**Data Cycle Operations**

**Some words to help your mental framework**

* All models are wrong, but some are useful
* Garbage in Garbage out: cleaner data will produce a better model, methods don’t fix bad data
* There is no right answer but there are wrong answers
* Explainable is important: You can get great statistical results but if they can’t be explained then your model may not be useful.
* We are not searching for answers but looking to build a model that can help us predict the answer and within as little chance that we may be wrong

**Business understanding**

* Assignment questions
* As we are not experts on the weather and this assignment is within a short time we will not need to focus on this much. However, some research may provide clues. Predictive modelling will likely tell us this anyway.

**Data Mining**

The data set we have been provided is mined data from some application. There may be more data somewhere with in but we are limited on time and its likely new findings will come from feature engineering.

**Data cleaning – We are here(mostly)**

This is already a very clean data set.

Formatting dates to solve the current issue of merging data sets is one barrier, only 3 dates left to fix. The current weather demand data set is reflective of this, and we are still able to use this for now. ~*assigned to Liam*

Filling NA’s

The most obvious data cleaning left to be done is filling NAs, this will be a factor for when we run predictive modelling but we can still derive insights to get ahead on our understanding.

Replacing with 0 for numeric columns or another character for character columns is likely the best option given our little time frame, or another method is finding rows of similar parameters and using the average to fill that in(numeric) or the most common/predictable occurrence. Placing a null value may be safer given the time frame but there are other methods if you have the capacity.

**Data exploration**

Assigning price categories for each day ~ Liam

We are half way here, we can start looking at things like correlation to understand the impact between different variables and filtering high low strength and directions. Plotting interesting relationships variables will be really helpful.

Data visualization is also helpful for exploring data and will provide visuals we can use in our report. As well contrasting or supporting visuals to our predictive model. Predictive modelling will do the majority of the work for us in this case.

Pair plots and heatmaps are great for correlations

We are able to work without the missing merged price demand data for now, once the formatting issue is fixed we can replace the CSV file and re run the code to see how this effects results etc.