NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

# MH3511 Data Analysis with Computer

## Group Project

Energy Performance of the Building Sector in Singapore

*Abstract:*
*Climate change has caused extreme weather conditions worldwide and is a prevalent issue in today's society. In Singapore, the building sector consumes up to 38% of the nation's electricity, which definitely forms a significant part of Singapore's energy consumption. It is thus important for Singapore to formulate policies to improve the energy efficiency of our buildings. One of them is the use of Green Mark standard to assess the energy performance of buildings and awarding the buildings performing well with a Green Mark status, with different levels of ratings in accordance to the level of performance. Hence, we would like to determine the relationship between energy use intensity of buildings and the buildings' features through basic data analysis techniques, to find out which features improve their energy efficiency, enabling them to be awarded the Green Mark status.*

# Table of Contents

**1. Introduction** **2**

**2. Data Description** **3**

**3. Description and Cleaning of the Dataset** **3**

3.1 Summary Statistics for the Main Variable of Interest : 2018 Energy Use Intensity, *X2018energyusintensity* 3

3.2 Summary Statistics for other variables 5

3.2.1 Building Type, *buildingtype* 5

3.2.2 Green Mark Status, *greenmarkstatus* 5

3.2.3 Green Mark Rating, *greenmarkrating* 6

3.2.6 Gross Floor Area, *grossfloorarea* 7

3.2.7 2017 Energy Use Intensity, *X2017energyuseintensity* 7

3.3 Final Dataset for Analysis 8

**4. Statistical Analysis** **8**

4.1 Correlation between log(*X2018energyusintensity*) and other Continous Variable 8

4.2 Statistical Tests 9

4.2.1 Does Energy Use Intensity reduce before Greenmark Status is awarded in 2018? 9

4.2.2 Relation between Energy Use Intensity (*X2018energyusintensity*) and Building Size (*buildingsize*) 10

4.2.3 Relation between Energy Use Intensity in 2018(*X2018energyusintensity*) and Gross Floor Area(*grossfloorarea*) 12

4.2.4 Relation between Energy Use Intensity(*X2018energyuseintensity*) and Green Mark Rating (*greenmarkrating*) 13

4.3.5 Relation between energy use intensity(*X2018energyuseintensity*) and building type (*buildingtype*) 15

**5. Conclusion and Discussion** **17**

**6. Appendix** **18**

# 1. Introduction

Climate change has caused extreme weather conditions to be experienced worldwide. We all have a role to play in mitigating climate change by reducing carbon emissions. Singapore's building sector consumes up to 38% of the nation's electricity. With a focus on addressing the environmental impact caused by buildings, the Building and Construction Authority (BCA) of Singapore formulates and charts green building policies to track and improve the energy efficiency of the built environment in Singapore under our Green Building Master plan.

The Green Mark standard is an assessment of environmental performance of a building for thermal, lighting, indoor air quality and energy efficiency. Green mark points are awarded for incorporating sustainable green features and practices. There are two main groupings of assessment criteria :

1. Energy Efficiency (where a minimum of 30 points must be obtained)
2. Other Green requirements - Water Efficiency, Indoor Environmental Quality, Sustainable Operation & Management (where a minimum of 20 points must be obtained)

In our project, a dataset containing the Energy Use Intensity (EUI) in year 2017, 2018 for different types of buildings and the respective Green Mark Rating is used. Based on this dataset we want to answer the following questions around Building Energy use intensity :

1. Does energy use intensity reduce before greenmark status is awarded in 2018
2. Does building size affect the energy use intensity
3. Does the gross floor area affect the energy use intensity
4. Does greenmark rating affect the energy use intensity
5. Does building type affect the energy use intensity

This report will cover the data descriptions and analysis using R language. For each of our research objectives, we performed statistical analysis and drew conclusions in the most appropriate approach, together with explanations and elaborations.

# 2. Data Description

The dataset, titled "Annual Energy Consumption from Singapore Buildings", is obtained from the online data science community kaggle.com. The original data consists of 1 csv data frame titled "listing-of-building-energy-performance-data-for-commercial-buildings.csv". The dataset consisted of 11 columns and 1244 rows. The dataset was originally posted on data.gov.sg, the official government database records for Singapore and is open to the public for study and research.

Before proceeding to data analysis we first performed a preliminary data cleaning to ensure that:

- Irrelevant columns such as "buildingname", "buildingaddress" and "voluntarydisclosure" were removed.
- All rows with NA data except "greenmarkrating" and "greenmarkyearaward" were eliminated

- Redundant information was removed such as “legislated” in “greenmarkrating” column

After all the preparation, 976 observations with 8 variables are retained for analysis:

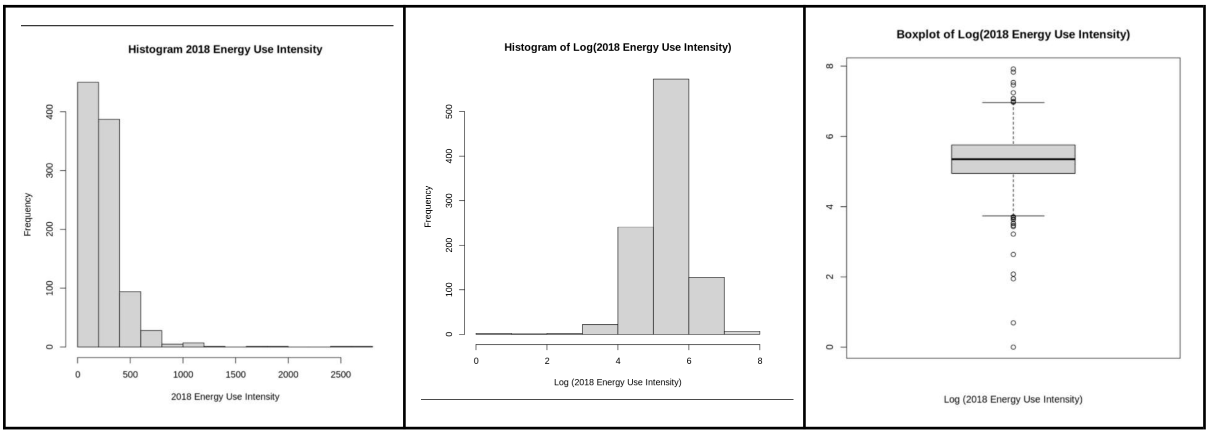
1. buildingtype : Type of Building (Office, Retail, Mixed Development, Hotel etc.)
2. greenmarkstatus : Green Mark Status of the Building (Yes or No)
3. greenmarkrating : Green Mark Rating Level of the Building (Platinum, Gold, GoldPlus, Certified, NA)
4. greenmarkyearaward: Year the Building was Awarded the Green Mark Status
5. buildingsize : Categorical Size of the Building (Large, Small)
6. grossfloorarea : Gross Floor Area of the Building
7. X2017energyuseintensity : Energy Use Intensity of the Building in 2017
8. X2018energyusintensity : Energy Use Intensity of the Building in 2018

# 3. Description and Cleaning of the Dataset

In this section, we shall look into the data in more detail. Each variable is investigated to look for possible outliers, and/or to perform a transformation to avoid highly skewed data.

