Demographic Electricity Consumption

This smart meter data has been taken from two demographic user groups, Mosaic G and Mosaic I. These were obtained from the Customer Revolution Network smart meter trial as it recorded not only the energy consumption of each participant, but also broke them down into separate demographics based on household type. For each demographic, electricity consumption has been recorded for a 30 minute period beginning at 4 am and 10 am on the same day; Monday the 9th January 2012. The data will be statistically analysed using Matlab and conclusions drawn.

Question 1

	Mean (kWh)	Standard Deviation	Confidence level (kWh)
Mosaic G 4am	0.214	0.184	0.214 +/- 0.022
Mosaic G 10am	0.468	0.600	0.468 +/- 0.071
Mosaic I 4am	0.181	0.176	0.181 +/- 0.011
Mosaic I 10am	0.483	0.546	0.483 +/- 0.036

Table 1: Mean, Standard Deviation and Confidence level for both demographics at 4am and 10am

Mean

The mean is calculated by dividing the sum of all energy consumptions in a particular data set by the number of elements in that set. This can be achieved in Matlab using the mean() function. From the data in Table 1 we can see that while Mosaic G has a higher energy consumption in the early morning, by mid morning Mosaic I were consuming more electricity on average. However, both saw a sharp increase by a factor of 2.2 and 2.7 respectively as people began to wake up and make use of appliances. Based on the mean there doesn't appear to be a dramatic difference in energy consumptions between the two Mosaics.

Standard Deviation

Standard deviation is a measure of the variation in a set of values. It is calculated using the std() function in matlab, which implements the equation:

$$S = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}} \quad (1)$$

where $X_i = Single\ data\ value, \bar{X} = mean, n = sample\ size$

It can be seen that for both Mosaic G and I, standard deviation is approximately equal in the early morning, and in both cases is over three times higher at 10am than at 4am. This is likely because very few people are awake at 4am, and the only energy being consumed is by appliances running overnight. At 10am on a Monday some people will be at work and some, such as retirees or stay-at-home parents, might be at home, and in some homes washing machines or dishwashers may be running. Hence there is far more variation in potential energy consumption by mid morning. Mosaic I's standard deviation at 10am is slightly less at 0.55 as opposed to 0.6, potentially suggesting that this demographic's energy habits are more similar.

Confidence Interval

The confidence level is the range in which 95% of values lie. Since there are more than 30 values in the dataset, central limit theorem can be used. The matlab norminv(0.975) function returns the confidence level, which can be used to calculate the confidence interval:

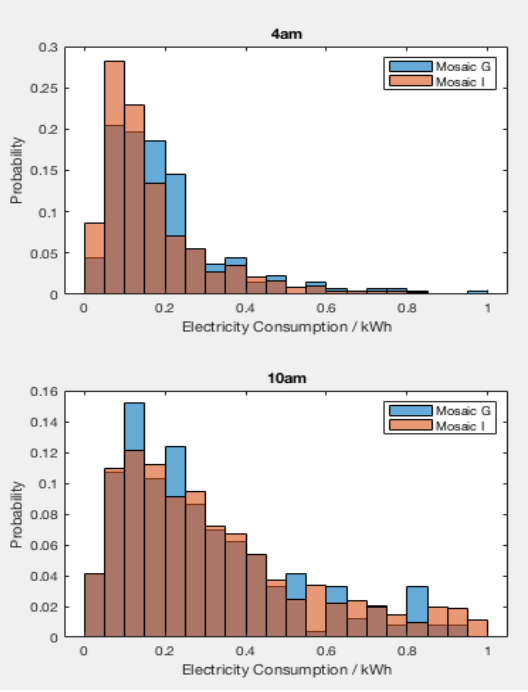
$$CI = \bar{X} \pm Z\frac{\sigma}{\sqrt{n}} \quad (2)$$

where σ = standard deviation and Z = confidence level

For both mosaics, the absolute confidence interval was much greater at 10am than 4am. For mosaic G the confidence interval a factor of 3.2 greater at 10am and 3.3 times greater for Mosaic I, suggesting a much greater variation in usage at 10am which is to be expected since the range of activity should be less at 4am as almost everyone is asleep.

Question 2

Density histograms were plotted in order to visually compare the energy consumption distributions of the 4 datasets. The data was reduced to only include energy use below one kilowatt hour, as a few extreme outliers would skew the x axis and make the main body of data less visible.



