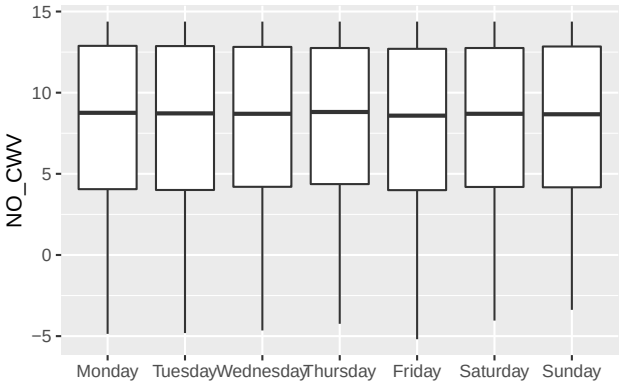
EDA

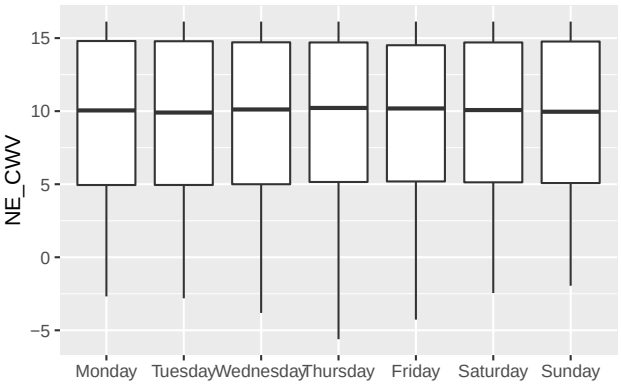
Daniel Dennis

27/05/2021

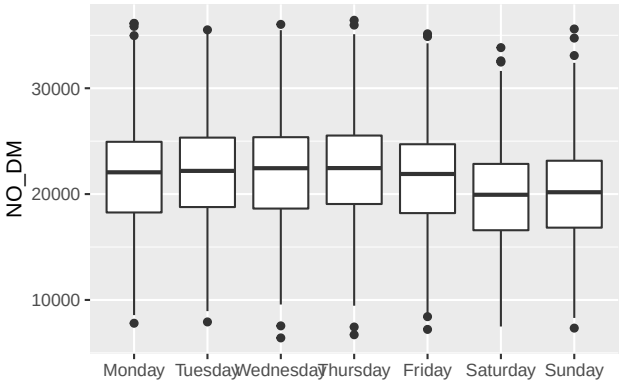
To begin, the data was loaded into RStudio from the .csv file which contained 4031 observations. There were four entries that exclusively featured 'NA' values, plus a further entry with 'NA' corresponding to all four variables (NO_CWV, NO_DM, NE_CWV, NE_DM). These entries were omitted from the data, then a check was performed for duplicated dates, which there were none of. The entries of the 'Date' column were transformed from DD-MONTH-YY format to YYYY-MM-DD format using the lubridate package, allowing the exploratory data analysis process to begin.

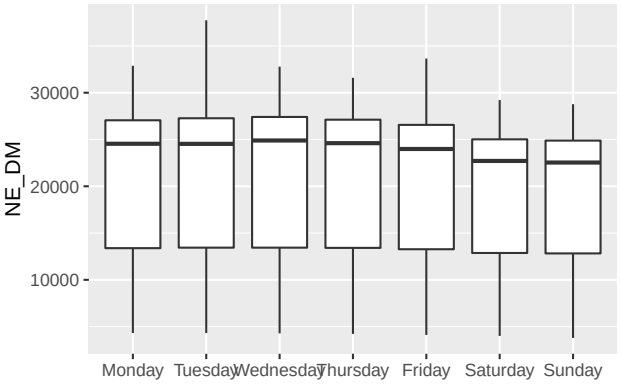
To understand the composite weather variable (CWV), box plots were created (figure 1 and 2) which were grouped into the days of the week. These show that the CWV is constant throughout the week for both regions, which is to be expected since the day of the week has no known relationship with the weather.