## 3.1 Summary Statistics for the Main Variable of Interest : 2018 Energy Use Intensity

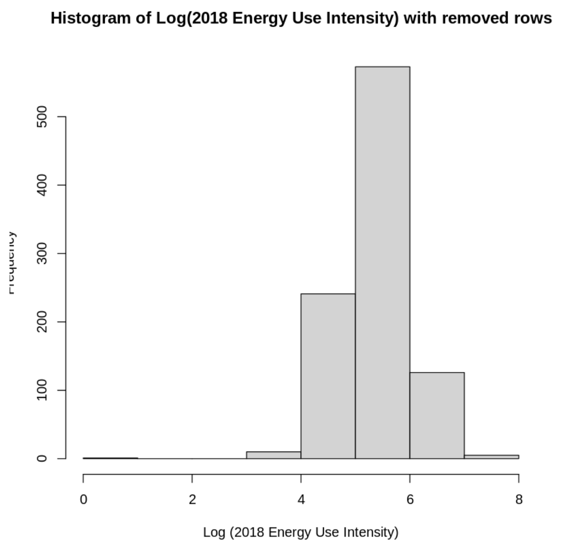
The following plots show the overall distribution of the variable *X2018energyusintensity*



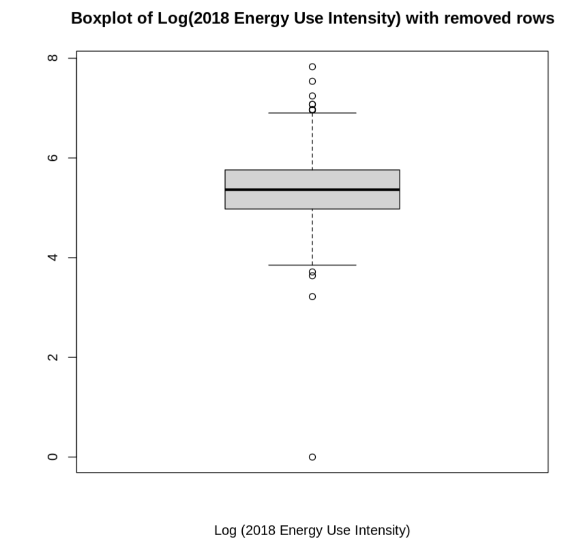
It appears that the variable *X2018energyuseintensity* is highly left skewed, hence we apply a log-transofrmation (base e) to the variable. The log transformed data appears to have quite a few outlying values at both the left and right tail. Upon further investigation, we notice that there are a few values with Energy Use Intensity extremely high even though they are Green Mark Buildings and few values with Energy Use Intensity extremely low even though they are not Green Mark Buildings. Therefore, we removed these values, approximately 2.07% of the data.

The Histogram and Boxplot of the log-transformed variable, with the outliers removed are shown below with summary statistics. The dataset is now more symmetric but it still has a few outliers - this could be due to the buildings performing extremely well in their Energy Use Consumption after receiving the Green Mark Award as well as buildings doing poorly in their Energy Use Consumption without having received the Green Mark Award.

We shall proceed to the next section with this trimmed dataset.



