Project 2

Austin Dollar

11/15/2021

Introduction and Domain

The data set I chose for this project is the deer Harvest Statistics for the State of Idaho, with data able to be obtained for the 2001-2020 deer seasons. This data keeps track of each individual Zone, or area of the state, called units, the amount of hunters hunting in any given year, the amount of successful harvests, among other data statistics for hunting such as harvest animal size. In terms of past studies, there has been studies on the deer hers themselves, but there has been little in regards to the context of hunting, which is important. It is important as these statistics can help us identify the overall trends of human and hunter impact on the deer population of Idaho can be gauged and predicted within the contexts of hunting, showing us things such as a trend in the amount of hunters, the amount of harvests, the hunt success rate, among other pieces of information that can help us predict trends in hunting overall, and how the quality and amount of deer harvested are affected over time. I am interested in this as I am a passionate hunter here in California, however, the data regarding California is only available in a PDF format, so I was forced to use data from Idaho, which was offered in a csv format.

Data Set

We will be using the Idaho data set available at

https://idfg.idaho.gov/ifwis/huntplanner/stats/.

The Idaho Data sets are broken up by year, in this case, we are utilizing the years 2015- 2020, which are to to be merged below in the "Tidy Data" section. The variables, as read in, however, are included below:

- Take.Method - The method use to take/harvest the deer; categorical variable of the following categories, which are sorted by their respective seasons: Any Weapon = General Season, can use any method of take, Archery = Archery Weapon only season, Muzzleloader = Muzzleloader only season, All Weapons Combined = Combine all seasons for a total.

- Unit - A categorical variable that describes the "Unit" in which the hunter was hunting and/or the deer was harvested. A Map of respective units is included in the following link:

https://fishandgame.idaho.gov/ifwis/portal/sites/ifwis/files/maps/2014_RegionsAndGMUs_8.5x11.png

- Harvest - Numerical Variable that gives the number of harvested deer in the observation.
- Hunters - Numerical variable that gives the amount of hunters in an observation.
- Success% - Numerical variable for the percent harvest success for the observation.
- Days - Numerical variable for total days hunted in the observation.
- Antlered - Numerical variable for antlered deer harvested in the observation.
- Antlerless - Numerical variable for antlerless deer harvested in the observation.

• %4+Pts - Numerical Variable for the percentage of harvested deer with greater than or equal to 4 points on the antlers.

• %5+Pts - Numerical Variable for the percentage of harvested deer with greater than or equal to 5 points on the antlers.

• %Whitetail - Numerical Variable for the percentage of harvested deer that are Whitetail (rather than Mule deer)

Tidying data

Due to the fact that these Data sets are in different files, we have to do some merging and tidying. I merged all the data sets for the years 2015-2020 by adding a variable called "year" to the data set before the merge. Additionally, due to the difficulty in sorting out the 60 plus Units of the state, I divided it into larger regions, which is another variable I added. A link to the image used to map these units to regions is included in the link below:

https://fishandgame.idaho.gov/ifwis/portal/sites/ifwis/files/maps/2014_RegionsAndGMUs_8.5x11.png

These new variables, added to the new Idaho dataset are included below:

• Year - Numerical Variable for the year the data was recorded

• Region - Categorical variable for region to consolidate units

Goal and Data Science Question

We will be exploring the Idaho data set in an effort to better understand the overall trends in different areas of hunting. This data comes directly from the department of Fish and Game in the state of Idaho, which can be obtained at he following link:

https://idfg.idaho.gov/ifwis/huntplanner/stats/

The link above links to the statistics for the 2020 season, the data for the years 2001-2020 can be seen and downloaded via clicking the "Deer" tab on the website, with the datasets I am using being listed under "General Season." This data seems to be reliable, coming from a government source, with its only possible source of fault, is that it does not account for poachers, as that would be an unreported statistic, as it is illegal. So while my results will ring true for legal hunters, it does not account for criminal activity.

I combined the data for the years 2015 to 2020 to one larger data set, to help better model trends over time, and if this five year time span is not enough to obtain a reliable model, I will continually add years until an accurate model can be obtained.

The three main data science questions that seem the most interesting and that can be addressed with this data are as follows:

1. Does "Average days hunted" contribute to and/or correlate with, hunter success, and what does it tell us?
2. Can we predict the trend (and is there a trend) in overall harvested deer antler size over time? IE; Is the percentage of pts greater than 4 and greater than 5 trending in an obvious direction?
3. How does the total amount of deer harvested trend over time, and what are the contributing factors of this result?

The main question that I will focus on for this project, and attempt to model is the second question, which, again, is as follows:

Can we predict the trend (and is there a trend) in overall harvested deer antler size over time? IE; Is the percentage of pts greater than 4 and greater than 5 trending in an obvious direction?

