# STEEL INDUSTRY PROJECT

*Project report submitted*

*In partial fulfillment of the requirement for the degree of*

# Bachelor of Technology

**In**

# Computer Science And Engineering By

**Sahil (2001010054) Under the guidance of**

# Ms. Pallavi Pandey (Assistant Professor)



**Department of**

# School of Engineering and Technology

**K. R. Mangalam University, Gurugram – 122003 July – 2022**

# DECLARATION

We declare that this written submission represents our ideas in our own words and where other’s ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all the principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will because for disciplinary action by the Institute and canal so evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed. We further declare that if there is any violation of the intellectual property right or copyright, my supervisor and university should not be held responsible for the same.

Sahil 2001010054 (Signature)

Place: K.R. Mangalam University

# ACKNOWLEDGEMENT

## “Enthusiasm is the feet of all progress, with it there is accomplishment and Without it there are only slits alibis.”

Acknowledgment is not a ritual but is certainly an important thing for the successful completion of the project. At the time when we were made to know about the project, it was really tough to proceed further as we were to develop the same on a platform which was new to us. More so, the coding part seemed tricky that it seemed to be impossible for us to complete the work within the given duration.

We really feel indebted in acknowledging the organizational support and encouragement received from the university.

The task of developing this system would not have been possible without the constant help of our faculty members and friends. We take this opportunity to express our profound sense of gratitude and respect to those who helped us throughout the duration of this project.

We express our gratitude to our supervisor Ms. Pallavi Pandey for giving his valuable time and guidance to us.

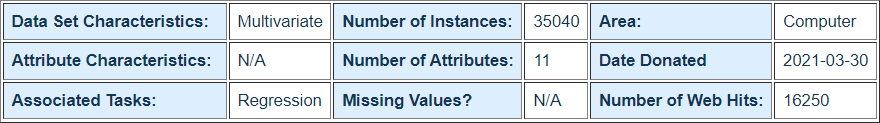
Place: - K.R. Mangalam University Sahil

# ABSTRACT

The South Korean Company produces several types of coils, steel plates, and iron plates. The information on electricity consumption is held in a cloud-based system.

The information on the energy consumption of the industry is stored on the website of the KOREAN ELECTRIC POWER CORPORATION (pccs.kepco.go.kr), we are going to predict the usage in kilowatt per hour which is also our target that is depending on various factors mentioned as features in dataset below.

## *Source*: UCI ML Repository (from where data is taken)



We have used python3 and jupyter notebook for our project. We have used different python libraries for different purposes like NumPy and Pandas for reading and analyzing the data, Matplotlib and Seaborn for plotting and visualizing the data in a more understandable manner, Sweetviz for univariate analysis and Scikitlearn for applying various

ML algorithms in order to check their accuracies for predicting the most accurate model for our given data.

***Key Words*: Python3, Jupyter Notebook, NumPy, Pandas, Matplotlib, Seaborn, Sweetviz, Scikitlearn.**

# INTRODUCTION

The data we are using for analysis is of a South Korean Company which produces several types of coils, steel plates, and iron plates on the daily bases. The data mainly consists of information regarding the electricity consumption of the company used by the machinery for production and other processes. Basically, this data is stored in a cloud-based system. The information on the energy consumption of the industry is stored on the website of the KOREAN ELECTRIC POWER CORPORATION (pccs.kepco.go.kr). The data is openly available for public use. The data is in the form of a CSV file containing 35040 rows and 11 columns. The feature description is given below: -

1. Date: it contains the date and time at which data is recorded
2. Usage\_kWh: Industry's power consumption continuous data. (target column)
3. Lagging\_Current\_Reactive.Power\_kVarh: There is a phase difference between voltage and current where load current lags the supply voltage.
4. Leading\_Current\_Reactive\_Power\_KVarh: The load current leads the supply voltage by certain phase angle.
5. CO2 (tCO2): Tonnes of Carbon Dioxide in ppm
6. Lagging\_Current\_Power\_Factor
7. Leading\_Current\_Power\_Factor
8. NSM: Number of Seconds from midnight
9. Week Status: Categorical (Weekend (0) or a Weekday (1))
10. Day\_of\_week: Categorical Monday to Sunday
11. Load\_Type: Categorical Light Load, Medium Load, Maximum Load

The target column is the Usage\_kWh, as it represents the most significant and appropriate data for the analysis and prediction of electricity consumption by the Steel Industry.

**Numpy** is the core library for scientific computing in Python. It provides a high-performance multidimensional array object, and tools for working with these arrays.

**Pandas** is a fast, powerful, flexible and easy to use open-source data analysis and manipulation tool, built on top of the Python programming language.

**Matplotlib** is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack.

**Seaborn** is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

**Scikit-learn** is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

**Sweetviz** is an open-source Python library that generates beautiful, high-density visualizations to kickstart EDA (Exploratory Data Analysis) with just two lines of code. Output is a fully self- contained HTML application.

Using the given data and applying Machine Learning models, we are going to predict the electricity consumption by the industry in the future.

# LITERATURE REVIEW

## Steps of data analysis: -

* **Business Understanding –** this phase consists of a very precise specification of the problem together with methods of evaluating the achievement of the goal.

Problem➜ Model ➜ Solution

Our goal is to create a model of the problem, which we use in turn to find a solution. The model should be precise enough to make the solution meaningful, otherwise, we will make too many assumptions and approximations which will make the solution far from real and meaningless. It should be added that the model is not the same as the problem, but only a representation of reality. Therefore, the model should be precise enough to find a solution and should be general enough not to include irrelevant variables. A correct solution to the problem involves a correct understanding of the problem. Real business problems are rarely obvious. Therefore, an iterative approach to the project with multiple repetitions of individual phases is necessary for the solution to be considered acceptable.

In our case the data is already meaningful to some extent as it is already stored on the website by the kepco in a useful and understandable way.

