WEST NILE VIRUS

CURBING THE EPIDEMIC

Eric, June, Rebecca, Matt, Tze Ling

Disease & Treatment Agency

Societal Cures In
Epidemiology and New
Creative Engineering

Problem Statement

Due to a recent outbreak of West Nile Virus (WNV), the Chicago Department of Public Health has set up a surveillance and control system.

As part of the efforts to curb the spread of WNV, our agency has been tasked with deriving an effective plan to deploy pesticides throughout the city.

Contents

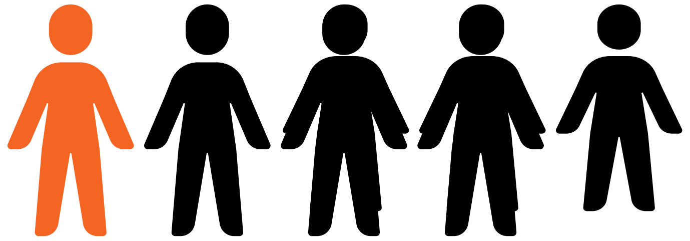
1. Introduction & Data Cleaning (Tze Ling)
2. EDA (June)
3. Feature Engineering & EDA (Matt)
4. Modelling (Eric)
5. Cost-Benefit Analysis, Conclusion and Recommendations (Rebecca)

WNV in US

Top mosquito-borne disease in US

Concern in Illinois where cases have surpassed other states

The Problem



1 in 5 people will develop West Nile fever

Symptoms include fever, headache, tiredness, and body aches, nausea, vomiting, occasionally with a skin rash and swollen lymph glands.

1 in 150

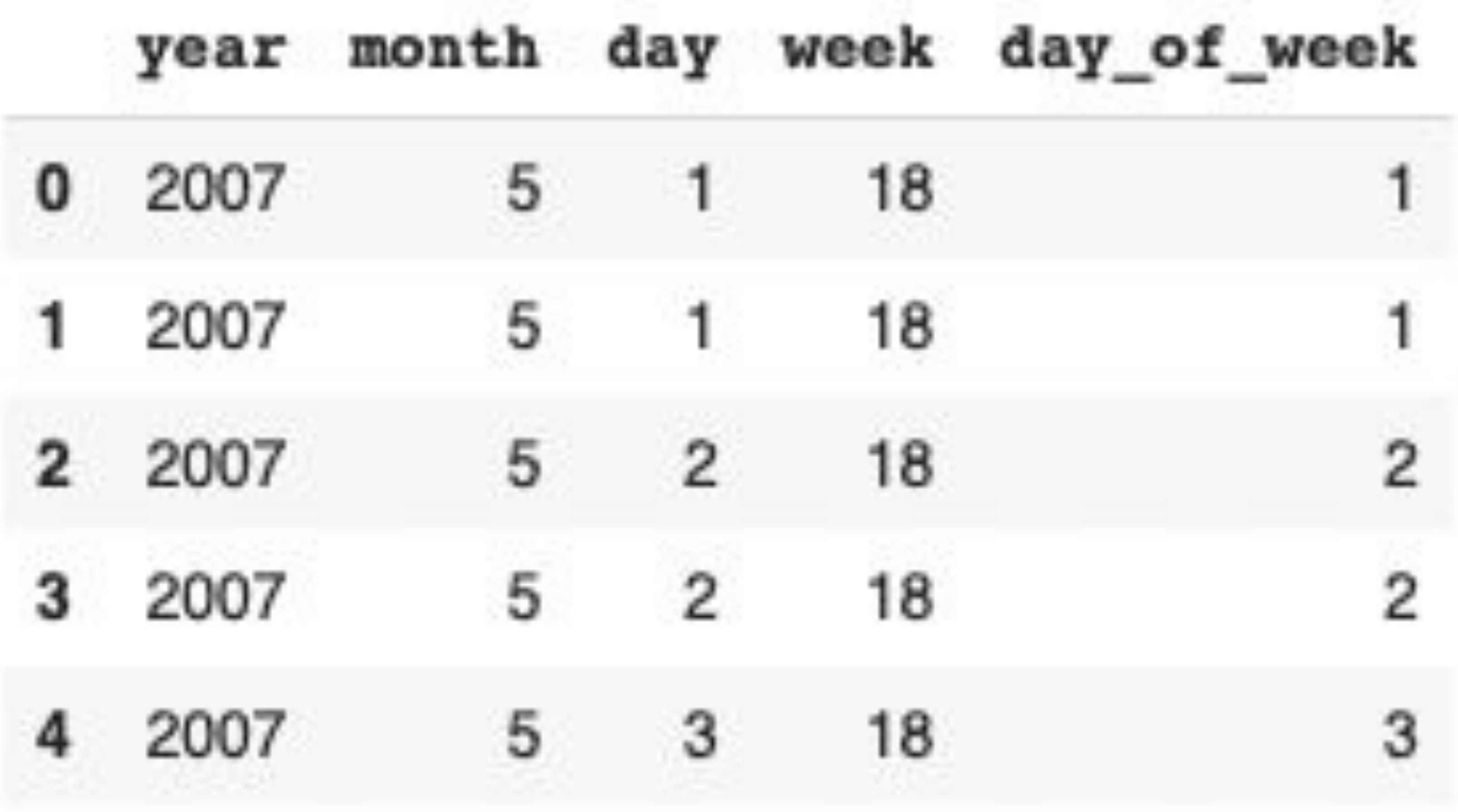
Develop serious neuroinvasive illnesses

1 in 10

with serious neuroinvasive illnesses pass away

Data Cleaning - General

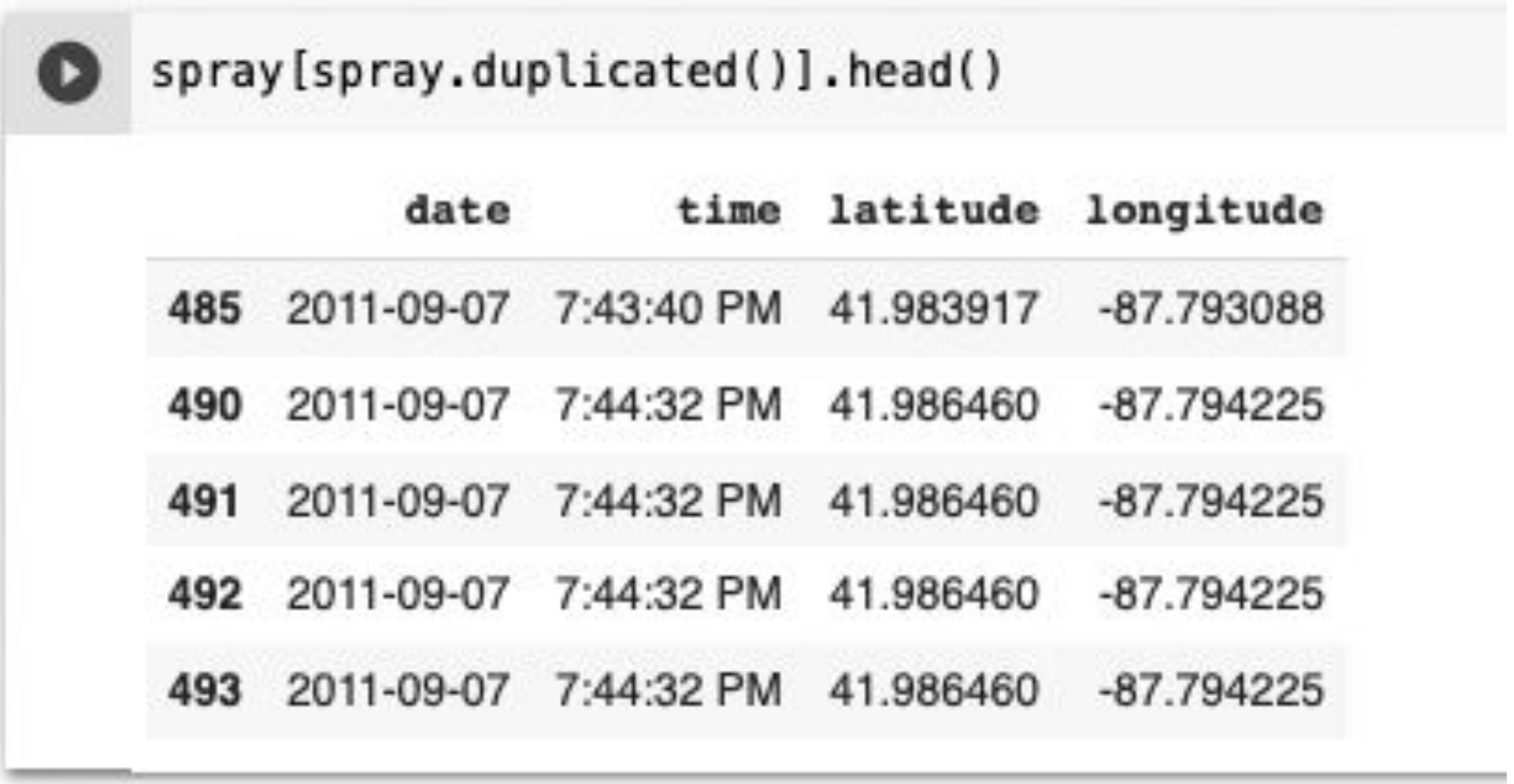
- For all datasets year, month, day, week and day of week were added



	year	month	day	week	day_of_week
0	2007	5	1	18	1
1	2007	5	1	18	1
2	2007	5	2	18	2
3	2007	5	2	18	2
4	2007	5	3	18	3

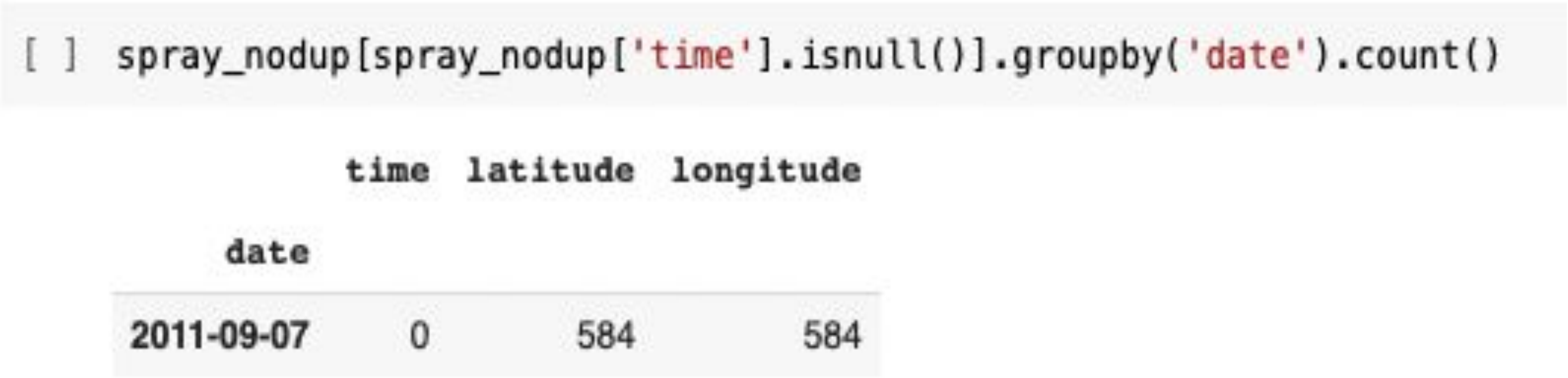
Data Cleaning - Spray dataset

- Duplicates and null values were dropped



```
spray[spray.duplicated()].head()
```

```
           date         time   latitude   longitude
485  2011-09-07   7:43:40 PM  41.983917  -87.793088
490  2011-09-07   7:44:32 PM  41.986460  -87.794225
491  2011-09-07   7:44:32 PM  41.986460  -87.794225
492  2011-09-07   7:44:32 PM  41.986460  -87.794225
493  2011-09-07   7:44:32 PM  41.986460  -87.794225
```

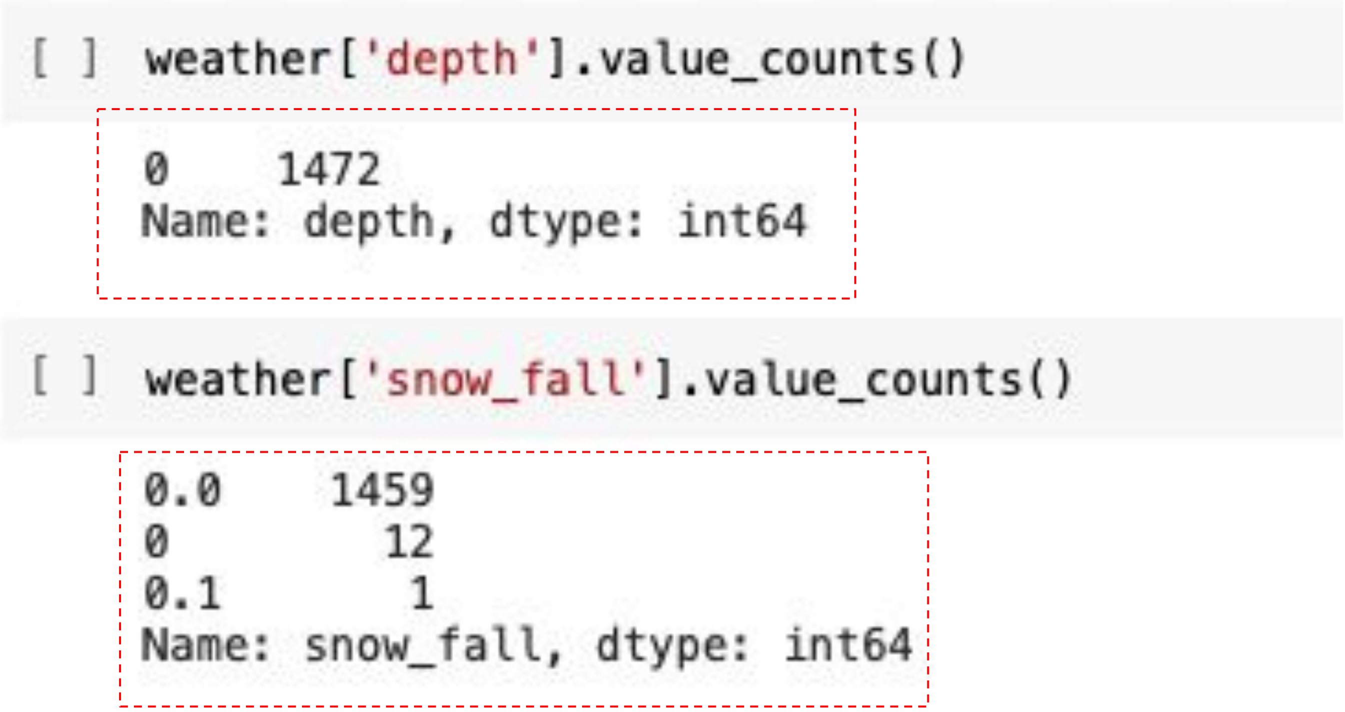


```
[ ] spray_nodup[spray_nodup['time'].isnull()].groupby('date').count()
```

```
            time  latitude  longitude
date
2011-09-07     0       584        584
```

Data Cleaning - Weather dataset

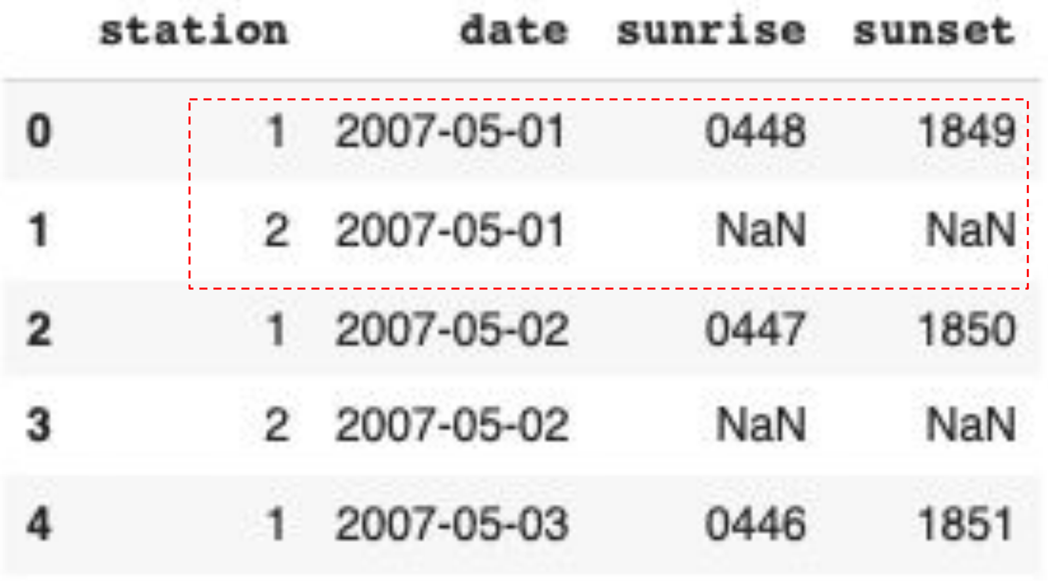
- Columns dropped
  - “water1’ had 100% null values
  - ‘depth’ and ‘snow_fall’ consist of nearly all zeros other than null values
  - ‘code_sum’ had other proxies such as temperature and humidity-related data



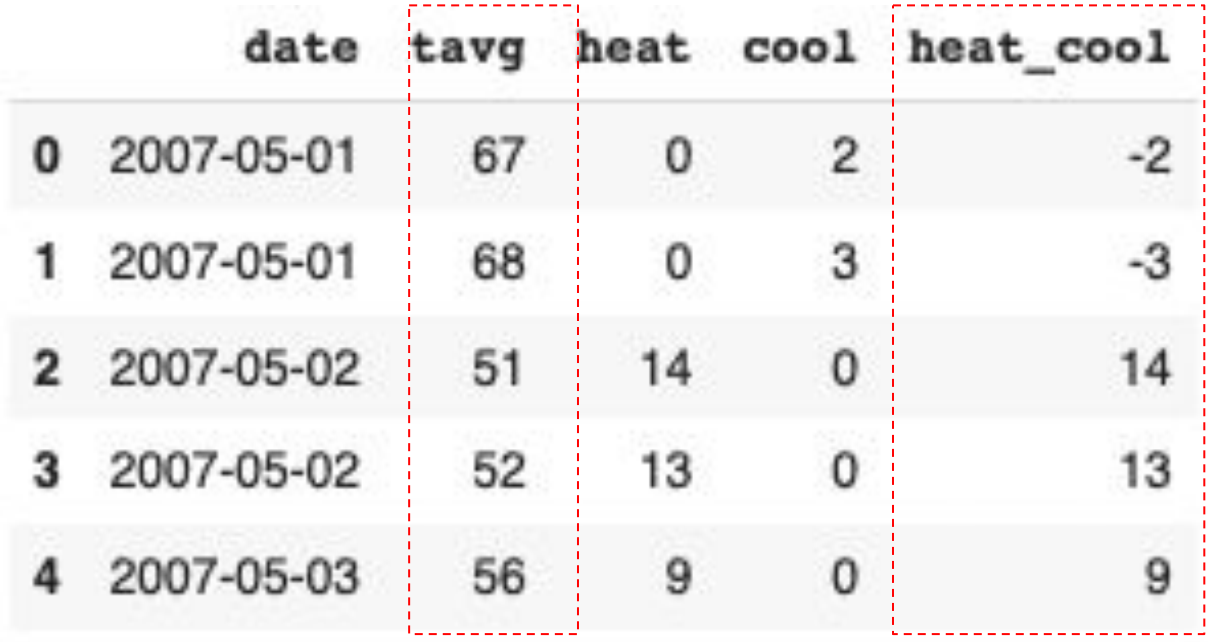
```
[ ]  weather['depth'].value_counts()

0     1472
Name: depth, dtype: int64

[ ]  weather['snow_fall'].value_counts()

0.0     1459
0         12
0.1        1
Name: snow_fall, dtype: int64
```

Data Cleaning - Weather dataset

- Imputed null values (1)
  - ‘sunrise’ & ‘sunset’ imputed values from the other station as they are located in the same city
  - ‘tavg’ used (‘tmax’+’tmin’)/2 for imputing values
  - ‘heat’ & ‘cool’ imputed with difference ‘tavg’ and base temperature
  - ‘depart’ used the difference between ‘tavg’ and normal temperature



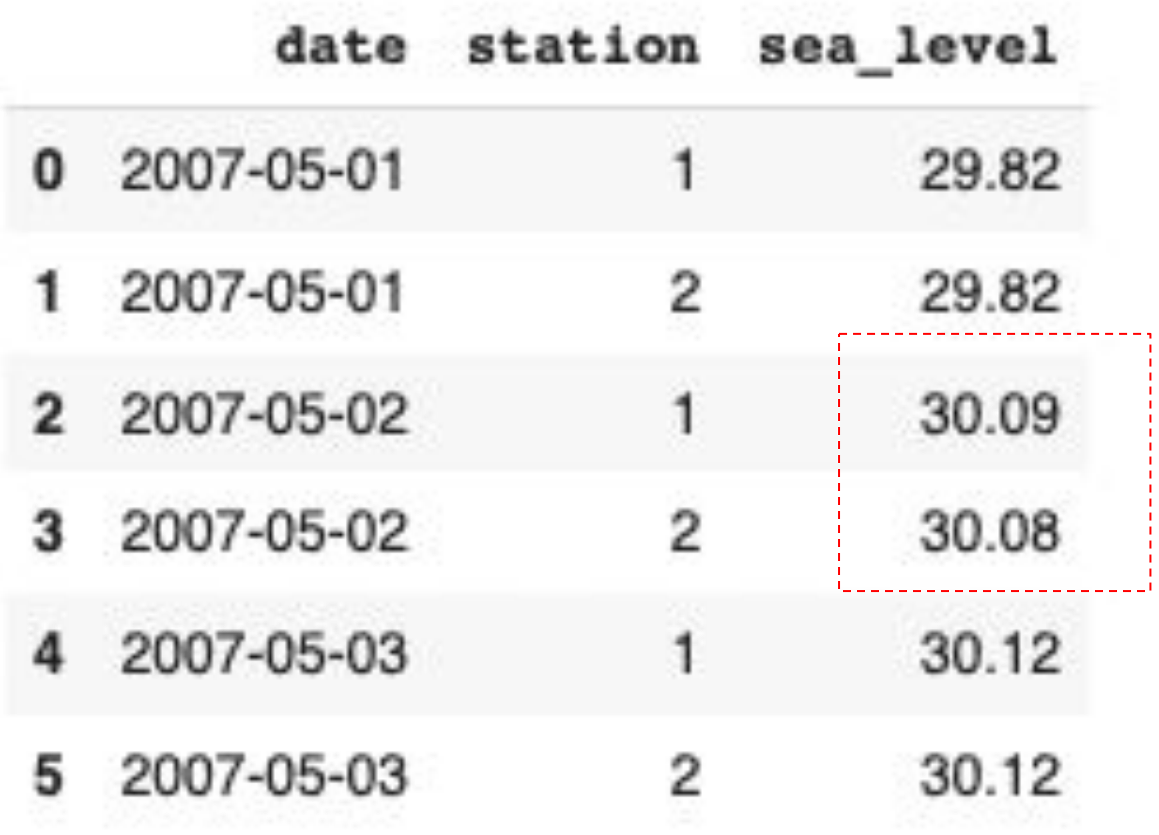
	station	date	sunrise	sunset
0	1	2007-05-01	0448	1849
1	2	2007-05-01	NaN	NaN
2	1	2007-05-02	0447	1850
3	2	2007-05-02	NaN	NaN
4	1	2007-05-03	0446	1851