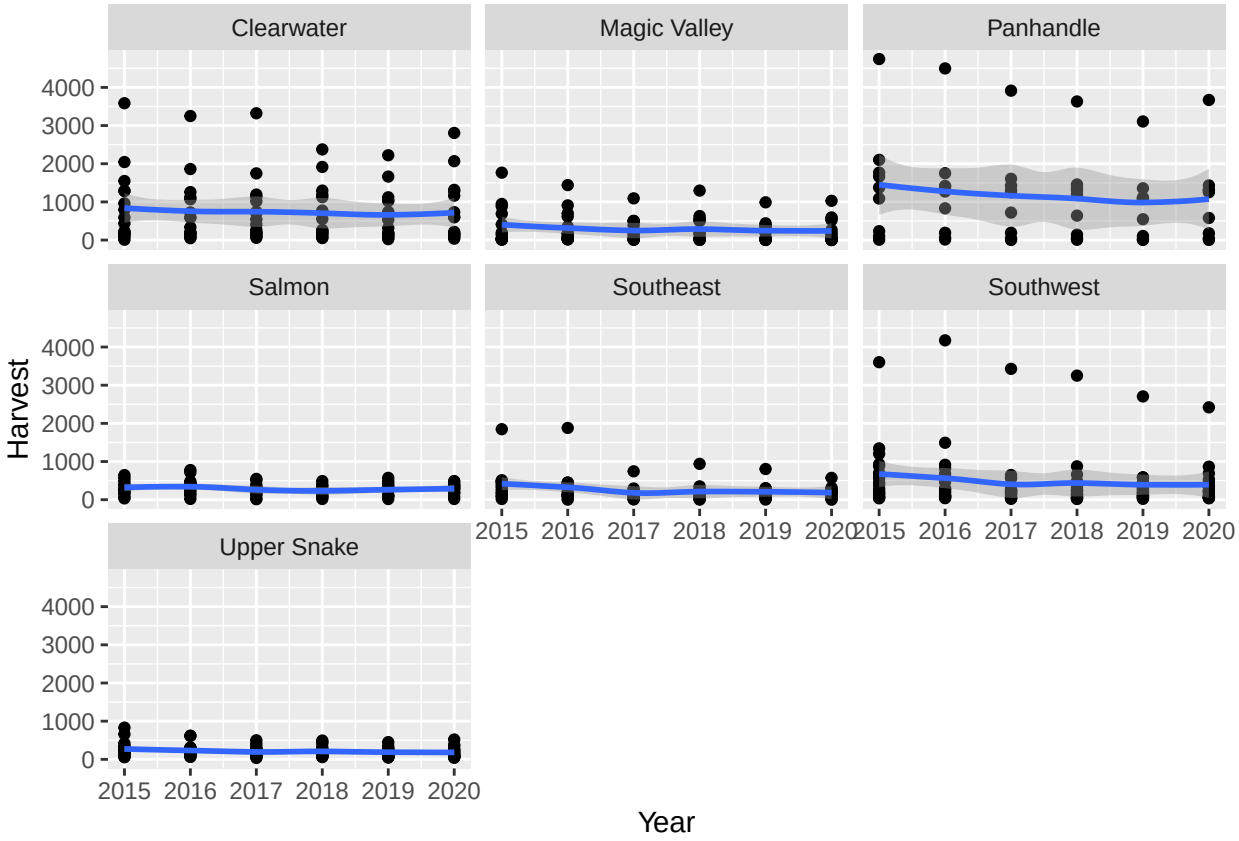
This question is going to be my focus, as I believe that in answering it, we must also answer question three, as it has relevant data in this situation. Overall, this model's parameters should make sense, showing a trend in deer size over time. This is because the metrics of percent greater than 4 pts, and percent greater than 5 points are a metric that represents overall deer health. As healthy, mature deer will have more points on their antlers, therefore measuring this metric is a decent way to measure overall health.The one point of confusion, however may be what that percentage means, which is why solving for trends in total population will be a good supplement in an attempt to better explain the results. Possible limitations, as stated before, could be the amount of data we have does not have enough information for an accurate model. However, this can be remedied by obtaining and merging more years into our data set.

Visual Representation of the relevant data (pre model)

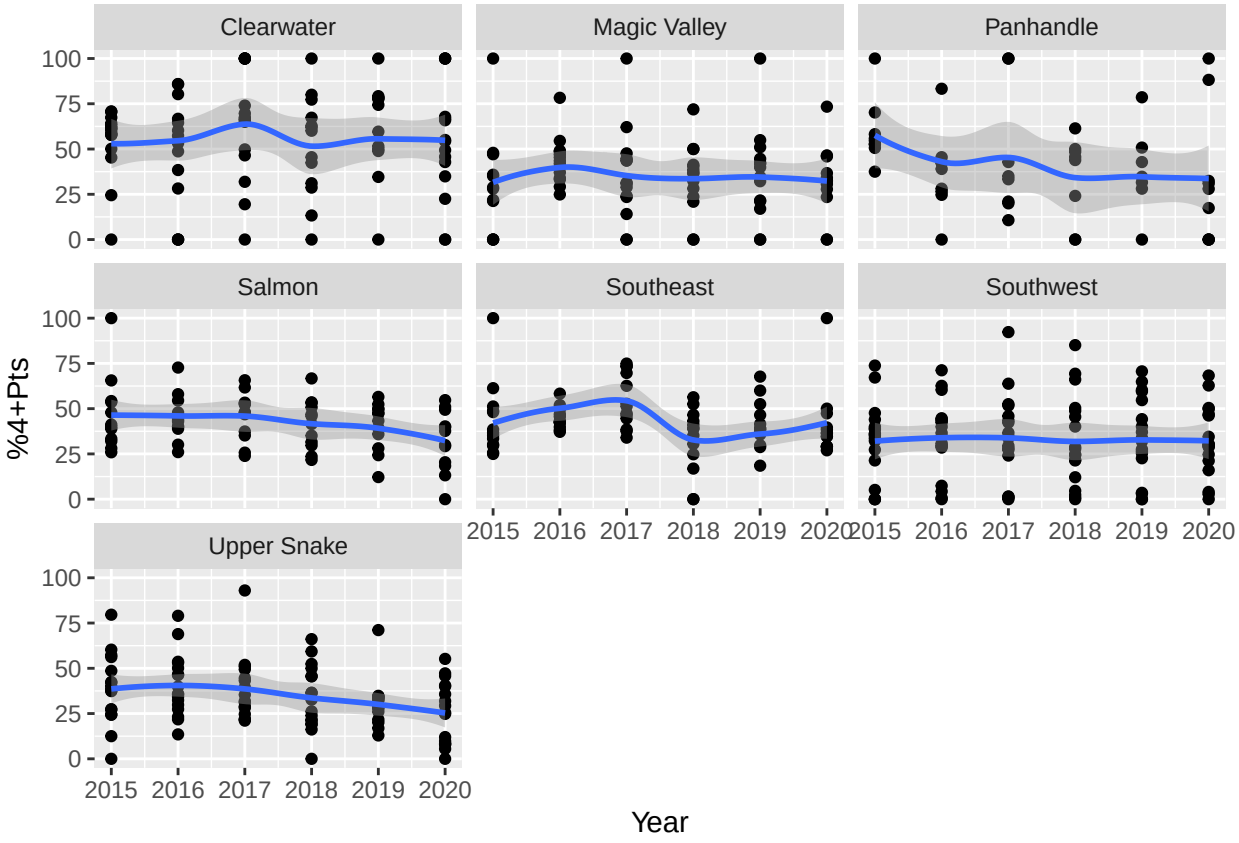
For our data science question, there are three main graphs that we would be interested in looking at, which is the trend in total deer harvests over time, as well as the trend in the percentage of those deer that have greater than 4 points on their antlers, and greater than 5 points respectively. Additionally, the graphs will be broken up by region, which will tell us if trends are localized to certain areas, or spread throughout the entire state. These graphs are included below, in order to give a visual representation of the data to be modeled.

