In this exercise, we want to learn a little about how you problem solve and get a sense of your coding style. The data you’ll be working with is real data, so this exercise is very similar to the data cleaning work you’ll do on the team. Because we want to get a sense of how you think, clearly documented code is more important than a perfect solution. Please try to spend no more than two hours on this exercise, and be prepared to discuss your approach in the interview.

This exercise involves cleaning real data about jail populations in 10 counties over time. Please send a cleaned data file, the code used to produce it, and a brief explanation of your approach. (For the last part, inline comments are fine, as are short sentences, illustrative charts, or anything else you used to solve this problem.)

The [attached dataset](https://drive.google.com/file/d/1HmI-CkujM_9LOS75ZN0GQmi84zoGu0_c/view?usp=sharing) is a real dataset of raw data reported by jails to the Bureau of Justice Statistics over the years. The columns are as follows:

fips: (FIPS code) A 5 digit unique identifier for every county in the county. A comprehensive list of all current FIPS codes can be found [here](https://www.nrcs.usda.gov/wps/portal/nrcs/detail/national/home/?cid=nrcs143_013697)

facility\_name: Taken from the most recent year a facility submitted data to BJS, the most recent name of the jail

year: The year in which the data was collected

rated\_capacity: The capacity of the jail facility

total\_confined\_pop: A single day total population count of the jail, usually taken on June 30th of a given year

adp: Average daily jail population (annual average)

confined\_women: A single day total population count of the number of women incarcerated in the jail, usually taken on June 30th of a given year

confined\_men: A single day total population count of the number of men incarcerated in the jail, usually taken on June 30th of a given year

admissions\_year: Total jail admissions in a given year

admissions\_week: Total jail admissions in a given week of the year (often a week in June)

admissions\_day: Total jail admissions in a given day of the year (often June 30th of a given year)

discharge\_year: Total jail discharges in a given year

discharge\_week: Total jail discharges in a given week of the year (often a week in June)

discharge\_day: Total jail discharges in a given day of the year (often June 30th of a given year)

The data is collected through surveys distributed to the jails. It is commonplace for those filling out the surveys to put numbers in the wrong columns or just write nonsense. A good way to try to understand if a data point is valid data or nonsense is to look at other columns that might be related. For example, if the total\_confined\_pop dramatically increases one year one could check the adp and see if that also increased. If there is a mismatch between those two variables one could check to see if the rated capacity increased, suggesting the jail might have expanded, or look up the jail and see if there are any news reports from that year that might explain the data anomalies. The same follows for the gender variables, the confined\_women and confined\_men should add up to the total\_confined\_pop as those numbers were taken on the same day (although it is pretty common for the numbers to be off by a little). Also, there are rarely more people classified as women than as men in a jail unless it is designated a women’s facility.

For the admissions and discharge data we would like you to condense the 3 columns for each variable into one yearly admissions variable. Therefore the output datafile should have 4 fewer columns than the datafile attached. Annual admissions and discharge will always be much higher than adp and total\_confined\_pop as many people get booked and spend days or weeks in jail before they are discharged.

**Assignment**: take the given dataset and create a cleaner dataset for researchers.

* Read in datafile
* Make all admissions and discharge data comparable on the year level and drop weekly and daily data
* Write code that identifies anomalies in the data (for example, depending on the spread of the data, we often use a 20-40% year over year change threshold to find anomalies, feel free to use this or your own way to find anomalies)
* Compare anomalous data to other columns or research the facility in order to understand if data should be omitted
* Interpolate all data points that were either originally missing or you decided should be omitted (by default we use linear interpolation but for some fields such as gender that add up to another field we might take that into account- there is no wrong way to do this we just want to see how you think)
* Write data to new file

There is a lot to accomplish in this exercise, but please limit yourself to 2 hours. It is okay if you have not found all the data errors in the allotted time, as we’re just trying to get a sense of the code you write and how you think about problems such as data cleaning. Please write all code in Python. We usually use the pandas library when doing this kind of work, if you are familiar it would be great if you could show us that, otherwise work with whatever is most comfortable for you, pandas is not a necessity and can be picked up easily on the job. Please feel free to reach out with any questions at all.