CA\_datapile.xlsx was added 2 weeks ago. Source: https://www.counties.org/post/datapile

* Has up to date demographic data for California counties: sex, race, poverty and unemployment, median income
* Complemented our other sources and let us ask research questions about how county-level demographics intersect with patient outcome (mortality)

For regression we needed a quantitative outcome variable of interest.

Decided to take the IMI data and pull out cases and deaths due to heart attack by hospital and standardized by calculating # of heart attack cases/1000 and # of heart attack deaths/1000.

(IMI data had a lot of this type of information grouped by disease/treatment strategy. Heart attack data had some of the least missingness and it seemed like a well-behaved variable – heart attacks are well-characterized and onset quickly. Treatment plans have also not changed radically in the last 5-10 years compared to diseases like pancreatic cancer. We hypothesized that hospitals used similar treatments for heart attacks and major differences in mortality might be due to hospital quality and patient demographics)

Cleaning was straightforward – we transformed strings to factor variables and extracted features using regex. Of note: each hospital had a categorical performance rating and the rating had 3 levels: worse, better, and as expected.

Tidying – we took the thousands of rows of hospital, case, and demographic data and distilled it down to about a hundred entries. Each row is a unique combination of the cases/death rates per 1000 in hospitals in that particular county with that particular rating, with the additional hospital HPSA status and demographic info.

But generally the goal was to acquire data to construct models where we infer how death rates are affected by hospital ratings/HPSA status and county demographics, controlling for case rates, and we did this successfully.

*We removed hospital observational rows with missing entries in hospital quality* and note that this leads to bias*.* Here are other options we could have explored for missing data: mean imputation (we chose to not do this because this shrinks variance), multivariate imputation using the MICE algorithm in R to backfill missing entries (chose not to due to time constraints, but a good future direction).