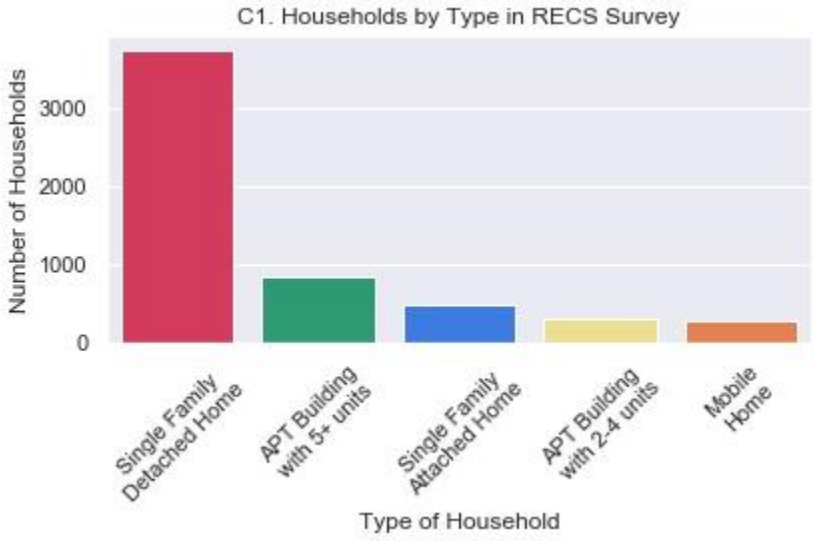
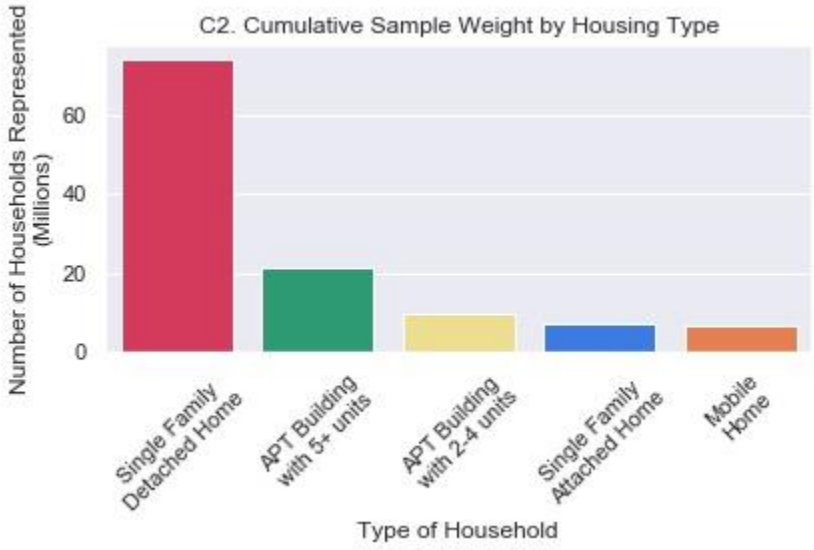
Capstone Project 1: Data Story

Springboard Data Science Career Track
Ishmail Grady
April 2019

Before developing a predictive model, some exploratory analysis was conducted in order to better understand the dataset and develop intuition that will inform model construction. Given that this dataset contains observations for a representative sample of US households, the first step in becoming familiar with the data is understanding what household types the dataset is comprised of.



