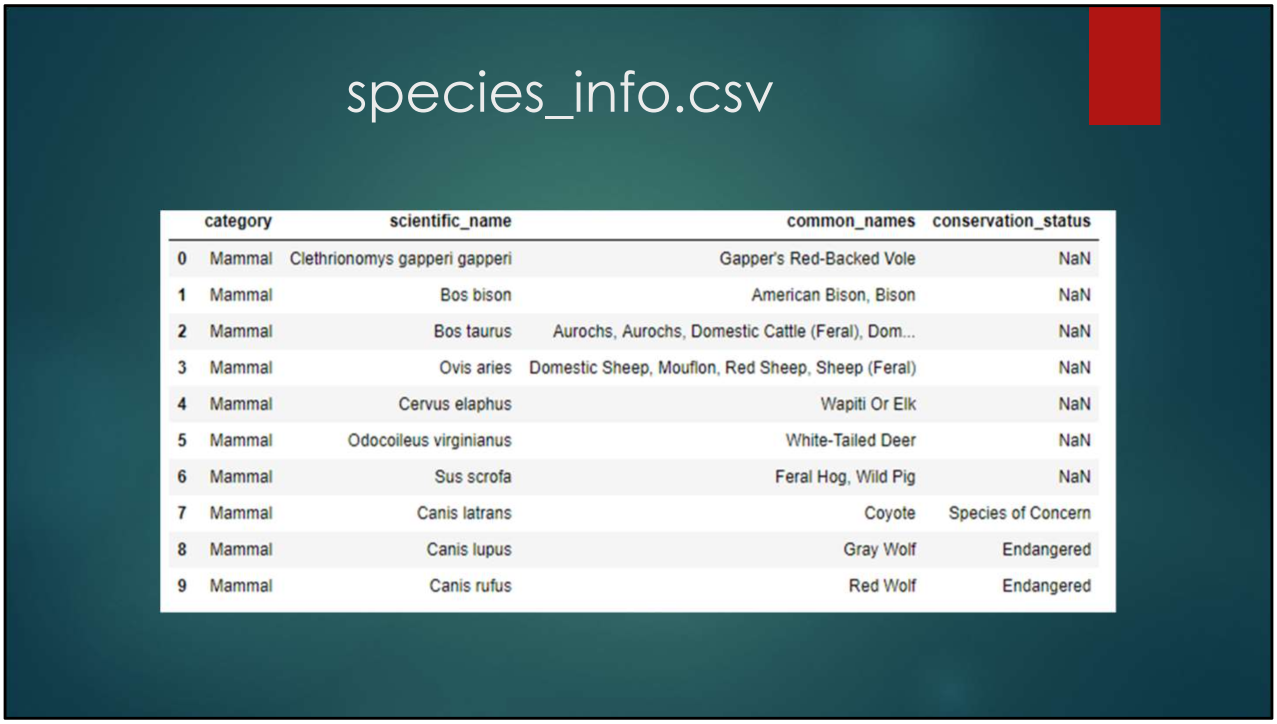
National Park Biodiversity

ANALYSIS OF SPECIES IN AMERICA'S NATIONAL PARKS

EVAN DE BROUX

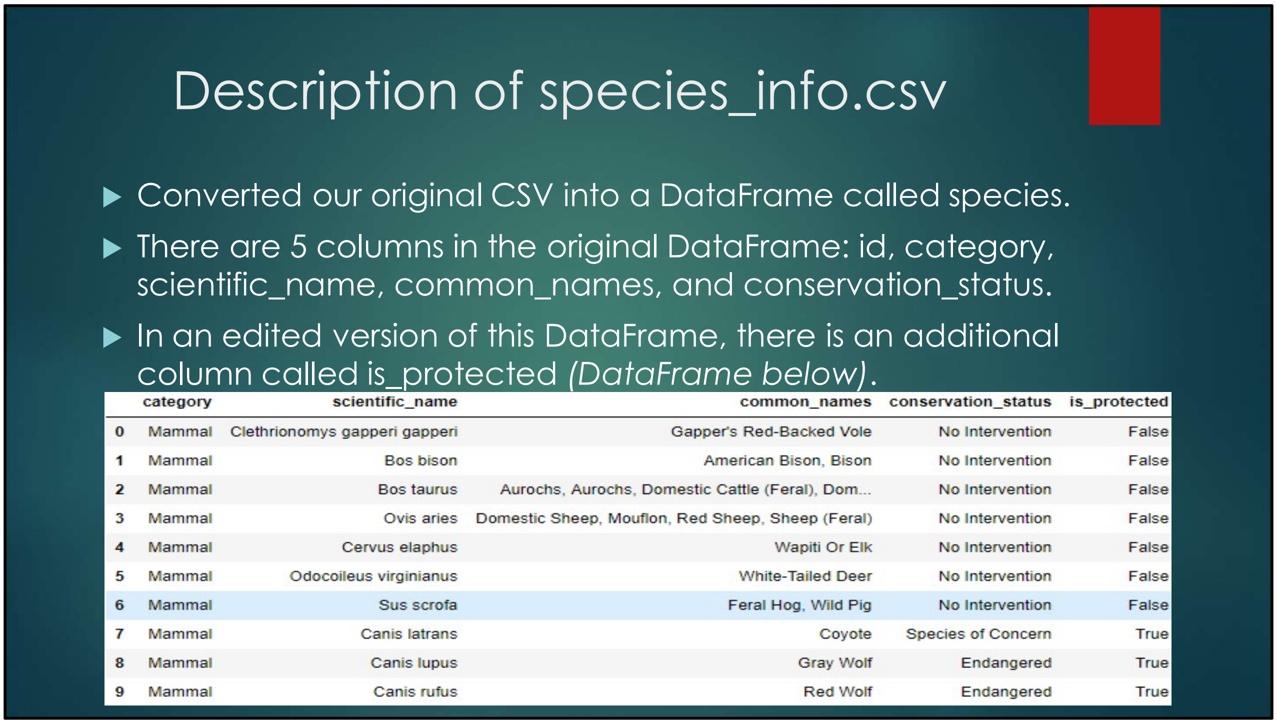
Hello, my name is Evan De Broux, and today's presentation will provide some insight into the biodiversity of species in the United States National Parks system. I was asked to provide analysis of two comma separated values files, or CSV files, given to me, species_info.csv and observations.csv, and I carried out these analysis using Python in Jupyter Notebook. I will provide descriptions of the two CSV files, the data analysis using significance tests on endangered statuses between categories of species, my recommendations to help protect these categories of species, a section on sample size determination for the foot and mouth disease study for sheep, and a section containing graphs to help visualize the data I am presenting. Now without further ado, let us begin.



