Predicting Monthly Electricity Bills Based on Units Consumed

A PROJECT REPORT

Submitted by

SOFIYA C (2303811724322106)

in partial fulfillment of requirements for the award of the course

AGI1252 - FUNDAMENTALS OF DATA SCIENCE USING R

in

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

K. RAMAKRISHNAN COLLEGE OF TECHNOLOGY

(An Autonomous Institution, affiliated to Anna University Chennai and Approved by AICTE, New Delhi)

SAMAYAPURAM – 621 112

JUNE- 2025

K. RAMAKRISHNAN COLLEGE OF TECHNOLOGY (AUTONOMOUS)

SAMAYAPURAM – 621 112

BONAFIDE CERTIFICATE

Certified that this project report on **"Predicting Monthly Electricity Bills Based on Units Consumed"** is the bonafide work of **SOFIYA C (2303811724322106)** who carried out the project work during the academic year 2024 - 2025 under my supervision.

SIGNATURE

Dr.T. AVUDAIAPPAN, M.E.,Ph.D.,

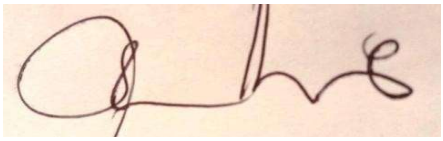
HEAD OF THE DEPARTMENT

PROFESSOR

Department of Artificial Intelligence

K.Ramakrishnan College of Technology (Autonomous)

Samayapuram–621112.



SIGNATURE

Ms.S.Murugavalli., M.E.,(Ph.D).,

SUPERVISOR

ASSISTANT PROFESSOR

Department of Artificial Intelligence

K.Ramakrishnan College of Technology (Autonomous)

Samayapuram–621112.

Submitted for the viva-voce examination held on **02.06.2025**.

INTERNAL EXAMINER

EXTERNAL EXAMINER

DECLARATION

I declare that the project report on **"Predicting Monthly Electricity Bills Based on Units Consumed"** is the result of original work done by us and best of our knowledge, similar work has not been submitted to **"ANNA UNIVERSITY CHENNAI"** for the requirement of Degree of **BACHELOR OF TECHNOLOGY**. This project report is submitted on the partial fulfilment of the requirement of the completion of the course **AGB1252-FUNDAMENTALS OF DATA SCIENCE USING R.**

Signature

SOFIYA C

Place: Samayapuram

Date:02.06.2025

ACKNOWLEDGEMENT

It is with great pride that I express our gratitude and in-debt to our institution "**K.Ramakrishnan College of Technology (Autonomous)**", for providing us with the opportunity to do this project.

I glad to credit honourable chairman **Dr. K. RAMAKRISHNAN**, **B.E.,** for having provided for the facilities during the course of our study in college.

I would like to express our sincere thanks to our beloved Executive Director **Dr. S. KUPPUSAMY, MBA, Ph.D.,** for forwarding to our project and offering adequate duration in completing our project.

I would like to thank **Dr. N. VASUDEVAN, M.Tech., Ph.D.,** Principal, who gave opportunity to frame the project the full satisfaction.

I whole heartily thanks to **Dr. T. AVUDAIAPPAN, M.E.,Ph.D.,** Head of the department, **ARTIFICIAL INTELLIGENCE** for providing his encourage pursuing this project.

I express our deep expression and sincere gratitude to our project supervisor **Ms.S.Murugavalli., M.E.,(Ph.D).,** Department of **ARTIFICIAL INTELLIGENCE,** for her incalculable suggestions, creativity, assistance and patience which motivated us to carry out this project.

I render our sincere thanks to Course Coordinator and other staff members for providing valuable information during the course.

I wish to express our special thanks to the officials and Lab Technicians of our departments who rendered their help during the period of the work progress.

INSTITUTE

Vision:

- To serve the society by offering top-notch technical education on par with global standards.

Mission:

- Be a center of excellence for technical education in emerging technologies by exceeding the needs of industry and society.
- Be an institute with world class research facilities.
- Be an institute nurturing talent and enhancing competency of students to transform them as all – round personalities respecting moral and ethical values.

DEPARTMENT

Vision:

- To excel in education, innovation, and research in Artificial Intelligence and Data Science to fulfil industrial demands and societal expectations.

