Dengue Fever

Prediction of disease cases in Puerto Rico and Peru

Group 6: Alisa Liu, Freddy Chen, Jessica Quansah, Keani Schuller, Zhicheng Zhong

01

Introduction

Context and Goals

Context

- Dengue fever is a disease transmitted by mosquitoes commonly found in tropical & subtropical regions
- Symptoms in mild cases are flu, including fever, rash, muscle and joint pain
- Severe cases can result in complications including severe bleeding, low blood pressure, and even death

- Due to its mosquito-borne nature, the spread of dengue fever is closely linked to climate factors such as temperature and precipitation
- While the relationship between dengue and climate is intricate, an increasing number of scientists assert that climate change is likely to cause shifts in the distribution of the disease
- This could have significant public health implications worldwide

Context

- Given that its symptoms resemble many other feverish sicknesses, dengue fever cases are severely under-reported
- Many people develop either mild symptoms or no symptoms, making it difficult to actually monitor the active number of cases, especially in countries that are already lacking healthcare resources
- The WHO reports that the number of dengue fever cases has skyrocketed in the last few decades, classifying the outbreaks as an epidemic
- With a model that can predict the number of future dengue fever cases, these countries can gain more visibility on the number of cases of dengue fever and its patterns
- This will allow them to better allocate their health care resources and help reduce the impact of dengue fever

Target Audience

Public Health Officials

Predictive data can allow for efficient resource allocation and better health strategies

Epidemiologists & Researchers

Equip researchers with predictive models to ensure accurate and up-to-date public health monitoring

Healthcare Organizations

Optimize their resource allocation, patient care, staffing, and operational efficiency

Stakeholders

General Population

The Puerto Rican and Peruvian general population are all at risk of contracting dengue fever without proper measure being taken

Medical Staff

Medical professionals can be better equipped to treat dengue fever cases if they are better informed

Policy Makers

To minimize dengue fever cases, policy makers must be well informed to invest in infrastructure to accomplish this and support medical professionals

CDC

The CDC has global influence on policy makers and the decisions of health organizations

Project Goals

- To predict the total number of dengue fever cases in San Juan and Iquitos given various variables on temperature, precipitation, vegetation index and humidity
- Compare different time series forecasting models to find the best model for creating these models
- Use these predictive models to enable the establishment of early warning systems for disease outbreaks, ensuring timely responses and resource allocation in the most high-risk areas of San Juan and Iquitos
- Encourage data-driven decision-making in public health policy and healthcare organizations by providing a clear report on dengue fever case findings

02

Dataset

Description of dataset and key variables

Dataset Description

- The dataset is a time series dataset consisting of 1457 observations on the number of weekly dengue fever cases and the environmental data of that week
- 937 from San Juan, Puerto Rico
- 520 from Iquitos, Peru
- Data spans for San Juan are from 1990 - 2008 and that of Iquitos are from 2000-2010
- It includes different variables on the weekly environment in each city, such as maximum and minimum temperature, humidity, and precipitation
- The dataset variables will be used to predict the future number of cases based on the environmental situation

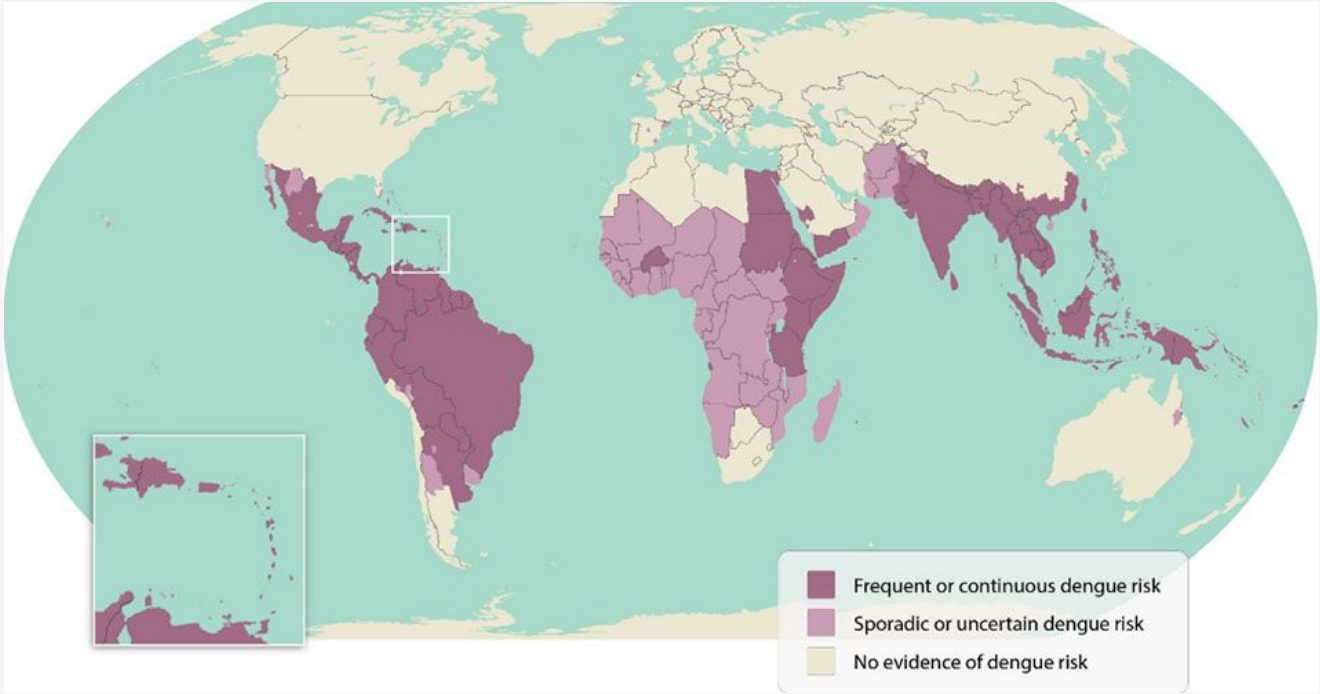
Prior Knowledge

Disease such as dengue, often has strong correlations with meteorological conditions:

- Temperature: "Warm temperatures facilitated rapid virus replication, warm conditions (>36°C) also increased mosquito mortality and led to an overall decrease in transmission potential"
- Precipitation: Rainfall can create breeding conditions for mosquitoes, "mosquitoes lay their eggs in standing water" especially
- Humidity: "Experimental work has thus far shown generally positive effects of increased relative humidity on mosquito survival and desiccation tolerance, production and development of eggs and mosquito activity (up to 90% relative humidity)"
- Vegetation (NDVI): Dense vegetation can provide resting sites for adult mosquitoes. "Mosquitoes seek certain natural resting site habitats such as understory vegetation, tree cavities, rock crevices or animal burrows, and show strong species-specific preferences for the type of resting site habitat"

Prior Knowledge

This is also evidenced from the map as we see that areas characterized by humidity, high temperatures and tropical species exhibit higher dengue fever risk



Key Variables Overview

- Target Variable: total_cases
- Predictive Variables (20) :
  - Temperature-related metrics:
    - station_max_temp_c
    - station_min_temp_c
    - station_avg_temp_c
    - station_diur_temp_rng_c
    - reanalysis_max_air_temp_k
    - reanalysis_min_air_temp_k
    - reanalysis_avg_temp_k
    - reanalysis_air_temp_k
    - reanalysis_tdtr_k
  - Precipitation measures:
    - station_precip_mm
    - precipitation_amt_mm
    - reanalysis_sat_precip_amt_mm
    - reanalysis_precip_amt_kg_per_m2
  - Humidity-related metrics:
    - reanalysis_dew_point_temp_k
    - reanalysis_relative_humidity_percent
    - reanalysis_specific_humidity_g_per_kg
  - Vegetation indices
    - ndvi_se
    - ndvi_sw
    - ndvi_ne
    - ndvi_nw

Key Variables

- Temperature
  - Our dataset includes daily climate data as measure by NOAA's GHCN which records daily climate summaries from over 100,000 weather stations globally. So our data set contains temperature and precipitation data from the station closest to this city
  - The temperature data provided by the station is measured in degrees Celsius and we have on a weekly basis, the **maximum**, **minimum** and **average** temperatures
  - From the station, we also have **diurnal** temperature (this measures the difference between maximum and minimum temperatures in the station - gives us some indication of the stability of the weather)
  - Our dataset also contains data from NOAA's NCEP which forecasts based on a scientific model taking into consideration Earth's oceans, land etc.
  - The temperature data provided from this model is measured in Kelvin
  - Similar to the weather station data, here is as well **max**, **min**, **average** and **diurnal** temperature. We also have air temperature forecasts as well.

Key Variables

- Precipitation
  - Similar to temperature data on precipitation is provided by NOAA's GHCN from the weather stations as well as NCEP's climate forecasts. There is also an additional precipitation estimate from PERSSIAN
  - Three of the variables for total precipitation are in millimeters and one in kg per m2 which can also be converted to mm for standardization purposes

- Humidity
  - Humidity is also measured by two data points provided by NCEP's climate forecasting data.
  - The first datapoint is mean relative humidity which is a ratio of atmospheric pressure present relative to what would be present if the air were to be saturated; expressed in percentage
  - The dataset also contianes mean specific humidity which measures the weight of water vapor present in one unit weights of air. Here measured is g/kg

Key Variables

- Vegetation Index (captured by Normalized Difference Vegetation Index (NDVI))
  - NDVI captures the healthiness of the vegetation in an area (i.e how green the vegetation is in a given area. It is computed as an index ranging normally from -1 to 1 with higher positive values denoting more green areas and more negative denoting less green areas
  - There are four key variables capturing Vegetation Index in our dataset based on ordinal directions -SouthWest (ndvi_sw), South East (ndvi_se), NorthWest (ndvi_nw) , NorthEast (ndvi_ne)
  - These variables are important because the literature tells us that areas with greener/ higher ndvis should correlate with the total number of cases of dengue fever found in the area.

Business Unit & Importance

- The primary unit of analysis remains the (year, week_of_year) timescale for each city.
- It is important to the following aspects:
  - Public Health: Timely predictions can help allocate medical resources (effective allocation of resources in high-risk areas.), conduct public health campaigns, or implement preventive measures.
  - Research: The dataset and prediction results can provide insights for academic research in epidemiology and public health

03

Data Exploration

Process and interesting Insights

Summary Statistics: Overall

	count	mean	std	min	25%	50%	75%	max
ndvi_ne	1262.0	0.142294	0.140531	-0.406250	0.044950	0.128817	0.248483	0.508357
ndvi_nw	1404.0	0.130553	0.119999	-0.456100	0.049217	0.121429	0.216600	0.454429
ndvi_se	1434.0	0.203783	0.073860	-0.015533	0.155087	0.196050	0.248846	0.538314
ndvi_sw	1434.0	0.202305	0.083903	-0.063457	0.144209	0.189450	0.246982	0.546017
precipitation_amt_mm	1443.0	45.760388	43.715537	0.000000	9.800000	38.340000	70.235000	390.600000
reanalysis_air_temp_k	1446.0	298.701852	1.362420	294.635714	297.658929	298.646429	299.833571	302.200000
reanalysis_avg_temp_k	1446.0	299.225578	1.261715	294.892857	298.257143	299.289286	300.207143	302.928571
reanalysis_dew_point_temp_k	1446.0	295.246356	1.527810	289.642857	294.118929	295.640714	296.460000	298.450000
reanalysis_max_air_temp_k	1446.0	303.427109	3.234601	297.800000	301.000000	302.400000	305.500000	314.000000
reanalysis_min_air_temp_k	1446.0	295.719156	2.565364	286.900000	293.900000	296.200000	297.900000	299.900000
reanalysis_precip_amt_kg_per_m2	1446.0	40.151819	43.434399	0.000000	13.055000	27.245000	52.200000	570.500000
reanalysis_relative_humidity_percent	1446.0	82.161959	7.153897	57.787143	77.177143	80.301429	86.357857	98.610000
reanalysis_sat_precip_amt_mm	1443.0	45.760388	43.715537	0.000000	9.800000	38.340000	70.235000	390.600000
reanalysis_specific_humidity_g_per_kg	1446.0	16.746427	1.542494	11.715714	15.557143	17.087143	17.978214	20.461429
reanalysis_tdtr_k	1446.0	4.903754	3.546445	1.357143	2.328571	2.857143	7.625000	16.028571
station_avg_temp_c	1413.0	27.185783	1.292347	21.400000	26.300000	27.414286	28.157143	30.800000
station_diur_temp_rng_c	1413.0	8.059328	2.128568	4.528571	6.514286	7.300000	9.566667	15.800000
station_max_temp_c	1436.0	32.452437	1.959318	26.700000	31.100000	32.800000	33.900000	42.200000
station_min_temp_c	1442.0	22.102150	1.574066	14.700000	21.100000	22.200000	23.300000	25.600000
station_precip_mm	1434.0	39.326360	47.455314	0.000000	8.700000	23.850000	53.900000	543.300000
total_cases	1456.0	24.675137	43.596000	0.000000	5.000000	12.000000	28.000000	461.000000

- On average, the precipitation in San Juan and Iquitos was around 39 millimeters of rain
- The was a large spread of the number of total cases for the cities
- The hottest temperature overall was around 32 degrees Celsius
- The diurnal temperature, or difference between the high and lows of each day, was around 8 degrees, meaning it stays quite hot overall

Summary Statistics: San Juan

	count	mean	std	min	25%	50%	75%	max
ndvi_ne	745.0	0.057925	0.107153	-0.406250	0.004500	0.057700	0.111100	0.493400
ndvi_nw	887.0	0.067469	0.092479	-0.456100	0.016425	0.068075	0.115200	0.437100
ndvi_se	917.0	0.177655	0.057166	-0.015533	0.139283	0.177186	0.212557	0.393129
ndvi_sw	917.0	0.165956	0.056073	-0.063457	0.129157	0.165971	0.202771	0.381420
precipitation_amt_mm	927.0	35.470809	44.606137	0.000000	0.000000	20.800000	52.180000	390.600000
reanalysis_air_temp_k	930.0	299.163653	1.236429	295.938571	298.195000	299.254286	300.132857	302.200000
reanalysis_avg_temp_k	930.0	299.276920	1.218637	296.114286	298.300000	299.378571	300.228571	302.164286
reanalysis_dew_point_temp_k	930.0	295.109519	1.569943	289.642857	293.847857	295.464286	296.418929	297.795714
reanalysis_max_air_temp_k	930.0	301.398817	1.258927	297.800000	300.400000	301.500000	302.400000	304.300000
reanalysis_min_air_temp_k	930.0	297.301828	1.294705	292.600000	296.300000	297.500000	298.400000	299.900000
reanalysis_precip_amt_kg_per_m2	930.0	30.465419	35.628055	0.000000	10.825000	21.300000	37.000000	570.500000
reanalysis_relative_humidity_percent	930.0	78.568181	3.389488	66.735714	76.246071	78.667857	80.963214	87.575714
reanalysis_sat_precip_amt_mm	927.0	35.470809	44.606137	0.000000	0.000000	20.800000	52.180000	390.600000
reanalysis_specific_humidity_g_per_kg	930.0	16.552409	1.560923	11.715714	15.236429	16.845714	17.858571	19.440000
reanalysis_tdtr_k	930.0	2.516267	0.498892	1.357143	2.157143	2.457143	2.800000	4.428571
station_avg_temp_c	930.0	27.006528	1.415473	22.842857	25.842857	27.228571	28.185714	30.071429
station_diur_temp_rng_c	930.0	6.757373	0.835993	4.528571	6.200000	6.757143	7.285714	9.914286
station_max_temp_c	930.0	31.607957	1.717297	26.700000	30.600000	31.700000	32.800000	35.600000
station_min_temp_c	930.0	22.600645	1.506277	17.800000	21.700000	22.800000	23.900000	25.600000
station_precip_mm	930.0	26.785484	29.325811	0.000000	6.825000	17.750000	35.450000	305.900000
total_cases	936.0	34.180556	51.381372	0.000000	9.000000	19.000000	37.000000	461.000000

- There is less vegetation in north east and north west of San Juan
- The total cases in San Juan was the most spread out predictor
- The station's precipitation and the analysis' total precipitation are also more spread out than other predictors
- The average temperature was around 27 degrees, with a maximum high of 31 degrees, and a minimum low of 22 degrees Celsius

