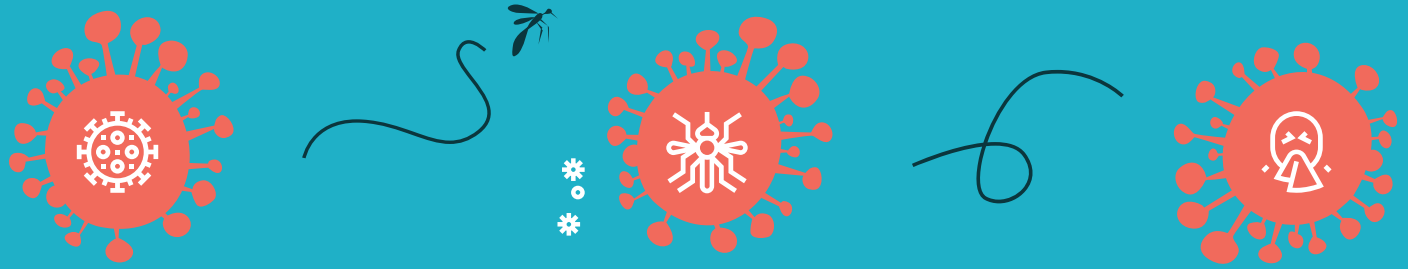
Project 4
West Nile Virus Prediction

Table of contents

01.
About the disease

West Nile Virus

About the disease



What?

- Single-stranded RNA virus that causes West Nile fever
- Leading cause of mosquito-borne disease in US

Seasonality

- Cases occur during mosquito season.
- Starts in summer, continues through fall

Transmission

Mosquitoes become infected when they feed on infected birds then spread it to other animals and humans through mosquito bite

Symptoms of the disease

80%
No symptoms

Most infected people show no symptoms

20%
Febrile illness (Fever)

Headache, body aches, joint pains, vomiting, diarrhea or rash.

1%
Serious Symptoms

High fever, neck stiffness, stupor, disorientation, coma, convulsions, tremors, muscle weakness, vision loss, numbness and paralysis.

Fatal

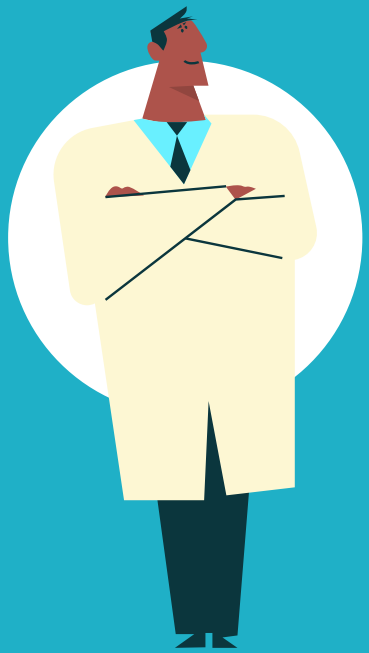
Affect central nervous system: encephalitis(inflamation of the brain) or Meningitis.

Treatment

01

Medications?

- No specific medications for WNV
- Antibiotics do not treat virus
- Rest, fluids and OTC pain medications may relieve some symptoms



02

Vaccines?

No vaccines available

Problem Statement

The west nile virus is a mosquito-borne disease plaguing the Chicago area. It results in multiple costs including but not limited to:

1. Medical treatment required;

2. Lost productivity as people miss work due to being ill; and

3. Pain and suffering of people afflicted by it.

There may be a way to reduce the costs inflicted by the virus by targeting mosquito populations with pesticides or otherwise. However, we should target mosquito populations when west nile virus is most prevalent.

Objectives

1. Build a model which predicts when west nile virus is present in mosquito populations.

2. If the model predicts the presence of west nile virus, action can be taken to reduce or exterminate mosquito populations.

Scope

1. Use the data to feed various machine learning models to achieve the objectives.

2. Select the model with the best ROC-AUC score.

Success Metrics and Targets

01

AUC Score

Predict the time and location at which various mosquito species will be detected as carriers of the West Nile virus.

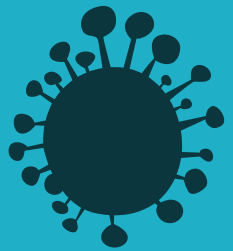
02

Recommendations

Propose measures to reduce the transmission of West Nile virus in the specified regions.

02.
Processing the Data

(OOP is OP)

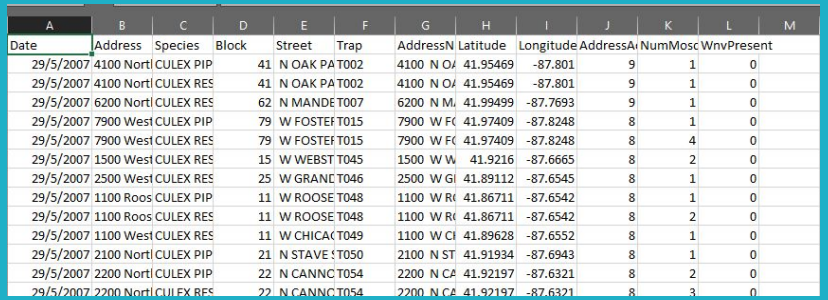


Dataset (Kaggle)

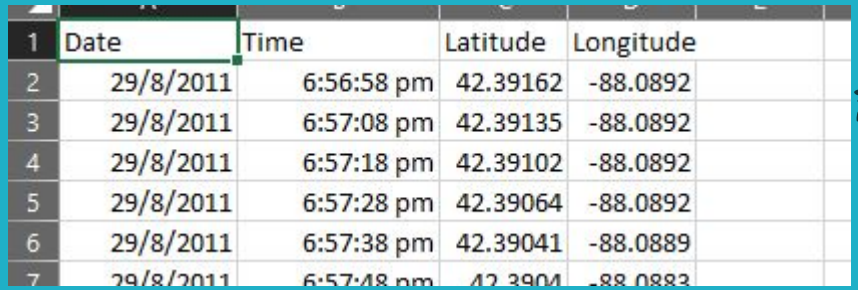
Dataset	Features	Years
Weather	temperature, weather codes, precipitation, pressure, wind, and other related information, along with station and date details.	2007-2014
Spray	date, time, latitude and longitude	2011 & 2013
Train	date, address, mosquito species, location details, trap data, and the presence of West Nile Virus (WNV) in mosquitoes.	2007, 2009, 2011, 2013
Test	Same as Train. Excluding Number of Mosquitos and presence of WNV	2008, 2010, 2012, 2014

Processing the Data

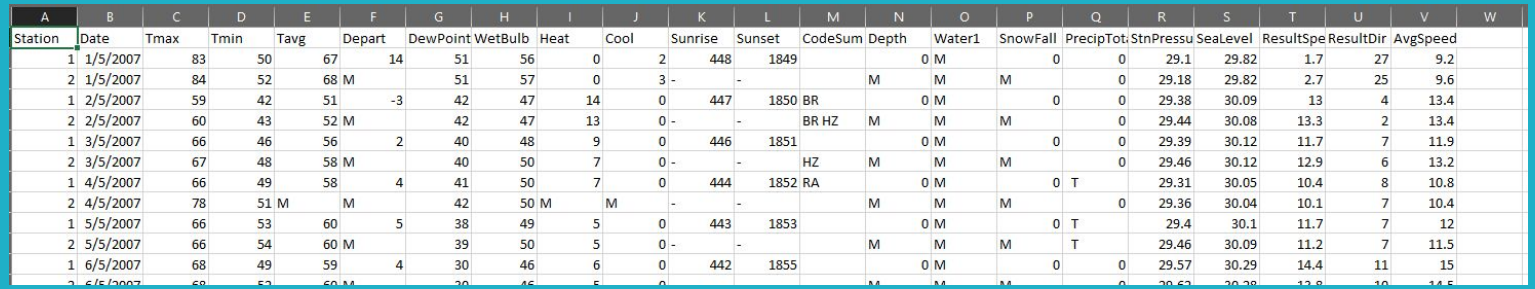
The data did not arrive neatly packaged:



Date	Address	Species	Block	Street	Trap	AddressN	Latitude	Longitude	AddressA	NumMosc	WnvPresent
29/5/2007	4100 North	CULEX PIP	41	N OAK PA	T002	4100 N O/	41.95469	-87.801	9	1	0
29/5/2007	4100 North	CULEX RES	41	N OAK PA	T002	4100 N O/	41.95469	-87.801	9	1	0
29/5/2007	6200 North	CULEX RES	62	N MANDE	T007	6200 N M/	41.99499	-87.7693	9	1	0
29/5/2007	7900 West	CULEX PIP	79	W FOSTEI	T015	7900 W F(	41.97409	-87.8248	8	1	0
29/5/2007	7900 West	CULEX RES	79	W FOSTEI	T015	7900 W F(	41.97409	-87.8248	8	4	0
29/5/2007	1500 West	CULEX RES	15	W WEBST	T045	1500 W W	41.9216	-87.6665	8	2	0
29/5/2007	2500 West	CULEX RES	25	W GRANI	T046	2500 W G	41.89112	-87.6545	8	1	0
29/5/2007	1100 Roos	CULEX PIP	11	W ROOSE	T048	1100 W R(	41.86711	-87.6542	8	1	0
29/5/2007	1100 Roos	CULEX RES	11	W ROOSE	T048	1100 W R(	41.86711	-87.6542	8	2	0
29/5/2007	1100 West	CULEX RES	11	W CHICA(	T049	1100 W C(	41.89628	-87.6552	8	1	0
29/5/2007	2100 North	CULEX PIP	21	N STAVE	T050	2100 N ST	41.91934	-87.6943	8	1	0
29/5/2007	2200 North	CULEX PIP	22	N CANNC	T054	2200 N C/	41.92197	-87.6321	8	2	0
29/5/2007	2200 North	CULEX RES	22	N CANNC	T054	2200 N C/	41.92197	-87.6321	8	3	0



