West Nile Story

Presented by: Samuel, Zhi Sheng, Lindy

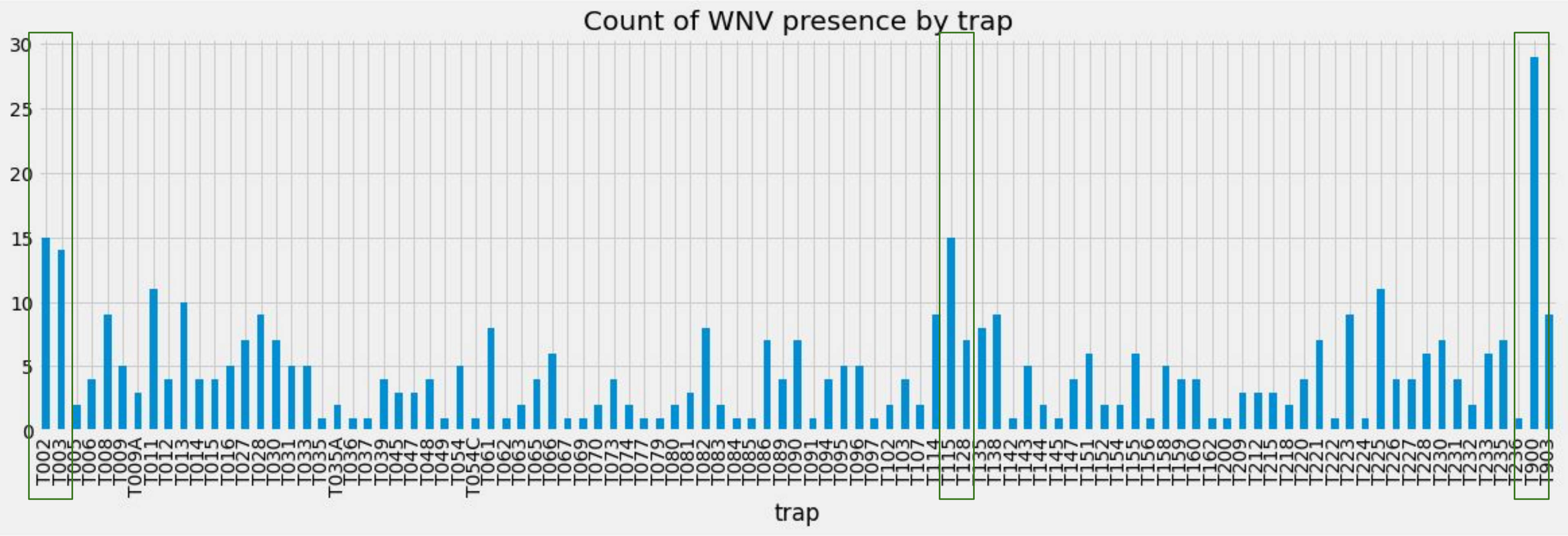
Problem Statement

1. As an employee of Disease And Treatment Agency, division of Societal Cures In Epidemiology and New Creative Engineering (DATA-SCIENCE), we are tasked to better understand the mosquito population and advise on appropriate interventions which are beneficial and cost-effective for the city.

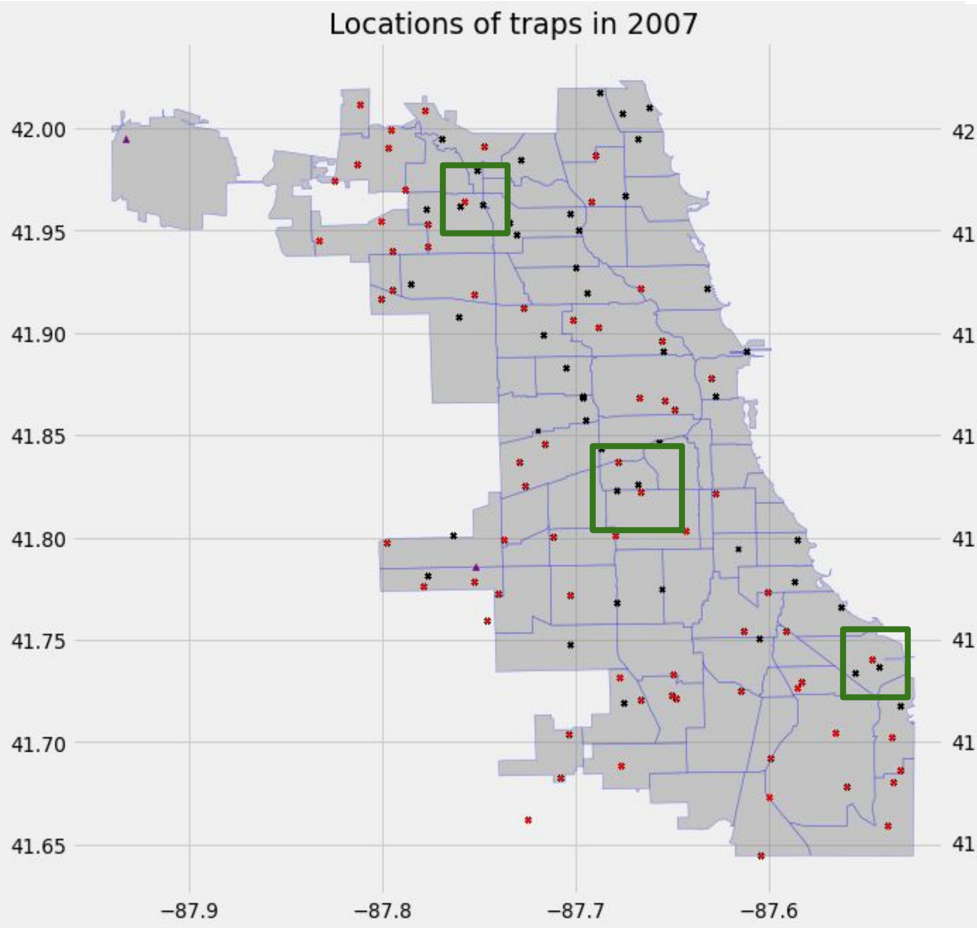
2. Through this exploration, we hope to:

- Identify features which are most important to predict presence of West Nile Virus (which can be done by ranking the coefficients of each feature in a logistic regression model)
- Predict the probability of West Nile Virus by location to provide decision makers an effective plan to deploy pesticides throughout the city, which consequently can help to reduce cost.

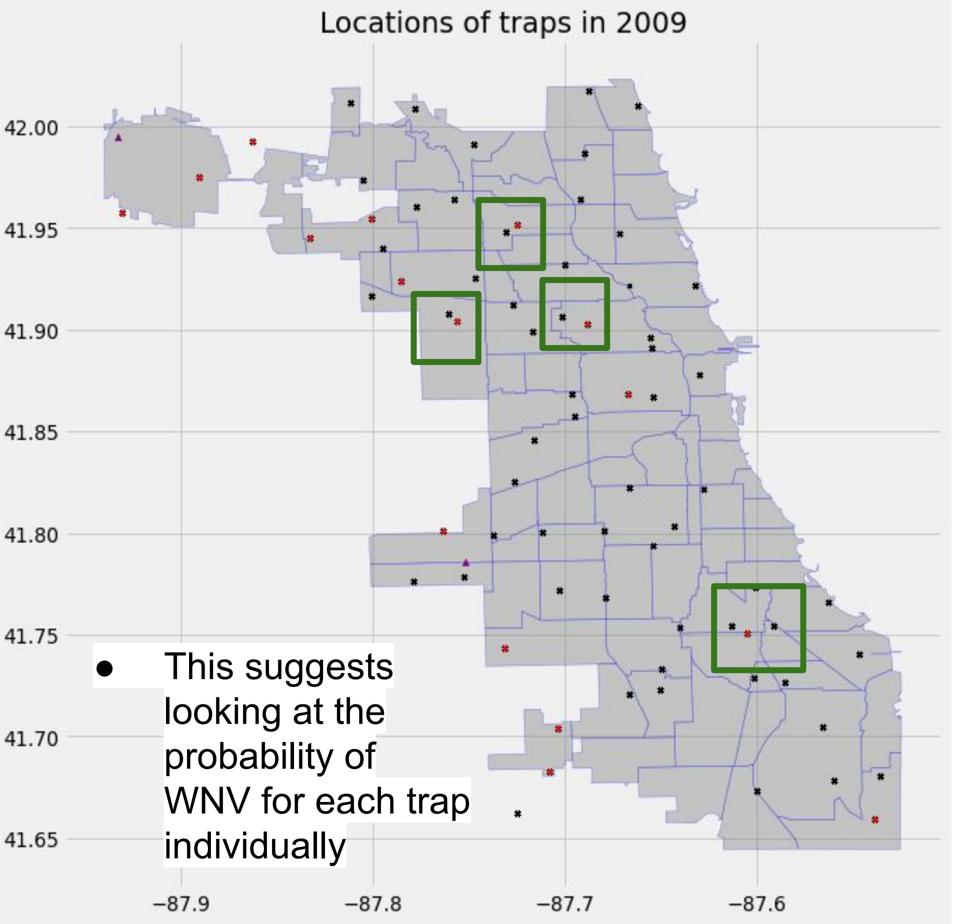
EDA: WNV has the highest occurrence in trap T900, T115, T002, T003, and T011



EDA: Hotspots may not be clustered geographically



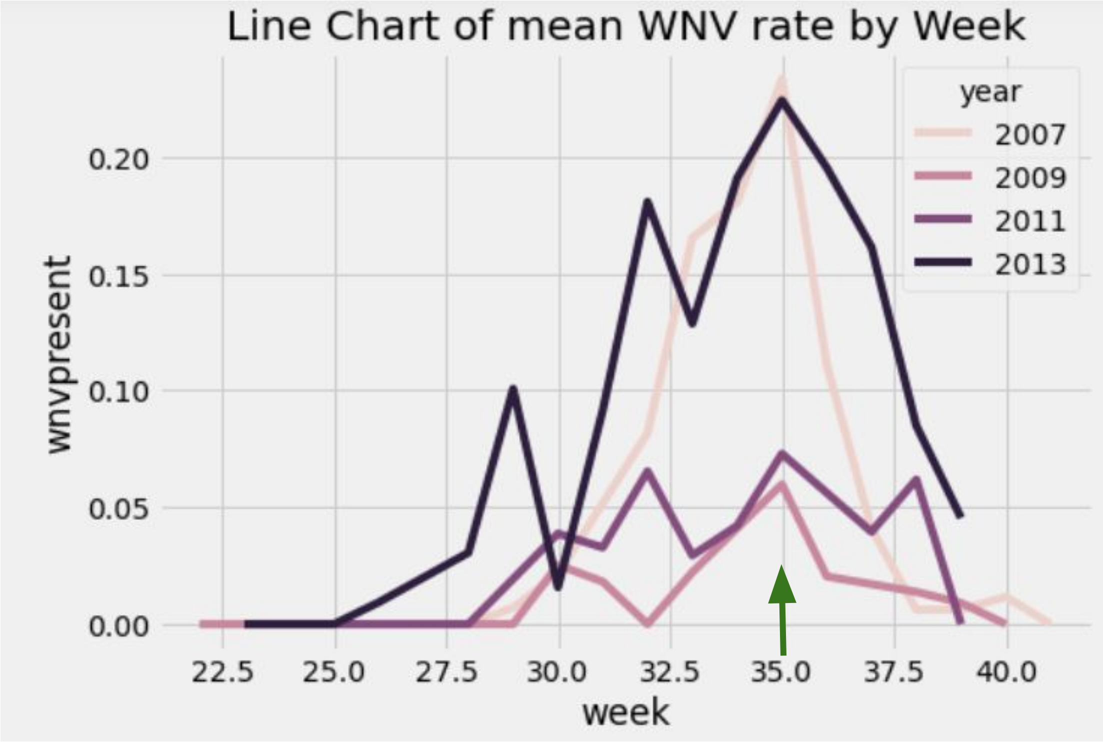
Locations of traps in 2007



Locations of traps in 2009

- This suggests looking at the probability of WNV for each trap individually

EDA: WNV peaking in week 35 could be due to dispersion of migratory birds, summer activities, and longer days

