Illinois Prison Population Final Report
Members: Kacper Mocarski, Raudel Vargas, Quang Le, Eddie Sanchez, Thoong Tran
Link to Github: Github
Link to Google Drive: Google Drive
Important Note: We could not figure out a solution to upload a file larger than 25 mb to Github, so it is missing our last two extra graphs. There is a zip file in Github that will have the full colab. Or, Google Drive also has everything we've made.

Introduction and Data:

During the beginning of the semester, we chose to explore the Illinois Prison Population and all of the issues, if any, to see if we could uncover any trends and provide possible solutions. We thought that this was a good idea to pursue because prisons around the entire United States share common issues: high incarceration rates, overpopulation, recidivism, and too much funding. Illinois supposedly experienced a lot of these issues too, so we wanted to see if this was true for our state, and see the entire history of this issue. The prison system is not an easy system to understand, but analyzing different trends with the data would allow us to gain deeper understanding, and give us the ability to pick out specific areas that need reform.

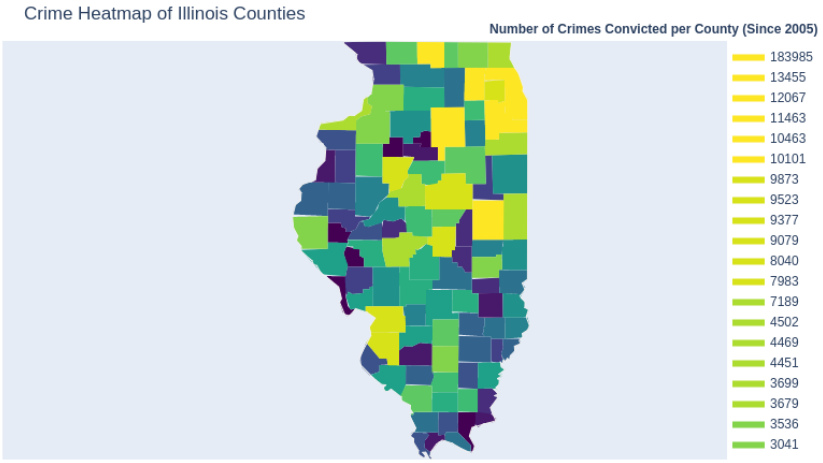
We gathered all of our data from the Illinois Department of Corrections, which contained data (as excel files) for every single month from January 2005 to June 2023. We decided to only use the excel files for December of each year, as there were a lot of repeated columns. It made it easier for us as it limited the number of files we needed to import and clean, while also keeping a majority of rows that were crucial for our analysis. We combined these files into one major CSV file, and did some basic cleaning to make sure the data was all in the same format. An issue with this data was that somewhere along the way in recording it, they would change the format in some of the columns, so we needed to make sure that it was as consistent as possible.

With an area that we were interested in developed and a data pool acquired, we needed to determine what exactly we wanted to discover, which led us to this problem question:
In Illinois, which crimes are the most committed/convicted, who is being convicted for those crimes, how do sentences differ based on who committed that crime, and how likely is a crime to be repeated?

Where is this taking place?

First, we wanted to see if there were any specific areas (counties) in Illinois that struggled the most with a large number of convictions. These counties and their institutions would be the main victims of overcrowding and would give us a general understanding of how each county compares to one another.

