Finding Bigfoot: A Statistical Approach

Andrew Blelloch, Adidev Jhujunwala, Theodore Johnson

Abstract

People have been on the lookout for Bigfoot as far back as the late 1800s, and a good many claim to have laid eyes upon the forbidden beast. Blurry photos and anecdotal accounts of the creature have helped fuel the search for decades as thousands of official reported sightings have been made across the country.

Data Science has rarely been considered amongst the believers (and possibly non-believers) of Bigfoot, but our group was curious to see what a more scientific approach could reveal. Let us start off with this - proving the existence or nonexistence of Bigfoot is not the primary concern. Myth or fact, Bigfoot sightings still persist, year after year, and it is not our interest to disprove anecdotal evidence. More importantly, we were curious to see what factors were correlated with bigfoot sightings at the county level, which could give us a more accurate picture of the characteristics of a region with high levels of sightings.

Our final objective is to create a model that can somewhat accurately predict the number of sightings in a county over any given 4-year period. To those who are trying to find bigfoot, maybe this model could help narrow down the counties to search. To those who do not believe in bigfoot, may this model serve as an indicator of counties with many believers. And to those who *fear* Bigfoot, this model could be extremely useful in avoiding the beast.

Background

The legend of Bigfoot has its roots in rural, Pacific Northwestern towns if the 19th century. However, the modern popularization of Bigfoot began in the mid-20th century, gaining widespread attention in the 1950s and 1960s. The term “Bigfoot” itself emerged in 1958 following the discovery of large, mysterious footprints in Northern California.

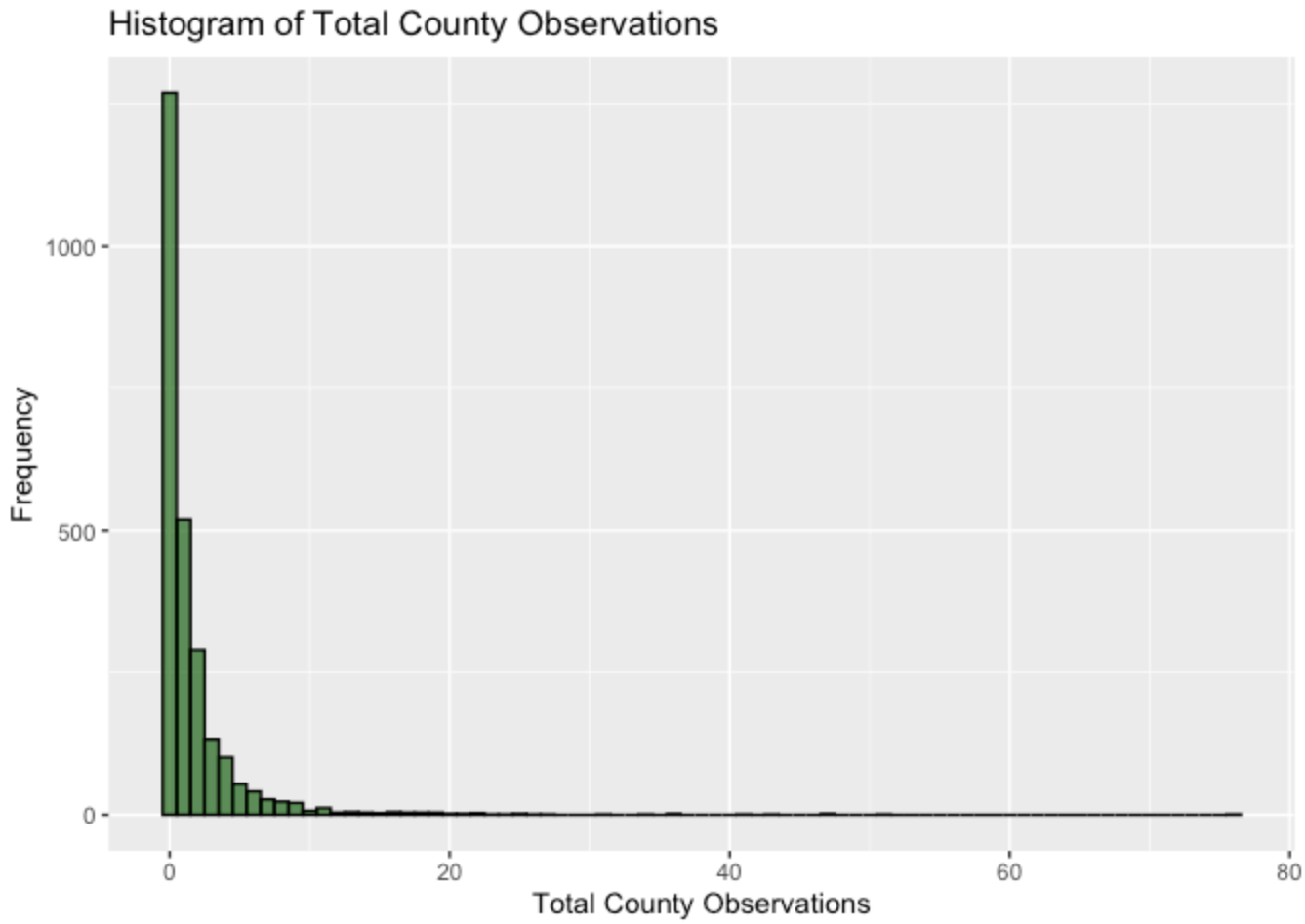
Numerous alleged sightings of Bigfoot have been reported across the United States and Canada, with additional anecdotal accounts coming from other parts of the world. The descriptions of Bigfoot vary, but common characteristics include a tall, hairy creature resembling an ape or human.

