**Elevation Prediction of Soil Sample**

The EDA conducted on the USGS data set was informative and provided much insight. Some, but not limited to the relationships between soil characteristics and elevation. I chose 5 variables to use to try and predict the elevation a soil sample was taken from. The variables used were lyr\_top, lyr\_bot, lyr\_mass, pro\_depth\_water, pro\_bedrock\_depth, and pro\_elevation. Of these, 2 of these variables were close to useless because they had so many outliers, they had very few valid points. With this many outliers I decided to throw those two out and use the remaining for my regression model. Those two were pro\_depth\_water and pro\_bedrock\_depth. The statistics highlighted spread and central tendencies of each. I then completed a PMF comparison fo lyr\_mass showing variations based on depth. I then completed a CDF of pro\_elevation showing how the elevation values were spread across the dataset. I then completed a regression model, and as expected some variables showed a stronger correlation than others. Lastly, I completed a hypothesis test using t-test results. The t-test showed strong statistical evidence to reject the null hypothesis.

Some items could have been missed during this test. There were so many variables to choose from in this data set. I may choose different variables next time. One thing I could do as well, as I got comfortable with, is create new variables with existing variables, to make a more focused approach.

Additional variables I would like to include would possibly be pH, water saturation, amount of fertilizer or other organic matter. I would also like to use another country’s data covering the same variables. If I was to include those in my regression model it could possibly make it more accurate.

One incorrect assumption was that the variables had a linear relationship with elevation. There were some moderate correlations, but additional non-linear correlation models could be completed. Additionally, I could have removed pro\_depth\_water and pro\_bedrock\_depth and chose other variables with more values, but I wanted a more real world feel. That idea could have been “too much” and should have gone the road more travelled. This rolls into areas of uncertainty as well. Those two variables probably caused more uncertainty, and my findings would have been more robust and statically significant.