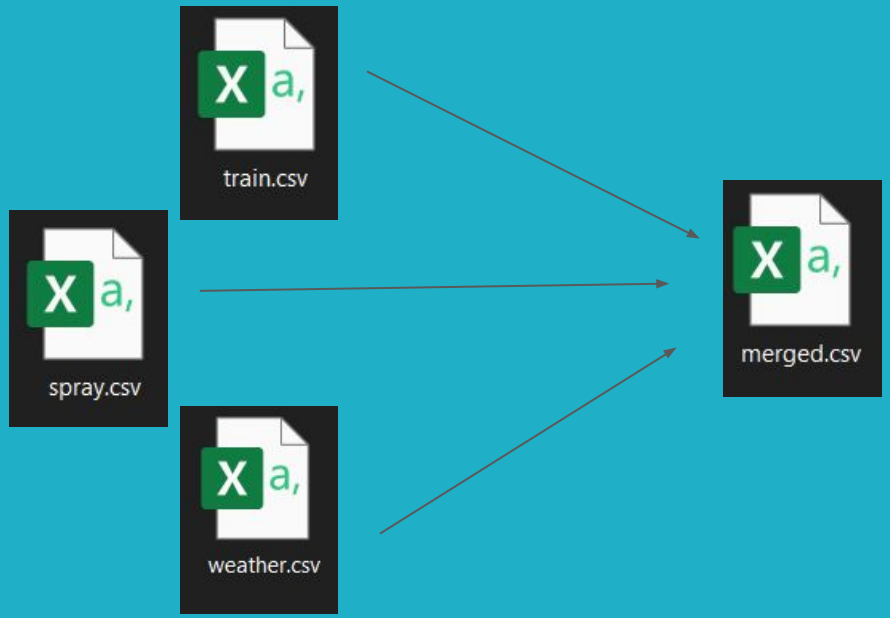
Date	Time	Latitude	Longitude
29/8/2011	6:56:58 pm	42.39162	-88.0892
29/8/2011	6:57:08 pm	42.39135	-88.0892
29/8/2011	6:57:18 pm	42.39102	-88.0892
29/8/2011	6:57:28 pm	42.39064	-88.0892
29/8/2011	6:57:38 pm	42.39041	-88.0889
29/8/2011	6:57:48 pm	42.3904	-88.0883



Station	Date	Tmax	Tmin	Tavg	Depart	DewPoint	WetBulb	Heat	Cool	Sunrise	Sunset	CodeSum	Depth	Water1	SnowFall	PrecipTot	StnPressu	SeaLevel	ResultSpe	ResultDir	AvgSpeed
1	1/5/2007	83	50	67	14	51	56	0	2	448	1849		0	M	0	0	29.1	29.82	1.7	27	9.2
2	1/5/2007	84	52	68	M	51	57	0	3	-	-		M	M	M	0	29.18	29.82	2.7	25	9.6
1	2/5/2007	59	42	51	-3	42	47	14	0	447	1850	BR	0	M	0	0	29.38	30.09	13	4	13.4
2	2/5/2007	60	43	52	M	42	47	13	0	-	-	BR HZ	M	M	M	0	29.44	30.08	13.3	2	13.4
1	3/5/2007	66	46	56	2	40	48	9	0	446	1851		0	M	0	0	29.39	30.12	11.7	7	11.9
2	3/5/2007	67	48	58	M	40	50	7	0	-	-	HZ	M	M	M	0	29.46	30.12	12.9	6	13.2
1	4/5/2007	66	49	58	4	41	50	7	0	444	1852	RA	0	M	0	T	29.31	30.05	10.4	8	10.8
2	4/5/2007	78	51	M	M	42	50	M	M	-	-		M	M	M	0	29.36	30.04	10.1	7	10.4
1	5/5/2007	66	53	60	5	38	49	5	0	443	1853		0	M	0	T	29.4	30.1	11.7	7	12
2	5/5/2007	66	54	60	M	39	50	5	0	-	-		M	M	M	T	29.46	30.09	11.2	7	11.5
1	6/5/2007	68	49	59	4	30	46	6	0	442	1855		0	M	0	0	29.57	30.29	14.4	11	15

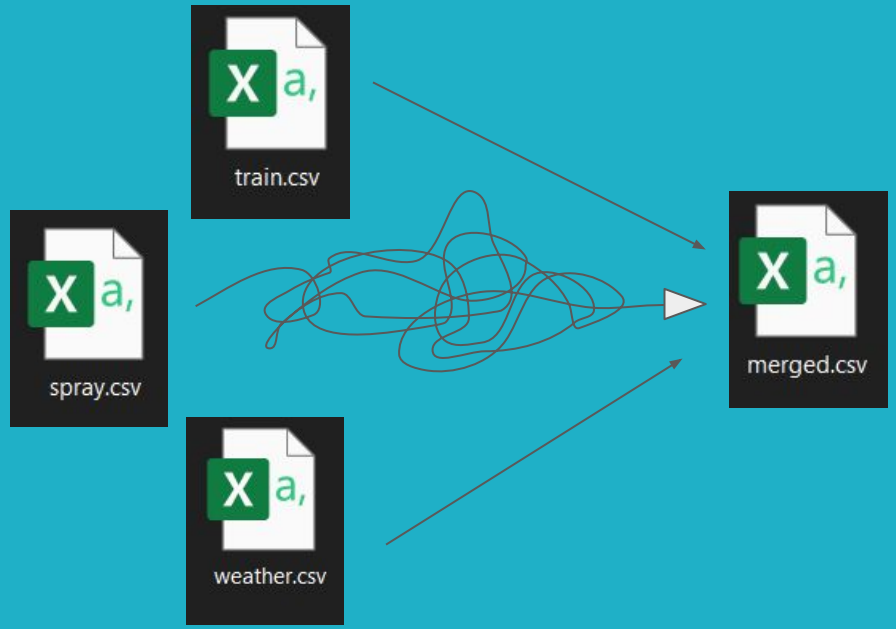
Processing the Data

It had to be merged:



Processing the Data

Well, not quite, it looked a lot more like this:



Processing the Data

Why? The pesticide in use is Zenivex

(https://www.cmmcp.org/pesticide-information/pages/zenivex-e4-etofenprox),

and should result in a lowered mosquito population for a few days. However, we have been reliably informed that mosquitoes can fly, *i.e.* mosquitoes from other areas can fly in. Therefore, we created a new column with the following criteria:

1. A set of coordinates with 1km of the spray coordinates is regarded as sprayed; and

2. Spraying must have occurred within 3 days prior to the trap being checked.

Processing the Data

Automating the cleaning process.

Why?

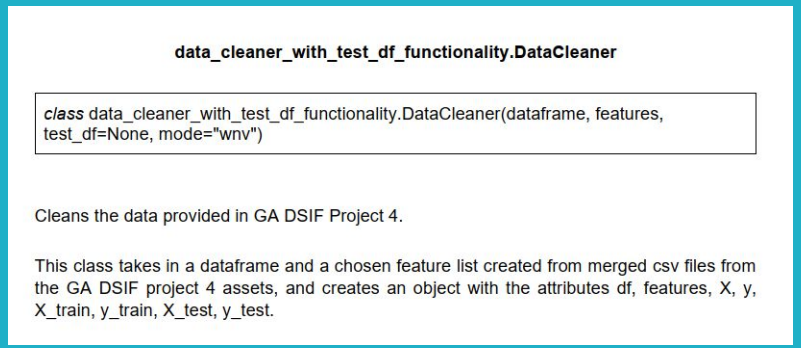
Mosquito control is an ongoing project, as new data comes, we do not want to repeat work.

Solution?

A data cleaner class that takes in:

1. Dataframe of training data;
2. Features to be used;
3. Dataframe of testing data (optional).

Fully documented!



data_cleaner_with_test_df_functionality.DataCleaner

class data_cleaner_with_test_df_functionality.DataCleaner(dataframe, features, test_df=None, mode="wnv")

Cleans the data provided in GA DSIF Project 4.

This class takes in a dataframe and a chosen feature list created from merged csv files from the GA DSIF project 4 assets, and creates an object with the attributes df, features, X, y, X_train, y_train, X_test, y_test.

Processing the Data

Where do we start?

First, we need to figure out which features to use.

An automated script was written to get the feature set which produced the best ROC-AUC score, which gave us a list of 13 features:

```
['Species', 'Tmax', 'Tmin', 'Tavg', 'DewPoint', 'WetBulb', 'Cool', 'PrecipTotal',
'ResultSpeed', 'ResultDir', 'StnPressure', 'SeaLevel', 'month']
```

Surprisingly, spraying was irrelevant, and in fact detrimental to ROC-AUC score.

Processing the Data

Then, an entire package of functions was written.

Each function:

1. If requiring the use of an sk-learn scaler, instantiated and saved a scaler for that column to fit on train data, to subsequently transform test data.

2. If operating on categorical variables, depends upon instantiating an instance of sk-learn's one hot encoder when the datacleaner object is instantiated. *I.e.*, the same fitted encoder can be applied against the optional test dataset if passed.

Why go through all the extra work

We could test a whole range of models, cleaning the training data in a single line of code, with another line to clean the test data.

If further data (in the same format) is collected for future analysis, the same class can be used with no changes.

