

**EDA for Final Project**

Masters of Professional Studies in Informatics, Northeastern University

ALY 6040: Data Mining Applications

Prof: Harpreet Sharma

**Group Members-**

|  |
| --- |
| Abhirami Muthukrishnan |
| Anusha Ranga |
| Mohammed Saif Wasay |
| Shawn Njoroge |
| Shreya Pinagani |

# INTRODUCTION

An overview of the preprocessing and exploratory data analysis (EDA) procedures carried out on a dataset comprising information on household electricity use is given in this study. The collection, which has more than 2 million rows, includes information on power use, including submetering data, voltage, and global active power.

# DATASET OVERVIEW

The dataset contains the following columns:

1. Date
2. Time
3. Global Active Power (kW)
4. Global Reactive Power (kW)
5. Voltage (V)
6. Global Intensity (A)
7. Sub-metering 1, 2, and 3 (Watt-hours)

Minute after minute, the data was gathered, yielding a massive dataset of 2,075,259 observations. After additional investigation, it was discovered that the data contained some possible outliers and missing numbers.

# MISSING DATA ANALYSIS

There were missing values in the dataset, especially in the voltage and submetering columns. The visualization (refer to Figure 1) indicates how many values are missing for every column. To maintain consistency and guard against bias resulting from skewed data, missing values were imputed using the median of each column. The graph makes it evident some columns—Global Active Power and Global Reactive Power, in particular—have comparatively few or no missing values, while others, like Voltage, have significant missing data.

# SUMMARY STATISTICS

Summary statistics were computed for the important numeric variables (See Figure 2). These comprise, among other things, the global intensity, voltage, and mean, median, and interquartile ranges for global active power and intensity. The results are broken down as follows:

1. Global Active Power illustrates the variation in power consumption with a mean of 1.09 kW and a standard deviation of 1.05 kW. The value at the median is 0.602 kW.
2. The mean voltage of 240.84 V and the standard deviation of 3.22 V for voltage show consistent voltage levels throughout the dataset.
3. The mean global intensity is 4.6 A, but there is a lot of variability because there are times when the intensity is very high.
4. The low median values of the sub-metering variables indicate that these appliances are frequently not used.

# OUTLIER DETECTION

To find outliers in the dataset, boxplots were utilized (See Figure 4 and 5). Outliers were visible in the boxplots, especially in the columns for global active power and voltage (See Figure 4). In order to ensure that extreme results were handled appropriately and lessen their impact on the analysis, the outliers were capped at the 95th percentile. The boxplots (refer to Figure 5) visually verify the data distribution and demonstrate that the core dataset was not distorted by capping.

# PEAK ELECTRICITY USAGE PATTERNS

By combining the data by hour of the day, peak and off-peak electricity usage was analyzed. A distinct pattern of electricity consumption was found by the investigation, with the evening (7-8 PM) and morning (7 AM) periods of high demand (See Figure 6). This pattern aligns with the usual home energy consumption, when households use more energy throughout the evenings.

# AVERAGE ELECTRICITY CONSUMPTION BY SUB-METERING

Figure 7 displays the average daily electricity use for each of the three sub-metering categories: Sub-metering 1, Sub-metering 2, and Sub-metering 3. Each sub-metering category represents a different electrical system or equipment in a residence, and watt-hours are used to assess their use.   
  
Similar to the general power consumption pattern, the blue-colored Sub-metering 3 shows the largest and most constant usage during the day, culminating in a noticeable peak in the evening. This suggests that the monitored devices or systems of Sub-metering 3, which are probably related to large kitchen appliances or necessities like air conditioning and heating, contribute significantly to the total amount of energy utilized in households.

In contrast, Sub-metering 1 (red) and Sub-metering 2 (green) show lower and more constant usage throughout the day, but a little higher usage in the afternoon and evening. These kinds of submetering categories most likely keep an eye on smaller, less complex equipment or systems. Sub-metering 1 could be used for minor kitchen appliances or lighting, while Sub-metering 2 could be used for household electronics like computers or entertainment systems.   
  
Overall, this graph highlights the systems that drive peak electricity usage and offers important insights into the breakdown of electricity consumption by individual household appliances. By identifying which sub-metering categories contribute the most to power consumption, energy management strategies can be modified to enhance efficiency and decrease unnecessary usage during peak hours.

