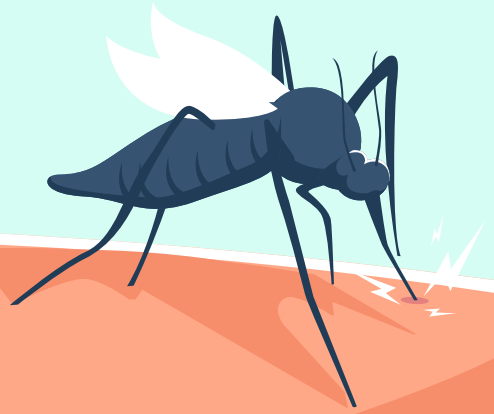
DATA-BACKED SOLUTIONS FOR COMBATING WNV IN CHICAGO

Machine-Learning Approach
to minimise cost and
maximise benefit



Presented by Data Nine-Nine

BREAKING DOWN THE NUMBERS

25,769

HOSPITALIZED

Number of Hospitalization cases from 1999 to 2022

2,773

DEATH

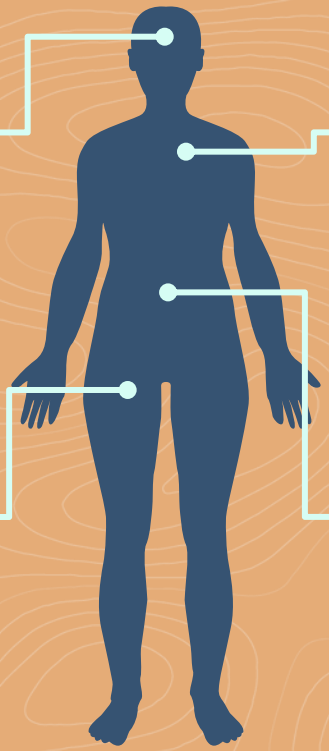
Number of Death cases from 1999 to 2022

Source: https://www.cdc.gov/westnile/statsmaps/historic-data.html

SYMPTOMS OF WEST NILE VIRUS

HEADACHE, DISORIENTATION

SWOLLEN LYMPH GLAND, STIFF NECK

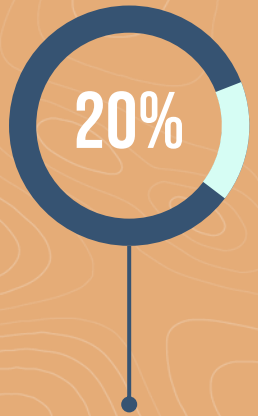


MUSCLE PAIN, RASH

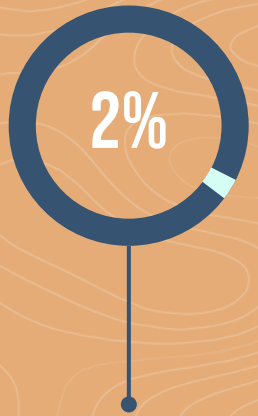
NAUSEA

Source: https://www.vdci.net/

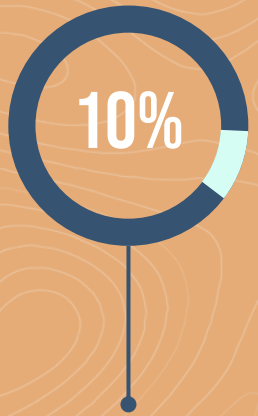
WEST NILE VIRUS



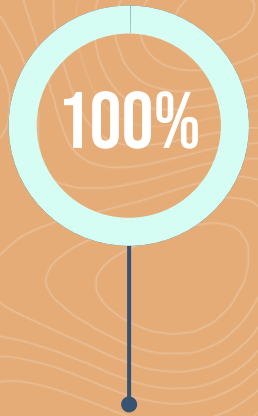
1 in 5 develop symptoms



1 in 50 Age > 60 developed severe illness



1 in 10 with severe illness die.

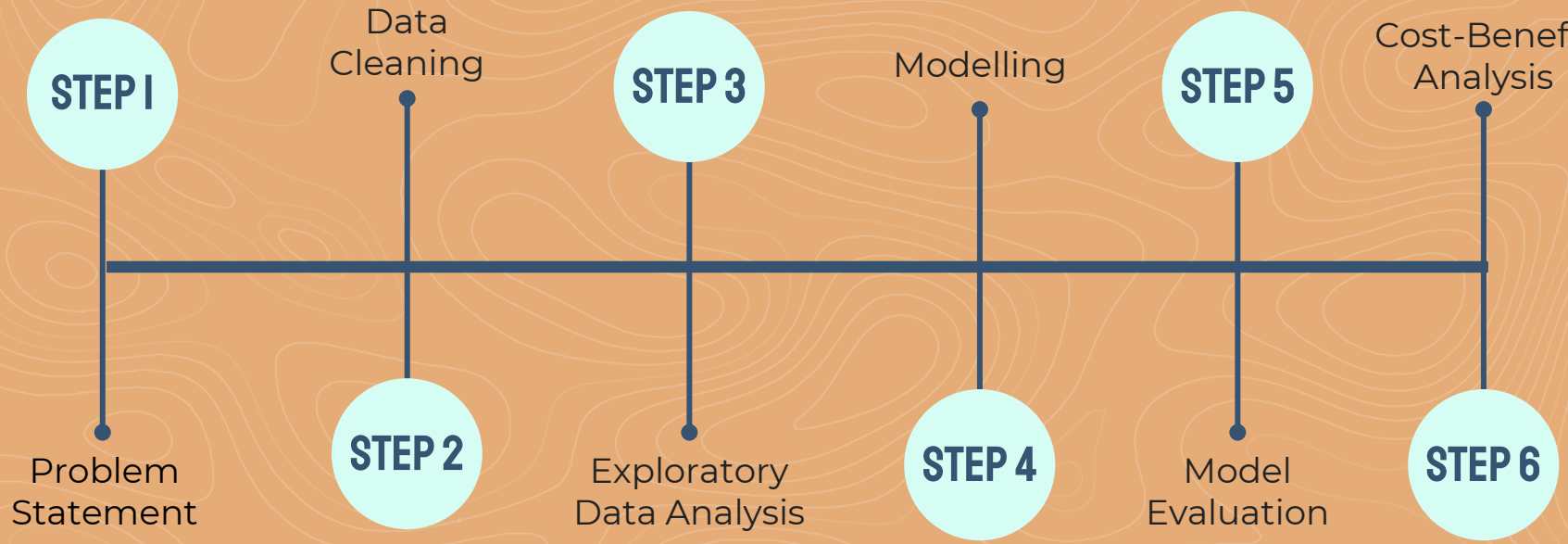


100% possible to reduce the risk with preventive actions!

Source: https://www.cdc.gov/westnile/index.html

PREVENTION IS KEY!

METHODOLOGIES



01 PROBLEM STATEMENT

You could enter a subtitle here if you need it

PROBLEM STATEMENT

We, **Data Nine-Nine**, have been engaged as a third-party consulting firm by the **Centre for Disease Control and Prevention (CDC)** to collaborate on a comprehensive review of their West Nile Virus (WNV) control efforts.
Our objective is to:

(1) Build Machine Learning model to predict the presence of WNV

(1) Provide valuable insights and recommendations to further enhance their strategies.

02 DATA CLEANING

Data Preparation prior to analysis and modeling

DATASETS

	2007	2008	2009	2010	2011	2012	2013	2014
train_df	May - Oct		May - Oct		Jun - Sep		Jun - Sep	
test_df		Jun - Sep		Jun - Oct		Jun - Sep		Jun - Oct
weather_df	May - Oct	May - Oct	May - Oct	May - Oct	May - Oct	May - Oct	May - Oct	May - Oct
spray_df					Aug - Sep		Jul - Sep	

DATA CLEANING

10,506

ROWS BEFORE

number of mosquitos
capped at 50 per row

8,610

ROWS AFTER

groupby to sum total
number of mosquitoes

AGGREGATE MOSQUITOES DATA FROM TRAIN_DF

135,039 Total number of mosquitoes caught over 4 years.

2,206 Highest number of mosquitoes caught in a day.

