Data project 1: Birds, birds, birds

Jilli Violano ([redacted])

Part 1: Critical Thinking

1. When it comes to noise and uncertainty, there are many places where this data set can go wrong. First, the maximum count of individuals coming and going can be very difficult to count at one time, especially if there are multiple species of birds. There is also no indication of the time taken or the time of day taken. Some birds may be more active at night or in the morning, than other birds, and thus may be miscounted and represented. Lastly, some birds may be misrepresented as other birds, especially if there is only a short period of observation, leading to over counting of one species and under counting of another. Also, not all birds are seed eating bird and some are omnivores or carnivores (think hawks), and thus will not capture that population very well. There can also be typos in the data as well.

2. FeederWatch probably chooses to account for maximum individuals because some birds can come and go, and since birds are hard to determine individually from each other, it will most likely over count the true amount of individual birds, since some can go and come back. This insures that there is not an over counting of a species, assuming the noise from above is minimized.

3. The data will heavily rely on the earlier amount of present birds in our estimate, and location of those birds being present. For instance, if we are more likely to find cranes ducks to water, someone in suburbia near no water will probably not count them as much as someone closer to the coast, and since this data is self reported/volunteered, FeederWatch cannot randomly select participants to cover those areas, so where we think there is change over time, it might just be under representation or over representation as more representation is accounted for. On top of this, many more individuals in 2020 and 2021 have more free time, as the pandemic hit during that time and was able to observe more often, as interacting in person was frowned upon (and thus encouraged to spend more time with birds).

Part 2: Working with the data

1. In 2011, there were 70 unique locations, while in 2021, there were 119 unique bird feeder locations.
2. Upon observing both of the data frames, there is a bit of a change with the average birds being seen. There is a general difference in average counts before and average counts after, which gives us generally smaller maximum flock sizes in 2021 than 2011, which makes sense, since there were less reports in response. Another observable part is the naming of some of the species do not remain consistent over the years- for instance, Goldfinch is noted as "Spinus sp. (goldfinch sp.)" and "Lawrence's Goldfinch" in 2011 and "Lesser Goldfinch" in 2021. These might be the same bird but accounted differently. Another place where there is more weight is that the "Spinus sp. (goldfinch sp.)" and "Lawrence's Goldfinch" are whole numbers, which implies that there were likely only one entry, which does not give us enough information to determine average. Included, the average count of birds has gone down for the top 5 birds, as in 2011, they were typically above 10 and in 2021, they were (mostly) below 10. On both list, we can see that the Cedar Waxwing, Wild Turkey, and Goldfinches stay in those top spots, while the Red Winged Blackbird is outranked by the Rock Pigeon and Pine Siskin.

