Capstone Project 1: Data Wrangling

Springboard Data Science Track
Ishmail Grady
April 2019

Data Collection

The main dataset used to develop this model will come from the microdata of the 2015 Residential Energy Consumption Survey (RECS) Survey conducted by the U.S. Energy Information Administration and is publicly available for download as a CSV file. The raw data set contains 5,686 observations and 759 column variables. Given the large number of variables, a codebook that contains a description and variable type for all variables, and response labels and encodings for all categorical variables.

Data Cleaning

The raw dataset is relatively clean and tidy. Each row in the main dataset represents a respondent in the survey and each column corresponds to distinct survey questions or parameters of the survey construction. All entries column variables contain the expected data types. One column in the original dataset, corresponding to a localized conversion factor for natural gas from cubic feet to BTU. There is another column in the dataset that corresponds to an indicator variable for natural gas usage in the household of the respondent. If the household does not use natural gas, the localized conversion factor column is blank for any given row. Thus, these values are not truly missing, but will be considered as not applicable.

The survey was designed to be statistically representative of all US households. Each observation has an associated weight that corresponds to the number of households the observation represents. Thus, using these weights will ensure that any outliers in the data will be handled appropriately.

There is a subset of column variables that may contain imputed values where data was not available. The dataset contains 217 columns that represent imputation flags for 222 columns variables (some imputation flag columns correspond to multiple variables). These imputation flags indicate if the corresponding variable was imputed or not. A separate dataset was constructed that contains observations without imputed values. To do this, for each observation if an imputation flag column indicated an imputed value, the field for the corresponding variable was replaced with NaN.

There were 20 questions in the survey that have “Don’t Know” as a possible answer (encoded as -9 in the dataset) the values in the corresponding columns of the dataset were replaced with NaN for all questions that were answered with “Don’t Know”.

The rows with the new missing values will remain in the dataset in order to leverage all of the available data. An appropriate predictive model that accounts for missing values will be used in this project. The codebook that contains a description and variable type for all variables, and response labels and encodings for all categorical variables is given in excel file. The information contained in this file was extracted and saved for future use as it will be a valuable tool for understanding the large number of variables. Pickle files containing dictionaries for variable descriptions, variable labels, and mappings for variable responses were made.