|  |  |
| --- | --- |
|  | 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** 2](#_Toc37688631)

[**II.** **Data Preparation** 2](#_Toc37688632)

[**a.** **Data Source** 2](#_Toc37688633)

[**b.** **Data Quality** 2](#_Toc37688634)

[**c.** **Challenges** 3](#_Toc37688635)

[**d.** **Tools and Code Required for Analysis** 3](#_Toc37688636)

[**III.** **Analysis** 5](#_Toc37688637)

[**Overall Trends, Correlations and/or Patterns** 5](#_Toc37688638)

[**Attribute #1: Precipitation** 10](#_Toc37688639)

[**Attribute #2: Temperature** 12](#_Toc37688640)

[**Attribute #3: Humidity** 16](#_Toc37688641)

[**Attribute #4: Vegetation** 17](#_Toc37688642)

[**IV.** **Conclusions** 19](#_Toc37688643)

[**V.** **Appendix** 20](#_Toc37688644)

[**Appending A – Summary of Fill Approaches for NaN Values** 20](#_Toc37688645)

[**Appendix B – Mapping NDVI Values to Vegetation Type** 22](#_Toc37688646)

[**Appendix C – Vegetation Plots: NDVI for Each City Overtime** 23](#_Toc37688647)

# **Introduction**

Dengue fever is a mosquito-borne disease that currently 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. These were the main environmental data points:

* 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 high number of Dengue Fever cases in San Juan and Iquitos?**

The analysis explores:

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

# **Data Preparation**

### **a.** **Data Source**

The data source was from open data on the DrivenData Competition website:

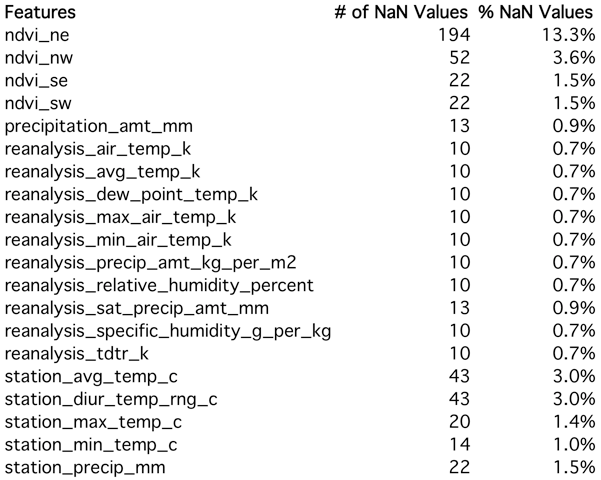
<https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/page/82/>

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 was overall 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.



### **c.** **Challenges**

There were 3 main challenges encountered upon examining the dataset:

1. Understanding and interpreting what environmental factors were;
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 other variables

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 then dropped from the analysis. This is further discussed in section “**III. Analysis**” of this report.

### **d.** **Tools and Code Required for Analysis**

#### **Joining Labels dataset to the Features dataset**

The “total\_cases” data field was located in the labels data set “dengue\_labels\_train.csv” file. A new DataFrame was created to join the “total\_cases'' field to the features dataset “dengue\_features\_train.csv”. The join was executed on a unique key that combined 'city','year' and 'weekofyear'. Refer to Notebook 1 section “2.a. Understanding the Data'' for the code.

#### **Filing in NaN values**

Once the features and labels dataset Referring to Notebook 1 section “2.b. Handling Missing **Data**”, since the environmental features would be specific to the location, filling methods were applied given the City as this would prevent imputing one city's data into another. To do this, the dataset was split into two different DataFrames based on City. Once the dataset was split into the respective Cities, 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 table outlining the fill approach.

#### **Converting Date into DateObject for Time Series**

The data field “week\_start\_date” was initially an “object” data type. To use this field in any time series analysis, the data type needed to be converted to a DateObject data type. Refer to Notebook 1 section “2.c.1 Transforming Data” for the code.

#### **Created a New Data Fields:**

#### **Mapping NDVI Values to Vegetation Type**

Further discussed in section “III. Analysis - Attribute #4” of this report, the NDVI values ('ndvi\_ne', 'ndvi\_nw', 'ndvi\_se' and 'ndvi\_sw') were mapped to the type of vegetation. 4 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 an extract 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 |

#### **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:



# **Analysis**

### **Overall Trends, Correlations and/or Patterns**

San Juan has more Dengue fever cases compared to Iquitos. Referring to Figure 1(a), in total San Juan has more cases than Iquitos. But as shown in Figure 1(b) the higher number of cases in San Juan was partially because Dengue cases in San Juan started in 1990. This is 10 years before Iquitos started having Dengue fever cases. This observation is also shown in Figure 1(c).

