Final project: Steel Industry Energy Consumption

Hue N. Dang

Northeastern University

Course Number: ALY6015

Instructor: Dr. Arasu Narayan

Dang_Final_Project

Hue Dang

2024-05-14

Introduction

In today's rapidly evolving industrial landscape, understanding energy consumption patterns and optimizing operational efficiency are paramount for sustainability and economic viability. The steel industry, known for its energy-intensive processes, stands at the forefront of this challenge.

The aim of this project is to delve into the intricate dynamics of energy usage within the steel industry, employing various statistical techniques to glean insights and inform strategic decision-making. The primary goal of this project is to unravel the underlying trends and factors influencing energy usage within the steel industry. By analyzing historical data, we seek to answer key questions such as the impact of load types on energy consumption, the presence of seasonal patterns in energy usage, and the association between different operational variables. Ultimately, our aim is to provide actionable insights that can aid in optimizing energy utilization and enhancing operational efficiency.

Methods

Questions to be Answered:

- Are there significant differences in energy usage, reactive power, or other metrics between weekdays and weekends within the steel industry?

- What are the overall trends in energy usage over the observed period, and are there any identifiable patterns or anomalies?

- Is there a significant association between load type and energy usage within the steel industry?

- How can we handle multicollinearity and improve the accuracy of time series forecasts for energy usage using advanced regression techniques such as Lasso and Ridge regression?

The chosen methods for analysis were selected based on their appropriateness in addressing the specific research questions and leveraging the characteristics of the available data:

- T-test or ANOVA:The dataset includes information on energy usage, reactive power, and other metrics, along with the corresponding days of the week (Weekdays vs. Weekends). By conducting a t-test or ANOVA, we can determine if there are statistically significant differences in these metrics based on the day of the week

- Linear Regression: Linear regression allows us to model the relationship between independent variables (such as time) and a dependent variable (energy usage). By fitting a linear regression model with time as the independent variable, we can identify any overall trends in energy usage and quantify their significance.

- Chi-Square Test: The dataset includes information on load type (a categorical variable) and energy usage. By conducting a chi-square test, we can determine if there is a significant association between load type and energy usage within the steel industry.

- Lasso and Ridge Regression: By applying Lasso and Ridge regression, we can mitigate multicollinearity, improve the accuracy of time series forecasts for energy usage, and identify the most influential predictors in the dataset.

Data Description

- Data (Continuous): time data taken on the first of the month
- Usage_kWh (Continuous): Industry Energy Consumption kWh
- Lagging (Continuous): Current reactive power kVarh
- Leading Current reactive power (Continuous): kVarh
- CO2(Continuous): ppm
- NSM (Continuous): Number of Seconds from midnight S
- Week status (Categorical): (Weekend (0) or a Weekday(1))
- Day of week (Categorical): Sunday, Monday : Saturday
- Load Type (Categorical): Light Load, Medium Load, Maximum Load