species_info.csv

	category	scientific_name	common_names	conservation_status
0	Mammal	Clethrionomys gapperi gapperi	Gapper's Red-Backed Vole	NaN
1	Mammal	Bos bison	American Bison, Bison	NaN
2	Mammal	Bos taurus	Aurochs, Aurochs, Domestic Cattle (Feral), Dom...	NaN
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	NaN
4	Mammal	Cervus elaphus	Wapiti Or Elk	NaN
5	Mammal	Odocoileus virginianus	White-Tailed Deer	NaN
6	Mammal	Sus scrofa	Feral Hog, Wild Pig	NaN
7	Mammal	Canis latrans	Coyote	Species of Concern
8	Mammal	Canis lupus	Gray Wolf	Endangered
9	Mammal	Canis rufus	Red Wolf	Endangered

We will begin with a description and observations from our first dataset without any reformatting, species_info.csv. Now this table is a DataFrame of the first ten rows of species_info.csv.



Description of species_info.csv

- Converted our original CSV into a DataFrame called species.
- There are 5 columns in the original DataFrame: id, category, scientific_name, common_names, and conservation_status.
- In an edited version of this DataFrame, there is an additional column called is_protected *(DataFrame below)*.

	category	scientific_name	common_names	conservation_status	is_protected
0	Mammal	Clethrionomys gapperi gapperi	Gapper's Red-Backed Vole	No Intervention	False
1	Mammal	Bos bison	American Bison, Bison	No Intervention	False
2	Mammal	Bos taurus	Aurochs, Aurochs, Domestic Cattle (Feral), Dom...	No Intervention	False
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False
4	Mammal	Cervus elaphus	Wapiti Or Elk	No Intervention	False
5	Mammal	Odocoileus virginianus	White-Tailed Deer	No Intervention	False
6	Mammal	Sus scrofa	Feral Hog, Wild Pig	No Intervention	False
7	Mammal	Canis latrans	Coyote	Species of Concern	True
8	Mammal	Canis lupus	Gray Wolf	Endangered	True
9	Mammal	Canis rufus	Red Wolf	Endangered	True

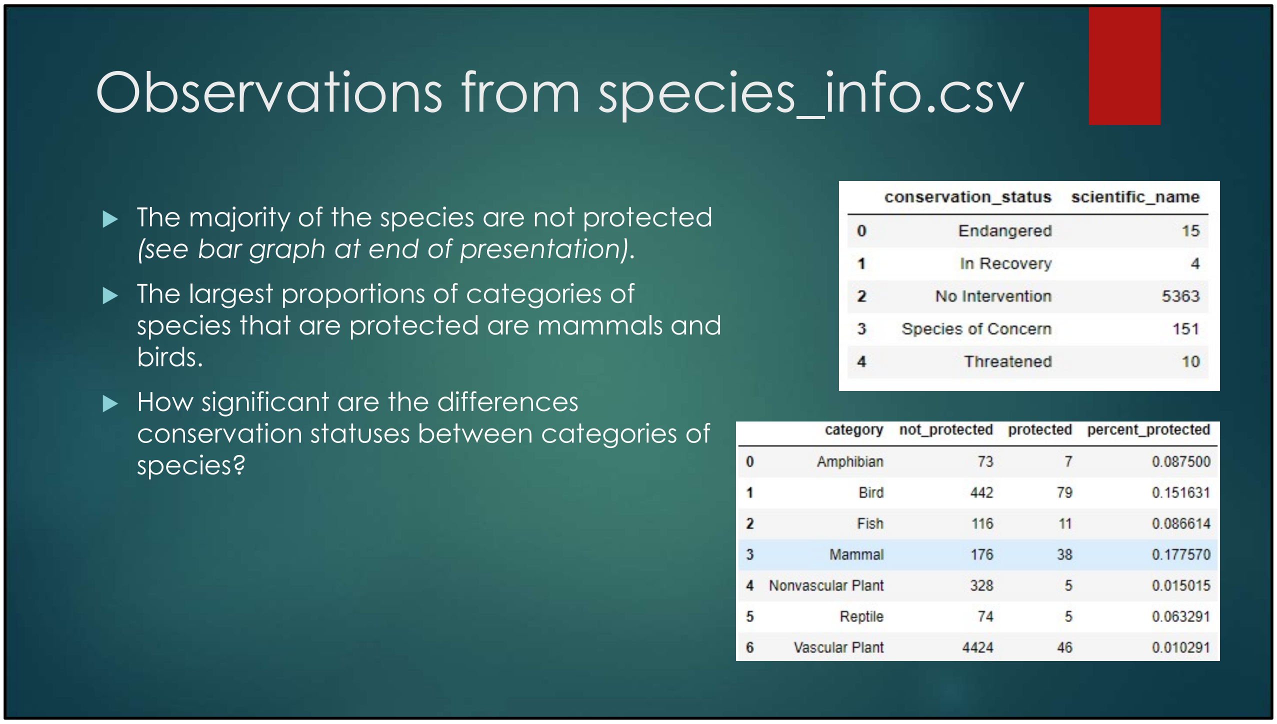
We started by converting our original CSV into a DataFrame called species. There are 5 columns in the species_info.csv DataFrame. Starting from the left, there is an id column which serves as a column of values generated by Jupyter Notebook that serves as a unique identifier for each row, i.e. species.

