# Predicting Dengue Case Numbers Using Climate Data:

## Milestone Report

This project is a study of the effect climate data has on the number of cases of dengue in a week developed off of the data provided by several US departments. The data provided is a compilation of different survey systems recording different features of the climate and geographic surveys of two different cities: San Juan, Puerto Rico, and Iquitos, Peru. Both cities have experienced outbreaks, periods where the reported number of cases in a week is a significant deviation higher than the average number of reported weekly cases. The data spans across several years, with each dataset encompassing different year intervals, but the same features are present in each dataset. This project is an exploration of the interaction that the climate features have on the case numbers, to model the independent features in such a way to be able to get approximate forecasts of the case numbers for a testing set given the appropriate features with attention given to the performance of the models during the outbreak periods for each city.

This project was inspired by a competition this coalition of government agencies held in 2015 to build a prediction system that helps with understanding how these features play a role in the case numbers and to be able to better prepare a response or strategy when a forecasted period indicates a spike in case numbers of dengue to provide adequate aid for the cities affected. With a successful prediction model, the response to outbreaks of dengue can prove critical to providing a proper response in aiding the civilians and societies afflicted with necessary actions to either help these cities recover from the widespread plague undercutting the health of these societies or provide guidance for steps and actions that can be put in places by these cities to reduce the number of cases, reduce the number of people exposed, implement strategies for mosquito population control, or build up the inventory of medication and supplies needed to treat those afflicted.

The dataset was obtained from the DengAI: Predicting Disease Spread competition hosted by DrivenData.org, which itself is a representation of the Dengue Forecasting competition hosted by the coalition of different US departments: Dept. of Health and Human Services, Dept. of Defense, Dept. of Commerce, and the Dept. of Homeland Security, with support from the Pandemic Prediction and Forecasting Science and Technology Interagency Working Group. The dataset provided is split into 3 CSV files, a training set of the feature data, a training set of the target variable total\_cases, and a testing set of the feature variables used to grade the performance of the submissions to this competition against a testing set of the total cases held out by the hosts and not made public. This project will utilize the first two CSVs to model and predict since the other CSV doesn’t include the true case numbers for these weeks.

The features provided in the dataset includes several climate features including temperature, humidity, dew point, vegetation index, and precipitation variables from the different government agencies in the council that originally hosted this competition as well as some time data associated with when the sample was taken. Some of the data sources include a reanalysis of the same data variables provided by different sources though in different units of measure. There are weeks with missing values for some of these variables to a varying degree. Some preprocessing was necessary to be able to model appropriately.

The month that each sample fell in was extracted from the week\_start\_date feature and dummy variables were created to represent the month labels as binary features rather than one categorical feature since many models aren’t able to interpret categorical variables. Several features were of the same process surveyed differently, and so the units of these features were aligned, which may not be necessary but it does make for easier interpretation. This was done before dealing with missing values to easily determine the imputation performance. A more in-depth imputer was necessary to provide a more realistic pattern in the climate features rather than using a simple approach like filling with the average value of the feature. Other common imputation techniques such as forward filling the previously valid value are fine for features with only a couple of missing values but features with longer ranges of missing values would result in a flat trend, which is isn't very realistic and could throw off the models ability to learn on the variances of these features, giving potentially worse fit and a decrease in prediction accuracy. Additionally, some years had entries for the 53rd week of the year which were the first week of that year mislabeled. With the basic preprocessing and data cleaning done, more preprocessing will be performed later per model requirements depending on the assumptions on the input data and the algorithm used.

Mosquitoes are the vector for the dengue viruses, and its entomology is affected by external factors such as temperature in their local ecosystem as well as standing water sources available for the larvae to mature in. Thus, it may prove useful to understand the precipitation and temperature features on a deeper level. There is not only a difference in the case patterns but of their climate patterns between Iquitos and San Juan. Looking at the average air temperature between the two cities, we see larger deviations from the mean for Iquitos than San Juan, which could prove some importance. Looking at the distribution of this feature, we see San Juan is bimodally distributed and Iquitos seems more normally distributed. This is good to know for when building models, should this feature be incorporated. To get an idea of how well this is representative of the population, I used bootstrapping to draw replicates from these samples to observe what a mean value of this feature could be for each city. Drawing bootstrap replicates and utilizing the Central Limit Theorem supported the hypothesis of the difference in mean average air temperature between the cities is nearly identical for the average air temperature between the cities, suggesting that the observed data is representative of the difference of the population from these two cities. With this test supporting the hypothesis that the mean values are close, we are affirmed this result did not occur by chance.

Standing water is a hotbed of mosquito activity, and so, precipitation data may prove to be of some importance to predicting cases of dengue. I was able to find some information on what happened during those peak outbreak seasons in San Juan, but less so for the peak infection times in Iquitos. I wanted to look at how these precipitation measurements were different between the two cities as I suspected that San Juan, being in the Caribbeans, had precipitation records closer to that of Iquitos than our sample data showed. By utilizing a permutation test of means, the results showed that the two mean replicate precipitation values were much closer together than shown in our data, suggesting that both locations had close to equal amounts of precipitation on average. As seen in the EDA portion, there doesn’t seem to be a feature variable that is strongly correlated with our target variable of total cases per week, so knowing how these two features that can affect mosquito populations strongly are related and represented is important to know.