

MIS 536

# WNV Mosquito Test Results Analytics

---

Team 4

Fall 2019

## Data Analytics Project - Checkpoint 1

1. How does your chosen topic and identified data and supporting material satisfy each one of the 5 criteria below? Please see the explanation provided above for each criterion in the “Five Criteria for Appropriate Data ” guideline above.

- a. **Importance**

West Nile virus can cause a fatal neurological disease in humans. However, approximately 80% of people who are infected will not show any symptoms. Vaccines are available for use in horses but not yet available for people. (**Source: WHO**) West Nile virus (WNV) is the leading cause of mosquito-borne disease in the United States. Effective prevention of human WNV infections depends on the development of comprehensive, integrated mosquito surveillance and control programs in areas where the virus occurs. The data is about the West Nile Virus (WNV) Mosquito test results in the city of Chicago. The analytics aims to reduce the spread of WNV virus in Chicago by predicting the outbreak of the virus in mosquitoes. This would essentially help the Chicago Department of Public Health in preventing transmission of the virus.

- b. **Availability**

The data is publicly available and accessible online by URL:

<https://data.cityofchicago.org/Health-Human-Services/West-Nile-Virus-WNV-Mosquito-Test-Results/jqe8-8r6s/data>

The data is updated very recently. The latest update was on 26th September 2019.

- c. **Documentation**

This program includes the collection of mosquitoes from traps located throughout the city; the identification and sorting of mosquitoes collected from these traps; and the testing of specific species of mosquitoes for WNV. The details of each column in the dataset are recorded in URL:

<https://data.cityofchicago.org/Health-Human-Services/West-Nile-Virus-WNV-Mosquito-Test-Results/jqe8-8r6s>

- d. **Support**

The data is provided by the Chicago Department of Public Health. So the data source is reliable.

- e. **Size**

The data is 4,984 kilobytes in size, with 29,489 observations and 18 variables.

2. Describe your data properties, including the following, as much as possible.
  - a. **Data format (tabular, database or file format, etc.)**  
The data is available in a structured table in .csv file format.
  - b. **Data tables (how many, their content/organization, etc.)**  
The data consists of the WNV test of mosquito from the year 2007 - 2019. Each row represents the test record. It has 29,489 observations and 18 variables of information.
  - c. **Data columns (most important ones, etc.)**  
There are 18 columns and the most important column are,  
"Result" - which indicates if the presence or absence of the WNV virus in a mosquito.  
"Block" - which gives information on the locations of the traps.  
"Species" - the information on the species of mosquitoes.  
"Test Date" - more information on the time of the observation.
  - d. **Data rows (unit of observation, count, etc.)**  
There are 29,489 rows of test records every 20th week to 40th week from 2007 to 2019
3. Describe your data variables. Please use distribution statistics (mean, median, mode, percent missing, etc.) and distribution charts.
  - a. Categorical variables (nominal or ordinal)
  - b. Numerical variables (binary or interval)
  - c. Potential target variable

Our Data Variables are as follows:

Data Variable	Description	Data Type
<b>SEASON YEAR</b>	Years that the WNV test is performed	Categorical - Ordinal
<b>WEEK</b>	Weeks that the WNV test is performed	Categorical - Ordinal
<b>TEST ID</b>	ID for the test	Label Variable
<b>BLOCK</b>	Address of the trap location	Categorical - Nominal
<b>TRAP</b>	The ID of the trap	Categorical - Nominal
<b>TRAP_TYPE</b>	Type of trap	Categorical - Nominal
<b>TEST DATE</b>	Data when the test is performed	Categorical - Ordinal
<b>NUMBER OF MOSQUITOES</b>	Number of mosquitoes caught in this trap	Numerical - Discrete
<b>RESULT</b>	Presence or Absences of WNV in mosquitoes	Categorical - Nominal
<b>SPECIES</b>	Mosquito species that was caught	Categorical - Nominal
<b>LATITUDE</b>	Latitude of the trap	Numerical - Interval
<b>LONGITUDE</b>	Longitude of the trap	Numerical - Interval
<b>LOCATION</b>	Latitude and longitude of the trap	Numerical - Interval
<b>Wards</b>	The number of wards placed	Numerical - Discrete
<b>Census Tracts</b>	The population of a county	Numerical - Discrete
<b>Zip Codes</b>	Zip Code where the trap is placed	Categorical - Nominal
<b>Community Areas</b>	Community Areas where the trap is placed	Numerical - Discrete
<b>Historical Wards 2003-2015</b>	The number of wards placed 2003 to 2015	Numerical - Discrete

SEASON.YEAR	WEEK	TEST.ID	BLOCK
Min. :2007	Min. :20.00	Min. :20000	100XX W OHARE AIRPORT : 2936
1st Qu.:2009	1st Qu.:28.00	1st Qu.:27656	127XX S DOTY AVE : 785
Median :2012	Median :31.00	Median :35018	101XX S STONY ISLAND AVE: 623
Mean :2012	Mean :31.01	Mean :35033	41XX N OAK PARK AVE : 593
3rd Qu.:2016	3rd Qu.:35.00	3rd Qu.:42458	52XX S KOLMAR AVE : 510
Max. :2019	Max. :40.00	Max. :49785	70XX W ARMITAGE AVE : 480
			(Other) :23318
TRAP	TRAP_TYPE	TEST.DATE	NUMBER.OF.MOSQUITOES
T115 : 785	CDC : 1256	08/15/2007 12:08:00 AM: 276	Min. : 1.00
T002 : 589	GRAVID :27719	08/03/2012 12:08:00 AM: 245	1st Qu.: 2.00
T138 : 551	OVI : 1	08/21/2014 12:08:00 AM: 238	Median : 5.00
T114 : 500	SENTINEL: 269	07/27/2012 12:07:00 AM: 237	Mean :12.37
T151 : 480		08/14/2014 12:08:00 AM: 227	3rd Qu.:16.00
T008 : 473		07/09/2012 12:07:00 AM: 215	Max. :77.00
(Other):25867		(Other) :27807	
RESULT	SPECIES	LATITUDE	LONGITUDE
negative:26767	CULEX PIPIENS/RESTUANS:13204	Min. :41.65	Min. :-87.85
positive: 2478	CULEX RESTUANS : 9991	1st Qu.:41.73	1st Qu.: -87.75
	CULEX PIPIENS : 4858	Median :41.83	Median :-87.69
	CULEX TERRITANS : 897	Mean :41.84	Mean :-87.69
	CULEX SALINARIUS : 216	3rd Qu.:41.94	3rd Qu.: -87.63
	CULEX TARSALIS : 42	Max. :42.02	Max. :-87.53
	(Other) : 37	NA's :4398	NA's :4398
	LOCATION	Wards	Census.Tracts
	: 4398	Min. : 1.00	Min. : 6.0
(41.66238672759086, -87.59017972751752) : 785		1st Qu.:19.00	1st Qu.:176.0
(41.956298856118664, -87.79751744482932): 589		Median :29.00	Median :391.0
(41.71054240215372, -87.58455893336821) : 551		Mean :28.78	Mean :406.6
(41.79821072626856, -87.73692496319906) : 500		3rd Qu.:42.00	3rd Qu.:637.0
(41.91613471854847, -87.80109280863755) : 480		Max. :50.00	Max. :787.0
(Other) :21942		NA's :4398	NA's :4398
Zip.Codes	Community.Areas	Historical.Wards.2003.2015	
Min. : 4299	Min. : 4.00	Min. : 1.00	
1st Qu.:21202	1st Qu.:24.00	1st Qu.:20.00	
Median :21861	Median :48.00	Median :34.00	
Mean :19863	Mean :42.24	Mean :31.71	
3rd Qu.:22254	3rd Qu.:61.00	3rd Qu.:45.00	
Max. :22620	Max. :77.00	Max. :53.00	
NA's :4398	NA's :4398	NA's :4398	

Figure 1: Summary R output of WNV mosquito data table

As you can see from Figure 1 above, attributes of Latitude, Longitude, Wards, Census,Tracts,Zip.Codes, Community Areas, and Historical wards are all missing 4398 values, which is 14.91% of the whole data.

The target variables of the dataset are listed below which aims to gain a deeper understanding of the result of the West Nile virus (WMV).

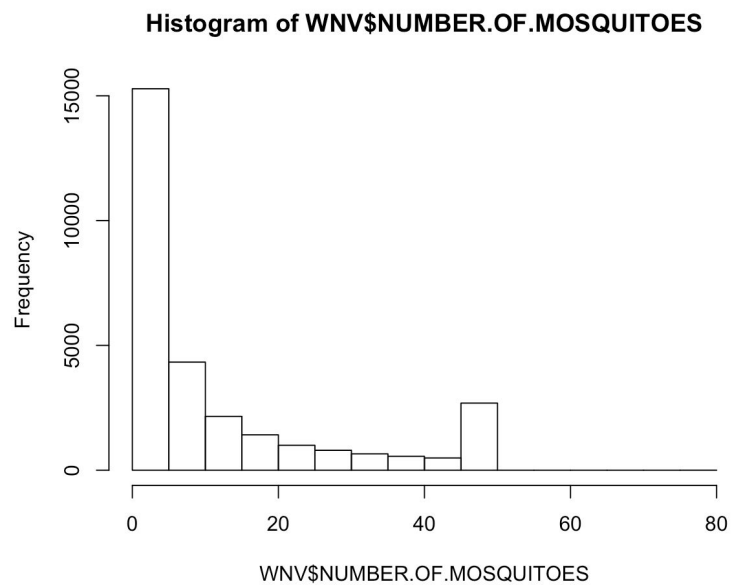
## NUMBER OF MOSQUITOES

SPECIES

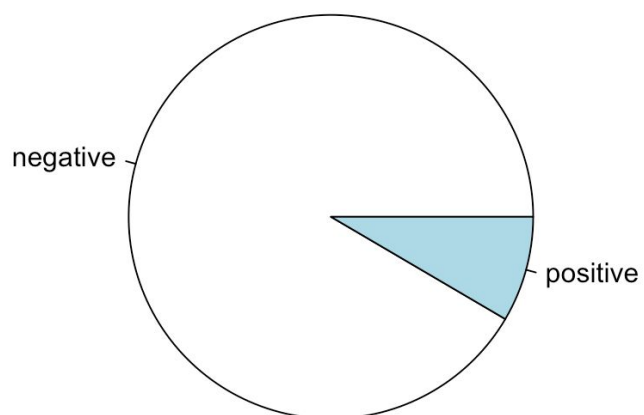
TRAP\_TYPE

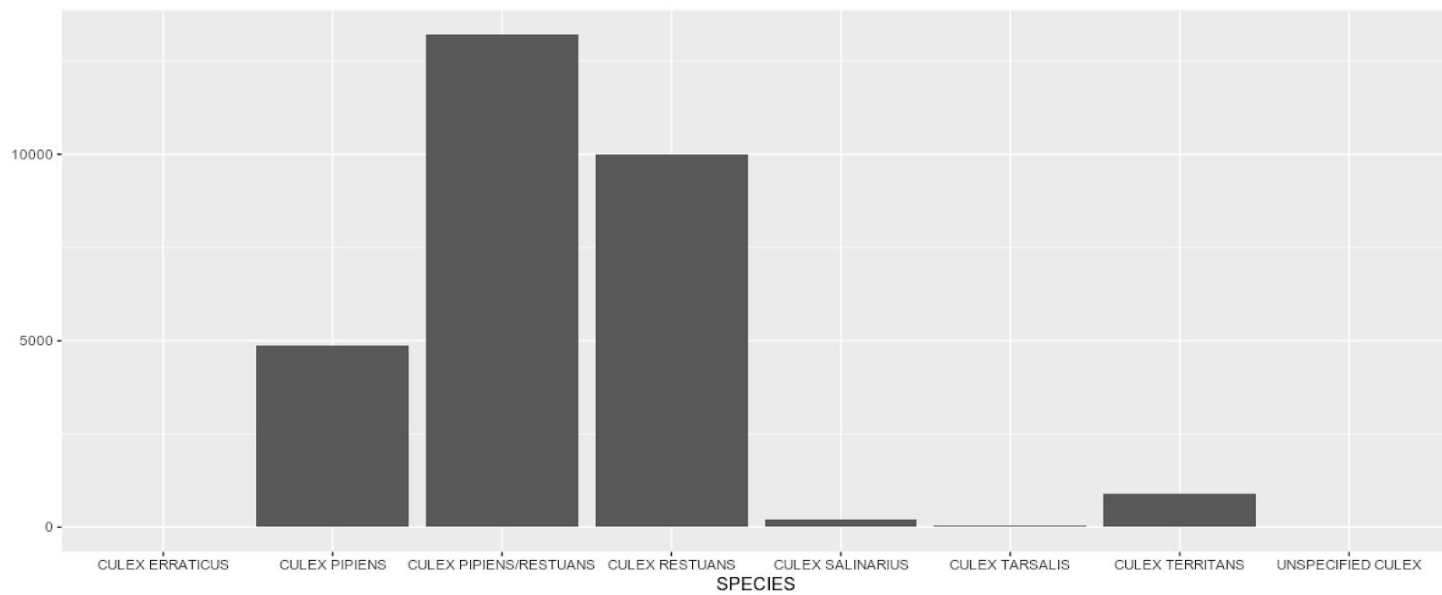
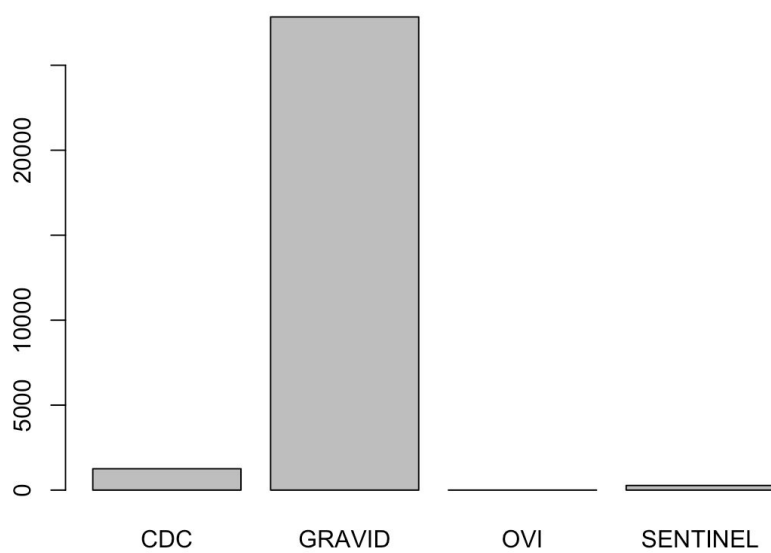
RESULT

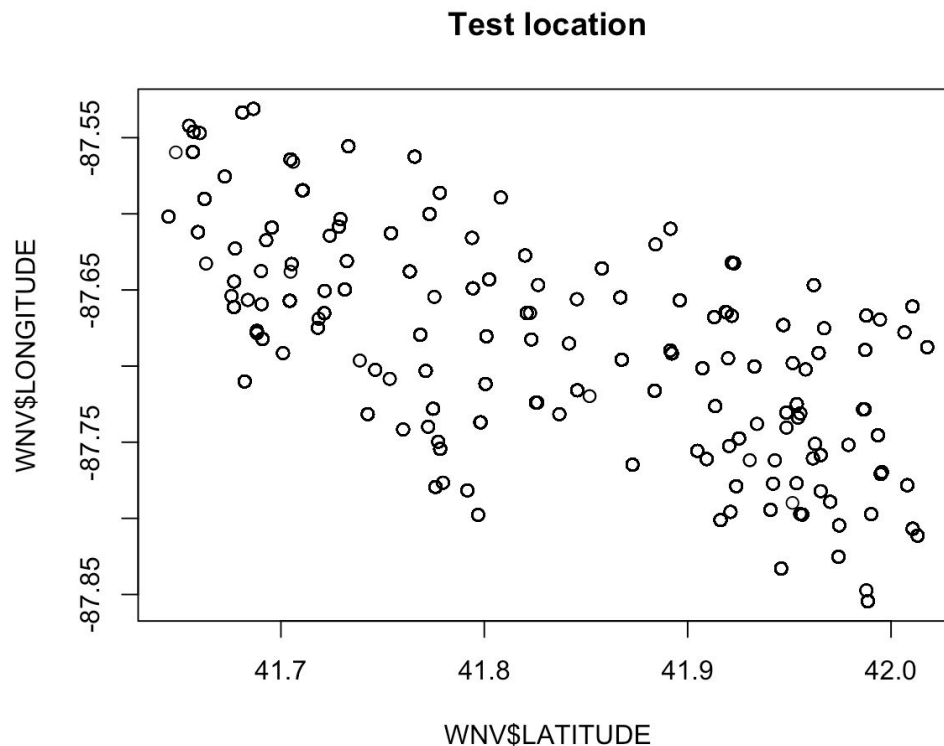
Location



## Test Result of WNV



**Barplot of Trap Type**




4. Propose potential questions you will answer with or insights you will gain from your data analytics.

Some of the questions we are looking to answer are as follows,

Which zip codes have the most cases?  
What are the most prominent species?  
Which year had the most negative result?  
What led to this? (Like special weather or events happened, fewer traps?)  
How can this be replicated? (Predict what will happen in the future)  
Which year had the most positive result? What led to this?  
How to resolve this?  
What time of the month is the infection the most?  
What time (months/quarter) has the most number?  
Generate a graph comparing a number of mosquitoes to Species type?  
Finding the ratio of the species from N number of mosquitos  
Finding the correlation between no mosquitoes and no. of species.  
Generate a graph showing the areas where species are most prominent?  
Which trap type was used the most?





Which trap is most efficient?

Is weather/rain associated with the number of mosquitos?

See the trend of results, increasing or decreasing year on year. Visually through hot zoning areas.

Can weather and area information be used to predict areas prone to infection and high time of infection?