Infosys Springboard
Internship

# Electricity demand and price forecasting

19 July 2024

by Praveena G

# CONTENTS :

# PROBLEM STATEMENT :

- With the growing need for accurate energy forecasting, this project explores the use of machine learning to predict both electricity demand and price.
- The focus lies in building predictive models that incorporate diverse data sources, including historical demand patterns, weather variables, and other relevant factors.
- The goal is to support smart grid operations and facilitate data-driven decision-making in the energy market.

# DATA INTRODUCTION :

**ENERGY DATASET**

1. Data shape : (35064,29)
2. Energy generation from various resources
3. Forecast for solar and wind generation
4. Actual price data
5. Actual load data

Granularity : Hourly

**WEATHER DATASET**

1. Data shape : (178396,17)
2. It contains data of Temperature
3. Wind speed
4. Weather conditions
5. Humidity

and so on ..

Granularity : Hourly

# EXPLORATORY DATA ANALYSIS :

Steps involved in EDA are :

1. Merging weather dataset and energy dataset.
2. Filling empty rows with forward fill (ffill) and backward fill (bfill) techniques.
3. Identify and handle outliers (OUTLIERS DETECTION AND REPLACEMENT).

**OUTCOME :**

After data preprocessing we get , CLEANED AND MERGED DATASET.

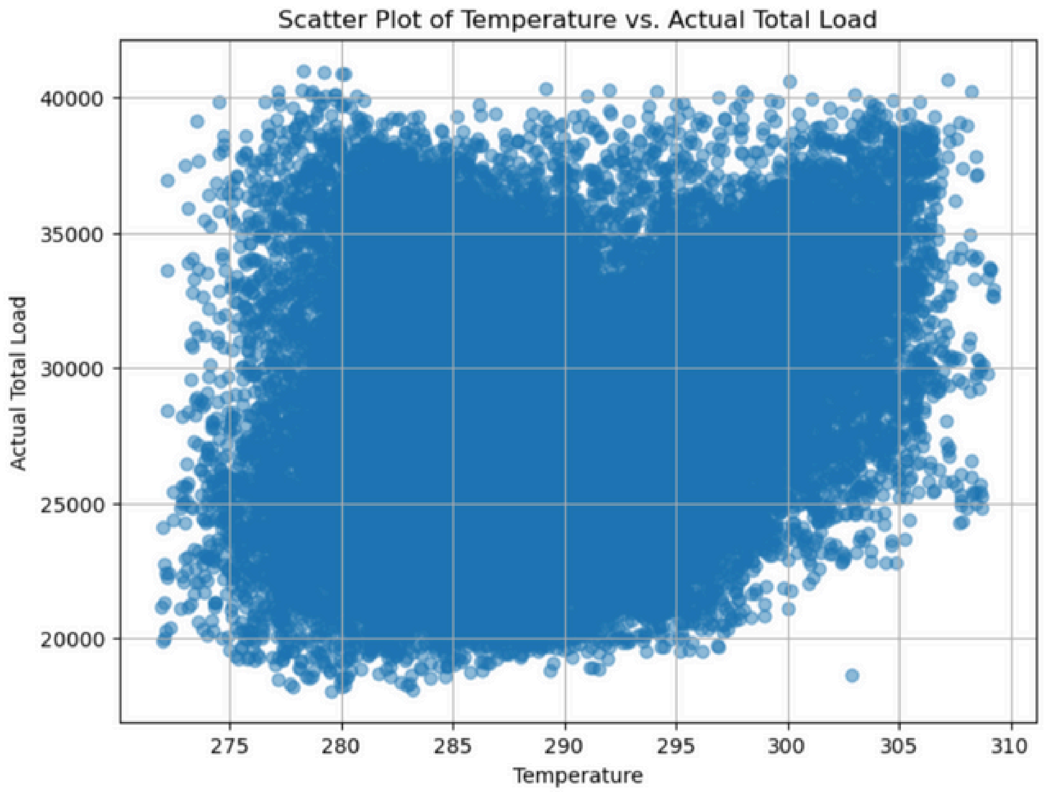
**SHAPE : (35064,36)**

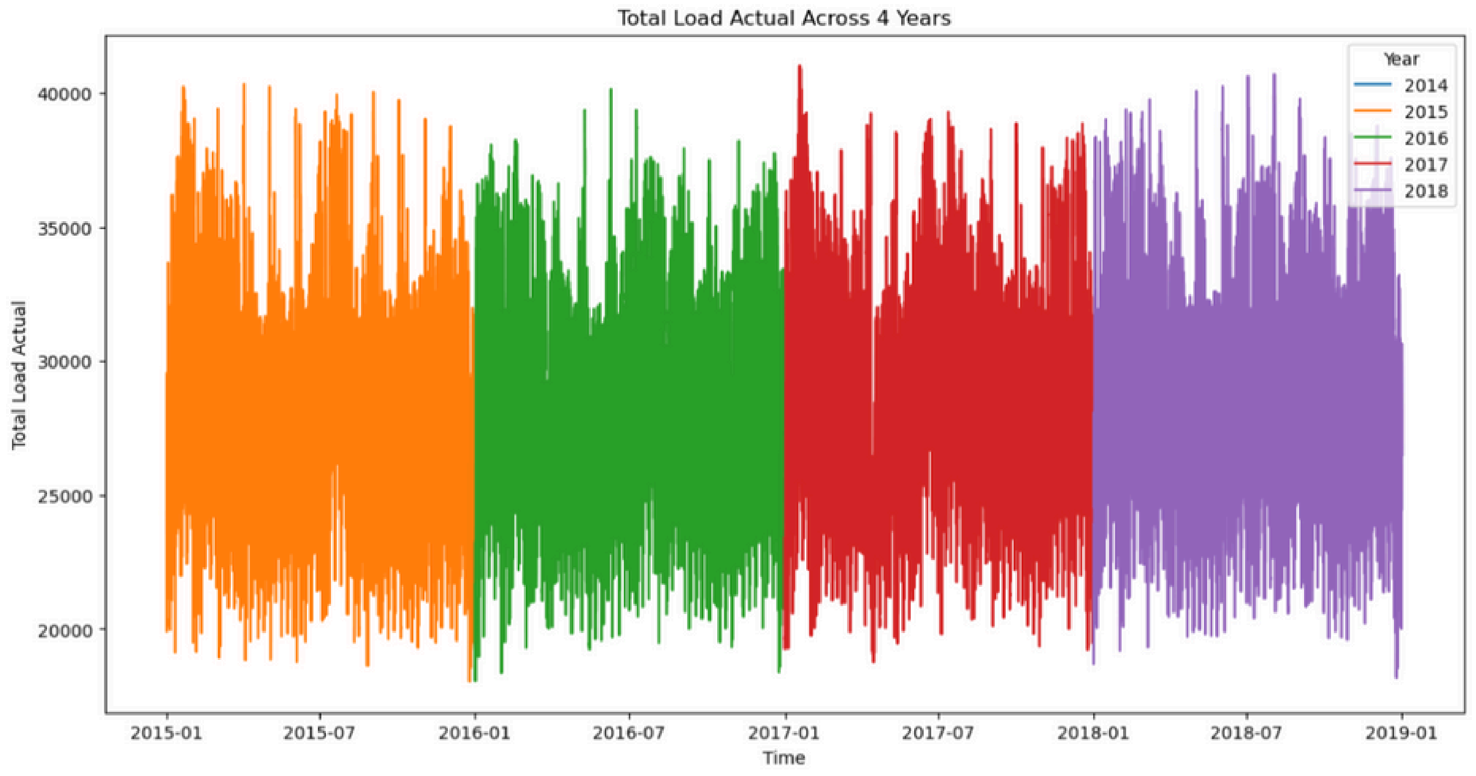
# FEATURE ENGINEERING :

Its all about creating new features , combining existing features, encoding categorical variables, interaction features , dimensionality reduction.

