KLE Technological University
Hubballi

A Course Project Report on

“Data-driven Impact Analysis on Dengue Fever with Climate Factors ”

A Course Project Report Submitted in Partial Fulfillment of the Requirement for the Course of

Exploratory Data Analysis

in

4th Semester of Computer Science and Engineering

by

KAVYA K MORAB	02FE21BCS039
HARSHVARDHAN PANDEY	02FE21BCS032
NEHA B SHELLIKERI	02FE21BCS051
KAVERI B HANABAR	02FE21BCS038

Under the guidance of

Dr. Santosh Pattar

Assistant Professor,
Department of Computer Science and Engineering,
KLE Technological University’s Dr. MSSCET, Belagavi.

KLE Technological University’s
Dr. M. S. Sheshgiri College of Engineering and Technology,
Belagavi – 590 008.

July 2023

DECLARATION

We hereby declare that the matter embodied in this report entitled "**Data-driven Impact Analysis on Dengue Fever with Climate Factors**" submitted to KLE Technological University for the course completion of Exploratory Data Analysis (21ECSC210) in the 4th Semester of Computer Science and Engineering is the result of the work done by us in the Department of Computer Science and Engineering, KLE Dr. M. S. Sheshgiri College of Engineering, Belagavi under the guidance of Dr. Santosh Pattar, Assistant Professor, Department of Computer Science and Engineering. We further declare that to the best of our knowledge and belief, the work reported here in doesn't form part of any other project on the basis of which a course or award was conferred on an earlier occasion on this by any other student(s), also the results of the work are not submitted for the award of any course, degree or diploma within this or in any other University or Institute. We hereby also confirm that all of the experimental work in this report has been done by us.

Belagavi – 590 008
Date : 22-July-2023

Kavya K Morab (02FE21BCS039)	Harshvardhan Pandey (02FE21BCS032)
Neha B Shellikeri (02FE21BCS051)	Kaveri B Hanabar (02FE21BCS038)

KLE TECH KLE
TECHNOLOGICAL UNIVERSITY
Creating Value, Leveraging Knowledge
Belagavi Campus

Dr.M.S.Sheshgiri College of Engineering & Technology
Department of Computer Science & Engineering

CERTIFICATE

This is to certify that the project entitled “Data-driven Impact Analysis on Dengue Fever with Climate Factors” submitted to KLE Technological University’s Dr. MSSCET, Belagavi for the partial fulfillment of the requirement for the course - Exploratory Data Analysis (21ECSC210) by Kavya K Morab(02FE21BCS039), Harshvardhan Pandey(02FE21BCS032),Neha B Shellikeri(02FE21BCS051) and Kaveri B Hanabar(02FE21BCS038), students in the Department of Computer Science and Engineering, KLE Technological University’s Dr. MSSCET, Belagavi, is a bonafide record of the work carried out by them under my supervision. The contents of this report, in full or in parts, have not been submitted to any other Institute or University for the award of any other course completion.

Belagavi – 590 008
Date : 22-July-2023

Dr. Santosh Pattar	Prof. Priyanka Gavade
(Course Teacher)	(Course Coordinator)

Dr. Rajashri Khanai
(Head of the Department)

Abstract

Problem Statement:

In this project, we aim to provide the impact of dengue disease based on environmental variables describing changes in temperature, precipitation, vegetation, and more through data-driven insights (i.e., EDA), so as to enhance safety measures.

Solution to Problem Statement:

1.Collection: Gather historical data on dengue cases, environmental variables (temperature, precipitation, vegetation), and any other relevant data sources. This could include public health records, meteorological data, satellite imagery, and other relevant sources.

2.Data Preprocessing: Clean and preprocess the collected data. This involves handling missing values, normalizing variables, and resolving any inconsistencies or errors in the data. Ensure the data is in a format suitable for analysis.

3.Exploratory Data Analysis (EDA): Conduct a comprehensive EDA on the collected data. Analyze the relationships between dengue cases and environmental variables. Explore patterns, correlations, and trends in the data. Visualize the data using graphs, charts, and maps to gain insights.

4.Feature Selection: Identify the most influential features (environmental variables) that are strongly correlated with dengue cases. Use statistical techniques or machine learning algorithms to assess the importance of each variable. Select the relevant features for your predictive model.

5.Predictive Modeling: Build a predictive model using the selected features and historical dengue data. You can use various machine learning algorithms such as decision trees, random forests, or gradient boosting algorithms. Split the data into training and testing sets for model validation.

Results:

1. Correlation with Environmental Variables

2. Effect of Temperature on Dengue Transmission
3. Effect of Precipitation on Dengue Incidence
4. Role of Vegetation in Dengue Spread
5. Temporal Trends and Patterns
6. Identification of High-Risk Areas
7. Understanding Diurnal Temperature Range

Contents

List of Figures

List of Tables

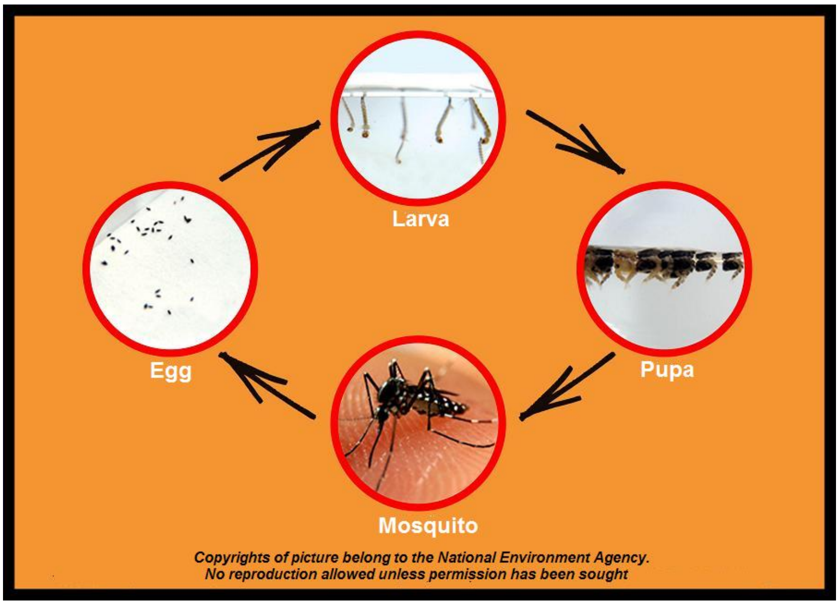
Chapter 1

Introduction

1.1 Background

• Dengue fever is a mosquito-borne disease that occurs in tropical and sub-tropical parts of the world. Dengue viruses are spread to people through the bites of infected Aedes species mosquitoes.

• In mild cases, symptoms are similar to the flu: fever, rash, and muscle and joint pain. In severe cases, dengue fever can cause severe bleeding, low blood pressure, and even death.



Copyrights of picture belong to the National Environment Agency.
No reproduction allowed unless permission has been sought

FIGURE 1.1: Dengue life cycle

- Because it is carried by mosquitoes, the transmission dynamics of dengue are related to climate variables such as temperature and precipitation.

- In recent years, historically the disease has been most prevalent in Southeast Asia and the Pacific islands. These days many of the nearly half billion cases per year are occurring in Latin America.

1.2 Problem Statement

In this project, we aim to provide the impact of dengue disease based on environmental variables describing changes in temperature, precipitation, vegetation, and more through data-driven insights (i.e., EDA), so as to enhance safety measures

1.2.1 Objectives

1. Quantify the Impact of Dengue Disease.

2. Identify Relevant Environmental Variables.

3. Develop Predictive Models.

4. Enhance Safety Measures and Preparedness.

Chapter 2

Knowing the Dataset

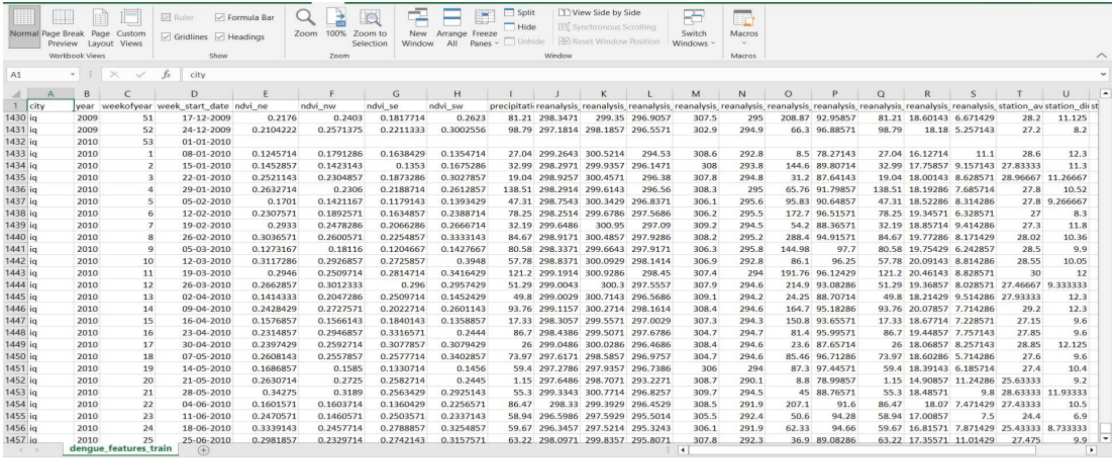
2.1 About the Dataset

The provided dataset is a time-series dataset containing information related to dengue disease and various environmental variables for two cities: San Juan (sj) and Iquitos (iq). Each row represents data for a specific week, and the data spans multiple years. The dataset includes columns such as maximum temperature, minimum temperature, average temperature, total precipitation, diurnal temperature range, dew point temperature, relative humidity, specific humidity, and normalized difference vegetation index (NDVI) values for different pixels around the city centroids. Additionally, there are reanalysis data columns that provide alternative measurements for certain environmental variables. The dataset's focus appears to be on understanding the impact of environmental factors on dengue outbreaks, as it includes weather-related information and NDVI, which can serve as proxies for vegetation and ecosystem health. The dataset's time-series nature enables the analysis of temporal patterns and trends, potentially valuable for predictive modeling and public health interventions to mitigate the effects of dengue outbreaks in these cities.

