Capstone Project 1

Predicting Household Energy Consumption

Project Proposal

HOUSEHOLD ENERGY CONSUMPTION PREDICTION TOOL

- The sociopolitical effects of climate change are shifting the focus of energy sector. Utility companies are facing pressure to be more efficient and environmentally sustainable
- How utility companies communicate with their customers about their energy consumption will play a crucial role in how their businesses are able to adapt to this changing market
- This project will serve as a solution to help energy providers enhance their customer communication by providing a predictive model that will allow its customers to predict their energy consumption based on the characteristics of their household

Utility Company and Customer Benefits

HOUSEHOLD ENERGY CONSUMPTION PREDICTION TOOL

Utility Company Benefits

- It will allow providers to compete with other energy management solutions that can threaten the relevancy of their customer communications.
- Provides a means of collecting data on customers to build better datasets that for applications such as peak load management and targeted marketing for goods and services.

Customer Benefits

- Analyze how various household decisions can affect their energy use, thus providing a means to minimize their consumption and save money
- Compare their actual consumption to predicted consumption in order to gauge how energy efficient their household is.

Data Wrangling

Data Source

HOUSEHOLD ENERGY CONSUMPTION PREDICTION TOOL

- 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
- This survey is a national sample of housing units that are considered primary residences, as defined by the U.S. Census Bureau
- The survey results contain data on 5,686 randomly selected households across the nation. This sample was statistically designed to represent 118.2 million households throughout the country
- The dataset contains two main types of information: household characteristics and consumption & expenditures
  - Household characteristics data covers many areas such as appliances, electronics, space heating, household demographics, and more
  - Consumption & expenditures data contains information on the fuel type(s) used, the end uses of the fuel associated with the various household characteristics, and the dollar values of the energy used

Data Cleaning

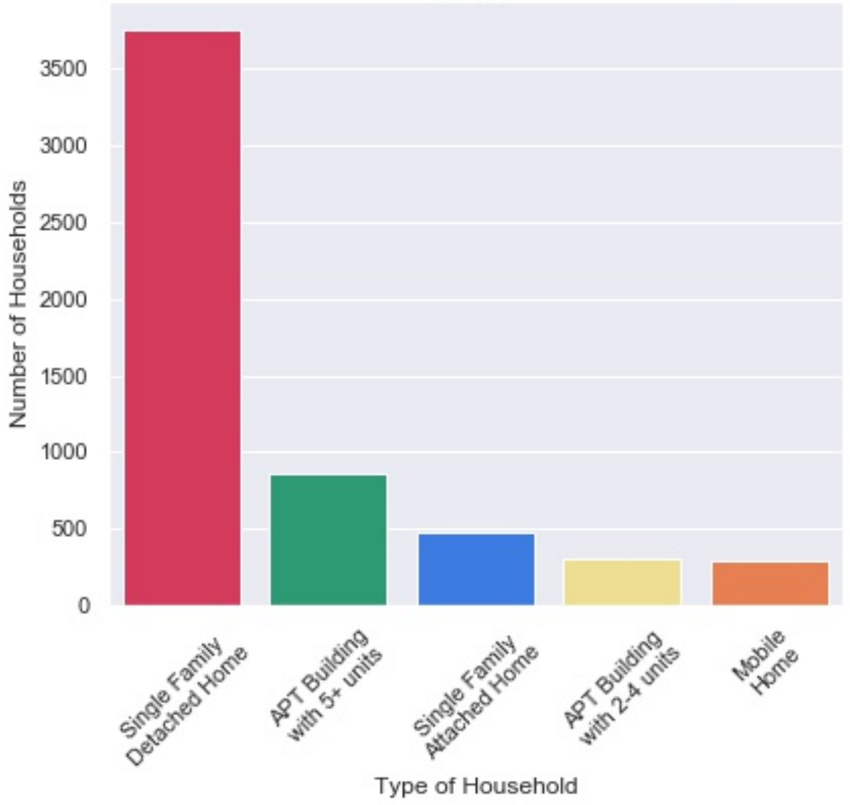
HOUSEHOLD ENERGY CONSUMPTION PREDICTION TOOL