Summary Statistics: Iquitos

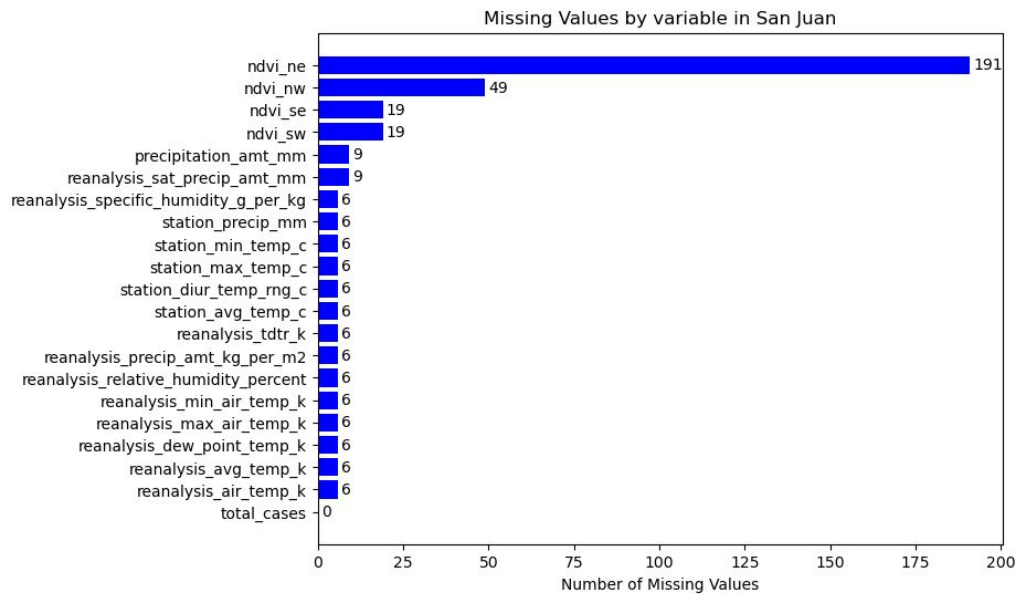
	count	mean	std	min	25%	50%	75%	max
ndvi_ne	517.0	0.263869	0.081370	0.061729	0.200000	0.263643	0.319971	0.508357
ndvi_nw	517.0	0.238783	0.076751	0.035860	0.179540	0.232971	0.293929	0.454429
ndvi_se	517.0	0.250126	0.077354	0.029880	0.194743	0.249800	0.302300	0.538314
ndvi_sw	517.0	0.266779	0.086345	0.064183	0.204129	0.262143	0.325150	0.546017
precipitation_amt_mm	516.0	64.245736	35.218995	0.000000	39.105000	60.470000	85.757500	210.830000
reanalysis_air_temp_k	516.0	297.869538	1.170997	294.635714	297.092500	297.822857	298.649286	301.637143
reanalysis_avg_temp_k	516.0	299.133043	1.332073	294.892857	298.221429	299.121429	300.123214	302.928571
reanalysis_dew_point_temp_k	516.0	295.492982	1.417229	290.088571	294.593929	295.852143	296.548571	298.450000
reanalysis_max_air_temp_k	516.0	307.082752	2.382980	300.000000	305.200000	307.050000	308.700000	314.000000
reanalysis_min_air_temp_k	516.0	292.866667	1.663069	286.900000	291.975000	293.050000	294.200000	296.000000
reanalysis_precip_amt_kg_per_m2	516.0	57.609864	50.286555	0.000000	24.065000	46.440000	71.072500	362.030000
reanalysis_relative_humidity_percent	516.0	88.639117	7.583889	57.787143	84.295000	90.917143	94.563929	98.610000
reanalysis_sat_precip_amt_mm	516.0	64.245736	35.218995	0.000000	39.105000	60.470000	85.757500	210.830000
reanalysis_specific_humidity_g_per_kg	516.0	17.096110	1.445769	12.111429	16.102857	17.428571	18.180357	20.461429
reanalysis_tdtr_k	516.0	9.206783	2.448525	3.714286	7.371429	8.964286	11.014286	16.028571
station_avg_temp_c	483.0	27.530933	0.921769	21.400000	27.000000	27.600000	28.100000	30.800000
station_diur_temp_rng_c	483.0	10.566197	1.535496	5.200000	9.500000	10.625000	11.655000	15.800000
station_max_temp_c	506.0	34.004545	1.325261	30.100000	33.200000	34.000000	34.900000	42.200000
station_min_temp_c	512.0	21.196680	1.260327	14.700000	20.600000	21.300000	22.000000	24.200000
station_precip_mm	504.0	62.467262	63.245958	0.000000	17.200000	45.300000	85.950000	543.300000
total_cases	520.0	7.565385	10.765478	0.000000	1.000000	5.000000	9.000000	116.000000

- Vegetation more equally distributed in Iquitos
- The area with the least vegetation was the north-west of the city
- The mean average Temperature in Iquitos is around 27 degrees Celsius
- The station's precipitation was the most spreaded variable since it had the highest standard deviation as well
- The precipitation measures were followed by the total number of cases which had a standard deviation of 10

Missing Values Overall

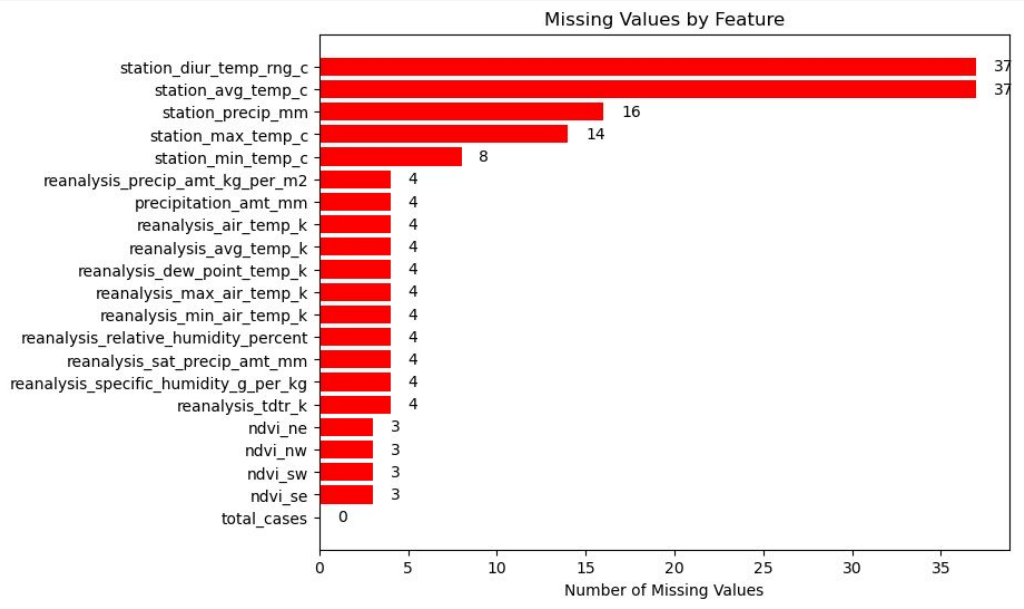
- By checking the missing values in the dataset, we expect the major preprocessing step to be imputation
- Whether replacing them with means, medians or other values will decided based on based on their distribution, stationarity check and model performance

Missing Values for San Juan



- The majority of missing values for San Juan are found within the four vegetation indices
- Notably found in the vegetation index for the north-east of the city

Missing Values for Iquitos



- The majority of missing values for San Juan comes from weather condition variables, such as its temperature and precipitation measure
- Notably in the diurnal temperature range and the average temperature

Stationarity Check for All Variables

sj_stationary

```
{'ndvi_ne': True,
 'ndvi_nw': True,
 'ndvi_se': True,
 'ndvi_sw': True,
 'precipitation_amt_mm': True,
 'reanalysis_air_temp_k': True,
 'reanalysis_avg_temp_k': True,
 'reanalysis_dew_point_temp_k': True,
 'reanalysis_max_air_temp_k': True,
 'reanalysis_min_air_temp_k': True,
 'reanalysis_precip_amt_kg_per_m2': True,
 'reanalysis_relative_humidity_percent': True,
 'reanalysis_sat_precip_amt_mm': True,
 'reanalysis_specific_humidity_g_per_kg': True,
 'reanalysis_tdtr_k': True,
 'station_avg_temp_c': True,
 'station_diur_temp_rng_c': True,
 'station_max_temp_c': True,
 'station_min_temp_c': True,
 'station_precip_mm': True,
 'total_cases': True}
```

iq_stationary

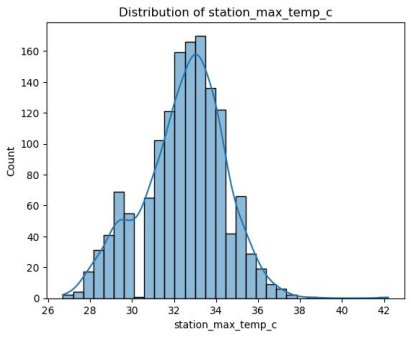
```
{'ndvi_ne': True,
 'ndvi_nw': True,
 'ndvi_se': True,
 'ndvi_sw': True,
 'precipitation_amt_mm': True,
 'reanalysis_air_temp_k': True,
 'reanalysis_avg_temp_k': True,
 'reanalysis_dew_point_temp_k': True,
 'reanalysis_max_air_temp_k': True,
 'reanalysis_min_air_temp_k': True,
 'reanalysis_precip_amt_kg_per_m2': True,
 'reanalysis_relative_humidity_percent': True,
 'reanalysis_sat_precip_amt_mm': True,
 'reanalysis_specific_humidity_g_per_kg': True,
 'reanalysis_tdtr_k': True,
 'station_avg_temp_c': True,
 'station_diur_temp_rng_c': True,
 'station_max_temp_c': True,
 'station_min_temp_c': True,
 'station_precip_mm': True,
 'total_cases': True}
```

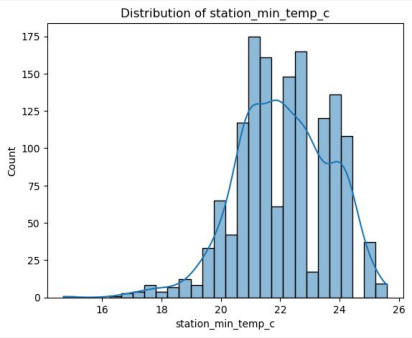
We can proceed with mean & mode imputation

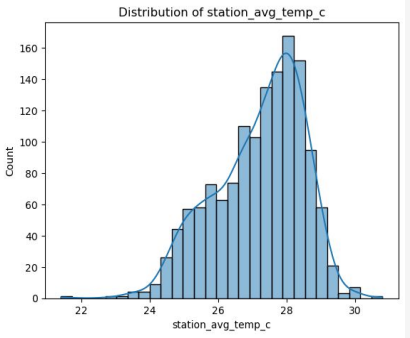
Distributions

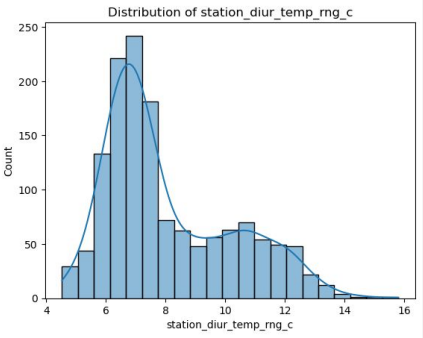
- Visualizations were conducted to study the distribution of each predictor
- When looking at the distributions, we see there are two type of distributions: quasi-normalized distribution with different kurtosis and right skewed distributions.
- Kurtosis measures the “tailedness” of a predictor’s distribution
- Tailedness, or having tails in a certain form, is usually how outliers occur
    - A sharp peak suggests high kurtosis (leptokurtic), implying many values are close to the mean with some notable outliers.
    - A flatter histogram suggests low kurtosis (platykurtic), meaning the values are generally far from the mean.
- Distributions help us better understand the seasonality and extremity of the predictors

