```
# Load required libraries
install.packages("randomForest")
install.packages("keras3")
install.packages("xgboost")
install.packages("ggplot2")
install.packages("corrplot")
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

also installing the dependencies ‘RcppTOML’, ‘here’, ‘png’, ‘config’, ‘tfautogra


Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)
```

```
library(keras3)
library(xgboost)
library(randomForest)
library(ggplot2)
library(corrplot)
```

```
randomForest 4.7-1.2

Type rfNews() to see new features/changes/bug fixes.


Attaching package: ‘ggplot2’


The following object is masked from ‘package:randomForest’:

    margin


corrplot 0.94 loaded
```

```
# Load and summarize the dataset
cat("Loading datasets...\n")
```

```
dataset_files <- c("/content/dataset_files/station00.csv")  # Add other files if nee
```

```
Loading datasets...
```

```
# Initialize empty list to store datasets
datasets <- list()


# Loop through each dataset file and load into a list
for (file in dataset_files) {
  dataset <- read.csv(file)  # Load each dataset
  datasets[[file]] <- dataset  # Store in the list
  cat("Loaded dataset: ", file, "\n")
}
```

```
Loaded dataset:  /content/dataset_files/station00.csv
```

```
# Combine all datasets into one large dataset
combined_dataset <- do.call(rbind, datasets)
cat("Combined dataset dimensions: ", dim(combined_dataset), "\n")
```

```
Combined dataset dimensions:  28896 15
```

```
# Show dataset names and first few rows
cat("Column names of dataset:\n")
print(names(combined_dataset))
cat("First few rows of the dataset:\n")
print(head(combined_dataset, 5))
```

```
Column names of dataset:
 [1] "date_time"         "nwp_globalirrad"   "nwp_directirrad"  
 [4] "nwp_temperature"   "nwp_humidity"      "nwp_windspeed"    
 [7] "nwp_winddirection" "nwp_pressure"      "lmd_totalirrad"   
[10] "lmd_diffuseirrad"  "lmd_temperature"   "lmd_pressure"     
[13] "lmd_winddirection" "lmd_windspeed"     "power"            
First few rows of the dataset:
                                                 date_time nwp_globalirrad
/content/dataset_files/station00.csv.1 2018-08-15 16:00:00               0
/content/dataset_files/station00.csv.2 2018-08-15 16:15:00               0
/content/dataset_files/station00.csv.3 2018-08-15 16:30:00               0
/content/dataset_files/station00.csv.4 2018-08-15 16:45:00               0
/content/dataset_files/station00.csv.5 2018-08-15 17:00:00               0
                                       nwp_directirrad nwp_temperature
/content/dataset_files/station00.csv.1               0           22.78
/content/dataset_files/station00.csv.2               0           22.75
/content/dataset_files/station00.csv.3               0           22.71
/content/dataset_files/station00.csv.4               0           22.64
/content/dataset_files/station00.csv.5               0           22.57
                                       nwp_humidity nwp_windspeed
/content/dataset_files/station00.csv.1        96.85          4.28
/content/dataset_files/station00.csv.2        96.91          4.30
/content/dataset_files/station00.csv.3        96.95          4.28
```

```
/content/dataset_files/station00.csv.4        97.12          4.28
/content/dataset_files/station00.csv.5        97.15          4.33
                                       nwp_winddirection nwp_pressure
/content/dataset_files/station00.csv.1            339.41      1007.27
/content/dataset_files/station00.csv.2            337.27      1007.27
/content/dataset_files/station00.csv.3            334.47      1007.48
/content/dataset_files/station00.csv.4            331.52      1007.39
/content/dataset_files/station00.csv.5            329.78      1007.09
                                       lmd_totalirrad lmd_diffuseirrad
/content/dataset_files/station00.csv.1              0                0
/content/dataset_files/station00.csv.2              0                0
/content/dataset_files/station00.csv.3              0                0
