# Capstone Project I milestone report

# 1. Introduction

[West](http://www.cdc.gov/westnile/) Nile Virus (WNV) is a disease caused by virus which is most commonly spread to humans through infected mosquitoes. Around 20% of people who become infected with the virus develop symptoms ranging from a persistent fever, to serious neurological illnesses that can result in death. WNV is a disease spread from infected mosquitoes to humans. For many patients, mild symptoms go away on their own but for people with more severe diseases such as meningitis or encephalitis, these symptoms can become life-threatening. People over the age of 70 and with chronic conditions such as weakened immune systems or high blood pressure are at most risk if they are infected with WNV. Chicago has experienced one of the highest levels of WNV risk in this decade. According to the daily herald, 290 Illinois residents have been reported to be ill from this disease, the highest recorded number in Chicago since 2012. The cost of WNV varies from hospital to hospital. However, on average, each admitted patient to the hospital for WNV costs approximately $25,000. On average, the cost for each patient that receives chronic care for WNV is $22,000.

Chicago has had trouble controlling the spread of WNV. By 2004 the City of Chicago and the Chicago Department of Public Health (CDPH) had established a comprehensive surveillance and control program that is still in effect today. As a result, traps were set up to capture mosquitoes. Every week from late spring through the fall, these traps are frequently checked to see if any of the captured mosquitoes carry the virus. Knowing which traps are more likely to have WNV present is important as it provides insight to the city as to where the city is best served to spray for the eradication of mosquitoes. Spraying can have adverse effects on the environment, and it is also very expensive, thus it’s necessary to target specific areas and at specific times. Our goal is to predict where in Chicago WNV occurs to help the city prepare accordingly. The City of Chicago and the Chicago Department of Public Health (CDPH) can use such a model to get information on when and where the city will spray airborne pesticides to control adult mosquito populations. This in turn will play a significant role in the prevention and control of the disease.

# 2. Data Acquisition and Cleaning

The data set was acquired from the Kaggle Competition. Originally, the data has three datasets, the first dataset contains WNV testing & location information, the second dataset contains information about weather, and the third dataset contains spray data. We decided to exclude the spray data as there is not enough information about the sprays used to help guide our project. We checked the cleaned the dataset one by one and then merged them together.

The first dataset (testing & location data) has 12 columns and 10506 rows. After we dropped the unnecessary location related columns ('Trap id', 'Address', 'Block', 'Street', 'Latitude', 'Longitude', 'Address Accuracy'), we are left with only 4 columns namely date that the WNV test is performed, species of mosquitos, number of mosquitoes caught, WNV presence in mosquitoes. We checked all the 6 columns for the missing values of and none of them have missing values.

The second dataset (weather data) It has 2944 rows and 22 columns. We dropped 4 unnecessary columns, and as a result we are left with 18 columns including Date, maximum temperature, minimum temperature, average temperature, departure from normal temperature, Dew Point, Wet Bulb, total precipitation. Since majority of the columns data types is object, we changed them to numeric.

Even though the data have no null values, there were many 'M', 'I’ values in most of the variables that indicate missing values, and we can also see that there ' T' values for the total precipitation column indicating trace amount. We converted the ' T' to 0.001. The data provided had weather report from different weather stations daily. Since the weather measured temperatures in different regions of Chicago, we averaged both weather data. Before we change the 'M' & '-' to missing values, we preferred to get the average of the two stations as most of the missing values are from station 2. As a result, we will have a smaller number of missing values. After we averaged the weather data of the two stations and removed the ‘Station’ column; Now all the variables have 1472 rows, except for ‘station pressure’ which has 1 missing value(nan). Hence, we filled the missing value by the mean of its column. Finally, we merged both datasets by their ‘Date’ column. 10506 rows and 24 columns. More details on acquiring, cleaning, merging, parsing these datasets can be found in this IPython notebook.