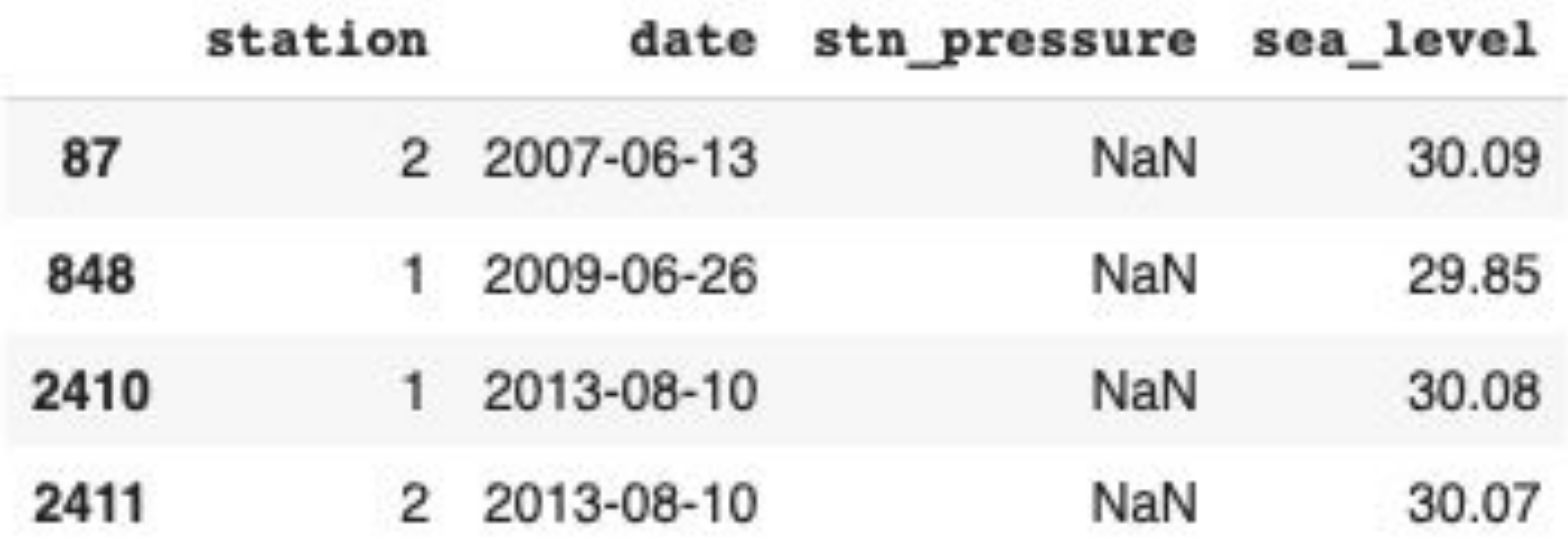
	date	tavg	heat	cool	heat_cool
0	2007-05-01	67	0	2	-2
1	2007-05-01	68	0	3	-3
2	2007-05-02	51	14	0	14
3	2007-05-02	52	13	0	13
4	2007-05-03	56	9	0	9

Data Cleaning - Weather dataset

- Imputed null values (2)
  - ‘sea_level’ imputed from other station as they had negligible difference
  - ‘stn_pressure’ imputed with other station with +/- 0.05



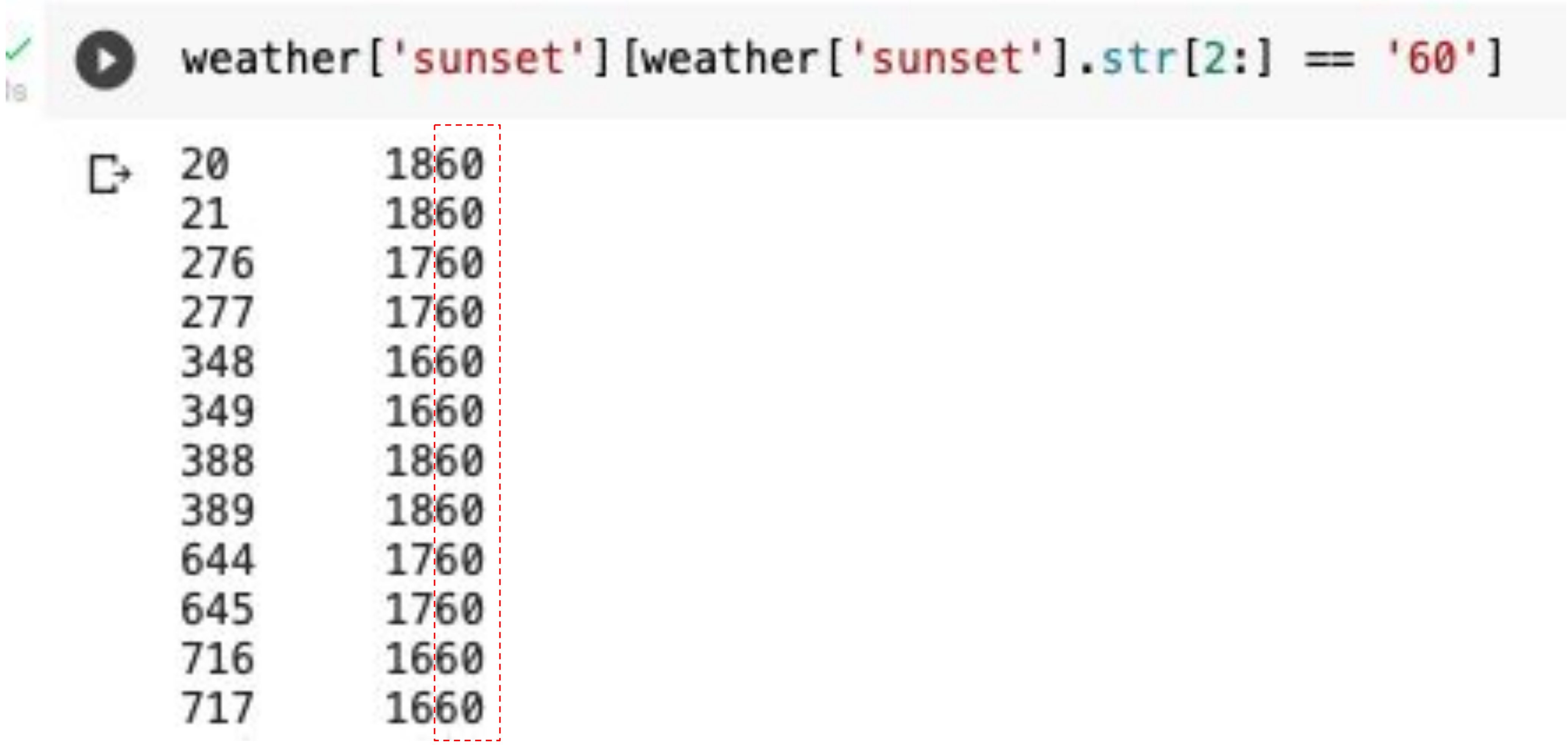
	date	station	sea_level
0	2007-05-01	1	29.82
1	2007-05-01	2	29.82
2	2007-05-02	1	30.09
3	2007-05-02	2	30.08
4	2007-05-03	1	30.12
5	2007-05-03	2	30.12



	station	date	stn_pressure	sea_level
87	2	2007-06-13	NaN	30.09
848	1	2009-06-26	NaN	29.85
2410	1	2013-08-10	NaN	30.08
2411	2	2013-08-10	NaN	30.07

Data Cleaning - Weather dataset

- Cleaning of outliers
    - ‘sunset’ had columns that ended with 60 instead of 00

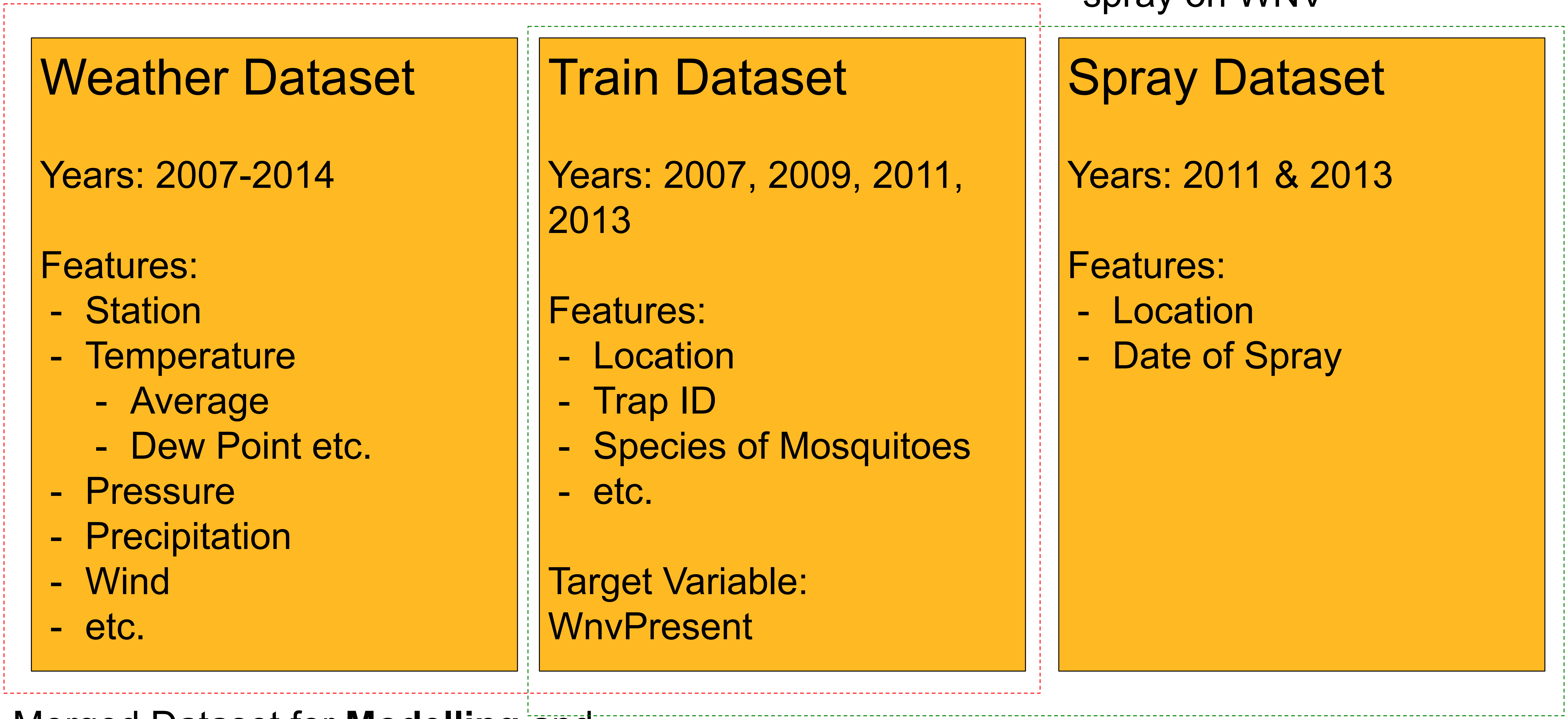


```
weather['sunset'][weather['sunset'].str[2:] == '60']
```

```
20      1860
21      1860
276     1760
277     1760
348     1660
349     1660
388     1860
389     1860
644     1760
645     1760
716     1660
717     1660
```

Datasets

Data Visualisation of effect of spray on WNV



Weather Dataset

Years: 2007-2014

Features:

- Station
- Temperature
  - Average
  - Dew Point etc.
- Pressure
- Precipitation
- Wind
- etc.

Train Dataset

Years: 2007, 2009, 2011, 2013

Features:

- Location
- Trap ID
- Species of Mosquitoes
- etc.

Target Variable: WnvPresent

Spray Dataset

Years: 2011 & 2013

Features:

- Location
- Date of Spray

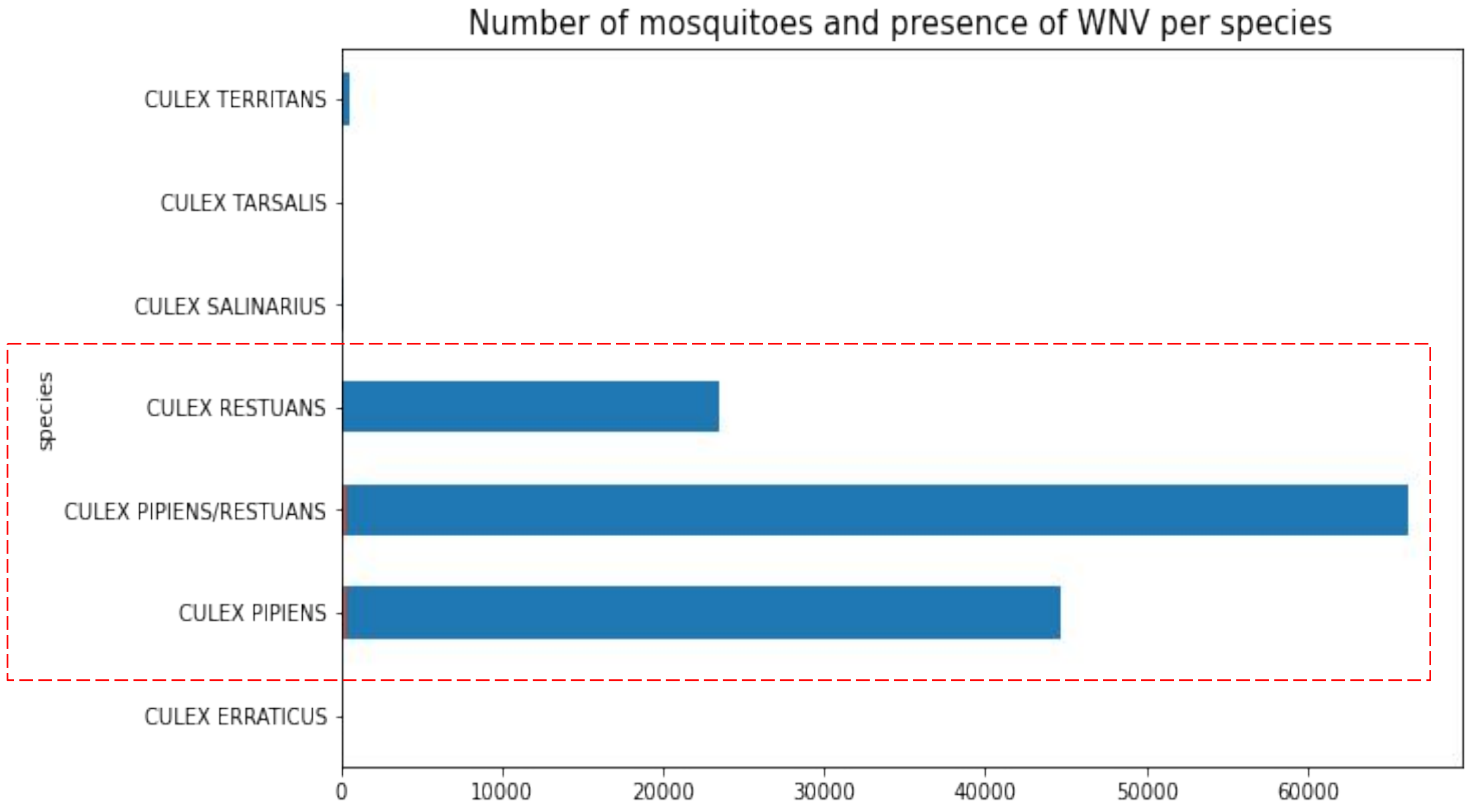
Merged Dataset for **Modelling** and prediction of virus incidence

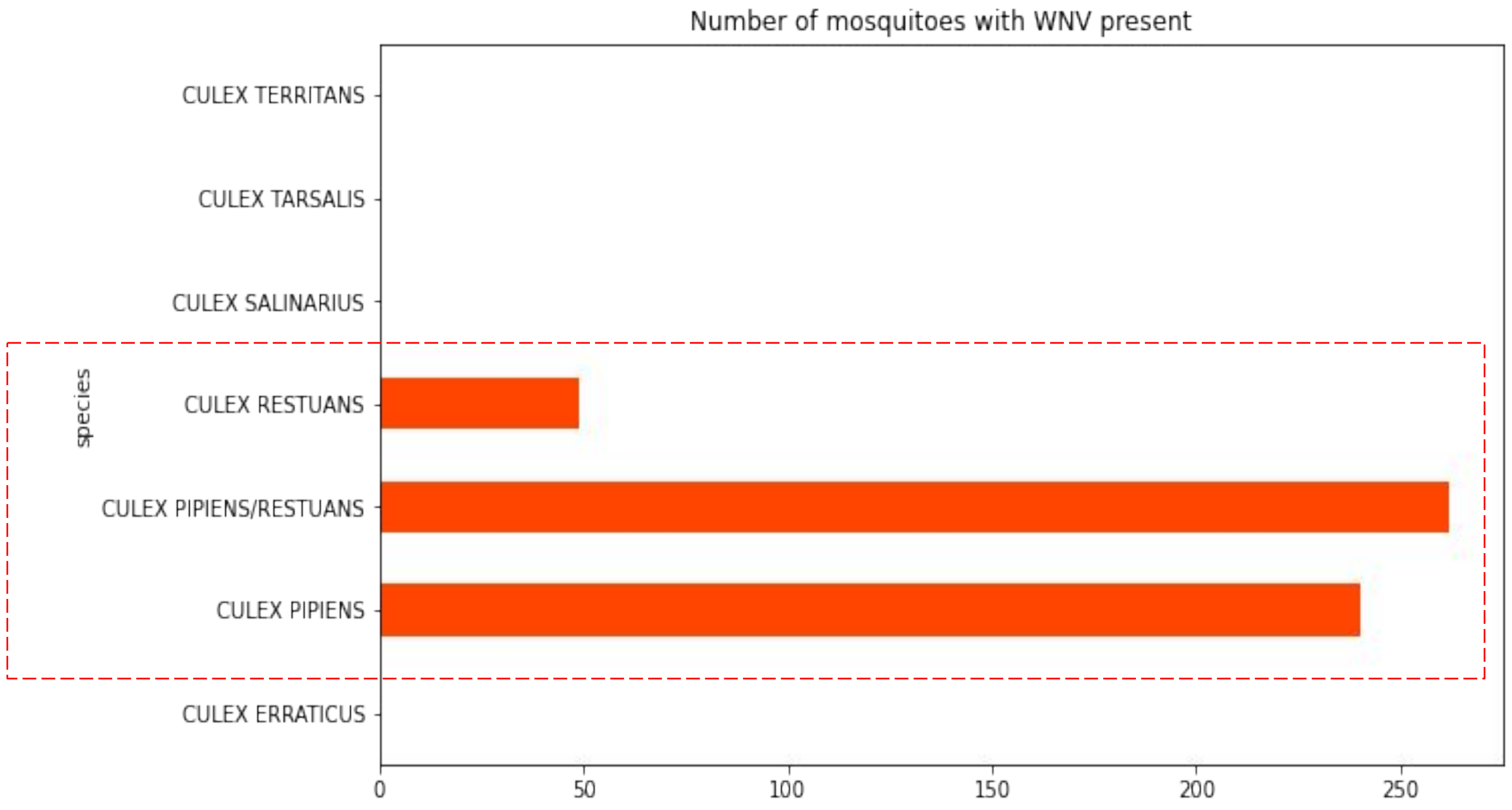
Exploratory Data Analysis (EDA)

- Mosquito Species
- Seasonality/ Time Periods
- Spray & Trap effectiveness and locations
- Weather impact on mosquitos
  - Temperature, Humidity, Precipitation.

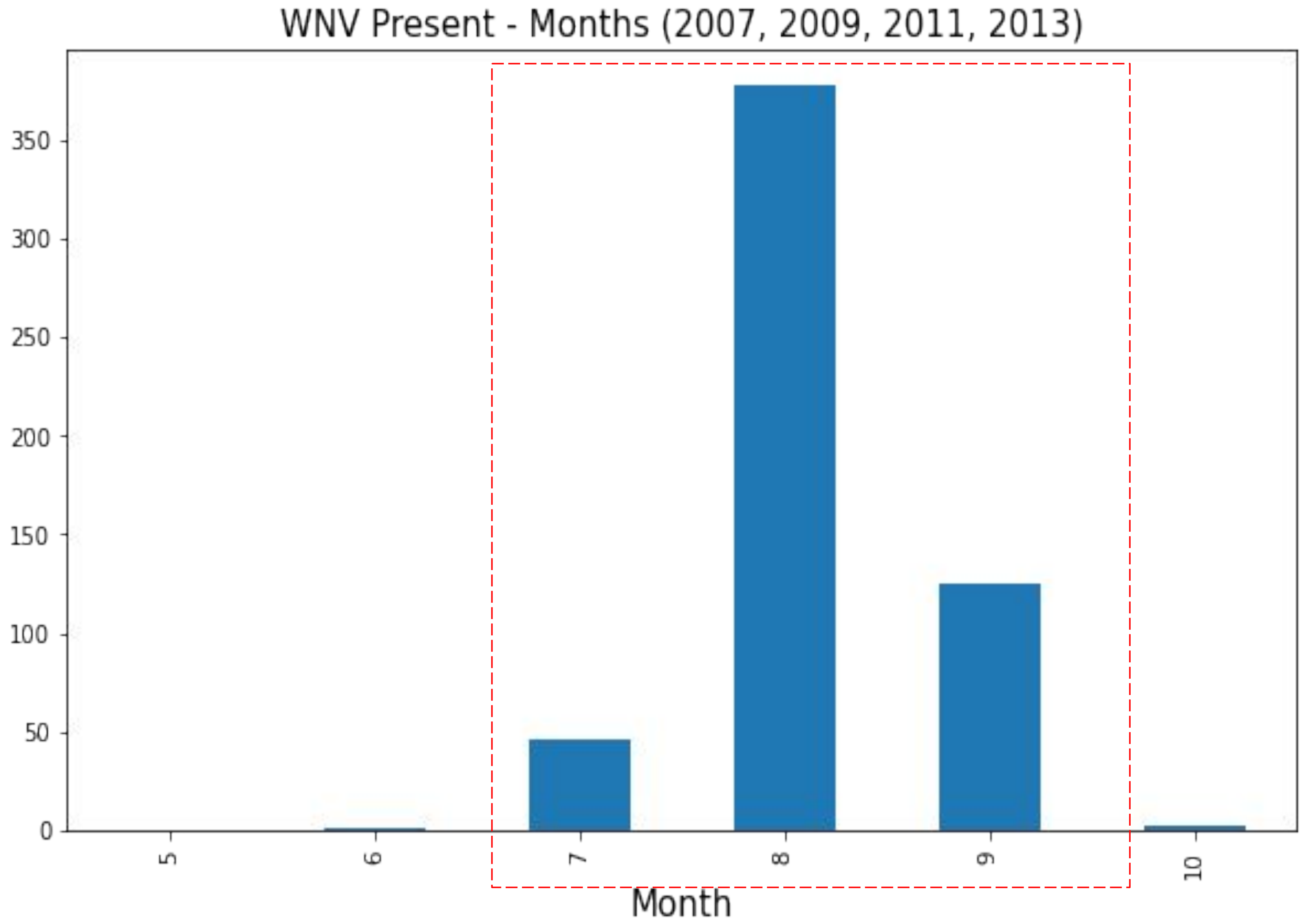
*Analysis is done only based on July to November in years: 2007, 2009, 2011, 2013 data that was collected.

