ANALYZING THE SPREAD OF DENGUE FEVER IN CONTRAST TO CHANGES IN ENVIRONMENT VARIABLES

BACKGROUND:

Dengue fever which is also known as breakbone fever is transmitted through Aedes mosquito bites which carry the dengue virus resulting into illness noticeable through symptoms such as headache, muscle pain, vomiting, to name but a few. We aim at analyzing the relationship between the transmission of dengue fever has with the changes in various environment variables/ factors using a structured dataset from driven data [1].

OVERVIEW OF THE DATA:

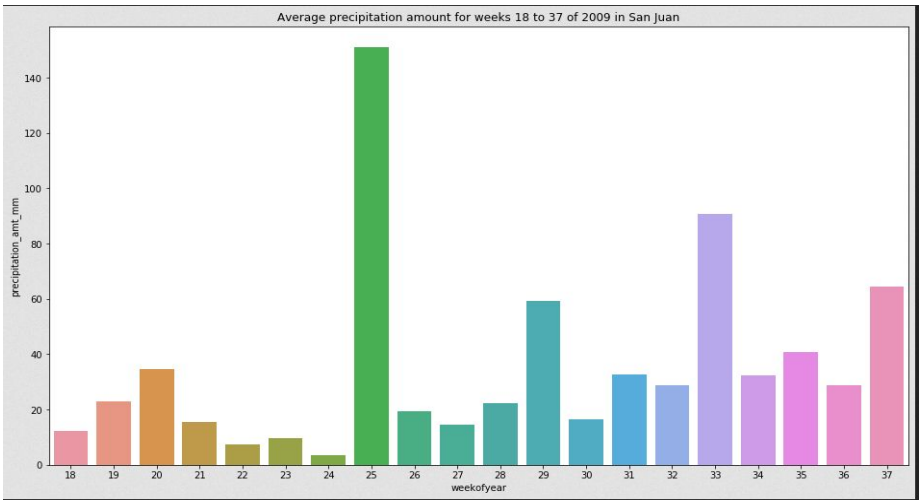
The data in question is a structured dataset obtained from the driven data website[1] and comprises 25 feature columns and 1456 rows. Each row in the dataset is indexed by *(city, year and weekofyear)* and the dataset focuses majorly on two cities, i.e San Juan and Iquitos which are respectively represented as *SJ* and *IQ*. The *weekofyear* just like any other dates indicated in the dataset are given in the yyyy-mm-dd format. The features consist of environmental factors such as temperature, humidity, precipitation, diurnal temperature ranges and the Normalised Difference Vegetation Index (NDVI); which is used to quantify vegetation cover of the target area under observation from remote satellites. Various measurements of spread(range, maximum and minimum) and center(mean) are also recorded for each of the numerical features. All this data is collected over the course of twenty years from 1990 to 2010. The number of dengue fever patients recorded for every week are also included in each row of data.

Through analysis, we aim at getting insights into whether higher temperatures and humidity are associated with a higher number of dengue fever cases reported or more dengue fever cases are associated with a higher NDVI and low precipitation levels.

References:

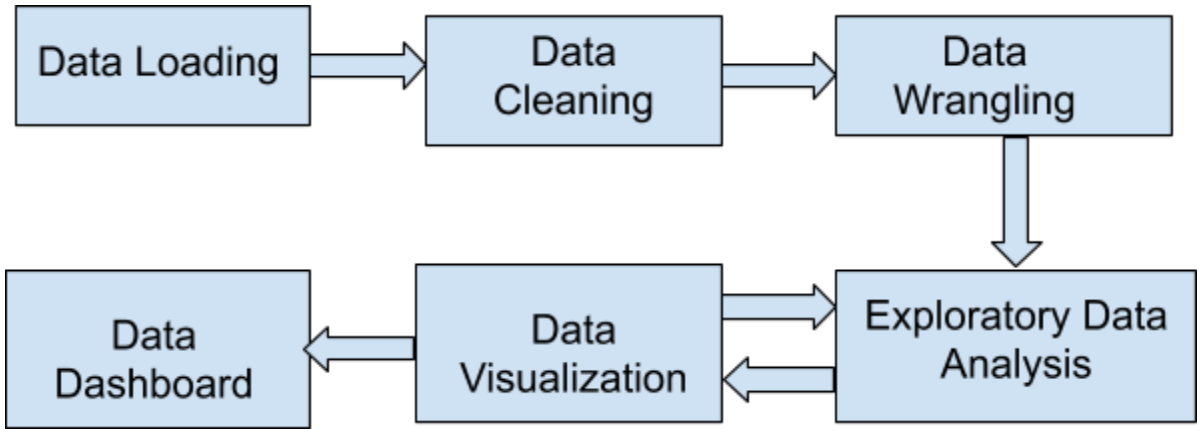
[1] https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/

A bar graph showing average precipitation amount against weeks of the year 2009 for San Juan



Regression analysis is one way we'll assert the truthfulness of the relationships between various dependent and independent variables in the dataset and through visualization, we will graphically illustrate the insights obtained from the data.

WORKFLOW (DATA PIPELINE):



Data pipelines as in the one above integrate various processing elements/steps into a single unit of parallel processes so as to eliminate repetition of individual processes for similar data at a later stage. The processes in the pipeline above are elaborated on below.

Data loading is the process through which data will be read into a data analysis tool such as the pandas python library which is our preferred choice due to its flexibility and seamless integration with the python language.

Data cleaning will be performed on the data upon loading and here inconsistent, inaccurate and irrelevant data such as missing values will be detected and corrected through the use of the pandas and numpy libraries. In data wrangling, the data is to be restructured into consistent formats/data types to prevent clashing data types errors that is to say, all categorical data is to be encoded as text and numerical data as integers.

Exploratory data analysis is the mechanism through which the attributes of the data will be established and preliminary insights into the data made with the aid of summary statistics, hypotheses will be tested and assumptions into the data asserted. In data visualization, graphical illustrations of the insights from the data will be made with the help of the matplotlib and seaborn libraries. Bar plots will aid in scenarios comparisons are made on categorical data such as in the plot above. Line plots will help visualize the change in different variables overtime and scatter plots will be useful in showing the relationship between independent and dependent variables therein aiding in regression/correlation analysis.