Analyzing U.S. Crime Data | Technical Report

Muhammed Khalid	201901493
Samaa Khair	201901481
Eman Allam	201900903
Abdelmonem Ali	201800276

University of Science and Technology at Zewail City

2022

Supervised by:
Dr. Mahmoud Abdelaziz

I. Introduction

The aim of this report is to explain the structure of the project such as:

- file hierarchy
- general flow
- 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.

This report will be divided into six parts, wherein each part, the team member responsible for that part will write about his part.

II. Part 1

- **Data Collection:** In this part, I used the API of each data to request accessing it. I stored it in CSV files and data frames to clean it and use it in the analysis. Regarding the FBI data, I created a function "*get_FBI_data*" to get the data of each offense in each state to store it all in the "*results*" array. Finally, stored it in a CSV file.
- **Data Cleaning:** We cleaned it by removing any nulls and duplicates, replacing the numerically-encoded categories with their descriptive string, and expecting the missed categorical information by its mode or mean.

III. Part 2

This part is aimed to answer several questions regarding the relationships between categories for exploratory purposes

A. Question One

To answer the attached question, I grouped the "FBI DataFrame" by columns of offense and count to sum them in order to be able to get the top five most frequent offense categories. Next, by looping through an array of years I calculated the average

Results are shown through a bar plot with years from 1990 to 2021 on the x-axis and frequencies on the y-axis

B. Question Two

By looping through the states in "FBI DataFrame", the rate of each violent crime was calculated and summed then divided by the total crime rate.

Results are shown through a bar plot representing state abbreviations on the x-axis and frequencies of violent crimes on the y-axis.

C. **Question Three**

Same approach as question two but this time looping on years and checking only the rates of homicide offenses and for violent crimes in general.
Results are shown through a bar plot with years on the x-axis and rates on the y-axis.

D. **Question Four**

Grouping "CVS Select –Personal Victimization Dataset", a table that links columns of age, sex, race/ethnicity, and type of crime together as well as calculates the frequency of each combination was created to be able to detect the combination of demographics that has the highest frequency of crime.
Results were shown in a tabular form showing the whole range of combinations and printing the probability of each. Additionally, the top and least 10 combinations were presented on a bar plot.

E. **Question Five**

Similar to question four, the same approach of calculation and representation was taken. However, this question was concerned with the criminal's demographics not the victim's. Columns "offenderage", "offendersex" and "offtracenew" were used for assessment from "CVS Select –Personal Victimization Dataset".

F. **Question Six**

Relationships between education levels and gross household income were separately bar-plotted showing their relationship with the rate of victimization. The rate of victimization was calculated by dividing the count of victim respondents and the general whole population from each education/income subcategory.

IV. Part 3

In this part, I was responsible for answering some questions using the cleaned and loaded datasets from part 1. There were exactly five questions. For the first 3 questions I used the personal victimization dataset while for the last two questions, I used the Georgia dataset.
For the first question:
First, I dropped the Do not know and Residue results from the notify column as they were irrelevant to the question asked, then I used group by function to make a new dataframe. The attributes given to the groupby function were the newoff column and notify column. The newoff column already contains all of the non fatal crime and the notify column contains whether the crime was reported or under-reported. To visualize the result I used a bar plot to draw the resultant dataframe. For the second part of the first question. I used the same technique used in the first part except for that the column tested with notify now was direl column which contains information about offender victim relationships.
For the second and third questions:

The function "*combine_demographic*" was created to combine age , gender and race in one column. Then I used the groupby function again to group demographics with newcrime column. To print the resultant dataframe i used to_markdown() attribute in the pandas dataframe to be able to see each demographic with the corresponding violent victimization frequency. Finally, I used a bar plot to visualize the result.
For question 3, it was exactly the same as the second question except that the column tested with demographic was the treatment column.
For question 4 and 5, Georgia dataset was used. For both questions, I dropped any row that contains False in the *recidivism_within_3years* column to be able to count both *prison_offense* and *age_at_release* then put the result in a bar plot. As an alternative , and to use normalization, I used group by function again to count both true and false and normalize the result.