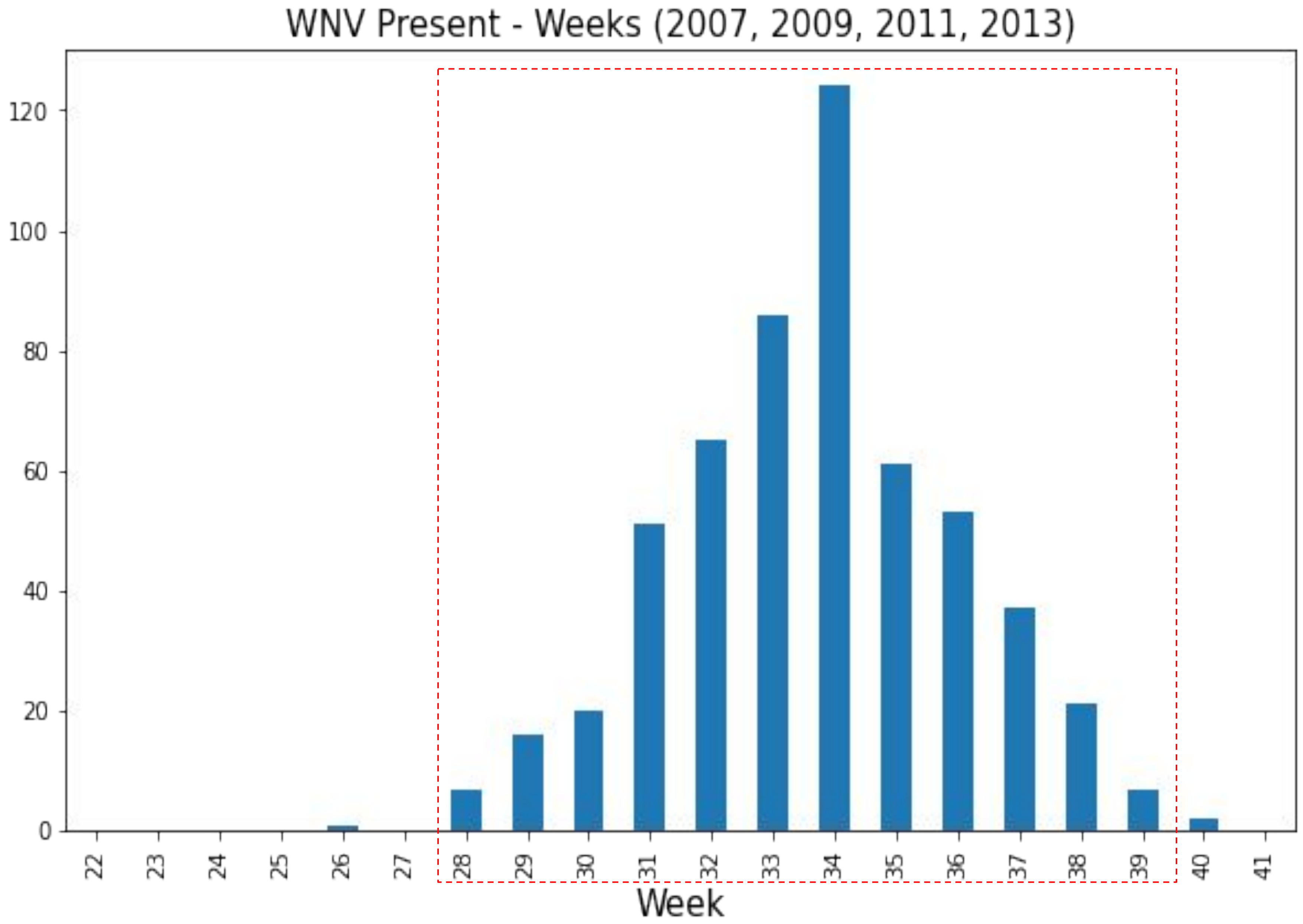
Mosquito Species



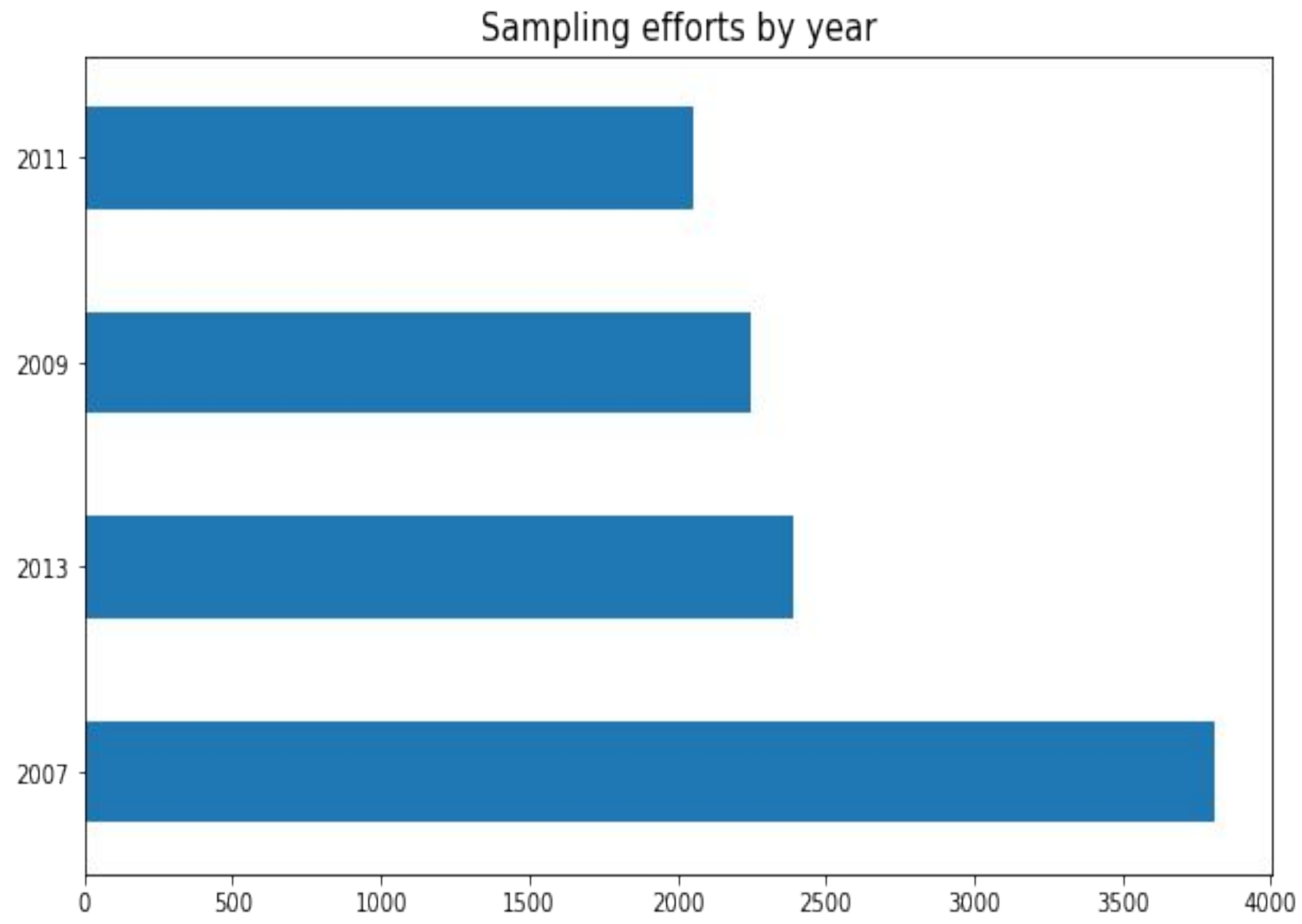


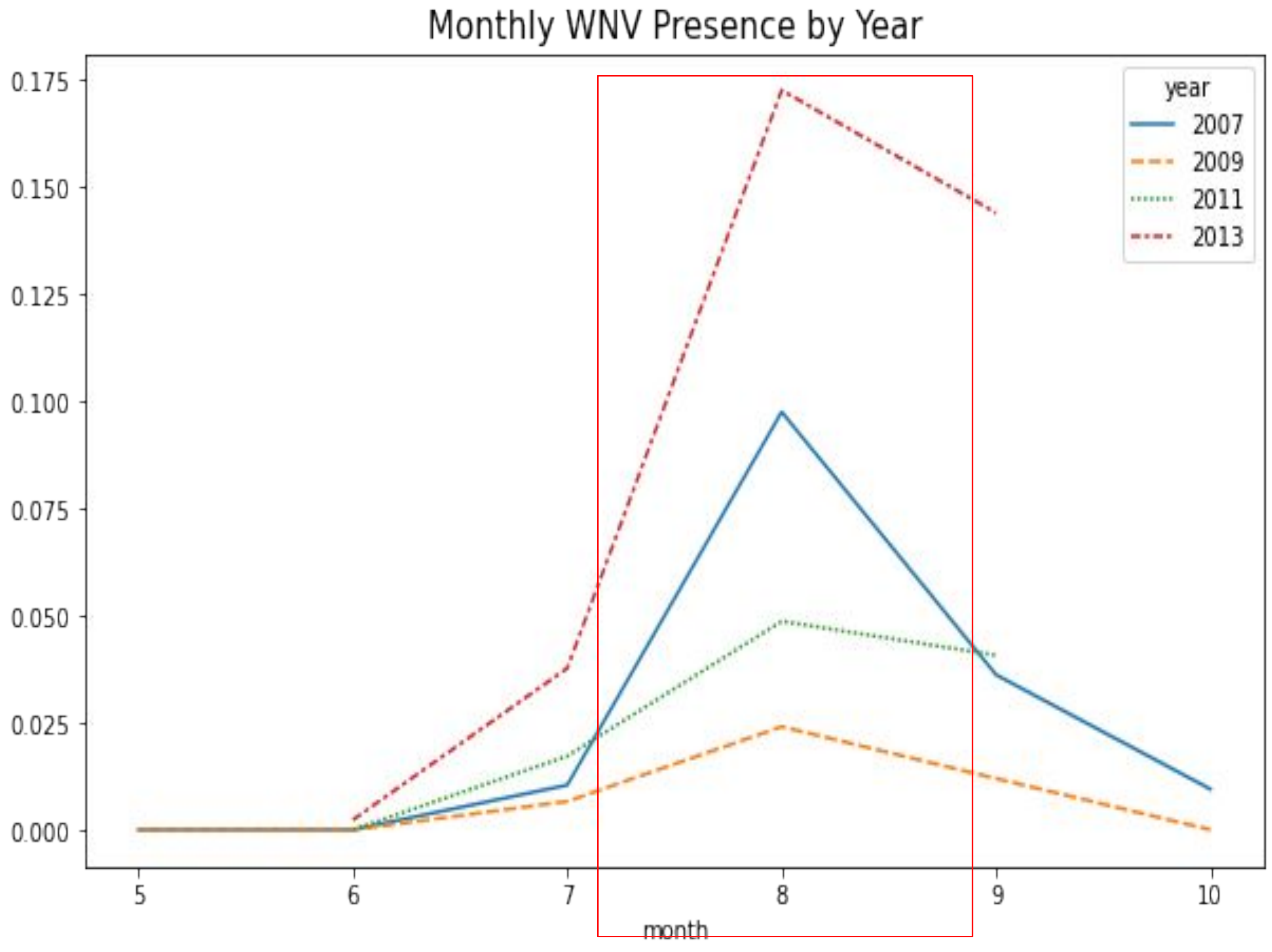
WNV over Time


WNV Present - Months (2007, 2009, 2011, 2013)

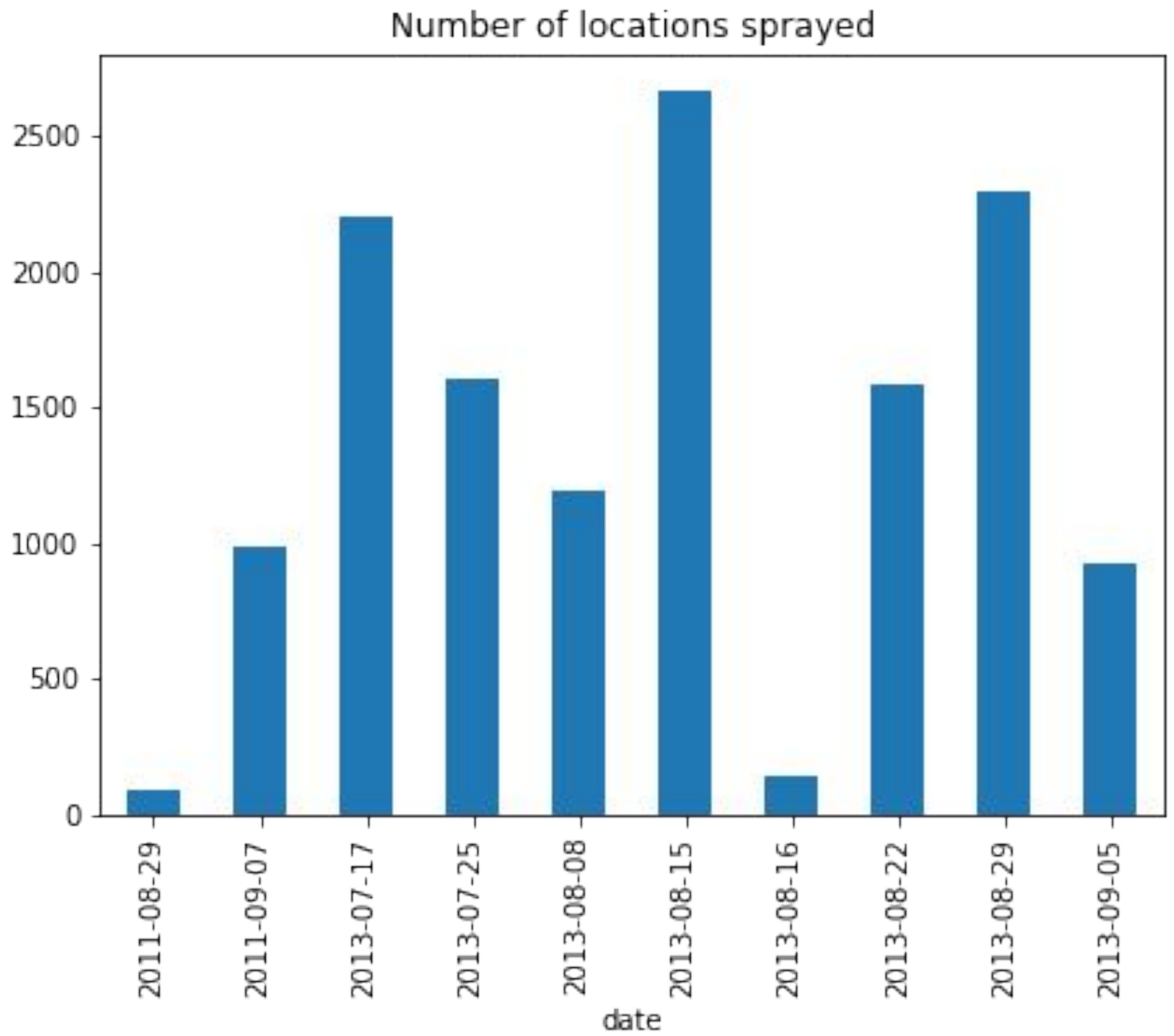

WNV Present - Weeks (2007, 2009, 2011, 2013)

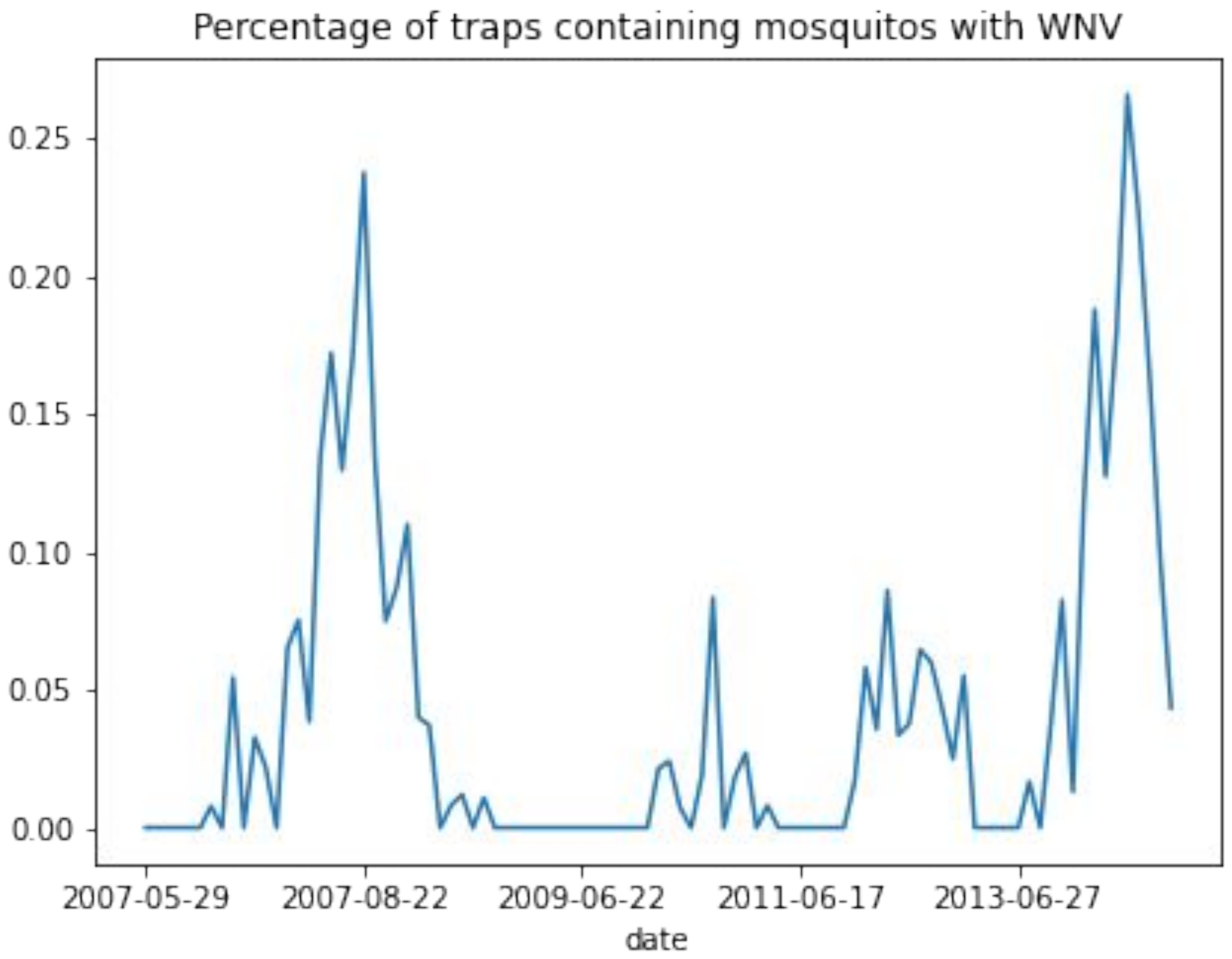
Trends


Sampling efforts by year
2011
2009
2013
2007
0
500
1000
1500
2000
2500
3000
3500
4000

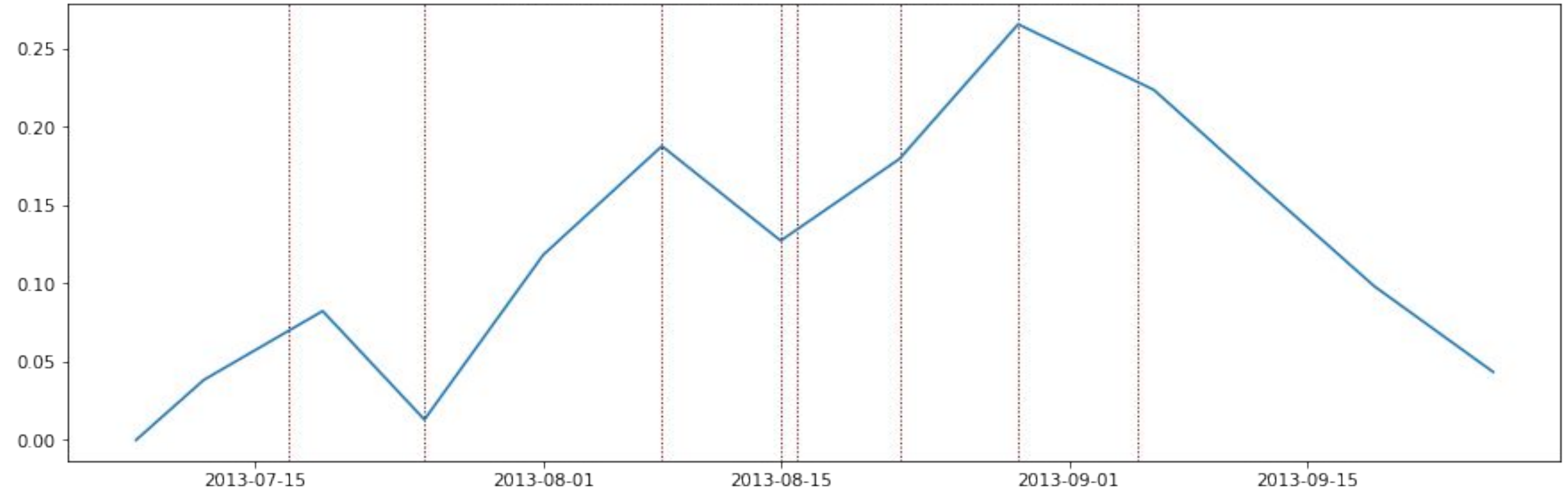

Monthly WNV Presence by Year
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
5
6
7
8
9
10
month
year
2007
2009
2011
2013

