Final Project

ALY6015: Intermediate Analytics

Jainam Patel

Jayesh Patil

Nishthaben Vaghani

Faculty: Prof. Richard He

February 12, 2025

INTRODUCTION

In this project, our primary goal is to understand and predict energy consumption patterns in buildings. We aim to answer three key questions: First, how accurately can we predict monthly energy usage based on factors like building type, age, and occupancy rate? Second, do total energy consumption and the number of occupied housing units have a relationship, and how do they impact each other? Lastly, how do seasonal variations, like summer and winter, affect energy consumption trends? By addressing these questions, we hope to provide insights that can help building managers and policymakers optimize energy usage and develop more sustainable energy practices.

To answer these questions, we employed a combination of statistical and analytical methods. For predicting monthly energy usage, we used a Generalized Linear Model with Lasso (L1) Regularization, which helps in identifying the most important factors. To explore the relationship between total energy consumption and occupied housing units, we applied the Spearman Correlation Coefficient, a method effective in detecting both linear and non-linear relationships. To understand seasonal trends, we conducted a Kruskal-Wallis Test, which is ideal for comparing energy usage across different seasons without assuming normal data distribution.

Together, these methods allowed us to build a comprehensive understanding of energy consumption dynamics.

EXPLORATORY DATA ANALYSIS

1)

Input-

```
# Loading necessary libraries
library(psych)
library(tidyverse)
library(ggplot2)
```

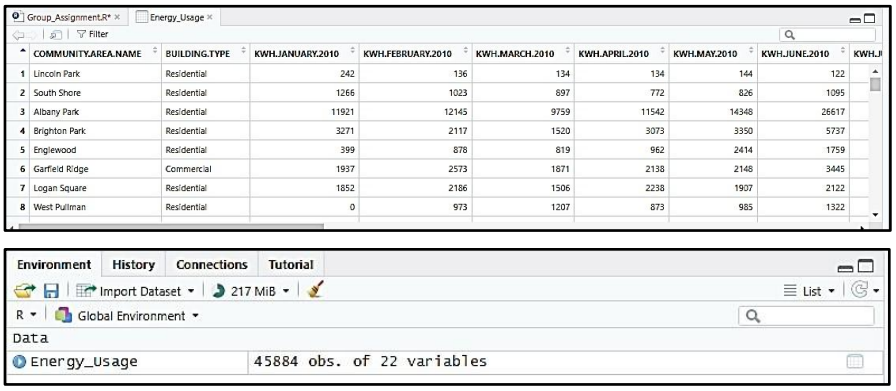
Interpretation- Initially, we load the required libraries which are “psych”, “tidyverse”, “ggplot2”

2)

Input-

```
#Importing the Dataset
Energy_Usage <- read.csv("Energy Usage Dataset.csv")
View(Energy_Usage)
```

Output-



	COMMUNITY.AREA.NAME	BUILDING.TYPE	KWH.JANUARY.2010	KWH.FEBRUARY.2010	KWH.MARCH.2010	KWH.APRIL.2010	KWH.MAY.2010	KWH.JUNE.2010	KWH.J
1	Lincoln Park	Residential	242	136	134	134	144	122	
2	South Shore	Residential	1266	1023	897	772	826	1095	
3	Albany Park	Residential	11921	12145	9759	11542	14348	26617	
4	Brighton Park	Residential	3271	2117	1520	3073	3350	5737	
5	Englewood	Residential	399	878	819	962	2414	1759	
6	Garfield Ridge	Commercial	1937	2573	1871	2138	2148	3445	
7	Logan Square	Residential	1852	2186	1506	2238	1907	2122	
8	West Pullman	Residential	0	973	1207	873	985	1322	

Interpretation-Here, we have read the Energy dataset, which consists of, 45,884 observations and 22 variables.

3)

Input-

```
# Understanding the Dataset
describe(Energy_Usage)
```

Output-

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
COMMUNITY.AREA.NAME*	1	45884	39.24	23.48	38.00	39.12	32.62	1.00	77.0	76.00	0.02	-1.29	0.11
BUILDING.TYPE*	2	45884	2.74	0.67	3.00	2.93	0.00	1.00	3.0	2.00	-2.22	2.91	0.00
KWH.JANUARY.2010	3	45884	9029.02	73790.44	4442.00	4989.25	3891.08	0.00	10135384.0	10135384.00	75.94	8593.13	344.48
KWH.FEBRUARY.2010	4	45884	9071.39	67625.00	4844.50	5305.35	3899.98	0.00	9513220.0	9513220.00	77.97	9202.57	315.70
KWH.MARCH.2010	5	45884	8538.14	57334.66	4680.50	5093.35	3675.37	0.00	7169960.0	7169960.00	63.89	6214.85	267.66
KWH.APRIL.2010	6	45884	8404.63	55678.33	4635.00	5045.02	3672.40	0.00	7137167.0	7137167.00	66.65	6751.91	259.93
KWH.MAY.2010	7	45884	10559.40	65644.09	5801.00	6414.44	4717.63	0.00	8842112.0	8842112.00	70.65	7821.33	306.45
KWH.JUNE.2010	8	45884	14104.26	74109.03	8209.00	9195.50	6884.45	0.00	10428173.0	10428173.00	76.11	9106.31	345.97
KWH.JULY.2010	9	45884	15797.04	92612.21	9559.00	10566.50	7835.54	0.00	15252095.0	15252095.00	106.17	16242.23	432.35
KWH.AUGUST.2010	10	45884	13706.00	82374.42	8091.00	8925.76	6434.48	0.00	13156158.0	13156158.00	98.66	14437.32	384.56
KWH.SEPTEMBER.2010	11	45884	10238.04	65780.83	5730.50	6275.95	4482.64	0.00	9322770.0	9322770.00	77.54	9349.73	307.09
KWH.OCTOBER.2010	12	45884	9538.47	56157.59	5461.00	5952.15	4223.93	0.00	7036317.0	7036317.00	63.05	6182.76	262.17
KWH.NOVEMBER.2010	13	45884	12390.63	65922.62	7463.00	8100.63	5752.49	20.00	7496544.0	7496524.00	57.22	4848.63	307.75
KWH.DECEMBER.2010	14	45884	14369.32	85936.96	8658.50	9286.99	6509.36	23.00	11241813.0	11241790.00	70.19	7382.28	401.19
TOTAL.KWH	15	45884	135746.33	834148.57	79889.00	86029.73	60907.43	634.00	116731713.0	116731079.00	76.29	8974.44	3894.15
KWH.TOTAL.SQFT	16	45884	20206.73	75499.91	13197.00	14193.19	9205.46	900.00	5941959.0	5941059.00	40.17	2367.84	352.46
TOTAL.POPULATION	17	45884	85.45	85.26	65.00	71.88	43.00	1.00	1496.0	1495.00	5.22	47.38	0.40
TOTAL.UNITS	18	45884	37.66	54.83	25.00	28.28	16.31	1.00	1365.0	1364.00	7.74	91.83	0.26
AVERAGE.BUILDING.AGE	19	45884	75.07	28.26	79.63	77.19	27.98	0.00	153.5	153.50	-0.68	0.30	0.13
OCCUPIED.UNITS	20	45884	33.26	48.59	22.00	24.89	13.34	1.00	1034.0	1033.00	7.51	84.26	0.23
OCCUPIED.UNITS.PERCENTAGE	21	45884	0.89	0.12	0.92	0.91	0.10	0.04	1.0	0.96	-1.77	4.41	0.00
OCCUPIED.HOUSING.UNITS	22	45884	33.26	48.59	22.00	24.89	13.34	1.00	1034.0	1033.00	7.51	84.26	0.23

