

Daily state-wise & district-wise analysis of soil moisture



PROJECT REPORT SUBMITTED TO
Symbiosis Institute of Geoinformatics

FOR PARTIAL FULFILLMENT OF THE M. Sc. DEGREE

By
Harsh Kishor Thakur

PRN 22070243051
M.Sc. (Data Science and Spatial Analytics)

BATCH 2022-24

Symbiosis Institute of Geoinformatics

Symbiosis International (Deemed University)
5th Floor, Atur Centre
Gokhale Cross Road
Model Colony
Pune – 411016
Maharashtra
India

January 2023

Acknowledgment

I would like to convey my heartfelt gratitude to Dr. Moushami Dasgupta for her tremendous support and assistance in completing my project. I would also like to thank her, for providing me with this beautiful opportunity to work on a project with the topic “Daily state-wise & district-wise analysis of soil moisture”. The completion of the project would not have been possible without her help and insights.

I am grateful to all those with whom I have had the pleasure to work during this and other related projects. Nobody has been more important to me in the pursuit of this project than the members of my family. I would like to thank my parents, whose love and guidance are with me in whatever I pursue. They are the ultimate role models.

- Harsh Thakur

Abstract

The project describes the **daily soil moisture level of Indian districts** over the year 2022. The data is fully monitored by the Indian government, the **Indian department of Ministry of Jal Shakti**, the **Department of Water Resources**, and the **National Water Informatics Centre**. The data is fully monitored and released under **National Data Sharing and Accessibility Policy (NDSAP)**.

The **purpose of the research is to observe and analyze the data under study and came out with good information like forecasting soil moisture, forecasting droughts, forecasting floods, analysis of the behavior of rainfall, soil temperature, seasonal conditions, etc**

The project report tells us about the **soil moisture behavior over the year state-wise and district-wise**, how it is pre-processed, what patterns are observed, what is the methodology for the analysis, which model is used to determine the outcome, what is unique information we get from the data, how data behaves over the years and the months, which method is used to collect the samples of the soil moisture, etc.

Tools used to carry out the task is – **Python** (For main coding and analysis purposes), **Dash** (To make an interactive dashboard to visualize graphs), and **Tableau** (Report and graphs formation)

Introduction

The main area under study is the **agriculture domain**. Today, in our country, most land is under agriculture, so a detailed study of soil is highly required. That's why the study of soil moisture levels is also the main aspect. The **daily analysis of soil moisture in Indian districts over the year 2022** was an important study aimed at understanding the variation in soil moisture levels across the country and its impact on agriculture and food security. The study aimed to provide insights into the factors affecting soil moisture levels and to make recommendations to improve soil moisture levels in India. The data is taken from the various departments of the Indian government. The data contains **Volumetric Soil Moisture data for states/UTs and districts**. The Soil Moisture data is calculated based on the output of the **VIC (Variable Infiltration Capacity model) model run by NRSC**. The data is mainly collected by putting the **moisturization calculator under 15cm below the soil**. The daily data of **766 districts in India for 2022** is collected. The data was collected on a daily basis and analyzed to understand the variation in soil moisture levels. The data was also analyzed to understand the impact of various factors such as climate, rainfall, and human activities on soil moisture levels. The tools used to carry out the task are – Python (For main coding and analysis purposes), Dash (To make an interactive dashboard to visualize graphs), and Tableau (Report and graphs formation). The purpose of the research is to observe and analyze the data under study and came out with good information like forecasting soil moisture, forecasting droughts, forecasting floods, analysis of the behavior of rainfall, soil temperature, seasonal conditions, etc. The project report objectives analyze the soil moisture behavior over the year state-wise and district-wise, how it is pre-processed, what patterns are observed, what is the methodology for the analysis, which model is used to determine the outcome, what is unique information we get from the data, how data behaves over the years and the months, which method is used to collect the samples of the soil moisture, etc.

Data Sources

The datasets are downloaded from the official **Indian government data portal** (data.gov.in)

Link: - <https://data.gov.in/catalog/soil-moisture?page=1>

There are .csv files for the soil moisture level according to each month. To study the soil moisture over the year 2022, I've downloaded **12 different .csv files for all 12 months**.

The data comes under **the Ministry of Jal Shakti, the Department of Water Resources, and the National Water Informatics Centre**.

The data is released under **National Data Sharing and Accessibility Policy (NDSAP)**.

Primarily, the data is collected by putting a soil moisturizer calculator under 15cm below the soil, to calculate the soil moisture level.

WHY CHOOSE THIS DATA: -

India is mainly an **agriculture-oriented nation**, there is almost **53.7 % of the total land is under agriculture**. If I do the proper analysis and link this analysis with the other sources like Phosphorous, nitrogen, and hydrogen levels of soil or with the plant diseases or with irrigation maintenance then it will be so much more beneficial for the farmers' who want better yield in their farms.

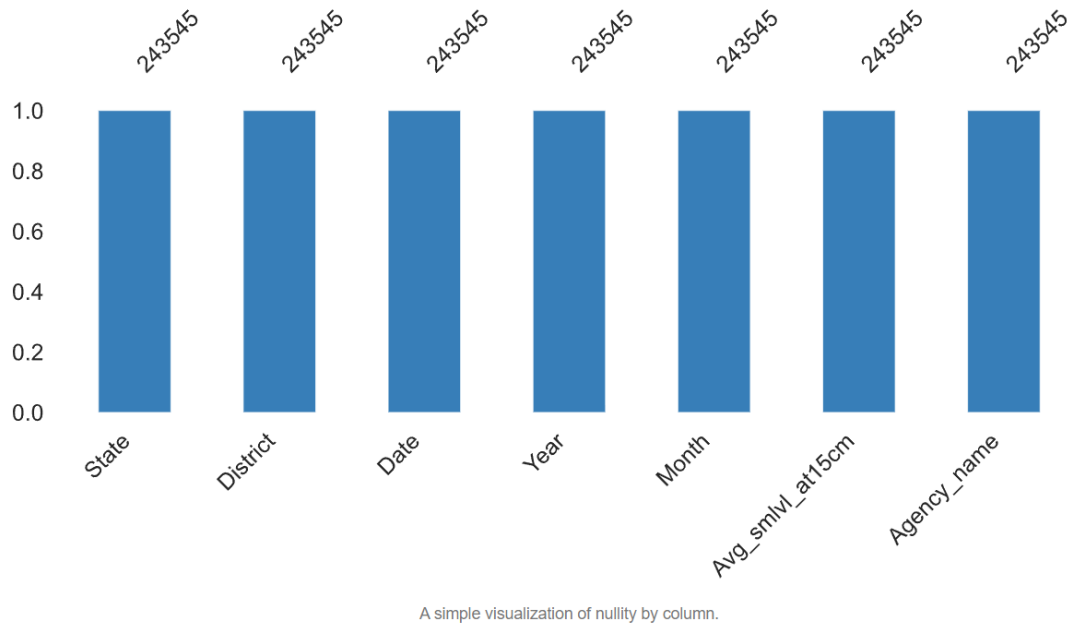
Data Cleaning

A dataset's flaws, irregularities, and missing values are found and fixed during data cleaning, often referred to as data cleansing and database scrubbing. Data cleaning aims to improve the data's correctness and quality so that it is more suited for analysis & decision-making. Duplicate value elimination, management of missing data, outlier detection, etc. are the key components of data cleaning.