7 Species of mosquitoes caught

FEATURE ENGINEERING

MIN/MAX/AVG TEMP

Split by station & average between the 2 station

DAY LENGTH

The number of daylight in hours & minutes derived from sunrise & sunset

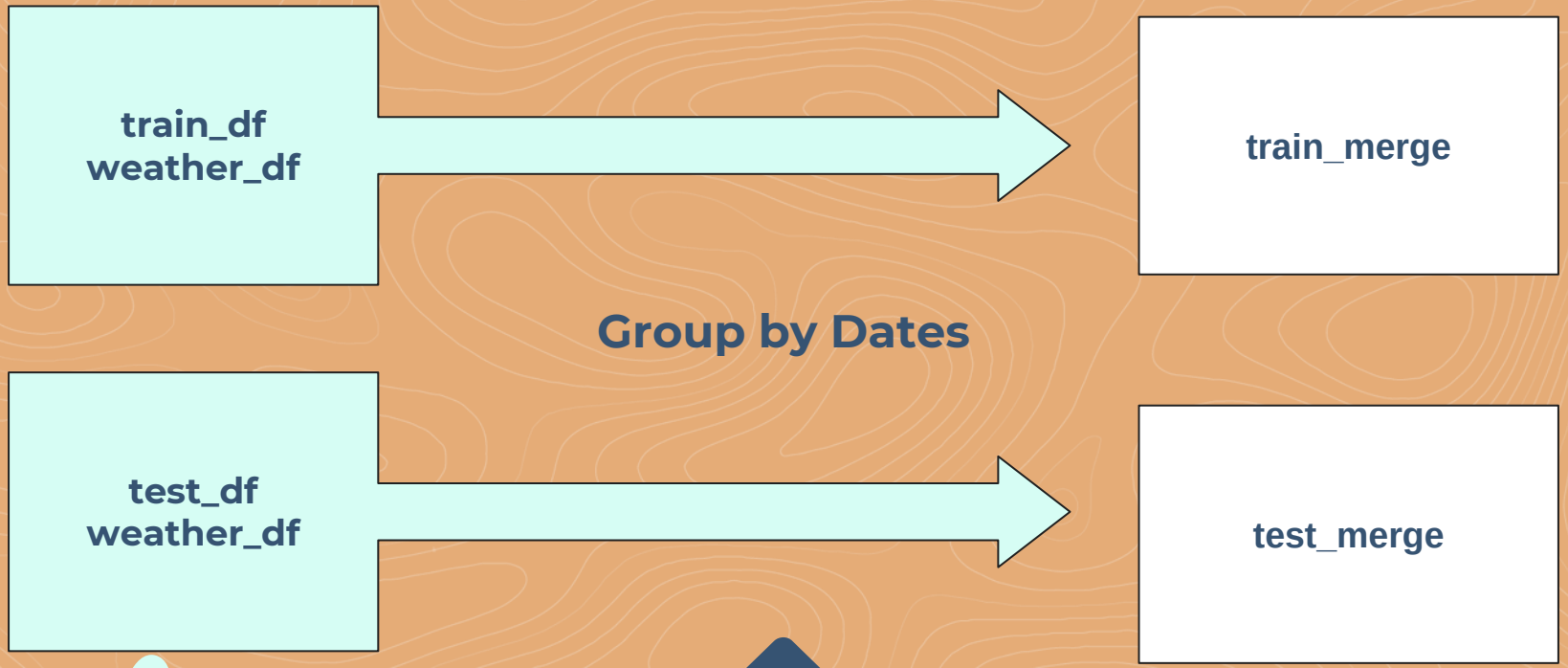
LAGGED TEMP/HUMIDITY/RAINFALL

Weather data lagged by up to 4 weeks

PRECIPITATION

Split by station & average between the 2 station

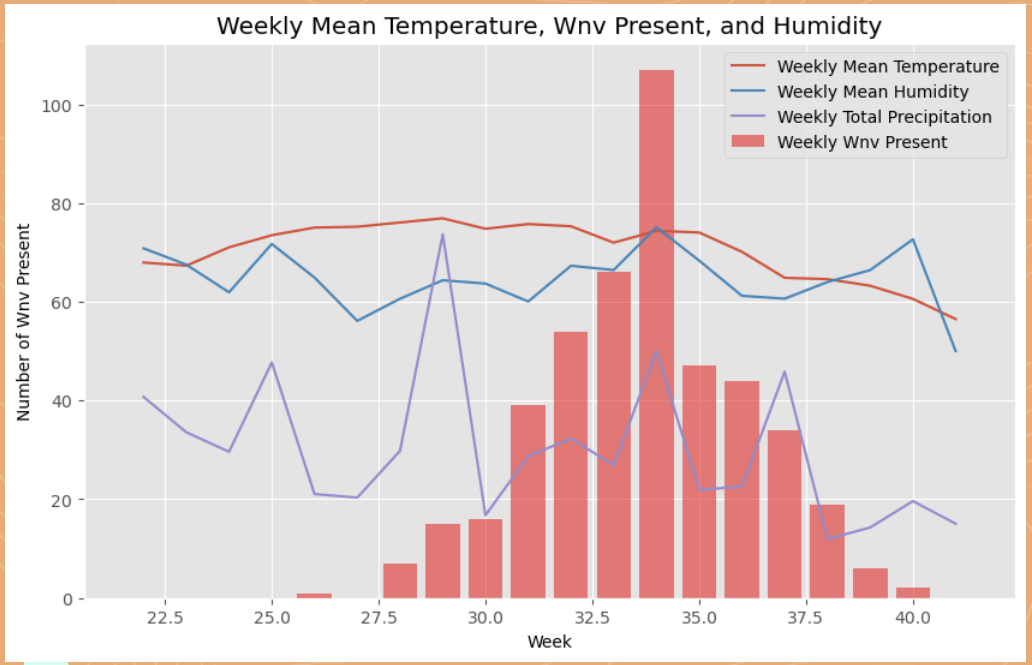
DATA MERGING

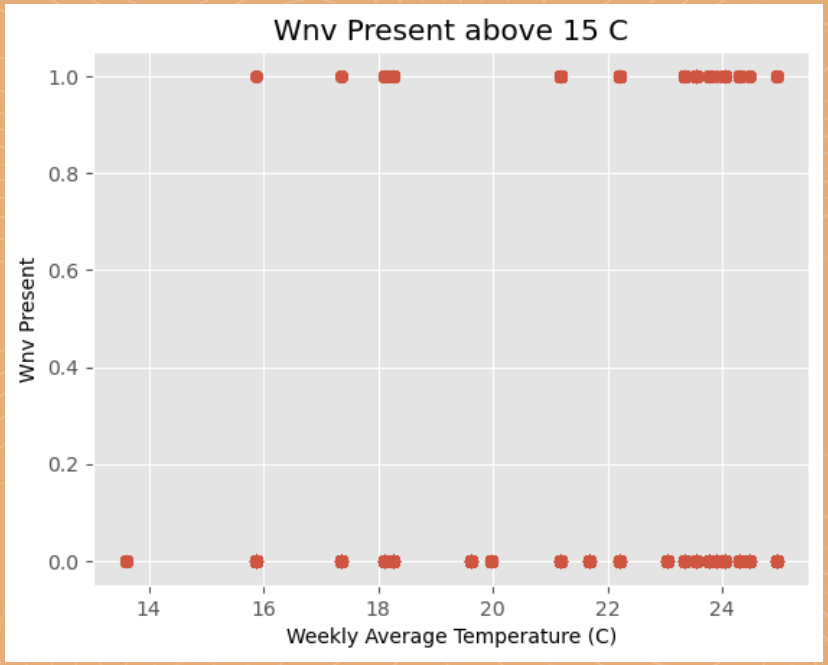


03 EXPLORATORY DATA ANALYSIS

Data Visualisation & Analysis

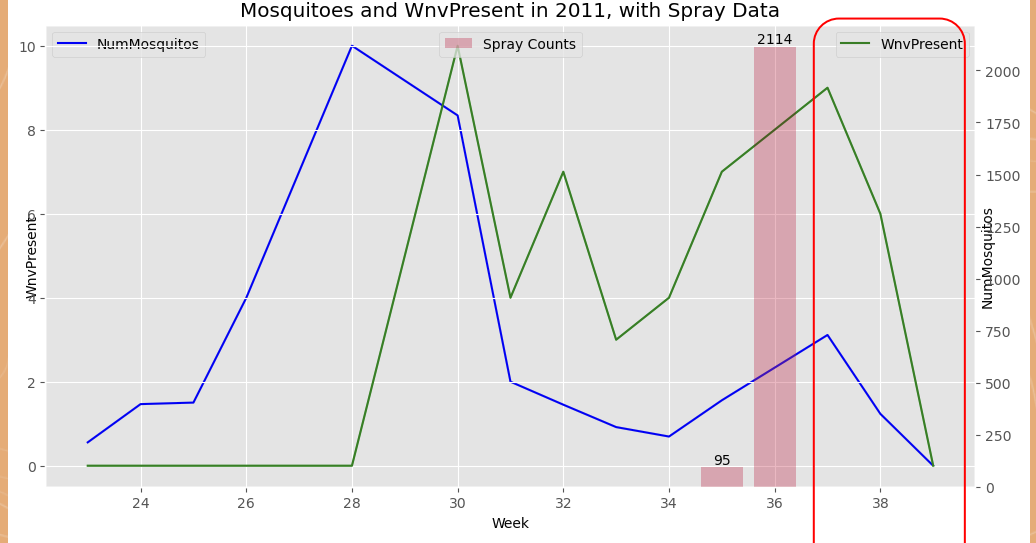
WEATHER CONDITION VS WNV CASES

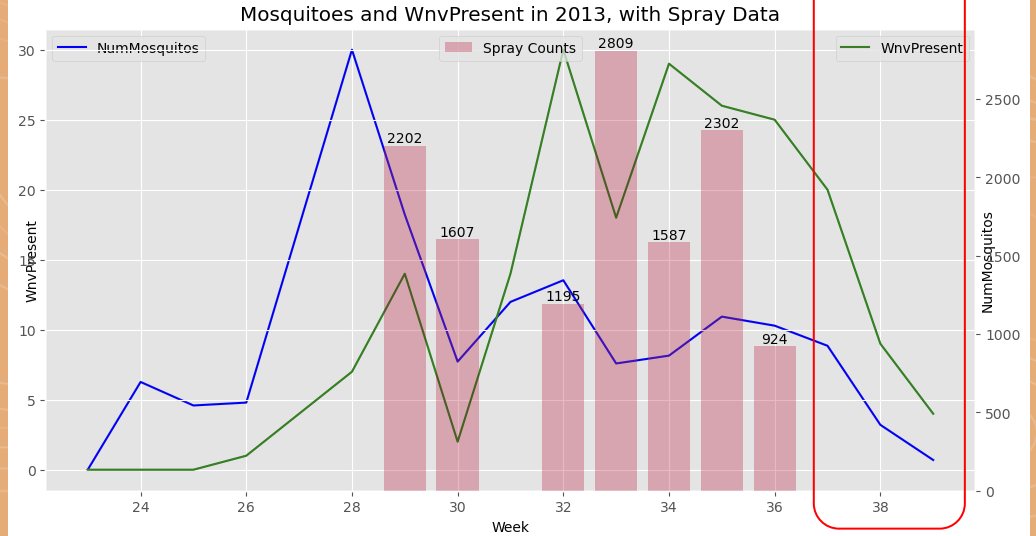




Note: Graph only shows correlation, not causation

ARE MOSQUITO SPRAYS EFFECTIVE?



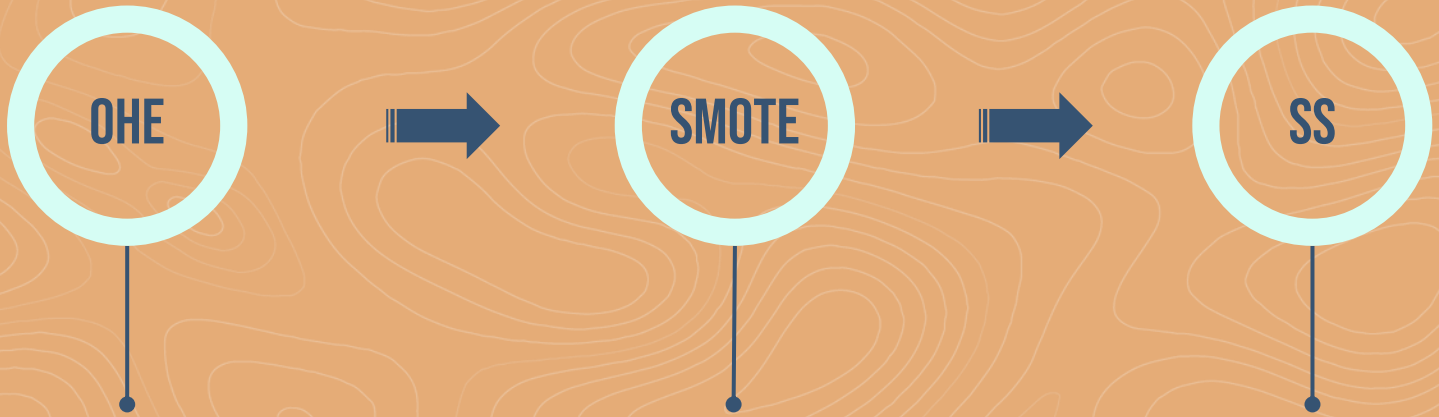


RED BAR IS SPRAY

04 MODELLING

Preprocessing & Model Deployment

MODELLING PREPROCESSING



One Hot Encoder

Transforms **Categorical** Data Into **Numerical**

Synthetic Minority Oversampling Technique

Balance the Target Class to a **50:50** ratio