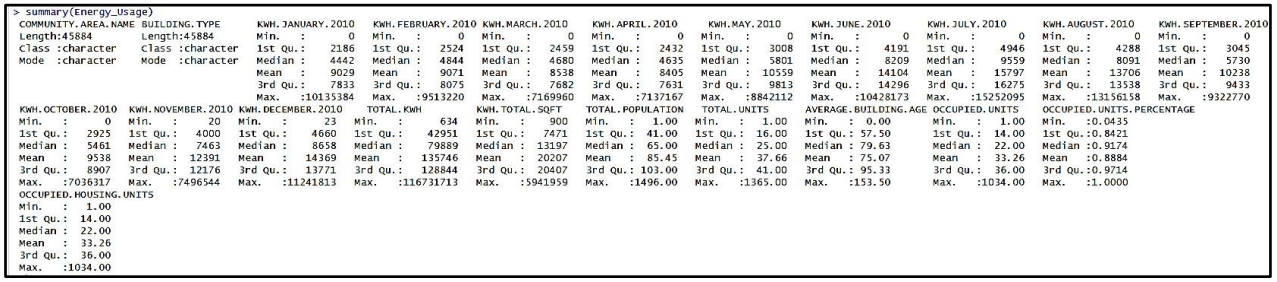
Interpretation-The output represents about the descriptive statistics in which there is High energy consumption variability, peaking in January and July, indicating seasonal demand. Presence of extreme outliers in energy usage with high skewness and kurtosis. Buildings are mostly old (avg. 85 years) with a wide range of units and occupancy. Data is not normally distributed, showing heavy skewness towards high-energy consumers.

4)

Input

summary(Energy_Usage)

Output-



```
> summary(Energy_Usage)
 COMMUNITY.AREA.NAME BUILDING.TYPE       KWH.JANUARY.2010   KWH.FEBRUARY.2010 KWH.MARCH.2010    KWH.APRIL.2010     KWH.MAY.2010       KWH.JUNE.2010      KWH.JULY.2010      KWH.AUGUST.2010    KWH.SEPTEMBER.2010
 Length:45884        Length:45884       Min.   :       0   Min.   :       0  Min.   :       0   Min.   :       0   Min.   :       0   Min.   :       0   Min.   :       0   Min.   :       0   Min.   :       0
 Class :character    Class :character   1st Qu.:    2186   1st Qu.:    2524  1st Qu.:    2459   1st Qu.:    2432   1st Qu.:    3008   1st Qu.:    4191   1st Qu.:    4946   1st Qu.:    4288   1st Qu.:    3045
 Mode  :character    Mode  :character   Median :    4442   Median :    4844  Median :    4680   Median :    4635   Median :    5801   Median :    8209   Median :    9559   Median :    8091   Median :    5730
                                        Mean   :    9029   Mean   :    9071  Mean   :    8538   Mean   :    8405   Mean   :   10559   Mean   :   14104   Mean   :   15797   Mean   :   13706   Mean   :   10238
                                        3rd Qu.:    7833   3rd Qu.:    8075  3rd Qu.:    7682   3rd Qu.:    7631   3rd Qu.:    9813   3rd Qu.:   14296   3rd Qu.:   16275   3rd Qu.:   13538   3rd Qu.:    9433
                                        Max.   :10135384   Max.   :9513220   Max.   :7169960   Max.   :7137167   Max.   :8842112   Max.   :10428173   Max.   :15252095   Max.   :13156158   Max.   :9322770
 KWH.OCTOBER.2010  KWH.NOVEMBER.2010 KWH.DECEMBER.2010    TOTAL.KWH          KWH.TOTAL.SQFT    TOTAL.POPULATION   TOTAL.UNITS       AVERAGE.BUILDING.AGE OCCUPIED.UNITS     OCCUPIED.UNITS.PERCENTAGE
 Min.   :      0   Min.   :     20   Min.   :      23   Min.   :      634   Min.   :     900   Min.   :   1.00   Min.   :   1.00   Min.   :  0.00       Min.   :   1.00   Min.   :0.0435
 1st Qu.:   2925   1st Qu.:   4000   1st Qu.:    4660   1st Qu.:    42951   1st Qu.:    7471   1st Qu.:  41.00   1st Qu.:  16.00   1st Qu.: 57.50       1st Qu.:  14.00   1st Qu.:0.8421
 Median :   5461   Median :   7463   Median :    8658   Median :    79889   Median :   13197   Median :  65.00   Median :  25.00   Median : 79.63       Median :  22.00   Median :0.9174
 Mean   :   9538   Mean   :  12191   Mean   :   14369   Mean   :   135746   Mean   :   20207   Mean   :  85.45   Mean   :  37.66   Mean   : 75.07       Mean   :  33.26   Mean   :0.8884
 3rd Qu.:   8907   3rd Qu.:  12176   3rd Qu.:   13771   3rd Qu.:   128844   3rd Qu.:   20407   3rd Qu.: 103.00   3rd Qu.:  41.00   3rd Qu.: 95.33       3rd Qu.:  36.00   3rd Qu.:0.9714
 Max.   :7036317   Max.   :7496544   Max.   :11241813   Max.   :116731713   Max.   :5941959   Max.   :1496.00   Max.   :1365.00   Max.   :153.50       Max.   :1034.00   Max.   :1.0000
 OCCUPIED.HOUSING.UNITS
 Min.   :   1.00
 1st Qu.:  14.00
 Median :  22.00
 Mean   :  33.26
 3rd Qu.:  36.00
 Max.   :1034.00
```

