CMPT 318
CYBERSECURITY
Course Project Report
Group 7

Wei Yao
Neil Mukesh Shah
Karan Sachdeva
Harry Preet Singh
Praneer Shrestha

Abstract

Critical Infrastructures such as a power grid, thermal plants, water metering etc produce large amounts of time series data. This project aimed to explore the data of electric consumption in some foreign electrical power grid and figure out point and contextual anomalies. Point anomalies were found out by general statistical methods such as standard deviation, comparing means, maximum and minimum values. Feature selection was based on correlation coefficients of the training data set.

Contextual anomalies were determined by building hidden markov models. The models were built upon training data to replicate normal behavior and said models were used on test data where their log likelihoods were compared to evaluate anomalous behaviour.

Based on our models, and the log likelihoods we determined that the test data sets is either quite anomalous or our models are not an accurate representation of the normal behaviour.

Contents

1. **Introduction……...3**
2. **Background Information…….....3**
3. **General Data Exploration……....4**
    a. **Correlation Coefficients……...11**
    b. **Comparing Training Data with Test Data……..12**
4. **Finding Point Anomalies…….....21**
5. **Finding Contextual Anomalies……..46**
6. **Precision and Recall……...49**

Introduction

"In light of increasing cyber threats, especially advanced persistent threats, and existing vulnerabilities that expose critical infrastructure to a variety of adversarial scenarios, the project explores behavior based intrusion detection methods used for cyber situational analysis of automated control processes." (Course Project).

This project is based on the dataset for electrical household consumption. To begin, the training dataset is investigated to understand underlying trends for certain time windows, and significance of the various characteristic features by statistical comparisons. Then, using the training data, an attempt to find anomalies in the given test data is performed using two approaches. The first approach aims to find point anomalies, while the second approach aims to detect contextual anomalies by building a Hidden Markov model (HMM); a behavior-based intrusion detection method.

Background Information

Due to the exponential progress towards Internet of Things (IOT)[1] and intensening of automation, cyberattacks are increasingly routine and sophisticated. Automation is vital for the continuous operation of critical infrastructure[2] and the services it provides because it enhances cost efficiency, quality of service delivery and safe operation of critical assets. However, increasing reliance on automation, and having a large system of interrelated computing devices in effect increases the attack surface for advanced persistent threats[3] and amplifies the risk of cascading effects.

The threat landscape has evolved to the point that risks that were once considered unlikely began occurring with regularity. This ongoing trend can be attributed to higher maturity of attack tools and methods, increased exposure, increased motivation of attackers, and better detection tools enabling more visibility.

[1] A system of interrelated computing devices, mechanical and digital machines, objects that are provided with unique identifiers (UIDs) and the ability to transfer data over a network without requiring human-to-human or human-to-computer interaction.
[2] Refer to processes, systems, facilities, technologies, networks, assets and services essential to the health, safety, security or economic well-being of people and the effective functioning of government.
[3] A set of stealthy and continuous computer hacking process, often orchestrated by a person or group targeting a specific entity (either a private organization, a state for both business or political motives).

In recognition of the limitations of signature-based analysis of network traffic, defenders have shifted their focus to more behavioral-based anomaly detection. This is because signature-based detection schemes contained a time lag between the initial attack and availability of either a signature or patch. Furthermore, behavior-based anomaly detection could overcome the limitation of being able to alert only on specific signature matches, and they have the potential to discover evidence of zero-day exploits[4] through identifying unusual behaviors.

General Data Exploration

The primary purpose in exploring the data was to identify an interesting time window and select significant characteristic features in order to create a suitable probabilistic model that could represent 'normal' system behavior.

We were given a dataset of approximately three years organized by a date and time column with characteristic features such as, global_active_power, global_reactive_power, global_intensity and voltage to represent household electricity consumption.

For reference the details about the data are as follows:

- **Date:** Date in format dd/mm/yyyy
- **Time:** Time in format hh:mm:ss
- **Global_active_power:** Household global minute-averaged active power (in kilowatts)
- **Global_reactive_power:** Household global minute-averaged reactive power (in kilowatts)
- **Voltage:** Minute-averaged voltage (in volts)

The columns corresponding to Sub_metering_1, Sub_metering_2, Sub_metering_3 were determined to be insignificant as its values would not be able to create probabilistic models to represent normal behavior and determine anomalies.

Hence, deciding to focus on the significant features, we plotted each feature's average value with the date to examine the overall underlying patterns of power consumption.

[4] A cyber-attack that occurs on the same day a weakness is discovered in software. At that point, it's exploited before a fix becomes available from its creator.

Figure 3.1.1: Plot of date (x-axis) vs average global_active_power (y-axis) of entire training data

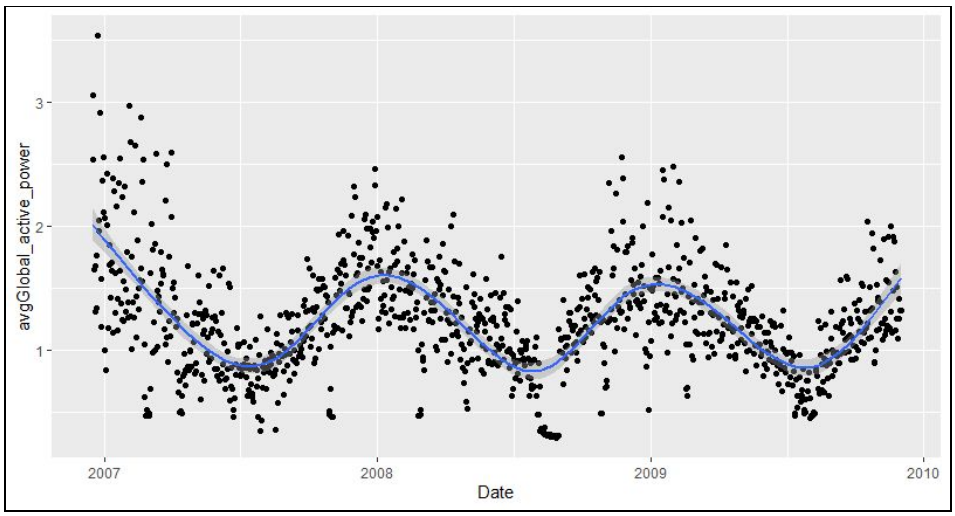


Figure 3.1.2: Plot of date (x-axis) vs average global_reactive_power (y-axis) of entire training data

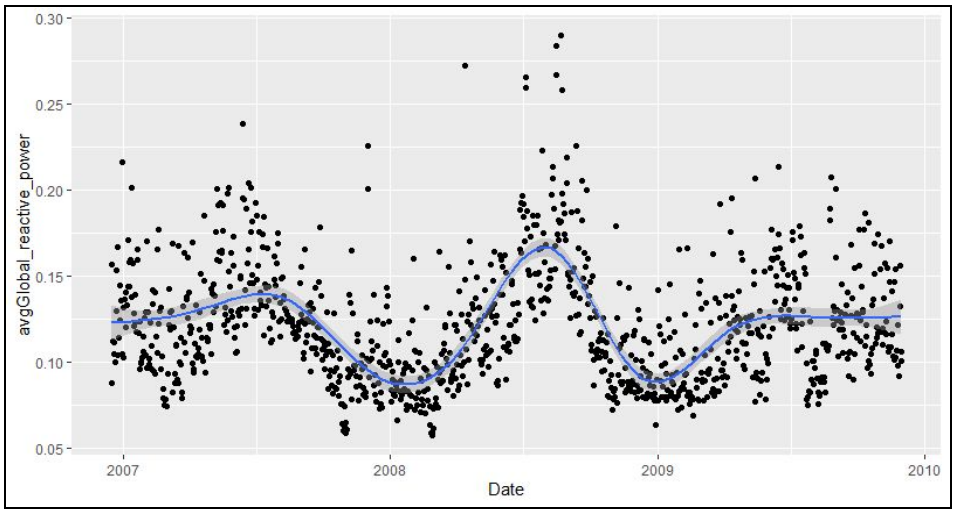


Figure 3.2.3: Plot of date (x-axis) vs average global_intensity (y-axis) of entire training data

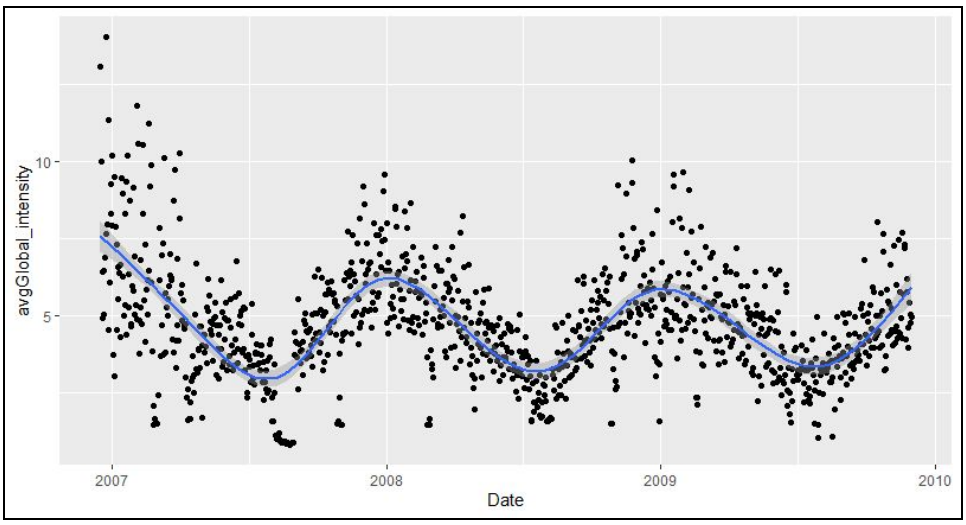
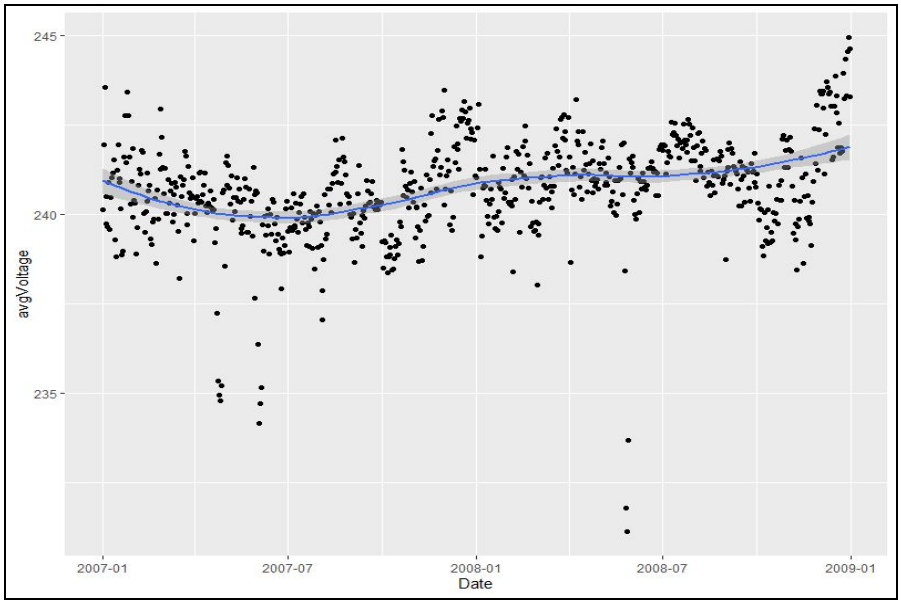


Figure 3.1.4: Plot of date vs average voltage of entire training data

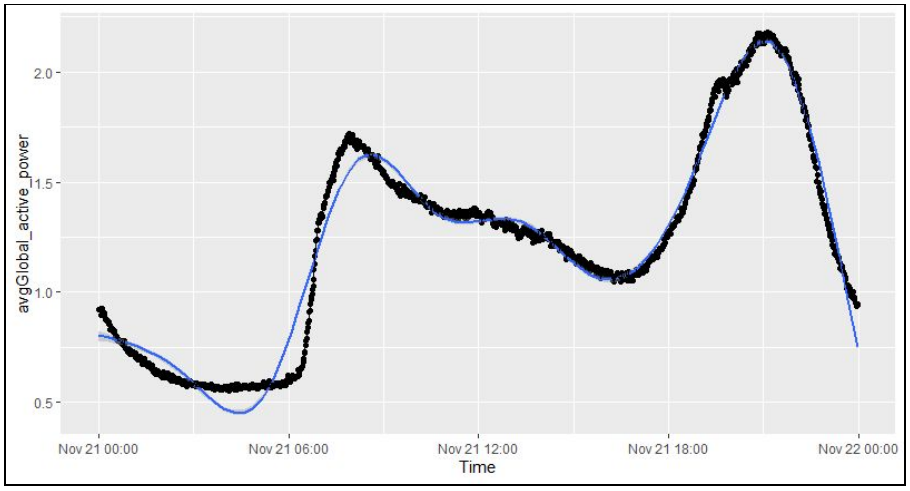


We suspected the underlying trend of dips and peaks in Figure 3.1 and 3.3 were caused by seasonal change. This is because the line of best fit experienced dips during Summer months and rose during the Winter months. By observing the best fit line these two figures, the features (global_active_power and global_intensity) seem to be highly correlated. Furthermore, the global_reactive_power seems to be inversely related to the former two features.

By observing these plots, we were interested in the first two features, global_active_power, global_reactive_power. This is because we thought based on the plots above, they shared some relationship with each other, while voltage did not.

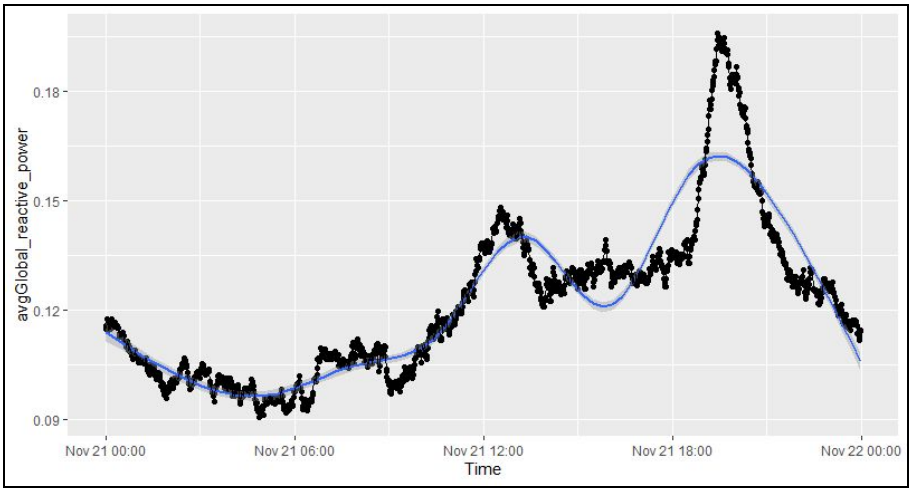
Our next steps were exploring the data to find an interesting time frame to use. Initially, we attempted to subset the data into four seasons, however, we realized that this over summarized the data. Hence, we plotted the average of the features against time in order to observe any general patterns of the features depending on time. In other words, we aggregated and averaged the features based on each minute time interval.

