

**Final Project Report**

**“Analysis of Household Power Consumption Patterns and Anomalies: Insights through Exploratory Data Analysis and Machine Learning”**

Abhirami Muthukrishnan

Anusha Ranga

Saif Wasay Mohammed

Shawn Njoroge

Shreya Pinagani

College of Professional Studies, Northeastern University

ALY6040: Data Mining Applications

Dr. Harpreet Sharma

October 23, 2024

**Introduction**

This report provides a comprehensive analysis of household power consumption aimed at identifying patterns of peak and off-peak usage to optimize energy efficiency. Utilizing over 2 million records, the analysis includes data preprocessing, visualization, and the application of both machine learning techniques such as a decision tree, SVM, and clustering methods, like K-means. By segmenting households based on energy use variables such as Global\_active\_power and Voltage, the report identifies distinct consumption groups, offering targeted recommendations for promoting sustainable energy practices and supporting demand-based pricing strategies. Insights gained can help households reduce costs while optimizing their energy consumption.

**Analysis**

The data cleaning process involves several essential steps to prepare the dataset for analysis. First, key columns like Global\_active\_power and Voltage are converted to numeric data types to ensure compatibility. The Date and Time columns are combined into a DateTime column using the lubridate package, enabling time-series analysis. Missing values are identified using colSums(is.na()) and visualized with gg\_miss\_var(), then imputed with the median to preserve data integrity. Outliers in variables such as Voltage and Global\_active\_power are detected through boxplots and capped at the 5th and 95th percentiles to minimize their influence. Finally, numeric columns are standardized using the scale() function to enhance comparability and improve modeling accuracy.

**Exploratory Data Analysis (EDA)**

1. **Descriptive Statistics Summary Table**

The descriptive statistics table summarizes the central tendencies and variability in the dataset. For **Global Active Power**, the mean is 1.09 kW, with a median of 0.602 kW, indicating that the data is slightly right-skewed, as the mean is higher than the median. The high standard deviation of 1.05 kW suggests considerable variability in power consumption. **Global Reactive Power** has a low mean of 0.12 kW, and the median is even lower at 0.1 kW, reflecting its smaller contribution to overall power usage. **Voltage** is fairly consistent, with a mean of 240.84 volts and a narrow interquartile range, showing stability in voltage supply. **Global Intensity** has a mean of 4.6 amps but shows a large spread with a standard deviation of 4.42 amps, highlighting fluctuations in power demand. The **Sub-Metering** variables show that **Sub-Metering 3** has significantly higher average consumption, suggesting a larger energy load in this area compared to the other sub-metered regions.

1. **Patterns of Peak and Off-Peak Electricity Usage**

This line graph shows the average Global Active Power usage over the course of a day, helping identify peak and off-peak electricity usage periods. The graph reveals two major peaks, one in the morning (around 10:00) and another in the evening (around 19:00), which likely correspond to common household activities such as cooking and the use of major appliances. The off-peak periods occur during the late night and early morning, where power consumption is significantly lower. These insights into peak usage times can be useful for managing energy efficiency and could inform demand-side management strategies, such as shifting non-essential usage to off-peak hours to reduce costs.

1. **Boxplots of Various Variables Before Handling Outliers**

The boxplots before handling outliers illustrate how extreme values affected the dataset. **Global Active Power** shows a wide spread with multiple outliers above the upper quartile, indicating a high variability in energy consumption. **Global Intensity** similarly exhibits numerous outliers, with some values being exceptionally high, skewing the dataset. **Global Reactive Power** also has significant outliers, showing a wide spread in the values. The **Voltage** variable shows a slightly less pronounced effect of outliers, but there are still some values outside the expected range. This visualization highlights the importance of handling outliers to ensure a more accurate analysis, as the presence of extreme values can distort statistical interpretations.

1. **Boxplots of Various Variables After Handling Outliers**

The first set of boxplots shows the distribution of variables after outliers were handled. For **Global Active Power**, the interquartile range (IQR) is much more concentrated, indicating that most observations fall within a predictable range, improving the quality of the data for analysis. **Global Intensity** has a tighter distribution, suggesting that after outlier removal, the variability in intensity is reduced. **Global Reactive Power** shows a more consistent spread of values with less variability, while **Voltage** remains stable within a narrow range. This visualization confirms that the data is cleaner and more suitable for analysis after outlier management, with the removal of extreme values that could otherwise skew results.

1. **Patterns of Peak and Off-Peak Electricity Usage**

This line graph shows the average **Global Active Power** usage over the course of a day, helping identify peak and off-peak electricity usage periods. The graph reveals two major peaks, one in the morning (around 10:00) and another in the evening (around 19:00), which likely correspond to common household activities such as cooking and the use of major appliances. The off-peak periods occur during the late night and early morning, where power consumption is significantly lower. These insights into peak usage times can be useful for managing energy efficiency and could inform demand-side management strategies, such as shifting non-essential usage to off-peak hours to reduce costs.

1. **Average Electricity Consumption by Sub-Metering**

This graph breaks down the household's electricity consumption by sub-metering, showing the usage patterns throughout the day for three sub-metered areas. **Sub-Metering 1** and **Sub-Metering 2** show relatively lower and steady levels of consumption, with slight increases during the afternoon. **Sub-Metering 3**, on the other hand, exhibits a much higher average consumption, peaking during the evening, which correlates with overall household consumption patterns. The sharp rise in **Sub-Metering 3** usage suggests it could be linked to a particular energy-intensive activity or appliance. This analysis provides insights into which sub-metering zones consume the most electricity, offering opportunities to target specific areas for energy savings.

