A Study on the Effect of Home Performance Improvement Program on Energy and Cost Savings in the State of New York

Capstone Project – Final Report

Analysis By

Tanuj Chauhan
Aravind Mohan
Saili Kotturu
Ganugula Satya Ravi Teja

Reported By

Tanuj Chauhan

PGP-DSE Bangalore 2020-21

Project Mentor & Research Supervisor: Mr. Ramkumar Manoharan

Great Lakes Institute of Management
https://www.greatlakes.edu.in/

Table of Contents

Acknowledgements

We wish to place on record our deep appreciation for the help provided to us by our Mentor Mr. Ramkumar Manoharan, who helped us narrow down on the choice of the Project as well as the scope and focus area of the project.

We would also like to place on record our appreciation for the guidance provided by Mr. Regi Mathew and Mr. Lakshmi Prasad for giving us valuable feedback and being a source of inspiration in helping us to work on this project.

We also cannot express enough thanks to the faculty of Data Science and Engineering in Great Learning Academy who helped up prepare, learn and understand on how to apply great analytical thinking.

Also, I would like to express my special thanks to Great Lakes Institute of Management and Great Learning Academy for making this project possible by giving us the space, time and resources required during the course of this project.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

Date: April 17, 2021
Place: Bangalore

Tanuj Chauhan
Saili Kotturu
Aravind Mohan
Ganugula Satya Ravi Teja

Summary

Home Performance with Energy Star or known as HPwES program was introduced by the New York State Energy Research and Development Authority and Energy Star Foundation by U.S. Environmental Protection Agency in the state of New York in 2007, which still has an on-going impact evaluation till this date.

Households who participated in the program were provided ways to upgrade energy and heating applications by which the State of New York hoped to achieving in power conservation and sustainable ways of power consumption.

The objective is to identify the most optimal performance solutions regarding energy efficiency and cost savings. Evaluating if the measures taken by the program have been successful in reducing the overall economic costs and reducing energy needs depending upon the performance after the end of the first year and after measure implementations in place.

Study needed to produce analytical understanding on the success of the program on the larger scale and to see if such programs subjected to increasing efficiency can be useful for the future regarding home performance in the state of New York or in other states.

Through exploratory data analysis we understood the imbalances in a dataset and sorted the dataset to analyse for Single Family Home Performance done on Building Shell over homes having Natural Gas Pre-Retrofitted Home Heating Fuel type with only 1 Unit served by the program in the year of 2018, 2019 and 2020.

Here, we used Regression Analysis to understand the target which was annual estimate of dollars saved. We chose to model with outliers and without outliers. From the regression analysis Decision Tree with frequency encoding was chosen to the be most optimal model for us.

We found out that Region, Billing Month, MMBtu savings, Total Project Cost, Electric Utility, kWh saved each year and year-wise comparison turned out to be the most important features which will affect the performance of the program.

1. Introduction:

1.1 Industry Review

The New York Home Performance Program, administered by NYSERDA, is a comprehensive retrofit program encourages home and building owners and tenants of existing one - to four - family homes to implement comprehensive energy efficiency-related improvements and technologies by contractors accredited by the Building Performance Institute.

The program is designed to offer enhanced assistance to low - to moderate income households. The "Assisted" component of the program is available to residents with up to 80% of area median income, or 80% of state median income, whichever is higher for the county. Need a study to produce analytical understanding on the success of the program on the larger scale and to see if such programs subjected to increasing efficiency can be useful for the future regarding home performance in the state of New York.

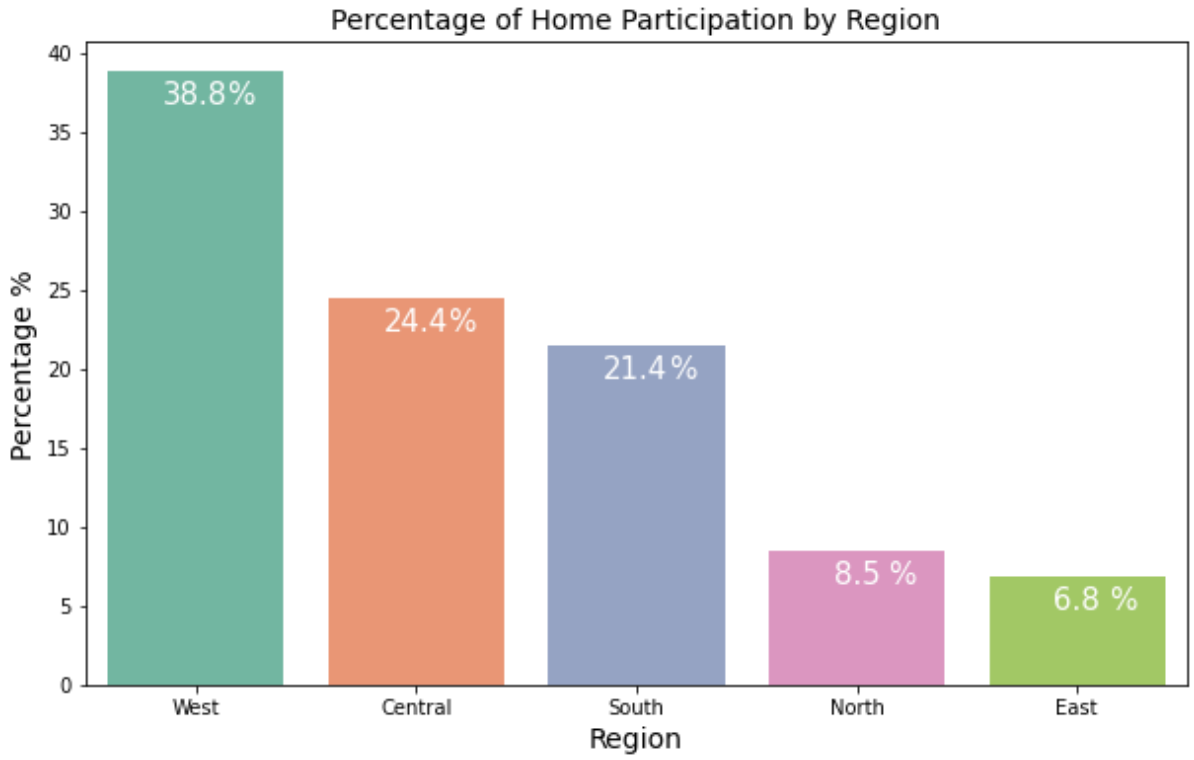


fig. 1.1: percentage of home participation in the program by region

1.2 Past Practices

The purpose of the impact evaluation done was to establish first year evaluated gross and evaluated net energy savings for PY2007 and 2008 participants. The evaluation was designed to estimate the realization rate (RR), i.e., the ratio of the evaluated gross savings to the NYSERDA-reported savings and develop estimates of the net-to-gross (NTG) components free ridership (FR) and spillover (SO). The NTG components were combined with the RR to calculate net savings. Electric demand savings, kW, was estimated based upon the evaluated gross and net electric energy savings.

Savings by major measure group were estimated, providing some insight into whether specific measure groups are more or less likely to achieve the expected savings. In addition, participants were surveyed concerning the reasons they replaced equipment or installed measures and information concerning non- energy impacts.

The evaluation included a billing analysis of all PY2007 and 2008 participants with sufficient billing history, a restricted billing analysis of a sample of approximately 600 households, and causality study (i.e., attribution to the Program to obtain a net savings estimate).

1.3 Current Practices

HPwES (Home Performance with Energy Star) program is different from utility energy efficiency programs in that NYSERDA, rather than the utility, is delivering services to the participants. This led to an unanticipated complexity by adding a layer to the process of obtaining the billing records, resulting in additional attrition due to the fact that some HPwES participants could not be identified in the utility billing systems. In addition, National Fuel, Central Hudson Gas and Electric and Saint Lawrence Gas were not able to provide billing data at all. For the electric model, the impact was minor, with about 1% of HPwES projects removed from the model for this reason. However, about 33% of potential gas model participants were customers of the three utilities who did not provide any billing data (primarily National Fuel).

In addition, once the modeling was underway, it became clear that the billing data from two utilities, NYSEG and RG&E), contained many unidentified estimated reads and reconciliations, thus breaking the direct relationship between consumption and

the weather impacts during the specific billing periods. The final models were run both with and without data representing these two utilities. The statistical reliability of the analysis dropped dramatically and the estimated savings from the model were substantially lower when all utilities were included in the model. Since the inclusion of NYSEG and RG&E had such a deleterious effect on the reliability of the regression results, the final evaluated savings are based on the model without data from these two utilities.

Overall, the attrition (the removal of participants from the regression model due to insufficient billing data) was substantial and all but three utilities were removed from the billing analyses. For the most part, the reasons for attrition, such as failure to locate specific HPwES participants in the utility billing systems, are likely to be random and would not be expected to introduce bias into the results. However, to the extent that entire utilities were removed, there could be unintended consequences in that specific large contractors may have been also eliminated from the analysis. The largest potential source of bias was the removal of NYSEG, RG&E and National Fuel as utilities with many Program participants which led to loss of three of the larger contractors. The other utilities not in the final regression model account for only a very small fraction of program activity.

2. Dataset and Domain

2.1 Dataset

2.1.1 Data Dictionary

Column Name	Description	Type
Reporting Period	The time period covered by the dataset	object
Project ID	Unique identifier for project	object

Project County	Name of county for project location	object
Project City	Name of city for project location	object
Project ZIP	ZIP code for project location	object
Gas Utility	Name of gas utility for project location. If blank, then utility was not reported, or project location is not served by a gas utility	int64
Electric Utility	Name of electric utility for project location	object
Project Completion Date	Date final project completion paperwork was reviewed and approved by Program	object
Total Project Cost	Cost of project (USD). NYSERDA incentive currently at 100% of the total project cost. Total Project Costs less than $100 often reflects mileage-only billing for projects with minor work scope	float64
Pre-Retrofit Home Heating	Indicates the pre-retrofit primary heating fuel type. Either coal, electricity, kerosene, natural gas, oil, other, pellets, propane, or wood	object
Year Home Built	Home construction date. Blank cells indicate data not reported by the contractor	object
Size Of Home	Square footage of home. Blank cells indicate data not reported by the contractor	object

Number of Units	Number of units served by the Program. Data may include exceptions to the One-to-Four units, which were approved by NYSERDA on a case-by-case basis	float64
Job Type	Indicates whether the project includes only electric reduction measures (Electric Reduction) or is a comprehensive (Home Performance) project including both electric and heating efficiency improvements	object
Type Of Dwelling	General home category describing the dwelling as Single Family, 2-4 Family, Multi Family, or Manufactured/Mobile Home	object
Measure Type	Measure classification describing primary project improvement defined as Combination-Home Performance, Combination-Electric Reduction, Heating Repair/Replacement, Refrigerator/Freezer Replacement, CFL/LED Lighting, Shell, Shower Head Replacement, or Other	object
Estimated Annual kWh Savings	Annual post-retrofit modeled electric savings estimate in kWh. Negative numbers represent projects with post-retrofit increase in electric consumption, typically from fuel conversions or ancillary savings. Projects with zero kWh represent projects with only health and safety measures, and customer efficiency education	float64

Estimated Annual MMBtu Savings	Annual post-retrofit modeled MMBtu savings based on primary fuel type. Negative numbers represent projects with post-retrofit increase in MMBtu consumption, typically from fuel conversions or ancillary savings. Projects with zero MMBtu represent projects with only health and safety measures, and customer efficiency education	float64
First Year Modeled Project Energy Savings $ Estimate	Estimated post-retrofit first year dollar savings (USD). Negative numbers represent projects with estimated post-retrofit first year dollar expenses, typically occurring when non-energy work was completed such as health and safety improvements, or when work was done in conjunction with another, net positive energy savings project. Projects with zero energy savings dollars represent projects with only health and safety measures, and customer efficiency education	float64
Location 1	Open Data/Socrata-generated geocoding information	object

2.1.2 Dataset Summary

Organization	New York State Energy Research and Development Authority
Time Period	Beginning January 1, 2018

Posting Frequency	Quarterly
Dataset Owner	New York State Energy Research and Development Authority
Contact Information	openny@nyserda.ny.gov
Coverage	Statewide
Granularity	ZIP Code

2.1.3 Variable Categorization

Type	Variables	Count
Numerical	'Project ZIP', 'Total Project Cost', 'Number Of Units', 'Estimated Annual kWh Savings', 'Estimated Annual MMBtu Savings', 'First Year Modeled Project Energy Savings $ Estimate'	6
Categorical	'Reporting Period', 'Project ID', 'Project County', 'Project City', 'Gas Utility', 'Electric Utility', 'Project Completion Date', 'Pre-Retrofit Home Heating Fuel Type', 'Year Home Built', 'Size Of Home', 'Job Type', 'Type Of	14

	Dwelling', 'Measure Type', 'Location 1'	

2.2 Pre-Processing Analysis

2.2.1 Missing Values

Total number of Missing Values: 16915

Variable	Percentage of Missing Values
Year Home Built	39.668%
Gas Utility	26.720%
Number Of Units	00.004%

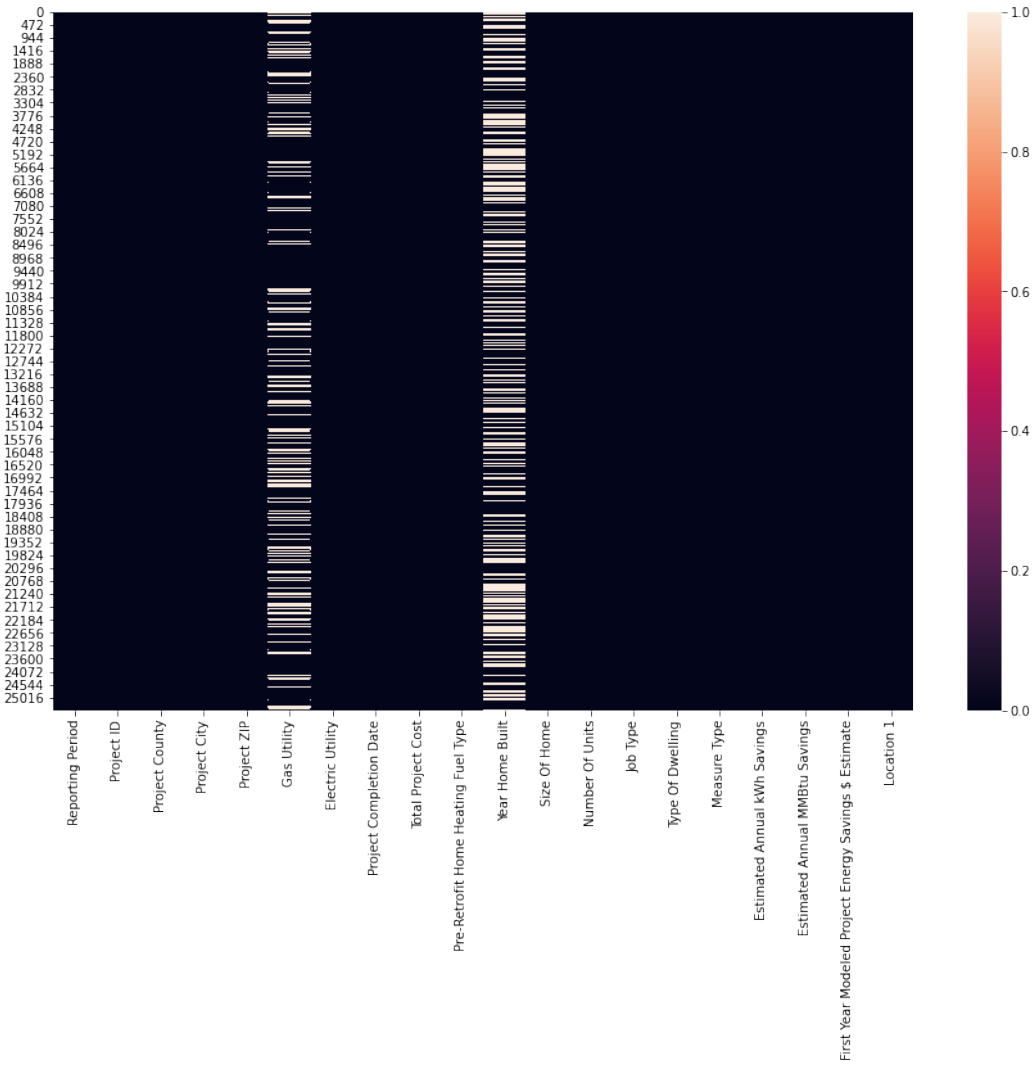


fig. 2.1: heatmap visualization of missing values in the dataset

2.2.2 Feature dType Correction

Checking for DataFrame information showed:

- 'Size of Home' is wrongly specified to be 'object' type.
- To be changed to ‘numeric’ type.

2.2.3 Creating New Features

New features created from existing features:

- Project Completion Year from Project Completion Date
- Project Month from Project Completion Date
- Latitude from Location 1
- Longitude from Location 1
- Type of Home Size by using quantiles on Size of Home
- Region by grouping Project County

2.2.4 Removing Unnecessary Features

Following features were removed for being redundant:

- Reporting Period
- Project ID
- Project ZIP

2.3 Project Justification

2.3.1 Project Statement

The project tries to identify the different means by which home owners and contractors in the state of New York are incentivized to use and provide energy efficient methods through the program administered by the NYSERDA. The purpose of the project is to create a clear picture of how to develop cost effective solutions and methods to incentivize contractors and home owners to build energy efficient systems for heating in the State of New York.

2.3.2 Complexity

Approach is to identify the most optimal performance solutions regarding energy efficiency and cost savings by creating regression models over the savings and using machine learning tree regression techniques like Decision Tree Regression and/or Random Forest Regression. Identifying which of the Project Counties were affected the most by the program in terms of energy and cost savings.