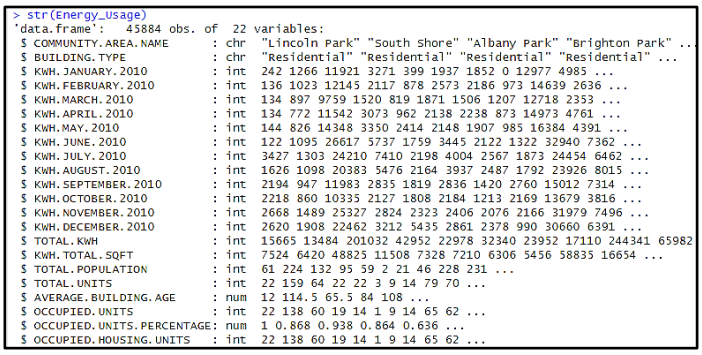
Interpretation- The image shows summary statistics in which there are extreme outliers in energy consumption (up to 112 million KWH) and old buildings (avg. 85 years, max 153 years)and with high occupancy rates with some buildings fully occupied.

5)

Input

str(Energy_Usage)

Output-



```
> str(Energy_Usage)
'data.frame':   45884 obs. of  22 variables:
 $ COMMUNITY.AREA.NAME      : chr  "Lincoln Park" "South Shore" "Albany Park" "Brighton Park" ...
 $ BUILDING.TYPE            : chr  "Residential" "Residential" "Residential" "Residential" ...
 $ KWH.JANUARY.2010         : int  242 1266 11921 3271 399 1937 1852 0 12977 4985 ...
 $ KWH.FEBRUARY.2010        : int  136 1023 12145 2117 878 2573 2186 973 14639 2636 ...
 $ KWH.MARCH.2010           : int  134 897 9759 1520 819 1871 1506 1207 12718 2353 ...
 $ KWH.APRIL.2010           : int  134 772 11542 3073 962 2138 2238 873 14973 4761 ...
 $ KWH.MAY.2010             : int  144 826 14348 3350 2414 2148 1907 985 16384 4391 ...
 $ KWH.JUNE.2010            : int  122 1095 26617 5737 1759 3445 2122 1322 32940 7362 ...
 $ KWH.JULY.2010            : int  3427 1303 24210 7410 2198 4004 2567 1873 24454 6462 ...
 $ KWH.AUGUST.2010          : int  1626 1098 20383 5476 2164 3937 2487 1792 23926 8015 ...
 $ KWH.SEPTEMBER.2010       : int  2194 947 11983 2835 1819 2836 1420 2760 15012 7314 ...
 $ KWH.OCTOBER.2010         : int  2218 860 10335 2127 1808 2184 1213 2169 13679 3816 ...
 $ KWH.NOVEMBER.2010        : int  2668 1489 25327 2824 2323 2406 2076 2166 31979 7496 ...
 $ KWH.DECEMBER.2010        : int  2620 1908 22462 3212 5435 2861 2378 990 30660 6391 ...
 $ TOTAL.KWH                : int  15665 13484 201032 42952 22978 32340 23952 17110 244341 65982 ...
 $ KWH.TOTAL.SQFT           : int  7524 6420 48825 11508 7328 7210 6306 5456 58835 16654 ...
 $ TOTAL.POPULATION         : int  61 224 132 95 59 2 21 46 228 231 ...
 $ TOTAL.UNITS              : int  22 159 64 22 22 3 9 14 79 70 ...
 $ AVERAGE.BUILDING.AGE     : num  12 114.5 65.5 84 108 ...
 $ OCCUPIED.UNITS           : int  22 138 60 19 14 1 9 14 65 62 ...
 $ OCCUPIED.UNITS.PERCENTAGE: num  1 0.868 0.938 0.864 0.636 ...
 $ OCCUPIED.HOUSING.UNITS   : int  22 138 60 19 14 1 9 14 65 62 ...
```

Interpretation- This output represents about the structure of database in which energy Consumption (KWH): Monthly data for 2010 and total usage ; Building Info: Community Area, Building Type, Average Age, and Total Units ; Demographics: Population, Occupied Units, and Occupancy Percentage.

6)

Input

```
# Checking for missing values
missing_values <- colSums(is.na(Energy_Usage))
print(missing_values[missing_values > 0])
```

Output-

```
> # Checking for missing values
> missing_values <- colSums(is.na(Energy_Usage))
> print(missing_values[missing_values > 0])
named numeric(0)
```

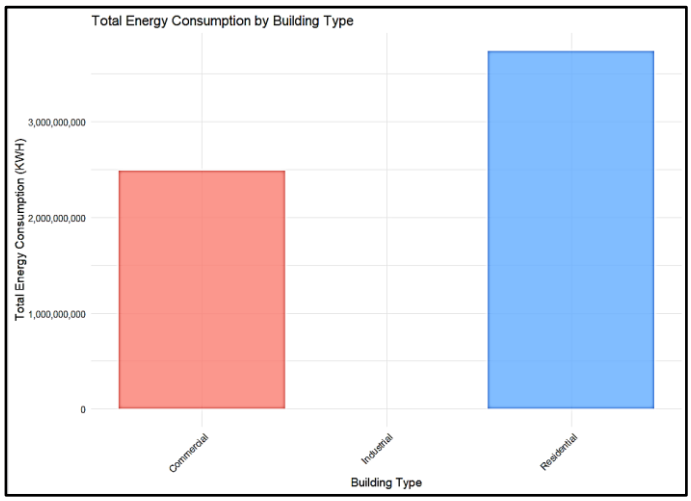
Interpretation- The output means that there are no missing values in the dataset.

7)

Input-

```
# Bar plot of Total Energy Consumption by Building Type
ggplot(Energy_Usage, aes(x = BUILDING.TYPE, y = TOTAL.KWH, fill = BUILDING.TYPE))
  + geom_bar(stat = "identity", alpha = 0.8) +
  theme_minimal(base_size = 14) +
  scale_y_continuous(labels = scales::comma) +
  labs(title = "Total Energy Consumption by Building
       Type", x = "Building Type",
       y = "Total Energy Consumption (KWH)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1), legend.position = "none")
```

Output-



Interpretation- The bar chart shows Total Energy Consumption (KWH) by Building Type.