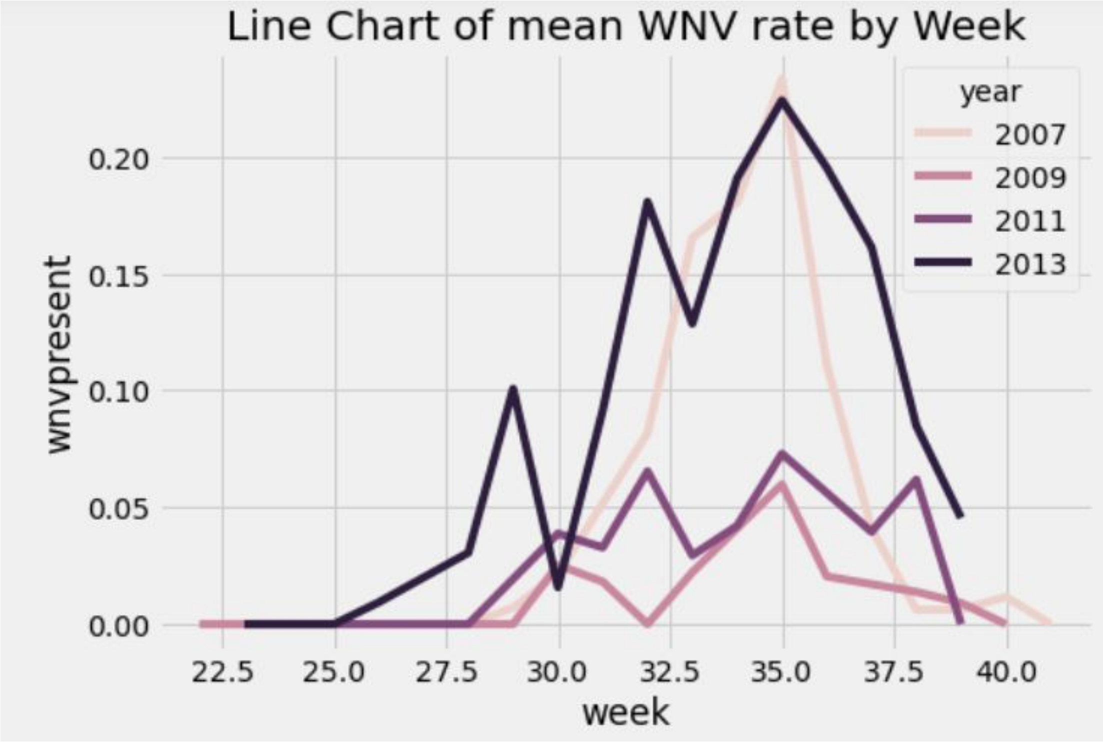


- July dispersion of robins (*Turdus migratorius*) is associated with shift among *Culex* mosquitoes from avian to human hosts
- More human activities during the summer seasons
- Increased activity of mosquitoes with longer days (mosquitos are active during dawn and dusk)

→ Increased human cases of WNV

Reference source: Association of West Nile virus illness and urban landscapes in Chicago and Detroit (https://ij-healthgeographics.biomedcentral.com/articles/10.1186/1476-072X-6-10)

EDA: WNV does not increase linearly with week

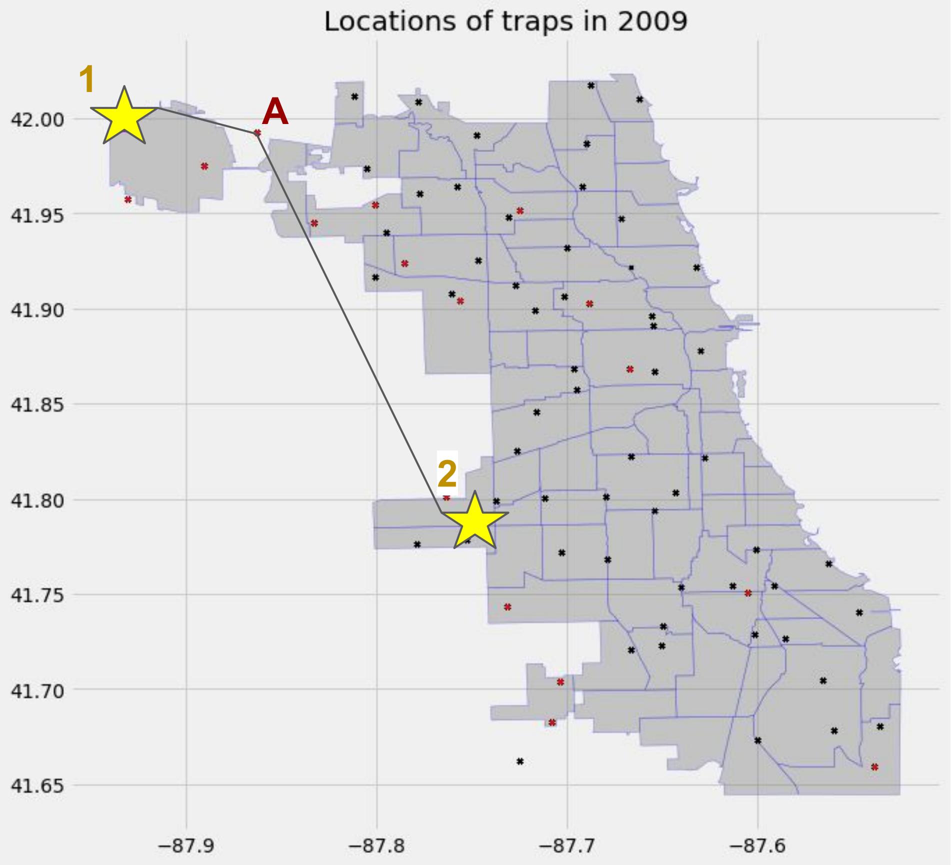


- WNV increases from week 25, peaked at week 35 and dips thereafter
- Considering non-linear trend between WNV and week, we dummified weeks to account for the non-linear relationship

Assign weather readings of traps based on the closest station

- Results are from two weather stations in Chicago which are quite far apart

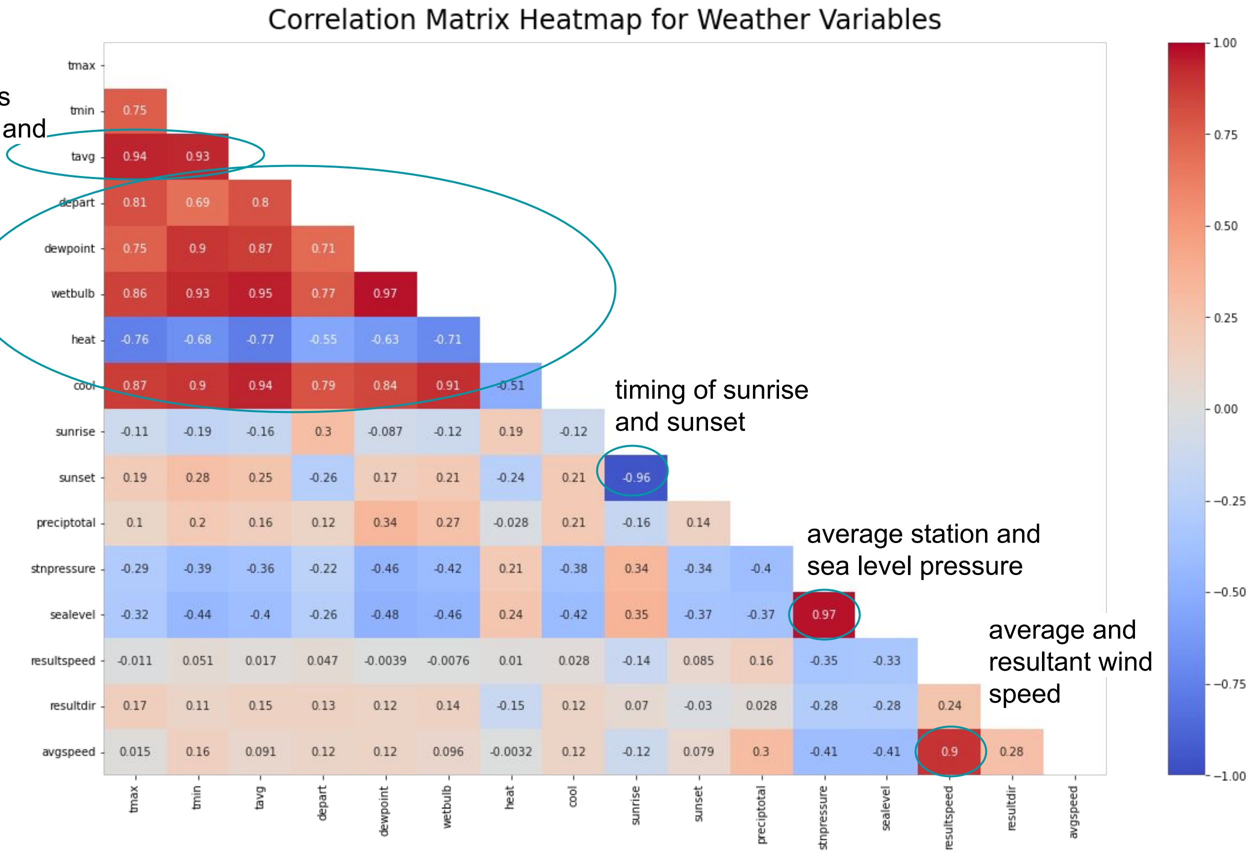
- To better factor the weather conditions at each trap location, we assume that it follows the weather conditions collected at the nearest station
  - E.g. Weather condition at trap A is based on weather condition measured at Substation 1



