The problem for my capstone project is building a model to predict the number of cases of dengue using various environmental factors including temperature, moisture and vegetation. The model is being built for a competition hosted by drivendata as part of the Prevent the Next Pandemic Initiative launched by the white house. The data set has already been cleaned and was split into training and testing sets. The training set was split up into dependant variable data and independent variable data. Only the dependant variable data was provided for testing. The predictions are to be submitted as part of the competition. The competition does not allow for the usage of other datasets. I have started of my capstone project by conducting an exploratory data analysis. I looked at the data distribution for the various variables and found that eleven of the variables had skewed distributions which may be something to keep in mind when deciding what kind of machine learning algorithm to use. I also plotted a correlation matrix to see which variables had strong correlations to the number of cases and plotted scatter plots for the five variables against the number of cases to look more closely at their relationship. Since the model will be used to predict the number of cases, which is a continuous variable, I will use regression when creating a model instead of a classifier. I can't use linear regression however since the scatterplots show that the data isn't linear.