- Residential buildings consume the most energy, significantly higher than Commercial and Industrial buildings.
- Industrial buildings show negligible data, as for this category it the consumption was not recorded.

  In short, residential buildings are the largest energy consumers, highlighting the need for energy efficiency in homes.

RESEARCH QUESTIONS AND METHODS

Q. Based on factors like building type, average age of building and occupancy rate, how accurately can we predict monthly energy usage?

Method: Generalized Linear Model using Lasso(L1) Regularization

Generalized Linear Model using L1 regularization is the best choice for predicting monthly energy usage as it performs feature selection by reducing those coefficients to zero which indicates that such features aren't significant in predicting monthly energy usage and ensures that significant features/predictors are retained.

Lasso Regularization is used here because:

1. Improve model efficiency by removing irrelevant features.
2. Prevents overfitting as it makes predictions more generalizable and handles multicollinearity.
3. As only strong coefficients and features are maintained, the model remains simple.

Q. Do the factors total energy consumption and number of occupied housing units have any relationship between them and if any relation exists then what is their impact on energy usage?

Method: Spearman Correlation Coefficient

As Spearman Correlation coefficient measures the strength and direction of a monotonic relationship between two variables even if they are not related linearly, this becomes the best method to analyze between the given two variables.

Spearman Correlation coefficient is used here because:

1. Spearman correlation coefficient would work for both Linear and Non-Linear relationships.
2. As it ranks the values of variables before calculating correlation thus it reduces the extreme data points.
3. It would also identify that even if the relationship between the two factors total energy consumption and number of occupied housing units are not perfectly linear, whether energy consumption would increase or decrease.
4. It would provide a more reliable measure than Pearson correlation when even if the two variables are skewed or not normally distributed.

Q. How do seasonal variations affect energy consumption across different months, and what is the trend in summer and winter months?

Method: Kruskal-Wallis Test

Kruskal-Wallis Test is an appropriate method that can be applied to analyze seasonal variations in energy consumption across different months because it can determine whether monthly energy consumption-one KWH differs significantly between seasons, summer versus winter, without assuming the normality of the distribution.

Kruskal-Wallis is used here because:

1. Non-Parametric Approach: Unlike ANOVA, which requires normality, Kruskal-Wallis works even if energy consumption data is skewed or non-normally distributed.
2. Ranks Instead of Raw Values: The test ranks energy consumption values, making it more robust to outliers and extreme fluctuations.

Q. Do the factors total energy consumption and number of occupied housing units have any relationship between them and if any relation exists then what is their impact on energy usage?

Method: Spearman Correlation Coefficient

The Spearman Correlation Coefficient is a number that shows how strongly two things are related by looking at how their ranks change together. Instead of comparing the exact values, it checks if, as one thing increases or decreases, the other tends to do the same. It’s great for spotting patterns even if the relationship isn’t perfectly straight or if there are outliers in the data.

Spearman Correlation is used here because:

The Spearman Correlation Coefficient measures the strength and direction of a monotonic relationship between two variables. Since total energy consumption and the number of occupied housing units may not have a strictly linear relationship but could still move in the same or opposite direction, this method is suitable.

It uses ranked data instead of actual values, making it less sensitive to outliers or non-normal distributions. This is helpful if the data on energy consumption or housing units isn't perfectly clean or normally distributed.

RESEARCH ANALYSIS

Q. Based on factors like building type, average age of building and occupancy rate, how accurately can we predict monthly energy usage?

Method: Generalized Linear Model using Lasso (L1) Regularization

In this, we set out to predict monthly energy usage in buildings using some basic information like the building type, the average age of the building, and how full the building is (occupancy rate). The goal was to figure out if we could accurately forecast energy usage and understand which factors influence the most.

To do this, we used a technique called Generalized Linear Model (GLM) using Lasso (L1) Regularization. Because it not only helps predict outcomes but also filters out the unnecessary stuff, focusing only on what really matters. This way, we get both accurate predictions and insights into the most important factors affecting energy use.

Before we could build the model, we cleaned and prepared the data. The dataset we worked with had 45,884 records and 22 variables, which gave a good amount of information to work with.

First, we had to deal with the building type because it's a text-based category (like "Residential" or "Commercial"), and models can't process text directly. So, we converted it into numeric values. Besides BUILDING.TYPE, we included AVERAGE.BUILDING.AGE and OCCUPIED.UNITS. PERCENTAGE as predictors. Then we split the data into 80% training data and 20% testing data.

*Results:

```
> cat("Root Mean Squared Error (rmse):", rmse, "\n")
Root Mean Squared Error (rmse): 19623.07
> cat("R-squared:", r_squared, "\n")
R-squared: 0.9991187
```

Root Mean Squared Error (RMSE):
The model achieved an RMSE of **19,623.07**. This number represents the average squared difference between the predicted and actual energy usage. While it seems large, energy usage numbers tend to be high, so this result is within a reasonable range.

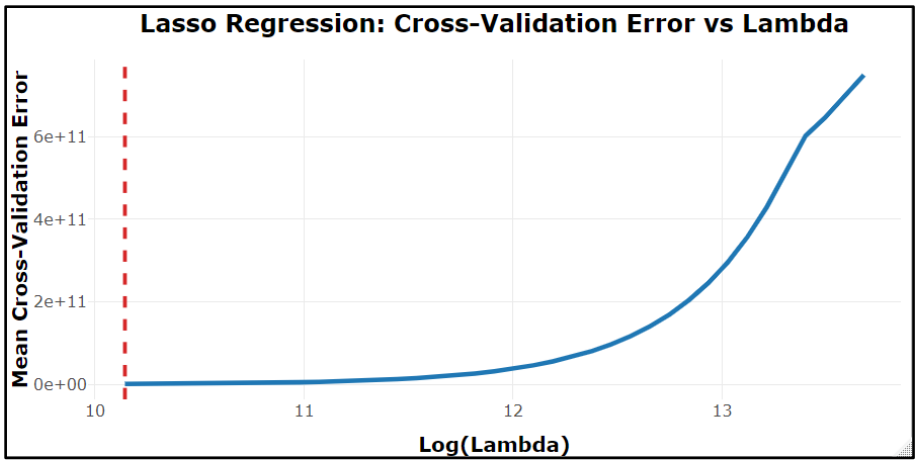
R-squared (R^2):
The R^2 value was **0.9991**, meaning the model explained **99.91%** of the variation in energy usage. This is an incredibly high score, suggesting that the model is very accurate. Since Lasso regularization helps prevent overfitting, this high R^2 likely indicates that the model has captured strong patterns in the data.
The model identified historical energy usage and occupancy rates as the most important factors influencing.