The Soil moisture data consist of Missing values, incorrectly inputted data, and outliers. The process is discussed below-

1. Handling missing values in the data: -

All the missing values are from the **state of Andaman and Nicobars**. In the study, it came forward that, there is **very little land is which under agriculture** and there is no scope to calculate the soil moisture out there. So, here we directly **dropped the missing value rows** from the data. After dropping we again checked the missing values using the **missing value detection graph**. Data cleaning helps to improve the quality of the data and to avoid biases and errors in the analysis. It is important to perform data cleaning regularly to ensure that the data remains **accurate and reliable for decision-making**.



2. Outlier Analysis: -

There are almost more than **300 values that are very high (above 10000)** and more than **600 values that are negative (below -100)**. The **higher values indicate that there was a flood** during that region and **lower values indicate that there was a drought** in that region. So these values are generalized as – **a. flood values = 90**

b. drought values = 0

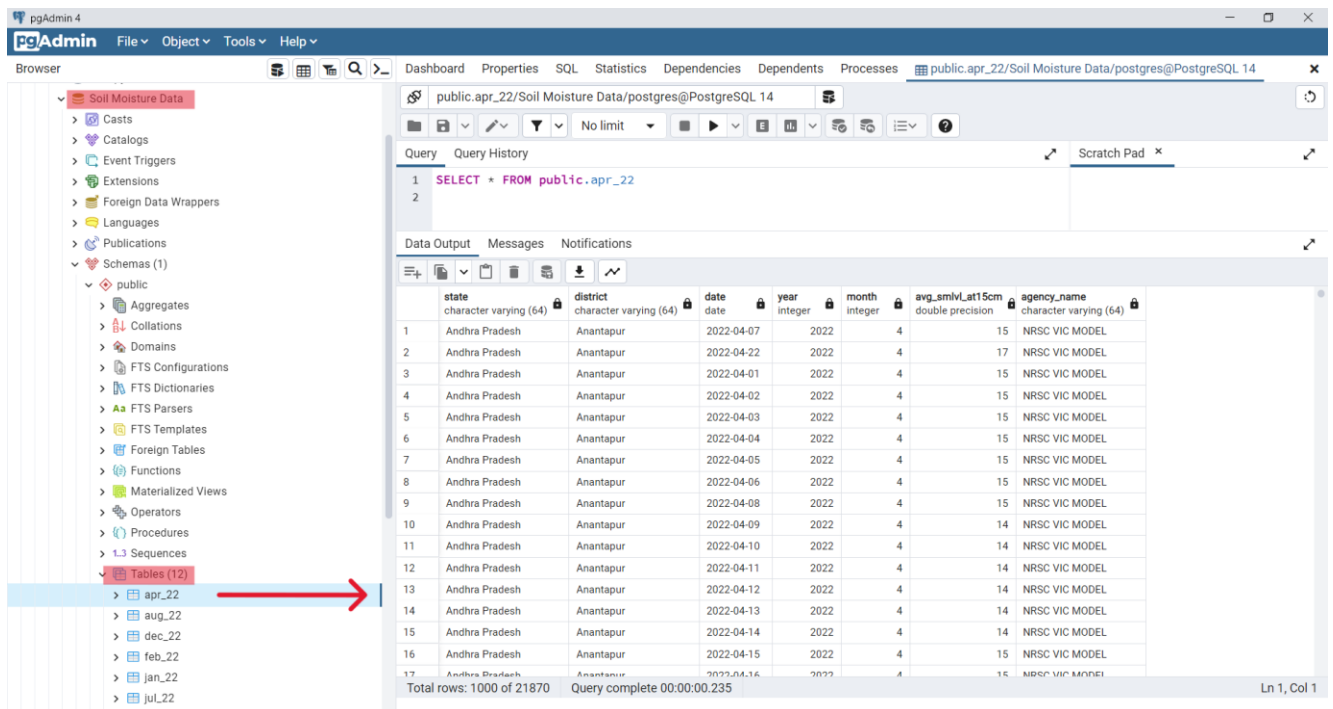
3. Data standardization =

There are almost **90-100 values in the State column that are inputted as 90**, which are of no use. So, we directly **remove those values** from the data.

Data Pre-processing

Preparing raw data in a format appropriate for analysis is known as data preprocessing, and it is a crucial stage in the process of data analysis. Data preprocessing aims to enhance the data quality, lower the possibility of biases and mistakes in the analysis and ready the data to be analyzed. Data preprocessing is crucial since it aids in ensuring the correctness and quality of the data and helps prevent biases and inaccuracies in the analysis. It is an essential phase in the analysis of data and aids in preparing the data for usage in a variety of applications, including machine learning, predictive modeling, and data visualization. Data preparation primarily entails data integration, the transformation of data, and storage.