Spray & Trap locations

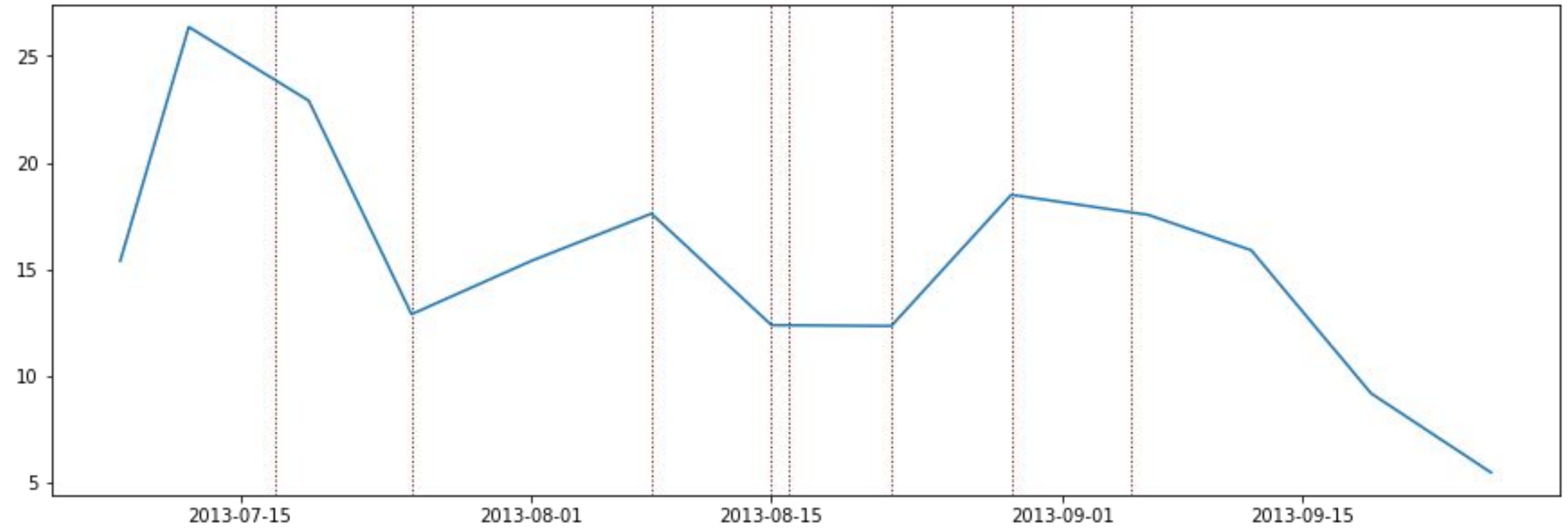




Spray Data

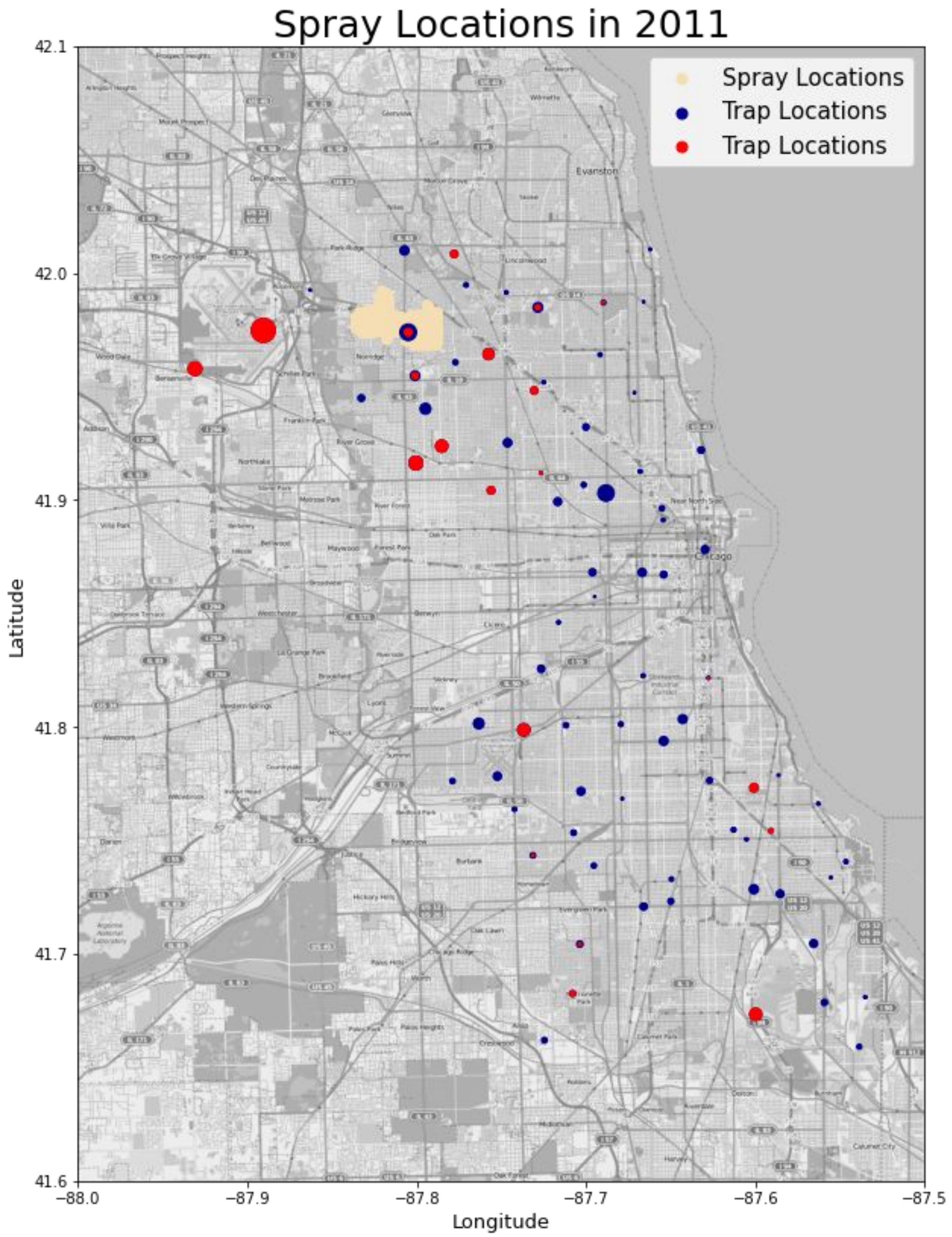


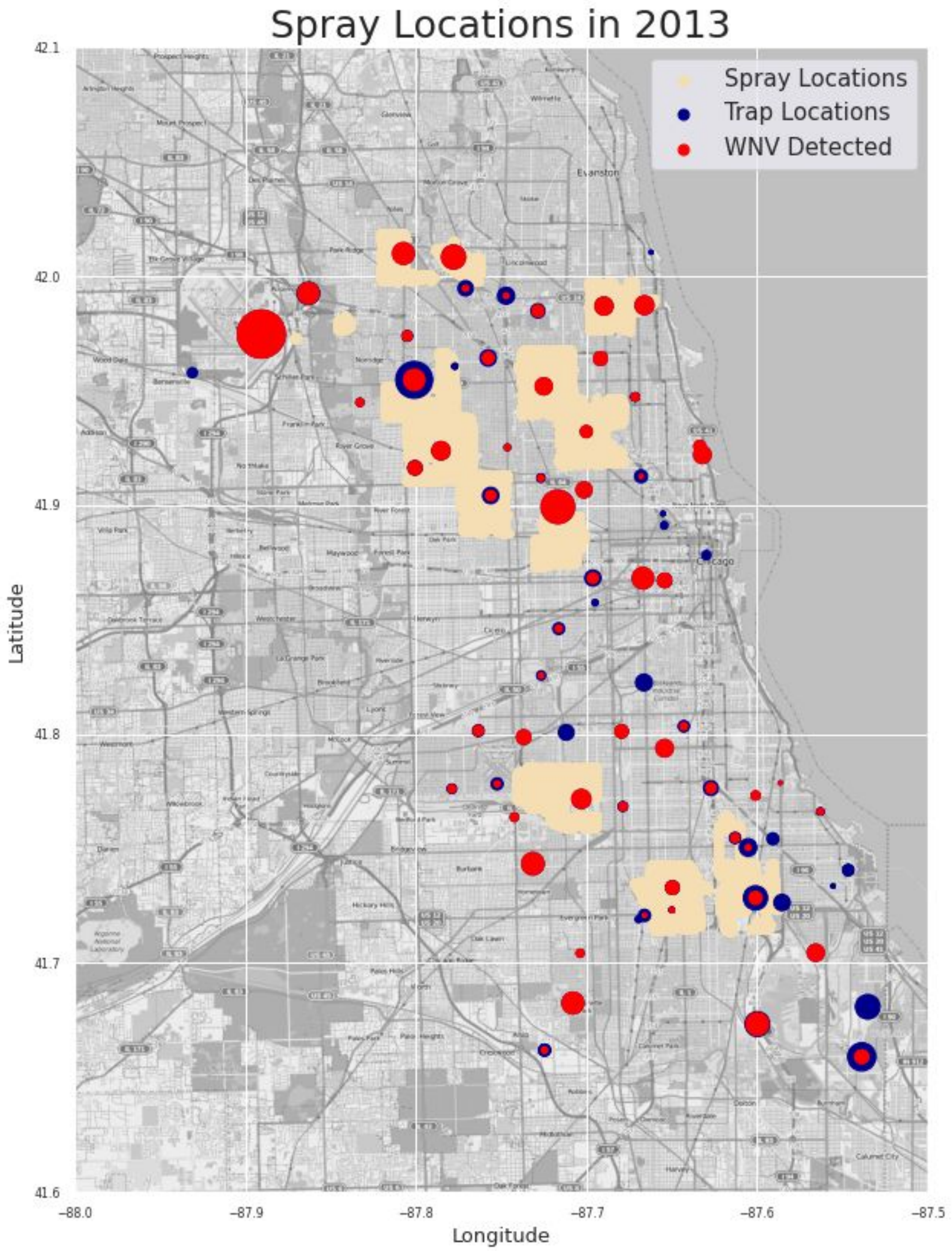
Percentage of traps containing mosquitos with WNV (2013)



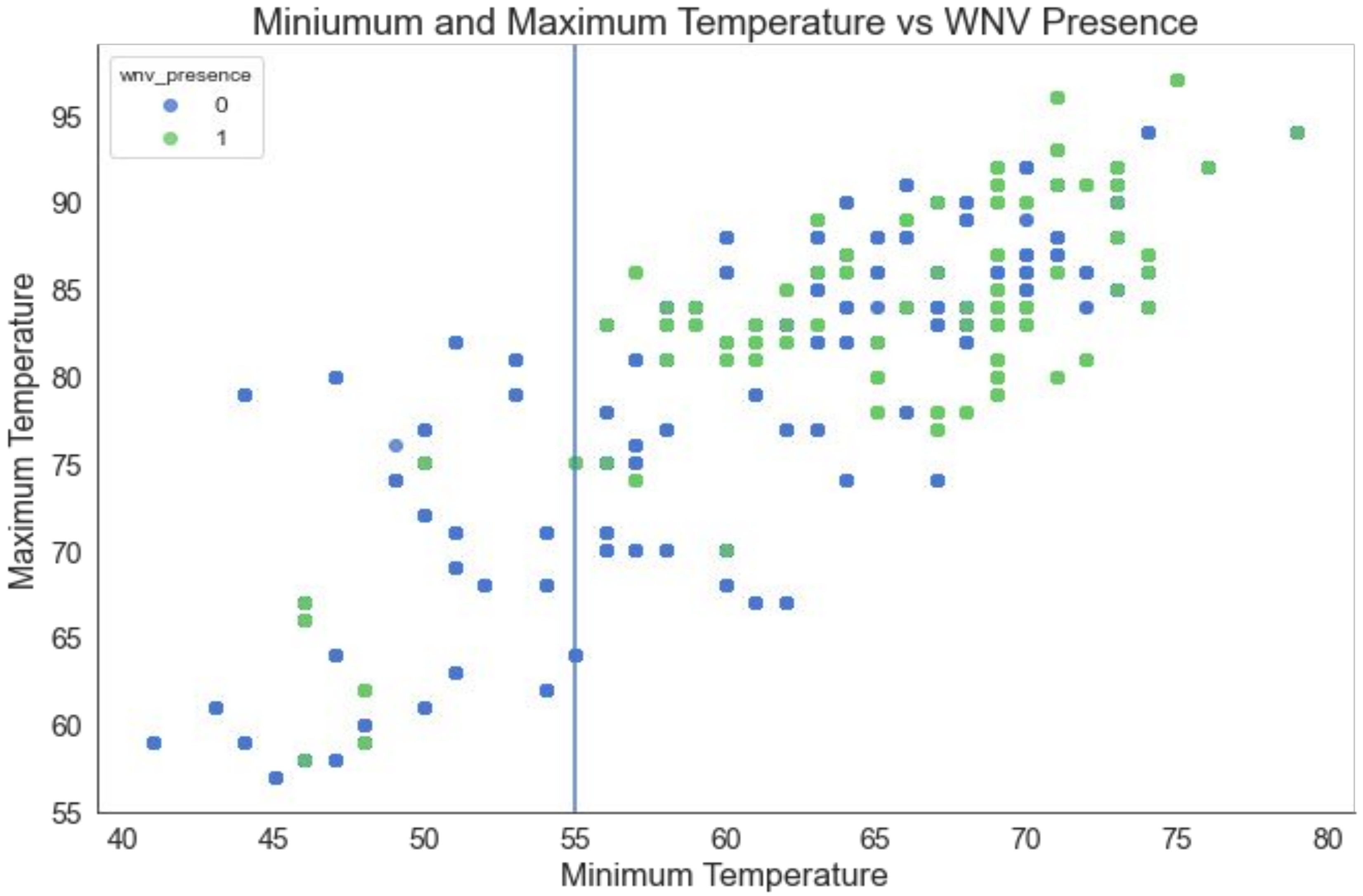
Average number of mosquitos per trap (2013)

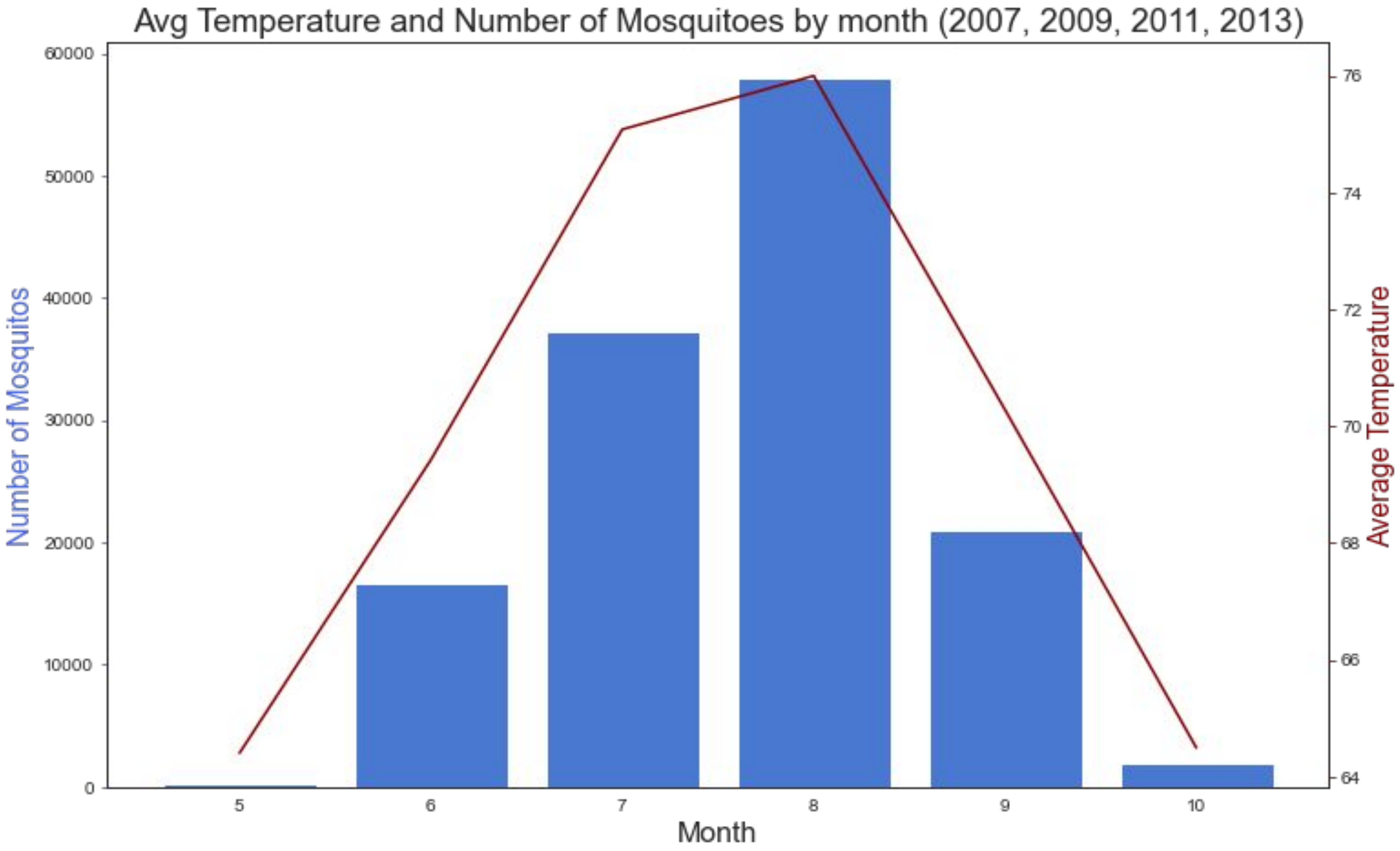
Spray & Trap locations


Spray Locations in 2011

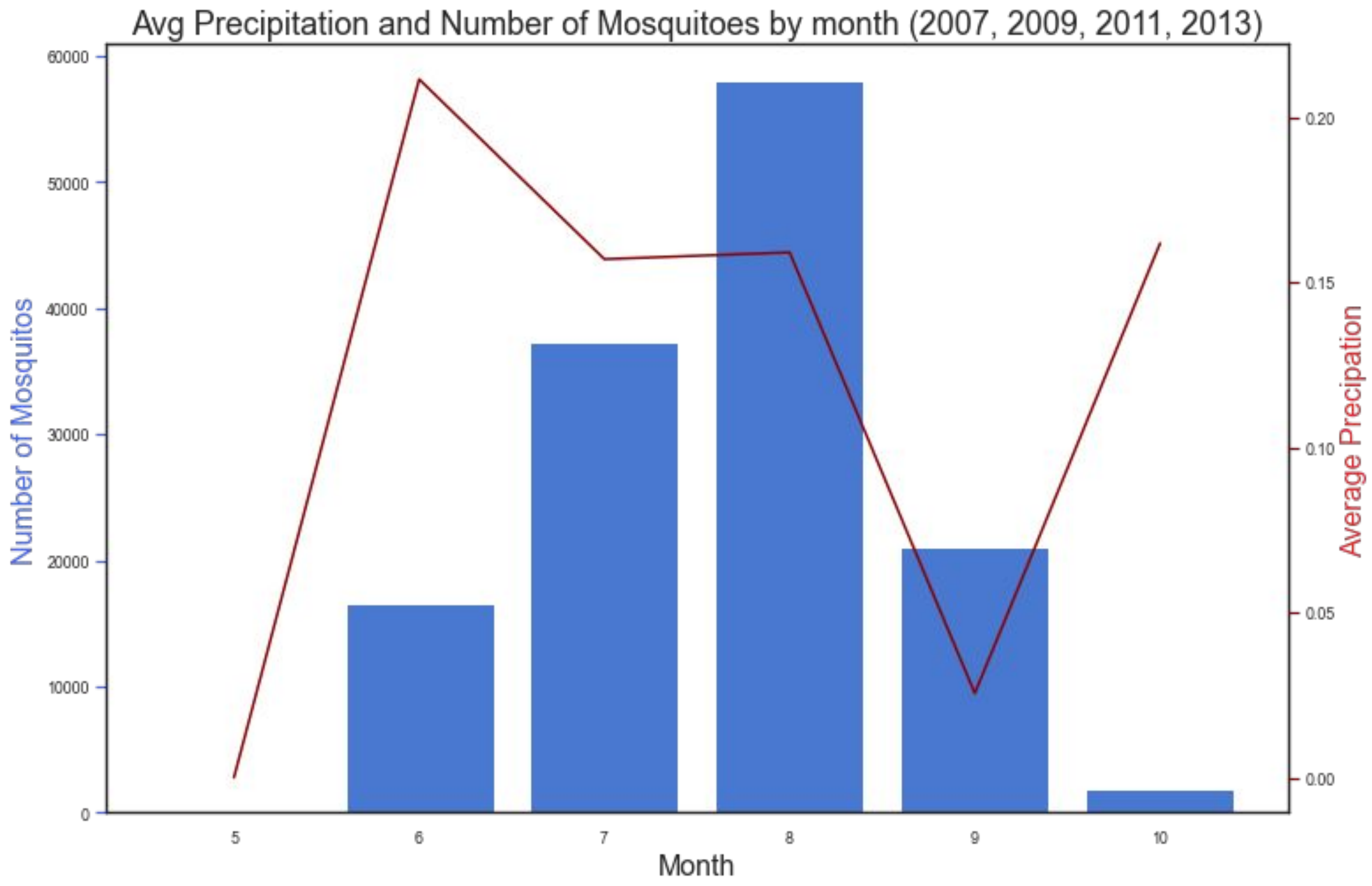

Spray Locations in 2013

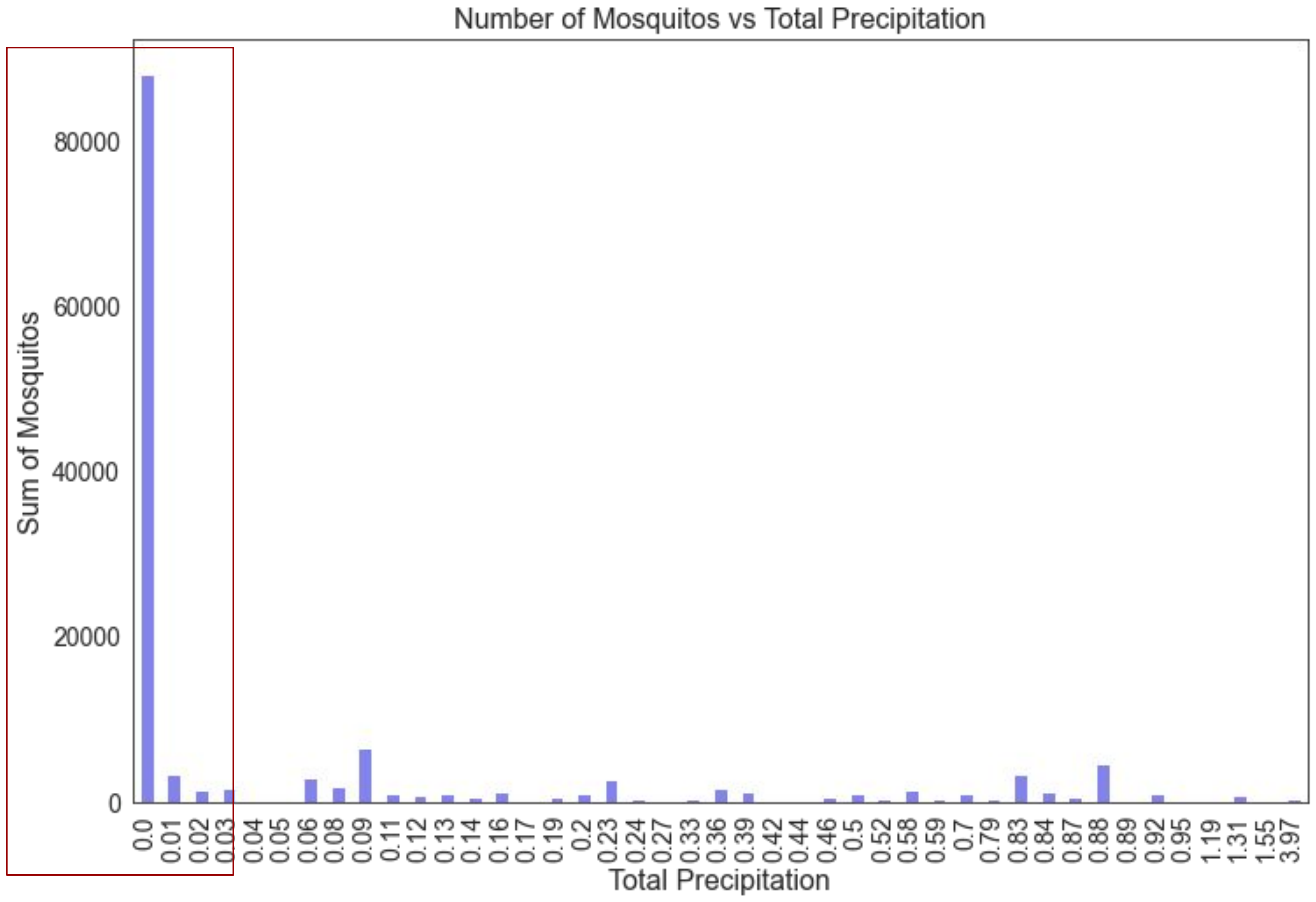
Weather Conditions (Temp)



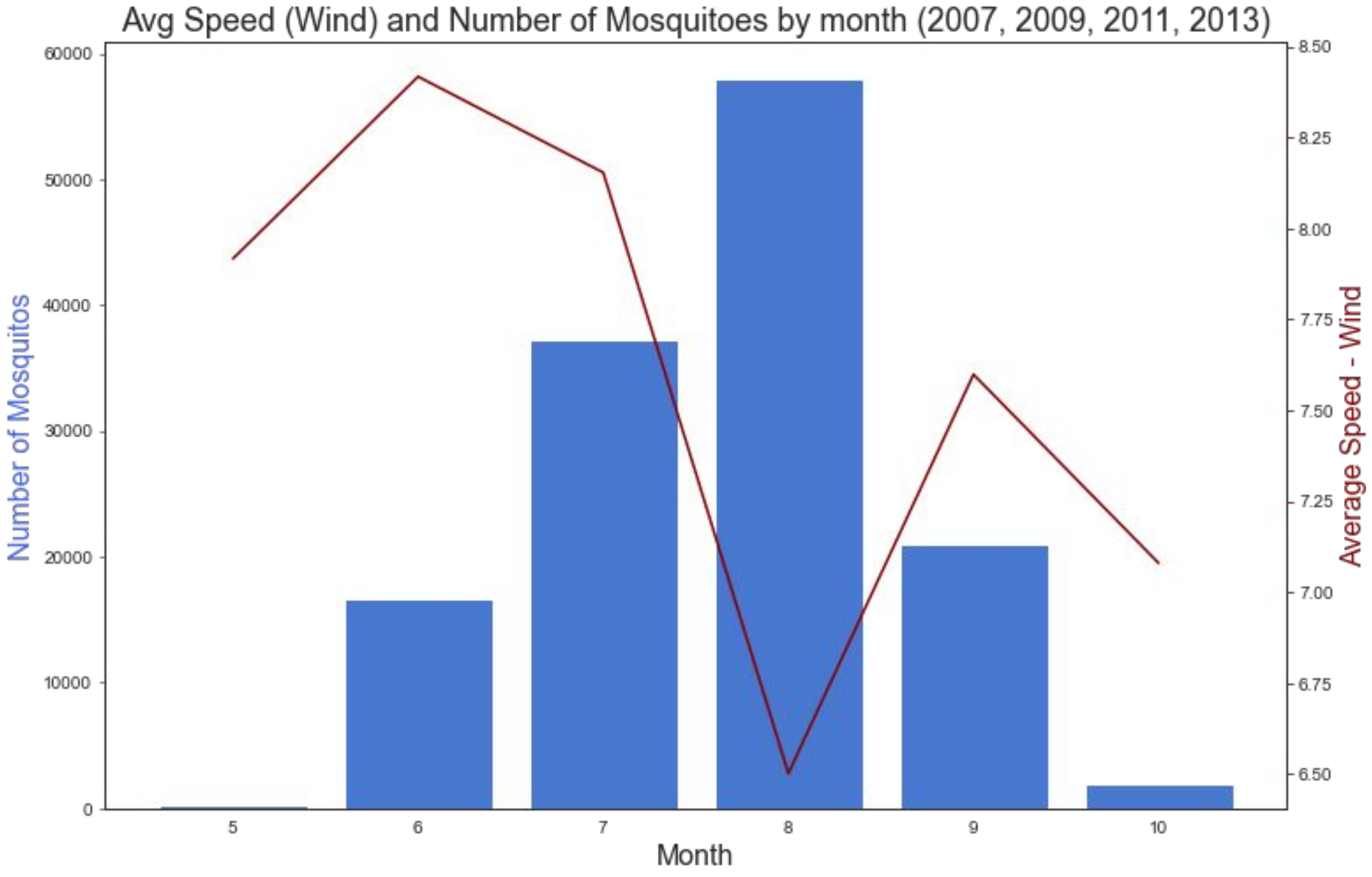


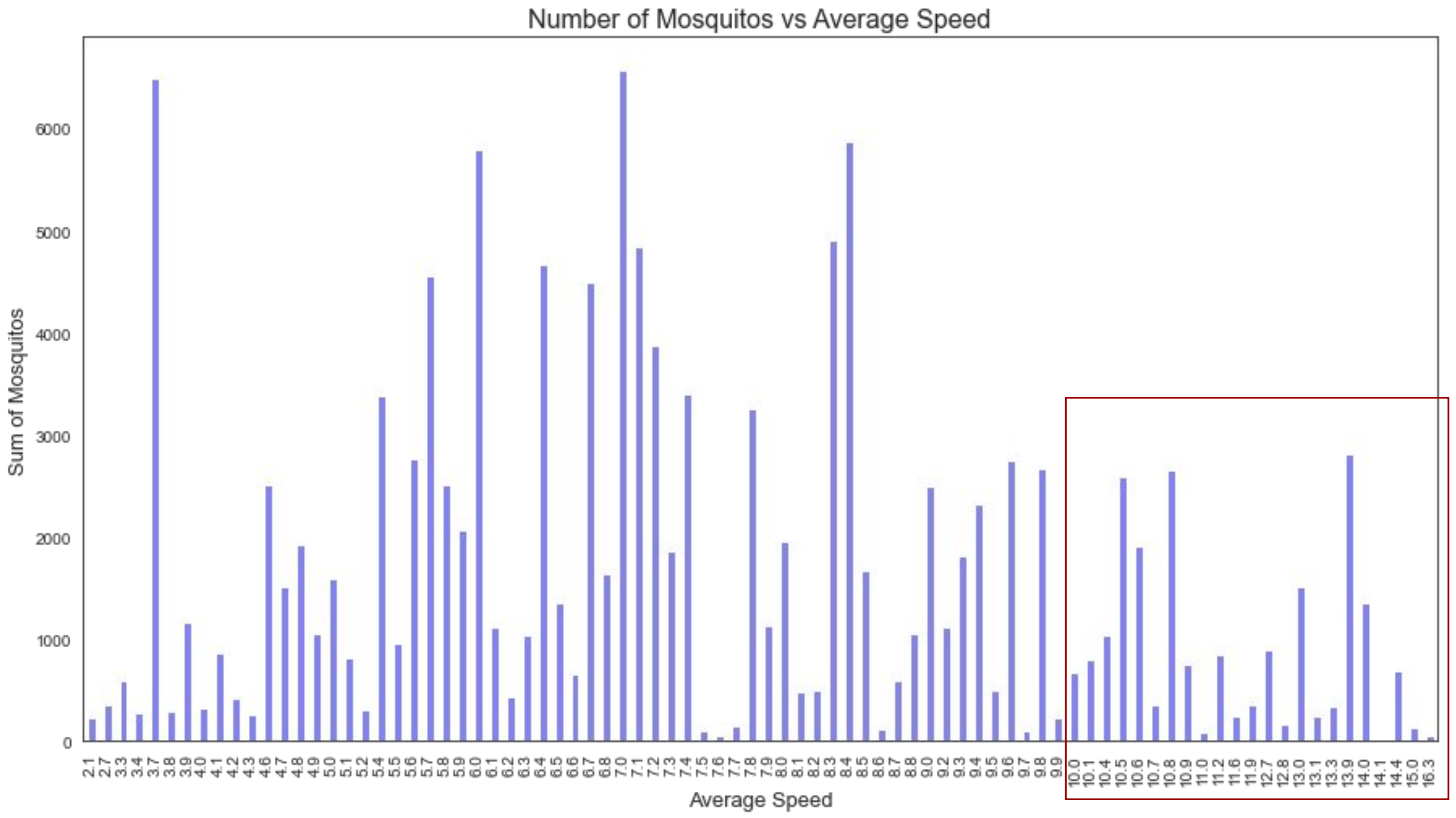
Weather Conditions (Precipitation)