Histogram of Log(2018 Energy Use Intensity) with removed rows

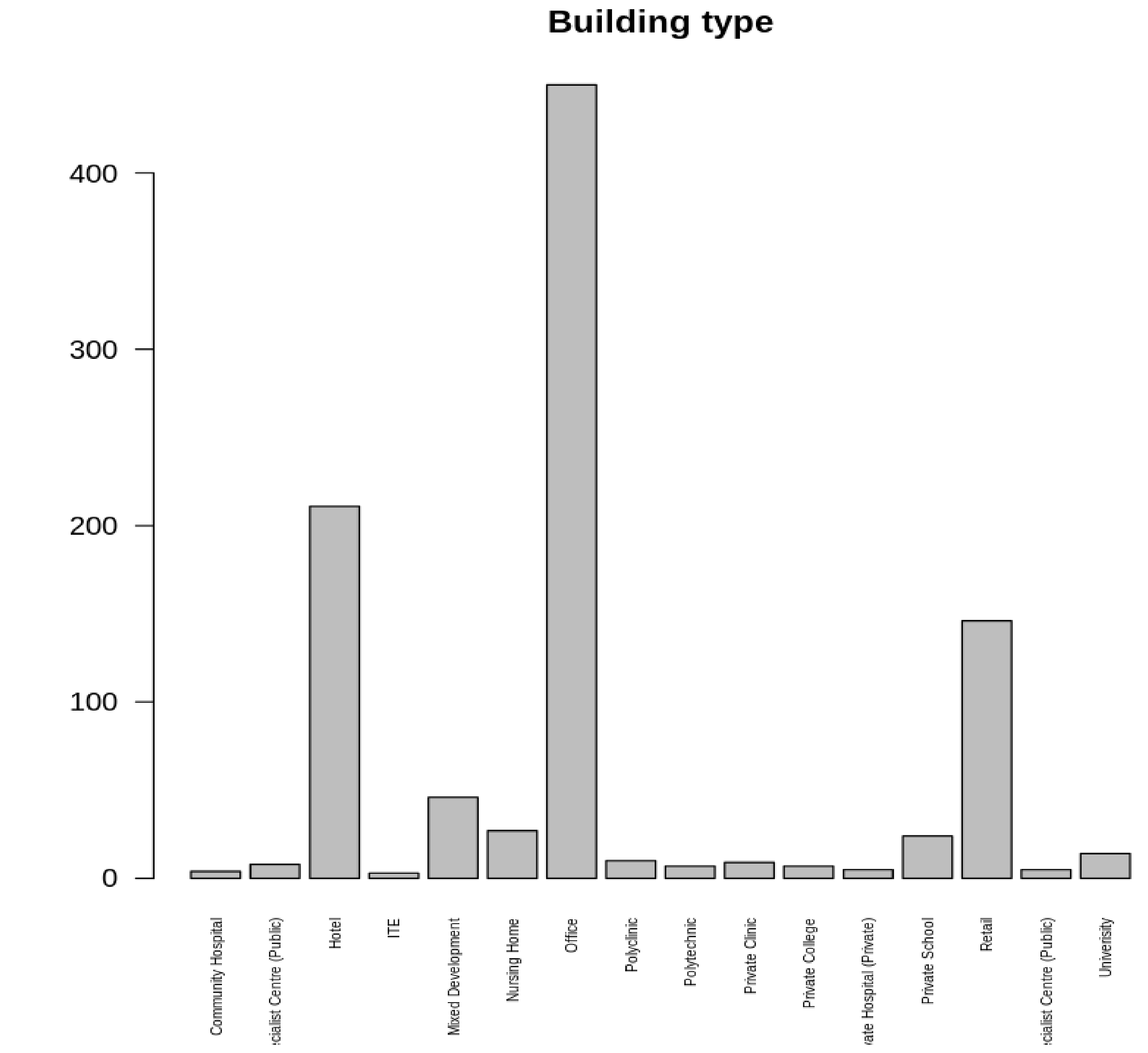


Boxplot of Log(2018 Energy Use Intensity) with removed rows

```
Min.  1st Qu. Median  Mean   3rd Qu.  Max.
1.0   145.0   213.5   258.4   317.2  2516.0
```

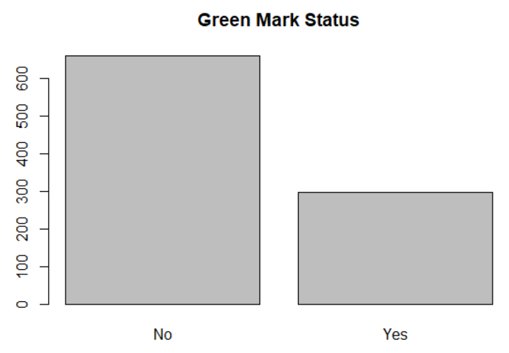
## 3.2 Summary Statistics for other variables

### 3.2.1 Building Type, *buildingtype*



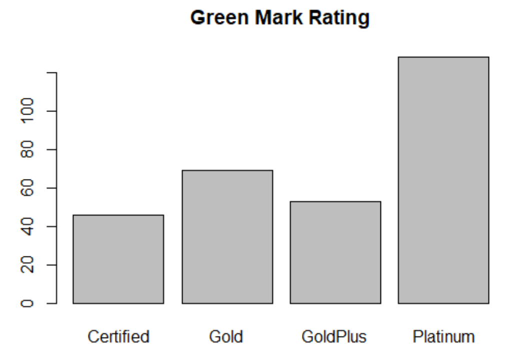
- Most of the buildings fall under the categories “Office”, “Hotel” and “Retail”.
- Removed TCM Clinic since it only had one value and was redundant

3.2.2 Green Mark Status, *greenmarkstatus*



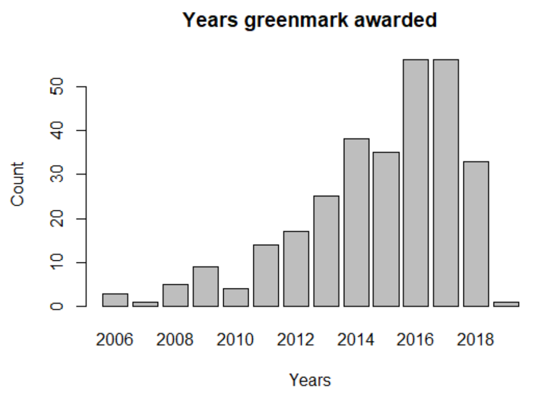
- No outliers were removed.
- A larger number of buildings were not awarded the green mark status.

### 3.2.3 Green Mark Rating, *greenmarkrating*



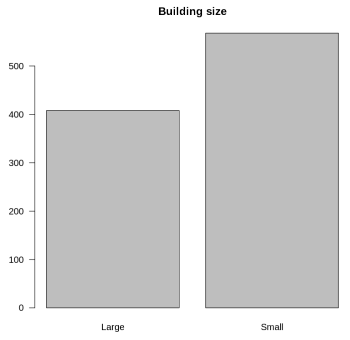
- No outliers were removed.
- Majority of the buildings that were awarded the green mark status had a platinum rating, which is the highest rating.

### 3.2.4 Green Mark Year Award, *greenmarkyearaward*

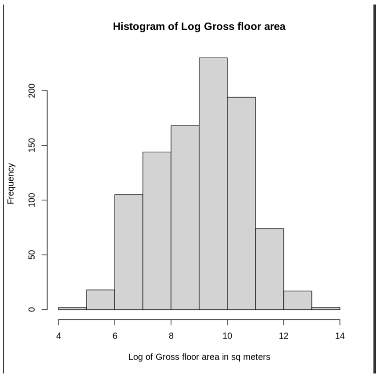
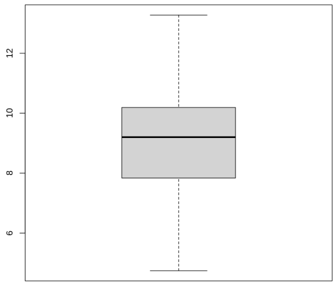


- No outliers were removed
- There seems to be more green mark awarded in the years 2016-2017

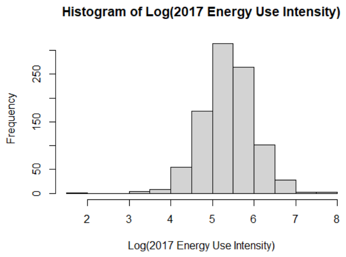
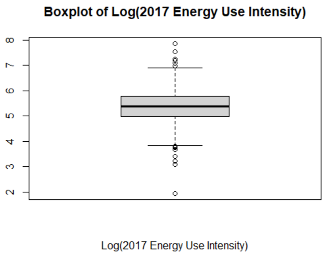
### 3.2.5 Building Size, *buildingsize*

| | |
|---|---|
|  | - No outliers were removed<br>- There are more smaller sized buildings |

### 3.2.6 Gross Floor Area, *grossfloorarea*

| | | |
|---|---|---|
|  |  | - The log-transformation (base e) is applied<br>- One outlier log(grossfloorarea) > 14 was removed |

