1. **Data Exploration: Load the dataset and perform initial exploratory data analysis (EDA) to understand the structure, summary statistics, and distribution of variables.**

Ans: The dataset has been loaded using geopandas.The attributes of the dataset (45040x17) are as follows:

- Regions: A total of 45040 villages (uniquely identified using rajas_id) across 245 unique block regions (identified using block_code_census). The villages are uniquely identified using geometry parameters. The coordinate system is EPSG:4326.
- Demographic: Population density has been specified at the block level.
- Healthcare Facilities: This has been specified at the block level.
- Climatic parameters: Rainfall data, PM2.5, ultraviolet index, land surface temperature during the day and night across block regions is provided.
- Health Parameters:
    - Non-communicable Diseases: Number of cases of heart disease, cancer, lung problems, smoker have been specified at the village level.
    - Communicable Diseases: Number of cases of Tb
    - Child and Maternal Health: Number of maternal anaemia cases and child malnutrition cases

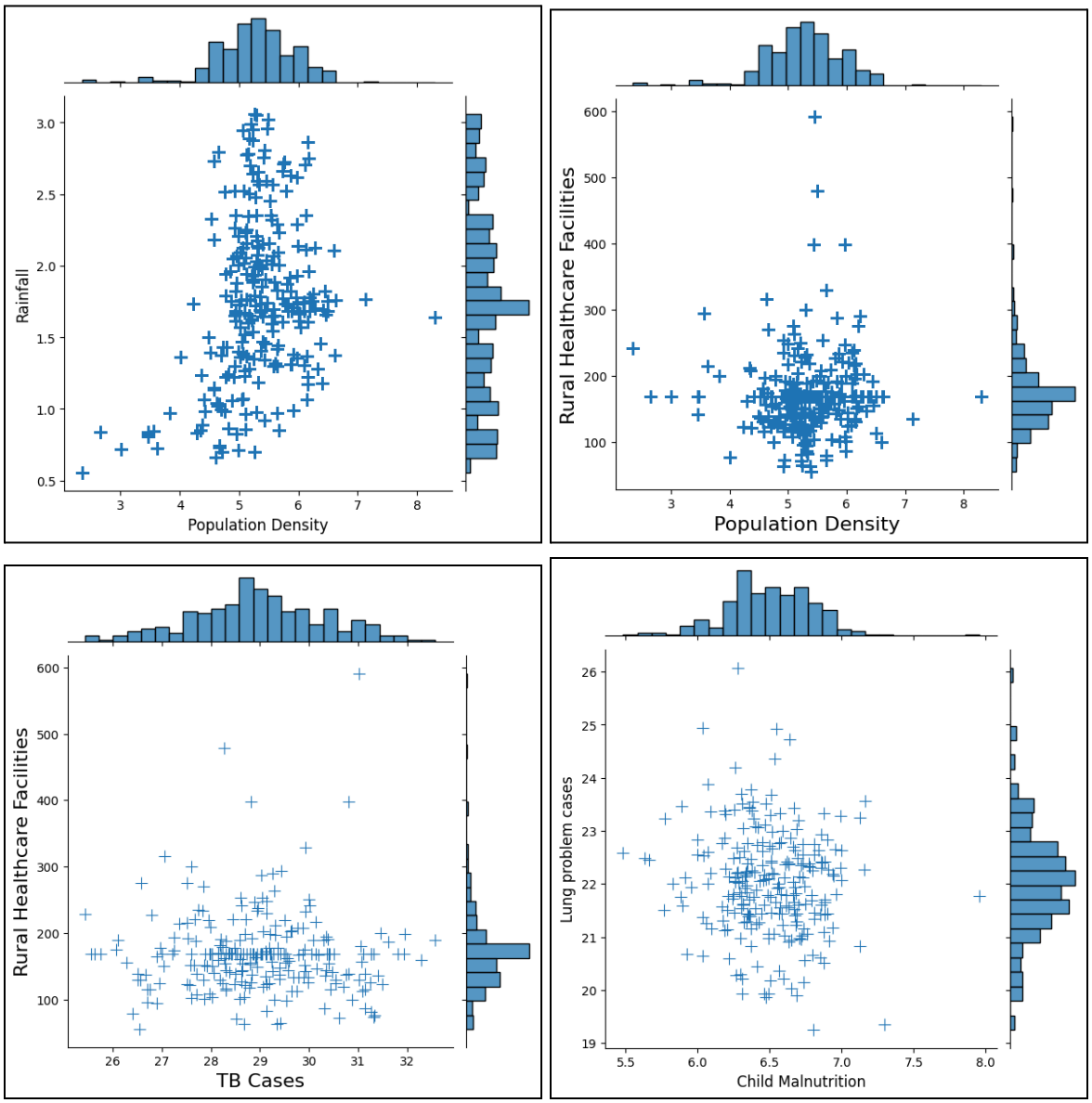
The summary statistics of health parameters are following:

	count	mean	std	min	25%	50%	75%	max
number_of_tb_cases	45040.0	28.94	17.08	0.0	14.0	29.0	44.0	58.0
number_of_heart_disease_cases	45040.0	9.48	5.77	0.0	4.0	10.0	14.0	19.0
number_of_cancer_cases	45040.0	6.99	4.32	0.0	3.0	7.0	11.0	14.0
number_of_child_malnutrition_cases	45040.0	6.50	4.03	0.0	3.0	6.5	10.0	13.0
number_of_maternal_anemia_cases	45040.0	10.54	6.35	0.0	5.0	11.0	16.0	21.0
number_of_smoker_cases	45040.0	52.29	30.70	0.0	26.0	52.0	79.0	105.0

