Final Data Analysis and Report

Trent Meyer

2022-12-14

Introduction

Our data comes from the New York State Department of Health public database, which was completed over fourteen years from 2008 to 2021. They collected samples at sites throughout the 62 counties in New York state, recording the total ticks collected, tick population density, along with percentages of those ticks collected who were found to have specific bacteria or parasites.

With data over many years, it would be interesting to look at if there is a difference in the Tick.Population.Density between the beginning year (2008) and the end year (2021). Independence between the two population densities would be more reasonable since the years are not closer together. Richard E.L. Paul, et. al., explain how their experiment yielded no changes in adult tick populations over time (2016). This would lead a look into the year and density to see if there is actually a difference or not.

One idea to look into is that certain regions of New York State that may be more urbanized than others to see if the Tick.Population.Density differs compared to more rural regions. Paul (2016) describes that ticks' natural habitat are often forests where they come into contact with their hosts. Regions such as the North Country or Southern Tier, which are more rural compared to say NYC, have larger areas of forests which could mean higher tick population density.

With the urbanization of many areas, ticks often find their way into suburban areas through nearby forests which house the ticks being used by domesticated pets and children (Paul 2016). This would lead a discussion about whether region and year are affecting the tick population density for the years 2008 and 2021. If areas are becoming more urbanized over time, that means that more ticks are possibly finding their way into cities.

Many individuals in rural areas are aware of ticks, how to get rid of them, and where to find them. However, many individuals in cities are not as aware of ticks, and may not realize that ticks can carry bacteria or parasites that can cause harmful diseases. Kowalec et. al., explain that species of bacteria or parasites can "differ between natural and urban areas"(2017). With this in mind, we can expect a possible difference between these bacteria and parasites while looking at years. Taking into account the urbanization of many areas, it could be possible to see more ticks with specific bacteria or parasites being found in more areas than say a decade ago.

Analysis

Let's clear the environment, load in packages, and read in the tick data:

```
rm(list = ls())
library(tidyverse)
library(here)
library(ggfortify)
library(lubridate)
tick_adult <- read.csv(here("Data", "Deer_Tick_Surveillance__Adults__Oct_to_Dec__excluding_Powassan_viru
```

Let's take a look at the data!

```
head(tick_adult)
```

```
##   Year     County Total.Sites.Visited Total.Ticks.Collected
## 1 2021     Albany                   4                   366
## 2 2021 Chautauqua                   3                   199
## 3 2021   Dutchess                   5                   438
## 4 2021     Warren                   4                    88
## 5 2020     Albany                   4                   114
## 6 2020    Chemung                   1                   271
##   Tick.Population.Density Total.Tested B..burgdorferi....
## 1                   85.83          168               53.6
## 2                   36.70          107               56.1
## 3                   66.80           51               47.1
## 4                   13.85           87               41.4
## 5                   37.15          103               71.8
## 6                  256.90           50               56.0
##   A..phagocytophilum.... B..microti.... B..miyamotoi....
## 1                   12.5            9.5              2.4
## 2                    0.0            0.0              0.0
## 3                   11.8           11.8              5.9
## 4                    6.9            9.2              2.3
## 5                   13.6            8.7              0.0
## 6                   16.0            0.0              2.0
##             County.Centroid
## 1 (42.5882713, -73.9740136)
## 2 (42.3042159, -79.4075949)
## 3 (41.7550085, -73.7399512)
## 4 (43.5551053, -73.8381388)
## 5 (42.5882713, -73.9740136)
## 6 (42.1552807, -76.7471788)
```

```
glimpse(tick_adult)
```

```
## Rows: 547
## Columns: 11
## $ Year                    <int> 2021, 2021, 2021, 2021, 2020, 2020, 2020, 2020~
## $ County                  <chr> "Albany", "Chautauqua", "Dutchess", "Warren", ~
## $ Total.Sites.Visited     <int> 4, 3, 5, 4, 4, 1, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ Total.Ticks.Collected   <int> 366, 199, 438, 88, 114, 271, 75, 131, 157, 51,~
## $ Tick.Population.Density <dbl> 85.83, 36.70, 66.80, 13.85, 37.15, 256.90, 21.~
## $ Total.Tested            <int> 168, 107, 51, 87, 103, 50, 69, 44, 50, 50, 50,~
## $ B..burgdorferi....      <dbl> 53.6, 56.1, 47.1, 41.4, 71.8, 56.0, 68.1, 59.1~
## $ A..phagocytophilum....  <dbl> 12.5, 0.0, 11.8, 6.9, 13.6, 16.0, 10.1, 36.4, ~
## $ B..microti....          <dbl> 9.5, 0.0, 11.8, 9.2, 8.7, 0.0, 0.0, 0.0, 0.0, ~
## $ B..miyamotoi....        <dbl> 2.4, 0.0, 5.9, 2.3, 0.0, 2.0, 1.4, 2.3, 0.0, 0~
## $ County.Centroid         <chr> "(42.5882713, -73.9740136)", "(42.3042159, -79~
```

```
str(tick_adult)
```

```
## 'data.frame':    547 obs. of  11 variables:
```

```
##  $ Year                   : int  2021 2021 2021 2021 2020 2020 2020 2020 2020 2020 ...
##  $ County                 : chr  "Albany" "Chautauqua" "Dutchess" "Warren" ...
##  $ Total.Sites.Visited    : int  4 3 5 4 4 1 3 1 1 1 ...
##  $ Total.Ticks.Collected  : int  366 199 438 88 114 271 75 131 157 51 ...
##  $ Tick.Population.Density: num  85.8 36.7 66.8 13.8 37.1 ...
##  $ Total.Tested           : int  168 107 51 87 103 50 69 44 50 50 ...
##  $ B..burgdorferi....     : num  53.6 56.1 47.1 41.4 71.8 56 68.1 59.1 48 46 ...
##  $ A..phagocytophilum.... : num  12.5 0 11.8 6.9 13.6 16 10.1 36.4 0 6 ...
##  $ B..microti....         : num  9.5 0 11.8 9.2 8.7 0 0 0 0 0 ...
##  $ B..miyamotoi....       : num  2.4 0 5.9 2.3 0 2 1.4 2.3 0 0 ...
##  $ County.Centroid        : chr  "(42.5882713, -73.9740136)" "(42.3042159, -79.4075949)" "(41.7550085
```

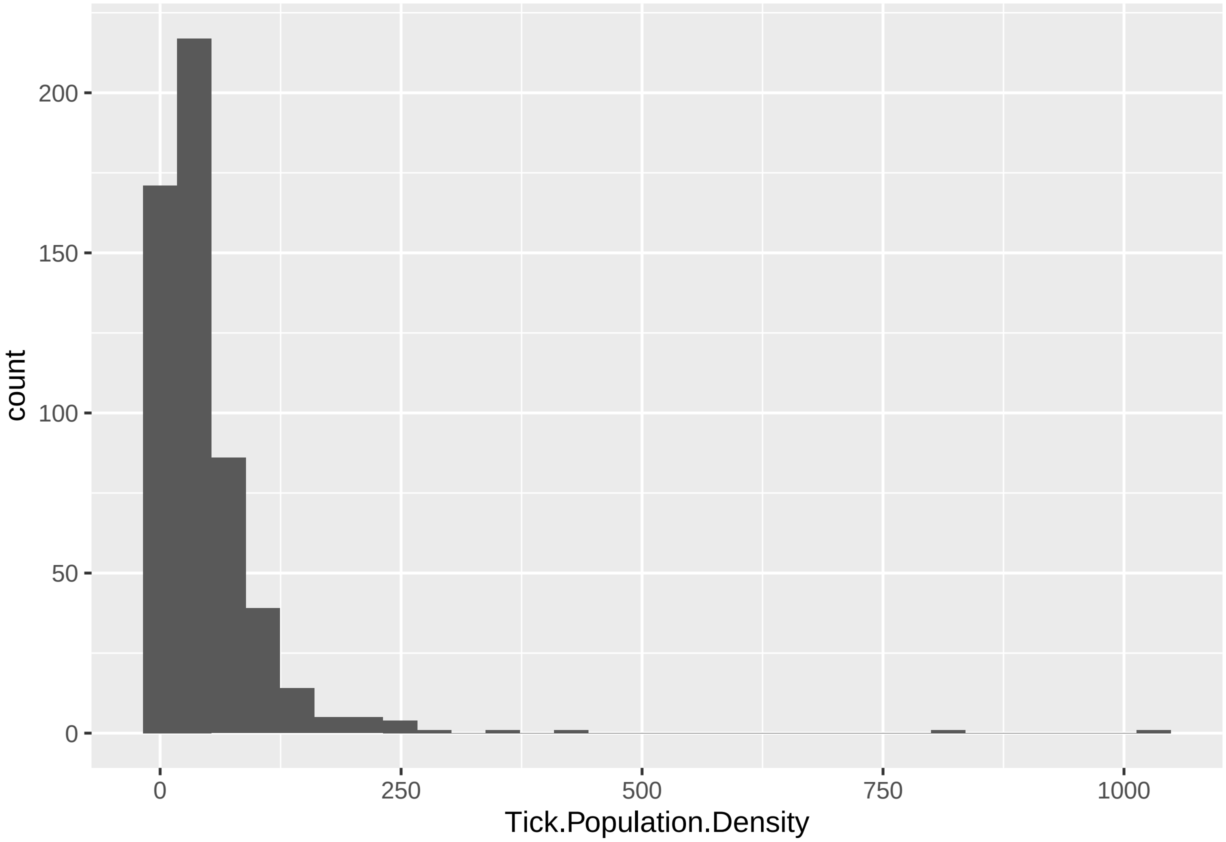
Let's rename the parasite and bacteria variables, and also read in County and Year as factors instead of characters:

```
tick_adult <- tick_adult %>%
  rename(B.burgdorferi = "B..burgdorferi....",
         A.phagocytophilum = "A..phagocytophilum....",
         B.microti = "B..microti....",
         B.miyamotoi = "B..miyamotoi....") %>%
  mutate(County = as.factor(County),
         Year = as.factor(Year))
```

Tick population density makes sense as a response variable, rather than total tick population. The number of total sites they collected from is different for each observation, meaning that total tick population would vary based on the number of sites. However, tick population density would give us more information as to how densely populated the ticks are in that specific county.