EDA: Many weather features have high collinearity with each other are excluded

Average temperature is simple average of min and max temperature

Measures related to temperature



Correlation Matrix Heatmap for Weather Variables

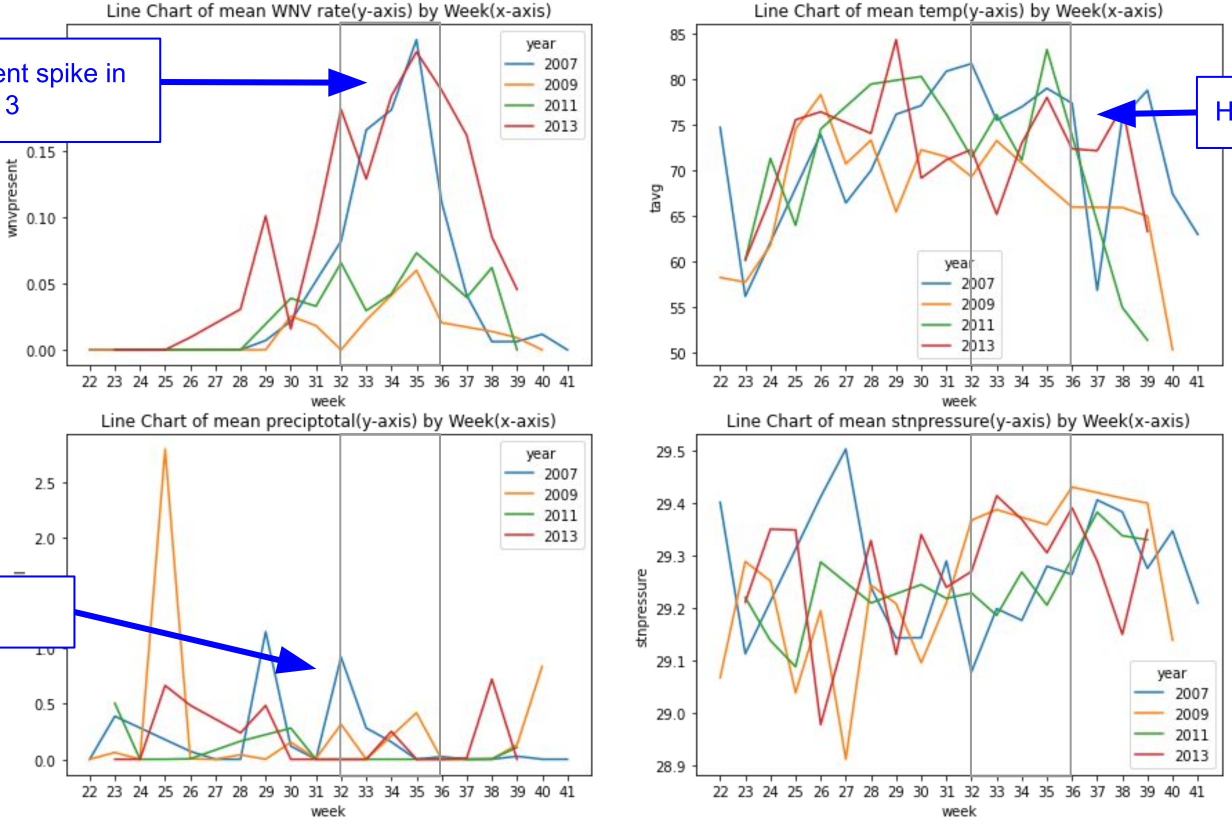
	tmax	tmin	tavg	depart	dewpoint	wetbulb	heat	cool	sunrise	sunset	preciptotal	stnpressure	sealevel	resultspeed	resultdir	avgspeed
tmax																
tmin	0.75															
tavg	0.94	0.93														
depart	0.81	0.69	0.8													
dewpoint	0.75	0.9	0.87	0.71												
wetbulb	0.86	0.93	0.95	0.77	0.97											
heat	-0.76	-0.68	-0.77	-0.55	-0.63	-0.71										
cool	0.87	0.9	0.94	0.79	0.84	0.91	-0.51									
sunrise	-0.11	-0.19	-0.16	0.3	-0.087	-0.12	0.19	-0.12								
sunset	0.19	0.28	0.25	-0.26	0.17	0.21	-0.24	0.21	-0.96							
preciptotal	0.1	0.2	0.16	0.12	0.34	0.27	-0.028	0.21	-0.16	0.14						
stnpressure	-0.29	-0.39	-0.36	-0.22	-0.46	-0.42	0.21	-0.38	0.34	-0.34	-0.4					
sealevel	-0.32	-0.44	-0.4	-0.26	-0.48	-0.46	0.24	-0.42	0.35	-0.37	-0.37	0.97				
resultspeed	-0.011	0.051	0.017	0.047	-0.0039	-0.0076	0.01	0.028	-0.14	0.085	0.16	-0.35	-0.33			
resultdir	0.17	0.11	0.15	0.13	0.12	0.14	-0.15	0.12	0.07	-0.03	0.028	-0.28	-0.28	0.24		
avgspeed	0.015	0.16	0.091	0.12	0.12	0.096	-0.0032	0.12	-0.12	0.079	0.3	-0.41	-0.41	0.9	0.28	

timing of sunrise and sunset

average station and sea level pressure

average and resultant wind speed

EDA: WNV spike was more pronounced in 2007 and 2013 due to higher temperature and precipitation


Line Chart of mean WNV rate(y-axis) by Week(x-axis)

Line Chart of mean temp(y-axis) by Week(x-axis)

Line Chart of mean preciptotal(y-axis) by Week(x-axis)

Line Chart of mean stnpressure(y-axis) by Week(x-axis)

More prominent spike in 2007 and 2013

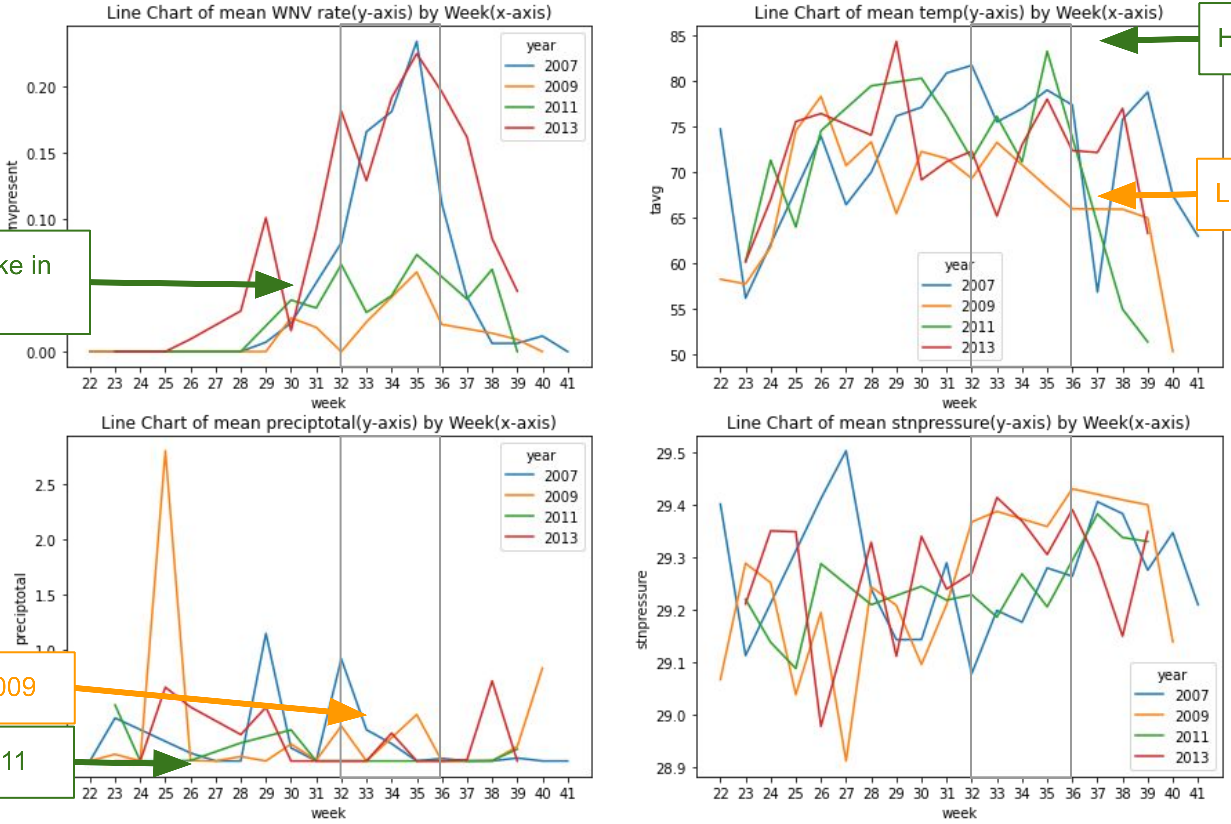
High temperature

High precipitation

- This suggests that interaction between weather features can be important

Reference source: When Is Mosquito Season In Your State? (https://www.mosquitomagnet.com/articles/mosquito-season)

EDA: WNV spike was less pronounce in 2009 and 2011 due to only high temperature or high precipitation



High temperature in 2011

Low temperature in 2009

Less prominent spike in 2009 and 2011

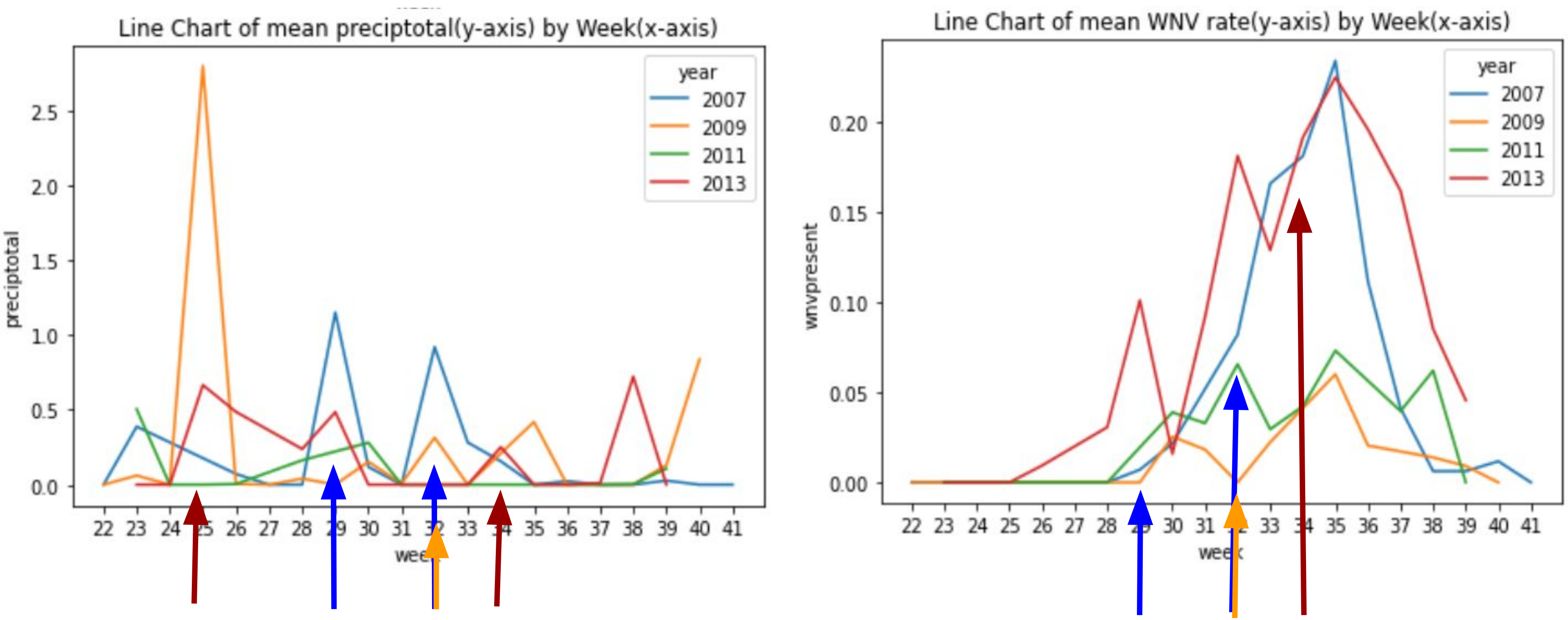
High precipitation in 2009

Low precipitation in 2011

- This suggests that interaction between weather features can be important

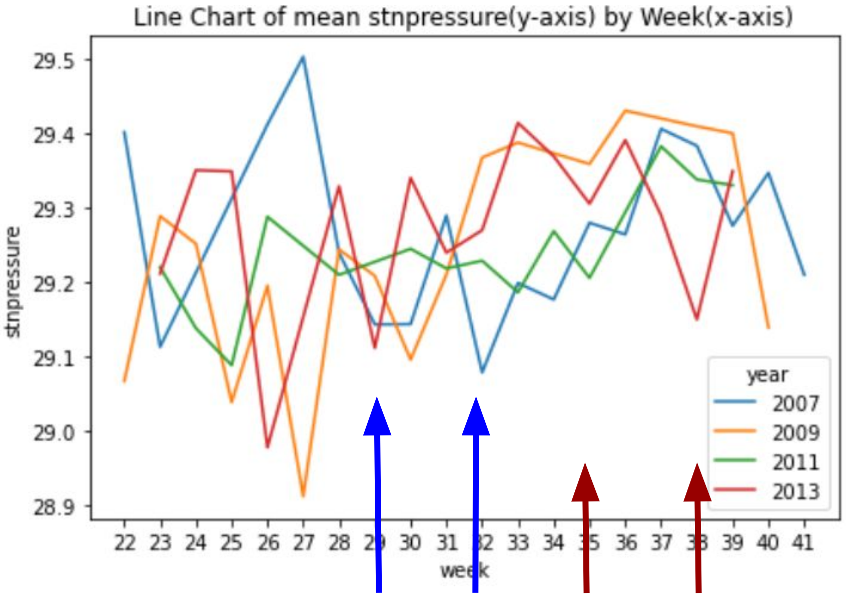
Reference source: When Is Mosquito Season In Your State? (https://www.mosquitomagnet.com/articles/mosquito-season)