/content/dataset_files/station00.csv.4              0                0
/content/dataset_files/station00.csv.5              0                0
                                       lmd_temperature lmd_pressure
/content/dataset_files/station00.csv.1            25.9       1006.3
/content/dataset_files/station00.csv.2            25.9       1006.2
/content/dataset_files/station00.csv.3            25.8       1006.3
/content/dataset_files/station00.csv.4            25.6       1006.3
/content/dataset_files/station00.csv.5            25.7       1006.3
                                       lmd_winddirection lmd_windspeed power
/content/dataset_files/station00.csv.1               353           1.1     0
/content/dataset_files/station00.csv.2               330           0.9     0
/content/dataset_files/station00.csv.3                 1           1.9     0
/content/dataset_files/station00.csv.4               309           0.4     0
/content/dataset_files/station00.csv.5               335           1.6     0
```

```
# Summary and structure of dataset
cat("Summary of combined dataset:\n")
print(summary(combined_dataset))
cat("Structure of combined dataset:\n")
print(str(combined_dataset))
```

```
Summary of combined dataset:
  date_time         nwp_globalirrad nwp_directirrad nwp_temperature
 Length:28896       Min.   :  0.0   Min.   :  0.0   Min.   :-14.01
 Class :character   1st Qu.:  0.0   1st Qu.:  0.0   1st Qu.:  2.29
 Mode  :character   Median :  0.0   Median :  0.0   Median : 10.51
                    Mean   :168.4   Mean   :147.8   Mean   : 11.06
                    3rd Qu.:305.5   3rd Qu.:259.2   3rd Qu.: 19.58
                    Max.   :942.8   Max.   :885.6   Max.   : 41.09
  nwp_humidity    nwp_windspeed    nwp_winddirection  nwp_pressure
 Min.   :  5.07   Min.   : 0.050   Min.   :  0.03    Min.   : 987.8
 1st Qu.: 23.33   1st Qu.: 2.070   1st Qu.: 89.50    1st Qu.:1007.6
 Median : 35.05   Median : 3.140   Median :186.89    Median :1015.4
 Mean   : 40.83   Mean   : 3.539   Mean   :184.02    Mean   :1014.8
 3rd Qu.: 54.46   3rd Qu.: 4.510   3rd Qu.:280.99    3rd Qu.:1021.1
 Max.   :100.00   Max.   :15.980   Max.   :360.00    Max.   :1040.8
 lmd_totalirrad   lmd_diffuseirrad lmd_temperature   lmd_pressure
 Min.   :   0.0   Min.   :  0.00   Min.   :-13.50   Min.   : 988.5
 1st Qu.:   0.0   1st Qu.:  0.00   1st Qu.:  2.80   1st Qu.:1006.8
 Median :   0.0   Median :  0.00   Median : 11.40   Median :1014.6
 Mean   : 167.4   Mean   : 96.43   Mean   : 11.51   Mean   :1014.4
 3rd Qu.: 275.0   3rd Qu.:138.00   3rd Qu.: 20.50   3rd Qu.:1020.8
 Max.   :1122.0   Max.   :927.00   Max.   : 36.80   Max.   :1043.8
 lmd_winddirection lmd_windspeed        power       
 Min.   :  0       Min.   : 0.000   Min.   :0.0000  
 1st Qu.:114       1st Qu.: 0.400   1st Qu.:0.0000  
 Median :177       Median : 1.100   Median :0.0000  
 Mean   :181       Mean   : 1.471   Mean   :0.8311  
 3rd Qu.:265       3rd Qu.: 2.100   3rd Qu.:1.3957  
 Max.   :359       Max.   :13.300   Max.   :5.5230  
Structure of combined dataset:
'data.frame':	28896 obs. of  15 variables:
 $ date_time        : chr  "2018-08-15 16:00:00" "2018-08-15 16:15:00" "2018-08-
 $ nwp_globalirrad  : num  0 0 0 0 0 0 0 0 0 0 ...
 $ nwp_directirrad  : num  0 0 0 0 0 0 0 0 0 0 ...
 $ nwp_temperature  : num  22.8 22.8 22.7 22.6 22.6 ...
 $ nwp_humidity     : num  96.8 96.9 97 97.1 97.2 ...
 $ nwp_windspeed    : num  4.28 4.3 4.28 4.28 4.33 4.39 4.44 4.55 4.74 5 ...
 $ nwp_winddirection: num  339 337 334 332 330 ...
 $ nwp_pressure     : num  1007 1007 1007 1007 1007 ...
 $ lmd_totalirrad   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ lmd_diffuseirrad : int  0 0 0 0 0 0 0 0 0 0 ...
 $ lmd_temperature  : num  25.9 25.9 25.8 25.6 25.7 ...
 $ lmd_pressure     : num  1006 1006 1006 1006 1006 ...
 $ lmd_winddirection: int  353 330 1 309 335 343 5 342 331 359 ...
 $ lmd_windspeed    : num  1.1 0.9 1.9 0.4 1.6 1.1 1.4 1.5 2.5 2.8 ...
 $ power            : num  0 0 0 0 0 0 0 0 0 0 ...
NULL
```

