Predicting West nile virus

West nile virus

Commonly spread to humans via infected mosquitoes

20%

Around 20% of people who become infected with the virus develop symptoms ranging from a persistent fever, to serious neurological illnesses that can result in death.

BACKGROUND

2002

The first human cases of West Nile virus were reported in Chicago.

2004

The City of Chicago and the Chicago Department of Public Health (CDPH) had established a comprehensive surveillance and control program that is still in effect today.

Every week from late spring through the fall, mosquitos in traps across the city are tested for the virus. The results of these tests influence when and where the city will spray airborne pesticides to control adult mosquito populations.

Problem statement

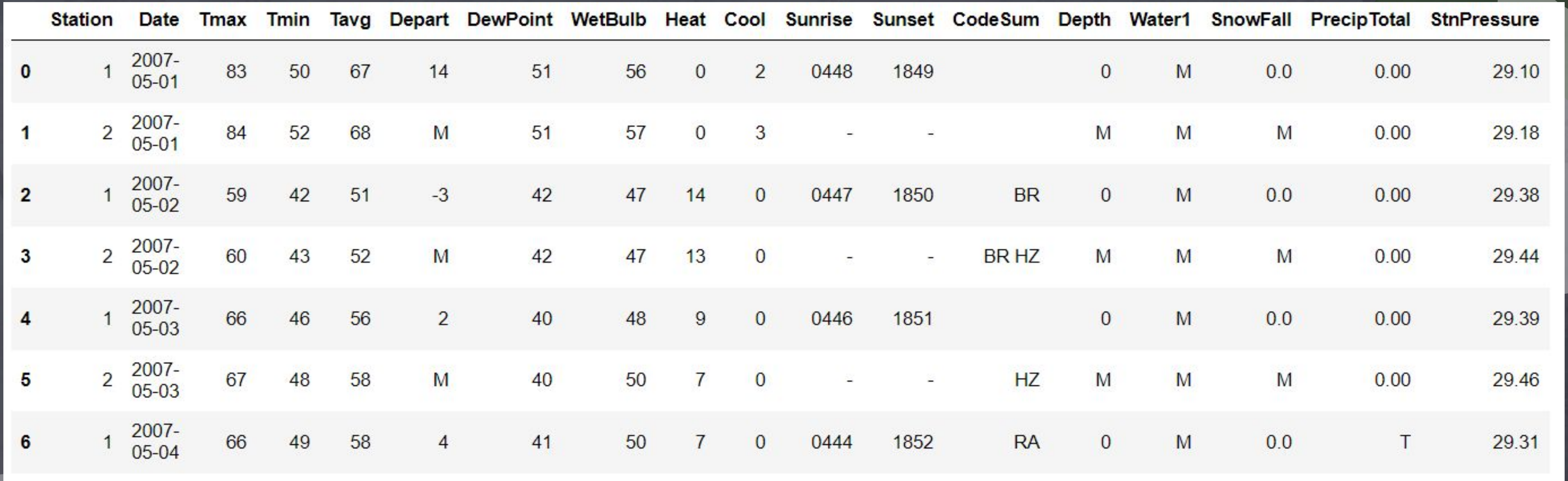
Data

- Weather
- Location
- Testing
- Spraying

PREDICT

- When will mosquitoes test positive for the virus?
- Where will mosquitoes test positive for the virus?

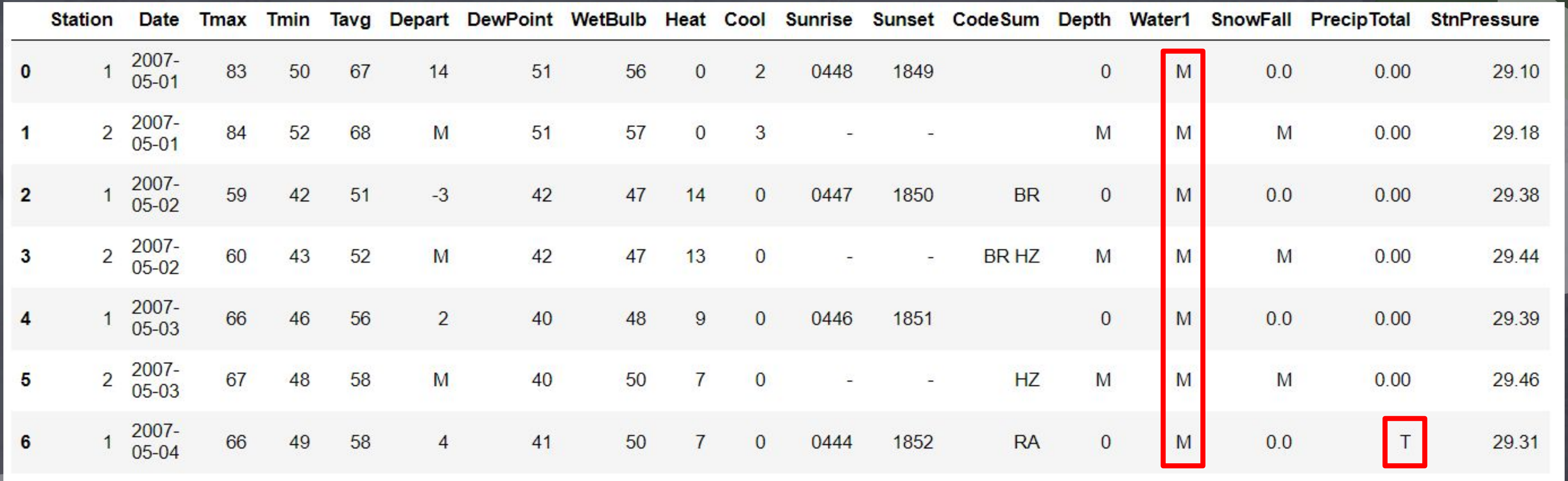
Weather Data overview



	Station	Date	Tmax	Tmin	Tavg	Depart	DewPoint	WetBulb	Heat	Cool	Sunrise	Sunset	CodeSum	Depth	Water1	SnowFall	PrecipTotal	StnPressure
0	1	2007-05-01	83	50	67	14	51	56	0	2	0448	1849		0	M	0.0	0.00	29.10
1	2	2007-05-01	84	52	68	M	51	57	0	3	-	-		M	M	M	0.00	29.18
2	1	2007-05-02	59	42	51	-3	42	47	14	0	0447	1850	BR	0	M	0.0	0.00	29.38
3	2	2007-05-02	60	43	52	M	42	47	13	0	-	-	BR HZ	M	M	M	0.00	29.44
4	1	2007-05-03	66	46	56	2	40	48	9	0	0446	1851		0	M	0.0	0.00	29.39
5	2	2007-05-03	67	48	58	M	40	50	7	0	-	-	HZ	M	M	M	0.00	29.46
6	1	2007-05-04	66	49	58	4	41	50	7	0	0444	1852	RA	0	M	0.0	T	29.31

Weather data collected from 2 stations on the same day. Goes in the order Day 1 - Station 1, Day 1 - Station 2, Day 2 - Station 1, Day 2 - Station 2 etc.

Weather Data overview

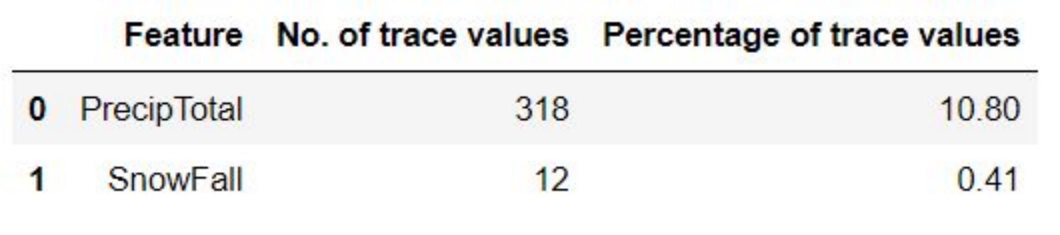


	Station	Date	Tmax	Tmin	Tavg	Depart	DewPoint	WetBulb	Heat	Cool	Sunrise	Sunset	CodeSum	Depth	Water1	SnowFall	PrecipTotal	StnPressure
0	1	2007-05-01	83	50	67	14	51	56	0	2	0448	1849		0	M	0.0	0.00	29.10
1	2	2007-05-01	84	52	68	M	51	57	0	3	-	-		M	M	M	0.00	29.18
2	1	2007-05-02	59	42	51	-3	42	47	14	0	0447	1850	BR	0	M	0.0	0.00	29.38
3	2	2007-05-02	60	43	52	M	42	47	13	0	-	-	BR HZ	M	M	M	0.00	29.44
4	1	2007-05-03	66	46	56	2	40	48	9	0	0446	1851		0	M	0.0	0.00	29.39
5	2	2007-05-03	67	48	58	M	40	50	7	0	-	-	HZ	M	M	M	0.00	29.46
6	1	2007-05-04	66	49	58	4	41	50	7	0	0444	1852	RA	0	M	0.0	T	29.31

