Dengue Fever Cases in San Juan and Iquitos

Group 4: Akram Ahmadi, Yewon Kang, Haroon Malik, Tammy Nguyen, Fani Hsieh

4/14/20

Table of Contents

I. Introduction

Dengue fever is a mosquito-borne disease that commonly occurs in tropical and sub-tropical parts of the world.

Environmental data was collected by various U.S. Federal Government agencies for San Juan, Puerto Rico and Iquitos, Peru. The main environmental data points observed are:

- Temperature
- Precipitation
- Humidity
- Vegetation

This dataset is currently being used in a competition hosted by DrivenData with the following problem statement: “Can you predict local epidemics of Dengue fever?”

The scope of our analysis will be focused on understanding the environmental variables and determining whether there are key environmental features that lead to a higher number of Dengue fever cases. Our problem statement is: **“What environmental features contribute to a higher number of Dengue Fever cases in San Juan and Iquitos?**

The analysis explores:

1. Whether there are key environmental factors overall that contribute to the high number of Dengue Fever cases and;
2. Whether there are key environmental differences between San Juan and Iquitos.

II. Data Preparation

a. Data Source

The data was from open data on the DrivenData Competition website:
https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/page/82/

To access the datasets, a user account needed to be set-up and the user needed to join the competition. Once this was set-up, there were 2 datasets that need to be downloaded:

- “dengue_features_train.csv”
- “dengue_labels_train.csv”

b. Data Quality

The data quality overall was good. Referring to the section “**2.a. Understanding the Data**”, 20 out of the 24 data columns contain missing data but NaN values for most data fields were generally below 1% for most of the data fields.

Features	# of NaN Values	% NaN Values
ndvi_ne	194	13.3%
ndvi_nw	52	3.6%
ndvi_se	22	1.5%
ndvi_sw	22	1.5%
precipitation_amt_mm	13	0.9%
reanalysis_air_temp_k	10	0.7%
reanalysis_avg_temp_k	10	0.7%
reanalysis_dew_point_temp_k	10	0.7%
reanalysis_max_air_temp_k	10	0.7%
reanalysis_min_air_temp_k	10	0.7%
reanalysis_precip_amt_kg_per_m2	10	0.7%
reanalysis_relative_humidity_percent	10	0.7%
reanalysis_sat_precip_amt_mm	13	0.9%
reanalysis_specific_humidity_g_per_kg	10	0.7%
reanalysis_tdtr_k	10	0.7%
station_avg_temp_c	43	3.0%
station_diur_temp_rng_c	43	3.0%
station_max_temp_c	20	1.4%
station_min_temp_c	14	1.0%
station_precip_mm	22	1.5%

c. Challenges

There were 3 main challenges encountered upon examining the dataset:

1. Understanding and interpreting the environmental factors;
2. Figuring out which values to use in the case where there were multiple variations of the same environmental factors (mainly temperature and precipitation values); and
3. How to appropriately incorporate the values into the analysis and understanding how the variables interact with one another.

To gain a better understanding of the specific environmental terminology and metrics, supplemental research was conducted. In the case where there were multiple variations of the same environmental factors, all related values were examined. If the values resulted in consistent patterns or didn't show any significance, the values were dropped from the analysis. This is further discussed in section "**III. Analysis**" of this report.

d. Tools and Code Required for Analysis

i. Joining Labels dataset to the Features dataset

The "`total_cases`" data field is located in the "dengue_labels_train.csv" file. A new DataFrame was created to combine "`total_cases`'' to the features dataset, "dengue_features_train.csv". The join combined the two datasets using the common columns '`city`', '`year`', and '`weekofyear`'. Refer to Notebook 1 Section 2.a Understanding the Data for the code.

ii. Filing in NaN values

As the environmental features are specific to the location, the dataset was split into two DataFrames filtered by city to avoid imputing one city's data into the other. Once split, different approaches were used to fill in the NaN values depending on the type of environmental factor (shown in the next table). Refer to Appendix A for a summary outlining the fill approach.

iii. Converting Date into DateObject for Time Series

The data field "`week_start_date`" was initially an "object" datatype. To use this field in any time series analysis, the datatype needed to be converted to a DateObject datatype. Refer to Notebook 1 Section 2.c.1 Transforming Data for the code.

iv. Created a New Data Fields:

a. Mapping NDVI Values to Vegetation Type

The NDVI values ('`ndvi_ne`', '`ndvi_nw`', '`ndvi_se`' and '`ndvi_sw`') were mapped to the type of vegetation. Details of the analysis are found in Section III. Analysis – Attribute #4. Four new columns were created and merged to the main DataFrame using the lambda function to assign the vegetation type if the value satisfied the rules outlined below (refer to Appendix B for an excerpt of the code).

Vegetation Type	Logic Rules
Water	x<-0.1
Barren	x<=0.1 and x>=-0.1
Grassland	x<=0.4 and x>0.1
Tropical	x<=1.0 and x>0.4
Unknown	Everything else

b. Mapping City Initials to Full City Name

In the original dataset, the City was indicated by initials:

City Initials	City Name
sj	San Juan
iq	Iquitos

The following code was used to map the City initials to the City name:

```
df_sj = df_merge[df_merge["city"] == "San Juan"]
df_iq = df_merge[df_merge["city"] == "Iquitos"]
```

III. Analysis

Overall Trends, Correlations and/or Patterns

San Juan has more Dengue fever cases compared to Iquitos. According to *Figure 1 (a)*, San Juan had more total cases than Iquitos. But *Figure 1 (b)* reveals that San Juan has a higher number of total cases because they started recording the occurrence of Dengue cases 10 years before Iquitos. This observation is also shown in *Figure 1 (c)*.

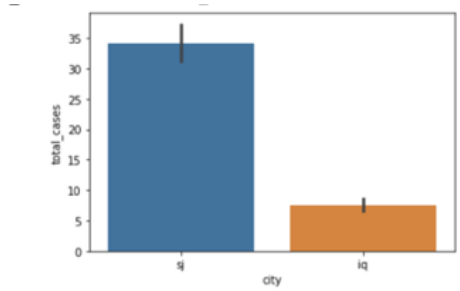


Figure 1 (a) Bar Chart of Total Cases by City

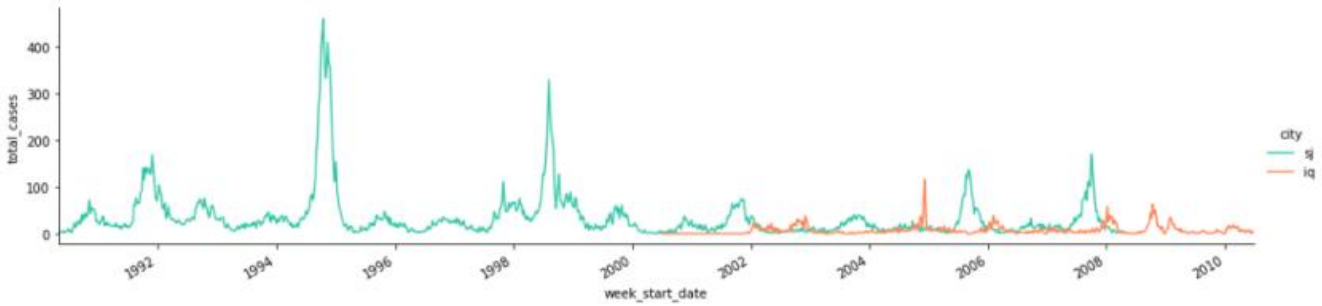


Figure 1 (b) Total Cases by City Over Time

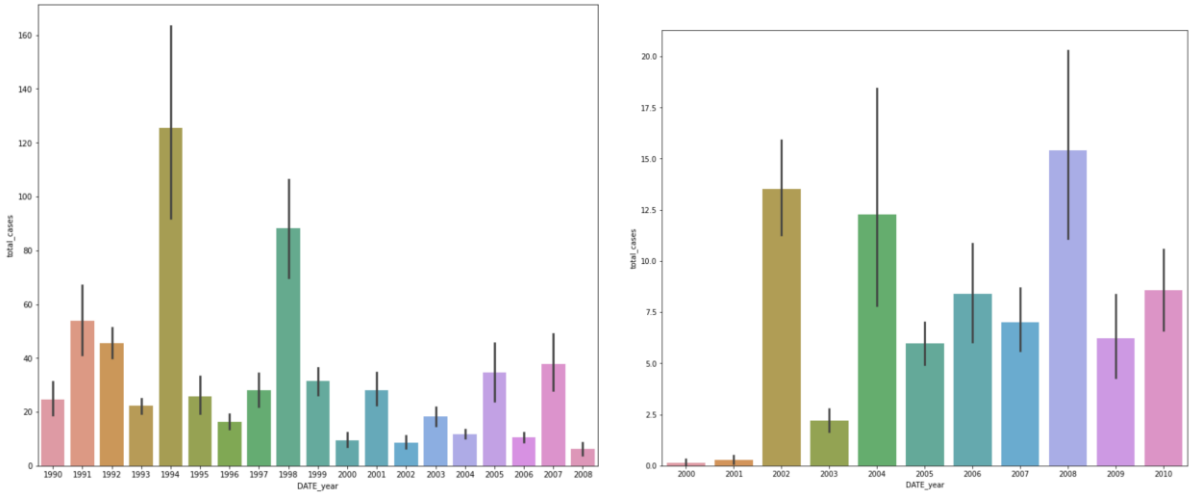


Figure 1 (c) Total Cases by in San Juan (Left) and Iquitos (Right) Over Time

In the left plot of *Figure 1 (c)*, the number of Dengue fever cases in San Juan peaked in 1994 and 1998. In comparison, Dengue cases were significantly lower in subsequent years, specifically 2000, 2002, 2004 and 2008. San Juan had no recorded cases in 2009 and 2010.

Right plot of *Figure 1 (c)* shows that there were no cases of the disease in Iquitos until 2000. Then followed by a sharp increase in cases in 2002. The next three time series plots show how the total cases trend between the weeks of a given year for each city.