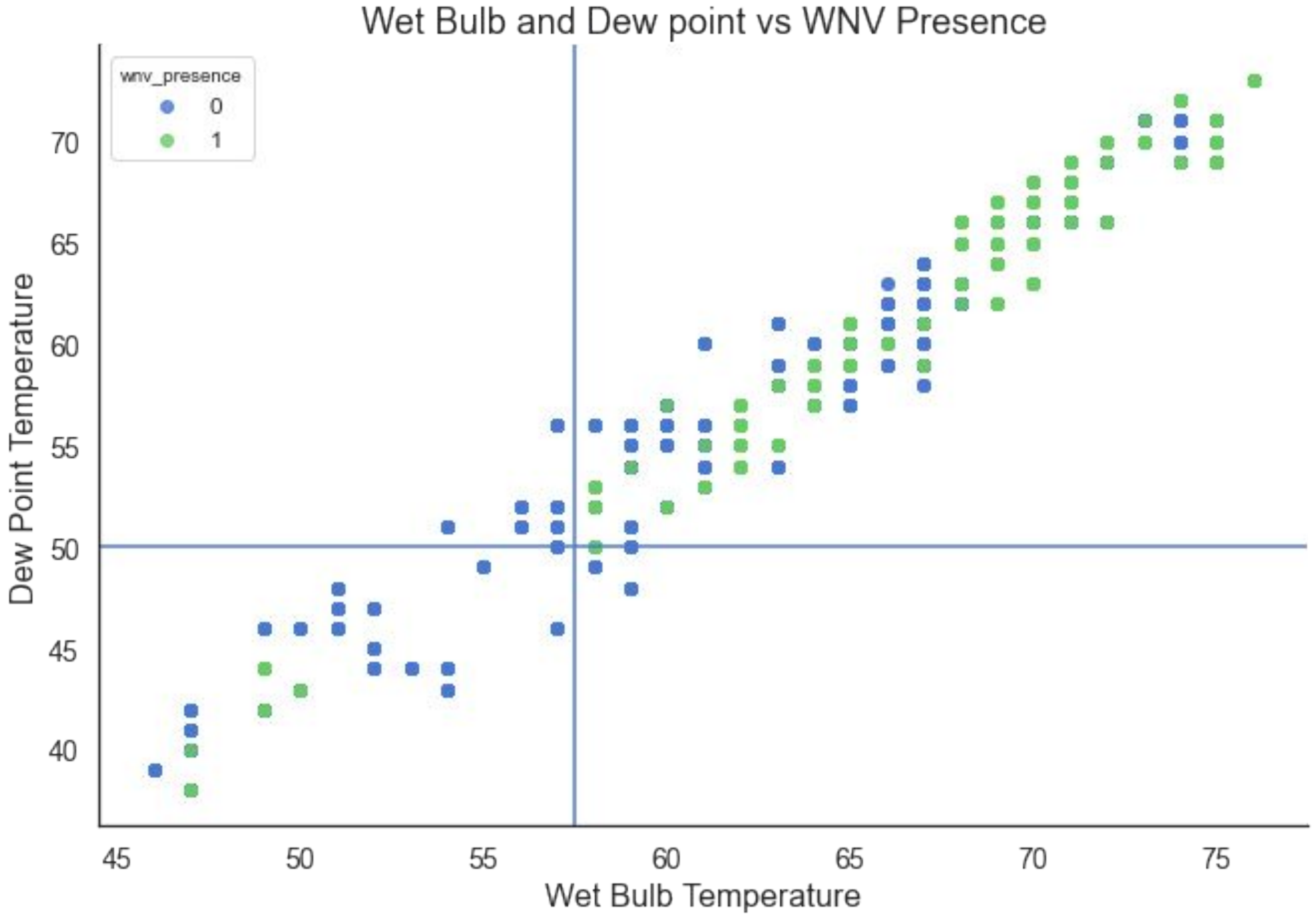


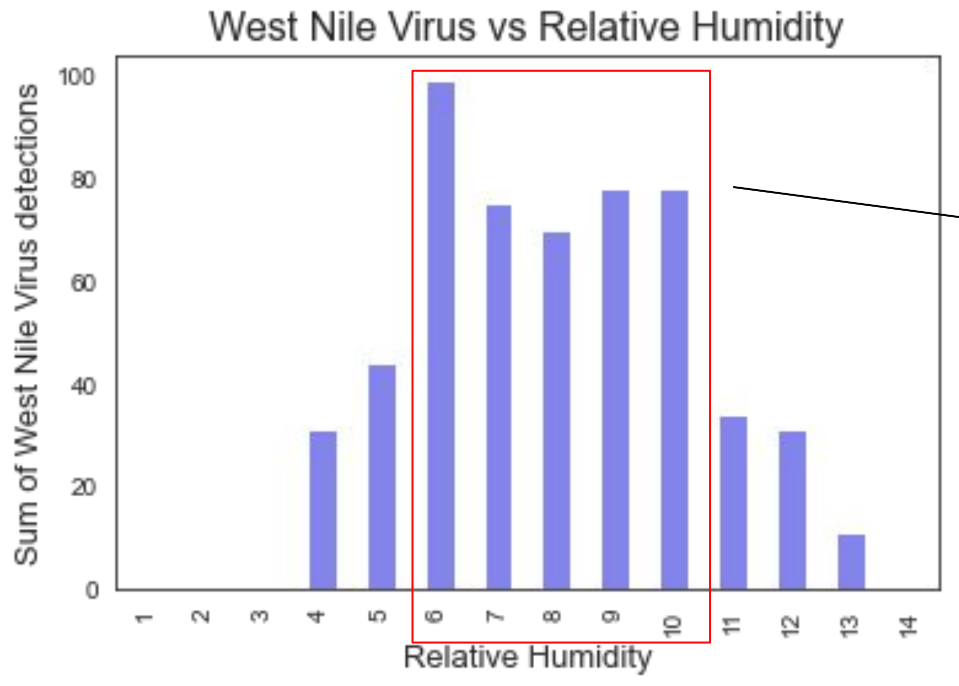
Weather Conditions (Wind)



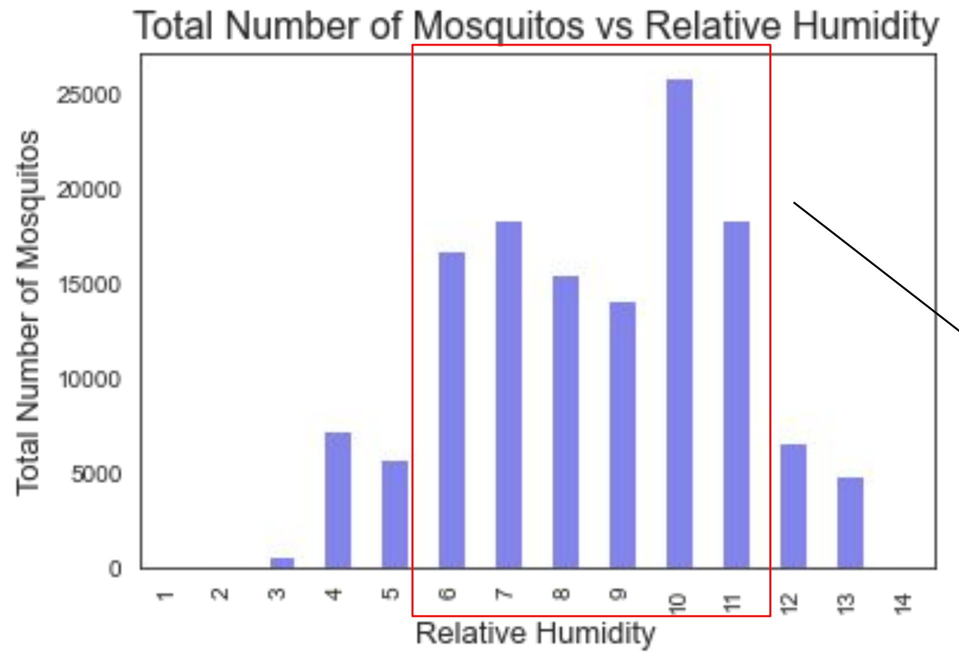


Weather Conditions (Humidity)





39% - 66%



33% - 66%

Relative Humidity Value

=

Tavg - WetBulb

Feature Engineering

Merged Weather, Train and Test data

Found Several Highly Correlated Features

Correlated Features will affect Model performance

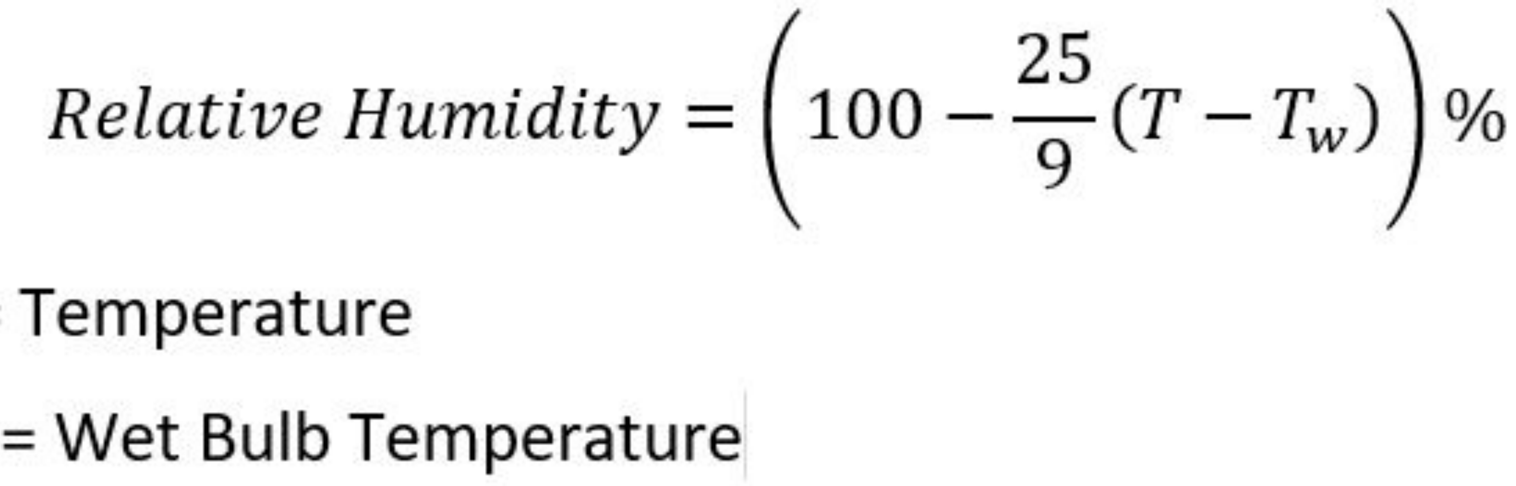
Feature Engineering: Humidity

Based on Magnus Approximation

Linear Ratio Formula Cross Verified with Thermodynamic[1]

Uses Average Temperature and Wet Bulb

Applied for both Train and Test Data



$$Relative\ Humidity = \left(100 - \frac{25}{9}(T - T_w)\right)\%$$

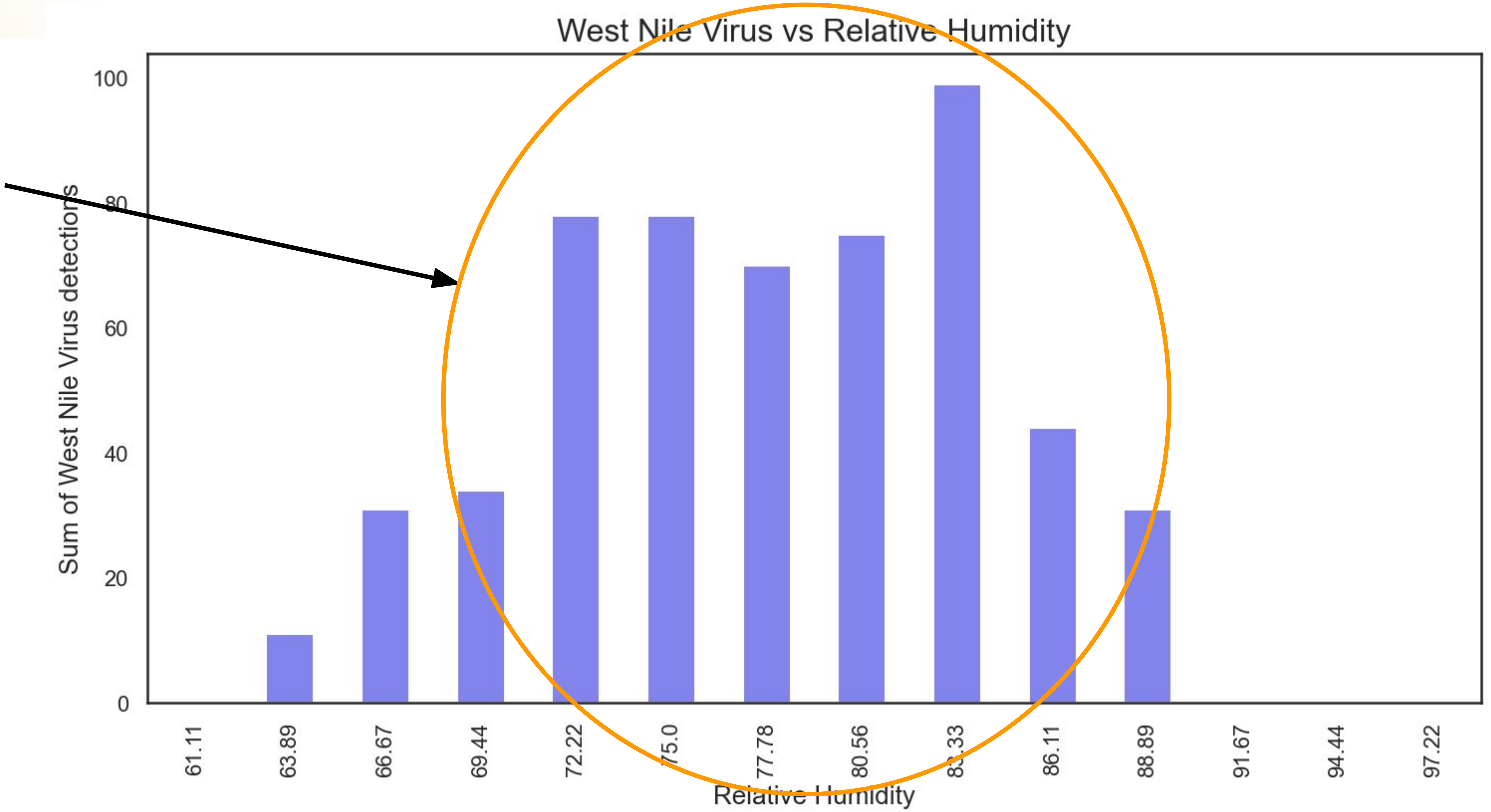
T = Temperature

T_w = Wet Bulb Temperature

[1] Çengel Yunus A., Boles, M. A., &amp; Kanoğlu Mehmet. (2016). *Thermodynamics: An engineering approach.* McGraw-Hill Education.

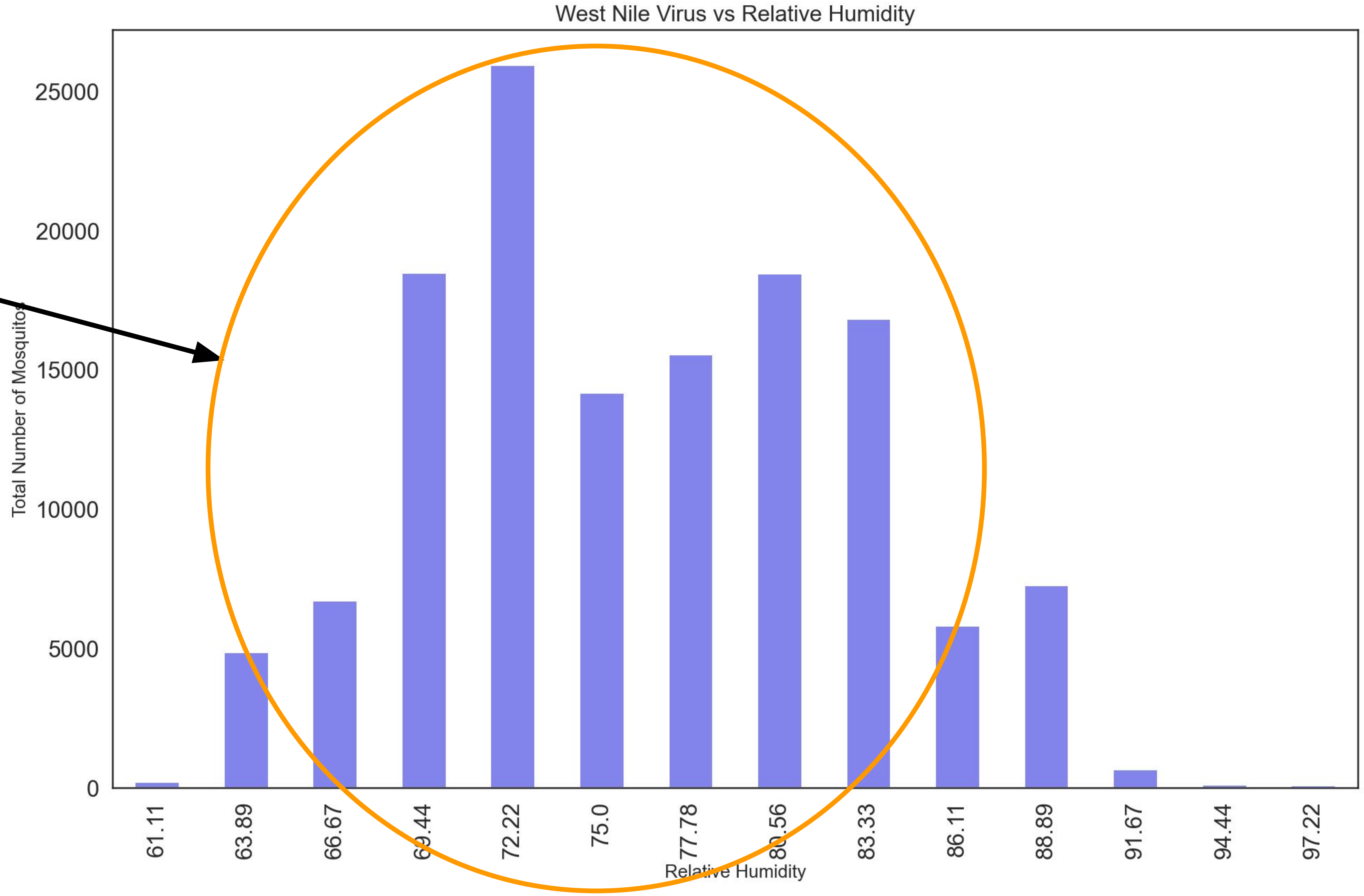
Detection of WNV against Humidity

WNV appears to thrive on about 66% to 89% Relative Humidity



Number of Mosquitoes vs Humidity

Mosquitoes appears to be prevalent between 64% to 83% relative humidity.



Other Correlated Features

Heatmap also pointed us to these strongly correlated features:

- Heatcool, Relative Humidity and Depart
- Result Speed and Average Speed
- Station Pressure and Sea Level
- Sunset and Sunrise

Interaction Features

Decided against dropping these data points to preserve model accuracy

Temperature, humidity and wind speed do play a role in their breeding

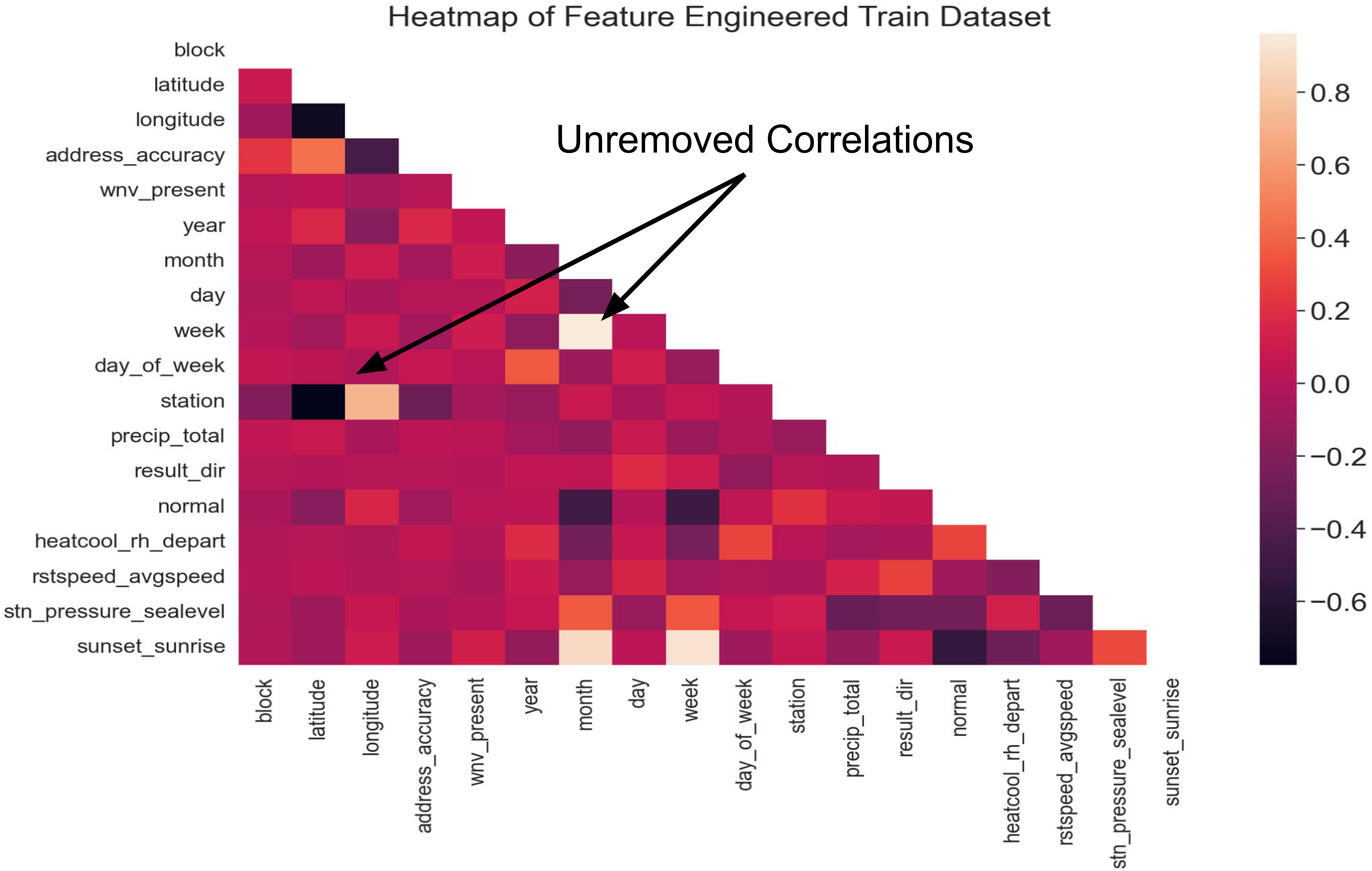
Multiplied each correlated data points together as interaction features