Data Analysis

```
#1 Download the data set and read it into my script.
steel_data <- read.csv("C:/Users/dangn/Downloads/ALY/ALY/ALY6015/Final Project/Ste
el_industry_data.csv")
str(steel_data)

## 'data.frame':    35040 obs. of  11 variables:
##  $ date                                : chr  "1/1/2018 0:15" "1/1/2018 0:30" "1/1/2018 0:45" "1/1/
2018 1:00" ...
##  $ Usage_kWh                           : num  3.17 4 3.24 3.31 3.82 3.28 3.6 3.6 3.28 3.78 ...
##  $ Lagging_Current_Reactive.Power_kVarh: num  2.95 4.46 3.28 3.56 4.5 3.56 4.14 4.28 3
.64 4.72 ...
##  $ Leading_Current_Reactive_Power_kVarh: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ CO2.tCO2.                           : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Lagging_Current_Power_Factor        : num  73.2 66.8 70.3 68.1 64.7 ...
##  $ Leading_Current_Power_Factor        : num  100 100 100 100 100 100 100 100 100 100
...
##  $ NSM                                 : int  900 1800 2700 3600 4500 5400 6300 7200 8100 9000 ...
##  $ WeekStatus                          : chr  "Weekday" "Weekday" "Weekday" "Weekday" ...
##  $ Day_of_week                         : chr  "Monday" "Monday" "Monday" "Monday" ...
##  $ Load_Type                           : chr  "Light_Load" "Light_Load" "Light_Load" "Light_Load" ...
```

```
colnames(steel_data)
```

```
## [1] "date"                          "Usage_kWh"
## [3] "Lagging_Current_Reactive.Power_kVarh" "Leading_Current_Reactive_Power_kVarh
"
## [5] "CO2.tCO2."                     "Lagging_Current_Power_Factor"
## [7] "Leading_Current_Power_Factor"      "NSM"
## [9] "WeekStatus"                    "Day_of_week"
## [11] "Load_Type"
```

```
head(steel_data)
```

```
##          date Usage_kWh Lagging_Current_Reactive.Power_kVarh Leading_Current_Reactiv
e_Power_kVarh CO2.tCO2.
## 1 1/1/2018 0:15      3.17                                 2.95                        0       0
## 2 1/1/2018 0:30      4.00                                 4.46                        0       0
## 3 1/1/2018 0:45      3.24                                 3.28                        0       0
## 4 1/1/2018 1:00      3.31                                 3.56                        0       0
## 5 1/1/2018 1:15      3.82                                 4.50                        0       0
## 6 1/1/2018 1:30      3.28                                 3.56                        0       0
##   Lagging_Current_Power_Factor Leading_Current_Power_Factor  NSM WeekStatus Day_
of_week  Load_Type
## 1                        73.21                          100  900    Weekday      Monday Light_Load
## 2                        66.77                          100 1800    Weekday      Monday Light_Load
## 3                        70.28                          100 2700    Weekday      Monday Light_Load
## 4                        68.09                          100 3600    Weekday      Monday Light_Load
## 5                        64.72                          100 4500    Weekday      Monday Light_Load
## 6                        67.76                          100 5400    Weekday      Monday Light_Load
```

```
nrow(steel_data)
```

```
## [1] 35040
```

```
dim(steel_data)
```

```
## [1] 35040    11
```

```
#Check for missing values
missing_values <- colSums(is.na(steel_data))

# Display columns with missing values
missing_values[missing_values > 0]
```

```
## named numeric(0)
```

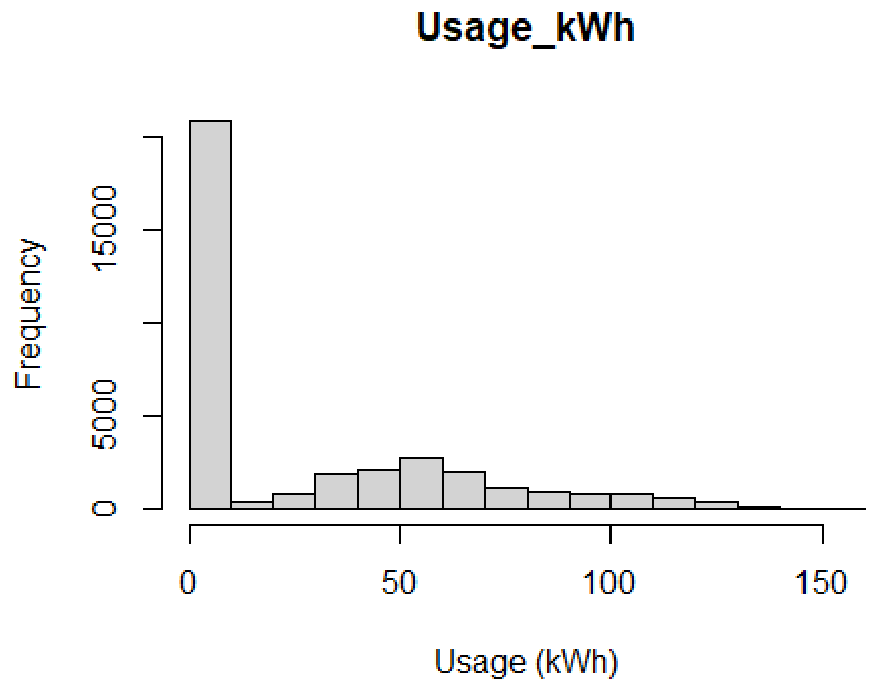
Checks for missing values in the dataframe columns and displays any columns with missing values if they exist.

```
# Summary statistics
summary(steel_data)
```