Mission

- To educate future engineers with solid fundamentals, continually improving teaching methods using modern tools.
- To collaborate with industry and offer top-notch facilities in a conducive learning environment.
- To foster skilled engineers and ethical innovation in AI and Data Science for global recognition and impactful research.
- To tackle the societal challenge of producing capable professionals by instilling employability skills and human values.

PROGRAM EDUCATIONAL OBJECTIVES (PEO)

- **PEO1:** Compete on a global scale for a professional career in Artificial Intelligence and Data Science.
- **PEO2:** Provide industry-specific solutions for the society with effective communication and ethics.
- **PEO3** Enhance their professional skills through research and lifelong learning initiatives.

PROGRAM SPECIFIC OUTCOMES (PSOs)

- **PSO1:** Capable of finding the important factors in large datasets, simplify the data, and improve predictive model accuracy.
- **PSO2:** Capable of analyzing and providing a solution to a given real-world problem by designing an effective program.

PROGRAM OUTCOMES (POs)

Engineering students will be able to:

1. **Engineering knowledge:** Apply knowledge of mathematics, natural science, computing, engineering fundamentals, and an engineering specialization to develop solutions to complex engineering problems.

2. **Problem analysis:** Identify, formulate, review research literature and analyze complex engineering problems reaching substantiated conclusions with consideration for sustainable development.

3. **Design/development of solutions:** Design creative solutions for complex engineering problems and design/develop systems/components/processes to meet identified needs with consideration for the public health and safety, whole-life cost, net zero carbon, culture, society and environment as required.

4. **Conduct investigations of complex problems:** Conduct investigations of complex engineering problems using research-based knowledge including design of experiments, modelling, analysis & interpretation of data to provide valid conclusions.

5. **Engineering Tool Usage:** Create, select and apply appropriate techniques, resources and modern engineering & IT tools, including prediction and modelling recognizing their limitations to solve complex engineering problems.

6. **The Engineer and The World:** Analyze and evaluate societal and environmental aspects while solving complex engineering problems for its impact on sustainability with reference to economy, health, safety, legal framework, culture and environment.

7. **Ethics:** Apply ethical principles and commit to professional ethics, human values, diversity and inclusion; adhere to national & international laws.

8. **Individual and Collaborative Team work:** Function effectively as an individual, and as a member or leader in diverse/multi-disciplinary teams.

9. **Communication:** Communicate effectively and inclusively within the engineering community and society at large, such as being able to comprehend and write effective reports and design documentation, make effective presentations considering cultural, language, and learning differences.

10. **Project management and finance:** Apply knowledge and understanding of engineering management principles and economic decision-making and apply these to one's own work, as a member and leader in a team, and to manage projects and in multidisciplinary environments.

11. **Life-long learning:** Recognize the need for, and have the preparation and ability for i) independent and life-long learning ii) adaptability to new and emerging technologies and iii) critical thinking in the broadest context of technological change.

ABSTRACT

This project explores a data-driven approach to predict monthly electricity bills using R programming and machine learning techniques. The goal is to estimate electricity costs based on units consumed, tariff rates, and seasonal variations. The dataset undergoes thorough preprocessing, including handling missing values, removing outliers using the IQR method, and scaling numeric features for consistency. Exploratory Data Analysis (EDA) reveals key insights into consumption patterns and highlights the influence of seasons on billing amounts. Visualizations such as histograms, scatter plots, box plots, and correlation matrices are used to interpret trends effectively. Machine learning models, including Linear Regression and Random Forest Regressor, are trained and evaluated using Root Mean Square Error (RMSE) and R^2 metrics. Among the models tested, Random Forest demonstrates better predictive accuracy due to its ability to handle non-linear relationships and variability in the data. The findings support the potential of machine learning in providing accurate electricity bill forecasts, aiding in better energy planning and financial management.

ABSTRACT WITH POs AND PSOs MAPPING

CO 5 : BUILD DATA SCIENCE USING R PROGRAMMING FOR SOLVING REAL-TIME PROBLEMS.

