Emily Cunningham
INFO 628: Data Librarianship and Management
Fall 2024
Final Project Write-up

Federal Bureau of Prisons: Federal Inmate Complaints

Data Management Plan (v3, 12/16/24)

Storage and Backup Procedures

Following the 3-2-1 rule, data will be securely stored in three separate locations including Google Drive, the desktop of my laptop, and a GitHub repository. This approach will ensure data redundancy.

File Formatting and Naming

All raw data will be stored as CSV files. Python scripts will be saved as .py files as well as Jupyter Notebooks (.ipynb), which may help people who are less familiar with coding to see how the script works. R scripts will similarly be saved as .r files as well as Jupyter Notebooks for the same reason. Data visualizations will be saved as PDF files.

Files will use PascalCase and will be named logically. If files need versioning, they will keep the same number prefix and title, but a version number will be appended to the end of the file name (e.g., _v02). Files names will be kept short and will not have spaces or special characters.

Files will also be versioned in their GitHub repository, where changes in code can be tracked and modified with descriptive commit messages to document which changes were made when.

Project Organization and Structure

There will be three folders: data, src, and results. All raw data that is provided for the user will be found in the data folder. The src folder will contain all Python and R scripts, along with the Jupyter Notebooks. The results folder will contain 2 subfolders: 1) a data folder where the Python script outputs will be saved and 2) an analysis folder where the R outputs will be saved. The data folder will contain CSVs with subsets of the original data provided by BOP. The results folder will contain CSV reports and PDF visualizations which will help summarize the data for a specific state.

Documentation

Scripts will be saved as Jupyter Notebooks so that markdown cells and inline comments can serve as a form of in-depth documentation.

A README file will be created for the GitHub repository, which will outline the entire project, and provide context as to what each file is, where it came from, and how it can be used. Jupyter Notebooks will be uploaded with their output, so that viewers can see the intended output and compare their results.

The Data Liberation Project provided a data dictionary for the column headers as well as some code translations for the facilities and complaint subjects. These documents will be cleaned up and some documents will be consolidated.

I will also keep a private research journal where I will keep track of each step taken during the research process, that way I can assure that the README file has information about various programs and settings that are necessary in order to get the scripts to work properly.

Methodology (10/21/24)

Data Source

I will be using a dataset titled Federal Inmate Complaints, which was published in July 2024 by the Data Liberation Project, which is an “initiative to identify, obtain, reformat, clean, document, publish, and disseminate government datasets of public interest.” The Data Liberation Project filed a request to the Federal Bureau of Prisons (BOP) asking for a copy of all database records pertaining to the organization’s Administrative Remedy Program, which allows inmates to seek formal reviews of issues relating to any aspect of their confinement. The BOP was not able to release the entire set of records, due to privacy concerns, but they did agree to release a substantial subset of data points for each of the 1.78 million complaints/appeal submissions made between January 2000 and May 2024.

Research Question

I am interested in investigating whether there is a correlation between the type of complaint and the amount of time it takes to be resolved. Similarly, is there a correlation between the type of complaint and the likelihood that it will be rejected?

Methodology

This study will employ a quantitative methodology, given the nature of the data. I will use a Python script to gather the New York subset of data. I will use OpenRefine to tidy the data and replace headers codes with their plain text version so that the data makes more sense to me. I will then conduct a statistical analysis using R to identify potential correlations, accompanied by visualizations that effectively illustrate the findings.

Data Gathered & Created

The Data Liberation Project provided a CSV file with all complaints released by BOP between 2000 and 2024. Because the CSV has over 1.7 million rows, I thought it would be better to focus on a subset of the data, so I decided to only evaluate complaints that took place in facilities located in the state of New York. The facility codes provided by BOP/The Data Liberation Project only came in the form of a PDF and did not include information about the facilities’ cities or states. I am currently working to create a CSV file that has facility codes, facility types, facility names, facility cities, and facility states so that the facility data can be in a more usable format.

I wrote a python script that uses my facility code CSV to create a set of facility codes for facilities located in New York. The script then uses that set to find all complaints that took place in a New York facility and writes those complaints into a new CSV, which will be my primary source of data for this project.

My next steps are to use OpenRefine to wrangle the CSV and get a better understanding of the data and to start experimenting with R to analyze the data.

Analysis (11/26/24)

This semester I started off thinking I would use the BOP Federal Inmate Complaint Dataset to do more more advanced statistical analysis using R, but during the long process of making the dataset more usable for myself, I switched gears to focus more on making the dataset more usable for the general public.

I am continuing to focus on NY submissions, but I am adding inline comments to the Python script that could enable someone with basic coding knowledge to create a similar enriched and expanded dataset for another state.

Enriching the Dataset

I have enriched the original dataset by adding several columns that make it easier to understand the complaint more holistically, taking each submission into account. The subcount column contains the number of total submissions associated with that one case number. The rejcount column counts the total number of rejected submissions associated with that one case number. The cldclocount column counts the total number of submissions with a “Closed Denied” or “Closed Other” status update for that one case number. The clgacccount column counts the total number of submissions with a “Accepted” or “Closed Granted” status update for that one case number. These few columns can help users better understand the history of the complaint, without having to sort by case number to count the number of submissions. I also added three columns that help users better understand the time frame of the complaint from beginning to the most recent status update. The earliestsitdtrcv column shows the earliest submission date for all submissions with the same case number. The latestsdtstat column shows the latest date a status was assigned for all submissions with the same case number. The daysbetween column calculates number of days between earliestsitdtrcv and latestsdtstat, which shows users how long an inmate spent trying to get their complaint resolved.

