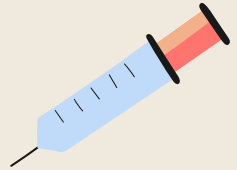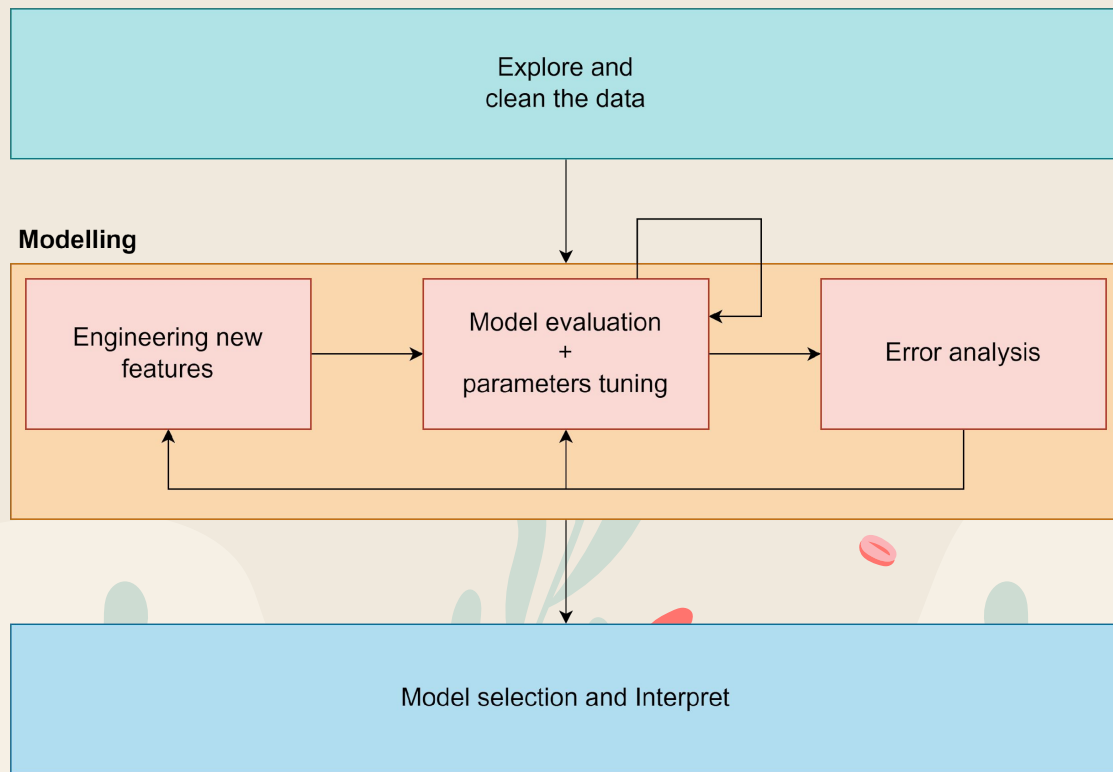# Overview and Problem

Chicago Department of Public Health (CDPH) concerned about West Nile Virus epidemic:

- The potential rate of West Nile virus presence in Chicago

- Where the presence of West Nile virus is observed?
  - We assume that it may originate from one point and then spread to nearby areas

- When we observed the presence of West Nile virus?
  - We aim to analyze past data to identify the week or month with the highest virus prevalence

Process

Explore and
clean the data

Modelling

Engineering new
features

Model evaluation
+
parameters tuning

Error analysis

Model selection and Interpret

Exploratory Data Analysis

We have data about traps that collect mosquitoes, a weather dataset, and a spray schedule.

Understanding that WNV comes from mosquitoes, the purpose of spraying is to reduce mosquito populations in the treated area.

Additionally, the exploration indicates that weather also affects the spread of mosquitoes in this case.

Overview of Datasets

```
RangeIndex: 10506 entries, 0 to 10505
Data columns (total 12 columns):
 #   Column                  Non-Null Count    Dtype
---  ------                  --------------    -----
 0   Date                    10506 non-null    datetime64[ns]
 1   Address                 10506 non-null    object
 2   Species                 10506 non-null    object
 3   Block                   10506 non-null    int64
 4   Street                  10506 non-null    object
 5   Trap                    10506 non-null    object
 6   AddressNumberAndStreet  10506 non-null    object
 7   Latitude                10506 non-null    float64
 8   Longitude               10506 non-null    float64
 9   AddressAccuracy         10506 non-null    int64
 10  NumMosquitos            10506 non-null    int64
 11  WnvPresent              10506 non-null    int64
```
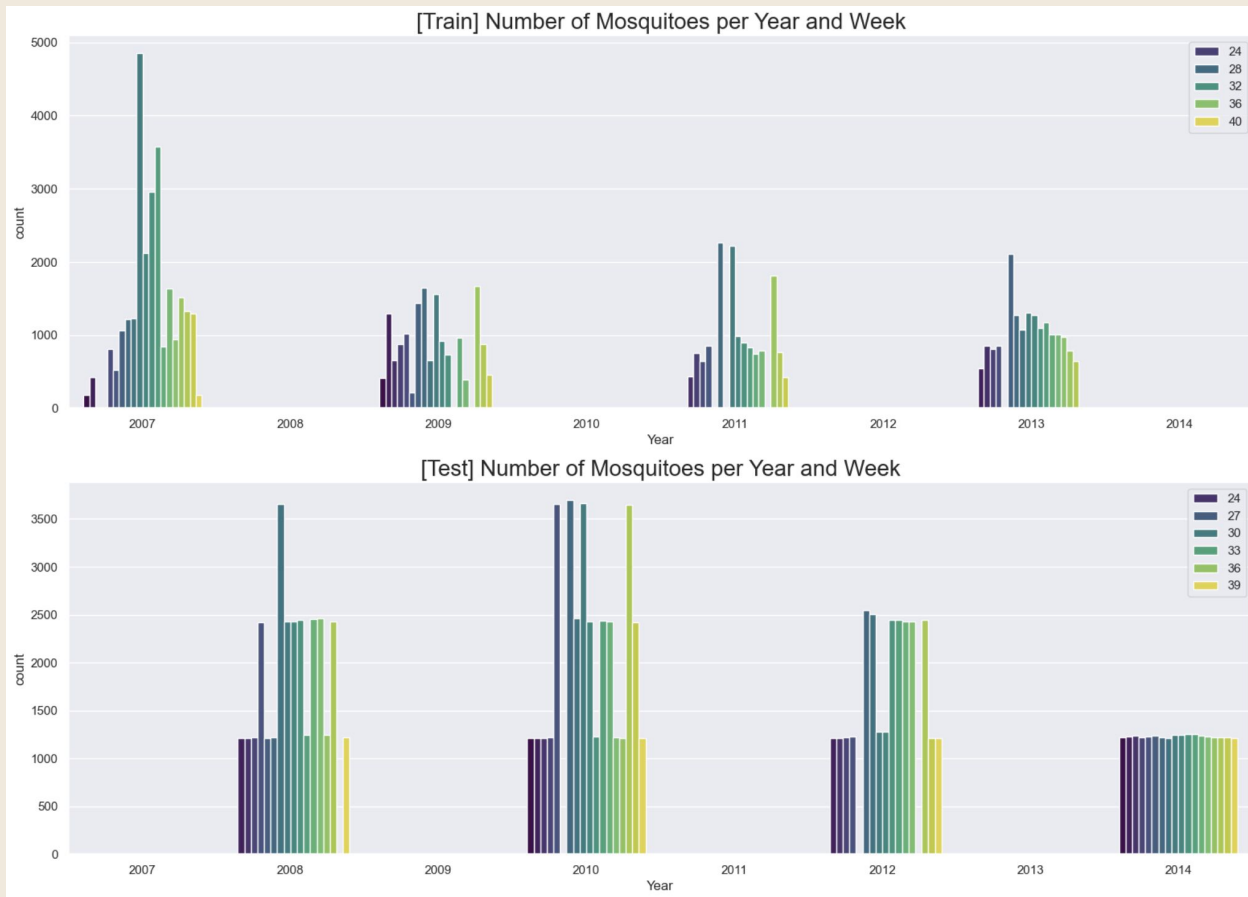
```
RangeIndex: 116293 entries, 0 to 116292
Data columns (total 11 columns):
 #   Column                  Non-Null Count     Dtype
---  ------                  --------------     -----
 0   Id                      116293 non-null    int64
 1   Date                    116293 non-null    object
 2   Address                 116293 non-null    object
 3   Species                 116293 non-null    object
 4   Block                   116293 non-null    int64
 5   Street                  116293 non-null    object
 6   Trap                    116293 non-null    object
 7   AddressNumberAndStreet  116293 non-null    object
 8   Latitude                116293 non-null    float64
 9   Longitude               116293 non-null    float64
 10  AddressAccuracy         116293 non-null    int64
```
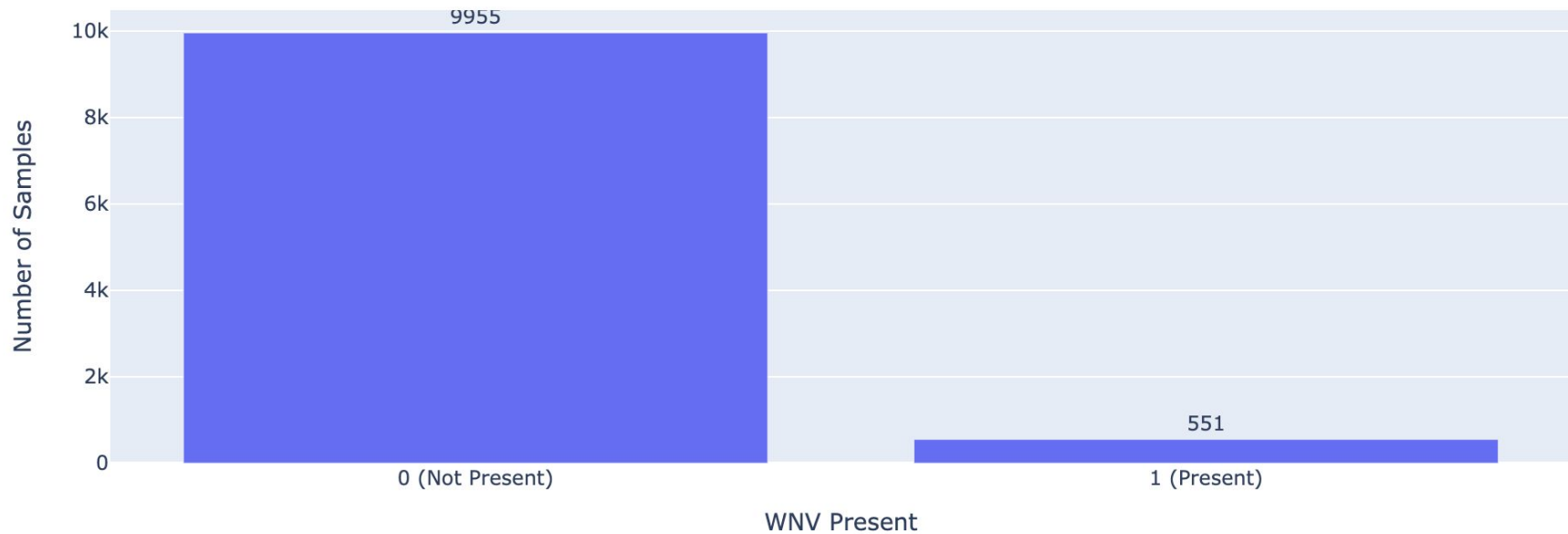
Remark: on train and test data amount on 10506 and 116293 rows
And `NumMosquitos` is missing on test dataset

