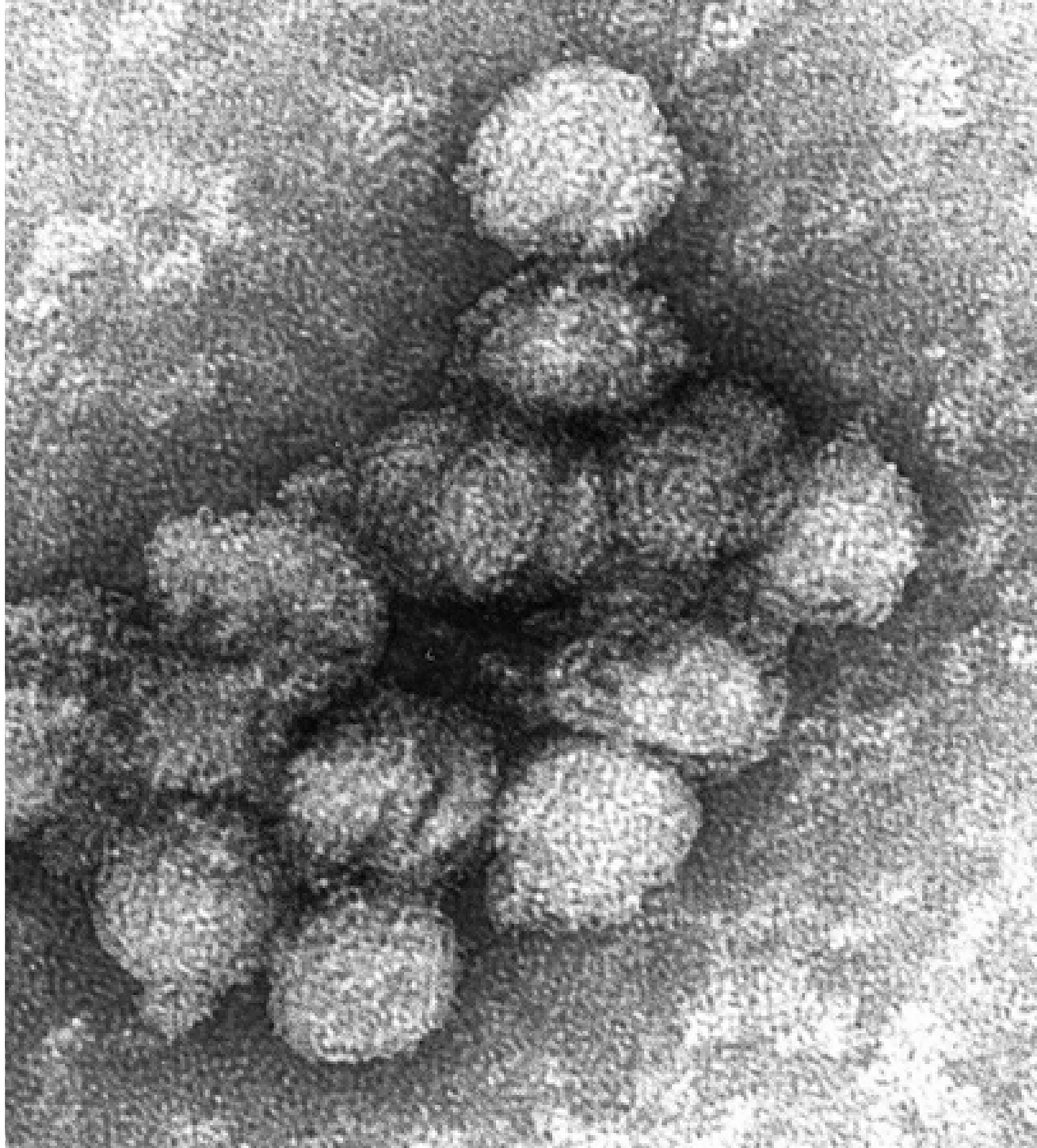
Department of
Data Science

Chicago

West Nile Virus

Data Scientist
Lee Mei, Supriya, Wandi

31 July 2021



Problem Statement

In recent years, there has been a surge in reported West Nile Virus (WNV) in the city of Chicago.

The **Public Health Department** took measures to set up a surveillance and control system, hoping that it will give them some insights from the mosquito population data collected over time.

While common control measures such as pesticides are necessary for the fight for public health, there are concerns to be considered such as cost and safety

As data scientists in the **(DATA-SCIENCE) department**, we are task to derive an effective strategy to curb the problem (while considering the effects of taken actions)

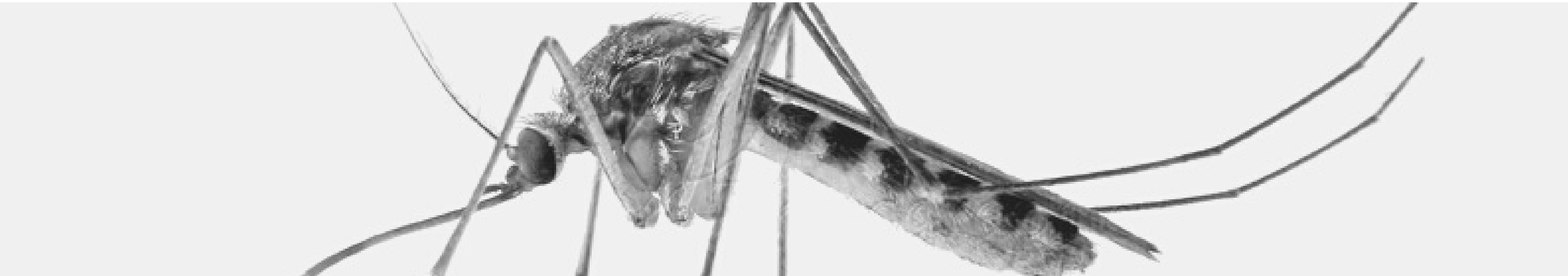
Brief History

West Nile virus (WNV) is a single-stranded RNA virus that causes West Nile fever.

A member of the family Flaviviridae which also contains the Zika virus, Dengue virus, and Yellow Fever virus.

The virus is primarily transmitted by mosquitoes, mostly commonly, Culex species. Mosquitoes become infected when they feed on infected birds. Infected mosquitoes then spread West Nile virus to people and other animals by biting them.

To date, humans and horses both exhibit disease symptoms from the virus and the first detection of the virus is dated back in 1973, Uganda and then to rest of the world.



Timeline in United States

1999

First detected in North America in 1999

Similar cases were reported in other hospitals

Requested assistance to help identify the cause

2001

WNV spreads geographically towards the eastern coast of US

Concerns of epidemic potential

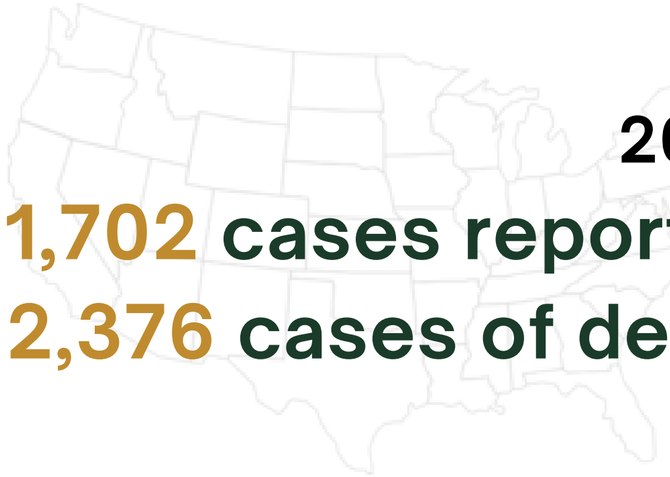
2002

WNV spreads and degree of transmission was unprecedented

4156 cases and 256 death reported

2003 >

WNV was at its peak in 2007 based on number of cases reported over the years



2019

51,702 cases reported

2,376 cases of death

Objectives

- Understanding the current status of wnv over the years (based on the dataset)
- Develop a model to predict the likelihood of WNV present in Chicago
- Understand the cost and benefits of deploying a model to combat WNV?
- Understand what other strategies can be adopted by the Department of Public Health?

" Prevention is better than cure "

Datasets

1

Train Data

- Test results showing the presence of West Nile Virus in the mosquitoes found in traps laid by health workers in Chicago. Data includes location of the traps, no. of mosquitoes, species of mosquitoes, and presence of WNV. Data is present from May to October for years 2007, 2009, 2011, 2013.

Train.csv (10,506 X 12)

2

Weather Data

- Dataset from NOAA of the weather conditions of 2007 to 2014, during the months of the tests. Observations from two different stations.

