



Predicting Dengue Disease

Iquitos, Peru & San Juan, Puerto Rico

Table of Contents

The Abstract.....	2
Literature Review	2
Article #1 DengAI Predicting Disease Spread ML CS 6375.501	3
Article #2 Using LASSO to predict Dengue occurrence by Preethi Subramanian	3
Article #3 Development of a mechanistic dengue simulation model for Guangzhou	4
Article #4 DengAI: Predicting Disease Spread - Adharsh Rajendran, Mithil Gotarne	4
Dataset.....	5
Iquitos, Peru	5
San Juan, Puerto Rico.....	6
Variables (parameters):	6
Feature data example	7
Approach Machine Learning Models for Predicting	8
Multi Linear regression with LASSO.....	8
Random forest (RF)	8
Support vector machine (SVM).....	8
Polynomial regression.....	8
Flow diagram of the process.....	9
Overview and insights on the Dataset.....	10
Importing Datasets, Python Packages	11
Importing the different libraries	11
Exploratory Analysis.....	12
• Univariate	12
• Bivariate.....	14
• Multivariate.....	17
Data Modeling & Validation.....	22
Multiple Linear Regression with LASSO.....	22
Random Forest	23
Support Vector Machine.....	24
Polynomial Regression (Non-linear)	25
Results & Conclusion	26
Reference.....	27

The Abstract

Dengue fever is an acute mosquito-borne disease, transmitted when bitten by Aedes mosquitoes with any one of the four dengue viral serotypes. It occurs mostly in the tropical and sub-tropical part of the world, which covers over 100 countries and 300 million people which equates to 40% of the world population at risk of contracting this fever. I was born in Bangladesh a tropical-monsoon country and growing up, we experienced the continuous threat of mosquito bites specially in winter or wet season. This can cause catastrophic and an epidemic concern in some countries with poor living conditions and current Covid-19 epidemic situation makes it even harder for population in these zones.

Weather plays a huge factor in the outbreak of dengue, as mosquitoes lay their eggs in areas of stagnant water and Warm, damp weather. If we can study these meteorological factors and their relationship to mosquitos, we can provide awareness and people can take necessary precautions from future outbreaks. I will be using different classifications, **python** visualization functions and tools that we used in our data science and predictive analytics course to help me guide and gain insights on dengue cases. I will be using the 3 datasets provided in “DengAI: Predicting Disease Spread” competition in Kaggle hosted by DrivenData (<https://www.kaggle.com/qcnquyen/dengai-predicting-disease-spread>). The datasets include different environmental data on San Juan, Peru & Iquitos, Puerto Rico between (1990-2010) by various U.S. Federal Government agencies. I will be using the Training and test dataset for **Data Mining and knowledge discovery** to gain insights, on what are the different meteorological factors that increases misquotes which leads to more cases in San Juan, Puerto Rico and Iquitos, Pero. I will be using jupyter notebook from anaconda distribution and concepts of python from Udemy.

Literature Review

Dengue Fever (DF) is caused by dengue virus (DENV), quiet often occurs in the tropical and subtropical zones around the world. Just like Covid-19 it has flue like symptoms and might cause fever, rash, muscle pain and joint pain. If DF is serious it might cause severe bleeding, pressure, low blood and even death.

The virus is transmitted to humans through the bites of infected female mosquitoes, primarily the **Aedes aegypti mosquito**. Other species within the Aedes genus can also act as vectors, but their contribution is secondary to Aedes aegypti. The Aedes genus are the type of mosquitoes are liable for spreading and transmitting Dengue Virus, and some other arboviruses such as yellow fever virus (YFV), chikungunya virus (CHIKV), and the infamous Zika virus (ZIKV) around the world. During the trans-Atlantic voyages between

1500s and 1700s, the *Aedes aegypti* arrived in the New World. The Dengue Virus (DV) has become a real issue in South America as the mosquito born disease in the continent.

[Article #1 DengAI Predicting Disease Spread ML CS 6375.501](#)

Based on an estimation of a model, it indicates that about 390 million cases per year for dengue virus infection (95% credible interval 284–528 million) [*Bhatt, S., et al Nature, 2013*], of which 96 million (67–136 million) manifest clinically (with any severity of disease). 3.9 billion people are at risk of infection with Dengue Virus per research on DF prevention [*J. E Cogan, 23, June 2020*]

Per this Article they have focused on gradient boosting from all the other ML techniques they have tested. Using gradient boosting an ensemble technique which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. Per scope of the project, the target was to achieve the best Mean Absolute Error and their model received a value of 4.27 for city Iquitos and 11.41 for San Juan. The Model was able to predict a better result for Iquitos. They have other techniques such as KNN, but they were getting higher MAE for both the cities.

[Article #2 Using LASSO to predict Dengue occurrence by Preethi Subramanian](#)

Although it is a complex relationship between Dengue Fever and climate, a growing number of scientists argue that climate change is likely to produce distributional shifts that will have significant public health implications worldwide [*Lindsay P. Campbell, 2015*].

This article focused on predicting the dengue cases using the (LASSO) Least Absolute Shrinkage Selector Operator. “LASSO is a type of penalized regression which can perform feature selection to prevent over-fitting”. It overcome the weaknesses of the Artificial Neural network (ANN) methodologies using time series forecasting models and other models. “Multiple LASSO models were created for each disease in different countries for 1-26 weeks

ahead. The optimal constraint parameters were determined using ten-fold cross validation in R. The performance of the other models waned beyond 4 weeks’ horizon” [*Dr.Preethi Subramanian, 2019*]. The meteorological variables were found to add more value in temperate region where cyclic patterns were observed for the diseases and climatic data, in comparison to the more stable climate in equatorial region.

Article #3 Development of a mechanistic dengue simulation model for Guangzhou

This Article is on Chinese Lab data, I wanted to compare and get insights on a completely different data set and gain more knowledge on the models applied for prediction. Dengue infection in China has increased dramatically in recent years. Guangdong province

(main city Guangzhou) accounted for more than 94% of all dengue cases in the 2014 outbreak. In China, dengue cases have been recorded each year for the past 25 years. [*Cheng Q et al. (2016)*].

The data was taken from 4 different serotype of Dengue disease. All tests were done in between 2004 and 2010 with controlled temperature and supervised but mimicking real world climate. The Correlation coefficient (r) values were the following (0.28, 0.37, 0.32 and 0.75). “The validity of the dengue simulation model was analyzed by using cross-correlation analysis to compare the simulated case data with the reported dengue cases”. The best model fit that was considered for this study had the highest positive predictive value (PPV) of 0.83. Simple linear regression was used to test the probability of outbreaks occurring each year in the same season in Guangzhou.

After evaluating the results, it was concluded the simulation model did not work using the multiple serotypes. “First, there is a possible existing bias in the reporting of actual dengue cases which can lead to under-reporting of the magnitude of an outbreak. Another reason is the set simulated population size which differs from the actual population size, due to computer simulation capacities of the dengue model used” [*G. Mincham 2019*].

Article #4 DengAI: Predicting Disease Spread - Adharsh Rajendran, Mithil Gotarne

Recently dengue fever is spreading all around the world, historically the disease has been most prevalent in Southeast Asia and Pacific regions. These days nearly billion Dengue fever cases has been reported in Latin America [*Luis Villar, M.D, 2015*]

1. Random Forest

A useful tool to learn about feature analysis and comparison.

2. XGBoost

XGBoost is an implementation of gradient boosted decision trees designed for speed and performance that is dominative competitive machine learning

3. Neural Networks

This basic tool uses classification and clustering tools. As this tool is a developed pattern recognition device, we decided practicing with this would provide a great learning experience. On this article three different ML models were used to see the best performance. “Each of the models had very similar scores ranging from 25 to

32. The only major difference created was separating the cities 'datasets during the preprocess. The initial assumption was XGBoost would overwhelm the other two methods chosen. However, it only slightly outperformed.” [Github - Adharsh Rajendran, 2019]

Through previous observations and researches, it is found out that dengue outbreaks occur mostly in a particular time of year under certain type of conditions. If we can predict the cases during the peak season along with the metrological factors that are most common in this season can help the health officials take proper measurements before the season arrives to control the situation.

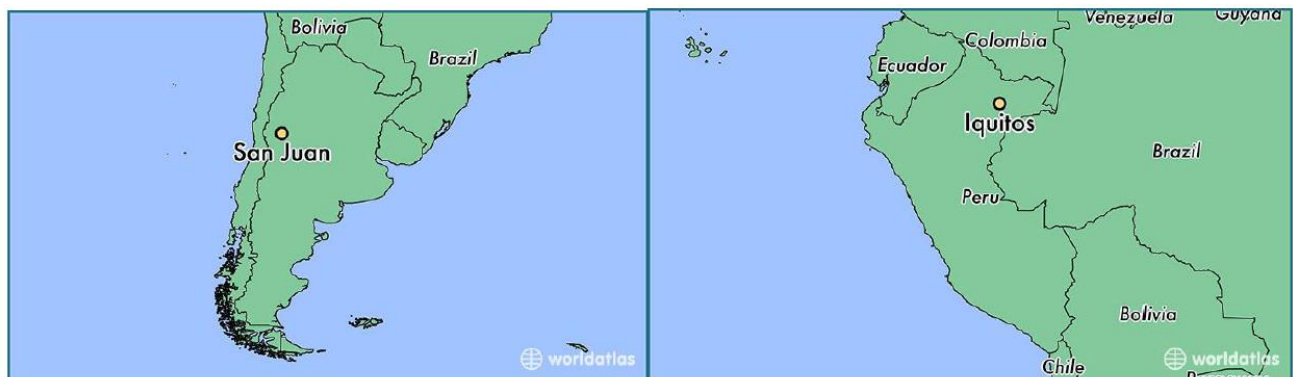
Dataset

I have downloaded the data set from Using environmental data collected by various **U.S. Federal Government agencies**—from the Centers for Disease Control and Prevention to the National Oceanic and Atmospheric Administration in the U.S. Department of Commerce. This project is part of a Data Science competition on drivendata.org webpage

Iquitos, Peru

Torres, Orduna, Piña-Pozas, Vázquez-Vega and Sarti [Torres, J.R..2017] distinguished that Iquitos has equatorial climate with and dengue outbreak in rainy season around March. The Andean area experienced heavy downpour induced by the La Niña phenomenon in early 2011 as reported by the International Federation of Red Cross and Red Crescent Societies (IFRC) [Peru and Bolivia, 2011]. Consequently, Loreto (the region where Iquitos was located) was the worst hit area in the country’s 2011 most serious dengue crisis IFRC [Peru and Bolivia, 2011].

The city of Iquitos sits at the Amazon Rivers of northeast Peru and the confluence of [Nanay, Itaya. Stoddard et. al (2014)] research data reported from a laboratory confirmed dengue dynamics between 2000 and 2010. Their studies are split within 3 seasons: trimester I, II, and III. The warmest temperature recorded with Maximum and mean temperature were between November and April. Over all the years, rainfall was highest between 2003 and 2008 with significantly less rainfall subsequent years.



San Juan, Puerto Rico

Sougata, Acebedo and Chua [Sougata, D. 2017] mentioned that San Juan reported that due to higher tropical monsoon climate and populations density there was an increase of Dengue cases which stood at 2.5 times more compared to Iquitos, Peru. Puerto Rico went through an epidemic during the period of 2007 and 2010 but systematic review conducted

[Torres, J.R..2017] did not support local weather as a crucial factor in explaining the changes in the annual case in Puerto Rico.

