**I. INTRODUCTION AND DATA EXPLORATION**

Our Project is on predicting the **“total\_cases”** of Dengue based on **“DengAI: Predicting Disease Spread” dataset** which is an active driven data competitions dataset. Here we have a set of climate information (precipitation, temperature, vegetation) from the two cities: San Juan (sj) and Iquitos (iq) with total cases of dengue count by city, year and week of the year. We aim at making a complete analysis on the DengAI dataset to find the total number of Dengue affected cases in the given two cities with respect to set of climate variables as mentioned above.

The ***DengAI: Predicting Disease Spread*** dataset is taken from an active DrivenData Competition and the link of which is given below:  
<https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/>

**BACKGROUND READING:**-

Dengue is known to be transmitted seasonally, especially in the rainy season when the creation of stagnant pools allows for more breeding grounds for disease-bearing mosquitoes. Every few years, however, these spikes burst into epidemics, which are still more or less sporadic. Moreover, climate change is expected to permit the entry of disease agents into new territories that are growing increasingly temperate, thus infecting populations as yet unfamiliar with the disease. The unpredictability and scale of dengue causes it to remain an issue of wide concern.

The relationship between dengue transmission and the environment is particularly tenuous, with a complex network of variables and interactions. The study **Climate and Dengue Transmission: Evidence and Implications** (<http://ehp.niehs.nih.gov/wp-content/uploads/121/11-12/ehp.1306556.pdf>) reviews the following results:

1. Increases in **temperature** have been found to be favorible to the faster reproduction of mosquitoes and a shorter extrinsic incubation period -- that is, it takes less time for an infected mosquito to spread the virus to another. But high temperatures have also been found to limit mosquitoes' blood feeding and adult development, among others.

2. Mosquitoes exposed to smaller **diurnal temperature ranges (DTR's)** -- which are the difference between the min and max temperatures within the day -- were more likely to become infected and survive longer lifespans. (Notably, smaller DTR's are another expected offshoot of climate change -- <https://www.ipcc.ch/ipccreports/tar/wg2/index.php?idp=422>). However, longer DTR's may also shorten extrinsic incubation periods.

3. The mosquitoes' choice of breeding locations depends on the amount of sunlight in an area and the **presence of trees**, which, by virtue of the shade they provide, control water temperature and evaporation.

4. Higher **precipitation** is generally associated with higher aedes populations, though too much rainfall could watch out habitats and breeding sites. However, drier seasons could also indirectly lead to population growth, as families tend to increase water storage at these times, leading to an abundance of breeding sites. Precipitation still exhibits less of an impact than the nutritional resources present in the area, over which young mosquitoes have to compete for survival.

5. The survival of eggs depends on the rate of **evaporation** as caused by the delicate interaction between temperature and precipitation.

6. While population sizes have been correlated with the number of dengue cases, the relationship between them has not been shown to the statistically significant, for a number of plausible reasons, such as herd immunity, or the fact that the temperature may support population growth but is not conducive to viral transmission.

7. Good predictive variables for total number of cases are **minimum, maximum, and mean temperature** as well as **relative humidity and wind velocity**, while precipitation is a valid indicator for their timing. The direction of these associations depend hugely on the local climate. **El Nino Southern-Oscillation**, or ENSO, indices and **sea surface temperatures** have also been shown to be strong predictors.

Climate change, by causing higher temperatures, would affect the relationships above, and thus, disease transmission.

**II. DATASET DESCRIPTION**

City and date indicators.

1. city: 'sj' for San Juan and 'iq' for Iquitos

2. year:

3. weekofyear:

4. week\_start\_date: the start date of each week, as given in dd-mm-yyyy format

NOAA's GHCN daily climate data weather station measurements: NOAA is the U.S.' National Oceanic and Atmospheric Association, and the GHCN (or the Global Historical Climatology Network) is their database integrating climate reports across land and sea stations around the world. All temperature values here are in degrees Celsius.

