

# Bond Factors for CT Pretrial Detainees

15 December 2018

## Abstract


For the following project reports, analysis will start with the file created in this paper, detainees\_classed\_offenses.csv. This file was created using the most recent documented record for each identifier and each “latest admission date,” the presumed date of the most recent arrest.

The new “classed” dataframe contains 28,808 rows and 13 columns. Three of these columns determine the setup of the dataframe: download\_date (date of the last record), identifier (detainee), and latest\_admission\_date (either arrest date or original entry into system). Seven columns are directly chosen from the groupby over the first three: race, gender, age, bond\_amount, offense, facility, and detainer. The last three columns are calculated in the earlier steps: offense\_class, days, and counts.

## Cleaning Steps

First, I read in the download\_date and latest\_admission\_date columns as dates while reading in the file. Then I lowercased the column names and changed spaces to underscores to make calling columns easier. Next, I changed to category types for all columns with a discrete number of options - gender (2), race (5), facility (50), and detainer (9).

I needed two additional pieces of information, the days detained and the felony/misdemeanor classification, if any provided. I created a time math column (days) subtracting the latest\_admission\_date from the download\_date for days detained and a backup column (counts) to count the number of rows in which each detainee appears. Then, I separated out the last two characters of the offense column for the offense\_class column. Connecticut has Class



A, B, C, and D felonies and misdemeanors. They can also have U (unclassified) or no class for both felony and misdemeanor offenses for a total of 10 classification options plus NaN when no classification was provided in the dataset.

Once these columns were created, I needed to condense the data. The dataset contained one row for each detainee each day over the ~27 month download and started at 2,677,123 rows. In order to preserve the most relevant information, a new dataframe was created using a groupby over the identifier (detainee) and the latest admission date (to account for repeat offenders) and populated from the data on the maximum download date for the groupby. This final information was chosen as charges, offense classifications, and bond amounts can change due to too many factors to control for, including prosecutor negotiation, placeholder values, and new or inadmissible evidence. This condensation brought the list to 40,147 unique rows and 29,531 unique detainees, or just under 1.4 stays per detainee.

One additional step, separating the rows that recorded felony or misdemeanor from the rows that left this information off, resulted in two final dataframes. The dataframe “not\_classed” will be preserved but will not be used in further steps. The dataframe “classed” brings our total number of rows to 28,808 with 22,663 unique detainees, or just under 1.3 stays per detainee.

## Missing Values

As this was a dataset owned by a state department of corrections and subject to strict recordkeeping requirements, I didn’t need to account for unexpected missing data in the rows. Below is a summary of the two columns I felt necessary for further exploration.

Although not “missing,” the information on felony or misdemeanor and associated class was included as a two-digit code at the end of the offense value. These values have been pulled out into a separate column so I can use them to divide the data for analysis.

About 28% of rows don’t have a felony or misdemeanor classifier after condensing. (60% of those are violation of probation arrests.) Since this project focuses on felony and misdemeanor detainees, these rows will be preserved but not used in further analysis.

I’ve also added “days” and “counts” columns that provide information on the number of days the detainee was incarcerated before bailing out, receiving a verdict, or having charges dropped.

Detainees still in the system on the download date and detainees admitted before an as-yet-undecided date will be filtered out for any length-of-stay analysis.

## Outliers

After cleaning, 93 offenses have bond amounts less than \$100 (16 have less than \$10). (From API: "BOND AMOUNT: Amount of bond for which the inmate is being held. In some instances, for particularly low (less than \$100), this bond amount may be considered a place holder [sic] value.") I don't think all of the sub-\$100 bond amounts are placeholders since we're using data from the last day of detention, and I haven't decided what I want to do with them, specifically in light of the law that we're examining. The law was put in place to reduce cash bonds for lower-income detainees.

Also after cleaning, 619 detainees have latest admission dates (ostensibly the date of the arrest) that predate 2014. From the API: "LATEST ADMISSION DATE: Most recent date in which the inmate has been admitted. In some instances, this may reflect an original date of admission to a correctional facility. Generally, if a date is more than one year old, an inmate should not be considered to have been held for the entire duration of that time." These rows are also useless for length-of-stay determinations, but they may be useful in other situations.

## Future Considerations

Two calculations present issues at this time. The "counts" column counts the individual rows in which a detainee shows up in the original data, and the "days" column performs time math on `download_date` and `latest_admission_date`. The state admits problems with the `latest_admission_date`, and at least one identifier has two gaps of about two weeks each between contiguous rows with no change in the admission date. One potential cause could be that the detainee was moved to a city or federal facility and was temporarily out of the state system. If this is the case, the detainee's information is still valid. This will be assumed true for this project and treated as a single stay.

I'm also using the age on the last day of the stay, and may consider using the age at arrest in future iterations.