Standard Scaler

Transform all features into similar **Scale** & **Distribution**.

ONE HOT ENCODER

3

CATEGORICAL DATA

[Species, Street, Trap]

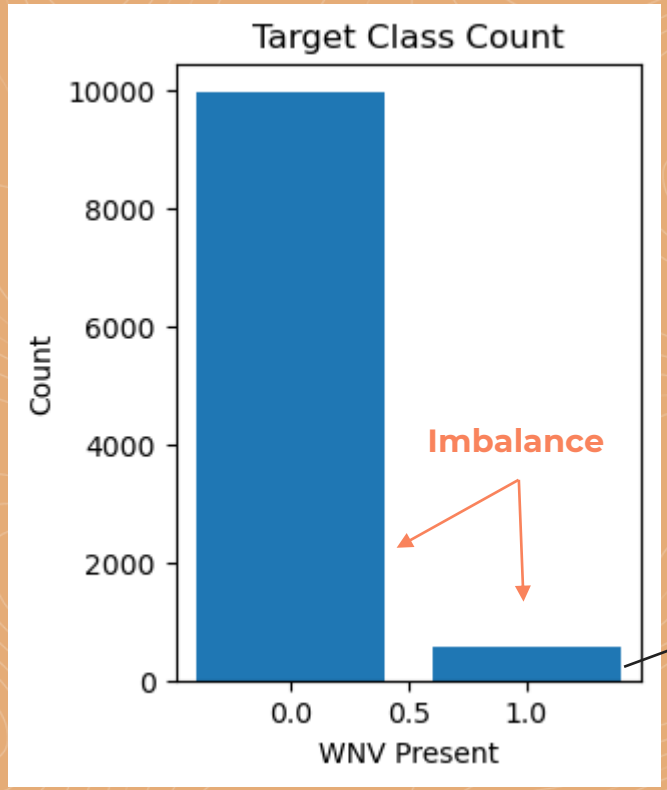
DUMMIFICATION

260+

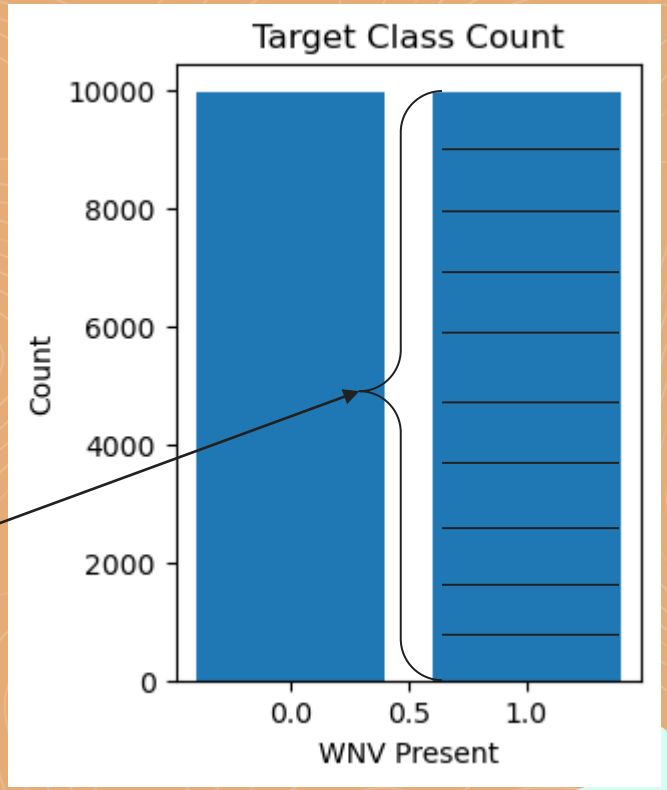
NUMERICAL DATA

7 species
128 streets
136 traps

SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE)



SMOTE



MODELLING WORKFLOW

MODEL USED

- Logistic Regression
- Random Forest Classifier
- XGBoost Classifier
- AdaBoost Classifier

HYPERPARAMETERS TUNING

- 400 Fits of LR
- 200 Fits of RFC
- 1200 Fits of XGB
- 300 Fits of ABC

ENSEMBLE

- Voting Classifier

05 MODEL EVALUATION

Model Performance Review

METRICS

01
SENSITIVITY

Measures the proportion of actual positive cases correctly identified by the model

02
SPECIFICITY

Measures the proportion of actual negative cases correctly identified by the model

03
AUC

Area Under the ROC Curve (AUC) measure of how well the classifier can distinguish between the positive and negative classes.

