Predicting Recidivism in Connecticut

Michael DiMattia
dimattiami@gmail.com
Eastern Connecticut State University
Senior Research CSC-450: Spring 2017 (Dr. Dancik)

Abstract

In the past ten years, all states combined reported annual spendings of $50 billion (fluctuating +-$2 billion per year) on jail inmates. Leveraging computers to find arrest patterns will help law enforcement predict what crimes to expect at any given time, which may reduce crime and lead to less money spent on imprisonment. This research has been conducted with Connecticut convictions to help understand the type of crimes that occur in the area that this research was conducted. Due to the fact that the Connecticut's conviction dataset is not available for direct download, custom software tools must be created to scrape the CT judicial website to get our data in a structured format that can be stored in a SQL database and analyzed with R programming. Information that can be gathered through analysis consists of (but is not limited to) common crimes in any given area and/or season, relationships between crimes, and estimations of recidivism for criminals.

Introduction

Crime is inevitable in society, and sanctions for criminals are not cheap. Between 2002 and 2010 all states combined ran up a costly bill that ranged between $48.4 and $53.4 billion per year [1]. It costs ~$29,000 per year to keep one offender incarcerated, and ~$2,750 per year for post-release monitoring programs per offender [2]. To decrease the expensive sanctions cost, states would rather free prisoners on a post-release monitoring program such as parole. Post-release monitoring programs may sound like a win-win situation, however, many of these parolee's are not reformed from these programs and are likely to be rearrested and be a detriment to society.

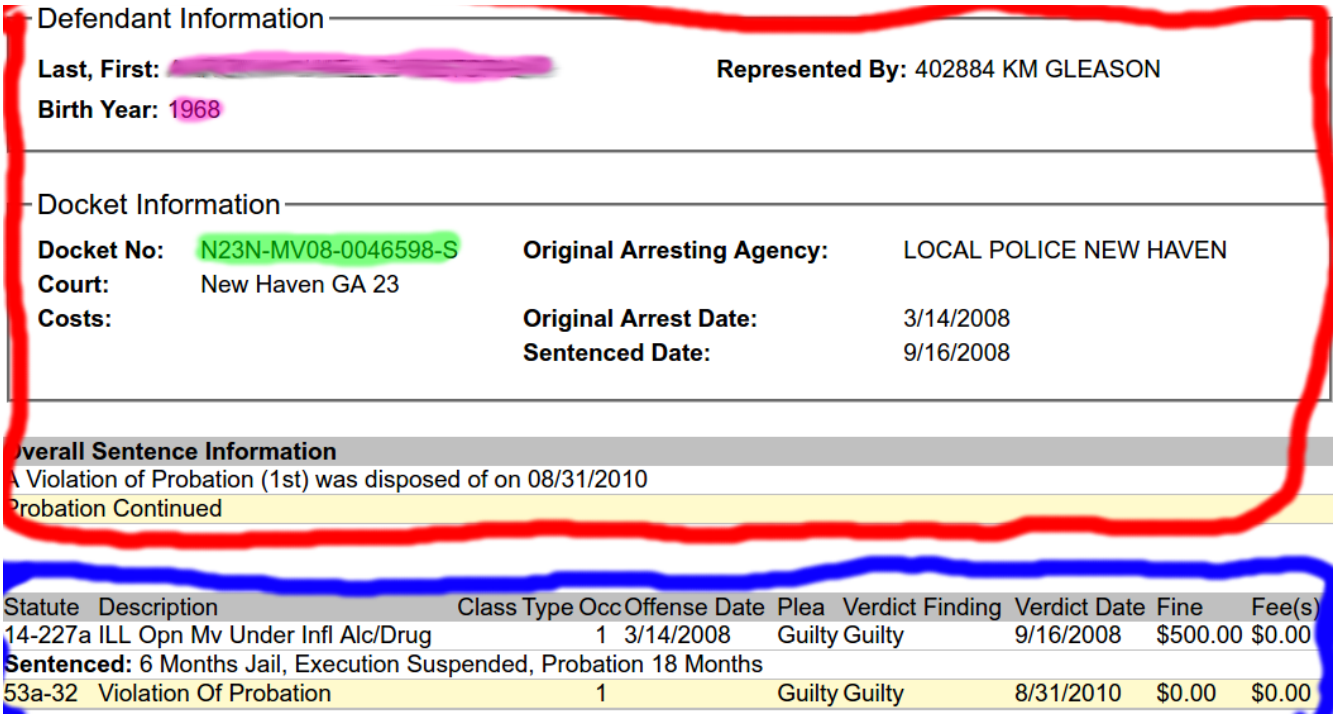
There are two main ways offenders are released from prison: conditionally and unconditionally. Unconditional releases simply occur when a prisoners' sentence ends – there is no post-release supervision. Conditional releases are a bit more complicated and can be either discretionary or mandatory. Discretionary releases require prisoners to be screened by a parole board to see if they've been reformed and are ready to go back into society. Mandatory releases are obtained by prisoners who were on a program while serving their prison sentence that allowed them to earn "good-time credits" for an early release. Prisoners released on discretionary parole are less likely to be arrested than mandatory and unconditional release offenders – 60% of unconditional and mandatory release offenders were rearrested within two years of release, and 54% of discretionary parolee's were arrested within two years after their release [3]. In addition, the Bureau of Justice Statistics reported in 2005 that out of all the ex-parolees that committed a crime within 5 years of their release, 36.8% of these offenders were rearrested within 6 months of release, and 56.7% were rearrested within their first year of release [4].