03. Exploratory Data Analysis

Number of mosquitoes

135,039

Total Mosquitoes

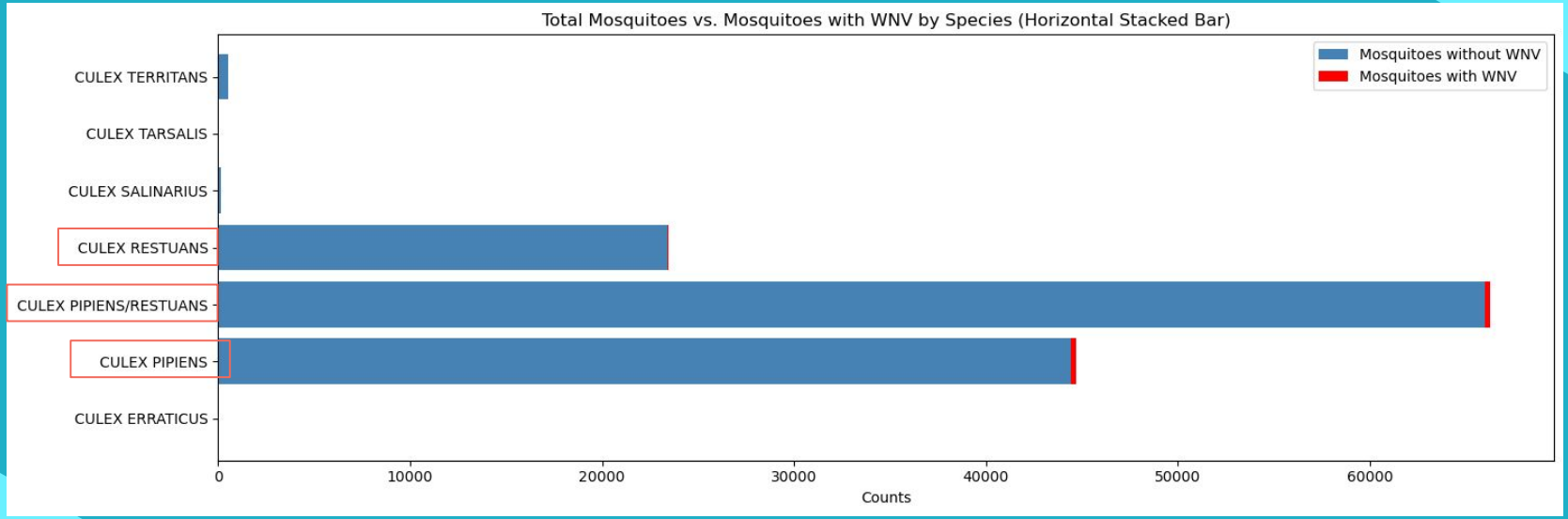
14,519

WNV present

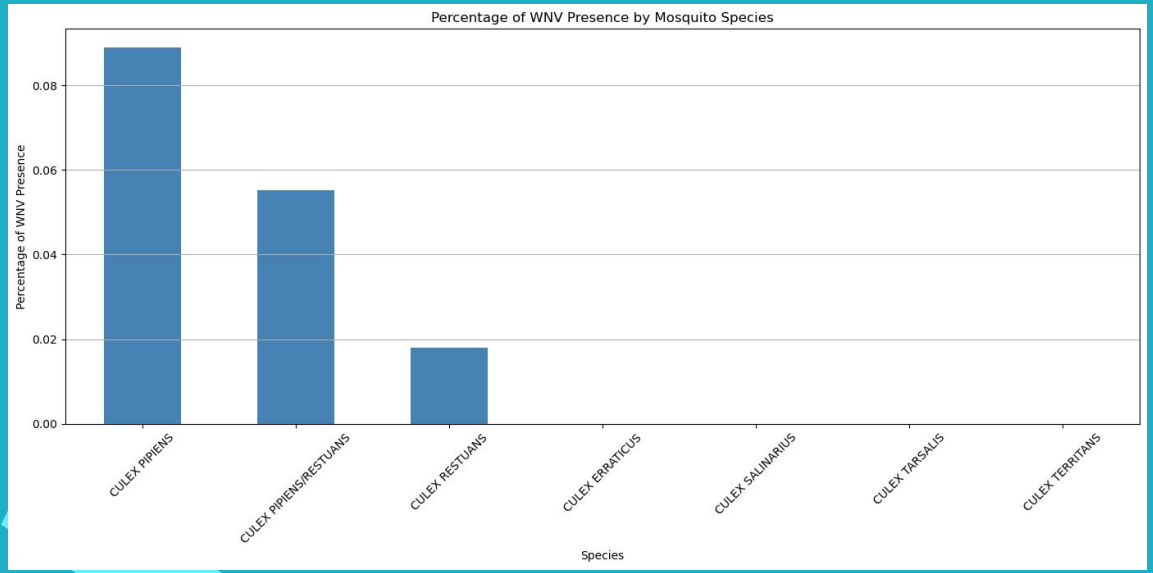
120,520

No WNV present

Only 2 out of 7 Species, WNV present



Which Species is the carrier?



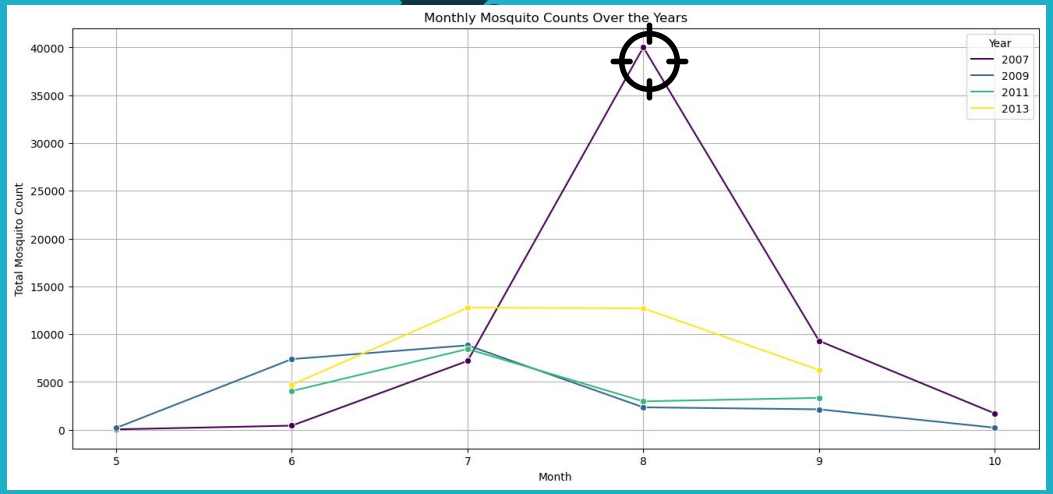
Culex Pipiens

Mosquito breed infected with the most WNV virus

Culex Restuans

Less likely to be infected

Mosquito Count



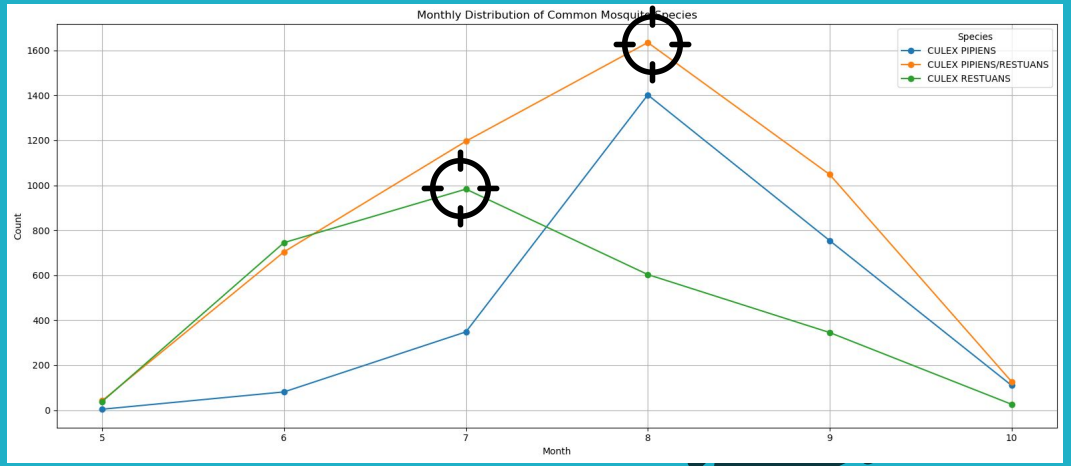
All Mosquito Species

Seasonal pattern:

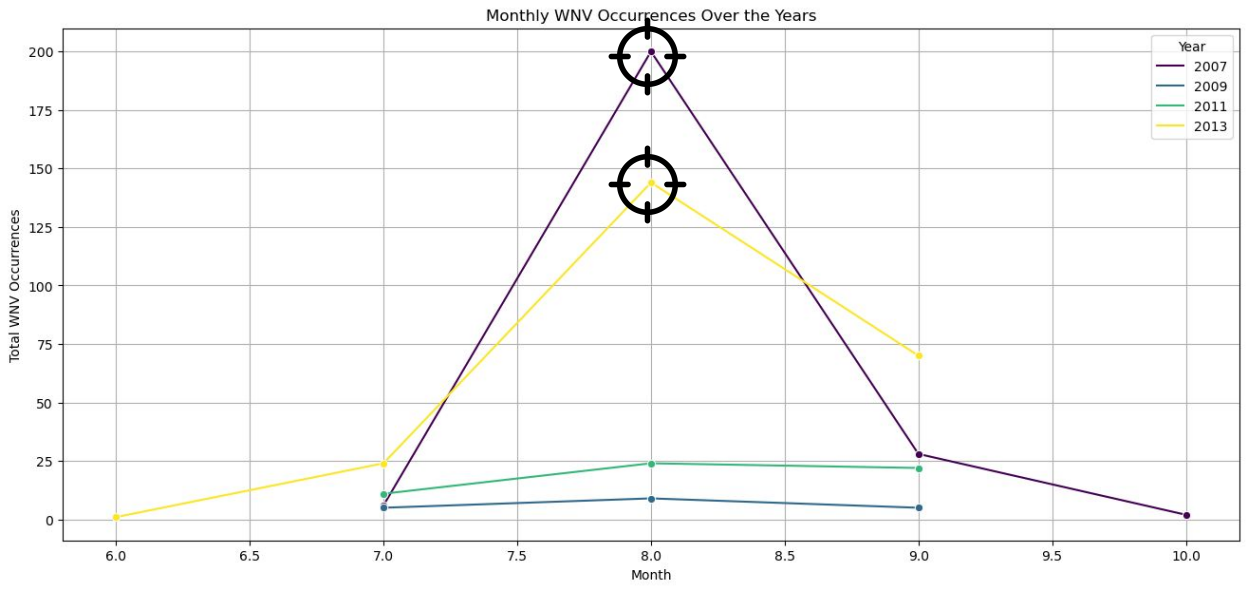
- rise in May,
- peak around July and August
- decrease through September and October.

WNV infected Mosquito

- Culex Pipiens: appear in May, peaks in August
- Culex Restuans: appear in May, but its peak is earlier, around June and July.
- Culex Pipiens/Restuans (Hybrid/Unspecified): consistent presence from June to August without a clear peak.
- Different species might have slightly different active periods.

