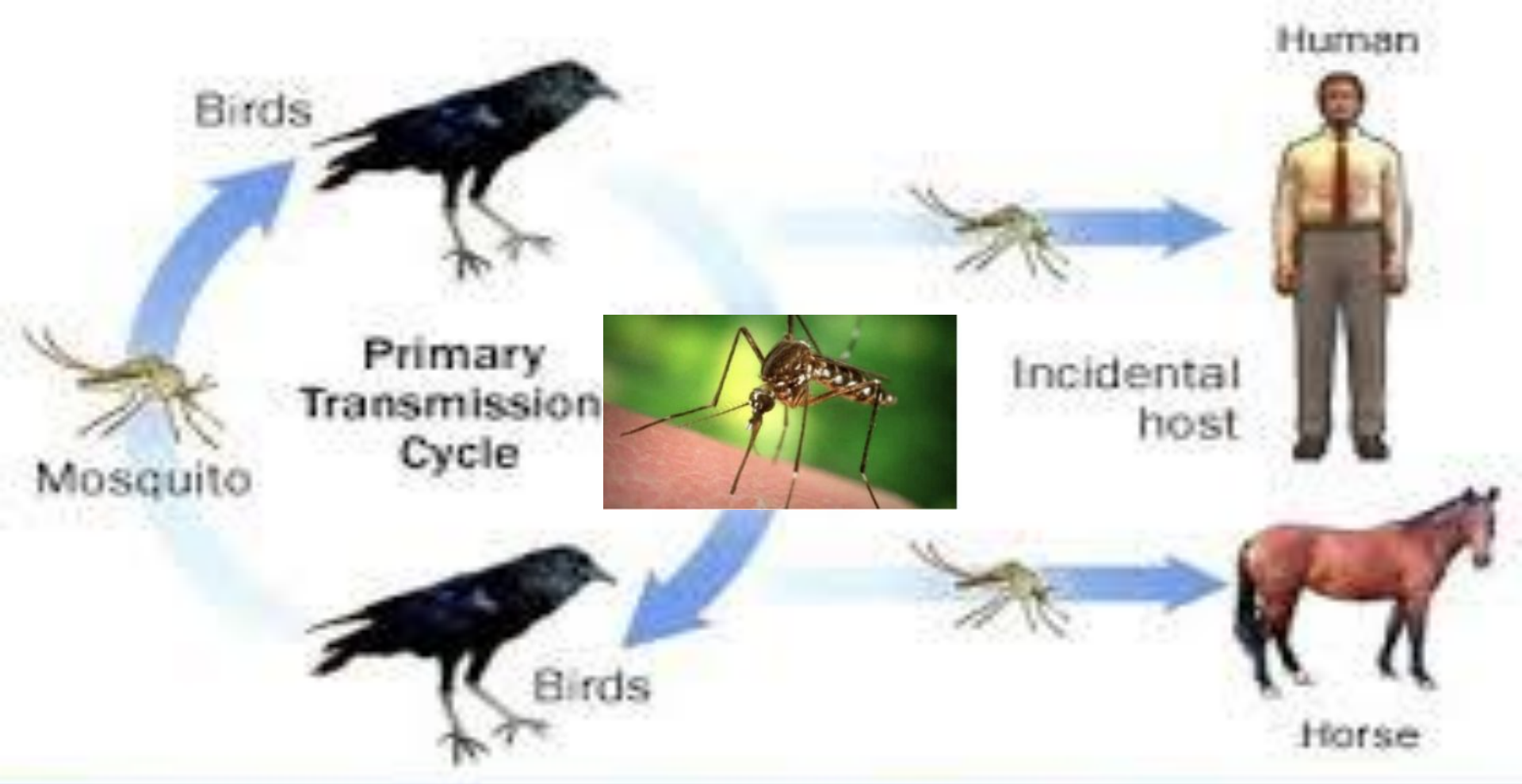
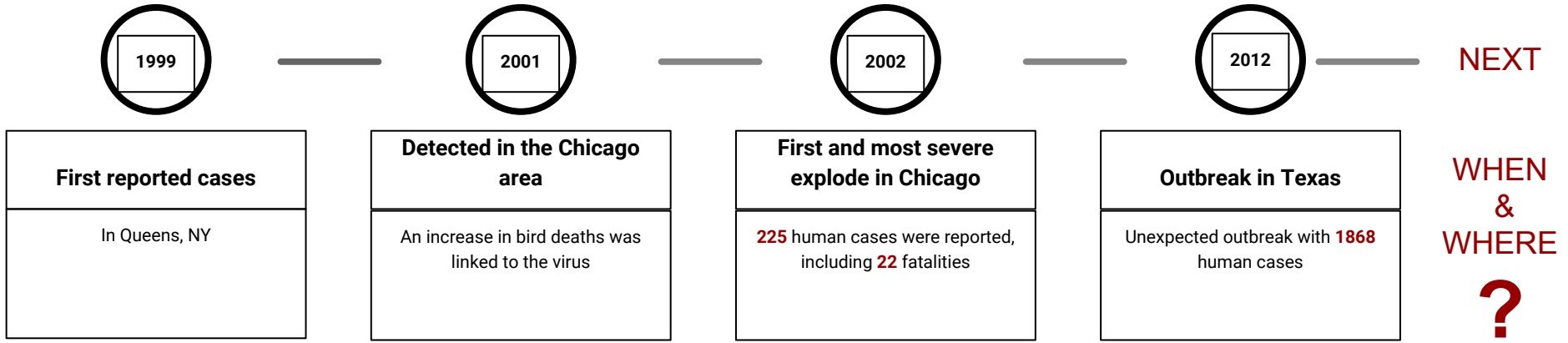
West Nile Virus Prediction

