Capstone 1 H1B LCA Petition Data Wrangling

Fortunately, the dataset provided by Kaggle was already relatively clean. With that said, I did do some cleaning and rearranging of my own with the analyses that I am intending to run in mind. The data were downloaded from Kaggle as a csv and on import using Pandas one of the columns imported as “Unnamed: 0.” I found the online documentation for the dataset and determined that this should be “CASE\_NUMBER” so I renamed it and then made that the index field for the data frame.

The next step was to inspect the CASE\_STATUS field. This is the field that contains the outcome of the LCA process and is also the field that we are trying to predict. For these reasons, I wanted to give it special attention. After doing some searching online, it turns out that there are only four valid values for this field. These are: certified, certified-withdrawn, withdrawn, and denied. For the purposes of this analysis, I decided only to keep certified, certified-withdrawn, and denied. While it is helpful to know the outcome when an application was certified but then withdrawn, it is not helpful for prediction if all we know about the application is that it was withdrawn. Therefore, I believe that it is appropriate to drop records where CASE\_STATUS is “withdrawn.” I also wanted to condense these results so that I had, essentially, a binary success failure field. I created a new field (“CERTIFIED”) where all records with CASE\_STATUS values of “certified” or “certified-withdrawn” have the value “certified” and records with the CASE\_STATUS value of “denied” have the value “denied.”

Other cleanup included changing the datatype of YEAR from a float to an integer. I split the field WORKSITE into two separate fields using a comma as the delimiter. I then applied the strip function to both columns to ensure that there were no leading or trailing spaces leftover from the split function. Finally, I created a new column (LOG\_WAGE). This is populated with log10 value of the PREVAILING\_WAGE field. PREVAILING\_WAGE has values that span from 0 to over 10^9 so I determined that it would be easier to visualize this spread on a log-scale.

In addition to the cleanup of the data obtained from Kaggle, I generated two other datasets via calls to the FCC Area API and the Census Population Estimates API. First, I obtained a unique list of latitude and longitude pairs from the cleaned H1B dataset. I then used these to call the FCC Area API. This resulted in a dataset that contains latitude, longitude, census block Federal Information Processing Standard (FIPS) code, population of the census block, and other county and state identifiers including the county FIPS code. I then used this dataset and, with the unique list of county-FIPS codes, called the Census Population Estimates API. The resulting dataset included the population of each county in my dataset. Finally, I merged these two datasets so that I had one overall csv containing location information including FIPS codes for counties and states, as well as county population, and census block population.