

Disease Spread

Predicting the Spread of Dengue using Machine Learning

Capstone II Project: Milestone Report 1

Springboard Data Science Career Track

Mark Rojas

March 2019

1. Problem Statement

Dengue fever is a disease caused by any one of four closely related dengue viruses (**DENV 1, DENV 2, DENV 3, or DENV 4**). Dengue cannot be spread directly from person to person. The viruses are transmitted to humans by the bite of an infected *Aedes aegypti* or *Aedes albopictus* mosquito.

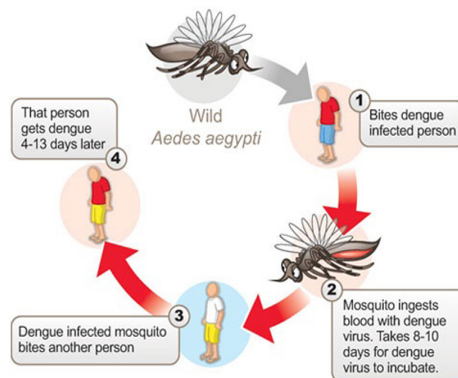


Figure 1. Human Transmission Cycle of Dengue Virus

In mild cases, symptoms are similar to the flu, however, in severe cases, dengue fever can cause severe bleeding, low blood pressure, and even death. As many as **400 million people are infected yearly with an estimated 500,000 people hospitalized with severe dengue**, also known as dengue hemorrhagic fever (DHF) (<https://www.draper.com/news-releases/predicting-dengue-fever-outbreaks-machine-learning>).

Outbreaks of dengue occur primarily in areas where *Ae. aegypti* or *Ae. albopictus* mosquitoes live. This includes most tropical urban areas of the world. Dengue viruses may also be introduced into areas by travelers who become infected while visiting other areas of the tropics where dengue commonly exists.

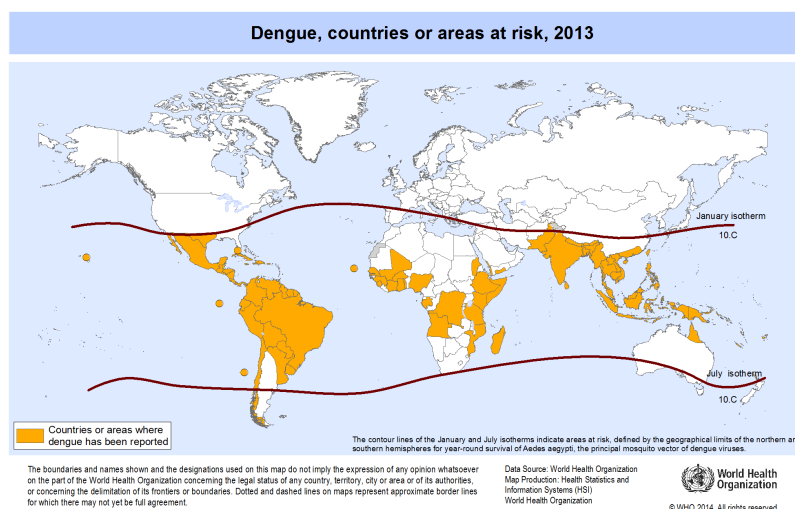


Figure 2. Dengue Fever Risk Map 2013.

(Source: WHO -http://gamapserver.who.int/mapLibrary/Files/Maps/Global_DengueTransmission_ITHRiskMap.png)

Since dengue is carried by mosquitoes, the transmission dynamics of dengue are related to climate variables such as temperature and precipitation. Although the relationship to climate is complex, a growing number of scientists argue that climate change is likely to produce distributional shifts that will have significant public health implications worldwide.

More recently, dengue fever has been spreading. 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.** And the problem is only growing worse. **Over the past 50 years, dengue cases have increased thirty-fold, and currently half of the world's population, some 3.9 billion people, are at risk.**