- 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
- 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
- The dataset contains 217 columns that represent imputation flags for 222 columns variables (some imputation flag columns correspond to multiple variables). These values were considered missing entries
- There were 20 questions in the survey that have “Don’t Know” as a possible answer. These values were also considered to be missing
- 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

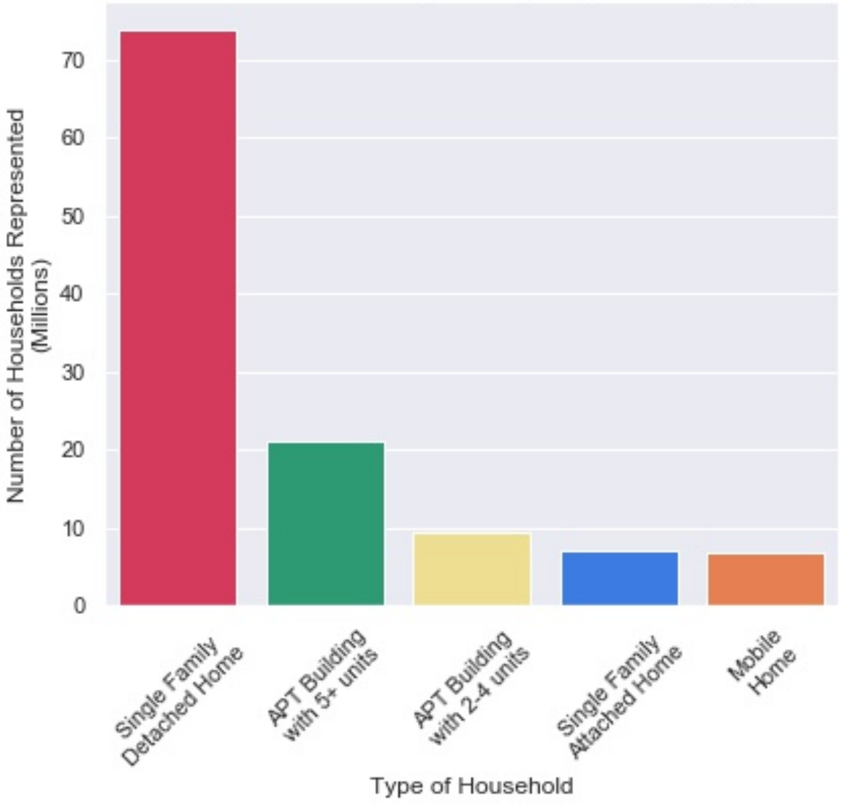
Exploratory Data Analysis

Exploratory Data Analysis (1 of 4)

- 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.
- The 3,752 observations of Single Family Detached households in the dataset are statistically representative of 73.8 million American households.

