Richard Swann | 823902440
Jason Cao | 200107583
Matt Fung | 301329703

CMPT318 Assignment 3

Q1. Data Exploration

	Global active power	Global reactive power	Voltage	Global intensity	Sub metering 1	Sub metering 2	Sub metering 3
Global active power	1.000	0.140	-0.045	0.722	0.282	0.297	0.444
Global reactive power	0.140	1.000	0.058	0.267	0.135	0.142	0.084
Voltage	-0.045	0.058	1.000	-0.003	-0.023	-0.016	0.001
Global intensity	0.722	0.267	-0.003	1.000	0.488	0.449	0.618
Sub metering 1	0.282	0.135	-0.023	0.488	1.000	0.059	0.108
Sub metering 2	0.297	0.142	-0.016	0.449	0.059	1.000	0.090
Sub metering 3	0.444	0.084	0.001	0.618	0.108	0.090	1.000

Table 1. Correlation Matrix of Variables

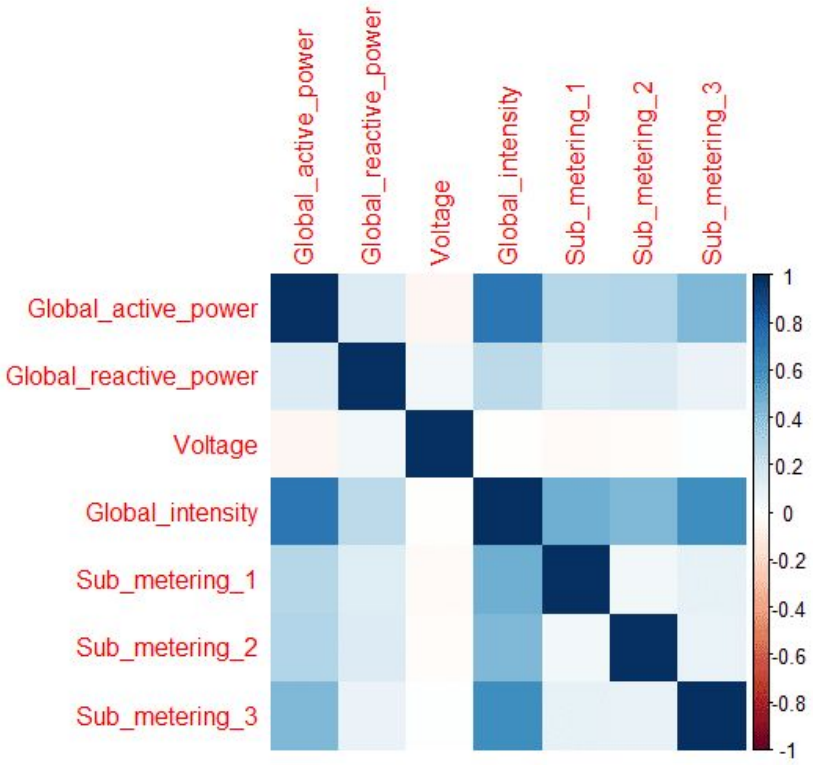

Figure 1. Correlation Matrix[2] Depicting the Correlation For Each Disjoint Pair of Features A - G