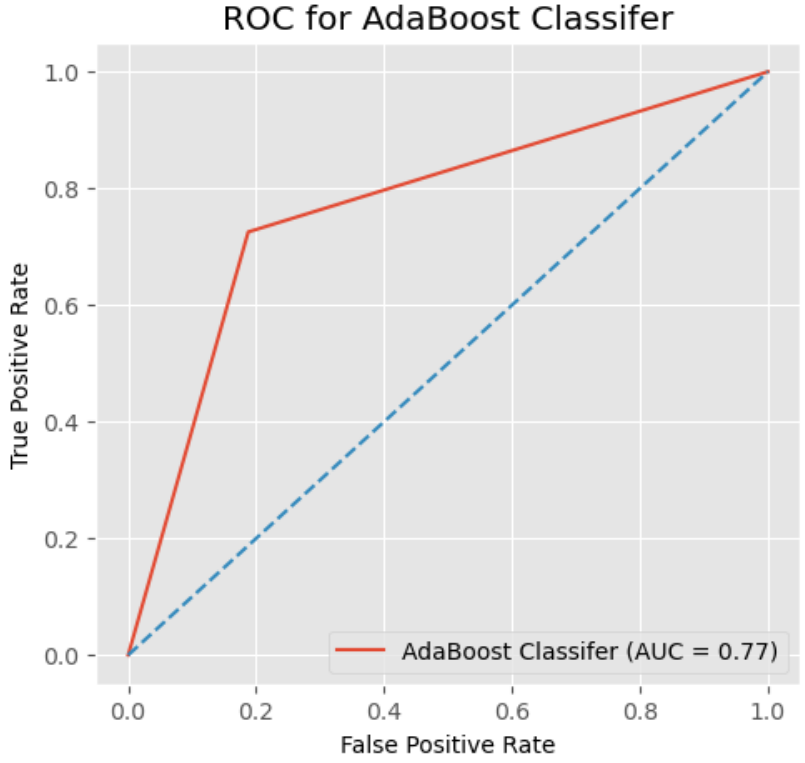
MODEL SCORE

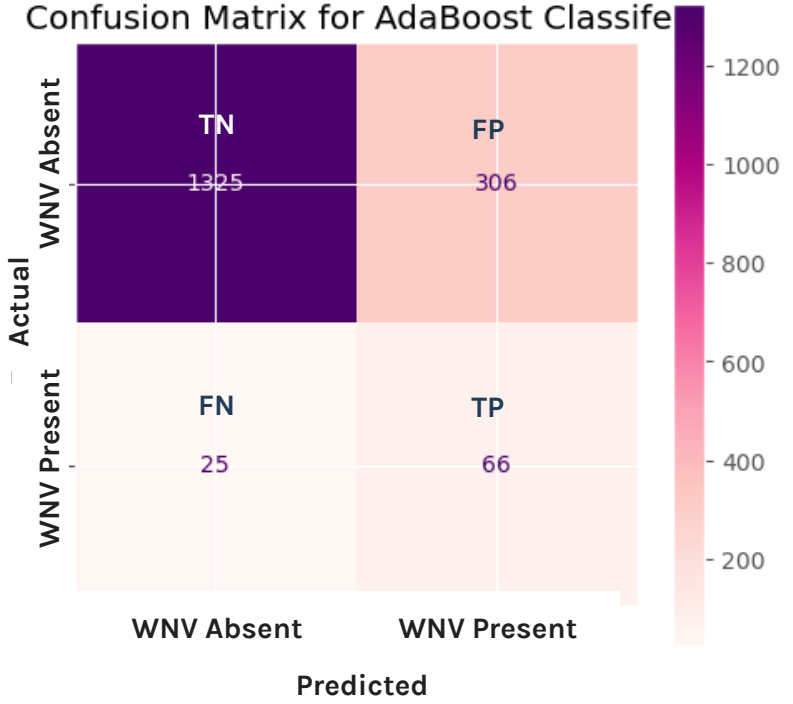
	TPR[1]	TNR[2]	Train ROC	Test ROC
Logistic Regression (Baseline)	0.7802	0.7419	0.8670	0.8269
Random Forest Classifier	0.8352	0.7069	0.8597	0.8552
XGBoost Classifier	0.1319	0.9853	0.9238	0.8733
AdaBoost Classifier	**0.7253**	**0.8124**	**0.8259**	**0.8564**
Voting Classifier	0.6703	0.8443	0.8987	0.8645

1 - True Positive Rate or Sensitivity. 2 - True Negative Rate or Specificity. 3 - Public Score (AUC)

AdaBoost Hyperparameters	**N Estimator - 300**	**Learning Rate - 0.05**	**K-neighbour in SMOTE - 3**

AUC PLOT & CONFUSION MATRIX





06 COST-BENEFIT ANALYSIS

Treatment & Prevention Proposal

FACTUAL TREATMENT NUMBERS FROM 1999 TO 2022

56,569 — Reported cases of WNV Disease

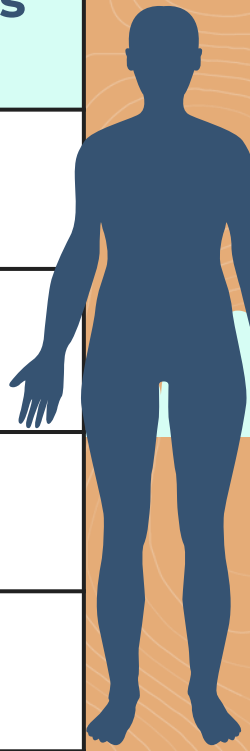
~45.5% — Requires hospitalisation

~50.6% — With Neurologic Invasive Disease

Source from CDC : https://www.cdc.gov/westnile/statsmaps/historic-data.html

MEDIAN HOSPITALISED FINANCIAL COSTS

	Initial Costs /Person	Long Term Costs /Person	Total Costs / Person
Fever	$4,617	$2,271	$6,888
Meningitis	$7,942	$10,556	$18,498
Encephalitis	$20,105	$8,055	$28,160
Acute Flaccid Paralysis	$25,117	$22,628	$47,745



Sources from National Library of Medicine : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3945683/

COST OF CONTROLLING MEASURES

AERIAL SPRAY

$1471 / KM^2

Aerial Spray

SENTINEL CHICKENS

$203 / FLOCK

Chickens placed at WNV area to detect WNV

MOSQUITOES TRAP

$ 120 - $180 / TRAP

Price data from Amazon

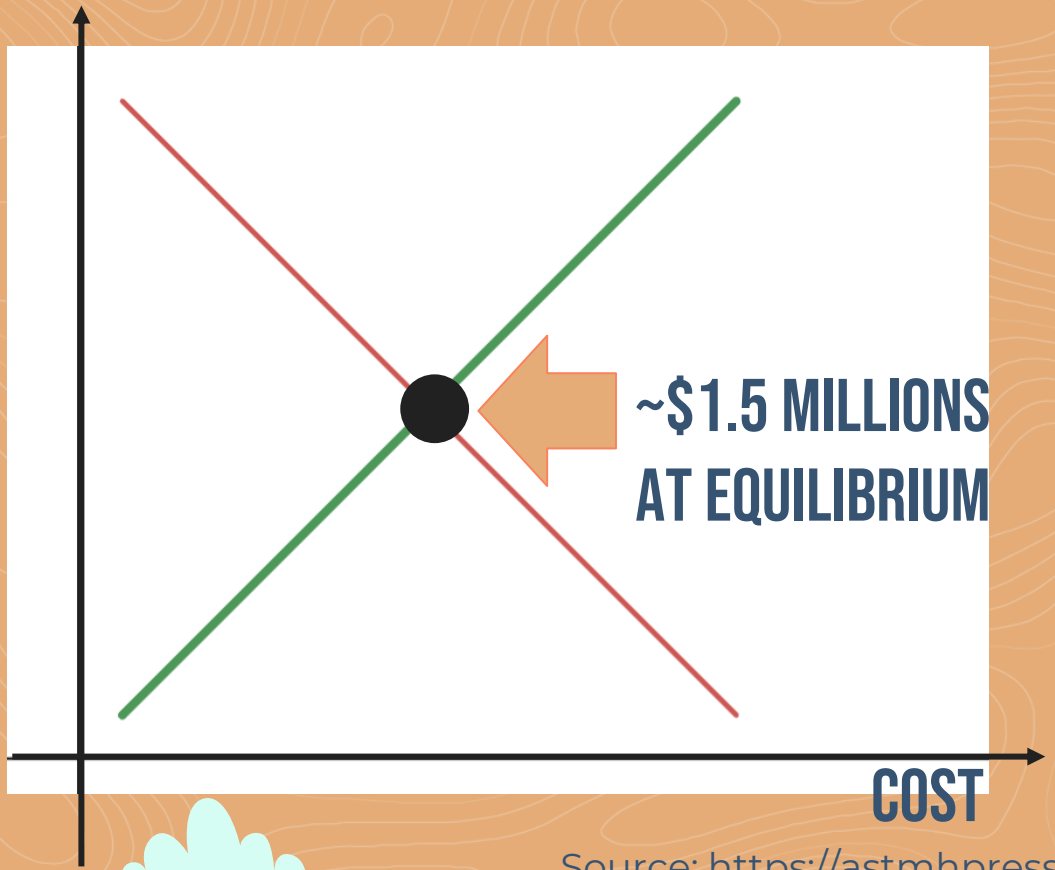
WEBSITE MAINTENANCE

$200 - $4500 / MONTH

Price data from Hostinger.com

Source from National Library of Medicine : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322011/