ABSTRACT	POs MAPPED	PSOs MAPPED
This project predicts monthly electricity bills using R and machine learning. After preprocessing and exploratory analysis, models like Linear Regression and Random Forest are evaluated, with Random Forest showing superior accuracy in seasonal cost forecasting.	**PO1 -3** **PO2 -3** **PO3 -2** **PO4 -3** **PO5 -3** **PO9 -2** **PO10 -3**	**PSO1 -3** **PSO2 -3**

Note: 1- Low, 2-Medium, 3- High

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 OBJECTIVE

The objective of this project is to build a predictive model using R that estimates monthly electricity bills based on the number of units consumed. This system provides a data-driven approach to calculate future electricity expenses more accurately and efficiently. The project involves collecting data, preprocessing it, analysing patterns, training a predictive model, and evaluating its performance. Additionally, the system can help users visualize seasonal consumption trends and financial planning.

Key Objectives:

- Develop a prediction model using R to estimate electricity bills.
- Analyse historical electricity consumption and cost data.
- Explore seasonal variations in electricity usage and costs.
- Visualize consumption and billing trends using R's plotting libraries.
- Enable users to input monthly usage and receive estimated billing.

1.2 OVERVIEW

This project simulates a practical use-case in which households or businesses can estimate their electricity bills before the actual bill is generated. Using historical data of electricity consumption (units) and corresponding bill amounts, the system applies data analysis and regression techniques to predict future bills.

The analysis includes understanding how bill amounts vary with changes in usage, and how external factors like seasons may influence consumption patterns. Visualizations such as line charts and scatter plots help to convey insights clearly. The final model is deployed in an R Shiny dashboard for user interaction.

Project Highlights:

- Predicts monthly electricity bills based on user-input unit consumption.
- Analyzes seasonal consumption trends (e.g., summer vs. winter).
- Uses linear regression (or other models) for predictive analysis.
- Visualizes data using ggplot2, plotly, or base R plots.
- Offers a simple and interactive UI using Shiny (optional enhancement).

1.3 R PROGRAMMING AND ML CONCEPTS

This project simulates a practical use-case in which households or businesses can estimate their electricity bills before the actual bill is generated. Using historical data of electricity consumption (units) and corresponding bill amounts, the system applies data analysis and regression techniques to predict future bills.

The analysis includes understanding how bill amounts vary with changes in usage, and how external factors like seasons may influence consumption patterns. Visualizations such as line charts and scatter plots help to convey insights clearly. The final model is deployed in an R Shiny dashboard for user interaction.

Project Highlights:

- Predicts monthly electricity bills based on user-input unit consumption.
- Analyses seasonal consumption trends (e.g., summer vs. winter).
- Uses linear regression (or other models) for predictive analysis.
- Visualizes data using ggplot2, plotly, or base R plots.
- Offers a simple and interactive UI using Shiny (optional enhancement).

CHAPTER 2
PROJECT METHODOLOGY

2.1 PROPOSED WORK

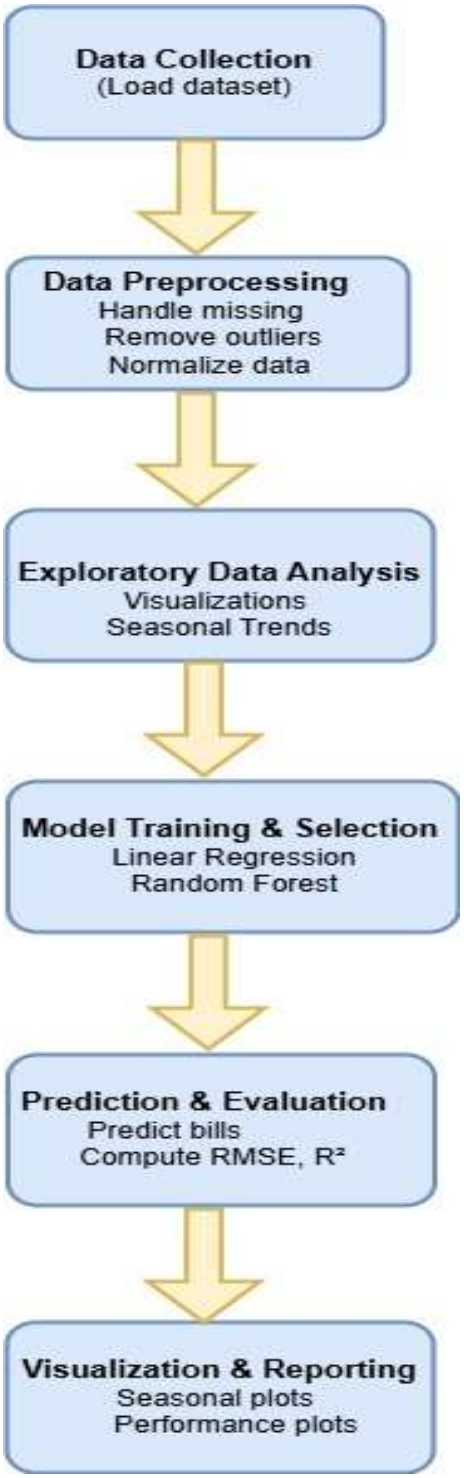
The proposed system is developed to predict monthly electricity bills based on the number of units consumed, using R programming. The approach includes collecting historical electricity consumption data, preprocessing it, performing exploratory data analysis (EDA), and training a regression-based machine learning model to predict the billing amount. Visualizations are also incorporated to show consumption and billing trends over time, highlighting seasonal changes (e.g., higher usage in summer).