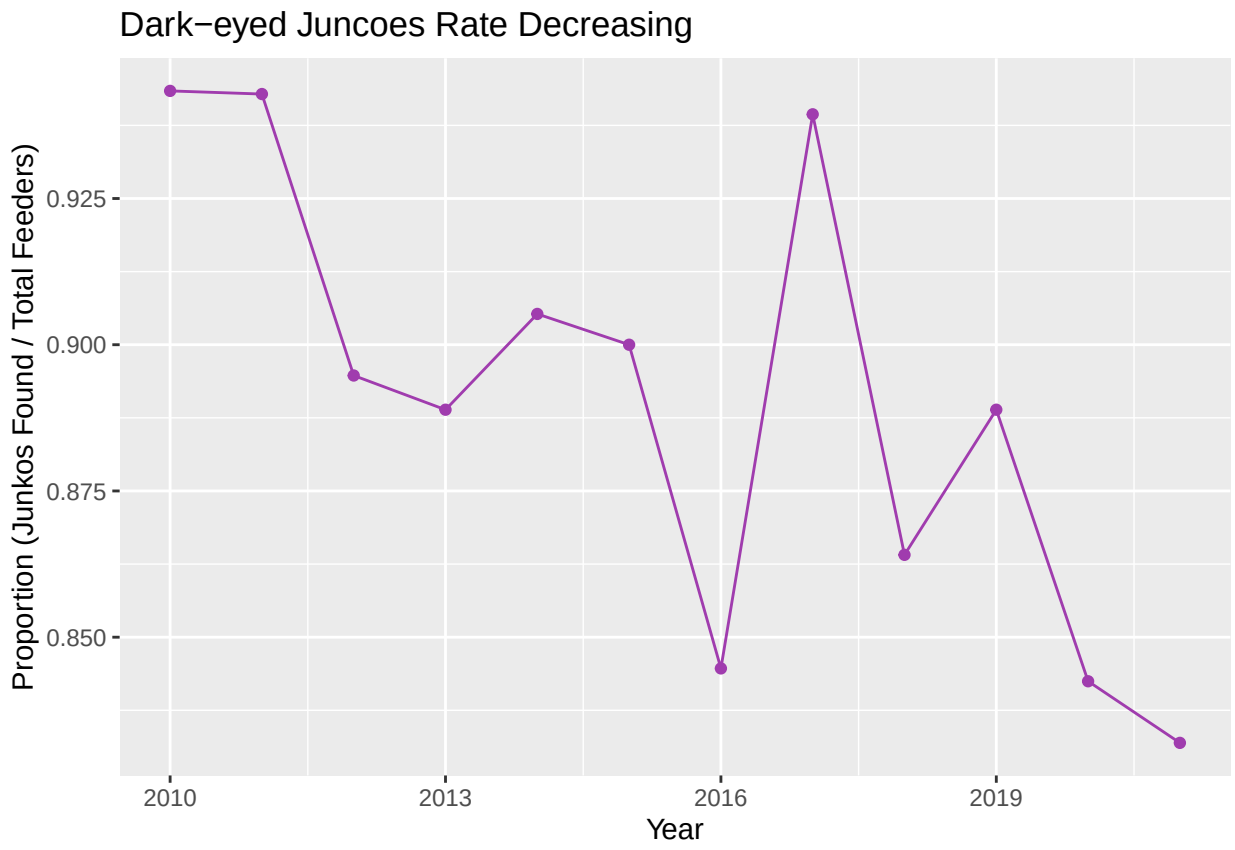
2011:

```
## # A tibble: 5 x 2
##   species_name                avgct
##   <chr>                       <dbl>
## 1 Cedar Waxwing                13.8
## 2 Wild Turkey                  14.0
## 3 Lawrence's Goldfinch         14
## 4 Spinus sp. (goldfinch sp.)   18
## 5 Red-winged Blackbird         19.2
```

2022

```
## # A tibble: 5 x 2
##   species_name                avgct
##   <chr>                       <dbl>
## 1 Lesser Goldfinch             5.75
## 2 Pine Siskin                  6.45
## 3 Wild Turkey                  7.32
## 4 Rock Pigeon (Feral Pigeon)   7.42
## 5 Cedar Waxwing               14.7
```

3. Over time, there has been a general trend downwards with the exception of a spike in 2017. The general range of values is from about .95 to .80.

