Disease Spread

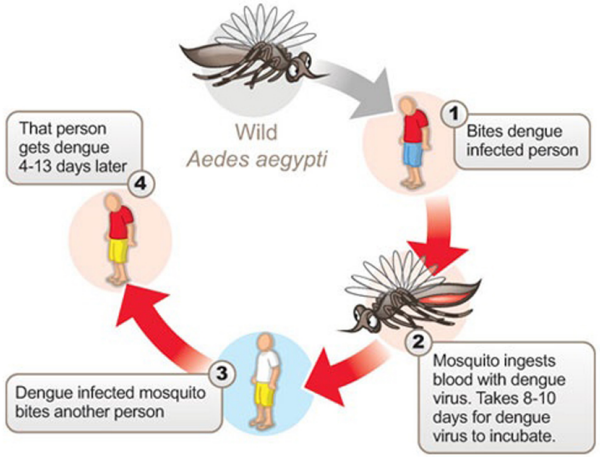
Predicting the Spread of Dengue

Springboard Data Science Career Track

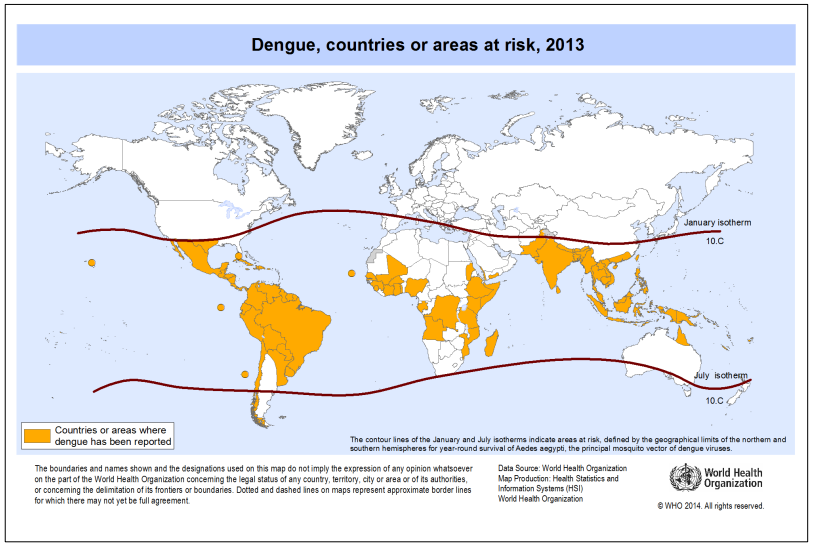
– Mark Rojas

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:

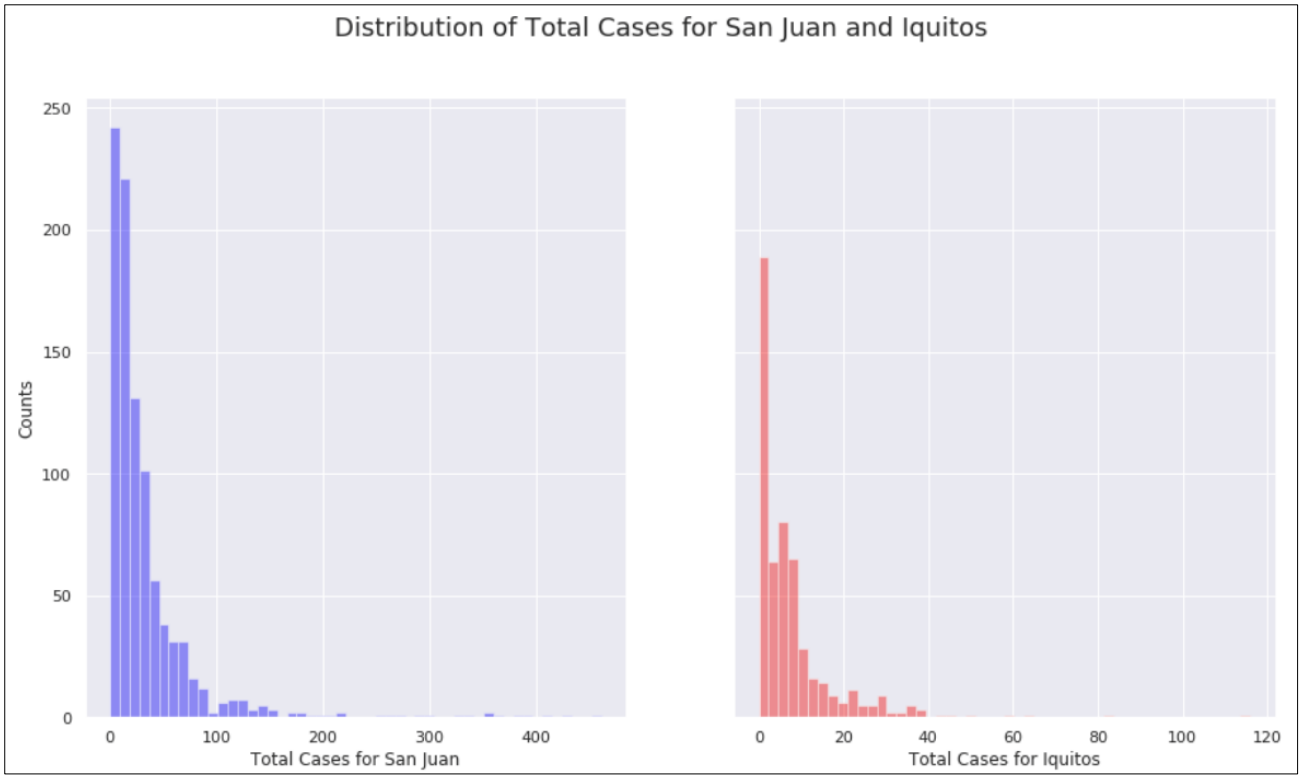
- For **San Juan**, **20** of the **24** features contained between **6 – 191 null values**
- For **Iquitos**, **20** of the **24** features contained between **3 – 37 null values**

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)

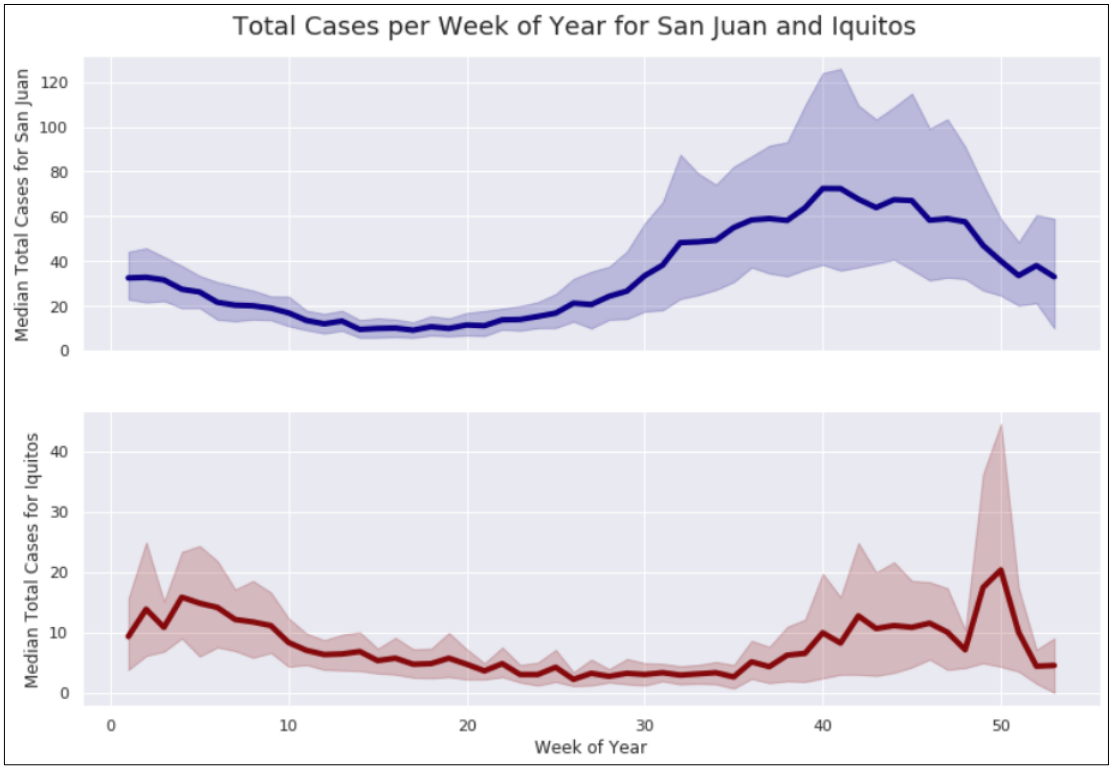


- Positively Skewed distributions
- Predictive model will need to be optimized for outliers
- Labeled data will need to be normalized

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

Exploratory Data Analysis

Total Cases over 53-week Period



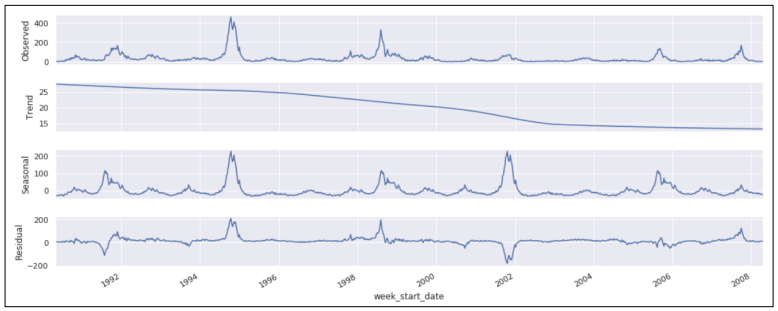
- **More cases** reported at the **end of the year** in **San Juan** than in **beginning of year**
- Similar trend in number of cases reported in **Iquitos** as for **San Juan**

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

Exploratory Data Analysis

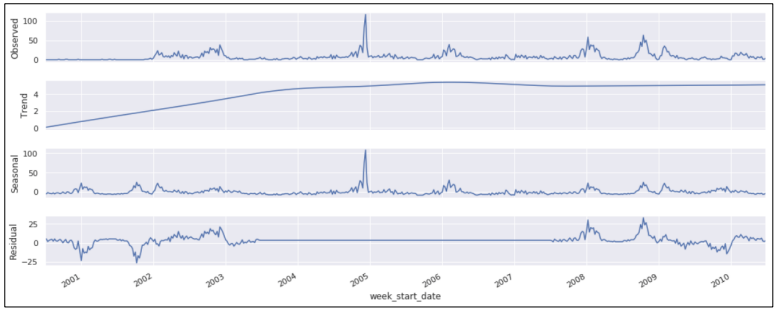
Time Series Decomposition

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.

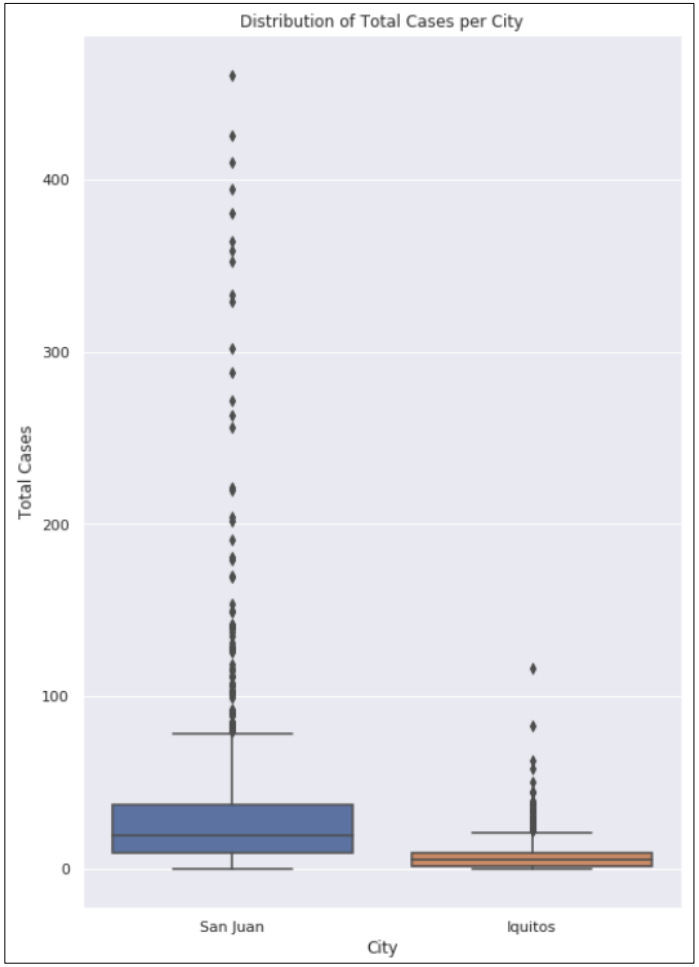
- **20 outliers** in the **San Juan** label data
- **7 outliers** in the **Iquitos** label data

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

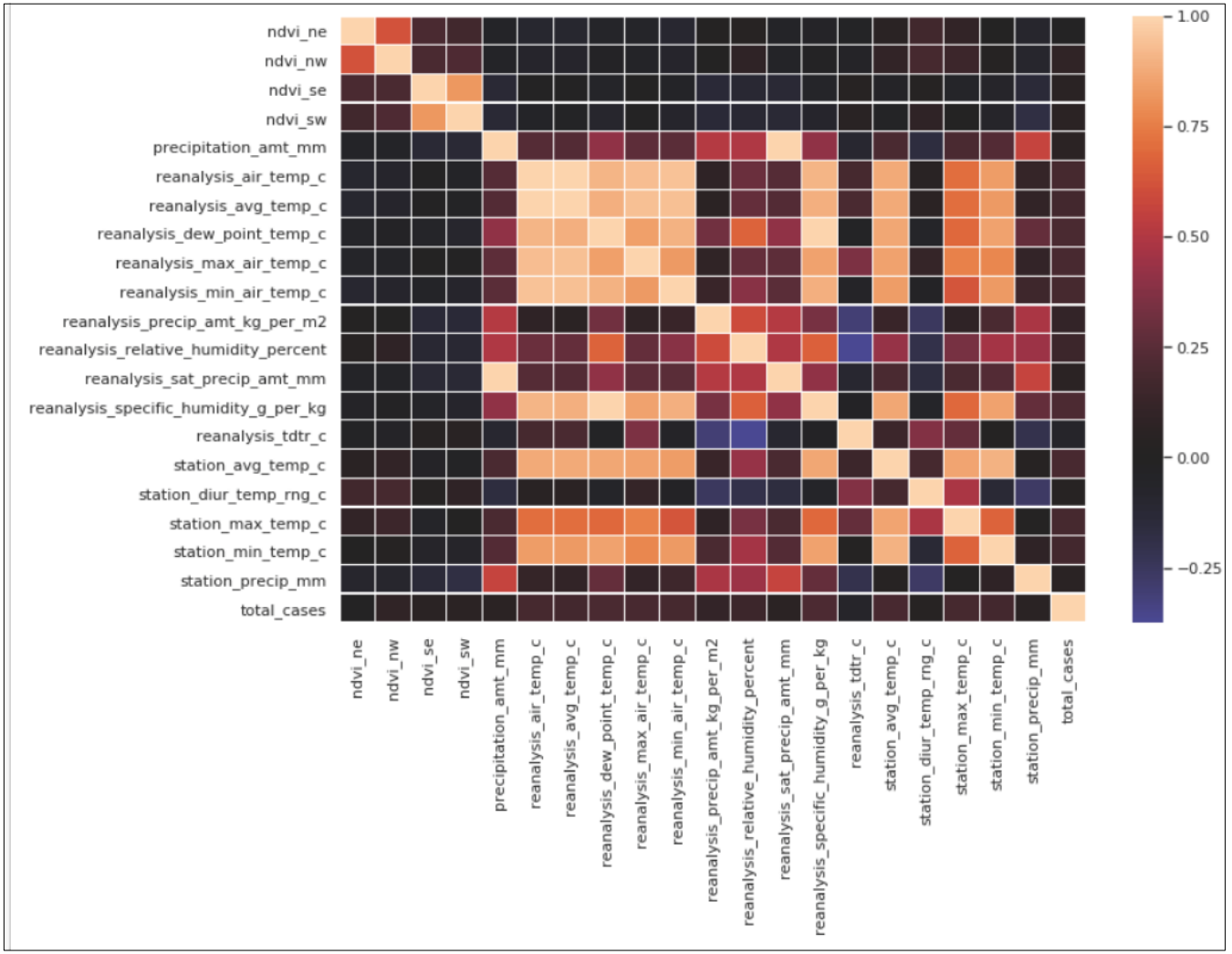


Exploratory Data Analysis

Correlation: Features vs. Total Cases

San Juan

Correlation Matrix

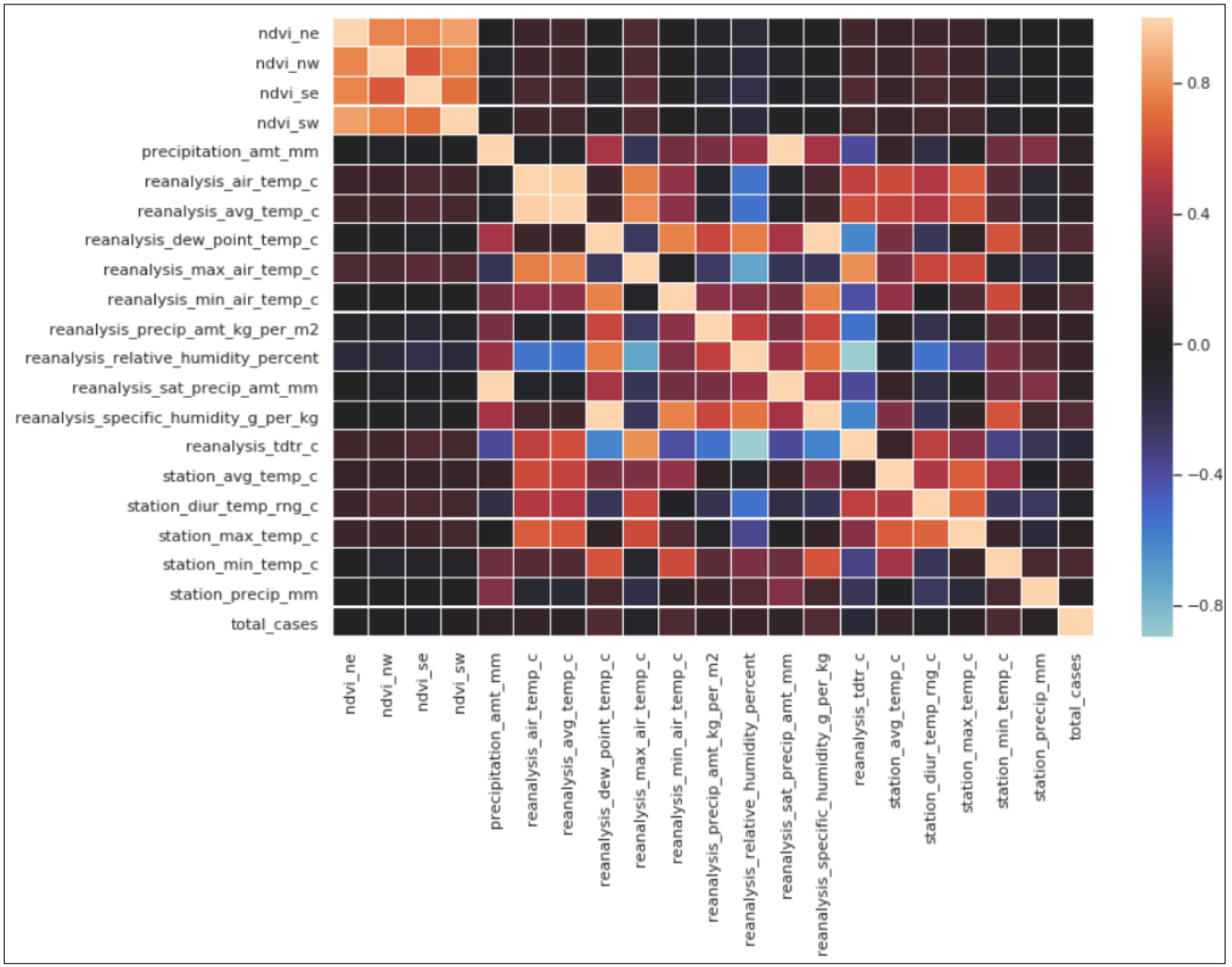


Exploratory Data Analysis

Correlation: Features vs. Total Cases

Iquitos

Correlation Matrix

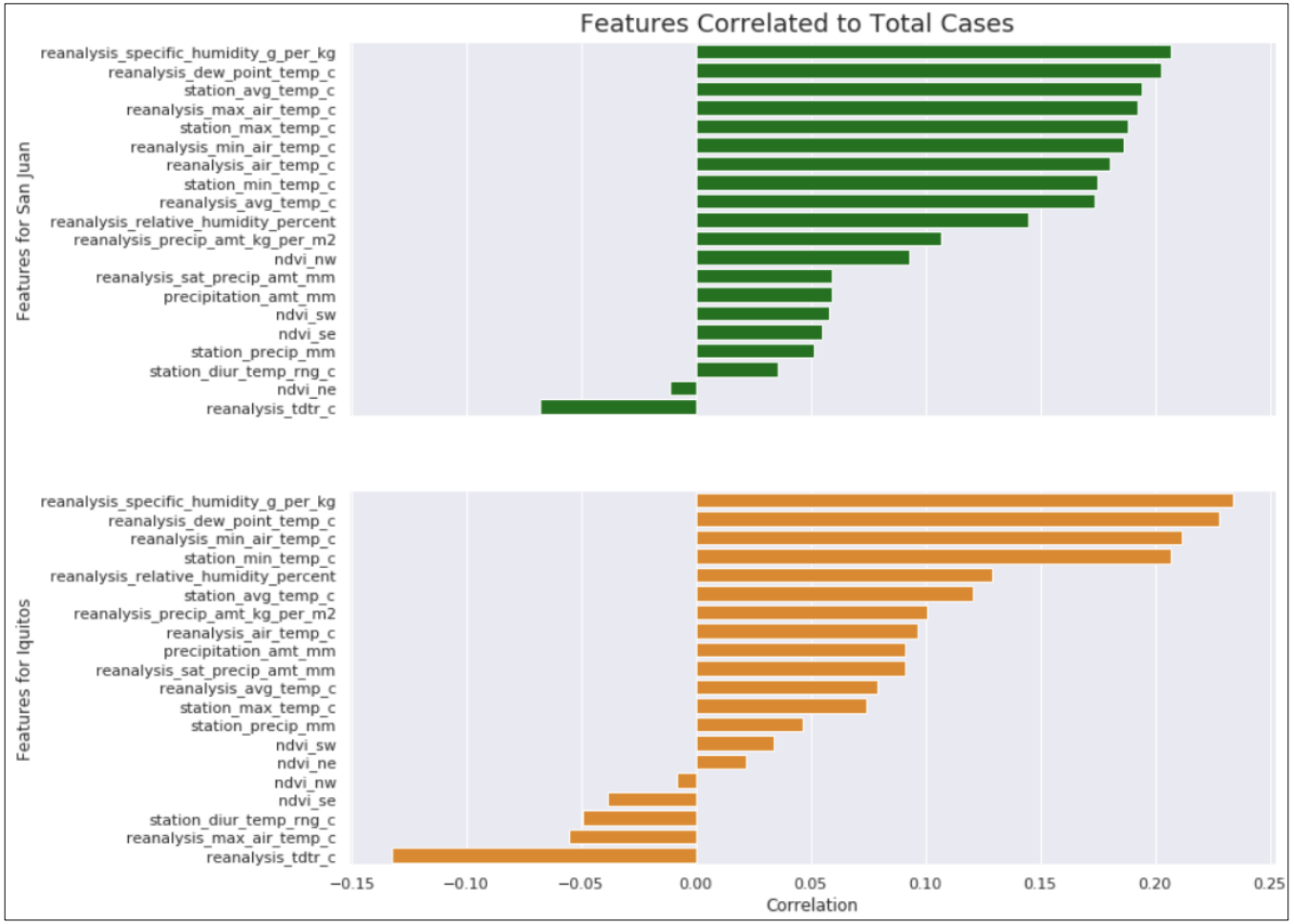


Exploratory Data Analysis

Features Correlated to Total Cases

Moisture in the Air!!

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



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)
  - 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

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.

2 and 3 week lag between features and labels required for incubation and symptom periods.

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!