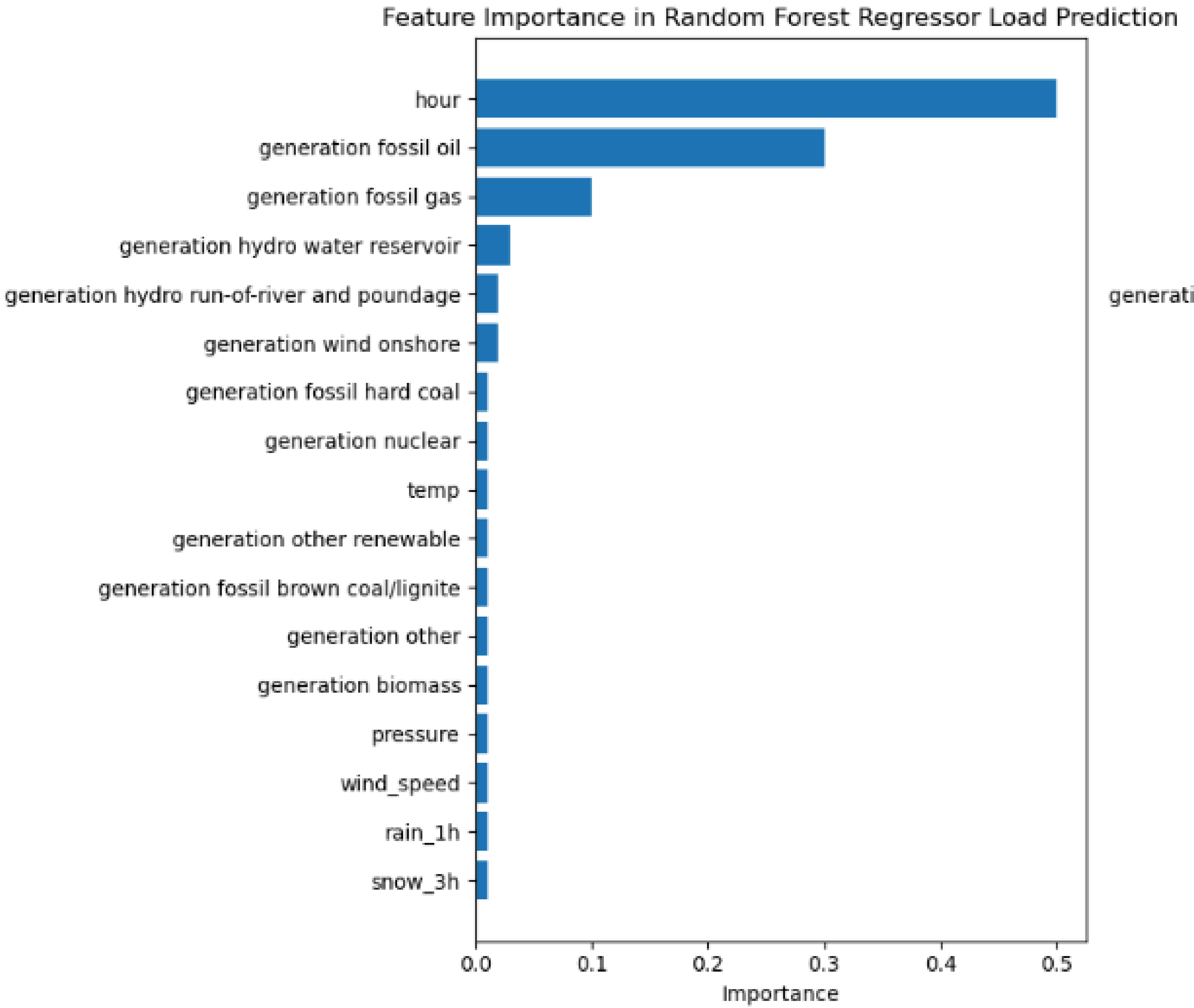
New features :

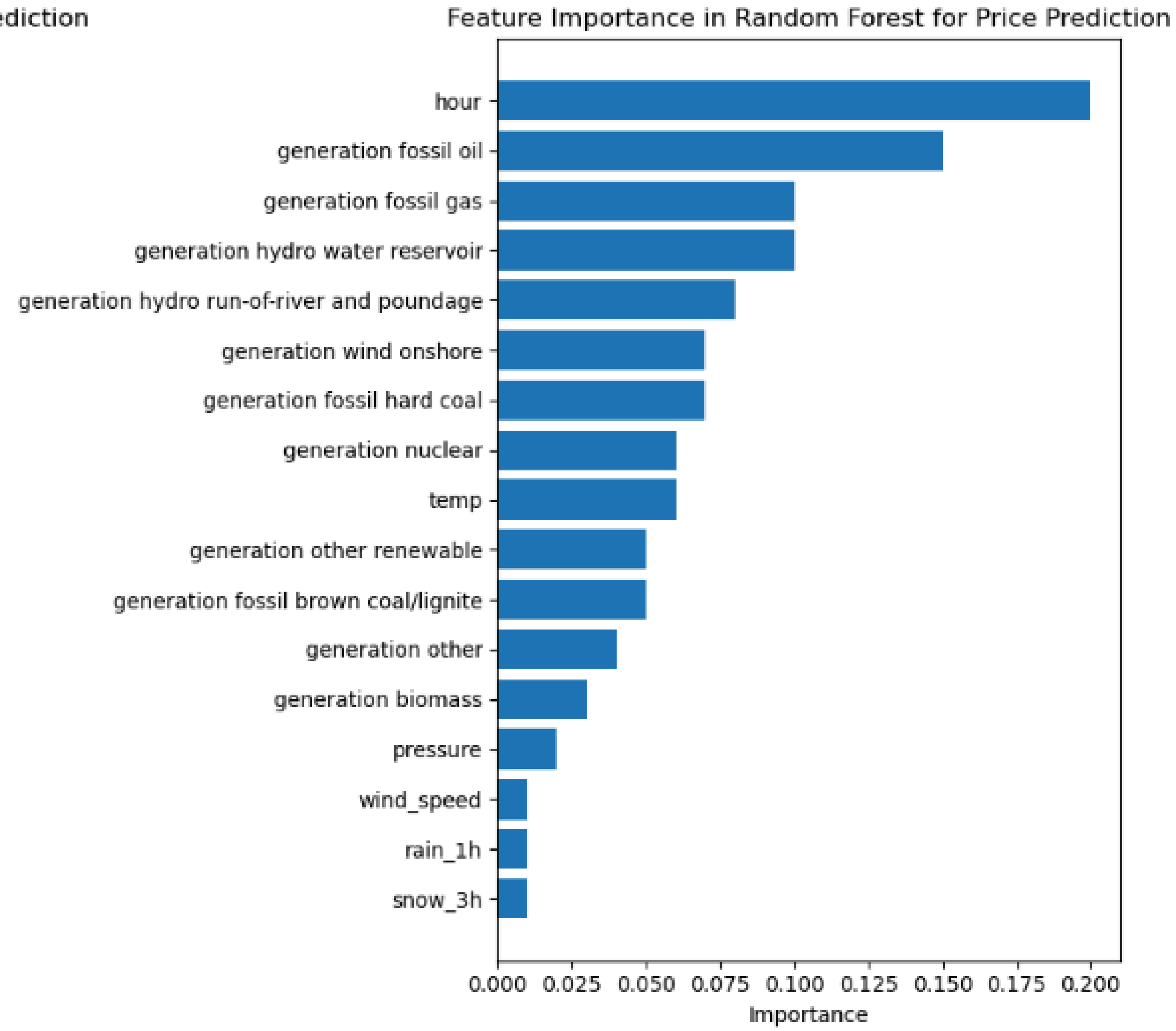
1. day_of_week : Extracted day of the week
2. month : specifies separate columns for each month
3. hour : demand and price will be categorized based on hour
4. temp_cubed : cubing the existing temp feature in the dataframe



