



PREDICTING BUILDING ENERGY CONSUMPTION IN THE UNITED STATES

STAT8101 – SAMPLING DESIGN AND SURVEY




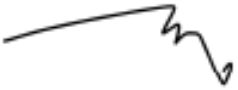


Lecturers: Georgy Sofronov
Hassan Bahrami

Prepared by: Gunj Chinbat 44197462
Jiyeon Kim 47518723
Nhu Duong 48040622
Niyati Niyati 47943319
Parinya Sodsai 47817283

Macquarie University
Sydney, Australia 2024

DECLARATION

We declare that this paper is an original work submitted to STAT8101 Sampling Design and Survey by the following group members who have all actively made a contribution. The declaration of each student along with their specific contribution is as follows:

Section	Student Name	Student ID	Signature
Project Description	Gunj Chinbat	44197462	
Data Description & Survey Methodology	Niyati Niyati	47943319	
Exploratory Data Analysis	Parinya Sodsai	47518723	
Data Cleaning and Transformation	Jiyeon Kim	47518723	
Regression Analysis	Nhu Duong	48040622	
Discussion, Conclusion & Final Editing	Gunj Chinbat	44197462	

ABSTRACT

This paper investigates the key factors influencing energy consumption in buildings across the United States and aims to develop a predictive model that can be used to forecast future energy consumption. By analysing a comprehensive dataset of building characteristics and energy usage provided by the 2018 Commercial Buildings Energy Consumption Survey (CBECS), we seek to identify the key determinants of energy consumption. Through regression analysis, we will construct a statistical model that can accurately predict energy usage based on these factors. Discussion on the findings from the exploratory analysis and the regression analysis is provided, followed by a recommendation for future research works.

TABLE OF CONTENTS:

Table of Contents

1. Introduction	4
2. Data Description	4
3. Survey Methodology	5
4. Data Analysis	6
5. Discussion and Conclusion	14
6. References	15
7. Appendix	15

1. Introduction

Energy consumption is a fundamental aspect of human activity, powering everything from our homes and workplaces to transportation and industry. As the global population continues to grow and economies expand, the demand for energy is also increasing at an unprecedented rate. This rising energy consumption has significant implications for both environmental sustainability and economic development including climate change, air pollution, resource depletion, and high costs. In the United States, buildings account for a significant portion of the nation's total energy usage, making it imperative to conduct analyses to understand the factors driving this consumption. Findings from these analyses can then be served as a valuable tool for policymakers, building managers, and energy utilities in making informed decisions regarding energy efficiency, renewal energy integration, and grid management.

2. Data Description

The **2018 Commercial Buildings Energy Consumption Survey (CBECS)** is a public-use microdata file that contains detailed information on individual commercial buildings across the United States. The dataset includes 6,436 records with each record representing a single building from the sample. This sample is designed to represent an estimated 5.9 million commercial buildings across the 50 states and the District of Columbia. The variables consist of several aspects of building characteristics, climate zones, and energy consumption, each providing insights into the energy profile of different commercial buildings.

For our project, we have selected the following key variables from the CBECS dataset:

Table 1: Data Description

Variable	Description
ELCNS	The total electricity consumed by the building, measured in kWh
CENDIV	Census division classification - 1=New England; 2=Middle Atlantic; 3=East North Central; 4=West North Central; 5=South Atlantic; 6=East South Central; 7=West South Central; 8=Mountain; 9=Pacific
REGION	The geographic region where the building is located - 1=Northeast; 2=Midwest; 3=South; 4=West
SQFT	The total area of the building, measured in square feet
FINALWT	Nonresponse Adjusted Weight

- i. **ELCNS (Total Electricity Consumption):** This variable measures the total electricity consumed by the building, recorded in kilowatt-hours (kWh). This is a crucial variable

for analysing the energy efficiency of buildings and understanding the relationship between building characteristics and energy consumption.

- ii. **CENDIV (Census Division):** This variable identifies the specific census division where the building is located, providing a more granular classification than the REGION variable. It helps further explore regional differences in building energy consumption and size, allowing for more detailed geographic analysis.
- iii. **REGION:** This variable categorises the geographic region where the building is located. It allows us to analyse patterns or trends in energy consumption and building size across different regions of the U.S.
- iv. **SQFT (Square Foot):** This variable records the total area of the building in square feet. The size of the building is a key factor in energy consumption analysis, as larger buildings typically have higher energy usage. This variable allows us to investigate the correlation between building size and electricity consumption.
- v. **FINALWT (Final Weight):** The **FINALWT** variable provides the nonresponse-adjusted weight assigned to each record. It adjusts for nonresponse bias and ensures that the sample accurately represents the population of commercial buildings in the U.S. This variable is critical for ensuring that our analyses yield statistically valid inferences that can be generalised to the larger population of U.S. commercial buildings.

3. Survey Methodology

In the United States, an estimated 5.9 million commercial buildings existed in 2018. Surveying all of these buildings would have been impractical and costly, so a probability sampling method was used to represent the entire population. To ensure a statistically valid sample, each building had a known and unique probability of selection. A key step in this process involved creating a sampling frame, comprising two components: the area frame and the list frames. Since no comprehensive list of U.S. commercial buildings exists, the area frame was constructed by dividing the country into geographical segments. Trained staff to identify all commercial buildings in selected areas in person or using a Geographic Information System (GIS) tool for virtual listing. This multi-stage area probability sampling ensured that all buildings in the selected geographic areas had a chance of being included.

The construction of the area frame began by dividing the U.S. into 687 Primary Sampling Units (PSUs), with each PSU representing a county or a group of counties. PSUs were chosen based on the level of local commerce, which is closely correlated with the number of commercial

buildings. This selection method, called Probability Proportionate to Size (PPS) sampling, ensured that PSUs with the most commercial activity, such as major cities, were selected with certainty. A total of 151 PSUs were selected. These PSUs were then further divided into Secondary Sampling Units (SSUs), composed of census tracts, to allow for more detailed building selection. In total, 8,559 SSUs were identified, and a PPS sample of 764 SSUs was drawn to represent the commercial building population.

In these selected SSUs, the final stage of building identification was conducted. Traditionally, staff would walk or drive through the area to list all commercial buildings. However, in 2018, a new virtual listing technique was introduced, allowing buildings to be identified remotely using GIS, satellite imagery, and other databases. This method was tested against traditional field listing and was found to be highly efficient, reducing costs and time while maintaining quality. Approximately 256,000 buildings were identified through this combination of field and virtual listing.

To ensure adequate representation of large buildings (over 200,000 square feet), which tend to consume more energy, a list frame was used in addition to the area frame. This list was compiled from five different sources, including administrative databases of hospitals, federal buildings, airports, and other large facilities. After processing and de-duplicating these lists, around 23,000 large buildings were added to the sampling frame.