category lists the type of species the animal is. There are 7 different categories for each species: 'Mammal’, 'Bird’, 'Reptile’, 'Amphibian’, 'Fish’, 'Vascular Plant’, and 'Nonvascular Plant’.

scientific_name is a column that gives the scientific name of each species from our biodiversity dataset. There are 5541 unique species according to a count of unique Latin names in the scientific_name column.

common_names is a column that lists the common name of each species. As you can see from our table shown on the slide, some species may have multiple or one common name. Also, it appears that some common names may overlap even if there are different scientific names of the species, which is not uncommon.

conservation_status lists the conservation status of each species. The original table contained values of nan (not a number), 'Species of Concern’, 'Endangered’, 'Threatened’, and 'In Recovery’. In our edited version of the table, I realized that nan meant that the species was not protected, so I changed those species’ conservation status to ‘No Intervention’. I also added a column called is_protected that is False when the conservation status is ‘No Intervention’ and True otherwise.



Observations from species_info.csv

- The majority of the species are not protected *(see bar graph at end of presentation).*
- The largest proportions of categories of species that are protected are mammals and birds.
- How significant are the differences conservation statuses between categories of species?

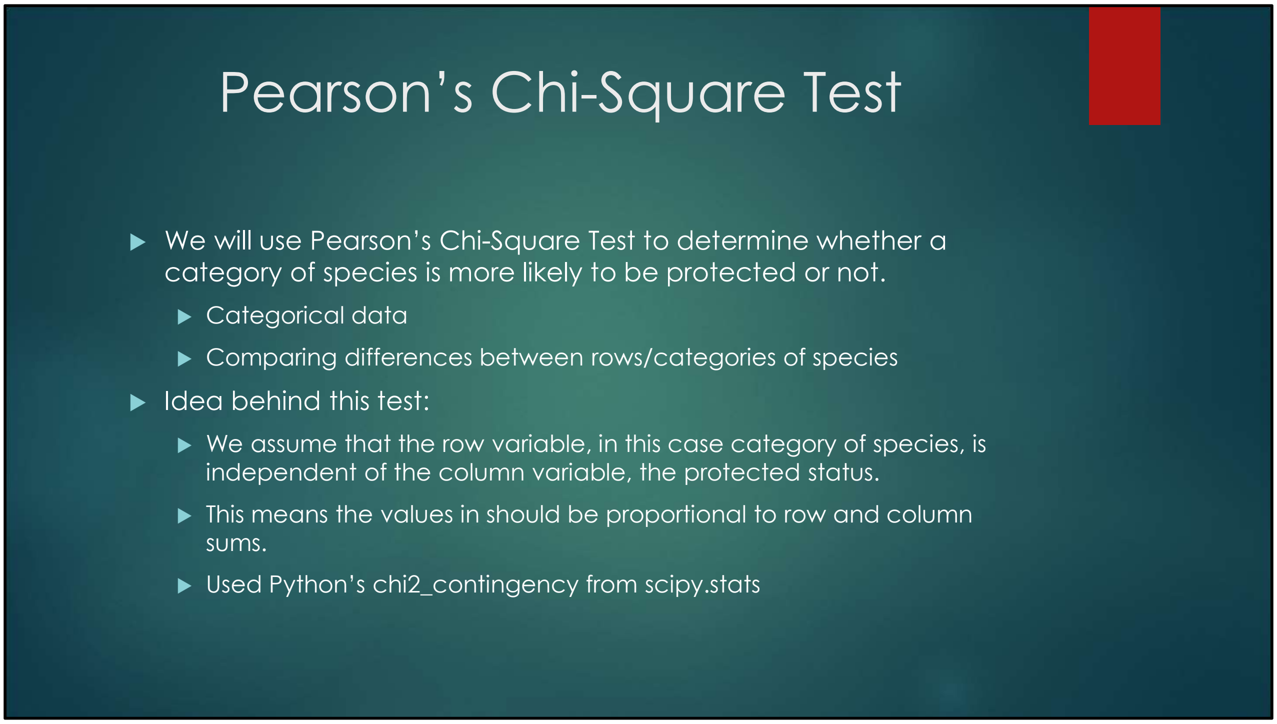
	conservation_status	scientific_name
0	Endangered	15
1	In Recovery	4
2	No Intervention	5363
3	Species of Concern	151
4	Threatened	10

	category	not_protected	protected	percent_protected
0	Amphibian	73	7	0.087500
1	Bird	442	79	0.151631
2	Fish	116	11	0.086614
3	Mammal	176	38	0.177570
4	Nonvascular Plant	328	5	0.015015
5	Reptile	74	5	0.063291
6	Vascular Plant	4424	46	0.010291

There are some observations I would like to share with you quickly. As you can see from the upper table, most species do not have some sort of protected status. There is a nice bar chart at the end of the presentation that helps us effectively visualize just how massive the difference between protected statuses of the species in our dataset. This is mainly traced to the fact that the majority of plants do not have any sort of protected status.

This is not to say that some categories of species do not have large proportions of species that are protected. Reptiles, amphibians, birds, fish, and mammals have larger proportions of species that are protected as shown in the lower table. The largest percentages that are protected by category are mammals and birds at 17.8% and 15.2%, respectively.

We have to ask ourselves, is there a way to measure just how likely it is that the differences in proportions between protected statuses between categories are observed by chance? Is it possible that the difference between reptiles and mammals occurred by chance? What about mammals and birds?