Let's graph tick population density to see what the distribution looks like.

```
ggplot(data = tick_adult, aes(x = Tick.Population.Density)) +
  geom_histogram()
```

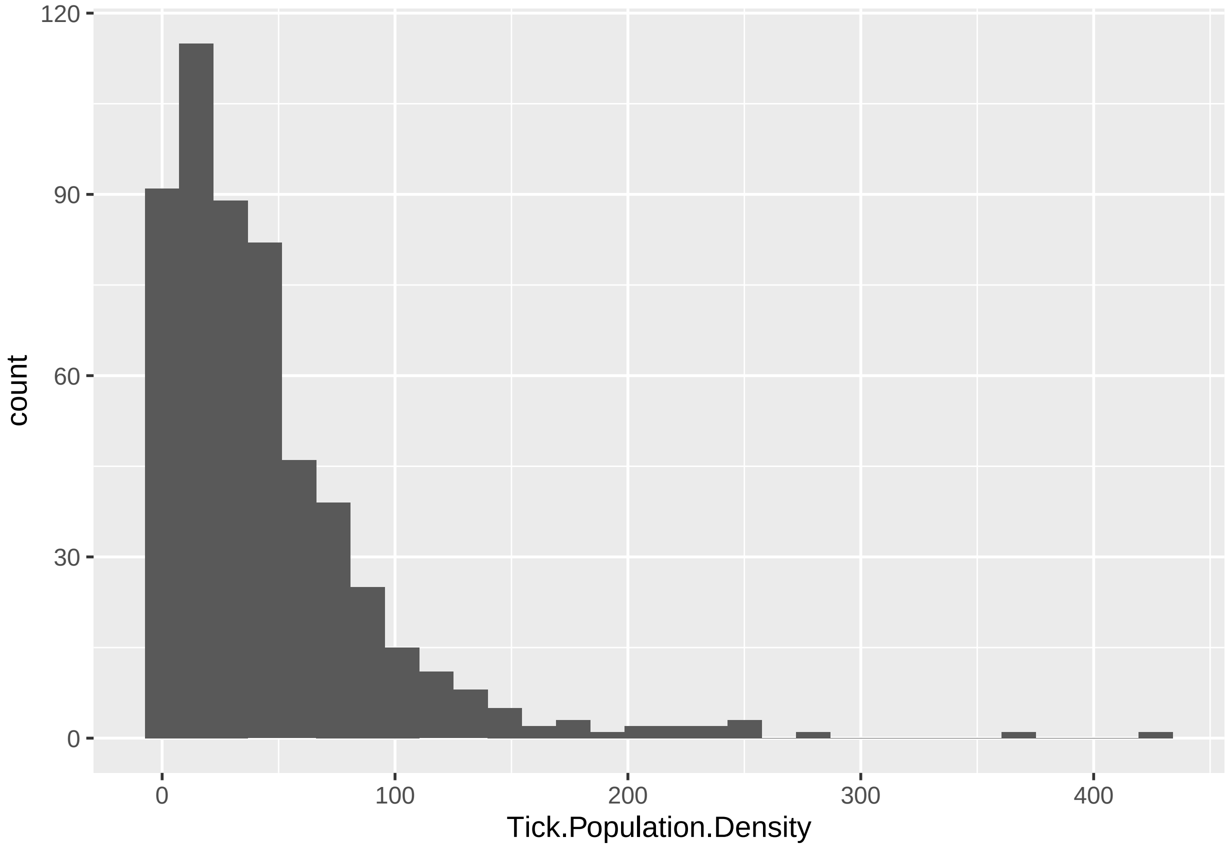


We see that there are two that have a Tick.Population.Density > 750. I am not sure if this is a mistake, or if they actually found this high of a tick population density. Let's remove these observations from the data.

```
tick_adult <- tick_adult %>% filter(Tick.Population.Density < 750)
```

Let's replot the distribution of Tick.Population.Density:

```
ggplot(data = tick_adult, aes(x = Tick.Population.Density)) +
  geom_histogram()
```



We can see that the data is right-skewed.

```
summary(tick_adult)
```

```
##       Year          County    Total.Sites.Visited Total.Ticks.Collected
##  2014   : 52   Albany  : 14   Min.   : 1.000      Min.   :    0.0
##  2015   : 52   Columbia: 14   1st Qu.: 1.000      1st Qu.:   54.0
##  2020   : 50   Dutchess: 14   Median : 2.000      Median :  101.0
##  2016   : 49   Monroe  : 14   Mean   : 2.494      Mean   :  169.6
##  2017   : 49   Orange  : 14   3rd Qu.: 3.000      3rd Qu.:  158.0
##  2021   : 45   Oswego  : 14   Max.   :28.000      Max.   :11260.0
##  (Other):247   (Other) :460
##  Tick.Population.Density  Total.Tested    B.burgdorferi    A.phagocytophilum
##  Min.   :  0.00          Min.   :  0.00   Min.   :  0.00   Min.   : 0.000
##  1st Qu.: 14.09          1st Qu.: 50.00   1st Qu.: 37.50   1st Qu.: 0.000
##  Median : 32.45          Median : 51.00   Median : 50.30   Median : 3.800
##  Mean   : 45.16          Mean   : 75.02   Mean   : 46.75   Mean   : 6.295
##  3rd Qu.: 60.00          3rd Qu.: 98.00   3rd Qu.: 58.90   3rd Qu.:11.600
##  Max.   :426.70          Max.   :500.00   Max.   :100.00   Max.   :44.100
##                                           NA’s   :27       NA’s   :27
##    B.microti       B.miyamotoi     County.Centroid
##  Min.   : 0.000   Min.   : 0.000   Length:544
##  1st Qu.: 0.000   1st Qu.: 0.000   Class :character
##  Median : 0.000   Median : 0.500   Mode  :character
##  Mean   : 2.912   Mean   : 1.168
##  3rd Qu.: 4.000   3rd Qu.: 2.000
##  Max.   :32.000   Max.   :10.000
##  NA's   :27       NA's   :218
```

We will use tick population density as the response for most of the statistical tests.

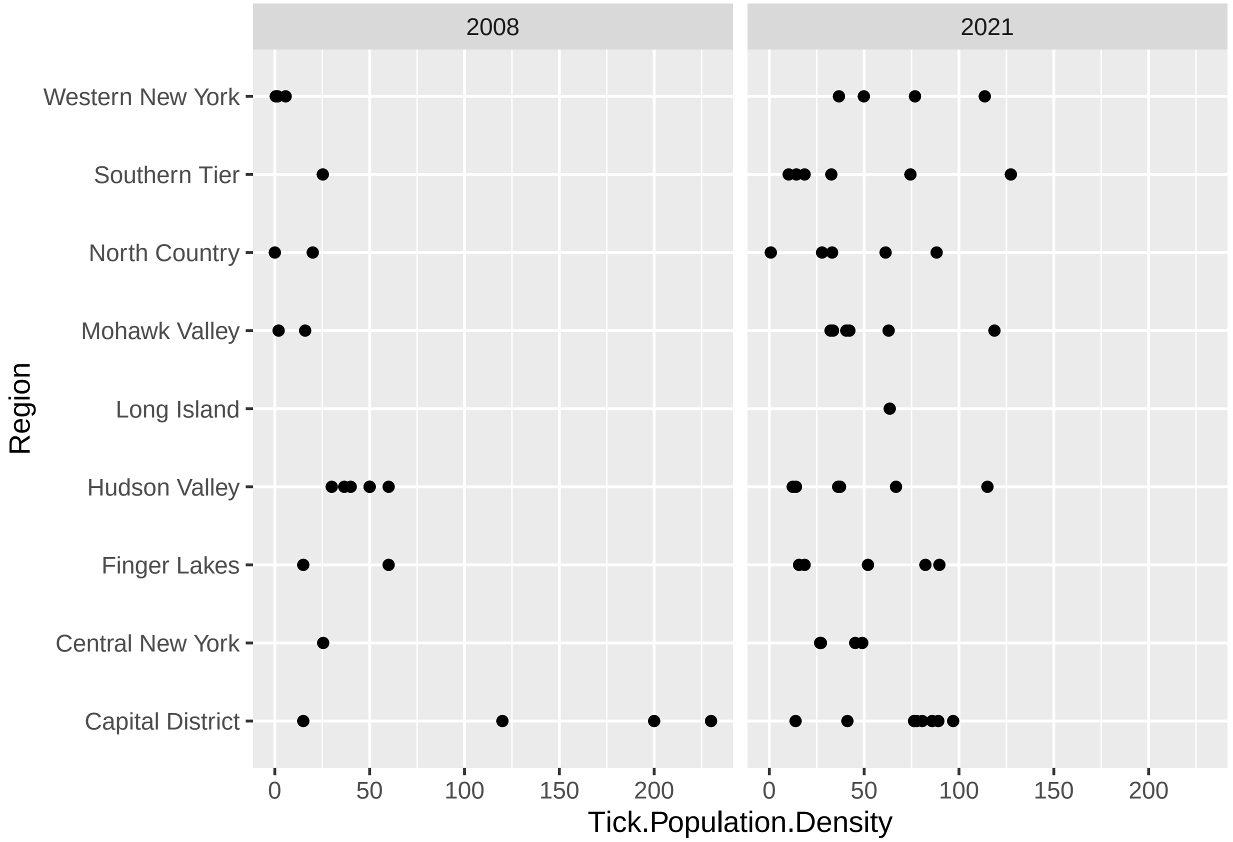
Is there a difference in the Tick.Population.Density between the years 2008 and 2021 and Region?

```
tick_adult_small <- tick_adult %>% filter(Year == "2008" | Year == "2021") %>% mutate(Region = case_whe
  County == "Suffolk" | County == "Nassau" ~ "Long Island",

  County == "Brooklyn" | County == "Bronx" | County == "Manhattan" |
    County == "Staten Island" | County == "Queens" ~ "New York City",

  County == "Dutchess" | County == "Orange" | County == "Putnam" |
    County == "Rockland" | County == "Sullivan" | County == "Ulster" |
    County == "Westchester" ~ "Hudson Valley",

  County == "Niagara" | County == "Erie" | County == "Chautauqua" |
    County == "Cattaragus" | County == "Cattaraugus" |
    County == "Allegany" ~ "Western New York",

  County == "Orleans" | County == "Genesee" | County == "Wyoming" |
    County == "Monroe" | County == "Livingston" | County == "Wayne" |
    County == "Ontario" | County == "Yates" | County == "Seneca" ~ "Finger Lakes",

  County == "Steuben" | County == "Schuyler" | County == "Chemung" |
    County == "Tompkins" | County == "Tioga" | County == "Chenango"
  | County == "Broome" | County == "Delaware" ~ "Southern Tier",

  County == "Cortland" | County == "Cayuga" | County == "Onondaga" |
    County == "Oswego" | County == "Madison" ~ "Central New York",

  County == "St. Lawrence" | County == "St Lawrence" | County == "Lewis" |
    County == "Jefferson" | County == "Hamilton" | County == "Essex" |
    County == "Clinton" | County == "Franklin" ~ "North Country",

  County == "Oneida" | County == "Herkimer" | County == "Fulton" |
    County == "Montgomery" | County == "Otsego" |
    County == "Schoharie" ~ "Mohawk Valley",

  County == "Albany" | County == "Columbia" | County == "Greene" |
    County == "Warren" | County == "Washington" | County == "Saratoga" |
    County == "Schenectady" | County == "Rensselaer" ~ "Capital District"
))

tick_adult_total <- tick_adult_small %>% group_by(Region, Year) %>% summarise(total_density = sum(Tick.

tick_adult_total
```

```
## # A tibble: 17 x 3
## # Groups:   Region [9]
##    Region           Year  total_density
##    <chr>            <fct>         <dbl>
##  1 Capital District 2008          565
##  2 Capital District 2021          562.
##  3 Central New York 2008           25.5
##  4 Central New York 2021          148.
##  5 Finger Lakes     2008           75
##  6 Finger Lakes     2021          258.
##  7 Hudson Valley    2008          267.
##  8 Hudson Valley    2021          282.
##  9 Long Island      2021           63.5
## 10 Mohawk Valley    2008           18
## 11 Mohawk Valley    2021          330.
## 12 North Country    2008           20
## 13 North Country    2021          211.
## 14 Southern Tier    2008           25.3
## 15 Southern Tier    2021          278.
## 16 Western New York 2008            7.7
## 17 Western New York 2021          277.
```

We see quite a few increases in some regions such as CNY, Finger Lakes, and Mohawk Valley from 2008 to 2021. Let's try to plot this and see if there is a difference.

```
ggplot(data = tick_adult_small, aes(x = Region, y = Tick.Population.Density)) +
  geom_point() +
  coord_flip() +
  facet_wrap(~Year)
```

