Emily Cunningham
INFO 628: Data Librarianship and Management
Fall 2024
Final Project Check-in #1

New York Federal Inmate Complaints

Project Overview and Methodology

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.

Files thus far:

- 01_RawComplaintFilings.csv (Provided)
  - The raw complaint filing data provided by BOP/The Data Liberation Project
- 02_FacilityList1.pdf (Provided)
  - A list of facility codes and names provided by BOP/The Data Liberation Project
- 03_FacilityList2.pdf (Provided)
  - A second list of facility codes and names provided by BOP/The Data Liberation Project
- 04_FacilityCodes_All.csv (Created)
  - A consolidated list of all facility codes, names, cities, and states, created by me
- 05_CreateNYComplaints.ipynb (Created)
  - A python script that creates the NYComplaints.csv
- 06_NYFacilities.csv (Created by Python script)
  - A CSV of all facilities in New York. An output of 05_CreateNYComplaints.ipynb
- 07_NYComplaints.csv (Created by Python script)
  - A CSV of all complaints that took place in a New York facility. An output of 05_CreateNYComplaints.ipynb

Access files from my GitHub repository:
https://github.com/emilyalcu/2023-10-20_federal-inmate-complaints-ny

Data Management Plan

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 and R scripts will be saved as Jupyter Notebooks (.ipynb) in order to offer researchers inline documentation, which should enhance the clarity and reproducibility of the code. Scripts will also be saved as .txt files, to ensure long term accessibility, in the case that .ipynb files become unopenable in the future. Data visualizations will be saved as PDF files.

Files will use PascalCase and will be named logically. Each file will have a number prefix (01_, 02_, 03_, etc), which will help me and future users to understand the order in which the various documents are used/created. 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. All files are currently nested in one single folder. This may change as the project progresses.

Files will also be versioned in their GitHub repository, where changes in code can be tracked and modified. I will write more descriptive commit messages that can help others and myself.

Project Organization and Structure

I hope to keep my number of files fairly low, so that they can all live in one project folder. My use of number prefixes in the file names should help future users better understand the workflow organization and streamline its future use. As the project's complexity potentially increases, I may switch to an organizational scheme that includes four subfolders (src, data, results, docs).

Documentation

Scripts will be saved as Jupyter Notebooks so that 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 provides some codebooks for the column headers, facility codes, and complaint subject codes. Some of these will have to be cleaned up a little bit. The raw codebooks along with my wrangled versions will all be included in the project folder.

I will also keep a 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.