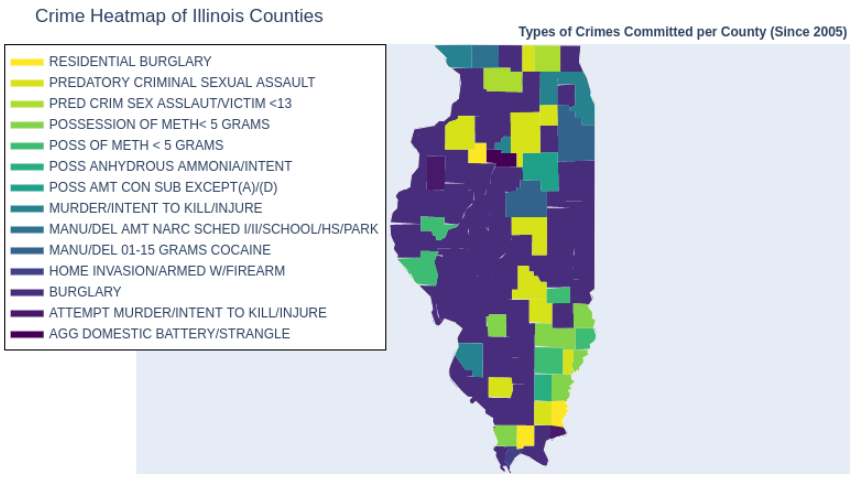
To understand where these crimes are taking place, we created a heatmap of Illinois separated by its counties. We grouped our data by “Sentencing County” and “Sentencing Date”, and filtered out the convictions from before 2005. We then got the count of each “Sentencing County”, and this gave us our graph.



What we found was that the most crimes were convicted in the counties surrounding Chicago. The most convictions happened in Cook county, which had 170,530 more convictions than the second highest county. The only outlier to this finding was Champaign, who also had a high number of convictions, but we determined that this was because it has a large college population, so a lot of crime is possible with the size of the population. In conclusion, however, we found the counties near Chicago, which are also some of the highest populated counties, have the most convictions in Illinois.

Which crimes are the most convicted?

In order to answer this question, we utilized the same type of graph to be able to see the difference between the counties with the number of crimes convicted and which crimes are the most popular. In order to create this graph, we followed a similar process with only keeping convictions in 2005 and later. We then grouped the county and holding offense, and calculated the count for each holding offense. Then, we kept the holding offense with the highest crime for each county, ending up with 102 rows (for 102 counties), and this was the result.

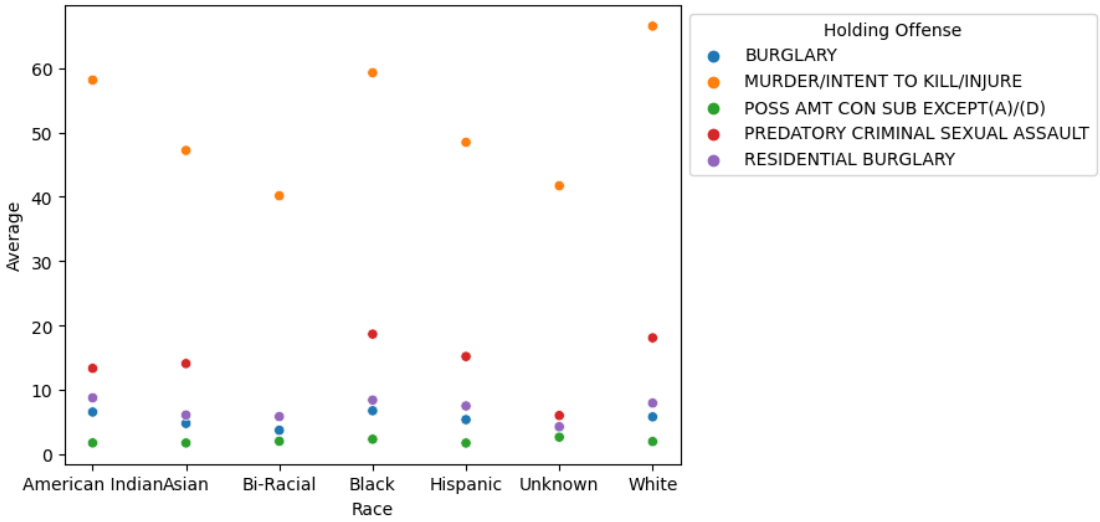


As a result, we can see that counties closer to Chicago have crimes related to murder or sexual assault as their most committed crimes. There are a sprinkle of burglaries and drug possession near that area, but the more populated counties have crimes related to harming others. Additionally, counties farther away from Chicago with less population typically have crimes related to burglary (central Illinois) or crimes related to possession of some illegal substance, whether that is drugs or firearms (south east Illinois).