The search for evidence of Bigfoot has led to the collection of footprint casts, blurry photographs, and occasional videos purported to depict the creature. However, skepticism surrounds the validity of much of this evidence, and scientific scrutiny has been challenging due to the lack of concrete, verifiable proof.

Data

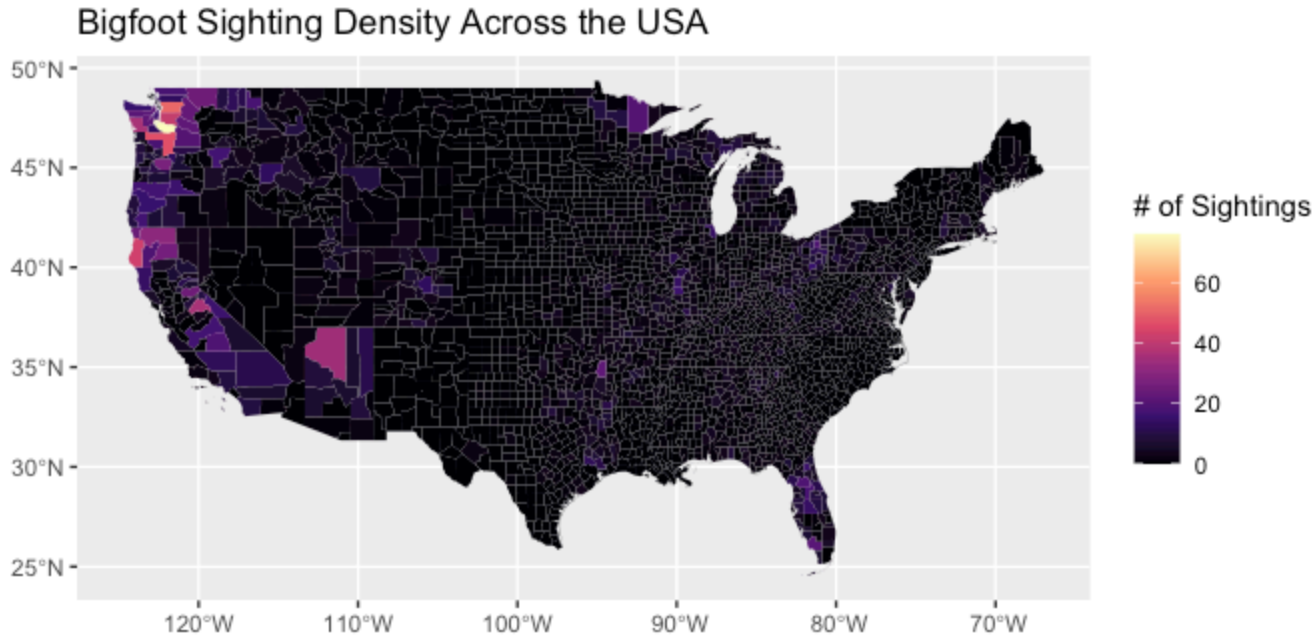
Our data for the reports came from the BRFO (https://www.bfro.net/) database, which contained a csv of all their collected reports stretching from the early 1900s up to 2023. For simplicity, we cut this data to just observations from 2008-2012. For the variables we chose to look at, it wasn’t quite possible to pair each county with the data from any year between 1900 and 2023. It made sense to stick to a range where we could find plenty of reliable data, mostly from surveys conducted in 2010.