# 3. Data Exploration

## 3.1 Introduction to the cleaned data

Our cleaned data has 10506 rows and 24 columns. Our target variable is WNV positivity which is a categorical variable with 0 representing WNV not present and 1 representing presence of WNV in the mosquito. The 'Date' feature by itself 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 ‘Date’ feature from our dataset. We will go through most of the for features of the dataset to explore their relationship with WNV positivity rate. Details about each field can be found in Kaggle and *‘noaa\_weather\_qclcd\_documentation.pdf’.*

## 3.2 WNV positivity rate

Out of the 10506 tests for WNV presence in mosquitos only 551 (5.24%) indicate the presence of WNV in the mosquitos. This does not seem like a large number, but such rare occurrence can cause devastating disease outbreak. Therefore, it is important to understand when, where and how these rare events occur. In the following few sub-sections, we will go through many interesting features and explore the trend for WNV positivity rate. To start with, we plot the total number of WNV positives with the calendar variables extracted from the Date variable. We have five calendar variables such as quarter, month, week, day of week, and day of year.

## 3.3 Summary Statistics

Majority of the columns means except ('DewPoint', 'StnPressure', 'SeaLevel', 'ResultSpeed', 'ResultDir', 'AvgSpeed') are different from the median represented by the 50th percentile. Besides, there is notably large difference between mean and 75th percentile in 'Heat', 'NumMosquitos', 'PrecipTotal'. Therefore, we can understand that there are outliers in these variables stated above. To be more specific we can also check the difference b/n the 95th percentile and maximum value, as well as the 5th percentile and minimum value. There is big difference between 95 percentile and maximum value for 'Heat', 'PrecipTotal', & 'ResultSpeed'. This indicates the presence of extreme outliers in these three variables. There is no big difference between the 5th percentile and minimum value for all the variables

## 3.4 Visual exploratory data analysis

Regarding the skewness of the distribution of each feature, we can see from the figure below that 'StnPressure', and 'SeaLevel' appear to be normally distributed; while 'PrecipTotal', 'Heat', and 'NumMosquitos' are extremely skewed to the right. On the other hand, 'Sunrise', 'NumMosquitos', and 'WetBulb' appear to be bimodal, whereas 'Cool', 'SunSet', 'ResultSpeed', and 'ResultDir' appear to be multimodal. The remaining features are slightly skewed to either right or left.

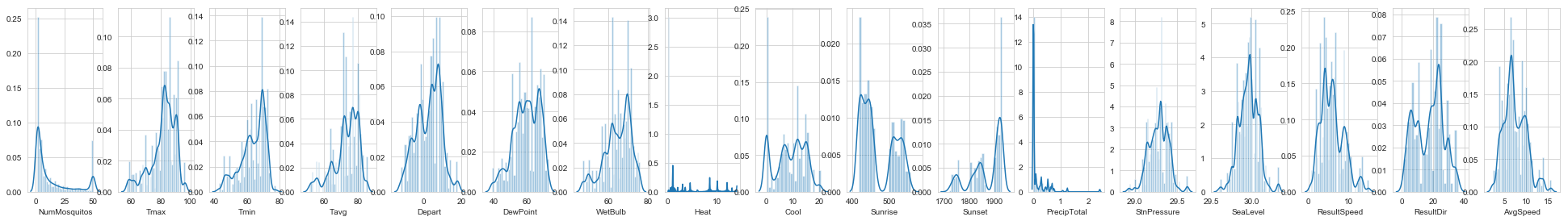


Figure 1: Distribution of all features.

A box plot (or box-and-whisker plot) shows the distribution of quantitative data in a way that facilitates comparisons between variables. The box shows the quartiles of the dataset while the whiskers extend to show the rest of the distribution.

