DATA 603: FINAL REPORT PROJECT TITLE: MULTIPLE REGRESSION ANALYSIS OF ENERGY CONSUMPTION IN THE CITY OF CALGARY

Group 13: CODE 404

Team members:

1.Khushi Himanshu Dave

2. Shashank Kumar Srivastava

3. Zheyu Song

4. Jannatul Naeema

INTRODUCTION:

1.1. Motivation

The domain that we will be working with for this project is Energy and sustainability of Calgary. Energy plays a fundamental role in our lives; everything requires energy in one form or another. Canada is in the top five of natural gas producers in the world; two-thirds of which come from Alberta. In 2017, the energy sector made up 9.2%, or $175 Billion, of Canada's Gross Domestic Product (GDP) whereas, in Alberta the energy sector contributed 21.61% of provincial GDP. This is significantly more than in the rest of Canada; the oil and gas sector make up a major part of economic activity in Alberta. As one of the major cities in Alberta, Calgary has long been known as an energy city and took on a lot of initiatives to encourage in becoming more energy efficient in the long run. In 2008, the City of Calgary developed the Sustainable Buildings Partnership Program to improve the performance of existing city infrastructure and support the sustainable building policy. The purpose of this program is to identify and improve the efficiency of existing corporate infrastructure. These is proposed to be done using audits, alternative energy technologies, conservation, and energy efficiency upgrades. We focused on addressing this context and investigate into energy consumption situation at City of Calgary.

1.2.Objectives

In our project we would like to analyze the energy consumption situations at different structure and facilities at City of Calgary. The goal of this project is to predict future energy use for the buildings and investigate the effects of different variables. We will perform multiple linear regression and we will use R studio to analyze the topic. This study is important to assess if the energy use of buildings and structures are in aligned with the sustainable building policy. Through this investigation, we aim to understand better energy efficiency and we aim to provide new insights as to whether the energy efficiency need to be improved.

METHODOLOGY:

2.1 Dataset Dataset 1: Building Energy Benchmarking – City of Calgary The first dataset we are using for this project is Building Energy Benchmarking Data from the City of Calgary webpage.

This is a open dataset and available for public use; the reference of dataset is included in the reference section of this report. This dataset is open data in tabular format collected annually, collected over period 2019-2021 with 297 rows and 23 columns. The City of Calgary's Commercial and Institutional Building Energy Benchmarking Program facilitates in measuring and tracking the energy performance of any commercial, institutional or non-profit organization with a building of any size that is located within the city boundary. This dataset contains building energy and greenhouse gas emission performance information for a subset of properties owned and operated by the City of Calgary. All energy and greenhouse gas emission metrics are calculated by ENERGY STAR Portfolio Manager using monthly, whole-building energy consumption data billed between January 1st and December 31st. This initiative is significant in obtaining standardized information on building energy consumption, energy costs and greenhouse gas emissions and also assisting in becoming eligible for ENERGY Star® Certification. This dataset is collected from City of Calgary open data website, available for public use. In this dataset, there are 17 different types of property built from 1896 to 2018 including fire station, ice rink, office, recreation center, heated swimming pool etc. Dataset 2: Current and Historical Alberta Weather Station Data - ACIS The weather dataset was collected by the Alberta Climate Information Service (ACIS) from a variety of meteorological stations operated by various government agencies. For the investigation of this dataset, we will be focusing on the variables: Date, Air Temp. Avg. (°C), Relative Humidity Avg. (%), Precip. Accumulated (mm), Wind Speed 10 m Avg. (km/h) for our predictive model designing and interpretation, and all the variables are quantitative. The dataset recorded daily temperature from 2014 to 2022 at Calgary International CS weather station. We will merge dataset 1 and dataset 2; we will take all variables from dataset 1, and from dataset 2 we will add precipitation, and calculate and then add the columns of Corn Heat Unit, Heating Degree Days, Cooling Degree Days. We will use these four columns from dataset 2 along with 20 columns from dataset 1 to build full model for energy use.

2.2 Approach

We imported the data in RStudio for our analysis by first using the "read.csv" and created the 'dataset' data frame.

The data we needed for our regression analysis included the dependent variable Site Energy as a function of twenty independent variables, we will build full model first with all these variables. Then from this point forward we will try to improve our model. All variables in the dataset are as follows: 1. Site Energy: The annual amount of all the energy a property consumes on-site, regardless of the source, dependent numerical variable 2. Property type: We have 17 different types of properties for this dataset, each type of property has multiple buildings; among the types of properties there are 107 fire-stations, 66 office buildings, 30 ice rinks, 24 fitness centers and few other different kinds of properties at different locations. This is categorical independent variable. 3. Number of Buildings: This represents how many buildings are present at a certain property, numerical independent variable. 4. Year built: This indicates either at which year the property was constructed or at which year the most recent major renovation was done including a complete interior redesign, categorical independent variable. 5. Property GFA: This includes the total property gross floor area, numerical independent variable. 6. Energy Star Score: A measure of how well a property is performing relative to similar properties, when normalized for climate and operational characteristics. This is categorical independent variable. 7. Weather Normalized Site Energy Use (GJ): This indicates the energy use a property would have consumed during 30-year average weather conditions, numerical independent variable. 8. Site EUI (GJ/m²): The Site Energy Use divided by the property square meters, numerical variable, numerical independent variable. 9. Weather Normalized Site EUI (GJ/m²): The Weather Normalized Site Energy Use divided by the property square meters, numerical independent variable. 10. Source Energy Use (GJ): The total amount of all the raw fuel required to operate a property, including losses that take place during generation, transmission, and distribution of the energy, numerical independent variable. 11. Weather Normalized Source Energy Use (GJ): The source energy use your property would have consumed during 30-year average weather conditions, numerical independent variable. 12. Source EUI: The Source Energy Use divided by the property square meters, numerical independent variable. 13. Weather Normalized Source EUI (GJ/m²): The Weather Normalized Source Energy Use divided by the property square meters, numerical independent variable. 14. CO2 Emission: Total Emissions is the sum of Direct Emissions and Indirect Emissions, numerical independent variable. 15. CO2 Emissions Intensity: Total GHG Emissions divided by the property square meters, numerical independent variable. 16. Direct GHG Emissions: Direct Emissions of CO2 as greenhouse gas are emissions associated with onsite fuel combustion (e.g. combustion of natural gas or fuel oil), numerical independent variable. 17. Direct GHG Emissions Intensity: Direct GHG Emissions divided by the property square meters, numerical independent variable. 19. Natural Gas: Total annual Natural Gas used annually, numerical independent variable. 20. Electricity Use – Generated from Onsite Renewable System: The total amount of energy produced by onsite solar/wind), numerical independent variable. 21. Year Ending: The last day of the 12-month reporting period, numerical independent variable. 22. Precipitation: The recorded rainfall and snowfall, numerical value. 23. Corn Heat Unit: Temperature-based index often used by farmers and agricultural researchers to estimate whether the climate is warm enough (but not too hot) to grow corn, numerical. 24. Heating Degree Days: Heating Degree Days are equal to the number of degrees Celsius a given day's mean temperature is below 18 °C, numerical. 25. Cooling Degree Days: Cooling Degree Days (CDD) are equal to the number of degrees Celsius a given day's mean temperature is above 18 °C. For example, if the daily mean temperature is 21 °C, the CDD value for that day is equal to 3 °C. If the daily mean temperature is below 18 °C, the CDD value for that day is set to zero, numerical.

MODELLING PLAN:

We intend to create a multiple regression model for forecasting future Site energy use by using the procedures we have learned from the Data 603 course materials. To check for multicollinearity, we will first construct a first-order model using all the predictor variables and use variance inflation factors (VIF). Then, using a screening method called stepwise regression, we shall ascertain which independent variables in the list are the key determinants of Y. To verify the outcome, we will now check the test statistics (t-test) for each of the various coefficients. Additionally, an all-possible-regressions selection technique will be used to choose the "best" regression model. After removing some variables from the model that are not crucial, the model will be improved by considering interaction terms and/or a high order multiple regression model. We shall perform the following steps in our model to reach the final conclusion and the best model: 1. Creating the full model 2. Using VIF technique and Multicollinearity and possible elimination of correlated variables. 3. Stepwise Regression 4. Global F- Test 5. Individual T-Test , with alpha value = 0.05. 6. Interaction models, and testing for significant variables and repeating the process till best model is found. 7. Decide the best model based on interactions 8. Higher order model is found. 9. Test assumptions: Linearity, Independence, Equal Variance, Normality and Outlier assumptions 10. Based on the output, we will decide whether to do Box-cox transformations or not.

On following these steps, we will achieve our best model for Site Energy prediction. Justification for using these methods: To find the best model from any given Full model, when we follow these steps, we end up with our Best model. These steps are conducted to be very sure of our new reduced Best Model.

We will have our Outcome Variable: Site_Energy, and Predictor Variables: NumberofBuildings, Emissions_CO2, Natural_Gas, Property_GFA in our best mode, as observed from the conclusion. We choose Site_Energy as our outcome variable as its value depends on all the other variables.

For the division of workload: Khushi Himanshu Dave: VIF and elimination,Interaction models, step wise regression, Normality assumption, Box-cox Transformation, Theory of project. Zheyu Song: F and T-tests, higher order model, Equal Variance assumption, Box-cox Transformation , Final model and interpretation, Theory of project. Shashank Kumar Srivastava: Interaction models, Linear assumptions, Normality assumption, Box-cox Transformation, Final model and interpretation, Theory of project. Jannatul Naeema: worked on Full model, VIF and elimination, higher order model, Equal Variance assumption, Box-cox Transformation, Theory of project.

WORKING ON OUR DATASET:

```
dataset = read.csv('cleaned603dataset-dataset603_1.csv')
head(dataset)
```

	PropertyID <int>	Property_Type <chr>	NumberofBuildings <int>	Year_Built <int>	Property_GFA <dbl>	Site_Energy <dbl>
1	6169481	Office	1	1981	7770	10118.0
2	6305956	Office	1	1974	6681	4792.7
3	6506773	Office	1	2008	14548	11983.9
4	6731628	Office	1	2017	5223	3653.5
5	6867796	Office	1	1990	540	506.1

	PropertyID <int>	Property_Type <chr>	NumberofBuildings <int>	Year_Built <int>	Property_GFA <dbl>	Site_Energy <dbl>
6	8854296	Office	1	1979	17468	14092.5

6 rows | 1-7 of 25 columns

```
options(scipen = 999)
dataset <- na.omit(dataset)
```