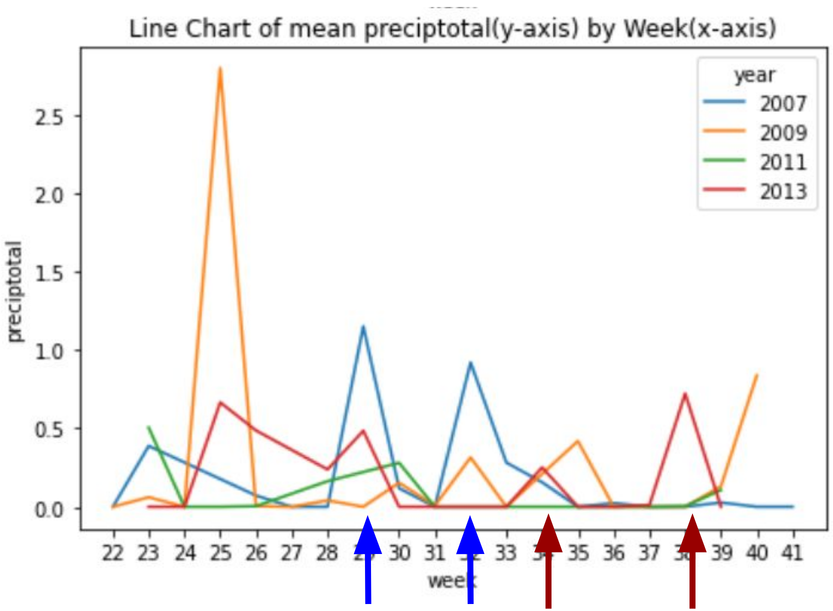
EDA: Total precipitation has a lagged-effect on WNV



- When precipitation increases, WNV surges and reaches a new height in about a week
  - This is due to an increase in overall humidity and standing water in the environment suitable for mosquito reproduction
- This suggests that we should explore lagging precipitation (e.g. by 7 or 10 days)

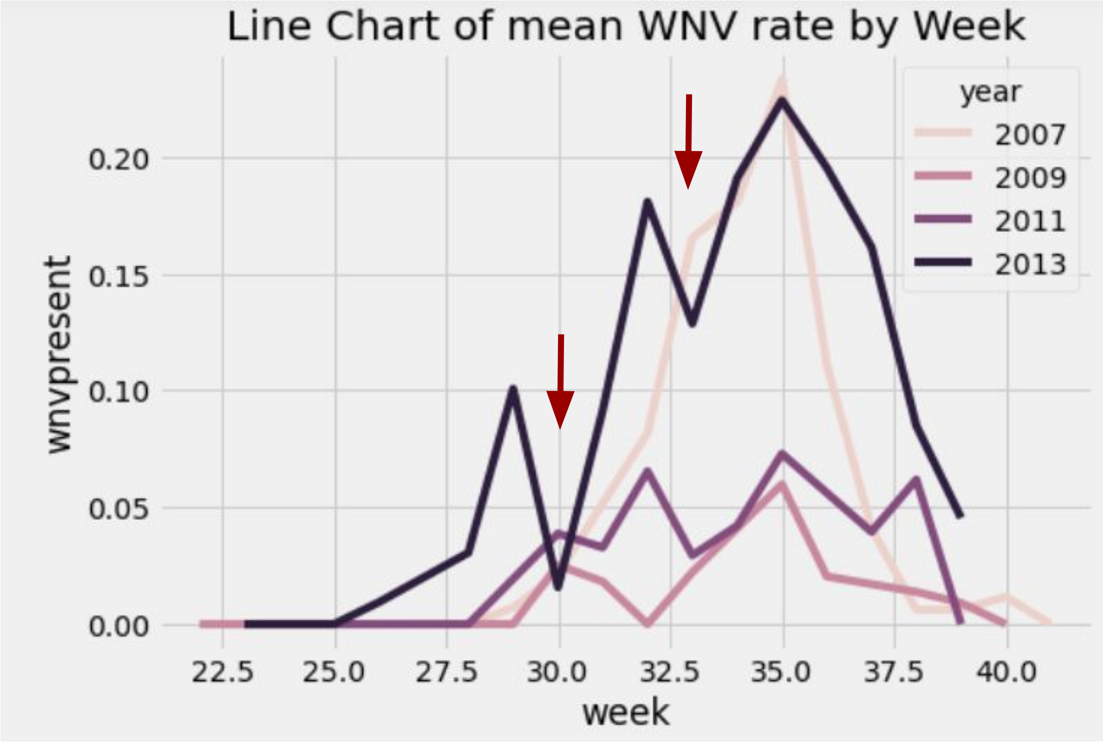
EDA: Station pressure has an inverse relationship with precipitation and consequently also has a lagged-effect on WNV





- When pressure decreases, precipitation increases
    - This is due to a rise in air movement, causing water vapour to condense, forming clouds and precipitation
- This suggests that we should explore lagging station pressure (e.g. by 7 or 10 days)

EDA: Spray should be included in the model to better account for the true effect of weather conditions

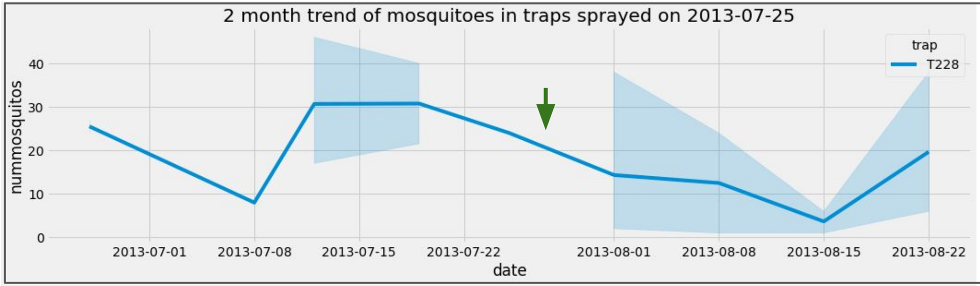


- WNV dipped in week 30 and week 33 in 2013 which coincide with date of spray

- While WNV is affected by weather conditions, it is also affected by human intervention (e.g. spray)

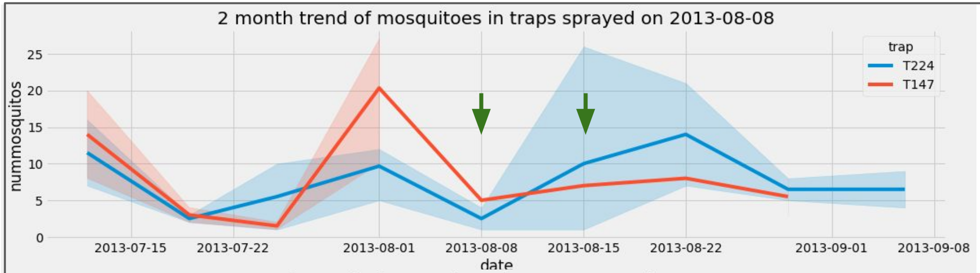
- Spray should included in the model to better account for the true effect of weather

EDA: Effect of spray in reducing mosquitoes is effective on the day of spray and persists even several days after spray

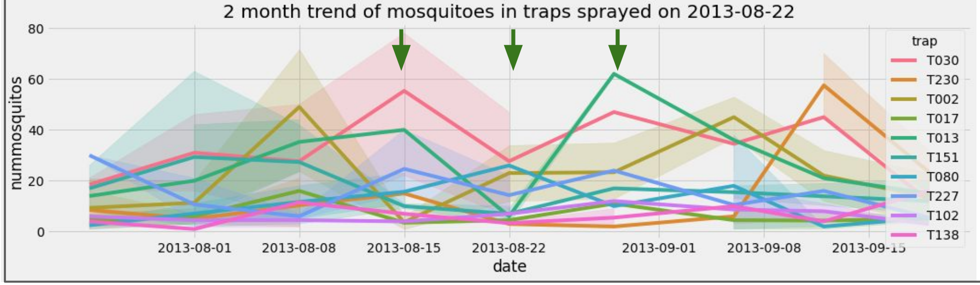


2 month trend of mosquitoes in traps sprayed on 2013-07-25

- Number of mosquitoes at trap T228 decreased gradually in the next 20 days after spray



2 month trend of mosquitoes in traps sprayed on 2013-08-08



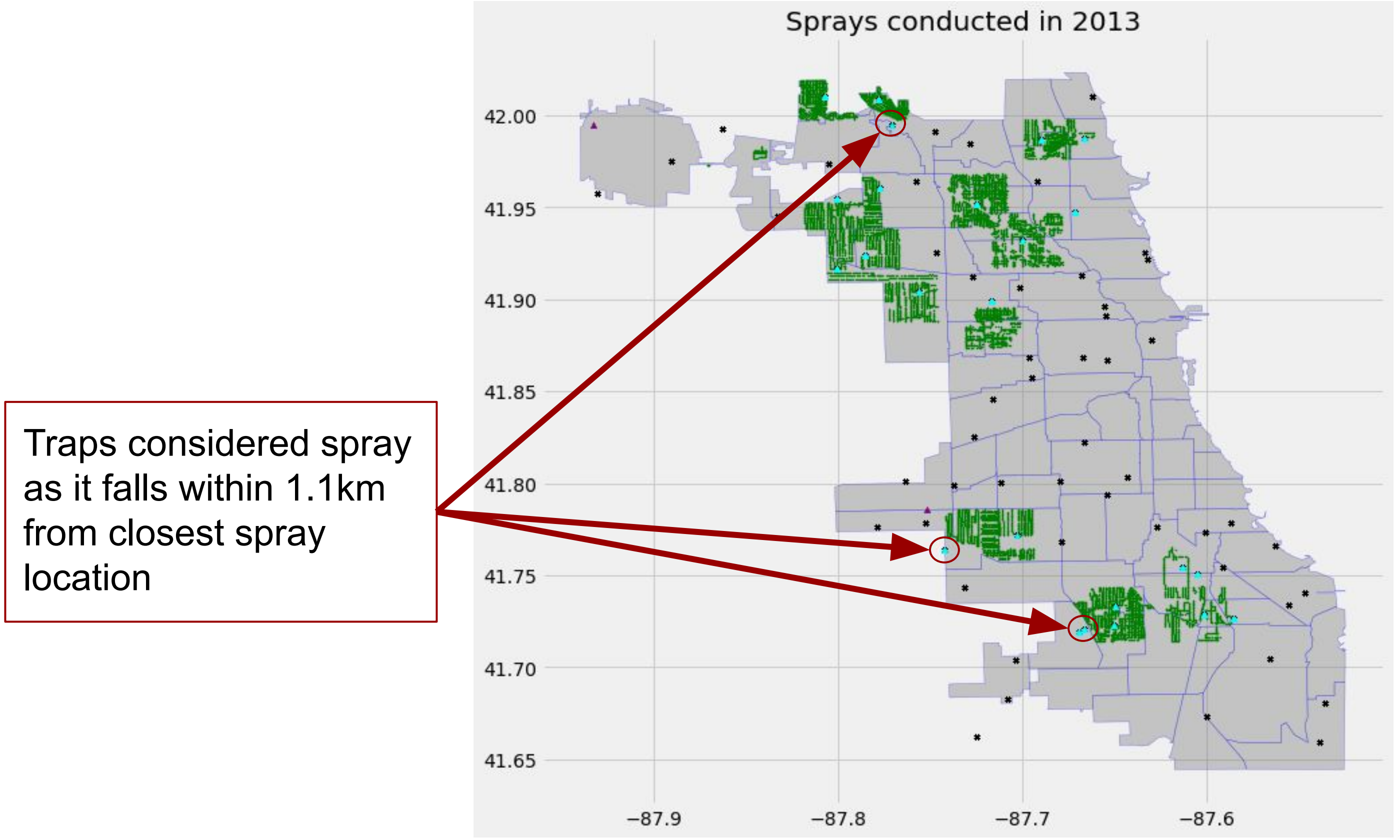
2 month trend of mosquitoes in traps sprayed on 2013-08-22

- Weekly spray was administered in Aug 2013 for 4 consecutive weeks

- Number of mosquitoes at trap T224 and T147 dipped on the day of spray, with dampened growth over the next 2 weeks. Traps not sprayed 2 weeks later.