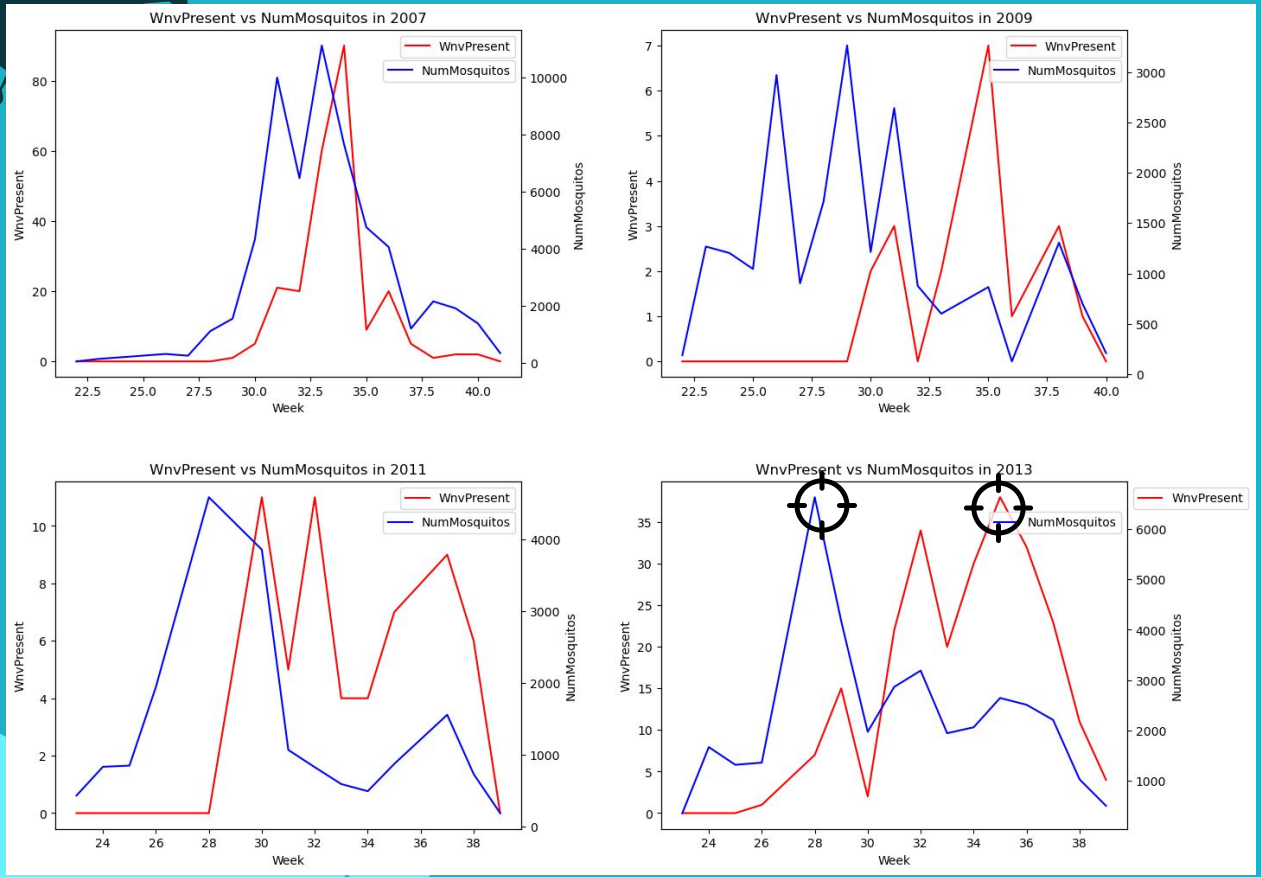


WNV Monthly Occurance



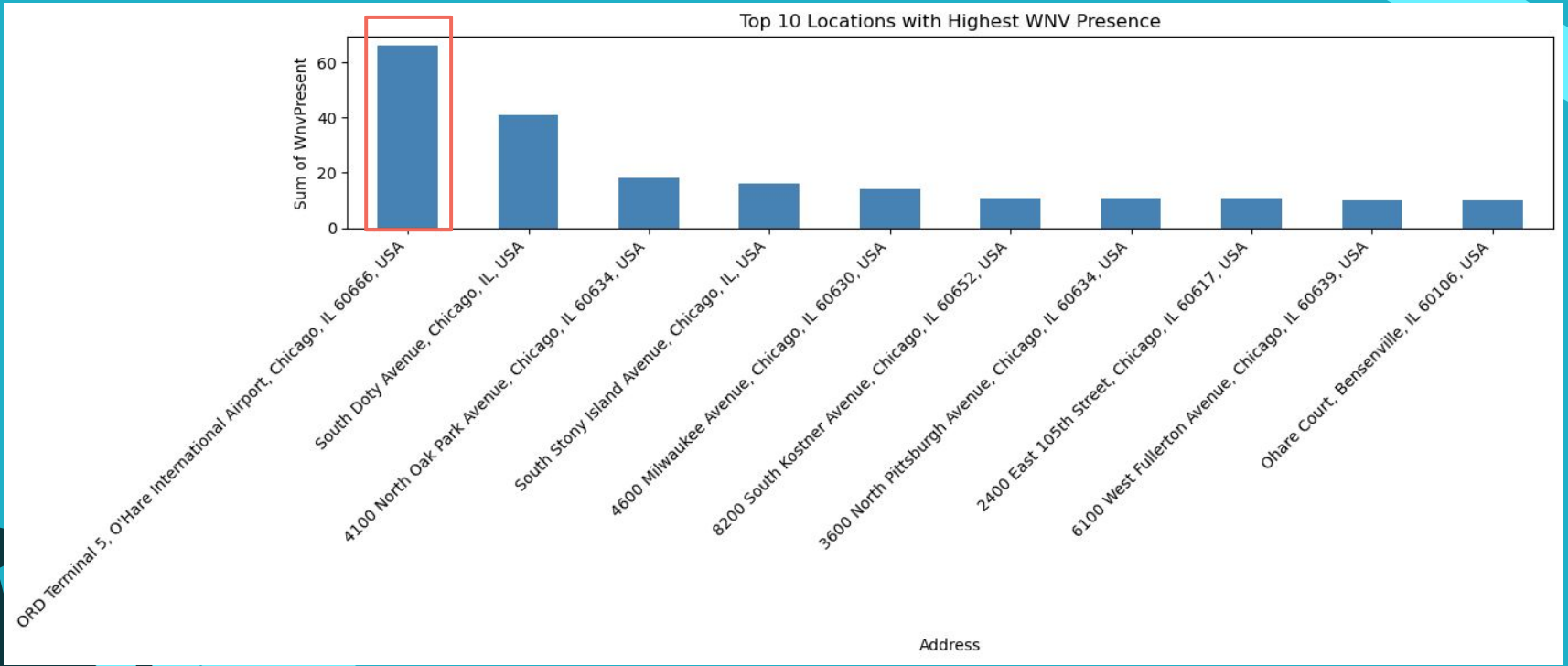
Common peak

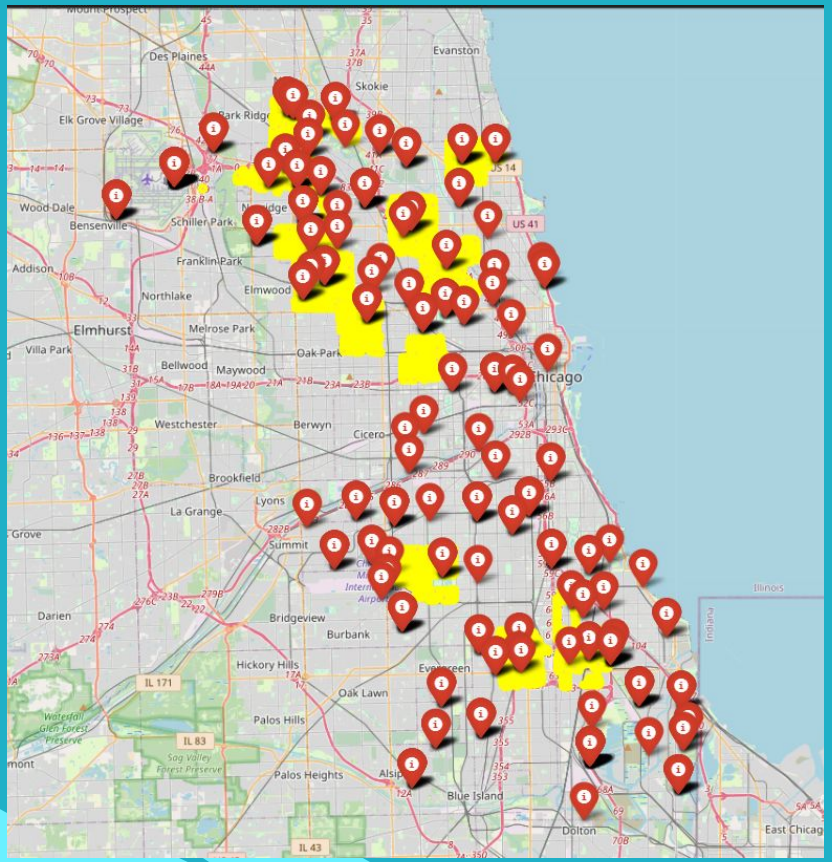
Combination of all species with infected WNV generally reaches its peak in August before it decreases



- Seasonal trend to mosquito population
- Lag observed between mosquito peaks and WNV detections aligns with the biological cycle of the virus, from acquisition from birds to human transmission and eventual detection after symptom onset

