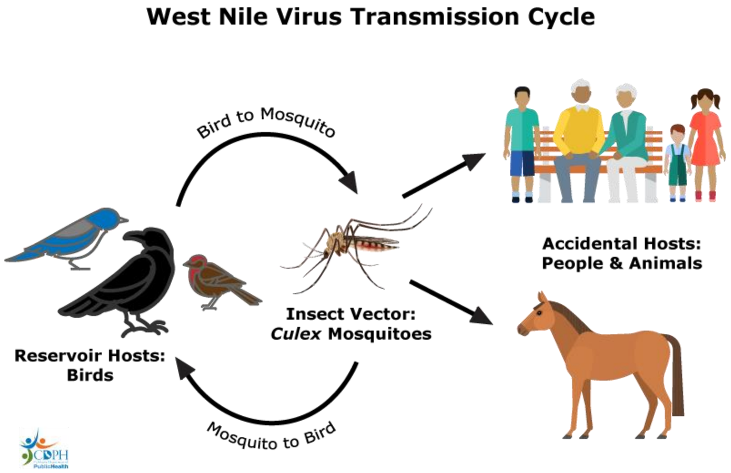
West Nile Virus Prediction

DSI Project 4
Monthon, Peerawat, Piyapon

Problem Statement

WNV have effect on economic in Chicago. So, we will try to predicting WNV
And prevent it to maximize cost benefit for USA government. Furthermore, we will
specify the factor which impact the WNV the most and when to spray in Chicago.

What is West Nile Virus



Transfusions, transplants, and mother-to-child. During WNV transmission season, all donated blood is checked for WNV before being used. The risk of getting WNV through blood transfusions and organ transplants is very small, and should not prevent people from receiving units of blood for medical conditions or for other circumstances (American Red Cross). Transmission during pregnancy from mother-to-baby or transmission to an infant via breastfeeding is extremely rare. **Not through touching.** WNV is not spread through casual contact, such as touching or kissing a person with the virus.

Source: https://www.msmosquito.org/west-nile-virus

effect of West nile virus

Febrile illness

About 1 in 5 people who are infected develop a fever with other symptoms such as headache, vomiting, diarrhea, or rash.

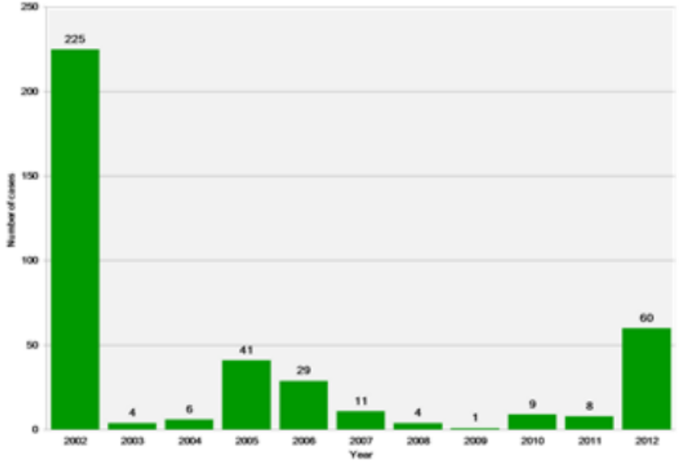
Serious symptoms

About 1 in 150 people who are infected develop a severe illness affecting the central nervous system

Source: https://www.cdc.gov/westnile/symptoms/index.html

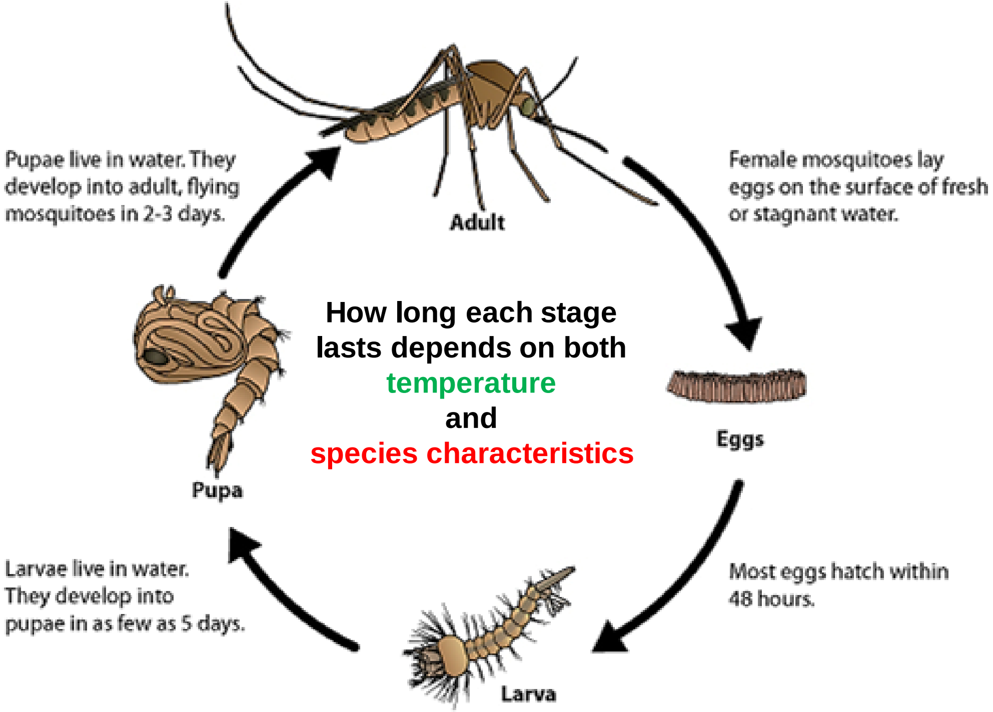
Number of cases 10 years

Figure 1. Number of reported confirmed and probable cases of West Nile virus among Chicago residents by year, 2002-2012.



- In 2002 don’t spray the pesticides therefore, the spike number of WNV cases
- The average 2003 – 2012 is around 17 cases per year

Mosquitos life cycle


Adult
Female mosquitoes lay eggs on the surface of fresh or stagnant water.
Eggs
Most eggs hatch within 48 hours.
Larva
Larvae live in water. They develop into pupae in as few as 5 days.
Pupa
Pupae live in water. They develop into adult, flying mosquitoes in 2-3 days.

How long each stage lasts depends on both temperature and species characteristics

Culex tarsalis,
(a common mosquito in California)

14 days at 70° F
10 days at 80° F.

some species
(naturally adapted)

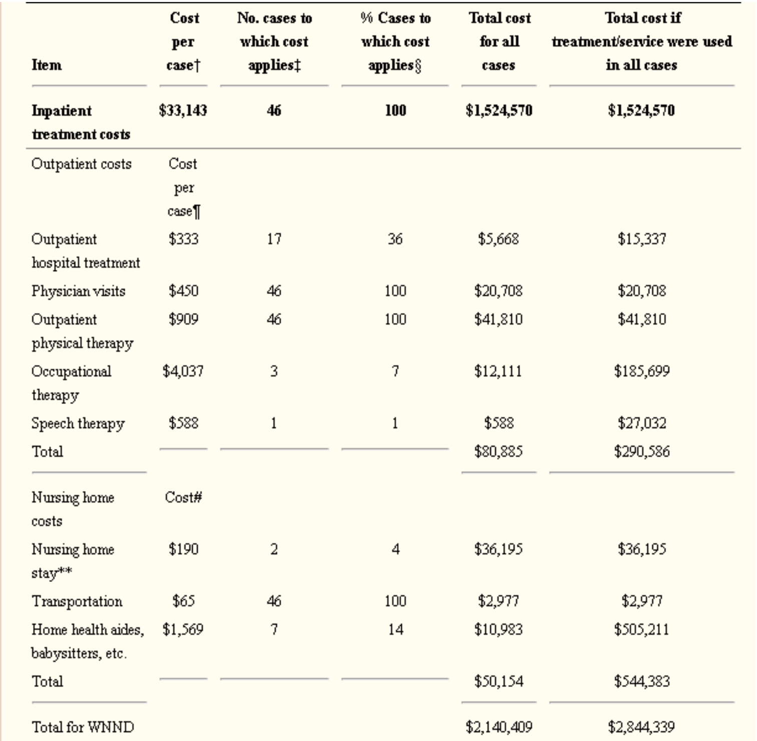
Minimum life cycle : 4 Days
maximum: life cycle : 30 DAYS.

Source: https://www.cdc.gov/dengue/resources/factsheets/mosquitolifecyclefinal.pdf

Data

Cost analysis of WNV

Cost estimate



Item	Cost per case†	No. cases to which cost applies‡	% Cases to which cost applies§	Total cost for all cases	Total cost if treatment/service were used in all cases
Inpatient treatment costs	**$33,143**	**46**	**100**	**$1,524,570**	**$1,524,570**
Outpatient costs	Cost per case¶				
Outpatient hospital treatment	$333	17	36	$5,668	$15,337
Physician visits	$450	46	100	$20,708	$20,708
Outpatient physical therapy	$909	46	100	$41,810	$41,810
Occupational therapy	$4,037	3	7	$12,111	$185,699
Speech therapy	$588	1	1	$588	$27,032
Total				$80,885	$290,586
Nursing home costs	Cost#				
Nursing home stay**	$190	2	4	$36,195	$36,195
Transportation	$65	46	100	$2,977	$2,977
Home health aides, babysitters, etc.	$1,569	7	14	$10,983	$505,211
Total				$50,154	$544,383
Total for WNND				$2,140,409	$2,844,339

A total of 46 WNND cases occurred in Sacramento County in 2005. Costs were ≈$33,143 per inpatient and ≈$6,317 per outpatient for all treatments (Table 2). Cost for each WNND patient estimated to have spent time in a nursing home was ≈$18,097. Productivity loss during symptomatic WNND cost $10,800 per patient <60 years of age and $7,500 per patient >60 years of age (Table 3). Total medical costs accrued by all WNND patients was ≈$2,140,409; total costs for all cases (medical cost plus productivity loss) was ≈$2,844,338.

Benefit of preventing a case of WNV in humans: $27,000 - $133,000 with a mean of $33,000.

Source: https://www.ncbi.nlm.ni
https://www.sciencedirect.com/science/article/pii/S0264410X16311495

Cost analysis of WNV

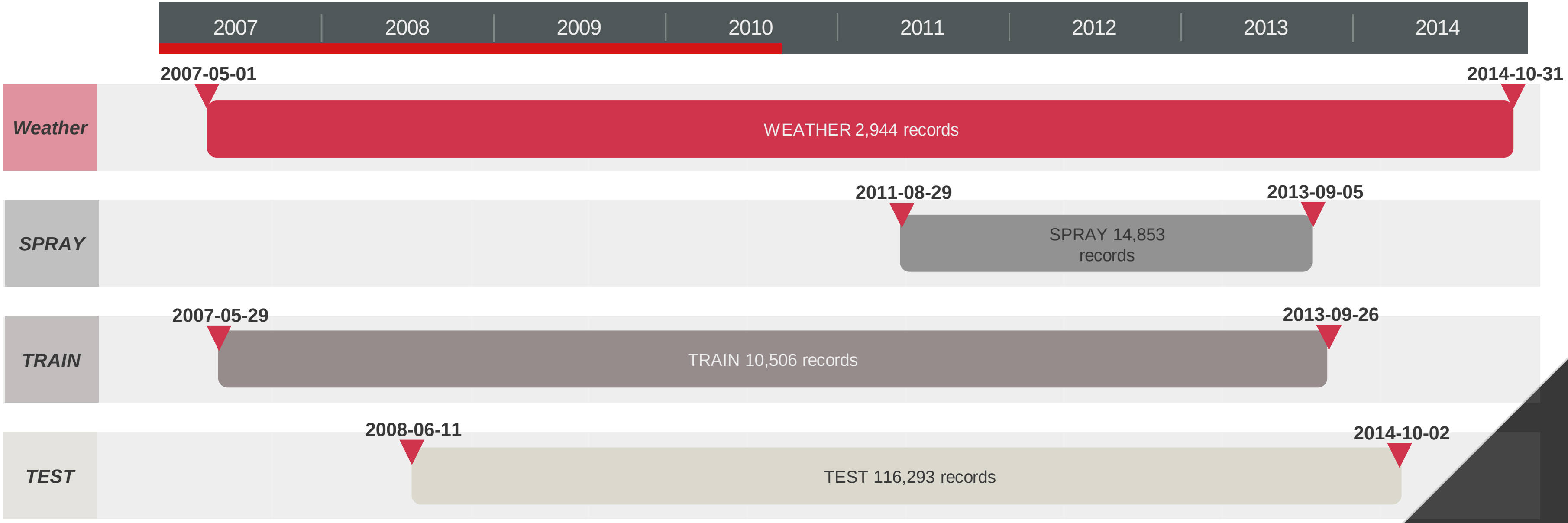
Cost SPRAY

The chemical used is Zenivex, applied at a rate of 1.5 fluid ounces per acre. That measure is approved by the U.S. EPA to control mosquitoes in outdoor residential and recreational areas.