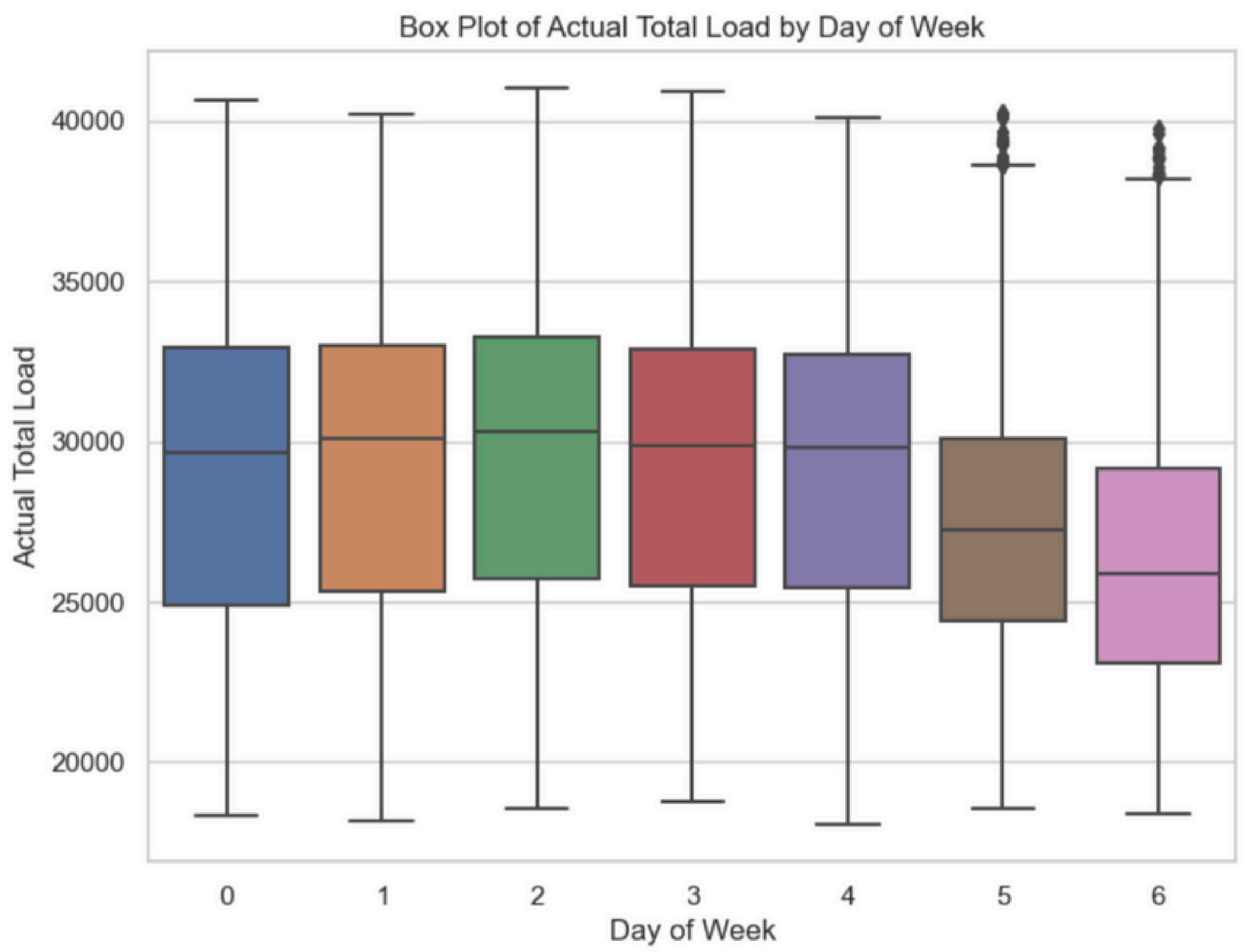
# FEATURE SELECTION:



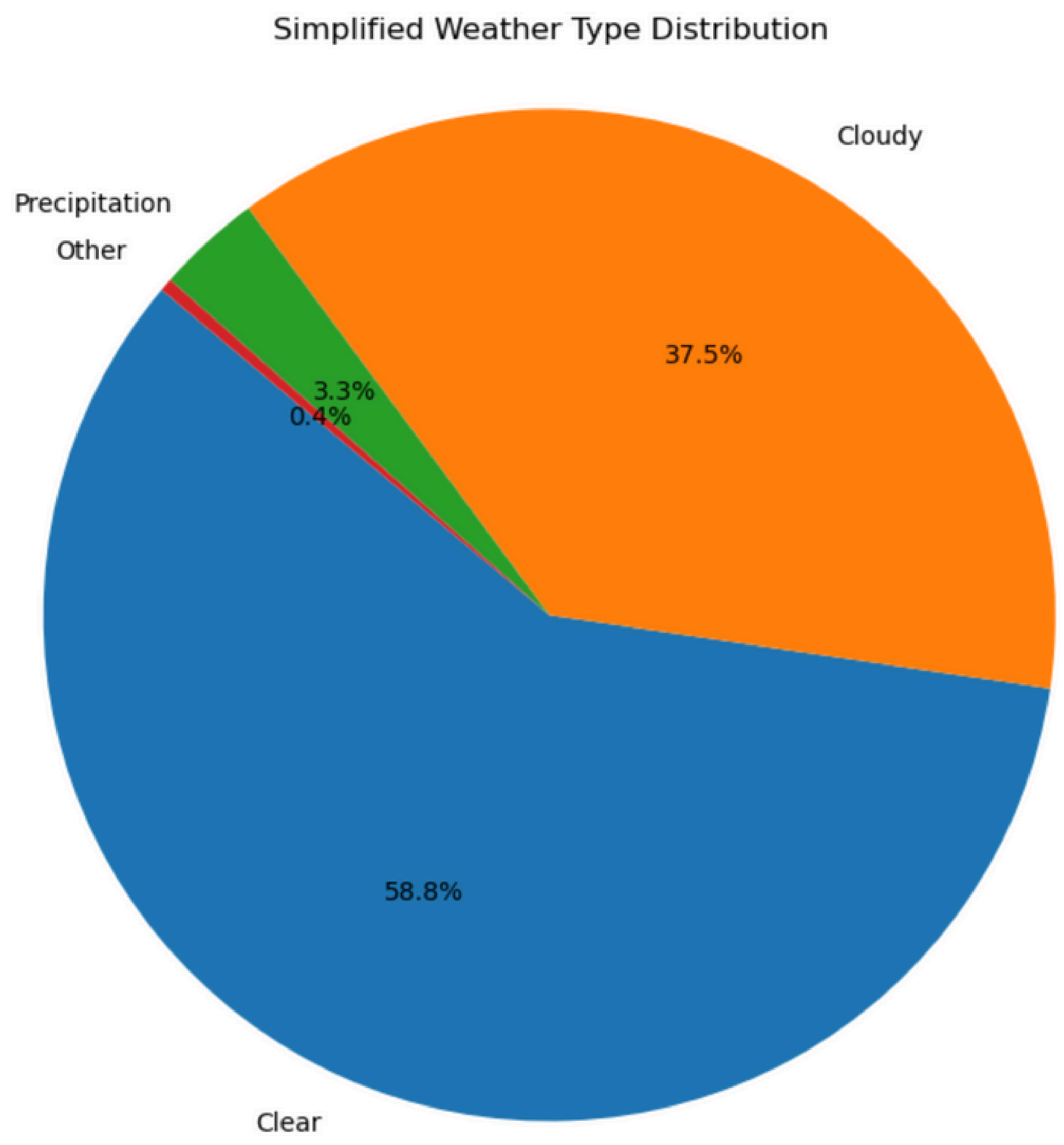




These are the important features has been selected for price actual and total load actual

# DATA VISUALIZATION:

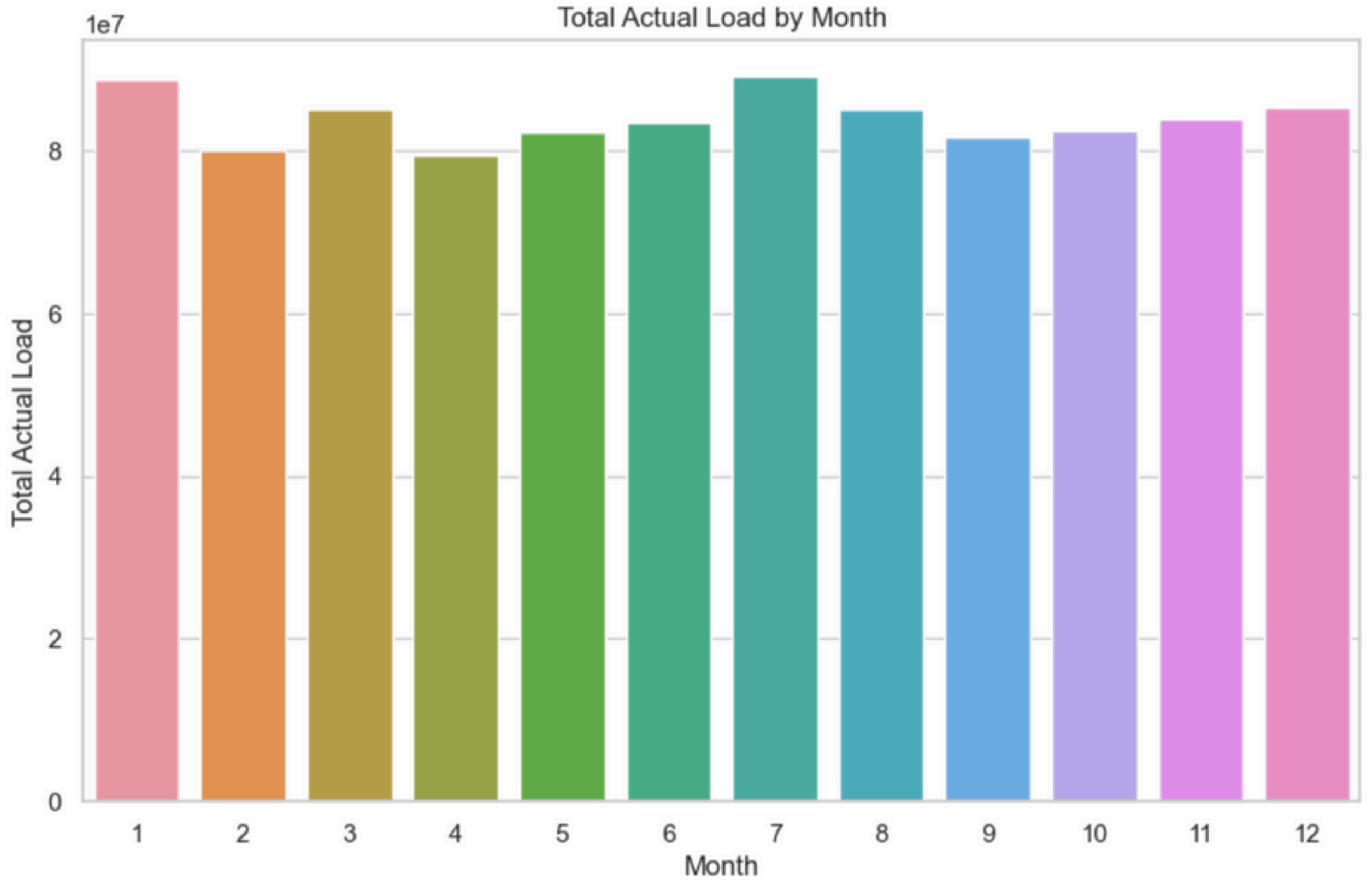



Displays the BOX PLOT of actual total load by days of week
(0-Mon , 1-Tue , 2-Wed , 3-Thurs , 4-Fri , 5-Sat , 6-Sun)
High load is seen in weekdays than weekends

