Disease Spread

Predicting the Spread of Dengue

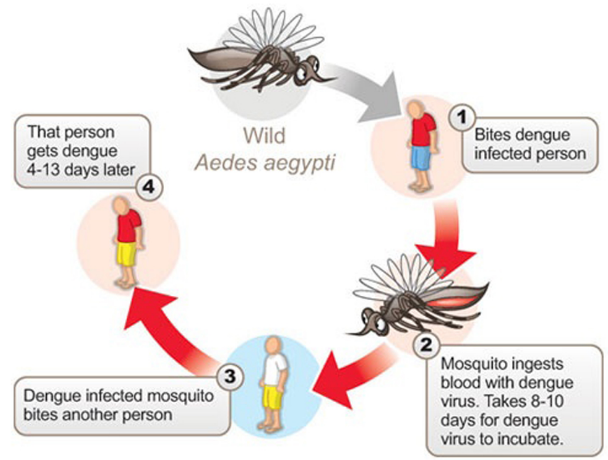
Springboard Data Science Career Track

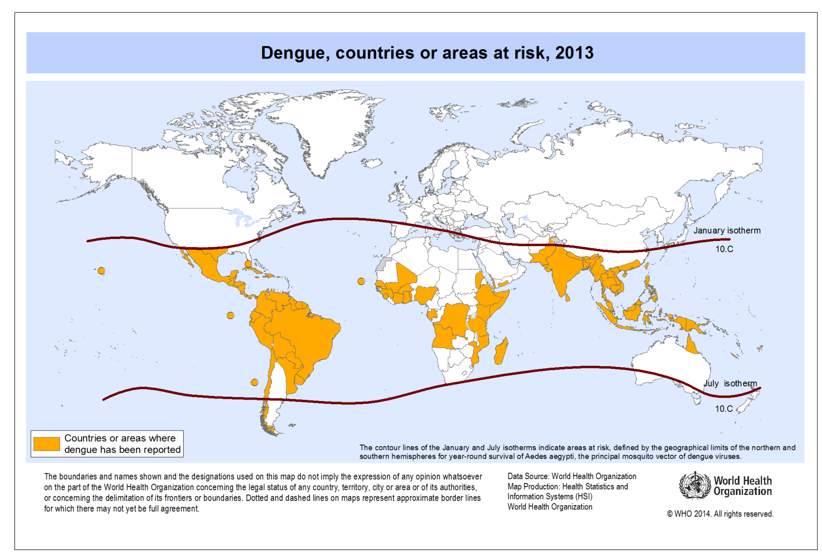
– Mark Rojas

Problem Statement

- Dengue is a mosquito-borne disease
- **400 million infected yearly**
- Symptoms similar to flu, however,
    - **severe cases, including dengue fever, can result in death!**
- **500,000 hospitalized with severe dengue**
- 3.9 Billion People at Risk!!!

- In tropical and sub-tropical parts of the world
- Believed to be related to climate variables:
    - **Temperature | Precipitation | Humidity**





Goal

Using climatological data, predict the number of dengue fever cases reported each week in:

Iquitos, Peru

San Juan, Puerto Rico

Data Collection

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

Dengue Surveillance Data:

- 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:

- 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

Data Cleaning and Wrangling

The data includes **1,456 observations** with **24 features** consisting of 2-object and 22-numerical features. **936** observations are from **San Juan** data set while the other **520** observations come from the **Iquitos** data.

Missing and Null Data:

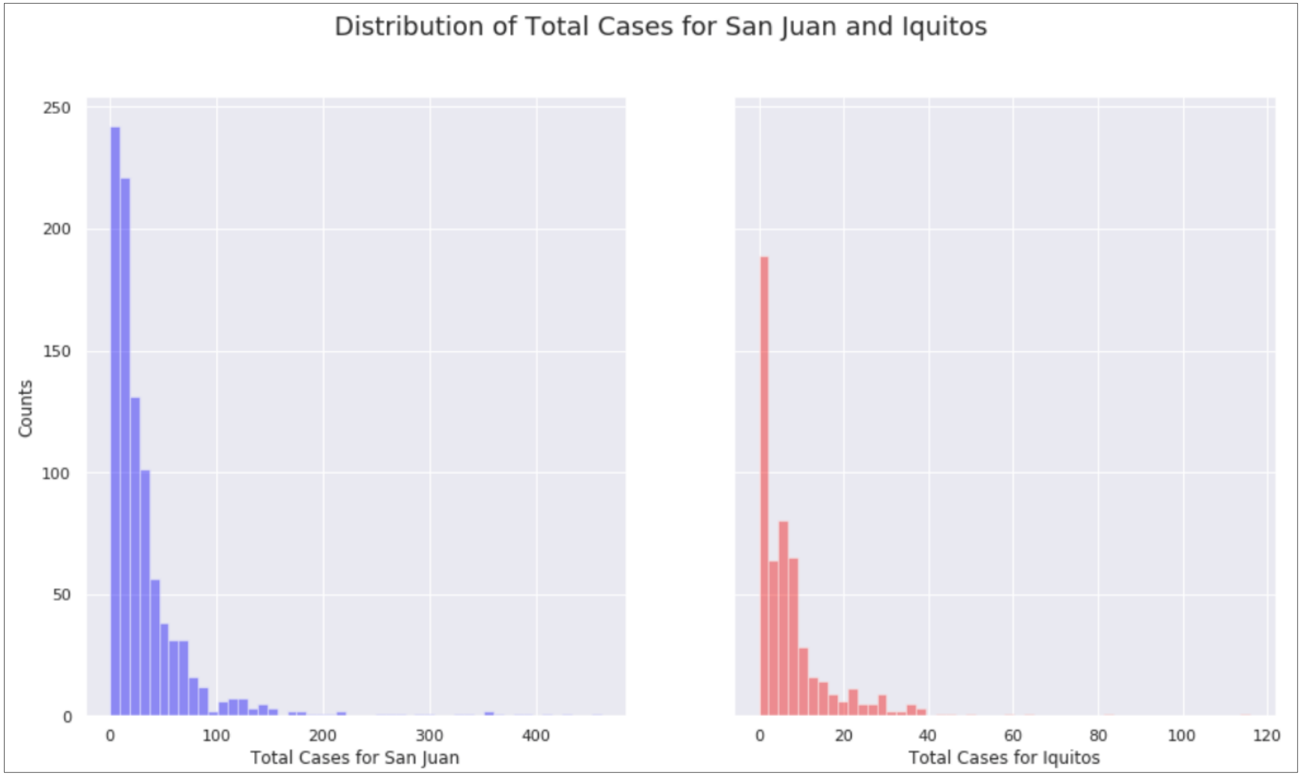
- For **San Juan**, **20** of the **24** features contained between **6 – 191 null values**
- For **Iquitos**, **20** of the **24** features contained between **3 – 37 null values**

Null values were imputed with **median** values for each ‘month’ and ‘year’. Median values were used rather than means because of outliers (described in the exploratory data analysis).

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°C`**. Columns were renamed accordingly to reflect unit change.