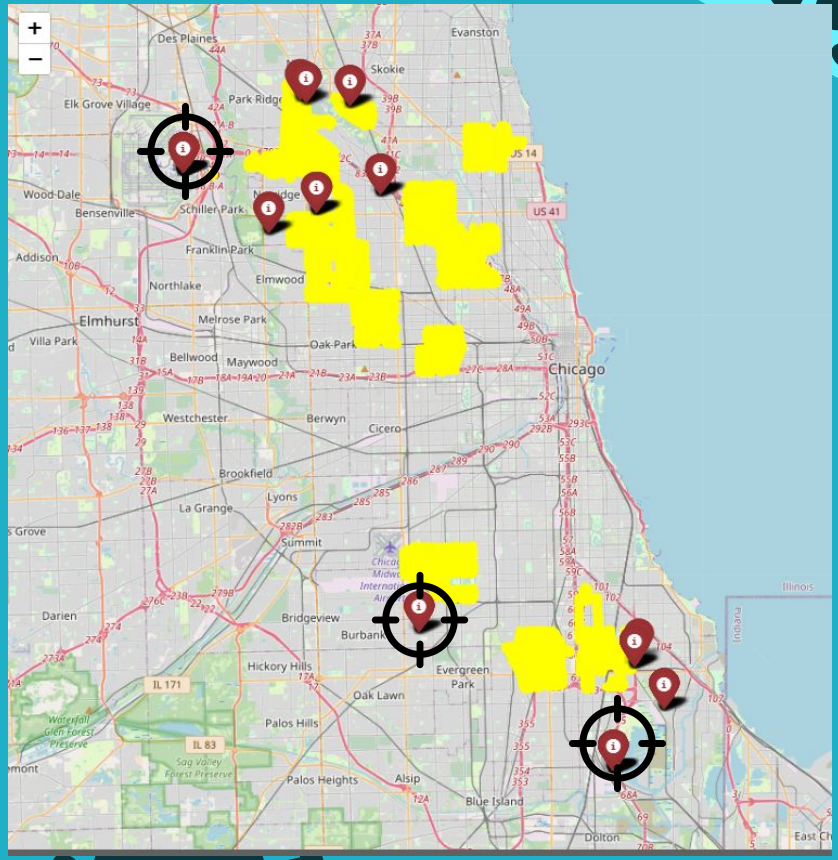
Top Locations with WNV present





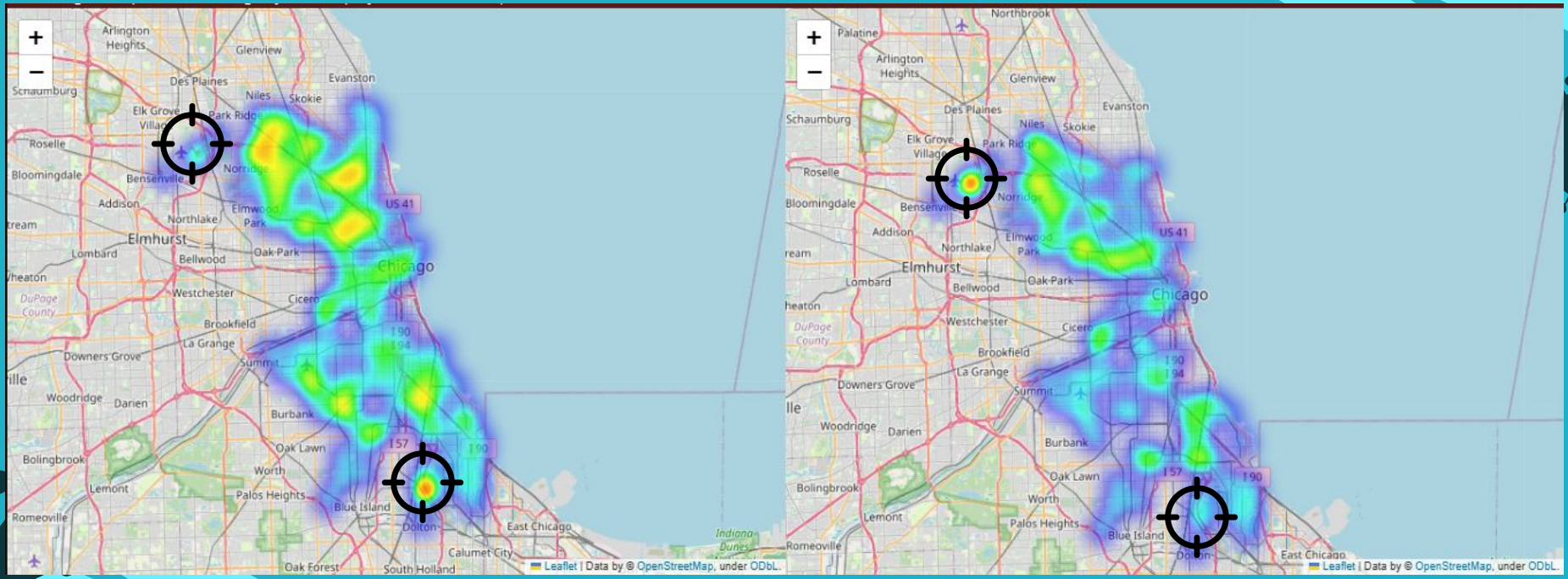
WNV Locations

Top 10 WNV Locations



Yellow-
Spray
area

Correlation ≠ Causation

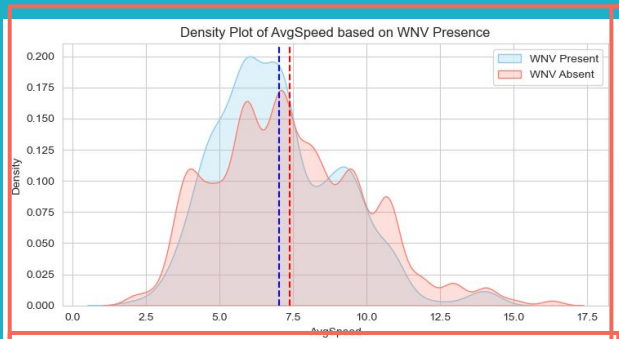


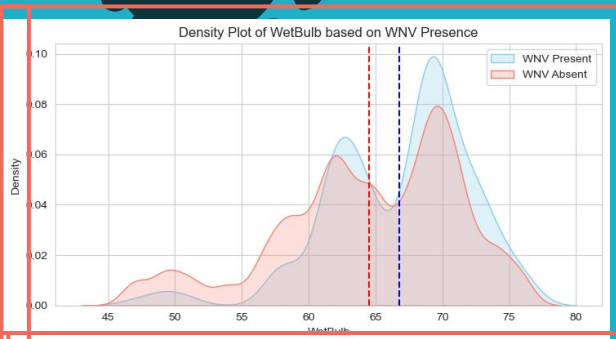
Mosquito areas

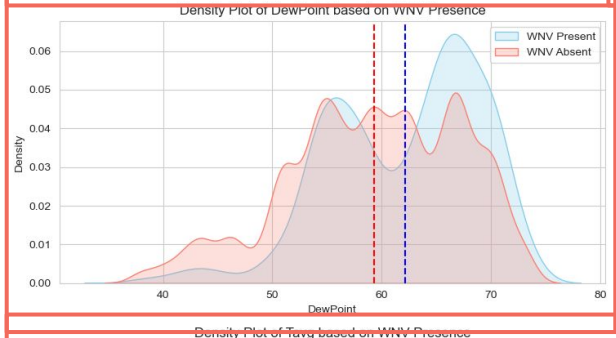
The brighter the area, the more mosquitoes in the area

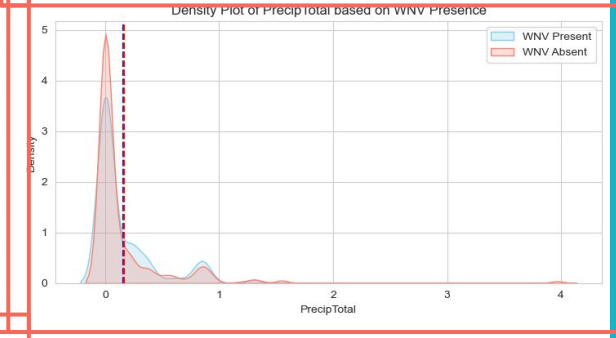
WNV infected Mosquito

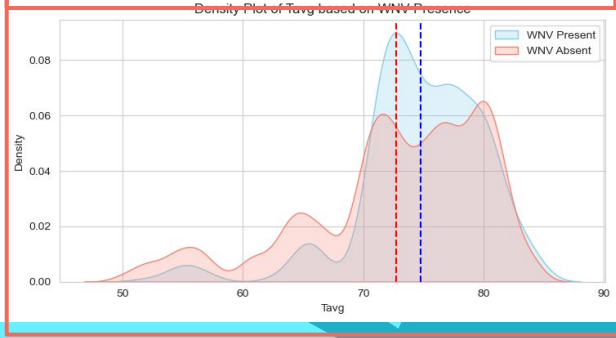
Areas with high WNV count are brighter in colour











Weather Corr

Seasonality

WNV more prevalent during warmer months

Humidity

Virus thrive in humid condition with higher Dew point

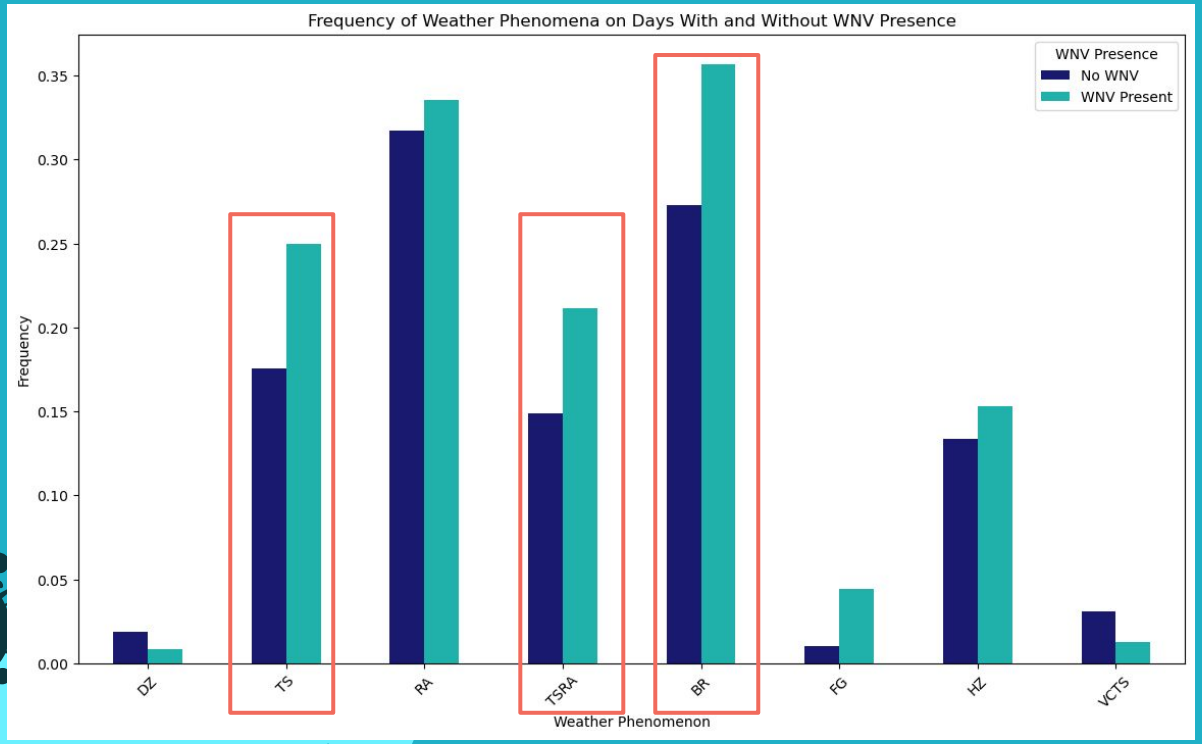
Wind Speed

Days with low wind speed, more WNV activity

Precipitation

Most days, regardless of virus presence, have little to no precipitation.

