For my DSC540 term project, I aim to examine the trends and interconnections between crime and incarceration in the USA across recent decades. The initial dataset is sourced from Wikipedia, presenting a "list of US states and territories by violent crime rate," covering the years 2011 to 2020. Comprising 12 columns and 53 rows, the definition of violent crime (according to the article) consists of murder and non-negligent manslaughter, rape, robbery, aggravated assault, and gang violence

The second dataset, downloaded from Kaggle, titled "crime_and_incarceration_by_state," amalgamates crime and incarceration statistics in each US state from 2001 to 2016. This comprehensive dataset features 17 columns and 816 rows. The respective columns are jurisdiction, includes_jail, year, prisoner_count, crime_reporting_change, crimes_estimated, state_population, violent_crime_total, murder_manslaughter, rape_legacy, rape_revised, robbery, agg_assault, property_crime_total, burglary, larceny, vehicle_theft.

The final dataset will be obtained through the FBI Crime Data API, specifically the "GET/arrest/national{offense}" section, with 31 columns ranging from aggravated assault, arson and burglary to vagrancy, vandalism and weapons: carrying, possessing, etc.” It also allows you to pull up the data for each of the US states. In the spirit of transparency, I’m still trying to figure out how to use this API.

Although I initially considered using "years" as the relational factor, I ultimately opted for "states" as the common denominator across all datasets to facilitate meaningful comparisons and analyses.

- Website: https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_violent_crime_rate
- Flat file: https://www.kaggle.com/datasets/christophercorrea/prisoners-and-crime-in-united-states/data
- API: https://cde.ucr.cjis.gov/LATEST/webapp/#/pages/docApi

After meticulously gathering and assessing the three datasets, it is evident to me that each of them possess a commendable level of solidity. So to speak. The collected datasets in my opinion exhibit a high degree of accuracy, consistency and completeness, not to mention that there seems to be almost no instance of missing or erroneous information. Which to me suggests that there might be minimal cleaning and transforming to do. It seems that I might have my task cut out of me in trying to figure out what to clean and how to transform the data. If anything, the true challenge it appears, lies in the impending task of merging these datasets into a unified dataset. For me, it is this process that is likely to demand careful consideration and meticulous attention to detail, as ensuring seamless integration while preserving the integrity of the information will be crucial.

Analyzing data related to crime and incarceration in the United States over the years carries significant ethical implications. One concern revolves around potential biases in the data collection process, reflecting systemic inequalities within the criminal justice system. Discrepancies in law enforcement practices, sentencing, and social dynamics can skew the dataset, potentially perpetuating stereotypes or reinforcing existing prejudices. Moreover, issues of privacy and consent arise when dealing with individual-level crime data, as the release and utilization of such information may compromise the rights and confidentiality of those involved. Responsible handling of this data requires transparency, anonymization, and a meticulous approach to avoid perpetuating societal injustices or stigmatizing certain communities. Additionally, ethical considerations extend to the potential misuse of data, raising questions about the balance between public interest, research, and the protection of individuals' rights in the realm of criminal justice data analysis.