Similarly, the summary statistics of demographic, infrastructural and climatic parameters are following:

	count	mean	std	min	25%	50%	75%	max
rain	45040.0	1.72	0.64	0.56	1.22	1.72	2.16	3.06
population_density	45040.0	5.28	0.70	2.35	4.93	5.28	5.68	8.30
rural_facilities	45040.0	170.18	55.04	55.00	138.00	168.00	191.00	591.00
land_surface_temperature_during_day	45040.0	15194.43	123.46	14770.00	15128.00	15194.43	15277.00	15537.00
land_surface_temperature_during_night	45040.0	14488.07	103.71	14145.00	14438.00	14498.00	14555.00	14871.00
ultraviolet_index	45040.0	9.29	0.46	8.26	9.06	9.30	9.69	10.09
pm_2.5	45040.0	40.36	7.88	26.38	35.08	37.94	42.74	68.49

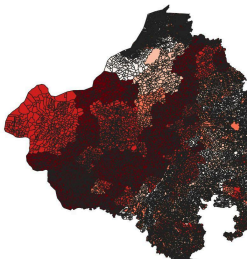
Rainfall is fairly distributed across the region, with some areas registering very little rainfall. As shown in the joint-distribution plot below, the healthcare facilities in rural areas are not properly distributed to account for the population density variation and number of cases across the regions. The cases of both nutritional deficiencies and NCDs (like lung cases) seem to be equally salient across the region.



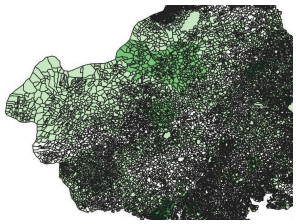
2. **Geospatial Visualization: Visualize the geographic distribution of variables such as rainfall, population density, healthcare facilities, etc., on a map using appropriate visualization libraries**

Ans: The demographic, climatic and health parameters have been visualised using QGIS.

The population density increases from west(arnd 3-4) to east (6-7). In the southern and south-eastern region, it's densely populated except a few pockets/clusters due to hilly regions.In the noth-eastern regions, the population density rises again.

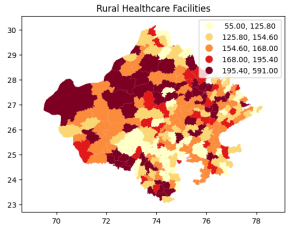


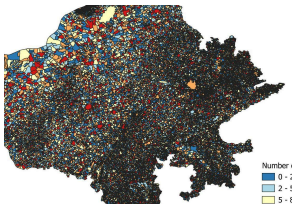
The south-eastern and north-eastern region shows higher rainfall compared to the other regions. Rainfall decreases westward. Land-surface temperature during the day shows distinct regional variations with the western, north-western and south-western region showing relatively higher temperature. This seems to change for Land-surface temperature during the night where the southern, south-eastern and south-western region shows higher temperature.



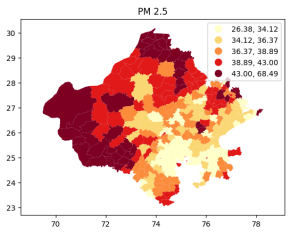
The western, north-western and south-western region shows higher PM2.5 concentration due to prevailing wind dusts.

The healthcare facilities in rural areas seem to be relatively higher in number in the northern regions compared to the southern regions, particularly south eastern regions.In the north-western region, there seems to be a block with relatively higher number of healthcare facilities compared to the nearby regions.