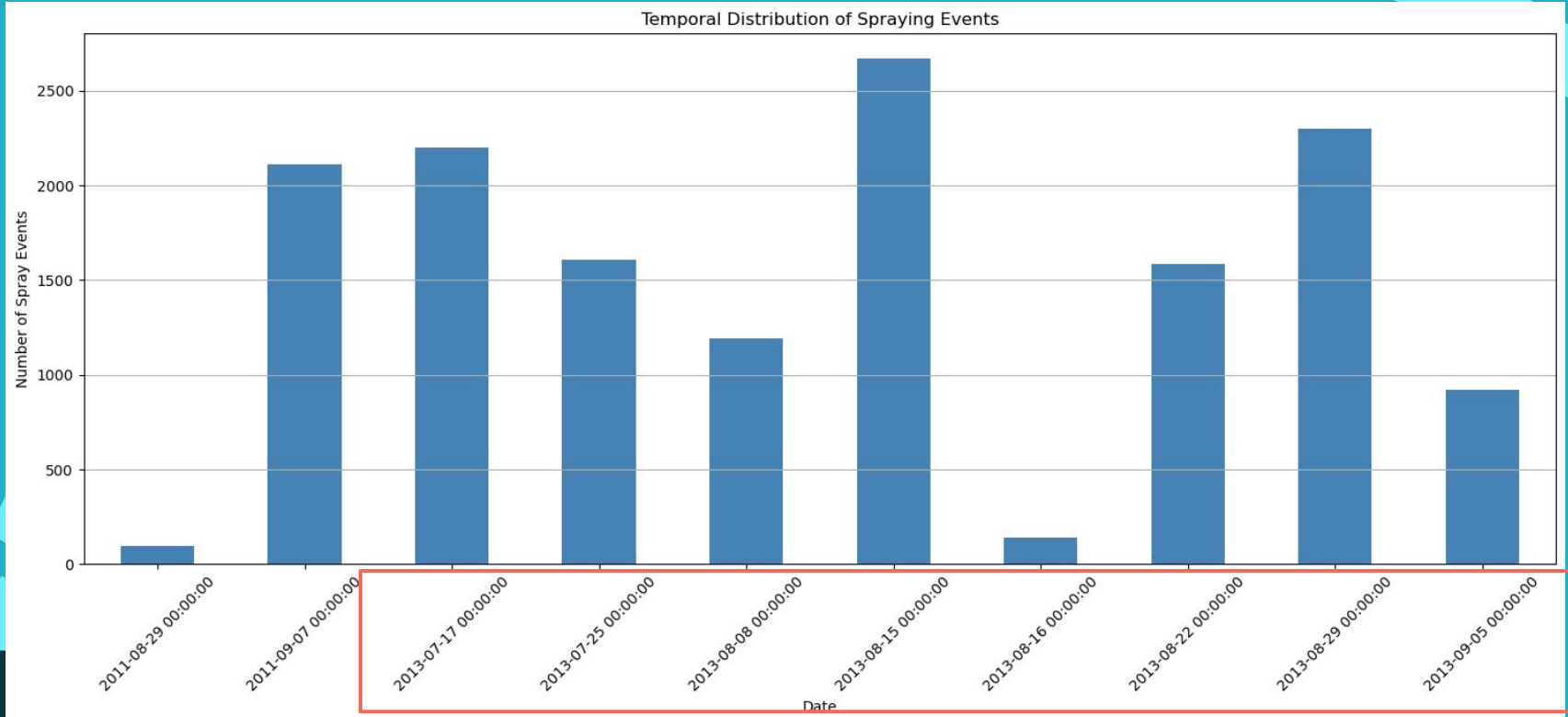
WNV vs Weather Phenomena



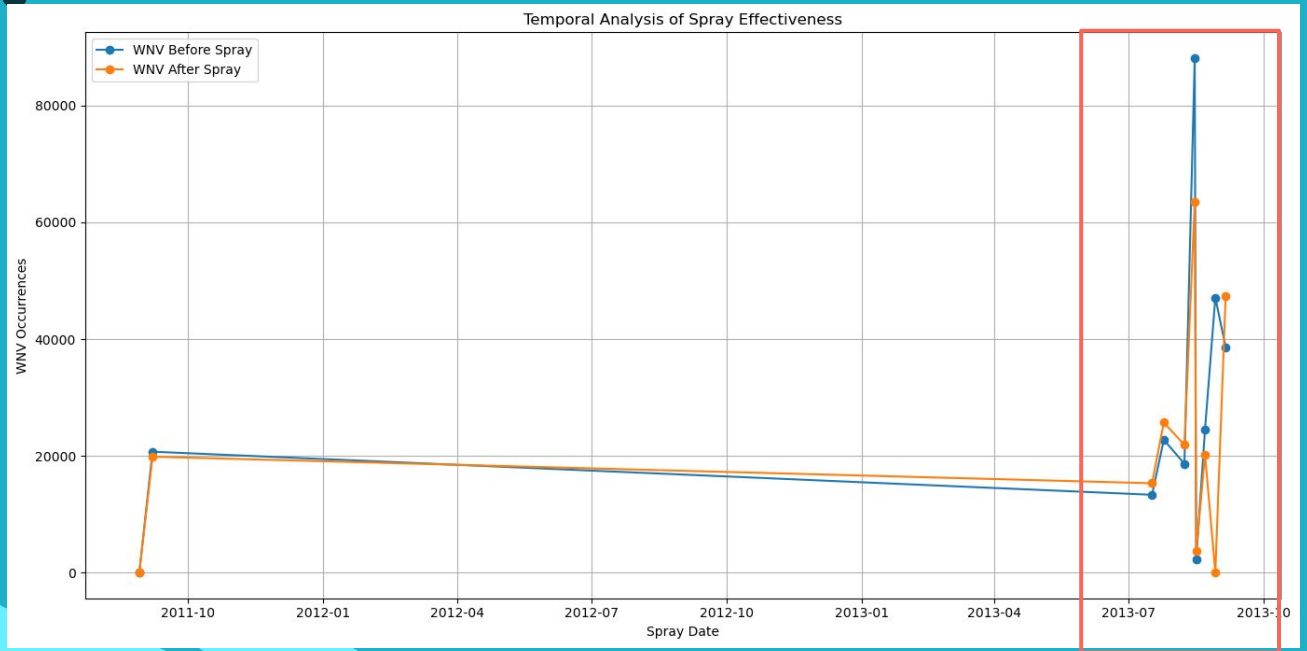
Code	Weather Condition
TSRA	Thunderstorm Rain
FG	Fog
DZ	Drizzle
VCTS	Vicinity Thunderstorms
TS	Thunderstorm
HZ	Haze
BR	Mist
RA	Rain

- **Thunderstorm Rain (TSRA) and Thunderstorm (TS), Rain (RA) and Mist(BR)** show a noticeable increase in frequency on days with WNV present.
- Weather conditions play a role in the presence of WNV, possibly due to creating ideal breeding grounds for mosquitoes.

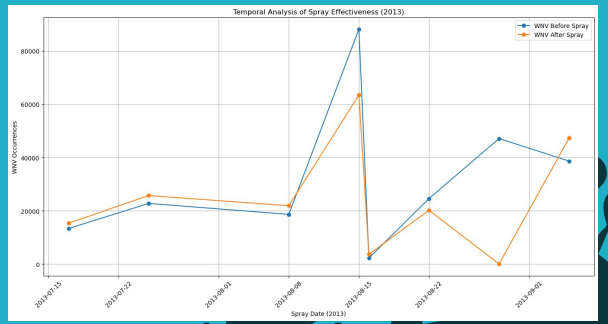
Total Spray Count



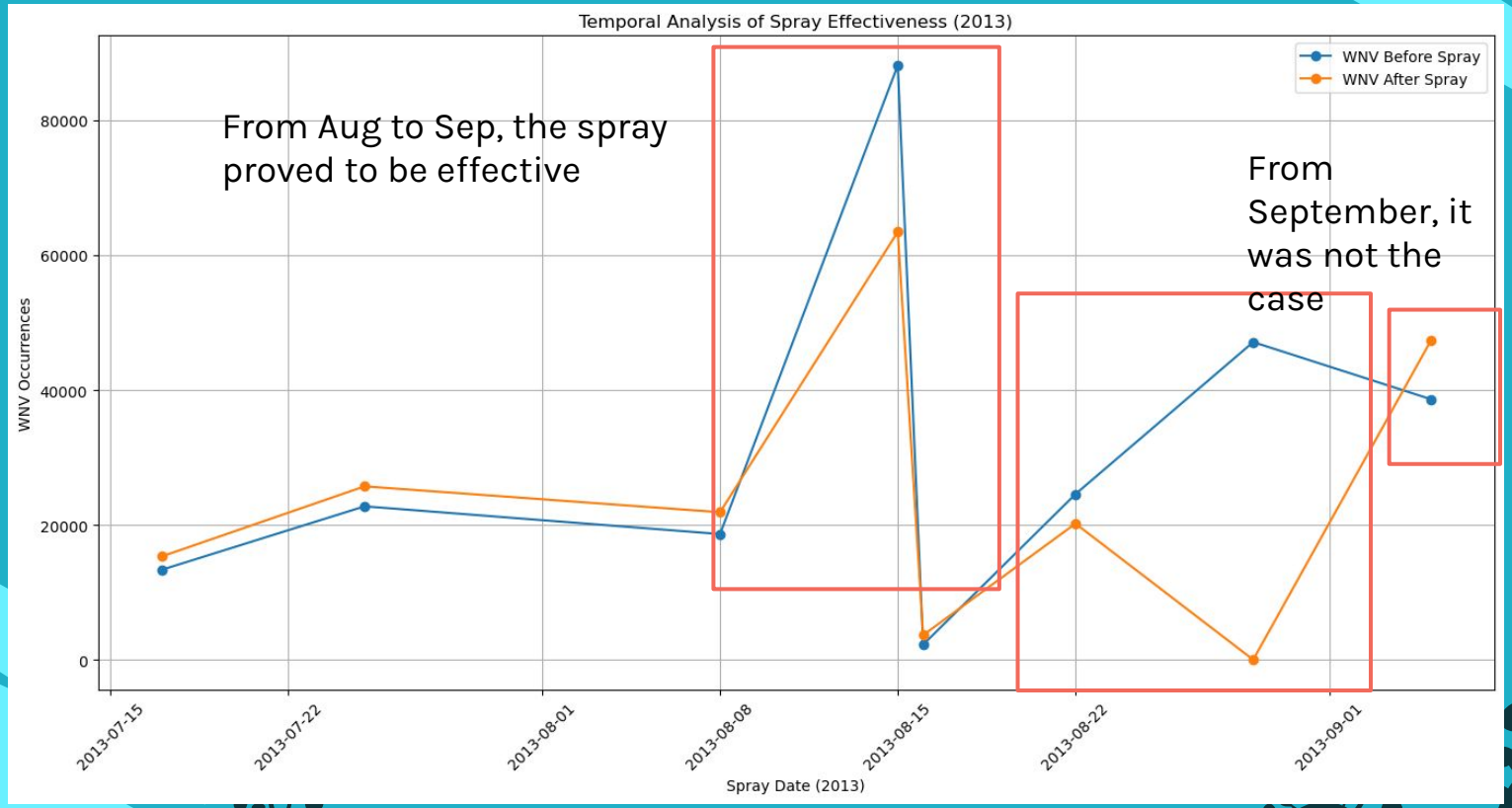
Spraying Insecticide Effective?

