Comprehensive Analysis and Insights on Climate and Health Intersection

TASK 1

LISA Analysis on Heart Disease Cases

Objective: Analyze spatial autocorrelation and clustering of heart disease cases.

LISA is a statistical method used to analyze **spatial autocorrelation**—how similar or different values are in neighboring locations. It helps identify **clusters (hotspots & cold spots)** and **outliers** in geographic data.

Moran's I measures global spatial autocorrelation, determining if values (e.g., heart disease cases) are clustered (+1), randomly distributed (0), or dispersed (-1). **Local Moran's I** identifies specific clusters: High-High (hotspots), Low-Low (cold spots), High-Low and Low-High (outliers), or Not Significant (no clear pattern).

Let's begin our analysis by computing the **Global Moran's I** to assess spatial autocorrelation in heart disease cases. We'll further explore this by visualizing the **Moran's I scatter plot** and the **Kernel Density Estimation (KDE) of its reference distribution**, offering deeper insights into spatial patterns.

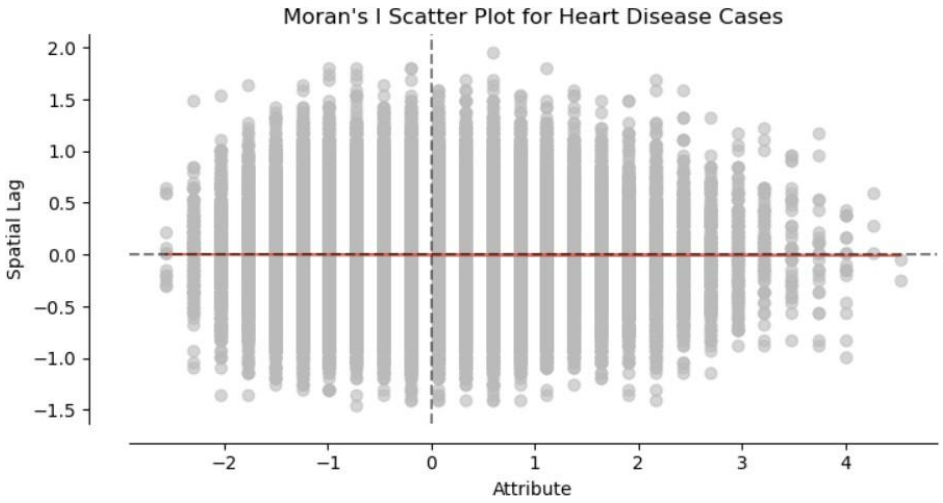
```
# Global Moran's I
moran = esda.Moran(heart_df['heart_disease_cases'], w)
print("\n--- Global Moran's I ---\n")
print(f"Moran's I: {moran.I:.4f}, p-value: {moran.p_sim:.4f}")
```

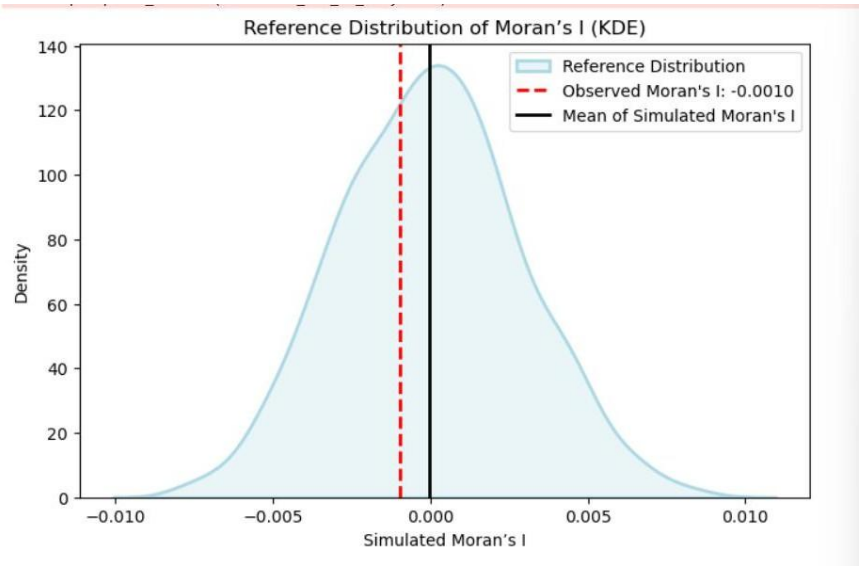
```
--- Global Moran's I ---

Moran's I: -0.0017, p-value: 0.2830
```

- **Moran's I Interpretation (-0.0017):** The value is close to 0, indicating almost no spatial autocorrelation in heart disease cases. Since it's slightly negative, it suggests a **very weak dispersion** (dissimilar values near each other), but the effect is negligible.
- **P-Value (0.2830):** A high p-value (>0.05) means the result is not statistically significant. This suggests that the spatial distribution of heart disease cases is **random** rather than clustered or dispersed.

Conclusion: Data does not show significant spatial clustering or dispersion. The heart disease cases appear to be randomly distributed across regions.

Now, let's take a look at the **Moran's I Scatter Plot for Heart Disease Cases** and the **Reference Distribution of Moran's I (KDE)** to better understand these values.





Moran's I Scatter Plot:

- The scatter plot shows the relationship between heart disease cases (x-axis) and their spatial lag (y-axis).
- The nearly horizontal red regression line suggests a lack of strong spatial autocorrelation.
- Points are widely spread, reinforcing the weak clustering pattern.

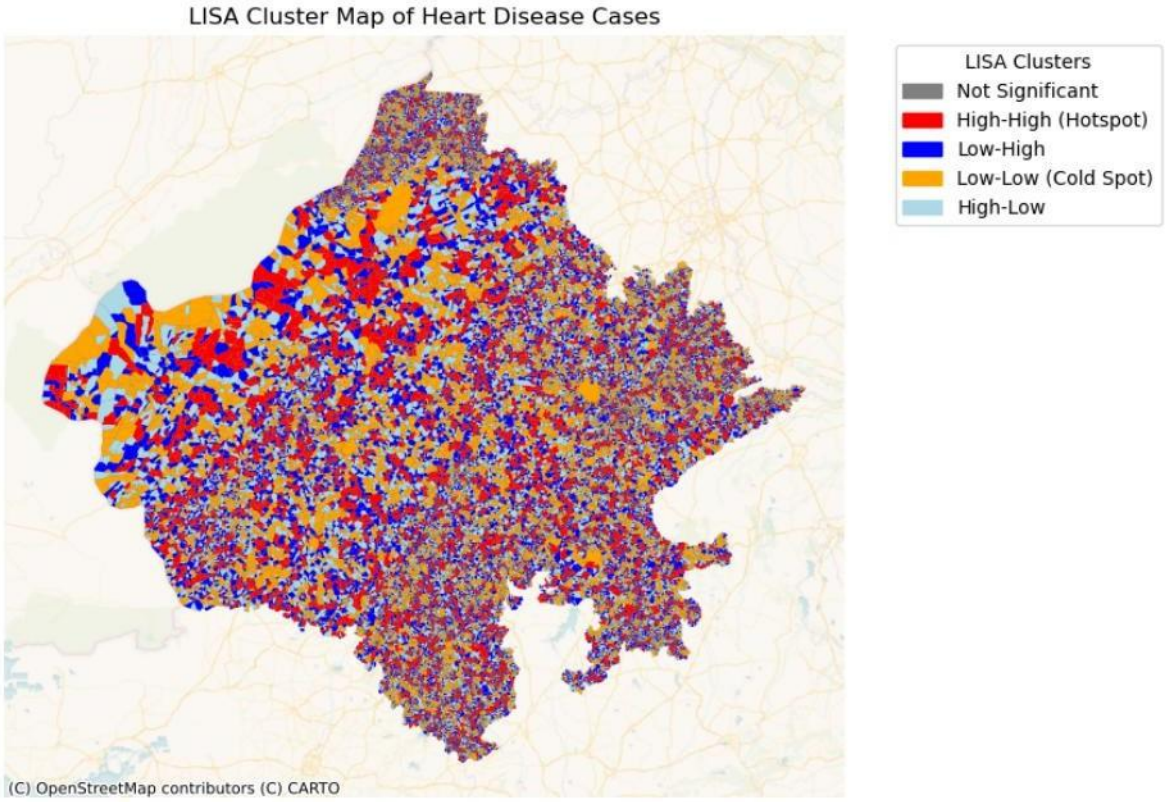
Reference Distribution of Moran's I (KDE)

- This plot compares the observed Moran's I (-0.0010, red dashed line) with a simulated null distribution.