M = Missing

T = Trace amounts

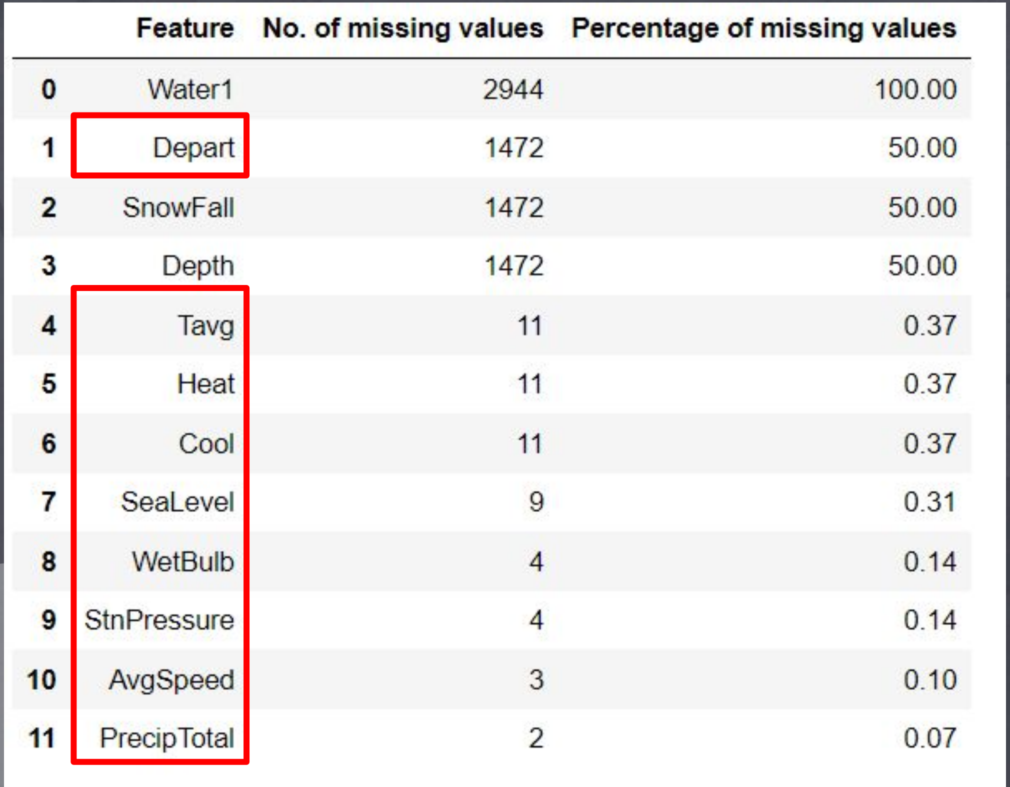
Trace amounts values



	Feature	No. of trace values	Percentage of trace values
0	PrecipTotal	318	10.80
1	SnowFall	12	0.41

Replace all Trace amounts values with the numerical value 0.01 to represent a non-zero small amount

Missing values

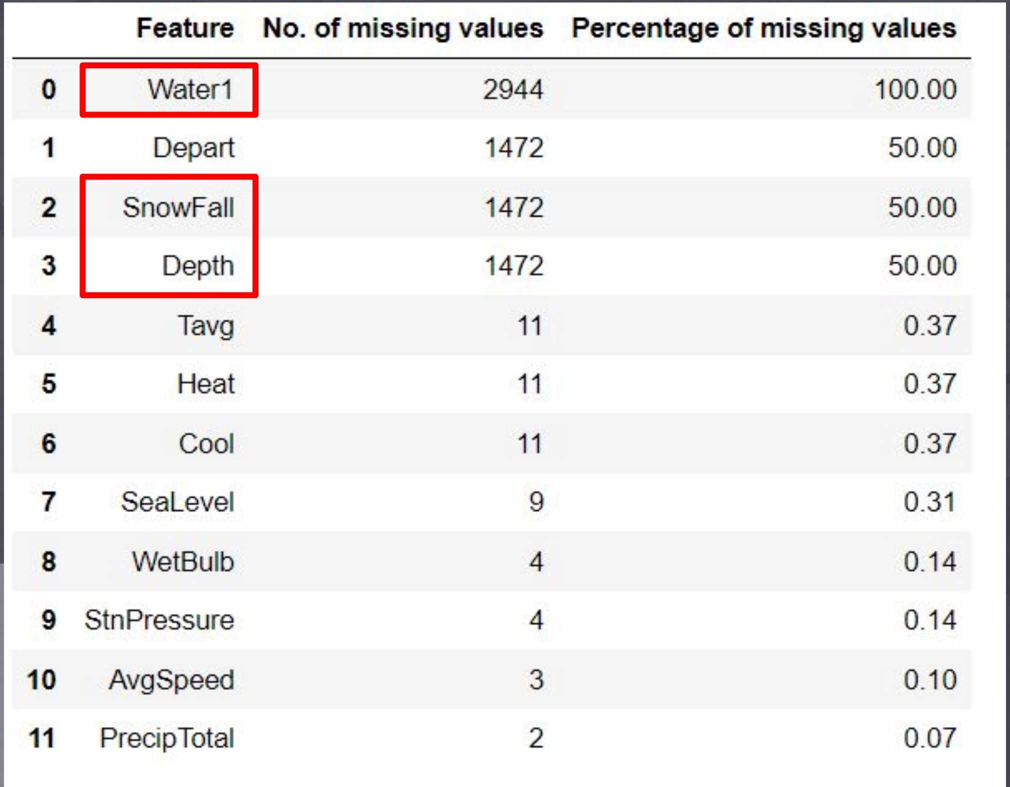


	Feature	No. of missing values	Percentage of missing values
0	Water1	2944	100.00
1	Depart	1472	50.00
2	SnowFall	1472	50.00
3	Depth	1472	50.00
4	Tavg	11	0.37
5	Heat	11	0.37
6	Cool	11	0.37
7	SeaLevel	9	0.31
8	WetBulb	4	0.14
9	StnPressure	4	0.14
10	AvgSpeed	3	0.10
11	PrecipTotal	2	0.07

For the selected Features, for the same day, either Station 1 or Station 2 values were missing.

Impute by assuming Stations 1 and 2 have the same value or same reference value on the same day.

Missing values



	Feature	No. of missing values	Percentage of missing values
0	Water1	2944	100.00
1	Depart	1472	50.00
2	SnowFall	1472	50.00
3	Depth	1472	50.00
4	Tavg	11	0.37
5	Heat	11	0.37
6	Cool	11	0.37
7	SeaLevel	9	0.31
8	WetBulb	4	0.14
9	StnPressure	4	0.14
10	AvgSpeed	3	0.10
11	PrecipTotal	2	0.07

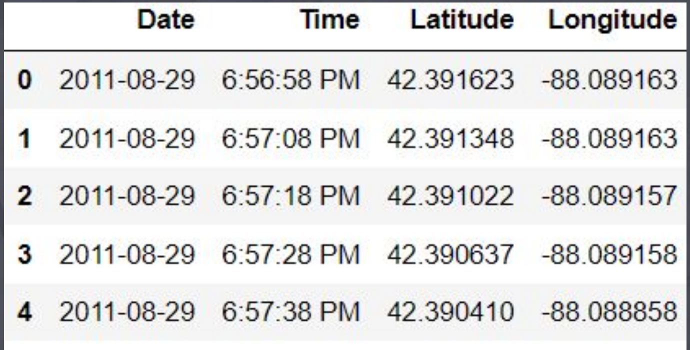
Water1 is removed as 100% of the values are missing.

Since SnowFall is a type of precipitation, it should be taken into account in PrecipTotal.

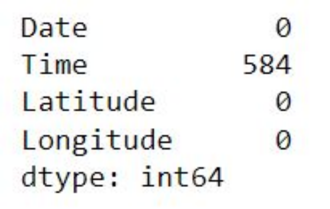
Also, there are no rows where both SnowFall and PrecipTotal are missing, so it should be safe to remove SnowFall.

Additionally, since Depth is a measure of the amount of SnowFall, and in our case has as many missing values as Snowfall, Depth will be removed as well.

