# AOM: Identification and Association with Soil Sample Features

Kevin Nguyen

September 20, 2022

## 1 Executive Summary

The most potent greenhouse gas in terms of affecting the temperature of the Earth is methane. Given its potency, researchers want to know how microorganisms on the sea floor process and convert methane into other compounds via a process known as anaerobic oxidation of methane. Both of the client's questions were properly addressed in this report. The first question was answered using robust binary logistic regression, while the second question was answered using a test of association using Kendall's $\tau$. We can predict the presence of AOM with a reasonable level of accuracy using logistic regression. Furthermore, according to the test of association, Sed Depth, $CH_4$, $NO_2$, and $NH_4$ are associated with AOM.

## 2 Introduction

### 2.1 General Background

There are various greenhouse gases responsible for climate change, but methane is the most potent in terms of altering the Earth's climate. In many regions of the world, methane is a significant source of energy for life on the sea floor and provides food for a wide variety of microorganisms. Since methane is a greenhouse gas, researchers want to understand precisely how these microorganisms process and convert it into other chemical compounds. Anaerobic oxidation of methane (AOM) is one of the mechanisms for this conversion and is used to detect microbial activity.

Radio-tracer studies can be used to measure AOM, but they are quite expensive, time-consuming, and subject to particular licensing requirements. As a result, a group of researchers at the University of Georgia received funding to collect soil samples from the Gulf of Mexico floor in an effort to discover proxies that can serve as AOM estimators to streamline the process. Many chemical compounds from soil samples can be easily tested “at-sea,” as opposed to AOM, which requires a shore-based facility. The soil samples were measured by four separate ships at 21 different sites, each of which can be classified as one of three types of site. The soil samples were tested for chemical compounds almost immediately. The samples are then maintained well below freezing to prevent any further chemical activity before being analyzed for AOM and other measures that can only be performed at the facility.

### 2.2 Objectives

This consultation's primary objective is to provide answers to the following two questions:

1. Can we predict AOM or identify the presence of AOM with any sort of accuracy using ”at-sea” measurements?
2. What attributes of a soil sample are associated with AOM?

## 3 Project Approach

### 3.1 Data Description

The client's raw data is composed of 275 observations and 14 features. The data contains the following features:

- Sample ID: The unique ID assigned to each soil sample collected at sea.
- Ship ID: The ship the soil sample was collected on, denoted by a letter from A through D.
- Sed Depth: The depth of the soil sample, with the oily layer on top of the sea floor serving as the baseline. This feature's values are ”Oily Layer” and a set of varying ranges denoting the distance below the oily layer in centimeters.
- Site: The location of the soil sample, denoted by a letter from A through E.

- Site Type: The type of site the soil sample was collected from. Sites can be classified as either "shelf", "abyssal", or "oil seep".

- AOM: The rate of anaerobic oxidation of methane. The units for AOM are not specified, but are on the same scale for all observations.

- $CH_4$: The methane concentration in a soil sample expressed in micromoles.

- Sulfide: The sulfide concentration in a soil sample expressed in millimoles.

- $SO_4$: The sulfate concentration in a soil sample expressed in millimoles.

- $NO_3$: The nitrate concentration in a soil sample expressed in micromoles.

- $NO_2$: The nitrogen dioxide concentration in a soil sample expressed in micromoles.

- $NH_4$: The ammonium concentration in a soil sample expressed in micromoles.

- POC: The amount of particulate organic carbon present in the soil sample. The units for POC are not specified, but are on the same scale for all observations.

- DOC: The amount of dissolved organic carbon present in the soil sample. The units for DOC are not specified, but are on the same scale for all observations.

## 3.2 Data Cleaning

The following six actions were taken to process the raw data:

1. Sample ID, Ship ID, Site, POC, and DOC were removed.

2. Outliers within AOM were removed.

3. Site Type was converted into a factor data type.

4. In Sed Depth, midpoints were imputed for the varying ranges, and 0 was imputed for "Oily Layer".

5. AOM was converted into a binary variable to produce a new feature.

6. The NA values in $CH_4$, Sulfide, $SO_4$, $NO_3$, $NO_2$, and $NH_4$ were imputed with predicted values.

### 3.2.1 Removal of Irrelevant Columns

Sample ID was removed because the soil samples are independent and identically distributed. Furthermore, we already know there are 275 observations. When AOM is the subject of concern, there is no need for a feature that simply counts up to 275. Ship ID was removed because the client could not provide information about the type of ships the soil samples were collected on. The only way to distinguish the ships is by their labeling, which is designated by a letter A through D. As a result, even if a ship was discovered to be significant, we have no idea what type of ship it is. The justification for removing Site is the same as the justification for removing Ship ID. The client was unable to provide geographical information about these sites, such as specific coordinates, and the only way to distinguish the sites is through their labeling, which is signified by a letter A through U. Therefore, even if a site was determined to be significant, its location is unknown. POC and DOC were removed because, according to the client, they aren't ”at-sea” measurements and are therefore irrelevant for this analysis.