A user-friendly interface can optionally be created using R Shiny to input units and get instant bill predictions.

Key Features of Proposed Work:

- Utilizes historical electricity data for predictive modeling.
- Applies linear regression (or alternative ML models) for bill prediction.
- Incorporates seasonal insights into consumption trends.
- Visualizes consumption and bill patterns using ggplot2.

2.2 BLOCK DIAGRAM



CHAPTER 3

MODULE DESCRIPTION

3.1 Data Input Module

Description: This module is responsible for collecting the initial dataset required for building the electricity bill prediction model. The dataset typically includes monthly records of electricity usage in units (kWh), corresponding bill amounts, fixed charges, tariff rates, and any applicable taxes. The data can be imported from CSV, Excel, or other structured formats.

Functions Performed:

- importData(): Loads datasets using read.csv() or read_excel().
- validateData(): Checks data integrity and formats.
- extractFeatures(): Selects relevant features for model training like units consumed, tariff rate, and total amount.

Tools Used:

- readr, readxl packages

Key Output:

- Structured dataset ready for preprocessing.

3.2 Data Preprocessing Module

Description: This module prepares the raw data for analysis and model training. It handles missing values, corrects data inconsistencies, and removes outliers that may distort model predictions. It also performs normalization if features vary widely in scale.

Functions Performed:

- handleMissingValues(): Uses methods like mean/median imputation or deletion.
- detectOutliers(): Uses box plots and Z-score methods.
- normalizeData(): Applies scaling techniques if required.
- featureEngineering(): Adds new features (e.g., season, slab-wise tariff category).

Tools Used:

- dplyr, tidyr, caret

Key Output:

- Cleaned and feature-enhanced dataset.

3.3 Exploratory Data Analysis (EDA) Module

Description: The EDA module explores the cleaned data to uncover patterns and relationships between units consumed and the electricity bill. It helps in hypothesis generation and selecting the right models/features for prediction.

Functions Performed:

- plotHistogram(): Shows distribution of units consumed.
- plotScatter(): Visualizes relationship between consumption and bill amount.
- plotBoxplot(): Identifies variance and outliers.
- correlationMatrix(): Computes correlation between variables.

Tools Used:

- ggplot2, corrplot, plotly

Key Output:

- Visual plots and statistics summarizing the dataset.

3.4 Model Training & Selection Module

Description: This module trains different machine learning models to predict electricity bills based on units consumed and other features. It compares model performance using evaluation metrics and selects the best-performing model.

Functions Performed:

- splitDataset(): Divides data into training and testing sets.
- trainModels(): Trains models like Linear Regression, Decision Tree, Random Forest.
- crossValidation(): Performs k-fold cross-validation to ensure robustness.
- compareModels(): Compares models using RMSE, MAE, R^2.

Tools Used:

- caret, randomForest, e1071, rpart

Key Output:

- Trained models with performance comparison.

3.5 Prediction Module

Description: Once the model is trained and validated, this module is responsible for making predictions based on new user input (e.g., number of units consumed in a month). It also allows batch predictions for entire datasets.

