**General Descriptions of Data from EDA**

1. Describe the data. What does it represent? What types are present? What does each data points' distribution look like? Discuss these questions, and your own, with your partners, and document your conclusions.
2. What kind of cleaning is needed? Document any potential issues that will need to be resolved.

**Problem Statement:**

We are attempting to predict a binary value(yes/no) for a test of the West Nile Virus in various traps set around the city of Chicago in odd numbered years between 2007-2013. We will be using a variety of predictive variables including weather, the presence of spraying to combat the virus, geographical location, and yearly/monthly/weekly trends to the information.

Our hypothesis is that these variables will have a predictive effect related to the presence of a positive test for West Nile Virus in Chicago on a trap by trap basis.

**Project Planning:**

Our goals for the project are to create a model which beats the baseline accuracy of 95% (represented by the total # of negative tests across the sample data, therefore we should strive to beat a “node” which guesses negative and achieves a 95% accuracy score). Our model will focus initially on using species type, and a geographical indicator of block (combined with a metric which shows the total # of positive tests from a given block). We would also like to show seasonal trends inherent in the information.

Optimistically, we would like to combine weather information along with a higher degree of geographical specificity using latitude and longitude into our model, but this assumes we are able to work with this data in an efficient and effective way. However, doing so would allow us to leverage the .csv file which documents all preventive sprays to guard against disease bearing mosquitos.

**General Organizational Shit:**