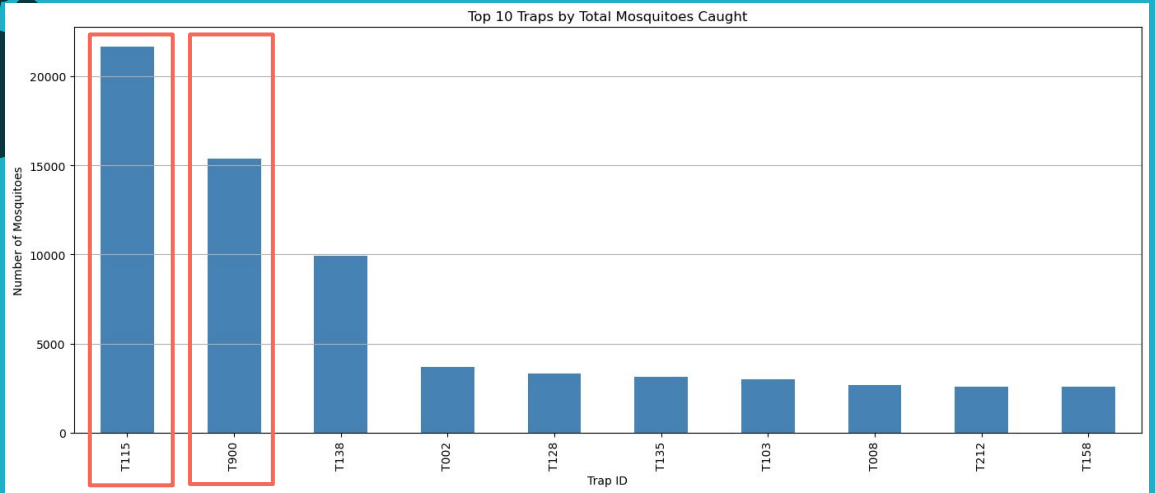


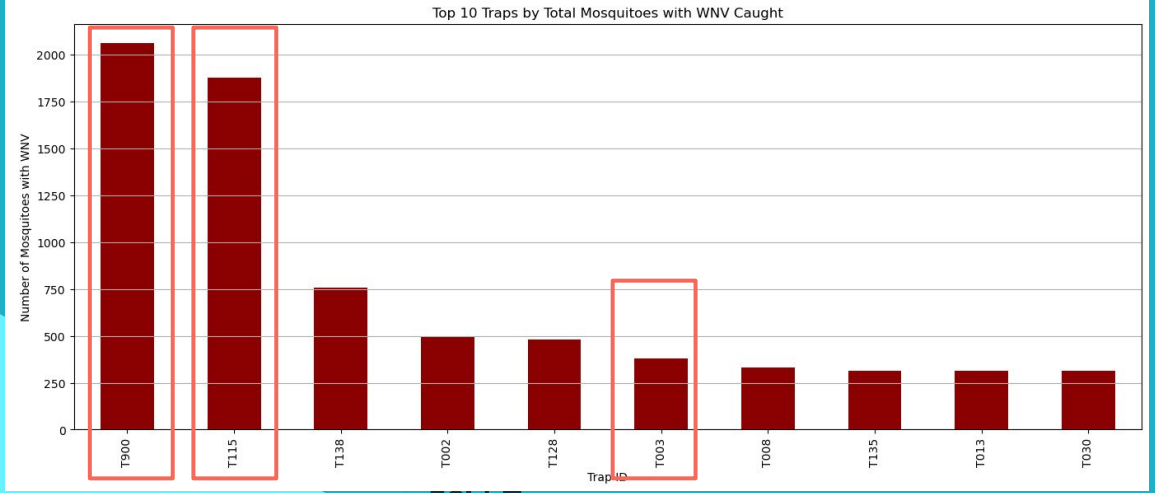
From the graph, most spraying activity was done in 2013



Spraying Insecticide Effective?





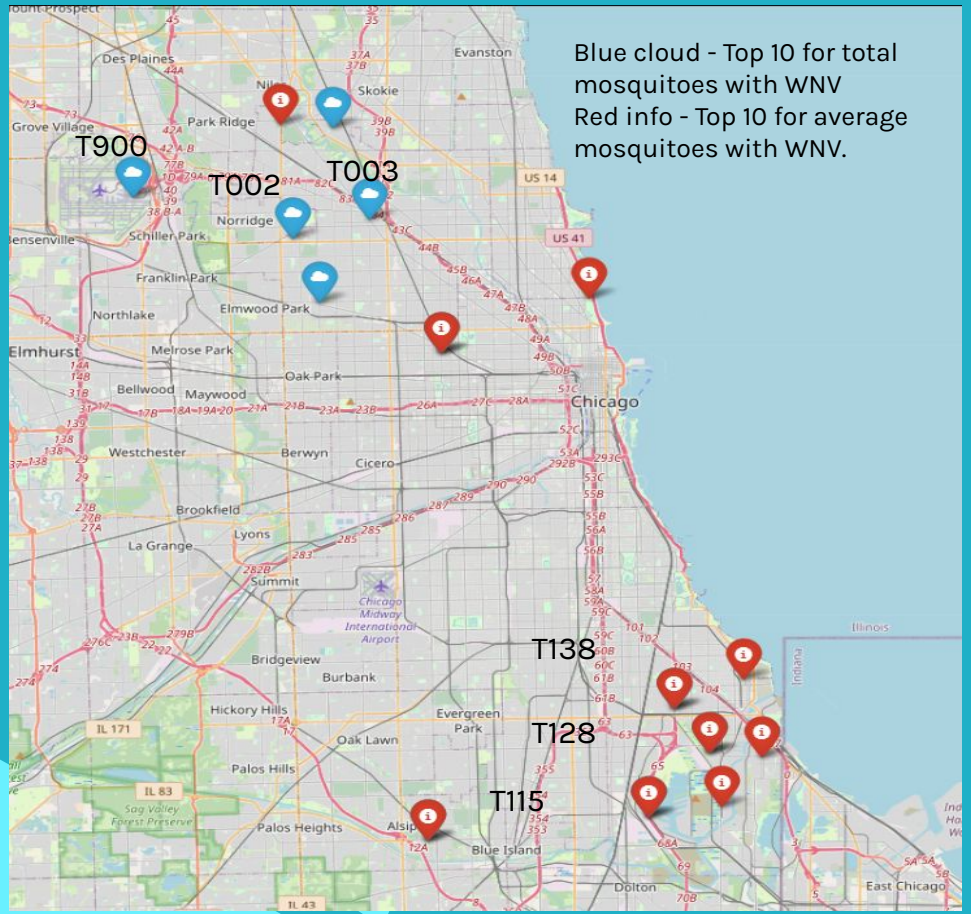


Traps

T115 caught the most mosquitoes but is the second highest WNV

T900 shows the opp

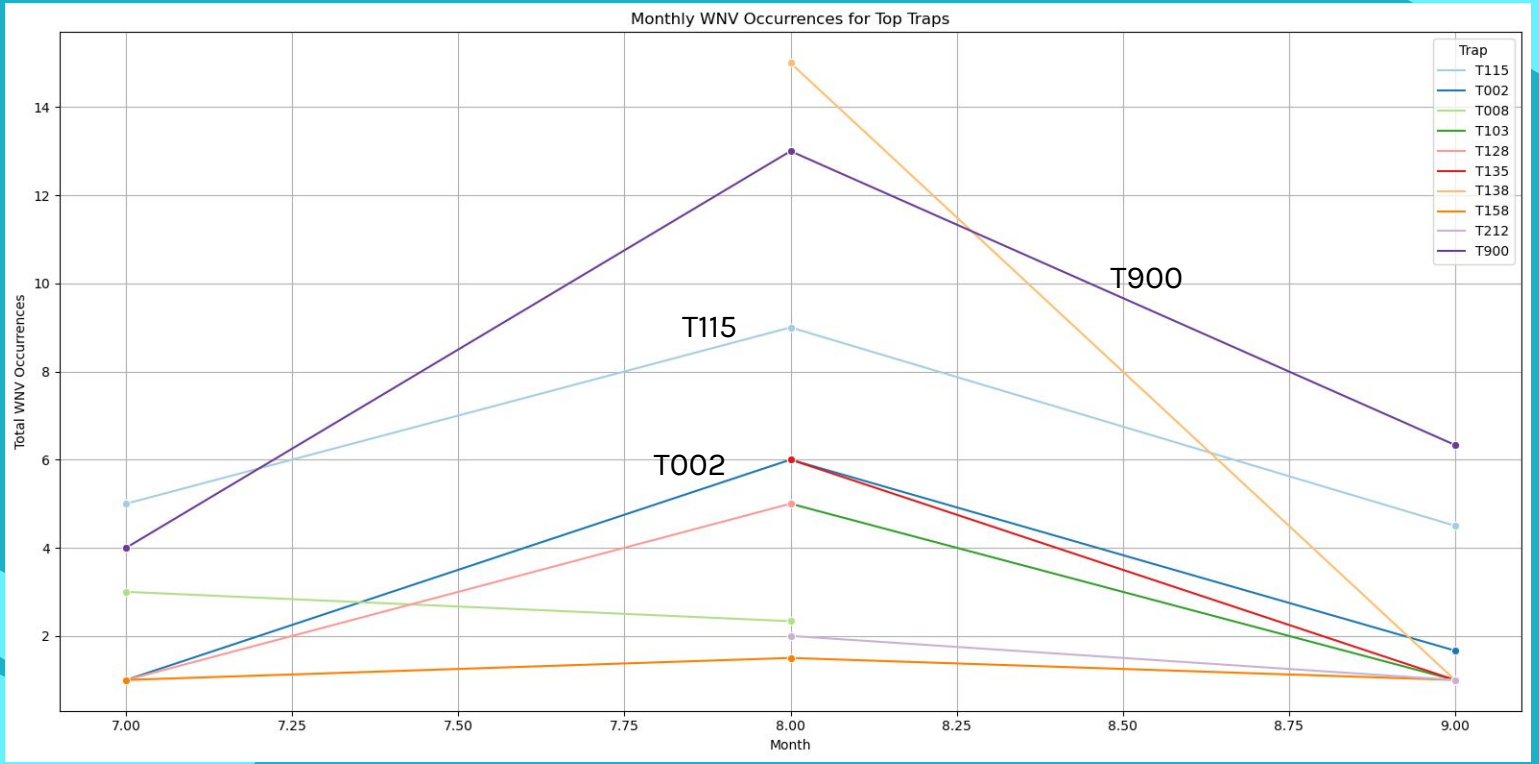
T003 was not in the top 10 traps but is in the top traps for mosquito wnv carrier



Higher Average Count doesnt necessarily indicate a higher wnv count.

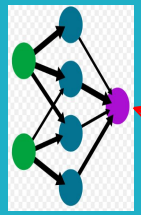
T900 may have the highest no of mosquitoes but on average they are not.