Before models are generated, here is a visual representation of what the key data relationships in this data set look like. We have graphs representing the trend in the total harvests over time as a baseline to then look at the percentage of those totals that have 4 antler points or greater, then likewise, the same for 5 or more antler points:

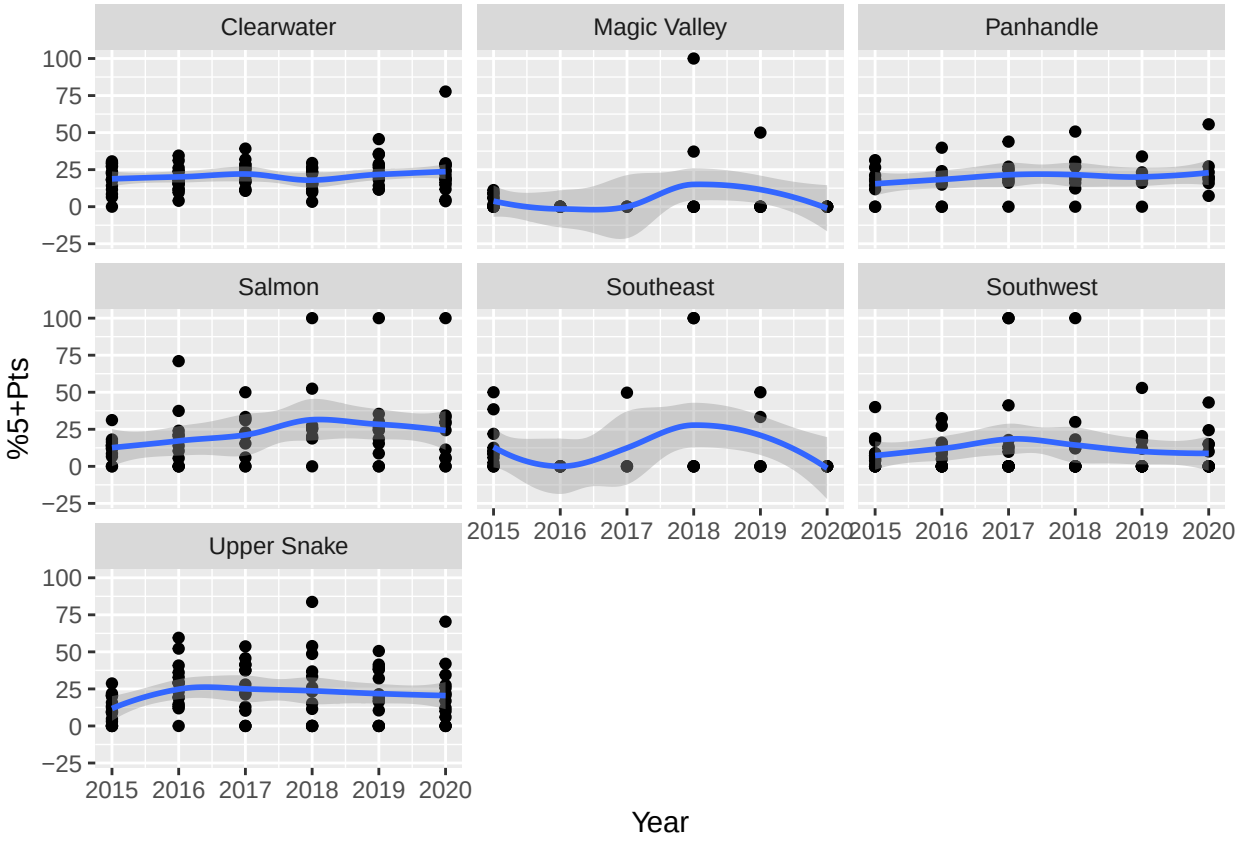
Graph of total harvests over time, sorted by region



Graph of percent greater than 4 points (%4+Pts) over time, sorted by region



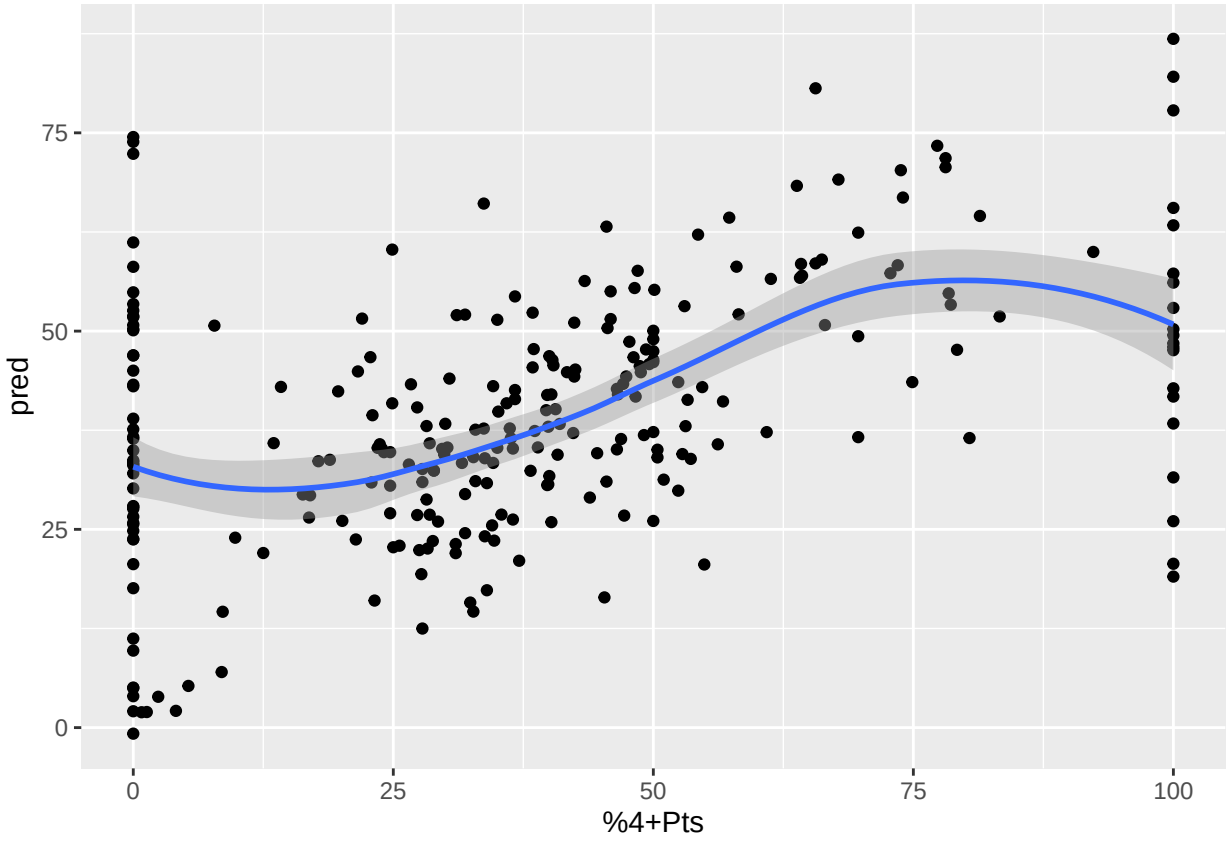
Graph of percent greater than 5 points (%5+Pts) over time, sorted by region