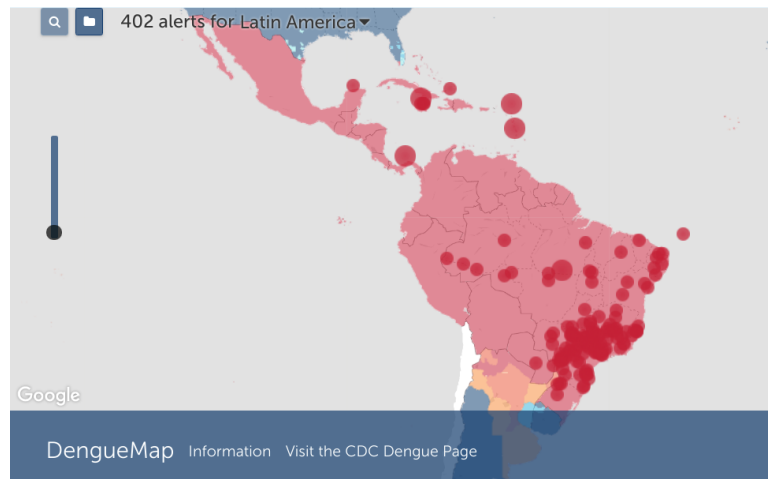


Figure 3: Dengue Notifications/Alerts for Latin America
402 alerts for the month of January 2019
(Source - <https://www.healthmap.org/outbreaksnearme>)

1.1. Goal

Using environmental data collected by various US Federal Government agencies, **the objective is to predict the number of dengue fever cases reported each week in San Juan, Puerto Rico and Iquitos, Peru based on environmental variables describing changes in temperature, precipitation, humidity, vegetation, and more.**

1.2. Client

Accurate dengue predictions would help several departments in the **US Federal Government (Department of Health and Human Services, Department of Defense, and Department of Commerce)**, **public health workers and people around the world** where mosquitoes are present.

2. Data

2.1. Collection

The data for this competition comes from multiple sources aimed at supporting the **Predict the Next Pandemic Initiative**.

Dengue Surveillance Data from:

- Centers for Disease Control and Prevention
- Department of Defense's Naval Medical Research Unit 6
- Armed Forces Health Surveillance Center
- Peruvian government and US universities

Environmental and Climate Data from:

- National Oceanic and Atmospheric Administration in the US Department of Commerce

File	Description	Format
Training Data Features	The features for the training data set	CSV
Training Data Labels	The number of dengue cases for each row in the training data set	CSV
Test Data Features	The features for the testing data set	CSV

Table 1: Data available from DrivenData.org

(Source - <https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/data>)

A complete list of features and their descriptions can be found here:

https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/page/82/#features_list

2.2. Cleaning and Transforming

The data consists of **1,456 observations** and **24 features**.

- **936 observations** are from **San Juan**
- **520 observations** are from **Iquitos**

Missing and Null Values

- For **San Juan**, 20 of the 24 features contained between **6 – 191** null values.
- For **Iquitos**, 20 of the 24 features contained between **3 – 37** null values.

San Juan, Puerto Rico		Iquitos, Peru	
Features	Null Values	Features	Null Values
14	6	4	3
2	9	11	4
2	19	1	8
1	49	1	14
1	191	1	16
		2	37

Table 2: Number of Features with Null Values for San Juan and Iquitos

Because our goal is to assess the impact of environmental variables on reported cases of Dengue infection over a given time scale, removing observations with null values may hinder the predictive capabilities. To avoid this, null values were imputed with median values for each 'month' and 'year'. Median values were used rather than means because of outliers (explained further in the EDA section).

Unit Conversion

For both San Juan and Iquitos, NOAA's GHCN temperatures are in Celsius (degree) while NOAA's NCEP temperatures are in Kelvin (non-degree) measurements. To avoid potential issues with scaling, NOAA's NCEP temperature Kelvin units were converted to Celsius using the formula: $0K - 273.15 = -273.1^{\circ}C$. Columns were renamed accordingly to reflect unit change.

3. Exploratory Data Analysis (EDA)

3.1. Training Labels – Total Cases

First thing we want to look at are the training labels (total_cases) we intend to predict. We want to ensure there are no missing or null values.

Null value counts for Training Labels

San Juan: 0

Iquitos: 0

Because we did not remove rows (observations) from the feature set due to missing or null values, we do not need to modify the training labels.

Descriptive Statistics

In Table 3, on average, we tend to see **22% more** cases of Dengue virus reported in San Juan than in Iquitos.