Overview of Datasets

Overview of Datasets

Class Distribution in Train Data

Overview of Datasets

Weather collect every for 8 years with 2 station records

Remark: weather not collect for every single day

```
RangeIndex: 2944 entries, 0 to 2943
Data columns (total 22 columns):
 #   Column        Non-Null Count   Dtype
---  ------        --------------   -----
 0   Station       2944 non-null    int64
 1   Date          2944 non-null    object
 2   Tmax          2944 non-null    int64
 3   Tmin          2944 non-null    int64
 4   Tavg          2944 non-null    object
 5   Depart        2944 non-null    object
 6   DewPoint      2944 non-null    int64
 7   WetBulb       2944 non-null    object
 8   Heat          2944 non-null    object
 9   Cool          2944 non-null    object
 10  Sunrise       2944 non-null    object
 11  Sunset        2944 non-null    object
 12  CodeSum       2944 non-null    object
 13  Depth         2944 non-null    object
 14  Water1        2944 non-null    object
 15  SnowFall      2944 non-null    object
 16  PrecipTotal   2944 non-null    object
 17  StnPressure   2944 non-null    object
 18  SeaLevel      2944 non-null    object
 19  ResultSpeed   2944 non-null    float64
 20  ResultDir     2944 non-null    int64
 21  AvgSpeed      2944 non-null    object
```
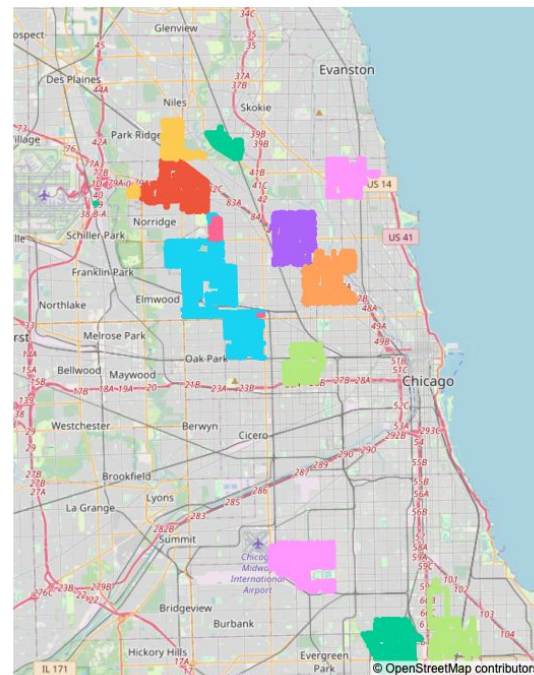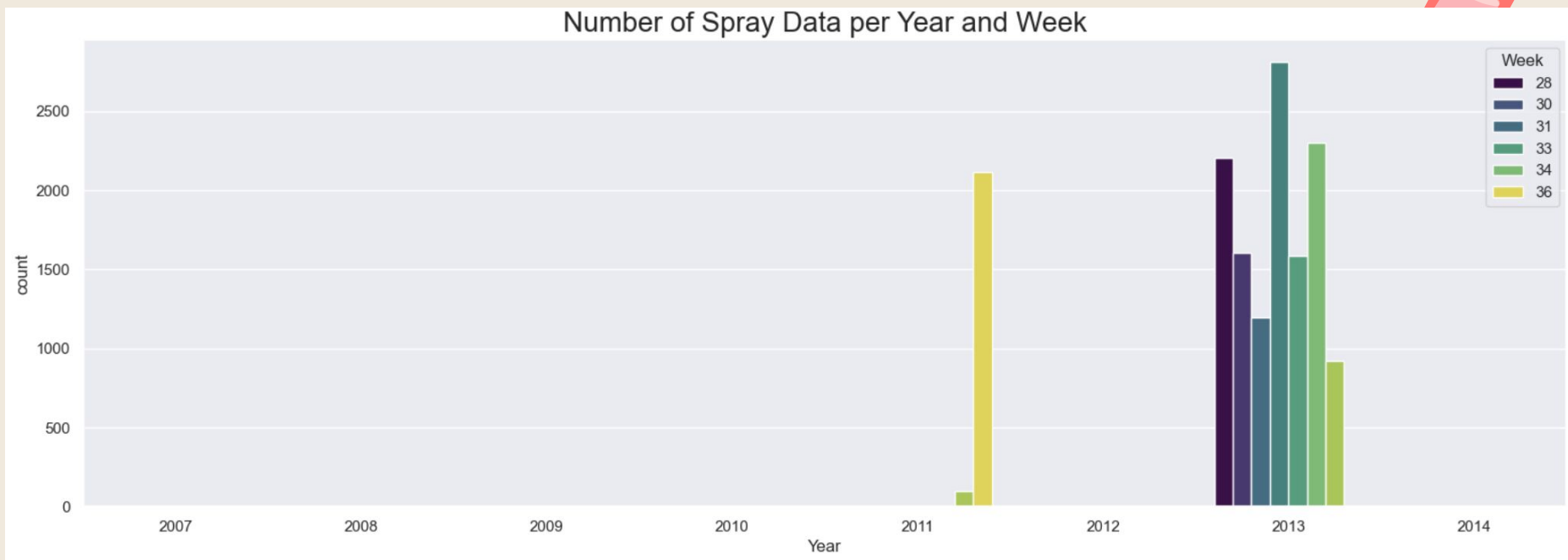
Overview of Datasets