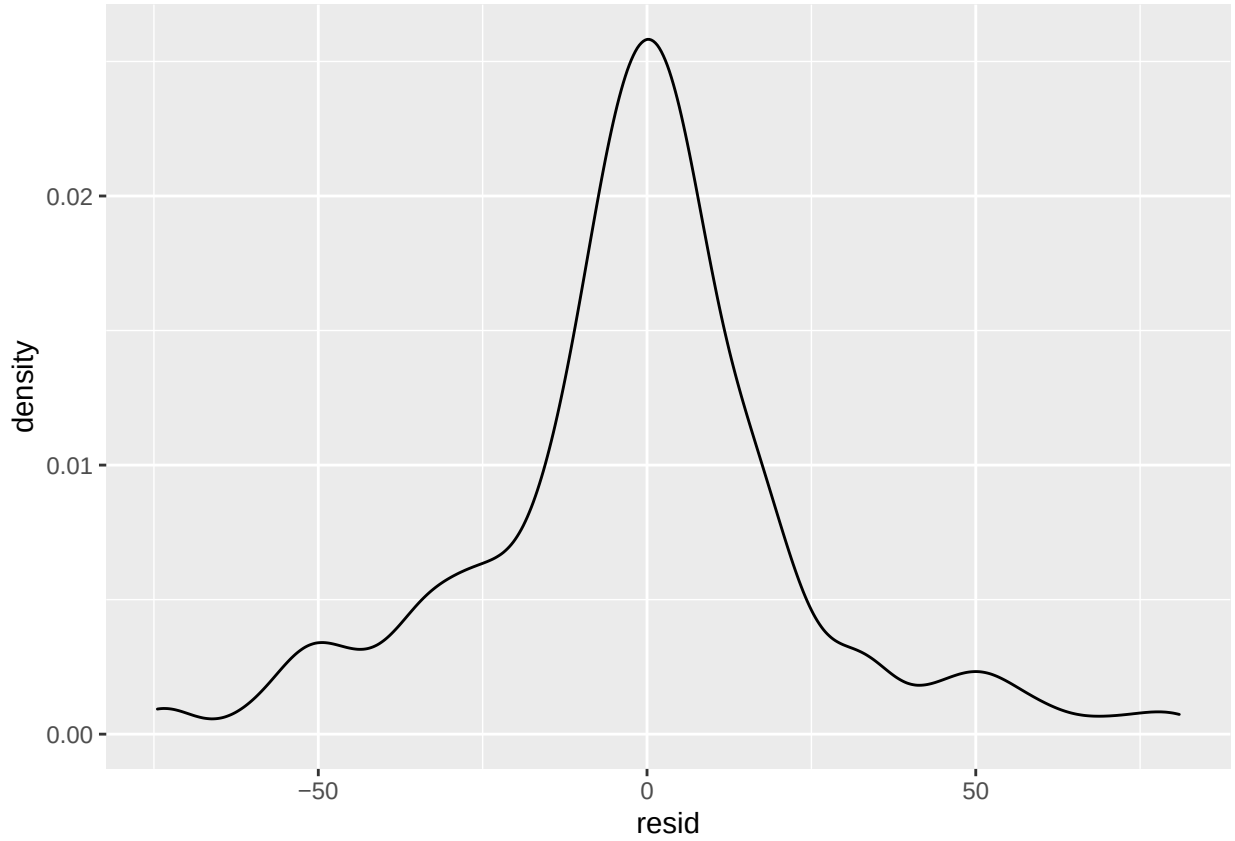
Model Generation and conclusions

To generate the model, I will generate the model on the entire deer population, rather than breaking it into regions, as that will be too many models. Creating one general model for the whole state in general will map data well enough to generally fit and predict individual regions as well. Below, I will generate models of the two main points of prediction, %4+Pts and %5+Pts, using an 80% training set, and a 20% testing set. Both of these metrics will give us an idea of the overall health of the Idaho deer herds, as the amount of antlers is largely an indicator of health. While there is much more nuance in reality, for the sake of this project, we can say that the percentage of harvested deer that have 4 points or greater is an indicator of the percentage of the deer herd that can be deemed at least "healthy". Likewise, the statistic that measures a greater hurdle, the percentage of deer with 5 or greater points, correlates to the percentage of deer that can be deemed "very healthy."