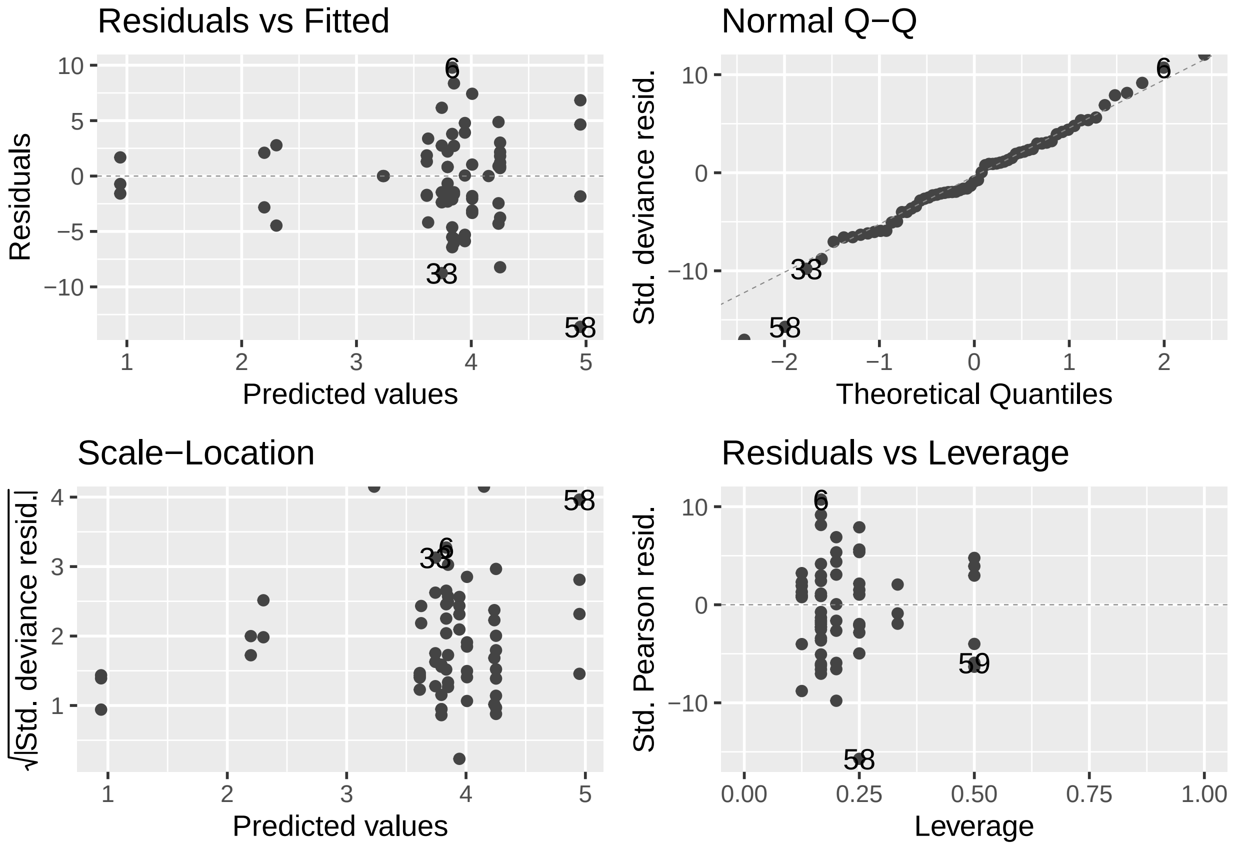


It looks like the data in 2021 is spread out a bit more, with more of the points being at around 50. 2008 has many points under 50, however Capital District is the only one higher than 100. This leads me to think that for regions besides Capital District, we might be able to see an interactions between year and region. I am going to try to fit a generalized linear model and predict the Tick.Population.Density collected given year, region, and an interaction between the two.

```
full_mod <- glm(Tick.Population.Density ~ Region + Year + Region*Year, data = tick_adult_small, family
summary(full_mod)
```

```
##
## Call:
## glm(formula = Tick.Population.Density ~ Region + Year + Region *
##     Year, family = "poisson", data = tick_adult_small)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -13.6098   -2.7329   -0.3383    2.1489    9.7861
##
## Coefficients: (1 not defined because of singularities)
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       4.95053    0.04207 117.673  < 2e-16 ***
## RegionCentral New York           -1.71185    0.20245  -8.456  < 2e-16 ***
## RegionFinger Lakes               -1.32619    0.12290 -10.791  < 2e-16 ***
## RegionHudson Valley              -1.15654    0.07430 -15.565  < 2e-16 ***
## RegionLong Island                -0.10039    0.13241  -0.758    0.448
## RegionMohawk Valley              -2.75331    0.23943 -11.500  < 2e-16 ***
## RegionNorth Country            -2.64795    0.22752 -11.638  < 2e-16 ***
## RegionSouthern Tier            -1.71973    0.20321  -8.463  < 2e-16 ***
## RegionWestern New York         -4.00792    0.36282 -11.047  < 2e-16 ***
## Year2021                       -0.69925    0.05959 -11.735  < 2e-16 ***
## RegionCentral New York:Year2021  1.07345    0.22251   4.824 1.40e-06 ***
## RegionFinger Lakes:Year2021     1.01948    0.14407   7.076 1.48e-12 ***
## RegionHudson Valley:Year2021    0.75463    0.10417   7.245 4.34e-13 ***
## RegionLong Island:Year2021           NA         NA      NA       NA
## RegionMohawk Valley:Year2021    2.51021    0.24927  10.070  < 2e-16 ***
## RegionNorth Country:Year2021    2.13994    0.24141   8.864  < 2e-16 ***
## RegionSouthern Tier:Year2021    1.30258    0.21605   6.029 1.65e-09 ***
## RegionWestern New York:Year2021  3.99415    0.37018  10.790  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2212.1  on 65  degrees of freedom
## Residual deviance: 1180.6  on 49  degrees of freedom
## AIC: Inf
##
## Number of Fisher Scoring iterations: 5
```

```
autoplot(full_mod, smooth.colour = NA)
```



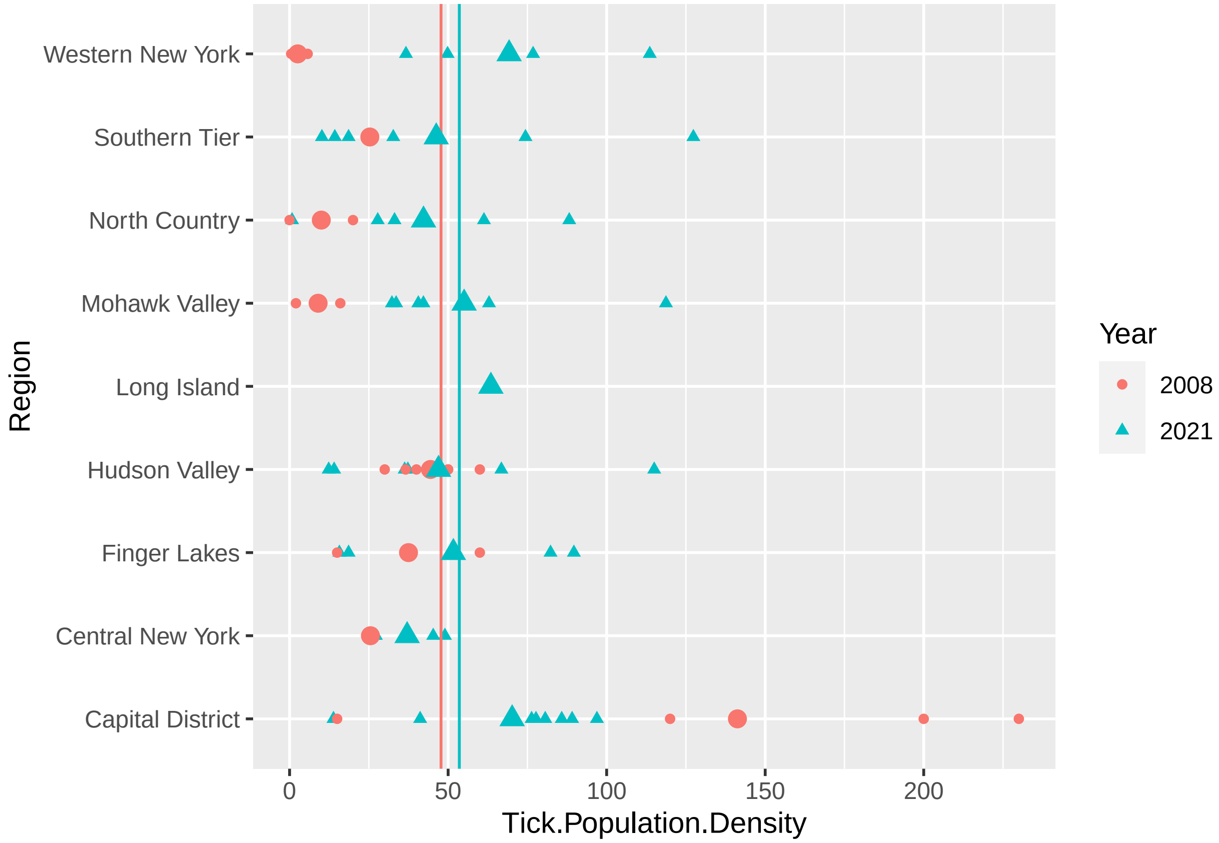
I added “family =”poisson” ” which helped with the normality assumption, as before it was not normally distributed. THe other assumptions are okay, not great however.

We can see that all of the interactions between Year and Region are significant, except Long Island, all with very small p-values.

Let's replot:

```
tick_adult_08 <- tick_adult_small %>% filter(Year == "2008")

tick_adult_08_mean <- tick_adult_08 %>% group_by(Year) %>% summarise(mean_density = mean(Tick.Populatio

tick_adult_08_region_mean <- tick_adult_08 %>% group_by(Year, Region) %>% summarise(mean_density = mean

tick_adult_21 <- tick_adult_small %>% filter(Year == "2021")

tick_adult_21_mean <- tick_adult_21 %>% group_by(Year) %>% summarise(mean_density = mean(Tick.Populatio

tick_adult_21_region_mean <- tick_adult_21 %>% group_by(Year, Region) %>% summarise(mean_density = mean
```

```
ggplot(data = tick_adult_small, aes(x = Tick.Population.Density, y = Region)) +
  geom_point(aes(shape = Year, colour = Year)) +
  geom_vline(data = tick_adult_08_mean, aes(xintercept = mean_density), colour = "#F8766D") +
  geom_vline(data = tick_adult_21_mean, aes(xintercept = mean_density), colour = "#00BFC4") +
  geom_point(data = tick_adult_08_region_mean, aes(x = mean_density, y = Region), size = 3, shape = 16,
  geom_point(data = tick_adult_21_region_mean, aes(x = mean_density, y = Region), size = 3, shape = 17,
```



We can see that the North Country and Mohawk Valley are very similar, with lots of overlap.

We will also use this model to see if there is a difference between: - the years 2008 and 2021 - region alone

Is there significant difference in Tick.Population.Density between the first year (2008), and last year (2021)?