* **Data Mining -** Data mining is the process of sorting through large data sets to identify patterns and relationships that can help solve business problems through data analysis. Data mining techniques and tools enable enterprises to predict future trends and make more-informed business decisions. Data mining is a key part of data analytics overall and one of the core disciplines in data science, which uses advanced analytics techniques to find useful information in data sets.
* **Data Cleaning -** Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process, so you know you are doing it the right way every time.
* **Exploratory Data Analysis -** Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate. Originally developed by American mathematician John Tukey in the 1970s, EDA techniques continue to be a widely used method in the data discovery process today.

* **Data preprocessing -** Data preprocessing is a step in the data mining and data analysis process that takes raw data and transforms it into a format that can be understood and analyzed by computers and machine learning. Raw, real-world data in the form of text, images, video, etc., is messy. Not only may it contain errors and inconsistencies, but it is often incomplete, and doesn’t have a regular, uniform design. Machines like to process nice and tidy information – they read data as 1s and 0s. So, calculating structured data, like whole numbers and percentages is easy. However, unstructured data, in the form of text and images must first be cleaned and formatted before analysis.
* **Feature Engineering and predictive model -** Feature engineering is a crucial part of predictive modeling success. The goal of predictive modeling is to identify how likely a subject (such as a customer or prospect) is to perform a desired action. A predictive model is a combination of attributes (also known as features) that predicts the likelihood of an outcome.

Feature engineering is the process of refining raw data and identifying the most predictive attributes to use in modeling. When applied to marketing use cases, feature engineering supports the creation of predictive models that produce actionable insights around high-value customers, such as propensity to churn or acceptance rates for a product offer.

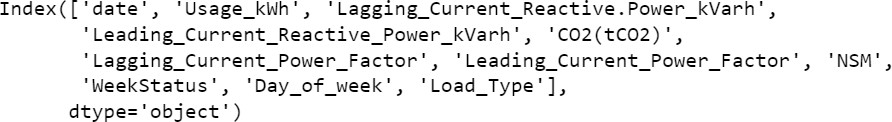
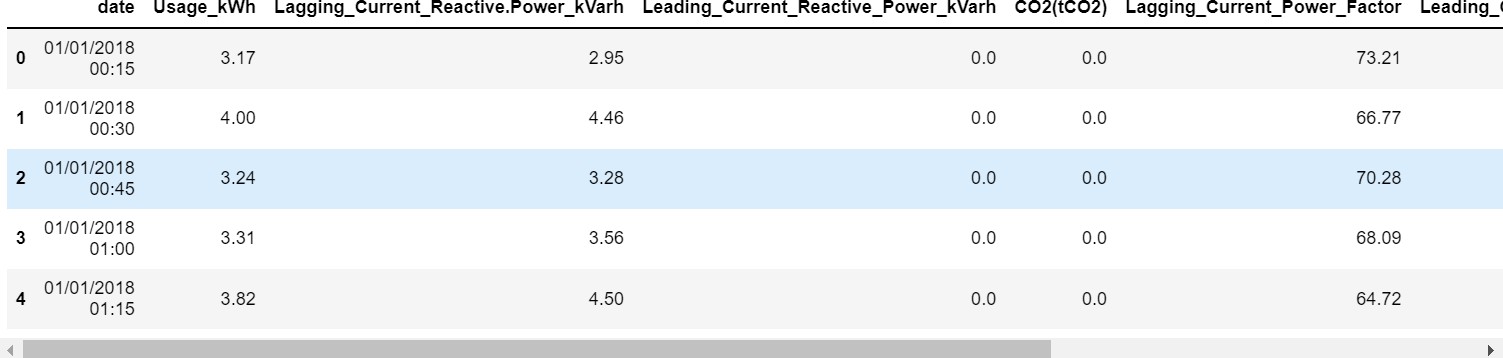
We have used the 5 ML model for predicting the most accurate data: -

* 1. Linear Regression
  2. K Nearest Neighbors Regression
  3. Decision Tree Regression
  4. Random Forest Regression
  5. Support Vector Machine Regression

# PROBLEM FORMULATION AND OBJECTIVES

**Problem Statement:** We need to predict energy usage in small scale coal, steel and manufacturing industries present in South Korea on basis of various input data columns.

Sample Dataset:

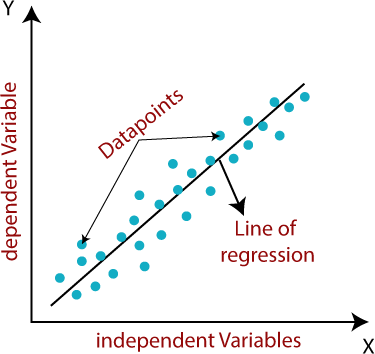


**Objective:** Main objective is to conserve energy as well as optimal utilization of natural resources. Electricity is made by non-renewable resources which are depleting assets.

# METHODOLOGY OF THE PROJECT

1. **Linear Regression** - Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc. Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



Mathematically, we can represent a linear regression as: y=a0+a1x+ε

**Here,**

Y = Dependent Variable (Target Variable)

X = Independent Variable (predictor Variable)

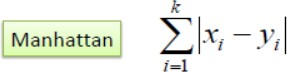
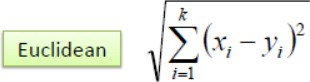
a0 = intercept of the line (Gives an additional degree of freedom)

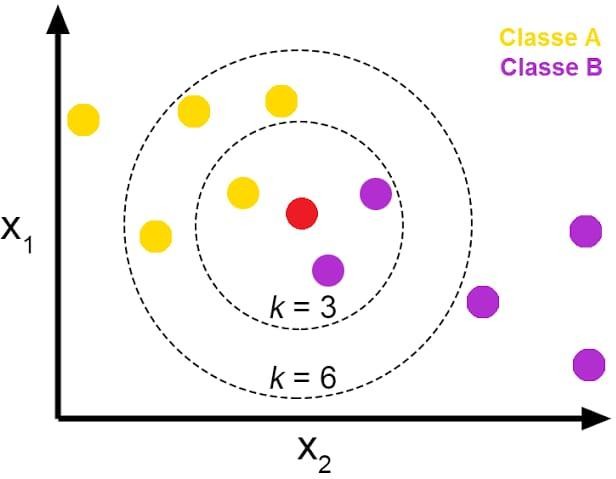
a1 = Linear regression coefficient (scale factor to each input value).