- Most traps saw decrease in number of mosquitoes within 1 week from the day of spray (e.g. T030, T013, T227)

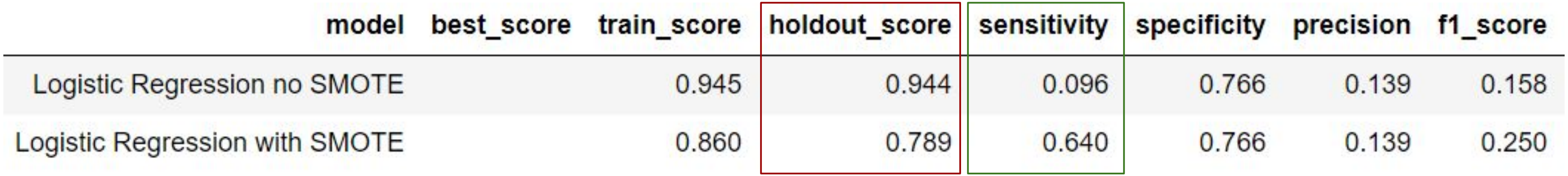
Assign trap as sprayed if there was a spray within 1.1km (0.01 degree in coordinates) in the past 7 days



Traps considered spray as it falls within 1.1km from closest spray location

Model Selection: Using our base model (*Logistic Regression*), using SMOTE increases sensitivity from 10% to 64%

- **Why Logistic Regression**: Ability to quantify impact of features on probability of WNV
- **Why SMOTE**: Ensured that minority class is represented
  - WNV is highly imbalanced (~5.5% with WNV)
  - Stratify proportionally to ensure same proportion of WNV in train and holdout dataset
  - Use Synthetic Minority Oversampling TEchnique (SMOTE) to balance class distribution - randomly increasing minority class by replicating them



model	best_score	train_score	holdout_score	sensitivity	specificity	precision	f1_score
Logistic Regression no SMOTE		0.945	0.944	0.096	0.766	0.139	0.158
Logistic Regression with SMOTE		0.860	0.789	0.640	0.766	0.139	0.250

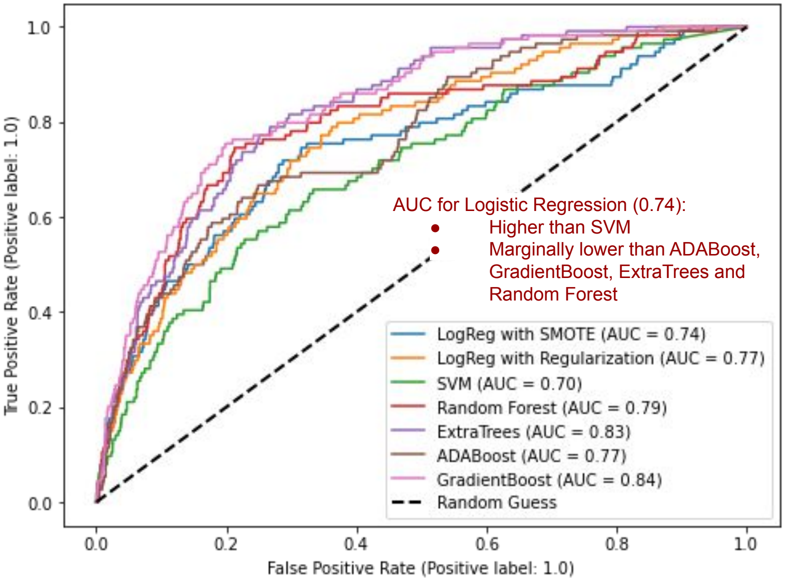
- Accuracy without SMOTE is 94.5%, close to proportion of WNV absent
- Accuracy decreases with SMOTE, indicating lesser TP

- Sensitivity increases with SMOTE to 64%, indicating that ~64% of WNV+ traps are captured by our model
- Use SMOTE across all exploratory models

Model Selection: Benchmark LogReg against SVM, Decision Trees & Neural Network

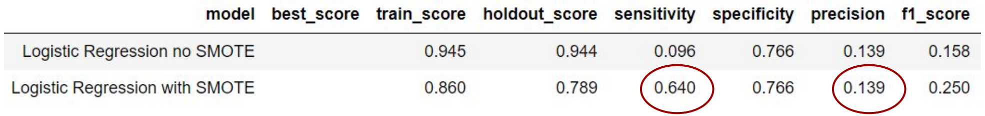
- Evaluate models based on sensitivity and precision
  - Sensitivity: Pick up more WNV+ to provide early/preemptive intervention (i.e. spray)
  - Precision: Better manage spraying cost by increasing correctness amongst our positive predictions

Comparing with Random Forest, ExtraTrees, SVM, ADABoost and Gradient Boost

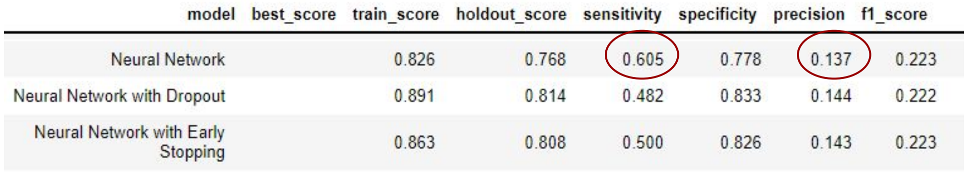


AUC for Logistic Regression (0.74):
- Higher than SVM
- Marginally lower than ADABoost, GradientBoost, ExtraTrees and Random Forest

Comparing with Neural Network



model	best_score	train_score	holdout_score	sensitivity	specificity	precision	f1_score
Logistic Regression no SMOTE		0.945	0.944	0.096	0.766	0.139	0.158
Logistic Regression with SMOTE		0.860	0.789	0.640	0.766	0.139	0.250



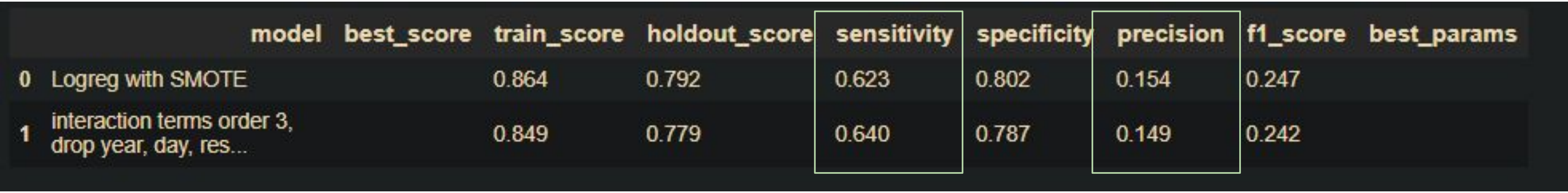
model	best_score	train_score	holdout_score	sensitivity	specificity	precision	f1_score
Neural Network		0.826	0.768	0.605	0.778	0.137	0.223
Neural Network with Dropout		0.891	0.814	0.482	0.833	0.144	0.222
Neural Network with Early Stopping		0.863	0.808	0.500	0.826	0.143	0.223

Sensitivity and precision are comparable with Neural Network

Final Model: **Logistic Regression with SMOTE**
- Ability to quantify impact of features on probability of WNV

Model Tuning: Further model tuning of Logistic Regression increase sensitivity by 2%-pt

- Further tuning to our selected model (Logistic Regression with SMOTE)
  - Removed Year, Day of Week
    - Still keep Week as it may account for the migratory pattern of birds that carry WNV
  - Removed wind direction
  - Feature engineering for interaction terms on 4 key weather features based on EDA:
    - Average Temperature
    - Precipitation (Lag 10 days)
    - Station Pressure
    - Average Wind Speed

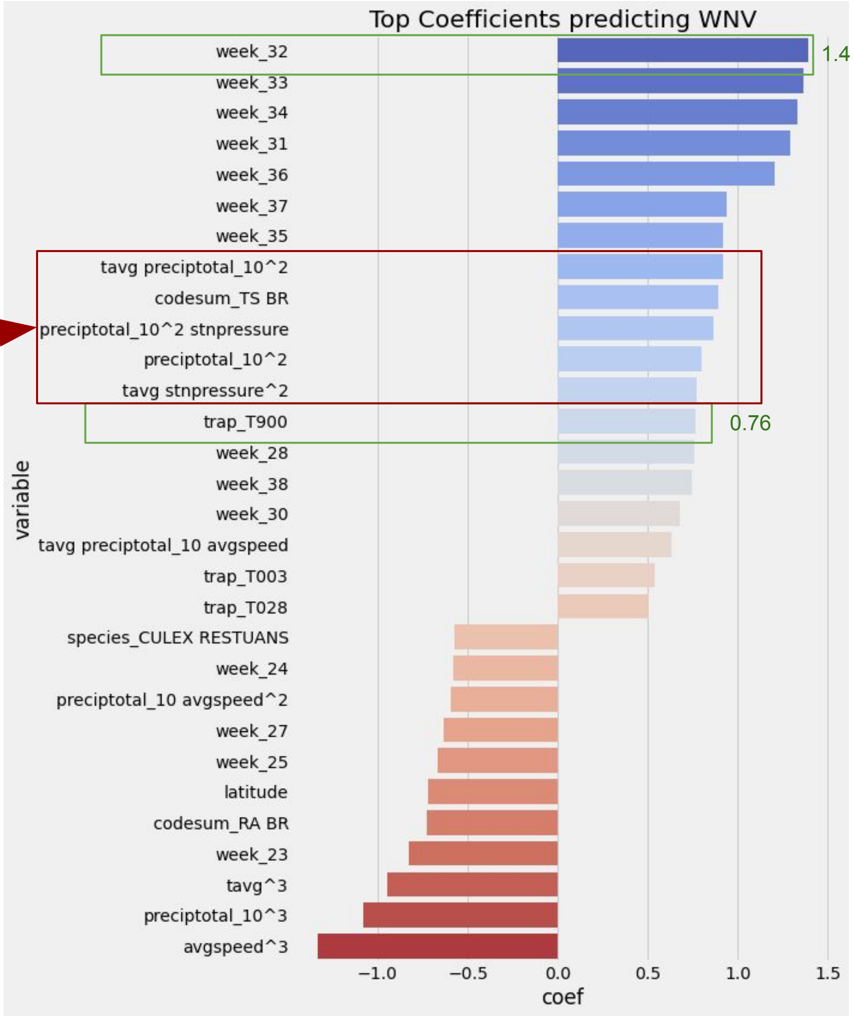


	model	best_score	train_score	holdout_score	sensitivity	specificity	precision	f1_score	best_params
0	Logreg with SMOTE		0.864	0.792	0.623	0.802	0.154	0.247	
1	interaction terms order 3, drop year, day, res...		0.849	0.779	0.640	0.787	0.149	0.242	

