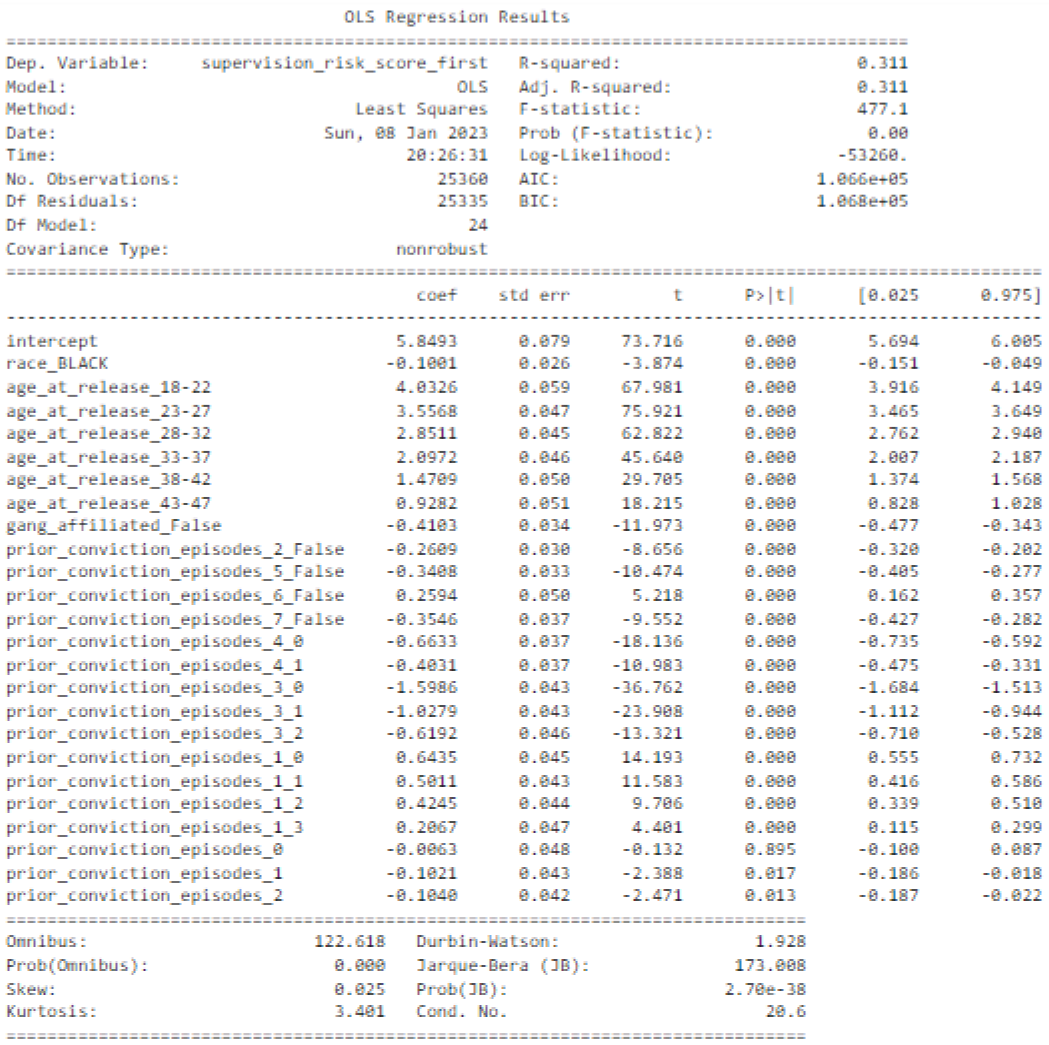
So, the feature of interest are:

```
interest_features = ['supervision_risk_score_first',
'race','age_at_release', 'gang_affiliated'
,'prior_conviction_episodes','prior_conviction_episodes_1','prior_conviction_episodes_2','prior_conviction_episodes_3','prior_conviction_episodes_4'
,'prior_conviction_episodes_5','prior_conviction_episodes_6','prior_conviction_episodes_7']
```

