Features that Help Explain and Predict Dengue Outbreaks in San Juan (Puerto Rico) and Iquitos (Peru)

Final report prepared for the Data Analysis and Interpretation Specialization
November 2017

Introduction to the Research Question

The purpose of this study is to identify features that best explains and predicts new cases of Dengue disease. This is not a comprehensive study of scientific root causes that lead to the disease, but rather a study exploring historic observational data.

As a data science practitioner I have interest in using data to help solve people's problems. Half billion cases of this disease is diagnosed per year around the globe. This study can help in the understanding about the features that are most influential for an outbreak of the disease, and can be of service to help prevent more cases and even save peoples lives.

Methods

Sample

The data contains 1456 samples divided between the two cities that involve the research question. San Juan contains 936 observations and Iquitos 520. The samples are independent and represent the measurement of the different observed variables in a given week of the year. Entries for San Juan included weekly measures from April 30, 1990 to April 22, 2008, while entries about Iquitos included weekly measures from July 1, 2000 to June 25, 2010.

The data set is provided by the U.S. Centers for Disease Control and prevention, as well as the Department of Defense's Naval Medical Research Unit 6 and the Armed Forces Health Surveillance Center, in collaboration with the Peruvian government and U.S. universities. The data set is freely available at www.drivendata.com.

Measures

The response variable of this dataset is the total cases of Dengue disease during a week of a given year.

The explanatory variables that will be part of the analysis are: city, week of the year, year, maximum temperature, minimum temperature, average temperature, total precipitation, diurnal temperature range, mean dew point temperature, mean air temperature, mean relative humidity, minimum air temperature, maximum air temperature and average air temperature.

Analyses

The distribution of the predictor and response variables were evaluated by examining the frequency counts of categorical variables. Continuous numeric variables were examined by descriptive statistics such as mean, standard deviation, quantiles and histograms.
In terms of bivariate analyses, scatter plots, box plots, Pearson correlation and analysis of variance (ANOVA) were used to test bivariate associations between individual predictors and the response variable (total number of Dengue fever cases in a given week).

A linear regression model was developed to better understand the relationship of the variables that can help predict Dengue disease cases. Although the linear model captured the relationship and could be used to predict new cases, a random forest algorithm was also developed since it has higher predictive ability for new cases of Dengue disease.

Results

Descriptive Statistics

The table below shows some descriptive statistics for the variables in the dataset that will be considered in the analysis. The response variable “total_cases” has a mean of approx. 27 cases per week with a standard deviation of approx. 44 cases.

VARIABLE	COUNT	MEAN	STD DEV	MIN	MAX	NMISS
precipitation_amt_mm	1443	45,76	43,72	0,00	390,60	13
reanalysis_air_temp_k	1446	298,70	1,36	294,64	302,20	10
reanalysis_avg_temp_k	1446	299,23	1,26	294,89	302,93	10
reanalysis_dew_point_temp_k	1446	295,25	1,53	289,64	298,45	10
reanalysis_max_air_temp_k	1446	303,43	3,23	297,80	314,00	10
reanalysis_min_air_temp_k	1446	295,72	2,57	286,90	299,90	10
reanalysis_precip_amt_kg_per_m2	1446	40,15	43,43	0,00	570,50	10
reanalysis_relative_humidity_percent	1446	82,16	7,15	57,79	98,61	10
reanalysis_sat_precip_amt_mm	1443	45,76	43,72	0,00	390,60	10
reanalysis_specific_humidity_g_per_kg	1446	16,75	1,54	11,72	20,46	10
reanalysis_tdtr_k	1446	4,90	3,55	1,36	16,03	10
station_avg_temp_c	1413	27,19	1,29	21,40	30,80	43
station_diur_temp_rng_c	1413	8,06	2,13	4,53	15,80	43
station_max_temp_c	1436	32,45	1,96	26,70	42,20	20
station_min_temp_c	1442	22,10	1,57	14,70	25,60	14
station_precip_mm	1434	39,33	47,46	0,00	543,30	22
total_cases	1456	24,68	43,60	0,00	461,00	0

Please notice that the last column shows the number of values missing in the data set. Since there are not many missing values for each variable, they were treated. The method of choice was “mean imputation", where the mean of each variable grouped by "city and month" was imputed in case of a missing value. This method was chosen having into consideration the condition of each location and the seasonality, since there are months were the number of cases are higher (fig. 1). This leads to assume that the conditions are not the same across the entire year. The ANOVA results in the "Bivariate Analysis" section shows more indicators that support this decision.