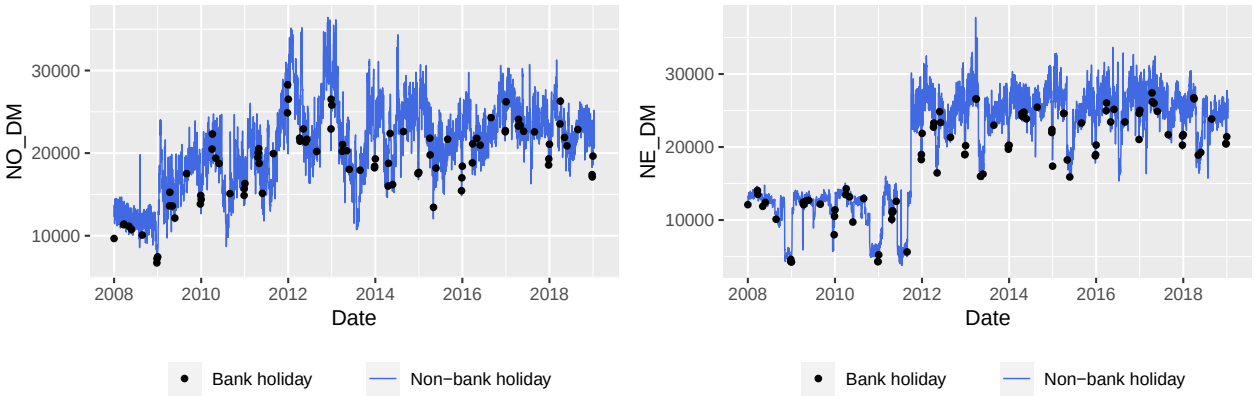


To help understand how the daily metered demand (DM) is affected by days of the week, more box plots were created (figure 3 and 4). These show that the demand is largely similar on weekdays before decreasing at the weekend. This is likely due to industrial users closing or reducing capacity on weekends, thus requiring less gas to power machinery and heat factories, among other things. Figure 3 shows that the North's quartiles have a smaller range than those of the North East (figure 4), meaning that the North's DM observations are generally closer to the median. The North East has a higher median DM for all days of the week, so requires a greater supply of gas than the North, on average.