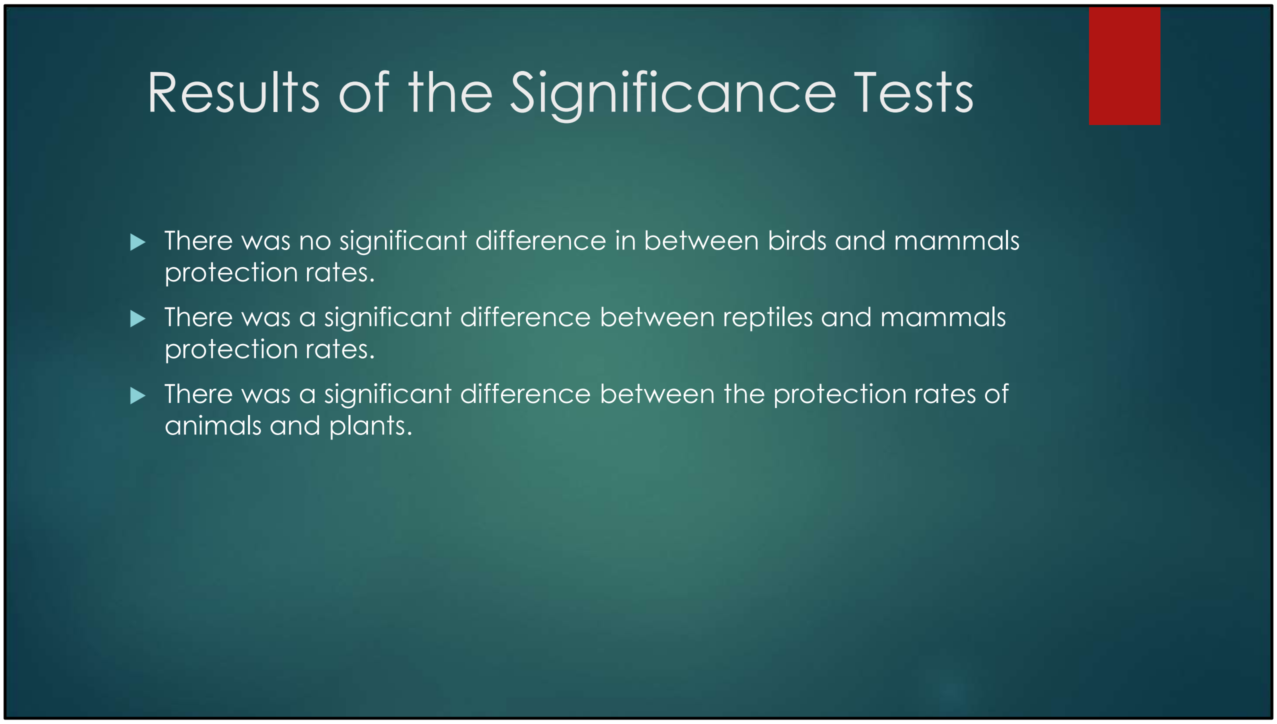
Pearson’s Chi-Square Test

- We will use Pearson’s Chi-Square Test to determine whether a category of species is more likely to be protected or not.
    - Categorical data
    - Comparing differences between rows/categories of species
- Idea behind this test:
    - We assume that the row variable, in this case category of species, is independent of the column variable, the protected status.
    - This means the values in should be proportional to row and column sums.
    - Used Python’s chi2_contingency from scipy.stats

To perform the analysis of our data, I used Pearson’s Chi-Square test to perform the analysis. There are 2 reasons behind this:
1. Our data is categorical
2. We are comparing differences between rows/categories of species. This test is specifically designed to test for these differences in categorical data for large sample sizes with large enough cell counts.

The idea behind this test is that we assume the row variable, the category of species, is independent of column variable, the protected status. This means the values should be proportional to the row and column sums. For example, the expected value of mammals that are protected is equal to the total number of species of mammal times the total number of species that are protected divided by the number of species we are comparing. If there are small deviations from this expected value for each cell, then it is fairly unlikely that there is some dependency between the column variable and the row variable.

I used Python’s chi2__contingency from scipy.stats to compare the data.



Results of the Significance Tests

- There was no significant difference in between birds and mammals protection rates.
- There was a significant difference between reptiles and mammals protection rates.
- There was a significant difference between the protection rates of animals and plants.