ε = random error

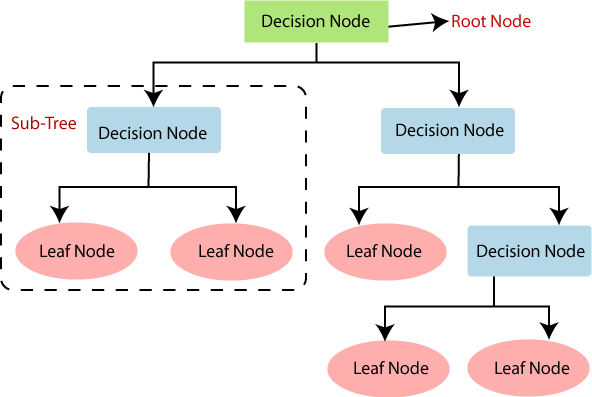
The values for x and y variables are training datasets for Linear Regression model representation.

1. **K-Nearest Neighbour Regression** - KNN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same *neighborhood*. The size of the neighborhood needs to be set by the analyst or can be chosen using cross- validation (we will see this later) to select the size that minimizes the mean-squared error. While the method is quite appealing, it quickly becomes impractical when the dimension increases, i.e., when there are many independent variables. Mathematical formulas for distance in KNN and graphical representation: -





1. **Decision Tree Regression** - Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, Internal nodes represent the features of a dataset, branches represent the decision rules, and each leaf node represents the outcome. There are two nodes, which are the Decision and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed based on features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. In order to build a tree, we use the CART (Classification and Regression Tree) algorithm.A decision tree simply asks a question and based on the answer (Yes/No), it further splits the tree into subtrees.

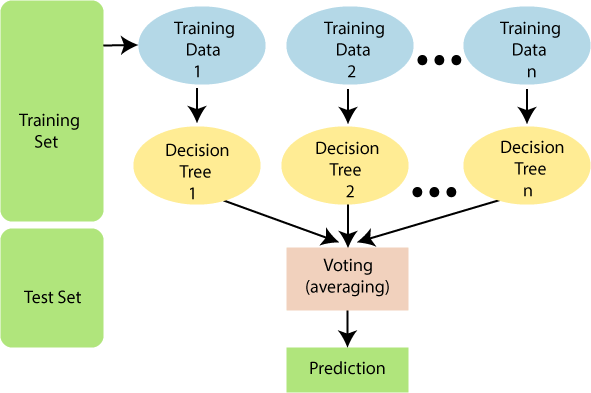


1. **Random Forest Regression** - It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

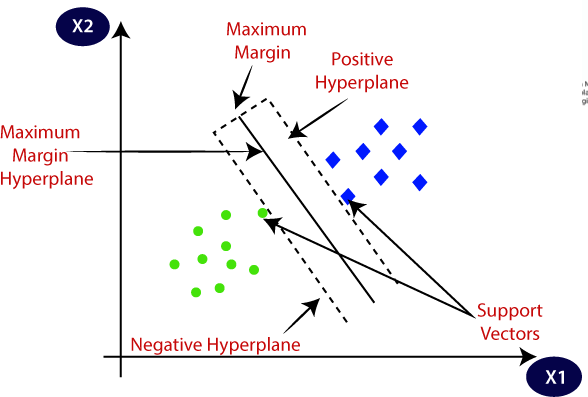
The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

The diagram below explains the working of the Random Forest algorithm:



1. **Support Vector Machine Regression** - The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence algorithm is termed as Support Vector Machine. Consider the diagram below in which there are two different categories that are classified using a decision boundary or hyperplane:



# IMPLEMENTATION AND RESULT

Business Understanding:

As we already discussed above, it is a dataset which contains the power consumption information of small-scale iron, coal and manufacturing industries in South Korea, hence we need to predict the usage of electricity in kilo watt per hour for future reference and to predict electricity consumption, if any new company starts its manufacturing unit.

Data Mining:

First of all, the necessary libraries are imported which includes numpy, pandas, matplotlib, seaborn, scikit learn packages and sweetviz.

Data in the form of comma separated values has been imported into jupyter notebook with the help of pandas command,

Data information is also mentioned below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| # |  | Column | Non-Null Count | Dtype |
| 0 |  | date | 35040 non-null | object |
| 1 |  | Usage\_kWh | 35040 non-null | float64 |
| 2 |  | Lagging\_Current\_Reactive.Power\_kVarh | 35040 non-null | float64 |
| 3 |  | Leading\_Current\_Reactive\_Power\_kVarh | 35040 non-null | float64 |
| 4 |  | CO2(tCO2) | 35040 non-null | float64 |
| 5 |  | Lagging\_Current\_Power\_Factor | 35040 non-null | float64 |
| 6 |  | Leading\_Current\_Power\_Factor | 35040 non-null | float64 |
| 7 |  | NSM | 35040 non-null | int64 |
| 8 |  | WeekStatus | 35040 non-null | object |
| 9 |  | Day\_of\_week | 35040 non-null | object |

10 Load\_Type 35040 non-null object

Data Cleaning:

Count of null values is checked in this part; hence no null values find in above dataset, further we will remove outliers which is also the part of data cleaning.

Exploratory Data Analysis:

Describe method is used to gather useful statistical information based on different features present in dataset.

