Deliverable descriptions:

Students should have some log, whether it is a markdown file, text file, or iPython notebook, describing the observations and decisions they made along the way. This should be submitted to your instructor prior to your final presentation.

\*\* This is in ‘collect thoughts’ state. It will be cleaned up prior to submission.

1. Data file examination - Train and Test files:
   1. WnvPresent is what is being predicted.
   2. Test file doesn’t have information on # of mosquitos present like the train file does, so can’t use that information for analysis.
   3. About 8 types of mosquitos involved. Some are more prevalent. Although some carry the virus significantly more than others, data and outside research indicate can’t rule out that any type of mosquito would have the file.
   4. The columns with location information (address, block, trap Id, longitude / latitude) are largely redundant with each other. We decided to take location as a categorical variable since it is unlikely that there is a linear relationship between longitude and latitude and Wnv being present. That is, Wnv present likely vary up and down as longitude or latitude increases. We used TrapId as our categorical variable. 148 traps around the city
   5. The observations (rows) in the files have multiple rows for the same trap on the same day. We combined the rows to reflect one observation per trap per day that we had information.
   6. We do not have trap observations for every day.
   7. The train and test files are for alternate years. This allows us to train a model with ALL the available data for a set of years and see how well it performs in different years.
2. Data file examination – Weather file:
   1. Information was provided for each day during the warmer months of the year. All days included in the train and test file were included in the weather file.
   2. Information was provided from 2 weather stations (O’Hare at Midway airports.) There were a lot of null values, particularly from one of the weather stations. We decided to combine the data from the 2 stations into one record. Where nulls existed, we replaced it with information from the other weather station or from adjacent days. We considered trying to extrapolate weather data for each latitude / longitude for each day, but that was deemed a lot of effort for questionable value, since summer weather patterns often in small cells (like thunderstorms), not a continuous pattern.
   3. Any rainfall observations of ‘T’ (trace) were set to a very small number.
   4. (\*\*\* Chaim, did we drop any features?)
3. Data file examination – Spray file:
   1. Nothing significant found
4. Other data considered:
   1. We considered bringing in elevation data for each latitude and longitude based on the hypothesis that moisture might collect in lower elevations. We prioritized this effort as lower priority after we made the most of the provided data.
5. Hypotheses and Feature Engineering
   1. Based on online research on the mosquito breeding, and spread of the Wnv, we hypothesized that mosquito breeding, and thus Wnv, would be affected by warmth and moisture, both currently, and over the preceding period. (\*\*\*\*We could use a plot showing if Wnv was more prevalent when more mosquitos were present).
   2. Based on this we added features for precipitation (rate of moisture added), wet bulb (rate that moisture removed), and cooling degree days (heat) over the last 7, 30, and 90 days. Note, that the 30 day features had the last 7 days removed, and the 90 day features had the last 30 days removed to make them more independent of each other.
   3. Variable selection: We looked at feature importances in the Random Forest models, and the confidence intervals in the Logistic Regression models. We removed features with lower feature importance. \*\*\*Regularization, plotting, preliminary regressions, and other efforts to remove variables?
6. Choice of Models
   1. Logistic regression – generally performs reasonably well and gives us interpretability which is useful for inferring which variables are more or less important, and makes the model more describable to stakeholders.
   2. Random forest – generally perform reasonably well and provide a cross check to logistic regression with some loss of interpretability
   3. XGBoost – based on work of others, these models perform very well
   4. Are we doing grid searches to tune parameters – y
7. Best model was random forest