**Machine Learning Modeling:**

Model 1: **Decision Tree**

* Dataset Splitting and Balancing: To assess the performance of the decision tree model, the dataset was split into training and testing sets. Using a random sampling method, 70% of the data was allocated for training, while the remaining 30% was used for testing. This approach ensures that the model can generalize its predictions to new, unseen data. After splitting, the distribution of the target variable (Usage\_Category) was examined to ensure that both classes (Peak and Off-Peak) were adequately represented, which helps prevent the model from favoring one class over the other.
* Training the Decision Tree: The decision tree classifier was trained using the rpart package, with Usage\_Category as the predicted variable based on selected features like Global\_active\_power and Voltage. The model's decision-making process, which involves splitting the data based on these features' thresholds, was visualized using the rpart.plot package. For instance, one key split occurred at Global\_active\_power < -0.49, helping to distinguish between Peak and Off-Peak usage (Fig 1).
* Making Predictions and Model Evaluation: After training, the model was applied to the test dataset to predict the Peak and Off-Peak usage categories using the predict() function with type = "class" to generate categorical predictions. To assess the model's performance, a confusion matrix (Fig 2) was created, comparing the actual usage categories in the test set with the predicted ones. This matrix allowed for the calculation of key performance metrics, including accuracy, sensitivity, and specificity.
* Performance Metrics and Inference: The confusion matrix (Fig 2) demonstrates that the model achieved perfect performance, with an accuracy of 1, indicating no misclassifications in predicting Peak or Off-Peak usage. Other performance metrics, such as sensitivity (the ability to correctly classify Off-Peak usage) and specificity (the ability to correctly classify Peak usage), were also both 1. This means the model performed flawlessly in identifying both usage categories. The balanced accuracy of 1 further confirms that the model's predictions were unbiased across the classes. These results highlight the decision tree's strong effectiveness in classifying energy consumption patterns, providing valuable insights that can inform energy efficiency strategies and demand-based pricing initiatives.

Model 2 **Clustering**

* Initial Analysis & Data Sampling: This initial analysis was performed on the full dataset, which included a large number of records, resulting in an extended runtime that made the process inefficient. To address this, a sampling method was employed, where one week of data from each month was selected for analysis. This sampling reduced the data volume while preserving representative energy consumption patterns. After sampling, the K-means clustering algorithm was applied, leading to the identification of the same five clusters. This approach allowed for faster computation while still capturing distinct household energy usage behaviors, enabling meaningful insights into consumption patterns. We selected a subset of one week from each month for analysis to reduce data size and focus on a representative sample.
* Elbow Method: To determine the optimal number of clusters, we used the elbow method by calculating the total within-cluster sum of squares (WSS) for different values of k. The elbow plot (Fig 3) helped identify where increasing k provided diminishing returns in reducing WSS, indicating the ideal cluster number. We performed K-means clustering for both 5 clusters (Cluster1) and 6 clusters (Cluster2). By adding an extra cluster, the 6-cluster model is able to capture more detailed patterns in energy consumption, particularly in the variables Sub\_metering\_1, Sub\_metering\_2, and Sub\_metering\_3, which relate to kitchen, laundry, and heating appliances.
* Cluster Visualization: The scatter plots showed how the clusters were separated based on Global\_active\_power and Voltage, offering a visual understanding of the cluster groupings for both Cluster1 (Fig 4) and Cluster2 (Fig 5). In the scatter plots comparing Global\_active\_power vs. Voltage, the 6-cluster model shows a finer separation, where clusters represent more distinct ranges of power consumption and voltage levels. Some clusters with similar behaviors in the 5-cluster model have now been split into more distinct groups, highlighting subtler differences in household energy consumption. Boxplots for each variable by cluster were generated to compare the distributions across clusters, highlighting how different features vary between clusters (Fig 6&7). The boxplots for the 6-cluster model (Fig 7) show more variation across variables like Global\_intensity, Voltage, and the sub-metering features. This suggests that the additional cluster provides a more detailed classification of households, allowing for better differentiation between high, medium, and low energy consumers, as well as more specific patterns in appliance use (reflected in sub-metering variables).
* Cluster Quality: The calculated BSS/TSS ratio indicates that the 6-cluster model (Fig 9) offers a better partition of the data compared to the 5-cluster model (Fig 8). The 6-cluster model offers more granular differentiation between households. For example, the additional cluster captures variations in energy consumption related to heating and cooling appliances (Cluster 6), which were not as clearly separated in the 5-cluster model. Voltage continues to show less variation across clusters, confirming that it is not a primary factor in distinguishing households. However, other variables, such as sub-metering values, show clearer trends across the six clusters. Households in the sixth cluster exhibit distinct patterns of energy use, particularly with appliances like water heaters and air conditioners, which were previously combined with other energy consumption behaviors.
* Inference: The 6-cluster model provides more targeted recommendations. Households in Cluster 6, for example, could benefit from smart thermostats or energy-efficient heating systems, whereas households in Cluster 5 (with high laundry-related energy consumption) should consider using energy-efficient appliances during off-peak hours. The additional cluster allows for more precise intervention strategies, making it easier to tailor energy-saving programs to specific household behaviors.