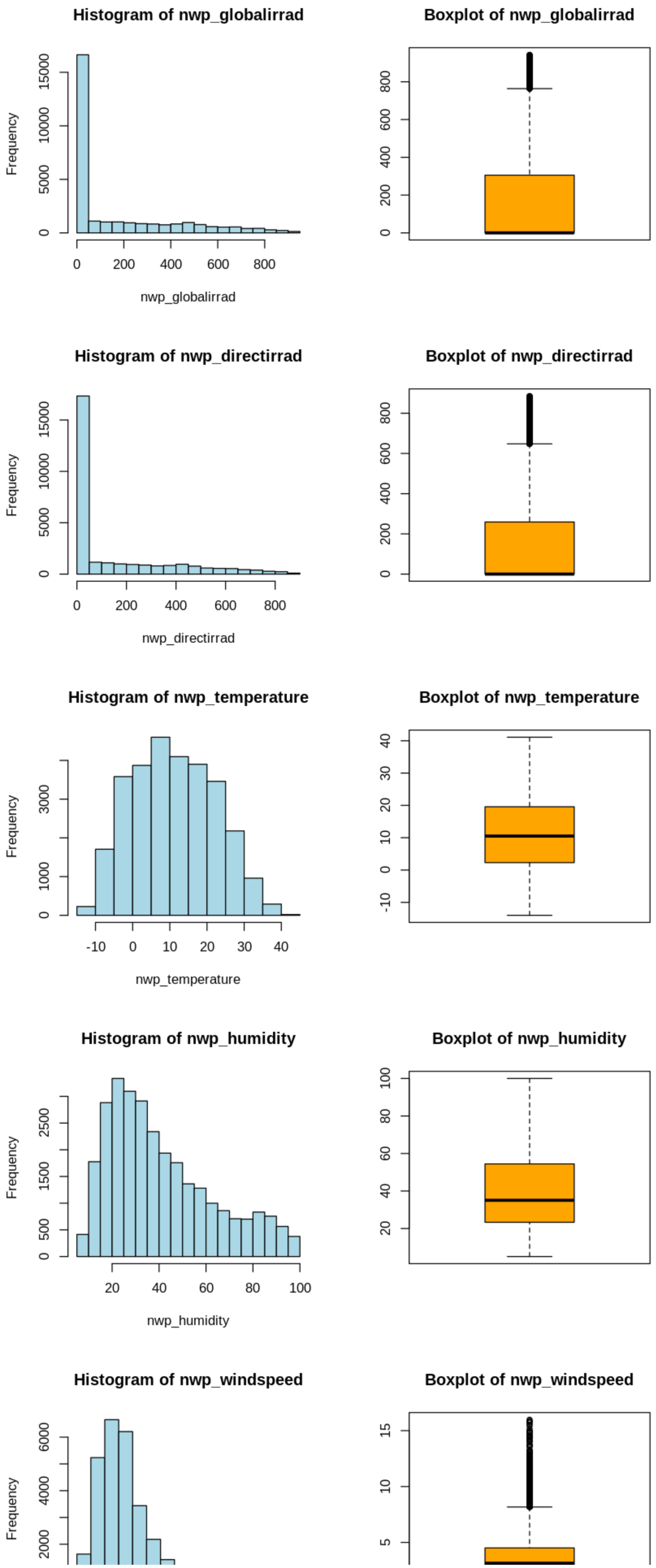
```
# Checking for missing values
cat("Checking for missing values...\n")
missing_values <- sum(is.na(combined_dataset))
cat("Total missing values in dataset: ", missing_values, "\n")
```