Yanxia Li



West Nile Virus (WNV): Chief among deadly Mosquito-borne threats in the US

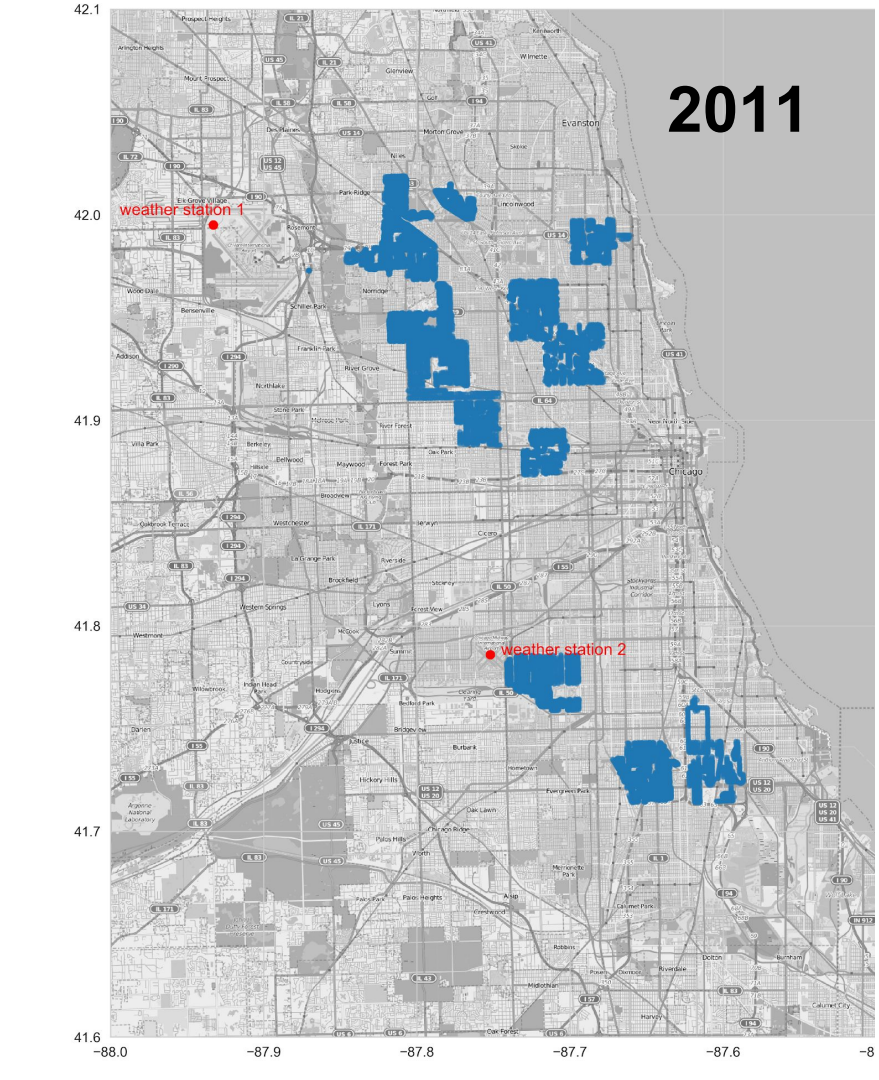


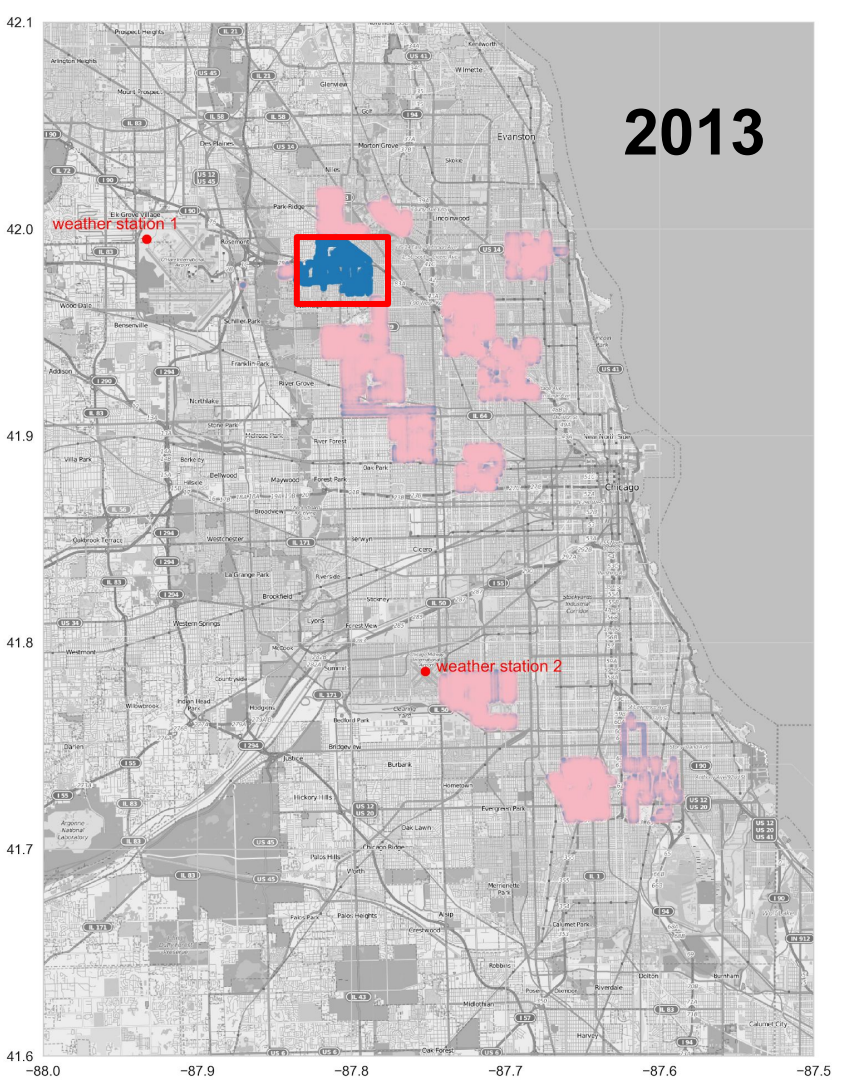
WNV Facts:

- **20%** of people who become infected develop symptoms ranging from a persistent fever, to serious neurological illness that can result in death
- Currently, **NO** vaccine or specific treatment for WNV
- On the population level, **community-based mosquito control programs** are the most effective tool to prevent the spread of WNV
- However, these programs are typically inadequately funded and the effectiveness of these control measures can be difficult to assess.

Chicago program:

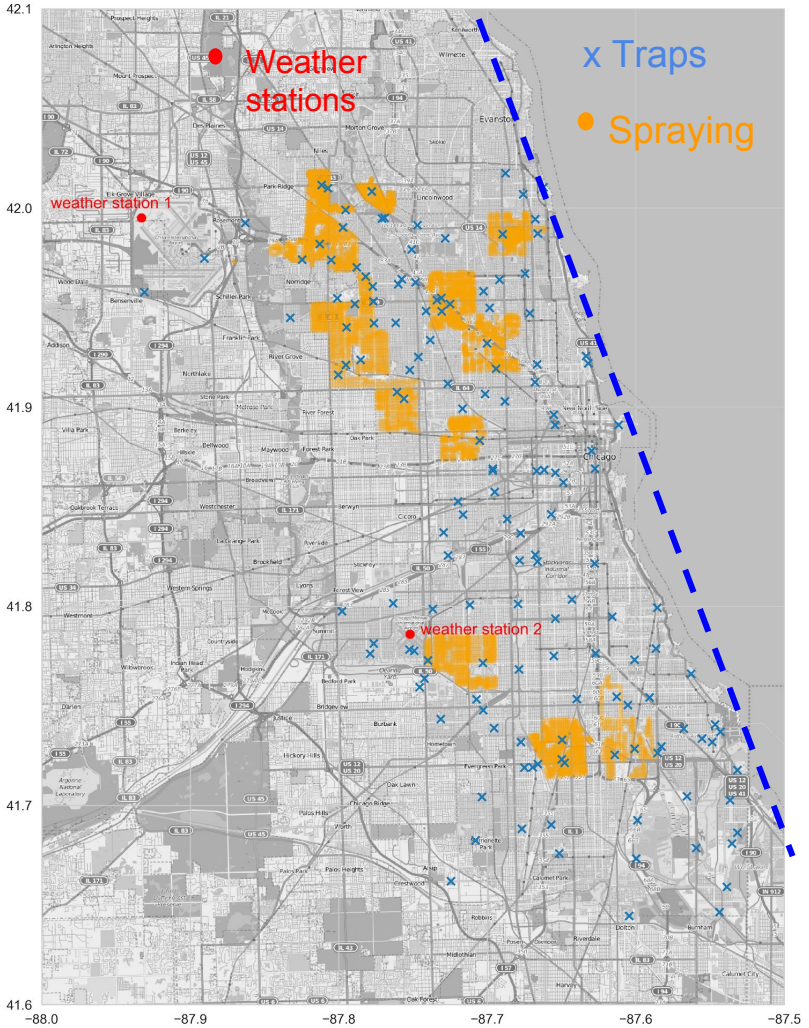
- The Chicago Department of Public Health established a comprehensive surveillance and control program to fight the spread of WNV
  - Larvicide in stormwater drains
  - DNA tests of mosquitoes
  - Spraying when WNV is present


2011
weather station 1
weather station 2
42.1
42.0
41.9
41.8
41.7
41.6
−88.0
−87.9
−87.8
−87.7
−87.6
−8


2013
weather station 1
weather station 2
42.1
42.0
41.9
41.8
41.7
41.6
−88.0
−87.9
−87.8
−87.7
−87.6
−87.5

- At any given time there are 60+ traps
- The traps are collected twice/week.
- Batches of up to 50 mosquitoes are

DNA tested.



The goal is to predict WNV presence, in order to help more efficiently and effectively allocate resources towards preventing transmission of this potentially deadly virus.

Datasets:

- Weather:

From NOAA, 2007 to 2014, during the months of the tests

- Traps:

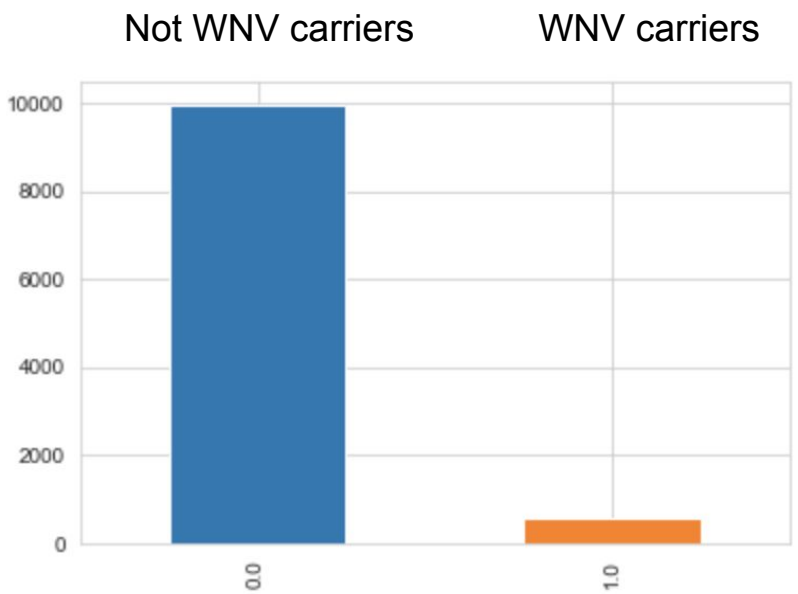
Where and when they put the traps

- Spraying:

EDA: inspect each feature

WnvPresent: Target Variable

```
1  df['WnvPresent'].value_counts().plot.bar()
```

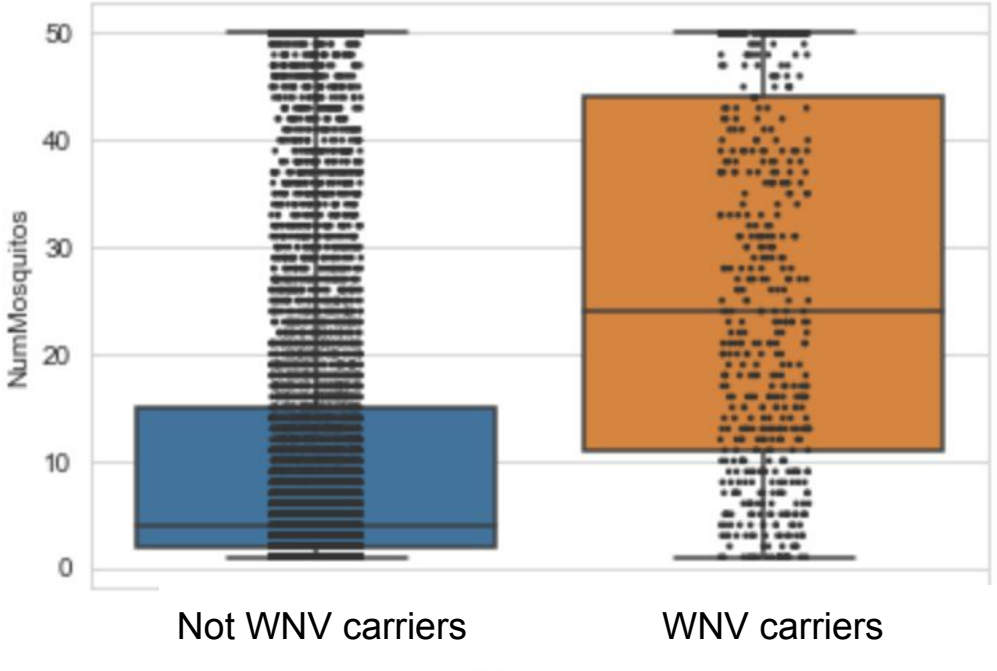


About 95% (10k) of the Mosquitoes are not Wnv carriers and 5% (0.5k) are found to have WNV.