Expanding the Dataset

As I was initially evaluating the data I had a really difficult time understanding what I was looking at because of the various codes used as column headers and values. I decided to create an expanded version of the dataset with translations of all of the codes for my own personal use, but then I thought that if an expanded dataset is helpful for me, it will likely be helpful for someone else as well. I included cleaned up code dictionaries and the python script which creates the expanded dataset.

Creating a Subset of Unique Complaints

When I was starting to do some basic summary statistics on the dataset, I realized that if I was trying to determine which complaint type was the most common in NY facilities between January 2000 and May 2024, my data was going to be skewed. Many complaints have more than one submission associated with them, and some complaints have over 10 submissions. I needed to create a subset of the NY submission data that contains only unique complaints. I chose to only

include the submission with the most recent status update date for each case number. With the addition of the subcount, rejcount, cldclocount, and clgacccount columns, I could get a pretty good idea about the history of the complaint, just by looking at the most recent submission.

I also decided to make an expanded version of the unique NY complaint dataset with translations of all of the codes for the columns and values.

Summary Statistics and Visualizations

When I finally got the datasets to a good place where I could understand the data and start to think about what questions I could ask of the data, I decided to change directions with the statistical analysis. I first wanted to do some more advanced statistical analysis, but I thought that if my data cleanup was more geared towards helping the general public view the data, then I would like to gear my analysis towards that same audience as well. I decided it would be more in line with my goal to instead try to create some visualizations based on simple statistical analyses to make the data easier to understand and to showcase what kinds of questions users could ask in regards to other states/regions. I have started playing with R to make some of these visualizations, but the learning curve feels quite steep. I may switch my approach and use either the Matplotlib or Plotly Python libraries.

Visualizations

- What is the average number of rejected submissions per complaint case number?
- What are the most common “status reasons” for rejected submissions?
- What percentage of unique complaints only have rejected submissions, i.e. the complaint was never properly addressed and therefore the inmate never received any resolution or explanation for why the issue could not be resolved.
- Which complaint types had the most “closed denied” or “closed other” status updates, i.e. which complaints were most likely to be denied.
- Which complaint types had the most “closed granted” or “closed accepted” status updates, i.e. which complaints were most likely to be granted/resolved.
- Which complaint types have the largest average daysbetween value?
- For each of the top 10 complaint types, what percentage are “closed denied”/”closed other” vs “closed granted”/”accepted” vs only “rejected” without being closed?
- Which NY facility has the largest average daysbetween value?

Conclusion

Initially I intended on doing more intense quantitative analysis on the BOP dataset, but as I realized how complex and extensive the dataset was, I thought that I might not be the best person to analyze the trends. I changed the focus of my project to making this dataset more accessible to the general public.

At my last check-in, I still thought that I would have to keep the scope of the project narrowed down to just submissions made in New York State facilities, but as I completed the script, I realized that I could allow users to select which state to evaluate, which also meant that I could make an option to create enriched/expanded datasets for the entire dataset as well.

I also wanted to provide some basic reports and visualizations to help make sense of the enriched and expanded datasets. I mostly focused on the most common Primary Subjects and Secondary Subjects for the unique complaints.

I was able to find some simple data on the total population per year for all of BOP. I used that data to create two additional visualizations which show changes in numbers of unique complaints and Primary Reasons over the years, while taking the fluctuations in population into account.

I struggled near the end of the project when checking to see if my code was reproducible. Originally I wanted to provide the user with the BOP/Data Liberation Project dataset, so I uploaded the dataset to GitHub as a LFS (Large File Storage) object, which seemed fine at the beginning of the project. When I tried to run the script on another computer, I realized I needed to include instructions for downloading Visual Studio Code, Python, and R, especially since my intended audience is the general public, most of whom I would assume are not super familiar with these software applications. I then realized that I would need to include instructions on downloading the LFS, which included installing Homebrew, which started to feel unnecessarily complex. I resolved to allow the user to download the data on their own, so I hope that the Data Liberation Project keeps that link live for a while! This project was very useful in helping me understand the difficulty of creating projects which are reproducible on other systems. There were many small issues that required troubleshooting, and I would have never known they existed had I not tried to start fresh on a different computer.

I also thought initially that I would upload the NY outputs, that way the user could understand what should happen after running the scripts. If I had more time, I would want to add PDFs of the NY visualizations to the Wiki and maybe also find a way to include a few rows of the CSV outputs.

Additionally, if I were to spend more time on this project (which I may in the future!) I would like to expand the Wiki to include more information about the nuances of the data. I would also want to create a more expansive documentation document, but that does feel unnecessary, given the Data Liberation Project's documentation.

Lastly, I would like to find more in depth data relating to the populations of each facility each year. This would better allow me and other users to compare the number of complaints and types of complaints from state to state or facility to facility, which is not possible with the data currently.

Overall I am pleased with the work I have completed and if I can clean up the Wiki, I would like to reach out to the Data Liberation Project to see if my scripts may be useful to others!