1. Full Model

Generating the full model:

```
fullmodel <-lm(Site_Energy~factor(Property_Type)+NumberofBuildings+Year_Built+Property_GFA+WeatherNormalizedSiteEnergyUse+SiteEUI+WeatherNormalizedSiteEUI+SourceEnergyUse+WeatherNormalizedSourceEnergyUse+SourceEUI+WeatherNormalizedSource.EUI+Emissions_CO2+Emissions_Intensity+DirectGHGEmissions+DirectGHGEmissionsIntensity+Electricity+Natural_Gas+YearEnding+Precipitation+Corn_Heat_Unit+Heating_Degree_Days+Cooling_Degree_Days, data=dataset)

summary(fullmodel)
```

```
##
## Call:
## lm(formula = Site_Energy ~ factor(Property_Type) + NumberofBuildings +
##     Year_Built + Property_GFA + WeatherNormalizedSiteEnergyUse +
##     SiteEUI + WeatherNormalizedSiteEUI + SourceEnergyUse + WeatherNormalizedSourceEnergyUse +
##     SourceEUI + WeatherNormalizedSource.EUI + Emissions_CO2 +
##     Emissions_Intensity + DirectGHGEmissions + DirectGHGEmissionsIntensity +
##     Electricity + Natural_Gas + YearEnding + Precipitation +
##     Corn_Heat_Unit + Heating_Degree_Days + Cooling_Degree_Days,
##     data = dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33550 -0.09134  0.00039  0.08570  0.26818
##
## Coefficients: (6 not defined because of singularities)
##                                                                    Estimate
## (Intercept)                                                    28.7659833559
## factor(Property_Type)Fire Station                               0.0884312169
## factor(Property_Type)Fitness Center/Health Club/Gym             0.0790569638
## factor(Property_Type)Heated Swimming Pool                       0.1346219522
## factor(Property_Type)Ice/Curling Rink                           0.1174324193
## factor(Property_Type)Indoor Arena                               0.0590910332
## factor(Property_Type)Mixed Use Property                         0.1954009644
## factor(Property_Type)Museum                                     0.2062339421
## factor(Property_Type)Non-Refrigerated Warehouse                 0.0095253775
## factor(Property_Type)Office                                     0.1118708267
## factor(Property_Type)Other - Public Services                    0.0574617102
## factor(Property_Type)Other - Recreation                        -0.0151777547
## factor(Property_Type)Performing Arts                            0.1271807846
## factor(Property_Type)Repair Services (Vehicle, Shoe, Locksmith, etc.)  0.0554254507
## factor(Property_Type)Self-Storage Facility                      0.0833856660
## factor(Property_Type)Social/Meeting Hall                       -0.0205276682
## NumberofBuildings                                              -0.0112749263
## Year_Built                                                      0.0007085271
## Property_GFA                                                    0.0000030241
## WeatherNormalizedSiteEnergyUse                                  0.0002073388
## SiteEUI                                                        -0.3217617728
## WeatherNormalizedSiteEUI                                        1.8022868624
## SourceEnergyUse                                                 1.0000705939
## WeatherNormalizedSourceEnergyUse                               -0.0001716851
## SourceEUI                                                      -0.5867624667
## WeatherNormalizedSource.EUI                                    -1.6118909700
## Emissions_CO2                                                  -3.9332832830
## Emissions_Intensity                                             0.0148954005
## DirectGHGEmissions                                              0.1392791912
## DirectGHGEmissionsIntensity                                    -0.0006164748
## Electricity                                                    -0.0007812684
## Natural_Gas                                                     0.1847144677
## YearEnding                                                     -0.0149726708
## Precipitation416.3                                              0.0042762500
## Precipitation832.8                                                        NA
## Corn_Heat_Unit2,193.00                                                    NA
## Corn_Heat_Unit2,305.00                                                    NA
## Heating_Degree_Days4,970.20                                               NA
## Heating_Degree_Days5,335.70                                               NA
## Cooling_Degree_Days                                                       NA
##                                                                  Std. Error
## (Intercept)                                                    36.9383800074
## factor(Property_Type)Fire Station                               0.1940649302
## factor(Property_Type)Fitness Center/Health Club/Gym             0.1930821320
## factor(Property_Type)Heated Swimming Pool                       0.1967569867
## factor(Property_Type)Ice/Curling Rink                           0.1905503498
## factor(Property_Type)Indoor Arena                               0.1741265864
## factor(Property_Type)Mixed Use Property                         0.2105177042
## factor(Property_Type)Museum                                     0.2031812912
## factor(Property_Type)Non-Refrigerated Warehouse                 0.1874345452
## factor(Property_Type)Office                                     0.1909226028
## factor(Property_Type)Other - Public Services                    0.1981112004
## factor(Property_Type)Other - Recreation                         0.2080727773
## factor(Property_Type)Performing Arts                            0.2131517542
## factor(Property_Type)Repair Services (Vehicle, Shoe, Locksmith, etc.)  0.1827942412
## factor(Property_Type)Self-Storage Facility                      0.2006679983
## factor(Property_Type)Social/Meeting Hall                        0.2080363526
## NumberofBuildings                                               0.0410252323
## Year_Built                                                      0.0003879379
## Property_GFA                                                    0.0000049179
## WeatherNormalizedSiteEnergyUse                                  0.0001970934
```

```
## SiteEUI                                                          1.5643393527
## WeatherNormalizedSiteEUI                                         1.3056300718
## SourceEnergyUse                                                  0.0000848891
## WeatherNormalizedSourceEnergyUse                                 0.0001739649
## SourceEUI                                                        1.8329667633
## WeatherNormalizedSource.EUI                                      1.2387033594
## Emissions_CO2                                                    0.0014083659
## Emissions_Intensity                                              0.0125502821
## DirectGHGEmissions                                               0.2607609398
## DirectGHGEmissionsIntensity                                      0.0062696224
## Electricity                                                      0.0000003097
## Natural_Gas                                                      0.0133901000
## YearEnding                                                       0.0182724484
## Precipitation416.3                                               0.0383345691
## Precipitation832.8                                                         NA
## Corn_Heat_Unit2,193.00                                                     NA
## Corn_Heat_Unit2,305.00                                                     NA
## Heating_Degree_Days4,970.20                                                NA
## Heating_Degree_Days5,335.70                                                NA
## Cooling_Degree_Days                                                        NA
##                                                                   t value
## (Intercept)                                                         0.779
## factor(Property_Type)Fire Station                                   0.456
## factor(Property_Type)Fitness Center/Health Club/Gym                 0.409
## factor(Property_Type)Heated Swimming Pool                           0.684
## factor(Property_Type)Ice/Curling Rink                               0.616
## factor(Property_Type)Indoor Arena                                   0.339
## factor(Property_Type)Mixed Use Property                             0.928
## factor(Property_Type)Museum                                         1.015
## factor(Property_Type)Non-Refrigerated Warehouse                     0.051
## factor(Property_Type)Office                                         0.586
## factor(Property_Type)Other - Public Services                        0.290
## factor(Property_Type)Other - Recreation                            -0.073
## factor(Property_Type)Performing Arts                                0.597
## factor(Property_Type)Repair Services (Vehicle, Shoe, Locksmith, etc.)  0.303
## factor(Property_Type)Self-Storage Facility                          0.416
## factor(Property_Type)Social/Meeting Hall                           -0.099
## NumberofBuildings                                                  -0.275
## Year_Built                                                          1.826
## Property_GFA                                                        0.615
## WeatherNormalizedSiteEnergyUse                                      1.052
## SiteEUI                                                            -0.206
## WeatherNormalizedSiteEUI                                            1.380
## SourceEnergyUse                                                11780.903
## WeatherNormalizedSourceEnergyUse                                   -0.987
## SourceEUI                                                          -0.320
## WeatherNormalizedSource.EUI                                        -1.301
## Emissions_CO2                                                  -2792.799
## Emissions_Intensity                                                 1.187
## DirectGHGEmissions                                                  0.534
## DirectGHGEmissionsIntensity                                        -0.098
## Electricity                                                    -2523.068
## Natural_Gas                                                        13.795
## YearEnding                                                         -0.819
## Precipitation416.3                                                  0.112
## Precipitation832.8                                                     NA
## Corn_Heat_Unit2,193.00                                                 NA
## Corn_Heat_Unit2,305.00                                                 NA
## Heating_Degree_Days4,970.20                                            NA
## Heating_Degree_Days5,335.70                                            NA
## Cooling_Degree_Days                                                    NA
##                                                                   Pr(>|t|)
## (Intercept)                                                          0.437
## factor(Property_Type)Fire Station                                    0.649
## factor(Property_Type)Fitness Center/Health Club/Gym                  0.683
## factor(Property_Type)Heated Swimming Pool                            0.494
## factor(Property_Type)Ice/Curling Rink                                0.538
## factor(Property_Type)Indoor Arena                                    0.735
## factor(Property_Type)Mixed Use Property                              0.354
## factor(Property_Type)Museum                                          0.311
## factor(Property_Type)Non-Refrigerated Warehouse                      0.960
## factor(Property_Type)Office                                          0.558
## factor(Property_Type)Other - Public Services                         0.772
## factor(Property_Type)Other - Recreation                              0.942
## factor(Property_Type)Performing Arts                                 0.551
## factor(Property_Type)Repair Services (Vehicle, Shoe, Locksmith, etc.)  0.762
## factor(Property_Type)Self-Storage Facility                           0.678
## factor(Property_Type)Social/Meeting Hall                             0.921
## NumberofBuildings                                                    0.784
```

```
## Year_Built                                                    0.069
## Property_GFA                                                  0.539
## WeatherNormalizedSiteEnergyUse                                0.294
## SiteEUI                                                       0.837
## WeatherNormalizedSiteEUI                                      0.169
## SourceEnergyUse                                          <0.0000000000000002
## WeatherNormalizedSourceEnergyUse                              0.325
## SourceEUI                                                     0.749
## WeatherNormalizedSource.EUI                                   0.194
## Emissions_CO2                                            <0.0000000000000002
## Emissions_Intensity                                           0.236
## DirectGHGEmissions                                            0.594
## DirectGHGEmissionsIntensity                                   0.922
## Electricity                                              <0.0000000000000002
## Natural_Gas                                              <0.0000000000000002
## YearEnding                                                    0.413
## Precipitation416.3                                            0.911
## Precipitation832.8                                               NA
## Corn_Heat_Unit2,193.00                                           NA
## Corn_Heat_Unit2,305.00                                           NA
## Heating_Degree_Days4,970.20                                      NA
## Heating_Degree_Days5,335.70                                      NA
## Cooling_Degree_Days                                              NA
##
## (Intercept)
## factor(Property_Type)Fire Station
## factor(Property_Type)Fitness Center/Health Club/Gym
## factor(Property_Type)Heated Swimming Pool
## factor(Property_Type)Ice/Curling Rink
## factor(Property_Type)Indoor Arena
## factor(Property_Type)Mixed Use Property
## factor(Property_Type)Museum
## factor(Property_Type)Non-Refrigerated Warehouse
## factor(Property_Type)Office
## factor(Property_Type)Other - Public Services
## factor(Property_Type)Other - Recreation
## factor(Property_Type)Performing Arts
## factor(Property_Type)Repair Services (Vehicle, Shoe, Locksmith, etc.)
## factor(Property_Type)Self-Storage Facility
## factor(Property_Type)Social/Meeting Hall
## NumberofBuildings
## Year_Built                                                    .
## Property_GFA
## WeatherNormalizedSiteEnergyUse
## SiteEUI
## WeatherNormalizedSiteEUI
## SourceEnergyUse                                          ***
## WeatherNormalizedSourceEnergyUse
## SourceEUI
## WeatherNormalizedSource.EUI
## Emissions_CO2                                            ***
## Emissions_Intensity
## DirectGHGEmissions
## DirectGHGEmissionsIntensity
## Electricity                                              ***
## Natural_Gas                                              ***
## YearEnding
## Precipitation416.3
## Precipitation832.8
## Corn_Heat_Unit2,193.00
## Corn_Heat_Unit2,305.00
## Heating_Degree_Days4,970.20
## Heating_Degree_Days5,335.70
## Cooling_Degree_Days
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1246 on 257 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.575e+11 on 33 and 257 DF,  p-value: < 0.00000000000000022
```