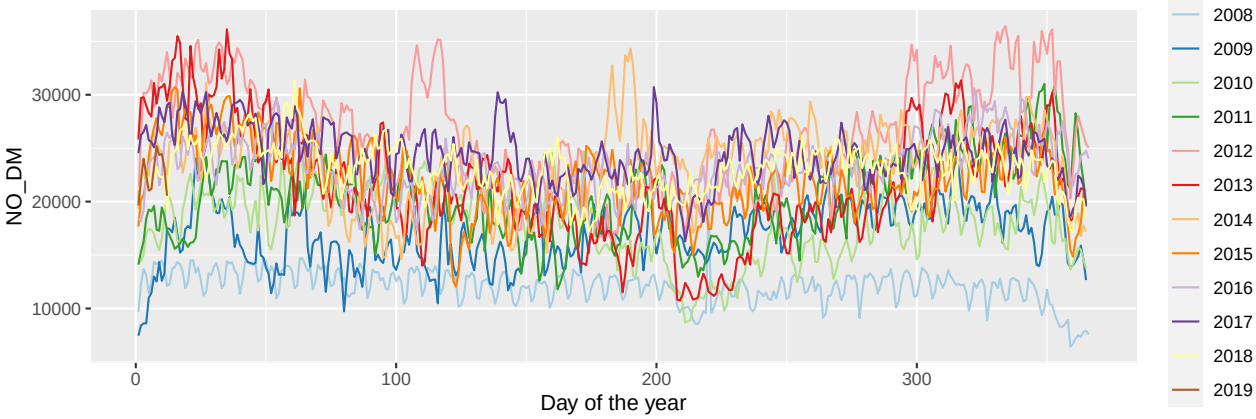


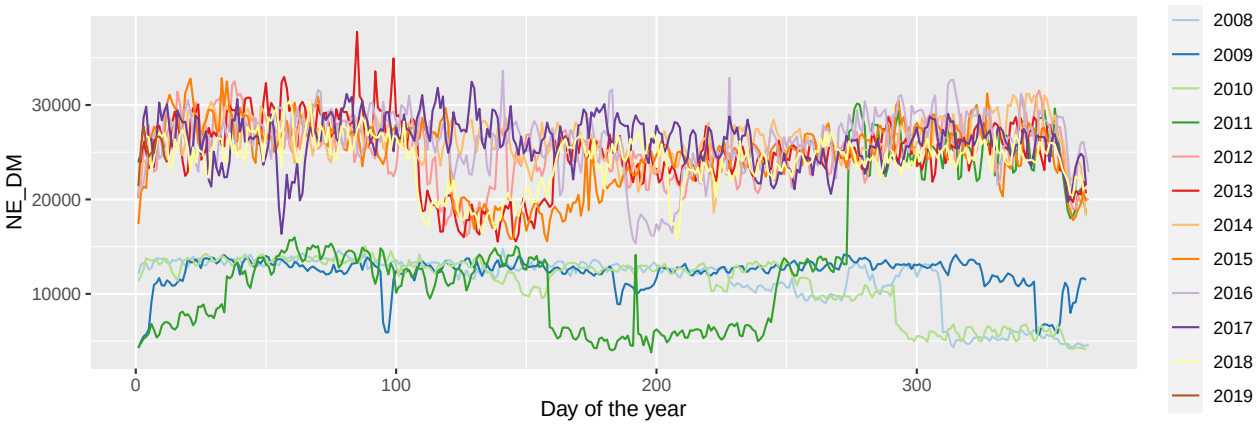


How daily metered demand is changing over time will be a vital factor in forecasting for the future. Figure 5 and 6 show that the demand on bank holidays is consistently lower than it is on non-bank holidays for both regions. There may be some seasonality to the data, however it is hard to see due to both having a general upwards trend over the 10-year period. Figure 5 shows the daily metered demand in the North increasing between 2008 and 2012, peaking in late 2012/early 2013, before plateauing through to 2019. Figure 6 shows the daily metered demand in the North East following a static path from 2008 to mid 2011, before sharply increasing until the inception of 2012, before again plateauing through to 2019. This is an indication that we might wish to use a model that allows step changes in the overall level and its variability. The steps are most clear in figure 6, with there being four general levels the data fall into: 5000, 13,000, 20,000 and 28,000 (all approximations). This pattern is less obvious in figure 5, where the data exhibits periods of linear growth with some step changes in the gradient dispersed throughout.

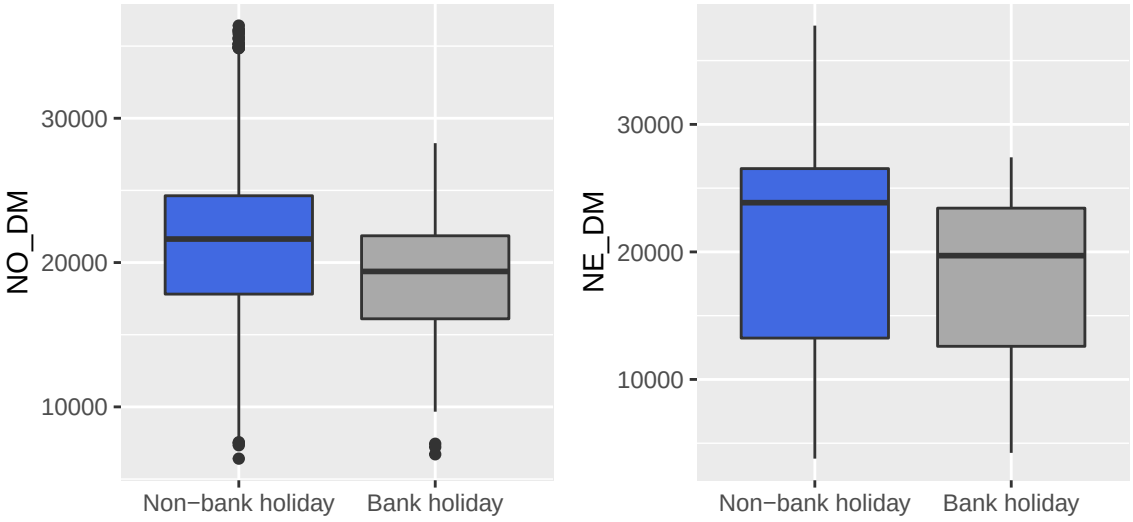


To see how the demand for each year changes, the annual curves can be superimposed on one another. Note that January 1st of each year is represented by 1 on the x-axis, while December 31st will fall on either day 365 or 366. Figures 7 and 8 show how the demand varies throughout the year for the entire 11-year period. There is evidence that demand is lower towards the middle of the year, which coincides with higher temperatures and CWV values, especially in the graph for the North. This suggests that weather influences the demand for gas from industrial consumers.

