

# Are Blacklegged Ticks the Upcoming Infestation? (Analysis on the Recent Rise in Black Legged Tick sightings in Toronto, Canada)

Adel Boufama 1001599059

2020-12-22

## Abstract

Black Legged Ticks (BLTs) pose an increasing threat to public health due to their bites having the potential to cause Lyme disease from the bacteria (*Borrelia burgdorferi*). Analyzing a dataset which includes Tick location data from the city of Toronto, Ontario, Canada, it is found that this dataset provides inconclusive results by using a logistic regression against the number of positive ticks found. Despite this, there are some notable findings from this data which will be discussed in this report. Some next steps would be to conduct a more thorough study resulting in a better dataset to analyze this topic further.

## Introduction

In recent years, you might have been noticing more news coverage and concern about ticks and the dangers they pose. This is not just your imagination, Black Legged Ticks (BLTs) have indeed been increasing substantially in population, especially in the past decade or so. This is not a good sign and a sobering fact because BLT bites can cause Lyme Disease through transmission of the bacteria *Borrelia burgdorferi* which is also on the rise as a result. Lyme Disease can be very debilitating and long lasting if it is not properly treated and turns chronic. Some people take years or even decades to recover from chronic Lyme Disease (Giorgi, 2018).

Some of the reasons that BLT populations are rising are climate change due to ticks thriving in warmer weather and deer population fluctuations as deer are the primary hosts of these ticks. Nevertheless, prevention is better than the cure, especially with the current medical capabilities where not all Lyme disease cases are treatable.

Due to the importance of this threat to public health, a study was done to analyze data from Toronto's tick surveillance to find out if tick numbers are indeed increasing in the city, which locations in the cities are more likely to have ticks and whether or not those ticks have the Lyme disease bacteria *Borrelia burgdorferi*. This report covers the dataset I use in detail, the type of mathematical model I use to analyze the data, the results obtained and discussion pertaining to the results and next steps.

## Data

The data set chosen was from Toronto's tick surveillance program. This data contains information on the number of blacklegged ticks, their locations and the number of them that carry the bacteria that causes Lyme disease (*Borrelia burgdorferi*). The dataset covers years 2013 through 2019 and the sample size is 191, meaning that the dataset contains 191 different locations that were searched for ticks. The ticks are found by tick dragging which is a process of collecting ticks in a specific location, it is done in the spring and fall. There does not seem to be a specific sampling technique used but perhaps it is closest to something resembling Cluster Sampling. This is because tick dragging is done in various locations around the city so that presumably, there is the same probability of finding ticks each time tick dragging is done in a particular location.

The population is every single Blacklegged Tick in the city of Toronto, the frame is all the Blacklegged Ticks in the locations that were visited to collect the ticks for this dataset. Finally, the sample is all the Blacklegged Ticks that were ultimately collected, studied and ended up in the dataset used for this report. Some of the key features of this dataset are the year, exact location of each tick in longitude and latitude values, the parks they were found at and whether or not the tick has the bacteria that causes Lyme disease. Some of the issues with this dataset are related to the consistency of data. It seems that there were not a lot of BlackLegged Ticks collected for most of the locations. This is the case because the tick dragging process can easily miss locations where there are a very small number of ticks and it can be an issue because small outliers might skew any correlation that we try to make.

The variables in this dataset that will be used for the main part of this analysis are Latitude, Longitude, Total BLTs, Positive\_Found and Year.

The Latitude and Longitude variables are pretty straightforward, they help find if there is a trend with location for number of ticks found or # positive. Each is in a numerical value.

The Total BLTs variable contains the total number of Black Legged Ticks found in any particular location, this variable is a combination of BLT Larvae and BLT Adults and Nymphs, I chose total BLTs over the other two or all together because the goal is not to analyze the life stage each BLT is at, but the overall number and trends.

I will create the variable: Positive\_Found from the older variable "# Positive" which is the number of ticks in a location that test positive for the Lyme disease causing bacteria (*Borrelia burgdorferi*). It is created as the binary response variable for the logistic regression. This was chosen because only ticks that contain the Lyme disease bacteria pose a threat and most of the locations have 0 positive findings which makes sense for this variable to be binary.

Year is simply the year 2013 to 2019 those ticks were collected. It is important to analyze this variable to see if there is a trend with time.