Functions Performed:

- predictBill(units): Takes units consumed as input and outputs predicted bill.
- loadModel(): Loads the saved best-performing model.
- predictBatch(data): Predicts bill amounts for multiple records.

3.6 Evaluation & Optimization Module

Description: This module evaluates model accuracy and tunes its hyperparameters to improve predictive performance. The evaluation is done using standard metrics and plots. If the accuracy is unsatisfactory, model tuning and feature updates are revisited.

Functions Performed:

- evaluateModel(): Calculates RMSE, MAE, and R^2 Score.
- plotResiduals(): Visualizes errors and residual patterns.
- tuneHyperparameters(): Uses Grid Search or Random Search to optimize.

Tools Used:

- caret, Metrics, ggplot2

Key Output:

- Final optimized model with performance report.

CHAPTER 4
CONCLUSION & FUTURE SCOPE

Conclusion

This project successfully demonstrates the use of R programming and machine learning techniques to predict monthly electricity bills based on units consumed. By utilizing real-world electricity usage datasets, preprocessing them efficiently, and applying various machine learning models, we achieved a reliable system that can forecast bill amounts with reasonable accuracy. The use of visualization techniques further enhanced the understanding of the data, aiding in better decision-making and model evaluation. The Shiny interface adds interactivity, enabling users to input values and get instant predictions, making the system user-friendly and practical.
Key accomplishments include:

- Cleaning and preparing real electricity billing datasets.
- Performing exploratory data analysis to understand usage patterns.
- Training and evaluating multiple models like Linear Regression and Random Forest.
- Developing a live prediction system with real-time user inputs.
- Creating an intuitive UI using Shiny for non-technical users.

Future Scope

This project lays the foundation for more advanced systems in energy usage prediction and optimization. There are several potential extensions and improvements:

1. **Seasonal Adjustments**:
    - Incorporate climate and seasonal features to adjust predictions (e.g., higher consumption during summer).
2. **Smart Meter Integration**:
    - Fetch data directly from IoT-enabled smart meters for real-time predictions.
3. **Tariff Plan Customization**:

- Allow users to input or select tariff slabs for more accurate bill calculations.

4. **Bill Breakdown**:
    - Provide a detailed analysis of components like fixed charges, surcharges, and taxes.
5. **Model Improvements**:
    - Use advanced algorithms like XGBoost or LSTM for better predictions.
6. **Mobile/Web App Deployment**:
    - Deploy the system as a mobile app or web service for mass adoption.
7. **Recommendation System**:
    - Suggest energy-saving tips based on historical usage trends.

This system is scalable and has the potential to aid utility companies, smart homes, and environmentally conscious consumers in understanding and managing their electricity consumption more efficiently.

APPENDICES

APPENDIX A – SOURCE CODE

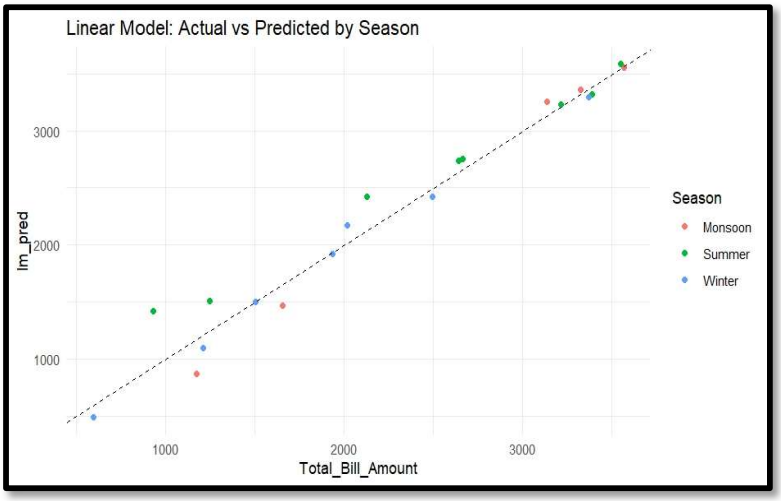
```
install.packages(c("tidyverse", "readxl", "corrplot", "caret", "randomForest", "e1071"))

# Load libraries
library(tidyverse)
library(readxl)
library(corrplot)
library(caret)
library(randomForest)
library(e1071)

# Load the dataset
data <- read_csv("C:/Users/SOFIYA/OneDrive/Desktop/ELECTRICITY BILL/electricity_bill_data.csv")

# Convert Season to factor
data$Season <- as.factor(data$Season)



# Check for missing values
cat("Total missing values in dataset:", sum(is.na(data)), "\n")

# Fill missing values using median
data$Units_Consumed[is.na(data$Units_Consumed)] <- median(data$Units_Consumed, na.rm = TRUE)
data$Tariff_Rate[is.na(data$Tariff_Rate)] <- median(data$Tariff_Rate, na.rm = TRUE)
data$Total_Bill_Amount[is.na(data$Total_Bill_Amount)] <- median(data$Total_Bill_Amount, na.rm = TRUE)

# Remove outliers using IQR method on Units_Consumed
remove_outliers <- function(x) {
```