- The observed value is close to zero and aligns with the null distribution's mean (black line), suggesting randomness.
- A non-significant p-value further confirms that heart disease cases do not exhibit significant spatial clustering or dispersion.

Conclusion: There is no strong evidence of spatial clustering of heart disease cases—cases appear randomly distributed across the study area.

Let's explore the **LISA Cluster Map** to uncover spatial patterns in heart disease cases. This map highlights **hotspots** (high-risk areas) and **cold spots** (low-risk areas), aiding in targeted healthcare strategies. Identifying these clusters helps optimize resource allocation and design effective interventions for better public health outcomes.



Insights & Recommendations

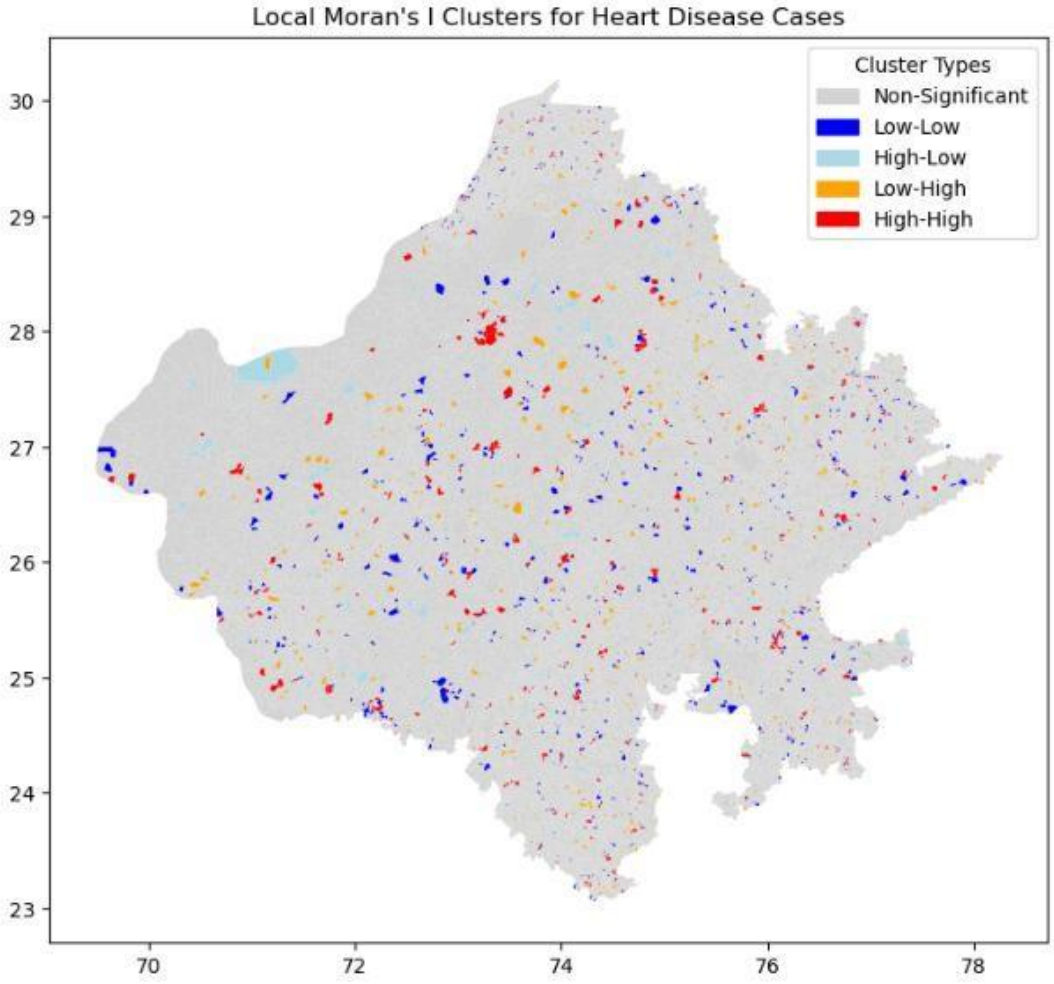
- **Hotspots (Red):** High heart disease cases clustered together, needing urgent interventions like better healthcare access, lifestyle education, and screening programs.
- **Cold Spots (Orange):** Low-case areas, potentially due to healthier lifestyles or better medical facilities.
- **Outliers (Blue & Light Blue):** Disparities in healthcare or environmental factors may explain anomalies—further investigation is needed.

- **Gray Areas:** Random distribution with no significant clustering.

Conclusion

Targeted healthcare policies, awareness campaigns, and preventive measures can help reduce heart disease prevalence, focusing on high-risk regions while sustaining low-risk zones.

This **Local Moran's I Cluster Map** reveals spatial patterns of heart disease cases, highlighting **statistically significant clusters.**



High-High (red) areas indicate disease hotspots, while **Low-Low (blue)** areas show regions with lower prevalence. **High-Low (light blue)** and **Low-High (orange)** areas represent spatial outliers. Understanding these patterns aids in targeted healthcare interventions and resource allocation.

Insights & Recommendations

1. **High-High clusters**: These regions need **urgent medical attention**, awareness programs, and improved healthcare facilities.
2. **Low-Low clusters**: Maintain preventive strategies to sustain low prevalence.

3. **Spatial outliers**: Investigate socio-economic and environmental factors influencing these trends.
4. **Policy Focus**: Implement localized health programs and screenings in high-risk areas.

Based on the comprehensive visualizations and insights derived from the data, the final recommendations addressing the given problem are as follows:

Recommendations:

1. **Targeted Healthcare Interventions:** Prioritize high-risk (High-High) areas for improved medical infrastructure, lifestyle education, and screening programs.
2. **Sustain Low-Prevalence Areas:** Continue preventive measures and healthcare initiatives in Low-Low regions to maintain low heart disease rates.
3. **Investigate Spatial Outliers:** Examine socio-economic and environmental factors contributing to anomalies in High-Low and Low-High regions.
4. **Policy Development:** Implement localized health policies focusing on high-risk clusters while monitoring areas with random distribution patterns.

TASK 2

Climate Health Vulnerability Index

Objective: Develop a Climate Health Vulnerability Index (CHVI) to rank regions based on vulnerability to climate-related health risks.

The Climate Health Vulnerability Index (CHVI) assesses and ranks regions based on their susceptibility to climate-related health risks. It helps policymakers prioritize interventions by identifying high-risk areas.