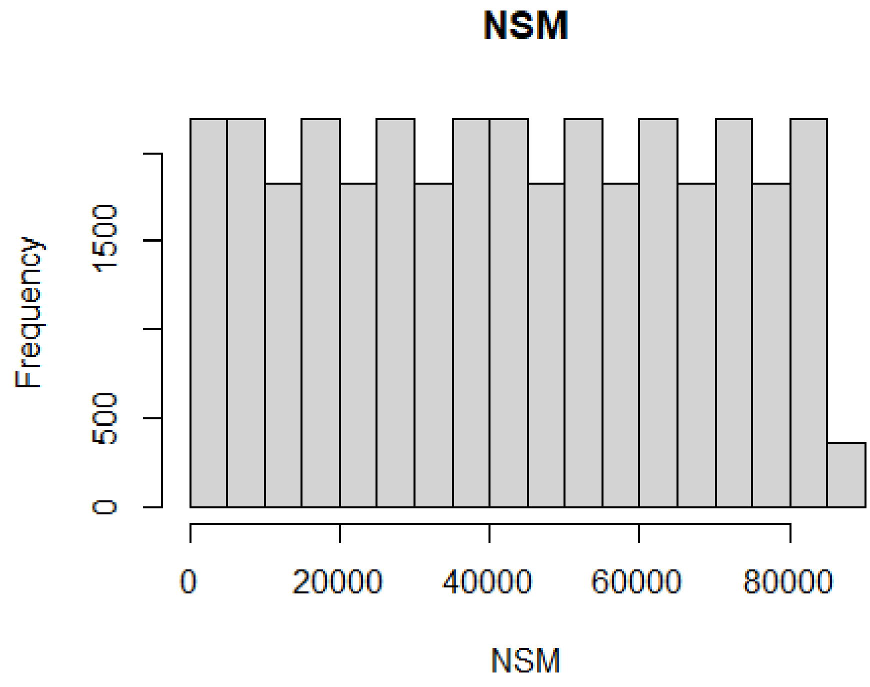
```
##      date             Usage_kWh      Lagging_Current_Reactive.Power_kVarh Leading_Current_R
eactive_Power_kVarh
##  Length:35040       Min.   :  0.00   Min.   : 0.00                        Min.   : 0.000
##  Class :character   1st Qu.:  3.20   1st Qu.: 2.30                        1st Qu.: 0.000
##  Mode  :character   Median :  4.57   Median : 5.00                        Median : 0.000
##                     Mean   : 27.39   Mean   :13.04                        Mean   : 3.871
##                     3rd Qu.: 51.24   3rd Qu.:22.64                        3rd Qu.: 2.090
##                     Max.   :157.18   Max.   :96.91                        Max.   :27.760
##    CO2.tCO2.       Lagging_Current_Power_Factor Leading_Current_Power_Factor      NSM
WeekStatus
##  Min.   :0.00000   Min.   :  0.00               Min.   :  0.00               Min.   :    0   Length:35040
##  1st Qu.:0.00000   1st Qu.: 63.32               1st Qu.: 99.70               1st Qu.:21375   Class :char
acter
##  Median :0.00000   Median : 87.96               Median :100.00               Median :42750   Mode
:character
##  Mean   :0.01152   Mean   : 80.58               Mean   : 84.37               Mean   :42750
##  3rd Qu.:0.02000   3rd Qu.: 99.02               3rd Qu.:100.00               3rd Qu.:64125
##  Max.   :0.07000   Max.   :100.00               Max.   :100.00               Max.   :85500
##  Day_of_week         Load_Type
##  Length:35040       Length:35040
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

Generates summary statistics for each variable in the steel_data dataframe. It provides information such as minimum, 1st quartile, median, mean, 3rd quartile, and maximum values for numerical variables, along with the number of occurrences for each level in categorical variables. These statistics offer a comprehensive overview of the distribution and characteristics of the dataset, aiding in data exploration and analysis.