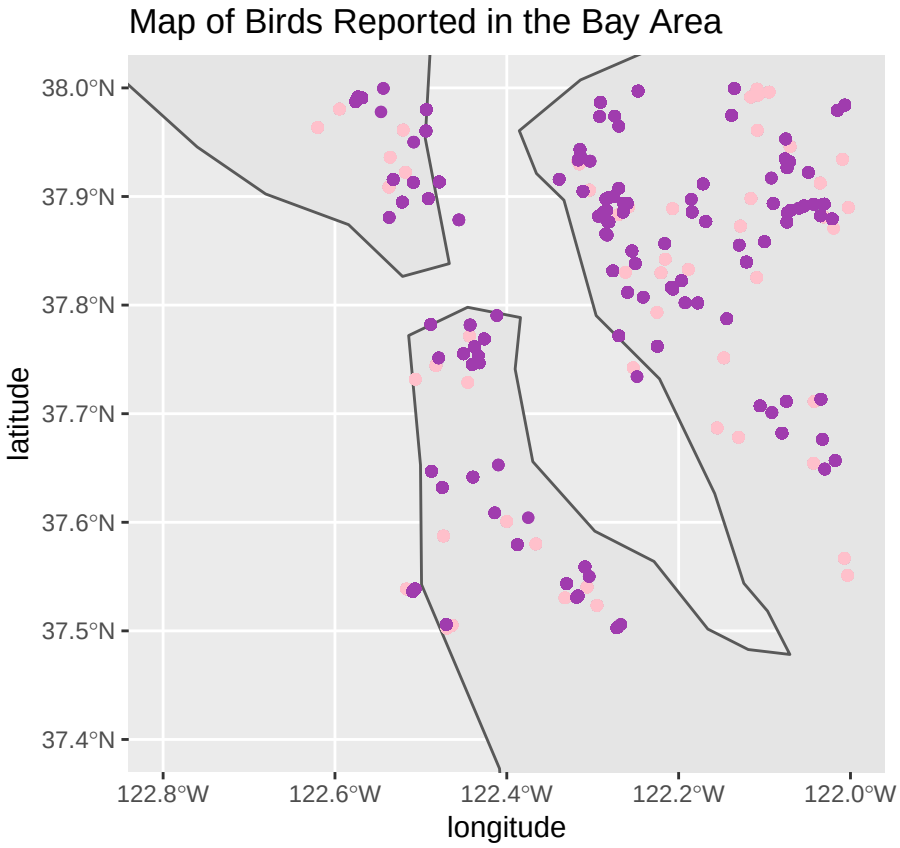


Part 3: EDA

First I want to look at the birds, we have right off the bat, to get an idea of what is captured.

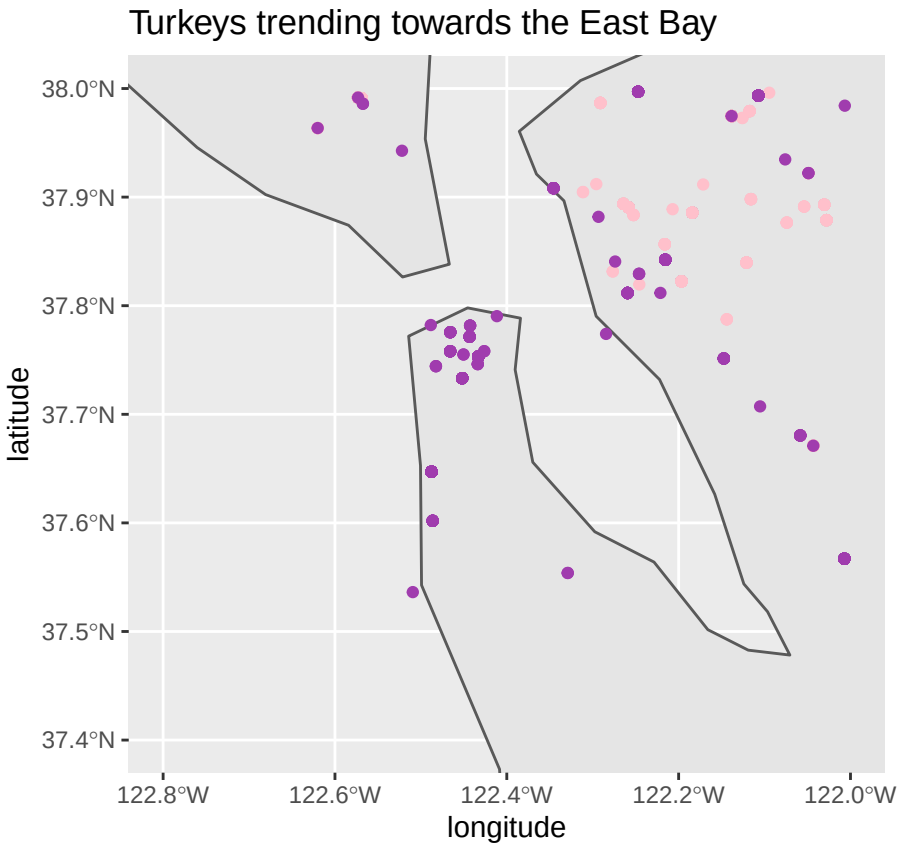
```
## # A tibble: 164 x 2
##    species_name          avgct
##    <chr>                 <dbl>
##  1 Accipiter sp.          1
##  2 Acorn Woodpecker       1.74
##  3 African Collared-Dove  1
##  4 Allen's Hummingbird    1.19
##  5 American Crow          3.32
##  6 American Goldfinch     4.73
##  7 American Kestrel       1
##  8 American Pipit         2
##  9 American Robin         3.77
## 10 Anna's Hummingbird     1.68
## # ... with 154 more rows
```