The soil moisture data was stored in a local drive which is in the .csv format, There is a need to standardize the data source to **maintain the database** and to handle the **dynamic data in the future**. So, I have attached my data to the **PostgreSQL database** and created a table consisting of values under each month.



	state	district	date	year	month	avg_smlvl_at15cm	agency_name
1	Andhra Pradesh	Anantapur	2022-04-07	2022	4	15	NRSC VIC MODEL
2	Andhra Pradesh	Anantapur	2022-04-22	2022	4	17	NRSC VIC MODEL
3	Andhra Pradesh	Anantapur	2022-04-01	2022	4	15	NRSC VIC MODEL
4	Andhra Pradesh	Anantapur	2022-04-02	2022	4	15	NRSC VIC MODEL
5	Andhra Pradesh	Anantapur	2022-04-03	2022	4	15	NRSC VIC MODEL
6	Andhra Pradesh	Anantapur	2022-04-04	2022	4	15	NRSC VIC MODEL
7	Andhra Pradesh	Anantapur	2022-04-05	2022	4	15	NRSC VIC MODEL
8	Andhra Pradesh	Anantapur	2022-04-06	2022	4	15	NRSC VIC MODEL
9	Andhra Pradesh	Anantapur	2022-04-08	2022	4	15	NRSC VIC MODEL
10	Andhra Pradesh	Anantapur	2022-04-09	2022	4	14	NRSC VIC MODEL
11	Andhra Pradesh	Anantapur	2022-04-10	2022	4	14	NRSC VIC MODEL
12	Andhra Pradesh	Anantapur	2022-04-11	2022	4	14	NRSC VIC MODEL
13	Andhra Pradesh	Anantapur	2022-04-12	2022	4	14	NRSC VIC MODEL
14	Andhra Pradesh	Anantapur	2022-04-13	2022	4	14	NRSC VIC MODEL
15	Andhra Pradesh	Anantapur	2022-04-14	2022	4	14	NRSC VIC MODEL
16	Andhra Pradesh	Anantapur	2022-04-15	2022	4	15	NRSC VIC MODEL
17	Andhra Pradesh	Anantapur	2022-04-16	2022	4	15	NRSC VIC MODEL
Total rows: 1000 of 21870			Query complete 00:00:00.235				Ln 1, Col 1

I've created a python code to handle dynamic data. We can collect real-time data by using google forms or using any other format. After we take it into the excel sheet or in the excel .csv format. After it should be passed from the code attached below to get updated in the database.

Code –

```
for i in data.index:
    vals=[data.at[i,col] for col in list(data.columns)]
    q1= "insert into data values('%s','%s')"% (vals[0],vals[1])
    cursor.execute(q1)

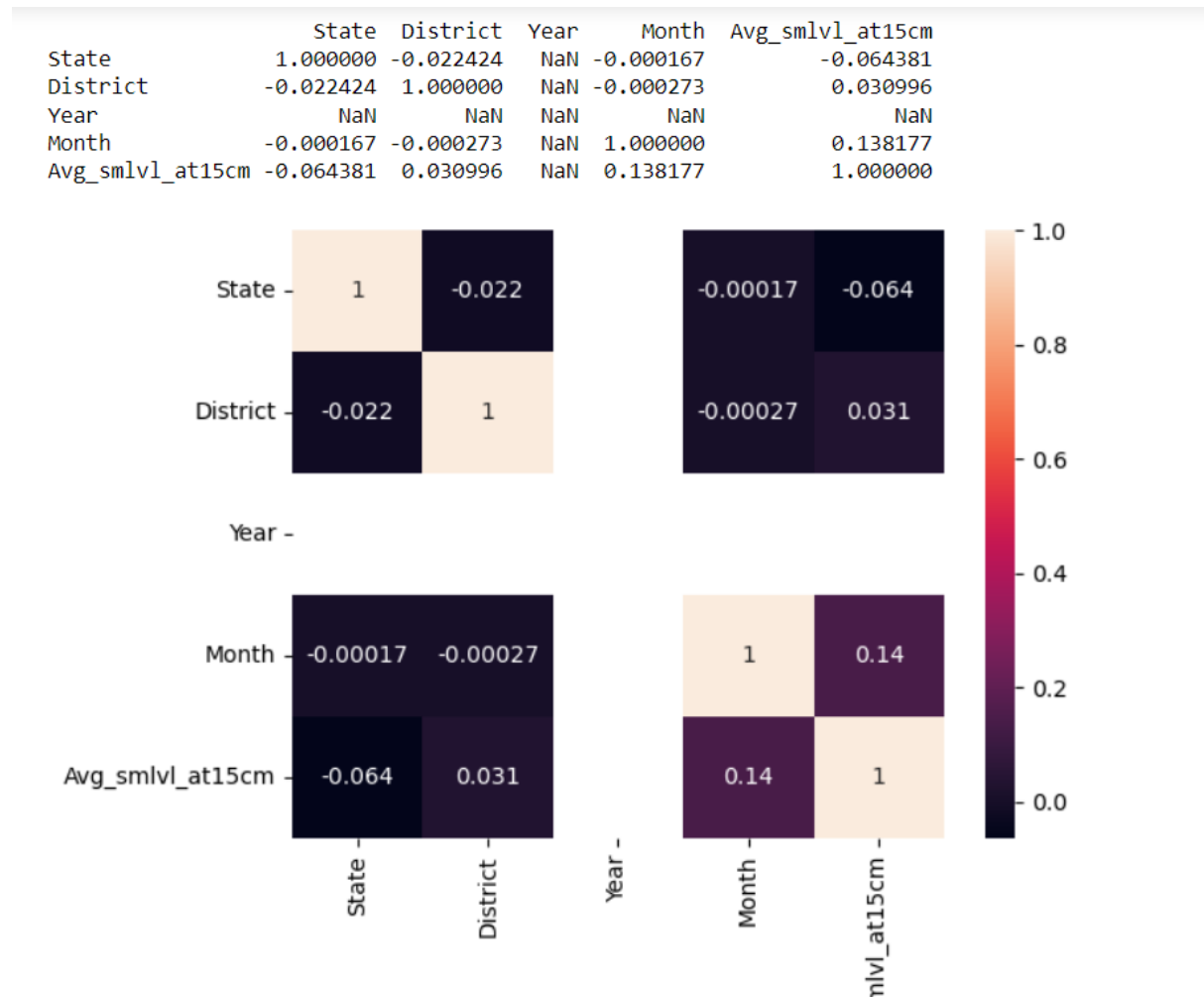
connection.commit()
```

In the data, the **dates are labeled as the string datatype**, so we have changed the datatype to the DATE.

Also, there is a **need of resetting the index as the index column is doubled**.

Label encoding, our data has **many categorical values** so to build the model on the data we need to apply the label encoding technique. Label encoding is the process of transforming variables with a limited range of values, such as category variables, into numerical values. Each distinct category in label encoding is given a distinct numeric value, for example, an integer or a collection of binary values. Because many machine learning algorithms cannot directly operate on categorical data, label encoding is utilized in data analysis. Label encoding makes it possible for algorithms to analyze and predict based on category input by transforming it into numerical values.

Correlation analysis, Correlation analysis is a statistical method used to determine the strength of a relationship between two or more variables. The goal of correlation analysis is to determine if a relationship exists between the variables and to what extent the variables are related to each other. The correlation between the variables is shown below using the heatmap.



Data Analysis

Examining, purifying, converting, and modeling data to provide actionable insights and enhance decision-making is just the process of data analysis. In order to analyze patterns and correlations, extract meaningful information from data, and enable the creation of new knowledge, data analysis is done.

Data analysis techniques used –

1. **Grouping the data state-wise –**

The data contains many districts from one state, if we have 10 districts in 1 state then the **state is getting repeated** 10 times in the column. So we have to group the states using the **python function groupby()**. It **increases the credibility** and readability of the data and helps in analysis purposes.

2. **Grouping the data District-wise –**

The data contains many districts and there are a minimum of **30 dates under every district**. So we have to group the districts using the python function groupby(). It **increases the credibility** and readability of the data and helps in analysis purposes.

3. **Pivot table –**

A pivot table is a data summarization tool used in spreadsheet programs. It allows for the reorganization and summarizing of a large amount of data into a concise and readable format. The pivot table takes simple column-wise data as input and groups the entries into a two-dimensional table with rows and columns. The table can be rearranged by dragging and dropping field labels to create different summaries of the data.

4. **State-wise and district-wise data frames –**

It is necessary to create district-wise data frames **for the plotting purpose**. To increase the credibility and readability of the data which helps in analysis purposes. It also helps in creating plots for **further analysis, visualization, and in data modeling**. I have **created a universal code** in which only we **have to input the dataset and it will return the dictionary of data frames**.

Code –

```
data_df={ }  
for name, group in data.groupby('District'):  
    data_df[str(name)] = group
```

5. Statistical summary =

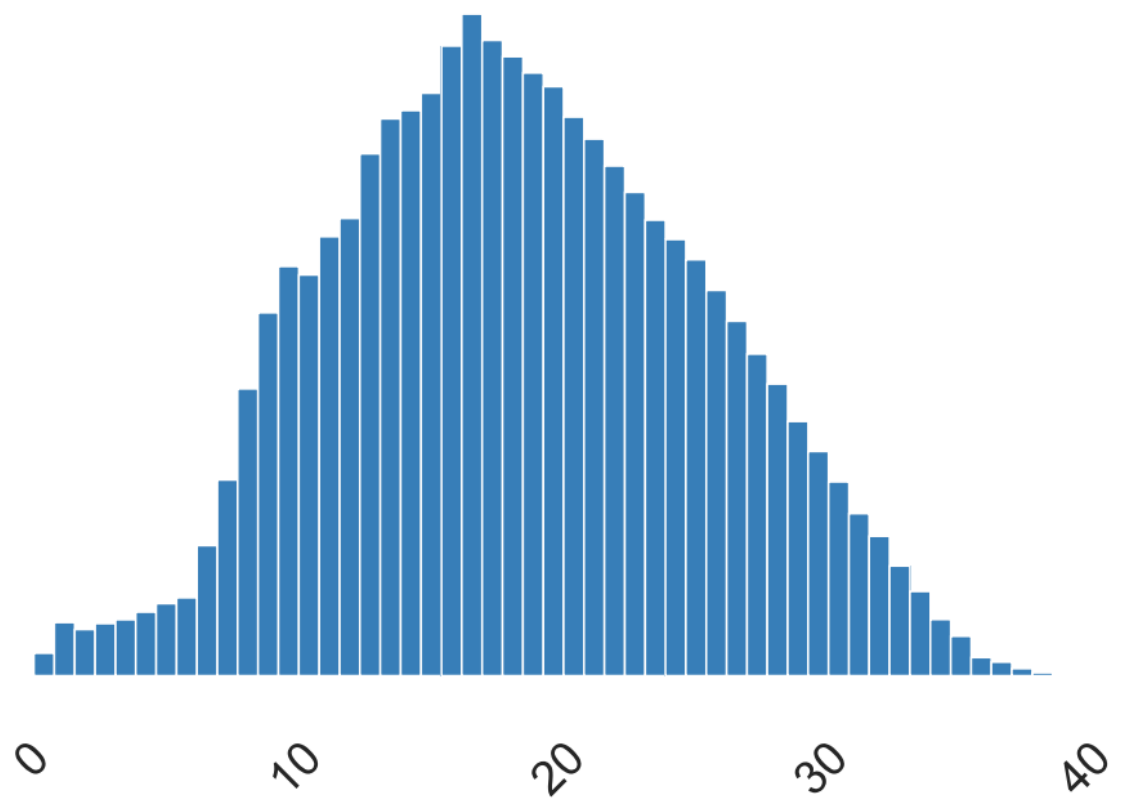
We need to calculate the statistical measures of the columns which are useful in model building. So that there are no anomalies affected while creating a model.

Statistical summary for target variable -

Avg_smlvl_at15cm
Real number (ℝ)

Distinct	242801
Distinct (%)	99.7%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	18.497918
Minimum	0.0109059
Maximum	38.72092
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	1.9 MiB

Normalization graph for target variable –



Quantile measures for target variable –

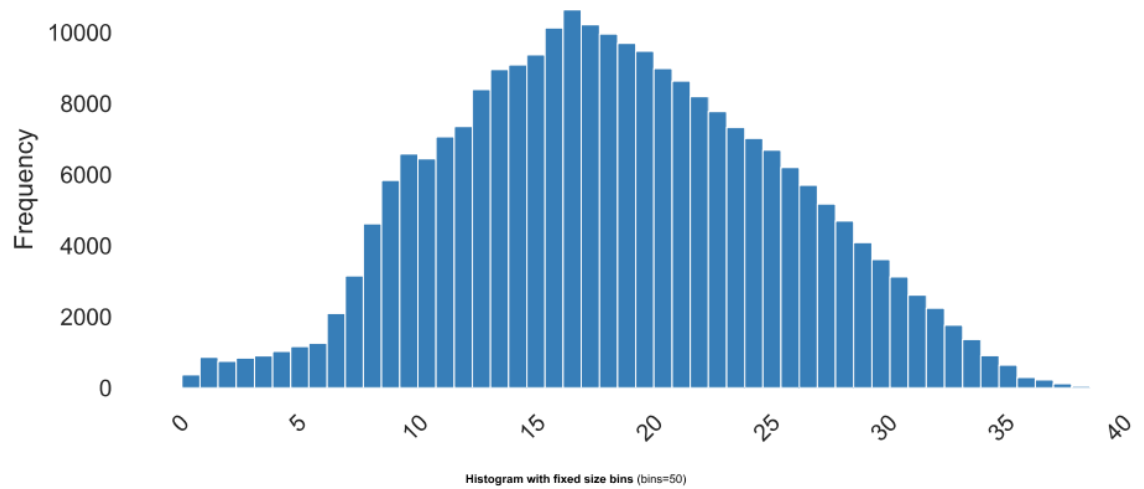
Quantile statistics

Minimum	0.0109059
5-th percentile	7.7145509
Q1	13.360136
median	18.181571
Q3	23.545498
95-th percentile	30.477271
Maximum	38.72092
Range	38.710014
Interquartile range (IQR)	10.185362