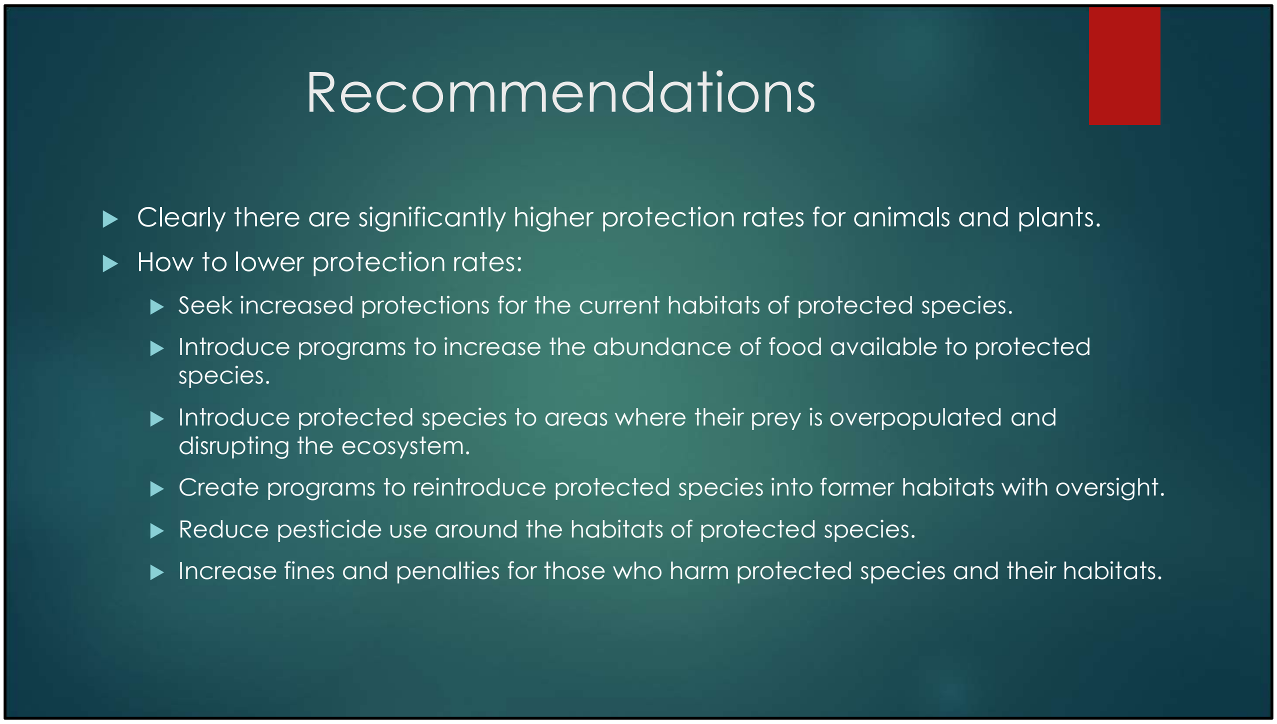
I ran three tests to check for significant differences between the protection rates of certain categories. When I run the scipy.stats program chi2_contingency in Jupyter Notebook, this will return 4 items. The first is a chi-square score, which is a summation of the difference in actual cell counts and expected cell counts squared. The second is a p-value, which is a probability that the data we are testing occurs due to chance. Lower p-values generally indicate that the data is unlikely to occur, and thus, for this specific test, we can say that there is evidence suggesting that the relationship between the row and column variables is not independent, i.e. there is a significant difference in protection rates for different species by category. The third and fourth items are the degrees of freedom and the expected value contingency table. The degrees of freedom is defined to be the number of factors in the final calculation of the statistic that are allowed to vary. For each of the tests I ran, there was only 1 degree of freedom. The expected value contingency table is the values we would expect if the row and column variables were independent of each other.

When I describe the results, I am only going to mention the p-value. If the p-value is below some predetermined significance level, in my case alpha = 0.05, then we would reject the null hypothesis and say there is a significant difference between protection rates of categories of species.

The first test I ran between mammals and birds yielded a p-value of 0.4459. This indicates that there was a 44.6% chance that these results occurred randomly. Thus, we cannot say that there is a significant difference between the protection rates of mammals and birds.

The second test between mammals and reptiles yielded a p-value of 0.0233 however. Since this is below our significance threshold, we would say that there is a significant difference between the protection rates of mammals and reptiles.

The third test I ran between animals and plants yielded a p-value of 1.62 x 10^-93. This is well below the significance threshold, so we would say there is a significant difference between plant protection rates and animal protection rates.



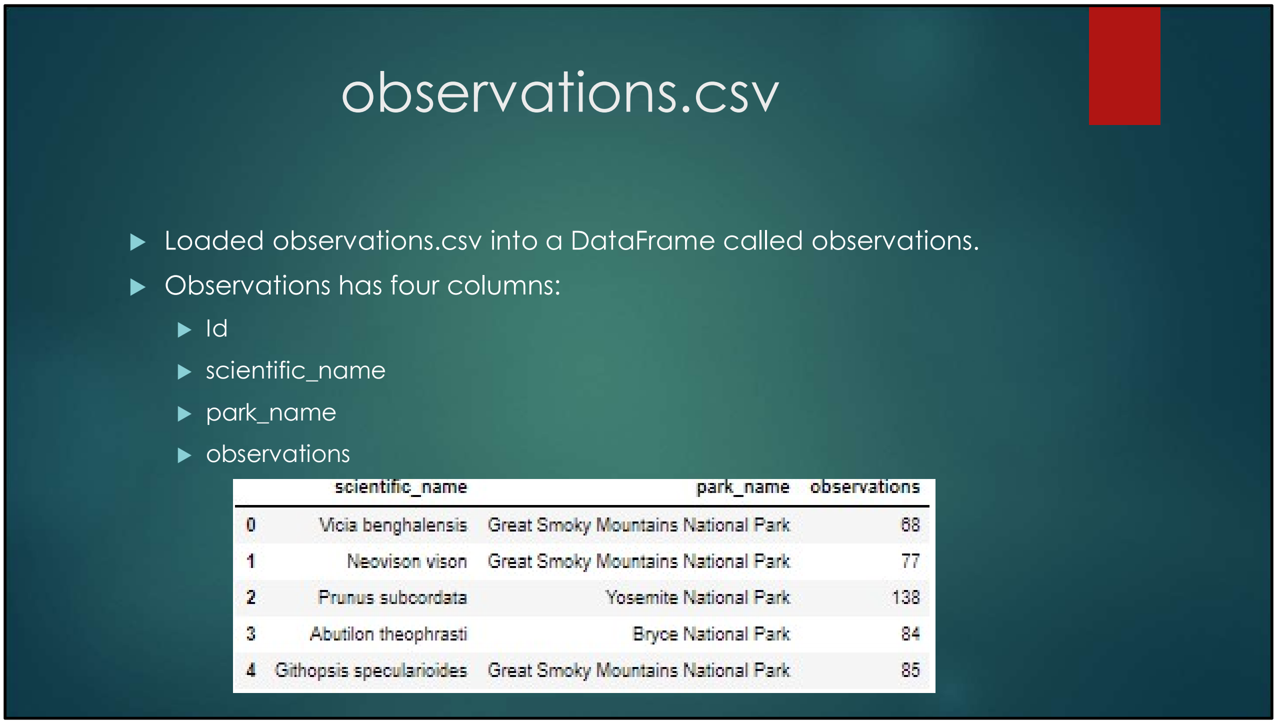
Recommendations

- Clearly there are significantly higher protection rates for animals and plants.
- How to lower protection rates:
  - Seek increased protections for the current habitats of protected species.
  - Introduce programs to increase the abundance of food available to protected species.
  - Introduce protected species to areas where their prey is overpopulated and disrupting the ecosystem.
  - Create programs to reintroduce protected species into former habitats with oversight.
  - Reduce pesticide use around the habitats of protected species.
  - Increase fines and penalties for those who harm protected species and their habitats.