My initial reaction is overwhelmed since there are so many, but we notice the data is a little messy, as some count different colorings of the same species as different species (Dark-eyed Junco has itself, but also an Oregon and Slate Colored variant), or do not use proper naming (ie: hawk sp. on row 73). On top of this, there were types of birds, such as Owls that are generally nocturnal and carnivorous, and therefore would not usually be seen at a bird feeder, unless its feasting on squirrels ransacking the feeders. After getting an idea of these birds and their corresponding entries, I wanted to map where they were located to get a better idea.



For this particular visualization, I wanted to map out the locations of the birds reported on the maps and

see the difference between spread of locations in both 2011 (pink) and 2021 (purple). As expected, there is a fair amount of clustering in urban areas, which makes sense given most people are probably observing their home bird feeder. Another observation this tells is that there is data from not only San Francisco peninsula, but also from East bay and North bay, of which have way different terrains and thus may give us some information on which birds like to remain where. I do that with my next graph– where I compare Wild Turkeys and Rock Pigeon (Feral Pigeon) sightings.



As an avid lover of the turkeys on campus, especially when I visit my friends in Stern Hall, I wondered to myself if they ever found themselves vacationing in San Francisco. As I observed closer, the Turkeys (pink) were exclusively in North/East Bay in sightings, while other species, such as the Pigeon (purple), found itself in more urban areas typically, with a large cluster near San Francisco, but as expected. This outcome shocked me however with the turkeys, since my sister, who lives in Monterey Bay, sees them fairly often and she is less than a mile from the coast. One explanation for why this might be is because of the extensive land beyond her, while West Bay is limited in land due to the water surrounding it, making predators hunt down turkeys faster. I feel that terrain also has a say in the behavior, as (fun fact!) Turkeys tend to sleep in trees to avoid predators on the ground (coyotes and wolves to name a few). Because of the urbanization of San Francisco and most of the land there, there might not be many places for the Turkeys to rest out of harms way, where as where my sister lives, the land is a lot less developed and is more wild life centered.