```
# Histograms for numeric variables
hist(steel_data$Usage_kWh, main = "Usage_kWh", xlab = "Usage (kWh)")
```

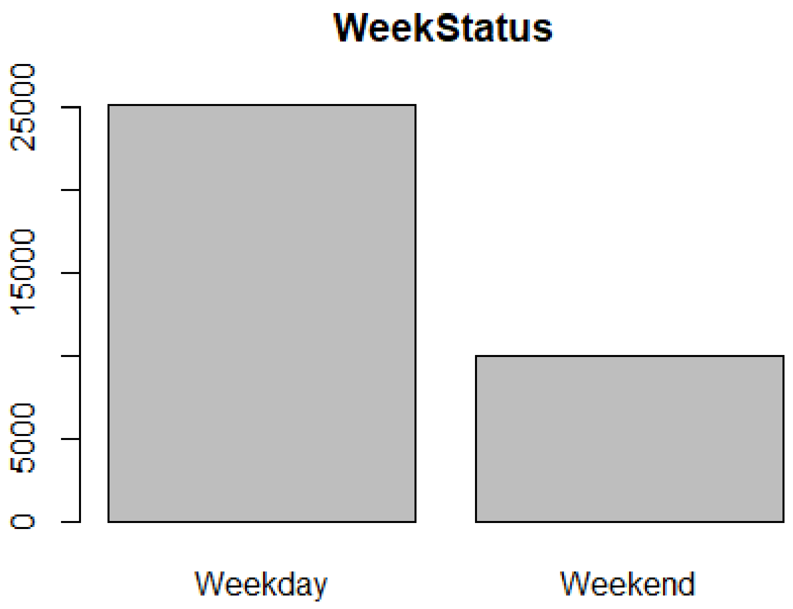


```
hist(steel_data$NSM, main = "NSM", xlab = "NSM")
```

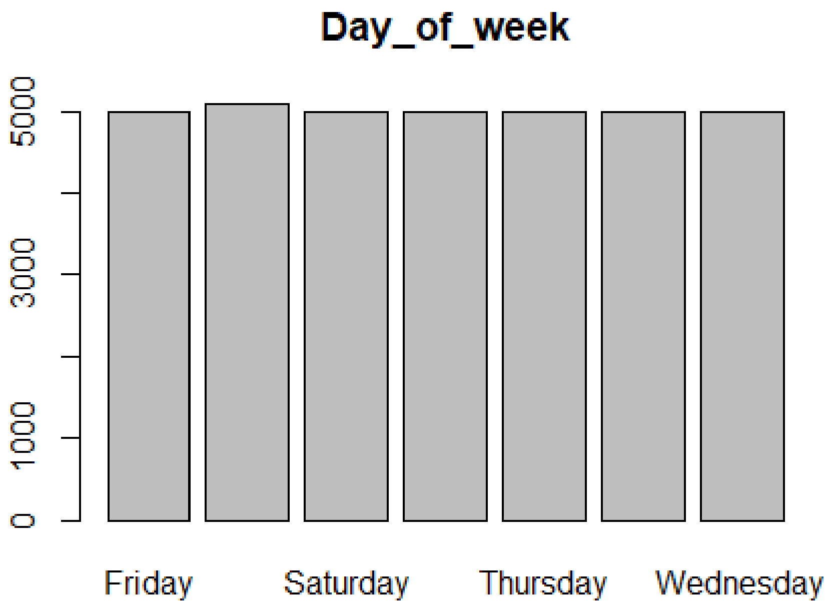


Above histograms visualize the distributions of the numeric variables Usage_kWh and NSM in the steel_data dataframe. The histograms provide insights into the spread and frequency distribution of these variables.