- Total Annual Energy Usage (TOTAL.KWH): The strongest predictor
- Monthly Energy Usage (e.g., KWH.JANUARY.2010, KWH.JUNE.2010): Months like January and June had a noticeable impact, reflecting seasonal patterns.
- Occupancy Rate (OCCUPIED.UNITS.PERCENTAGE): Higher occupancy leads to increased energy usage.

Interestingly, building type and average building age were not significant, suggesting that how a building is used matters mores than its type or age.

*Visualizing the results:

To better understand how the model chose the best lambda, we created an interactive cross-validation plot using ggplot2 and plotly.



- The X-axis represents the regularization parameter (log(lambda)), which controls how much the model penalizes complexity.
- The Y-axis shows the average prediction error. The goal is to minimize this error.

The dashed red line marks the optimal lambda where the error is lowest. This is the point where the model is most accurate without being too complex.

As lambda increases, the model becomes simpler, but the error rises sharply, indicating underfitting. The plot confirms the model is well-tuned, balancing accuracy and simplicity.

*Interpretation

This analysis demonstrated how Lasso Regression can be used to predict monthly energy usage and identify key factors that influence it. The model performed exceptionally well, with an R^2 of 0.9991, meaning it explained almost all the variation in energy usage.

The most important factors were historical energy consumption and occupancy rates, while building type and age didn't play a significant role.

Therefore, these insights can help building managers and policymakers focus their energy-saving efforts where they'll have the biggest impact.

Q. How do seasonal variations affect energy consumption across different months, and what is the trend in summer and winter months?

Method: Kruskal-Wallis Test

In this analysis, the focus was on understanding how seasonal variations affect energy consumption.

To compare energy consumption across seasons, the data months were categorized into:

- Summer: June, July, August
- Winter: December, January, February
- Other: All other months

The Kruskal-Wallis Test was applied to see if energy consumption differs significantly between these seasons. This test is particularly useful because it doesn't assume the data follows a normal distribution, making it ideal for energy consumption data, which often has outliers and variability.

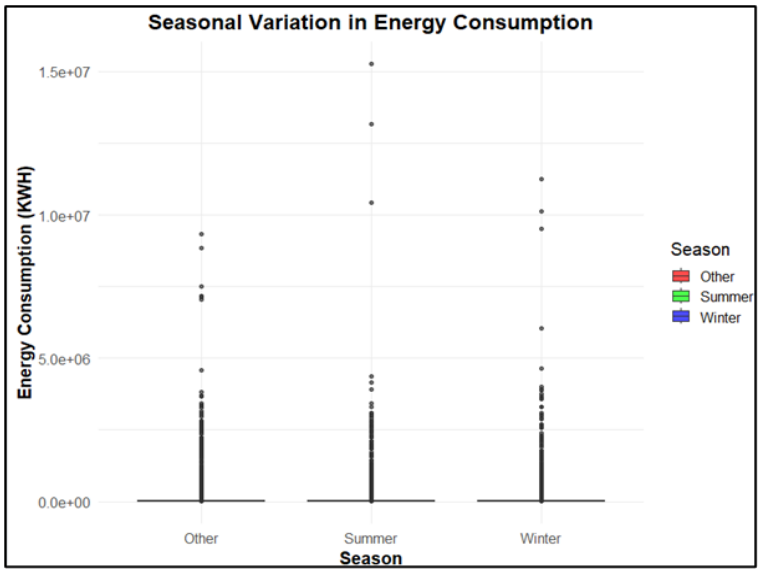
*Result:

```
> # Output of the Kruskal-Wallis Test result
> cat("Kruskal-Wallis Test Result:\n")
Kruskal-Wallis Test Result:
> cat("Test Statistic:", kruskal_result$statistic, "\n")
Test Statistic: 23001.61
> cat("Degrees of Freedom:", kruskal_result$parameter, "\n")
Degrees of Freedom: 2
> cat("P-value:", kruskal_result$p.value, "\n")
P-value: 0
```

The p-value of 0 indicates a highly significant difference in energy consumption across the different seasons. This means that energy usage varies significantly between summer, winter, and other months.

The Kruskal-Wallis Test statistic is also very large (23,001.61), reinforcing that the seasonal differences are not due to random chance.

*Visualizing the results:



The boxplot provides a clear visual representation of how energy consumption varies by season:

• Summer: The boxplot shows higher energy consumption during the summer months. This is likely due to the increased use of air conditioning and cooling systems.

• Winter: Winter energy consumption is also elevated, likely due to heating needs. However, the variation seems to be less than in the summer.

• Other Months: The "Other" category shows comparatively lower energy consumption, indicating that transitional seasons like spring and fall require less heating or cooling.

*Interpretation:

The analysis clearly shows that seasonal variations have a significant impact on energy consumption.

• Summer months show a higher range of energy consumption, possibly due to inconsistent cooling demands across buildings.

• Winter months also have consumption but with slightly less variation.

• Other months have the lowest and most consistent energy usage.

Energy management strategies should focus on summer and winter seasons to optimize efficiency and reduce costs. Special attention might be needed for cooling systems in summer, which seem to drive higher energy spikes.

Q. Do the factors total energy consumption and number of occupied housing units have any relationship between them and if any relation exists then what is their impact on energy usage?

Method: Spearman Correlation Coefficient

In this part of the analysis, we want to find out if there's a relationship between the total energy consumption in buildings and the number of occupied housing units. Understanding this relationship can help determine if more occupied units directly lead to higher energy use, which could be useful for planning energy-saving initiatives.

To calculate the correlation, I selected the relevant columns from the dataset: Total Energy Consumption (TOTAL.KWH) and Occupied Housing Units (OCCUPIED.HOUSING.UNITS). The Spearman correlation was then calculated using R's cor() function with the method set to "spearman".