ECONOMIST METRICS



~$56 MILLIONS / YEAR
FOR UNITED STATES

2.67% POPULATION
CHICAGO / US

~$1.5 MILLIONS / YEAR
FOR CHICAGO

Source: https://astmhpressroom.wordpress.com/journal/february-2014/

CONTROLLING MEASURES AT A GLANCE

	Minimalist	Economist	Life-Saving
Protective Measures	Educational Campaign on Website	Machine Learning Social Media Campaign + Prediction for Detection + Spray Decision	Mass Educational Campaign on Social Media
Early Detection			Sentinel Chicken + Mosquitoes Surveillance
Spread Control	Spray after first Human Case		Spray after Early Detection Method

Sources from National Library of Medicine : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3367363/

SUMMARY OF APPROACHES

	MINIMALIST APPROACH	ECONOMIST APPROACH	LIFE-SAVING APPROACH
PROS	Easy to implement Cost-saving	Cost-effective approach	Proactive and Diligent approach
CONS	Reactive Measures	Does not cover all grounds	Difficult to implement Expensive
INTENT	Minimalising complexity in implementing the measures	Maximising benefit per dollar spent over preventive measures	Maximising potential life saves Minimalising number of mosquitos

POTENTIAL COST REDUCTION BY MACHINE LEARNING

~3.53 MILLIONS

MACHINE LEARNING

~78.5% REDUCTION

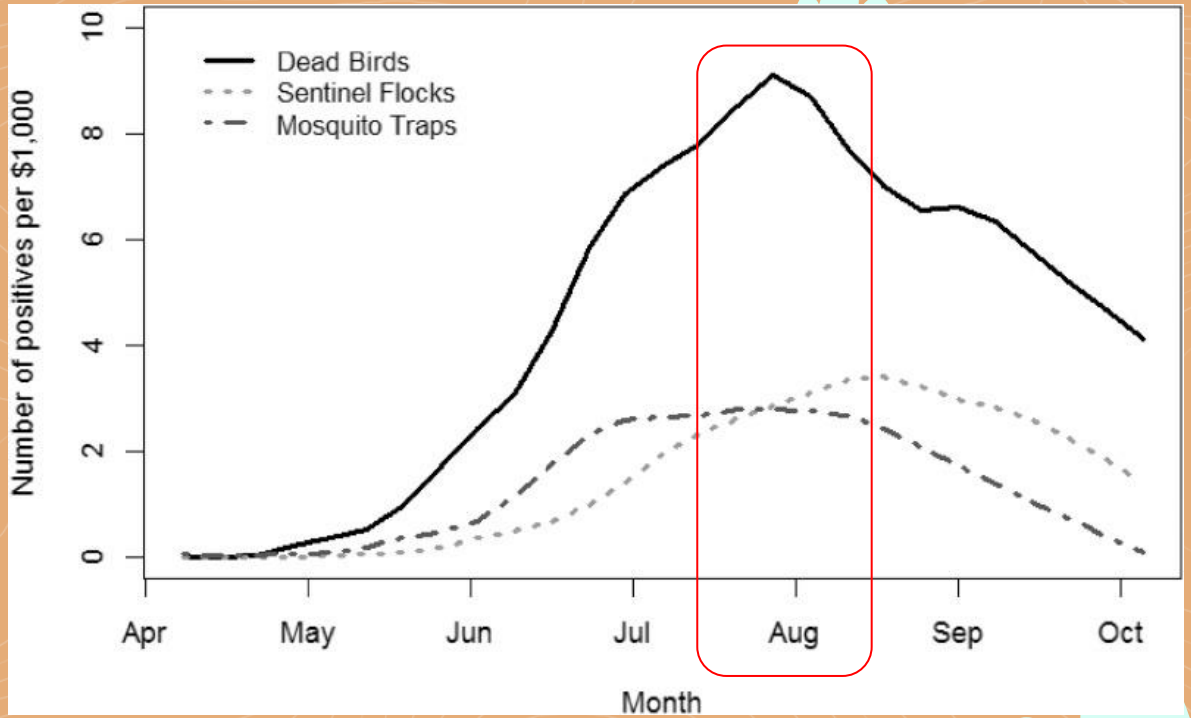
~814 THOUSANDS

Source from National Library of Medicine : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322011/

TIMELINESS OF INTERVENTION

AUGUST

BEST RETURN PER DOLLAR SPENT



Source from National Library of Medicine : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322011/

RECOMMENDATIONS

(1) TIME YOUR EFFORT

- May to October: Economist Approach
- November to April: Minimalist Approach

(2) FOCUS ON CULEX PIPIENS POPULATED AREAS

(3) INITIATE LOCALISED SOCIAL MEDIA CAMPAIGN

RECOMMENDATIONS

EFFECTIVENESS

PROTECTIVE EDUCATIONS > EARLY DETECTION > INTERVENTION > MEDICAL

PRICE TO PAY

PROTECTIVE EDUCATIONS < EARLY DETECTION < INTERVENTION < MEDICAL

THANK YOU

AND SAY NO TO LIFE LOSS

Data Nine-Nine

1. Ming Fatt
2. Jasmine
3. JJ
4. Willson
5. Wenxi

APPENDIX

ALL DISEASE NUMBERS

Year: 1999-2022

Type of case: All disease cases

Human Disease Cases	Hospitalizations	Deaths
56,569	**25,769**	**2,773**
Cases from year(s) and type of case selected above	*Hospitalizations from year(s) and type of case selected above*	*Deaths from year(s) and type of case selected above*

Source from CDC : https://www.cdc.gov/westnile/statsmaps/historic-data.html

NEUROINVASIVE DISEASE NUMBERS

Year: 1999-2022

Type of case: Neuroinvasive disease cases

Human Disease Cases	Hospitalizations	Deaths
28,614	**20,937**	**2,632**
Cases from year(s) and type of case selected above	*Hospitalizations from year(s) and type of case selected above*	*Deaths from year(s) and type of case selected above*

Source from CDC : https://www.cdc.gov/westnile/statsmaps/historic-data.html

HYPERPARAMETERS TUNING

	Parameter 1	Parameter 2	Parameter 3	Parameter 4
Logistic Regression	Penalty	Inverse of Regularization Strength	K-neighbour in SMOTE	
Random Forest Classifier	Max Depth	N Estimator	K-neighbour in SMOTE	
XGBoost Classifier	Max Depth	Learning Rate	N Estimator	K-neighbour in SMOTE
AdaBoost Classifier	N Estimator	Learning Rate	K-neighbour in SMOTE	

CONTROLLING MEASURES AT A GLANCE

	Minimalist	Economist	Life-Saving
Protective Measures	Educational Campaign on Website	Machine Learning Campaign on Social Media	Mass Educational Campaign on Social Media
Early Detection	Sentinel Chicken	Dead Bird Surveillance	Mosquitoes Surveillance
Spread Control	Spray after first Human Case	Spray according to Machine Learning	Spray after Early Detection Method

Sources from National Library of Medicine : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3367363/