Weather.csv (2,944 X 22)

3

Spray

- GIS data for their spray efforts of the city of Chicago in 2011 and 2013

Spray.csv (14,836 X 4)

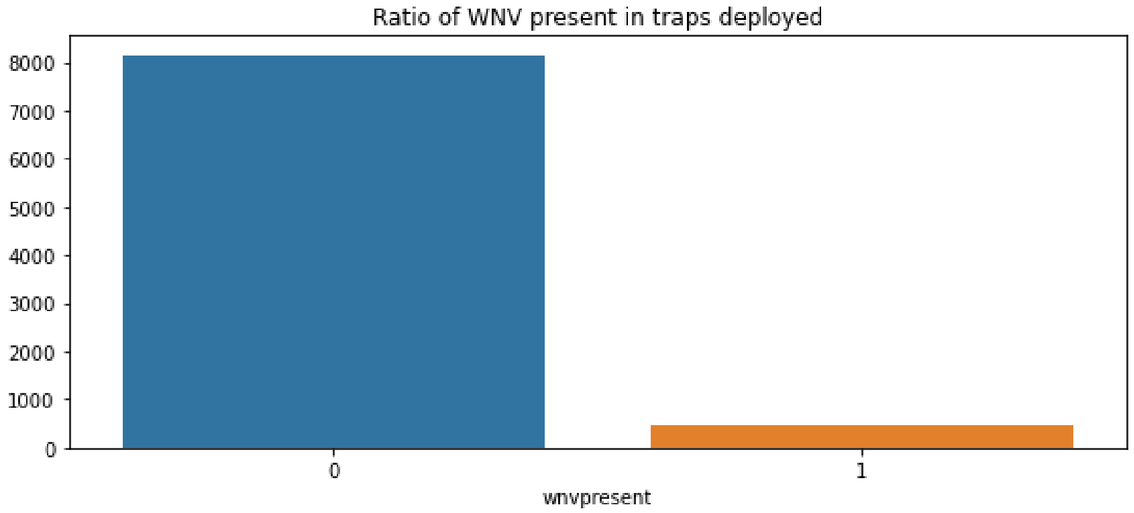
Part 1:

Data Cleaning & Imputation

Train Dataset

1

- **Train.csv**
  - 10,506 records
  - 12 columns
  - No Null rows
  - 1,896 duplicate records: Traps with more than 50 mosquitoes split into multiple records
    - Split rows were combined together
  - Imbalanced Data



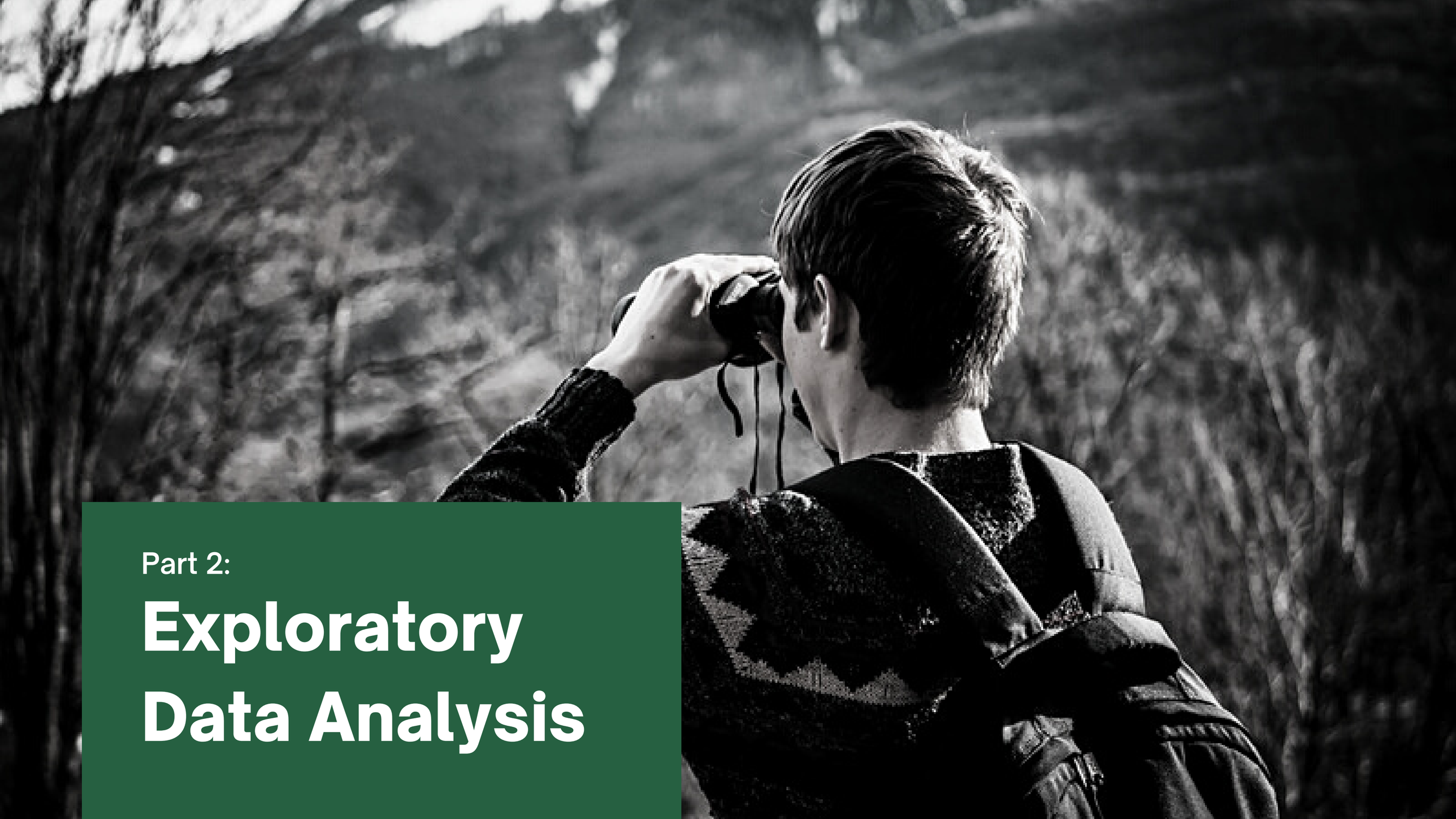
Weather Dataset

2

- **Weather.csv**
  - 2,944 records
  - 22 columns
  - No duplicate records
  - Missing Data
    - **Tavg, Heat, Cool**
      - Imputed using values from other features
    - **Sunrise, Sunset, WetBulb, Sealevel, Avgspeed**
      - Imputed from values for the same day from the other station
    - **Depth, Water1, Snowfall, Depart**
      - Dropped
  - Merging Train dataset with the Weather based on the station located nearest to the traps

Spray Dataset

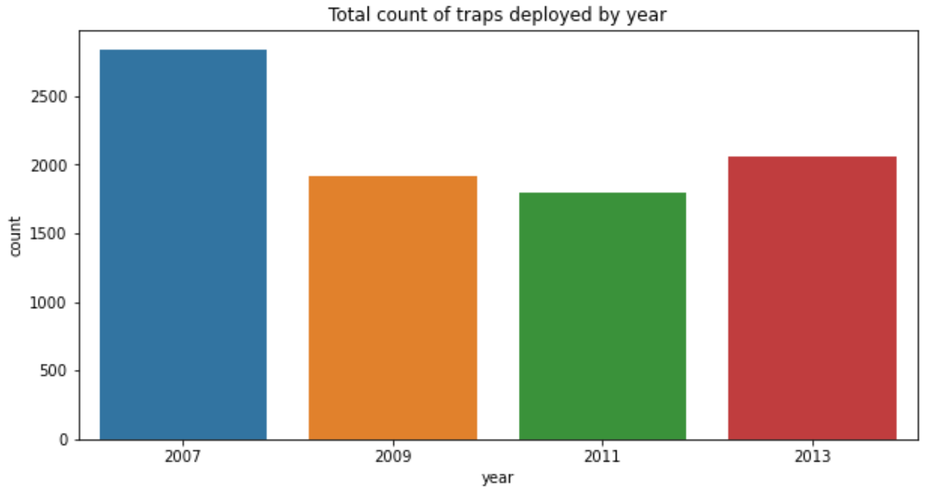
- **Spray.csv**
  - 14,836 records
  - 4 columns
  - 541 duplicate records dropped
  - Missing Data
    - 584 null values in Time column
    - Dropped Time column and removed duplicates