The combined area and list frames were used to draw a sample of 16,000 buildings, with larger buildings sampled at higher rates to account for their higher energy consumption. The base weight of every building in the sample was determined by inversely calculating its chance of selection. To ensure the final sample was representative of all commercial buildings in the United States, these base weights were adjusted for non-responses and ineligible buildings. Finally, the adjusted base weights of the sampled buildings were summed to estimate the total number of commercial buildings in the United States.

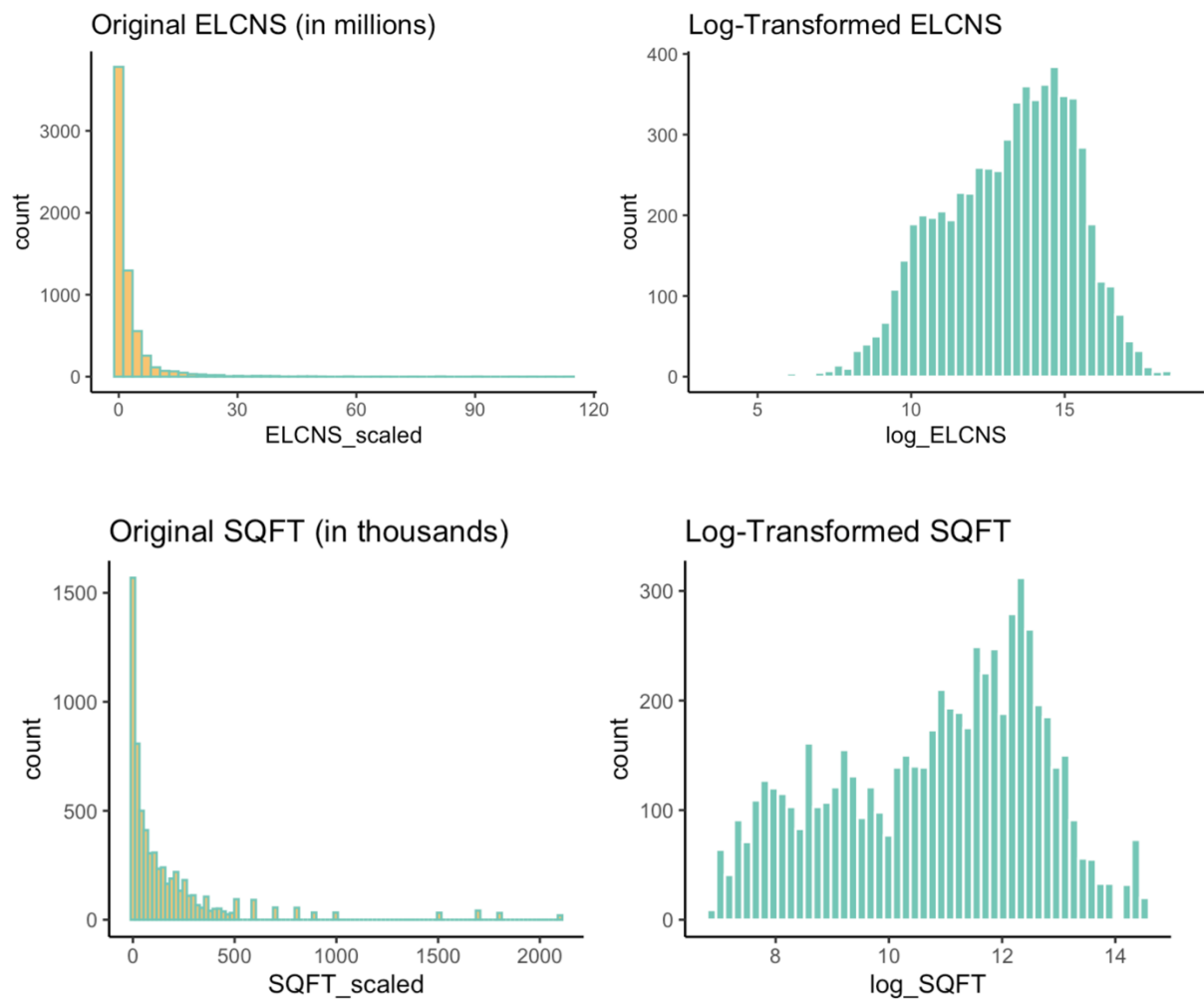
4. Data Analysis

4.1. Descriptive Statistics

4.1.1. Exploring continuous variables

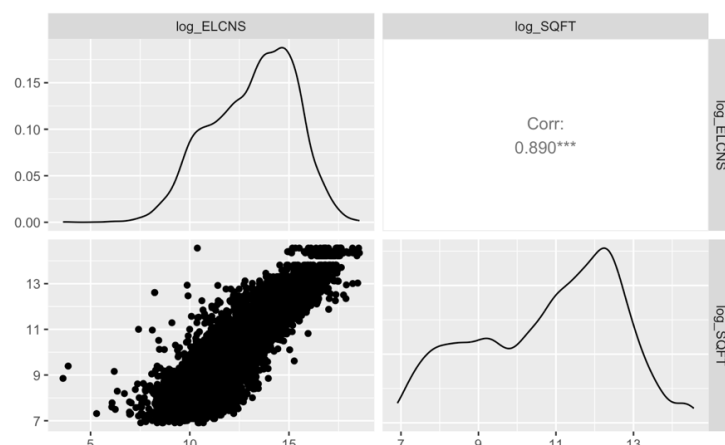
The histogram plots of **ELCNS** and **SQFT** revealed that both variables were heavily right-skewed, suggesting a transformation is required to achieve a normal distribution which is more

suitable for statistical modelling. We applied log-transformation to these variables and plotted the histograms of them to show the comparison.



Correlation Between Log-Transformed ELCNS and Log-Transformed SQFT

To explore the relationship between $\log(\text{ELCNS})$ and $\log(\text{SQFT})$, we calculated the **correlation coefficient** and plotted the scatter plot.



The **correlation coefficient** is **0.890**, indicating a strong positive relationship between the two variables. The scatter plot (bottom-left) shows a clear **positive linear relationship** which explains that the electricity consumption tends to increase as the square footage of a building increases.