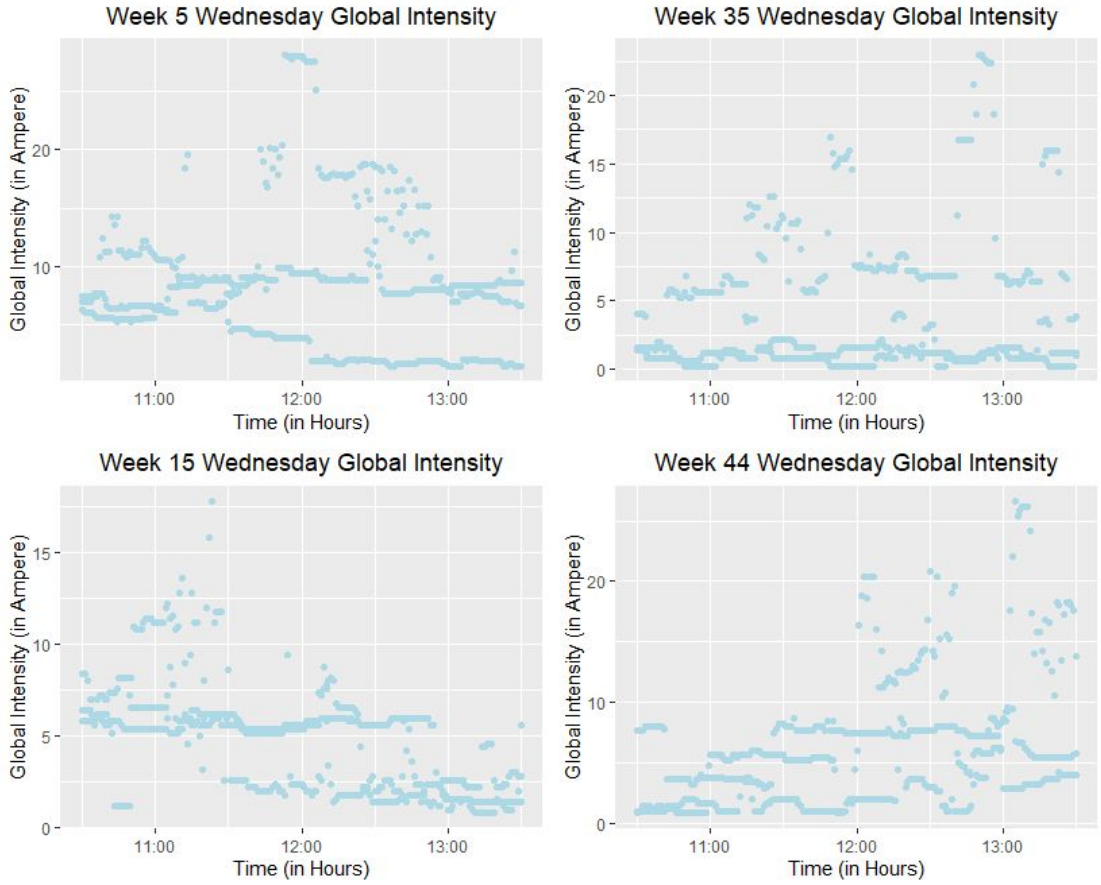

Figure 2. Diagram Depicting the Plots of Global Intensity for Wednesday For Weeks 5, 15, 35 and 44

