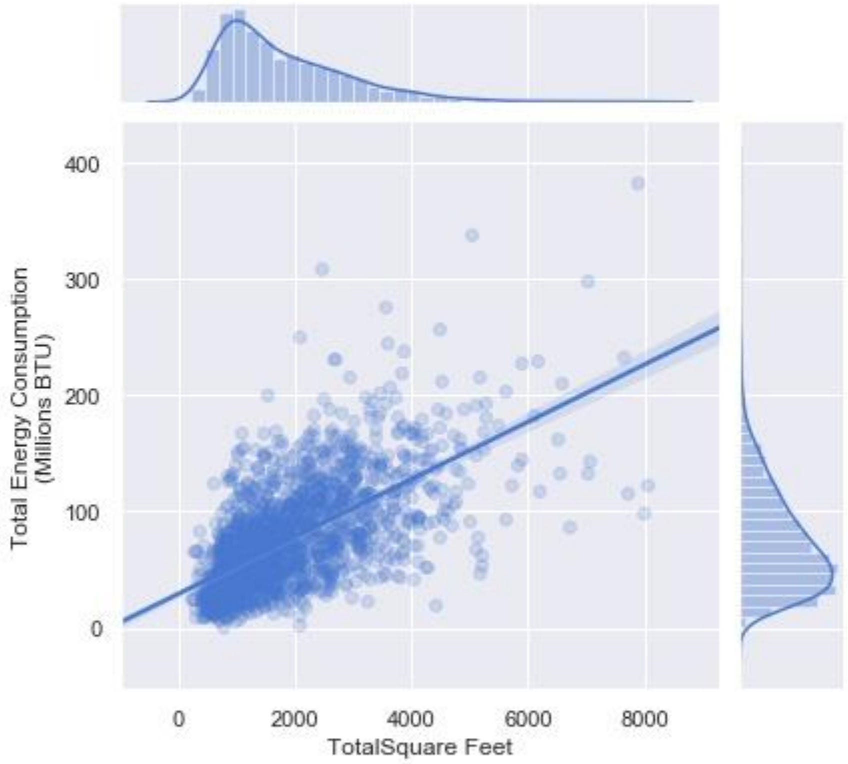
# Capstone Project 1:
# Exploratory Data Analysis

Springboard Data Science Career Track
Ishmail Grady
May 2019

To get a better understanding of the dataset before developing a predictive model, exploratory data analysis was conducted with a focus on identifying variables that have a correlation with total energy consumption. Furthermore, verifying the statistical significance of these correlations will be important in making sure that the features included in the model truly generalize the data.

One particular variable of interest is the total square feet of a household. This variable has the strongest absolute pairwise correlation of all the variables in the dataset. As shown in Figure C4, there is a moderate positive correlation between total square feet of a household and total energy consumption with a Pearson correlation coefficient of .64. A positive correlation between these variables make sense for a variety of reasons. For example, we can expect that households with more square footage to possibly use more energy for heating in cooling. Larger households might also have more inhabitants and more inhabitants could increase the use of electronics household appliances.

C4. Total Square feet v. Total Energy Consumption



Though this correlation is not surprising it is still important to verify the statistical significance of this correlation within the dataset. After verifying the significance, it will be useful to quantify a range of for the true population correlation between these variables

To verify the significance of the of the sample correlation between total square feet and total energy consumption, a z-test was performed using a Fisher transformation with the following experiment design:

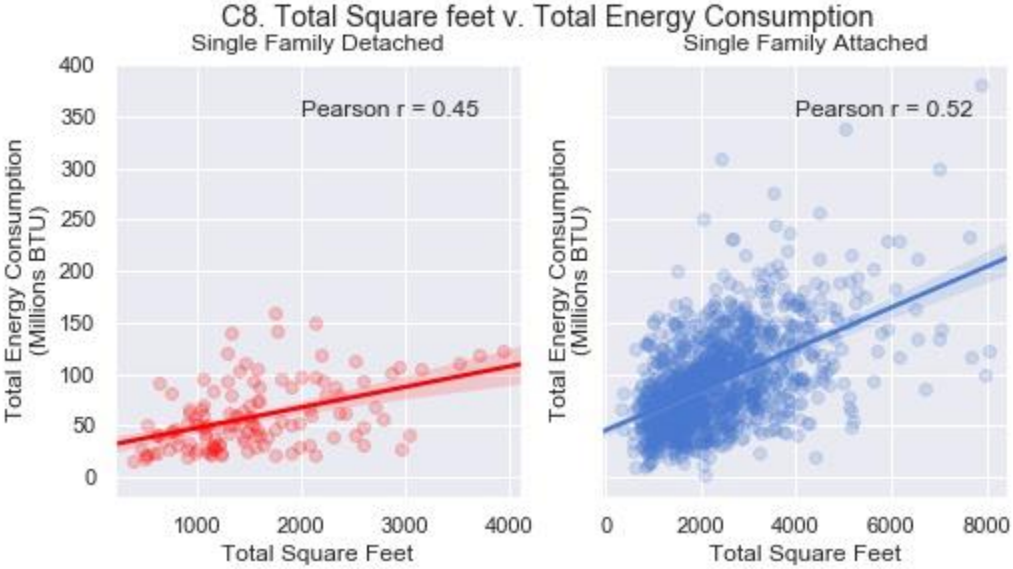
$$H_o: r = 0$$

$$H_a: r \neq 0$$

$$\alpha = 0.05$$

The result of the experiment yielded a p-value of near 0, thus we reject the null hypothesis that there is no correlation between total square feet of a household and total energy consumption in favor of the alternative hypothesis. This suggests that there is positive correlation between total square feet of a household and total energy consumption among all households. A 95% confidence interval for the true correlation is yields a range of [0.62 - 0.65].

Now that the correlation between total square feet and total energy consumption is understood among all households, it will be interesting to observe how the correlation varies across how households. Figure C8 below shows regressions plots of total square feet and total energy consumption for Single Family Detached households and Single Family Attached household to illustrate the difference the difference in correlation.



It is important to note that there are significantly more Single Family Attached households (3752) than Single Family Attached households (479) in the dataset. The correlation for both groups is weaker than the correlation for the entire sample. The correlation for Single Family Attached households is stronger than Single Family Attached homes. To test the significance of this difference in correlation among the two groups, a z-test was performed using a Fisher transformation with the following experiment design:

$$H_o: r_a = r_d$$
$$H_a: r_a > r_d$$
$$\alpha = 0.05$$

The p-value of the above z-test was 0.027 which is less than the significance level, thus we reject the null hypothesis that the correlation between total square feet of a household and total energy consumption for Single Family Detached Homes and Single Family Attached Homes is equal in favor of the alternative hypothesis. This suggests that the correlation for Single Family Detached homes is greater than that of Single Family Attached Homes.

In conclusion, the analysis detailed above will assist in selecting features that will be used in the predictive model. Having confidence in the relationship between the size of a household and the energy it will consume will drive intuition for selecting variables that are related to household size. Incorporating Additional household characteristics should help better predict energy consumption and account for the difference between household types.