Model 3 **Support Vector Machines**

* Feature Engineering: A new target variable, Malfunction, was created by labeling data points where Global\_active\_power exceeded 5.0 kW as 1 (indicating a potential malfunction) and the rest as 0 (normal operation). Non-essential columns, such as Date, Time, and Datetime, were removed from the dataset since they were not required for prediction purposes. The Malfunction variable was then transformed into a factor, turning this task into a binary classification problem.
* Data Partitioning: The dataset was divided into two sets: 70% for training and 30% for testing, using the createDataPartition() function from the *caret* package. This ensures that the model is trained on a subset of the data and its performance is later evaluated on unseen data.
* SVM Model Training: The Support Vector Machine (SVM) model was trained using the svm() function, with Malfunction as the target variable and the remaining features as predictors. A radial kernel was used to account for non-linear relationships between the features.
* Model Predictions & Evaluation: Once the model was trained, predictions were made on the test dataset using the predict() function. The model's performance was evaluated using a confusion matrix (Fig 10), created with the confusionMatrix() function, which provided key performance metrics such as accuracy, precision, and recall. These metrics offer insights into how well the model identified inefficiencies or malfunctions.
* Visualizations: A bar plot (Fig 11) was generated using geom\_bar() to compare actual malfunctions in the test set with the predicted values, visually displaying the counts of each category. Additionally, a scatter plot (Fig 12) was created to illustrate the relationship between Global\_active\_power and Voltage, with colors representing the predicted classifications, helping to visualize how the SVM model distinguishes between malfunctioning and normally functioning appliances.
* ROC Analysis: The ROC curve (Fig 13) was plotted using the roc() function from the *pROC* package, allowing for a graphical evaluation of the model’s true positive and false positive rates. This visualization helps understand the balance between sensitivity and specificity for the classifier.
* Model Development and Evaluation: After training the SVM model with the training data, predictions were generated on the test data. The model's performance was assessed using a confusion matrix, yielding metrics such as accuracy, precision, and recall. The high accuracy score suggests that the model was effective at distinguishing between normal and malfunctioning appliances.

Model 4 **Logistic Regression**

* Model training and Evaluation: The dataset was divided into training (80%) and testing (20%) sets to train and evaluate the model. A multinomial logistic regression model was fitted using the glmnet package, with cross-validation (cv.glmnet) identifying the optimal regularization parameter (lambda.min). The model’s performance was evaluated using a confusion matrix (Fig 14), which revealed how accurately it classified households into different consumption categories (High, Low, Medium).
* Model Performance: The model achieved a high accuracy of 99.6%, demonstrating its strong capability to classify households into the three energy consumption categories. The confusion matrix showed that most classifications were accurate, with only a few misclassifications, such as some "Medium" households being categorized as "High."
* Variable Relationships: Global\_active\_power emerged as the primary driver for segmentation across the consumption categories (Low, Medium, High). Other features like Voltage and Global\_intensity may also influence these categories, but further exploration, such as assessing feature importance, is necessary to determine their roles.
* Visualizations: A heatmap (Fig 15) of the confusion matrix visualized the accuracy of the model’s predictions, with color intensity representing the frequency of correct or incorrect classifications for each Actual-Predicted pair. Additionally, a distribution plot (Fig 16) of Global\_active\_power across the consumption categories provided insights into its influence on the segmentation.
* Inference: The multinomial logistic regression model performed well in classifying households by energy consumption, with high accuracy. This model can be used to create tailored energy-saving programs and continuously monitor household energy usage over time. Incorporating additional features and ongoing monitoring will further improve the model's precision and its contribution to business objectives.

Model 5 **Associate Mining**

This analysis uses the Apriori algorithm to mine association rules from sub-metering data, aiming to uncover combinations of appliances that consume significant amounts of energy. Such insights can aid both energy providers and homeowners in optimizing energy usage to mitigate peak loads. The preprocessing stage involved cleaning incomplete cases and categorizing continuous energy readings into discrete bins labeled as "High" or "Low," making the data suitable for mining. The Apriori method was applied with a confidence threshold of 0.5 and a minimum support of 0.01, yielding 26 association rules (Table 1), which were ranked by lift to identify the most meaningful relationships. A network graph (Fig 17) was generated to visualize appliance interactions during peak demand periods.

The analysis uncovered 26 rules, with the top 10 revealing notable energy usage patterns. The most prominent rule showed that high consumption in Sub\_metering\_3 (related to air conditioning or water heating) frequently coincided with high consumption in Sub\_metering\_1 (associated with kitchen appliances). This pattern appeared in 1.94% of the data and had a lift of 1.82, indicating that this combination of appliance usage is 82% more likely to occur together than by random chance. Other key findings included a rule that highlighted simultaneous high usage in Sub\_metering\_1 with low consumption in Sub\_metering\_2, suggesting that kitchen and water heating appliances are often used together, while laundry appliances are used at different times. Another pattern identified a correlation between high values in both Sub\_metering\_2 and Sub\_metering\_3, though this rule had a slightly lower lift. The relationships among appliances were visually represented, with larger, darker nodes indicating stronger and more frequent consumption patterns, providing actionable insights for energy consumption management.

**Interpretations**

This analysis applied multiple machine learning models to explore different aspects of household energy consumption. The Decision Tree model effectively predicted peak vs. off-peak usage, with visual insights from the decision process and strong performance metrics, aiding in energy management. Clustering identified six distinct household consumption patterns, offering a detailed segmentation that can inform more targeted interventions. The Support Vector Machine (SVM) model accurately classified appliances as normal or malfunctioning, with key features like Global\_active\_power and Global\_intensity driving the predictions. Although highly effective, incorporating more granular data could further enhance the model's robustness. The Logistic Regression model demonstrated high accuracy in classifying households into low, medium, and high consumption categories, with Global\_active\_power as the most influential feature. This model has practical applications for targeted energy-saving initiatives, and additional data could further improve its precision. Finally, Association Rule Mining (Apriori method) uncovered key patterns in appliance usage, such as a significant association between high usage of kitchen appliances (Sub\_metering\_1) and water heaters or air conditioners (Sub\_metering\_3), providing actionable insights for optimizing energy usage.