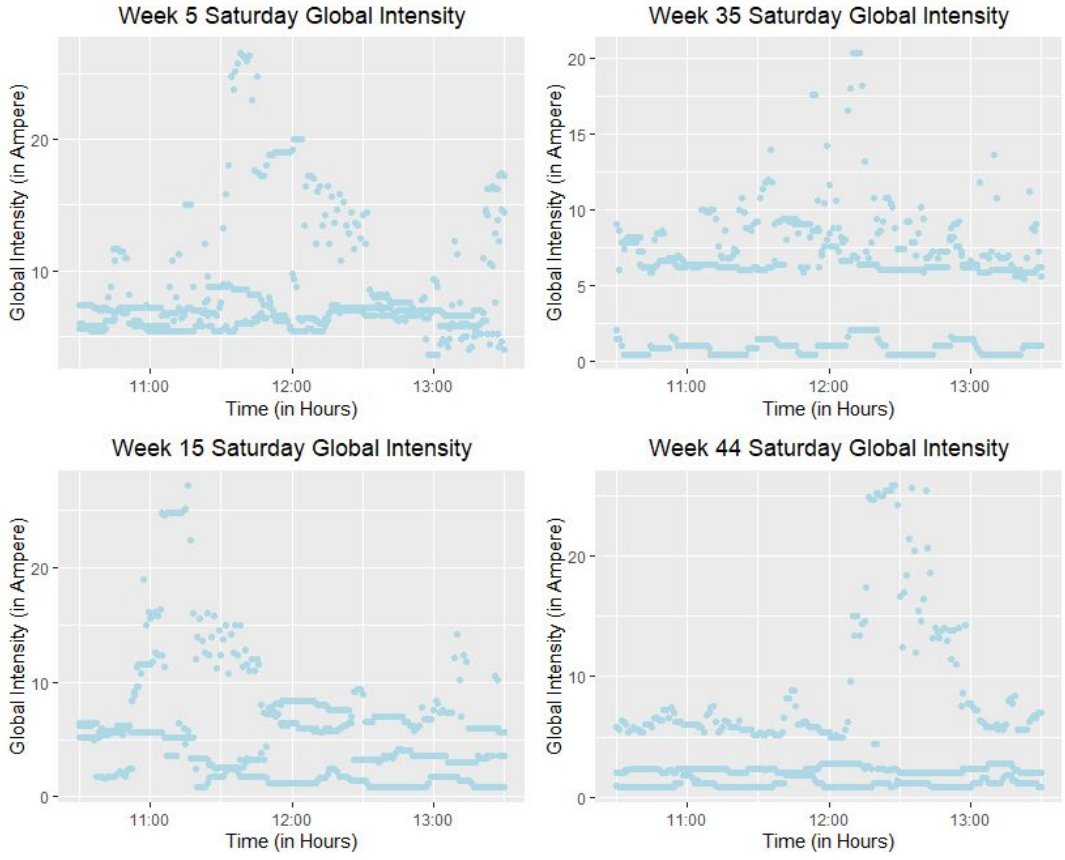

Figure 3. Diagram Depicting the Plots of Global Intensity for Saturday For Weeks 5, 15, 35 and 44

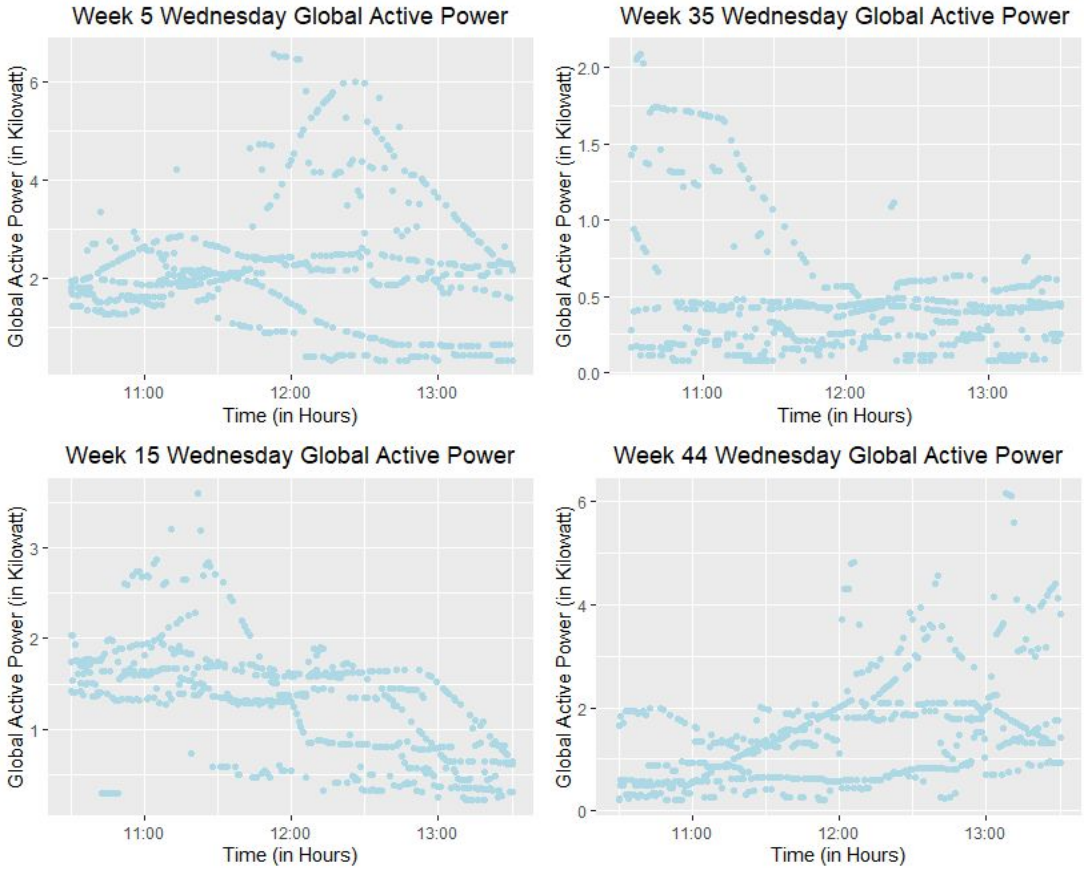

Figure 4. Diagram Depicting the Plots of Global Active Power for Wednesday For Weeks 5, 15, 35 and 44