Number of Weather Data per Year and Week
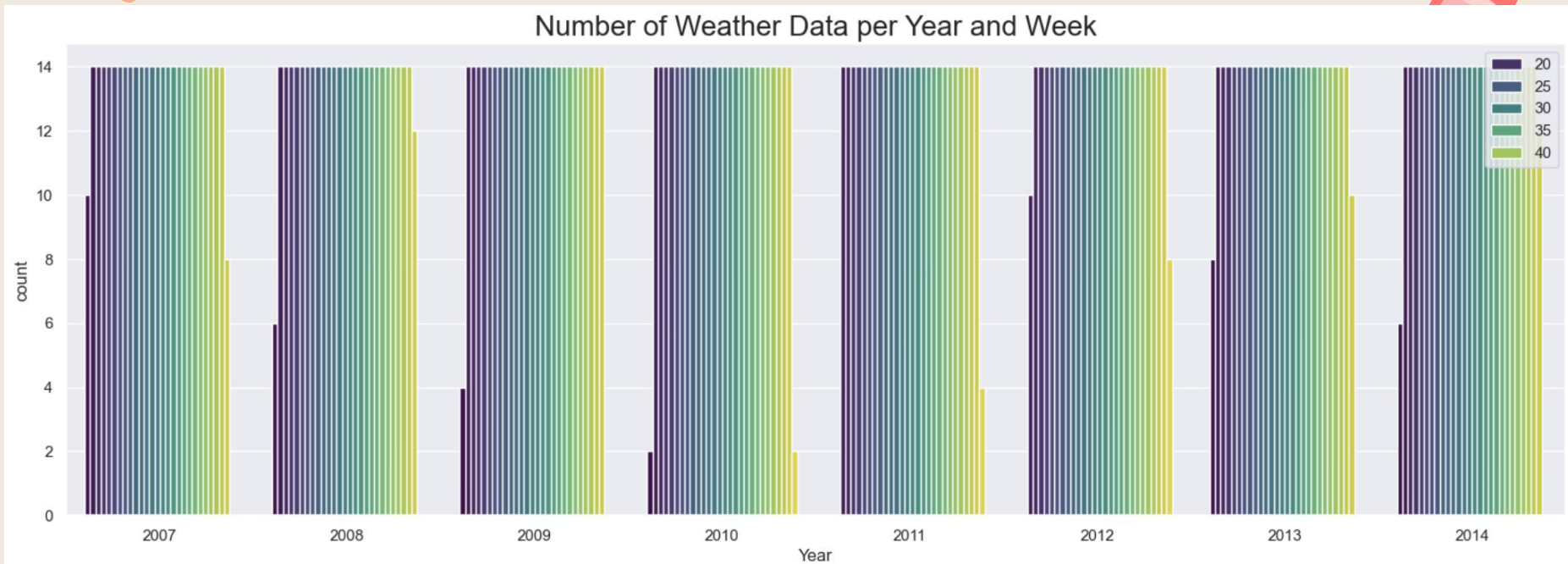
Overview of Datasets

```
RangeIndex: 14835 entries, 0 to 14834
Data columns (total 4 columns):
 #   Column     Non-Null Count   Dtype
---  ------     --------------   -----
 0   Date       14835 non-null   object
 1   Time       14251 non-null   object
 2   Latitude   14835 non-null   float64
 3   Longitude  14835 non-null   float64
```
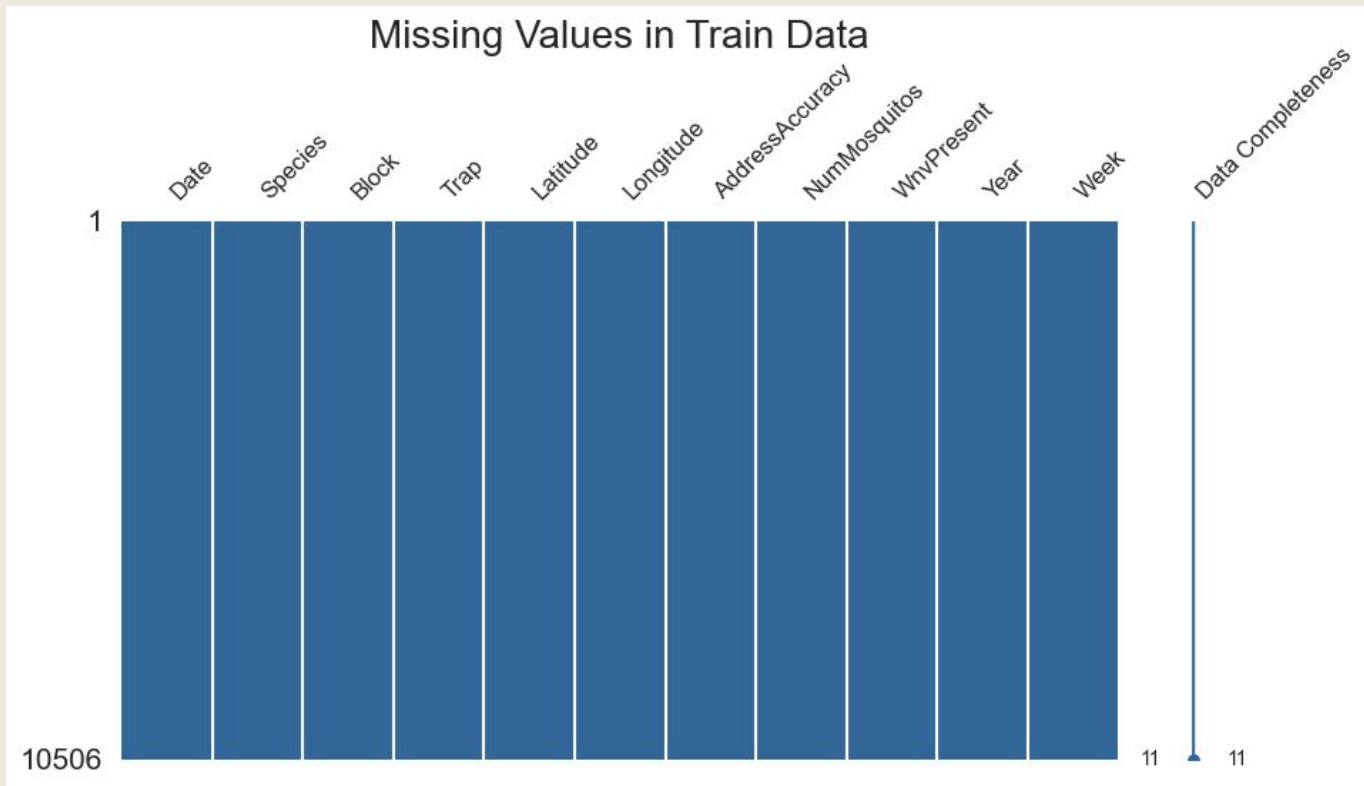
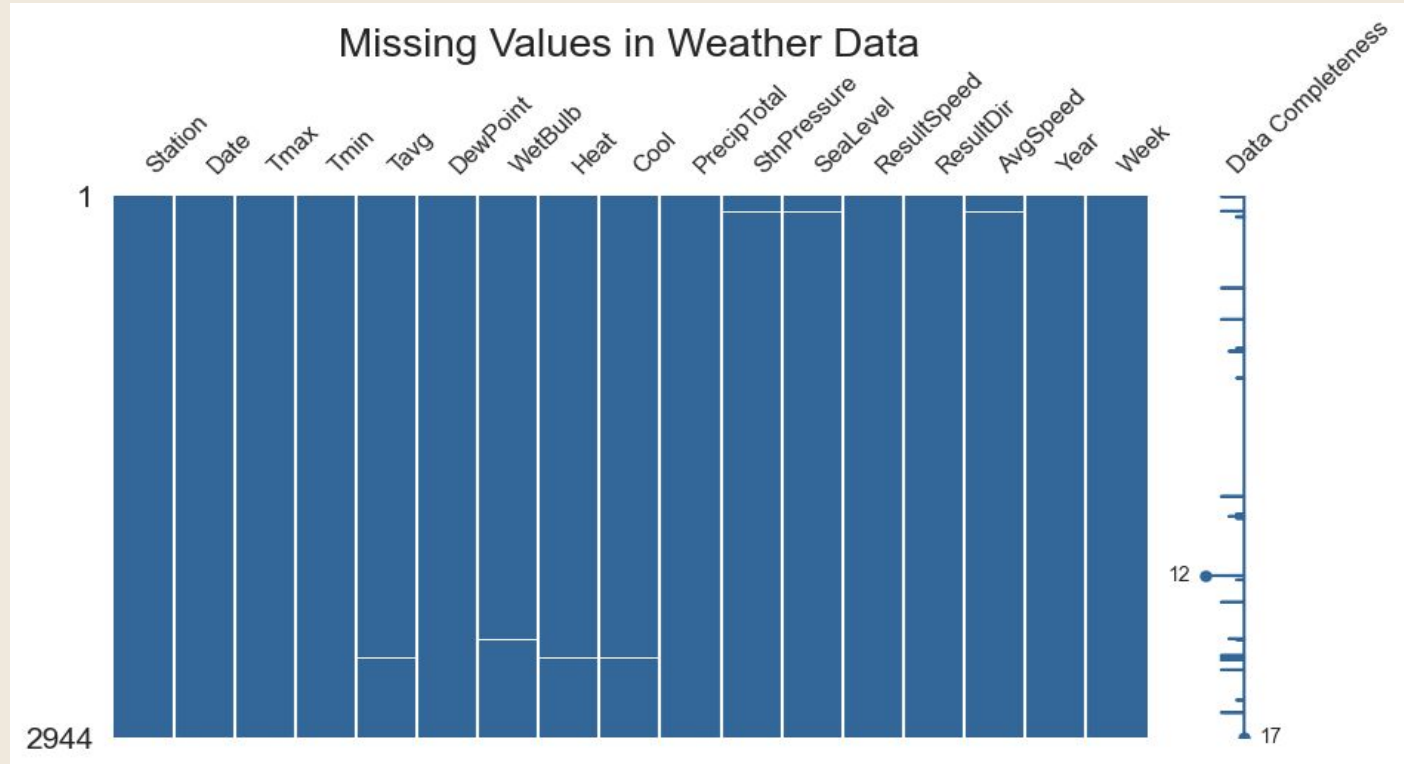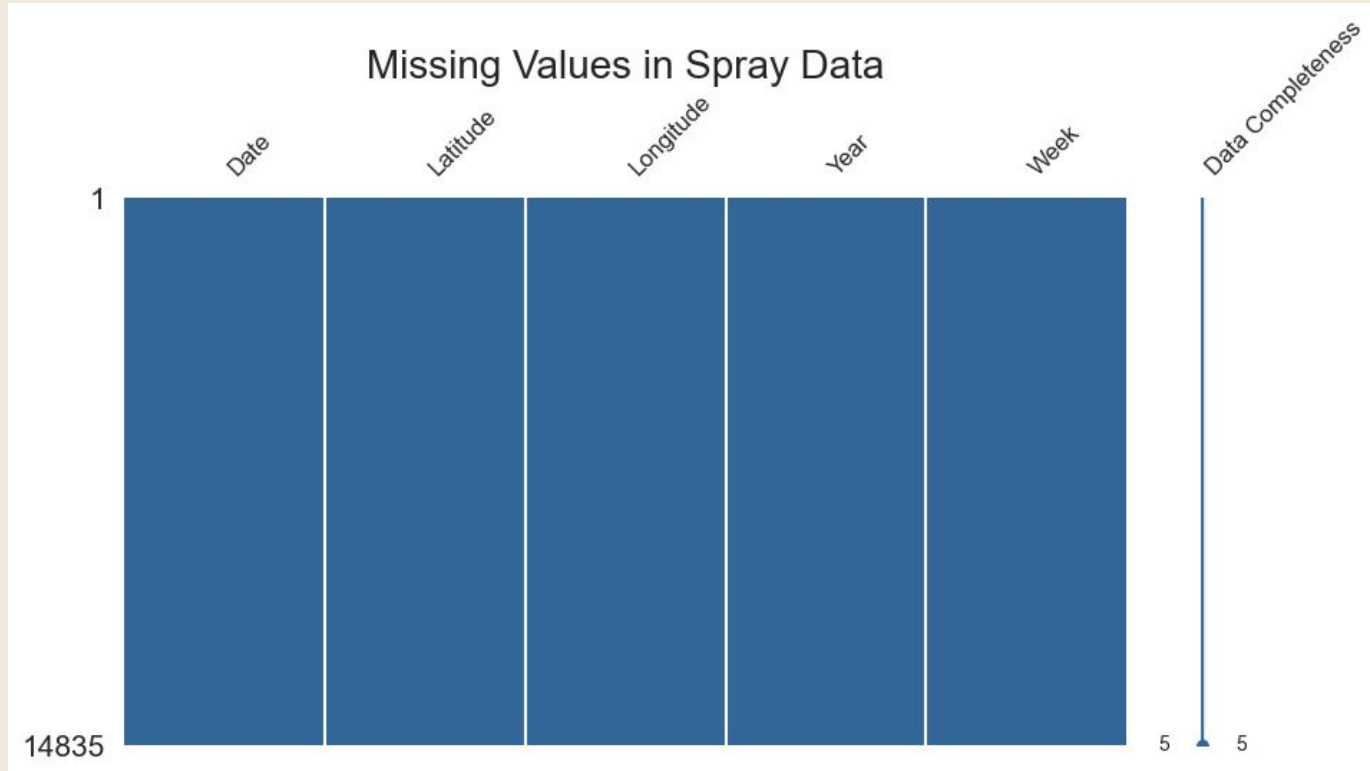Spray collect location and when the sprayed to kill mosquito