## Model

The model used for this analysis is a Logistical Regression. This models the relationship between a binary response variable and one or more explanatory variables. It is great for scenarios where one might need to predict a binary outcome given several factors that can be continuous or categorical. For example, when determining whether or not it will rain tomorrow in a city given factors that might influence the weather or in my case, whether or not a location has Ticks that test positive for Lyme disease bacteria given the location coordinates, the number of total BLTs found, and the year they were found. Putting this data in a logistic regression model can allow me to see if any of the variables above have a relationship with whether or not there are BLTs that have the Lyme disease bacteria.

Mathematically, this model is represented as:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

Notation Summary:

- $p$  is the probability of the event of interest occurring. In my case, this is whether or not infected Ticks are found.
- Beta 0 ( $B_0$ ) is the intercept of the y axis which represents the log odds of ticks being found to have been positive for Lyme disease bacteria given the default factors for every explanatory variable.  $x_1 = x_2 = \dots = x_k = 0$ .
- The rest of the coefficients  $B_1, B_2, \dots, B_k$  represent the change in log odds for every one unit increase or factor in the respective  $x_1, x_2, \dots, x_k$ .
- The predictor variables can be numerical or categorical

Based on the mathematical formula, logistical regression does not fit a line on the data plot like linear regression does, instead it fits an 'S' shaped natural logarithmic based function to the data. This model fits the curve using maximum likelihood estimation.

The main reason this model was used is because it is useful when working with the categorical variables that are in this dataset. An important variable was added to represent the response variable that was necessary for this model to work. It was the 'Positive\_Found' response variable, this was discussed in the Data section above, it replaces the "# Positive" variable. This was necessary because logistical regression requires the response variable to be binary. Finally, the software used to run this model is R run on RStudio.

Logistic regressions are great for probabilistic interpretation, they are less prone to over-fitting when dealing with fewer parameters and can be updated easily. A weakness however, is that this model tends to perform poorly when dealing with a high number of parameters.