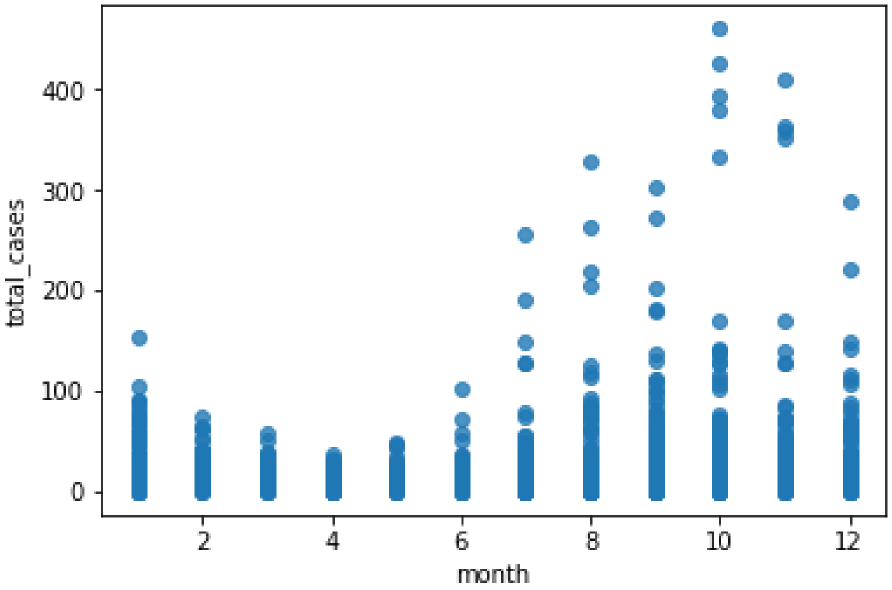

Fig. 1 Total Dengue Cases Distributed per Months

Bivariate Analysis

The variable that has the highest Pearson correlation coefficient with the response variable is "reanalysis_min_air_temp_k" which indicates the minimum air temperature in degrees Kelvin. The Pearson correlation coefficient is 0.3257 having a p-value of 2.3810e-37. The plot in the fig.2 shows that there is a presence of outliers in the response variable "total_cases", however their influence in the regression line is small since most of data points are concentrated below the 100th mark:

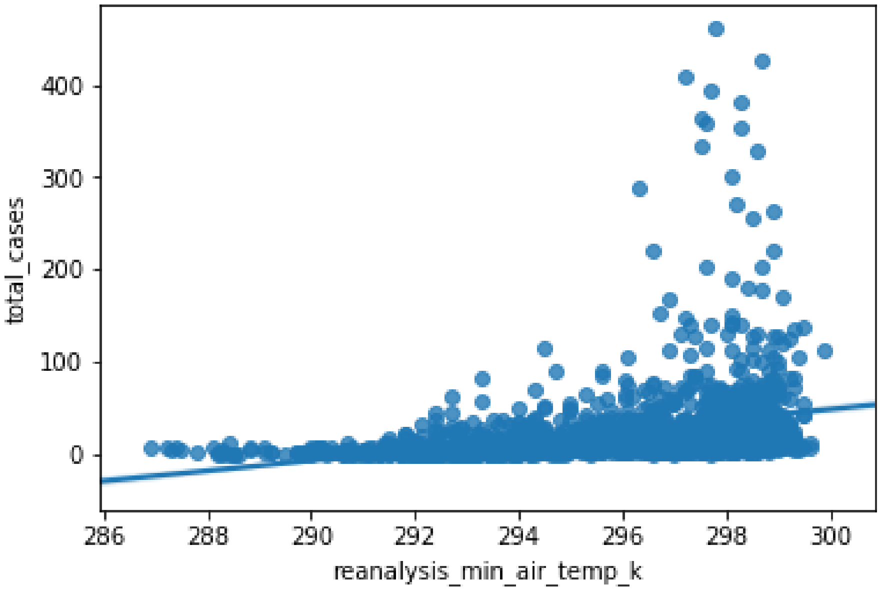

Fig. 1 Total Dengue Cases per Minimum Air Temperature

Decision was made to “trim” the dataset to keep the highest outliers out of the regression. The formula used to identify outliers was based on interquartile range, where:

*Upper Bound Outlier Cutoff = Q3 + (Q3 – Q1) * 1.5*

In the case of the response variable total_cases the cutoff for outliers would be 28 + (28 -5) * 1.5 = 62.5. Based on the fact that as air temperature arises, the conditions for the reproduction of the mosquito increases, decision was made to increase the cutoff value to 100 cases. This way the dataset can be trimmed of outliers, but still hold some influential observations that will be significant for the analysis.

Although the correlation is weak, we see a pattern in fig. 1: most cases of the disease happen when the minimum weekly temperature is between 296 and 300 degrees Kelvin. An analysis of variance (ANOVA) illustrates that this difference in temperature is also related to the time of the year. In tropical countries the warmest months are roughly between October through March. Fig. 3 illustrates this difference. The results for the ANOVA model showed an F-statistic of 13.81, p-value of 1.68e-25 and R^2 of 0.088.

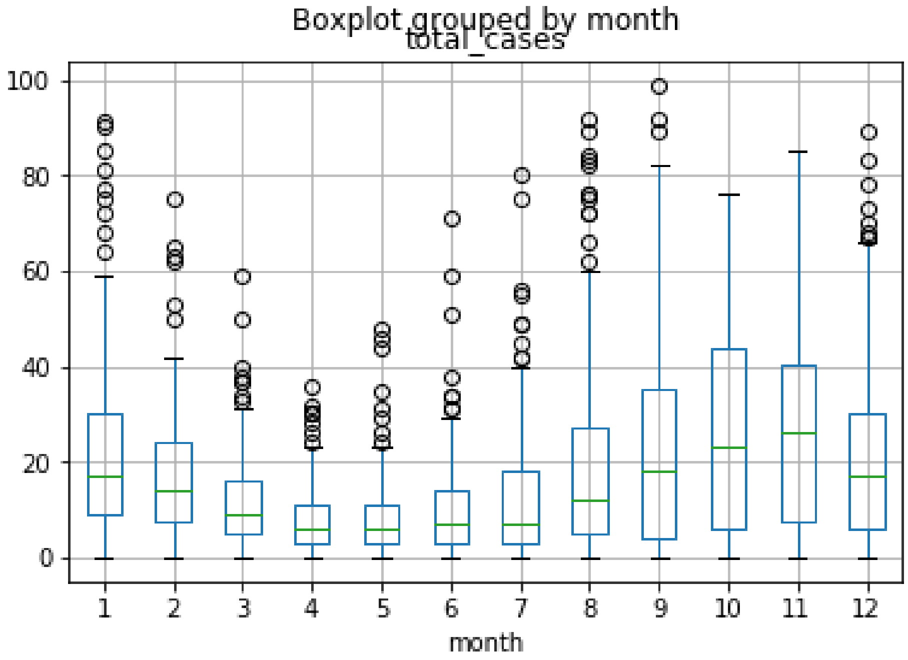

Fig. 3 Box Plot of Total Cases per Month

The ANOVA analysis also showed that there is a significant difference in the number of cases per city. San Juan presented significantly more weekly cases than Iquitos. This confirms the decision taken to group the variables by “city and month” when taking the mean for imputation discussed above. The ANOVA results were: F-statistic of 136.2, p-value of 3.87e-30 and R^2 of 0.085. Fig. 4 illustrates the difference:

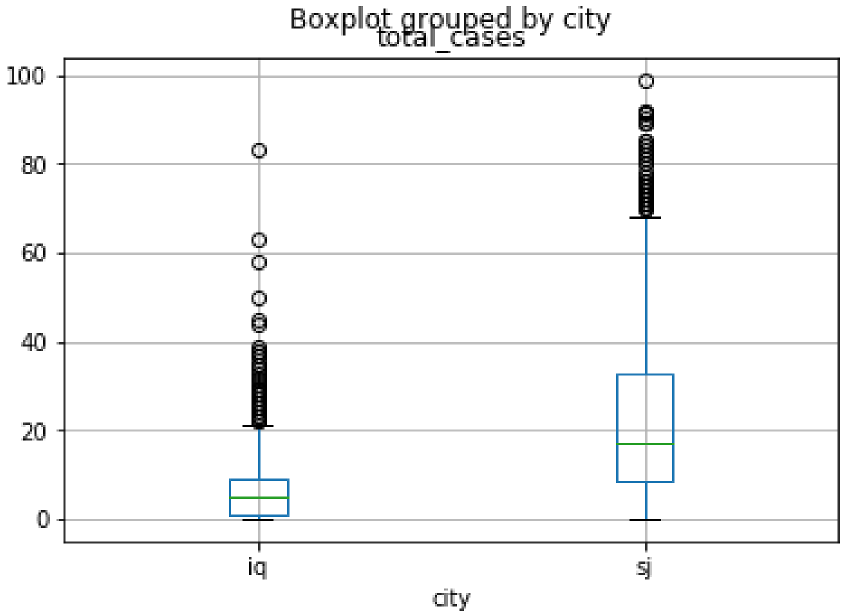

Fig. 3 Box Plot of Total Cases per City

Regression Analysis

Several linear models were tested for this analysis, including stepwise methods where variables are added/ removed automatically from the model by a script. The highest R² obtained was around 0.563. This indicates that the combination of the categorical variables *city, year, month* and continuous variables *precipitation_amt_mm, reanalysis_air_temp_k, reanalysis_dew_point_temp_k, reanalysis_max_air_temp_k, reanalysis_min_air_temp_k, reanalysis_relative_humidity_percent, reanalysis_tdtr_k, station_avg_temp_c* and *station_min_temp_c* account for roughly 56% of the variation in the total cases of Dengue fever in a given week. Please refer to the appendix for a extensive list of the model coefficients.

Most of the model's residuals fall within 2 standard deviation of the mean, which indicates an acceptable overall fit:

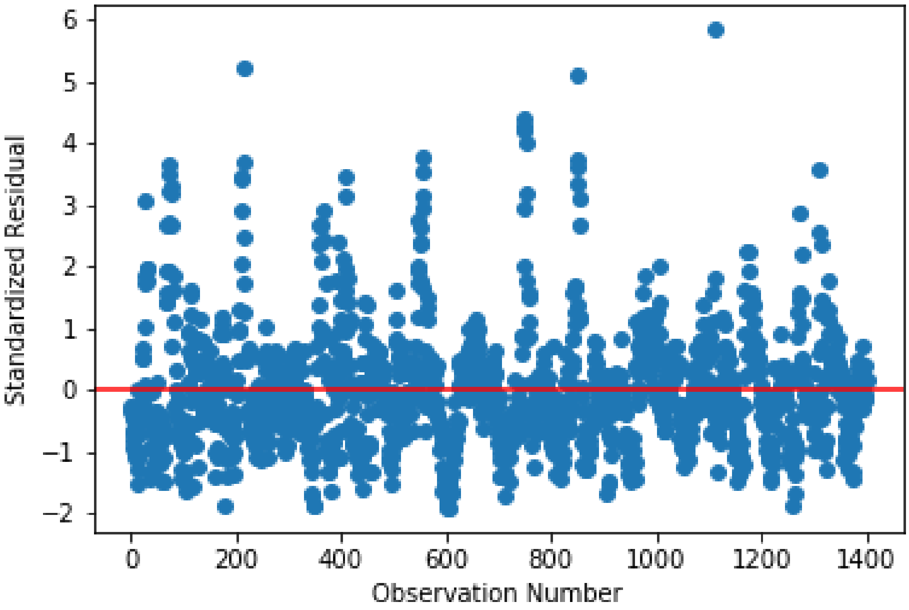

Fig. 4 Scatter plot of model's residuals

Random Forest Model

The previous linear model is good for analysis purposes and to understand which combination of environmental variables play an important role in new Dengue fever cases. To predict new cases, a random forest model was developed using all variables. This model result as a better predictor with an accuracy of 0.987 in the train data set and of 0.617 in the test data set. Fig. 5 shows the most important features in the random forest model:

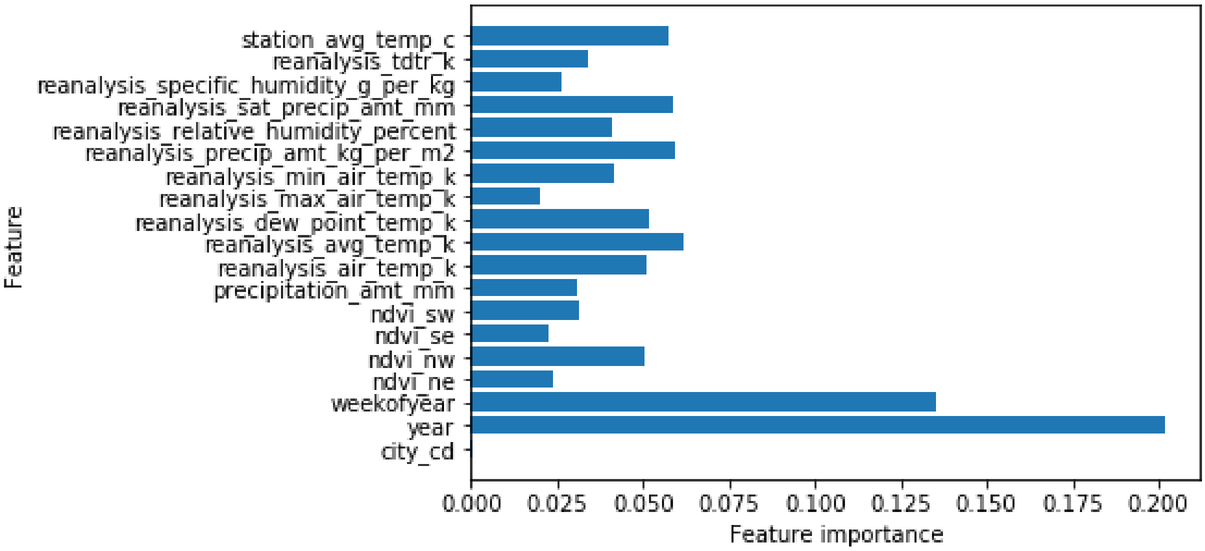

Fig. 5 Feature Importance in Random Forest Model

Conclusions/Limitations

This analysis attempted to show the steps involved in the exploratory data analysis and modeling of predicting the number of cases of Dengue fever in Iquitos and San Juan. The steps outlined in this document are the most relevant among all the analyses and data manipulation that were conducted to arrive at these results.

The best linear model had a combination of 12 variables to explain the relationship of the explanatory variables with the total cases of Dengue fever in the 2 cities of the observational data. The regression coefficients in the appendix show that the environmental variables that most influence in the increase of the response variable are in short: increase in average temperature, increase in precipitation amount, increase in air temperature and increase in humidity. These variables are in line with a thorough study published by the National Center for Biotechnology Information entitled: "*Effects of Weather Factors on Dengue Fever Incidence and Implications for interventions in Cambodia*"[1]. Knowing the weather conditions and time of the year that are most suitable for the reproduction of the mosquito can be beneficial in the effective combat of the disease.

Additionally, a Random Forest model was developed to best predict new dengue fever cases in a given week of the year. This can also be beneficial as means of monitoring a possible outbreak.

[1] Available on https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4784273/ accessed Nov. 2017

A quick note regarding the limitations of this study. First, in order to increase the predictive ability of the model, data should continue to be gathered so that the model can be fine tuned when its accuracy start to decrease. Second, it should be noted that during the course of the time, preventive measures have been taken by the government of these 2 cities, which led to less cases of the disease even in months were more cases were observed in previous years. This is really good news since it means that less people got infected. This fact may also explain why the linear regression model did not present a higher adjusted R².