V. Part 4

In this part, I decided to use a two-tailed independent t test using stats.ttest_ind built in function in python.
To use this function, I needed to have two means to compare. The null hypothesis is that there is no difference between the mean violent crime rate between states with strict firearm control laws and states with less strict firearm control. The alternative hypothesis is that there is a difference between the mean violent crime rate between states with strict firearm control laws and states with less strict firearm control. To perform this task, I used firearm dataset and FBI dataset. First, I chose to take the year 2020 as my test year, so I dropped all other years. To determine what is the strict firearm and which is not, I calculated the mean using mean function and decided that any firearm above this mean will be strict and any below or equal will be non strict. Then I appended both datasets together [modified firearm data set and FBI dataset] and calculated the mean number of crimes for strict states and mean number of crimes for non strict states and compared both with stats.ttest_ind that i mentioned above.
The reason why I chose the two-tailed independent t test is because this test is appropriate as it allows us to compare the mean violent crime rate states with strict firearm control laws to the mean violent crime rate of states with less strict firearm control laws. Additionally, the t test assumes that the data is normally distributed and independent, which is reasonable to assume given that we are comparing two means of two separate groups.

VI. Part 5

Data set: The recidivism in Georgia dataset
Required: regression model that predicts the
Offender's supervision risk score based on :
- All prior convictions.
- Offender's race.

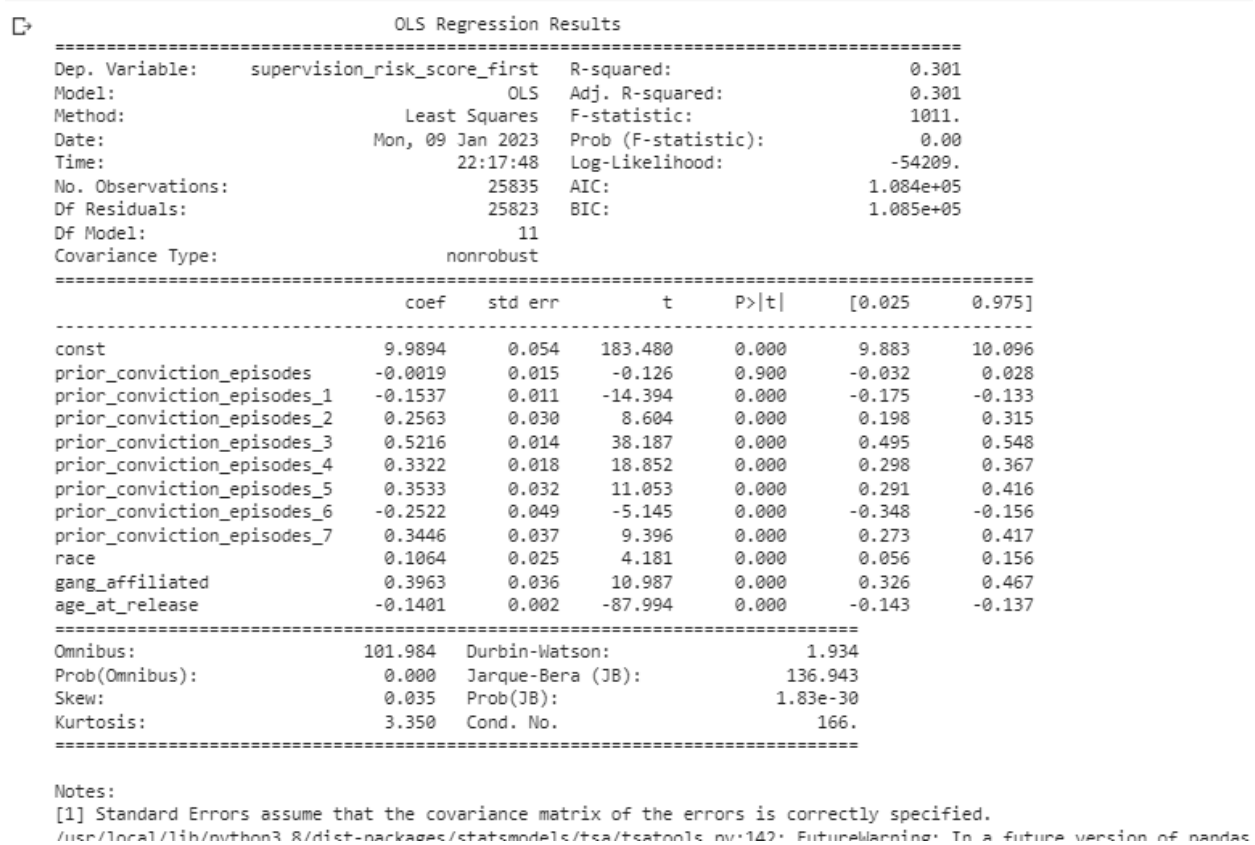
- Offender’s gang affiliation.
- Offender’s age at release.

