Challenge:
West Nile Virus Detection

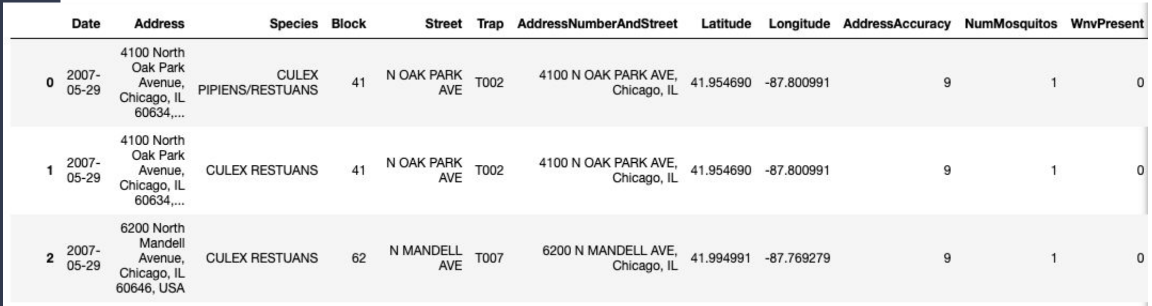
From Kaggle

Presented by:
Marco Santos

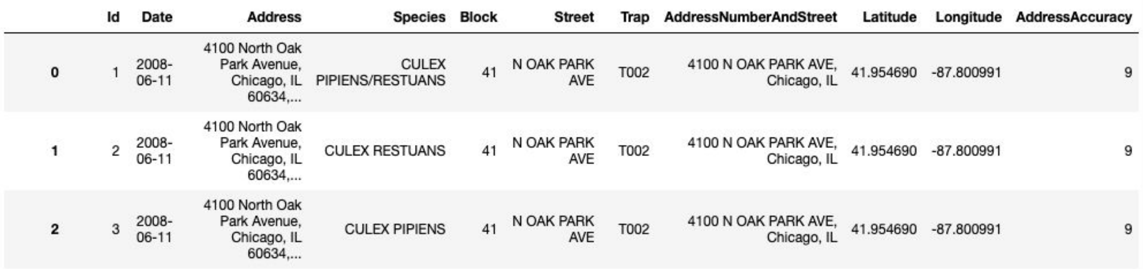
The Challenge

Analyzing weather data and GIS data and predicting whether or not the West Nile virus is present, for a given time, location, and species.

The Dataset



	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

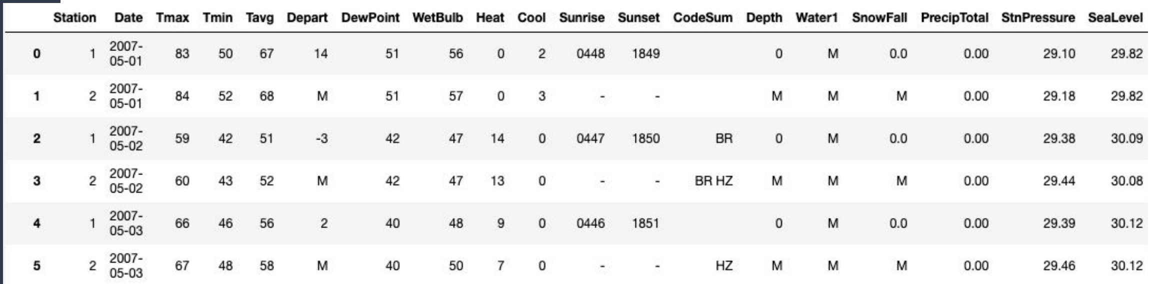


	Id	Date	Address	Species	Block	Street	Trap	AddressNumberAndStreet	Latitude	Longitude	AddressAccuracy
0	1	2008-06-11	4100 North Oak Park Avenue, Chicago, IL 60634,...	CULEX PIPIENS/RESTUANS	41	N OAK PARK AVE	T002	4100 N OAK PARK AVE, Chicago, IL	41.954690	-87.800991	9
1	2	2008-06-11	4100 North Oak Park Avenue, Chicago, IL 60634,...	CULEX RESTUANS	41	N OAK PARK AVE	T002	4100 N OAK PARK AVE, Chicago, IL	41.954690	-87.800991	9
2	3	2008-06-11	4100 North Oak Park Avenue, Chicago, IL 60634,...	CULEX PIPIENS	41	N OAK PARK AVE	T002	4100 N OAK PARK AVE, Chicago, IL	41.954690	-87.800991	9

Files to work with:

- Train.csv
- Test.csv
- Weather.csv

Used Pandas to view and clean



	Station	Date	Tmax	Tmin	Tavg	Depart	DewPoint	WetBulb	Heat	Cool	Sunrise	Sunset	CodeSum	Depth	Water1	SnowFall	PrecipTotal	StnPressure	SeaLevel
0	1	2007-05-01	83	50	67	14	51	56	0	2	0448	1849		0	M	0.0	0.00	29.10	29.82
1	2	2007-05-01	84	52	68	M	51	57	0	3	-	-		M	M	M	0.00	29.18	29.82
2	1	2007-05-02	59	42	51	-3	42	47	14	0	0447	1850	BR	0	M	0.0	0.00	29.38	30.09
3	2	2007-05-02	60	43	52	M	42	47	13	0	-	-	BR HZ	M	M	M	0.00	29.44	30.08
4	1	2007-05-03	66	46	56	2	40	48	9	0	0446	1851		0	M	0.0	0.00	29.39	30.12
5	2	2007-05-03	67	48	58	M	40	50	7	0	-	-	HZ	M	M	M	0.00	29.46	30.12

Data Cleaning and Engineering

Dealing with Weather Problems

- DataFrame for Weather contained two stations.
- Some columns were missing data (usually from station #2).
- Dates were repeated for each station.
- Both stations had their own measurements for some columns.

Solving the Weather Problem

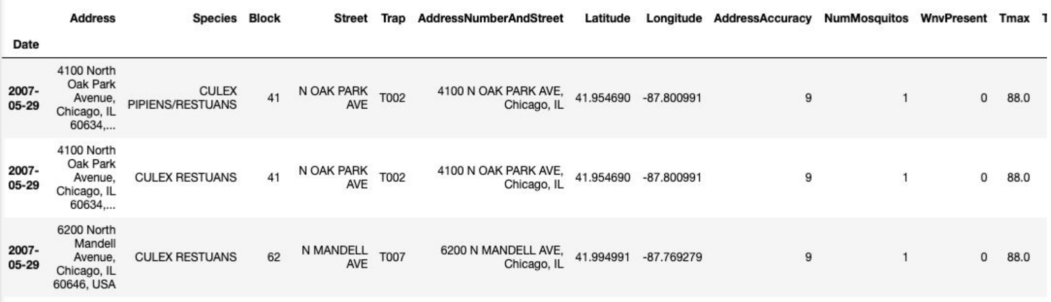
- Dropped columns: **CodeSum** and **Station**.
- CodeSum contained many missing values.
- Taking the average between the two stations made the Station column obsolete.
- Since the stations both represented the weather, averaging the values between them seemed appropriate.

New Weather Feature

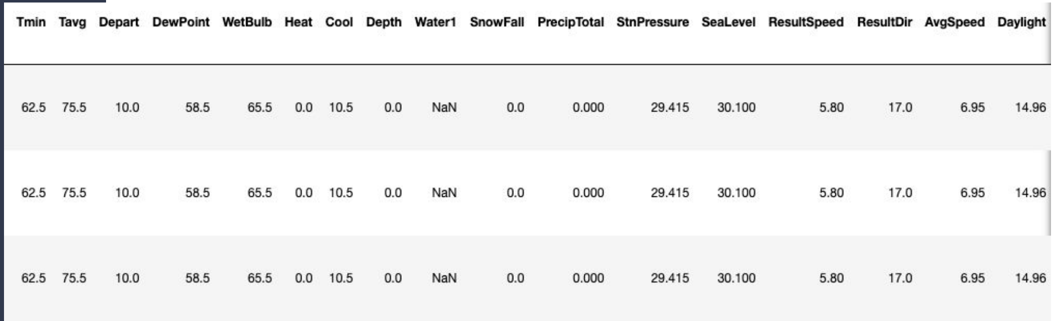
- Two columns: **Sunrise** and **Sunset**. Contained the sun’s time in 24hr format.
- New column was created by *subtracting* ***Sunrise*** *values from* ***Sunset*** *values, then dividing by 100*.
- New column/feature was created called **Daylight** which had the length of time, in hours, of the sun’s presence.

Joining the two DataFrames

- Joined/concatenated the **train** DF and the newly formatted **weather** DF on their shared **Dates**.



	Address	Species	Block	Street	Trap	AddressNumberAndStreet	Latitude	Longitude	AddressAccuracy	NumMosquitos	WnvPresent	Tmax	T
Date													
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	88.0	
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	88.0	
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	88.0	



Tmin	Tavg	Depart	DewPoint	WetBulb	Heat	Cool	Depth	Water1	SnowFall	PrecipTotal	StnPressure	SeaLevel	ResultSpeed	ResultDir	AvgSpeed	Daylight
62.5	75.5	10.0	58.5	65.5	0.0	10.5	0.0	NaN	0.0	0.000	29.415	30.100	5.80	17.0	6.95	14.96
62.5	75.5	10.0	58.5	65.5	0.0	10.5	0.0	NaN	0.0	0.000	29.415	30.100	5.80	17.0	6.95	14.96
62.5	75.5	10.0	58.5	65.5	0.0	10.5	0.0	NaN	0.0	0.000	29.415	30.100	5.80	17.0	6.95	14.96

Newly Created DF

- Contained the columns from the recently formatted data but also the original train.csv columns.
- Needed more formatting/cleaning.

New Issues with this DataFrame:

- Most of the features dealt with location.
- Dates themselves could be a feature.
- A lot of redundant features involving the streets and addresses.

Fixing problems with the new DF

- Tried to create a new feature called **Zipcode** from the **Address** column.
- Many Addresses did not contain the zip code in the dataset.
- Decided to rely on the **Latitude** and **Longitude** columns for location data.

Fixing problems with the new DF

Dropping columns related to the Address:

- Address
- Block
- Street
- AddressNumberAndStreet
- AddressAccuracy

What remained:

- Species
- Trap
- Latitude
- Longitude
- NumMosquitos
- WnvPresent

New Engineered Features: Month

- Created a new column called **Month** derived from slicing the **Date** column.
- Retrieved the Month from the *yyyy-mm-dd* format of the Date column.
- Renamed the months to their respective names.
- Only had recorded monthly data from the summer and fall months.

New Engineered Features: Lat&Long

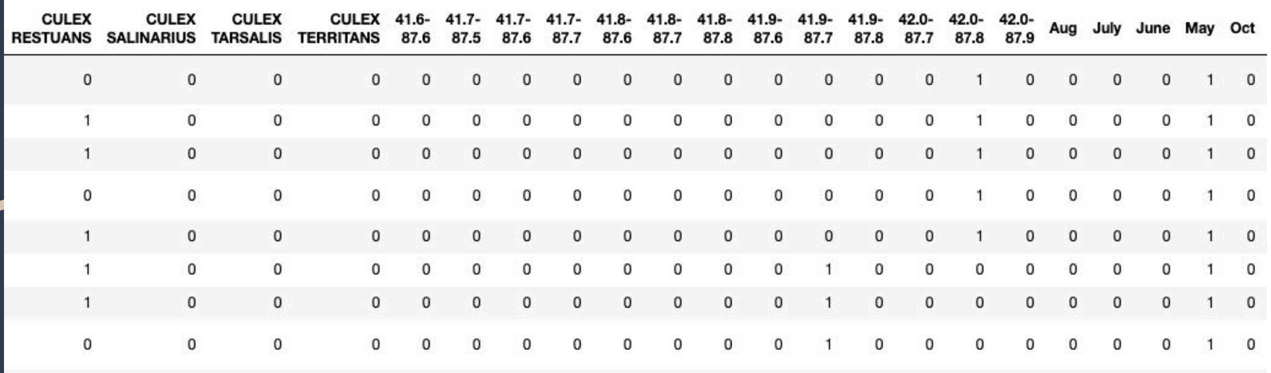
- Combined both the **Latitude** and **Longitude** columns as one column.
- Rounded both numbers to one decimal point.
- Which created thirteen unique locations.
- Combined together in string format to create the **Lat&Long** column.
- Dropped the Latitude and Longitude afterwards.

New Engineered Features: One-Hot Encoding

Three Feature columns contained categorical data:

- Month
- Lat&Long
- Species

Opted to One-Hot Encode each of the features.



CULEX RESTUANS	CULEX SALINARIUS	CULEX TARSALIS	CULEX TERRITANS	41.6-87.6	41.7-87.5	41.7-87.6	41.7-87.7	41.8-87.6	41.8-87.7	41.8-87.8	41.9-87.6	41.9-87.7	41.9-87.8	42.0-87.7	42.0-87.8	42.0-87.9	Aug	July	June	May	Oct
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0
1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0

Finishing the Cleaning and Formatting

Dropping any NaNs that remain

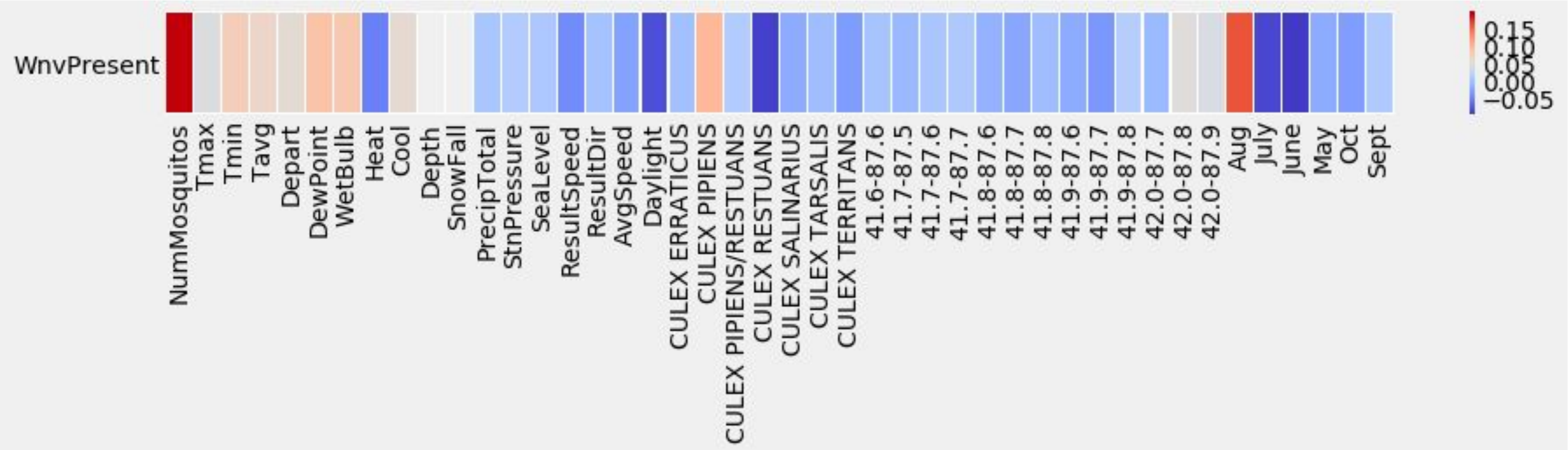
- Ended up being only one column that contained only NaN values: **Water1**

Exported the Final DF:

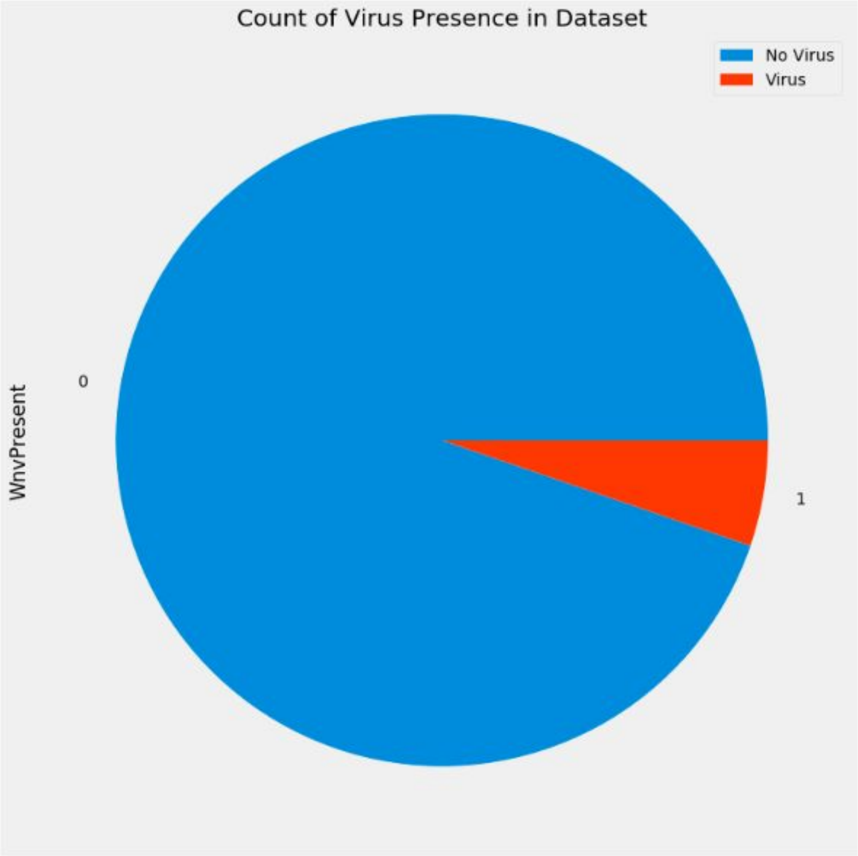
- Pickled the final DF for use in EDA and Feature Selection.

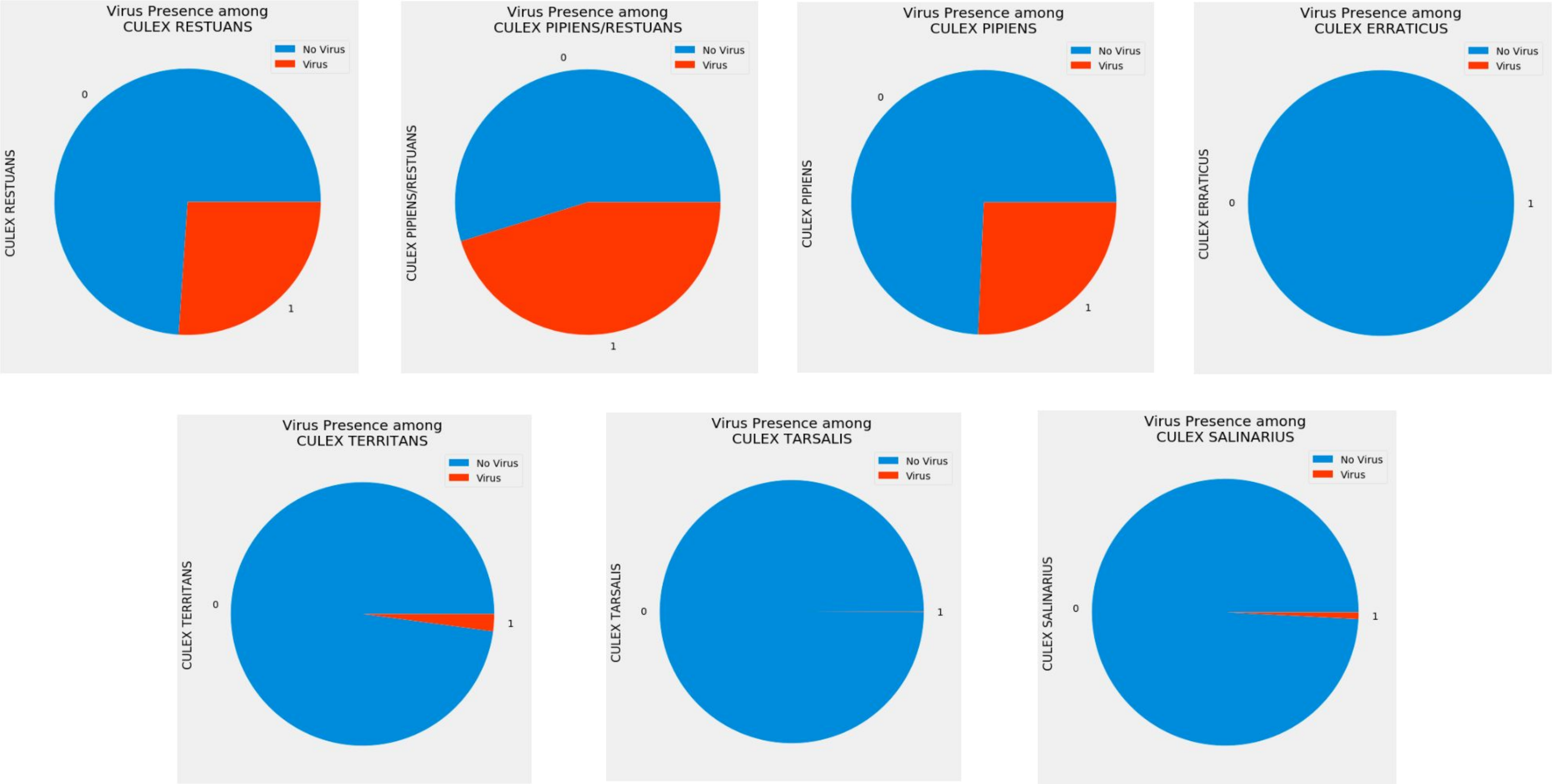
Data Exploration and Analysis

Checked for Correlation



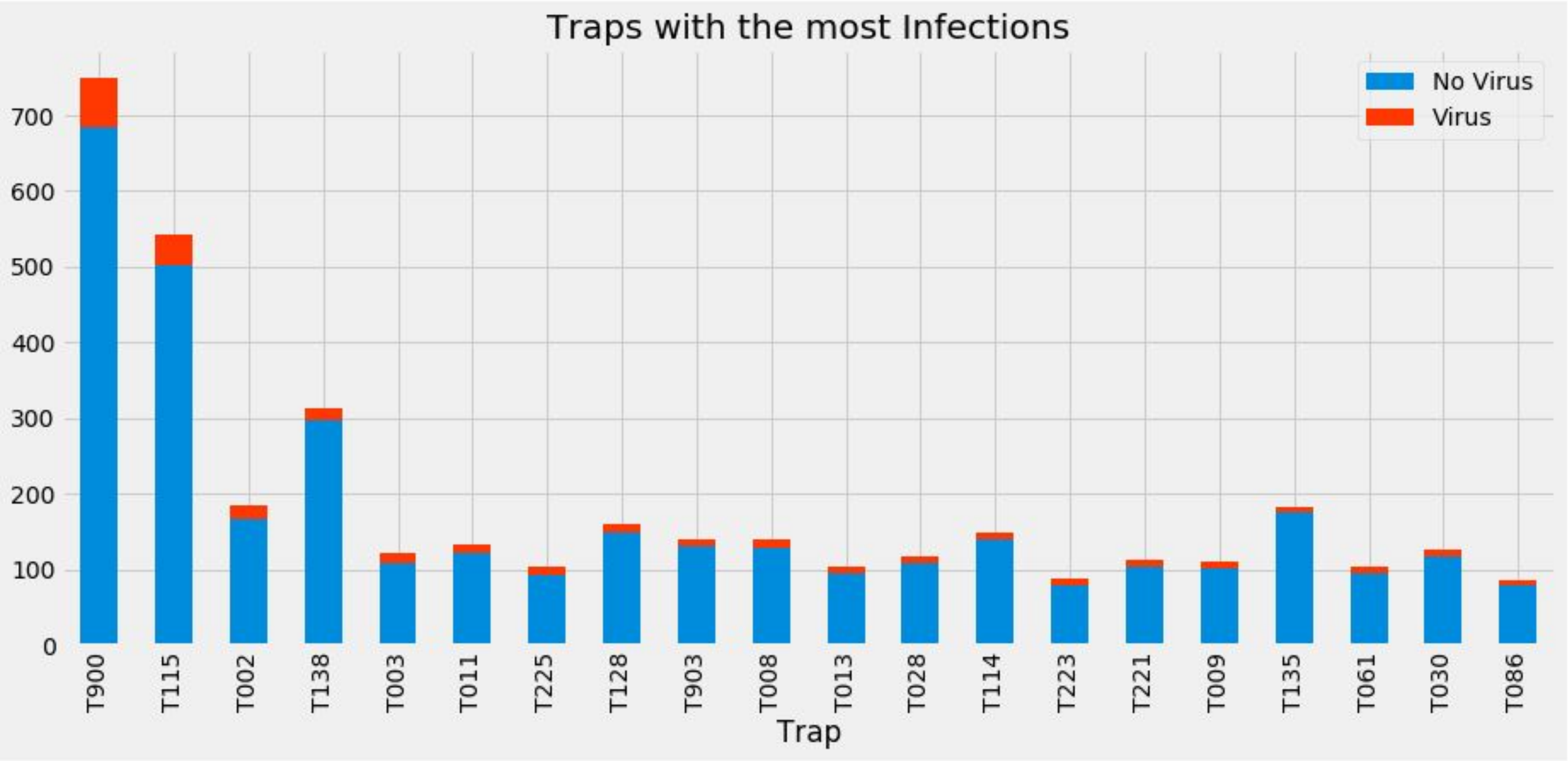
Class Balance



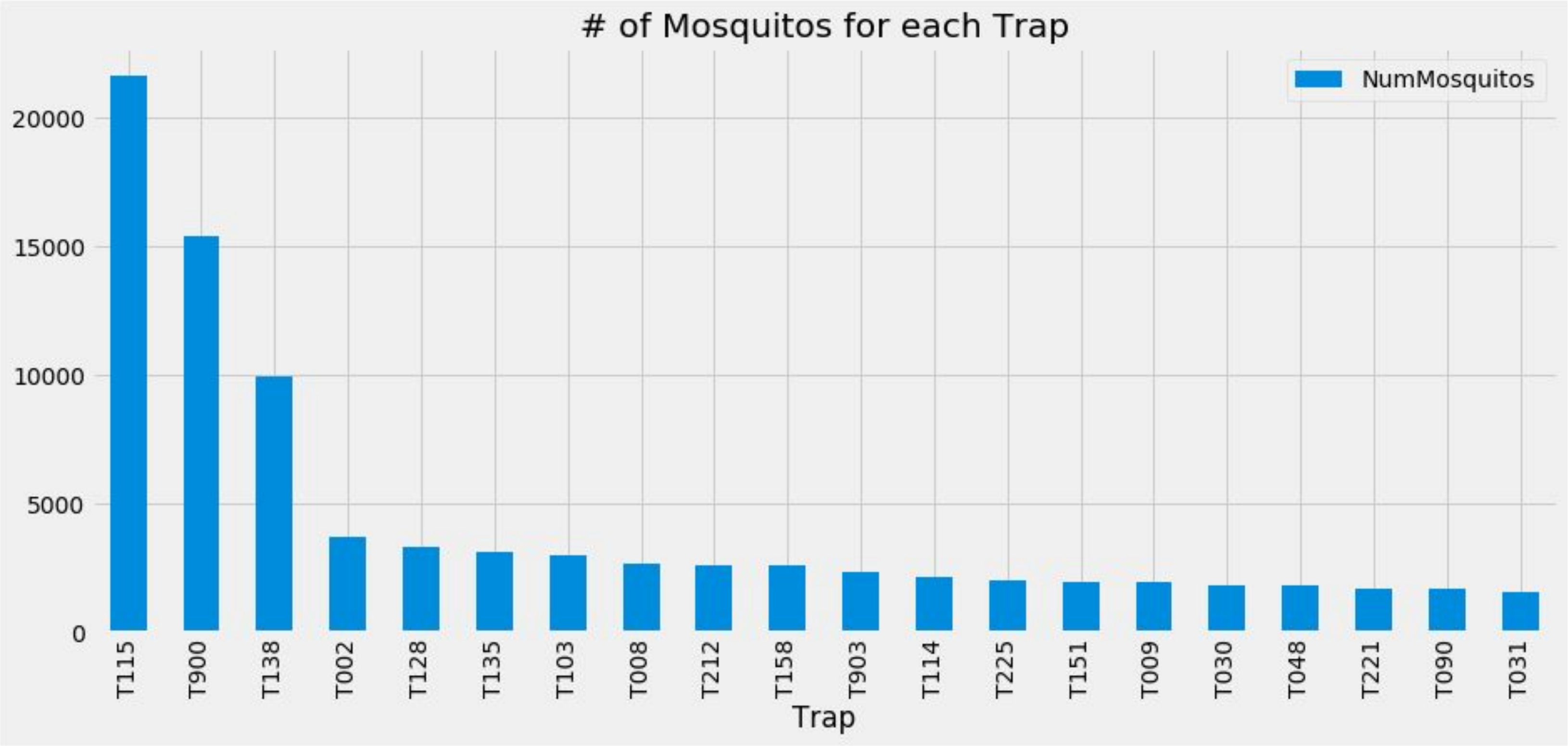


Count of Virus Presence Among Each Species

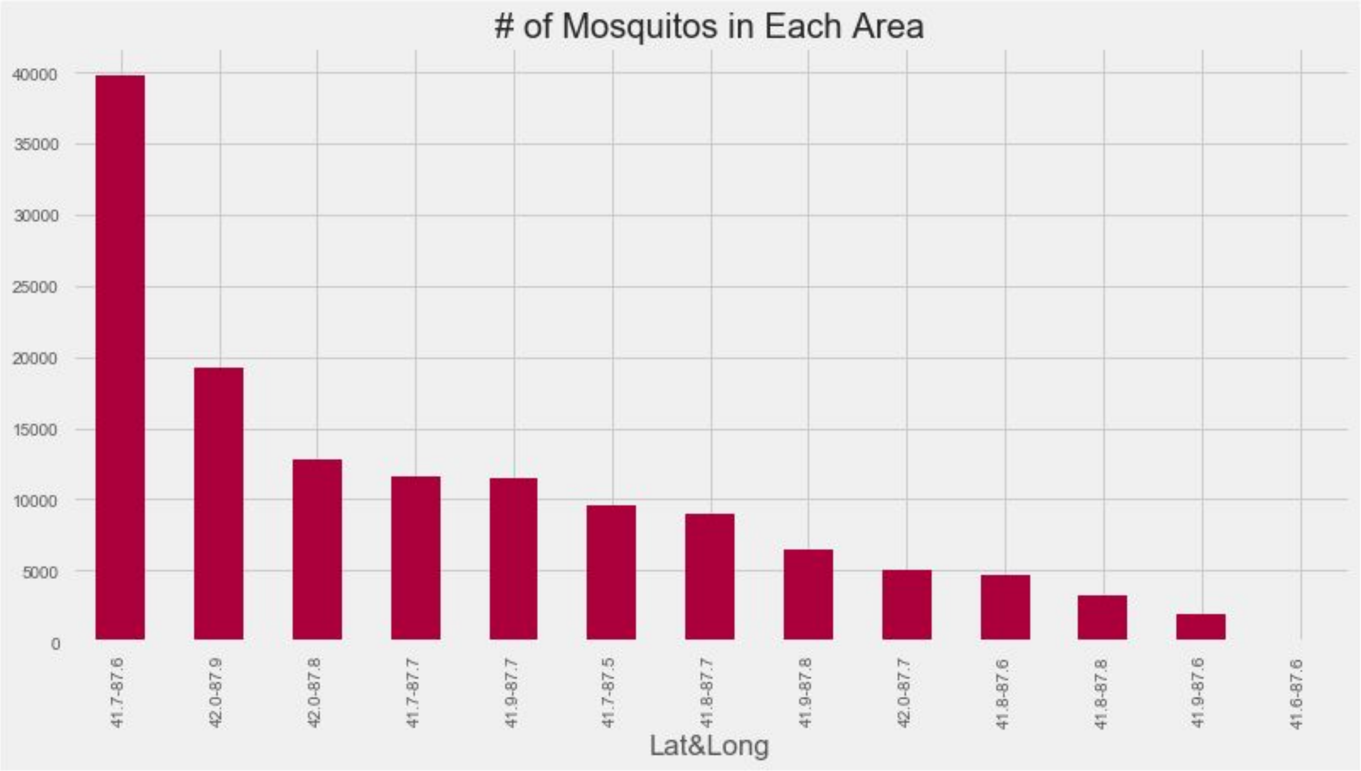
Finding Traps with Infections



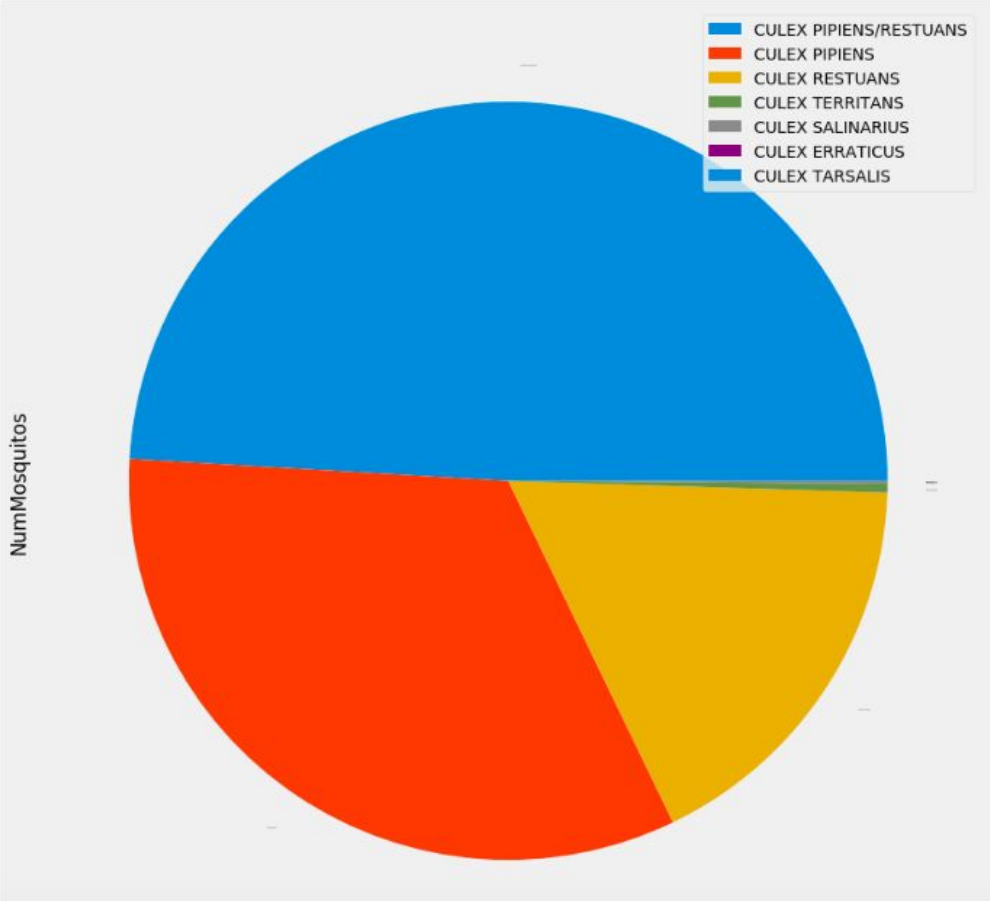
Quantity of Mosquitos Found in Each Trap



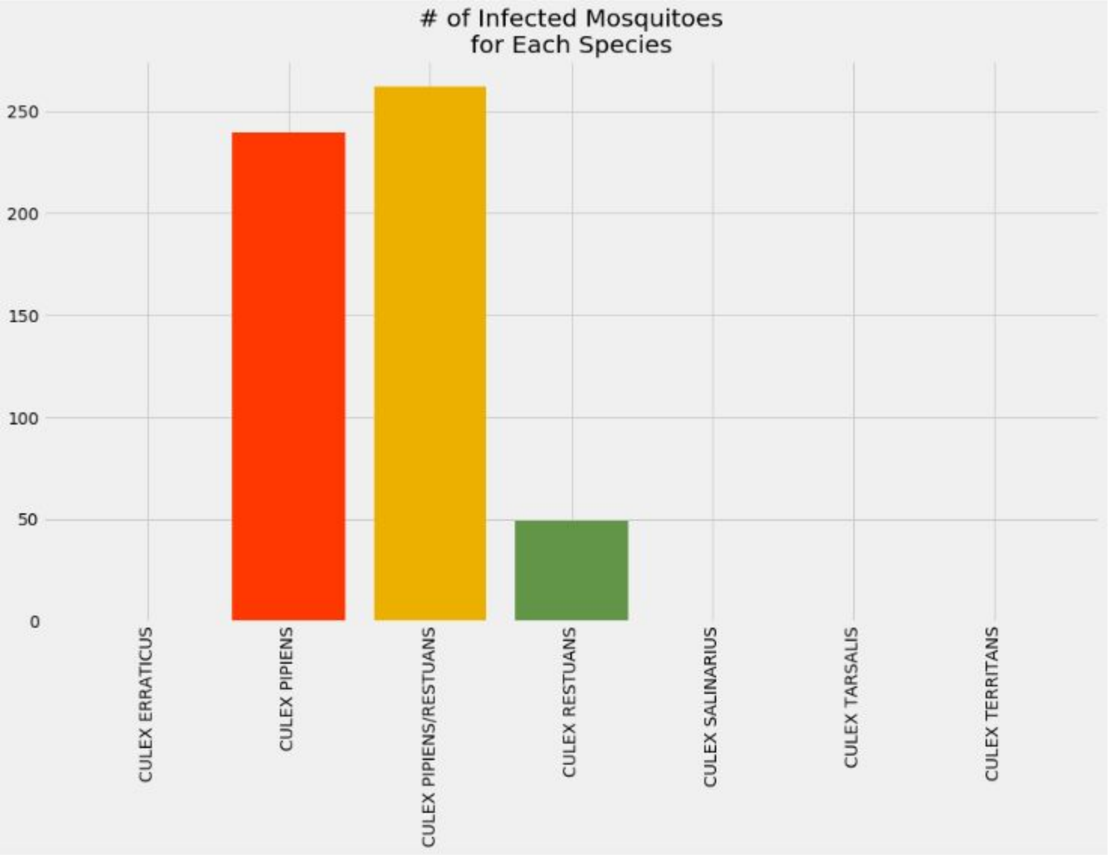
Number of Mosquitos in Each Area



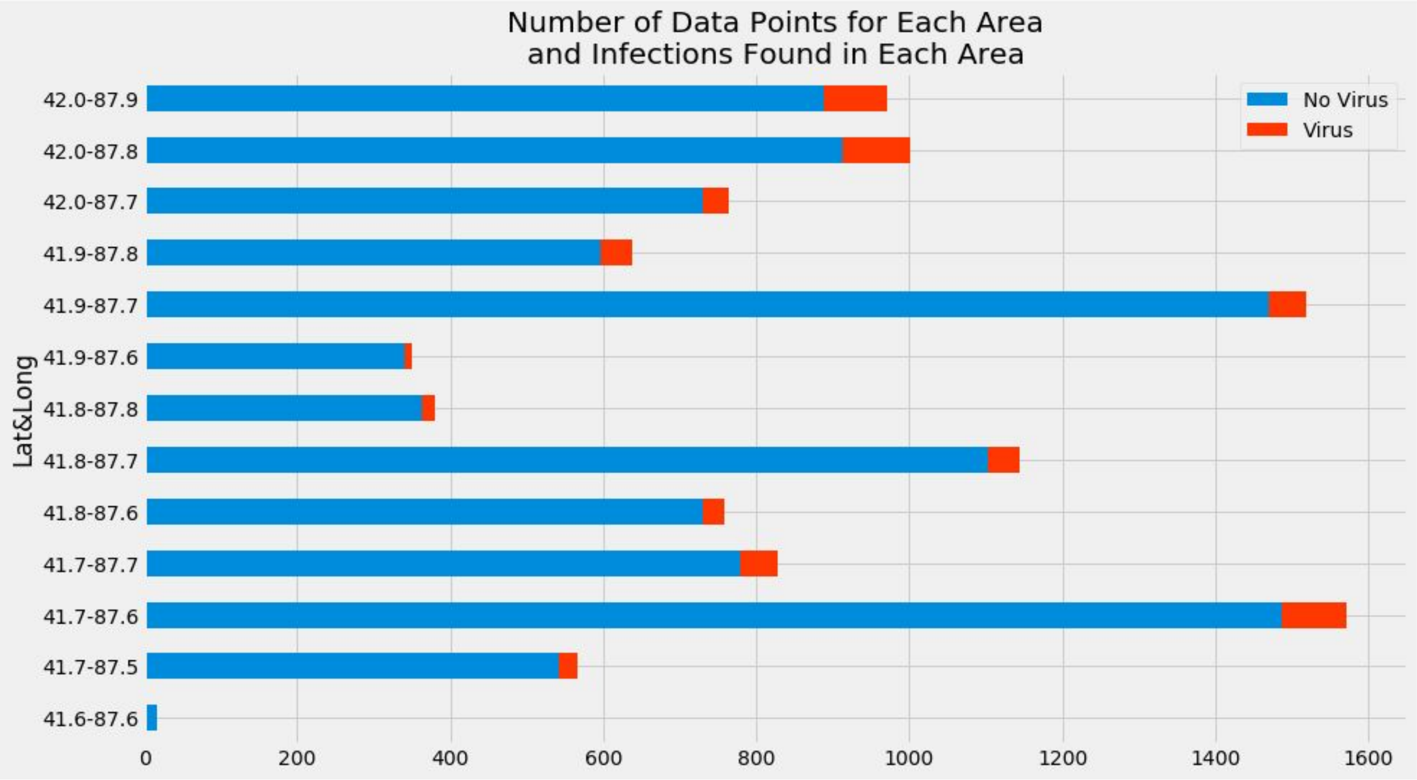
Presence of Each Species in the Dataset



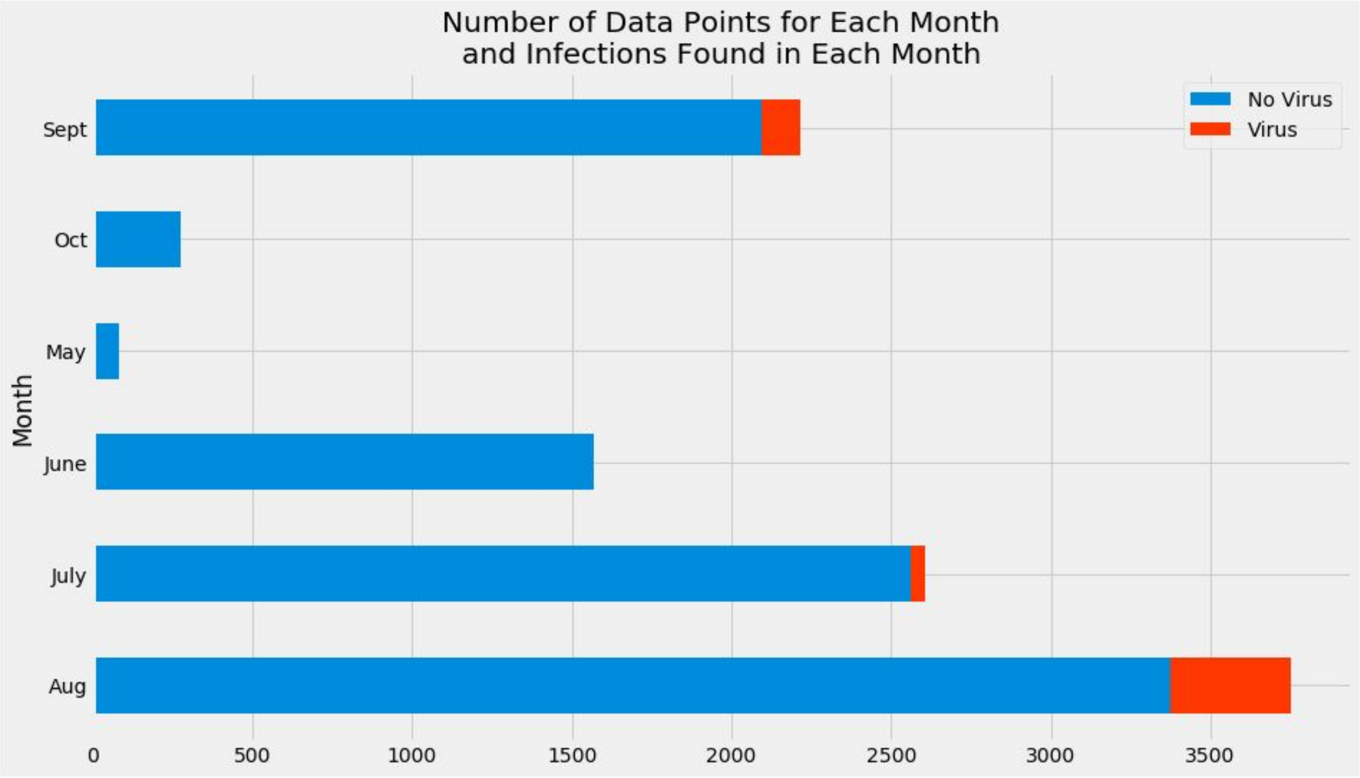
Infections Found among Each Species



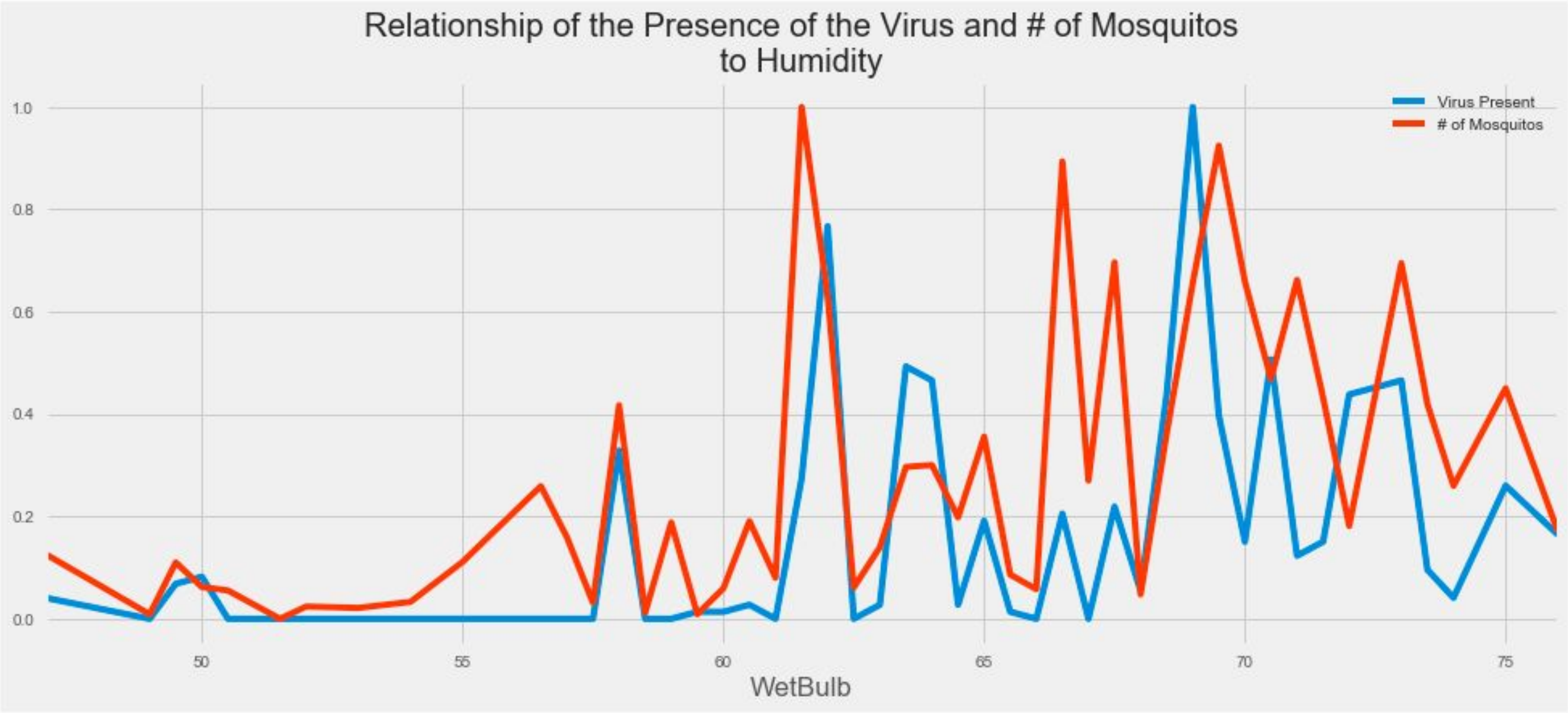
Data and Infections for Each Area



Data and Infections for Each Month



Virus Presence, Mosquitos, and Humidity



Feature Selection

Dropping Features

Dropping columns already one-hot encoded:

- Species
- Month
- Lat&Long
- Dropping **Trap** as well because it is a derivative of the location.
- Dropping **NumMosquitos** because it is not recorded in the test.csv.

Dropping Features

- Checked the Dataset for any null or NaN values that may remain.
- Found that **SnowFall** and **PrecipTotal** had null values.
- Dropped both.

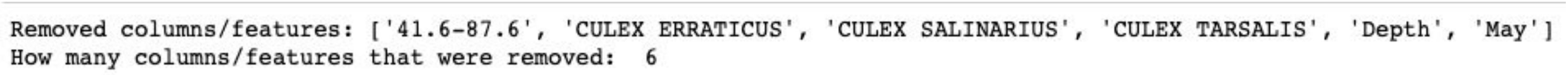
2 Different Approaches to Feature Selection

1. Using **Variance Threshold**

2. Using **Feature Importances** from ExtraTreesClassifier

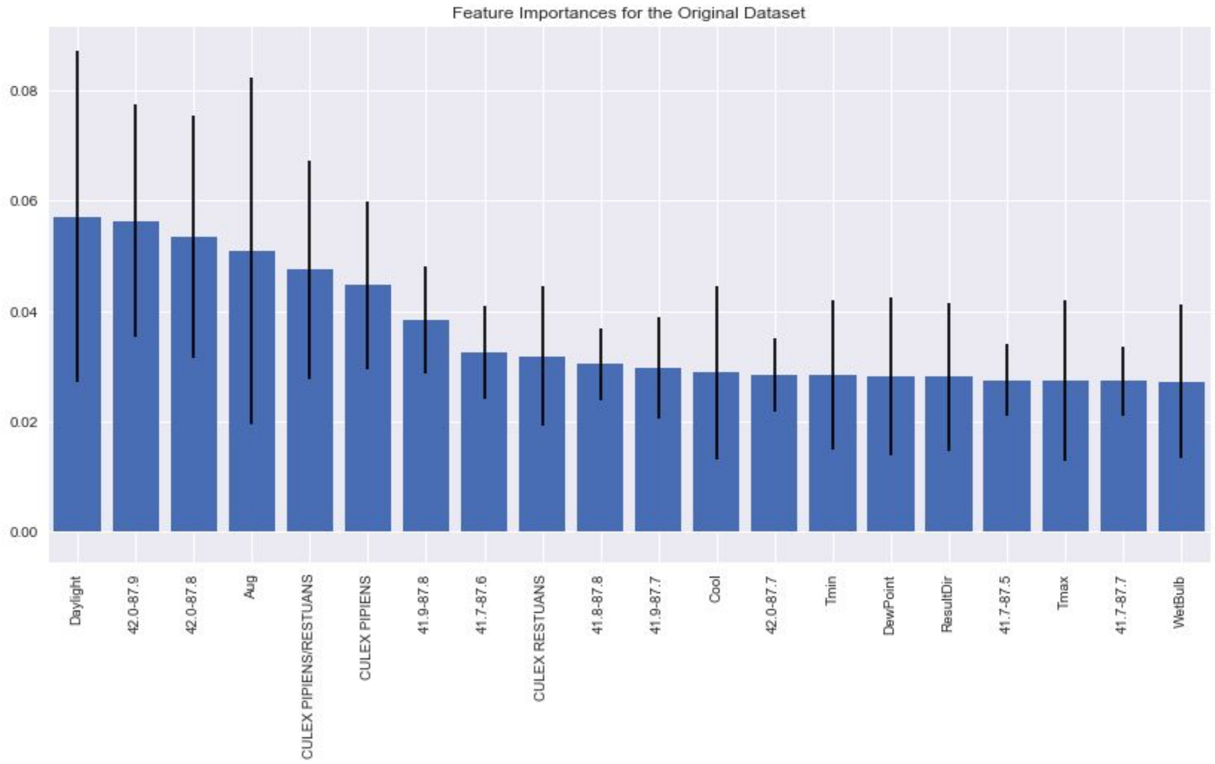
Variance Threshold

Using Variance Threshold of .06 to remove six features:



```
Removed columns/features: ['41.6-87.6', 'CULEX ERRATICUS', 'CULEX SALINARIUS', 'CULEX TARSALIS', 'Depth', 'May']
How many columns/features that were removed:  6
```

Feature Importances



Decreasing the DF size

- For Variance Threshold, kept only the remaining columns after features with low variance were removed.
- For Feature Importances, kept only the top 20 most important features from the feature columns.

Data Modeling

Data to Model

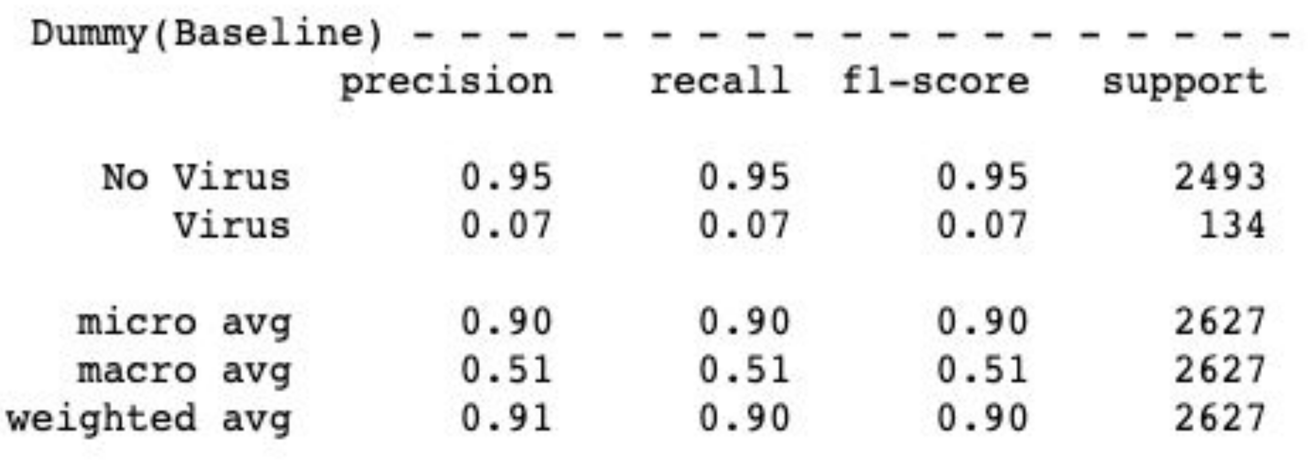
- Two different Datasets depending on method of Feature Selection
    - A Variance Dataset
    - An Important Features Dataset
- Tested out each model on both datasets to find the optimum dataset.

Pipeline Creation and Models

- Used various models to determine the best performing one.
- Fitted and predicted with each model but using default parameters.

- Each model was evaluated with the **Precision** and **Recall** metric instead of **Accuracy**.

Baseline Model (Classification Report)



```
Dummy(Baseline) - - - - - - - - - - - - - - - - - - - -
               precision    recall  f1-score   support

    No Virus       0.95      0.95      0.95      2493
       Virus       0.07      0.07      0.07       134

   micro avg       0.90      0.90      0.90      2627
   macro avg       0.51      0.51      0.51      2627
weighted avg       0.91      0.90      0.90      2627
```

Top Three Models

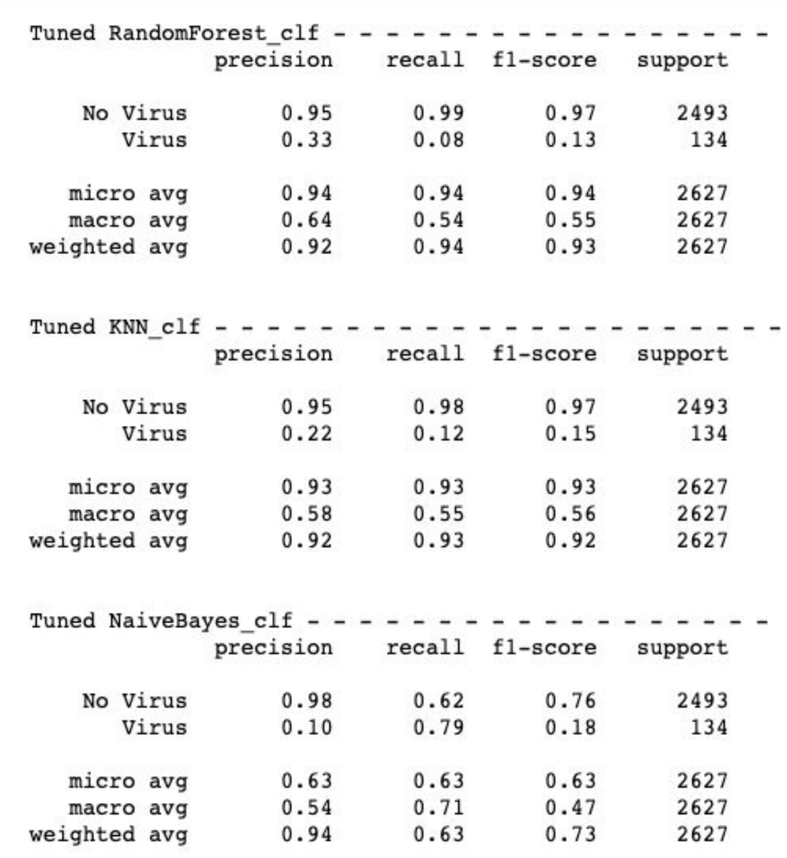
1. NaiveBayes (*ComplementNB*)
2. KNN
3. Random Forest

Grid Searching Parameters

- Grid Searched the Top Three Models with *f1_macro* as the scoring metric.
- **F1** because of necessary balance between Precision and Recall
- **Macro** because of the imbalance classes.

Grid Search the Top Three Models

Classification Reports for Each Tuned Model

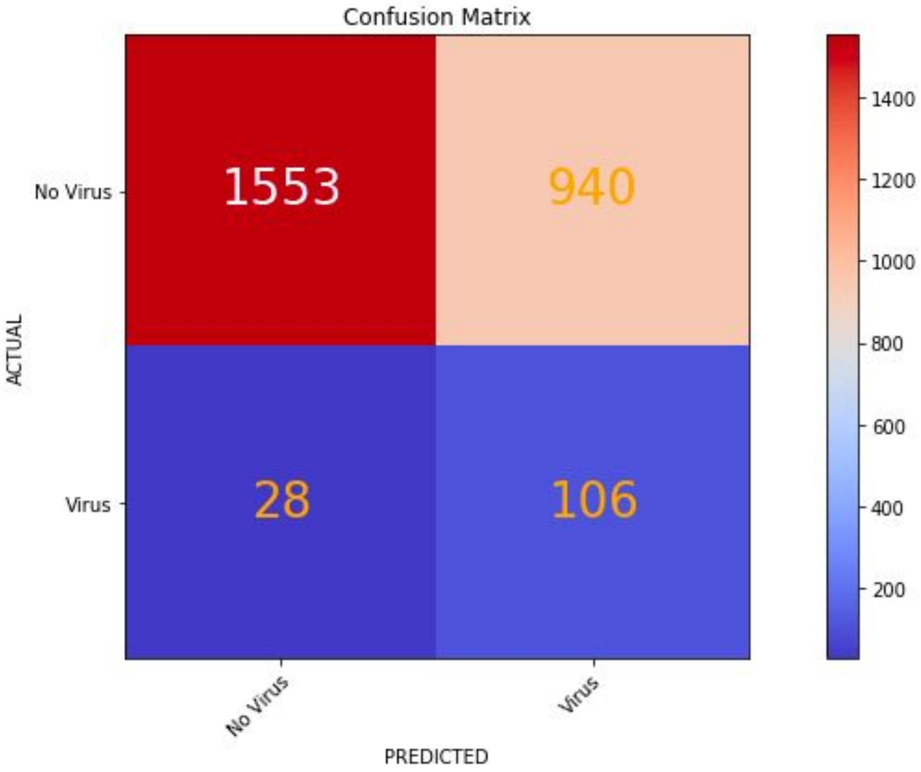


```
Tuned RandomForest_clf - - - - - - - - - - - - - - - - - -
              precision    recall  f1-score   support

    No Virus       0.95      0.99      0.97      2493
       Virus       0.33      0.08      0.13       134

   micro avg       0.94      0.94      0.94      2627
   macro avg       0.64      0.54      0.55      2627
weighted avg       0.92      0.94      0.93      2627


Tuned KNN_clf - - - - - - - - - - - - - - - - - - - - - - -
              precision    recall  f1-score   support

    No Virus       0.95      0.98      0.97      2493
       Virus       0.22      0.12      0.15       134

   micro avg       0.93      0.93      0.93      2627
   macro avg       0.58      0.55      0.56      2627
weighted avg       0.92      0.93      0.92      2627


Tuned NaiveBayes_clf - - - - - - - - - - - - - - - - - -
              precision    recall  f1-score   support

    No Virus       0.98      0.62      0.76      2493
       Virus       0.10      0.79      0.18       134

   micro avg       0.63      0.63      0.63      2627
   macro avg       0.54      0.71      0.47      2627
weighted avg       0.94      0.63      0.73      2627
```

Naive Bayes Model

- Best Performing Model considering the evaluation metrics used.
- ComplementNB was used because it is particularly suited for the imbalanced dataset.

Confusion Matrix for the Naive Bayes Model



Next Steps

Potential Improvements

- Other forms of feature selection using the other tree classifiers, L1 based selection, univariate selection.
- Use a Neural Network possibly.
- Could possible grid search every model tested to see if there were any surprises.
- Possibly more data to use.

Questions?