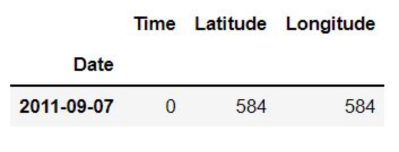
SPRAY DATA OVERVIEW



	Date	Time	Latitude	Longitude
0	2011-08-29	6:56:58 PM	42.391623	-88.089163
1	2011-08-29	6:57:08 PM	42.391348	-88.089163
2	2011-08-29	6:57:18 PM	42.391022	-88.089157
3	2011-08-29	6:57:28 PM	42.390637	-88.089158
4	2011-08-29	6:57:38 PM	42.390410	-88.088858



```
Date            0
Time          584
Latitude        0
Longitude       0
dtype: int64
```

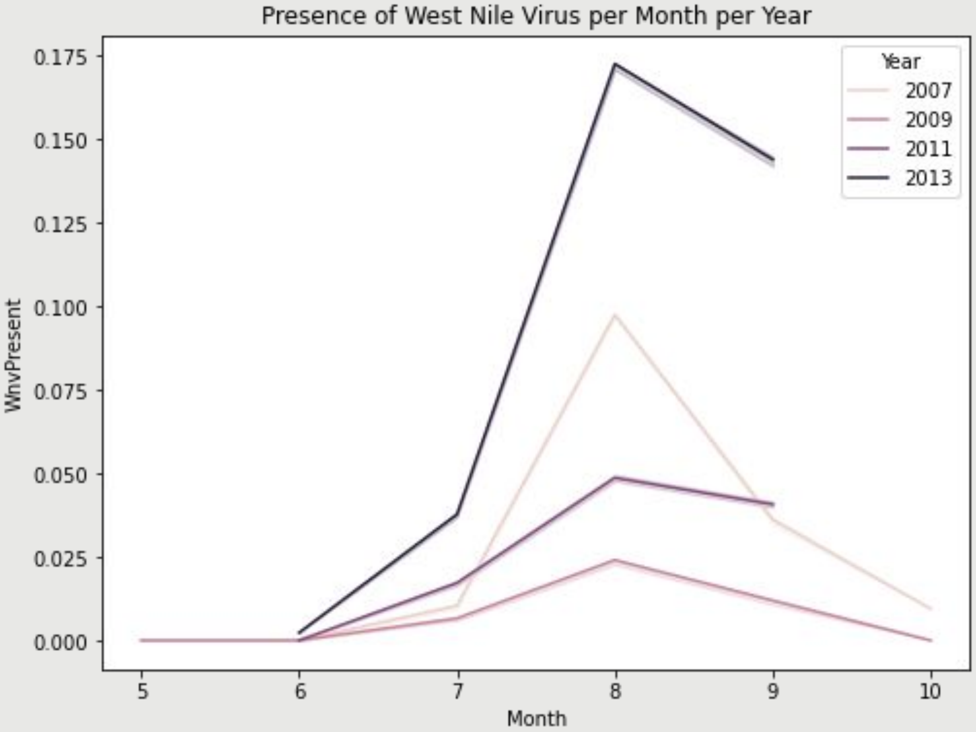


Date	Time	Latitude	Longitude
2011-09-07	0	584	584

Spray data shows the time and location of sprays carried out.

584 time values are missing and they all come from 2011-09-07.

As 584 is only around 4% of the total data, null values were removed from the data.