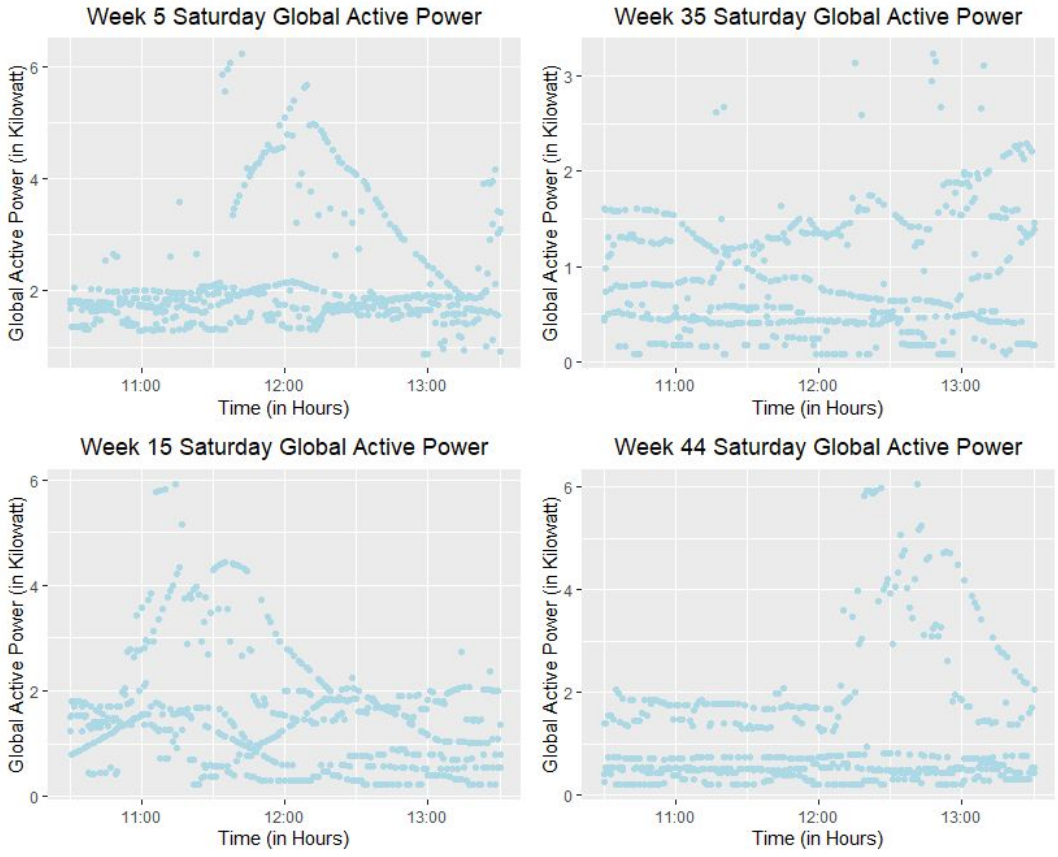

Figure 5. Diagram Depicting the Plots of Global Active Power for Saturday For Weeks 5, 15, 35 and 44

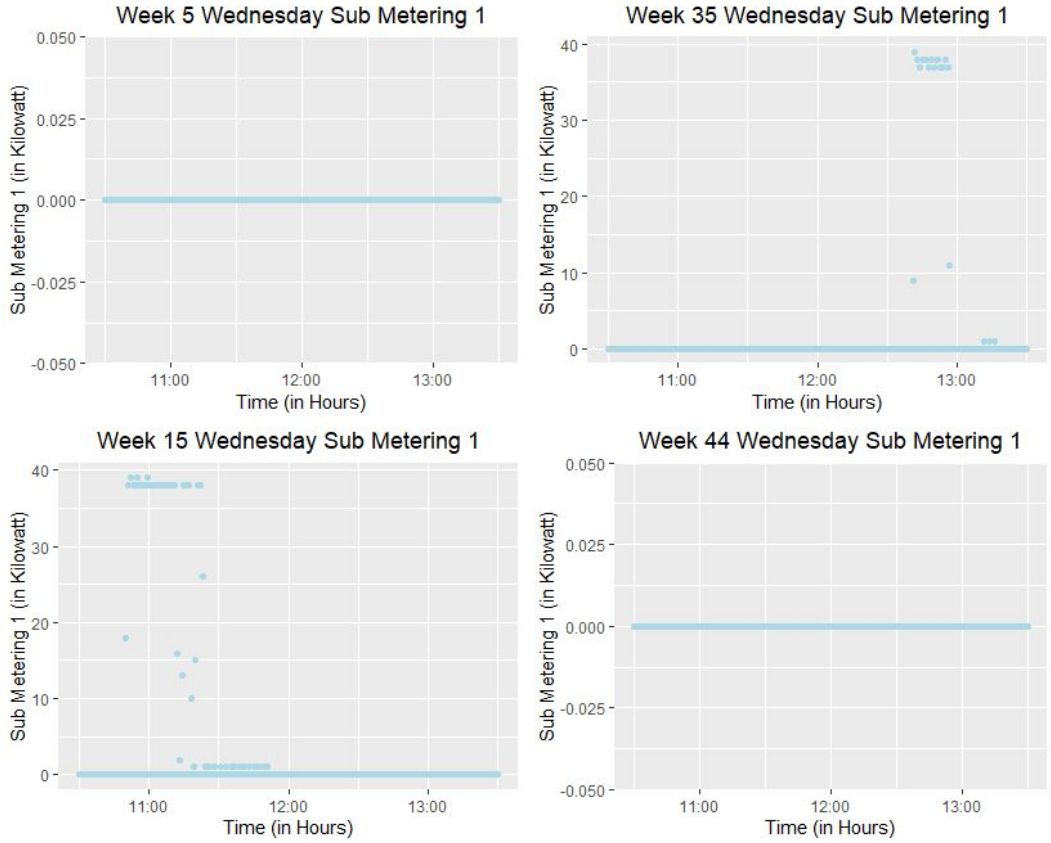

Figure 6. Diagram Depicting the Plots of Sub Metering 1 for Wednesday For Weeks 5, 15, 35 and 44

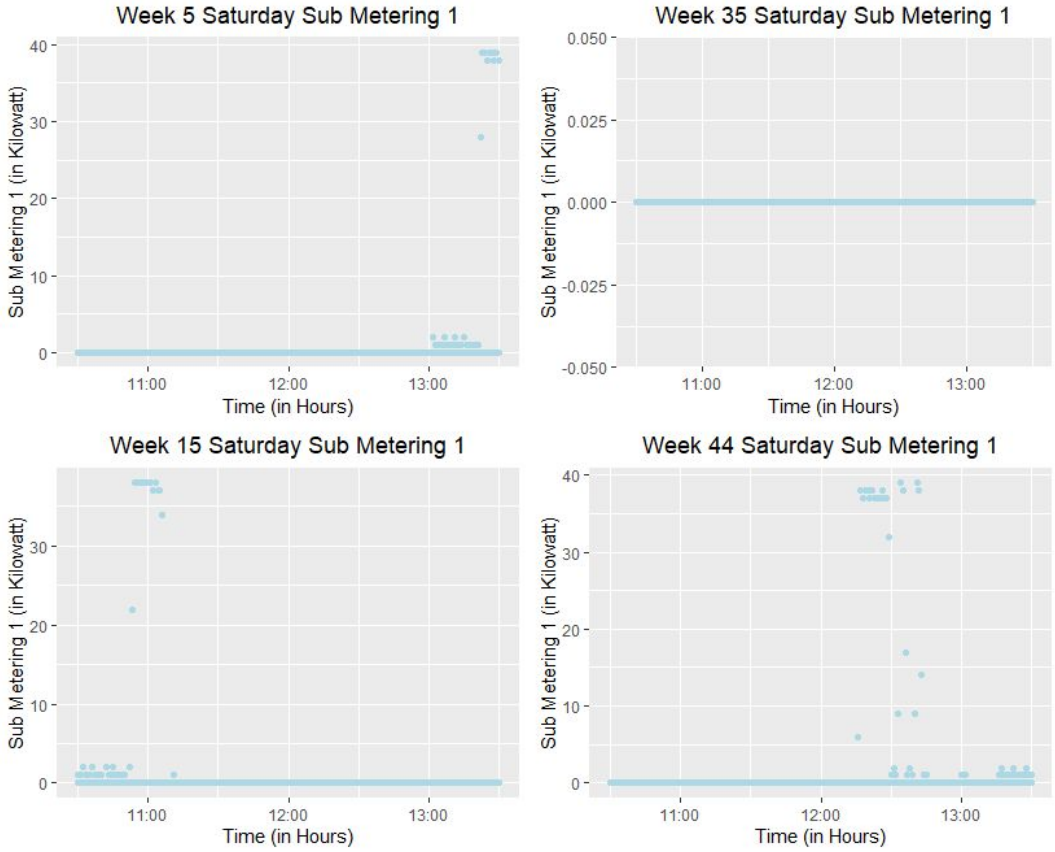

Figure 7. Diagram Depicting the Plots of Sub Metering 1 for Saturday For Weeks 5, 15, 35 and 44