San Juan	count	mean	std	min	25%	50%	75%	max
year	936	1998.83	5.21	1990	1994	1999	2003	2008
weekofyear	936	26.50	15.02	1.0	13.75	26.5	39.25	53.0
total_cases	936	34.18	51.38	0.0	9.00	19.0	37.00	461.0
Iquitos	count	mean	std	min	25%	50%	75%	max
year	520	2005.0	2.92	2000	2002.8	2005	2007.3	2010
weekofyear	520	26.50	15.03	1.0	13.75	26.5	39.25	53.0
total_cases	520	7.57	10.77	0.0	1.00	5.0	9.00	116.0

Table 3: Descriptive Statistics for Training Labels

Distributions – Total Cases

For both San Juan and Iquitos, we see **Positively Skewed** (right skewed) distributions for the number of Total Cases (**Figure 4**). This indicates that the selected predictive model will need to be optimized for outliers or the labeled data will need to be normalized. One approach to normalize the labeled data will be to apply **Logarithmic Transformation**.

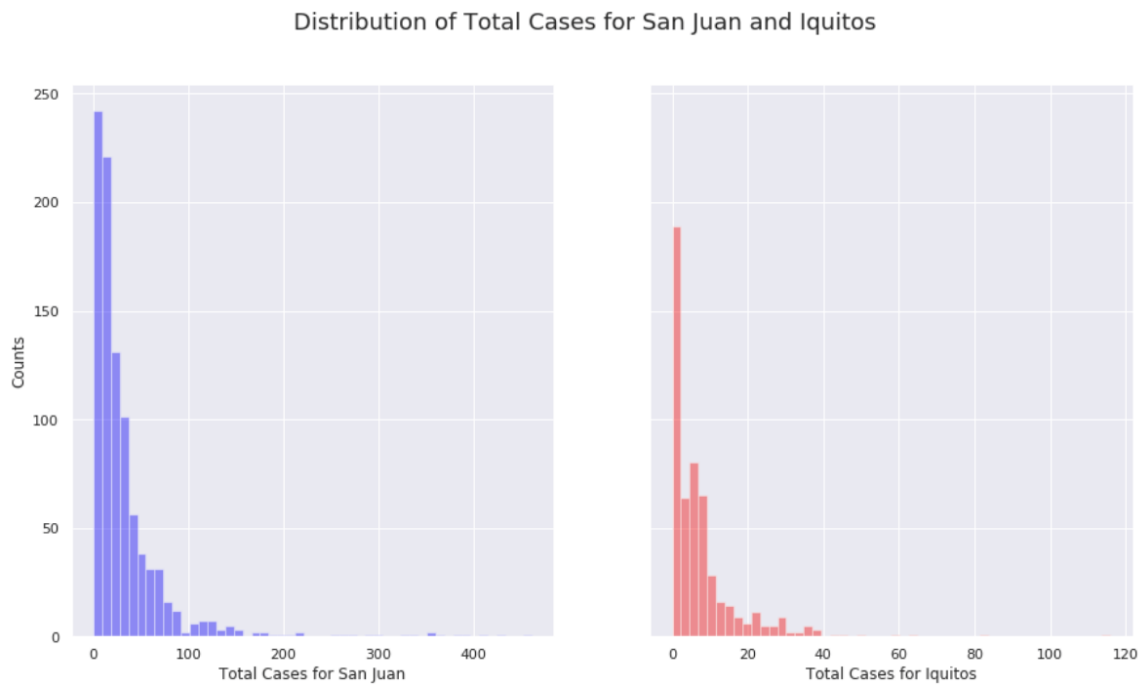


Figure 4: Distribution of Total Cases Reported for San Juan and Iquitos

Distributions – Total Cases per Week of Year (53 weeks)

Are Total Cases of Dengue reported happening during the same week of the year for San Juan and Iquitos? To check this, we grouped 'total_cases' by 'weekofyear', shown in Figure 5.