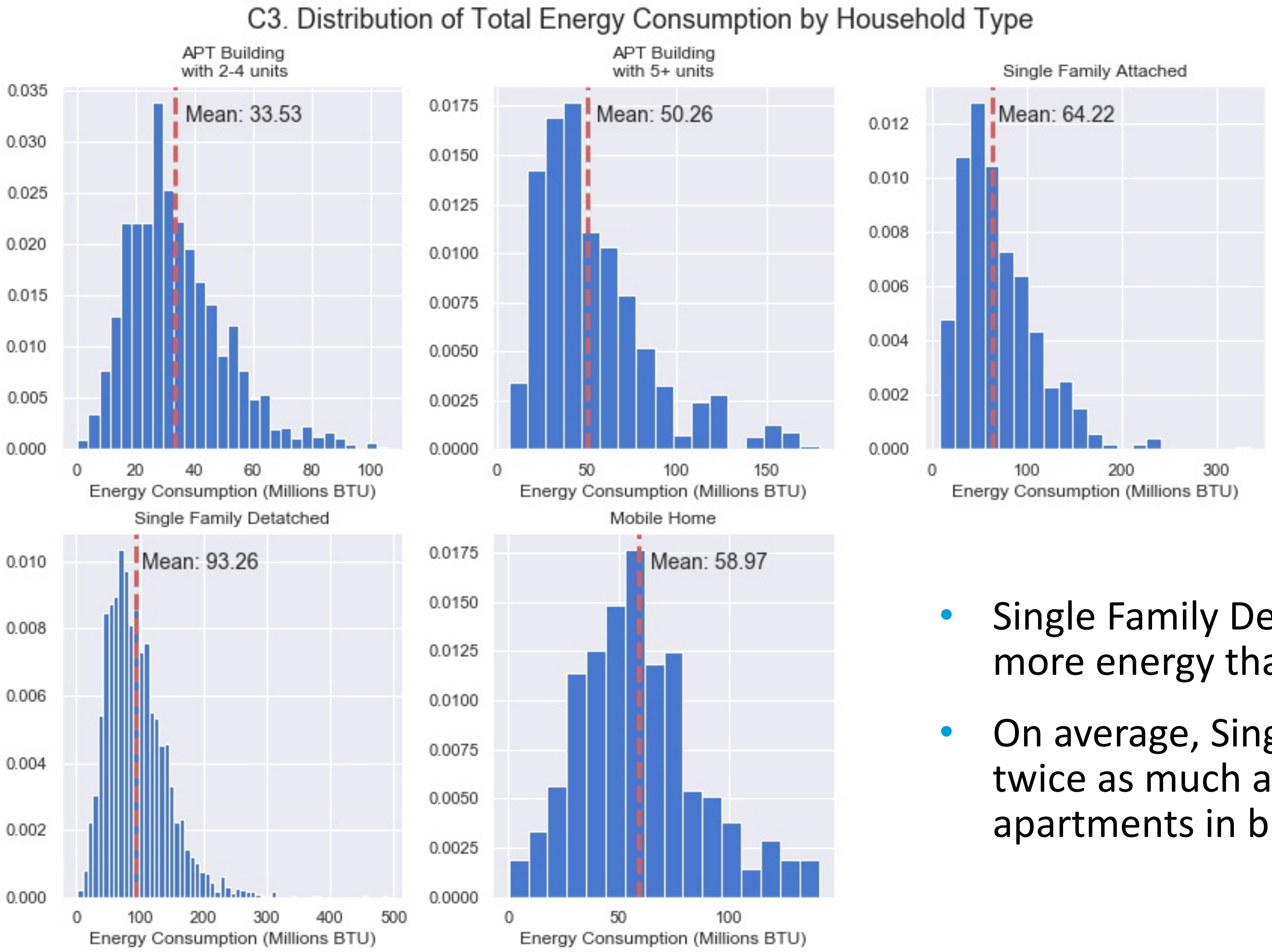


C1. Households by Type in RECS Survey



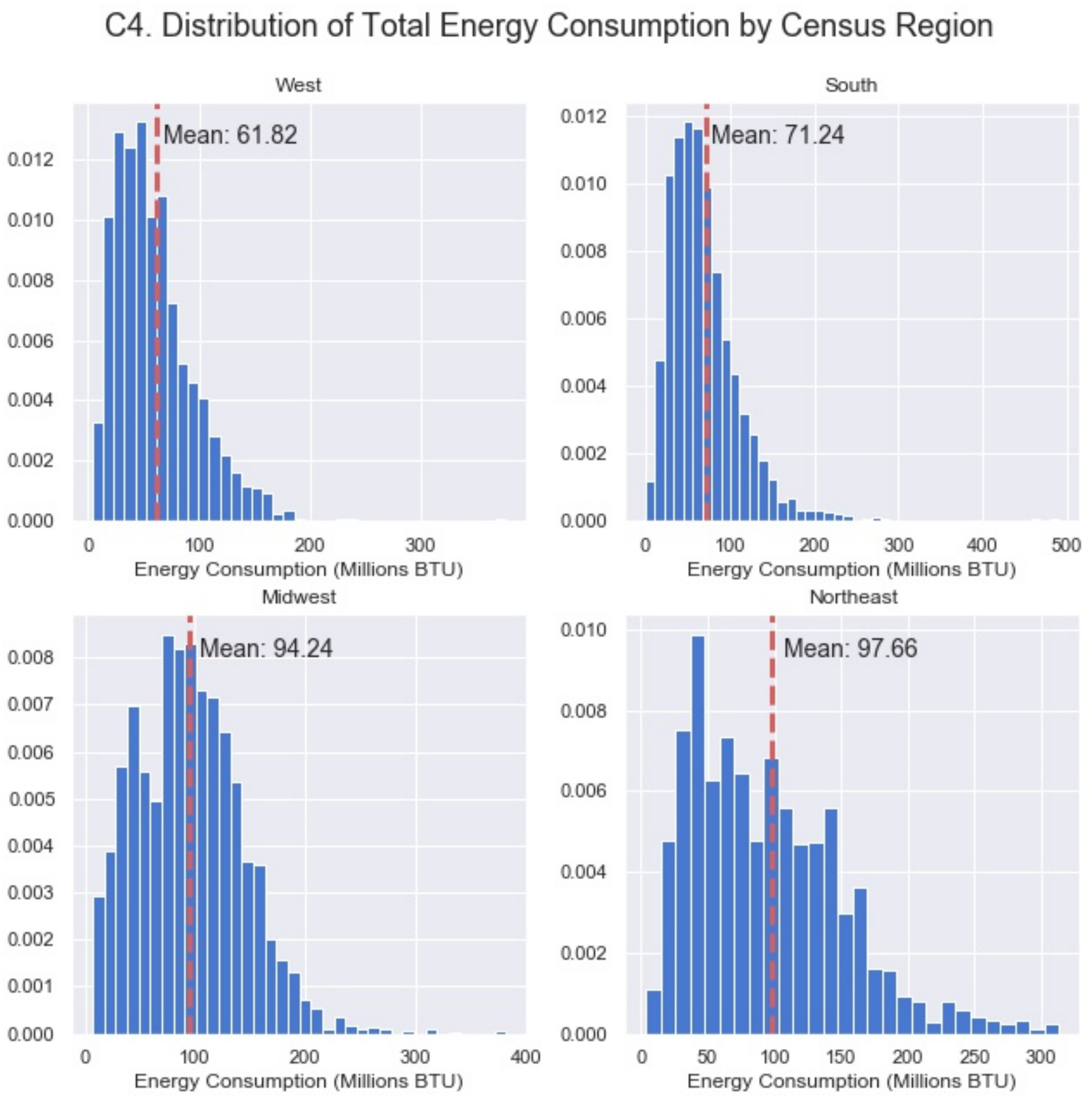
C2. Cumulative Sample Weight by Housing Type

Exploratory Data Analysis (2 of 4)



- Single Family Detached households consume significantly more energy than other household types.
- 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.

Exploratory Data Analysis (3 of 4)

