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

Capstone Project – Interim Report

Submitted By:

Tanuj Chauhan
Aravind Mohan
Saili Kotturu
Ganugula Satya Ravi Teja

Table of Contents

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.

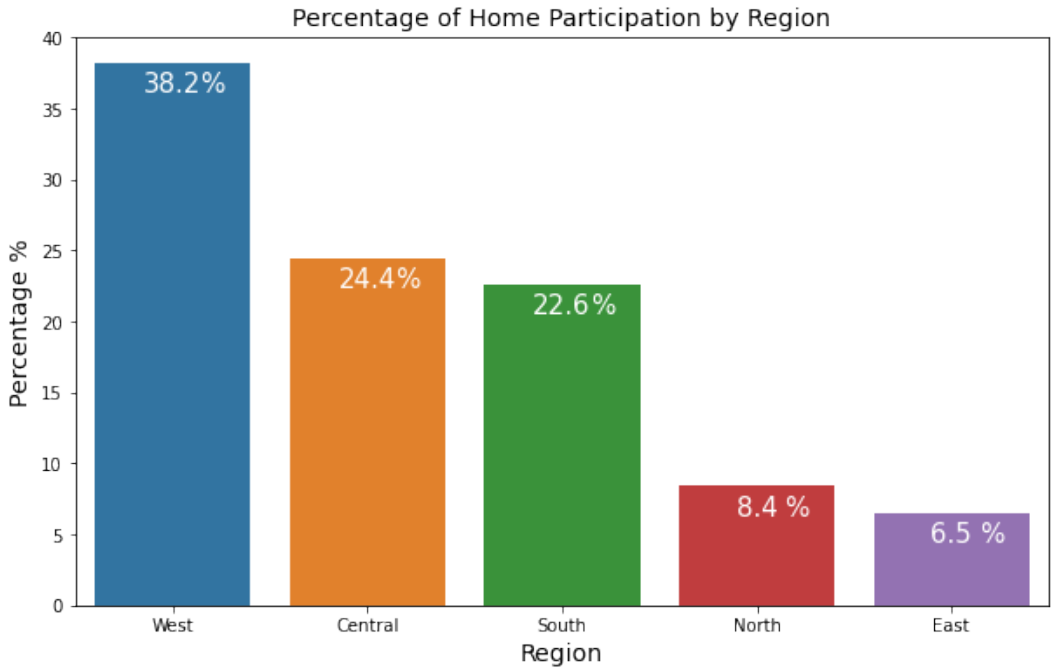


fig. 1.1: Percentage of home participation in the program by region

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

2.2 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 Dwelling', 'Measure Type', 'Location 1'	14

2.3 Pre-Processing Data Analysis

2.3.1 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%

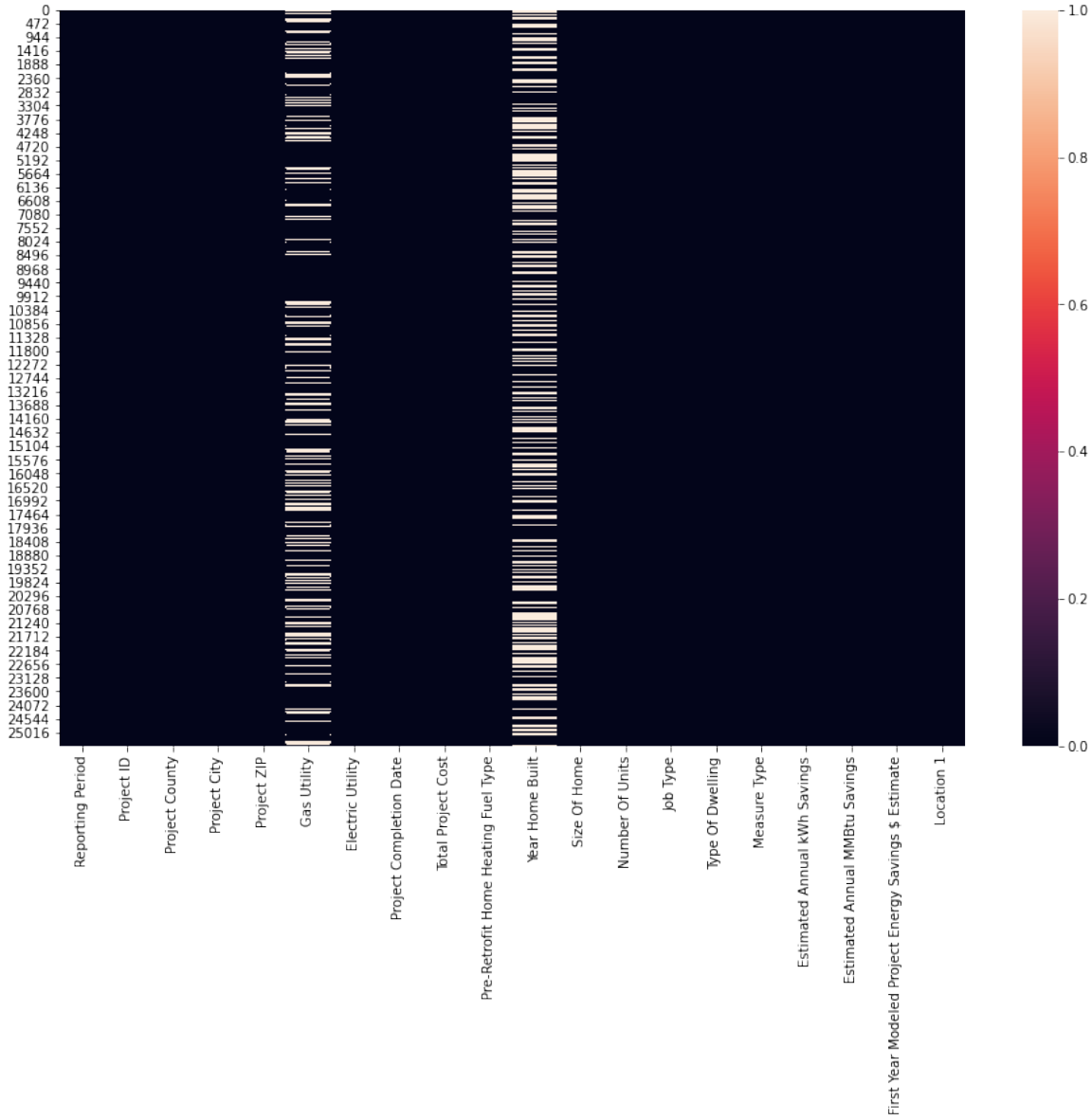


fig. 2.1: Heatmap visualization of missing values in the dataset

2.3.2 Feature dType Correction

Checking for DataFrame information showed:

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

2.3.3 Creating New Features

New features created from existing features:

- Project Completion Year
- Latitude
- Longitude
- Type of Home Size
- Region

2.3.4 Removing Unnecessary Features

Following features were removed for being redundant:

- Reporting Period
- Project ID
- Project ZIP

2.4 Project Justification

2.4.1 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.