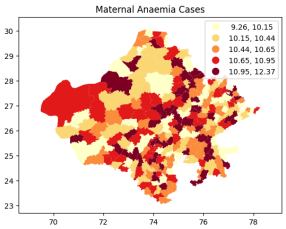


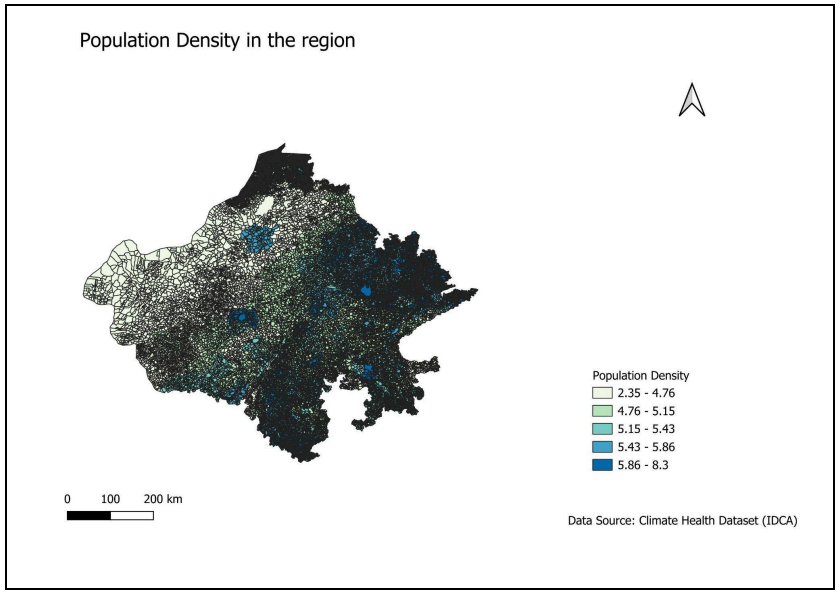


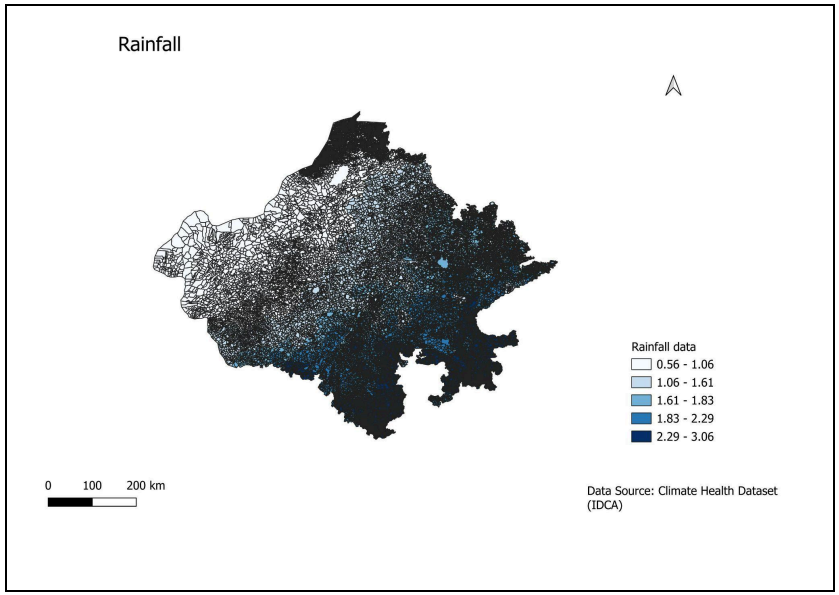
The number of Tb cases and nutritional deficiencies in children and mothers seem to be higher in north-eastern, south-eastern and southern regions compared to the western and north-western regions. There seems to be a spatial cluster in the south-western region with significant number of cases

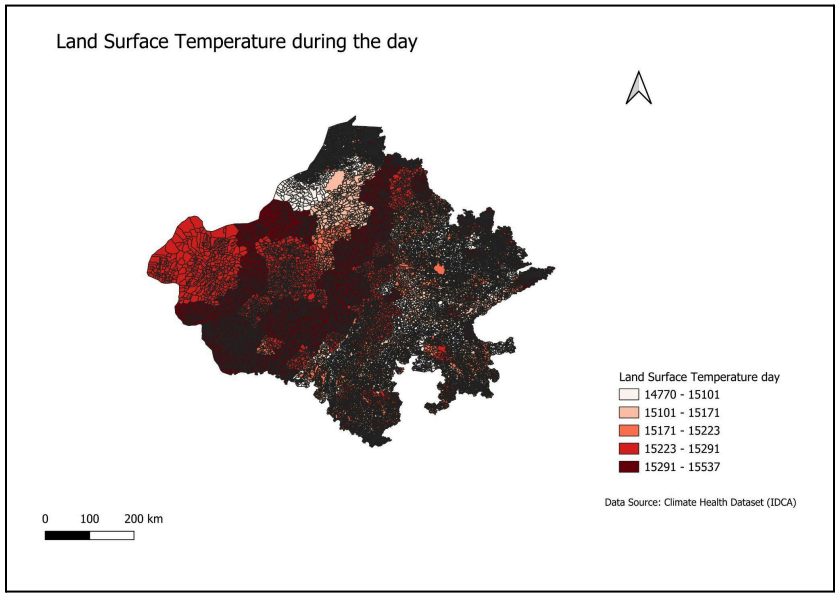


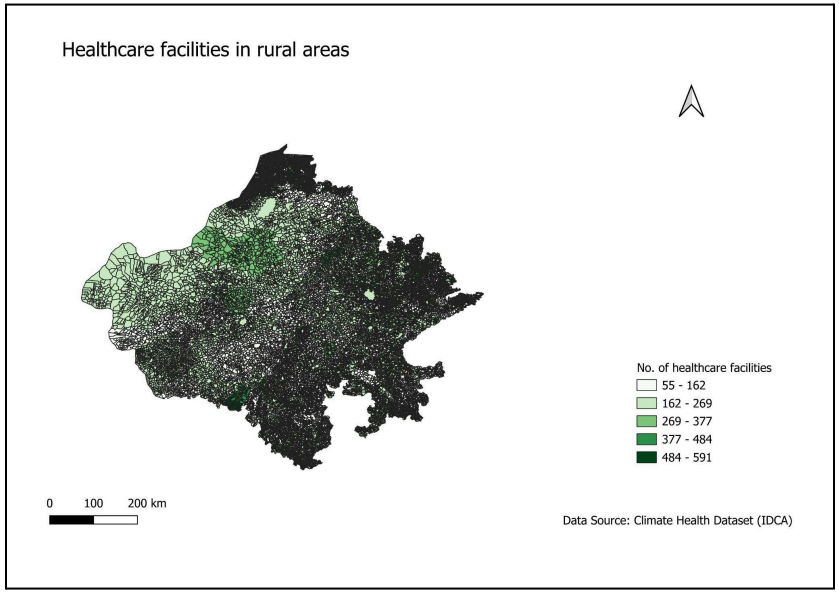
Since the village-level data is quite granular, the geographical visualisation has been attempted at the block level to understand the regional variation. This analysis distinctly shows the regional variation, particularly for climatic variables such as rainfall, PM2.5 etc.