Applying Weights to Correlation

A comparison of the weighted and unweighted summary for **log(ELCNS)** and **log(SQFT)** and the relationship between them revealed important differences:

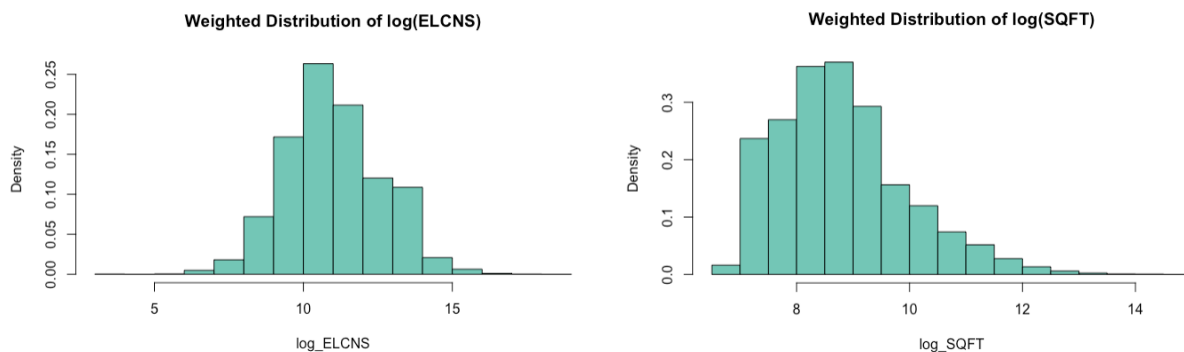


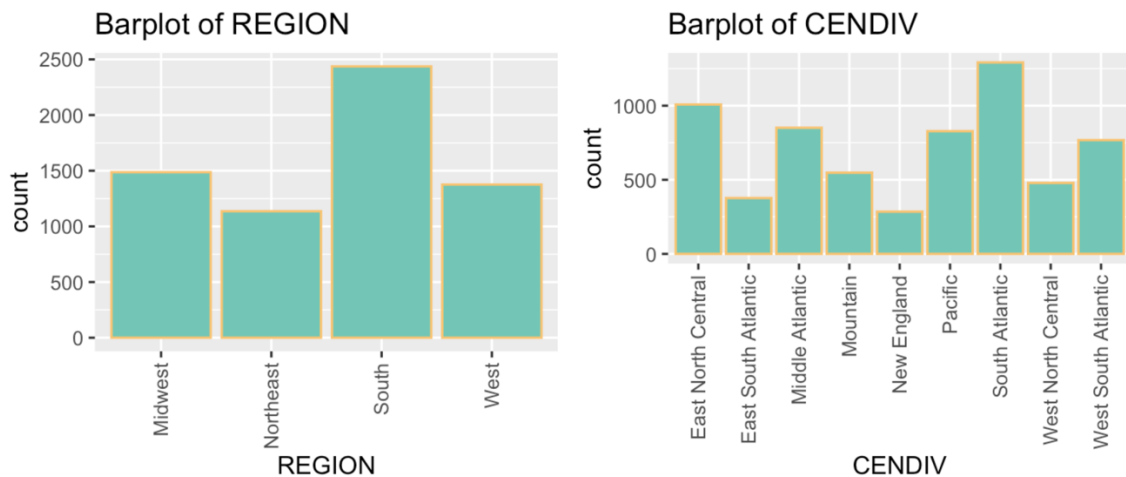
Table 2: Weighted Correlation

Variable	Mean	Variance	Weighted_correlation
Unweighted (log(ELCNS))	13.17	4.43	0.61
Weighted (log(ELCNS))	10.99	2.57	

The reduction in the correlation from **0.89** (unweighted) to **0.61** (weighted) indicates that the strength of the relationship between building size and electricity consumption decreases when we adjust for survey design and nonresponse bias. This suggests that the unweighted data may have overestimated the strength of the relationship due to overrepresentation of certain groups. The weighted analysis provides a more accurate picture by ensuring that each observation's influence is proportional to its representation in the population.

4.1.2. Exploring categorical variables

In this analysis, we focus on the variables **REGION** and **CENDIV** to investigate geographic influences on electricity consumption patterns. The bar plots below summarise the distribution of buildings across these geographic categories.



The bar plot of REGION shows that South region has the most representation in this dataset, followed by Midwest, West and Northeast. On the other hand, the bar plot of CENDIV illustrates that South Atlantic division has the highest number of observations while New England has the least.

4.2. Hypothesis Testing

4.2.1. Two-Way Table between REGION and CENDIV

Two-way table between REGION and CENDIV

	New England	Middle Atlantic	East North Central	West North Central	South Atlantic	East South Central	West South Central	Mountain	Pacific
Northeast	285	851	0	0	0	0	0	0	0
Midwest	0	0	1008	479	0	0	0	0	0
South	0	0	0	0	1292	377	768	0	0
West	0	0	0	0	0	0	0	548	828

The two-way table reveals significant geographic trends in building distribution across regions and census divisions. This could be due to a **multicollinearity** between **REGION** and **CENDIV** as they represent overlapping geographic hierarchies. **CENDIV** is essentially a more granular breakdown of **REGION**, which may introduce redundant information into the model.

- The Northeast shows strong representation in New England and Middle Atlantic, indicating a well-distributed sample in this area.
- The Midwest has notable counts primarily in East North Central.
- The South is heavily represented in the South Atlantic, highlighting concentrated energy consumption patterns, while the West shows balanced representation across the Mountain and Pacific divisions.

These insights can inform energy management strategies tailored to specific geographic needs.

4.2.2. Chi-Square Test

Given the potential for multicollinearity between REGION and CENDIV, we aim to further investigate whether the association between these two geographic variables is statistically significant. To do so, we will conduct a Chi-square test to quantify the relationship between REGION and CENDIV.

H0: There is no association between REGION and CENDIV. Thus, the distribution of floors is independent of building types.

H1: There is an association between REGION and CENDIV. Thus, the distribution of floors depends on building types.

```
Pearson's Chi-squared test with simulated p-value (based on
2000 replicates)

data: two_way_table
X-squared = 19308, df = NA, p-value = 0.0004998
```

The results indicate a strong association between REGION and CENDIV. This reinforces the concern of multicollinearity, suggesting that only one of these two variables should be included in the further analysis. We will only include CENDIV in our regression analysis.

4.3. Handling Missing Data

4.3.1. Nonresponse rate

In the original survey, the nonresponse rate was calculated using a partially weighted response rate, which accounted for subsampling procedures. The overall partially weighted response rate for the survey was 54.7%, with nonresponse rates varying depending on the type of buildings surveyed.

For our specific analysis, we checked for missing values across the selected variables. the nonresponse rate for ELCNS was calculated as 1.23%, indicating that a small proportion of buildings was missing ELCNS values. The figure below provides a visualization of the missing data pattern:

Figure 1. Missing data visualisation - ELCNS

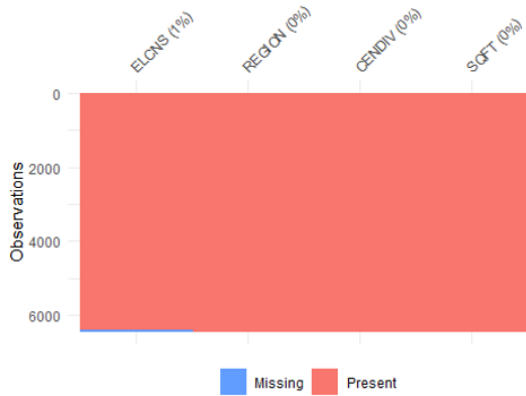


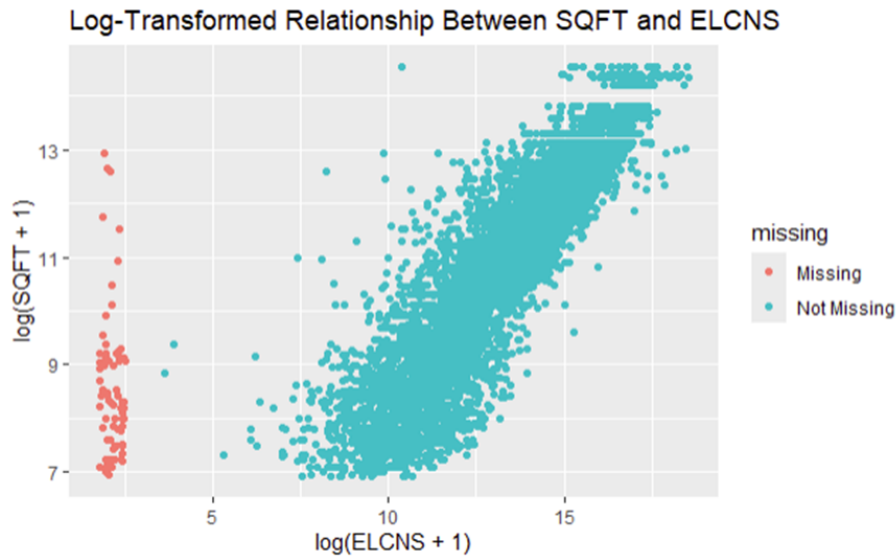
Table 3: Missing Value Summary

Variable	Missing_Count
ELCNS	79
REGION	0
CENDIV	0
SQFT	0

4.3.2. Association with Other Variables

We assessed whether the missingness in ELCNS was associated with other key variables. In Figure 2, red dots on the left represent observations where ELCNS is missing. They are scattered across the range of building sizes, suggesting that missingness in ELCNS might be random and not directly linked to SQFT. This means the missing values are unlikely to introduce bias related to building size.

Figure 2. Log-Transformed relationship between



A Chi-square test was used to check for relationships between the missingness and categorical variables. The results showed:

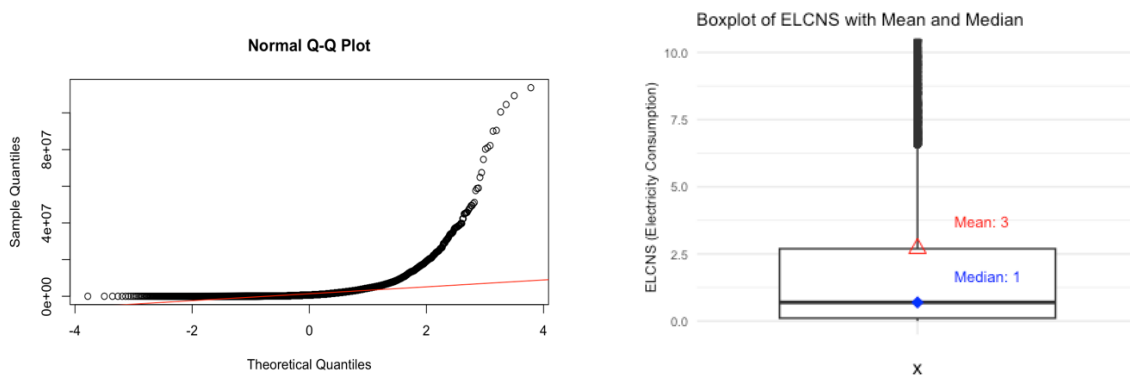
- **No significant relationship** between **REGION** and missing ELCNS values ($p = 0.2681$).

- **A significant relationship** between **CENDIV** and missing ELCNS values ($p = 0.03942$), suggesting that missingness in electricity consumption is associated with specific divisions, as defined by CENDIV.

4.3.3. Type of Nonresponse

To determine the type of nonresponse, we conducted a Little's MCAR test on the dataset excluding the weight variable FINALWT. The test result returned a $p\text{-value} = 0.0000137$, which is statistically significant. This implies that the missing data in our dataset is not completely random. Therefore, the missing data is likely Missing at Random (MAR) or Missing Not at Random (MNAR), depending on observed variables, and must be accounted for in our analysis.

4.3.4. Imputation for Missing Values



As the Figure 3 displays the boxplot and Q-Q plot for ELCNS, the mean is 2,681,229, while the median is 695,373. These confirmed the presence of extreme outliers and a non-normal distribution, justifying the use of weighted median imputation for ELCNS. Furthermore, given the association between CENDIV and missingness in ELCNS, we chose to perform group-wise median imputation based on CENDIV. This approach ensured that missing electricity consumption values were imputed based on the median consumption within each specific geographic division.

4.4. Regression Analysis

4.4.1. Model Selection Process

We employed a forward model selection method to develop the best performing model and used AIC as our selection metric since AIC is more appropriate in finding a prediction model

while BIC is useful in selecting a model that is correct. Table 4 shows the result of the selection process.

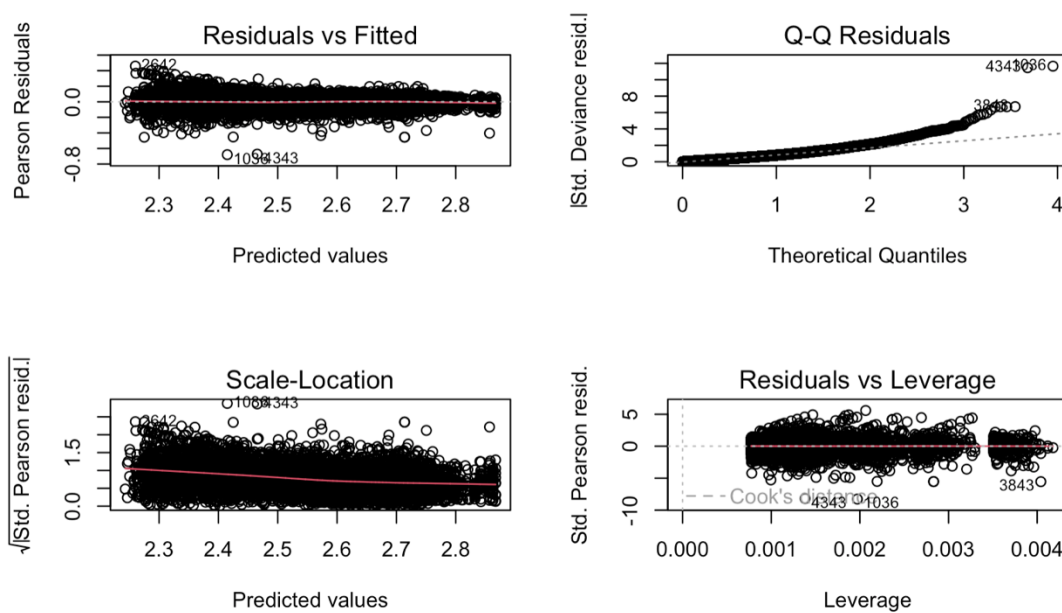
Table 4: Forward Model Selection Output

Model	Model_Type	Link_Function	AIC
$\log(\text{ELCNS}) \sim \text{CENDIV}$	GLM	Log	28128.72
$\log(\text{ELCNS}) \sim \log(\text{SQFT})$	GLM	Identity	19686.87
$\log(\text{ELCNS}) \sim \log(\text{SQFT})$	GLM	Inverse	19685.11
$\log(\text{ELCNS}) \sim \log(\text{SQFT})$	GLM	Log	19572.04
$\log(\text{ELCNS}) \sim \log(\text{SQFT})$	Cluster Sampling Design (with replacement)	-	31780.00
$\log(\text{ELCNS}) \sim \log(\text{SQFT}) + \text{CENDIV}$	GLM	-	19500.00