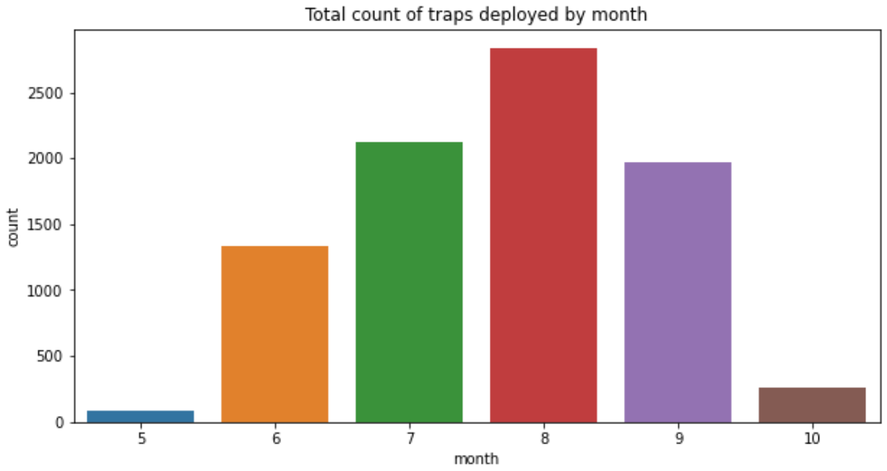


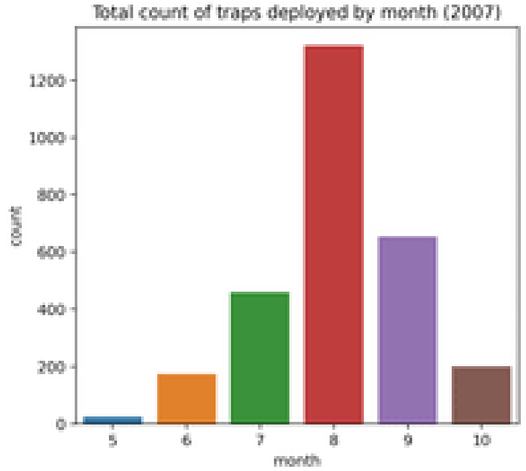
Part 2:

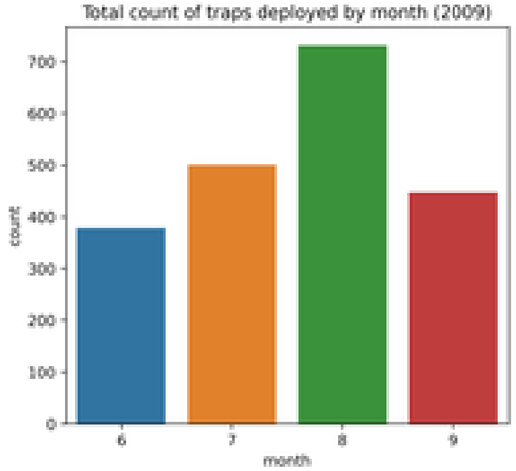
Exploratory Data Analysis

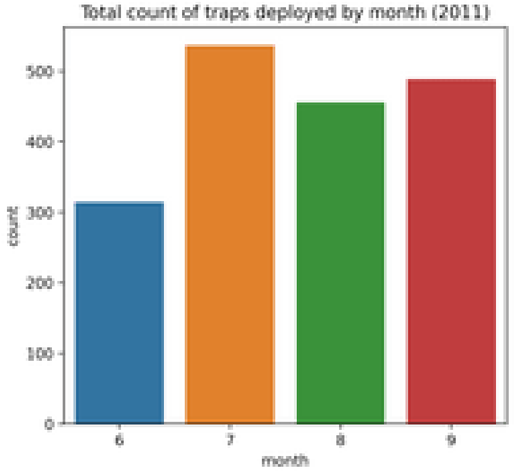
Distribution of traps by year and month

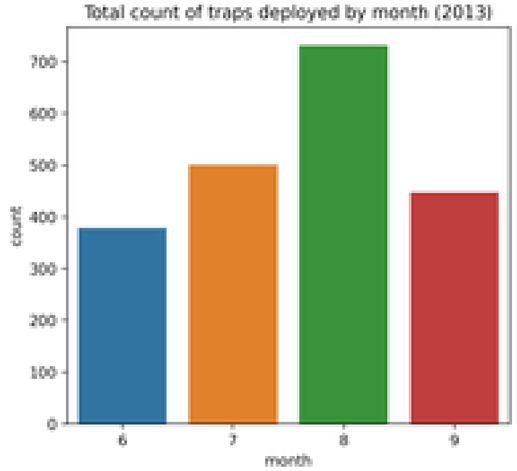




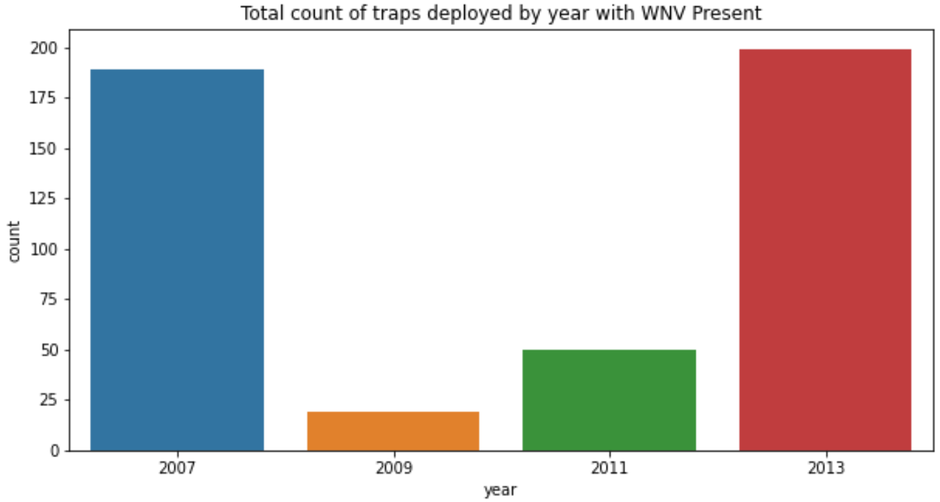


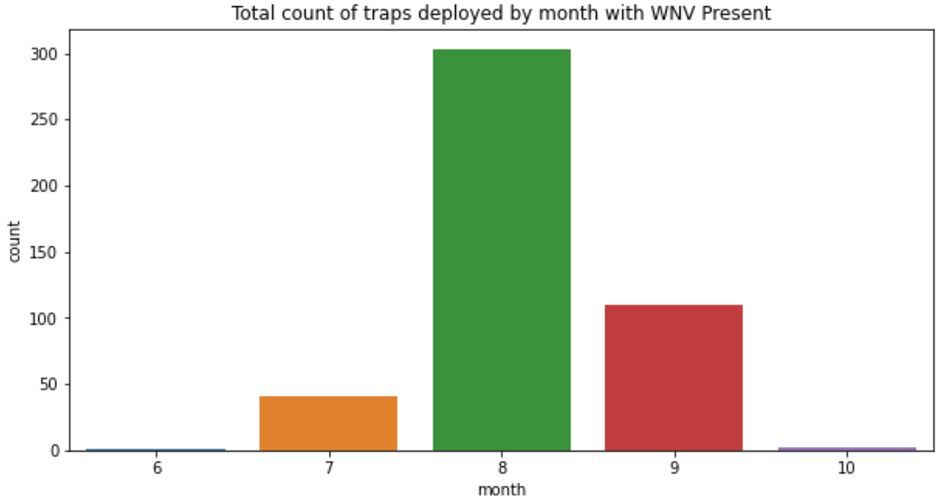


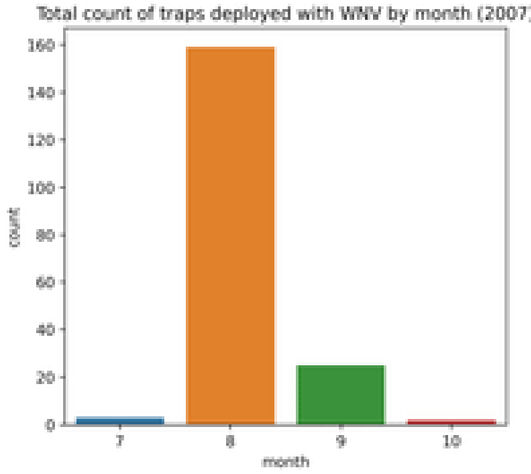


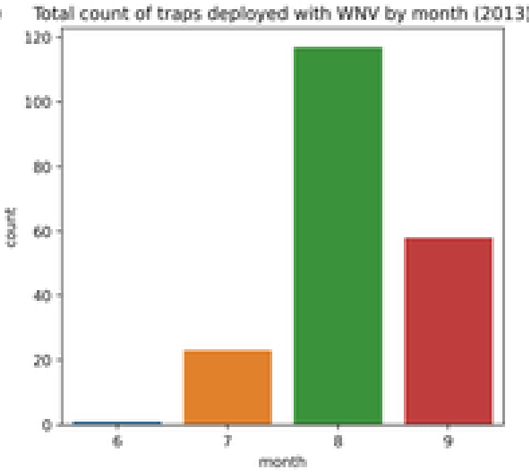


Distribution of traps with West Nile Virus by year and month