Part 4: Parameter estimation

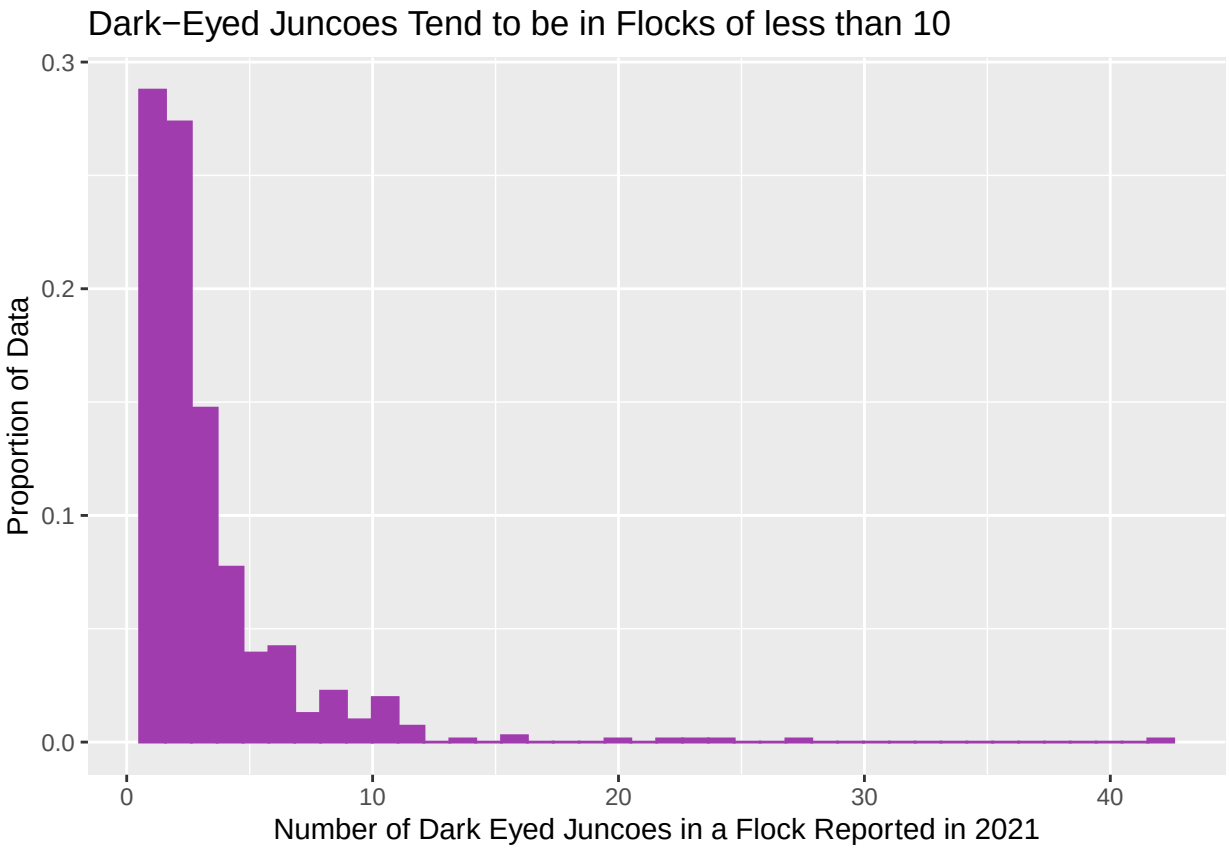
1. The population we are trying to capture with this data set is the birds at various feeders in the Bay Area.

2. This data is very much biased. Although in Part 1 I covered this extensively, I will highlight the most important parts. First, the data is subject to bias due to self reported data. This causes noise such as