### 3.2.7 2017 Energy Use Intensity, *X2017energyuseintensity*

| | | |
|---|---|---|
|  |  | - The log-transformation (base e) is applied<br>- No outliers were removed |

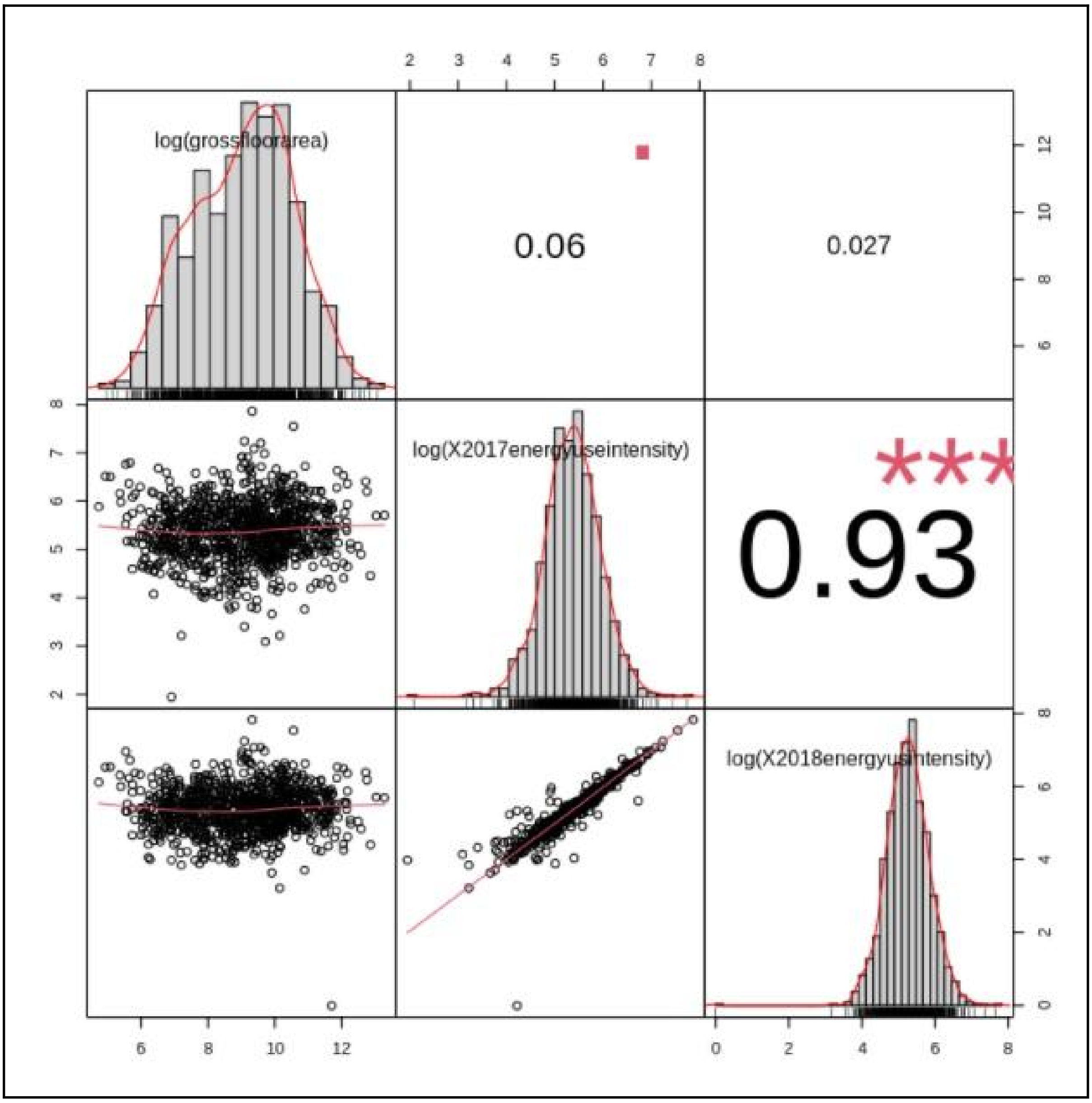
### 3.3 Final Dataset for Analysis

Based on the above analysis, the dataset is further reduced to 954 observations with the suggested transformations. Namely, log-transformation (base e) to be applied to *grossfloorarea, X2018energyusintensity* and *X2017energyuseintensity.*

# 4. Statistical Analysis

For all statistical tests, we will be using a confidence level of 0.05.

### 4.1 Correlations between *log(X2018energyusintensity)* and other Continous Variables



Scatter plots and correlation coefficients are useful in studying the possible linear relationships between the gross floor area and the Energy Use Intensity in 2017 and 2018.

From the plots, it appears that *log(grossfloorarea)* is not well related to either of the Energy Use Intensities in 2017 or 2018.

Among the Energy Use Intensities, there is an interesting observation from this tabulation :

- *log(X2018energyusintensity)* and *log(X2017energyuseintensity)* are highly correlated (r = 0.93)
    - This is as expected as the two columns are the Energy Use Intensity in consecutive years.
- *log(grossfloorarea)* is positively related to both *log(X2018energyusintensity)* (r = 0.027) and *log(X2017energyuseintensity)* (r = 0.06)
    - This tells us that as floor area increases there is some correlation with the increase in the Energy Use Intensity of the building.

We shall perform some statistical tests to confirm some of our observations in the next section.

## 4.2 Statistical Tests

### 4.2.1 Does Energy Use Intensity(*X2017energyuseintensity* ,*X2018energyusintensity*) reduce before Greenmark Status is awarded in 2018(*greenmarkyearaward* )?