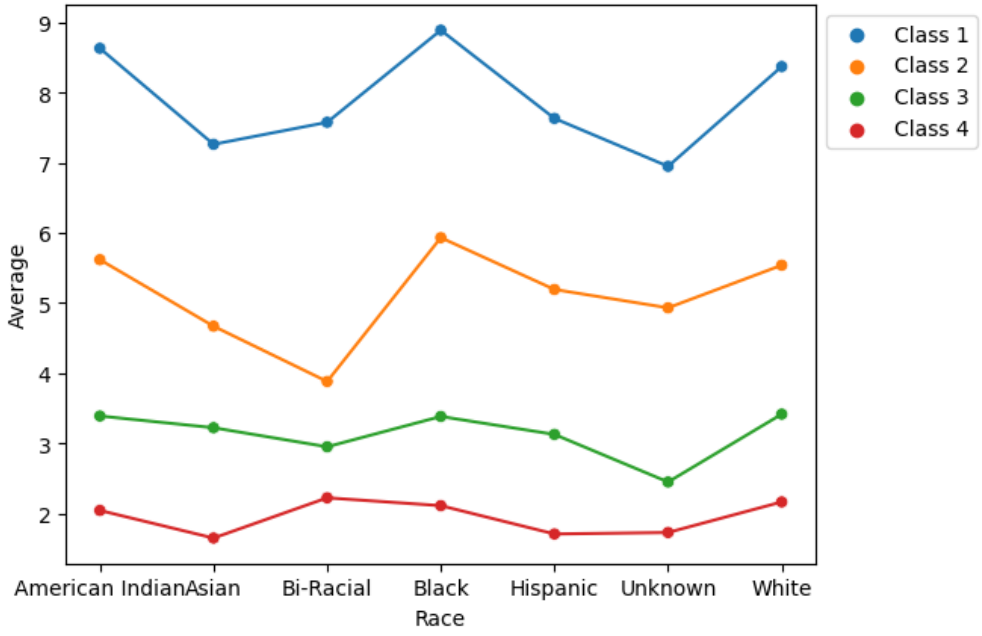
Who is being convicted for those crimes and how do their sentences differ?

The data we have from the Illinois Department of Corrections classified each convict by their race, and they had seven main groups: American Indian, Asian, Black, Hispanic, White, Bi-Racial, and Unknown. For this project, we mainly focused on race as a classification for who is being convicted for crimes. This was mainly because when researching this topic initially, one of the main concerns was over racial disparity in sentencing. There wasn't much discussion over issues in sentencing years for gender, so we decided to only categorize by race.

To begin, we wanted to see the average sentence for each race with the top five convicted crimes in Illinois. We did this by first getting the top five crimes by totaling the "Holding Offense" and how often someone was convicted for that crime. We then grouped the columns "Race" and "Holding Offense", and calculated the average for "Sentence Years" for each grouping. This process gave us this output.