Figure 3.1.5 Plot of time (x-axis) vs average global_active_power (y-axis)



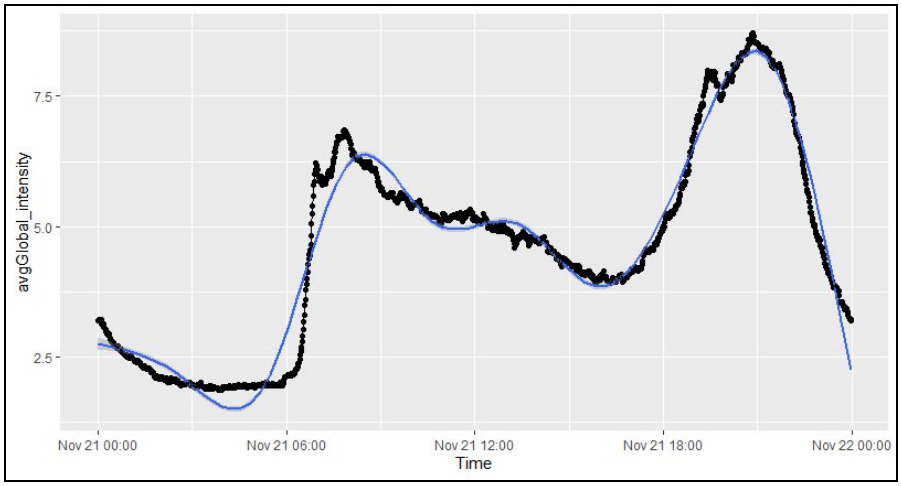
Note the x-axis labels are not an accurate representation of the graphed plot. The x-values represent time from 00:00 to 24:00

Figure 3.1.6 Plot of time (x-axis) vs average global_reactive_power (y-axis)



Note the x-axis labels are not an accurate representation of the graphed plot. The x-values represent time from 00:00 to 24:00

Figure 3.1.7 Plot of time (x-axis) vs average global_intensity



Note the x-axis labels are not an accurate representation of the graphed plot. The x-values represent time from 00:00 to 24:00

Figures 3.5 to 3.7 were very interesting since all three shared a similar trend with two peaks. However, with our initial plots we had expected the global_reactive_power to contain an inverse relation with global_active_power and global_reactive_power.

The steep peaks at approximately 06:00 and 21:00 in Figures 3.5 and 3.7 generated heavy interest. Our initial thoughts and assumptions were that 06:00 was a general waking up time for the population thus the increase in power usage as people generally use electric kitchen appliances and turn on lights once they are awake. We hypothesized that the second even higher peak at 21:30, may be a time that the general population watch television and engage in their electronic devices before going to sleep.

The following trends in the figures above interested us in subsetting the data into mornings and nights.

To investigate subsetting data into mornings and nights, we attempted to plot the averaged features against the date where the averaged features were within the morning (06:00 to 12:00) or night (18:00 to 24:00) time frame.

Figure 3.1.8 Plot of date (x-axis) vs averaged morning time global_active_power

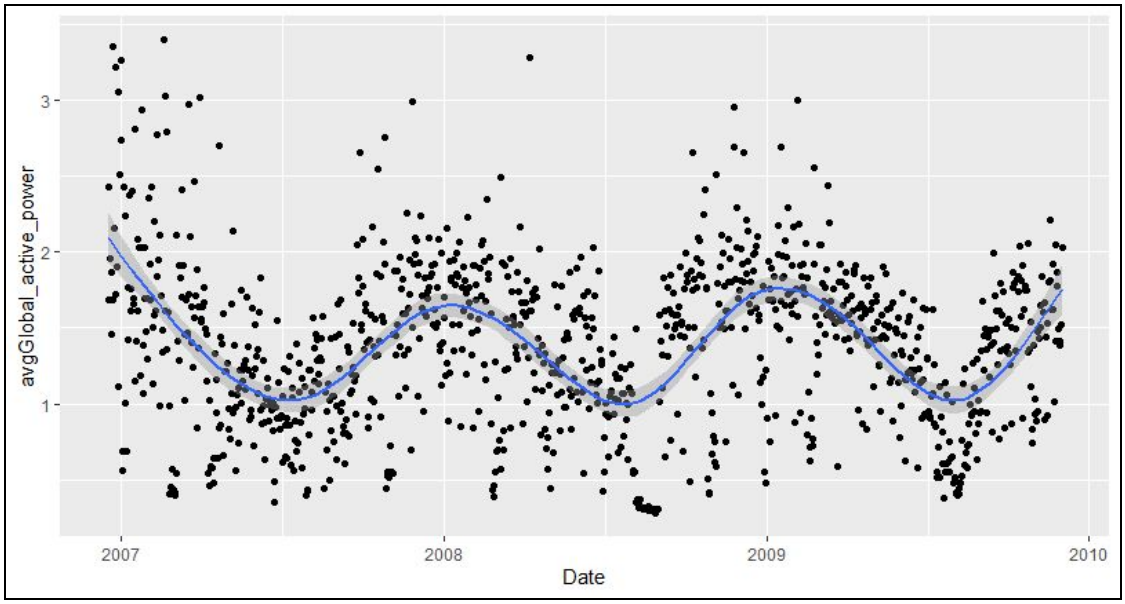


Figure 3.1.9 Plot of date (x-axis) vs averaged morning time global_reactive_power

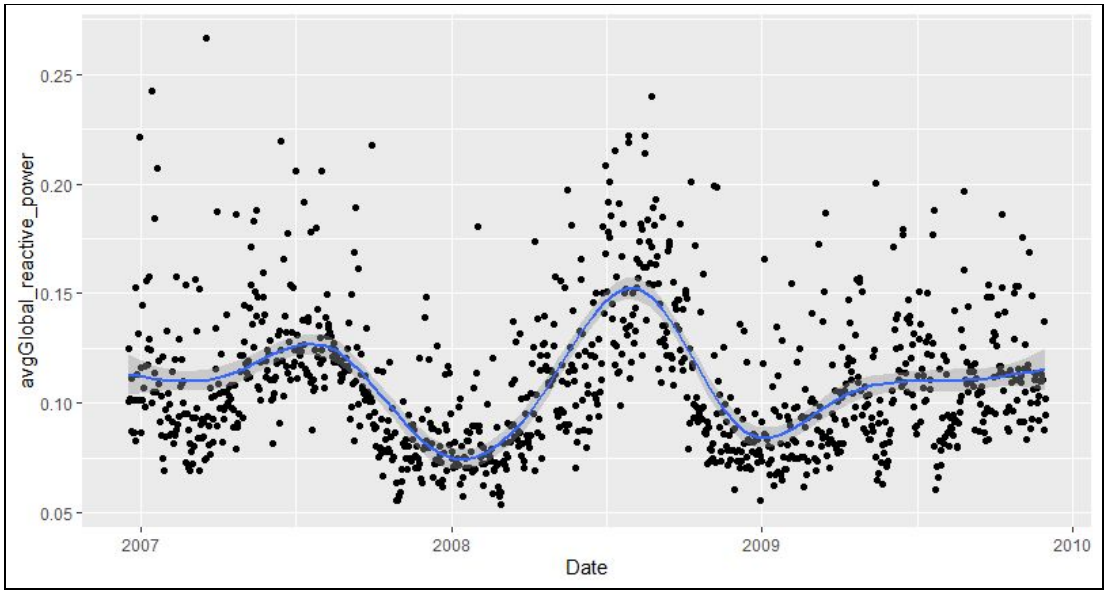
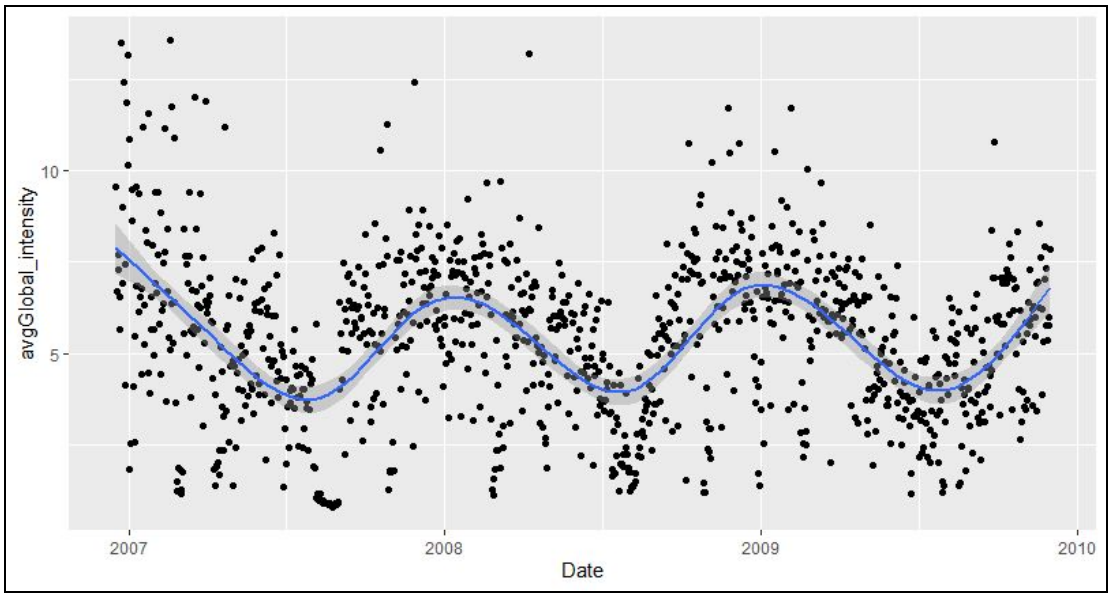


Figure 3.2.1 Plot of date (x-axis) vs averaged morning time global_intensity



We were also curious about if any differences existed in household power consumption of the days of the week (Mon-Sun). Therefore we plotted our selected features by aggregating and averaging the data into the days of the week.

For the following figures, please note that the days of the week are represented through numeric values where 0 is Sunday and 6 is Saturday.

Figure 3.2.2 Plot of day (x-axis) vs global_active_power (y-axis)

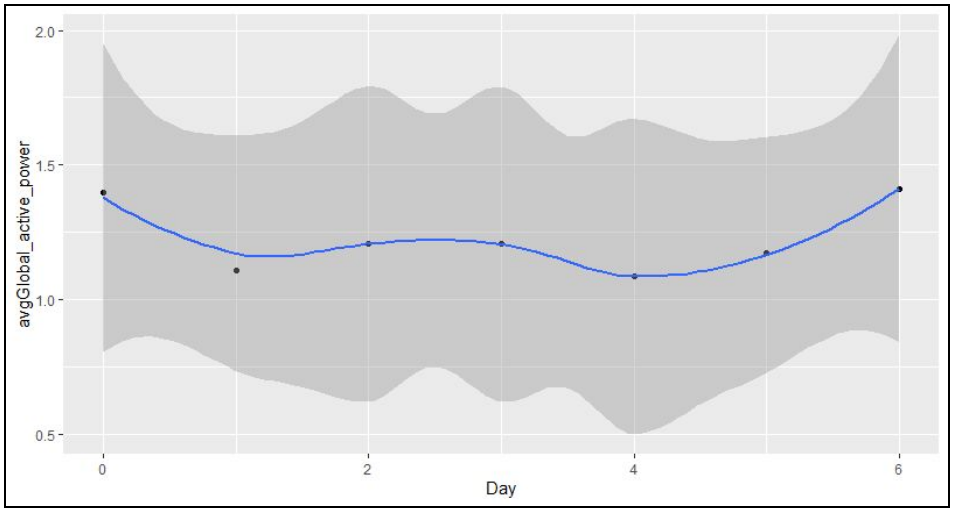


Figure 3.2.3 Plot of day (x-axis) vs global_reactive_power (y-axis)

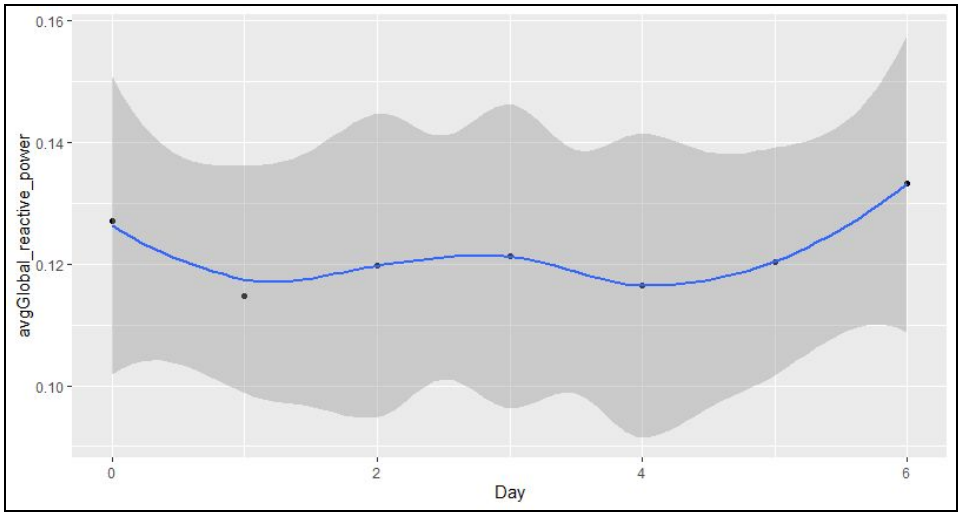
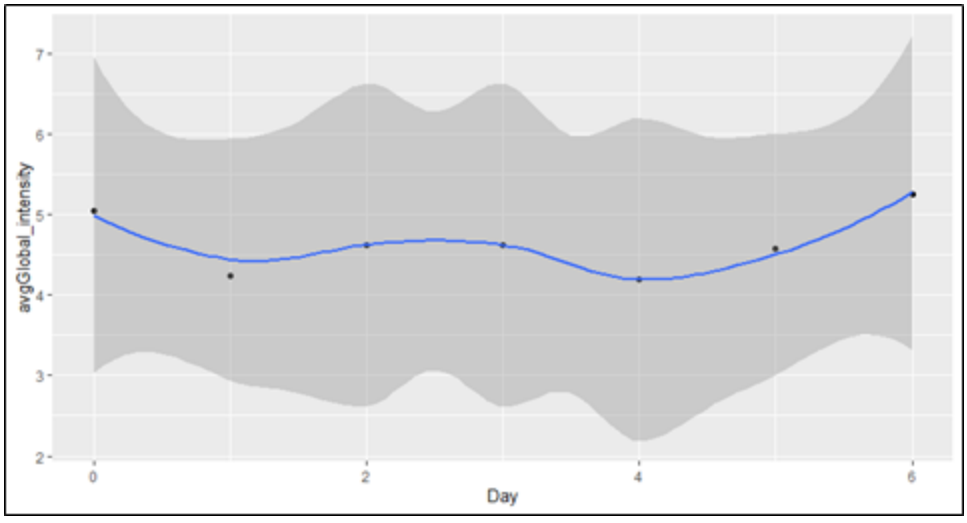


Figure 3.2.4 Plot of day (x-axis) vs global_intensity (y-axis)



Although, the following graphs showed us some differences in the power consumption for each specific day, we wanted to compare timeframes that could be more meaningful. Therefore noticing that the weekends generally corresponded to higher power consumption, we agreed to delve further in analyzing the following subsets:

- Weekday Mornings (Mon-Fri from 06:00 to 12:00)
- Weekday Nights (Mon-Fri from 18:00 to 24:00)
- Weekend Mornings (Sat-Sun from 06:00 to 12:00)
- Weekend Nights (Sat-Sun from 18:00 to 24:00)

Correlation Coefficients of Features in Train Data

Global Active Power and Global Reactive Power = **0.1226429**

Global Active Power and Global Intensity = **0.7347049**

Global Active Power and Voltage = **-0.3847512**

Global Reactive Power and Voltage = **-0.1225639**

Global Reactive Power and Global Intensity = **0.2487322**

Voltage and Global Intensity = **-0.5060405**

We observe that the correlation between Global Active Power and Global Reactive Power is weak while Global Active Power and Global Intensity have a strong linear relationship. Further we also discover that Voltage is negatively correlated with a weak value to Global Active Power and Reactive Power while it has a strong negative correlation with Global Intensity.

