Renewable Energy and Weather:

Analyzing The Relationship Between Change in Energy and Weather Conditions

Gloria Xu and Sylvia Ozoede

Non-technical Summary:

Our analysis aimed to understand the complex relationship between weather patterns and renewable energy output. We wanted to see if we could form a good idea of the relationship between the difference in energy consumption in Watt-hours, using variables such as the Global Horizontal Irradiance (the amount of sun hitting the Earth's surface), temperature in degrees Celsius, atmospheric pressure, humidity, wind speed, the amount of precipitation, the amount of snowfall, the amount of clouds in the sky, the sunlight time in minutes, the day length, weather type, hour, and month. The data was collected every fifteen minutes from December 31st, 2016, to August 31st, 2022, with a total of 196, 778 observations. We applied a couple of techniques to see which method was best for analyzing energy consumption, Ridge, Elastic Net, LASSO, and Principal and Confirmatory Factor Analysis.

All three regularized models performed well on unseen data, and the results were very similar across methods. Elastic net produced the lowest prediction error, and LASSO had a nearly identical result, and the highest R squared value. This meant that predicting the change in energy consumed using the weather variables was feasible even though the differences between the best models were very small.

Using principal factor analysis revealed that weather doesn't just act as individual variables; 5 components were identified as influencers of the environment surrounding our energy systems. The Daily Cycle, which is the most predictable factor, combining time of day, humidity levels, and total daylight. Storm Activity, where cloud cover, rainfall, and air pressure group together to create "storm profiles" that impact solar efficiency. Solar Intensity combines temperature and GHI to measure actual energy potential. Seasonal

Shifts, where the month and snowfall create a distinct “winter” signature. Wind Dynamics, where wind speed and air pressure form a separate category that measures wind-based assets independently of solar factors.

Solar energy is highly correlated with “Daily Cycle” (time and humidity), making it more predictable than wind, which operates on its own, independent of the “Wind Dynamics” factor. When the “Storm Activity” factor is high, solar efficiency drops more sharply than when it is simply evening time. Because wind and solar are driven by different underlying factors, they provide a natural balance. When the “Solar Intensity" factor is low, the “Wind Dynamics" factor often remains unaffected, ensuring a more stable total energy output. Weather is messy and full of “noise”, but it follows patterns that are mappable. By understanding these factors , we can better predict energy dips and build a more stable renewable energy grid.

Technical Summary:

After looking at the data, we noticed that many of the observations had delta energy at 0 and GHI at 0, these would occur mainly when the sun is not out, indicated by the category isSun being equal to 0. It would make more sense for us to use delta energy as the response variable when the sun is out, therefore, we took the data and only looked at when isSun is equal to 1. This allowed us to filter out most of the delta energy equals 0 and GHI equals 0.

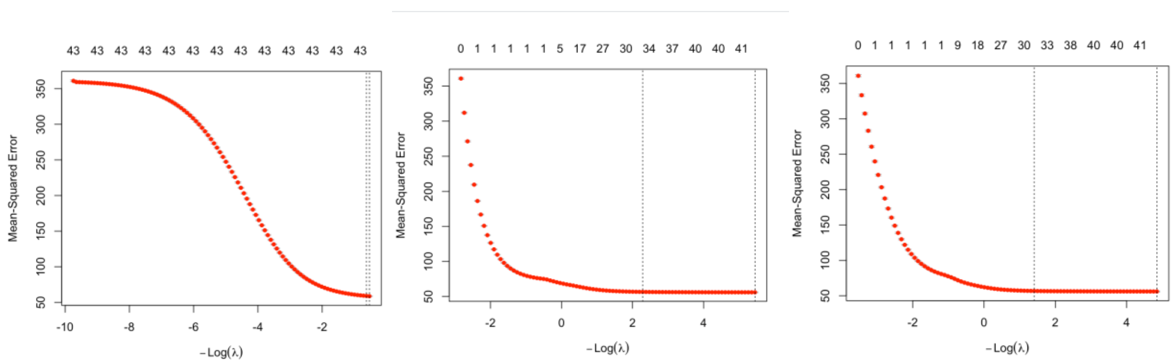
The filtered dataset was then cleaned to remove non-numeric columns. To address the skewness of some of the variables and stabilize variance, a couple of transformations were applied. Log transformation was applied to the wind_speed variable and square root transformations were applied to Energy.deltha.Wh and GHI.