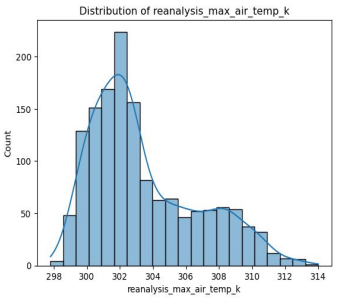
Distributions: Temperature Predictors

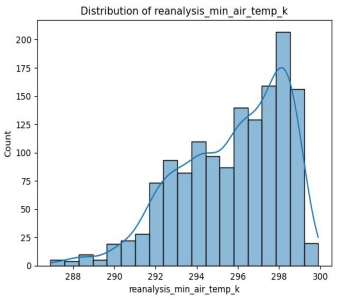


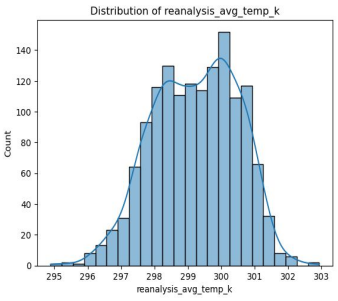


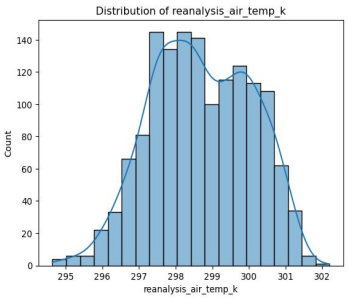


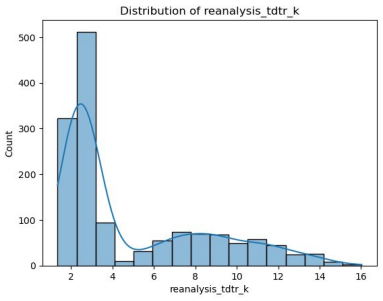




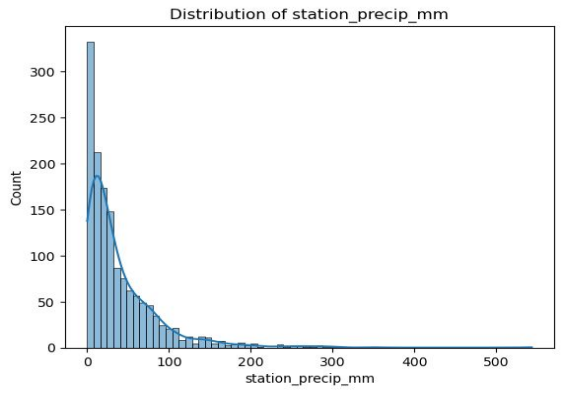


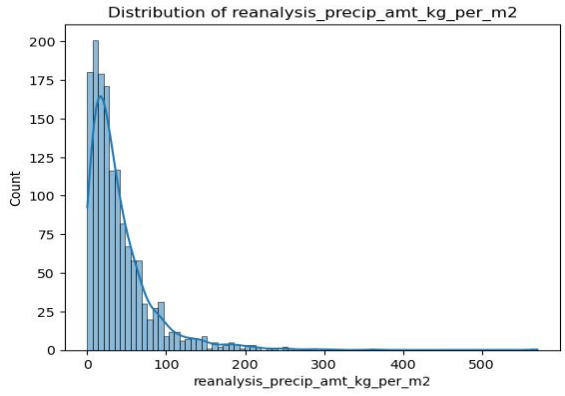


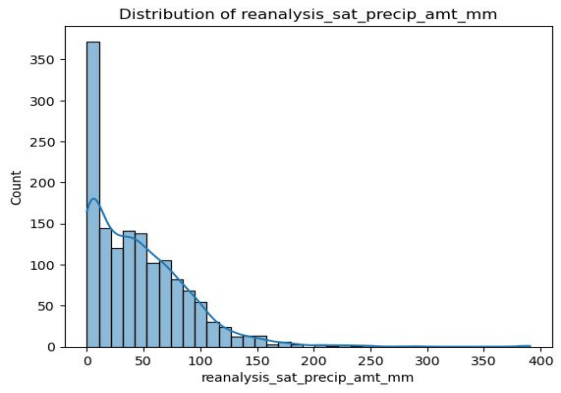


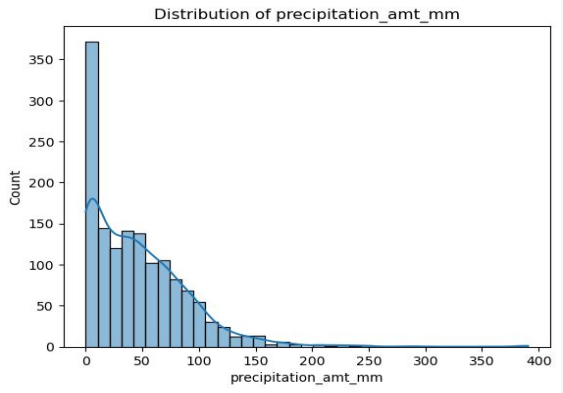


Distributions: Precipitation Predictors

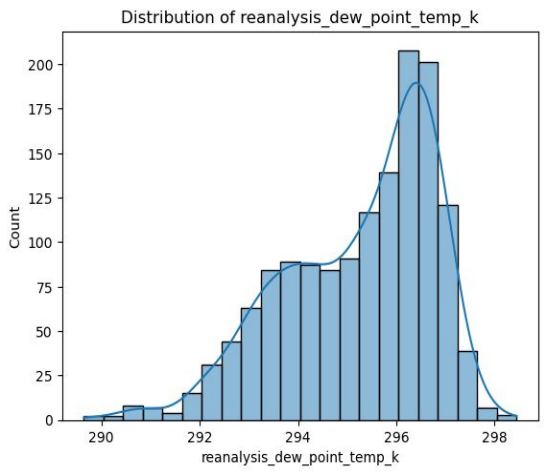


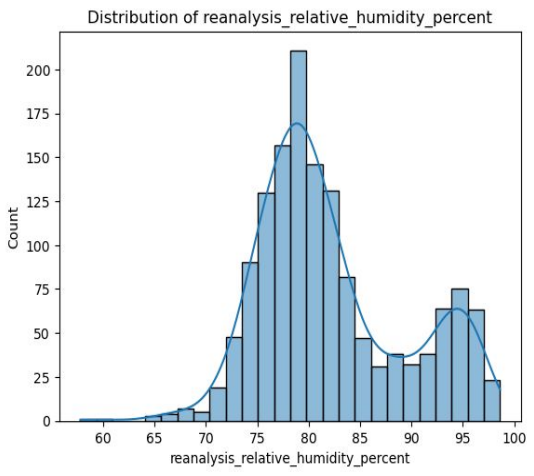


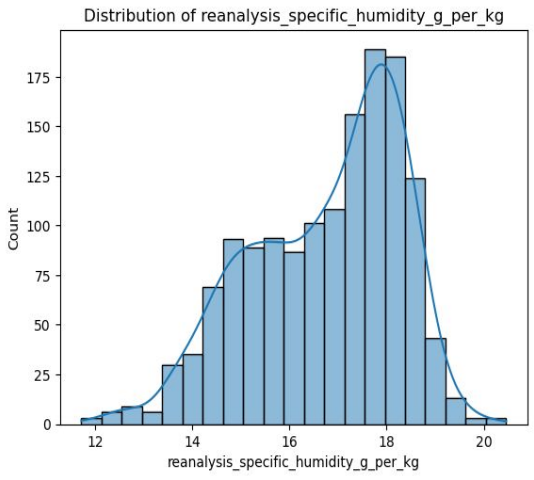




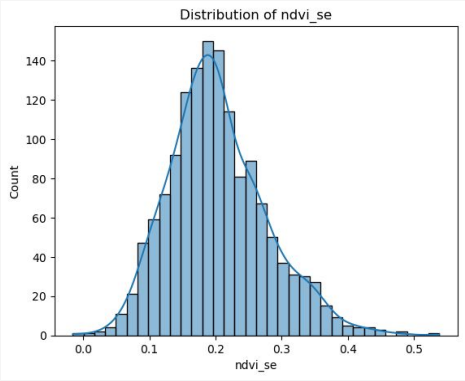
Distributions: Humidity Predictors

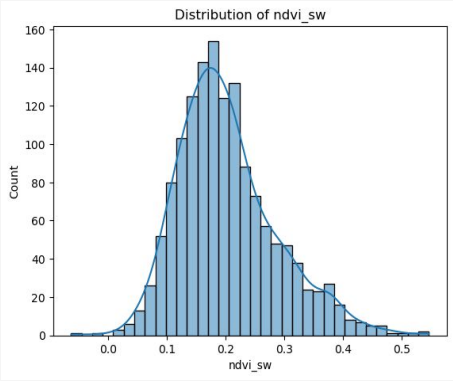


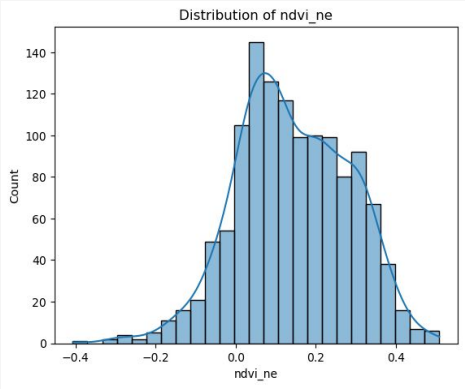


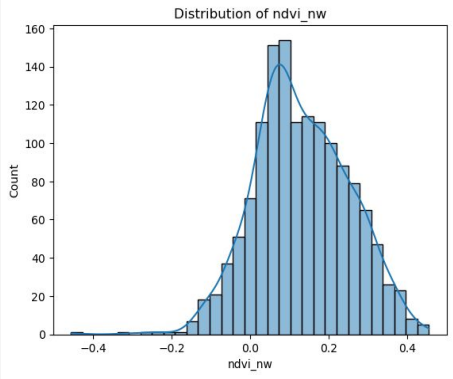


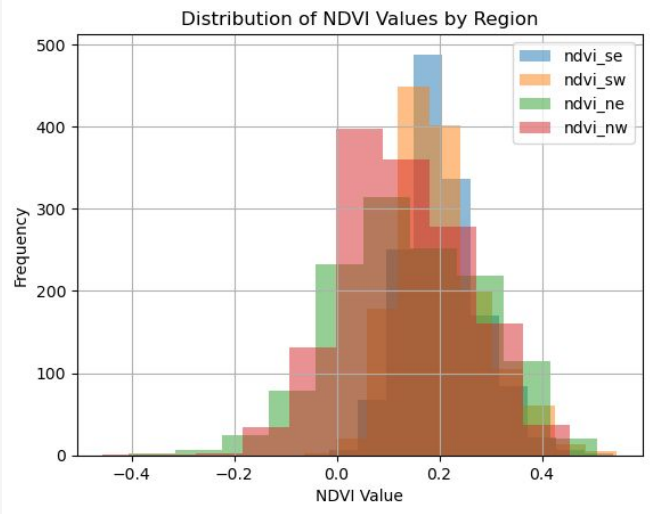
Distributions: Vegetation Predictors



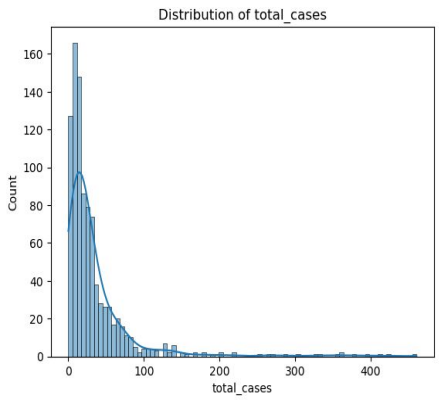


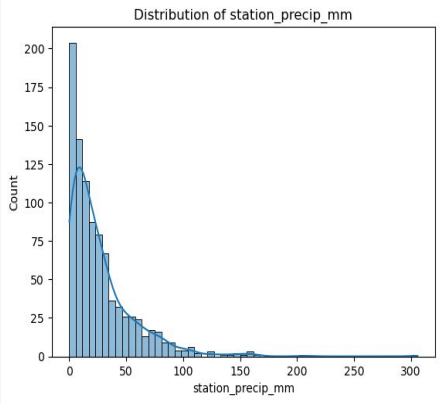


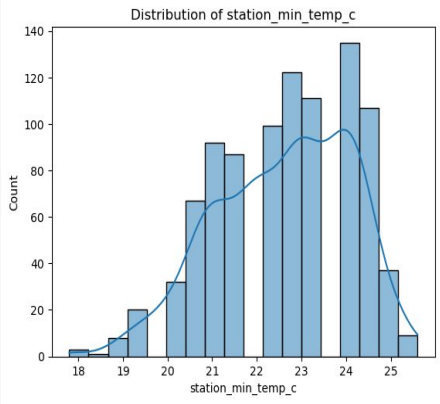


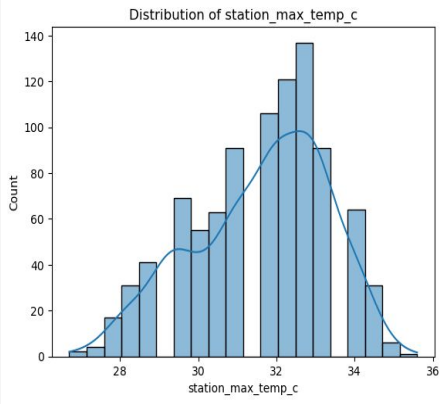


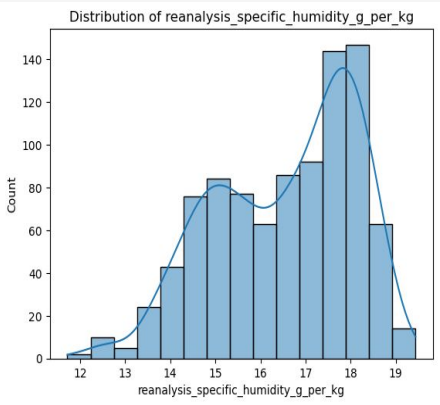
Distributions of “SJ” City

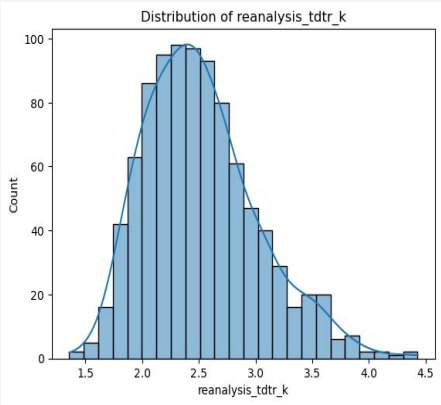


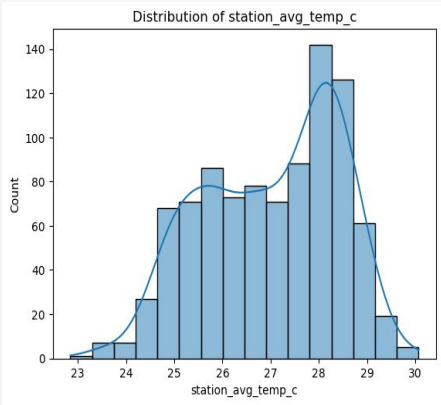


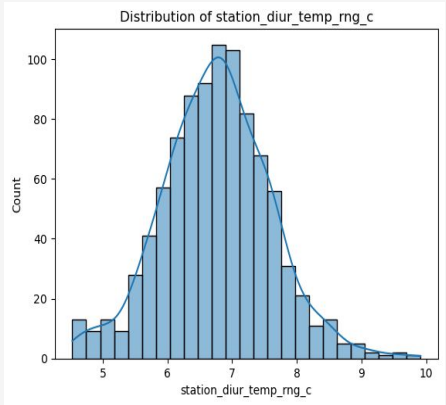




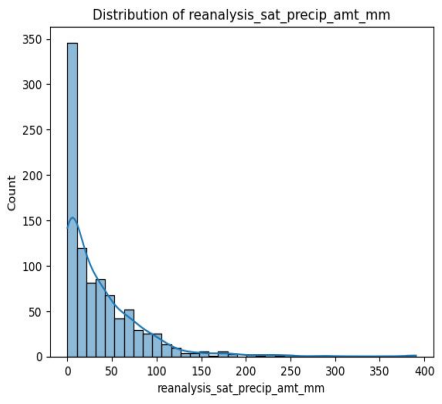


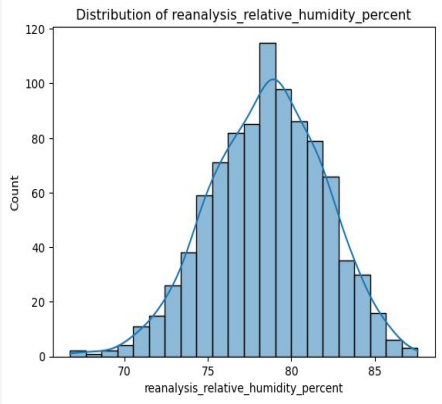


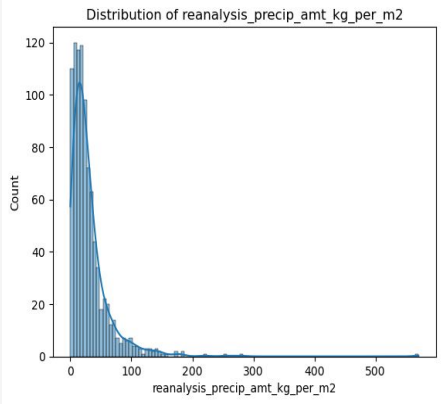


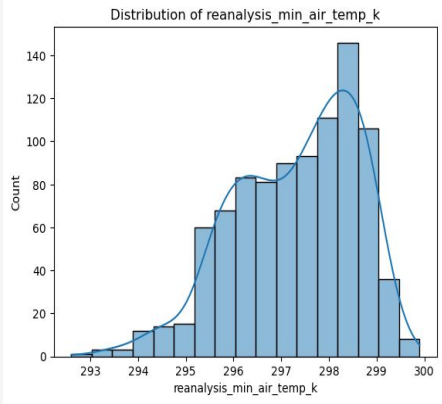


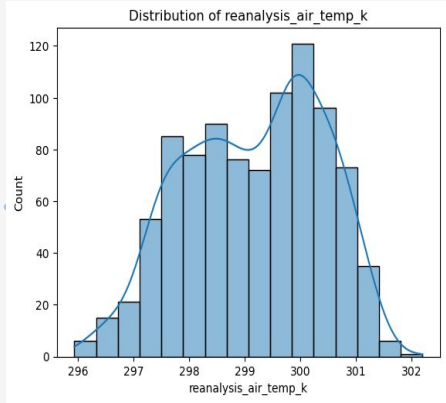
Distributions of “SJ” City

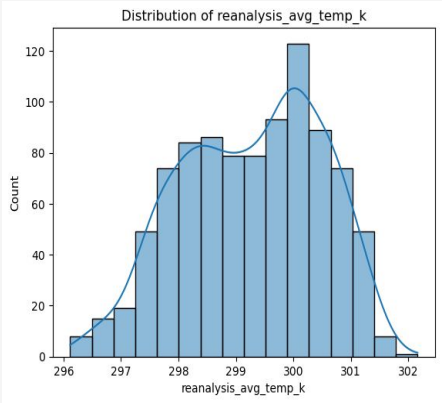


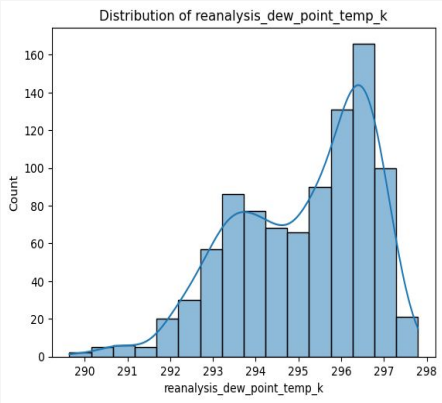


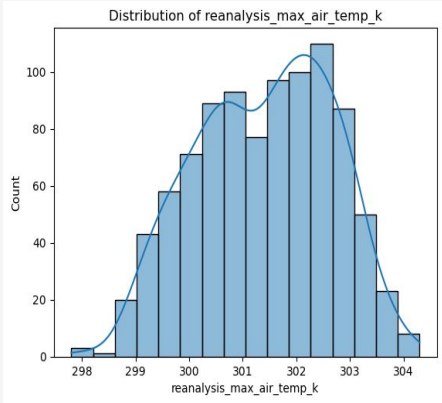




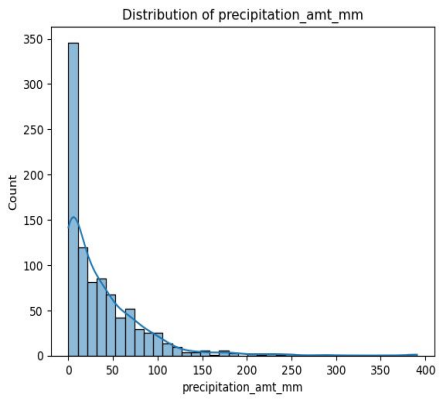