2.3.3 Project Outcome

The objective is to identify if the measures taken by the program have been successful in reducing the overall economic costs and reducing energy needs depending upon the performance after the end of the first year after measure implementation. Energy Studies identify and analyze opportunities to make buildings more efficient, which lowers associated energy costs.

2.3.4 Methodology

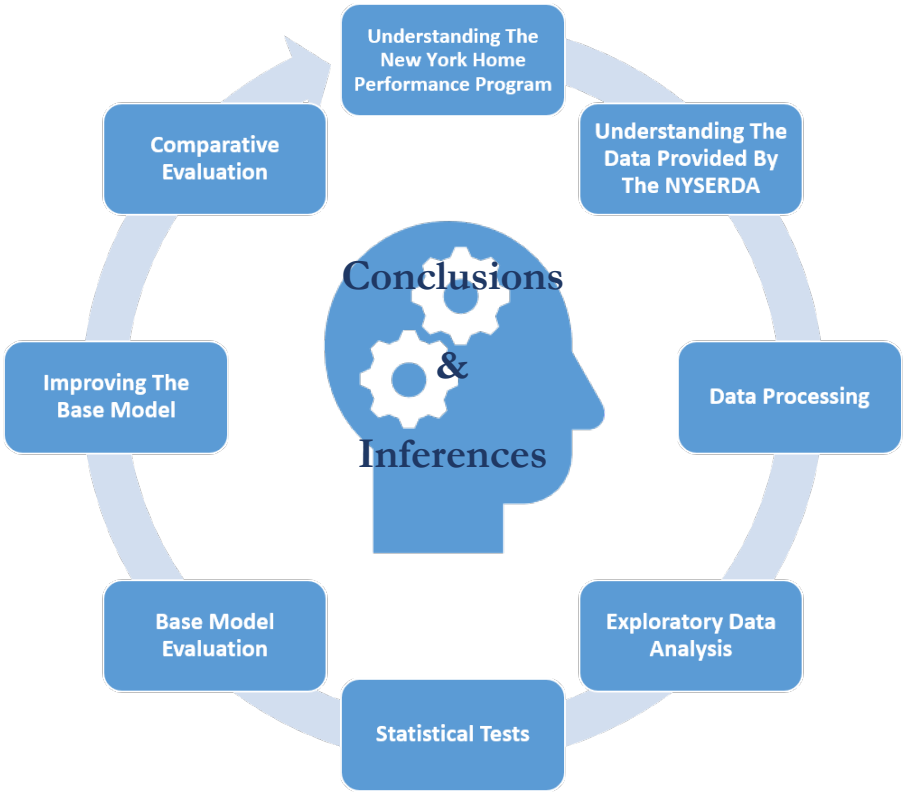


fig. 2.2: methodology map for the project

Before starting any analysis over a project, we first have to understand the project's purpose. If we fail to do so, then our inferences may become insignificant because our evaluation can be completely inconsequential towards bringing any viable improvement to the project.

Now, when we have a certain knowledge about what the project or the program is actually doing, we can move ahead in understanding the dataset. We have to understand about what the data represents. We have understood the basic features of our dataset in the previous slide.

After Understanding the data, we move ahead with processing it. At this stage it's all about removing inconsistencies like understanding and treating the null values, wrongly assigned data types to variables, outlier detection and removing unnecessary features if needed. We can also create new features by engineering the dataset although a good amount of domain knowledge is needed to achieve the best feature engineering to the respective dataset.

Next would be about finding different relationships between the features. I have named this part as exploratory data analysis to reduce confusion in the work flow even though Data Processing is also a part of EDA. Here we create different plots and evaluate for the best inferences. We can evaluate on the basis of target variables or by using the correlation between the features.

To understand the significance of the data, we can do so by understanding it's distribution. We can use statistical methods to achieve this. Like shapiro for testing normality or levene for testing the equality of variance within samples.

Then we move towards building a base model geared towards regression methods. After evaluating initial metrics, we will try to improve it either by tuning our base model or by using other regression algorithms.

After our models have been built, we will evaluate the performance by comparing their metrics. If the model or evaluation does not seem to be effective towards the program, we go back again to understanding the program.
If everything goes fine and we are happy with the performance of the final model, we will conclude the project, create inferences and provide our client: NYSERDA with probable improvement solutions.

3. Exploratory Data Analysis

3.1 Sorting Dataset

3.1.1 Correlation Analysis

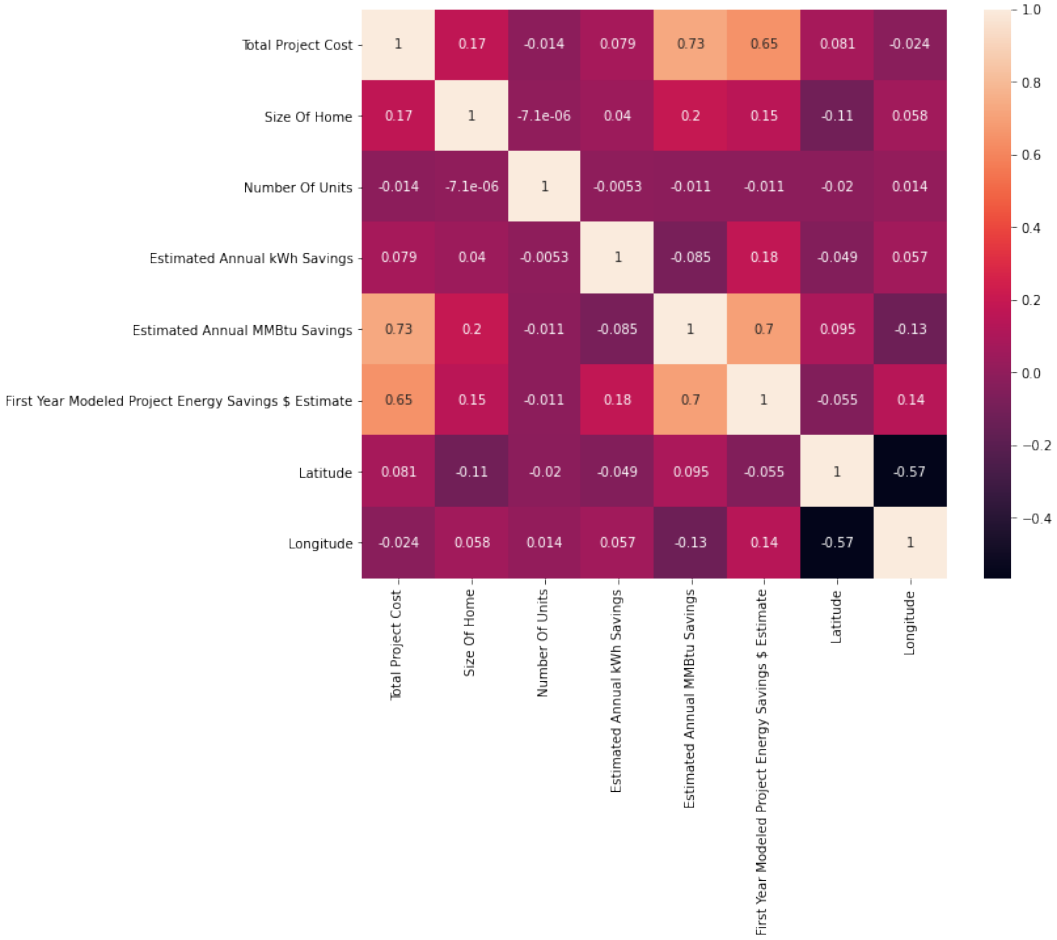


fig. 3.1: heatmap visualization of correlation between the numerical features

- 'Estimated Annual MMBtu Savings' is positively correlated with the 'Total Project Cost' at 0.73.

- 'Estimated Annual MMBtu Savings' is positively correlated with 'First Year Modeled Project Energy Savings $ Estimate' at 0.70.

- 'First Year Modeled Project Energy Savings $ Estimate' is positively correlated with 'Total Project Cost' at 0.65.

- Every other feature has a very weak correlation with each other.

3.1.2 Analyzing Imbalances

3.1.2.1. Project Completion Year

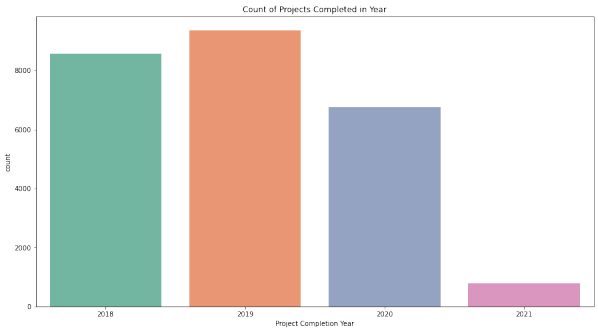


fig. 3.2: bar graph showing the count of projects completed in which year

- 2018 participants have the most projects completed., followed by 2019, 2020 and then 2021 in which most of the projects are still on going.

- 2021 having the least participating homes due to being the current ongoing year for evaluation.

- Due to billing months being incomplete, the year 2021 will hinder the analysis if taken in training. Hence, we remove participants from the year 2021.

3.1.2.2. Job Type

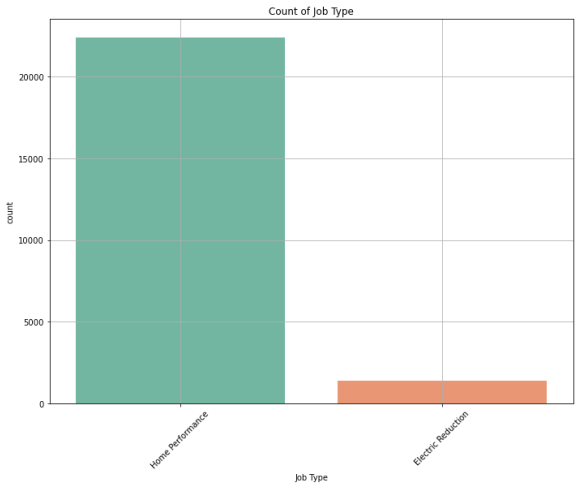


fig. 3.3 bar graph showing the count of different types of Project Job

Job Type tells us whether the project includes only electric reduction measures (Electric Reduction) or is a comprehensive (Home Performance) project including both electric and heating efficiency improvements.

- In the year of 2018, 2019 and 2020, Electric Reduction Jobs haven't been applied much by the NYSERDA.
- Homer Performance Job Type seems to be the most applied comprehension to the project.
- Not much can be inferred from the Electric Reduction job done in the project due to the participants being very low in the year of 2018, 2019 and 2020.

3.1.2.3. Measure Type

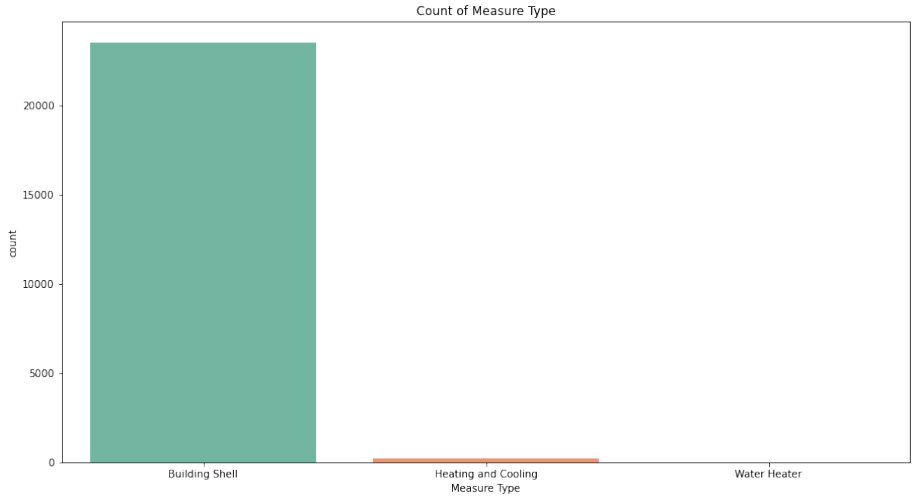


fig. 3.4: bar graph showing the count of different types of Measuring used.

Measure Type is a classification describing primary project improvement defined as Combination-Home Performance, Combination-Electric Reduction, Heating Repair/Replacement, Refrigerator/Freezer Replacement, CFL/LED Lighting, Shell, Shower Head Replacement, or Other.

- In the year of 2018, 2019 and 2020, Building Shell type of Measure has been applied throughout the program by the NYSERDA.
- Not much inferences can be gathered from Heating and Cooling Measure Type and Water Heater Measure type.

- We will remove the participants who have been provided Heating and Colling and Water Heater Measure Types by the program.

3.1.2.4. Type of Dwelling

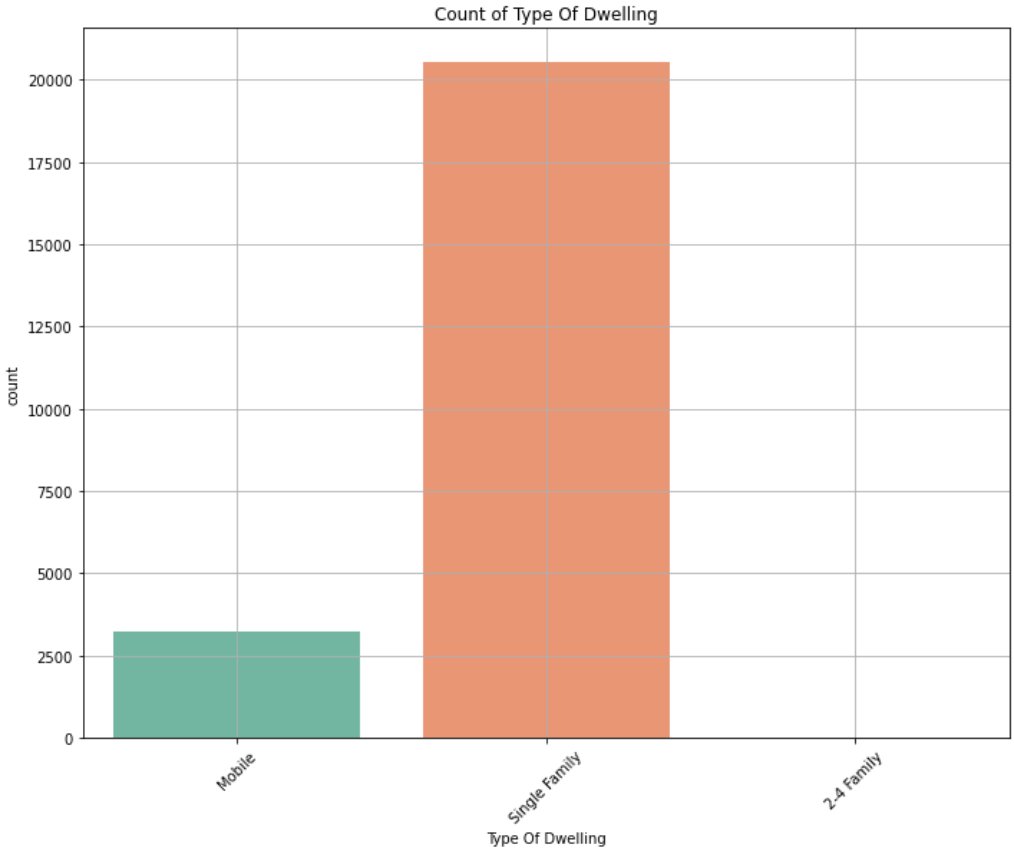


fig. 3.5: bar graph showing the count of different types of dwelling.

This is a general home category describing the dwelling as Single Family, 2-4 Family, Multi Family, or Manufactured/Mobile Home.

- In the year of 2018, 2019 and 2020, NYSERDA have collected most of the billing data from Single Family homes.
- Since, in these three years most of the most common households is being the Single-Family households, getting inferences for performance of the program may get difficult from either Mobile or 2-4 Family households.

- We will analyse the performance for the single-family households and compare the metrics after building the model.

3.1.2.5. Pre-Retrofit Home Heating Fuel Type

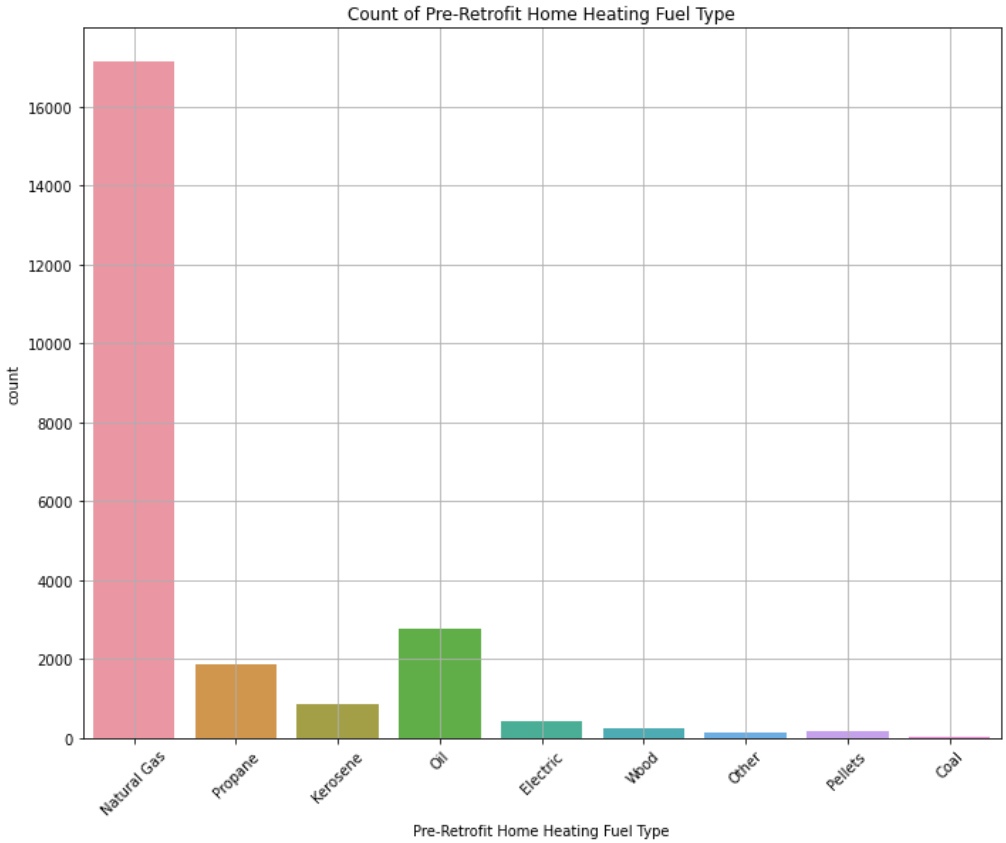


fig. 3.6: bar graph showing the count of different types of pre-retrofitted home heating fuel types.