For univariate models, it is worthwhile to include variables with high correlation with each other as the model will explain the variance a lot better with the BIC value being less.
Though, for multivariate models it might be counter productive to include highly correlated variables as they might cause the BIC value to go up.

Comparing Training Data Statistics with Test Data Statistics

After careful consideration of time, we intend to explore the analysis of only one features which are **Global Active Power**.
In total, we ran 5 iterations of our code (because of 5 test data sets and 1 feature for each).

TRAINING DATA STATISTICS FOR GLOBAL ACTIVE POWER

1. WEEKDAY MORNINGS

STATISTIC	VALUE
MEAN	**1.33828**
STANDARD DEVIATION	**0.9547039**
MAXIMUM	**9.178**
MINIMUM	**0.078**

2. WEEKDAY NIGHTS

STATISTIC	VALUE
MEAN	**1.714151**
STANDARD DEVIATION	**1.303915**
MAXIMUM	**0.076**
MINIMUM	**10.65**

3. WEEKEND MORNINGS

STATISTIC	VALUE
MEAN	**1.384935**
STANDARD DEVIATION	**0.9609538**
MAXIMUM	**7.992**

MINIMUM	**0.078**

4. WEEKEND NIGHTS

STATISTIC	VALUE
MEAN	**1.783993**
STANDARD DEVIATION	**1.353077**
MAXIMUM	**9.666**
MINIMUM	**0.078**

TEST DATA 1 STATISTICS FOR GLOBAL ACTIVE POWER

1. WEEKDAY MORNINGS

STATISTIC	VALUE
MEAN	**1.767354**
STANDARD DEVIATION	**1.168029**
MAXIMUM	**10.82134**
MINIMUM	**-2.35546**

2. WEEKDAY NIGHTS

STATISTIC	VALUE
MEAN	**1.968807**
STANDARD DEVIATION	**1.319557**
MAXIMUM	**9.82279**
MINIMUM	**-2.33054**

3. WEEKEND MORNINGS

STATISTIC	VALUE
MEAN	**1.780838**
STANDARD DEVIATION	**1.18055**
MAXIMUM	**8.614621**
MINIMUM	**-1.999361**

4. WEEKEND NIGHTS

STATISTIC	VALUE
MEAN	**1.981679**
STANDARD DEVIATION	**1.359031**
MAXIMUM	**-2.015539**
MINIMUM	**11.67644**

TEST DATA 2 STATISTICS FOR GLOBAL ACTIVE POWER

1. WEEKDAY MORNINGS

STATISTIC	VALUE
MEAN	**1.764119**
STANDARD DEVIATION	**1.166756**
MAXIMUM	**11.00891**
MINIMUM	**-2.391212**

2. WEEKDAY NIGHTS

STATISTIC	VALUE
MEAN	**1.977447**
STANDARD DEVIATION	**1.328857**
MAXIMUM	**10.92837**
MINIMUM	**-2.29971**

3. WEEKEND MORNINGS

STATISTIC	VALUE
MEAN	**1.798599**
STANDARD DEVIATION	**1.168606**
MAXIMUM	**8.978198**
MINIMUM	**-2.052982**

4. WEEKEND NIGHTS

STATISTIC	VALUE
MEAN	**1.960694**
STANDARD DEVIATION	**1.353938**
MAXIMUM	**-2.228937**
MINIMUM	**11.26668**

Test Data 3

TEST DATA 3 STATISTICS FOR GLOBAL ACTIVE POWER

1. WEEKDAY MORNINGS

STATISTIC	VALUE
MEAN	**1.767354**
STANDARD DEVIATION	**1.168029**
MAXIMUM	**10.82134**
MINIMUM	**-2.35546**

2. WEEKDAY NIGHTS

STATISTIC	VALUE
MEAN	**1.968807**
STANDARD DEVIATION	**1.319557**
MAXIMUM	**9.82279**
MINIMUM	**-2.33054**

3. WEEKEND MORNINGS

STATISTIC	VALUE
MEAN	**1.780838**
STANDARD DEVIATION	**1.18055**
MAXIMUM	**8.614621**
MINIMUM	**-1.999361**

4. WEEKEND NIGHTS

STATISTIC	VALUE
MEAN	**1.981679**
STANDARD DEVIATION	**1.359031**
MAXIMUM	**-2.015539**
MINIMUM	**11.67644**

Test Data 4

TEST DATA 4 STATISTICS FOR GLOBAL ACTIVE POWER

1. WEEKDAY MORNINGS

STATISTIC	VALUE
MEAN	**4.414696**
STANDARD DEVIATION	**3.101496**
MAXIMUM	**30.80559**
MINIMUM	**-7.01554**

2. WEEKDAY NIGHTS

STATISTIC	VALUE
MEAN	**4.927866**
STANDARD DEVIATION	**3.512181**
MAXIMUM	**27.54501**
MINIMUM	**-6.28146**

3. WEEKEND MORNINGS

STATISTIC	VALUE
MEAN	**4.452307**
STANDARD DEVIATION	**3.149045**
MAXIMUM	**25.55583**
MINIMUM	**-5.99808**

4. WEEKEND NIGHTS

STATISTIC	VALUE
MEAN	**4.962424**
STANDARD DEVIATION	**3.610587**
MAXIMUM	**28.08875**
MINIMUM	**-5.4787**

TEST DATA 5 STATISTICS FOR GLOBAL ACTIVE POWER

1. WEEKDAY MORNINGS

STATISTIC	VALUE
MEAN	**4.422102**
STANDARD DEVIATION	**3.119735**
MAXIMUM	**30.93146**
MINIMUM	**-7.06638**

2. WEEKDAY NIGHTS

STATISTIC	VALUE
MEAN	**4.922643**
STANDARD DEVIATION	**3.513732**
MAXIMUM	**27.436**
MINIMUM	**-6.28146**

3. WEEKEND MORNINGS

STATISTIC	VALUE
MEAN	**4.447604**
STANDARD DEVIATION	**3.136112**
MAXIMUM	**25.84386**
MINIMUM	**-5.27935**

4. WEEKEND NIGHTS

STATISTIC	VALUE
MEAN	**4.957628**
STANDARD DEVIATION	**3.608183**
MAXIMUM	**29.92784**
MINIMUM	**-5.4787**

CONCLUSION FOR TEST DATA vs TRAINING DATA

We found the statistics of Test 1 data to be the same as the Test 3 data. Hence we expect same anomalies for both these data groups. Further, we also find that the Test Data Sets 4,5 are generally spread over a wider range of values. We conclude this by observing their minimum and maximum values and their respective standard deviation. While the Test Data Sets 1,2,3 are spread over a narrow range.
We also consider the fact that there are no negative instances in Training Set but there are a number of negative instances in the test data sets.

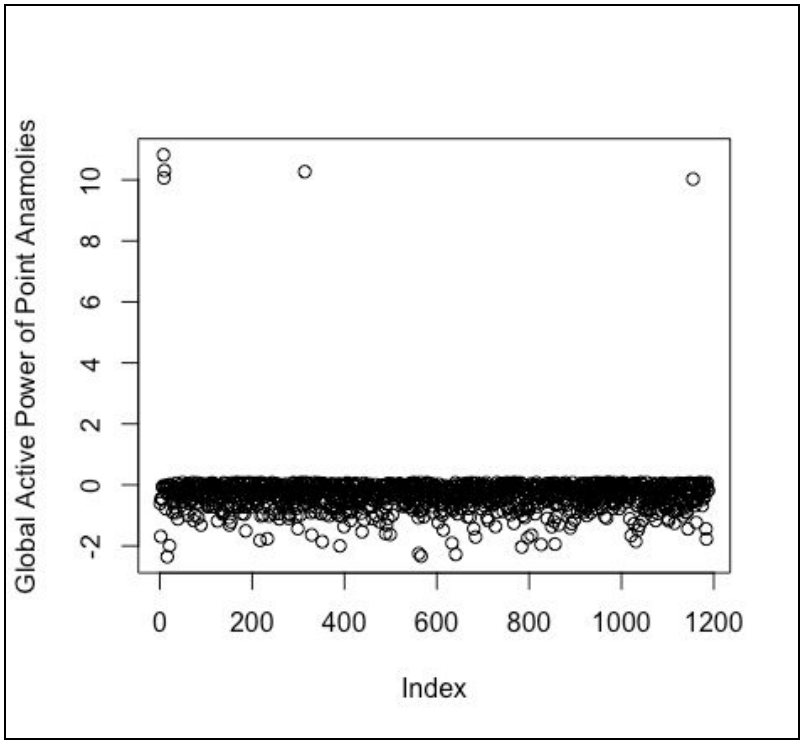
Phase 2:

Approach 1 - Part 1: Finding Point Anomalies

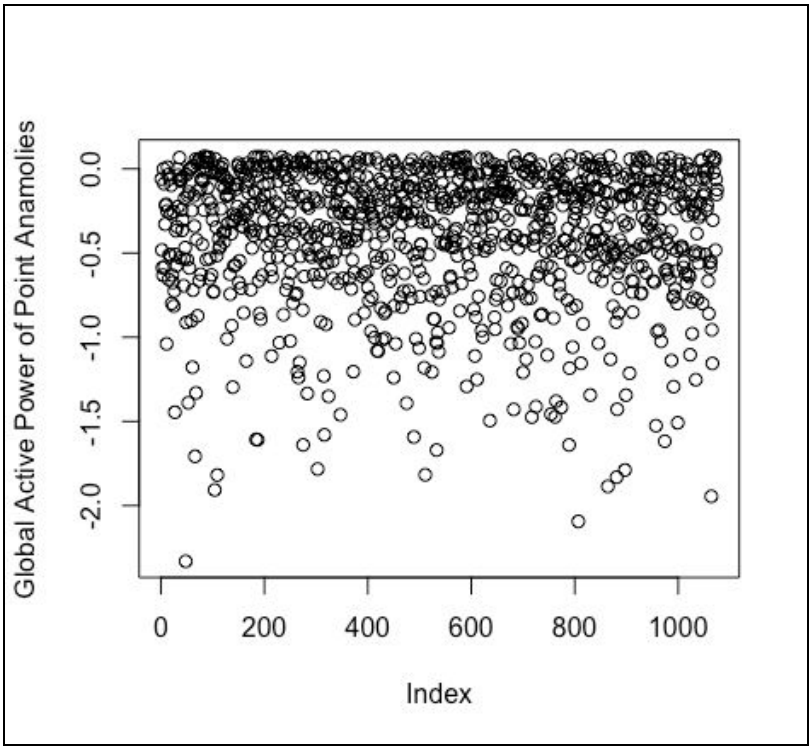
Number of Point Anomalies in Test Data 1 for Global Active Power

Time Window	No. of Point Anomalies
Weekday Mornings	1188
Weekday Nights	1075
Weekend Mornings	484
Weekend Nights	442