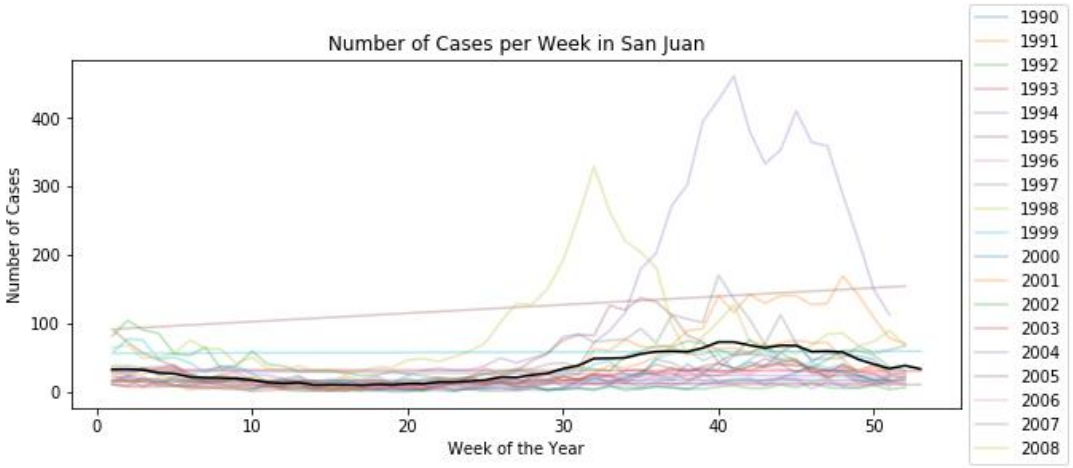


Figure 2 (a) - Total Cases by Week of Year in San Juan

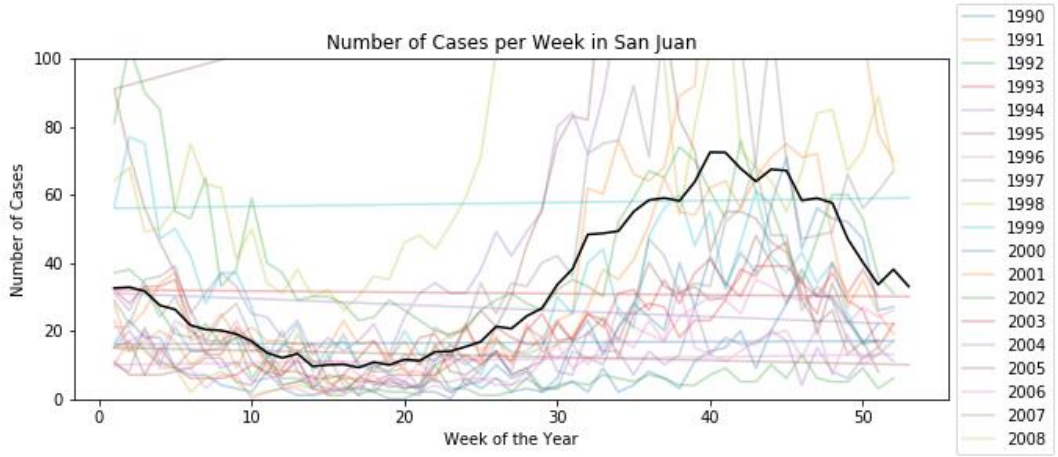


Figure 2 (a-1) - Total Cases by Week of Year in San Juan

Note: The y-axis was adjusted to reduce the max of the range to focus on the total cases pattern.

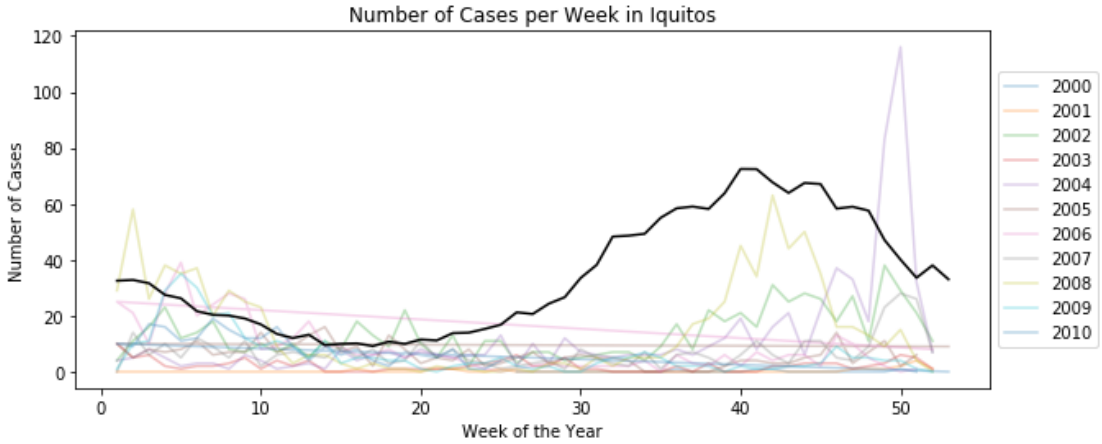


Figure 2 (b) - Total Cases by Week of Year in Iquitos

Figure 2 (a) shows that Dengue fever cases started appearing in San Juan in 1990. Colour plots indicate the trend by week of year for each year from 1990 to 2008 and the black line represents the average of total cases every week of year. The y-axis in *Figure 2 (a-1)* was adjusted to better emphasize the black average total cases line. Dengue fever cases in San Juan peaked in week 40 in 1994. Dengue fever cases in Iquitos only started appearing 10 years after cases showed up in San Juan. As shown in *Figure 2 (b)*, Dengue fever cases in Iquitos only begun appearing in 2000. Both cities have the same pattern in total cases throughout the year. There are decreasing trends in total cases from January to March then the number of cases begin increasing from April to October. Cases then decrease again from November to December. This trend is further highlighted in *Figure 3 (a)* and *Figure 3 (b)*.

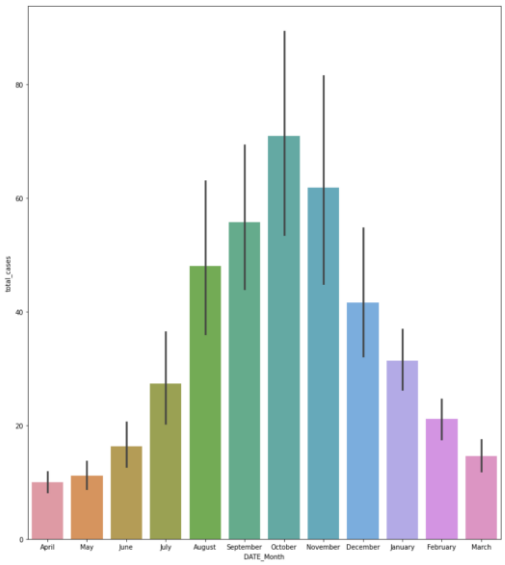


Figure 3(a) - Total Cases by Month of the Year in San Juan

The total cases of disease increase between the months of April to October, which corresponds to the wet summer season and hotter temperatures[1]. The decline in cases from November to March corresponds with the dry season and cooler temperatures[2]. Precipitation, Temperature and Humidity are further explored in the upcoming Attribute sections of this report.

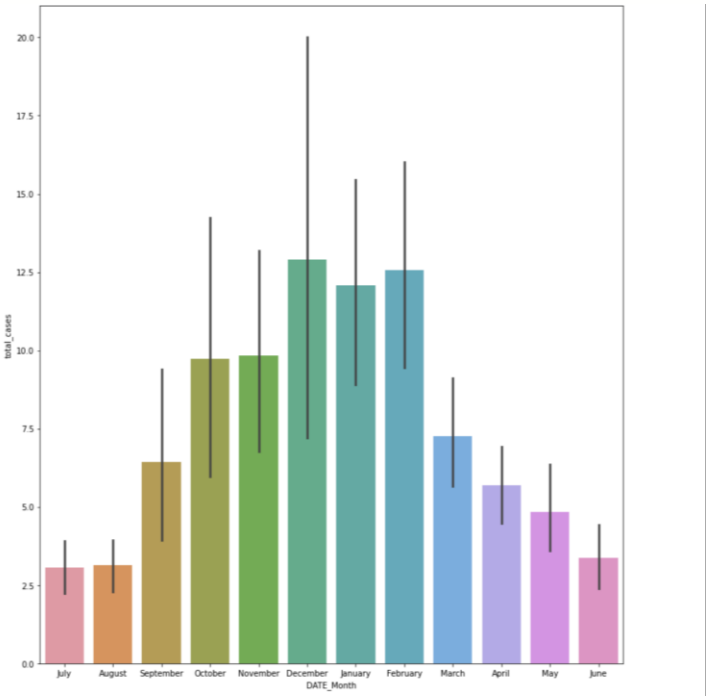


Figure 3 (b) - Total Cases by Month of the Year in Iquitos

Between September and February, which are the peak of the wet season[3], there is an increasing trend of Dengue cases followed by a decline in cases from March to August. This is in line with the dry season that begins in June[4].

The following heat map (*Figure 4*) illustrates how all the environmental attributes in the dataset correlate with other attributes.

[1] Wikipedia. Climate of Puerto Rico. Retrieved from URL: https://en.wikipedia.org/wiki/Climate_of_Puerto_Rico
[2] ibid
[3] Weather Spark. Average Weather in Iquitos. Retrieved from URL: https://weatherspark.com/y/24250/Average-Weather-in-Iquitos-Peru-Year-Round
[4] ibid