**Recommendations**

To improve energy management, it is recommended to implement targeted energy efficiency programs, particularly focusing on high-consumption areas like laundry and kitchen appliances. Encouraging the adoption of energy-efficient appliances and promoting better practices can significantly reduce overall consumption. Additionally, households should be encouraged to install smart monitoring systems to track key metrics like Global\_active\_power and Global\_intensity, providing real-time feedback. Appliance-specific monitoring can further help identify inefficiencies, allowing for personalized recommendations. Homeowners should be educated on optimizing usage schedules, shifting high-energy appliance use to off-peak hours to help balance demand. Future studies should also incorporate external factors like weather, household demographics, and occupancy to enhance predictive accuracy. Continuous model updates using new data will help maintain the relevance of the analysis as household behavior and appliance technology evolve. Moreover, public awareness campaigns should be launched to raise awareness about the importance of energy efficiency and its impact on overall consumption.

**Conclusion**

This report presents a comprehensive analysis of household energy consumption using a variety of machine learning models and techniques, each providing valuable insights into different aspects of energy use. The Decision Tree model successfully predicted peak and off-peak energy usage, offering a clear decision-making process and high accuracy, which supports more informed energy management decisions. K-means clustering segmented households into distinct groups based on consumption patterns, with five clusters identified as optimal, highlighting high-consumption areas like kitchens and laundry rooms, where targeted energy-saving strategies could be implemented. The Support Vector Machine (SVM) model accurately classified appliances as normal or malfunctioning based on energy consumption, indicating that monitoring metrics such as Global\_active\_power and Global\_intensity could improve energy efficiency. The Logistic Regression model effectively categorized households into low, medium, and high consumption groups, proving useful for designing targeted energy-saving programs, though further enhancements in data granularity and features could improve its performance. Finally, Association Rule Mining (Apriori Method) uncovered key patterns in appliance usage, identifying simultaneous high usage in kitchen and water heating appliances, which can inform smarter scheduling and energy usage recommendations.

Overall, the combination of these models provides a holistic understanding of household energy consumption, offering actionable insights that can guide energy efficiency programs, the adoption of smart technologies, and optimized appliance usage. Future work could benefit from incorporating external factors like weather, demographics, and real-time monitoring to further enhance the precision and utility of these models.

**References**

* Unknown. (October 7, 2024) Individual household electric power consumption. UCI Machine Learning Repository

(<https://archive.ics.uci.edu/dataset/235/individual+household+electric+power+consumption>)

* Eugenia Anello. (March 21, 2023). K-Means Clustering in R Tutorial. Datacamp

(<https://www.datacamp.com/tutorial/k-means-clustering-r>)