CHVI Components:

CHVI is a weighted composite index based on three dimensions:

1. **Exposure (Environmental Risks)** – NO_2 density, $PM_{2.5}$ index, land surface temperature, rainfall
2. **Sensitivity (Health Vulnerability)** – Child malnutrition, maternal anemia, smoker cases

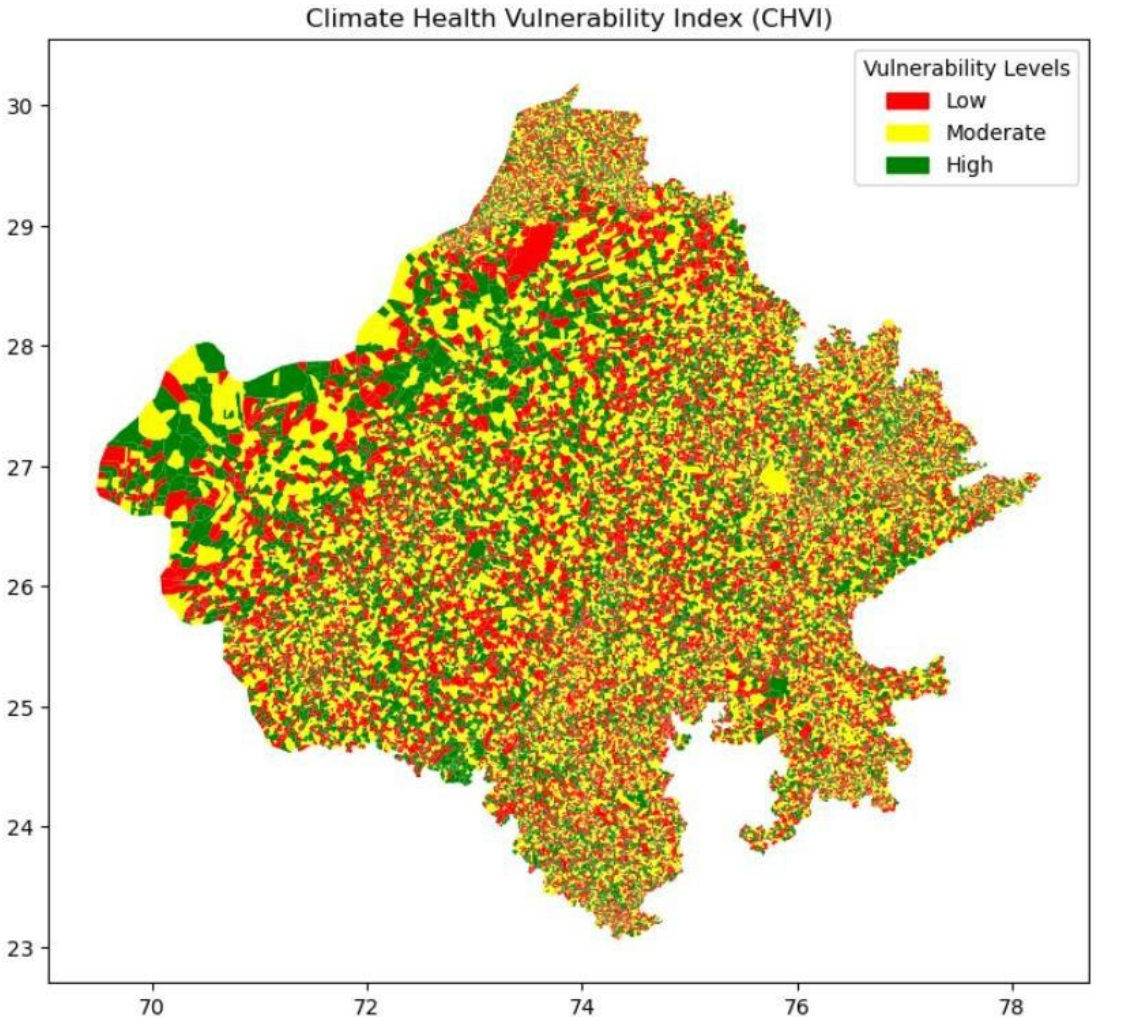
3. **Adaptive Capacity (Resilience)** – Rural healthcare, NDVI (vegetation), NDWI (water availability)

Each factor is normalized and weighted to compute CHVI, ranking regions into **low, moderate, and high vulnerability.**

Significance:

- Guides policy on climate-health adaptation.
- Supports risk mapping via geospatial analysis.
- Enhances resilience by informing healthcare and infrastructure planning.

CHVI is a crucial tool for proactive climate-health management and sustainable development.



Insights, Findings, and Recommendations from CHVI Visualization

Insights

- Most regions have moderate vulnerability (yellow), with high-risk clusters (red) in central and southern areas.

- High-vulnerability areas likely face poor healthcare, high pollution, and low vegetation/water availability.
- Low-vulnerability zones (green) in the northwest suggest better infrastructure and resilience.

Recommendations

- **Strengthen Healthcare & Pollution Control** – Improve medical facilities and reduce air pollution in high-risk areas.
- **Enhance Climate Resilience** – Increase green cover, water management, and public health awareness.
- **Data-Driven Monitoring** – Regularly update CHVI data and track changes for targeted interventions.

Focused actions can mitigate climate-health risks and improve regional resilience.

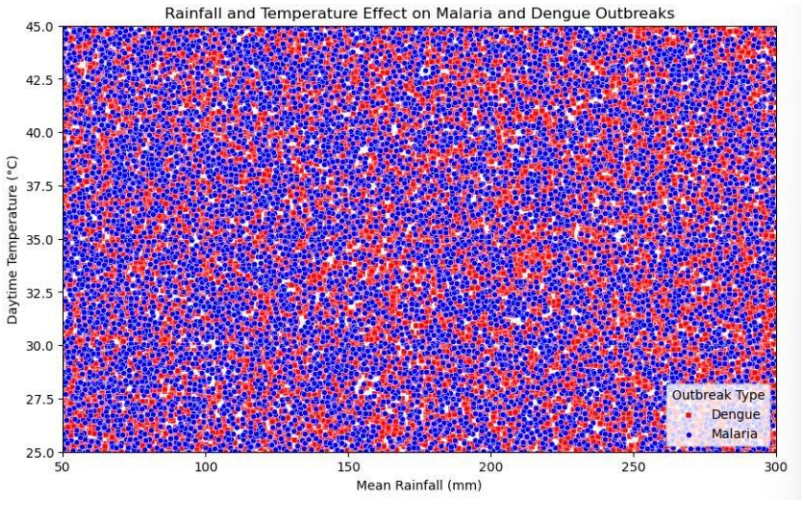
TASK 3

Malaria and Dengue Hotspot Analysis

Objective: Identify malaria and dengue hotspots using rainfall and temperature.

The **Getis-Ord Gi*** statistic is a spatial analysis tool used to detect clustering of high or low values. In this task, it helps identify malaria and dengue hotspots based on rainfall and temperature. By combining hotspot analysis with regression, we assess how these environmental factors influence disease spread, validating our hypothesis statistically.

Let us now begin our task with exploratory data analysis, focusing on the effects of rainfall and temperature on dengue and malaria outbreaks.



Insights from the Visualization:

1. **Rainfall Influence:**
    - Malaria outbreaks (red squares) seem to increase with higher rainfall, supporting the hypothesis that malaria thrives in waterlogged conditions where mosquitoes breed.
    - Dengue outbreaks (blue circles) appear more evenly distributed across different rainfall levels, indicating that standing water (even in lower rainfall regions) may be sufficient for dengue transmission.
2. **Temperature Effect:**
    - Malaria cases are more concentrated in moderate temperature ranges (27-37°C), aligning with the optimal survival conditions for Anopheles mosquitoes.
    - Dengue cases persist even at higher temperatures (up to 42°C), suggesting that Aedes mosquitoes can tolerate a wider temperature range.
3. **Coexistence of Outbreaks:**
    - Both diseases have overlapping regions, but malaria seems more dominant in areas with high rainfall, whereas dengue shows widespread distribution across varying conditions.
    - This implies that controlling malaria may require better drainage systems, while dengue control should focus on localized mosquito breeding sources.

Recommendations:

- **Targeted Prevention:** Malaria control should prioritize high-rainfall regions, while dengue efforts should be widespread, focusing on eliminating small breeding sites.
- **Temperature-Based Surveillance:** Health authorities can implement temperature-based risk models to predict outbreak severity and allocate resources accordingly.
- **Integrated Vector Management:** Combining measures like larvicides, bed nets (for malaria), and public awareness about stagnant water (for dengue) will be more effective in outbreak mitigation.