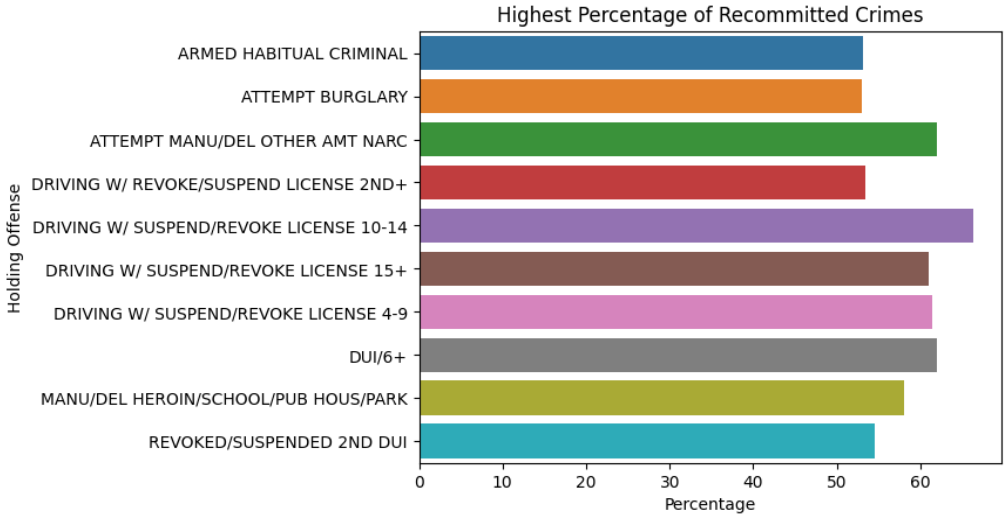
What we see from this graph is that there isn't as much racial disparity in the Illinois prison system as we first thought there would be. Residential burglary, burglary, and possession of controlled substances (Class 4 felony) all have comparable sentence averages, while murder and predatory criminal sexual assault begin to see slight differences in these averages. However, this data does not suggest any sort of racial disparity in the top five most convicted crimes. We weren't satisfied with this as an answer, so we wanted to continue investigating. With further research, we discovered that racial disparity in sentencing is truly seen in the prison system for lower level crimes. In our data, we have this sort of classification under "Crime Class" which has crime classes Class X (the most severe crimes), Class 1, Class 2, Class 3, and Class 4 (the least severe crimes). So, we followed a similar structure to the previous, except this time we grouped all crimes into their crime classes, instead of getting a top five crimes list. This is the resulting graph.



This graph shows us something very similar to what we saw in the previous graph. The lower class crimes have little to no difference in average sentences, but that difference slightly increases as you increase the level of the crimes. However, it seems that the main victims of high sentence averages are blacks and whites. These two races have the highest average sentences in each category we've observed.

How likely is a crime to be repeated?

The final area we wanted to understand was recommitted crimes in Illinois. The Illinois Prison system is a victim of high recidivism, and we wanted to uncover what crimes lead to this. We began by grouping "Holding Offense" and "Admission Type", and getting the count for each grouping. Then, we wanted to calculate which holding offenses had the highest percentage of "Discharged & Recommitted", so we divide the number where "Admission Type" was "Discharged & Recommitted" by the total number of admission types for each holding offense. This process gave us this graph.



According to this graph, we can see that the highest recommitted crimes can be categorized into three groups: Driving when you're not supposed to (impaired or suspended license), possession of illegal objects (drugs or weapons), and burglary. These crimes having over 50% recommitted rates is too high. With the number of crimes that this data gave us (over 700,000 rows of data), it is not just a few people recommitting these crimes. These crimes are being committed by a majority of convicts, and they are going out and recommitting them again.

ML Analysis

We also performed various machine learning analyses to see what sort of potential findings we could uncover within our data.

For one of the analyses, we wanted to focus on the columns Admission Type, Crime Class, Holding Offense, and Sentence Years. We wanted to train the data on Admission Type, Crime Class, and Sentence Years to see if our classifier could predict the Holding Offense. We used a Linear SVC classification, and we fit the data using StandardScaler, purely because the data seemed too big and the classifier would take too long to run when I ran it with just SVC and no Standard Scalar.

We found that this classifier gave us a 76% accuracy on the Holding Offense it would give us. This means that we can get a fairly accurate result when we are only given Crime Class, Holding Offense, and Sentence Years.