Spray Area

Date
- 2011-08-29 00:00:00
- 2011-09-07 00:00:00
- 2013-07-17 00:00:00
- 2013-07-25 00:00:00
- 2013-08-08 00:00:00
- 2013-08-15 00:00:00
- 2013-08-16 00:00:00
- 2013-08-22 00:00:00
- 2013-08-29 00:00:00
- 2013-09-05 00:00:00

© OpenStreetMap contributors

Overview of Datasets

Number of Spray Data per Year and Week

Missing Values

Missing Values in Train Data

Missing Values (Cont.)

Missing Values in Weather Data

Missing Values (Cont.)

Missing Values in Spray Data

Data Dictionary

Train Data

Data Name	Description	Type	Example
Id	Show ID	String	123456
Date	Show Date	Datetime	20020-01-01
Address	Approximate Address of the trap location	String	4100 North Oak
Species	Mosquito's Species	String	CULEX
Block	Building Block	Integer	41
Street	Street Name	String	N OAK PARK
Trap	Trap Code Number	String	T002
AddressNumberAndStreet	Address and Street	String	4100 N OAK PARK AVE
Latitude	Show Latitude	String	41.867108
Longitude	Show Longitude	String	-87.654224
count_prev_week_records	Count Virus Present Previous Week	Boolean	0,1
Wnvpresent	Show West Nile Virus Present	Boolean	0,1

Weather Data

Data Name	Description	Type	Example
Station	Show Station Number	String	1
Date	Show Date	Datetime	2007-01-01
Tavg	Temperature Average	String	65
StnPressure	Station Pressure	Float	22.12
ResultDir	Show the wind direction	Integer	23
AvgSpeed	Show Average Wind Speed	Float	20.5

Spray Data

Data Name	Description	Type	Example
Date	Show Date	Datetime	2011-01-01
Latitude	Latitude	float	42.391623
Longitude	Longitude	float	-88.089163

Spray data limitation

Spray data is limited to the year 2013, and we decide to impute it only for that year, there is a challenge when it comes to predicting the test data.
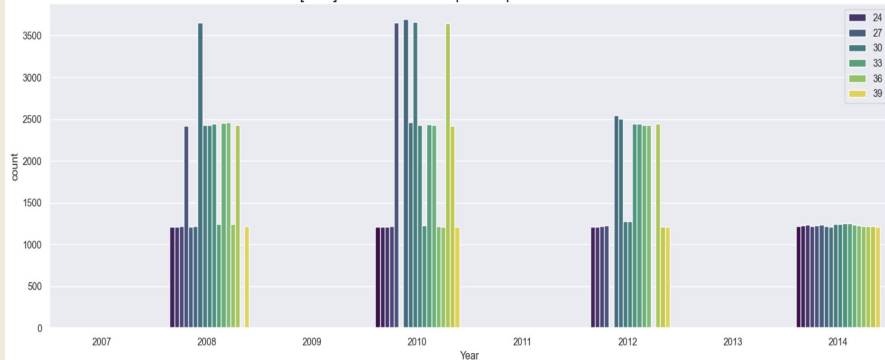
Since our model will have no information about spraying in other years, any impact of spraying on mosquito activity or the West Nile Virus would be unaccounted for.

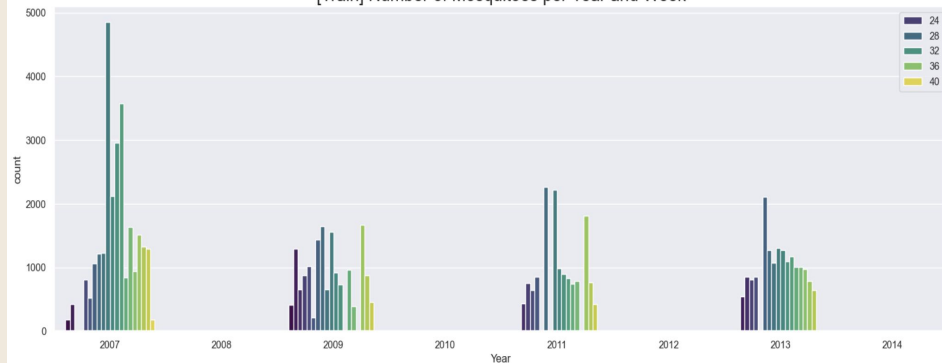So we decide to **exclude** Spray data from model

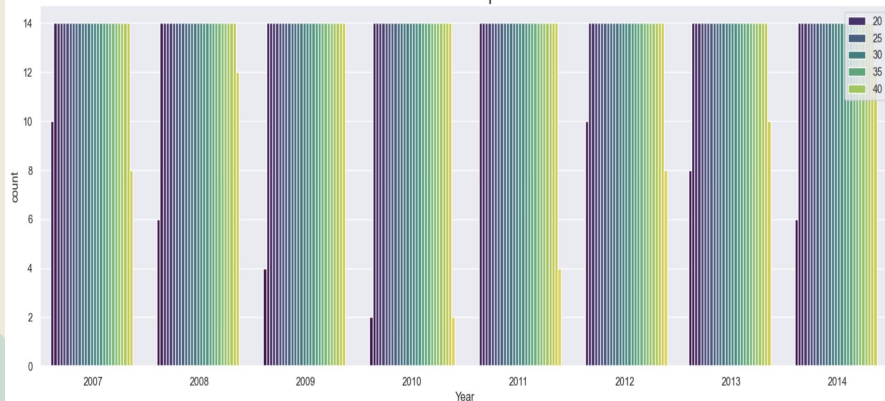From Limited Data, and make more complex to Model Interpretability

# Spray data limitation
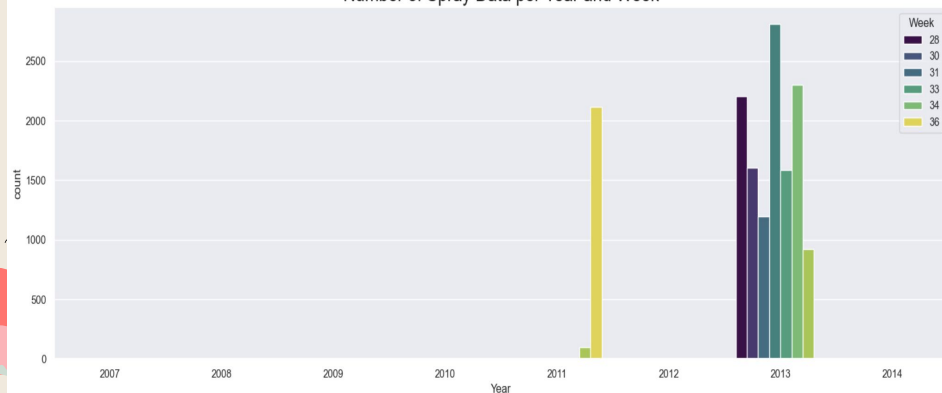
Weather Station

As weather data is collected from two stations,

we split the data into two groups and observed differences. However, we found no significant distinctions in the histogram.