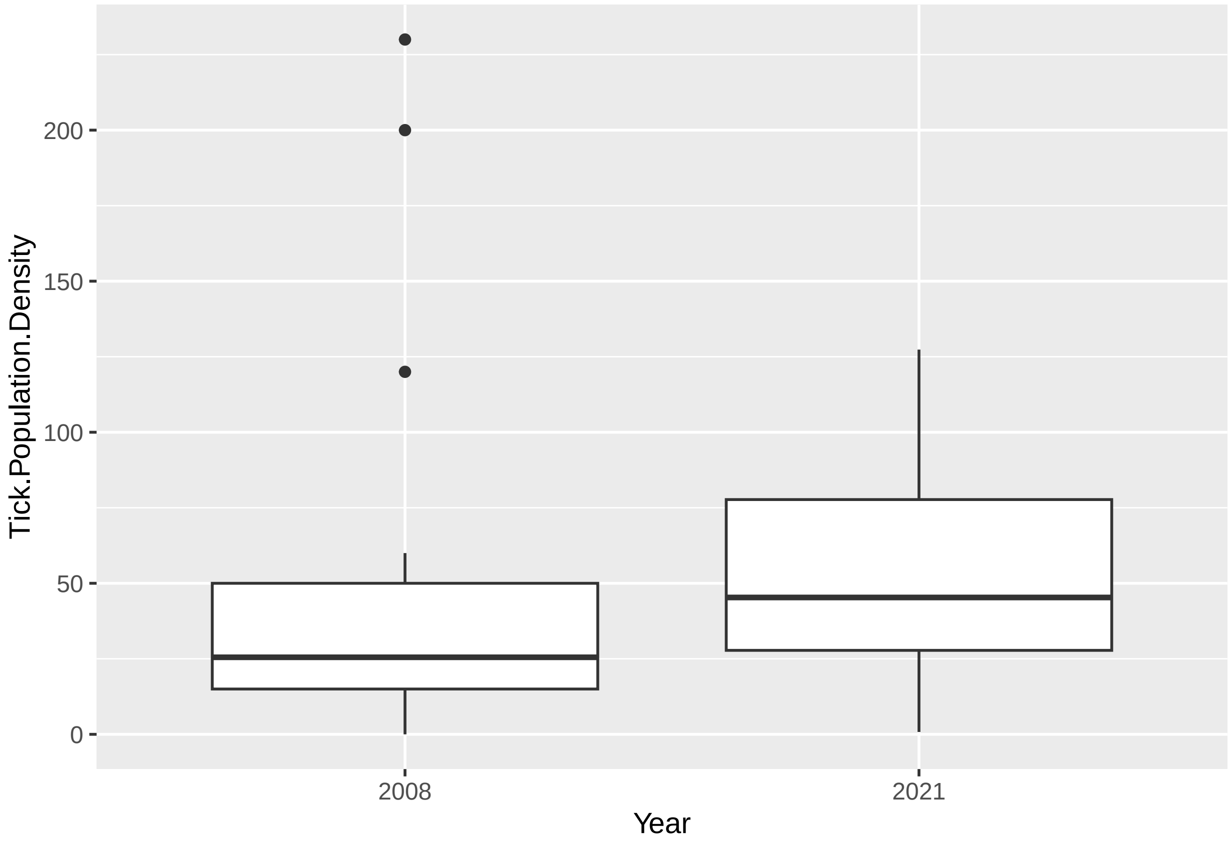
I already subsetted the data to only have the years 2008 and 2021. Let's see what the means are:

```
tick_adult_small %>% group_by(Year) %>% summarise(mean_density = mean(Tick.Population.Density, na.rm = T
```

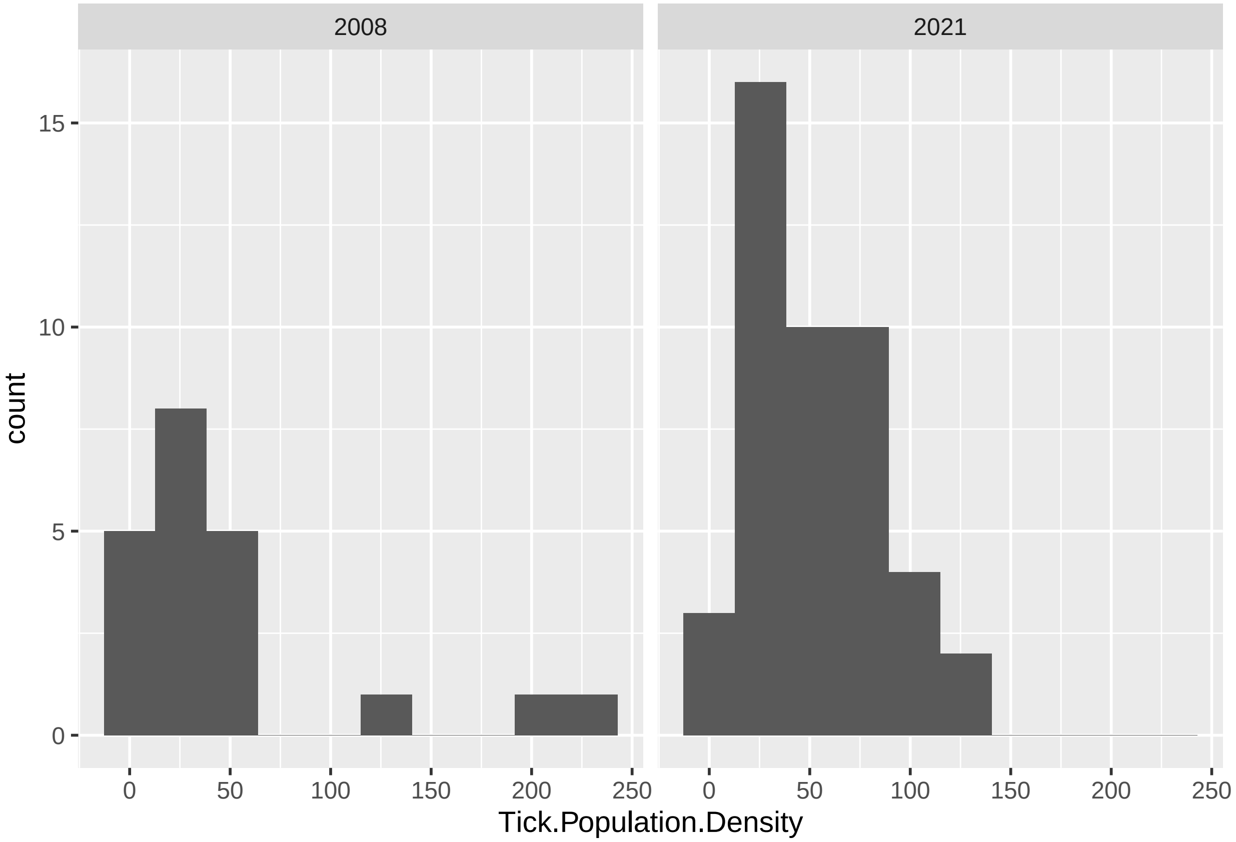
```
## # A tibble: 2 x 2
##   Year  mean_density
##   <fct>        <dbl>
## 1 2008          47.8
## 2 2021          53.5
```

We can see that 2021 is a bit higher than 2008! I now want to see if there is a significant difference between the two years.

```
ggplot(data = tick_adult_small, aes(x = Year, y = Tick.Population.Density)) +
  geom_boxplot()
```



```
ggplot(data = tick_adult_small, aes(x = Tick.Population.Density)) +
  geom_histogram(bins = 10) +
  facet_wrap(~Year)
```



There is a bit of overlap in the spread of each year, however it looks like the mean is higher for 2021 compared to 2008. These plots are much easier to read with the few observations out that were making it difficult to read and compare between years. The histograms show that the peaks are in slightly different location on the x-axis.

I will be using the two-sample t test that was included in my full_mod, with the following hypotheses:

–Null Hypothesis: There is no true difference in mean Tick.Population.Density between 2008 and 2021.

–Alternative Hypothesis: There is a true difference in mean Tick.Population.Density between 2008 and 2021.

```
summary(full_mod)
```

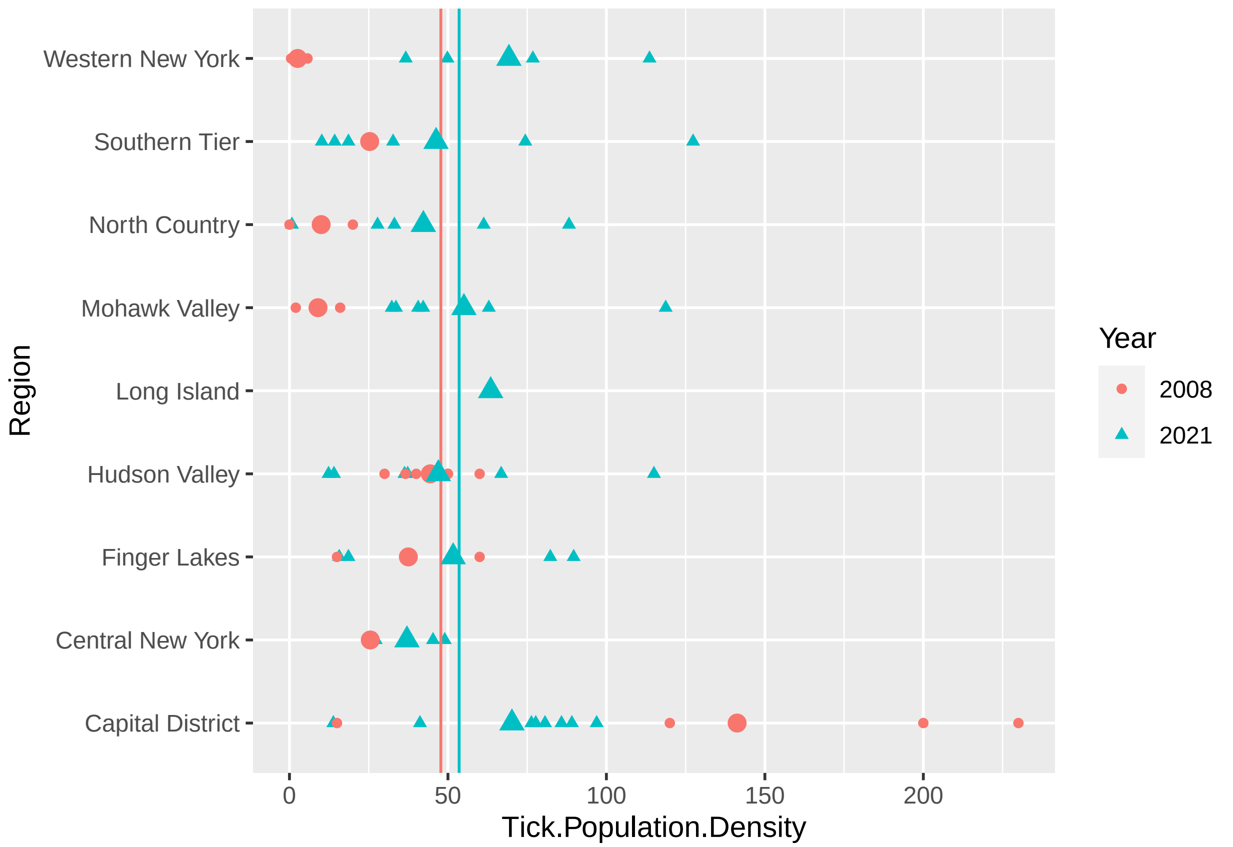
```
##
## Call:
## glm(formula = Tick.Population.Density ~ Region + Year + Region *
##     Year, family = "poisson", data = tick_adult_small)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -13.6098   -2.7329   -0.3383    2.1489    9.7861
##
## Coefficients: (1 not defined because of singularities)
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      4.95053    0.04207 117.673  < 2e-16 ***
## RegionCentral New York          -1.71185    0.20245  -8.456  < 2e-16 ***
## RegionFinger Lakes              -1.32619    0.12290 -10.791  < 2e-16 ***
## RegionHudson Valley             -1.15654    0.07430 -15.565  < 2e-16 ***
## RegionLong Island               -0.10039    0.13241  -0.758    0.448    
## RegionMohawk Valley             -2.75331    0.23943 -11.500  < 2e-16 ***
## RegionNorth Country             -2.64795    0.22752 -11.638  < 2e-16 ***
## RegionSouthern Tier             -1.71973    0.20321  -8.463  < 2e-16 ***
## RegionWestern New York          -4.00792    0.36282 -11.047  < 2e-16 ***
## Year2021                        -0.69925    0.05959 -11.735  < 2e-16 ***
## RegionCentral New York:Year2021  1.07345    0.22251   4.824 1.40e-06 ***
## RegionFinger Lakes:Year2021      1.01948    0.14407   7.076 1.48e-12 ***
## RegionHudson Valley:Year2021     0.75463    0.10417   7.245 4.34e-13 ***
## RegionLong Island:Year2021            NA         NA      NA       NA    
## RegionMohawk Valley:Year2021     2.51021    0.24927  10.070  < 2e-16 ***
## RegionNorth Country:Year2021     2.13994    0.24141   8.864  < 2e-16 ***
## RegionSouthern Tier:Year2021     1.30258    0.21605   6.029 1.65e-09 ***
## RegionWestern New York:Year2021  3.99415    0.37018  10.790  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2212.1  on 65  degrees of freedom
## Residual deviance: 1180.6  on 49  degrees of freedom
## AIC: Inf
##
## Number of Fisher Scoring iterations: 5
```

We can see from this output that we have a p-value of < 2e-16, which is *very* small, meaning we are able to reject the null hypothesis, and we see evidence that there is a true difference in mean Tick.Population.Density between 2008 and 2021.