2. Multicollinearity Assumption

VIF: The variance inflation factor (VIF), a straightforward test for multicollinearity in your regression model, can be used. The variance inflation factor (VIF) determines the existence and magnitude of correlations between independent variables.

```
library(mctest)
```

```
imcdiag(fullmodel, method='VIF')
```

```
## Warning in summary.lm(lm(x[, i] ~ x[, -i])): essentially perfect fit: summary
## may be unreliable

## Warning in summary.lm(lm(x[, i] ~ x[, -i])): essentially perfect fit: summary
## may be unreliable
```

```
##
## Call:
## imcdiag(mod = fullmodel, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##                                                                      VIF
## factor(Property_Type)Fire Station                                164.6769
## factor(Property_Type)Fitness Center/Health Club/Gym               52.8531
## factor(Property_Type)Heated Swimming Pool                         21.7378
## factor(Property_Type)Ice/Curling Rink                             62.8992
## factor(Property_Type)Indoor Arena                                  5.7957
## factor(Property_Type)Mixed Use Property                            8.4714
## factor(Property_Type)Museum                                        7.8912
## factor(Property_Type)Non-Refrigerated Warehouse                   19.7267
## factor(Property_Type)Office                                      119.7581
## factor(Property_Type)Other - Public Services                      35.9486
## factor(Property_Type)Other - Recreation                            8.2758
## factor(Property_Type)Performing Arts                               8.6847
## factor(Property_Type)Repair Services (Vehicle, Shoe, Locksmith, etc.)       12.6411
## factor(Property_Type)Self-Storage Facility                         7.6972
## factor(Property_Type)Social/Meeting Hall                           8.2729
## NumberofBuildings                                                  2.4799
## Year_Built                                                         1.7984
## Property_GFA                                                      47.1596
## WeatherNormalizedSiteEnergyUse                                203187.5547
## SiteEUI                                                        75885.4147
## WeatherNormalizedSiteEUI                                       53331.9624
## SourceEnergyUse                                                56200.6560
## WeatherNormalizedSourceEnergyUse                              237489.0434
## SourceEUI                                                     154227.7597
## WeatherNormalizedSource.EUI                                    70886.1664
## Emissions_CO2                                                  72431.0767
## Emissions_Intensity                                            35425.4031
## DirectGHGEmissions                                          640057341.2989
## DirectGHGEmissionsIntensity                                      2335.0530
## Electricity                                                      2431.5276
## Natural_Gas                                                 640077035.9937
## YearEnding                                                             Inf
## Precipitation416.3                                                     Inf
## Precipitation832.8                                                     Inf
## Corn_Heat_Unit2,193.00                                                 Inf
## Corn_Heat_Unit2,305.00                                                 Inf
## Heating_Degree_Days4,970.20                                            Inf
## Heating_Degree_Days5,335.70                                            Inf
## Cooling_Degree_Days                                                    Inf
##                                                             detection
## factor(Property_Type)Fire Station                                   1
## factor(Property_Type)Fitness Center/Health Club/Gym                 1
## factor(Property_Type)Heated Swimming Pool                           1
## factor(Property_Type)Ice/Curling Rink                               1
## factor(Property_Type)Indoor Arena                                   0
## factor(Property_Type)Mixed Use Property                             0
## factor(Property_Type)Museum                                         0
## factor(Property_Type)Non-Refrigerated Warehouse                     1
## factor(Property_Type)Office                                         1
## factor(Property_Type)Other - Public Services                        1
## factor(Property_Type)Other - Recreation                             0
## factor(Property_Type)Performing Arts                                0
## factor(Property_Type)Repair Services (Vehicle, Shoe, Locksmith, etc.)       1
## factor(Property_Type)Self-Storage Facility                          0
## factor(Property_Type)Social/Meeting Hall                            0
## NumberofBuildings                                                   0
## Year_Built                                                          0
## Property_GFA                                                        1
## WeatherNormalizedSiteEnergyUse                                      1
## SiteEUI                                                             1
## WeatherNormalizedSiteEUI                                            1
## SourceEnergyUse                                                     1
## WeatherNormalizedSourceEnergyUse                                    1
## SourceEUI                                                           1
## WeatherNormalizedSource.EUI                                         1
## Emissions_CO2                                                       1
## Emissions_Intensity                                                 1
## DirectGHGEmissions                                                  1
## DirectGHGEmissionsIntensity                                         1
## Electricity                                                         1
```

```
## Natural_Gas                                                                1
## YearEnding                                                                 1
## Precipitation416.3                                                         1
## Precipitation832.8                                                         1
## Corn_Heat_Unit2,193.00                                                     1
## Corn_Heat_Unit2,305.00                                                     1
## Heating_Degree_Days4,970.20                                                1
## Heating_Degree_Days5,335.70                                                1
## Cooling_Degree_Days                                                        1
##
## Multicollinearity may be due to factor(Property_Type)Fire Station factor(Property_Type)Fitness Center/Health Club/Gym fac
tor(Property_Type)Heated Swimming Pool factor(Property_Type)Ice/Curling Rink factor(Property_Type)Non-Refrigerated Warehouse
factor(Property_Type)Office factor(Property_Type)Other - Public Services factor(Property_Type)Repair Services (Vehicle, Sho
e, Locksmith, etc.) Property_GFA WeatherNormalizedSiteEnergyUse SiteEUI WeatherNormalizedSiteEUI SourceEnergyUse WeatherNorm
alizedSourceEnergyUse SourceEUI WeatherNormalizedSource.EUI Emissions_CO2 Emissions_Intensity DirectGHGEmissions DirectGHGEm
issionsIntensity Electricity Natural_Gas YearEnding Precipitation416.3 Precipitation832.8 Corn_Heat_Unit2,193.00 Corn_Heat_U
nit2,305.00 Heating_Degree_Days4,970.20 Heating_Degree_Days5,335.70 Cooling_Degree_Days regressors
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## ==================================
```

The VIF values are ideally supposed to be less than 5 for us to not take any corrective action. However, here in our fullmodel, we can observe that the VIF values are well above 5. As we suspected, there is considerable multicollinearity in our data! To solve this problem, we just drop the problematic variables. Since the presence of multicollinearity suggests that the information that this variable gives about the response is redundant in the presence of the other variables, this may typically be done without much harm to the regression model. After eliminating the problematic variables manually, we end up with: NumberofBuildings, Year_Built, Property_GFA, Natural_Gas, Emissions_CO2. But, we will not use Year_Built as it generates a Time-Series dependancy, which we want to avoid.

```
newmodel <-lm(Site_Energy~NumberofBuildings+Property_GFA+Natural_Gas+Emissions_CO2,data=dataset)
summary(newmodel)
```

```
##
## Call:
## lm(formula = Site_Energy ~ NumberofBuildings + Property_GFA +
##     Natural_Gas + Emissions_CO2, data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3671.1    -6.6   210.7   305.8  3779.9
##
## Coefficients:
##                       Estimate   Std. Error t value             Pr(>|t|)
## (Intercept)         702.127112   206.876859   3.394             0.000786 ***
## NumberofBuildings -1183.156115   191.397757  -6.182        0.00000000218 ***
## Property_GFA          0.153195     0.013026  11.761 < 0.0000000000000002 ***
## Natural_Gas           0.614171     0.005016 122.447 < 0.0000000000000002 ***
## Emissions_CO2         6.389525     0.108854  58.698 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 883.2 on 286 degrees of freedom
## Multiple R-squared:  0.9972, Adjusted R-squared:  0.9972
## F-statistic: 2.579e+04 on 4 and 286 DF,  p-value: < 0.00000000000000022
```

```
imcdiag(newmodel, method='VIF')
```

```
##
## Call:
## imcdiag(mod = newmodel, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##                      VIF detection
## NumberofBuildings 1.0748         0
## Property_GFA      6.5876         0
## Natural_Gas       1.7884         0
## Emissions_CO2     8.6158         0
##
## NOTE:  VIF Method Failed to detect multicollinearity
##
##
## 0 --> COLLINEARITY is not detected by the test
##
## ===================================
```

Here, we can observe that Property_GFA and Emissions_CO2 have a VIF>5, but they are important variables to us in the prediction of Site_Energy, and therefore we decide to keep them as a part of our New Model.

3. Stepwise Regression:

Stepwise regression is the iterative process of building a regression model step by step while choosing independent variables to be included in the final model. After each iteration, the potential explanatory factors are successively added or removed, and the statistical significance is tested.

```
library(olsrr)
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
stepmod=ols_step_both_p(newmodel,pent = 0.1, prem = 0.3, details=TRUE)
```

```
## Stepwise Selection Method
## ---------------------------
##
## Candidate Terms:
##
## 1. NumberofBuildings
## 2. Property_GFA
## 3. Natural_Gas
## 4. Emissions_CO2
##
## We are selecting variables based on p value...
##
##
## Stepwise Selection: Step 1
##
## - Emissions_CO2 added
##
##                            Model Summary
## ----------------------------------------------------------------------
## R                       0.915       RMSE                  6731.577
## R-Squared               0.838       Coef. Var               83.820
## Adj. R-Squared          0.837       MSE                45314123.578
## Pred R-Squared          0.806       MAE                    1853.342
## ----------------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                    ANOVA
## --------------------------------------------------------------------------------
##                       Sum of
##                      Squares         DF        Mean Square       F          Sig.
## --------------------------------------------------------------------------------
## Regression    67611714435.593          1    67611714435.593    1492.067    0.0000
## Residual      13095781714.069        289       45314123.578
## Total         80707496149.662        290
## --------------------------------------------------------------------------------
##
##                                  Parameter Estimates
## ----------------------------------------------------------------------------------------------------
##         model       Beta    Std. Error    Std. Beta      t        Sig        lower       upper
## ----------------------------------------------------------------------------------------------------
##   (Intercept)    148.905       444.249                  0.335    0.738    -725.469    1023.279
## Emissions_CO2     10.918         0.283        0.915    38.627    0.000      10.362      11.474
## ----------------------------------------------------------------------------------------------------
##
##
##
## Stepwise Selection: Step 2
##
## - Natural_Gas added
##
##                            Model Summary
## ----------------------------------------------------------------------
## R                       0.998       RMSE                  1160.494
## R-Squared               0.995       Coef. Var               14.450
## Adj. R-Squared          0.995       MSE                1346747.276
## Pred R-Squared          0.994       MAE                    609.930
## ----------------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                    ANOVA
## --------------------------------------------------------------------------------
##                       Sum of
##                      Squares         DF        Mean Square        F          Sig.
## --------------------------------------------------------------------------------
## Regression    80319632934.301          2    40159816467.150    29819.861    0.0000
## Residual        387863215.362        288        1346747.276
## Total         80707496149.662        290
## --------------------------------------------------------------------------------
##
##                                  Parameter Estimates
## ----------------------------------------------------------------------------------------------------
##         model        Beta    Std. Error    Std. Beta       t        Sig        lower       upper
## ----------------------------------------------------------------------------------------------------
##   (Intercept)    -491.951        76.870                  -6.400    0.000    -643.250    -340.653
```

```
## Emissions_CO2       7.490         0.060        0.628    124.490    0.000       7.371       7.608
##   Natural_Gas       0.591         0.006        0.490     97.139    0.000       0.579       0.603
## ----------------------------------------------------------------------------------------------
##
##
##
##                            Model Summary
## ----------------------------------------------------------------
## R                       0.998       RMSE                1160.494
## R-Squared               0.995       Coef. Var             14.450
## Adj. R-Squared          0.995       MSE                1346747.276
## Pred R-Squared          0.994       MAE                    609.930
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                      ANOVA
## ---------------------------------------------------------------------------------
##                        Sum of
##                       Squares         DF        Mean Square        F          Sig.
## ---------------------------------------------------------------------------------
## Regression    80319632934.301          2    40159816467.150    29819.861    0.0000
## Residual        387863215.362        288        1346747.276
## Total         80707496149.662        290
## ---------------------------------------------------------------------------------
##
##                                      Parameter Estimates
## ----------------------------------------------------------------------------------------------------
##         model        Beta    Std. Error    Std. Beta       t        Sig         lower       upper
## ----------------------------------------------------------------------------------------------------
##   (Intercept)    -491.951        76.870                -6.400    0.000    -643.250    -340.653
## Emissions_CO2       7.490         0.060        0.628    124.490    0.000       7.371       7.608
##   Natural_Gas       0.591         0.006        0.490     97.139    0.000       0.579       0.603
## ----------------------------------------------------------------------------------------------------
##
##
##
## Stepwise Selection: Step 3
##
## - Property_GFA added
##
##                           Model Summary
## ----------------------------------------------------------------
## R                       0.998       RMSE                 938.731
## R-Squared               0.997       Coef. Var             11.689
## Adj. R-Squared          0.997       MSE                881215.394
## Pred R-Squared          0.996       MAE                    580.080
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                      ANOVA
## ---------------------------------------------------------------------------------
##                        Sum of
##                       Squares         DF        Mean Square        F          Sig.
## ---------------------------------------------------------------------------------
## Regression    80454587331.529          3    26818195777.176     30433.19    0.0000
## Residual        252908818.133        287         881215.394
## Total         80707496149.662        290
## ---------------------------------------------------------------------------------
##
##                                      Parameter Estimates
## ----------------------------------------------------------------------------------------------------
##         model        Beta    Std. Error    Std. Beta       t        Sig         lower       upper
## ----------------------------------------------------------------------------------------------------
##   (Intercept)    -524.419        62.236                -8.426    0.000    -646.916    -401.922
## Emissions_CO2       6.234         0.113        0.523     55.383    0.000       6.012       6.455
##   Natural_Gas       0.616         0.005        0.511    115.789    0.000       0.606       0.627
##  Property_GFA       0.168         0.014        0.103     12.375    0.000       0.142       0.195
## ----------------------------------------------------------------------------------------------------
##
##
##
##                           Model Summary
## ----------------------------------------------------------------
## R                       0.998       RMSE                 938.731
## R-Squared               0.997       Coef. Var             11.689
```

```
## Adj. R-Squared          0.997       MSE                881215.394
## Pred R-Squared          0.996       MAE                   580.080
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                    ANOVA
## --------------------------------------------------------------------------------
##                       Sum of
##                      Squares         DF        Mean Square        F          Sig.
## --------------------------------------------------------------------------------
## Regression    80454587331.529          3    26818195777.176     30433.19    0.0000
## Residual        252908818.133        287         881215.394
## Total         80707496149.662        290
## --------------------------------------------------------------------------------
##
##                                    Parameter Estimates
## ----------------------------------------------------------------------------------------------------
##         model        Beta    Std. Error    Std. Beta       t        Sig        lower       upper
## ----------------------------------------------------------------------------------------------------
##   (Intercept)    -524.419        62.236                -8.426    0.000    -646.916    -401.922
## Emissions_CO2       6.234         0.113        0.523     55.383    0.000       6.012       6.455
##   Natural_Gas       0.616         0.005        0.511    115.789    0.000       0.606       0.627
##  Property_GFA       0.168         0.014        0.103     12.375    0.000       0.142       0.195
## ----------------------------------------------------------------------------------------------------
##
##
##
## Stepwise Selection: Step 4
##
## - NumberofBuildings added
##
##                          Model Summary
## ----------------------------------------------------------------
## R                       0.999       RMSE                883.216
## R-Squared               0.997       Coef. Var            10.998
## Adj. R-Squared          0.997       MSE                780070.100
## Pred R-Squared          0.996       MAE                   507.858
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                    ANOVA
## --------------------------------------------------------------------------------
##                       Sum of
##                      Squares         DF        Mean Square        F          Sig.
## --------------------------------------------------------------------------------
## Regression    80484396101.175          4    20121099025.294    25793.963    0.0000
## Residual        223100048.487        286         780070.100
## Total         80707496149.662        290
## --------------------------------------------------------------------------------
##
##                                       Parameter Estimates
## ----------------------------------------------------------------------------------------------------
##             model         Beta    Std. Error    Std. Beta       t        Sig         lower       upper
## ----------------------------------------------------------------------------------------------------
##       (Intercept)      702.127       206.877                   3.394    0.001      294.933    1109.321
##     Emissions_CO2        6.390         0.109        0.536     58.698    0.000        6.175       6.604
##       Natural_Gas        0.614         0.005        0.509    122.447    0.000        0.604       0.624
##      Property_GFA        0.153         0.013        0.094     11.761    0.000        0.128       0.179
## NumberofBuildings    -1183.156       191.398       -0.020     -6.182    0.000    -1559.883    -806.429
## ----------------------------------------------------------------------------------------------------
##
##
##
##                          Model Summary
## ----------------------------------------------------------------
## R                       0.999       RMSE                883.216
## R-Squared               0.997       Coef. Var            10.998
## Adj. R-Squared          0.997       MSE                780070.100
## Pred R-Squared          0.996       MAE                   507.858
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                    ANOVA
```

```
## --------------------------------------------------------------------------------
##                       Sum of
##                      Squares         DF        Mean Square        F          Sig.
## --------------------------------------------------------------------------------
## Regression    80484396101.175          4    20121099025.294    25793.963    0.0000
## Residual        223100048.487        286         780070.100
## Total         80707496149.662        290
## --------------------------------------------------------------------------------
##
##                                        Parameter Estimates
## ----------------------------------------------------------------------------------------------------------------
##             model         Beta    Std. Error    Std. Beta       t        Sig         lower       upper
## ----------------------------------------------------------------------------------------------------------------
##       (Intercept)      702.127       206.877                   3.394    0.001      294.933    1109.321
##     Emissions_CO2        6.390         0.109        0.536     58.698    0.000        6.175       6.604
##       Natural_Gas        0.614         0.005        0.509    122.447    0.000        0.604       0.624
##      Property_GFA        0.153         0.013        0.094     11.761    0.000        0.128       0.179
## NumberofBuildings    -1183.156       191.398       -0.020     -6.182    0.000    -1559.883    -806.429
## ----------------------------------------------------------------------------------------------------------------
##
##
##
##
## Final Model Output
## ------------------
##
##                           Model Summary
## ------------------------------------------------------------------
## R                       0.999       RMSE                 883.216
## R-Squared               0.997       Coef. Var             10.998
## Adj. R-Squared          0.997       MSE                780070.100
## Pred R-Squared          0.996       MAE                   507.858
## ------------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                     ANOVA
## --------------------------------------------------------------------------------
##                       Sum of
##                      Squares         DF        Mean Square        F          Sig.
## --------------------------------------------------------------------------------
## Regression    80484396101.175          4    20121099025.294    25793.963    0.0000
## Residual        223100048.487        286         780070.100
## Total         80707496149.662        290
## --------------------------------------------------------------------------------
##
##                                        Parameter Estimates
## ----------------------------------------------------------------------------------------------------------------
##             model         Beta    Std. Error    Std. Beta       t        Sig         lower       upper
## ----------------------------------------------------------------------------------------------------------------
##       (Intercept)      702.127       206.877                   3.394    0.001      294.933    1109.321
##     Emissions_CO2        6.390         0.109        0.536     58.698    0.000        6.175       6.604
##       Natural_Gas        0.614         0.005        0.509    122.447    0.000        0.604       0.624
##      Property_GFA        0.153         0.013        0.094     11.761    0.000        0.128       0.179
## NumberofBuildings    -1183.156       191.398       -0.020     -6.182    0.000    -1559.883    -806.429
## ----------------------------------------------------------------------------------------------------------------
```

```
summary(stepmod$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3671.1    -6.6   210.7   305.8  3779.9
##
## Coefficients:
##                       Estimate   Std. Error t value             Pr(>|t|)
## (Intercept)         702.127112   206.876859   3.394             0.000786 ***
## Emissions_CO2         6.389525     0.108854  58.698 < 0.0000000000000002 ***
## Natural_Gas           0.614171     0.005016 122.447 < 0.0000000000000002 ***
## Property_GFA          0.153195     0.013026  11.761 < 0.0000000000000002 ***
## NumberofBuildings -1183.156115   191.397757  -6.182        0.00000000218 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 883.2 on 286 degrees of freedom
## Multiple R-squared:  0.9972, Adjusted R-squared:  0.9972
## F-statistic: 2.579e+04 on 4 and 286 DF,  p-value: < 0.00000000000000022
```