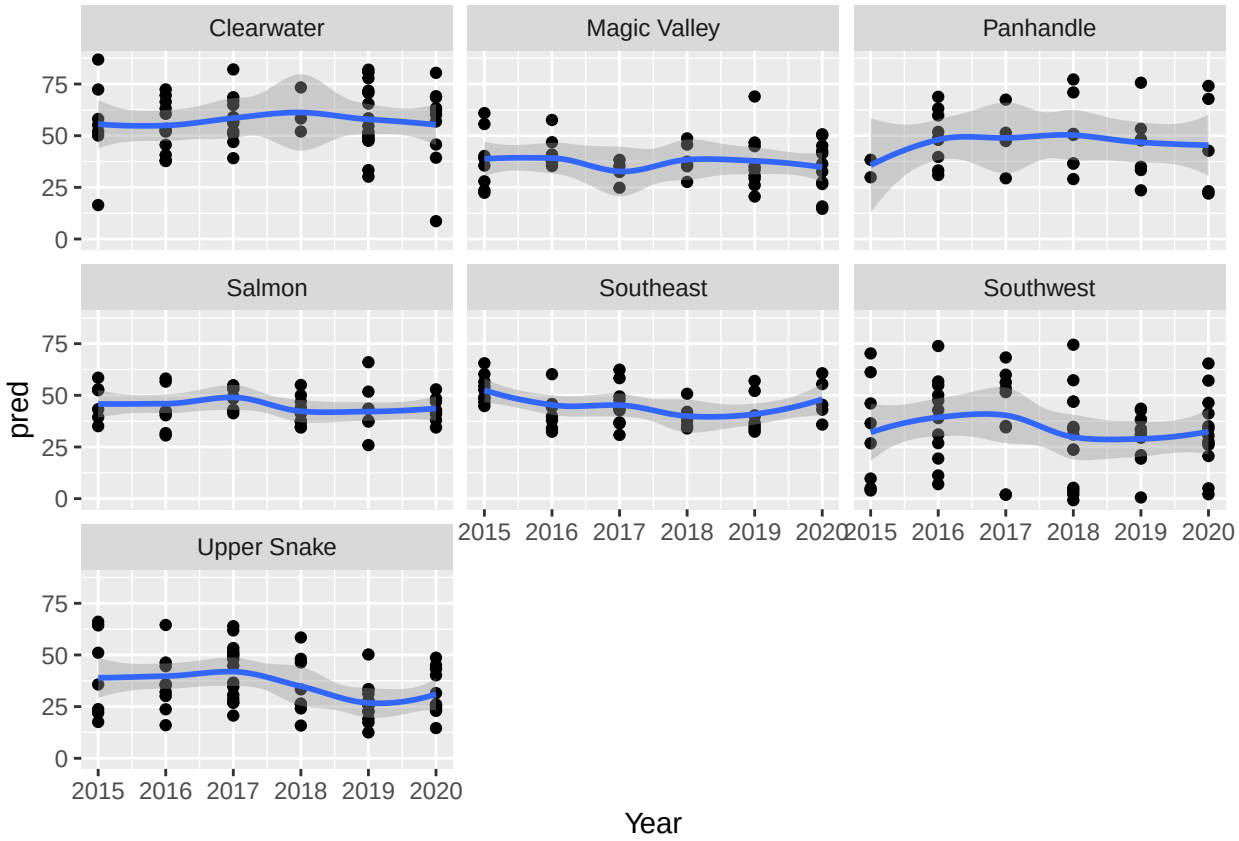
Model generation of the %4+Pts variable

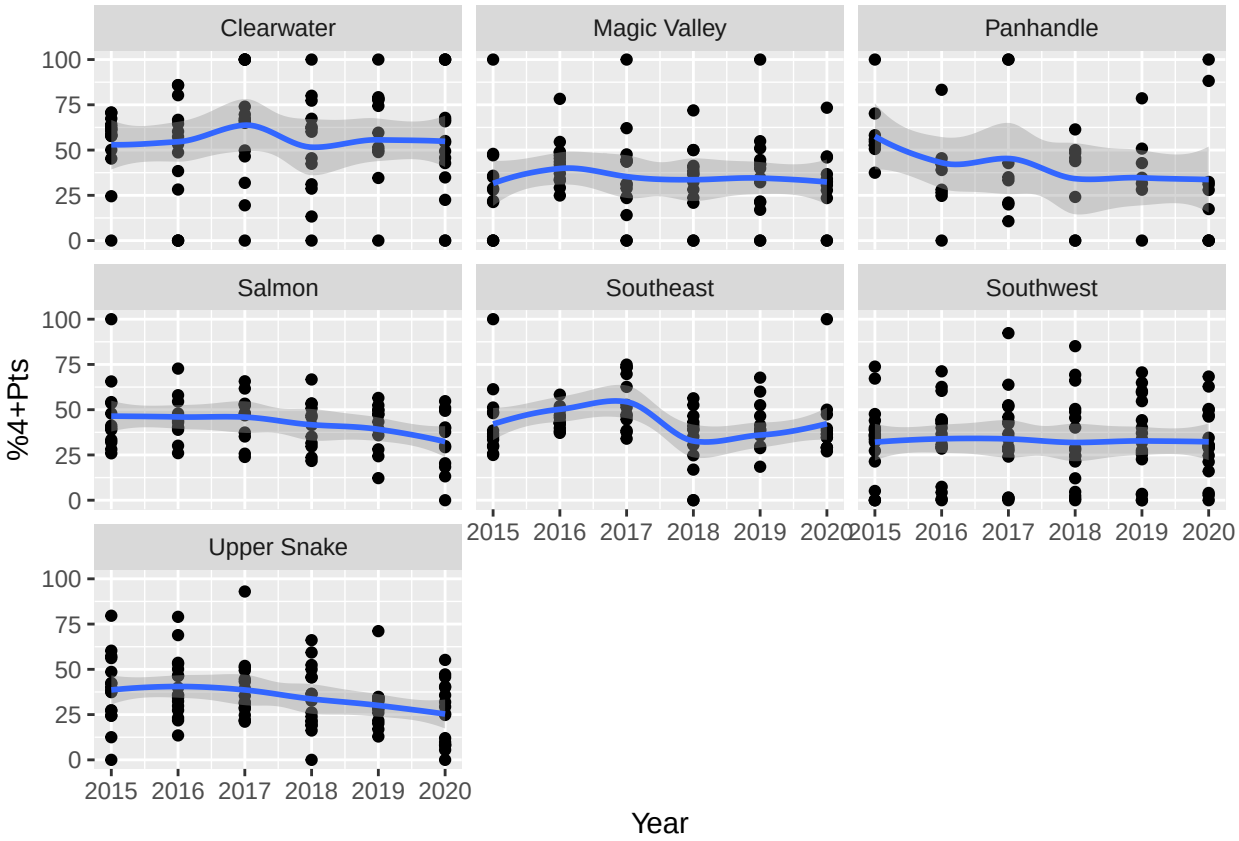


Above, we have a plot of the true values of the test set on the x axis, with the correlated predicted value on the y axis. From looking at the graph, we can determine that it is pretty accurate, as the general graphical trend is 1:1 linear, with a slight curve to it, meaning that the predicted value and true value are, generally speaking, the same. We can see that there is a lot of points at 0 and 100%, which is as expected, as these are percentages. Next, we will add residuals and plot them to ensure our model is as accurate as possible.



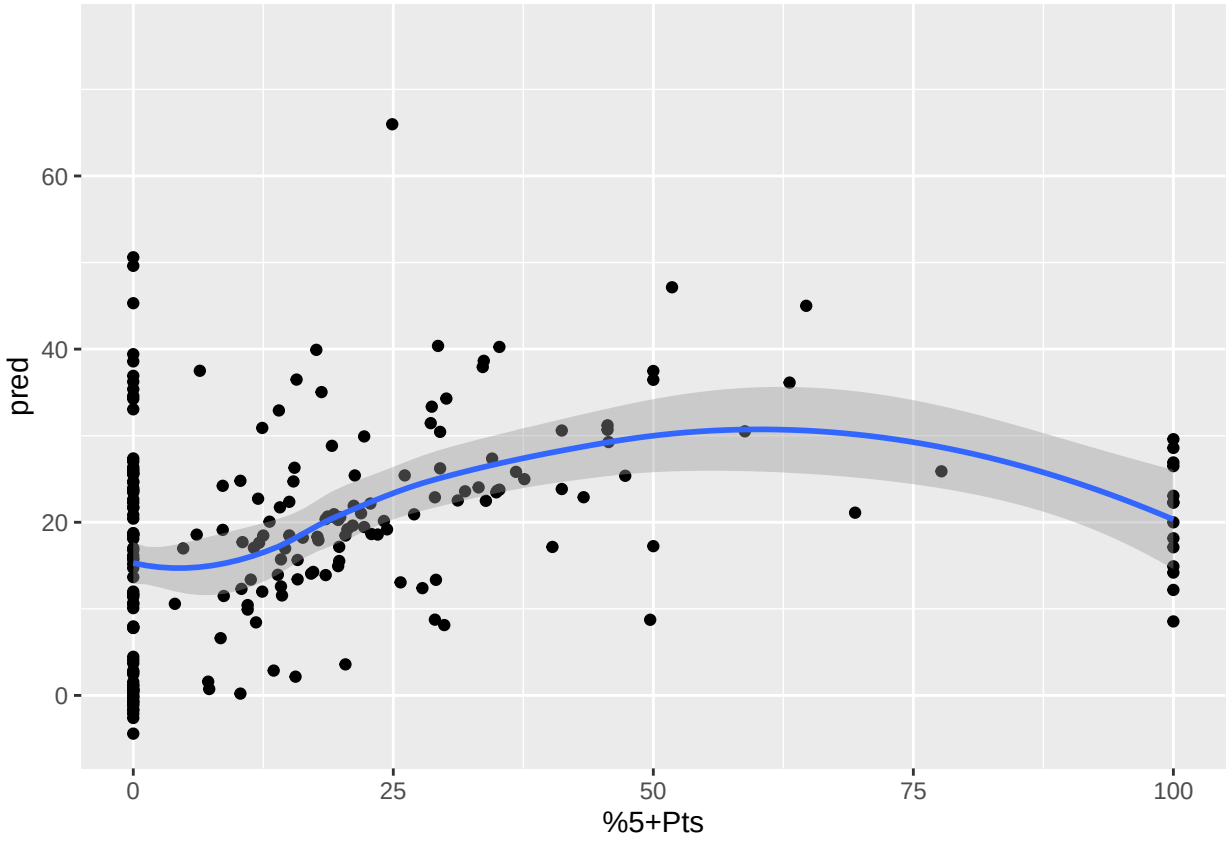
Our graphical representation of residuals, shown above, is as good as I could get it. However, my density is really good, as it is normally distributed, centered around zero. This gives us really good results, even though the model became more complicated than anticipated, as it requires more data than just the year. In order to accurately predict the amount of deer that have 4 or greater points, we will not only need the year, but the Region, amount of Hunters, Take Method, and Unit as well. These values are actually changeable by the government, which controls the amount of hunting permits allotted for specific regions. Using this data, one could find the ideal amount of permits to give out in different regions to best maintain a healthy deer population. We can look at the predictions of the test set as a representation of a trend over time, shown below:


Clearwater
Magic Valley
Panhandle
Salmon
Southeast
Southwest
Upper Snake
pred
0
25
50
75
2015
2016
2017
2018
2019
2020
Year