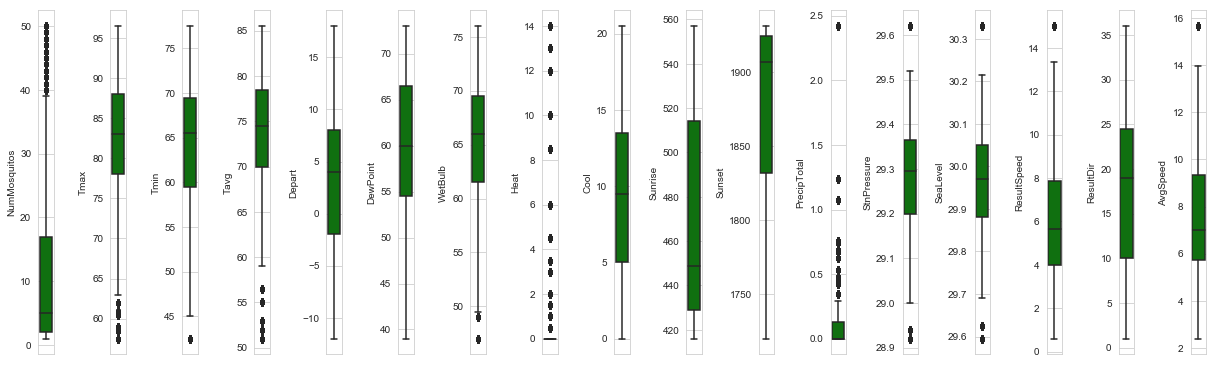


Figure 2: Box plot for all features

The black dot above or below the whiskers indicate the presence of outlier. 'DewPoint', 'Cool', 'Sunrise', 'Sunset', & 'ResultDir' are the features without outliers. All other columns show outliers.

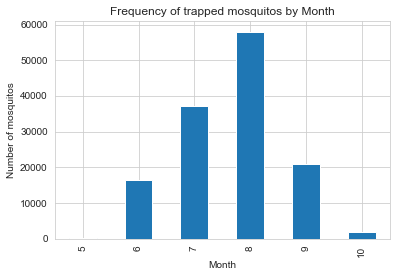


Figure 3: WNV positivity rate by month.

The highest WNV positivity rate was recorded in August followed by July, September, and June. WNV positivity rate is high in the third quarter, in August and September, which is in line with time of year when mosquito populations will be largest (August, July, June & September). Figure 3 shows the number of mosquitos trapped by month.

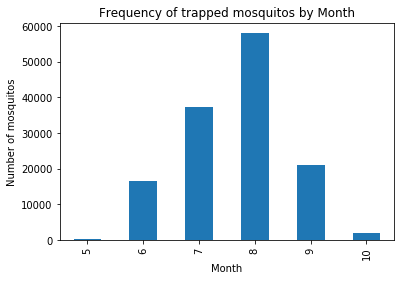


Figure 4: Number of mosquitoes caught in a trap by month.

On top of that, WNV positivity rate is high in Thursday, Tuesday and Wednesday as shown in figure 4. 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.

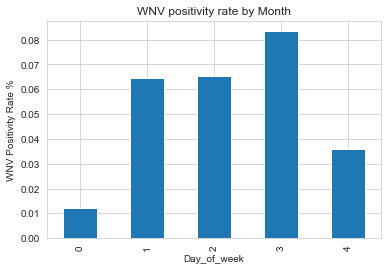


Figure 5: WNV positivity rate by Month.

We have seven species of Culex mosquitoes, namely *'CULEX PIPIENS/RESTUANS', 'CULEX RESTUANS', 'CULEX PIPIENS', 'CULEX SALINARIUS', 'CULEX TERRITANS', 'CULEX TARSALIS',* and *'CULEX ERRATICUS'*. *'CULEX PIPIENS/RESTUANS', 'CULEX RESTUANS', 'CULEX PIPIENS'* were the three dominant species. 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.