**Appendix**

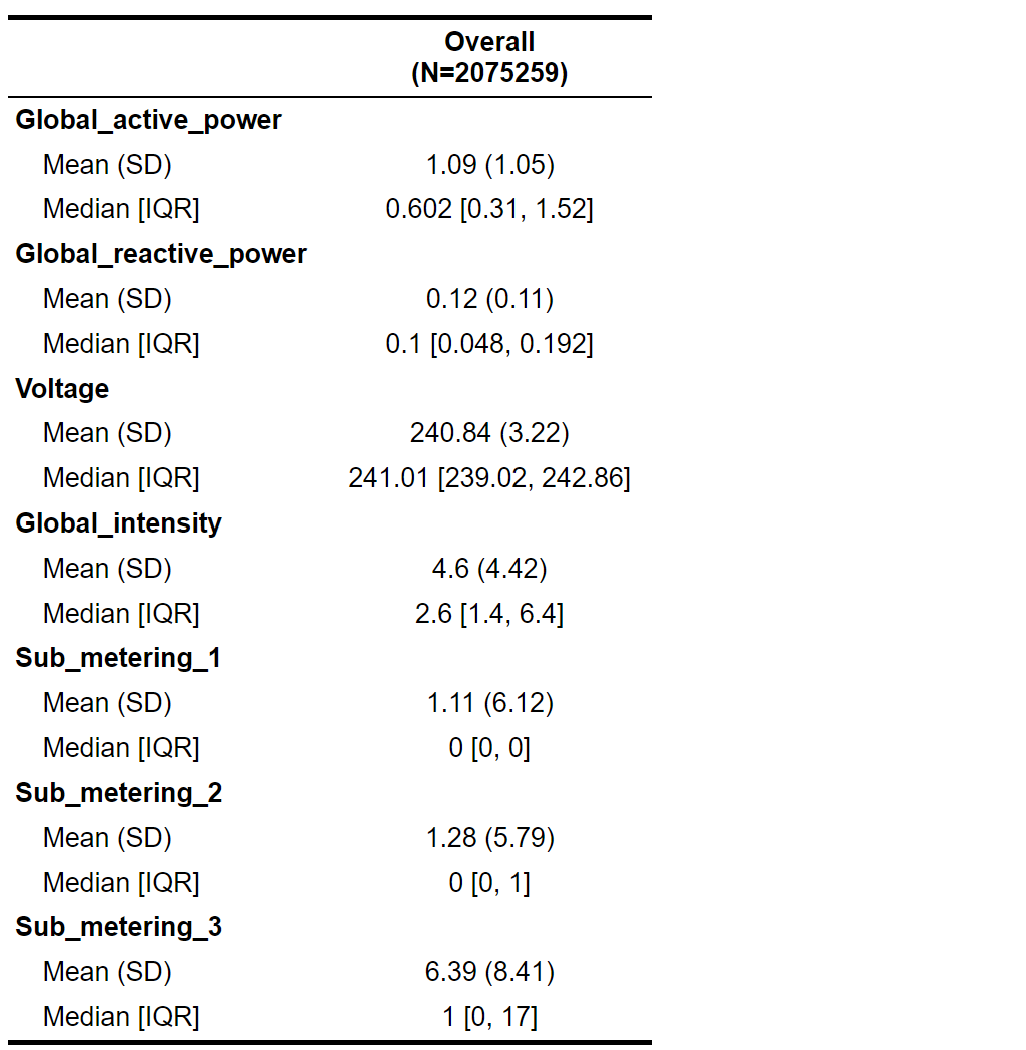
****

Fig 1 **Summary Statistics**

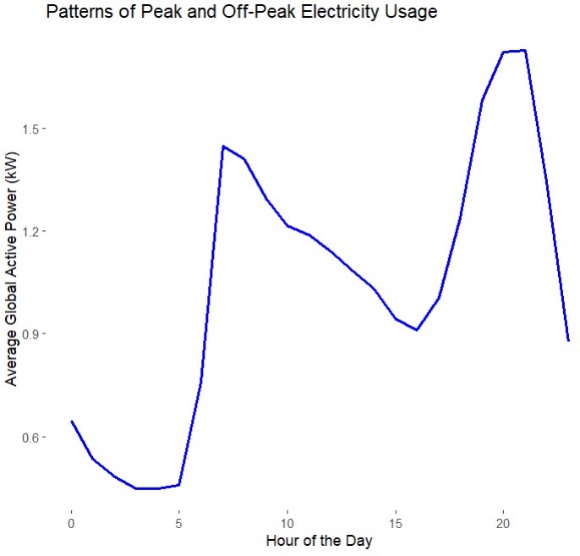
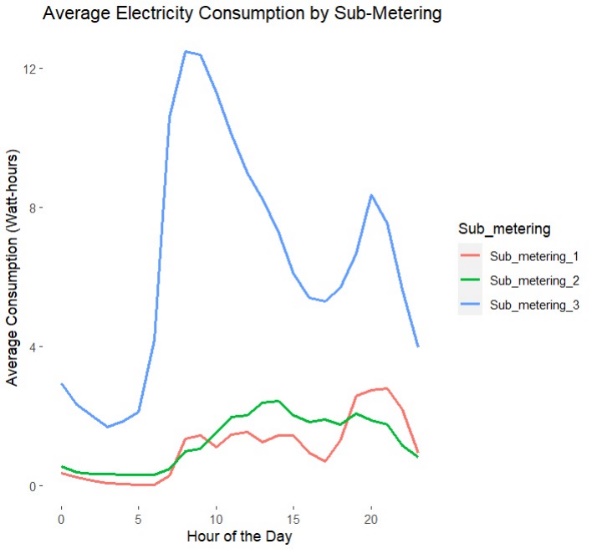
****

Fig 2 **Electricity Usage (Peak and Off-Peak)**



**Fig 3** Average Electricity Consumption by Sub-Metering

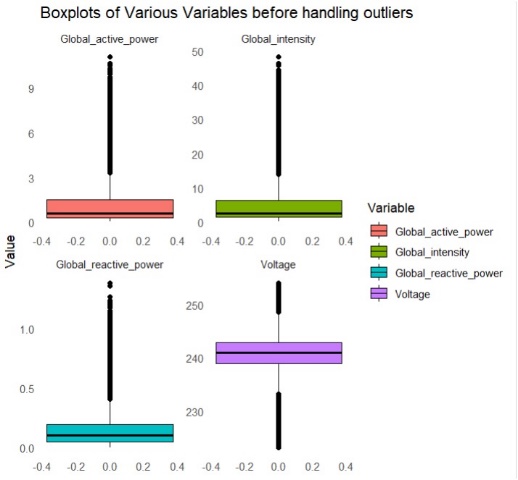
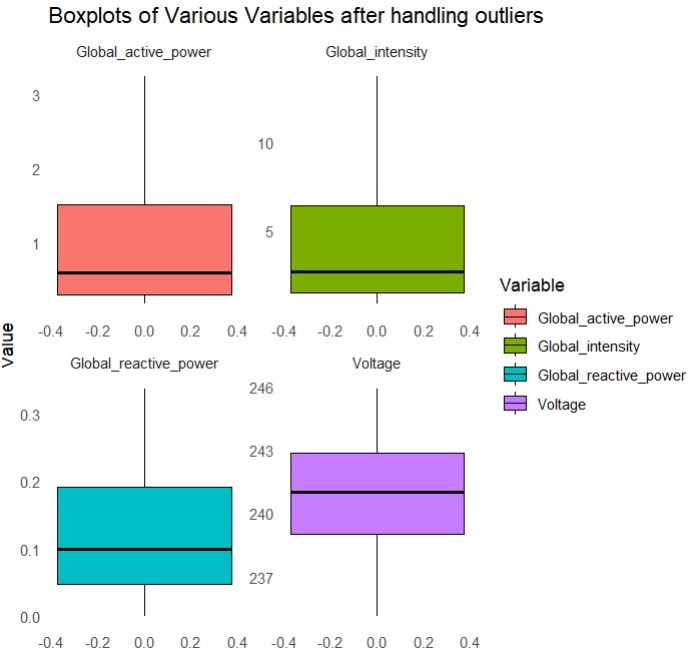