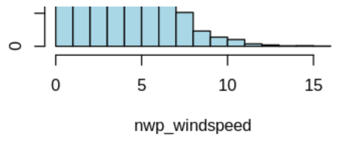
```
Checking for missing values...
Total missing values in dataset:  0
```

```
# ---------- Data Visualization using par() and corrplot() ----------
cat("Visualizing dataset distributions...\n")
```

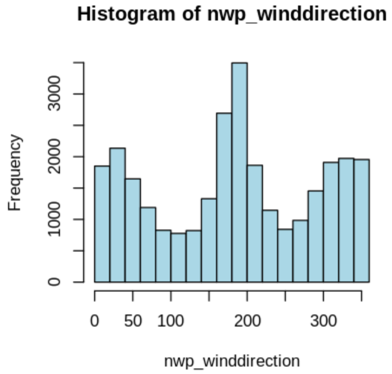
```
Visualizing dataset distributions...
```

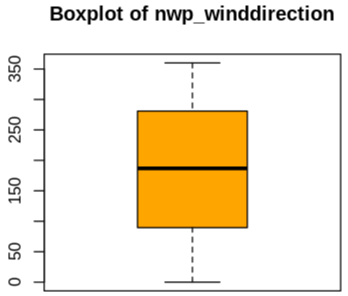
```
# Using par() to create multiple plots in one graphic
par(mfrow=c(2, 2))  # 2 rows, 2 columns of plots
for (i in 2:(ncol(combined_dataset)-1)) {
  column_name <- names(combined_dataset)[i]
  hist(combined_dataset[[i]], main = paste("Histogram of", column_name), xlab = colu
  boxplot(combined_dataset[[i]], main = paste("Boxplot of", column_name), col = "ora
}
```

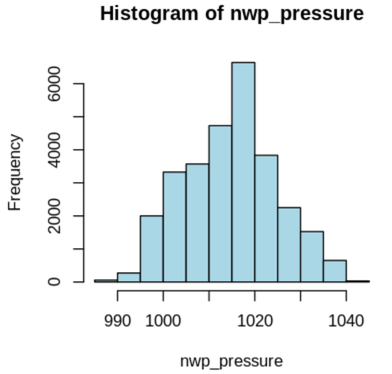

Histogram of nwp_globalirrad
Frequency
0
5000
10000
15000
0
200
400
600
800
nwp_globalirrad
Boxplot of nwp_globalirrad
0
200
400
600
800
Histogram of nwp_directirrad
Frequency
0
5000
10000
15000
0
200
400
600
800
nwp_directirrad
Boxplot of nwp_directirrad
0
200
400
600
800
Histogram of nwp_temperature
Frequency
0
1000
3000
-10
0
10
20
30
40
nwp_temperature
Boxplot of nwp_temperature
-10
0
10
20
30
40
Histogram of nwp_humidity
Frequency
0
500
1500
2500
20
40
60
80
100
nwp_humidity
Boxplot of nwp_humidity
20
40
60
80
100
Histogram of nwp_windspeed
Frequency
2000
4000
6000
Boxplot of nwp_windspeed
5
10
15

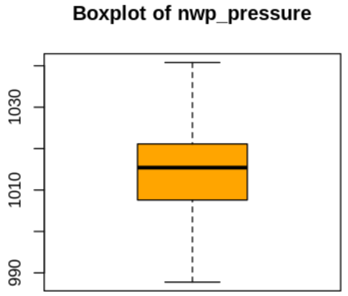

0
0
5
10
15
nwp_windspeed

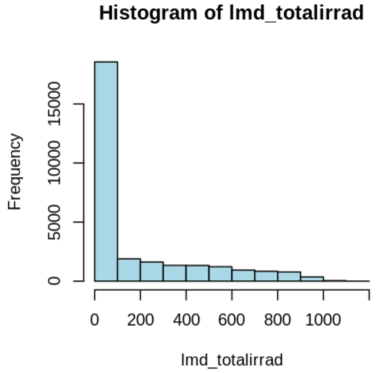
0

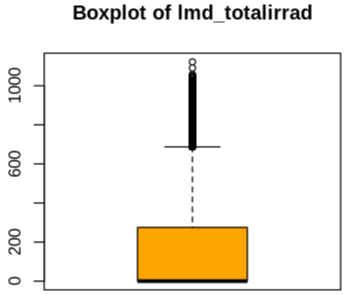

Histogram of nwp_winddirection
Frequency
0
1000
2000
3000
0
50
100
200
300
nwp_winddirection

