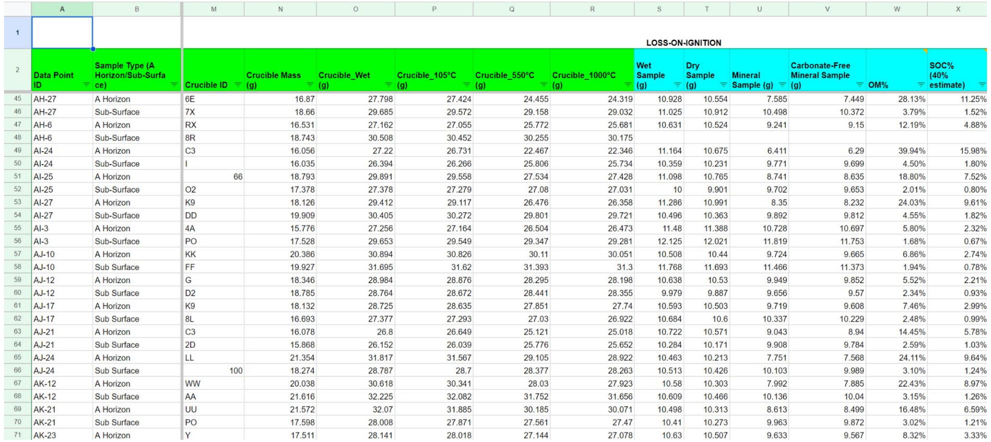
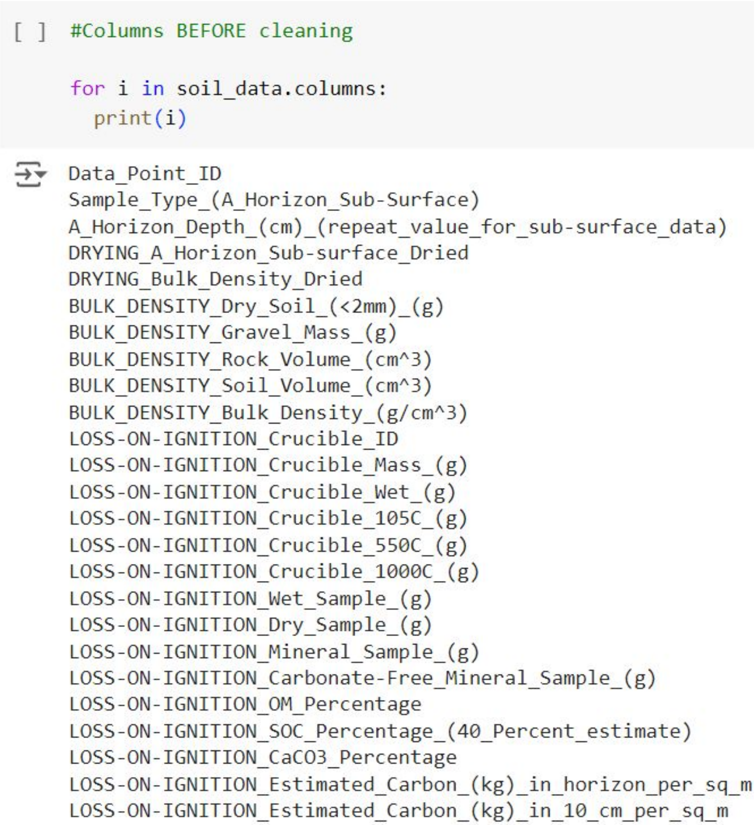
Data 151 Project Part 2

