Biodiversity Capstone

Jonathan Roberts

Tasks

- species_info.csv
- Analyze and plot conservation status by species
  - Final thoughts/analysis
- Investigate endangered species
  - Test for significance
  - Final thoughts/analysis
- Observations.csv
- Foot and mouth reduction effort sample size analysis
  - Isolate sheep observations
  - Final thoughts/analysis
- Conclusion

species_info.csv

I found this CSV lists four columns describing species in the National Parks:

1. Category
   a. Mammal, Bird, Reptile, Amphibian, Fish, Vascular Plant, Nonvascular Plant
2. Scientific Name (ex. Ovis aries)
3. Common Names (ex. Sheep)
4. Conservation Status
   a. Species of Concern - declining population or appears to be in need of conservation.
   b. Threatened - vulnerable to endangerment in the near future.
   c. Endangered - seriously at risk of extinction.
   d. In Recovery - formerly Endangered, but currently not in danger of extinction throughout all or a significant portion of its inhabitable range.

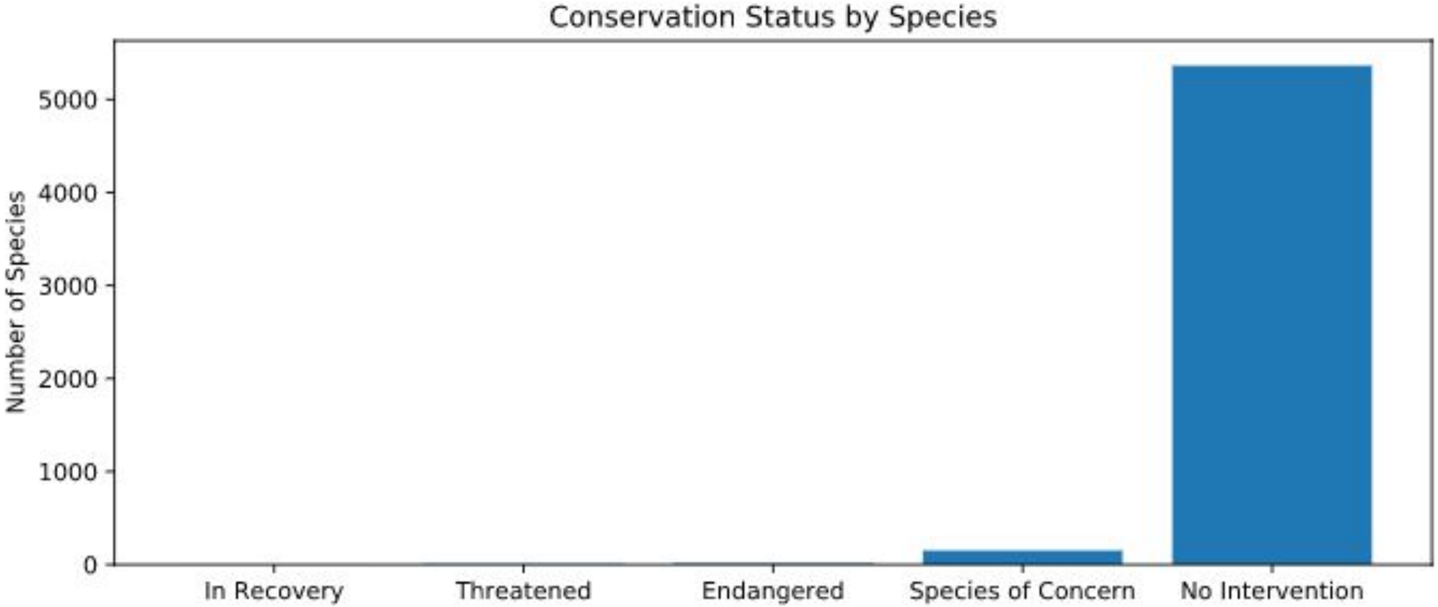
Listed was a total of **5,541** different species.

species_info.csv Analysis/Thoughts

First, I isolated the total of each conservation status, and listed each number. These totals were far lower than the total species I knew the CSV to contain. I then assigned the NaN or “conservation status equal to none” values to reflect a status of “No Intervention”. This allowed us to use this NaN number as a percentage of the overall species total, and to accurately visualize the data in the CSV, based on conservation status. I did this by creating a bar chart for the numbers which can be found on the next slide.