Consequently, we then combine the data from these two stations into an average dataset.

* If any is missing from one station, uses data from another station instead

Weather Station 1

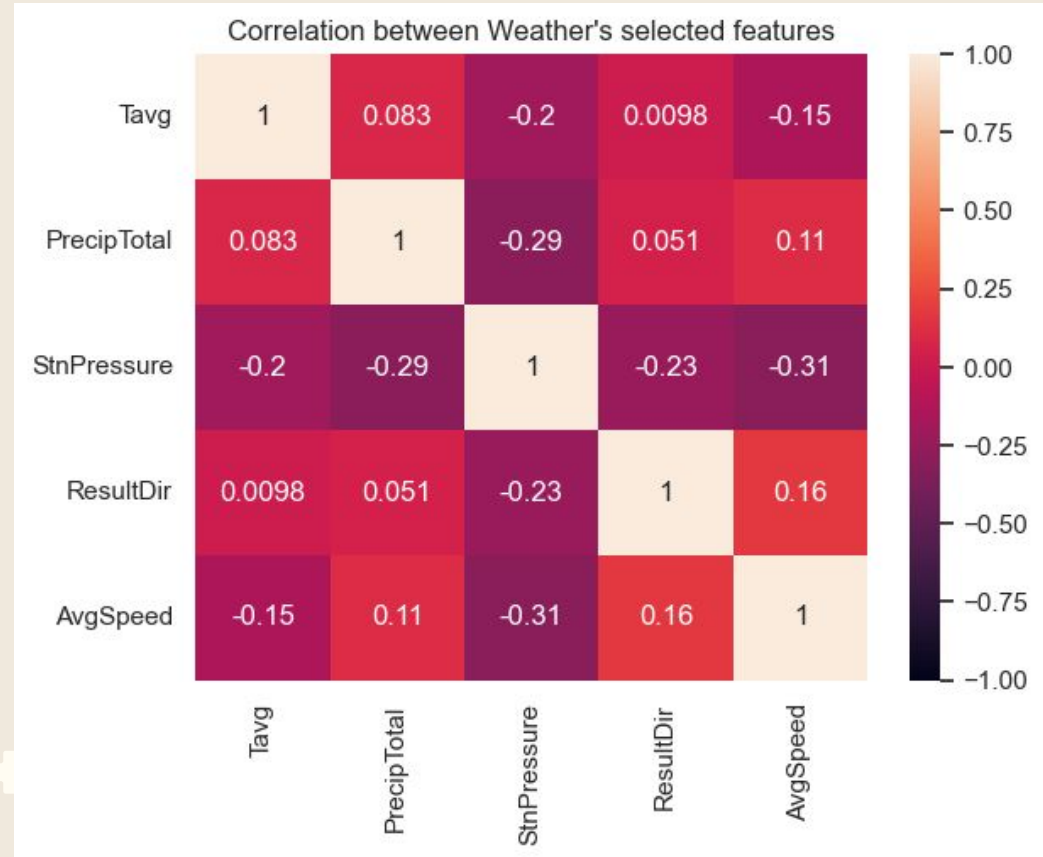

Weather Station 2
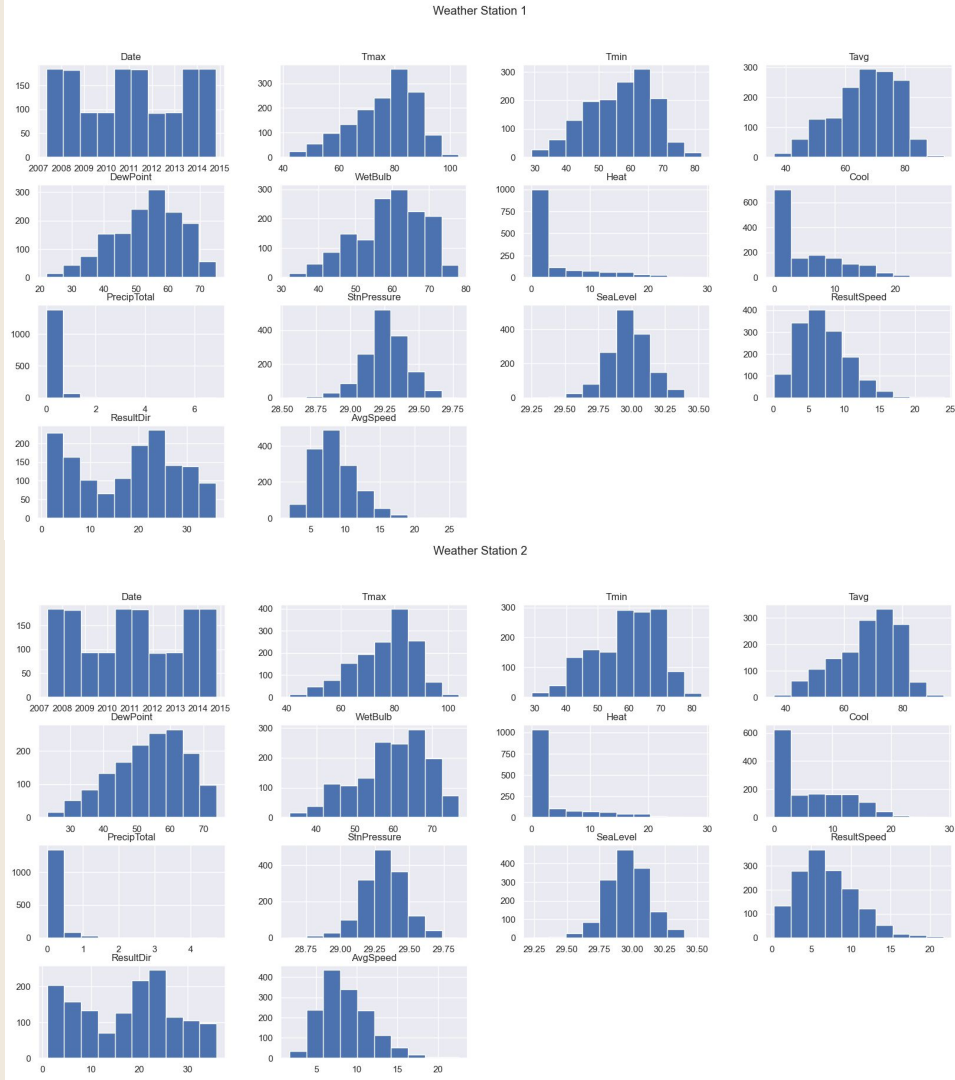
Weather Correlation

After exploring the correlation among features,

we remove those that strongly correlate with each other and choose only the features that represent these groups.