The data are categorical, we dealt with this in two different ways →
1- OneHot encoding and drop the last column to avoid the multicollinearity as it could be predicted 100% from the other columns

2- label encoding : encode each variable with its number for ex. 3 or more with 3 and False with 0 and true with 1

Then, the intercept is added and stats model of least squared is used to fit on the data
The result from ONE HOT ENCODING:



```
                            OLS Regression Results
========================================================================================
Dep. Variable:     supervision_risk_score_first   R-squared:                       0.311
Model:                                      OLS   Adj. R-squared:                  0.311
Method:                           Least Squares   F-statistic:                     477.1
Date:                          Sun, 08 Jan 2023   Prob (F-statistic):               0.00
Time:                                  20:26:31   Log-Likelihood:                -53260.
No. Observations:                         25360   AIC:                         1.066e+05
Df Residuals:                             25335   BIC:                         1.068e+05
Df Model:                                    24
Covariance Type:                      nonrobust
=====================================================================================================
                                        coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------
intercept                             5.8493      0.079     73.716      0.000       5.694       6.005
race_BLACK                           -0.1001      0.026     -3.874      0.000      -0.151      -0.049
age_at_release_18-22                  4.0326      0.059     67.981      0.000       3.916       4.149
age_at_release_23-27                  3.5568      0.047     75.921      0.000       3.465       3.649
age_at_release_28-32                  2.8511      0.045     62.822      0.000       2.762       2.940
age_at_release_33-37                  2.0972      0.046     45.640      0.000       2.007       2.187
age_at_release_38-42                  1.4709      0.050     29.705      0.000       1.374       1.568
age_at_release_43-47                  0.9282      0.051     18.215      0.000       0.828       1.028
gang_affiliated_False                -0.4103      0.034    -11.973      0.000      -0.477      -0.343
prior_conviction_episodes_2_False    -0.2609      0.030     -8.656      0.000      -0.320      -0.202
prior_conviction_episodes_5_False    -0.3408      0.033    -10.474      0.000      -0.405      -0.277
prior_conviction_episodes_6_False     0.2594      0.050      5.218      0.000       0.162       0.357
prior_conviction_episodes_7_False    -0.3546      0.037     -9.552      0.000      -0.427      -0.282
prior_conviction_episodes_4_0        -0.6633      0.037    -18.136      0.000      -0.735      -0.592
prior_conviction_episodes_4_1        -0.4031      0.037    -10.983      0.000      -0.475      -0.331
prior_conviction_episodes_3_0        -1.5986      0.043    -36.762      0.000      -1.684      -1.513
prior_conviction_episodes_3_1        -1.0279      0.043    -23.908      0.000      -1.112      -0.944
prior_conviction_episodes_3_2        -0.6192      0.046    -13.321      0.000      -0.710      -0.528
prior_conviction_episodes_1_0         0.6435      0.045     14.193      0.000       0.555       0.732
prior_conviction_episodes_1_1         0.5011      0.043     11.583      0.000       0.416       0.586
prior_conviction_episodes_1_2         0.4245      0.044      9.706      0.000       0.339       0.510
prior_conviction_episodes_1_3         0.2067      0.047      4.401      0.000       0.115       0.299
prior_conviction_episodes_0          -0.0063      0.048     -0.132      0.895      -0.100       0.087
prior_conviction_episodes_1          -0.1021      0.043     -2.388      0.017      -0.186      -0.018
prior_conviction_episodes_2          -0.1040      0.042     -2.471      0.013      -0.187      -0.022
==============================================================================
Omnibus:                      122.618   Durbin-Watson:                   1.928
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              173.008
Skew:                           0.025   Prob(JB):                     2.70e-38
Kurtosis:                       3.401   Cond. No.                         20.6
==============================================================================
```