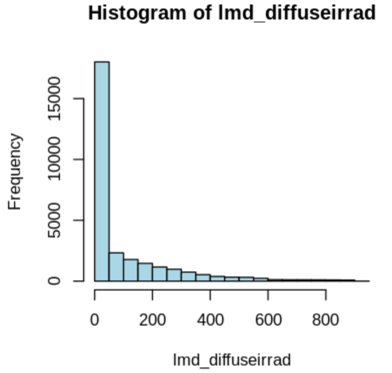

Boxplot of nwp_winddirection
0
50
150
250
350

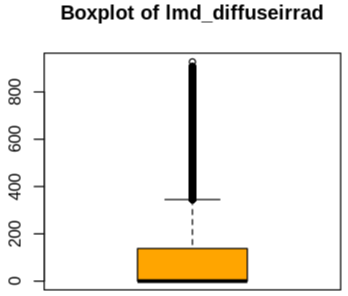

Histogram of nwp_pressure
Frequency
0
2000
4000
6000
990
1000
1020
1040
nwp_pressure

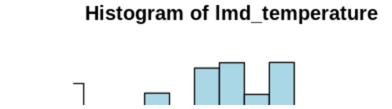

Boxplot of nwp_pressure
990
1010
1030

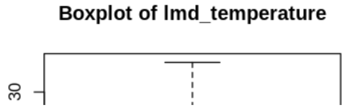

Histogram of lmd_totalirrad
Frequency
0
5000
10000
15000
0
200
400
600
800
1000
lmd_totalirrad

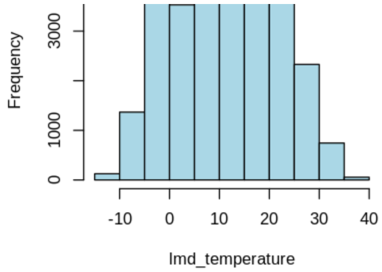

Boxplot of lmd_totalirrad
0
200
600
1000

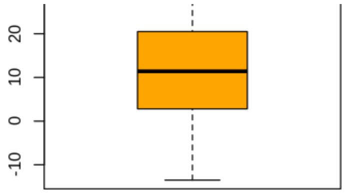

Histogram of lmd_diffuseirrad
Frequency
0
5000
10000
15000
0
200
400
600
800
lmd_diffuseirrad