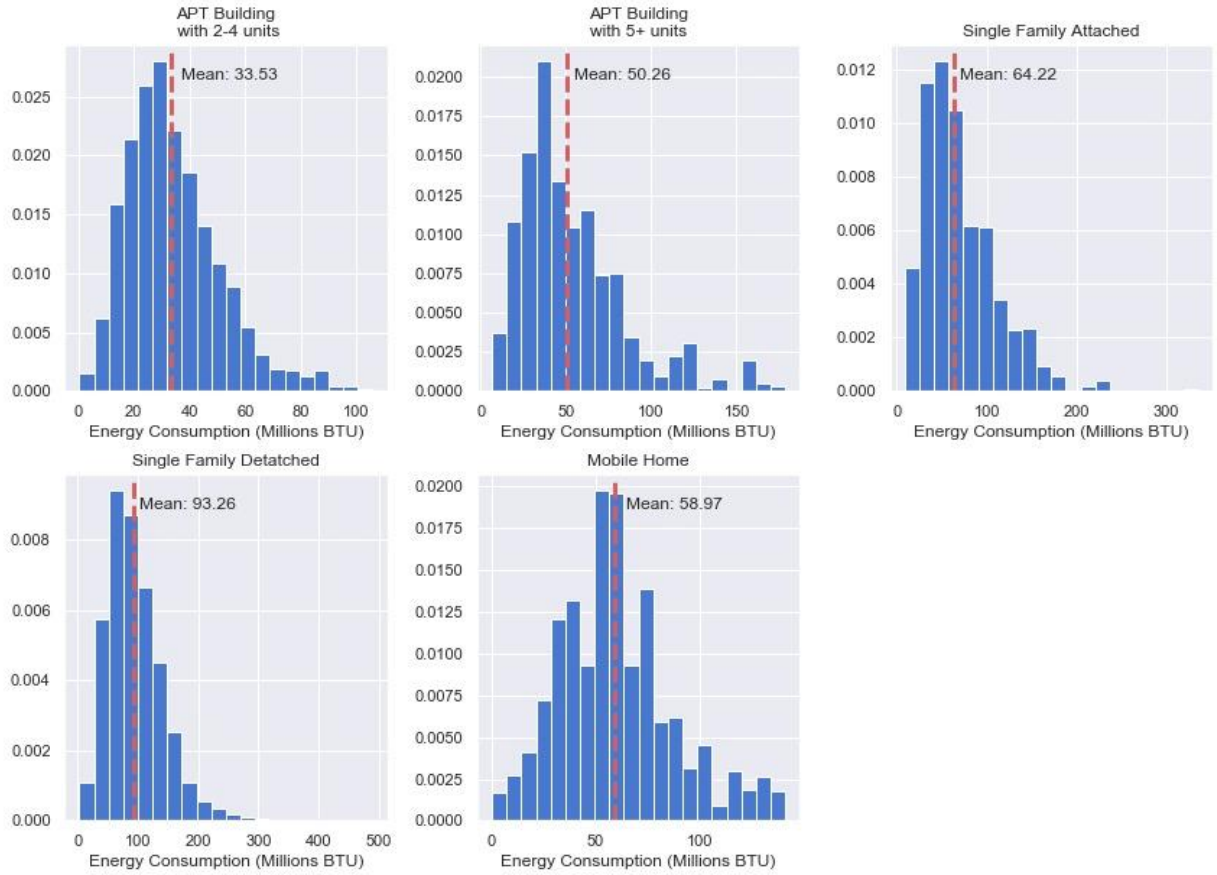


As shown in Figure C1, the majority of households in the dataset are single family detached homes and these households make up 66% of the 5686 observations in the dataset. Figure C2, shows

the cumulative sample weights of the observations by housing type. The amount of observations for each household type are roughly related to the cumulative weights of their observations as shown when comparing Figure C1 and C2. The 3,752 observations of single family detached households in the dataset are statistically representative of 73.8 million American households.

It is logical that the type of household would play a factor into how much energy it will consume. This intuition needs to be verified and it will be useful to know energy consumption differs across household types.