As mentioned previously there is clearly a higher amount of protected species of animals than plants. Specifically, there is a significant difference between mammals and reptiles, but we will wish to try to reduce the protected species percentages by getting these species of this list.

My recommendations include:

1. Seek increased protections for the current habitats of protected species.
2. Introduce programs to increase the abundance of food available to protected species.
3. Introduce protected species to areas where their prey is overpopulated and disrupting the ecosystem. An example of this is the reintroduction of wolves to Yellowstone National Park where they were brought in to quell the Elk population. Obviously, the impact of reintroducing that species on other types of prey should also be considered.
4. Create programs to reintroduce protected species into former habitats with appropriate oversight.
5. Reduce pesticide use around habitats of all protected species.
6. Increase fines and penalties for those who harm protected species and their habitats.

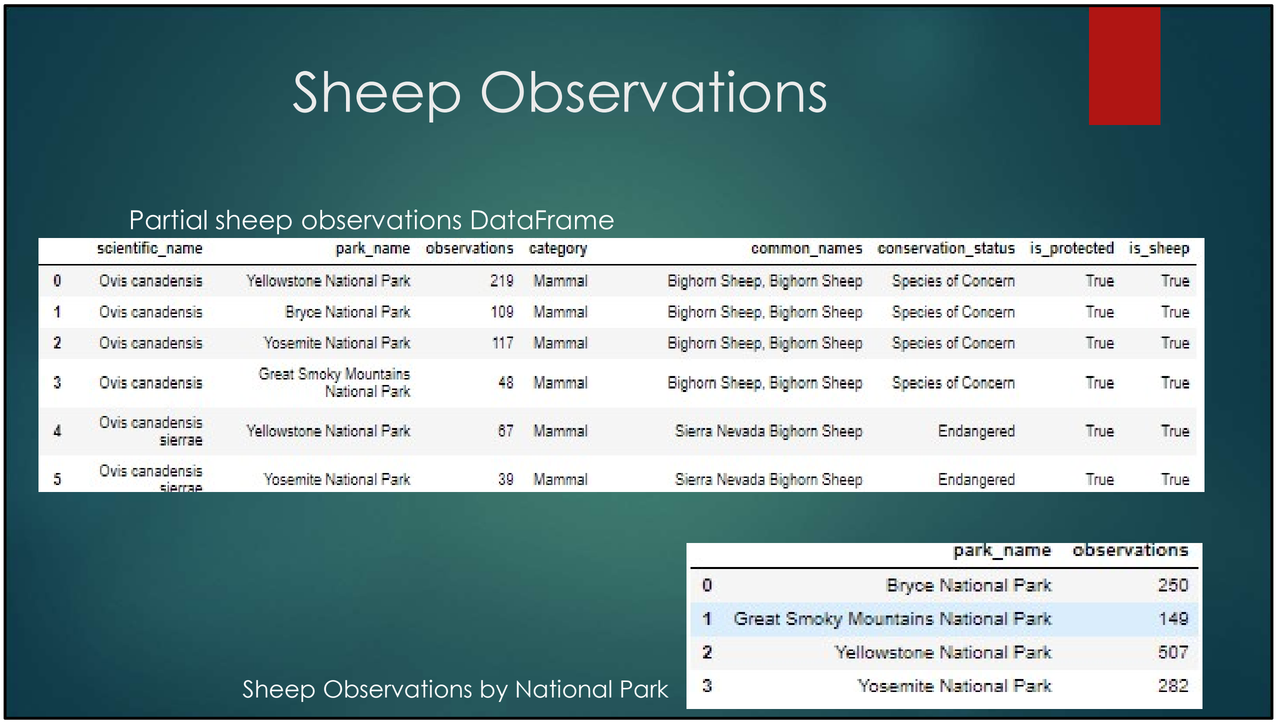


observations.csv

- Loaded observations.csv into a DataFrame called observations.
- Observations has four columns:
  - Id
  - scientific_name
  - park_name
  - observations

	scientific_name	park_name	observations
0	Vicia benghalensis	Great Smoky Mountains National Park	68
1	Neovison vison	Great Smoky Mountains National Park	77
2	Prunus subcordata	Yosemite National Park	138
3	Abutilon theophrasti	Bryce National Park	84
4	Githopsis specularioides	Great Smoky Mountains National Park	85

I was also given a CSV of observations of species at national parks in the last week. This CSV was loaded into a DataFrame called observations. This DataFrame has four columns: an id column like the one described in the species DataFrame from before, scientific_name which contains the scientific name of each observed species, park_name which is the name of the park the species was observed in, and observations which is the number of times the species was observed.