Comments

The chosen observed variables are Global Intensity, Global Active Power and Sub Metering 1. From Table 1, we can see that Global Intensity has a strong correlation with Global Active Power and a moderate correlation with Sub Metering 1. We can also see that Global Active Power and Sub Metering 1 have a weak correlation. Hence, this is a good combination of observed response variables as the strength of the correlation between the three response variables with each other varies with different pairings. With the observed variables determined, we then found a time representative window on weekdays and weekend days that show a clearly recognizable electricity consumption pattern over a time period of several hours. The time representative window chosen is between 10:30 - 13:30. During this window, we can see there is a surge in energy consumption. This pattern holds true for the weekdays and weekends for all three observed response variables. Due to having to take in account of finding multiple weeks where this pattern holds true for weekdays and weekend days for all three observed variables, there are certain cases where this pattern is not evident. For weeks 5 and 44 for weekdays and week 35 for weekends, this pattern does not hold for Sub Metering 1. However, it should be noted that for these exceptions, the value for Sub Metering 1 was a NA in the datafile. We replaced all NAs with zeros in the datafile and therefore this is why these exceptions occurred. Apart from this, the pattern is evidently seen for all the observed response variables on weekdays and weekends.

Q2. Model Training

The provided dataset is partitioned into weekdays training set & test set and weekends training set & test set. According to the recommendation from the instructor, the first 3 years of the data are considered to be the training set while the last year of the data are the test set.

In training the HMMs, a number of different states across a range from 10 states to 20 states are used to find a model with the optimal state.

An optimal model is neither overfitted nor underfitted on the dataset. The following criterion ensures an optimal model is selected: find the state with a negative Log-likelihood close to 0 and a large BIC.

Weekdays HMM Models				Weekends HMM Models		
States	BICs	logLik		States	BICs	logLik
10	320820.8	-159555.9		10	199110.0	-98769.4
11	307130	-152561.3		11	199481.1	-98817.9
12	319868.9	-158770		12	191277.1	-94568.3
13	303741.5	-150534.4		13	186896.3	-92219.7
14	277645.9	-137303.1		14	185025.6	-91115.7
15	262545.1	-129557.7		15	178756.1	-87801.6
16	259879.5	-128018.4		16	177448.1	-86957.8
17	285621.4	-140671.4		17	195688.8	-95877.8
18	-520699.3	262718.3		18	173109.1	-84377.1
19	235904.8	-115342.9		19	169722.5	-82462.4
0	0	0		20	172872.9	-83805.6

Table 2. A list of HMM models for both Weekdays and Weekends with their corresponding BICs and LogLik

HMM states (each representing a corresponding model) are graphed against their corresponding BICs and Log Likelihood as a method to select the optimal HMM model.