Lilu Smith, Owen Galicia, Jack Colwell, Santiago Gutierrez Morales

Original Data and Columns



	A	B	M	N	O	P	Q	R	S	T	U	V	W	X
1									LOSS-ON-IGNITION					
2	Data Point ID	Sample Type (A Horizon/Sub-Surface)	Crucible ID	Crucible Mass (g)	Crucible_Wet (g)	Crucible_105°C (g)	Crucible_550°C (g)	Crucible_1000°C (g)	Wet Sample (g)	Dry Sample (g)	Mineral Sample (g)	Carbonate-Free Mineral Sample (g)	OM%	SOC% (40% estimate)
45	AH-27	A Horizon	6E	16.87	27.798	27.424	24.455	24.319	10.928	10.554	7.585	7.449	28.13%	11.25%
46	AH-27	Sub-Surface	7X	18.66	29.685	29.572	29.158	29.032	11.025	10.912	10.498	10.372	3.79%	1.52%
47	AH-6	A Horizon	RX	16.531	27.162	27.055	25.772	25.681	10.631	10.524	9.241	9.15	12.19%	4.88%
48	AH-6	Sub-Surface	8R	18.743	30.508	30.452	30.255	30.175						
49	AI-24	A Horizon	C3	16.056	27.22	26.731	22.467	22.346	11.164	10.675	6.411	6.29	39.94%	15.98%
50	AI-24	Sub-Surface	I	16.035	26.394	26.266	25.806	25.734	10.359	10.231	9.771	9.699	4.50%	1.80%
51	AI-25	A Horizon	66	18.793	29.891	29.558	27.534	27.428	11.098	10.765	8.741	8.635	18.80%	7.52%
52	AI-25	Sub-Surface	O2	17.378	27.378	27.279	27.08	27.031	10	9.901	9.702	9.653	2.01%	0.80%
53	AI-27	A Horizon	K9	18.126	29.412	29.117	26.476	26.358	11.286	10.991	8.35	8.232	24.03%	9.61%
54	AI-27	Sub-Surface	DD	19.909	30.405	30.272	29.801	29.721	10.496	10.363	9.892	9.812	4.55%	1.82%
55	AI-3	A Horizon	4A	15.776	27.256	27.164	26.504	26.473	11.48	11.388	10.728	10.697	5.80%	2.32%
56	AI-3	Sub-Surface	PO	17.528	29.653	29.549	29.347	29.281	12.125	12.021	11.819	11.753	1.68%	0.67%
57	AJ-10	A Horizon	KK	20.386	30.894	30.826	30.11	30.051	10.508	10.44	9.724	9.665	6.86%	2.74%
58	AJ-10	Sub Surface	FF	19.927	31.695	31.62	31.393	31.3	11.768	11.693	11.466	11.373	1.94%	0.78%
59	AJ-12	A Horizon	G	18.346	28.984	28.876	28.295	28.198	10.638	10.53	9.949	9.852	5.52%	2.21%
60	AJ-12	Sub Surface	D2	18.785	28.764	28.672	28.441	28.355	9.979	9.887	9.656	9.57	2.34%	0.93%
61	AJ-17	A Horizon	K9	18.132	28.725	28.635	27.851	27.74	10.593	10.503	9.719	9.608	7.46%	2.99%
62	AJ-17	Sub Surface	8L	16.693	27.377	27.293	27.03	26.922	10.684	10.6	10.337	10.229	2.48%	0.99%
63	AJ-21	A Horizon	C3	16.078	26.8	26.649	25.121	25.018	10.722	10.571	9.043	8.94	14.45%	5.78%
64	AJ-21	Sub Surface	2D	15.868	26.152	26.039	25.776	25.652	10.284	10.171	9.908	9.784	2.59%	1.03%
65	AJ-24	A Horizon	LL	21.354	31.817	31.567	29.105	28.922	10.463	10.213	7.751	7.568	24.11%	9.64%
66	AJ-24	Sub Surface	100	18.274	28.787	28.7	28.377	28.263	10.513	10.426	10.103	9.989	3.10%	1.24%
67	AK-12	A Horizon	WW	20.038	30.618	30.341	28.03	27.923	10.58	10.303	7.992	7.885	22.43%	8.97%
68	AK-12	Sub Surface	AA	21.616	32.225	32.082	31.752	31.656	10.609	10.466	10.136	10.04	3.15%	1.26%
69	AK-21	A Horizon	UU	21.572	32.07	31.885	30.185	30.071	10.498	10.313	8.613	8.499	16.48%	6.59%
70	AK-21	Sub Surface	PO	17.598	28.008	27.871	27.561	27.47	10.41	10.273	9.963	9.872	3.02%	1.21%
71	AK-23	A Horizon	Y	17.511	28.141	28.018	27.144	27.078	10.63	10.507	9.633	9.567	8.32%	3.33%