Now, let's delve deeper into the correlation between temperature and rainfall by calculating the Variance Inflation Factor (VIF).

```
vif = pd.DataFrame()
vif['Feature'] = X.columns
vif['VIF'] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
print("\n=== Variance Inflation Factor (VIF) ===\n", vif)
```

```
=== Variance Inflation Factor (VIF) ===
           Feature        VIF
0            const  24.493860
1    mean_rainfall   1.000014
2  lst_day_celsius   1.000014
```

Interpretation of Variance Inflation Factor (VIF) Results:

1. **VIF for mean_rainfall and lst_day_celsius:**
    - Both features have a VIF close to **1.000014**, indicating **no multicollinearity** between them.
    - This means that rainfall and temperature are independent predictors, making them reliable for logistic regression modelling.
2. **VIF for const (Intercept):**
    - A high VIF (24.49) for the constant term is expected and does not indicate multicollinearity issues.
    - It results from the intercept being correlated with the mean of the independent variables.

Conclusion:

- The low VIF values confirm that **rainfall and temperature can be used together** in the logistic regression model without concerns about redundancy.
- The model is likely to perform well without issues related to collinearity between predictors.

Based on the conclusions derived, the **logistic regression model** and its summary are as follows:

```
Optimization terminated successfully.
         Current function value: 0.613632
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:          outbreak_type   No. Observations:                45040
Model:                          Logit   Df Residuals:                    45037
Method:                           MLE   Df Model:                            2
Date:                Sun, 16 Mar 2025   Pseudo R-squ.:                0.004026
Time:                        15:15:42   Log-Likelihood:                -27638.
converged:                       True   LL-Null:                       -27750.
Covariance Type:            nonrobust   LLR p-value:                 3.051e-49
==============================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const              0.6135      0.051     12.134      0.000       0.514       0.713
mean_rainfall      0.0018      0.000     14.853      0.000       0.002       0.002
lst_day_celsius   -0.0017      0.001     -1.217      0.223      -0.005       0.001
==============================================================================
```

Insights from Logistic Regression Results:

1. **Model Performance:**
    - The **Pseudo R-squared (0.004026)** is quite low, indicating that rainfall and temperature alone may not strongly explain malaria and dengue outbreaks.

- The **log-likelihood (-27638)** and **LLR p-value (3.051e-49)** suggest that the model is statistically significant.

2. **Effect of Predictors:**
   - **Rainfall (mean_rainfall)** has a **positive and statistically significant effect (p < 0.001)** on the outbreak type, meaning higher rainfall is associated with a higher likelihood of malaria compared to dengue.
   - **Temperature (lst_day_celsius)** has a **negative but non-significant effect (p = 0.223)**, suggesting that temperature alone may not be a strong predictor in distinguishing between malaria and dengue outbreaks.

Conclusion:

- Higher rainfall increases the probability of malaria outbreaks compared to dengue.
- Temperature does not significantly impact the outbreak type.
- Other factors (e.g., humidity, stagnant water) should be considered for better predictive accuracy.

The key **interpretations and insights** derived from the model are as follows:

```
=== Model Interpretation ===
Log-Likelihood: -27638.002671651917
Pseudo R-squared: 0.004025670264845571
Variable Coefficients:
const              0.613522
mean_rainfall      0.001761
lst_day_celsius   -0.001732
dtype: float64
```

Further Interpretation of Model Results:

1. **Log-Likelihood (-27638.0027)**
   - A lower (more negative) log-likelihood suggests that the model does not fit the data very well. While it is statistically significant, its explanatory power remains low.
2. **Pseudo R-squared (0.0040267)**
   - This very low value implies that **only 0.4% of the variability** in the outbreak type (malaria vs. dengue) is explained by rainfall and temperature.
   - This suggests that **additional features** (e.g., humidity, population density, water stagnation, mosquito breeding conditions) might be needed to improve the model.
3. **Variable Coefficients:**

- **Constant (0.6135):** The baseline log-odds of malaria outbreak when both predictors are zero.
- **Rainfall (0.00176, positive):** A **small but significant** positive effect, meaning **increased rainfall slightly raises the probability of malaria outbreaks** over dengue.
- **Temperature (-0.00173, negative):** A **slightly negative but non-significant** effect, suggesting that **higher daytime temperatures might reduce the likelihood of malaria outbreaks compared to dengue**, but this result is not strong.

Key Takeaways:

- Rainfall is a significant predictor, **positively influencing malaria outbreaks**.
- Temperature does not strongly affect outbreak type in this model.
- The model explains **very little variation** in outbreak type, **indicating the need for additional features**.
- **Next Steps:** Consider adding **humidity, elevation, mosquito population data, or socioeconomic factors** for a better prediction model.

Now, we will conduct hypothesis testing to evaluate the validity and relevance of our model, assessing whether the observed relationships between variables are statistically significant or simply due to random chance

```
=== Hypothesis Testing ===
Rainfall Coefficient: 0.0018, P-Value: 0.0000
- The effect of rainfall on malaria outbreaks is statistically significant.
```

Interpreting the Hypothesis Testing Results:

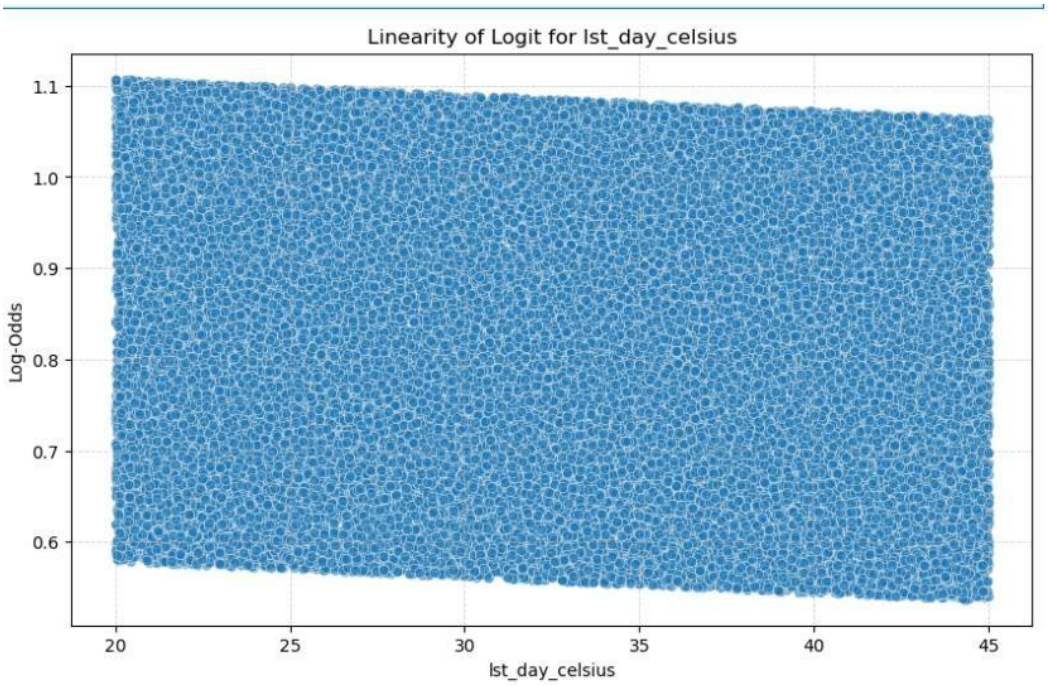
1. **Rainfall Coefficient (0.0018):**
   - This positive coefficient suggests that an **increase in rainfall slightly increases the likelihood of malaria outbreaks**.
2. **P-Value (0.0000):**
   - Since **p-value < 0.05**, we **reject the null hypothesis** (which assumes no effect of rainfall).
   - This confirms that **rainfall has a statistically significant impact on malaria outbreaks**.