Let's replot:

```
tick_adult_08 <- tick_adult_small %>% filter(Year == "2008")

tick_adult_08_mean <- tick_adult_08 %>% group_by(Year) %>% summarise(mean_density = mean(Tick.Populatio

tick_adult_08_region_mean <- tick_adult_08 %>% group_by(Year, Region) %>% summarise(mean_density = mean

tick_adult_21 <- tick_adult_small %>% filter(Year == "2021")

tick_adult_21_mean <- tick_adult_21 %>% group_by(Year) %>% summarise(mean_density = mean(Tick.Populatio

tick_adult_21_region_mean <- tick_adult_21 %>% group_by(Year, Region) %>% summarise(mean_density = mean
```

```
ggplot(data = tick_adult_small, aes(x = Tick.Population.Density, y = Region)) +
  geom_point(aes(shape = Year, colour = Year)) +
  geom_vline(data = tick_adult_08_mean, aes(xintercept = mean_density), colour = "#F8766D") +
  geom_vline(data = tick_adult_21_mean, aes(xintercept = mean_density), colour = "#00BFC4") +
  geom_point(data = tick_adult_08_region_mean, aes(x = mean_density, y = Region), size = 3, shape = 16,
  geom_point(data = tick_adult_21_region_mean, aes(x = mean_density, y = Region), size = 3, shape = 17,
```

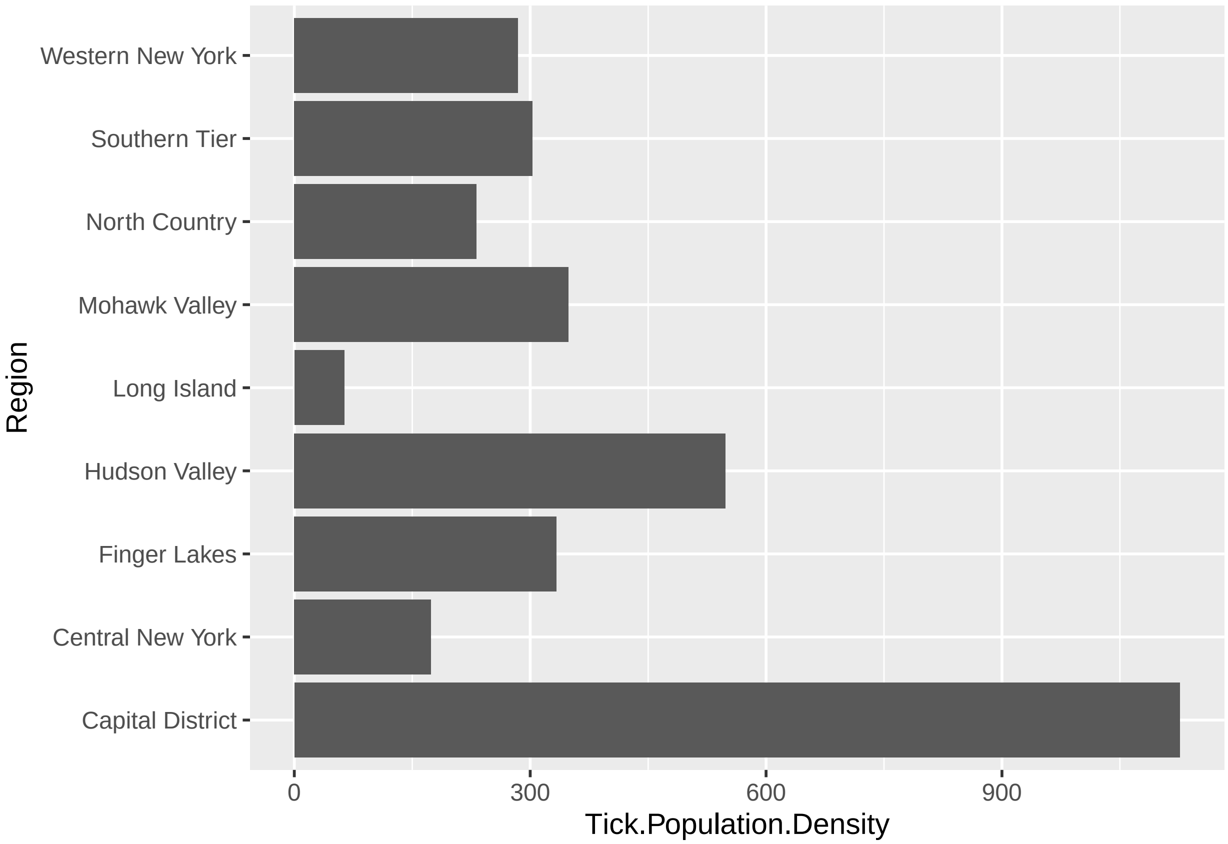


We can see here that there looks like a difference between 2008 and 2021, which we saw with our very small p-value.

Is there an association between Tick.Population.Density and Region (in 2008 and 2021)?

Let’s plot the Regions compared to Tick.Population.Density:

```
ggplot(data = tick_adult_small, aes(x = Region, y = Tick.Population.Density)) +
  geom_col() +
  coord_flip()
```



It looks like there are possible differences, with Capital Region being much higher, and also possibly Hudson Valley have a higher density? Long Island seems to have a pretty low tick density.

We are going to use the model from before, using the following hypotheses:

– Ho: There is no difference in Region mean Tick.Population.Density

– Ha: There is a difference in Region mean Tick.Population.Density

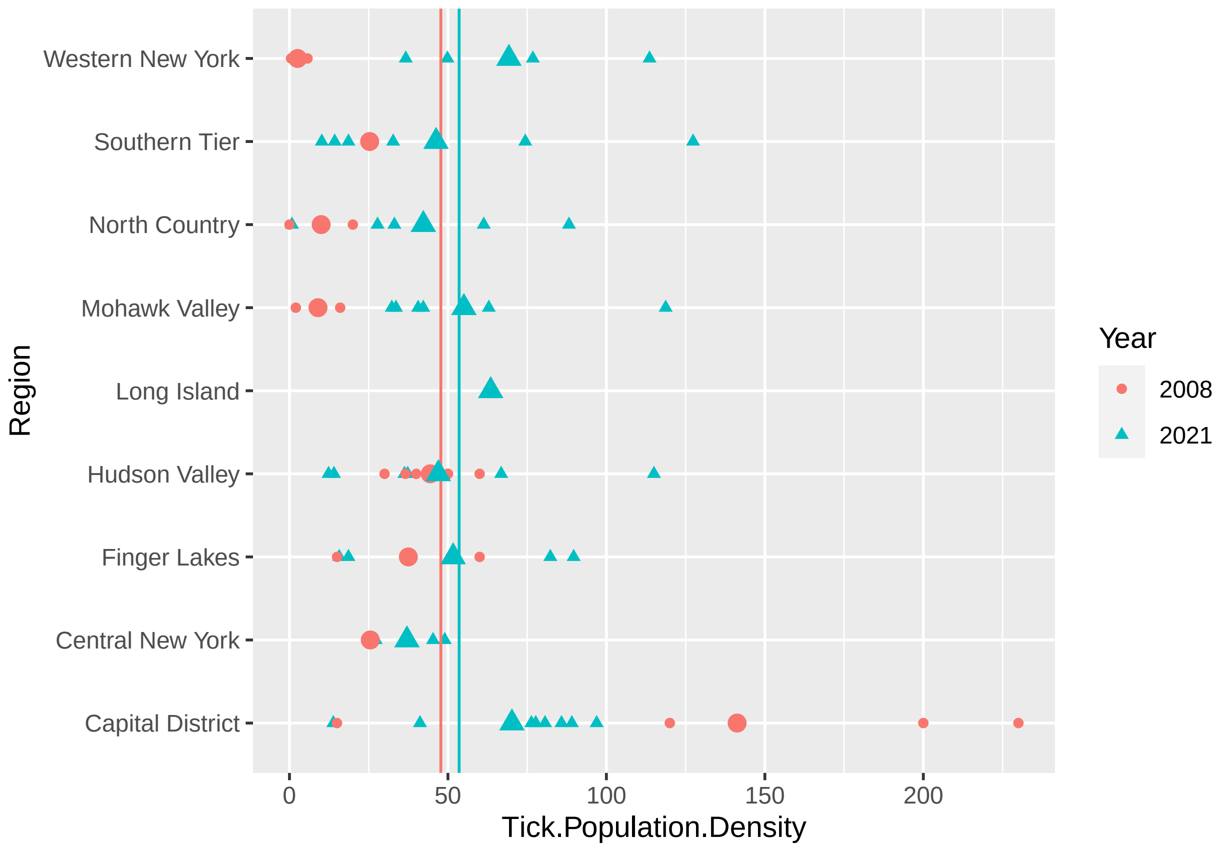
```
summary(full_mod)
```

```
##
## Call:
## glm(formula = Tick.Population.Density ~ Region + Year + Region *
##     Year, family = "poisson", data = tick_adult_small)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -13.6098   -2.7329   -0.3383    2.1489    9.7861
##
## Coefficients: (1 not defined because of singularities)
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       4.95053    0.04207 117.673  < 2e-16 ***
## RegionCentral New York           -1.71185    0.20245  -8.456  < 2e-16 ***
## RegionFinger Lakes               -1.32619    0.12290 -10.791  < 2e-16 ***
## RegionHudson Valley              -1.15654    0.07430 -15.565  < 2e-16 ***
## RegionLong Island                -0.10039    0.13241  -0.758    0.448
## RegionMohawk Valley             -2.75331    0.23943 -11.500  < 2e-16 ***
## RegionNorth Country             -2.64795    0.22752 -11.638  < 2e-16 ***
## RegionSouthern Tier             -1.71973    0.20321  -8.463  < 2e-16 ***
## RegionWestern New York          -4.00792    0.36282 -11.047  < 2e-16 ***
## Year2021                        -0.69925    0.05959 -11.735  < 2e-16 ***
## RegionCentral New York:Year2021  1.07345    0.22251   4.824 1.40e-06 ***
## RegionFinger Lakes:Year2021      1.01948    0.14407   7.076 1.48e-12 ***
## RegionHudson Valley:Year2021     0.75463    0.10417   7.245 4.34e-13 ***
## RegionLong Island:Year2021            NA         NA      NA       NA    
## RegionMohawk Valley:Year2021     2.51021    0.24927  10.070  < 2e-16 ***
## RegionNorth Country:Year2021     2.13994    0.24141   8.864  < 2e-16 ***
## RegionSouthern Tier:Year2021     1.30258    0.21605   6.029 1.65e-09 ***
## RegionWestern New York:Year2021  3.99415    0.37018  10.790  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2212.1  on 65  degrees of freedom
## Residual deviance: 1180.6  on 49  degrees of freedom
## AIC: Inf
##
## Number of Fisher Scoring iterations: 5
```

We have the Capital District as the reference group, and we can see that all of the p-values are < 2e-16, except for Long Island. For the others, since we have such a small p-value, we can reject the null hypothesis, and decide that there is a difference in the specific Region mean Tick.Population compared to the Capital District.