This is an unbalanced classification problem. Be careful with model selection and cross-validation.

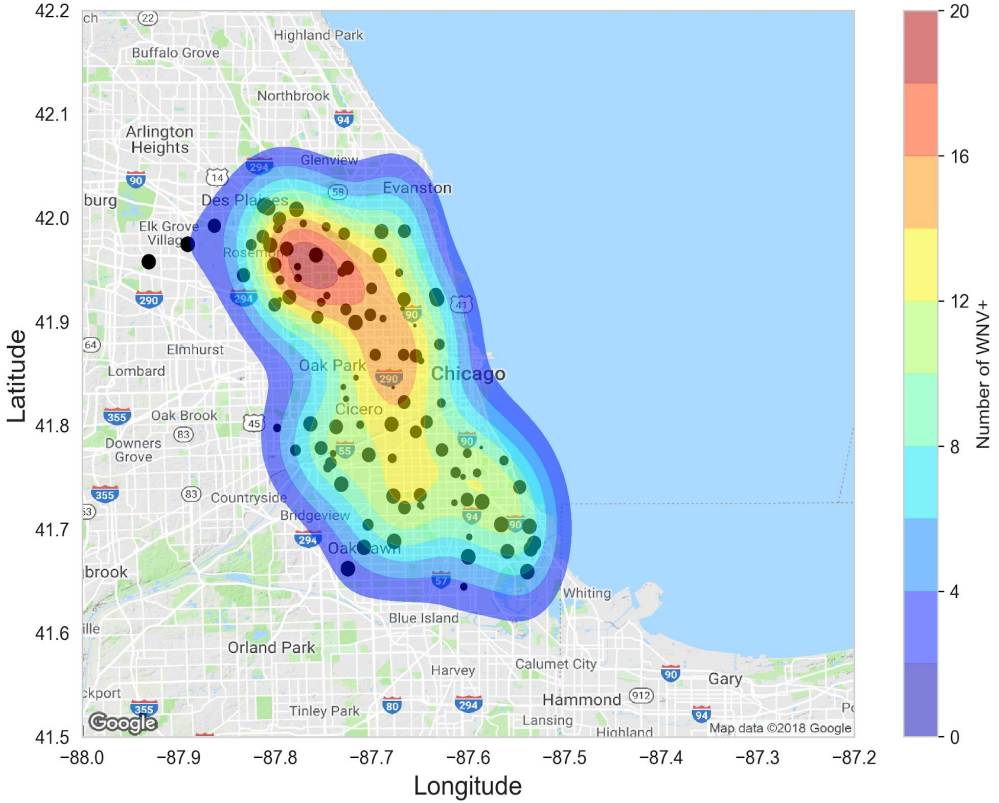
“NumMosquitos”

Strips of observations on top of a box plot

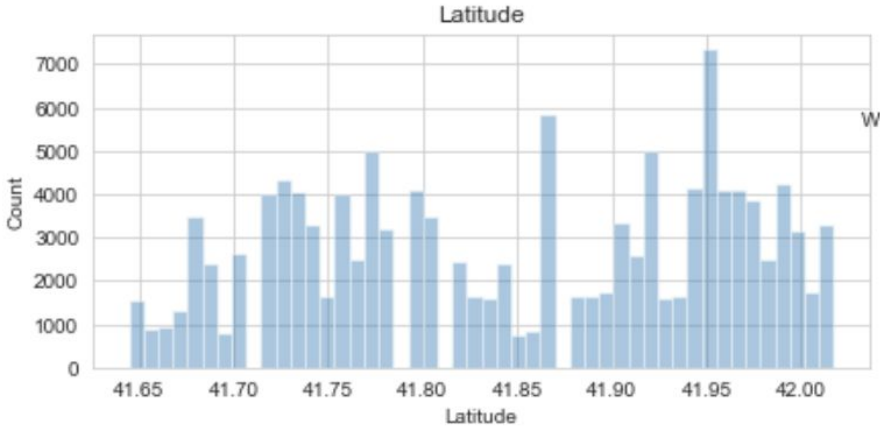


1. Only a small amount of WNV carriers among all mosquitoes
2. WNV carriers have a larger median number of mosquitoes
3. WNV carriers have a larger range of “NumMosquitos” values, 10 - 40.
4. It *looks* quite different, but are the distributions statistically different?

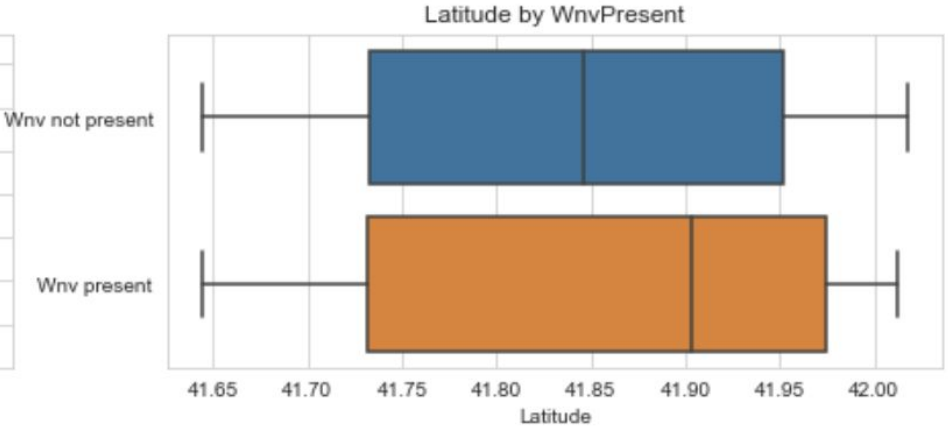
Geographic variations of the number of mosquitoes?



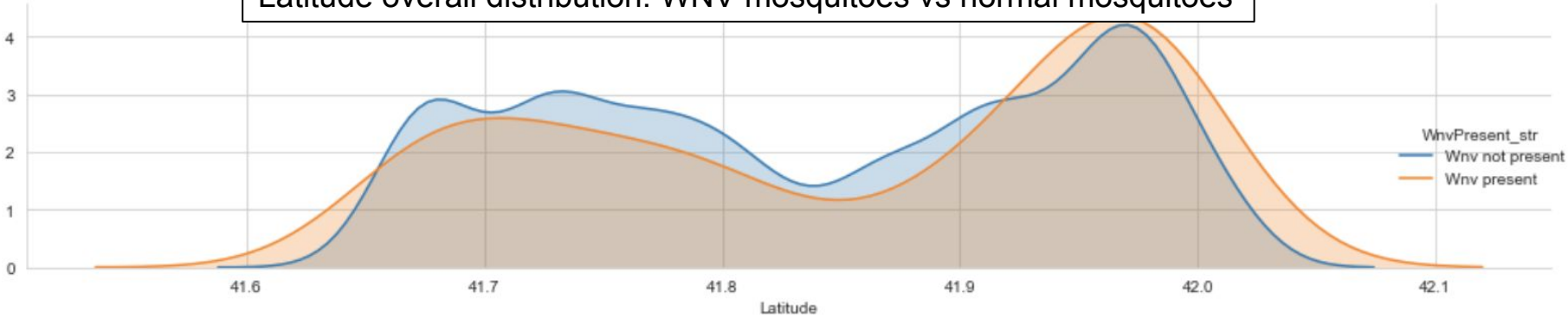
Latitude overall distribution



WNV more at higher latitudes?

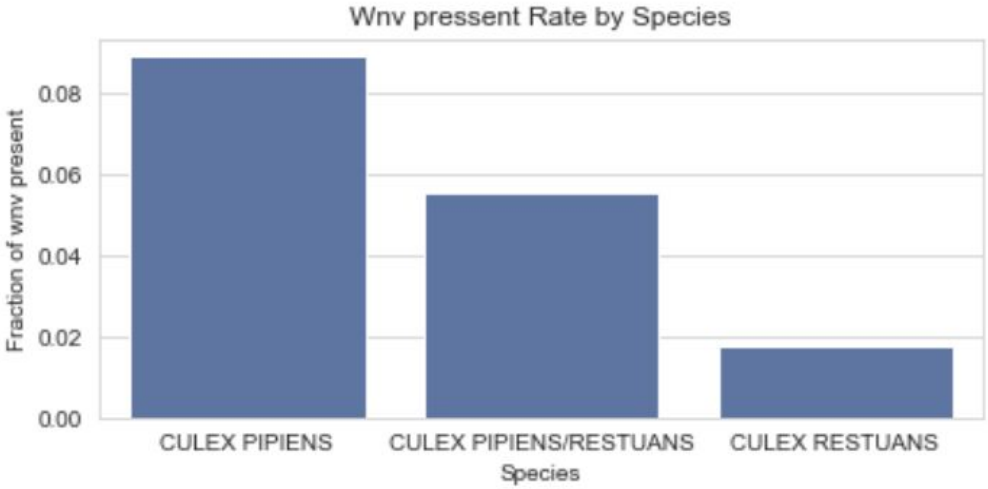


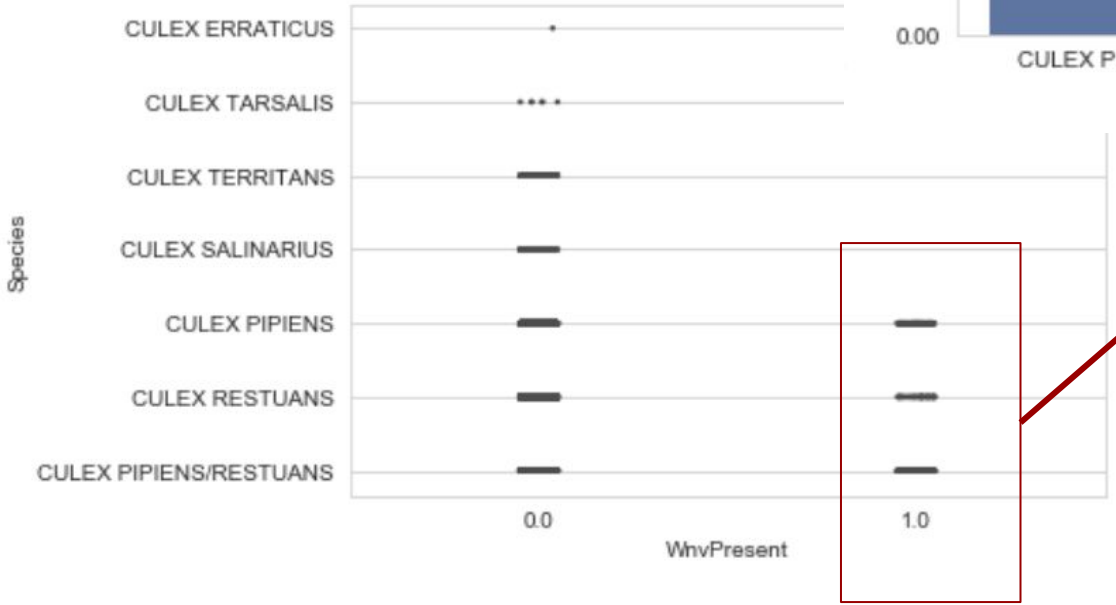
Latitude overall distribution: WNV mosquitoes vs normal mosquitoes