## Results

### Logistic Regression Summary Statistics (Figure 1)

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##
## Call:
## glm(formula = Positive_Found ~ as.factor(Year) + Longitude +
##      Latitude + `Total BLTs`, family = "binomial", data = BLT_Data)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.7513  -0.2041  -0.1826  -0.0452   2.6696
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    40.9992   6568.0280   0.006 0.995019
## as.factor(Year)2014  -0.2387   7223.0211   0.000 0.999974
## as.factor(Year)2015  14.1420   6522.6389   0.002 0.998270
## as.factor(Year)2016  12.1202   6522.6390   0.002 0.998517
## as.factor(Year)2017  15.0586   6522.6387   0.002 0.998158
## as.factor(Year)2018  15.2780   6522.6387   0.002 0.998131
## as.factor(Year)2019  15.1902   6522.6387   0.002 0.998142
## Longitude         0.8137     5.4728   0.149 0.881804
## Latitude          0.1015    10.0525   0.010 0.991945
## `Total BLTs`      0.5839     0.1772   3.295 0.000983 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 100.118  on 190  degrees of freedom
## Residual deviance:  36.971  on 181  degrees of freedom
## AIC: 56.971
##
## Number of Fisher Scoring iterations: 17
```

Based on these summary statistics, we can see that there isn't any significant relationship between any of the variables and whether or not ticks with Lyme are found. The only relationship is with Total BLTs which makes sense as the more ticks found, the more likely there would be ticks with Lyme. If no ticks are found, it is impossible that ticks with Lyme would be found.

I believe that the lack of significant relationships has to do with weaknesses in the dataset such as a lack of data points and incomplete data. Despite it being not significant enough, we can see that based on the data, the chance of finding Ticks with Lyme disease bacteria is higher in more recent years compared to 2013 and 2014 especially. We can also see that longitude seems to have a slightly stronger relationship with # positive compared to latitude although both are not significant enough to say there is a relationship.

As a result of not finding a lot of significant relationships using this model, I decided to use some plots to examine the relationships further.

Figure 2: Scatterplot of Total number of BLTs versus Number of positive BLTs found in that location (with Line Fit)

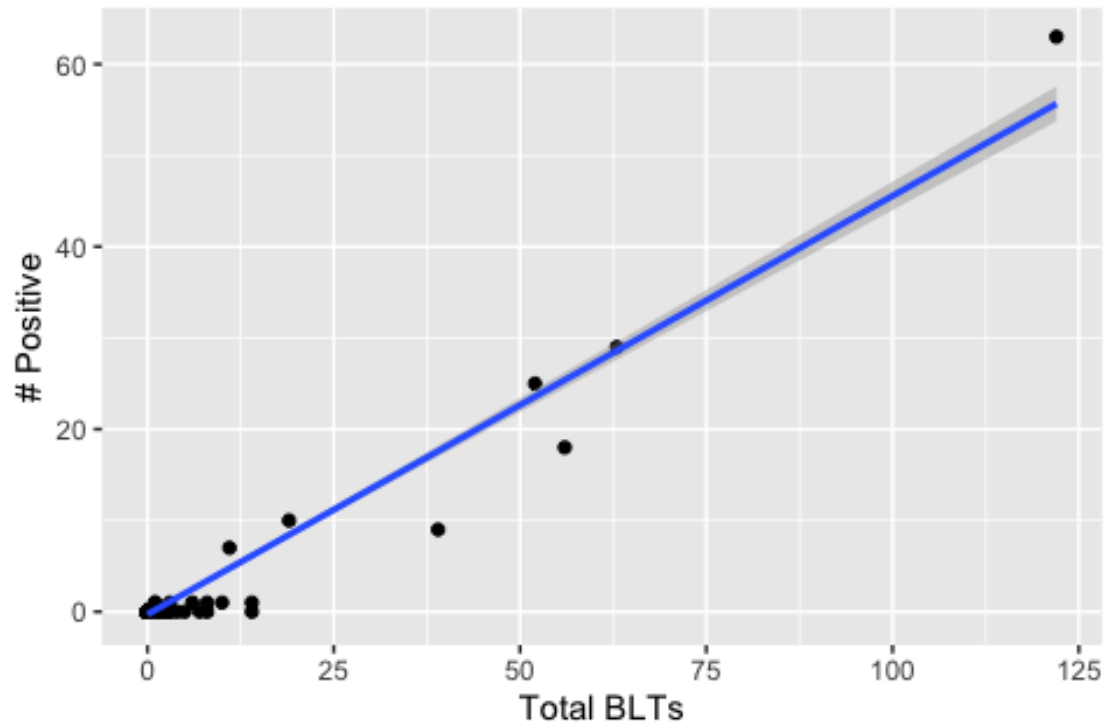


Figure 3: Scatterplot of Longitude value versus Number of positive BLTs found found in that location (with Line Fit)