C3. Distribution of Total Energy Consumption by Household Type

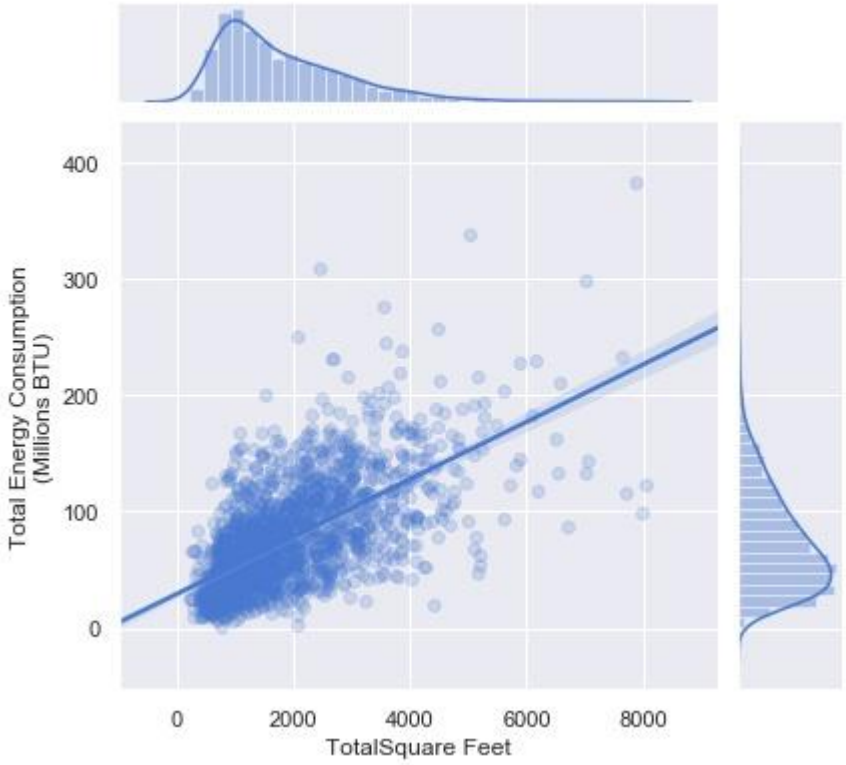


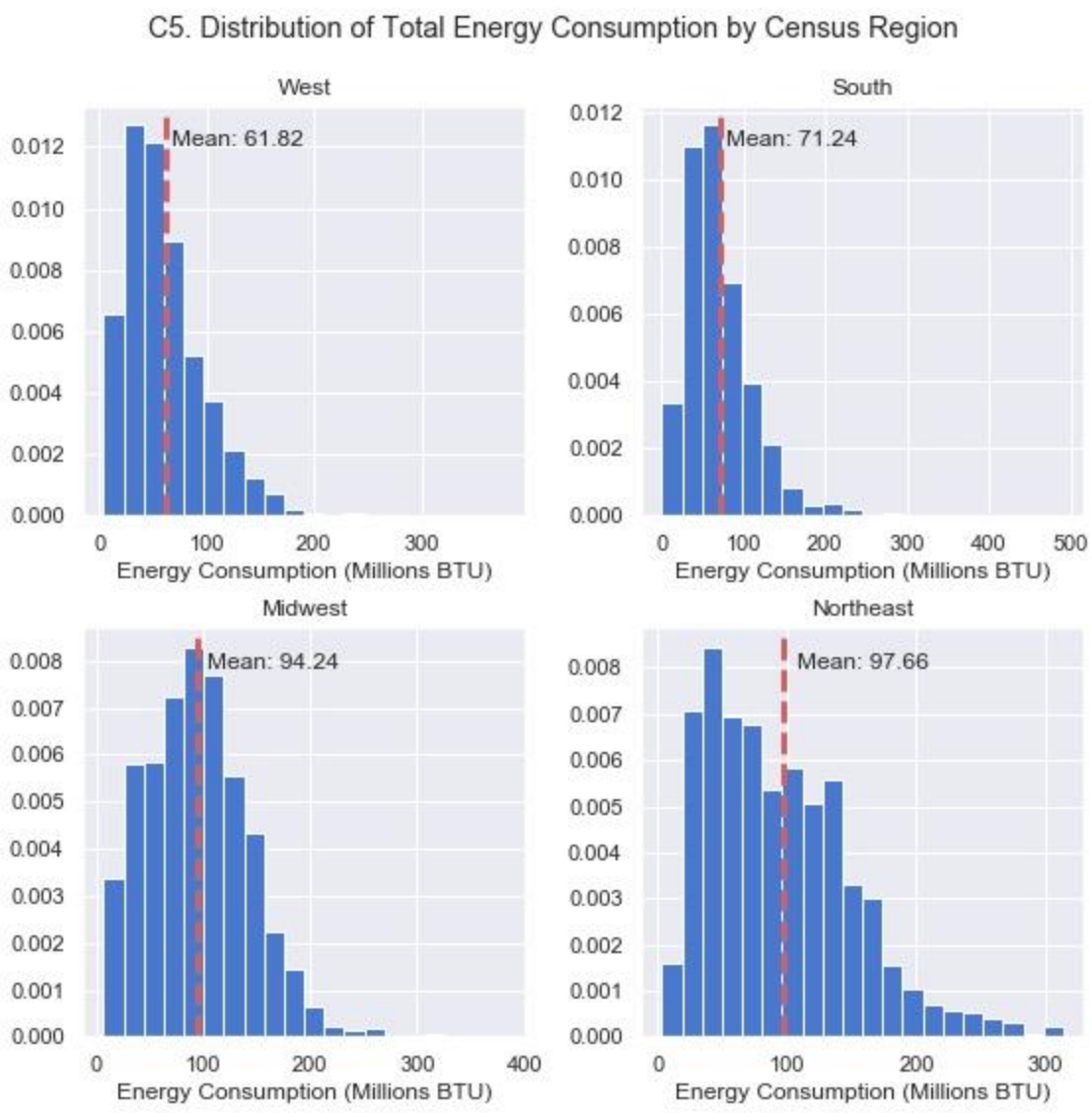
According to Figure C3, Single family Detached households consume significantly more energy than other household types. The energy consumption of these households also have a significantly larger variance. As shown in Figure C3, the distribution of energy consumption for each household type is positively skewed, with the exception of mobiles homes which have a more symmetric shape. On average, Single family detached home consumes over twice as much as apartments in buildings with 2-4 units and apartments in buildings with 5+ units. It is surprising to see that mobile homes have a higher mean total energy consumption than apartments, some additional investigation may be necessary to determine the exact cause of this.

