Anaerobic Oxidation of Methane:

Methane (CH4) is an important energy source for sea-floor life in many parts of the world and feeds many micro-organisms in these areas. Since methane is a greenhouse gas, there is interest in understanding the precise way in which the micro-organisms process this methane and convert it into other compounds. One of the mechanisms for this conversion is called anaerobic oxidation of methane (AOM). AOM serves as one indicator of microbial activity. Rates of carbon and sulfur assimilation in sea-floor sediments are highly variable in deep seafloor sediments and can be drastically different over centimeter distances.

Radio-tracer studies which are used to measure AOM are relatively expensive, time consuming, and require special licensing to perform. Therefore, our goal is to find proxies to serve as estimators of **AOM** to streamline the work. Determining the feasibility of doing so is part of your task here.

A group of researchers at the University of Georgia received funding to collect soil samples from the floor of the Gulf of Mexico. These data include measurements taken by four different ships at 21 different sites, which can be described as one of three site types: shelf, abyssal, or oil seep. Each observation also notes the depth of the sample: some samples were taken from the oily layer on top of the sea floor, others from various depths. Overall, there are n=275 observations.

Each observation includes measures of CH4, NO3, NO2, and NH4 (measured in micromoles) as well as Sulfide and SO4 (measured in millimoles), all of which can be measured while the boat is underway. AOM, POC, and DOC are also given here, but these can only be measured at special labs on shore.

Your task is to answer two main questions:

1. Can we predict AOM with any sort of accuracy using “at-sea” measurements? Notably, AOM =0 for more than half of the observations, so even identifying the presence of AOM would be useful
2. What attributes of a soil sample are associated with AOM?

A few notes:

- There are missing values in several places in the data. You’ll need to decide how to handle them.
- Many variables have skewed distributions.