Another analysis we ran was Logistic Regression, where we would try to see how accurately we can predict the "Admission Type" of a convict given categories "Race", "Holding Offense", and "Sentence Years". We separated the "Admission Type" into two groups, 1 if the type was "Discharged & Recommitted" and 0 for anything else. We were able to get an overall accuracy of about 70%, and our model seems to perform better when it comes to predicting class 0 (non-recommitted) compared to predicting for class 1 (recommitted).

Findings

With this data, graphs, and machine learning analyses, what should we take away from this project?

This project showed us that a lot of these convictions are happening a lot in higher populated counties, especially in and around Chicago. These areas have a much higher population, so it isn't a big surprise that more crime can happen here. However, Illinois is not well equipped to deal with these crimes in terms of prison space. Too many people are being sentenced in these areas, and a lot of space is being occupied as a result. We were also able to see what crimes are the most popular in these counties, and discovered that counties towards the middle and border of Illinois away from the northwest have the burglary or possessions of some sort of drugs as their most committed crimes since 2005. Counties closer to Chicago have crimes related to murder or sexual assault as their most committed crimes.

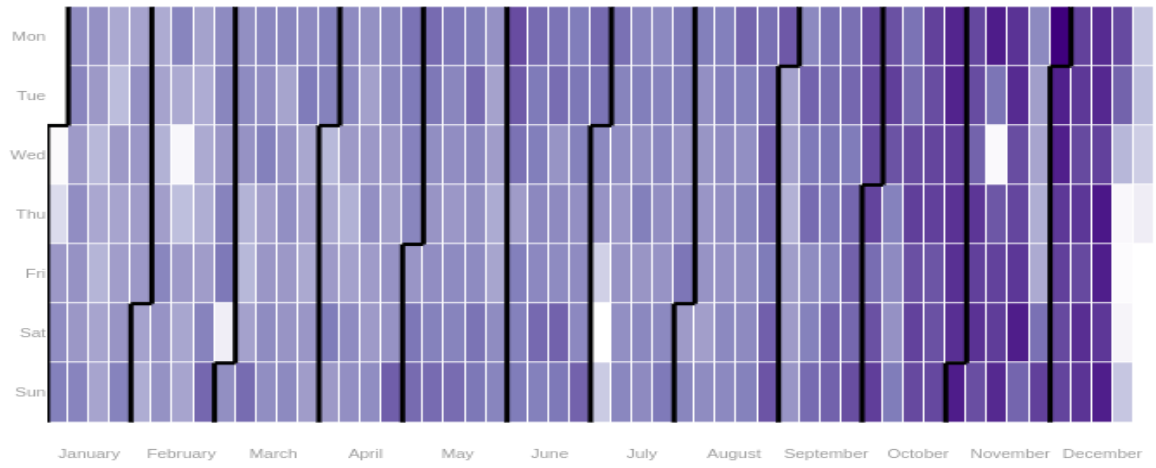
We were also able to discover that although the Illinois prison system struggles with a large number of convicts, it does not struggle with high rates of racial disparity in sentencing. Through our visualizations, we were able to see that there wasn't any evidence that showed minorities receive harsher sentences on average than non-minorities, rather we saw that two races were on the receiving end of larger sentence durations: whites and blacks. This could be due to the fact that they are the top two most common races that get convicted in Illinois, so their

averages could be increased because they have more convictions. Overall, we do not believe that Illinois struggles with unfair sentences based on race.

A final discovery we think is crucial is the high rates of recidivism. Obviously, there are crimes that need to be punished with prison time, such as murder and sexual assault, but prison time is not the solution for every crime. We saw this with our graph on the most popular recommitted crimes. All of these crimes had over 50% reconviction, and there were more that didn't make this top ten. This discovery calls for grouping these kinds of crimes and developing plans to combat the type of crime, whether it is rehabilitation for crimes related to drugs, or higher fines and driving school for impaired driving.