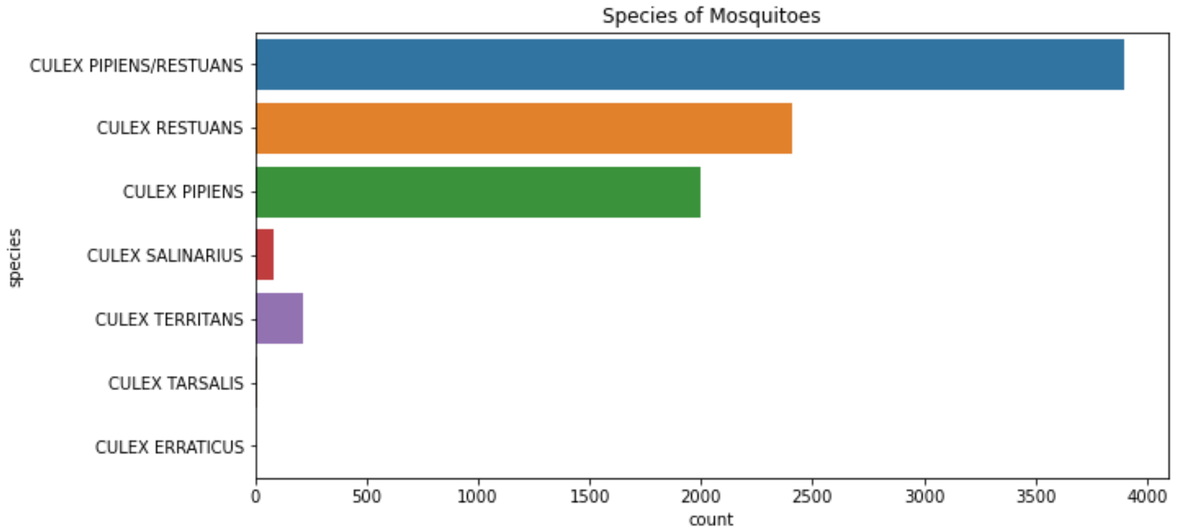


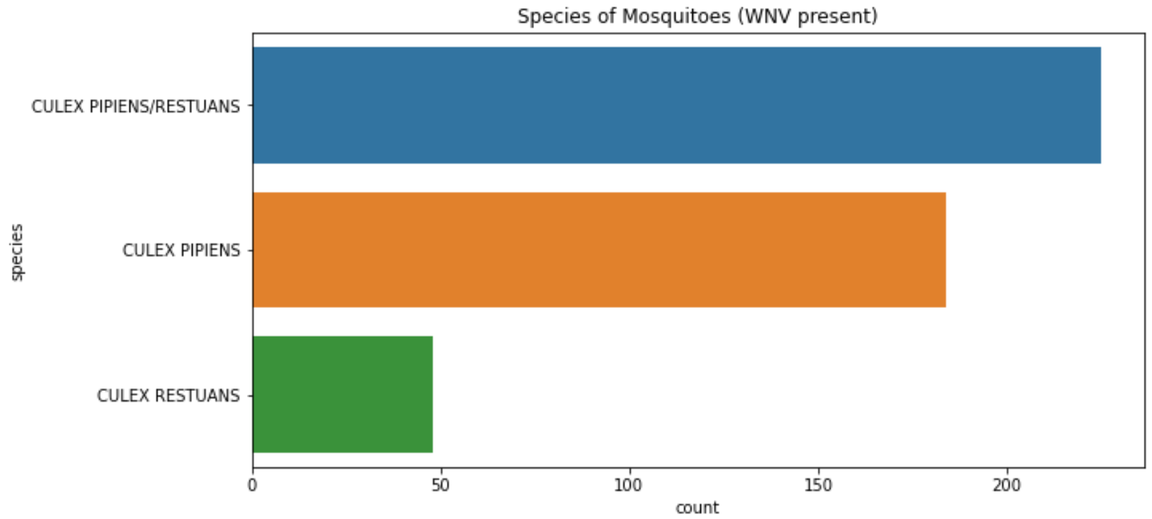




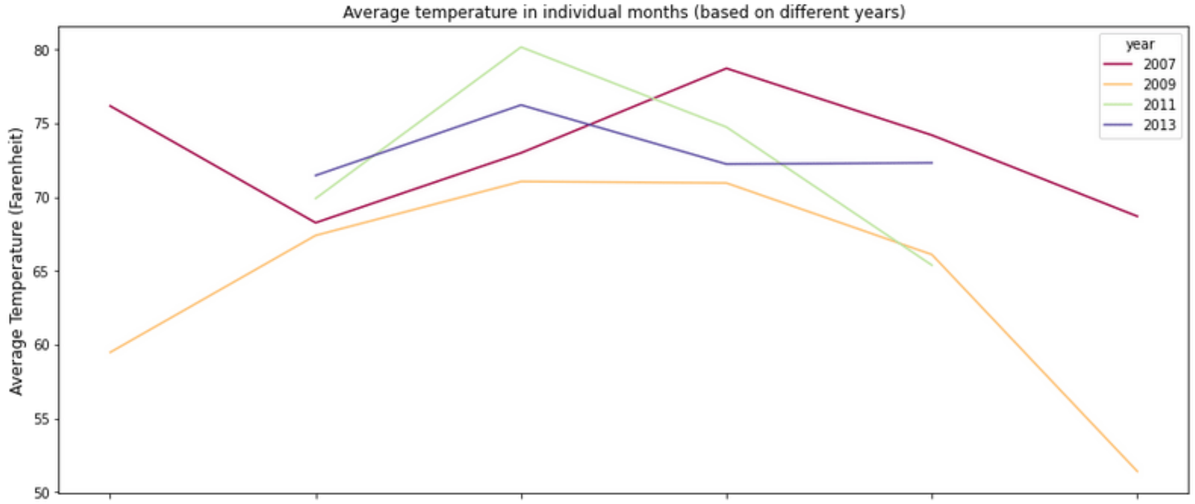


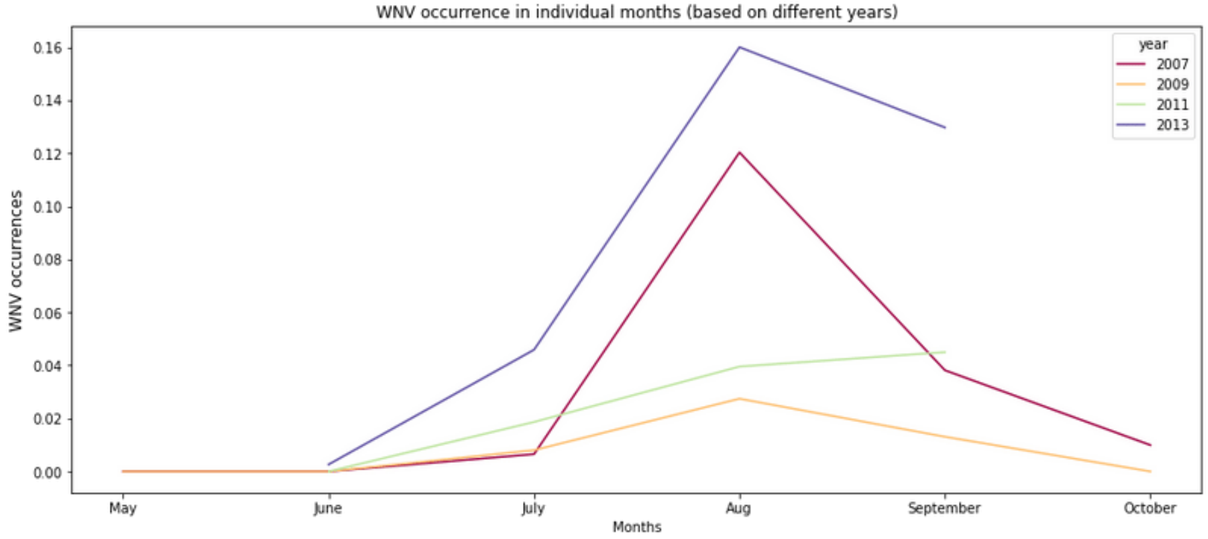
Species of mosquitoes found in traps



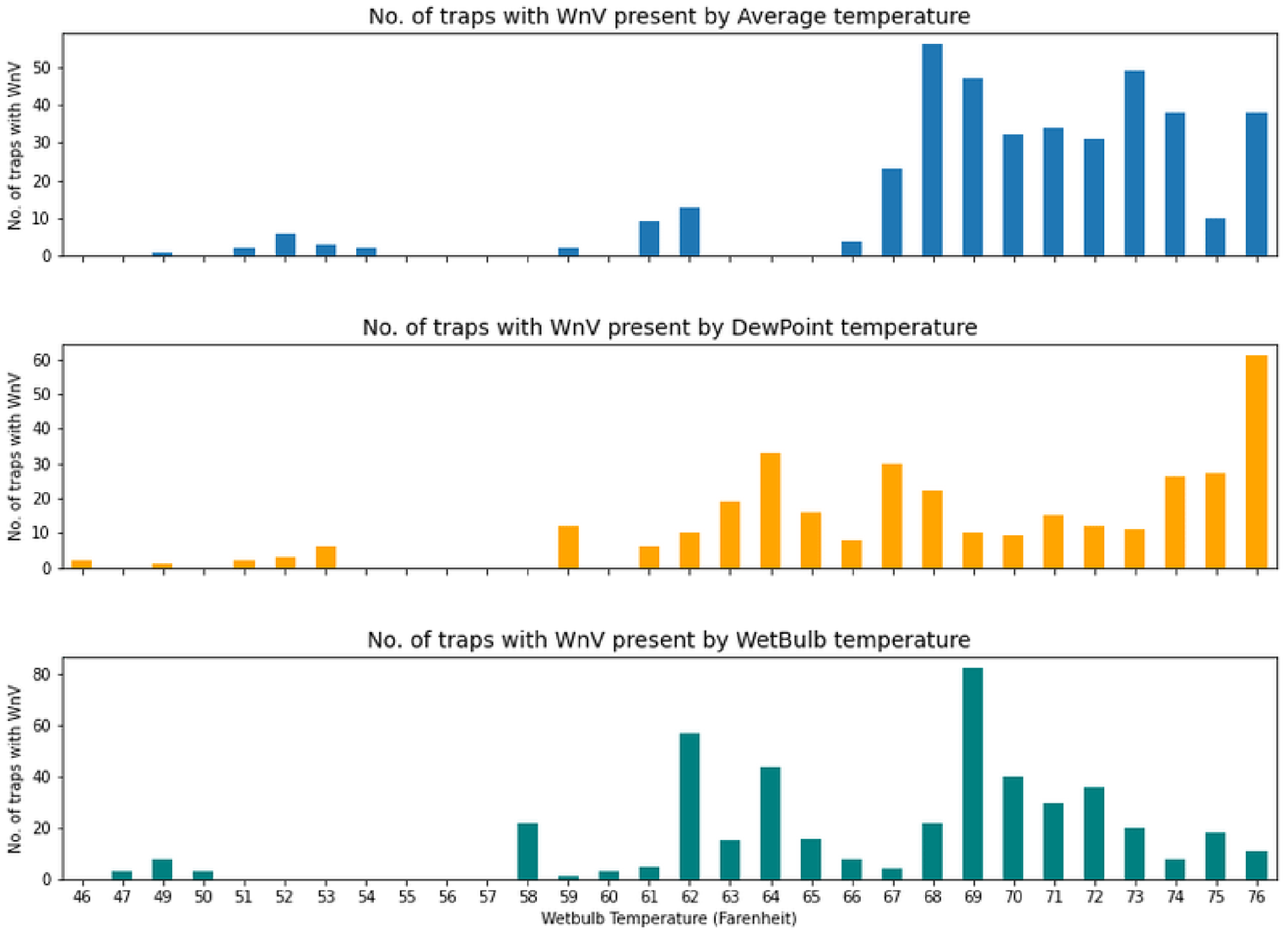


Temperature vs WNV Occurrence by year

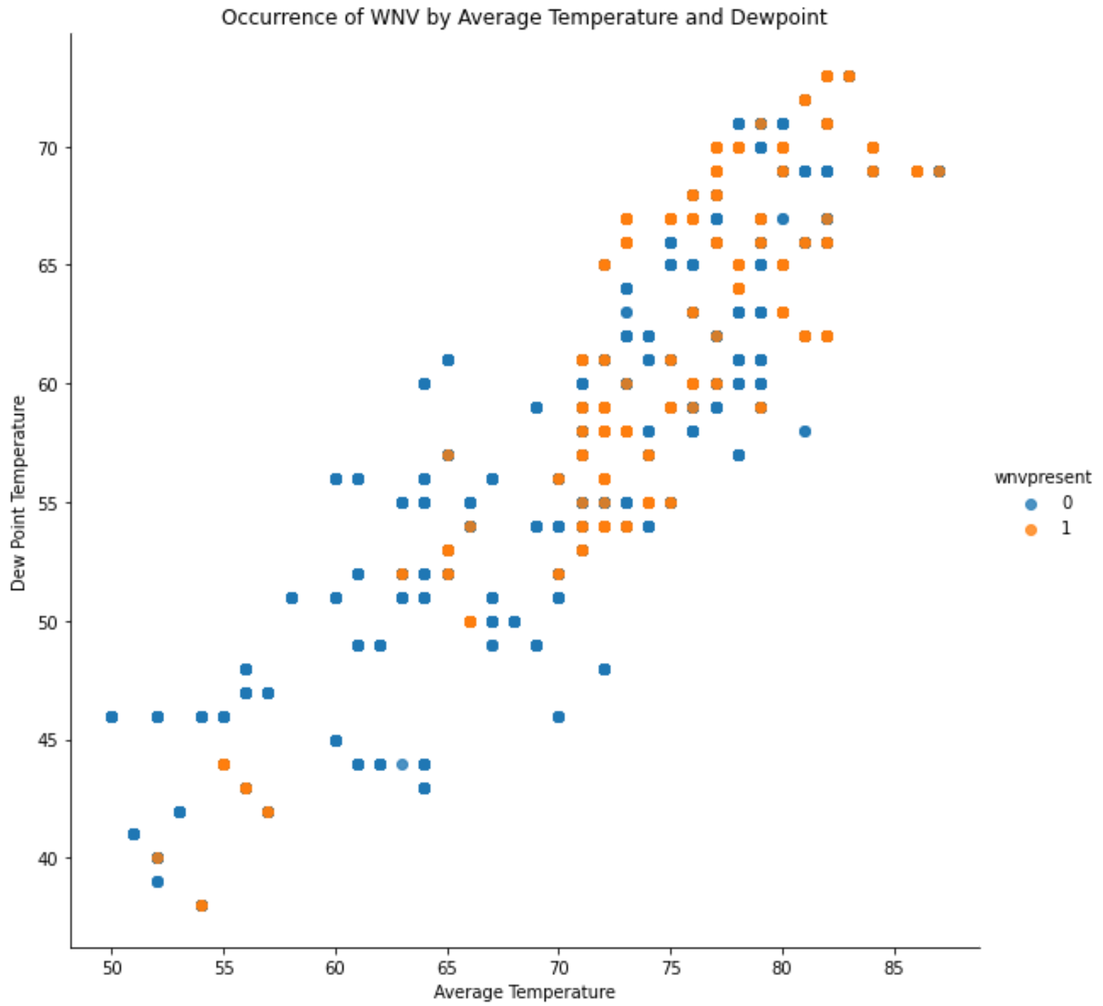




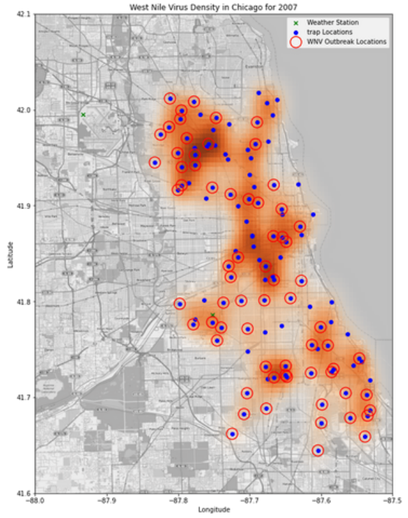
Presence of WNV in traps by Average, DewPoint and WetBulb temperatures