Fig Boxplots of various variables before handling Outliers

****

**Fig 4** Boxplots of various variables after handling Outliers

A diagram of a tree

Description automatically generated

Fig 5 **Decision Tree**

Fig 6 **Confusion matrix of Decision tree**

A screenshot of a computer

Description automatically generated

Fig 7 Elbow method

A graph of a number of clusters

Description automatically generated

Fig 8 **Scatter plot of Cluster 1**

A chart of multiple colored dots

Description automatically generated with medium confidence

Fig 9 **Scatter plot of Cluster 2**

A chart of multiple colored dots

Description automatically generated

Fig 10 **Box plot of Cluster 1**

A screenshot of a graph

Description automatically generated

Fig 11 **Box plot of Cluster 2**

A screenshot of a graph

Description automatically generated

Fig 12 **Quality of Cluster 1**

A screenshot of a computer code

Description automatically generated

Fig 13 **Quality of Cluster 2**

A screenshot of a computer

Description automatically generated

Fig 14 **Confusion matrix of SVM**

A screenshot of a computer

Description automatically generated

Fig 15 **Bar graph of malfunction(actual) and predicted**

A graph with red and blue squares

Description automatically generated

Fig 16 **Scatter plot of power and voltage**

A graph of a graph showing a red and green chart

Description automatically generated

Fig 17 **ROC curve of SVM**

A graph of a function

Description automatically generated with medium confidence

Fig 18 **Confusion matrix of Logistic Regression**

A number and text on a white background

Description automatically generated

ModelAccuracy:  