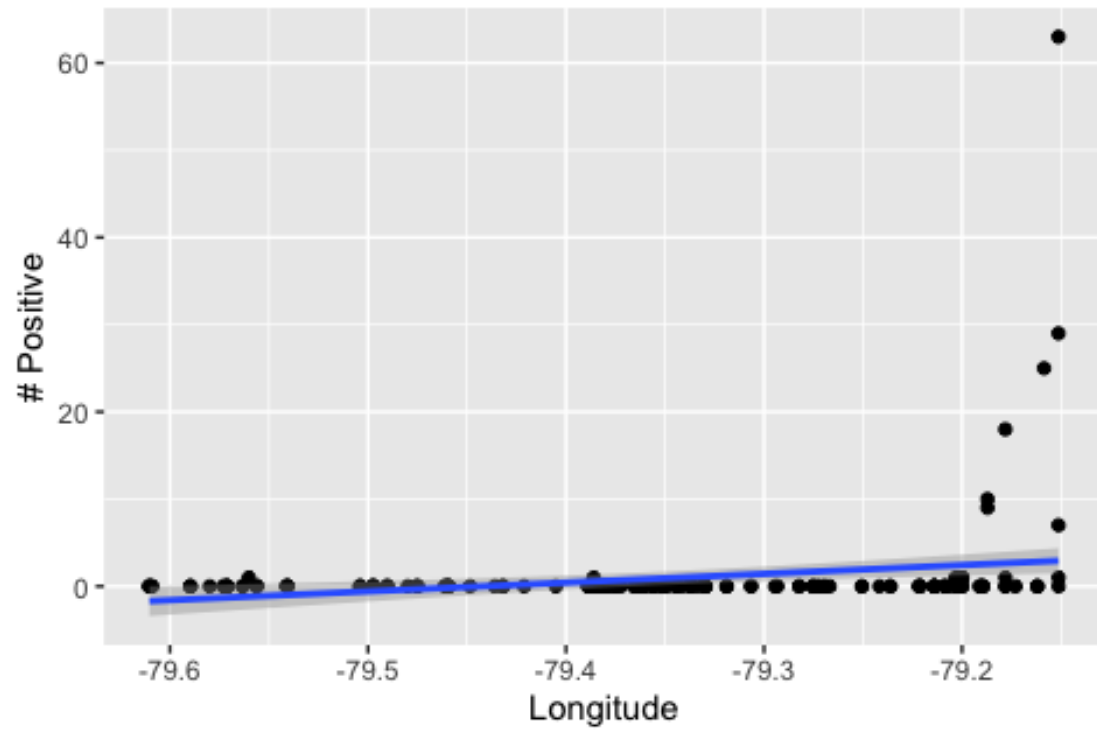


Figure 4: Scatterplot of Latitude value versus Number of positive BLTs found found in that location (with Line Fit)

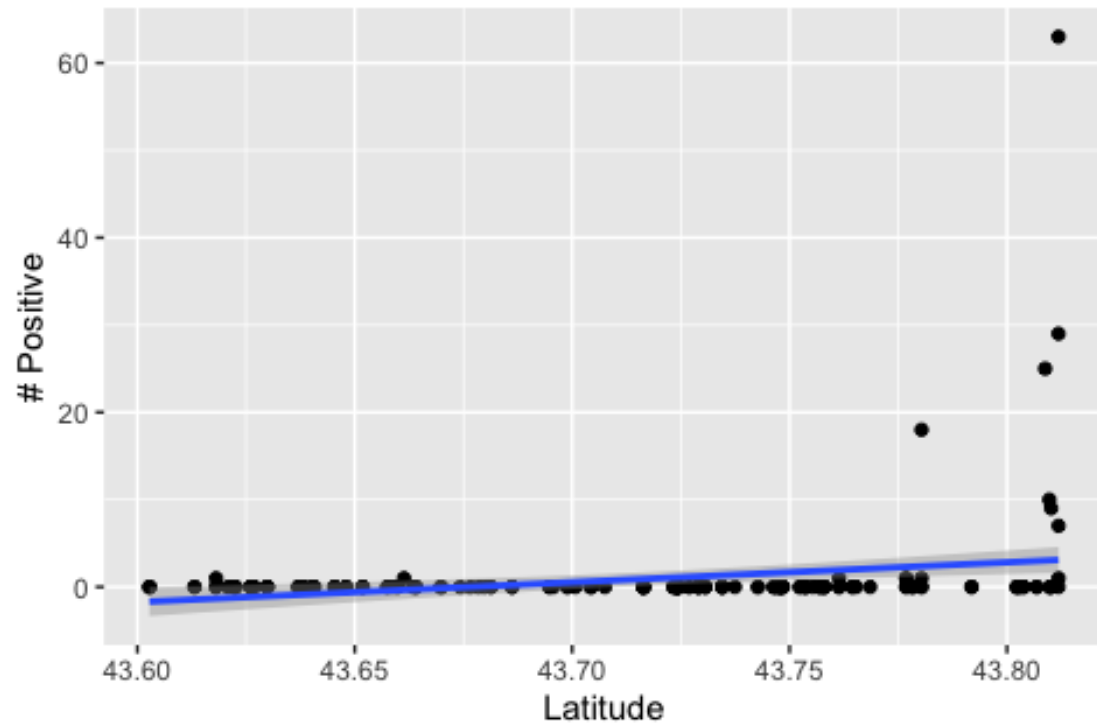
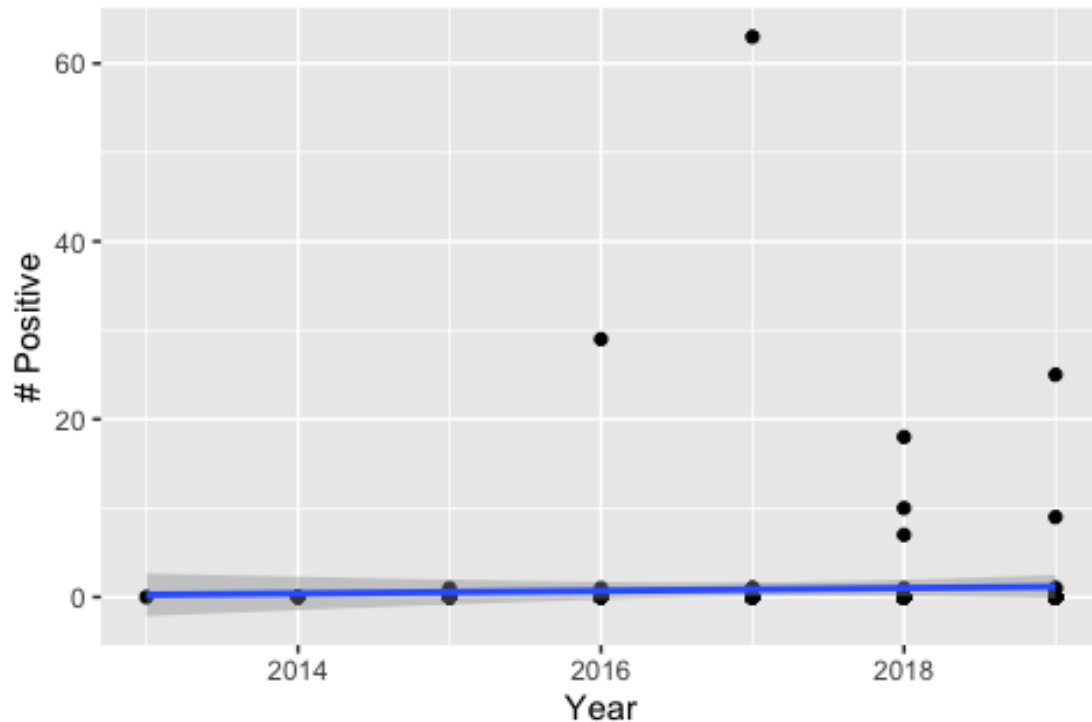




Figure 5: Scatterplot of Year versus Number of positive BLTs found found in that location (with Line Fit)



Based on the scatter plots above, you can see the relationship more clearly, it is now obvious that the relationship between total number of BLTs found and the number of positive cases of BLTs carrying the Lyme disease bacteria are related. The reasons for that are straightforward, generally the more BLTs you find, the more likely more of them would test positive.

The latitude vs # positive (figure 3) and longitude vs # positive (figure 4) look like they have a rather similar situation going on. For longitude, the more east you go in the city, the more likely you will find outlier locations where there are many more ticks found. Similarly for latitude, the more north you go, the more likely you find outlier locations where there are many more ticks. These two do not show an obvious relationship but it is interesting how the high outlier locations are found only near the northern and eastern areas of the city.

For Figure 5, finding a relationship between year and positive BLTs found, we find that in the later years, there are more cases of locations with very high number of ticks positive with lyme.

## Discussion

The general conclusion for this analysis is that the dataset does not offer a clear enough picture for any solid relationships between number of BLTs testing positive and any other useful variable included. The results of this analysis have given some clues to where a correlation or relationship might be but no solid evidence thus far.

Something interesting that was noticed was relating to the parks that BLTs were spotted at. Looking into this further, it seems that most or all of the outliers are from one or two parks (Rouge Park being one).

Here are a couple of tables, the first (Figure 6) with top 10 parks for total number of BLTs, the second (Figure 7) is top 10 parks for total number of BLTs with the Lyme disease causing bacteria.

(Figure 6) Top 10 Locations with  
Highest Number of BLTs Found

	Park Location	Total BLTs	# Positive
1	Rouge Park	122	63
2	Rouge Park	63	29
3	Colonel Danforth Trail	56	18
4	Rouge Park: Glen Eagle Vista Trail	52	25
5	Upper Rouge Trail Park	39	9
6	Upper Rouge Trail Park	19	10
7	Morningside Park	14	0