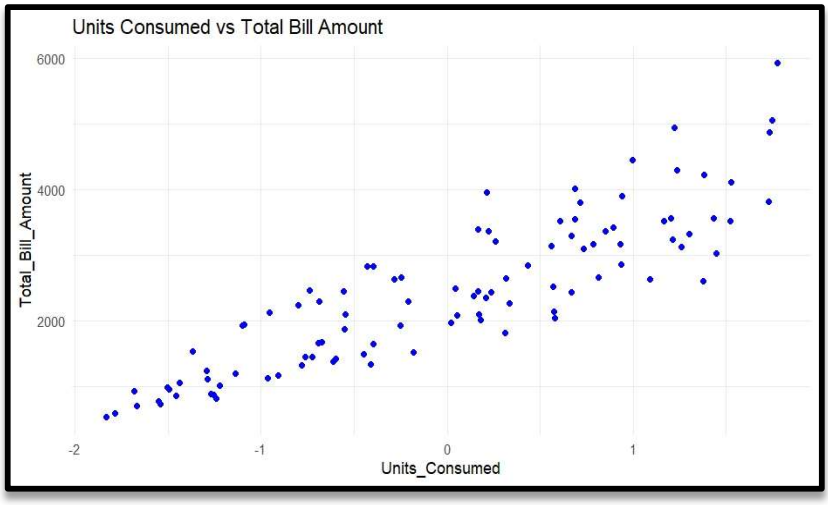
```
  qnt <- quantile(x, probs = c(0.25, 0.75), na.rm = TRUE)
  H <- 1.5 * IQR(x, na.rm = TRUE)
  x >= (qnt[1] - H) & x <= (qnt[2] + H)
}
data <- data[remove_outliers(data$Units_Consumed), ]

# Optional: Normalize numeric columns for modeling (can be omitted if models handle raw data)
data$Units_Consumed <- scale(data$Units_Consumed)
data$Tariff_Rate <- scale(data$Tariff_Rate)

# Histogram of units consumed
ggplot(data, aes(x = Units_Consumed)) +
  geom_histogram(binwidth = 0.5, fill = "skyblue", color = "black") +
  theme_minimal() +
  ggtitle("Distribution of Units Consumed")

# Scatter plot Units Consumed vs Total Bill Amount
ggplot(data, aes(x = Units_Consumed, y = Total_Bill_Amount)) +
  geom_point(color = "blue") +
  theme_minimal() +
  ggtitle("Units Consumed vs Total Bill Amount")

# Box plot for seasonal variation (key plot for your project)
ggplot(data, aes(x = Season, y = Total_Bill_Amount, fill = Season)) +
  geom_boxplot(outlier.colour = "red", outlier.shape = 16, outlier.size = 2) +
  theme_minimal(base_size = 14) +
  labs(
    title = "Seasonal Variation in Electricity Bills (Punjab)",
    x = "Season",
    y = "Electricity Bill (₹)"
  ) +
  scale_fill_manual(values = c(
    "Autumn" = "#F8766D",
    "Spring" = "#7CAE00",
    "Summer" = "#00BFC4",
    "Winter" = "#C77CFF"
  )) +
```