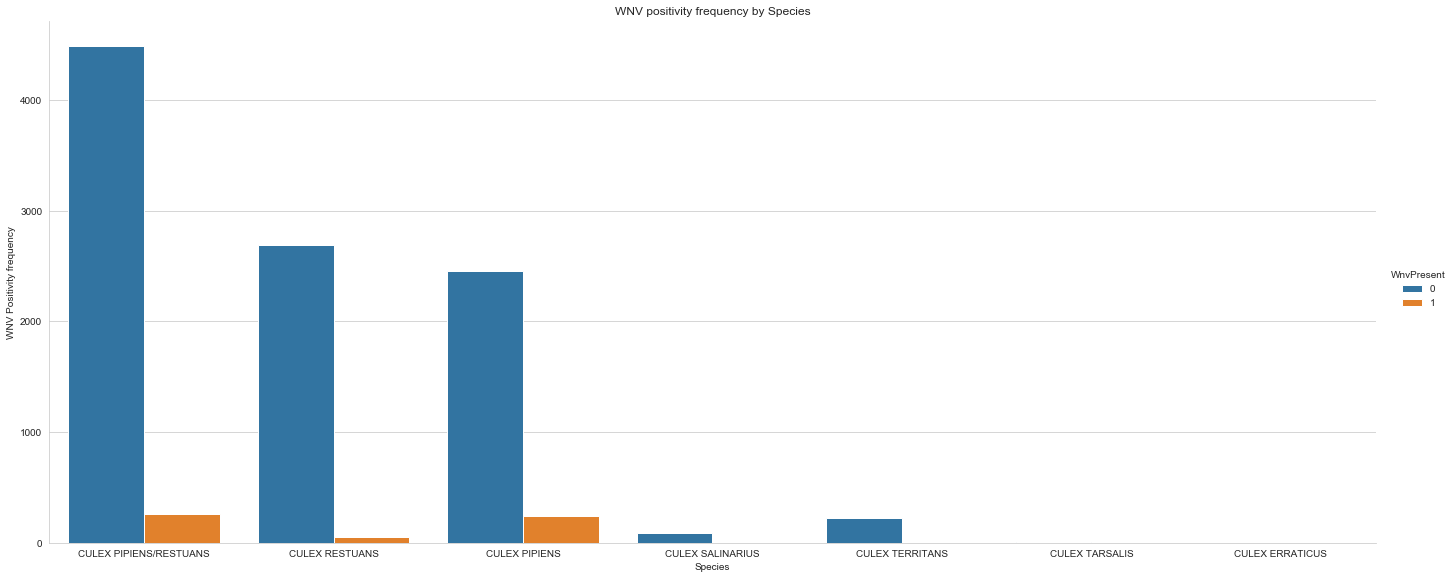


Figure 6: WNV positivity frequency by Species

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.

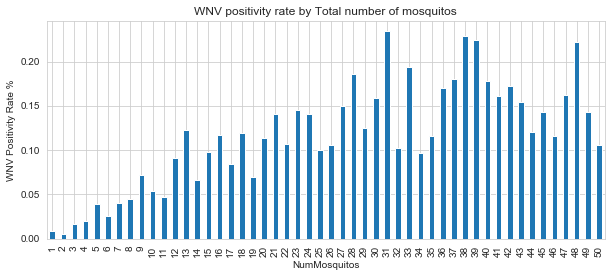


Figure 7: WNV positivity frequency by Species by Total number of mosquitoes in a trap

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.

#### Weather Factors

There are many weather factors such as temperature, precipitation, dew point, pressure, windspeed, wind direction, humidity etc. but we will focus on only some factors here to keep the discussion short. A detailed data exploration can be found in this IPython notebook. The temperature data is described in four different forms namely maximum temperature, minimum temperature, average temperature, departure from normal temperature.

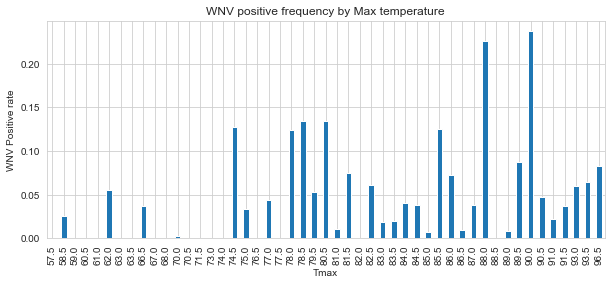


Figure 8: WNV positive frequency by Maximum temperature

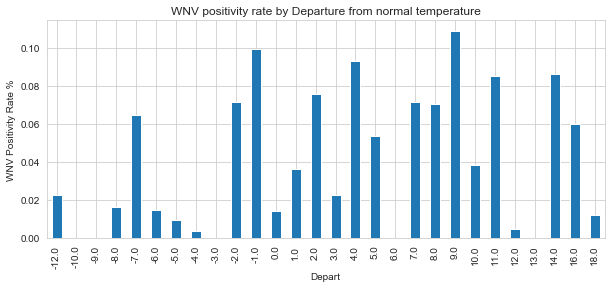


Figure 9: WNV positivity rate by Departure from normal temperature

The above two figures illustrate that 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 WNV than cold and wet.