Let's replot again:

```
ggplot(data = tick_adult_small, aes(x = Tick.Population.Density, y = Region)) +
  geom_point(aes(shape = Year, colour = Year)) +
  geom_vline(data = tick_adult_08_mean, aes(xintercept = mean_density), colour = "#F8766D") +
  geom_vline(data = tick_adult_21_mean, aes(xintercept = mean_density), colour = "#00BFC4") +
  geom_point(data = tick_adult_08_region_mean, aes(x = mean_density, y = Region), size = 3, shape = 16,
  geom_point(data = tick_adult_21_region_mean, aes(x = mean_density, y = Region), size = 3, shape = 17,
```



We can see from this plot that we can see differences between the regions' means compared to the Capital District, as the Capital District has very high mean Tick.Population.Density.

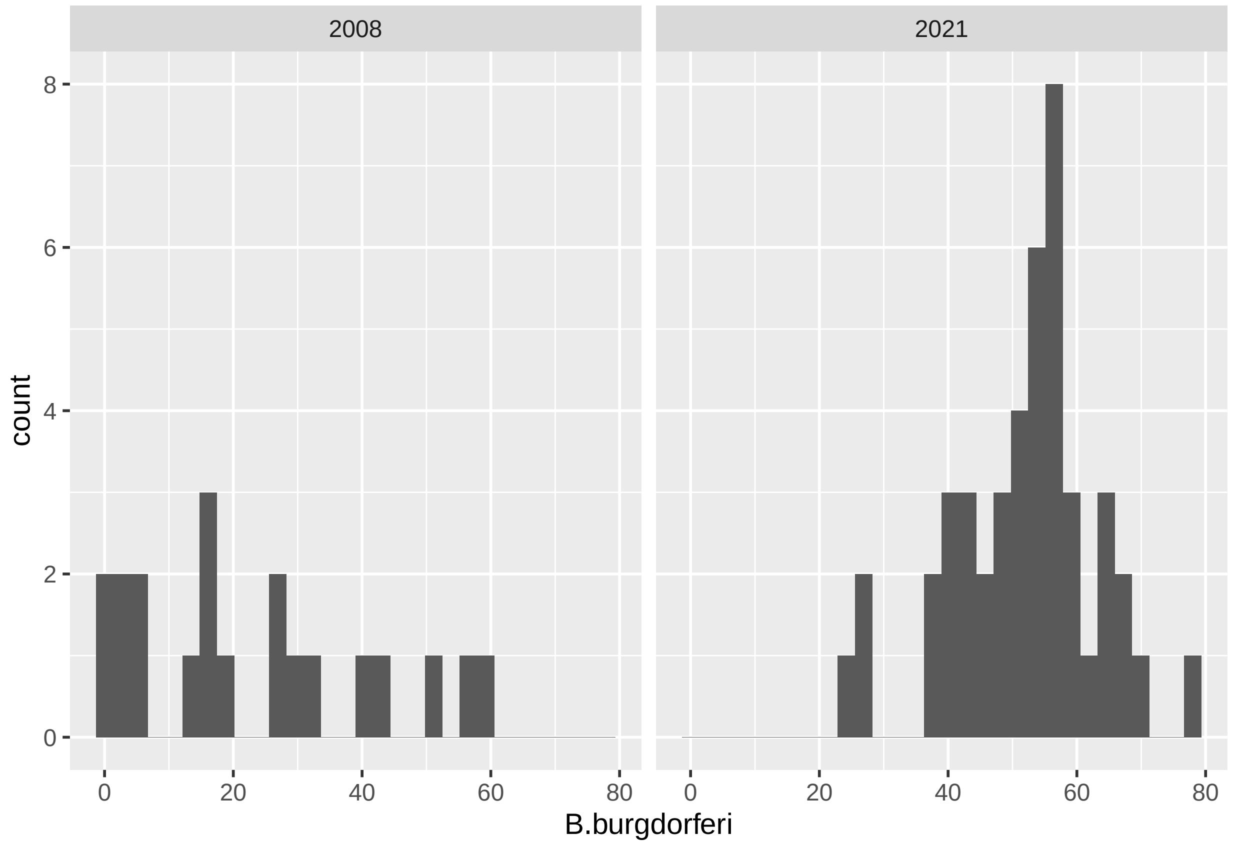
```
library(emmeans)
em <- emmeans(full_mod, "Region")
contrast(em, "pairwise", adjust = "Tukey")
```

```
##  contrast                            estimate     SE  df z.ratio p.value
##  Capital District - Central New York   1.1751 0.1113 Inf  10.563  <.0001
##  Capital District - Finger Lakes       0.8164 0.0720 Inf  11.334  <.0001
##  Capital District - Hudson Valley      0.7792 0.0521 Inf  14.961  <.0001
##  Capital District - Long Island        nonEst     NA  NA      NA      NA
##  Capital District - Mohawk Valley      1.4982 0.1246 Inf  12.021  <.0001
##  Capital District - North Country      1.5780 0.1207 Inf  13.073  <.0001
##  Capital District - Southern Tier      1.0684 0.1080 Inf   9.890  <.0001
##  Capital District - Western New York   2.0108 0.1851 Inf  10.864  <.0001
##  Central New York - Finger Lakes      -0.3587 0.1257 Inf  -2.854  0.1001
##  Central New York - Hudson Valley     -0.3959 0.1154 Inf  -3.431  0.0175
##  Central New York - Long Island        nonEst     NA  NA      NA      NA
##  Central New York - Mohawk Valley      0.3231 0.1617 Inf   1.998  0.5444
##  Central New York - North Country      0.4028 0.1587 Inf   2.539  0.2132
##  Central New York - Southern Tier     -0.1067 0.1492 Inf  -0.715  0.9986
##  Central New York - Western New York   0.8357 0.2118 Inf   3.946  0.0026
##  Finger Lakes - Hudson Valley         -0.0372 0.0783 Inf  -0.476  0.9999
##  Finger Lakes - Long Island            nonEst     NA  NA      NA      NA
##  Finger Lakes - Mohawk Valley          0.6818 0.1376 Inf   4.953  <.0001
##  Finger Lakes - North Country          0.7615 0.1341 Inf   5.679  <.0001
##  Finger Lakes - Southern Tier          0.2520 0.1228 Inf   2.052  0.5069
##  Finger Lakes - Western New York       1.1944 0.1941 Inf   6.154  <.0001
##  Hudson Valley - Long Island           nonEst     NA  NA      NA      NA
##  Hudson Valley - Mohawk Valley         0.7190 0.1283 Inf   5.602  <.0001
##  Hudson Valley - North Country         0.7987 0.1245 Inf   6.414  <.0001
##  Hudson Valley - Southern Tier         0.2892 0.1123 Inf   2.576  0.1968
##  Hudson Valley - Western New York      1.2316 0.1876 Inf   6.565  <.0001
##  Long Island - Mohawk Valley           nonEst     NA  NA      NA      NA
##  Long Island - North Country           nonEst     NA  NA      NA      NA
##  Long Island - Southern Tier           nonEst     NA  NA      NA      NA
##  Long Island - Western New York        nonEst     NA  NA      NA      NA
##  Mohawk Valley - North Country         0.0798 0.1683 Inf   0.474  0.9999
##  Mohawk Valley - Southern Tier        -0.4298 0.1595 Inf  -2.695  0.1494
##  Mohawk Valley - Western New York      0.5126 0.2191 Inf   2.340  0.3182
##  North Country - Southern Tier        -0.5095 0.1564 Inf  -3.258  0.0309
##  North Country - Western New York      0.4329 0.2169 Inf   1.996  0.5464
##  Southern Tier - Western New York      0.9424 0.2101 Inf   4.485  0.0003
##
## Results are averaged over the levels of: Year
## Results are given on the log (not the response) scale.
## P value adjustment: tukey method for comparing a family of 9 estimates
```

We can see the various comparisons between specific regions, and I am from the North Country, which is significant compared to all besides Mohawk Valley.

Is there a difference between mean B.burgdorferi percentage in the years 2008 and 2021?

```
ggplot(data = tick_adult_small, aes(x = B.burgdorferi)) +
  geom_histogram() +
  facet_wrap(~Year)
```

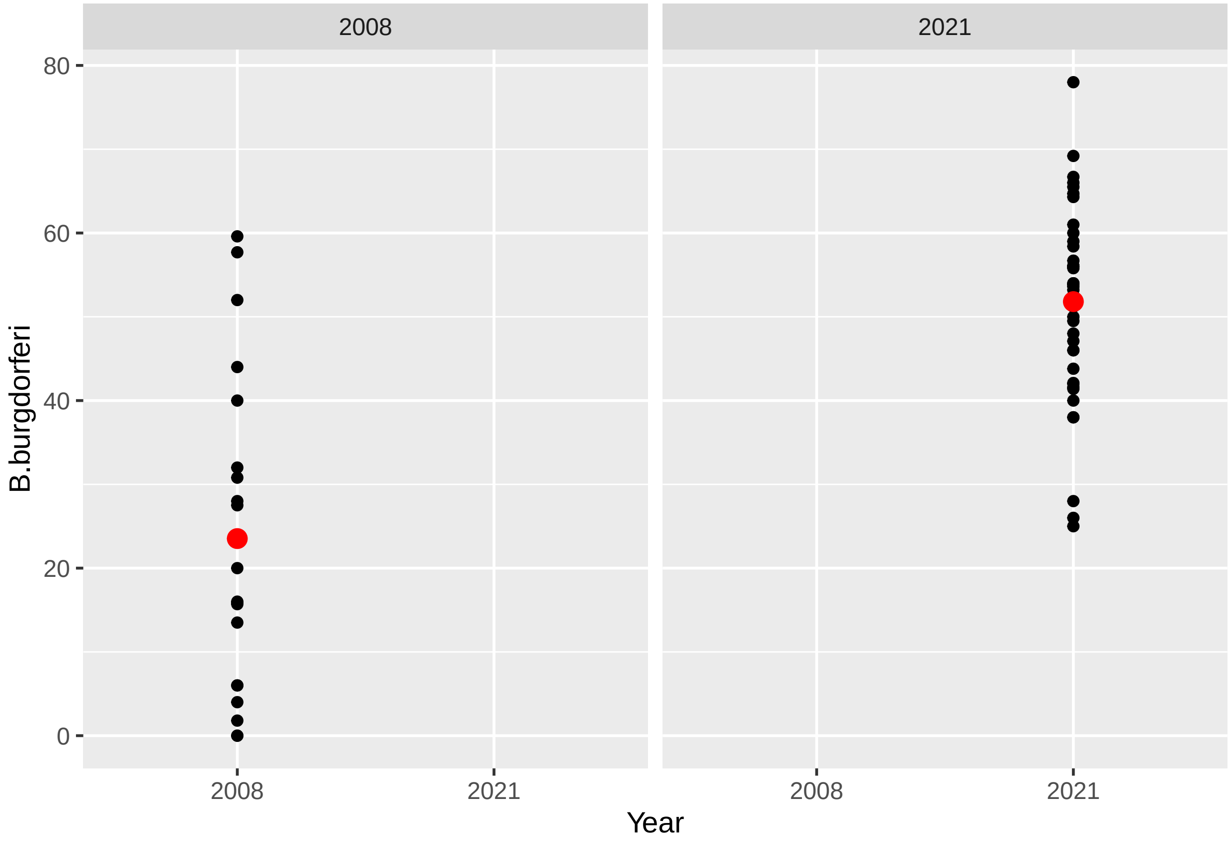


There looks to be a difference in where the peaks are on the y-axis, so this leads me to believe that there could be a significant difference in means between the years.