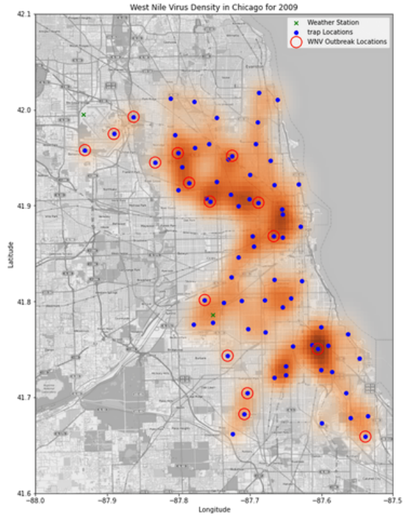


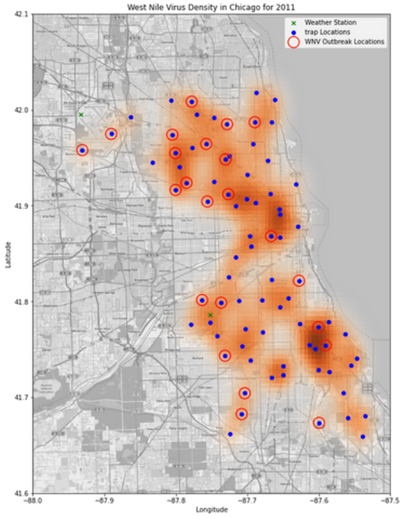
No. of traps with WNV in by Average and Dewpoint Temperatures

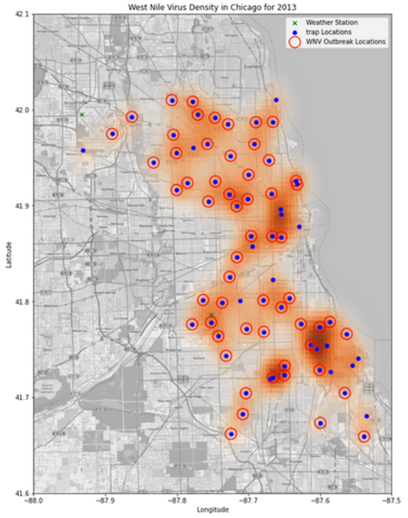


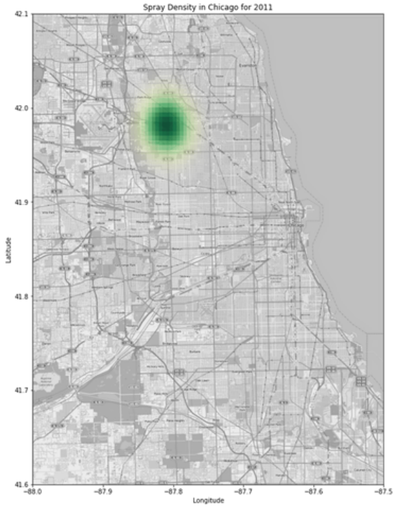
WNV Density in Chicago

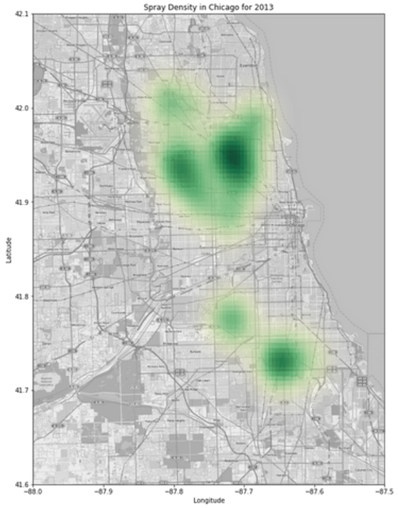




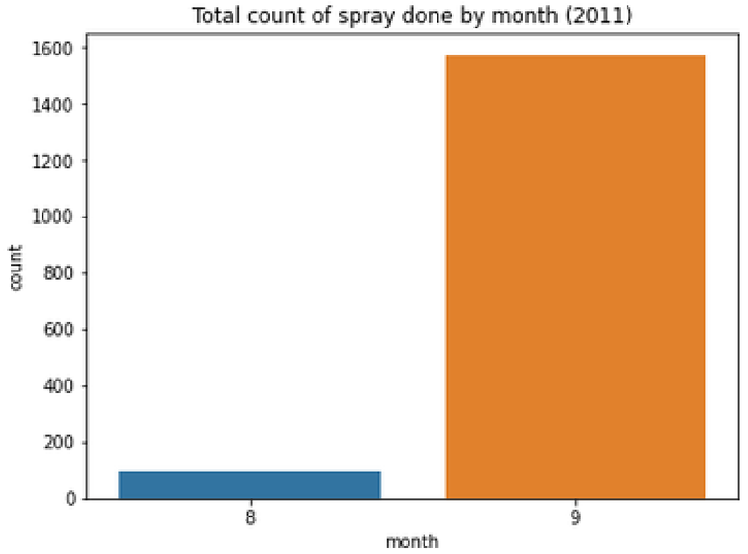


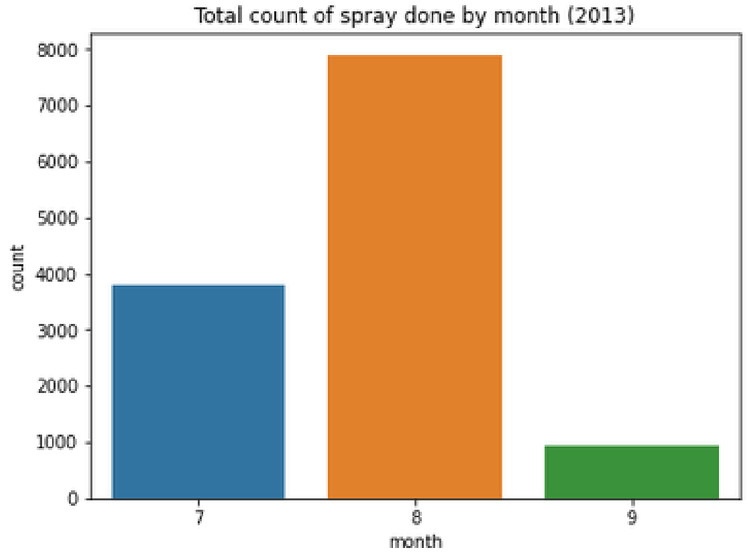


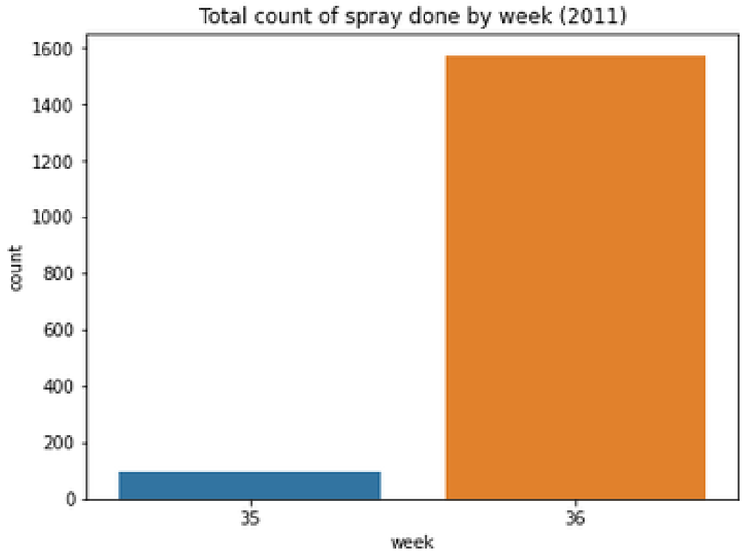


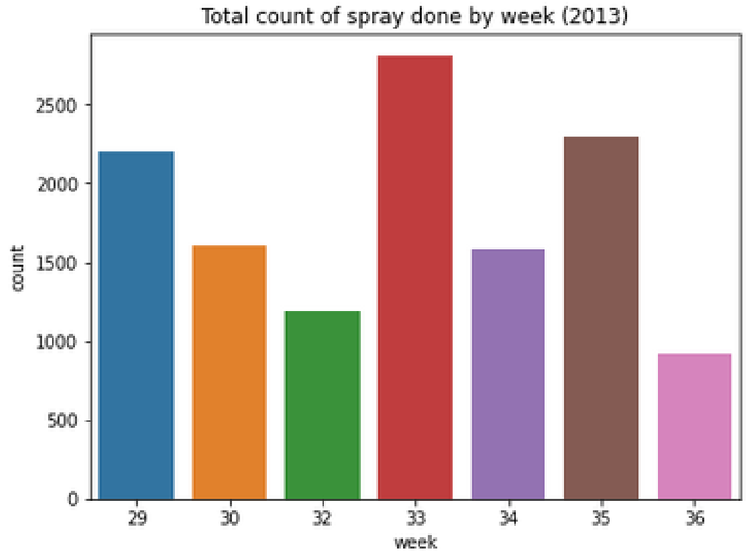


Spray Distribution by Year, Month, and Week









Part 3:

Modelling

Data Modelling Process

06 *Data Modelling*

ML(classification) Models with Hyperparameters

05 *Preprocessing*

SMOTE, StandardScaler

04 *Feature Selection*

Select top 40 features correlated with Wnvpresent

03 *Handling Multi-Colinearity*

Drop correlated features - week, tmin, tmax, dewpoint, wetbulb, sealevel, resultspeed