After doing stepwise regression, we get our final model as: Site_Energy = 702.127112 + 6.389525(Emissions_CO2) + 0.614171(Natural_Gas) + 0.153195(Property_GFA) - 1183.156115(NumberofBuildings)

Our full model:

$$

$$Y_{SiteEnergy} == \beta_0 + \beta_1 X_{EmissionsCO2} + \beta_2 X_{Natural_Gas} + \beta_3 X_{Property_GFA} + \beta_4 X_{NumberofBuildings} + \epsilon$$

$$

4.Global F Test on Full Model:

Any statistical test with an F-distribution for the test statistic under the null hypothesis is known as an F-test. In order to determine which statistical model better represents the population from which the data were sampled, it is most frequently applied when contrasting models that have been fitted to data sets.

Hypothesis Statement for Individual T-test:

$$H_0 : \beta_i = 0$$
$$H_a : \text{at least one } \beta_i \text{ is not zero } (i = 1, 2, 3, 4)$$

We set up the significance level at 0.05($\alpha = 0.05$).

Full Model Global F Test:

```
#Full Model Test
fullmodel <-lm(Site_Energy~NumberofBuildings+Property_GFA+Emissions_CO2+ Natural_Gas, data=dataset)
reg<-lm(Site_Energy~1, data=dataset) # Model with only intercept

anova(fullmodel,reg) # We compare the NULL model with the full model
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	286	223100048	*NA*	*NA*	*NA*	*NA*
2	290	80707496150	-4	-80484396101	25793.96	0

2 rows

By using global F test, the output shows that $F_{cal} = 25794$ with df= -4, and $p - value < 0.00000000000000022 < \alpha = 0.05$,indicating that we should clearly reject the null hypothesis. It provides compelling evidence against the null hypothesis. The Global F-test suggests that at least one of the independent variables must be related to Site Energy. Based on the p-value, we also have extremely strong evidence that at least one of the independent variables is associated with increased Site Energy.

After we check the global F-test and reject the null hypothesis, we are checking the test statistics for the individual coefficients and particular subsets of the full model test in the following steps.

5.Individual Coefficients Test:

To evaluate whether there is a significant difference between the means of two groups and their relationships, a t-test is an inferential statistic that is used. When data sets contain unknown variances and a normal distribution, such as the data set obtained from tossing a coin 100 times, t-tests are utilised. In order to evaluate statistical significance, the t-test, a test used for hypothesis testing in statistics, uses the t-statistic, the values of the t-distribution, and the degrees of freedom.

Hypothesis Statement for Individual T-test:

$$H_0 : \beta_i = 0$$
$$H_a : \beta_i \neq 0 (i = 1, 2, 3, 4)$$

We set up the significance level at 0.05($\alpha = 0.05$).

```
summary(fullmodel)
```

```
##
## Call:
## lm(formula = Site_Energy ~ NumberofBuildings + Property_GFA +
##     Emissions_CO2 + Natural_Gas, data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3671.1    -6.6   210.7   305.8  3779.9
##
## Coefficients:
##                       Estimate   Std. Error t value             Pr(>|t|)
## (Intercept)         702.127112   206.876859   3.394            0.000786 ***
## NumberofBuildings -1183.156115   191.397757  -6.182       0.00000000218 ***
## Property_GFA          0.153195     0.013026  11.761 < 0.0000000000000002 ***
## Emissions_CO2         6.389525     0.108854  58.698 < 0.0000000000000002 ***
## Natural_Gas           0.614171     0.005016 122.447 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 883.2 on 286 degrees of freedom
## Multiple R-squared:  0.9972, Adjusted R-squared:  0.9972
## F-statistic: 2.579e+04 on 4 and 286 DF,  p-value: < 0.00000000000000022
```

From the out put, It shows that the variables NumberofBuildings has $t_{cal} = -6.182$ with the $p - value = 0.00000000218 < 0.05$, Property_GFA has $t_{cal} = 11.761$ with the $p - value < 0.0000000000000002 < 0.05$, Emissions_CO2 has $t_{cal} = 58.698$ with the $p - value < 0.0000000000000002 < 0.05$, Natural_Gas has $t_{cal} = 122.447$ with the $p - value < 0.0000000000000002 < 0.05$, indicating that we should clearly reject the null hypothesis for variable NumberofBuildings, Property_GFA,Emissions_CO2 and Natural_Gas, which means that we should keep those variables to the model at the significance level at $\alpha = 0.05$.

Interaction Model

Interaction models A particular trait of three or more variables known as an interaction in statistics occurs when two or more variables interact to impact a third variable in a non-additive way. In other words, the two factors interact to produce a result that is greater than the combination of their individual effects.

Individual Coefficients Test (T-tests) on Interaction Term:

For testing an interaction term in regression model, we use the Individual Coefficients Test (t-test) method. *Hypothesis Statement:*

$$H_0 : \beta_i = 0$$
$$H_a : \beta_i \neq 0$$

$(i = NumberofBuildings * EmissionsCO2, NumberofBuildings * NaturalGas, NumberofBuildings * PropertyGFA, EmissionsCO2 *$

We set up the significance level at 0.05($\alpha = 0.05$).

```
#T-tests:
#Hypothesis:
#H0= Bi = 0
#Ha: Bi != 0 (i = 1,2,3...)
interacmodel <-lm(Site_Energy~(NumberofBuildings+Emissions_CO2+Natural_Gas+Property_GFA)^2, data=dataset)
summary(interacmodel)
```