San Juan in Puerto Rico annually reports to have an average air surface temp of 24-29 °C, with average precipitation of about 1800 mm. During the dry season (0-50 mm occurs between March and June), air temperatures fluctuate between 36-40 °C, while the rainy seasons report 30-35 °C. [Laureano-Rosario et.al (2018)] used AI networks (ANNs) to predict dengue fever cases between 1994-2012

Variables (parameters):

Below is the list of meteorological parameters and an example of assigned values

The following set of information has been provided on a (year, weekofyear) timescale:
(Where appropriate, units are provided as a unit suffix on the feature name.)

City and date indicators

- `city` – City abbreviations: `sj` for San Juan and `iq` for Iquitos
- `week_start_date` – Date given in yyyy-mm-dd format

NOAA's GHCN daily climate data weather station measurements

- `station_max_temp_c` – Maximum temperature in °C
- `station_min_temp_c` – Minimum temperature in °C
- `station_avg_temp_c` – Average temperature in °C
- `station_precip_mm` – Total precipitation
- `station_diur_temp_rng_c` – Diurnal temperature range in °C

PERSIANN satellite precipitation measurements (0.25x0.25 degree scale)

- `precipitation_amt_mm` – Total precipitation

NOAA's NCEP Climate Forecast System Reanalysis measurements (0.5x0.5 degree scale)

- `reanalysis_sat_precip_amt_mm` – Total precipitation
- `reanalysis_dew_point_temp_k` – Mean dew point temperature in °C
- `reanalysis_air_temp_k` – Mean air temperature in °C
- `reanalysis_relative_humidity_percent` – Mean relative humidity
- `reanalysis_specific_humidity_g_per_kg` – Mean specific humidity

- `reanalysis_precip_amt_kg_per_m2` – Total precipitation
- `reanalysis_max_air_temp_k` – Maximum air temperature in
- `reanalysis_min_air_temp_k` – Minimum air temperature in
- `reanalysis_avg_temp_k` – Average air temperature in
- `reanalysis_tdtr_k` – Diurnal temperature range

Satellite vegetation - Normalized difference vegetation index (NDVI) - NOAA's CDR Normalized Difference Vegetation Index (0.5x0.5 degree scale) measurements

- `ndvi_se` – Pixel southeast of city centroid
- `ndvi_sw` – Pixel southwest of city centroid
- `ndvi_ne` – Pixel northeast of city centroid
- `ndvi_nw` – Pixel northwest of city centroid

Feature data example

For example, a single row in the dataset, indexed by (city, year, weekofyear)- (San Juan, Puerto Rico, 1994, 18), has the following values below:

<code>week_start_date</code>	5/7/1994
<code>total_cases</code>	22
<code>station_max_temp_c</code>	33.3 °C
<code>station_avg_temp_c</code>	27.75714286 °C
<code>station_precip_mm</code>	10.5
<code>station_min_temp_c</code>	22.8 °C
<code>station_diur_temp_rng_c</code>	7.7 °C
<code>precipitation_amt_mm</code>	68
<code>reanalysis_sat_precip_amt_mm</code>	68
<code>reanalysis_dew_point_temp_k</code>	295.2357143
<code>reanalysis_air_temp_k</code>	298.9271429
<code>reanalysis_relative_humidity_percent</code>	80.35285714
<code>reanalysis_specific_humidity_g_per_kg</code>	16.62142857
<code>reanalysis_precip_amt_kg_per_m2</code>	14.1
<code>reanalysis_max_air_temp_k</code>	301.1
<code>reanalysis_min_air_temp_k</code>	297
<code>reanalysis_avg_temp_k</code>	299.0928571
<code>reanalysis_tdtr_k</code>	2.671428571
<code>ndvi_location_1</code>	0.1644143
<code>ndvi_location_2</code>	0.0652
<code>ndvi_location_3</code>	0.1321429
<code>ndvi_location_4</code>	0.08175

Approach Machine Learning Models for Predicting

Specific goal for this project is to predict dengue cases for San Juan, Puerto Rico and Iquitos, Peru. Hence, a model needs to be developed by analyzing the relationship between dengue cases and climate data to predict peak time of dengue outbreak and maximum weekly incidence. Understanding these patterns will help to better develop effective public health strategies to combat the potential increases in dengue outbreak.

Multi Linear regression with LASSO is a standard statistical approach that allows questions that consider the role(s) of multiple independent variables in a single dependent variable to be answered (Nathans, 2012). Three types of work can be done by MLR: (1) defining relationships between dependent variables and independent variables, (2) Estimating the values of the dependent variables based on independent variables' observed values, (3) identifying independent variables affecting dependent variables (Schneider 2010; Jeon, 2015).

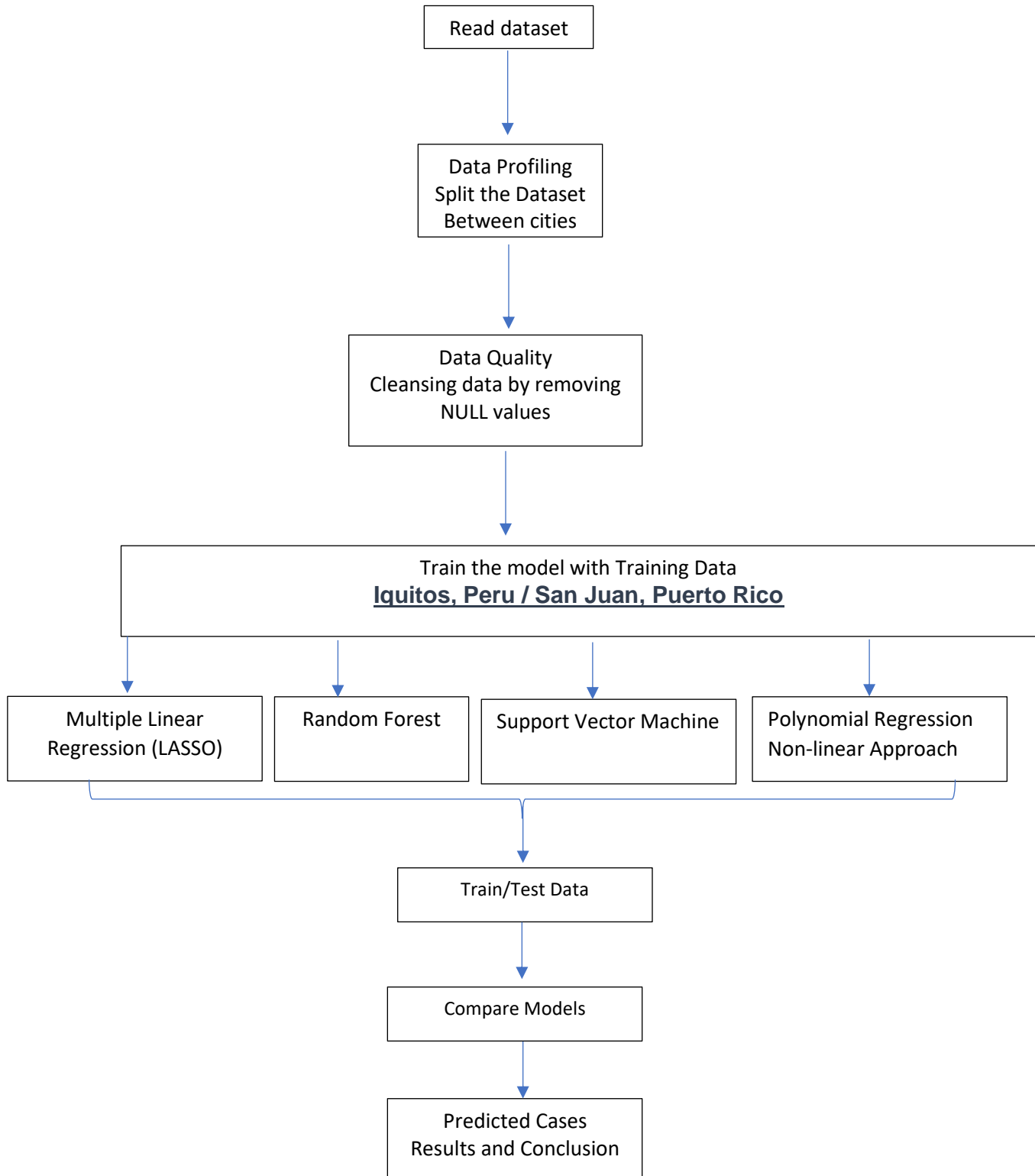
Least Absolute Selection and Shrinkage LASSO operator is a linear regression that which can perform feature selection to prevent over-fitting by using shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. LASSO ridge regression is good for fewer parameters and small datasets. This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model.

Random forest (RF) is an ensemble classifier and consisting of many Decision Tress which makes it look like a forest of many trees. The forest then chooses the classification of having the most 'votes' classification outcome) or the average of all trees in the forest (for numeric classification outcome). Random Forest can reduce the variance resulted from having on decision tree as the random forest algorithm considers the outcomes from many decision trees.

Support vector machine (SVM) algorithm can classify both linear and non-linear data. It first maps each data item into an n-dimensional feature space where n is the number of features. It then identifies the hyperplane that separates the data items into two classes while maximizing the marginal distance for both classes and minimizing the classification errors.

Polynomial regression fits a nonlinear Polynomial Relapse may be a shape of direct relapse in which the relationship between the free variable x and subordinate variable y is modeled as an nth degree polynomial. Polynomial relapse fits a non-linear relationship between the value of x and the comparing conditional cruel of y, signified $E(y|x)$

Flow diagram of the process



Overview and insights on the Dataset

There are Two target cities and but for CIND820 I will be merging both the cities to make the dataset bigger. I will be running the models on the combined list if 1457 records. I will need to run some basic functions to see the number of Null values, Mean, medium, Standard deviation etc. thorough using basic *describe()* function from Python. In the data set the records are sorted by city and year, followed by weeks and other metrological factors, and ending with Total Cases per week.

Mean relative & mean specific humidity, in % & kg, respectively, are described. Minimum, maximum, & mean of air temperatures in Kelvin (K) are provided for each city. An interesting set of variables are the Normalized difference Vegetation Indices (NDVI), derived from remote sensing data closely linked to drought conditions. NDVI is a measure of plant wellbeing based on how the plant reacts light at certain frequencies, i.e. a calculation of vegetation wellbeing. This esteem ranges from -1 to 1. Negative values correspond to dead plants or lifeless objects, sound plants have positive lists.

The relationship between the dengue epidemic and greenness indexes, such as normalized difference vegetation index (NDVI) or enhanced vegetation index (EVI), is not consistent. Some studies indicate that the dengue epidemic reveals a positive association with vegetation, while others have found that low vegetation cover areas present increased dengue incidence rates. This inconsistency may be explained by regional differences. So, I will run 4 ML models with the NDVI parameters and without it and see which one gives a better Mean Absolute Error (MAE) and R.

Since we are predicting dengue cases which is the outcome and is the dependent variable and it is based on multiple parameters (independent variables), this will be a multivariate analysis. Multivariate analysis is a more complex form of statistical analysis technique and used when there are more than two variables.

Principal component Analysis examination is beneficial in huge datasets because it decreases their dimensionality, while making strides interpretability at a fetched of minimizing data misfortune. This strategy creates new uncorrelated factors which maximizes change in progression (*Jollie & Cadima, 2016*). Ahmed & Siddiqui (2014) perform PCA on Dengue cases with 5 climatic variables: wind speed, precipitation, most extreme temperature, least temperature, and relative stickiness.