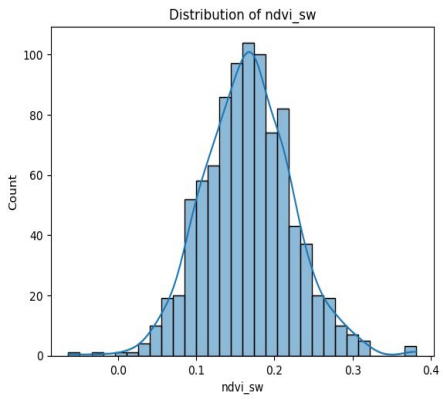


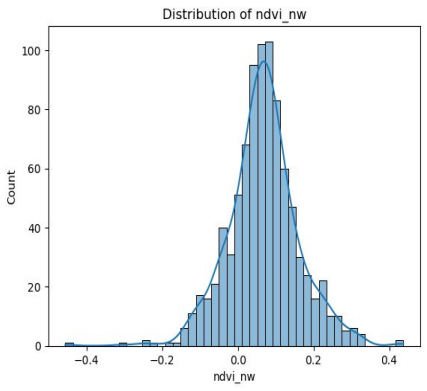


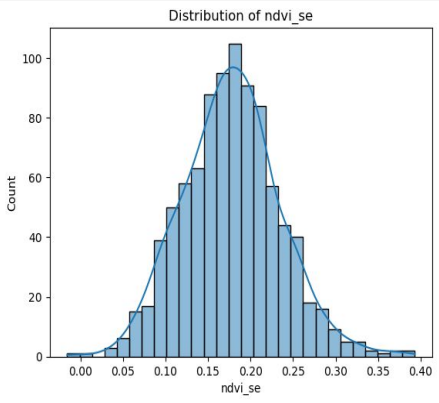


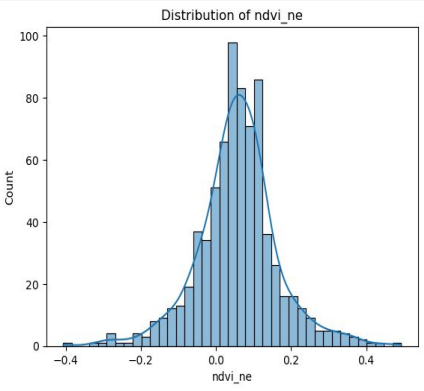
Distributions of “SJ” City



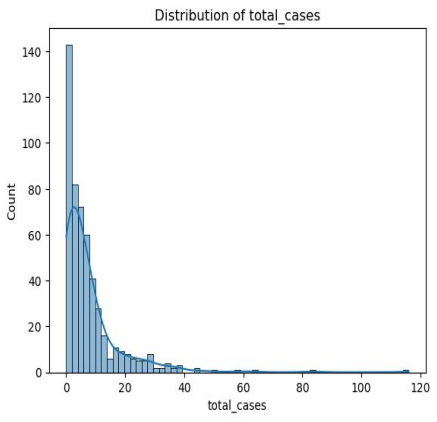


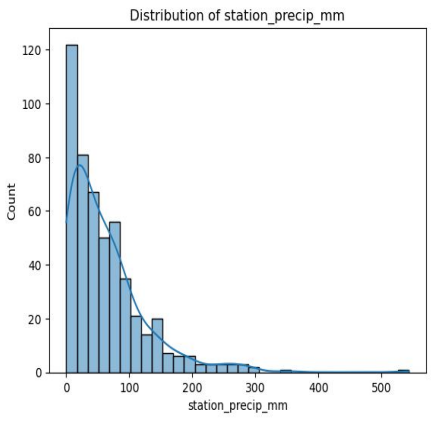


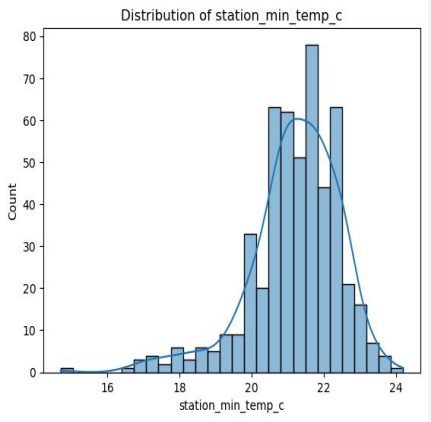


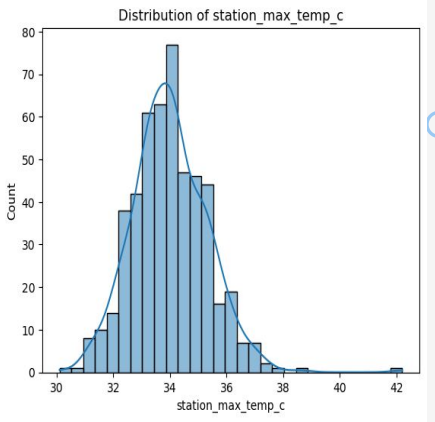


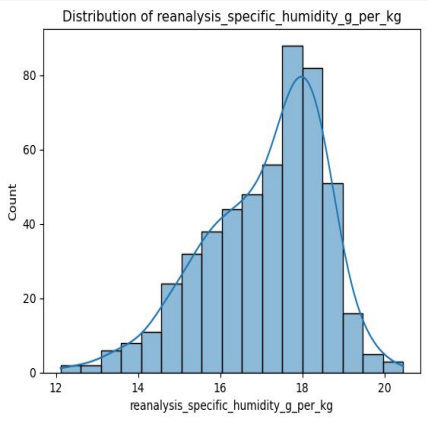
Distributions of “IQ” City

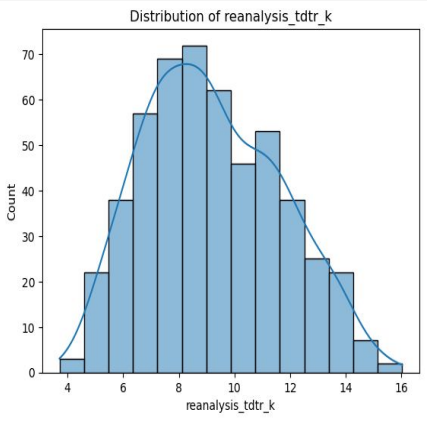


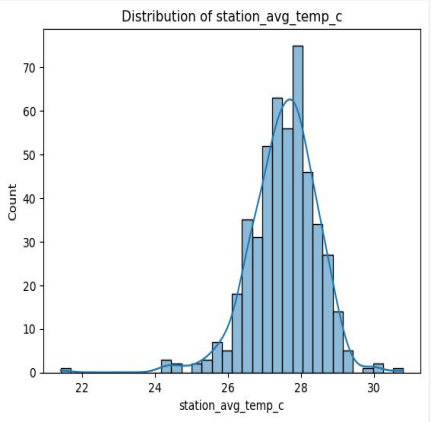


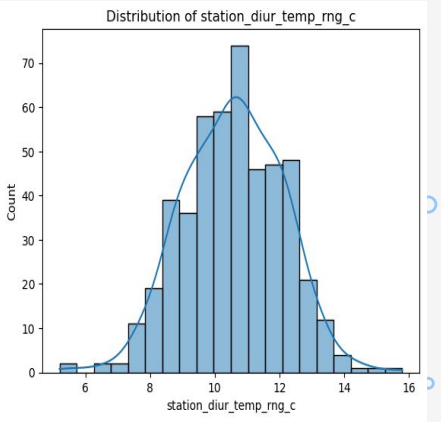




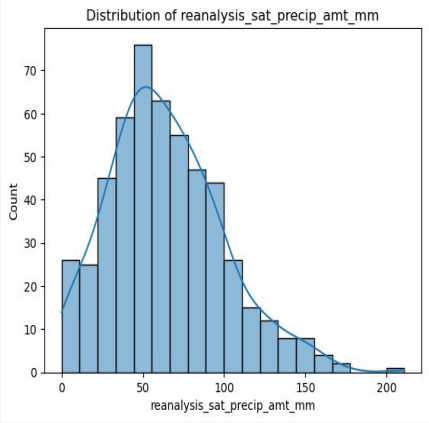


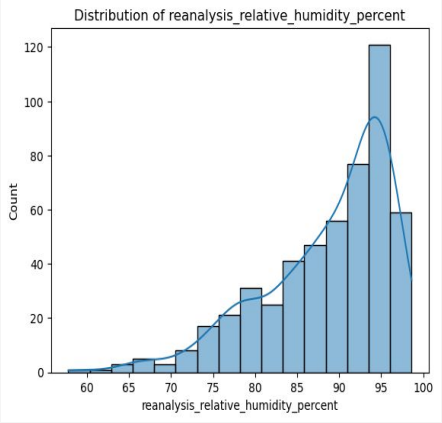


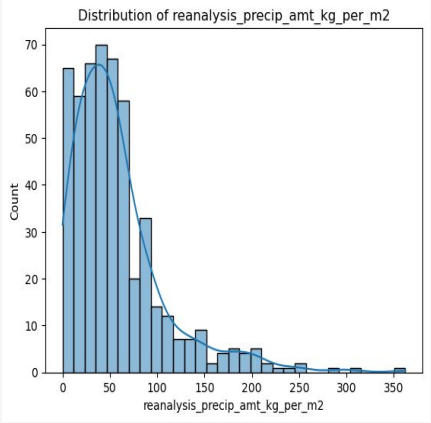


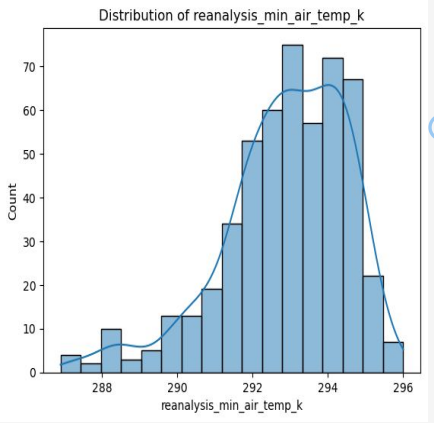


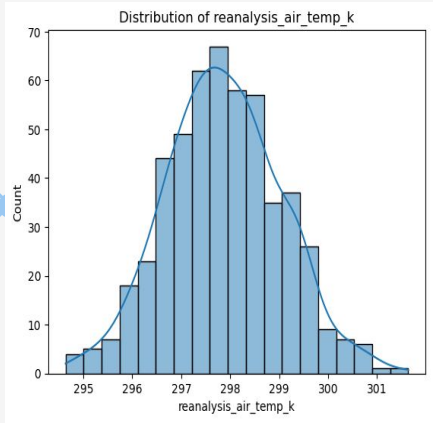
Distributions of “IQ” City

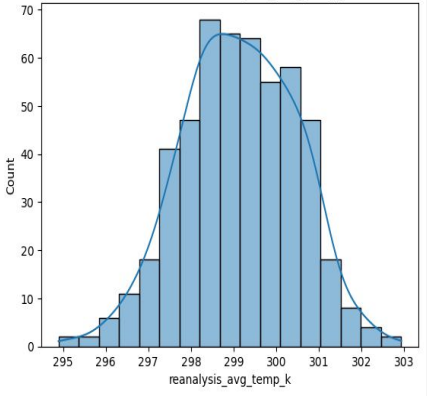


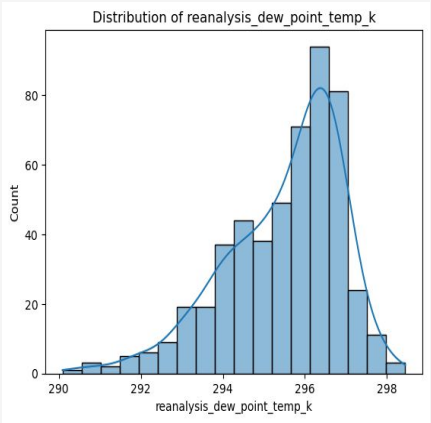


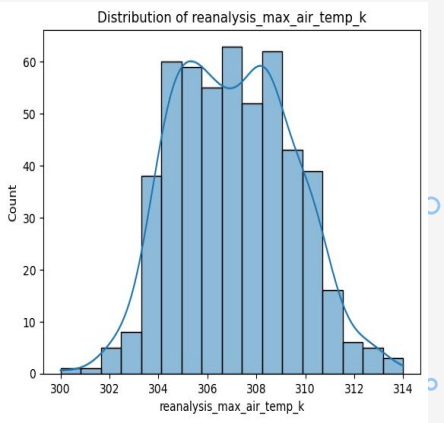




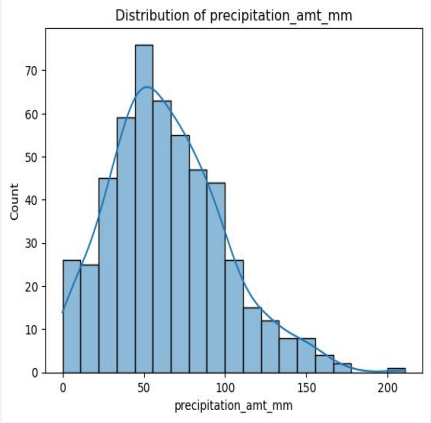


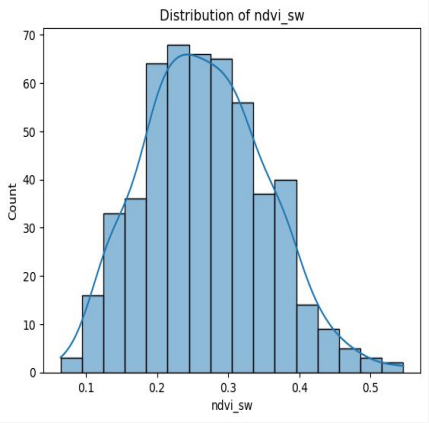


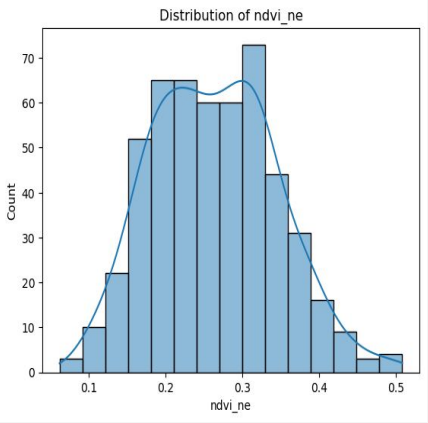


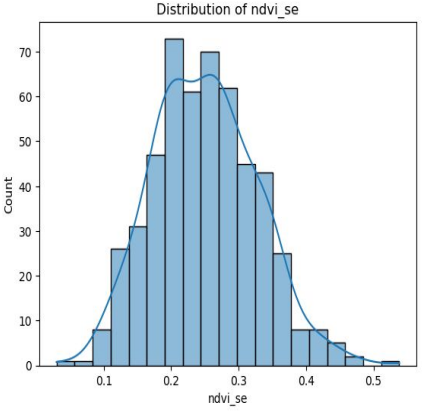


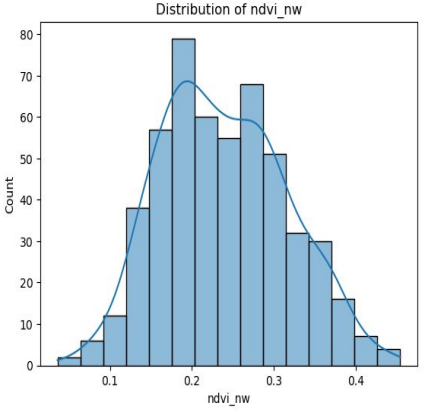
Distributions of “IQ” City



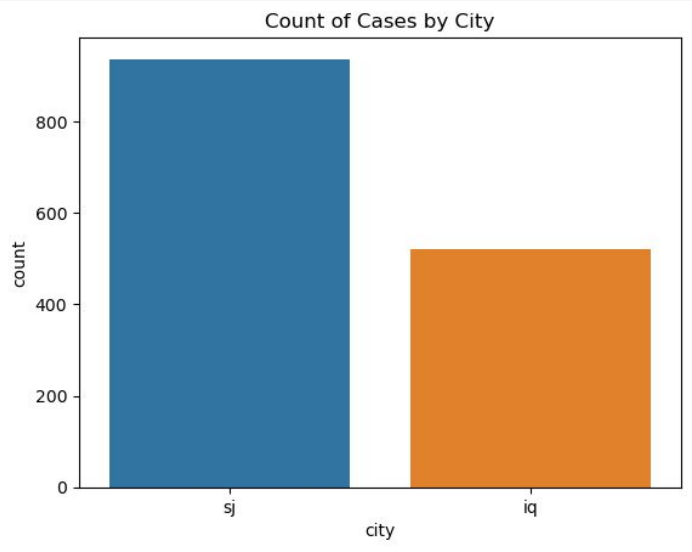


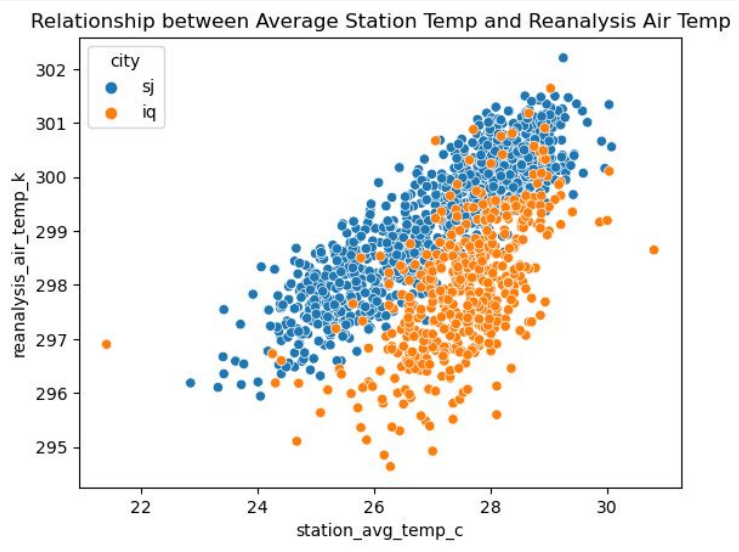






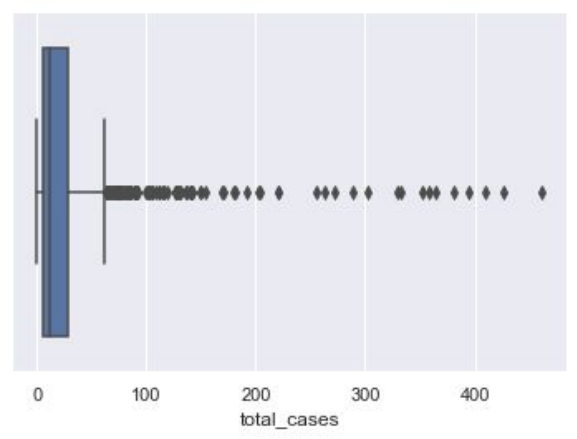
Histogram & Scatter Plot





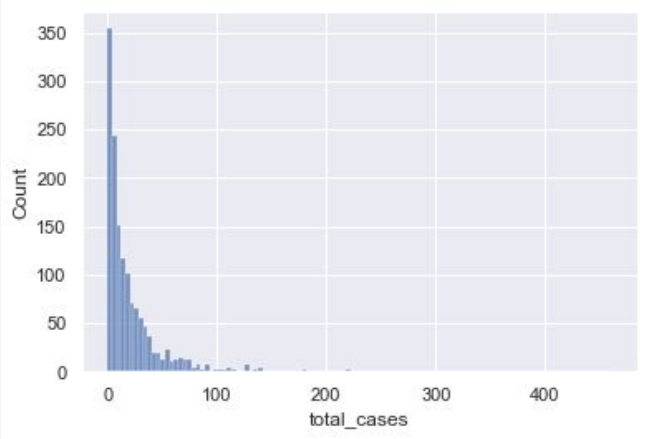
As we can tell from the scatter plot, San Juan has a higher value in temperature factors in general; and from the histogram, San Juan has more disease cases. We assume that there's a positive relationship between the temperature predictors and the target variable total_cases.

Total Cases Histogram & Boxplot