This will lead me to run a two-sample t-test with the following hypotheses:

Null Hypothesis: There is no difference in mean B.burgdorferi percentage between the years 2008 and 2021. Alternative Hypothesis: There is a difference in mean B.burgdorferi percentage between the years 2008 and 2021.

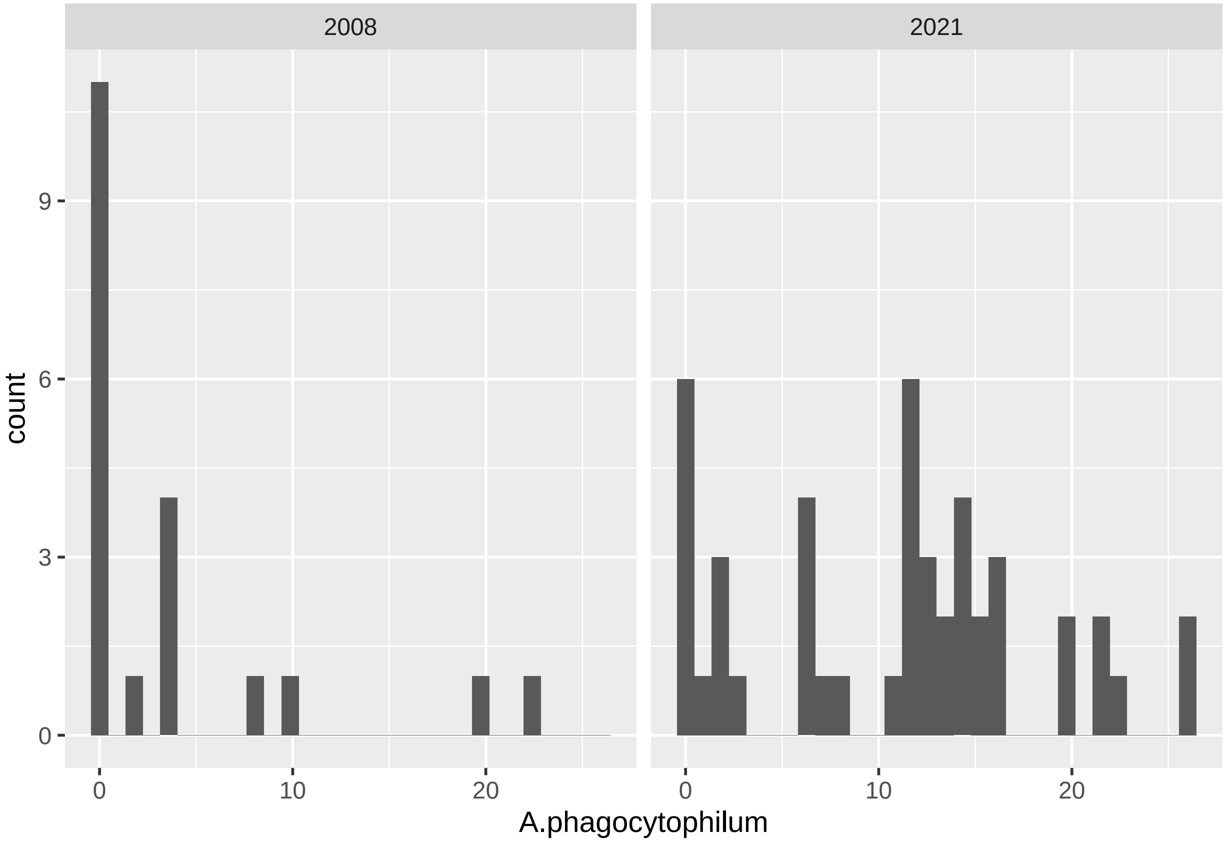
```
bburg_ttest <- t.test(B.burgdorferi ~ Year, data = tick_adult_small)

tick_adult_bburg_mean <- tick_adult_small %>% group_by(Year) %>% summarise(mean = mean(B.burgdorferi, na
ggplot(data = tick_adult_small, aes(x = Year, y = B.burgdorferi)) +
  geom_point() +
  geom_point(data = tick_adult_bburg_mean, aes(x = Year, y = mean), colour = "red", size = 3) +
  facet_wrap(~Year)
```



We can see a p-value of 2.034679e-06, which is very small, and leads us to determine that we can reject our null, and see evidence that there is a significant difference between the mean B.burgdorferi percentage in the years 2008 and 2021.

Is there a difference between mean A.phagocytophilum percentage in the years 2008 and 2021?

```
ggplot(data = tick_adult_small, aes(x = A.phagocytophilum)) +
  geom_histogram() +
  facet_wrap(~Year)
```



Here it looks like the peak is higher for 2021, meaning the mean A.phagocytophilum percentage could possibly be significantly different compared to 2008. I will now test to see if there is a significant difference in the means.

This will lead me to run a two-sample t-test with the following hypotheses:

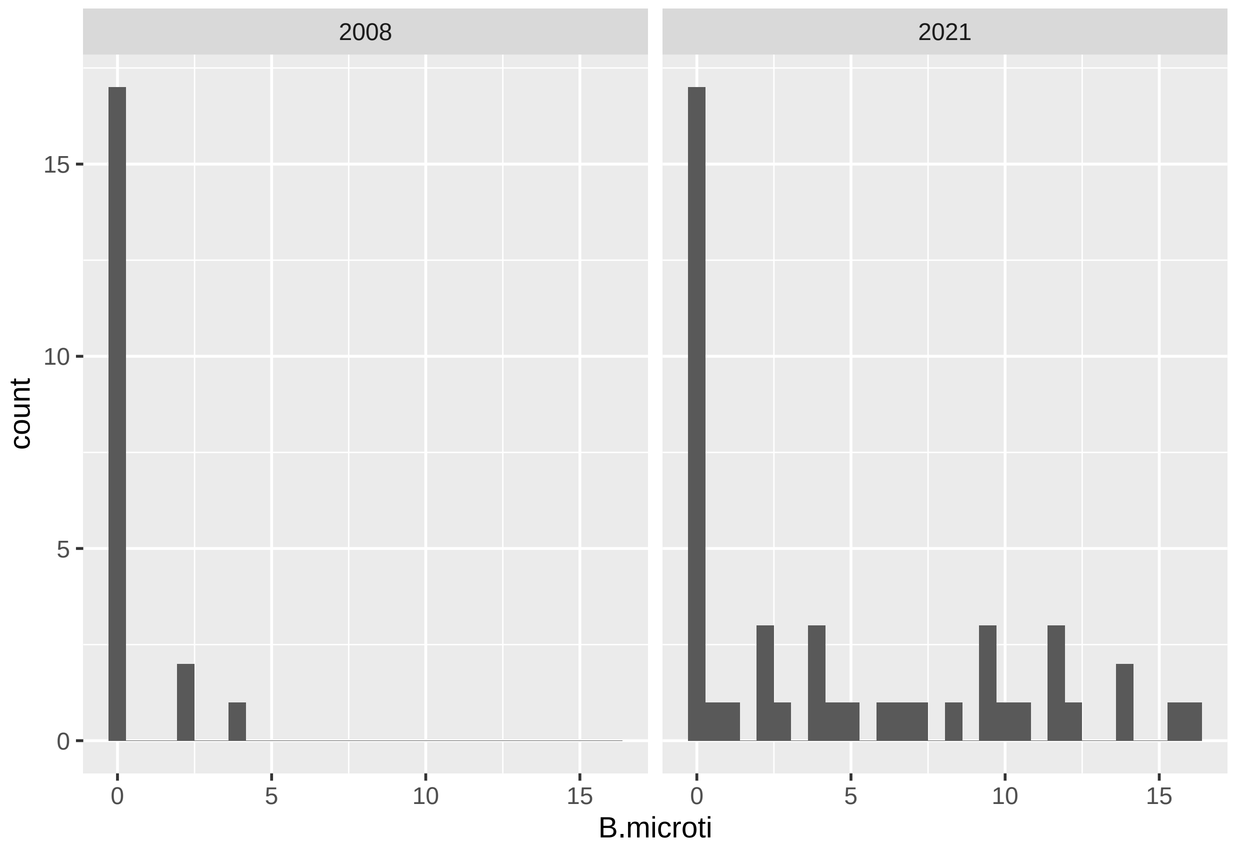
Null Hypothesis: There is no difference in mean A.phagocytophilum percentage between the years 2008 and 2021. Alternative Hypothesis: There is a difference in mean A.phagocytophilum percentage between the years 2008 and 2021.

```
aphago_ttest <- t.test(A.phagocytophilum ~ Year, data = tick_adult_small)
```

We can see a p-value of 0.0004769, which is small, and leads us to determine that we can reject our null, and see evidence that there is a significant difference between the mean A.phagocytophilum percentage in the years 2008 and 2021.

Is there an association between mean B.microti percentage in the years 2008 and 2021?

```
ggplot(data = tick_adult_small, aes(x = B.microti)) +
  geom_histogram() +
  facet_wrap(~Year)
```

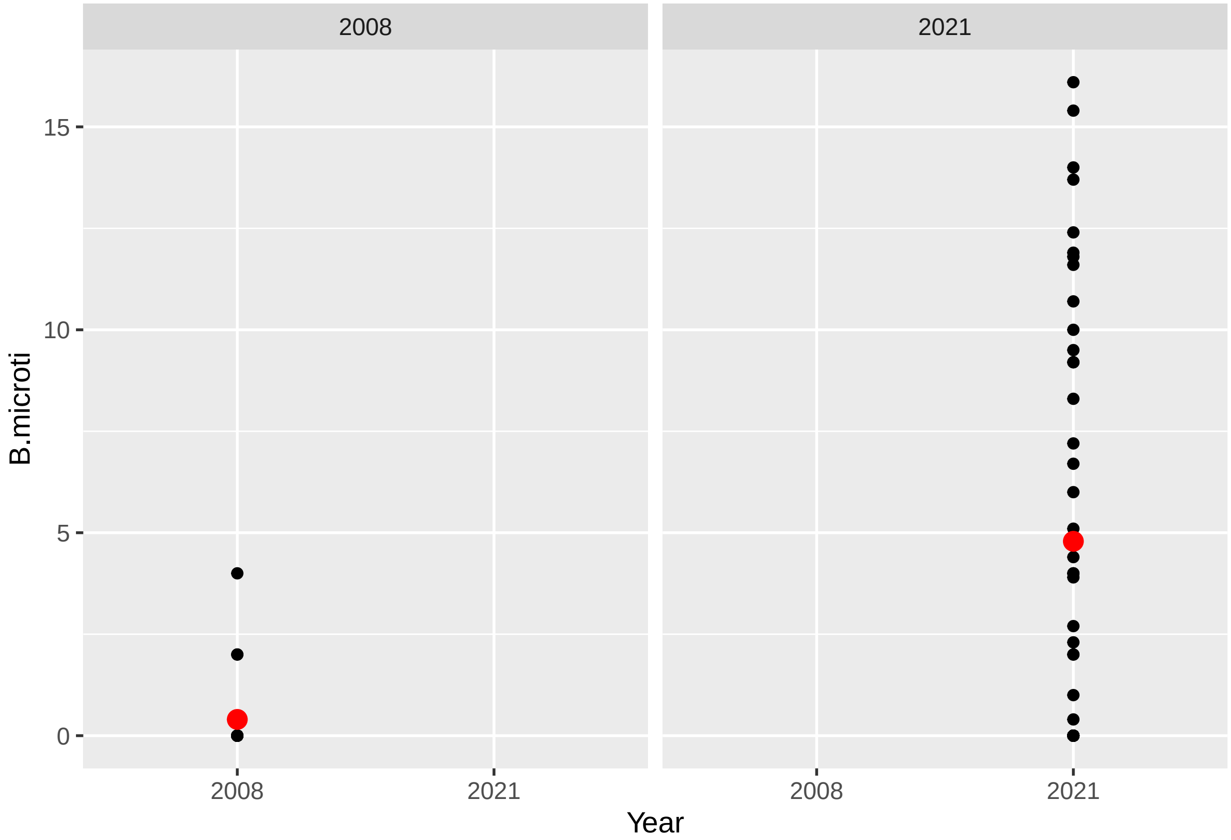


There is more spread for 2021, however the peaks both seem to be at or close to 0%. I think this one would show less significance than the other questions regarding bacteria.