Conclusion:

- Rainfall **positively influences malaria outbreaks**, meaning **higher rainfall is associated with a greater likelihood of malaria outbreaks compared to dengue**.

- However, the small coefficient indicates that **while statistically significant, the actual effect size is quite small**.
- To strengthen the model, **consider adding more environmental or epidemiological factors** (e.g., humidity, mosquito breeding conditions).

The following plot shows the linearity of the logit for *lst_day_celsius*, verifying the assumption of a linear relationship between this variable and the log-odds in the logistic regression model.



Interpretation of Logit Linearity Plot for lst_day_celsius

1. **Purpose of the Plot**
    - This plot checks the assumption of **log-linearity** in logistic regression.
    - It visualizes how lst_day_celsius relates to the **log-odds** (logit of predicted probabilities).
2. **Observations**
    - The scatterplot **does not show a clear linear pattern**.
    - The log-odds appear to **slightly decrease** as temperature (lst_day_celsius) increases, but the relationship is weak.
3. **Implications**
    - Since the relationship is **not strongly linear**, lst_day_celsius might **not be a good predictor** of malaria outbreaks.

- The variable could be **transformed (e.g., polynomial terms, interaction effects) or reconsidered** in the model.

Based on the comprehensive visualizations and insights derived from the data, the final recommendations addressing the given problem are as follows:

Recommendations:

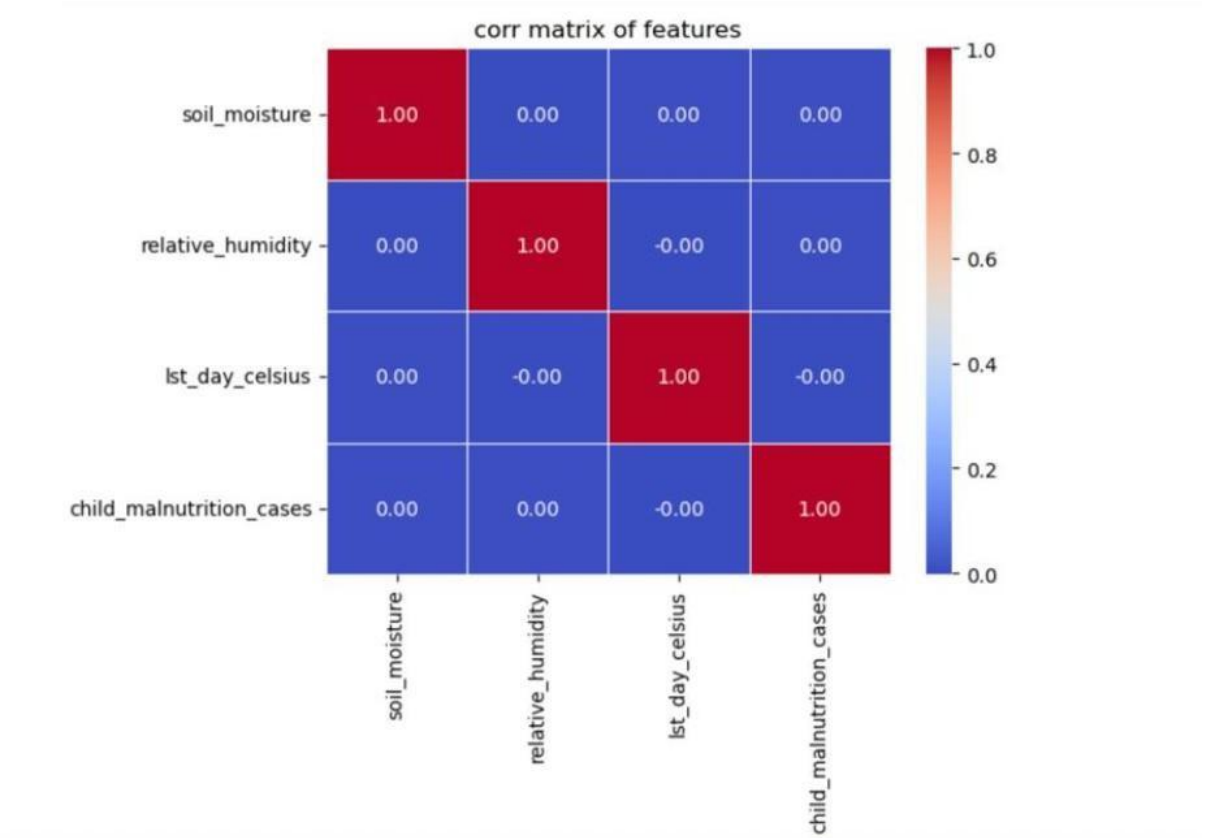
1. **Targeted Disease Control:** Focus malaria prevention in high-rainfall regions and implement widespread dengue control by eliminating breeding grounds.
2. **Temperature-Based Surveillance:** Establish risk prediction models using temperature to anticipate and respond to future outbreaks.
3. **Integrated Vector Management**: Use a combined approach (e.g., larvicides, bed nets, and public awareness) for more effective vector control.
4. **Model Enhancement**: Incorporate additional variables (e.g., humidity and socio-economic factors) to improve the accuracy of outbreak predictions.

TASK 4

Child Malnutrition with soil moisture, relative humidity and temperature

Objective: Examine how soil moisture, relative humidity, and temperature influence child malnutrition**.**

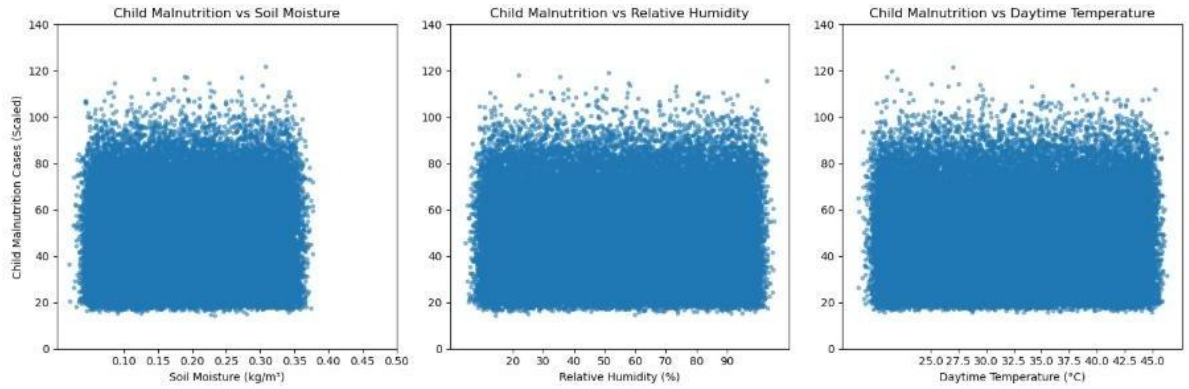
First, we will start by finding the **correlation coefficient** between the number of malnutrition cases, soil moisture, relative humidity, and temperature to check the dependency of malnutrition cases on these three factors.



By observing the Heatmap we can draw the following insights:

- The correlation between **child_malnutrition_cases** and **soil_moisture, relative_humidity, and lst_day_celsius** is **0.00** or very close to 0.
- This indicates **no linear relationship** between child malnutrition cases and these environmental factors.

These conclusions are again solidified by the following scatter plots, which indicate that there is **no clear increasing or decreasing trend**, suggesting that soil moisture, relative humidity, and temperature may not have a strong linear correlation with child malnutrition cases.



Since it is difficult to determine the relationship between variables using a simple heatmap or scatter plot, we now create **a Linear Regression Model** to better understand the relationships and dependencies of these factors.