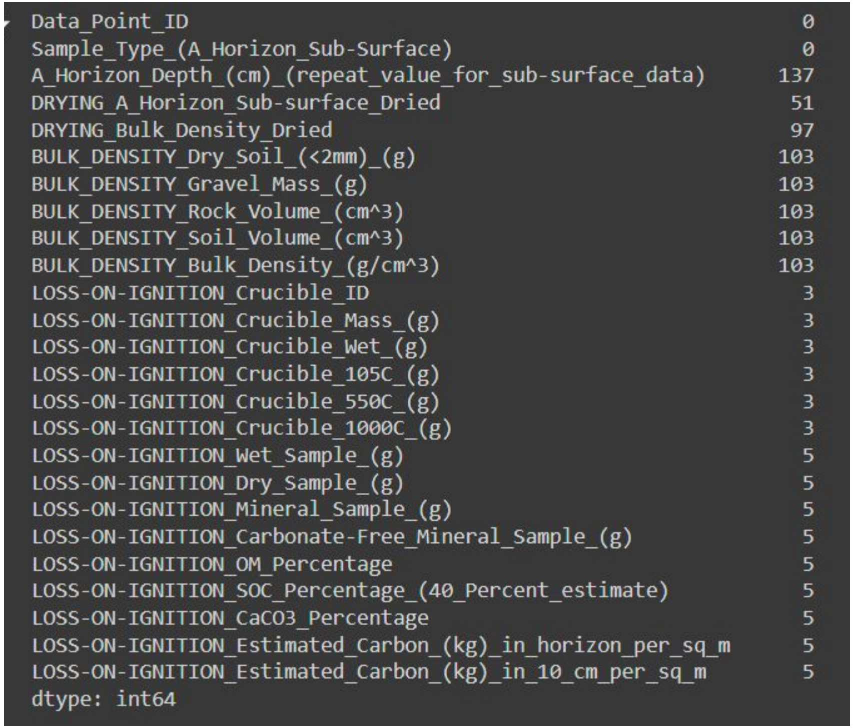


```
[ ]  #Columns BEFORE cleaning

     for i in soil_data.columns:
       print(i)
```

```
Data_Point_ID
Sample_Type_(A_Horizon_Sub-Surface)
A_Horizon_Depth_(cm)_(repeat_value_for_sub-surface_data)
DRYING_A_Horizon_Sub-surface_Dried
DRYING_Bulk_Density_Dried
BULK_DENSITY_Dry_Soil_(<2mm)_(g)
BULK_DENSITY_Gravel_Mass_(g)
BULK_DENSITY_Rock_Volume_(cm^3)
BULK_DENSITY_Soil_Volume_(cm^3)
BULK_DENSITY_Bulk_Density_(g/cm^3)
LOSS-ON-IGNITION_Crucible_ID
LOSS-ON-IGNITION_Crucible_Mass_(g)
LOSS-ON-IGNITION_Crucible_Wet_(g)
LOSS-ON-IGNITION_Crucible_105C_(g)
LOSS-ON-IGNITION_Crucible_550C_(g)
LOSS-ON-IGNITION_Crucible_1000C_(g)
LOSS-ON-IGNITION_Wet_Sample_(g)
LOSS-ON-IGNITION_Dry_Sample_(g)
LOSS-ON-IGNITION_Mineral_Sample_(g)
LOSS-ON-IGNITION_Carbonate-Free_Mineral_Sample_(g)
LOSS-ON-IGNITION_OM_Percentage
LOSS-ON-IGNITION_SOC_Percentage_(40_Percent_estimate)
LOSS-ON-IGNITION_CaCO3_Percentage
LOSS-ON-IGNITION_Estimated_Carbon_(kg)_in_horizon_per_sq_m
LOSS-ON-IGNITION_Estimated_Carbon_(kg)_in_10_cm_per_sq_m
```

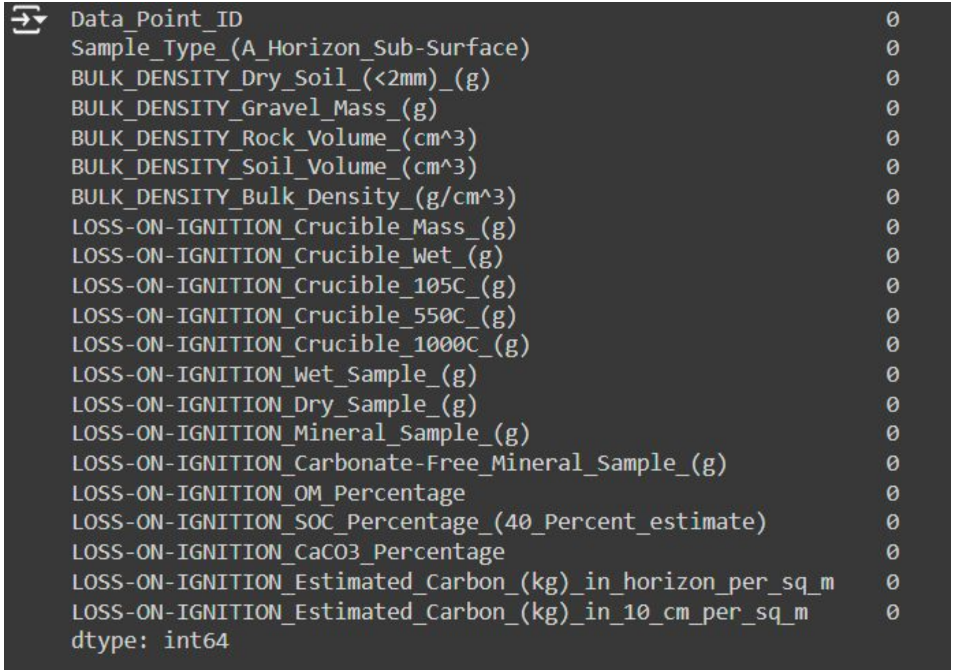
Null rows:

Before data cleaning



```
Data_Point_ID                                                0
Sample_Type_(A_Horizon_Sub-Surface)                          0
A_Horizon_Depth_(cm)_(repeat_value_for_sub-surface_data)   137
DRYING_A_Horizon_Sub-surface_Dried                          51
DRYING_Bulk_Density_Dried                                   97
BULK_DENSITY_Dry_Soil_(<2mm)_(g)                           103
BULK_DENSITY_Gravel_Mass_(g)                               103
BULK_DENSITY_Rock_Volume_(cm^3)                            103
BULK_DENSITY_Soil_Volume_(cm^3)                            103
BULK_DENSITY_Bulk_Density_(g/cm^3)                         103
LOSS-ON-IGNITION_Crucible_ID                                 3
LOSS-ON-IGNITION_Crucible_Mass_(g)                           3
LOSS-ON-IGNITION_Crucible_Wet_(g)                            3
LOSS-ON-IGNITION_Crucible_105C_(g)                           3
LOSS-ON-IGNITION_Crucible_550C_(g)                           3
LOSS-ON-IGNITION_Crucible_1000C_(g)                          3
LOSS-ON-IGNITION_Wet_Sample_(g)                              5
LOSS-ON-IGNITION_Dry_Sample_(g)                              5
LOSS-ON-IGNITION_Mineral_Sample_(g)                          5
LOSS-ON-IGNITION_Carbonate-Free_Mineral_Sample_(g)           5
LOSS-ON-IGNITION_OM_Percentage                               5
LOSS-ON-IGNITION_SOC_Percentage_(40_Percent_estimate)        5
LOSS-ON-IGNITION_CaCO3_Percentage                            5
LOSS-ON-IGNITION_Estimated_Carbon_(kg)_in_horizon_per_sq_m   5
LOSS-ON-IGNITION_Estimated_Carbon_(kg)_in_10_cm_per_sq_m     5
dtype: int64
```

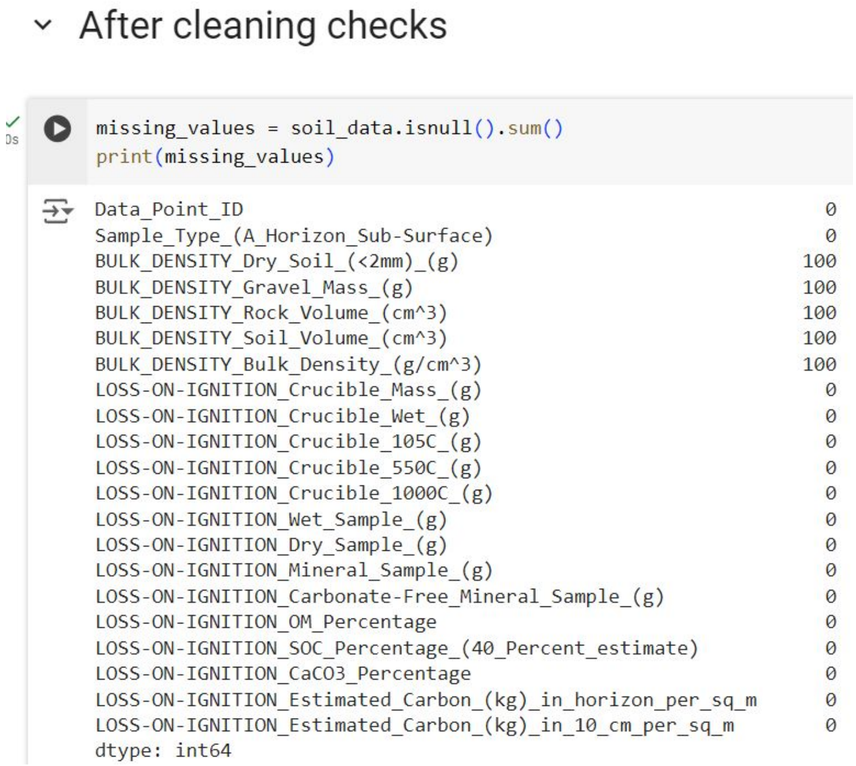
After data cleaning



```
Data_Point_ID                                                0
Sample_Type_(A_Horizon_Sub-Surface)                          0
BULK_DENSITY_Dry_Soil_(<2mm)_(g)                             0
BULK_DENSITY_Gravel_Mass_(g)                                 0
BULK_DENSITY_Rock_Volume_(cm^3)                              0
BULK_DENSITY_Soil_Volume_(cm^3)                              0
BULK_DENSITY_Bulk_Density_(g/cm^3)                           0
LOSS-ON-IGNITION_Crucible_Mass_(g)                           0
LOSS-ON-IGNITION_Crucible_Wet_(g)                            0
LOSS-ON-IGNITION_Crucible_105C_(g)                           0
LOSS-ON-IGNITION_Crucible_550C_(g)                           0
LOSS-ON-IGNITION_Crucible_1000C_(g)                          0
LOSS-ON-IGNITION_Wet_Sample_(g)                              0
LOSS-ON-IGNITION_Dry_Sample_(g)                              0
LOSS-ON-IGNITION_Mineral_Sample_(g)                          0
LOSS-ON-IGNITION_Carbonate-Free_Mineral_Sample_(g)           0
LOSS-ON-IGNITION_OM_Percentage                               0
LOSS-ON-IGNITION_SOC_Percentage_(40_Percent_estimate)        0
LOSS-ON-IGNITION_CaCO3_Percentage                            0
LOSS-ON-IGNITION_Estimated_Carbon_(kg)_in_horizon_per_sq_m   0
LOSS-ON-IGNITION_Estimated_Carbon_(kg)_in_10_cm_per_sq_m     0
dtype: int64
```