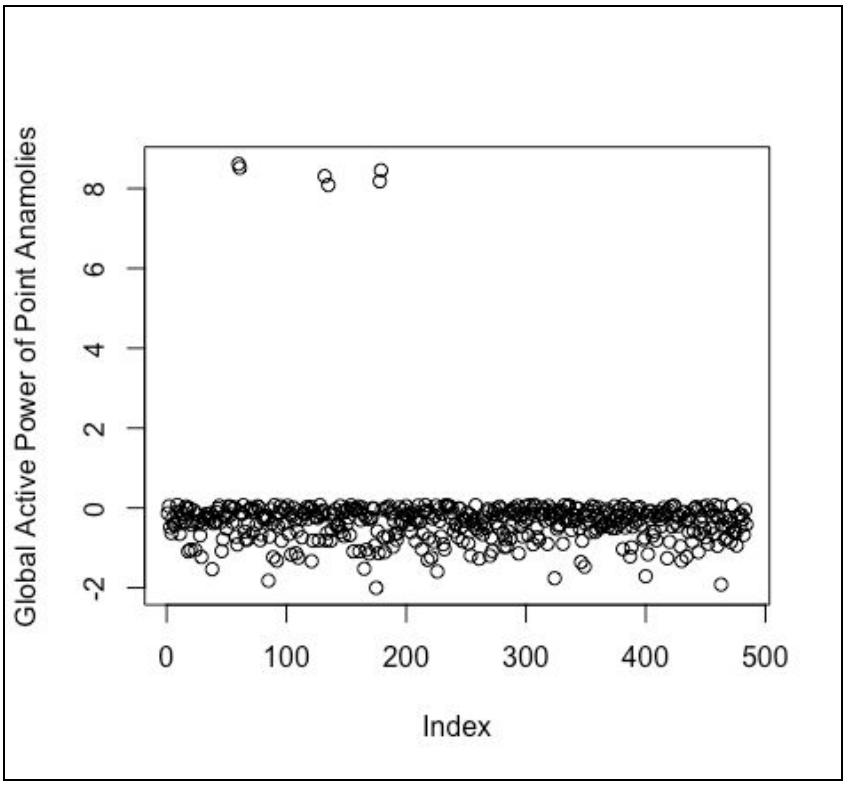
Weekday Mornings Test Data 1



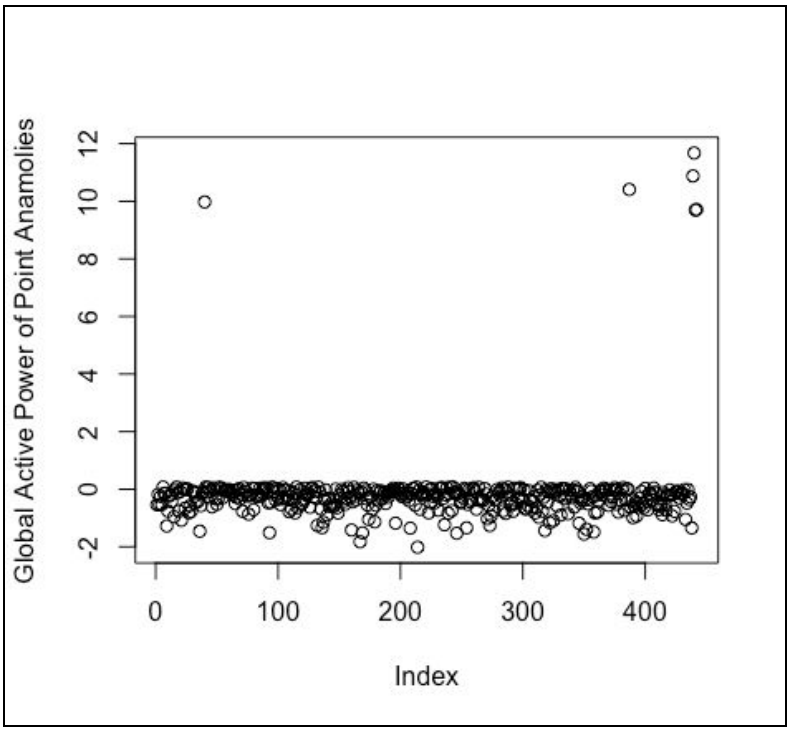
Weekday Nights Test Data 1



Weekend Mornings Test Data 1



Weekend Nights Test Data 1

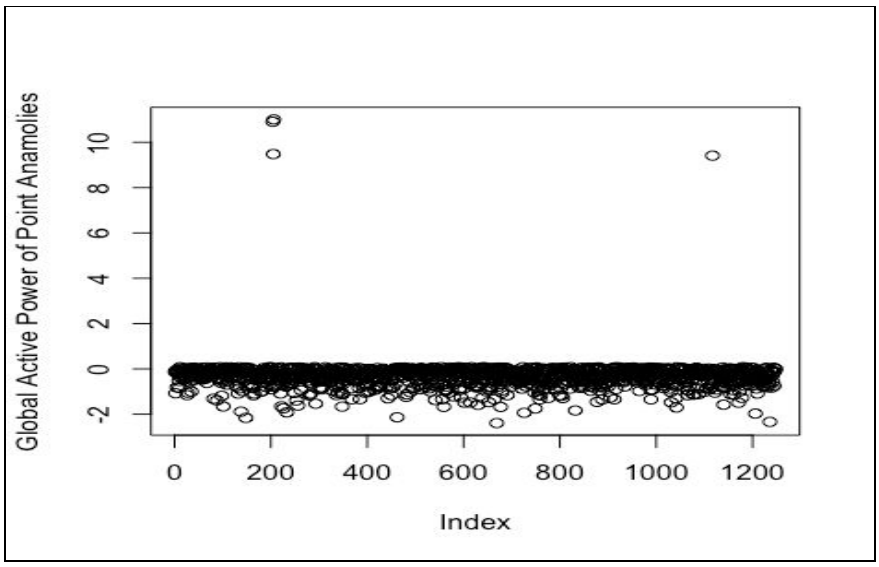


Number of Point Anomalies in Test Data 2 for Global Active Power

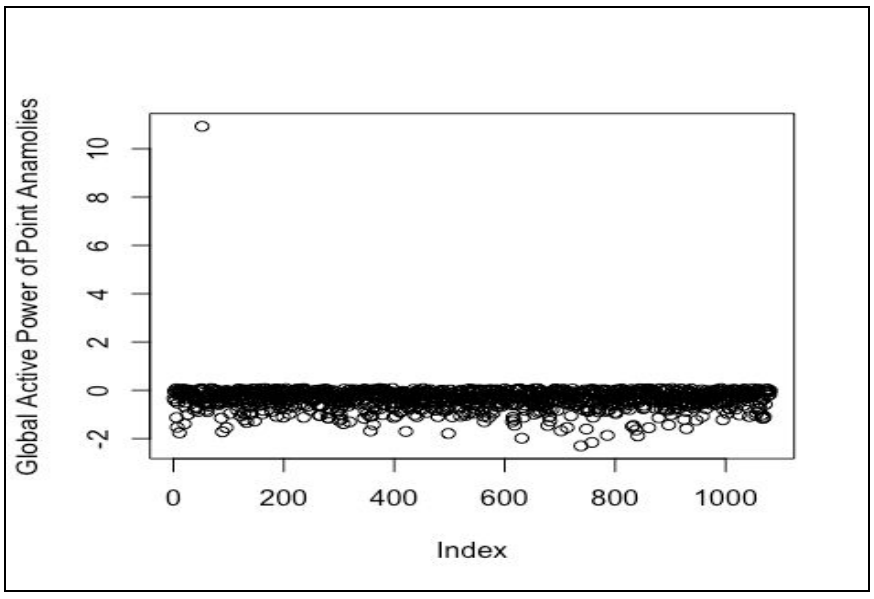
Time Window	No. of Point Anomalies
Weekday Mornings	1248
Weekday Nights	1081
Weekend Mornings	460
Weekend Nights	441

?

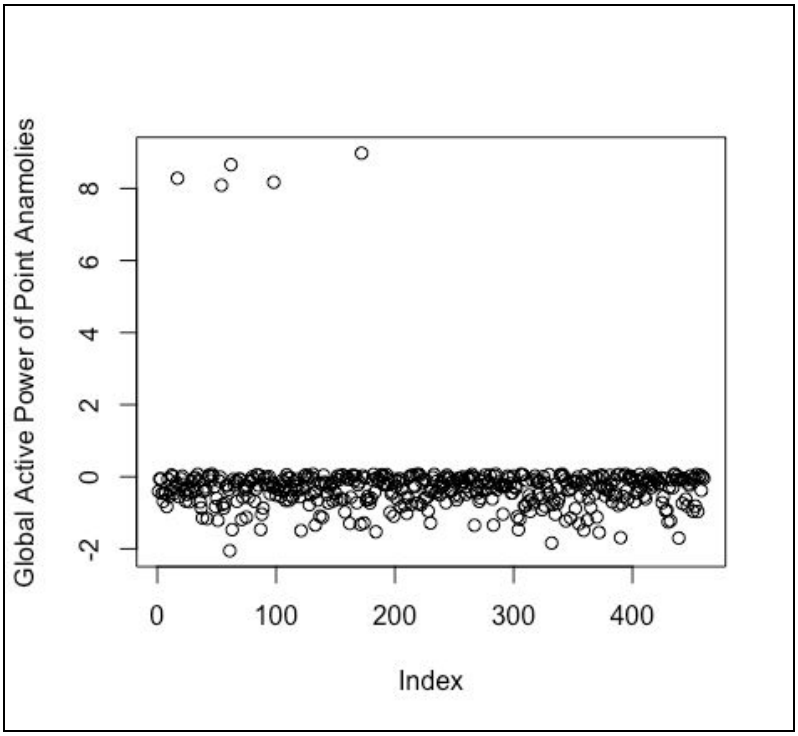
Weekday Mornings Test Data 2



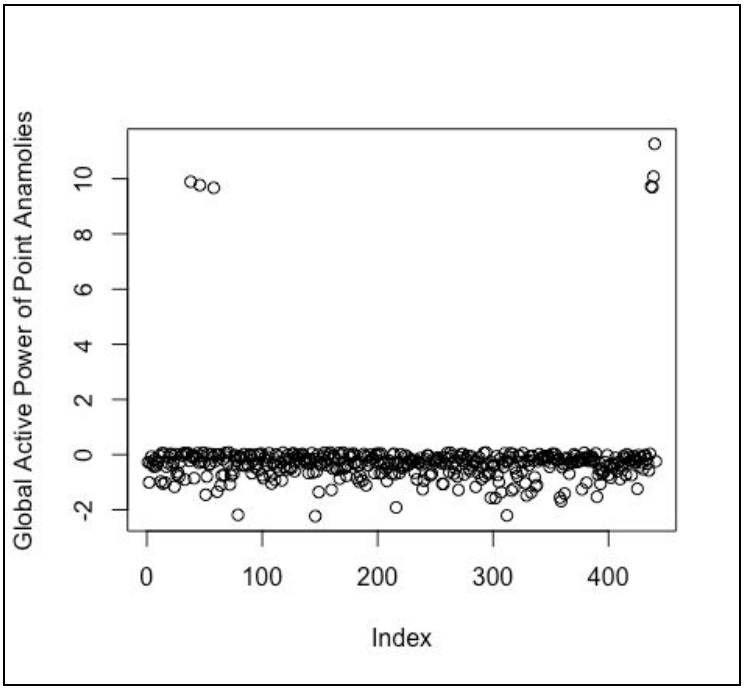
Weekday Nights Test Data 2



Weekend Mornings Test Data 2



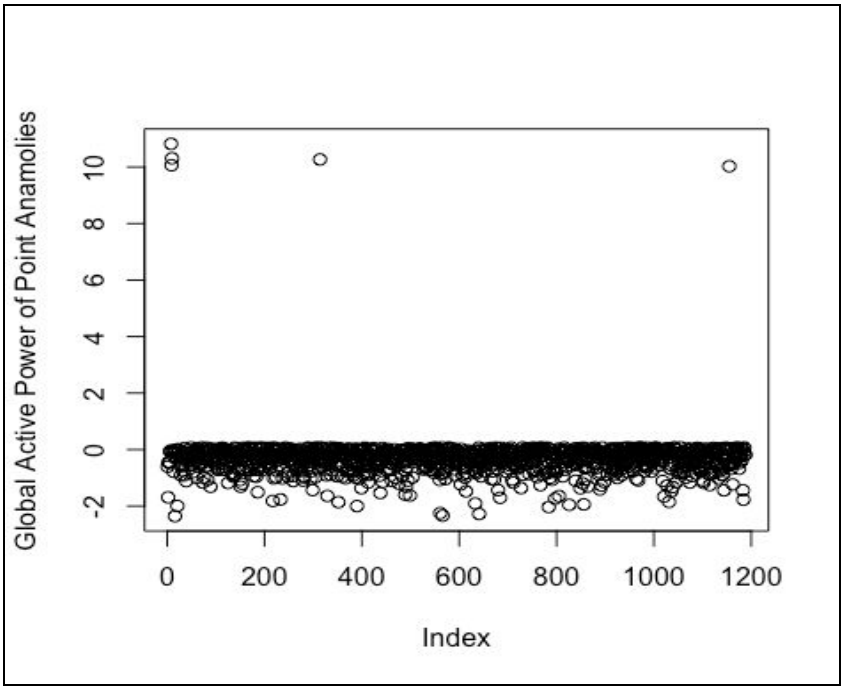
Weekend Nights Test Data 2



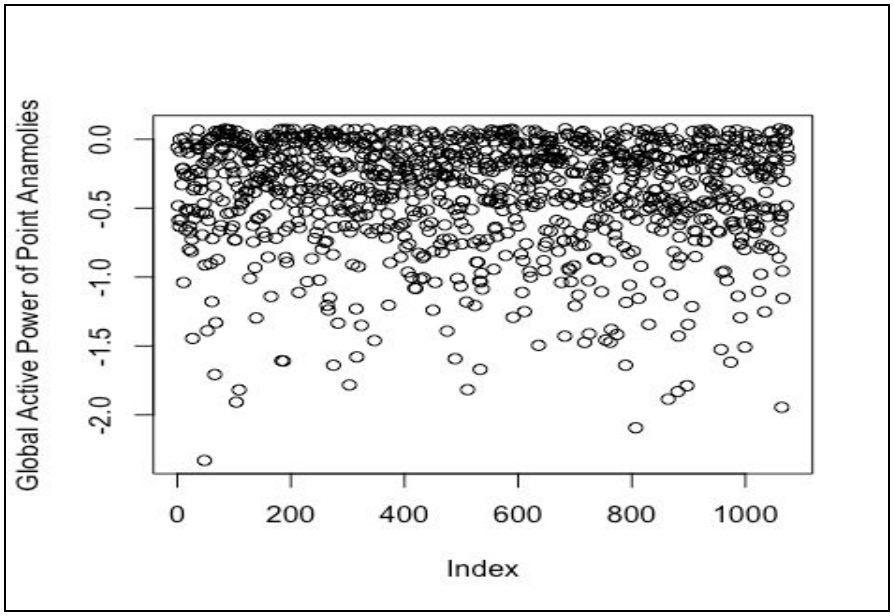
Number of Point Anomalies in Test Data 3 for Global Active Power

Time Window	No. of Point Anomalies
Weekday Mornings	1188
Weekday Nights	1075
Weekend Mornings	484
Weekend Nights	442