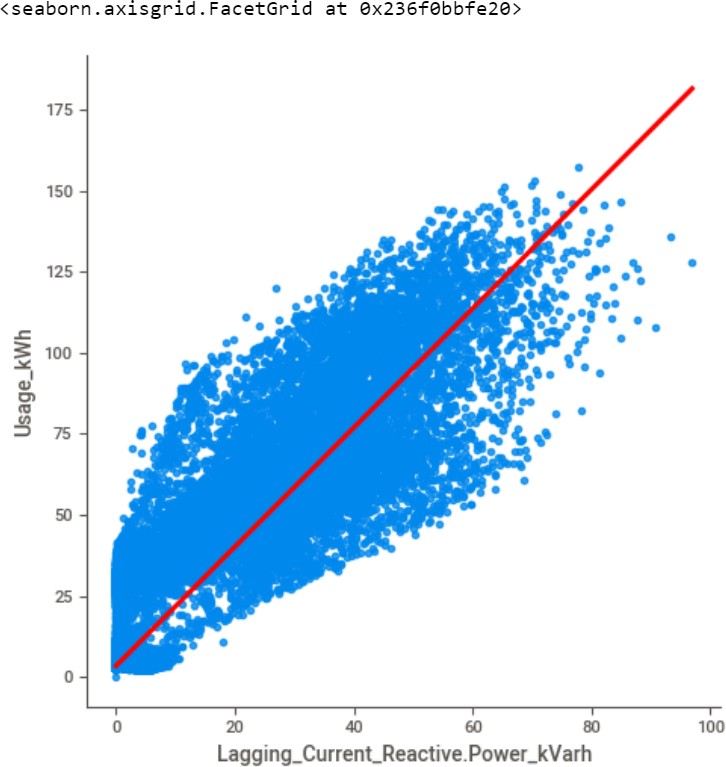
Seaborn histplot is used to visualize histogram plot along with Gaussian distribution so that one can understand how symmetrical the feature is.

Hence for same purpose we used skew method to check symmetry. If the value lies in range of

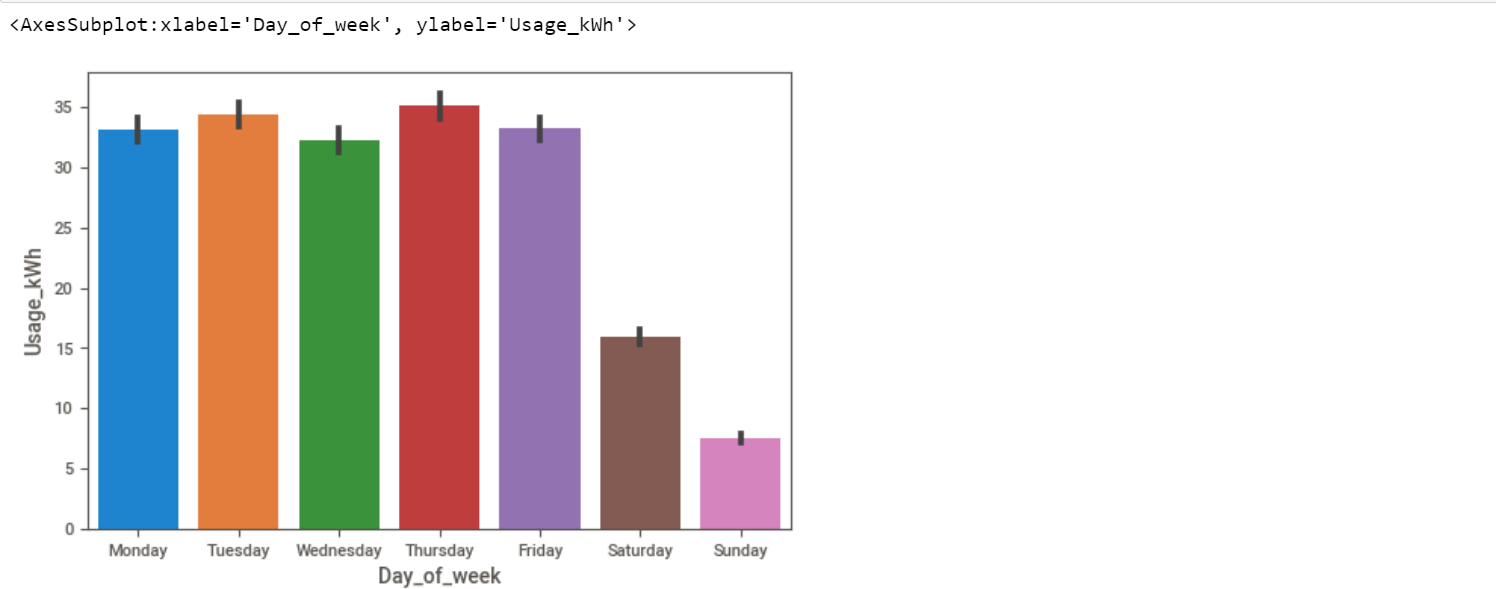
-0.5 to 0.5 it is meant to be fairly symmetrical, if it lies between -1 to -0.5 and 0.5 to 1, it is said to be somewhat symmetrical, beyond -1 or 1 is meant to be highly asymmetrical.

For the univariate analysis and to gather association between each column we used automatic data visualization and analysis tool known as sweetviz.

