1. Perform pre-processing on the full dataset here: <https://www.kaggle.com/sohier/calcofi>
2. Perform feature selection (decide what columns should be included in your analysis). How did you figure out what features were important?
3. Perform multiple linear regression on the dataset. Be sure to display the final resulting equation with the coefficients and intercept in markdown.
4. Which regression approach had the “best” performance? Be sure to compare it to what you did last week.
5. Work with the diabetes dataset to perform multiple logistic regression.
6. What are the other ways to assess the performance of the model besides accuracy? What approach makes the most sense with the diabetes dataset? Why?
7. Tune your model to improve performance based on the performance metric you identified in question 6. Look at the documentation for logistic regression in statsmodel and/or sklearn and see if you can tune the model performance based on available parameters. Please note: data scientists need to constantly learn new approaches to optimizing models. This is one way of doing that.
8. Is the KNN model or the multiple logistic regression model more performant for the diabetes dataset? How can you tell?