Prisoners are often released early from sentences on monitoring programs like parole, however, many of these early-released prisoners end up back in prison. Releasing inmates prematurely allows them another chance to negatively affect society.

The Connecticut judicial website (http://www.jud2.ct.gov/crdockets/parm1.aspx) holds information about Connecticut court cases for the past 10 years. Each case is identified by a unique docket number. Docket numbers are composed of the court code, type of offense, year of offense, case number, and a suffix. Cases can be looked up by docket number, which will give a detailed summary regarding the court case. This detailed summary includes defendant information, docket information, and charge information. Information about the defendant consists of defendant first and last name, birth year, and representing attorney. Docket information tells the arresting agency (typically police

departments in the town that the crime was committed,) original arrest date, and sentenced date. Finally, all charges are listed with statute number, description, class (A, B, C, etc.,) type (felony/misdemeanor,) occurrences, offense date, plea, verdict finding, verdict date, and any fines/fees that the defendant was ordered to pay. **Figure 1** is a screenshot of a random docket on the CT judicial website that contains the aforementioned data. The data enclosed in the red box is 'metadata' about the case, blue box is the actual charges that the offender was charged with, green is the docket number I used as a join key between the case and metadata SQL tables, and purple is the composite key used to uniquely identify offenders.

In this paper, we will explore the process of creating a queryable database containing Connecticut judicial records, along with tools/models that allows the analysis and prediction of crimes in Connecticut. These tools will weigh offender's likelihood of being rearrested, which will affect how an offending criminal is sentenced.



Defendant Information

Last, First: [redacted] **Represented By:** 402884 KM GLEASON

Birth Year: 1968

Docket Information

Docket No: N23N-MV08-0046598-S **Original Arresting Agency:** LOCAL POLICE NEW HAVEN

Court: New Haven GA 23

Costs: **Original Arrest Date:** 3/14/2008

Sentenced Date: 9/16/2008

verall Sentence Information

A Violation of Probation (1st) was disposed of on 08/31/2010

Probation Continued

Statute	Description	Class	Type	Occ	Offense Date	Plea	Verdict Finding	Verdict Date	Fine	Fee(s)
14-227a	ILL Opn Mv Under Infl Alc/Drug			1	3/14/2008	Guilty	Guilty	9/16/2008	$500.00	$0.00
Sentenced: 6 Months Jail, Execution Suspended, Probation 18 Months										
53a-32	Violation Of Probation			1		Guilty	Guilty	8/31/2010	$0.00	$0.00

Figure 1: random docket on the Connecticut judicial system website

Materials and Methods

Preparation

This project is split into two large components: preparing the court records and analyzing the court records. Preparing the court records is required before analysis may be performed on this dataset. Data preparation in this project is essentially an ETL (Extract Transform Load) process. The extract phase of this project consists of creating a web scraper for the CT judicial website that retrieves all court records and then saves them in HTML format on disk. The CT judicial website uses JavaScript postbacks for their court record pagination, I resorted to using the PhantomJS "scriptable webkit" to inject JavaScript code to help me navigate the website.