5. station\_max\_temp\_c: Maximum temperature

6. station\_min\_temp\_c: Minimum temperature

7. station\_avg\_temp\_c: Average temperature,

8. station\_precip\_mm: Total precipitation

9. station\_diur\_temp\_rng\_c: Diurnal temperature range

PERSIANN satellite precipitation measurements (0.25x0.25 degree scale): PERSIANN, on the other hand, is the Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks, as developed by UC Irvine's Center for Hydrometeorology and Remote Sensing (CHRS). As its name might suggest, the system uses neural networks to estimate rainfall rate at a given geographic location, so it may be interesting to see how the values here differ from those given by NOAA's different measurements.

10. precipitation\_amt\_mm: Total precipitation

NOAA's NCEP Climate Forecast System Reanalysis measurements (0.5x0.5 degree scale): NCEP are NOAA's National Centers for Environmental Prediction. At its simplest, the CFS, or the Climate Forecast System, is a model of the interaction among the earth's lands, oceans, and temperature based on hourly data. All temperature values here are in Kelvin.

11. reanalysis\_sat\_precip\_amt\_mm: Total precipitation (expressed in millimeters)

12. reanalysis\_dew\_point\_temp\_k: Mean dew point temperature (The temperature at which air would have to cool in order to reach saturation)

13. reanalysis\_air\_temp\_k: Mean air temperature

14. reanalysis\_relative\_humidity\_percent: Mean relative humidity (The amount of water vapor in the air, expressed as the percentage of the amount needed for the air to be saturated at the same temperature)

15. reanalysis\_specific\_humidity\_g\_per\_kg: Mean specific humidity (The amount of water vapor in the air, with respect to the total mass of air + water vapor)

16. reanalysis\_precip\_amt\_kg\_per\_m2: Total precipitation (expressed as kg per meters squared)

17. reanalysis\_max\_air\_temp\_k: Maximum air temperature

18. reanalysis\_min\_air\_temp\_k: Minimum air temperature

19. reanalysis\_avg\_temp\_k: Average air temperature

20. 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: The NDVI is an indicator that measures the presence of green vegetation on a given pixel of land surfaces. This is done by searching for the distinct wavelengths of sunlight absorbed (visible) and reflected (near-infrared) by plants for photosynthesis. The values here range between 0 and 0.8, with NDVI's between 0.3 and 0.8 indicating the presence of vegetation, and those below 0.3 bare soils. I'm expecting that higher vegetation would at least be correlated with higher numbers of cases.

21. ndvi\_se – Pixel southeast of city centroid

22. ndvi\_sw – Pixel southwest of city centroid

23. ndvi\_ne – Pixel northeast of city centroid

24. ndvi\_nw – Pixel northwest of city centroid

And lastly the target variable.

25. total\_cases: the number of cases within the timeframe for a given city

Total number of attributes = 24 (predictors)  
Total number of instances = 1456

Indicators mentioned in “background reading” that aren't covered in the above are *rate of evaporation, ENSO indices, and sea surface temperatures*.

**III. PREPROCESSING**

**Splitting the data into two datasets based on city:**

**San Juan City:**

****

**Iquitos City:**

****

San Juan is in coastal area where as Iquitos is located much inside to the coastal region. So we assume that the occurrence of dengue is different in both regions and we split the data into two datasets based on city attribute.

**Missing Values**: Read in the data and *summary* command in R gives insight into minimum and maximum values of the attribute and missing values (or NA’s) in the attribute.

> summary(train)