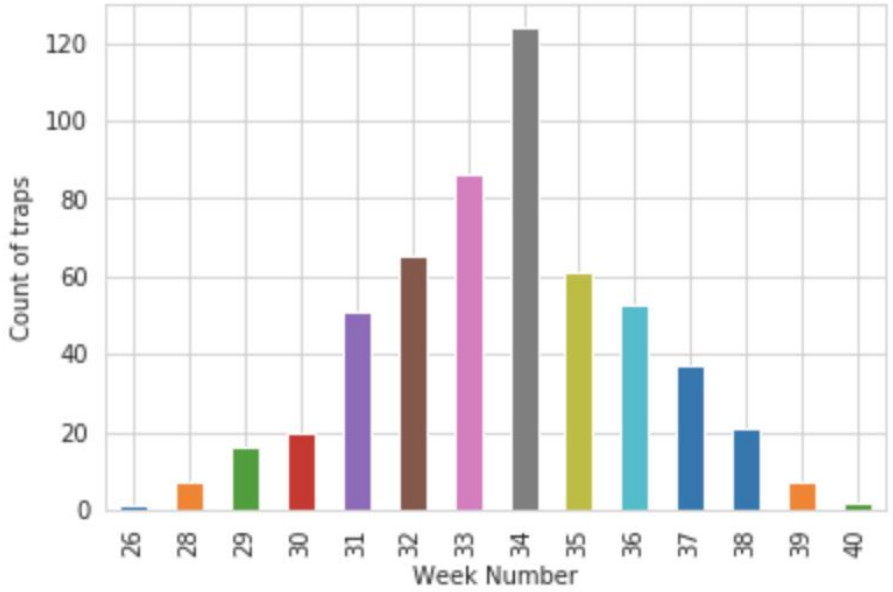
Species

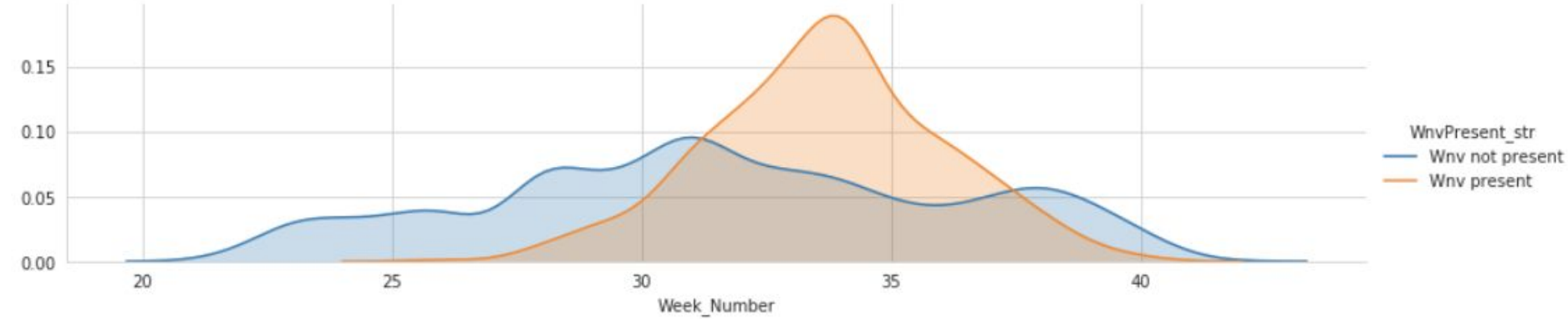
Culex Pipiens & Culex Restuans are the top 2 species as WNV carriers.

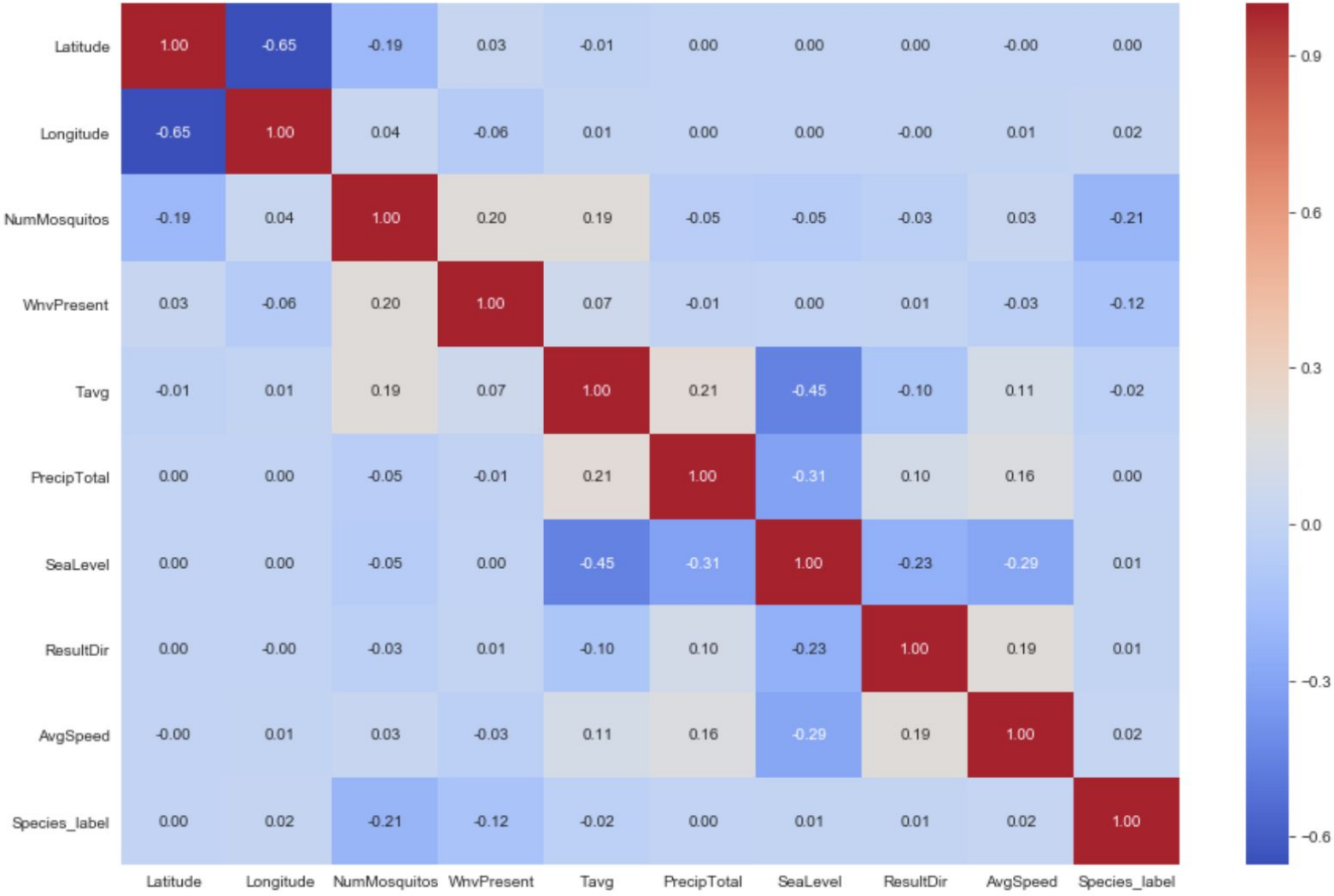




Seasonal effects






Latitude 1.00 -0.65 -0.19 0.03 -0.01 0.00 0.00 0.00 -0.00 0.00
Longitude -0.65 1.00 0.04 -0.06 0.01 0.00 0.00 -0.00 0.01 0.02
NumMosquitos -0.19 0.04 1.00 0.20 0.19 -0.05 -0.05 -0.03 0.03 -0.21
WnvPresent 0.03 -0.06 0.20 1.00 0.07 -0.01 0.00 0.01 -0.03 -0.12
Tavg -0.01 0.01 0.19 0.07 1.00 0.21 -0.45 -0.10 0.11 -0.02
PrecipTotal 0.00 0.00 -0.05 -0.01 0.21 1.00 -0.31 0.10 0.16 0.00
SeaLevel 0.00 0.00 -0.05 0.00 -0.45 -0.31 1.00 -0.23 -0.29 0.01
ResultDir 0.00 -0.00 -0.03 0.01 -0.10 0.10 -0.23 1.00 0.19 0.01
AvgSpeed -0.00 0.01 0.03 -0.03 0.11 0.16 -0.29 0.19 1.00 0.02
Species_label 0.00 0.02 -0.21 -0.12 -0.02 0.00 0.01 0.01 0.02 1.00
Latitude
Longitude
NumMosquitos
WnvPresent
Tavg
PrecipTotal
SeaLevel
ResultDir
AvgSpeed
Species_label
0.9
0.6
0.3
0.0
−0.3
−0.6