### 3.2.2 Removal of Outliers

Outliers are typically unexpected irregularities that happen during the data creation process. Outliers may therefore weaken the impact of any trends an analysis might uncover. In order to avoid bias, outliers must be addressed. We determined an observation is an outlier if its AOM value $> 171$. This threshold provided the best balance between accuracy and significance. Using this criterion, we identified four observations as outliers, which we then removed from the data.

### 3.2.3 Conversion to Factor Data Type for Site Type

Site Type was initially a character data type. We wish to use this feature as a categorical predictor, so we converted this feature to a factor data type using the `as.factor()` function.

### 3.2.4 Imputation of Midpoints in Sed Depth

Closer examination of Sed Depth revealed that the values in Sed Depth can fall into 24 different categories. A post-analysis interpretation will be difficult to make sense of with this many categories. Therefore, we decided it would be better to make this feature a numeric data type by imputing the midpoint for varying ranges and the value of 0 for “Oily Layer” as it serves as a baseline for the different ranges in this feature. For example, “6-9 cm” is a possible category for a value in this feature. Thus, the midpoint is $\frac{6+9}{2} = 7.5$ (cm).

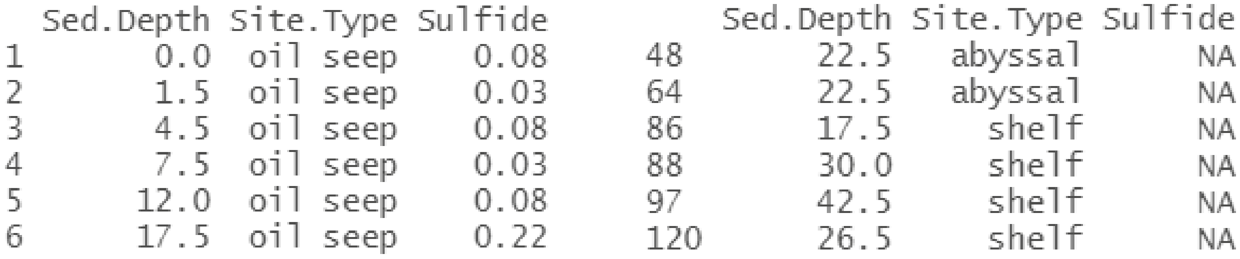
However, even after the midpoints were imputed, the feature remained a character data type. To convert the feature to a numeric data type, we utilized the `as.numeric()` function.

### 3.2.5 Conversion to Binary Variable for AOM

There is a poor chance of creating a relatively accurate model for forecasting AOM because 154 out of 275 AOM observations, or more than half of the AOM observations, have a value of 0. When AOM = 0, it signifies that the AOM process could not be detected by the device because it was below the detection threshold, which is unknown to the client and us. Therefore, we chose to identify the presence of AOM instead. We added a new feature named "Presence" that converts AOM into a binary variable. In Presence, a value is labeled as 1 if AOM $\geq 0.01$ and 0 if AOM $< 0.01$ by using the `ifelse()` function. Given that 0.01 is the lowest AOM value other than 0 in the data, 0.01 was chosen as the threshold.

### 3.2.6 Imputation of Predicted Values

In addition to the 24 NA values in AOM, there are 77 NA values spread among the other features. The sample size is already quite small, thus it would be detrimental to omit any observations with missing values. Upon further inspection, we found that neither Sed Depth nor Site Type have any values missing. Rather than removing any observations with missing values, we may leverage the fact that we have all of the values for these two features for every observation to predict the missing values in the other features. The features with missing values are $CH_4$, Sulfide, $SO_4$, $NO_3$, $NO_2$, and $NH_4$. To predict the missing values in these features, we divide each feature into two data frames: one with no missing values for the respective feature and another with just missing values for respective feature. All data frames will also contain the Sed Depth and Site Type values associated with the relevant observation. The two data frames for the Sulfide feature are depicted in Figure 1 as an example.



```
  Sed.Depth Site.Type Sulfide
1       0.0  oil seep    0.08
2       1.5  oil seep    0.03
3       4.5  oil seep    0.08
4       7.5  oil seep    0.03
5      12.0  oil seep    0.08
6      17.5  oil seep    0.22
```

No Missing Values for Sulfide

```
    Sed.Depth Site.Type Sulfide
48       22.5   abyssal      NA
64       22.5   abyssal      NA
86       17.5     shelf      NA
88       30.0     shelf      NA
97       42.5     shelf      NA
120      26.5     shelf      NA
```

Only Missing Values for Sulfide

Figure 1: First Six Observations of Sulfide Data Frames

Then, for each feature, we constructed a model that predicts the missing values using a combination of the `predict()` function and `rlm()` function from the `MASS` package. Following that, we imputed the missing values with the predicted values.

## 3.3 Exploratory Data Analysis (EDA)

We began the EDA process by plotting a histogram of AOM, the response, to visualize its distribution. Figure 2 illustrates the values in the response are heavily right-skewed. There is a significant peak where AOM = 0, which is not surprising given that more than half of the observations had an AOM value of 0.