- Further increase sensitivity by 2%-pt
- Precision is comparable

Top features are weeks, weather and specific traps

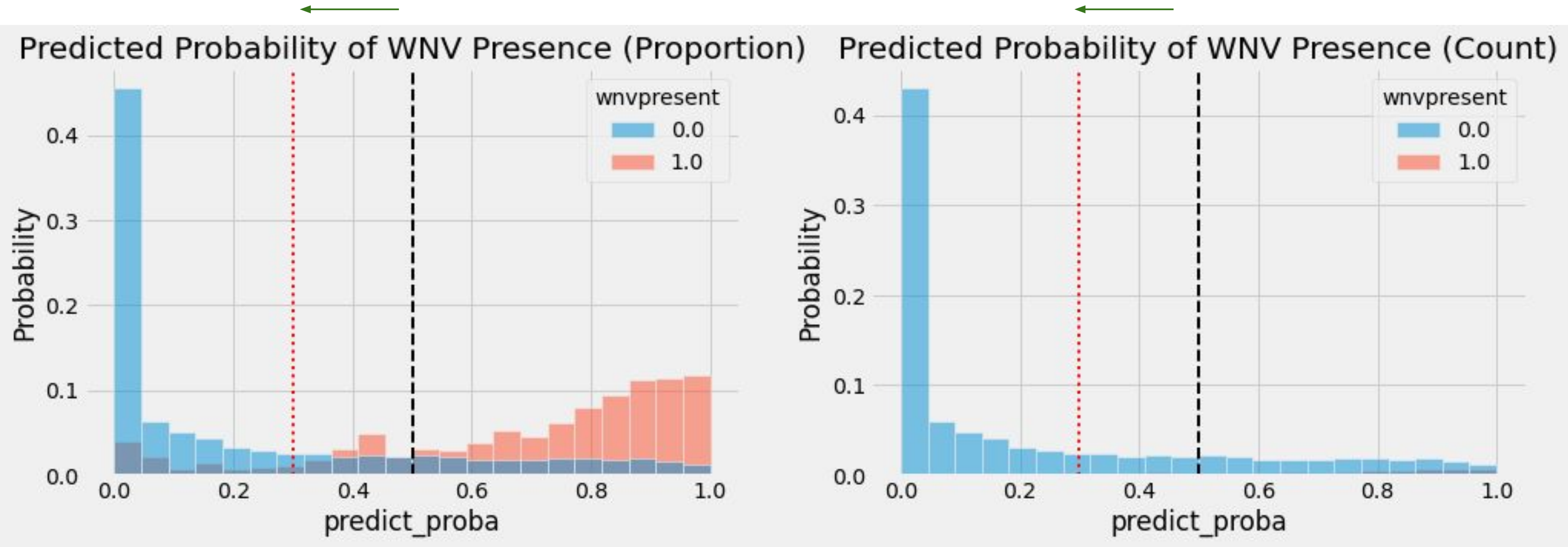
- Most important features:
  - Time of the year (week)
  - A combination of weather conditions (weather interaction terms)
  - Certain traps (e.g. T900)
- Interpreting coefficients*:
  - Probability of WNV increases by 4 times^ in week 32
  - Probability of WNV increases by 2.1 times^ at trap T900



* Interpretation of weather features are more difficult due to interaction terms. Details on interpreting interaction terms in logistic regression can be found here: https://www.cantab.net/users/filimon/cursoFCDEF/will/logistic_interact.pdf

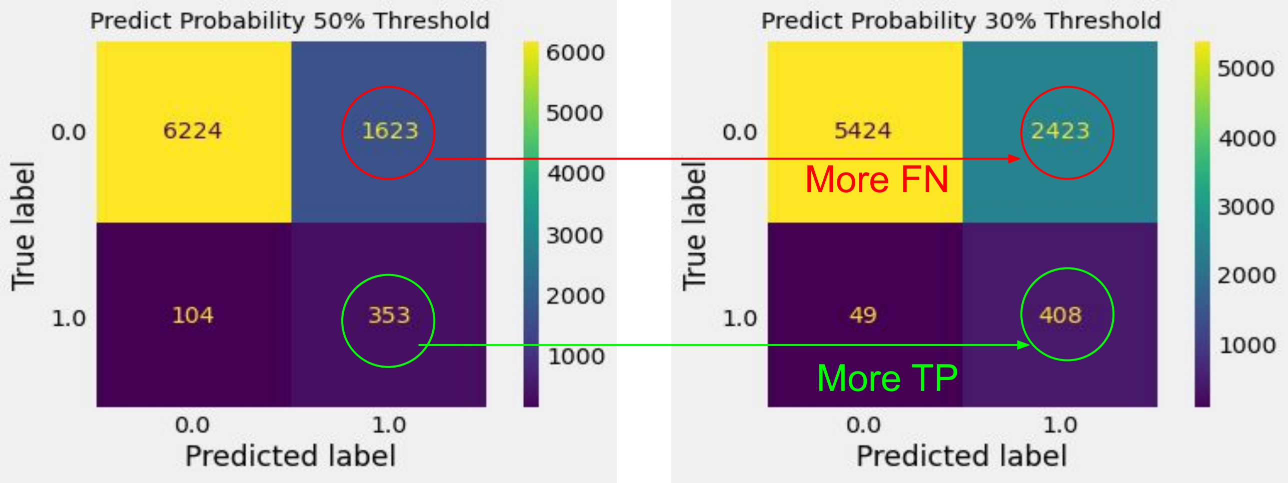
^ Probability of WNV increases by 4 times (exp(1.4)) and 2.1 times (exp(0.76)) in week 32 and at trap T900 respectively

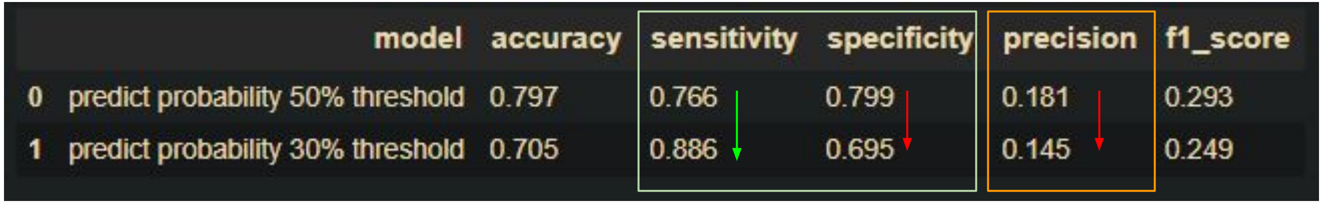
Model Tuning: Optimizing Prediction Probability Threshold



Reduced the threshold of prediction probability from 50% to 30%

Model Tuning: Sensitivity increased by 12%-point





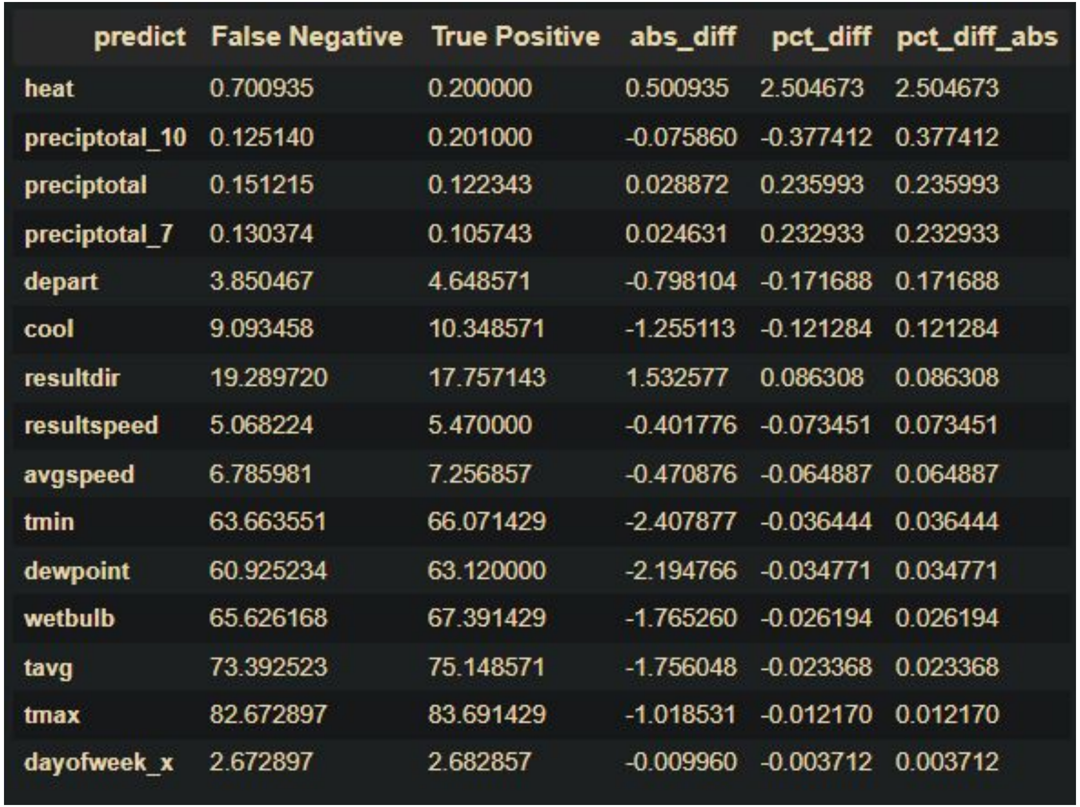
	model	accuracy	sensitivity	specificity	precision	f1_score
0	predict probability 50% threshold	0.797	0.766	0.799	0.181	0.293
1	predict probability 30% threshold	0.705	0.886	0.695	0.145	0.249

- Trade-off between Sensitivity (+8%) and Specificity (-10%)
- Slight decrease in precision

Error Analysis of False Negative

- Differences between the mean of the features

- Further study could look into more interaction terms and varying lag in weather features



predict	False Negative	True Positive	abs_diff	pct_diff	pct_diff_abs
heat	0.700935	0.200000	0.500935	2.504673	2.504673
preciptotal_10	0.125140	0.201000	-0.075860	-0.377412	0.377412
preciptotal	0.151215	0.122343	0.028872	0.235993	0.235993
preciptotal_7	0.130374	0.105743	0.024631	0.232933	0.232933
depart	3.850467	4.648571	-0.798104	-0.171688	0.171688
cool	9.093458	10.348571	-1.255113	-0.121284	0.121284
resultdir	19.289720	17.757143	1.532577	0.086308	0.086308
resultspeed	5.068224	5.470000	-0.401776	-0.073451	0.073451
avgspeed	6.785981	7.256857	-0.470876	-0.064887	0.064887
tmin	63.663551	66.071429	-2.407877	-0.036444	0.036444
dewpoint	60.925234	63.120000	-2.194766	-0.034771	0.034771
wetbulb	65.626168	67.391429	-1.765260	-0.026194	0.026194
tavg	73.392523	75.148571	-1.756048	-0.023368	0.023368
tmax	82.672897	83.691429	-1.018531	-0.012170	0.012170
dayofweek_x	2.672897	2.682857	-0.009960	-0.003712	0.003712

Cost-Benefit Analysis

- Spraying of pesticides can help to reduce the population of mosquitoes, which can carry WNV.
- Zenivex is a mosquito adulticide chosen by the Chicago government
    - Extremely low toxicity to mammals
    - Cost of spraying Zenivex is **0.67 cents per acre**

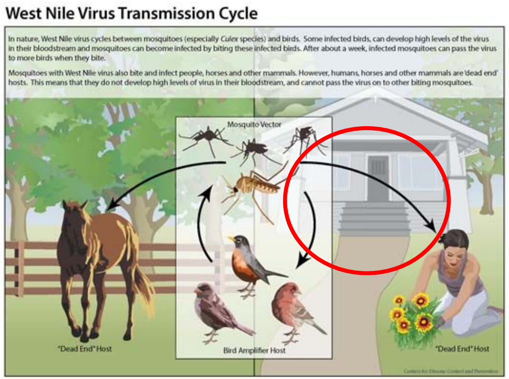
- Assumption is that it is extremely efficacious in reducing mosquitoes.