Price is $10,800 per gallon or $4.2 per acres

Source: https://chicago.cbslocal.com/2017/08/30/spray-mosquitoes-far-south-side-west-nile-prevention/

SUMMARY OF DATA



2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014

Weather

2007-05-01

WEATHER 2,944 records

2014-10-31

SPRAY

2011-08-29

SPRAY 14,853 records

2013-09-05

TRAIN

2007-05-29

TRAIN 10,506 records

2013-09-26

TEST

2008-06-11

TEST 116,293 records

2014-10-02

Data Dictionary: clean_train.csv / clean_test.csv

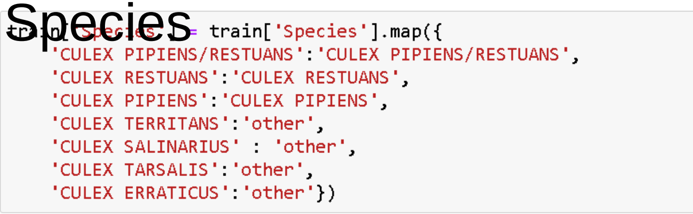
Variables	Description	Example Value
date	Date which investigate trap.	2007-05-29
species	Species of mosquitos	CULEX RESTUANS / other
trap	Unique trap ID	T002, T015
latitude	Latitude of trap	41.954690
longitude	Longitude of trap	-87.800991
addressaccuracy	Accuracy of lat/long	8 / 9
nummosquitos	Number of mosquitos found in trap	1 / 25 / 50
week	Week of the year	0 / 24 / 52

Data Dictionary: clean_weather.csv

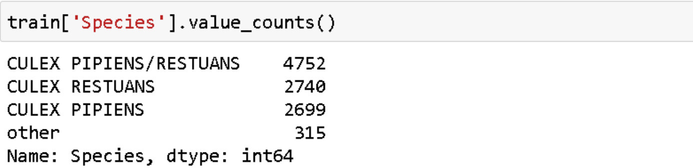
Variables	Description	Example Value
tavg	Temperature average on that day	45.0 / 67.0
depart	measure of climate change but tells us nothing about the effects of climate change.	-3 / 6 / 14
dewpoint	temperature to which air must be cooled to become saturated	29 / 35 / 51
heat	measure of how hot it really feels when relative humidity is factored in with the actual air temperature.	0 / 9 / 23
cool	measure of how cool it really feels when relative humidity is factored in with the actual air temperature.	0 / 6 / 25
dewpoint	the air needs to be cooled to (at constant pressure) in order to achieve a relative humidity	27 / 50 / 74
sealevel	the atmospheric pressure at sea level at a given location	29.23 / 30.05
averagespeed	Average speed of the wind	3.9 / 14.5 / 17.8
codesum	Weather event. For instance, SN – snow. We convert into 0 if no event and 1 if any event occur	0 / 1
snowfall	Height of snow	0 / 0.005 / 0.1
preciptotal	Measurement of water – rain / snow / blizzard / etc.	0.000 / 0.030 / 0.040
resultdir	Wind direction	2 / 4 / 25 / 27

DATA CLEANING

TRAIN - Species

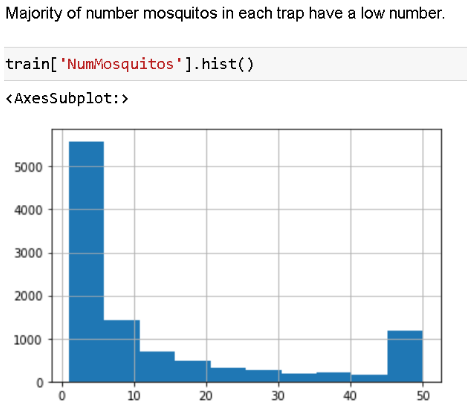


```
train['Species'] = train['Species'].map({
    'CULEX PIPIENS/RESTUANS':'CULEX PIPIENS/RESTUANS',
    'CULEX RESTUANS':'CULEX RESTUANS',
    'CULEX PIPIENS':'CULEX PIPIENS',
    'CULEX TERRITANS':'other',
    'CULEX SALINARIUS' : 'other',
    'CULEX TARSALIS':'other',
    'CULEX ERRATICUS':'other'})
```



```
train['Species'].value_counts()
```

```
CULEX PIPIENS/RESTUANS    4752
CULEX RESTUANS            2740
CULEX PIPIENS             2699
other                      315
Name: Species, dtype: int64
```

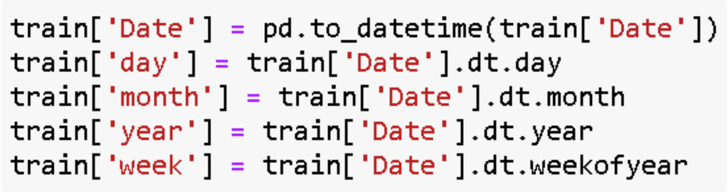


Majority of number mosquitos in each trap have a low number.

```
train['NumMosquitos'].hist()
```

```
<AxesSubplot:>
```

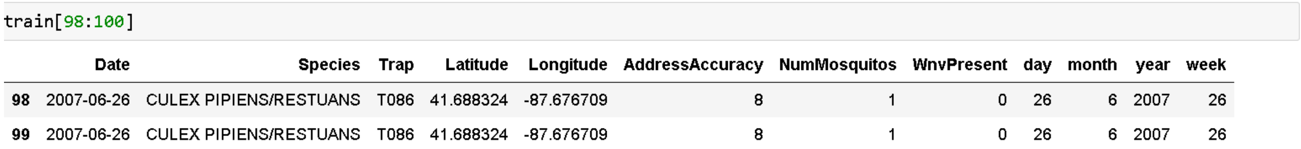
TRAIN – Date



```
train['Date'] = pd.to_datetime(train['Date'])
train['day'] = train['Date'].dt.day
train['month'] = train['Date'].dt.month
train['year'] = train['Date'].dt.year
train['week'] = train['Date'].dt.weekofyear
```

DATA CLEANING

TRAIN – duplicated row



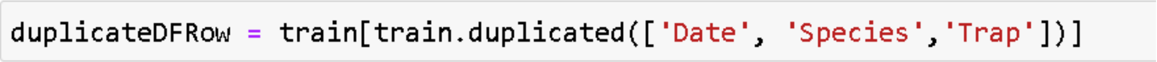
```
train[98:100]
```

	Date	Species	Trap	Latitude	Longitude	AddressAccuracy	NumMosquitos	WnvPresent	day	month	year	week
98	2007-06-26	CULEX PIPIENS/RESTUANS	T086	41.688324	-87.676709	8	1	0	26	6	2007	26
99	2007-06-26	CULEX PIPIENS/RESTUANS	T086	41.688324	-87.676709	8	1	0	26	6	2007	26

- If number of mosquitos go above 50 will generate new record
- Some are 2 record in same day with 4 nummosquitos and 6 nummosquitos

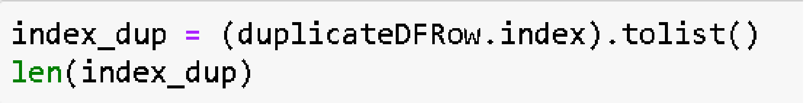
DATA CLEANING

TRAIN – duplicated row



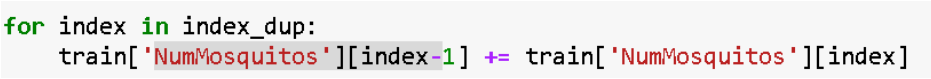
```
duplicateDFRow = train[train.duplicated(['Date', 'Species','Trap'])]
```

- Looking for duplicate row



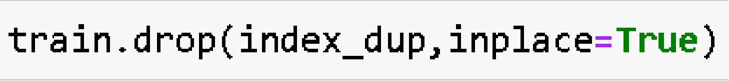
```
index_dup = (duplicateDFRow.index).tolist()
len(index_dup)
```

- Take the index



```
for index in index_dup:
    train['NumMosquitos'][index-1] += train['NumMosquitos'][index]
```

- Sum number of duplicated row together

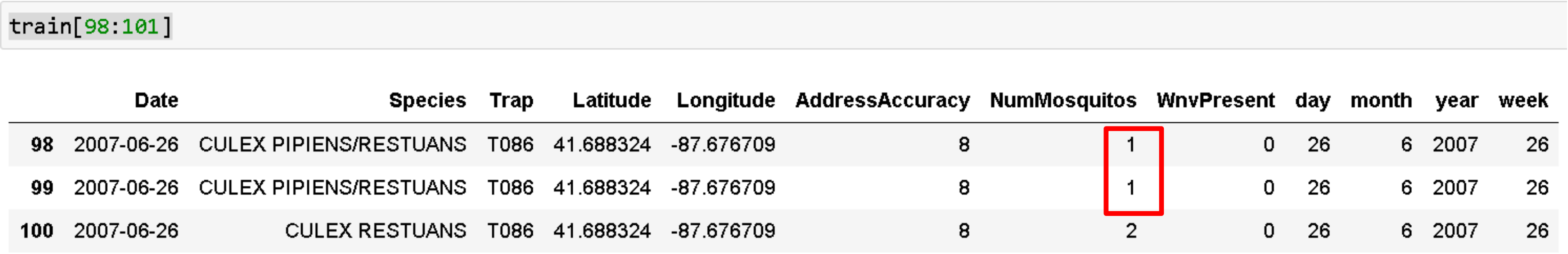


```
train.drop(index_dup,inplace=True)
```

- Drop index which duplicated

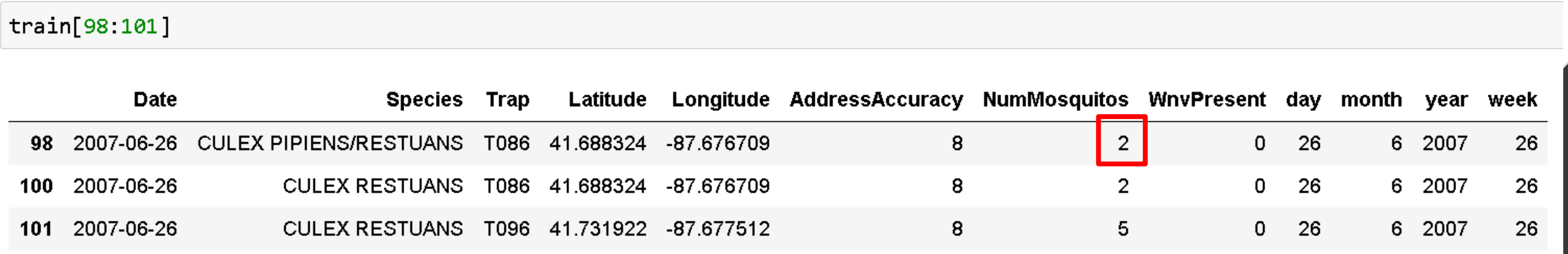
DATA CLEANING

TRAIN – duplicated row



```
train[98:101]
```

	Date	Species	Trap	Latitude	Longitude	AddressAccuracy	NumMosquitos	WnvPresent	day	month	year	week
98	2007-06-26	CULEX PIPIENS/RESTUANS	T086	41.688324	-87.676709	8	1	0	26	6	2007	26
99	2007-06-26	CULEX PIPIENS/RESTUANS	T086	41.688324	-87.676709	8	1	0	26	6	2007	26
100	2007-06-26	CULEX RESTUANS	T086	41.688324	-87.676709	8	2	0	26	6	2007	26