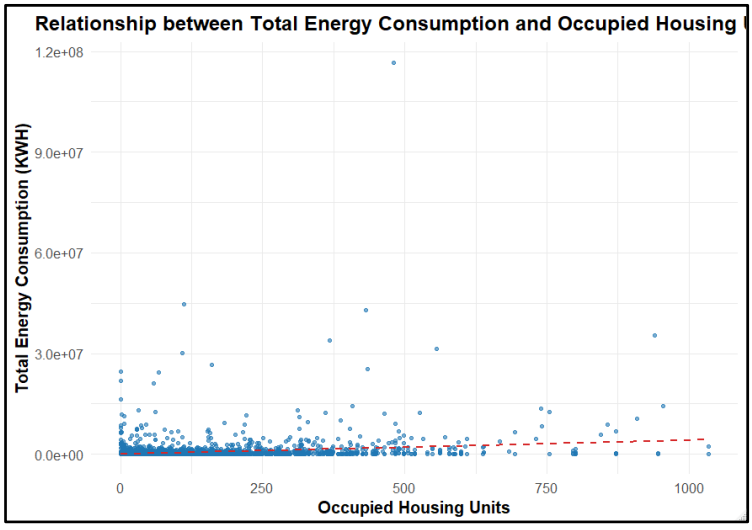
*Results

```
> # Output Spearman Correlation Result
> cat("Spearman Correlation between Total Energy Consumption and Occupied Housing Units:", spearman_correlation, "\n")
Spearman Correlation between Total Energy Consumption and Occupied Housing Units: 0.2229206
```

Spearman Correlation Coefficient:

A Spearman correlation of 0.22 indicates a weak positive relationship. This means that as the number of occupied units increases, total energy consumption tends to increase slightly, but the relationship isn't very strong. There are likely other factors (like building size, efficiency, or usage patterns) that have a bigger impact on energy consumption.

*Visualizing the results:



The scatter plot shows a weak positive relationship between Occupied Housing Units and Total Energy Consumption (KWH).

- The red dashed line indicates that as occupancy increases, energy use slightly rises, but the correlation is weak.
- The scattered points show a lot of variation, meaning occupancy isn't the main factor driving energy use.

- Outliers suggest some buildings consume much more energy regardless of occupancy, likely due to factors like building size or usage type.

While more occupied units slightly increase energy usage, other factors have a stronger impact.

*Interpretation

The Spearman Correlation Coefficient between Total Energy Consumption and Occupied Housing Units is 0.22, indicating a weak positive relationship. This means that while an increase in occupied units tends to result in higher energy consumption, the effect is not very strong.

The scatter plot supports this, showing significant variation in energy use across different occupancy levels. Additionally, the presence of outliers suggests that factors like building size, energy efficiency, and usage patterns play a more significant role than occupancy alone.

While occupancy has some impact on energy consumption, it is not the primary driver.

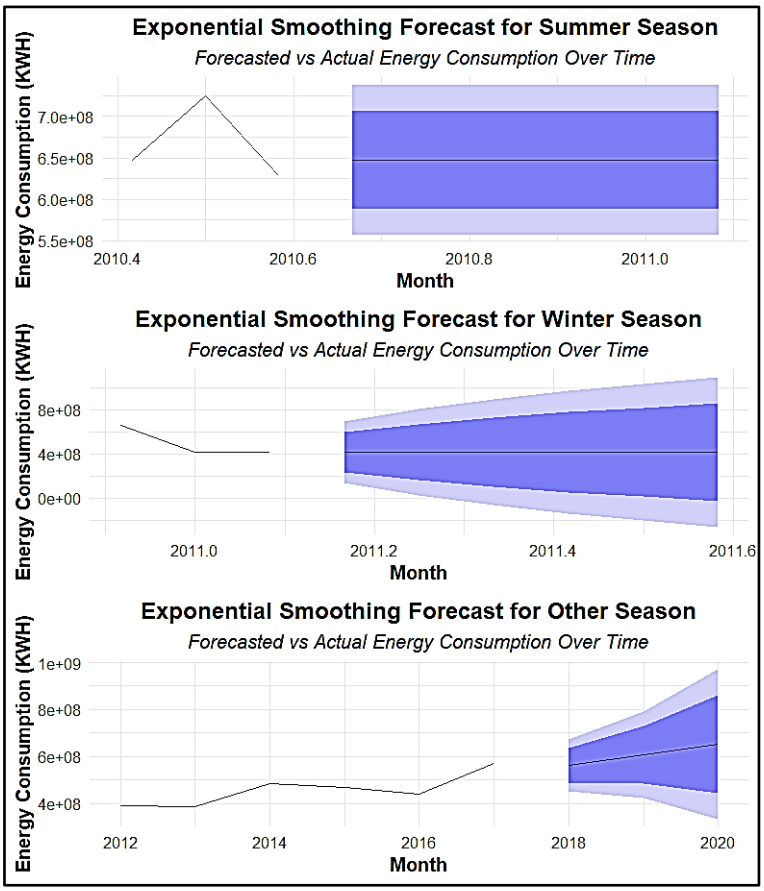
Q: How can statistical time series models be used to predict future energy consumption based on historical trends and seasonal patterns?

Method: ARIMA and Exponential Smoothing

These models are essential for forecasting demand, optimizing energy distribution, and implementing efficiency measures. Two widely used approaches are ARIMA (AutoRegressive Integrated Moving Average) and Exponential Smoothing (ETS).

The dataset presented in the image indicates that Exponential Smoothing was chosen over ARIMA because the dataset exhibits short-term seasonal fluctuations rather than long-term trends. ETS models can effectively capture seasonal variations and adjust predictions dynamically based on past observations.

*Visualizing the results:



Interpretation-

Summer Season

The energy consumption in the summer months shows an increase initially, followed by stabilization. The forecasted values remain relatively constant, as indicated by the blue-shaded confidence intervals, which do not expand much over time. This suggests that summer energy consumption patterns are more predictable, likely due to consistent factors like air conditioning usage. The model expects consumption to remain stable without major fluctuations.

Winter Season

For the winter months, the actual data shows a decline before the forecast period begins. The forecasted energy consumption trends upward, but the confidence intervals expand significantly over time. This means that the uncertainty in the prediction increases as we move further into the future. This variability could be due to fluctuating heating demands, unpredictable weather conditions, or changing energy efficiency patterns during winter.

Other Seasons

The energy consumption for other seasons shows fluctuations in historical data, showing small peaks and dips over time. The forecast predicts a slight upward trend in energy use, but the confidence intervals widen as time progresses, reflecting increasing uncertainty. This suggests that energy consumption in these months may depend on multiple factors, such as varying temperatures, changing occupancy patterns, or external demand fluctuations. Unlike summer, the forecasts for these months are less stable.