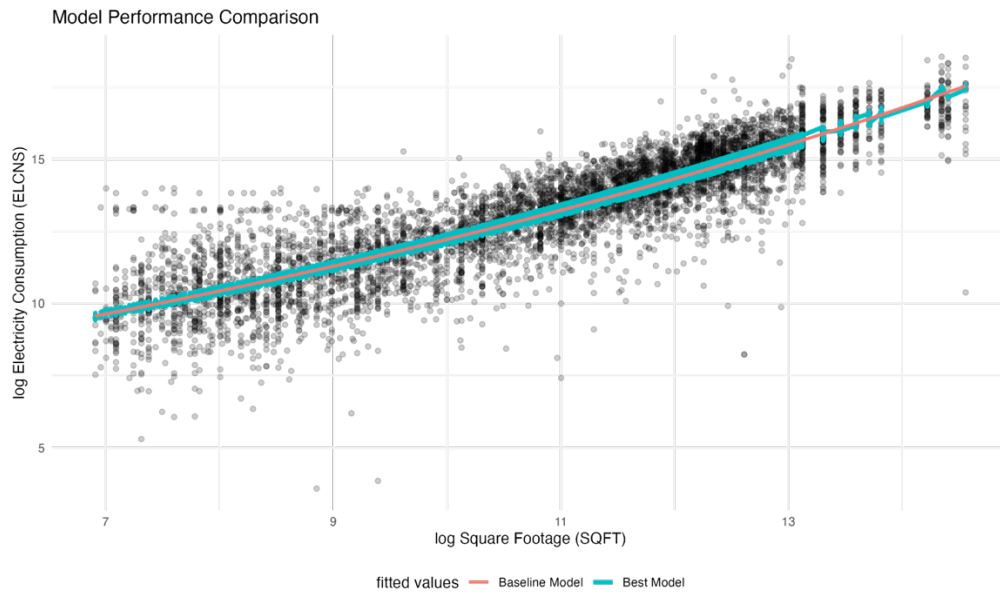
We can see the results above that the model with $\log(\text{ELCNS}) + \log(\text{SQRT}) + \text{CENDIV}$ has the least AIC of 19,500. This is then selected as our final model.

4.4.2. Diagnostic checks

Now that we have selected our final model, we will conduct diagnostic checks to evaluate the validity of our model.



The diagnostic plots show that the model is quite good. The Residual vs Fitted plot shows all the points scattered randomly about zero reference line. Therefore, the assumption of linear relationship holds.



The model performance comparison shows that the best model $\log(\text{ELCNS}) + \log(\text{SQFT}) + \text{CENDIV}$ performs better than the base model. It captures more of the variation of the data compared to the baseline model. Overall, our best model shows that it can be used to predict the electricity usage of the building.

5. Discussion and Conclusion

We have conducted an exploratory data analysis into some of the key characteristics of buildings for energy consumption and carried out a regression analysis to develop a model that can be used to predict future energy usage. The exploratory analysis revealed a strong correlation between the energy consumption and the size of the buildings, explained in SQFT in this dataset. The geographic location of a specific building is also found to be a significant predictor of the total usage. For our regression analysis, we applied a forward model selection method with AIC as the performance metric. The method found that a model consisting of square footage and census division to be the best model with the least AIC score. We then conducted diagnostic checks to evaluate the validity of our model which showed the model was a good fit as the residuals are randomly scattered about zero reference line.

In future studies into energy consumption of buildings, researchers could consider additional factors such as occupancy and usage patterns, buildings orientation in relation to solar exposure, insulation and heating, ventilation, and air conditioning (HVAC) systems to improve the accuracy of the model.

6. References

U.S. Energy Information Administration. (n.d.). *Commercial Buildings Energy Consumption Survey (CBECS)*. Retrieved from U.S. Energy Information Administration:
<https://www.eia.gov/consumption/commercial/data/2018/index.php?view=microdata>

7. Appendix

```
# Load necessary libraries
library(dplyr)
library(tidyr)
library(VIM)
library(gridExtra)
library(kableExtra)
library(tidyverse)
library(GGally)
library(ggplot2)
library(car)
library(DHARMA)
library(MASS)
library(janitor)
library(survey) # applying survey design, handling weights
library(readr)
library(broom)

# Import your dataset
data <- read.csv("cbeecs2018_final_public.csv")

data_desc <- data.frame(
  Variable = c("ELCNS", "CENDIV", "REGION", "SQFT", "FINALWT"),
  Description = c("The total electricity consumed by the building, measured in kWh",
                  "Census division classification - 1=New England; 2=Middle Atlantic; 3=East North Central; 4=West North Central; 5=South Atlantic; 6=East South Central; 7=West South Central; 8=Mountain; 9=Pacific",
                  "The geographic region where the building is located - 1=Northeast; 2=Midwest; 3=South; 4=West",
                  "The total area of the building, measured in square feet",
                  "Nonresponse Adjusted Weight")
)

knitr::kable( data_desc[1:5, 1:2], digits = 3,
              booktabs = TRUE, format = 'latex',
              position = "h!", caption = "Data Description" ) %>%
  kableExtra::kable_styling( position = "center") %>%
  column_spec(1, "1in") %>%
  column_spec(2, "5in")
```



```

# Select relevant columns
selected_columns <- c(
  "ELCNS", "SQFT", "REGION", "CENDIV", "FINALWT"
)
extracted_data <- data %>% dplyr::select(all_of(selected_columns))

# Convert categorical variables to factors
extracted_data <- extracted_data %>%
  mutate(
    REGION = as.factor(REGION),
    CENDIV = as.factor(CENDIV),
  )

extracted_data <- extracted_data %>%
  mutate(REGION = ifelse(REGION == 1, "Northeast", REGION),
    REGION = ifelse(REGION == 2, "Midwest", REGION),
    REGION = ifelse(REGION == 3, "South", REGION),
    REGION = ifelse(REGION == 4, "West", REGION))

extracted_data <- extracted_data %>%
  mutate(CENDIV = ifelse(CENDIV == 1, "New England", CENDIV),
    CENDIV = ifelse(CENDIV == 2, "Middle Atlantic", CENDIV),
    CENDIV = ifelse(CENDIV == 3, "East North Central", CENDIV),
    CENDIV = ifelse(CENDIV == 4, "West North Central", CENDIV),
    CENDIV = ifelse(CENDIV == 5, "South Atlantic", CENDIV),
    CENDIV = ifelse(CENDIV == 6, "East South Atlantic", CENDIV),
    CENDIV = ifelse(CENDIV == 7, "West South Atlantic", CENDIV),
    CENDIV = ifelse(CENDIV == 8, "Mountain", CENDIV),
    CENDIV = ifelse(CENDIV == 9, "Pacific", CENDIV)
  )

head(extracted_data, 10)

extracted_data <- extracted_data %>%
  mutate(ELCNS_scaled = ELCNS/1000000)

extracted_data <- extracted_data %>%
  mutate(SQFT_scaled = SQFT/1000)

head(extracted_data, 10)

g1 = extracted_data |> ggplot(aes(x = ELCNS_scaled)) +
  geom_histogram(bins=50,
    color = '#73c6b6', fill = '#f8c471') +
  #scale_color_grey()+scale_fill_grey() +
  labs(title = "Original ELCNS (in millions)") +
  theme_classic()

extracted_data <- extracted_data %>%
  mutate(log_ELCNS = log(ELCNS))

g2 = extracted_data |> ggplot(aes(x = log_ELCNS)) +
  geom_histogram(bins=50,
    color = 'white', fill = '#73c6b6') +