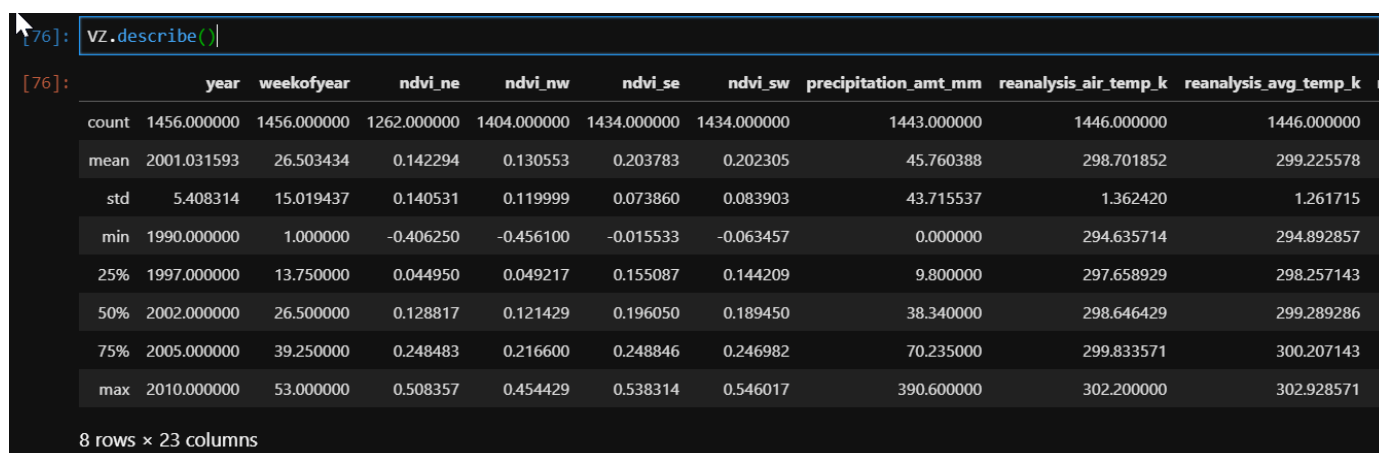
Importing Datasets, Python Packages

Importing the different libraries

- There are Three datasets and we will Importing “**Total_Cases**” from “*DengAI_Predicting_Disease_Spread_Training_Data_Labels*” file and join it to the attributes to “*DengAI_Predicting_Disease_Spread_Training_Data_Features*”.
- As the larger part of the information is climate information, there will be different similitudes with a few of the highlights. Dropping the columns with the high variance and highlights with heavy commonalities is the strategy I am choosing to extend exactness.
- We will need to import some specific Python libraries for our data analysis, first we are importing NumPy which is used for working with arrays. NumPy is also capable of working in linear algebra, Fourier transform, and matrices domain. NumPy stands for Numerical Python.
- Pandas library is written for Python which is used for data manipulating and analysis. It offers data structures and operations for manipulating numerical tables and time series.
- Imported matplotlib.pyplot, which is a collection of function and each Pyplot function makes a figure, makes a plotting region in a figure, plots a few lines in a plotting range, beautifies the plot with names, etc.
- Seaborn is a Python information visualization library based on matplotlib. It gives a high-level interface for drawing appealing and enlightening factual design.
- Matplotlib is basically sent for essential plotting. Visualization utilizing Matplotlib by and large comprises of bars, pies, lines, scramble plots and so on. Seaborn, on the other hand, gives a assortment of visualization designs. It employment less language structure and has effectively curiously default themes.
- Package train_test_split from Sklearn, which splits the dataset into two separate data set, train set and a test set. With this function we don't need to split it manually, we need to let the function know the percentage of train and test ratio through parameters.
- Sklearn.linear_model fits a linear model with coefficients $w = (w_1, \dots, w_p)$ to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation. Parameters fit_interceptbool, default=True.

Exploratory Analysis

- Initially I wanted to split the following *“DengAI_Predicting_Disease_Spread__Training_Data_Features”* data set per city but due the size of the data set, I kept it as is and ran the ML models on one dataset.
- Exploratory Data Analysis (EDA)** is the process of analyzing Univariate, Bivariate and Multivariate data sets. EDA involves introductory information investigation of the data sets before running any ML models on it. The goal of EDA is to gain insight about the data and its underlying structure. It helps in identifying the independent and dependent variables and their impact on the outcome.
 - Some of the steps of EDA that I took are as follows:
 - Removing the null values and replacing them by mean of the attribute
 - Finding the outliers in the dataset
 - Choosing the attribute and finding a way to measure the performance
 - Define and estimate different parameters for the data and find out the associated confidence intervals.
 - The function *describe()* will give us essential insights, such as the **Mean, Median, Mode**, and any **Outliers** of the information. It'll too give us progressed stats in a single data-frame for example **standard deviation, percentiles** and **min** and **max** values (Fig:1).



```
[76]: VZ.describe()
```

	year	weekofyear	ndvi_ne	ndvi_nw	ndvi_se	ndvi_sw	precipitation_amt_mm	reanalysis_air_temp_k	reanalysis_avg_temp_k
count	1456.000000	1456.000000	1262.000000	1404.000000	1434.000000	1434.000000	1443.000000	1446.000000	1446.000000
mean	2001.031593	26.503434	0.142294	0.130553	0.203783	0.202305	45.760388	298.701852	299.225578
std	5.408314	15.019437	0.140531	0.119999	0.073860	0.083903	43.715537	1.362420	1.261715
min	1990.000000	1.000000	-0.406250	-0.456100	-0.015533	-0.063457	0.000000	294.635714	294.892857
25%	1997.000000	13.750000	0.044950	0.049217	0.155087	0.144209	9.800000	297.658929	298.257143
50%	2002.000000	26.500000	0.128817	0.121429	0.196050	0.189450	38.340000	298.646429	299.289286
75%	2005.000000	39.250000	0.248483	0.216600	0.248846	0.246982	70.235000	299.833571	300.207143
max	2010.000000	53.000000	0.508357	0.454429	0.538314	0.546017	390.600000	302.200000	302.928571

8 rows x 23 columns

Fig:1

- I will be looking at the parameters in the dataset at **Univariate, Bivariate** and **Multivariate** level. The examination of Univariate information is the simplest form of examination since the data examines as if there is only one variable.
- Univariate** analysis, to find the number of records available for San Juan and Iquitos
 #dn_iq represents **Iquitos**

```
dn_iq = IRD_MLR[IRD_MLR['city']=='iq']
```

```
len(dn_iq)
```

520

```
[14]: #dn_iq represents Iquitos
      dn_iq = IRD_MLR[IRD_MLR['city']=='iq']
      len(dn_iq)

[14]: 520
```

#dn_sj represents **San Juan**

```
dn_sj = IRD_MLR[IRD_MLR['city']=='sj']
```

```
len(dn_sj)
```

936

```
[15]: #dn_iq represents Iquitos
      dn_sj = IRD_MLR[IRD_MLR['city']=='sj']
      len(dn_sj)

[15]: 936
```

The average Humidity in San Juan and Iquitos is between 70 and 85 and 90 and 97.

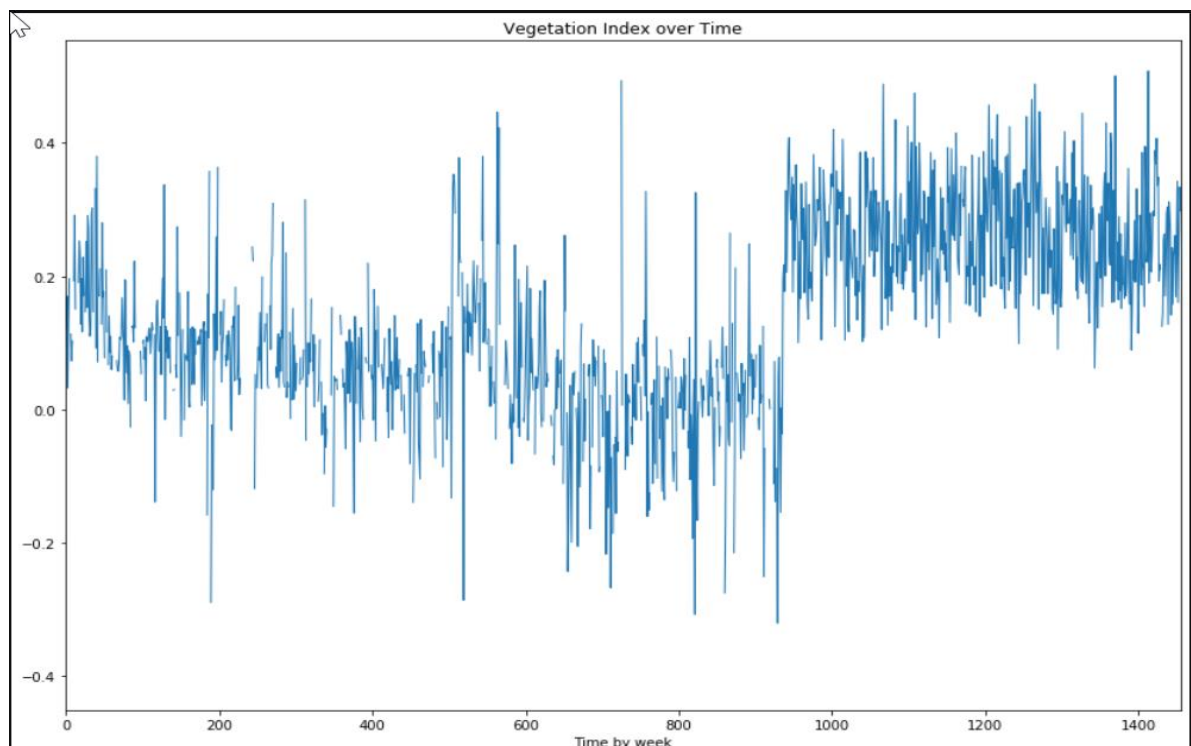


Fig:2

Fig:2 shows the ndvi_ne, vegetation index over time in a time series plot.

- **Bivariate** data involve two variables, the analysis of this type of data is mostly an analysis to find out the relationship between the two variables. Here in Target attribute which is e.g. **“Total_cases”** vs **“weekofyear”**. Using jointplot diagram (Fig:3) from Seaborn), we can see that most of the case are denser between 25th and 50th weeks, which means the Dengue cases are high around the 25th week of a year and the cases starts to fall from 50th week.

