**Capstone Project 1: Statistical Data Analysis of WNV (West Nile Virus) dataset**

Our dataset contains two datasets namely the main dataset that has information including mosquito species, trap, WNV positivity, date of trap, as well as weather dataset that includes weather condition during the specified period. Our target variable is WNV positivity which is a categorical variable with 0 representing no WNV and 1 representing presence of WNV in the mosquito. The 'Date' feature byitself is not helpful for our statistical analysis. However, we can extract many variables such as 'Day\_of\_week', 'Day\_of\_year', 'Week\_of\_year', 'Month', and 'Quarter' that can be important to our prediction model. Hence, we extracted these five features and removed the ‘data’ feature from our dataset. Besides as our data is a timeseries data we have dome feature engineering on all the features of the weather data by creating 1 -14 days lag variable for each weather variables,

Regarding the association of the features with the WNV, WNV positivity rate is high in the third quarter, during August and September, which is in line with time of year when mosquito populations will be largest (August, July, June & September). On top of that, WNV positivity rate is high in Thursday, Tuesday and Wednesday. However, we don't have any data that was obtained during the weekends (Saturday, and Sunday). Mosquito traps were done from Monday to Friday only. Knowing which mosquito species are more likely to carry the virus will be useful if the species tend to exist in different areas. CULEX PIPIENS/RESTUANS (262), CULEX PIPIENS (240), & CULEX RESTUANS (49) were the only species carrying WNV. Later in the machine learning part, we will filter our dataset to these three species only since the rest species have null value and as a result have no relevance in building prediction model. The above chart depicts that as the number of trapped mosquitoes increase WNV positivity rate also increases. This finding is straight forward that the more sample we have the more variety of mosquito species we get and consequently we will have more WNV positivity rate.

On the other hand, as temperature increases the positivity rate for WNV also increases. This could be explained by the fact that hot and dry conditions are more favorable for West Nile virus than cold and wet. The distribution of WNV positivity rate over 'DewPoint' and 'WetBulb' is similar. This may indicate these two variables may be highly correlated. Besides, the positivity rate for WNV increases with both 'DewPoint' and 'WetBulb'. The possible explanation for the association between Wet bulb and WNV positivity is that, Wet bulb is an indicator of evaporation rate, thus decreases in moisture, thus less moisture tends to be favorable for West Nile virus.

We have seven species of Culex mosquitoes, namely 'CULEX PIPIENS/RESTUANS', 'CULEX RESTUANS', 'CULEX PIPIENS', 'CULEX SALINARIUS', 'CULEX TERRITANS', 'CULEX TARSALIS', and 'CULEX ERRATICUS'. Out of which only three species (CULEX PIPIENS/RESTUANS (262), CULEX PIPIENS (240), & CULEX RESTUANS (49) ) were carrying WNV. The rest four species have null value for WNV positivity. Hence, we will be forced to remove these four species as they have no relevance to our prediction algorithm. As a result, 865 observations are filtered out and our dataframe contains 10085 rows, and 51 columns.

In order to check the correlation of features with one another or with the target variable we have done correlation matrix.

1. Regarding the correlation among the features:

- 'Sunrise' has strong negative correlation with 'Sunset' whereas it has strong positive relationship with 'Month'.

- 'AvgSpeed' has strong positive correlation with 'ResultSpeed'

- 'Heat' has strong positive correlation with many variables including 'Tavg', 'Tmax', 'WetBulb','Tmin', 'DewPoint', 'Depart', 'Cool'.

- 'Tmin' has strong positive correlation with many variables including 'Tavg', 'WetBulb', 'DewPoint', 'Cool', 'Depart', 'Tmax'.

- 'Tavg' has strong positive correlation with many variables including 'Tmax', 'Cool', 'WetBulb','Tmin', 'DewPoint', 'Depart'.

- 'Tmax' has strong positive correlation with many variables including 'Tavg', 'Cool', 'WetBulb', 'Depart', 'Tmin', 'DewPoint'.

- 'Depart' has strong positive correlation with many variables including 'Tmax', 'Tavg', 'Cool', 'WetBulb', 'DewPoint', 'Tmin'.

- 'Cool' has strong positive correlation with many variables including 'Tavg', 'WetBulb', 'Tmax', 'DewPoint', 'Depart'

2. Regarding the correlation of the features with the target variable:

- 'NumMosquitos', 'Day\_of\_year, 'Sunrise', 'Week\_of\_year', 'Sunrise\_lag\_14', & 'Month' are the top 6 variables that have correlation with the target variable. However, there is not any feature with strong or moderate correlation; all of them have weak or no correlation. Whereas 'ResultDir', & 'PrecipTotal' have no correlation with the targer variable.