Mosaic G and Mosaic I were plotted on the same axes for 4am and 10am. The number of bins would normally be determined by taking the square root of n, but since we're plotting multiple histograms on a single axis a standard number has been used to make comparisons more clear.

Distributions were similar for both mosaics. At 4am mosaic G has a slightly longer tail and appears to have a higher probability of energy consumption around 0.2kWh than Mosaic I, with a lower, longer distribution. This supports the findings in part 1 suggesting they have similar standard deviations, with Mosaic G having a higher mean electricity consumption.

Distributions for 10am were much lower and longer than the 4am distribution. Both mosaics peak at a similar electricity consumption than before, but have far thicker tails, with consumptions being much more varied than before. The distributions for the two Mosaics are very similar, however mosaic I has a greater proportion of higher consumption customers whereas G seems to have more at the lower end around 0.2kWh. Again these findings support the results from part 1, with very similar means, and Mosaic I having a larger standard deviation to account for its slightly more even spread of probability.

Figure 1: Histograms for the 4 data sets

Question 3

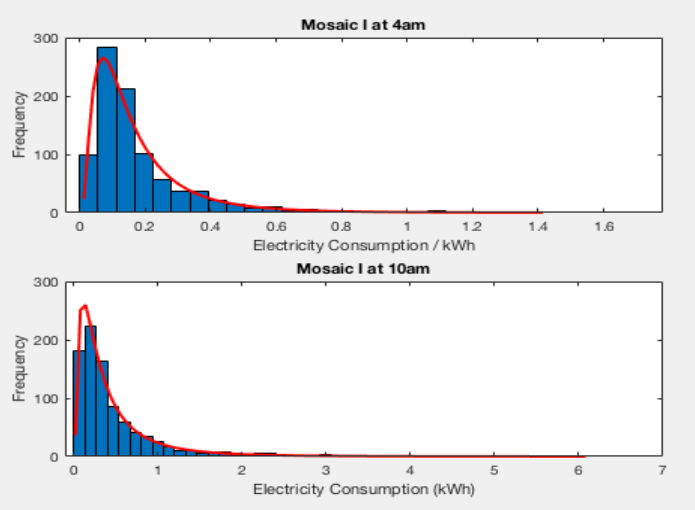


Figure 2: Histograms fitted with lognormal curves

In figure 2, the sample data for mosaic I at 4am and 10am has been represented as a histogram, with a lognormal model fitted over the top as a means of comparison. This was achieved using the histfit() function in matlab. As a first impression, the data appears to fit the model well in both cases, with the exception of the peak frequency value being slightly too low at 10am.

To find out whether the lognormal distribution could accurately be used to model and make assumptions these two datasets, the number of customers predicted by the model to use over 1 kilowatt hour of energy in the 30 minutes after 4am and 10am was compared to the actual number of people who used over 1 kilowatt hour in each case.

The predicted values were arrived at by using the lognfit() function in matlab which returns the mean and variance of the fitted lognormal distribution. These can be used to find the z values using the equation:

$$z = \frac{\log(x) - \mu}{\sigma} \tag{3}$$

Table 2: Comparison between predicted number of customers using over 1kWh vs actual number

	4am	10am
Number of customers predicted to use more than 1kWh	5	104
Number of customers who actually used more than 1kWh	7	106

With z, the number of customers with consumption above 1 can be found by using the normcdf() with z as the input variable, subtracting the answer from 1 and multiplying this by the number of customers in that dataset.

As can be seen in table 2, the predictions are impressively close to the true values. Both were 2 below the real value. With a larger sample size this prediction would likely become more accurate.

Appendix

[1] En.wikipedia.org. (2019). *Log-normal distribution*. [online] Available at: https://en.wikipedia.org/wiki/Log-normal distribution [Accessed 15 Jan. 2019].
[2] Mathsisfun.com. (2019). *Standard Deviation Formulas*. [online] Available at: https://www.mathsisfun.com/data/standard-deviation-formulas.html [Accessed 13 Jan. 2019].
[3] Stat.yale.edu. (1997). *Confidence Intervals*. [online] Available at: http://www.stat.yale.edu/Courses/1997-98/101/confint.htm [Accessed 13 Jan. 2019].