Data Cleaning

- Removed most of the omitted data because we can’t do much with empty data cells
- Added formulas to cells where the functions weren’t applied
- Dropped 3 rows with missing values in all columns (the data was not able to be collected in the field for those entries)



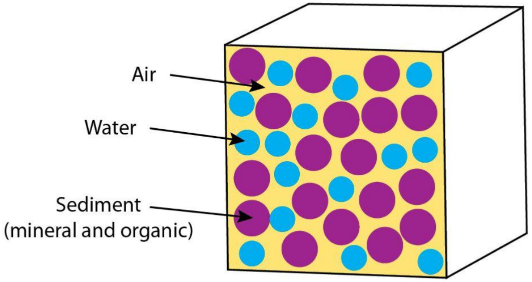
After cleaning checks

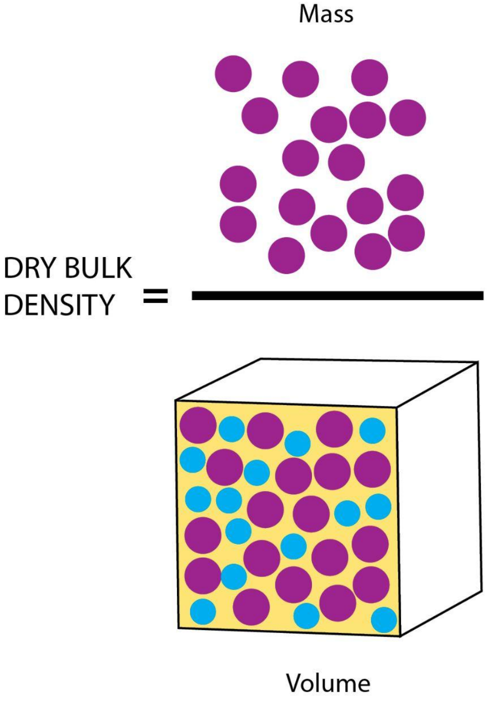
```
missing_values = soil_data.isnull().sum()
print(missing_values)
```

```
Data_Point_ID                                                   0
Sample_Type_(A_Horizon_Sub-Surface)                             0
BULK_DENSITY_Dry_Soil_(<2mm)_(g)                              100
BULK_DENSITY_Gravel_Mass_(g)                                  100
BULK_DENSITY_Rock_Volume_(cm^3)                               100
BULK_DENSITY_Soil_Volume_(cm^3)                               100
BULK_DENSITY_Bulk_Density_(g/cm^3)                            100
LOSS-ON-IGNITION_Crucible_Mass_(g)                              0
LOSS-ON-IGNITION_Crucible_Wet_(g)                               0
LOSS-ON-IGNITION_Crucible_105C_(g)                              0
LOSS-ON-IGNITION_Crucible_550C_(g)                              0
LOSS-ON-IGNITION_Crucible_1000C_(g)                             0
LOSS-ON-IGNITION_Wet_Sample_(g)                                 0
LOSS-ON-IGNITION_Dry_Sample_(g)                                 0
LOSS-ON-IGNITION_Mineral_Sample_(g)                             0
LOSS-ON-IGNITION_Carbonate-Free_Mineral_Sample_(g)              0
LOSS-ON-IGNITION_OM_Percentage                                  0
LOSS-ON-IGNITION_SOC_Percentage_(40_Percent_estimate)           0
LOSS-ON-IGNITION_CaCO3_Percentage                               0
LOSS-ON-IGNITION_Estimated_Carbon_(kg)_in_horizon_per_sq_m      0
LOSS-ON-IGNITION_Estimated_Carbon_(kg)_in_10_cm_per_sq_m        0
dtype: int64
```