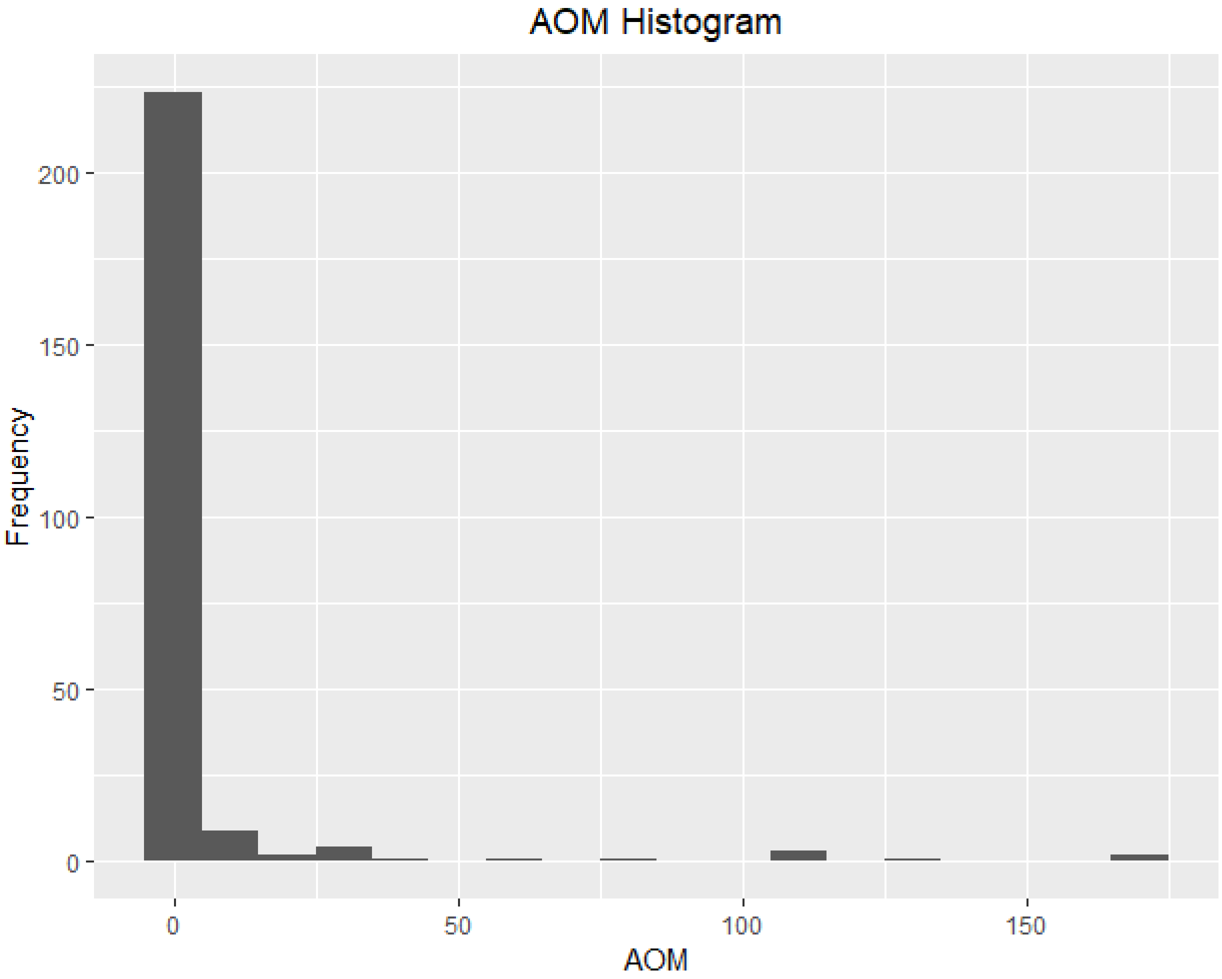


Figure 2: Distribution of AOM

We then created a Q-Q plot, which compares the residual distribution to the theoretical quantile of a normal distribution. On the Q-Q plot, right-skewed data will have an upward curve. Figure 3 further illustrates how heavily skewed the response is to the right. Thus, the data for AOM does not meet the assumption of homoscedasticity.