The findings were that there is a very low percentage of the overall CSV in need of conservation.

Conservation Status by Species Bar Chart



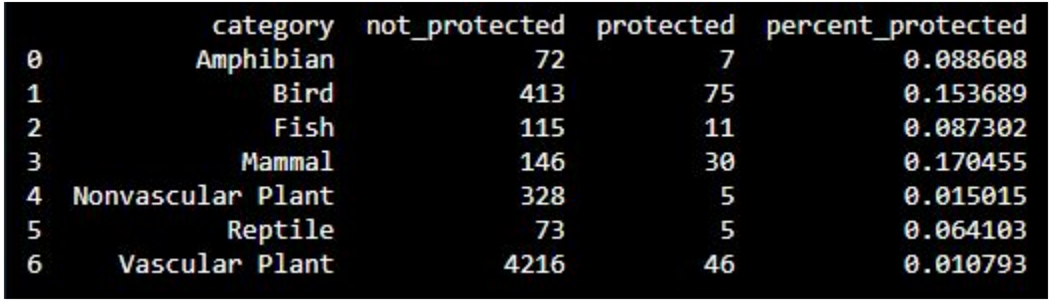
species_info.csv Analysis/Thoughts

This led me to the next question of “Are certain types of species more likely to be endangered?”

I created a pivot table that identified whether or not a species was under No Intervention or not and grouped by the species category, allowing me to then calculate this as a percentage of its own total of protected category mates, versus not protected.

What I found was that **yes,** Mammals are the most likely to be endangered, as reflected by the aforementioned pivot table on the following slide.

Endangered Pivot Table



```
            category  not_protected  protected  percent_protected
0          Amphibian             72          7           0.088608
1               Bird            413         75           0.153689
2               Fish            115         11           0.087302
3             Mammal            146         30           0.170455
4  Nonvascular Plant            328          5           0.015015
5            Reptile             73          5           0.064103
6     Vascular Plant           4216         46           0.010793
```

Testing the significance and results

In order to test if, in fact, Mammals are more likely to be endangered than the next closest category, Birds, I used a Chi-Squared Test.

I created a contingency based off the protected and not protected numbers for each, then saved the p-value for the test and ran the numbers for mammals against reptiles in order to have more than one comparison.

The results of the first test gave us a p-value of 0.688, which we then CANNOT reject the null hypothesis of the difference in percentages was a result of chance. So, mammals being more likely than birds was the result of chance.

The second test pval of 0.038 means we reject the null hypothesis and can conclude that certain types of species are more likely to be endangered than others.

Observations.csv

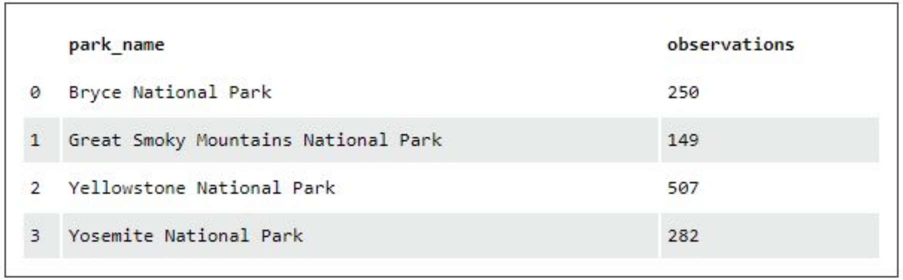
The National Park Service sent me a second dataset to analyze. Conservationists have been recording sightings of different species at several national parks for the past 7 days. Their observations were recorded in observations.csv.