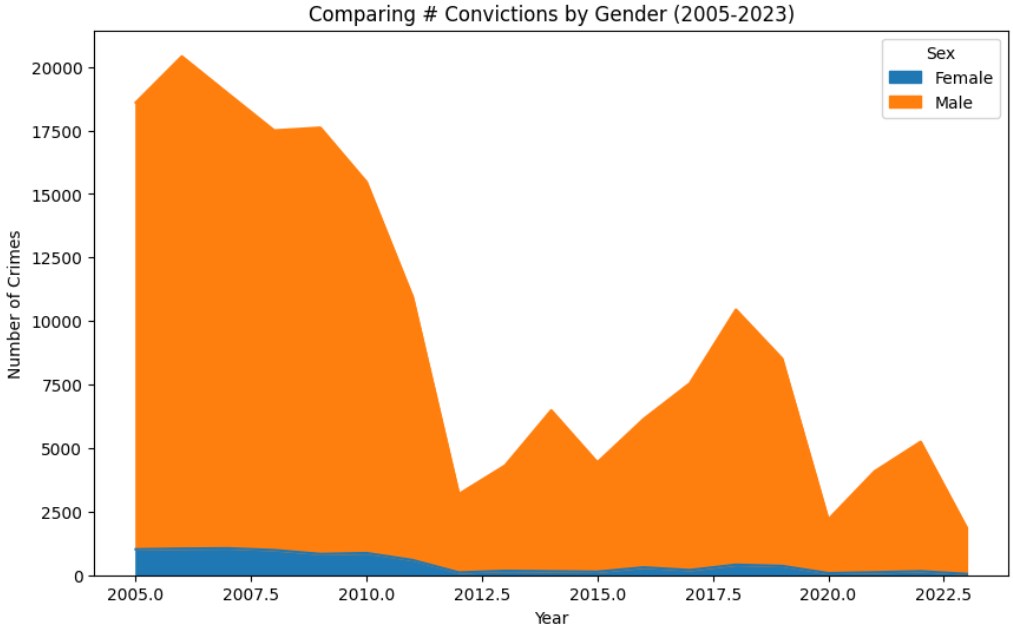
Additional Work and Findings

Some of the additional work that we did explored questions outside of our original problem question, but would still give us insight on information about the problem as a whole. The first additional exploration was trying to see what days in the year are most popular for convictions in Illinois, all counties. We grouped each custody date and got the count of every row with the same custody date, and we used a calplot in order to plot this data.



The way colab could print this image makes it seem distorted, however we still gain valuable insight for it. The first of each month starts in the top box on the furthest left side of the month, and then the days count down. From this graph, we see that the majority of convictions happen towards the end of the year, from about mid-October to mid-December. This was very interesting to see, and it is almost true with what we thought would happen. We hypothesized that the most convictions would happen around the holidays, and although this isn't true for all holidays, it is true for certain ones, especially near Thanksgiving and Christmas.

Throughout our project, we also focused on race as a way to see *who* the person getting convicted is, which was intentional. There wasn't a lot of concern over gender when it came to sentencing, only racial disparity. However, this is what the data showed us.



The number of males being sentenced to prison each year is significantly higher than women. Generally, it is expected that more men commit crimes than women, but convicting almost 10x the number of women is very large. It's difficult to come up with a way to decrease this number, as there are many factors that contribute to this. However, we felt like it was important to highlight the fact that the male population makes up a majority of these convictions.

Reflection

Overall, we think this was an interesting and engaging project to work on. It was very insightful to see the prison system in Illinois, what kinds of crimes are being convicted, how long sentences are for certain types of crimes, and just get an overall understanding of how our prison system is set up.

However, a difficulty within our project was the type of information we got from the data. A lot of the data was useful and helped us answer our main question, but it lacked additional, useful columns to help us create a lot of great visualizations for the data. When coming up with ideas to represent the data, we found ourselves coming up with repetitive ideas or graphs that were too similar to ones we already created. If we were to restart this project, we would probably try to find an additional resource we could connect with this current data to see if we were able to uncover more unique information.