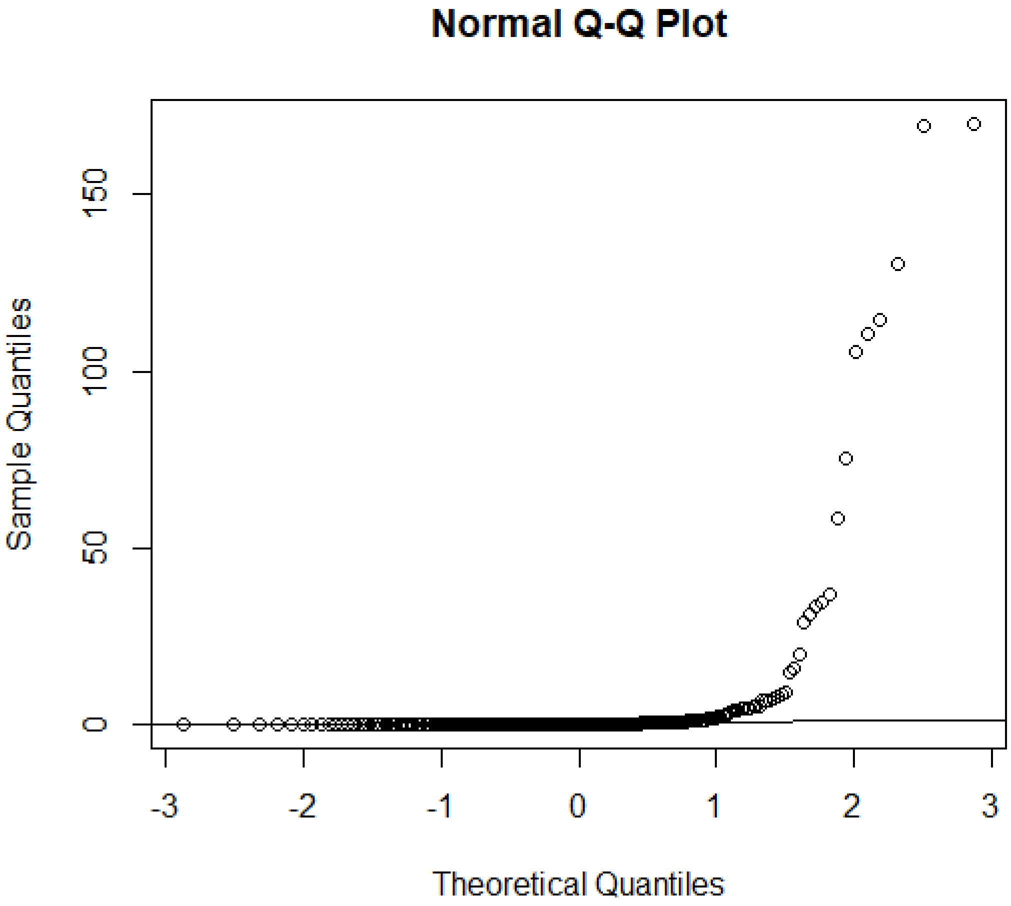


Figure 3: Q-Q Plot of AOM

To analyze the group means of the one factor feature we have, we subsequently generated a boxplot for Site Type. According to Figure 4, it appears Site Type caused noticeable differences in log (AOM). It appears that soil samples from oil seeps resulted in a higher AOM value. The medians of abyssal and oil seep were fairly comparable. However, the sampling was severely imbalanced, with 160 oil seep samples and only 23 abyssal samples. It is therefore still possible that their medians are different. Conversely, there is a clear difference between the medians of the oil seep and shelf. Furthermore, when examined closely enough, the notches in their boxplots, which indicate the approximate 95% confidence intervals for the medians, marginally overlap. In practice, we may argue that their notches do not overlap, and because the notches in their boxplots do not overlap, we can conclude with 95% confidence that their true medians differ.