Sheep Observations

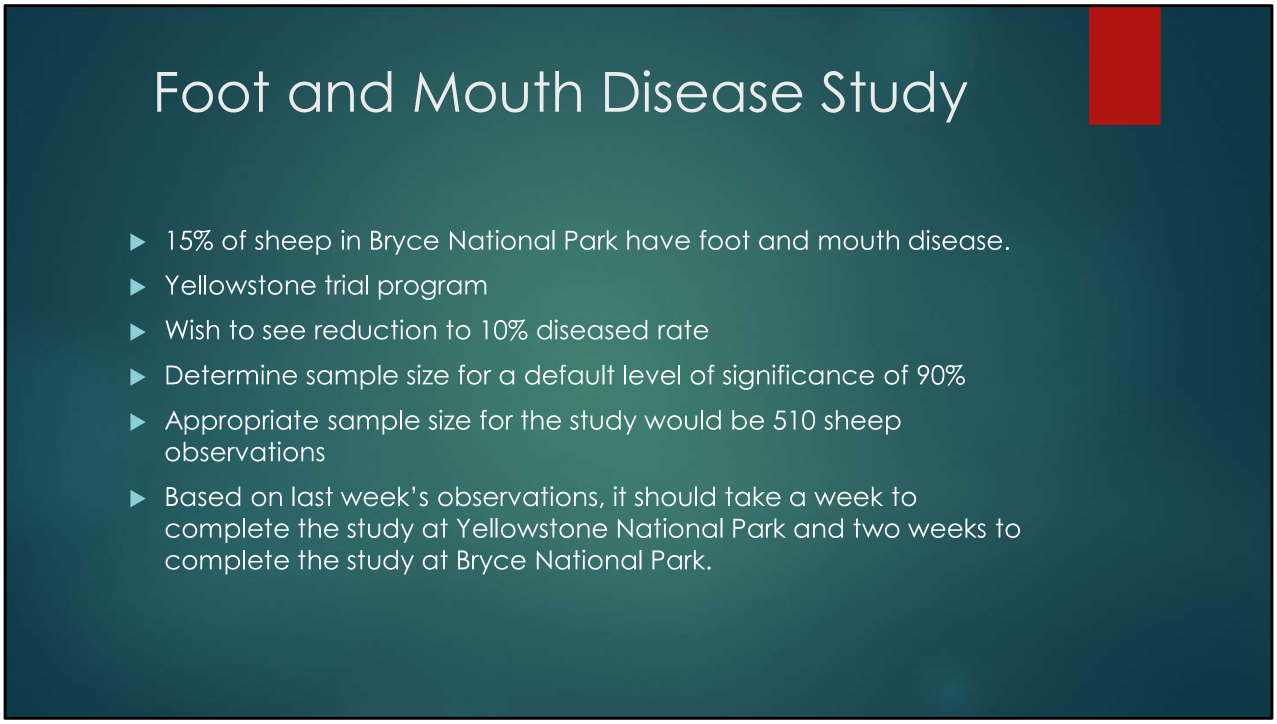
Partial sheep observations DataFrame

	scientific_name	park_name	observations	category	common_names	conservation_status	is_protected	is_sheep
0	Ovis canadensis	Yellowstone National Park	219	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True
1	Ovis canadensis	Bryce National Park	109	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True
2	Ovis canadensis	Yosemite National Park	117	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True
3	Ovis canadensis	Great Smoky Mountains National Park	48	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True
4	Ovis canadensis sierrae	Yellowstone National Park	67	Mammal	Sierra Nevada Bighorn Sheep	Endangered	True	True
5	Ovis canadensis sierrae	Yosemite National Park	39	Mammal	Sierra Nevada Bighorn Sheep	Endangered	True	True

Sheep Observations by National Park

	park_name	observations
0	Bryce National Park	250
1	Great Smoky Mountains National Park	149
2	Yellowstone National Park	507
3	Yosemite National Park	282

I was able to combine the species DataFrame containing sheep (the mammal) with the observations DataFrame using merge and sorting operations. From the partial DataFrame displayed on the left, this sheep DataFrame contains a combination of the observations DataFrame and the species DataFrame. This table contains only sheep that are mammals, which is obvious given that the is_sheep column is True in all rows and the category is listed as mammal for all rows as well. There is also a bar graph of all the observations of sheep by national park at the end of this presentation. By table of sheep observations in the lower right of the slide, it is clear that Yellowstone had the most sheep observations last week, followed by Yosemite, Bryce, and Great Smoky Mountains National Parks.