This will lead me to run a two-sample t-test with the following hypotheses:

Null Hypothesis: There is no difference in mean B.microti percentage between the years 2008 and 2021. Alternative Hypothesis: There is a difference in mean B.microti percentage between the years 2008 and 2021.

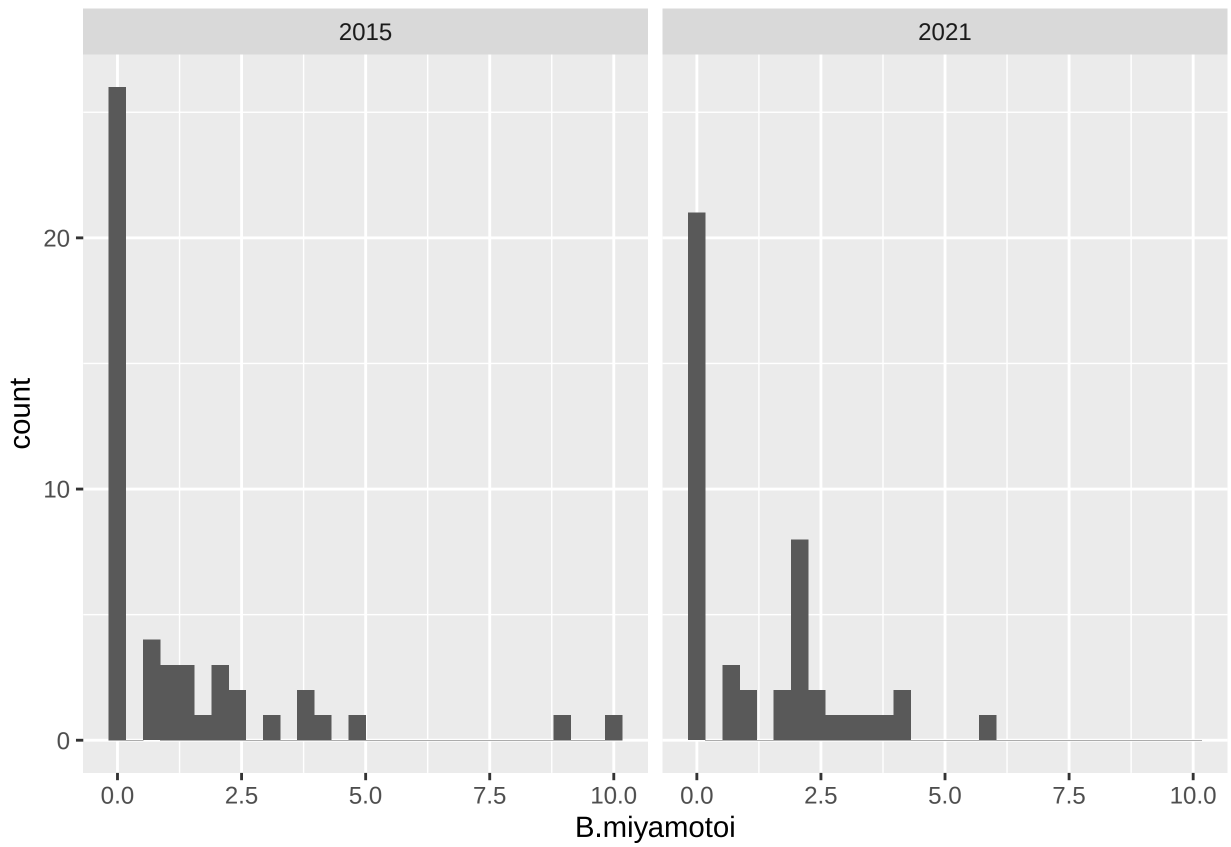
```
bmicroti_ttest <- t.test(B.microti ~ Year, data = tick_adult_small)
tick_adult_microti_mean <- tick_adult_small %>% group_by(Year) %>% summarise(mean = mean(B.microti, na.
ggplot(data = tick_adult_small, aes(x = Year, y = B.microti)) +
  geom_point() +
  geom_point(data = tick_adult_microti_mean, aes(x = Year, y = mean), colour = "red", size = 3) +
  facet_wrap(~Year)
```



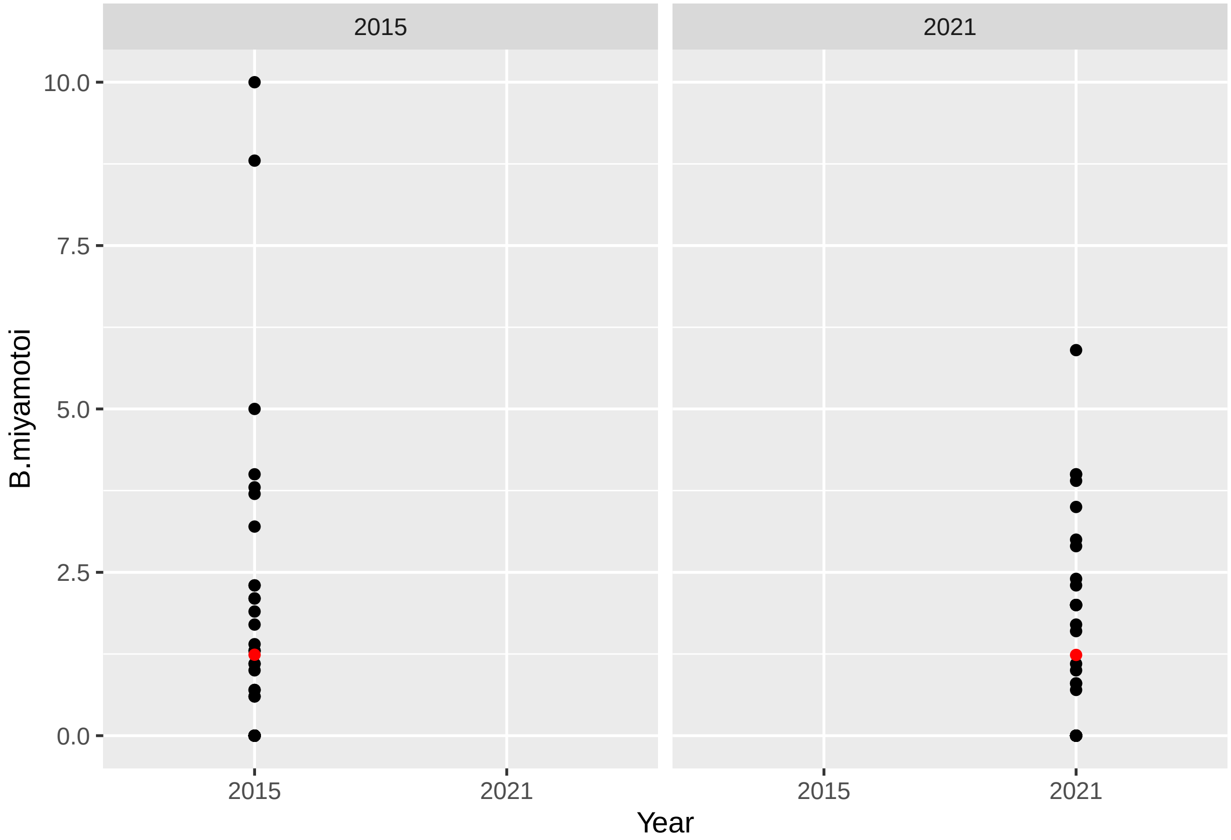
We can see a p-value of 1.829822e-06, which is small, and leads us to determine that we can reject our null, and see evidence that there is a significant difference between the mean B.microti percentage in the years 2008 and 2021.

Is there an association between mean B.miyamotoi percentage in the years 2015 and 2021 (there was no data for 2008-2014)?

```
tick_adult_small2 <- tick_adult %>% filter(Year == "2015" | Year == "2021")
tick_adult_small2_mean <- tick_adult_small2 %>% group_by(Year) %>% summarise(mean = mean(B.miyamotoi, na
ggplot(data = tick_adult_small2, aes(x = B.miyamotoi)) +
  geom_histogram() +
  facet_wrap(~Year)
```



```
ggplot(data = tick_adult_small2, aes(x = Year, y = B.miyamotoi)) +
  geom_point() +
  geom_point(data = tick_adult_small2_mean, aes(x = Year, y = mean), colour = "red") +
  facet_wrap(~Year)
```



This one is not as obvious that there could be a difference in the means. They both have peaks around zero, however there are some values in 2015 greater than 7.5%, which might make it have a higher mean, and cause the difference to be significance.

This will lead me to run a two-sample t-test with the following hypotheses:

Null Hypothesis: There is no difference in mean B.miyamotoi percentage between the years 2015 and 2021. Alternative Hypothesis: There is a difference in mean B.miyamotoi percentage between the years 2015 and 2021.

```
bmiya_ttest <- t.test(B.miyamotoi ~ Year, data = tick_adult_small2)
```

This is a very high p-value (0.9932), meaning that there is not enough evidence to reject the null that there was no difference between the mean B.miyamotoi percentage between the years 2015 and 2021.

Biological Summary

We have found that there was a significant difference in tick population density between the years 2008 and 2015, which is different from Paul (2016) stating that there was no difference in tick population over time. Paul (2016) also describes how ticks are spreading into more urban areas with urbanization and the spread of their hosts. We can see that the tick population density test by region was significant, we just need to see if possibly doing a two-sample t-test between a more urban and more rural region would lead us to more evidence of region affecting density.

When looking at region and year, we could see differences in the mean tick population desnity between almost all regions when being compared to the Capital Region, except for Long Island. We would have to look to see if the Capital Region is more urban, however we know for certain that some more rural regions like the North Country, have a signficant difference compared to the Capital Region. The interaction terms are being compared to the Capital Region, and the only non-signficant ones were between year and CNY and Southern Tier. Both of these regions are more rural, with some urban areas. Paul (2016) explained how areas with forests or even now more suburban areas close to forests are experiencing higher tick population density.

We found that for all bacteria and parasites, they were all signficant except for B.miymaotoi. If we compared this by region, we might be able to see a difference in mean percentage of ticks with B.miyamotoi for the years. Kowalec saw differences between urban and rural areas, and with more urbanization we can see a signficant difference between the years, however it would be helpful to look at regions that are more urban versus rural.

Challenges

I have had to think back to things I have learned in statistics classes up to two years ago. I do not think that I am interpreting the glm model with the interaction terms correctly. It was brought up using interaction terms, however, I think this model has too many parameters, so I focused on 2008 and 2021, and even after that it still has too many parameters.

I think for the first test using year and density, I decided to just use two years, the beginning (2008) and the end (2021) of the data in order to simplify things. I tried to do a time series plot, but could not get the data in the correct format. I also attempted to put the Tukey letters onto a plot, however kept on getting a lot of errors that did not make sense when I looked them up.

Works Cited

Kowalec, M., Szewczyk, T., Welc-Falęciak, R., Siński, E., Karbowiak, G., & Bajer, A. (2017). Ticks and the city - are there any differences between city parks and natural forests in terms of tick abundance and prevalence of spirochaetes?. Parasites & vectors, 10(1), 573. https://doi.org/10.1186/s13071-017-2391-2

Paul, R.E.L., Cote, M., Le Naour, E. et al. Environmental factors influencing tick densities over seven years in a French suburban forest. Parasites Vectors 9, 309 (2016). https://doi.org/10.1186/s13071-016-1591-5