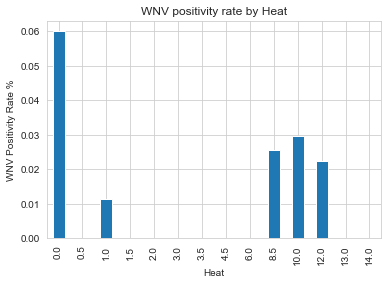


Figure 10: WNV positivity rate by Heat

WNV positivity is associated with low heating days and high cooling days. Low heating days and high cooling days are an indication of hot weather which is favorable to the WNV growth. Majority of the WNV positivity rates occurred at 0.00 heating degree days.

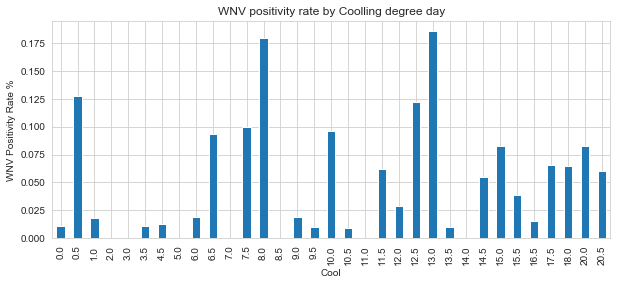


Figure 11: WNV positivity rate by Cooling degree day

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 is that, Wet bulb is an indicator of evaporation rate, thus decreases in moisture, thus less moisture tends to be favorable for WNV.

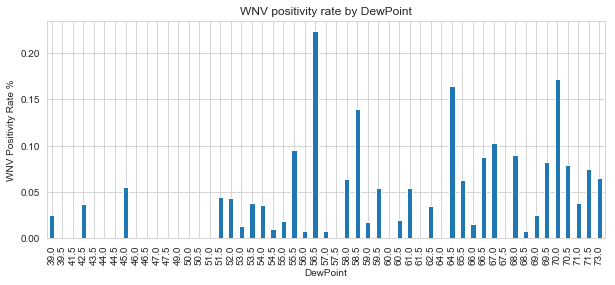


Figure 12 WNV positivity rate by Dew Point.

The WNV positivity rate increases slightly with total precipitation, though the data looks bimodal. Precipitation represents an increase in moisture, which can be favorable weather condition for the growth of mosquito population.

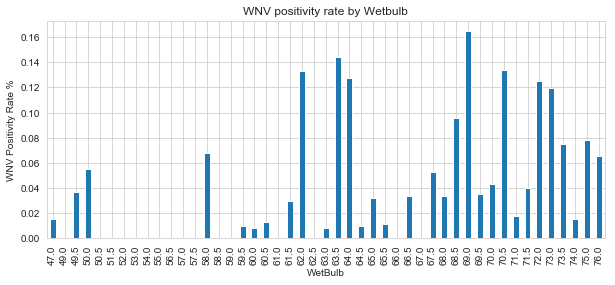


Figure 13: WNV positivity rate by Wet bulb

## 3.5 Outliers Handling

Data values that are either less than the 5th percentile or above the 95th percentile are removed from the dataset. We removed 107 rows that lies in the mentioned range. We used the boundary of 5th percentile and 95th percentile because it is the maximum interval that includes all the WNV positives.  We tried to use different intervals such as 25th and 75th percentile, 10th and 90th percentile, but all of them excludes all the 551 WNV positives which in turn would make prediction impossible.