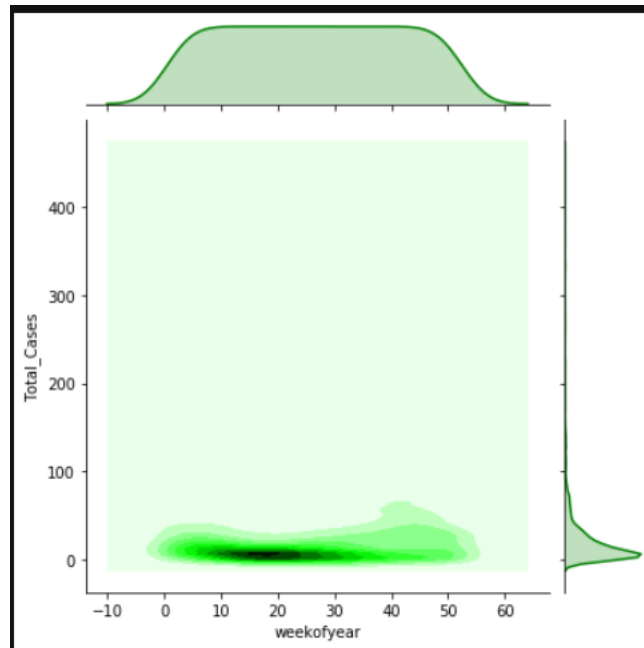


Fig:3

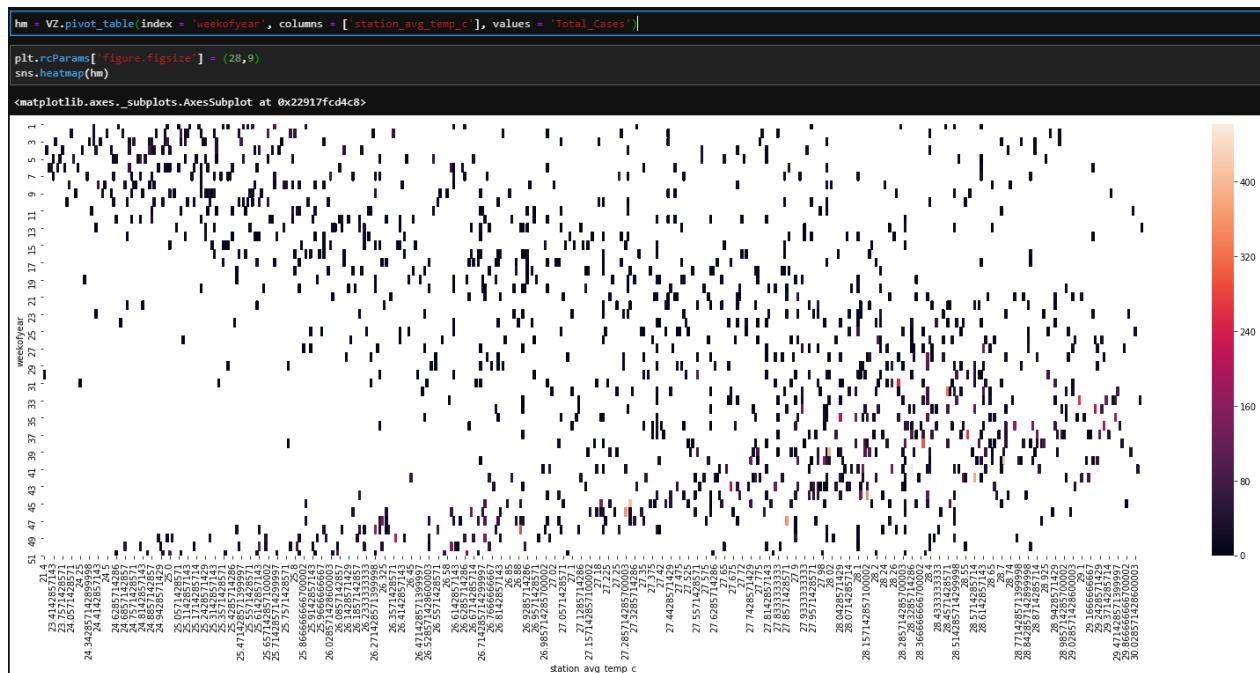


Fig:4

From the Fig:4 plot heatmap, shows how the values of total cases during a particular week of a year on a certain temperature. From this plot we can see for example when the temperature is 27.28 and 27.32 and between 45th and 46th week the Cases are above 400, which are like the outliers as average number of cases per week is around 27.

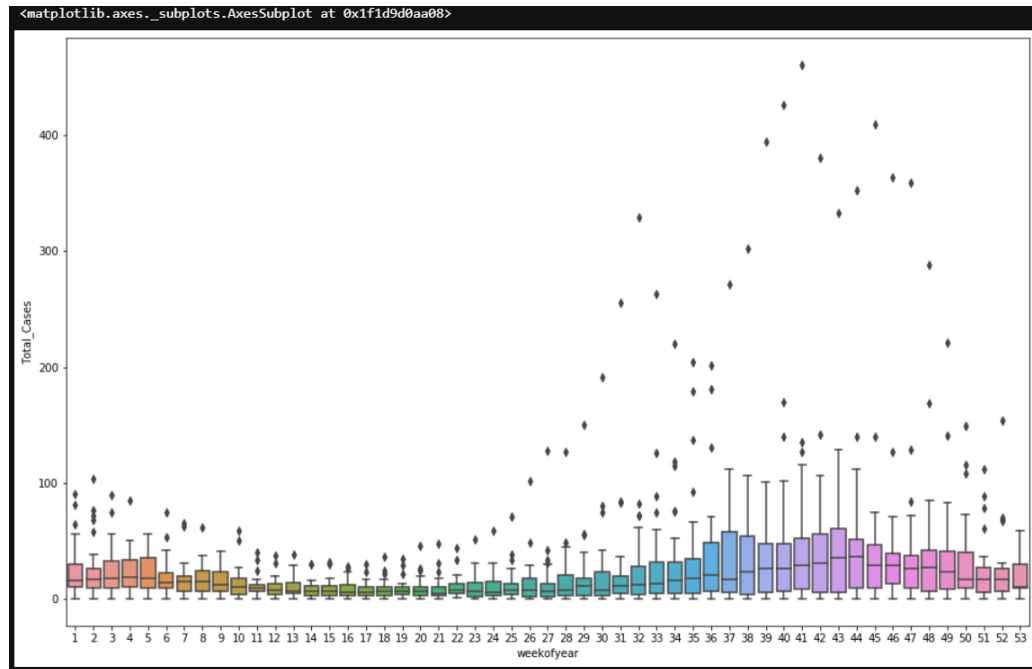


Fig:5

Since our main goal is to predict Dengue Cases, it is vital that we find as many data points we can dig out through EDA analysis that shows the effect of the independent factors on Dengue cases and the through different plots. In the box plot diagram in Fig:5, we see that there are lot of outlier values, specially between week 25 and 49. Fig:6 shows Total_Cases vs per year, we can see that there was a spike in cases during 1994, 1998 and around 2005 & 2007.

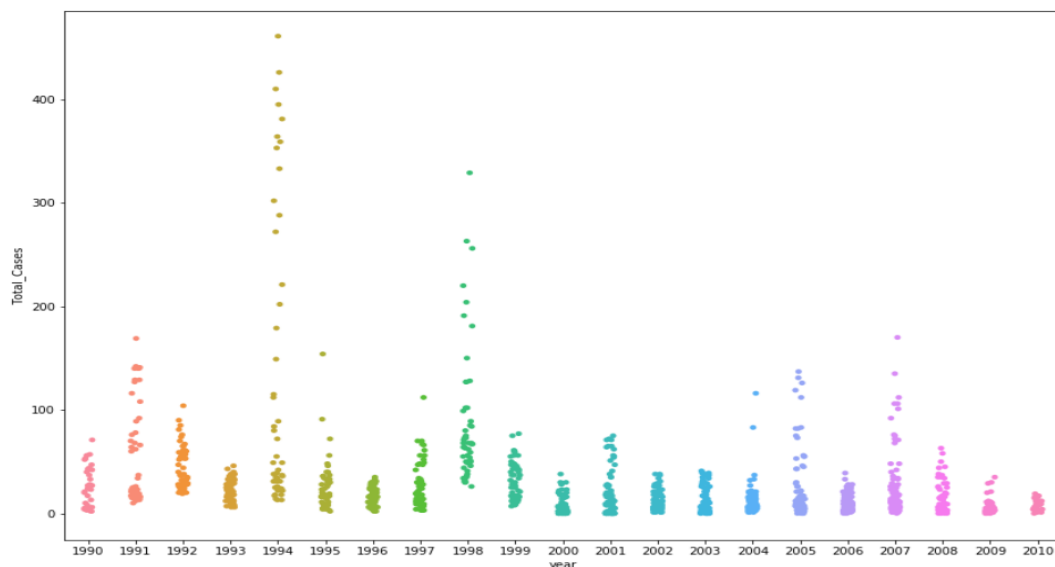


Fig:6

While comparing the effect of temperature on Total cases, it seems like high number of cases are occurring when the average temperature is between 26.5°C to 29°C.