Even though there are not many outliers, their prediction has a huge impact on model performance.

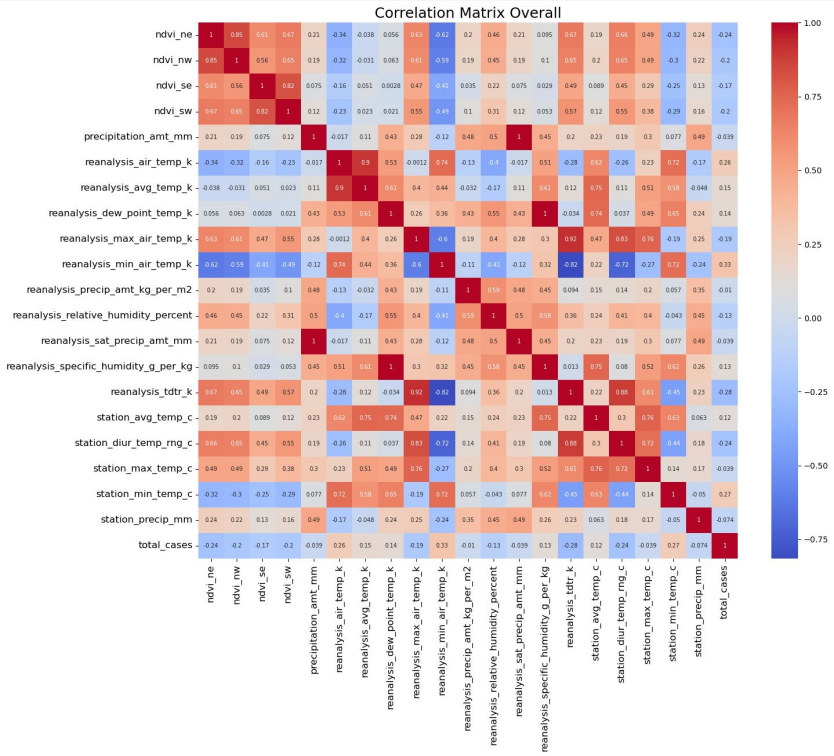
In this case, the outliers are important instances that show the spikes in weekly dengue fever cases.



Shows the total number of dengue fever cases each week.

Indicates that the majority of dengue fever cases are below 100 cases per week.

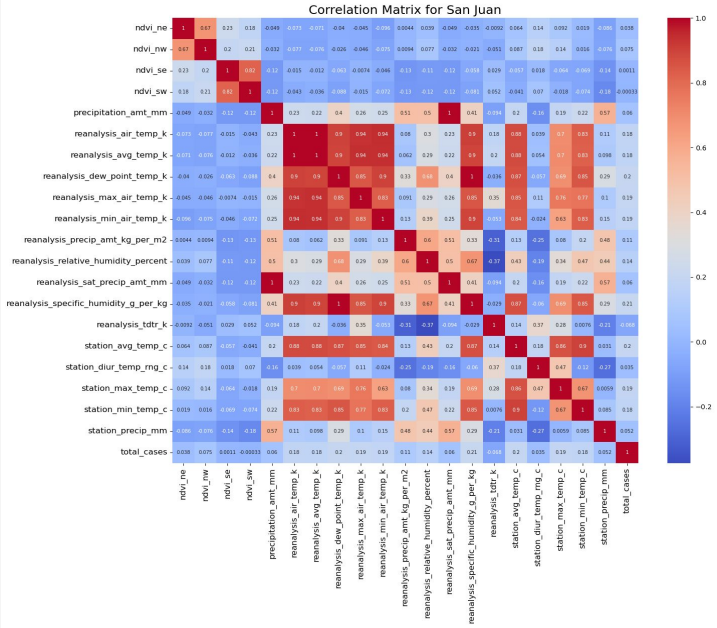
Correlation Heat Map

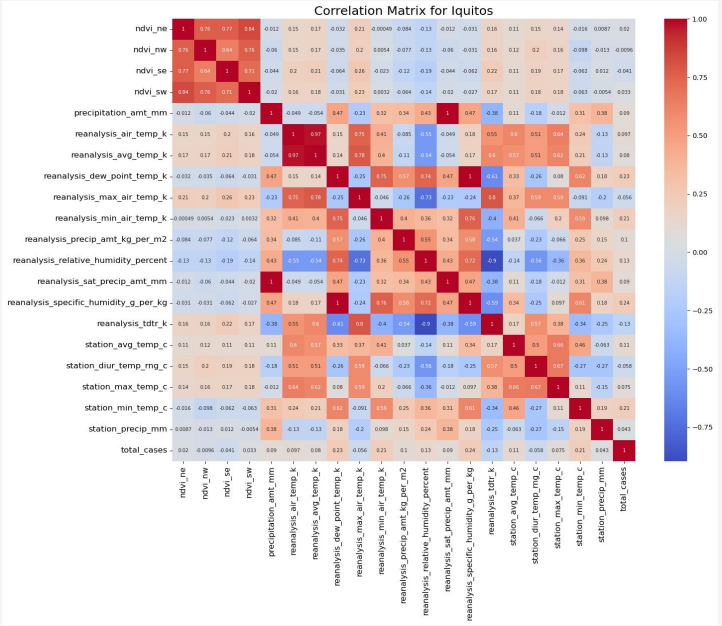


This correlation heat map visualizes the strength of the correlations between each variable in the dataset.

By looking at the correlation between total cases and each predictor, we can see the strength that each predictor has with the number of cases of dengue fever. The strongest were the air temperatures (minimum and average) and the precipitation levels.

Correlation Heat Map by city



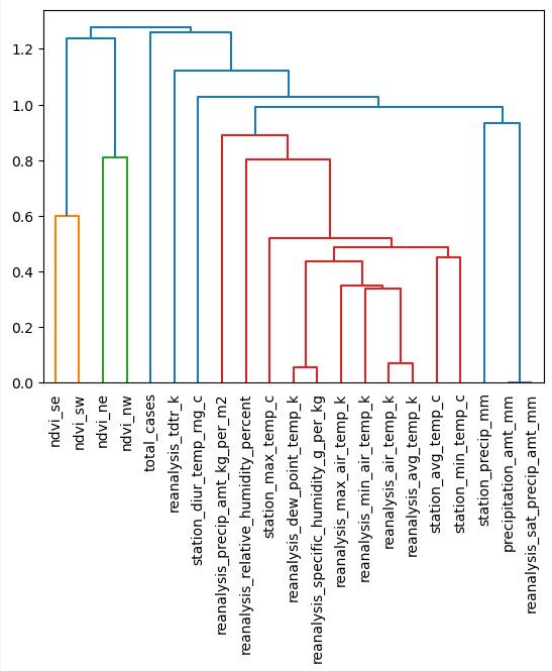


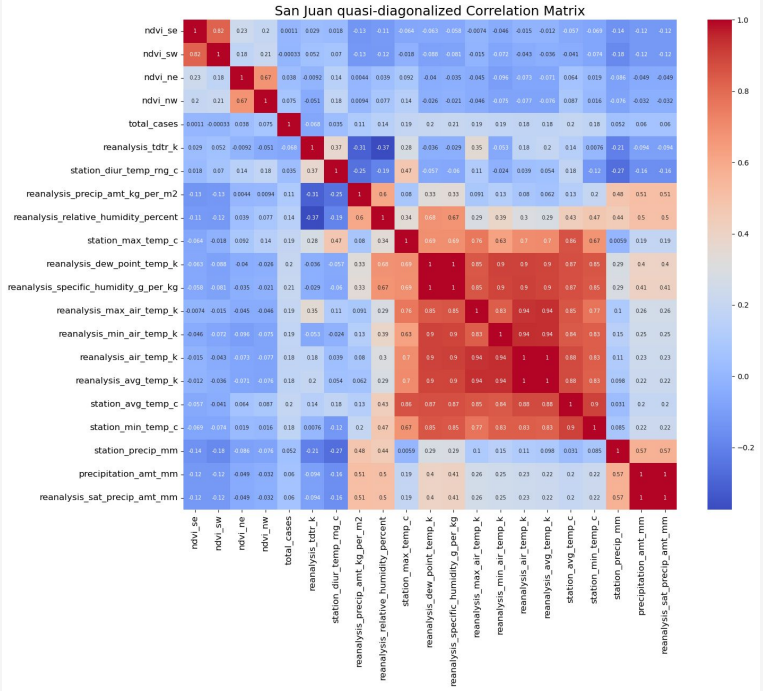
- High correlation between temperature metrics, high correlation between temperature and humidity
- Corr(reanalysis_specific_humidity_g_per_kg, temperature metrics) very high, indicating a hot humid climate, which may accelerate the spread of Dengue

- High correlation among vegetation predictors and among temperature predictors
- High correlation between humidity and precipitation.
- Total case is slightly more correlated to temperature and humidity

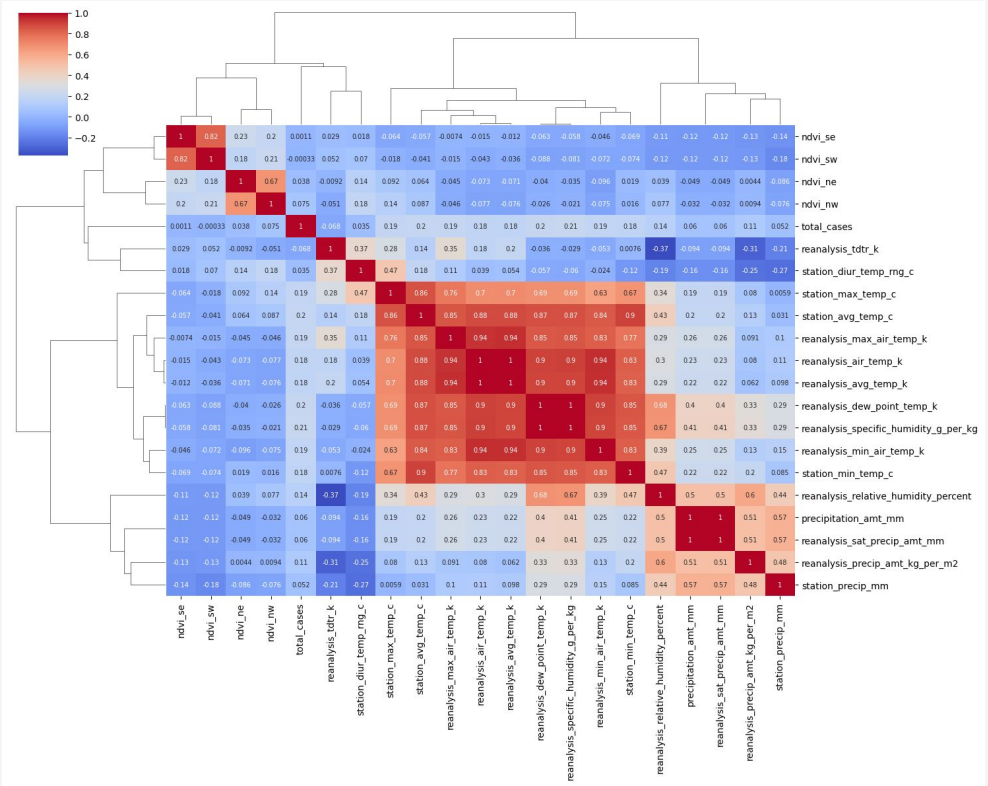
Quasi-Diagonalized Correlation Matrix: SJ

- To make the map more readable, we transformed it into a quasi-diagonalized matrix
- This means the strongest correlations follow the diagonal of the matrix