With the scraped content and a stable DB schema for the court records, the transformation phase begins. BeautifulSoup 4 (an HTML parsing library for Python) is used to transform each HTML document into a line (or record) in a CSV file. This CSV file is loaded directly into MySQL (a popular,

free and open source database.) Once MySQL has ingested all of the CSV data, the analysis process starts.

Before continuing to the analysis process, I'd like to point out that midway through the analysis phase, I realized I needed to go back to the preparation phase to 'beef up' my records, which would allow me to perform additional analysis methods on the dataset. A Java program was created to iterate through the local MySQL database, which created several boolean fields for each court case. These fields corresponded to: hasDrugCharges, hasWeaponCharge, hasBurglaryCharge, hasExecutionSuspended, and hasProbation.

Understanding the dataset

Analysis in this project occurs in three different modules. The first module consisted of creating dashboards for the dataset. These dashboards helped illustrate the amount and type of data I was working with. These dashboards also supported 'drilling down' for a more detailed view of court records that a business user would be interested in viewing. These graphs and dashboards were modeled in Excel and the Business Analytic tool, Tableau.

Finding patterns

The second module entailed using the IBM SPSS Modeler tool for data mining. Given the dataset and user-specified input/output fields, IBM Modeler helps identify the best predictive models for complex relationships in the dataset. After running my original dataset through the IBM SPSS Modeler with unsatisfying results, I decided I needed to go back to the data preparation phase to add boolean fields that would help predict patterns.

Crime correlation

The final module consisted of using R to finalize my analysis. R utilized the data gained from modules 1 and 2 to create appropriate data models to perform statistical tests on. R allowed me to utilize the previously-created boolean values of the dataset to create a contingency table, which helped me find correlations between crimes that are likely to occur together. A fisher test was then applied to these contingency tables to statistically prove and support my findings.

Based on the information and patterns found throughout the analysis phase, a recidivism calculator has been generated that will give each offender a 'recidivism score.' Higher recidivism scores represent higher chances of the offender being rearrested. The formula for this calculation consists of three components of an offender: amount of previous cases, amount of times in jail, and probation violations. In conjunction with the aforementioned factors, a logistic regression model was used against a boolean value, representing whether or not that offender was re-sentenced within five years after their initial arrest. The exact formula for calculating recidivism scores is revealed in the results section of this paper.

All of the tools created and used for preparing the dataset used in this project are located on my GitHub at https://github.com/dimattiami/CT-Judicial-Analysis. The dataset in MySQL format (dump.txt) used in this project is available for download on my webserver at http://osbuddy.pro/ct-jud/.

Results

The data preparation process is illustrated in **Figure 2.** After eliminating any duplicate cases, 308,360 out of 408,205 cases were retained, with a total of of 536,991 charges across 141,017 different offenders. **Table 1** represents a breakdown of the amount of charges in each court case, and the corresponding amount of occurrences.

Figure 2: data preparation process

Table 1: Frequency table for number of number of charges in each court case

# of charges	1	2	3	4	5	6	7	8	9	10	>10
Frequency	207,672	61,916	26,947	7138	3084	928	380	128	52	35	36

Several tests were performed on this dataset to reproduce previous findings, including a correlation between drug and weapon charges (**table 2**.) Only .014% of drug charges contained weapons. There were no major correlations found between drugs and weapons in this dataset, which is discussed further in the discussion section of this paper. However, after running Fisher's exact test for count data, a small p-value of 0.00958 was calculated. A similar test was conducted to find correlations between weapon and burglary charges (**table 3**,) with the majority of weapon charges not being related to burglaries with a p-value of <2.2e-16.

Table 2: 98.5% of drug charges did not contain weapons, while only .015% of drug charges contained weapons.

	weaponfalse	**weapontrue**
drugfalse	0.0	0.0
drugtrue	0.98534916	0.01465084

Table 3: 5.9% of weapon charges were related to burglaries

	burglaryfalse	**burglaryfalse**
weaponfalse	0.0	0.0
weapontrue	0.94076935	0.05923065

Figure 3 represents the recidivism calculator that was discussed in the previous section. Higher recidivism scores represent higher chances of an offender to be rearrested. Recidivism score is calculated by multiplying three times the amount of previous cases an offender has, multiplying 2 by the amount of times the offender has been in jail, and then the amount of times an offender has probation violations multiplied by 1.5. The results from this calculator were then applied against whether or not an offender was rearrested in a linear regression model.

recidivism score = 3*amt of previous cases + 2*timesInJail + 1.5*probationViolationOccurences

Figure 3: weighted recidivism calculator

Figure 4 is a tree-map that represents the most occurring crimes per season (darker color/bigger square means more occurrences.) The biggest issue in summer for law enforcement was assault of the 3rd degree, spring and winter had the most arrests for driving under the influencing, and autumn's largest issue was larceny of the 6th degree.