Fig 19 **Heat map of Confusion matrix of Logistic Regression**

A blue square with white text

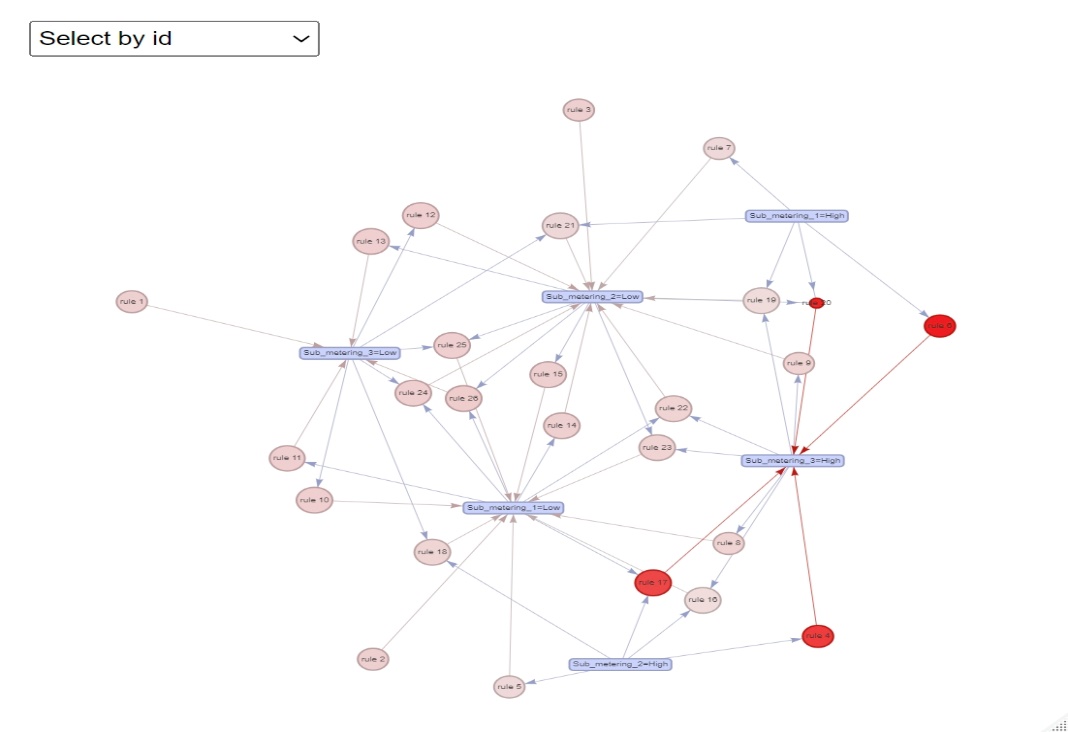
Description automatically generated

Fig 20 **Distribution of Power by consumption strategy**

A graph of a number of power

Description automatically generated

Fig 21 **Network graph of relationships**



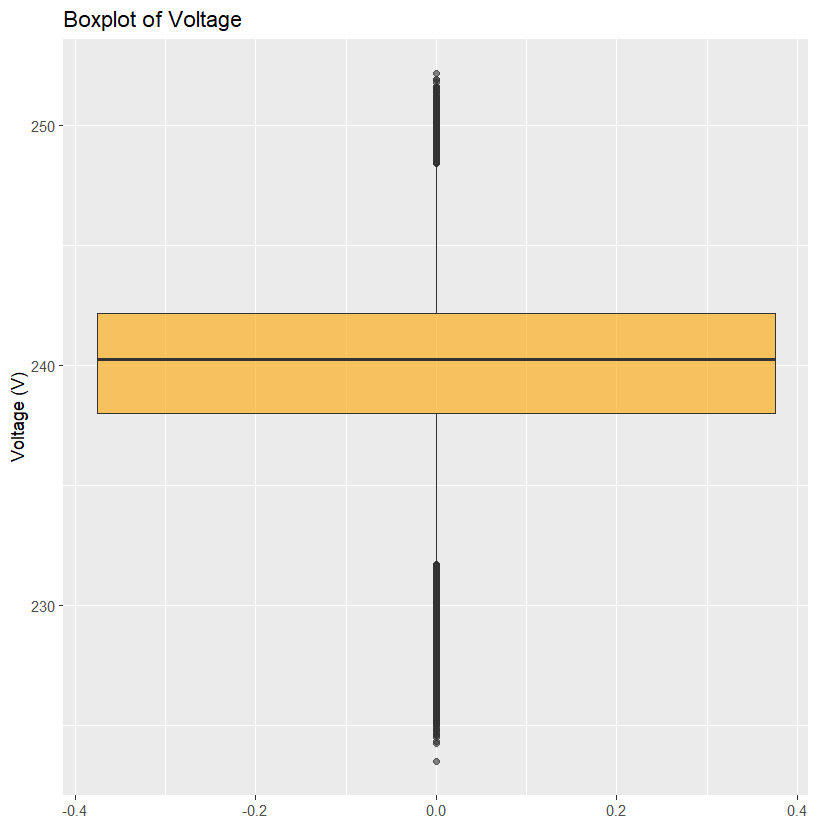
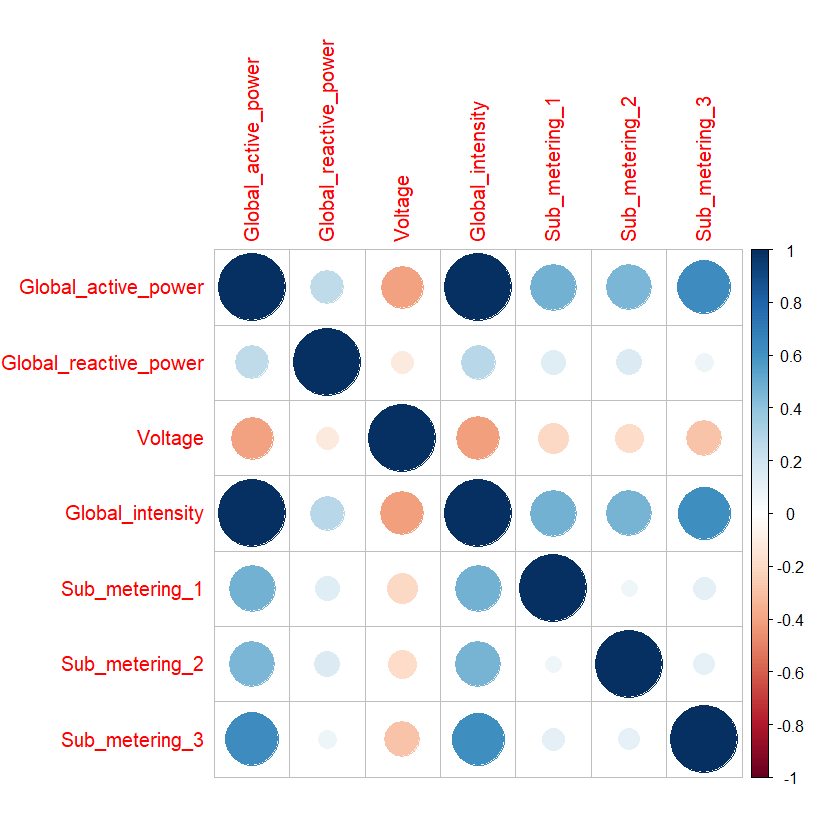
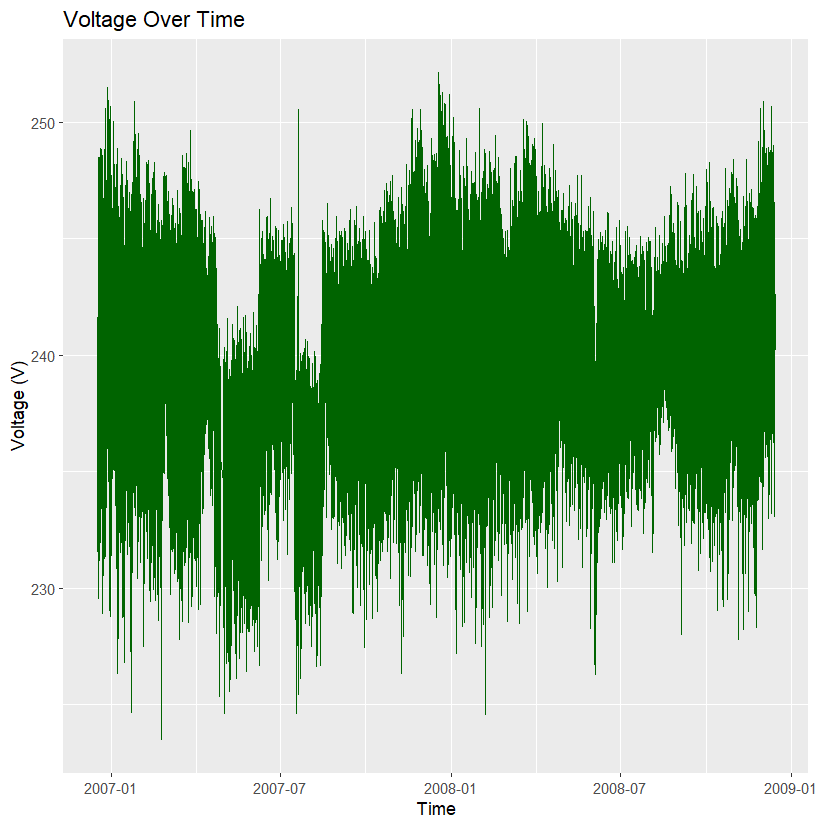