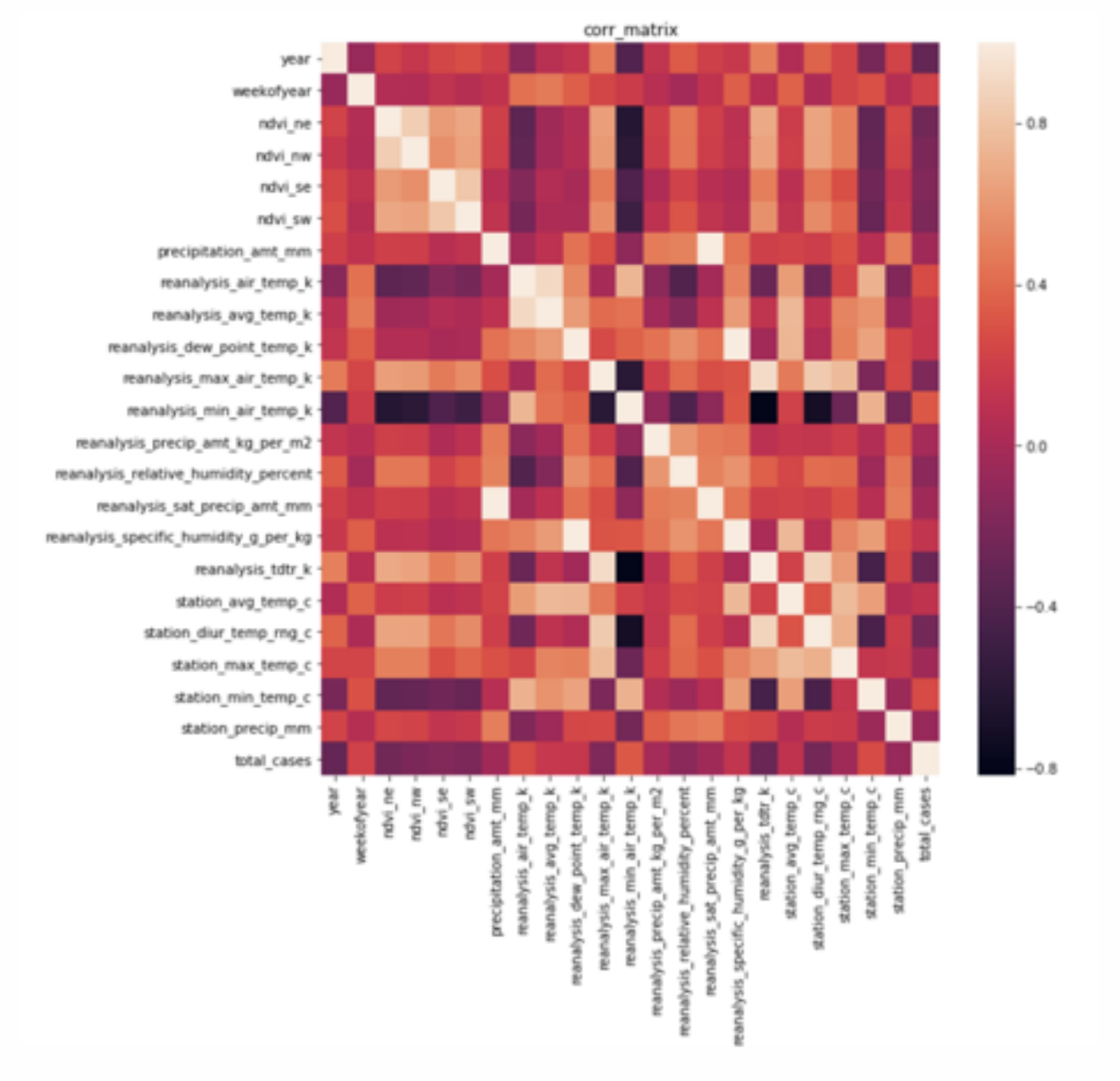


Figure 4 - Correlation Heat Map with All Environmental Factors

Many of the temperature measures are strongly correlated, which is expected. The overall number of Dengue cases (‘`total_cases`’) does not have many obvious strong correlations. Many of the environmental variables are much more strongly correlated with other environmental variables. For example, the precipitation variables bear little to no correlation to `total_cases`, but highly correlated with the humidity variables. The NDVI vegetation index also only has weak correlation with other variables. These correlations help to validate some of the observations in the subsequent attribute analysis.

Attribute #1: Precipitation[5]

There are 4 features that represent precipitation in the dataset. According to the drivendata.org, the difference between the precipitation variables is the units used.

- `precipitation_amt_mm`
- `reanalysis_precip_amt_kg_per_m2`
- `reanalysis_sat_precip_amt_mm`
- `station_precip_mm`

1. Total precipitation over months and Total Cases

The x-axis represents the month of year and y-axis represents the average of each month's total precipitation features in both *Figure 5 (a)* and *Figure 5 (b)*. The blue line is the total cases.

In San Juan, all 4 precipitation features are all relative and show similar trends with the months of the year. Precipitation levels decreased from January to March, increased from March until October and then decreased from October to December. There are ups and downs in the "`station_precip_mm`" feature but generally all precipitation features show an increasing trend from March onwards. Interestingly, the total cases feature has the same trend as the precipitation features. Total cases decreased between January and March and started increasing from March to October, then decreased again from October to December. According to this plot, precipitation and the number of Dengue cases are strongly related.

However, Iquitos experiences a completely opposite trend in precipitation features and total cases. This observation supports the fact that the wet season for Iquitos is opposite to the wet season in San Juan. In Iquitos, precipitation features decreased from January to February, then increased February to March, decreased from March until August and increased from August to December. In terms of the total cases plot, the number of cases is relatively consistent throughout the year and there doesn't appear to be clear patterns related to total precipitation amount.

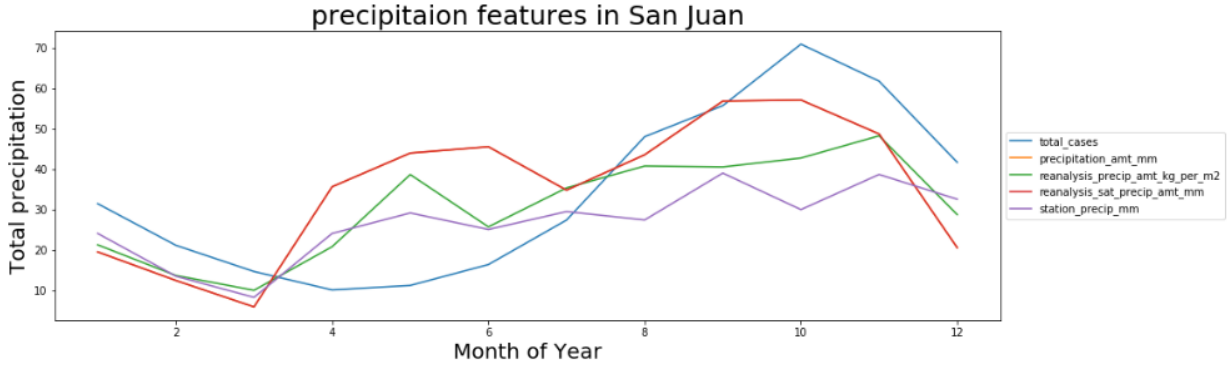


Figure 5 (a) Total precipitation over months and total cases in San Juan

[5] Refer to Notebook 2 for the detailed analysis and code.

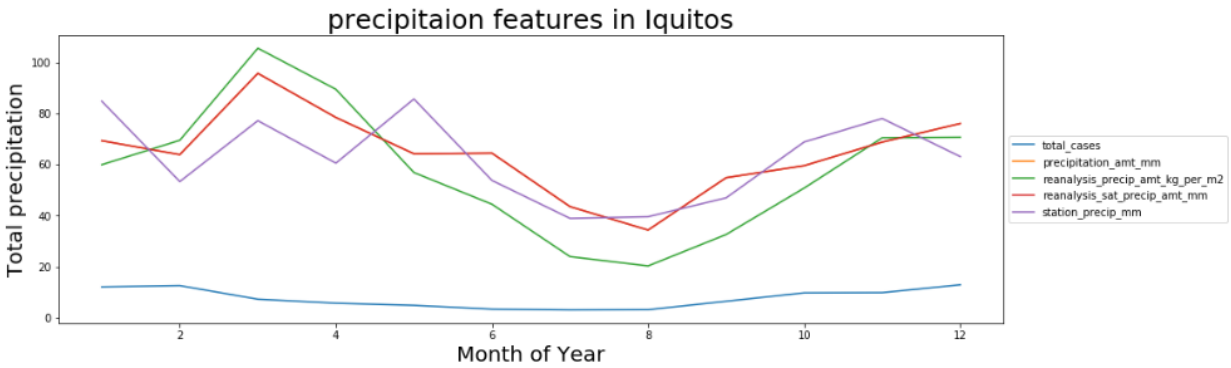


Figure 5 (b) - Total precipitation over months and total cases in Iquitos

2. Total precipitation features over years and total cases

Figure 5 (c) and *Figure 5 (d)* share the same x-axis and y-axis, representing year and average total precipitation respectively. The blue line indicates total cases.

In San Juan (*Figure 5 (c)*), there were two peaks in the number of Dengue cases in 1994 and 1998. Precipitation features have seasonality patterns in the plot and total cases followed those seasonal patterns. Even though precipitation features and total cases behave seasonally, they do not necessarily move in the same direction every year so it is hard to conclude a relationship.

In Iquitos (*Figure 5 (d)*), there were no clear patterns in precipitation features, nor total cases year over year. The plots are relatively flat. Compared to San Juan, the average of total precipitation in Iquitos is higher throughout the year but there is a lower number of recorded Dengue cases.

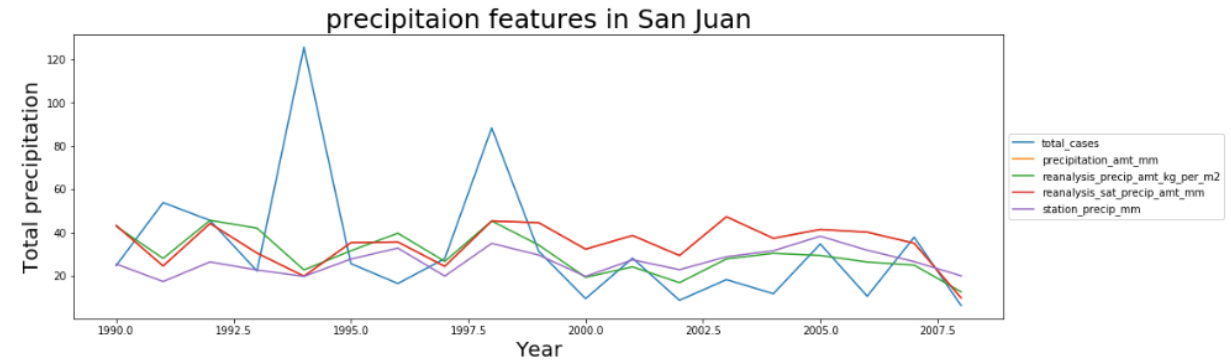


Figure 5 (c) - Total precipitation over years and total cases in San Juan

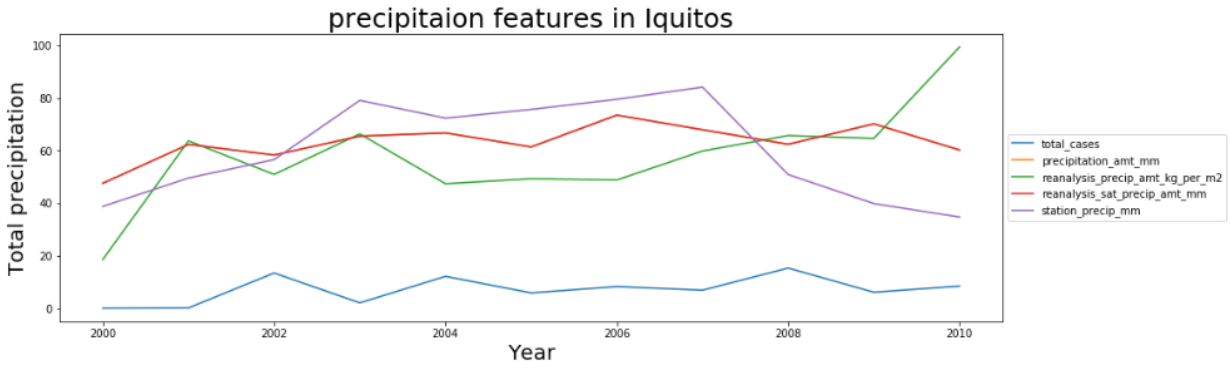
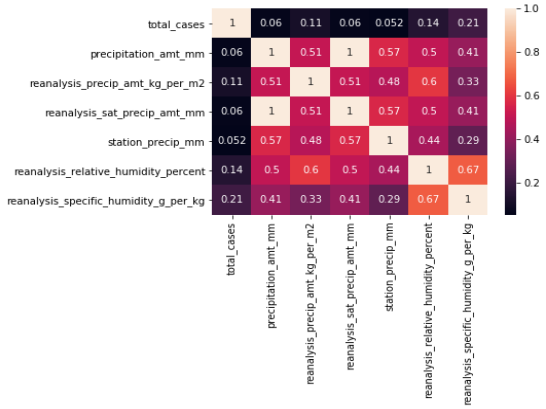