02 *One Hot Encoding*

Species, Traps

01 *Feature Engineering*

Relative Humidity, Handling Codesum, Polynomial Features

01

Feature Engineering

- Calculate Relative Humidity based on Dew Point
- Polynomial Features - Using top 5 most correlated features in the dataset
  - WetBulb * Sunrise
  - Week * WetBulb
  - Month * WetBulb
  - Week* Dewpoint
  - Month*DewPoint

04

Feature Selection

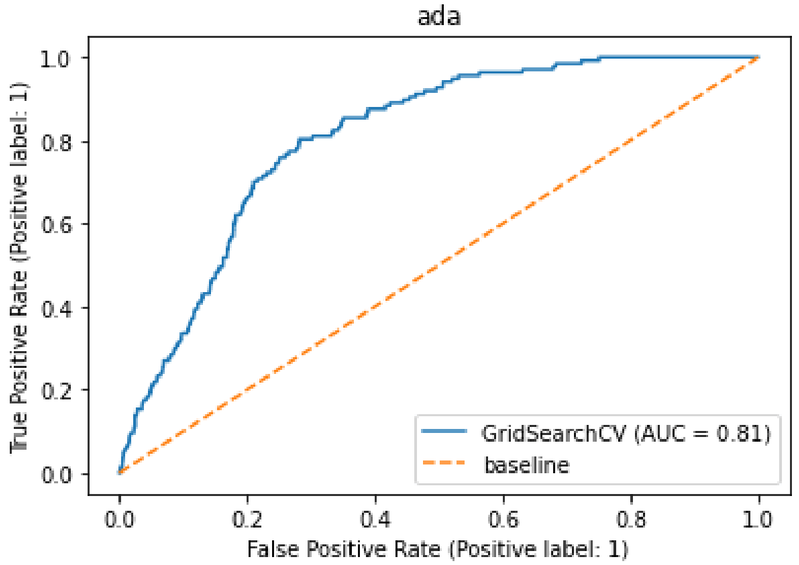
- Find correlation of all features with WNV Present
- Select top 40 features out of all features

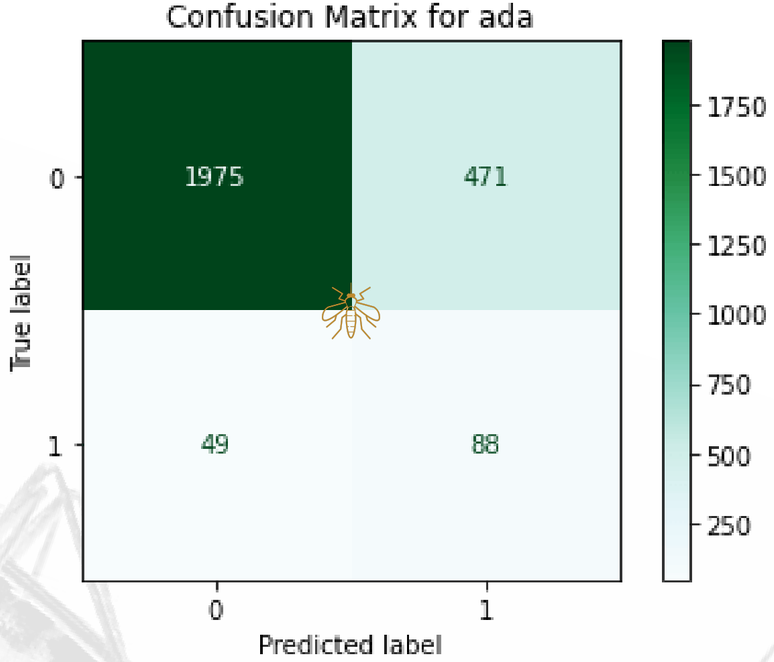
Machine Learning (Classification) Models

Models	Hyperparameters	Recall	ROC AUC
Logistic Regression	C: 4.8 max_iter: 5000 n_jobs: 3 penalty: l2	0.4817	0.7065
K Nearest Neighbors	n_neighbors: 3	0.4233	0.0758
Random Forest	max_depth: None n_estimators: 120	0.3430	0.8032
AdaBoost	learning_rate: 0.5 n_estimators: 200	0.6423	0.8082
Gradient Boosting	learning_rate: 0.5 n_estimators: 200	0.2487	0.8330

Part 4:

Modelling Evaluation



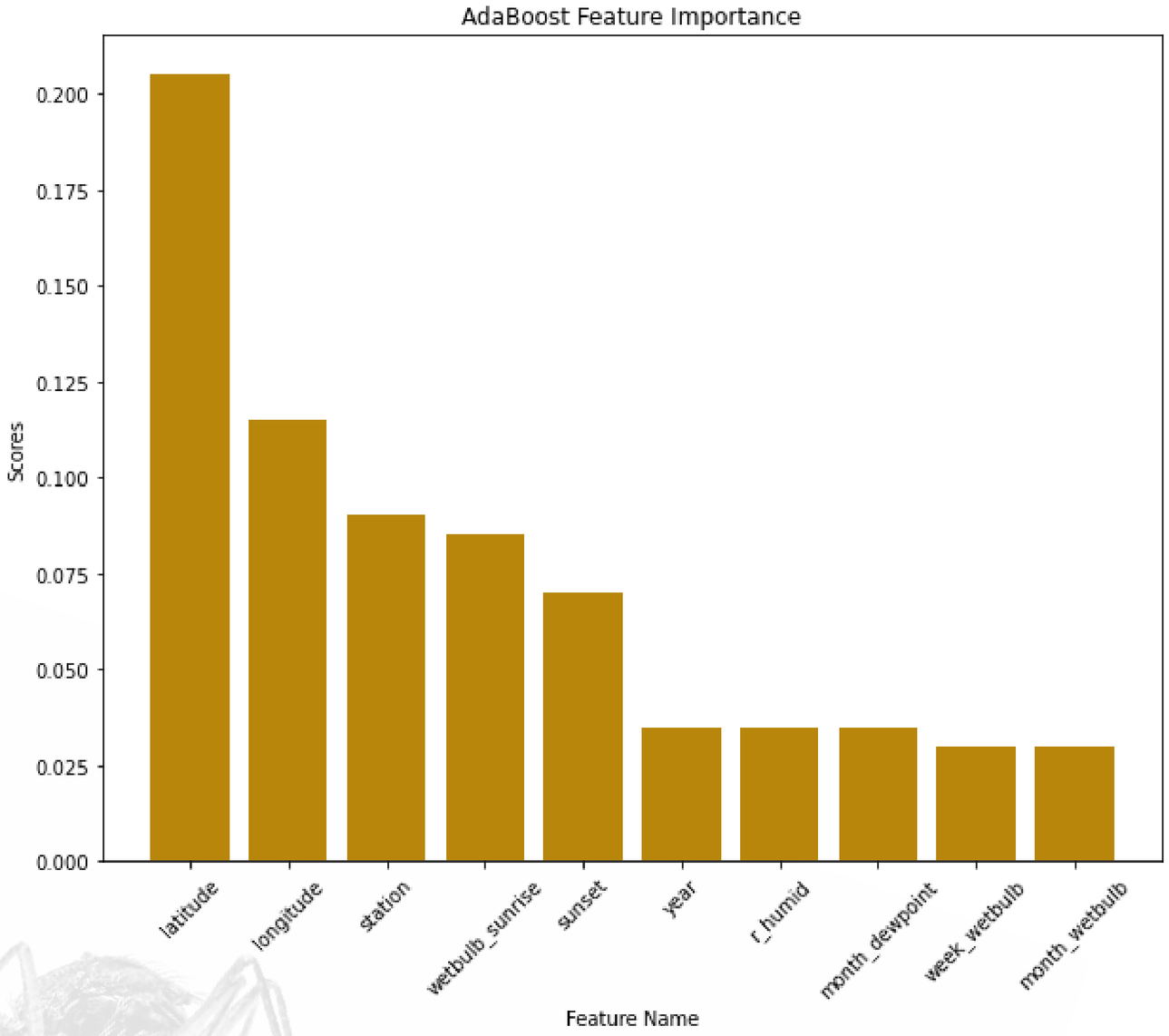


Modelling Evaluation

Choosing best algorithm based on ROC AUC and Recall scores

The AUC - ROC curve is a performance measurement for the classification problems threshold settings of 0.5

The recall is the measure of our model correctly identifying True Positives



Feature Importance

Relative importance of each feature when making a prediction

Importance are mainly based on:

- Locations
- Temperature
- Humidity

Part 5:

Conclusion & Recommendations

Conclusion

CDPH | Department of Data Science

'Best' model to predict WNV present:

- AdaBoost model

WNV is prevalent depending on:

- Location
- Higher average temperature and humidity
    - **An environment where mosquitoes thrive!**

Recommendation

CDPH | Department of Data Science

Utilise the model to implement targeted mosquito control measures:

- To date, there is no evidence of human-to-human transmission
- With predictive modelling, decision-makers can be guided to treat targeted areas
- For e.g. our model predicted the following WNV occurrences (i.e. 15% on average)

Year	Total locations	Number of WNV cases predicted	Percentage of cases
2008	30,498	3,806	13%
2010	36,557	4,408	12%
2012	27,115	4,866	18%
2014	22,123	4,050	18%

Cost Benefit Analysis

CDPH | Department of Data Science

- Since first discovered in 1999, WNV has cost ~$778 million in both medical cost and lost in productivity, i.e. $21,000 per case*

- With the help of predictive modelling, decision makers can be guided to deploy mosquito control measures

Cost (without modelling)
Treatment cost of
$365,000**

Benefit (with modelling)

- Treatment cost of **$55,000**** (i.e. ~ 15% based on predicted WNV occurrence)

- The remaining **85%** of the budget can be freed up for other critical use

Note:
**Source: www.sciencedaily.com*
*** Cost based on per acre per year with the assumption that mosquito treatment cost ~$1,000 per acre as quoted in www.fixr.com*

Other Recommendations

CDPH | Department of Data Science

More Control Measures

Regular checks on mosquito breeding areas (weekly & fortnightly)

Monitoring of potential hotspots and their surrounding environment

Removal of mosquito breeding areas

R&D Investment

Researching on methods and technologies that may help potentially suppress mosquitoes breeding

Example:

- ***Aedes aegypti experiment***

Community Participation

The community can play a role in reporting potential breeding areas to the Department of Public Health.

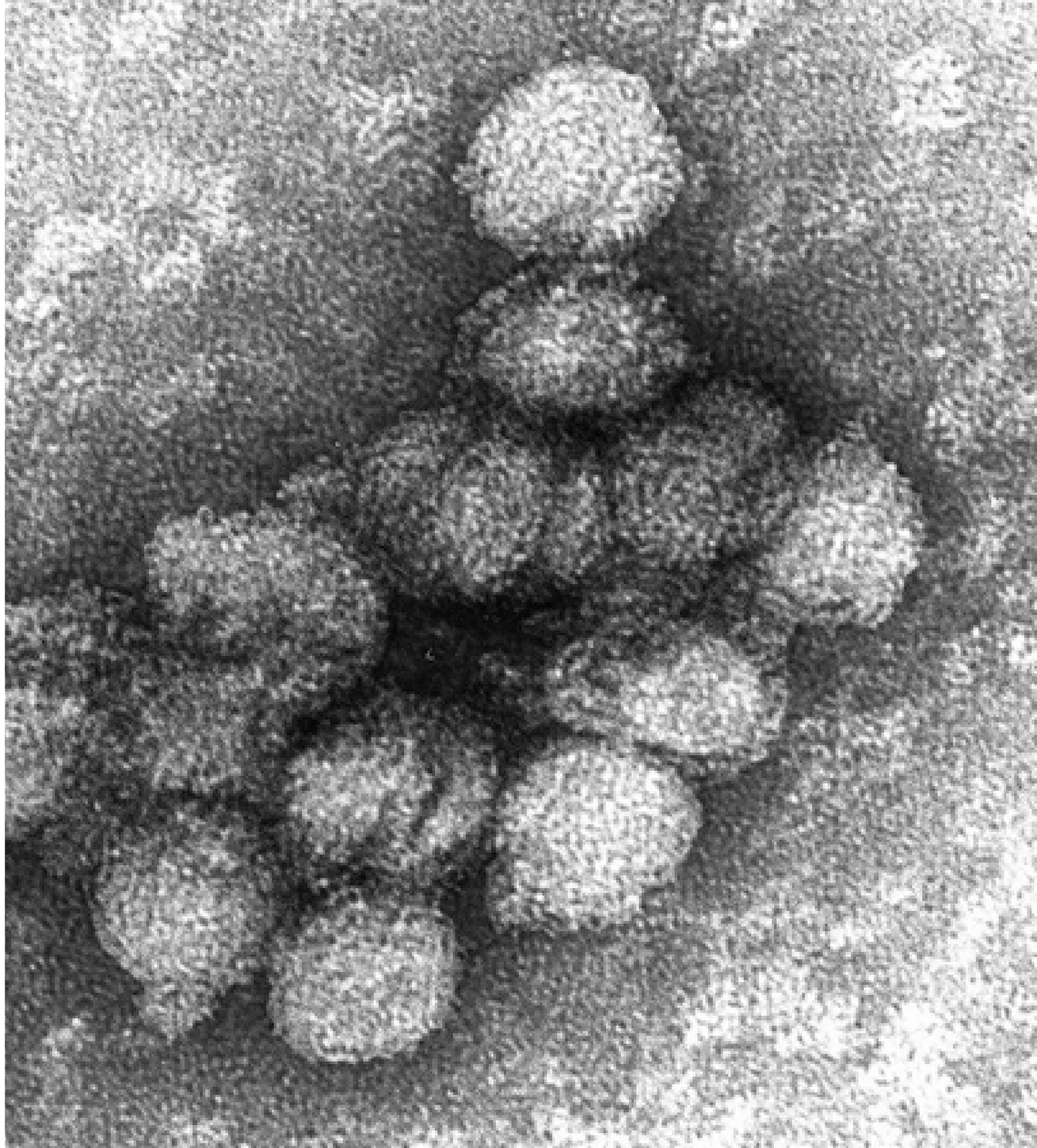
Educating the community on severity of the West Nile virus and how they can protect themselves

Department of
Data Science

Thank you

Contact Us

For any questions or
clarifications on our report.



Appendix 1 - Severity of the WNV

“ In about 80% of infected people have few to no sign of symptoms”

“ Recovery in severe cases may take weeks to months ”

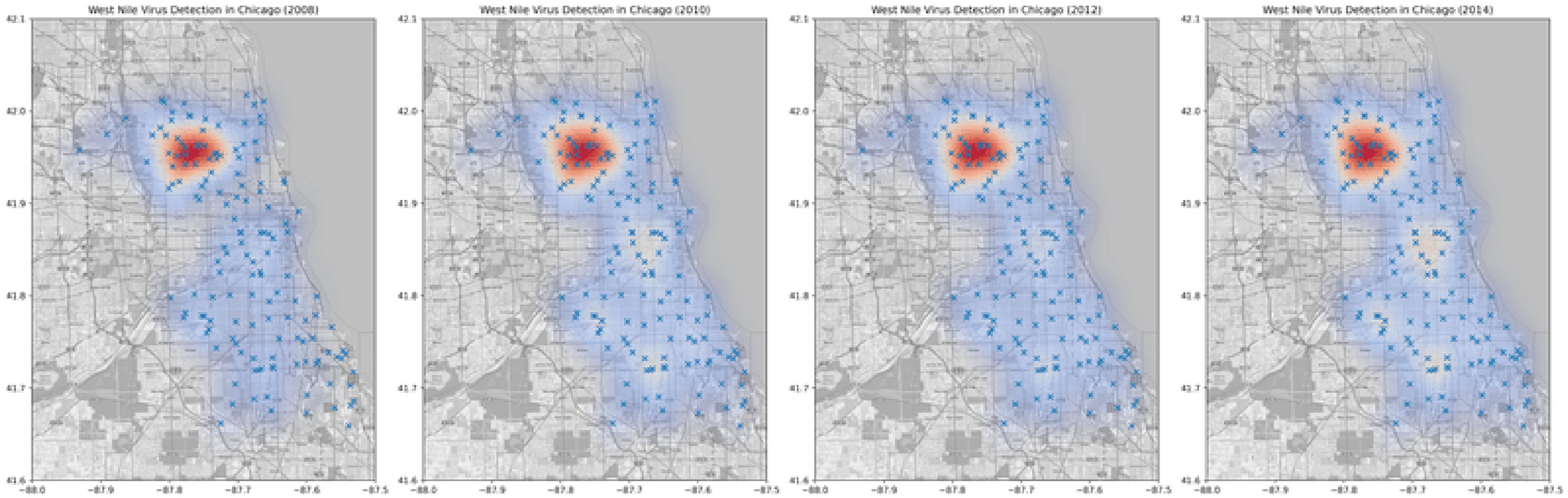
“ The risk of death among those in whom the nervous system is affected is about 10%. ”

No vaccine for humans

Appendix 2

CDPH | Department of Data Science

Predictions of WNV present:



Hotspot areas:

West Ohare Airport, West Higgins Avenue, West Foster Avenue, North Pittsburgh Avenue, West Armitage Avenue