



**Artificial Intelligence and Data Science Department  
Houari Boumediene University of Science and Technology**

Project Report on :  
**(Data Mining : Data Analysis and Preprocessing)**

**Submitted by**  
Benzemrane Lydia / Ghamraci Rahil  
Session: 2024-2025

# Contents

<b>General Introduction</b>	<b>1</b>
<b>Chapter 1:     Data manipulation.</b>	<b>2</b>
<b>Chapter 2:     Analysis of the characteristics of the dataset attributes.</b>	<b>9</b>
<b>Chapter 3:     Data integration , Data reduction, Normalization and Discretization .</b>	<b>16</b>
<b>General Conclusion</b>	<b>23</b>

# List of figures

1.1	summary of the country dataset	2
1.2	Algeria polygones	3
1.3	Soil dataset global description	5
1.4	Soil dataset columns description part 1	5
1.5	Soil dataset columns description part 2	5
1.6	Algeria soil-polygons map	6
1.7	Climate dataset global description	7
1.8	Climate dataset columns description	7
1.9	Climate and soil data map	8
2.1	Soil attributes summary part 1	9
2.2	Soil attributes summary part 1	9
2.3	Boxplots with outliers	10
2.4	Histograms for soil attributes	11
2.5	Soil attributes scatter plots	12
2.6	Climate attributes summary	13
2.7	Climate symmetric box plots	13
2.8	Climate skewed and outliers box plots	14
2.9	skewed distribution histograms for climate	14
2.10	normal distribution histograms for climate	15
2.11	Climate attributes scatter plots	15
3.1	Outlier handling for soil attributes	16
3.2	Outlier handling for climate attributes	16
3.3	Final merged dataset	17
3.4	Final merged dataset part 2	18
3.5	Correlation heatmap	18
3.6	Seasonal Aggregation result	19
3.7	Min-Max normalization formula	20
3.8	Min-Max normalization	20
3.9	S-score formula	21
3.10	Z-score normalization	21

## **General Introduction**

The interplay between soil and climate attributes is fundamental to understanding environmental processes, ecosystem health, and sustainable land management. These interactions influence agricultural productivity, water cycles, and carbon storage, making them critical to addressing global challenges like climate change and food security. This project focuses on creating a comprehensive dataset to analyze soil-climate relationships, paving the way for data-driven environmental research and decision-making.

To achieve this, diverse datasets were merged and harmonized to ensure consistency and reliability. The project emphasizes the importance of maintaining spatial and temporal alignment while addressing challenges such as outliers, missing values, and data variability. By creating a well-structured dataset, we aim to facilitate analyses that reveal how climatic factors like temperature, precipitation, and surface pressure shape soil properties over time and space.

This report provides an overview of the project's goals and highlights the critical preprocessing steps undertaken, such as integration, normalization, and data reduction. These steps were designed to preserve the precision and integrity of the data, ensuring its readiness for meaningful analysis and interpretation.

# Chapter 1

## Data manipulation.

The first step in this project involves loading and exploring a geospatial dataset to gain an understanding of the attributes related to country boundaries worldwide. Using Python libraries such as geopandas, we import a Shapefile that contains detailed polygons and metadata representing various countries. This initial analysis focuses on inspecting the structure and content of the dataset, which is essential for guiding further data processing and analysis steps.

### 1.1 Importing, visualizing, and saving the contents of our datasets

#### 1.1.1 Visualizing the global country data

In the first part, we load and inspect a geospatial dataset containing global country boundaries. Using the geopandas library, we read a Shapefile that contains polygons for each country, along with metadata.

```
Country Data:
<class 'geopandas.geodataframe.GeoDataFrame'>
RangeIndex: 27272 entries, 0 to 27271
Data columns (total 11 columns):
#   Column      Non-Null Count  Dtype
---  -
0   AREA        27272 non-null  float64
1   PERIMETER   27272 non-null  float64
2   CNT1M_1     27272 non-null  int64
3   CNT1M_1_ID  27272 non-null  int64
4   FAO_NAME    27272 non-null  object
5   FAO_CODE    27272 non-null  int32
6   UN_CODE     27272 non-null  int32
7   ISO_CODE    27270 non-null  object
8   CNTRY_NAME  27272 non-null  object
9   ISO3_CODE   27272 non-null  object
10  geometry    27272 non-null  geometry
dtypes: float64(2), geometry(1), int32(2), int64(2), object(4)
memory usage: 2.1+ MB
None
```

	AREA	PERIMETER	CNT1M_1	CNT1M_1_ID	FAO_NAME	FAO_CODE	UN_CODE	\
0	649.422101	900.013634	2	1	Greenland	85	304	
1	0.002037	0.262890	3	2	Greenland	85	304	
2	0.052469	1.545726	4	3	Greenland	85	304	
3	0.142221	4.510463	5	4	Greenland	85	304	
4	0.000403	0.147693	6	5	Greenland	85	304	

	ISO_CODE	CNTRY_NAME	ISO3_CODE	\
0	GL	Greenland	GRL	
1	GL	Greenland	GRL	
2	GL	Greenland	GRL	
3	GL	Greenland	GRL	
4	GL	Greenland	GRL	

	geometry
0	POLYGON ((-40.12595 82.95437, -40.06486 82.960...
1	POLYGON ((-28.66035 83.43806, -28.68495 83.440...
2	POLYGON ((-39.14128 83.28785, -39.2557 83.2895...
3	POLYGON ((-38.84585 83.11409, -38.90443 83.115...
4	POLYGON ((-41.16922 83.33663, -41.22647 83.342...

Figure 1.1: summary of the country dataset

### 1.1.2 Output Analysis:

- We observe that there are 27,272 records and 11 attributes. The geometry column holds polygonal shapes, essential for visualizing or filtering the dataset geographically.
- 1. The dataset includes columns such as AREA and PERIMETER, which are numeric descriptors of each country's shape, and CNTRY-NAME, which provides the country name.
  2. The geometry column contains the polygon shapes, which can be visualized later using mapping tools.

### 1.1.3 Isolating Algerian Data

In the next step of our analysis, we filter the dataset to focus exclusively on Algeria. Given the global scope of the original Shapefile, narrowing down the data to a specific country is essential for our targeted exploration and preprocessing.

To achieve this, we apply a conditional filter on the FAO-NAME column, selecting only rows labeled "Algeria" in order to create a new GeoDataFrame, Algeria-GDF. This GeoDataFrame contains all relevant records and attributes for Algeria, including:

- AREA: Represents the area of each polygon in the dataset.
- PERIMETER: The perimeter length for each polygon.
- ISO-CODE and ISO3-CODE: International standard codes identifying Algeria.
- Geometry: The geometric data for Algeria, defined as polygons, each describing a region within the country.

	AREA	PERIMETER	CNT1M_1_	CNT1M_1_ID	FAO_NAME	FAO_CODE	UN_CODE	ISO_CODE	CNTRY_NAME	ISO3_CODE	geom
12822	213.434727	73.435570	12824	11914	Algeria	4	12	DZ	Algeria	DZA	POLY ((7.5 37.0 7.5 37.0
13013	0.000072	0.033402	13015	11939	Algeria	4	12	DZ	Algeria	DZA	POLY ((-0.8 35.7 -0.8 35.77
13017	0.000111	0.044937	13019	11942	Algeria	4	12	DZ	Algeria	DZA	POLY ((-1.1 35.7 -1.1 35.72

Figure 1.2: Algeria polygons

The resulting output shows several entries, indicating multiple polygons that represent different parts of Algeria's territory.

#### **1.1.4 Soil Data Import and Initial Overview**

##### **Geometric Validation and Cleaning of Soil Data**

To ensure the accuracy and reliability of spatial analyses in this study, it was essential to validate and clean the geometric data in the soil dataset. The geometry field in the dataset is provided in Well-Known Text (WKT) format, which stores spatial information about the soil properties across various regions in Algeria. However, spatial data can often contain invalid or improperly formatted entries, which can disrupt geospatial operations if left unaddressed.

To validate and clean the WKT geometries, the following approach was taken:

1. Each WKT entry was first checked to confirm it was correctly formatted as a POLYGON using the `is-valid-wkt` function. This function ensured that only entries with valid polygon formatting were processed further
2. For each valid WKT string, the `clean-wkt-load` function attempted to load the geometry. If a geometry was found to be invalid after loading, the function attempted to correct minor topological issues using the `buffer(0)` technique. This operation creates a small buffer around the geometry, often resolving self-intersections and other minor errors. Geometries that could not be fixed were excluded from the final dataset to maintain data quality.
3. After cleaning, all rows with `None` geometries were removed, ensuring that only valid spatial data was retained for analysis.
4. Finally, the cleaned data was converted into a `GeoDataFrame`, enabling efficient geospatial analysis using Python's `geopandas` library.

##### **Soil Dataset Summary and Description**

next, To gain an overview of the cleaned soil dataset, a descriptive analysis was performed. This analysis summarized the structure and column data types of the dataset.

Rows Number	291
Columns Number	26
Usage of memory	59.109375 ko
Data types	[object, float64, geometry]

Figure 1.3: Soil dataset global description

	Name	Values not null	Type
0	CNT_FULLNAME	291	object
1	sand % topsoil	291	float64
2	sand % subsoil	291	float64
3	silt % topsoil	291	float64
4	silt% subsoil	291	float64
5	clay % topsoil	291	float64
6	clay % subsoil	291	float64
7	pH water topsoil	291	float64
8	pH water subsoil	291	float64
9	OC % topsoil	291	float64
10	OC % subsoil	291	float64
11	N % topsoil	291	float64
12	N % subsoil	291	float64
13	BS % topsoil	291	float64
14	BS % subsoil	291	float64
15	CEC topsoil	291	float64
16	CEC subsoil	291	float64
17	CEC clay topsoil	291	float64

Figure 1.4: Soil dataset columns description part 1

18	CEC Clay subsoil	291	float64
19	CaCO3 % topsoil	291	float64
20	CaCO3 % subsoil	291	float64
21	BD topsoil	291	float64
22	BD subsoil	291	float64
23	C/N topsoil	291	float64
24	C/N subsoil	291	float64
25	geometry	291	geometry

Figure 1.5: Soil dataset columns description part 2



The dataset, imported as `soil-df`, includes 295 entries and 26 columns, capturing various soil characteristics such as sand, silt, and clay percentages, pH levels, organic carbon (OC), base saturation (BS), cation exchange capacity (CEC), calcium carbonate ( $\text{CaCO}_3$ ), and bulk density (BD) in both topsoil and subsoil layers. Each entry also includes a geometry column, providing polygon coordinates for spatial analysis.

- **Data Integrity:** All 295 entries are non-null across the 26 columns, indicating that there are no missing values.
- **Column Types:** The dataset consists mainly of float64 types for soil properties and an object type for the geometry column, which will support further geospatial analyses

### 1.1.5 Visualization of Soil Data and Algeria Polygons Boundary

To understand the spatial distribution of soil data within Algeria, a boundary plot was generated to visually compare the soil data polygons with the geographical boundary of Algeria. This visualization helps confirm that the soil data polygons are properly situated

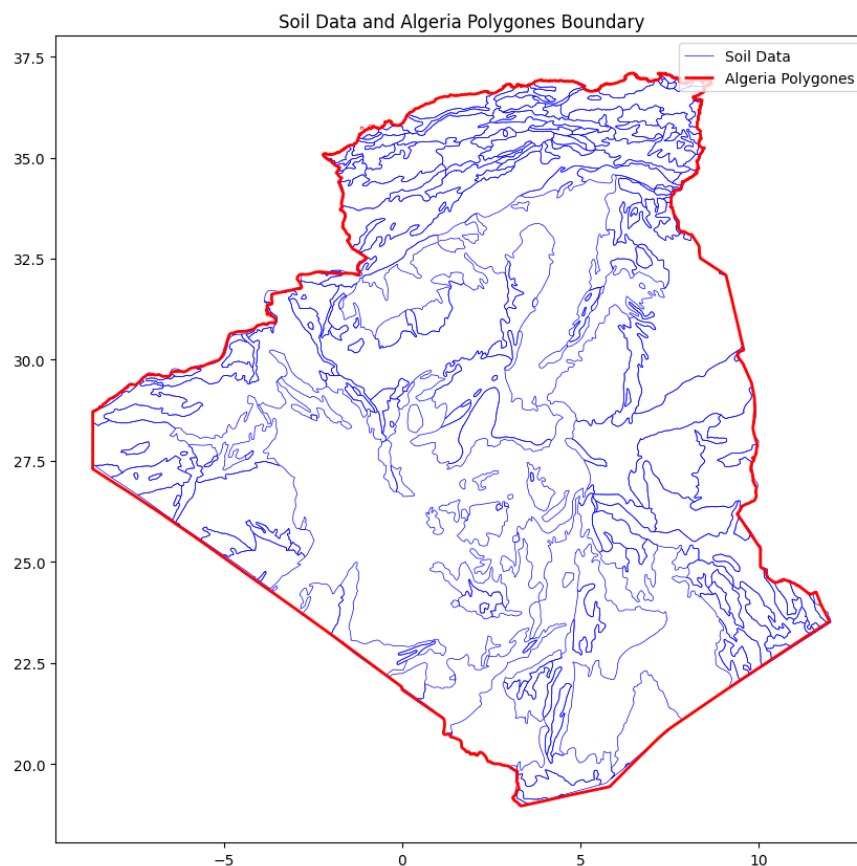


Figure 1.6: Algeria soil-polygons map

within Algeria's borders, ensuring that the dataset's spatial component aligns with the study area

### 1.1.6 Exploration and Analysis of Climate Dataset

To further understand environmental factors within Algeria, we explored a climate dataset with multiple parameters, including pressure, air quality, rainfall, snow, temperature, and wind speed. This dataset was processed and analyzed using various Python libraries, including xarray, rioxarray, and geopandas.[1]

Rows Number	12929760
Columns Number	11
Usage of memory	757603.125 ko
Data types	[datetime64[ns], float64, int32, float32, geom...

Figure 1.7: Climate dataset global description

	Name	Values not null	Type
0	time	12929760	datetime64[ns]
1	lon	12929760	float64
2	lat	12929760	float64
3	spatial_ref	12929760	int32
4	PSurf	7498560	float32
5	Qair	7498560	float32
6	Rainf	7498560	float32
7	Snowf	7498560	float32
8	Tair	7498560	float32
9	Wind	7498560	float32
10	geometry	12929760	geometry

Figure 1.8: Climate dataset columns description

This summary provided insights into the structure of the climate dataset and its memory footprint. Key climate attributes (PSurf, Qair, Rainf, Snowf, Tair, Wind) had a substantial number of non-null values, indicating good coverage within Algeria's boundaries. However, certain columns had missing values, which will be handled later on.

### 1.1.7 Visualization of Climate Data and Soil Data

In this part of the analysis, we visualized a subset of the climate dataset to better understand the spatial distribution of key climate variables (such as temperature, air quality, rainfall, and wind) across Algeria. A random sample of 100,000 valid climate data points was selected, with missing values removed for crucial variables (Tair, Qair, Rainf, Wind, PSurf, Snowf). These selected points were plotted as red dots on a map of Algeria, superimposed over the boundaries of soil data (shown in blue).

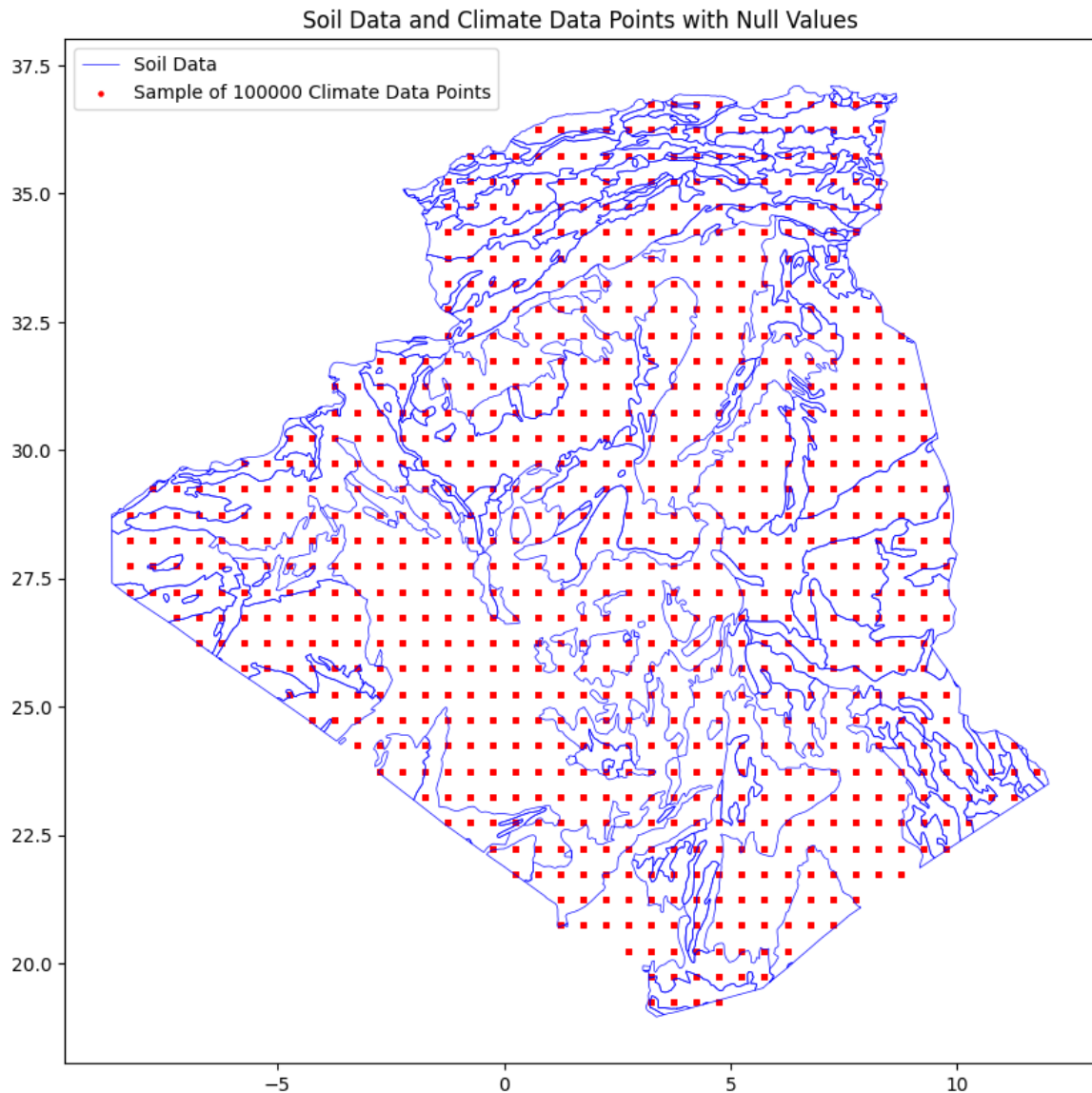


Figure 1.9: Climate and soil data map

## Chapter 2

### Analysis of the characteristics of the dataset attributes.

#### 2.1 Soil Attributes Summary Report

This section summarizes the key soil attributes, focusing on statistical measures, data distribution, and visualizations. We will look at central tendency measures, outliers, and the distribution of data through box plots, bar charts, and histograms.

##### 2.1.1 Soil Attributes measures of central tendency

We calculate the mean, median, and mode to understand the typical values of soil attributes, such as sand, silt, clay, and organic content. These measures help reveal if the data is skewed or clustered around certain values.

	Attribute	Mean	Median	Mode	Q1	Q3	IQR	Has Outliers	Missing Values	Unique Values
0	sand % topsoil	32.935423	36.0900	0.0	0.0	50.390	50.390	No	0	97
1	sand % subsoil	32.054615	36.4900	0.0	0.0	46.550	46.550	No	0	96
2	silt % topsoil	18.081759	18.5460	0.0	0.0	26.220	26.220	No	0	98
3	silt% subsoil	17.972852	19.1000	0.0	0.0	27.900	27.900	No	0	98
4	clay % topsoil	18.263509	22.7600	0.0	0.0	25.200	25.200	No	0	95
5	clay % subsoil	19.149845	22.3200	0.0	0.0	27.240	27.240	No	0	94
6	pH water topsoil	5.253718	7.3000	0.0	0.0	7.705	7.705	No	0	79
7	pH water subsoil	5.340416	7.4600	0.0	0.0	7.810	7.810	No	0	84
8	OC % topsoil	0.439228	0.4475	0.0	0.0	0.678	0.678	Yes	0	96
9	OC % subsoil	0.253943	0.2870	0.0	0.0	0.400	0.400	No	0	87
10	N % topsoil	0.068581	0.0630	0.0	0.0	0.104	0.104	Yes	0	77
11	N % subsoil	0.036186	0.0340	0.0	0.0	0.058	0.058	Yes	0	53
12	BS % topsoil	62.683677	85.6000	0.0	0.0	95.800	95.800	No	0	77
13	BS % subsoil	64.830859	91.5000	0.0	0.0	96.000	96.000	No	0	70
14	CEC topsoil	10.093003	9.7500	0.0	0.0	14.440	14.440	Yes	0	96
15	CEC subsoil	9.465024	8.6600	0.0	0.0	14.125	14.125	Yes	0	93

Figure 2.1: Soil attributes summary part 1

16	CEC clay topsoil	38.541375	49.4000	0.0	0.0	58.050	58.050	No	0	88
17	CEC Clay subsoil	32.435773	42.3000	0.0	0.0	49.300	49.300	No	0	79
18	CaCO3 % topsoil	6.176900	3.5800	0.0	0.0	10.980	10.980	Yes	0	91
19	CaCO3 % subsoil	8.849361	5.1150	0.0	0.0	16.185	16.185	Yes	0	95
20	BD topsoil	0.989313	1.3400	0.0	0.0	1.470	1.470	No	0	52
21	BD subsoil	0.977835	1.3400	0.0	0.0	1.430	1.430	No	0	47
22	C/N topsoil	7.331100	9.3500	0.0	0.0	10.450	10.450	No	0	56
23	C/N subsoil	6.899381	8.5000	0.0	0.0	10.600	10.600	No	0	65

Figure 2.2: Soil attributes summary part 1

The data suggests no missing values for most attributes, and only a few show potential outliers, indicating relatively clean data for the majority of the attributes.

### 2.1.2 Soil attributes box plots and outliers.

Box plots provide a visual representation of soil attribute distribution, highlighting the median, IQR, and outliers. Outliers indicate extreme values that could signify unusual soil conditions worth further investigation.

#### Box plots with visible outliers

in this section, we will be visualizing the attributes with outliers. For the other box plots,

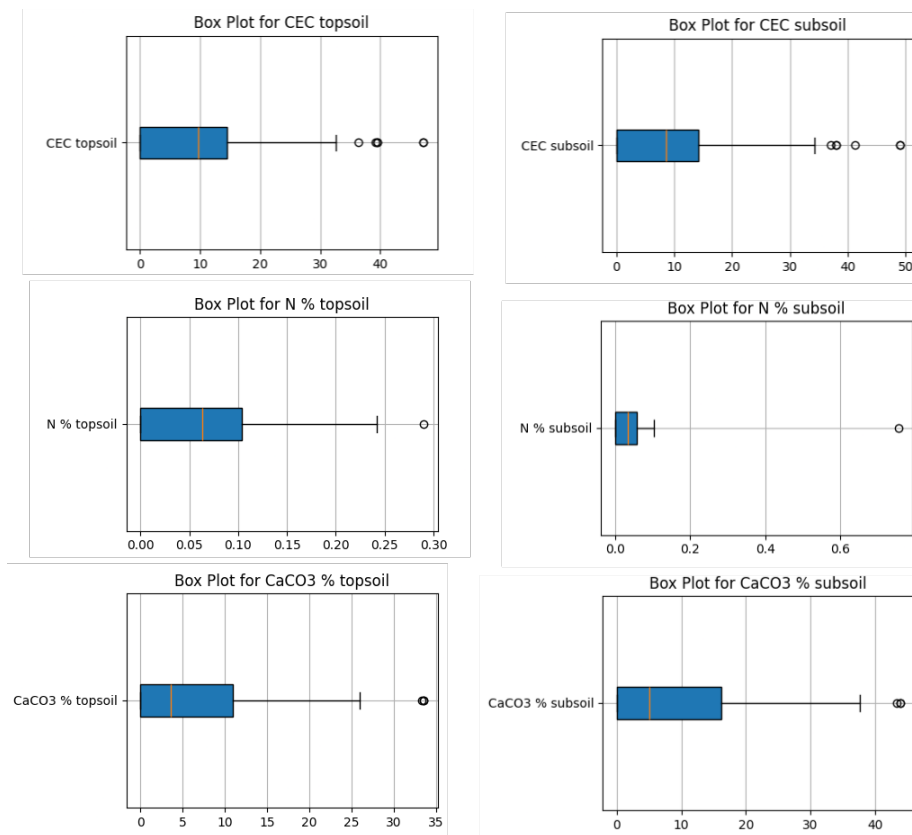


Figure 2.3: Boxplots with outliers

aside from those displaying visible outliers, the data does not appear to follow a normal distribution or exhibit symmetry.

### 2.1.3 Soil attributes histograms and data distribution.

Histograms show the distribution of continuous soil attributes like pH and organic content. They help assess the spread of data and identify patterns, such as skewness or gaps in the data.

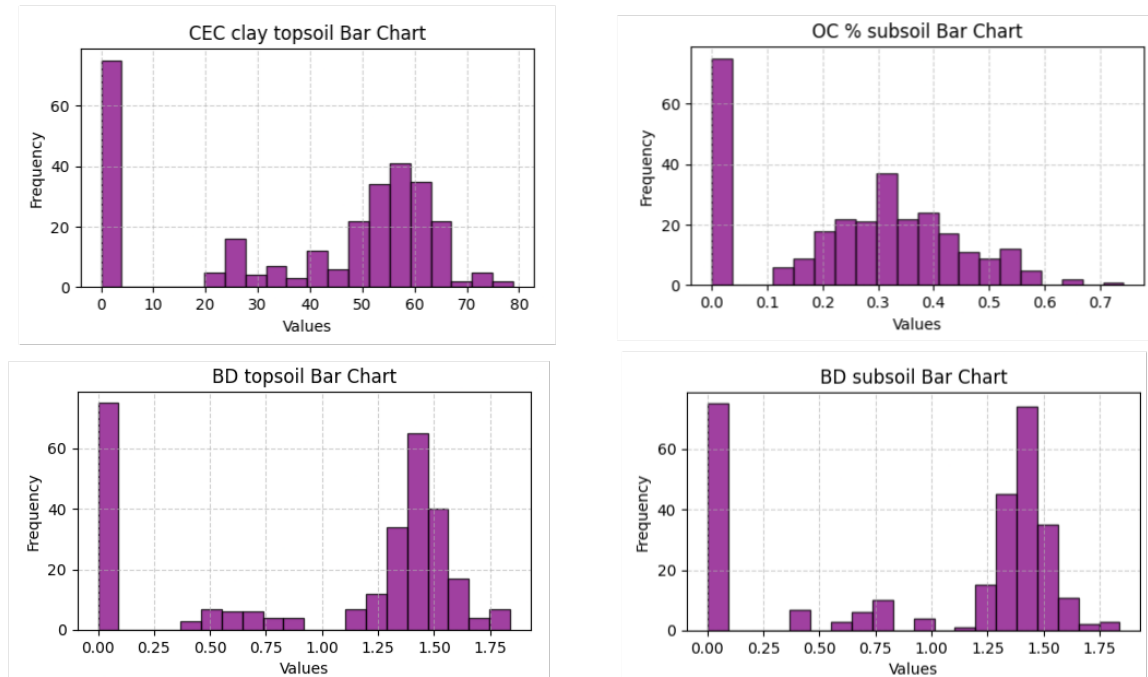


Figure 2.4: Histograms for soil attributes

- **Symmetry:** The distribution appears nearly symmetric, suggesting that most of the data points are evenly distributed around the central value. This implies a balanced spread of the attribute values on either side of the mean or median.
- **Outliers:** However, the presence of outliers is noticeable as isolated bars far from the main cluster of the distribution. These outliers, although sparse, could indicate rare or extreme values that may influence statistical measures like the mean and standard deviation.

#### 2.1.4 Soil attributes Scatter Plots Analysis

Scatter plots are a useful graphical tool for visualizing the relationship between two continuous variables. Each point in the plot represents a pair of values for the variables being analyzed, allowing us to quickly assess patterns, trends, and correlations in the data.

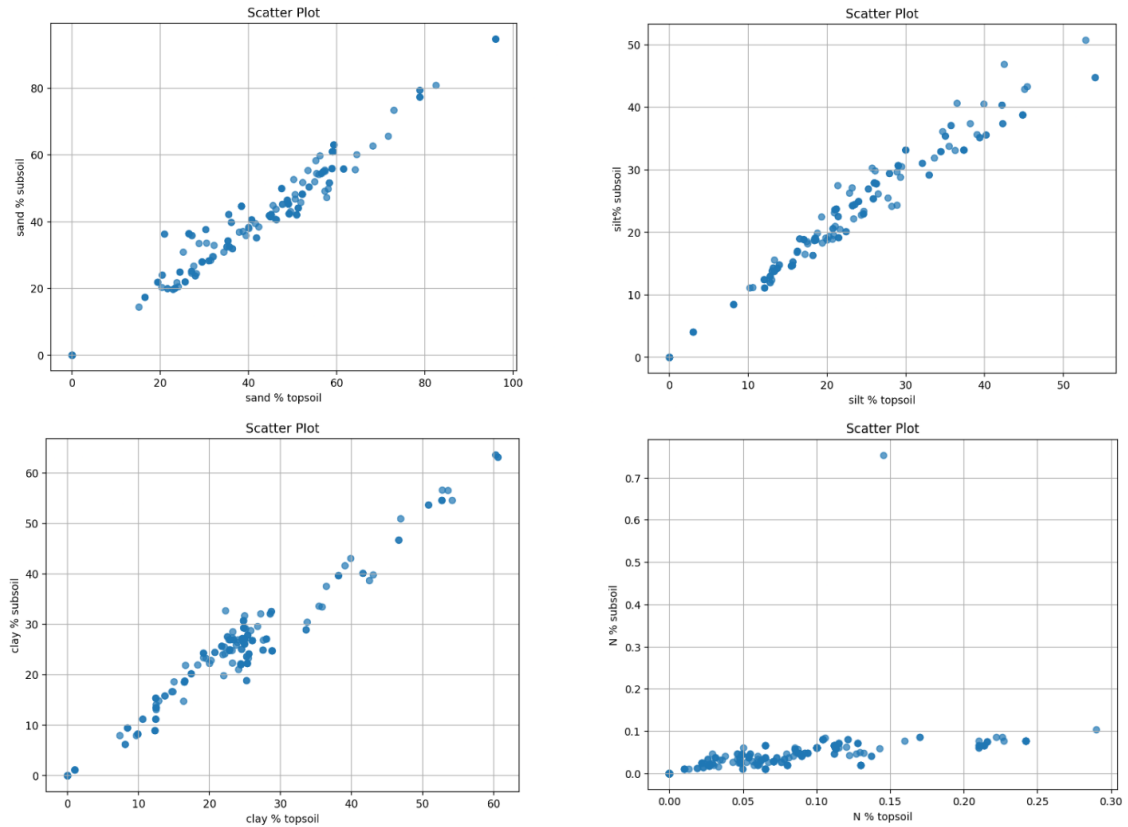


Figure 2.5: Soil attributes scatter plots

As observed in the scatter plots for the variables Clay %, Sand %, and Silt % for both the topsoil and subsoil, there is a clear upward trend, unlike the scatter plot for N %, indicating a positive linear correlation between each variable pair. This suggests that as one variable increases, the others tend to increase as well. Given this consistent relationship, we can conclude that these variables may be interdependent, and it might be possible to use one as a substitute for the others in certain analyses or models.

## 2.2 Climate Attributes Summary Report

This section focuses on climate data attributes, such as temperature and rainfall. Similar to soil attributes, we analyze these using central tendency measures, outliers, and visualizations.

### 2.2.1 Climate Attributes measures of central tendency

We calculate the mean, median, and mode to understand typical climate conditions, such as temperature and rainfall patterns. These measures help identify seasonal trends that may impact soil and vegetation.

	Attribute	Mean	Median	Mode	Q1	Q3	IQR	Has Outliers	Missing Values	Unique Values
0	lon	1.750000e+00	1.750000	-8.25, -7.75, -7.25, -6.75, -6.25, -5.75, -5.2...	-3.250000	6.750000	10.000000	No	0	41
1	lat	2.800000e+01	28.000000	19.25, 19.75, 20.25, 20.75, 21.25, 21.75, 22.2...	23.625000	32.375000	8.750000	No	0	36
2	PSurf	9.487716e+04	95569.117188	97212.2578125	93079.195312	97283.812500	4204.617188	Yes	5431200	1770194
3	Qair	4.690343e-03	0.004132	0.003966310992836952, 0.004317568149417639	0.002996	0.005816	0.002820	Yes	5431200	6075571
4	Rainf	2.350112e-06	0.000000	0.0	0.000000	0.000000	0.000000	Yes	5431200	144365
5	Snowf	3.864786e-08	0.000000	0.0	0.000000	0.000000	0.000000	Yes	5431200	3812
6	Tair	2.964559e+02	296.686310	300.2471923828125	288.702972	303.969948	15.266975	Yes	5431200	1418678
7	Wind	4.043839e+00	3.779118	4.110576152801514	2.645402	5.162067	2.516665	Yes	5431200	6312312

Figure 2.6: Climate attributes summary

- Many attributes, including lon, lat, Tair, and Wind, have reliable data with no missing values and exhibit relatively normal distributions or low skew.
- However, attributes like PSurf, Qair, Rainf, and Snowf present challenges due to significant missing values and skewed distributions.
- The presence of outliers in several variables (such as PSurf, Qair, Rainf, and Snowf) suggests the need for careful data cleaning or outlier handling, especially in cases where the outliers may be errors or anomalies.

### 2.2.2 Climate attributes box plots and outliers.

Box plots for climate attributes show the distribution of data and help identify outliers, such as extreme temperatures or unexpected rainfall, that may affect soil and agricultural conditions.

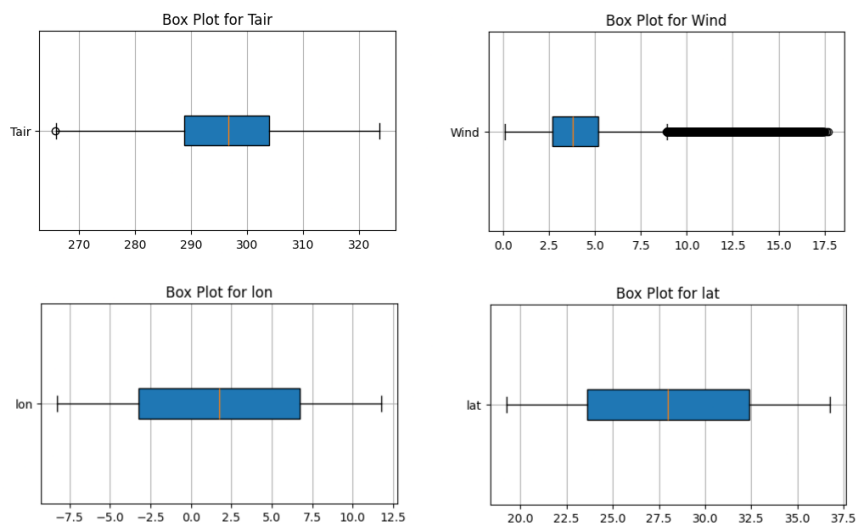


Figure 2.7: Climate symmetric box plots



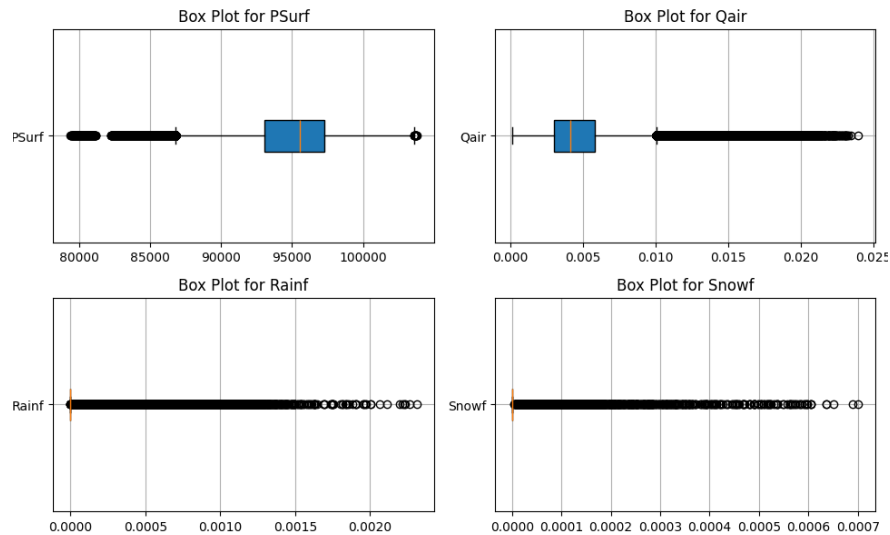


Figure 2.8: Climate skewed and outliers box plots

From the box plot, we can confirm that our initial observations from the data summary were accurate, particularly regarding the distribution of the data and the identification of attributes with outliers.

### 2.2.3 Climate attributes histograms and data distribution.

Histograms for climate data show the distribution of variables like temperature and precipitation, helping to identify trends or anomalies in the dataset.

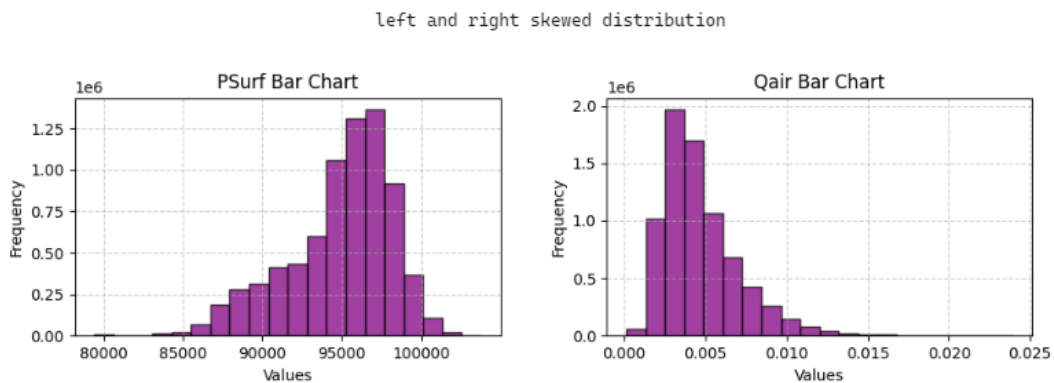


Figure 2.9: skewed distribution histograms for climate

Similarly, the histograms validate our initial observations from the data summary, confirming the distribution patterns of the data and the presence of outliers in certain attributes.

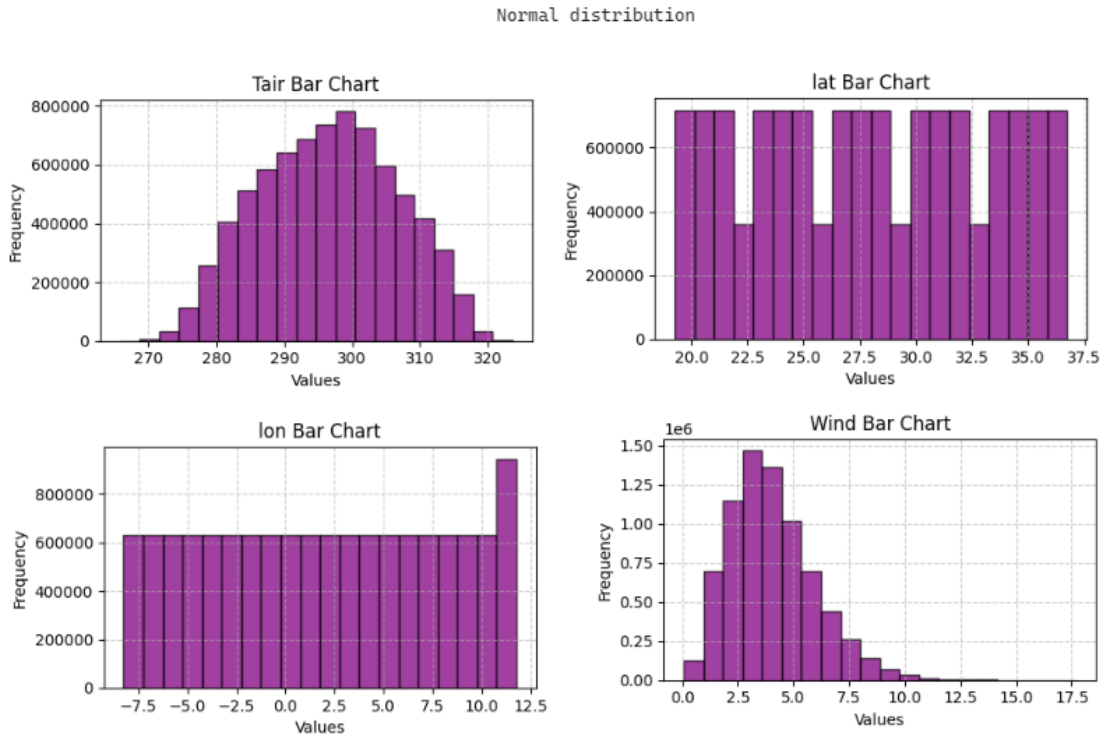


Figure 2.10: normal distribution histograms for climate

#### 2.2.4 Climate attributes Scatter Plots Analysis

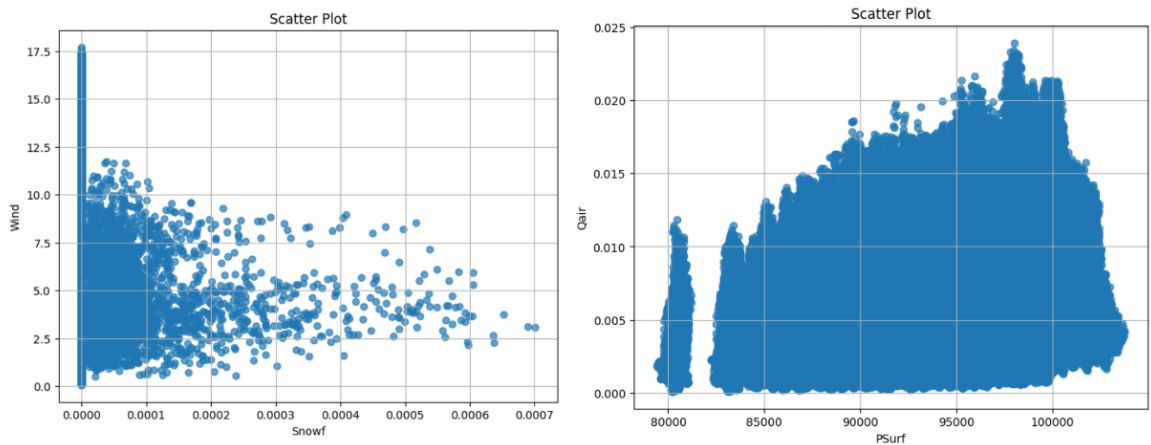


Figure 2.11: Climate attributes scatter plots

- **Left Plot (Wind vs. Snowf):** Low snowfall dominates, with wind speed varying widely at low snowfall and stabilizing as snowfall increases.
- **Right Plot (Qair vs. PSurf):** A positive trend exists, with distinct clusters (pressure bands), suggesting granularity in surface pressure data and a correlation between higher surface pressure and increased Q

## Chapter 3

### Data integration , Data reduction, Normalization and Discretization .

#### 3.1 Outlier Handling

Outlier handling is performed separately for soil and climate datasets to address their unique characteristics.

- First, we detect outliers using the IQR method.
- Then, we will replace outliers using configurable strategies (mean, median, mode, quantile capping, or random values).

For the soil attributes, all replacement strategies gave the same result seen below. For

```
OC % topsoil: 1 outliers detected.  
OC % topsoil: Outliers after handling: 0  
N % topsoil: 1 outliers detected.  
N % topsoil: Outliers after handling: 0  
N % subsoil: 1 outliers detected.  
N % subsoil: Outliers after handling: 0  
CEC topsoil: 6 outliers detected.  
CEC topsoil: Outliers after handling: 0  
CEC subsoil: 6 outliers detected.  
CEC subsoil: Outliers after handling: 0  
CaCO3 % topsoil: 3 outliers detected.  
CaCO3 % topsoil: Outliers after handling: 0  
CaCO3 % subsoil: 3 outliers detected.  
CaCO3 % subsoil: Outliers after handling: 0
```

Figure 3.1: Outlier handling for soil attributes

the climate attributes, replacing the outliers with the median value has shown the best results, as seen below.

```
PSurf: 125547 outliers detected.  
PSurf: Outliers after handling: 0  
Qair: 264733 outliers detected.  
Qair: Outliers after handling: 0  
Rainf: 145080 outliers detected.  
Rainf: Outliers after handling: 0  
Snowf: 3813 outliers detected.  
Snowf: Outliers after handling: 0  
Tair: 1 outliers detected.  
Tair: Outliers after handling: 0  
Wind: 130202 outliers detected.  
Wind: Outliers after handling: 0
```

Figure 3.2: Outlier handling for climate attributes

## 3.2 Discretization and Handling Missing Values

### 3.2.1 Discretization

- Discretization was not used in this project because the soil attributes are continuous variables, and preserving their precision is essential for accurate analysis and modeling. It is typically applied when simplifying data for interpretability, handling non-linear relationships, or aligning with domain-specific thresholds, none of which were necessary for this study's objectives.

### 3.2.2 Handling Missing Values

The handling-missing-values function offers multiple imputation strategies

- Ignoring NaN values.
- Replacing with statistical measures (mean, median).
- Using Bayesian methods to impute values based on distribution assumptions.

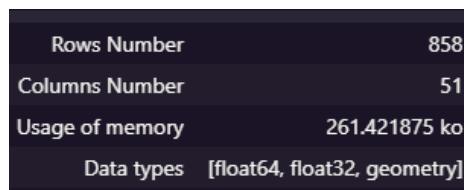
The different strategies were only applied to the climate attributes because the soil dataset has no missing values, for the rest of the project we went with the replacement with the median strategy.

## 3.3 Data Integration

Data integration in this project aligns soil and climate datasets geographically to create a unified dataset for analysis. This step ensures spatial consistency and facilitates meaningful correlations between soil attributes and climatic conditions.

for this step we used the `gpd.sjoin` from GeoPandas, it checks if soil points fall within climate regions and keeps only overlapping records, ensuring spatial alignment.

Seasonal aggregation of climate attributes was performed before data integration to simplify the process and ensure efficiency.



Rows Number	858
Columns Number	51
Usage of memory	261.421875 ko
Data types	[float64, float32, geometry]

Figure 3.3: Final merged dataset

	lon	lat	autumn_PSurf	spring_PSurf	summer_PSurf	winter_PSurf	autumn_Qair	spring_Qair	summer_Qair	winter_Qair	...	CEC topsoil	CEC subsoil	CEC clay topsoil	CEC Clay subsoil
16	-8.25	27.25	96395.804688	96220.125000	96089.164062	96951.468750	0.005147	0.004590	0.006067	0.004074	...	12.58	11.94	57.1	46.8
17	-8.25	27.75	96290.093750	96105.710938	95972.968750	96852.187500	0.005093	0.004534	0.006074	0.004016	...	9.98	10.31	57.8	44.8
18	-8.25	28.25	95736.570312	95542.250000	95417.031250	96283.453125	0.004946	0.004288	0.006044	0.003795	...	9.98	10.31	57.8	44.8
19	-8.25	28.75	96331.351562	96131.632812	95971.125000	96908.789062	0.004576	0.003775	0.005432	0.003525	...	10.87	10.06	56.7	45.8
52	-7.75	27.25	96943.312500	96762.343750	96602.812500	97524.000000	0.005038	0.004342	0.005875	0.004074	...	0.00	0.00	0.0	0.0

5 rows × 51 columns

Figure 3.4: Final merged dataset part 2

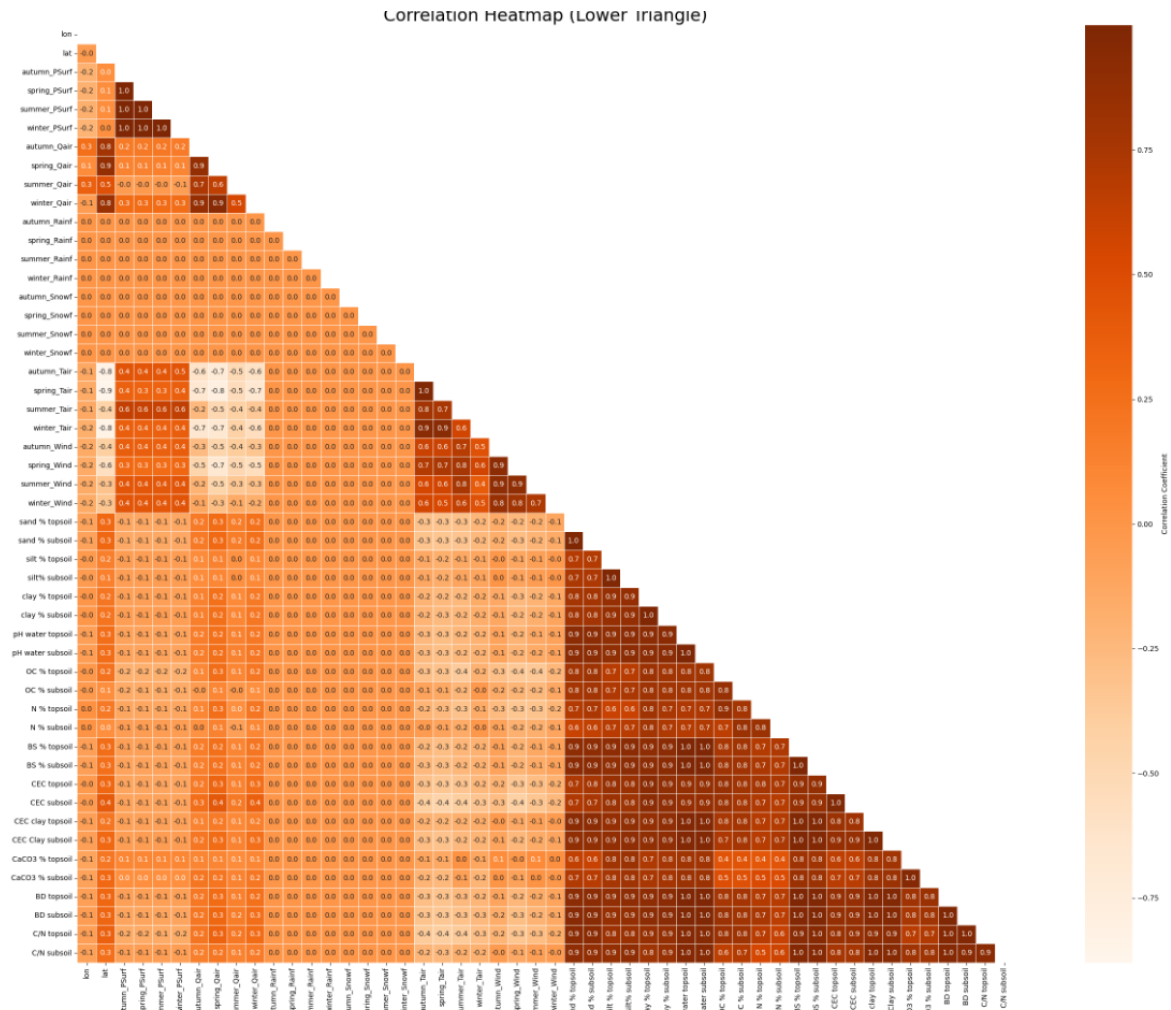


Figure 3.5: Correlation heatmap

Darker colors indicate a strong correlation, suggesting that changes in one attribute are closely linked to another, while lighter colors represent weak correlations, where variables are independent

### 3.4 Data Reduction

Data reduction techniques reduce the dataset's dimensionality and size while preserving key patterns.

	lon	lat	autumn_PSurf	spring_PSurf	summer_PSurf	winter_PSurf	autumn_Qair	spring_Qair	summer_Qair	winter_Qair	...	winter_Snowf	autumn_Tair	spring_Tair	summer_Tair	winter_Tair	autumn_W
0	-8.25	19.25	95569.117188	95569.117188	95569.117188	95569.117188	0.004132	0.004132	0.004132	0.004132	...	0.0	296.68631	296.68631	296.68631	296.68631	3.779
1	-8.25	19.75	95569.117188	95569.117188	95569.117188	95569.117188	0.004132	0.004132	0.004132	0.004132	...	0.0	296.68631	296.68631	296.68631	296.68631	3.779
2	-8.25	20.25	95569.117188	95569.117188	95569.117188	95569.117188	0.004132	0.004132	0.004132	0.004132	...	0.0	296.68631	296.68631	296.68631	296.68631	3.779
3	-8.25	20.75	95569.117188	95569.117188	95569.117188	95569.117188	0.004132	0.004132	0.004132	0.004132	...	0.0	296.68631	296.68631	296.68631	296.68631	3.779
4	-8.25	21.25	95569.117188	95569.117188	95569.117188	95569.117188	0.004132	0.004132	0.004132	0.004132	...	0.0	296.68631	296.68631	296.68631	296.68631	3.779

5 rows × 17 columns

Figure 3.6: Seasonal Aggregation result

### 3.4.1 Seasonal Aggregation:

- The data-reduction function transforms climate data into seasonal averages, reducing temporal granularity and focusing on meaningful patterns.
- Longitude (lon) and latitude (lat) values are rounded to reduce positional precision and redundancy.
- Outputs are grouped by season, longitude, and latitude, and stored in a GeoDataFrame for geospatial compatibility.

### 3.4.2 Vertical and Horizontal Reduction:

We will apply the vertical and horizontal reduction to the climate data

### 3.4.3 Horizontal Reduction

- **Method:** Eliminated rows with missing values to ensure complete datasets.
- **Result:** Since the dataset already had clean and complete entries, no rows were removed, leaving the total unchanged at 858.

### 3.4.4 Vertical Reduction

In the Data Reduction step, we applied vertical reduction using the "Low Variance Columns" method. However, the climate attributes were removed because they exhibited low variance across the dataset, which means they had little variation or were constant over the records. This is likely due to the aggregation of climate data into seasonal averages, which reduced the variability. Given that these variables do not significantly contribute to the analysis due to their lack of variance, we decided not to apply vertical reduction to them, as their removal would have removed important context for understanding soil-climate interactions.

### 3.5 Normalization

the dataset was normalized using two common techniques: Min-Max Normalization and Z-Score Normalization.

#### 3.5.1 Min-Max Normalization

Min-Max normalization was applied to the dataset using the following formula:

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Figure 3.7: Min-Max normalization formula

Where: X is the original value, Xmin and Xmax are the minimum and maximum values of the attribute, respectively.

This technique scales the data to a range between 0 and 1, which is particularly useful when attributes have varying ranges. Below is a preview of some attributes after Min-Max normalization:

	lon	lat	autumn_PSurf	spring_PSurf	summer_PSurf	winter_PSurf	\
16	0.000	0.457142	0.627260	0.617731	0.625790	0.645963	
17	0.000	0.485714	0.619851	0.609750	0.617565	0.639257	
18	0.000	0.514285	0.581057	0.570449	0.578209	0.600844	
19	0.000	0.542857	0.622743	0.611558	0.617434	0.643080	
52	0.025	0.457142	0.665632	0.655551	0.662152	0.684633	
53	0.025	0.485714	0.664879	0.654134	0.660555	0.684753	
54	0.025	0.514285	0.586702	0.575660	0.582273	0.607276	
55	0.025	0.542857	0.563571	0.552027	0.558476	0.584503	
56	0.025	0.571428	0.652205	0.640180	0.644218	0.673543	
87	0.050	0.428571	0.653650	0.643876	0.649534	0.671874	
88	0.050	0.457142	0.709593	0.698738	0.704139	0.728750	
89	0.050	0.485714	0.670258	0.659032	0.664909	0.690301	
90	0.050	0.514285	0.621593	0.609807	0.615662	0.642269	
91	0.050	0.542857	0.524395	0.512752	0.519371	0.545363	
92	0.050	0.571428	0.595018	0.582996	0.587824	0.616333	
122	0.075	0.400000	0.613182	0.603873	0.608566	0.630218	
123	0.075	0.428571	0.640643	0.630456	0.635566	0.658650	
124	0.075	0.457142	0.704335	0.692955	0.697938	0.723343	
125	0.075	0.485714	0.645217	0.633511	0.639385	0.665120	
126	0.075	0.514285	0.614334	0.602188	0.607882	0.634772	

Figure 3.8: Min-Max normalization

### 3.5.2 Z-Score Normalization

Z-Score normalization, also known as standardization, was applied using the following formula:

$$Z = \frac{X - \mu}{\sigma}$$

Figure 3.9: S-score formula

Where X is the original value, U is the mean of the attribute, O is the standard deviation of the attribute.

This method scales the data to have a mean of 0 and a standard deviation of 1, which is useful when the distribution of the data is not uniform or has outliers. Below is a preview of a few attributes after applying Z-Score normalization:

	lon	lat	autumn_PSurf	spring_PSurf	summer_PSurf	\
16	-2.516734	-0.221097	0.445383	0.430282	0.448337	
17	-2.516734	-0.104429	0.410214	0.391794	0.408487	
18	-2.516734	0.012238	0.226061	0.202253	0.217826	
19	-2.516734	0.128905	0.423940	0.400514	0.407855	
52	-2.401808	-0.221097	0.627535	0.612678	0.624494	
	winter_PSurf	autumn_Qair	spring_Qair	summer_Qair	winter_Qair	...
16	0.476755	-0.007348	0.496834	1.223306	0.733869	...
17	0.444913	-0.077901	0.440049	1.231053	0.668869	...
18	0.262509	-0.270089	0.193272	1.197538	0.421698	...
19	0.463066	-0.752399	-0.320935	0.511401	0.120001	...
52	0.660376	-0.149664	0.248152	1.008132	0.733487	...
	CEC topsoil	CEC subsoil	CEC clay topsoil	CEC Clay subsoil	\	
16	0.886948	0.987362	0.764477	0.874049		
17	0.422883	0.676656	0.792036	0.775990		
18	0.422883	0.676656	0.792036	0.775990		
19	0.581736	0.629002	0.748729	0.825020		
52	-1.358412	-1.288604	-1.483587	-1.420529		

Figure 3.10: Z-score normalization

## 3.6 Justification of the preprocessing steps

The preprocessing steps applied in this project were carefully chosen to ensure the quality, consistency, and suitability of the data for analysis. Each preprocessing method directly contributes to the effectiveness of the analysis [3] [2]



### 1. **Outlier Handling**

Outlier handling was critical for ensuring that extreme values did not distort the analysis and model performance. We used the Interquartile Range (IQR) method to detect outliers. By handling outliers appropriately, we ensured that the dataset remained free from undue influence by anomalous values, enabling more accurate and meaningful analysis.

### 2. **Discretization and Handling Missing Values**

While discretization is a common preprocessing step in many studies, it was not applied in this project due to the nature of the soil attributes. Soil attributes are continuous variables, and discretizing them would have resulted in the loss of critical precision.

Because the soil dataset had no missing values, the focus was on imputing missing data for the climate attributes. The median replacement strategy was chosen, as it is a robust method that minimizes the impact of outliers, ensuring a reliable and consistent dataset for the analysis.

### 3. **Data Integration**

We used GeoPandas `gpd.sjoin` to spatially align soil and climate datasets, ensuring geographic consistency. Seasonal aggregation of climate data simplified the process and focused the analysis on key trends, facilitating efficient integration.

### 4. **Data Reduction**

We reduced temporal complexity by applying seasonal aggregation and rounding longitude/latitude values. This minimized redundancy while preserving important information, making the dataset more manageable for analysis.

### 5. **Normalization**

To ensure consistent model performance, Min-Max normalization and Z-Score normalization were applied. These techniques standardized the data by scaling values and centering the distribution, ensuring that no single attribute dominated the analysis.

By executing these preprocessing steps, we ensured the dataset was cleaned, standardized, and harmonized for accurate analysis. Outlier handling prevented distortion, and avoiding discretization preserved the precision of continuous variables. Data integration maintained spatial consistency, while reduction techniques simplified the dataset without losing key information. Normalization ensured all attributes were comparable in scale, laying a strong foundation for analyzing soil-climate interactions. Data transformation was not performed as the data was already in the format required for analysis, allowing us to focus directly on preprocessing and integration.

## **General Conclusion**

This project aimed to develop a reliable dataset for studying the interactions between soil and climate attributes, which are essential for understanding environmental processes and sustainable land management. Through several key preprocessing steps, we transformed raw data into a cohesive structure, ensuring its quality and suitability for analysis.

The preprocessing phase focused on handling outliers, integrating soil and climate datasets, and normalizing the data. Outliers were addressed using the IQR method, ensuring no distortion in the analysis. Climate and soil data were integrated using spatial alignment, ensuring accurate pairing of climate attributes with soil observations. The normalization methods of Min-Max and Z-Score were applied to standardize the data, facilitating better comparisons across variables.

Additionally, missing values in the climate dataset were handled using median imputation, which preserved the data's integrity. Data reduction techniques simplified the dataset without losing critical information, and the decision to forgo discretization ensured the preservation of continuous soil data.

The resulting dataset, combining spatially aligned and normalized soil and climate data, is now ready for further analysis. It provides valuable insights into the relationship between climate conditions and soil properties, with potential applications in agricultural management and environmental studies. This work lays a solid foundation for future research on how climate impacts soil and ecosystems, contributing to better land management practices.

## References

- [1] Petru Buzulan. Read and analyze netcdf4 files with xarray in python. Medium, n.d. Available at <https://medium.com/@buzulan.petru/read-and-analyze-netcdf4-files-with-xarray-in-python-1c714fad8a66>.
- [2] Sagar Khandelwal. What are the best practices for data preprocessing in mining? LinkedIn, n.d. Available at <https://www.linkedin.com/advice/3/what-best-practices-data-preprocessing-mining-skills-data-mining-35hnc>.
- [3] Yifei Ren. Data preprocessing for data mining, 2013.