```
##
## Call:
## lm(formula = Site_Energy ~ (NumberofBuildings + Emissions_CO2 +
##     Natural_Gas + Property_GFA)^2, data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1012.0   -45.6   -19.7    19.4  3328.0
##
## Coefficients:
##                                     Estimate     Std. Error t value
## (Intercept)                     -94.7097644699 392.9567782134  -0.241
## NumberofBuildings                95.3410856191 389.5627874334   0.245
## Emissions_CO2                     5.2559291330   0.1776056839  29.593
## Natural_Gas                       0.6356456233   0.0375272849  16.938
## Property_GFA                      0.1501775882   0.0423196419   3.549
## NumberofBuildings:Emissions_CO2  -0.1029419638   0.1498594656  -0.687
## NumberofBuildings:Natural_Gas     0.0436008071   0.0353678947   1.233
## NumberofBuildings:Property_GFA   -0.0708970683   0.0396449875  -1.788
## Emissions_CO2:Natural_Gas         0.0000123024   0.0000015975   7.701
## Emissions_CO2:Property_GFA        0.0000241965   0.0000008809  27.469
## Natural_Gas:Property_GFA         -0.0000026951   0.0000001841 -14.642
##                                             Pr(>|t|)
## (Intercept)                                 0.809717
## NumberofBuildings                           0.806838
## Emissions_CO2                   < 0.0000000000000002 ***
## Natural_Gas                     < 0.0000000000000002 ***
## Property_GFA                                0.000454 ***
## NumberofBuildings:Emissions_CO2             0.492700
## NumberofBuildings:Natural_Gas               0.218692
## NumberofBuildings:Property_GFA              0.074809 .
## Emissions_CO2:Natural_Gas          0.000000000000234 ***
## Emissions_CO2:Property_GFA      < 0.0000000000000002 ***
## Natural_Gas:Property_GFA        < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 391.7 on 280 degrees of freedom
## Multiple R-squared:  0.9995, Adjusted R-squared:  0.9994
## F-statistic: 5.258e+04 on 10 and 280 DF,  p-value: < 0.00000000000000022
```

From the out put, It shows that the interaction terms $EmissionsCO2 * NaturalGas$ has $t_{cal} = 7.701$ with the $p - value = 0.000000000000234 < 0.05$, $EmissionsCO2 * PropertyGFA$ has $t_{cal} = 27.469$ with the $p - value < 0.0000000000000002 < 0.05$, and $NaturalGas * PropertyGFA$ has $t_{cal} = -14.642$ with the $p - value < 0.0000000000000002 < 0.05$ indicating that we should clearly reject the null hypothesis which means that we should add the above interaction terms to the model at $\alpha = 0.05$.

However, the interaction terms $NumberofBuildings * EmissionsCO2$ has $t_{cal} = -0.687$ with the $p - value = 0.492700 > 0.05$, $NumberofBuildings * NaturalGas$ has $t_{cal}=1.233 $ with the $p - value = 0.218692 > 0.05$, indicating that we should clearly not reject the null hypothesis which means that we should not add the above interaction terms to the model at $\alpha = 0.05$.

The interaction terms $NumberofBuildings * PropertyGFA$ has $t_{cal} = -1.788$ with the $p - value = 0.074809 > 0.05$, is lies in the grey zone.

After including the interaciton terms we can see that not every interaction term contributes to the model. We are removing NumberofBuildings:Emissions_CO2, NumberofBuildings:Natural_Gas but we are going to keep NumberofBuildings:Property_GFA as it lies in the grey zone. Hence we are going to do the T-test again with the predictors that are significant to the model. For the rest of the predictors p-value is less than the alpha value(0.05) hence, we can reject our null hypothesis and accept the alternative.

Individual Coefficients Test (T-tests) on Interaction Term:

Hypothesis Statement:

$$H_0 : \beta_i = 0$$
$$H_a : \beta_i \neq 0$$

$(i = NumberofBuildings * PropertyGFA, EmissionsCO2 * NaturalGas, EmissionsCO2 * PropertyGFA, NaturalGas * PropertyGFA)$

We set up the significance level at 0.05($\alpha = 0.05$).

```
#Hypothesis:
#H0= Bi = 0
#Ha: Bi != 0 (i = 1,2,3...)
interacmodel1 <-lm(Site_Energy~NumberofBuildings+Emissions_CO2+Natural_Gas+Property_GFA+NumberofBuildings*Property_GFA+Emiss
ions_CO2*Natural_Gas+Emissions_CO2*Property_GFA+Natural_Gas*Property_GFA, data=dataset)

summary(interacmodel1)
```

```
##
## Call:
## lm(formula = Site_Energy ~ NumberofBuildings + Emissions_CO2 +
##     Natural_Gas + Property_GFA + NumberofBuildings * Property_GFA +
##     Emissions_CO2 * Natural_Gas + Emissions_CO2 * Property_GFA +
##     Natural_Gas * Property_GFA, data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1001.3   -49.0   -19.2    18.7  3341.8
##
## Coefficients:
##                                     Estimate      Std. Error t value
## (Intercept)                   -168.8310349319  263.7047253788  -0.640
## NumberofBuildings              167.5886730373  259.9071206976   0.645
## Emissions_CO2                    5.1595052172    0.0660808040  78.079
## Natural_Gas                      0.6804319968    0.0080185943  84.857
## Property_GFA                     0.1188225012    0.0318151677   3.735
## NumberofBuildings:Property_GFA  -0.0409164942    0.0282224066  -1.450
## Emissions_CO2:Natural_Gas        0.0000120316    0.0000015810   7.610
## Emissions_CO2:Property_GFA       0.0000242558    0.0000008766  27.672
## Natural_Gas:Property_GFA        -0.0000026798    0.0000001827 -14.666
##                                            Pr(>|t|)
## (Intercept)                                0.522545
## NumberofBuildings                          0.519580
## Emissions_CO2                  < 0.0000000000000002 ***
## Natural_Gas                    < 0.0000000000000002 ***
## Property_GFA                               0.000227 ***
## NumberofBuildings:Property_GFA             0.148229
## Emissions_CO2:Natural_Gas         0.000000000000413 ***
## Emissions_CO2:Property_GFA     < 0.0000000000000002 ***
## Natural_Gas:Property_GFA       < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 391.4 on 282 degrees of freedom
## Multiple R-squared:  0.9995, Adjusted R-squared:  0.9994
## F-statistic: 6.581e+04 on 8 and 282 DF,  p-value: < 0.00000000000000022
```

From the out put, It shows that the interaction term $NumberofBuildings * PropertyGFA$ has $t_{cal} = -1.450$ with the $p - value = 0.148229 > 0.05$, indicating that we should clearly not reject the null hypothesis which means that we should not add the above interaction terms to the model at $\alpha = 0.05$.

The interaction terms $EmissionsCO2 * NaturalGas$ has $t_{cal}=7.610 $ with the $p - value = 0.000000000000413 < 0.05$, $EmissionsCO2 * PropertyGFA$ has $t_{cal}=27.672 $ with the $p - value < 0.0000000000000002 < 0.05$, and $NaturalGas * PropertyGFA$ has $t_{cal}=-14.666 $ with the $p - value < 0.0000000000000002 < 0.05$ indicating that we should clearly reject the null hypothesis which means that we should keep the above interaction terms to the model at $\alpha = 0.05$.

After removing the insignificant terms yet keeping the grey zone predictor, we see that it also becomes insignificant to the model. Hence we can now safely remove the predictor which once was in the grey zone. Therefore, we are left with NumberofBuildings, Emissions_CO2, Natural_Gas+Property_GFA, Emissions_CO2:Natural_Gas, Emissions_CO2:Property_GFA, Natural_Gas:Property_GFA as our final predictors for our model. As for these predictors the p-value is less than the alpha value(0.05) hence, we can reject our null hypothesis and accept the alternative.

Individual Coefficients Test (T-tests) on Interaction Term:

Hypothesis Statement:

$$H_0 : \beta_i = 0$$
$$H_a : \beta_i \neq 0$$

$$(i = EmissionsCO2 * NaturalGas, EmissionsCO2 * PropertyGFA, NaturalGas * PropertyGFA)$$

We set up the significance level at 0.05($\alpha = 0.05$).

```
#Hypothesis:
#H0= Bi = 0
#Ha: Bi != 0 (i = 1,2,3...)
interacmodel2 <-lm(Site_Energy~NumberofBuildings+Emissions_CO2+Natural_Gas+Property_GFA+Emissions_CO2*Natural_Gas+Emissions_
CO2*Property_GFA+Natural_Gas*Property_GFA, data=dataset)

summary(interacmodel2)
```

```
##
## Call:
## lm(formula = Site_Energy ~ NumberofBuildings + Emissions_CO2 +
##     Natural_Gas + Property_GFA + Emissions_CO2 * Natural_Gas +
##     Emissions_CO2 * Property_GFA + Natural_Gas * Property_GFA,
##     data = dataset)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -965.5  -53.4  -22.3   19.9 3358.7
##
## Coefficients:
##                                  Estimate      Std. Error t value
## (Intercept)                 188.8801999814   93.2548459558   2.025
## NumberofBuildings          -185.4587942484   91.0135231331  -2.038
## Emissions_CO2                 5.1848117794    0.0638576446  81.193
## Natural_Gas                   0.6779761710    0.0078528760  86.335
## Property_GFA                  0.0744549739    0.0087160491   8.542
## Emissions_CO2:Natural_Gas     0.0000122403    0.0000015775   7.759
## Emissions_CO2:Property_GFA    0.0000242883    0.0000008780  27.664
## Natural_Gas:Property_GFA     -0.0000026627    0.0000001827 -14.575
##                                        Pr(>|t|)
## (Intercept)                              0.0438 *
## NumberofBuildings                        0.0425 *
## Emissions_CO2              < 0.0000000000000002 ***
## Natural_Gas                < 0.0000000000000002 ***
## Property_GFA               0.000000000000000826 ***
## Emissions_CO2:Natural_Gas  0.000000000000156309 ***
## Emissions_CO2:Property_GFA < 0.0000000000000002 ***
## Natural_Gas:Property_GFA   < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 392.2 on 283 degrees of freedom
## Multiple R-squared:  0.9995, Adjusted R-squared:  0.9994
## F-statistic: 7.492e+04 on 7 and 283 DF,  p-value: < 0.00000000000000022
```

From the out put, It shows that the interaction terms $EmissionsCO2 * NaturalGas$ has $t_{cal}=7.759 $ with the $p-value = 0.000000000000156309 < 0.05$, $EmissionsCO2 * PropertyGFA$ has $t_{cal}=27.664 $ with the $p-value < 0.0000000000000002 < 0.05$, and $NaturalGas * PropertyGFA$ has $t_{cal}=-14.575 $ with the $p-value < 0.0000000000000002 < 0.05$ indicating that we should clearly reject the null hypothesis which means that we should keep the above interaction terms to the model at $\alpha = 0.05$.

All the remaining predictors are significant to the model. Their p-value is much less than the alpha value(0.05). Hence, we can reject our null hypothesis and accept the alternative.

Interaction Term Partial F-tests:

After fitting a model with all interactions, we dropped non-significant interaction terms. Final estimation model obtained is the interacmodel3.

$$Y_{SiteEnergy} = \beta_0 + \beta_1 X_{EmissionsCO2} + \beta_2 X_{NaturalGas} + \beta_3 X_{PropertyGFA} + \beta_4 X_{NumberofBuildings} + \beta_5 X_{EmissionsCO2*NaturalGas} + \beta_6 X_{EmissionsCO2*PropertyGF} + \beta_7 X_{NaturalGas*PropertyGFA} + \epsilon$$

To confirm that we should drop all those interaction terms together, we perform a partial F-test.

Hypothesis Statement:

$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \ldots = \beta_p = 0 (\text{Interaction terms are not significant})$$
$$H_a : \text{at least one } \beta_i \neq 0 (\text{At least one interaction term is significant})$$

We set up the significance level at 0.05($\alpha = 0.05$).

```
anova(interacmodel2,interacmodel1)
```

Res.Df <dbl>	RSS <dbl>	Df <dbl>	Sum of Sq <dbl>	F <dbl>	Pr(>F) <dbl>
1	283	43527112	*NA*	*NA*	*NA*
2	282	43205083	1	322028.6	2.101884	0.148229

2 rows

It gives a p-value of $0.1482 > 0.05$, indicating that we should clearly not to reject the null hypothesis, which confirms that we do not have enough evidence to keep those non-significant interaction terms in the model. Hence, we can conclude that the interaction model below is the best fitted model:

$$\widehat{Y_{SiteEnergy}} = \beta_0 + \beta_1 X_{EmissionsCO2} + \beta_2 X_{NaturalGas} + \beta_3 X_{PropertyGFA} + \beta_4 X_{NumberofBuildings} + \beta_5 X_{EmissionsCO2*NaturalGas} + \beta_6 X_{EmissionsCO2*PropertyGF} + \beta_7 X_{NaturalGas*PropertyGFA}$$

Higher Order

In the previous steps, the stepwise regression procedure declared that the best one-variable predictor of site energy usage is emissions CO2. Hence, we decided to add quadratic term to the model based on it.

For testing an higher order term in regression model, we use the Individual Coefficients Test (t-test) method.