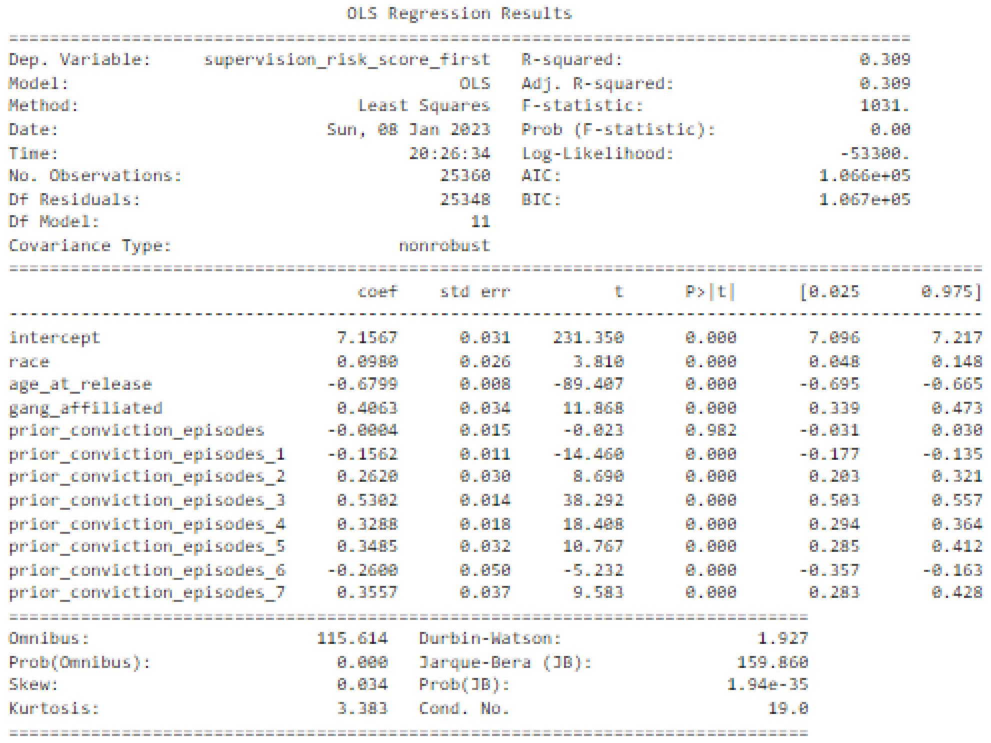
It seems the p_values are all statistically significant except for prior_convection_episodes0 with very small betas which is logically acceptable.
so , all the parameters are good except it.

#graphing the heatmap of the correlation between the variables,there is high correlation between the variables

To asses the model :
We check the r_squared which is 0.3 not a good model for prediction.However the error follows our assumptions of Homocidicity normal distribution 0 mean and specific variance (unbiased).
Check the residual vs the y_hat

Check the residual distribution against orders to check if there is a pattern of residuals with orders or not.