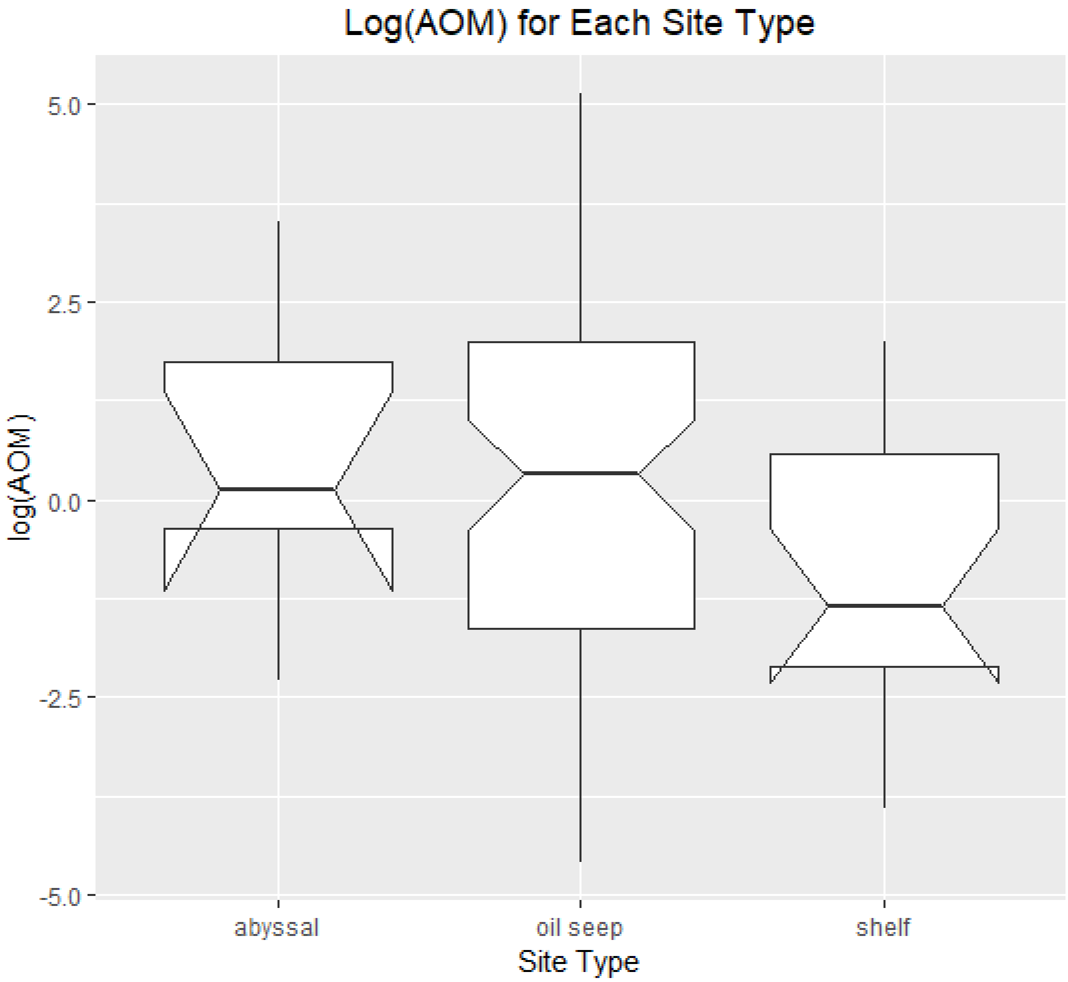
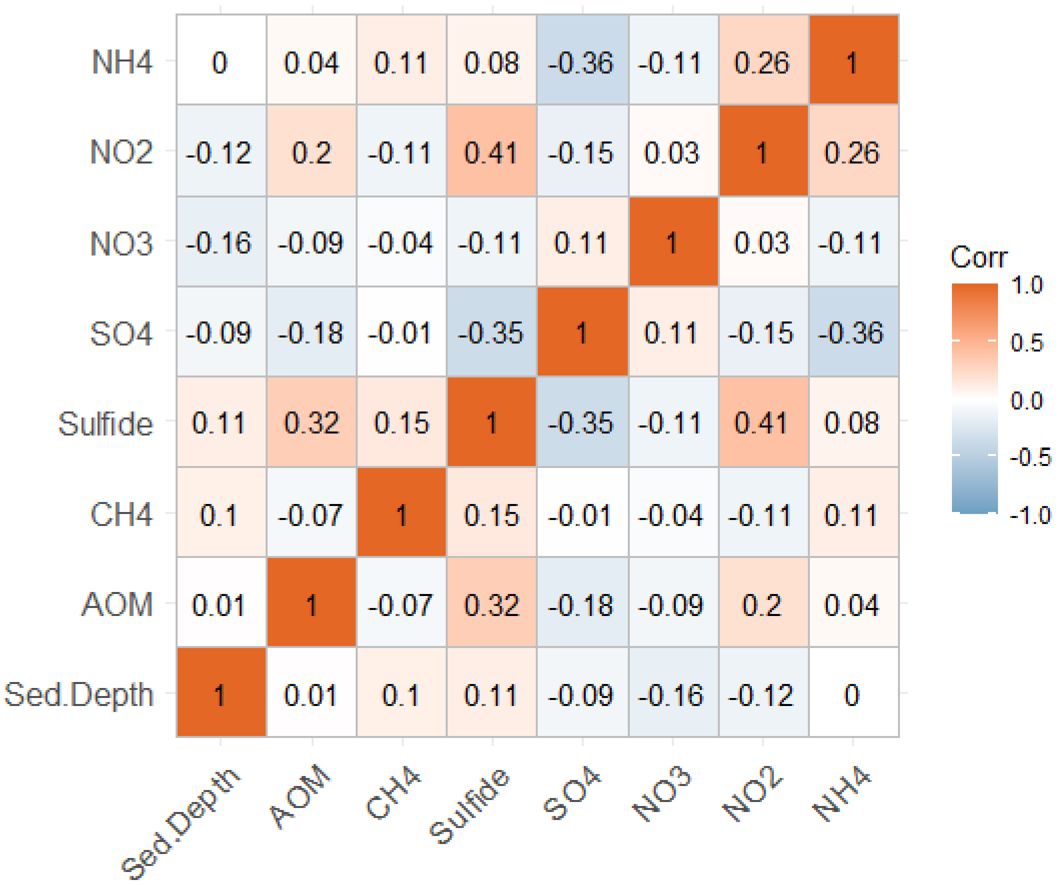


Figure 4: Site Type Group Means for log(AOM)

Finally, we created a correlation matrix of the quantitative features to see how all potential pairs of quantitative features correlate with one another. We are particularly interested in the correlation coefficients between AOM and all of the other features. Figure 5 illustrates there is no strong correlation between AOM and any of the other quantitative features. However, it does have some correlation with SO4, NO2, and sulfide. These correlation coefficients suggest that these three features may aid in addressing the client's two questions.