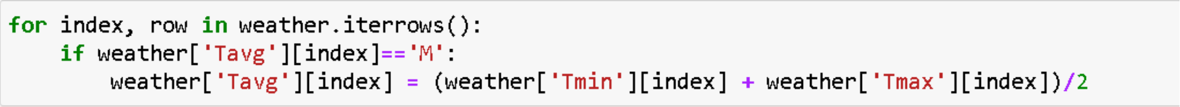


```
train[98:101]
```

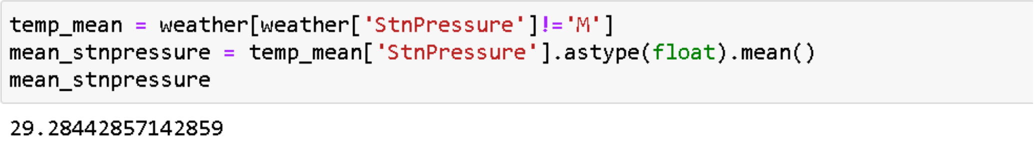
	Date	Species	Trap	Latitude	Longitude	AddressAccuracy	NumMosquitos	WnvPresent	day	month	year	week
98	2007-06-26	CULEX PIPIENS/RESTUANS	T086	41.688324	-87.676709	8	2	0	26	6	2007	26
100	2007-06-26	CULEX RESTUANS	T086	41.688324	-87.676709	8	2	0	26	6	2007	26
101	2007-06-26	CULEX RESTUANS	T096	41.731922	-87.677512	8	5	0	26	6	2007	26

DATA CLEANING

WEATHER – Missing value

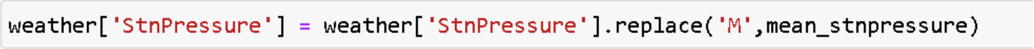


```python
for index, row in weather.iterrows():
    if weather['Tavg'][index]=='M':
        weather['Tavg'][index] = (weather['Tmin'][index] + weather['Tmax'][index])/2
```



```python
temp_mean = weather[weather['StnPressure']!='M']
mean_stnpressure = temp_mean['StnPressure'].astype(float).mean()
mean_stnpressure
```

```
29.28442857142859
```



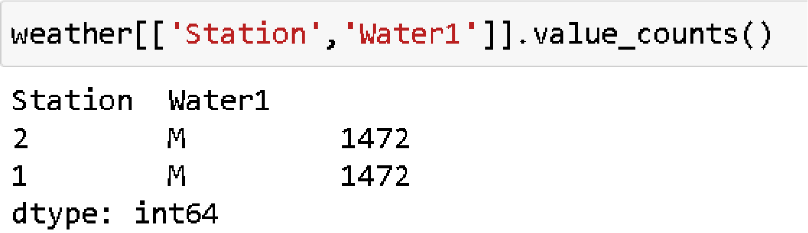
```python
weather['StnPressure'] = weather['StnPressure'].replace('M',mean_stnpressure)
```

1. Filling with average value / mean value

- Tavg
- StnPressure
- SeaLevel
- AvgSpeed

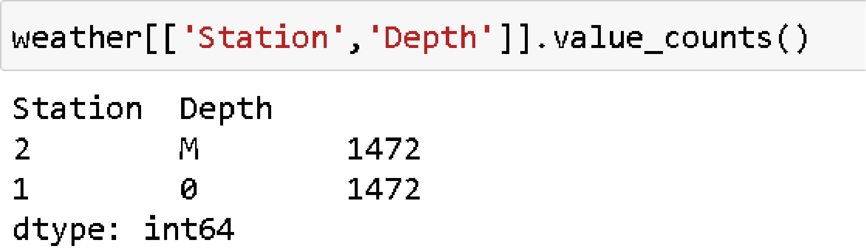
DATA CLEANING

WEATHER – Missing value



```
weather[['Station','Water1']].value_counts()
```

```
Station  Water1
2        M         1472
1        M         1472
dtype: int64
```



```
weather[['Station','Depth']].value_counts()
```

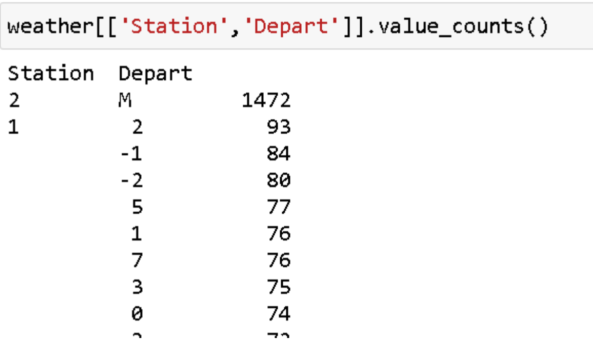
```
Station  Depth
2        M         1472
1        0         1472
dtype: int64
```

2. Dropout columns due to it is all missing value

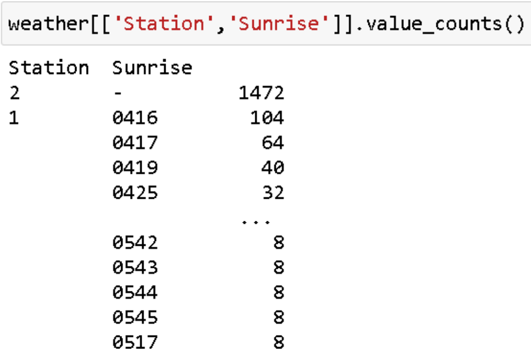
- Water1
- Depth

DATA CLEANING

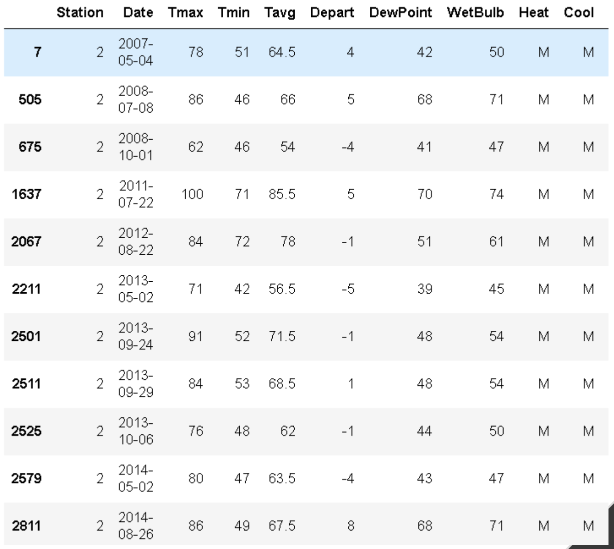
WEATHER – Missing value



```
weather[['Station','Depart']].value_counts()

Station  Depart
2        M          1472
1          2          93
          -1          84
          -2          80
           5          77
           1          76
           7          76
           3          75
           0          74
```



```
weather[['Station','Sunrise']].value_counts()

Station  Sunrise
2        -          1472
1        0416        104
         0417         64
         0419         40
         0425         32
                    ...
         0542          8
         0543          8
         0544          8
         0545          8
         0517          8
```



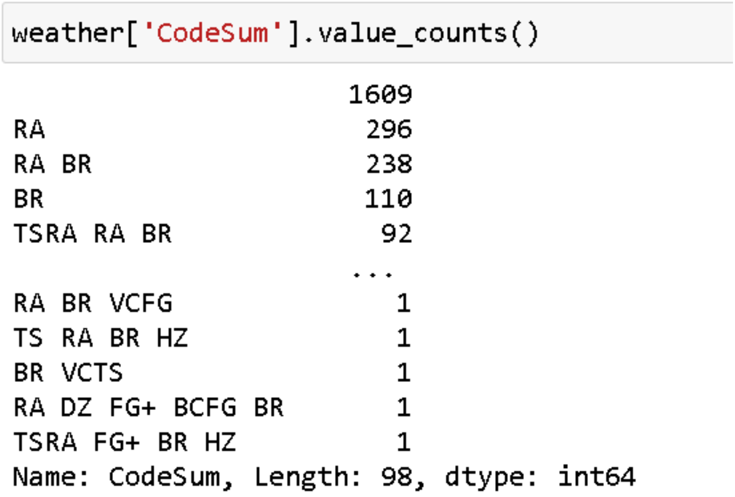
	Station	Date	Tmax	Tmin	Tavg	Depart	DewPoint	WetBulb	Heat	Cool
7	2	2007-05-04	78	51	64.5	4	42	50	M	M
505	2	2008-07-08	86	46	66	5	68	71	M	M
675	2	2008-10-01	62	46	54	-4	41	47	M	M
1637	2	2011-07-22	100	71	85.5	5	70	74	M	M
2067	2	2012-08-22	84	72	78	-1	51	61	M	M
2211	2	2013-05-02	71	42	56.5	-5	39	45	M	M
2501	2	2013-09-24	91	52	71.5	-1	48	54	M	M
2511	2	2013-09-29	84	53	68.5	1	48	54	M	M
2525	2	2013-10-06	76	48	62	-1	44	50	M	M
2579	2	2014-05-02	80	47	63.5	-4	43	47	M	M
2811	2	2014-08-26	86	49	67.5	8	68	71	M	M

3. Filling with other station

- Depart
- Heat / Cool
- Sunset / Sunrise
- Wetbulb

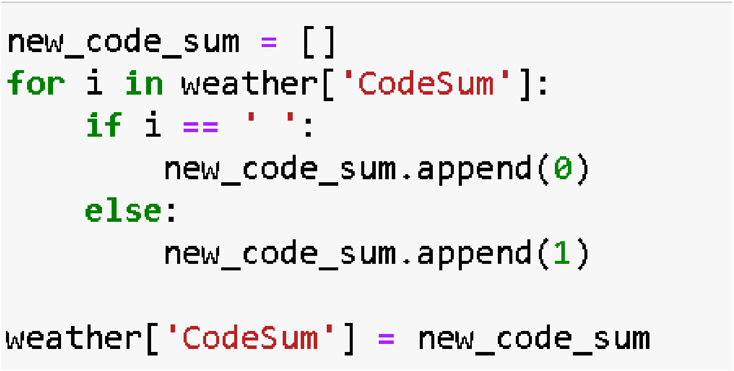
DATA CLEANING

WEATHER – CodeSum



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

```
                      1609
RA                     296
RA BR                  238
BR                     110
TSRA RA BR              92
                      ...
RA BR VCFG               1
TS RA BR HZ              1
BR VCTS                  1
RA DZ FG+ BCFG BR        1
TSRA FG+ BR HZ           1
Name: CodeSum, Length: 98, dtype: int64
```

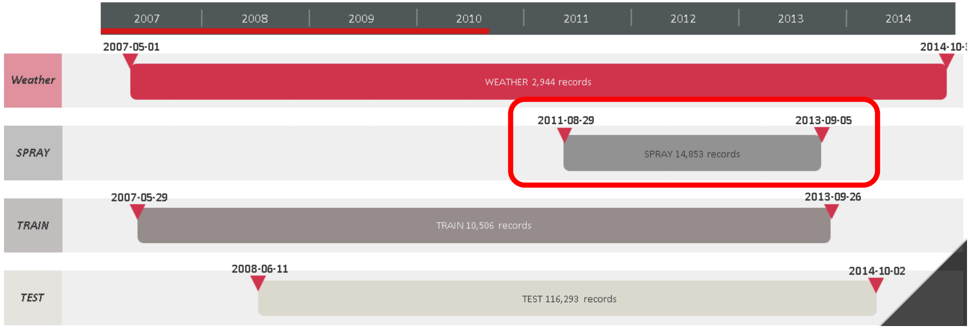


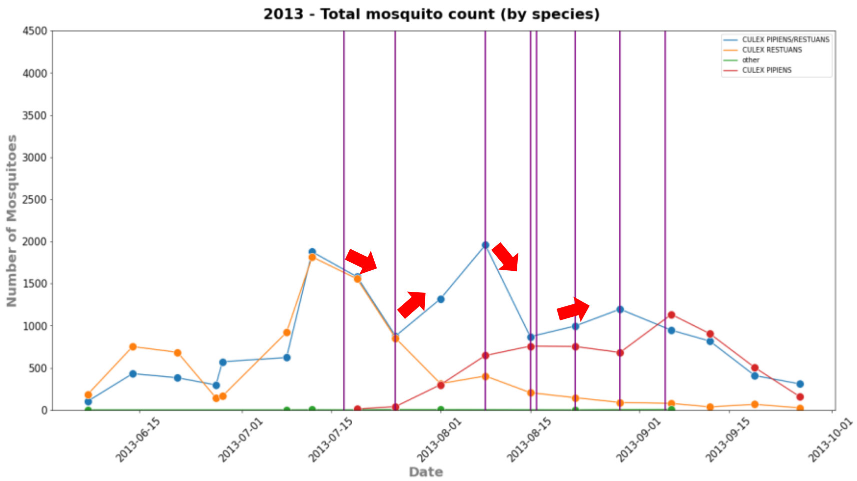
```
new_code_sum = []
for i in weather['CodeSum']:
    if i == ' ':
        new_code_sum.append(0)
    else:
        new_code_sum.append(1)

weather['CodeSum'] = new_code_sum
```

CodeSum represent event that occur in that day whether it be, snow, rain, duststrom, freezing, etc. Therefore, we will mapping into 0 if no event occur in that they, and 1 if any event occur

EDA

Why we not use SPRAY.csv

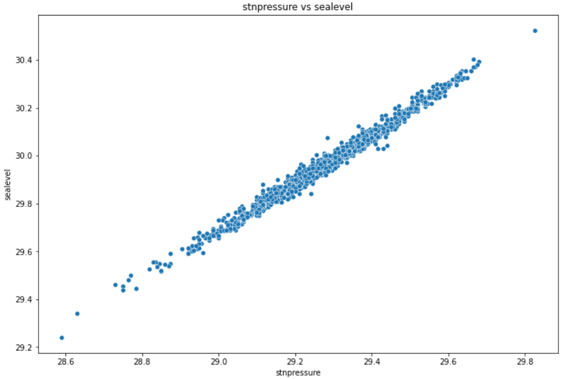


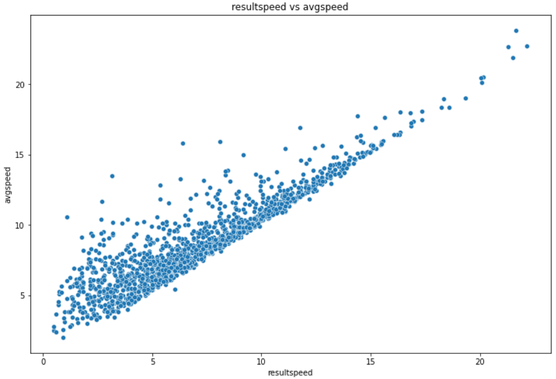


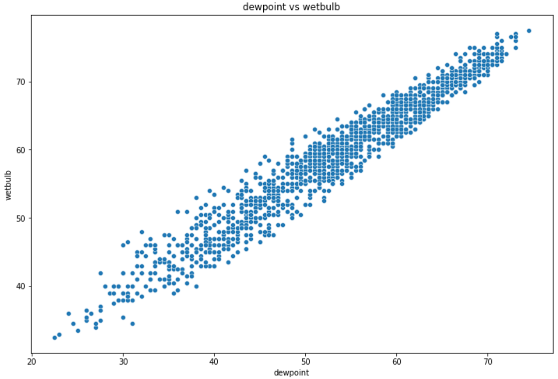
1. Data train occur 5 years and spray have only 2 years if we merge together data will be lose.

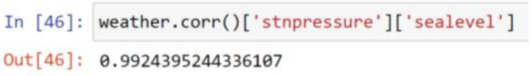
2. Other reason is as you can see on spray record after spray in area number of mosquitos sometimes increase sometimes decrease. So, we think spray will not significantly effect

Data Correlate redundant variable?

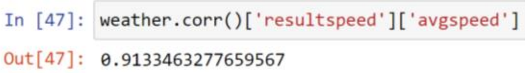




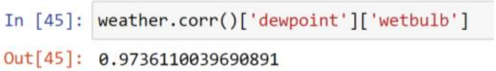




```
In [46]: weather.corr()['stnpressure']['sealevel']
Out[46]: 0.9924395244336107
```



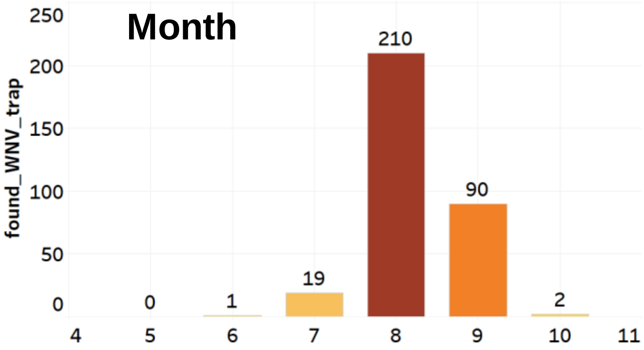
```
In [47]: weather.corr()['resultspeed']['avgspeed']
Out[47]: 0.9133463277659567
```

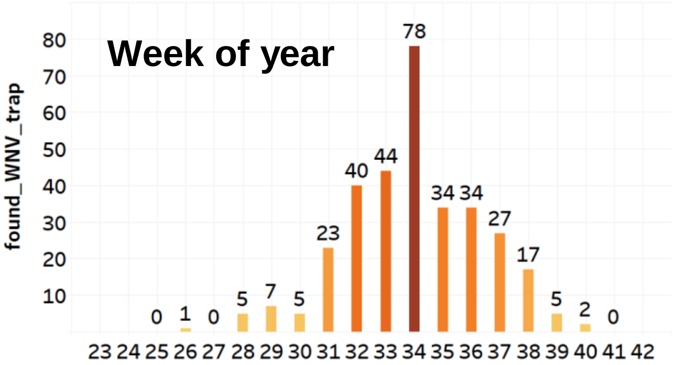


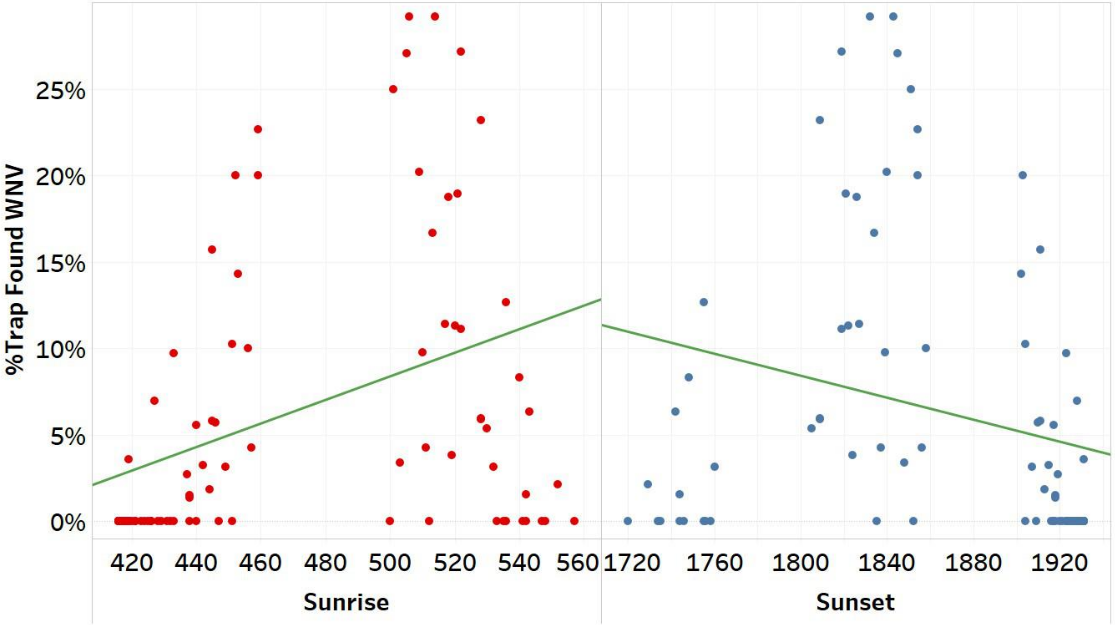
```
In [45]: weather.corr()['dewpoint']['wetbulb']
Out[45]: 0.9736110039690891
```

- We drop tmax / tmin becuase using only tavg to decrease redundant data
- We drop wetbulb because have a strong correlate with dewpoint (to decrease number of feature).
- We drop stnpressure because have a strong correlate with sealevel (to decrease number of feature).
- We drop resultspeed because have a strong correlate with averagespeed (to decrease number of feature).

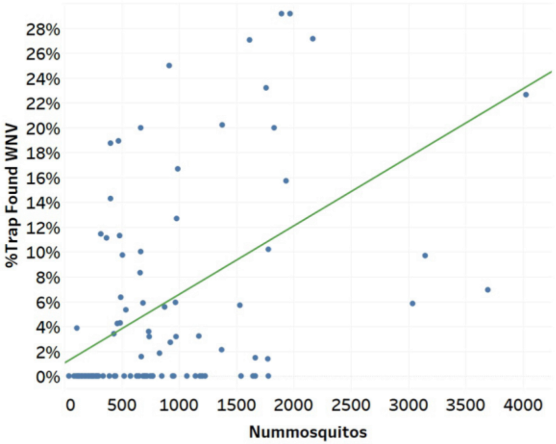
Most effected feature



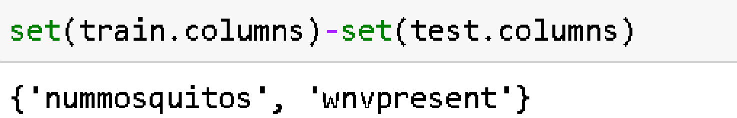




Number of Mosquitos V.S. WNVpresent



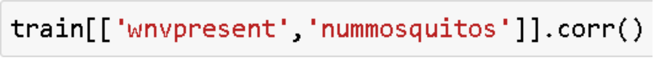
Need Prediction of Number of Mosquitos in **Test.csv**



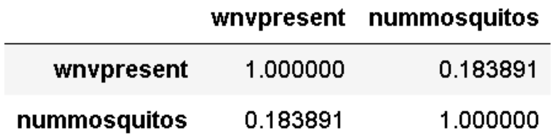
```
set(train.columns)-set(test.columns)
```

```
{'nummosquitos', 'wnvpresent'}
```

Due to we looking through the number of mosquitos have some relation on WNV. We will try to predict number of mosquitos on file test and using it as a feature in model.



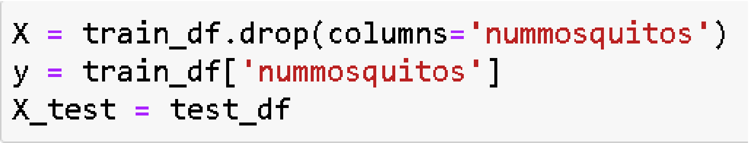
```
train[['wnvpresent','nummosquitos']].corr()
```



	wnvpresent	nummosquitos
wnvpresent	1.000000	0.183891
nummosquitos	0.183891	1.000000

NUMMOSQuITOS

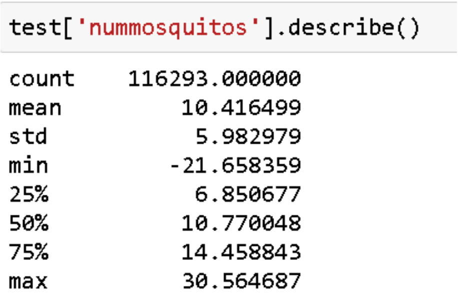
Predicting numosquitos on test file



```
X = train_df.drop(columns='nummosquitos')
y = train_df['nummosquitos']
X_test = test_df
```

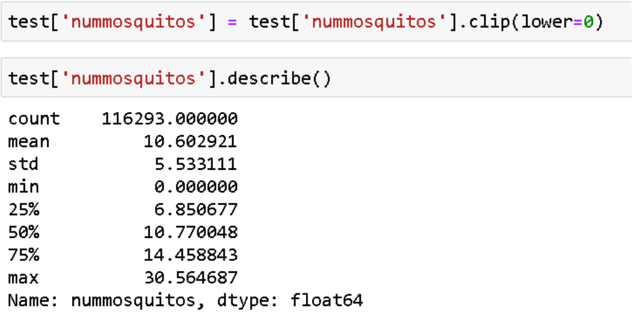
Set target variable as a nummosquitos and the rest as a predictors.

After try on LinearRegression, Ridge, Lasso. Ridge perform the best (least error). However, it’s some wrong prediction number of mosquitos can’t be negative. There fore, I map lower than 0 to be 0 and round up to be number of mosquitos on test file



```
test['nummosquitos'].describe()
```

```
count    116293.000000
mean         10.416499
std           5.982979
min         -21.658359
25%           6.850677
50%          10.770048
75%          14.458843
max          30.564687
```



```
test['nummosquitos'] = test['nummosquitos'].clip(lower=0)
```

```
test['nummosquitos'].describe()
```

```
count    116293.000000
mean         10.602921
std           5.533111
min           0.000000
25%           6.850677
50%          10.770048
75%          14.458843
max          30.564687
Name: nummosquitos, dtype: float64
```

Model & Tuning

Maximize Sensitivity

If we predicting that area not have WNV but it’s actually have (False negative)

- **Cost that government pay to cure patient and improductive $ 33,000 per cases**
- **While if we spray in all Chicago areas the costs will be $610,260 ($4.2 per acre)**

With out Spray

2002 with out Spray, found 225 WNV cases

Cost of ameliorate and cure is
225 x $33,000 = **$7,425,000**

With Spray

Total Area Chicago 145,300 acres
Cost for Spray
145,300 x 4.2= $610,260
Not included wage

Average WMV case in (2003-2012)
is 17 cases per year
Cost of ameliorate and cure is
17 x $33,000 = $561,000
Total cost is **$1,171,260**

Source: https://www.chicago.gov/content/dam/city/depts/cdph/statistics_and_reports/CDInfo_2013_JULY_WNV.pdf
https://www.sciencedirect.com/science/article/pii/S0264410X16311495

CLASSIFICATION MODEL

Goal : Minimize False Negative

Minimize **wrong prediction that West Nile virus is not present but actually it’s occur in that area**

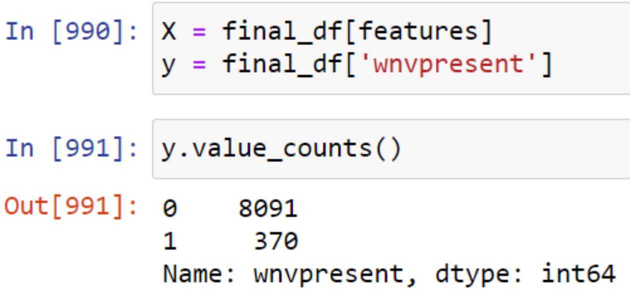
CLASSIFICATION MODEL

West Nile Virus Present

0: West Nile Virus Not Present

1: West Nile Virus Present

Assign Variable

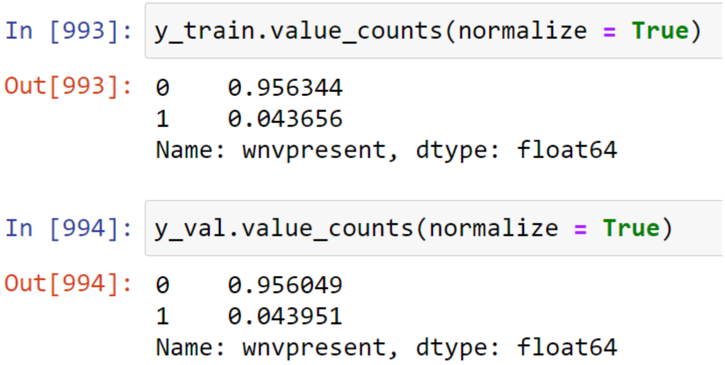


```
In [990]: X = final_df[features]
          y = final_df['wnvpresent']

In [991]: y.value_counts()

Out[991]: 0    8091
          1     370
          Name: wnvpresent, dtype: int64
```

Split Data

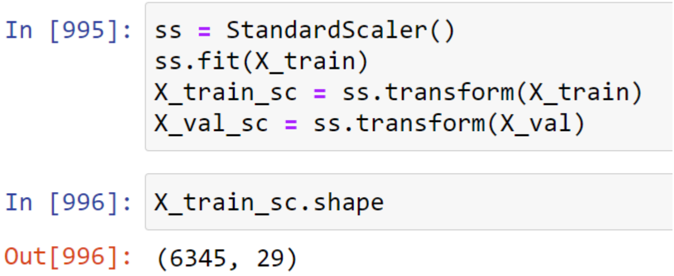
- X_train
- X_val (X_test of train data)
- y_train
- y_val (y_test of train data)

Target Variable



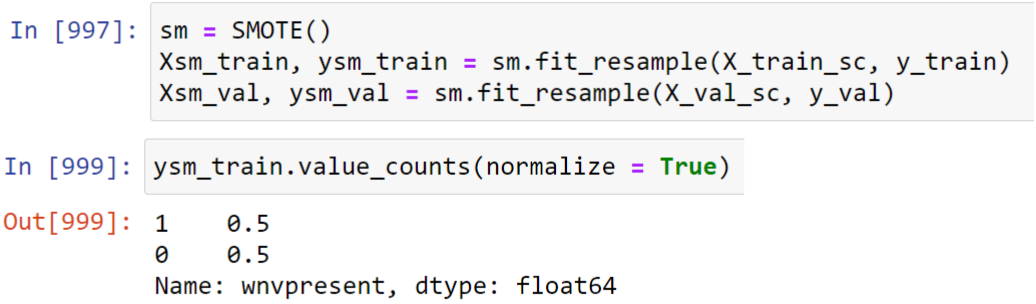
```
In [993]: y_train.value_counts(normalize = True)

Out[993]: 0    0.956344
          1    0.043656
          Name: wnvpresent, dtype: float64

In [994]: y_val.value_counts(normalize = True)

Out[994]: 0    0.956049
          1    0.043951
          Name: wnvpresent, dtype: float64
```

Standardization



```
In [995]: ss = StandardScaler()
          ss.fit(X_train)
          X_train_sc = ss.transform(X_train)
          X_val_sc = ss.transform(X_val)

In [996]: X_train_sc.shape

Out[996]: (6345, 29)
```

Balance Target Variable



```
In [997]: sm = SMOTE()
          Xsm_train, ysm_train = sm.fit_resample(X_train_sc, y_train)
          Xsm_val, ysm_val = sm.fit_resample(X_val_sc, y_val)

In [999]: ysm_train.value_counts(normalize = True)

Out[999]: 1    0.5
          0    0.5
          Name: wnvpresent, dtype: float64
```

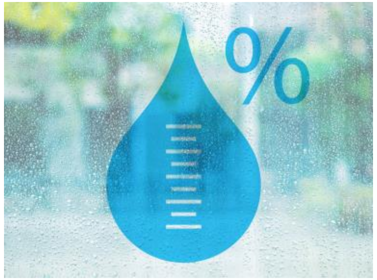
Model 1 : All features in file train and weather

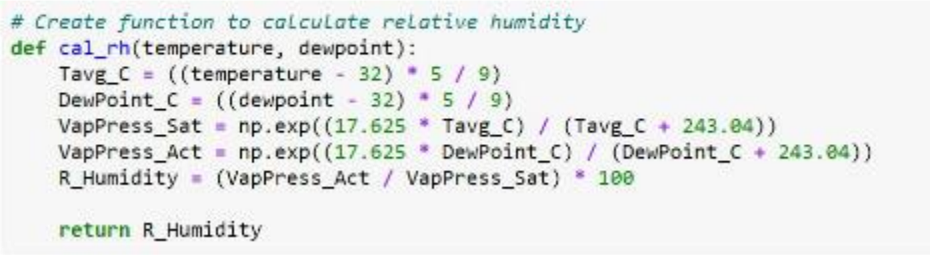
Model & accuracy

Model	Train Score	CrossVal Score	Test Score	AUC	Sensitivity	False Negative	Kaggle Score
Baseline	50%	50%	50%	-	-	-	
Adaboost	89.4%	88.7%	88.7%	0.95	91.3%	176	0.668
LR	80.3%	80.1%	83.6%	0.90	89.9%	205	0.732
Naive-Bayes	69.9%	70.1%	69.7%	0.82	73.4%	538	0.668
Random Forest	100%	96.2%	83.7%	0.95	71.4%	579	0.733
KNeighbors	95.6%	93.4%	76.4%	0.81	63.0%	747	0.584

Feature engineering

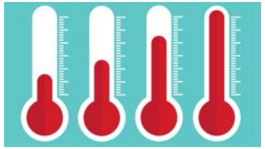
1. Add Humidity

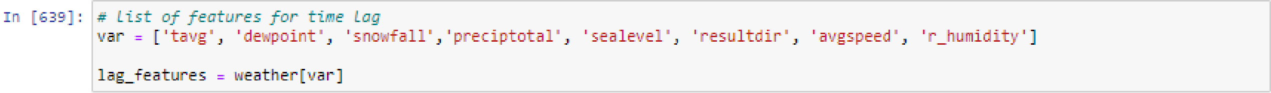




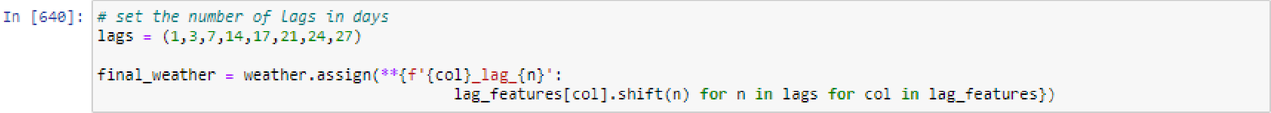
```
# Create function to calculate relative humidity
def cal_rh(temperature, dewpoint):
    Tavg_C = ((temperature - 32) * 5 / 9)
    DewPoint_C = ((dewpoint - 32) * 5 / 9)
    VapPress_Sat = np.exp((17.625 * Tavg_C) / (Tavg_C + 243.04))
    VapPress_Act = np.exp((17.625 * DewPoint_C) / (DewPoint_C + 243.04))
    R_Humidity = (VapPress_Act / VapPress_Sat) * 100

    return R_Humidity
```

2. Add Lagging of time of features





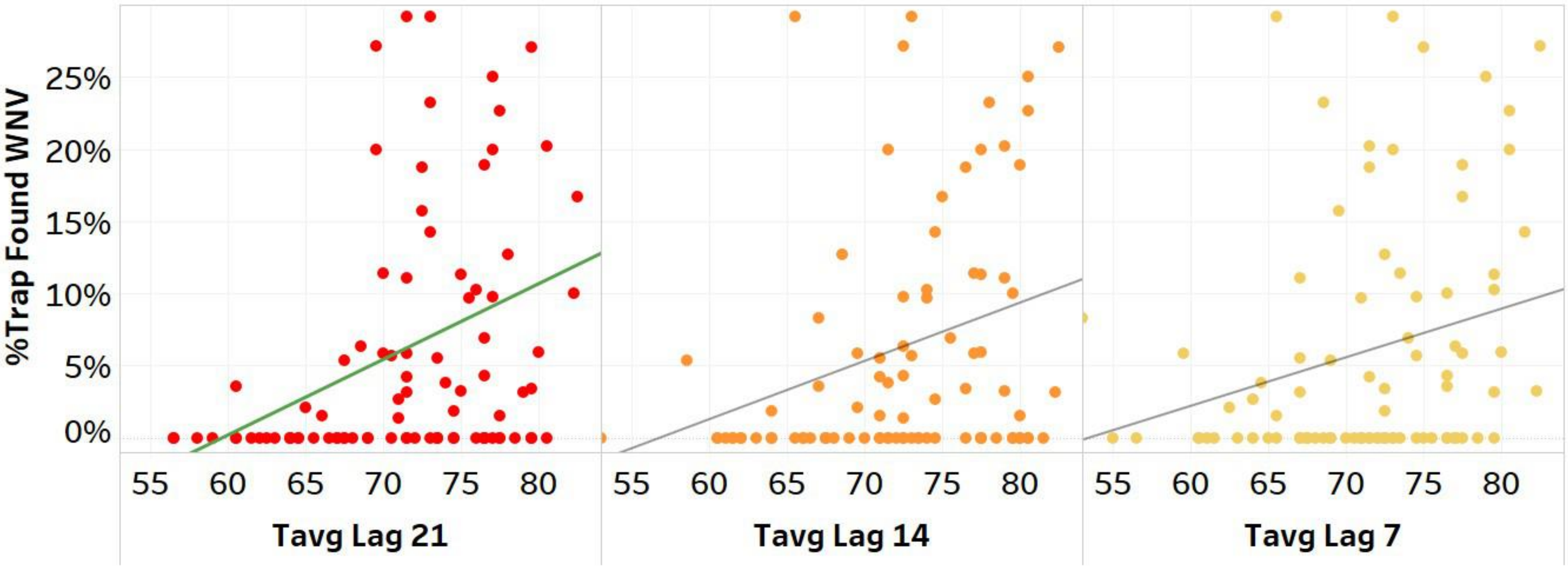
```
In [639]: # list of features for time lag
          var = ['tavg', 'dewpoint', 'snowfall','preciptotal', 'sealevel', 'resultdir', 'avgspeed', 'r_humidity']

          lag_features = weather[var]
```



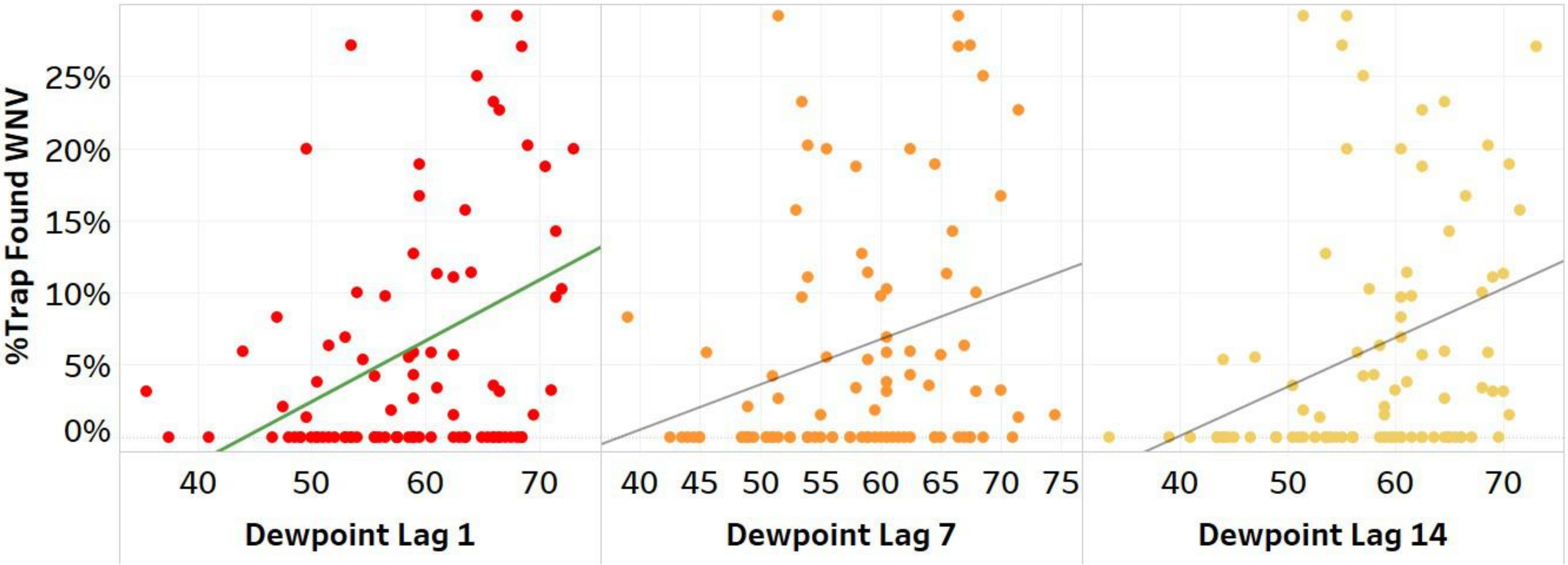
```
In [640]: # set the number of lags in days
          lags = (1,3,7,14,17,21,24,27)

          final_weather = weather.assign(**{f'{col}_lag_{n}':
                                              lag_features[col].shift(n) for n in lags for col in lag_features})
```

Source: https://sciencing.com/relative-humidity-7611453.html

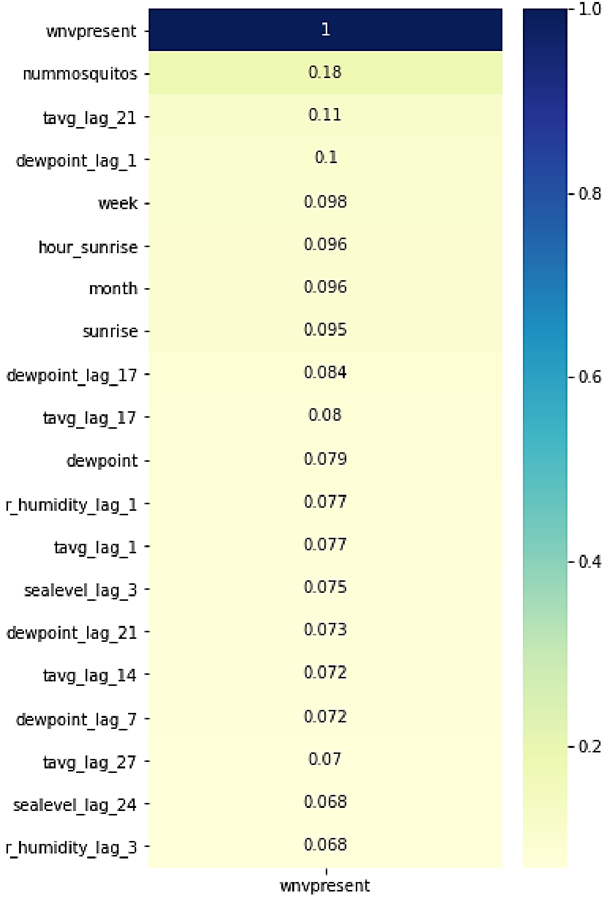
Tavg V.S. WNVpresent



Dewpoint V.S. WNVpresent



High correlate Features – (top 40 Features use in Model)



Number of mosquitos

Dewpoint previous 1 day

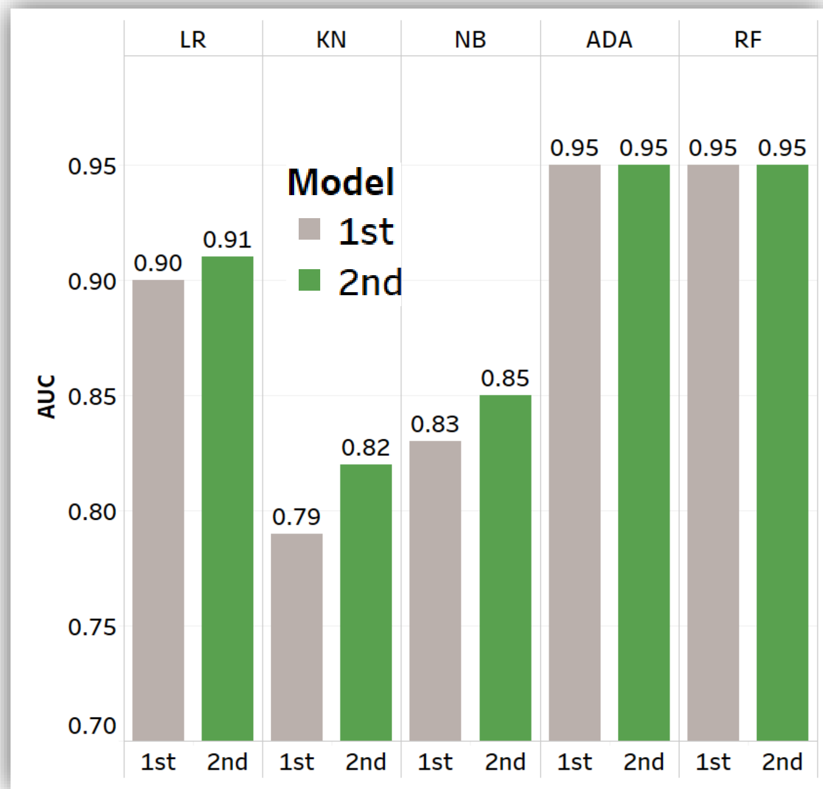
Temperature previous 21 day

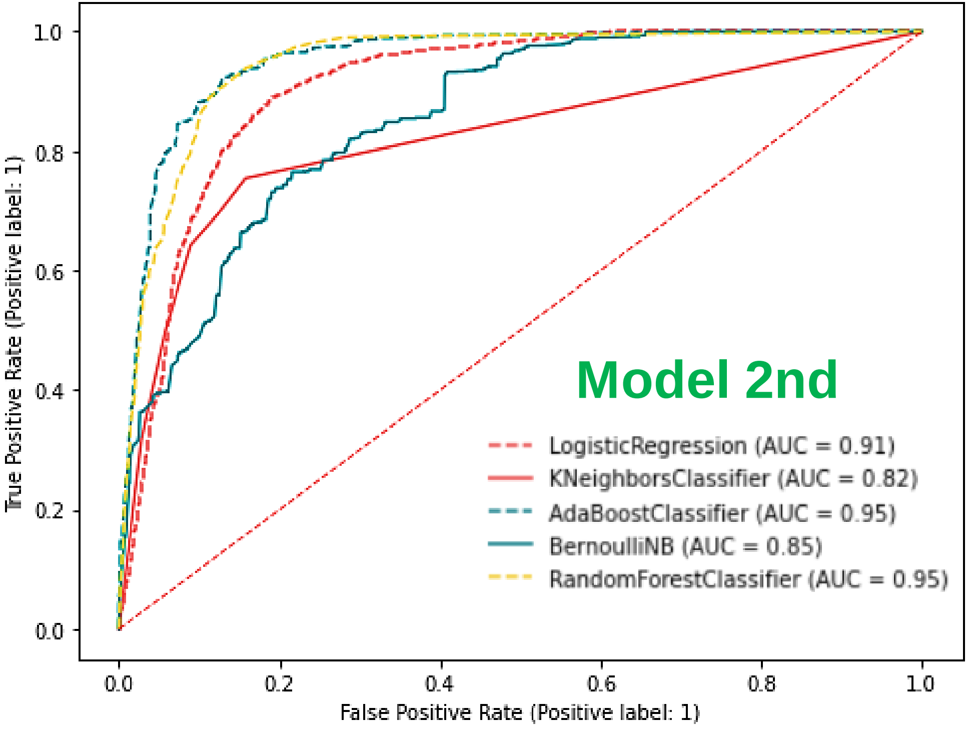
Humidity previous 1 day

ROC CURVE

- Adding humidity and time lag features

- All models have **higher AUC score,** better on distinguishing between classes.



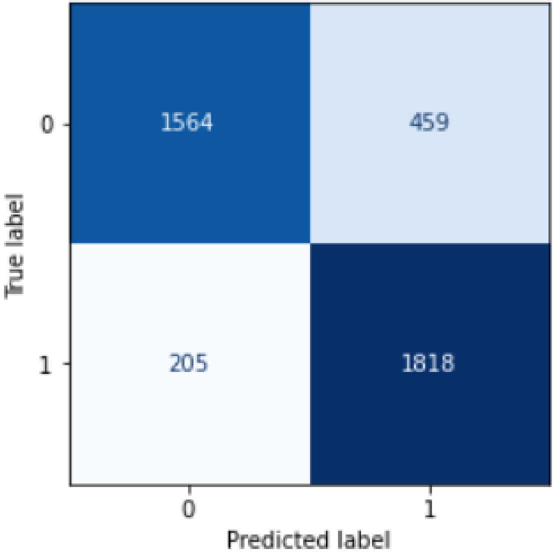


Model 2: TOP high corr feature from lag and humidity

Model & accuracy

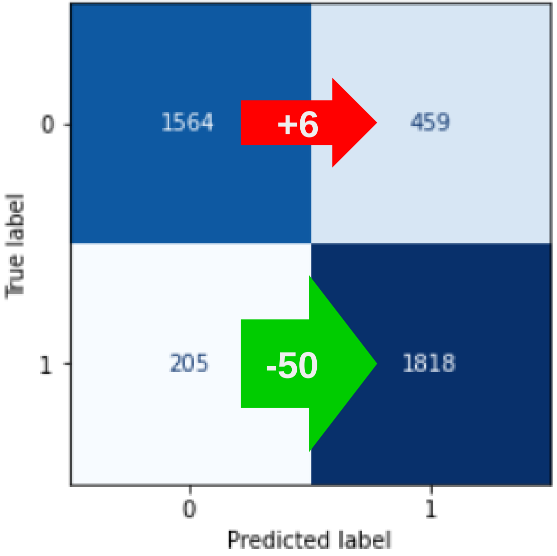
Model	Train Score	CrossVal Score	Test Score	AUC	Sensitivity	False Negative	Kaggle Score
Baseline	50%	50%	50%	-	-	-	
LR	81.3%	81.1%	84.7%	0.91	**92.3%**	**155**	**0.782**
Adaboost	89.5%	89.0%	89.7%	0.95	91.8%	165	0.629
Naive-Bayes	73.5%	73.7%	76.3%	0.86	81.0%	385	0.646
Random Forest	99.9%	94.4%	80.9%	0.95	66.9%	670	0.716
KNeighbors	94.7%	92.3%	78.2%	0.81	65.3%	702	0.537

Big improve in Sensitivity, small degrade in Specificity

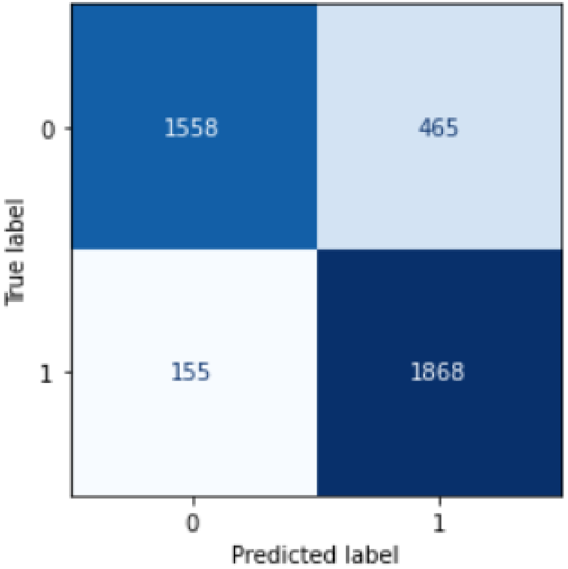


True label \ Predicted label	0	1
0	1564	459
1	205	1818

LR Model 1



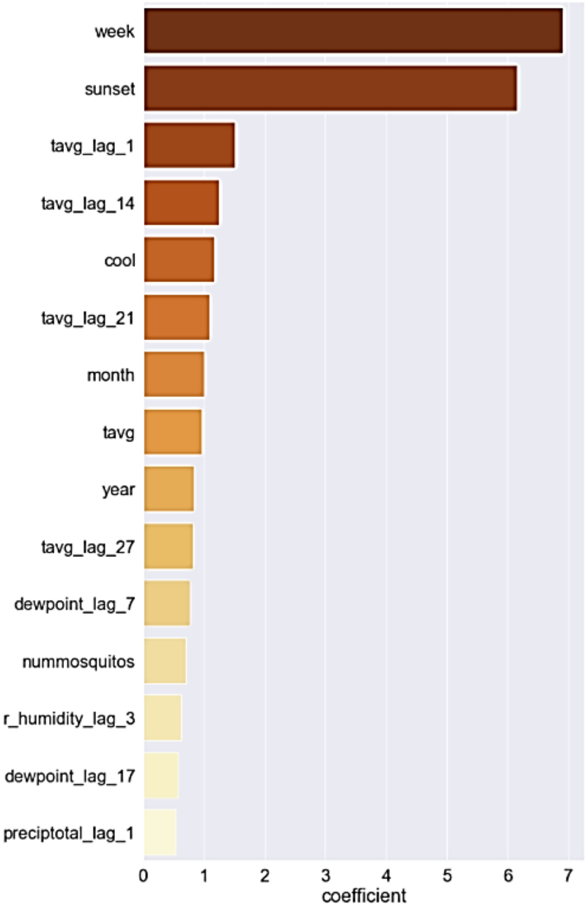
True label \ Predicted label	0	1
0	1564	459
1	205	1818

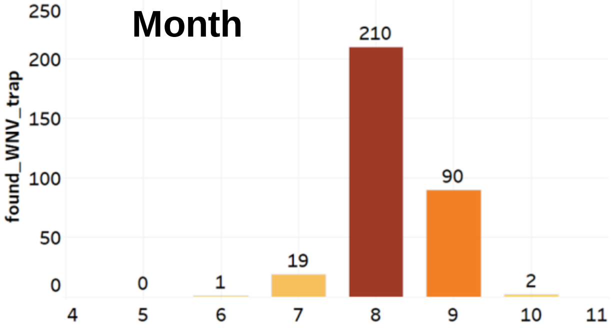


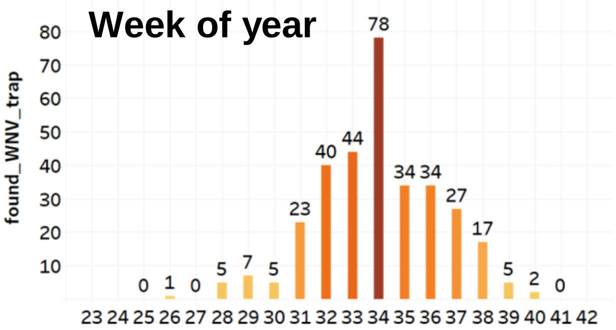
True label \ Predicted label	0	1
0	1558	465
1	155	1868

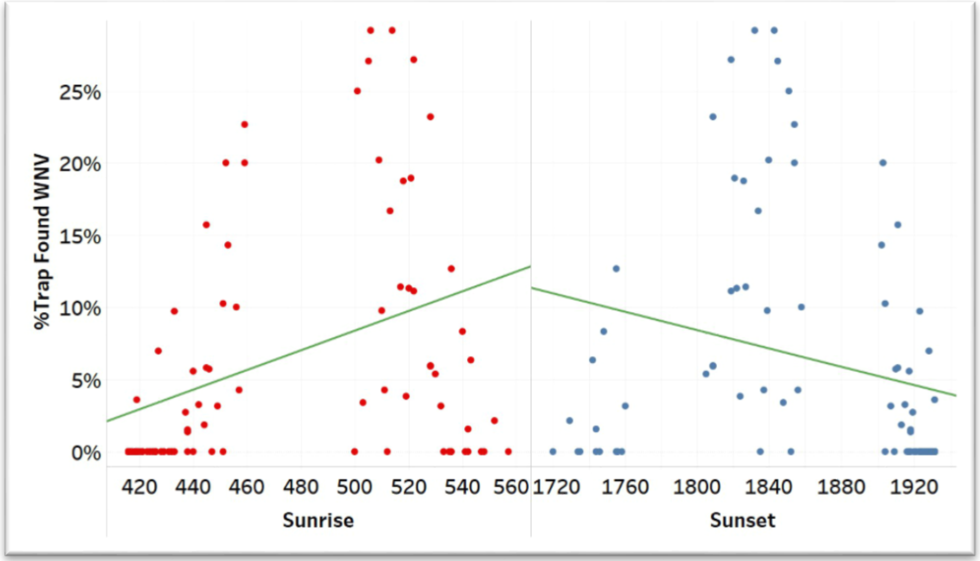
LR Model 2

TOP Coefficient Features – [Seasonal features]

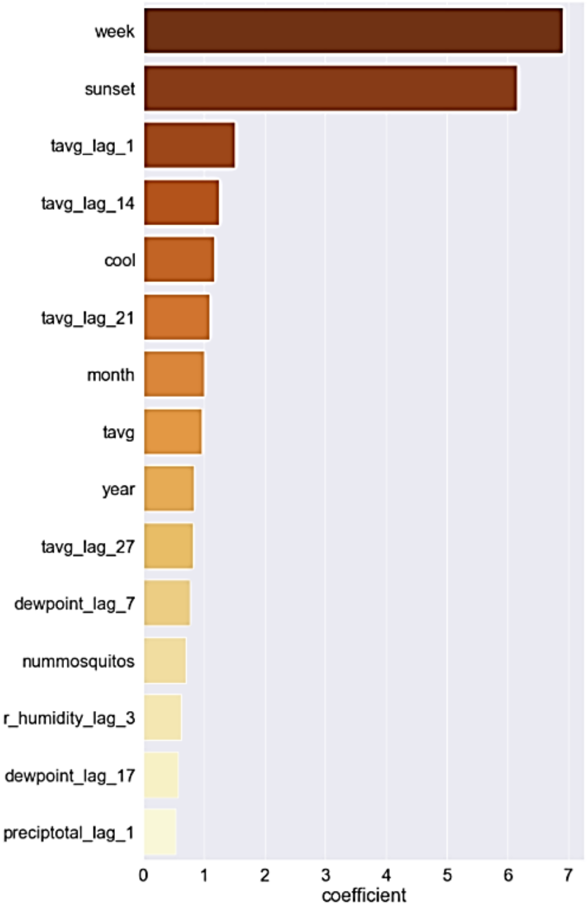


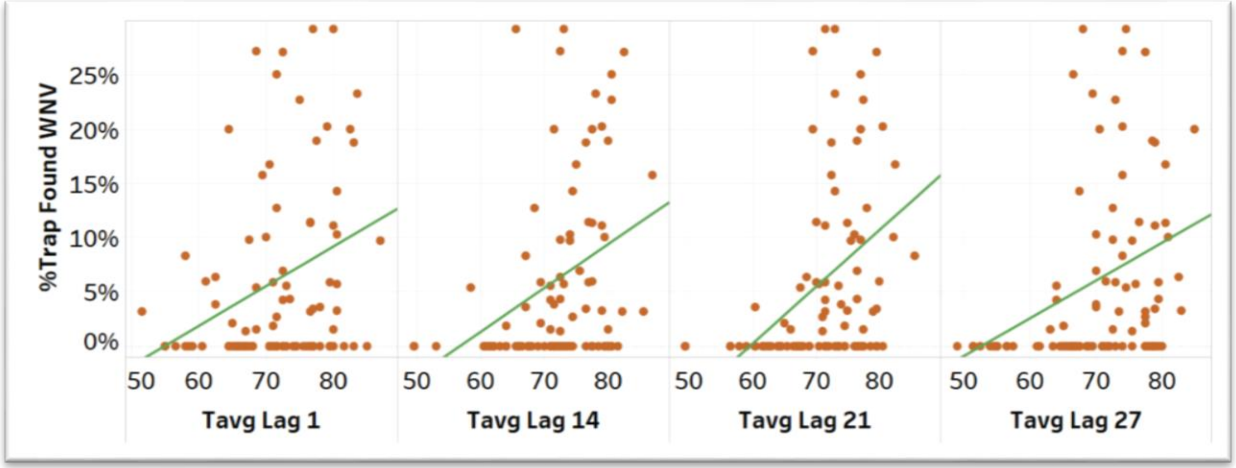

Month

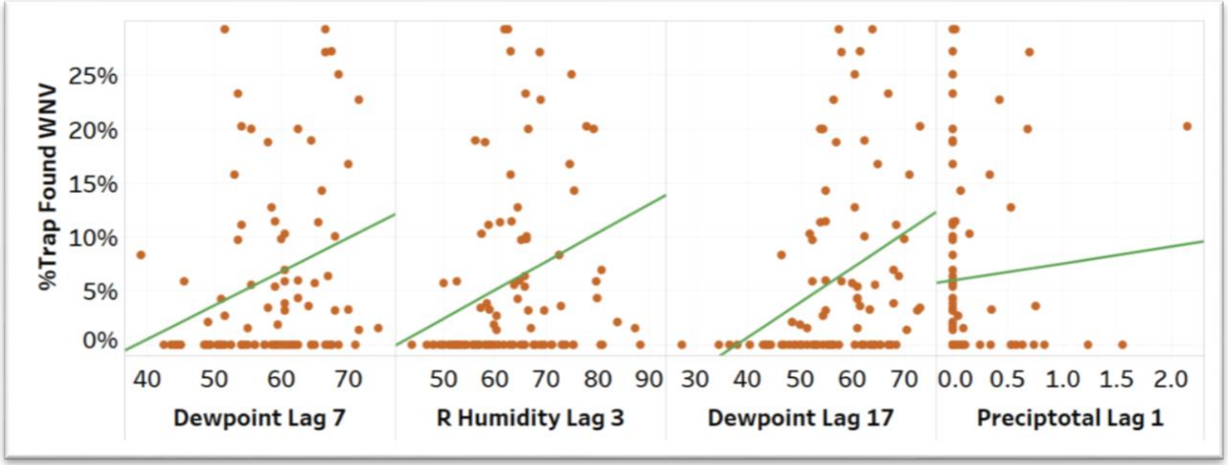

Week of year



TOP Coefficient Features – [Lagging Time features]







Conclusion

Conclusion

- WNV is highly seasonal, most occur in end of July and mid of September (week31-week38)
- WNV outbreak are more serious in summer (high temperature) and high perciption (humidity, dewpoint, etc.)
- Therefore, we need to spray before week 31 to maximize cost benefit.

Best scoring model

- LogisticRegression()
- Kaggle score: **0.782**

Furthermore improvement

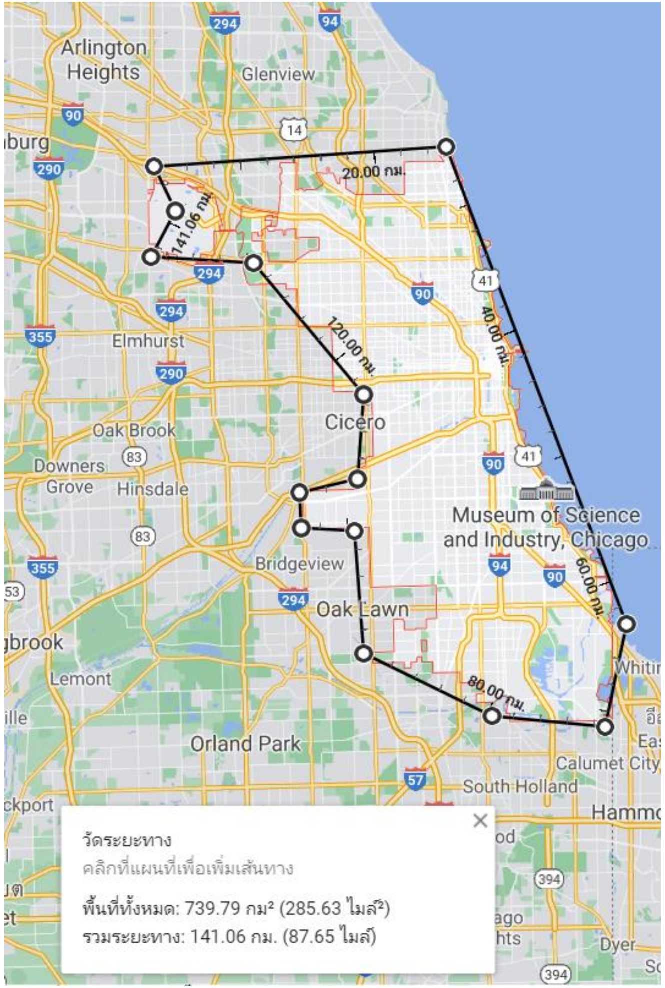
- Go deep down in weather: streaks of weather like rain 7 days in a row
- Spatial area correlation is neighbor area effect.
- Spatial time series correlation is neighbor temperature effect or not.

Thank You

Do you have any question?

C

i



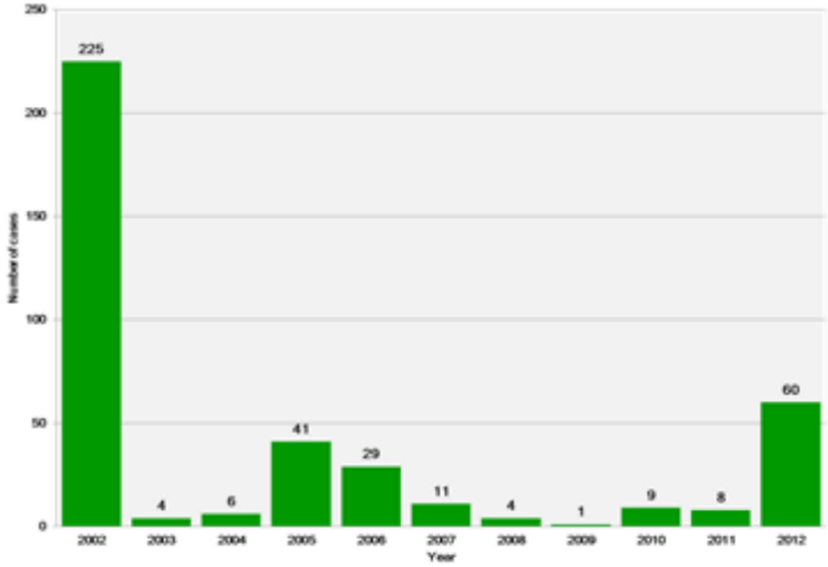
วัดระยะทาง

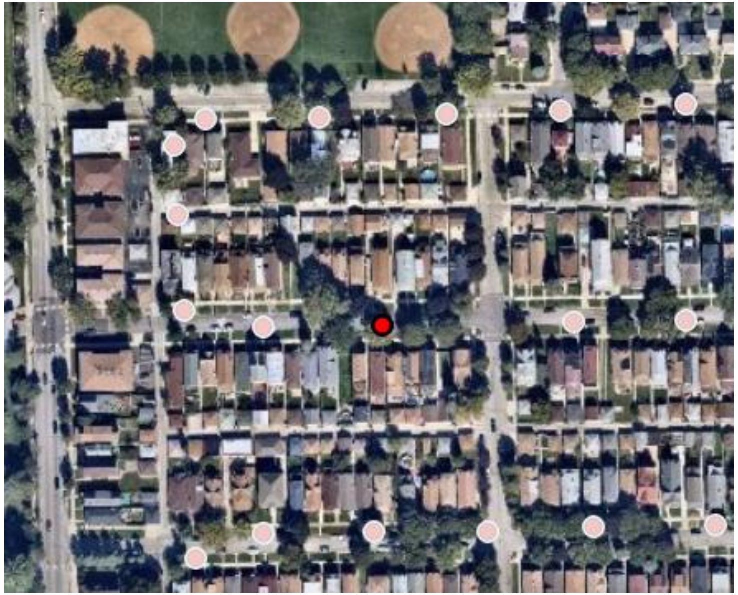
คลิกที่แผนที่เพื่อเพิ่มเส้นทาง

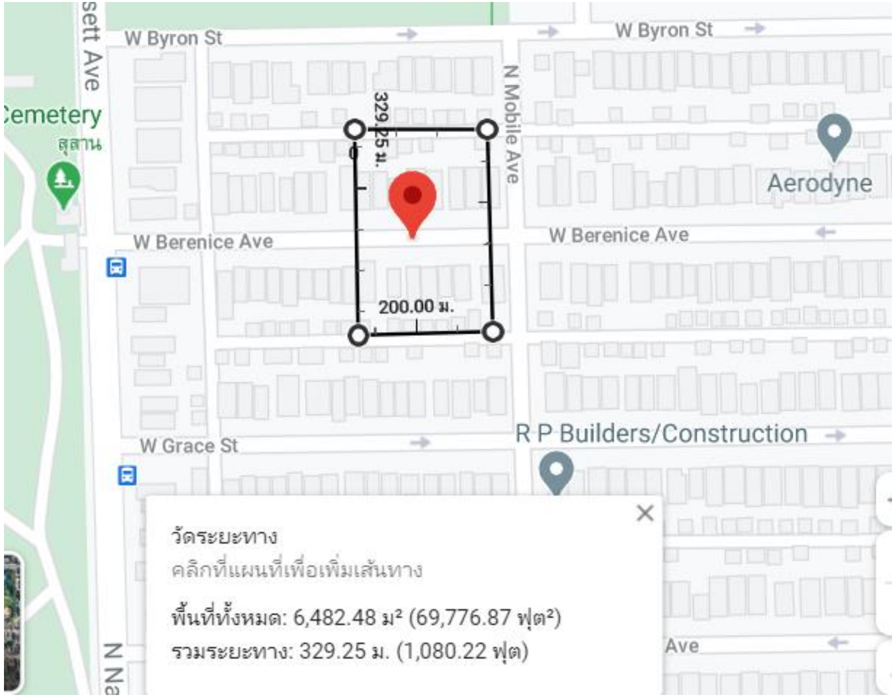
พื้นที่ทั้งหมด: 739.79 กม² (285.63 ไมล์²)

รวมระยะทาง: 141.06 กม. (87.65 ไมล์)

Figure 1. Number of reported confirmed and probable cases of West Nile virus among Chicago residents by year, 2002-2012.







วัดระยะทาง

คลิกที่แผนที่เพื่อเพิ่มเส้นทาง

พื้นที่ทั้งหมด: 6,482.48 ม² (69,776.87 ฟุต²)

รวมระยะทาง: 329.25 ม. (1,080.22 ฟุต)