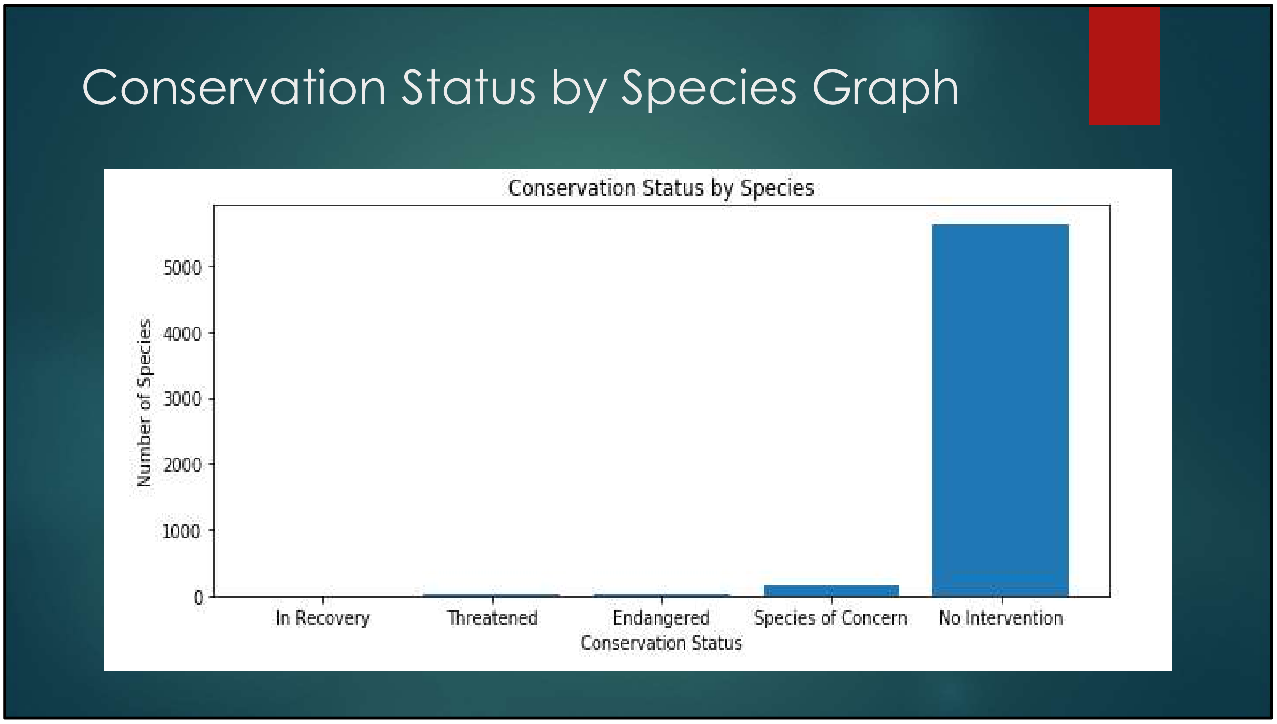


Foot and Mouth Disease Study

- 15% of sheep in Bryce National Park have foot and mouth disease.
- Yellowstone trial program
- Wish to see reduction to 10% diseased rate
- Determine sample size for a default level of significance of 90%
- Appropriate sample size for the study would be 510 sheep observations
- Based on last week's observations, it should take a week to complete the study at Yellowstone National Park and two weeks to complete the study at Bryce National Park.

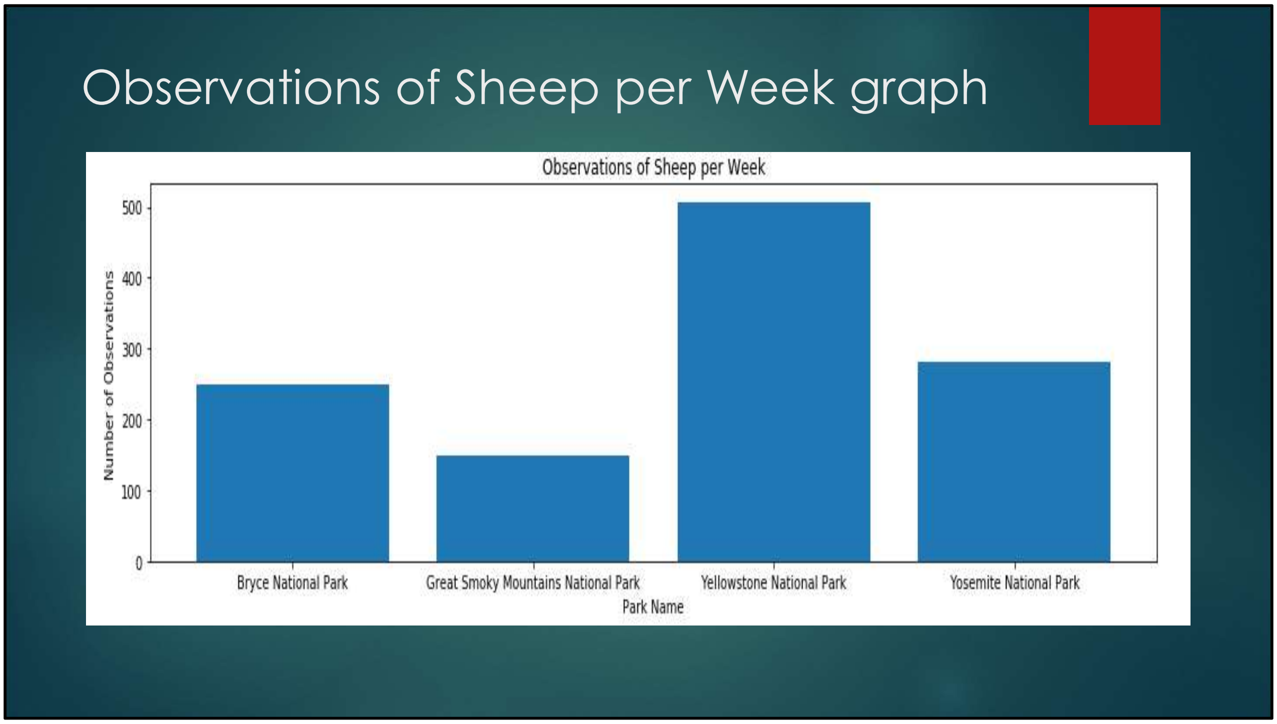
Our scientists know that 15% of sheep in Bryce National Park suffer from foot and mouth disease. Park rangers from Yellowstone National Park wish to know whether their program to reduce foot and mouth disease is working, and wish to detect a reduction in 5% for diseased rates. Some simple math would give us that the minimum detectable effect the rangers want to be aware of is 33.3%. Using a baseline rate of 15% and a default level of significance of 90%, I determined the sample size per variant using the Optimizely sample size calculator. The appropriate sample size was 510 sheep for the study.

If we harken back to the previous slide. We will remember that there were 507 observations of sheep in the last 7 days at Yellowstone and 250 observations of sheep at Bryce. This means that if we observe sheep at these national parks according to last week's rate, the study would take roughly 1 week to complete at Yellowstone National Park and roughly 2 weeks to complete at Bryce National Park.



Conservation Status by Species Graph

This is the bar graph mentioned in slide 4. As noted before and as my chart clearly demonstrates, the majority of species are not being protected at this point in time.



Observations of Sheep per Week graph

This is the graph mentioned in slide 9. It is a bar chart of all the observations of sheep in the four national parks. As the previously stated and shown by the graph now, Yellowstone had the most sheep observations, followed by Yosemite, Bryce, and the Great Smoky Mountains, respectively.

Questions?

That concludes my presentation. If anyone wishes to view the code I inputted to Jupyter Notebook, I please send me an email and I will send you the program as a Python program. Are there any other questions at this time?

Thank you very much.