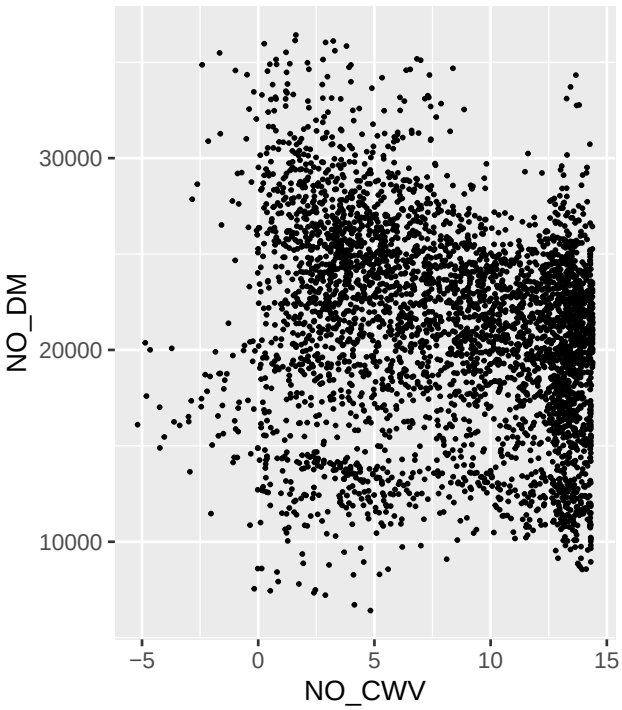


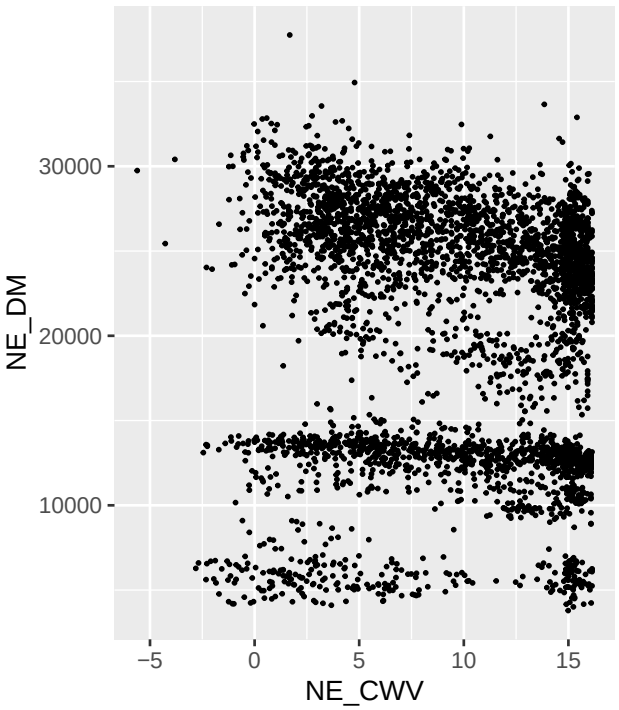


To further investigate the effects of bank holidays on daily metered demand, figures 9 and 10 were produced. These box plots clearly show that bank holidays correspond to lower daily demand for gas than non-bank holidays, as well as covering a smaller range of values across the data. The reason behind the former is that industrial users are more likely pause operations on bank holidays, thus requiring less gas, while the reason behind the latter is that the number of bank holidays is very small in comparison to non-bank holidays, so you would expect them to cover a smaller range of values.