2.4.2 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.

2.4.3 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.

3. Exploratory Data Analysis

3.1 Variable Relationships

3.1.1 Correlation Analysis

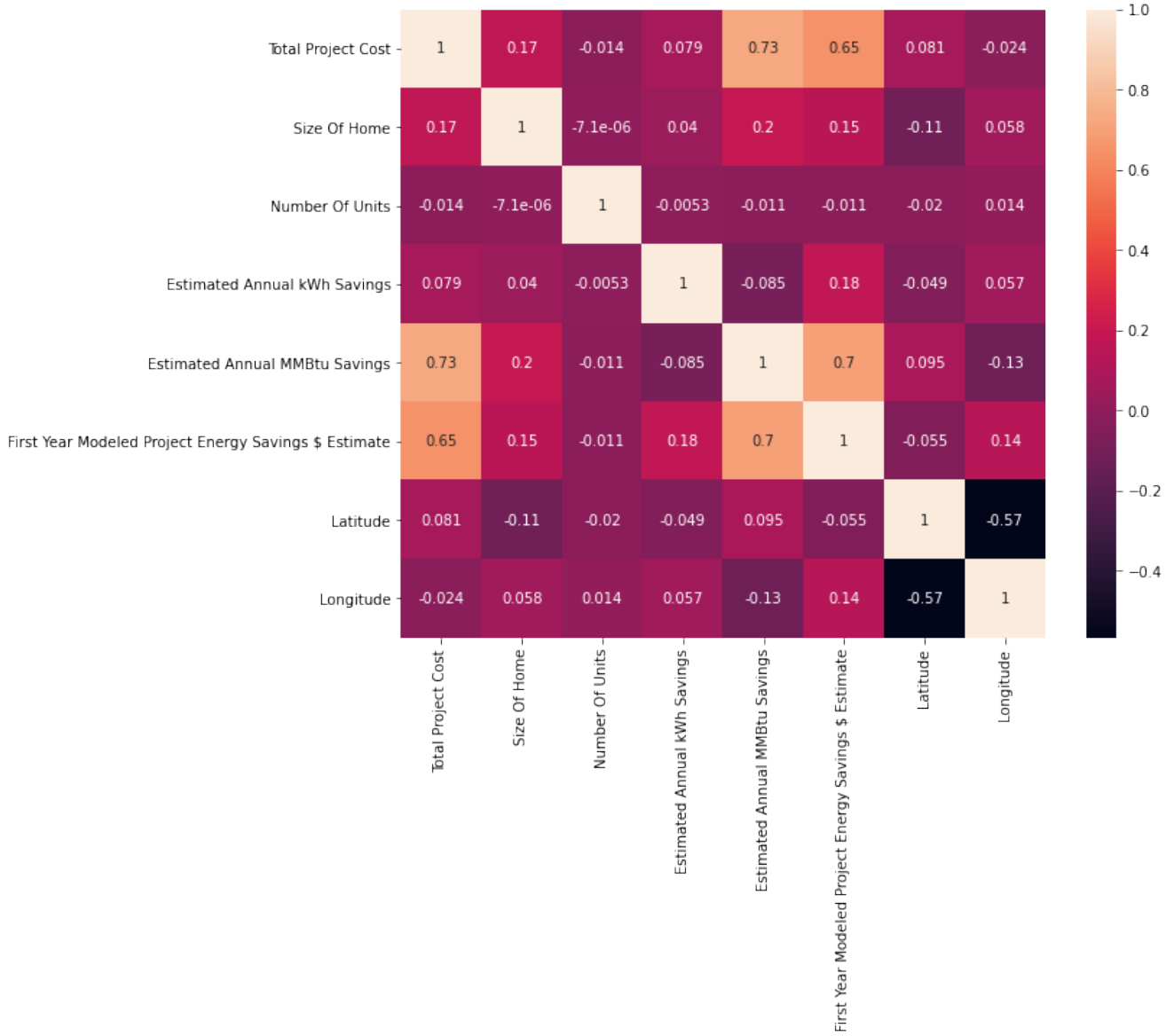


fig. 3.1: Heatmap visualization of correlation between the numerical features

- '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 Univariate Analysis

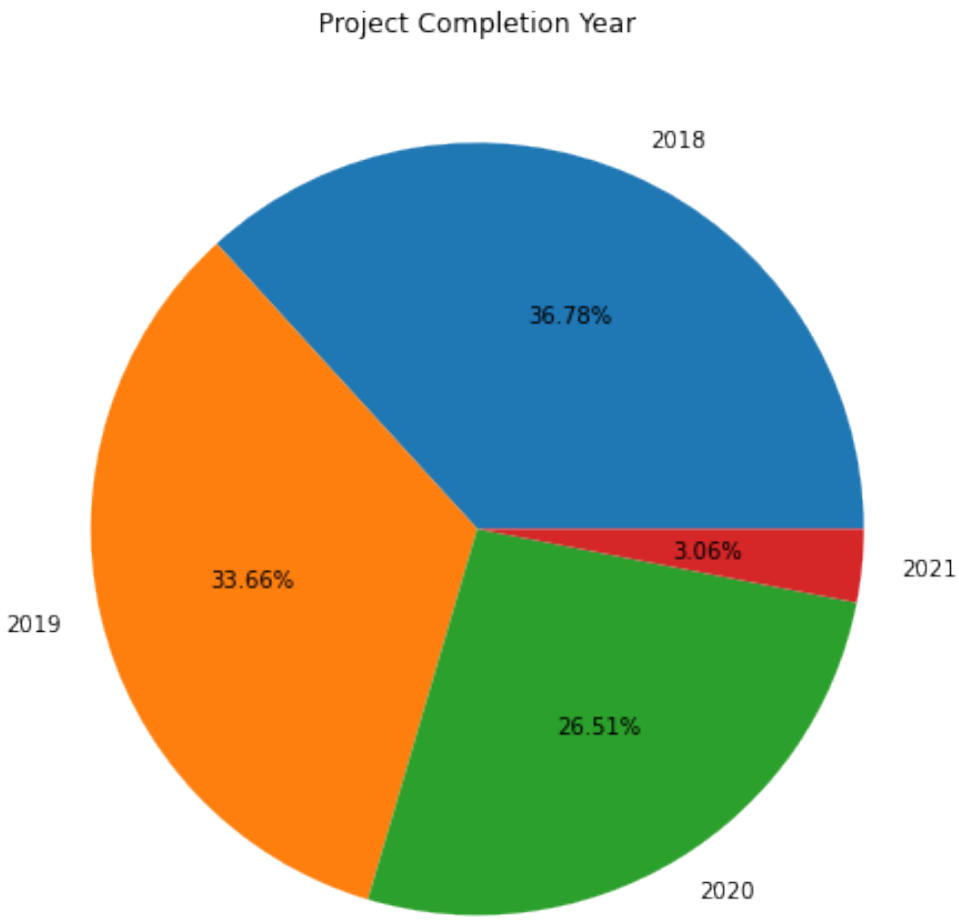


fig. 3.2: Pie Chart Showing the percentages of Projects Completed within what years

