Round #1

* If I wanted to categorize a bunch of students based on data points about them, but did not have a specific target, would this be a supervised or unsupervised problem?
* If you had a an X variable (Class below) that was structured as below, how might you structure this for a machine learning problem if you were predicting Grade.

| Class | Grade |
| --- | --- |
| Freshman | A |
| Sophmore | B |
| Freshman | B |

* In the previous dataset, if I was predicting their letter grade, what would be an appropriate evaluation metric?

Round #2

* If you wanted to find what words are important to a given document, while discounting words that are used commonly across all documents, which sklearn function would you use?
* Which sklearn function is it important to run on your data prior to using the KMeans algorithm for clustering?

1. What is the primary package for machine learning in Python?

2. Why is using training and testing data important?

3. What type of data is Grade from the first round?

Round #3

2. What are the 2 biggest features (uses) of Git?

3. What type of visual would be appropriate if you wanted to measure the relationship between a categorical and a continuous variable? Please show an example.

4. What is the primary CLI (Command Line Interface) for Mac OS? For Windows?

5. What CLI command (Terminal or CMD) would I use to list the files in a given directory

What is 1000 %% 5? What about 1000 %% 3?

What would happen if I ran the following code in R?

Var\_a <- “10”

Var\_b <- Var\_a + 10

Round #4

1. Write 2 for-loops to loop over the following list and prints the result +1, one which iterates over the VALUE, one which iterates over the INDEX

2. From the table above with grade and class, print the head of the column Grade

3. If I wanted to test if ***nationality*** had any impact on ***count of blood cells***, which statistical test might I use?

a.

ANOVA

b.

Mean

c.

Correlation

d.

crosstab

4. If I had a HUGE dataset, which Pandas was struggling with memory-wise, is there an alternative package that might be useful?

5. If was doing a supervised regression problem with a numeric/continuous target variable, what 2 models that we learned in class would be appropriate?