What did I do?
first : calling and cleaning data
I checked null values and filled it with many methods
In Numerical data, I checked the skewness of the data. If it was skewed, I filled in the median. If not, I filled with mean

In categorical data I filled with forward and backward methods

Before fitting the model, I mapped the categorical variables to numbers to be able to fit it

I chose the linear regression model .
The results and P- values are



OLS Regression Results

Dep. Variable:	supervision_risk_score_first	R-squared:	0.301
Model:	OLS	Adj. R-squared:	0.301
Method:	Least Squares	F-statistic:	1011.
Date:	Mon, 09 Jan 2023	Prob (F-statistic):	0.00
Time:	22:17:48	Log-Likelihood:	-54209.
No. Observations:	25835	AIC:	1.084e+05
Df Residuals:	25823	BIC:	1.085e+05
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	9.9894	0.054	183.480	0.000	9.883	10.096
prior_conviction_episodes	-0.0019	0.015	-0.126	0.900	-0.032	0.028
prior_conviction_episodes_1	-0.1537	0.011	-14.394	0.000	-0.175	-0.133
prior_conviction_episodes_2	0.2563	0.030	8.604	0.000	0.198	0.315
prior_conviction_episodes_3	0.5216	0.014	38.187	0.000	0.495	0.548
prior_conviction_episodes_4	0.3322	0.018	18.852	0.000	0.298	0.367
prior_conviction_episodes_5	0.3533	0.032	11.053	0.000	0.291	0.416
prior_conviction_episodes_6	-0.2522	0.049	-5.145	0.000	-0.348	-0.156
prior_conviction_episodes_7	0.3446	0.037	9.396	0.000	0.273	0.417
race	0.1064	0.025	4.181	0.000	0.056	0.156
gang_affiliated	0.3963	0.036	10.987	0.000	0.326	0.467
age_at_release	-0.1401	0.002	-87.994	0.000	-0.143	-0.137

Omnibus:	101.984	Durbin-Watson:	1.934
Prob(Omnibus):	0.000	Jarque-Bera (JB):	136.943
Skew:	0.035	Prob(JB):	1.83e-30
Kurtosis:	3.350	Cond. No.	166.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
/usr/local/lib/python3.8/dist-packages/statsmodels/tsa/tsatools.py:142: FutureWarning: In a future version of pandas

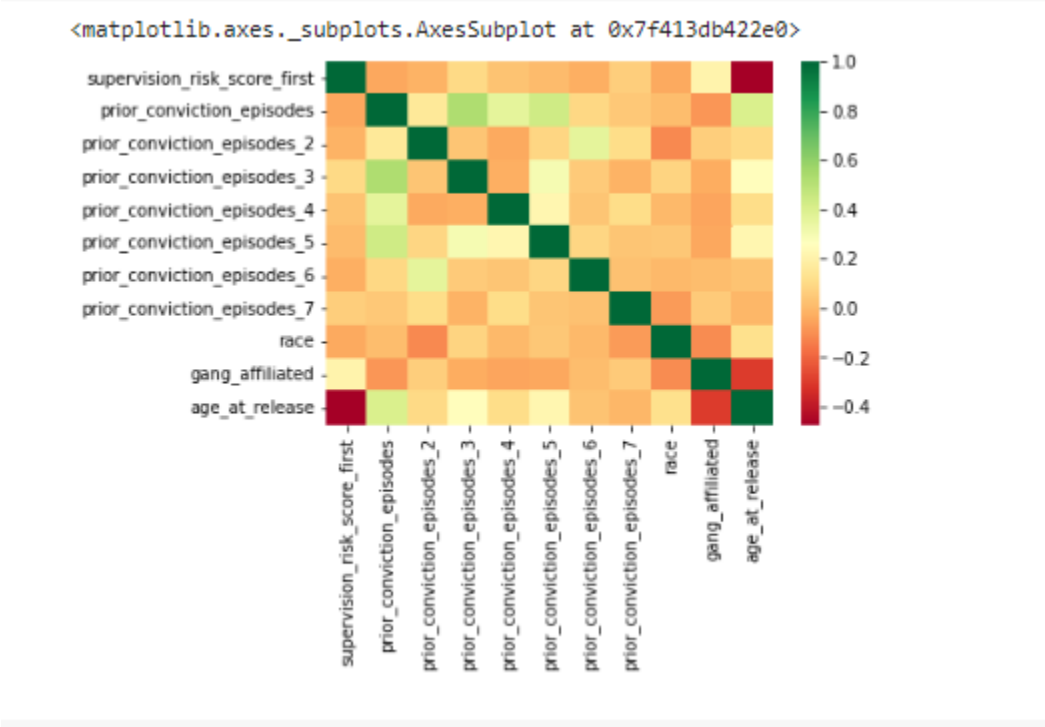
To determine which variables are good predictors of the variabilities in the target, you can look at the p-values of the individual coefficients in the linear regression model.