| | Sed.Depth | AOM | CH4 | Sulfide | SO4 | NO3 | NO2 | NH4 |
|---|---|---|---|---|---|---|---|---|
| NH4 | 0 | 0.04 | 0.11 | 0.08 | -0.36 | -0.11 | 0.26 | 1 |
| NO2 | -0.12 | 0.2 | -0.11 | 0.41 | -0.15 | 0.03 | 1 | 0.26 |
| NO3 | -0.16 | -0.09 | -0.04 | -0.11 | 0.11 | 1 | 0.03 | -0.11 |
| SO4 | -0.09 | -0.18 | -0.01 | -0.35 | 1 | 0.11 | -0.15 | -0.36 |
| Sulfide | 0.11 | 0.32 | 0.15 | 1 | -0.35 | -0.11 | 0.41 | 0.08 |
| CH4 | 0.1 | -0.07 | 1 | 0.15 | -0.01 | -0.04 | -0.11 | 0.11 |
| AOM | 0.01 | 1 | -0.07 | 0.32 | -0.18 | -0.09 | 0.2 | 0.04 |
| Sed.Depth | 1 | 0.01 | 0.1 | 0.11 | -0.09 | -0.16 | -0.12 | 0 |

Figure 5: Correlation Matrix of Quantitative Features

## 3.4 Methodologies

We implemented logistic regression, a supervised machine learning model, to answer the client's first question. Based on a given data set of independent features, logistic regression calculates the likelihood that an event will occur. A logit transformation is applied on the odds, the probability of success divided by the probability of failure, in logistic regression. The transformed odds are often referred to as log odds. In this model, the beta parameters, or coefficients, are often estimated via maximum likelihood estimation. This algorithm evaluates various beta values over a number of iterations to get the best match for the log odds. In order to determine the most accurate parameter estimate, logistic regression aims to maximize the log likelihood function, which is produced by all of these iterations. Following the identification of the best coefficients, the conditional probabilities for each observation can be computed, logged, and summed to provide a predicted probability. In the case of binary classification, a probability of less than 0.5 predicts 0 and a probability of more than or equal to 0.5 predicts 1. Due to the data's extreme skewness, as shown in Figure 1, we chose to employ a robust version of logistic regression. Additionally, Figure 2 demonstrated that the data violated the homoscedasticity assumption, which would result in inconsistent estimation of the beta parameters and their associated standard errors if we used a non-robust version.

To respond to the client's second question, we implemented a test for association between paired samples based on Kendall's $\tau$, a Pearson's product moment correlation coefficient. Kendall's $\tau$ is a nonparametric measure of the strength and direction of association between two features measured on at least an ordinal scale. If the data fails one or more of the assumptions of this Pearson's product-moment correlation coefficient test, it is regarded as a nonparametric alternative. It is also regarded as an alternative to the nonparametric Spearman rank-order correlation coefficient, especially when the data has a small sample size with many tied ranks.

# 4 Results

## 4.1 Robust Binary Logistic Regression

First, we divided the data into a training set and a test set using an 80/20 split using the `sample.int()` function. We also dropped AOM because it is associated with Presence, and including AOM in the model will result in multicollinearity. The next step was to estimate the following model, with Presence as the response, using the `glmrob()` function from the `robustbase` package:

$$\hat{y} = \beta_0 + \hat{\beta}_1 SedDepth + \hat{\beta}_2 SiteType + \hat{\beta}_3 CH_4 + \hat{\beta}_4 Sulfide + \hat{\beta}_5 SO_4 + \hat{\beta}_6 NO_3 + \hat{\beta}_7 NO_2 + \hat{\beta}_8 NH_4$$

Afterwards, we used the `predict()` function to calculate the predicted probabilities of the Presence values in the test data set. Then, we created a confusion matrix using the `table()` function. A confusion matrix is a table that is used to summarize and visualize a classification algorithm's performance. The confusion matrix is made up of four main properties (numbers) that are utilized to define the classifier's measuring metrics. These four numbers are:

1. True Positive (TP): The observation is predicted to be positive and is in fact positive.

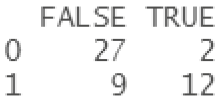
2. False Positive (FP): The observation is predicted to be positive but is actually negative.

3. True Negative (TN): The observation is predicted to be negative and is in fact negative.

4. False Negative (FN): The observation is predicted to be negative but is actually positive.

*Note: In our case, positive indicates being classified as 1, and negative indicates being classified as 0.*

A confusion matrix is shown in Figure 6 as an example. In Figure 6, TP = 12, FP = 2, TN = 27, and FN = 9.



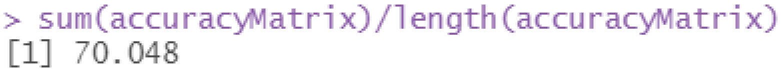
```
  FALSE TRUE
0    27    2
1     9   12
```

Figure 6: Confusion Matrix

These four numbers can be used to calculate the accuracy of our robust binary logistic regression model.

$$Accuracy = \frac{TN + TP}{TN + FP + FN + TP} * 100 = \frac{27 + 12}{27 + 2 + 9 + 12} * 100 = 78\%$$