Reason for Data Split

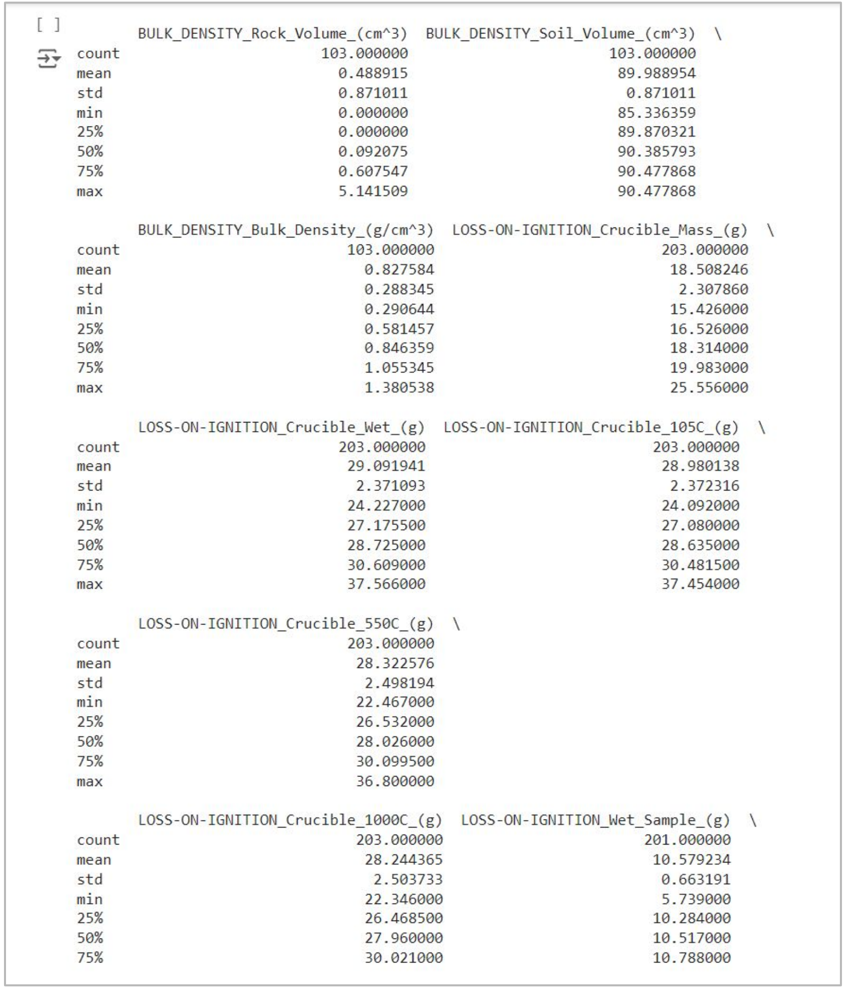
- Bulk density is a measurement that can only be taken in the A horizon and therefore there will be missing data for the subsurface horizon in the complete dataset
- Broke the data into two subsets. One containing the sampling points (only A horizon measurements) with bulk density measurement, and the other containing the remaining data.





Five Number Summary

- Added a five number summary to see the overall distribution of the data
    - Need data to be normally distributed for statistical tests done later