A correlation analysis using the "ellipse" method indicated significant multicollinearity among solar-related variables (GHI, temp, and sunlight duration) and an inverse relationship between humidity and temperature. The correlation plot also showed that the response variable had a strong positive relationship with GHI, indicating that higher GHI is associated with large positive changes in energy consumption. The inverse relationship between humidity and change in energy consumption suggests that less favorable weather conditions reduce energy generation. The presence of multicollinearity made ordinary least square regression (OLS) less attractive as a route of analysis because of the overlap among predictors, therefore we selected regularized regression to understand how delta energy can be predicted, and PCA to understand if there are underlying relationships between variables.

Analysis 1 - Regularized Regression:

Ridge, LASSO, and Elastic net were appropriate because regularization helps stabilize estimates when multicollinearity is present. The response variable was the transformed change in energy consumption, and the predictors, transformed GHI, transformed wind speed, temperature, pressure, humidity, rainfall, snowfall, cloud cover, weather type, hour, month etc. Dummy variables were created for the categorical inputs. The data was split into 70% training and 30% testing sets, and cross validation was used within the training set to select the optimal penalty parameter, lambda λ.

Ridge keeps all predictors in the model but shrinks their coefficients. LASSO can shrink some coefficients to exactly zero, making it useful for variable selection. Elastic Net combines both penalties. Cross-validation plots showed that LASSO and Elastic Net reached very similar error levels, while Ridge also performed very well.

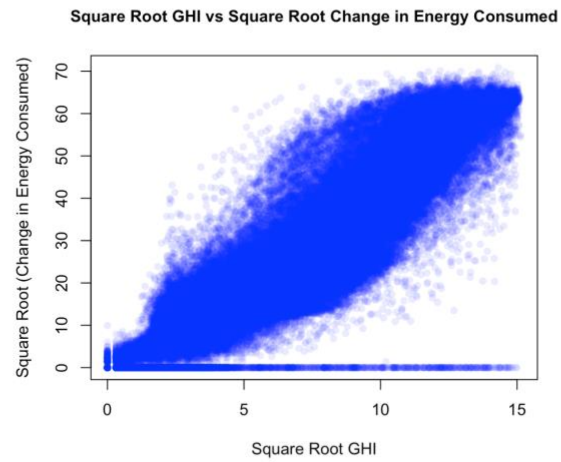


Ridge LASSO Elastic Net

Model	RMSE	R^2
Ridge	7.513691	0.8432622
Lasso	7.340141	0.854192
Elastic Net	7.339853	0.8504310

These results show that all three methods performed similarly, with Elastic Net having the smallest RMSE and LASSO having the largest R squared. The difference between LASSO and Elastic Net was practically negligible.

The predictive signal in the dataset is fairly strong and stable since the regularized methods all captured that signal well suggesting that the relationship between weather conditions and change in energy consumption is real. Elastic net is slightly preferable for the goal of purely predicting since it had the lowest test error.

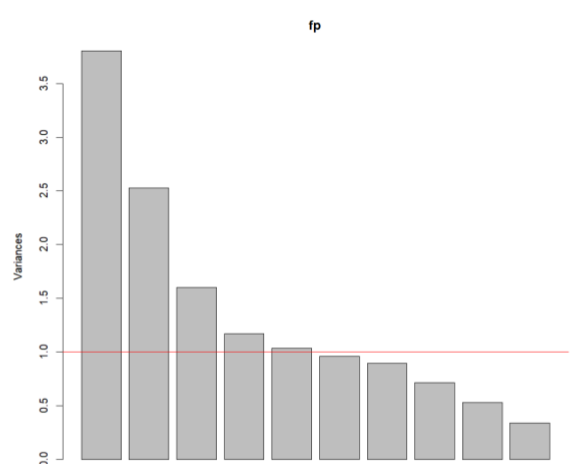


The analyses show that change in energy consumption is strongly associated with GHI, while also responding to humidity, clouds, and related weather conditions. The regression results indicate that the response variable can be predicted well, especially once transformations are applied and multicollinearity is handled appropriately. The factor-based analysis compliments this by showing that many of these predictors naturally group together.

These findings suggest that change in energy consumption forecasting can benefit from incorporating multiple weather measurements at once, particularly solar and atmospheric conditions. The analysis shows why simple models may struggle due to the predictors overlapping and having their effects intertwined. This makes regularized modeling especially useful because it balances prediction accuracy with stability in the correlated inputs.

Analysis 2 – PCA, EFA, CFA:

The first of the three distinct multivariate methodologies was Principal Component Analysis. PCA was used as our primary variable reduction technique. By analyzing the variance-covariance matrix of the scaled data, we identified the underlying dimensions of the dataset. To select the number of factors, we used the Kaiser Criterion and the Scree Plot test. 5 components



were chosen based on both methods, which explains about 73% of the total variance.

To interpret the factors found from PCA analysis, we used Exploratory Factor Analysis with Promax Rotation. We chose an oblique rotation because weather variables (like pressure and wind) are naturally correlated; forcing them to be perfectly independent would have resulted in a less realistic model Using a loading cutoff of 0.4, this revealed 5 latent factors: Daily Cycle, Storm Activity, Solar Intensity, Seasonal Shifts, and Wind Dynamics. We then tested the 5-component theory using Confirmatory Factor Analysis. We used the lavaan package to validate the EFA results on a separate test set, creating a structural model to see if the proposed factors did well with new data.

We used a manual-hybrid selection method, starting with PCA to identify the strongest factors and refining them into a CFA model. The latent variables are defined here:

```
fmodel <- '
  TimeOfDay =~ SunlightTime.daylength + sunlightTime + hour + humidity
  Storm =~ rain_1h + clouds_all + weather_type
  Solar =~ GHI + temp + humidity + dayLength
  Seasonality =~ month + snow_1h
  WindDynamics =~ pressure + wind_speed'
```

To ensure the model wasn't just "overfitting" the training data, we implemented a 70/30 training/test split. The CFA was run on the training data and then validated against the test data. While the factor loadings are high with many values over 0.90, the high RMSEA of 0.285 statistically suggests that the model should be rejected as a perfect fit. It indicates that while the five factors are strong, however, there is "unexplained noise" in weather data (likely due to the extreme values of rainfall and snowfall) that a linear factor model cannot fully explain.

Statistically, the separation of Wind Dynamics and Solar Intensity is the most important finding. It confirms that the underlying atmospheric drivers for wind and solar are not collinear; this justifies the expansion of hybrid wind-solar farms to minimize energy volatility. Variables like SunlightTime.daylength and hour, and GHI and temp are so highly correlated that they provide redundant information. In the future, to reduce costs, we can prioritize one variable over the other. While we did not use LDA, our factor loadings for weather_type and clouds_all effectively group samples into “Stable” vs “Volatile” environmental categories, which can be used to trigger energy storage reserves.

Conclusion:

Through this two-phased analysis, we have demonstrated that the relationship between weather conditions and changes in energy consumption is both statistically significant and structurally complex. By combining predictive modeling with latent factor discovery, we can view weather beyond a collection of isolated variables and instead identify it as a series of integrated “profiles” that influence energy output.

Our first analysis established that energy change is highly predictable. By utilizing Ridge, LASSO, and Elastic Net Regression, we successfully navigated the high degree of multicollinearity in the atmospheric data. The superior performance of the Elastic Net model confirms that while GHI is the primary driver of energy change, its predictive power is significantly enhanced when balanced against a suite of correlated secondary variables like humidity, cloud cover, and temperature.

The second analysis provided the reasoning behind these predictive results. By identifying the five latent factors (Daily Cycle, Storm Activity, Solar Intensity, Seasonal Shifts, and Wind Dynamics), we revealed the underlying architecture of the environment. The clear statistical separation between “Wind Dynamics” and “Solar Intensity” provides our most practical finding; it confirms that wind and solar assets are driven by distinct atmospheric pressures. The lack of collinearity between wind and solar drivers justifies the implementation of hybrid energy farms to mitigate volatility and ensure a stable total energy supply.

While high RMSEA in our CFA suggests that extreme weather events introduce noise that linear models cannot fully capture, the overall alignment between our predictive and structural analysis is clear. We conclude that effective energy forecasting requires a holistic approach. Regularized models provide the necessary precision for day-to-day operations, while factor-based analysis offers a roadmap for grid stability and resource optimization. In the future, energy systems can be made more resilient by prioritizing these environmental “signatures” over redundant individual metrics.

Appendix A: Gloria Xu

Overall, I was responsible for the initial data filtering and implementation of the transformations. I performed the PCA, EFA, and CFA and generated the scree plots to justify the 5-factor approach. I specifically chose the Promax rotation and interpreted the loading matrix, which allowed me to name the factors. I also conducted the training/test split to ensure our CFA results weren't just the result of overfitting the general data.

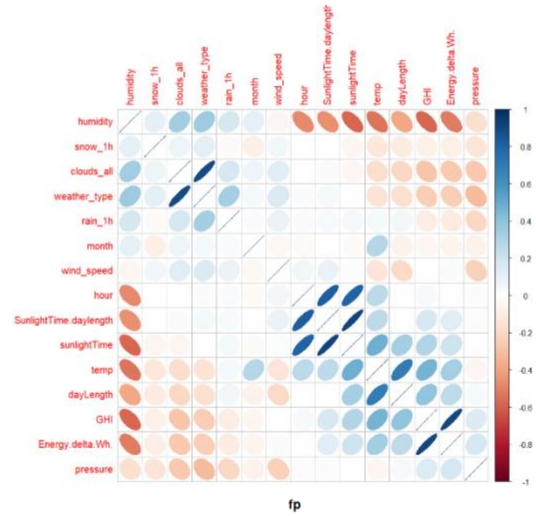
Throughout the project, I tried many different variations of the data, where the data that was kept was the full hour when isSun equaled 1 and creating new variables that categorized the hour data. However, I realized that keeping the full hour would mean more 0 values, so I decided to just use the data when isSun equals 0. The new variable I created categorized the hours to times of the day, such as dawn, early morning, noon, and dusk. After running the corrplot with the new variable, I saw that it has the same exact relationship as the other variables as SunlightTime/daylength variable, so I decided the new category did not bring anything new to the analysis.

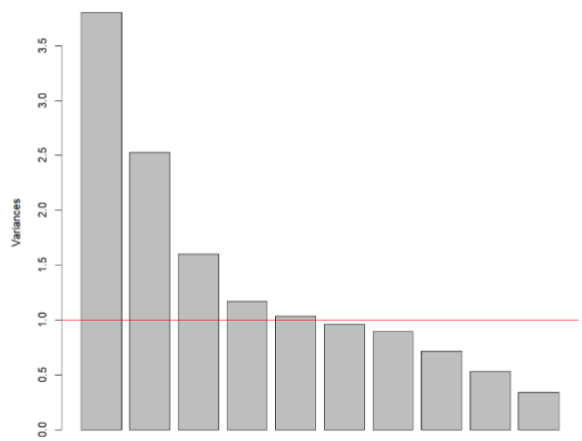
I was proactive about getting our work done in a timely manner and tried to communicate with my group member to ensure that we could meet our deadlines. I reached out to them often and offered help when I could, even with prior obligations.

The most significant lesson I learned was the reality of RMSEA. In a perfect world, we look for values that are less than 0.05, but in high-frequency weather data, getting a "perfect" fit is almost impossible. I realized that a model can be practically useful for categorization even if it doesn't meet the statistical thresholds for a perfect fit.

Below I have included the R-code that I used as well as some outputs that I used in my analysis:

```r
# #final data
ndata <- read.csv("renewable_filtered2.csv")
fdata <- ndata[,-c(1, 11, 18, 19)]
head(fdata)

fdata$wind_speed <- log(fdata$wind_speed + 1)
fdata$Energy.delta.Wh. <- sqrt(fdata$Energy.delta.Wh. + 1)
fdata$GHI <- sqrt(fdata$GHI + 1)
hist.data.frame(fdata)

fcor <- cor(fdata)
corrplot(fcor, method = "ellipse", order = "AOE")

#PCA on whole data
fp = prcomp(fdata[, -c(1)], scale = T)
plot(fp)
abline(h = 1, col = "red")
summary(fp)

fprinds <- principal(fdata[, -c(1)], nfactors = 5, rotate = "promax")
print(fprinds$loadings, cutoff = 0.4, sort = T)
summary(fprinds)

print(fprinds$STATISTIC)
RMSEA(fprinds$STATISTIC, 31,108528, alpha = 0.1)
#RMSEA = 0.285, which indicates that the model should be rejected

#split training and test set

fdata <- as.matrix(fdata)

set.seed(365)

training <- sample(nrow(fdata), nrow(fdata) * 0.7)
traindata <- fdata[training, ]
testdata <- fdata[-training, ]

#Cross-validation with PCA and EFA

fpt = prcomp(traindata[, -c(1)], scale = T)
plot(fpt)
abline(h = 1, col = "red")
summary(fpt)

ftprinds <- principal(traindata[, -c(1)], nfactors = 5, rotate = "promax")
print(ftprinds$loadings, cutoff = 0.4, sort = T)
summary(ftprinds)

#CFA
fmodel <- '
  TimeOfDay =~ SunlightTime.daylength + sunlightTime + hour + humidity
  Storm =~ rain_1h + clouds_all + weather_type
  Solar =~ GHI + temp + humidity + dayLength
  Seasonality =~ month + snow_1h
  WindDynamics =~ pressure + wind_speed'

trainfit <- cfa(model = fmodel, data = traindata)
fitmeasures(trainfit)
testfit <- cfa(model = fmodel, data = testdata)
fitmeasures(testfit)
```






```
Loadings:
                       RC1    RC3    RC2    RC5    RC4
sunlightTime           0.886
SunlightTime.daylength 0.999
hour                   0.986
GHI                           0.649
temp                          0.837
dayLength                     0.959
rain_1h                              0.557
clouds_all                           0.891
weather_type                         0.957
pressure                                    -0.583
wind_speed                                   0.878
month                                               0.936
humidity               -0.420 -0.418
snow_1h                                             -0.403

                 RC1   RC3   RC2   RC5   RC4
SS loadings    3.011 2.431 2.415 1.261 1.192
Proportion Var 0.215 0.174 0.172 0.090 0.085
Cumulative Var 0.215 0.389 0.561 0.651 0.736
```

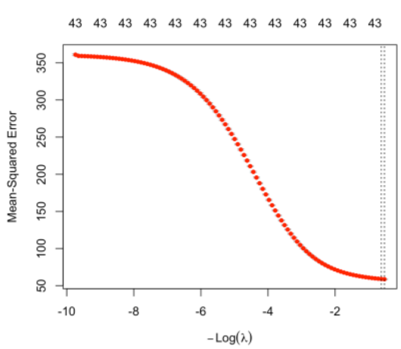
```
Factor analysis with Call: principal(r = traindata[, -c(1)], nfactors = 5, rotate = "promax")

Test of the hypothesis that 5 factors are sufficient.
The degrees of freedom for the model is 31  and the objective function was  2.51
The number of observations was  75969  with Chi Square =  190635  with prob <  0

The root mean square of the residuals (RMSA) is  0.05

 With component correlations of
      RC1   RC3   RC2   RC5   RC4
RC1  1.00  0.35 -0.15  0.10 -0.03
RC3  0.35  1.00 -0.25  0.03  0.03
RC2 -0.15 -0.25  1.00  0.22  0.16
RC5  0.10  0.03  0.22  1.00 -0.10
RC4 -0.03  0.03  0.16 -0.10  1.00
```

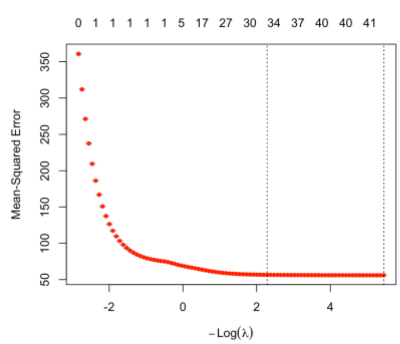
Appendix B: Sylvia Ozoede

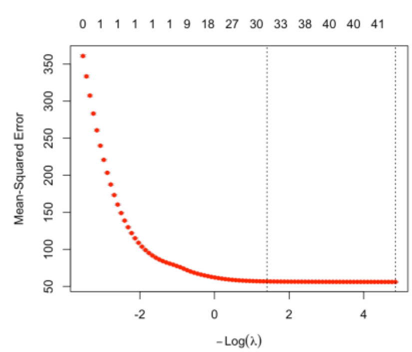
My main contribution to this project was the regression and regularization portion. I explored the relationship between the change in energy consumption variable and the weather predictors. Moreover, both Gloria and I individually created the heatmap correlation and log transformations for our own understanding, so we just used mine were used in the presentation since I made the slides containing them at the time. I went ahead to fit Ridge, LASSO, and Elastic net models, and created the exploratory visuals used to justify the regularized regression. In addition, I presented and recorded the complete slide presentation, including both project sections since my teeammate had work.

From my part, we were able to conclude that GHI was one of the strongest predictors of the response while humidity and cloud-related variables showed weaker or negative relationships.

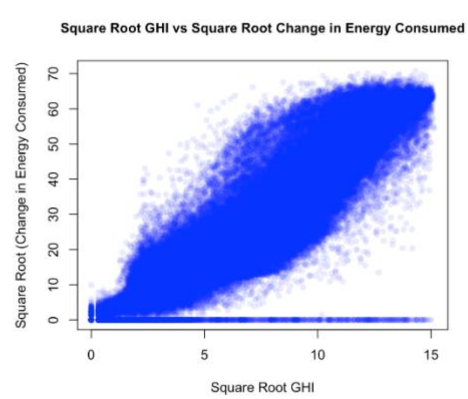
I was able to learn how important exploratory analysis and variable transformations are when you want to work with a dataset. I also learnt that when interpreting regularized regression techniques, we should focus on practical differences rather than only choosing a winner by the numbers.

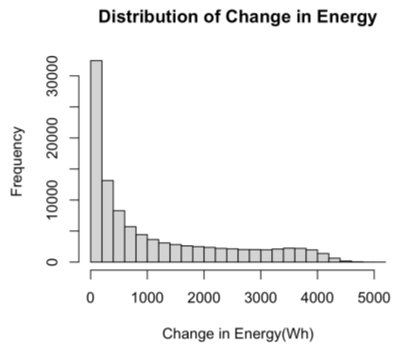


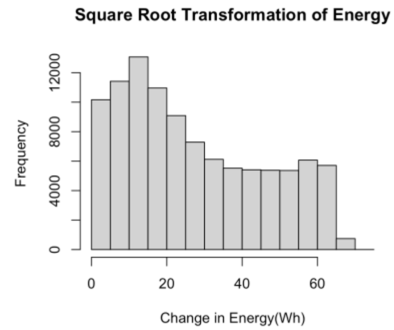


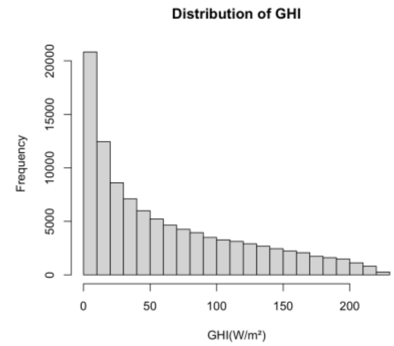


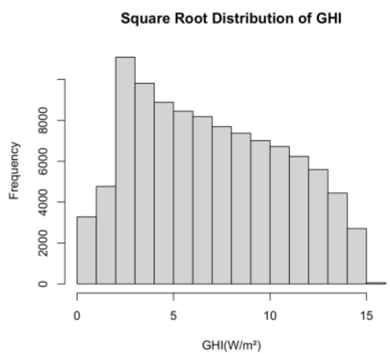
Model	RMSE	R^2
Ridge	7.513691	0.8432622
Lasso	7.340141	0.854192
Elastic Net	7.339853	0.8504310

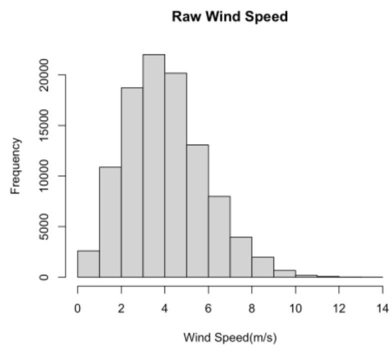


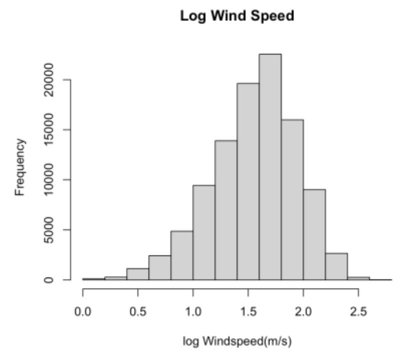

Distribution of Change in Energy
Frequency
0
10000
20000
30000
0
1000
2000
3000
4000
5000
Change in Energy(Wh)

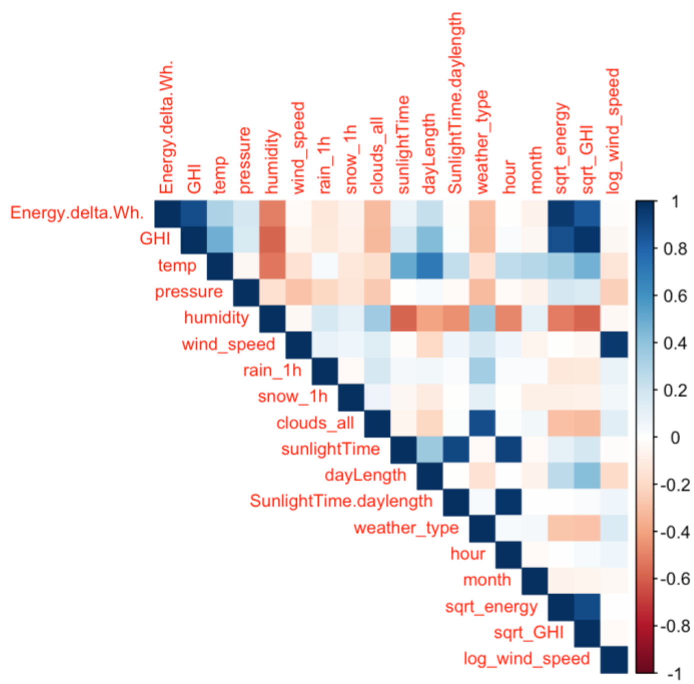

Square Root Transformation of Energy
Frequency
0
4000
8000
12000
0
20
40
60
Change in Energy(Wh)


Distribution of GHI
Frequency
0
5000
10000
15000
20000
0
50
100
150
200
GHI(W/m²)


Square Root Distribution of GHI
Frequency
0
2000
4000
6000
8000
0
5
10
15
GHI(W/m²)


Raw Wind Speed
Frequency
0
5000
10000
15000
20000
0
2
4
6
8
10
12
14
Wind Speed(m/s)


Log Wind Speed
Frequency
0
5000
10000
15000
20000
0.0
0.5
1.0
1.5
2.0
2.5
log Windspeed(m/s)