Correlation between Weather's features
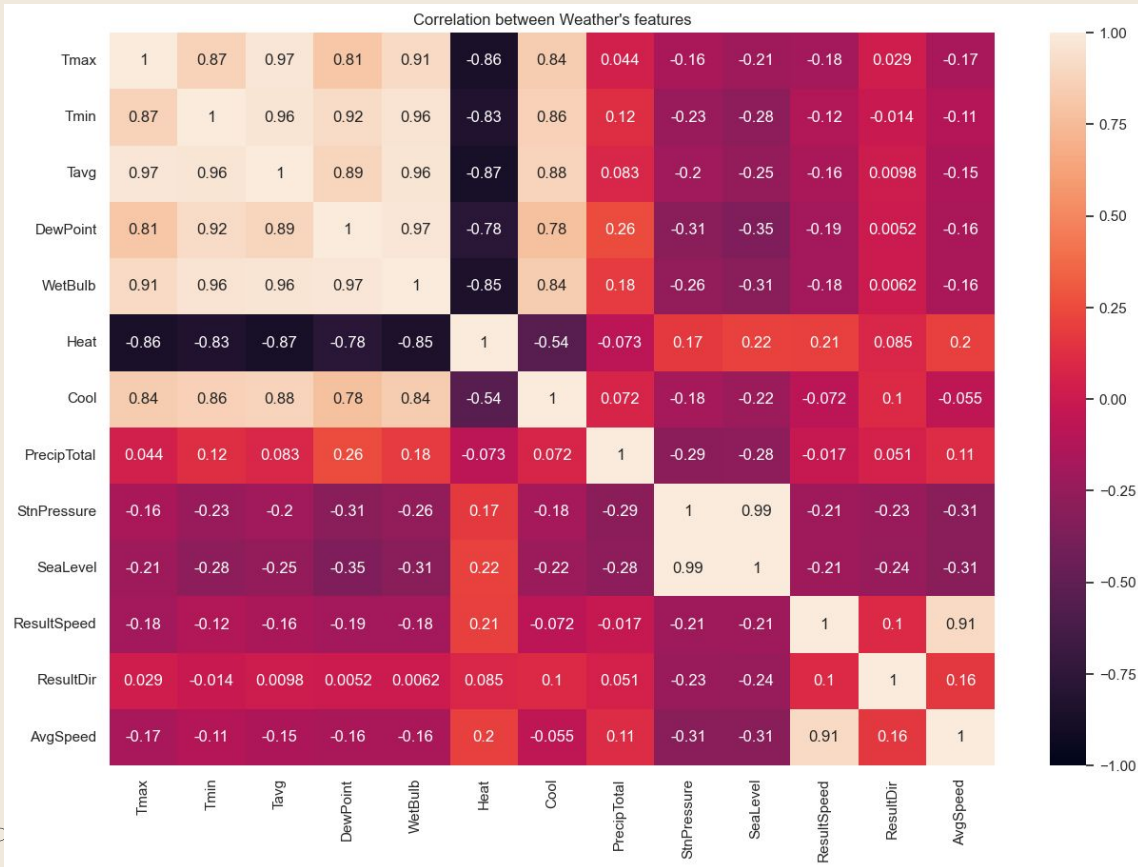
Weather Correlation

Correlation after drop some features

Correlation between Weather's selected features

Weather data correction

Replace malform data eg. M, T, - that found in weather dataset

M from missing replace with NaN
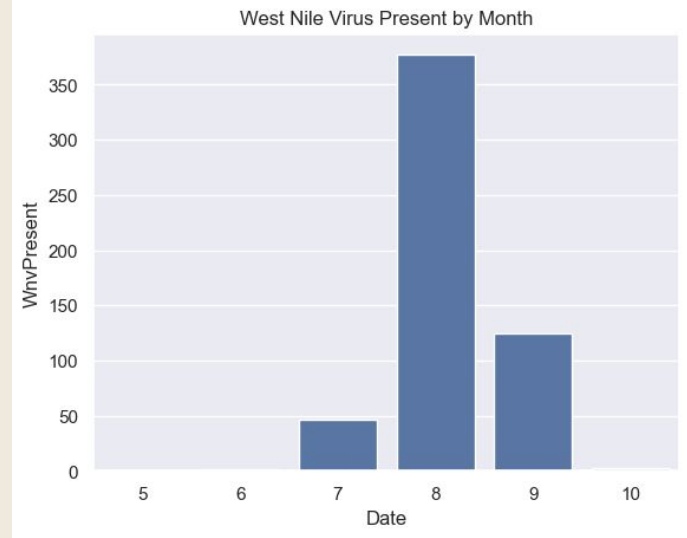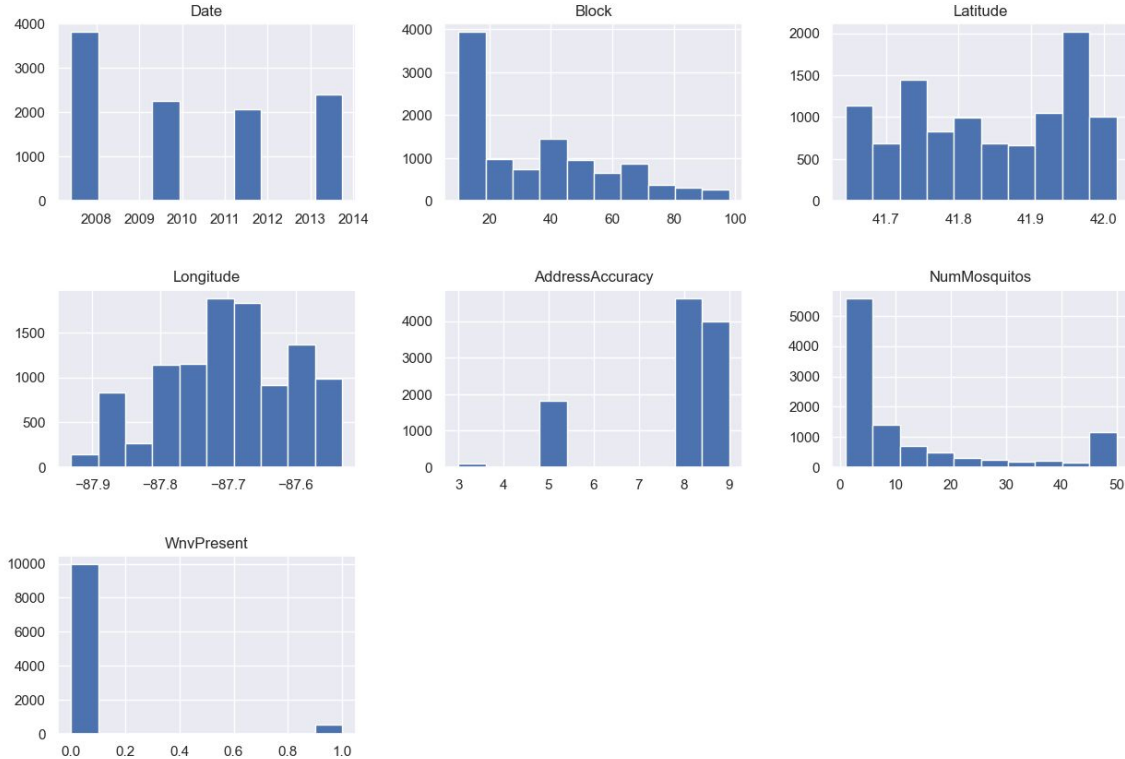
T from Trace replace with small number 0.001
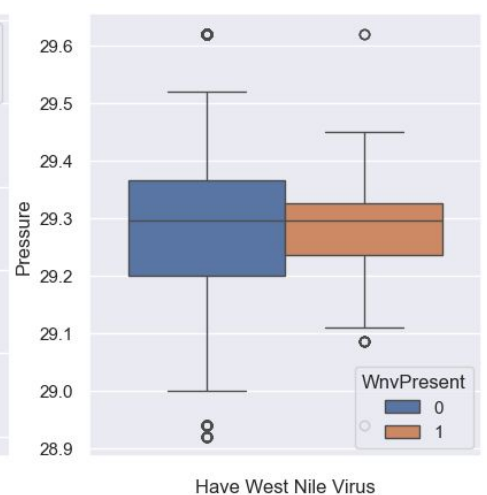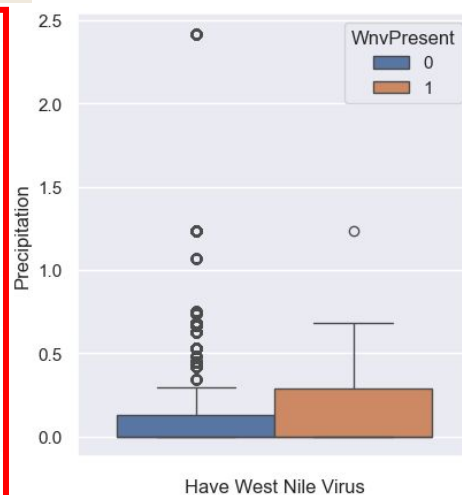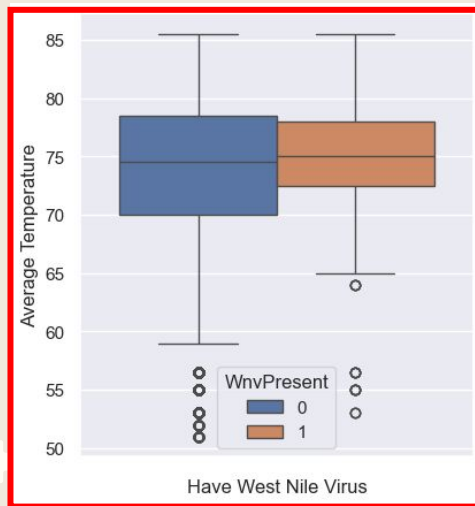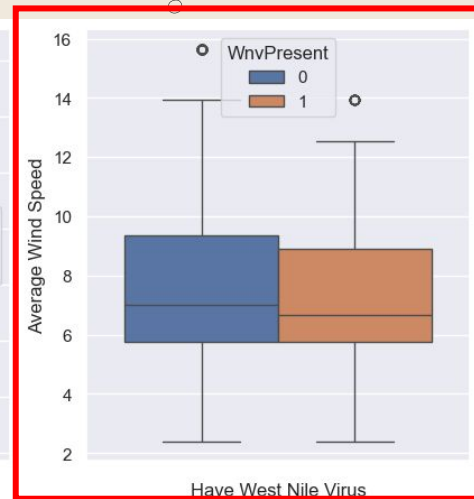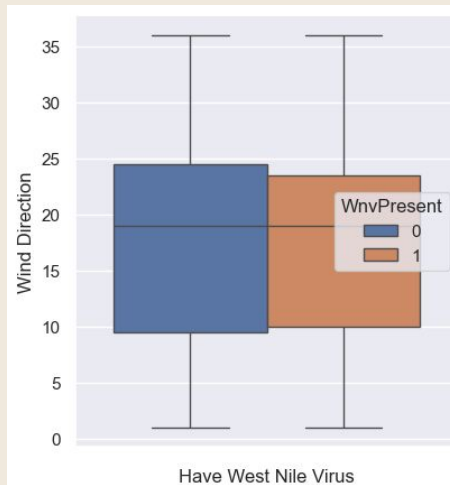
- replace with NaN