We repeated the same steps on label encoding and gave the same results but with different parameters: all are statistically significant except prior_conviction_episodes



```
                                OLS Regression Results
==============================================================================
Dep. Variable:     supervision_risk_score_first   R-squared:                       0.309
Model:                                      OLS   Adj. R-squared:                  0.309
Method:                           Least Squares   F-statistic:                     1031.
Date:                          Sun, 08 Jan 2023   Prob (F-statistic):               0.00
Time:                                  20:26:34   Log-Likelihood:                -53300.
No. Observations:                         25360   AIC:                         1.066e+05
Df Residuals:                             25348   BIC:                         1.067e+05
Df Model:                                    11
Covariance Type:                      nonrobust
===============================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
intercept                       7.1567      0.031    231.350      0.000       7.096       7.217
race                            0.0980      0.026      3.810      0.000       0.048       0.148
age_at_release                 -0.6799      0.008    -89.407      0.000      -0.695      -0.665
gang_affiliated                 0.4063      0.034     11.868      0.000       0.339       0.473
prior_conviction_episodes      -0.0004      0.015     -0.023      0.982      -0.031       0.030
prior_conviction_episodes_1    -0.1562      0.011    -14.460      0.000      -0.177      -0.135
prior_conviction_episodes_2     0.2620      0.030      8.690      0.000       0.203       0.321
prior_conviction_episodes_3     0.5302      0.014     38.292      0.000       0.503       0.557
prior_conviction_episodes_4     0.3288      0.018     18.408      0.000       0.294       0.364
prior_conviction_episodes_5     0.3485      0.032     10.767      0.000       0.285       0.412
prior_conviction_episodes_6    -0.2600      0.050     -5.232      0.000      -0.357      -0.163
prior_conviction_episodes_7     0.3557      0.037      9.583      0.000       0.283       0.428
==============================================================================
Omnibus:                      115.614   Durbin-Watson:                   1.927
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              159.860
Skew:                           0.034   Prob(JB):                     1.94e-35
Kurtosis:                       3.383   Cond. No.                         19.0
==============================================================================
```

The model behaves the same as the one hot encoding

The Min_Max scaling is tried on the data before fitting to see the change in the graph of residuals and the y_hat , the r squared.As expected r_squared is higher and the points are collected over the line 0 in the graph.However, the normal distribution, became biased.

6 Bonus Task

Depending on :

```
interest =
['employment_exempt','prior_arrest_episodes_drug','prior_arrest_episodes_violent','prior_arrest_episodes_misd','program_attendances','violations_instruction','gender','dependents','education_level','prison_years','prison_offense','recidivism_within_3years','race', 'age_at_release',
'gang_affiliated'
,'prior_conviction_episodes','prior_conviction_episodes_1','prior_conviction_episodes_2','prior_conviction_episodes_3','prior_conviction_episodes_4','prior_conviction_episodes_5','prior_conviction_episodes_6','prior_conviction_episodes_7','supervision_risk_score_first']
```

Using neural network we got accuracy = 67%

If we assumed the arrest in year 1 ,2 ,3 features not in forms of the target:

First,recidivism_arrest in {year1,year2,year3} were found to indicate whether recidivism_arrest_within_3years column is true or not.

- After counting the number of returns to the prison in them, year1, and year2 got the most impact, less people returned in the third year, so year1,and year2 were taken to be two of the chosen features.
- Also 'gender','dependents','education_level', and 'prison_offense' were chosen to be in the predicting features.
- One hot encoding was used to spread the categorical data and get better feature engineering before the model step.
- After many models trials, neural network with 2 hidden layers and 0.01 as the correction factor were found to be the best model to fit the data with test accuracy = 90%

Function:

```
def get_count(state,offense)
```

To get the count of the offenses in each year using the API

To be used in multi_threading

Limitations:

- Interview limitations in the collection of victimization dataset
- The population dataset of victimization is unbalanced over years and quarters. The household with their people should be followed for 3 years. The people are less each quarter, but the difference is acceptable.The last 3 years have more bias in the dataset as they have more records.
- Normalization in part 2 Q 5 because the demographic offender's data is different in victimization from the population

- It is a limitation from the data set of victimization to know the race before 2012 as shown in the curve and they are the most number of offenders.
- It is a limitation to know the age of offenders for most non fatal crimes as there are large part which we do not know.

Challenges:

- The visualization coloring was difficult
- Visualization could be better to plot some curves with others in the same figure like bar plot and line plot to show the trends

Assumptions:

- All the data sets collection are randomly collected without bias
- In firearm dataset, the weight for each rule is correctly distributed and not biased
- Some assumptions related to the normal curve in regression
- The bonus task when fitting the model, the features of interest are assumed to be the best variables.
- Residues are neglected not the least number in some curves