Before wrapping up my results, since our school is located in Willimantic, I'd like to reflect on the arrests made by the Willimantic police department. First thing that I'd like to point out are the age of offenders at the time of their arrests. The top ten ages of offenders in Willimantic are 26, 22, 23, 24, 25, 30, 21, 28, 27, and 29 with occurrences (arrest counts) of 149, 149, 138, 136, 130, 129, 126, 122, 118, 115, respectively; people in their mid-twenties are likely to offend in Willimantic. The top five offenses that occurred in Willimantic were breach of peace 2nd degree (504 occurrences,) failure to appear 2nd degree (393 occurrences,) driving under the influence (356 occurrences,) assault 3rd degree (314 occurrences,) and possession of narcotics (312 occurrences.)

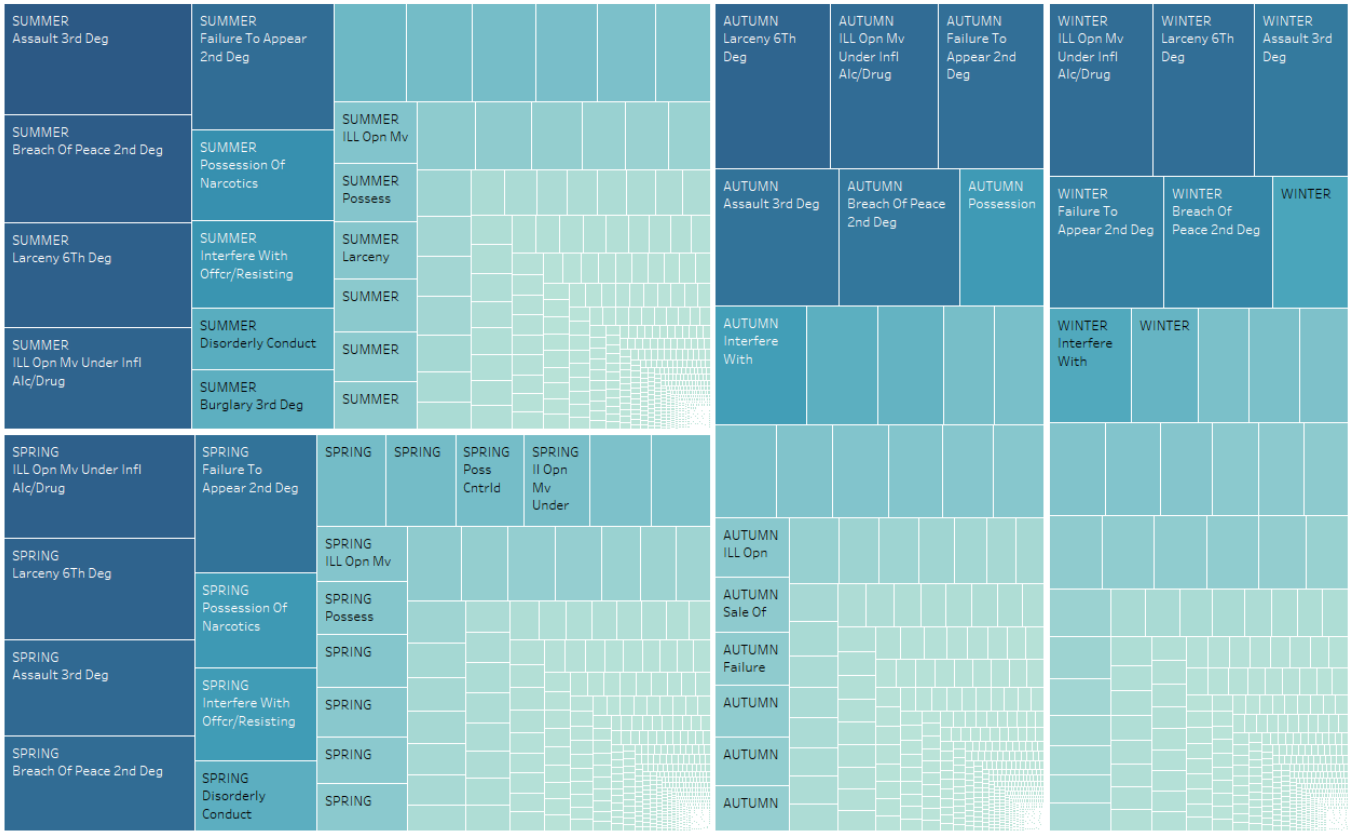


Figure 4: Season and offenses by crime type

Discussion

During the initial scraping process, HTML documents for each record were saved locally. Though saving an HTML document for each record is not as space efficient as parsing the HTML up front (transformation method #1) and then writing the values of each field to disk (CSV format, etc.,) it is the most reliable method of extraction in this situation. Firstly, if records are ever lost from my database, the judicial website will not need to be scraped again. Secondly, I've noticed inconsistencies between court records - some court records have fields that others do not. Having the HTML records on my disk will allow me to easily update my database structure if new fields are discovered.

One limitation to this research is that not all judicial records were collected, mainly because of docket number parsing issues. Cases always have a docket number associated with it, which always contains a suffix character (S,T,A, or 0-9; I am still unsure of what the suffix represents.) This was an issue that occurred during the scraping process with PhantomJS; all records with an S suffix were scraped but the others were not. After eliminating any duplicate cases, 308,360 out of 408,205 records were retained. Time restrictions for this project prevented me from going back to correct this issue. I felt that having ~75% of the dataset would be a fair amount of the dataset to work with.

It should be pointed out is that there were several mistakes in the CT judicial database itself. For example, if the docket number H12M-CR12-0239071-S is searched on the CT judicial website, 102 charges for that single case are returned. Most of these 102 charges are duplicates, but it is important to be aware that data sources are not always perfect. If the proper modifications (such as duplication removal) are not performed on the dataset before analysis, results may be inaccurate. I was able to remove a decent amount of duplicate cases in my dataset, however, I did not notice all of the dataset errors until after my analysis was complete.

When analyzing individual offenders, the only data provided by the data source for offender identification are first name, last name, and birth year. This project therefore utilized the three