This csv contains two columns:

1. Scientific name
2. Park name

Observations.csv Analysis/Thoughts

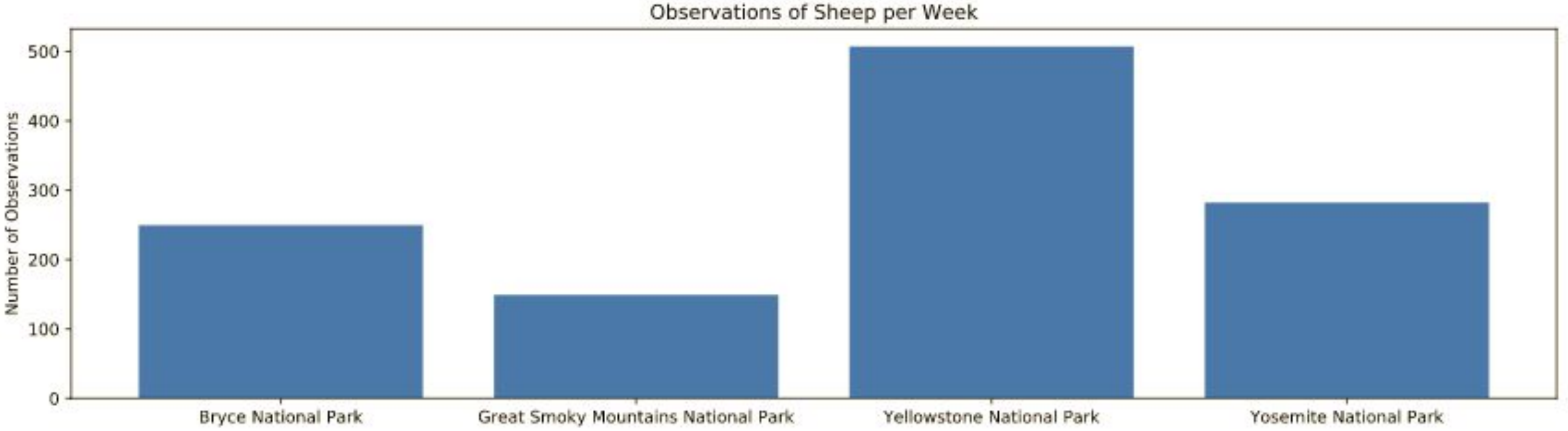
I was instructed to narrow this csv down to total sheep observations and grouped them by park name, which yielded the following:



	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 plotted these numbers on a bar chart found on the next slide.

Observations of Sheep per Week Bar Chart



Yellowstone Foot and Mouth Reduction Effort

Following the implementation of a program to reduce the rate of foot and mouth disease, the park rangers at Yellowstone wanted my assistance in testing whether or not their program was working.

Wanting to detect reductions of at least 5% points, with confidence, I calculated the minimum sample size of sheep needed to be observed from each park using last years baseline conversion rate (percent of sheep with F&M) of 15% from Bryce National Park.

With the standard statistical significance of 90% and the minimum detectable effect being 100*x / baseline or equal to 33.33, I calculated that the sample size needed is 510 (using the Optimizely Calculator, the Codecademy one is bugged) .

I then combine this number with the observation numbers per of sheep per week for each park in the sheep observations bar chart.

Foot and Mouth Reduction Effort - Results/Thoughts

With the sample size needed being 510, per the rate of sheep observed per week at each park, scientists would need the following amount of time at each:

Yellowstone = 510 / 507 or about one week

Bryce = 510 / 250 or about two week

The scientists could stay an extra day to ensure they get enough observations and to postpone sitting inside an office all day like I am now.

Summary/Conclusion

Data analysis can be very powerful when used correctly. With only a small amount of data, I am able to extract several useful nuggets of information based on the priorities given to me as well as visualize this information in order to better understand and report my results.

And so ends this course. What a ride.