CONTROLLING MEASURES AT A GLANCE

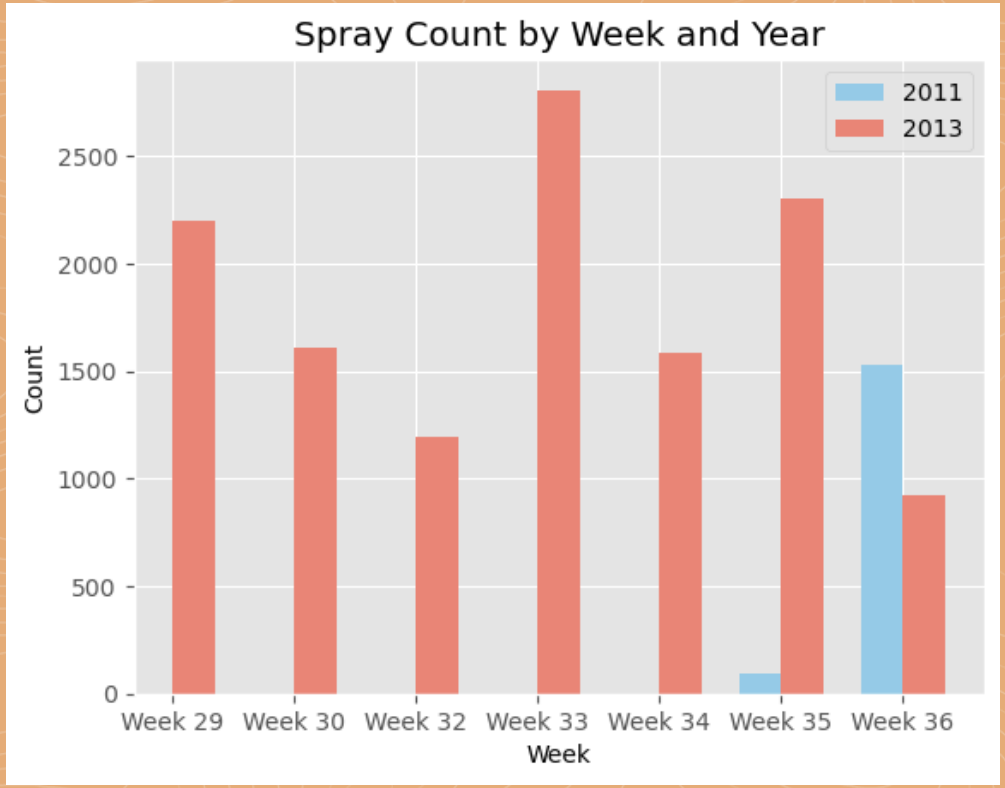
More $$

	Minimalist	Economist	Life-Saving
Protective Measures	Educational Campaign on Website	Machine Learning Campaign on Social Media	Mass Educational Campaign on Social Media
Early Detection	-	Perform Model Prediction using Machine Learning	Sentinel Chicken + Mosquitoes Surveillance
Spread Control	Spray after first Human Case	Spray according to Machine Learning	Spray after Early Detection Method

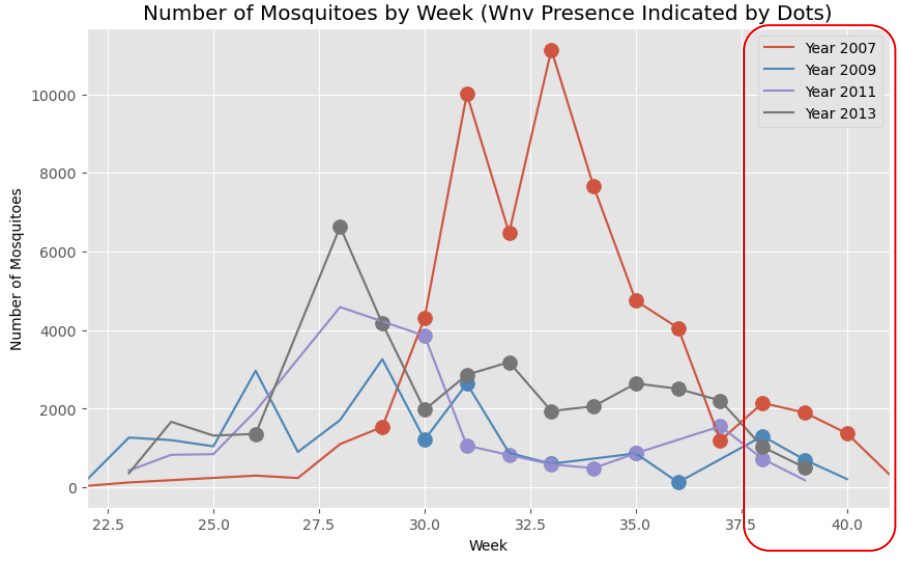
Sources from National Library of Medicine : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3367363/

ARE THE SPRAYING EFFORTS EFFECTIVE?

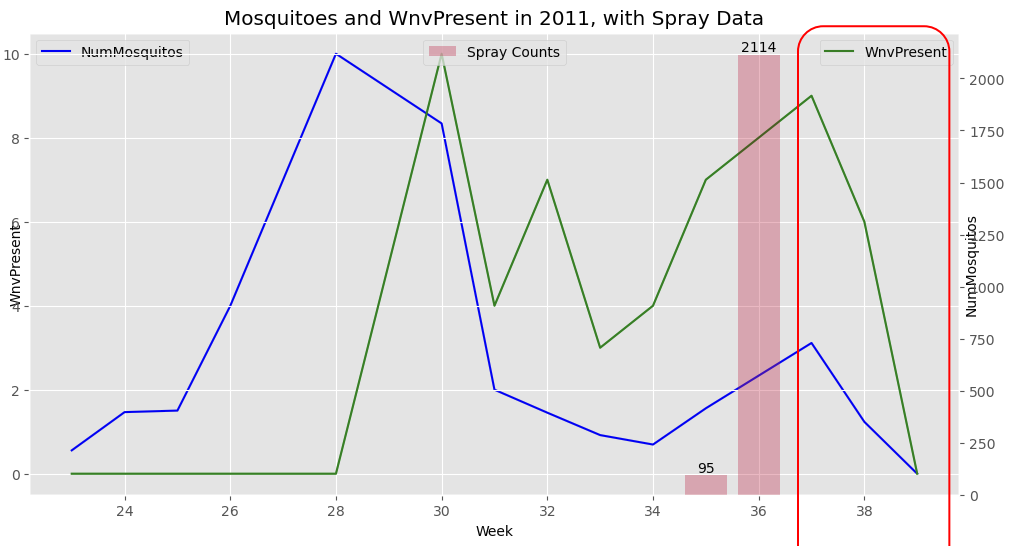
THE SPR

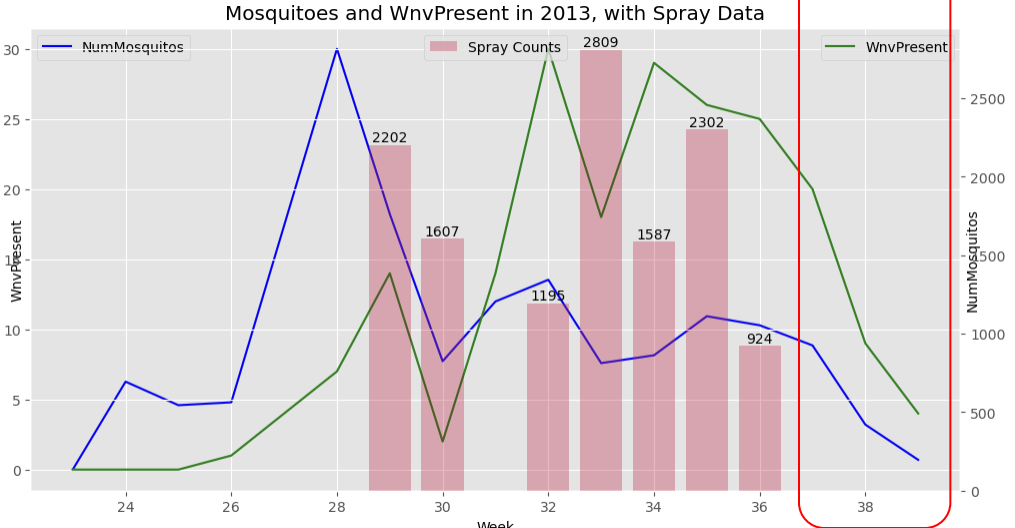


IS SPRAY EFFECTIVE?