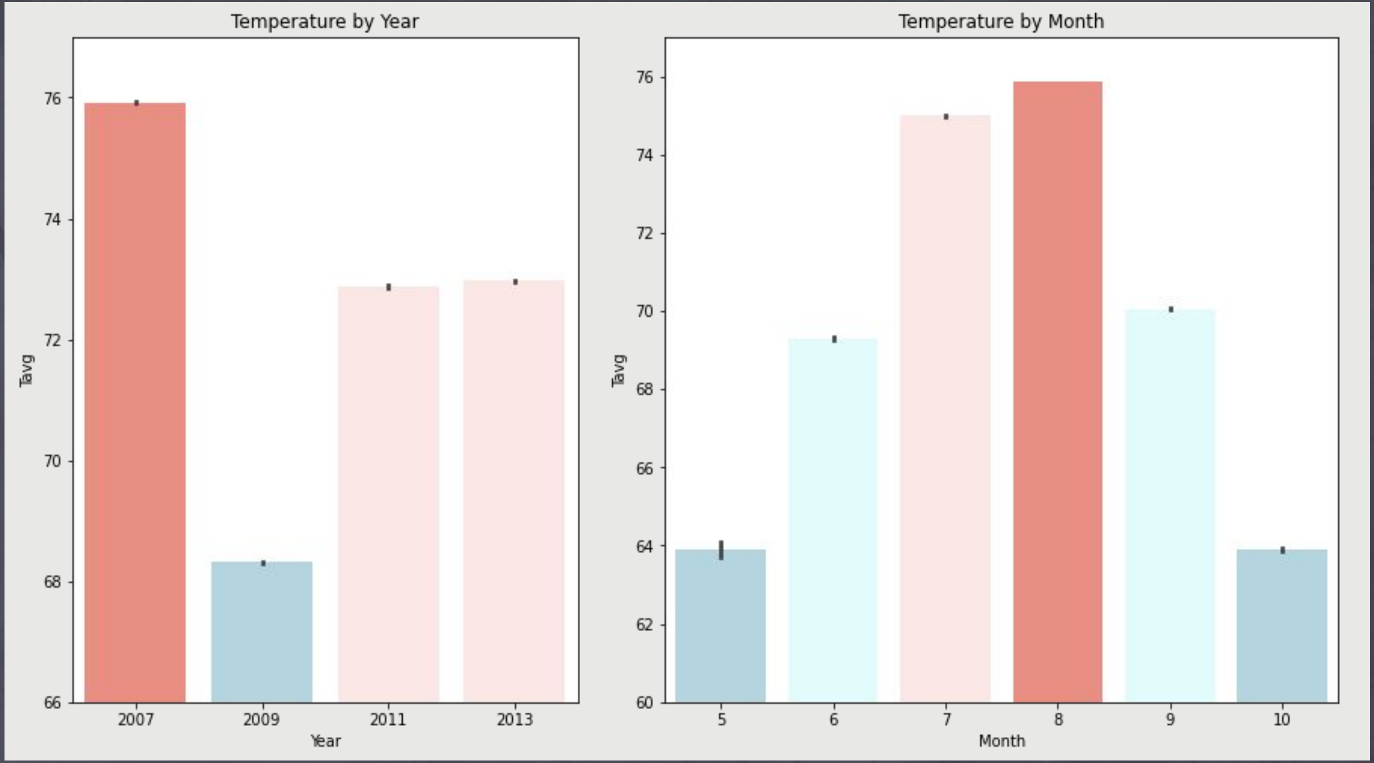


Exploratory Data analysis

WNV Occurrence

- Highest - 2013
- Lowest - 2009
- Yearly spikes in August

Temperature by year/month

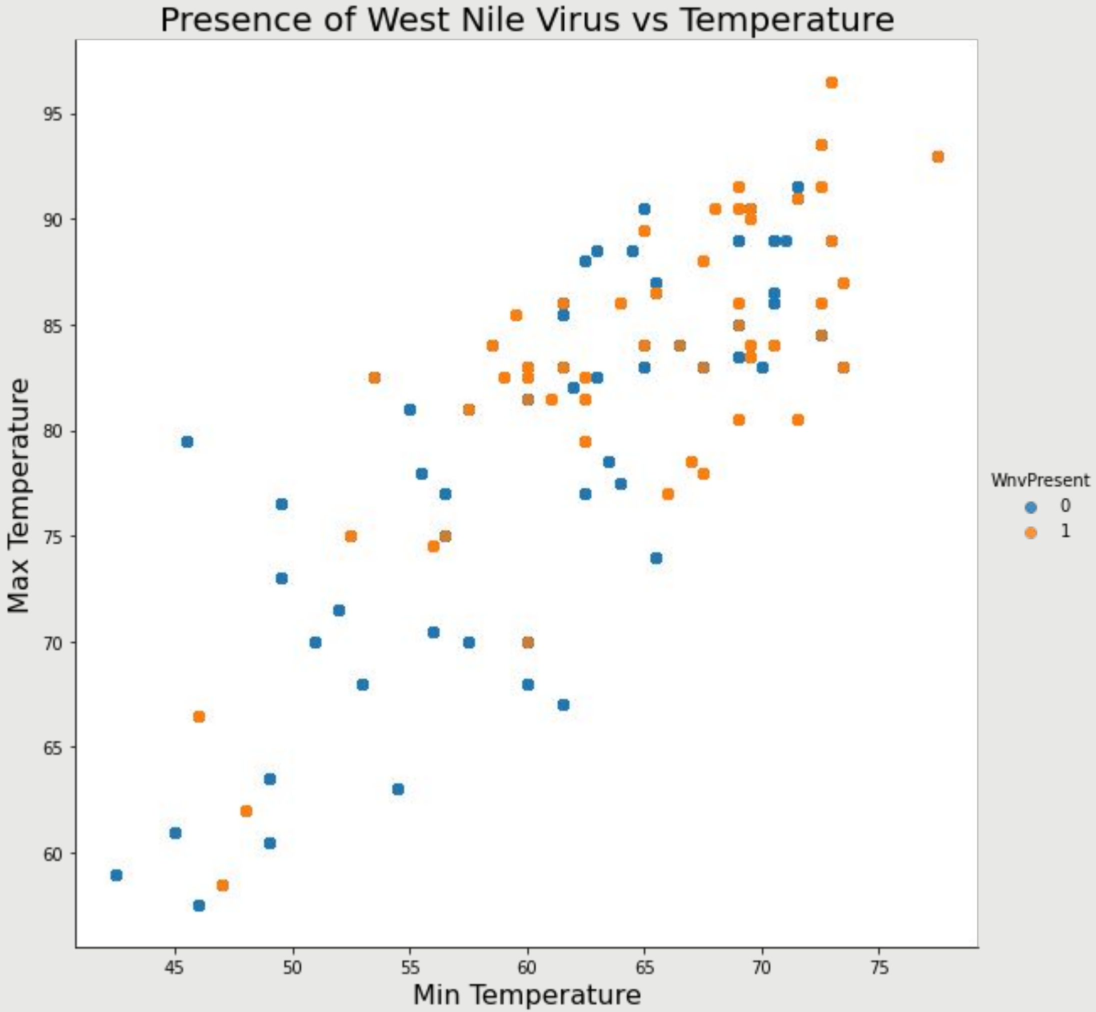


Temperature by **Year**

- Highest - 2007
- Lowest - 2009

Temperature by **Month**

- Highest - August
- Lowest - May/October



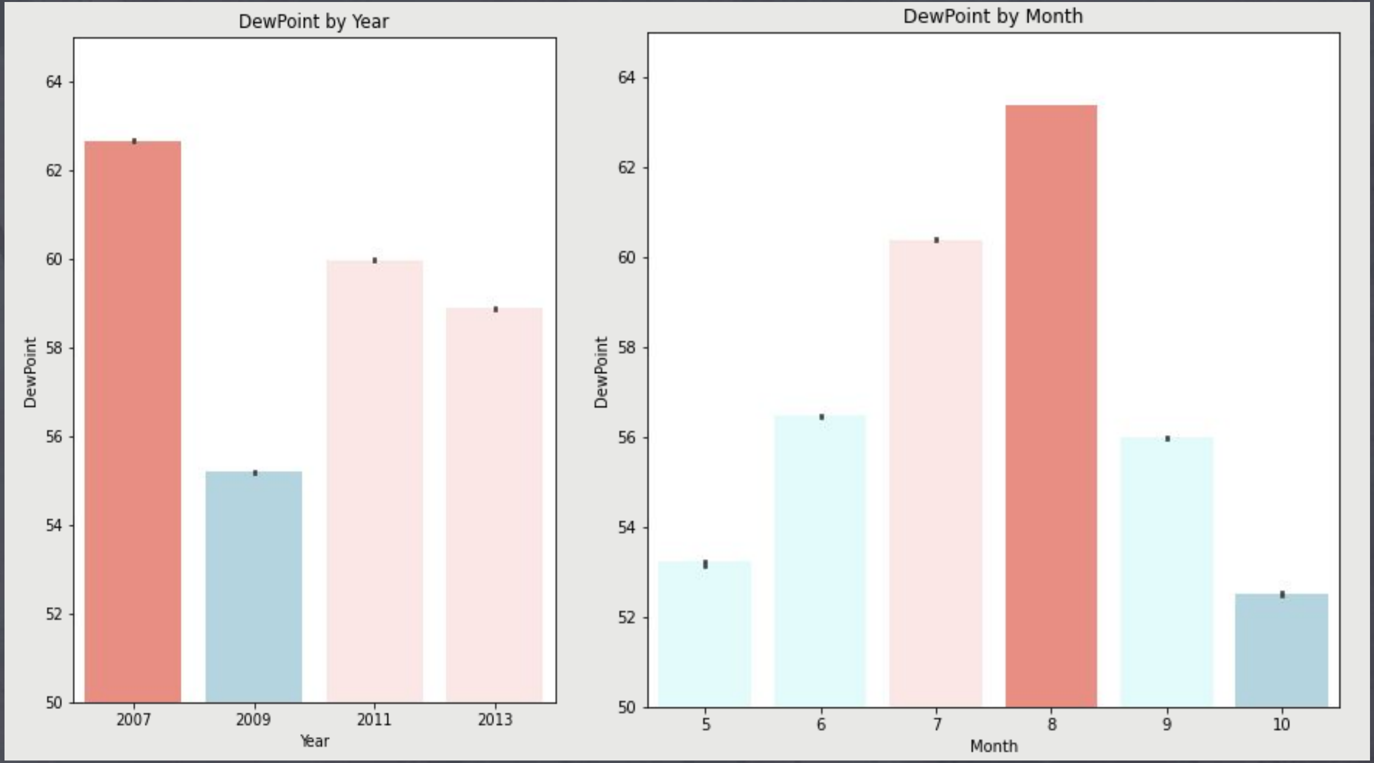
WNV vs Temperature

Increase in temperature

⇒ Increase in WNV cases

⇒ Positive correlation between temperature & WNV

Dewpoint by year/month

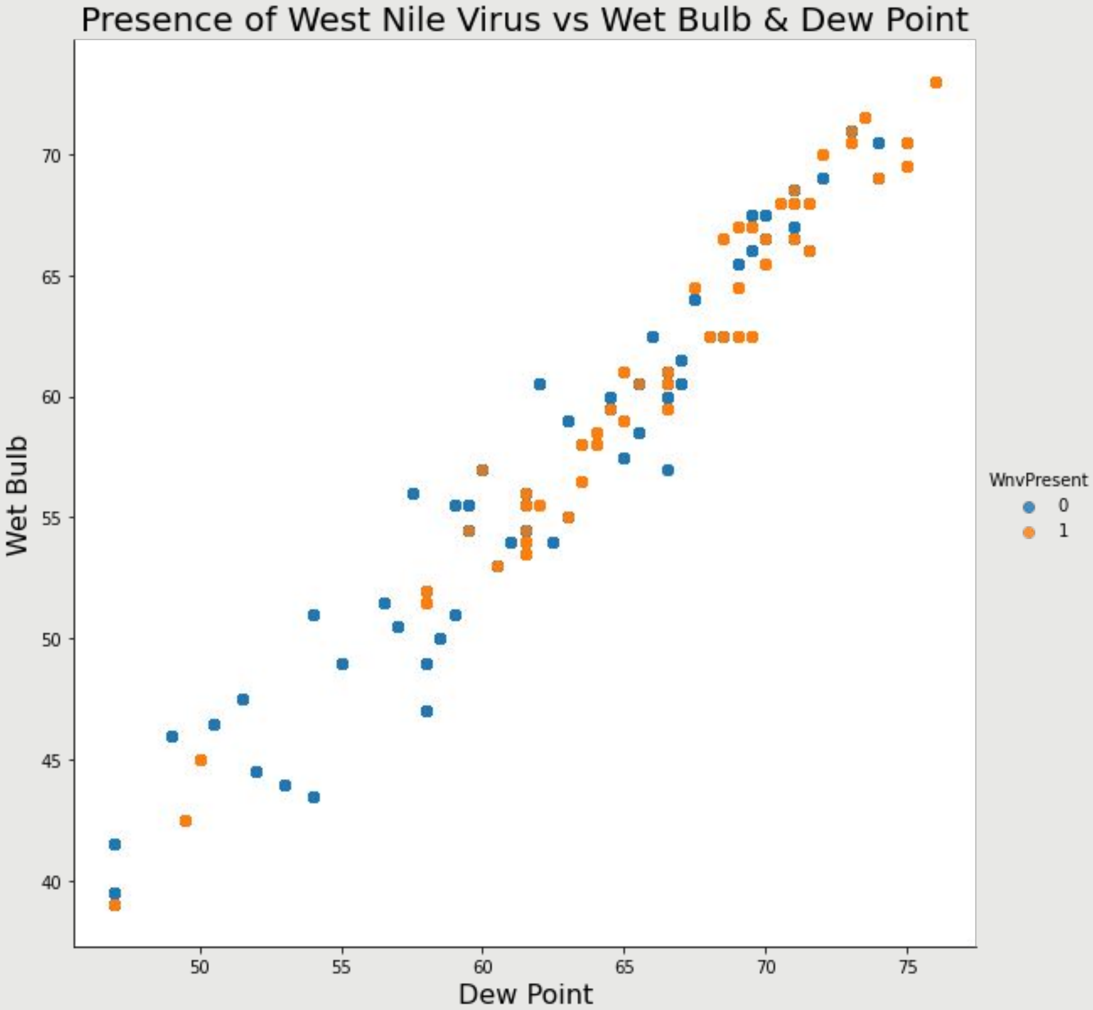


DewPoint by **Year**

- Highest - 2007
- Lowest - 2009

DewPoint by **Month**

- Highest - August
- Lowest - October