- **Look to balance spray cost with medical and opportunity costs of Chicago residents.**

Cost-Benefit Analysis- WNV infection calculation

- By looking at specific years,
  - Proportion of traps that have recorded WNV viruses.
  - Proxy for actual infection during that year
  - Other factors are population size and an Infection Rate

Year	Prevalence WNV in traps	Population Size	People infected by WNV	Infection Rate
2007	0.070182	2,811,035	10	5.069 e-05
2009	0.010096	2,824,064	1	3.507e-05
2011	0.029709	2,700,741	8	9.9705e-05
2013	0.097263	2,706,101	37	14.0576e-05



West Nile Virus Transmission Cycle

In nature, West Nile virus cycles between mosquitoes (especially *Culex* species) and birds. Some infected birds, can develop high levels of the virus in their bloodstream and mosquitoes can become infected by biting these infected birds. After about a week, infected mosquitoes can pass the virus to more birds when they bite.

Mosquitoes with West Nile virus also bite and infect people, horses and other mammals. However, humans, horses and other mammals are 'dead end' hosts. This means that they do not develop high levels of virus in their bloodstream, and cannot pass the virus on to other biting mosquitoes.

Centers for Disease Control and Prevention

Infection Population =

WNV Prevalence in traps * Population * <u>Infection Rate</u>

Average infection rate:

8.15107 e-05

Cost-Benefit Analysis - Medical Costs

- The cost of patients that have visited or been hospitalised due to WNV
  - West Nile fever (WNF, influenza symptoms)
    - Occurs in 20% of patients
    - Costs
      - Diagnostics visit: ~$167
      - Diagnostic test: ~$135
      - Productivity loss: ~$955 over 5 days etc
  - West Nile neuroinvasive disease (WNND)
    - Occurs in 1 out of 150 patients
    - Costs
      - Outpatient and inpatient costs
      - Nursing homes
      - Productivity loss etc.
      - Average cost over 46 patients surveyed: **~$57,070 etc**
  - Average cost for hospitalised or inpatient patients in US from 1999 through 2012

$778 million / 37,088 patients =

$20,977 per patient

Cost-Benefit Analysis - **50% Threshold (default)**

- The cost of patients that have visited or been hospitalised due to WNV
    - West Nile fever (WNF, influenza symptoms)
        - Occurs in 20% of patients
        - Costs
            - Diagnostics visit: ~$167
            - Diagnostic test: ~$135
            - Productivity loss: ~$955 over 5 days etc.
    - West Nile neuroinvasive disease (WNND)
        - Occurs in 1 out of 150 patients
        - Costs
            - Outpatient and inpatient costs
            - Nursing homes
            - Productivity loss etc.
            - Average cost over 46 patients surveyed: **~$57,070**
    - Average cost for hospitalised or inpatient patients in US from 1999 through 2012

$778 million / 37,088 patients =
$20,977 per patient

There is an estimated 44.44 patients
Total cost =
44.44 * $20,977 * **0.772** (Sensitivity) = **$719,737**

Assuming 1Km radius to be sprayed around our traps:

608 traps* $1.656 * 1000m * 1000m * 3.1415 *1.656/1000m^2

Total cost to spray = $316,225

Total amount saved = $403,512

Cost-Benefit Analysis - **30% Threshold**

- The cost of patients that have visited or been hospitalised due to WNV
  - West Nile fever (WNF, influenza symptoms)
    - Occurs in 20% of patients
    - Costs
      - Diagnostics visit: ~$167
      - Diagnostic test: ~$135
      - Productivity loss: ~$955 over 5 days etc.
  - West Nile neuroinvasive disease (WNND)
    - Occurs in 1 out of 150 patients
    - Costs
      - Outpatient and inpatient costs
      - Nursing homes
      - Productivity loss etc.
      - Average cost over 46 patients surveyed: **~$57,070**
  - Average cost for hospitalised or inpatient patients in US from 1999 through 2012

$778 million / 37,088 patients =
$20,977 per patient

There is an estimated 54.20 patients
Total cost =
54.2 * $20,977 * **0.893** (Sensitivity) = **$1,015,410**

Assuming 1Km radius to be sprayed around our traps:

706 traps* $1.656 * 1000m * 1000m * 3.1415 *1.656/1000m^2

Total cost to spray = $382,799

Total amount saved = $632,611

Summary of Cost-Benefit Analysis

	Sensitivity	Number of positives	Number of true positives	Traps, WNV = 1	Cost saved for patients	Cost incurred for spray	Total savings
50% Threshold (Default)	**0.772**	44.4 patients	**34.3 patients**	**608**	719,737	**316,225**	403,511
30% Threshold	**0.893**	54.2 patients	**48.4 patients**	**736**	1,015,410	**382,799**	632,611
						Difference	229,100

Conclusion and Recommendations

- Probability of WNV increases exponentially from week 31 to 35, especially when there is a combination of high temperature and precipitation (rainfall). Certain traps have higher chance of contracting WNV (e.g. T900, T003, T028).

- We recommend using Logistic Regression with 30% threshold to predict WNV
  - 89% of WNV+ is being captured our model (based on sensitivity)
  - 15% of our positive predictions are correct
  - More cost-effective for the city
    - Benefits to public health are economically significant when compared to the cost of eradicating mosquitoes
    - About 14 more true positives can be prevented in 2014, preventing medical costs
    - Cost of spray is lower than short and long term medical costs

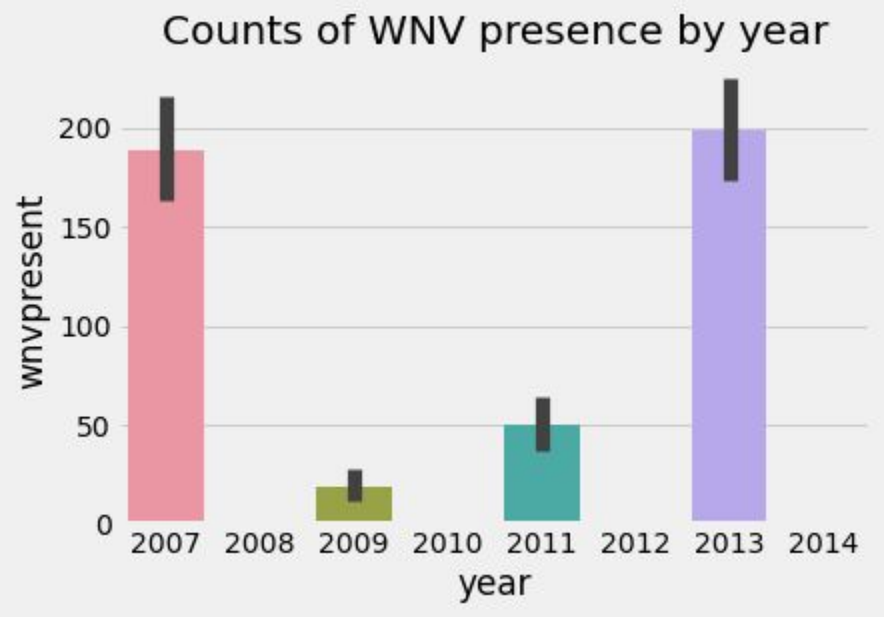
Next Steps

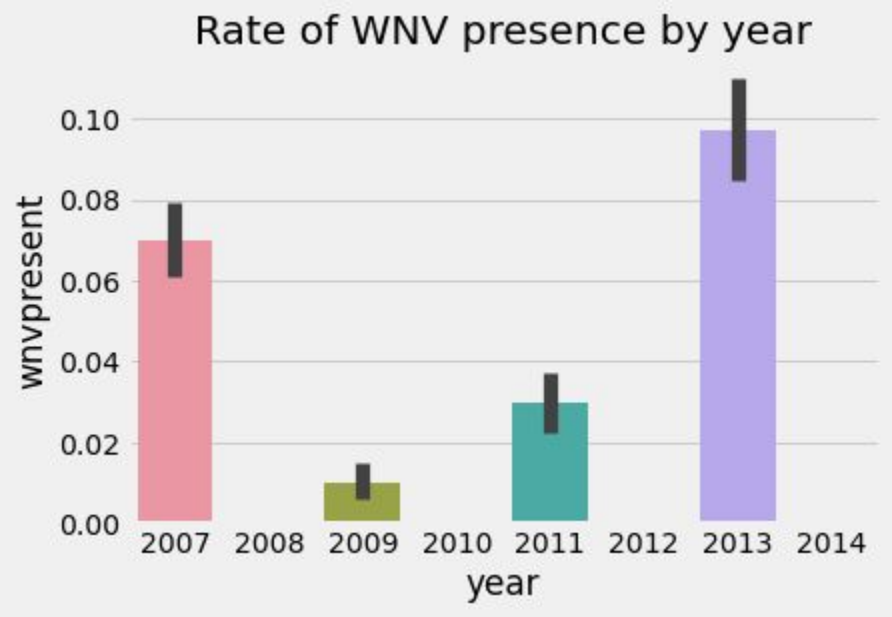
- Further tune parameters to improve accuracy
  - More weather features and their interaction
  - Grouping of traps based on risk level (likelihood of WNV)

Thank you

Annex

EDA: WNV has the highest occurrence in 2007 and 2013.