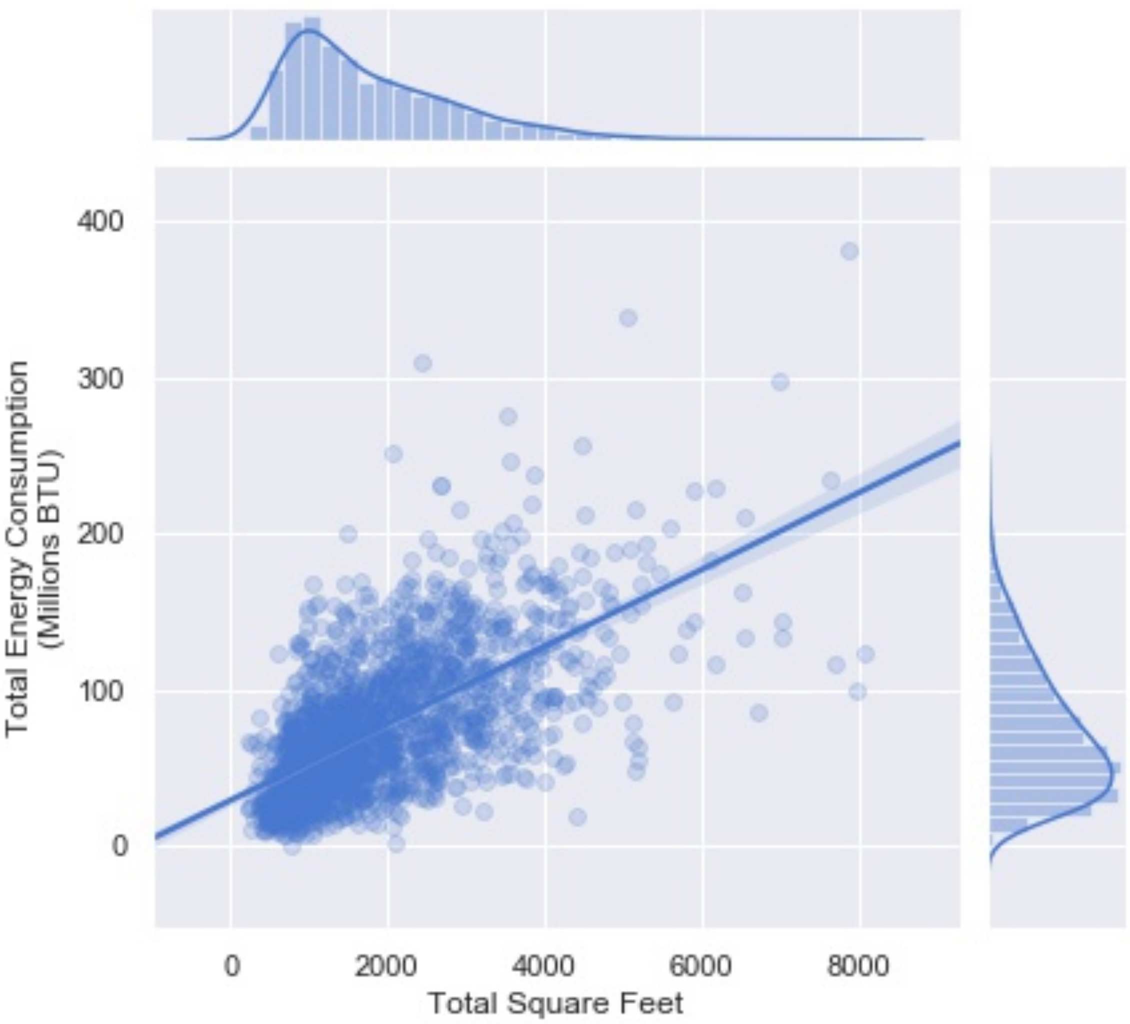


- All four regions have positively skewed distribution. The distributions of energy consumption in the Northeast and Midwest appear to be similar however the distributions of the South and West have significantly lower means
- This may suggest that total energy consumption is closely related to the region that a household is in and that households in colder climates consume more energy than those in warmer climates

Statistical Inference

Statistical Inference (1 of 4)

C5. Total Square Feet v. Total Energy Consumption



- Total square feet of a household and total energy consumption two variables have a pairwise Pearson correlation coefficient of .64.
- The correlation between total square feet and total heated square feet is the highest correlation to total energy consumption of any numeric household characteristic
- This correlation between these variables should be evaluated for significance and and dependence on other household characteristics

Statistical Inference (2 of 4)