The data revealed a pretty skewed range of sightings within each county. About half of all counties had 1 or fewer sightings, and many of those counties don’t claim to have seen Bigfoot at all.

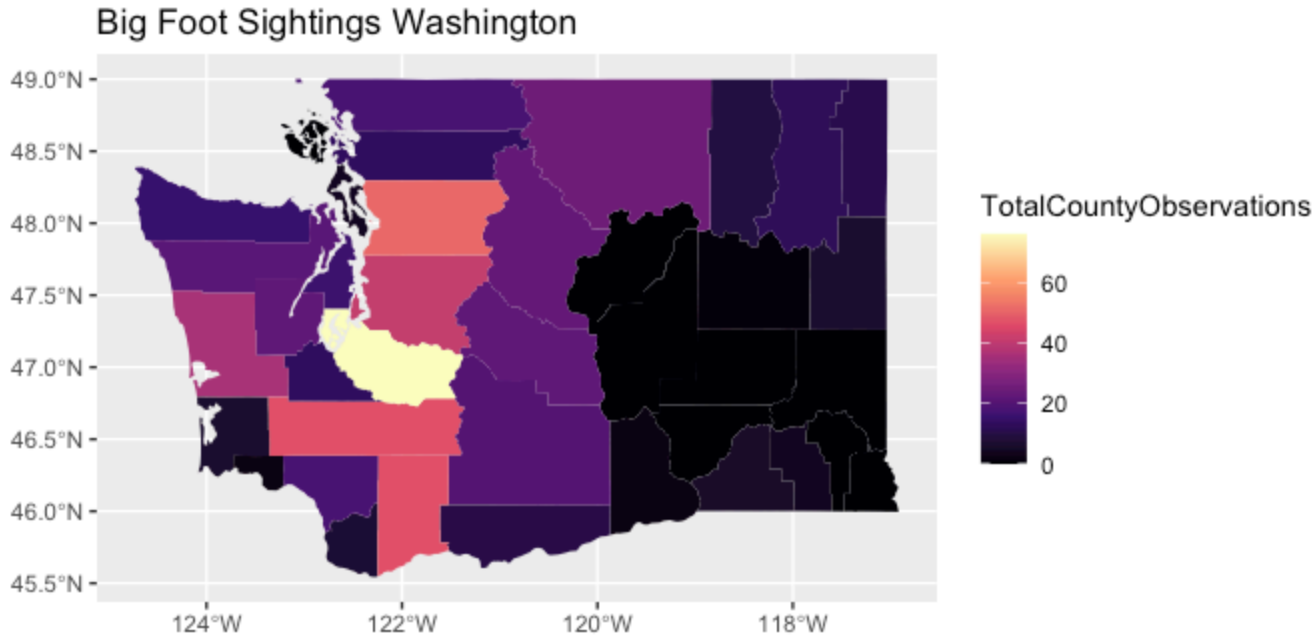


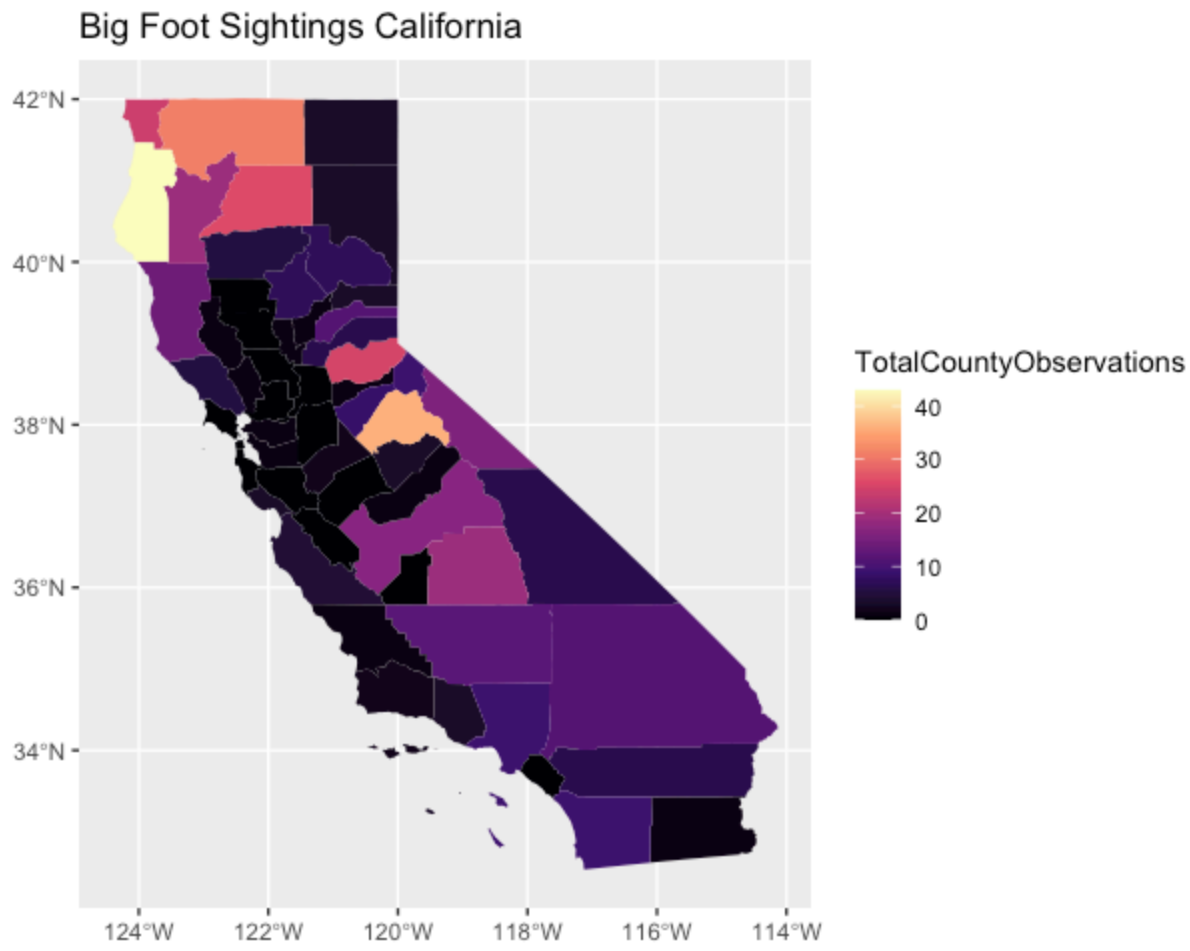

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   1.000   1.735   2.000  76.000
```

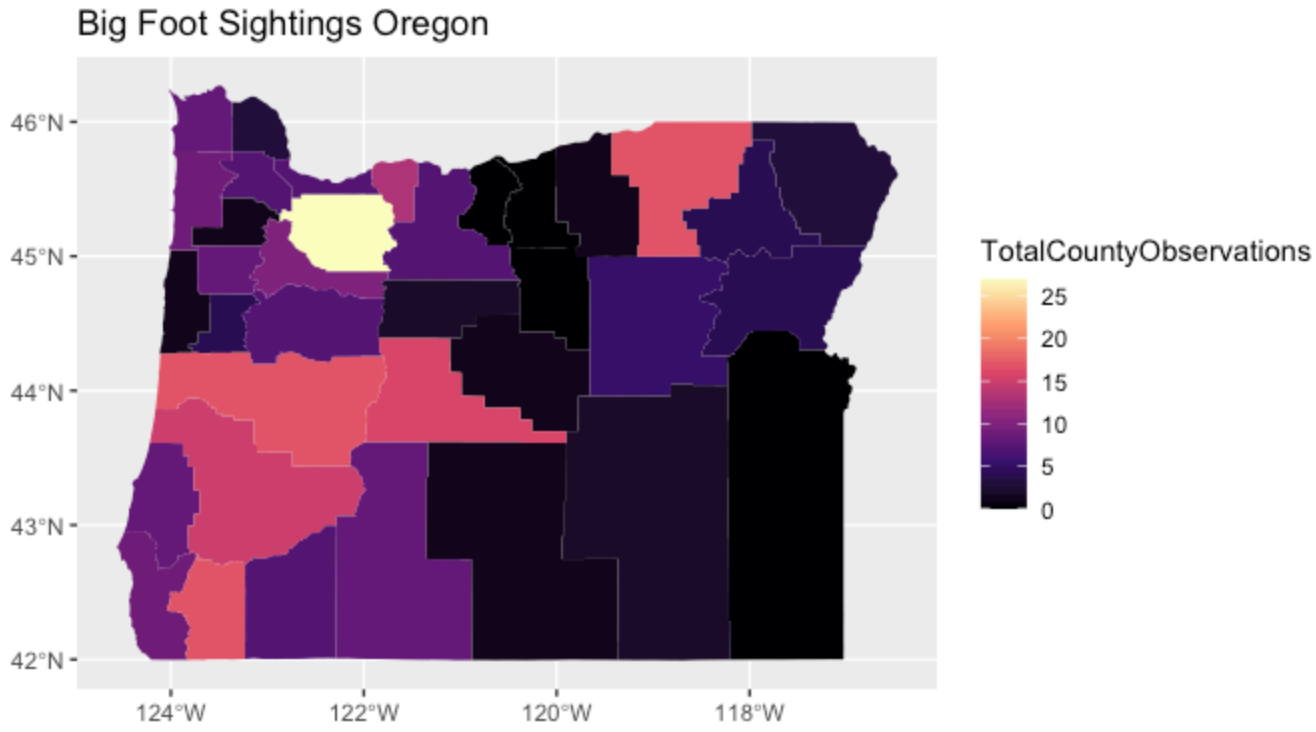
This posed an obvious issue for the accuracy of our model which we will discuss more in detail later on. However, it was far from surprising. We knew after looking at the data that specific regions, mainly those in the pacific northwest were the most popular. After all, that is where the myth of Bigfoot originated. Not to mention the heavily wooded areas of the pacific Northwest, to our understanding, are considered to be the “natural habitat” of the sasquatch. Here, we visualized the density of sightings on a national level:



The vast majority of sightings are concentrated in Washington, Oregon, and California, with small patches of sightings across certain regions of the Midwest and the South. Here's a closer look at those three states:


Big Foot Sightings Washington
49.0°N
48.5°N
48.0°N
47.5°N
47.0°N
46.5°N
46.0°N
45.5°N
124°W
122°W
120°W
118°W
TotalCountyObservations
60
40
20
0


Big Foot Sightings California
42°N
40°N
38°N
36°N
34°N
124°W
122°W
120°W
118°W
116°W
114°W
TotalCountyObservations
40
30
20
10
0



Variables and Data Collection

Still, there is some variation across counties that isn't totally explained by location. Why do certain counties in Washington have higher reports than others? Why does Florida have more observations than any other state in the South? Our goal was to choose variables that might pick up the causes this variation from county to county.

First, we were interested in economic variables that may affect sightings at the county level. We found that **median income** and **unemployment** were both statistically significant at the county level in predicting bigfoot sightings. Demographic variables like **race**, **county population**, and **poverty percent** were found to be statistically significant as well. Most of this data was harnessed from the 2010 census.

Out of pure curiosity there were a couple variable we hypothesized would be significant in predicting sightings that we found in a county level health survey across the US. The first was the **% Excessive Drinking**, that is, the percentage of people who admitted to binge drinking in a given county. **% Completed High School** was also a variable we were interested in including, as we hypothesized that maybe counties with lower levels of education would be more interested in pursuing and reporting sightings of a mythical creature.

The last one we pulled from the health data set was a bit of a long shot, but we started to notice in the descriptions for a lot of the reports that they occurred during a drive, off of a highway or a road. We wrote the following code to get a more exact idea of how many highway or road sightings there were in our 733 observations.

```
descriptions <- reports_08_to_12$`Location Details`
contains_highway <- grepl('Highway', descriptions, ignore.case = TRUE)
contains_road <- grepl('Road', descriptions, ignore.case = TRUE)
contains_highway_or_road <- contains_highway | contains_road
count_highway_or_road <- sum(contains_highway_or_road)
count_highway_or_road
```