In general, a variable is considered to be a good predictor if it has a low p-value (less than 0.05) in the model. This means that there is a significant relationship between the variable and the target, and the variable can be used to predict the target.

On the other hand, a variable is considered to be a bad predictor if it has a high p-value (greater than 0.05) in the model. This means that there is no significant relationship between the variable and the target, and the variable cannot be used to predict the target.

good predictors : all variables are good predictors except prior_conviction_episodes

Some of these predictors correlated with each other (the result on the notebook)



Limitations and challenges:
In our case, an R-squared value of 0.301 indicates that about 30.1% of the variance in the target variable is explained by the model. This means that the model is able to capture some of the variability in the target, but not all of it.

There could be several reasons why the model has a low R-squared value. Some possible reasons include:

- The relationship between the predictors and the target is not linear.
- There is a lot of noise in the data, making it difficult to accurately predict the target.
- The model is underfitting or overfitting the data.
- There is multicollinearity between the predictors

VII. Part 6

Data set: The recidivism in Georgia dataset
Required: classifier to predict the likelihood of recidivism
within 3 years of release based on the state of Georgia recidivism records.

What did I do?
first : calling and cleaning data
I checked null values and filled it with many methods
In Numerical data, I checked the skewness of the data. If it was skewed, I filled with median. If not, I filled with mean

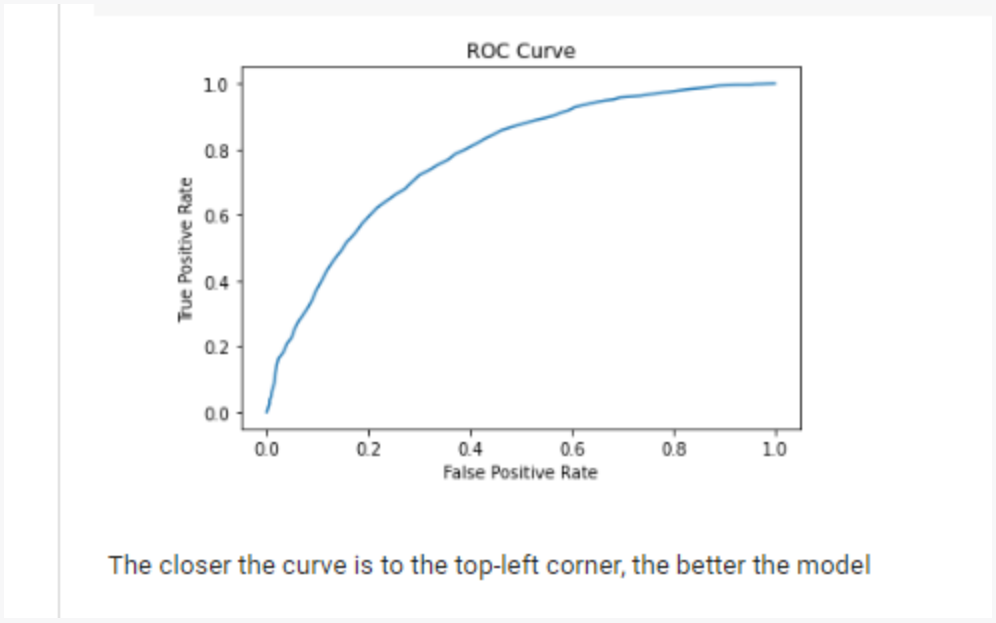
In categorical data I filled with forward and backward methods

Before fitting the model,I did a one hot encoding to categorical data and dropped the last 3 columns

Model: random forest

I decided to use random forest because:

- Random forests are powerful and accurate: Random forest classifiers are known for their high accuracy and ability to handle large and complex datasets. They work by creating an ensemble of decision trees, each of which is trained on a random subset of the data, and then combining their predictions to make a final decision. This process helps to reduce overfitting and improve the generalization of the model.
- They can handle a mix of data types: Random forests can handle a mix of numerical and categorical data, which is often the case in real-world datasets. This makes them a good choice for predicting recidivism, which may involve a variety of different types of data, such as demographic information, criminal history, and social support.
- They are easy to interpret: Random forests are relatively easy to interpret, as you can visualize the decision trees and see how different features contribute to the final prediction. This can be helpful when trying to understand the factors that influence recidivism and identify areas for intervention.



Limitations and future work:
A test set accuracy of 0.73 for a random forest model indicates that the model is able to correctly predict the target class for about 73% of the test set examples.

To improve the model's test set accuracy, we can try the following:

- Collect more data and use it to train the model.
- Use different hyperparameters for the model.
- Try different feature engineering techniques to create more relevant features for the model.
- Use a different model altogether.