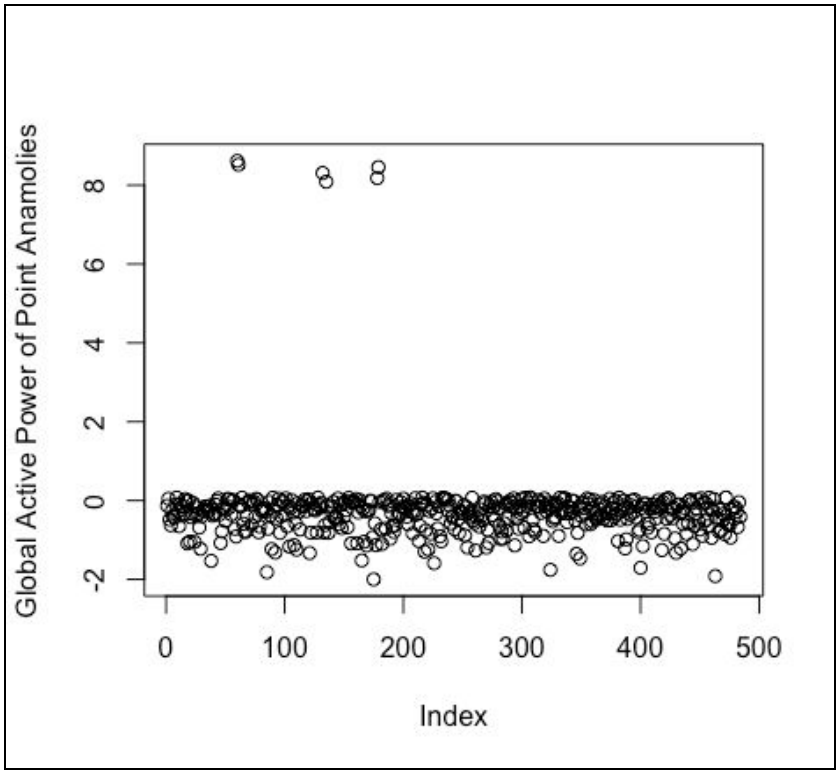
Weekday Mornings Test Data 3



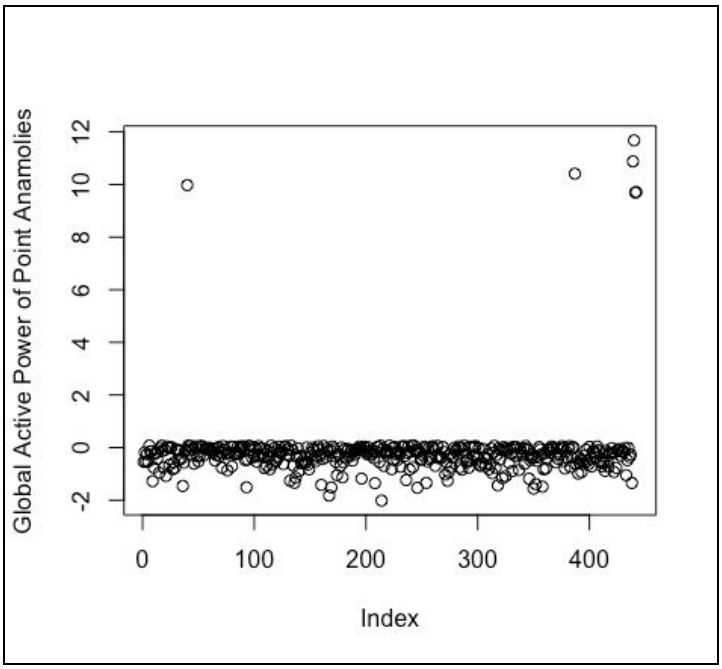
Weekday Nights Test Data 3



Weekend Mornings Test Data 3



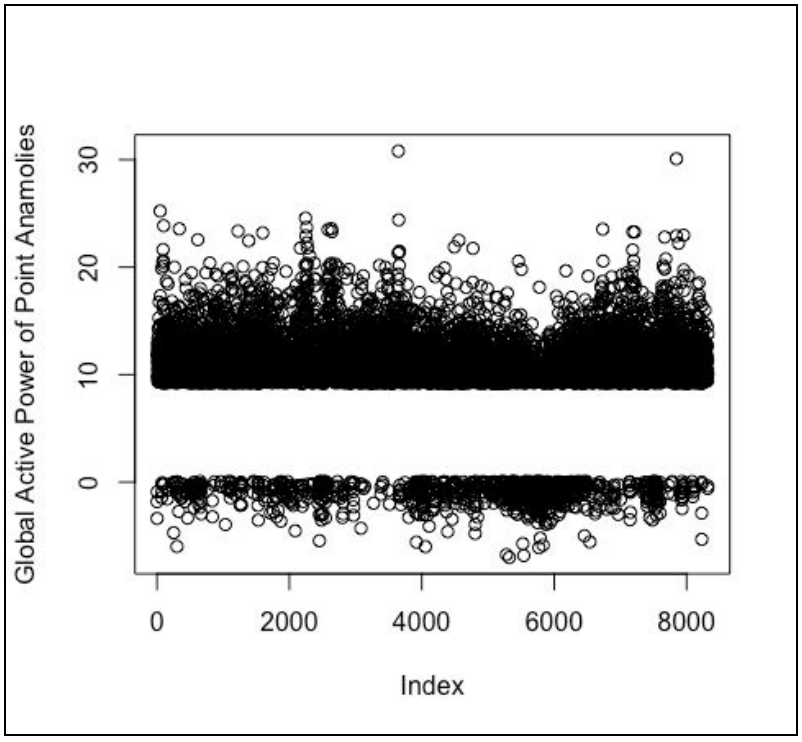
Weekend Nights Test Data 3



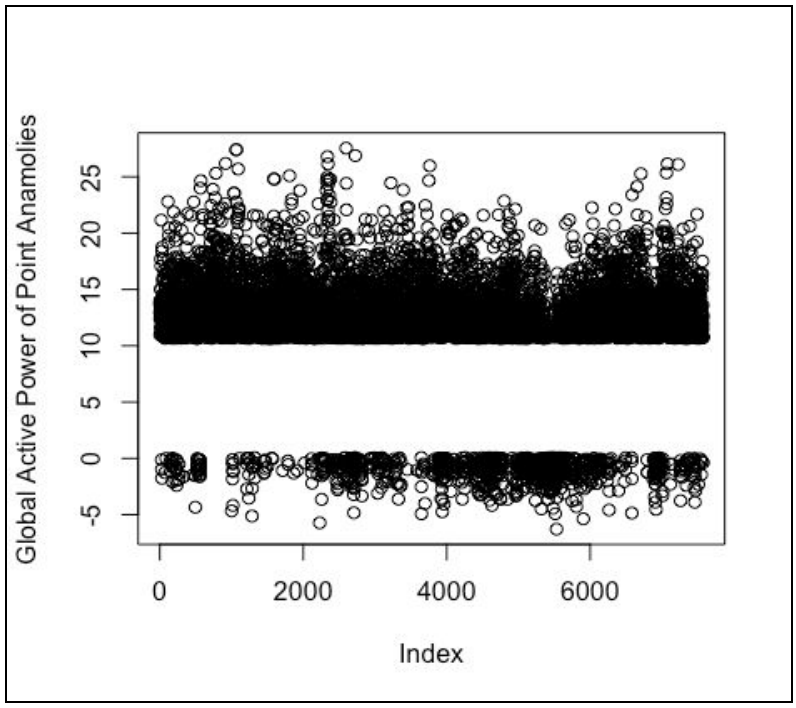
Number of Point Anomalies in Test Data 4 for Global Active Power

Time Window	No. of Point Anomalies
Weekday Mornings	8321
Weekday Nights	7574
Weekend Mornings	5147
Weekend Nights	4254

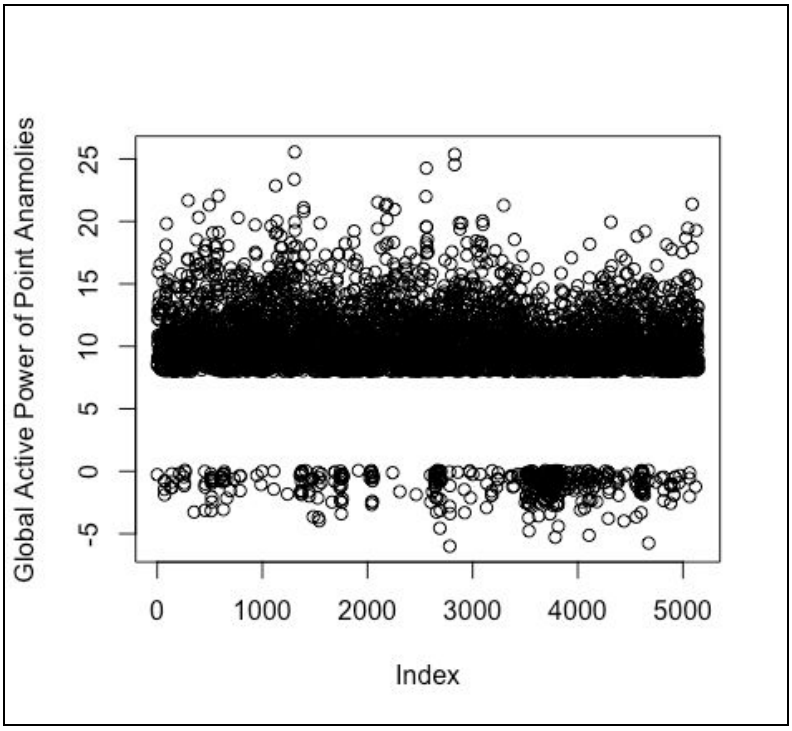
Weekday Mornings Test Data 4



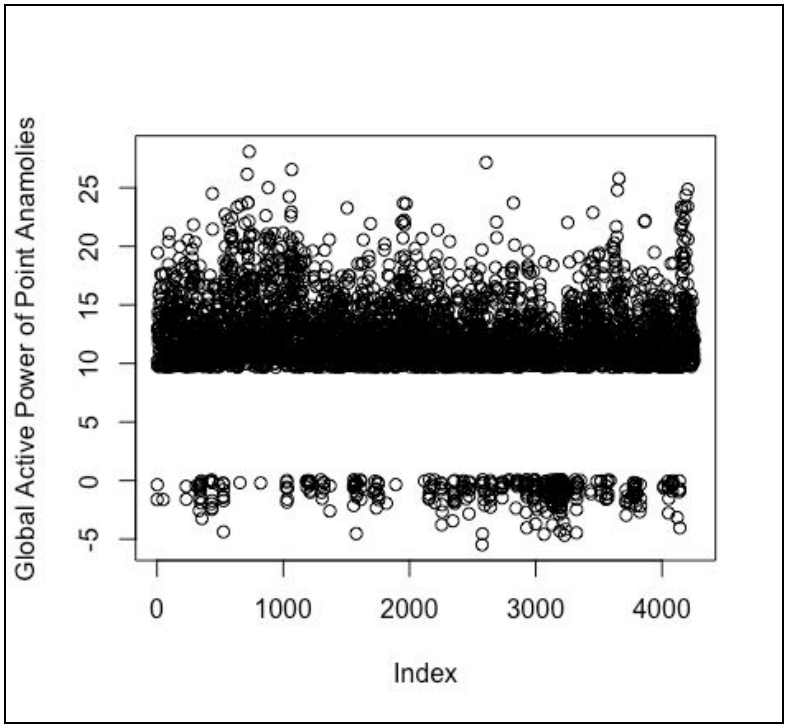
Weekday Nights Test Data 4



Weekend Mornings Test Data 4



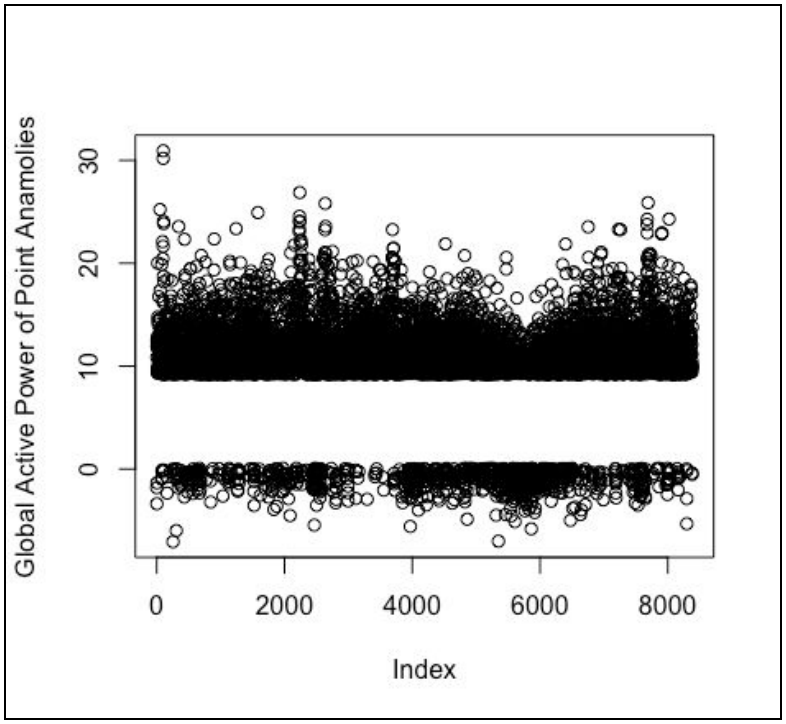
Weekend Nights Test Data 4



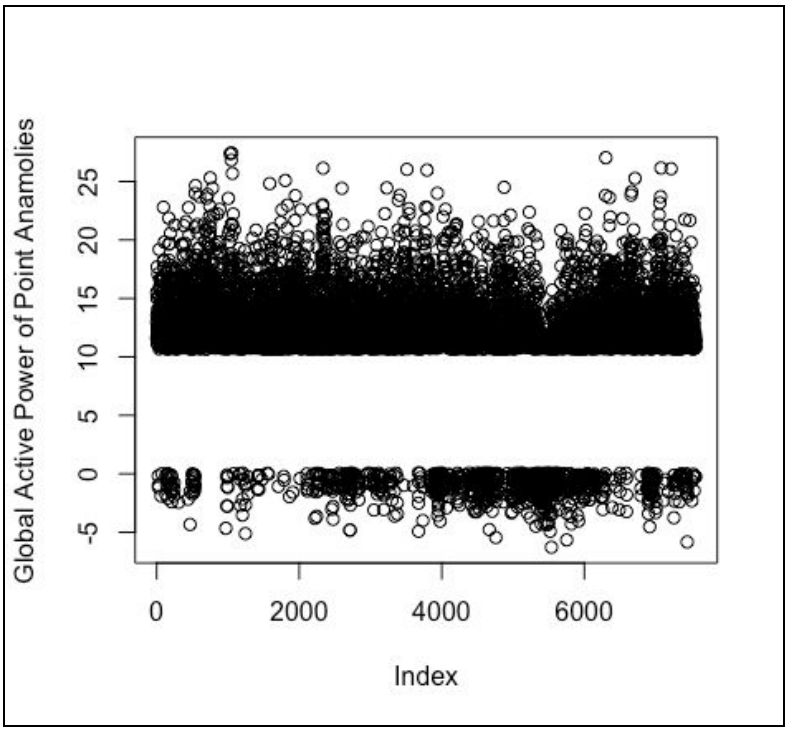
Number of Point Anomalies in Test Data 5 for Global Active Power

Time Window	No. of Point Anomalies
Weekday Mornings	8383
Weekday Nights	7563
Weekend Mornings	5205
Weekend Nights	4235

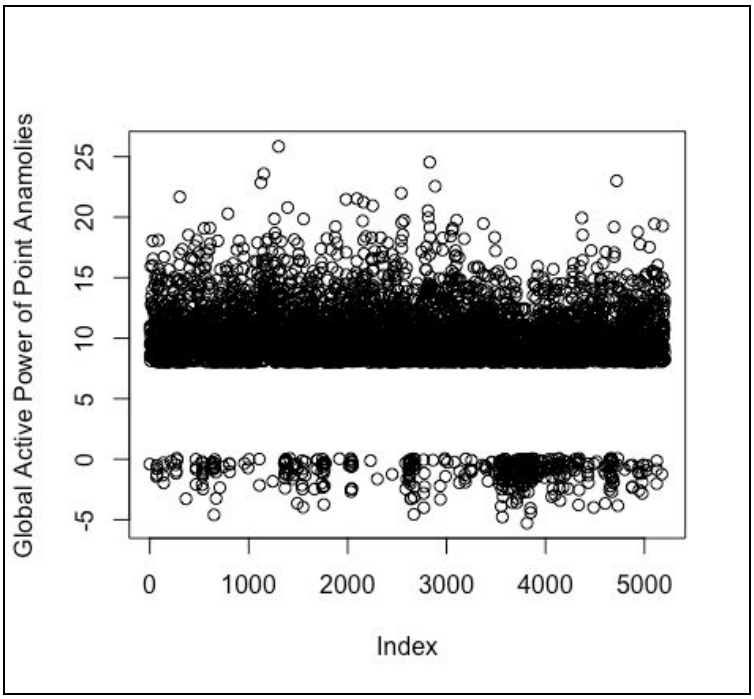
Weekday Mornings Test Data 5



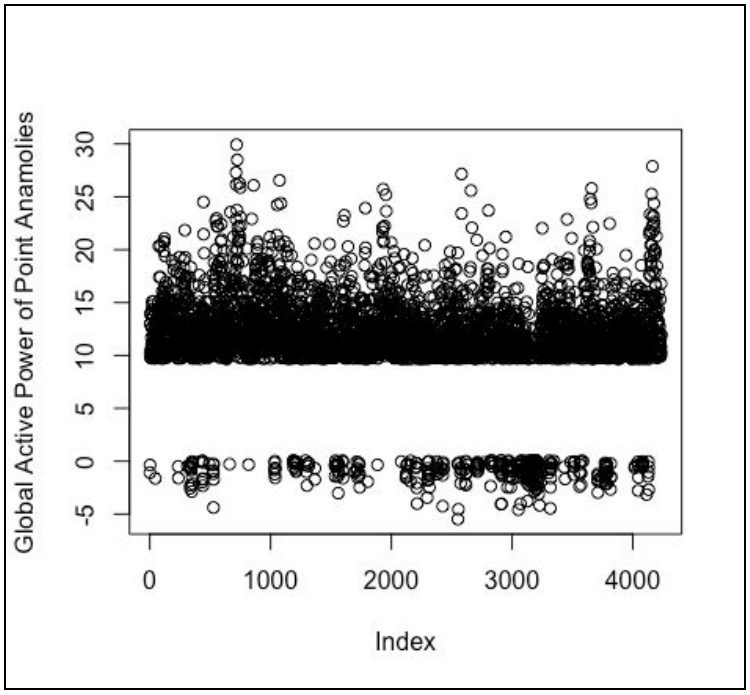
Weekday Nights Test Data 5



Weekend Mornings Test Data 5



Weekend Nights Test Data 5



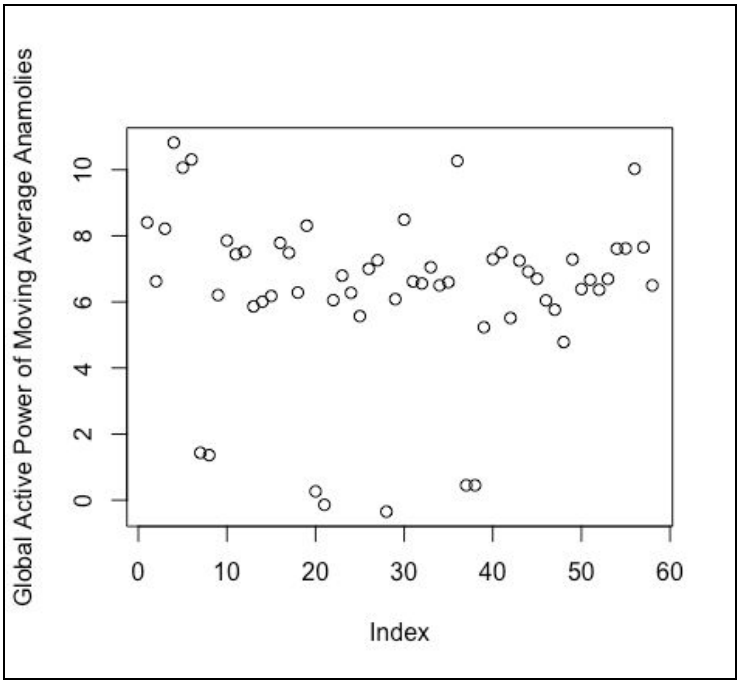
Approach 1 - Part 2: Finding Moving Average Anomalies