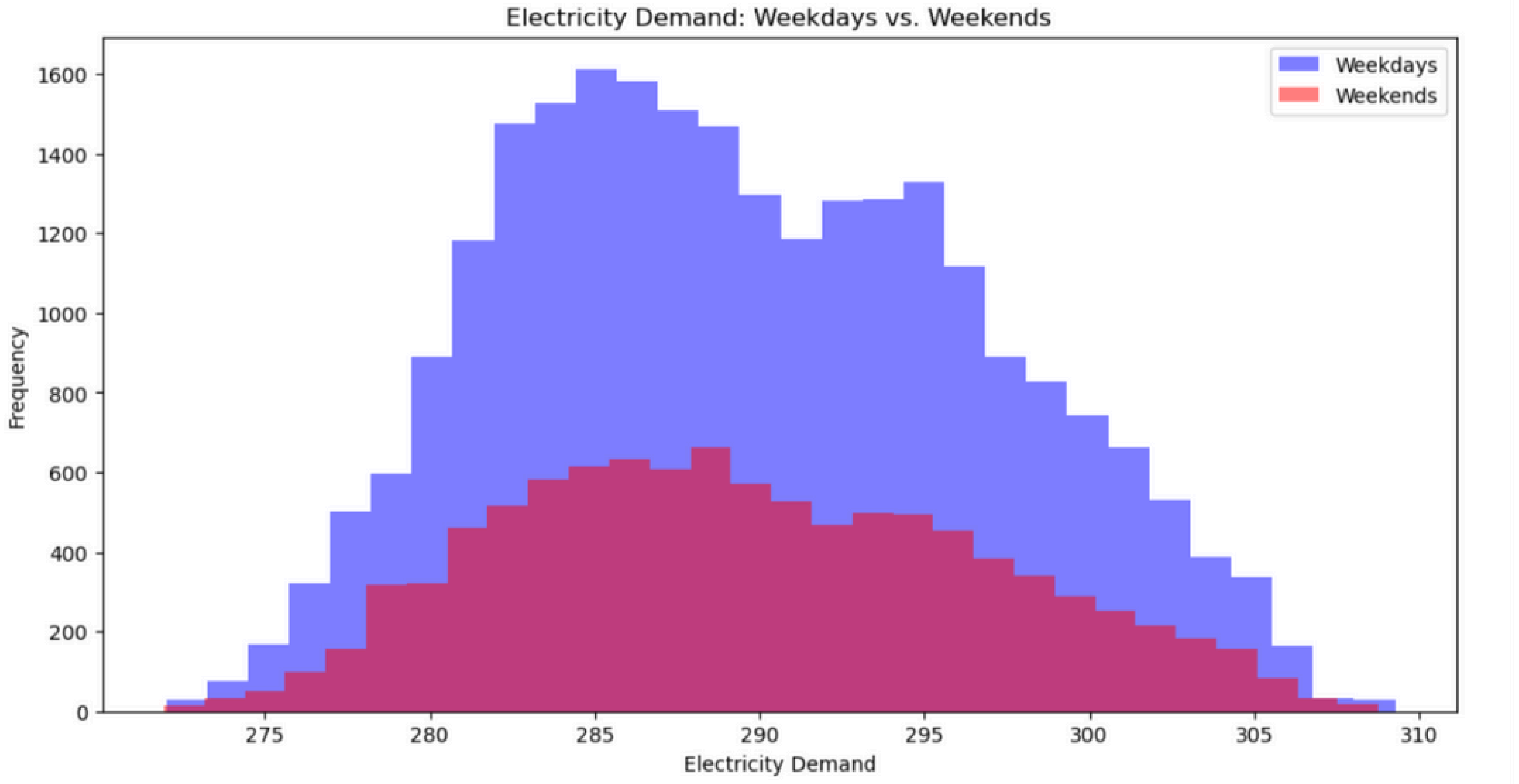



The pie chart shows that clear weather is the most frequent, accounting for 58.8% of the observations, followed by cloudy weather at 37.5%. Precipitation and other weather types are much less common, making up 3.3% and 0.4%, respectively.

# DATA VISUALIZATION: Contd...

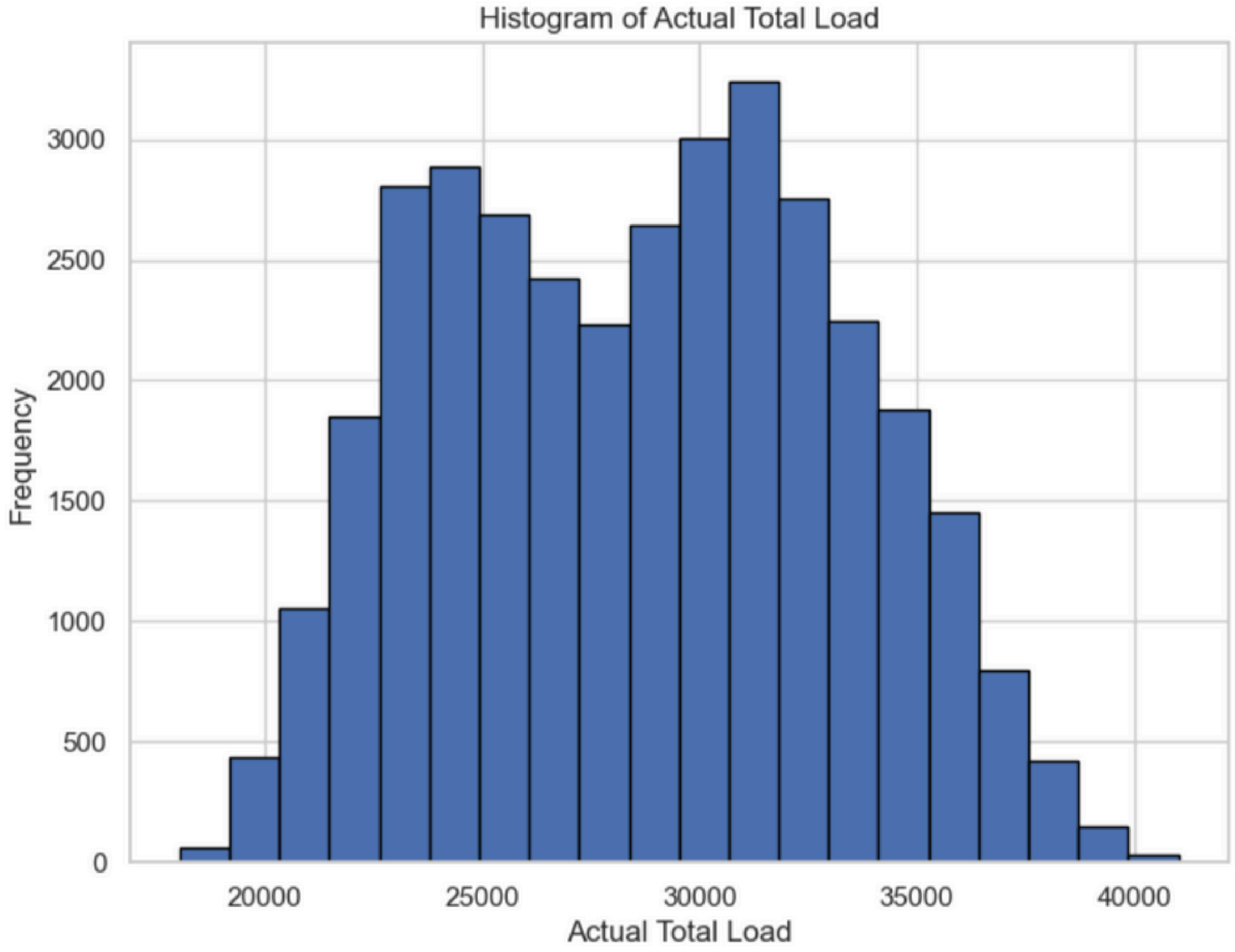



Displays the total load over month .
In January and July , the actual total load is same and high.

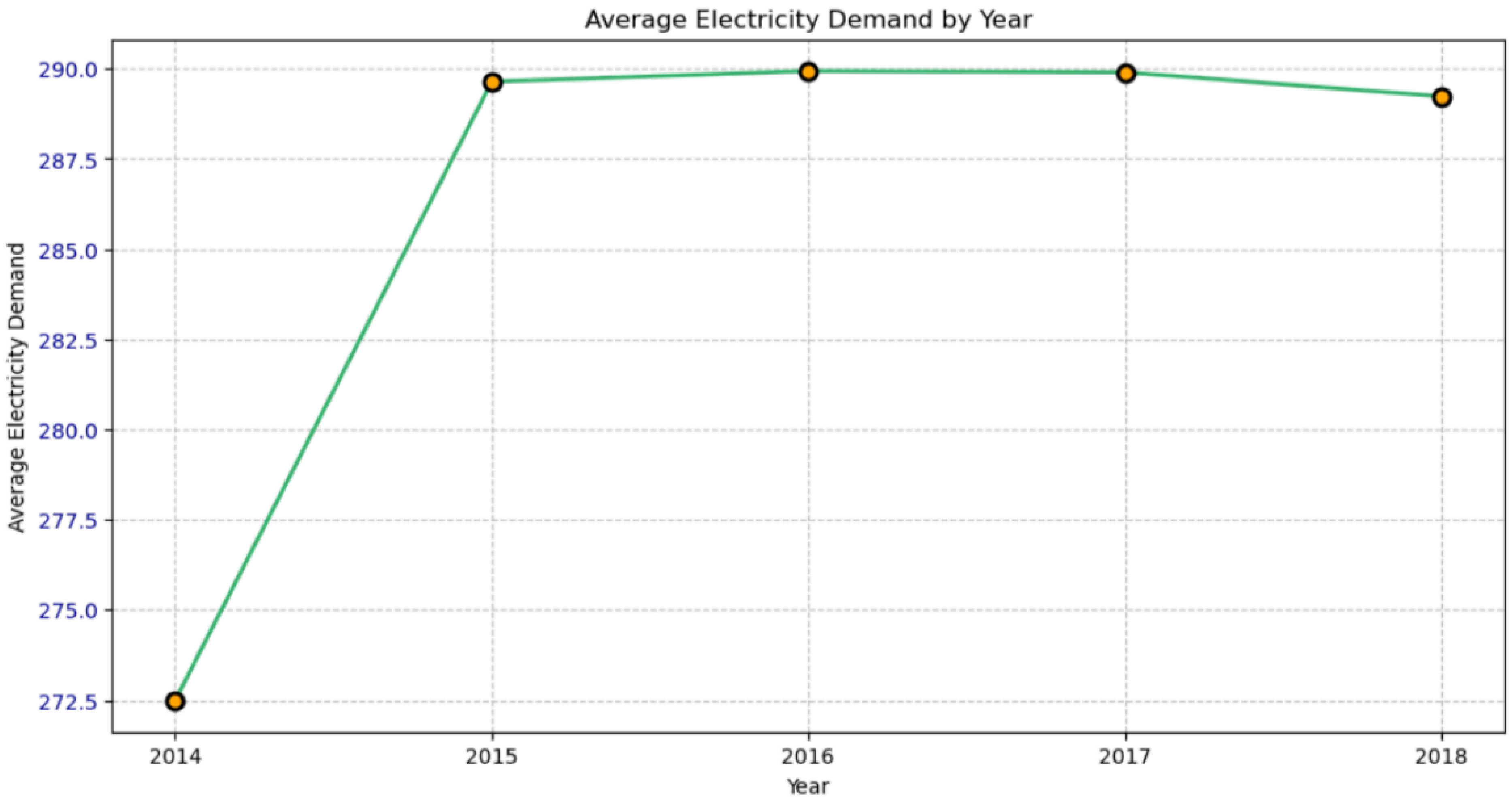



Electricity demand in weekdays is higher than weekends

# DATA VISUALIZATION: Contd...



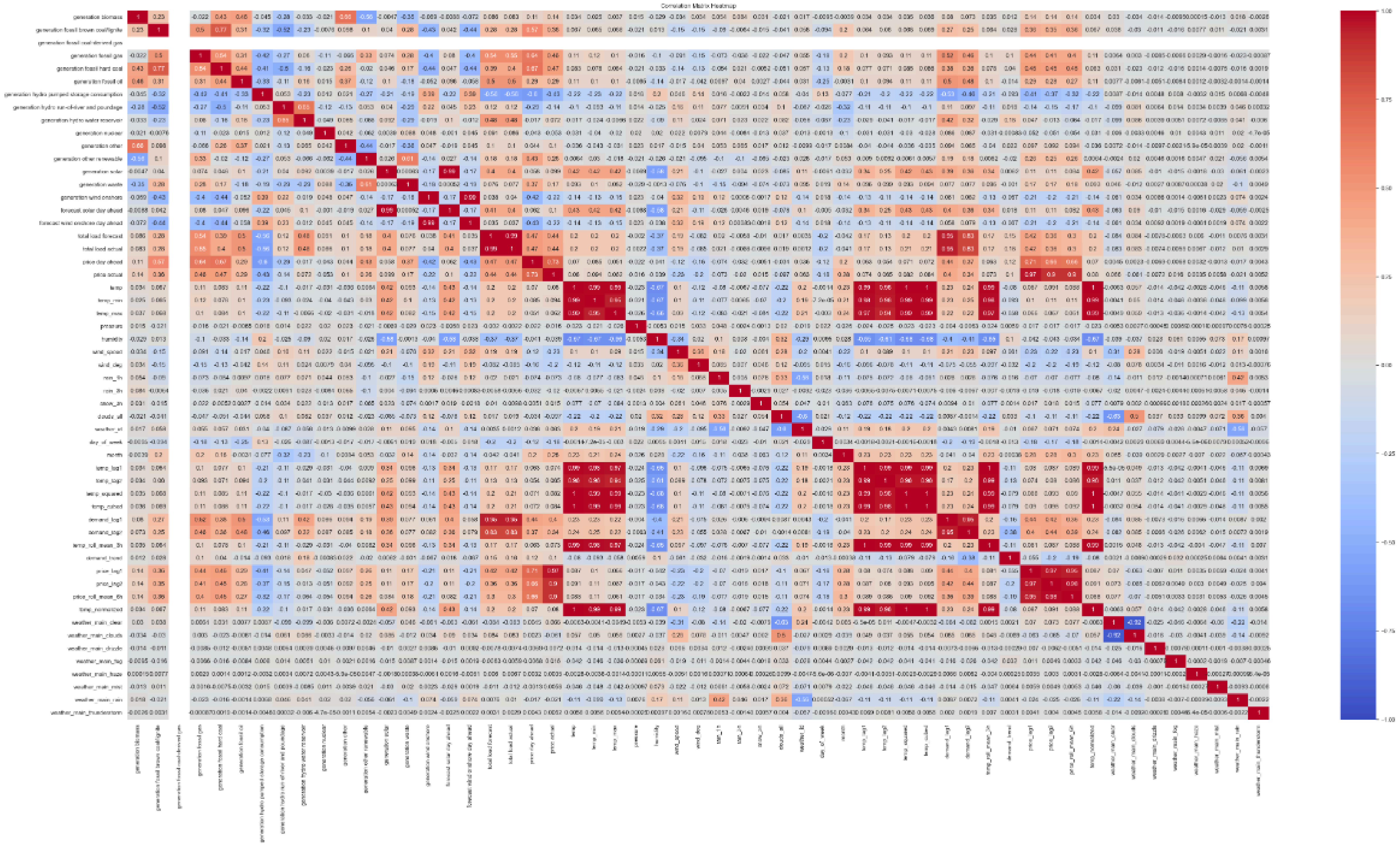

This histogram shows the frequency distribution of the actual total load values, indicating that most observations are concentrated around the 25,000 to 35,000 range. The distribution appears roughly symmetrical with a peak frequency around 30,000.




It depicts the lineplot of average electricity demand by year. In 2014 , demand is low and drastically increases in the upcoming year. In 2016 and 2017 , the electricity demand is higher.

# DATA VISUALIZATION: Contd...

- **CORRELATION MATRIX** represents the correlation coefficients between pairs of variables in your dataset.
- Each cell in the heatmap shows the correlation between two variables, with values ranging from -1 to 1.
- 1 indicates a perfect positive correlation.
- -1 indicates a perfect negative correlation.
- 0 indicates no correlation.
- The colors range from blue to red.
- Red signifies a strong positive correlation.
- Blue signifies a strong negative correlation.
- Gray signifies no correlation.



# MODEL SELECTION:

**LINEAR REGRESSION:**
Price Actual:
MAE: 1.8951413538247457
RMSE: 2.7729891452599817
MAPE: 3.57%
$R^2$: 0.9621335153916698
Adjusted $R^2$: 0.9620414881953379

Total Load Actual:
MAE: 629.9630720265048
RMSE: 936.964649851538
MAPE: 2.22%
$R^2$: 0.9579087623556332
Adjusted $R^2$: 0.9578004348924364

**GRADIENT BOOSTING:**
Price Actual:
MAE: 1.7532585221005914
RMSE: 2.518151730857104
$R^2$: 0.9687735625577445
Adj. $R^2$: 0.9685
MAPE: 3.32%

Total Load Actual :
MAE: 264.2988105895638
RMSE: 371.6877611373267
$R^2$: 0.9933762824354007
Adj. $R^2$: 0.9933,
MAPE: 0.93%

**GRADIENT BOOSTING with HP:**
Price Actual:
MAE: 6.869667081472161
RMSE: 8.617007885580326
$R^2$: 0.6343449041232151
Adj. $R^2$: 0.6337180668159978
MAPE: 13.744072213654668%

Total Load Actual:
MAE: 319.3294680619043
RMSE: 449.7800898294249
$R^2$: 0.990300580619867
Adj. $R^2$: 0.9902839530437868
MAPE: 1.1203310870844125%

# MODEL SELECTION: Contd...

**RANDOM FOREST**:
Price Actual Model :
MAE: 2.92897037512061
RMSE: 4.005268442963803
MAPE: 5.56%
$R^2$: 0.921000942989641
Adjusted $R^2$: 0.9209896751167255

Total Load Actual Model :
MAE: 385.3031758327563
RMSE: 556.1501386350349
MAPE: 1.34%
$R^2$: 0.9851704018225922
Adjusted $R^2$: 0.985168286632437

**RANDOM FOREST with HP:**
Price Actual:
MAE: 4.032533988046175
RMSE: 5.5366502304004594
$R^2$: 0.8490431494519777
Adj. $R^2$: 0.8487843662796096
MAPE: 8.076816618768614%

Total Load Actual:
MAE: 265.8557869195027
RMSE: 402.4526857395628
$R^2$: 0.9922344013009821
Adj. $R^2$: 0.9922210888460694
MAPE: 0.9328605479192709%

**LSTM:**
Price Actual:
RMSE: 2.5176233217052197
MAE: 1.799009959591994
$R^2$: 0.9687866662987997
Adjusted $R^2$: 0.9686973833076635
MAPE: 3.59%

Total Load Actual :
RMSE: 433.471487805356
MAE: 302.102087195208
$R^2$: 0.990991212211731
Adjusted $R^2$: 0.9909654433679431
MAPE: 1.07%

# MODEL COMPARISON TABLE:

| Model | Price Actual MAE | Price Actual RMSE | Price Actual MAPE | Price Actual $R^2$ | Price Actual Adj. $R^2$ |
|---|---|---|---|---|---|
| Linear Regression | 1.895 | 2.773 | 3.57 | 0.962 | 0.962 |
| Gradient Boosting | 1.753 | 2.518 | 3.32 | 0.969 | 0.968 |
| Random Forest | 2.929 | 4.005 | 5.56 | 0.921 | 0.921 |
| Gradient Boosting (Tuned) | 2.028 | 2.893 | 3.81 | 0.959 | 0.633 |
| Random Forest (Tuned) | 1.708 | 5.53 | 3.27 | 0.97 | 0.97 |
| LSTM | 1.737 | 2.57 | 3.36 | 0.97 | 0.97 |