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

3.1.3 Bivariate Analysis

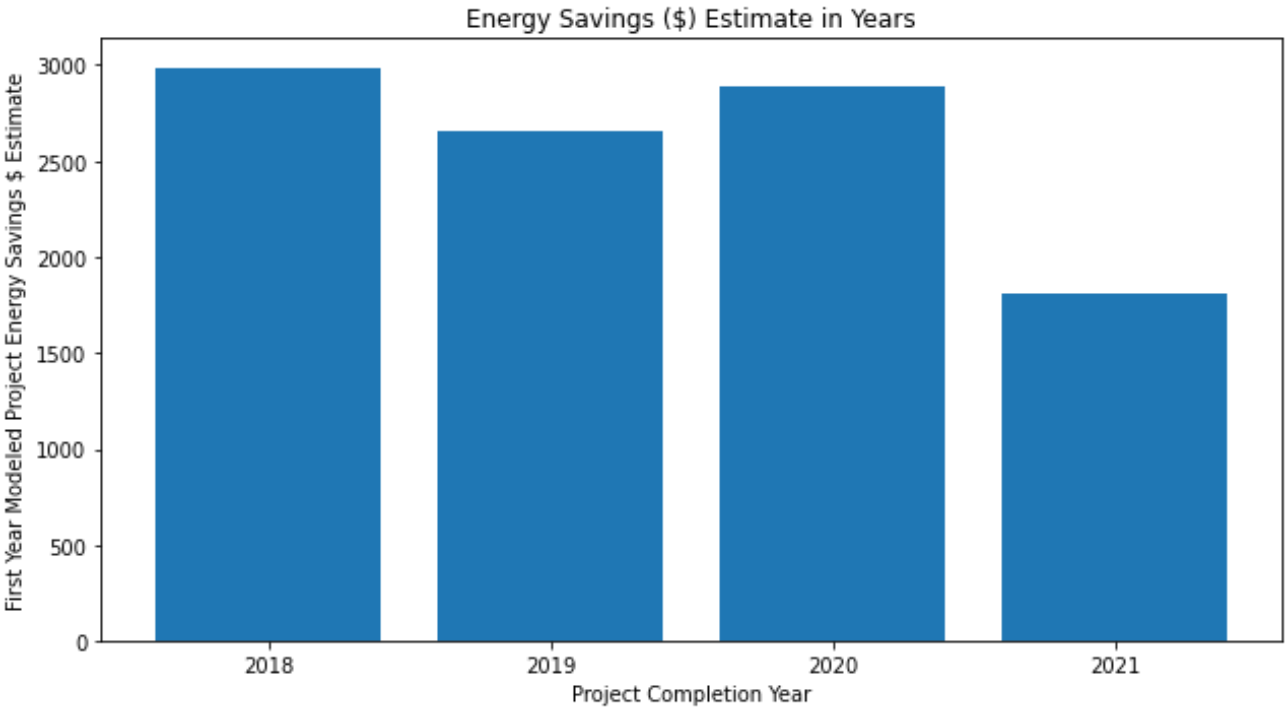


fig. 3.3: Energy savings estimation in dollars ($) through years

- We can see the projects completed in 2020 had the most economical impact of all the years.
- Projects done in 2020 had a better economic performance than the year 2018 even though the number of projects completed in 2018 were the highest.
- Projects completed in 2021 have performed the least economically, stating to the fact that most of the projects are still on-going.

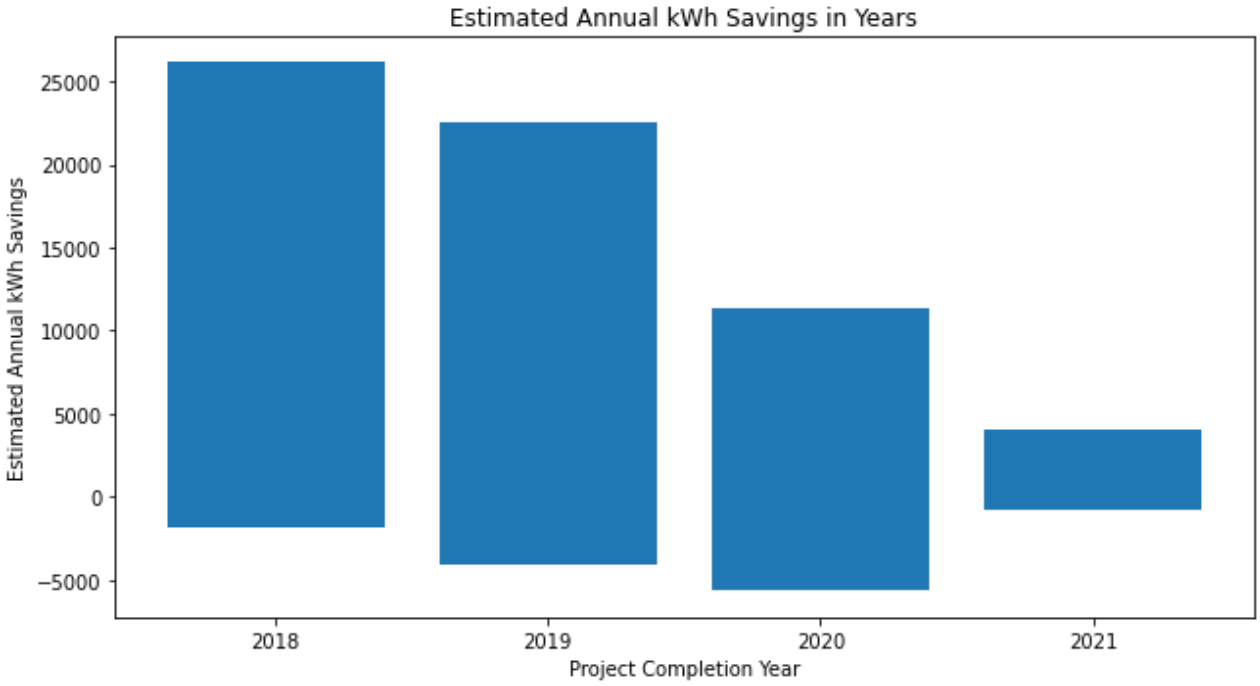


fig. 3.4: Annul energy savings estimation in energy units (kWh) through years

- The projects completed in the year 2018 had the most impact over energy savings and the projects will be deemed the most energy efficient of all.
- The projects completed in the year 2019 are trailing behind the projects completed in 2018 in terms of energy efficiency and it is also having more energy consumption too. Please note that, the negative energy savings noted through the billing cycles can be attributed to energy consumption during different parts of the months, for example: In 2 or more families with kids, more energy is consumed during vacations.
- The projects completed in 2021 have faired pretty well in terms of negative energy savings even though the billing cycle was conducted during lockdowns which happened due to the global effect Covid-19, rendering families to consume more energy.

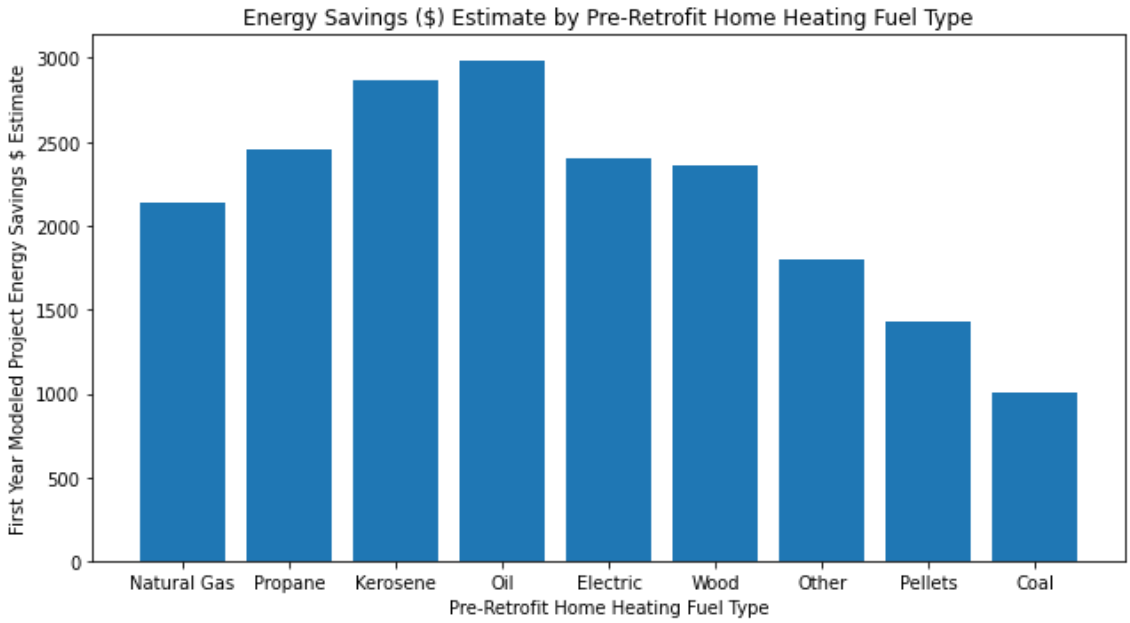


fig. 3.5: Energy savings estimations in dollars ($) by Pre-Retrofitting Heating Fuel Type

- The homes with Oil as the type of pre-retrofit home heating fuel were certainly fared better through the program as the economical savings have been the highest after retrofitting to a more energy efficient method.
- Kerosene Type is trailing very little behind Oil Type.
- Whereas, Coal Type has the worst performance of all.

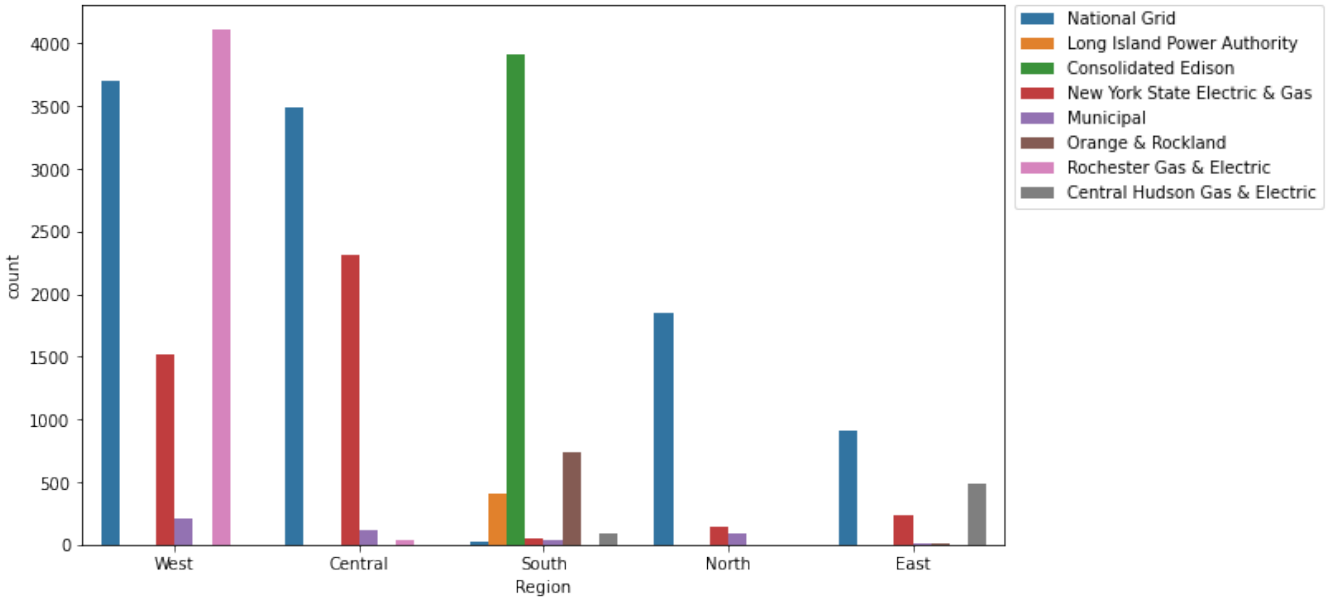


fig. 3.6: Count of home participation in the program by Region based on their type Electric Utility used

- In West, the Rochester Gas & Electric is the most used energy providing service.
- In Central New York, it's the National Grid.
- In the Southern region, Consolidated Edison energy company has the most homes covered.
- The National Grid is again the most used energy provider in the North and East region.

3.1.4 Multivariate Analysis

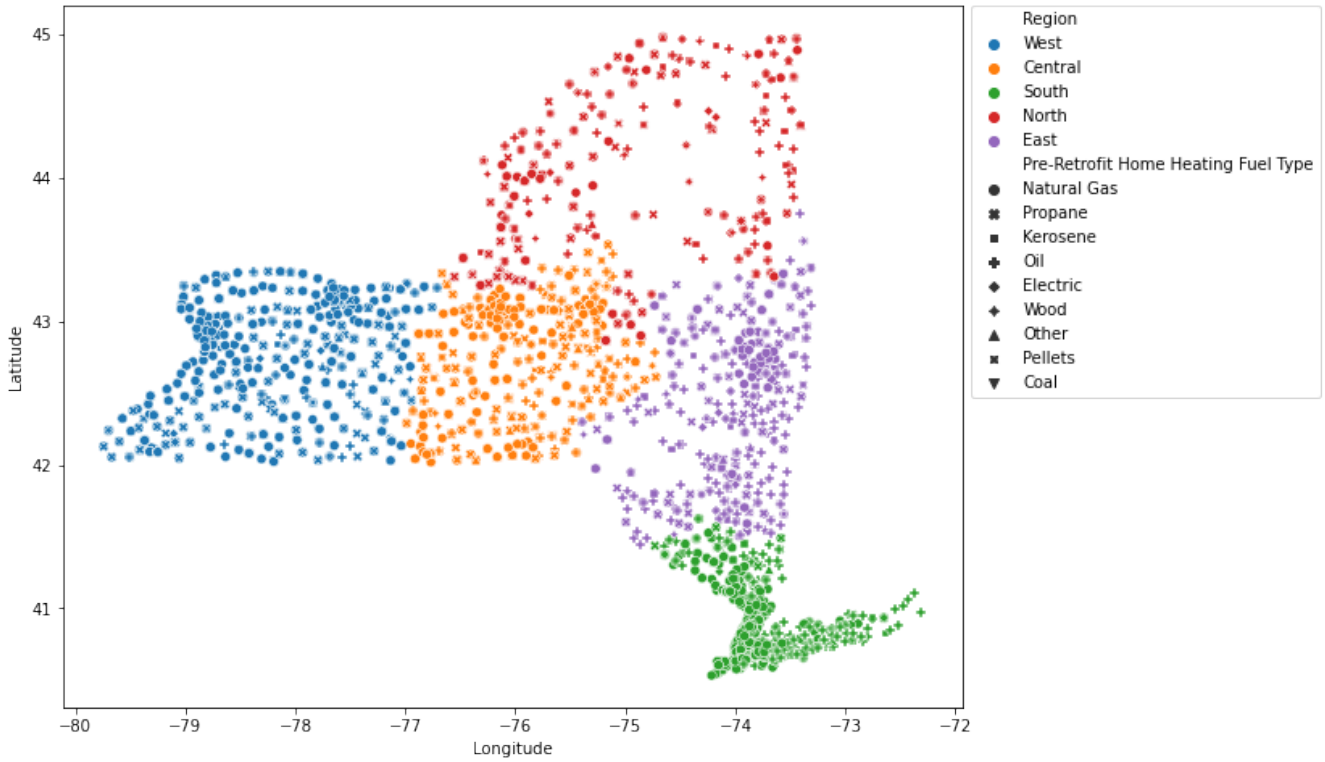
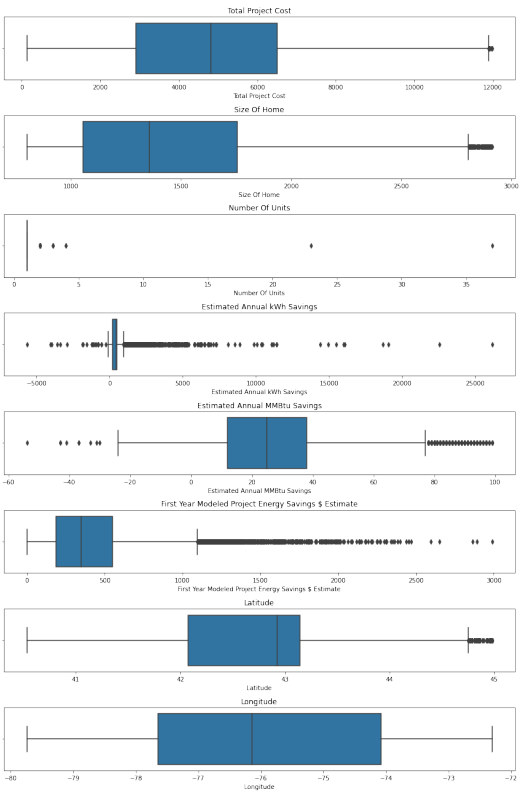


fig. 3.7: Scatter plotting Latitude and Longitude to visualize home participants based on Region and the Pre-Retrofit Home Heating Fuel Type

- Natural Gas is the most used pre-retrofitting home heating fuel type in the state.
- Western region has the most consolidated use of the program.
- Southern Region has the most densely populated projects completed or under works.

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

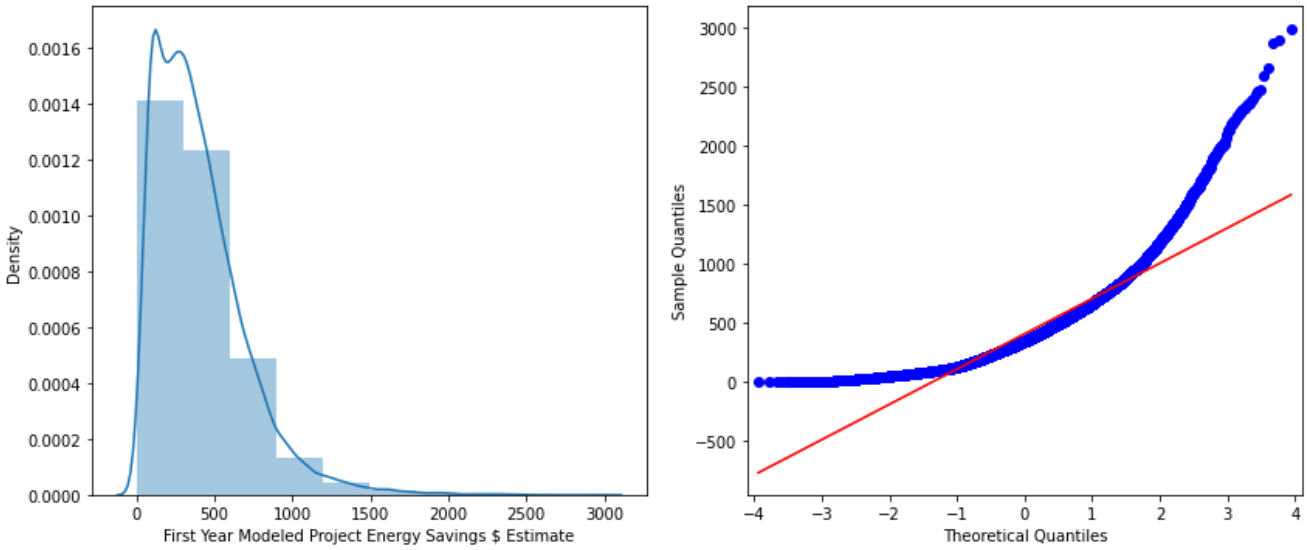


fig. 3.1: Distribution plot for First Year Modeled Project Energy Savings $ Estimate

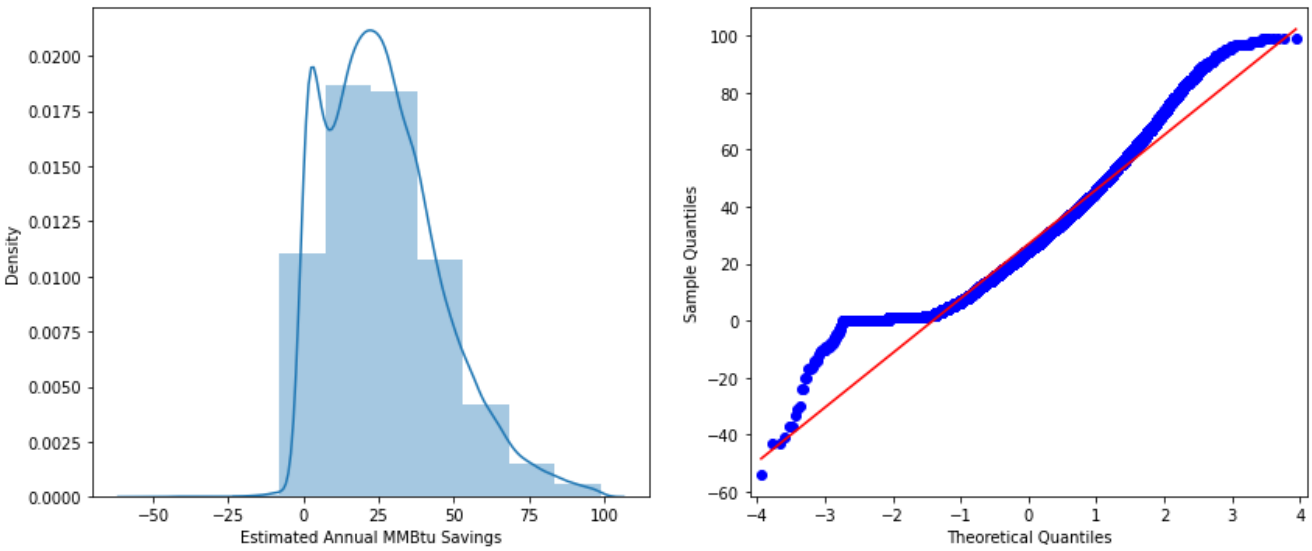


fig. 3.2: Distribution plot for Estimated Annual MMBtu Savings

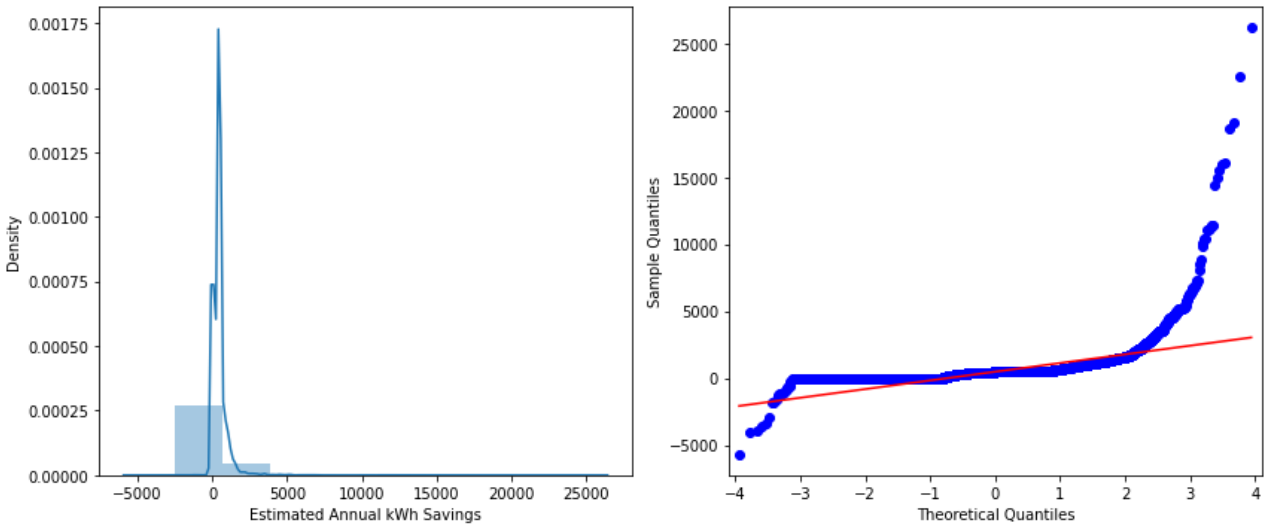


fig. 3.3: Distribution plot for Estimated Annual kWh Savings

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 Jarque-Bera tests.

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 Model

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
Pre-Retrofit Home Heating Fuel Type_Natural Gas	0.148
Electric Utility_Orange & Rockland	0.097
Measure Type_Heating and Cooling	0.067

5. Conclusions

5.1 Testing for Assumptions for OLS Base Model

5.1.1 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 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.

	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

5.1.2 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.

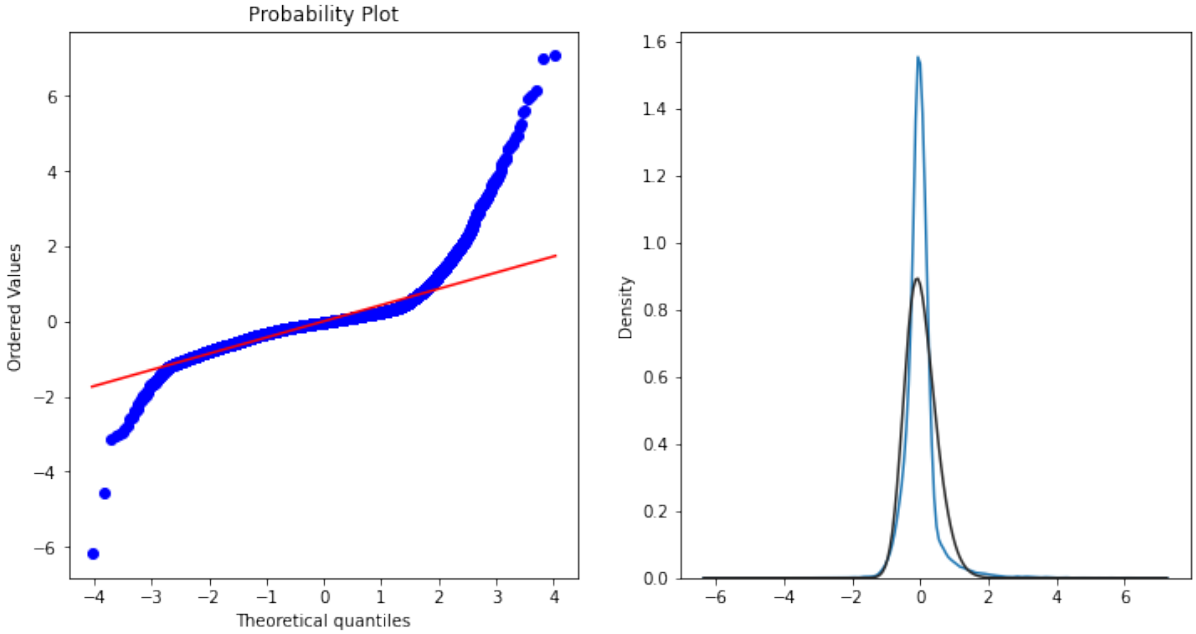


fig. 5.1: Q-Q plot and the distribution plot for the residuals.

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.

5.1.3 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.

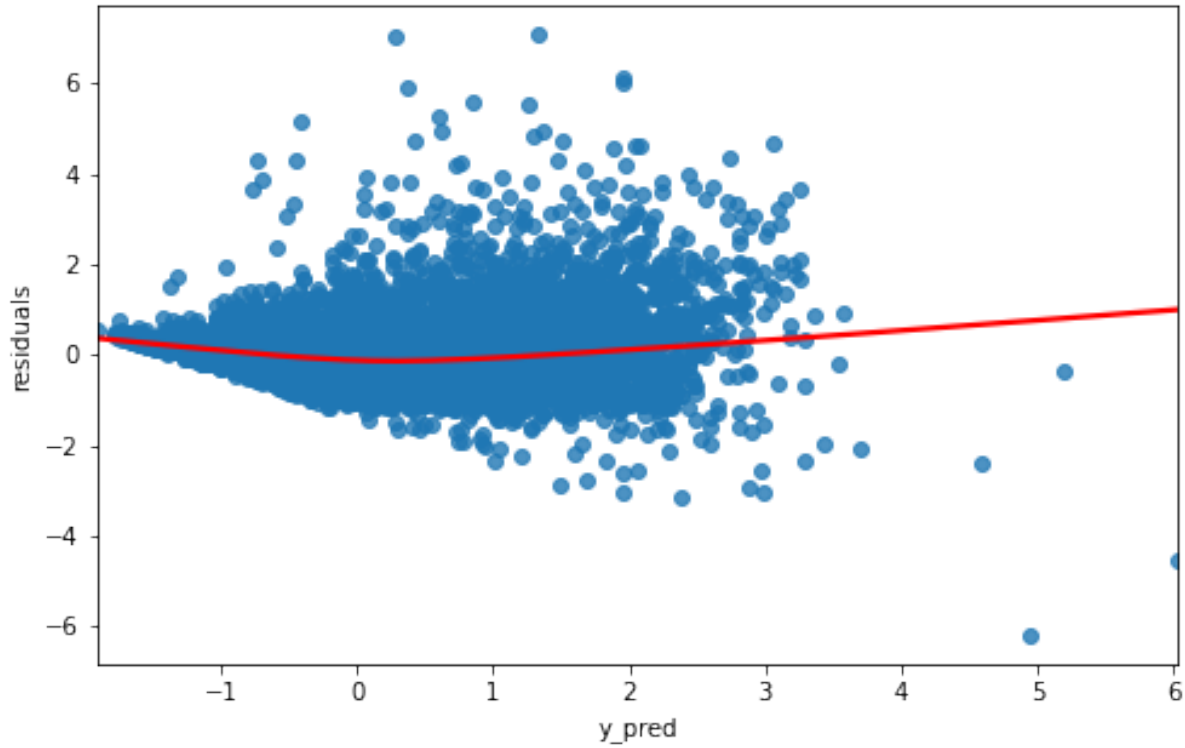


fig. 5.2: Scatter plot for the predicted target variable vs the residuals to check for homoscedasticity.

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.

5.1.3.1 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 are not constant.

5.1.4 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.

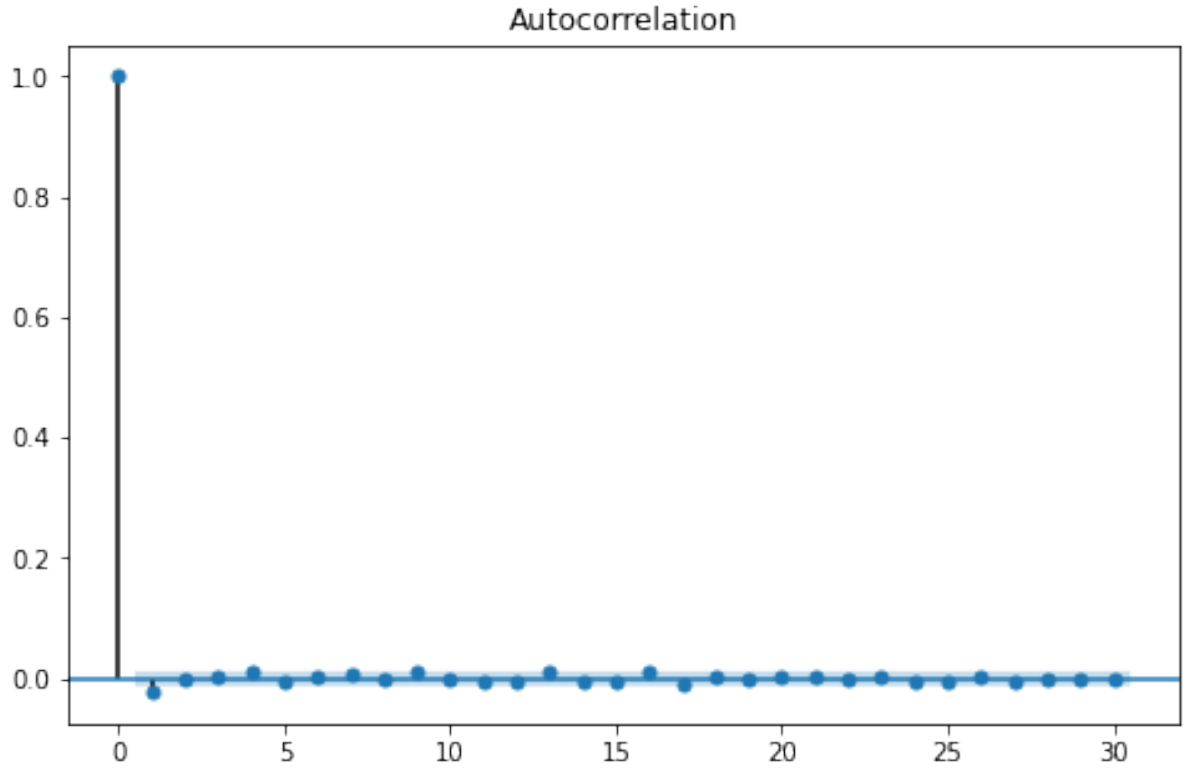


fig. 5.3: acf_plot for checking Auto-Correlation through the residuals

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.

5.1.5 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.

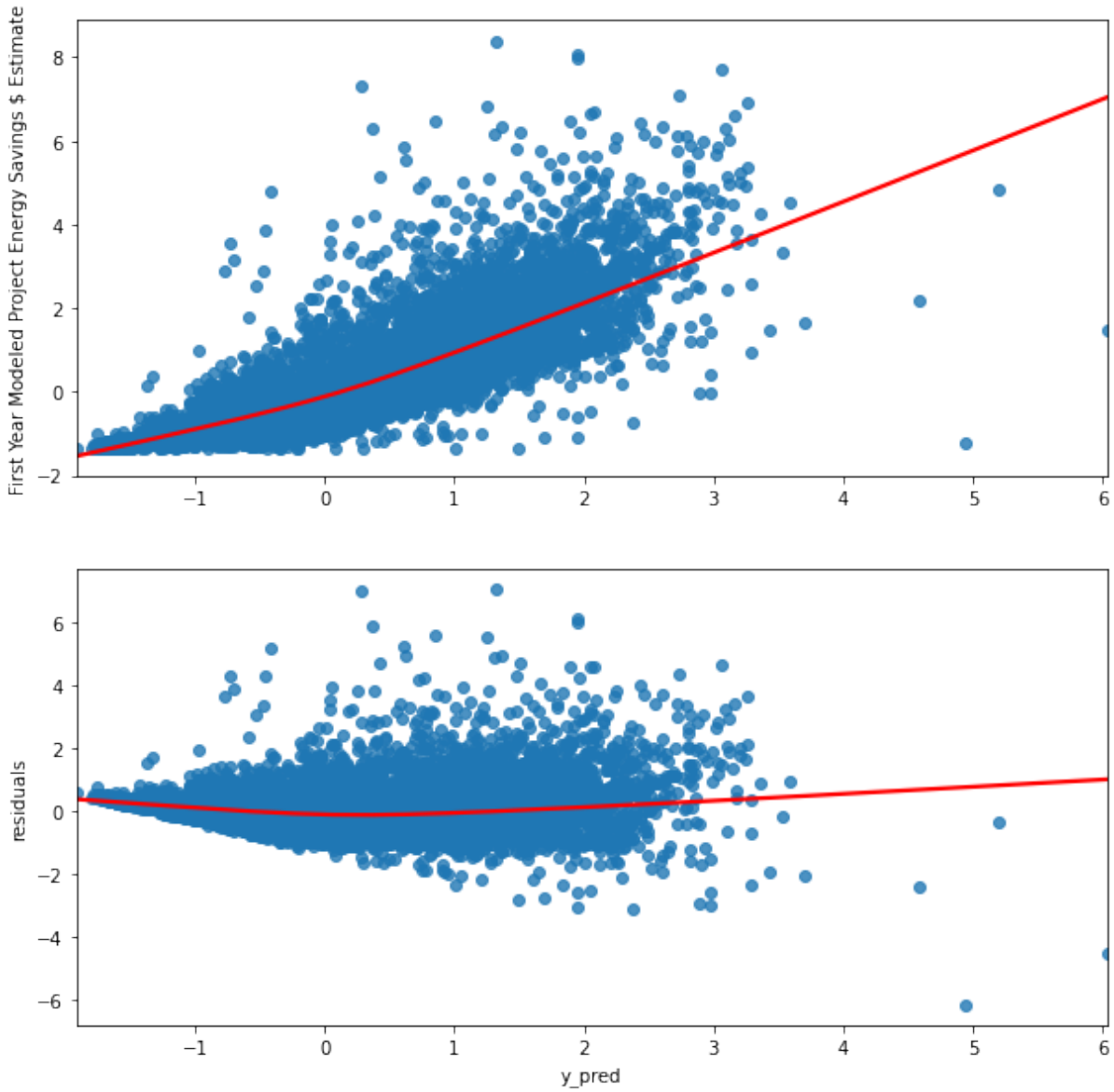


fig. 5.4: Scatter plot for checking the linear relationship y and y_pred.

5.1.5.1 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.

5.2 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.

6. Further Approach

6.1 Further Feature Engineering

Trying to achieve better modelling with more feature engineering by creating new features like seasons to give better analysis over the usage of electricity and heat. Will give more inferences over power consumptions as the participants might have a variable usage over the course of the year. This would also help us to understand more about the attrition in the dataset (refer 1.3).

6.2 Further Exploratory Data Analysis

With better feature engineering, we could achieve better plots which could help us in a better and more robust comparative study by using different types of plots over new features.

6.3 Improving Regression Models

Other than the OLS Model we can make better models by other Regressor Techniques like K-NN (K-Nearest Neighbors) and Tree Regressors like Random Forrest and Decision Tree Regressors. We can then compare the results through RMSE (root mean square error) or R Squared significance. As the models will pose a problem of over-fitting, we can use regularization techniques to prevent it. Lasso and Ridge regularization techniques are widely used for linear regression models.

We can then try to get the best parameters for the model that works the best for us and tune the parameters it to achieve an ideal or close to ideal model to attain probable inferences which could help us understand the energy consumption after the program and to understand how this program could further help the state of New York and other states in optimizing energy efficiency.

END OF INTERIM REPORT