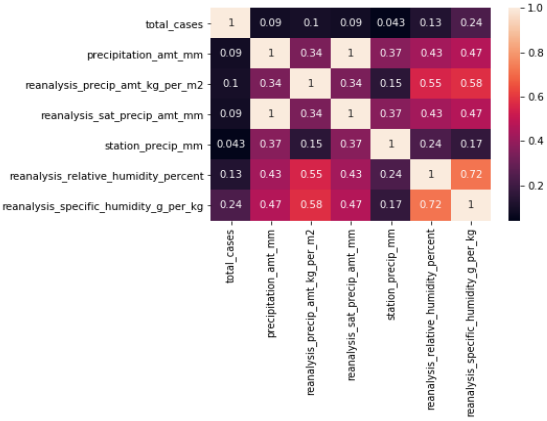
Figure 5 (d) - Total precipitation over years and total cases in Iquitos

Looking at the correlation between the precipitation features and total cases in both cities, there was a very low correlation between the four different precipitation features and total cases. The correlation heatmap confirms what we have observed from the previous plots which is precipitation and total cases are not highly correlated.

Correlation heat map in San Juan



Correlation heatmap in Iquitos



Attribute #2: Temperature[6]

Dengue fever is considered a tropical/subtropical disease because mosquitos live in greater abundance and are more active in hotter climates. The Dengue-carrying mosquitoes tend to feed at sunrise and sunset and have been observed to bite more and reproduce more frequently at higher temperatures. Due to this correlation, we expect higher temperatures in the data to be associated with more recorded total cases of Dengue fever. [7]

[6] Refer to Notebook 3 and 4 for the detailed analysis and code.
[7] Dengue and severe Dengue. By World Health Organization. Retrieved URL from https://www.who.int/en/news-room/fact-sheets/detail/dengue-and-severe-dengue

There were 9 variables that meausre temperature. `station_max_temp_c` and other temperature related variables that begin with `station`, are derived from the National Oceanic and Atmospheric Administration (NOAA) Global Historical Climatology Network (GHCN). This data is recorded from land surface stations across the globe and goes through extensive quality assurance reviews as it is commonly used in other research capacities. The remaining five temperature variables, named `reanalysis`, are from NOAA National Centers for Environmental Protection (NCEP) Climate Forecast System Reanalysis (CFSR) where reanalysis is the systemic approach to collecting meteorological data over time. These variables are measured at about 1.2 metres from the ground and are more influenced by attitude, surface type, coastal or interior, elevation, and atmospheric or oceanic circulations than surface temperature.

Surface temperatures like `station` are more extreme because surfaces absorb energy (sunlight) during the day and warms up, and cools when it loses energy at night. Think about walking on a scalding pavement on a hot day where the ground is so hot you can cook an egg. In our dataset, `station` surface temperature variables have greater fluctuations and on average both hotter and cooler than the `reanalysis` air temperature. Both types of temperatures were investigated as surface and air temperatures are equally viable influences on Dengue mosquito activity.

Figure 6 (a) Correlation Heatmap shows that there is a low correlation between total cases of Dengue fever and the temperature variables with the largest correlation coefficient from station_avg_temp_c at 0.196563 for San Juan. *Figure 6 (b-1) and (b-2)* Rolling Averages of Total Dengue Cases and Air Temperature over time doesn't show anything insightful. The peak in total cases in 1994 to 1996 appears to react nearly independently to air temperature.

Figure 6 (c) Heatmap of Total Dengue Cases against average air temperature and average surface temperature shows that there is a clear trend where there are more cases when temperatures are on average 26°C or greater in San Juan. The relationship is less obvious in Iquitos.

Figure 6 (d-1) and (d-2) Regression Plot between Total Dengue Cases and average surface temperature shows the impact of extreme total cases as temperature increases. However, the vast majority of total cases are still under 150, or even 100 across the whole spectrum of recorded temperatures. *Figure 6 (e)* and *(f)* show station temperature statistics San Juan and Iquitos against a timeline for the months of the year with all the outputs accumulated over the entire dataset duration.

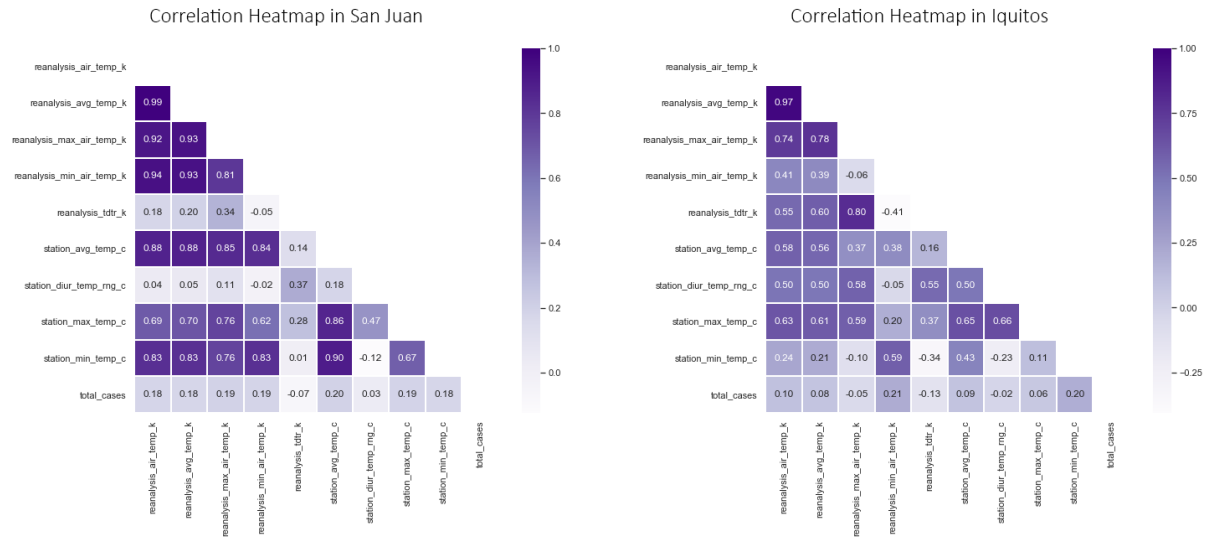


Figure 6 (a) - Correlation Heatmap between Total Cases and all Temperature variables

