STATS 506 Project

AUTHOR
Yang Han

Link to Github repository:

https://github.com/Marslalala/STATS506Final-Proojecct.git

Introduction

Research Question

Under the background of energy saving, many enterprises start to think about how to minimize energy consumption with minimal losses. Replacement of old-fashioned energy resources with renewable energy, such as solar power, has always been a potential solution. This report will display an investigation of the possibility of reducing the energy consumption of other types by the usage of solar energy for a commercial building in the U.S.

Data

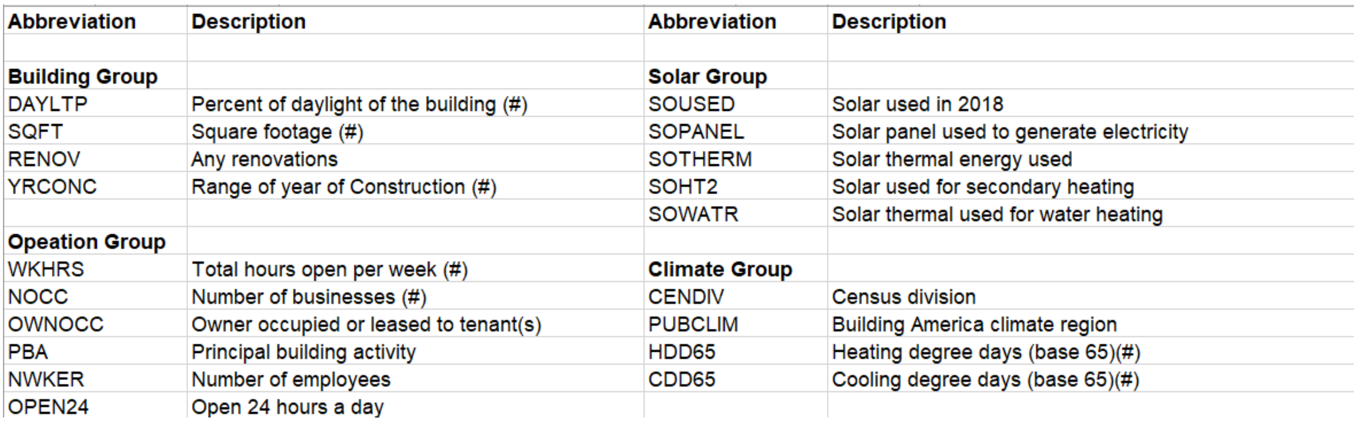
The sample data set comes from the US Energy Information Administration and it contains information on a sample of all commercial buildings in the US and their energy characteristics, consumption, and expenditures. This is a complex survey design, and the following analysis has accounted for this.

Statistical Method

Variable Selection & Transformation

Starting from the response, since we were considering the total consumption of other energy sources, we added up all the energy usage except solar energy for each observation and standardized it concerning the total square footage of all the space, yielding a new dependent variable TEUI (Total Annual Energy Usage Intensity) with unit **kBtu/ft^2*year**.

20 independent variables were selected as predictors and categorized into 5 different groups that we believed were most relevant to the energy consumption of a building. The table below gives details.



Abbreviation	Description	Abbreviation	Description
Building Group		**Solar Group**	
DAYLTP	Percent of daylight of the building (#)	SOUSED	Solar used in 2018
SQFT	Square footage (#)	SOPANEL	Solar panel used to generate electricity
RENOV	Any renovations	SOTHERM	Solar thermal energy used
YRCONC	Range of year of Construction (#)	SOHT2	Solar used for secondary heating
		SOWATR	Solar thermal used for water heating
Opeation Group			
WKHRS	Total hours open per week (#)	**Climate Group**	
NOCC	Number of businesses (#)	CENDIV	Census division
OWNOCC	Owner occupied or leased to tenant(s)	PUBCLIM	Building America climate region
PBA	Principal building activity	HDD65	Heating degree days (base 65)(#)
NWKER	Number of employees	CDD65	Cooling degree days (base 65)(#)
OPEN24	Open 24 hours a day		