Heatmap after Feature Engineering

Models better

Less Data Noise

Increases Correlation to WNV presence



Unremoved Correlations

These Correlated are not removed or engineered:

- Station and Lattitude
- Week and Month

Station has a fixed location and will thus always have a correlation with location data

Week and Month are a function of each other

Label Encoding

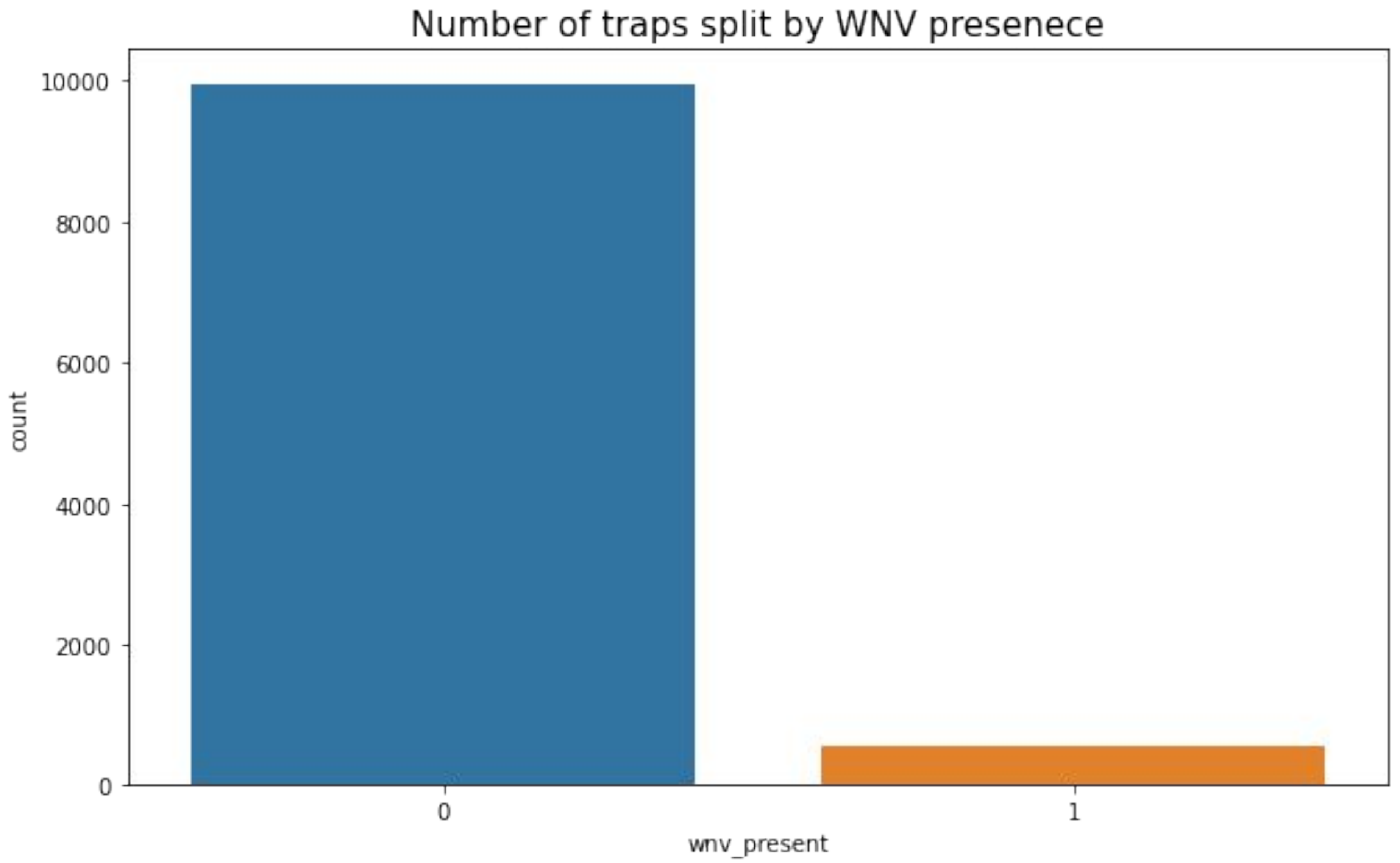
Species, Street and Traps are label-encoded

Allows us to categorise these data into numbers for modelling

Imbalanced Dataset

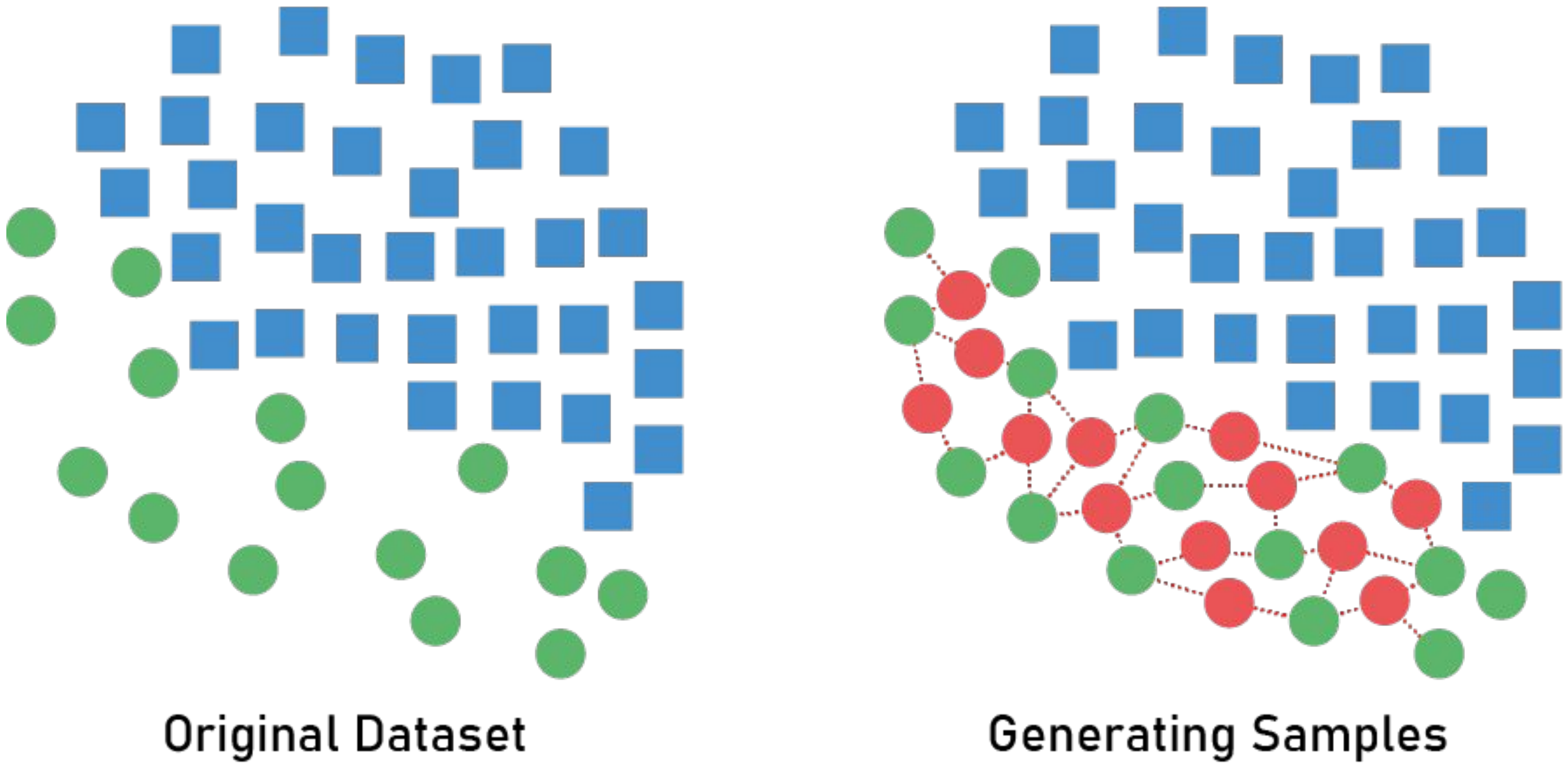
- Train dataset is imbalanced as

  >90% of the data is WNV absent



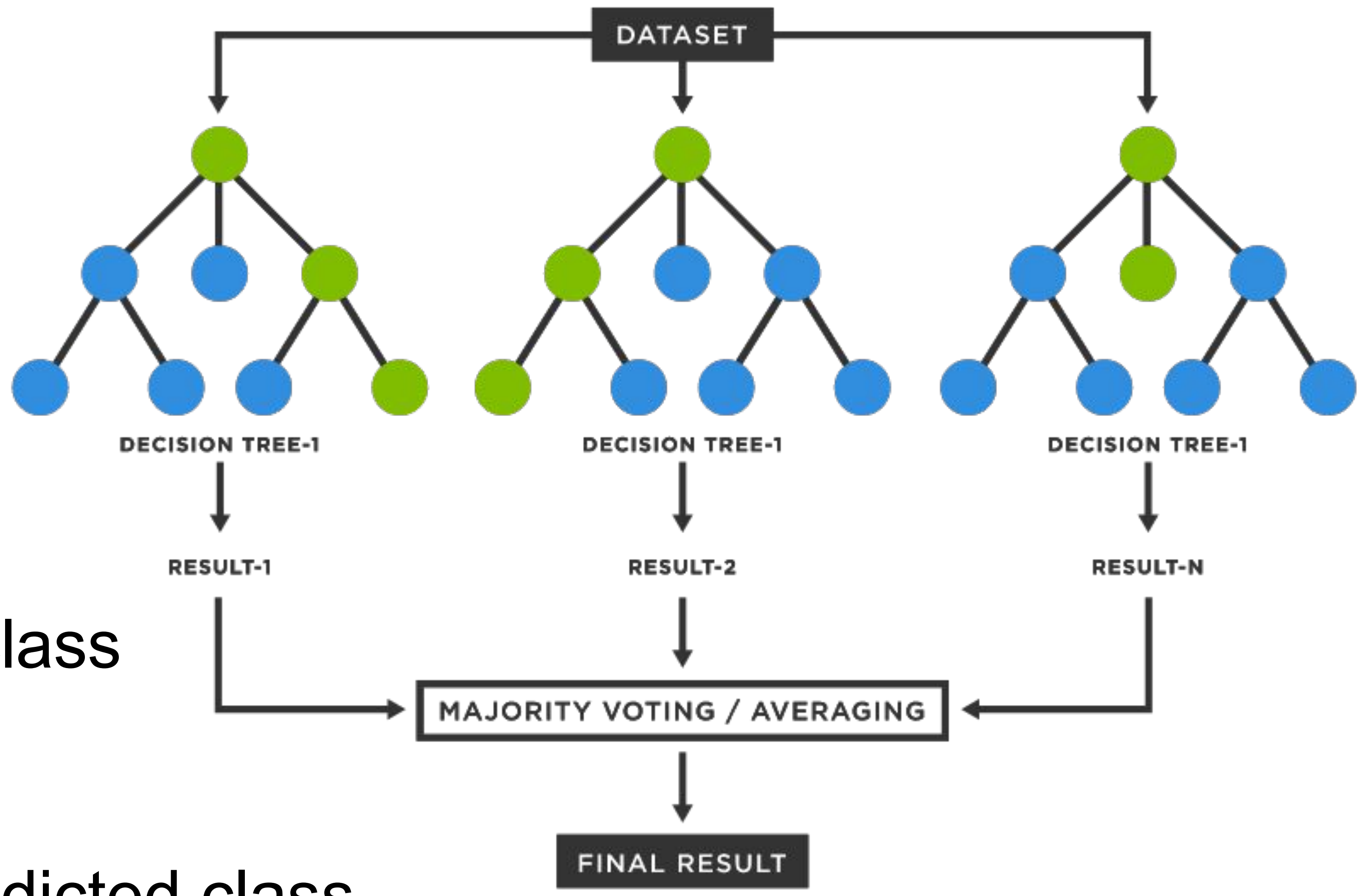
Modelling Workflow

- Data split into 80% train, 20% test
- Synthetic Minority Oversampling Technique (SMOTE)
- Coupled with a cross-validation and hyperparameter-tuning pipeline
- Results generation as per metrics



Models used

- **K-Nearest Neighbours** (baseline model)
  - Similar data points are grouped into the same class
- **Random Forest**
  - Ensemble of decision trees that vote for the predicted class
- **Support Vector Classifier**
  - Tries to divide the 2 classes of data with a hyperplane
- **Logistic Regression**
  - Fits data on a S-shaped curve to sort between two classes



DATASET

DECISION TREE-1

DECISION TREE-1

DECISION TREE-1

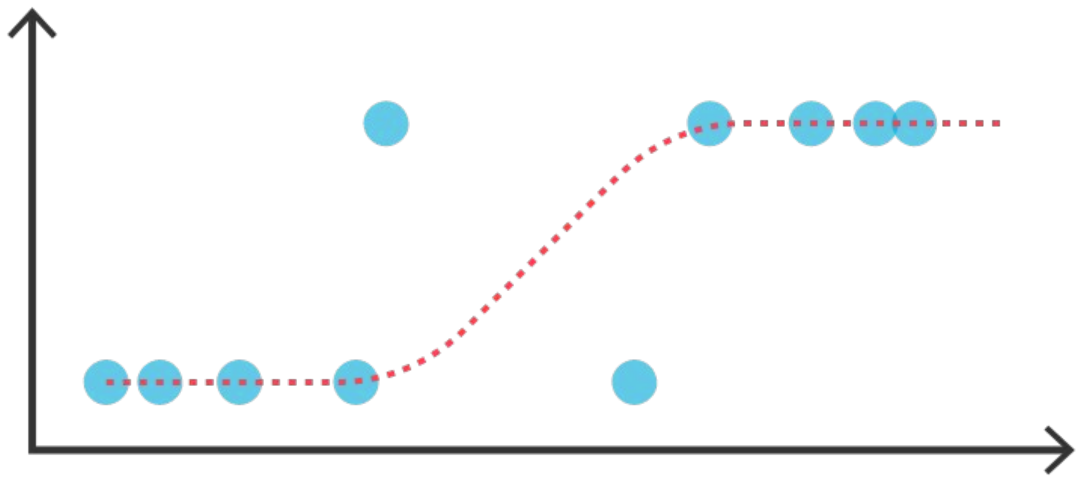
RESULT-1

RESULT-2

RESULT-N

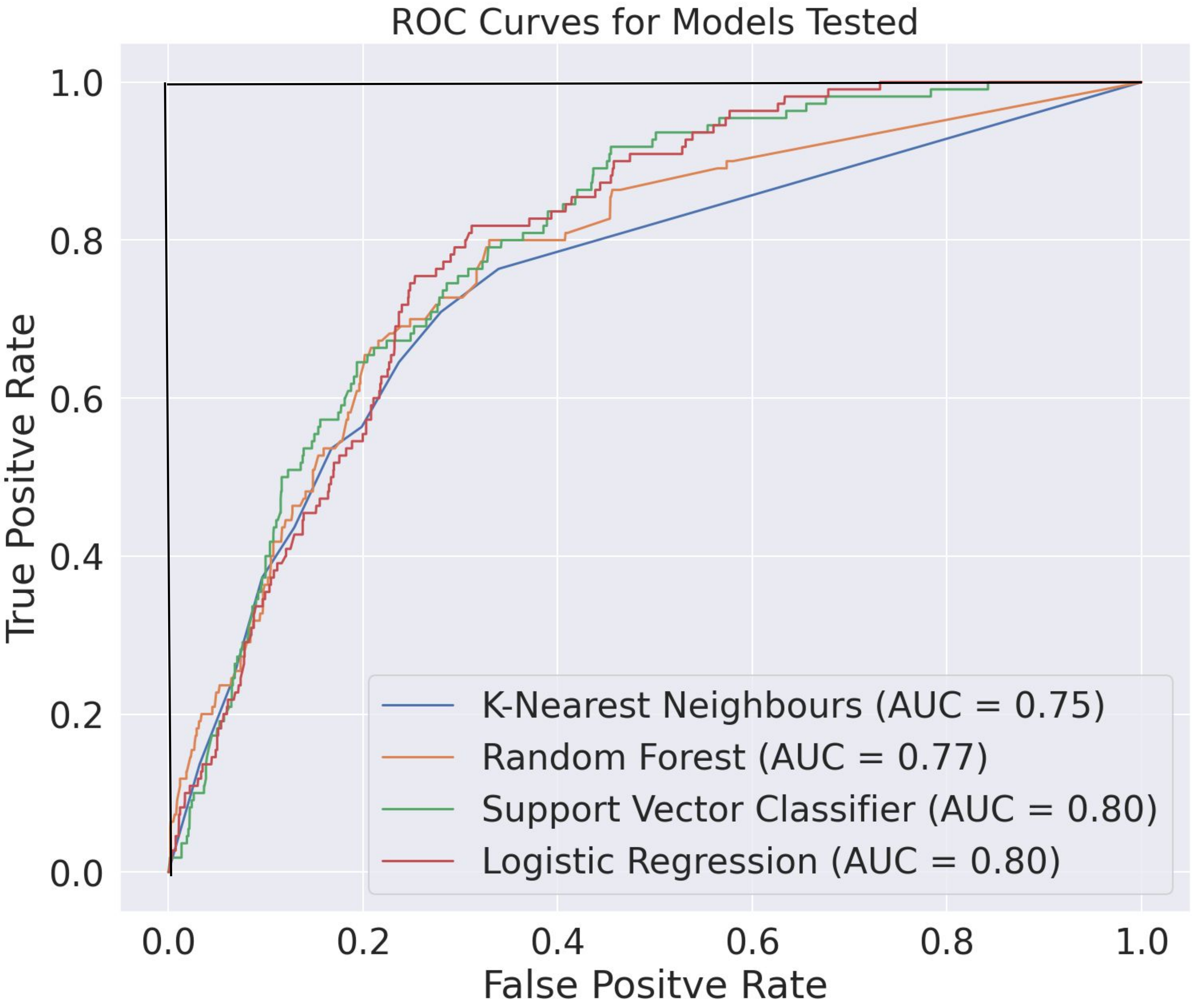
MAJORITY VOTING / AVERAGING

FINAL RESULT



ROC-AUC

- Receiver Operating Characteristic (ROC) curve
- Area Under Curve (AUC)
- AUC of 0.8 means there is an 80% chance that the model can distinguish between the two classes

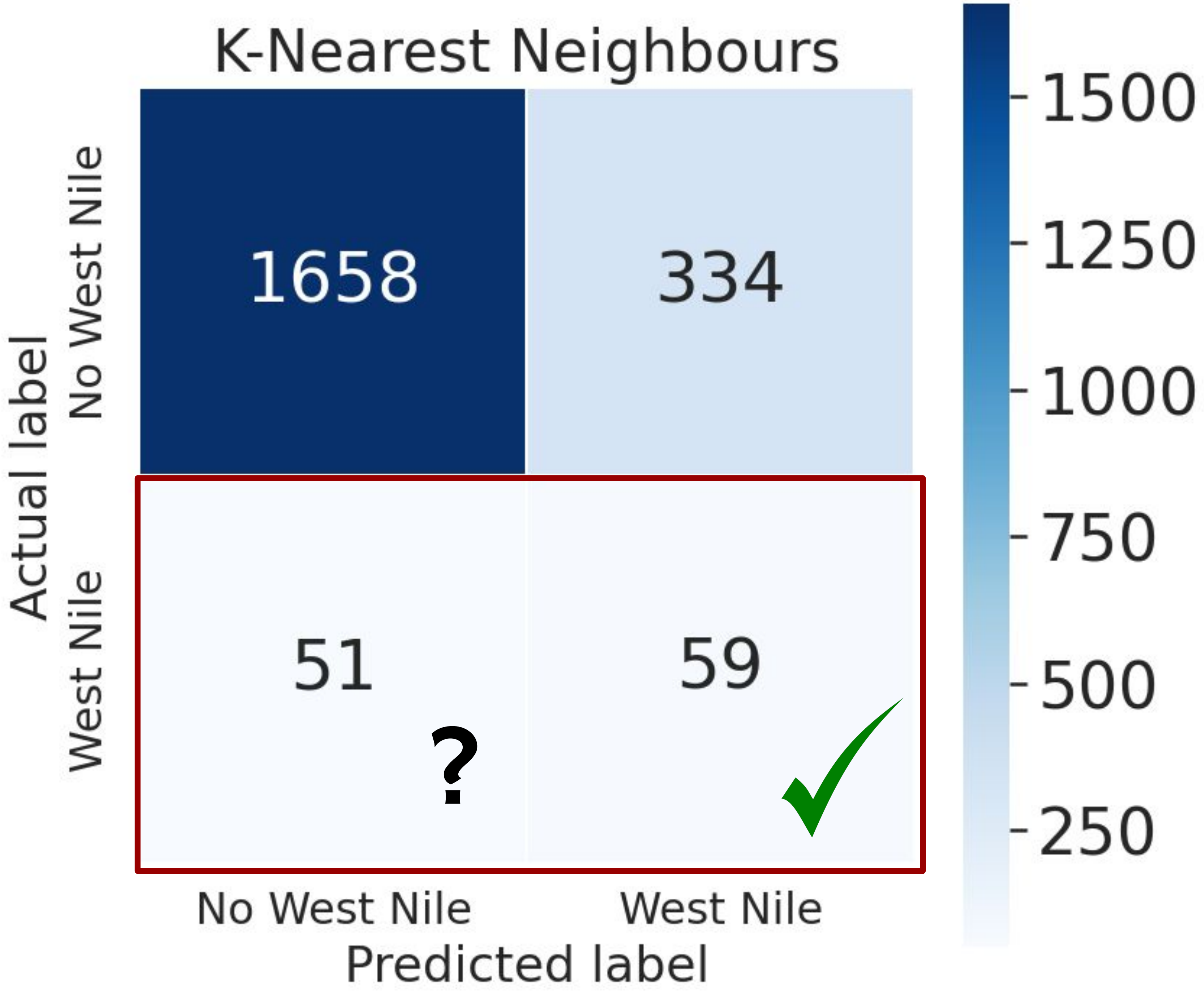


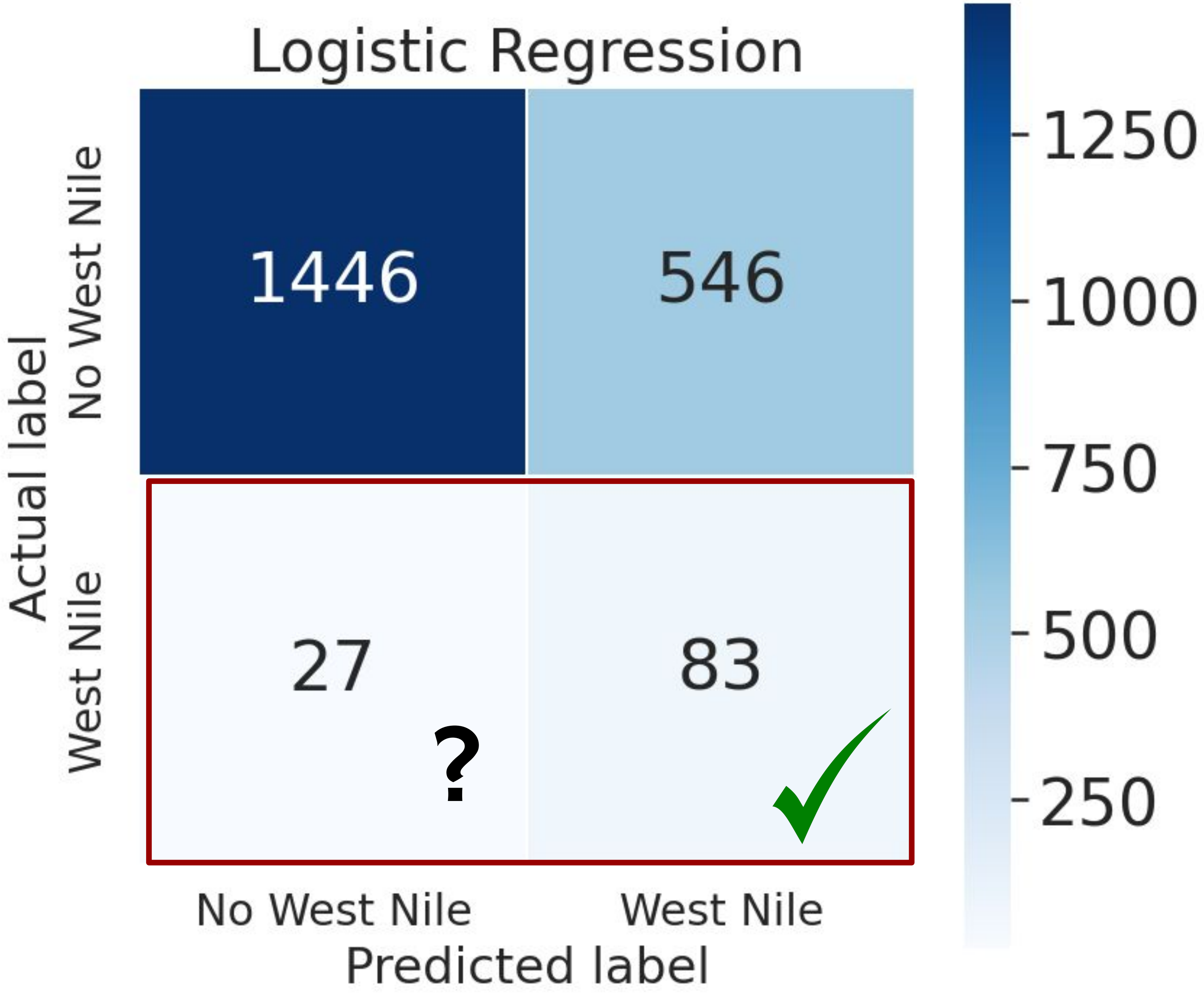
Model Results

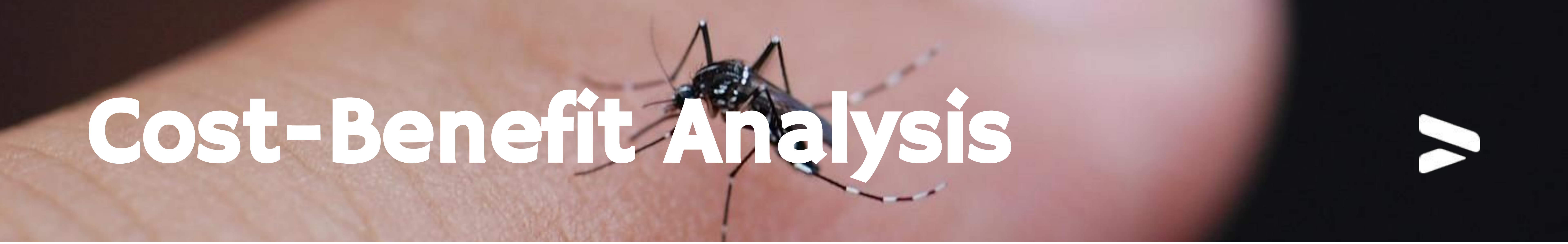
- ROC AUC is our primary metric, followed by recall **(Identified positives / All positives)**
- Logistic Regression performs well on both train and test for AUC, with a high recall