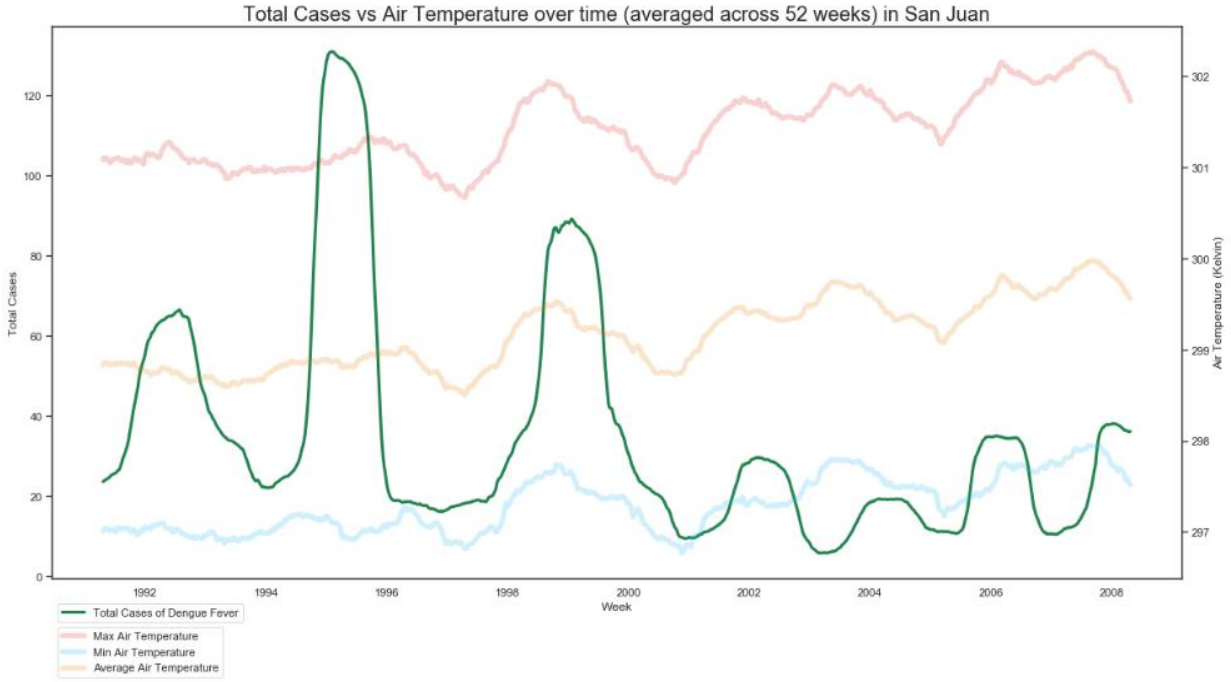


Figure 6 (b-1) - Rolling Average of Total Cases over time

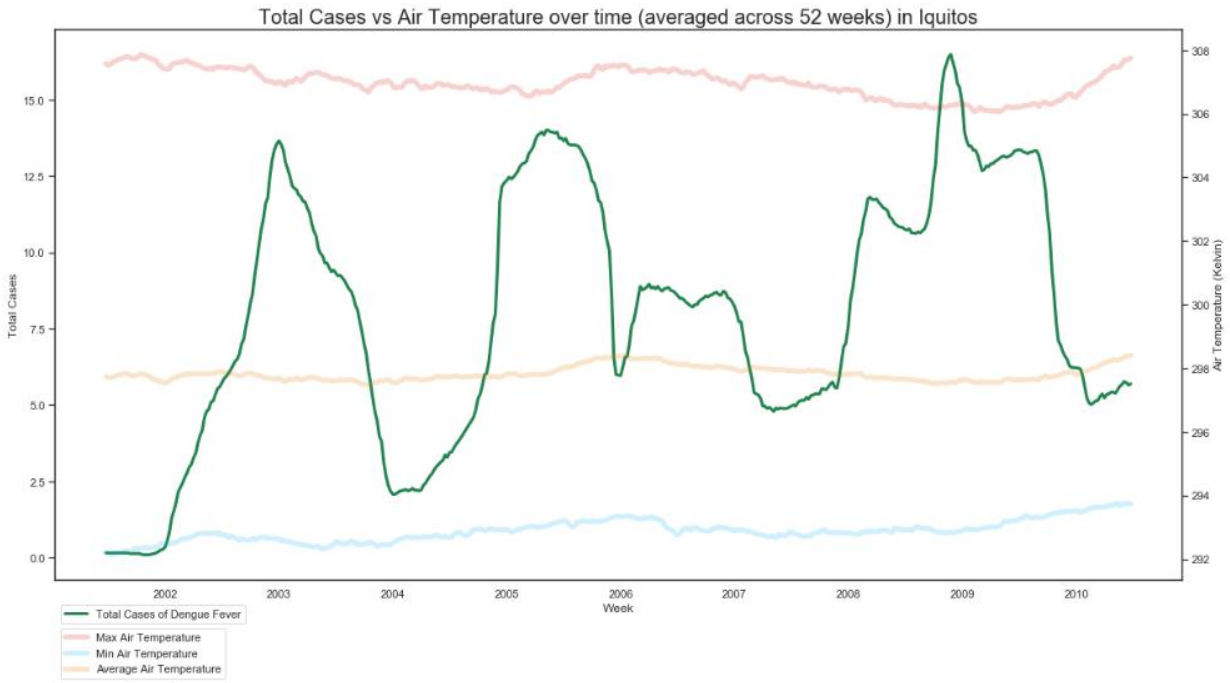


Figure 6 (b-2) - Rolling Average of Total Cases over time

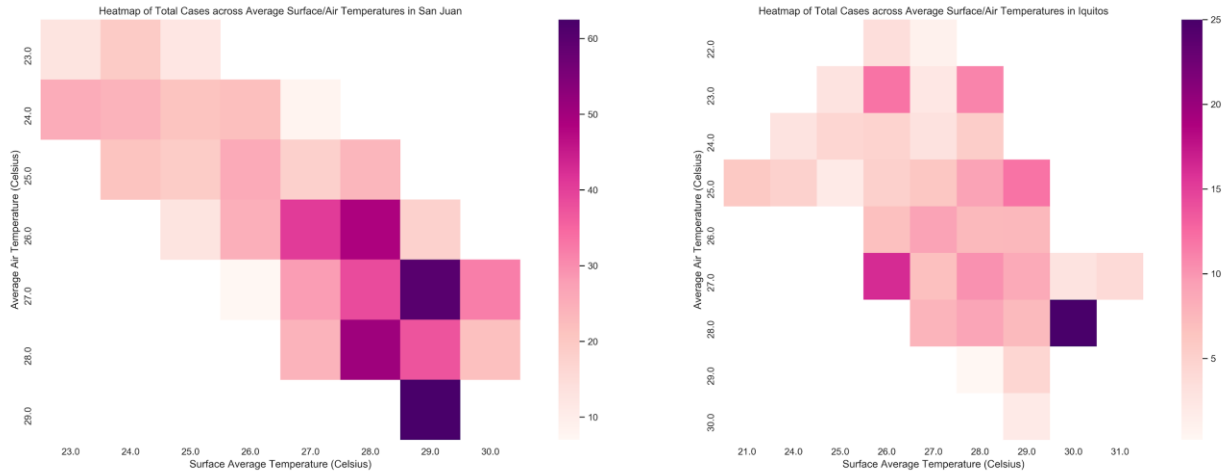


Figure 6 (c) - Heatmap of Total Cases between Average Surface and Air Temperature

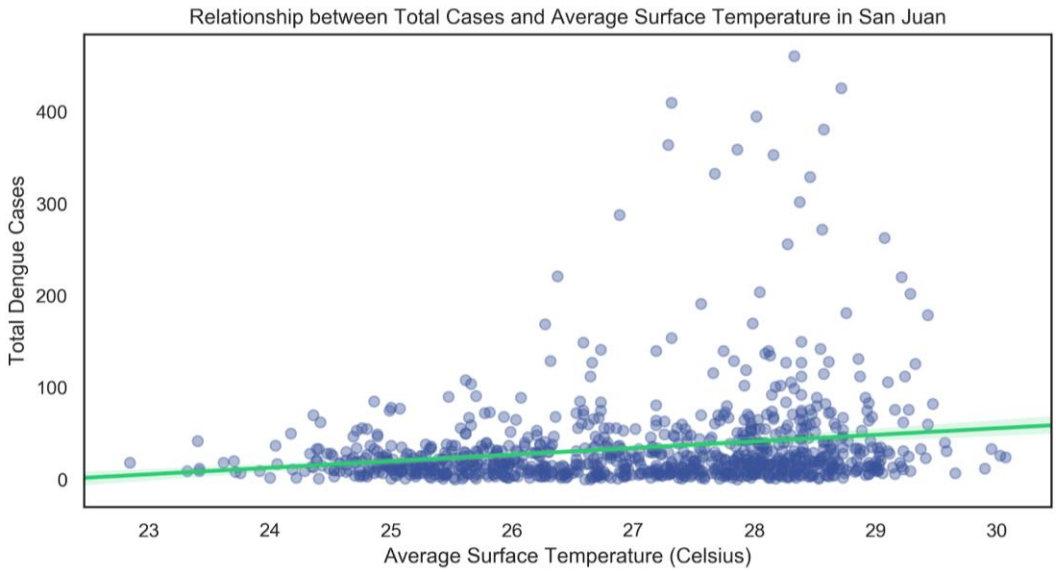


Figure 6 (d-1) - Regression Plot between Total Cases and Average Surface Temperature

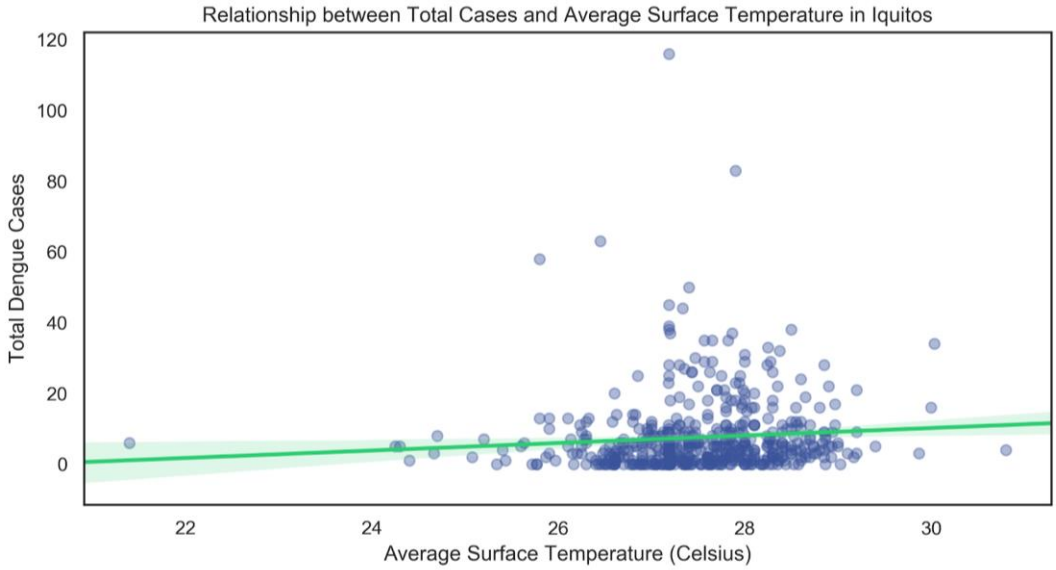


Figure 6 (d-2) - Regression Plot between Total Cases and Average Surface Temperature

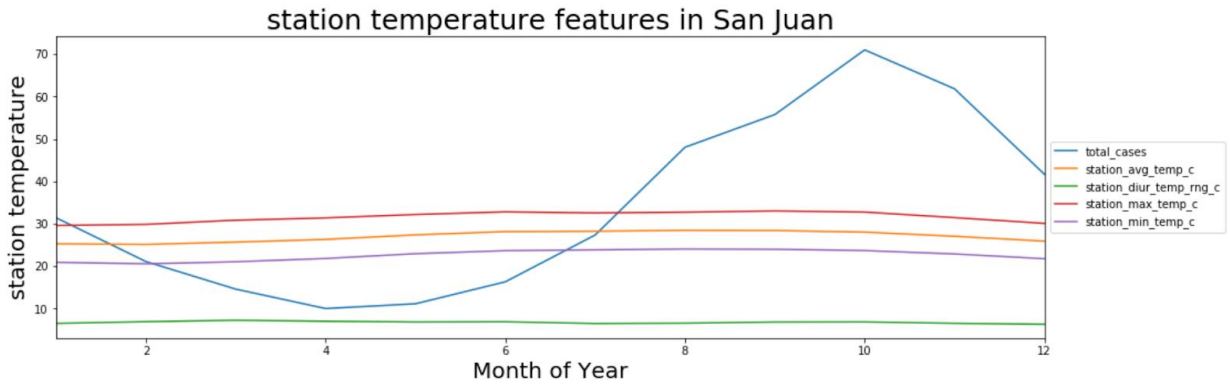


Figure 6 (e) - Mean of the station San Juan temperature values for the station attributes across the year related with Total cases

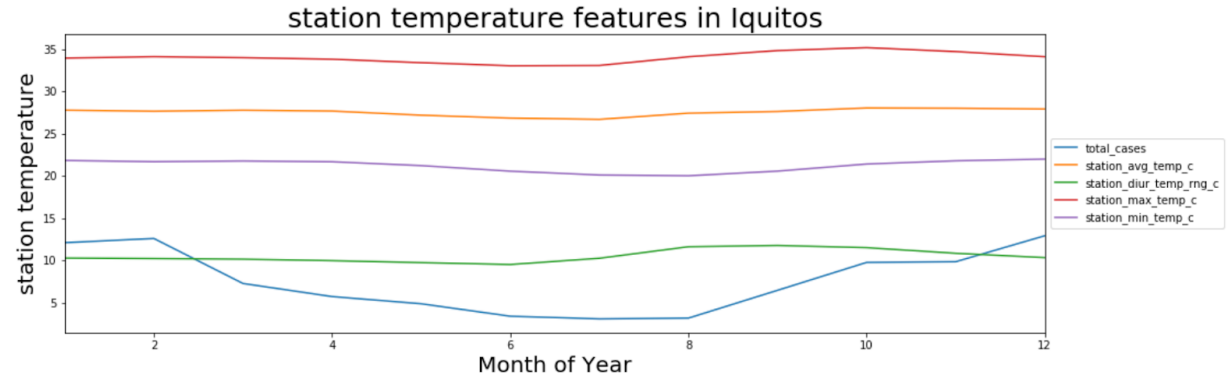


Figure 6 (f) - Mean of the station Iquitos temperature values for the station attributes across the year related with Total cases

Attribute #3: Humidity[8]

Shown in *Figure 7* below, humidity levels in Iquitos are generally higher than San Juan.