The variable *greenmarkyearaward* indicates the year in which a building achieved a green mark status. In order to verify whether a relationship between energy use intensity and green mark status being awarded in 2018 exists, we need to define a few more attributes:-

*energyuseintensitychange :* The change in energy use intensity from 2017 to 2018

*energyusedecrease :* A dichotomous categorical variable which indicates whether energy use intensity decreases from 2017 to 2018

In order to investigate the above, we first create a subset of the dataset under the condition that the green mark status has either been awarded in 2018 or not at all. We then define the two new attributes defined above as columns in this subset. We the use a chi-squared test to deduce whether the green mark status awarded in 2018 is independent of the decrease in energy use intensity under the following null and alternative hypotheses:-

$H_0$ : green mark status awarded in 2018 is independent of energy
use intensity decrease from 2017 to 2018

$H_1$ : green mark status awarded in 2018 is not independent of energy
use intensity decrease from 2017 to 2018

```
X-squared = 7.5698, df = 1, p-value = 0.005936
```

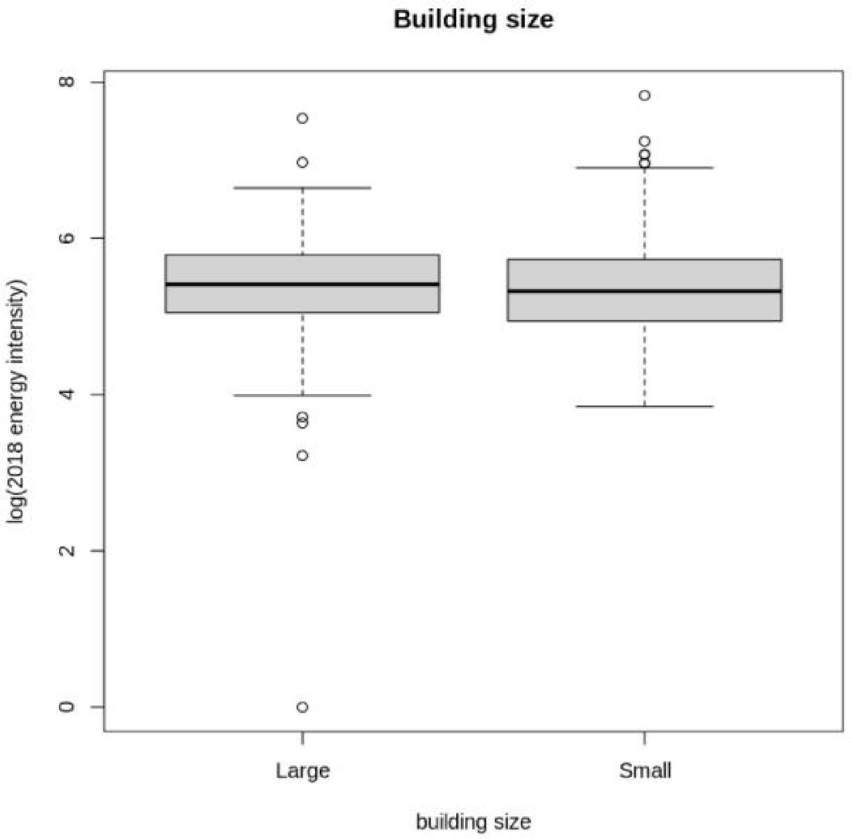
The chi-squared test returns a p-value of 0.005936, which shows that green mark status awarded in 2018 is dependent on energy use intensity decrease from 2017 to 2018 Now we find Pearson's Correlation coefficient to determine whether these variables are positive or negatively correlated.

```
t = 2.7625, df = 690, p-value = 0.005889

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:
 0.03029544 0.17773427

sample estimates:
      cor
0.1045895
```

Pearson's product-moment correlation returns a correlation coefficient of 0.1045895. Hence, we can say that the two variables are positively correlated.

### 4.2.2 Relation between Energy Use Intensity (*X2018energyusintensity*) and Building Size *(buildingsize)*

In this section we try to answer "Does building size affect the energy use intensity.
Since, building size (*buildingsize*) is a categorical variable, we use F-test to compare the variances of the *buildingsize* by log(*X2018energyusintensity*). After, which we use two sample t-test, taking into account whether or not the variance is the same of Large and Small buildings, to compare the difference in means of *buildingsize* by log(*X2018energyusintensity*).
The following is a box plot of *buildingsize* by log(*X2018energyusintensity)*.



Looking at the boxplot the mean of the building sizes appears to be the same. However, the spread does not seem very similar as there are a lot more extreme outliers present on small building size when compared to large. Hence, we use F-test to compare the variances of the *buildingsize* by *X2018energyusintensity*.

$H_0$ : var(small) = var(large) against $H_1$ : not all var are equal

```
F = 1.0012, num df = 400, denom df = 552, p-value = 0.986
```

F-test returns a p-value of 0.986 which shows that the variances are not significantly different for the building sizes. We now conduct a t-test taking into account that the variances of large and small buildings are not significantly different.

$H_0$ : $\mu_{small} = \mu_{large}$ against $H_1$ : not all means are equal

```
t = 1.3418, df = 952, p-value = 0.18
```

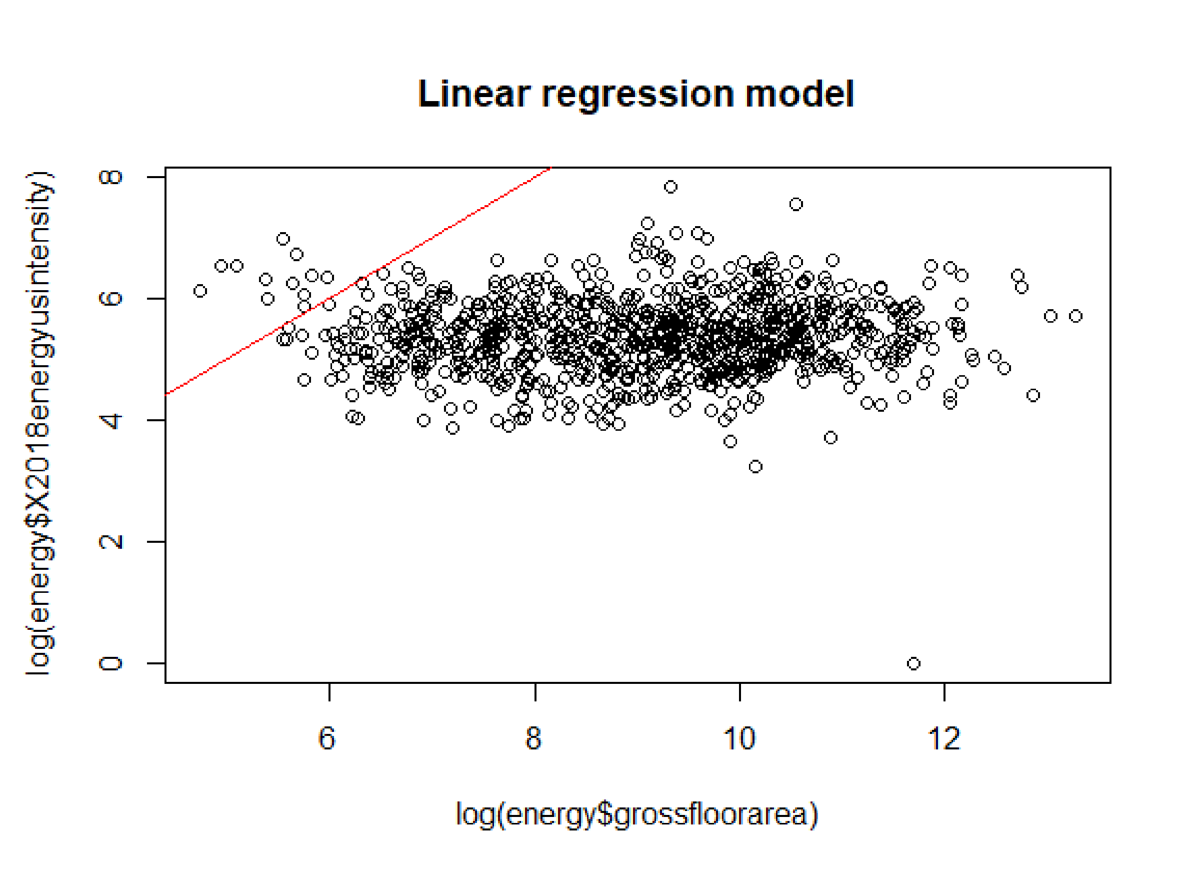
t-test returns a p-value of 0.18 which show that the means are not significantly different for the building sizes. Therefore, we conclude that the energy use intensity does not vary based on whether the building size (*buildingsize*) is small or large.