# RELATIONSHIP BETWEEN GLOBAL ACTIVE POWER AND VOLTAGE

The link between voltage and global active power (measured in kilowatts) is depicted in Figure 8. A negative association is seen, much like in the previous graph, where lower voltage levels are linked to higher active power utilization. The line of best fit illustrates that as global active power increases, the voltage tends to drop.  
  
This inverse relationship is crucial to comprehending how variations in voltage levels can result from increasing demand for electrical power, particularly during peak hours. The graph shows that while the worldwide active power typically ranges from 1 kW to 3 kW, the majority of the voltage values are focused between 237 V and 246 V. As appliances demand more power, the strain on the electrical system might lead to a fall in voltage, perhaps impacting the performance of devices.

For energy providers, this information is crucial in optimizing power distribution and mitigating voltage drops during high-demand periods. Implementing strategies to manage power loads can help ensure a consistent supply of electricity, maintaining stable voltage levels across households and improving overall energy efficiency.

# RELATIONSHIP BETWEEN GLOBAL INTENSITY AND VOLTAGE

The link between voltage, expressed in volts (V), and global intensity, measured in amperes (A), is depicted in Figure 9. These two variables have a negative association, as seen by the scatter plot, which also clearly indicates a downward trend in the voltage as the global intensity rises. According to this inverse relationship, voltage tends to decrease as household appliance current (global intensity) increases.  
  
This behavior is consistent with well-known electrical circuit concepts, such as Ohm's Law, which states that increased current can lower voltage in the presence of resistance in the system. The dense clustering of points at specified values shows frequent occurrences of these combinations, possibly reflecting regular working circumstances for household appliances. This pattern demonstrates how the electrical system experiences a proportional drop in voltage as more power-intensive devices are operated, which raises the global intensity.  
  
Maintaining voltage stability during times of heavy energy demand requires careful management of this relationship. The efficiency and lifespan of home appliances can be impacted by notable voltage changes during peak demand, which makes this information important for both energy providers and consumers.

# RELATIONSHIP BETWEEN GLOBAL INTENSITY AND GLOBAL ACTIVE POWER

Global intensity (A) and global active power (kW) have a significant positive linear association, as seen by the scatter plot that depicts their relationship. The closely packed spots along the blue trend line indicate that global active power grows proportionately to increases in global intensity. The direct correlation between current and power consumption in home appliances is reflected in this linearity: the more current an appliance pulls, the more active power it uses.   
  
There is little departure from the trend in the data points, suggesting that there is a dependable and predictable link between these factors. Power consumption stays relatively low at lower intensity levels (below 5 amperes), usually less than 1 kW. However, there is a discernible increase in power consumption with increasing intensity, notably over 5 amperes, which can reach up to 3 kW for higher current values. This pattern illustrates how household appliances should behave, with larger appliances requiring more current and using more active power.

# CONCLUSION AND NEXT STEPS

Important findings from the statistics on residential power consumption are highlighted in this EDA study. To enhance the quality of the data, outliers were capped and missing values were imputed. Subsequent actions entail utilizing data mining methodologies to unearth more profound understandings, like forecasting the highest point in electricity use, recognizing usage trends across various intervals, and investigating the correlation between energy usage and extraneous variables like meteorological conditions.

# REFERENCES

UCI Machine Learning Repository. (n.d.). Archive.ics.uci.edu. <https://archive.ics.uci.edu/dataset/235/individual+household+electric+power+consumption>

Amin, S. M., & Wollenberg, B. F. (2005). Toward a smart grid: Power delivery for the 21st century. *IEEE Power and Energy Magazine*, *3*(5), 34-41. https://doi.org/10.1109/MPAE.2005.1507024