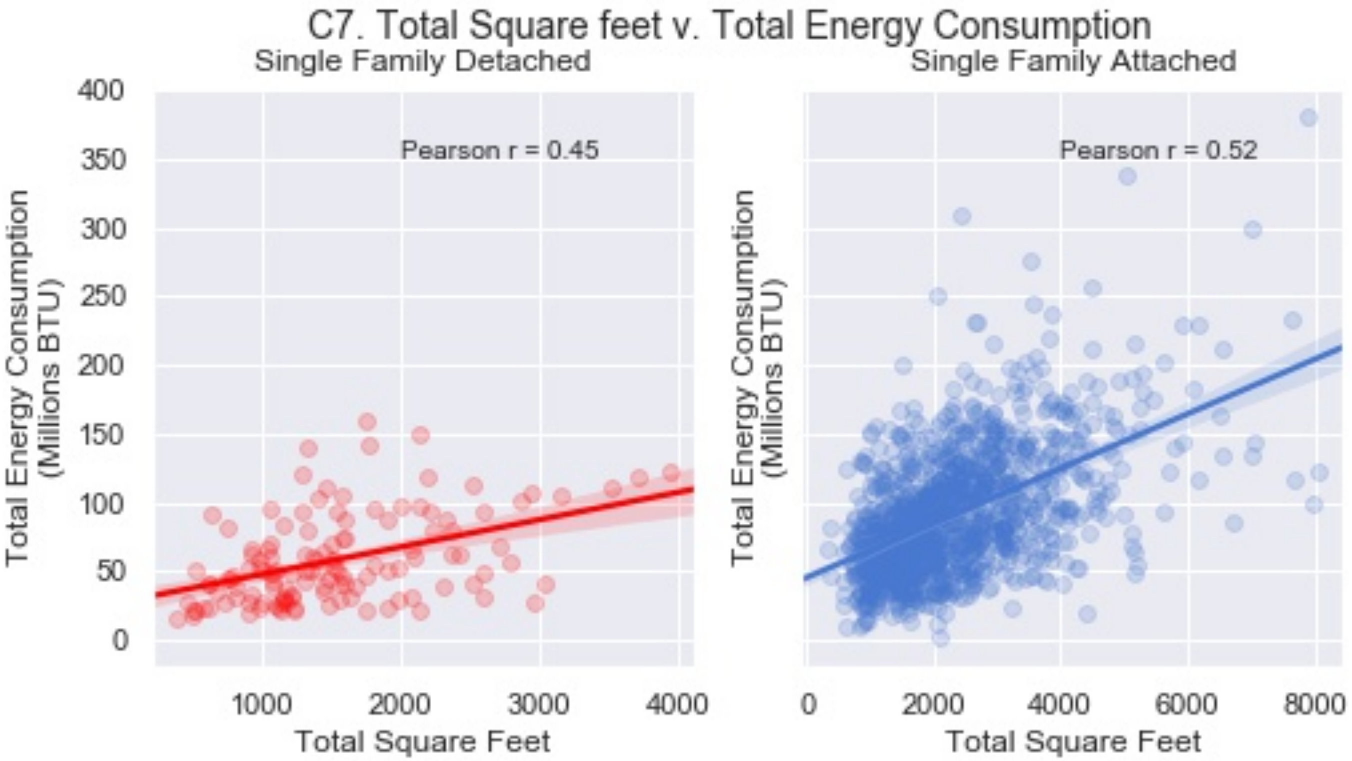
To verify the significance of the of the sample correlation between total square feet and total energy consumption, a z-test was performed using a Fisher transformation with the following experiment design:

- The null hypothesis is that the true correlation between total square feet and total energy consumption is zero.
- The alternative hypothesis is that the the the true correlation between total square feet and total energy consumption does not equal zero
- The significance level is set at 0.05

- The result of the experiment yielded a p-value of near 0, thus we reject the null 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]

Statistical Inference (3 of 4)

- The below shows regressions plots for Single Family Detached households and Single Family Attached households to illustrate the difference the difference in correlation between of total square feet and total energy consumption across household types
- 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.


C7. Total Square feet v. Total Energy Consumption

Statistical Inference (4 of 4)

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:

- The null hypothesis is that the true correlation between total square feet and total energy among Single Family Detached households and Single Family Attached household consumption is equal.
- The alternative hypothesis is that the the the true correlation between total square feet and total among Single Family Detached households is greater then that of Single Family Attached households
- The significance level is set at 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 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.