### 4.2.3 Relation between Energy Use Intensity in 2018 (*X2018energyusintensity*) and Gross Floor Area (*grossfloorarea*)

In this section, we determine whether the energy use intensity in 2018 is dependent on the gross floor area of the buildings. We perform a simple linear regression between *X2018energyusintensity* and log(*grossfloorarea*).

The regression model provides a p-value of 0.405 which indicates a non-statistically significant relationship between log(*X2018energyusintensity*) and log(*grossfloorarea*) at 0.05 level of significance. The R-squared value for this model is 0.0007286, which is less than 1%. The low R-squared value indicates that log(*grossfloorarea*) does not explain much about the variations in log(*X2018energyusintensity*). This further supports what we see in Section 4.1, whereby the linear correlation between log(*X2018energyusintensity*) and log(*grossfloorarea*) is only 0.027.

Therefore, we conclude that the gross floor area does not statistically affect the energy use intensity of buildings.The variable log(*grossfloorarea)* explains less than 1% variation in the energy energy use intensity in 2018.

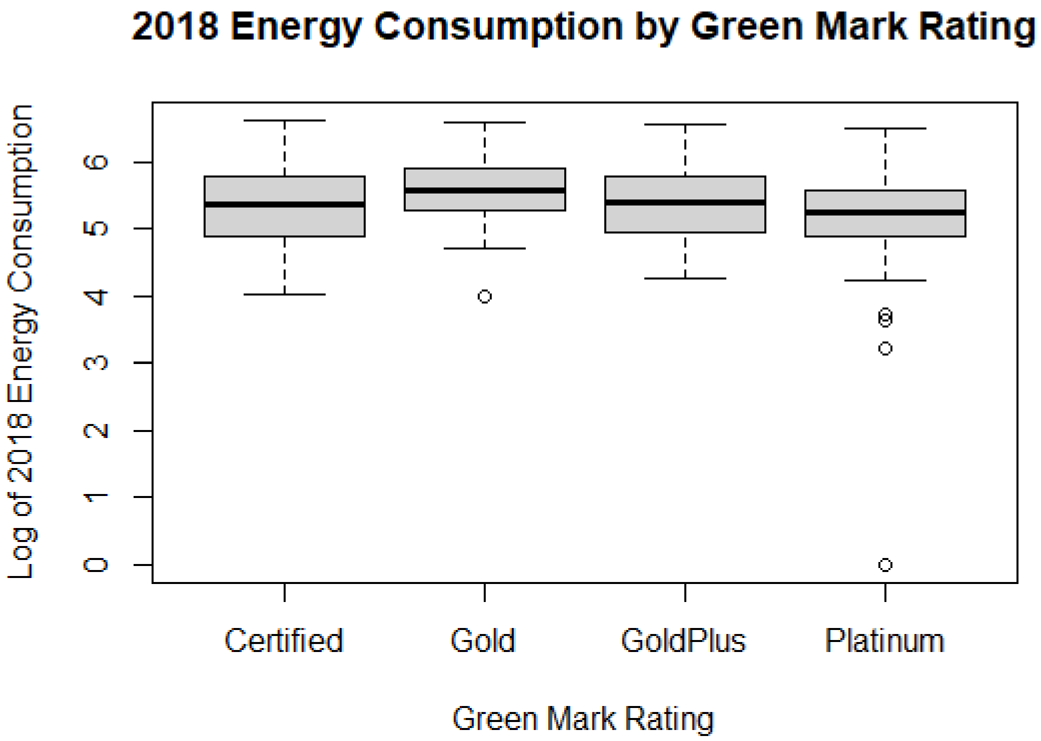


```
Call:
lm(formula = log(X2018energyusintensity) ~ log(grossfloorarea),
    data = energy)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.3950 -0.3890 -0.0039  0.3892  2.4608 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)          5.26989    0.11771  44.769   <2e-16 ***
log(grossfloorarea)  0.01070    0.01284   0.833    0.405    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6215 on 952 degrees of freedom
Multiple R-squared:  0.0007286, Adjusted R-squared:  -0.000321 
F-statistic: 0.6942 on 1 and 952 DF,  p-value: 0.405
```

### 4.2.4 Relation between Energy Use Intensity (*X2018energyuseintensity*) and Green Mark Rating (*greenmarkrating*)

A green mark rating is given to a building that has been awarded the green mark status.

Here we look at the boxplot of log(*X2018energyusintensity)* by (*greenmarkrating*).



There does not seem to be any significant difference in the means of the
log of the energy use intensity and there only seems to be a slight difference in their spread.

Since green rating is a categorical variable and there are more than 2 samples, we use ANOVA to compare the means between all groups and the pairwise.t.test to compare between each two groups.

$H_0$ : no significant difference between the means against $H_1$ : not all means are equal

```
                                         Df  Sum Sq  Mean Sq  F value   Pr(>F)
factor(onlyrating$greenmarkrating)   3    5.34   1.779    4.361   0.00505
**
```

The ANOVA test shows that the means are indeed not all equal to one another. Thus we further need to use the pairwise.t.test to see which pairs have the same mean and which do not.

