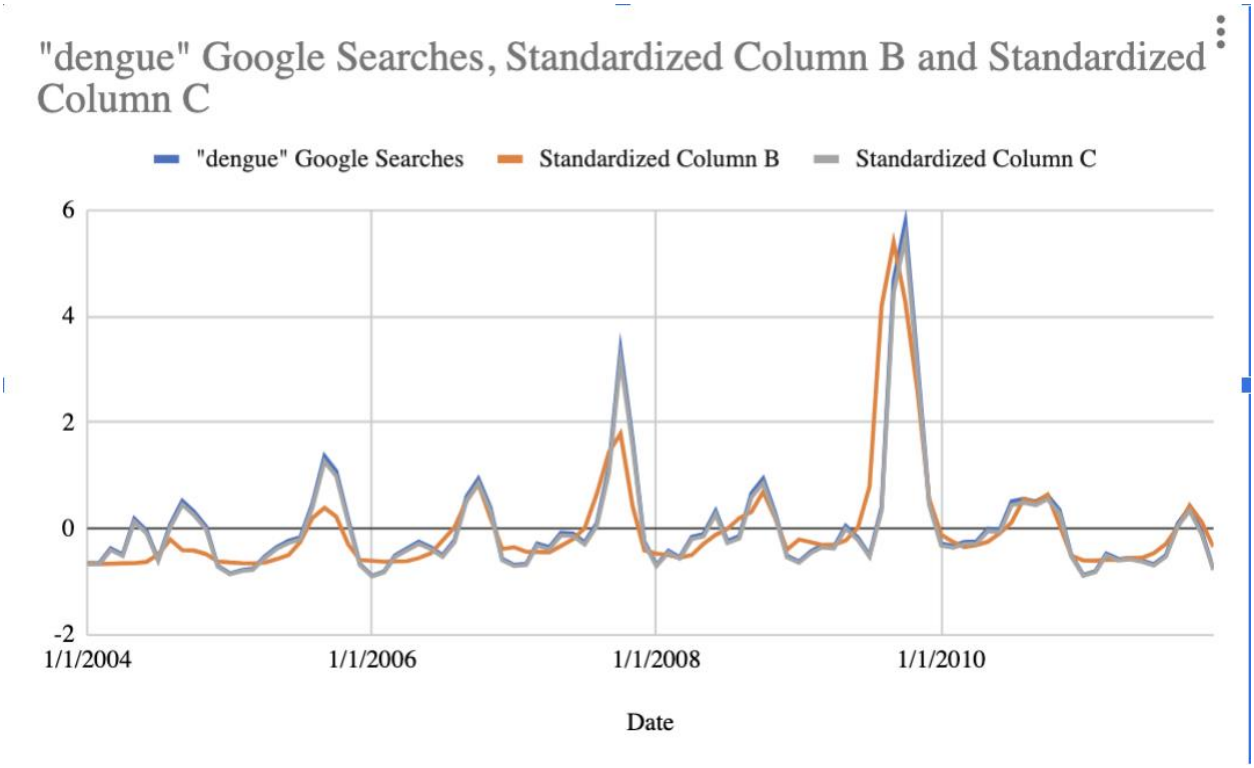
Dengue and Google Trends Modeling Exercise

Dengue fever is an infectious disease caused by a virus that is spread by mosquitoes. It affects millions of people in tropical environments around the world.

In this exercise, you'll be asked to construct a simple Excel version of the digital disease detection tool "Google Dengue Trends" – a Dengue analog to Google Flu Trends – for Mexico. You'll need to download the spreadsheet **`Dengue_trends_AM_111.xlsx`**. The first column represents the date (in months, from 2004 – 2011), the second column represents the number of cases of Dengue in Mexico (as reported by the Mexican Ministry of Health), and the third column represents a rescaled number of Google searches for the term "dengue" in Mexico each month.

In the fourth and fifth columns, you'll notice that we have gone ahead and *standardized* the data in columns two and three, respectively. When we develop models for machine learning, we often want to make sure that the components – or features – of our models are on the same scale so that they can be compared directly (i.e., in our case, we can see that column two contains raw values while column three contains scaled values). One way to do this is through standardization, which takes input data and scales it so that it has a mean of 0 and a standard deviation of 1; the formula we use is as follows: $Z = \frac{x-\mu}{\sigma}$, where x is each original data point, μ is the mean of the given data, σ is the standard deviation of the given data, and Z – also known as a standard score – is our rescaled data point that represents the number of standard deviations above or below the mean that a specific data point falls. A negative Z score means that the value of our original data point is less than the average value of the dataset.



1) Plot the standardized number of Dengue cases as a function of time. What trends do you notice in the curve?

2) In machine learning, we often talk about dividing our data into a training set and a testing set. When we want to "teach" a machine learning model how to do a task – in this case, ultimately forecasting Dengue cases – we need to give it examples of what we want it to learn. By showing the model lots of examples, it can learn to recognize patterns and to make predictions on new data that it hasn't seen before. The training set is this collection of examples, more specifically a collection of input-output pairs, and is a subset of the entire dataset.

   Let your training period span the period 2004 – 2006, inclusive (36 months). Fit a linear regression model that explains the standardized number of cases of Dengue as a function of Google searches for the term "dengue" (HINT: you can use Excel's "add trendline" option). What is the equation of the best-fit line? Do Google searches for the term "dengue" appear positively or negatively correlated with the number of cases of Dengue?

3) Use the equation of the line you obtained in 2) to plot the standardized number of Dengue cases as a function of standardized Google searches, predicted by your method during the training period. You may want to create another column in the spreadsheet that reflects these predicted values. Additionally, compare your results to the plot in 1) restricted to the same time period.

4) Now, for the prediction period spanning 2007 – 2011, inclusive, use the equation of the line you obtained in 2) to predict the number of Dengue cases as a function of Google searches for "dengue." Plot this prediction and compare it to the actual number of cases for the same time period.

5) Discuss the results you obtained in 1) – 4). How would you suggest improving this modeling approach? What other data sources would you propose using to increase confidence in the approach? Take into specific consideration the geography in question.