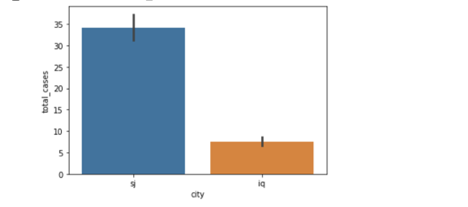


Figure 1 (a) Bar Chart of Total Cases by City

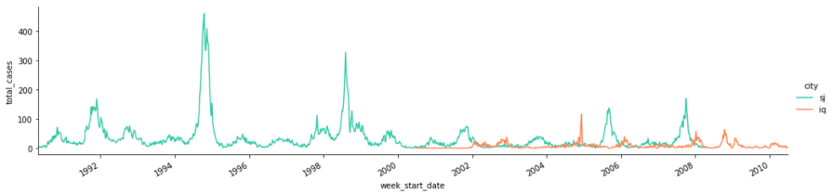
****

Figure 2 (b) Total Cases by City Over Time

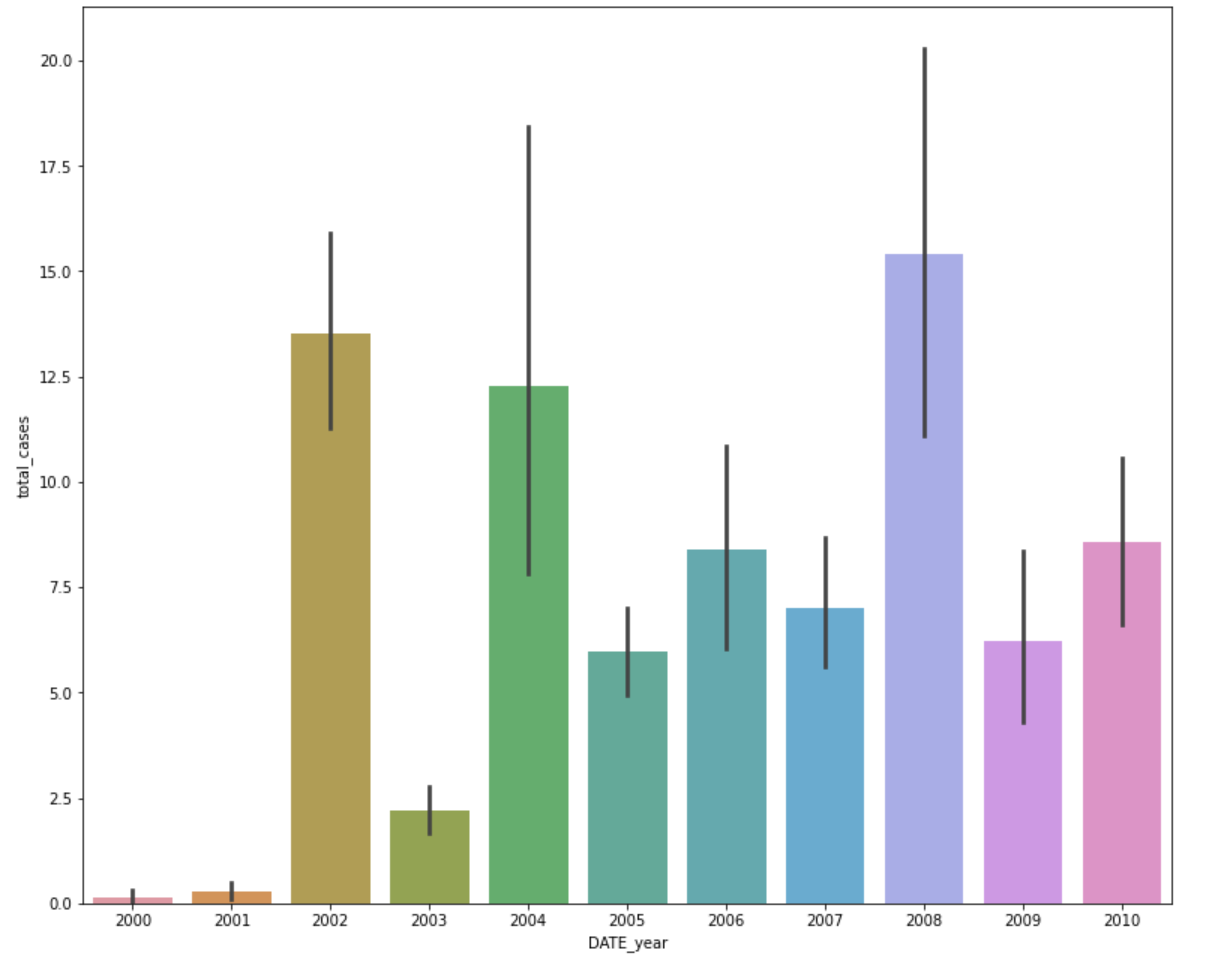
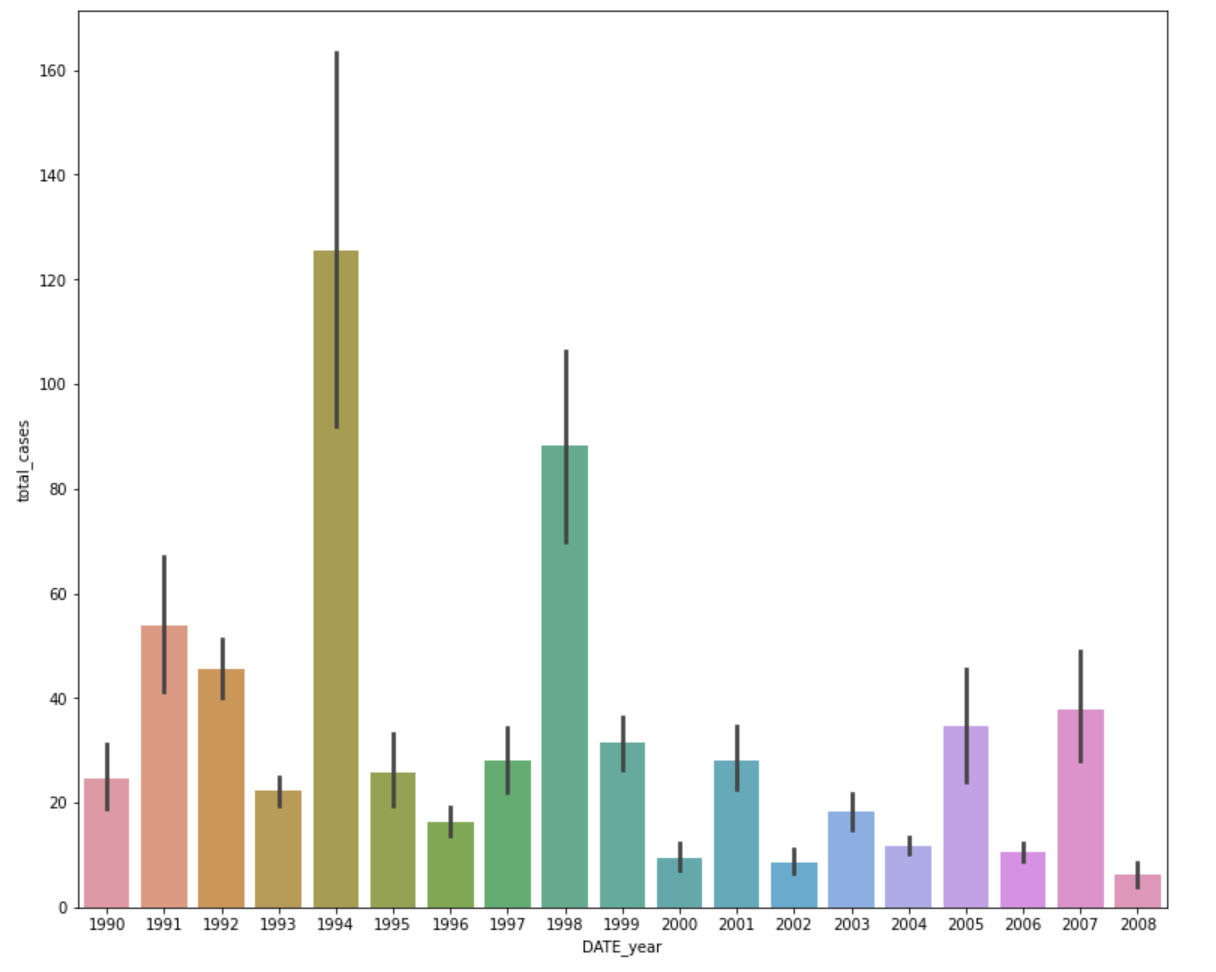


Figure 3 (c) Total Cases by in San Juan (Left) and Iquitos (Right) Over Time

Also shown in the left plot of Figure 1(c), the number of Dengue fever cases in San Juan peaked in 1994 and 1998. In comparison, in 2000, 2002, 2004 and 2008 Dengue cases were significantly lower. There were also no cases in 2009 and 2010 in San Juan.

In the city of Iquitos (right plot of Figure 1(c)), there were no cases of the disease until 2000. There was a sharp increase in cases in 2002.

The next 3 time series plots show how the total cases trend between the weeks of a given year for each city.

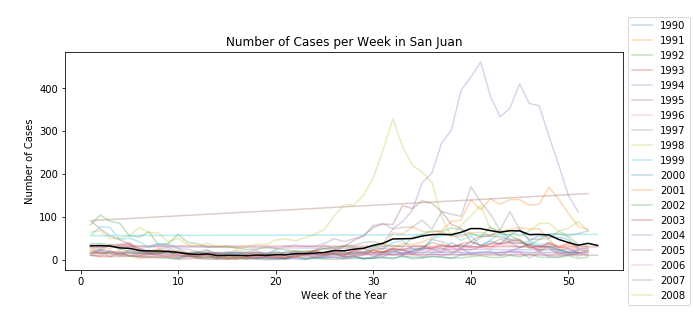


Figure 2(a) - Total Cases by Week of Year in San Juan

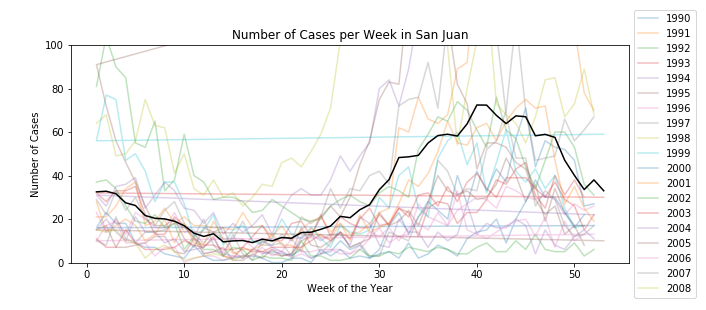


Figure 2(a-1) - Total Cases by Week of Year in San Juan

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

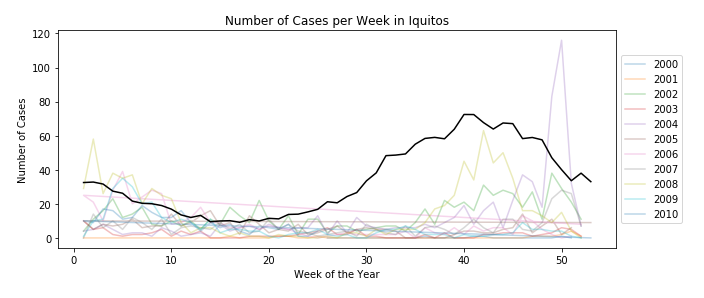


Figure 2(b) - Total Cases by Week of Year in Iquitos

Figure2(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 for the week of year. Figure 2(a-1) shows the total cases line trend better since the y-axis range was adjusted. 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 Figure2(b), Dengue fever cases in Iquitos started 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 after which the number of cases start increasing from April to October. Cases then decreased again from November to December. This trend is further highlighted in Figure 3(a) and Figure 3(b).