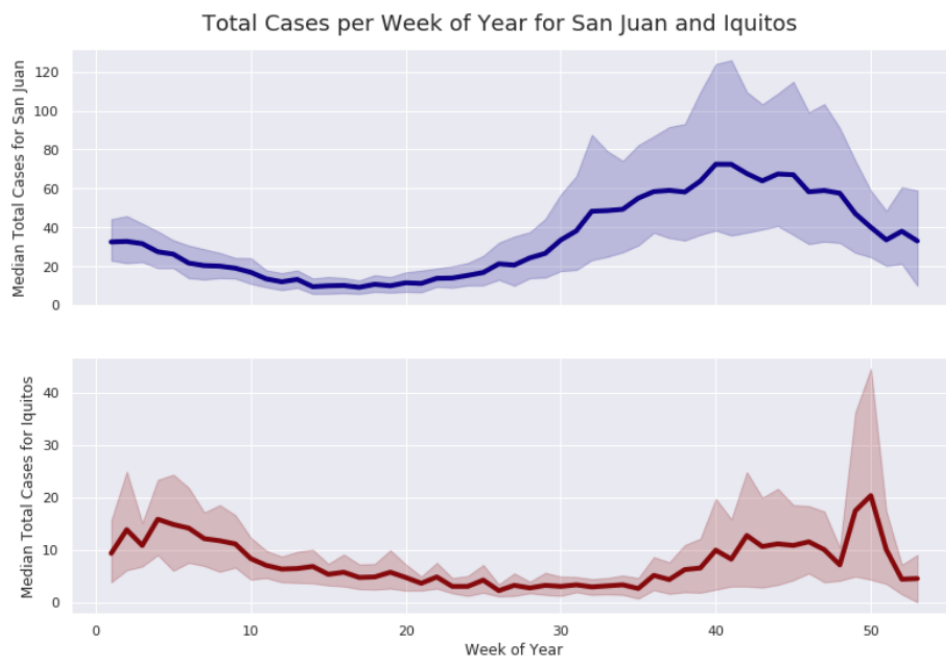


Figure 5: Distribution of Total Cases Reported over 53-week Period

We can see similar trends for number of cases reported in San Juan as for Iquitos. Most of the cases are reported at the end of the year, however, there are quite a few at the beginning of the year as well. To further illustrate a possible trend and/or seasonality, we can apply Decomposition Time Series to Total Cases for San Juan and Iquitos (Figure 6a, 6b).

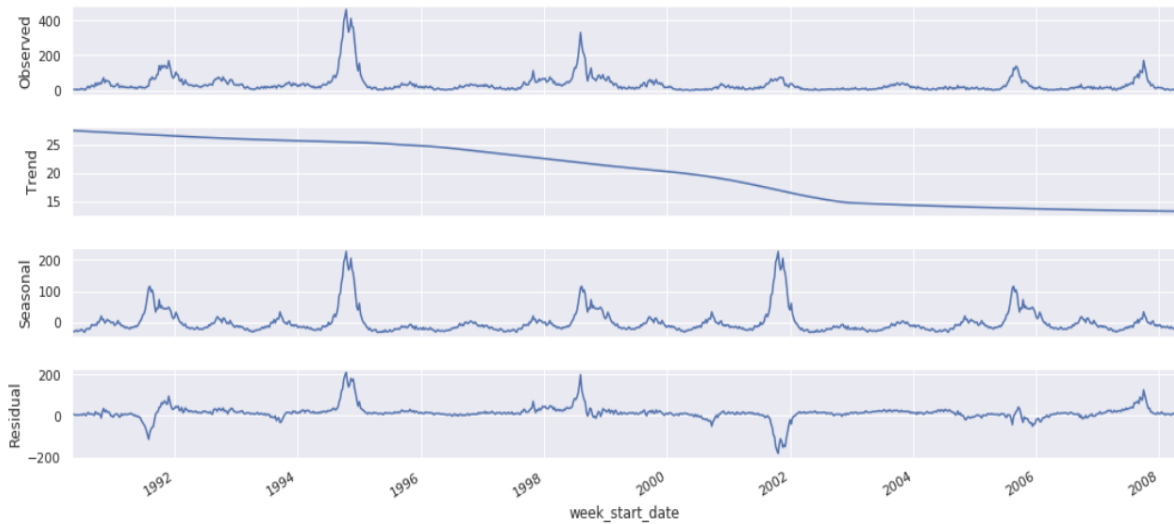


Figure 6a: Time Series Data for **San Juan** Decomposed into Trend and Seasonality

For San Juan (Figure 6a), we see what appears to be strong seasonality, supported by the years 1992-1995 and 1999-2002. We also see a downward trend that bottoms out around 2003-2004. There is also a cyclic behavior that corresponds to seasonality. Finally, the residual graph appears to show seasonality with cyclic behavior.