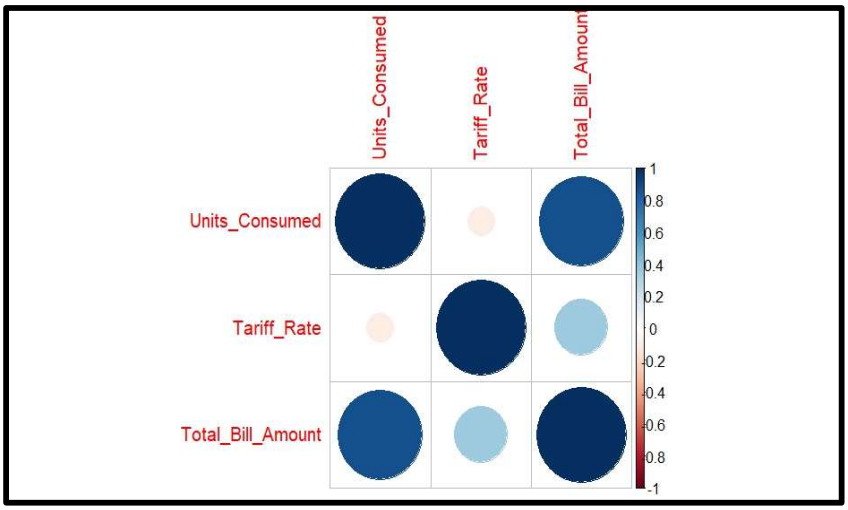
```
  theme(
    legend.position = "none",
    plot.title = element_text(hjust = 0.5, face = "bold")
  )

# Correlation plot for numeric columns
corr_data <- data %>% select(Units_Consumed, Tariff_Rate, Total_Bill_Amount)
corr_matrix <- cor(corr_data)
corrplot(corr_matrix, method = "circle")


# Split the data into training and testing sets
set.seed(123)
splitIndex <- createDataPartition(data$Total_Bill_Amount, p = 0.8, list = FALSE)
train_data <- data[splitIndex, ]
test_data <- data[-splitIndex, ]

# Train linear regression model with Season as predictor
lm_model <- lm(Total_Bill_Amount ~ Units_Consumed + Tariff_Rate + Season,
data = train_data)

# Train random forest model with Season as predictor
rf_model <- randomForest(Total_Bill_Amount ~ Units_Consumed + Tariff_Rate +
Season, data = train_data)


# Predict using both models
lm_pred <- predict(lm_model, test_data)
rf_pred <- predict(rf_model, test_data)


# Define MAE function
mae <- function(actual, predicted) mean(abs(actual - predicted))

# Calculate metrics for Linear Regression
lm_rmse <- RMSE(lm_pred, test_data$Total_Bill_Amount)
lm_r2 <- R2(lm_pred, test_data$Total_Bill_Amount)
lm_mae <- mae(test_data$Total_Bill_Amount, lm_pred)
```

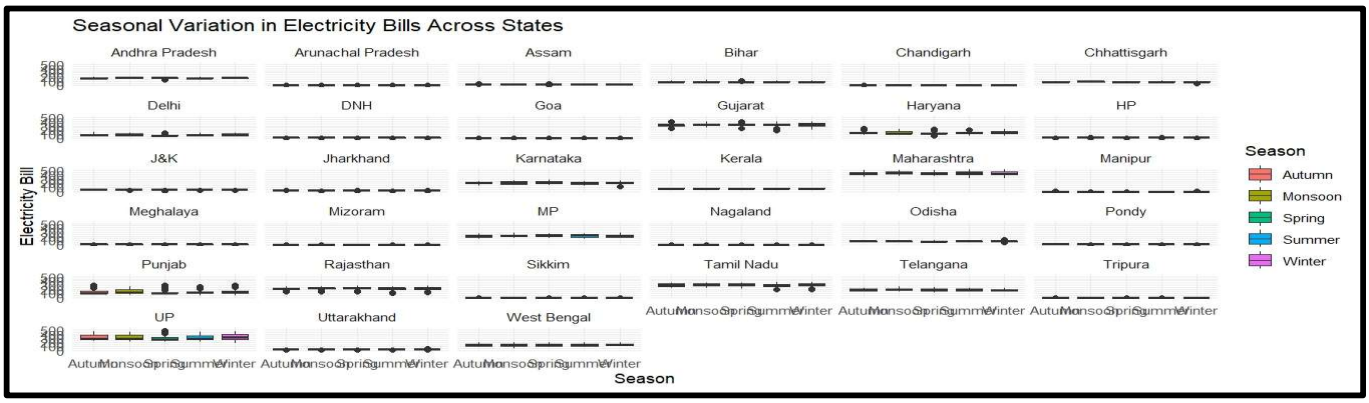
```
# Calculate metrics for Random Forest
rf_rmse <- RMSE(rf_pred, test_data$Total_Bill_Amount)
rf_r2 <- R2(rf_pred, test_data$Total_Bill_Amount)
rf_mae <- mae(test_data$Total_Bill_Amount, rf_pred)

cat("Linear Regression RMSE:", lm_rmse, " R²:", lm_r2, " MAE:", lm_mae, "\n")
cat("Random Forest RMSE:", rf_rmse, " R²:", rf_r2, " MAE:", rf_mae, "\n")

test_data$lm_pred <- lm_pred
test_data$rf_pred <- rf_pred

ggplot(test_data, aes(x = Total_Bill_Amount, y = lm_pred, color = Season)) +
  geom_point() +
  geom_abline(slope=1, intercept=0, linetype="dashed") +
  ggtitle("Linear Model: Actual vs Predicted by Season") +
  theme_minimal()

ggplot(test_data, aes(x = Total_Bill_Amount, y = rf_pred, color = Season)) +
  geom_point() +
  geom_abline(slope=1, intercept=0, linetype="dashed") +
  ggtitle("Random Forest: Actual vs Predicted by Season") +
  theme_minimal()
```

