# EDA

Description as given by Kaggle:
"In this competition, you will be analyzing weather data and GIS data and predicting whether or not West Nile Virus is present, for a given time, location, and species.

Every year from late-May to early-October, public health workers in Chicago setup mosquito traps scattered across the city. Every week from Monday through Wednesday, these traps collect mosquitos, and the mosquitos are tested for the presence of West Nile virus before the end of the week. The test results include the number of mosquitos, the mosquitos species, and whether or not West Nile virus is present in the cohort. "

Kaggle had particular emphasis on saying that in the "*test set, we ask you for all combinations/ permutations of possible predictions and are only scoring the observed ones.*"

Describe the data - Overview:
This dataset includes 4 CSV files related to the Chicago Metropolitan Area, representing 3 areas of investigation. The files *train.csv* and *test.csv* are delineated by a test observation for West Nile Virus present, grouped by dates, looking at specific mosquito traps. Accompanying each observation is the trap number, and the specific address location down to Longitude/Latitude with a caveat of how accurate is that location data ("AddressAccuracy"). *Train.csv* includes 2 additional columns, with 1 serving as our target to predict for *test.csv* - namely "NumMosquitos" and "WnvPresent." The goal is to use relevant information in this and other files to predict the likelihood that an observation will have a positive mosquito for West Nile Virus ("1" in column "WnvPresent".). Of note from Kaggle, not all locations are tested at all times, and records are only present when a particular species is found in certain traps at certain times.

*Spray.csv* includes information about spraying of mosquito pesticide, with each row representing a spray's date, time, and longitude/latitude.

*Weather.csv* includes NOAA weather observations per date at 2 stations. Station 1 is at Chicago O'Hare International Airport, and 2 is at Chicago Midway International Airport. Specific coordinates are listed at the end of this section for reference. The weather conditions are listed for 2007 to 2014 during the relevant periods of mosquito testing.

Airport Coordinates:
Station 1: ORD. Lat 41.995, Lon -87.993, Elev 662 ft above sea level
Station 2: MDW. Lat 41.786, Lon -87.752 Elev 612 ft above sea level

Describe the data - Data Types:
*Train.csv*: Overall no missing values. Most columns are related location strings for a specific "Trap". Likewise, most data is read in okay, with "Longitude" and "Latitude" correctly a float. "Date" imported as an object but should be a datetime series. As "Species" is of interest, important note that in addition to singular species there are combinations of "Pipiens/Restuans". Interestingly, 136 traps are noted, but 138 location values ("AddressNumberAndStreet", "Longitude", "Latitude"). Plotting distribution is mostly limited to visualizing the geographic trap locations, as other data are mostly label class.

*Test.csv*: Mostly the same. Same curious grouping of traps off by 2 from locations (149 traps, 151 locations). Also expectedly missing “NumMosquitos” and “WnvPresent.” Of note, “Species” has an additional grouping versus train called “Unspecified Culex,” but dummies can still be created for particular species.

*Spray.csv*: Date, time, and location data for mosquito sprays. “Time” has missing values, but eventual grouping by date alone will likely be most appropriate (likely very little gain from considering the specific time of day just in general). Additionally, time only seems to be missing on 1 date (2011-09-07 specifically) so may be of limited utility. Useful plotting is visualizing spray locations per date.

*Weather.csv*: Several columns, almost all with some sort of issue. Overall there were no “missing” values as every cell is filled with something, however, missing points are filled in with “M” or “-”. In particular Station 2 often is missing data. Overall “Sunrise” and “Sunset” are times and largely missing from Station 2, which could be reasonably imputed from Station 1. Further investigation revealed issues with DateTime conversion as some “Sunset” values used XX60, instead of moving up to the next hour unit at XX00 and were corrected. “CodeSum” has several “unique” values representing coded significant weather observations, but closer investigation shows that only a few codes are actually present. “Tavg" has missing values and should be an integer like “Tmax" and “Tmin" (Temperatures); can calculate by hand for any missing points. “WetBulb” seems logical to impute from the other station for those values. Plotting the average for investigation showed a somewhat normal or left-skewed distribution. “Heat” and “Cool” are comparing the average temperature for a day to a value of 65, in the context of indoor heating or cooling (as in, days above 65 have a calculated difference to the average temperature in “Cool” and below 65 have this difference to “Heat.”). Can be easily calculated but may not be a relevant feature as this presumably has little bearing on a Mosquito’s life. “Depth” is snow precipitation depth, but only values present are either “M” or 0 and can likely be safely discarded. “Water1” is also a precipitation measure, believed related to snowfall (water equivalents), but is entirely “M” so while it would likely be helpful, there is no helpful data here and can definitely be safely discarded. “Snowfall” has mostly 0 values or “M”, with 13 non-zero observations (eventually filled but is a column that can likely be dropped as a result. “PrecipTotal" has trace values and were imputed as 0.00001 after considering source below*. It also accounts for rain and snow in the precipitation amount, confirming that “Depth” and “Water1” can be safely removed (it encompasses those observations in terms of water amounts). “StnPressure and SeaLevel” represent atmospheric pressures and both have missing values; can be imputed manually after looking at trends between the stations. “ResultSpeed”, “ResultDir”, and “AvgSpeed” are related to wind direction and strength from a station. These are correctly recognized as int/float so can be used as is.

*https://www.stateclimate.org/sites/default/files/upload/pdf/journal-articles/2013_Adnan_et_al_2013.pdf

Key Assumptions:

- Spraying will actually have an effect on the mosquito population (effective pesticide with little or no resistance to the pesticide).
- Weather observations are accurate from that station with minimal input errors when a value is actually present (No “M” or “-” in that cell)
- Similar integrity of train and test files.