Threshold = 3 Times the Mean of Global Active Power in Training Data

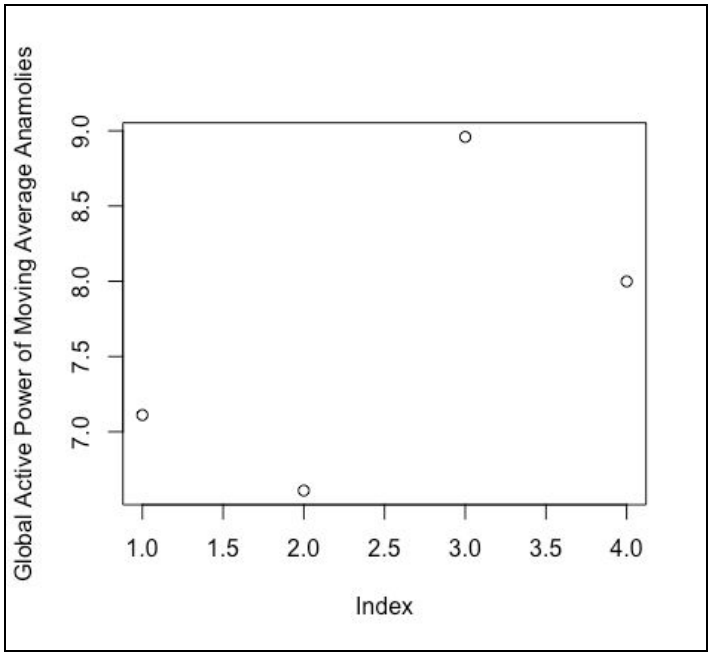
Number of Moving Average Anomalies in Test Data 1 for Global Active Power

Time Window	No. of Moving Average Anomalies
Weekday Mornings	58
Weekday Nights	4
Weekend Mornings	15
Weekend Nights	5

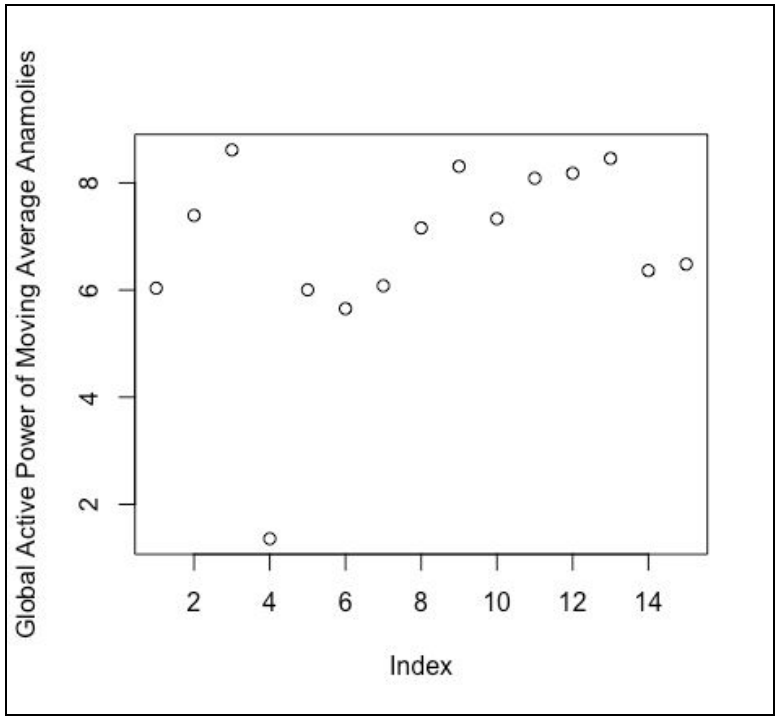
Weekday Mornings Test Data 1



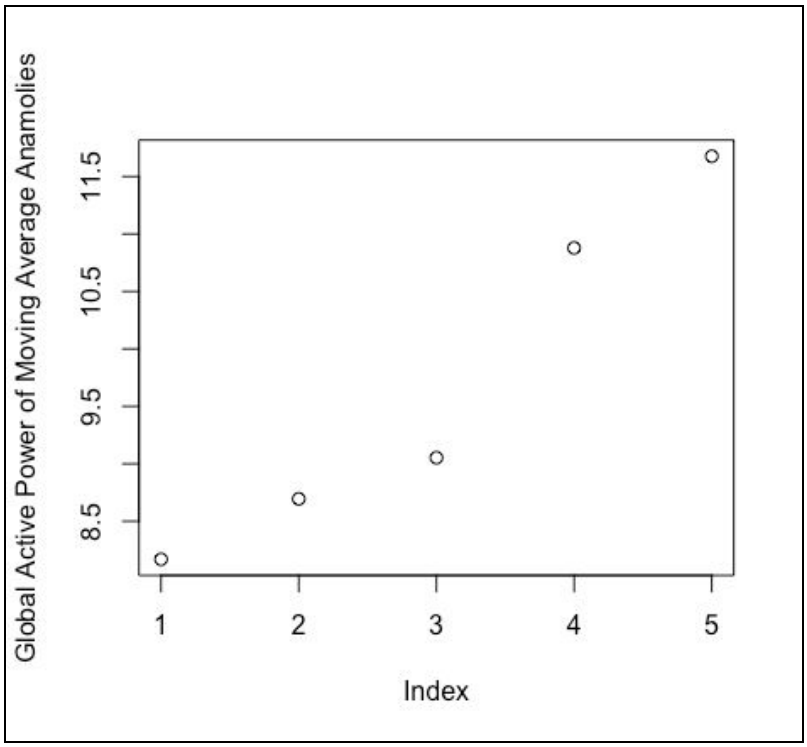
Weekday Nights Test Data 1



Weekend Mornings Test Data 1



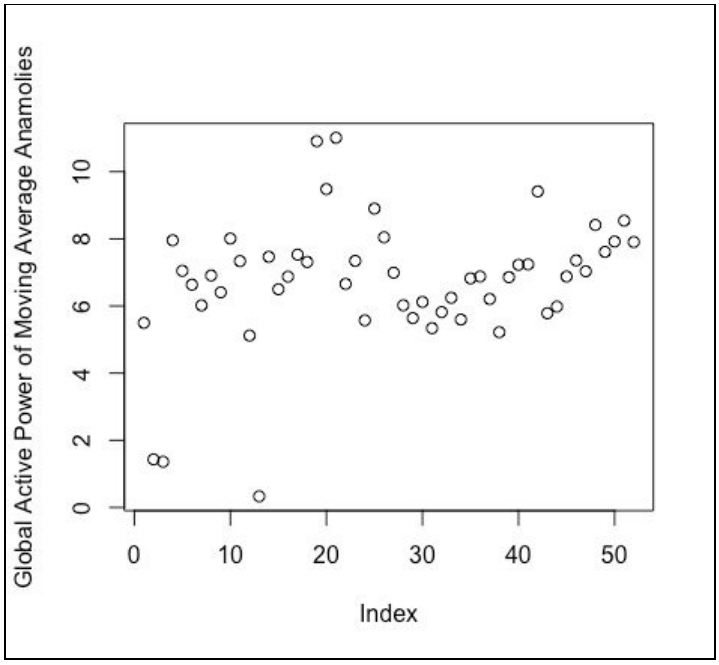
Weekend Nights Test Data 1



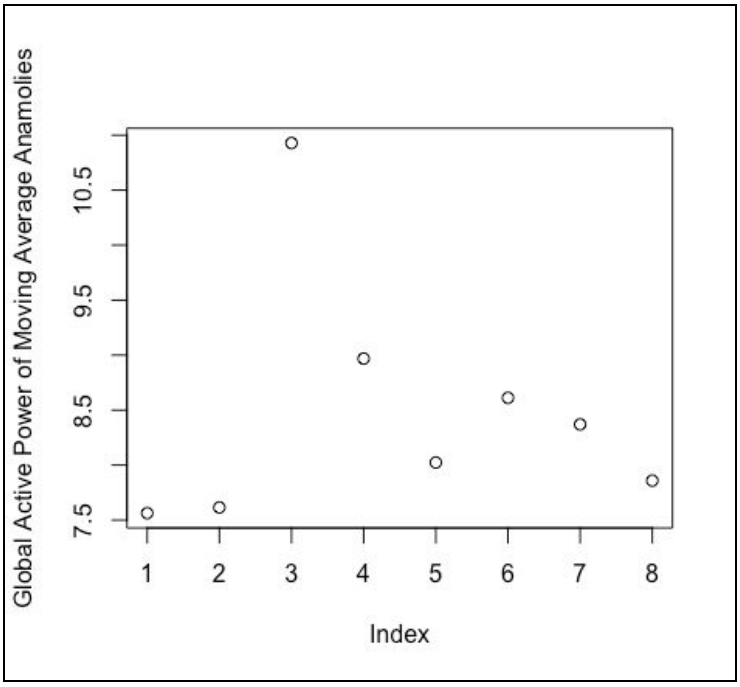
Number of Moving Average Anomalies in Test Data 2 for Global Active Power

Time Window	No. of Moving Average Anomalies
Weekday Mornings	52
Weekday Nights	8
Weekend Mornings	8
Weekend Nights	4

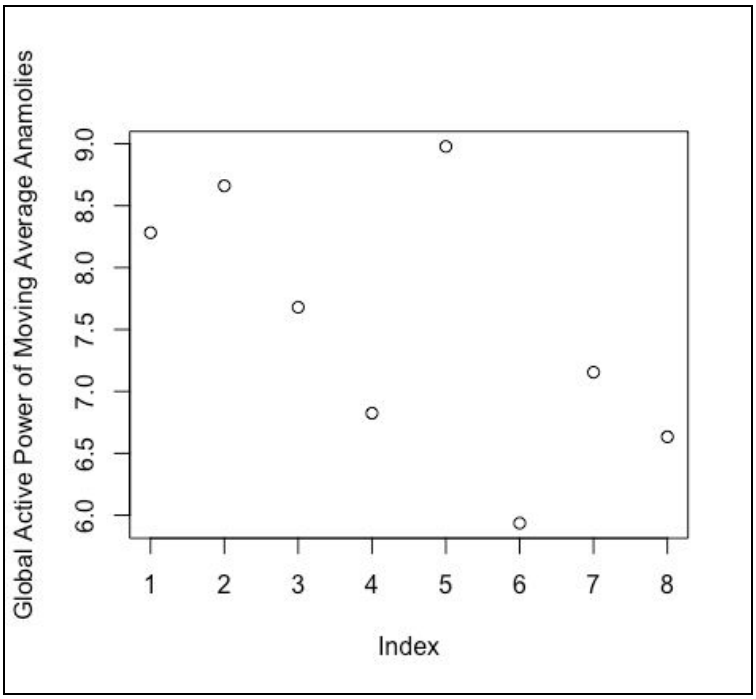
Weekday Mornings Test Data 2



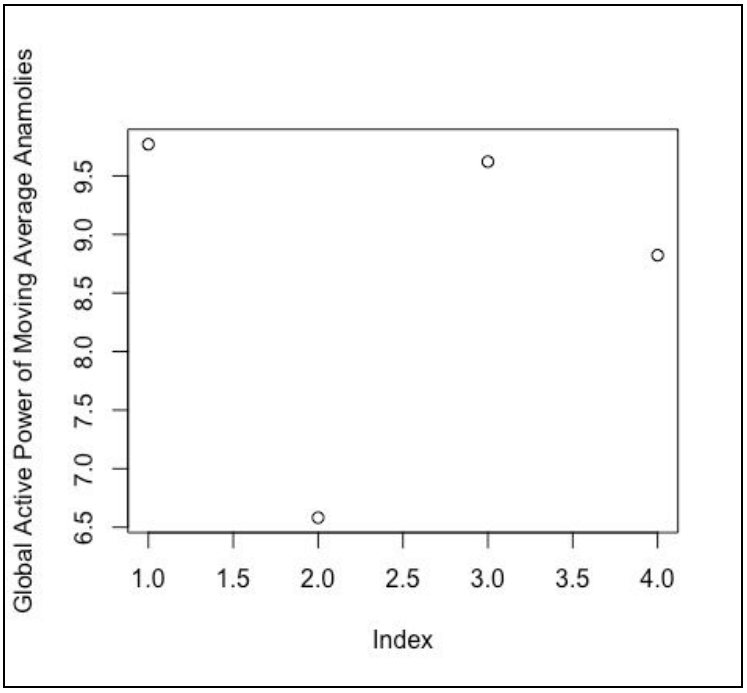
Weekday Nights Test Data 2



Weekend Mornings Test Data 2



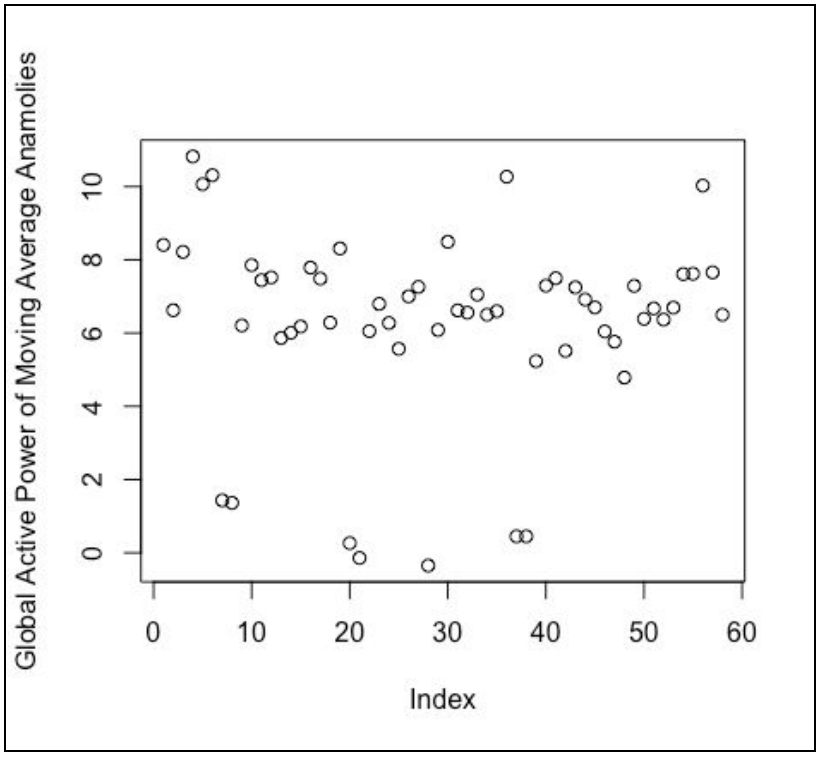
Weekend Nights Test Data 2



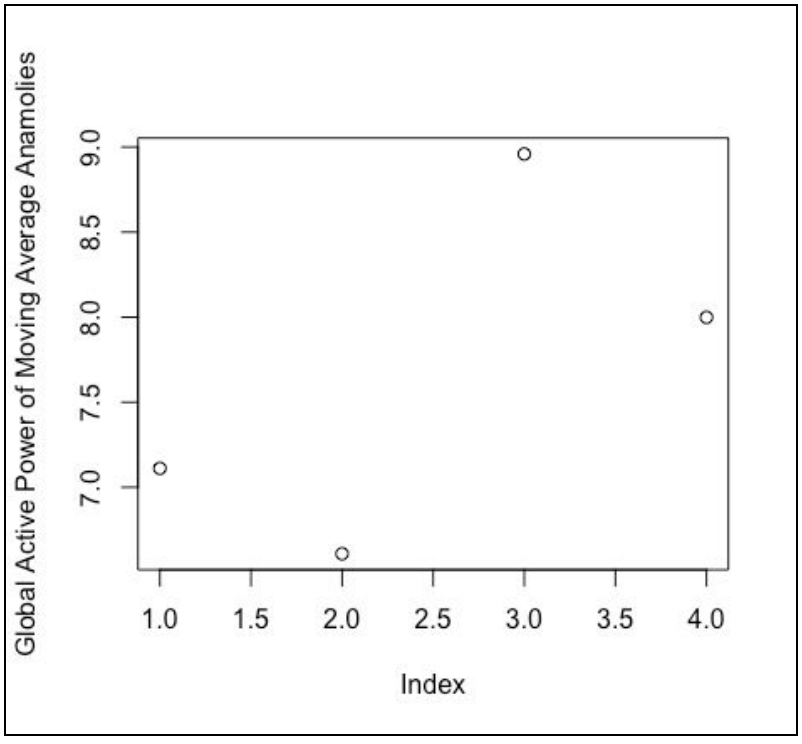
Number of Moving Average Anomalies in Test Data 3 for Global Active Power