city year weekofyear week\_start\_date ndvi\_ne

iq:520 Min. :1990 Min. : 1.00 01-01-2001: 2 Min. :-0.4062

sj:936 1st Qu.:1997 1st Qu.:13.75 01-01-2002: 2 1st Qu.: 0.0449

Median :2002 Median :26.50 01-01-2003: 2 Median : 0.1288

Mean :2001 Mean :26.50 01-01-2004: 2 Mean : 0.1421

3rd Qu.:2005 3rd Qu.:39.25 01-01-2005: 2 3rd Qu.: 0.2485

Max. :2010 Max. :53.00 01-01-2006: 2 Max. : 0.5084

(Other) :1444 **NA's :193**

ndvi\_nw ndvi\_se ndvi\_sw precipitation\_amt\_mm

Min. :-0.45610 Min. :-0.01553 Min. :-0.06346 Min. : 0.00

1st Qu.: 0.04922 1st Qu.: 0.15509 1st Qu.: 0.14421 1st Qu.: 9.80

Median : 0.12143 Median : 0.19605 Median : 0.18945 Median : 38.34

Mean : 0.13055 Mean : 0.20378 Mean : 0.20231 Mean : 45.76

3rd Qu.: 0.21660 3rd Qu.: 0.24885 3rd Qu.: 0.24698 3rd Qu.: 70.23

Max. : 0.45443 Max. : 0.53831 Max. : 0.54602 Max. :390.60

**NA's :52 NA's :22 NA's :22 NA's :13**

reanalysis\_air\_temp\_k reanalysis\_avg\_temp\_k reanalysis\_dew\_point\_temp\_k

Min. :294.6 Min. :294.9 Min. :289.6

1st Qu.:297.7 1st Qu.:298.3 1st Qu.:294.1

Median :298.6 Median :299.3 Median :295.6

Mean :298.7 Mean :299.2 Mean :295.2

3rd Qu.:299.8 3rd Qu.:300.2 3rd Qu.:296.5

Max. :302.2 Max. :302.9 Max. :298.4

**NA's :10 NA's :10 NA's :10**

reanalysis\_max\_air\_temp\_k reanalysis\_min\_air\_temp\_k reanalysis\_precip\_amt\_kg\_per\_m2

Min. :297.8 Min. :286.9 Min. : 0.00

1st Qu.:301.0 1st Qu.:293.9 1st Qu.: 13.05

Median :302.4 Median :296.2 Median : 27.25

Mean :303.4 Mean :295.7 Mean : 40.15

3rd Qu.:305.5 3rd Qu.:297.9 3rd Qu.: 52.20

Max. :314.0 Max. :299.9 Max. :570.50

**NA's :10 NA's :10 NA's :10**

reanalysis\_relative\_humidity\_percent reanalysis\_sat\_precip\_amt\_mm

Min. :57.79 Min. : 0.00

1st Qu.:77.18 1st Qu.: 9.80

Median :80.30 Median : 38.34

Mean :82.16 Mean : 45.76

3rd Qu.:86.36 3rd Qu.: 70.23

Max. :98.61 Max. :390.60

**NA's :10 NA's :13**

reanalysis\_specific\_humidity\_g\_per\_kg reanalysis\_tdtr\_k station\_avg\_temp\_c

Min. :11.72 Min. : 1.357 Min. :21.40

1st Qu.:15.56 1st Qu.: 2.329 1st Qu.:26.30

Median :17.09 Median : 2.857 Median :27.41

Mean :16.75 Mean : 4.904 Mean :27.19

3rd Qu.:17.98 3rd Qu.: 7.625 3rd Qu.:28.16

Max. :20.46 Max. :16.029 Max. :30.80

**NA's :10 NA's :10 NA's :43**

station\_diur\_temp\_rng\_c station\_max\_temp\_c station\_min\_temp\_c station\_precip\_mm