Train data distribution

Feature Selection Weather

- One that correlated with WNV
  - Average Temp.
  - Average Wind Speed
  - Uses moving average value

Feature Selection & Engineering
Train data - Location and Time

Location

Latitude

Longitude

Time - Extract from date

Weeknum (week of year)

Year

	Date	Address	Species	Block	Street	Trap	AddressNumberAndStreet	Latitude	Longitude	AddressAccuracy	NumMosquitos	WnvPresent
0	2007-05-29	4100 North Oak Park Avenue, Chicago, IL 60634,...	CULEX PIPIENS/RESTUANS	41	N OAK PARK AVE	T002	4100 N OAK PARK AVE, Chicago, IL	41.954690	-87.800991	9	1	0
1	2007-05-29	4100 North Oak Park Avenue, Chicago, IL 60634,...	CULEX RESTUANS	41	N OAK PARK AVE	T002	4100 N OAK PARK AVE, Chicago, IL	41.954690	-87.800991	9	1	0
2	2007-05-29	6200 North Mandell Avenue, Chicago, IL 60646, USA	CULEX RESTUANS	62	N MANDELL AVE	T007	6200 N MANDELL AVE, Chicago, IL	41.994991	-87.769279	9	1	0
3	2007-05-29	7900 West Foster Avenue, Chicago, IL 60656, USA	CULEX PIPIENS/RESTUANS	79	W FOSTER AVE	T015	7900 W FOSTER AVE, Chicago, IL	41.974089	-87.824812	8	1	0
4	2007-05-29	7900 West Foster Avenue, Chicago, IL 60656, USA	CULEX RESTUANS	79	W FOSTER AVE	T015	7900 W FOSTER AVE, Chicago, IL	41.974089	-87.824812	8	4	0

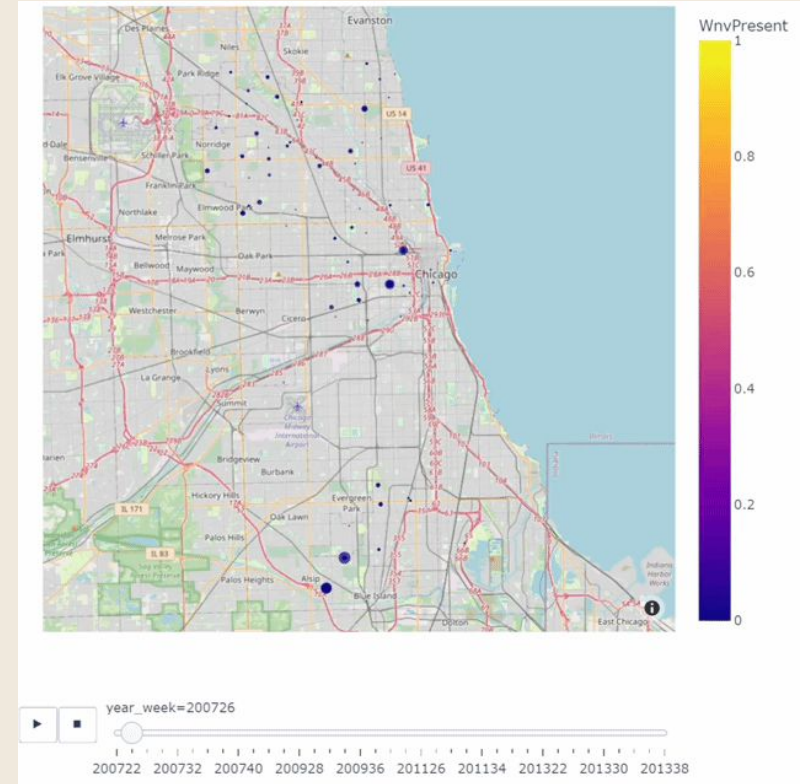# Feature Selection & Engineering Train data – Species
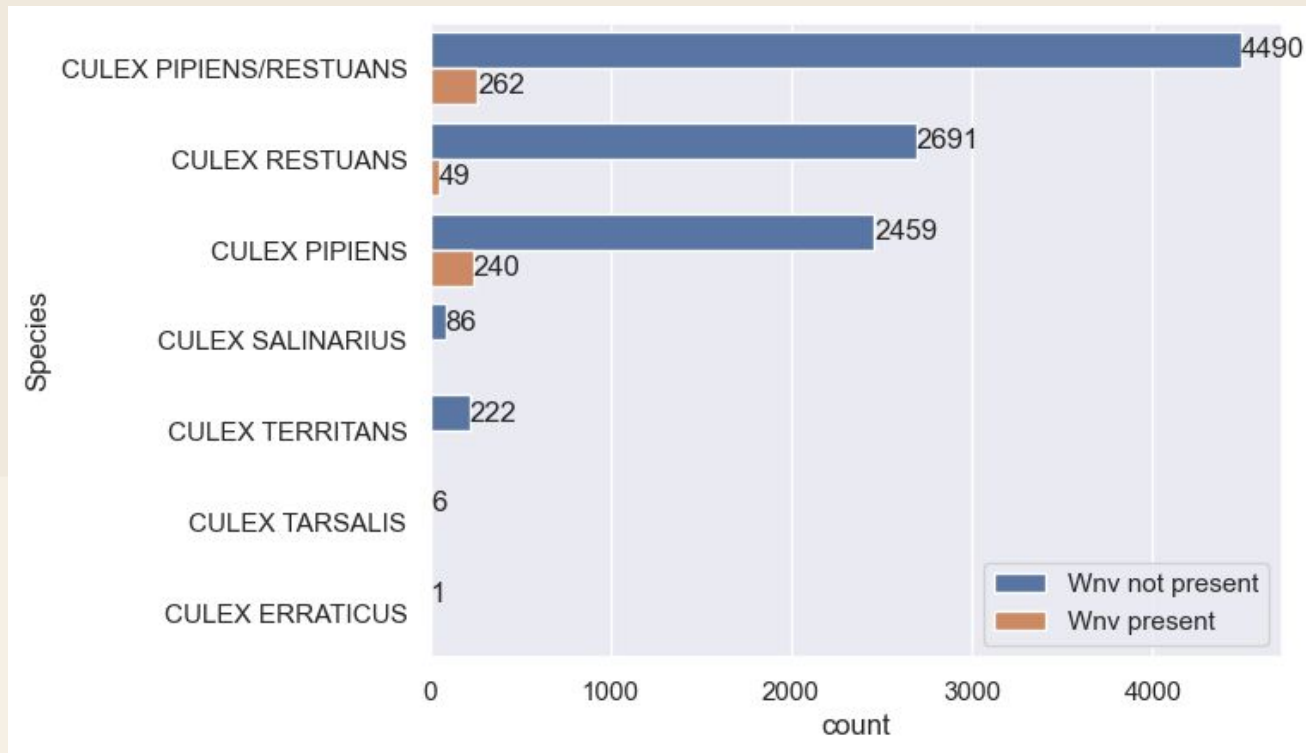
Feature Selection & Engineering Train data – Time related

- As the number of mosquitoes missing in test data

- Solution: Use records count instead

- Create the difference of number of records from previous and current week

Feature Selection & Engineering
Train data – West Nile virus presence proportion

- Calculate the number of records by species, location (lat/long) and week of year

- From the number of records, using the average number of records to calculate proportion (group by species and week of year)

	Species	Latitude	Longitude	weeknum	count_wnv_not_present	count_wnv_present
0	CULEX ERRATICUS	41.974689	-87.890615	35	1.0	0.0
1	CULEX PIPIENS	41.644612	-87.604498	30	1.0	0.0
2	CULEX PIPIENS	41.644612	-87.604498	31	1.0	0.0
3	CULEX PIPIENS	41.644612	-87.604498	32	1.0	0.0
4	CULEX PIPIENS	41.644612	-87.604498	33	2.0	0.0
...	...	...	...	...	...	...
4474	CULEX TERRITANS	42.006858	-87.675919	38	1.0	0.0
4475	CULEX TERRITANS	42.010412	-87.662140	23	1.0	0.0
4476	CULEX TERRITANS	42.010412	-87.662140	30	1.0	0.0
4477	CULEX TERRITANS	42.011601	-87.811506	39	1.0	0.0
4478	CULEX TERRITANS	42.017430	-87.687769	33	1.0	0.0