8	Morningside Park	14	1
9	Rouge Park	11	7
10	Rouge Park	10	1

## (Figure 7) Top 10 Locations with Highest Number of Positive BLTs

	Park Location	Total BLTs	# Positive
1	Rouge Park	122	63
2	Rouge Park	63	29
3	Rouge Park: Glen Eagle Vista Trail	52	25
4	Colonel Danforth Trail	56	18
5	Upper Rouge Trail Park	19	10
6	Upper Rouge Trail Park	39	9
7	Rouge Park	11	7
8	Rouge Park	10	1
9	Colonel Danforth Trail	8	1
10	Highland Creek Park	3	1

As we can see on Figure 6, the locations that are in Rouge Park appear for 7 of the top 10 locations in the table. Another park called Morningside Park shows up twice along with the Colonel Danforth Trail which shows up once. All of these locations seen in Figure 6 are at the east edge of the city so it makes sense that there were extreme outliers with higher longitude values.

For Figure 7, two locations stand out that were also in Figure 6, Rouge Park and Colonel Danforth Trail. What is interesting though is that Morningside Park does not appear at all and a new park, Highland Creed Park appears. Knowing this, we can see that Morningside Park has a rather low infection per tick ratio.

Examining the data in Figure 6 even further, it is found that there are a total of 485 ticks in the entire dataset and 400 are located in the top 10 locations combined. In addition, there were only 4 parks in Figure 6 which means one can drastically reduce their chances of encountering ticks if they simply avoid the 4 parks in Toronto: Colonel Danforth Trail, Morningside Park, Highland Creek Park and especially Rouge Park.

Most importantly though, with 485 ticks in total, 168 ticks out of the 485 in total were confirmed to be positive of the Lyme disease bacteria. By examining Figure 7, a more interesting finding is revealed. 163 out of the 168 ticks with Lyme disease show up in the table (Figure 7). This is 97% of the infected tick sightings. The most surprising thing is that there are only 3 parks included in this table, it can therefore be concluded by this dataset

that one can reduce one's chance of getting Lyme disease by 97% if they just avoid these 3 parks in Toronto: Rouge Park, Colonel Danforth Trail and Highland Creek Park. This is very interesting information because if people follow this guideline, the risk of lyme disease will drastically lower, given that this dataset is accurate to a certain degree.

