Process

In order to build a master data set, I first used the Hein Online data and matched it with Popnames, USLeg, and Significant Legislation Databases in order to make sure Hein was not missing any important acts. After realizing that the Hein Online Database could be improved by merging it with the GPO Database, I merged these two databases. They overlapped between 1951 and 2000, with GPO supplementing the years between 2000 and 2016 and Hein supplementing the years prior to 1951. Then, I rematched with Popames to see if the amount of unmatched data had been improved.

Problems Come Across

One problem I noticed is that the Hein Online database skipped some acts that appear on the pages in the table of contents. This made them not appear in the database that was originally collected, as I believe the database was compiled using the table of contents. These need to be included in the new master dataset if it was not supplemented by the GPO data. Another issue I encountered was with public law numbers and dates. Often, the GPO dates and the Hein dates would not align by a day or two, even though they were referring to the same law. This is because often the Hein data would be using the wrong public law number to obtain the date, even though we found a case where the GPO data had actually used the wrong date. There were 753 cases in which this happened, which still gives the data 99% accuracy. For now, the date as written in the GPO data is reflected in the set as “date_of_passage” and the date from Hein Online is reflected as “secondary_date.” Another problem that affected mostly laws in the middle period (about 345) occurred when I was matching the master set with Popnames. The popnames data would often choose a sal_page_start and a Title that occurred in the middle of a larger act, and thus the acts would not match. Sometimes, two acts were recorded in Popnames that would be considered one act in Hein. Generally, this happened when there were larger expenditure bills. Another problem was typos and general errors from the other databases. When there was a typo, I verified the true value and fixed the dataset, which is reflected in the code. An example of general errors was a few laws coded as from 1941 in the GPO database, even though it began in 1951. These laws were from 2007 and 2008, so I fixed that as well. Finally, some unmatched data occurred because of how the different databases definitionally defined laws. There were mixed private laws, amendments, proclamations, organization plans, etc. mixed into the other databases.