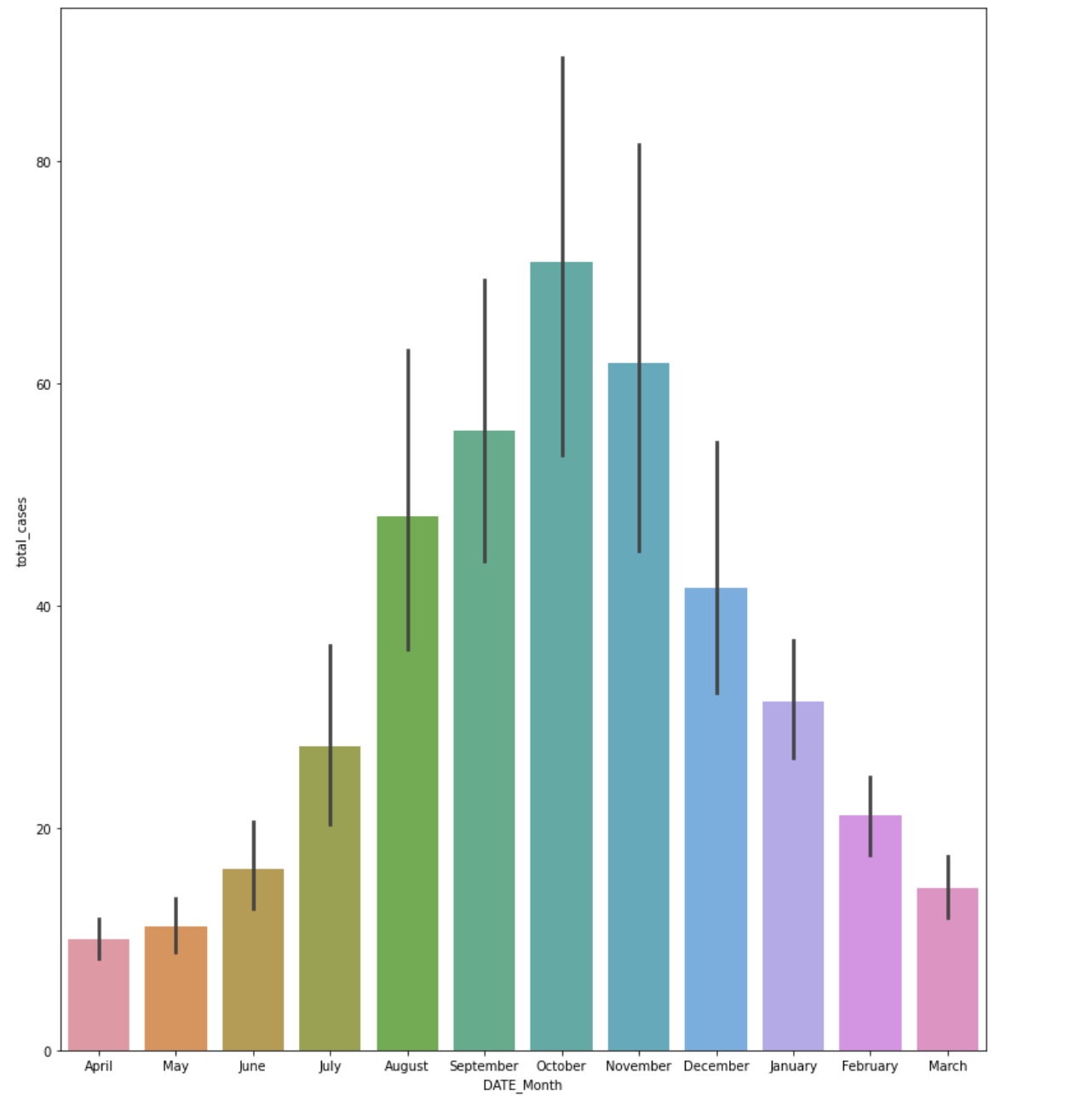


Figure 3(a) - Total Cases by Month of the Year in San Juan

Between the months of April to October (which corresponds with the wet summer season and hotter temperatures[[1]](#footnote-1)), the total cases of diseases are increasing followed by a decline in cases from November to March (which corresponds with the dry season and cooler temperatures[[2]](#footnote-2)). Precipitation, Temperature and Humidity are further explored in the upcoming Attribute sections of this report.

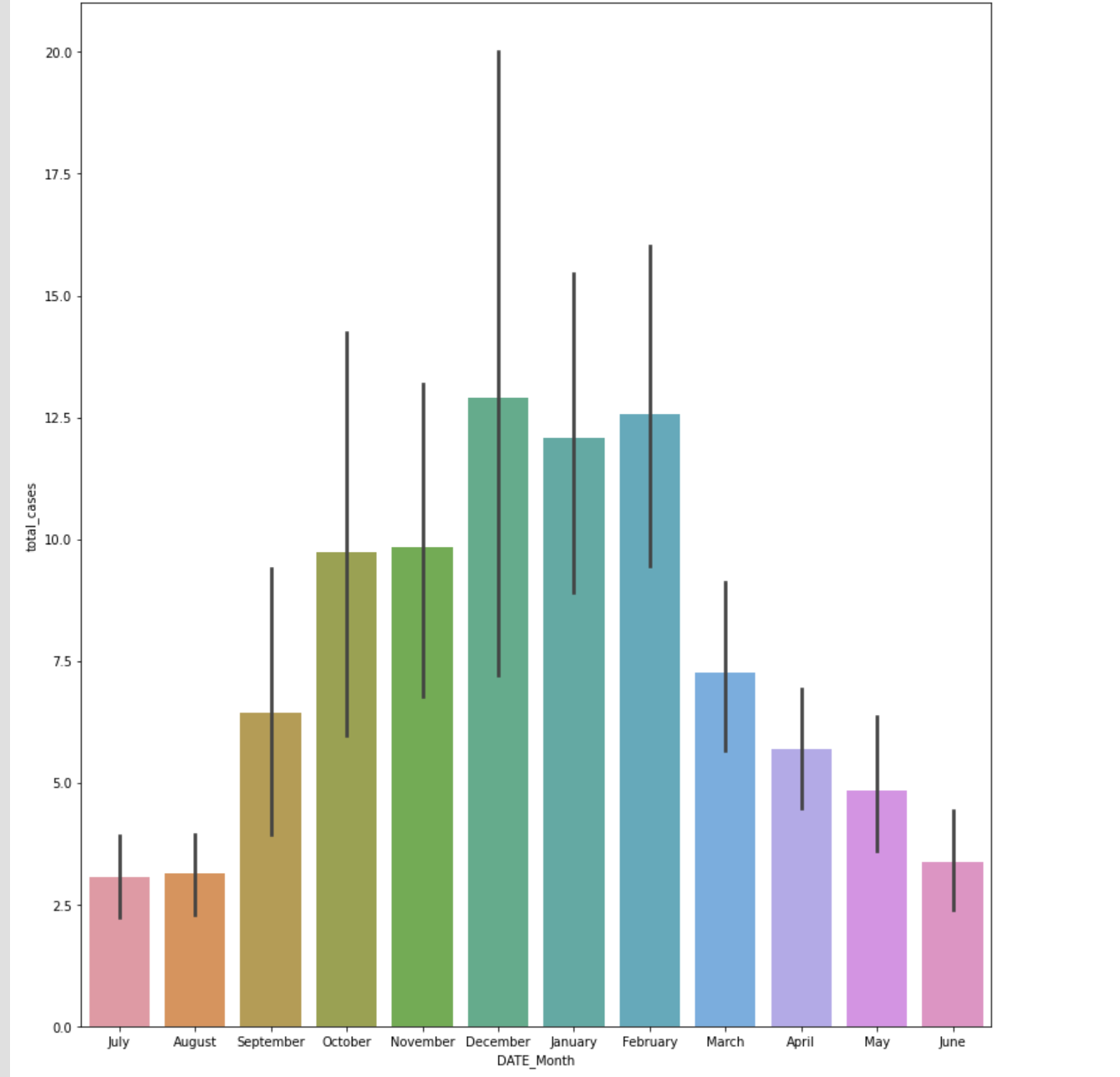


Figure 3(b) - Total Cases by Month of the Year in Iquitos

Between the months of September to February (which corresponds to the peak of the wet season[[3]](#footnote-3)), there is an increasing trend of Dengue cases followed by a decline in cases from March to August (which corresponds with the dry season which starts in June[[4]](#footnote-4)).

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

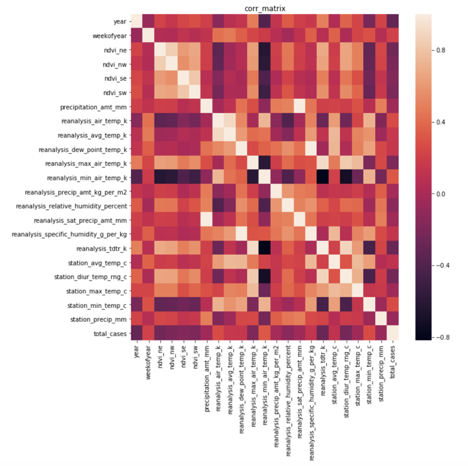
****

Figure 4 - Correlation Heat Map with All Environmental Factors

Many of the temperature measures are strongly correlated, which is expected, but overall the 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, precipitation variables bear little to no correlation to total\_cases, but there are strong correlations with 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]](#footnote-5)**

There are 4 features that represent precipitation in the dataset. According to the drivendata.org, all columns represent total precipitation amounts but in different units.

* Prcipitation\_amt\_mm
* Reanalysis\_precip\_amt\_kg\_per\_m2
* Reanalysis\_sat\_precip\_amt\_mm
* Station\_precip\_mm

1. **Total precipitation features over months and total cases**

Referring to Figure 5(a) and Figure 5(b), both graphs’ x-axis represents the month of year and y-axis represents the average of each months’ total precipitation features. The blue line indicates the total cases.

In San Juan, all 4 precipitation features are all relative and show similar trends between 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 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. Precipitation and the number of Dengue cases are strongly related based on this plot.

However, Iquitos has a completely opposite trend in precipitation features as well as the total cases. As discussed earlier, 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 all the way to 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.

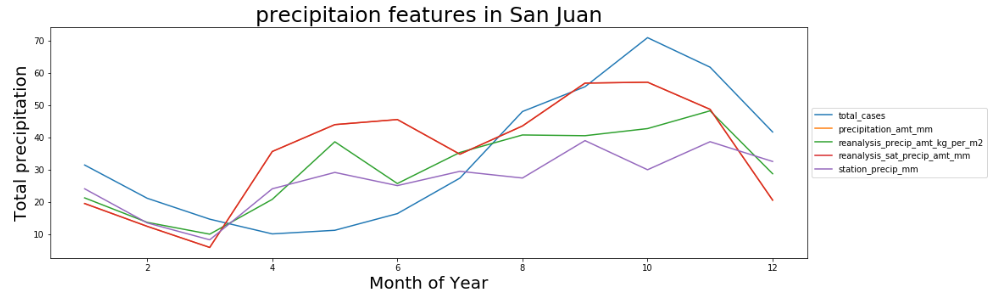


Figure 5(a) Total precipitation over months and total cases in San Juan

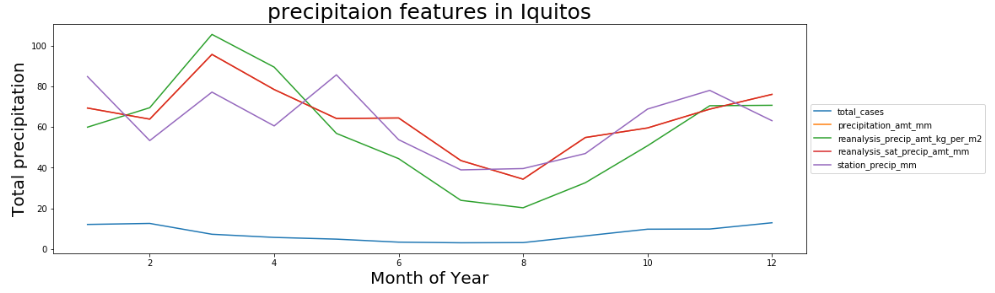


Figure 5(b) - Total precipitation over months and total cases in Iquitos

1. **Total precipitation features over years and total cases**

Referring to Figure 5(c) and Figure 5(d), both graphs’ x-axis represents the year and y-axis represents the average of each year’s total precipitation features. Also, 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 were also following those seasonal patterns. Even though precipitation features and total cases have the seasonality with similar patterns, 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 as well as the total cases year over year. The plots are relatively flat. Compared to San Juan, the average of total participation in Iquitos is higher throughout the year but there is a lower number of Dengue cases.

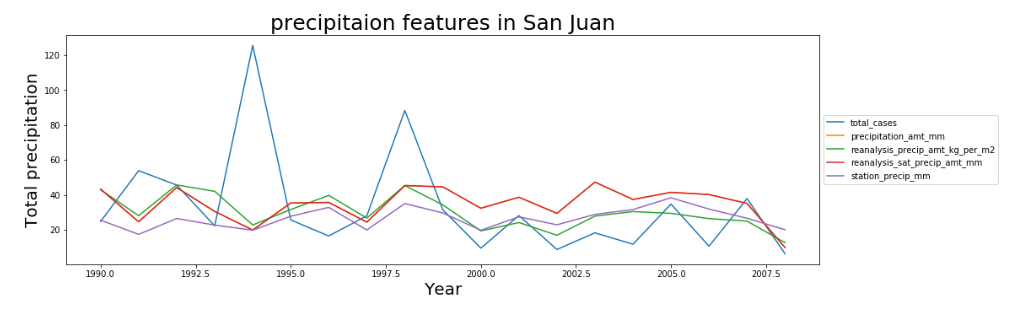


Figure 5(c) - Total precipitation over years and total cases in San Juan

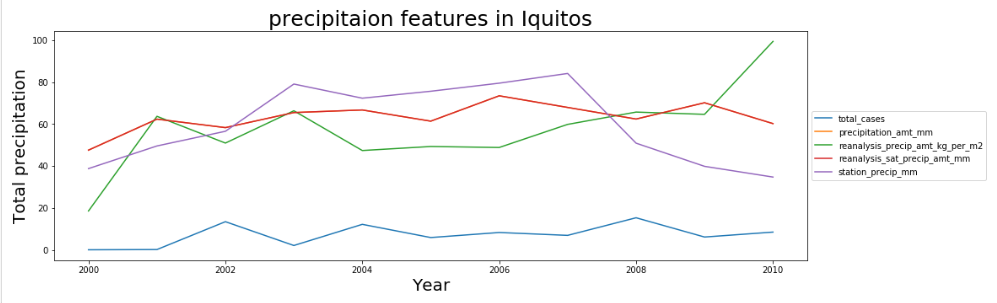
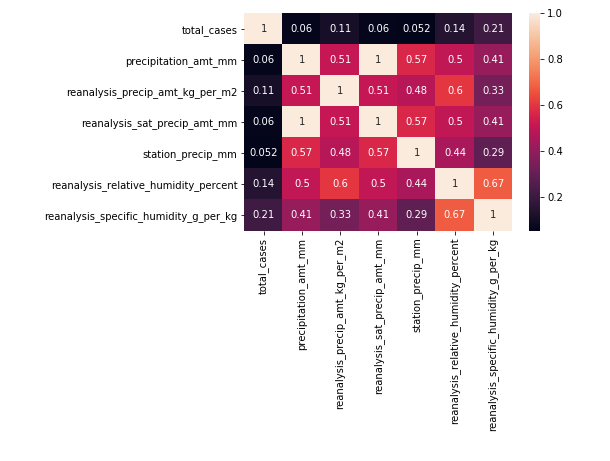


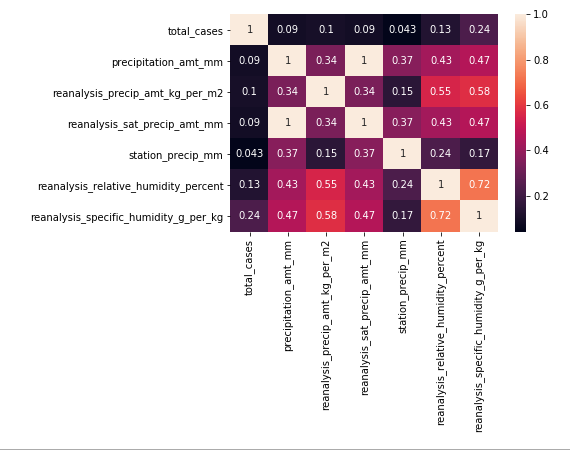
Figure 5(d) - Total precipitation over years and total cases in Iquitos

Looking at the correlation between the precipitation features and total cases, in both cities, there was a very low correlation between 4 different precipitation features and total cases. 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]](#footnote-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.

There were 9 variables that look at temperature. station\_max\_temp\_c and other variables that begin with station (that are temperature related (4)), 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 temperature variables, named reanalysis (5), 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 reanalysis air temperature. Both types of temperatures were investigated as surface and air temperatures are equally viable influences on Dengue mosquito activity.

Figure 2.1 Correlation Heatmap shows that there is 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.

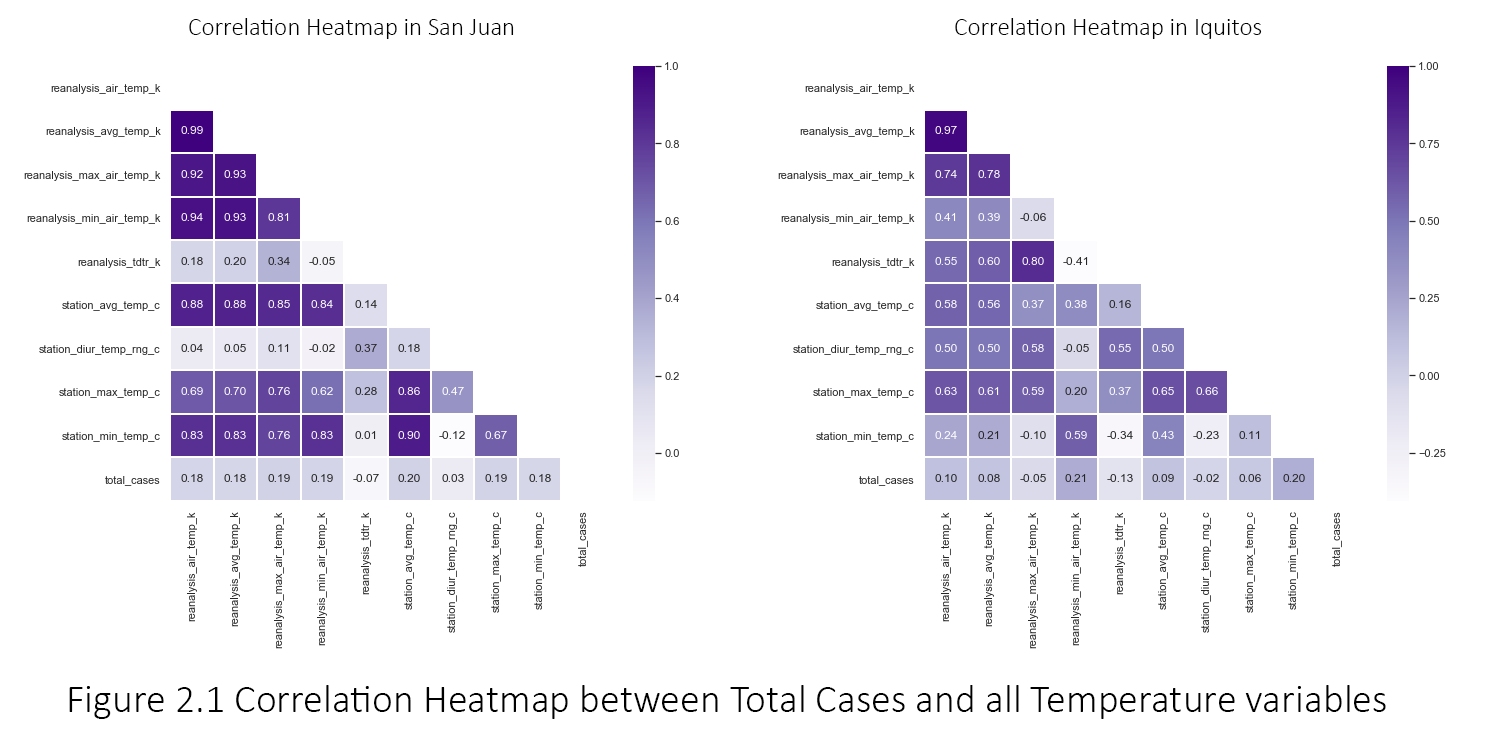
Figure 2.2 Rolling Averages of Total Dengue Cases and Air Temperature over time doesn’t show anything insightful. The peak in total cases in 1994 to 1996 appears to react nearly independently to air temperature.

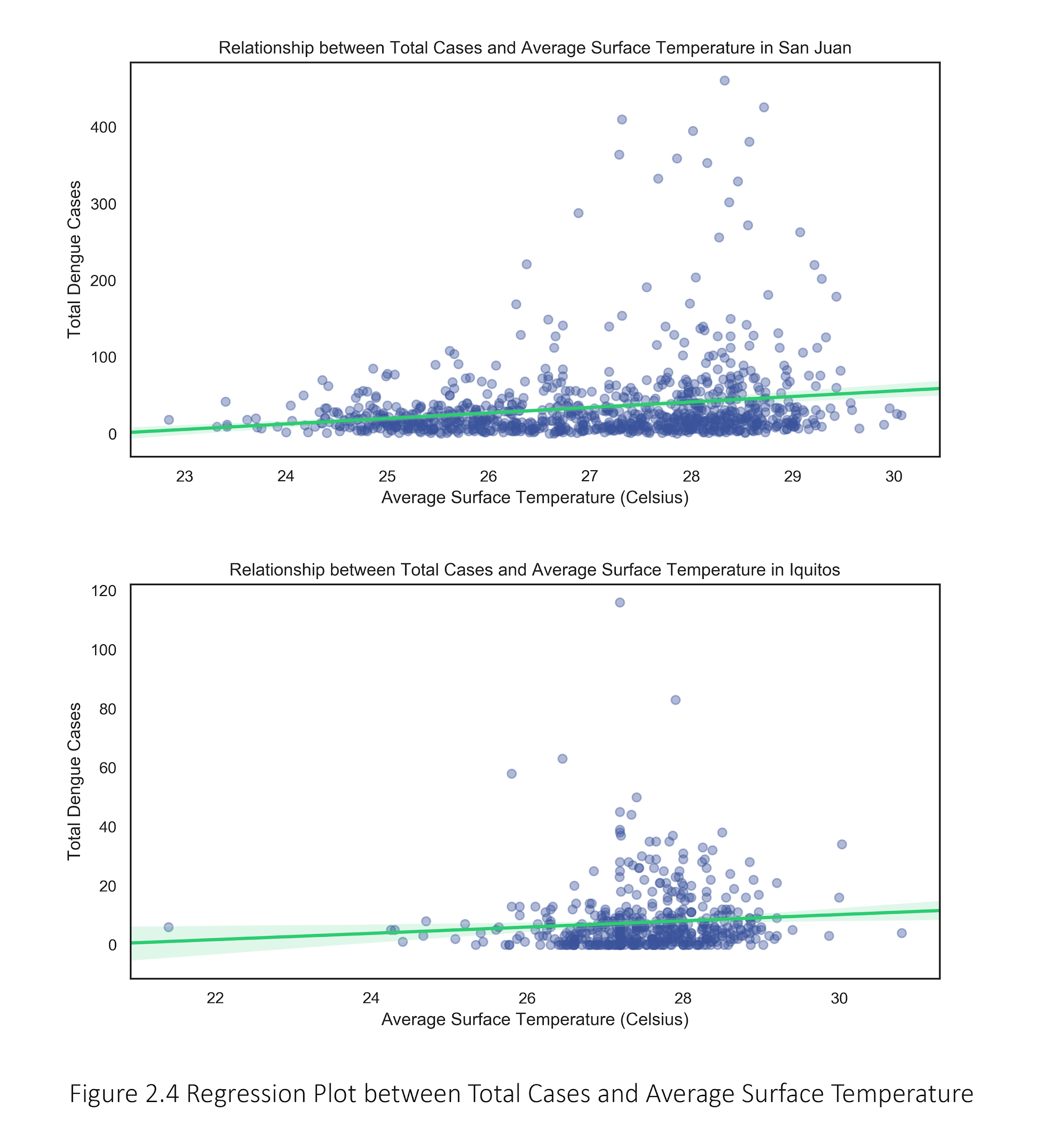
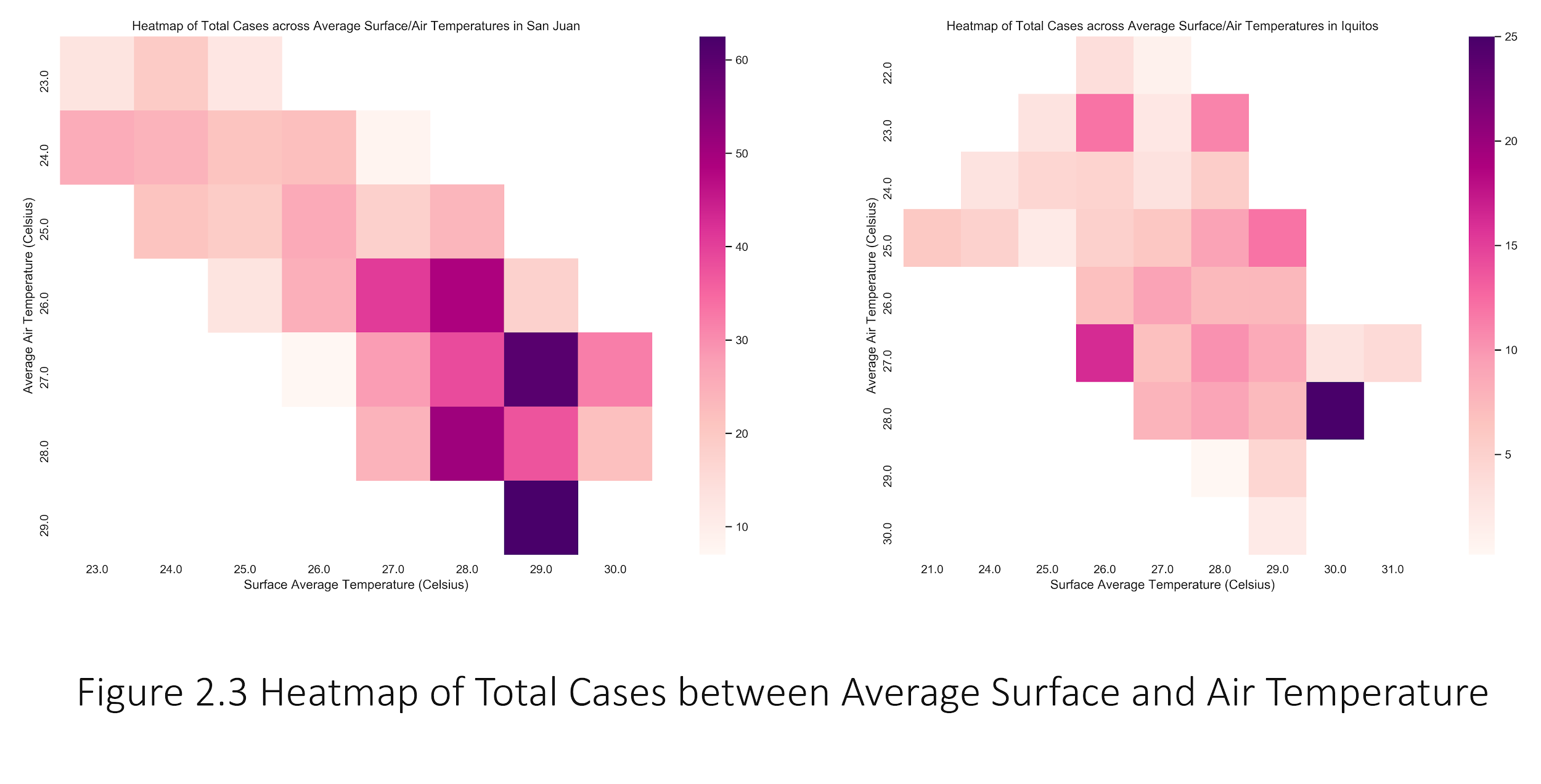
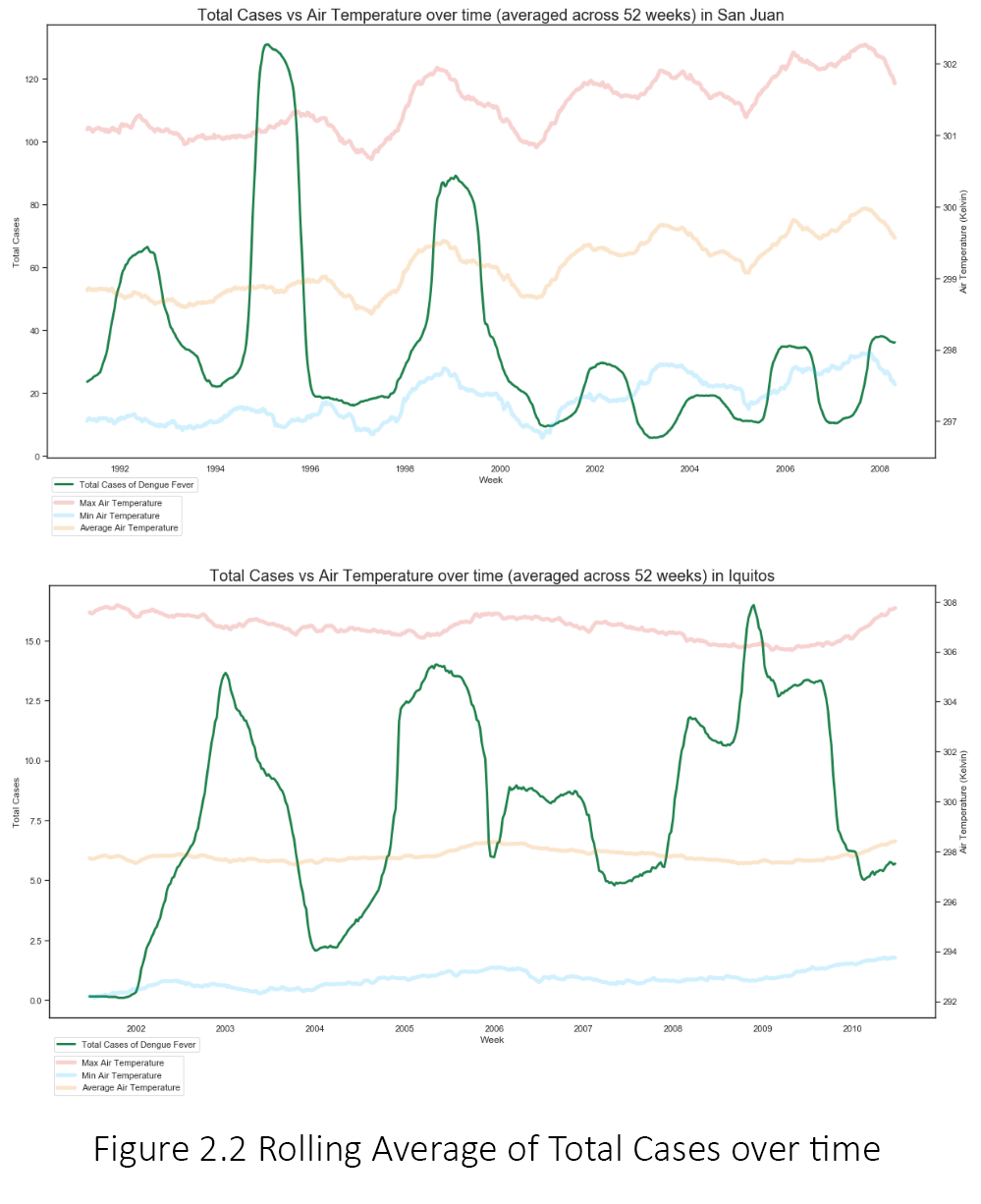
Figure 2.3 Heatmap of Total Dengue Cases against average air temperature and average surface temperature shows that there is a clear trend where there are more cases when temperatures are on average 26°C or greater in San Juan. The relationship is less obvious in Iquitos.

Figure 2.4 Regression Plot between Total Dengue Cases and average surface temperature shows the impact of extreme total cases as temperature increases. The vast majority of total cases are still under 150, or even 100.

Figure 2.5 Station temperature statistics (San Juan) against a timeline for the months of the year with all the outputs accumulated over the entire dataset duration.

Figure 2.6 Station temperature statistics (Iquitos) against a timeline for the months of the year with all the outputs accumulated over the entire dataset duration.





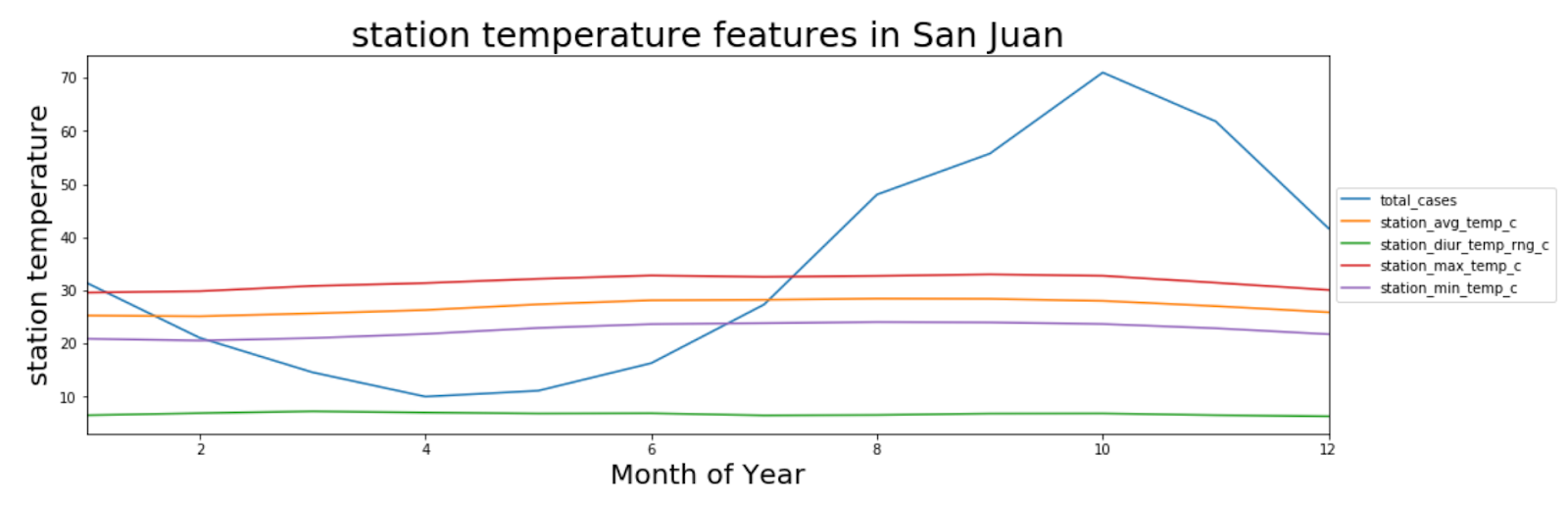
****

Figure 2.5 Mean of the station San Juan temperature values for the station attributes across the year related with Total cases.

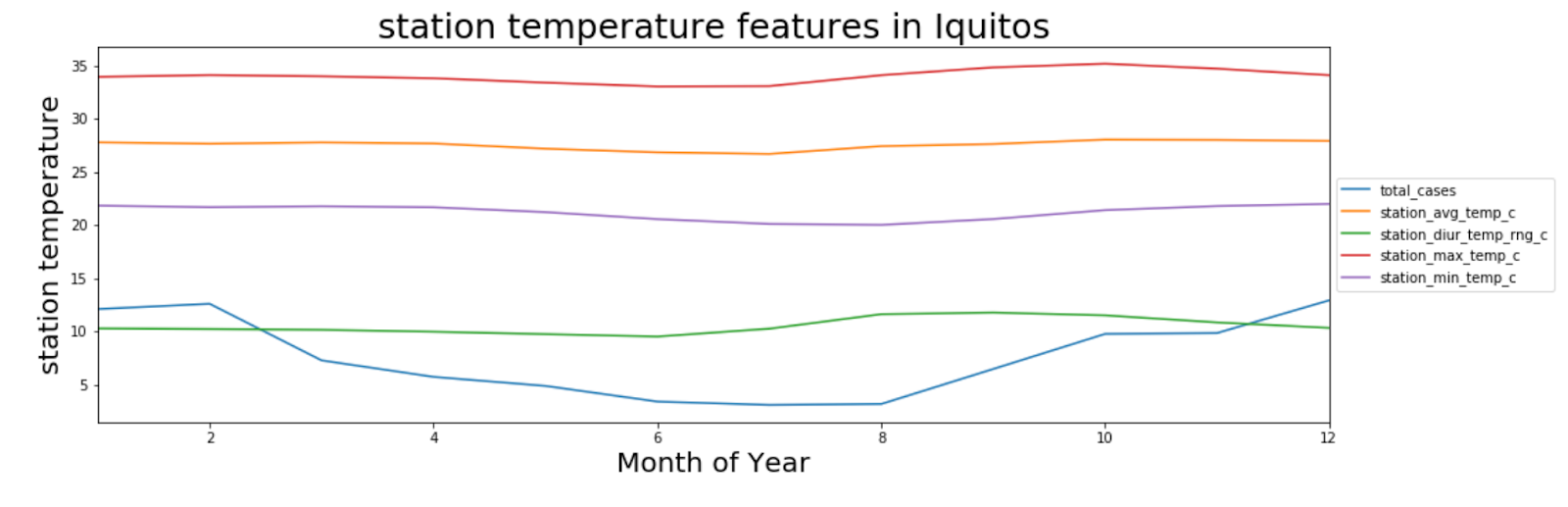
****

Figure 2.6 Mean of the station Iquitos temperature values for the station attributes across the year related with Total cases.

### **Attribute #3: Humidity[[7]](#footnote-7)**

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

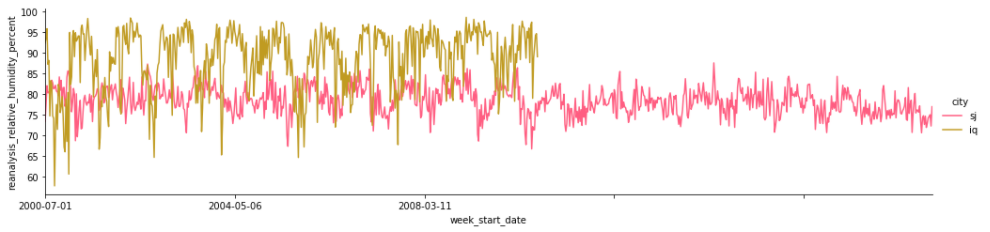


Figure 7 - Humidity % and Total Cases Over Time by City

In San Juan, humidity levels are quite seasonal. Referring to Figure 8(a) below, total cases (indicated by the blue line) 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.

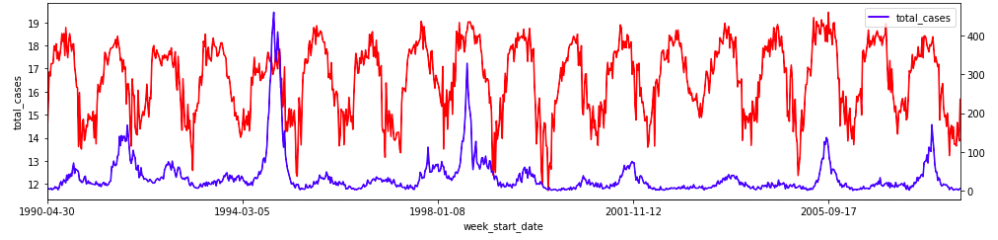


Figure 8(a) - Humidity g per kg and Total Cases Over Time in San Juan

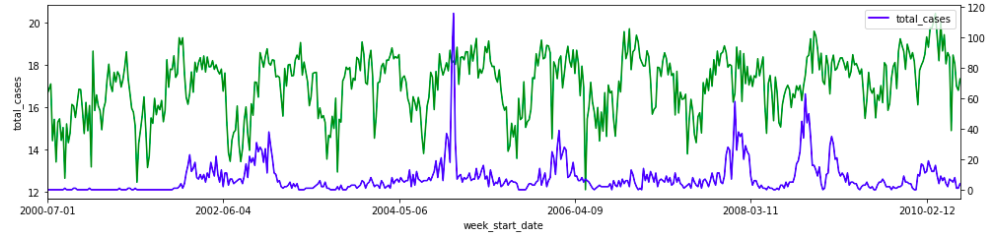


Figure 8(b) - Humidity g per kg and Total Cases Over Time in Iquitos

### **Attribute #4: Vegetation[[8]](#footnote-8)**

The vegetation data that was provided in the dataset was the NDVI index. NDVI stands for "Normalized Difference Vegetation Index".[[9]](#footnote-9)

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).

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, when the area is more Grassland it results in the highest number of Dengue cases.

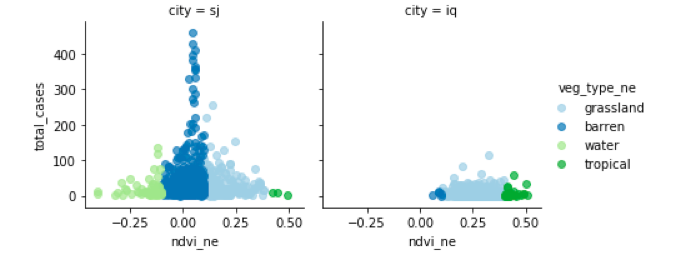


Figure 9(a) - Vegetation Types in NE Region

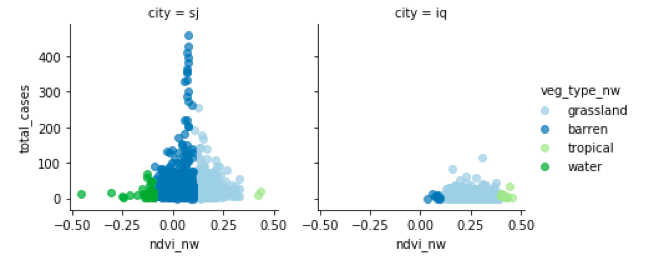


Figure 9(b) - Vegetation Types in NW Region

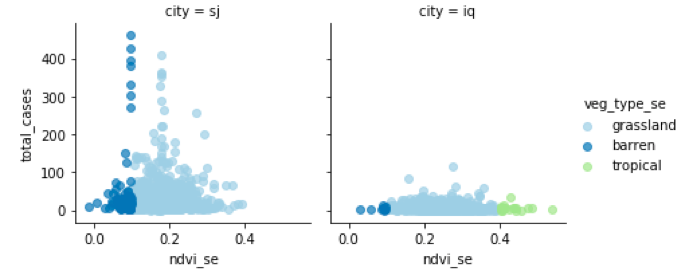


Figure 9(c) - Vegetation Types in SE Region

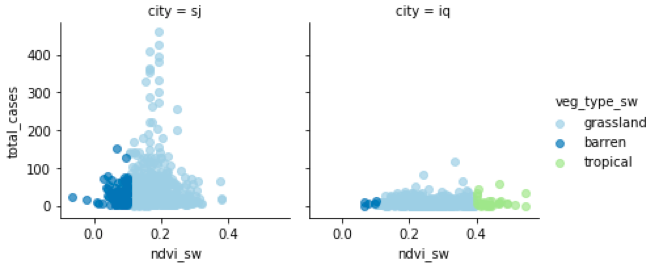


Figure 9(d) - Vegetation Types in SE Region

# **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**

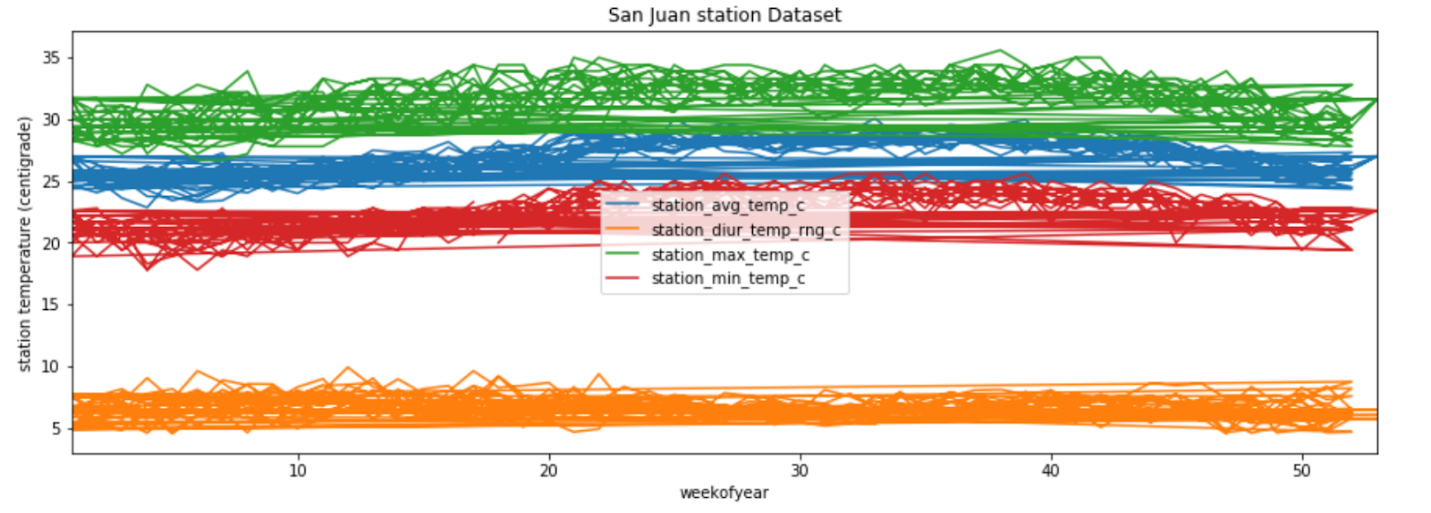
There was no direct correlation between Precipitation and Dengue Cases but has a stronger correlation to humidity.

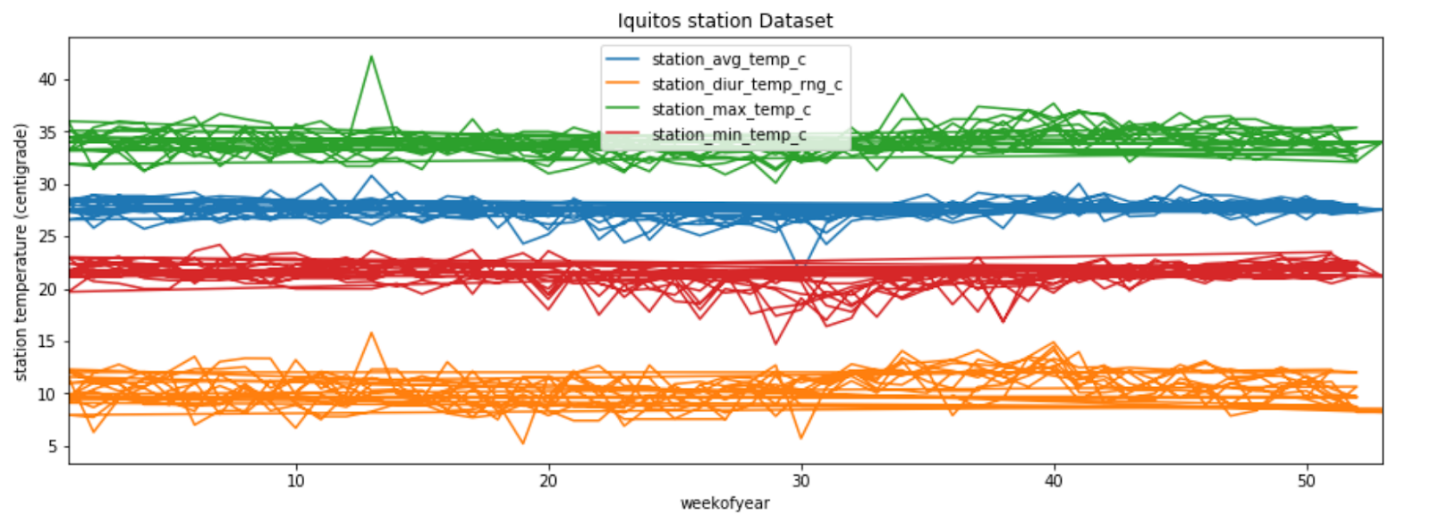
# **Appendix**

### **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.



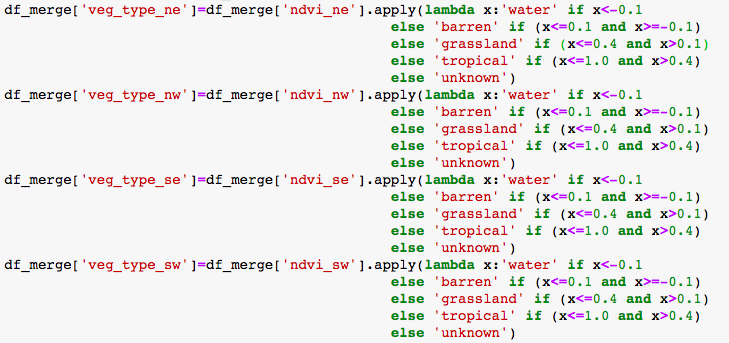


* + **Precipitation and Humidity:** For simplicity the mean for the respective cities was used to fill in the NaN values.

### **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 |



### **Appendix C – Vegetation Plots: NDVI for Each City Overtime**

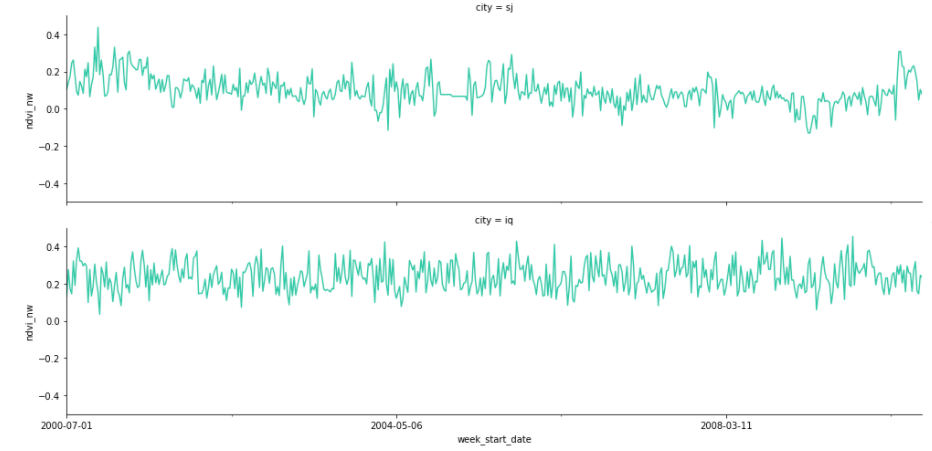


Figure A - NDVI for the North West (NW) Area Overtime

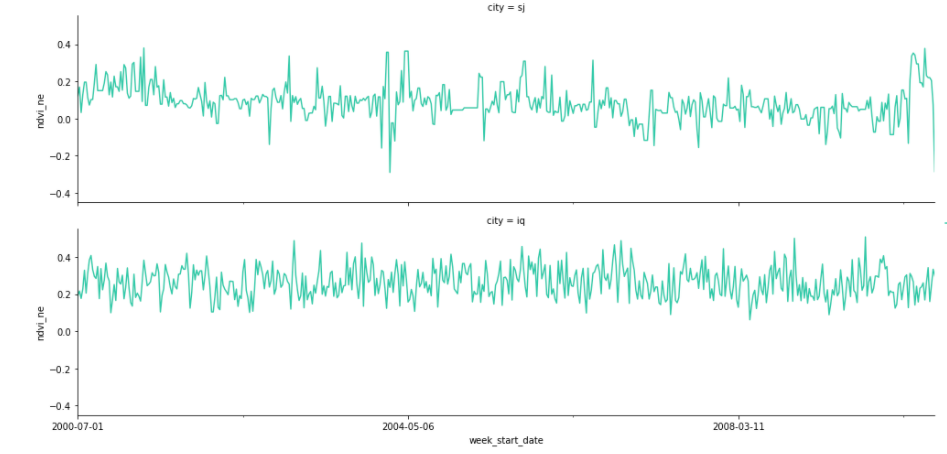


Figure B - NDVI for the North East (NE) Area Overtime

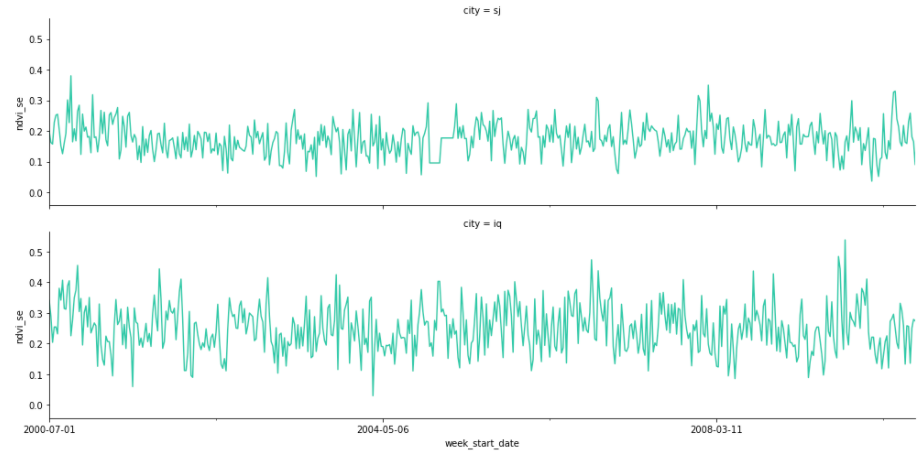


Figure C - NDVI for the South East (SE) Area Overtime

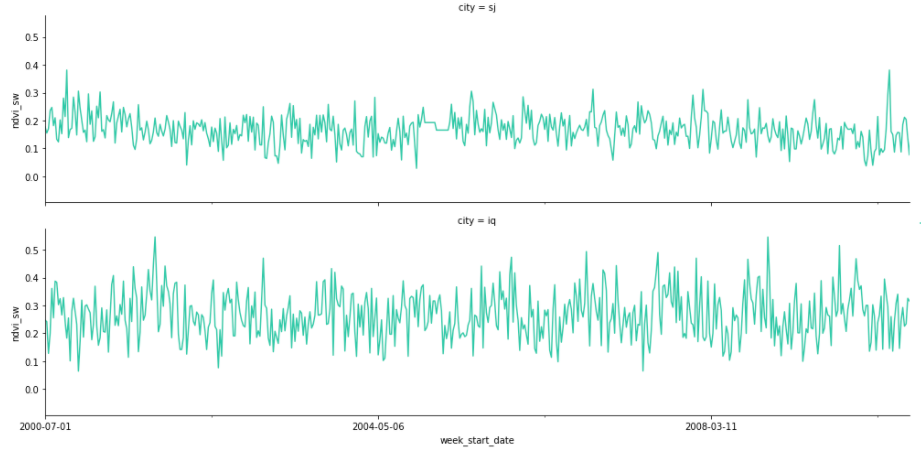


Figure D - NDVI for the South West (SW) Area Overtime

1. Wikipedia. Climate of Puerto Rico. Retrieved from URL: <https://en.wikipedia.org/wiki/Climate_of_Puerto_Rico> [↑](#footnote-ref-1)
2. ibid [↑](#footnote-ref-2)
3. Weather Spark. Average Weather in Iquitos. Retrieved from URL: <https://weatherspark.com/y/24250/Average-Weather-in-Iquitos-Peru-Year-Round> [↑](#footnote-ref-3)
4. ibid [↑](#footnote-ref-4)
5. Refer to Notebook 3 for the detailed analysis and code. [↑](#footnote-ref-5)
6. Refer to Notebook X for the detailed analysis and code. [↑](#footnote-ref-6)
7. Refer to Notebook 2 for the detailed analysis and code. [↑](#footnote-ref-7)
8. Refer to Notebook 2 for the detailed analysis and code. [↑](#footnote-ref-8)
9. Sentinel Hub by Sinergise. NDVI (Normalized Difference Vegetation Index). Retrieved from URL: <https://www.sentinel-hub.com/eoproducts/ndvi-normalized-difference-vegetation-index> [↑](#footnote-ref-9)