Different types of heating utilities primarily being used by the home owners. This variable indicates the pre-retrofit primary heating fuel type. Either coal, electricity, kerosene, natural gas, oil, other, pellets, propane, or wood.

- The types for home heating fuel have been very imbalanced in the year 2018, 2019 and 2020.

- This indicates that the regression model wouldn't perform well if other home heating fuel types are taken into consideration.

- Most homes have been using the Natural Gas home heating fuel type.

- Hence, our regression model will be based on the Natural Gas pre-retrofitted home heating fuel type.

3.1.3 Single Family Home Performance done on Building Shell over homes having Natural Gas Pre-Retrofitted Home Heating Fuel type with only 1 Unit served by the program in the year of 2018, 2019 and 2020

3.1.3.1. Region

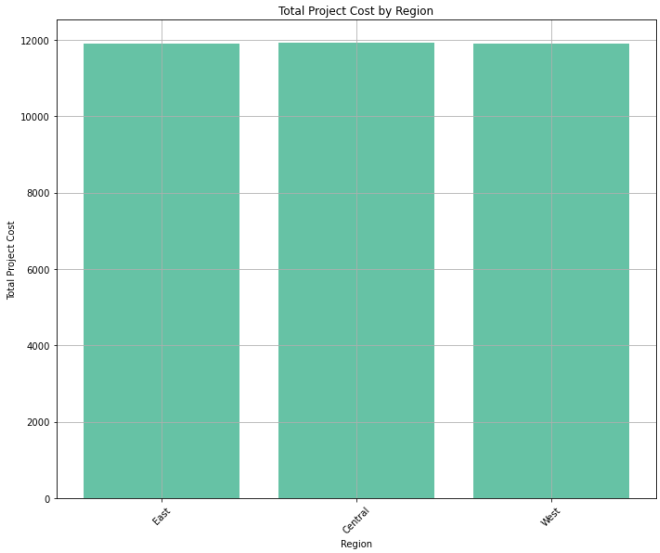


fig. 3.7: bar graph showing the relationship between Total Project Cost and Region.

- West has a greater number of houses with home performance measure taken, with no difference in project cost within the region.

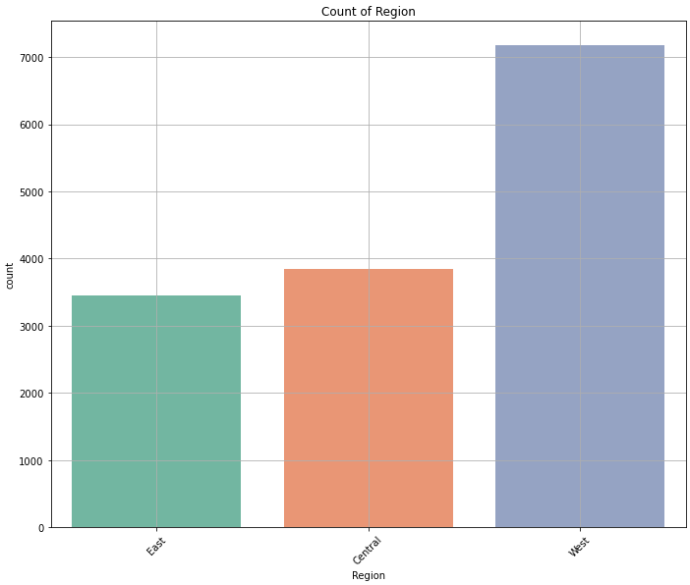


fig. 3.8: bar graph showing the count of homes based on the regions.

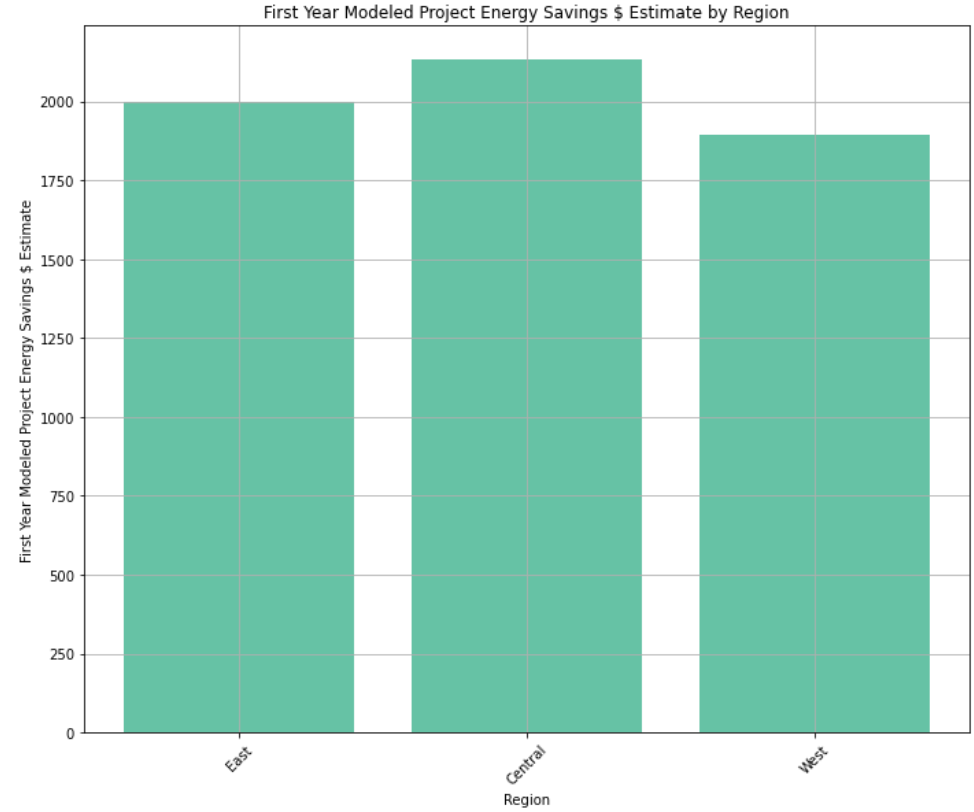


fig. 3.9: bar graph showing the amount of money saved in Regions.

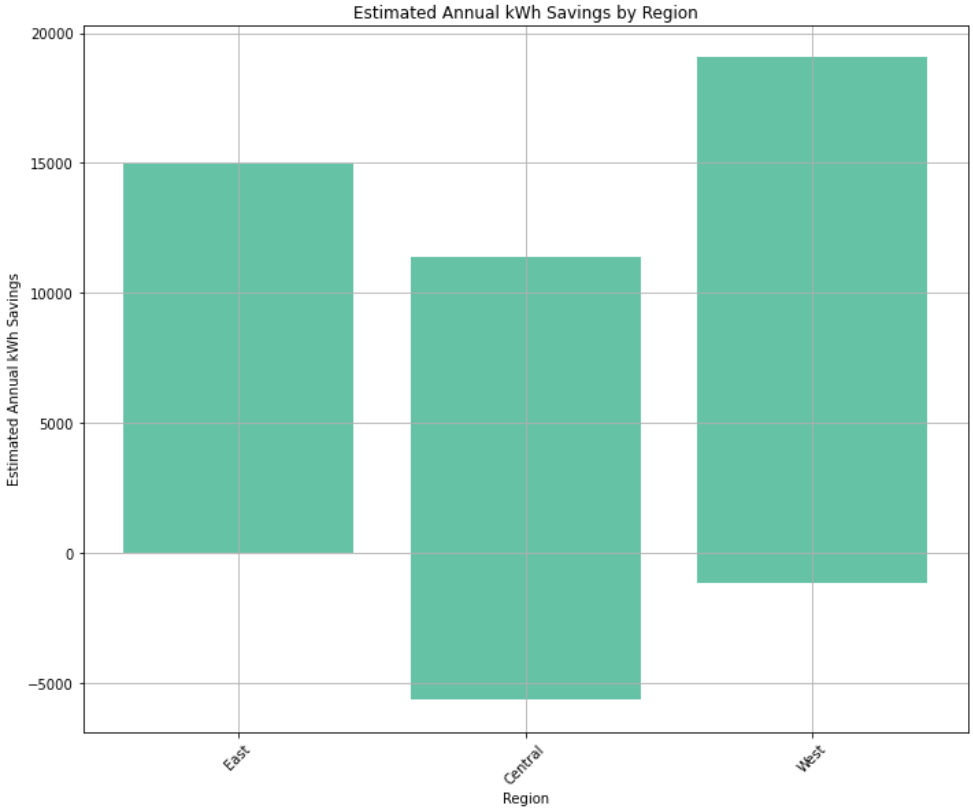


fig. 3.10: bar graph showing the amount of energy saved (kWh) in Regions.

- Central region has the highest energy savings followed by East and then West

- West with the greatest number of homes has the highest savings annually in terms of kWh.

- Central with the second the greatest number of homes has the highest loss in in annual energy savings whereas East has a higher amount of energy savings and lower amount of energy loss when compared to Central.

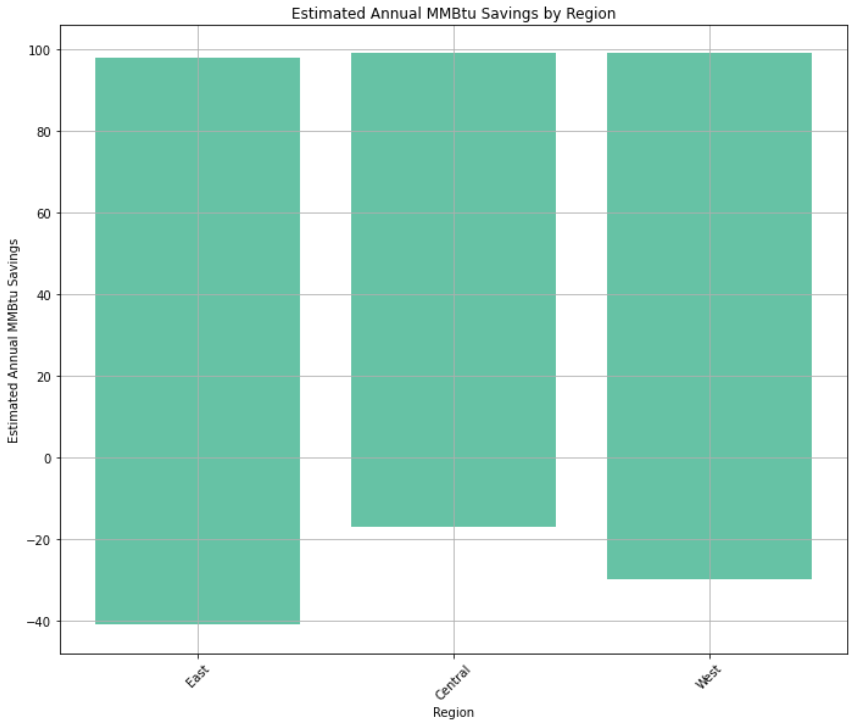


fig. 3.11: bar graph showing the amount of heat saved (MMBtu) in Regions.

- All regions with close to equal amount of heat or MMBtu savings but has a difference in heat loss where East has lost the most amount of heat followed by West and then Central.

3.1.3.2. Billing Month

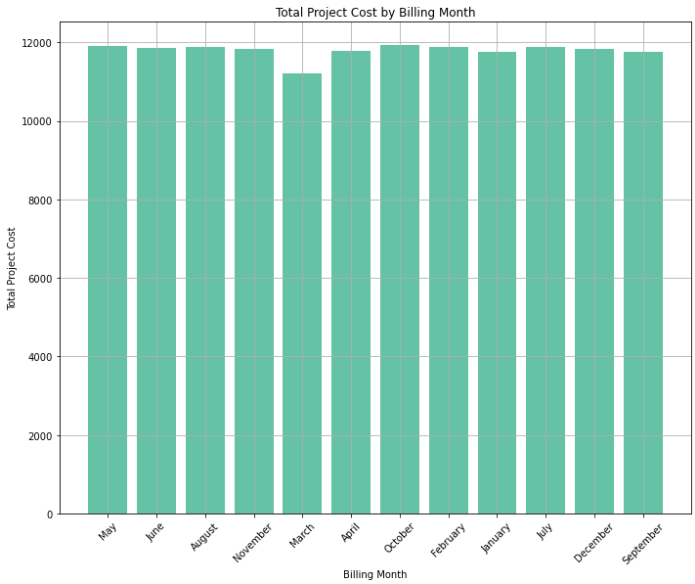


fig. 3.12: bar graph showing the relationship between Total Project Cost and the month of billing.

- March, August and October month have the greatest number of houses which have the Billing data produced with Home Performance measure

- Very small difference in project cost which are completed within the months.

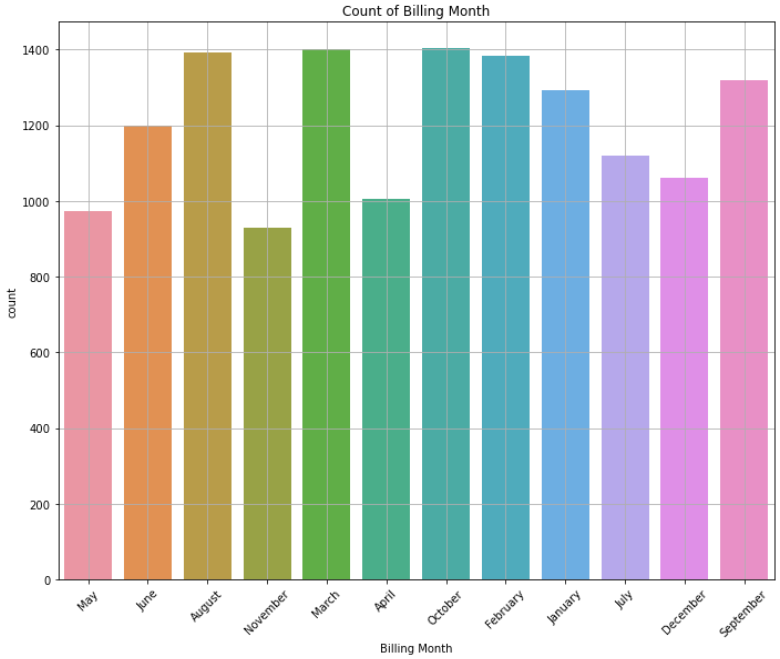


fig. 3.13: bar graph showing the count of homes based on the billing month.

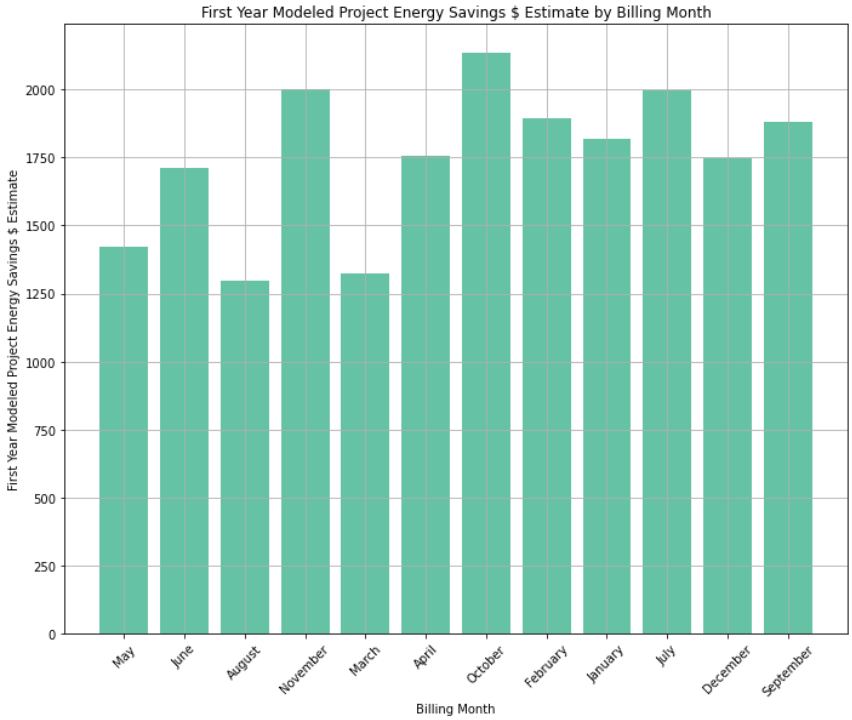


fig. 3.14: bar graph showing the amount of money saved in which months.