Engineering

1. train/test split: I use WNV detected on earlier dates to predict the outcome of WNV presence on later dates (ratio of 8:2)
2. Metrics: Mostly AUROC

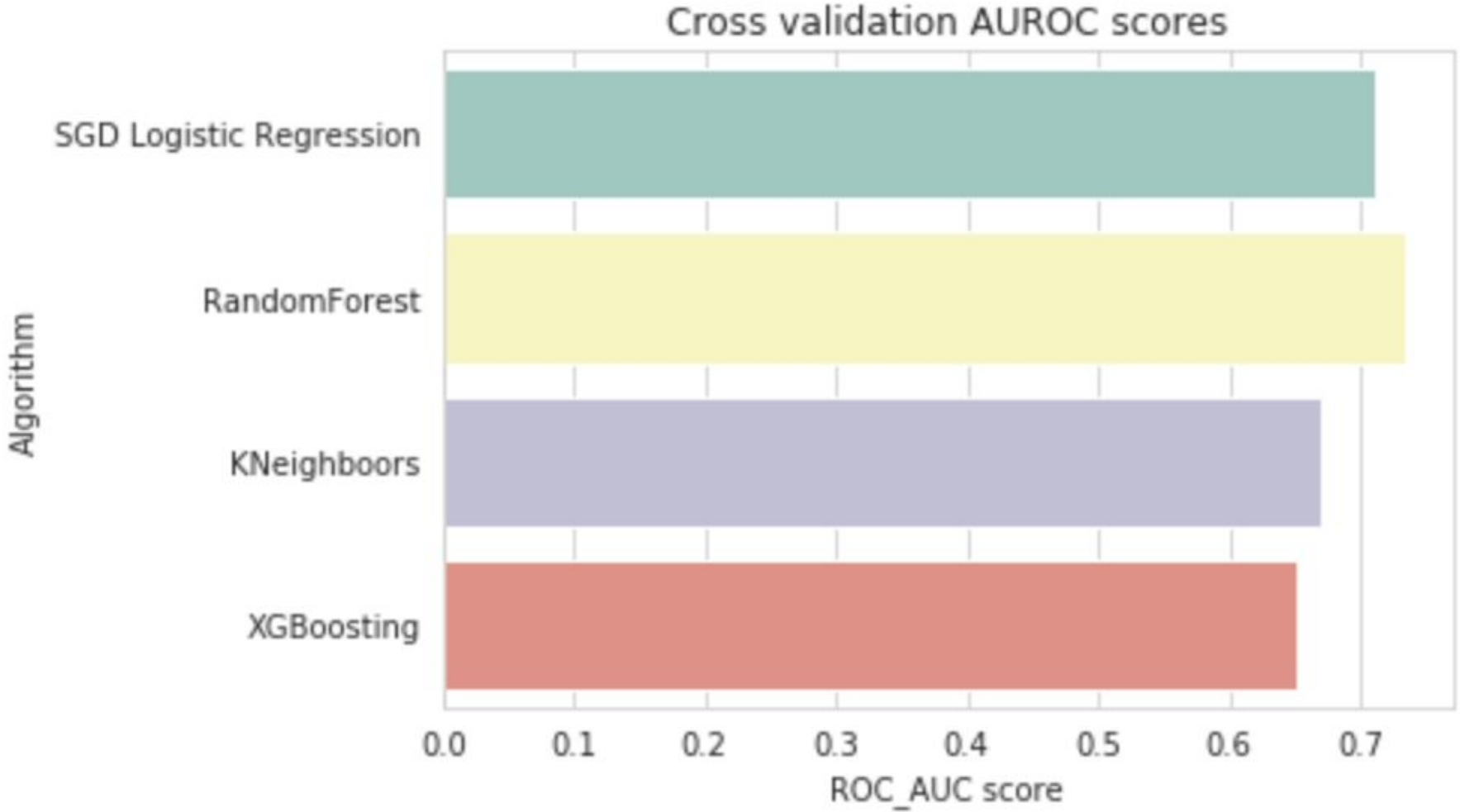
	Tavg	PrecipTotal	SeaLevel	ResultDir	AvgSpeed	trap_dist	Week_Number	Species_label
0	74.0	0.0	30.11	18.0	6.5	-0.272072	22	1
1	74.0	0.0	30.11	18.0	6.5	-0.272072	22	2
2	74.0	0.0	30.11	18.0	6.5	-0.025062	22	2
3	74.0	0.0	30.11	18.0	6.5	-0.346565	22	1
4	74.0	0.0	30.11	18.0	6.5	-0.346565	22	2

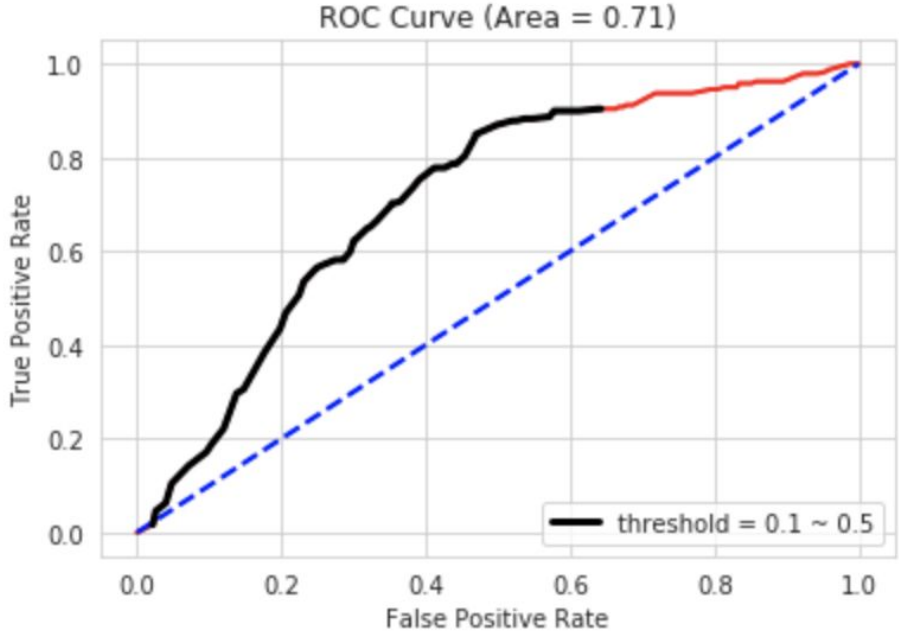
Factors most relevant to mosquito population?

	index	pearson_corr
1	Species_label	-0.211172
2	WnvPresent	0.196621
3	Tavg	0.191456
4	Latitude	-0.185226
5	SeaLevel	-0.053283
6	PrecipTotal	-0.050171
7	Longitude	0.038065
8	ResultDir	-0.026904
9	AvgSpeed	0.026834

Factors most relevant to WNV presence?

	index	pearson_corr
1	NumMosquitos	0.196621
2	Species_label	-0.120986
3	Tavg	0.067092
4	Longitude	-0.060360
5	Latitude	0.029043
6	AvgSpeed	-0.026863
7	ResultDir	0.008030
8	PrecipTotal	-0.007099
9	SeaLevel	0.000805


Cross validation AUROC scores
SGD Logistic Regression
RandomForest
KNeighboors
XGBoosting
Algorithm
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
ROC_AUC score


ROC Curve (Area = 0.71)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
threshold = 0.1 ~ 0.5

I chose a cut off of 0.25. Anything over 0.25, I said “this is a positive”. With that cut off, there is a 71% chance that model will be able to distinguish between positive class and negative class.

Feature importance

	weight
Species_label	-0.157
Week_Number	0.118
Tavg	0.113
SeaLevel	-0.066
trap_dist	-0.066
AvgSpeed	-0.056
PrecipTotal	0.024
ResultDir	-0.002

	importance
Week_Number	0.638500
Tavg	0.120113
SeaLevel	0.115582
Species_label	0.081706
PrecipTotal	0.020919
AvgSpeed	0.011568
ResultDir	0.009414
trap_dist	0.002198

Conclusion

- I applied ML to predict the probability of WNV presence. After training with 4 models (Logistic Regression, random forest, XG Boosting and KNN), I found that Random Forest classifier performs the best (highest AUROC score of 0.82) and the score on the testing dataset is slightly lower at 0.71.
- the most important features for predicting WNV are: Species, Week of the year, Temperature.

Alternative data

After researching on the problem, I think the following data could be helpful:

- Reported human WNV cases: time, neighborhood, etc
- The life expectancy of the mosquito
- The recovery time for an infected bird and the contact rate between mosquitoes and birds
- If having the above datasets, can we further model relationship between peak timing of infectious mosquitoes, total numbers of infected mosquitoes, and spillover infection of humans?

Backups

Spray Data

- The data quality is good, only 4% missing data in “Time”
- There are only 2 unique spraying days in 2011 and 8 spraying days in 2013.
- Spraying could reduce the number of mosquitos in the area and therefore might eliminate the appearance of WNV. However, I could imagine several difficulties in using this dataset
    - What is the effective area/radius of the spraying? And how long does it last?
    - The number difference before and after each spraying at each location?
    - Need to compare the spraying dates and the testing dates.
    - If I include this feature into my final data frame for modeling, how should I deal with the missing values?
- But, as a first-order check/comparison, probably I can just plot the spraying locations on the maps and check their distribution on the maps, etc.

Train/Test Data

- There are 2 variables missing in the testing data. One is our target variable, “WnvPresent” (0: NO, 1: YES), the other one is the number of mosquitoes (“NumMosquitos”). Later I will show that these two are correlated.
- I add the weather information for each “Date” in the train/test data.
- I stack the test and train data in order to do the same transformations and feature engineering to both. Before training the model, I will split them back.