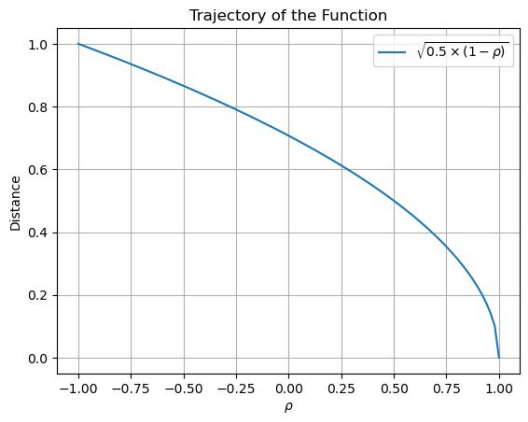


Quasi-Diagonalized Correlation Matrix: SJ



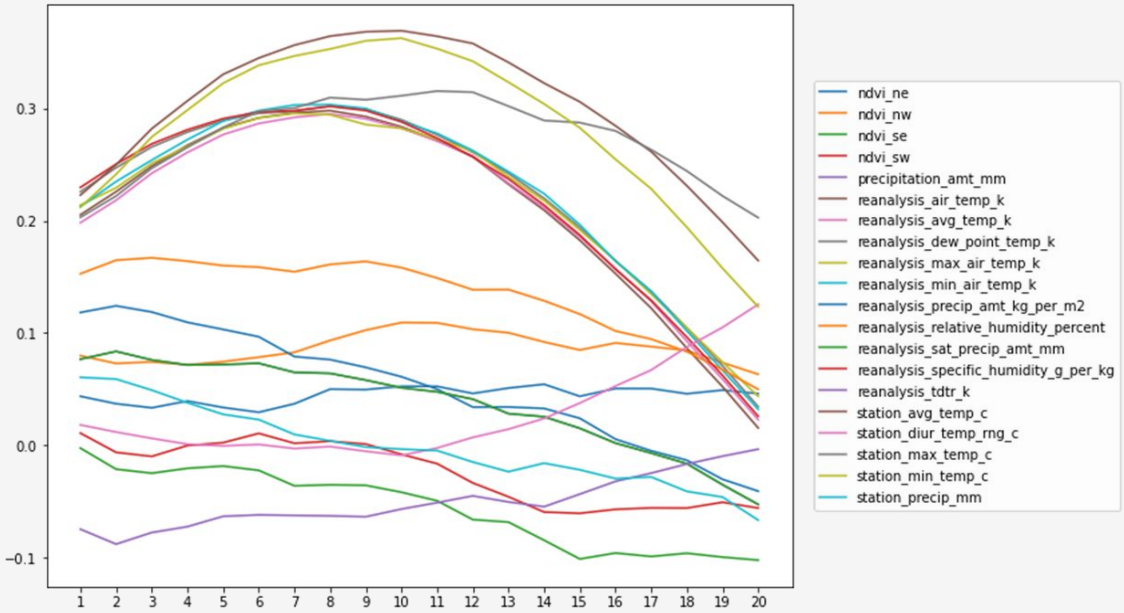
- Inter-predictor Distance Calculate based on correlation coefficient as

$$D(i,j) = \sqrt{0.5 * (1 - \rho(i,j))}$$



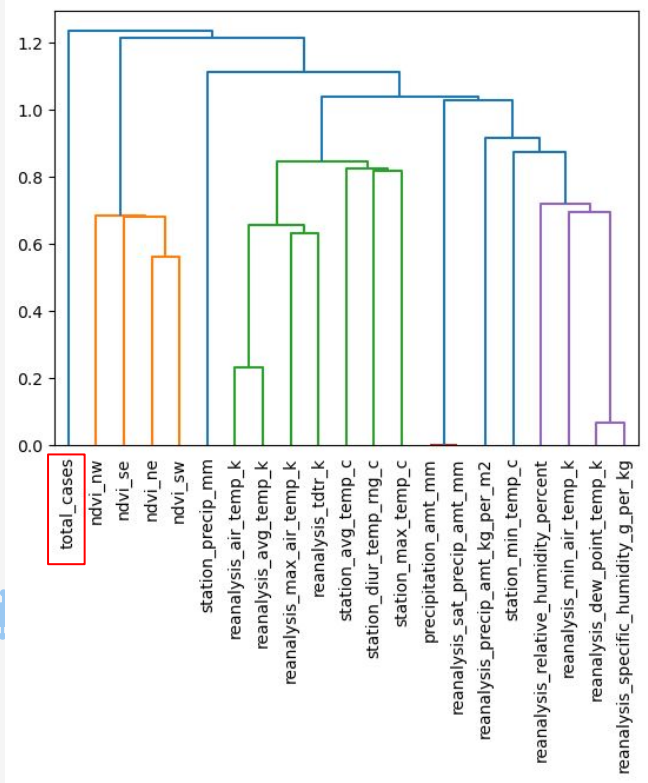
- Hierarchical clustering performed based on the distances and sort the order

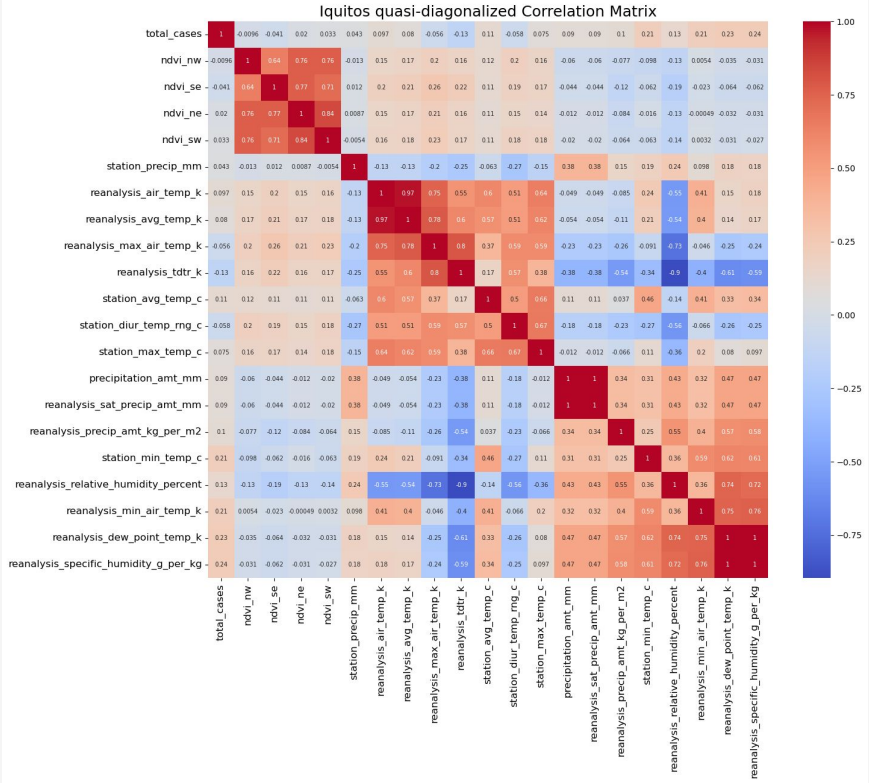
Correlation Time Lead Analysis: San Juan



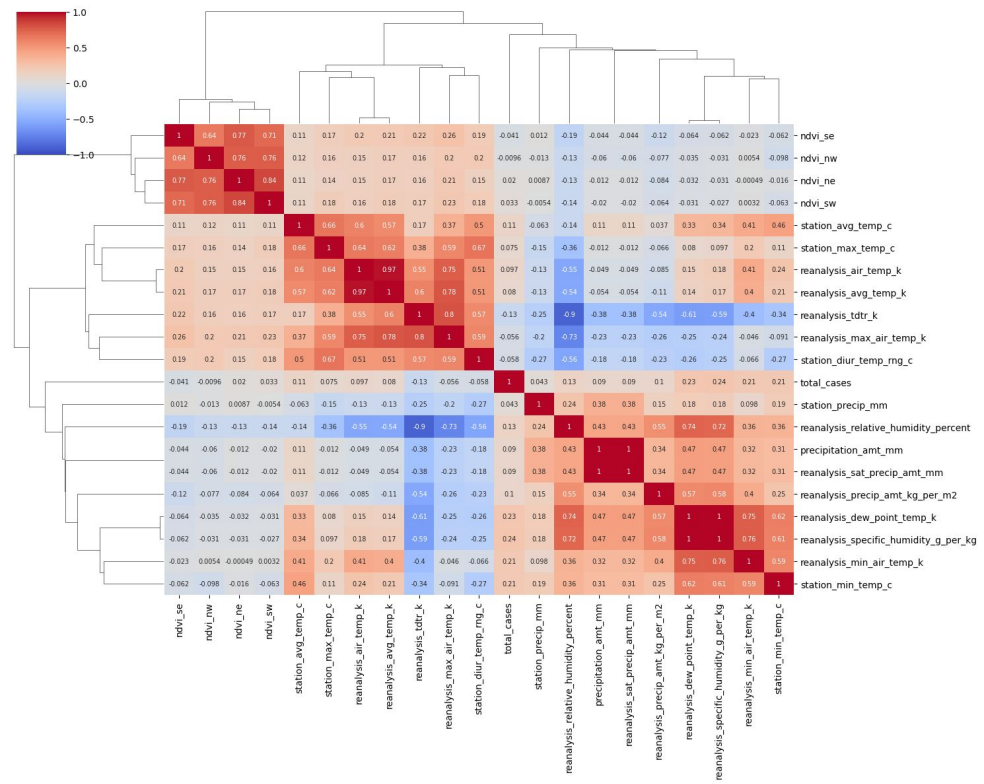
- For SJ, the strongest correlation of features with total cases occur at lead = 10 and then diminishes as lag increases.
- Effect from temperature/humidity change takes around 2 months to set in
- For station_diur_temp_rng_c, the strongest correlation occurs at lead = 20 weeks
- The temperature cluster is the same one from diagonal of the correlation matrix, very strong signal in the data

Quasi-Diagonalized Correlation Matrix: IQ



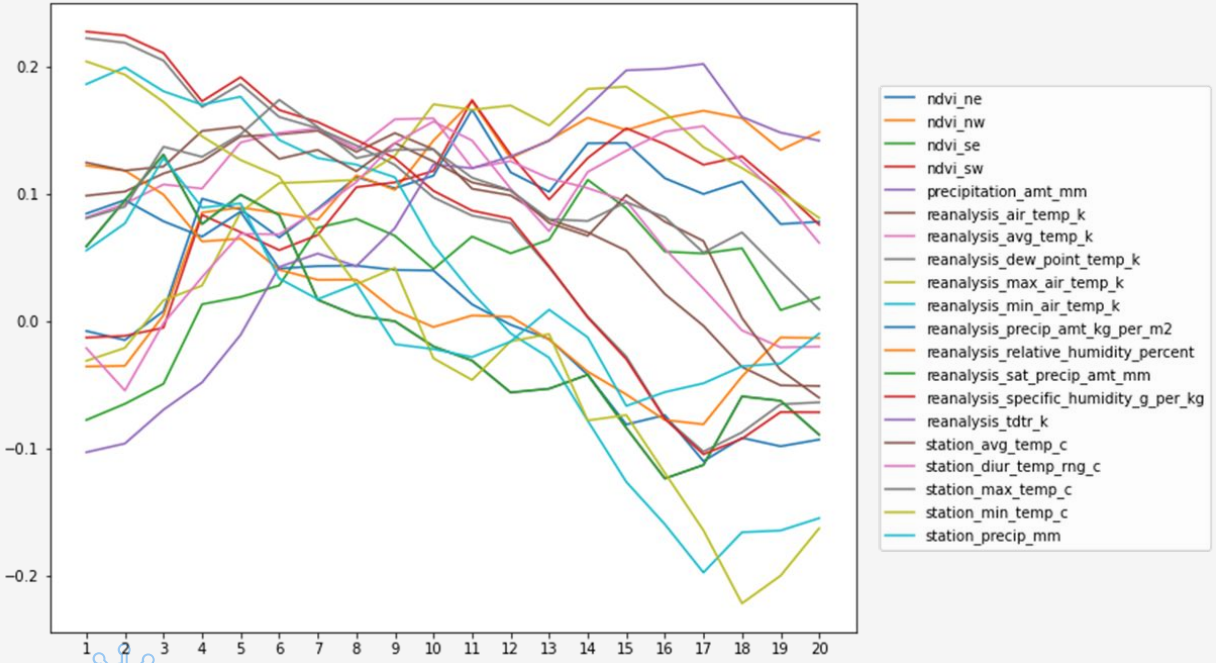


Quasi-Diagonalized Correlation Matrix: IQ



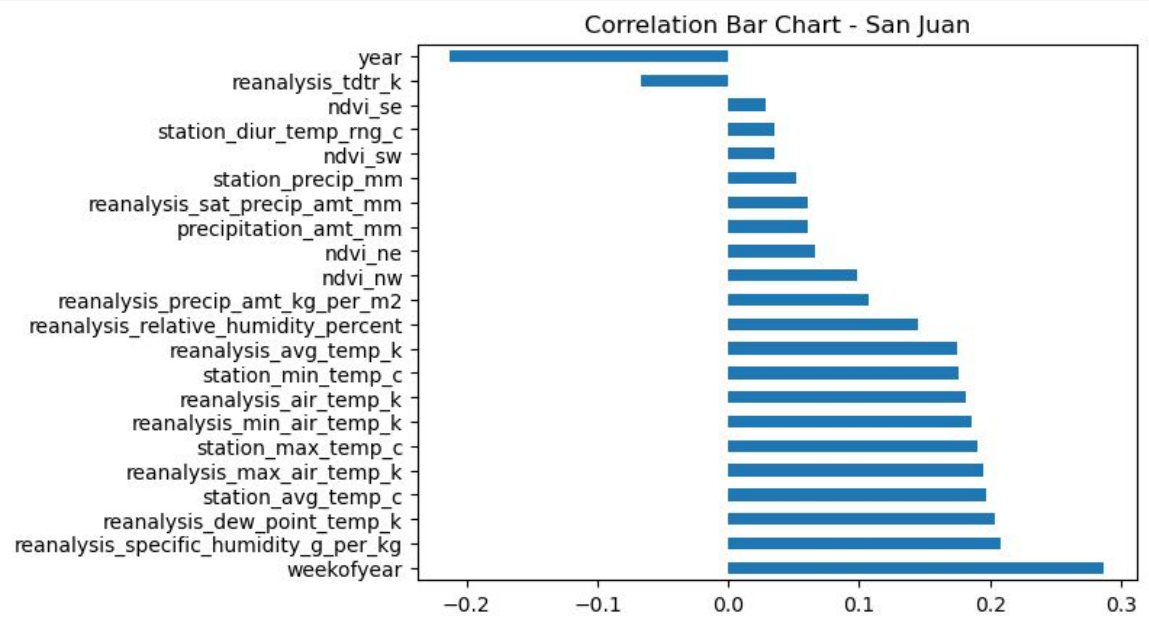
- Seaborn-embedded function plots the matrix slightly differently than the manually made one
- The clusters remain more or less the same
- In the manually made correlation matrix, total case is a separate cluster at the highest level
- In the seaborn matrix, total cases is closest to the precipitation cluster, confirming the previous finding that total case has similar distribution to precipitation
- However, the diagonal is less concentrated less that of San Juan, meaning more noise

Correlation Time Lead Analysis: Iquitos



- For Iquitos, the pattern is less obvious, indicating abundance of noise
- There is a peak at lead = 11 for some vegetation metrics

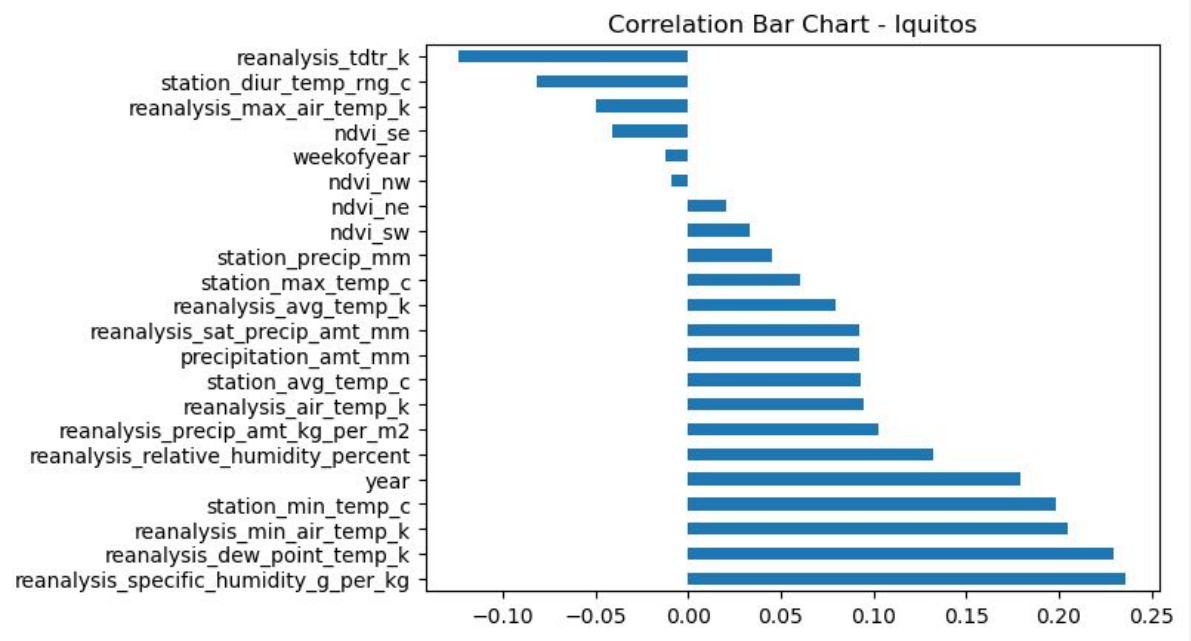
Correlation of predictors to target variable



We see here that the variables correlating with the target variable for San Juan include:

- Diurnal temperature
- Minimum Temperature (From Weather Station)
- Minimum Air Temperature
- Dew point Temperature
- Specific Humidity
- Year
- Week of the year

Correlation of predictors to target variable



We see here that the variables correlating with the target variable for Iquitos include:

- Diurnal temperature
- Minimum Temperature (From Weather Station)
- Minimum Air Temperature
- Dew point Temperature
- Specific Humidity
- year

04

Model

Model Building and Testing

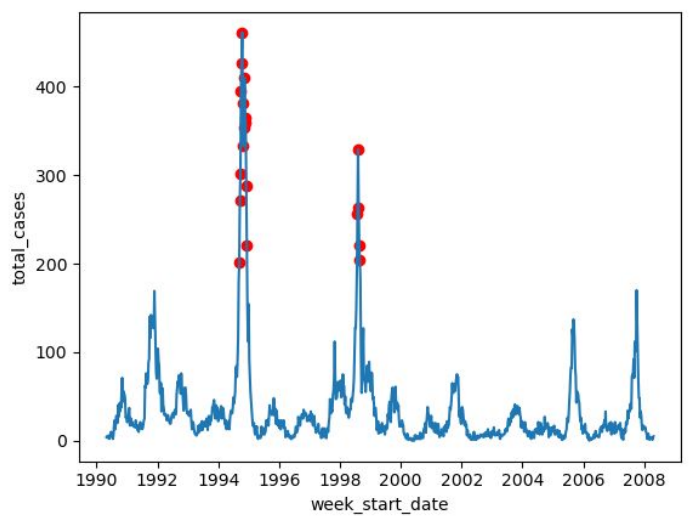
Time Series Data Modeling

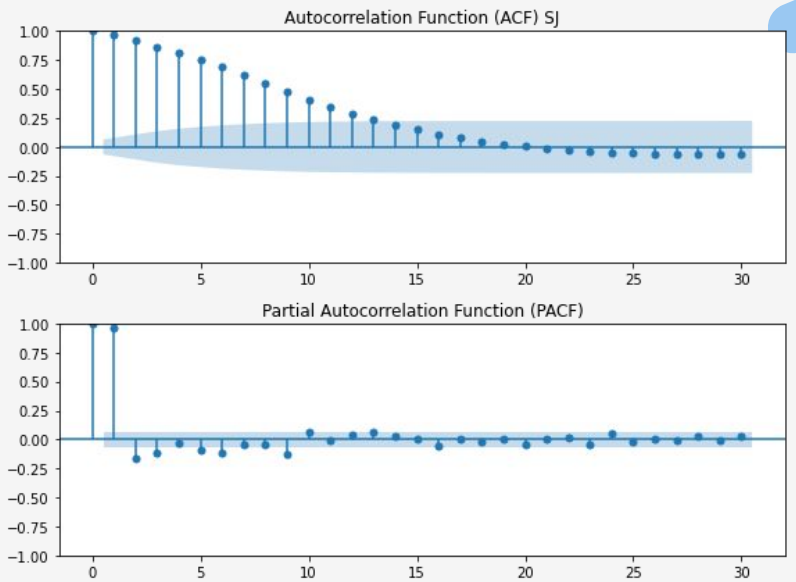
- ARIMA and SARIMA are time series forecasting models used to predict future values based on past observations.