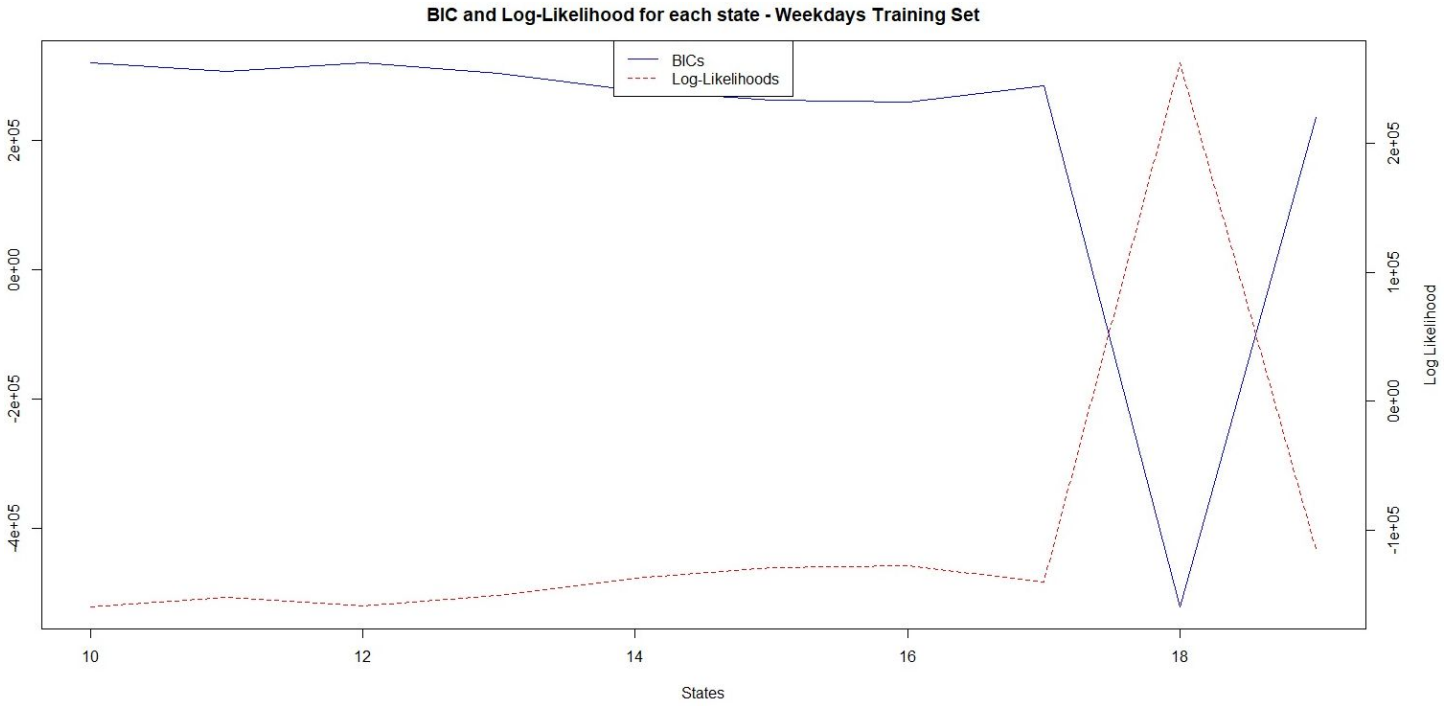

Figure 8. Graph depicting the BIC and Log-likelihood for each state from 10 to 20 for Weekdays

Based on Figure 8, the likely HMM candidate (i.e., a model with a negative Log-Likelihood that's close to 0 and a high positive BIC value) for Weekdays data is the one with states parameter = 16.