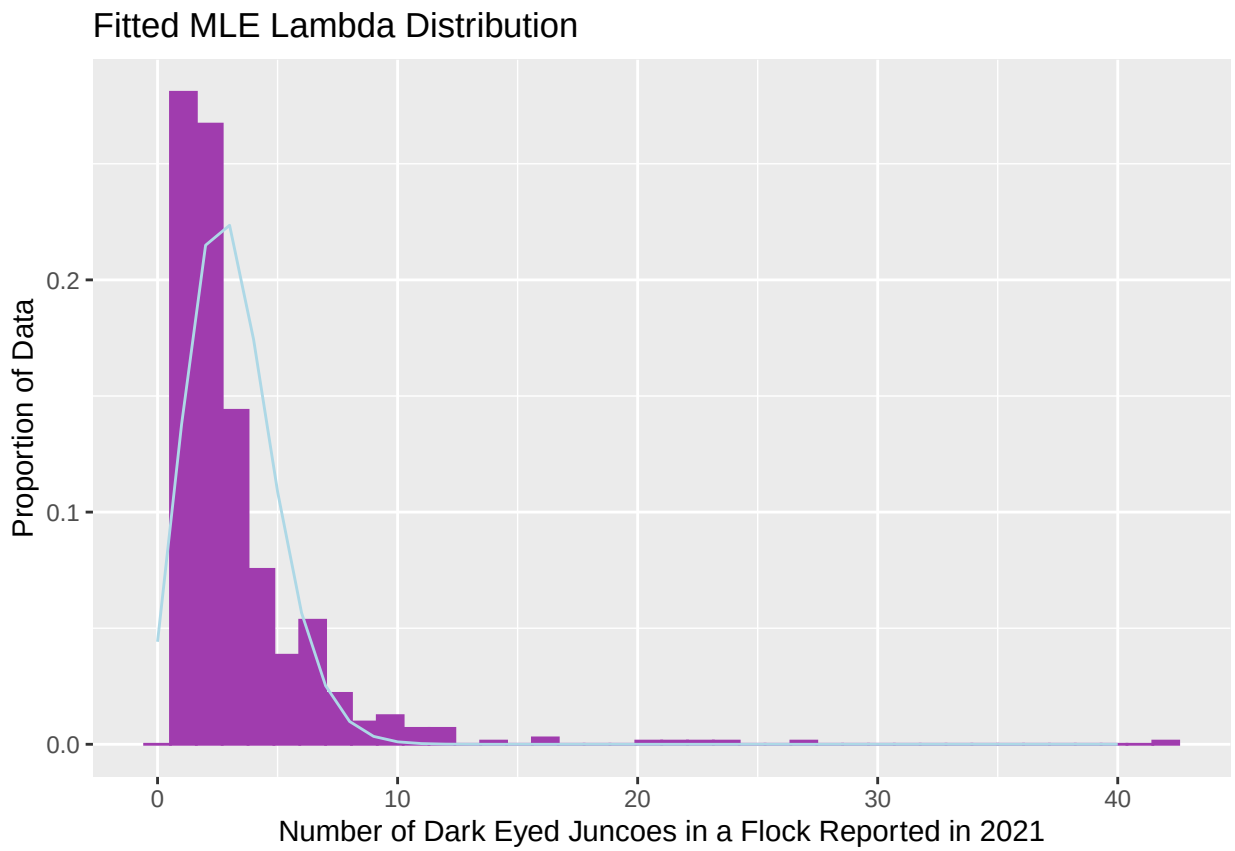
incorrect counting, identification of bird species, and entry of data. Second, due to the vicariousness of the locations, since this is a citizen science project, some areas might end up inadvertently clustering (such as neighbors bird watching and reporting similar data). Since there is no random selection, it is left up to this bias as a result.