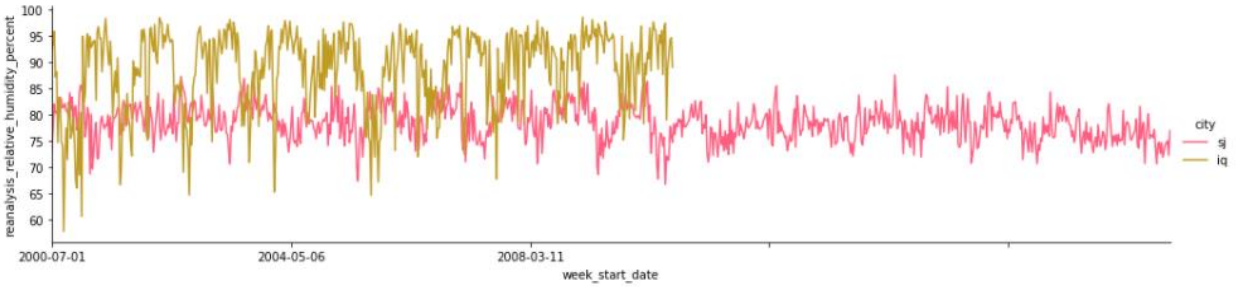


Figure 7 - Humidity % and Total Cases Over Time by City

[8] Refer to Notebook 5 for the detailed analysis and code.

Humidity levels in San Juan are quite seasonal. Total cases, as indicated by the blue line in *Figure 8 (a)* tend to spike when humidity reaches high levels. In Iquitos (*Figure 8(b)*), there isn't a clear relationship between total cases and humidity levels but the peak in Dengue fever cases in Iquitos does occur when humidity levels were at its highest.

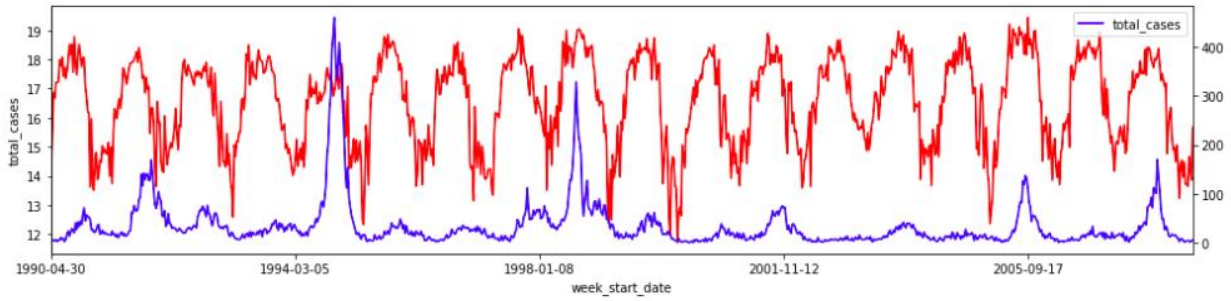


Figure 8 (a) - Humidity g per kg and Total Cases Over Time in San Juan

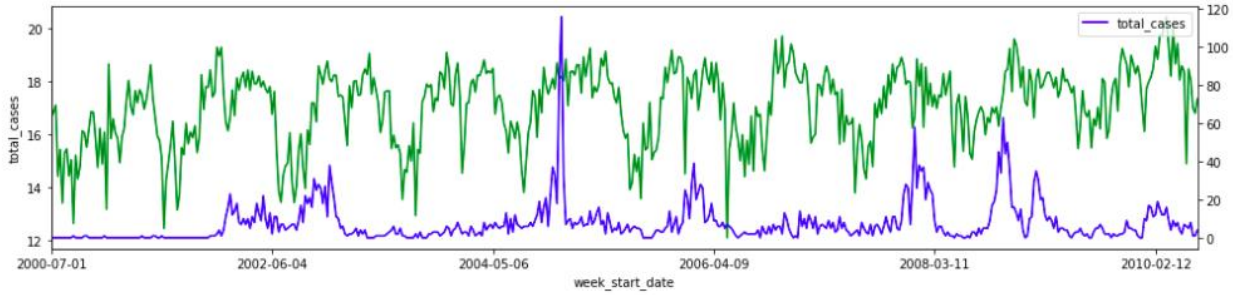


Figure 8 (b) - Humidity g per kg and Total Cases Over Time in Iquitos

Attribute #4: Vegetation[9]

The vegetation data that was provided in the dataset was the NDVI index. NDVI stands for "Normalized Difference Vegetation Index".[10]

The value of the NDVI index ranges from -1 to 1. The index value indicates the type of vegetation in the given area. These are the main types of vegetation:

a. Water - Negative values approaching -1
b. Barren areas of rock, sand, or snow - Values close to zero (-0.1 to 0.1)
c. Shrub/Grassland - Low, positive values (approximately 0.2 to 0.4.
d. Tropical Rainforest - High values (values approaching 1).

[9] Refer to Notebook 5 for the detailed analysis and code.
[10] Sentinel Hub by Sinergise. NDVI (Normalized Difference Vegetation Index). Retrieved from URL: https://www.sentinel-hub.com/eoproducts/ndvi-normalized-difference-vegetation-index

Referring to *Figure 9 (a)* to *Figure 9 (d)*, in San Juan, the North East and North West areas with the greatest number of cases occurring with the land are more barren. In the South East and South West areas of San Juan, Dengue fever cases occur when the area is more Grassland. In Iquitos, areas that are more Grassland results in the highest number of Dengue cases.

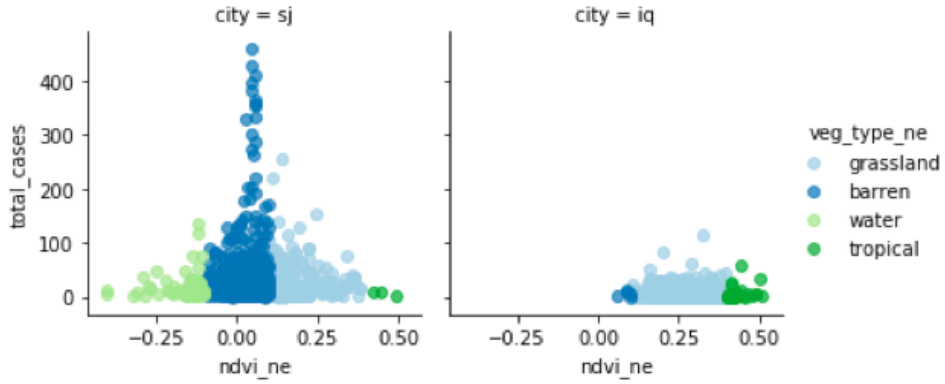


Figure 9 (a) - Vegetation Types in NE Region

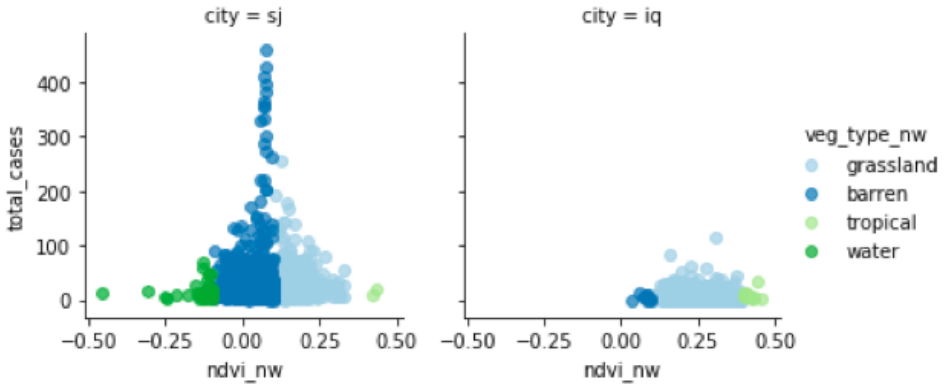


Figure 9 (b) - Vegetation Types in NW Region

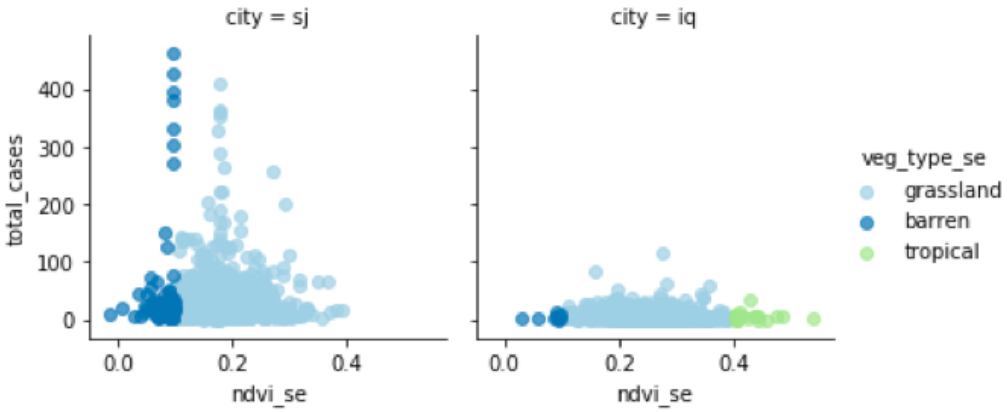


Figure 9 (c) - Vegetation Types in SE Region

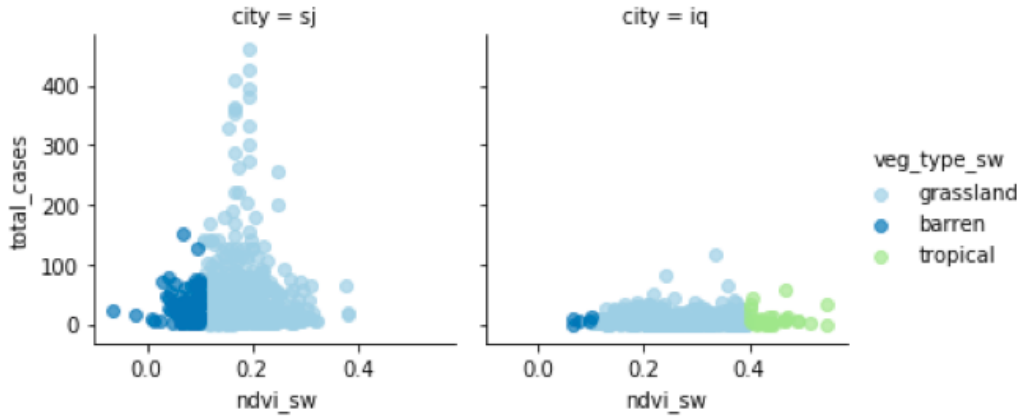


Figure 9 (d) - Vegetation Types in SE Region

IV. Conclusions

Based on the above analysis, these are the overall conclusions on what environmental factors contribute to a higher number of Dengue cases in San Juan and Iquitos:

- **Hot and Heavy**
  When the environment is humid and hot, this lends to a higher number of Dengue fever cases. Temperature is similar between San Juan and Iquitos, but the overall temperature is consistently high throughout the year (the mean for both cities was 27C). Further, as minimum temperatures, maximum temperatures, and average temperatures rise, the cases of Dengue fever tend to rise as well.

  The correlation strengths differ for each city, but overall humidity was most strongly correlated with the number of Dengue fever cases. This is accurate as mosquitos thrive in wet climates.

- **Precipitation is correlated with Humidity**
  There was no direct correlation between Precipitation and Dengue Cases but total cases has a stronger correlation to humidity.

V. Appendix

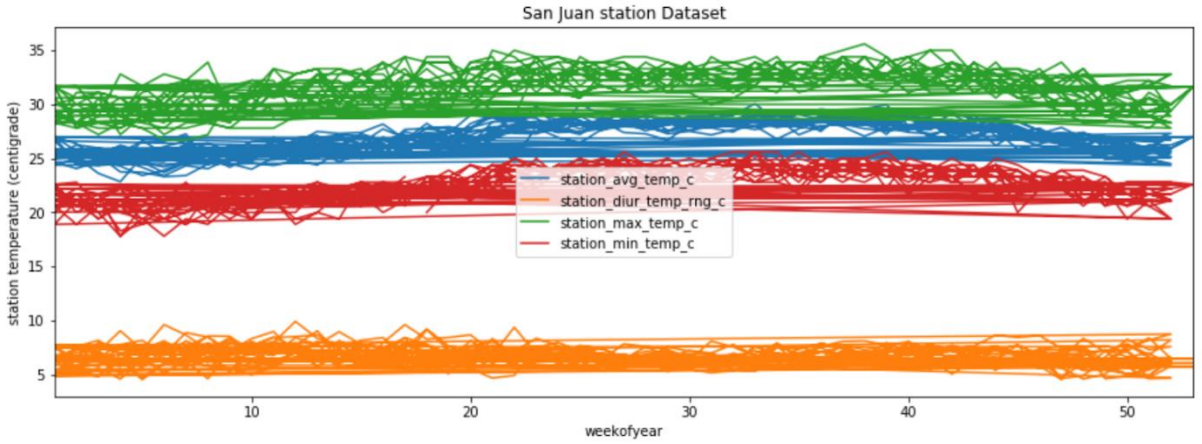
Appending A – Summary of Fill Approaches for NaN Values

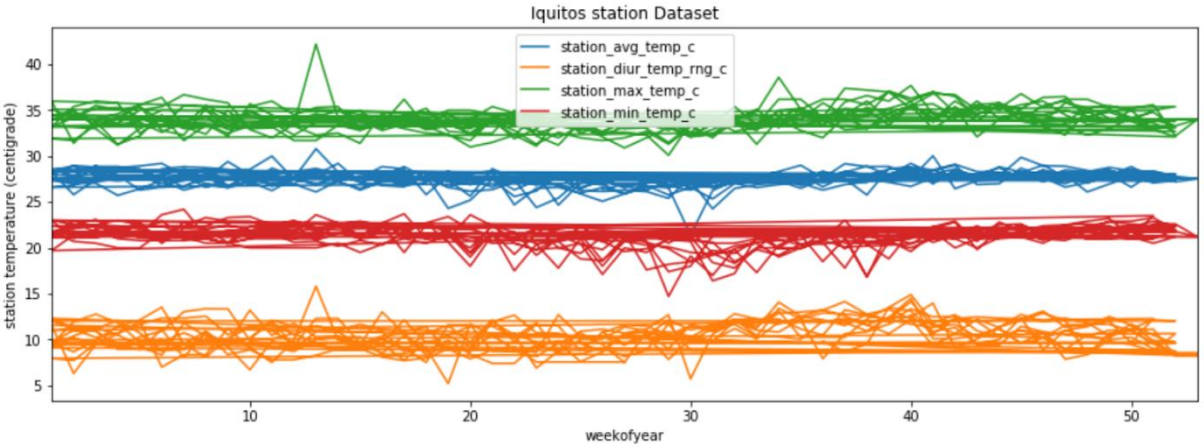
				Fill Approach	
#	Column Names	Type of Environmental Factor	# of NaN Values	Forward-Fill (ffil)	Mean
1	ndvi_ne	Vegetation	194	V	
2	ndvi_nw	Vegetation	52	V	
3	ndvi_se	Vegetation	22	V	
4	ndvi_sw	Vegetation	22	V	
5	precipitation_amt_mm	Precipitation	13		V
6	reanalysis_air_temp_k	Temperature	10	V	V
7	reanalysis_avg_temp_k	Temperature	10		V
8	reanalysis_dew_point_temp_k	Temperature	10		V
9	reanalysis_max_air_temp_k	Temperature	10		V
10	reanalysis_min_air_temp_k	Temperature	10		V
11	reanalysis_precip_amt_kg_per_m2	Precipitation	10		V
12	reanalysis_relative_humidity_percent	Humidity	10		V
13	reanalysis_sat_precip_amt_mm	Precipitation	13		V
14	reanalysis_specific_humidity_g_per_kg	Humidity	10		V
15	reanalysis_tdtr_k	Temperature	10	V	
16	station_avg_temp_c	Temperature	43		V
17	station_diur_temp_rng_c	Temperature	43		V
18	station_max_temp_c	Temperature	20		V
19	station_min_temp_c	Temperature	14		V
20	station_precip_mm	Precipitation	22		V

- **Vegetation** in a given city does not change drastically week-over-week (i.e. changes in vegetation should be gradual overtime). As shown in Appendix C, the NDVI values in the four regions of each city appear relatively stationary as no

clear increasing or decreasing trends persist overtime. Using the last observed value (forward fill) before the NaN value would be appropriate.

- The average value for the given city was used to fill in Temperature, Precipitation and Humidity.
    - **Temperature**: Referring to the figure below; the temperatures throughout the year generally close to the mean and no disparity throughout the weeks of the year on temperature variations. Therefore, using the mean to fill in the NaN values is appropriate.





    - **Precipitation and Humidity:** For simplicity the mean for the respective cities was used to fill in the NaN values.

Appendix B – Mapping NDVI Values to Vegetation Type

Original Columns	Added Columns
ndvi_ne	veg_type_ne
ndvi_nw	veg_type_nw
ndvi_se	veg_type_se
ndvi_sw	veg_type_sw

Vegetation Type	Logic Rules
Water	x<-0.1
Barren	x<=0.1 and x>=-0.1
Grassland	x<=0.4 and x>0.1
Tropical	x<=1.0 and x>0.4
Unknown	Everything else

```
df_merge['veg_type_ne']=df_merge['ndvi_ne'].apply(lambda x:'water' if x<-0.1
                                                  else 'barren' if (x<=0.1 and x>=-0.1)
                                                  else 'grassland' if (x<=0.4 and x>0.1)
                                                  else 'tropical' if (x<=1.0 and x>0.4)
                                                  else 'unknown')
df_merge['veg_type_nw']=df_merge['ndvi_nw'].apply(lambda x:'water' if x<-0.1
                                                  else 'barren' if (x<=0.1 and x>=-0.1)
                                                  else 'grassland' if (x<=0.4 and x>0.1)
                                                  else 'tropical' if (x<=1.0 and x>0.4)
                                                  else 'unknown')
df_merge['veg_type_se']=df_merge['ndvi_se'].apply(lambda x:'water' if x<-0.1
                                                  else 'barren' if (x<=0.1 and x>=-0.1)
                                                  else 'grassland' if (x<=0.4 and x>0.1)
                                                  else 'tropical' if (x<=1.0 and x>0.4)
                                                  else 'unknown')
df_merge['veg_type_sw']=df_merge['ndvi_sw'].apply(lambda x:'water' if x<-0.1
                                                  else 'barren' if (x<=0.1 and x>=-0.1)
                                                  else 'grassland' if (x<=0.4 and x>0.1)
                                                  else 'tropical' if (x<=1.0 and x>0.4)
                                                  else 'unknown')
```

Appendix C – Vegetation Plots: NDVI for Each City Overtime

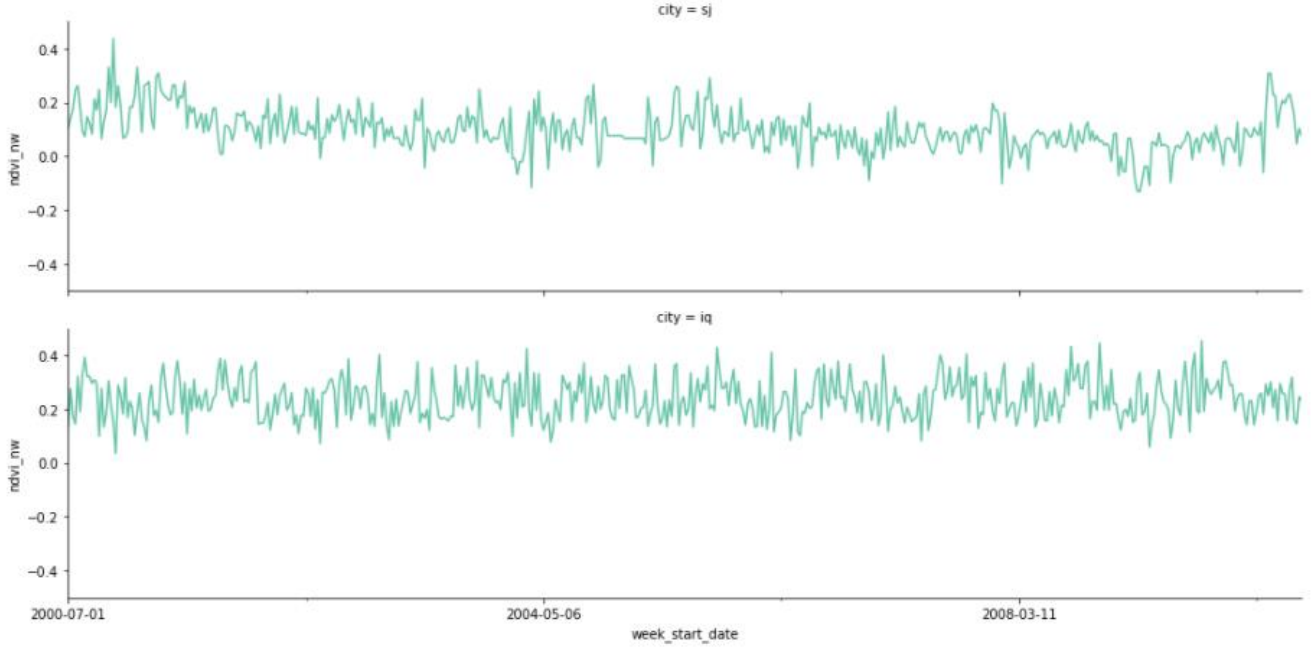


Figure A - NDVI for the North West (NW) Area Overtime

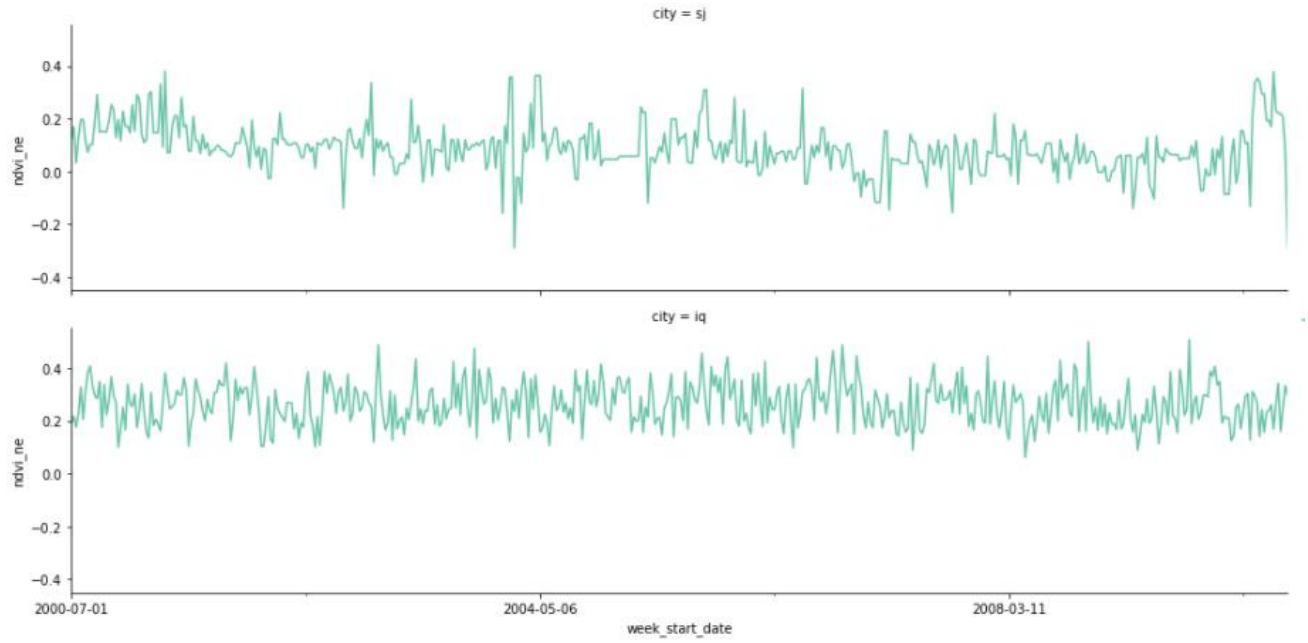


Figure B - NDVI for the North East (NE) Area Overtime

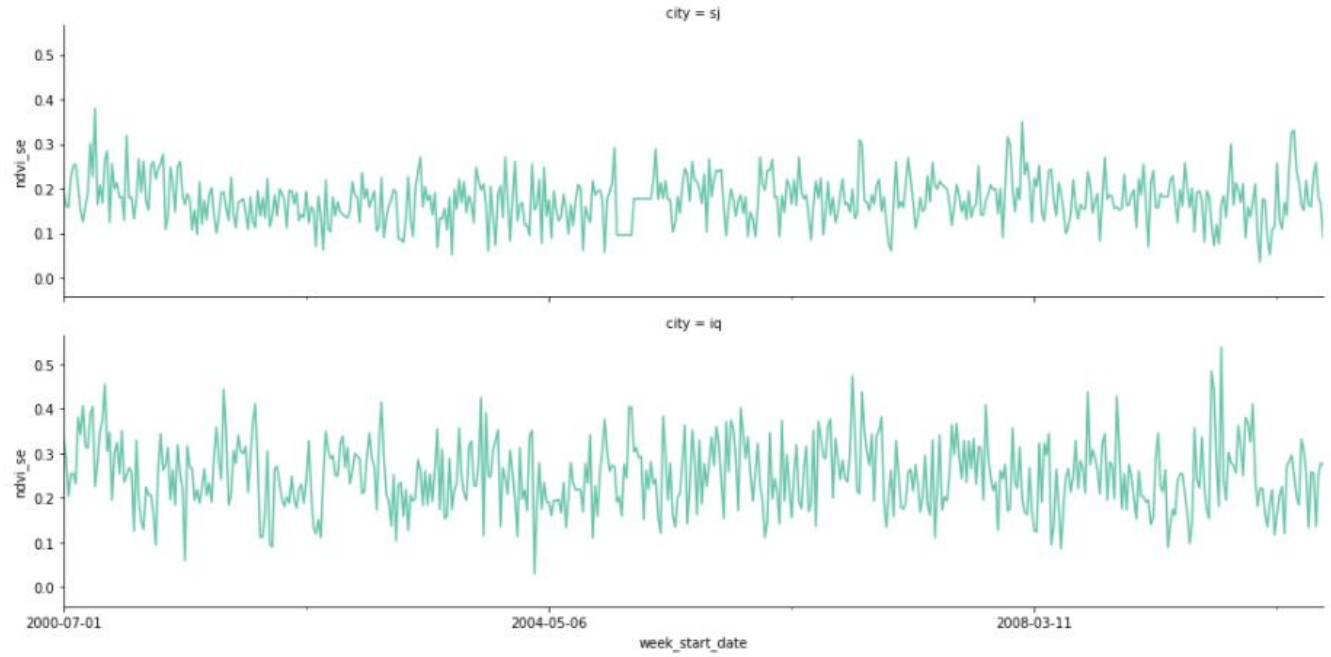


Figure C - NDVI for the South East (SE) Area Overtime

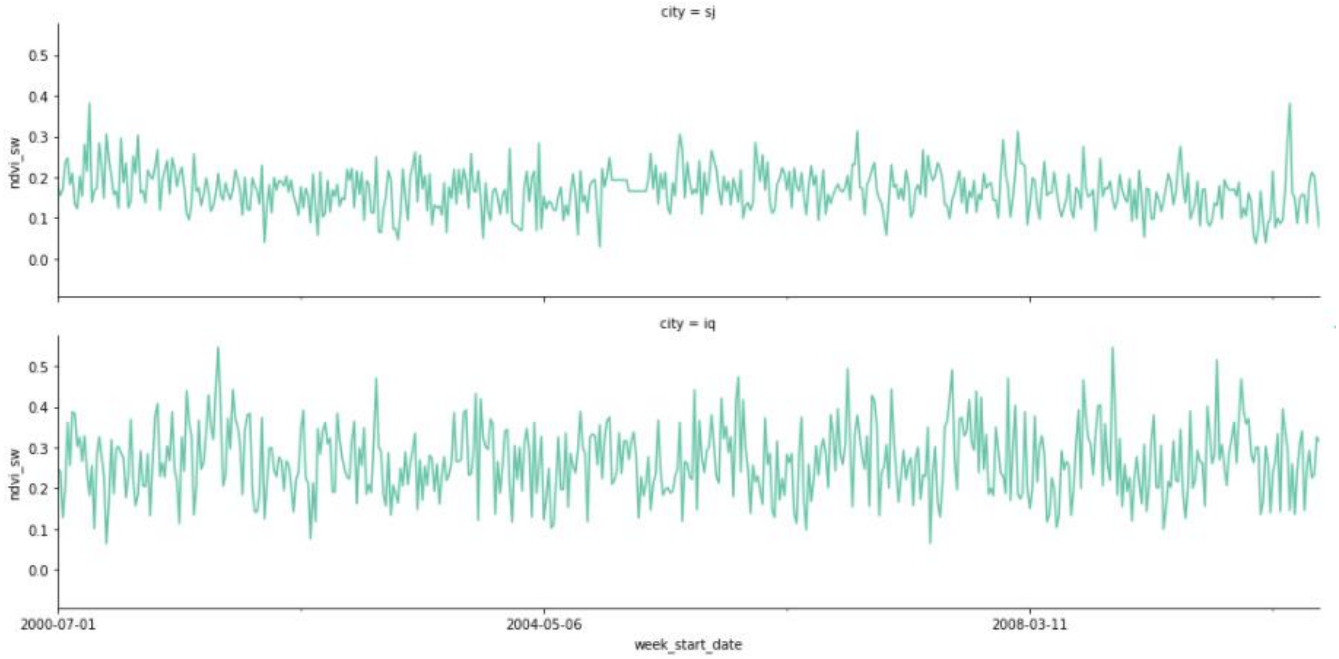


Figure D - NDVI for the South West (SW) Area Overtime