2.2 Content of the Dataset

The dataset contains information on dengue disease and various environmental variables for two cities, San Juan (sj) and Iquitos (iq). Each row represents data

[?]



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	
1	city	year	weekofyear	week_start_date	ndvi_ne	ndvi_nw	ndvi_se	ndvi_sw	precipitati	reanalysis_	reanalysis_	reanalysis_	reanalysis_	reanalysis_	reanalysis_	reanalysis_	reanalysis_	reanalysis_	reanalysis_	station_av	station_di	st
1430	iq	2009	51	17-12-2009	0.2176	0.2403	0.1817714	0.2623	81.21	298.3471	299.35	296.9057	307.5	295	208.87	92.95857	81.21	18.60143	6.671429	28.2	11.125	
1431	iq	2009	52	24-12-2009	0.2104222	0.2571375	0.2211333	0.3002556	98.79	297.1814	298.1857	296.5571	302.9	294.9	66.3	96.88571	98.79	18.18	5.257143	27.2	8.2	
1432	iq	2010	53	01-01-2010																		
1433	iq	2010	1	08-01-2010	0.1245714	0.1791286	0.1638429	0.1354714	27.04	299.2643	300.5214	294.53	308.6	292.8	8.5	78.27143	27.04	16.12714	11.1	28.6	12.3	
1434	iq	2010	2	15-01-2010	0.1452857	0.1423143	0.1353	0.1675286	32.99	298.2971	299.9357	296.1471	308	293.8	144.6	89.80714	32.99	17.75857	9.157143	27.83333	11.3	
1435	iq	2010	3	22-01-2010	0.2521143	0.2304857	0.1873286	0.3027857	19.04	298.9257	300.4571	296.38	307.8	294.8	31.2	87.64143	19.04	18.00143	8.628571	28.96667	11.26667	
1436	iq	2010	4	29-01-2010	0.2632714	0.2306	0.2188714	0.2612857	138.51	298.2914	299.6143	296.56	308.3	295	65.76	91.79857	138.51	18.19286	7.685714	27.8	10.52	
1437	iq	2010	5	05-02-2010	0.1701	0.1421167	0.1179143	0.1393429	47.31	298.7543	300.3429	296.8371	306.1	295.6	95.83	90.64857	47.31	18.52286	8.314286	27.8	9.266667	
1438	iq	2010	6	12-02-2010	0.2307571	0.1892571	0.1634857	0.2388714	78.25	298.2514	299.6786	297.5686	306.2	295.5	172.7	96.51571	78.25	19.34571	6.328571	27	8.3	
1439	iq	2010	7	19-02-2010	0.2933	0.2478286	0.2066286	0.2666714	32.19	299.6486	300.95	297.09	309.2	294.5	54.2	88.36571	32.19	18.85714	9.414286	27.3	11.8	
1440	iq	2010	8	26-02-2010	0.3036571	0.2600571	0.2254857	0.3333143	84.67	298.9171	300.4857	297.9286	308.2	295.2	288.4	94.91571	84.67	19.77286	8.171429	28.02	10.36	
1441	iq	2010	9	05-03-2010	0.1273167	0.18116	0.1204667	0.1427667	80.58	298.3371	299.6643	297.9171	306.3	295.8	144.98	97.7	80.58	19.75429	6.242857	28.5	9.9	
1442	iq	2010	10	12-03-2010	0.3117286	0.2926857	0.2725857	0.3948	57.78	298.8371	300.0929	298.1414	306.9	292.8	86.1	96.25	57.78	20.09143	8.814286	28.55	10.05	
1443	iq	2010	11	19-03-2010	0.2946	0.2509714	0.2814714	0.3416429	121.2	299.1914	300.9286	298.45	307.4	294	191.76	96.12429	121.2	20.46143	8.828571	30	12	
1444	iq	2010	12	26-03-2010	0.2662857	0.3012333	0.296	0.2957429	51.29	299.0043	300.3	297.5557	307.9	294.6	214.9	93.08286	51.29	19.36857	8.028571	27.46667	9.333333	
1445	iq	2010	13	02-04-2010	0.1414333	0.2047286	0.2509714	0.1452429	49.8	299.0029	300.7143	296.5686	309.1	294.2	24.25	88.70714	49.8	18.21429	9.514286	27.93333	12.3	
1446	iq	2010	14	09-04-2010	0.2428429	0.2727571	0.2022714	0.2601143	93.76	299.1157	300.2714	298.1614	308.4	294.6	164.7	95.18286	93.76	20.07857	7.714286	29.2	12.3	
1447	iq	2010	15	16-04-2010	0.1576857	0.1566143	0.1840143	0.1358857	17.33	298.3057	299.5571	297.0029	307.3	294.3	150.8	93.65571	17.33	18.67714	7.228571	27.15	9.6	
1448	iq	2010	16	23-04-2010	0.2314857	0.2946857	0.3316571	0.2444	86.7	298.4386	299.5071	297.6786	304.7	294.7	81.4	95.99571	86.7	19.44857	7.757143	27.85	9.6	
1449	iq	2010	17	30-04-2010	0.2397429	0.2592714	0.3077857	0.3079429	26	299.0486	300.0286	296.4686	308.4	294.6	23.6	87.65714	26	18.06857	8.257143	28.85	12.125	
1450	iq	2010	18	07-05-2010	0.2608143	0.2557857	0.2577714	0.3402857	73.97	297.6171	298.5857	296.9757	304.7	294.6	85.46	96.71286	73.97	18.60286	5.714286	27.6	9.6	
1451	iq	2010	19	14-05-2010	0.1686857	0.1585	0.1330714	0.1456	59.4	297.2786	297.9357	296.7386	306	294	87.3	97.44571	59.4	18.39143	6.185714	27.4	10.4	
1452	iq	2010	20	21-05-2010	0.2630714	0.2725	0.2582714	0.2445	1.15	297.6486	298.7071	293.2271	308.7	290.1	8.8	78.99857	1.15	14.90857	11.24286	25.63333	9.2	
1453	iq	2010	21	28-05-2010	0.34275	0.3189	0.2563429	0.2925143	55.3	299.3343	300.7714	296.8257	309.7	294.5	45	88.76571	55.3	18.48571	9.8	28.63333	11.93333	
1454	iq	2010	22	04-06-2010	0.1601571	0.1603714	0.1360429	0.2256571	86.47	298.33	299.3929	296.4529	308.5	291.9	207.1	91.6	86.47	18.07	7.471429	27.43333	10.5	
1455	iq	2010	23	11-06-2010	0.2470571	0.1460571	0.2503571	0.2337143	58.94	296.5986	297.5929	295.5014	305.5	292.4	50.6	94.28	58.94	17.00857	7.5	24.4	6.9	
1456	iq	2010	24	18-06-2010	0.3339143	0.2457714	0.2788857	0.3254857	59.67	296.3457	297.5214	295.3243	306.1	291.9	62.33	94.66	59.67	16.81571	7.871429	25.43333	8.733333	
1457	iq	2010	25	25-06-2010	0.2981857	0.2329714	0.2742143	0.3157571	63.22	298.0971	299.8357	295.8071	307.8	292.3	36.9	89.08286	63.22	17.35571	11.01429	27.475	9.9	

Figure 2.1: Snapshot of Dataset

for a specific week, with details such as maximum temperature, minimum temperature, average temperature, total precipitation, diurnal temperature range, dew point temperature, relative humidity, specific humidity, and normalized difference vegetation index (NDVI) values for different pixels around the city centroids. It also includes reanalysis data for certain environmental variables.

Table 2.1: Properties of the Dataset

Characteristics	**Values**
Name	Data-driven Impact Analysis on Dengue Fever with Climate Factors
Type	Time series dataset
Source	**https://rb.gy/qbo9a**
Number of instances	1456
Number of Features	25
Dataset Format	Comma Separated Values (CSV) Format

2.3 Features of the Dataset

There are a total of 25 features in this dataset. They are described as follows in Table 2.2. In Table 2.3, the details of the features are

summarized. For each feature, we list the type of the data, number of unique and missing values.

2.4 Data Distribution of the Features

In this section, we visualize the distributions of the different features of the dataset(as shown in Figure 2.2).

Table 2.2: Details of the Features in the Dataset.

Feature Name	Data Type	Distinct Values	Missing Values
city	Categorical	2	0
year	Interval	21	0
weekofyear	Interval	53	0
week start date	Nominal	1049	0
ndvi ne	Numeric	1214	194
ndvi nw	Numeric	1365	52
ndvi se	Numeric	1395	22
ndvi sw	Numeric	1388	22
precipitation amt mm	Numeric	1157	13
reanaysis air temp k	Numeric	1176	10
reanalysis avg temp k	Numeric	600	10
reanalysis dew point temp k	Numeric	1180	10
reanalysis max air temp k	Numeric	141	10
reanalysis min air temp k	Numeric	117	10
reanalysis precip amt kg per m2	Numeric	1039	10
reanalysis relative humidity percent	Numeric	1370	10
reanalysis sat precip amt mm	Numeric	1157	13
reanalysis precip humidity g per kg	Numeric	1171	10
reanalysis tdtr k	Numeric	519	10
station avg temp c	Numeric	492	43
station durinal temp rng c	Numeric	470	43
station max temp c	Numeric	73	20
station min temp c	Numeric	70	14
station precip mm	Numeric	663	22

Table 2.3: Description of the Features in the Dataset.

Feature Name	**Description**
city	City abbreviations: sj for San Juan and iq for Iquitos
week start date	Date given in yyyy-mm-dd format
station max temp c	Maximum temperature in Celsius
station min temp c	Minimum temperature in Celsius
station avg temp c	Average temperature in Celsius
station precip mm	Total precipitation in mm
station diurnal temp range c	Diurnal temperature range in Celsius
precipitation amt mm	Total precipitation in mm
reanalysis stat precip amt mm	Total precipitation in mm
reanalysis dew point temp k	Mean dew point temperature in K
reanalysis air temp k	Mean air temperature in K
reanalysis relative humidity percent	Mean relative humidity
reanalysis specific humidity g per kg	Mean specific humidity in g/kg
reanalysis precip amt kg per m2	Total precipitation in kg/m^2
reanalysis max air temp k	Maximum air temperature in K
reanalysis min air temp k	Minimum air temperature in K
reanalysis avg temp	Average air temperature in K
reanalysis tdtr k	Diurnal temperature range in K
ndvi se	Pixel southeast of city centroid
ndvi sw	Pixel southwest of city centroid
ndvi ne	Pixel northeast of city centroid
ndvi nw	Pixel northwest of city centroid
year	Year
week of the year	week of the year

2.5 Observations

List your observations from the dataset here.

- How are the features? All categorical? Mix?

  Features are mixed.

- Are there any missing values? If yes, are they large or small?

  Yes,large

- What is the range of data items? How are they distributed?

  It depends on features of the Dataset

- Are there any outliers?

  Yes

- Are any of the features skewed?

  Yes

- Does any of the features require normalization, scaling?

  Yes

- Overall what are the characteristics of your dataset?

Environmental variables

Missing values

Temperature range

Precipitation range

NDVI values

City wise data

2.6 Statistical Data Analysis

The mean,maximum,minmum,standard deviation and qurtiles of all features are given below:

1. **ndvi ne:**

   The mean of ndvi ne is 0.1422935374167987

   The max of ndvi ne is 0.5083571

   The min of ndvi ne is -0.40625

   The standard deviation of ndvi ne is 0.14053115314639272

The 25th percentile of ndvi ne is 0.044950000000000004

The 50th percentile of ndvi ne is 0.12881665

The 75th percentile of ndvi ne is 0.248483325

2. **ndvi nw:**

   The mean of ndvi nw is 0.13055257610470083

   The max of ndvi nw is 0.4544286

   The min of ndvi nw is -0.4561

   The standard deviation of ndvi nw is 0.11999906267761298

   The 25th percentile of ndvi nw is 0.0492166675

   The 50th percentile of ndvi nw is 0.1214286

   The 75th percentile of ndvi nw is 0.2166

3. **ndvi se:**

   The mean of ndvi se is 0.20378318902580195

   The max of ndvi se is 0.5383143

   The min of ndvi se is -0.01553333

   The standard deviation of ndvi se is 0.07385973904332568

   The 25th percentile of ndvi se is 0.155087475

   The 50th percentile of ndvi se is 0.19605

   The 75th percentile of ndvi se is 0.24884582500000002

4. **ndvi sw:**

   The mean of ndvi sw is 0.20230549071129708

   The max of ndvi sw is 0.5460167

   The min of ndvi sw is -0.06345714

   The standard deviation of ndvi sw is 0.08390267778770004

The 25th percentile of ndvi sw is 0.144208725

The 50th percentile of ndvi sw is 0.18945

The 75th percentile of ndvi sw is 0.24698215

5. **precipitation amt mm:**

   The mean of precipitation amt mm is 45.760388080388076

   The max of precipitation amt mm is 390.6

   The min of precipitation amt mm is 0.0

   The standard deviation of precipitation amt mm is 43.71553699164491

   The 25th percentile of precipitation amt mm is 9.8

   The 50th percentile of precipitation amt mm is 38.34

   The 75th percentile of precipitation amt mm is 70.235

6. **reanalysis air temp k:**

   The mean of reanalysis air temp k is 298.7018524007227

   The max of reanalysis air temp k is 302.2

   The min of reanalysis air temp k is 294.635714286

   The standard deviation of reanalysis air temp k is 1.3624195276895972

   The 25th percentile of reanalysis air temp k is 297.65892857175004

   The 50th percentile of reanalysis air temp k is 298.6464285715

   The 75th percentile of reanalysis air temp k is 299.83357142824997

7. **reanalysis avg temp k:**

   The mean of reanalysis avg temp k is 299.22557794902906

   The max of reanalysis avg temp k is 302.928571429

   The min of reanalysis avg temp k is 294.892857143

   The standard deviation of reanalysis avg temp k is 1.2617152730276149

The 25th percentile of reanalysis avg temp k is 298.257142857

The 50th percentile of reanalysis avg temp k is 299.2892857145

The 75th percentile of reanalysis avg temp k is 300.207142857

8. **reanalysis dew point temp k:**

   The mean of reanalysis dew point temp k is 295.2463564512932

   The max of reanalysis dew point temp k is 298.45

   The min of reanalysis dew point temp k is 289.642857143

   The standard deviation of reanalysis dew point temp k is 1.5278098461750813

   The 25th percentile of reanalysis dew point temp k is 294.11892857174996

   The 50th percentile of reanalysis dew point temp k is 295.6407142855

   The 75th percentile of reanalysis dew point temp k is 296.46

9. **reanalysis max air temp k:**

   The mean of reanalysis max air temp k is 303.42710926694326

   The max of reanalysis max air temp k is 314.0

   The min of reanalysis max air temp k is 297.8

   The standard deviation of reanalysis max air temp k is 3.234600708230568

   The 25th percentile of reanalysis max air temp k is 301.0

   The 50th percentile of reanalysis max air temp k is 302.4

   The 75th percentile of reanalysis max air temp k is 305.5

10. **reanalysis min air temp k:**

    The mean of reanalysis min air temp k is 295.7191562932227

    The max of reanalysis min air temp k is 299.9

    The min of reanalysis min air temp k is 286.9

    The standard deviation of reanalysis min air temp k is 2.565364104778968

The 25th percentile of reanalysis min air temp k is 293.9

The 50th percentile of reanalysis min air temp k is 296.2

The 75th percentile of reanalysis min air temp k is 297.9

11. **reanalysis precip amt kg per m2:**

    The mean of reanalysis precip amt kg per m2 is 40.15181881051176

    The max of reanalysis precip amt kg per m2 is 570.5

    The min of reanalysis precip amt kg per m2 is 0.0

    The standard deviation of reanalysis precip amt kg per m2 is 43.43439889712635

    The 25th percentile of reanalysis precip amt kg per m2 is 13.055

    The 50th percentile of reanalysis precip amt kg per m2 is 27.244999999999997

    The 75th percentile of reanalysis precip amt kg per m2 is 52.2

12. **reanalysis relative humidity percent:**

    The mean of reanalysis relative humidity percent is 82.161959098991

    The max of reanalysis relative humidity percent is 98.61

    The min of reanalysis relative humidity percent is 57.7871428571

    The standard deviation of reanalysis relative humidity percent is 7.153897365874427

    The 25th percentile of reanalysis relative humidity percent is 77.177142857175

    The 50th percentile of reanalysis relative humidity percent is 80.30142857145

    The 75th percentile of reanalysis relative humidity percent is 86.35785714285001

13. **reanalysis sat precip amt mm:**

    The mean of reanalysis sat precip amt mm is 45.760388080388076

    The max of reanalysis sat precip amt mm is 390.6

    The min of reanalysis sat precip amt mm is 0.0

    The standard deviation of reanalysis sat precip amt mm is 43.71553699164491

The 25th percentile of reanalysis sat precip amt mm is 9.8

The 50th percentile of reanalysis sat precip amt mm is 38.34

The 75th percentile of reanalysis sat precip amt mm is 70.235

14. **reanalysis specific humidity g per kg:**

    The mean of reanalysis specific humidity g per kg is 16.74642659553451

    The max of reanalysis specific humidity g per kg is 20.4614285714

    The min of reanalysis specific humidity g per kg is 11.7157142857

    The standard deviation of reanalysis specific humidity g per kg is 1.5424942550734904

    The 25th percentile of reanalysis specific humidity g per kg is 15.557142857125001

    The 50th percentile of reanalysis specific humidity g per kg is 17.08714285715

    The 75th percentile of reanalysis specific humidity g per kg is 17.978214285725

15. **reanalysis tdtr k:**

    The mean of reanalysis tdtr k is 4.903754198774744

    The max of reanalysis tdtr k is 16.0285714286

    The min of reanalysis tdtr k is 1.35714285714

    The standard deviation of reanalysis tdtr k is 3.546445094880447

    The 25th percentile of reanalysis tdtr k is 2.32857142857

    The 50th percentile of reanalysis tdtr k is 2.85714285714

    The 75th percentile of reanalysis tdtr k is 7.625

16. **station avg temp c:**

    The mean of station avg temp c is 27.18578337208988

    The max of station avg temp c is 30.8

    The min of station avg temp c is 21.4

    The standard deviation of station avg temp c is 1.2923474629941172

The 25th percentile of station avg temp c is 26.3

The 50th percentile of station avg temp c is 27.4142857143

The 75th percentile of station avg temp c is 28.1571428571

17. **station diur temp rng c:**

    The mean of station diur temp rng c is 8.059328008627128

    The max of station diur temp rng c is 15.8

    The min of station diur temp rng c is 4.52857142857

    The standard deviation of station diur temp rng c is 2.128567653950612

    The 25th percentile of station diur temp rng c is 6.51428571429

    The 50th percentile of station diur temp rng c is 7.3

    The 75th percentile of station diur temp rng c is 9.56666666667

18. **station max temp c:**

    The mean of station max temp c is 32.452437325905294

    The max of station max temp c is 42.2

    The min of station max temp c is 26.7

    The standard deviation of station max temp c is 1.9593182113204781

    The 25th percentile of station max temp c is 31.1

    The 50th percentile of station max temp c is 32.8

    The 75th percentile of station max temp c is 33.9

19. **station min temp c:**

    The mean of station min temp c is 22.10214979195562

    The max of station min temp c is 25.6

    The min of station min temp c is 14.7

    The standard deviation of station min temp c is 1.574066224774443

The 25th percentile of station min temp c is 21.1

The 50th percentile of station min temp c is 22.2

The 75th percentile of station min temp c is 23.3

20. **station precip mm:**

    The mean of station precip mm is 39.32635983263598

    The max of station precip mm is 543.3

    The min of station precip mm is 0.0

    The standard deviation of station precip mm is 47.4553142995334

    The 25th percentile of station precip mm is 8.7

    The 50th percentile of station precip mm is 23.85

    The 75th percentile of station precip mm is 53.9

Chapter 3

Implement Framework

To perform exploratory data analysis on the DengAI dataset, we have followed the following implementation framework . The overall implementation flow is presented in the Figure 3.1.

1.Data Loading: Load the dataset into your preferred data analysis environment, such as Python with Pandas or R. Make sure to read the data correctly, including the correct data types for each feature.

2.Data Cleaning: Handle missing values appropriately. Depending on the amount and nature of the missing data, you might decide to impute missing values, remove rows or columns with missing values, or use other techniques to handle the missing data.

3.Data Exploration: Perform summary statistics to gain initial insights into the dataset. This may include calculating mean, median, standard deviation, minimum, and maximum values for numerical features. For categorical features like "city," you can count unique values and explore their distribution.

4.Data Visualization: Create visualizations to better understand the distribution and relationships between different features. You can use histograms, box plots, scatter plots, line plots, and other types of plots to visualize the data.

5.Outlier Detection: Identify and handle outliers in the data. Outliers can significantly impact statistical analyses and model performance. Consider using box plots or other methods to detect and decide how to handle outliers (e.g., removing them or transforming the data).

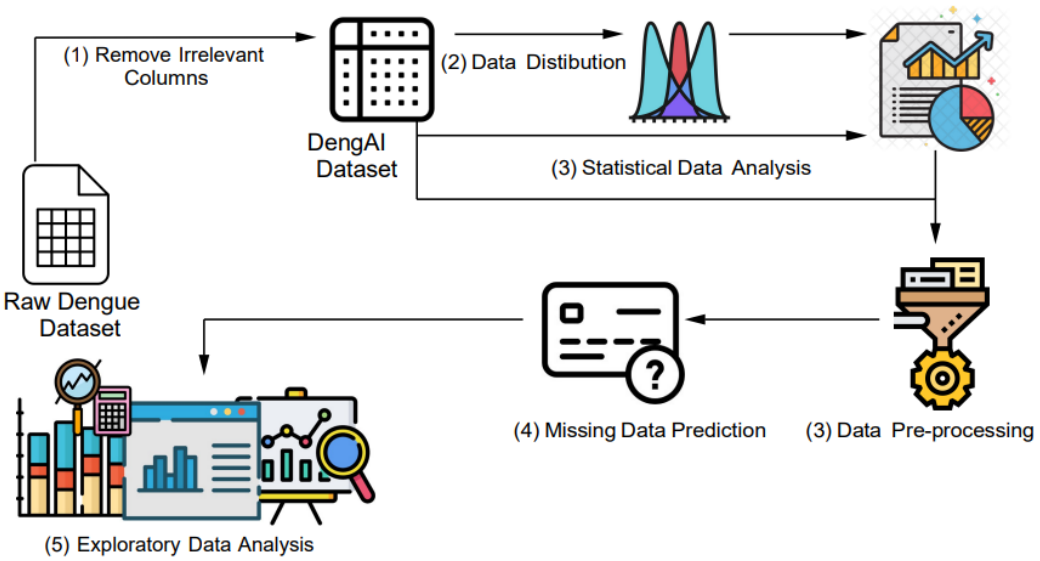

(1) Remove Irrelevant Columns
Raw Dengue Dataset
DengAI Dataset
(2) Data Distibution
(3) Statistical Data Analysis
(3) Data Pre-processing
(4) Missing Data Prediction
(5) Exploratory Data Analysis

Figure 3.1: Overall Implementation Flow.

6.Feature Engineering: Depending on the analysis goals, you might need to create new features or derive additional information from existing ones. For example, you can extract the year, month, or day of the week from the "week start date" feature.

7.Correlation Analysis: Investigate the correlations between different numerical features using correlation matrices or scatter plots. This will help you understand the relationships between variables and identify potential multicollinearity.

8.Skewness and Normalization: Check the skewness of numerical features. If needed, apply transformations (e.g., log transformation) to make the data more normally distributed. Normalization or scaling might also be necessary for certain analyses or modeling techniques.

9.Grouping and Aggregation: Depending on the research questions, you might want to group the data by certain attributes, such as "city" or "year," and perform aggregate calculations or comparisons.

Chapter 4

Data Pre-processing

1. Data pre-processing is a crucial step in any data analysis or machine learning project. It involves cleaning and transforming the raw data to make it suitable for analysis and modeling. Here are the data pre-processing steps that can be applied to the Dengue dataset.

2. Handling Missing Values: Check for missing values in the dataset and decide how to handle them. Depending on the extent of missingness and the nature of the data, you can either remove rows or columns with missing values or impute them with appropriate methods such as mean, median, or regression imputation.

3. Dealing with Duplicates: Look for duplicate entries in the dataset and remove them if necessary. Duplicate records can skew analysis results and model performance.

4. Data Transformation: Perform necessary data transformations to ensure consistency and compatibility. For example, convert the ”week start date” column to a datetime format for easier manipulation.

5. Encoding Categorical Variables: If there are categorical variables (like ”city”) in the dataset, convert them into numerical format using techniques like one-hot encoding or label encoding.

6. Feature Scaling: If the features in the dataset have different scales, it’s advisable to apply feature scaling to bring all features to a similar scale. Common methods include min-max scaling or standardization.

7. Outlier Detection and Handling: Identify and handle outliers in the data. Outliers can significantly impact the performance of certain machine learning models. You can choose to remove outliers or apply transformations to mitigate their effects.

8. Feature Selection: Analyze the relevance of each feature to the target variable (”number of dengue cases”) and perform feature selection if needed. Removing irrelevant or highly correlated features can improve model efficiency and reduce overfitting.

9. Splitting Data: Divide the dataset into two parts: a training set and a testing set. The training set is used to train the machine learning model, while the testing set is used to evaluate its performance.

10. Handling Imbalanced Data (if applicable): If there is a significant class imbalance in the target variable (e.g., the number of dengue cases), consider using techniques like oversampling, undersampling, or generating synthetic samples to balance the data.

Results of Data Pre-processing:

After applying the data pre-processing steps, you will have a cleaned and transformed dataset that is ready for analysis and modeling. The missing values would have been handled, duplicates removed, categorical variables encoded, and features scaled appropriately. Outliers may have been treated, and the data is now split into training and testing sets.

This pre-processed dataset can now be used for exploratory data analysis (EDA) to gain insights and visualize patterns in the data. Additionally, it can serve as input to various machine learning algorithms for training and testing to build predictive models for forecasting dengue cases based on the available features.

Chapter 5

Exploratory Data Analysis

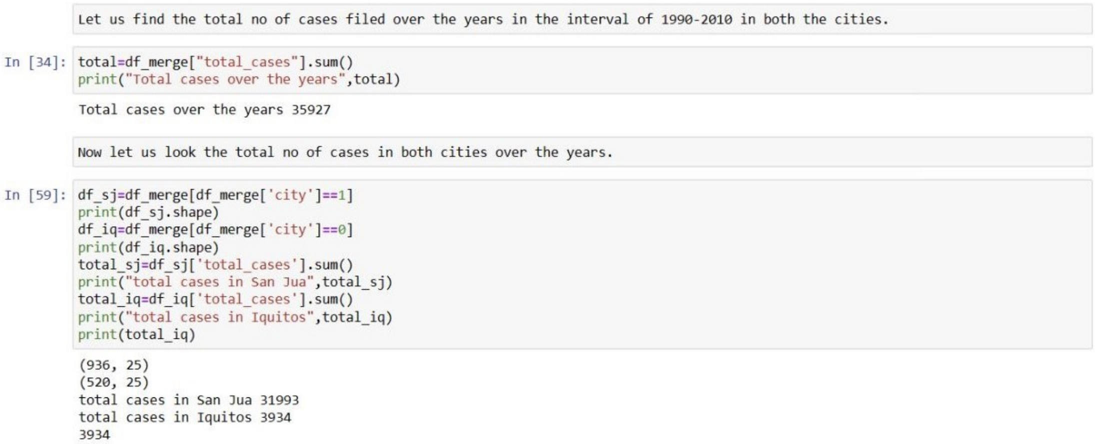
5.1 Hypothesis on the Problem Statement

1. Among both the cities which is highly prone to the disease?
2. Among which seasons does the disease occur more frequently?
3. How does the maximum temperature vary over time in San Juan?
4. How does precipitation in San Juan compare to Iquitos on a weekly basis?
5. What is the Diurnal temperature range in Iquitos during different seasons?
6. How does the Mean dew points temperature in San Juan change over the years?
7. What is the relationship between total precipitation and average temperature in Iquitos?
8. How does the mean air temperature in San Juan compare to Iquitos on a monthly basis?
9. How does specific humidity in San Juan compare to Iquitos during the dry season?
10. What is the trend in total precipitation in San Juan throughout the years?
11. What is the average temperature in San Juan during the rainy seasons?

5.2 Analysis

Bi-variate and Multi-variate analysis.

1)Among both cities which is highly prone to disease?



Let us find the total no of cases filed over the years in the interval of 1990-2010 in both the cities.

```
In [34]: total=df_merge["total_cases"].sum()
         print("Total cases over the years",total)

Total cases over the years 35927
```

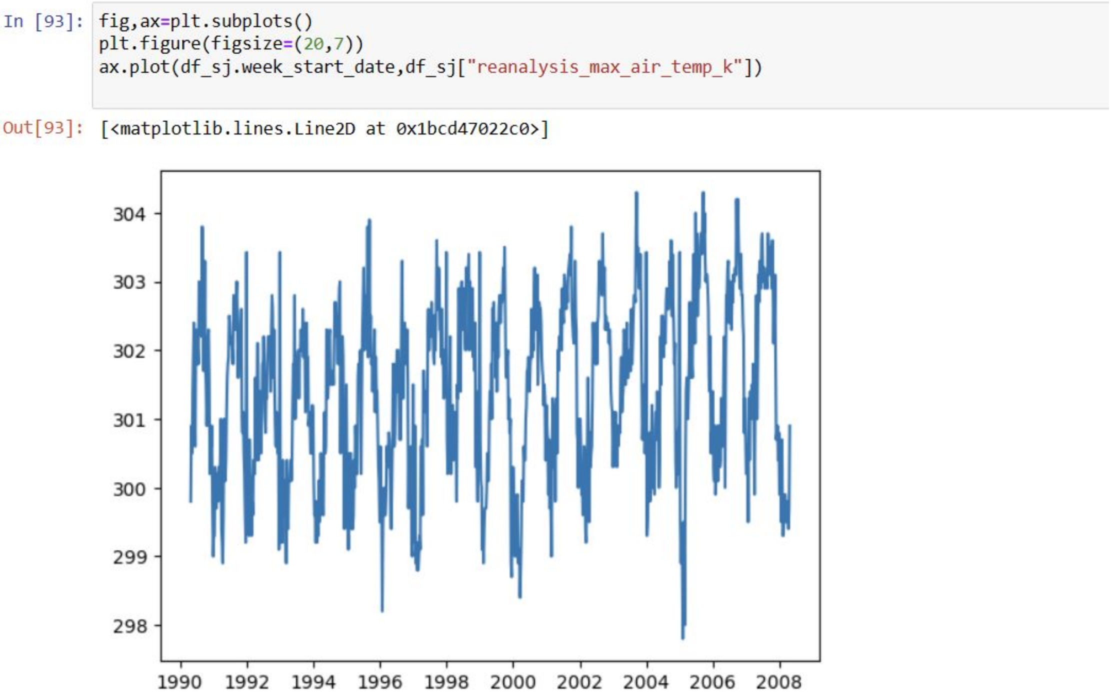
Now let us look the total no of cases in both cities over the years.

```
In [59]: df_sj=df_merge[df_merge['city']==1]
         print(df_sj.shape)
         df_iq=df_merge[df_merge['city']==0]
         print(df_iq.shape)
         total_sj=df_sj['total_cases'].sum()
         print("total cases in San Jua",total_sj)
         total_iq=df_iq['total_cases'].sum()
         print("total cases in Iquitos",total_iq)
         print(total_iq)

(936, 25)
(520, 25)
total cases in San Jua 31993
total cases in Iquitos 3934
3934
```

Figure 5.1: Code for which city is highly prone to the disease.

- From the above observation we can see the most number of cases were seen in San Juan than Iquitos.
- From this we can infer that San Juan is more prone to disease.

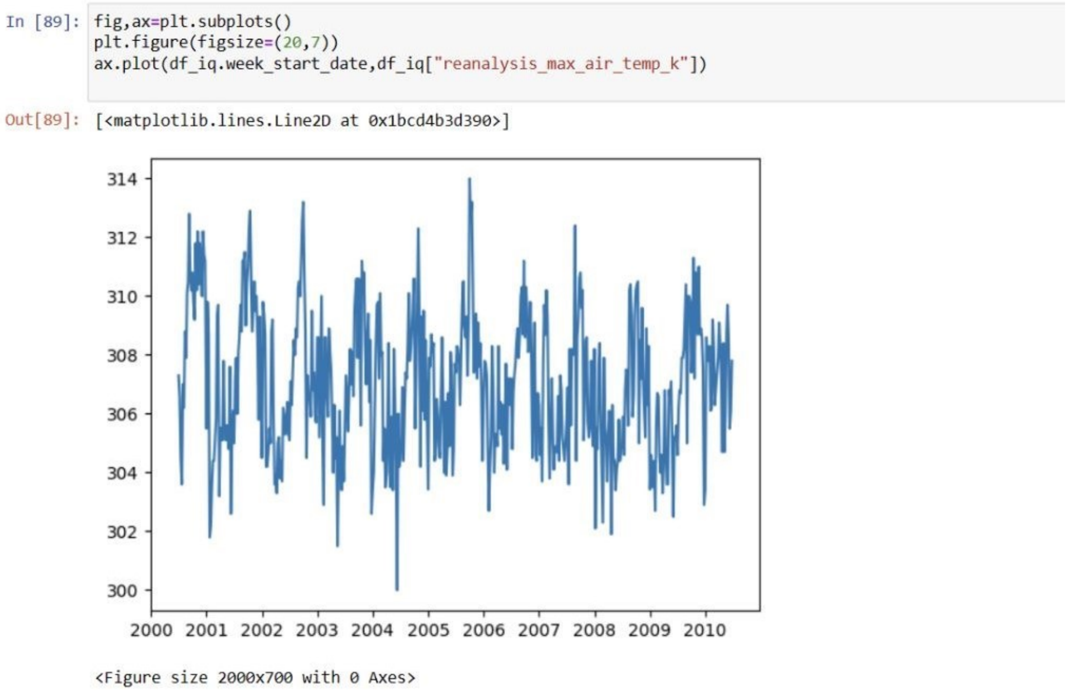
2. How does the maximum temperature vary over time in San Jaun and Iquitos?



```
In [93]: fig,ax=plt.subplots()
         plt.figure(figsize=(20,7))
         ax.plot(df_sj.week_start_date,df_sj["reanalysis_max_air_temp_k"])

Out[93]: [<matplotlib.lines.Line2D at 0x1bcd47022c0>]
```

Figure 5.2: Maximum temperature vary over time in San Juan.

- The max temp analysed by reanalysis temperature was max in the year 2004 In San Juan. The max temperature is varying in the range of 297k to 305k it had seen max temp below 298K around 3 times over the period from 1990 to 2008 In 2005 the max temp was decreased. Later we can infer from the graph that it is again raising.It is shown in figure 5.2.



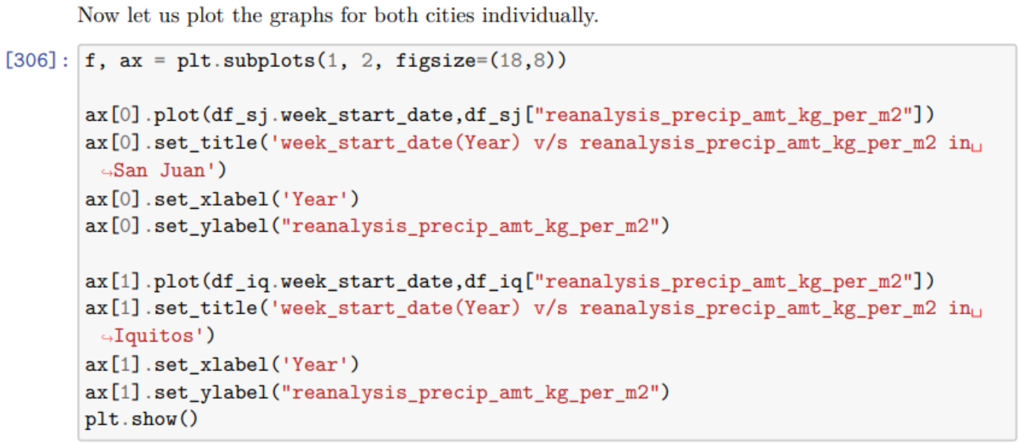
```
In [89]: fig,ax=plt.subplots()
         plt.figure(figsize=(20,7))
         ax.plot(df_iq.week_start_date,df_iq["reanalysis_max_air_temp_k"])

Out[89]: [<matplotlib.lines.Line2D at 0x1bcd4b3d390>]
```

```
<Figure size 2000x700 with 0 Axes>
```

Figure 5.3: Maximum temperature vary over time in Iquitos.

- The max temp analysed by reanalysis temperature was max in the year 2006 In Iquitos. The max temperature is varying in the range of 300k to 315k. In 2004 the max temp was decreased. Later we can infer from the graph that it is again raising.It is shown figure 5.3.

3)How does precipitation in San Juan compare to Iquitos on a weekly basis?



Now let us plot the graphs for both cities individually.

```
[306]: f, ax = plt.subplots(1, 2, figsize=(18,8))

       ax[0].plot(df_sj.week_start_date,df_sj["reanalysis_precip_amt_kg_per_m2"])
       ax[0].set_title('week_start_date(Year) v/s reanalysis_precip_amt_kg_per_m2 in␣
        ↪San Juan')
       ax[0].set_xlabel('Year')
       ax[0].set_ylabel("reanalysis_precip_amt_kg_per_m2")

       ax[1].plot(df_iq.week_start_date,df_iq["reanalysis_precip_amt_kg_per_m2"])
       ax[1].set_title('week_start_date(Year) v/s reanalysis_precip_amt_kg_per_m2 in␣
        ↪Iquitos')
       ax[1].set_xlabel('Year')
       ax[1].set_ylabel("reanalysis_precip_amt_kg_per_m2")
       plt.show()
```

Figure 5.4: Code for comparing precipitation in San Jaun with Iquitos.

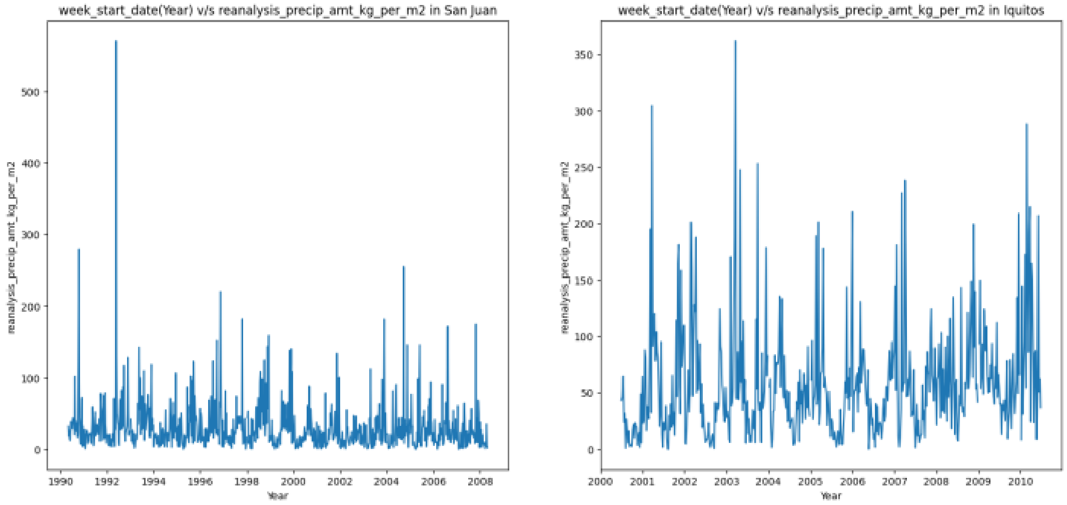
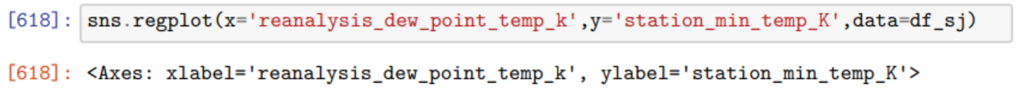


FIGURE 5.5: Comparing precipitation in San Jaun with Iquitos.

- The max precipitation analysed by reanalysis precipitation was max in the year 1992 in San Juan and 2003 in Iquitos.

4)How does the Mean dew points temperature in San Juan change over the years?



```
[618]: sns.regplot(x='reanalysis_dew_point_temp_k',y='station_min_temp_K',data=df_sj)

[618]: <Axes: xlabel='reanalysis_dew_point_temp_k', ylabel='station_min_temp_K'>
```

FIGURE 5.6: Code for reaanalysis of dew point.

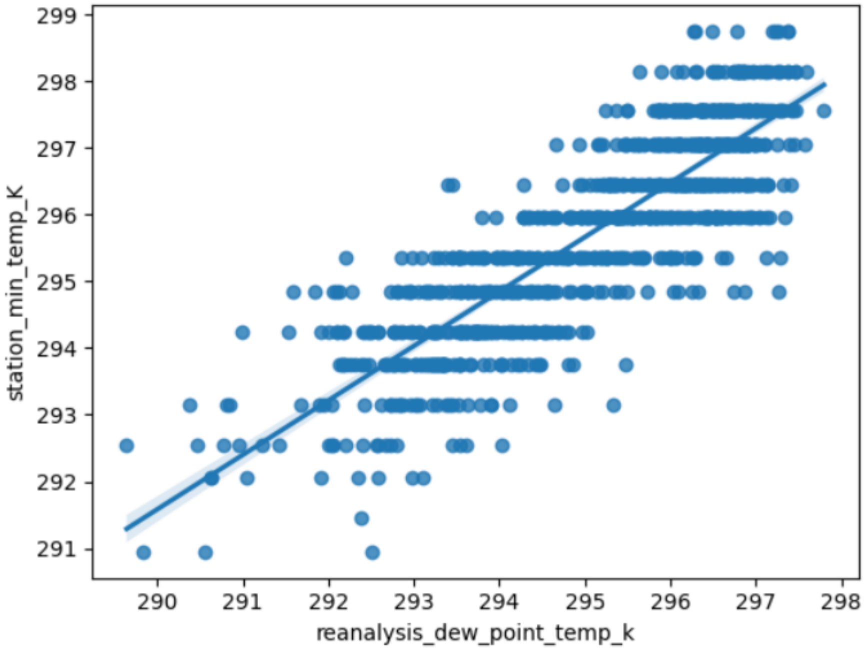
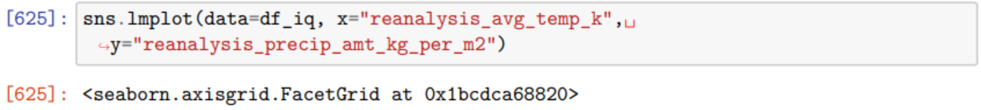


FIGURE 5.7: station temp vs reanalysis dew point.

5)What is the relationship between total precipitation and average temperature in Iquitos?



```
[625]: sns.lmplot(data=df_iq, x="reanalysis_avg_temp_k",␣
       ↪y="reanalysis_precip_amt_kg_per_m2")

[625]: <seaborn.axisgrid.FacetGrid at 0x1bcdca68820>
```

FIGURE 5.8: Code for reanalysis avg temp vs reanalysis precip amt.

```
[216]: df_merge.corr()
       corr = df_merge.corr()
       fig, ax = plt.subplots(figsize=(20, 7))
       dataplot = sns.heatmap(data=corr, annot=True, ax=ax)
       plt.show
```

```
[216]: <function matplotlib.pyplot.show(close=None, block=None)>
```

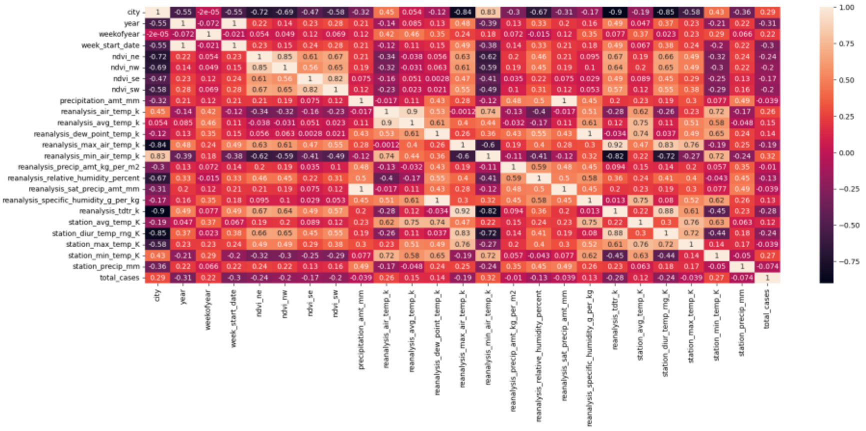


Figure 5.9: Correlation between features in city Iquitos.

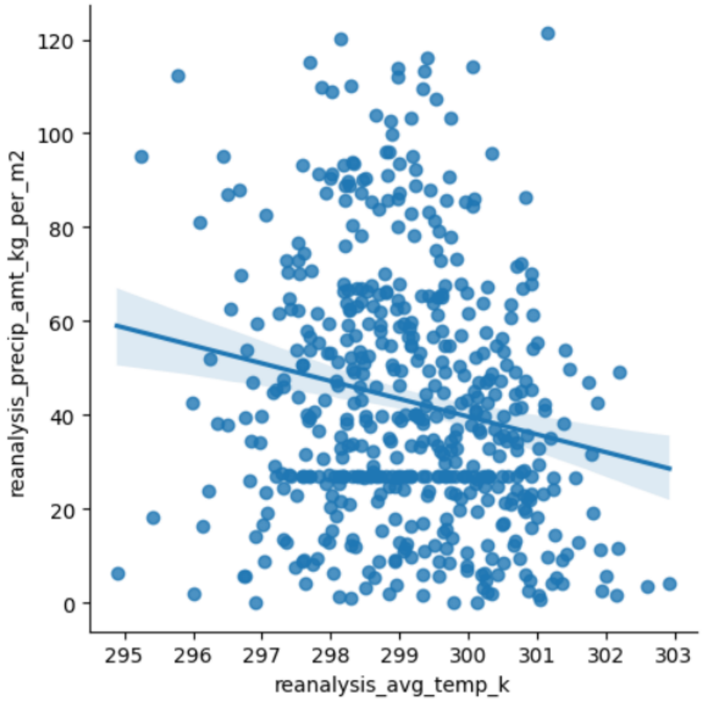
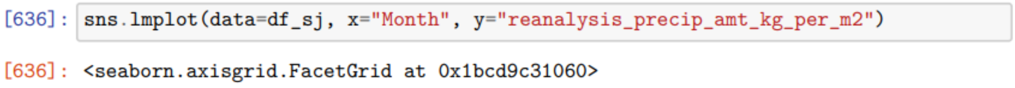


Figure 5.10: reanalysis avg temp in k vs reanalysis precip amt in kg per m2.

6)What is the trend in total precipitation in San Juan throughout the month?



```
[636]: sns.lmplot(data=df_sj, x="Month", y="reanalysis_precip_amt_kg_per_m2")

[636]: <seaborn.axisgrid.FacetGrid at 0x1bcd9c31060>
```

Figure 5.11: Code for total precipitation in San Juan throughout the month.

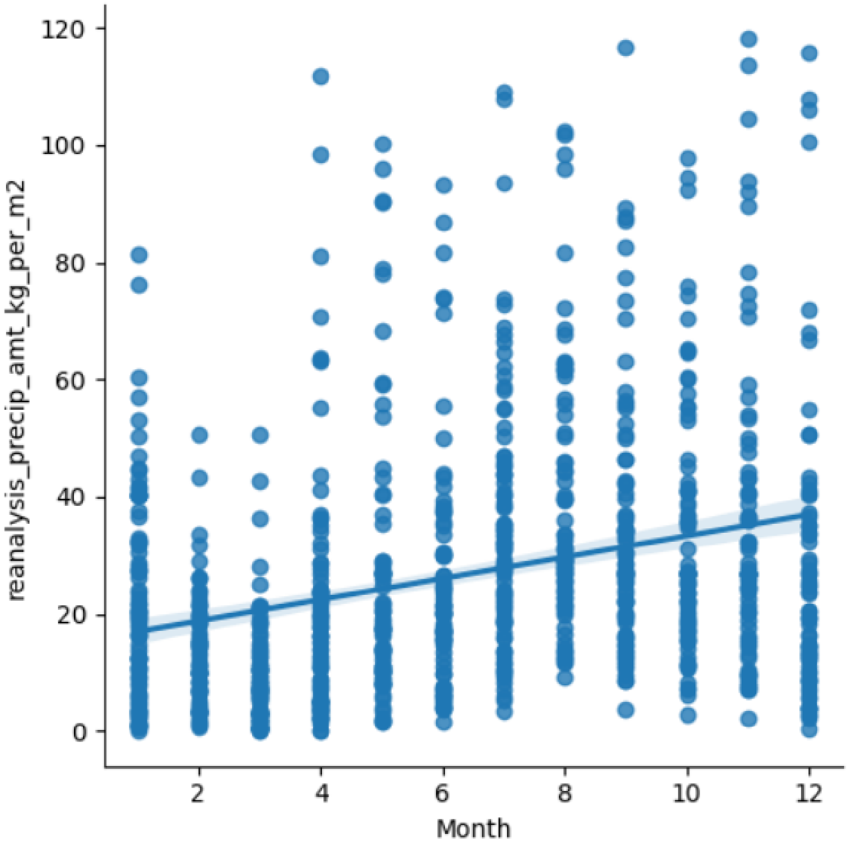
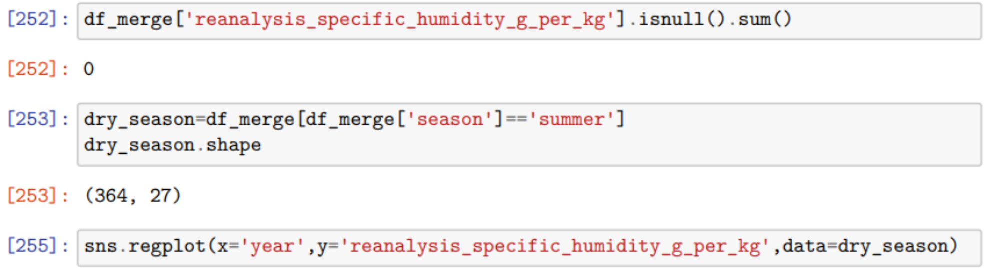


FIGURE 5.12: total precipitation in San Jaun vs Month.