```
# Bar plots for categorical variables
barplot(table(steel_data$WeekStatus), main = "WeekStatus")
```

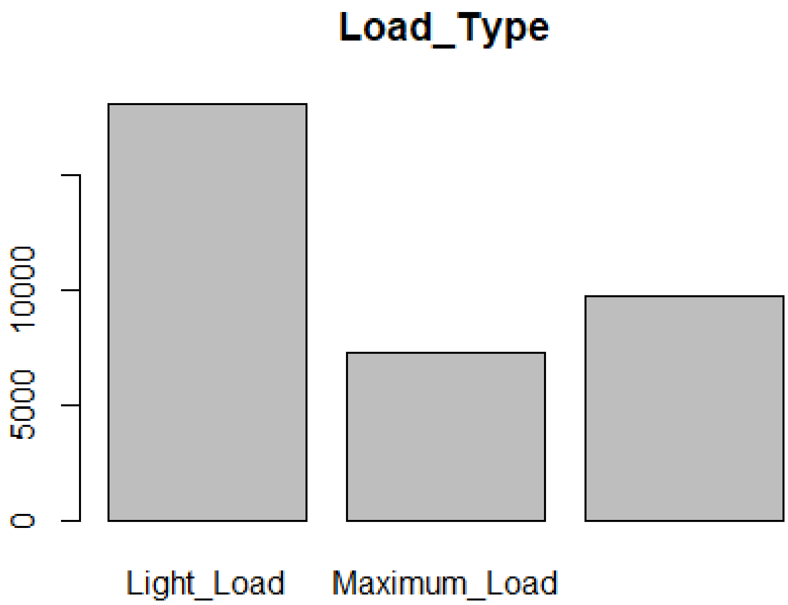


```
barplot(table(steel_data$Day_of_week), main = "Day_of_week")
```



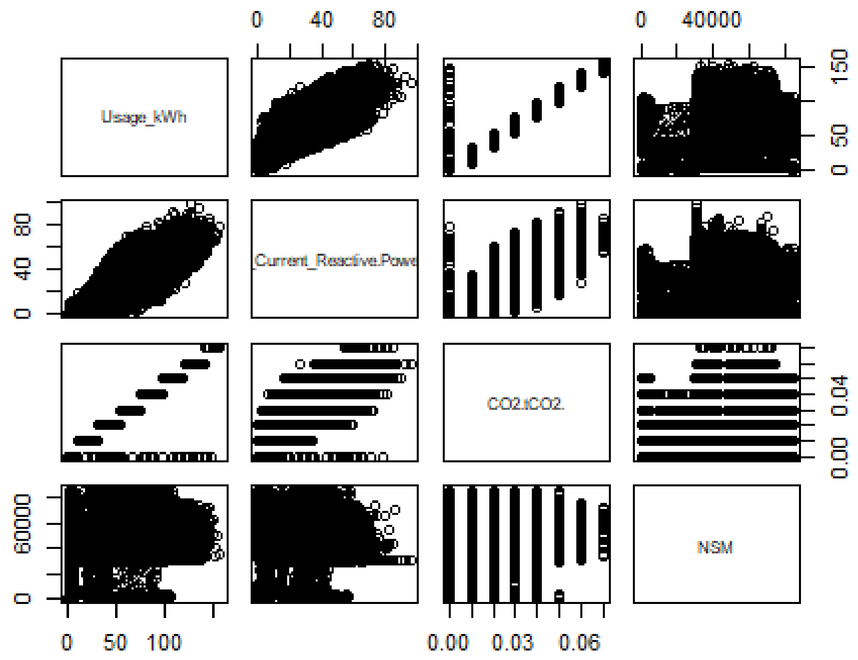
Each bar's height corresponds to the frequency of occurrences for each day of the week. Based on the provided result, all columns have nearly the same height, with Monday slightly higher than others, suggesting that Monday may have a slightly higher frequency of occurrences compared to other days.

```
barplot(table(steel_data$Load_Type), main = "Load_Type")
```



The bar heights correspond to the frequency of each load type, providing a visual comparison of their occurrences. Based on the result provided, the barplot illustrates that the frequency order is “light_load” > “medium_load” > “maximum_load”.