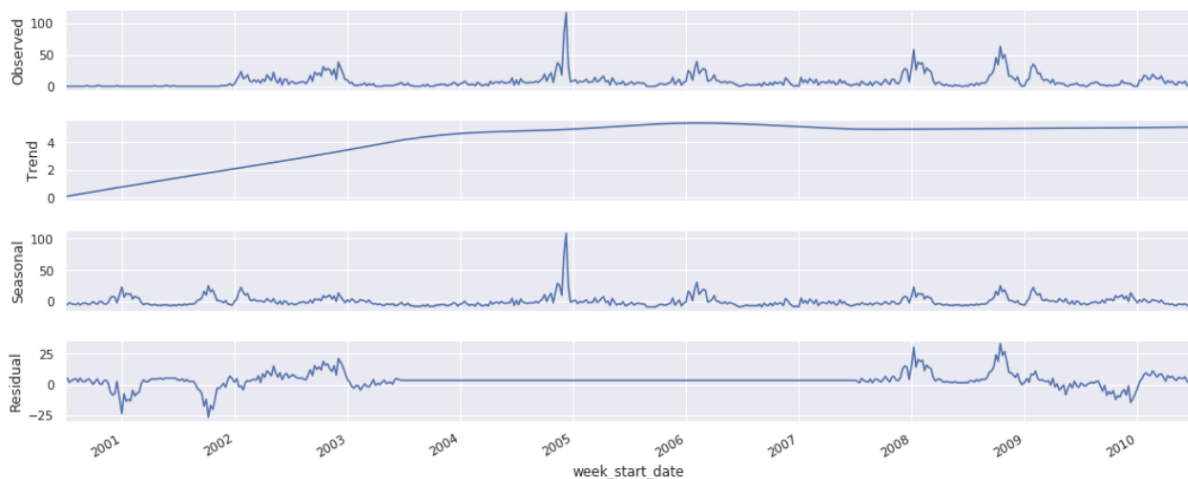


Figure 6b: Time Series Data for **Iquitos** Decomposed into Trend and Seasonality

For Iquitos (Figure 6b), we also see what appears to be strong seasonality, supported by the years 2001-2003 and 2008-2010. **Unlike San Juan, we see an upward trend that max's out around 2003-2004. This is interesting since it is during the downtrend in same years as San Juan.** The cyclic behavior tends to correspond to seasonality. The residual graph appears to show seasonality with cyclic behavior, however, from the middle of 2003-2008, we don't see any variation in the series.

Outliers

Using z-scores with a threshold > 3 to detect outliers, we were able to reveal that there are **20** outliers present in the **San Juan** label data and **7** outliers for **Iquitos** (Tables 4a, 4b).

index	week_start_date	year	weekofyear	total_cases
227	1994-09-10	1994	36	202
228	1994-09-17	1994	37	272
229	1994-09-24	1994	38	302
230	1994-10-01	1994	39	395
231	1994-10-08	1994	40	426
232	1994-10-15	1994	41	461
233	1994-10-22	1994	42	381
234	1994-10-29	1994	43	333
235	1994-11-05	1994	44	353
236	1994-11-12	1994	45	410
237	1994-11-19	1994	46	364
238	1994-11-26	1994	47	359
239	1994-12-03	1994	48	288
240	1994-12-10	1994	49	221
428	1998-07-23	1998	30	191
429	1998-07-30	1998	31	256
430	1998-08-06	1998	32	329
431	1998-08-13	1998	33	263
432	1998-08-20	1998	34	220
433	1998-08-27	1998	35	204

Table 4a: List of Outliers Detected for San Juan Training Labels

For San Juan (**Table 4a**), the 'outliers' appear to be influential as they are also spread across the x-axis (not just a single outlier for the year or month). Because the values of concern appear to be consistent for same months and year, we may need to investigate the presence of outliers by another method as confirmation.

index	week_start_date	year	weekofyear	total_cases
230	2004-12-02	2004	49	83
231	2004-12-09	2004	50	116
391	2008-01-08	2008	2	58
429	2008-09-30	2008	40	45
431	2008-10-14	2008	42	63
432	2008-10-21	2008	43	44
433	2008-10-28	2008	44	50

Table 4b: List of Outliers Detected for Iquitos Training Labels

For Iquitos (**Table 4b**), the 'outliers' appear to be somewhat influential as they are also spread across the x-axis (not just a single outlier for the year or month) with the exception of week of January 8, 2008.

If 'outliers' are influential observations, they will have a large impact on the estimated coefficients of a regression model.

3.2. Training Features

Next, we want to look at are the training features and determine if there is any correlation with Total Cases and/or other Features that may impact our model.

Features vs. Total Cases

First, we merged Total Cases from labeled data with feature data so we could visualize correlations using a correlation matrix (Figure 7a, 7b).

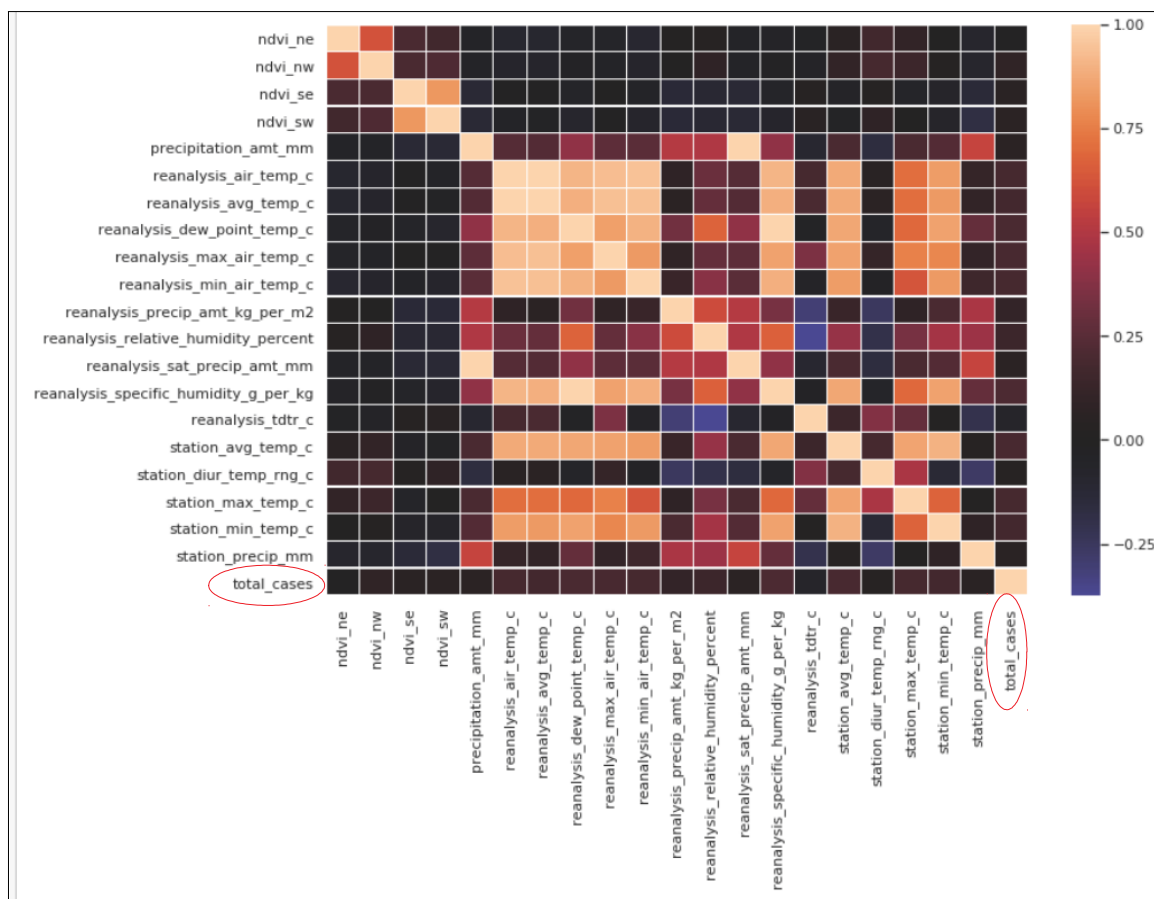


Figure 7a: Correlation Matrix for **San Juan** Training Features and Total Cases

For San Juan (**Figure 7a**), none of the features appear to be correlated with the labeled data (total_cases).

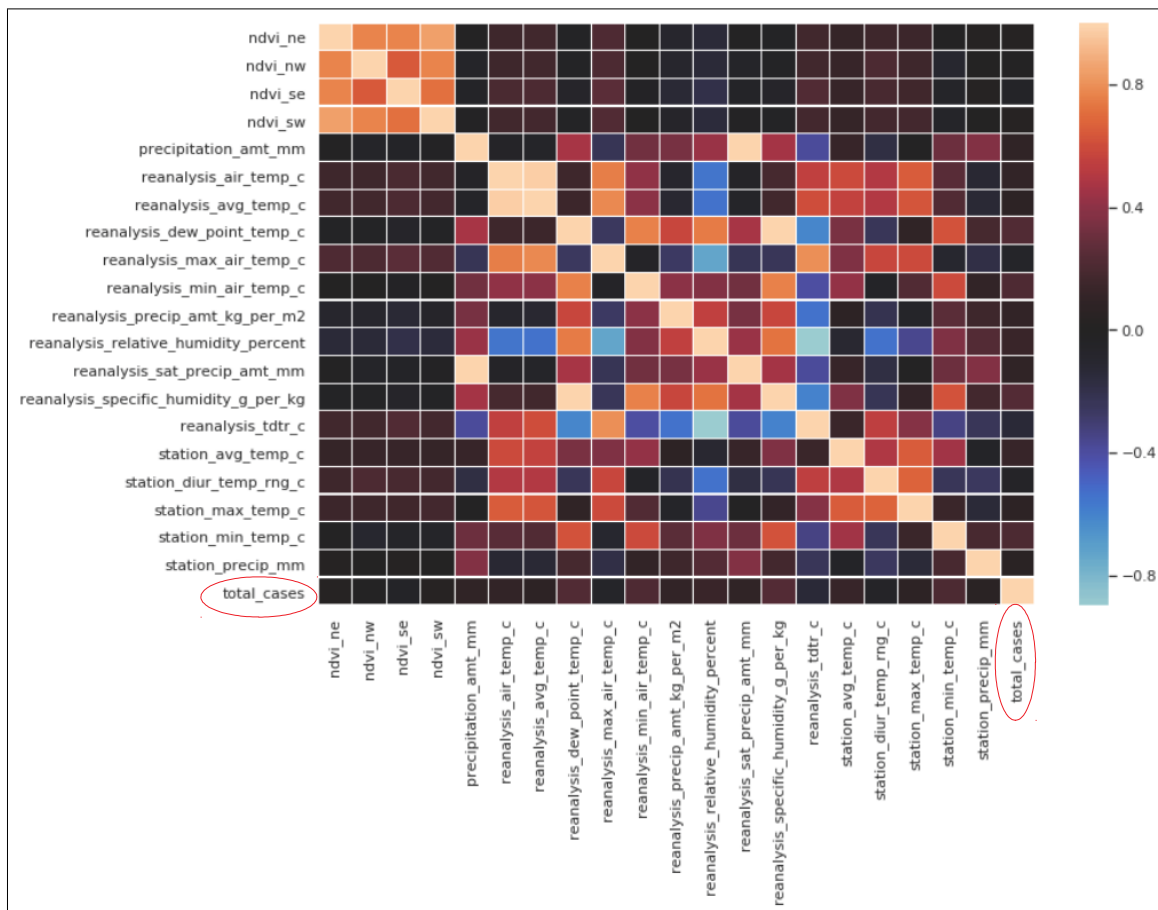


Figure 7b: Correlation Matrix for Iquitos Training Features and Total Cases

For Iquitos (**Figure 7b**), none of the features appear to be correlated with the labeled data (total_cases). While we did not see any features that are ‘highly’ correlated to total cases, in Figure 8, we can still take a look at the correlation values for features to total cases.

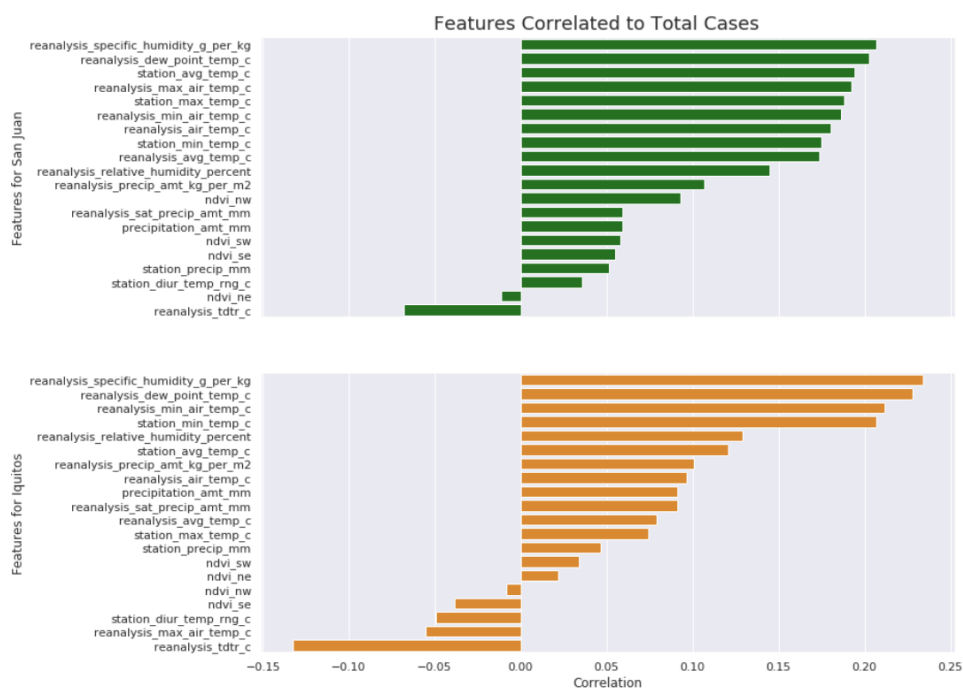


Figure 8: Features Correlation Values to Total Cases

In Figure 8, we see that for both San Juan and Iquitos, the most correlated features are **reanalysis_specific_humidity_g_per_kg** and **reanalysis_dew_point_temp_c**. This makes sense since mosquitoes lay their eggs in standing water where humidity and dew point would probably be highest.

Features vs. Features

There is some definite correlation between features. For example, for San Juan (**Figure 7a**), the vegetation index features, **ndvi_ne**, **ndvi_nw**, **ndvi_se**, and **ndvi_sw** are highly correlated. Also, we definitely see correlation between the temperature features. This is expected, since min, max, and avg temperatures should be related. There is some correlation between humidity and dew point. The diurnal temperature ranges are slightly negatively correlated with humidity and precipitation but just slightly.

For Iquitos (**Figure 7b**), once again, we see some correlation within the vegetation index features, **ndvi_ne**, **ndvi_nw**, **ndvi_se**, and **ndvi_sw**. Correlation between the temperature features are not as visually prevalent as they are for San Juan but we still see some correlation again between humidity and dew point. The diurnal temperature ranges are more negatively correlated with humidity and precipitation than for San Juan.

4. Summary

For the most part, the 24 features data and labeled data for both San Juan and Iquitos are fairly clean.

- Split feature data and labeled data into two groups (San Juan, Iquitos)
- We needed to impute missing and null values using median values
- Some values could not be imputed using values from same 'month' and 'year' since unavailable
 - This was resolved by using `fillna(...)` with forward-fill method

For San Juan, we have 936 observations data from 1990 – 2008. We have 520 observations data from 2000-2010 for Iquitos

- Approximately a 9 year overlap in observation between San Juan
- Distributions for Total Cases are positively skewed for San Juan and Iquitos
- We see Seasonality for both Time Series data
- There are outliers that need to be addressed for both cities
- Features are not 'highly' correlated to the labeled data (total_cases)
 - Highest correlated features to total_cases were:
 - reanalysis_specific_humidity_g_per_kg
 - reanalysis_dew_point_temp_c
 - **Moisture in the air!**
- We see some correlation in features vs. features that may provide some insights

5. Next Steps

Moving forward, we need to consider Demographics and Climate conditions for the two cities. This information may provide more insight into why we see opposite levels of reported cases during the year.

Excluding the time it takes for mosquitoes to breed (lay eggs and the eggs hatch), there is an incubation period, after ingestion, before for the dengue virus in mosquitoes can infect other humans. This is usually between 8-10 days after ingestion. Once a human is infected, it take another 4-13 days to show symptoms.

This will require a lag between features and labels to be inserted. A 2 and 3 week lag should be considered and compared when training the predictive model.

Performance measure to be used: **Mean Absolute Error (MAE)**

$$mae = \frac{\sum_{i=1}^n abs(y_i - \lambda(x_i))}{n}$$

A predictive model will need to be selected.

- Driven Data used **Negative Binomial Regression** to establish the benchmark score: **25.8173**
- Perform Feature Engineering
- Train model
- Tune parameters
- Test performance

We will beat this score!