```
## [1] 1373
```

```
count_highway_or_road/2554
```

```
## [1] 0.5375881
```

The variable **% Long Drive - Drives Alone** accounts for the percentage of people who have long drives home after work. The variable proved to be highly significant in predicting sightings, which wasn't surprising, given that more than 53% of our sightings occurred on highways.

Modeling

Poisson Model

We used Poisson modeling to predict the number of Bigfoot sightings in U.S. counties, by using various socio-economic factors as covariates. The choice of Poisson modeling was primarily due to its effectiveness in handling count data – specifically, the frequency of rare events in defined intervals, like Bigfoot sightings. By integrating socio-economic variables, such as population density, income levels, and educational attainment, we tried to find correlations that might influence the likelihood of reported sightings. The Poisson model's ability to deal with varying rates of occurrences – in this case, the sightings – across different counties, made it an ideal tool for this analysis. This approach not only allowed us to predict sighting frequencies but also offered insights into the socio-economic dynamics potentially associated with these reports.

While exploring the use of the Poisson model for our analysis, we encountered some limitations that needed addressing. We forsaw these issues as our model had a variance greater than the mean. Specifically, the model was not accurately predicting the number of zero counts - it underestimated the occurrence of zeros and overestimated the number of counties with Bigfoot sightings. To tackle this issue, we turned to both the Zero Inflated Poisson (ZIP) Model and the Negative Binomial (NB) Model. These approaches showed promising improvements: the ZIP model, in particular, was more effective in capturing the excess zeros in the data, and the NB model helped in reducing the standard deviation of our predictions, indicating a tighter fit to the observed data.

However, despite these enhancements, the models did not perfectly align with the actual data. They still fell short in predicting the exact number of zeros, although they notably outperformed the original Poisson model in this aspect. The ZIP model, with its focus on modeling the excess zeros, and the NB model, known for handling overdispersion in count data, together provided a more nuanced and accurate representation of our dataset than the standard Poisson model could achieve. Nonetheless, there remained a gap between our predictions and the true data, suggesting room for further refinement and exploration of more sophisticated modeling techniques or additional explanatory variables that might better capture the underlying patterns of Bigfoot sighting reports.

We compared the accuracy of the three different models by comparing their standard deviations - the closer to zero, supposedly the closer it is to the actual value. The most accurate model was the Zero Inflated Poisson, with a standard deviation of 1.67.

```
FullCounty2$difference_poisson <- abs(FullCounty2$predictions_poisson - FullCounty2$Tota
lCountyObservations)
mean(FullCounty2$difference_poisson)
```

```
## [1] 1.768207
```

```
FullCounty2$difference_Zinb <- abs(FullCounty2$predictions_Zinb - FullCounty2$TotalCount
yObservations)
mean(FullCounty2$difference_Zinb)
```

```
## [1] 1.927565
```

```
FullCounty2$difference_Zip <- abs(FullCounty2$predictions_Zip - FullCounty2$TotalCountyO
bservations)
mean(FullCounty2$difference_Zip)
```

```
## [1] 1.668755
```

Logit

We wanted to test our model in terms of the conventional metrics that we've been using in this class. Further, we wanted to adjust for outliers and to see if our data was actually significant. We decided that the simplest and probably the best way of doing this was to use a Logit model and then computing the accuracy, precision and recall statistics. We specifically thought that the calculated precision and recall statistic would be important to ensure that we weren't over or under predicting zeros and positive cases.

```
## # A tibble: 1 × 3
##   accuracy precision recall
##      <dbl>     <dbl>  <dbl>
## 1    0.661     0.650  0.706
```

Our data was as outlined above, each of our three metrics were roughly in the 65-70% range which we were happy with as we are trying to predict a phenomenon as random as a predicted bigfoot occurrence.

Analysis and Conclusion

Unsurprisingly, Bigfoot sightings are hard to predict. It made pretty clear sense that most sightings were clustered in the Pacific Northwest (where the myth first originated), but outside of that, it's still hard to say what creates the variation from county to county reports. If we had to say which model creates the best predictions, it's a close run between the logit and the zero inflated poisson.

With more time and access to better data, we'd want to look at more data that could pick up more of the variation in observations between counties. Specifically, our group discussed finding more data on geographic characteristics that could raise the number of sightings, such as forest land cover estimates or the number of animals that could possibly resemble Bigfoot. In short, there is still plenty of interesting information to uncover in the prediction of Bigfoot sightings.

Finally, we've concluded that it's incredibly difficult to apply statistics to myth-based observations. It's important to understand that our model doesn't actually predict where Bigfoot will be, but what type of characteristics exist within a county that lead to more *sightings*. What causes a sighting of Bigfoot might need a much deeper psychological evaluation of human beings that can't be tracked with county level statistics. Perhaps, mental health statistics could serve us well. Or, we should just accept that predicting Bigfoot sightings may be elusive as finding Bigfoot himself.