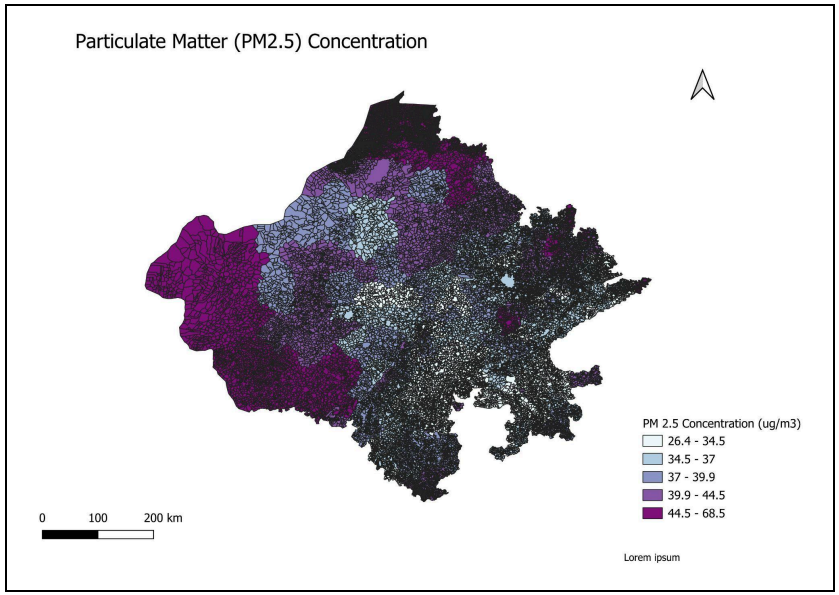


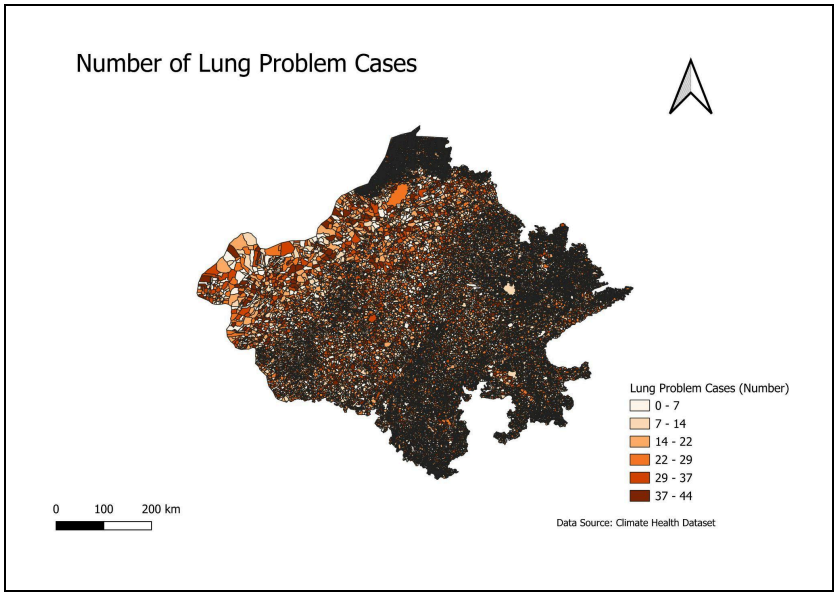

Population Density in the region
Population Density
2.35 - 4.76
4.76 - 5.15
5.15 - 5.43
5.43 - 5.86
5.86 - 8.3
0 100 200 km
Data Source: Climate Health Dataset (IDCA)

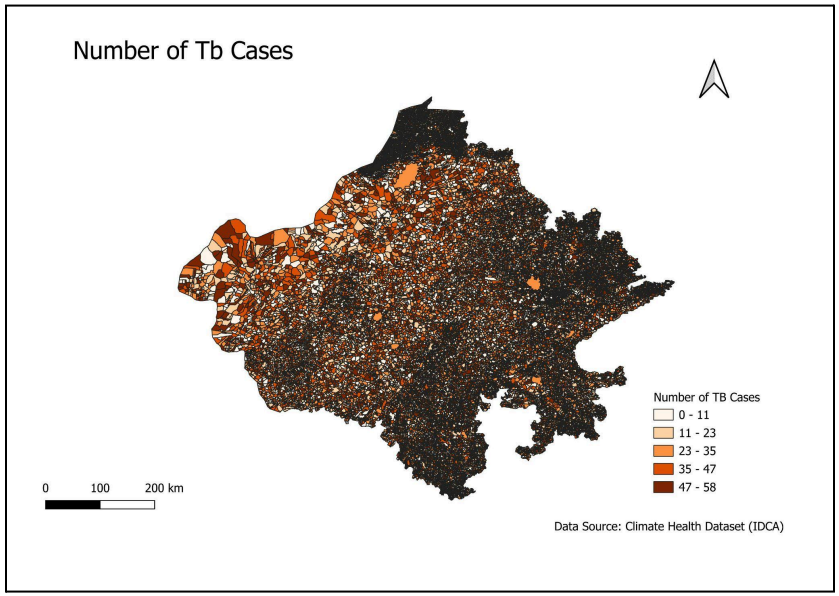

Rainfall
Rainfall data
0.56 - 1.06
1.06 - 1.61
1.61 - 1.83
1.83 - 2.29
2.29 - 3.06
0 100 200 km
Data Source: Climate Health Dataset (IDCA)