Time Window	No. of Moving Average Anomalies
Weekday Mornings	58
Weekday Nights	4
Weekend Mornings	15
Weekend Nights	5

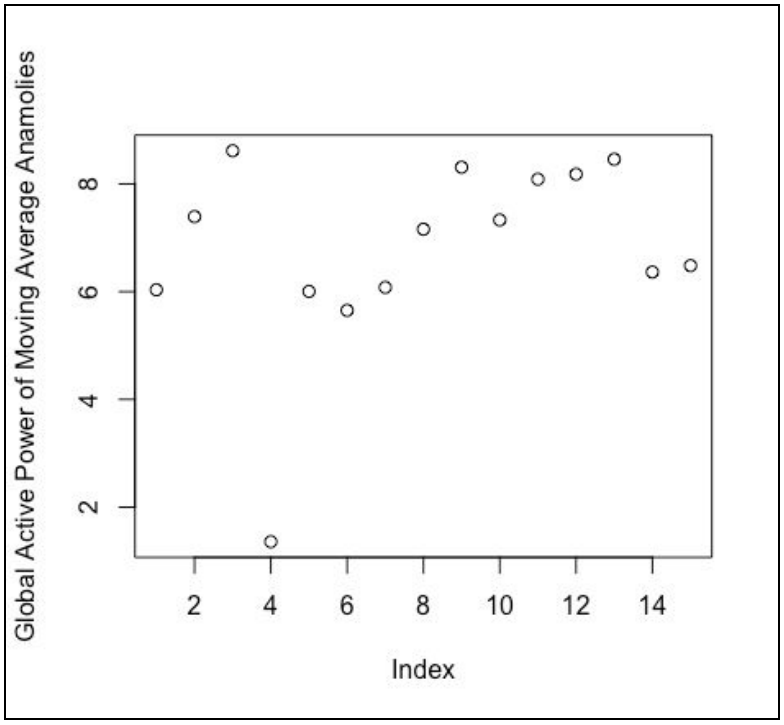
Weekday Mornings Test Data 3



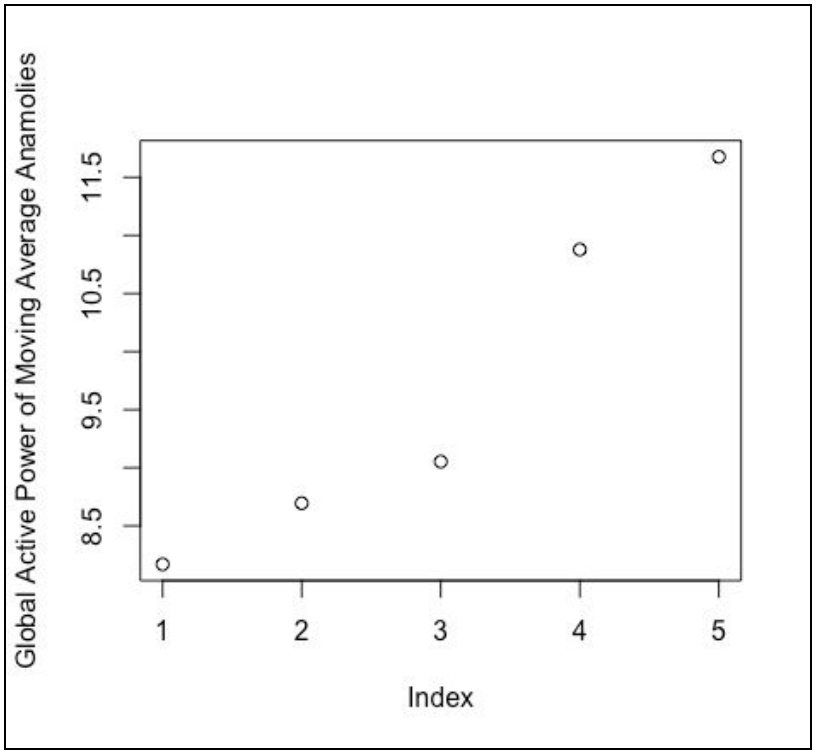
Weekday Nights Test Data 3



Weekend Mornings Test Data 3



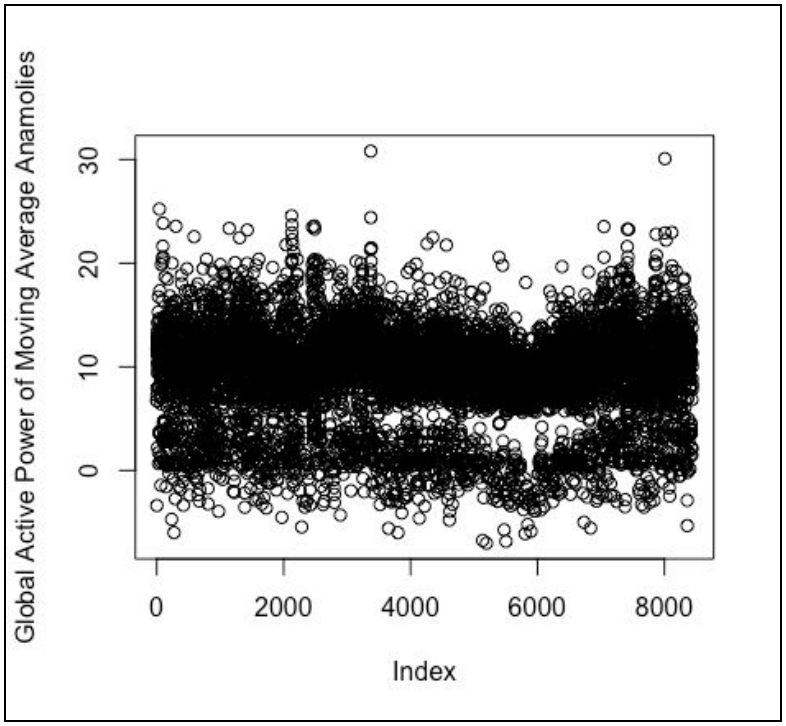
Weekend Nights Test Data 3



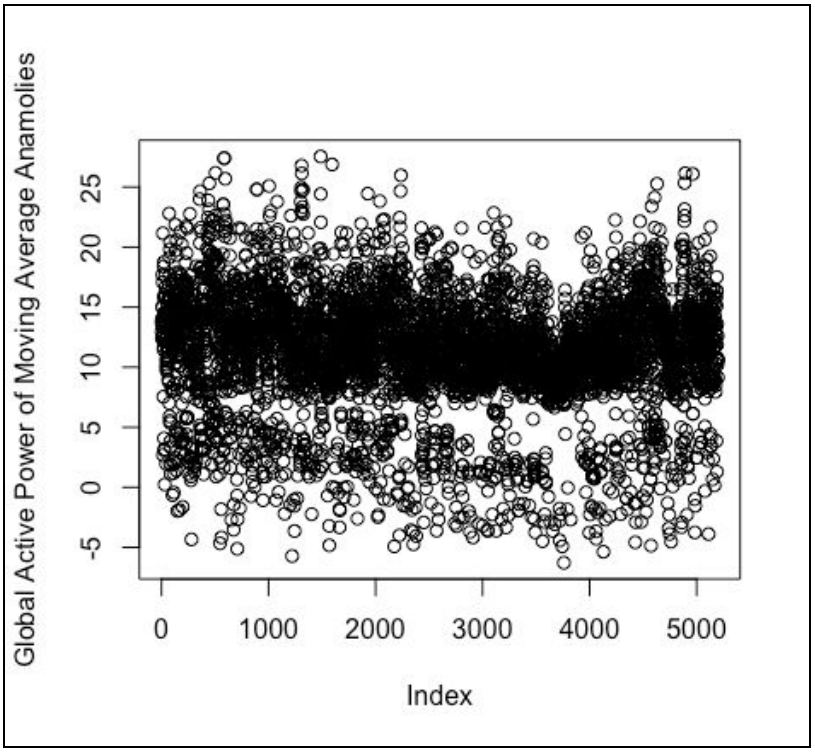
Number of Moving Average Anomalies in Test Data 4 for Global Active Power

Time Window	No. of Moving Average Anomalies
Weekday Mornings	8440
Weekday Nights	5199
Weekend Mornings	3139
Weekend Nights	1818