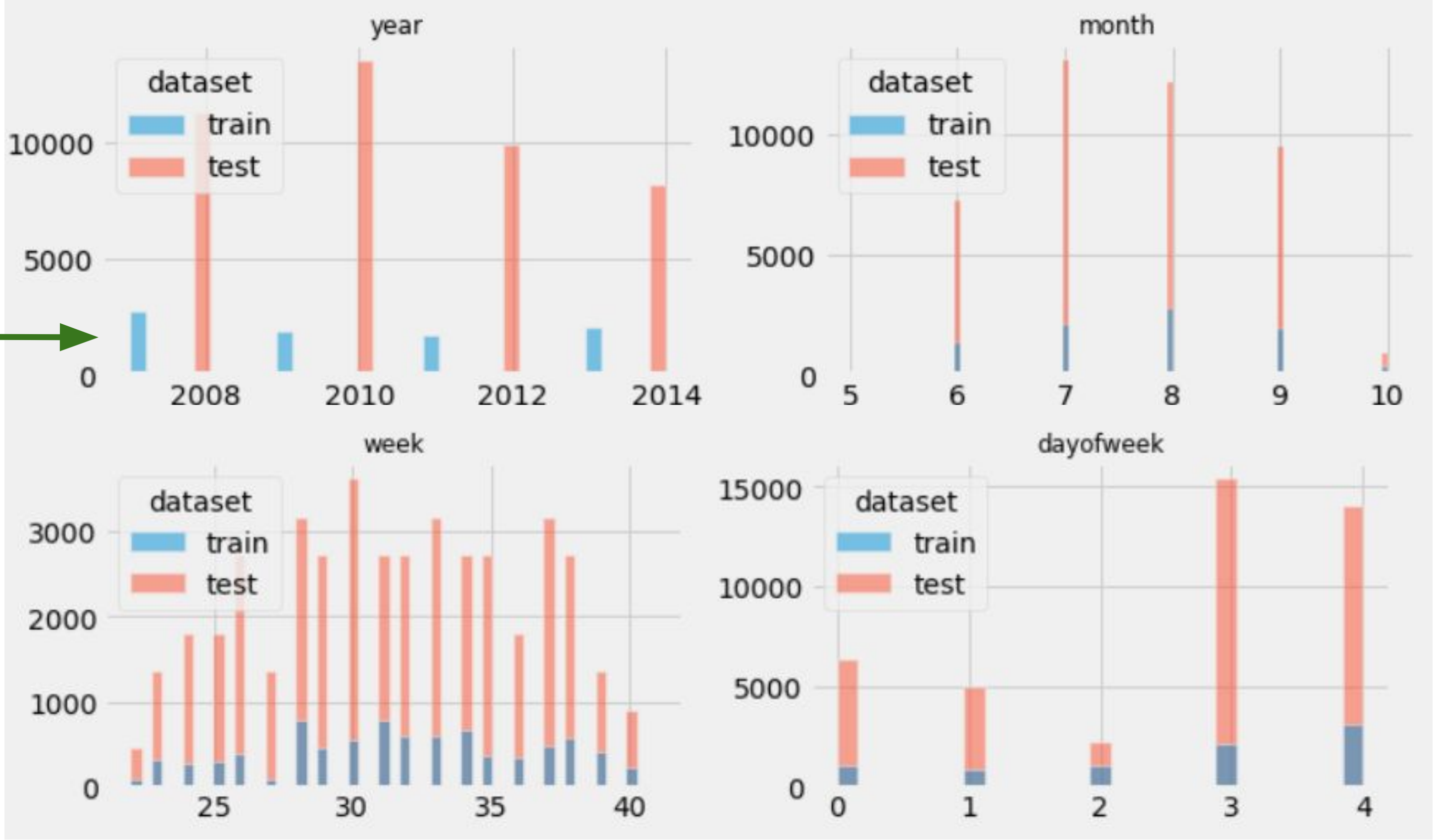




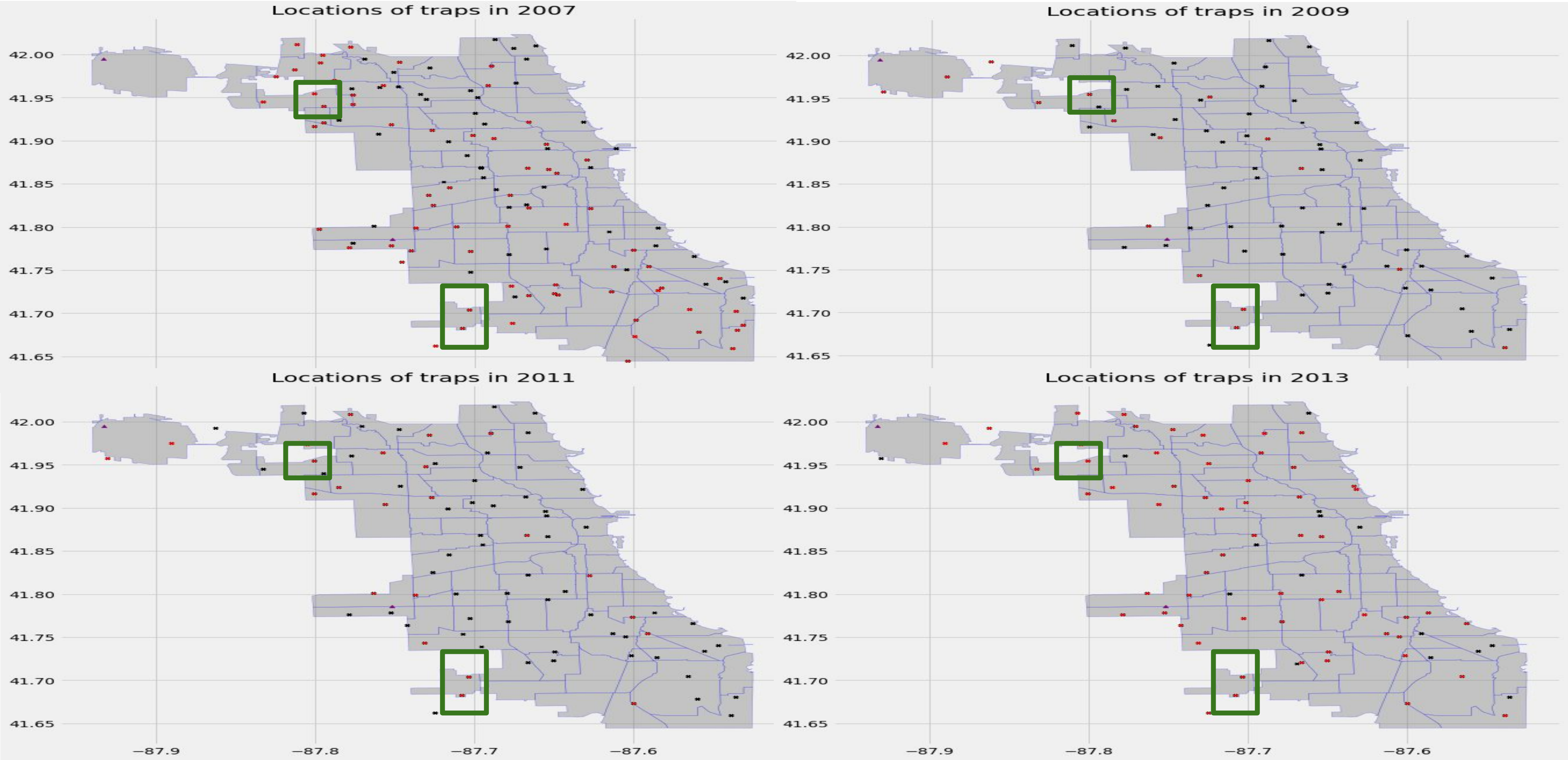
EDA: Training dataset only has data in the odd years, while test dataset only has data in the even years

- This suggests that year is a not useful feature

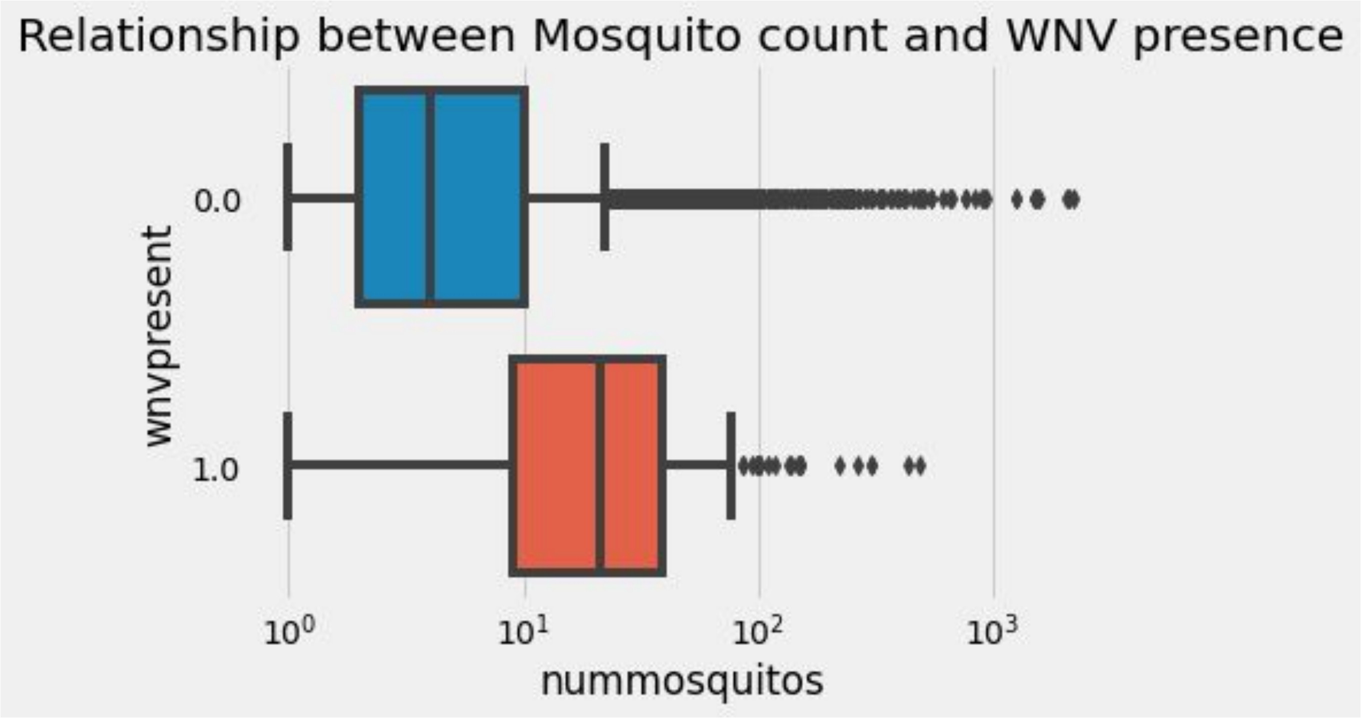
- We cannot use time series e.g. ARIMA to predict WNV



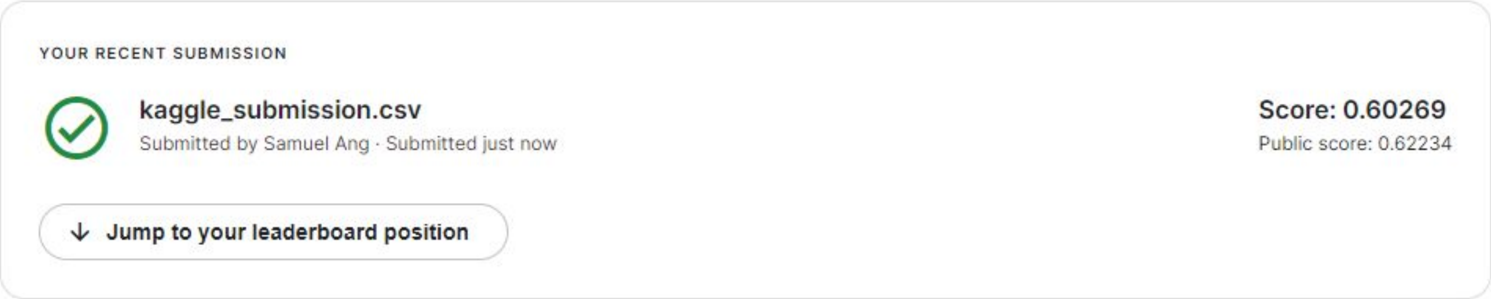
EDA: Some trap spots are consistent high-risk areas


Locations of traps in 2007
Locations of traps in 2009
Locations of traps in 2011
Locations of traps in 2013
42.00
41.95
41.90
41.85
41.80
41.75
41.70
41.65
−87.9
−87.8
−87.7
−87.6

EDA: We can look at the effect of spray on number of mosquitoes as it is correlated to WNV



Kaggle Submission Score



YOUR RECENT SUBMISSION

✓ kaggle_submission.csv

Submitted by Samuel Ang · Submitted just now

Score: 0.60269

Public score: 0.62234