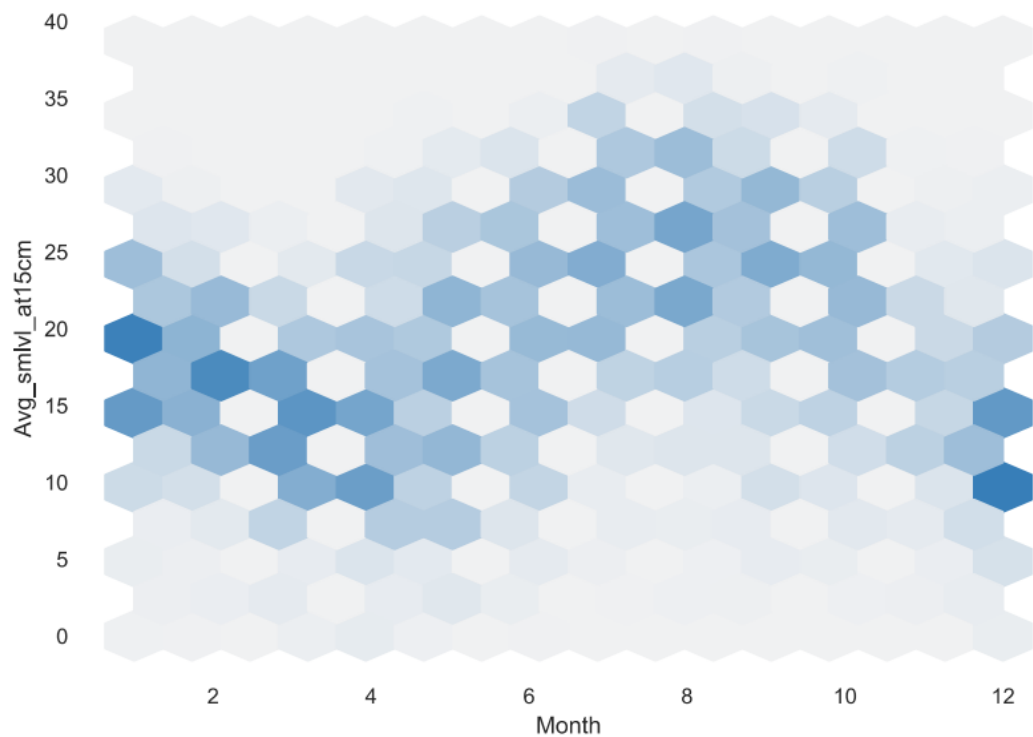
Descriptive statistics

Standard deviation	7.0324955
Coefficient of variation (CV)	0.38017768
Kurtosis	-0.47914055
Mean	18.497918
Median Absolute Deviation (MAD)	5.0640727
Skewness	0.09384603
Sum	4505075.4
Variance	49.455993
Monotonicity	Not monotonic

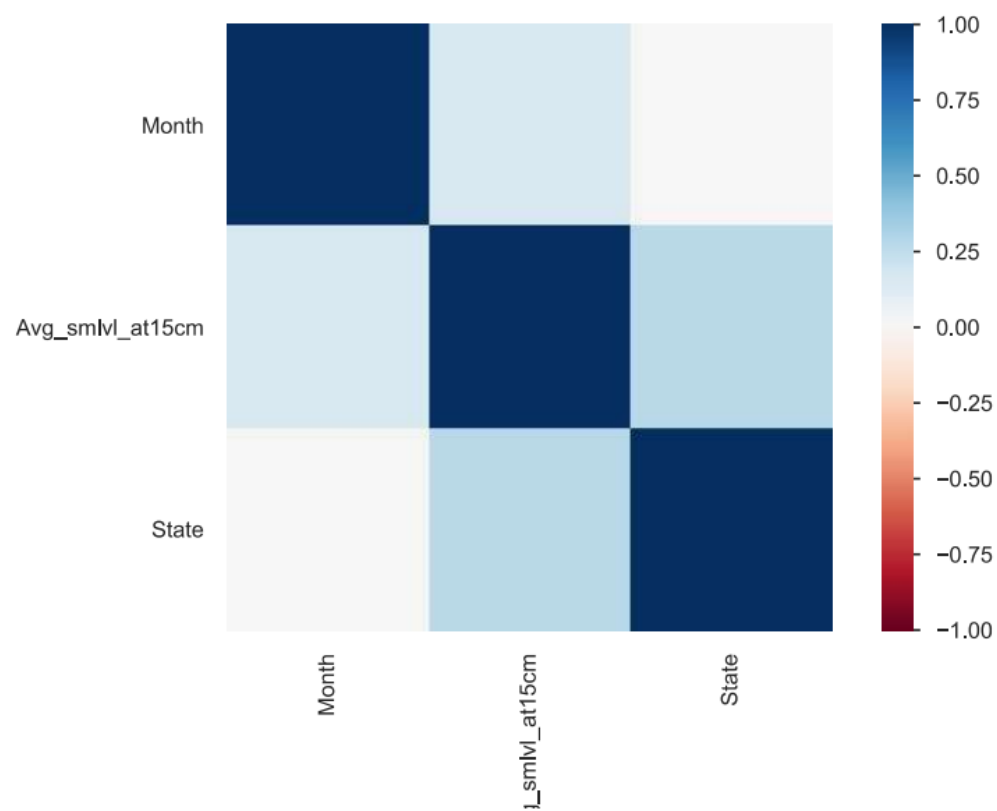
Frequency graph for values of target variable –