Classifier	Cross-validated Training ROC AUC Score	Testing ROC AUC score	Testing recall score	Testing accuracy score
Logistic Regression	0.80	0.80	0.75	0.73
Random Forest	0.79	0.77	0.25	0.90
Support Vector Classifier	0.81	0.80	0.55	0.83
K- Nearest Neighbors	0.78	0.75	0.54	0.82

Confusion Matrix







Cost-Benefit Analysis

Cost

Cost of pesticides

Benefit

Reduced medical costs

Reduced productivity costs

Reduced human suffering

Goal

Reduce cost-benefit ratio by targeting at-risk areas more effectively

Costs

Cost per acre

USD 0.92 per acre

X

Number of weeks

14 weeks over 3 months

X

Number of acres

149,800 acres

$2,172,842

Costs

Cost per acre

USD 0.92 per acre

x **Number of weeks**

14 weeks over 3 months

x **Number of acres**

149,800 acres

$2,172,842

Benefits

Mild effects

1 in 5 get mild effects

x **Medical cost**

$302 per patient

+ **Productivity loss**

$790 per worker

Severe effects

1 in 150 get severe effects

x **Medical cost**

$39,460 per patient

+ **Productivity loss**

$9150 per worker

$19,895 per patient

If 100 people get sick, the benefits would outweigh the costs

Key Insights

TIME OF YEAR

- Mosquito season peaks from July to September

WEATHER

- 50 - 80° F temperature
- 64 - 83 relative humidity
- Spike in mosquitos a period of time after rain

-

LOCATIONS

- Current spray efforts seem to have limited effect on containing outbreak
- Spray efforts not targeted to problem areas

Recommendations

TIME OF YEAR

- Monitor closely during July to end of September

WEATHER

- Weather forecasts should be used to direct spraying

LOCATIONS

- Our prediction model should be used to guide future spray campaigns

AN INTEGRATED SOLUTION

Develop a front-end application using our logistic regression model for scientists and biologists to gauge WNV probability when collecting mosquito samples.

Future Steps

- **More accurate data on weather should be gathered**
  - More localized weather data would improve model fit and prediction

- **Measure efficacy of other mosquito control methods**
  - Removing breeding habitats
  - Constructing structural barriers
  - Controlling mosquitos at the larval stage
  - Controlling adult mosquitos

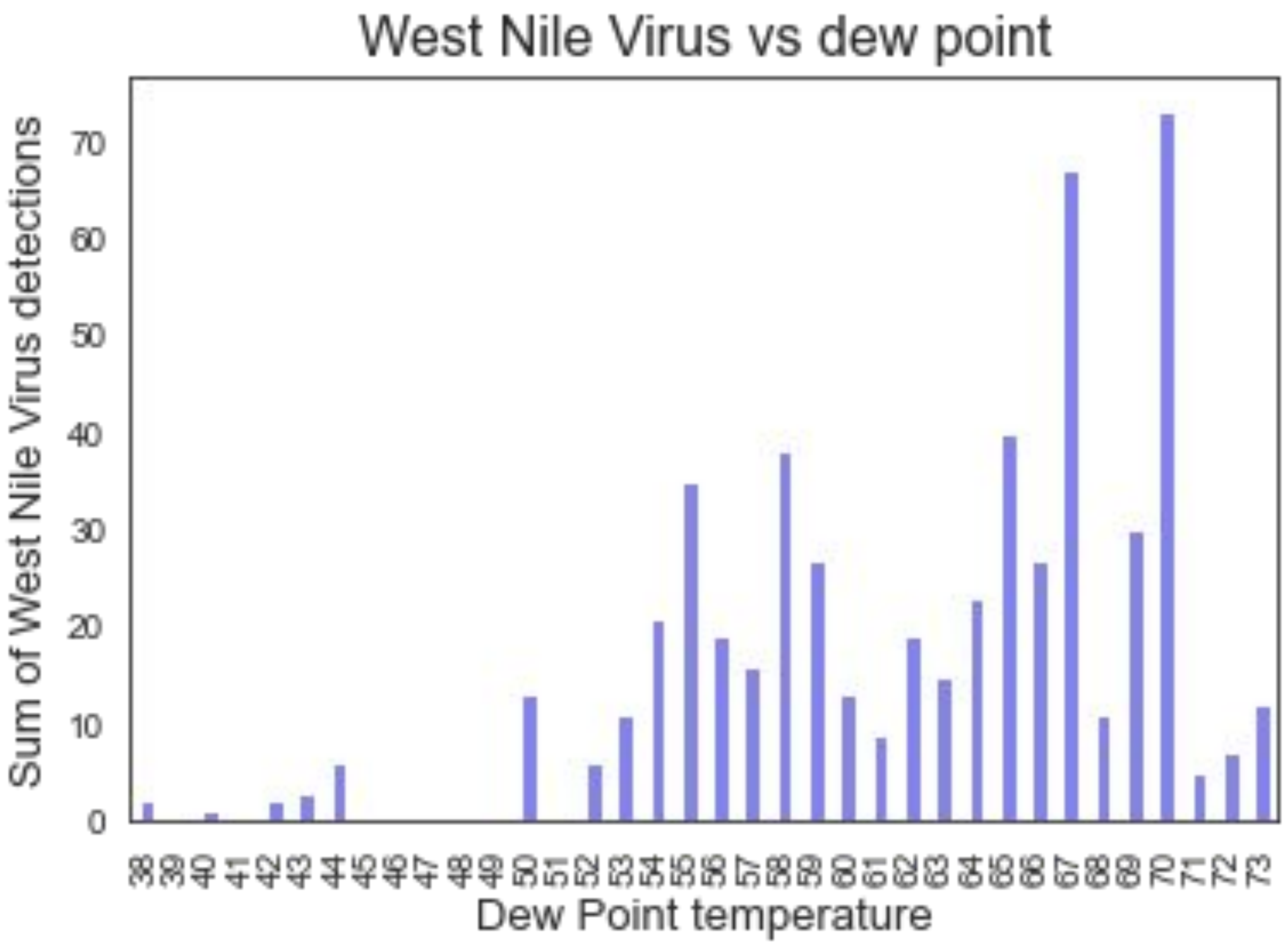
THANK YOU

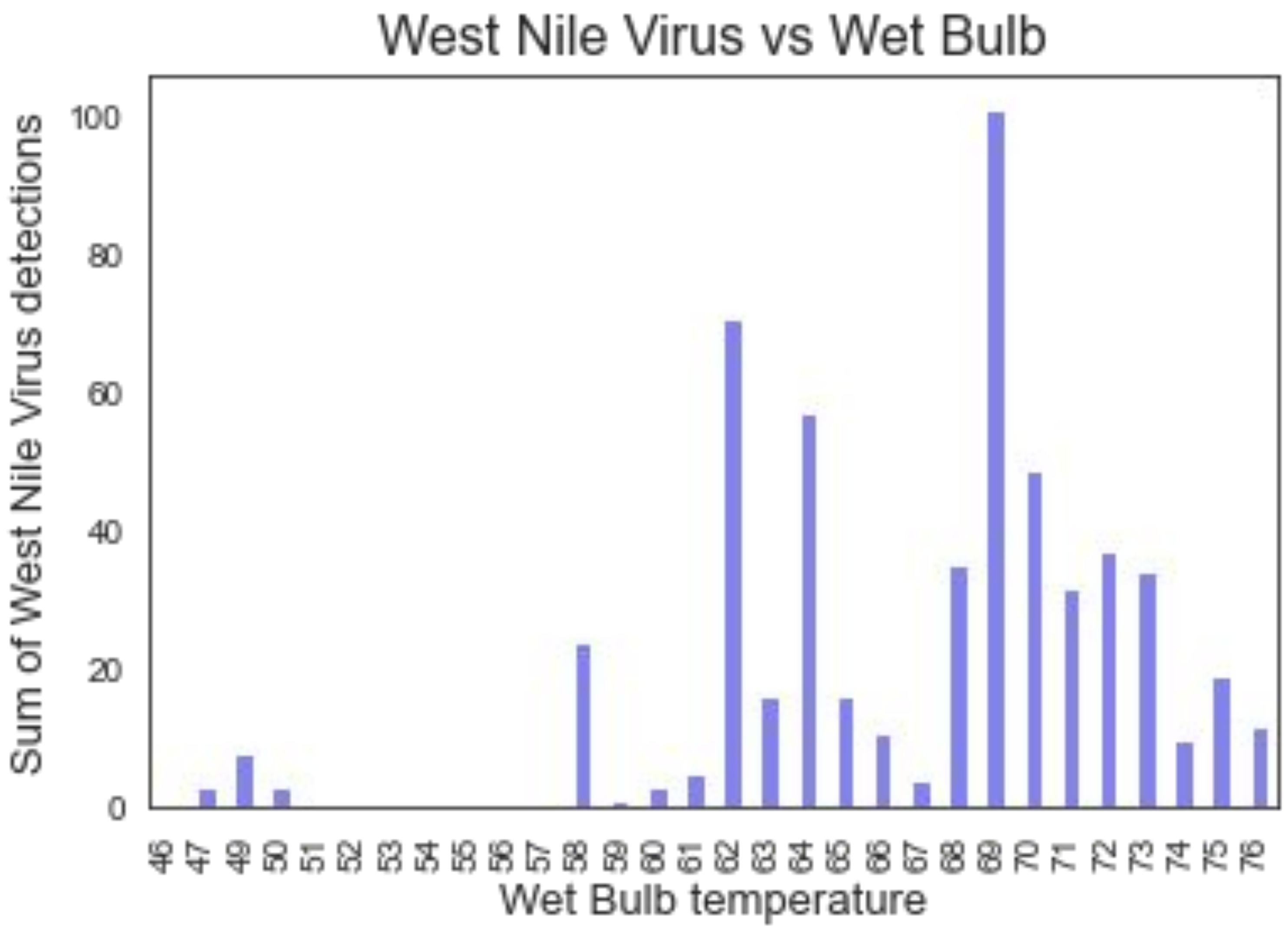
QUESTIONS & ANSWERS

Disease & Treatment Agency

Societal Cures In Epidemiology and New Creative Engineering

Appendix (EDA)

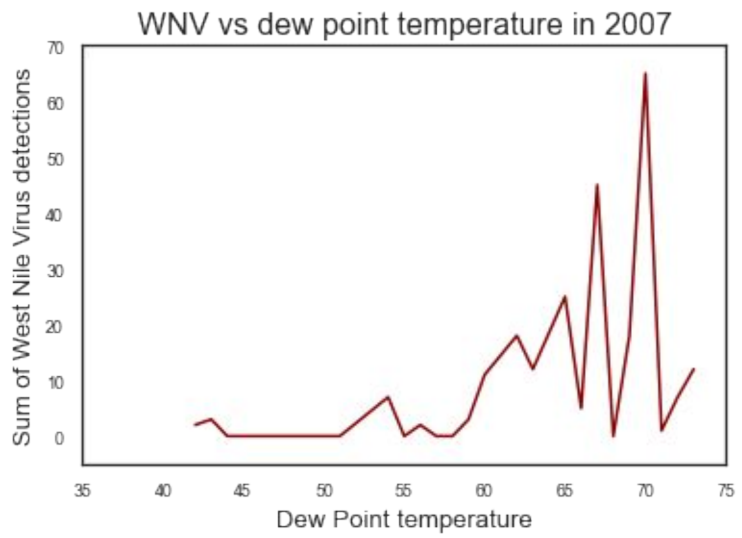


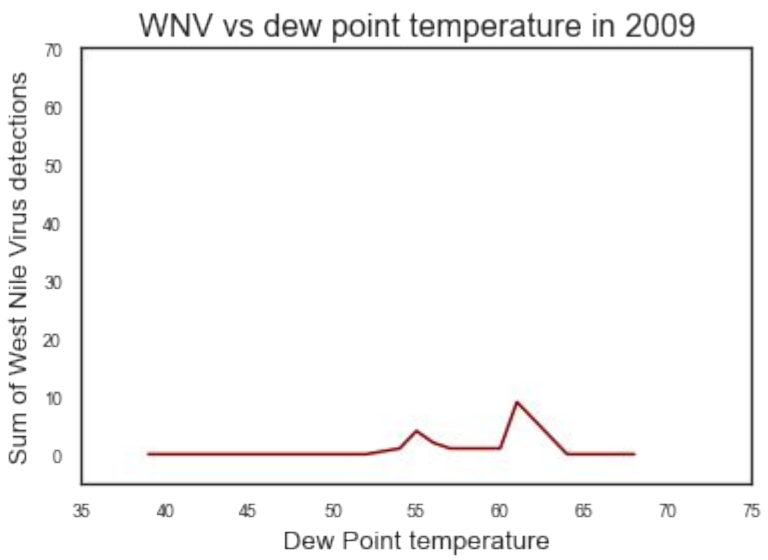


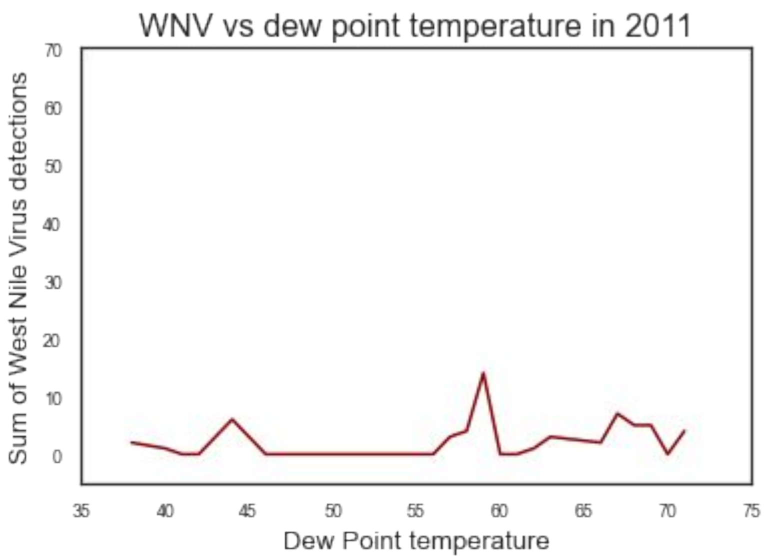
From the two graphs above we can see that there is a relationship between the number of WNV positive mosquitoes and WetBulb and DewPoint temperature.

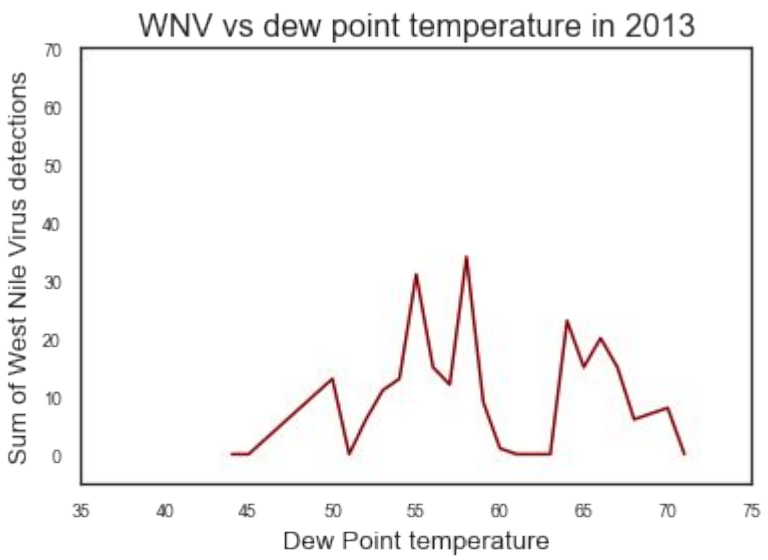
The higher the WetBulb and DewPoint temperatures, the higher the number of WNV positive mosquitoes.

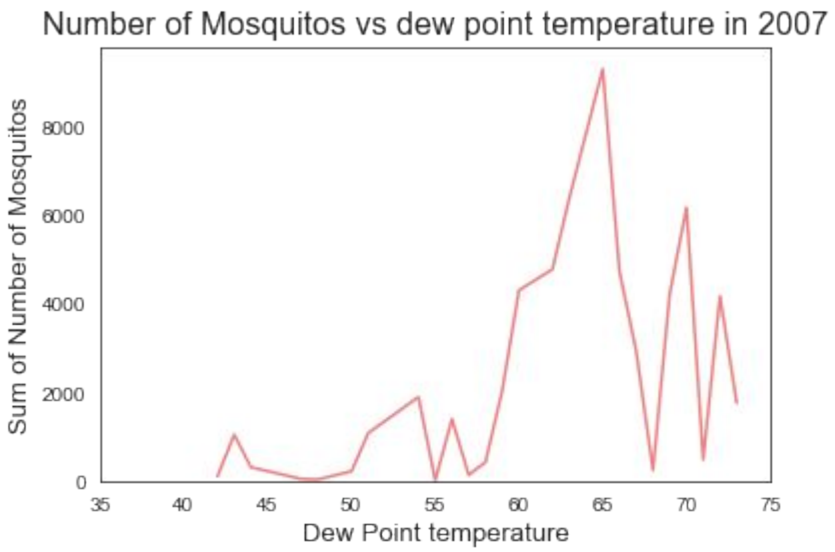
Appendix (EDA)

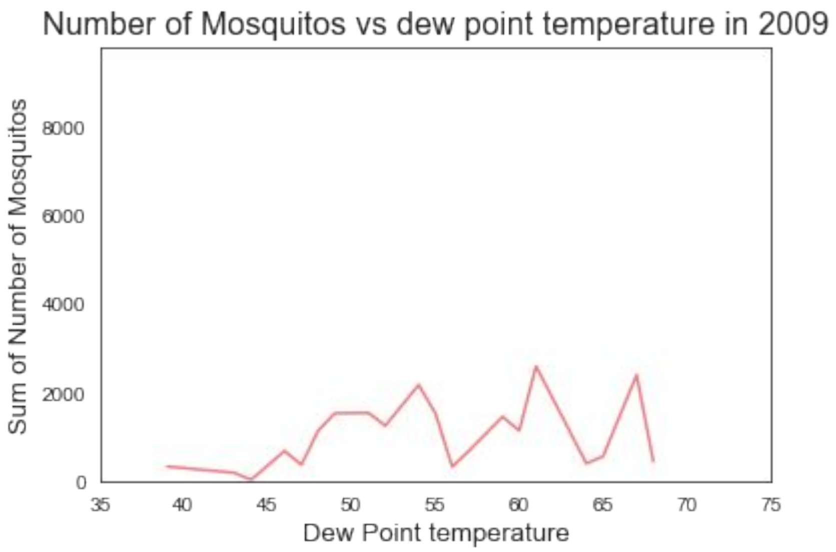

WNV vs dew point temperature in 2007

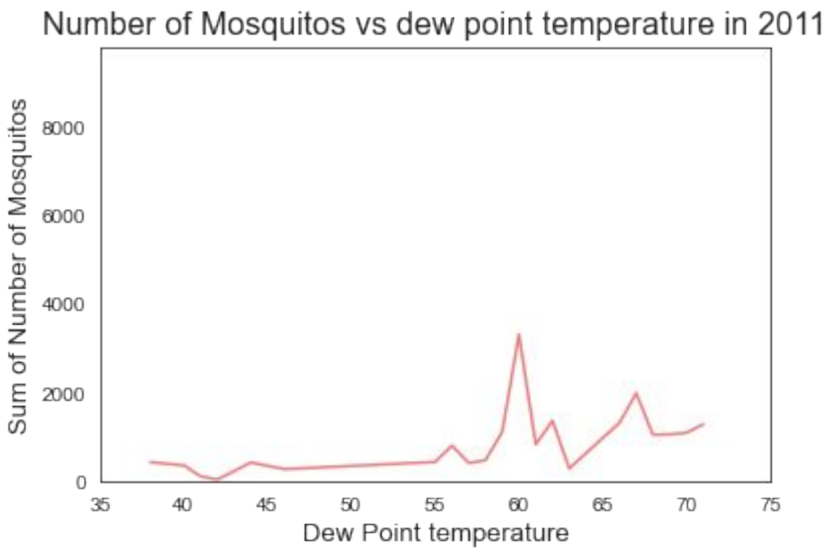

WNV vs dew point temperature in 2009

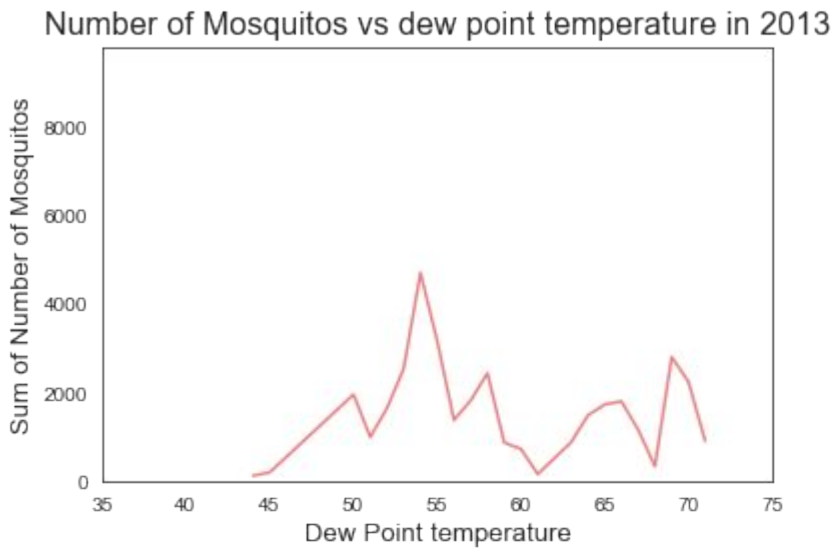

WNV vs dew point temperature in 2011


WNV vs dew point temperature in 2013

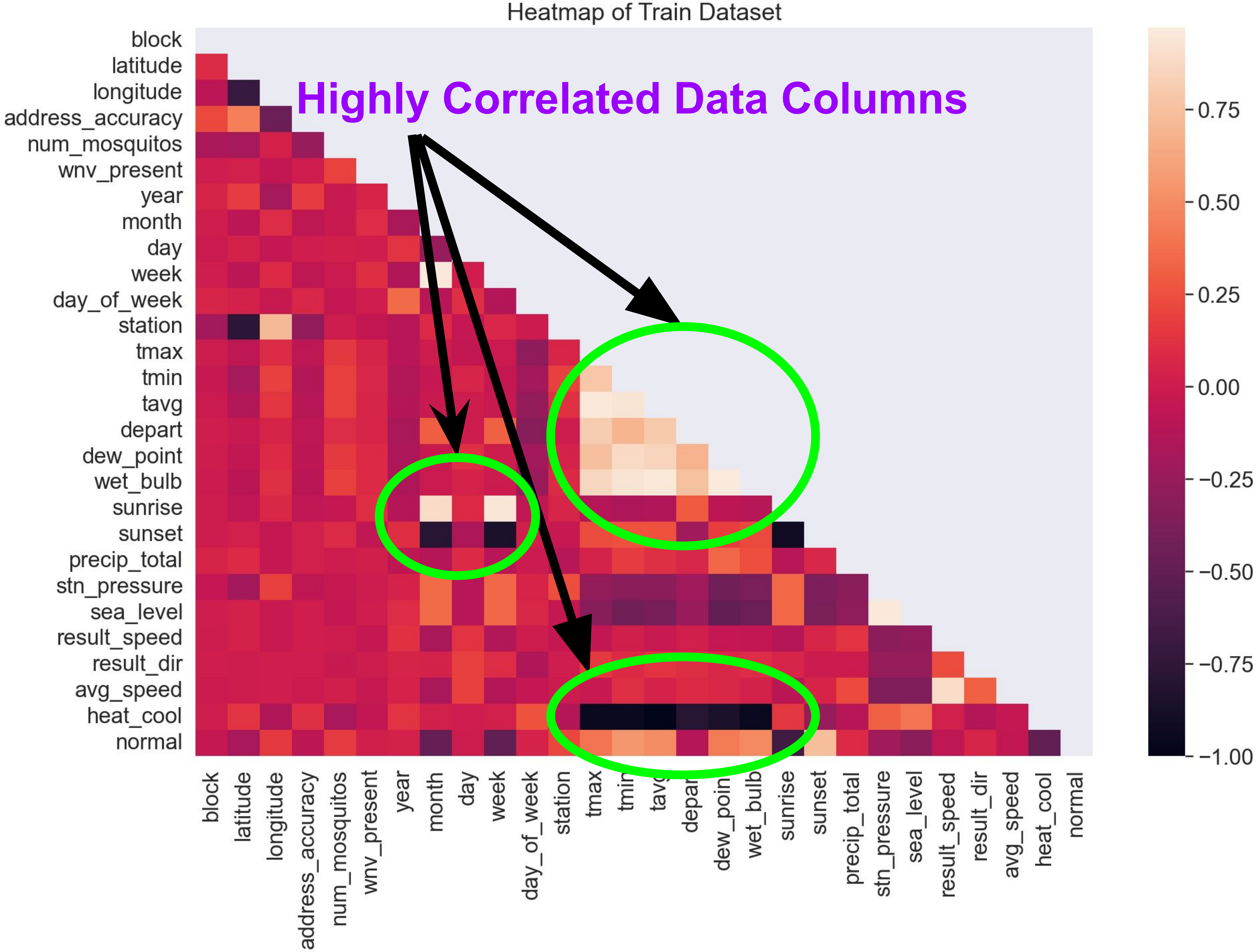

Number of Mosquitos vs dew point temperature in 2007


Number of Mosquitos vs dew point temperature in 2009


Number of Mosquitos vs dew point temperature in 2011


Number of Mosquitos vs dew point temperature in 2013

Appendix: Feature Engineering



Slide Title

- Point 1

Slide Title

- Point 1

Slide Title

Main point 1

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore

Main point 2

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore.