3. From coding results, I got an average of 3.12 for the average max individuals of Dark-eyed Juncoes.

4. Please refer to histogram below.

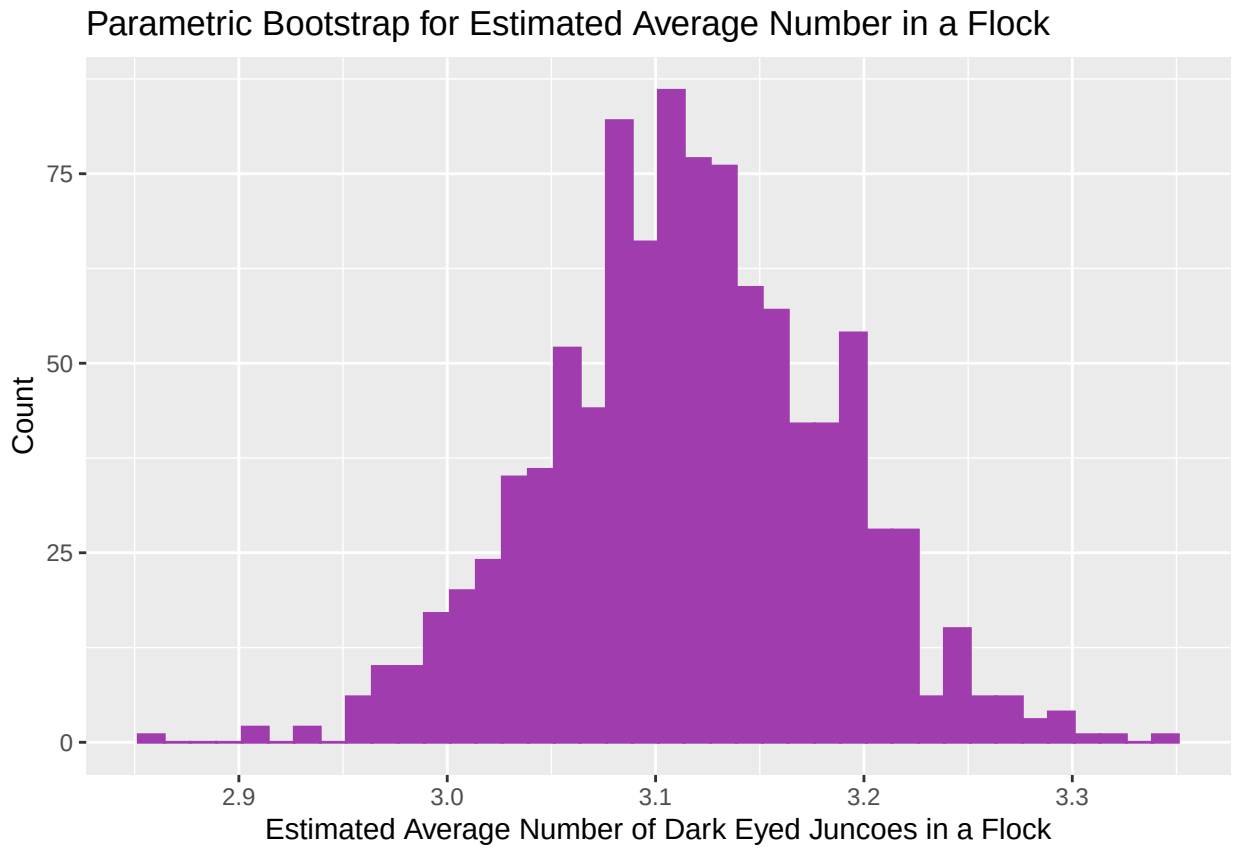


5. Using what we derived in lecture 5 of Stat 135, we know that the MLE estimate for λ is $\bar{X}$ and thus we can assume our fitted distribution follows a poisson(3.12). Visually looking at the data, I personally feel that it does not do a great job at capturing all the data, but it makes a rather rough estimate of the proportion of each flock size. Some areas that mightve contributed to the higher MLE estimate would be outliers, such as the couple about 40, which could sway our sample mean further to the right when it would fit better more snug to the left. Refer to histogram below.



6. Our estimated bias is 0.0041 and our estimated variance is 0.0046.

```
#reference: bootstrap_mean.R in course files
set.seed(13)
p_flock <- map_df(1:1000, function(i) {
  # sample from the approximated distribution
  b <- rpois(length(junco21$max_individuals), mean(junco21$max_individuals))
  # compute the sample mean of the parametric bootstrap sample
  data.frame(b_m = mean(b))
})
ggplot(p_flock, aes(x=b_m)) + geom_histogram(colour="#A03CAE", fill="#A03CAE", bins = 40) +
  xlab("Estimated Average Number of Dark Eyed Juncoes in a Flock") +
  ylab("Count") +
  labs(title = "Parametric Bootstrap for Estimated Average Number in a Flock")
```




```
#bias calc
3.12 - mean(p_flock$b_m)
```