aforementioned keys to create a composite 'PersonID' unique identifier for each person in the database. I noticed several times that offenders had similar names and birth years, but had different PersonID's. After further investigation, these people with similar PersonIDs were actually the same person – they just had misspellings in their name. Another issue that revolves around using first name, last name, and birth year as a unique identifier is that if two offenders have the same name and birth year (very unlikely,) there will be inconsistencies when analyzing individual offenders. A solution to this issue (which would guarantee unique identification,) would be to use offender social security number, which is most likely only available to the CT judicial system internally.

The recidivism calculator falls short in two aspects. The calculator takes into account offenders that were re-sentenced within five years after their initial arrest. Due to the small time frame of the dataset offered by the CT judicial system (past 10 years of arrest data,) many of the later offenders were unable to be followed up with because five-years of time hadn't elapsed since their arrest. A potential solution to this issue is to check if the offender already exists in the database when new arrest data is automatically ingested into the system (discussed in a few paragraphs.) The initial idea for the recidivism calculator came from The United States Bureau of Justice Statistics (BJS); BJS used prisoner data from 2005 to create a "prisoner recidivism analysis tool" to estimate recidivism rates based on factors such as gender, race, type of crime, and prior arrests [5]. Due to time restrictions, I was unable to calculate a recidivism score for each offender in the dataset, and then run those scores through a linear regression model to obtain statistical support for my calculator.

When looking for correlations between crimes, I based the correlating factors on sucessful correlations found in previous studies. The Bureau of Justice Statistics reported that 24% of drug offenders were charged with possession of a weapon during their arrest [6]. In my study, there was not a visible correlation between drug and weapon charges. I believe that the reason why my correlation results did not reflect the BJS's findings was because the BJS's study consisted of federal inmates, who tend to be more dangerous (more likely to be in possession of weapons) than offenders arrested at the state-level.

In addition to the cases utilized in my research, the CT judicial website also offers information about pending cases (http://www.jud2.ct.gov/crdockets/parm1.aspx.) An automated process could be created to scrape and ingest new court records from this pending case database into my local MySQL database. Having this automatic ingestion process would allow real-time trend analysis on a day-by-day (daily) time granularity.