Min. : 4.529 Min. :26.70 Min. :14.7 Min. : 0.00

1st Qu.: 6.514 1st Qu.:31.10 1st Qu.:21.1 1st Qu.: 8.70

Median : 7.300 Median :32.80 Median :22.2 Median : 23.85

Mean : 8.059 Mean :32.45 Mean :22.1 Mean : 39.33

3rd Qu.: 9.567 3rd Qu.:33.90 3rd Qu.:23.3 3rd Qu.: 53.90

Max. :15.800 Max. :42.20 Max. :25.6 Max. :543.30

**NA's :43 NA's :20 NA's :14 NA's :22**

total\_cases

Min. : 0.00

1st Qu.: 5.00

Median : 12.00

Mean : 24.68

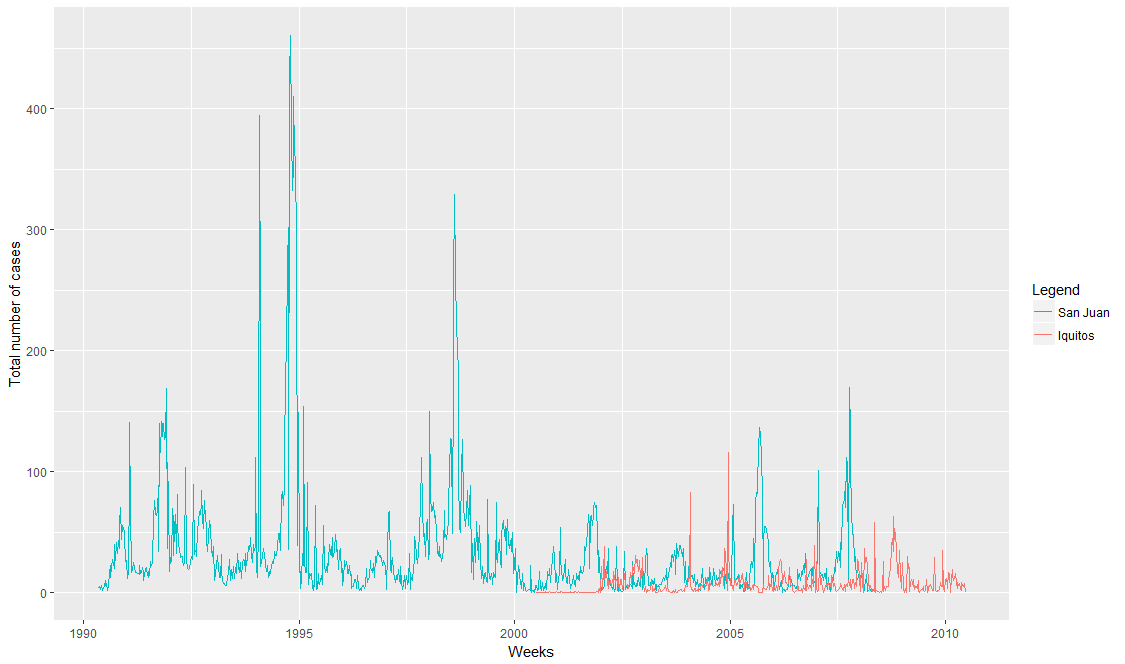
3rd Qu.: 28.00

Max. :461.00

Observation from summary: The count of NA's at the bottom of each variable indicates that there is an underlying pattern to the missing values for each data source. Nevertheless, **since these are variables that are known to follow seasonal trends, we can impute them by taking the most recent values (except for NDVI, because it has clusters of values that are all missing)**.

**We also converted all Kelvin temperatures to Celsius to maintain the consistency between different temperature measures.**