- Billing data produced during the month of October displayed the most energy savings in terms of dollars done in one year. With November and July months having the same saving estimate at the end of the year.

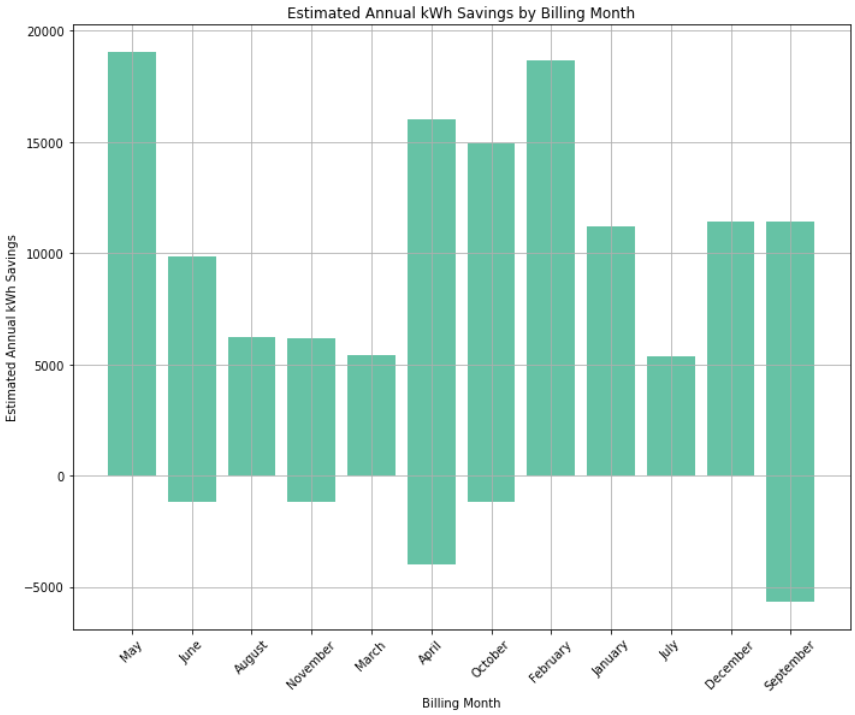


fig. 3.15: bar graph showing the amount of energy saved (kWh) in which months.

- Here, most energy in terms of kWh is saved within the month of May followed by the month of February

- Highest energy loss in terms of kWh has been see during the month of September followed by April.

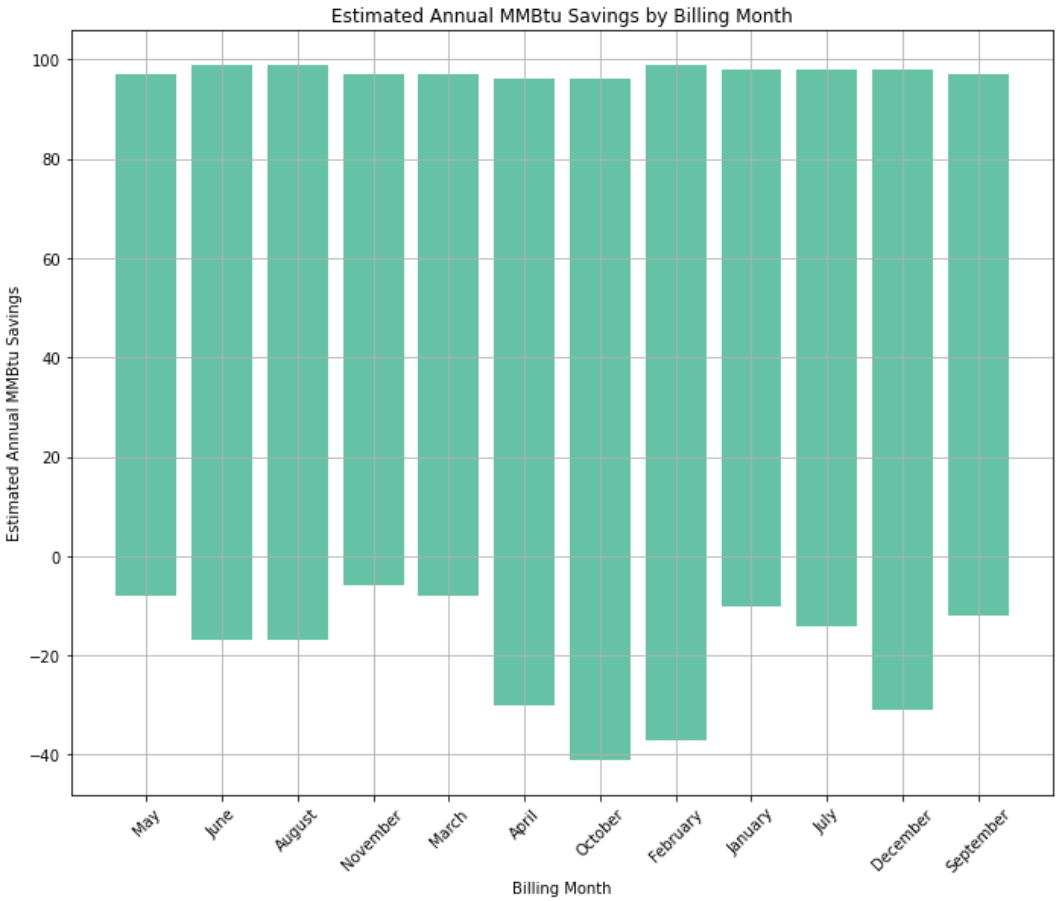


fig. 3.16: bar graph showing the amount of heat saved (MMBtu) in which months.

- All billing months have a close to equal amount of heat or MMBtu savings but have a difference in heat loss where the month of October has experienced the most amount of heat loss followed by February and then December and February.

3.1.3.3. Project Completion Year

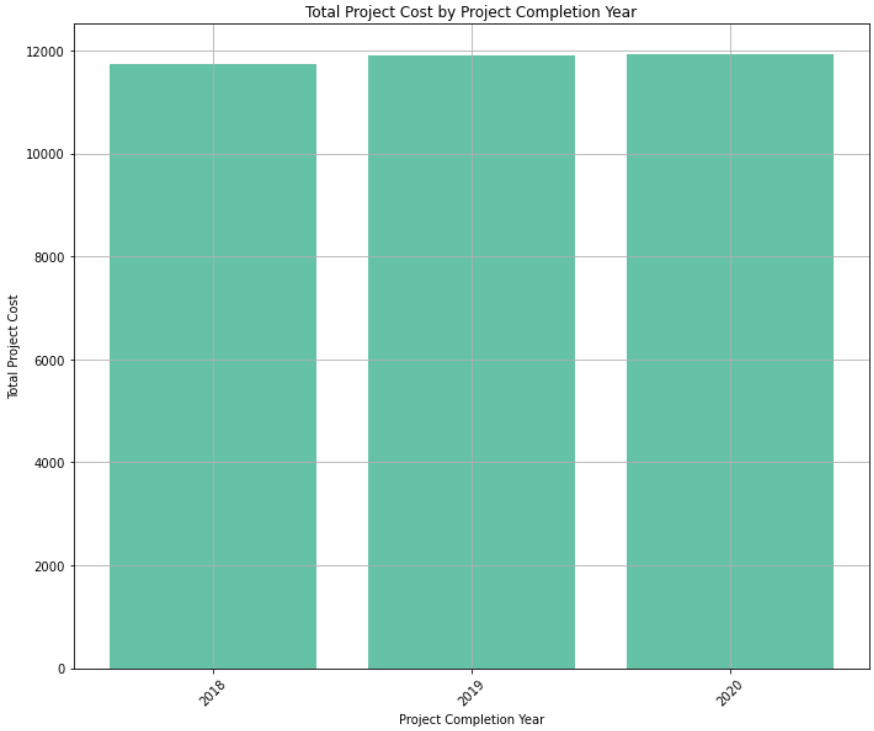


fig. 3.17: bar graph showing the relationship between Total Project Cost and the year of billing.

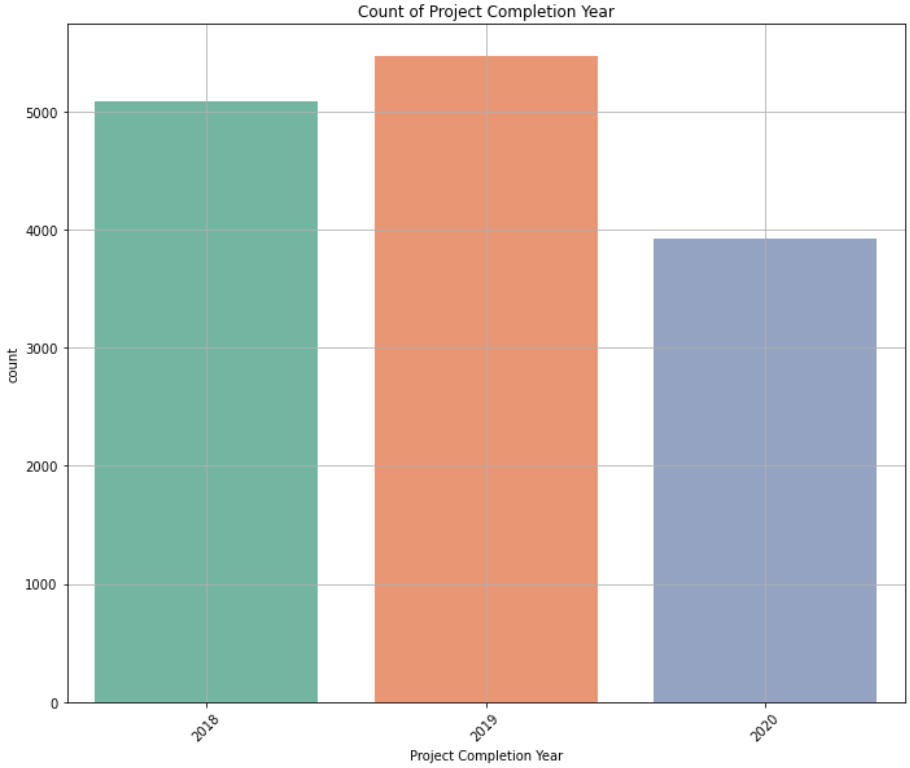


fig. 3.18: bar graph showing the count of homes based on the billing year.

- All the years taken for analysis have very close number of projects done where 2019 has the greatest number of projects done followed by 2018 and 2020.

- The Total Project Cost is close to being the same for all the three years. Keeping the count plot and the Project Cost in mind we will do our further analysis.

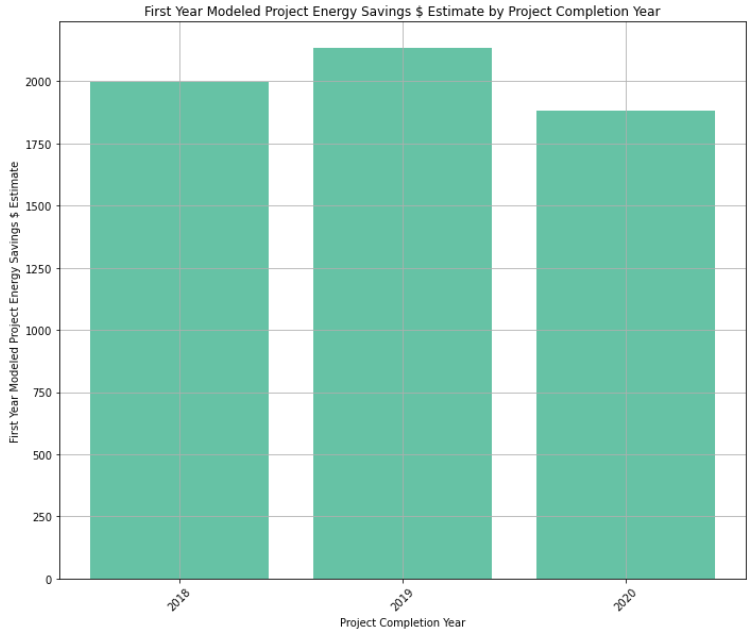


fig. 3.19: bar graph showing the amount of money saved in which year.

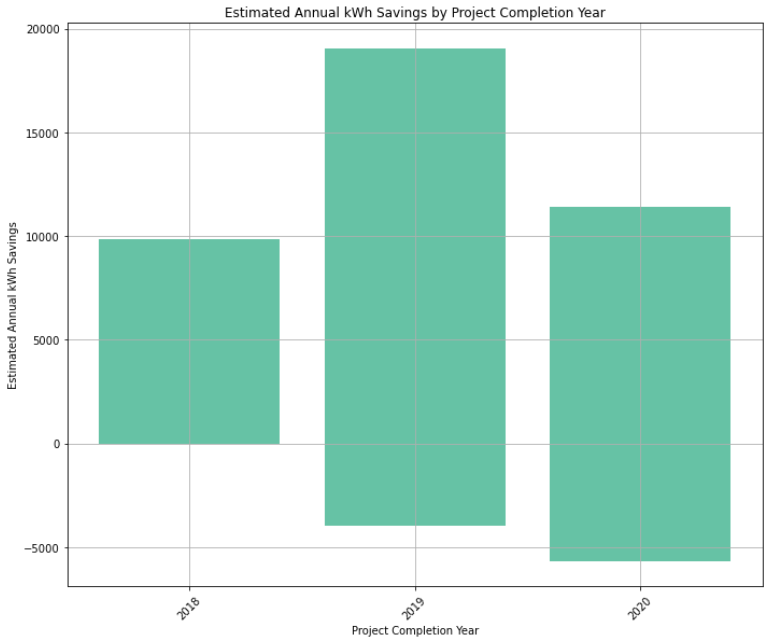


fig. 3.20: bar graph showing the amount of energy saved (kWh) in which year.

- Billing data produced in 2019 shown to be saving the most energy in terms of dollars followed by 2018 and then 2020.
- 2019 also has the most energy saved in terms of kWh within a year whereas 2018 and 2020 have the same amount saved. Although the year 2020 saw the most amount of energy loss in terms of kWh followed by 2019. Year 2018 saw no loss in energy in terms of kWh.

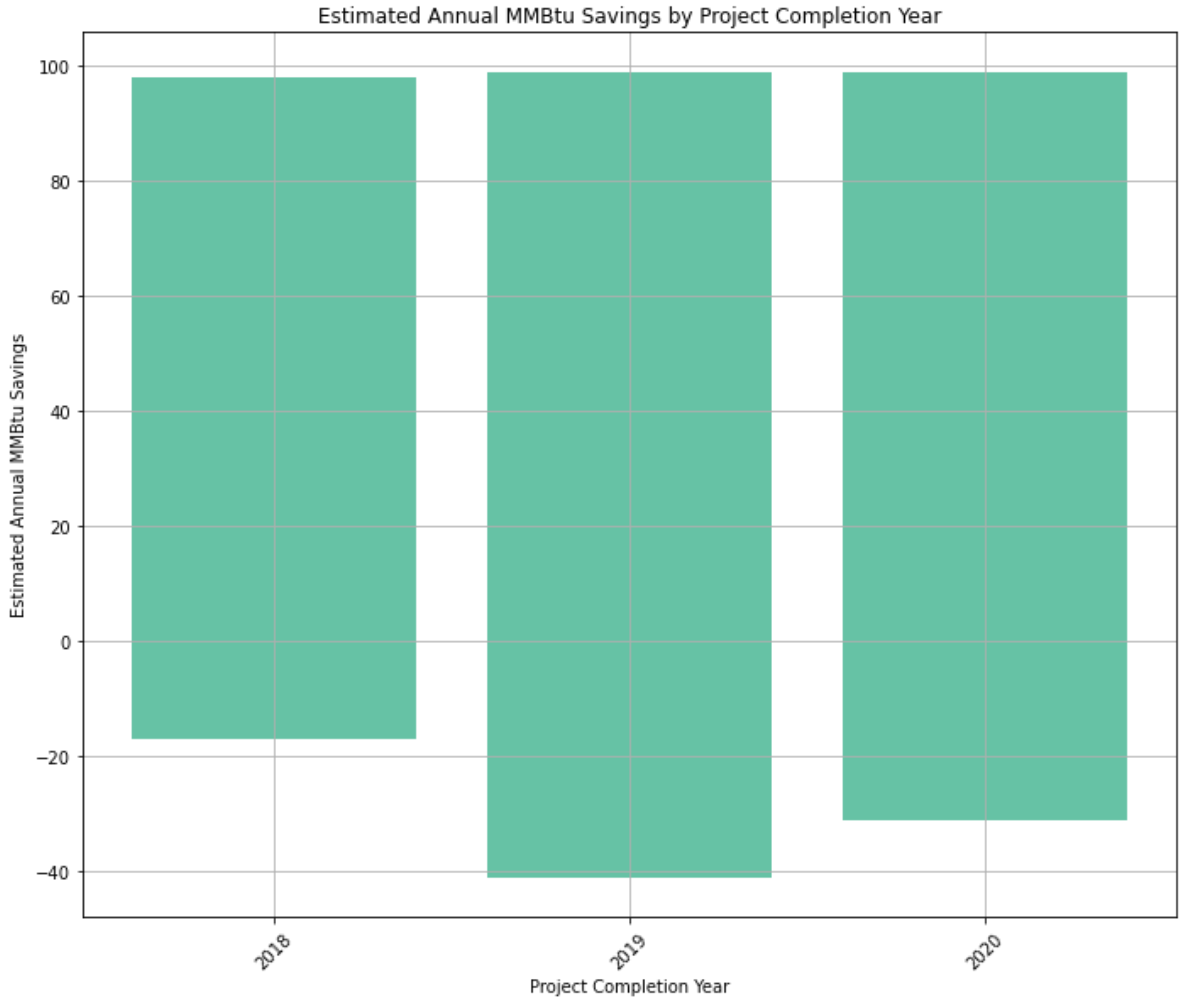


fig. 3.21: bar graph showing the amount of heat saved (MMBtu) in which year.