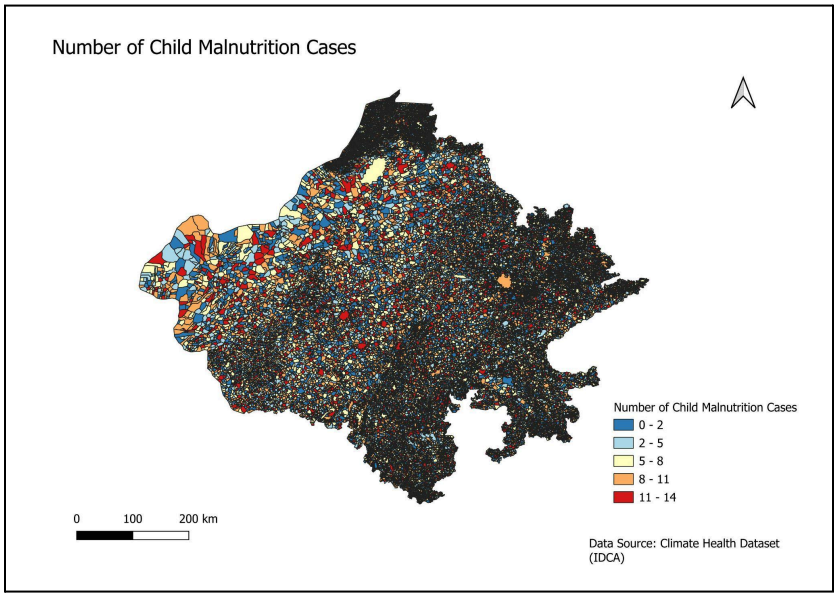

Land Surface Temperature during the day
Land Surface Temperature day
14770 - 15101
15101 - 15171
15171 - 15223
15223 - 15291
15291 - 15537
0 100 200 km
Data Source: Climate Health Dataset (IDCA)


Healthcare facilities in rural areas
No. of healthcare facilities
55 - 162
162 - 269
269 - 377
377 - 484
484 - 591
0 100 200 km
Data Source: Climate Health Dataset (IDCA)


Particulate Matter (PM2.5) Concentration
PM 2.5 Concentration (ug/m3)
26.4 - 34.5
34.5 - 37
37 - 39.9
39.9 - 44.5
44.5 - 68.5
0 100 200 km
Lorem ipsum


Number of Lung Problem Cases
Lung Problem Cases (Number)
0 - 7
7 - 14
14 - 22
22 - 29
29 - 37
37 - 44
0 100 200 km
Data Source: Climate Health Dataset

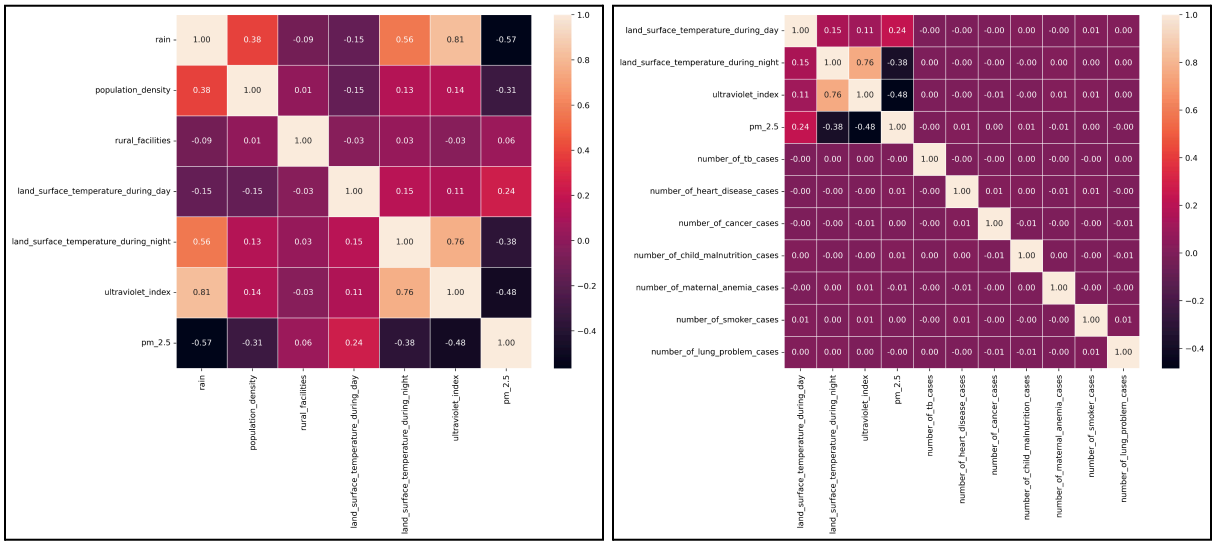

Number of Tb Cases
Number of TB Cases
0 - 11
11 - 23
23 - 35
35 - 47
47 - 58
0 100 200 km
Data Source: Climate Health Dataset (IDCA)