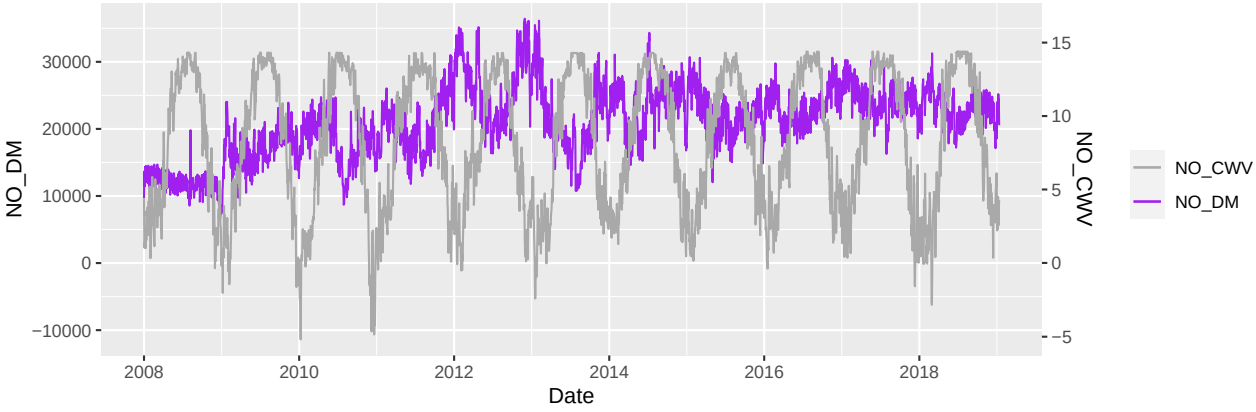
To explore the concept that the weather plays a part on gas demand, scatter plots were created showing the CWV against the DM. Figure 11 and 12 show that there a weak negative correlation between the CWV and DM for both regions. For the North East, the data is fairly distinctly clustered for the daily metered demand. The data tends to sit on four different levels, as seen in figure 6, explaining the four clusters seen in figure 12.

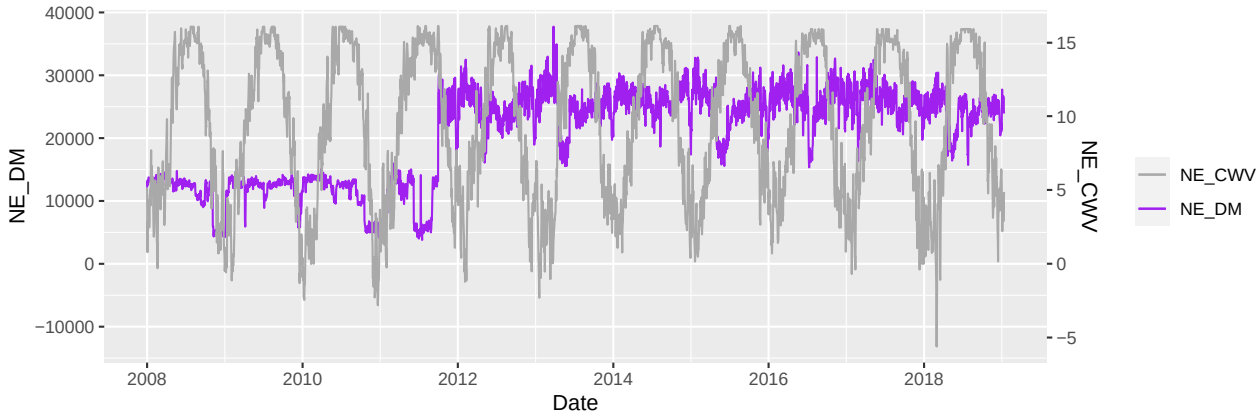




Add k-means clustering here?

The graphs showing the daily metered demand changing through time do not show any clear signs of seasonality, but by plotting the CWV data over the top of these may give a better insight. From figure 13 and 14, there appears to be some inverted relationship between the DM and the CWV. That is, as the CWV peaks, the DM is at a trough, then as the CWV troughs, the DM peaks. This is more pronounced in the North than the North East, especially in the period from 2012-2014. This suggests that in the winter, the daily metered demand tends to be at its highest for the year, while in summer it is at its lowest. This is likely due to the industrial users requiring more gas to heat their buildings due to lower temperatures/poorer weather.