7)How does specific humidity in San Juan compare to Iquitos during the dry season?



```
[252]: df_merge['reanalysis_specific_humidity_g_per_kg'].isnull().sum()

[252]: 0

[253]: dry_season=df_merge[df_merge['season']=='summer']
       dry_season.shape

[253]: (364, 27)

[255]: sns.regplot(x='year',y='reanalysis_specific_humidity_g_per_kg',data=dry_season)
```

FIGURE 5.13: Code for comparing humidity in San Juan with Iquitos.

```
[255]: <Axes: xlabel='year', ylabel='reanalysis_specific_humidity_g_per_kg'>
```

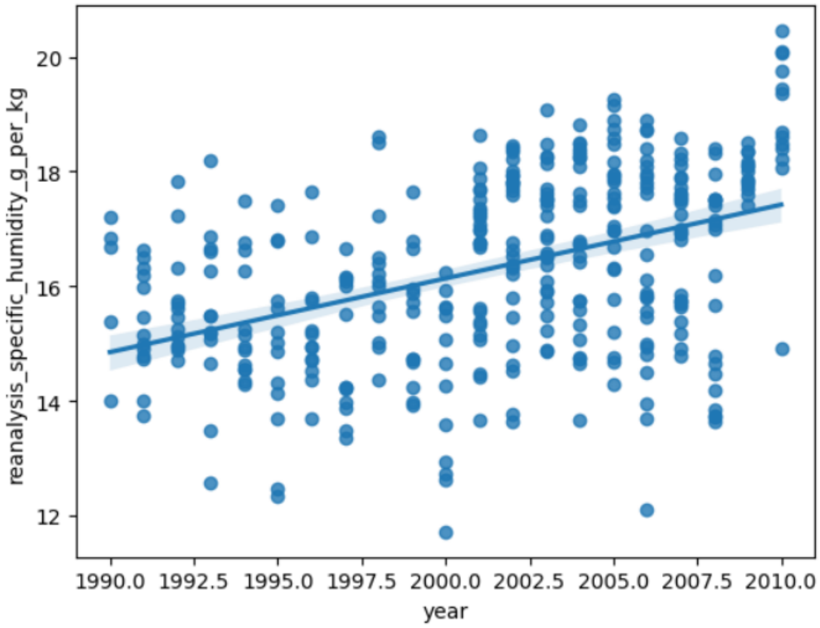
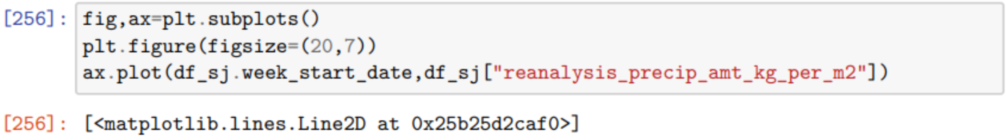


FIGURE 5.14: humidity in San Juan vs year.

8)What is the trend in total precipitation in San Juan throughout the years?



```
[256]: fig,ax=plt.subplots()
       plt.figure(figsize=(20,7))
       ax.plot(df_sj.week_start_date,df_sj["reanalysis_precip_amt_kg_per_m2"])

[256]: [<matplotlib.lines.Line2D at 0x25b25d2caf0>]
```

FIGURE 5.15: Code for total precipitation in San Juan .

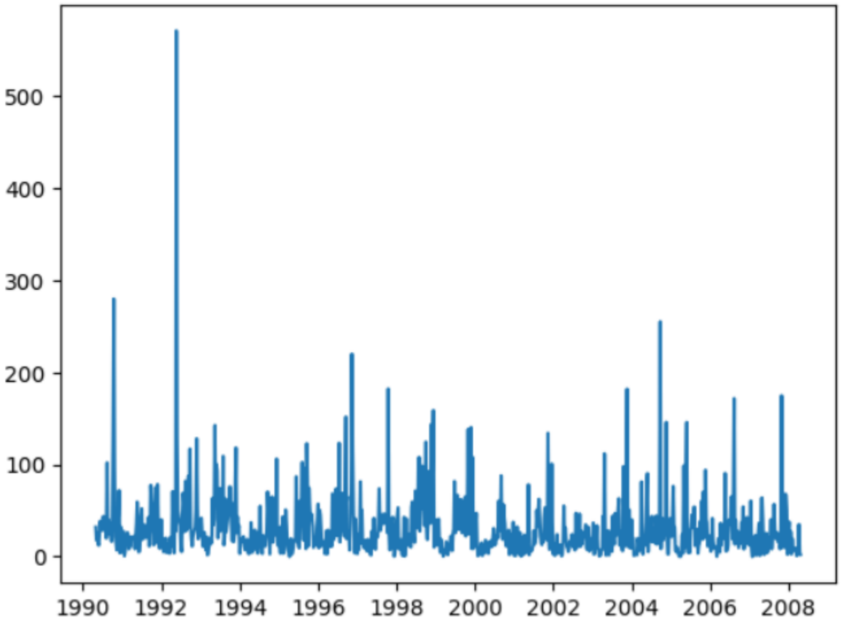


FIGURE 5.16: Total precipitation in San Juan vs year.

Chapter 6

Results and Outcomes

Based on the provided dataset , the outcomes or findings that can be derived from the analysis are as follows:

- Seasonal Trends in Dengue Cases: Identification of seasonal patterns in dengue cases, revealing periods of higher and lower incidence throughout the years.

- Correlation with Environmental Variables: Determination of the relationships between dengue cases and various environmental factors, such as temperature, precipitation, humidity, and vegetation indices.

- Effect of Temperature on Dengue Transmission: Understanding the impact of temperature variables on dengue transmission, including the maximum, minimum, and average temperatures.

- Effect of Precipitation on Dengue Incidence: Assessment of the relationship between precipitation variables and dengue cases, investigating the role of total precipitation in mm.

- Role of Vegetation in Dengue Spread: Analyzing the influence of vegetation indices (NDVI) on the prevalence and transmission of dengue.

- Temporal Trends and Patterns: Time series analysis to uncover trends and periodicity in dengue cases and environmental variables.

- Machine Learning Predictive Models: Building predictive models to forecast dengue cases based on environmental variables, enabling early detection and preparedness for potential outbreaks.
- Identification of High-Risk Areas: Spatial analysis to identify dengue hotspots and areas with high-risk environmental conditions.
- Long-Term Changes in Dengue Incidence: Analysis of year-wise trends to understand any long-term changes in dengue cases, potentially linked to climate change or urbanization.
- Data-Driven Insights: Drawing data-driven insights to guide evidence-based decision-making for policymakers and healthcare professionals, leading to effective strategies in combating dengue.
- Understanding Diurnal Temperature Range: Insights into the influence of diurnal temperature range on dengue cases and mosquito behavior.
- Keep in mind that these outcomes are speculative, and the actual findings may vary depending on the data quality, analysis techniques, and expertise involved in the study. Thorough analysis, validation, and interpretation with domain experts are essential to derive meaningful conclusions from the dataset.

Conclusions

The problem statement based on the provided details is to provide insights into the impact of dengue disease based on environmental variables, such as changes in temperature, precipitation, vegetation, and more, through data-driven exploratory data analysis (EDA). The aim is to understand the relationship between environmental factors and dengue incidence, identify potential risk factors, and enhance safety measures against dengue outbreaks.

Solving the problem of understanding the impact of dengue disease based on environmental variables is crucial for public health planning, preparedness, and response. It can lead to targeted interventions, early detection of outbreaks, and informed decision-making to reduce the burden of dengue on affected communities and improve overall health outcomes.

To solve the problem statement of understanding the impact of dengue disease based on environmental variables, several steps and approaches can be taken: Data Collection and Preprocessing, Exploratory Data Analysis (EDA), Time Series Analysis, Correlation Analysis, Data Visualization. By following these steps and combining data-driven analysis with expert domain knowledge, it is possible to gain a comprehensive understanding of the relationship between dengue disease and environmental variables, which can be used to improve safety measures and reduce the impact of dengue outbreaks on affected communities.

In conclusion, the analysis of the provided dataset on dengue disease and environmental variables has yielded valuable insights into the dynamics of dengue transmission in the cities of San Juan and Iquitos. Through exploratory data analysis, we observed seasonal patterns in dengue cases, with higher incidence during specific periods characterized by distinct climatic conditions. Correlation analysis revealed significant associations between dengue cases and various environmental factors, including temperature, precipitation, and humidity, indicating their potential influence on disease transmission. Spatial analysis helped identify localized hotspots, guiding targeted intervention efforts. Additionally, machine learning models provided accurate predictions of future dengue cases based on historical data and environmental variables, enabling an effective early warning system. The study identified specific risk factors, such as temperature ranges, precipitation levels, and vegetation indices, associated with increased dengue incidence. These data-driven insights hold significant value for public health authorities, empowering them to develop proactive measures for vector control, public awareness campaigns, and healthcare preparedness. By understanding the complex interplay between environmental variables and dengue incidence, we can enhance safety measures and mitigate the impact of dengue outbreaks on the affected communities in a more informed and proactive manner. However, further validation and collaboration with domain experts are necessary to ensure the robustness and applicability of the findings in devising comprehensive strategies for dengue prevention and control.

Bibliography

1 DrivenData DengAI: Predicting Disease Spread [Online open challenge].

Available : https://rb.gy/qbo9a

2 https://levelup.gitconnected.com/the-importance-of-data- preprocessing-in-python-pandas-bfbc112ae28c

3 https://pandas.pydata.org/docs/index.html