To develop a useful model, it will be important to understand how the variables are related to the primary dependent variable, total energy consumption. The joint plot below in Figure C4 shows the relationship between total square feet in the household and total energy consumption. There appears to a moderate positive correlation between total heated square feet and energy consumption. In fact, these two variables have a pairwise correlation coefficient of .64. Total heated square feet in a household and total energy consumption have a correlation coefficient of .58. The correlation is not very strong, however, total square feet and total heated square feet have the highest correlation to total energy consumption of any numeric household characteristic.

This variable will likely be an important feature to include in the final predictive model however more investigation is necessary to fully understand how this relationship changes when conditioned on other variables.

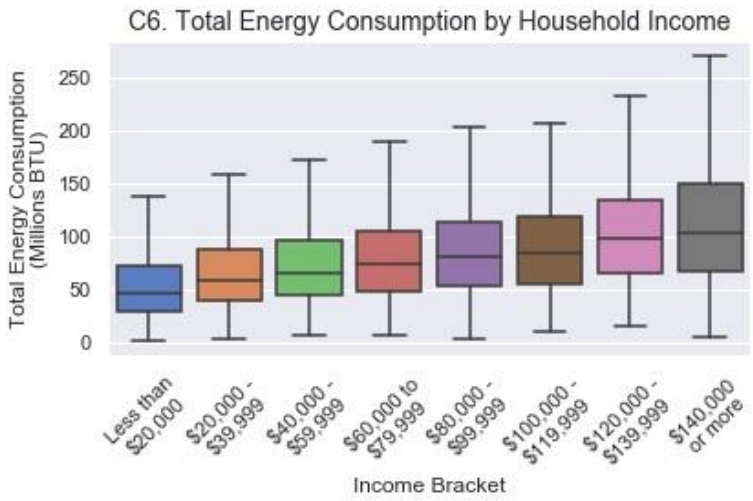
C4. Total Square feet v. Total Energy Consumption





Other important factors to consider in this model are geography and climate. According to Figure C5 shown above, there appears to be a significant difference in the means of the distributions of energy consumption across regions of the country. All four regions have positively skewed distribution. The distributions of energy consumption in the Northeast and Midwest appear to be similar however the distributions of the South and West have significantly lower means.

This may suggest that total energy consumption is closely related to the region that a household is in and that households in colder climates consume more energy than those in warmer climates. Further investigation should be done to quantify the significance of the difference in distribution and how other variable related to region or climate affect energy consumption.



The dataset also contains many demographic variables for each household including the yearly income of the household. Income can be a good indicator or consumer behavior and many different context so it is worth investigating how household incomes relates to energy consumption.

According to Figure C6, total energy consumption increases slightly with household income. There is little increase in energy consumption between adjacent income brackets, however the gradual increase in consumption across all brackets results in significant increases in consumption between non-adjacent brackets. The distribution of energy consumption within income brackets also becomes more widespread in higher income brackets. Further investigation should be done to understand what conclusions can be made about the relationship between energy consumption and household income in the context of other demographic variables such as family size, education level and age.