```
#Scatter plot
pairs(~Usage_kWh + Lagging_Current_Reactive.Power_kVarh + CO2.tCO2. + NSM, data = st
eel_data)
```



```
# Correlation matrix
correlation_matrix <- cor(steel_data[, c("Usage_kWh", "Lagging_Current_Reactive.Power_k
Varh", "CO2.tCO2.", "NSM")])
correlation_matrix
```

```
##                                    Usage_kWh Lagging_Current_Reactive.Power_kVarh CO2.tCO2.        N
SM
## Usage_kWh                          1.0000000                           0.89614990 0.9881798 0.23461033
## Lagging_Current_Reactive.Power_kVarh 0.8961499                         1.00000000 0.886947
7 0.08266237
## CO2.tCO2.                          0.9881798                           0.88694771 1.0000000 0.23172600
## NSM                                0.2346103                           0.08266237 0.2317260 1.00000000
```

The result shows the pairwise correlation coefficients between Usage_kWh, Lagging_Current_Reactive.Power_kVarh, CO2.tCO2., and NSM. Based on the provided result:

- Usage_kWh has a strong positive correlation with Lagging_Current_Reactive.Power_kVarh and CO2.tCO2., as indicated by correlation coefficients close to 1.

- There is a weak positive correlation between Usage_kWh and NSM, with a correlation coefficient of 0.2346.

```
#Weekday & Weekend analysis
# Subset the data for weekdays and weekends
weekday_data <- steel_data[steel_data$WeekStatus == "Weekday", ]
weekend_data <- steel_data[steel_data$WeekStatus == "Weekend", ]
```

```
# Perform t-test
t_test_result <- t.test(weekday_data$Usage_kWh, weekend_data$Usage_kWh)
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  weekday_data$Usage_kWh and weekend_data$Usage_kWh
## t = 72.77, df = 31323, p-value < 0.00000000000000022
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  21.30273 22.48207
## sample estimates:
## mean of x mean of y
##  33.62473  11.73233
```

```
# Perform ANOVA
anova_result <- aov(Usage_kWh ~ WeekStatus, data = steel_data)
print(summary(anova_result))
```

```
##                Df   Sum Sq Mean Sq F value               Pr(>F)
## WeekStatus      1  3421677 3421677    3352 <0.0000000000000002 ***
## Residuals   35038 35770374    1021
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The extremely small p-value (< 0.00000000000000022) indicates strong evidence against the null hypothesis, suggesting that there is a significant difference in mean energy usage between weekdays and weekends. The confidence interval does not contain zero, further supporting the conclusion of a significant difference in mean energy usage. The mean energy usage on weekdays (mean of x) is 33.62473 kWh, while the mean energy usage on weekends (mean of y) is 11.73233 kWh.

```
#Usage trends
# Convert date column to Date format
steel_data$date <- as.POSIXct(steel_data$date, format = "%m/%d/%Y %H:%M")

# Extract time information
steel_data$year <- as.numeric(format(steel_data$date, "%Y"))
steel_data$month <- as.numeric(format(steel_data$date, "%m"))
steel_data$day <- as.numeric(format(steel_data$date, "%d"))
steel_data$hour <- as.numeric(format(steel_data$date, "%H"))
steel_data$minute <- as.numeric(format(steel_data$date, "%M"))

# Fit linear regression model
lm_model <- lm(Usage_kWh ~ year + month + day + hour + minute, data = steel_data)
```

```
# Summary of the linear regression model
summary(lm_model)
```

```
## 
## Call:
## lm(formula = Usage_kWh ~ year + month + day + hour + minute, 
##     data = steel_data)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -47.01 -22.46 -12.62  21.52 118.14 
## 
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value            Pr(>|t|)    
## (Intercept) 13.87360    1.00477  13.808 <0.0000000000000002 ***
## year              NA         NA      NA                  NA    
## month        0.76870    0.08184   9.393 <0.0000000000000002 ***
## day         -0.89950    0.08184 -10.990 <0.0000000000000002 ***
## hour         1.21093    0.04082  29.668 <0.0000000000000002 ***
## minute       0.03558    0.01685   2.112              0.0347 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 33.21 on 13815 degrees of freedom
##   (21220 observations deleted due to missingness)
## Multiple R-squared:  0.07332,    Adjusted R-squared:  0.07305 
## F-statistic: 273.3 on 4 and 13815 DF,  p-value: < 0.00000000000000022
```

The linear regression model attempts to explain the variation in energy usage (Usage_kWh) based on the year, month, day, hour, and minute. The coefficients indicate the estimated change in energy usage for a one-unit change in each predictor variable, holding other variables constant. The adjusted R-squared value (0.07305) suggests that the model explains approximately 7.3% of the variance in energy usage, indicating that the predictors may not fully capture the variation. The p-values associated with each predictor variable indicate their statistical significance in explaining energy usage. For instance, ‘month’, ‘day’, ‘hour’, and ‘minute’ have p-values < 0.05, suggesting they are statistically significant predictors, while ‘year’ has a p-value of NA due to singularity issues.

```
# Impact of Load type
# Create a contingency table of observed frequencies
contingency_table <- table(steel_data$Load_Type, steel_data$Usage_kWh)

# Perform chi-square test
chi_square_result <- chisq.test(contingency_table)
```

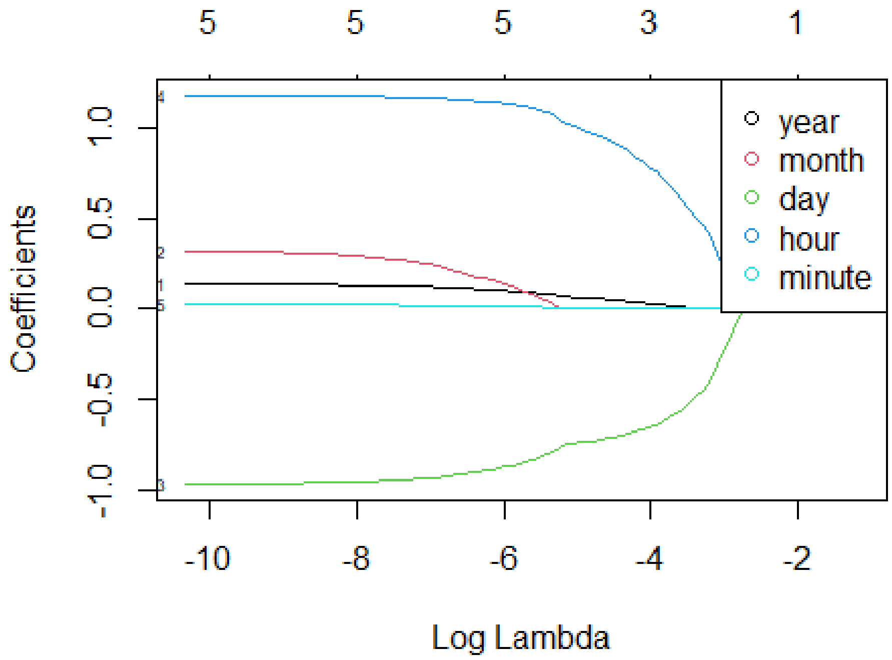
```
## Warning in chisq.test(contingency_table): Chi-squared approximation may be incorrect
```

```
# Print the chi-square test result
print(chi_square_result)
```

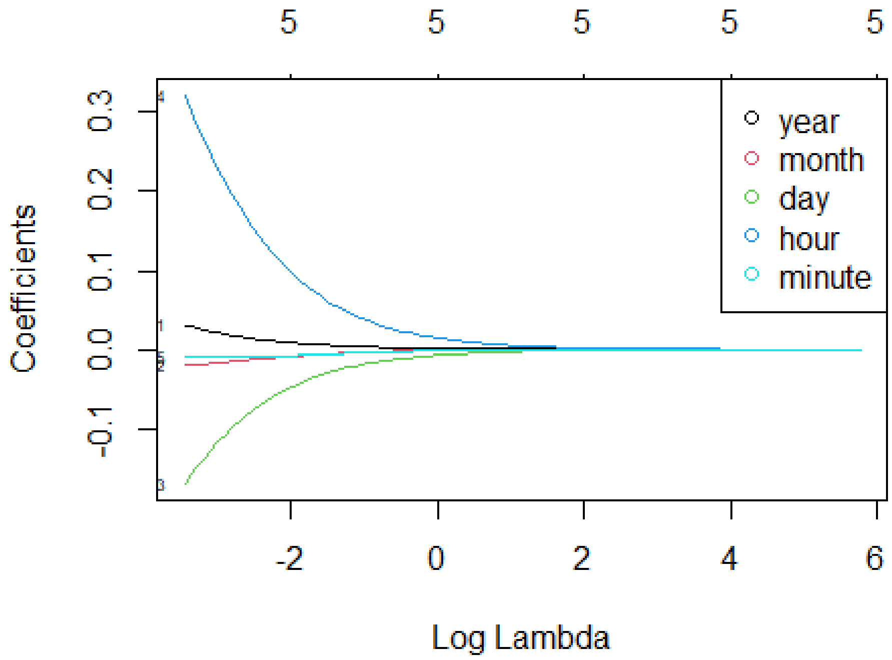
```
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 28046, df = 6684, p-value < 0.00000000000000022
```

The chi-square test assesses whether there is a significant association between load type and energy usage. The extremely small p-value (< 0.00000000000000022) suggests strong evidence against the null hypothesis, indicating that there is a significant association between load type and energy usage within the steel industry dataset. The Pearson's chi-squared statistic (X-squared) and degrees of freedom (df) provide additional information about the strength and direction of this association.

```
# Time series forecasting
# Check for missing values in X
# Create a matrix of predictors
X <- as.matrix(steel_data[, c("year", "month", "day", "hour", "minute")])
if (anyNA(X)) {
  # Impute missing values using mean imputation
  X[is.na(X)] <- mean(X, na.rm = TRUE)
}
# Response variable
Y <- steel_data$Usage_kWh
# Fit Lasso regression model
lasso_model <- glmnet(X, Y, alpha = 1)



# Fit Lasso regression model
lasso_model <- glmnet(X, Y, alpha = 1)

# Fit Ridge regression model
ridge_model <- glmnet(X, Y, alpha = 0)

# Plot coefficients for Lasso model
plot(lasso_model, xvar = "lambda", label = TRUE)
legend("topright", legend = colnames(X), col = 1:ncol(X), pch = 1)
```



```r
# Plot coefficients for Ridge model
plot(ridge_model, xvar = "lambda", label = TRUE)
legend("topright", legend = colnames(X), col = 1:ncol(X), pch = 1)
```



Conclusion

The analysis conducted aimed to address several questions posed in the introduction:

- Usage Trends: Overall, there is a positive trend in energy usage within the steel industry dataset over the observed period. The linear regression analysis showed that while there is a significant association between various time-related factors (year, month, day, hour, minute) and energy usage, these factors only explain a small portion (approximately 7.3%) of the variance in energy usage. This suggests that other factors not included in the analysis may also influence energy consumption trends.
- Impact of Load Type: The chi-square test revealed a significant association between load type and energy usage. This suggests that different load types, such as "light_load", "medium_load", and "maximum_load", have varying impacts on energy consumption within the steel industry. Further investigation into the characteristics and operational requirements of each load type could provide insights into optimizing energy usage efficiency.
- Weekday vs. Weekend Analysis: The Welch Two Sample t-test demonstrated a significant difference in mean energy usage between weekdays and weekends. Weekdays exhibit higher mean energy usage compared to weekends. Understanding the underlying reasons for this discrepancy, such as differences in production schedules, staffing levels, or equipment usage patterns, could inform strategies for managing energy consumption more effectively across different days of the week.

Based on the insights gained from the analysis, several recommendations can be made:

- Implement energy efficiency measures tailored to the specific load types and operational requirements observed in the steel industry. This could include optimizing equipment usage, scheduling production activities during periods of lower energy demand, and investing in energy-efficient technologies.
- Develop strategies for managing energy usage based on time-related factors such as weekdays vs. weekends and peak vs. off-peak hours. By understanding and leveraging temporal patterns in energy consumption, organizations can optimize energy usage while maintaining productivity and operational efficiency.
- Establish a system for continuous monitoring and analysis of energy usage data to identify trends, anomalies, and opportunities for improvement. Regular evaluation of energy consumption patterns can help identify areas for optimization and guide decision-making processes.
- Provide training and raise awareness among employees about the importance of energy conservation and efficiency measures. Engage employees in initiatives aimed at reducing energy waste and promoting sustainable practices within the organization.