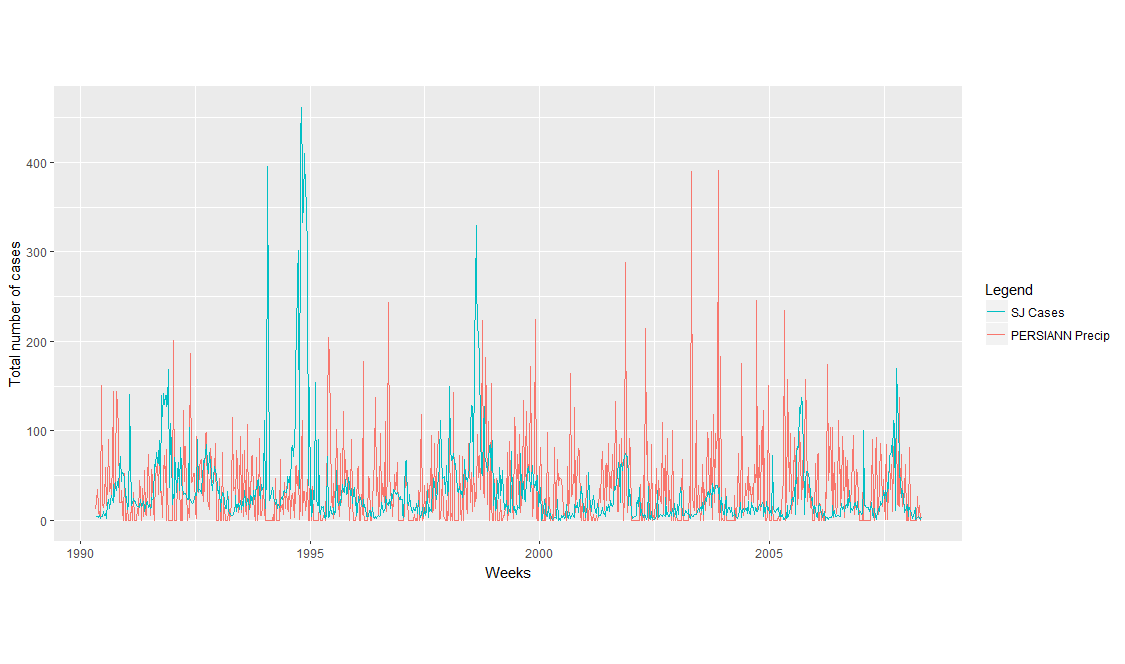
Total number of dengue cases vs Date



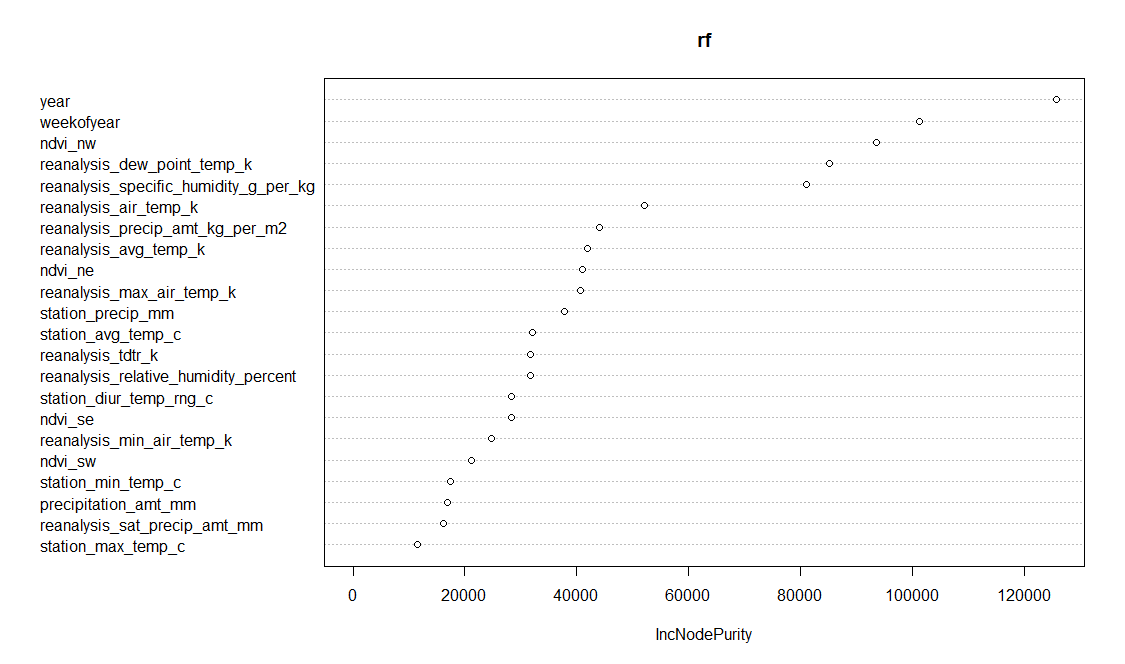
The above plot confirms the **strong seasonality of dengue transmission** and its punctuation by major outbreaks, as marked by sudden massive spikes in the plot. Their **spontaneity indicates that it might not be a good idea to predict these using time series analysis alone**. The difference in plot height between San Juan and Iquitos could be attributed to a big disparity in population size.

Correlation between number of cases and other climate variables

We see that none of the variables are significantly correlated with the total number of cases (although some of the temperature variables are correlated with each other, as expected). It may well be that our expected model is a combination of several of them, with no particular features having a strong impact on the outcome.