Number of Child Malnutrition Cases
Number of Child Malnutrition Cases
0 - 2
2 - 5
5 - 8
8 - 11
11 - 14
0 100 200 km
Data Source: Climate Health Dataset (IDCA)

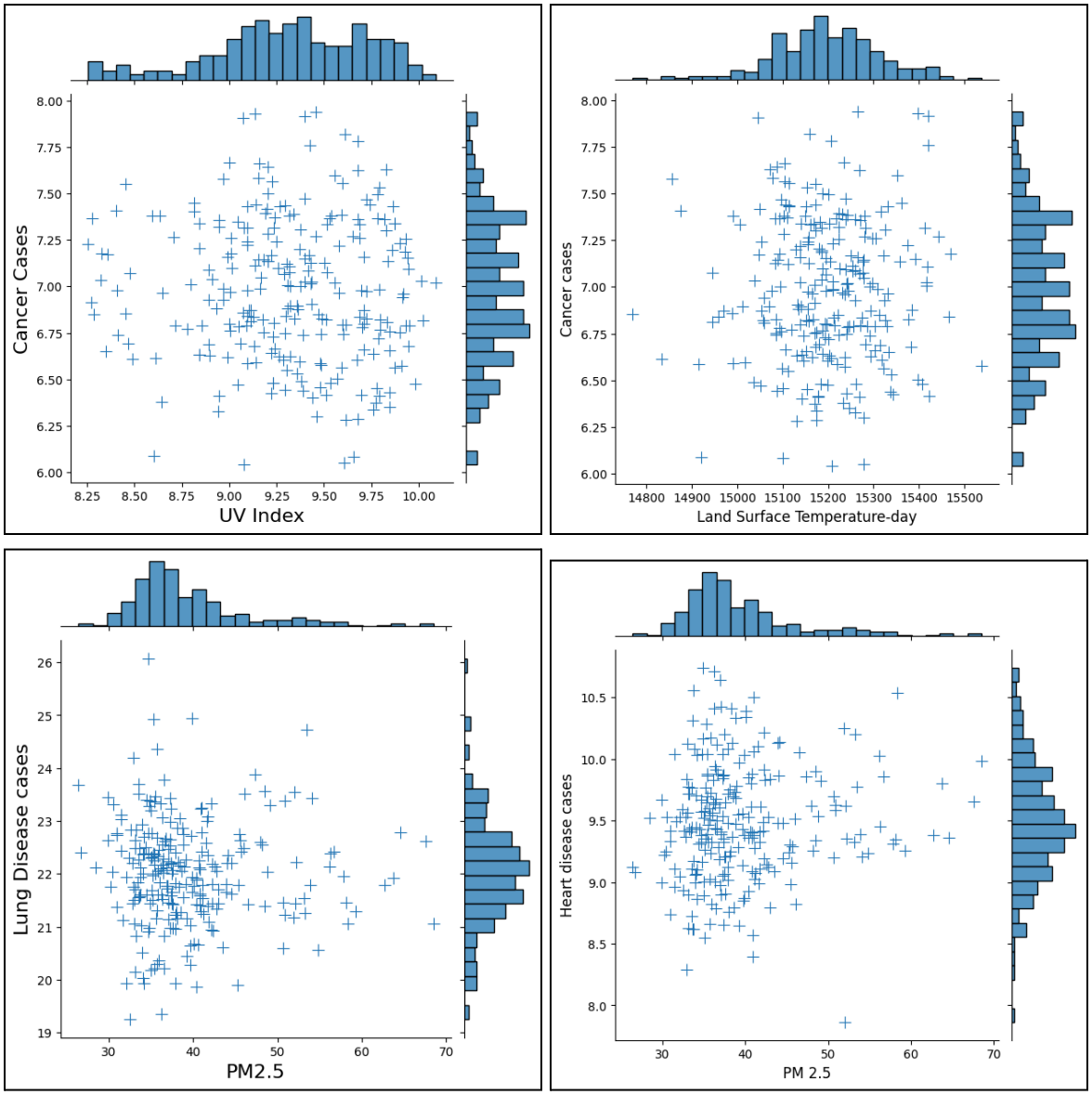
3. **Correlation Analysis: Analyze the correlation between climatic variables (rainfall, temperature, UV index, etc.) and health-related variables (number of TB cases, heart disease cases, etc.). Visualize these correlations using suitable techniques.**

Ans: The correlation between climatic variables and health-related variables have been undertaken using Pearson Correlation Coefficient. The pattern remains the same using the Spearman method. As it's evident, the following correlation seems somewhat significant:

- Between demographic and climatic variables:
    - Rain and Population Density (Positive);
    - PM2.5 and Population Density (Negative);
- Among Climatic variables:
    - Rain and PM2.5 (Negative)
    - Land Surface Temperature during the day and PM2.5 (Positive)
    - Land Surface Temperature during the day and UV index (Positive)
    - Land Surface Temperature during the night and PM2.5 (Negative)
    - PM 2.5 and Ultraviolet Index (Negative)
- Between Climatic variables and health-related variables: No significant Correlation



The joint-plot also shows no correlation between the climatic variables and health-related variables. Higher UV index has both higher as well as lower number of cancer cases. Similarly, the land-surface temperature (day) doesnt seem to have any correlation with the cancer cases. The Lung disease cases are more prevalent in the low PM2.5 regions.