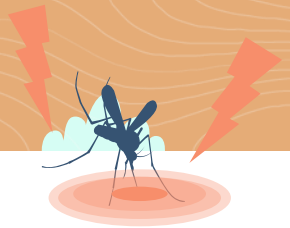


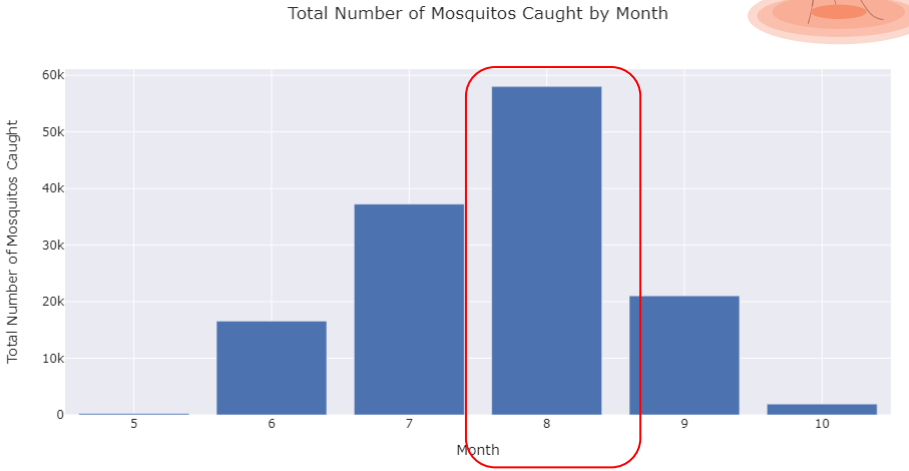
RED BAR IS SPRAY

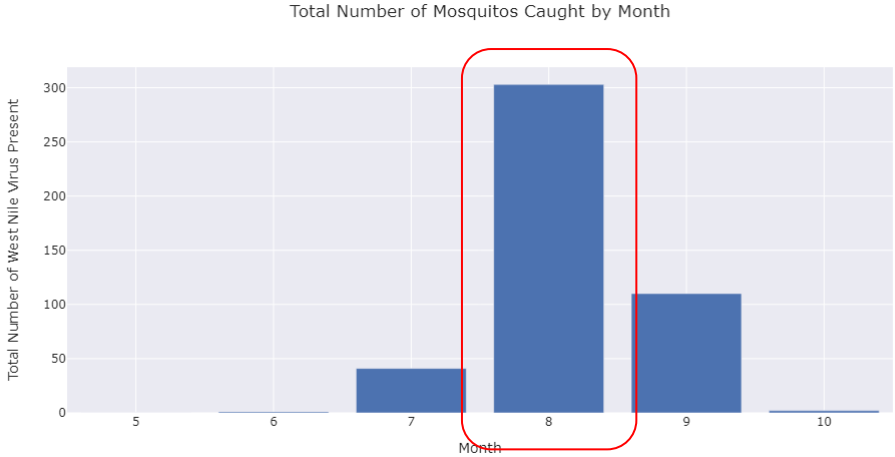




INSIGHTS

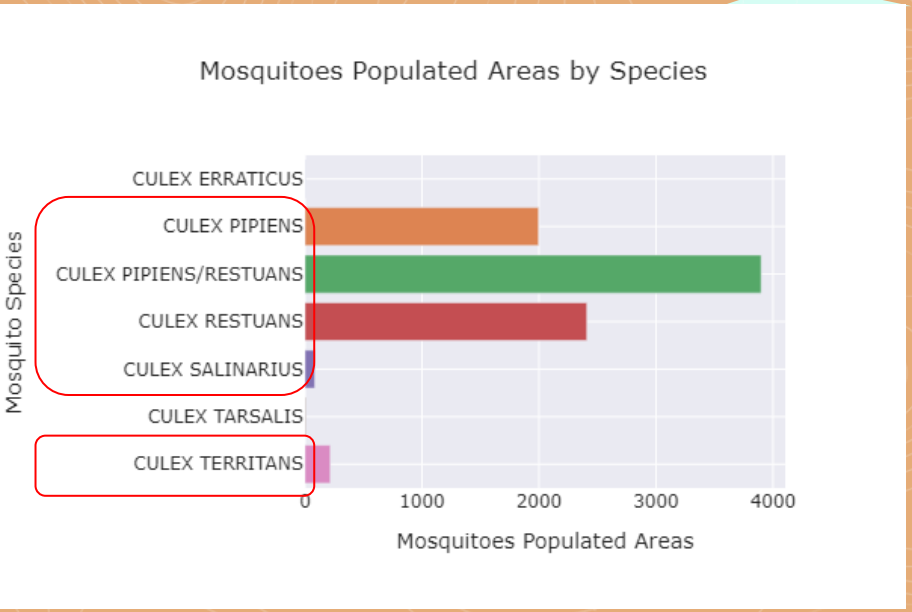


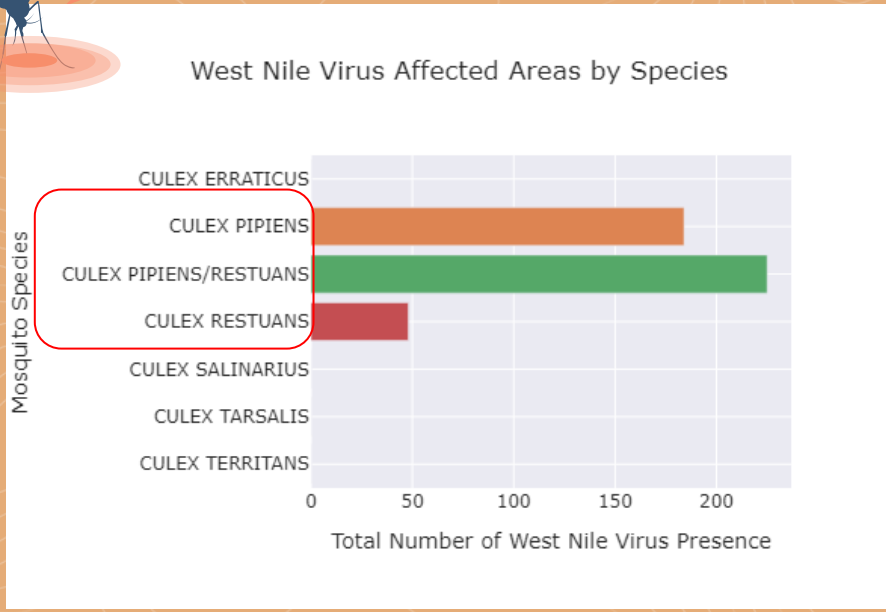




AUGUST HAS THE MOST MOSQUITOES CAUGHT AND MOST WNV PRESENT

INSIGHTS

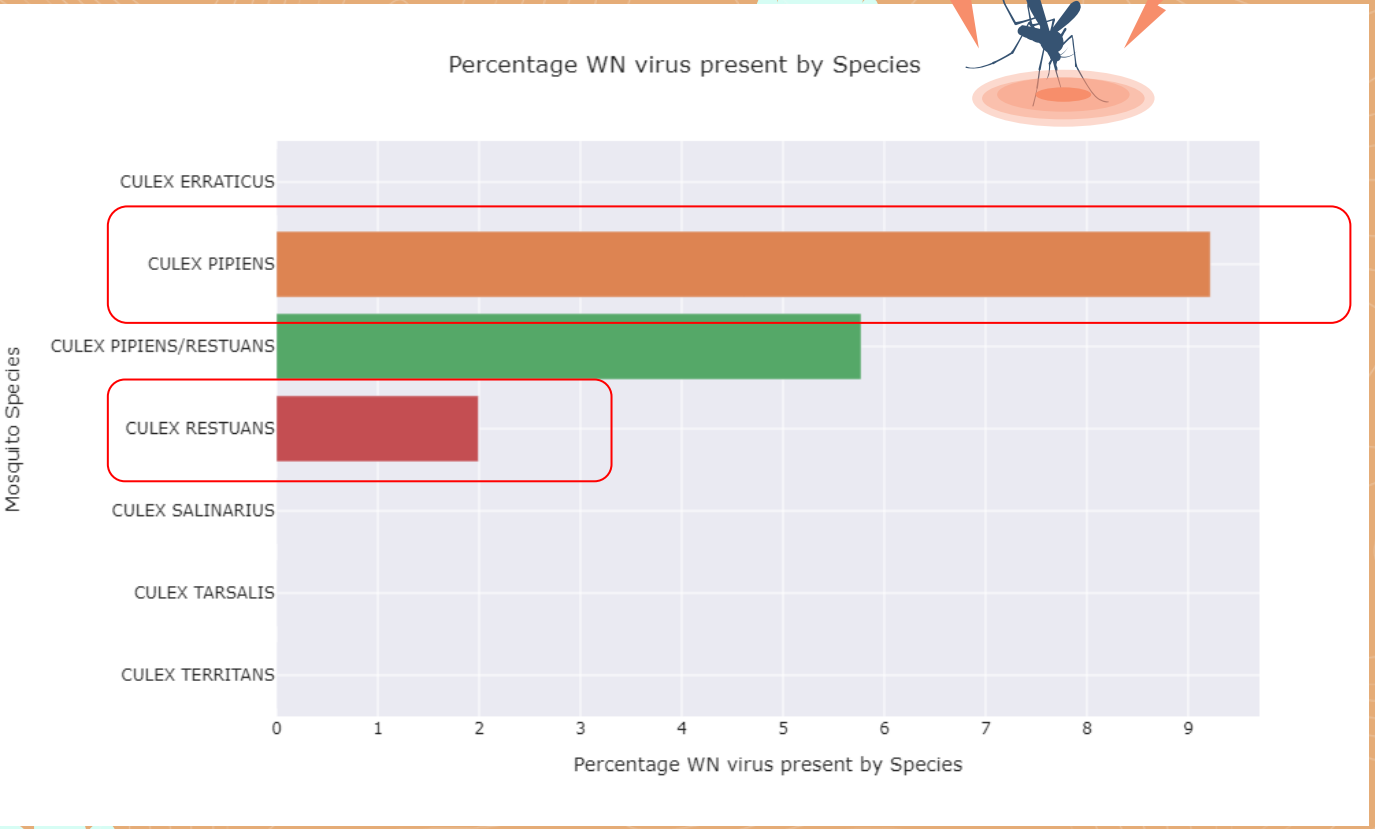




4 SPECIES OF MOSQUITOES CAUGHT IN CHICAGO

ONLY 2 SPECIES OF MOSQUITOES CARRIES WEST NILE VIRUS

DERIVED INSIGHTS



PERCENTAGE WNV

= (WNV AREA/MOSQUITO AREA)* 100

CULEX PIPIENS IS

5X

MORE LIKELY TO CARRY WNV THAN CULEX RESTUANS

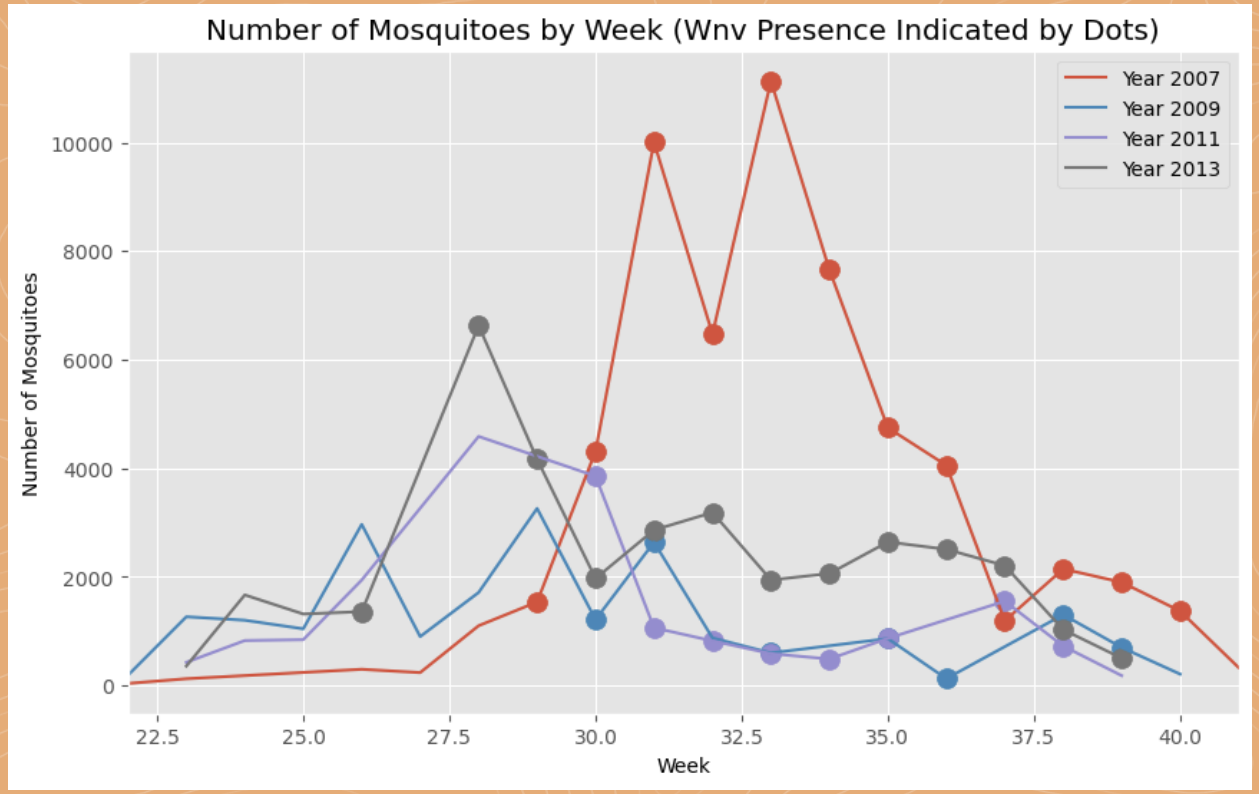
NUMBER OF MOSQUITOS CAUGHT WEEKLY

FIGURES SEEMS RANDOM

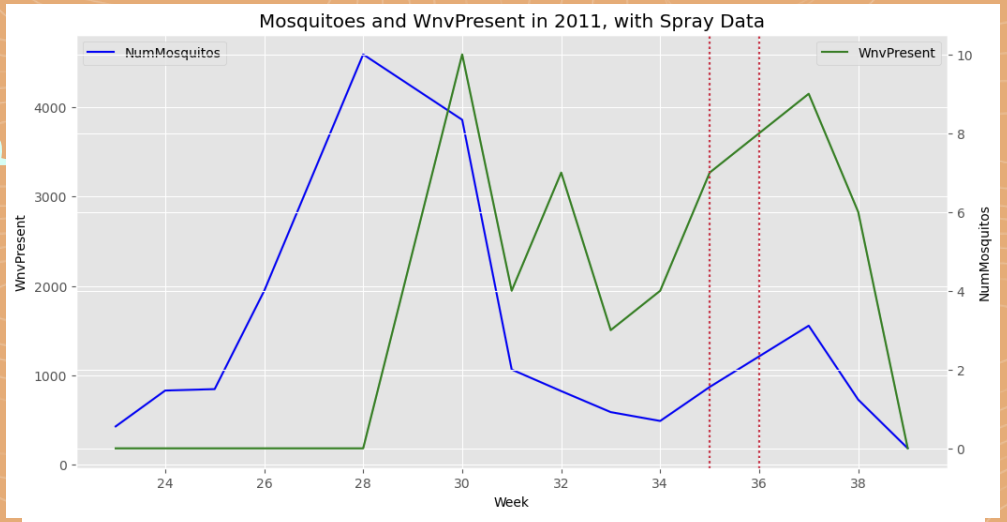
There are no clear tends observed.

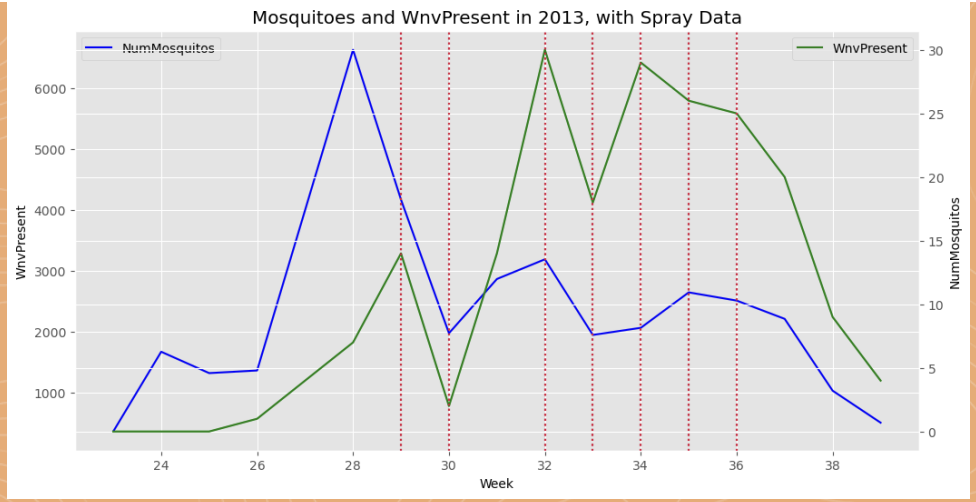
HIGH NUMBERS IN 2007

Year 2007 was observed to have higher number of mosquitos caught.



SPRAY


Mosquitoes and WnvPresent in 2011, with Spray Data
NumMosquitos
WnvPresent
WnvPresent
NumMosquitos
Week
0
1000
2000