- Savings produced in terms of heat or MMBtu shows the same amount in all three years, but shows difference in heat loss where 2019 being the year with the most loss in heat

3.1.3.4. Electric Utility

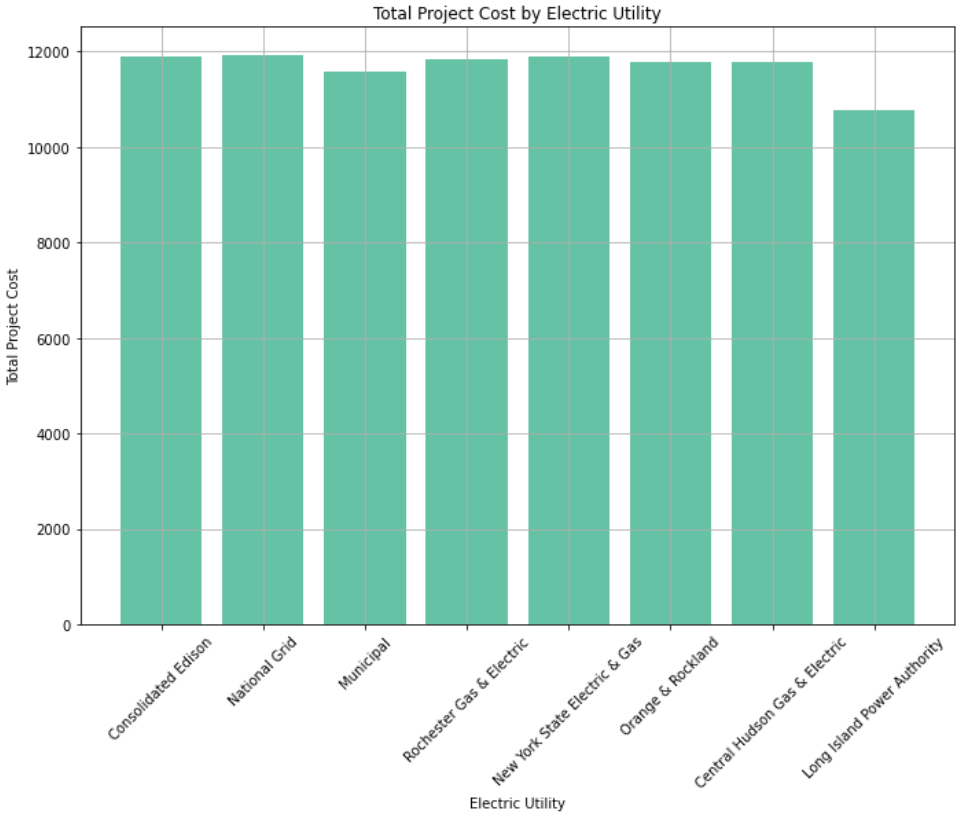


fig. 3.22: bar graph showing the Total Project Cost by Electric Utility.

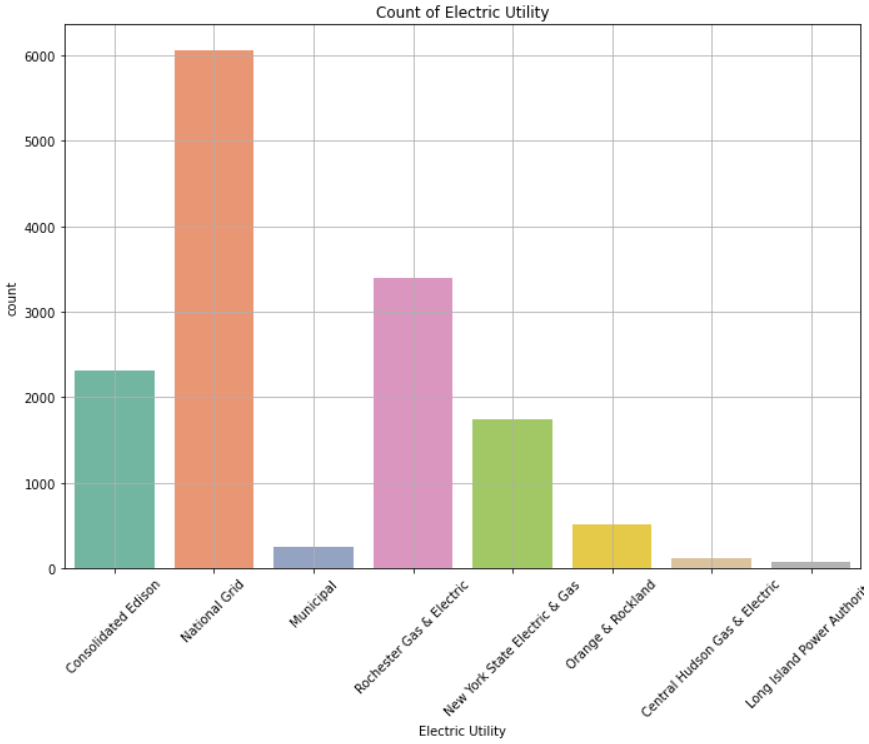


fig. 3.23: bar graph showing the count of Electric Utility used in the state.

- With a nearly equal Total Project Cost for different Electric Utilities, we see that National Grid oversees the greatest number of Homes in the state of New York followed by Rochester Gas & Electric.

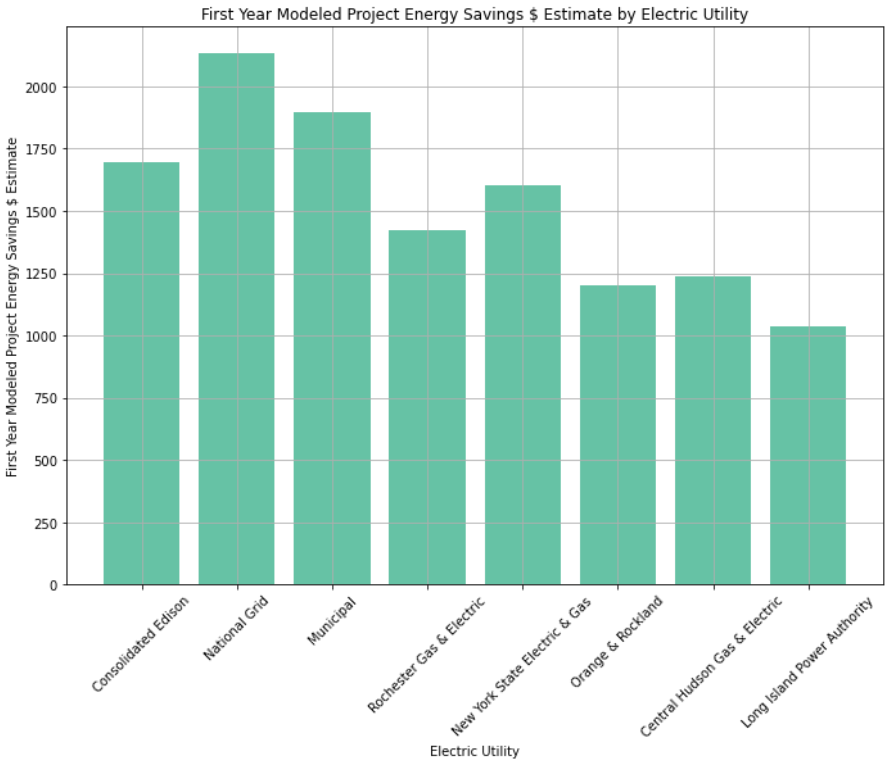


fig. 3.24: bar graph showing the amount of money saved with what Electric Utility.

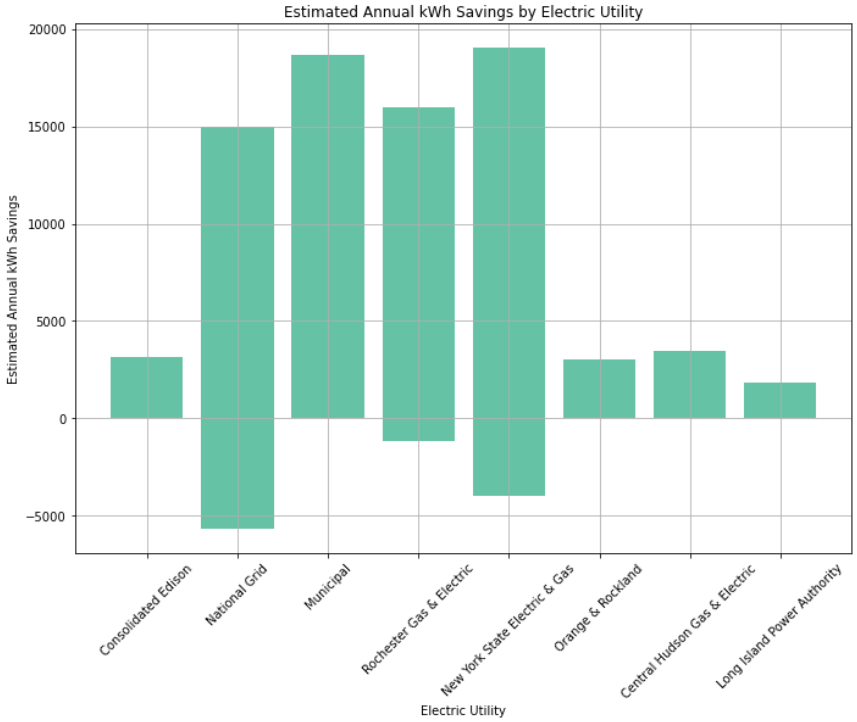


fig. 3.25: bar graph showing the total energy saved (kWh) with what Electric Utility.

- Most of the energy savings in terms of dollars is being produced by National Grid followed by Municipal Utility. Considering the count of Electric Utilities, here the performance of Municipal Utility is better than expected.

- In terms of saving the energy in kWh, we can see that Municipal Utility and New York State Electric & Gas is saving the most amount. But due to New York State Electric & Gas having negative savings we can deduce that the Municipal Utility is the best performing Electric Utility.

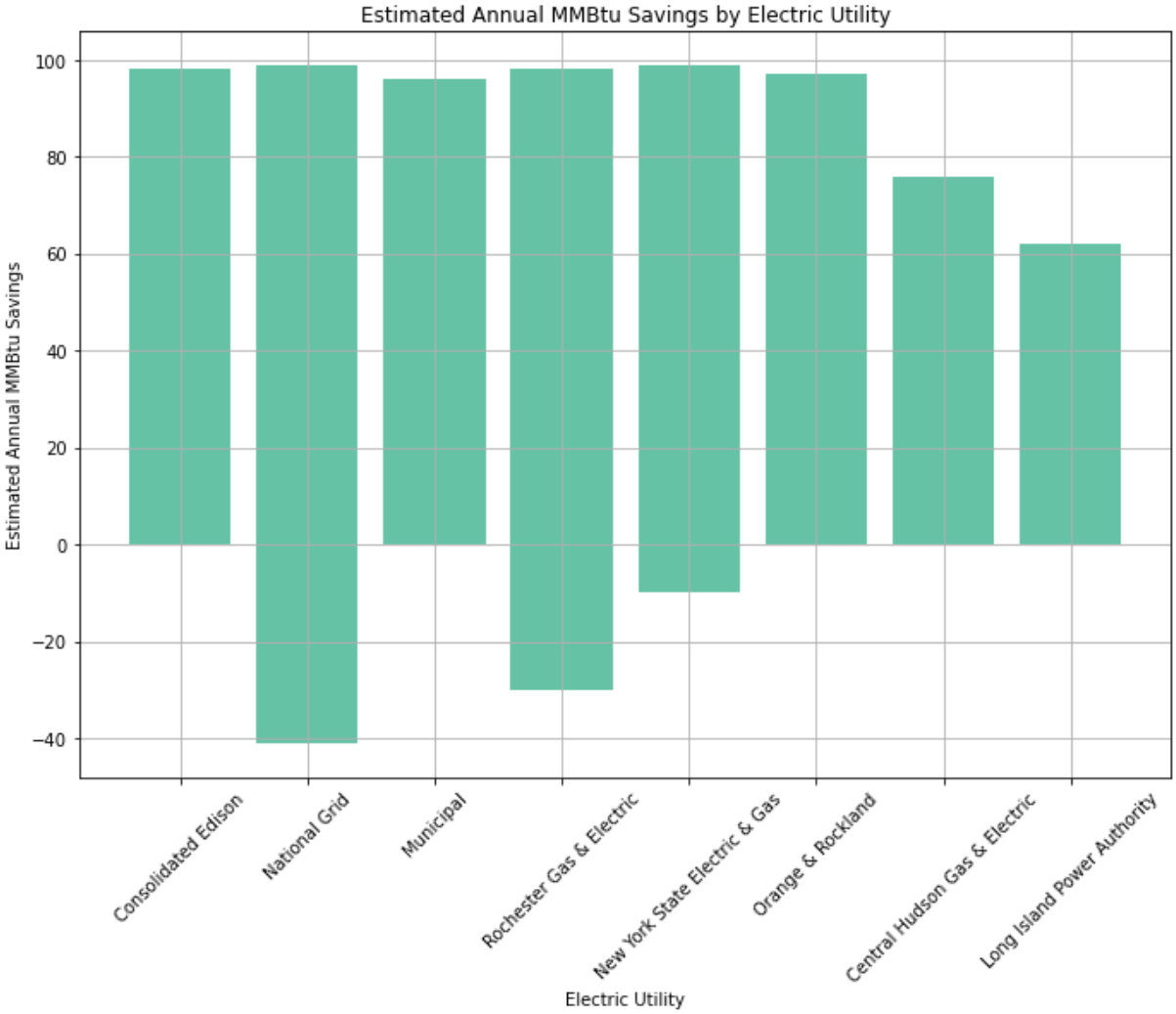


fig. 3.26: bar graph showing the total heat saved (MMBtu) with what Electric Utility.

- In terms of saving the heat in MMBtu, we can see that except Central Hudson Gas & Electric and Long island Power Authority are having lesser savings when compared to other utilities where all other utilities have nearly the same amount of heat saved.

- National Grid performance in saving both energy (kWh) and heat (MMBtu) is low as attrition arises due to the utility being state wide and also the most used utility.

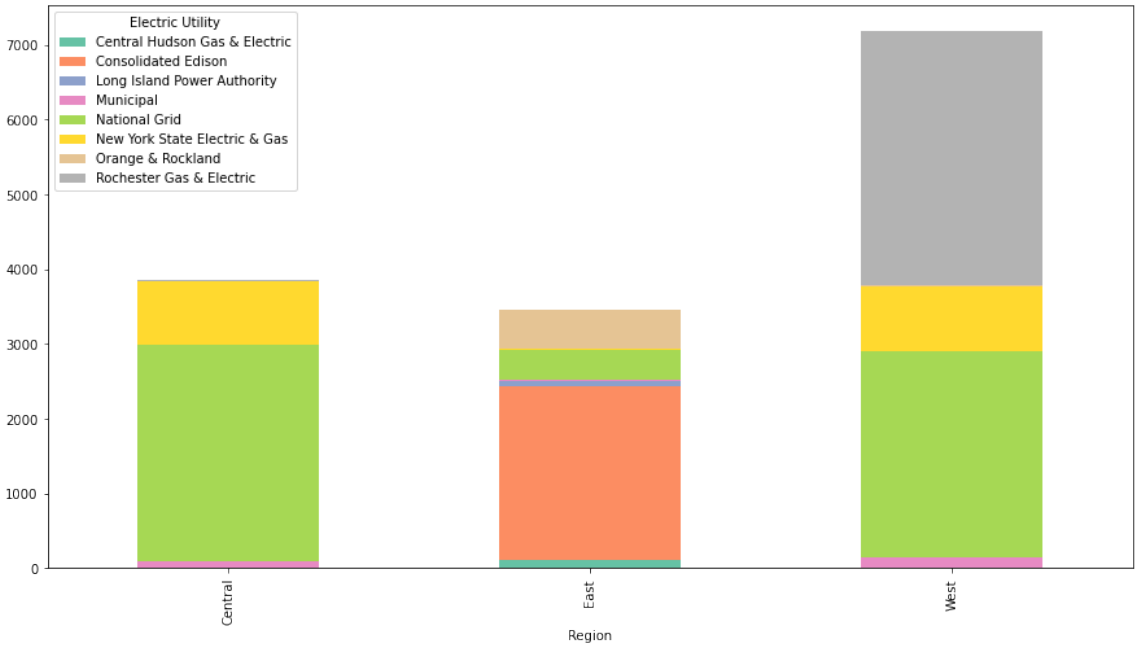


fig. 3.27: stacked bar graph showing different Electric Utilities in Regions.

- Here, we can see that West Region has a competition between Rochester Gas & Electric and National Grid. Consolidated Edison is the most used electric utility in the East Region and most of the West Region is covered by the National Grid.

3.1.3.5. Size of Home

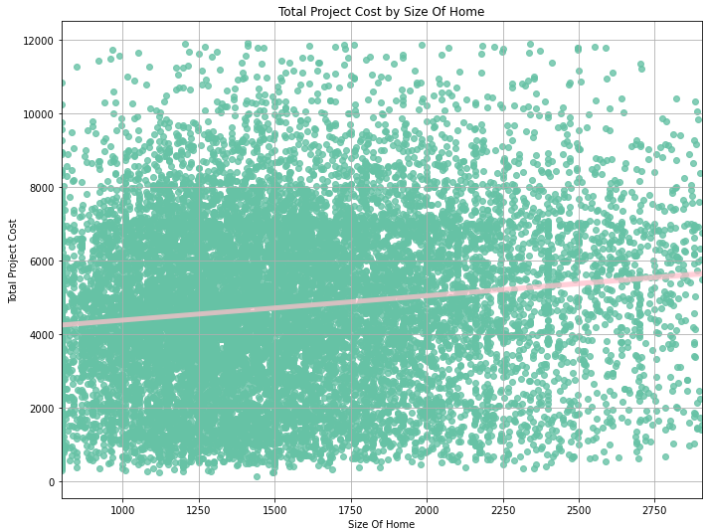


fig. 3.28: scatter plot for relationship between Total Project Cost by Size of Home.

- We need to figure out about how the size of home is reacting with the other variable to see if there is any importance.

- From the correlation plot we saw that it is not correlated with any features.

- From the scatter plot we can see that the Size of Home does not influence the Total Project Cost.

.1.3.6. Location

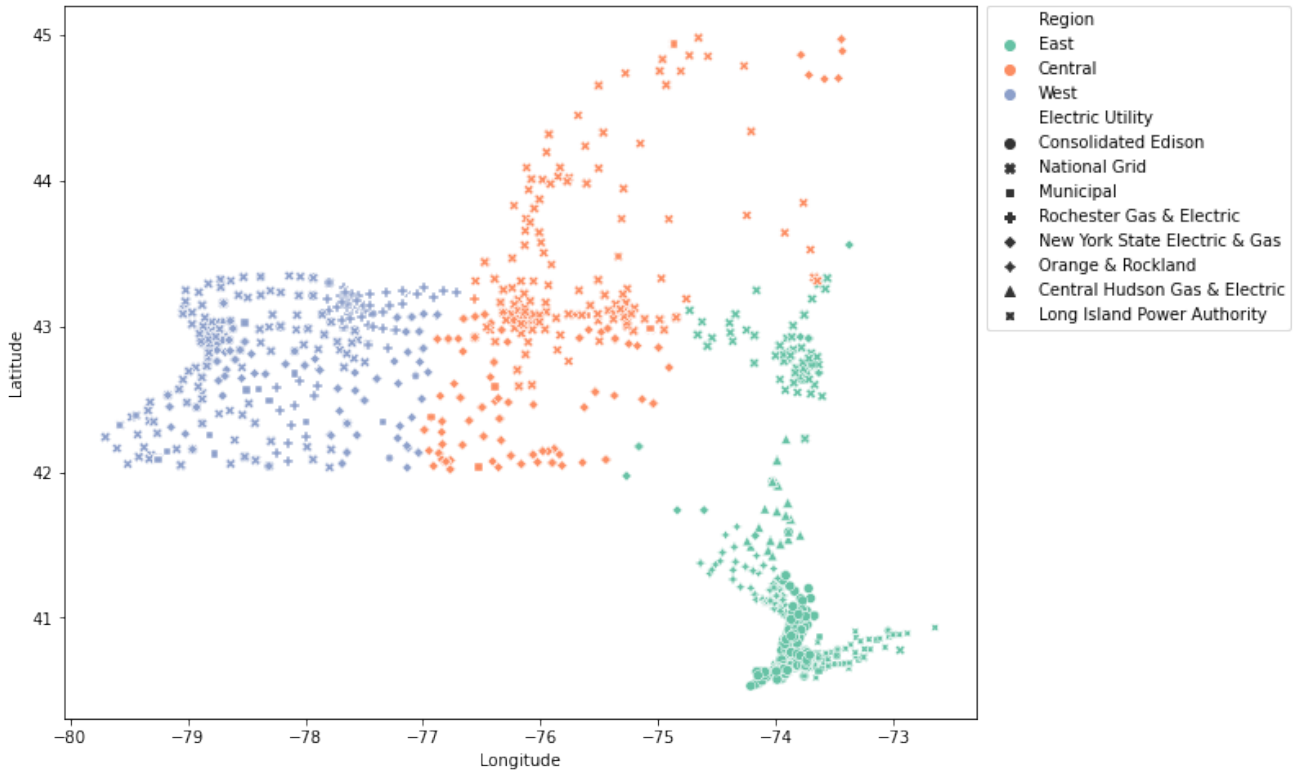
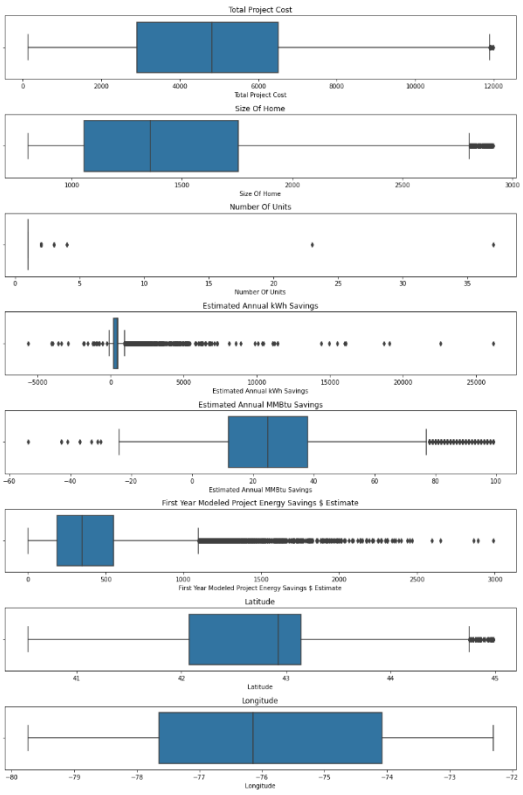


fig. 3.29: scatter plot showing the participants, regions and the Electric Utility used

Here you can see how we have divided the state into three regions with each participant's home Electric Utility.

3.2 Outlier Treatment



Checking for outliers from the box plot we can visualize the outliers. We should infer from the dataset on what variables the outliers exist or is it just extreme values. We inferred that on target variables like 'Estimated Annual kWh Savings', 'First Year Modeled Project Energy Savings $ Estimate' and 'Estimated Annual MMBtu Savings' we might be looking at extreme values. Removing values will add up to the increase in attrition. So, to treat them we dropped these target variables and treated through the IQR Method. The interquartile range is a measure of where the "middle fifty" is in a data set. Where a range is a measure of where the beginning and end are in a set, an interquartile range is a measure of where the bulk of the values lie. That's why it's preferred over many other measures of spread (i.e. the average or median) when reporting things like school performance or SAT scores.

The interquartile range formula is the first quartile subtracted from the third quartile:

IQR = Q3 – Q1.
Upper limit= Q3 + 1.5 * IQR
Upper limit = Q1 - 1.5 * IQR

Hence any values below the lower limit or above the upper limit were capped or removed based on each feature's variation.

3.3 Statistical Significance

3.3.1 Distribution Plots

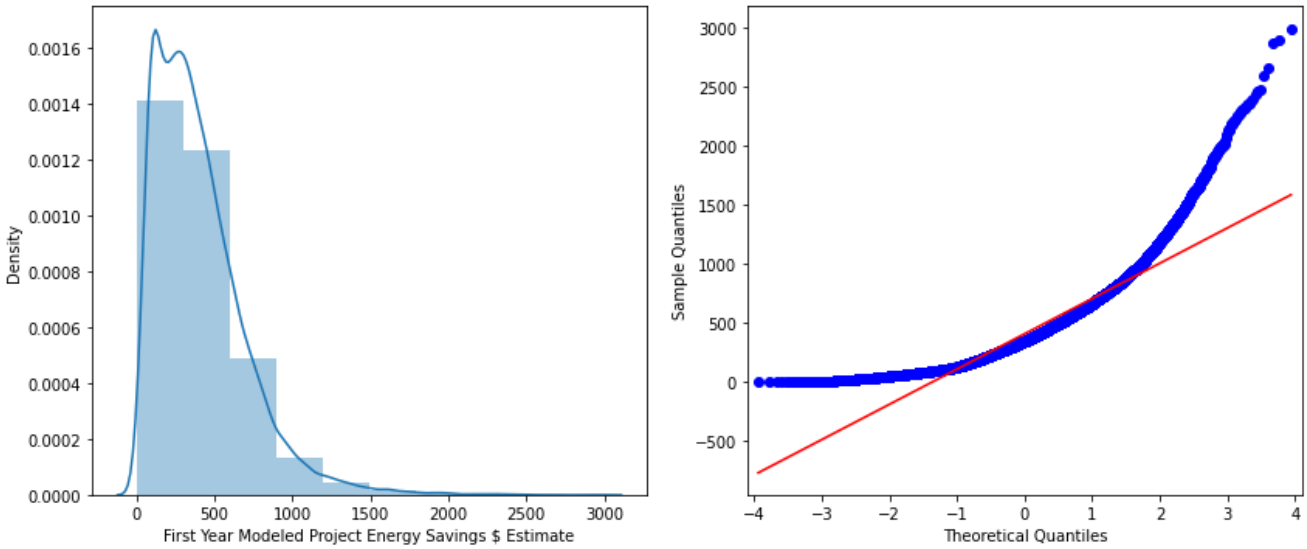


fig. 3.3.1: Distribution plot for First Year Modeled Project Energy Savings $ Estimate

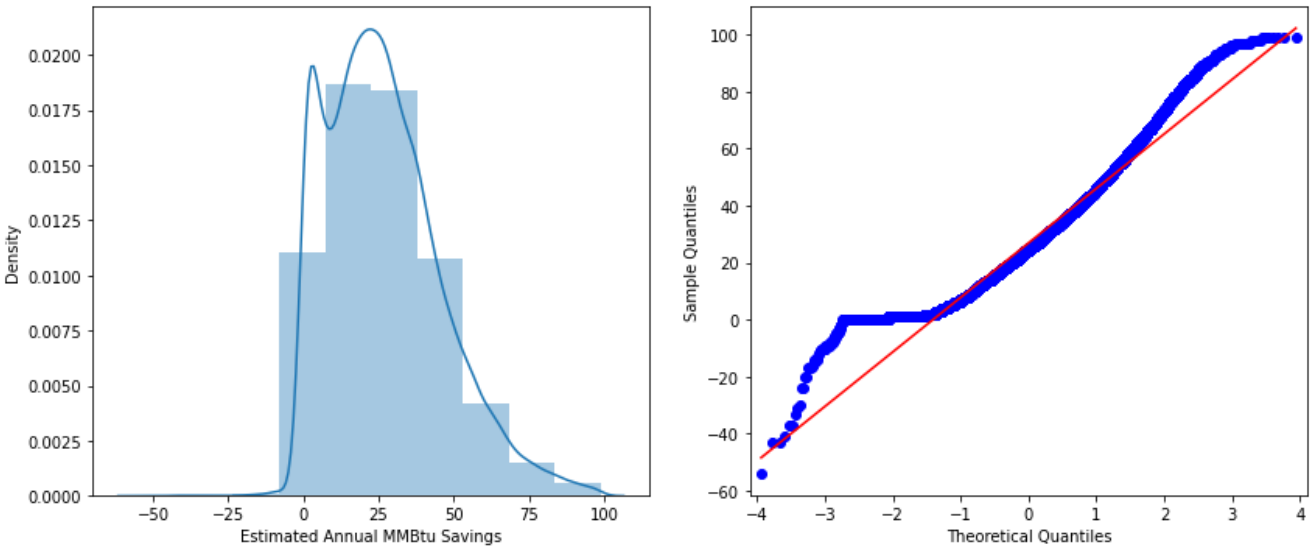


fig. 3.1.2: Distribution plot for Estimated Annual MMBtu Savings

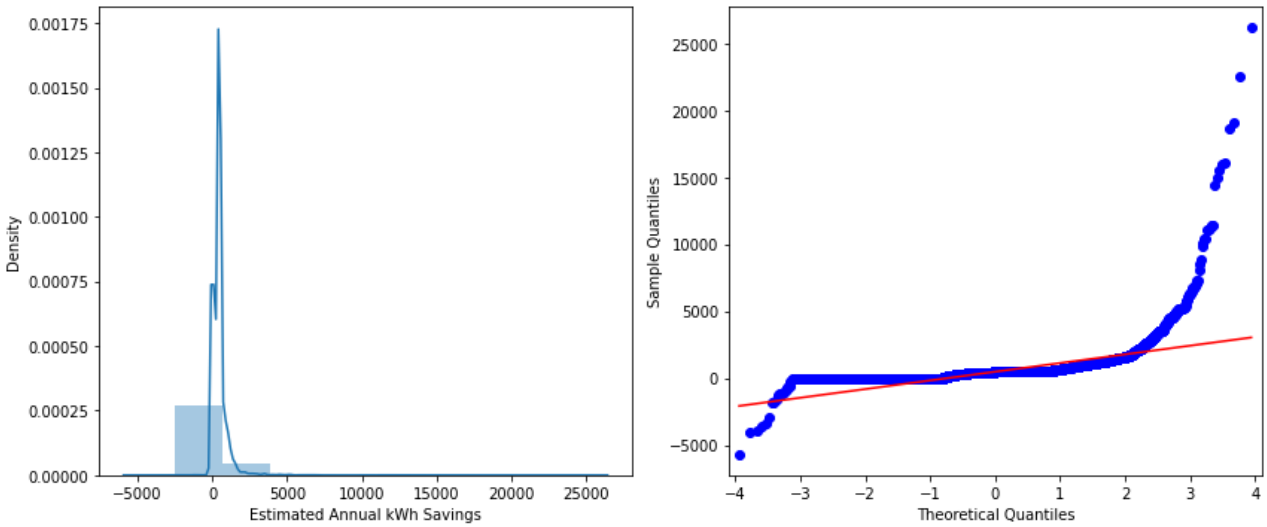


fig. 3.1.3: Distribution plot for Estimated Annual kWh Savings

We produced distribution plots for three variables to see the distribution of the data:

- The distribution for First Year Modeled Project Energy Savings $ Estimate and Estimated Annual MMBtu Savings are highly skewed towards the right.
- The distribution for Estimated Annual kWh Savings is slightly skewed towards the right.
- From the distribution plots we can see that the data is not normally distributed.

We will further check statistically for normality by doing Shapiro-Wilk test.

3.3.2 Shapiro-Wilk Test

Shapiro-Wilk test is conducted to check for normality. This test will be the most preferred as our population is large. For Shapiro-Wilk Test we will first create a null and an alternate hypothesis on the basis of normality.

Hypothesis:

H0: The data is normally distributed.

Ha: The data is not normally distributed.

Variable	Test Statistic	P-Value
First Year Modeled Project Energy Savings $ Estimate	0.882	0
Estimated Annual MMBtu Savings	0.956	0
Estimated Annual kWh Savings	0.956	0

The Shapiro-Wilk test produced Test Statistic and P-Value solutions to help us with the Hypothesis testing. Here the P-Value should be greater than 0.05 for us to fail reject the null hypothesis i.e., H0. As we can see, for all target variables the P-Value is showing to be 0. Hence, we reject the null hypothesis; accepting our assumption created by looking at the distribution plots and QQ-Plots that the data is not normally distributed.

4. Regression Analysis

4.1 Base OLS Models

4.1.1. With Outliers

OLS Regression Results

Dep. Variable:	First Year Modeled Project Energy Savings $ Estimate	**R-squared:**	0.761
Model:	OLS	**Adj. R-squared:**	0.761
Method:	Least Squares	**F-statistic:**	3241.
Date:	Wed, 10 Mar 2021	**Prob (F-statistic):**	0.00

Time:	09:38:19			**Log-Likelihood:**	-17920.	
No. Observations:	25477			**AIC:**	3.589e+04	
Df Residuals:	25451			**BIC:**	3.610e+04	
Df Model:	25					
Covariance Type:	nonrobust					
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	0.1445	0.026	-5.622	0.000	-0.195	-0.094
Electric Utility_Consolidated Edison	0.1641	0.017	9.388	0.000	0.130	0.198
Electric Utility_Long Island Power Authority	0.2868	0.029	-9.918	0.000	-0.344	-0.230
Electric Utility_Municipal	0.2623	0.029	-9.042	0.000	-0.319	-0.205
Electric Utility_National Grid	0.2031	0.020	-9.928	0.000	-0.243	-0.163
Electric Utility_New York State Electric & Gas	0.1732	0.021	-8.346	0.000	-0.214	-0.133
Electric Utility_Rochester Gas & Electric	0.1578	0.022	-7.046	0.000	-0.202	-0.114
Pre-Retrofit Home Heating Fuel Type_Electric	0.2667	0.023	11.691	0.000	0.222	0.311
Pre-Retrofit Home Heating Fuel Type_Kerosene	0.9253	0.019	50.003	0.000	0.889	0.962
Pre-Retrofit Home Heating Fuel Type_Oil	0.9131	0.010	87.359	0.000	0.893	0.934
Pre-Retrofit Home Heating Fuel Type_Other	0.8093	0.036	22.692	0.000	0.739	0.879
Pre-Retrofit Home Heating Fuel Type_Pellets	0.5090	0.037	13.892	0.000	0.437	0.581
Pre-Retrofit Home Heating Fuel Type_Propane	1.0840	0.013	84.613	0.000	1.059	1.109
Pre-Retrofit Home Heating Fuel Type_Wood	0.5192	0.029	17.732	0.000	0.462	0.577
Job Type_Home Performance	0.1034	0.016	6.283	0.000	0.071	0.136
Type Of Dwelling_Mobile	0.0307	0.011	2.793	0.005	0.009	0.052
Project Completion Year_2019	0.1192	0.007	-16.139	0.000	-0.134	-0.105
Project Completion Year_2020	0.2282	0.008	-27.470	0.000	-0.244	-0.212
Project Completion Year_2021	0.2818	0.019	-15.133	0.000	-0.318	-0.245
Region_North	0.0956	0.013	7.625	0.000	0.071	0.120
Region_South	0.1339	0.023	5.799	0.000	0.089	0.179