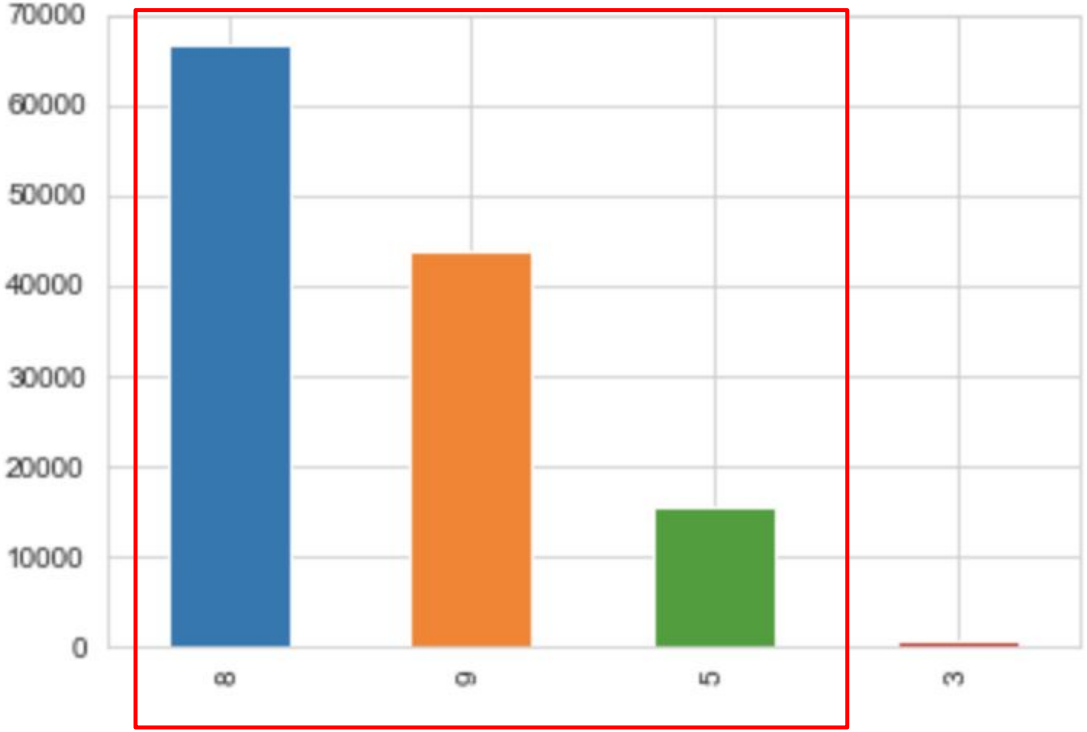
Weather

- Weather From NOAA, 2007 to 2014, during the months of the tests
- Messy data:
  - Missing values (stored as “M”, or “-”, or blank)
  - Wrong formatted values (“ T”)
  - Each date has 2 records, one for each station
  - There are certain information that are not recorded in station 2 (e.g., “sunset” time), and/or station 1 (e.g., “Water 1”)
- My plan is to merge this weather data into the train/test data on the same “Date”. So I need to merge the 2 stations’ records together. As there is no significant difference between the measurements from the 2 stations, I use station 1’s data. (Ideally, I could use the average of the 2 stations or find a better way than just throwing away station 2’s data.)

“Date”

- The "Date" range from May to September.
- The number of traps for each date ranges from just a few to more than a thousand for different days.
- The "Date" information is the “key” when merging with the weather dataset.

Address information (Block, Latitude, Longitude...)



98% of our address information is accurate (with accuracy value of 5 and above)

I will use the [Latitude, Longitude] for the locations, as it is more accurate and easier to handle than the categorical values.

If need to separate the area into different neighborhoods, can still do so using the Latitudes and Longitudes.

“NumMosquitos”

The pairwise correlation with the target variable “WnvPresent” is ~0.20!

Thoughts:

This feature is very interesting. It only exists in the training dataset but intuitively, it makes sense that the more mosquitos the higher possiblity of Wnvpresent.

- How to include this feature into the modeling?

I am thinking to first predict this variable for the testing dataset and then use the modeled values as feature into final prediction of the Wnvpresent.

- These test results are organized in such a way that when the number of mosquitos exceed 50, they are split into another record (another row in the dataset), such that the number of mosquitos are capped at 50.

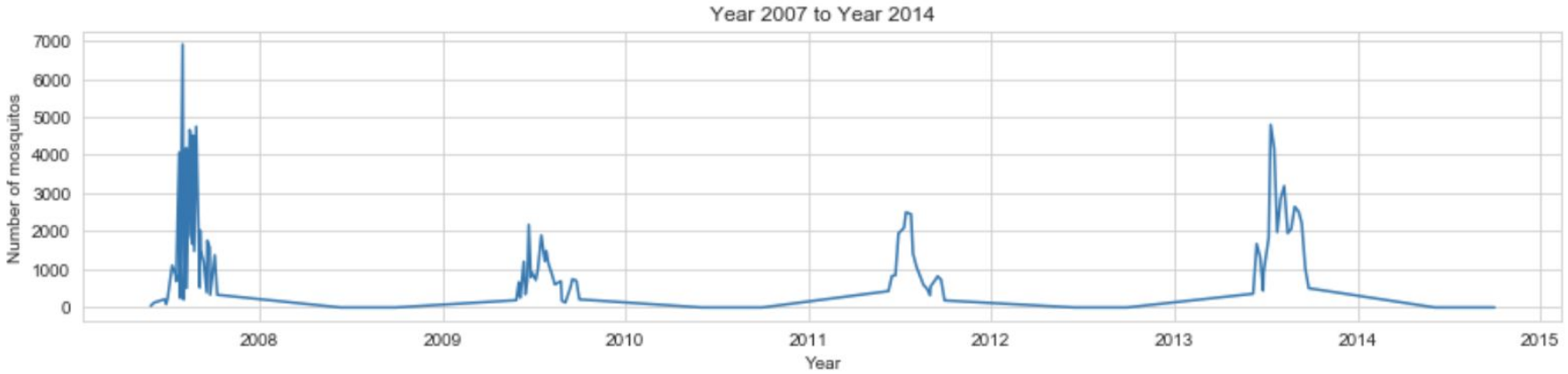
If I need to do regression on this feature, then i need to "count" it correctly for each trap.

Do the trap locations change with year/season?

I suspect that, at the beginning of each season, trap locations were spatially distributed throughout the county and guided by the historical presence of WNV.

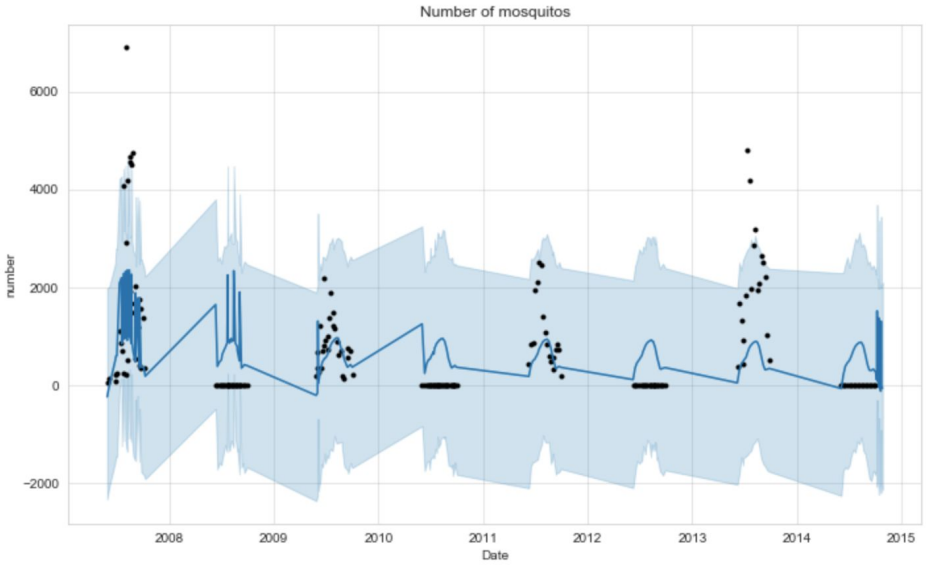
As the season progressed, mosquito monitoring was expanded within regions where WNV had been identified.

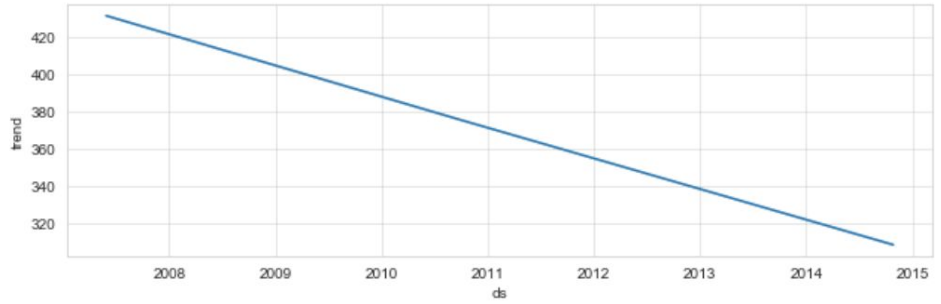
Is there any long-term change of the number of mosquitos?