The results of the Regression Model are:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     child_malnutrition_cases   R-squared:                       0.000
Model:                                 OLS   Adj. R-squared:                 -0.000
Method:                      Least Squares   F-statistic:                     0.1959
Date:                     Sun, 16 Mar 2025   Prob (F-statistic):              0.899
Time:                             15:18:59   Log-Likelihood:            -1.5078e+05
No. Observations:                    45040   AIC:                         3.016e+05
Df Residuals:                        45036   BIC:                         3.016e+05
Df Model:                                3
Covariance Type:                 nonrobust
==============================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const                 10.5488      0.181     58.281      0.000      10.194      10.904
soil_moisture          0.1568      0.374      0.419      0.675      -0.577       0.891
relative_humidity      0.0005      0.001      0.388      0.698      -0.002       0.003
lst_day_celsius       -0.0023      0.005     -0.513      0.608      -0.011       0.007
==============================================================================
Omnibus:                     2631.453   Durbin-Watson:                   1.981
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1738.177
Skew:                           0.364   Prob(JB):                         0.00
Kurtosis:                       2.369   Cond. No.                         802.
==============================================================================
```

1) **R-squared and Adjusted R-squared (0.000 and -0.000)**

- The R-squared value is essentially 0, indicating that the independent variables (soil moisture, relative humidity, and daytime temperature) explain **almost none** of the variation in child malnutrition cases.
- The Adjusted R-squared being negative confirms that adding these variables does not improve the model's explanatory power.

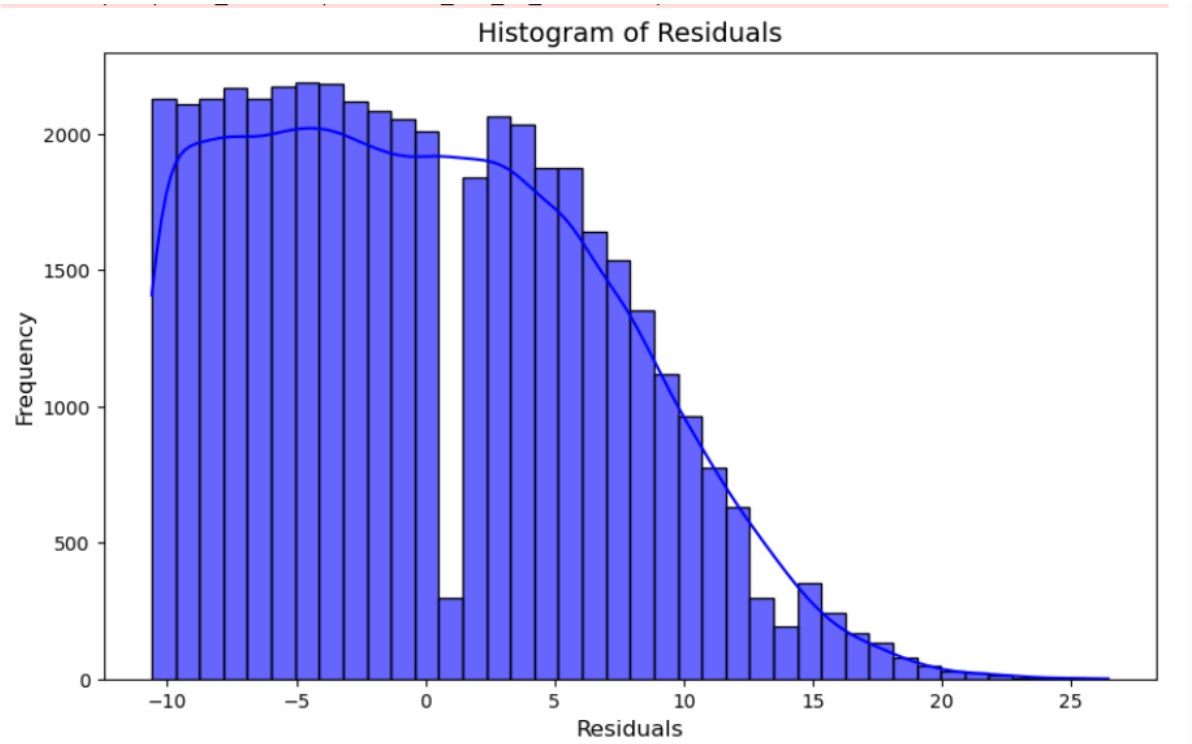
2) **P-values of Independent Variables**

- **Soil Moisture (P = 0.675)**, **Relative Humidity (P = 0.698)**, and **Daytime Temperature (P = 0.608)** all have **high p-values (greater than 0.05)**, indicating that none of these variables are statistically significant predictors of child malnutrition cases.
- This suggests that soil moisture, humidity, and temperature **do not have a meaningful linear relationship** with child malnutrition cases.

3) **F-statistic and Prob (F-statistic) (0.1959 and 0.899)**

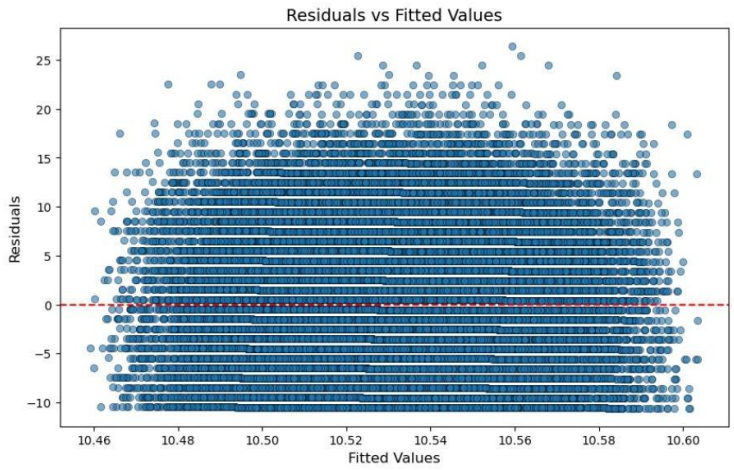
- A low F-statistic and a high p-value (0.899) indicate that the regression model as a whole is **not statistically significant**.
- This means that, collectively, the independent variables **fail to predict** malnutrition cases.

To further analyze this model, we will plot **a histogram of residuals against frequency** to gain a deeper understanding of its behaviour. The plot and insights are:



- **Non-Normal Distribution:** This histogram shows an asymmetric distribution with a heavy tail towards the right, indicating skewness.
- **Presence of Gaps:** There is a noticeable gap in the middle of the distribution, which is unusual. This could suggest missing values, data binning issues, or an inherent structure in the dataset that needs further investigation.
- **Heteroscedasticity Concerns:** If residuals exhibit an increasing spread as values increase, it indicates heteroscedasticity (variance of errors is not constant). This could mean that the model is not capturing some nonlinear relationships in the data.

To further evaluate the model's performance, we will visualize the relationship between residuals and fitted values using a scatter plot.



- No Clear Pattern → Suggests No Strong Non-Linearity
- High Variability in Residuals → Possible Heteroscedasticity
- **Presence of Outliers:** There are residuals far above 20 and below -10, which might indicate outliers or influential data points.

To reinforce the findings from the regression model and gain deeper insights into the data, we can perform an **ANOVA (Analysis of Variance)** test on the regression model. The results of the ANOVA test are as follows:

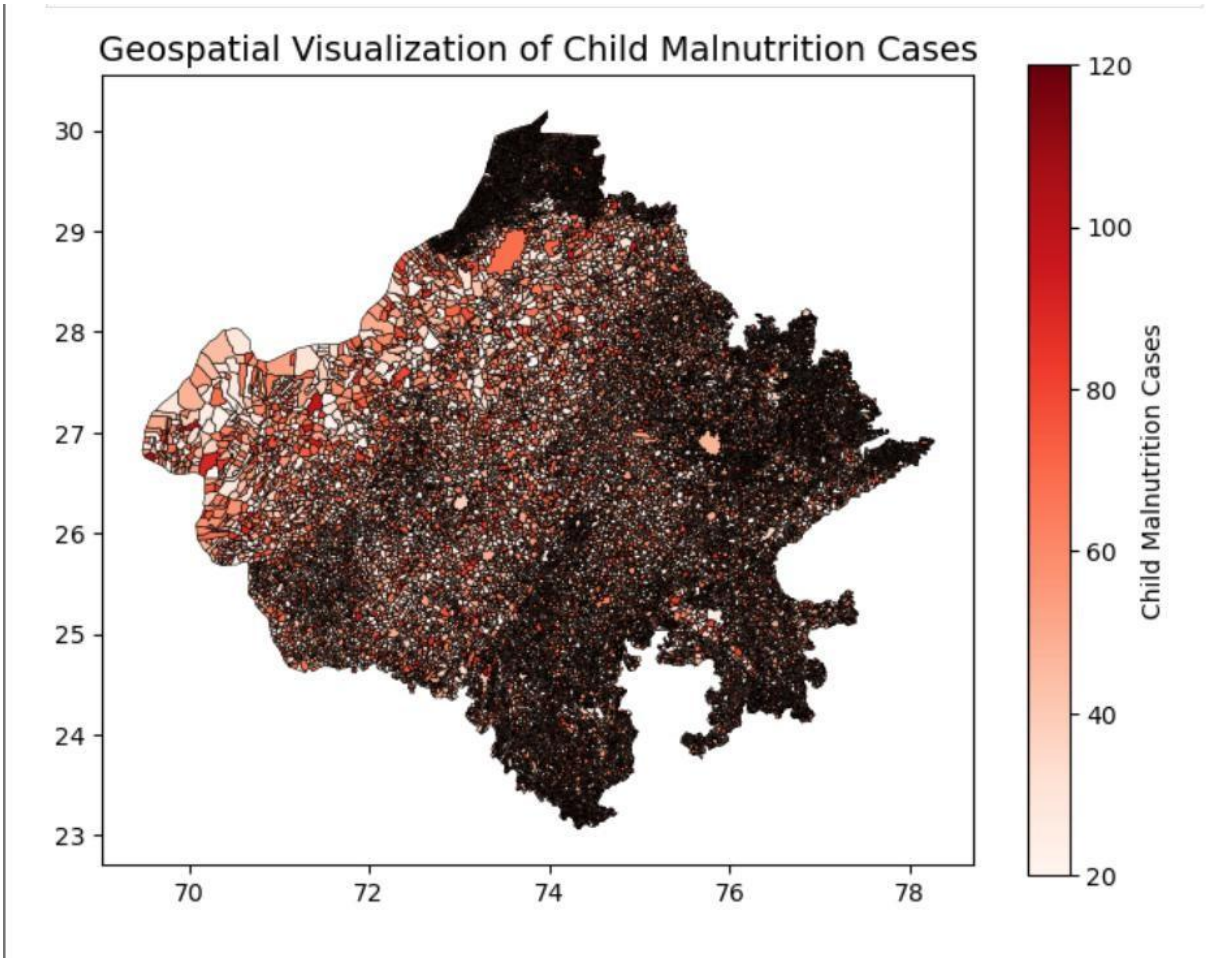
```
# Example: Suppose df has columns 'child_malnutrition_cases', 'soil_moisture', 'relative_hu
formula = "child_malnutrition_cases ~ soil_moisture + relative_humidity + lst_day_celsius"
model = smf.ols(formula=formula, data=df).fit()  # Fit the model