Region_West	0.0312	0.009	-3.497	0.000	-0.049	-0.014
Total Project Cost	0.1076	0.005	20.904	0.000	0.098	0.118
Size Of Home	0.0126	0.003	3.667	0.000	0.006	0.019
Estimated Annual kWh Savings	0.2127	0.003	66.875	0.000	0.206	0.219
Estimated Annual MMBtu Savings	0.6694	0.005	135.888	0.000	0.660	0.679

Omnibus:	17015.885	**Durbin-Watson:**	2.047
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	631932.766
Skew:	2.707	**Prob(JB):**	0.00
Kurtosis:	26.790	**Cond. No.**	25.8

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

The base OLS model shown above was modeled after applying backward elimination feature selection technique. Backward elimination technique removed following features based on their P-Value being greater than 0.05:

Features Removed	**P-Value > 0.05**
Type Of Dwelling_Single Family	0.948
Measure Type_Water Heater	0.651
Region_East	0.264
Number Of Units	0.237
	0.148

Pre-Retrofit Home Heating Fuel Type_Natural Gas	
Electric Utility_Orange & Rockland	0.097
Measure Type_Heating and Cooling	0.067

4.1.2. Without Outliers

OLS Regression Results

Dep. Variable:	First Year Modeled Project Energy Savings $ Estimate	**R-squared:**	0.919
Model:	OLS	**Adj. R-squared:**	0.919
Method:	Least Squares	**F-statistic:**	8477.
Date:	Sat, 17 Apr 2021	**Prob (F-statistic):**	0.00
Time:	10:12:09	**Log-Likelihood:**	-69650.
No. Observations:	12784	**AIC:**	1.393e+05
Df Residuals:	12766	**BIC:**	1.395e+05
Df Model:	17		
Covariance Type:	nonrobust		

Omnibus:	9660.065	**Durbin-Watson:**	1.998
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	464451.770
Skew:	3.171	**Prob(JB):**	0.00
Kurtosis:	31.839	**Cond. No.**	8.61e+04

4.1.3. With Outliers and Square Root Target

OLS Regression Results

Dep. Variable:	$sqr	R-squared:	0.903
Model:	OLS	Adj. R-squared:	0.903
Method:	Least Squares	F-statistic:	5859.
Date:	Sat, 17 Apr 2021	Prob (F-statistic):	0.00
Time:	10:12:09	Log-Likelihood:	-29604.
No. Observations:	14474	AIC:	5.926e+04
Df Residuals:	14450	BIC:	5.944e+04
Df Model:	23		
Covariance Type:	nonrobust		

Omnibus:	5274.940	Durbin-Watson:	2.008
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3081881.865
Skew:	-0.076	Prob(JB):	0.00
Kurtosis:	74.486	Cond. No.	1.25e+05

4.1.3. Without Outliers and Square Root Target

OLS Regression Results

Dep. Variable:	$sqr	R-squared:	0.929
Model:	OLS	Adj. R-squared:	0.929
Method:	Least Squares	F-statistic:	7962.
Date:	Sat, 17 Apr 2021	Prob (F-statistic):	0.00
Time:	10:12:10	Log-Likelihood:	-23249.
No. Observations:	12784	AIC:	4.654e+04
Df Residuals:	12762	BIC:	4.671e+04
Df Model:	21		
Covariance Type:	nonrobust		

Omnibus:	4658.192	Durbin-Watson:	1.992
Prob(Omnibus):	0.000	Jarque-Bera (JB):	145670.647
Skew:	1.123	Prob(JB):	0.00
Kurtosis:	19.384	Cond. No.	4.70e+04

4.1.4 Testing for Assumptions for OLS Base Model

4.1.4.1. Assumption 1: Multicollinearity

Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. Mathematically, the VIF for a regression model

	VIF
const	70.332706
Electric Utility_Consolidated Edison	4.554450
Electric Utility_Long Island Power Authority	1.467772
Electric Utility_Municipal	1.652786
Electric Utility_National Grid	10.707479
Electric Utility_New York State Electric & Gas	6.551557
Electric Utility_Rochester Gas & Electric	7.376853
Pre-Retrofit Home Heating Fuel Type_Electric	1.013593
Pre-Retrofit Home Heating Fuel Type_Kerosene	1.239887
Pre-Retrofit Home Heating Fuel Type_Oil	1.232799
Pre-Retrofit Home Heating Fuel Type_Other	1.018072
Pre-Retrofit Home Heating Fuel Type_Pellets	1.019139
Pre-Retrofit Home Heating Fuel Type_Propane	1.273880
Pre-Retrofit Home Heating Fuel Type_Wood	1.027196
Job Type_Home Performance	1.505131
Type Of Dwelling_Mobile	1.471898
Project Completion Year_2019	1.350748
Project Completion Year_2020	1.431213
Project Completion Year_2021	1.094434
Region_North	1.287373
Region_South	9.928254
Region_West	2.004891
Total Project Cost	2.822858
Size Of Home	1.249677
Estimated Annual kWh Savings	1.076972
Estimated Annual MMBtu Savings	2.583257

fig. 4.1: variance inflation factor of variables.

variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable. This ratio is calculated for each independent variable. A high VIF indicates that the associated independent variable is highly collinear with the other variables in the model.

4.1.4.2. Assumption 2: Normality of Residuals

Normality is the assumption that the underlying residuals are normally distributed, or approximately so. If the test p-value is less than the predefined significance level, you can reject the null hypothesis and conclude the residuals are not from a normal distribution. Here we will check the normality of residuals by using probplot or a Q-Q plot from stats module available in the SciPy library. We will also plot a distribution plot for the residuals and fit the normal imported from the same library through it.

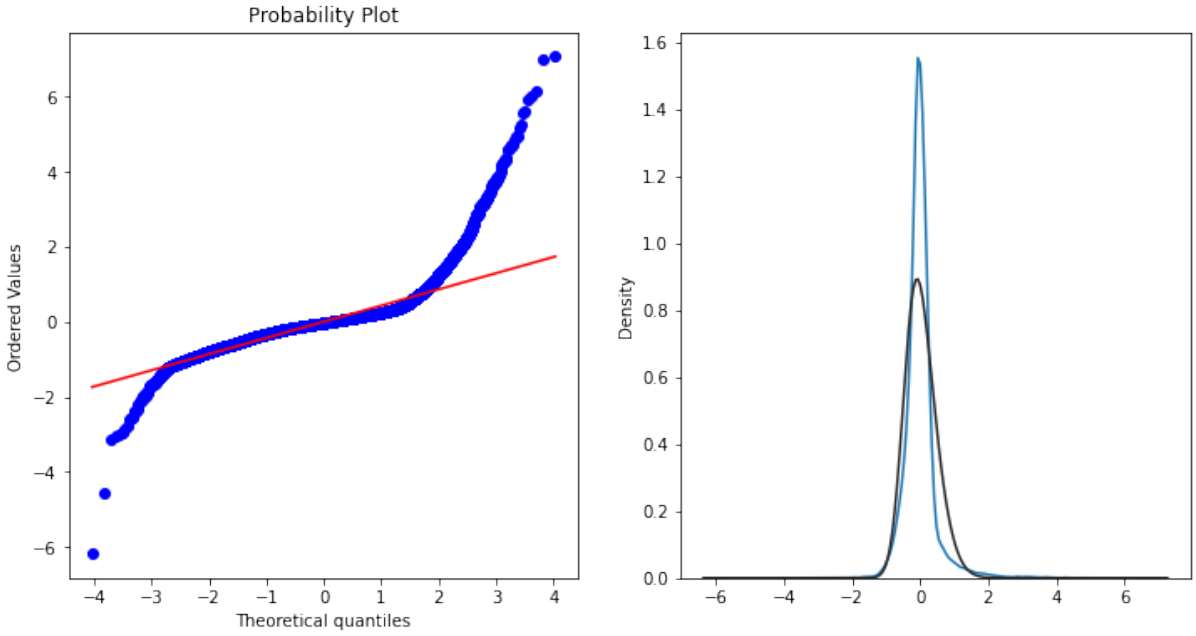


fig. 4.2: Q-Q plot and the distribution plot for the residuals.

The first thing that can be observed is the fact that points form a curve rather than a straight line, which usually is an indication of skewness in the sample data. Another way of interpreting the plot is by looking at the tails of the distribution. In this case, the considered Skew Normal distribution has a lighter left tail (less mass, points on the left side of Q-Q plot below the line) and heavier right tail (more mass, points on the right side of Q-Q plot above the line) than one could expect under Standard Normal distribution. The conclusion is that there is definitely more mass in the tails (indicating more negative and positive returns) than as assumed under Normality.

4.1.4.3. Assumption 3: Homoscedasticity

Homoscedasticity describes a situation in which the error term (that is, the "noise" or random disturbance in the relationship between the independent variables and the

dependent variable) is the same across all values of the independent variables. In simple terms, Homoscedasticity refers to whether the residuals are equally distributed, or whether they tend to bunch together at some values, and at other values, spread far apart.

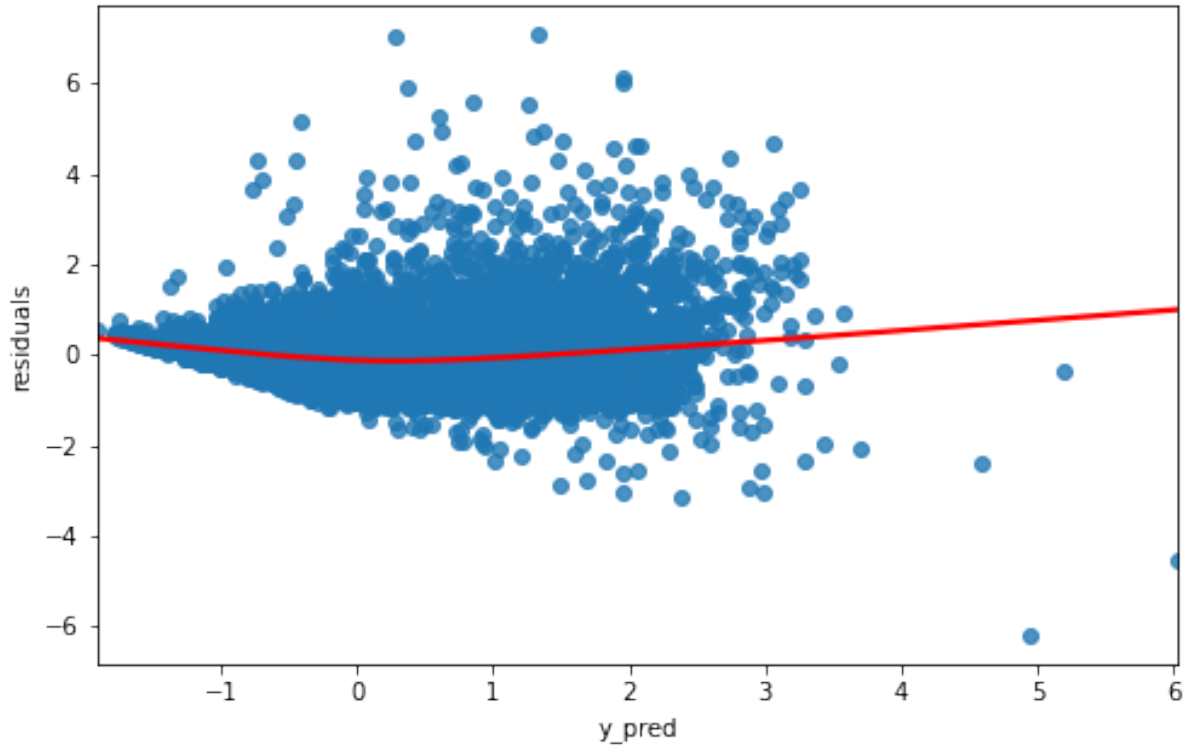


fig. 4.3: Scatter plot for the predicted target variable vs the residuals to check for homoscedasticity.

The scatter plot is good way to check whether the data are homoscedastic (meaning the residuals are equal across the regression line). The above scatter plot shows that the residuals are not homoscedastic (i.e., heteroscedastic). We will further check for this assumption with Goldfeld Quandt Test.

Goldfeld Quandt Test for Homoscedasticity

The Goldfeld Quandt Test is a test used in regression analysis to test for homoscedasticity. It compares variances of two subgroups; one set of high values and one set of low values. If the variances differ, the test rejects the null hypothesis that the variances of the errors are not constant.

Hypothesis:

H0: Variance of residuals is constant across the range of data.

Ha: Variance of residuals is not constant across the range of data.

Test Results:

F-Statistic	-	1.0940339895673918
P-Value	-	2.0360055000948948e-07
Type	-	Increasing

Since P-Value (2.0360055000948948e-07) is less than significance level, we will reject H0 to conclude that variance of residuals is not constant.

4.1.4.4. Assumption 4: Auto-Correlation

Autocorrelation is a type of serial dependence. Specifically, autocorrelation is when a time series is linearly related to a lagged version of itself. By contrast, correlation is simply when two independent variables are linearly related. We need to check for auto-correlation because if we try to do regression analysis on data with autocorrelation, then our analysis will be misleading. We will use acf_plot by using time series analysis module from Statsmodels.

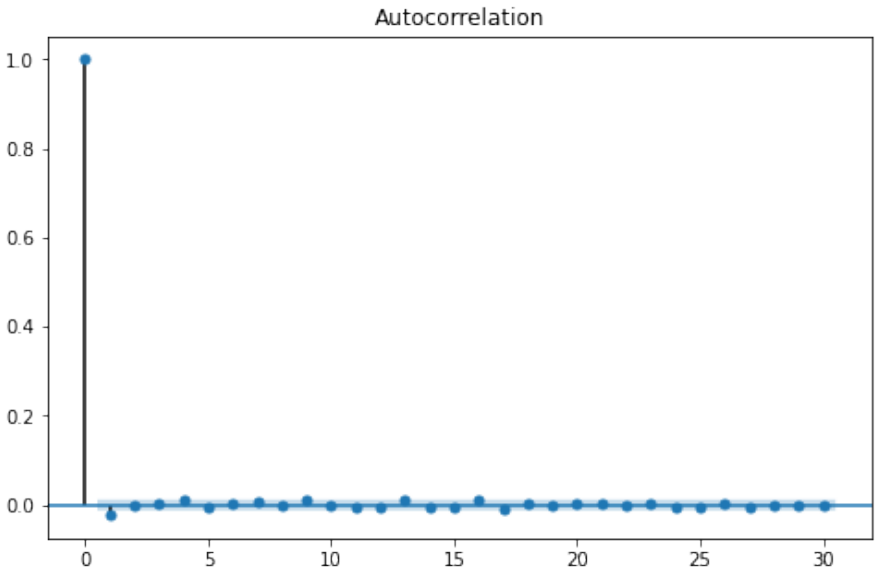


fig. 4.2.3: acf_plot for checking Auto-Correlation through the residuals

From this plot, we see that values for the ACF are within 95% confidence interval (represented by the solid gray line) for lags > 0, which verifies that our data doesn't have any autocorrelation.

4.1.4.5 Assumption 5: Linearity of Relationship

Here we will check the linear relationship of the predicted target variable and the target variable through scatter plots and a diagnostic measure called linear_rainbow from the stats module.

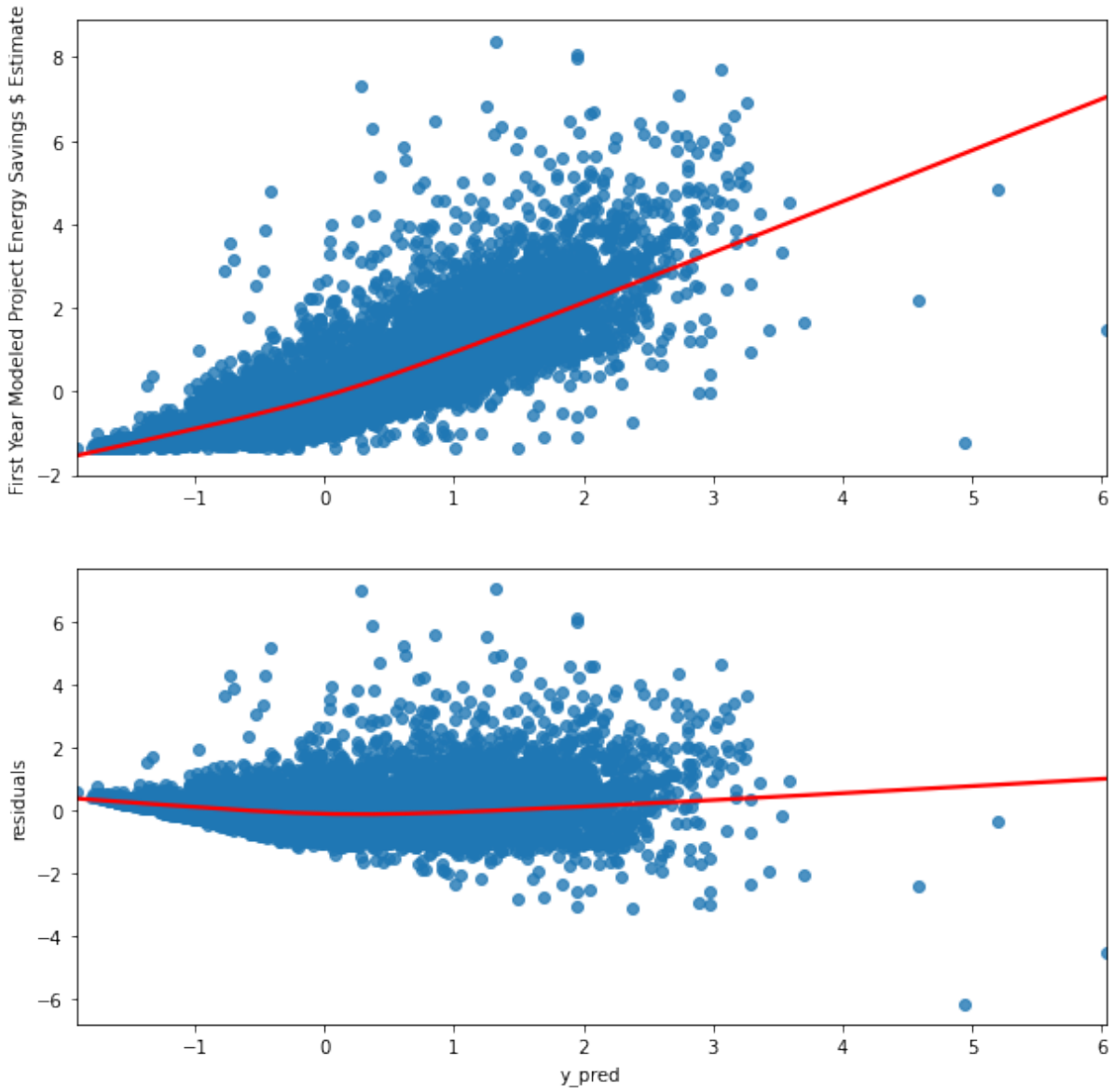


fig. 4.2.4: Scatter plot for checking the linear relationship y and y_pred.