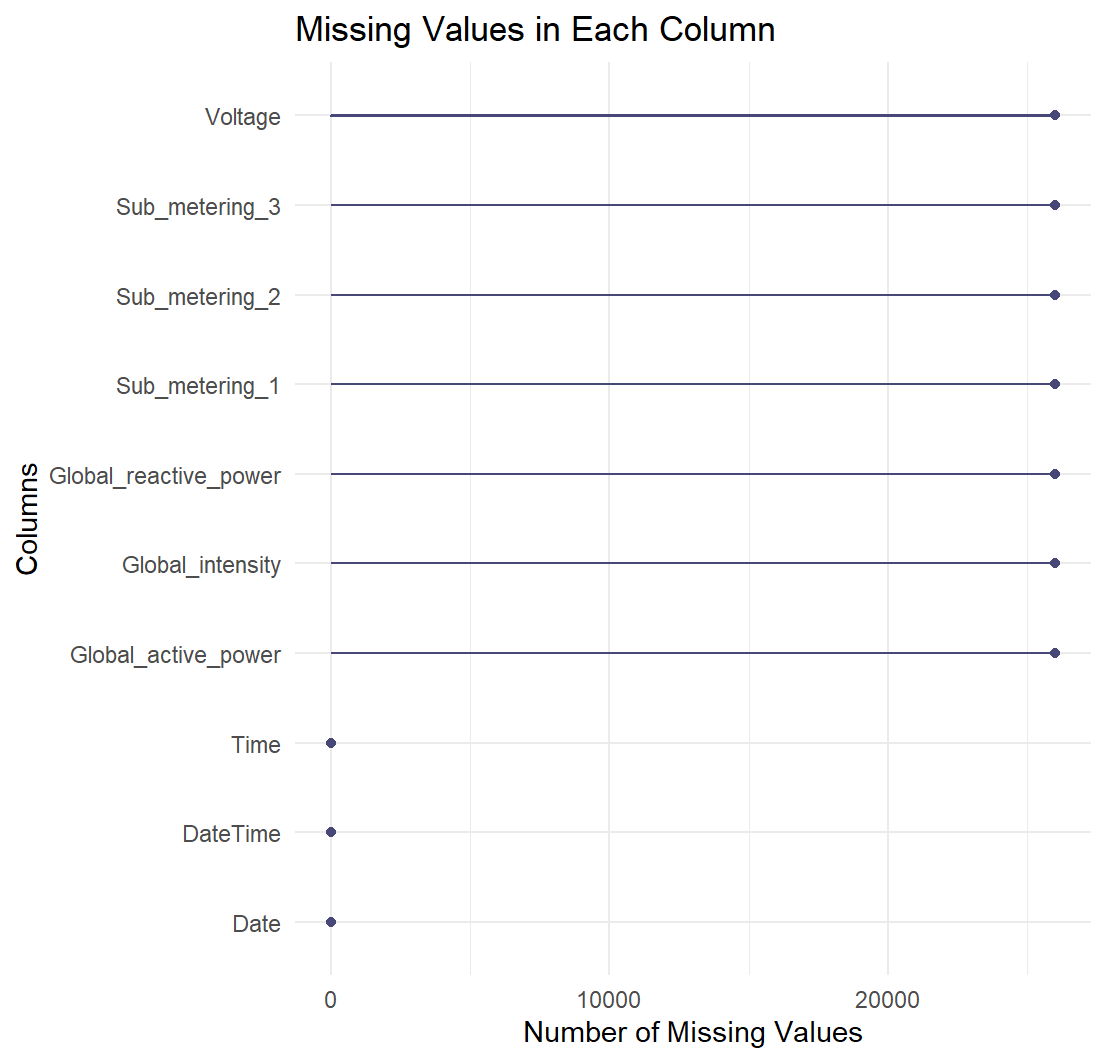
De Myttenaere, A., Golden, B., Le Grand, B., & Rossi, F. (2016). Mean absolute percentage error for regression models. *Neurocomputing*, *192*, 38-48. https://doi.org/10.1016/j.neucom.2015.12.114

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. Springer. <https://doi.org/10.1007/978-1-4614-7138-7>