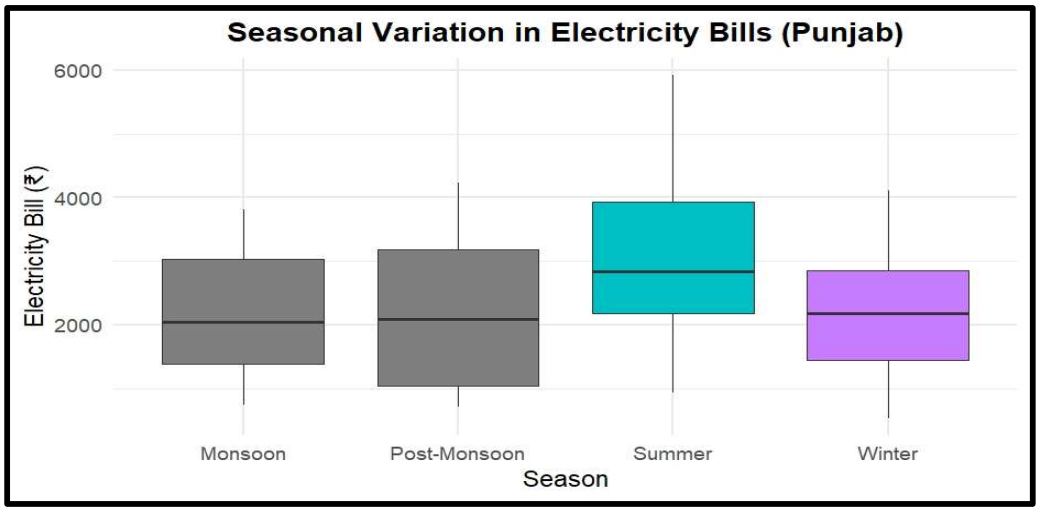
APPENDIX B – SCREENSHOTS








Seasonal Variation in Electricity Bills Across States
Andhra Pradesh
Arunachal Pradesh
Assam
Bihar
Chandigarh
Chhattisgarh
Delhi
DNH
Goa
Gujarat
Haryana
HP
J&K
Jharkhand
Karnataka
Kerala
Maharashtra
Manipur
Meghalaya
Mizoram
MP
Nagaland
Odisha
Pondy
Punjab
Rajasthan
Sikkim
Tamil Nadu
Telangana
Tripura
UP
Uttarakhand
West Bengal
Electricity Bill
Season
Autumn
Monsoon
Spring
Summer
Winter


Seasonal Variation in Electricity Bills (Punjab)
6000
4000
2000
Electricity Bill (₹)
Monsoon
Post-Monsoon
Summer
Winter
Season

REFERENCES:

- Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R* (2nd ed.). Springer.
- Wickham, H., & Grolemund, G. (2016). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324