Fig 22: **Box Plot of Voltage**



**Fig 23:** Correlation Matrix Global\_active\_power Vs Global\_intensity.



**Fig 24: Voltage Over Time. (Time Series Graph).**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **rules** | **support** | **confidence** | **coverage** | **lift** | **count** |
| {} => {Sub\_metering\_3=Low} | 0.65421039 | 0.65421039 | 1 | 1 | 1357656 |
| {} => {Sub\_metering\_1=Low} | 0.969255404 | 0.969255404 | 1 | 1 | 2011456 |
| {} => {Sub\_metering\_2=Low} | 0.971167936 | 0.971167936 | 1 | 1 | 2015425 |
| {Sub\_metering\_2=High} => {Sub\_metering\_3=High} | 0.016937163 | 0.587441923 | 0.0288321 | 1.698842 | 35149 |
| {Sub\_metering\_2=High} => {Sub\_metering\_1=Low} | 0.026273347 | 0.911254471 | 0.0288321 | 0.940159 | 54524 |
| {Sub\_metering\_1=High} => {Sub\_metering\_3=High} | 0.019384568 | 0.630503268 | 0.0307446 | 1.823373 | 40228 |
| {Sub\_metering\_1=High} => {Sub\_metering\_2=Low} | 0.028185879 | 0.916775073 | 0.0307446 | 0.943992 | 58493 |
| {Sub\_metering\_3=High} => {Sub\_metering\_1=Low} | 0.326405041 | 0.943941149 | 0.3457896 | 0.973883 | 677375 |
| {Sub\_metering\_3=High} => {Sub\_metering\_2=Low} | 0.328852447 | 0.951018878 | 0.3457896 | 0.979253 | 682454 |
| {Sub\_metering\_3=Low} => {Sub\_metering\_1=Low} | 0.642850362 | 0.982635513 | 0.6542104 | 1.013805 | 1334081 |
| {Sub\_metering\_1=Low} => {Sub\_metering\_3=Low} | 0.642850362 | 0.663241453 | 0.9692554 | 1.013805 | 1334081 |
| {Sub\_metering\_3=Low} => {Sub\_metering\_2=Low} | 0.642315489 | 0.981817927 | 0.6542104 | 1.010966 | 1332971 |
| {Sub\_metering\_2=Low} => {Sub\_metering\_3=Low} | 0.642315489 | 0.661384571 | 0.9711679 | 1.010966 | 1332971 |
| {Sub\_metering\_1=Low} => {Sub\_metering\_2=Low} | 0.942982057 | 0.972893267 | 0.9692554 | 1.001777 | 1956932 |
| {Sub\_metering\_2=Low} => {Sub\_metering\_1=Low} | 0.942982057 | 0.970977337 | 0.9711679 | 1.001777 | 1956932 |
| {Sub\_metering\_2=High,Sub\_metering\_3=High} => {Sub\_metering\_1=Low} | 0.01517931 | 0.896213264 | 0.0169372 | 0.924641 | 31501 |
| {Sub\_metering\_1=Low,Sub\_metering\_2=High} => {Sub\_metering\_3=High} | 0.01517931 | 0.57774558 | 0.0262733 | 1.670801 | 31501 |
| {Sub\_metering\_2=High,Sub\_metering\_3=Low} => {Sub\_metering\_1=Low} | 0.011094037 | 0.932671663 | 0.0118949 | 0.962256 | 23023 |
| {Sub\_metering\_1=High,Sub\_metering\_3=High} => {Sub\_metering\_2=Low} | 0.017626716 | 0.909316894 | 0.0193846 | 0.936313 | 36580 |
| {Sub\_metering\_1=High,Sub\_metering\_2=Low} => {Sub\_metering\_3=High} | 0.017626716 | 0.625373976 | 0.0281859 | 1.808539 | 36580 |
| {Sub\_metering\_1=High,Sub\_metering\_3=Low} => {Sub\_metering\_2=Low} | 0.010559164 | 0.929501591 | 0.01136 | 0.957097 | 21913 |
| {Sub\_metering\_1=Low,Sub\_metering\_3=High} => {Sub\_metering\_2=Low} | 0.311225731 | 0.953495479 | 0.326405 | 0.981803 | 645874 |
| {Sub\_metering\_2=Low,Sub\_metering\_3=High} => {Sub\_metering\_1=Low} | 0.311225731 | 0.946399318 | 0.3288524 | 0.976419 | 645874 |
| {Sub\_metering\_1=Low,Sub\_metering\_3=Low} => {Sub\_metering\_2=Low} | 0.631756325 | 0.982742427 | 0.6428504 | 1.011918 | 1311058 |
| {Sub\_metering\_2=Low,Sub\_metering\_3=Low} => {Sub\_metering\_1=Low} | 0.631756325 | 0.983560783 | 0.6423155 | 1.014759 | 1311058 |
| {Sub\_metering\_1=Low,Sub\_metering\_2=Low} => {Sub\_metering\_3=Low} | 0.631756325 | 0.669955829 | 0.9429821 | 1.024068 | 1311058 |

Fig 35 **Table 1 Rules from Association Mining**