Homework 2
Due: Feb 6, 2025

This homework uses data from a project sponsored by the Department of Energy (DOE). The data are quite limited, which often justifies the application of generative AI, but they are of good quality. They are posted on CANVAS (DATA_DOE_project). Check the file to determine how the measurements were stored. Our focus is on the measurements reported in Appendix 4.

Extract the pertinent measurements of 118 rock samples whose data are reported as "ROUTINE (AMBIENT)." The extracted measurements of each rock sample should include:

a. Sample depth (ft)
b. In-situ Klinkenberg permeability (mD)
c. Routine porosity
d. Mercury injection capillary pressure (psi) vs cumulative wetting phase saturation (% pore vol)

1. As part of data wrangling, remove all samples whose pertinent measurements are incomplete after checking the four categories (a, b, c, d). Report the number of removed samples.

2. Remove samples whose permeability values are outliers. Show how the outliers were determined and report the number of removed samples.

3. Show which feature-engineering methods help you better analyze the data. Elaborate your answer and provide relevant plots.

4. Present the cross plots of the first three properties (a, b, c).

5. Determine the correlation coefficients of the first three properties after feature engineering. Show whether the feature engineering really has significant effects on the coefficients.

6. Discuss whether the correlation coefficients capture the dependency in this example.