Main point 3

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore

Slide Title

- Point 1 and explanation
- Point 2 and explanation
- Point 3 and explanation

Slide Title

Main point I

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore

Main point 2

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore.

Main point 3

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore

Slide Title

1

Main point I

Explanation of main point 1. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor

2

Main point 2

Explanation of main point 2. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor

3

Main point 3

Explanation of main point 3. Lorem ipsum dolor sit amet, consectetur

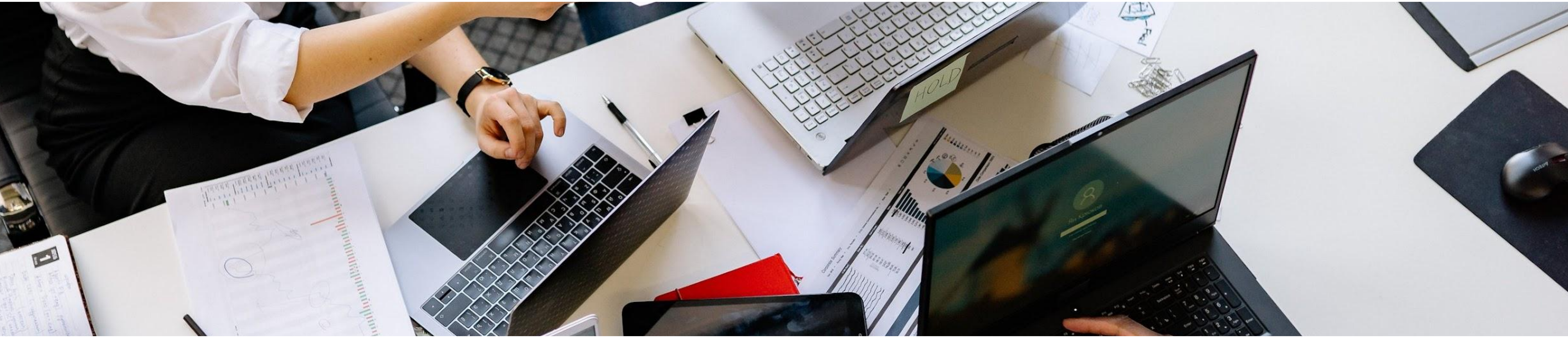
4

Main point 4

Explanation of main point 4. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod

Slide Title

- Point 1 and explanation
- Point 2 and explanation
- Point 3 and explanation



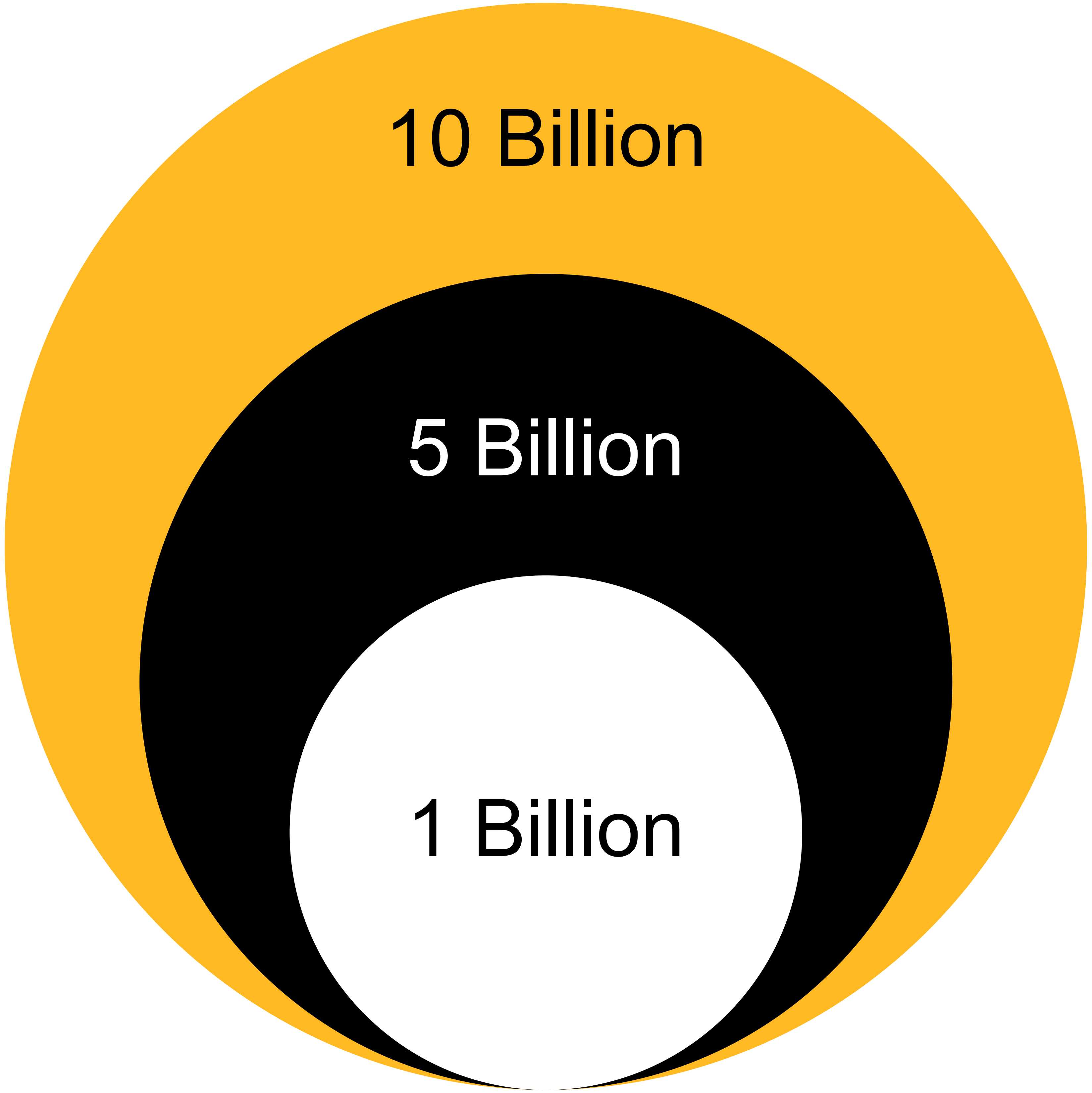
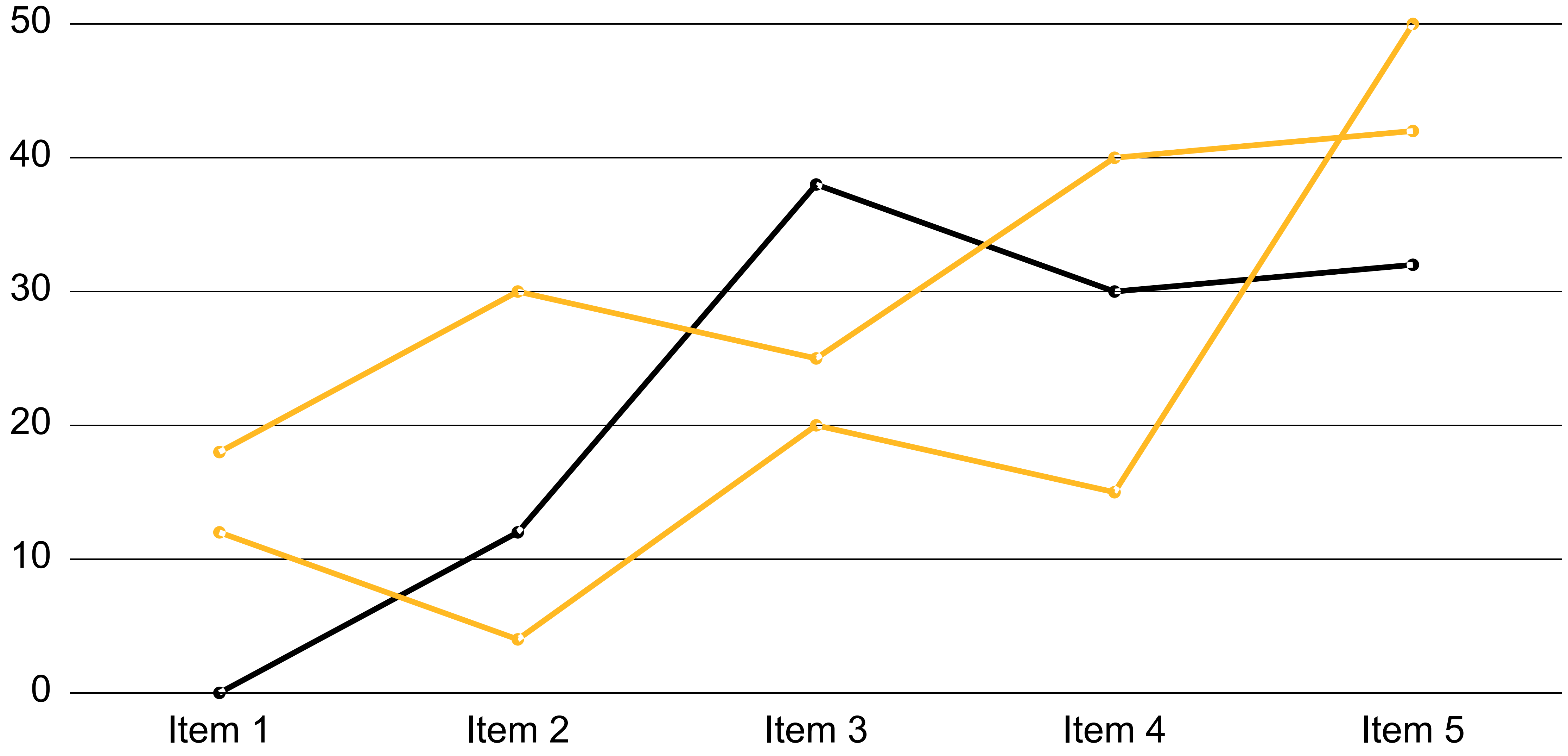


Chart Title

- Interpretation of chart and key takeaways
- Lorem ipsum

Chart Title

Subtitles (if applicable)



Main point

More verbose explanation of the chart interpretation and the main point.

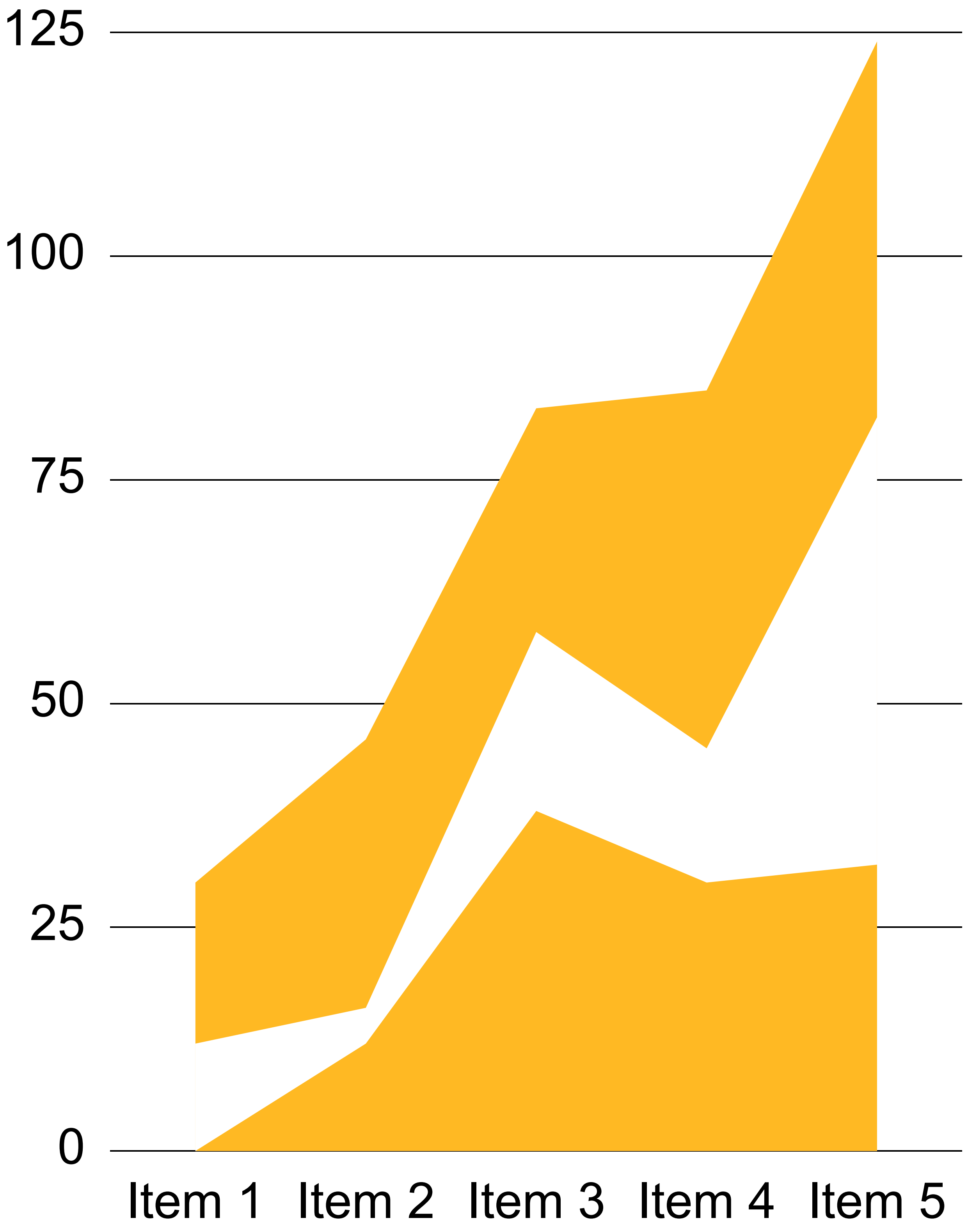
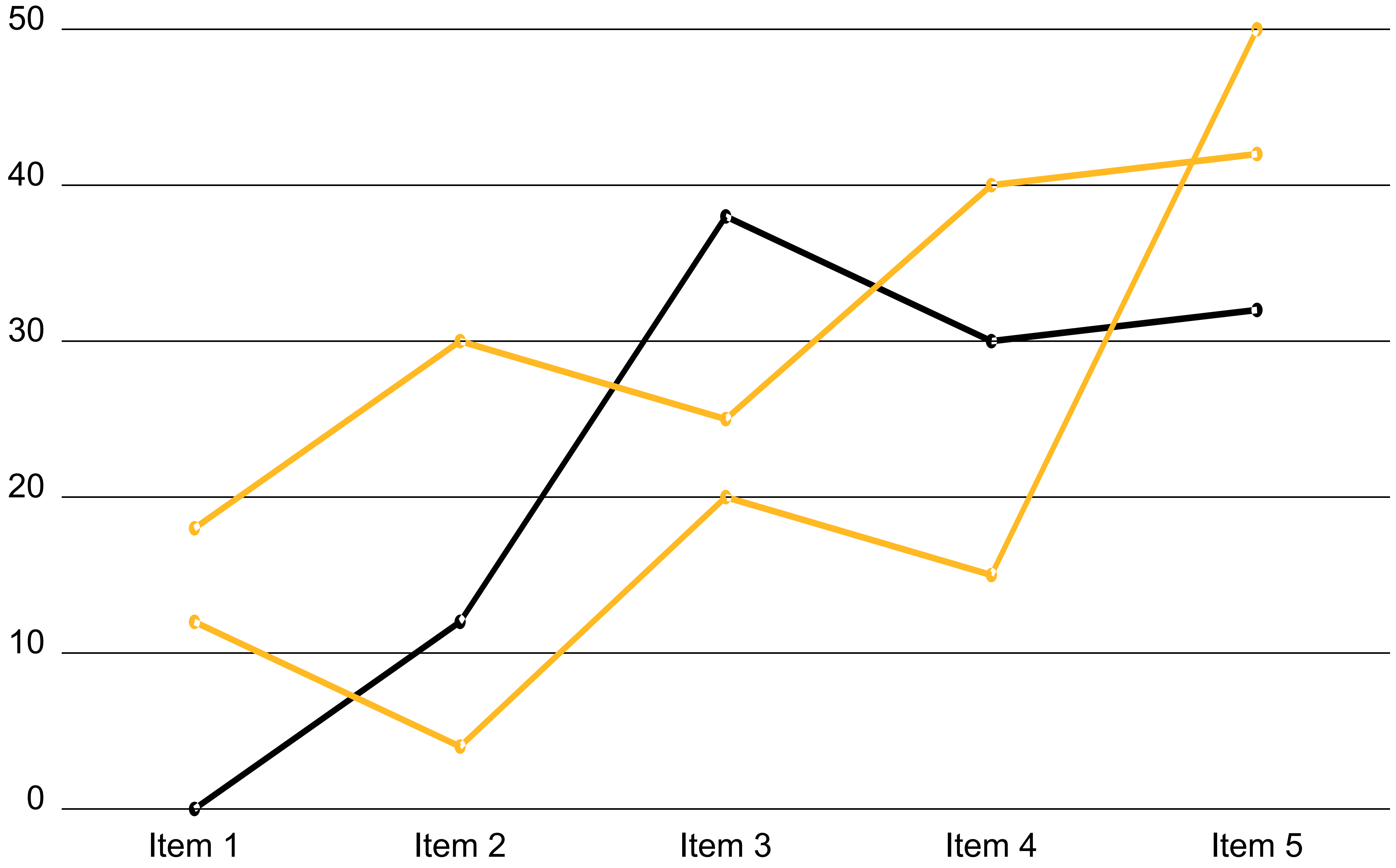


Chart Title

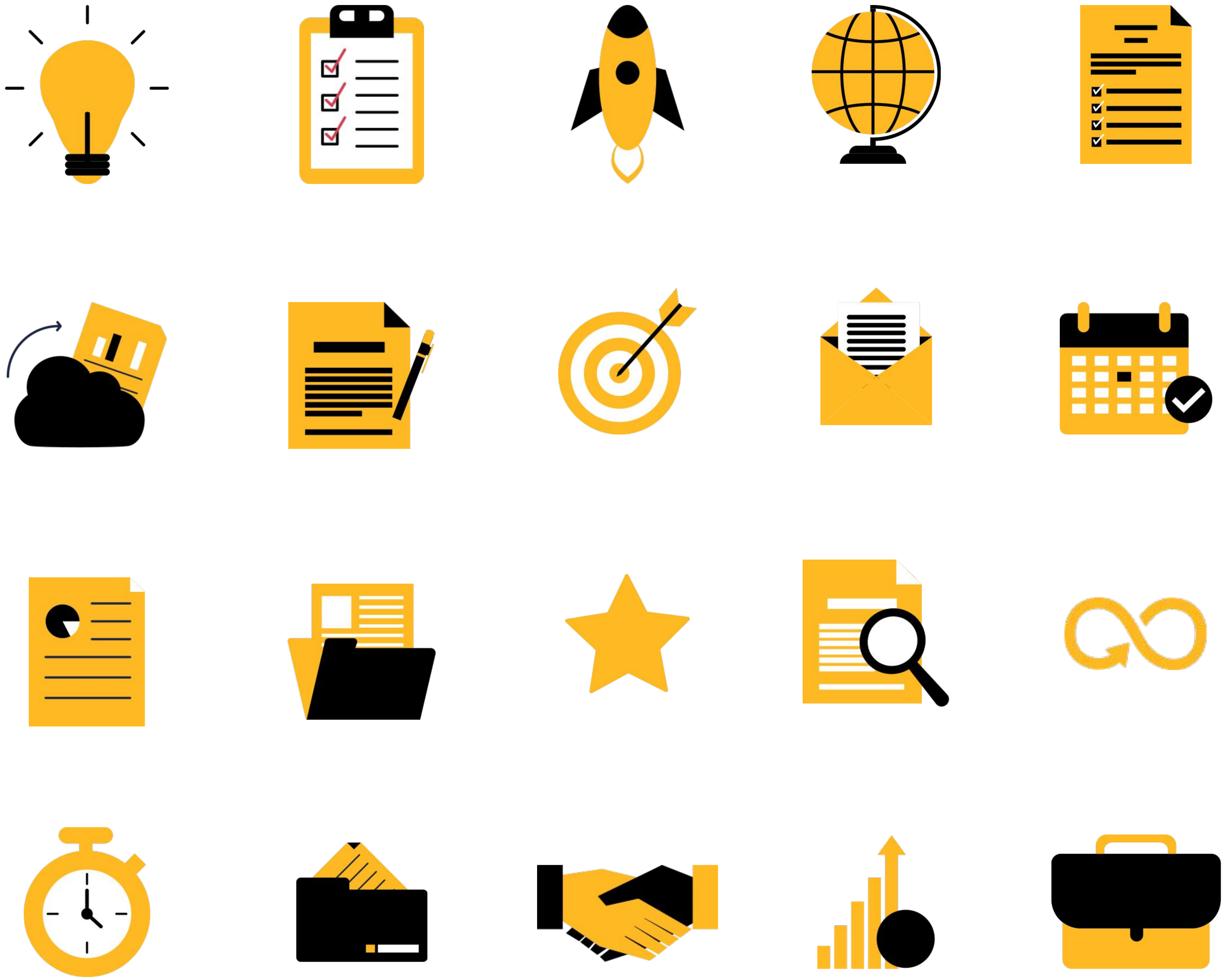
- Interpretation of chart and key takeaways
- Lorem ipsum

Chart Title

- Interpretation of chart and key takeaways
- Lorem ipsum



ICONS FOR USE IN SLIDES



Transmission of WNV

- Mosquito-borne disease
  - Spread through bite of an infected mosquito
  - Mosquitoes are infected from feeding on infected birds
- Culex mosquitoes are the main vectors of WNV
  - Especially Culex pipiens
- Mosquito population is highest in summer
  - Peaking in August and September