Energy.delta.Wh.
GHI
temp
pressure
humidity
wind_speed
rain_1h
snow_1h
clouds_all
sunlightTime
dayLength
SunlightTime.daylength
weather_type
hour
month
sqrt_energy
sqrt_GHI
log_wind_speed
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

```r
# Load Libraries
library(ggplot2)
library(GGally)
library(corrplot)
library(psych)
library(glmnet)

# Load Data
renewable <- read.csv("renewable_filtered3.csv")

# Basic inspection
head(renewable)
summary(renewable)
str(renewable)

# Check missing values
colSums(is.na(renewable))

# Transform variables
renewable$sqrt_energy <- sqrt(renewable$Energy.delta.Wh.)
renewable$sqrt_GHI <- sqrt(renewable$GHI)
renewable$log_wind_speed <- log1p(renewable$wind_speed)

# Remove constant numeric columns for correlation analysis
numeric_vars <- renewable[, sapply(renewable, is.numeric)]
numeric_vars <- numeric_vars[, sapply(numeric_vars,
                                      function(x) sd(x, na.rm = TRUE) > 0)]

# Correlation matrix
cor_matrix <- cor(numeric_vars)

corrplot(cor_matrix,
         method = "color",
         type = "upper",
         tl.cex = 0.8,
         number.cex = 0.6)

# Histograms for response and key transformed predictors
hist(renewable$Energy.delta.Wh.,
     main = "Distribution of Change in Energy",
     xlab = "Change in Energy(Wh)")

hist(renewable$sqrt_energy,
     main = "Square Root Transformation of Energy",
     xlab = "Change in Energy(Wh)")

hist(renewable$GHI,
     main = "Transformation of GHI",
     xlab = "GHI(W/m²)")

hist(renewable$sqrt_GHI,
     main = "Square Root Transformation of GHI",
     xlab = "GHI(W/m²)")

hist(renewable$wind_speed,
     main = "Raw Wind Speed",
     xlab = "Wind Speed(m/s)")

hist(renewable$log_wind_speed,
     main = "Log Wind Speed",
     xlab = "log Windspeed(m/s)")

# Scatterplot
plot(renewable$sqrt_GHI,
     renewable$sqrt_energy,
     pch = 16,
     col = rgb(0, 0, 1, 0.1),
     xlab = "Square Root GHI",
     ylab = "Square Root (Change in Energy)",
     main = "Square Root GHI vs Square Root Change in Energy",
     cex.main = 1.0)

plot(renewable$sqrt_GHI,
     renewable$sqrt_energy,
     pch = 16,
     col = rgb(0, 0, 1, 0.1),
     cex = 0.6,
     cex.lab = 1.2,
     cex.main = 1.3,
     xlab = "Square Root GHI",
     ylab = "Square Root Change in Energy (Wh)",
     main = "Square Root GHI vs Square Root Change in Energy (Wh)")

# Train & Test split
set.seed(123)

train_index <- sample(1:nrow(renewable), 0.7 * nrow(renewable))
train_data <- renewable[train_index, ]
test_data  <- renewable[-train_index, ]

# Treat coded variables as factors
train_data$weather_type <- factor(train_data$weather_type)
test_data$weather_type  <- factor(test_data$weather_type,
                                  levels = levels(train_data$weather_type))

train_data$hour <- factor(train_data$hour)
test_data$hour  <- factor(test_data$hour,
                          levels = levels(train_data$hour))

train_data$month <- factor(train_data$month)
test_data$month  <- factor(test_data$month,
                           levels = levels(train_data$month))

# Response = transformed energy delta
y_train <- train_data$sqrt_energy
y_test  <- test_data$sqrt_energy

# Predictors: exclude raw and transformed response,
# keep transformed GHI and transformed wind speed
x_train <- model.matrix(
  sqrt_energy ~ sqrt_GHI + temp + pressure + humidity +
    log_wind_speed + rain_1h + snow_1h + clouds_all +
    sunlightTime + dayLength + SunlightTime.daylength +
    weather_type + hour + month,
  data = train_data)[, -1]

x_test <- model.matrix(
  sqrt_energy ~ sqrt_GHI + temp + pressure + humidity +
    log_wind_speed + rain_1h + snow_1h + clouds_all +
    sunlightTime + dayLength + SunlightTime.daylength +
    weather_type + hour + month,
  data = test_data)[, -1]

# Ridge
ridge_model <- cv.glmnet(x_train, y_train, alpha = 0)
plot(ridge_model)
ridge_model$lambda.min
coef(ridge_model, s = "lambda.min")

# Lasso
lasso_model <- cv.glmnet(x_train, y_train, alpha = 1)
plot(lasso_model)
lasso_model$lambda.min
coef(lasso_model, s = "lambda.min")

# Elastic Net
elastic_model <- cv.glmnet(x_train, y_train, alpha = 0.5)
plot(elastic_model)
elastic_model$lambda.min
coef(elastic_model, s = "lambda.min")

# Test Set Predictions
ridge_pred <- predict(ridge_model, newx = x_test, s = "lambda.min")
lasso_pred <- predict(lasso_model, newx = x_test, s = "lambda.min")
elastic_pred <- predict(elastic_model, newx = x_test, s = "lambda.min")

# Evaluation Metrics
rmse <- function(actual, pred) {
  sqrt(mean((actual - pred)^2))}

r2 <- function(actual, pred) {
  1 - sum((actual - pred)^2) / sum((actual - mean(actual))^2)}

results <- data.frame(
  Model = c("Ridge", "LASSO", "Elastic Net"),
  RMSE = c(
    rmse(y_test, ridge_pred),
    rmse(y_test, lasso_pred),
    rmse(y_test, elastic_pred)),
  R2 = c(
    r2(y_test, ridge_pred),
    r2(y_test, lasso_pred),
    r2(y_test, elastic_pred)))

results
```