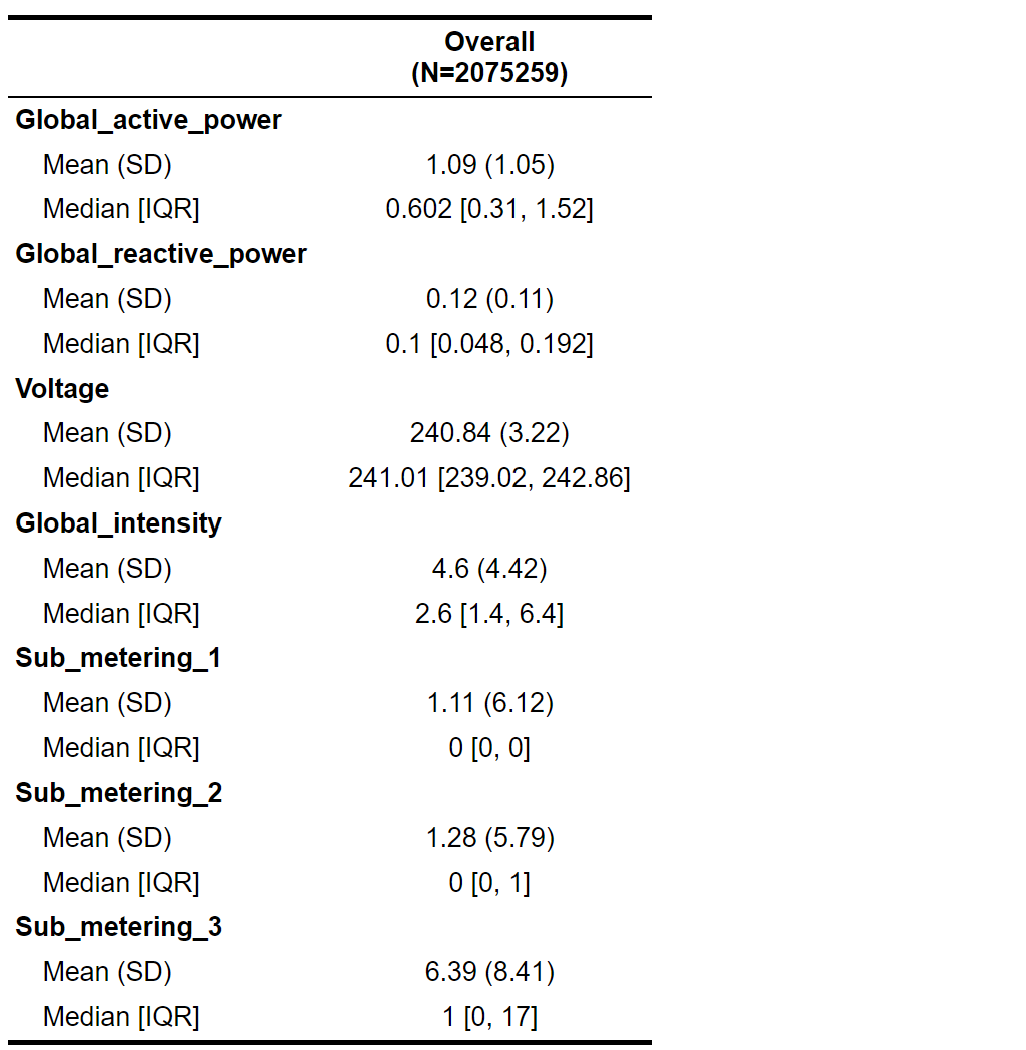
# APPENDIX

This section contains the visualizations referenced throughout the report. These images help provide a visual understanding of the data distribution, missing values, and patterns in electricity usage.

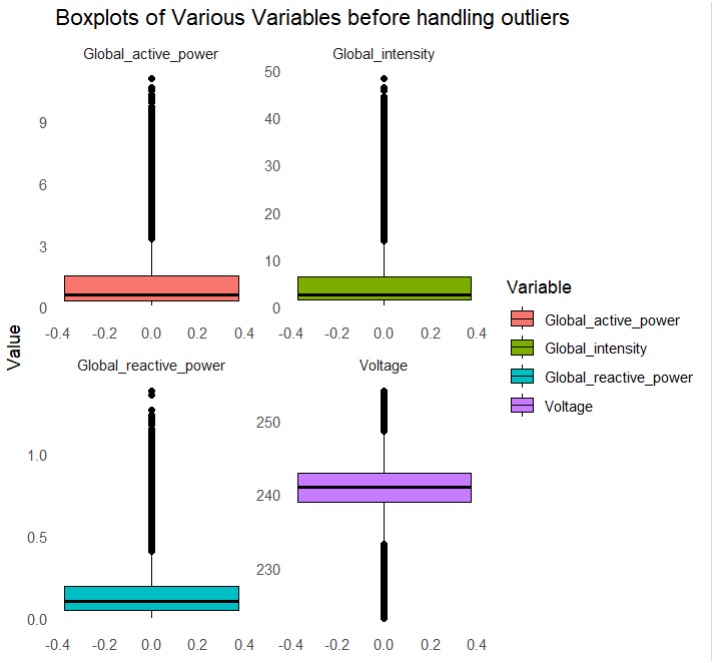
## Figure 1: Missing Values in Each Column