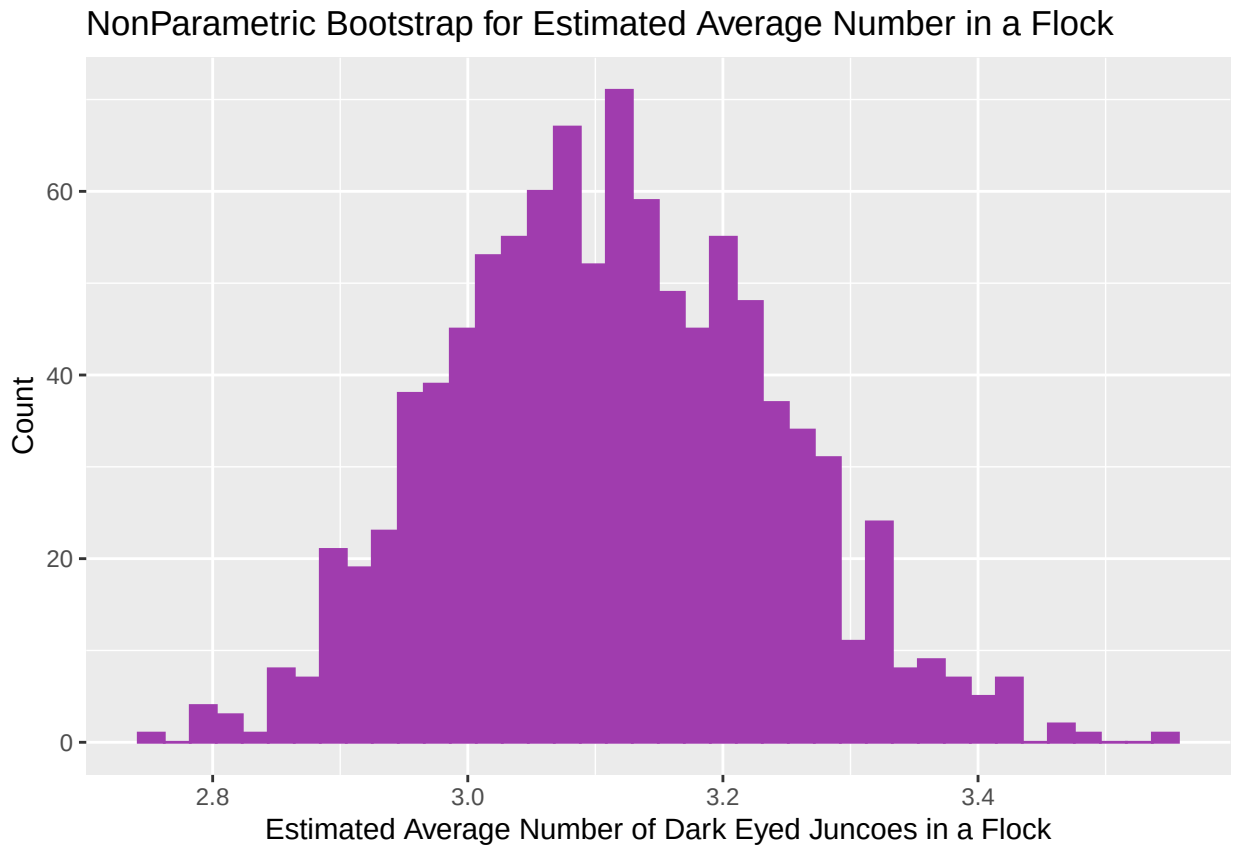
```
## [1] 0.004073746
```

```
#variance calc
var(p_flock$b_m)
```

```
## [1] 0.004569946
```

7. Our estimated bias is 0.0076 and our estimated variance is 0.0158.

```
#reference: bootstrap_mean.R in course files
set.seed(13)
np_flock = map_df(1:1000, function(i) {
# sample from the data WITH replacement
b_flock = sample(junco21$max_individuals, length(junco21$max_individuals), replace = TRUE)
# compute the sample mean of the bootstrap sample
data.frame(b_m = mean(b_flock))
})
#plot histogram of values
ggplot(np_flock, aes(x=b_m)) + geom_histogram(colour="#A03CAE", fill="#A03CAE", bins = 40) +
  xlab("Estimated Average Number of Dark Eyed Juncoes in a Flock") +
  ylab("Count") +
  labs(title = "NonParametric Bootstrap for Estimated Average Number in a Flock")
```




```
#bias calc
3.12 - mean(np_flock$b_m)
```

```
## [1] 0.007578171
```

```
#variance calc
var(np_flock$b_m)
```

```
## [1] 0.01582218
```

Part 5: Become a citizen scientist!

In this part, I noted the species, number seen at one time, location, time of day, and what bird was doing.

1. Wild Turkeys, 5, Stern Hall, sunset, walking through the brush in a group.
2. American Crow, 1, Stern Hall, sunset, cawing on a lightpost.
3. Rock Pigeon, 2, Stern Hall, a little before sunset, sifting through trash around the trashcans.
4. Dark-eyed Junco (YES! In the flesh!!), 1, Stern Hall, a little before sunset, picking up seeds on the ground.