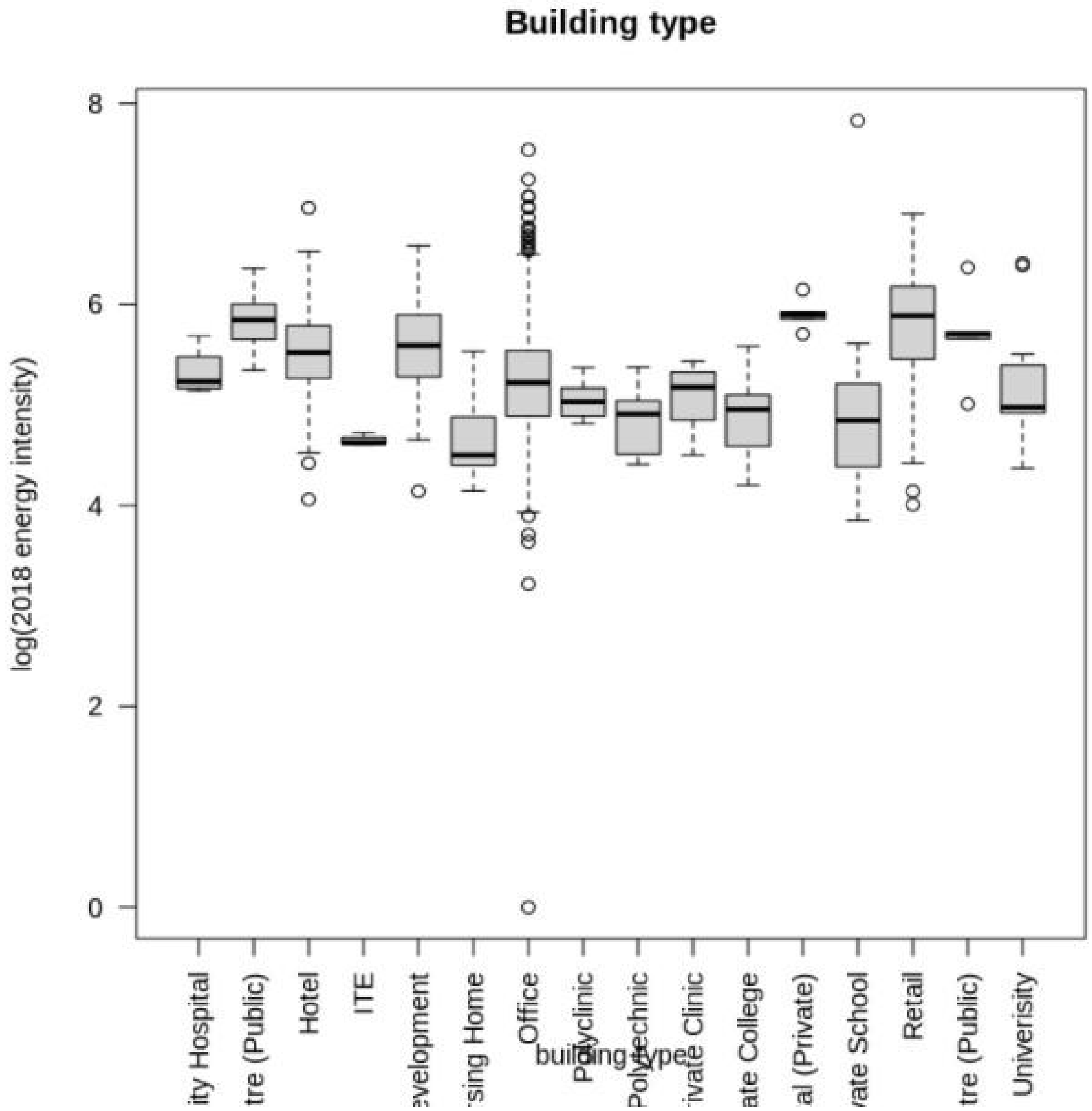
```
      Pairwise comparisons using t tests with pooled SD

data:  log(onlyrating$X2018energyusintensity) and
```

```
onlyrating$greenmarkrating

         Certified  Gold    GoldPlus
Gold     0.07250    -       -
GoldPlus 0.84693    0.09693 -
Platinum 0.25551    0.00037 0.15179

P value adjustment method: none
```

Through pairwise.t.test we can see that Certified and GoldPlus rated buildings have similar means while Gold and Platinum rated buildings have a considerably significant difference in means.Thus we conclude that energy use intensity depends on the green mark rating of the building.

### 4.3.5 Relation between energy use intensity (*X2018energyuseintensity*) and building type (*buildingtype*)

In this section we try to answer “Does building type affect the energy use intensity.
We use analysis of variance (ANOVA) to determine whether log(*X2018energyusintensity*) is different based on building type, since *buildingtype* is a categorical variable. The following plot illustrates the distributions of log(*X2018energyusintensity*) among the different building types (*buildingtype*).



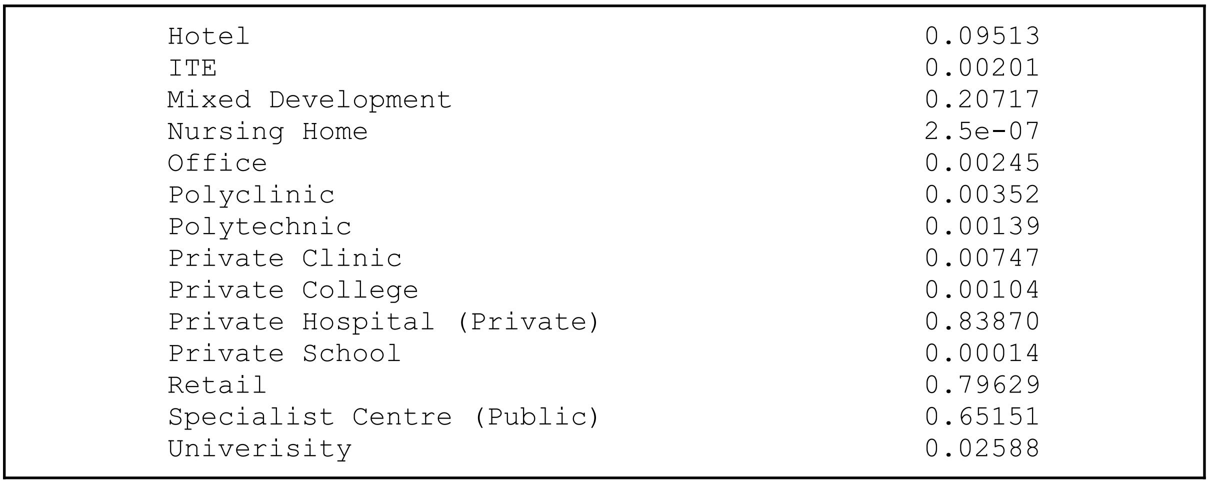
Looking at this boxplot, we see that the spread varies a lot based on the building type. Hence, the ANOVA test is appropriate for testing the equality of the means.

