Data Analysis of Environmental Parameters of Soil

Data-151: Intro to Data Science

10/18/2024

By: Jack Colwell, Owen Galicia, Santiago Gutiérrez-Morales, and Lilu Smith

I have neither given, or received, nor tolerated others’ use of unauthorized aid.

* *Jack Colwell, Owen Galicia, Santiago Gutiérrez-Morales, and Lilu Smith*

1. Progress Made

Since our last proposal we have retrieved the data for our project from Jack Colwell and his research team. We have brought the dataset into a Google colab environment using Python as the scripting language. Shown in figure 1 is the data columns before cleaning.

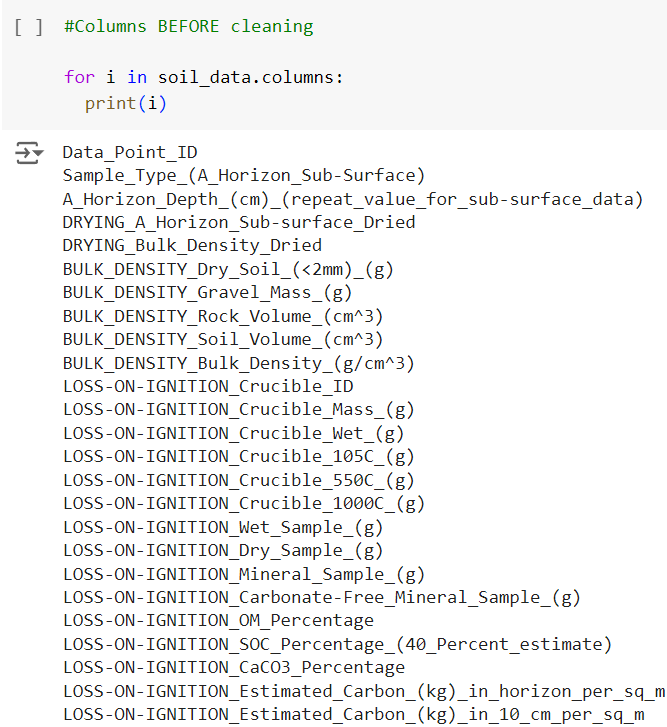


Figure 1. Data columns of our dataset before cleaning.

We then cleaned the data and removed columns that were simply markers of tasks to be done on the samples, as they are not important to analysis of the data. We noticed that in our original google spreadsheet that a few cell formulas had not been applied to those rows. That was fixed and helped to eliminate our missing values. Finally, we dropped three rows that had complete missing values, as those samples were not able to be collected correctly in the field. In doing so, we have now rectified all missing values as shown in figure 2.

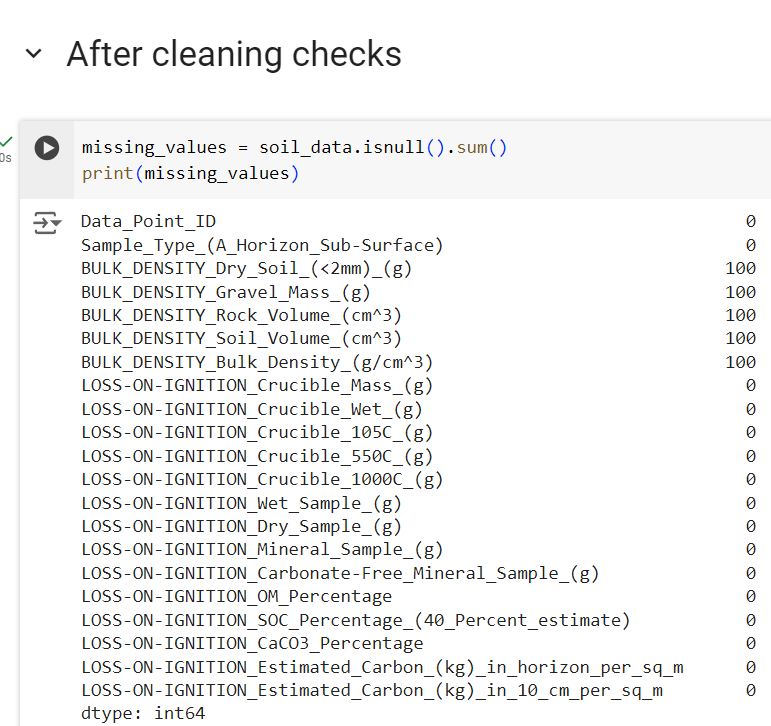


Figure 2. The data after all missing values had been rectified.

Furthermore, it was noted that bulk density is a measurement that can only be taken in the A horizon and therefore there will be missing data for the subsurface horizon in the complete dataset. Therefore, we broke the data into two subsets. One containing the sampling points (only A horizon measurements) with bulk density measurement, and the other containing the remaining data.

1. Analysis Conducted So Far

The analysis conducted so far includes a five number summary creation for all variables, as well as initial correlation matrix creation. We have done this for both of the datasets, the bulk density dataset with values and the other dataset without any bulk density data. The five number summary has shown us that the BULK\_DENSITY\_Dry\_Soil <2 mm has a large range, while most of the other variables are closer in range of 22 to 36 and 0 to 15 as a max value. Looking at the first quartile to the third quartile for many of the variables, the numbers are close together. This shows that the spread of the data is not very large so there is less variation and there is not a skewed distribution. From the correlation matrix, we can see that there are several negative correlations, meaning that as one variable increases, the other decreases, indicating these variables move in opposite directions. Along the diagonal, there is a high number of positive correlations, meaning that is something we should take into consideration when deciding on which variables to analyze more deeply. Future work includes the creation of box plots and deeper analysis of the correlation matrix.

A couple things that have been difficult throughout this process is knowing which variables to look at. We had to be very diligent with cleaning the data and getting rid of the missing value rows. Majority of our group is finding out that soil data has complex interdependencies so uncovering and properly modeling these relationships can be difficult when most of us do not have this initial knowledge about our data. Soil data typically involves a variety of metrics related to the physical, chemical, and biological properties of soil, such as pH, moisture, nutrient levels, organic matter, and bulk density. These metrics are essential for understanding soil health, fertility, and its suitability for plant growth. Analyzing soil data can be challenging due to the natural variability of soil across different regions, temporal changes, and sometimes one does not have the means to be able to collect samples due to weather conditions impacting the A horizon.

1. Initial Choices About Predictive Modeling

We are in the very early stages of predictive modeling. We are currently analyzing the correlation matrix of variables for both the datasets including bulk density, and the dataset that does not. As of present, there are some correlations that will need to be removed because of autocorrelation due to mathematical conversions. We are still planning on running ANOVA analysis with the highly correlated variables to see which specific categories impact the soil on being suitable for agriculture.

1. Data Reduction Methods

The data reduction method we used was sampling. This is creating a subset of data from the larger set to analyze. In this instance, we decided to create a subset of the bulk density dry soil column. Because this variable is only measured in the A horizon, the subsurface rows do not have a measurement, so we split the data into a non empty bulk density dataset and an empty bulk density dataset. This helped us create our correlation matrix between all the variables.

1. Remaining Work Schedule

We will continue to work on figuring out which predictive model is best for our data and finalize it. We should start writing our final report with an overview of the project, our data and methods used, and conclusions from our exploratory data analysis. Once this draft is complete, we will present our findings and receive feedback from our peers on any improvements to make. The paper will be revised and edited, then turned in.