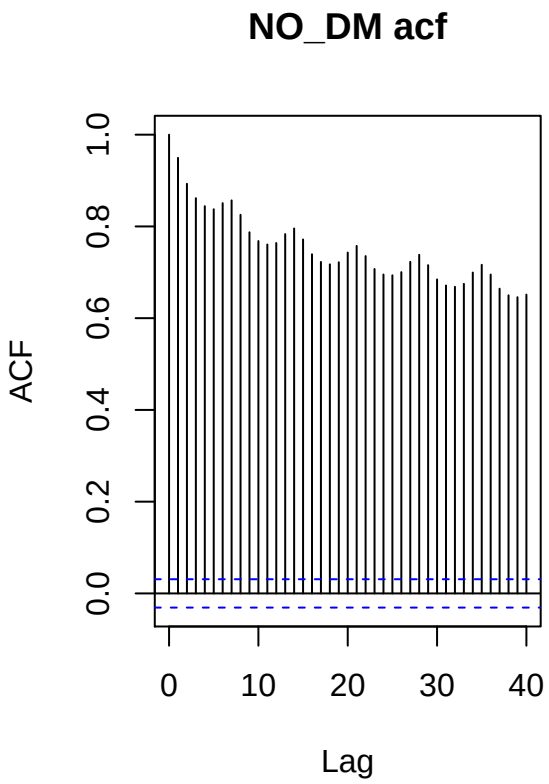


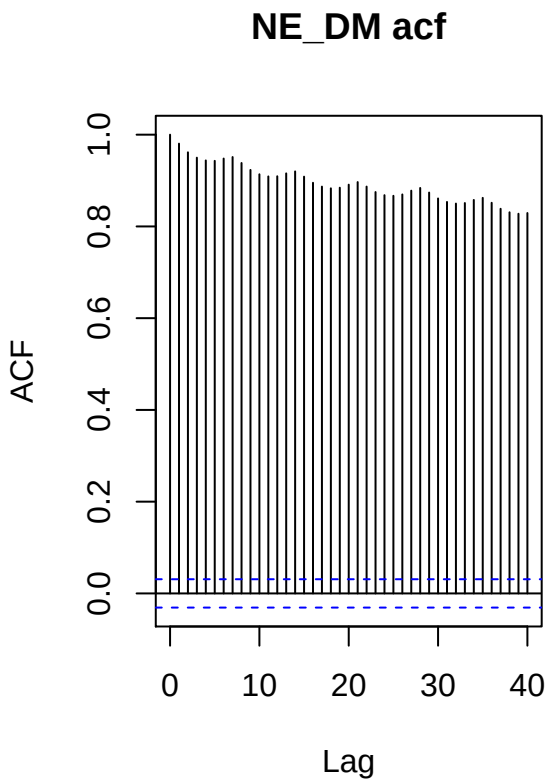


There do not appear to be any obvious correlations in the data from figure 9 and 10, so using correlation coefficients (Pearson and Spearman) may help give more insight. Table 1 shows these for the CWV vs the DM in the North and North East, then the DM in the North vs the North East. It shows a weak negative correlation across coefficients for the CWV vs the DM in both the North and North East. This indicates there is very little relationship between the CWV and the DM. The DM in the North vs the North East shows relatively strong positive correlations across the coefficients. This indicates that as the DM increases in one region, we would usually see an increase in the other region. All the p-values are very small and thus significant, so the null hypothesis that the p-values' corresponding correlation coefficient is equal to 0 is rejected.

	Pearson	Pearson p-value	Spearman	Spearman p-value
NO_CWV vs NO_DM	-0.260	<2.2e-16	-0.276	<2.2e-16
NE_CWV vs NE_DM	-0.0883	1.981e-08	-0.187	<2.2e-16
NO_DM vs NE_DM	0.695	<2.2e-16	0.714	<2.2e-16

The autocorrelation function of the DM is useful to gain an understanding of the correlation of points separated by various time lags. Figures 15 and 16 show the autocorrelation function of the DM in both regions, each exhibiting very slow decay. The data is clearly from a non-stationary process due to the varying mean from figures 5 and 6, with nearby observations being highly correlated.


NO_DM acf
ACF
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
Lag


NE_DM acf
ACF
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
Lag