```

```

  labs(title = "Log-Transformed ELCNS") +
  theme_classic()

grid.arrange(g1, g2, ncol=2)

g3 = extracted_data |> ggplot(aes(x = SQFT_scaled)) +
  geom_histogram(bins=100,
                 color = '#73c6b6', fill = '#f8c471') +
  #scale_color_grey()+scale_fill_grey() +
  labs(title = "Original SQFT (in thousands)") +
  theme_classic()

extracted_data <- extracted_data %>%
  mutate(log_SQFT = log(SQFT))

g4 = extracted_data |> ggplot(aes(x = log_SQFT)) +
  geom_histogram(bins=50,
                 color = 'white', fill = '#73c6b6') +
  labs(title = "Log-Transformed SQFT") +
  theme_classic()

grid.arrange(g3, g4, ncol=2)

extracted_data |> ggpairs(extracted_data, columns=c('log_ELCNS', 'log_SQFT
')) +
  theme_minimal()

# Group-wise median imputation based on CENDIV
imputed_data <- extracted_data %>%
  group_by(CENDIV) %>%
  mutate(ELCNS = ifelse(is.na(ELCNS), median(ELCNS, na.rm = TRUE), ELCNS))

# Check if missing values have been imputed
sum(is.na(imputed_data$ELCNS))

head(imputed_data)

```

5. Nonresponse Adjustments

FINALWT (Nonresponse Adjusted Weight) is specifically designed to adjust for nonresponse bias.

```

# Use imputed_data and apply Log transformation to ELCNS and SQFT
transformed_data <- imputed_data %>%
  mutate(log_ELCNS = log(ELCNS + 1), # Log transformation of ELCNS
         log_SQFT = log(SQFT + 1)) # Log transformation of SQFT

str(transformed_data)

# Define the survey design with clustering and weights
svy_design <- svydesign(ids = ~REGION + CENDIV,
                      data = transformed_data,
                      weights = ~FINALWT)

```

```

# 1. Weighted Distribution Graph for Log_ELCNS
svyhist(~log_ELCNS, design = svy_design,
        main = "Weighted Distribution of log(ELCNS)", col = "blue")

# 1. Weighted Distribution Graph for Log_SQFT
svyhist(~log_SQFT, design = svy_design,
        main = "Weighted Distribution of log(SQFT)", col = "green")

# 2. Weighted Correlation between Log_ELCNS and Log_SQFT
# Weighted means of Log_ELCNS and Log_SQFT
weighted_mean_log_ELCNS <- svymean(~log_ELCNS, svy_design)
weighted_mean_log_SQFT <- svymean(~log_SQFT, svy_design)

# Weighted variances of Log_ELCNS and Log_SQFT
weighted_var_log_ELCNS <- svyvar(~log_ELCNS, svy_design)
weighted_var_log_SQFT <- svyvar(~log_SQFT, svy_design)

# Weighted covariance between Log_ELCNS and Log_SQFT
cov_log_ELCNS_SQFT <- svyvar(~log_ELCNS + log_SQFT, svy_design)[1, 2]

# Calculate weighted correlation
weighted_correlation <- cov_log_ELCNS_SQFT / (sqrt(weighted_var_log_ELCNS)
* sqrt(weighted_var_log_SQFT))

# Print the weighted correlation
print(weighted_correlation)

cat("Weighted mean log_ELCNS:", weighted_mean_log_ELCNS, "\n")
cat("Weighted variance log_ELCNS:", weighted_var_log_ELCNS, "\n")
cat("Weighted mean log_SQFT:", weighted_mean_log_SQFT, "\n")
cat("Weighted variance log_SQFT:", weighted_var_log_SQFT, "\n")

# Custom function to calculate weighted correlation
weighted_corr <- function(x, y, weights) {
  # Calculate weighted means
  mean_x <- sum(weights * x) / sum(weights)
  mean_y <- sum(weights * y) / sum(weights)

  # Calculate weighted covariance
  cov_xy <- sum(weights * (x - mean_x) * (y - mean_y)) / sum(weights)

  # Calculate weighted standard deviations
  sd_x <- sqrt(sum(weights * (x - mean_x)^2) / sum(weights))
  sd_y <- sqrt(sum(weights * (y - mean_y)^2) / sum(weights))

  # Return weighted correlation
  return(cov_xy / (sd_x * sd_y))
}

# Applying the custom function to your data
weighted_correlation <- weighted_corr(transformed_data$log_ELCNS,
                                     transformed_data$log_SQFT,