Month-wise variation of values of target variable –



Correlation analysis of data –



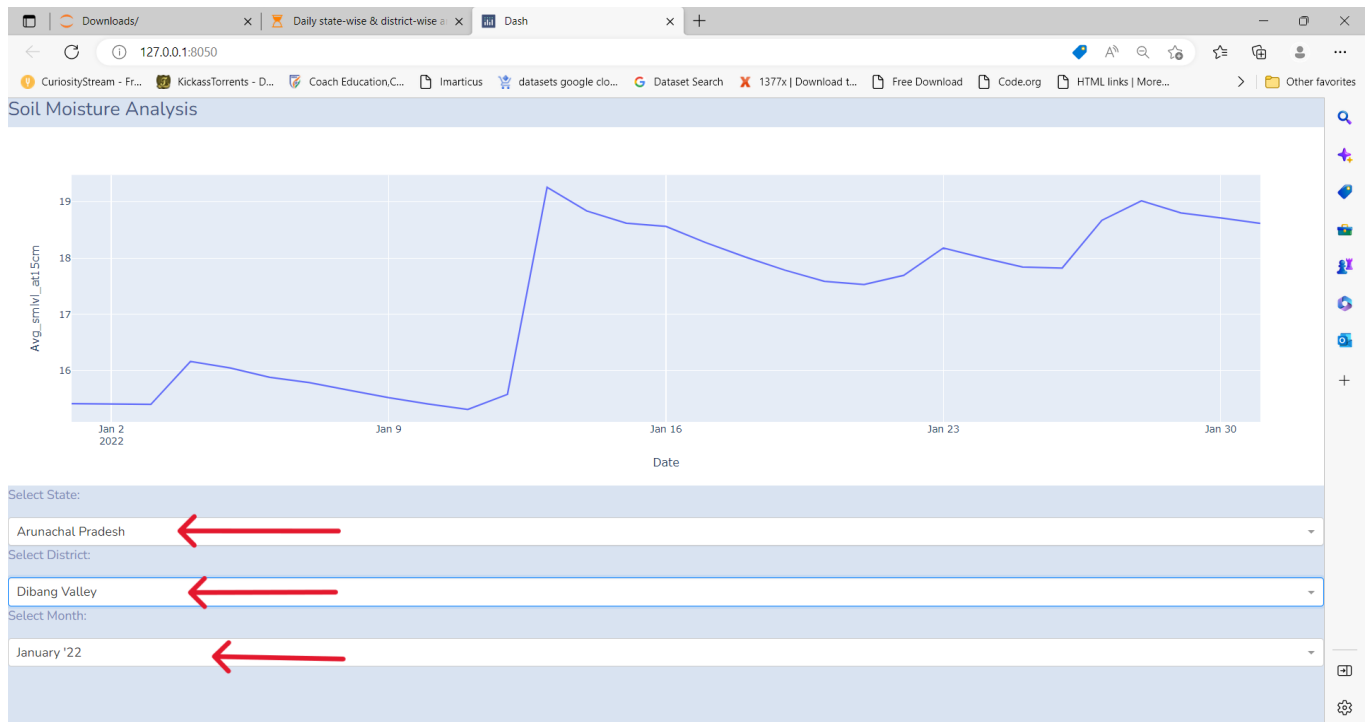
Data Visualization

To make information easier to comprehend and access, data visualization involves producing graphical representations of the data. Effective and efficient data communication of insights, patterns, and correlations is the aim of data visualization. Charts, graphs, maps, and dashboards are just a few of the various formats that data visualization may take. The audience, the desired insights, and the type of information all influence the visualization method used. Heat maps, scatter plots, histograms, bar charts, and line charts are a few examples of common visualization types.

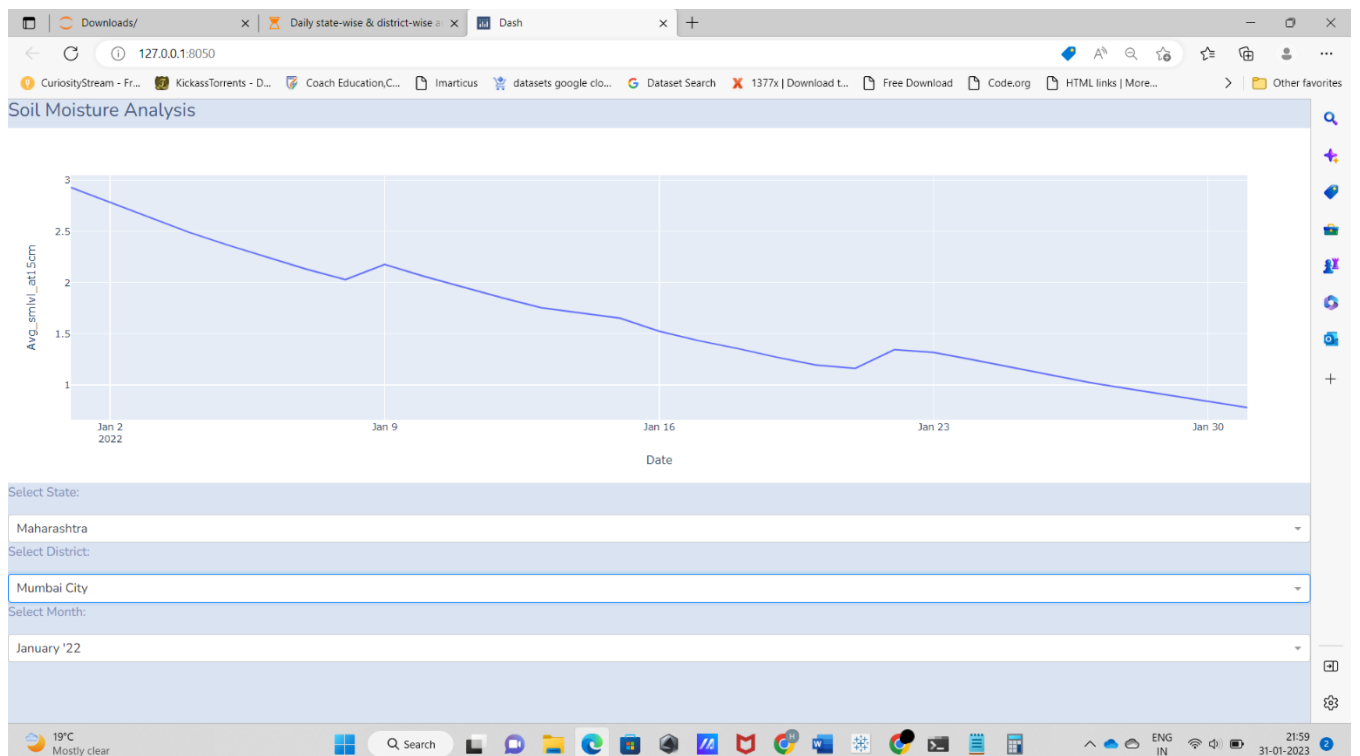
Here I have **used Plotly dash as a dynamic dashboard visualization** for graphs and **tableau to make a static dashboard** of data. Using **Plotly dash I nearly plot and analyse 9158 various graphs** of districts over the months and the year.

Some examples of dashboards created through Plotly dash for the **month-wise** display of time series graphs of soil moisture –

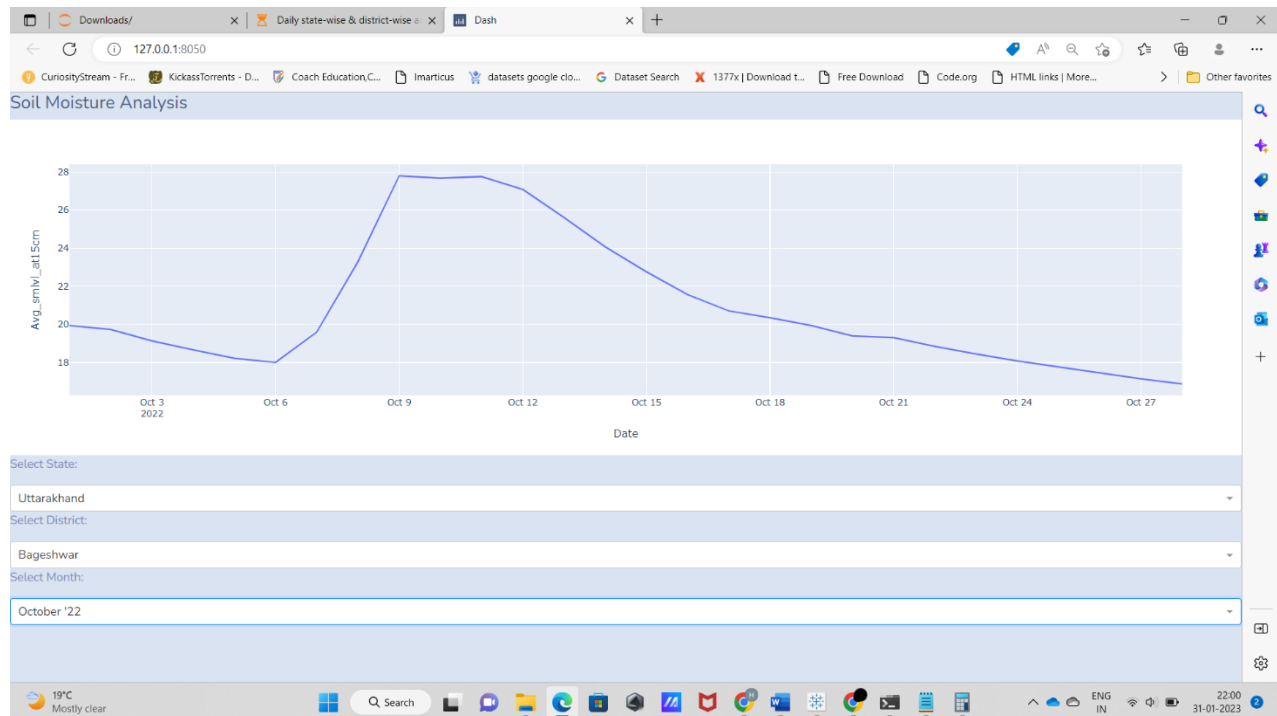
1. Graph of Dibang Valley district in Arunachal Pradesh of month January 2022



2. Graph of District Mumbai City in the state Maharashtra of for the month of January 2022

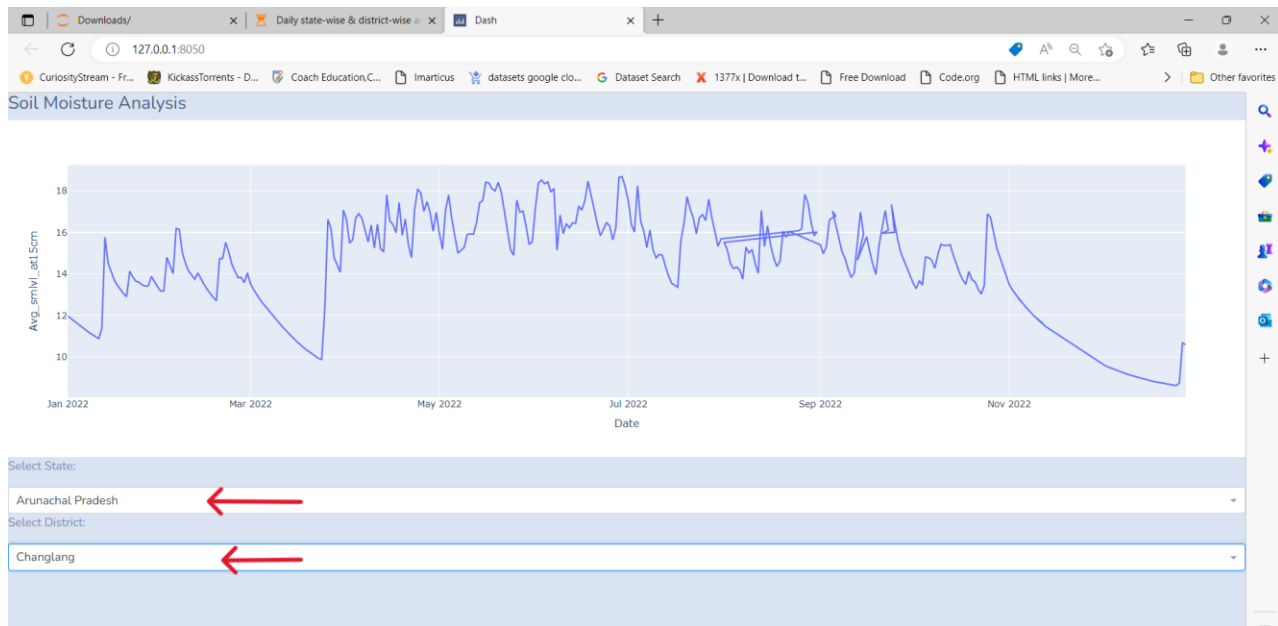


3. Graph of District Bageshwar of State Uttarakhand of month October 2022



Some examples of dashboards created through plotly dash for the **year-wise** display of time series graphs of soil moisture –

1. Graph of District Changlang of State Arunachal Pradesh for the year 2022 –



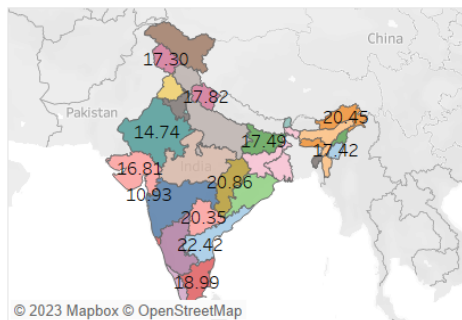
2. Graph of District Yawatmal of State Maharashtra for the year 2022 –



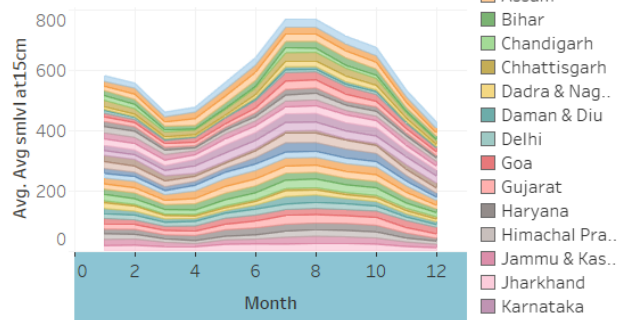
Dashboard of graphs **created through Tableau** –

Daily state-wise & district-wise analysis of soil moisture

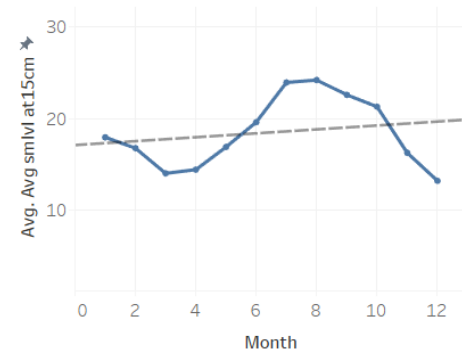
Statewise average soil moisture in India



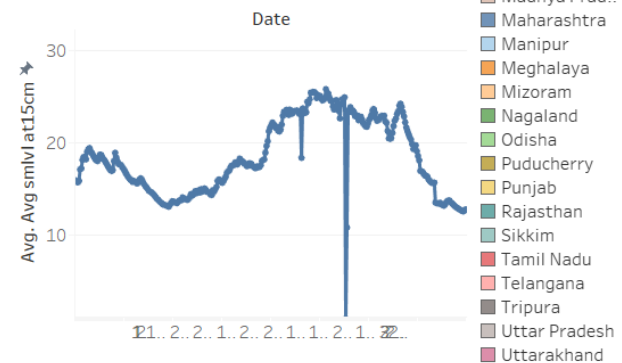
Statewise monthly depth of soil moisture



Monthly analysis of soil moisture



Daily analysis of soil moisture



Conclusion

Soil moisture analysis is **important to understand water availability in the soil** for crops and plants. The daily analysis of soil moisture in Indian districts over the year 2022 was conducted to understand the variation in soil moisture levels across the country.

Some of the key measures from the analysis –

1. There are some states which have **less rainfall**, as a result, the **soil moisture level is low**, they can be identified by a soil moisture level of **less than 10**.
For e.g. – **Gujrat, Rajasthan**, etc.
2. There are some states which having **high rainfall** as a result the **soil moisture level is high**, they can be identified by a soil moisture level **greater than 26**.
For e.g. – **West Bengal, Assam**, etc.
3. **Normal Rain** varies between the soil moisture **12 to 24**.
For e.g. – **Maharashtra, Karnataka**, etc.
4. The analysis gives us a rough **idea about which crops should be harvested** in particular districts. Of course, the soil's Phosphorous, Nitrogen levels matter.
5. The analysis determines the **type of soil**, which can be identified as muddy, wet, moist, normal, dry, or very dry.
6. The analysis gives us an idea about **weather conditions, atmospheric temperature, soil temperature, and crop recommendation** in the particular State or District.

The study was based on the following distinct findings:

1. Different locations and climate circumstances resulted in different soil moisture levels, with some regions having greater soil moisture levels than others.
2. **Soil moisture** levels were significantly **impacted by the monsoon season**, which resulted in **higher soil moisture** levels in areas with **heavy rainfall**.

3. Throughout the year, there were **droughts in some areas**, which resulted in **low levels of soil moisture** which had an impact on agricultural output and agriculture in those areas.

Based on these findings, the study recommended the following **steps to improve soil moisture** levels:

1. **Implementing rainwater harvesting techniques** to conserve and preserve soil moisture levels.
2. Encouraging farmers **to adopt drought-resistant crops and irrigation methods** to reduce the impact of droughts on soil moisture levels.
3. Monitoring soil moisture levels on a regular basis to understand the variation and to take timely measures to conserve soil moisture levels.

According to the study's findings, soil moisture analysis is crucial for the country's food security and the sustained growth of agriculture. For India to have a sustainable future, action must be taken to retain and sustain soil moisture levels.

Future Work

1. Crop recommendation: Soil moisture gives an idea about which crop should be harvested in a particular land. Obviously, the Phosphorous, Potassium, and Nitrogen levels of the soil must take into consideration.
2. Improved measurement techniques: Development of new and more accurate methods for measuring soil moisture, such as satellite remote sensing and drone-based sensing, to increase the accuracy and precision of soil moisture data.
3. Predictive modeling: Development of machine learning and artificial intelligence models to predict soil moisture levels based on historical data and other environmental variables, such as temperature and rainfall.
4. Prediction of flood and drought: As we can predict weather, rainfall, and temperature from the analysis we can also have predictions about drought and floods from the collective data.
5. Climate change impact analysis: Study of the impact of climate change on soil moisture levels and the implications for agriculture and water management.
6. Real-time monitoring: Development of real-time monitoring systems to track soil moisture levels in near real-time, allowing for timely and effective response to changing conditions.
7. Integration with other data sources: Integration of soil moisture data with other data sources, such as weather data, to gain a more comprehensive understanding of the soil moisture-weather relationship.

8. Decision support systems: Development of decision support systems to provide actionable information to farmers and water resource managers, such as the best times to irrigate crops or allocate water resources.
9. Interdisciplinary collaboration: Collaboration between scientists, engineers, and stakeholders to develop integrated solutions to address soil moisture-related challenges and to advance the field of soil moisture analysis.