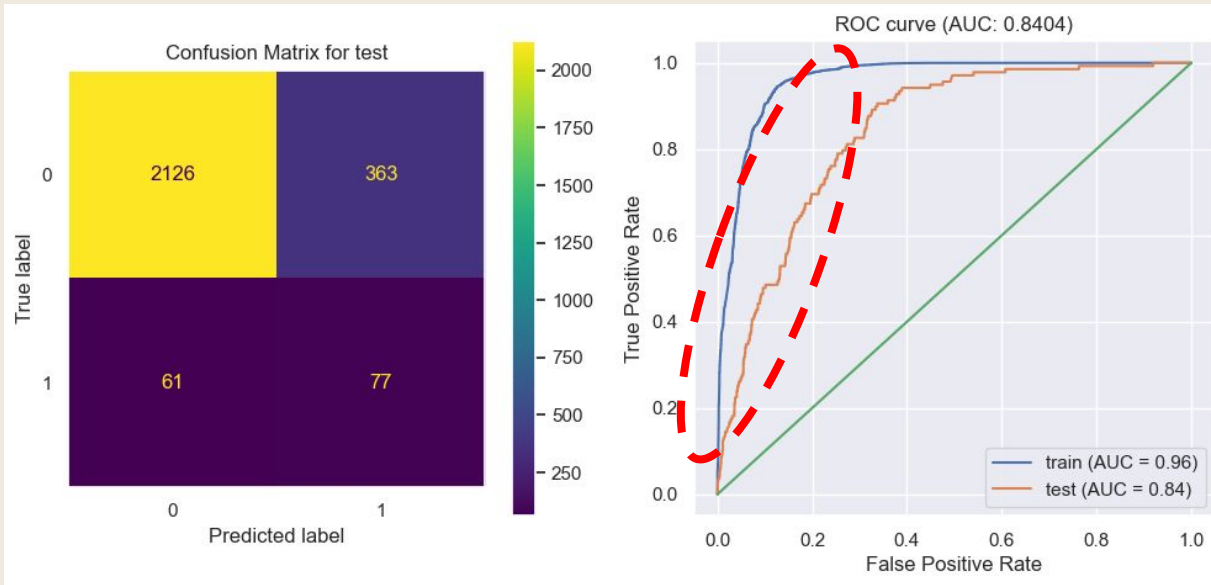|	Latitude	Longitude	wnv_present_proportion
0	41.644612	-87.604498	0.058824
1	41.659112	-87.538693	0.120000
2	41.662014	-87.724608	0.089054
3	41.673408	-87.599862	0.152443
4	41.678618	-87.559308	0.014252
...	...	...	...
133	42.008314	-87.777921	0.147783
134	42.009876	-87.807277	0.141176
135	42.010412	-87.662140	0.000000
136	42.011601	-87.811506	0.024390
137	42.017430	-87.687769	0.000000

Feature Selection - summary

- Average temp.

- Average wind speed

- Latitude

- Longitude

- Week of year

- Species (dummified)

- Diff # records, 1 week lagging

- Diff # records, 2 weeks lagging

- Average proportion of West Nile virus presence in each area

Model Selection

| Model Name	Recall	Train AUC	Test AUC	Features
Logistic Regression	0.74	0.80	0.78	**Number Feature**: Latitude, Longitude, weeknum, year, Tavg (Average Temperature), Avgspeed (Average Wind Speed)
Random Forest	0.60	0.96	0.84	
Regularized Greedy Forest (RGF)	0.52	0.97	0.84	**Category Feature**: Species **Engineering Feature**: count current week diff, count previous week diff and wnv present proportion
XGBoost	0.75	0.87	0.81	

Error Analysis

Random Forest

- Train and Test score is big different

- This ROC curve shows train and test lines

- If they are not near each other

- model is overfitted

Best Model

Logistic Regression
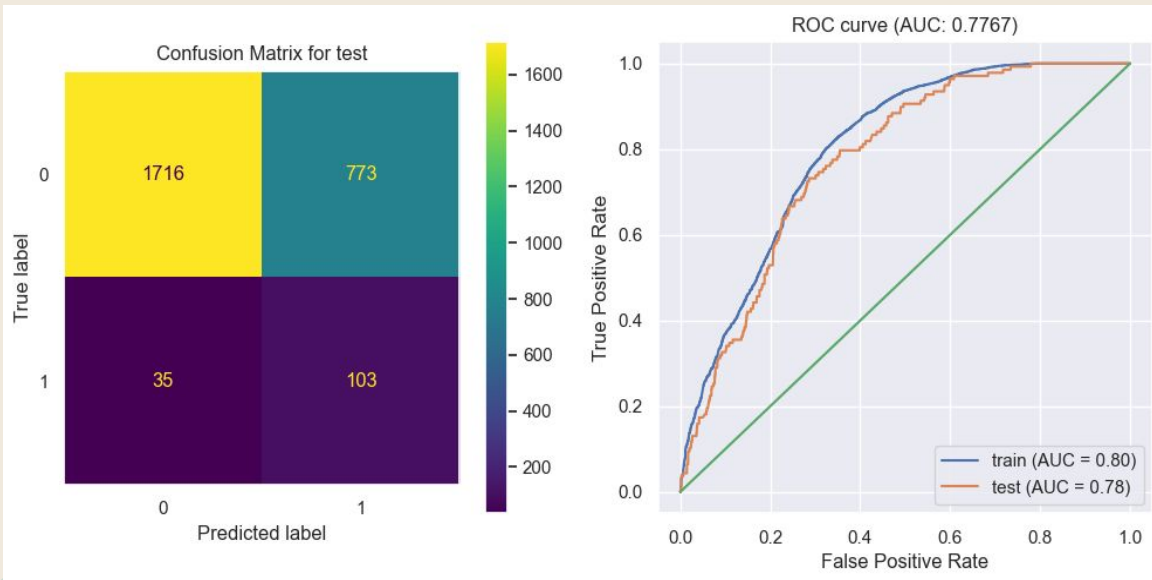
- These two lines are near each other

- The prediction is show it is probability to have West Nile Virus

Summary

- Model for prediction – using Logistic Regression
  - Accuracy: 0.73
  - Recall: 0.74

- Inferencing, probability of having WNV presence
  - Culex Pipiens: expected increase by 6.22 times
  - Culex Pipiens/Restuans: increase by 4.2 times
  - Culex Restuans: increase by 1.5 times
  - Weather condition
    - Higher temperature increase the chance by 1.15 times
   - Higher average wind speed will reduce the chance by 0.62 times

Recommendation

- Cost effective approach

- Yellow circles indicate the presence of the West Nile virus in the mosquitoes

- Optimize the use of resources by targeting areas predicted to have a higher risk

- 2014 Predictions (only July - Sep)
  - Employing various probability thresholds
  - Adjusting threshold percentage
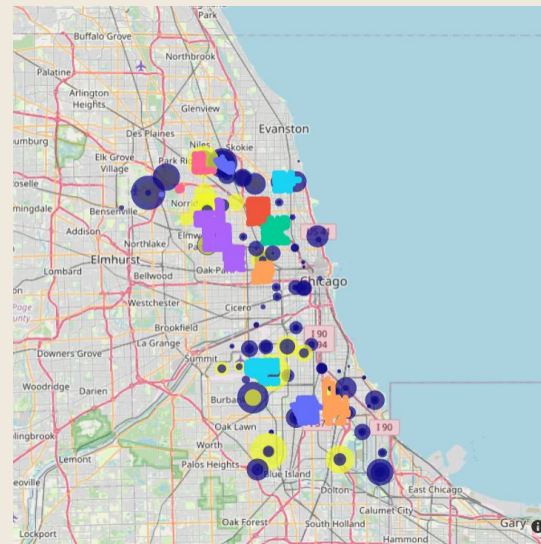
Previous approach
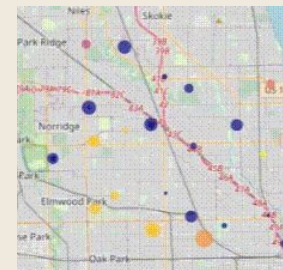
Color changes = sprayed nearby within 1 km with its decay time

	Threshold	# locations to be sprayed	Missing locations (+ error margin)
0	10.48%	151	0.0%
1	82.0%	151	2.0%
2	83.0%	144	6.64%
3	84.0%	137	11.27%
4	85.0%	125	19.22%
5	86.0%	110	29.15%
6	87.0%	95	39.09%
7	88.0%	75	52.33%
8	89.0%	53	66.9%
9	90.0%	30	82.13%

Challenges & Limitation

- Small dataset & Class Imbalance
  - As mentioned before, the training dataset has significantly less data than the test dataset, posing a challenge to our modeling efforts.
  - Due to the substantial imbalance in the 'WnvPresent' class, Synthetic Minority Over-sampling Technique (SMOTE) is employed to address this disparity.
- Different time periods
  - Feature lagging is applied to aggregate historical input data for modeling purposes.
  - Rolling Window Validation is utilized to iteratively train and predict for each year
- Model can only be applied to Chicago area where it has been trained on.
- It is essential to retrain the model annually to ensure its optimal performance
  - The discovery of new species also will be incorporated into the model, potentially leading to improved performance.

Thank You