Appendix

Linear Regression Model Output:

```
--------------------------------------------------------------------------------------
Variable                          coef   std err    t    P>|t|  [0.025   0.975]
--------------------------------------------------------------------------------------
Intercept                        258.3685 04.933  -1.261  0.208  660.388  143.651
C(month)[T.2]                     -4.5480  1.607  -2.830  0.005   -7.700   -1.396
C(month)[T.3]                    -10.9980  1.634  -6.730  0.000  -14.204   -7.792
C(month)[T.4]                    -15.7018  1.561 -10.056  0.000  -18.765  -12.639
C(month)[T.5]                    -16.1750  1.695  -9.542  0.000  -19.500  -12.850
C(month)[T.6]                    -14.1547  1.717  -8.242  0.000  -17.523  -10.786
C(month)[T.7]                     -9.6102  1.662  -5.784  0.000  -12.870   -6.351
C(month)[T.8]                     -0.2435  1.820  -0.134  0.894   -3.814    3.327
C(month)[T.9]                      3.0960  1.843   1.680  0.093   -0.518    6.711
C(month)[T.10]                     5.4658  1.799   3.038  0.002    1.936    8.995
C(month)[T.11]                     3.9785  1.748   2.276  0.023    0.549    7.408
C(month)[T.12]                     0.2049  1.660   0.123  0.902   -3.051    3.461
C(year)[T.1991]                   12.6505  2.915   4.339  0.000    6.931   18.370
C(year)[T.1992]                   21.9733  2.785   7.891  0.000   16.510   27.436
C(year)[T.1993]                   -0.3961  2.771  -0.143  0.886   -5.832    5.040
C(year)[T.1994]                   18.2796  3.075   5.945  0.000   12.247   24.312
C(year)[T.1995]                    0.5453  2.774   0.197  0.844   -4.897    5.987
C(year)[T.1996]                   -6.0451  2.772  -2.181  0.029  -11.483   -0.607
C(year)[T.1997]                    4.1748  2.776   1.504  0.133   -1.271    9.620
C(year)[T.1998]                   34.3718  2.962  11.605  0.000   28.562   40.182
C(year)[T.1999]                    8.8346  2.773   3.185  0.001    3.394   14.275
C(year)[T.2000]                  -12.2634  2.632  -4.660  0.000  -17.426   -7.101
C(year)[T.2001]                   -1.9819  2.531  -0.783  0.434   -6.948    2.984
C(year)[T.2002]                   -5.9842  2.534  -2.362  0.018  -10.955   -1.013
C(year)[T.2003]                   -7.6187  2.537  -3.003  0.003  -12.595   -2.642
C(year)[T.2004]                   -6.3387  2.544  -2.492  0.013  -11.329   -1.348
C(year)[T.2005]                   -1.4631  2.581  -0.567  0.571   -6.527    3.600
```

```
C(year)[T.2006]                   -7.8641  2.563  -3.068  0.002  -12.892   -2.836
C(year)[T.2007]                   -0.1226  2.577  -0.048  0.962   -5.179    4.933
C(year)[T.2008]                   -1.0777  2.750  -0.392  0.695   -6.472    4.317
C(year)[T.2009]                   -6.7946  2.906  -2.338  0.020  -12.496   -1.093
C(year)[T.2010]                    0.1678  3.543   0.047  0.962   -6.783    7.119
C(city)[T.sj]                      1.5753  5.218   0.302  0.763   -8.661   11.812
precipitation_amt_mm               0.0098  0.010   1.007  0.314   -0.009    0.029
reanalysis_air_temp_k              7.3813  4.839   1.525  0.127   -2.111   16.873
reanalysis_dew_point_temp_k       -3.3815  4.957  -0.682  0.495  -13.105    6.342
reanalysis_max_air_temp_k         -1.2978  0.475  -2.735  0.006   -2.229   -0.367
reanalysis_min_air_temp_k         -2.0833  0.596  -3.498  0.000   -3.252   -0.915
reanalysis_relative_humidity_perc  0.9455  1.090   0.868  0.386   -1.192    3.083
reanalysis_tdtr_k                 -1.3663  0.578  -2.363  0.018   -2.501   -0.232
station_avg_temp_c                 0.0279  0.674   0.041  0.967   -1.294    1.350
station_min_temp_c                 0.5546  0.466   1.189  0.235   -0.360    1.470
```