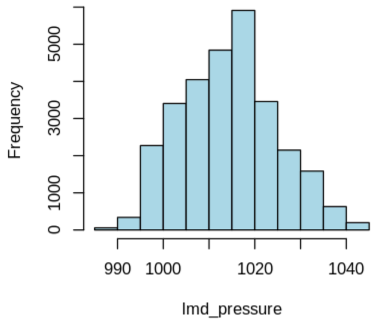

Boxplot of lmd_diffuseirrad
0
200
400
600
800

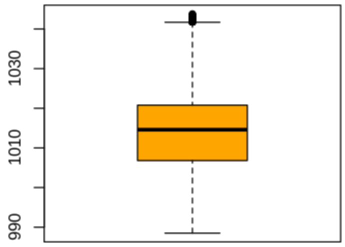

Histogram of lmd_temperature

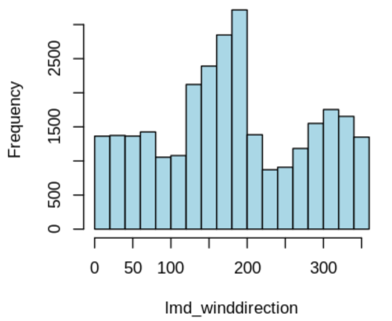

Boxplot of lmd_temperature
30

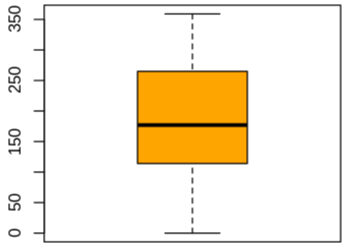


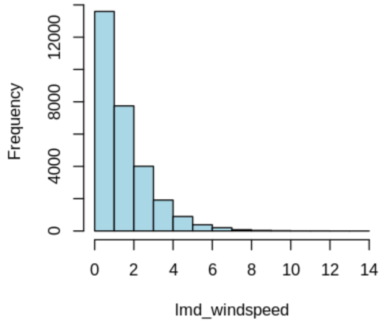


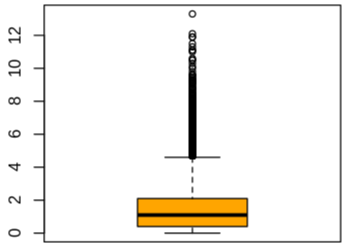

Histogram of lmd_pressure


Boxplot of lmd_pressure


Histogram of lmd_winddirection

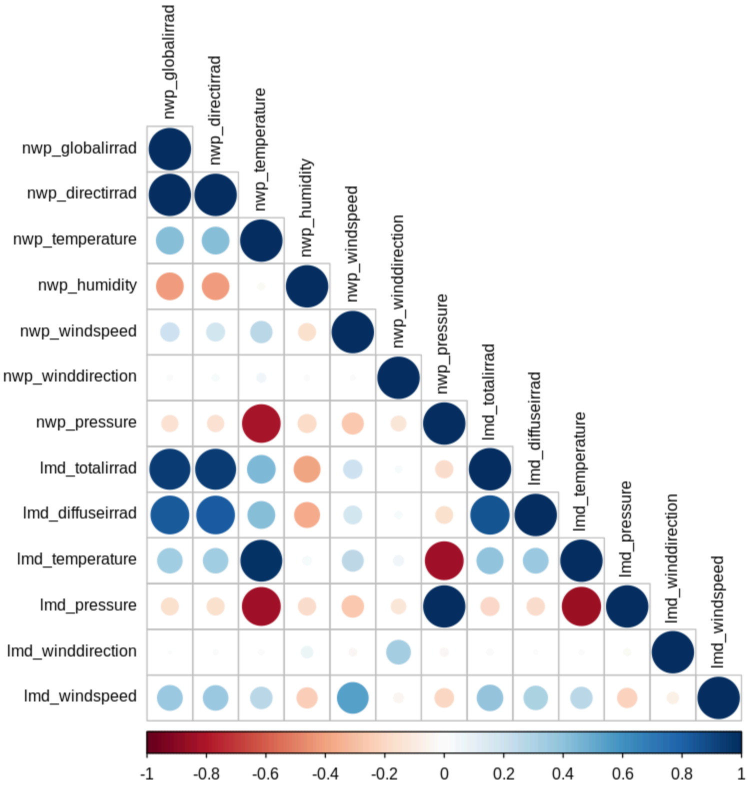

Boxplot of lmd_winddirection


Histogram of lmd_windspeed


Boxplot of lmd_windspeed

```
# Correlation plot for the dataset (for the first 13 columns)
cat("Creating correlation plot...\n")
cor_matrix <- cor(combined_dataset[, 2:(ncol(combined_dataset)-1)])
corrplot(cor_matrix, method = "circle", type = "lower", tl.col = "black", tl.cex = 0
```

```
Creating correlation plot...
```



```
# Subset the data to only use 300 samples
set.seed(42)
sampled_idx <- sample(1:nrow(combined_dataset), 300)
sampled_data <- combined_dataset[sampled_idx, ]


# Select middle 13 columns as features and last column as the target variable
if (ncol(combined_dataset) < 15) stop("Not enough columns in the dataset.")
X <- combined_dataset[, 2:(ncol(combined_dataset)-1)]  # Middle 13 columns as featur
y <- combined_dataset$power  # Last column 'power' as the target


# Split into 80% training and 20% testing
set.seed(42)
n <- nrow(X)
train_idx <- sample(1:n, size = 0.8 * n)
test_idx <- setdiff(1:n, train_idx)

X_train <- as.matrix(X[train_idx, ])
y_train <- y[train_idx]
X_test <- as.matrix(X[test_idx, ])
y_test <- y[test_idx]
```

```
cat("Training and test data split completed. Training rows: ", nrow(X_train), " Test
```

```
Training and test data split completed. Training rows:  23116  Testing rows:  57
```

```
# ---------- Train LSTM Model ----------
cat("Training LSTM model...\n")
X_train_lstm <- array(X_train, dim = c(nrow(X_train), ncol(X_train), 1))
X_test_lstm <- array(X_test, dim = c(nrow(X_test), ncol(X_test), 1))

lstm_model <- keras_model_sequential() %>%
  layer_lstm(units = 10, input_shape = list(ncol(X_train), 1), return_sequences = FA
  layer_dense(units = 1)

lstm_model %>% compile(
  optimizer = 'adam',
  loss = 'mse'
)

lstm_model %>% fit(
  X_train_lstm, y_train, epochs = 10, batch_size = 30, verbose = 1
)

lstm_train_pred <- lstm_model %>% predict(X_train_lstm)
lstm_test_pred <- lstm_model %>% predict(X_test_lstm)
```

```
Training LSTM model...
```

```
# ---------- Train XGBoost Model ----------
cat("Training XGBoost model...\n")
dtrain <- xgb.DMatrix(data = X_train, label = y_train)
dtest <- xgb.DMatrix(data = X_test)

xgb_params <- list(
  objective = "reg:squarederror",
  eta = 0.1,
  max_depth = 3
)

xgb_model <- xgb.train(params = xgb_params, data = dtrain, nrounds = 50)

xgb_train_pred <- predict(xgb_model, X_train)
xgb_test_pred <- predict(xgb_model, X_test)
```

```
Training XGBoost model...
```

```
# ---------- Train Random Forest Model ----------
cat("Training Random Forest model...\n")
```

```
rf_model <- randomForest(X_train, y_train, ntree = 100)

rf_train_pred <- predict(rf_model, X_train)
rf_test_pred <- predict(rf_model, X_test)
```

```
Training Random Forest model...
```

```