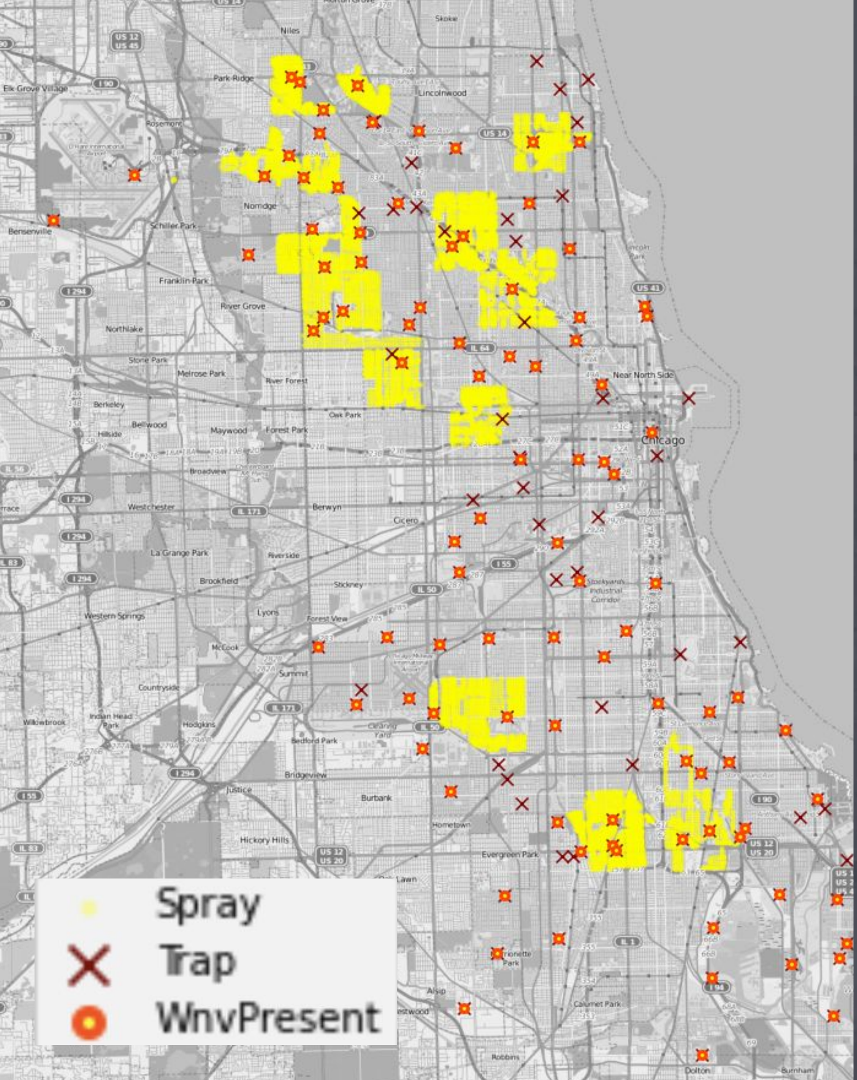


WNV vs Humidity

Increase in humidity

⇒ Increase in WNV cases

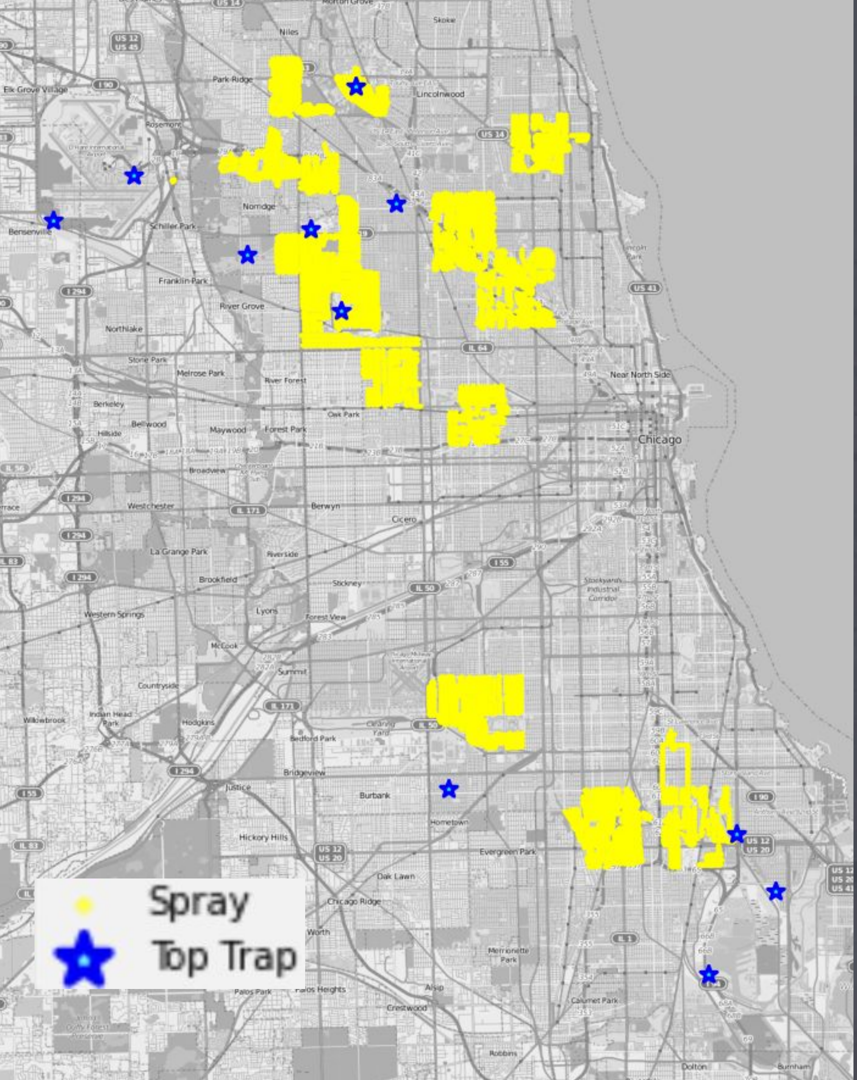
⇒ **Strong** positive correlation between humidity & WNV



Spray Locations

From the scatterplot,

- Most traps captured at least 1 WNV mosquito
- Most of the locations with WNV present were not sprayed

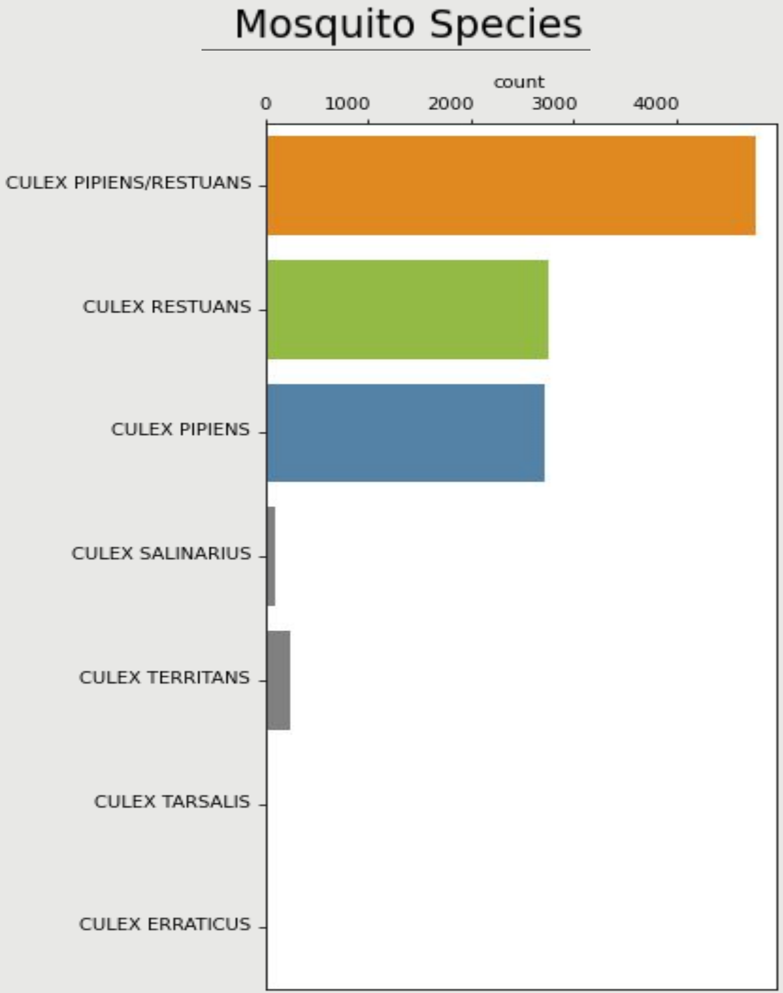


Spray Locations

- ‘Top Trap’: Traps (90 percentile) that caught the most number of WNV

  ⇒ ~40% of the total WNV count
- 3 out of 11 top traps were sprayed

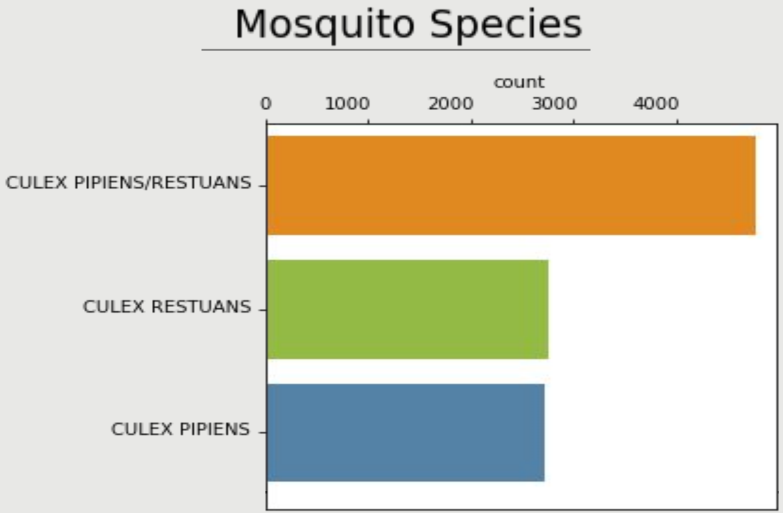
  ⇒ i.e. ~73% of the hotspots were not sprayed
- Spray efforts were not targeted at the right locations