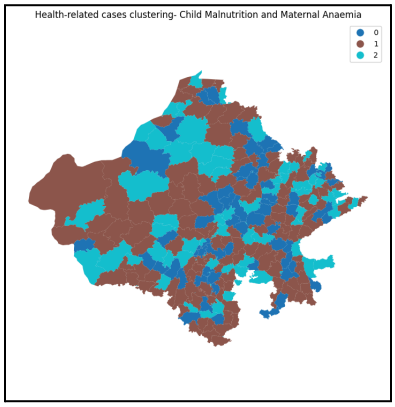
4. **Spatial Analysis: Conduct spatial analysis to identify spatial patterns or clusters of health-related cases across different regions. Use spatial clustering techniques or mapping tools to visualize these patterns.**

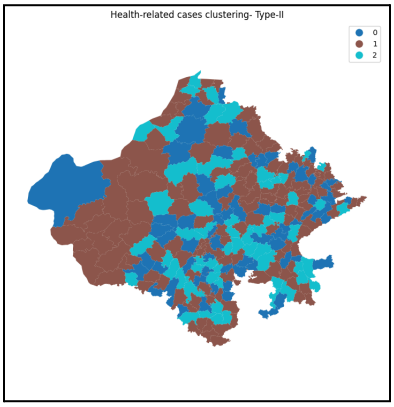
Ans: In order to conduct the spatial analysis, I utilised ‘dissolve’ to combine villages into block regions.

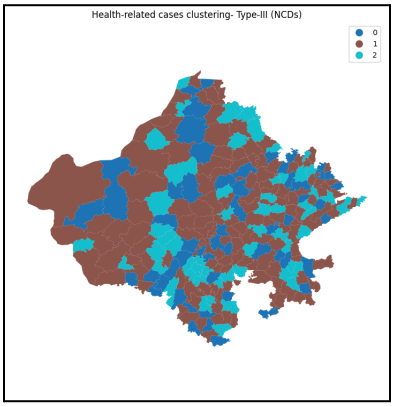
I have divided the health-related cases into three types:

- Type-1: Maternal Anaemia and Children Malnutrition cases-
- Type-2: Health-related cases for which national program has long been in existence and continues to be in focus (Tb-cases, Maternal Anaemia, Children Malnutrition)
- Type-3: Non-communicable diseases (Heart diseases, Lung problem cases, Cancer)

The spatial clustering (using KMeans), the three cluster properties (number of cases) and the elbow curve is shown in the chart below:



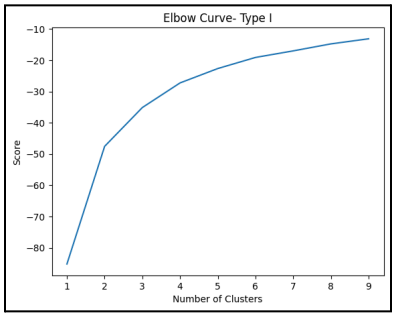


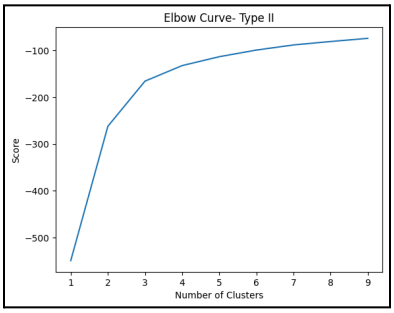


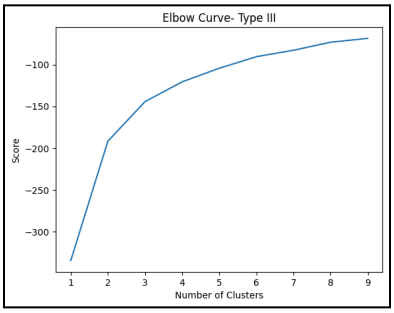
k5cls	number_of_child_malnutrition_cases	number_of_maternal_anemia_cases
0	6.59	11.21
1	6.47	10.56
2	6.47	9.93

k5cls	number_of_tb_cases	number_of_child_malnutrition_cases	number_of_maternal_anemia_cases
0	30.71	6.49	10.60
1	28.92	6.48	10.54
2	27.21	6.53	10.53

k5cls	number_of_heart_disease_cases	number_of_cancer_cases	number_of_lung_problem_cases
0	9.49	6.94	20.72
1	9.56	6.98	21.99
2	9.29	7.02	23.32



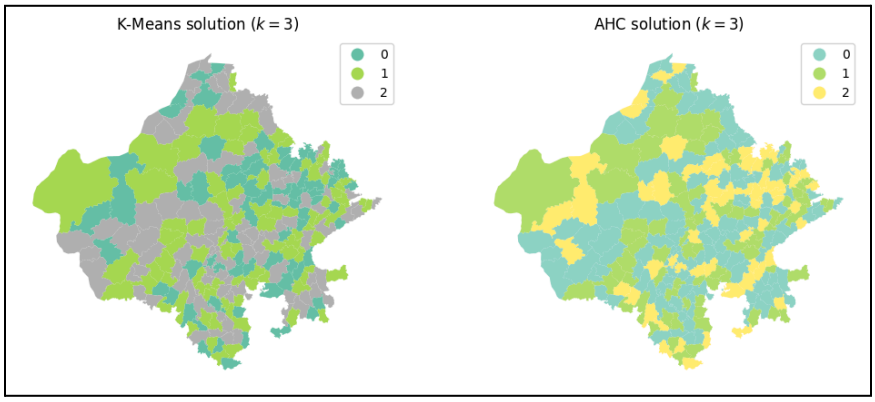




In case of Type-III health-related (health-related) cases, three-clusters differ in particularly the number of lung problem cases and heart cases. In case of TYpe-I health-related cases (nutritional deficiency), the number of maternal anaemia cases seem to be concentrated in north-eastern and southern regions.