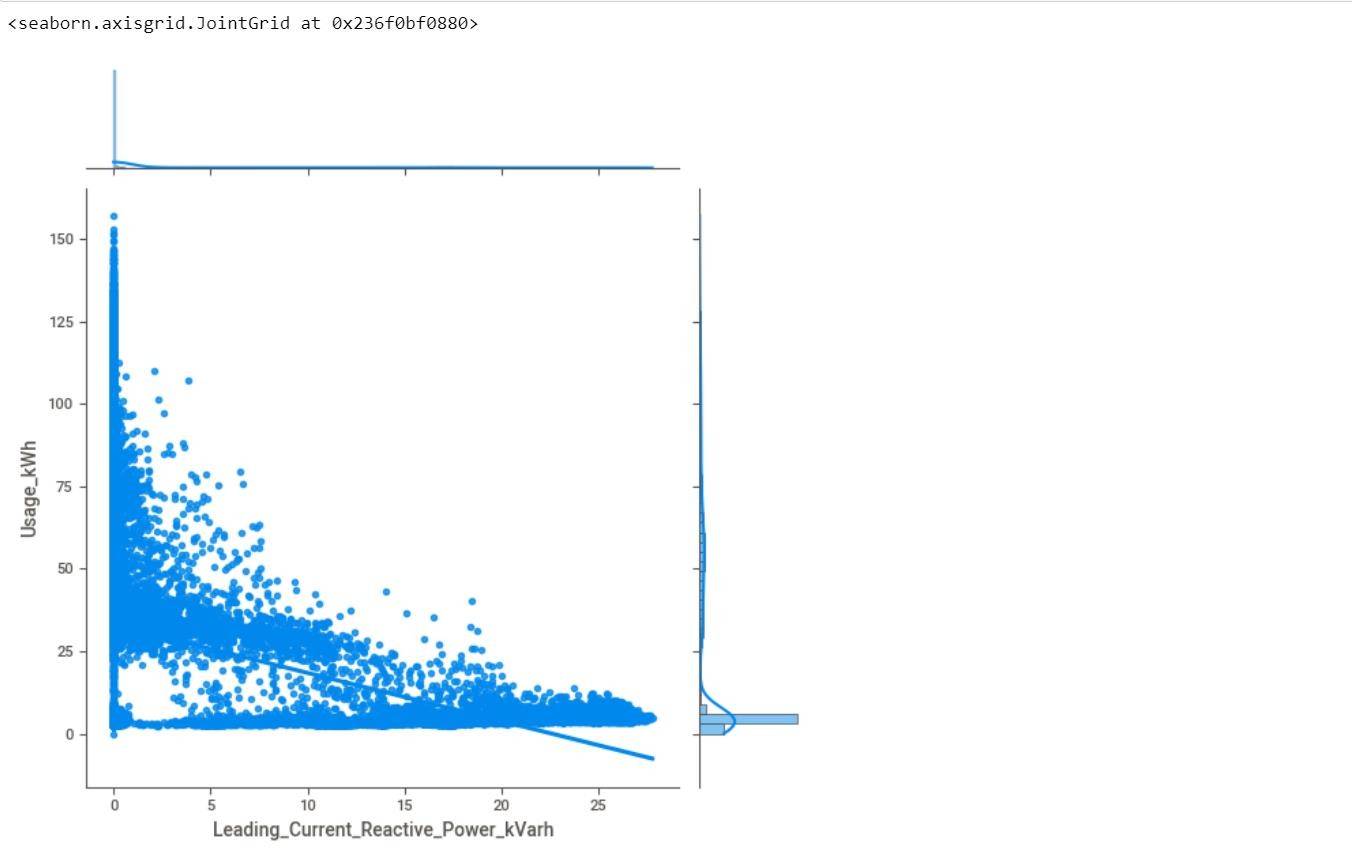
For multivariate analysis, first of all we started with linear plot along with scatter plot which is called lmplot to describe relation between Lagging\_Current\_Reactive.Power\_kVarh and Usage\_kWh.



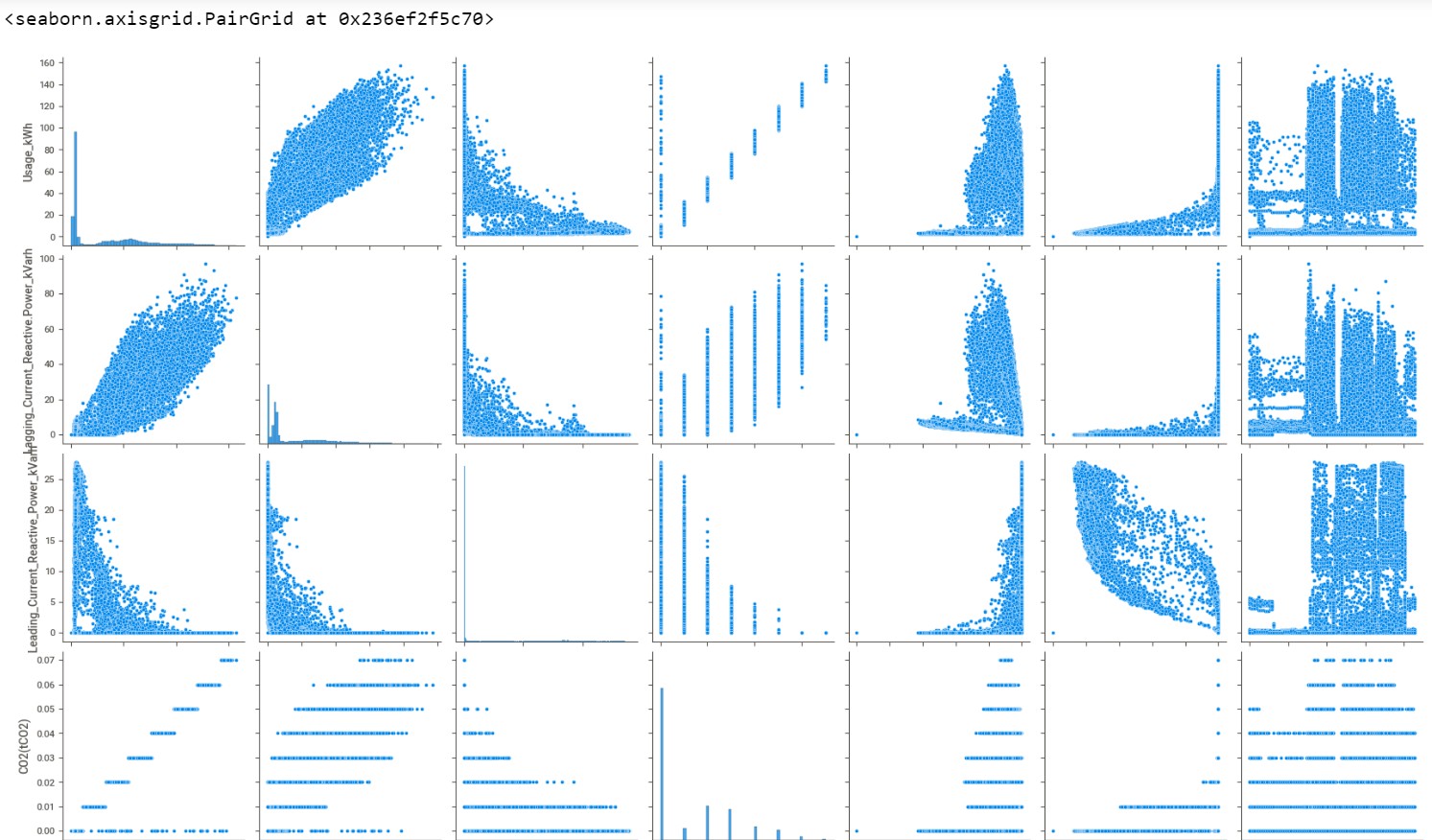
Barplot shows the average power consumption according to the different days of week.

