# Data Exploration

The dataset is comprised of a semi-contiguous observation period spanning from July 6, 2017 to August 6, 2018 (two days of data from roughly July 31, 2018 to August 1, 2018 are missing).

The time series data can be broken up into “voyages”, contiguous periods of engine activity (as judged by Fuel Consumption) and “stops”, contagious periods of idle engine presumably while in port.

The dataset has been hour level qualified and our target variable, Main Engine Fuel Consumption, represents the averaged burn rate over the hour. Because of the aggregate nature of the data, fuel consumption measurements at the beginning and end of voyages have been contaminated by having the engine off for an indeterminate period of time and will be excluded from analysis.

Stopping periods will similarly be excluded, as they will lead to overestimates in true model performance.

# Train Test Split Design

The model should be fit and evaluated using separate voyages to prevent highly temporally correlated data points from being used in both train and test sets leading to over-confidence in model performance.

This design makes it difficult to ensure our model will experience the full range of possible input conditions however so we will need to use cross validation to estimate our model robustness to this approach.

# Feature Engineering

The goal of this exercise can be interpreted in one of two ways:

1. Build a regression model that can project fuel consumption into the future given a variable set of conditions. This type of model would be most effective for route optimization. In this case data from future time points in the same voyage cannot be used in feature engineering but past data points are fair game.
2. Build a deconvolutional model to isolate the effects of independent variables like draft and wind from fuel burn estimates. This type of model would be optimal for quantitative evaluation of vessel performance (vessel optimization). In this case data from past and future time points can be used in feature engineering.

To preserve generality of the model we will use interpretation number one for feature engineering. Features to generate:

1. Difference in compass heading between current and previous hour data points – presumably turning will impact vessel drag and fuel burn, and may indicate if in port or not. Must normalize angle deltas to be within +/- 180 degree range and then use their sine component.
2. Trim as recommended by the appendix, thanks! Will normalize by mean and standard deviation.
3. Sine transformation of the rudder angle. Meaningful rudder deflections are limited by < +/-90 degrees anyway.
4. Rudder deflections without change in heading may be indicative of lateral ocean currents or winds, will use the normalized ratio rudder\_deflections/compass\_heading\_deltas as a feature. Larger negative ratios indicate strong lateral currents/winds.
5. Water depth readings of zero will be replaced by the max water depth value.
6. A number of ​water surface temperature readings indicate absolute zero. This reading typically spans an entire voyage. Will replace those readings with the rolling mean of the 24 nearest valid readings regardless of engine activity.