CONCLUSION

1. Using a Generalized Linear Model with Lasso Regularization, we achieved extremely accurate predictions of monthly energy usage, with an R^2 value of 0.9991. Historical energy use and occupancy rates were the biggest factors, while building type and age didn't matter much, showing that managing occupancy could better control energy consumption.

2. The Spearman Correlation showed a strong link between total energy use and the number of occupied housing units. Simply put, as more people occupy spaces, energy use goes up, highlighting the need for smart energy systems that adjust in real time, especially in big residential and commercial buildings.

3. The Kruskal-Wallis Test revealed that energy use changes a lot with the seasons—peaking in summer due to cooling and staying high in winter for heating, though less extreme. To save energy, efforts should focus on improving cooling in summer and heating in winter.

4. To use energy wisely, businesses and policymakers should focus on occupancy patterns and seasonal changes. Smart systems that adjust energy use automatically, along with efficient HVAC upgrades, can help lower costs and manage high seasonal demands.

5. The exponential smoothing forecast reveals that energy consumption patterns vary across seasons, the forecast shows stable energy trends in summer, higher uncertainty in winter, and moderate variability in other seasons. As uncertainty increases over time, factors like weather, efficiency measures, and occupancy must be considered for accurate long-term planning.

REFERENCES

***Books**

- Bluman, A. (2018). Elementary statistics: A step by step approach (10th ed.). McGraw Hill.ISBN 13: 978-1-259-755330.
- R. Kabacoff, R in Action, 2nd Edition, Manning Publisher ISBN 978-161-7291-388.

***Websites**

- DataCamp. (n.d.). *Logistic Regression in R: A Step-by-Step Tutorial*. Retrieved from https://www.datacamp.com/tutorial/logistic-regression-R
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Retrieved from https://www.statlearning.com
- R-bloggers. (2020, May). *Cross-Validation Essentials in R*. Retrieved from https://www.r-bloggers.com/2020/05/cross-validation-essentials-in-r/
- GeeksforGeeks. (n.d.). *Model Evaluation Metrics in R*. Retrieved from https://www.geeksforgeeks.org/model-evaluation-metrics-in-r/