- ARIMA (Autoregressive Integrated Moving Average) and SARIMA (Seasonal-ARIMA) require **Stationarity**

- The parameters of the ARIMA model include:
    - autoregressive term (p) - number of lags of predictor (total_cases) to be used as predictors
    - differencing (d) - minimum number of differencing needed to make the series stationary. If the series is already stationary, then d=0
    - moving average term (q) -refers to the number of lagged forecast errors that should go into the ARIMA Model.
- However, ARIMA models are limited in their ability to capture seasonality in the data.
- SARIMA (Seasonal ARIMA) model is able to capture that seasonality in the data by including other exogenous terms

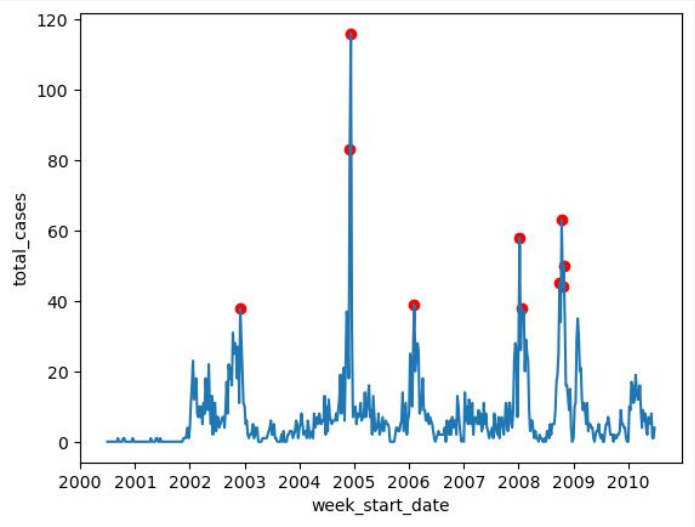
Target Variable Stationarity Check: SJ

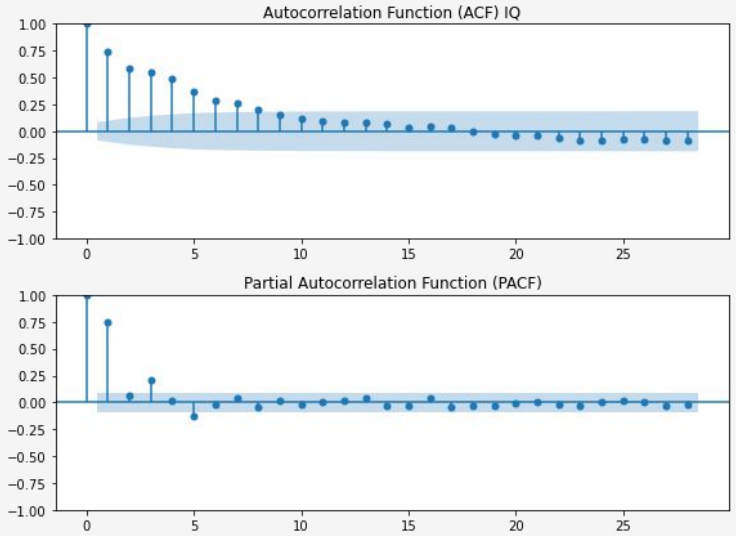




- As shown above, SJ's total cases time series is stationary. Therefore, difference d = 0.
- Its Augmented Dicket-Fuller test has a p value below 0.05.
- From ACF, we can infer that a moving average model with lag = 13 would fit, q = 13.
- From PACF, we can infer that an AutoRegressive model with lag = 9 would fit, p =9

Target Variable Stationarity Check: IQ



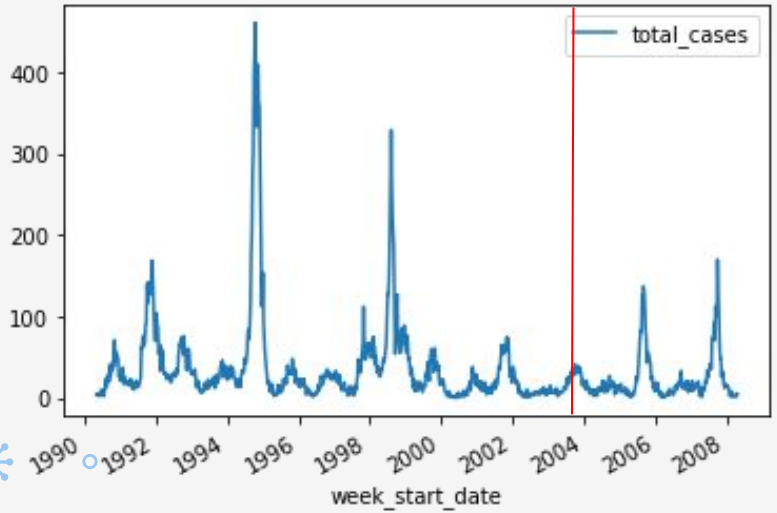


- As shown above, IQ's total cases time series is stationary. Therefore, difference d = 0.
- Its Augmented Dicket-Fuller test has a p value below 0.05.
- From ACF, we can infer that a moving average model with lag = 8 would fit, q = 8.
- From PACF, we can infer that an AutoRegressive model with lag = 4 would fit, p = 4.
- Outliers identified by Isolation Forest in the graph

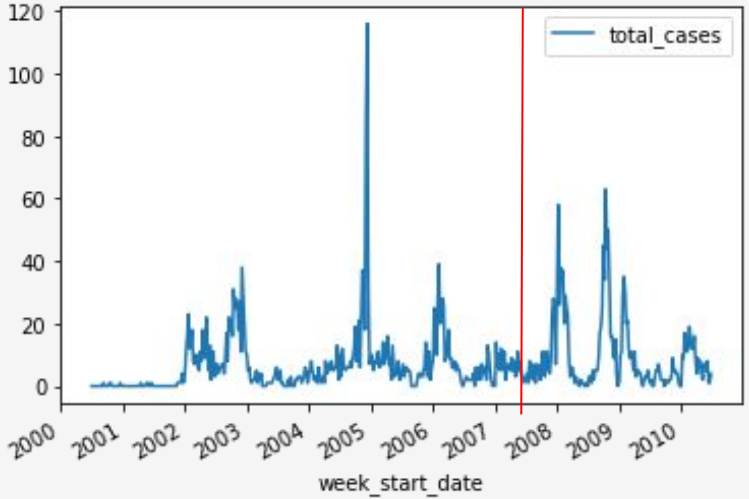
Model Testing

- In order to better test the models, the data was split into training periods and testing periods.

Train Test



Train Test



Ensemble Model

Simple Approach

Weight = [w1, w2, w3, ...]

Predictions = [model1, model2, model3,.....]

Final Prediction = Transpose(Weight) * Predictions = w1*model1+w2*model2+w3*model3+.....

Softmax Approach

The formula for the softmax function $\sigma(x)$ for a vector $x = \{x_0, x_1, \ldots, x_{n-1}\}$ is

$$\sigma(x)_j = \frac{e^{x_j}}{\sum_k e^{x_k}}$$

Predictions = [model1, model2, model3,.....]

Weight = [exp(model1)/Sum(exp(model1),exp(model2),exp(model3).....), exp(model2)/Sum(exp(model1),exp(model2),exp(model3).....), exp(model3)/Sum(exp(model1),exp(model2),exp(model3).....),....]

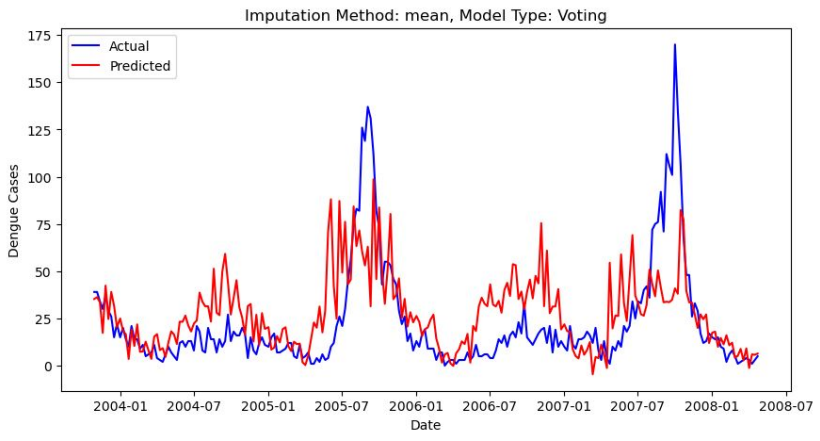
Final Prediction = Transpose(Weight) * Predictions

Model Testing: San Juan

This testing was conducted using Mean and Mode Imputation Superior to see what model yielded the best result.

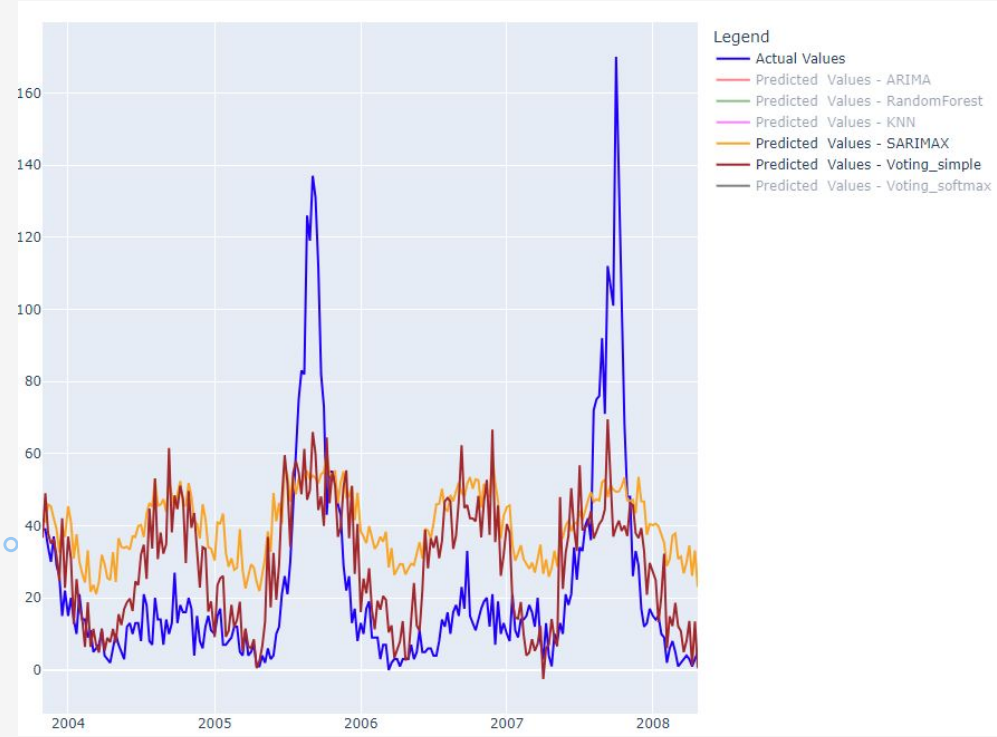
imputation	model	mse	mae	r2
mean	ARIMA	1061.5	28.329	-0.3027
	RandomForest	1693.2	31.587	-1.0779
	KNN	1400.3	27.162	-0.7185
	SARIMAX	785.71	23.435	0.0358
	Voting_simple	618.71	16.346	0.2407
	Voting_softmax	2307	40.712	-1.8312
mode	ARIMA	1061.5	28.329	-0.3027
	RandomForest	1415.9	28.581	-0.7375
	KNN	1395.5	27.049	-0.7126
	SARIMAX	820.25	23.937	-0.0066
	Voting_simple	674.03	17.769	0.1728
	Voting_softmax	2062	38.442	-1.5305
ffill	ARIMA	1061.5	28.329	-0.3027
	RandomForest	1384.1	27.875	-0.6985
	KNN	1400.2	27.176	-0.7183
	SARIMAX	934.3	25.853	-0.1466
	Voting_simple	902.91	24.005	-0.1081
	Voting_softmax	2237.4	40.801	-1.7457
bfill	ARIMA	1061.5	28.329	-0.3027
	RandomForest	1241.8	26.556	-0.524
	KNN	1411.3	27.148	-0.732
	SARIMAX	840.67	24.111	-0.0317
	Voting_simple	784.8	20.624	0.0369
	Voting_softmax	1976	37.833	-1.425
interpolation	ARIMA	1061.5	28.329	-0.3027
	RandomForest	1536.6	30.343	-0.8858
	KNN	1411.4	27.297	-0.7321
	SARIMAX	673.52	19.385	0.1734
	Voting_simple	818.56	22.189	-0.0045
	Voting_softmax	2226.9	40.588	-1.7328

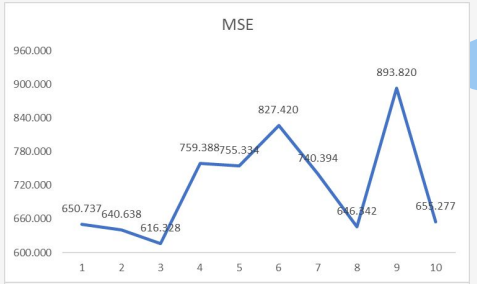
- As shown, SARIMA models captures stationarity and seasonality but fails to capture the spikes
- Machine Learning models (such as Random Forest and KNN) capture those volatile periods with high dengue occurrence, but it may exaggerate, similar to a false positive in classification
- Merging the two approaches would capture both the low periods and volatile periods, though not accurate but enough to alert a spike in Dengue cases
- Mean imputation gives the best result (ensemble model)

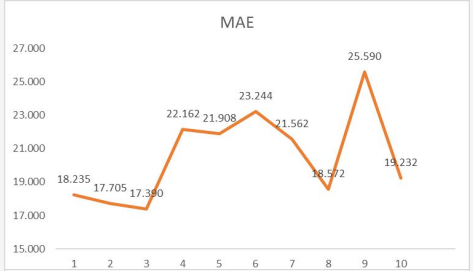


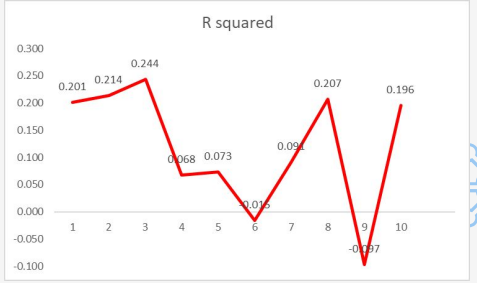
Different Time Leads: San Juan

- A lead of 3 weeks using a simple ensemble gives the best result in terms of R squared 0.244 vs 0.2406