```

```

transformed_data$FINALWT)
print(weighted_correlation)

# Scatter plot of log_ELCNS vs log_SQFT with points sized by weights
ggplot(transformed_data, aes(x = log_SQFT, y = log_ELCNS, size = FINALWT))
+
  geom_point(alpha = 0.6) + # Adding transparency for better visibility
  theme_minimal() + # Clean theme
  labs(
    title = "Weighted Scatter Plot of log_ELCNS vs log_SQFT",
    x = "log(SQFT) - Building Size",
    y = "log(ELCNS) - Electricity Consumption",
    size = "Weight (FINALWT)"
  ) +
  scale_size_continuous(range = c(1, 6)) # Adjust point size range

# Unweighted mean and variance for log_ELCNS
unweighted_mean_log_ELCNS <- mean(transformed_data$log_ELCNS, na.rm = TRUE)
unweighted_variance_log_ELCNS <- var(transformed_data$log_ELCNS, na.rm = TRUE)

# Unweighted mean and median for log_SQFT
unweighted_mean_log_SQFT <- mean(transformed_data$log_SQFT, na.rm = TRUE)
unweighted_variance_log_SQFT <- var(transformed_data$log_SQFT, na.rm = TRUE)

# Print results
cat("Unweighted mean log_ELCNS:", unweighted_mean_log_ELCNS, "\n")
cat("Unweighted variance log_ELCNS:", unweighted_variance_log_ELCNS, "\n")
cat("Unweighted mean log_SQFT:", unweighted_mean_log_SQFT, "\n")
cat("Unweighted variance log_SQFT:", unweighted_variance_log_SQFT, "\n")

g6 = extracted_data |> ggplot(aes(x = REGION)) +
  geom_bar(color = '#f8c471', fill = '#73c6b6') +
  #scale_color_grey()+scale_fill_grey() +
  labs(title = "Barplot of REGION") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

g7 = extracted_data |> ggplot(aes(x = CENDIV)) +
  geom_bar(color = '#f8c471', fill = '#73c6b6') +
  #scale_color_grey()+scale_fill_grey() +
  labs(title = "Barplot of CENDIV") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

grid.arrange(g6,g7,ncol=2)

mean_value <- mean(extracted_data$ELCNS, na.rm = TRUE)
median_value <- median(extracted_data$ELCNS, na.rm = TRUE)
cat("Mean:", mean_value, "\nMedian:", median_value)

# Load required libraries
library(knitr)

two_way_table <- table(extracted_data$REGION, extracted_data$CENDIV)

```

```

# Convert the two-way table to a data frame for better visualization
two_way_table_df <- as.data.frame.matrix(two_way_table)

# Print the table using kable with added styling for better contrast
kable(two_way_table_df, caption = "Two-way table between REGION and CENDIV") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"), full_width = F) %>%
  row_spec(0, bold = TRUE, color = "black", background = "lightgray") # Header row styling

# Perform Chi-Square test for association between PBA and NFLOOR
chi_square_test <- chisq.test(two_way_table, simulate.p.value = TRUE)

# Print the results of the Chi-Square test
print(chi_square_test)

# Q-Q plot to check for normality and outliers
qqnorm(extracted_data$ELCNS)
qqline(extracted_data$ELCNS, col = "red")

# Calculate mean and median
mean_value <- mean(extracted_data$ELCNS_scaled, na.rm = TRUE)
median_value <- median(extracted_data$ELCNS_scaled, na.rm = TRUE)
median_value

class(extracted_data$ELCNS_scaled)

# Boxplot with mean and median displayed as text
ggplot(extracted_data, aes(x = "", y = ELCNS_scaled)) +
  geom_boxplot() +
  coord_cartesian(ylim = c(0, 10)) +

  # Add mean point
  stat_summary(fun = mean, geom = "point", shape = 2, color = "red", size = 3,
    show.legend = TRUE) +

  # Add median point
  stat_summary(fun = median, geom = "point", shape = 18, color = "blue", size = 3,
    show.legend = TRUE) +

  # Annotate mean value above the point
  annotate("text", x = 1.1, y = mean_value + 1, label = paste("Mean:", round(mean_value, 0)),
    color = "red", size = 3, hjust = 0) +

  # Annotate median value below the point
  annotate("text", x = 1.1, y = median_value + 1, label = paste("Median:", round(median_value, 0)),
    color = "blue", size = 3, hjust = 0) +

```

```

# Adjust title and axis label sizes
ggtitle("Boxplot of ELCNS with Mean and Median") +
ylab("ELCNS (Electricity Consumption)") +
theme_minimal() +
theme(
  plot.title = element_text(size = 10), # Smaller title
  axis.title.y = element_text(size = 8), # Smaller y-axis label
  axis.text = element_text(size = 7)    # Smaller tick labels
)

# Regression Analysis
predictors_list = c("CENDIV", "SQFT")

run_gamma_regression <- function(var) {
  formula <- as.formula(paste("log(ELCNS) ~", var))

  model <- tryCatch({
    glm(formula, data = data, family = Gamma(link = "log"))
  }, error = function(e) {
    message("Error in fitting model for ", var, ": ", e$message)
    return(NULL)
  })

  if (!is.null(model)) {
    tidy_results <- tidy(model) %>% mutate(variable = var)
    aic_value <- AIC(model)
    return(list(tidy_results = tidy_results, aic = aic_value))
  } else {
    return(NULL)
  }
}

#read data

data <- transformed_data

head(data)

```

Check each predictors

#List of the predictors

```

predictors_list = c("CENDIV", "SQFT")

#Function run regression on all the variables
run_gamma_regression <- function(var) {
  formula <- as.formula(paste("log(ELCNS) ~", var))

  model <- tryCatch({
    glm(formula, data = data, family = Gamma(link = "log"))
  }, error = function(e) {
    message("Error in fitting model for ", var, ": ", e$message)
    return(NULL)
  })
}

```

```

if (!is.null(model)) {
  tidy_results <- tidy(model) %>% mutate(variable = var)
  aic_value <- AIC(model)
  return(list(tidy_results = tidy_results, aic = aic_value))
} else {
  return(NULL)
}
}

# Run regressions for all variables
results_glm_pli <- lapply(predictors_list, run_gamma_regression)

# Combine tidy results into a single dataframe
all_results_pli <- do.call(rbind, lapply(results_glm_pli, function(x) x$tidy_results))

# Extract AIC values
aic_values <- sapply(results_glm_pli, function(x) x$aic)

# Create a dataframe with variables and their aIC values
aic_df <- data.frame(variable = predictors_list, aic = aic_values)

# Combine the filtered results with BIC values
filtered_results <- all_results_pli %>%
  dplyr::select(variable, p.value, term) %>%
  filter(term != "(Intercept)") %>%
  mutate(p.value = round(p.value, 5)) %>%
  left_join(aic_df, by = "variable")

print(filtered_results[order(filtered_results$aic, decreasing = FALSE), ])

```

AUTO GLM

```

auto_run_glm <- function(data, response_var, list_variable, list_link_function) {
  library(broom)

  results <- data.frame()

  for (var in list_variable) {
    for (link in list_link_function) {
      formula <- as.formula(paste(response_var, "~", var))

      model <- tryCatch({
        glm(formula, family = Gamma(link = link), data = data)
      }, error = function(e) {
        return(NULL)
      })

      if (!is.null(model)) {
        model_summary <- summary(model)
        coef_table <- tidy(model)

        for (i in 1:nrow(coef_table)) {
          results <- rbind(results, data.frame(

```

```

        Variable = var,
        Link = link,
        Parameter = coef_table$term[i],
        Estimate = coef_table$estimate[i],
        P_value = round(coef_table$p.value[i],4),
        AIC = model$aic
      ))
    }
  }
}

return(results)
}

list_variable <- c("SQFT", "CENDIV")
list_link_function <- c("log", "inverse", "identity")
result_table <- auto_run_glm(data, "ELCNS", list_variable, list_link_function)
print(result_table)

result_table[order(result_table$AIC), ]