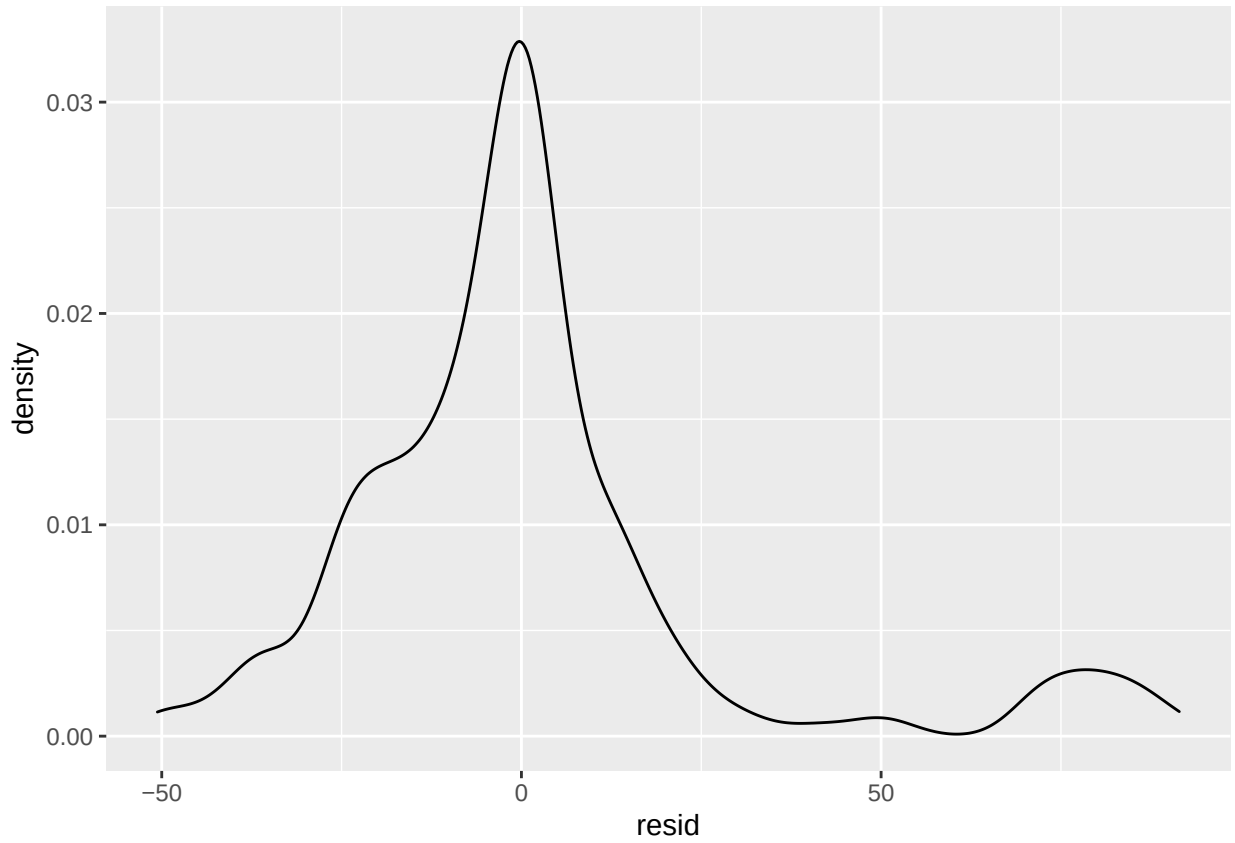


When observing these graphs, sorted by region, the trends are overall pretty much the same as the graph of the test data set, with some minimal differences. You can tell this by comparing the two above graphs, with the graph of predictions above the graph of the entire data set.The key is that the overall trend is the same, which is exactly what we want, as it says that the predictive model is accurate to the test data that we have. In the next section, we will use the same methods and variables to generate and test a model for the greater barrier, deer that have 5 or greater points, that can be deemed "very healthy."