Mosquito Species

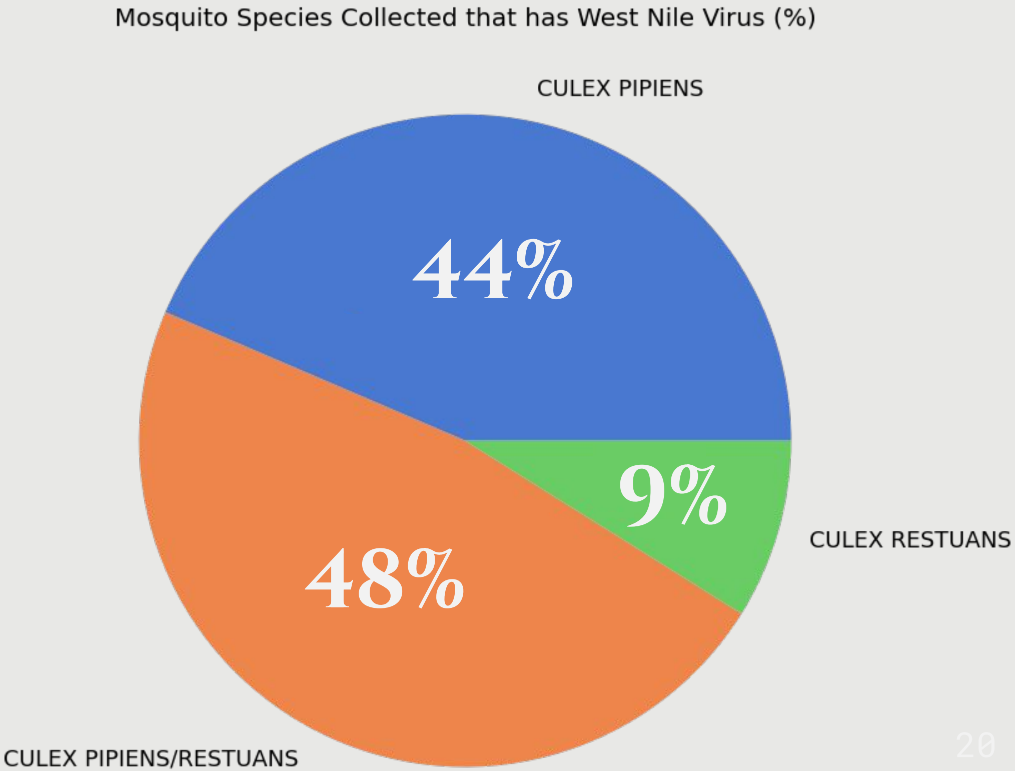
- 6 species in our dataset
- 2 out of the 6 species are carriers of WNV

Mosquito Species (WNV Carriers)

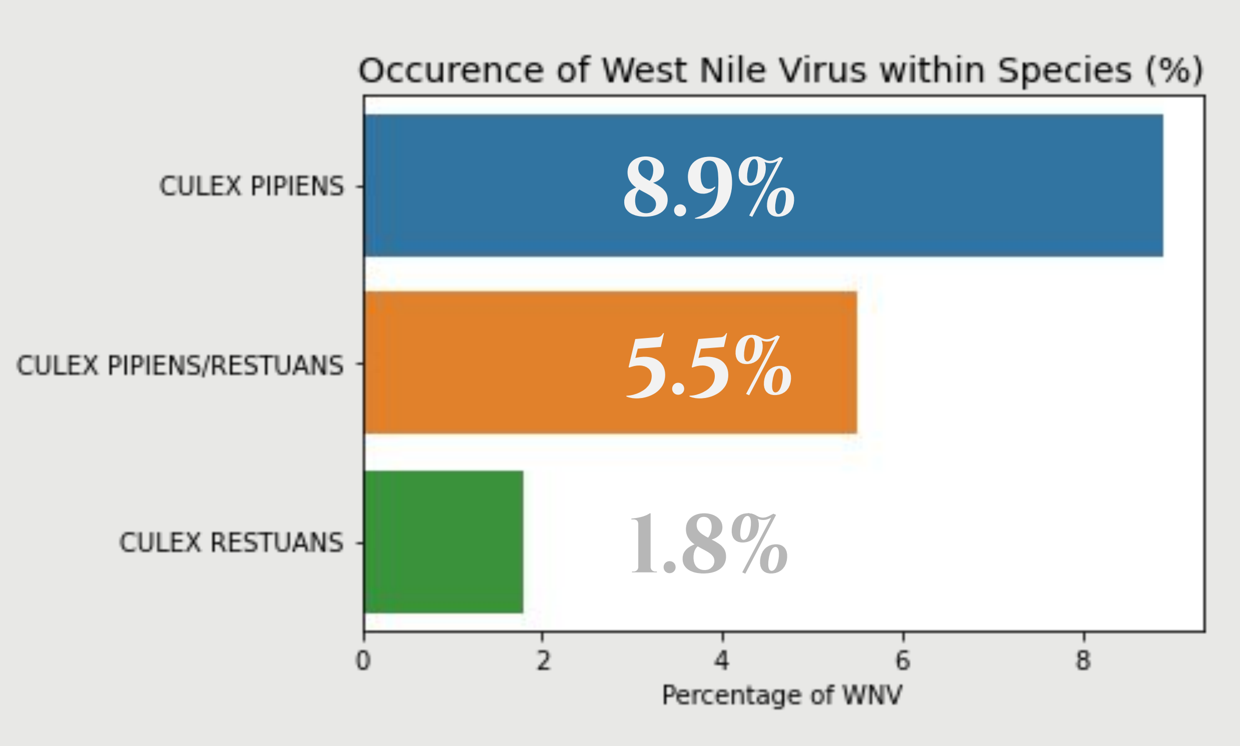


Despite similar species size

- % Culex Pipiens WNV carriers > % Culex Restuans WNV carriers



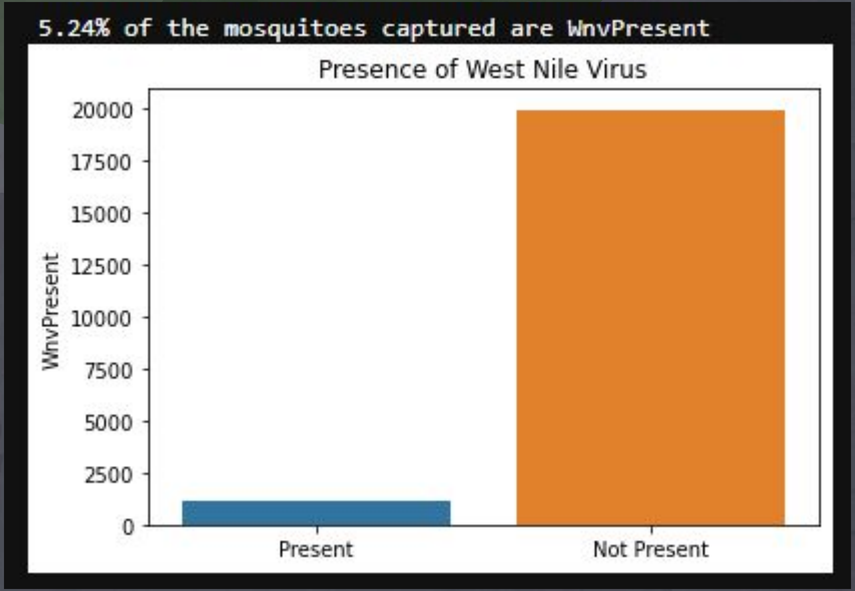
Occurrence of WNV within Mosquito Species



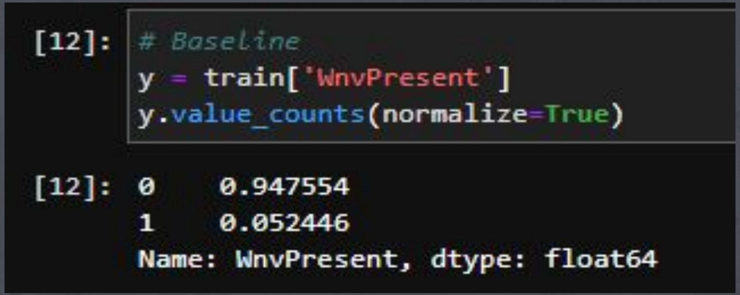
Within each species,

- Culex Pipiens likely to be the major vector of WNV
- Assigned weights proportional to the occurrence rate using ordinal encoding