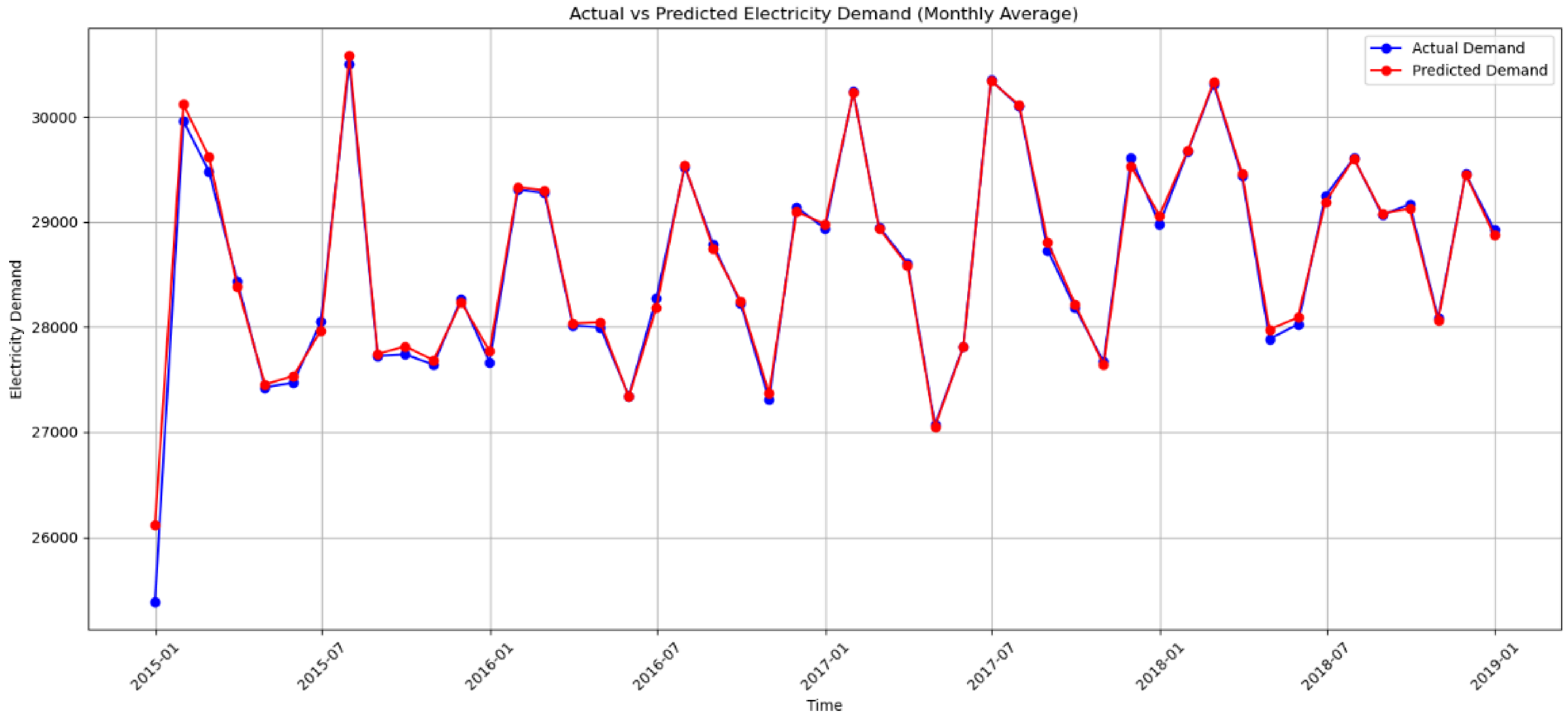
| Model | Total Load Actual MAE | Total Load Actual RMSE | Total Load Actual MAPE | Total Load Actual $R^2$ | Total Load Actual Adj. $R^2$ |
|---|---|---|---|---|---|
| Linear Regression | 629.963 | 936.965 | 2.22 | 0.958 | 0.958 |
| Gradient Boosting | 264.299 | 371.688 | 3.12 | 0.993 | 0.968 |
| Random Forest | 385.303 | 556.15 | 1.34 | 0.985 | 0.985 |
| Gradient Boosting (Tuned) | 3207.583 | 3956.007 | 8.58 | 0.25 | 0.99 |
| Random Forest (Tuned) | 284.745 | 402.43 | 1.0 | 0.991 | 0.991 |
| LSTM | 314.275 | 433.47 | 1.11 | 0.991 | 0.991 |

**FINAL MODEL SELECTION :**

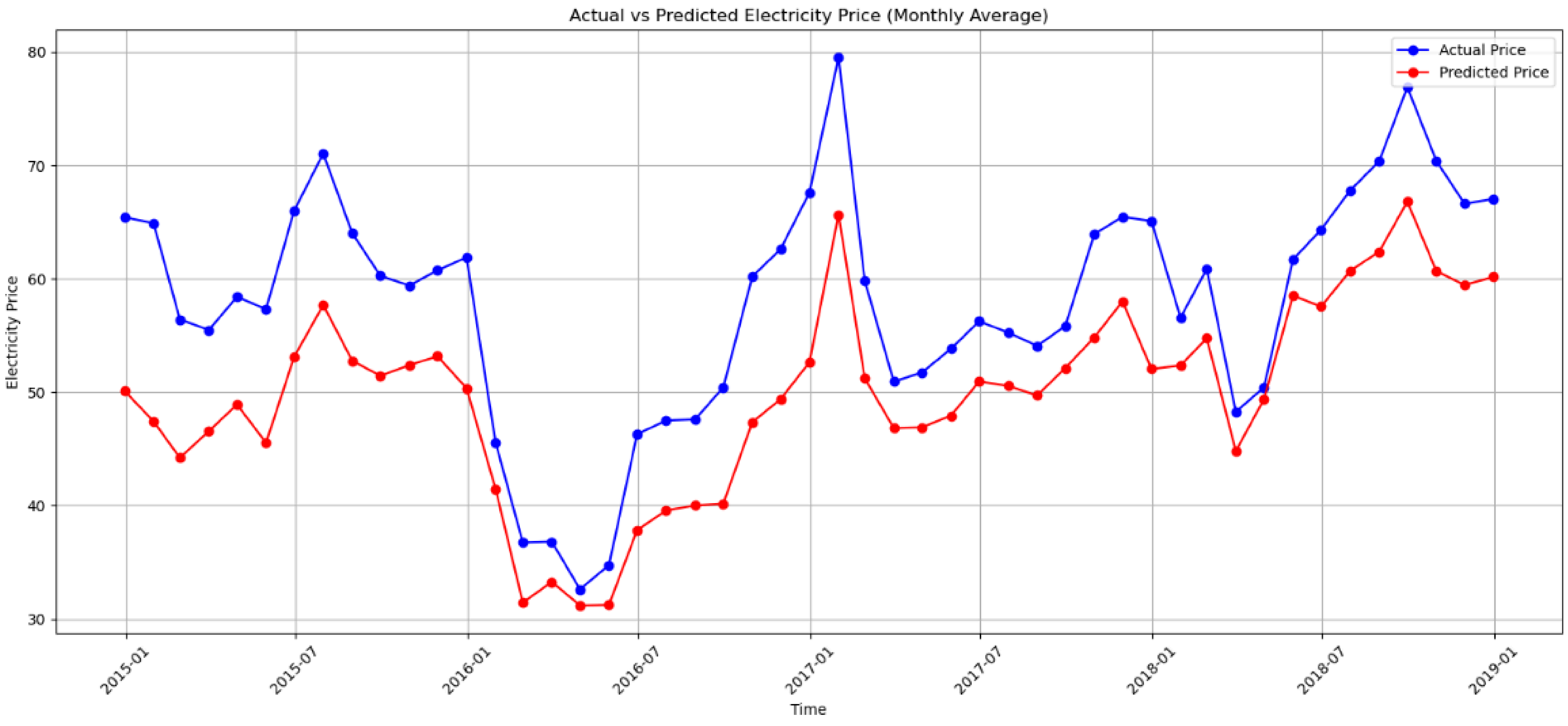
Random forest based on hyperparameter tuning for both **price and load**

# CONCLUSION:

- The actual and predicted electricity prices from  the below graph shows that while the predicted prices generally follow the trend of the actual prices, there are periods where they significantly diverge, especially around 2016 and early 2017.
- The actual and predicted electricity demand exhibit a much closer alignment, indicating that the model predicts demand more accurately than it does prices.
- Among the various models tested , RANDOM FOREST based on HYPERPARAMETER TUNING is the best model for both price and load.



# CONCLUSION: Contd...




THANK
YOU