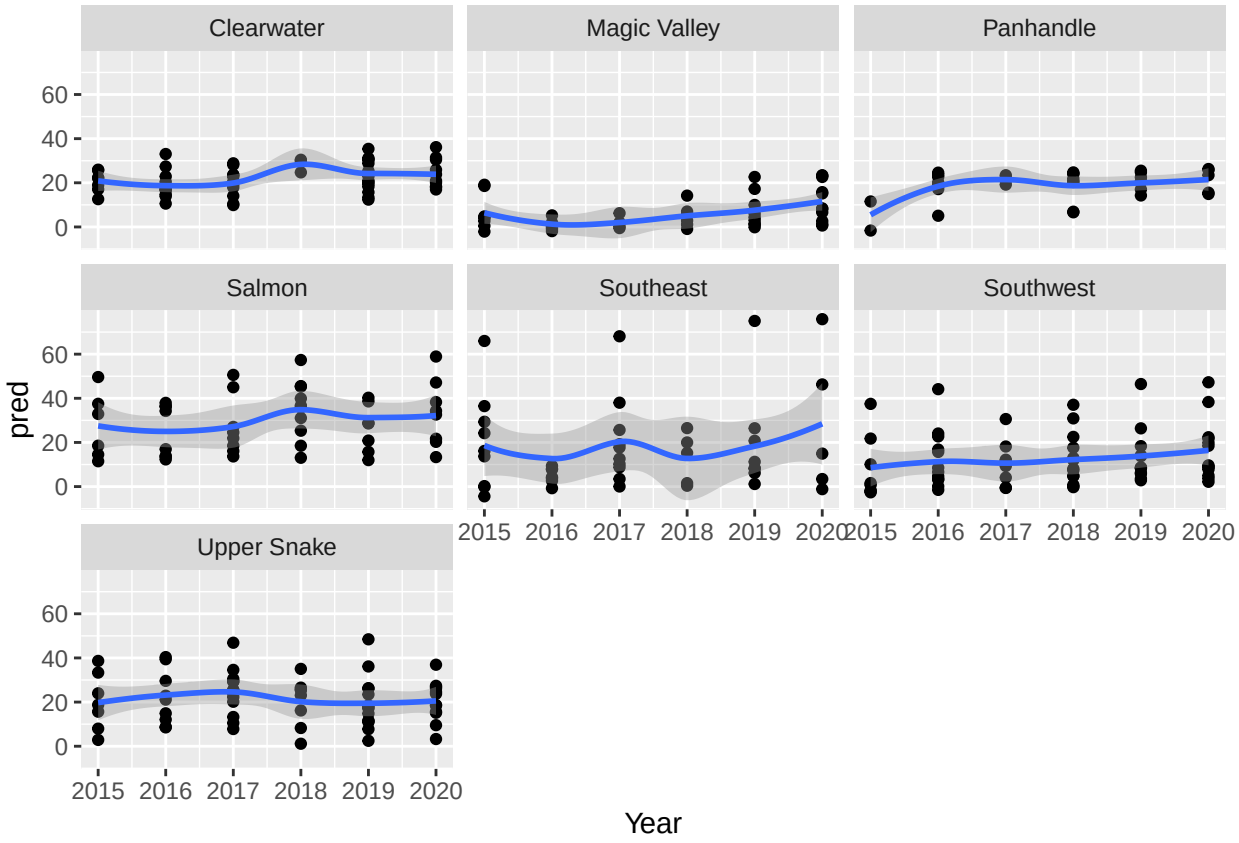
Model generation of the %5+Pts variable

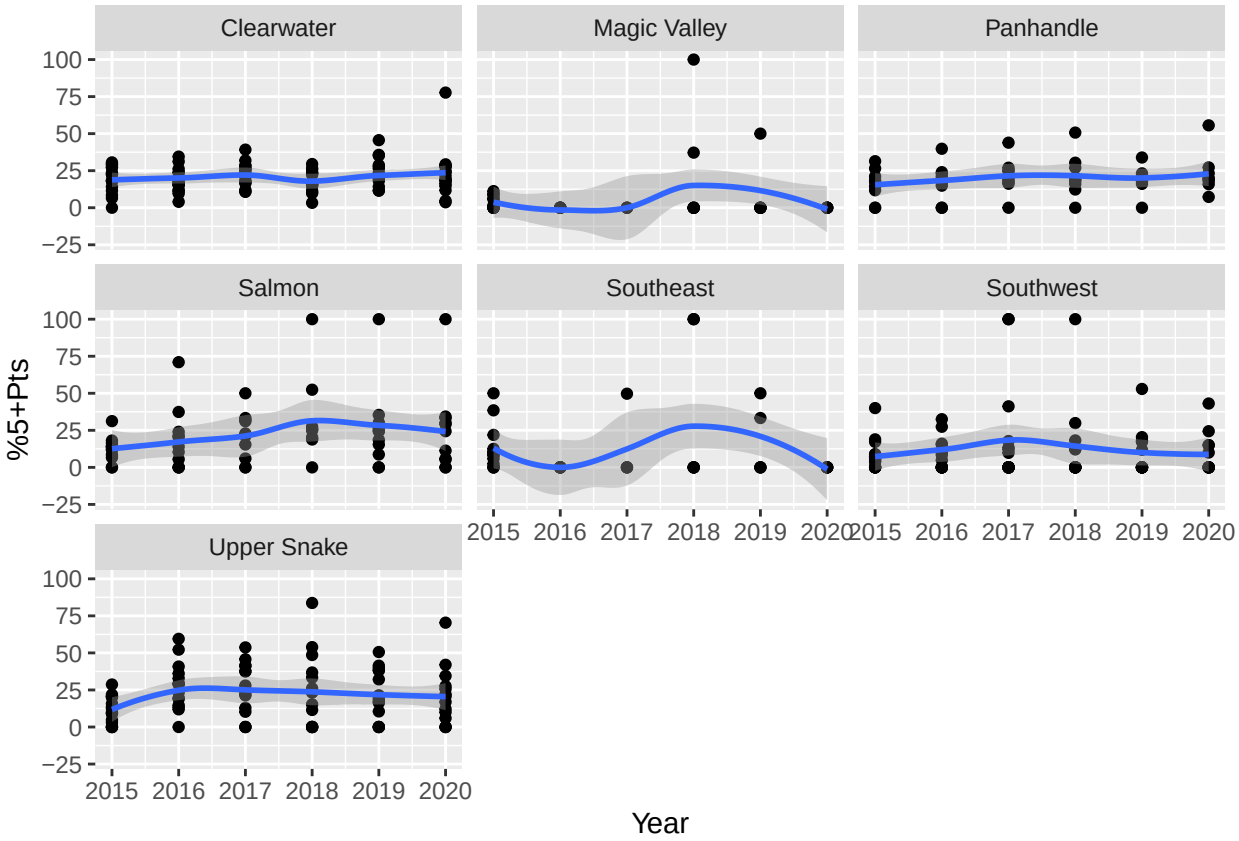


Above, we have a plot of the true values of the test set on the x axis, with the correlated predicted value on the y axis. From looking at the graph, we can determine that it is pretty accurate, as the general graphical trend is 1:1 linear, with a slight curve to it, meaning that the predicted value and true value are, generally speaking, the same. We can see that there is a lot of points at 0 and 100%, which is as expected, as these are percentages. Next, we will add residuals and plot them to ensure our model is as accurate as possible.



Our graphical representation of residuals, shown above, is as good as I could get it. The graph is relatively normally distributed, but has some skew due to the amount of values concentrated close to zero. The model became more complicated than anticipated, however, as it requires more data than just the year. In order to accurately predict the amount of deer that have 5 or greater points, we will not only need the year, but the Region, amount of Hunters, Take Method, and Unit as well. These values are actually changeable by the government, which controls the amount of hunting permits allotted for specific regions. Using this data, one could find the ideal amount of permits to give out in different regions to best maintain a healthy deer population. We can look at the predictions of the test set as a representation of a trend over time, shown below:


Clearwater
Magic Valley
Panhandle
Salmon
Southeast
Southwest
Upper Snake
pred
Year
0
20
40
60
2015
2016
2017
2018
2019
2020



When observing these graphs, sorted by region, the trends are overall pretty much the same as the graph of the test data set, with some minimal differences. You can tell this by comparing the two above graphs, with the graph of predictions above the graph of the entire data set. The key is that the overall trend is the same, which is exactly what we want, as it says that the predictive model is accurate to the test data that we have. Overall, both the 4 Point and 5 Point Models are pretty accurate, and these models can be used to predict the health of the deer population for future years.

Social and Ethical Ramifications

This data and predictive model has little Social Ramifications due to its scientific nature, however, it has great Ethical and even Environmental Ramifications. By having a predictive model that can anticipate the health of the overall deer herd, depending on the year and amount of hunters, as well as the method they use for hunting, both in total and per region, we can actually help keep the deer herds healthy. The government agencies that grant these permits can use the model to anticipate the amount of permits that they should allow in order to maintain a healthy, or even very healthy deer herd. Ethically and environmentally speaking, it will allow Idaho officials to help keep a healthy deer herd for generations to come.