Baseline model



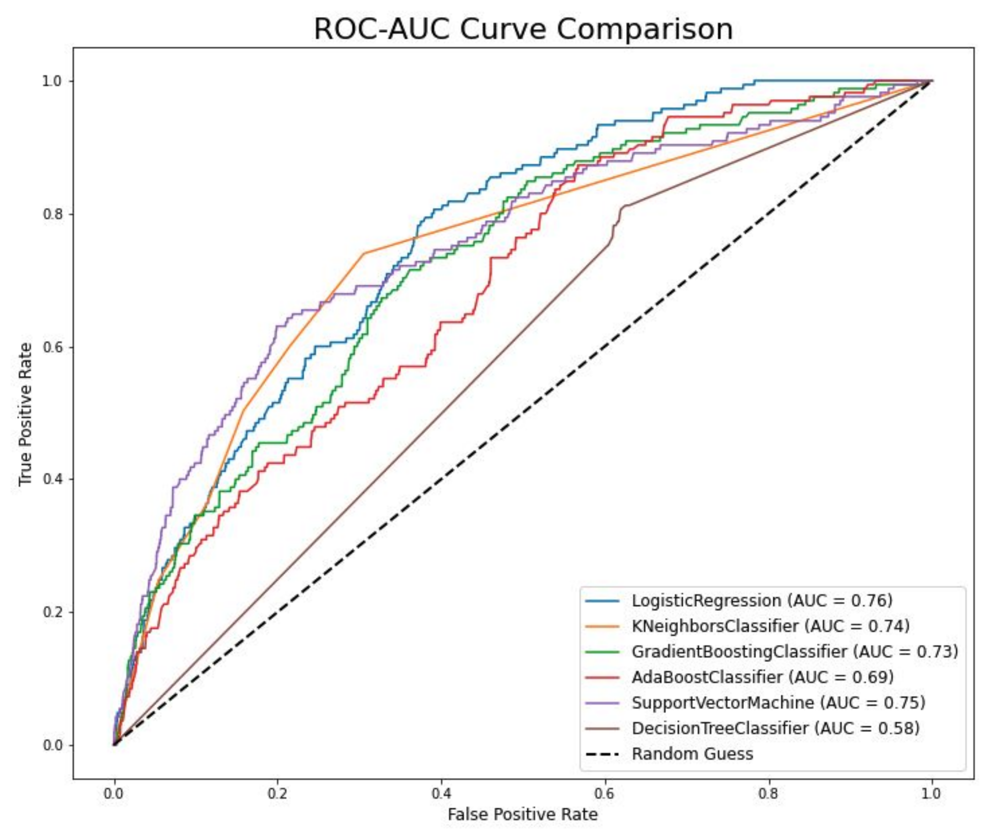
5.24% of the mosquitoes captured are WnvPresent



```
[12]: # Baseline
      y = train['WnvPresent']
      y.value_counts(normalize=True)

[12]: 0    0.947554
      1    0.052446
      Name: WnvPresent, dtype: float64
```

- Imbalanced dataset
- SMOTE is a commonly used oversampling method that attempts to balance class distribution

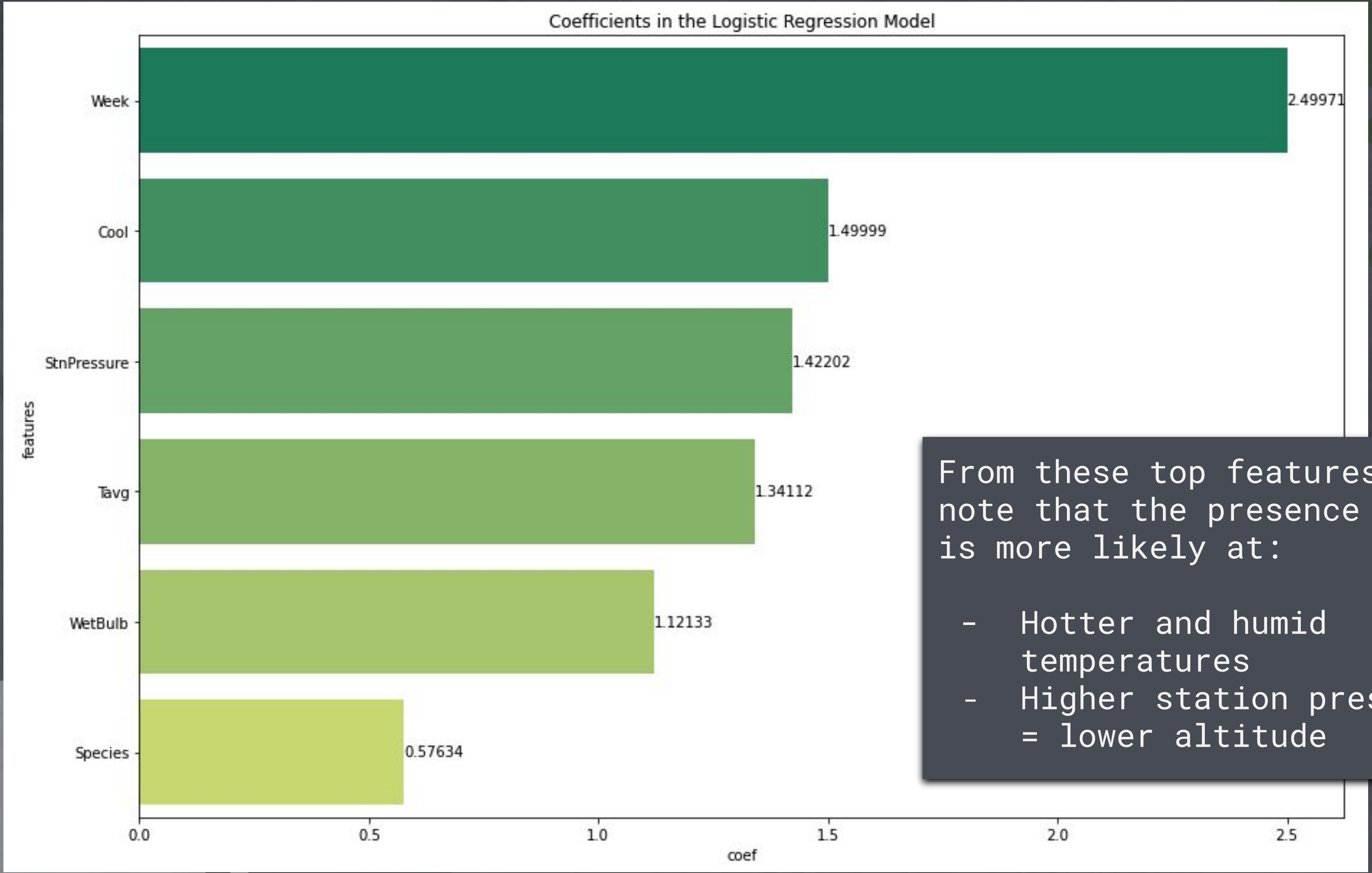


ROC-AUC

- Logistic Regression Model has the highest AUC score

0.68

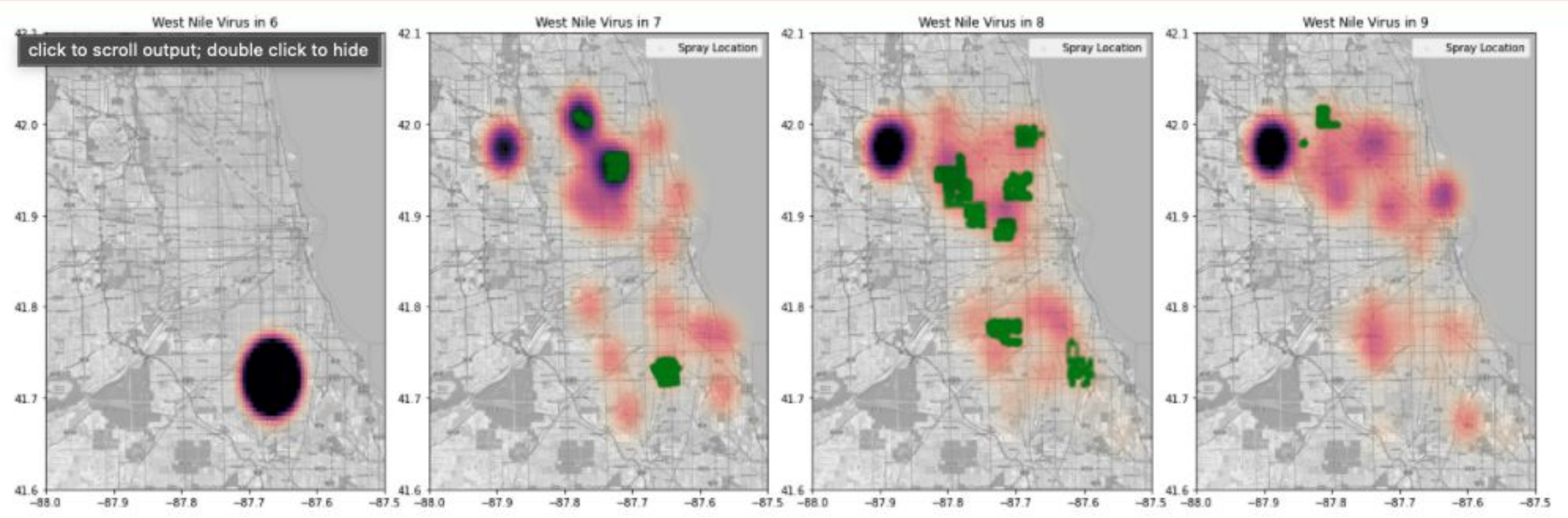
Kaggle Score



From these top features, we can note that the presence of WNV is more likely at:

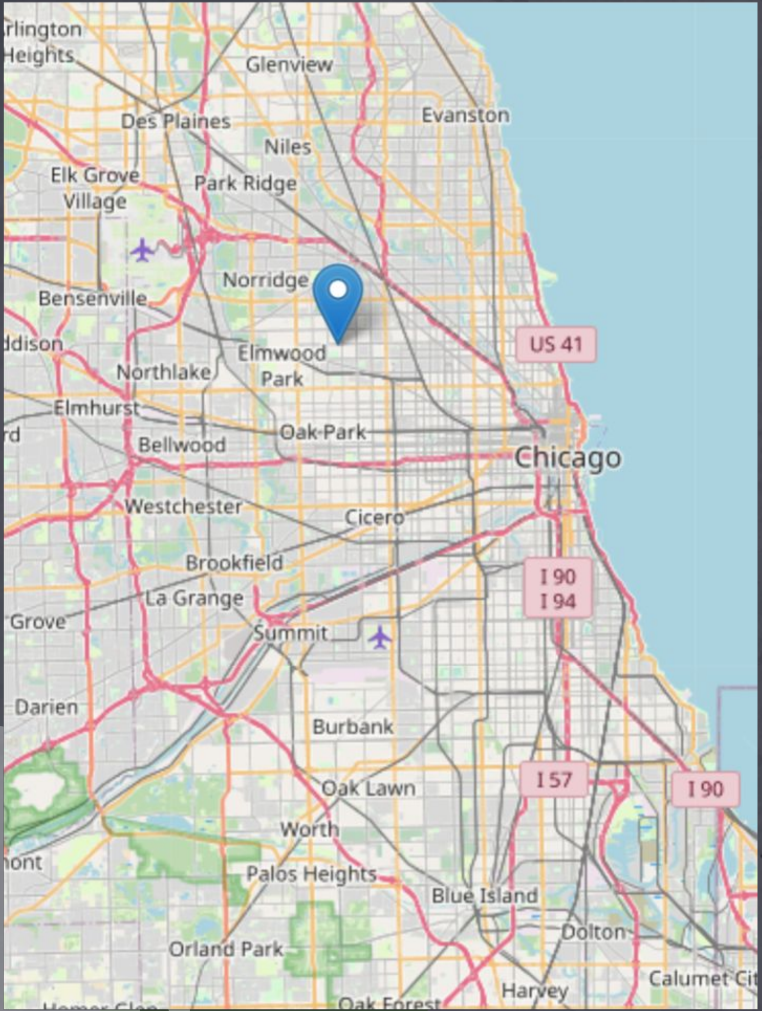
- Hotter and humid temperatures
- Higher station pressure = lower altitude

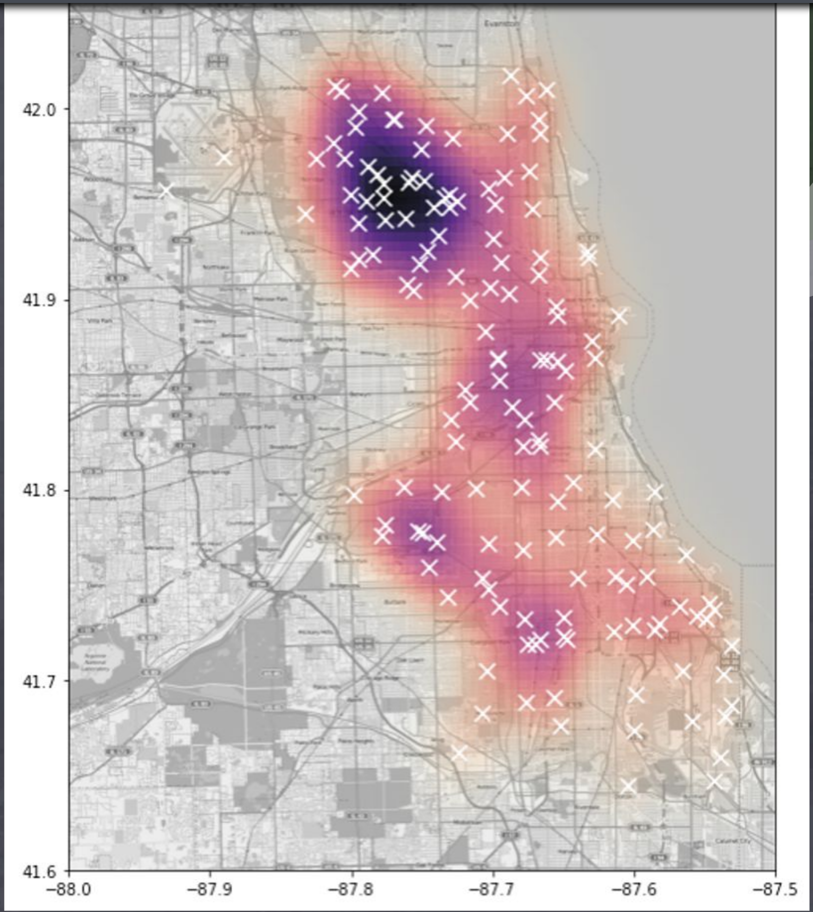
Monthly Sprays in 2013



The sprays are reactive but does decrease the probability densities post-spray.

Predicted Presence of WNV





v.Estimated Cost-Benefit Analysis

2.71 Million

Chicago Pop.

(2019)

149,800 acres

Chicago Area

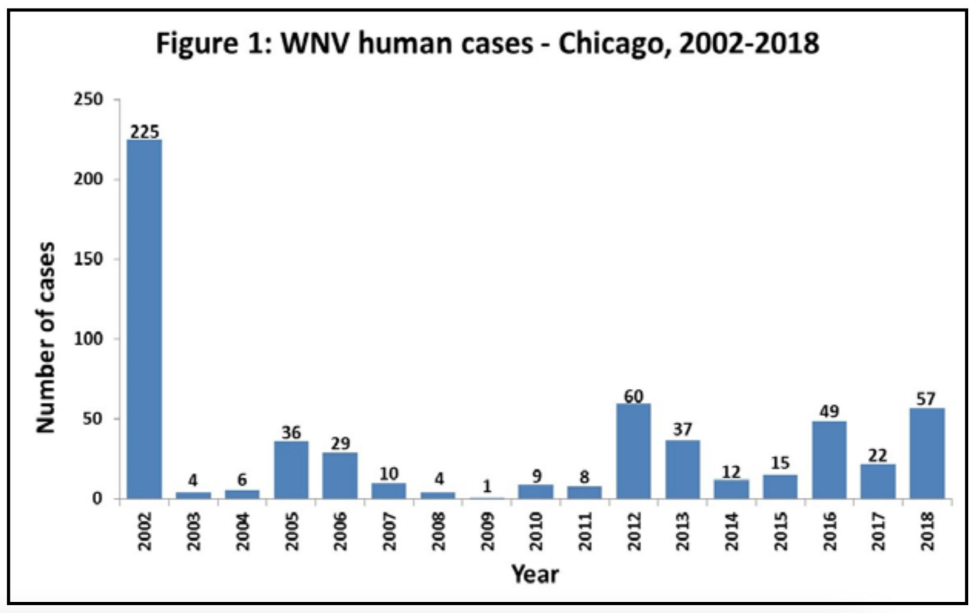


Figure 1: WNV human cases - Chicago, 2002-2018

Recommendations

Model

- Further tuning of available features
- More data on previous infected human clusters, bird clusters, water bodies etc. to supplement model predictions

Vector Control Measures

- House inspections on residences with unruly yards that could be potential breeding grounds
- Promote community support through public education

Spraying

- Early prevention in Northern Chicago
- Focus on green areas and still water bodies
- Utilise weekly surveillance report to supplement spray areas
- Further investigate on airport vector control strategies

Thank you!