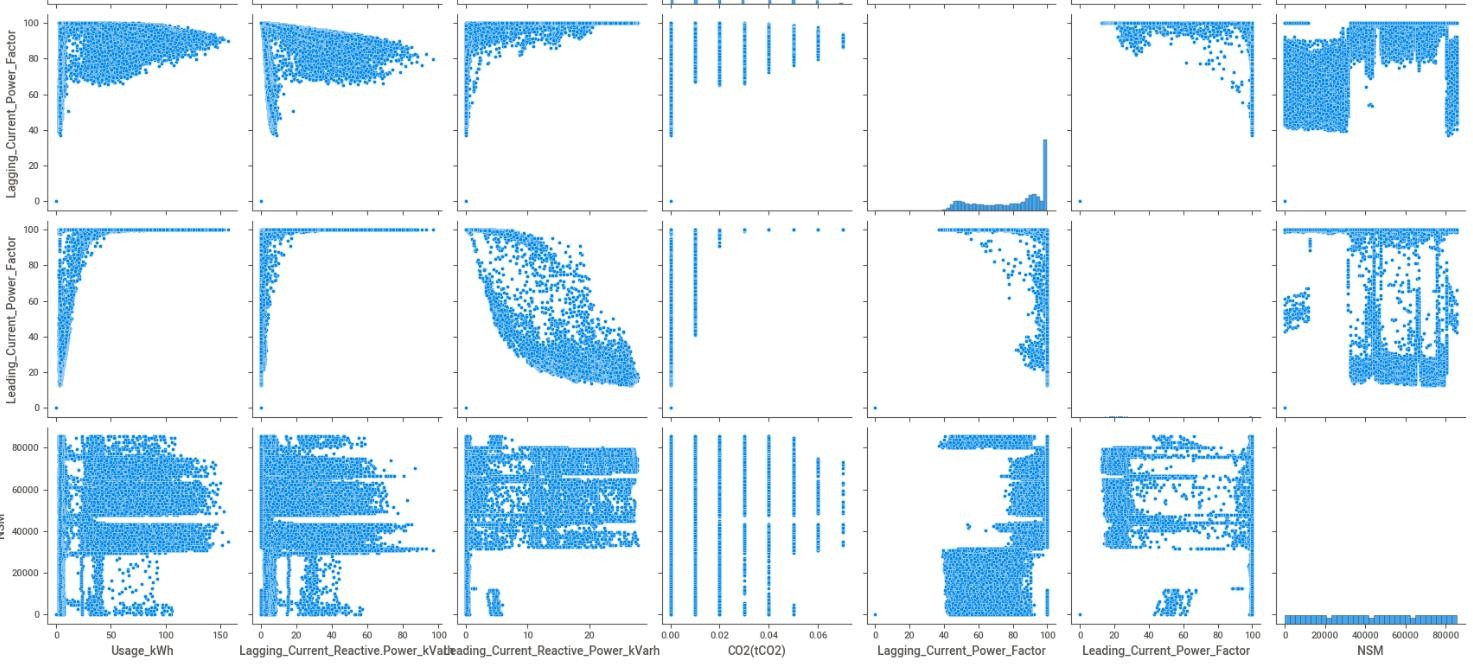


Here in this project two more plots have been used, first is jointplot which shows scatter plot as well as individual histogram for Leading\_Current\_Reactive\_Power\_kVarh and Usage\_kWh.

