Springboard Project 5.5 (Capstone Project Proposal: Data Cleaning)
Alexander Olden

Because one of my analytical approaches will be a logistic regression, my first data-cleaning step was to set the dependent variable (whether a patient was readmitted within 30 days or not) to a binary one. With that step completed, I was ready to check for any potential relationships between variables with high percentages of missing values and the dependent variable. Payer code and attending-physician specialty were each missing more than 50% of their values, so I checked for any relationships between their missing values and the outcome variable. Nothing emerged, so I dropped these two features. I also dropped weight as a feature, since more than 90% of its observations were missing.

The next issue I had to address was the presence of multiple encounters (hospital visits) for several patients. To maintain independence among observations, I sorted the data by encounter ID and kept only the lowest ID (which is the first one chronologically) for each patient. To further reduce bias, I removed encounters that resulted in death or discharge to a hospice facility.

After these steps to reduce bias, I looked at variables with smaller amounts of missingness: race and three diagnosis codes. Race had 3% of values missing, so I expected that dropping it would not present a serious issue, but I first checked to see if its missing values to had any association with the outcome variable. They did not, so I dropped race. The three diagnosis codes each had 2% of values missing, if not less, so I comfortably dropped them as well. These codes also had more than 50 potential values, making the case for missing at random quite strong.

With checks on randomness and bias complete, I proceeded to categorize the values in the three features showing diagnosis codes. The codes can be grouped according to the condition they pertain to and thus changed from numbers to descriptive terms. Using information provided with the initial dataset, I changed the codes into the medical conditions they pertain to. This adjustment will also expedite final analyses that rely on categorization (and one-hot encoding) and do not require the level of granularity that comes with initial diagnosis codes.

The final cleaning step that I took was to change all remaining missing values to NA, as they had initially been coded as question marks. From there, I checked on the balance of the remaining dataset and found a ratio of 10:1 for positive and negative results – certainly something I will keep in mind going forward.