## Weaknesses and Next Steps

As we have touched upon in the Data section, there are several weaknesses with the sampling approach which result in a less robust dataset. Tick dragging does not work well in locations where the number of ticks are low and thus many locations with 0 BLTs might actually have some, just in smaller numbers. If better methods were used for Tick collection, this dataset will be more robust and will result in more accurate findings from analysis. Another weakness is the limitations due to the small sample size.

Some next steps would be to find more datasets containing information on Black Legged Ticks (BLTs) and perhaps one that takes into account a larger land area, like looking at the tick sightings in North America based on latitude and longitude. Several other variables that should be included are related to deer population density and climate change. BLTs like to use deers as hosts and therefore are also found in higher numbers where deer populations are more dense. In addition BLTs also like hot and humid summer weather and cases of BLTs have been increasing according to the CDC in the past few decades which coincides with global temperature rise. All of these variables should be studied along with tick population growth and the relationship with changing ecosystems.

## Sources

“Open Data Dataset.” City of Toronto Open Data Portal, Toronto Public Health, 14 Feb. 2020, [open.toronto.ca/dataset/blacklegged-tick-surveillance/](https://open.toronto.ca/dataset/blacklegged-tick-surveillance/).

“Tick Surveillance.” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 27 May 2020, [www.cdc.gov/ticks/surveillance/index.html](https://www.cdc.gov/ticks/surveillance/index.html).

“Lyme and Other Tickborne Diseases Increasing.” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 22 Apr. 2019, [www.cdc.gov/media/dpk/diseases-and-conditions/lyme-disease/index.html](https://www.cdc.gov/media/dpk/diseases-and-conditions/lyme-disease/index.html).

Giorgi, A. (2018, September 17). Chronic (Persistent) Lyme Disease: Symptoms and Diagnosis. Retrieved December 23, 2020, from <https://www.healthline.com/health/lyme-disease-chronic-persistent>

“Grid v3.6.2.” Grid Package | R Documentation, R-Core [R-Core@R-Project.org](https://R-Core@R-Project.org), [www.rdocumentation.org/packages/grid/versions/3.6.2](https://www.rdocumentation.org/packages/grid/versions/3.6.2).

Rising concern of tick-borne diseases in Canada. (n.d.). Retrieved December 23, 2020, from <https://ncceh.ca/content/blog/rising-concern-tick-borne-diseases-canada>

“Tidyverse Packages.” Tidyverse, [www.tidyverse.org/packages/](https://www.tidyverse.org/packages/).