The plot exhibits phenomenon we may have come to expect from the literature -- the seasonality in precipitation somewhat mirrors that of dengue cases, but the lows in the dengue plot also fall on massive peaks in precipitation, which may be due to intense rainfall washing out mosquito breeding sites. Curiously, the two outbreaks between 1992 and 1995 were during periods with relatively low precipitation, suggesting the contribution of other variables.

We next plotted ranking of the variables in order of significance, by exploiting the algorithm used within random forests for deciding upon which variable to split.



Observations: weekofyear and year ranked as two of the most important.

We have deleted any features that are strongly correlated to others based on the corrplot.

We have removed the following attributes:

'city','precipitation\_amount\_mm','reanalysis\_avg\_temp\_k','reanalysis\_specific\_humidity\_g\_per\_kg',

\_diur\_temp\_rng\_c'

**IV. Feature Engineering:**

Feature engineering = transforming our features (or even come up with new ones) to make our machine learning algorithms more effective.

When you look at the total cases reported per week in the training files, can we say that those cases are the result of the meteorological conditions recorded in previous weeks? What would be the relationship between the reported cases for a given week and the data provided for the same week (if any)?

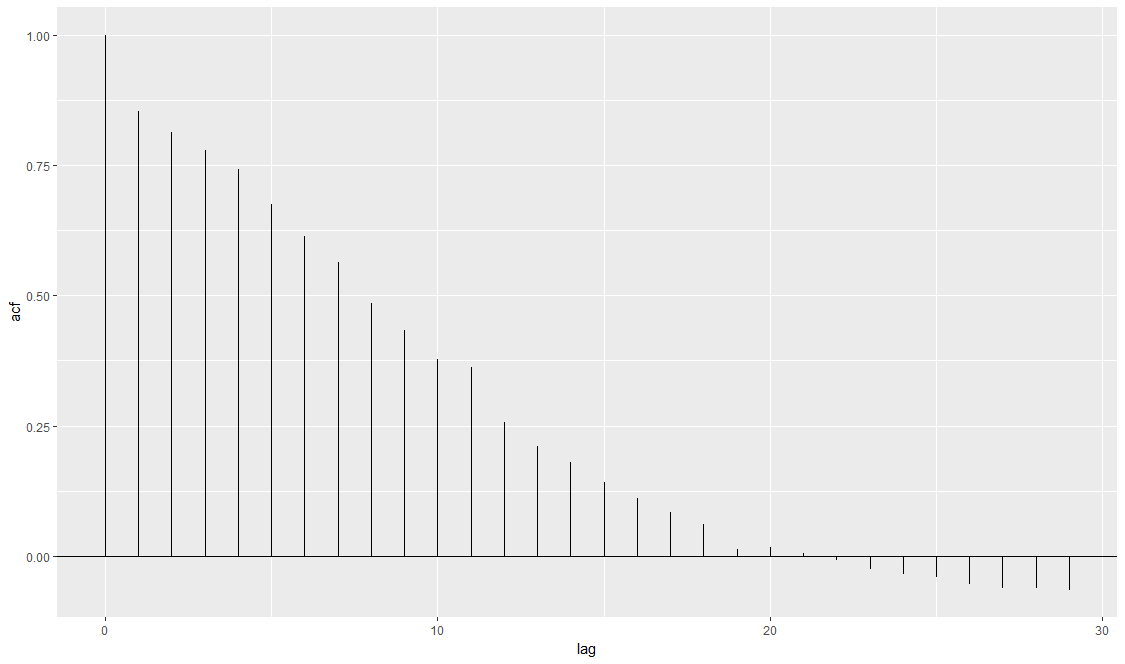
What I see here is that the infections reported for a specific week are not a direct cause-effect of the meteorological conditions recorded on that same week. This is mostly because the disease has an incubation period of 4-7 days, which means that the infection is related to the conditions in previous weeks.