The spatial clustering for overall health-related cases using K-Means are shown here. The brown-clusters are one with relatively higher cases of Tb cases, lung disease, children and maternal undernutrition. The clustering is attempted using Hierarchical Clustering algorithm as well (Agglomerative Cluster). The coherence score is calculated and the KMeans algorithm gives a better fit.

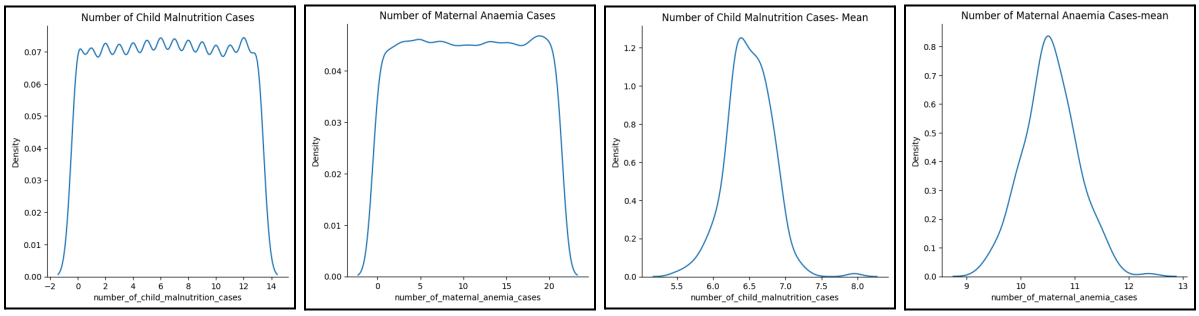
cluster type	
k5cls	27.612662
AHC_cls	25.671677



5. **Hypothesis Testing: Formulate a hypothesis (Does higher PM 2.5 concentration correlate with a higher number of lung diseases?), perform statistical tests, and present findings with appropriate visualizations.**

Ans: The hypothesis is regions with higher maternal anaemia lead to higher children malnutrition and vice versa.

H0: $\mu_a - \mu_b = 0$ Ha: $\mu_a - \mu_b \neq 0$ (Unpaired two-tailed)



Since the normality assumption is not valid, we need to use the nonparametric version of 2 group comparison for unpaired data: the Mann:Whitney U test. Since the p-value from the test is less than 0.05 (the level of significance), the null-hypothesis is rejected.

```
[36] def check_normality(data):
         test_stat_normality, p_value_normality=stats.shapiro(data)
         print("p value:%.4f" % p_value_normality)
         if p_value_normality <0.05:
             print("Reject null hypothesis >> The data is not normally distributed")
         else:
             print("Fail to reject null hypothesis >> The data is normally distributed")

     check_normality(ds2_d['number_of_maternal_anemia_cases'])

     p value:0.5862
     Fail to reject null hypothesis >> The data is normally distributed

     check_normality(ds2_d['number_of_child_malnutrition_cases'])

     p value:0.0022
     Reject null hypothesis >> The data is not normally distributed
```

```
40] ttest,pvalue = stats.mannwhitneyu(ds2_d['number_of_maternal_anemia_cases'],ds2_d['number_of_child_malnutrition_cases'], alternative="two-sided")
    print("p-value:%.4f" % pvalue)
    if pvalue <0.05:
        print("Reject null hypothesis")
    else:
        print("Fail to recejt null hypothesis")

    p-value:0.0000
    Reject null hypothesis
```

6. **Insights and Recommendations: Based on your analysis, provide insights or recommendations for policymakers or healthcare professionals. Identify areas of concern or potential interventions based on the observed patterns or correlations.**

Ans: The insights and corresponding areas of concern/interventions are the following:

- There doesnt seem to be significant correlation between the climatic and health variables. This simply implies that the climatic variables unilaterally donot exert significant influence on health variables for now. However, since the climatic variables are internally significantly correlated, the combined effect of climatic variables need to be analysed on health variables, particularly the non-communicable diseases. For example- heat wave and its effect.
    - Policy Recommendation: Account for additional health-related cases (through OPDs, health-camps etc.) due to climatic variables by strengthening database at the community level
    - Policy Recommendation: Undertake climate vulnerability assessment in the climate-adverse regions (western, south-western).

- There are certain regional pockets (comprising blocks) where the prevalence of certain health-related cases is more prevalent than others. This may be due to socio-cultural practices that need to be accounted for. Additionally, The regional variability at the village level didnt reveal patterns, however the right unit of analysis may be explored further.
    - Policy Recommendations: Allocate financial resources to expand and strengthen the healthcare delivery at the rural healthcare facilities

- There is a strong association between children and maternal nutrition based on hypothesis testing.
    - Policy Recommendations: Strengthen inter-departmental convergence and community outreach to address social determinants of health.