```

AUTO glm with log(ELCNS)

```

list_variable_log <- c("log(SQFT)", "CENDIV", "REGION")
list_link_function <- c("log", "inverse", "identity")
result_table_log <- auto_run_glm(data, "log(ELCNS)", list_variable_log, list_link_function)
print(result_table_log)

result_table_log[order(result_table_log$AIC), ]

```

Conclusion:

Single glm model with $\log(\text{ELCNS}) \sim \log(\text{SQFT})$ with log link has the lowest AIC. Therefore we will use it as baseline model.

survey design model

```

svy_design <- svydesign(ids = ~CENDIV, data = data, weights = ~FINALWT)

svyglm(log(ELCNS) ~ log(SQFT), design = svy_design)

svyglm(log(ELCNS) ~ CENDIV, design = svy_design)

svyglm(ELCNS ~ CENDIV, design = svy_design)

svyglm(log(ELCNS) ~ REGION, design = svy_design)

m1_svg <- svyglm(log(ELCNS) ~ log(SQFT) + CENDIV, design = svy_design)

m1_svg

svyglm(log(ELCNS) ~ log(SQFT) + REGION, design = svy_design)

```


Forward model selection

```
m1_sqft_cendiv <- glm(log(ELCNS) ~ log(SQFT) + CENDIV, family = Gamma(link = "log"), data=data)

m1_sqft_region <- glm(log(ELCNS) ~ log(SQFT) + REGION, family = Gamma(link = "log"), data=data)

m1_sqft_cendiv
m1_sqft_region

Anova(m1_sqft_region)

Anova(m1_sqft_cendiv)
```

Final model

```
final_model <- glm(log(ELCNS) ~ log(SQFT) + CENDIV, family = Gamma(link = "log"), data = data)

final_model
```

Models diagnostic

```
scaled.deviance <- final_model$deviance / summary(final_model)$dispersion
scaled.deviance
```

Scaled deviance = 6788.7119 at 6435 d.f

```
resdev_final_model <- residuals(final_model, type = "deviance")

library(patchwork)

par(mfrow = c(2,2))
plot(final_model)

library(gridExtra)

res_final_model <- simulateResiduals(fittedModel = final_model)
q_final_model <- residuals(res_final_model, quantileFunction = qnorm)
p1_f <- ggplot(tibble(q = q_final_model), aes(x = q)) +
  geom_histogram(aes(y = after_stat(density)), bins = 10, fill = "skyblue",
  color = "black") + geom_density(color = "red") +
  labs(title = "Histogram of Simulated Residuals", x = "Residuals", y = "Density")

p2_f <- ggplot(tibble(q = q_final_model), aes(sample = q)) +
  geom_qq() +
  geom_qq_line(color = "red") +
  labs(title = "Q-Q Plot of Simulated Residuals", x = "Theoretical Quantiles", y = "Sample")

grid.arrange(p1_f, p2_f, ncol = 2)
```

Plotting models

```
baseline_model <- glm(log(ELCNS) ~ log(SQFT), family = Gamma(link = "log"), data=data)

# Load required Libraries
library(ggplot2)

# Assuming 'data' is your dataset and models are already fitted

# Generate predictions
data$predicted_best <- predict(final_model, newdata = data, type = "response")
data$predicted_baseline <- predict(baseline_model, newdata = data, type = "response")

# Create a Long format dataset for plotting
plot_data <- tidyr::pivot_longer(data,
                                cols = c(ELCNS, predicted_best, predicted_baseline),
                                names_to = "model",
                                values_to = "value")

plot_data
```

Comparison models without S.E

```
# Create the plot
ggplot(data, aes(x = log(SQFT), y = log(ELCNS))) +
  # Add actual data points
  geom_point(alpha = 0.2, colour = "black") +
  # Add line for best model
  geom_line(aes(y = predicted_best, colour = "Best Model"), size = 1.5) +
  # Add line for baseline model
  geom_line(aes(y = predicted_baseline, colour = "Baseline Model"), size = 1) +
  # Customize colors and Labels
  #scale_color_manual(values = c("Best Model" = "orange", "Baseline Model" = "green"),
  #name = "Model") +
  # Set labels and title
  labs(title = "Model Performance Comparison",
        x = "log Square Footage (SQFT)",
        y = "log Electricity Consumption (ELCNS)") +
  # Customize theme
  scale_color_discrete(name = "fitted values")+
  theme_minimal() +
  theme(legend.position = "bottom")
  ggsave("model_performance_comparison.png", width = 10, height = 6, dpi = 300)
```

Comparison models with Standard Error

```
# Load required Libraries
library(ggplot2)
library(dplyr)
```

```

# Assuming 'data' is your original dataset and models are already fitted

# Create new_data with only ELCNS and SQFT
new_data <- data[, c("ELCNS", "SQFT")]

# Function to calculate prediction intervals
get_pred_int <- function(model, newdata, level = 0.95) {
  preds <- predict(model, newdata, se.fit = TRUE)
  crit <- qt((1 + level) / 2, df = df.residual(model))
  upr <- exp(preds$fit + crit * preds$se.fit)
  lwr <- exp(preds$fit - crit * preds$se.fit)
  fit <- exp(preds$fit)
  return(data.frame(fit = fit, upr = upr, lwr = lwr))
}

# Calculate prediction intervals
best_pred <- get_pred_int(final_model, data)
baseline_pred <- get_pred_int(baseline_model, data)

# Add predictions and intervals to the new_data
new_data <- new_data %>%
  mutate(
    best_fit = best_pred$fit,
    best_upr = best_pred$upr,
    best_lwr = best_pred$lwr,
    baseline_fit = baseline_pred$fit,
    baseline_upr = baseline_pred$upr,
    baseline_lwr = baseline_pred$lwr
  )

# Create the plot
ggplot(new_data, aes(x = log(SQFT))) +
  # Add actual data points
  geom_point(aes(y = log(ELCNS)), alpha = 0.2, colour = "black") +
  # Add line and ribbon for best model
  geom_line(aes(y = best_fit, colour = "Best Model"), size = 1.2) +
  geom_ribbon(aes(ymin = best_lwr, ymax = best_upr, fill = "Best Model"),
alpha = 0.2) +
  # Add line and ribbon for baseline model
  geom_line(aes(y = baseline_fit, colour = "Baseline Model"), size = 1) +
  geom_ribbon(aes(ymin = baseline_lwr, ymax = baseline_upr, fill = "Baseline
Model"), alpha = 0.2) +
  # Set labels and title
  labs(title = "Model Performance Comparison with Standard Errors",
    x = "log Square Footage (SQFT)",
    y = "log Electricity Consumption (ELCNS)") +
  # Customize theme and colors
  scale_color_manual(values = c("Best Model" = "cadetblue", "Baseline Model" = "coral"), name = "Fitted") +
  scale_fill_manual(values = c("Best Model" = "cadetblue", "Baseline Model" = "coral"), name = "Fitted") +
  theme_minimal() +
  theme(legend.position = "bottom")

```

```
# Save the plot (optional)
# ggsave("model_performance_comparison_with_se.png", width = 10, height =
6, dpi = 300)
```