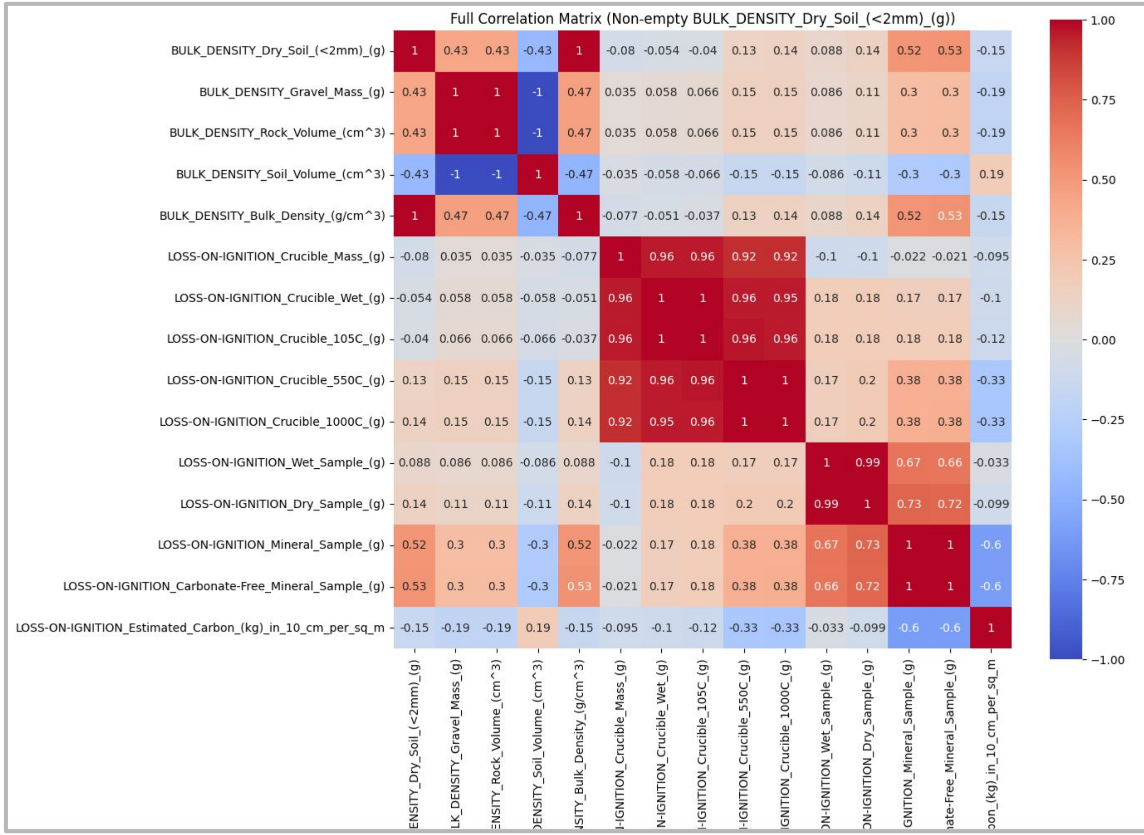


```
[ ]
          BULK_DENSITY_Rock_Volume_(cm^3)  BULK_DENSITY_Soil_Volume_(cm^3)  \
   count                       103.000000                       103.000000
   mean                          0.488915                        89.988954
   std                           0.871011                         0.871011
   min                           0.000000                        85.336359
   25%                           0.000000                        89.870321
   50%                           0.092075                        90.385793
   75%                           0.607547                        90.477868
   max                           5.141509                        90.477868

          BULK_DENSITY_Bulk_Density_(g/cm^3)  LOSS-ON-IGNITION_Crucible_Mass_(g)  \
   count                          103.000000                          203.000000
   mean                             0.827584                           18.508246
   std                              0.288345                            2.307860
   min                              0.290644                           15.426000
   25%                              0.581457                           16.526000
   50%                              0.846359                           18.314000
   75%                              1.055345                           19.983000
   max                              1.380538                           25.556000

          LOSS-ON-IGNITION_Crucible_Wet_(g)  LOSS-ON-IGNITION_Crucible_105C_(g)  \
   count                         203.000000                          203.000000
   mean                           29.091941                           28.980138
   std                             2.371093                            2.372316
   min                            24.227000                           24.092000
   25%                            27.175500                           27.080000
   50%                            28.725000                           28.635000
   75%                            30.609000                           30.481500
   max                            37.566000                           37.454000

          LOSS-ON-IGNITION_Crucible_550C_(g)  \
   count                          203.000000
   mean                            28.322576
   std                              2.498194
   min                             22.467000
   25%                             26.532000
   50%                             28.026000
   75%                             30.099500
   max                             36.800000

          LOSS-ON-IGNITION_Crucible_1000C_(g)  LOSS-ON-IGNITION_Wet_Sample_(g)  \
   count                           203.000000                       201.000000
   mean                             28.244365                        10.579234
   std                               2.503733                         0.663191
   min                              22.346000                         5.739000
   25%                              26.468500                        10.284000
   50%                              27.960000                        10.517000
   75%                              30.021000                        10.788000
```

Correlation Matrix

Correlation matrix of every category in our data.

- Most of the categories have a weak correlation, but there are a good number of categories that correlated with each other extremely well
- Some of this is expected, as some variables are just mathematic conversions of the other variables



Full Correlation Matrix (Non-empty BULK_DENSITY_Dry_Soil_(<2mm)_(g))

	ENSITY_Dry_Soil_(<2mm)_(g)	LK_DENSITY_Gravel_Mass_(g)	ENSITY_Rock_Volume_(cm^3)	DENSITY_Soil_Volume_(cm^3)	NSITY_Bulk_Density_(g/cm^3)	I-IGNITION_Crucible_Mass_(g)	N-IGNITION_Crucible_Wet_(g)	I-IGNITION_Crucible_105C_(g)	I-IGNITION_Crucible_550C_(g)	IGNITION_Crucible_1000C_(g)	ON-IGNITION_Wet_Sample_(g)	ON-IGNITION_Dry_Sample_(g)	GNITION_Mineral_Sample_(g)	nate-Free_Mineral_Sample_(g)	oon_(kg)_in_10_cm_per_sq_m
BULK_DENSITY_Dry_Soil_(<2mm)_(g)	1	0.43	0.43	-0.43	1	-0.08	-0.054	-0.04	0.13	0.14	0.088	0.14	0.52	0.53	-0.15
BULK_DENSITY_Gravel_Mass_(g)	0.43	1	1	-1	0.47	0.035	0.058	0.066	0.15	0.15	0.086	0.11	0.3	0.3	-0.19
BULK_DENSITY_Rock_Volume_(cm^3)	0.43	1	1	-1	0.47	0.035	0.058	0.066	0.15	0.15	0.086	0.11	0.3	0.3	-0.19
BULK_DENSITY_Soil_Volume_(cm^3)	-0.43	-1	-1	1	-0.47	-0.035	-0.058	-0.066	-0.15	-0.15	-0.086	-0.11	-0.3	-0.3	0.19
BULK_DENSITY_Bulk_Density_(g/cm^3)	1	0.47	0.47	-0.47	1	-0.077	-0.051	-0.037	0.13	0.14	0.088	0.14	0.52	0.53	-0.15
LOSS-ON-IGNITION_Crucible_Mass_(g)	-0.08	0.035	0.035	-0.035	-0.077	1	0.96	0.96	0.92	0.92	-0.1	-0.1	-0.022	-0.021	-0.095
LOSS-ON-IGNITION_Crucible_Wet_(g)	-0.054	0.058	0.058	-0.058	-0.051	0.96	1	1	0.96	0.95	0.18	0.18	0.17	0.17	-0.1
LOSS-ON-IGNITION_Crucible_105C_(g)	-0.04	0.066	0.066	-0.066	-0.037	0.96	1	1	0.96	0.96	0.18	0.18	0.18	0.18	-0.12
LOSS-ON-IGNITION_Crucible_550C_(g)	0.13	0.15	0.15	-0.15	0.13	0.92	0.96	0.96	1	1	0.17	0.2	0.38	0.38	-0.33
LOSS-ON-IGNITION_Crucible_1000C_(g)	0.14	0.15	0.15	-0.15	0.14	0.92	0.95	0.96	1	1	0.17	0.2	0.38	0.38	-0.33
LOSS-ON-IGNITION_Wet_Sample_(g)	0.088	0.086	0.086	-0.086	0.088	-0.1	0.18	0.18	0.17	0.17	1	0.99	0.67	0.66	-0.033
LOSS-ON-IGNITION_Dry_Sample_(g)	0.14	0.11	0.11	-0.11	0.14	-0.1	0.18	0.18	0.2	0.2	0.99	1	0.73	0.72	-0.099
LOSS-ON-IGNITION_Mineral_Sample_(g)	0.52	0.3	0.3	-0.3	0.52	-0.022	0.17	0.18	0.38	0.38	0.67	0.73	1	1	-0.6
LOSS-ON-IGNITION_Carbonate-Free_Mineral_Sample_(g)	0.53	0.3	0.3	-0.3	0.53	-0.021	0.17	0.18	0.38	0.38	0.66	0.72	1	1	-0.6
LOSS-ON-IGNITION_Estimated_Carbon_(kg)_in_10_cm_per_sq_m	-0.15	-0.19	-0.19	0.19	-0.15	-0.095	-0.1	-0.12	-0.33	-0.33	-0.033	-0.099	-0.6	-0.6	1

Our Choices of Predicted Models

Although the correlation matrix gives us an idea on which categories are significant, we will still need to run ANOVA analysis with the highly correlated variables to see which specific categories have a true significance on making the soil suitable for agriculture.

Difficulties

- Knowing which variables to look at
  - Soil has a lot of complex interdependencies and majority of the group is not skilled in this area
- Cleaning the data
  - Had to figure out why some entries did not have a value
  - Cells had missing functions applied

Remaining Work Schedule

10/18-11/13:

- Make a draft of our model, and start typing up the research paper
- Work on the final draft of our report and demonstrate our modeling skills and efforts

11/13-12/10:

- Continue to work on research paper and final presentation slides.
- Complete them and make them look presentable and clean up the whole project.

Questions?