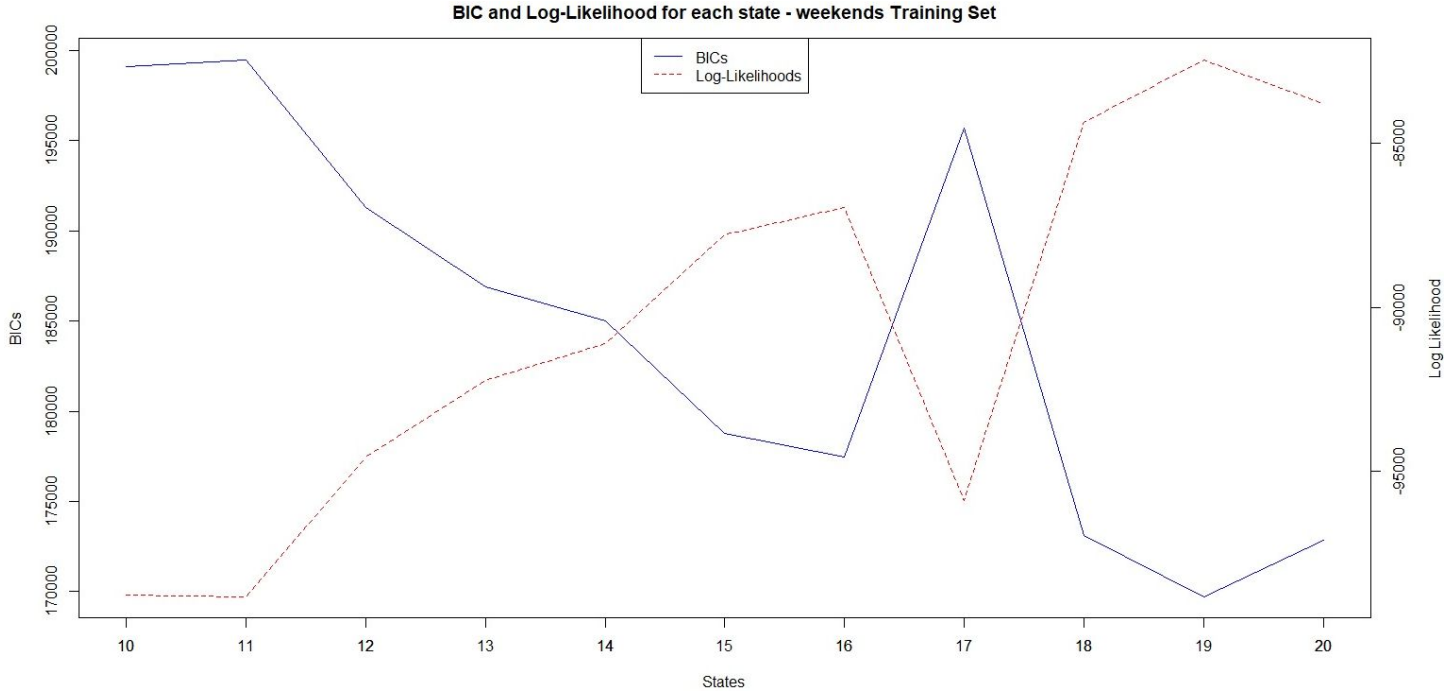

Figure 9. Graph depicting the BIC and Log-likelihood for each state from 10 to 20 for Weekends

Based on Figure 9, the likely HMM candidate (i.e., a model with a negative Log-Likelihood that's close to 0 and a high positive BIC value) for Weekdays data is the one with states parameter = 14.

Q3. Model Testing

After training the models and finding the best number of states for the chosen responses (based on log-likelihood and BIC), Weekdays and Weekends models are evaluated with their respective test data.

Parameters of trained models are fed to the test data to compute the log-likelihood values which are then normalized in order to be compared against the log-likelihood of the trained models.

Weekdays Model: state = 16	Weekdays Model: state = 14
Training: 'log Lik.' -189762.2 (df=335)	Training: 'log Lik.' -94037.76 (df=265)
Testing: 'log Lik.' -194487.8 (df=335)	Testing: 'log Lik.' -90553.72 (df=265)
Difference in log Lik value: 4,725.6	Difference in log Lik value: 3,484.04

Table 3. A comparison between the log-likelihood of training vs testing for Weekdays and Weekends models

For a fit model (not overfitted nor underfitted), the log-likelihood for the training set and test set should be close. Based on the results, both models are fit models.

Q4. a) Anomaly Detection via Log-Likelihood Method

With the completion of model training and validation, both Weekdays and Weekends models are used to detect the existence of anomalies in the second dataset to the same chosen time windows.

After splitting the second dataset into weekdays and weekends, each partition is assessed by their respective models i.e., parameters of Weekdays trained model is fed to the weekdays anomalies data to compute a log-likelihood value and parameters of Weekends trained model is fed to the weekends anomalies data to compute a log-likelihood value. The computed log-likelihood values are normalized for comparisons.

Weekdays Model: state = 16	Weekdays Model: state = 14 (used seed 1)
Training: 'log Lik.' -189762.2 (df=335)	Training: 'log Lik.' -89932.19 (df=265)
Testing: 'log Lik.' -142879.3 (df=335)	Testing: 'log Lik.' -148774.80 (df=265)
Difference in log Lik value: 46,882.9	Difference in log Lik value: 58,842.6

Table 4. A comparison between the log-likelihood of training vs anomaly for Weekdays and Weekends models

For anomaly detection, the log-likelihood for the training set and test set should be quite large. Based on the results, anomalies exist in both weekdays and weekends in the second dataset.

Q4. b) Moving Average

Using a moving average with thresholds approach doesn't lend itself to using multiple disjoint data blocks e.g. Monday from 10:00 to 13:30, followed by Tuesday from 10:00 to 13:30, … The data discontinuities at the end of each data block will cause problems. So it was decided to select a single weekday, Wednesday, and a single weekend day, Saturday, for evaluation.

Extending from our work in assignment 2 we will use Global_intensity as the predictor and moving average with a averaging window of 10 data points and using the "exponential" moving average function (The exponential moving average is a weighted moving average that reduces influences by applying more weight to recent data points () reduction factor 2/(n+1)).

For thresholds we will use T1 - 10%, T2 - 25% and T3 - 50% of the maximum Global_intensity value.

To detect anomalies we will consider the number of exceptions generated by the Training data for a given threshold as the baseline and compare the number of exceptions generated by the Test data and the number of exceptions generated by the Anomaly data for the same threshold to determine if there is any significant predictive capability using a moving average.

Threshold	**Training Data**	**Test Data**	**Anomaly Data**
T1	13	51	78
T2	4	33	26
T3	1	16	7

Table 5. Weekday Results

Threshold	**Training Data**	**Test Data**	**Anomaly Data**
T1	140	66	45
T2	101	18	13
T3	46	5	4

Table 6: Weekend Results

Conclusion

Using Moving Average and a Threshold has no obvious anomaly detection capability.