Rainbow Test for Linearity

The basic idea of the Rainbow test is that even if the true relationship is non-linear, a good linear fit can be achieved on a subsample in the "middle" of the data. The null hypothesis is rejected whenever the overall fit is significantly worse than the fit for the subsample. This test assumes residuals are homoscedastic and may reject a correct linear specification if the residuals are heteroskedastic.

Hypothesis:

H0: Fit of model using full sample = Fit of model using a central subset (linear relationship)

Ha: Fit of model using full sample is worse compared to fit of model using a central subset.

Test Results:

F-Statistic - 1.2389361586884111
P-Value - 7.385520188495804e-34

Since, P-Value (7.385520188495804e-34) is lower than significance level, we will reject the H0 to conclude that Fit of model using full sample is worse compared to fit of model using a central subset. We need to improve our model.

4.2 OLS Base Model Summary

Interpretation:

The R-squared value obtained from this model is 0.761 which means that the above model explains 76.1% of the variation in the First Year Modeled Project Energy Savings $ Estimate.

Durbin-Watson Test:

The test is used to check the autocorrelation between the residuals.

- If the Durbin-Watson test statistic is near to 2: no autocorrelation
- If the Durbin-Watson test statistic is between 0 and 2: positive autocorrelation
- If the Durbin-Watson test statistic is between 2 and 4: negative autocorrelation

The summary output shows that the value of the test statistic is close to 2 (= 2.047) which means there is no autocorrelation.

Jarque-Bera Test:

The test is used to check the normality of the residuals. Here, the p-value of the test is less than 0.05; that implies the residuals are not normally distributed**.**

'Cond. No':

(= 1) represents the Condition Number (CN) which is used to check the multicollinearity.

- If CN < 100: no multicollinearity
- If CN is between 100 and 1000: moderate multicollinearity
- If CN > 1000: severe multicollinearity

With Cond. No. = 25.8, it can be seen that there is mild multicollinearity in the data as seen through assumption testing.

4.4 Improving Base Regression Models

4.4.1. Model Selection

	Model_Name	Alpha (Wherever Required)	l1-ratio	R-Squared	Adj. R-Squared	Train_MSE	Test_MSE
0	Ridge Regression (alpha = 2 w/o Outliers)	2	-	0.921161	0.920932	3057.212791	3402.058394
1	Ridge Regression (alpha = 1 w/o Outliers)	1	-	0.921207	0.920977	3055.460206	3403.355072
2	ElasticNet Regression (GridSearchCV w/o Outliers)	0.000100	0.200000	0.921215	0.920986	3055.120552	3403.941905
3	Lasso Regression (w/o Outliers)	0.01	-	0.921209	0.920980	3055.361814	3404.595659
4	Ridge Regression (GridSearchCV w/o Outliers)	0.100000	-	0.921226	0.920996	3054.726546	3405.674397
5	Lasso Regression (GridSearchCV w/o Outliers)	0.100000	-	0.921226	0.920996	3054.724168	3405.832117
6	Linear Regression (w/o Outliers)	-	-	0.921226	0.920996	3054.717727	3406.024609
7	Decision Tree Regression (GridSearchCV w/o Outliers)	-	-	0.943764	0.943600	2180.736668	3437.121467
8	ElasticNet Regression (w/o Outliers)	0.1	0.01	0.909729	0.909466	3500.552609	3759.539320
9	Decision Tree Regression (GridSearchCV)	-	-	0.910975	0.910746	4610.063310	7201.719312
10	RandomForestRegressor (GridSearchCV)	-	-	0.911472	0.911244	4584.353997	7228.327757
11	Ridge Regression (alpha = 2)	2	-	0.877590	0.877275	6338.913356	7634.481186
12	Ridge Regression (alpha = 1)	1	-	0.877615	0.877300	6337.576732	7636.944111
13	ElasticNet Regression (GridSearchCV)	0.000100	0.200000	0.877619	0.877304	6337.398888	7637.557368
14	Lasso Regression (GridSearchCV)	0.100000	-	0.877619	0.877304	6337.394995	7637.819791
15	Lasso Regression	0.01	-	0.877619	0.877304	6337.394995	7637.819791
16	Ridge Regression (GridSearchCV)	0.100000	-	0.877626	0.877311	6337.025613	7640.436081
17	Linear Regression	-	-	0.877626	0.877311	6337.019066	7640.922190
18	ElasticNet Regression	0.1	0.01	0.868150	0.867810	6827.753459	8095.864889

fig. 4.4.1.1: Comparison of models through metrics – Selection 1

Algorithm	Encoding	Test RMSE
Base Model	(Irrespective)	92908
KNN Regressor (on original data)	One hot Encoding	146.3
KNN Regressor (on original data)	Frequency Encoding	155.2
Decision Tree Regressor (on original data)	One hot Encoding	139.8
Decision Tree Regressor (on original data)	Frequency Encoding	133.7
Decision Tree Regressor (with outliers)	One hot encoding	84.8
Random Forest Regressor (with outliers)	One hot encoding	84.9
Ridge Regressor (with outliers)	One hot encoding	87.3
Lasso Regressor and elastic net Regressor (with outliers)	One hot encoding	87.4
Ridge Regression(w/o outliers)	One hot encoding	58.3
ElasticNet Regression (w/o outliers)	One hot encoding	58.4
Lasso Regression (w/o outliers)	One hot encoding	58.4
Linear Regression (w/o outliers)	One hot encoding	58.3
Decision Tree Regression (w/o outliers)	One hot encoding	58.6
Decision Tree Regression (w/o outliers)	frequency encoding	56.3
KNN Regressor (w/o outliers)	One hot encoding	63.6
KNN Regressor (w/o outliers)	frequency encoding	73.6

fig. 4.4.1.2: Comparison of models through metrics – Selection 2

In the view of possible bias and sampling precision which are the two critical factors affecting the underlying reliability of evaluation results, the procedure that was taken in the above analysis to address these is: sampling precision can be ruled out in this case study as the dataset has the total population (participants with sufficient billing history) included for analysis and no samples were taken. In the view of bias, this will be very possible because participants without sufficient consumption history were excluded in the analysis. Also, the 'external influences' like weather, home occupancy patterns can create change in energy usage and affect the results of billing analysis. Since we are going with fixed effects regression model, external influences are not included in the feature set and so bias is very much possible.

In the view of the bias that's possible in the dataset, we have developed three datasets i.e.,

- Original data
- With a few extreme outliers removed
- With all possible outliers removed

We have done modelling on all these three datasets to understand the performance which in turn helps in understanding realization rates. Also, we have developed two feature sets one with one hot encoded categorical variable and another with frequency encoded categorical variables as tree-based models perform better with frequency encoded features.

The above table shows the results with all three datasets and different encoding techniques.

1. It is clear that Decision Tree Regressor on dataset with all outliers removed and frequency encoded performed better with 56.3 RMSE value.

2. Also, the rmse values of the three datasets overall are in the order without outliers > with partial outliers > original data in the order of performance.

3. Therefore, it is understandable since all simple to complex models on original data (with whole population, bias(attrition)) haven't performed better than that on other two datasets. indicating the modelling software shows a lower

realization rate in reality with whole population involved due to possible bias (-> attrition and external influences) in original data.

4. From the modelling analysis, it is also clear that 'Estimated MMBtu Savings', 'Total Project Cost' and 'Region' emerged important features which hasn't changed much over the years from 2007-08 to 2018 and very much in line with 2007-08 impact analysis though the realization rates have improved.

Overall, when the bias in the dataset can be addressed to some extent (by removing outliers etc.), then the results the modelling will give on any real-world input for the 'estimated 1st year annual savings' will be a better estimate with better realization rate in real-time.

5. Conclusions

5.1 Attrition:

Due to the uniqueness of the participants. The evaluation of the program can bring a lot of invariability where we would not be able to figure out the outcomes properly for the counties separately. These features can produce attrition in not sorted.

- Billing Month
- Evaluation Job type
- Type of Dwelling
- Measure Type

5.2 Improvements:

Region – West and central regions can be used to understand region-based program performance estimates.

Billing Month – Month of May, April and September showing a possibility of future attrition due to a big difference in electric usage.

Electric Utility – National Grid, Rochester Gas & Electric and Consolidated Edison are the most used Electric Utility by the participants in the State of New York.

Size of Home – Size of Home does not have an influence on the performance or the cost of project.

5.3 Feature Importance

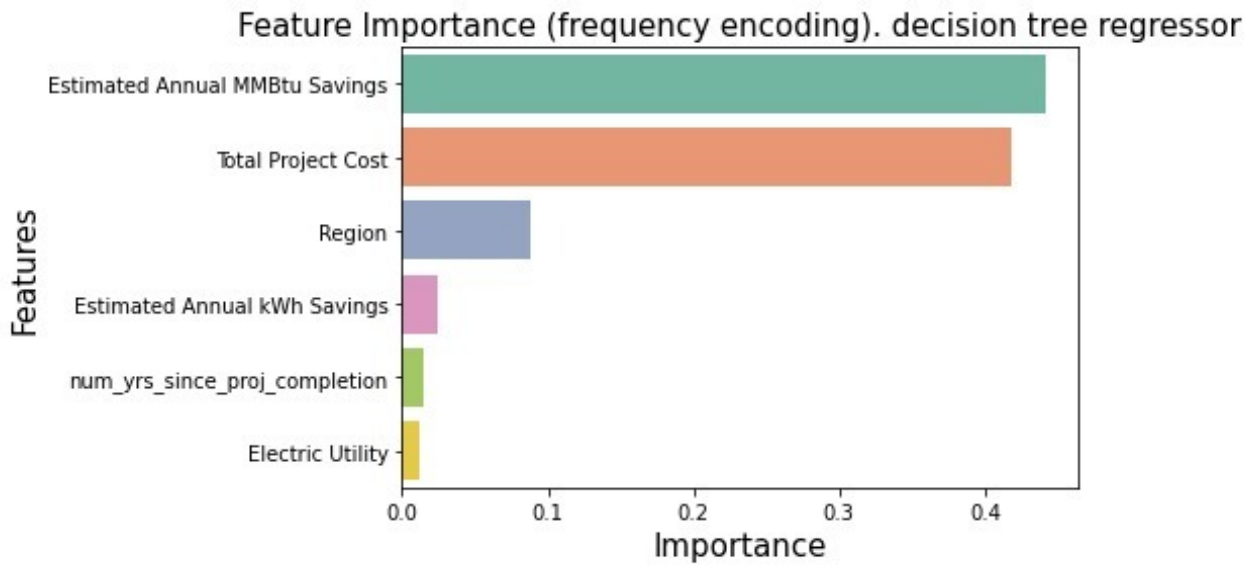


fig. 5.1.: Feature Importance from Decision Tree Regressor with frequency encoding

Feature importance shows us the average gain of variables with the frequency encoded Decision Tree Regressor. Here, we can see the following Features which are important for the analysis. The ranking is as follows:

- Estimated Annual MMBtu Savings
- Region
- Estimated Annual kWh Savings
- Electric Utility
- Total Project Cost

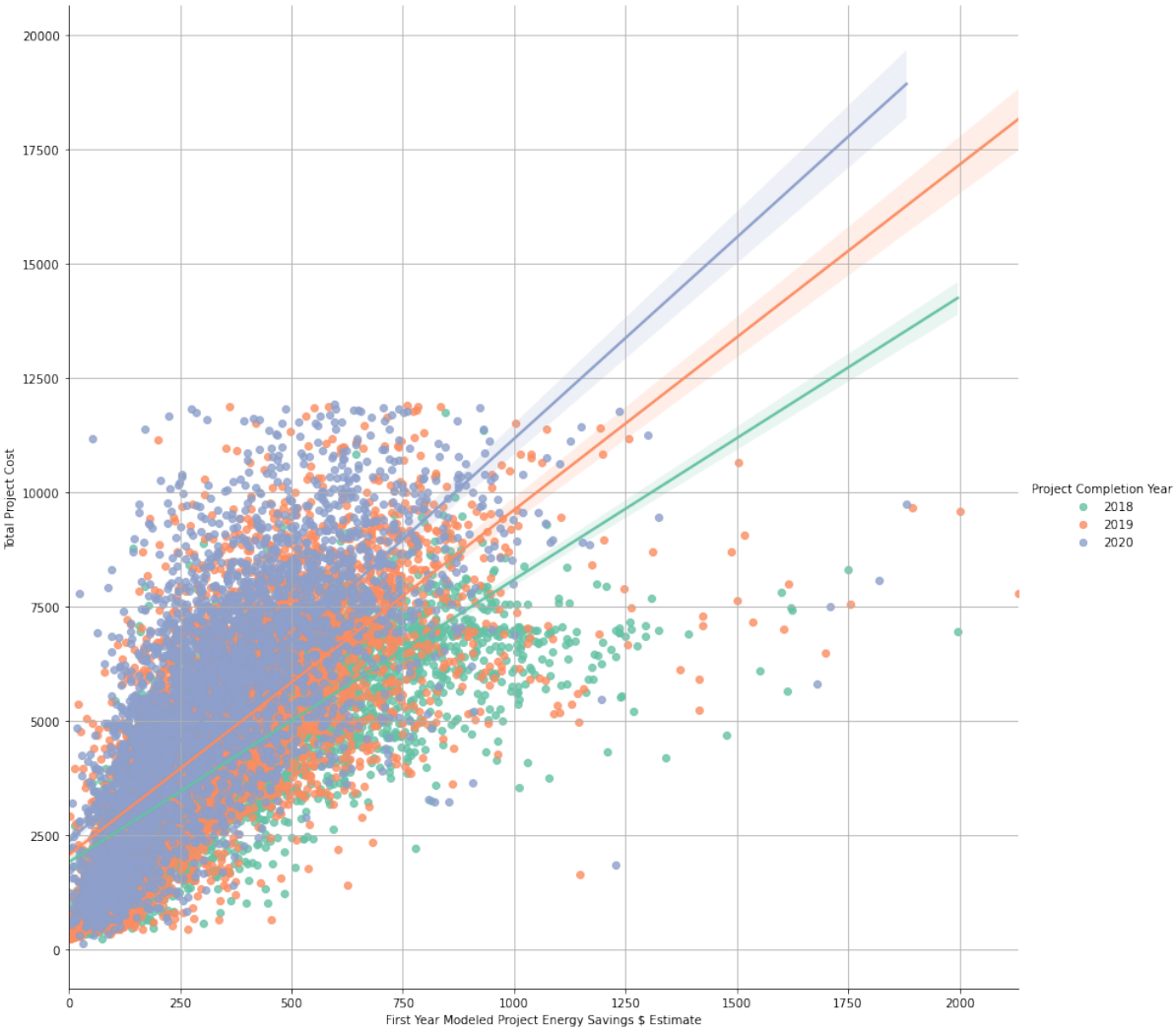


fig. 5.2.: scatter plot showing Total Project Cost by money saved in terms of Years

- Total Project Cost is subjected to increase each year.

- With the increase Project Cost there is also an increase in the First Year Modeled Project Energy Savings $ Estimate.

- Hence, every year the Project Cost will increase with the savings estimate. This is the expected trend which is to be followed.

References

- A Guide to Energy Efficient Concepts for New Residential Construction – Marilyn Kaplan, Christopher Sgroi, Priscilla Richards
- High Performance Residential Design Challenge – Gregory Pederick
- New Efficiency: New York – NYSERDA Department of Public Service
- Home Performance with ENERGY STAR® Program Impact Evaluation Report (2007-2008) – Carley Murray, Judeen Byrne
- Document number Residential Solar Energy Storage Analysis – NYSERDA, DNV KEMA Team
- Clean Energy Technologies Innovation Metrics 2012 – Jacques Roeth, Jennifer Ozawa, Nancy Chan
- 2020 Energy Conservation Code – New York City Energy Conservation Code Advisory Committee
- 2019 New York Getting to Zero Status Report, Status Report on Net Zero Energy and High-Performance Buildings – Janet Joseph
- Increasing Efficiency of Building Systems and Technologies 2015 – NYSERDA Quadrennial Technology Review
- New York City's 2020 Energy and Water Use Report – Urban Green Council, The Office of The City of New York Mayor Bill de Blasio