```
interacmodel2 <-lm(Site_Energy~NumberofBuildings+Emissions_CO2+I(Emissions_CO2^2)+I(Emissions_CO2^3)+Natural_Gas+Property_GF
A+Emissions_CO2*Natural_Gas+Emissions_CO2*Property_GFA+Natural_Gas*Property_GFA, data=dataset)

summary(interacmodel2)
```

```
##
## Call:
## lm(formula = Site_Energy ~ NumberofBuildings + Emissions_CO2 +
##     I(Emissions_CO2^2) + I(Emissions_CO2^3) + Natural_Gas + Property_GFA +
##     Emissions_CO2 * Natural_Gas + Emissions_CO2 * Property_GFA +
##     Natural_Gas * Property_GFA, data = dataset)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -967.3  -39.7    2.7   39.2 3224.7
##
## Coefficients:
##                                    Estimate        Std. Error t value
## (Intercept)                 149.718178262441   91.326756015509   1.639
## NumberofBuildings          -200.086431113004   89.100506340703  -2.246
## Emissions_CO2                 5.777410420664    0.154756832968  37.332
## I(Emissions_CO2^2)           -0.000225682523    0.000054376006  -4.150
## I(Emissions_CO2^3)            0.000000007538    0.000000003011   2.504
## Natural_Gas                   0.662884049277    0.008709747855  76.108
## Property_GFA                  0.040574086706    0.012157593104   3.337
## Emissions_CO2:Natural_Gas     0.000019266916    0.000002418796   7.965
## Emissions_CO2:Property_GFA    0.000038513764    0.000004279453   9.000
## Natural_Gas:Property_GFA     -0.000003159993    0.000000238654 -13.241
##                                      Pr(>|t|)
## (Intercept)                           0.10226
## NumberofBuildings                     0.02551 *
## Emissions_CO2               < 0.0000000000000002 ***
## I(Emissions_CO2^2)            0.0000440369703034 ***
## I(Emissions_CO2^3)                    0.01286 *
## Natural_Gas                 < 0.0000000000000002 ***
## Property_GFA                          0.00096 ***
## Emissions_CO2:Natural_Gas    0.0000000000000412 ***
## Emissions_CO2:Property_GFA < 0.0000000000000002 ***
## Natural_Gas:Property_GFA   < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 381.8 on 281 degrees of freedom
## Multiple R-squared:  0.9995, Adjusted R-squared:  0.9995
## F-statistic: 6.15e+04 on 9 and 281 DF,  p-value: < 0.00000000000000022
```

In conclusion, adding higher-order terms to a Linear regression model can improve its predictive power by allowing for more complex relationships between the dependent and independent variables. In this case, the addition of up to the power of 3 for the main effect Emissions_CO2 resulted in improved model performance. However, when the power of 4 was tested, the model did not pass the hypothetical test, therefore, should not be included in the final model. This shows that while the addition of higher-order terms can be beneficial, it is important to carefully evaluate their impact on the model and avoid overfitting. Overall, the use of higher-order terms can help to better capture the underlying relationships in the data and improve the accuracy of the model's predictions.

We will be using the Higher Order model for further calculations.

7.Linearity Assumption

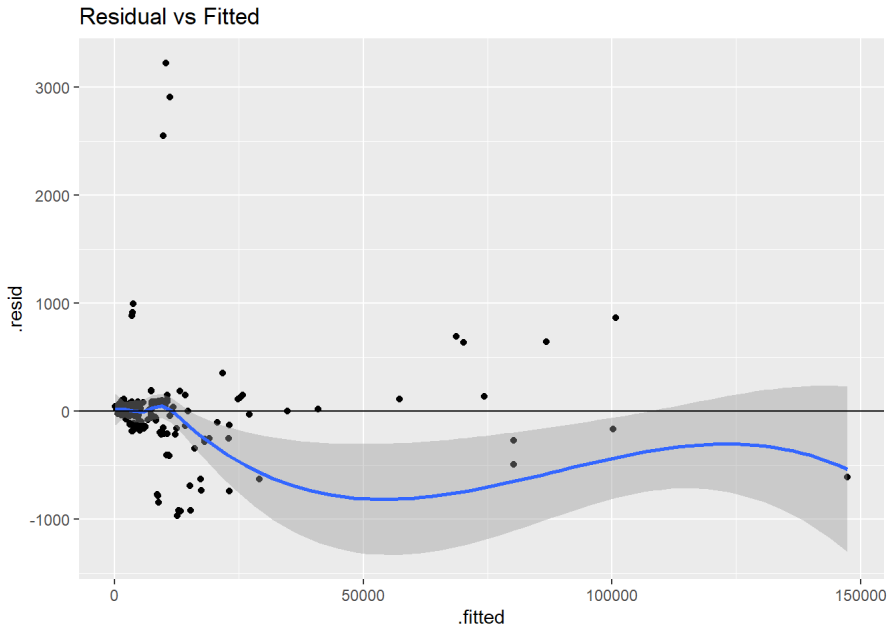
Linearity Assumption The linear regression model counts on a linear relationship existing between the predictors and the outcome. Almost all of the conclusions we get from the fit are dubious if the underlying relationship is not linear. Additionally, the model's forecast accuracy may suffer dramatically.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```
ggplot(interacmodel2, aes(x=.fitted, y=.resid)) +
  geom_point() + geom_smooth()+
  geom_hline(yintercept = 0)+
  ggtitle("Residual vs Fitted")
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



We can draw the conclusion that, compared to a basic linear regression model, the quadratic model more closely fits the data. Model interpretations are meaningless when the independent variable's range is exceeded. Despite the fact that the model seems to back up the data. The value of X must fall inside the range of the independent variable in order to generate a forecast for Y. Otherwise, the prediction won't have any real relevance.

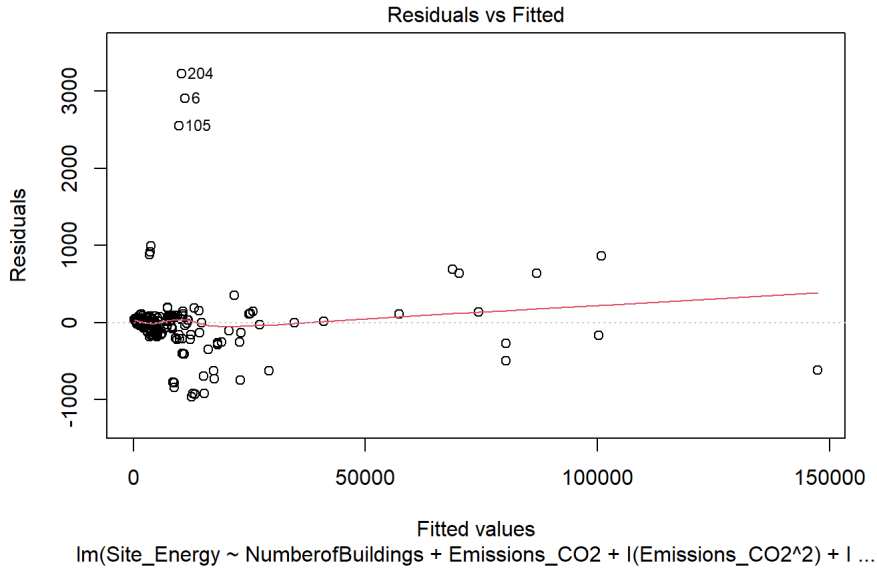
8.Independence Assumption

Independence Assumption When subsequent errors are correlated, the assumption of independent errors is broken. This often happens when time-series data, which are observations of data for both dependent and independent variables sequentially over a period of time, are used. Since the objects of our experiment were unrelated to time, we may be quite confident that the measurements are independent.

9.Equal Variance Assumption

Equal Variance Assumption Test: Uneven dispersal is a sign of heteroscedasticity. Heteroscedasticity in regression analysis is a systematic alteration in the distribution of the residuals over the range of measured values. Using a concave function to convert the response Y in response to this issue is one potential fix. We conduct the equal variance assumption test (Breusch- Pagan test).

```
plot(interacmodel2,which=1)
```




```
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.2.2
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 4.2.3
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
#H0: heteroscedacity is not present
#H1: Heteroscedacity is present
bptest(interacmodel2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  interacmodel2
## BP = 93.687, df = 9, p-value = 0.0000000000000002956
```

From our dataset, the output displays the Breusch-Pagan test that results from the cubic model. The p-value = 0.0000000000000002956 <0.05, indicating that we do reject the null hypothesis. Therefore, the test provides very strong evidence to suggest that heteroscedasticity does exist. Therefore, we do the Shapiro-Wilk test.

10.Normality Assumption

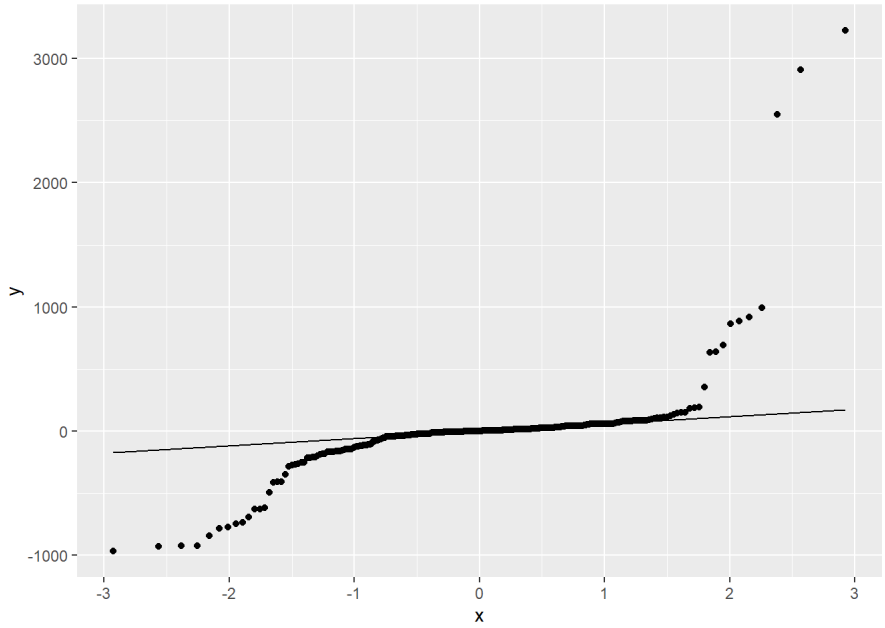
Normality Assumption Test: The residuals of the regression, or the errors between observed and predicted values, must be regularly distributed in order to perform a multiple linear regression analysis. By examining a histogram, a normal probability plot, or a Q-Q-Plot, this assumption can be verified. We conduct the Normality assumption test (Shapiro-Wilk Test).

```
#H0: the sample data are significantly normally distributed
#Ha: the sample data are not significantly normally distributed
shapiro.test(residuals(interacmodel2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(interacmodel2)
## W = 0.49317, p-value < 0.00000000000000022
```

Shapiro-Wilk normality test also confirms that the residuals are NOT normally distributed as the p-value= 0.00000000000000022< 0.05. It is alson confirmed in the normal Q-Q plot below:

```
#normal QQ plot
ggplot(dataset, aes(sample=interacmodel2$residuals)) +
  stat_qq() +
  stat_qq_line()
```



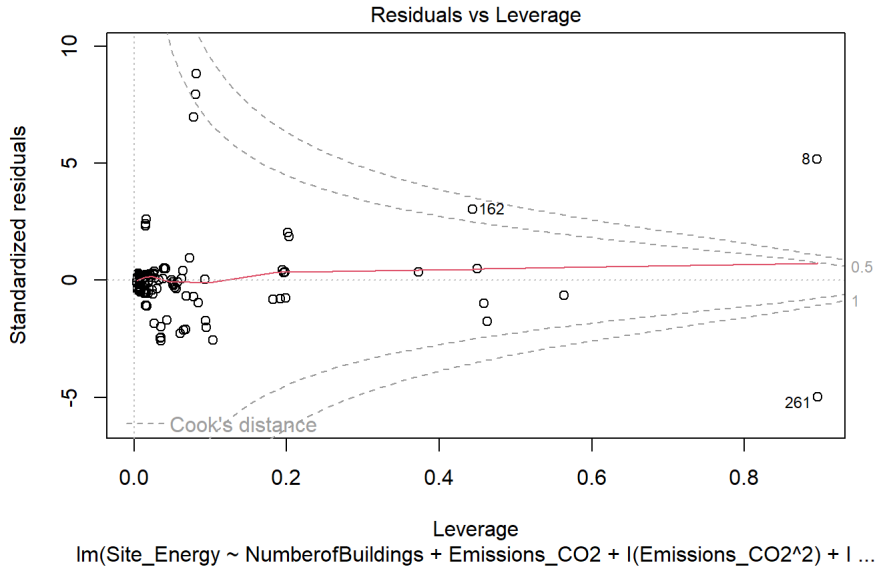
Non-normal and Heteroscedastic means we need to do the Box-Cox transformation in the forthcoming steps.

11.Outliers Test

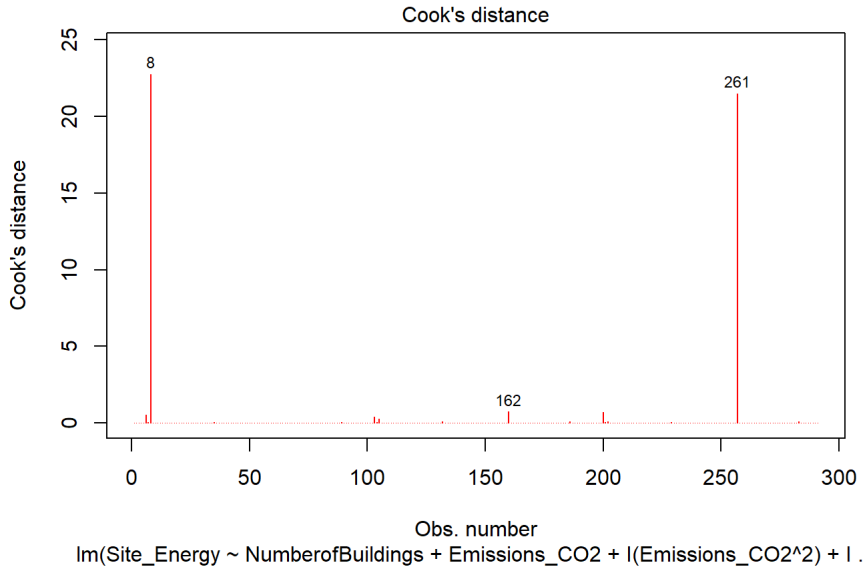
Outliers Test: A specific observation (Y, X1, X2,…, Xp) that deviates from the majority of the cases in the data collection is referred to as an outlier case. We can identify and assess outlier or influential points in a variety of ways. We use Cook's distance method. The Cook's distance Di is interpreted for the ith observation and quantifies the impact of eliminating a specific observation. Influential Outliers

