Data Wrangling: Solar Survey Capstone project

The data I chose to work with for my Springboard Capstone project was three separate datasets from surveys filled out by three different groups: households that have purchased solar array systems for their house, a group of households that are considering purchasing a solar array system, and a final group that has never considered purchasing a solar array system.

There are about 49 survey questions in common with all three of these datasets, and I thought it could be possible to find patterns in the answers of the households that had purchased solar array systems, and find similar patterns in the other groups, and predict who may be a future purchaser of a system.

I first brought in all three datasets into ‘adopters’, ‘considerers’, and ‘gps’ groups (terms originally used in the survey study). I then took a look at the features and datatypes of each of these datasets, and while there were from 120-160 features in these datasets, there were 49 that were in common with all three datasets, which would give me direct apples to apples comparisons of the different datasets. I then created new datasets of each containing only these comparable features.

I then began to look at the missing data within these datasets. Some features, such as the HOME feature in the adopters dataset was missing almost all data (95.32% missing data), along with 9 other features in that dataset, but the rest under 12% missing data. I decided to drop the HOME feature since so much data was missing, and then took a look at the other features to see what could be done to fill in the missing data. The other datasets were missing much less data, but I still had to decide on how to fill in that data as well.

Some features, such as the WINTER\_NOPV\_BINNED and SUMMER\_NOPV\_BINNED were, while categorical in nature and binned, could be averaged and filled in with that average. Some, such as GENDER and HAVE\_KIDS were so vastly more in one category (almost 3x more) that adding the few missing values could be put into the larger category without biasing the values too much. The rest had such even spread of categorical values, that I chose to use a forward fill to fill in the missing data, and then recheck the percentages to see if everything was still even. This seemed to keep the same balance to the percentages. I filled in the missing data for all three datasets in this way.

I also noticed that in some of the features there were outliers far away from the other values. These were all survey questions, somewhat ordinal in there ranking of the answers, from 1 to 5, with a catch all in some cases that the question was not answered, in such case a 99 or very high number was entered, resulting in these outliers. This greatly affected visualizing the distribution of the data through histograms, so I took any value 90 or higher and replaced them with NaN, so they would not plot and I could get better visualizations of the histogram distributions.

The data was now pretty much ready to visualize through histograms to look at their distributions. I was able to find some insights into these distributions, now that the data was filled in, and the outliers changed to NaNs.