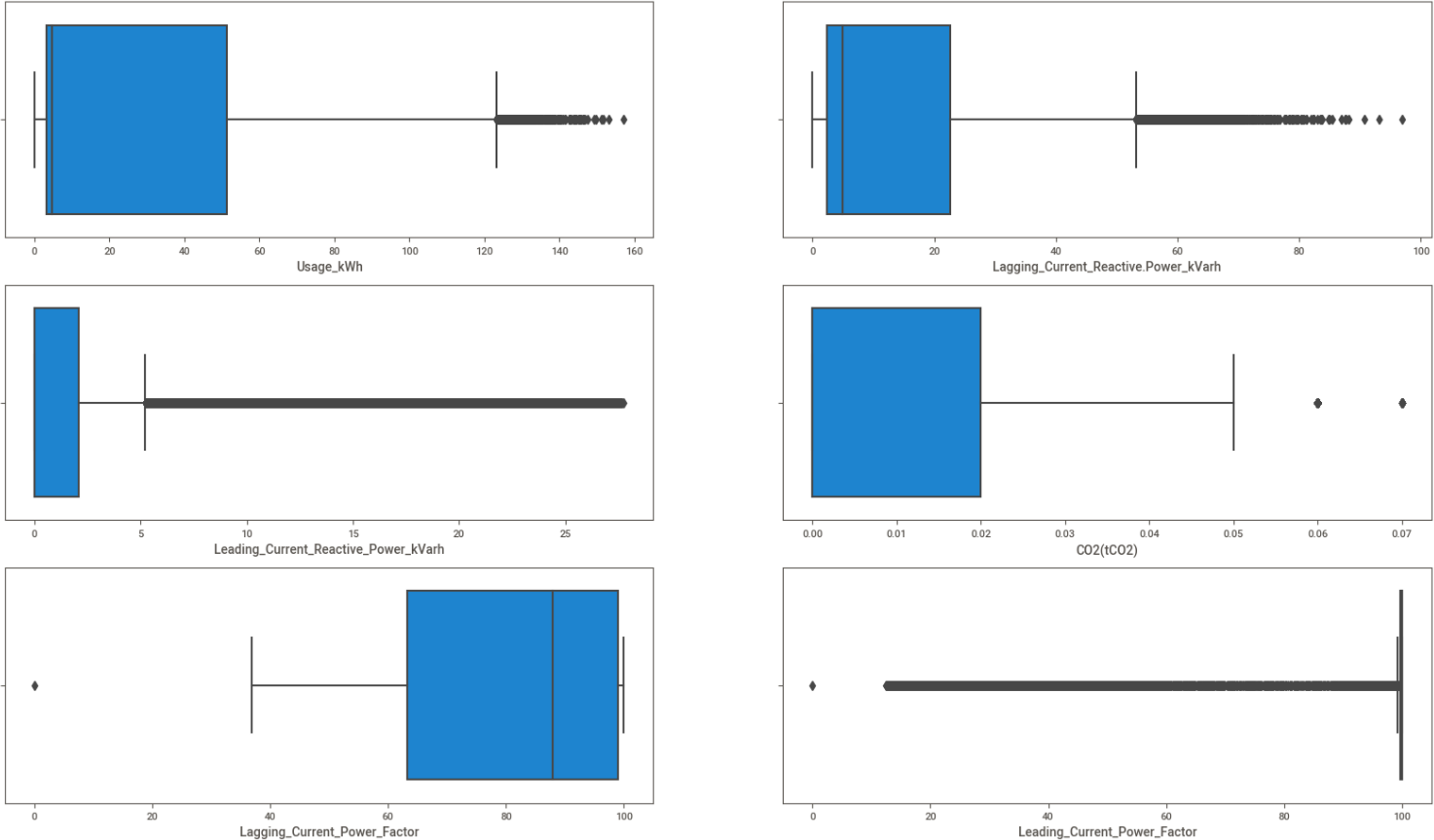


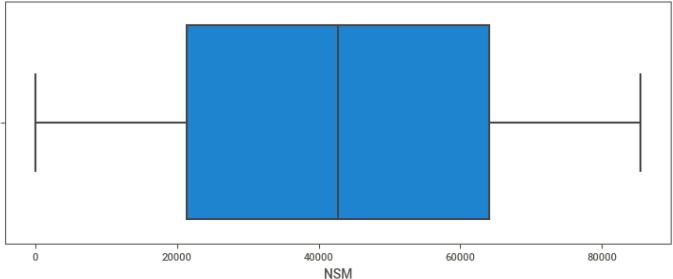
On the other hand, pairplot is used to understand the relationship between each numerical kind of column.





At last, we used boxplot to check whether an outlier is present in particular feature or not. If outliers are present, we tend to clean data by removing outliers. It is done by the concept of inner quartile range, upper fence and lower fence.



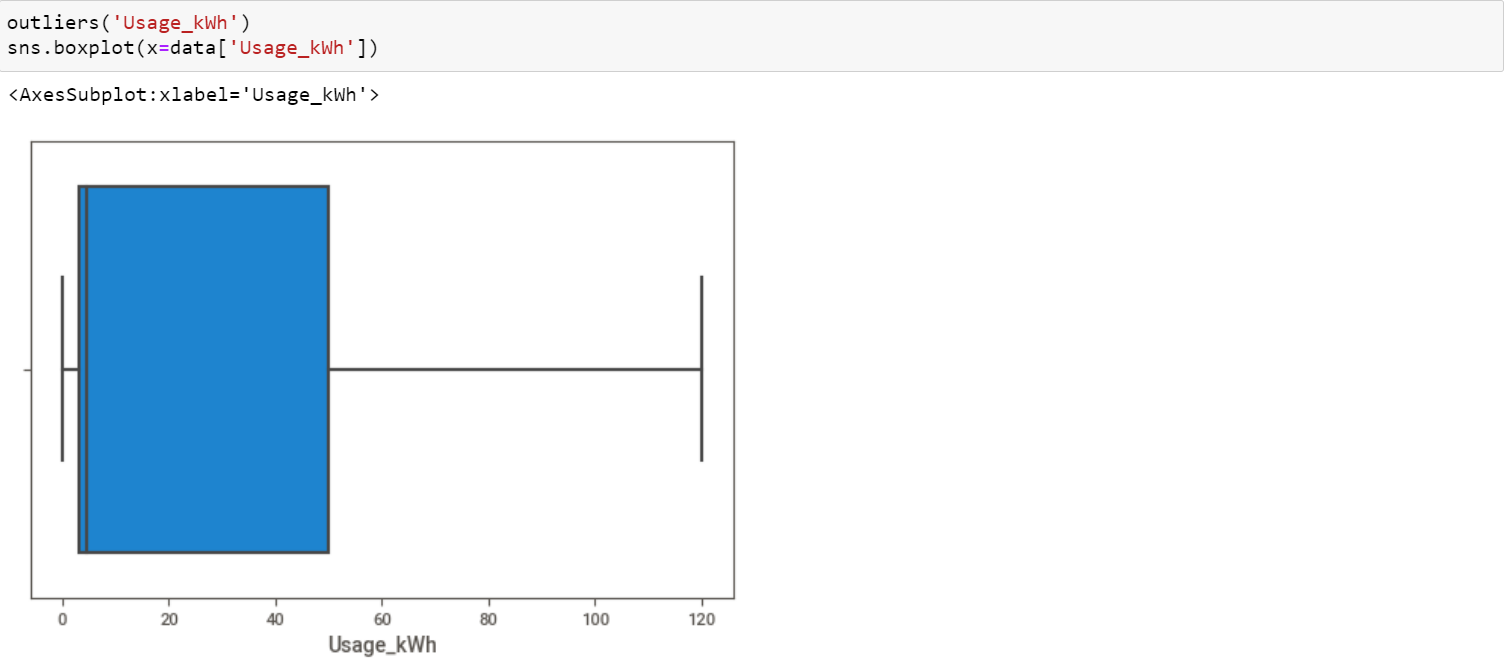


Formula as follows:

Quantile 1 which is 25 percentile is taken (Q1) Quantile 3 which is 75 percentile is taken (Q3) IQR (inner quantile range) =Q3-Q1 LowerFence=Q1-1.5\*IQR UpperFence=Q3+1.5\*IQR

Any value below lowerfence and above upperfence is considered to be outlier, which need to be dropped from dataset for better accuracy.

