1. What is in your data?

Federal defendant charges and sentences over fiscal year (October 2018 to October 2019) in the United States. The data set is sourced from the U.S. Sentencing Commission, which collects data on federal sentencing trends annually from all federal District Courts across the country. For the purpose of generalizability, and to avoid applying the conclusions to defendants convicted of unrelated offenses, we have narrowed down the population of interest to only defendants who had been convicted of trafficking either powder or crack cocaine. We are not distinguishing between the two groups of cocaine (crack versus powder) because following several amendments, the Sentencing Guidelines indicate that similar quantities of crack and powder cocaine should be sentenced similarly.

2. How will these data be useful for studying the phenomenon you're interested in?

We have two ideas for what we'll be diving deeper into. Since it's a massive data set, we want to narrow it down somehow. The first way we were thinking of doing that is by filtering the data to a crime genre (like cocaine trafficking) and discerning patterns based on explanatory variables like age, plea deal, conceal/open carry, district courts etc. We are particularly interested in plea deals based on our EDA suggesting that subjects who accept a plea deal have shorter sentences. We'd like to dig into this to see if, for example, the plea deal subjects are from a more affluent background than those of non plea deal. Another idea we have is to take all of the data (or maybe random sampling so we don't have an overwhelming amount of data) and discern patterns based on the same exploratory variables as above. This would also be interesting because we could examine, for example, the amount of plea deals between white collar crimes and drug trafficking.

3. What are the challenges you've resolved or expect to face in using them?

The dataset is very large, which could make it difficult to explore the data, but we hope to resolve this issue by narrowing down our focus. Additionally, the variable names are not intuitive and identifying them requires tedious investigation of the codebook. We are planning to rename the variables to make our research easier. There are also some missing values, and the code book does not address clearly how to go about handling these. We will have to come up with a method of cleaning the dataset to filter out these missings. The data dictionary is almost useless because it is extremely lengthy and requires a lot of sifting through to obtain the context needed.

The Sentencing Guidelines ensure that judges uniformly sentence defendants with identical criminal histories and severity of crime. However, despite the application of the Sentencing Guidelines, some disparities are still present in the data. Although we require further EDA and analysis to pinpoint the source of the disparity, sentencing is generally considered to be inherently unpredictable. Perhaps the capriciousness of sentencing decisions shows that the Guidelines are operating properly, as judges have the discretion to individualize sentences according to the circumstances of each defendant – such as the presence of mitigating or

aggravating factors. On the other hand, this makes creating a predictive model difficult, as the data may not fully capture all nuances in the observations.

KEY VARIABLES

Variable Name	Definition
SENTTCAP	Length of incarceration (prison sentence length) for individual defendants, in months. Missing values and zero terms are not included.
AGE	Age of the offender at the time of sentencing, in years.
MONSEX	Gender of the offender. Levels: 0 (male), 1 (female)
MONRACE	Offender's race, as recorded by the court.
NEWEDUC	The highest level of education completed by the offender. Levels: 1 (less than H.S. graduate) 3 (H.S. graduate) 5 (some college) 6 (college graduate) *May be re-coded in the EDA to just non-HS, HS, and college/above*
WGT1	Total quantity of drug carried; standardized in grams across all types.
BASEHI	Base Offense Level (unit: BOL), which corresponds to the severity of an offense. Higher BOL indicates more severity; for example, a charge of weapon enhancement increases the BOL and results in a longer sentence.
IS924C	Weapon enhancement charge; shows whether the defendant has a 18 U.S.C. § 924(c) conviction. Levels: 0 (no charge – no weapon enhancement), 1 (at least one charge of weapon enhancement)
NEWCNVTN	Shows whether the defendant's case is settled by plea agreement or trial. Levels: 0 (settled by plea agreement) 1 (trial; no distinction between bench and jury trials)
ADJOFLHI	Adjusted Offense Level (from BASEHI)
CITIZEN	Citizenship status of the offender. 1 = US Citizen 2 = Resident/Legal Alien 3 = Illegal Alien

	4 = Not a US Citizen or Alien Status is unknown 5 = Extradited Alien
Add more as we go	