# ---------- Stacking (Meta Learner) ----------
cat("Training meta-learner (linear regression)...\n")
stack_train <- data.frame(
  lstm = lstm_train_pred,
  xgb = xgb_train_pred,
  rf = rf_train_pred
)

stack_test <- data.frame(
  lstm = lstm_test_pred,
  xgb = xgb_test_pred,
  rf = rf_test_pred
)

meta_model <- lm(y_train ~ ., data = stack_train)

final_train_pred <- predict(meta_model, stack_train)
final_test_pred <- predict(meta_model, stack_test)
```

```
Training meta-learner (linear regression)...
```

```
# ---------- Model Evaluation ----------
cat("Evaluating models using RMSE and R-squared...\n")
rmse <- function(actual, predicted) {
  sqrt(mean((actual - predicted)^2))
}

r_squared <- function(actual, predictions) {
  ss_res <- sum((actual - predictions)^2)  # Residual sum of squares
  ss_tot <- sum((actual - mean(actual))^2)  # Total sum of squares
  1 - (ss_res / ss_tot)
}

train_rmse <- rmse(y_train, final_train_pred)
test_rmse <- rmse(y_test, final_test_pred)

cat("Train RMSE: ", train_rmse, "\n")
cat("Test RMSE: ", test_rmse, "\n")

train_r_squared_value <- r_squared(y_train, final_train_pred)
test_r_squared_value <- r_squared(y_test, final_test_pred)
```

```
cat("Train R-squared: ", train_r_squared_value, "\n")
cat("Test R-squared: ", test_r_squared_value, "\n")
```

```
Evaluating models using RMSE and R-squared...
Train RMSE:  0.05272727
Test RMSE:  0.1475873
Train R-squared:  0.9983231
Test R-squared:  0.9865005
```

```
# ---------- Condensed Summary of Linear Regression Model ----------
cat("Summary of the linear regression model:\n")

# Output model call
cat("Call:\n")
print(summary(meta_model)$call)

# Output key metrics of residuals (Min, 1Q, Median, 3Q, Max)
cat("Residuals (key statistics):\n")
residuals_summary <- summary(meta_model)$residuals
print(head(residuals_summary, 5))  # Show only the first 5 residual statistics

# Output significant coefficients with stars indicating significance
cat("Coefficients (significant predictors):\n")
coefficients <- summary(meta_model)$coefficients
print(coefficients[abs(coefficients[, "t value"]) > 2, ])  # Show only significant c

# Output key model fit metrics: RSE, R-squared, F-statistic
cat("Model Fit Metrics:\n")
cat("Residual standard error (RSE): ", summary(meta_model)$sigma, "\n")
cat("Multiple R-squared: ", summary(meta_model)$r.squared, "\n")
cat("F-statistic: ", summary(meta_model)$fstatistic[1], "\n")
```

```
Summary of the linear regression model:
Call:
lm(formula = y_train ~ ., data = stack_train)
Residuals (key statistics):
/content/dataset_files/station00.csv.27185
                               0.001555552
/content/dataset_files/station00.csv.28645
                               0.152380010
/content/dataset_files/station00.csv.18753
                              -0.009848704
/content/dataset_files/station00.csv.21657
                              -0.103271500
 /content/dataset_files/station00.csv.9290
                               0.002210861
Coefficients (significant predictors):
        Estimate  Std. Error   t value     Pr(>|t|)
lstm -0.02866712 0.002224339 -12.88793 7.109742e-38
xgb  -0.29571011 0.003248952 -91.01707 0.000000e+00
rf    1.32158493 0.002326961 567.94450 0.000000e+00
Model Fit Metrics:
Residual standard error (RSE):  0.05273183
```