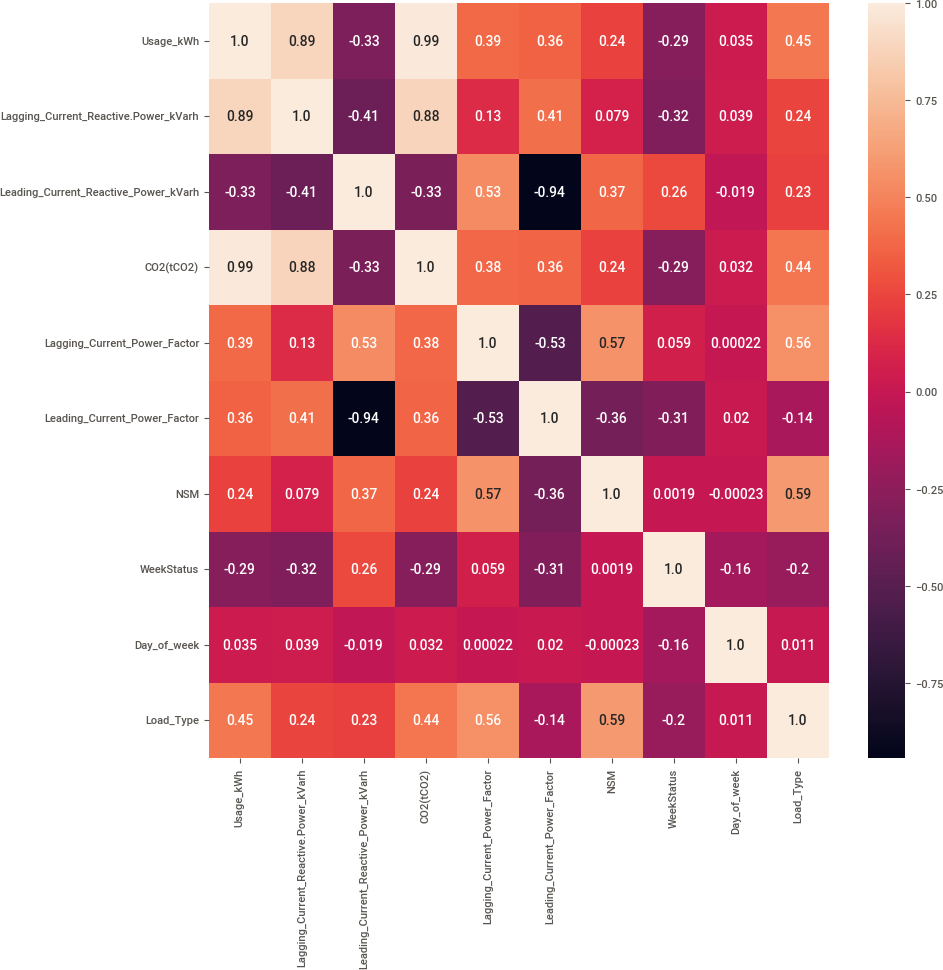
After applying the formula, we get the following box plot.



Data Preprocessing:

As we can see there are some categorical columns in data set which may impact target column, so to understand relation or to predict model they need to be converted into numerical column so here label encoder functionality is used.

Data.corr( ) shows the correlation between each column which is used further for feature engineering.



Feature Engineering:

The concept of feature engineering is to find only a few features whose importance are more, and which helps in selective prediction of model. Simply, our motive is to reduce dimensions.

Predicting model:

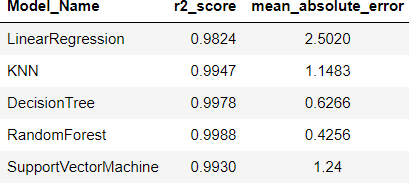
To predict model first we split the data set into an 80:20 ratio, which means 80% of the data will be used for training purposes and the rest of the amount is used for testing purposes.

Steps to predict model:

1. Need to decide and import the right model.
2. Train or make model learn or fit by using training data.
3. After that we can predict the model by passing x axis of test data.
4. After predicting one can check accuracy of model by analyzing predictive outcome and original outcome.
5. For regression type dataset we use r2\_score, mean\_absolute\_error and mean\_squared\_error to tally predictive outcome.

Here 5 algorithms are used:

1. Linear Regression
2. K Nearest Neighbors Regressor
3. Decision Tree Regressor
4. Random Forest Regressor
5. Support Vector Machine Regressor Conclusion:



Random Forest shows the highest r2 score of 99.88% and the most accurate prediction for power consumption.

# CONCLUSION AND FUTURE SCOPE

In order to predict the most accurate data, various machine learning models have been used namely Linear Regression, K Nearest Neighbors Regression, Decision Tree Regression, Random Forest Regression, Support Vector Machine Regression. Random forest regression, being at 99.88% provided the most accuracy.

Using this type of analysis, we can predict the electricity consumption on all days of the week. According to the results, government or the company can do better load management on the necessary days. And, they can have better management of production and transmission of electricity. These predictions can also help in reducing the pressure on the natural energy production resources and help in avoiding wastage of energy.

# REFERENCES

* [**https://www.javatpoint.com/**](https://www.javatpoint.com/)
* [**https://www.geeksforgeeks.org/**](https://www.geeksforgeeks.org/)
* [**https://www.python.org/**](https://www.python.org/)
* [**https://archive.ics.uci.edu/ml/index.php**](https://archive.ics.uci.edu/ml/index.php)
* [**https://www.tutorialspoint.com/index.htm**](https://www.tutorialspoint.com/index.htm)**l**