When analyzing the data, would it be convenient to somehow shift the data back one week

For accounting the spikes that are large outbreaks we take into account the contagiousness of dengue. A possible way to account for this is to build a model that progressively predicts a new value while taking into account the previous prediction. By training on the dengue outbreaks and then using the predicted number of patients in the week before, we can start to model this time dependence that the current model misses.

Since we are essentially dealing with seasonal time series here, we thought it would be good to add variables on past time lags for each observation.

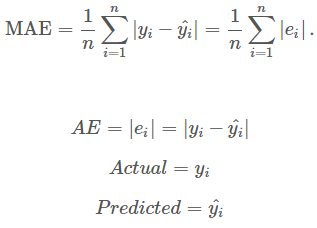
The autocorrelation plot below would confirm the correlation of the total number of cases at a given time with those of the past 15 weeks. We created five new features to show the total cases from the last five time lags for each observation.



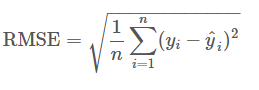
**V. Model Training and Validation**

**Performance Metrics Used: MAE, RMSE, R Squared and Predicted vs Actual plots**

**The mean absolute error (MAE) measures the closeness of forecasts or predictions to the actual outcomes. The mean absolute error is given by:**

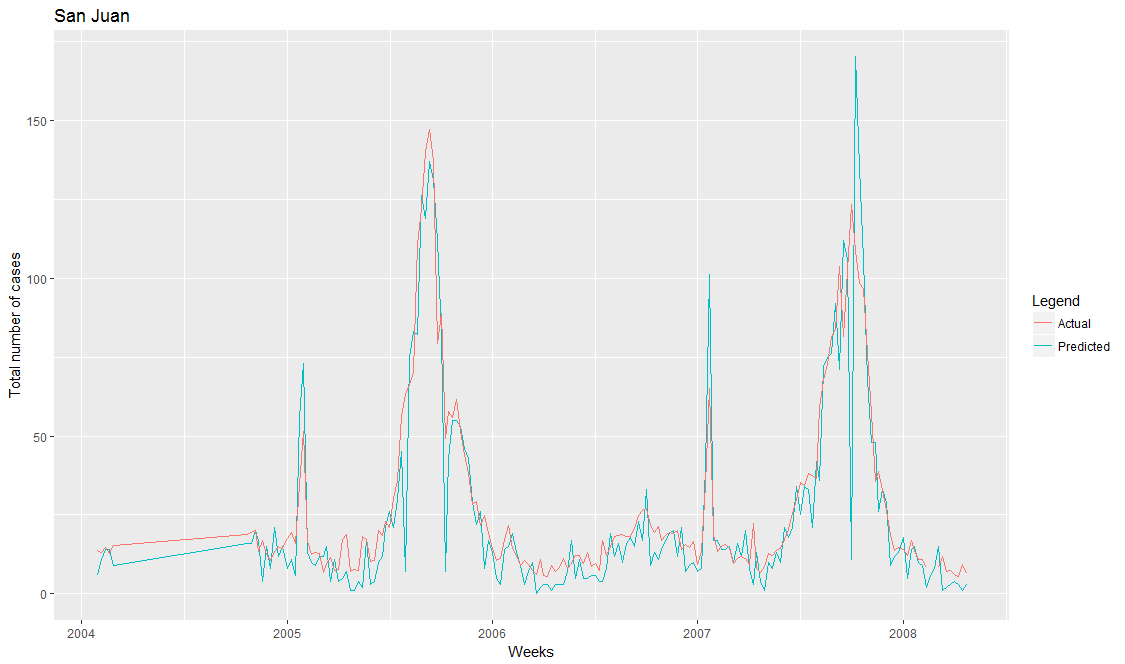


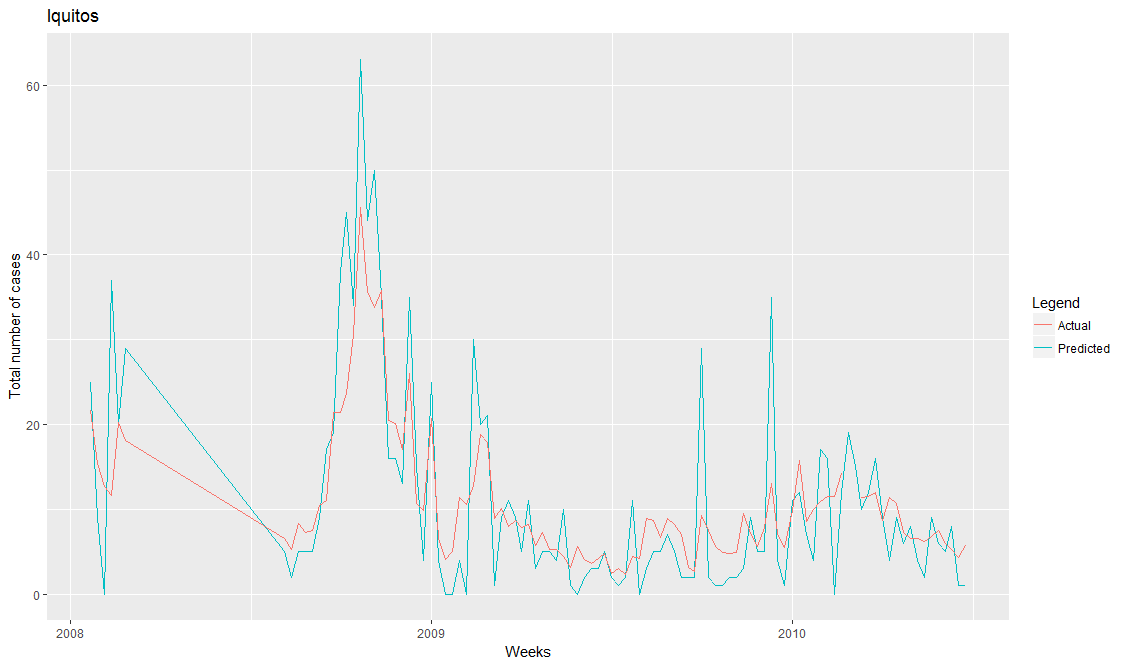
The square root of the mean/average of the square of all of the error. It is generally used as an error metric for numerical predictions.