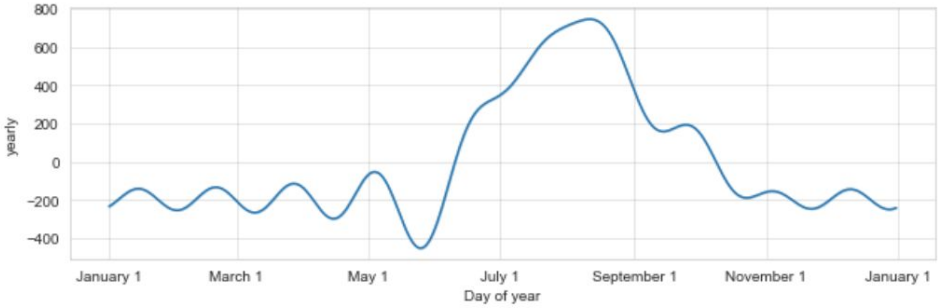


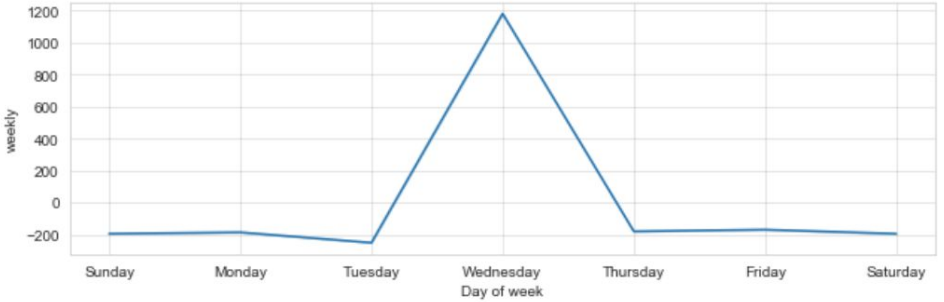
I further zoom in to each year in the training data (2007, 2009, 2011, 2013) and checked their variation along time. Hard to tell by eye.

I then performed a 1st order time-series analysis, using Facebook's Prophet package.


Number of mosquitos
number
6000
4000
2000
0
−2000
2008
2009
2010
2011
2012
2013
2014
2015
Date


trend
420
400
380
360
340
320
2008
2009
2010
2011
2012
2013
2014
2015
ds


yearly
800
600
400
200
0
−200
−400
January 1
March 1
May 1
July 1
September 1
November 1
January 1
Day of year


weekly
1200
1000
800
600
400
200
0
−200
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Day of week

Thoughts:

Here I will perform a time series analsis, 2 purpose:

- whethere there is a large-scale, long-term change of the number of the mosquitos that can be well-modeled with time series models.
- I earlier thought that I could use ML regression models to predict "NumMosquitos", but not I think I can also compare with the predictions from time series analysis.
- for the Time series analysis, here I am using Facebook's Prophet package. Based on my previous relevant project experience, I found it to be really "user-friendly", that is, easy to tune parameters. But the results are not necessarily better than ARIMA model.

If I decide to use time series to predict the number of mosquitos in testing dataset then I will further compare different time series models.

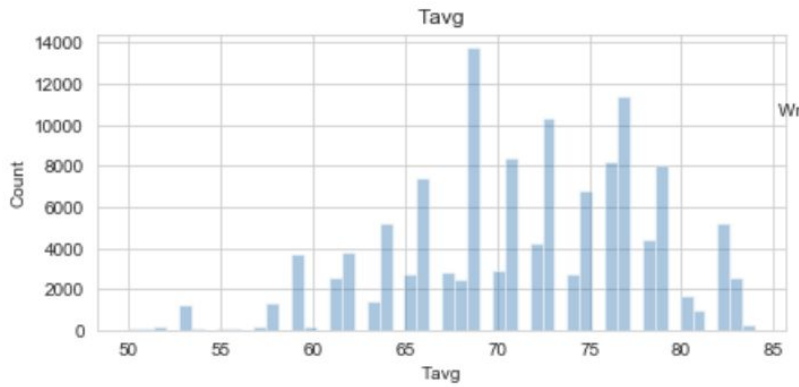
Also, it would be interesting to investigate the time series of temperature and precipitation as well. Especially a comparision of the trend of temperature with the number of mosquitos.

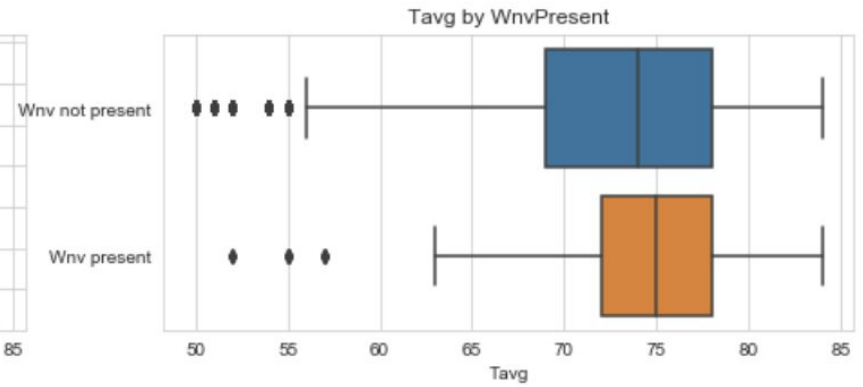
In the difference between the 2 distributions for different features significant?

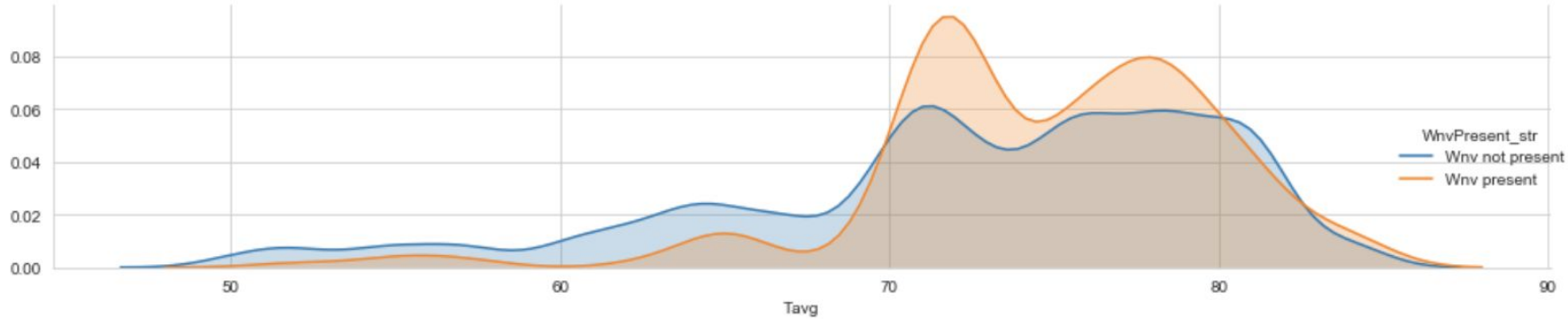
YES.

I performed K-S test and conclude that the features (that we see before) do have statistically different distributions for WNV carriers and normal mosquitoes.

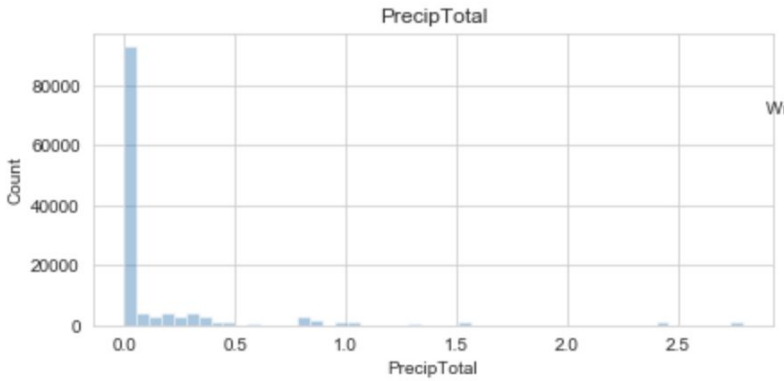
How does temperature affect the number of mosquitoes?