```
#Influential Outliers

plot(interacmodel2,which=5)
```




```
plot(interacmodel2,pch=18,col="red",which=c(4))
```




```
dataset[cooks.distance(interacmodel2)>0.5,]
```

	PropertyID <int>	Property_Type <chr>	NumberofBuildings <int>	Year_Built <int>	Property_GFA <dbl>	Site_Energy <dbl>
6	8854296	Office	1	1979	17468.0	14092.5
8	8854298	Office	1	1982	85941.0	87455.3
162	10417930	Distribution Center	1	2018	44228.3	101625.5
204	8854296	Office	1	1979	17468.0	13562.7
261	10417930	Distribution Center	1	2018	44228.3	146772.3

5 rows | 1-7 of 25 columns

12.Box-Cox Transformation

Box-Cox Transformation: When the Normality assumption and heteroscedastic assumption fail, we must do Box-Cox transformation.

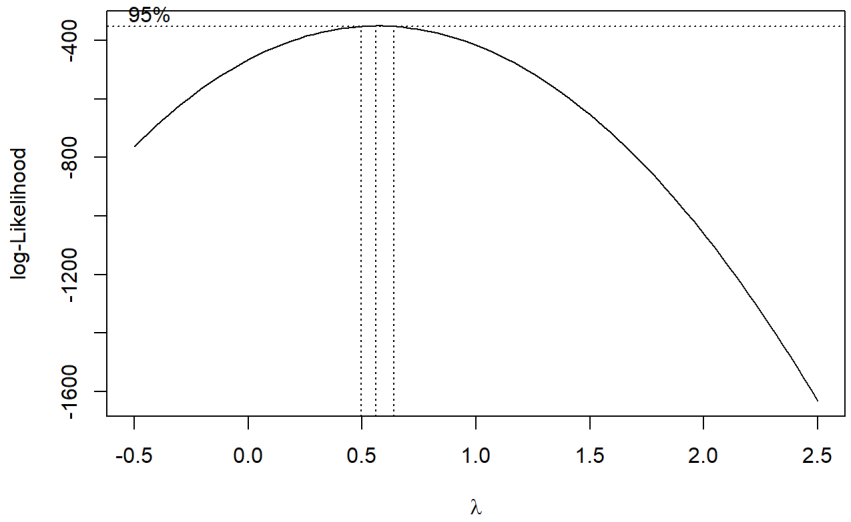
```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.2.3
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:olsrr':
##
##     cement
```

```
bc=boxcox(interacmodel2,lambda=seq(-0.5,3))
```



From performing the Box-Cox Transformation, we have the lambda value that can help us improve our model. We will be discussing this before our final presentation.

```
bestlambda=bc$x[which(bc$y==max(bc$y))]
bestlambda
```

```
## [1] 0.5606061
```

```
interacmodel2 <-lm(Site_Energy~NumberofBuildings+Emissions_CO2+I(Emissions_CO2^2)+I(Emissions_CO2^3)+Natural_Gas+Property_GF
A+Emissions_CO2*Natural_Gas+Emissions_CO2*Property_GFA+Natural_Gas*Property_GFA, data=dataset)
summary(interacmodel2)
```

```
##
## Call:
## lm(formula = Site_Energy ~ NumberofBuildings + Emissions_CO2 +
##     I(Emissions_CO2^2) + I(Emissions_CO2^3) + Natural_Gas + Property_GFA +
##     Emissions_CO2 * Natural_Gas + Emissions_CO2 * Property_GFA +
##     Natural_Gas * Property_GFA, data = dataset)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -967.3  -39.7    2.7   39.2 3224.7
##
## Coefficients:
##                                   Estimate        Std. Error t value
## (Intercept)                 149.718178262441   91.326756015509   1.639
## NumberofBuildings          -200.086431113004   89.100506340703  -2.246
## Emissions_CO2                 5.777410420664    0.154756832968  37.332
## I(Emissions_CO2^2)           -0.000225682523    0.000054376006  -4.150
## I(Emissions_CO2^3)            0.000000007538    0.000000003011   2.504
## Natural_Gas                   0.662884049277    0.008709747855  76.108
## Property_GFA                  0.040574086706    0.012157593104   3.337
## Emissions_CO2:Natural_Gas     0.000019266916    0.000002418796   7.965
## Emissions_CO2:Property_GFA    0.000038513764    0.000004279453   9.000
## Natural_Gas:Property_GFA     -0.000003159993    0.000000238654 -13.241
##                                        Pr(>|t|)
## (Intercept)                             0.10226
## NumberofBuildings                       0.02551 *
## Emissions_CO2                < 0.0000000000000002 ***
## I(Emissions_CO2^2)              0.0000440369703034 ***
## I(Emissions_CO2^3)                      0.01286 *
## Natural_Gas                  < 0.0000000000000002 ***
## Property_GFA                            0.00096 ***
## Emissions_CO2:Natural_Gas    0.0000000000000412 ***
## Emissions_CO2:Property_GFA < 0.0000000000000002 ***
## Natural_Gas:Property_GFA   < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 381.8 on 281 degrees of freedom
## Multiple R-squared:  0.9995, Adjusted R-squared:  0.9995
## F-statistic: 6.15e+04 on 9 and 281 DF,  p-value: < 0.00000000000000022
```

```
bcmodel2=lm((((Site_Energy^0.5606)-1)/0.5606)~NumberofBuildings+Emissions_CO2+I(Emissions_CO2^2)+I(Emissions_CO2^3)+Natural_
Gas+Property_GFA+Emissions_CO2*Natural_Gas+Emissions_CO2*Property_GFA+Natural_Gas*Property_GFA, data=dataset)
summary(bcmodel2)
```

```
##
## Call:
## lm(formula = (((Site_Energy^0.5606) - 1)/0.5606) ~ NumberofBuildings +
##     Emissions_CO2 + I(Emissions_CO2^2) + I(Emissions_CO2^3) +
##     Natural_Gas + Property_GFA + Emissions_CO2 * Natural_Gas +
##     Emissions_CO2 * Property_GFA + Natural_Gas * Property_GFA,
##     data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.483  -8.418   2.013  11.057 102.019
##
## Coefficients:
##                                  Estimate       Std. Error t value
## (Intercept)                72.2965213788745  4.0134895740509  18.013
## NumberofBuildings          -5.7714759365160  3.9156537343814  -1.474
## Emissions_CO2               0.2468022044408  0.0068010182638  36.289
## I(Emissions_CO2^2)         -0.0000517350634  0.0000023896341 -21.650
## I(Emissions_CO2^3)          0.0000000026758  0.0000000001323  20.223
## Natural_Gas                 0.0082225024113  0.0003827627711  21.482
## Property_GFA               -0.0009936064400  0.0005342834378  -1.860
## Emissions_CO2:Natural_Gas   0.0000000603862  0.0000001062976   0.568
## Emissions_CO2:Property_GFA  0.0000011673934  0.0000001880669   6.207
## Natural_Gas:Property_GFA   -0.0000000541350  0.0000000104880  -5.162
##                                      Pr(>|t|)
## (Intercept)                < 0.0000000000000002 ***
## NumberofBuildings                      0.142
## Emissions_CO2              < 0.0000000000000002 ***
## I(Emissions_CO2^2)         < 0.0000000000000002 ***
## I(Emissions_CO2^3)         < 0.0000000000000002 ***
## Natural_Gas                < 0.0000000000000002 ***
## Property_GFA                           0.064 .
## Emissions_CO2:Natural_Gas              0.570
## Emissions_CO2:Property_GFA        0.00000000193 ***
## Natural_Gas:Property_GFA          0.00000046380 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.78 on 281 degrees of freedom
## Multiple R-squared:  0.993,  Adjusted R-squared:  0.9928
## F-statistic:  4438 on 9 and 281 DF,  p-value: < 0.00000000000000022
```

After box-cox transformation the Residual standard error improved from 381.8 to 16.78. We can check for normality and heteroscedasticity again.

```
#Hypothesis for Heteroscedasticity
#Null Hypothesis, H0: Heteroscedasticity is not present
#Alternate Hypothesis, Ha: Heteroscedasticity is present
bptest(bcmodel2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  bcmodel2
## BP = 126, df = 9, p-value < 0.00000000000000022
```

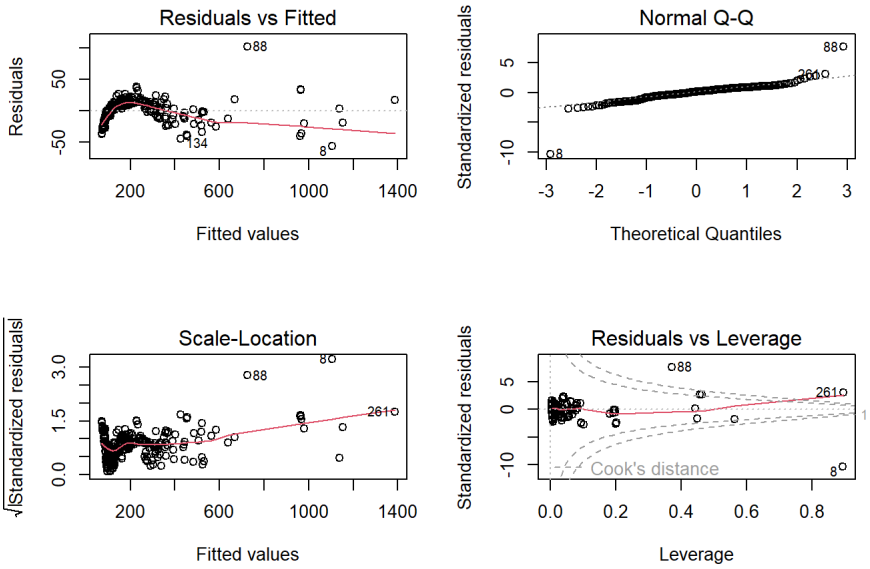
```
#Testing for Normality
shapiro.test(residuals(bcmodel2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(bcmodel2)
## W = 0.94256, p-value = 0.000000003178
```

From the output, the Breusch-Pagan test that result from the Box-Cox model, the p-value < 0.00000000000000022 which is <0.05, indicating that we reject the null hypothesis. Therefore, the test provide evidence to suggest that homoscedasticity does not exist but heteroscedasticity does exist in the model even after improvement with box-cox transformation.Additinally, Shapiro-Wilk normality test also confirms that the residuals are not normally distributed as the p-value=0.000000003178 >0.05. So Box-Cox Tranformation is helpful for our model to improve model by reducing RMSE, but it did not help with normality and homoscedasticity.

Plots of the model after Box Cox transformation

```
par(mfrow=c(2,2))
plot(bcmodel2)
```



In our case, the two interaction terms Emissions_CO2:Natural_Gas and Emissions_CO2:Property_GFA became non-significant after applying the Box-Cox transformation, which reduced the accuracy of our predictions. Therefore, we decided to revert to the original model for our final model instead of the model after the Box-Cox transformation.

RESULTS

Final Model and Interpreting Coefficients

After successfully conducting all these tests, our final best fitted model including main effects, interaction terms and higher order terms is expressed as:

$$\begin{aligned} Y_{SiteEnergy} &= \beta_0 + \beta_1 X_{EmissionsCO2} + \beta_2 X^2_{EmissionsCO2} + \beta_3 X^3_{EmissionsCO2} + \beta_4 X_{NaturalGas} + \beta_5 X_{PropertyGFA} \\ &+ \beta_6 X_{NumberofBuildings} + \beta_7 X_{EmissionsCO2*NaturalGas} + \beta_8 X_{EmissionsCO2*PropertyGF} + \beta_9 X_{NaturalGas*PropertyGFA} + \epsilon \end{aligned}$$

Final model expanded with all terms

$$\begin{aligned} \widehat{Y_{SiteEnergy}} &= 149.7182 + 5.7774 X_{EmissionsCO2} - 0.0002257 X^2_{EmissionsCO2} + 0.000000007 X^3_{EmissionsCO2} + 0.6629 X_{NaturalGas} \\ &+ 0.0406 X_{PropertyGFA} - 200.0864 X_{NumberofBuildings} + 0.0000192 X_{EmissionsCO2*NaturalGas} \\ &+ 0.0000385 X_{EmissionsCO2*PropertyGFA} - 0.0000031 X_{NaturalGas*PropertyGFA} \end{aligned}$$

Adjusted R-square and RMSE of Best Fitted Model

The adjusted R-squared, $R^2_{adj} = 0.9995$ implies that 99.95% of the variation in the response variable site energy is explained by this model containing the predictors emissions CO2, natural gas, property GFA, number of buildings, and the interactions trems $EmissionsCO2 * NaturalGas$, $EmissionsCO2 * PropertyGFA$, $NaturalGas * PropertyGFA$ as well as the second order term and third order term of emissions CO2.

$RMSE = 381.8$, this value indicates that the standard deviation of the unexplained variation in estimation of response variable site energy is 381.8 GJ.

Final model with EmissionsCO2 terms collected

$$\begin{aligned} \widehat{Y_{SiteEnergy}} &= 149.7182 + (5.7774 + 0.0000192 X_{NaturalGas} + 0.0000385 X_{PropertyGFA}) X_{EmissionsCO2} \\ &- 0.0002257 X^2_{EmissionsCO2} + 0.000000007 X^3_{EmissionsCO2} + 0.6629 X_{NaturalGas} \\ &+ 0.0406 X_{PropertyGFA} - 200.0864 X_{NumberofBuildings} \\ &- 0.0000031 X_{NaturalGas*PropertyGFA} \end{aligned}$$

Final model with NaturalGas terms collected

$$\begin{aligned} \widehat{Y_{SiteEnergy}} &= 149.7182 + 5.7774 X_{EmissionsCO2} - 0.0002257 X^2_{EmissionsCO2} + 0.000000007 X^3_{EmissionsCO2} \\ &+ (0.6629 + 0.0000192 X_{EmissionsCO2} - 0.0000031 X_{PropertyGFA}) X_{NaturalGas} \\ &+ 0.0406 X_{PropertyGFA} - 200.0864 X_{NumberofBuildings} \\ &+ 0.0000385 X_{EmissionsCO2*PropertyGFA} \end{aligned}$$

Final model with PropertyGFA terms collected

$$\begin{aligned}\widehat{Y_{SiteEnergy}} &= 149.7182 + 5.7774X_{EmissionsCO2} - 0.0002257X^2_{EmissionsCO2} + 0.000000007X^3_{EmissionsCO2} + 0.6629X_{NaturalGas} \\ &+ (0.0406 + 0.0000385X_{EmissionsCO2} - 0.0000031X_{NaturalGas})X_{PropertyGFA} \\ &- 200.0864X_{NumberofBuildings} + 0.0000192X_{EmissionsCO2*NaturalGas}\end{aligned}$$

Interpretation of Coefficients

There are four $\beta_i(i = EmissionsCO2, PropertyGFA, NaturalGas, NumberofBuildings)$ coefficients in our final model.

Explanations of the relationship between each coefficients and the response variable site energy are given below.

Note that in the following interpretation we ignore the higher order terms, due to their complexity.

$$\widehat{\beta_{EmissionsCO2}} = 5.7774 + 0.0000192X_{NaturalGas} + 0.0000385X_{PropertyGFA}$$

This equation value indicates that the effect of emissions CO2 on site energy (in GJ) changes by natural gas and property GFA. While all other main effects are held constant,increasing emissions CO2 by 1 metric tons leads to an increase in site energy by $5.7774 + 0.0000192X_{NaturalGas} + 0.0000385X_{PropertyGFA}$ GJ.

$$\widehat{\beta_{NaturalGas}} = 0.6629 + 0.0000192X_{EmissionsCO2} - 0.0000031X_{PropertyGFA}$$

This equation value indicates that the effect of natural gas on site energy (in GJ) changes by emissions CO2 and property GFA. While all other main effects are held constant,increasing natural gas by 1 GJ leads to an increase in site energy by $0.6629 + 0.0000192X_{EmissionsCO2} - 0.0000031X_{PropertyGFA}$ GJ.

$$\widehat{\beta_{PropertyGFA}} = 0.0406 + 0.0000385X_{EmissionsCO2} - 0.0000031X_{NaturalGas}$$

This equation value indicates that the effect of property GFA on site energy (in GJ) changes by emissions CO2 and natural gas. While all other main effects are held constant,increasing the property gross floor area by 1 m^2 leads to an increase in site energy by $0.0406 + 0.0000385X_{EmissionsCO2} - 0.0000031X_{NaturalGas}$ GJ.

$$\widehat{\beta_{NumberofBuildings}} = -200.0864$$

This equation value indicates that while all other main effects are held constant,increasing 1 buildings on the property leads to a decrease in site energy by -200.0864 GJ.

Predicted Site Energy Use

We use the model to make a prediction of energy usage as an example. In a scenario where the number of buildings is 2, the property gross floor area is 9300 m^2, emissions CO2 is 760 Metric Tons, and natural gas usage is 9500 GJ.

Checking Extrapolation:

```
library(mosaic)
```

```
## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2
```

```
##
## The 'mosaic' package masks several functions from core packages in order to add
## additional features.  The original behavior of these functions should not be affected by this.
```

```
##
## Attaching package: 'mosaic'
```

```
## The following objects are masked from 'package:dplyr':
##
##     count, do, tally
```

```
## The following object is masked from 'package:Matrix':
##
##     mean
```

```
## The following object is masked from 'package:ggplot2':
##
##     stat
```

```
## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var
```

```
## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

```
favstats(~NumberofBuildings, data=dataset)
```

min	Q1	median	Q3	max	mean	sd	n	missing
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
1	1	1	1	3	1.061856	0.2809258	291	0

1 row

```
favstats(~Emissions_CO2, data=dataset)
```

min	Q1	median	Q3	max	mean	sd	n	missing
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
16.8	123.4	240.8	695.6	10999.9	721.9388	1398.523	291	0

1 row

```
favstats(~Natural_Gas, data=dataset)
```

min	Q1	median	Q3	max	mean	sd	n	missing
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
13.2	886.6	1594.5	5157.65	145089	5271.179	13827.86	291	0

1 row

```
favstats(~Property_GFA, data=dataset)
```

min	Q1	median	Q3	max	mean	sd	n	missing
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
204.4	1101.1	1806.5	4189.8	85941	4797.688	10219.44	291	0

1 row

Extrapolation does not exist.

$$

$$\begin{aligned}\widehat{Y_{SiteEnergy}} &= 149.7182 + 5.7774X_{EmissionsCO2} - 0.0002257X^2_{EmissionsCO2} + 0.000000007X^3_{EmissionsCO2} + 0.6629X_{NaturalGas} \\ &+ 0.0406X_{PropertyGFA} - 200.0864X_{NumberofBuildings} + 0.0000192X_{EmissionsCO2*NaturalGas} \\ &+ 0.0000385X_{EmissionsCO2*PropertyGFA} - 0.0000031X_{NaturalGas*PropertyGFA} \\ \widehat{Y_{SiteEnergy}} &= 149.7182 + 5.7774 * (760) - 0.0002257 * (760)^2 + 0.000000007(760)^3 + 0.6629 * (9500) \\ &+ 0.0406 * (9300) - 200.0864 * (2) + 0.0000192 * (760 * 9500) \\ &+ 0.0000385 * (760 * 9300) - 0.0000031 * (9500 * 9300) \\ &= 10825.06491\end{aligned}$$

$$

In a scenario where the number of buildings is 2, the property gross floor area is 9300 m^2, emissions CO2 is 760 Metric Tons, and natural gas usage is 9500 GJ. In this scenario, the predicted site energy usage results are 10825.06491 GJ.

```
interacmodel3 <-lm(Site_Energy~NumberofBuildings+Emissions_CO2+Natural_Gas+Property_GFA+Emissions_CO2*Natural_Gas+Emissions_
CO2*Property_GFA+Natural_Gas*Property_GFA, data=dataset)


newdata = data.frame(NumberofBuildings= 2, Emissions_CO2=760, Property_GFA=9300, Natural_Gas = 9500  )
predict(interacmodel3,newdata,interval="predict")
```

```
##        fit      lwr      upr
## 1 10916.42 10121.64 11711.19
```

From the R command predict, with 95% confidence interval, the site energy usage is between 10121.64 GJ to 11711.19 GJ when the number of buildings is 2, the property gross floor area is 9300 m^2, emissions CO2 is 760 Metric Tons, and natural gas usage is 9500 GJ. Our result of 10825.06491 GJ which lies in the 95% confidence interval. Thus, it verifies our result.

DISCUSSION:

The site energy use can be predicted using a few variables that are the most significant. We studied the effects of NumberofBuildings, Emissions_CO2, Natural_Gas, Property_GFA on Site Energy variable. We initially expected that all variables from the original model would be responsible for the Site Energy value, but turned out not all were significant enough. So we conducted tests and assumptions to get rid of the insignificant variables, and then formed our final model as mentioned above. The issue we faced initially was that our R^2adj value was very close to 1, and that there was almost no scope for improvement. We, however, went ahead and conducted the tests and tried getting the R^2adj value from 0.9994 or higher, for practice purposes as suggested by Ms. Thuntida.

In this Statistical Analysis, we found the best model for prediction of Site Energy use. However, below are a few things we can do to improve our model and as a future scope: 1. Use of Time Series regression to check if model is improving. 2. Use All possible regression and drop more variables to get our best model. 3. Get the ANOVA table and interpret the meaning of the values. 4. By using lambda value from box-cox, we can improve our model.

REFERENCES:

1. Canada Energy Regulator-Energy in Canada [Online]. Available at: https://www.cer-rec.gc.ca/en/about/publications-reports/annual-report/2018/energy-in-canada.html# (https://www.cer-rec.gc.ca/en/about/publications-reports/annual-report/2018/energy-in-canada.html#):~:text=Canada%20is%20currently%20ranked%20the,and%20future%20needs%20of%20Canadians (Accessed November 6, 2022)
2. Sustainable Building Partnership Program, City of Calgary [Online]. Available at: https://www.calgary.ca/development/sustainable-building-partnership.html (https://www.calgary.ca/development/sustainable-building-partnership.html) (Accessed November 6, 2022)
3. Canada Energy Regulator - Provincial and Territorial Energy Profiles – Alberta [Online]. Available at: https://www.cer-rec.gc.ca/en/data-analysis/energy-markets/provincial-territorial-energy-profiles/provincial-territorial-energy-profiles-alberta.html (https://www.cer-rec.gc.ca/en/data-analysis/energy-markets/provincial-territorial-energy-profiles/provincial-territorial-energy-profiles-alberta.html) (Accessed November 6, 2022)
4. Building Energy Benchmarking - City of Calgary [Online]. Available at: https://data.calgary.ca/Environment/Building-Energy-Benchmarking-City-of-Calgary/8twd-upbv (https://data.calgary.ca/Environment/Building-Energy-Benchmarking-City-of-Calgary/8twd-upbv) (Accessed November 1, 2022)
5. Alberta Agriculture, Forestry and Rural Economic Development, Alberta Climate Information Service. Current and Historical Alberta Weather Station Data [Online]. Available at: https://acis.alberta.ca/acis/weather-data-viewer.jsp (https://acis.alberta.ca/acis/weather-data-viewer.jsp) (Accessed November 5, 2022)
6. University of Calgary Utilites, Thermal Comfort [Online]. Available at: https://www.ucalgary.ca/facilities/thermal-comfort (https://www.ucalgary.ca/facilities/thermal-comfort) (Accessed November 5, 2022)