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Classifier** | **ntress** | **mtries** | **maxdepth** | **MAE** | **RMSE** | **R2** |
| Random Forest | 1000 | 3 | 4 | Sj: 16.027  Iq: 4.605 | 30.4456  8.914 | 0.6937  0.2249 |
| Random Forest | 1000 | 4 | 4 | Sj: 15.258  Iq: 4.569 | 29.86  8.9308 | 0.705  0.222 |
| Random Forest | 1000 | 5 | 4 | Sj: 14.758  Iq: 4.537 | 29.459  8.938 | 0.713  0.220 |
| Random Forest | 1000 | 5 | 5 | Sj: 14.00  Iq: 2.286 | 28.66  8.88 | 0.728  0.230 |
| Random Forest | 1000 | 10 | 5 | Sj: 13.472  Iq: 4.546 | 28.73  9.11 | 0.722  0.190 |
| Random Forest | 1000 | 10 | 10 | Sj: 13.2106  Iq: 4.5516 | 28.66  9.101 | 0.728  0.192 |

**Predicted vs Actual cases plot:**





|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Classifier** | **Activation function** | **Hidden layers** | **MAE** | **RMSE** | **R2** |
| Deep Learning |  |  |  |  |  |
| Deep Learning |  |  |  |  |  |
| Deep Learning |  |  |  |  |  |
| Deep Learning |  |  |  |  |  |
| Deep Learning |  |  |  |  |  |
| Deep Learning |  |  |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Classifier** | **ntress** | **Stopping rounds** | **maxdepth** | **MAE** | **RMSE** | **R2** |
| Gradient Boosting |  |  |  |  |  |  |
| Gradient Boosting |  |  |  |  |  |  |
| Gradient Boosting |  |  |  |  |  |  |
| Gradient Boosting |  |  |  |  |  |  |
| Gradient Boosting |  |  |  |  |  |  |
| Gradient Boosting |  |  |  |  |  |  |