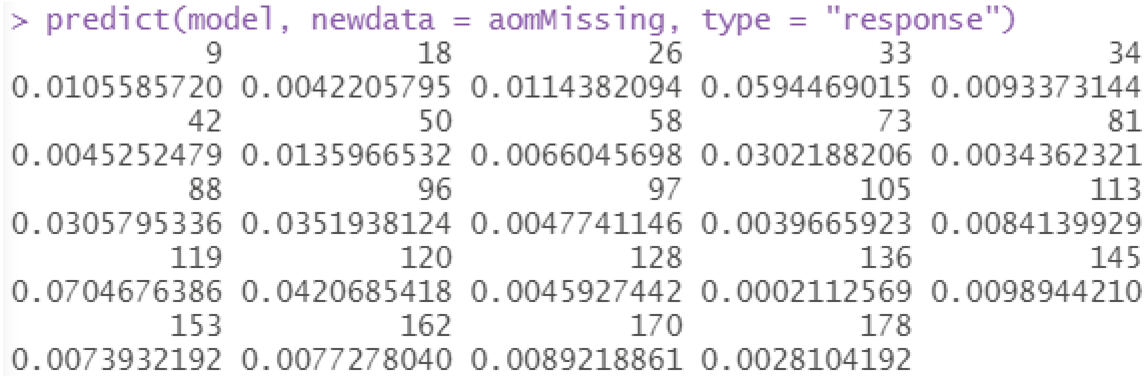
We repeated this process 999 times more, for a total of 1000 times, to obtain an average accuracy of our classification algorithm because splitting the data into a training and test set is a random process that is dependent on the seed of R's random number generator. To calculate the average accuracy, we generated an empty vector that held the computed accuracy for each iteration. We then used a for loop to repeat the procedure 1000 times, and for each iteration, the computed accuracy was added to the vector. Finally, we divided the sum of the now-complete vector by 1000, the number of iterations. Our results showed an average accuracy of 70.048%.



```
> sum(accuracyMatrix)/length(accuracyMatrix)
[1] 70.048
```

Figure 7: First Average Accuracy

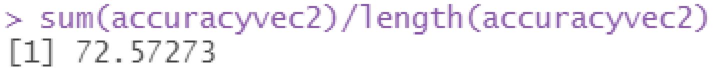
With a logistic regression model in hand, we can now utilize it to predict the missing values in AOM. We initially subset all data set observations for AOM that have a missing value. We can then forecast the missing values using the `predict()` function together with our logistic regression model and data subset. The `type = "response"` argument instructs R to output probabilities of the form P(Y = 1—X). Consequently, the values in Figure 8 depict the probabilities that the missing values in AOM are equal to 1. For a value to be classified as 1, it must be greater than or equal to 0.5. Since none of the values have a probability greater than or equal to 0.5, they will all be classified as 0. The missing values in AOM can then be imputed with 0. Our data set no longer contains any missing values.



```
> predict(model, newdata = aomMissing, type = "response")
           9           18           26           33           34 
0.0105585720 0.0042205795 0.0114382094 0.0594469015 0.0093373144 
          42           50           58           73           81 
0.0045252479 0.0135966532 0.0066045698 0.0302188206 0.0034362321 
          88           96           97          105          113 
0.0305795336 0.0351938124 0.0047741146 0.0039665923 0.0084139929 
         119          120          128          136          145 
0.0704676386 0.0420685418 0.0045927442 0.0002112569 0.0098944210 
         153          162          170          178 
0.0073932192 0.0077278040 0.0089218861 0.0028104192 
```

Figure 8: Predicted Probabilities of Missing Values in AOM

Given that we have a complete data set, we repeated the logistic regression process to see if we could construct a model with a greater average accuracy. Our results showed an average accuracy of 72.573%. Using the complete data set with the predicted values increased our average accuracy by approximately 2.5%.

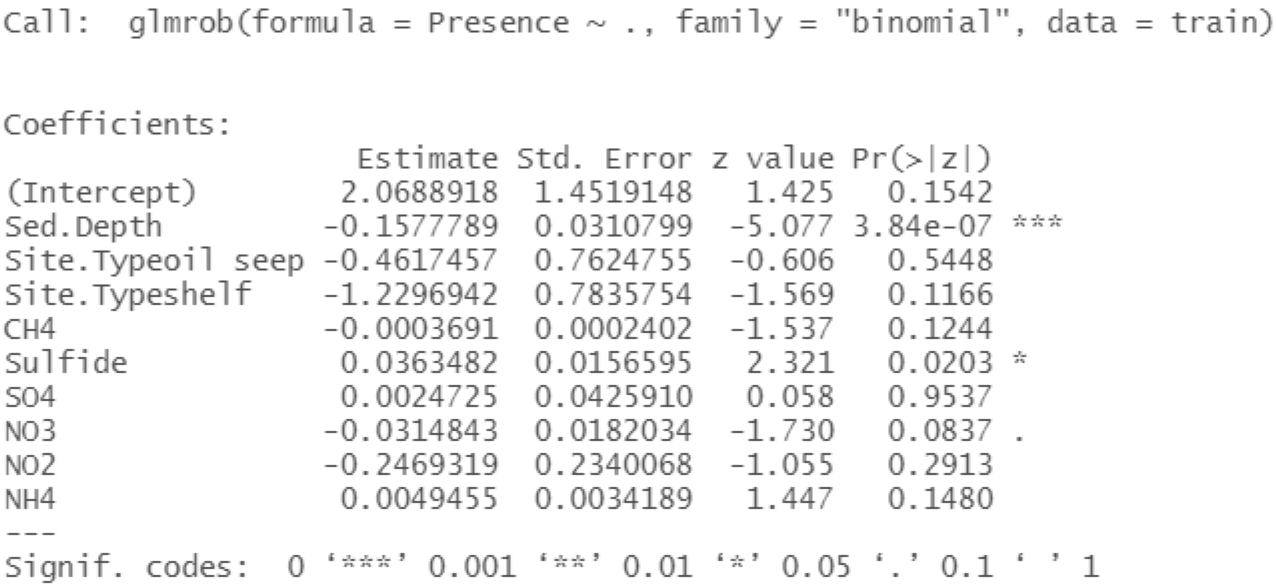


```
> sum(accuracyvec2)/length(accuracyvec2)
[1] 72.57273
```