# distinguishes numerical variables, and there were 8 numerical predictors and the rest were categorical.

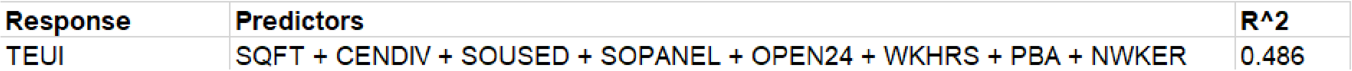
Approach

After the data cleaning process, I generated some descriptive statistics and fitted a one-factor linear regression model (Analysis of Variance) to investigate the relationship between the total energy usage intensity (TEUI) and the use of solar energy (SOUSED). I found that more than 75% of the observations have ETUI less than or equal to 98.97 kBtu/ft^2 per year, while the maximum usage intensity is 1710.84 kBtu/ft^2, which is quite large. This implied that it might be the case for both buildings using solar or not, there were other factors influencing the amount of TEUI of a building. From the analysis of variance, I found that the p-value was large, indicating that the two groups from SOUSED might not be giving any differences to TEUI (SOUSED is a categorical variable with 1 indicating solar energy was used in the building in 2018 and 2 indicating was not used).

Next, I generated ordinary least squares regression models and selected the best model using backward elimination. Before selecting the best model step-by-step, I believe it is better to check for collinearities among numerical variables to ensure the validity of the t-tests and numerical stability first and I did not find any significant correlations among variables, hence, I believe the p-values shown by model summary reflected the truth. By eliminating predictors one by one based on the insignificant p-values, eventually I got a model with 8 predictors listed in the table below. The diagnosis was carried out and I found that there were 3 outliers which were removed from the best model at last. The distribution of the errors in the model was heavy-tailed. Bootstrapping could be used to generate robust estimates of standard errors and confidence intervals. I could do this if any inference is needed. Since here we just need to interpret the model to draw a conclusion, this process can be saved.

Results

The best model is shown below.



Response	Predictors	R^2
TEUI	SQFT + CENDIV + SOUSED + SOPANEL + OPEN24 + WKHRS + PBA + NWKER	0.486

We can see that apart from the solar set predictors, the square footage of the building, the census division of the building, the operating mode, the number of workers in the building, and the main activities in the building are also crucial. The R^2 is 0.486 for this mode which indicates that nearly half of the variability observed is explained by this regression model, which is enough to be a good model for data from a complex survey.

Conclusion

We have done an investigation of how the use of solar energy reduces the energy consumption of other sources of a commercial building with some other controlled variables. In a simple linear regression model, there are no large differences in total energy usage intensity between the buildings that use or do not use solar energy. However, when the complex survey design is considered and with other controlled variables joining in, the ordinary least squares regression model gave a very significant p-value for those who use and do not use solar energy (SOUSED), implying a significant difference between these two groups. The second group of SOUSED represents those who do not use solar energy, and the positive estimate indicates this level would raise the TEUI. Hence we can conclude that the usage of solar energy can help a commercial building to reduce the usage of energy of other forms. As this conclusion is reflected by TEUI, it does not mean that a building would consume less energy compared with other buildings if it is using solar energy. The point is the change in consumption of other energies before and after the use of solar energy. Interestingly, those buildings that use solar panels for solar thermal or electricity would increase TEUI. There is one possibility: those buildings need solar panels for thermal energy or electricity requires an enormous energy amount, which is costly, so they choose to use solar panels to save energy costs, but their TEUI is high.

Some other predictors also influence TEUI significantly. One more worker in the building would result in an increase of TEUI by **3.491 * 10^-2** and one more hour people work in the building per week would result in an increase of TEUI by 0.591. TEUI is also heavily influenced by the principal activities in the building. We can see that PBA6, PBA15, and PBA16 (representing Food sales, Food service, and Impatient health care) have extremely significant p-values and large estimates need to be added to the intercept. This indicates that these three activities require more energy consumption than any other activities (more than 10 units of TEUI). This might be because refrigeration is needed for the storage of food and medicines.