What is our problem statement?

West Nile Virus is most commonly spread to humans through infected mosquitoes, and can cause a fatal neurological disease in about 20% of humans who are infected with it. The first humans were infected with West Nile Virus in Chicago in 2002. We have been tasked with using available data from certain years on trap locations, weather data, and pesticide spray data, to identify where and when West Nile Virus is likely to occur in other years. Due to limited availability of resources for testing traps and spraying pesticides, these predictions can be used to assist efforts and efficiently allocate resources to prevent the spread of West Nile Virus to humans in high-risk areas.

What can we learn from the data in order to make an educated hypothesis? What is our hypothesis?

Each observation in the dataset contains information about a particular trap and includes the Date, Address, Mosquito Species, Block, Street, Trap, Address Number and Street, Latitude, Longitude and Address Accuracy. The training data also included information about how many mosquitoes were caught in the trap, and whether or not West Nile Virus was present among the mosquitoes tested in that observation. The number of mosquitoes caught in each observations was capped at 50, so there are some rows that have the exact same observation repeated multiple times. For example, if a trap caught 342 mosquitoes in a day, there would be 6 rows with the same information (50 * 6 = 300) and one observation where all of the information is the same except that number of mosquitoes caught would be 42 rather than 50, and whether or not that observation of mosquitoes had West Nile Virus. This cap makes the number of repeated observations for a trap on a certain day a proxy for the number of mosquitoes caught by that trap on a day. These were counted by each trap and month, and then weighted by the total number of observations that had West Nile Virus present in a trap in a given month. For example, if 1 out of the 7 individual observations for the trap described above had West Nile Virus present, it would be assigned a weight of 1/7.

My preliminary knowledge of West Nile Virus is that it thrives in hot, dry conditions, so our hypothesis is that West Nile Virus will be most prevalent in the summer months of June, July, and August.

Separate information was also provided about weather and temperature and locations the city has sprayed in the past. Based on our hypothesis, we originally incorporated a weight of the average temperature and precipitation for a given month compared to the average during that month across all 8 years of available data. However, later discoveries in our model building showed that this had no positive effect on the predictability of our model. Given more time, it may have been useful for us to look at a rolling average temperature and level of precipitation for a range of dates leading up to the date that a trap was inspected, since we were eventually led to believe that there was some lag or possible incubation period after hot and dry conditions occur before West Nile Virus would show up in mosquitoes.