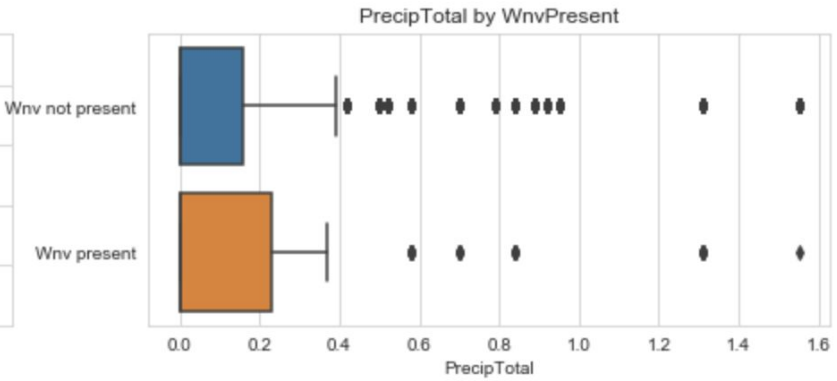


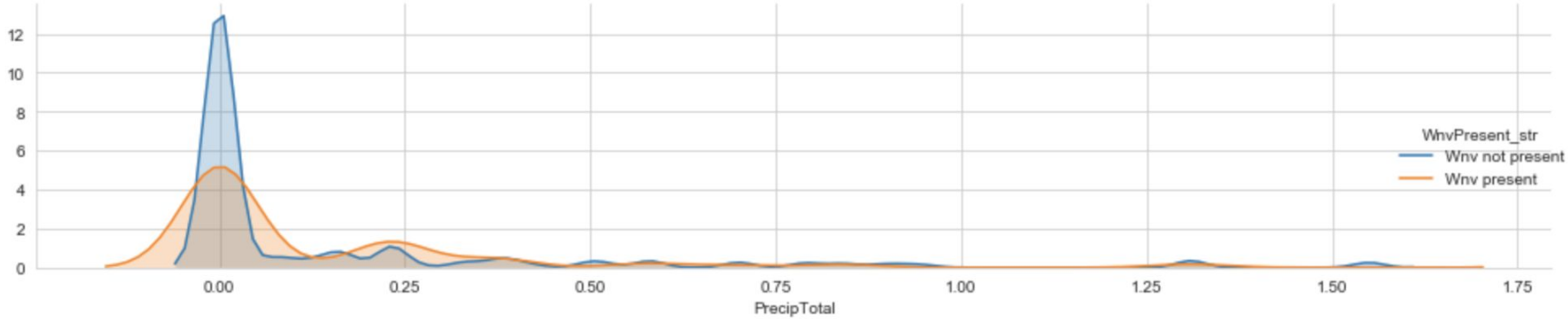




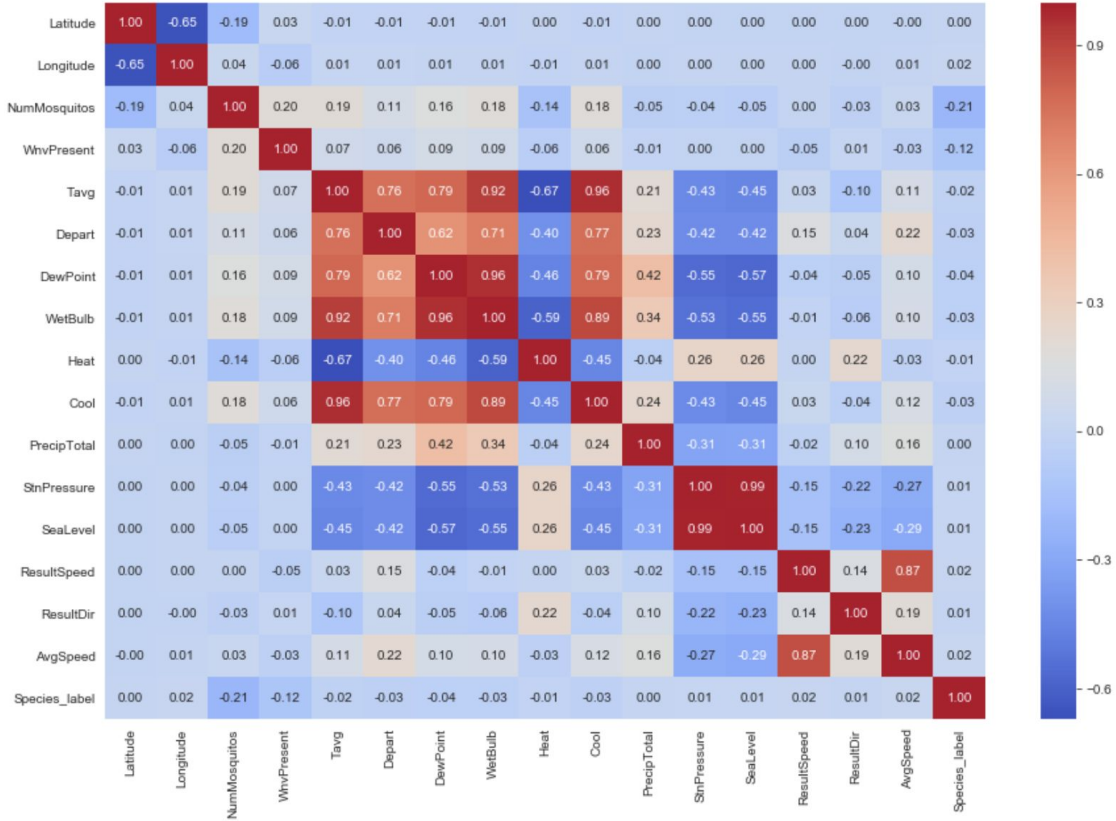
How about precipitation?







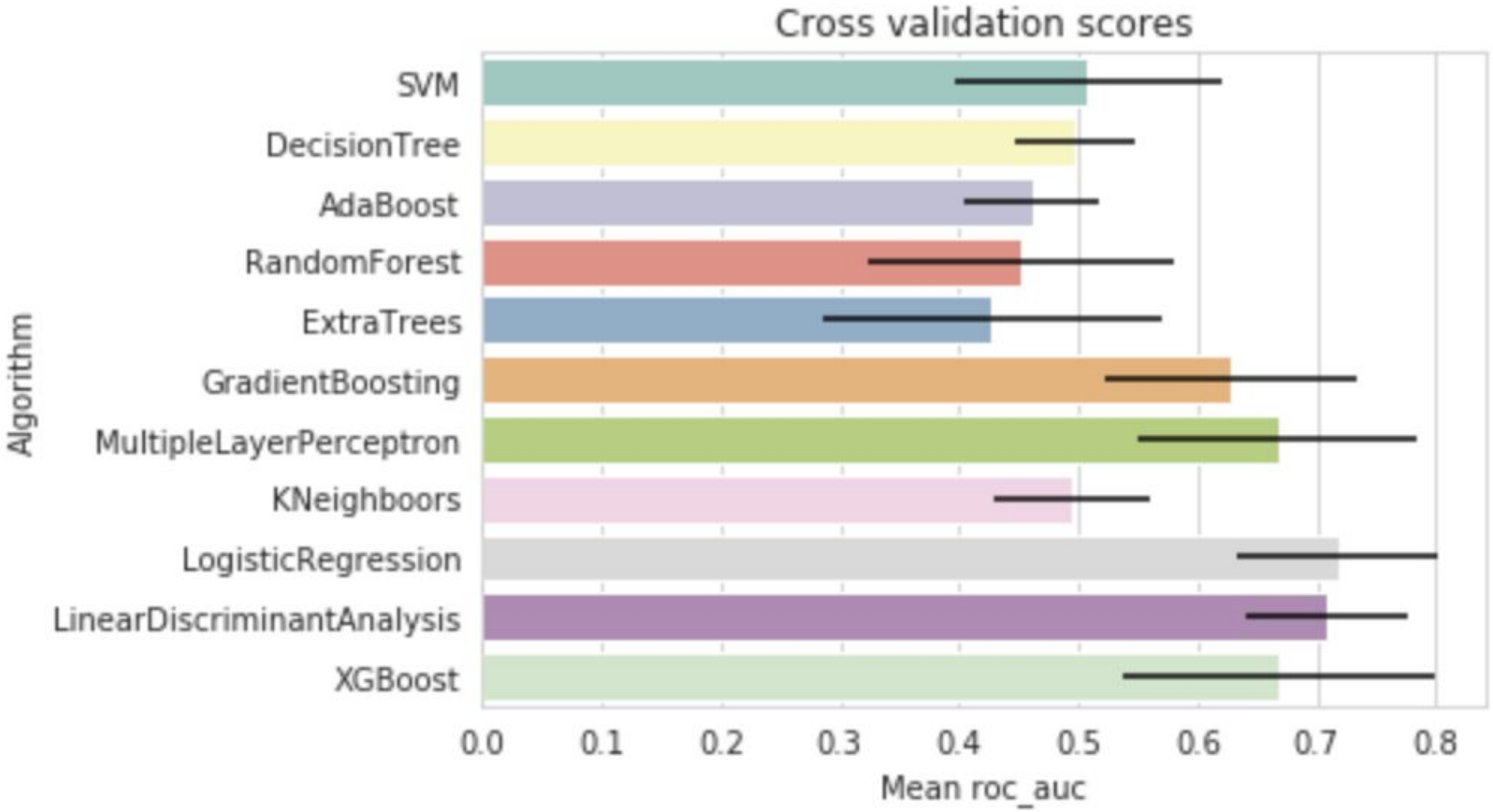
Any correlation between the features?

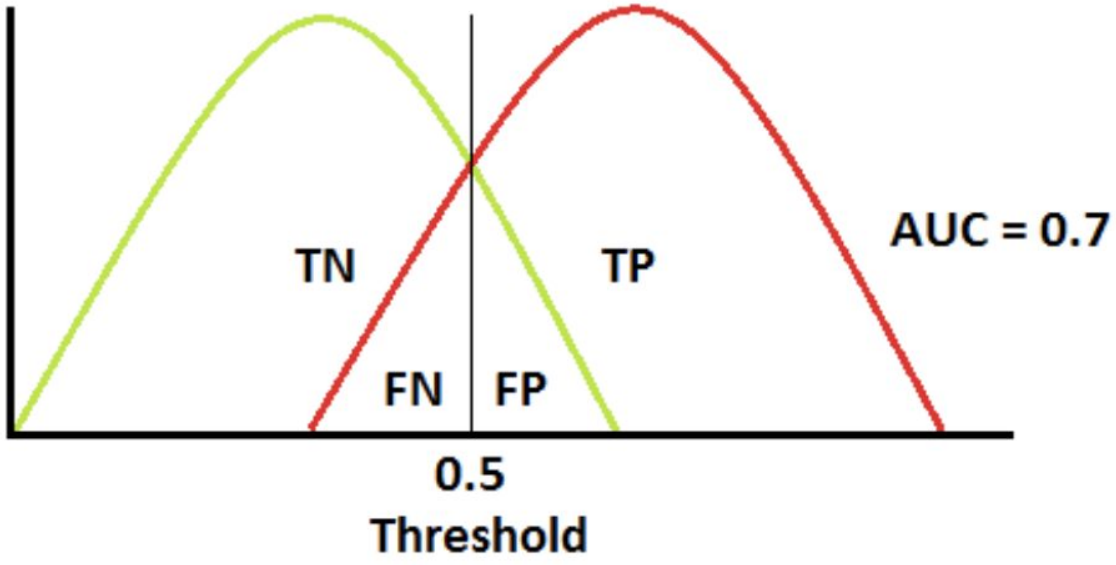


Correlated features:

"Tavg" with: "Depart" (0.76), "DewPoint" (0.79), "WetBulb"(0.92), "Cool" (0.96), "heat" (-0.67).

"Sealevel" with "Stnpressure" (0.99)
"AvgSpeed" with "ResultSpeed" (0.87)


Cross validation scores
SVM
DecisionTree
AdaBoost
RandomForest
ExtraTrees
GradientBoosting
MultipleLayerPerceptron
KNeighboors
LogisticRegression
LinearDiscriminantAnalysis
XGBoost
Algorithm
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Mean roc_auc



AUC = 0.7