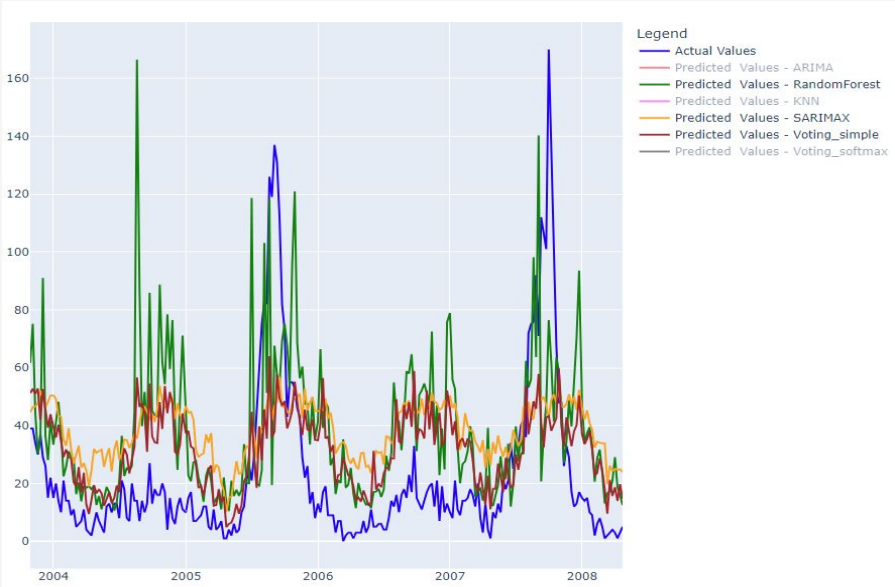






Reduced Predictor Sets: San Juan

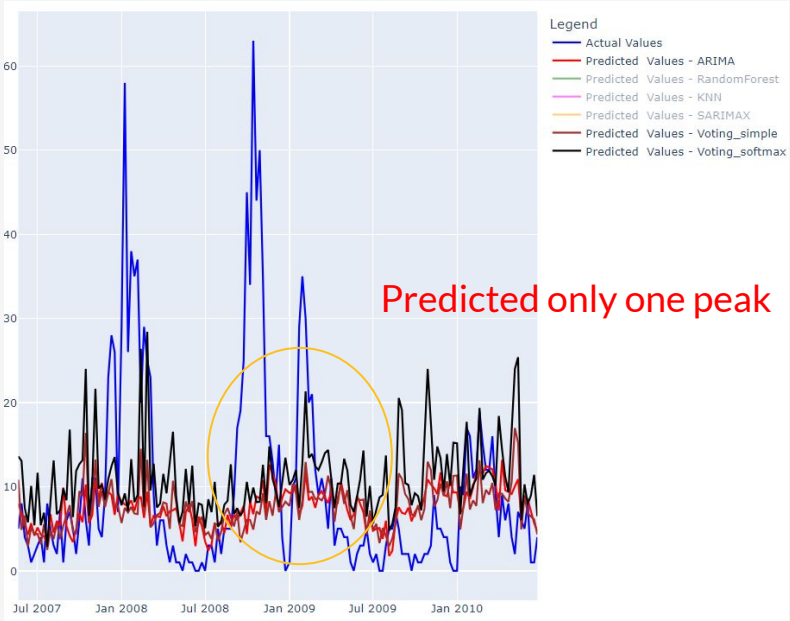
Based on the clustered correlation matrix, 'reanalysis_tdtr_k','ndvi_nw','reanalysis_precip_amt_kg_per_m2', 'precipitation_amt_mm','station_precip_mm','reanalysis_avg_temp_k', and 'station_avg_temp_c' are kept in training set, decent results under reduced training time with simple ensemble



imputation_method	model_type	mse	mae	r2
mean	ARIMA	1061.538383	28.328797	-0.302723
mean	RandomForest	1064.818154	24.240413	-0.306748
mean	KNN	1301.801709	26.589744	-0.597575
mean	SARIMAX	902.443888	26.174785	-0.107482
mean	Voting_simple	711.890347	20.657720	0.126366
mean	Voting_softmax	1866.465453	37.575005	-1.290533
mode	ARIMA	1061.538383	28.328797	-0.302723
mode	RandomForest	1093.860983	24.404024	-0.342390
mode	KNN	1301.357436	26.523932	-0.597030
mode	SARIMAX	895.055358	25.914368	-0.098415
mode	Voting_simple	841.485996	24.591313	-0.032674
mode	Voting_softmax	1912.324581	38.281104	-1.346811

Model Testing: Iquitos

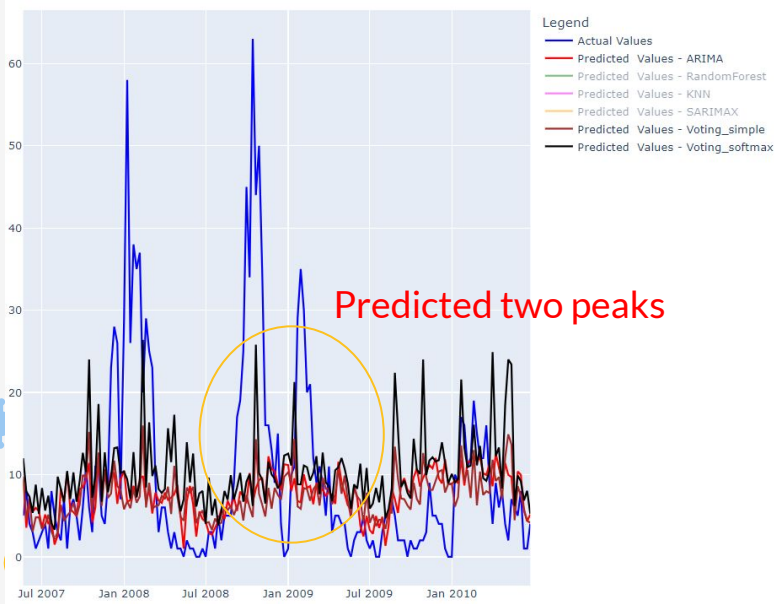
- For Iquitos, ARIMA with exogenous variables outperforms SARIMA with exogenous variables. SARIMA was then trained without exogenous predictors to obtain the stationary level
- Softmax ensemble captures some peak periods but fit badly



imputation	model	mse	mae	r2
mean	ARIMA	135.1519	7.084665	0.013423
	RandomForest	145.9922	7.692222	-0.06571
	KNN	170.3417	8.582716	-0.24345
	SARIMAX	147.9283	7.561395	-0.07984
	Voting_simple	141.913	7.430963	-0.03593
	Voting_softmax	154.254	8.759855	-0.12602
mode	ARIMA	143.7759	7.285124	-0.04953
	RandomForest	152.7751	7.684691	-0.11522
	KNN	170.5363	8.493827	-0.24488
	SARIMAX	147.9283	7.561395	-0.07984
	Voting_simple	146.071	7.523068	-0.06628
	Voting_softmax	156.8829	8.722983	-0.14521
ffill	ARIMA	135.6609	7.097148	0.009708
	RandomForest	151.5253	7.841173	-0.1061
	KNN	172.004	8.590123	-0.25559
	SARIMAX	147.9283	7.561395	-0.07984
	Voting_simple	144.4735	7.525608	-0.05462
	Voting_softmax	152.1673	8.751141	-0.11079
bfill	ARIMA	136.1437	7.038495	0.006183
	RandomForest	144.2463	7.670396	-0.05296
	KNN	172.0373	8.581481	-0.25583
	SARIMAX	147.9283	7.561395	-0.07984
	Voting_simple	143.0733	7.455147	-0.0444
	Voting_softmax	150.5349	8.409605	-0.09887
interpolation	ARIMA	136.4226	7.061819	0.004147
	RandomForest	144.5395	7.656481	-0.0551
	KNN	172.118	8.614815	-0.25642
	SARIMAX	147.9283	7.561395	-0.07984
	Voting_simple	142.9167	7.467759	-0.04326
	Voting_softmax	151.3914	8.565704	-0.10512

Reduced Predictor Set improved softmax

- Based on the clusters in the quasi-diagonalized matrix, the following predictors were retained: 'ndvi_se','ndvi_ne','reanalysis_avg_temp_k','reanalysis_tdtr_k','station_precip_mm','reanalysis_relative_humidity_percent','reanalysis_dew_point_temp_k','reanalysis_precip_amt_kg_per_m2','precipitation_amt_mm','reanalysis_min_air_temp_k'



ffill	ARIMA	135.660902	7.097148	0.009708
ffill	RandomForest	151.525293	7.841173	-0.106099
ffill	KNN	172.003951	8.590123	-0.255588
ffill	SARIMAX	147.928254	7.561395	-0.079841
ffill	Voting_simple	144.473548	7.525608	-0.054623
ffill	Voting_softmax	152.167317	8.751141	-0.110785

Entire Set

ffill	ARIMA	136.320808	7.169463	0.004890
ffill	RandomForest	153.524017	7.819979	-0.120689
ffill	KNN	164.786173	8.256790	-0.202900
ffill	SARIMAX	147.928254	7.561395	-0.079841
ffill	Voting_simple	143.172663	7.433999	-0.045127
ffill	Voting_softmax	146.591929	8.465177	-0.070086

New set

06

Results

Results & Interpretations

Interpretation of Results

What do they show?

- By collecting data on temperature, precipitation, humidity and vegetation, a prediction was made by a model trained on historical data on the trend of Dengue cases with a 3-week lead in San Juan & Iquitos
- By collecting data on the reduced predictor set for Iquitos, a signal can be given as to whether it would be a high or low period for dengue cases in Iquitos
- The results from our Simple Ensemble for San Juan show that the average absolute error between our prediction and the test set (MAE) in the model is around 16 cases
- The results from our Softmax Ensemble for Iquitos show that the average absolute error between our prediction and the test set (MAE) is around 8.4 cases
- The predictor that had the most impact on the spikes in cases in both the models for San Juan and Iquitos was precipitation, which makes sense since any form of water is a breeding ground for mosquitoes

Interpretation of Results

How is this useful?

- Since our results don't differ too much from reality, we can safely use them to predict when spikes of dengue fever cases are going to happen.
- By being able to know when spikes of dengue fever cases should occur, the necessary resources to mitigate these cases can be allocated within these cities.
- While it is difficult to perfectly predict disease cases perfectly, the model created can still allow stakeholders to make informed decisions.
- They can base themselves off our model to flag early warning signs, and this can help reduce the likelihood of death and severe symptoms.

07

Implications

Implications & Suggestions for Stakeholders

Implications for Stakeholders

Public Health Authorities:

- Early Warning and Resource Allocation: The predictive models can help public health authorities set up early warning systems, allowing them to respond to future epidemics in a timely manner. This permits the proactive allocation of resources such as medical supplies, manpower, and preventive measures.
- Improved Planning: Having access to precise projections can assist public health organisations in planning and implementing targeted interventions in high-risk areas, resulting in more efficient and successful public health policies.

Healthcare Organizations:

- Resource Allocation: The forecasts can be used by healthcare organisations to optimise resource allocation within their facilities. For example, if a specific area is predicted to see an increase in dengue fever cases, hospitals and clinics can plan for greater patient loads and provide resources accordingly.
- Access to early warnings assists healthcare professionals to improve their preparedness by ensuring they have the required medical staff, therapies, and facilities in place to deal with any epidemics.

Implications for Stakeholders

Local Governments:

- Policy Decision Support: Accurate forecasting and early warning systems give essential information for local governments to make informed public health policy decisions. Including but not limited to: implementing preventative steps, assigning funding, and collaborating with relevant institutions to handle expected health concerns.

Community:

- Increased Awareness: Early warnings enable people to take preventive measures, such as eliminating mosquito breeding areas or seeking medical assistance at the first signs of sickness.

Researchers and Scientists:

- Model Improvement: Comparing different time series forecasting models advances epidemiology and predictive modelling. The findings of this study can assist enhance the accuracy and reliability of models used to predict vector-borne diseases such as dengue fever.

Implications for Stakeholders

International Health Organizations:

- Global Health Surveillance: The project's findings may aid in global health surveillance efforts, particularly if the methods and models established can be implemented in other locations confronting similar health concerns. Collaboration with international health organisations could result in the dissemination of best practises and data-driven strategies.

Policy Makers:

- Data-Driven Decision-Making: In public health policy, the emphasis on data-driven decision-making can inspire policymakers to prioritise evidence-based policies. The project's findings may advocate for the incorporation of predictive modelling into policy creation to solve public health issues.

08

Appendix

References, Glossary, Extra Charts, etc.

Abbreviations & Definitions

- GHCN: Global Historical Climatology Network
- NOAA: National Oceanic and Atmospheric Administration
- NCEP: National Centers for Environmental Prediction
- PERSIANN: Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks
- Vegetation Indices Ex. ndvi_se: indicates the amount of vegetation in a certain area of the city (SE: south-east, NW: north-west, etc.)
- Diurnal Temperature Range: the difference between daily maximum and minimum temperatures
- Kurtosis: a measure of the “tailedness” of a predictor’s distribution
- Tailedness: having tails in a certain form ex. thin versus wide
    - Leptokurtic: high kurtosis shown by a sharp peak and values more concentrated around the mean
    - Platykurtic: low kurtosis shown by a flatter distribution and values further from the mean

Abbreviations & Definitions

- Univariate: data with observations of only a single characteristic or attribute
- Stationarity: stochastic process with a distribution that doesn't change over time. To be put more simply it is a time series with no trends or changes in variance, therefore not capturing seasonality.
- ARIMA: AutoRegressive Integrated Moving Average
    - It is a combination of autoregressive, differencing, and moving average components. The model is denoted as ARIMA(p, d, q), where "p" is the order of the autoregressive component, "d" is the order of differencing, and "q" is the order of the moving average component.
- SARIMAX: Seasonal AutoRegressive Integrated Moving Average with eXogenous factors
    - It is an extension of the SARIMA (Seasonal ARIMA) model that incorporates exogenous, or external, variables into the time series forecasting framework. Exogenous variables are additional independent variables that are not part of the time series but may influence its behavior.

Abbreviations & Definitions

- Augmented Dickey-Fuller test: testing to see if non-stationarity is present in a time series model (this is the null hypothesis). The alternative hypothesis is that the model shows stationarity or trend-stationarity.
- Lag: features included in a model that come from the target value of a previous time period. For example, in this case you can use the number of dengue fever cases from the week prior, and you would be using a lag = 1 (weeks). If you wanted to use the number of cases from the last month, you would use lag = 4.
- Lead: similar to lag, except the features are calculated from the target value from a future time period.
- Mean: the central tendency of the data
- Standard deviation: the spread of the data
- Min and Max: the range of the data
- 25th, 50th, and 75th percentile: give insights into the data distribution at each percentile

References

Brown, J. J., Pascual, M., Wimberly, M. C., Johnson, L. R., & Murdock, C. C. (2023). *Humidity - The overlooked variable in the thermal biology of mosquito-borne disease*. Ecology letters, *26*(7), 1029–1049. https://doi.org/10.1111/ele.14228

Center for Disease Control. (n.d.) *Dengue Around the World*. https://www.cdc.gov/dengue/areaswithrisk/around-the-world.html

Ebi, K. L., & Nealon, J. (2016). *Dengue in a changing climate*. Environmental Research, *151*, 115-123. https://doi.org/10.1016/j.envres.2016.07.026

Prabhakaran, S. (n.d.). *ARIMA Model – Complete Guide to Time Series Forecasting in Python*. Machine Learning Plus. https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/#:~:text=ARIMA%2C%20short%20for%20'Auto%20Regressive,used%20to%20forecast%20future%20values

Sauer, F. G., Grave, J., Lühken, R., Kiel, E. (2021). *Habitat and microclimate affect the resting site selection of mosquitoes*. Med Vet Entomol, *35*, 379-388. https://doi.org/10.1111/mve.12506

References

Tesla, B., Powers, J. S., Barnes, Y., Lakhani, S., Acciani, M. D., & Brindley, M. A. (2022). Temperate Conditions Limit Zika Virus Genome Replication. *Journal of virology*, *96*(10), e0016-522. https://doi.org/10.1128/jvi.00165-22

Truscott, P. (2017, May 31). *How Weather Affects Mosquito Activity*. Mosquito Buzz. https://blog.mosquito.buzz/how-weather-affects-mosquito-activity