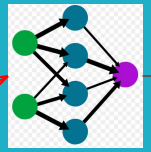
Consistent Peaks Among the Traps



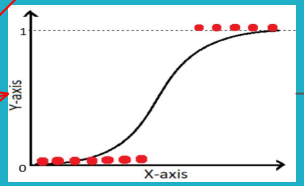
04.
Modelling

Modelling Methodology



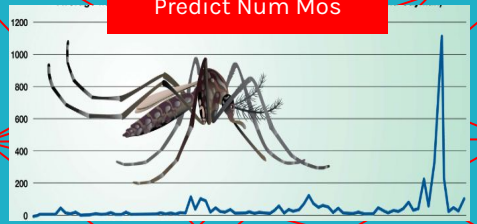


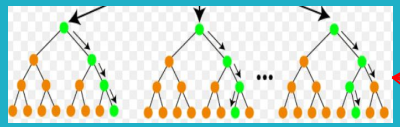
Neural Net

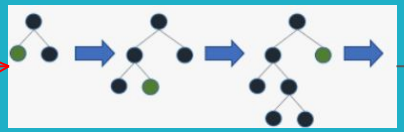


LogReg

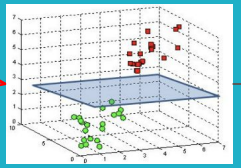
Predict Num Mos



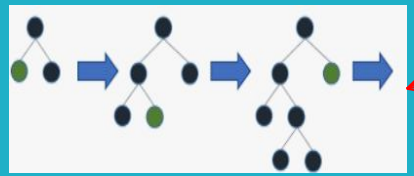


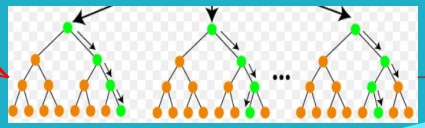


XGBoost



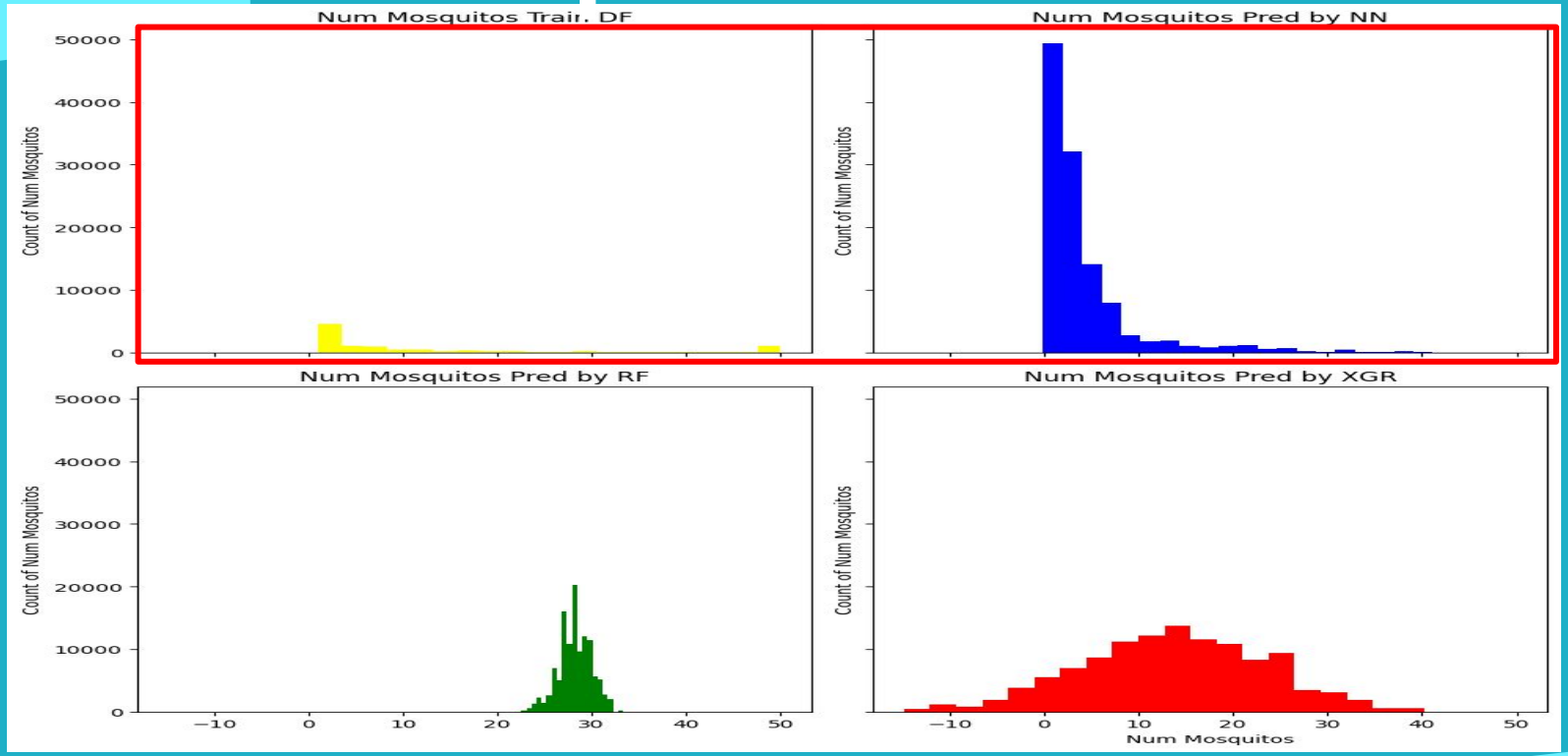
Support Vec





Rand Forest

Num Mosquitos Distribution



- Num Mosquitos Predictions differed noticeably across all models
- Num Mosquitos in train dataframe and predictions from neural network had the closest resemblance

Summary of Results

Model for NumMoSquitos	WnV Model	AUC (Train Data Set)	Kaggle Score	Best Params WNV Model
Neural Net Regression	Neural Net Classifier	0.87	0.54	N.A
Neural Net Regression	RandomForestClassifier	0.86	0.56	rc__max_depth': None, 'rc__max_samples': None, 'rc__min_samples_leaf': 5, 'rc__min_samples_split': 10, 'rc__n_estimators': 100
Neural Net Regression	Support Vector Classifier	0.86	0.51	'svc__C': 100, 'svc__gamma': 'scale', 'svc__kernel': 'rbf'
Neural Net Regression	Logistic Regression	0.87	0.54	'lr__C': 100, 'lr__l1_ratio': 0.75, 'lr__penalty': 'l2'
Neural Net Regression	XGB Classifier	0.87	0.57	'xgc__booster': 'gbtree', 'xgc__gamma': 0, 'xgc__learning_rate': 0.1, 'xgc__n_estimators': 500, 'xgc__reg_alpha': 1, 'xgc__reg_lambda': 1
RandomForestRegression	Neural Net Classifier	0.87	0.53	N.A
RandomForestRegression	RandomForestClassifier	0.87	0.56	'rc__max_depth': None, 'rc__max_samples': None, 'rc__min_samples_leaf': 5, 'rc__min_samples_split': 10, 'rc__n_estimators': 100
RandomForestRegression	Support Vector Classifier	0.86	0.53	'svc__C': 100, 'svc__gamma': 'scale', 'svc__kernel': 'rbf'
RandomForestRegression	Logistic Regression	0.87	0.55	'lr__C': 100, 'lr__l1_ratio': 0.25, 'lr__penalty': 'l2'
RandomForestRegression	**XGB Classifier**	**0.87**	**0.59**	**'xgc__booster': 'gbtree', 'xgc__gamma': 0, 'xgc__learning_rate': 0.1, 'xgc__n_estimators': 500, 'xgc__reg_alpha': 1, 'xgc__reg_lambda': 1**
XGB Regression	Neural Net Classifier	0.87	0.53	N.A
XGB Regression	RandomForestClassifier	0.87	0.56	'rc__max_depth': None, 'rc__max_samples': None, 'rc__min_samples_leaf': 5, 'rc__min_samples_split': 10, 'rc__n_estimators': 300
XGB Regression	Support Vector Classifier	0.86	0.52	'svc__C': 100, 'svc__gamma': 'scale', 'svc__kernel': 'rbf'
XGB Regression	Logistic Regression	0.87	0.54	'lr__C': 100, 'lr__l1_ratio': 0.25, 'lr__penalty': 'elasticnet'
XGB Regression	XGB Classifier	0.87	0.52	'xgc__booster': 'gbtree', 'xgc__gamma': 0, 'xgc__learning_rate': 0.1, 'xgc__n_estimators': 500, 'xgc__reg_alpha': 1, 'xgc__reg_lambda': 1

XGB Clinched top 2 places

Features may not be linearly seperable

Best params were at the higher end of provided options,, which suggests that it could be trying to compensate overfitting

05.
Cost Analysis

Treatment Cost

Although there was not many WNV cases reported in the past 3 years, but if the officials are complacent, the city is likely to face the high number of WNV cases as high as **229** in the year 2012

Asymptomatic

229 x 80% = approx 184 individuals

Serious & Fatal

229 x 1% = approx 2 individuals

Febrile illness (Fever)

229 - 184 - 2 = approx 43 individuals

Treatment Cost

517,502.67 USD

Total Treatment Cost

391,335.69 USD

Total Treatment Cost for Non-Severe Cases
43 x 9,100.83 USD

126,166.98 USD

Total Treatment Cost for Severe Cases
2 x 63,083.49 USD

166,355.09 USD

Loss of Income aside from footing treatment cost.
Each individual made an income loss of 3,696.78 USD
(3,696.78 x (43 + 2))

Total Loss

683,857.76 USD

Total Loss

517,502.67 USD

Total Treatment Cost

166,355.09 USD

Total Income Cost

Cost of Zenivex

Size of Chicago (land)

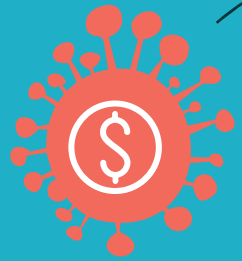
589.82 square km =145,747.67 acres

Spray Amount

1.5 ounces per acre
145,747.67 / 1.5 = 97,165.11 ounces

Convert to gallon

97,165.11 /128= 759.1 gallon



Cost of Zenivex

759.10 x 386.53 USD = **293,415.87 USD**

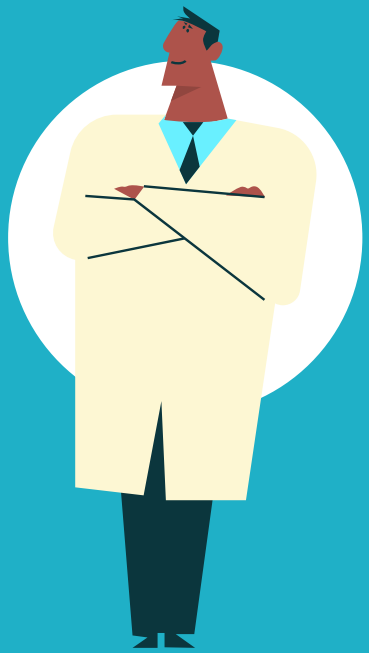
06.
Conclusion

Conclusion

01

Zenivex

293,415.87 USD



02

Total Loss

683,857.76 USD

Spraying efforts should not be stopped!

Conclusion

The presence of WNV virus is based on these factors:

- Numbers of mosquitoes.
- Temperature and humidity is higher than normal days.
- Lower average wind speed.
- During the months from May to August.
- Negligence of spraying at some of the areas with WNV presence.
- Mosquitoes species belonging to Culex Pipiens and Culex Restuans.

Conclusion

Additional Information

- The top traps located that has high mosquitoes count and WNV presence are T900 and T115.
- Thunderstorms and mist might contribute to no. of WNV cases.
- Spraying efforts should not be stopped especially when the cost of medical treatment will keep rising.

Recommendations

Do

To use data driven approach to reduce WNV by implementing preventive measures:

- Spray is during the month of July to August as these two months are the peak of the mosquitoes numbers.
- Ensure thorough spraying.
- Residents to wear loose-fitting clothes that can cover arms and legs during the month of July and August
- Educate the residents to prevent mosquitoes breeding, such as draining water out from anything that can collect rainwater.
- Enforce the laws and regulations for mosquitoes breeding.

Recommendations

Don’t

- Use the wrong concentration amount of insecticides.

Further Improvements

- More data on number of mosquitoes for more accurate modelling results.
- Improve model's performance by factoring in the time lag that was observed for WNV to be present when an increase in mosquitoes numbers are detected

Thank You!

Do you have any questions?

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik.