West Nile Virus Prediction

Predict west nile virus in mosquitos across the city of Chicago

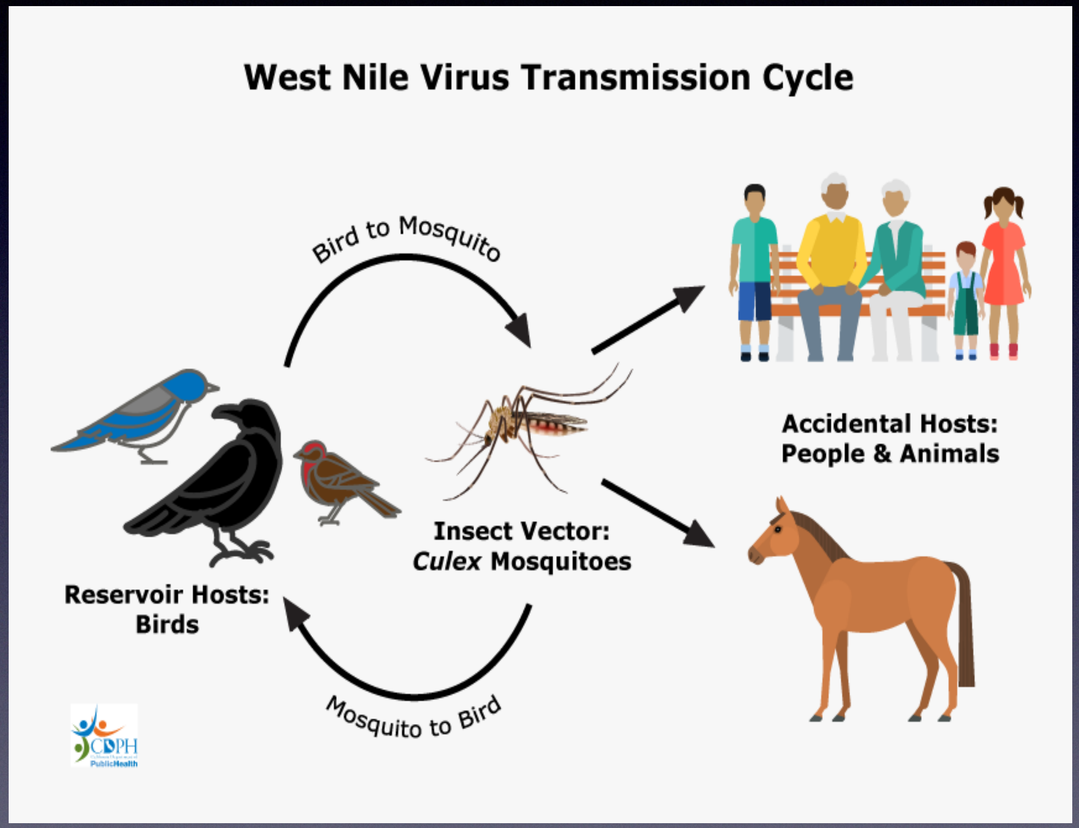
Bhuvanandini Ramesh

Table of Contents

Introduction - West Nile virus

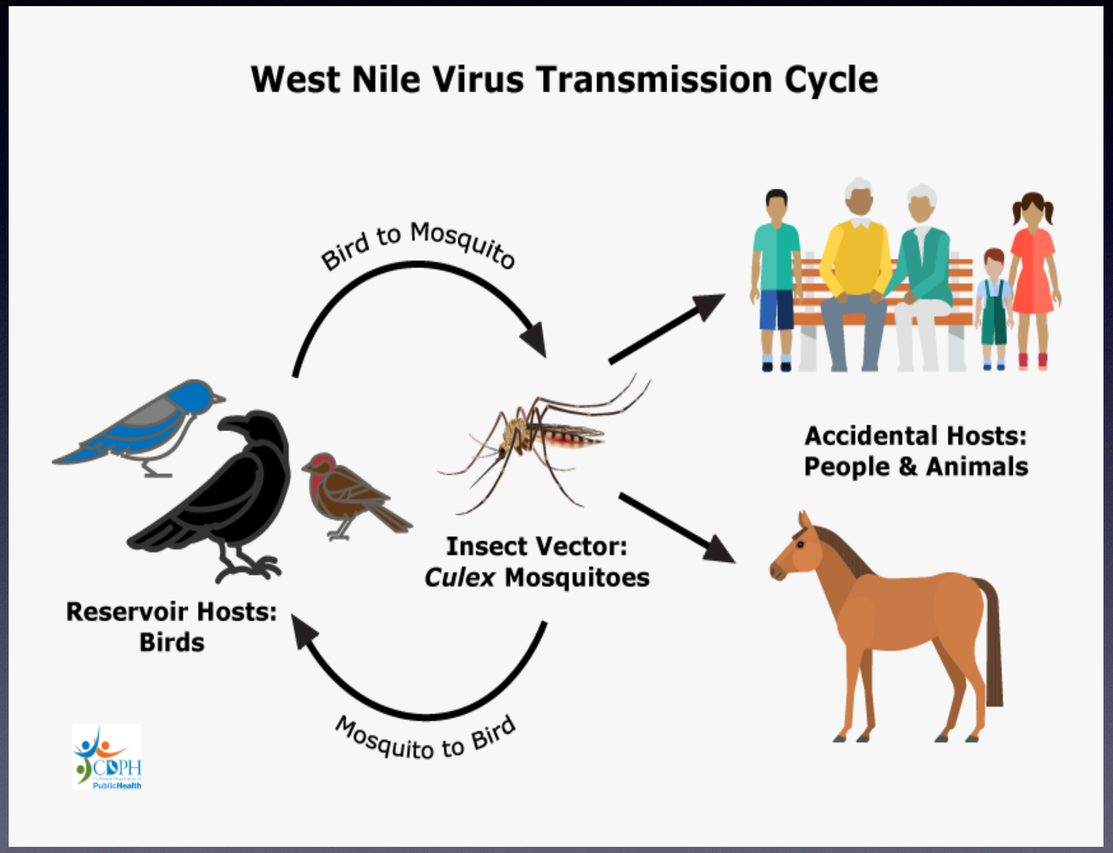
☆ The West Nile Virus (WNV) has been a serious problem for the United States since 1999.

☆ It is a mosquito-borne disease and about 1 in 150 affected people develop severe illness damaging the central nervous system.



Introduction - West Nile virus

- ☆Since 2002, Illinois and more specifically Chicago, has continued to suffer from multiple outbreaks of the West Nile Virus.
- ☆From 2005 to 2016, a total of 1,371 human WNV cases were reported within Illinois. Out of these total reported cases, 906 cases (66%) were from the Chicago region (Cook and DuPage Counties).



Introduction - Objective

This project is aimed at predicting outbreaks of the West Nile Virus. This will help the City of Chicago and Chicago Department of Public Health (CDPH) more efficiently and effectively allocate resources towards preventing transmission of this potentially deadly virus.

Data Sets

Spray

- 14,294 spray observations
- Data from 2011 & 2013
- 3 features - Location (Lat & Long) and Date

Weather

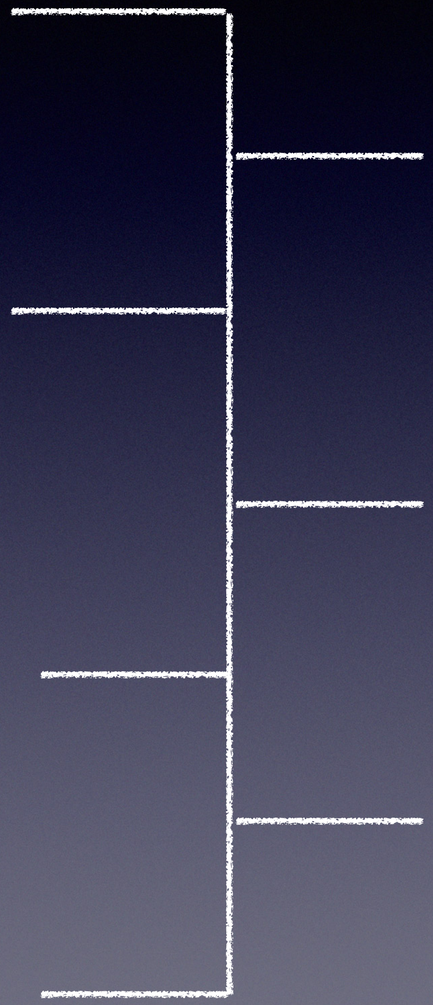
- 2944 observations from 2 stations
- Data from 2007 - 2014 everyday
- 21 features like Station, Date, Temperature, Dewpoint, etc.

Train

- 10,505 observations
- Data from 2007, 2009, 2011 & 2013
- 10 features like Location, Date, NumMosquitos, etc.

Test

- 116,293 observations
- Data from 2008, 2010, 2012 & 2014
- 9 features like Location, Date, species, etc.
- Id variable



Used these to perform EDA and Data Visualisations

Merged dataset for model training

Merged dataset for prediction

Workflow of the Solution

Data Cleaning → EDA → Feature Engineering → Model Building → Model Evaluation

★ Removal of outliers and imputing the missing values.

★ Perform data analysis to have a better understanding of the features and the prediction.

★ Create new features like

Relative Humidity,

Lag variables,

etc.

★ Model selection and fine tuning using SMOTE.

★ The built model is evaluated based on ROC curve and accuracy.

Data Cleaning

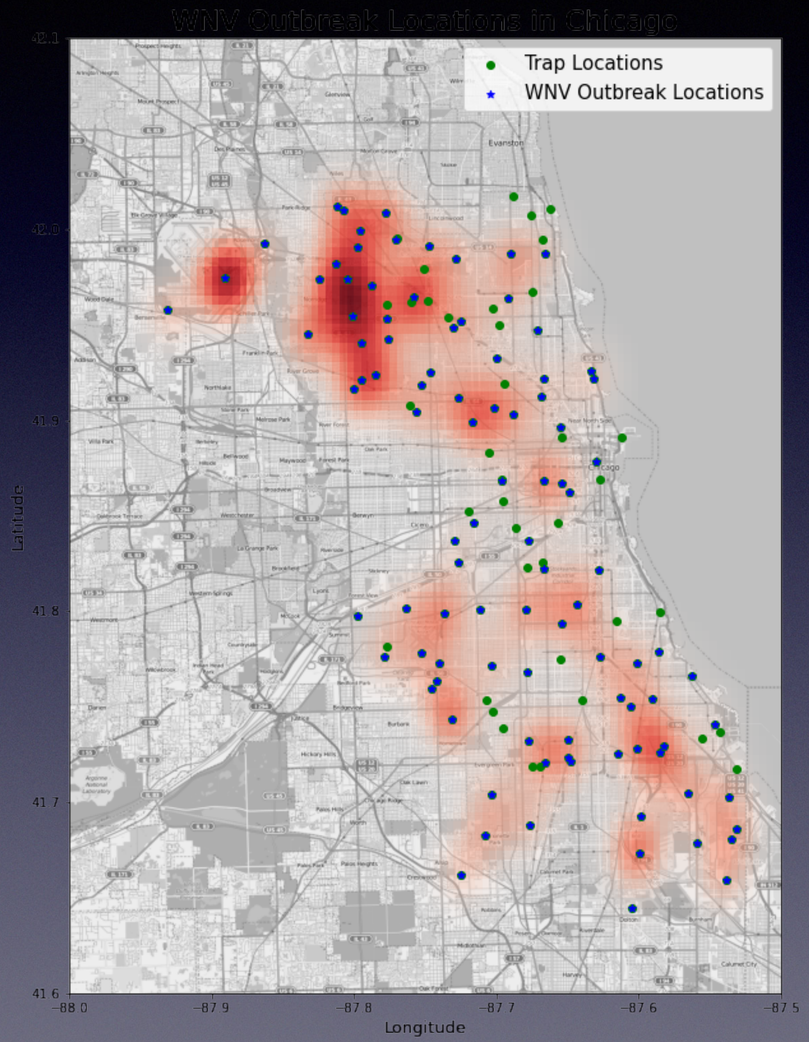
Weather data :
Station 1: CHICAGO O'HARE INTERNATIONAL AIRPORT Lat: 41.995 Lon: -87.933 Elev: 662 ft. above sea level
Station 2: CHICAGO MIDWAY INTL ARPT Lat: 41.786 Lon: -87.752 Elev: 612 ft. above sea level

- Imputed missing values for daily average temperature, Cool and Warm parameters with average of Tmax and Tmin.
- Calculated missing Depart from Station 2 with 30 year normal temperature based on Station 1 readings and Station 2 Tavg.
- Imputed missing values for WetBulb, PrecipTotal, StnPressure, SeaLevel, AvgSpeed using readings of station with non-missing value.
- Imputed 'T' or trace values as 0.01.
- Imputed Sunrise and Sunset for Station 2 with Station 1 values.

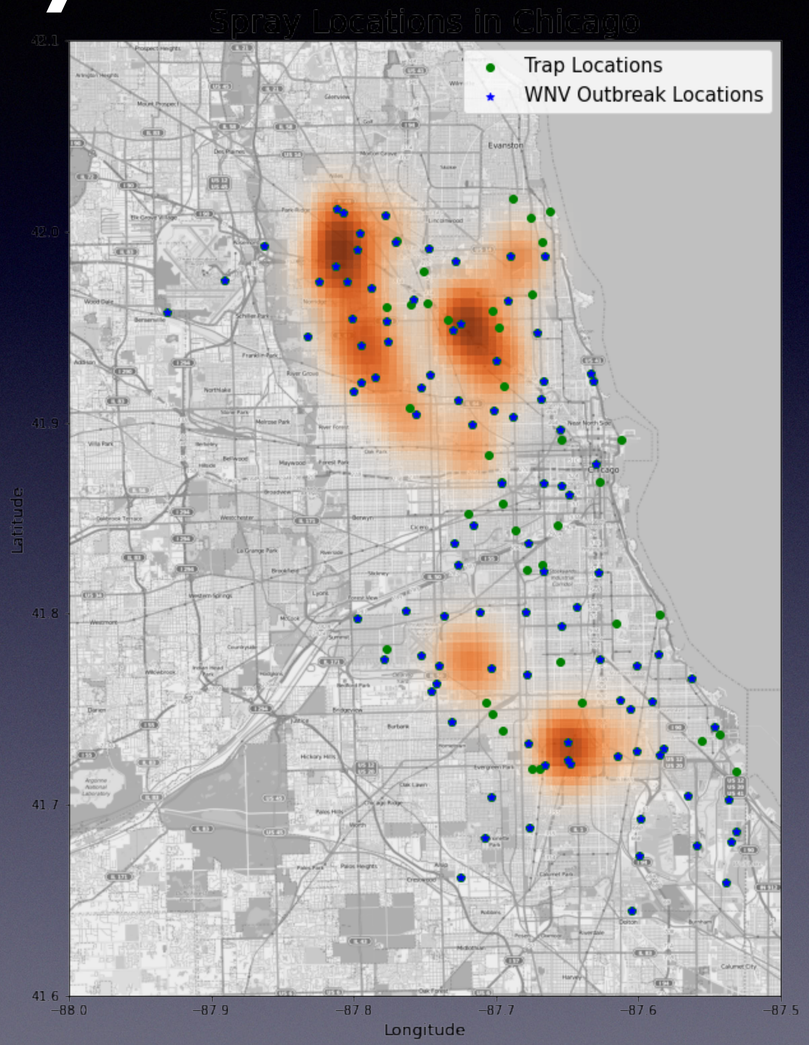
Data Cleaning

- Created new feature counting number of exceptional weather phenomena based on CodeSum.
- Created more interpretable features like Rain and Mist based on CodeSum.
- Changed Date from string object to datetime64.
- Dropped Water1, Depth, SnowFall due to high missing values.
- Transformed all features into float values Merged Station 1 and Station 2 by averaging values of each station.
- Extracted Year, Month, Week and Day of Week features.

Explanatory Data Analysis - Location

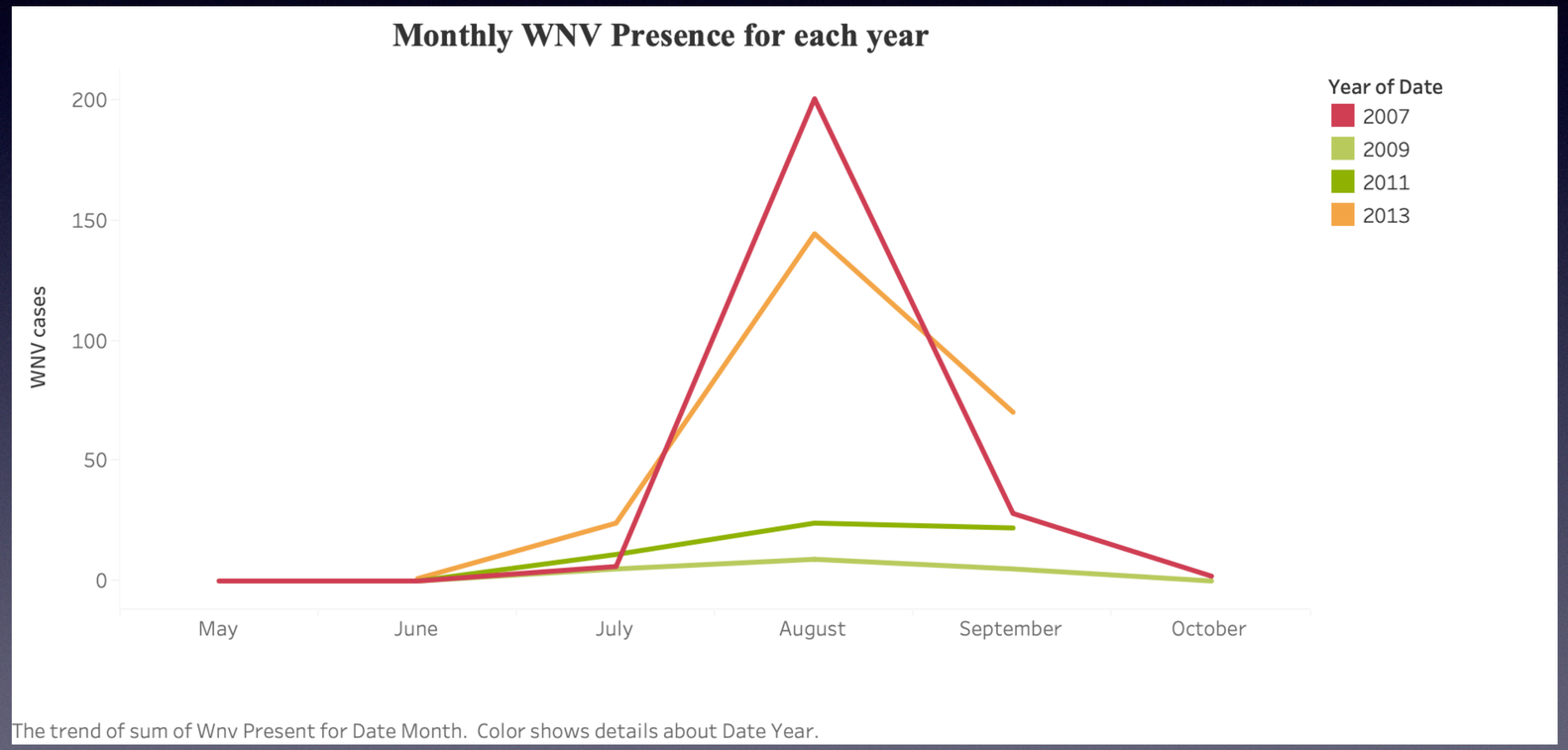


WNv Occurrence in Chicago



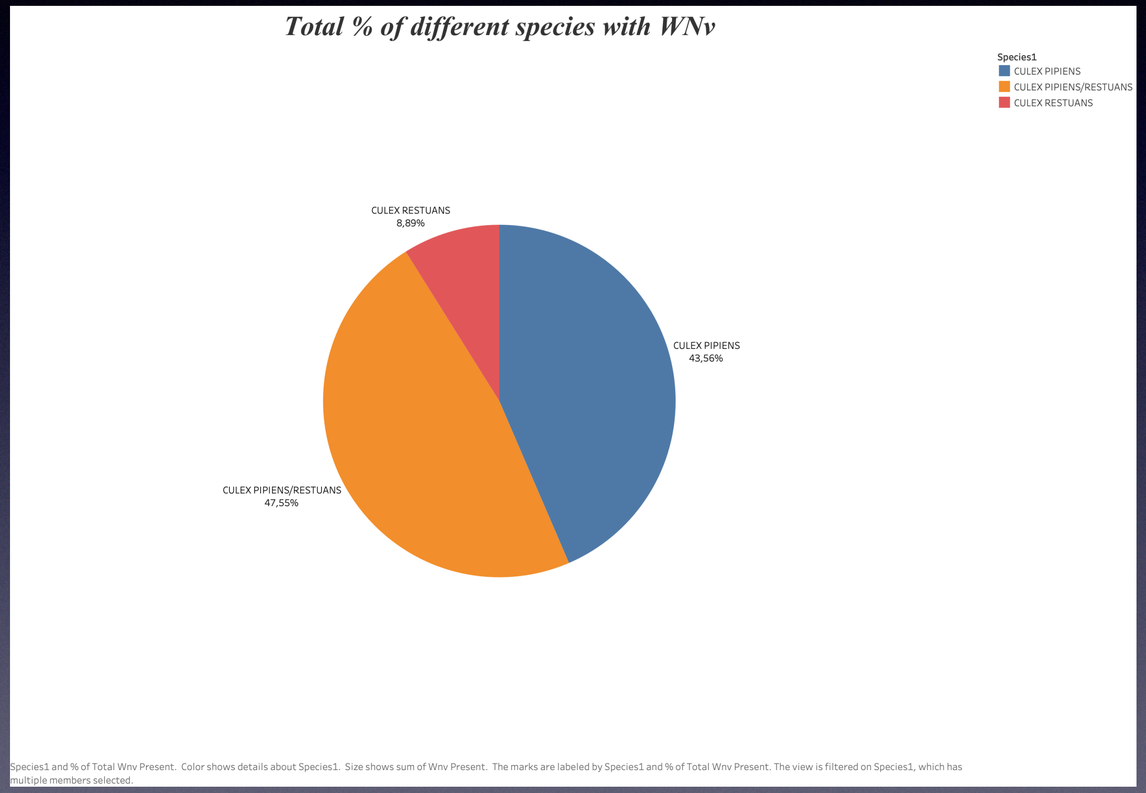
Spray Locations in Chicago

Exploratory Data Analysis - Time



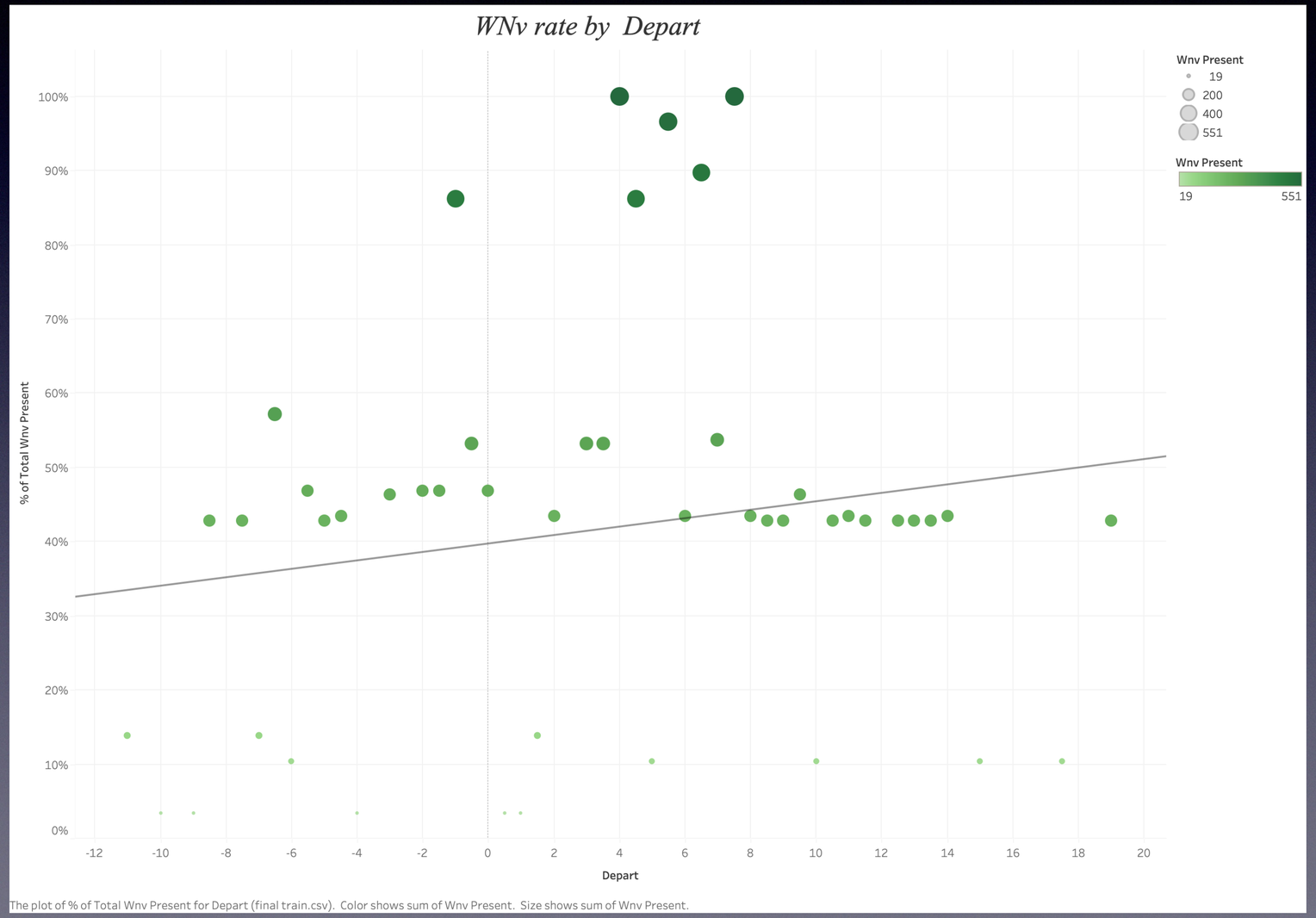
AUGUST peaks!

Explanatory Data Analysis - Species



CULEX PIPIENS/ RESTUANS - Major carriers

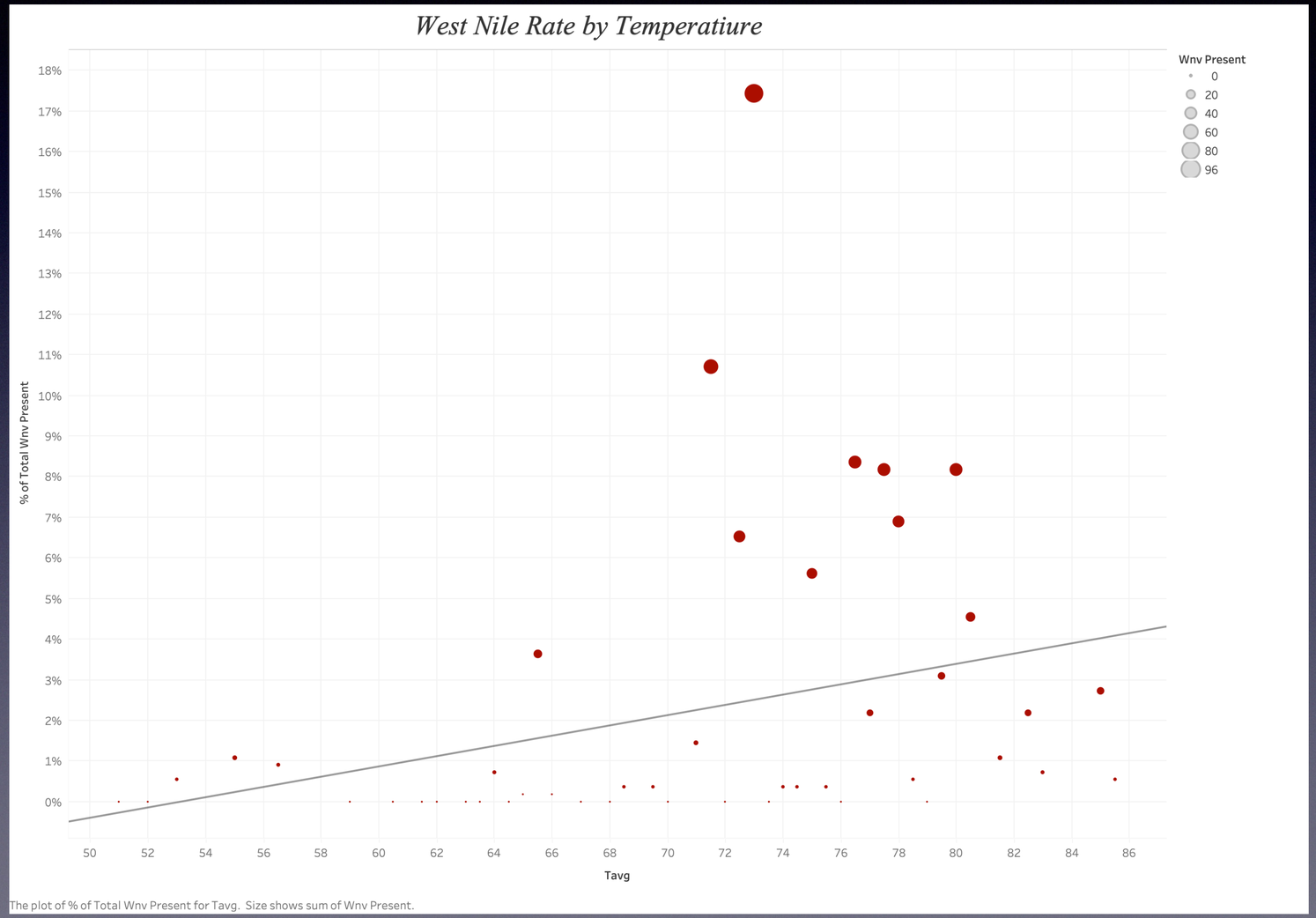
Exploratory Data Analysis - Depart



The plot of % of Total Wnv Present for Depart (final train.csv). Color shows sum of Wnv Present. Size shows sum of Wnv Present.

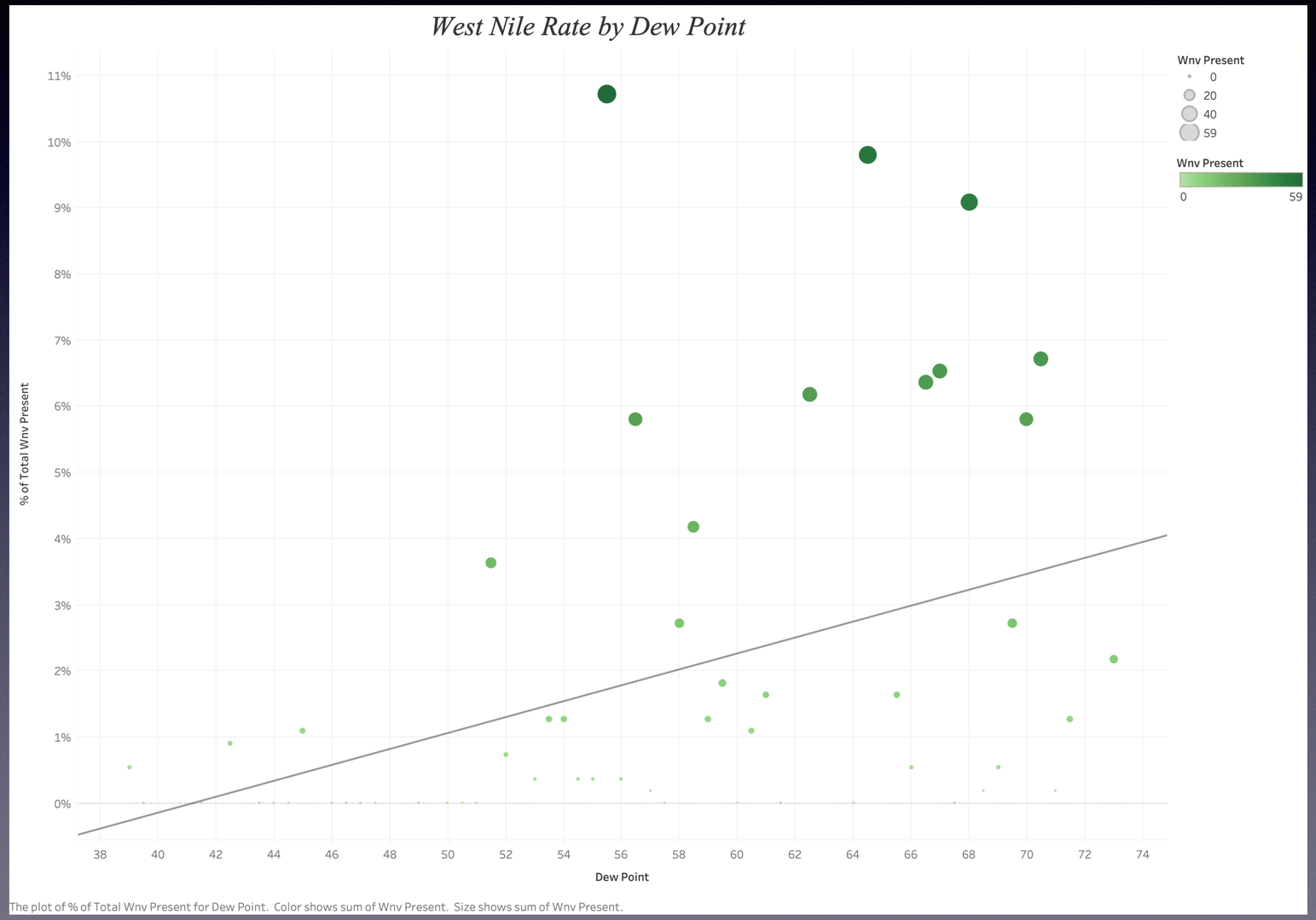
MORE depart
MORE WNv
occurrence

Explanatory Data Analysis - Temperature



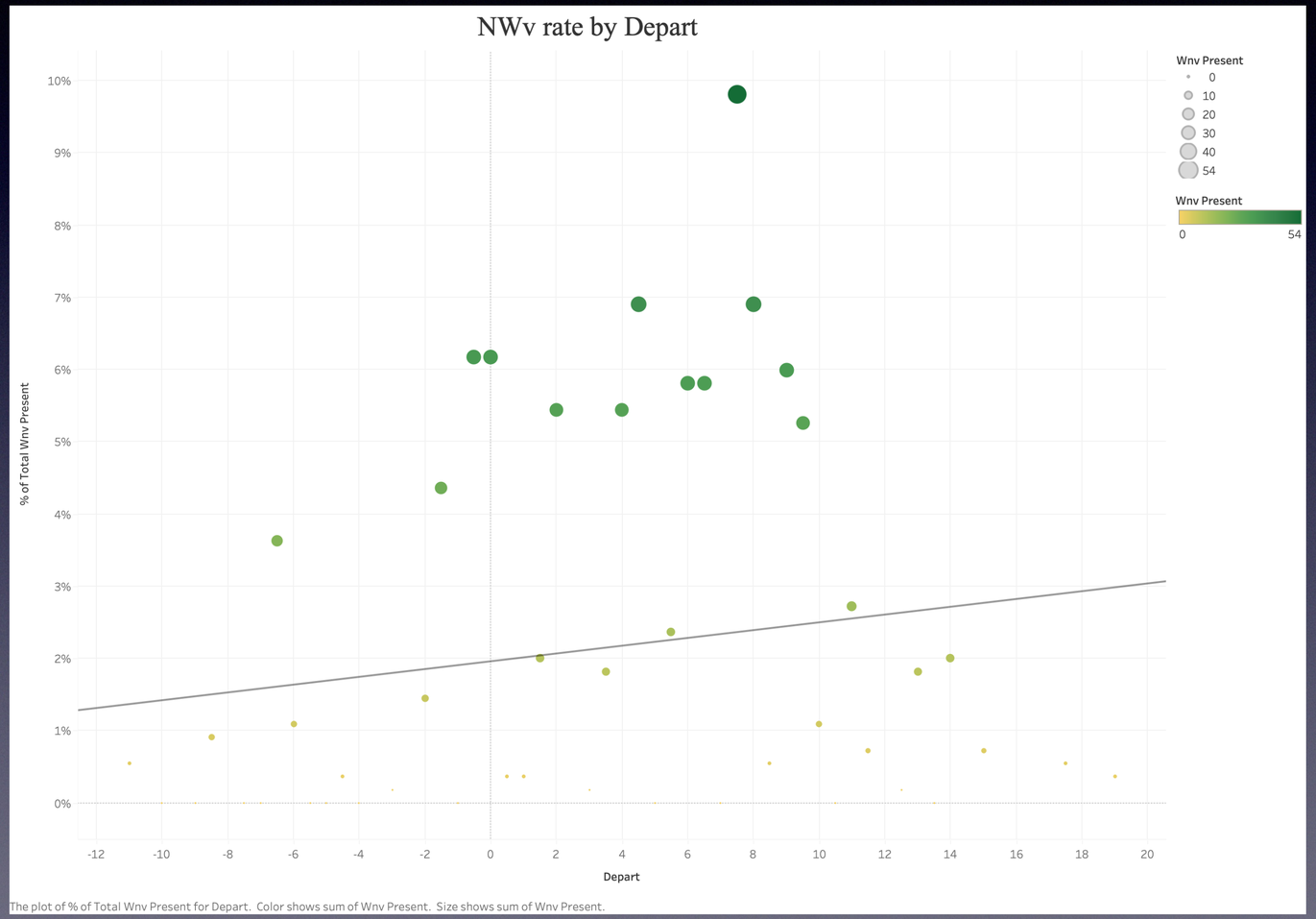
HIGHER temperature
HIGHER WNv rate

Explanatory Data Analysis - DewPoint



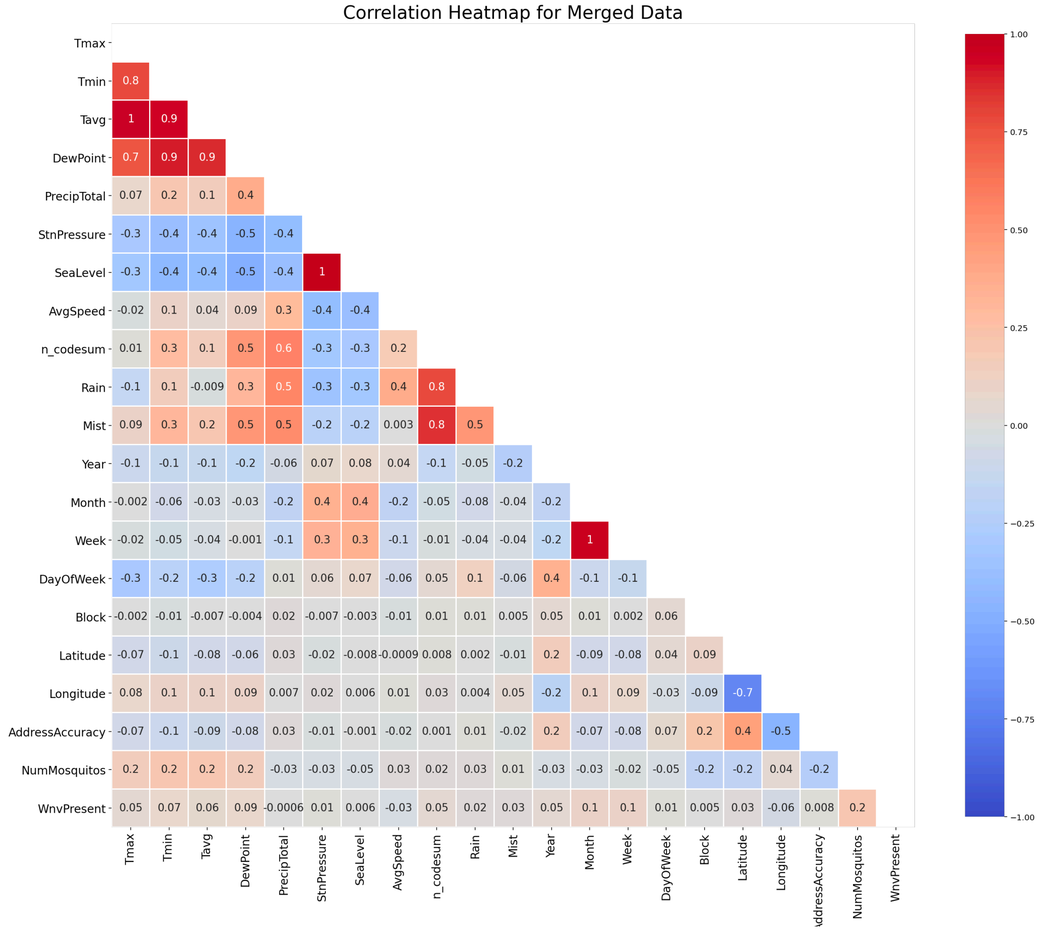
HIGHER dew HIGHER WNv rate

Explanatory Data Analysis - Precipitation



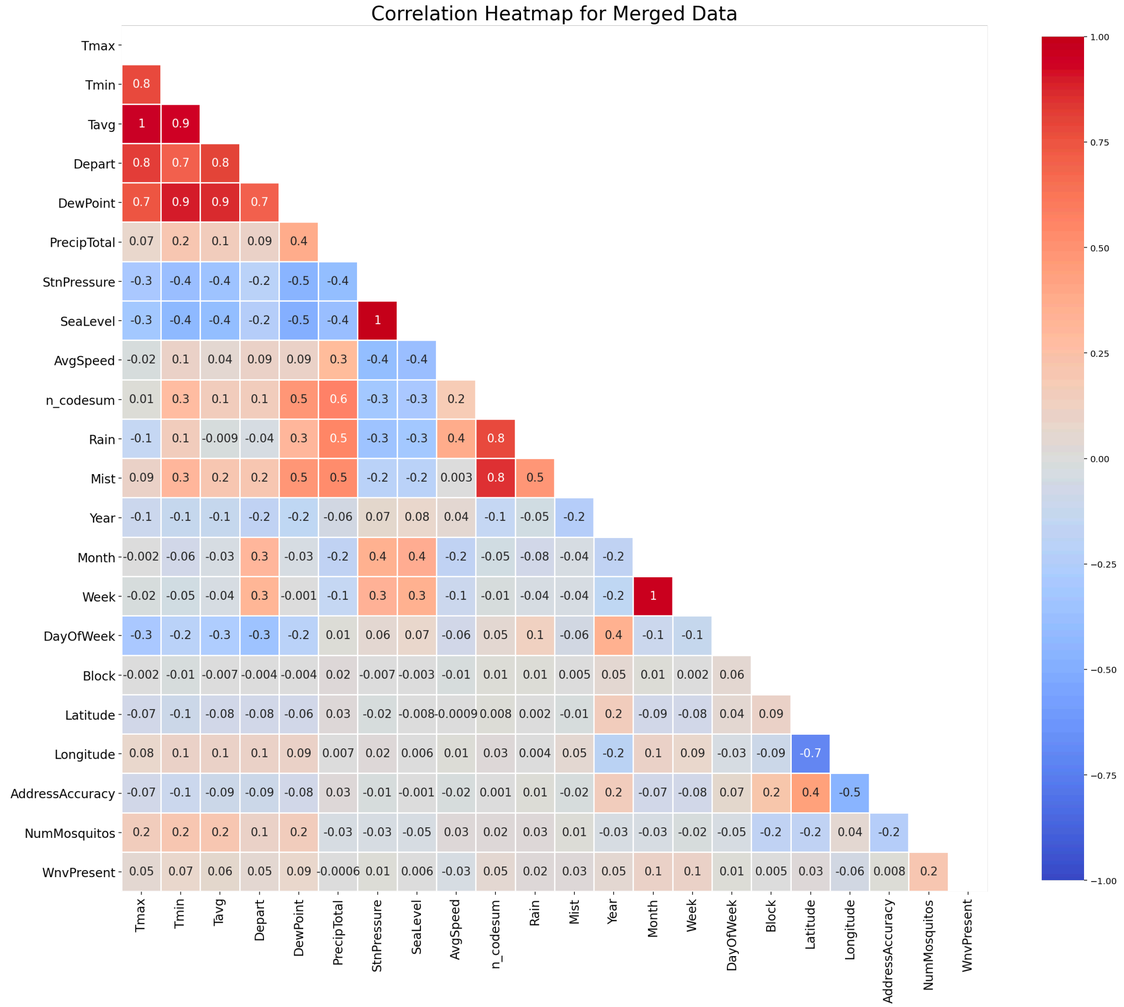
HIGHER precipitation HIGHER WNv rate

Exploratory Data Analysis



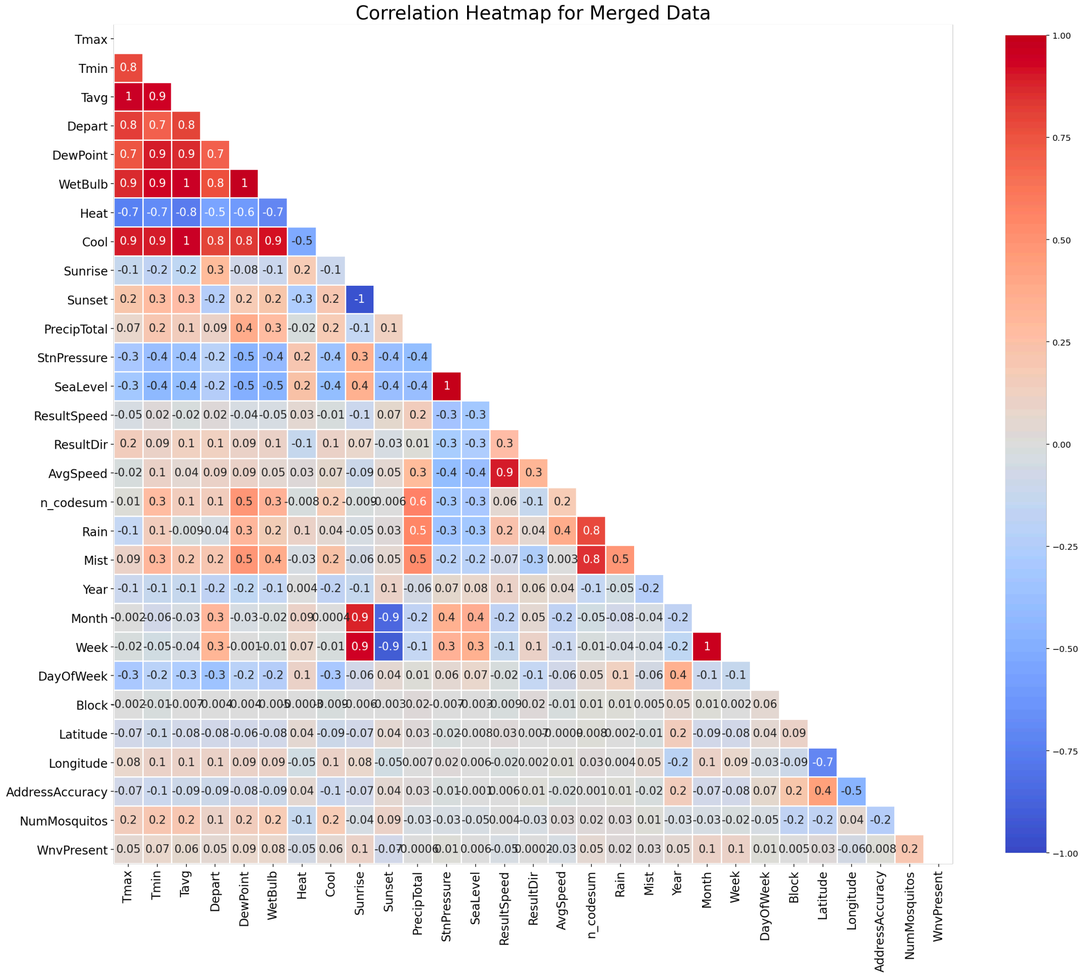
Heatmap without Depart, WetBulb, Heat, Cool, Sunrise, Sunset, Depth, Water1, SnowFall, ResultSpeed, ResultDir

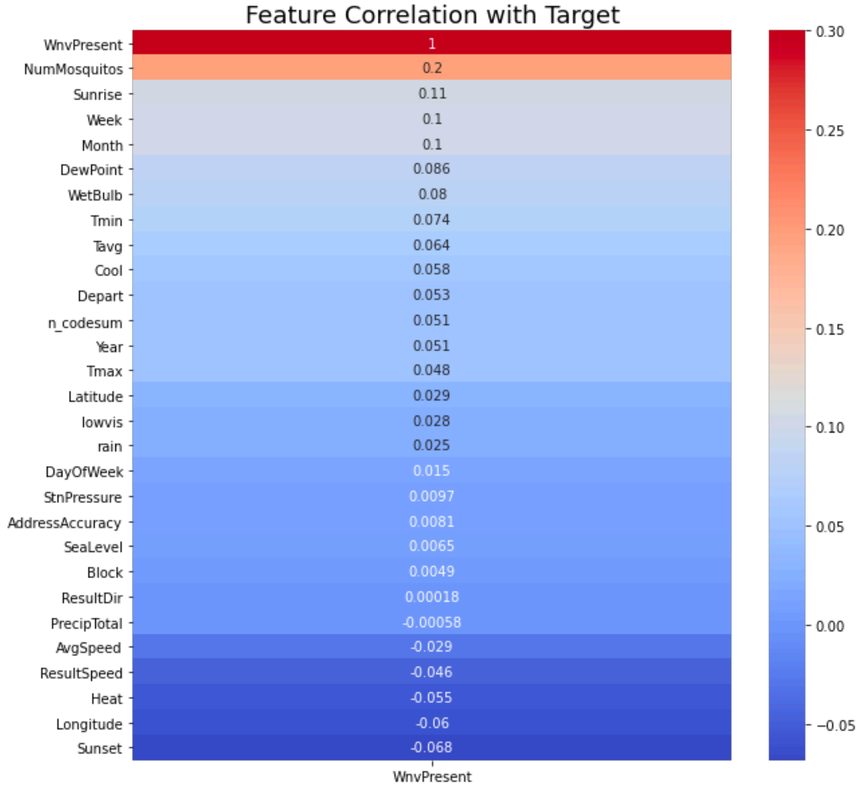
Exploratory Data Analysis



Heatmap without WetBulb, Heat, Cool, Sunrise, Sunset, Depth, Water1, SnowFall, ResultSpeed, ResultDir

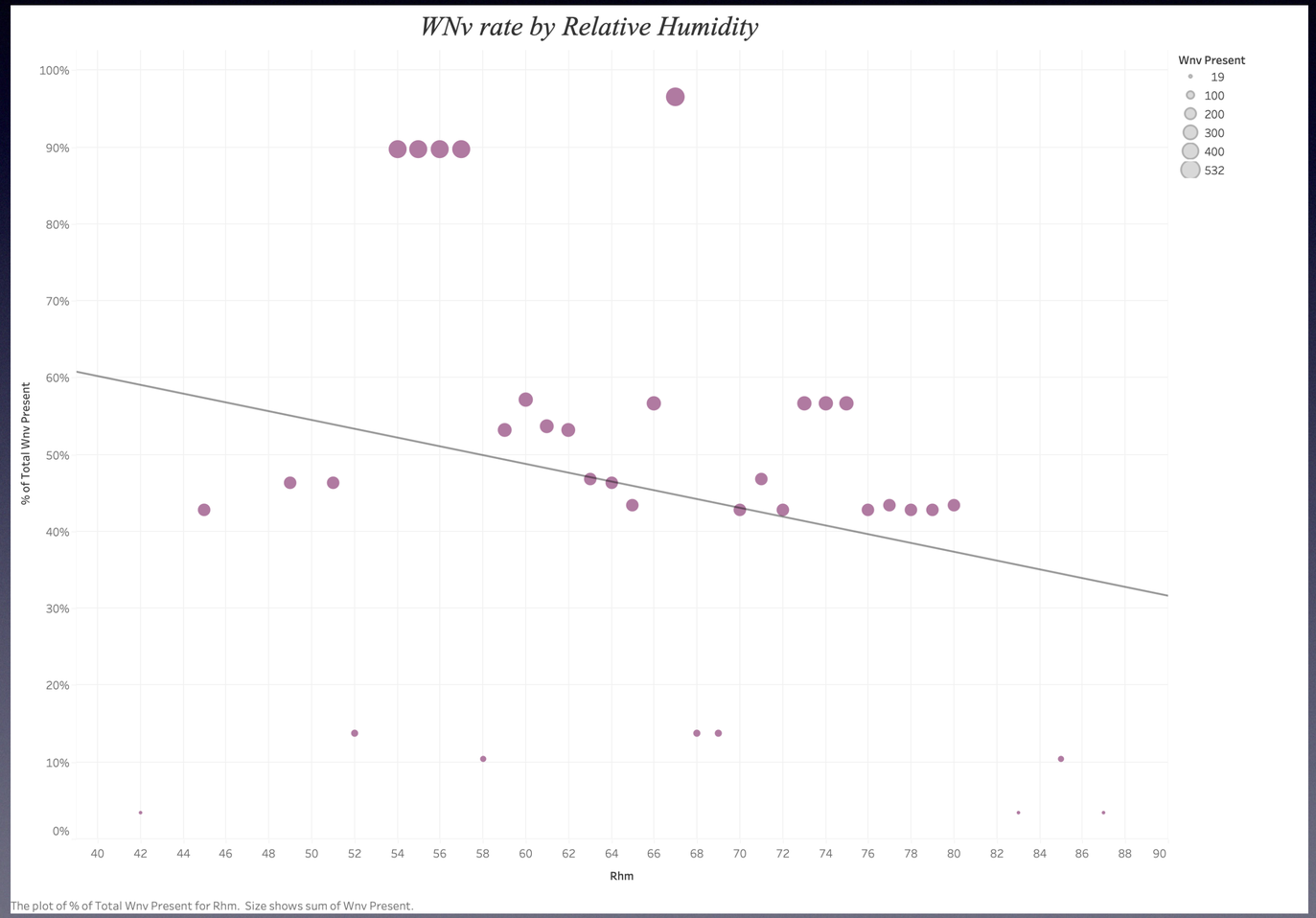
Exploratory Data Analysis





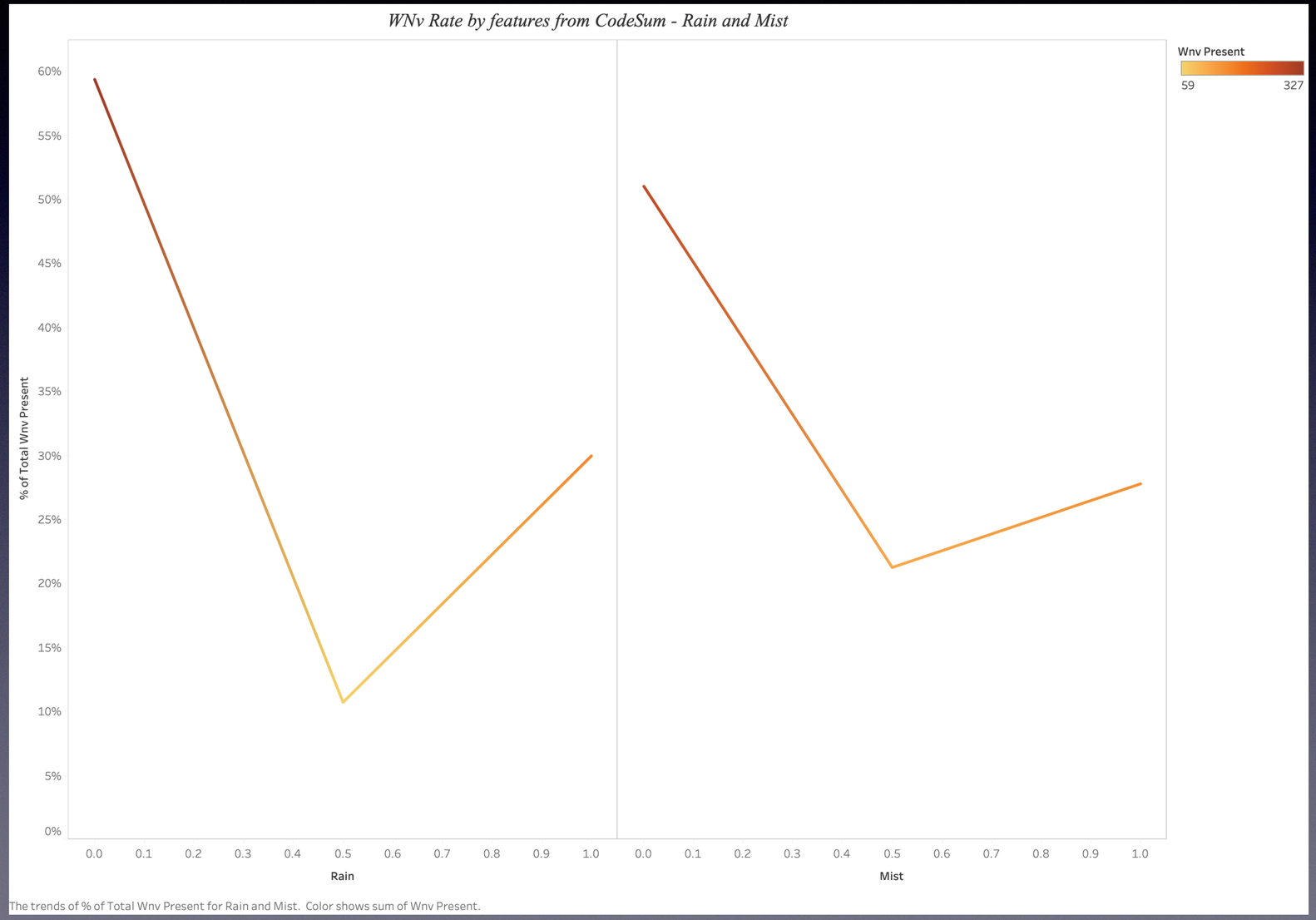
Heatmap with all the features and Feature Correlation

Feature Engineering- Relative Humidity



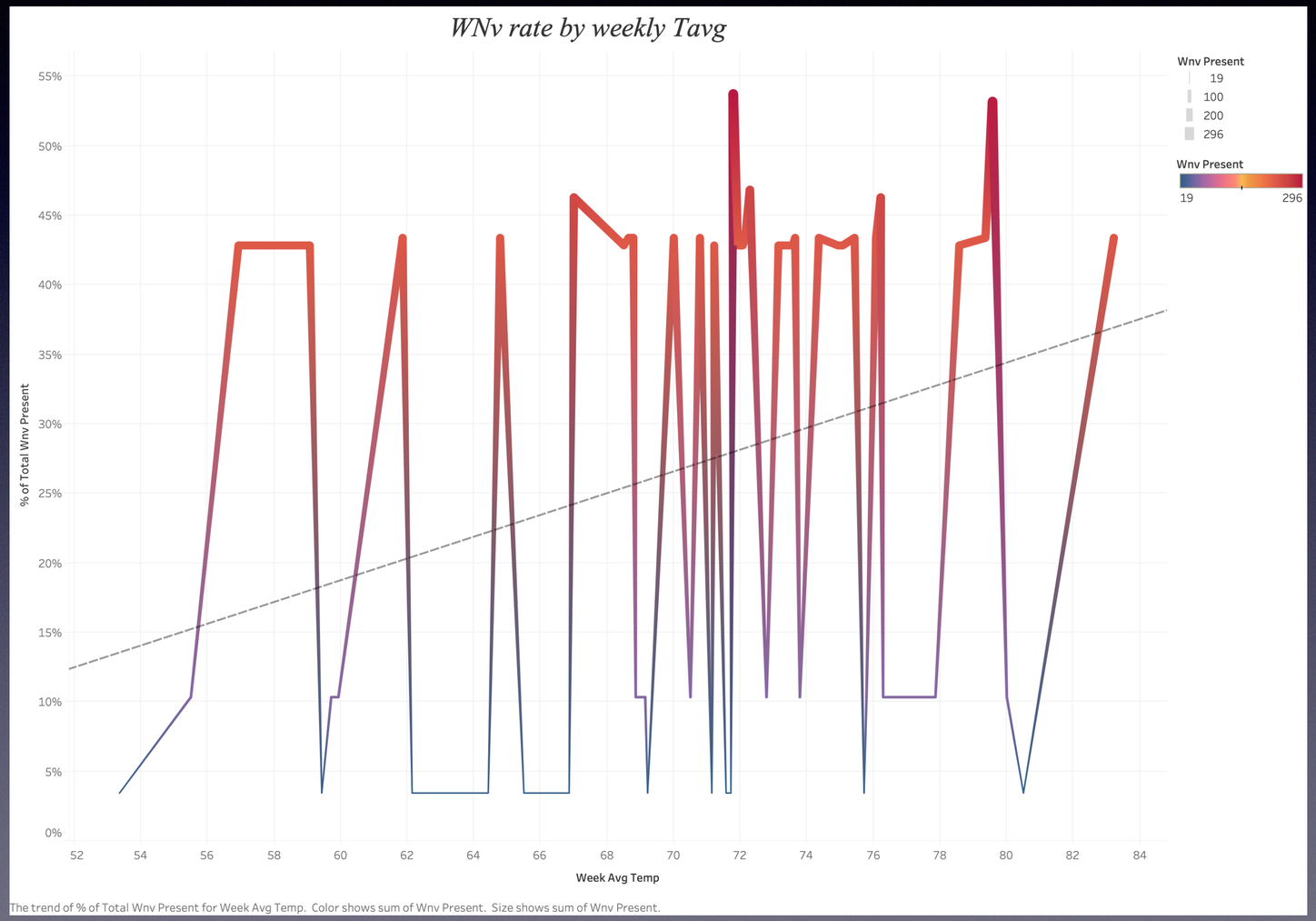
RAISE in Relative Humidity DECREASE in WNv rate

Feature Engineering - CodeSum



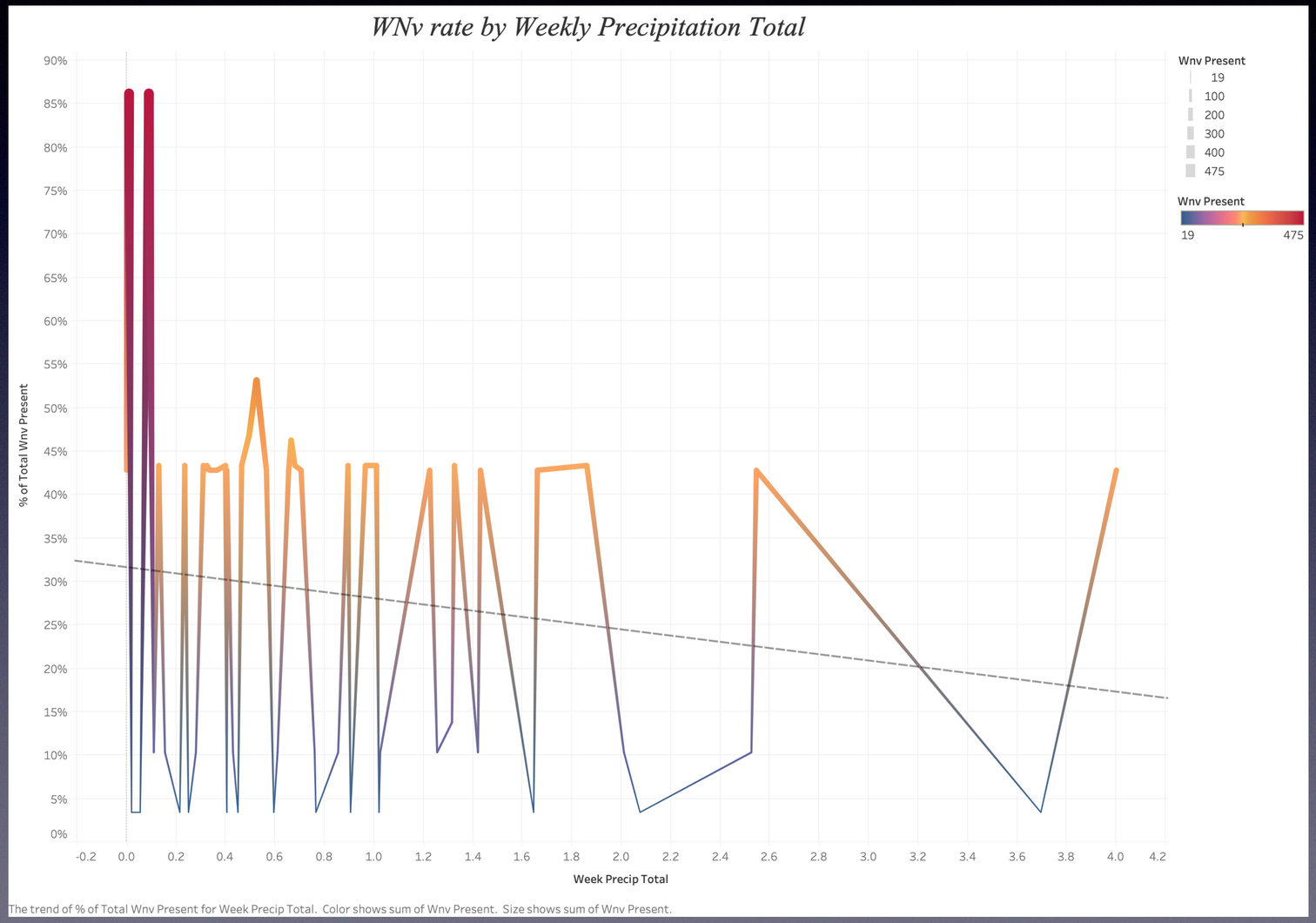
Decrease in WNv rate with Rain and Mist

Feature Engineering - Weekly Average Temperature



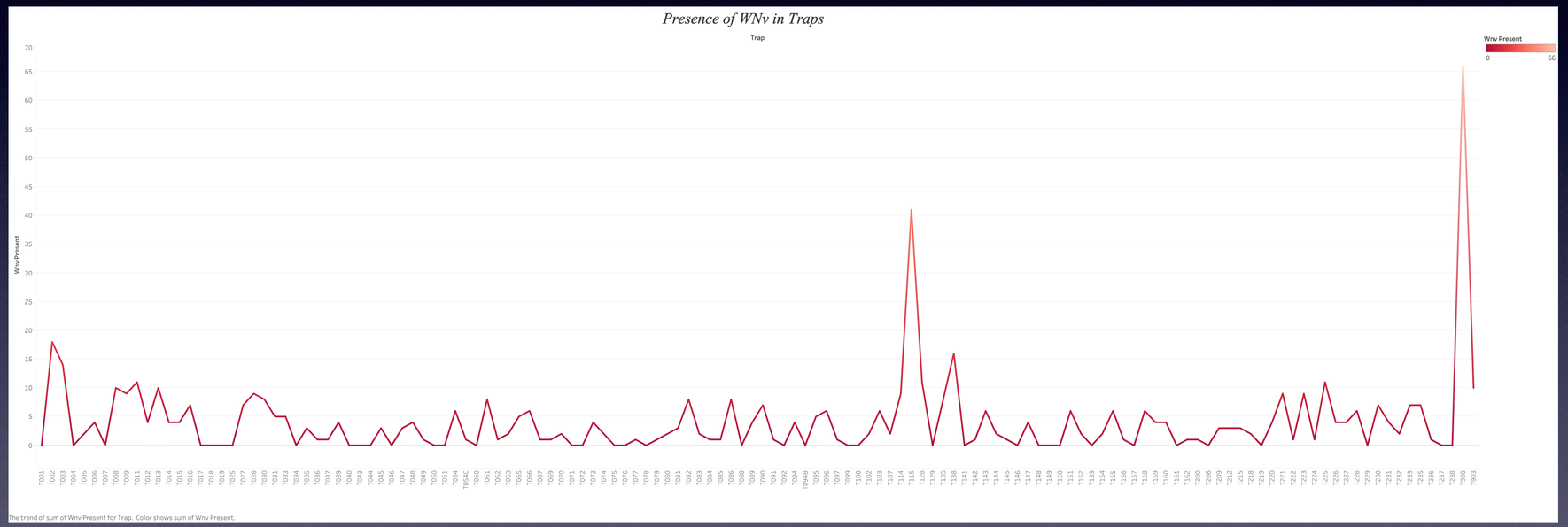
HIGHER temperature
HIGHER WNv rate

Feature Engineering - Cumulative Precipitation



HIGHER precipitation DECREASE in WNv rate

Feature Engineering - Traps



Graph showing the presence of WNv in Traps

Model Building

Model Building

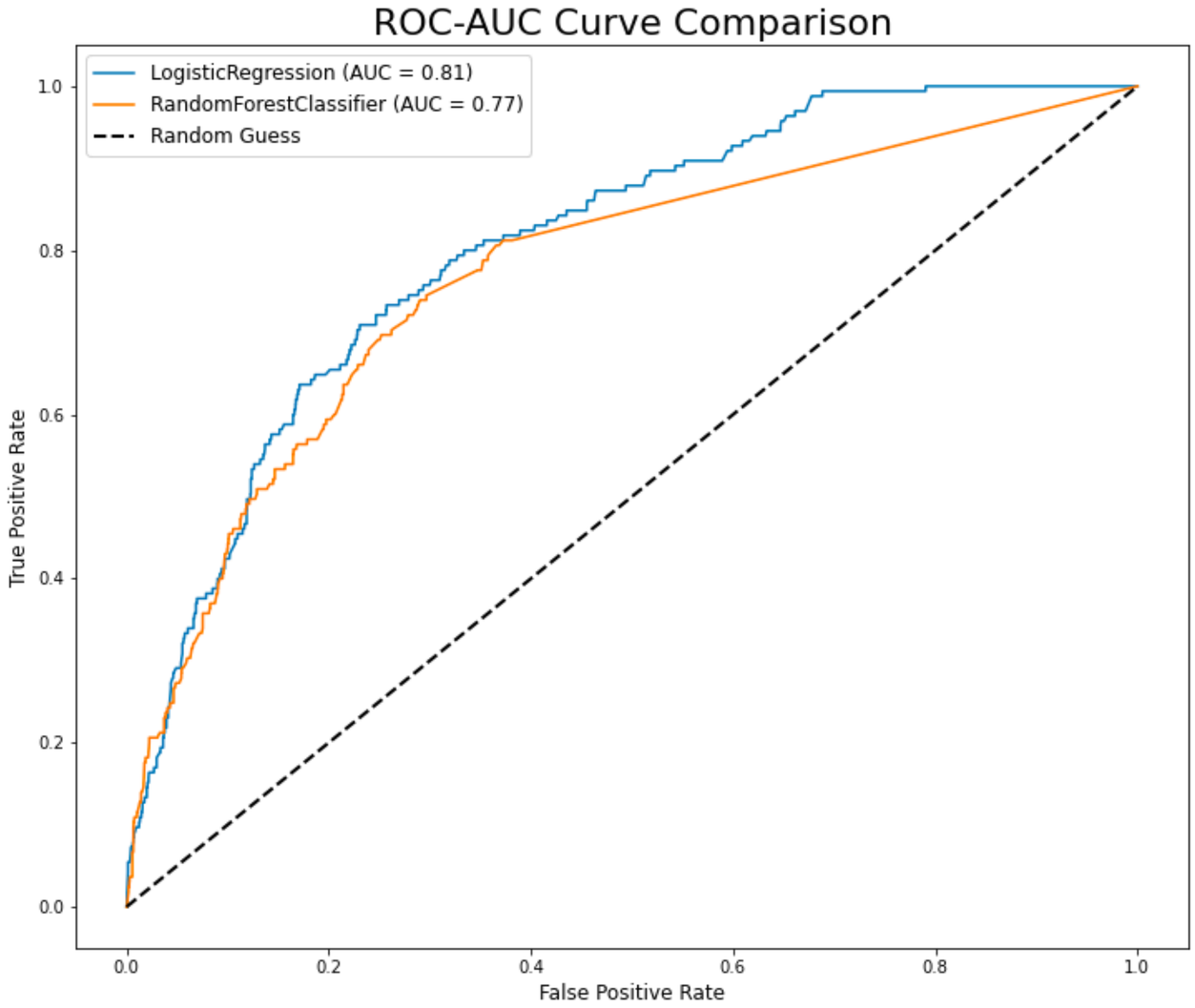
Weather

- ☆ Temperature
- ☆ Sunrise
- ☆ Dew Point
- ☆ PrecipTotal

Trap

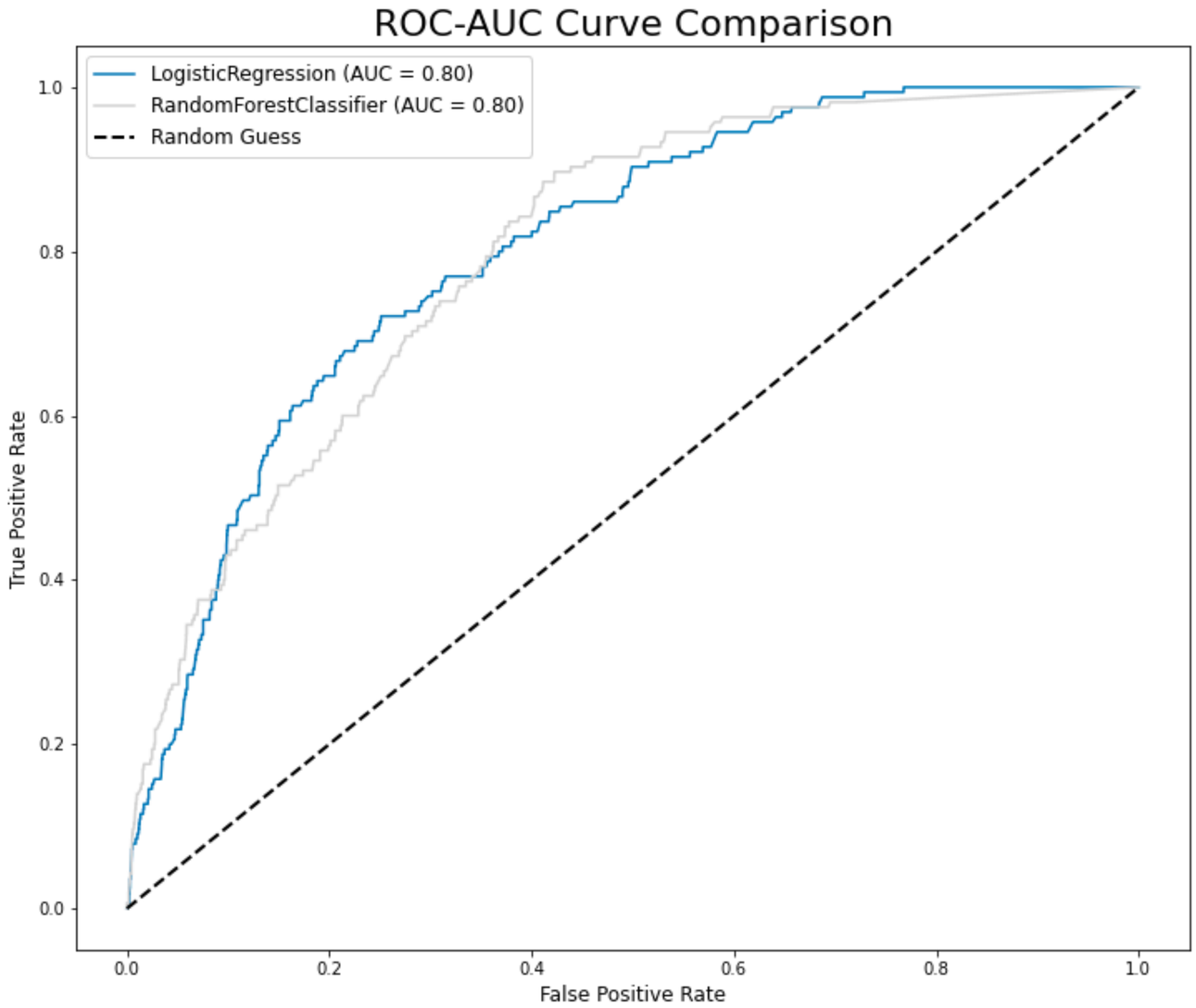
- ☆ Location
- ☆ Species
- ☆ Month

Baseline - Logistic Regression



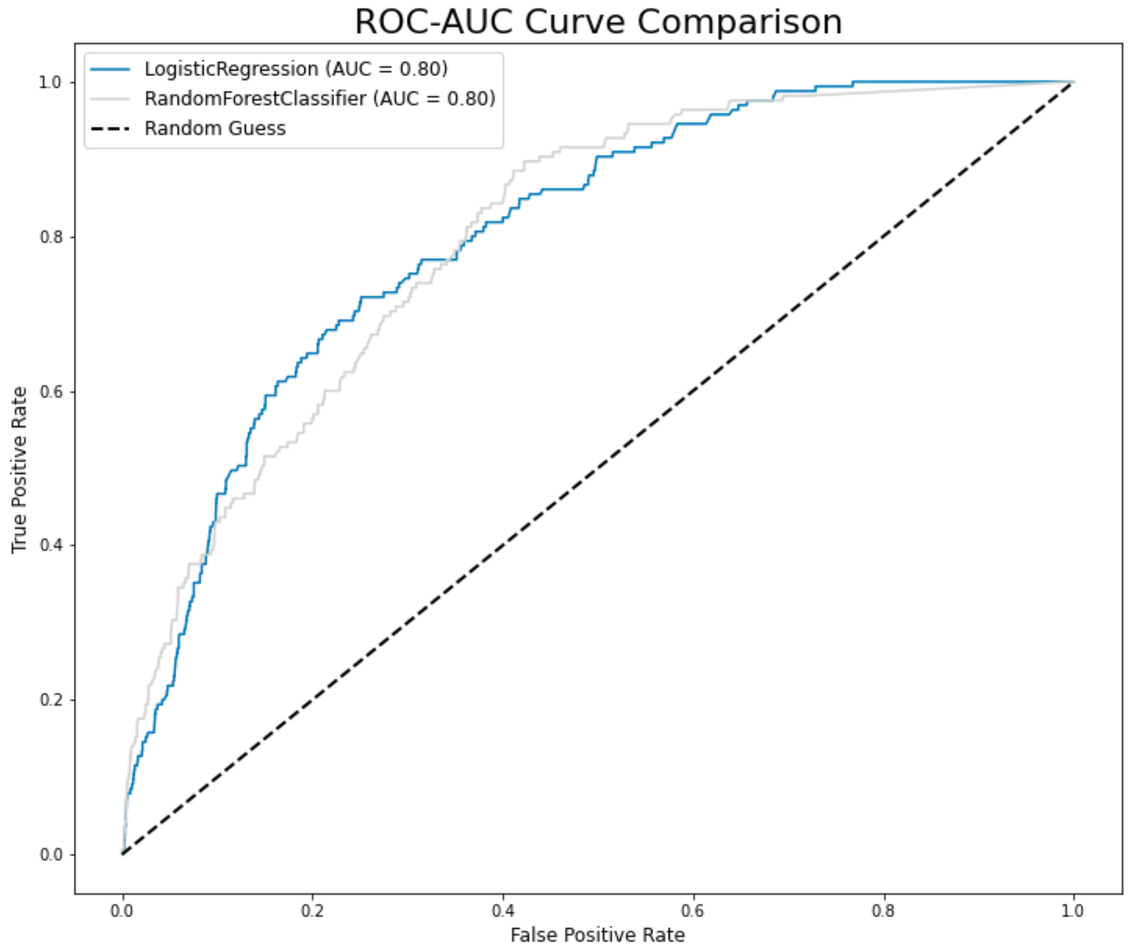
Score:
0.804625

Baseline- Random Tree Classifier



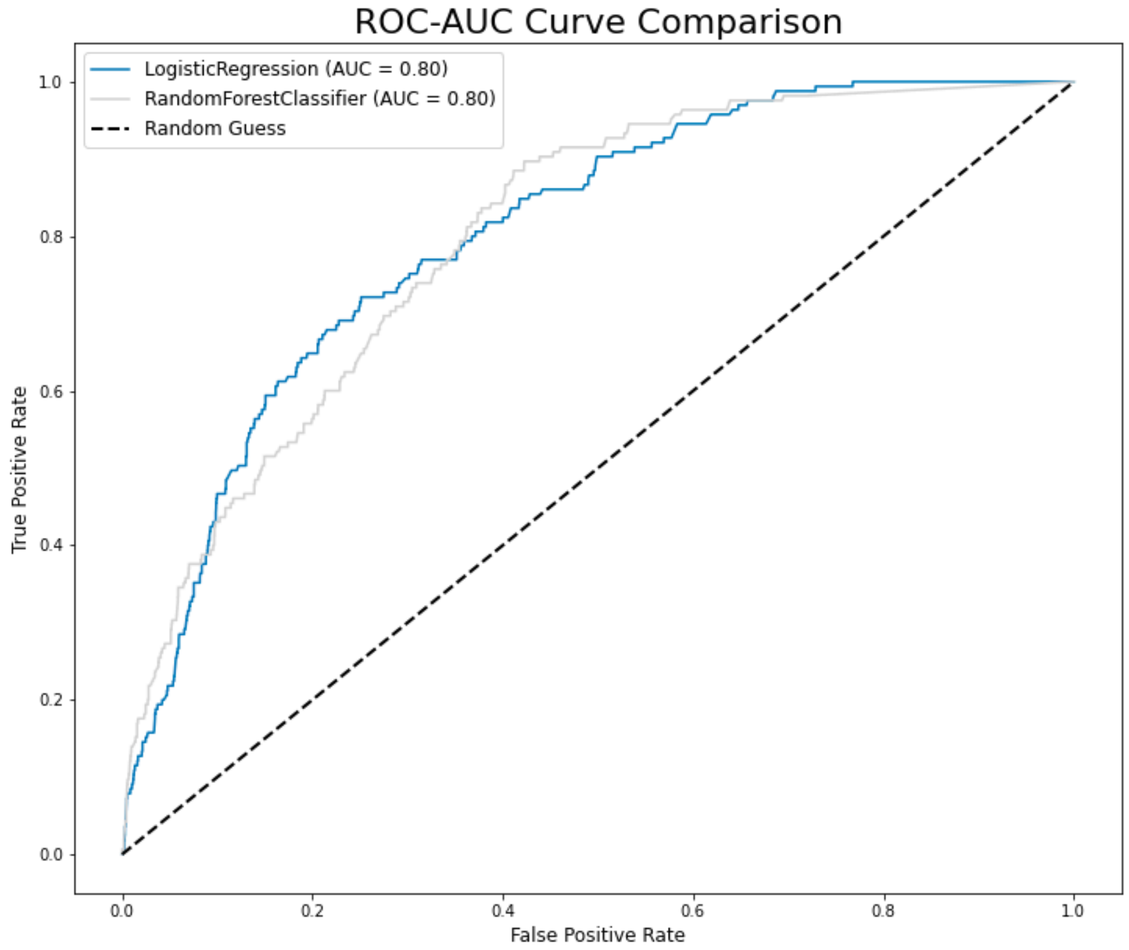
Score:
0.772755

SMOTE - Logistic Regression



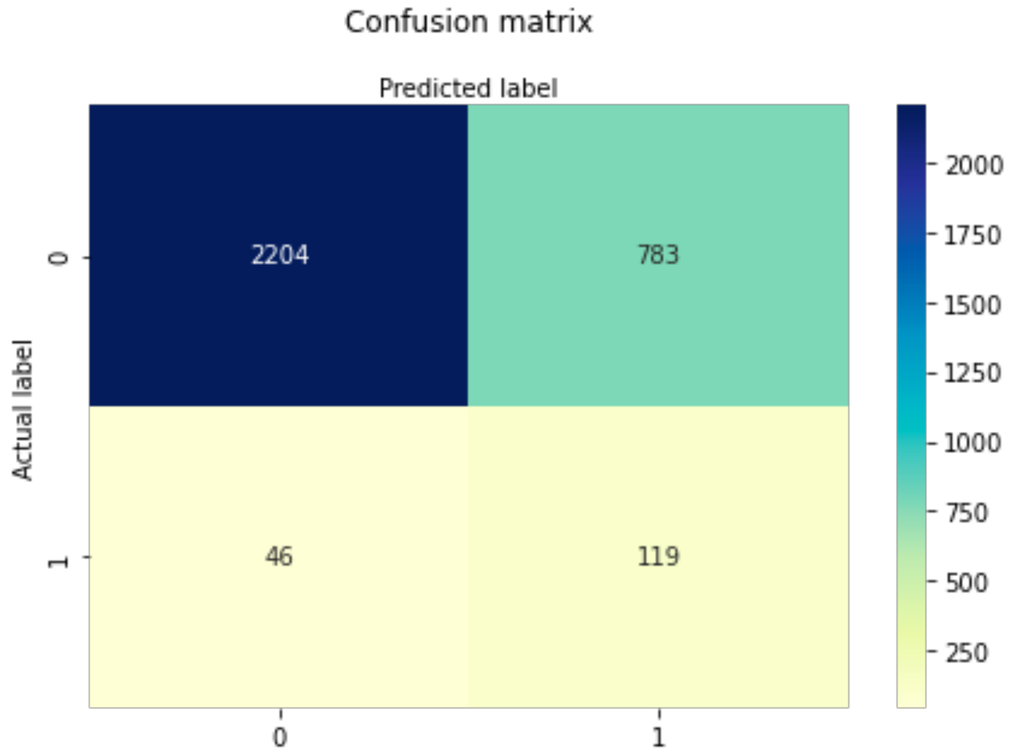
Score:
0.803751

SMOTE - Random Tree Classifier



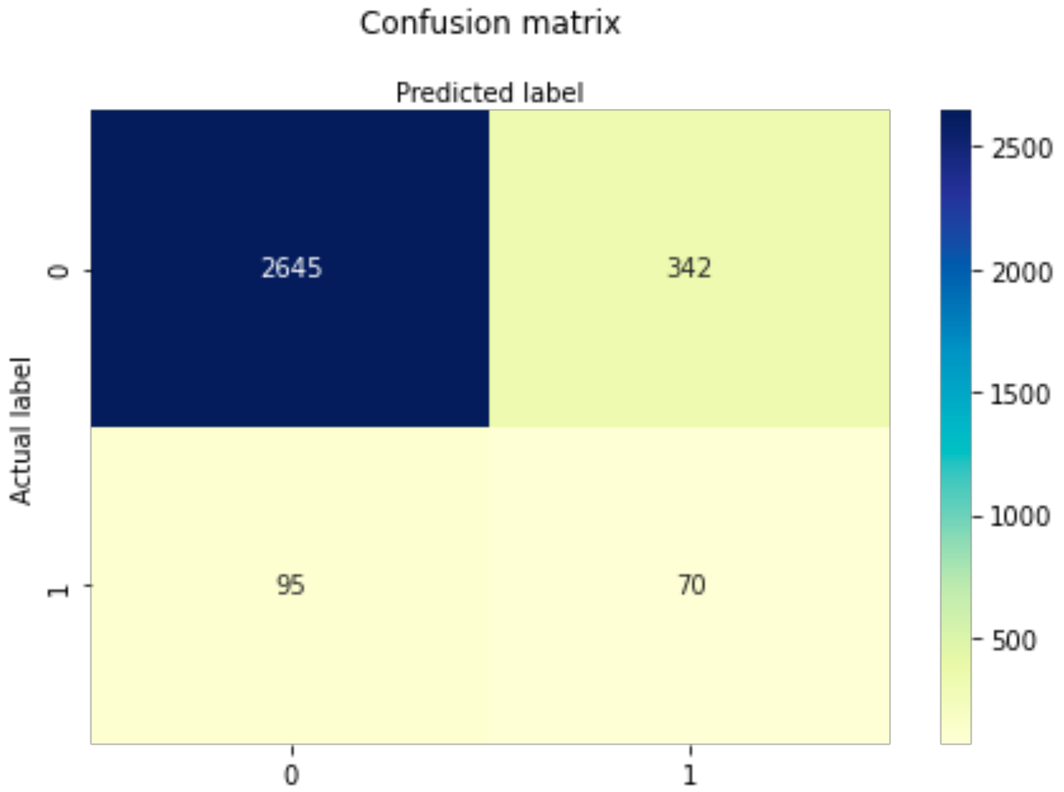
Score:
0.796314

Model Evaluation - Logistic Regression



Accuracy:
0.7369992

Model Evaluation - Random Tree Classifier



Accuracy:
0.861358

Summary

- Based on AUC scores Logistic Regression model should be selected.
- But when considering the accuracy of the predictions the Random Forest Classifier gives more accurate predictions.
- The final selected model was Random Classifier, with a test AUC of 0.796 and accuracy of 0.861.
- The WNv occurrence for the years 2008, 2010, 2012 and 2014 are predicted.
- Thus using the predicted values, we can further identify the potential locations of WNv outbreaks and help the Chicago Department of Public Health (CDPH)to take necessary precautions.
- These predictions might be helpful for further Analysis.

Tools Used

- Language Used: Python 3
- Tools and IDE: Tableau for EDA, Jupyter Notebook
- Libraries Used: Pandas, Numpy, Matplotlib, Seaborn, Sklearn, pandas_profiling, imblearn