3000
4000
0
2
4
6
8
10
24
26
28
30
32
34
36
38


Mosquitoes and WnvPresent in 2013, with Spray Data
NumMosquitos
WnvPresent
WnvPresent
NumMosquitos
Week
1000
2000
3000
4000
5000
6000
0
5
10
15
20
25
30
24
26
28
30
32
34
36
38

FEATURE IMPORTANCE LIST.

PLACEHOLDER
WHILE WE WAIT FOR SHAPLEY
TO BE OUT. BUT IT DOESNT SEEM
TO BE WORKING.

IF WE ARE USING THIS TABLE
→
THERE'S ONLY 3 VALUES.....

	features	coefficient
0	longitude	0.426667
1	latitude	0.123333
2	year	0.023333
3	block	0.000000
4	mixed_tmax	0.000000
5	street	0.000000
6	species	0.000000
7	stat_2_precip_total	0.000000
8	stat_2_tavg	0.000000
9	stat_2_tmin	0.000000
10	stat_2_tmax	0.000000
11	mixed_precip_total	0.000000
12	mixed_tmin	0.000000
13	sunrise_hours	0.000000
14	sunset_hours	0.000000
15	day_length_nearh	0.000000
16	day_length_mprec	0.000000
17	stat_1_precip_total	0.000000
18	stat_1_tavg	0.000000
19	stat_1_tmin	0.000000

CALCULATION

$$\frac{66+25+306}{1324+66+25+306} = 0.2306798373$$

$$\boxed{0.2306798373}_{\text{ans}} \cdot 3.53 = 0.8142998257$$