Exploratory Data Analysis
Training Labels (Total Cases)

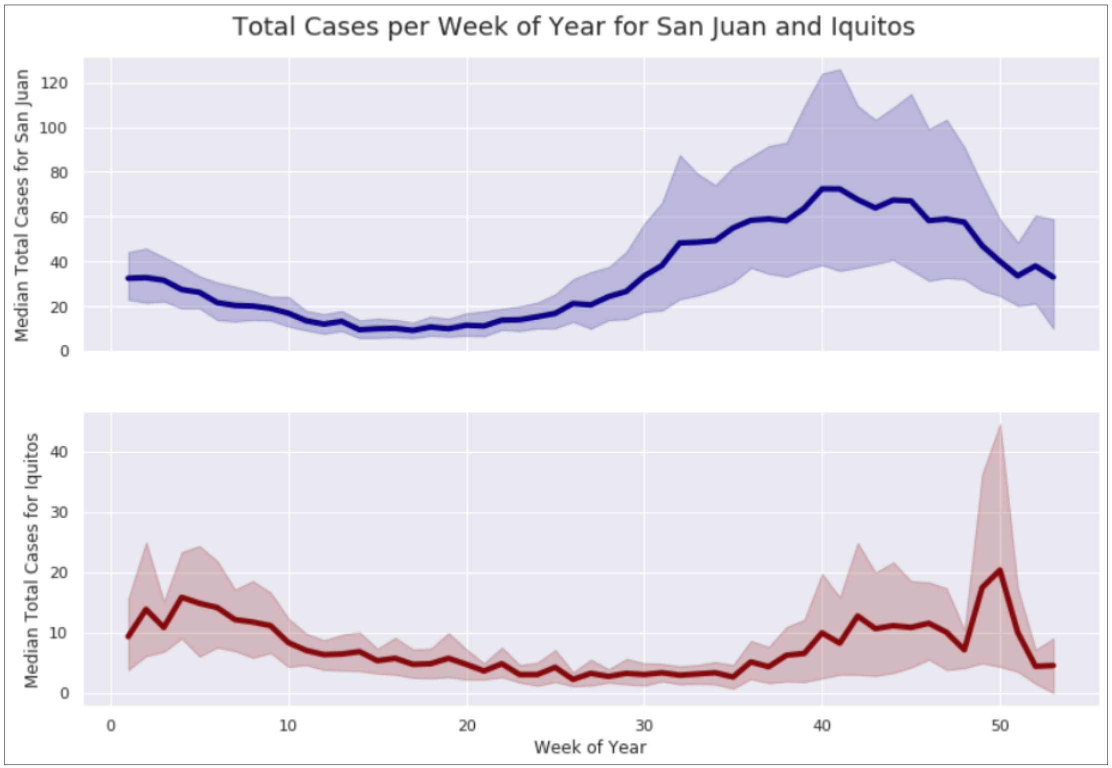


- Positively Skewed distributions
- Predictive model will need to be optimized for outliers
- Labeled data will need to be normalized

One approach to normalize the labeled data will be to apply Logarithmic Transformation.

Exploratory Data Analysis

Total Cases over 53-week Period



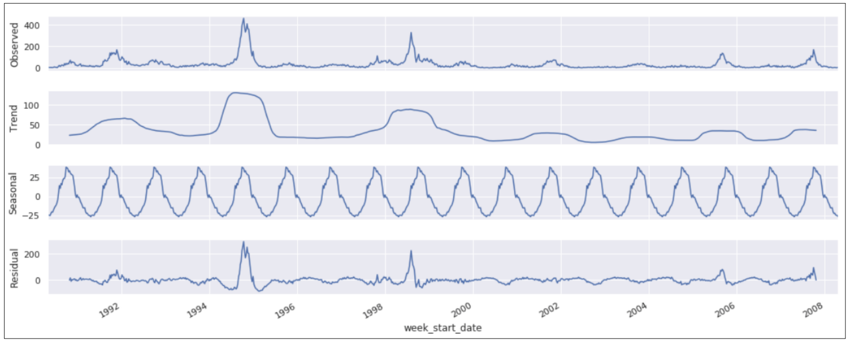
- **More cases** reported at the **end of the year** in **San Juan** than in **beginning of year**
- Similar trend in number of cases reported in **Iquitos** as for **San Juan**

To illustrate further, apply Decomposition Time Series to Total Cases for both cities.

Exploratory Data Analysis

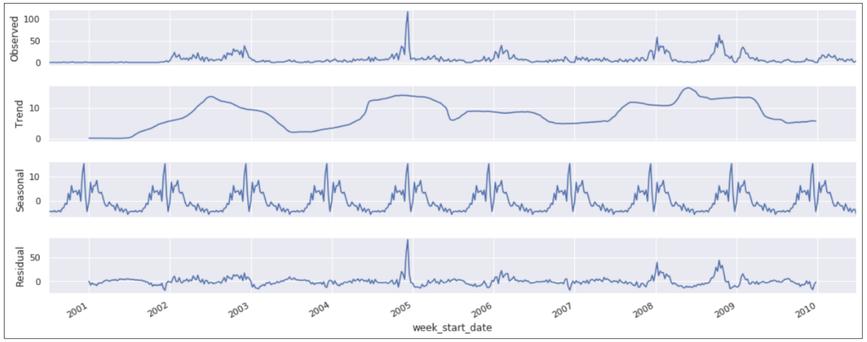
Time Series Decomposition

San Juan



- **Strong Seasonality**
- **Downward Trend**
- **Cyclic Behavior**
- **Residual shows seasonality with cyclic behavior**

Iquitos



- **Strong Seasonality**
- **Upward Trend**
- **Cyclic Behavior**
- **Residual shows seasonality with cyclic behavior**

Exploratory Data Analysis
Outliers for Training Labels

Z-scores with a threshold > 3 used to detect outliers.

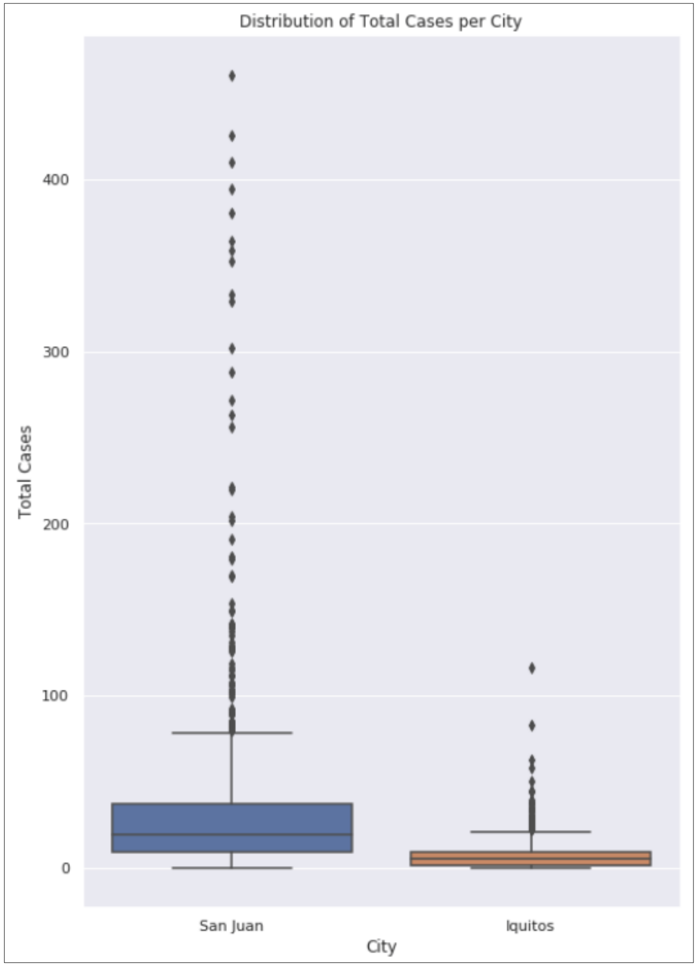
- **20 outliers** in the **San Juan** label data
- **7 outliers** in the **Iquitos** label data

San Juan

week_start_date	city	year	weekofyear	total_cases
1994-09-10	sj	1994	36	202
1994-09-17	sj	1994	37	272
1994-09-24	sj	1994	38	302
1994-10-01	sj	1994	39	395
1994-10-08	sj	1994	40	426
1994-10-15	sj	1994	41	461
1994-10-22	sj	1994	42	381
1994-10-29	sj	1994	43	333
1994-11-05	sj	1994	44	353
1994-11-12	sj	1994	45	410
1994-11-19	sj	1994	46	364
1994-11-26	sj	1994	47	359
1994-12-03	sj	1994	48	288
1994-12-10	sj	1994	49	221
1998-07-23	sj	1998	30	191
1998-07-30	sj	1998	31	256
1998-08-06	sj	1998	32	329
1998-08-13	sj	1998	33	263
1998-08-20	sj	1998	34	220
1998-08-27	sj	1998	35	204

Iquitos

week_start_date	city	year	weekofyear	total_cases
2004-12-02	iq	2004	49	83
2004-12-09	iq	2004	50	116
2008-01-08	iq	2008	2	58
2008-09-30	iq	2008	40	45
2008-10-14	iq	2008	42	63
2008-10-21	iq	2008	43	44
2008-10-28	iq	2008	44	50

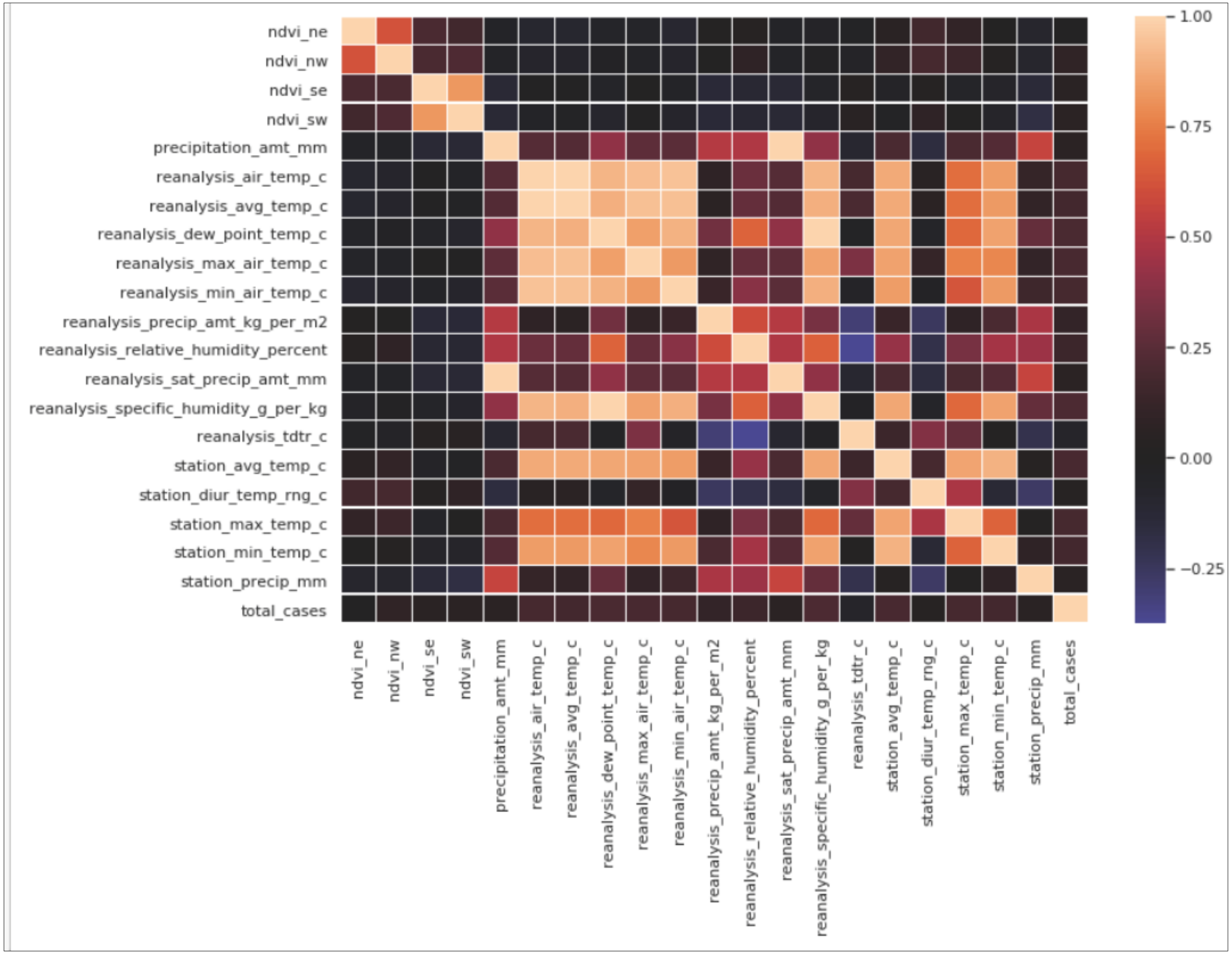


Exploratory Data Analysis

Correlation: Features vs. Total Cases

<u>San Juan</u>

Correlation Matrix

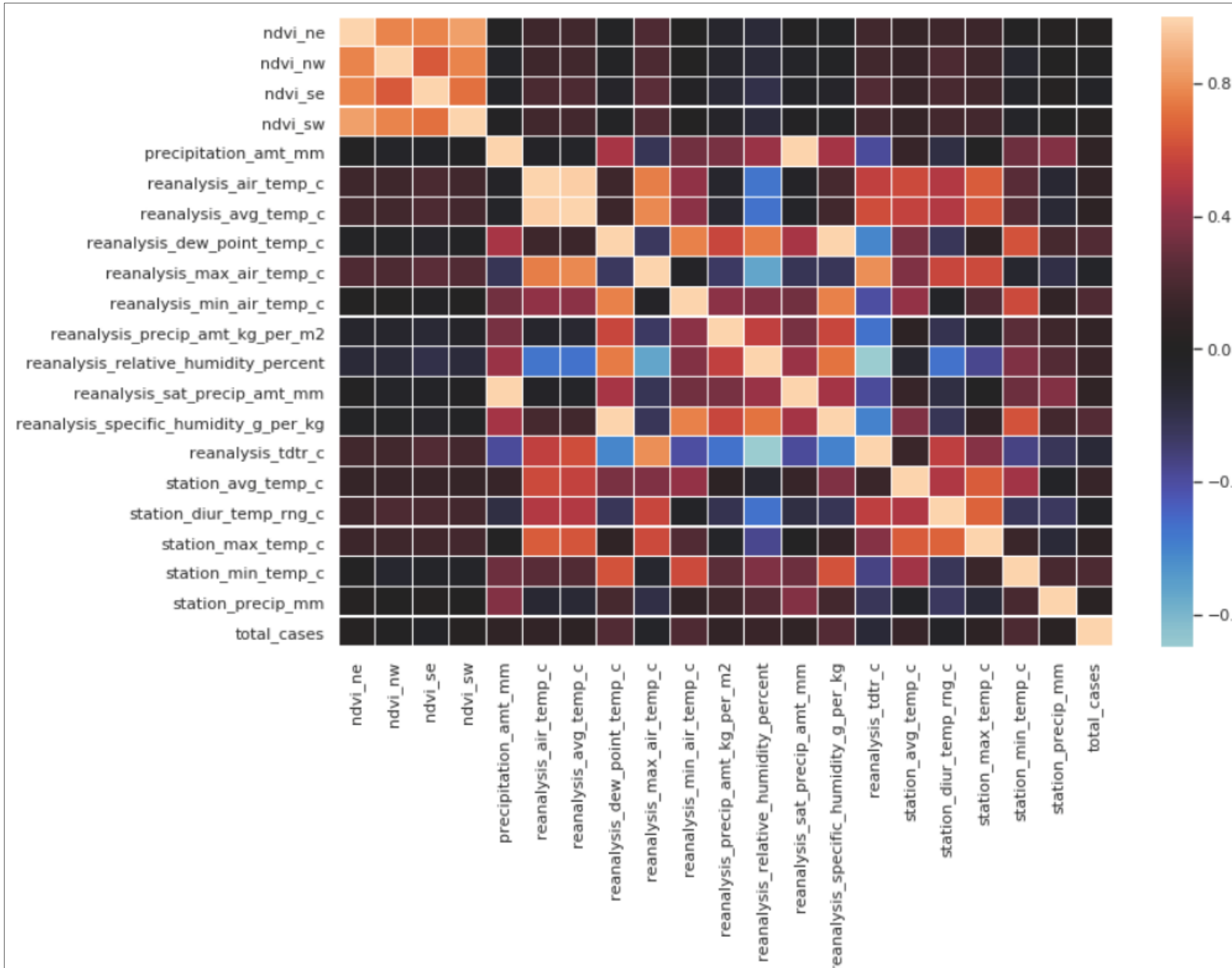


Exploratory Data Analysis

Correlation: Features vs. Total Cases

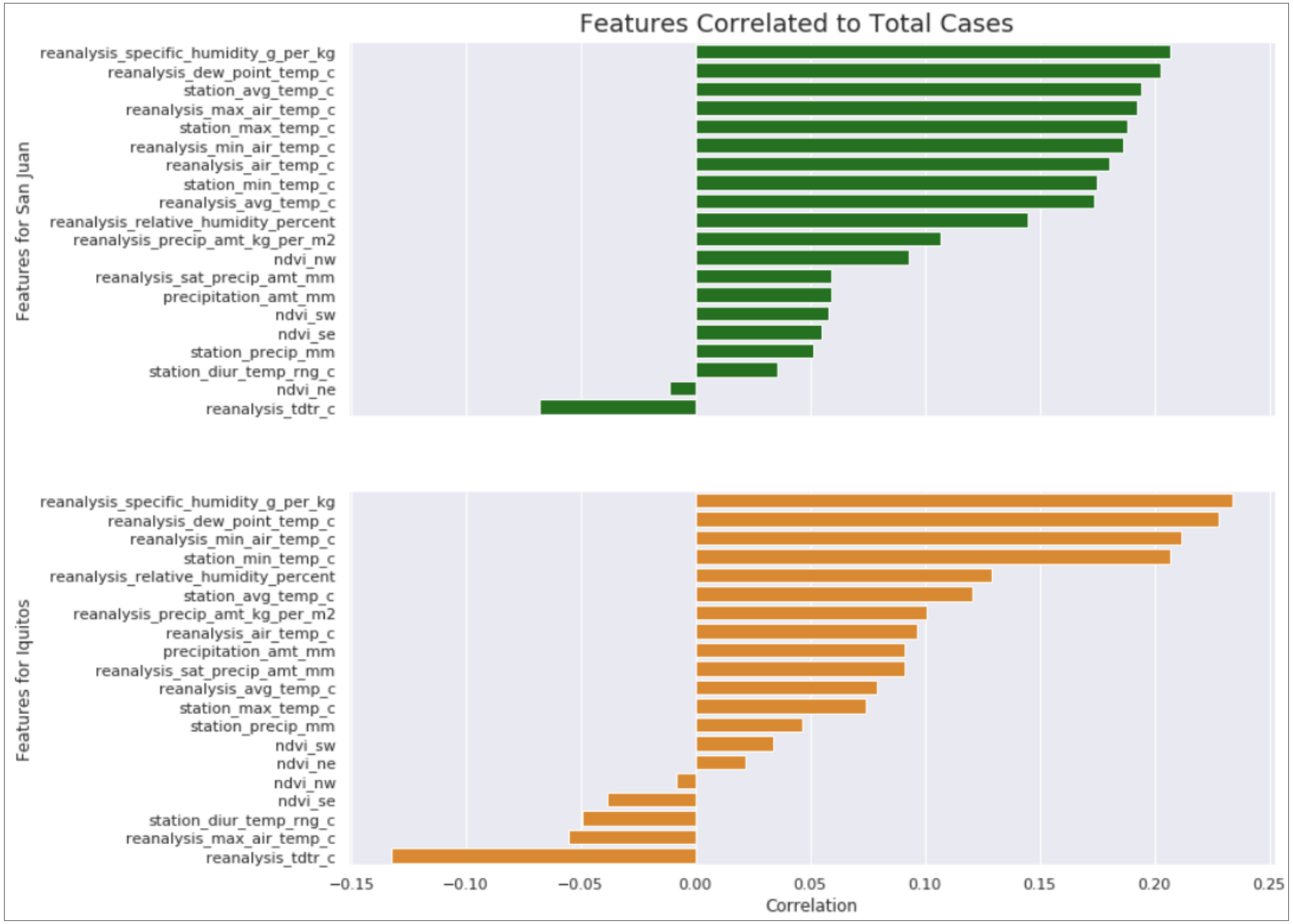
<u>Iquitos</u>

Correlation Matrix



Exploratory Data Analysis

Features Correlated to Total Cases



Moisture in the Air!!

1) **Reanalysis Specific Humidity**
2) **Reanalysis Dew Point Temp**

Predictive Modeling

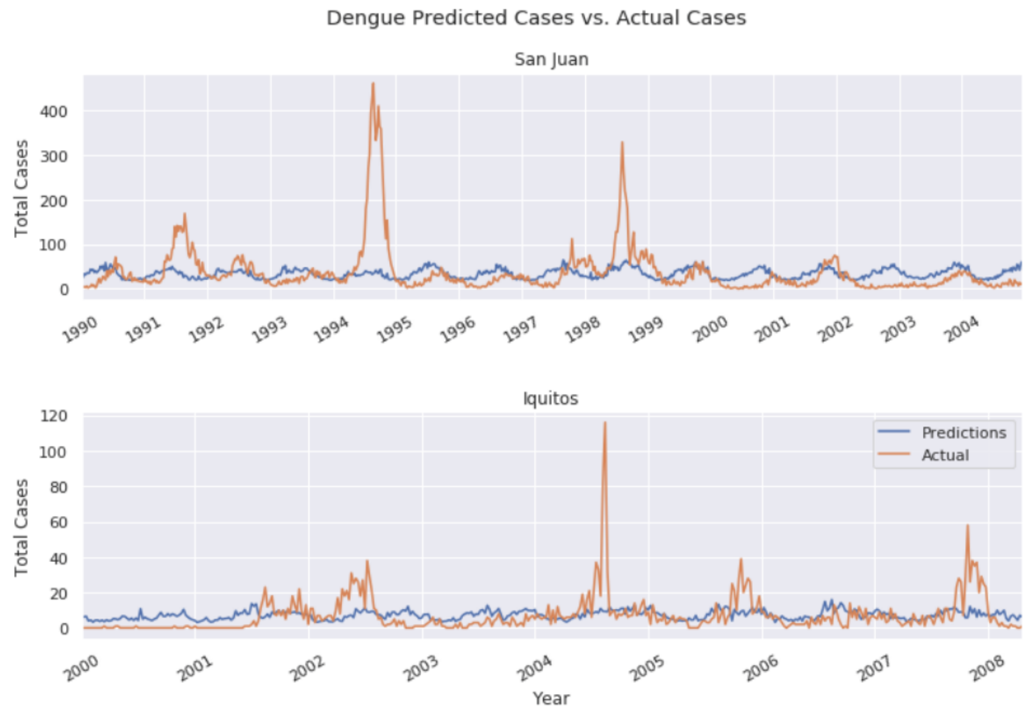
Generalized Linear Models - Benchmark

Poisson Regression

- `reanalysis_specific_humidity_g_per_kg`
- `reanalysis_dew_point_temp_c`

```
Test Poisson Regression Model using
Humidity and Dew Point.
--------------------
San Juan:
score =  24.5455

Iquitos:
score =  7.0481
--------------------
```



Variance >> Mean = Not Appropriate

Predictive Modeling

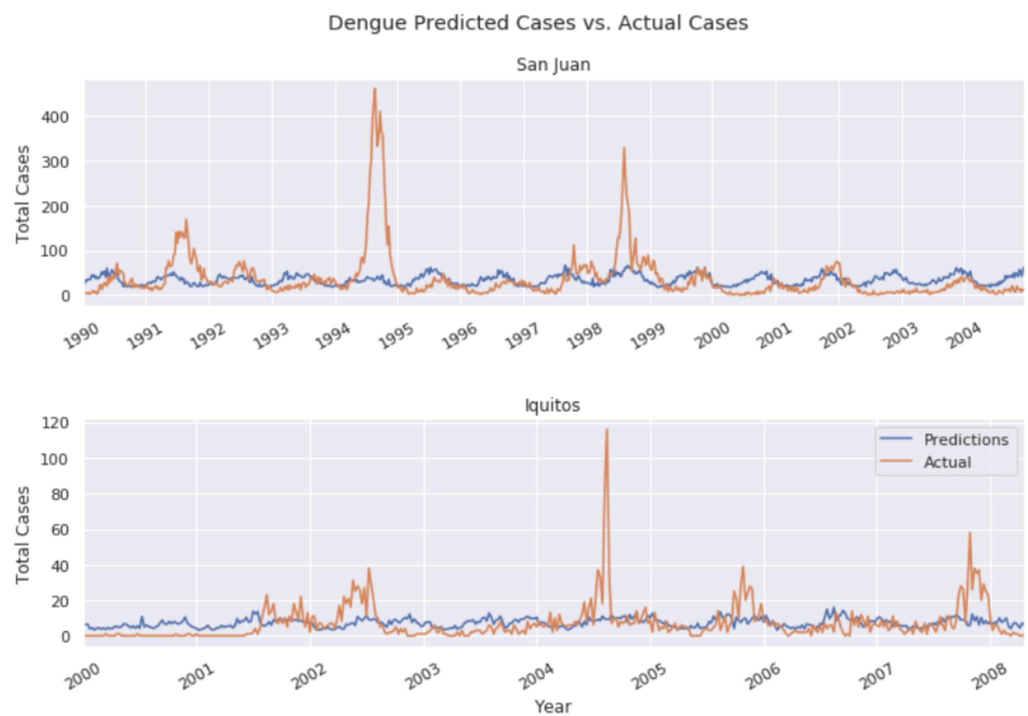
Generalized Linear Models - Benchmark

Negative Binomial Regression

- reanalysis_specific_humidity_g_per_kg
- reanalysis_dew_point_temp_c

```
Test Negative Binomial Regression Model
using Humidity and Dew Point.
--------------------
San Juan:
best alpha = 1.0
best score =  24.4011

Iquitos:
best alpha = 0.0001
best score =  7.0385
--------------------
```



Variance >> Mean = Appropriate

Time Series Lag Adjustment

Given optimal climate conditions when mosquitoes are most likely to breed and deposit eggs, **(1)** there is a period in which the eggs will need time to hatch, **(2)** incubation period after ingesting the Dengue virus, and **(3)** period before symptoms appear in human once infected.

This means that we will need to shift the data, inserting a **Time Series lag** accordingly.

(1) 8-10 days for *aedis aegypti* to develop from egg to full-grown mosquito

(2) 8-10 days for Dengue virus to incubate once ingested by mosquito

(3) 4-13 days for infected Human to show symptoms

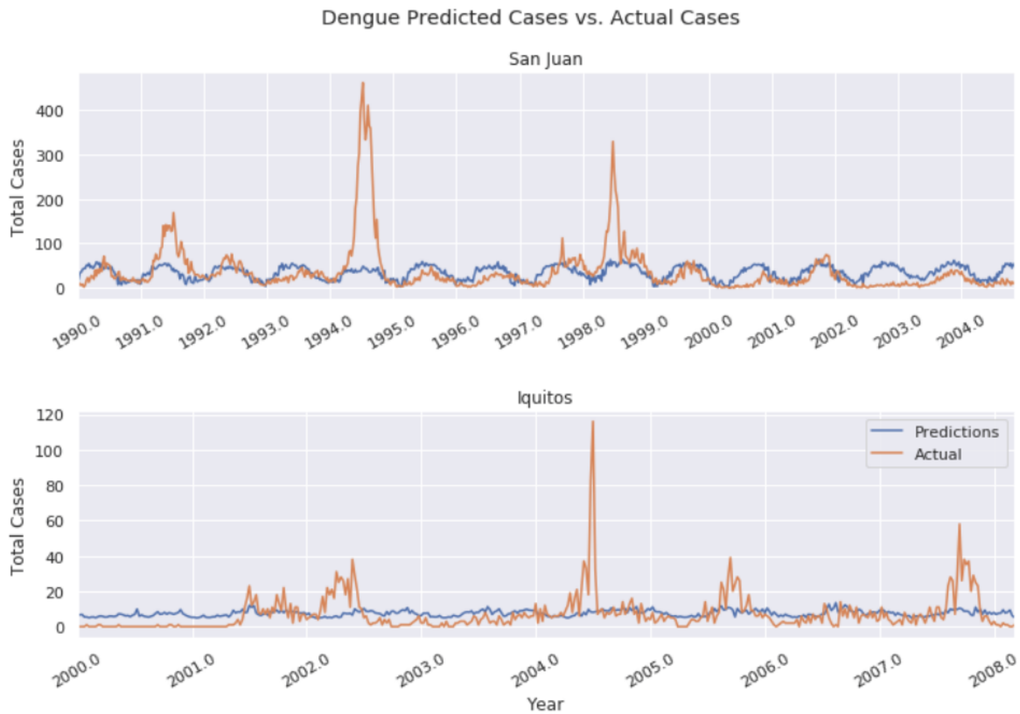
Time Series Lag Adjustment

Observations

- Applied **6-week shift** to San Juan and Iquitos Data

Resulted in an improved Mean Absolute Error (MAE) for San Juan at **23.4309**. The Predicted cases aligned better with Actual cases.

Because the MAE was poor at **8.6224** when applying a 6-week shift and the plot did not show any difference, **no shift was used for Iquitos in downstream analysis**.



Feature Selection

Feature scaling

The **RobustScaler** was chosen to scale feature data because of the presence of **outliers** in the data. The RobustScaler uses a similar method to the **MinMaxScaler**, shrinking the range between 0-1, however, instead it uses the **interquartile range** to handle the **outliers**.

Highly correlated features

Features that had a **99% correlation** were removed. For San Juan, there were three (3) features that were removed while Iquitos required that two (2) features be removed. The list of dropped features and the correlated features kept in the data set can be seen below.

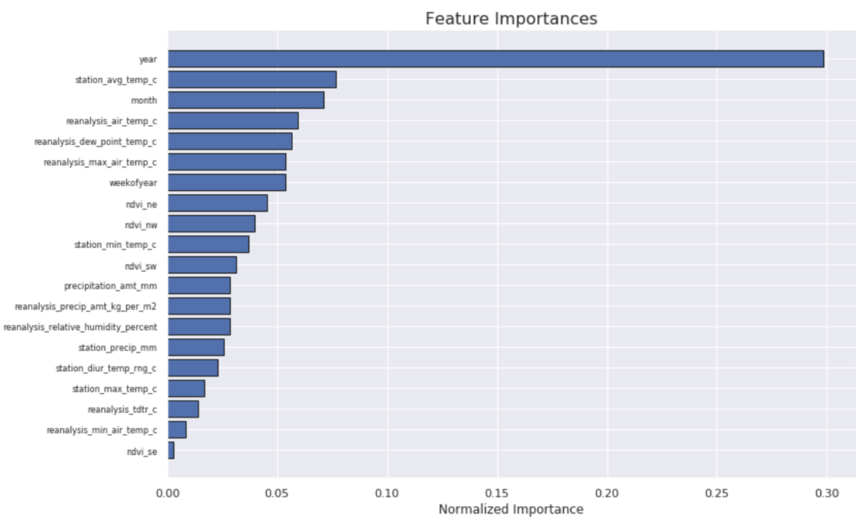
city	drop_feature (dropped)	corr_feature (kept)	corr_value
San Juan	reanalysis_avg_temp_c	reanalysis_air_temp_c	0.997268
San Juan	reanalysis_sat_precip_amt_mm	precipitation_amt_mm	1.000000
San Juan	reanalysis_specific_humidity_g_per_kg	reanalysis_dew_point_temp_c	0.998477
Iquitos	reanalysis_sat_precip_amt_mm	precipitation_amt_mm	1.000000
Iquitos	reanalysis_specific_humidity_g_per_kg	reanalysis_dew_point_temp_c	0.997894

Feature Selection

Feature importance

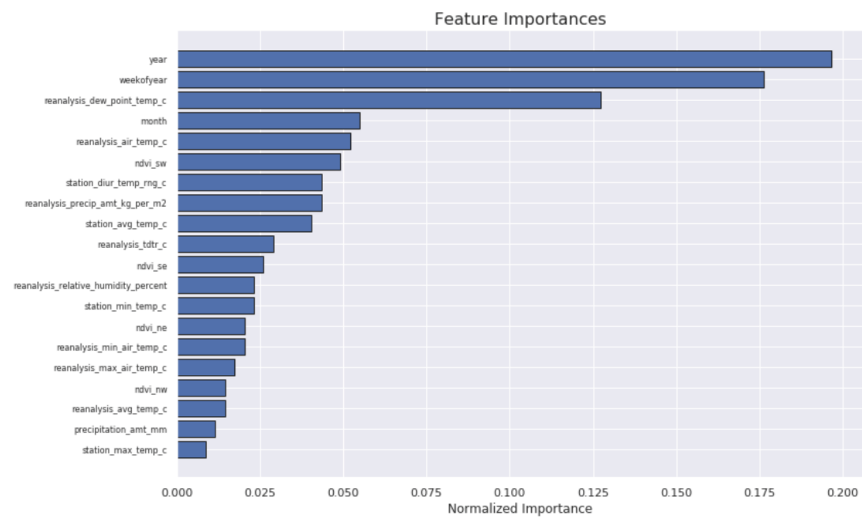
Features that provided zero contribution to **99% of the cumulative importance** were removed. **XGBoost Regressor** combined with validation to avoid 'overfitting' was used to determine feature importance.

San Juan



Top Feature Importance = **Year**

Iquitos



Top Feature Importance = **Year**
weekofyear
reanalysis_dew_point_temp_c

Negative Binomial Regression Model

As a result of feature selection and re-fitting our model using all the remaining features, for San Juan, we get a new score of **17.663** (**an improvement of -6.7381**).
However, for Iquitos, we actually get a new score of **9.6154** (**a decline of +2.5769**).

New Approach for Iquitos:

1) Create every combination of features as a model formula.
2) Select all 65,538 combinations that include:
   1) `‘month’ + ‘weekofyear’ + ‘ndvi_ne’` in the formulas
3) Test Negative Binomial Regression model for Iquitos using the 65,000+ formulas.
4) Select best alpha, score, and formula combination.

As a result, the best R-string model formula was:

```
‘total_cases ~ 1 + month + weekofyear + ndvi_ne + ndvi_nw + precipitation_amt_mm +
reanalysis_avg_temp_c + reanalysis_relative_humidity_percent + reanalysis_tdtr_c’
```

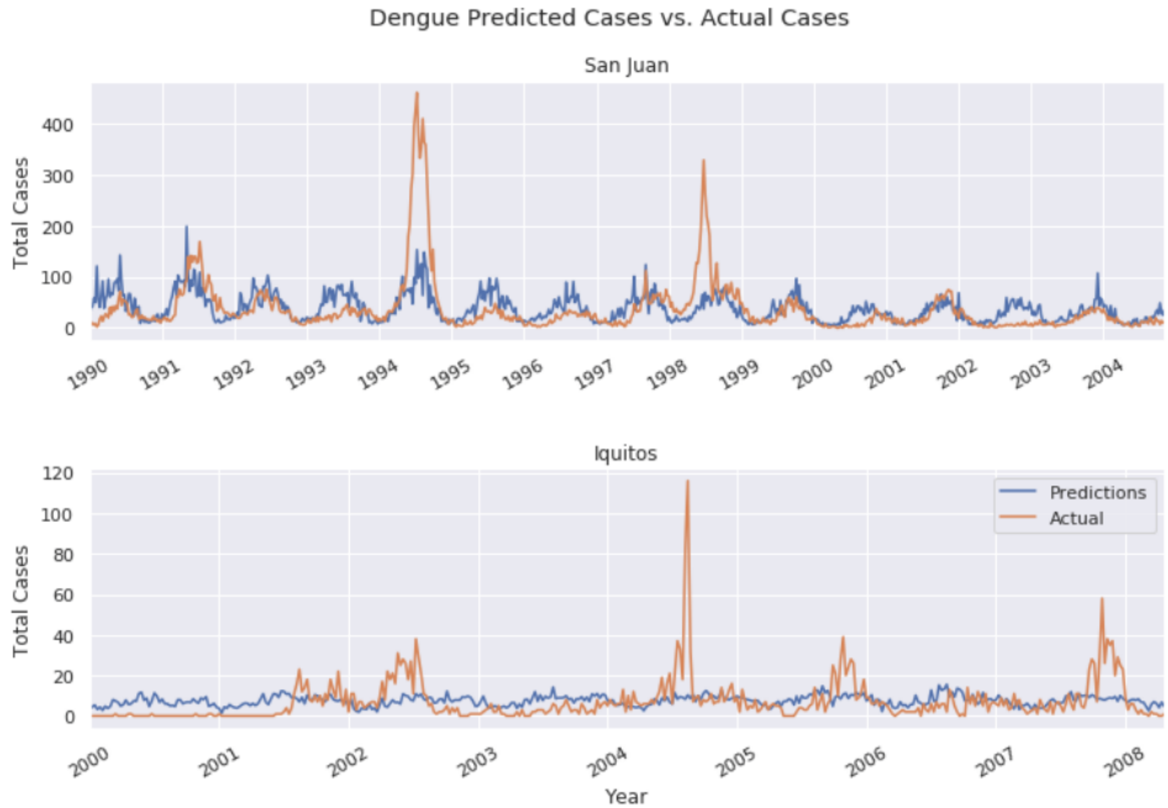
Negative Binomial Regression Model

Observations

Test Negative Binomial Regression Model using Best Model Formulas.

```
--------------------
San Juan:
best alpha = 0.0
best score = 17.663

Iquitos:
best alpha = 1.0
best score = 6.5865
--------------------
```



Improvement in predicting some of the slightly higher spikes in cases reported!

Advanced Predictive Modeling

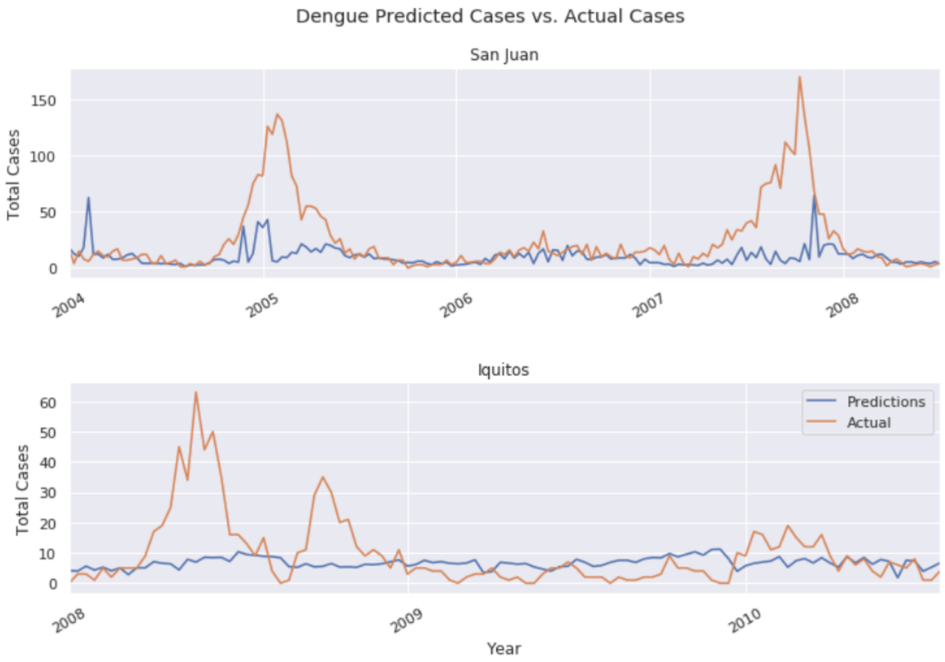
To further investigate if it is possible to improve **MAEs** for San Juan and Iquitos, we attempt to train the following **Advanced Predictive Models**:

- XGBoost
- Seasonal AutoRegression Integrated Moving Average (ARIMA)
- Deep Learning - Tensor Flow Neural Network

XGBoost

The first approach was to train an **XGBoost Regressor Model** using optimal parameters selected from 4,200 unique fits using an exhaustive **GridSearchCV**.

As a result, we were only able to generate an **MAE** of **17.0246** for **San Juan** (**which is an improvement**) and **7.1693** for **Iquitos**.



Seasonal ARIMA

ARIMA is a forecasting method for univariate time series data forecasting and while it can handle data with trends, it does not support time series with a seasonal component.

It adds three new hyperparameters to specify the **autoregression (AR)**, **differencing (I)** and **moving average (MA)** for the seasonal component of the series, as well as an additional parameter for the period of the seasonality.

Trend Elements

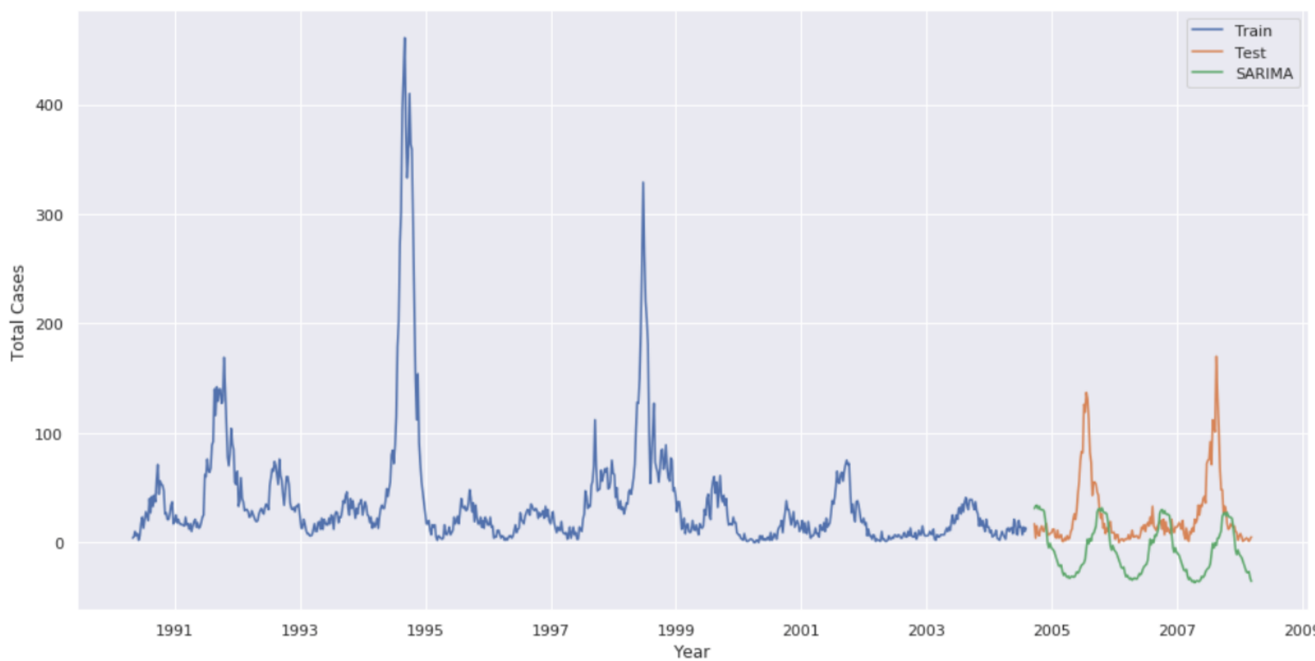
- **p:** Trend autoregression order
- **d:** Trend difference order
- **q:** Trend moving average order

Seasonal Elements

- **P:** Seasonal autoregressive order
- **D:** Seasonal difference order
- **Q:** Seasonal moving average order
- **m:** The number of time steps for a single seasonal period
  (52 weeks in our case)

Seasonal ARIMA
San Juan, Puerto Rico

Seasonal ARIMA for San Juan



Observation

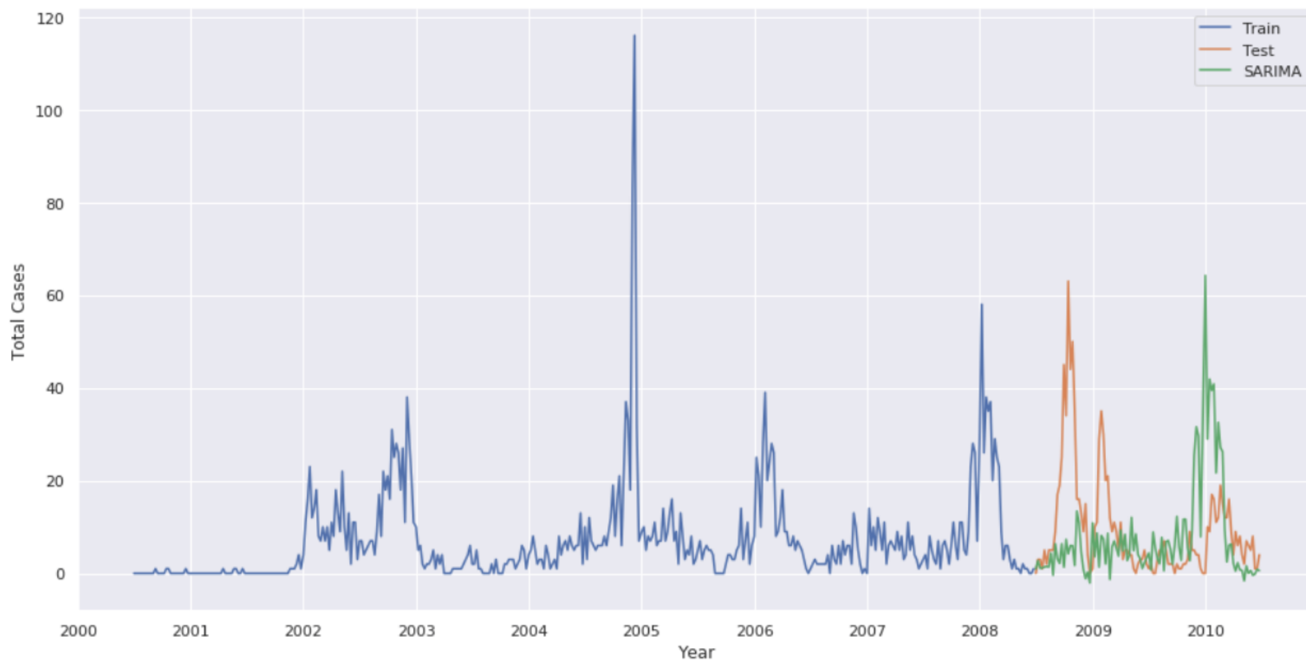
```
--------------------
San Juan:
MAE of 36.9862
--------------------
```

Seasonal ARIMA
Iquitos, Peru

Observation

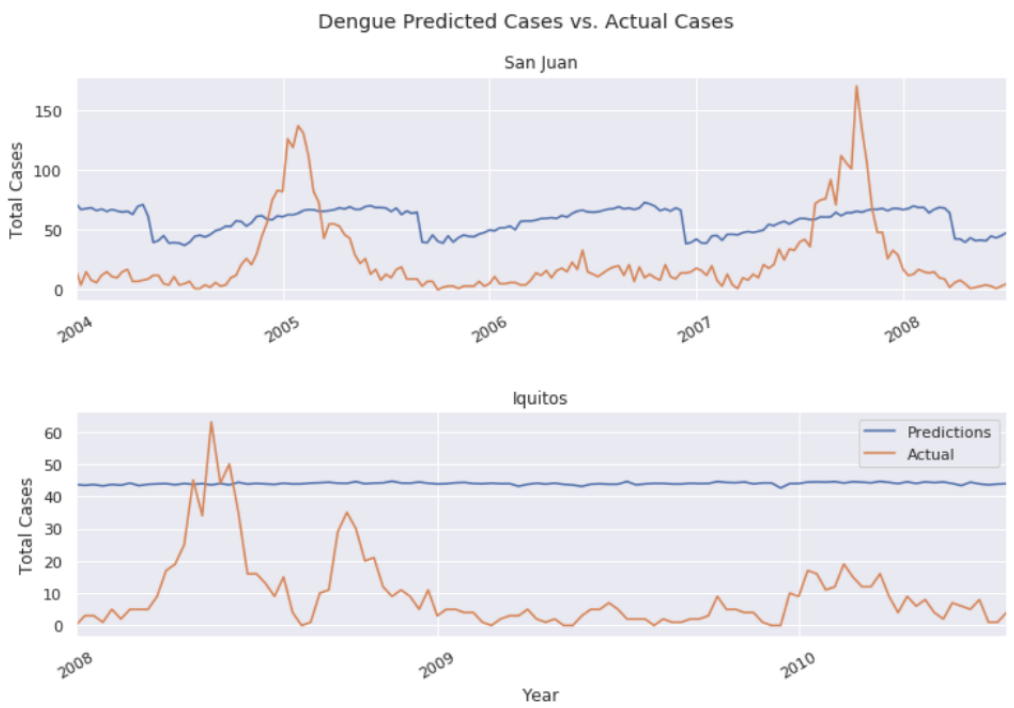
```
--------------------
Iquitos:
MAE of 10.1451
--------------------
```

Seasonal ARIMA for Iquitos



Deep Learning – TensorFlow Neural Networks

- We trained the model for 1,000 epochs, and recorded the training and validation accuracy in the history object.
- To further optimize the model and prevent overfitting, we applied **EarlyStopping** callback that tests a training condition for every epoch.
  - If a set amount of epochs elapsed without showing improvement, then it automatically stopped the training.
- **As a result, our MAE for San Juan and Iquitos were significantly worse at 41.60 and 34.96, respectively.**



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)
- Imputed missing and null values using median values
- Used `fillna(…)` with forward-fill method where medians could not be computed

- Approximately a 9 year overlap in observation between San Juan
- **Positively skewed** Distributions for Total Cases
- Seasonality for Time Series data
- Outliers Detected
- Features are not ‘highly’ correlated to the labeled data (total_cases)
- We see some correlation in features vs. features that may provide some insights

- San Juan required a **6-week shif**t to account for mosquito growth, ingest and incubation of dengue, transmission of dengue to human, and signs of symptoms

- Data was scaled using **Robust Scaler** and Features with high correlation and low importance were removed.

Findings

Additional models (**XGBoost**, **Seasonal ARIMA**, and **TensorFlow NN**) were implemented to train and test the data for both cities to try and improve overall **Mean Absolute Error (MAE)**.

As a result, the best model for **San Juan** was **XGBoost** with an **MAE** of **17.0246**.

	Model	Mean Absolute Error (MAE)
Naive Approach with Default Settings	Poisson Regression	24.5455
	Negative Binomial Regression	24.4011
6-week Shift Included	Negative Binomial Regression	23.4309
Shift + Feature Selection	Negative Binomial Regression	17.663
	XGBoost	**17.0246**
	SARIMA	36.9862
	TensorFlow NN	41.6

The **Negative Binomial Regression Model** using a custom model formula was the best model for **Iquitos** data with an **MAE** of **6.5865**.

	Model	Mean Absolute Error (MAE)
Naive Approach with Default Settings	Poisson Regression	7.0481
	Negative Binomial Regression	7.0385
	Negative Binomial Regression	8.6224
Feature Selection	Negative Binomial Regression	9.6154
Feat. Selection + Custom Model Formula	**Negative Binomial Regression**	**6.5865**
	XGBoost	7.1693
	SARIMA	10.1451
	TensorFlow NN	34.69

Final Comments

Overall, we were able to predict total cases for both San Juan and Iquitos with an improved measure of difference between Predicted vs. Actual Dengue cases reported.

We were also not able to forecast total cases 1 or more years in advance and when we encountered high spikes in total cases of Dengue reported.

Moving forward, we we should consider additional data such as 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.