$H_0$ : no significant difference between the means against $H_1$ : not all means are equal

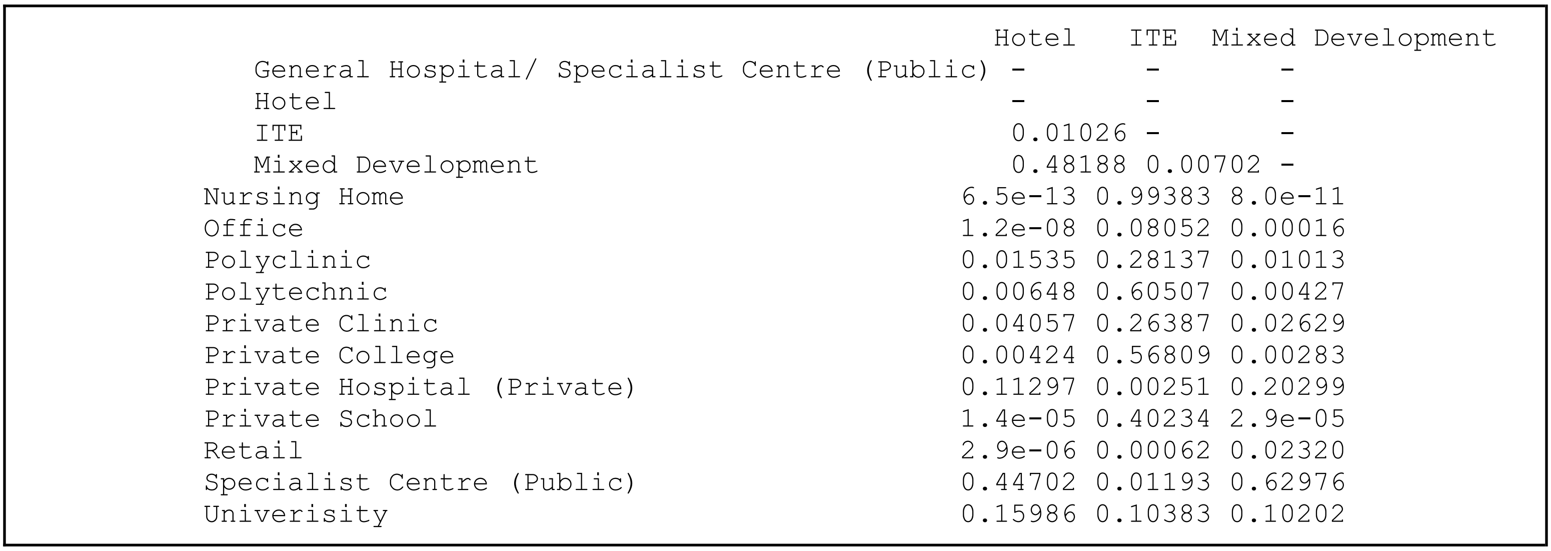
```
                              Df  Sum Sq Mean Sq F value Pr(>F)    
factor(energy$buildingtype)   15  67.54   4.503   14.06 <2e-16 ***
```

The F-test confirms that the means for the different building types are different. Therefore we use a pairwise.t.test to compute the p-values of the different building types in pairs. This will tell us which pairs don't not have an equal mean.

```
                                General Hospital/ Specialist Centre (Public)
General Hospital/ Specialist Centre (Public)    -
```



```
        Hotel                                        0.09513
        ITE                                          0.00201
        Mixed Development                            0.20717
        Nursing Home                                 2.5e-07
        Office                                       0.00245
        Polyclinic                                   0.00352
        Polytechnic                                  0.00139
        Private Clinic                               0.00747
        Private College                              0.00104
        Private Hospital (Private)                   0.83870
        Private School                               0.00014
        Retail                                       0.79629
        Specialist Centre (Public)                   0.65151
        Univerisity                                  0.02588
```



```
                                                           Hotel   ITE  Mixed Development
     General Hospital/ Specialist Centre (Public) -       -       -
     Hotel                                         -       -       -
     ITE                                           0.01026 -       -
     Mixed Development                             0.48188 0.00702 -
  Nursing Home                                  6.5e-13 0.99383 8.0e-11
  Office                                        1.2e-08 0.08052 0.00016
  Polyclinic                                    0.01535 0.28137 0.01013
  Polytechnic                                   0.00648 0.60507 0.00427
  Private Clinic                                0.04057 0.26387 0.02629
  Private College                               0.00424 0.56809 0.00283
  Private Hospital (Private)                    0.11297 0.00251 0.20299
  Private School                                1.4e-05 0.40234 2.9e-05
  Retail                                        2.9e-06 0.00062 0.02320
  Specialist Centre (Public)                    0.44702 0.01193 0.62976
  Univerisity                                   0.15986 0.10383 0.10202
```

Through pairwise t-test we see that the means of building types differ a lot based on which pair of building types are compared. However, some do have similar means such as Private Hospital and Hotel, there are also those pairs such as Retail and Hotel which have significantly different means. Through ANOVA and pairwise t-test we conclude that the energy use intensity depends on the building type.

# 5. Conclusion and Discussion

While we all strive to become more environmentally-friendly. We also strive to reduce our electricity consumption. Reducing electricity consumption does not just help the environment, it also helps reduce costs. To get a better understanding of what factors affect electricity use intensity we look at some factors that could affect energy use intensity with very little detail on what instruments (lightbulbs, kitchen appliances, chargers, etc) are consuming electricity.

We conclude that:

- Decrease in energy use from 2017 to 2018 and green mark status being awarded in 2018 are positively correlated
- Energy use intensity is not dependent on building size
- Gross floor area does not affect energy use intensity
- Energy use intensity depends on the green mark rating of the building.
- Building type affects energy use intensity

We note that while we were able to conclude on some factors driving electricity consumption. There are factors that have not been included such as appliances used, however, building type does shine some light on that missing information.

# 6. Appendix

https://colab.research.google.com/drive/1x9XbEzfF_e6Q6qsy8dJkk-jARqazFkjy?usp=sharing#scrollTo=v73gqY1oW7xx