# Now perform ANOVA
anova_results = sm.stats.anova_lm(model, typ=2)
print(anova_results)
```

```
                       sum_sq       df         F    PR(>F)
soil_moisture      8.306148e+00      1.0  0.175395  0.675363
relative_humidity  7.111941e+00      1.0  0.150178  0.698367
lst_day_celsius    1.246205e+01      1.0  0.263152  0.607965
Residual           2.132761e+06  45036.0       NaN       NaN
```

The insights that we can draw from this data are:

- **Sum of Squares (sum_sq):** The residual sum of squares (2.13 million) is much larger than the sum of squares for soil moisture (8.3), relative humidity (7.1), and temperature (1.2), indicating that most of the variance is unexplained by these variables.
- **F-statistic (F):** The F-values for **soil moisture (0.175), relative humidity (0.150), and temperature (0.263)** are **very low**, indicating that these variables explain very little variance in child malnutrition cases.
- **p-values (PR(>F)**): All p-values are much greater than 0.05, confirming that none of these variables are statistically significant in explaining child malnutrition cases.

To understand the spatial distribution of child malnutrition, we analyze geospatial data to identify high-risk regions and potential intervention areas.



- **Geographical Variation in Malnutrition Cases:** The map highlights varying levels of child malnutrition across different regions.
- **High-Risk Zones:** Clusters of dark red indicate areas with high malnutrition prevalence (above 100 cases). These regions may require **targeted interventions**, including nutrition programs, healthcare services, and awareness campaigns.
- **Spatial Distribution and Policy Implications:** The map suggests a need for localized policies rather than a one-size-fits-all approach. **Government programs** should focus more on darker regions while maintaining preventive measures in lighter areas.

Drawing insights from the visualizations analyzed, we derive the following conclusions and propose strategic recommendations to address the identified patterns.

Recommendations:

1:- Feature Engineering: Instead of using raw values of soil moisture, temperature, and humidity, try creating new features (e.g., **temperature anomalies, seasonal variations, or interaction terms**) to uncover hidden patterns.

2: Re-evaluate the factors influencing malnutrition: Since soil moisture, relative humidity, and temperature do not significantly impact malnutrition cases, other factors should be explored, such as socioeconomic status, food availability, healthcare access, and sanitation.

Conclusion:

The linear regression model confirms that **soil moisture, humidity, and temperature do not significantly influence child malnutrition cases**. This suggests that **other factors** should be investigated, and alternative modeling approaches (such as non-linear regression or machine learning models) may be more suitable for understanding the drivers of malnutrition.

Based on the comprehensive visualizations and insights derived from the data, the final recommendations addressing the given problem are as follows:

Recommendations:

1. **Explore Alternative Factors:** Investigate socio-economic conditions (e.g., food access, sanitation) as stronger predictors of malnutrition rather than environmental variables.
2. **Feature Engineering:** Develop advanced features such as seasonal variations and interaction terms to reveal complex relationships.
3. **Localized Policy Design:** Implement region-specific nutrition programs focusing on high-risk zones while maintaining preventive measures in low-risk areas.
4. **Model Enhancement:** Consider advanced modeling approaches (e.g., non-linear models or machine learning) for better predictive accuracy.

TASK 5

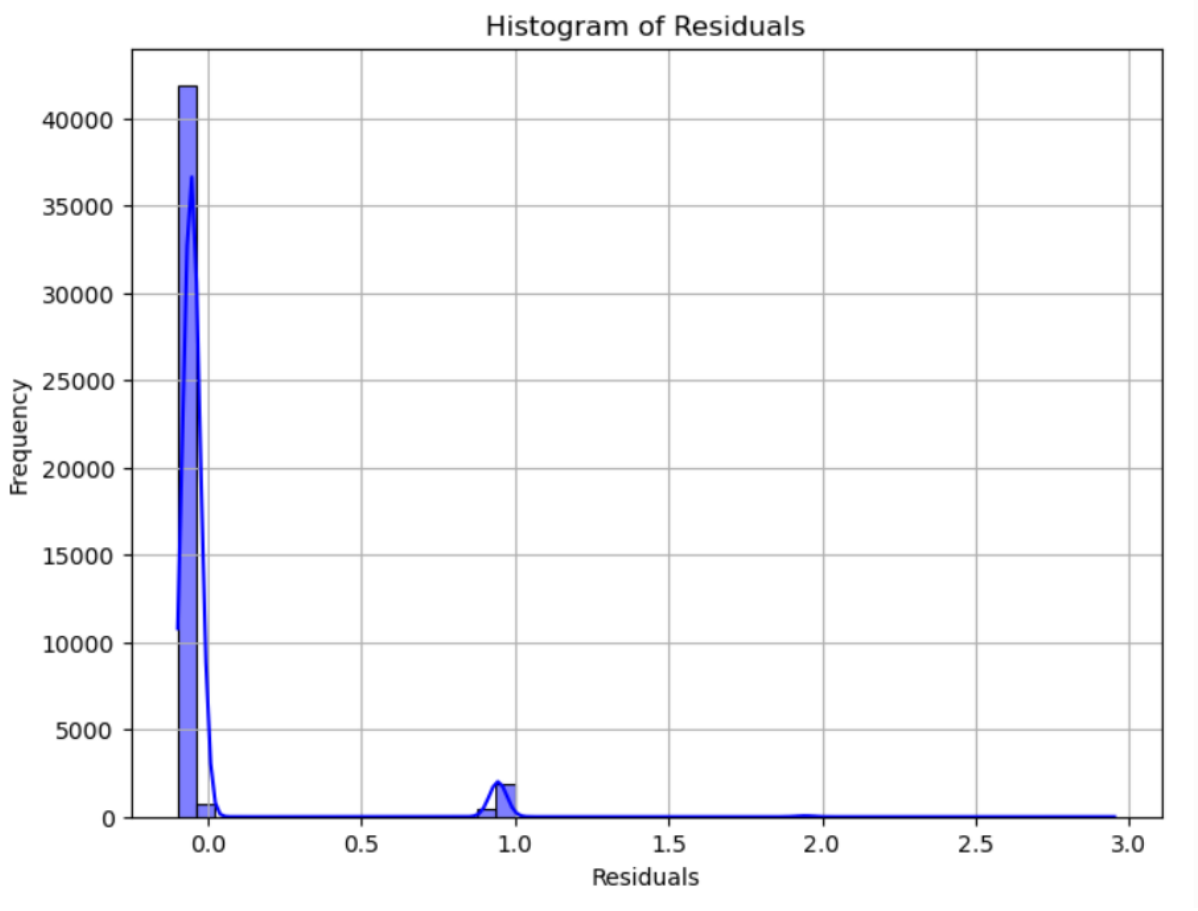
Tuberculosis and Population Density

Objective: Evaluate how population density and rural facilities affect TB Cases.

To find the relationship between rural facilities and population density with TB cases, we first create a **Negative Binomial regression model**. This model is well-suited for the analysis because TB cases are count data, which are typically non-negative and discrete.

The initial step in studying the relationship between variables using a Negative Binomial Regression model is to examine the ***Histogram of Residuals versus Frequency*.** Residuals represent the difference between the actual observed values and the values predicted by the model. Analyzing the residuals helps us identify how well the model fits the data and whether any systematic patterns or biases remain unaccounted for.