Another idea I had - but was unable to implement due to time restrictions - was to split the dataset into a training and testing set by arrest year. I would try to train a model with the first six years of criminal records to calculate recidivism rates for each offender, and then test the calculated recidivism rate against the testing set (cases following the first six years) to see how accurate the recidivism calculator is, based on offenders that were rearrested in the testing set years.

There were other factors that were not examined during this projects due to time restrictions and lack of data offered by the CT judicial website, such as gender. BJS analyzed criminals in parole programs and found that 48% of the female parolee's successfully completed their parole program, while only 39% of males completed their parole program [7]. In 2015, 59% of female federal prisoners and 49% of male federal prisoners were serving sentences for drug charges [8]. If we wanted to analyze Connecticut crimes based on gender, an additional field would be added to the database; gender. This field would hold characters 'M' for male or 'F' for female, which would be obtained by determining offender gender based on their first name. The main issue with finding gender based on a first name is unisex names; if CT judicial provided this in their database, gender could have easily been incorporated into this project.

Bibliography

[1] T. Kyckelhahn, (2012). State corrections expenditures, FY 1982-2010. *Bureau of Justice Statistics, Office of Justice Programs, US Department of Justice, Washington, DC*, 1-14. Available: https://www.bjs.gov/content/pub/pdf/scefy8210.pdf [Accessed: 22-Feb-2017].

[2] Justice Policy Institute Published: June 2, 2010, "For Immediate Release: How to Safely Reduce Prison Populations and Support People Returning to Their Communities," *For Immediate Release: How to Safely Reduce Prison Populations and Support People Returning to Their Communities — Justice Policy Institute*, 02-Jun-2010. [Online]. Available: http://www.justicepolicy.org/research/1919. [Accessed: 22-Feb-2017].

[3] Amy L Solomon, Vera Kachnowski, Avinash Bhati, "Does Parole Work? Analyzing the Impact of Postprison Supervison on Rearrest Outcomes," *Urban Institute*, 1-Mar-2005. [Online]. Available: http://webarchive.urban.org/UploadedPDF/311156_Does_Parole_Work.pdf. [Accessed: 22-Feb-2017].

[4] Matthew R. Durose, Alexia D. Cooper, Ph.D., Howard N. Snyder, Ph.D., *Bureau of Justice Statistics*, "Bureau of Justice Statistics (BJS) - Recidivism of Prisoners Released in 30 States in 2005: Patterns from 2005 to 2010 - Update," *Recidivism of Prisoners Released in 30 States in 2005: Patterns from 2005 to 2010 - Update*, 22-Apr-2014. [Online]. Available: https://www.bjs.gov/index.cfm?ty=pbdetail&iid=4986. [Accessed: 22-Feb-2017].

[5] Snyder, Howard N., Durose, Matthew R., Cooper, Alexia, and Mulako-Wangota, Joseph. Bureau of Justice Statistics. *Prisoner Recidivism Analysis Tool - 2005 (PRAT-2005)* at http://www.bjs.gov/recidivism_2005_arrest/. (02/04/2016) [Accessed: 22-Feb-2017].

[6] Timothy Hughes and Doris James Wilson, "Bureau of Justice Statistics (BJS) - Drug Offenders in Federal Prisons: Estimates of Characteristics Based on Linked Data," *Bureau of Justice Statistics (BJS) - Drug Offenders in Federal Prisons: Estimates of Characteristics Based on Linked Data*, 27-Oct-2015. [Online]. Available: https://www.bjs.gov/index.cfm?ty=pbdetail&iid=5437. [Accessed: 22-Feb-2017].

[7] Sam Taxy, Julie Samuels, William P. Adams, "Reentry Trends in the United States," *Bureau of Justice Statistics Reentry Trends in the U.S.*, 20-Aug-2003. [Online]. Available: https://www.bjs.gov/content/reentry/reentry.cfm. [Accessed: 22-Feb-2017].

[8] E. Ann Carson, Ph.D., Elizabeth Anderson, "Bureau of Justice Statistics (BJS) - Prisoners in 2015," Bureau of Justice Statistics (BJS) - Prisoners in 2015, 29-Dec-2016. [Online]. Available: https://www.bjs.gov/index.cfm?ty=pbdetail&iid=5869. [Accessed: 22-Feb-2017].