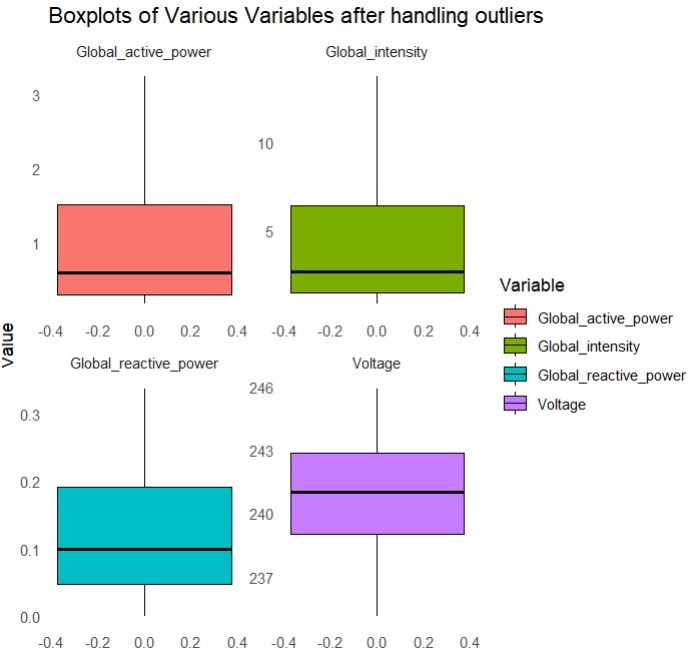
## Figure 2: Summary Statistics



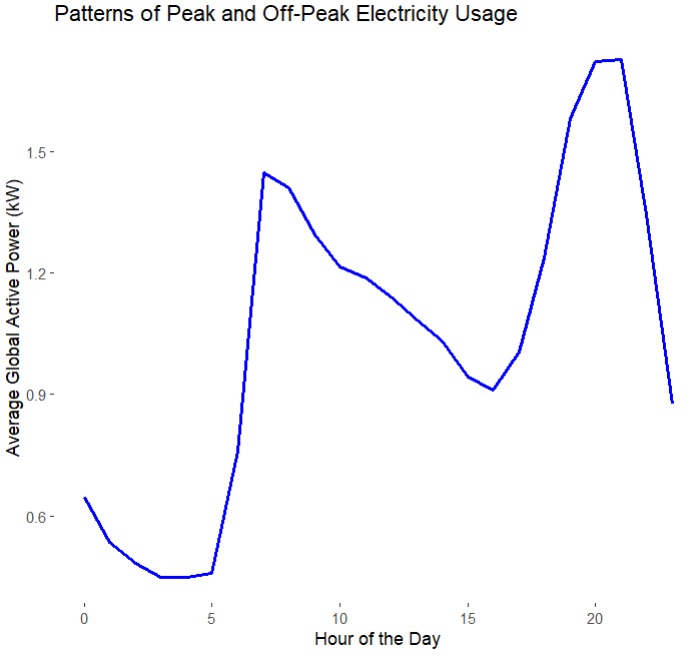
## Figure 3: Boxplot of Various Variables before handling outliers



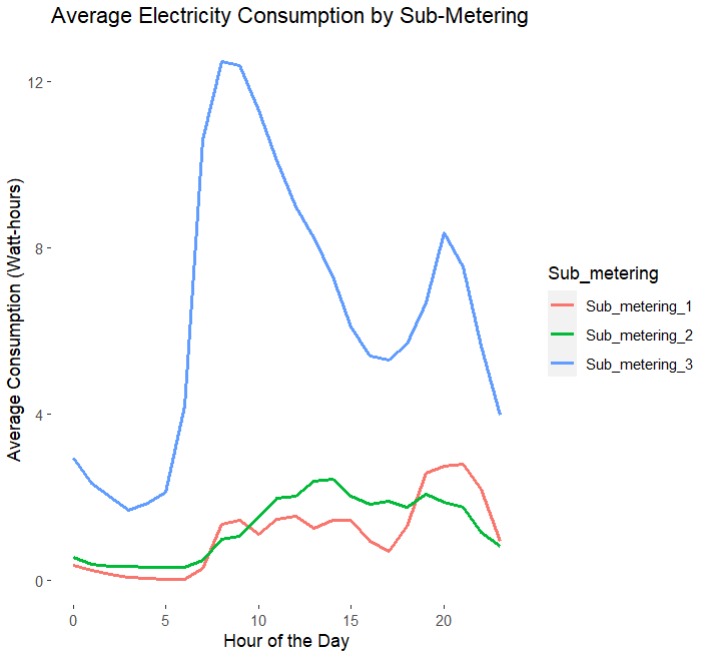
## Figure 4: Boxplot of Various Variables after handling outliers