The insights from this analysis are:



1) **Bimodal Distribution (Multiple Peaks):**

- The histogram shows **at least two clear peaks**—one close to zero and another around **1.0**, suggesting the residuals are **not normally distributed**.
- **Interpretation:** This may indicate the presence of **distinct sub-groups** in the data, meaning the model may not capture differences across **different regions or conditions**.

2) **Clustered Residuals (Model Misspecification):**

- The **second peak** suggests that a significant portion of the data has **systematic underestimation**, meaning the model **consistently under-predicts** TB cases in certain areas.
- **Insight:** There may be **missing variables** (e.g., environmental factors, healthcare access) that are influencing TB cases but are **not included** in the model.

3) O**utliers and Extreme Residuals:**

- There are a few residuals beyond **2.0**, indicating the model **fails** to predict TB cases accurately in **some regions**.
- **Insight:** Investigate these **outlier** areas to identify **unusual conditions** contributing to higher or lower TB incidence.

The results we obtained from the model are:

[86]:

Generalized Linear Model Regression Results

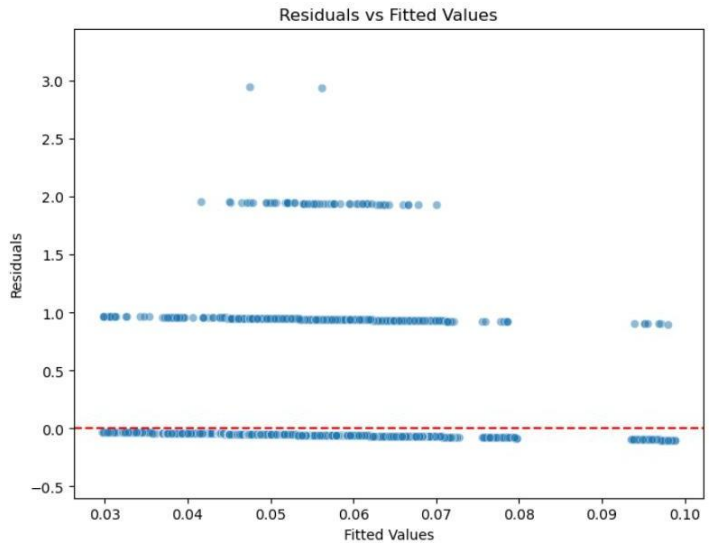
Dep. Variable:	tb_cases	**No. Observations:**	45040
Model:	GLM	**Df Residuals:**	45036
Model Family:	NegativeBinomial	**Df Model:**	3
Link Function:	Log	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-9658.5
Date:	Sun, 16 Mar 2025	**Deviance:**	12611.
Time:	15:22:14	**Pearson chi2:**	4.27e+04
No. Iterations:	6	**Pseudo R-squ. (CS):**	0.0008736
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.8385	0.185	-20.773	0.000	-4.201	-3.476
population_density	0.1900	0.031	6.149	0.000	0.129	0.251
rural_facilities	-0.0002	0.000	-0.598	0.550	-0.001	0.001
co_density	-0.1387	0.180	-0.772	0.440	-0.491	0.213

The insights that we can draw from this Model summary are:

- Since the value of population density coefficient is 0.1900 and p value is 0, this indicates **that an increase in population density is associated with an increase in TB cases**. Specifically, for each unit increase in population density, the expected log count of TB cases increases by 0.1900, holding other variables constant.
- The rural facilities coefficient and Co-density coefficient both have negative values signifying , that both these factors does not show a significant effect on TB Cases.
- The **Pseudo R-squared (CS)** value is 0.00008736 while the **Log-Likelihood** value is 9658.5. This suggests that the model explains very little of the variability in TB cases.

A deeper evaluation of the model's performance is conducted through the residuals vs. fitted values plot, which helps assess the adequacy of the model, detect potential patterns, and identify areas for improvement.



Insights from Residuals vs Fitted Values Plot

1. **Pattern in Residuals (Non-Randomness)**
    - The residuals appear in discrete bands rather than being randomly dispersed. This could indicate an issue with the model assumptions, such as overdispersion in the Negative Binomial model.
2. **Heteroscedasticity (Unequal Spread of Residuals)**
    - The spread of residuals does not appear uniform across fitted values. This suggests that variance is not constant, which could imply the need for further adjustments, such as including additional predictor variables or using a different model formulation.
3. **Presence of Outliers**
    - A few residuals are noticeably larger, especially above **2**. These points might be outliers, suggesting regions with unusually high or low TB cases that are not well explained by the model.

Based on the comprehensive visualizations and insights derived from the data, the final recommendations addressing the given problem are as follows:

Recommendations:

1. **Address Overdispersion:** Explore advanced models (e.g., Zero-Inflated Negative Binomial) to account for heterogeneous data patterns.
2. **Expand Predictors:** Include additional socio-economic and spatial variables (e.g., healthcare quality and literacy rates) for better model performance.

3. **Investigate Outliers:** Examine regions with extreme TB cases to identify unique causes (e.g., inadequate healthcare or environmental risks).
4. **Refine Surveillance:** Strengthen data collection and monitoring systems to capture the true variability of TB incidence and improve future modelling.

Report prepared by:

Shalvi Singh