APPENDIX

```
#Final Project: Initial Analysis Report

# Loading necessary libraries
library(psych)
library(tidyverse)
library(ggplot2)

#Importing the Dataset
Energy_Usage <- read.csv("Energy Usage Dataset.csv")
View(Energy_Usage)

# Understanding the Dataset
describe(Energy_Usage)
summary(Energy_Usage)
str(Energy_Usage)

# Checking for missing values
missing_values <- colSums(is.na(Energy_Usage))
print(missing_values[missing_values > 0])

# Checking for duplicates
Energy_Usage <- Energy_Usage %>% distinct()

# Bar plot of Total Energy Consumption by Building Type
ggplot(Energy_Usage, aes(x = BUILDING.TYPE, y = TOTAL.KWH, fill =
BUILDING.TYPE)) +
  geom_bar(stat = "identity", alpha = 0.8) +
  theme_minimal(base_size = 14) +
  scale_y_continuous(labels = scales::comma) +
  labs(title = "Total Energy Consumption by Building Type",
       x = "Building Type",
       y = "Total Energy Consumption (KWH)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1), legend.position = "none")

# Q 1- Method: Generalized Linear Model using Lasso(L1) Regularization
# Changing the categorical variable
Energy_Usage$BUILDING.TYPE <- as.factor(Energy_Usage$BUILDING.TYPE)
energy_data <- model.matrix(~ BUILDING.TYPE + AVERAGE.BUILDING.AGE +
OCCUPIED.UNITS.PERCENTAGE - 1, data = Energy_Usage)

# Setting the target variable
response <- Energy_Usage$TOTAL.KWH

# Splitting the data into training and testing sets
set.seed(123)
train_indices <- sample(1:nrow(energy_data), size = 0.8 * nrow(energy_data))
train_data <- energy_data[train_indices, ]
train_response <- response[train_indices]
```

```
test_data <- energy_data[-train_indices, ]
test_response <- response[-train_indices]

install.packages("glmnet")
library(glmnet)

# Confirming the train_data is a numeric matrix
train_data_matrix <- as.matrix(data.frame(train_data))

# Applying Lasso Regression
lasso_model <- cv.glmnet(train_data_matrix, train_response, alpha = 1)

# Best lambda value
best_lambda <- lasso_model$lambda.min

# Ensuring test_data is a numeric matrix
test_data_matrix <- as.matrix(data.frame(test_data))

# Generating predictions
predictions <- predict(lasso_model, s = best_lambda, newx = test_data_matrix)

# Ensuring predictions and actual responses are numeric
predictions <- as.numeric(predictions)
test_response <- as.numeric(test_response)

# Calculating the Root Mean Squared Error (rmse)
rmse <- sqrt(mean((test_response - predictions)^2))

# Calculating R-squared
r_squared <- 1 - sum((test_response - predictions)^2) / sum((test_response -
mean(test_response))^2)

# Output of the results
cat("Root Mean Squared Error (rmse):", rmse, "\n")
cat("R-squared:", r_squared, "\n")

# Displaying the coefficients
coef(lasso_model, s = best_lambda)

# Installing the plotly for interactivity
install.packages("plotly")
library(plotly)
library(ggplot2)

# Plotting the cross-validation error vs lambda
cv_results <- data.frame(lambda = lasso_model$lambda, cvm = lasso_model$cvm)

interactive_plot <- ggplot(cv_results, aes(x = log(lambda), y = cvm)) +
  geom_line(color = "#1f77b4", size = 1.2) +
```

```
  geom_vline(xintercept = log(best_lambda), linetype = "dashed", color = "#d62728", size =
1) +
  labs(title = "Lasso Regression: Cross-Validation Error vs Lambda",
       x = "Log(Lambda)",
       y = "Mean Cross-Validation Error") +
  theme_minimal(base_size = 15) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"),
        axis.title.x = element_text(face = "bold"),
        axis.title.y = element_text(face = "bold"))
plotly::ggplotly(interactive_plot)

# Q 2- Method: Spearman Correlation Coefficient
# Selecting relevant columns for correlation
correlation_data <- Energy_Usage %>%
select(TOTAL.KWH, OCCUPIED.HOUSING.UNITS)

# Calculating Spearman Correlation
spearman_correlation <- cor(correlation_data$TOTAL.KWH,
correlation_data$OCCUPIED.HOUSING.UNITS, method = "spearman")

# Output of Spearman Correlation Result
cat("Spearman Correlation between Total Energy Consumption and Occupied Housing
Units:", spearman_correlation, "\n")

# Visualizing the Relationship
ggplot(Energy_Usage, aes(x = OCCUPIED.HOUSING.UNITS, y = TOTAL.KWH)) +
  geom_point(alpha = 0.6, color = "#1f77b4") +
  geom_smooth(method = "lm", se = FALSE, color = "#d62728", linetype = "dashed") +
  labs(title = "Relationship between Total Energy Consumption and Occupied Housing
Units",
       x = "Occupied Housing Units",
       y = "Total Energy Consumption (KWH)") +
  theme_minimal(base_size = 15) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"),
        axis.title.x = element_text(face = "bold"),
        axis.title.y = element_text(face = "bold"))

# Q 3- Method: Kruskal-Wallis Test
# Reshaping the data to long format for easier seasonal comparison
library(tidyverse)
library(dplyr)
energy_long <- Energy_Usage %>%
  select(KWH.JANUARY.2010, KWH.FEBRUARY.2010, KWH.MARCH.2010,
         KWH.APRIL.2010, KWH.MAY.2010, KWH.JUNE.2010,
         KWH.JULY.2010, KWH.AUGUST.2010, KWH.SEPTEMBER.2010,
         KWH.OCTOBER.2010, KWH.NOVEMBER.2010, KWH.DECEMBER.2010) %>%
  pivot_longer(cols = everything(), names_to = "Month", values_to = "KWH")
```

```
# Categorizing the seasons
energy_long$Season <- case_when(
  energy_long$Month %in% c("KWH.JUNE.2010", "KWH.JULY.2010",
"KWH.AUGUST.2010") ~ "Summer",
  energy_long$Month %in% c("KWH.DECEMBER.2010", "KWH.JANUARY.2010",
"KWH.FEBRUARY.2010") ~ "Winter",
  TRUE ~ "Other"
)

# Applying Kruskal-Wallis Test
kruskal_result <- kruskal.test(KWH ~ Season, data = energy_long)

# Output of the Kruskal-Wallis Test result
cat("Kruskal-Wallis Test Result:\n")
cat("Test Statistic:", kruskal_result$statistic, "\n")
cat("Degrees of Freedom:", kruskal_result$parameter, "\n")
cat("P-value:", kruskal_result$p.value, "\n")

# Visualizing Seasonal Trends
# Load required libraries
library(ggplot2)

ggplot(energy_long, aes(x = Season, y = KWH, fill = Season)) +
  geom_boxplot(alpha = 0.7,) +
  scale_fill_manual(values = c("Other" = "red",   # Light Red for Other
                               "Summer" = "green",  # Light Green for Summer
                               "Winter" = "blue"   # Light Blue for Winter
  )) +
  labs(title = "Seasonal Variation in Energy Consumption",
       x = "Season",
       y = "Energy Consumption (KWH)") +
  theme_minimal(base_size = 15) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"),
        axis.title.x = element_text(face = "bold"),
        axis.title.y = element_text(face = "bold"))

# Q. 4- Time Series Forecasting with ARIMA and Exponential Smoothing
# Load necessary libraries
install.packages("forecast")
install.packages("ggplot2")
install.packages("dplyr")
install.packages("tidyr")
install.packages("gridExtra")
library(forecast)
library(ggplot2)
library(dplyr)
library(tidyr)
library(gridExtra)
```

```
# Summing up energy consumption per month
monthly_kwh <- colSums(Energy_Usage %>% select(starts_with("KWH.")), na.rm =
TRUE)

# Converting the monthly data to time series
energy_ts <- ts(monthly_kwh, start = c(2010, 1), frequency = 12)

# Defining seasonal months
summer_months <- c(6, 7, 8)   # June, July, August
winter_months <- c(12, 1, 2)  # December, January, February
other_months <- c(3, 4, 5, 9, 10, 11)  # Other months

# Extracting energy consumption for each season
summer_kwh <- monthly_kwh[summer_months]
winter_kwh <- c(monthly_kwh[12], monthly_kwh[1], monthly_kwh[2])
other_kwh <- monthly_kwh[other_months]

# Ensuring there are enough data points
ensure_min_data <- function(kwh_values) {
  if (length(kwh_values) < 3) {
    kwh_values <- rep(mean(kwh_values, na.rm = TRUE), 3)  # Fill with mean if too short
  }
  return(kwh_values)
}

summer_kwh <- ensure_min_data(summer_kwh)
winter_kwh <- ensure_min_data(winter_kwh)
other_kwh <- ensure_min_data(other_kwh)

# Creating time series with correct frequency
summer_ts <- ts(summer_kwh, start = c(2010, 6), frequency = 12)
winter_ts <- ts(winter_kwh, start = c(2010, 12), frequency = 12)
other_ts <- ts(other_kwh, start = c(2010, 3), frequency = 12)

# Applying Exponential Smoothing (ETS) models
summer_ets <- ets(summer_ts, model = "ZZZ")
winter_ets <- ets(winter_ts, model = "ZZZ")
other_ets <- ets(other_ts, model = "ZZZ")

# Forecasting with a longer horizon (h = 6)
forecast_horizon <- 6
summer_forecast <- forecast(summer_ets, h = forecast_horizon)
winter_forecast <- forecast(winter_ets, h = forecast_horizon)
other_forecast <- forecast(other_ets, h = forecast_horizon)

# Using Function to plot forecasts
plot_forecast <- function(forecast_obj, title_text) {
  autoplot(forecast_obj) +
    labs(title = title_text,
         subtitle = "Forecasted vs Actual Energy Consumption Over Time",
```

```
        x = "Month",
        y = "Energy Consumption (KWH)") +
    theme_minimal(base_size = 15) +
    theme(plot.title = element_text(hjust = 0.5, face = "bold"),
          plot.subtitle = element_text(hjust = 0.5, face = "italic"),
          axis.title.x = element_text(face = "bold"),
          axis.title.y = element_text(face = "bold"))
}

# Generating plots for each season
plot_summer <- plot_forecast(summer_forecast, "Exponential Smoothing Forecast for
Summer Season")
plot_winter <- plot_forecast(winter_forecast, "Exponential Smoothing Forecast for Winter
Season")
plot_other <- plot_forecast(other_forecast, "Exponential Smoothing Forecast for Other
Season")

# Arranging the plots in a single output (stacked format)
grid.arrange(plot_summer, plot_winter, plot_other, ncol = 1)
```