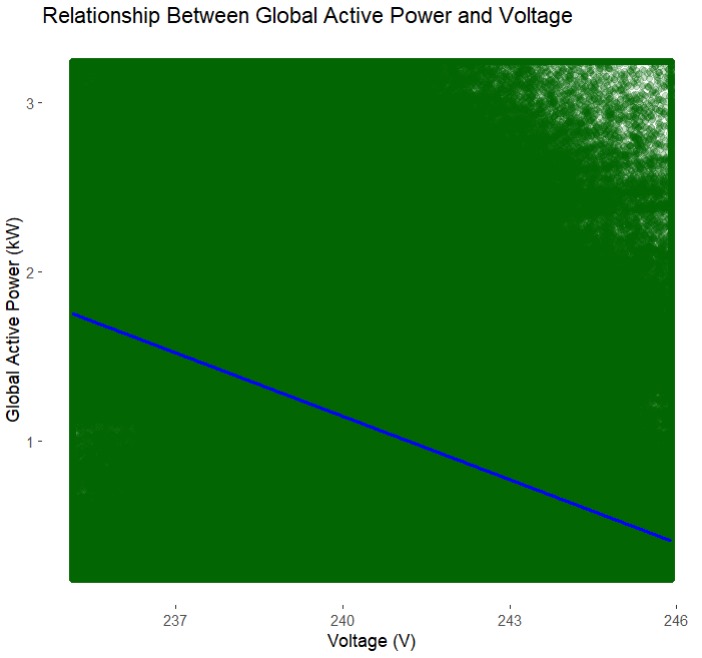
## Figure 5: Patterns of Peak and Off-Peak Electricity Usage



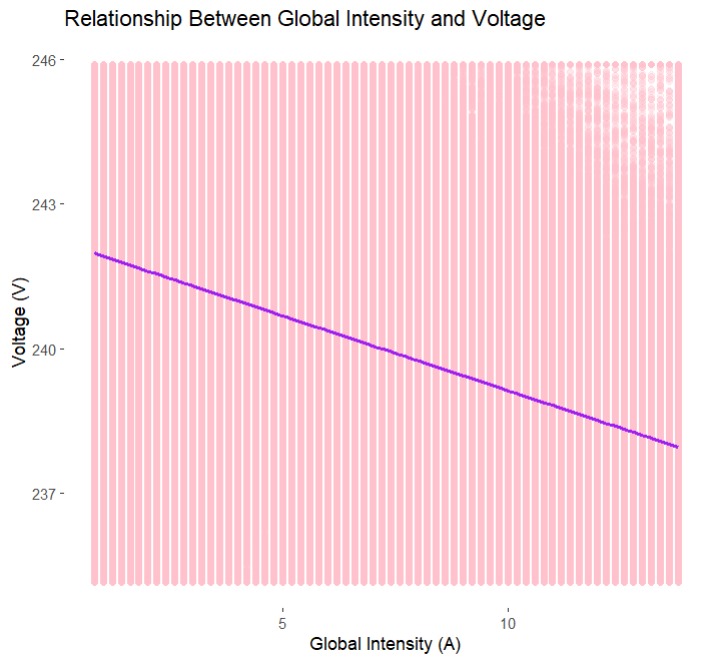
## Figure 6: Average Electricity Consumption by Sub-Metering



## Figure 7: Relationship between Global Active Power and Voltage



## Figure 8: Relationship between Global Intensity and Voltage



## Figure 9: Relationship Between Global Intensity and Global Power