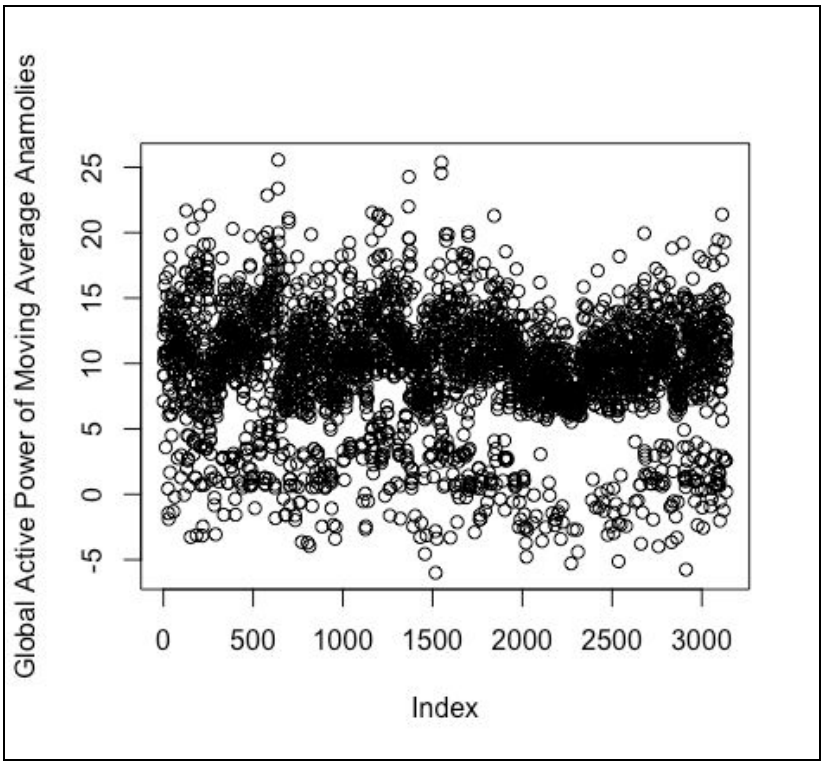
Weekday Mornings Test Data 4



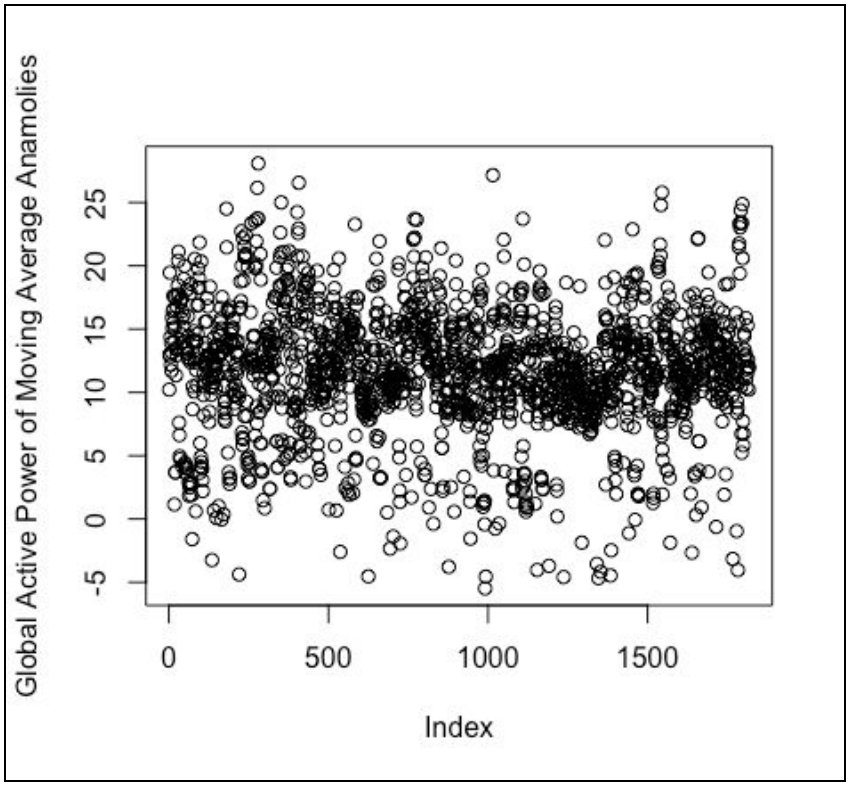
Weekday Nights Test Data 4



Weekend Mornings Test Data 4



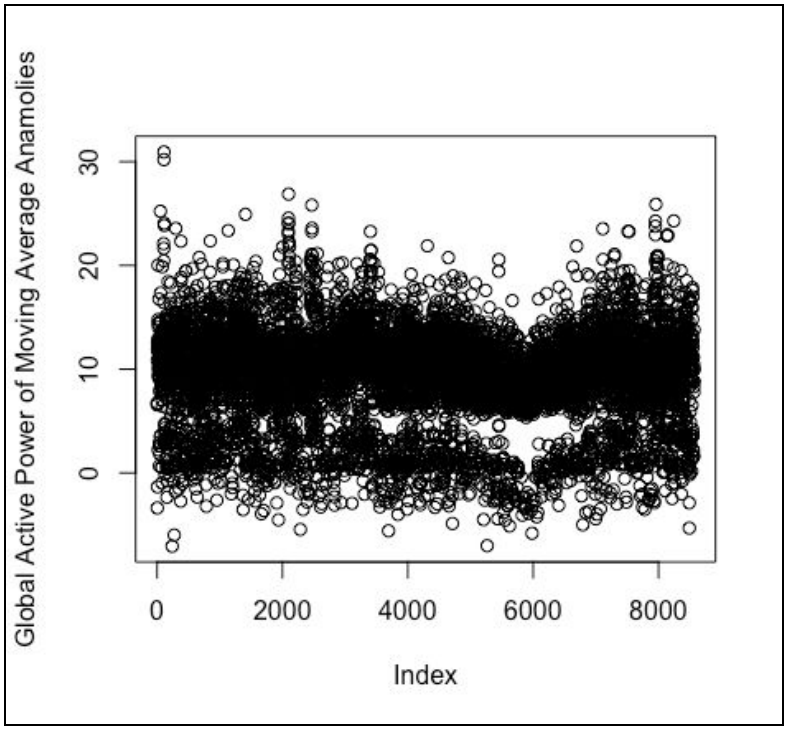
Weekend Nights Test Data 4



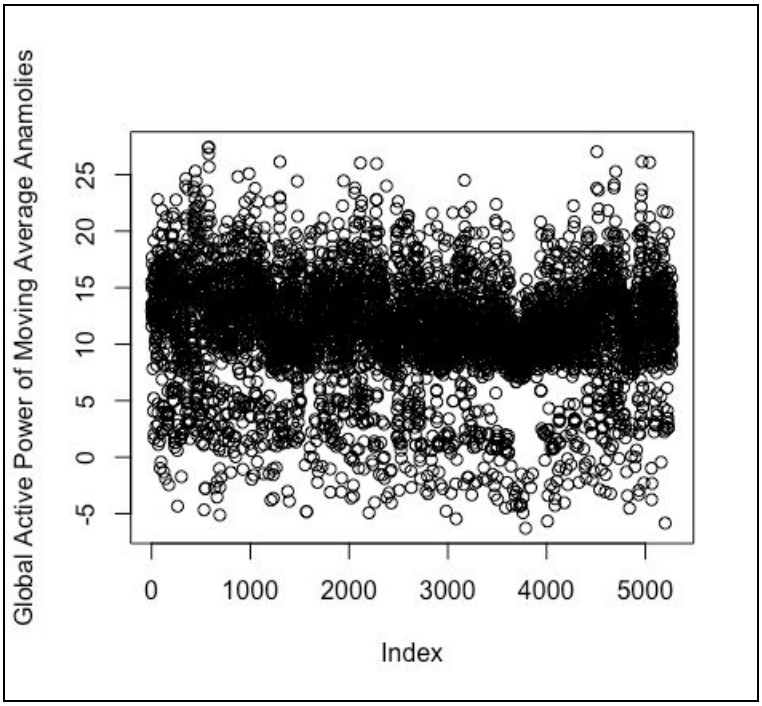
Number of Moving Average Anomalies in Test Data 5 for Global Active Power

Time Window	No. of Moving Average Anomalies
Weekday Mornings	8567
Weekday Nights	5277
Weekend Mornings	3148
Weekend Nights	1821

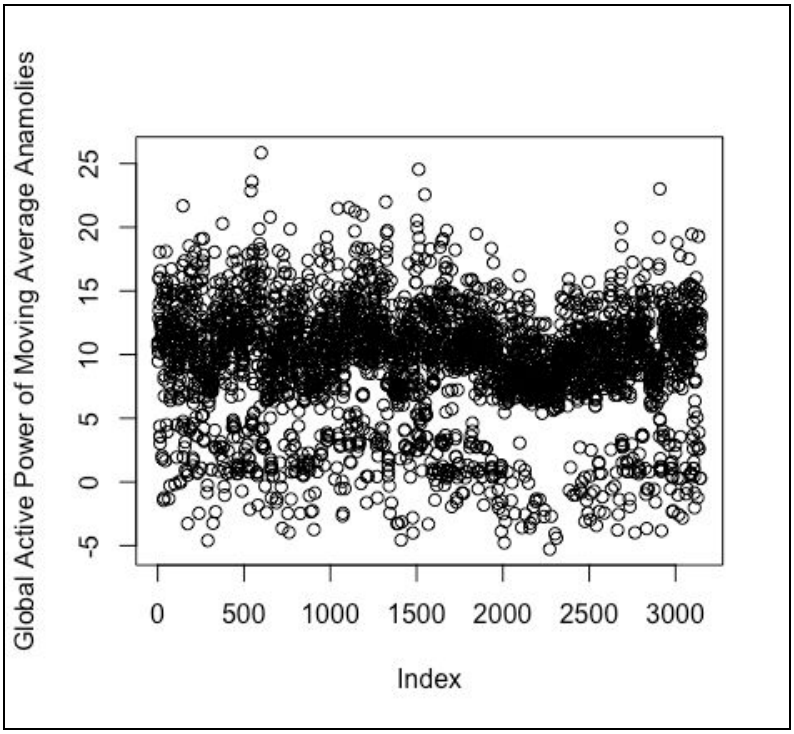
Weekday Mornings Test Data 5



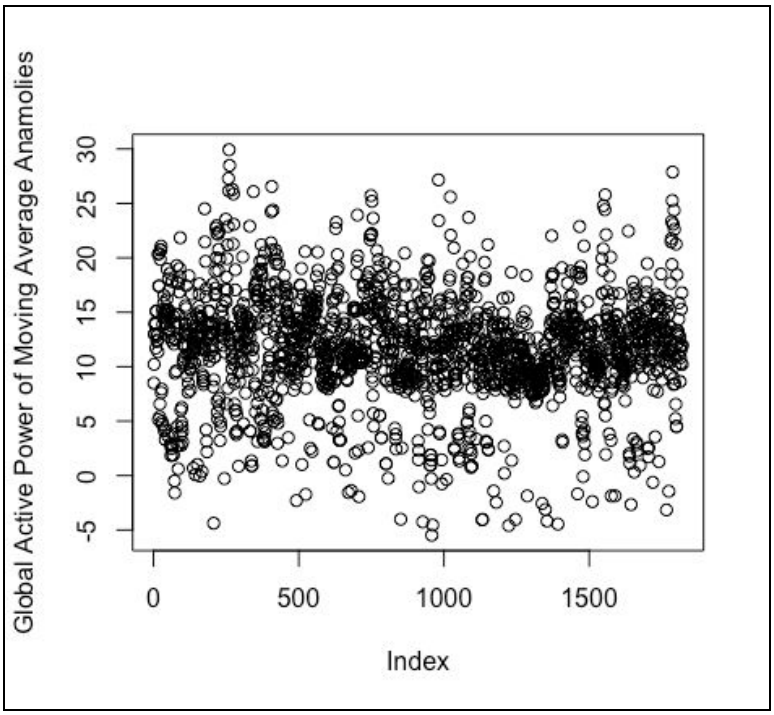
Weekday Nights Test Data 5



Weekend Mornings Test Data 5



Weekend Nights Test Data 5

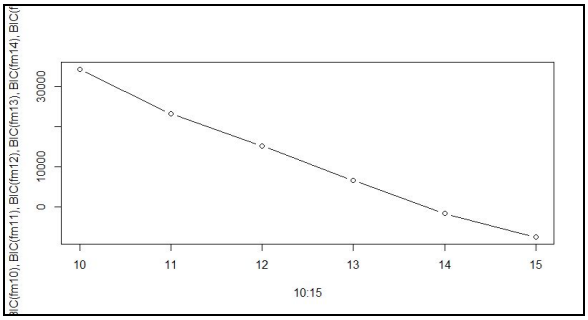


Phase 2:

Approach 2: Finding Contextual Anomalies

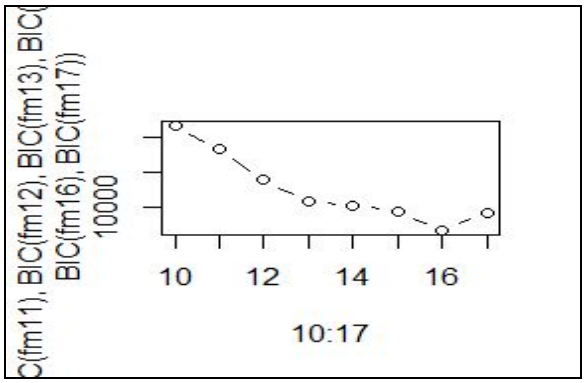
Our initial goal was to build a univariate Hidden Markov model (HMM) using the Global Active Power and a multivariate HMM using both Global Active Power and Global Intensity as they shared a high correlation coefficient. However, modelling multivariate HMMs brought up challenges with the program. When modelling multivariate HMMs, the models would not be able to be created after 6 states in our computing devices. Hence, we decided to focus on univariate data on Global Active Power.

Before comparing the results of the log-likelihood, we developed models that we thought would be the best by manipulating the number of states and observing both the BIC value and its log-likelihood.



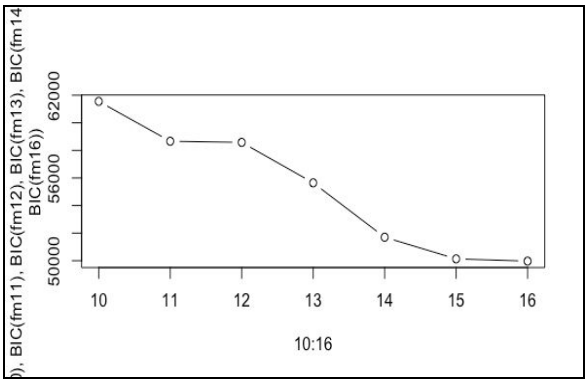
Number of states vs BIC value for Weekday Mornings

We decided on 13 as the ideal number of states because the model ceased to converge after 13 number of states.



Number of states vs BIC value for Weekend Mornings

We decided on 14 as the ideal number of states because the log-likelihood was a positive number after 14 as the number of states. Although the BIC was decreasing the log-likelihood did not make sense as it was a positive.



Number of states vs BIC value for Weekend Nights

We decided that 16 was the ideal number of states because the BIC value started to slightly increase after 16 as the number of states.

The following table below displays the results of the log-likelihood of the training data and the test data.

Subset	**Training**	**Test**
Weekdays Morning (Number of states = 13)	-2137.596	**TemporaryTest:** -148095.7 **Test 1:** -148095.7 **Test 2:** -149302.1 **Test 3:** -148095.7 **Test 4:** -650510.5 **Test 5:** -657725
Weekdays Night (Number of states = 16)	-55778.27	**TemporaryTest:** -139037.9 **Test 1:** -139037.9 **Test 2:** -138999 **Test 3:** -139037.9 **Test 4:** -535442.7 **Test 5:** -536319
Weekends Morning (Number of states = 14)	-3943.674	**TemporaryTest:** -56885.37 **Test 1:** -56885.37 **Test 2:** -57085.43 **Test 3:** -56885.37 **Test 4:** -239010.8 **Test 5:** -237294.2
Weekends Night (Number of states = 16)	-23374.12	**TemporaryTest:** -54248.82 **Test 1:** -69838.19 **Test 2:** -59237.82 **Test 3:** -44762.41 **Test 4:** 103032.54 **Test 5:** 102526.82

As the data suggests, the log likelihoods are extremely different for the training data and the test data. This suggests that either the test data contain a lot of contextual anomalies or that our model was not a good representation of normal behavior.

Interestingly, some of the log-likelihoods of the tests share the exact same value. We believe this means that our model still has value because there are tests with log-likelihoods much farther

apart compared to other test data. Perhaps the ratio of the differences may represent the amount of anomalous data points in the tests.

There are numerous ways to improve our model. One is to shorten the time period. This would lower the amount of total data, so the depmixS4 package may be able to operate better. We found that the depmixS4 package had high difficulty in modelling large datasets. To improve our model, we thought of choosing one day instead of weekdays vs weekends.

Another method may be to use better packages in R that modeled Hidden Markov models in a more sophisticated way. However, these packages are more complex to use which is the tradeoff.

Precision and Recall

Precision is the fraction of relevant instances among the retrieved instances. Whereas recall is the fraction of relevant instances that have been retrieved over total relevant instances. Both precision and recall are therefore based on an understanding and measure of relevance.

Intuitively, high precision means an algorithm returns substantially more relevant results than irrelevant ones, while high recall means that it returns most of the relevant results.

Although both are wanted, there is an inverse relationship between precision and recall, where it is possible to increase one at the cost of reducing the other.

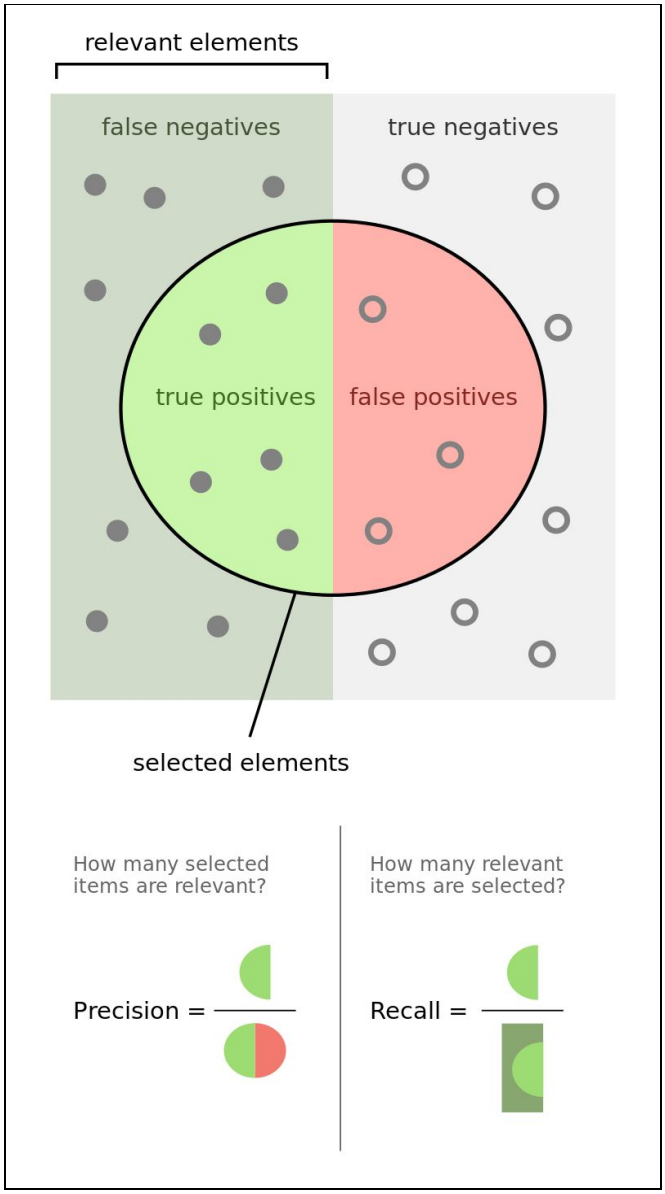
Therefore, depending on priorities, security systems can put more weight on one over the other. For example, critical infrastructures may need higher recall than precision because they have less tolerance for false negatives.

Classification context

For classification tasks, the terms true positives , true negatives, false positives and false negatives, compare the results of the classifier under test with trusted external judgments. The terms ‘positive’ and 'negative’ refer to the classifier's prediction (or expectation),
and the terms true and false refer to whether that prediction corresponds to the external judgment (a.k.a. the observation).

Alternative terminology:

- true positive (TP) — eqv. with hit
- true negative (TN) — eqv. with correct rejection
- false positive (FP) — eqv. with false alarm
- false negative (FN) — eqv. with miss



F Measure

The F measure (F1 score or F score) is a measure of a test's accuracy and is defined as the weighted harmonic mean of the precision and recall of the test.

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

It is approximately the average of the two when they are close, and is more generally the harmonic mean, which, for the case of two numbers, coincides with the square of the geometric mean divided by the arithmetic mean.