Figure 9: Second Average Accuracy

## 4.2 Test for Association Using Kendall's $\tau$

To determine whether any of the other features are associated to AOM, we used the `cor.test()` function with the `method = "kendall"` argument. Unfortunately, because Site Type is a categorical feature, we are unable to apply this function with it. However, based on our logistic regression model, we can observe that Site Type is not statistically significant at $\alpha = 0.05$, implying Site Type is unlikely to be associated with AOM. Figure 10 displays the logistic regression model's output.

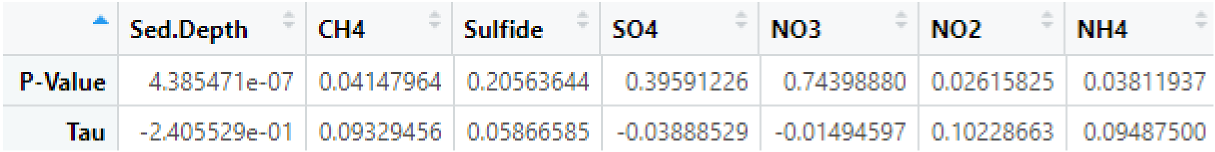


```
Call:  glmrob(formula = Presence ~ ., family = "binomial", data = train)


Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       2.0688918  1.4519148   1.425   0.1542
Sed.Depth        -0.1577789  0.0310799  -5.077 3.84e-07 ***
Site.Typeoil seep -0.4617457  0.7624755  -0.606   0.5448
Site.Typeshelf   -1.2296942  0.7835754  -1.569   0.1166
CH4              -0.0003691  0.0002402  -1.537   0.1244
Sulfide           0.0363482  0.0156595   2.321   0.0203 *
SO4               0.0024725  0.0425910   0.058   0.9537
NO3              -0.0314843  0.0182034  -1.730   0.0837 .
NO2              -0.2469319  0.2340068  -1.055   0.2913
NH4               0.0049455  0.0034189   1.447   0.1480
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 10: Output of Robust Binary Logistic Regression Model

Figure 11 is a table we created that contains the p-values and estimated $\tau$'s from the test for each feature. The $\tau$ coefficient can range from -1 to 1, with -1, 1, and 0 indicating a perfect negative relationship, a perfect positive relationship, and no relationship, respectively. The p-values for Sed Depth, $CH_4$, $NO_2$, and $NH_4$ are statistically significant at $\alpha = 0.05$, as shown in Figure 11. Based on these results, Sed Depth, CH4, NO2, and NH4 are associated with AOM. Sed Depth is negatively correlated with AOM, whereas CH4, NO2, and NH4 are positively correlated with AOM.



| | Sed.Depth | CH4 | Sulfide | SO4 | NO3 | NO2 | NH4 |
|---|---|---|---|---|---|---|---|
| P-Value | 4.385471e-07 | 0.04147964 | 0.20563644 | 0.39591226 | 0.74398880 | 0.02615825 | 0.03811937 |
| Tau | -2.405529e-01 | 0.09329456 | 0.05866585 | -0.03888529 | -0.01494597 | 0.10228663 | 0.09487500 |

Figure 11: Second Average Accuracy

# 5 Conclusions and Recommendations

## 5.1 Conclusions

In this report, we are able to adequately answer both of the client's questions. In response to the first question, while we were unable to predict AOM, we were able to identify its presence with moderate accuracy using ”at-sea” measurements by implementing robust binary logistic regression. An average accuracy of at least 80% is what we would ideally like for our model.

Even if our model's accuracy is less than ideal, at around 73% on average, it can still save a significant amount of money, time, and labor. In response to the second question, our results from the test for association using Kendall's Tau revealed that the attributes of a soil sample that are associated with AOM are Sed Depth, $CH_4$, $NO_2$, and $NH_4$. Of the four features, Sed Depth has the strongest correlation with AOM.

## 5.2 Recommendations

The following bullet points list the actions I advise taking to improve the model and, ideally, achieve an average accuracy of at least 80%:

- Identify the unit of measurement for AOM.
- Identify each site's specific coordinates and/or geographical information so Site can be used as a feature in the analysis.
- Identify the types of ships the soil samples were collected from so Ship ID can be used as a feature in the analysis.
- Determine the detection limit of the equipment used to measure AOM so that there are no more zero values. For analysis, a value of 0.001, for example, is preferable over 0. Request the technical details of the equipment the researchers utilized, if at all possible.
- To improve the logistic regression model's accuracy, continue adding more soil sample data to it. Machine learning algorithms, such as logistic regression, improve with more data.