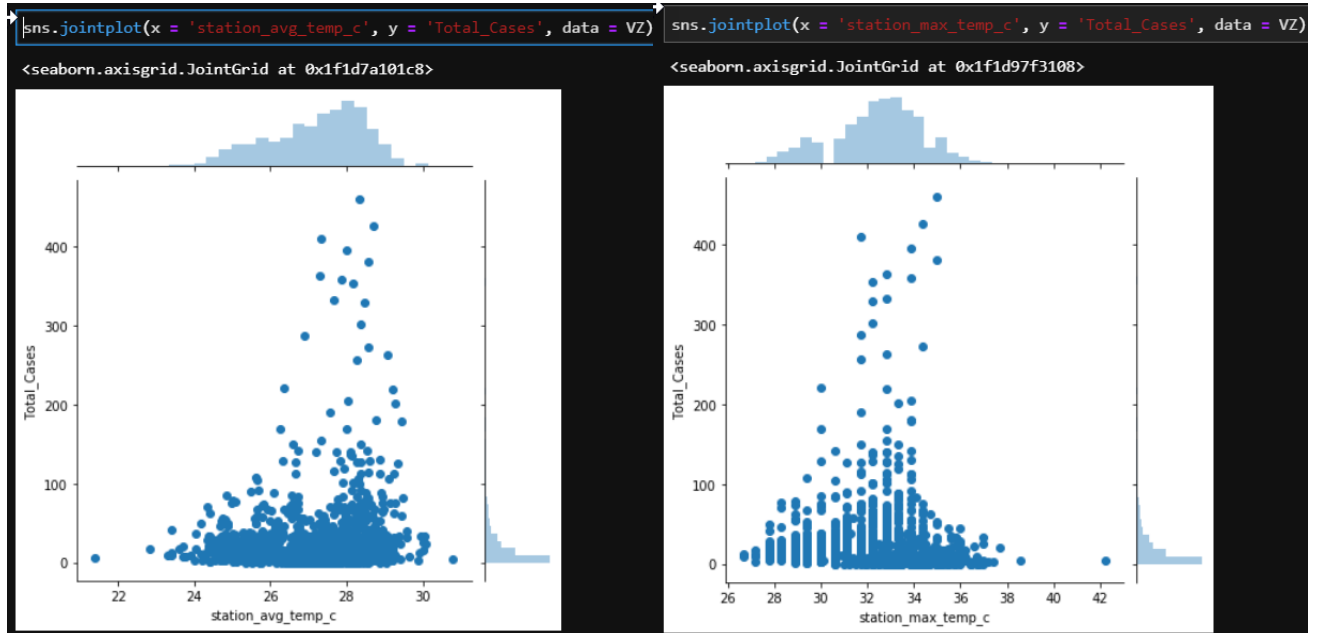


Fig:7

Fig:8

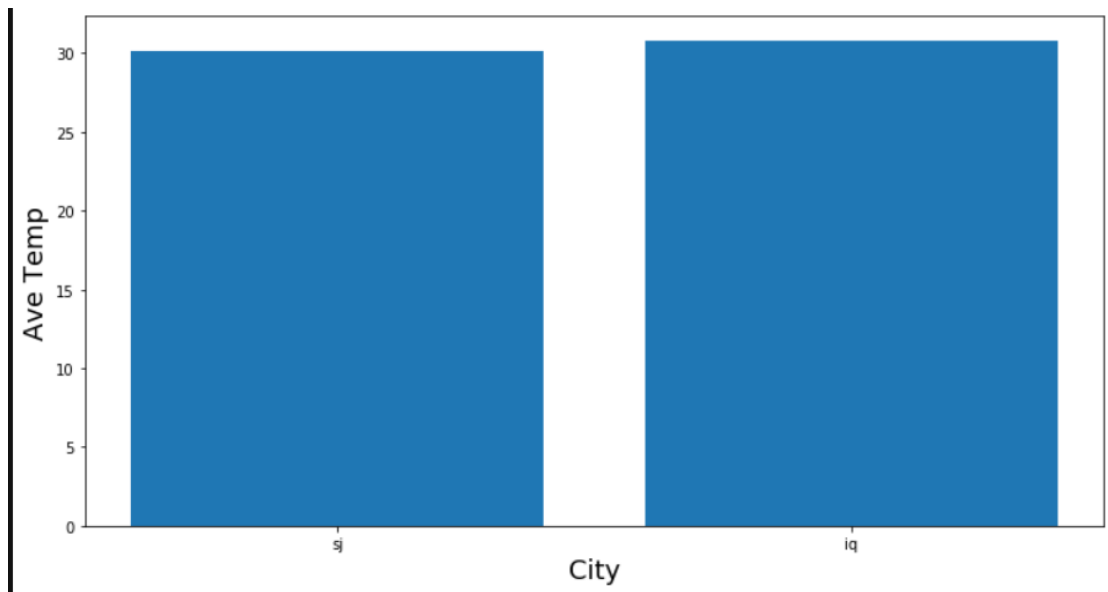


Fig:9

Fig:8, showing highest cases are between 27°C and 29°C and cases starts to pick up from 23°C. The Maximum temperature sometimes reaches to 36°C and 38°C but adding few extra Cases. Fig:9, showing the comparison of Avg temperature by weekly at San Juan and Iquitos, which tells us that even though we are dealing with two different cities, there are lot of similarities in the variables for the two cities.

- **Multivariate** data analysis involves three or more variables, it is like bivariate but contains more than one dependent variable. In the heatmap below (Fig:10), we see the NULL values, shown by the white lines. We have replaced the NULL values by the mean of each attribute.

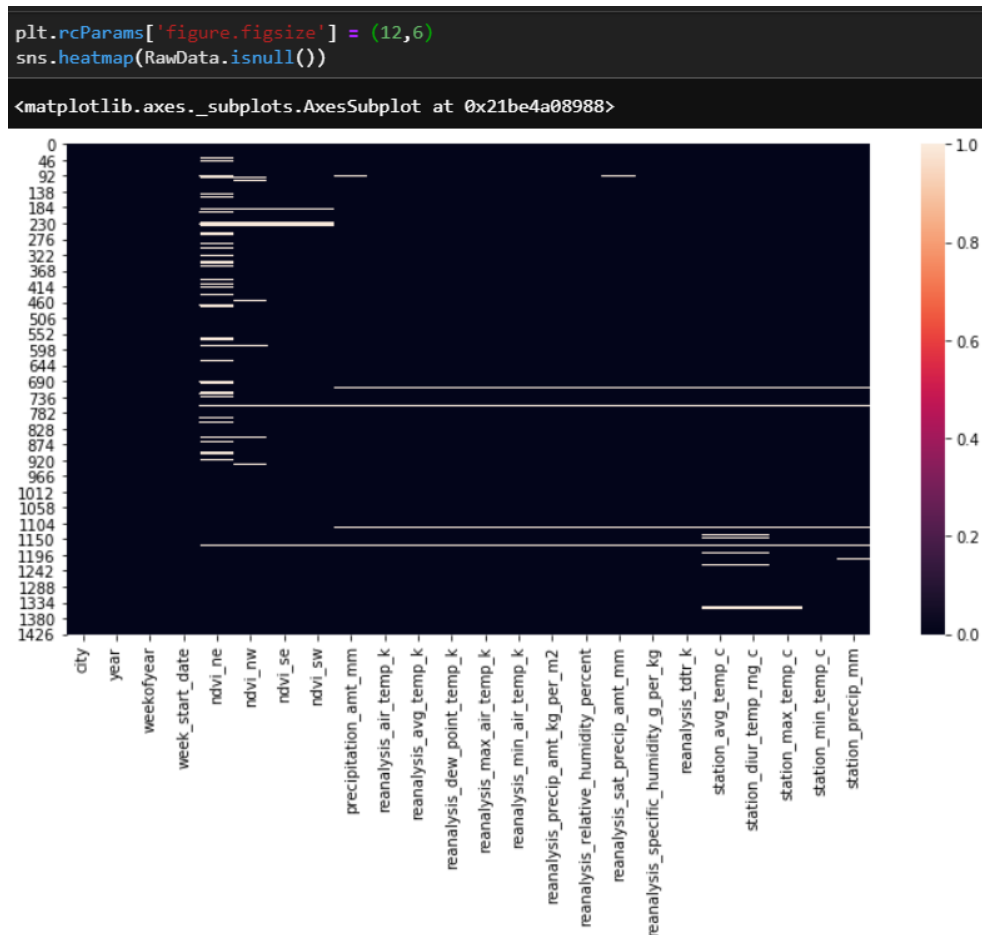


Fig:10

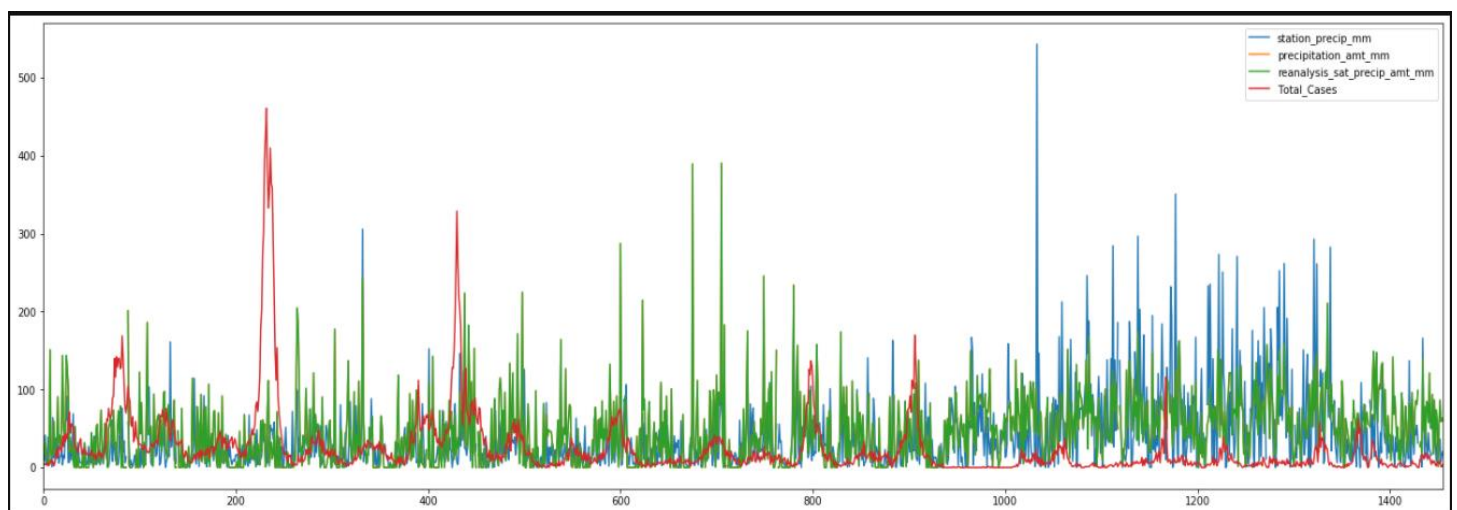


Fig:11

The timeseries plot below shows how the relationship between precipitation and Dengue Cases. when there is high rain fall there seems to be a spike in the Dengue cases. dln Fig:11, we can see that there is a positive correlation between Dengue Cases and precipitation, with higher precipitation causes more cases.

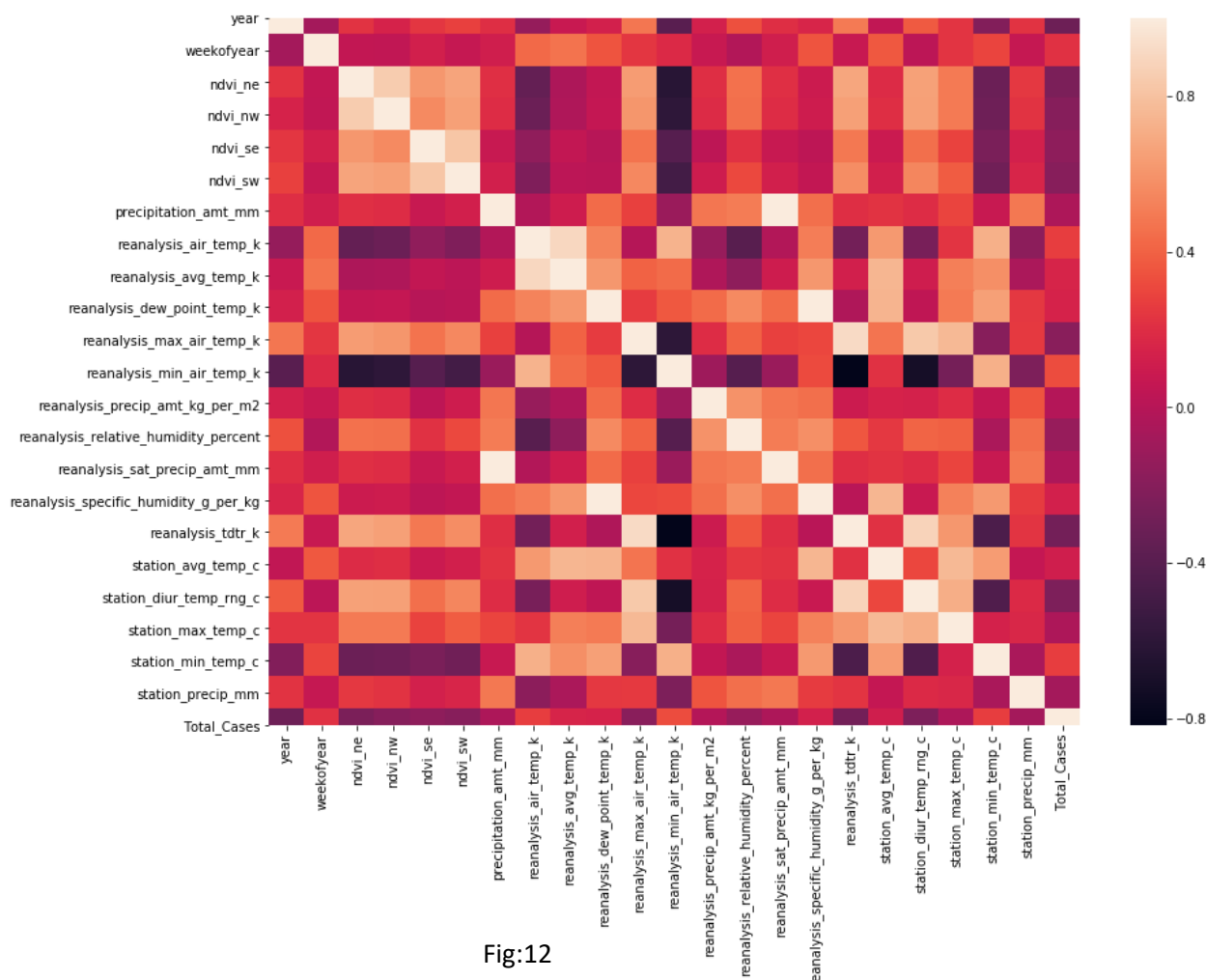


Fig:12

Correlation ranges from -1 to +1. Values closer to zero implies there's no linear relation between the two factors. The closer the values are to the 1 the relationship is the more emphatically correlated; and -1 mean stronger negative correlation. The following independent variables ('ndvi_nw', 'ndvi_ne', 'ndvi_se', 'ndvi_sw') seems to have strong negative correlation with the Total_Cases. And 'weekofyear', 'precipitation_amt_mm', 'station_max_temp_c', 'station_min_temp_c', 'station_avg_temp_c', 'station_diur_temp_rng_c') are some of the variables with positive correlation with 'Total_Cases'. In spite I have explained Principal Component Analysis, I did not perform it in my analysis since point requires normalization first. When running our introductory investigation, we observed weak correlations among the reaction variable. Implementing PCA would find the strongest metrological factors and determine the best linear combinations among them.

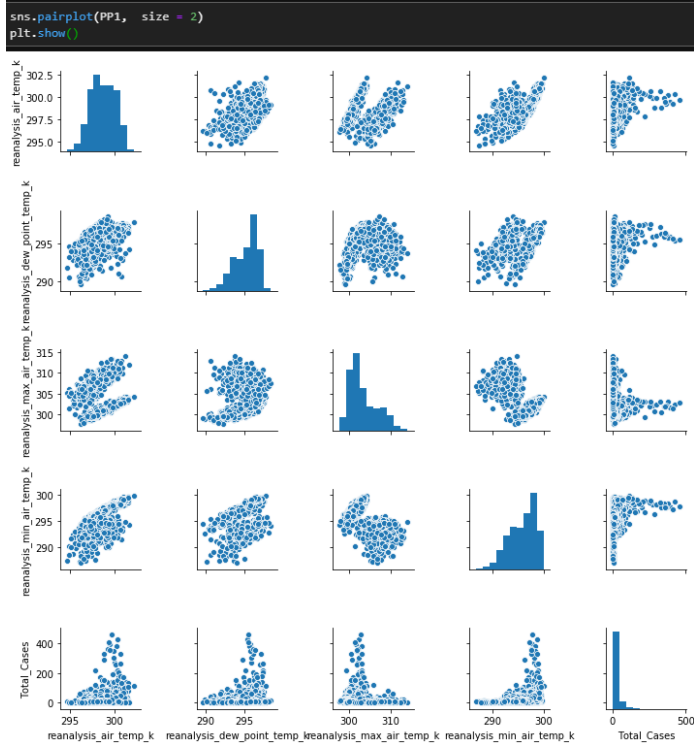


Fig:13

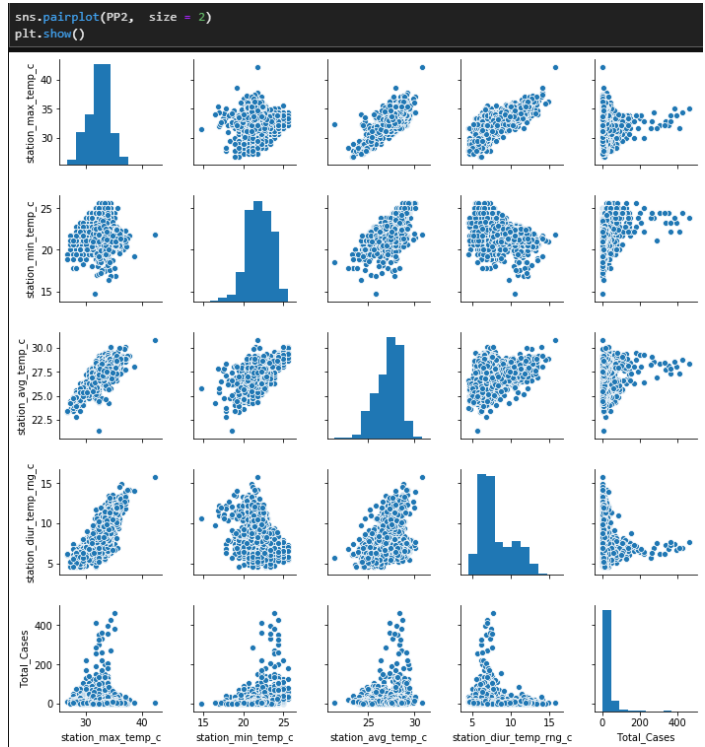


Fig:14

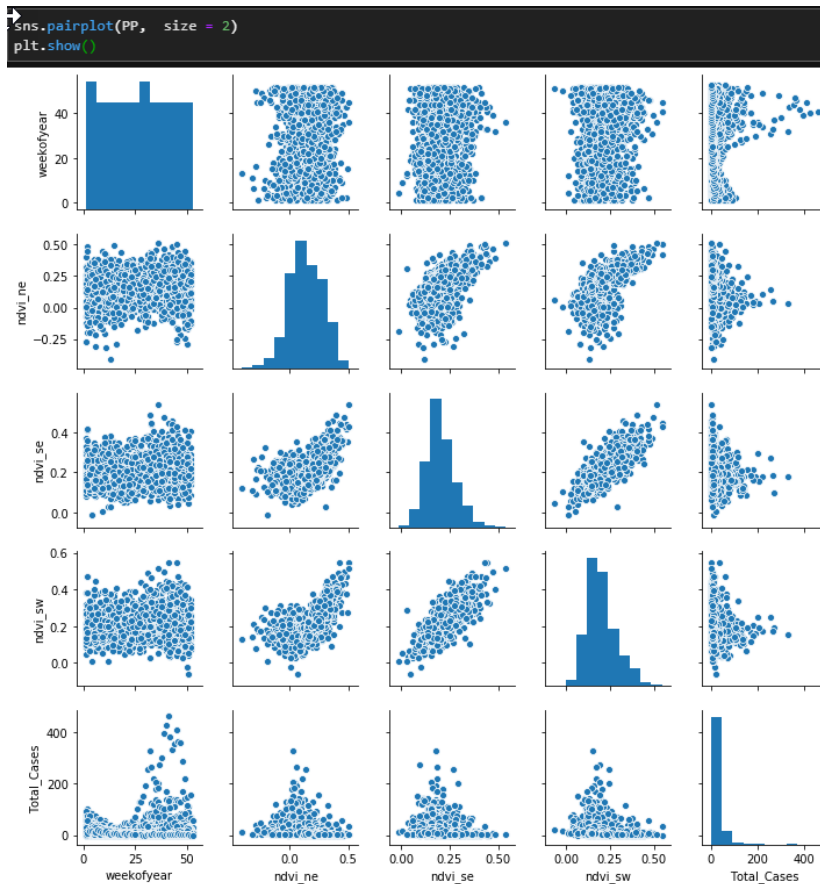


Fig:15

Fig:13, 14 & 15, are the Pairplot visualizations from seaborn library, showing the relationship between the independent parameters and Total dengue cases in San Juan and Iquitos. We can see how there some variables have positive linear relation and some negative linear relation.

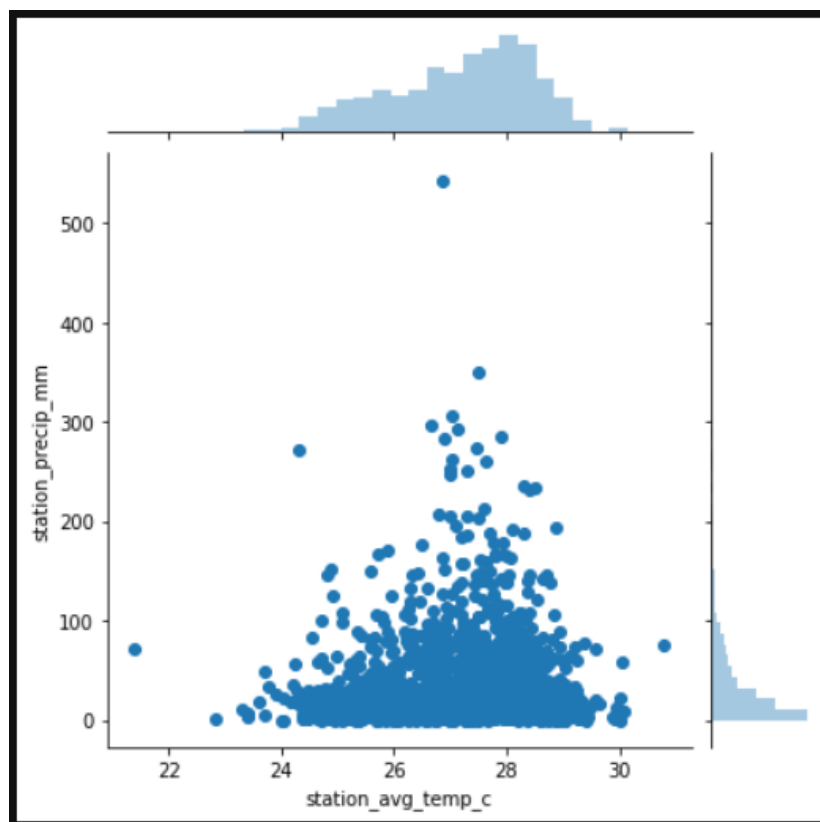


Fig:16

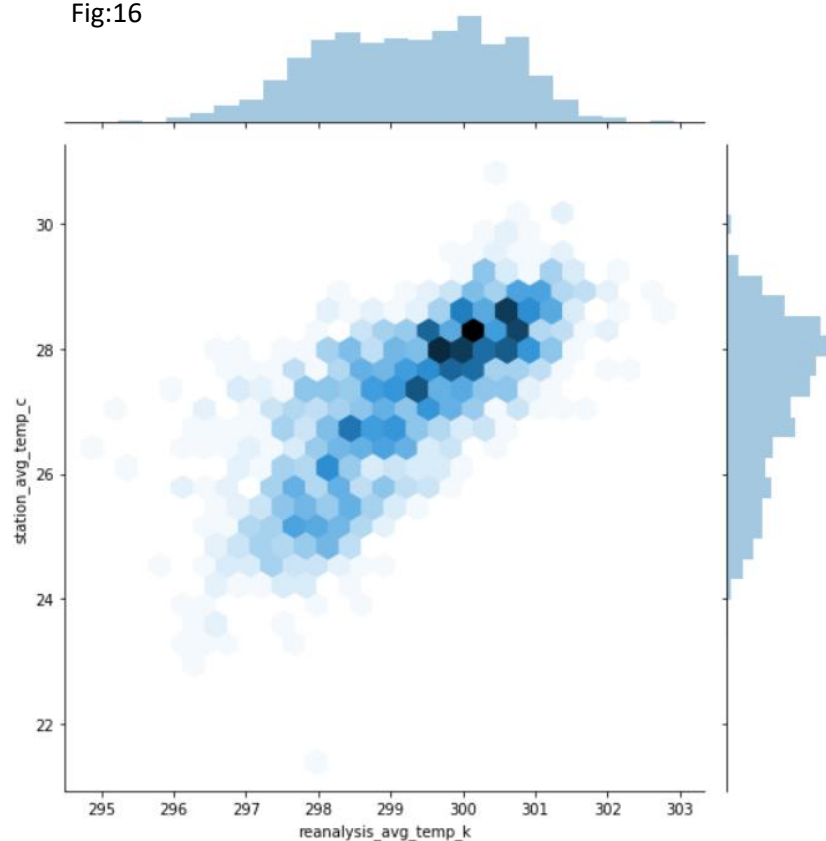


Fig:17

From the Exploratory Data Analysis, we can conclude that with higher Precipitation, temperature and lower NDVI values there is an increase in Dengue Cases. From the high number of cases between San Juan & Iquitos, our literature works propose that high temperatures do in truth influence the number of cases between both cities. Upon seeing the reviewing all the plots above, we will make a few presumptions. There has been a spike of cases around Q3 of 1994, with the most spiked number of cases detailed around the 460 mark. Around this time, there are to some high counts of Cases between the cities. Between 2000-2010, there has been a surge in cases between the two cities, with higher share of the numbers. We see an increase in total cases between 2005-2006 and 2007-2008. This correlates with the fact that the *Aedes aegypti* mosquito favor freshwater regions and uses saline waters for breeding site, as specified within the literature review; San Juan is found in coastal Puerto Rico and Iquitos in Peru somewhat closer to the ocean.

OLS Regression Results			
Dep. Variable:	Total_Cases	R-squared:	0.132
Model:	OLS	Adj. R-squared:	0.121
Method:	Least Squares	F-statistic:	11.50
Date:	Sun, 28 Mar 2021	Prob (F-statistic):	4.08e-33
Time:	20:18:55	Log-Likelihood:	-7458.7
No. Observations:	1456	AIC:	1.496e+04
Df Residuals:	1436	BIC:	1.506e+04
Df Model:	19		
Covariance Type:	nonrobust		

Fig:18

	coef	std err	t	P> t	[0.025	0.975]
Intercept	2313.4804	3660.602	0.632	0.527	-4867.221	9494.182
ndvi_ne	6.7917	15.221	0.446	0.656	-23.066	36.649
ndvi_nw	33.7881	16.922	1.997	0.046	0.594	66.982
ndvi_se	-9.9660	26.389	-0.378	0.706	-61.730	41.798
ndvi_sw	9.5346	25.350	0.376	0.707	-40.192	59.261
precipitation_amt_mm	-0.0124	0.016	-0.761	0.447	-0.044	0.020
reanalysis_air_temp_k	2.0187	14.868	0.136	0.892	-27.147	31.185
reanalysis_avg_temp_k	-12.9778	6.974	-1.861	0.063	-26.659	0.703
reanalysis_dew_point_temp_k	1.6143	16.247	0.099	0.921	-30.256	33.485
reanalysis_max_air_temp_k	1.1243	1.566	0.718	0.473	-1.947	4.195
reanalysis_min_air_temp_k	0.4816	2.155	0.224	0.823	-3.745	4.709
reanalysis_precip_amt_kg_per_m2	0.0299	0.034	0.871	0.384	-0.037	0.097
reanalysis_relative_humidity_percent	-3.0872	2.931	-1.053	0.292	-8.837	2.662
reanalysis_sat_precip_amt_mm	-0.0124	0.016	-0.761	0.447	-0.044	0.020
reanalysis_specific_humidity_g_per_kg	13.1563	12.663	1.039	0.299	-11.683	37.996
reanalysis_tdtr_k	-1.9908	2.056	-0.968	0.333	-6.024	2.043
station_avg_temp_c	-0.1361	2.519	-0.054	0.957	-5.077	4.804
station_diur_temp_rng_c	-2.0058	1.406	-1.426	0.154	-4.765	0.753
station_max_temp_c	2.8840	1.428	2.020	0.044	0.083	5.685
station_min_temp_c	-0.2240	1.682	-0.133	0.894	-3.524	3.076
station_precip_mm	0.0070	0.028	0.255	0.799	-0.047	0.061
Omnibus:	1550.835	Durbin-Watson:	0.112			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	94542.334			
Skew:	5.263	Prob(JB):	0.00			
Kurtosis:	41.048	Cond. No.	6.35e+16			

Fig:19

R-squared value 0.132 tells us how much variation is there in the metrological parameters explained by the model. So, 0.132 R-square means that the model explains 13.2% of variation within the dataset. Some of the parameters in the dataset has a p-value of less than the significance level (which is usually 0.05) and those parameters fits the data well, e.g. Avg Temp.

Adjusted R-squared can provide a more precise view of that correlation by also considering how many independent variables are added to a particular model. Adding more independent variables or predictors to a regression model tends to increase the R-squared value, which tempts makers of the model to add even more. This is called overfitting and can return an unwarranted high R-squared value. Adjusted R-squared is used to determine how reliable the correlation is and how much is determined by the addition of independent variables.

Our Adjusted R value is 0.121 that mean most of our independent variables are not significant in a regression model.

Data Modeling & Validation

From the hypotheses made approximately the information within the beginning steps (relationship, p-values, plots), deploys strategies and algorithm to test the data.

Test our dataset with numerous ML models to predict the Total Dengue Cases

1st Multilinear Regression

2nd Random Forest

3rd Support Vector Machine

4th Polynomial Regression non-linear approach

Each Model will be approved once I run train_test_split from the python Library

We will pick the best performed algorithm with the most accuracy and the least Mean Square Error (MSE)

MSE calculation will come towards the end of the project, after all tests have been made

We will submit our predictions on the "<https://www.drivendata.org/competitions/44>". The website's scoring metric is based on Mean Absolute Error; used to calculate the amount of error in the predictions and averages all the absolute errors.

$$MAE = \frac{1}{n} \sum_{i=1}^n |f_i - y_i|$$

Multiple Linear Regression with LASSO

We must clean the dataset and prepare for Multiple Linear Regression, after we prepare the data frame into preparing & testing splits (70% & 30%, separately) using the Python's function from Sklearn. Utilizing all conceivable climate features as the free factors.

```
x = MLR[['weekofyear', 'reanalysis_air_temp_k', 'reanalysis_dew_point_temp_k',
'reanalysis_max_air_temp_k', 'reanalysis_min_air_temp_k', 'reanalysis_relative_humidity_percent',
'reanalysis_sat_precip_amt_mm', 'reanalysis_specific_humidity_g_per_kg', 'station_max_temp_c',
'station_min_temp_c', ]]
```

```
y = MLR['Total_Cases'].values
```

From Sklearn I have imported the Linear Regression and fit the Trained data set with Linear Regression.

Once the Training data set portion is fitted with the MLR, I ran the score on the regression and I go only 0.1494595, which means my data is overfitted.

After I learned on the LASSO regression and how this can overcome this problem, imported LASSO regression from Sklearn fitted the it on the training dataset and the score this time was 0.03805 which was worse. And the mean_absolute_error was 20.989 and the mean_squared_error was 1670.82. Plotted a graph on the $(y_{\text{test}} - y_{\text{pred}})$ values to see curve. It overcome the weaknesses of many models' inflexibility to include a myriad of predictors and the challenging interpretability of some "black box" machine learning models.

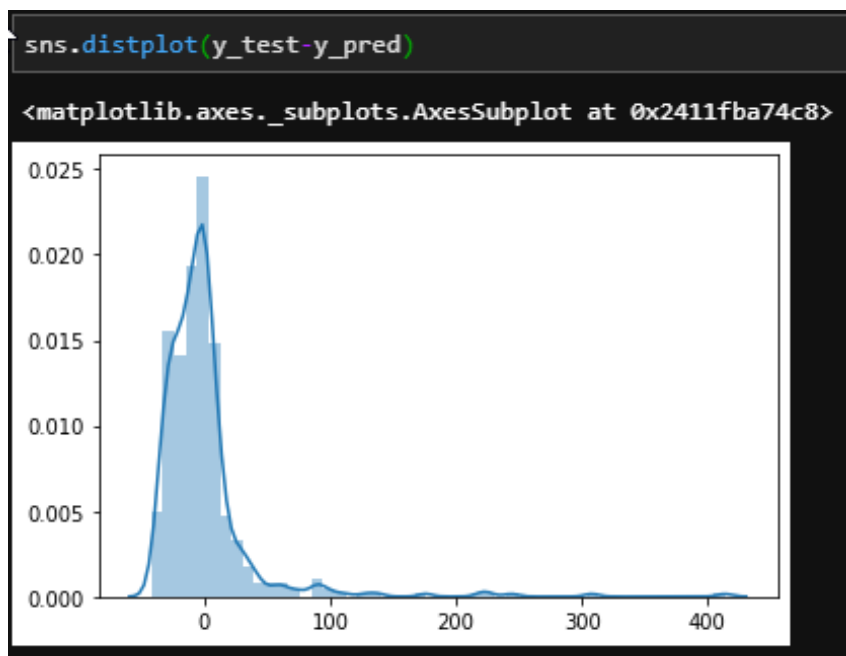


Fig:20

Random Forest

Random Forest takes variables at random and creates multiple decision trees at a time and makes it look like a forest. It simultaneously develops multiple trees in combination and averages the error to bring out the best possible results. The forest then chooses the classification of having the most 'votes' classification outcome) or the average of all trees in the forest (for numeric classification outcome). Random Forest can reduce the variance resulted from having on decision tree as the random forest algorithm considers the outcomes from many decision trees. Just like regression imported the Random Forest regressor library from Sklearn, using `n_estimators = 150`. `N_estimator` from Sklearn taking in a value which is the number of trees a person want to build before taking the maximum voting or averages of predictions. Higher number of trees give you better performance but makes the code slower.

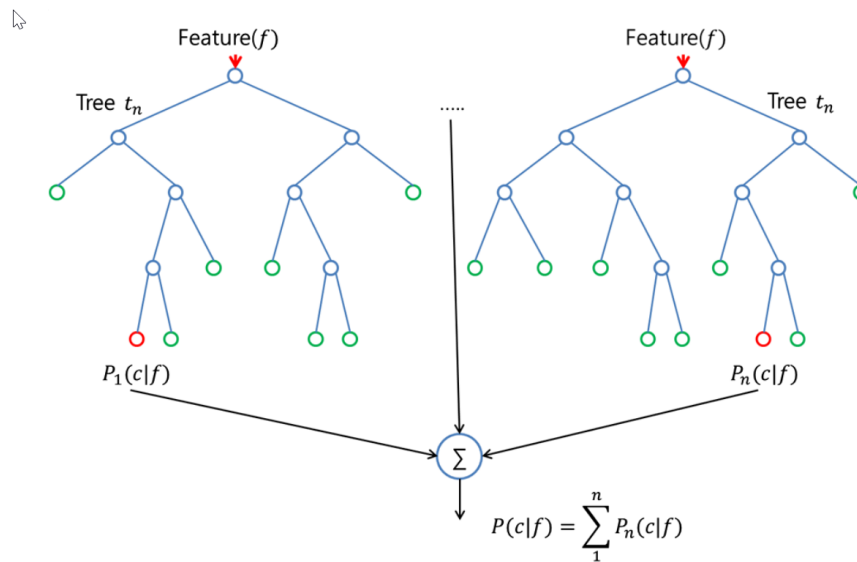


Fig:21

When I ran the regressor.score on the train dataset it gave me a very good score of 0.8696 (87%). Shows that it was able to build a strong training set model and the data is probably not over fitted.

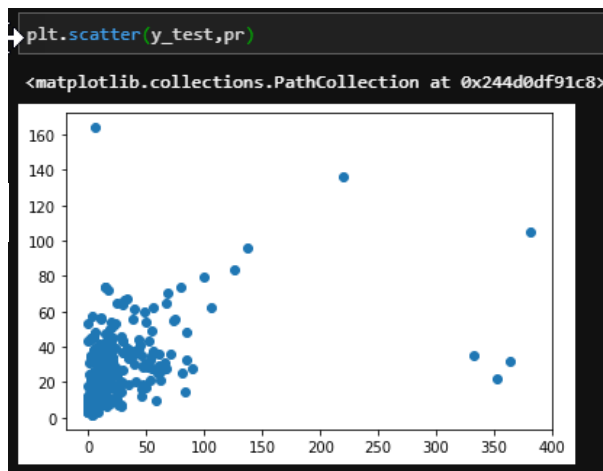


Fig:22

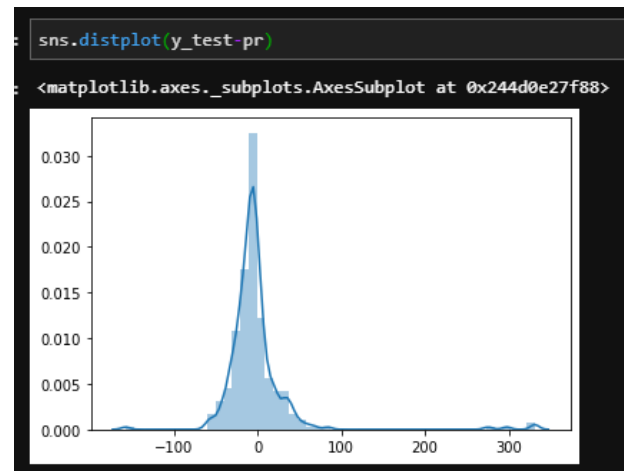


Fig:23

When we plot the Y_{test} vs prediction it gives somewhat a linear line and bell-shaped curve with a mean_absolute_error of 20.79. Since the numbers were strong in RF, I took an extra step and split the dataset per city and ran separately on each city. For San Juan the mean_absolute_error was 23.69 but Iquitos I got really low number of 5.65 which is amazing to see but not sure why RF performed better for Iquitos compared to San Juan.

Support Vector Machine

Support Vector Machine is non-parametric it won't train the network. The SVM algorithm tries to plot all the data in an n-dimensional hyper plane and applies the same logic on the test set, based on the reference created by the training set. To perform the classification, we then need to find the hyperplane that differentiates the two classes by the maximum margin. Imported the support vector machine SVM from

sklearn and set gamma = 10 and fitted the model to the split training data set. The score on the test set was only 0.055 but on the training dataset was almost perfect 0.99.

```
from sklearn.svm import SVC
```

```
SVM = SVC(gamma=10)
```

```
SVM.fit(x_train, y_train)
```

```
SVM.score(x_test, y_test)
0.0547945205479452

SVM.score(x_train, y_train)
0.9939862542955327
```

```
from sklearn import metrics
metrics.mean_absolute_error(y_test, y_pred)
20.74732451025732
```

The mean absolute error (MSE) was 20.747

Polynomial Regression (Non-linear)

From the last models we observed the MAE's were over 20 and the score was below 6%, so and looking at all the plots in the exploratory Data Analysis the curves were not all perfectly linear as well. Even though Random Forest performed better compared to MLR and SVM, this time I thought of taking a different approach for my last ML model which is a non-linear regression model (polynomial regression) method.

Imported polynomial function with degree value 3, which will generate all the features of degree 3 or lower.

Features: $x_1^0 x_2^3 + x_1^1 x_2^2 + x_1^2 x_2^1 + x_1^3 x_2^0$
 $+ x_1^2 + x_2^2 + x_1 x_2$
 $+ x_1 + x_2$
 $+ x_1^0 x_2^0$

```
from sklearn.preprocessing import PolynomialFeatures
poly = PolynomialFeatures(degree=2)
x_poly = poly.fit_transform(x_train)
poly.fit(x_train, y_train)

PolynomialFeatures(degree=2, include_bias=True, interaction_only=False,
order='C')
```

I have assigned x_poly as the new matrix that will take the new features from the fitted model. And then I imported the Linear regression and ran it using X-poly and y_train and predicted the dengue cases. The score on this model is 0.25 and MSE on this model is 19.37.

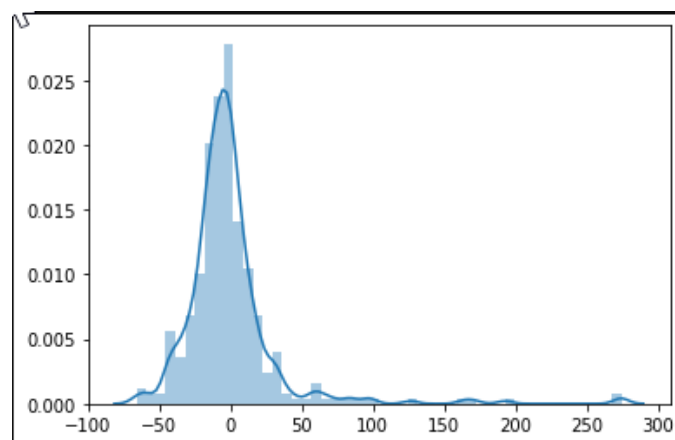


Fig:24

Results & Conclusion

The following table below shows the summary of the findings from running different models on the dataset.

Technique	Model	MAE	MSE	R Squared	np.sqrt(metrics.MSE)
MLR	LinearRegression linear_model.Lasso	20.988	1670.82	0.137	40.875
RF	RandomForestRegressor (n_estimators=150)	20.79	1807.34	0.186	51.79
SVM	SVM = SVC(gamma=10)	20.747	1762.17	-0.198	46.16
Polynomial Regression	PolynomialFeatures (degree=2)	19.37	1233.10	0.235	37.908

The evaluation of using R or Python outlines that R is most effective instrument that creates statistical analysis, information representation & control and demonstrate advancement & assessment easier and quicker, whereas Python needs numerous built-in information investigation highlights which makes statistical analysis and demonstrate improvement calm harder in Python. But since I have taken R during my course at Ryerson already and learned a bit on R, I wanted to use Python for my project as I have never used this platform. I spent most of time in 2020 WFH and learning Python from different sources like Udemy etc. Python is the most excellent when it comes to create prescient models for applications and generation server where tall preparing speed and ease of integration are anticipated. But as a multipurpose language, Python is the foremost fabulous when it comes to make prescient models for applications and era server where tall planning speed and ease of integration are expected.

Initially I thought Multilinear regression was going to outperform all the other models but between Support vector machine which can run both linear and non-linear and Random Forest techniques, all the techniques produced very close Mean Squared Error value. The 4th technique Polynomial non-linear Regression outperformed rest of the techniques and were able to deliver better MAE from the other models. The 3rd Random Forest regressor shows. Since competition on the www.drivendata.org/competitions/44/ allows multiple submission and I will start by submitting 19.37 as MSE using Polynomial Regression. We could spend more time on refining the in-depth analysis, to find the best independent variables to use for prediction. But for my analysis I have found the Normalized difference vegetation index (NDVI) with Precipitation and temperature around 28°C had stronger correlation compared to other variables. So if the local disease prevention team in place starts to prepare before the monsoon season arrives, that would help reduce and control dengue cases.

Reference

- (J. E Cogan, 23, June 2020) <https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue#.YCBhJOTGLml.gmail>
- Bhatt, S., et al., *The global distribution and burden of dengue*. *Nature*, 2013. 496(7446): p. 504–507.
- Brady, O.J., et al., *Refining the global spatial limits of dengue virus transmission by evidence-based consensus*. *PLOS Neglected Tropical Diseases*, 2012. 6(8): p. e1760.
- Lindsay P. Campbell, Caylor Luther, David Moo-Llanes, Janine M. Ramsey, Rogelio Danis-Lozano and A. Townsend Peterson Published:05 April 2015.
<https://doi.org/10.1098/rstb.2014.0135>
- G. Mincham 2019 <https://www.cambridge.org/core/journals/epidemiology-and-infection/article/development-of-a-mechanistic-dengue-simulation-model-for-guangzhou/7AA81F08BA5C6F9FA74F9F50E09B9388>
- Luis Villar, M.D., Gustavo Horacio Dayan, M.D., José Luis Arredondo-García, M.D., Doris Maribel Rivera, M.D., Rivaldo Cunha, M.D., Carmen Deseda, ...*N Engl J Med* <https://www.nejm.org/doi/10.1056/NEJMoa1411037>
- Nathans, L. L., Oswald, F. L. and Nimon, K. (2012) 'Interpreting multiple linear regression: A guidebook of variable importance', *Practical Assessment, Research & Evaluation*, 17(9).
- Dr.Preethi Subramanian, 2019 ..*Periodicals of Engineering and Natural Sciences*
<http://pen.ius.edu.ba/index.php/pen/article/viewFile/442/328>
- Schneider, A., Hommel, G. and Blettner, M. (2010) 'Linear regression analysis: part 14 of a series on evaluation of scientific publications', *Deutsches Ärzteblatt International*, 107(44), p. 776.
- <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0226841>
- <https://medium.com/python-in-plain-english/how-i-solved-the-dengue-fever-competition-question-using-adaboostingregressor-a9e82cb25f95>
- Torres,J.R., Orduna,T.A., Piña-Pozas,M., Vázquez-Vega,D. and Sarti,E. *Epidemiological Characteristics of Dengue Disease in Latin America and in the Caribbean: A Systematic Review of the Literature*. *Journal of Tropical Medicine*, Vol. 2017, March, 2017. [Online serial]. Available at: <https://doi.org/10.1155/2017/8045435>. [Accessed Dec. 22, 2018].
- *Peru and Bolivia: Dengue outbreak - DREF operation n° MDR46001* 21 September 2011. *The International Federation of Red Cross and Red Crescent Societies*, 2011. [Online]. Available at: https://reliefweb.int/sites/reliefweb.int/files/resources/Full_Report_2391.pdf. [Accessed Dec. 5, 2018].

- *Peru and Bolivia: Dengue outbreak - DREF operation n° MDR46001 18 February 2011. The International Federation of Red Cross and Red Crescent Societies, 2011. [Online]. Available at: https://reliefweb.int/sites/reliefweb.int/files/resources/1507AAFA08C7F39E8525783B00764030-Full_Report.pdf. [Accessed Dec. 5, 2018].*
- Sougata,D., Acebedo,C.M.L. and Chua,M.C.H. An ensemble prediction approach to weekly Dengue cases forecasting based on climatic and terrain conditions. *J. Health. Soc. Sci.*, Vol. 2, no. 3, pp. 257-272, 2017.
- Laureano-Rosario, A., Duncan, A., Mendez-Lazaro, P., Garcia-Rejon, J., Gomez-Carro, S.,Farfan-Ale, J., . . . Muller-Karger, F. (2018, May). Application of arti_cial neural networks for dengue fever outbreak predictions in the northwest coast of yucatan, mexico and san juan, puerto rico. *Tropical Medicine and Infectious Disease*, 3 (1), 5.doi: 10.3390/tropicalmed3010005
- Stoddard, S. T., Wearing, H. J., Reiner, R. C., Morrison, A. C., Astete, H., Vilcarromero, S., . . . et al. (2014). Long-term and seasonal dynamics of dengue in iquitos, peru. *PLoS Neglected Tropical Diseases*, 8 (7). doi: 10.1371/journal.pntd.0003003
- Guzman MG and Harris E (2015) Dengue. *The Lancet* 385, 453–465.
- Cheng Q et al. (2016) Climate and the timing of imported cases as determinants of the dengue outbreak in Guangzhou, 2014: evidence from a mathematical model. *PLoS Neglected Tropical Diseases* 10, p. e0004417
- <https://github.com/mithilgotarne/DengAI-Predicting-Disease-Spread>
- Cover T, Hart P. Nearest neighbor pattern classification. *IEEE Trans Inf Theory*. 1967;13(1):21–7.
- <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6163306/>