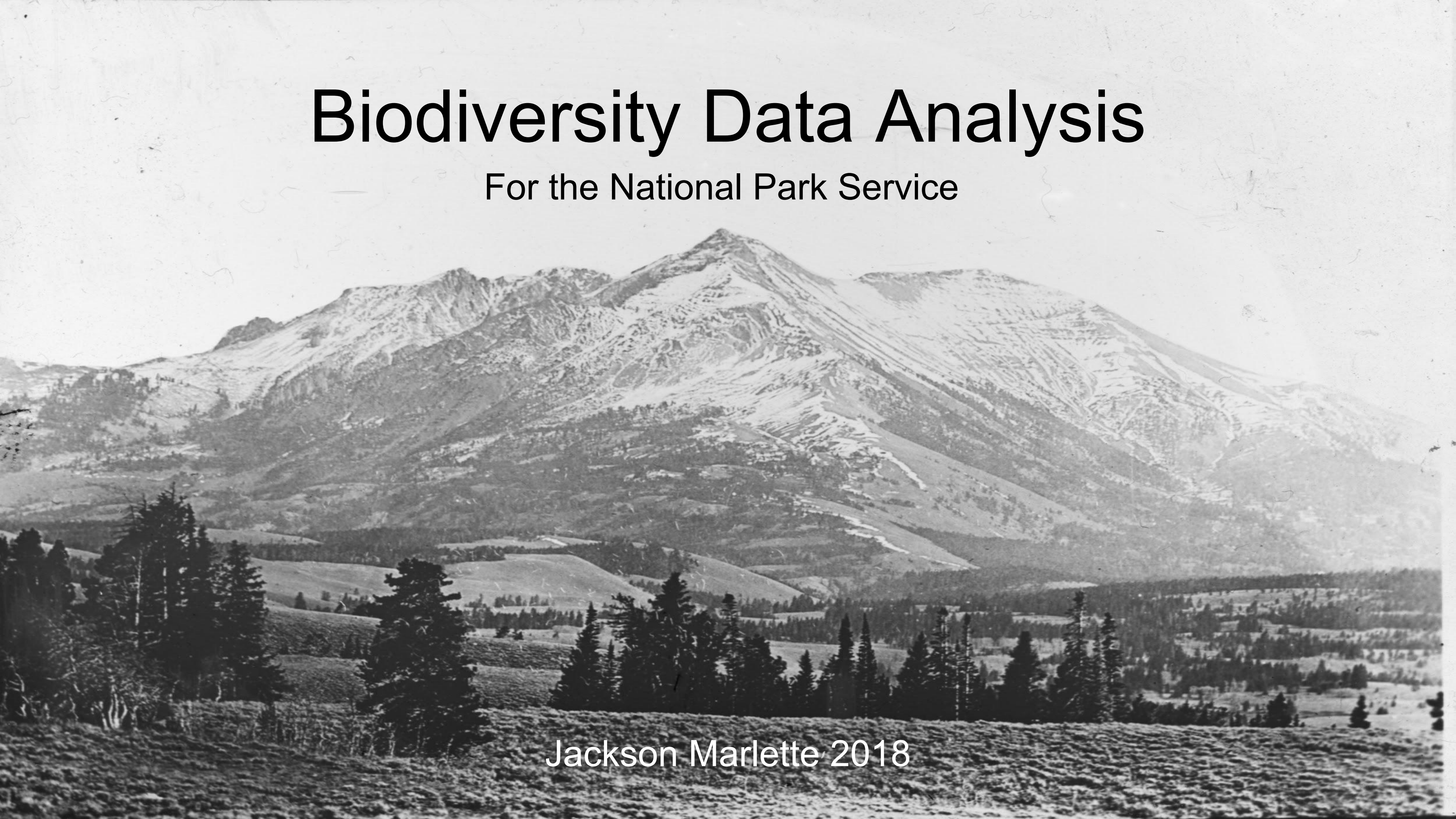
Biodiversity Data Analysis

For the National Park Service



Jackson Marlette 2018

Summary of Data in species_info.csv

The data from species_info.csv presents 5541 species divided into four columns:

- Category (Mammal, Bird, Reptile, Amphibian, Fish, and Vascular Plant),
- Scientific Names
- Common Names
- Conservation Status

The column ‘Conservation Status’ provides the primary focus for the first part of our analysis. Its values are:

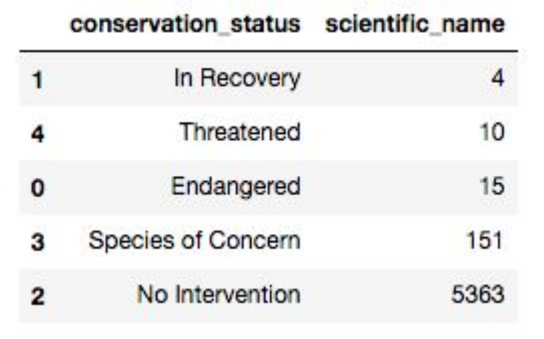
- Species of Concern
- Threatened
- Endangered
- In Recovery

After initial inspection, all of the null values for Conservation Status have been filled with a fifth category:

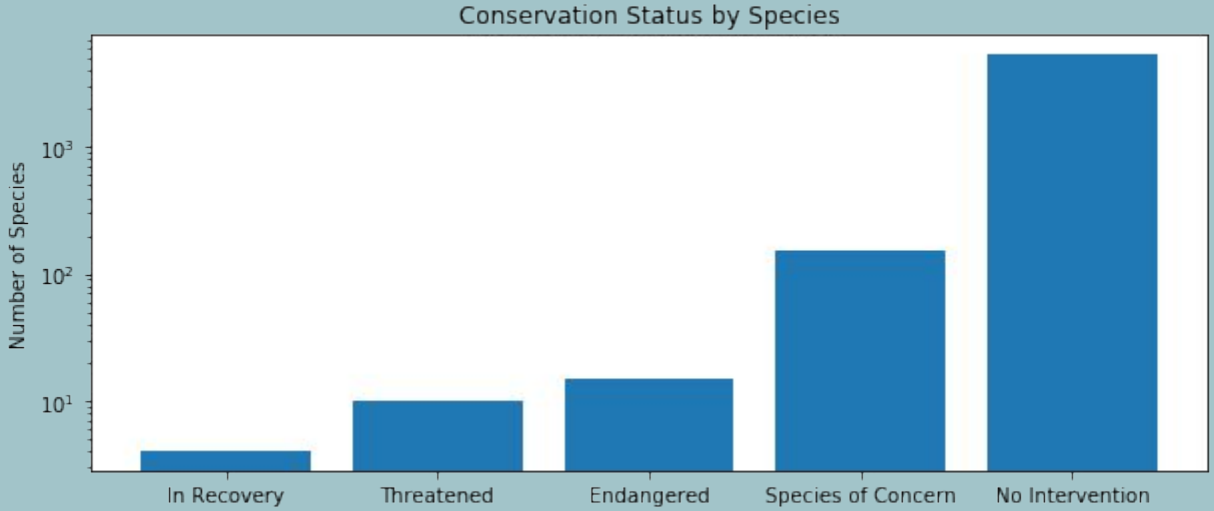
- No Intervention

Conservation Status by Species

When we group by conservation status, we find that species classed as ‘No Intervention’ dramatically outnumber those that are requiring any level of protection.



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



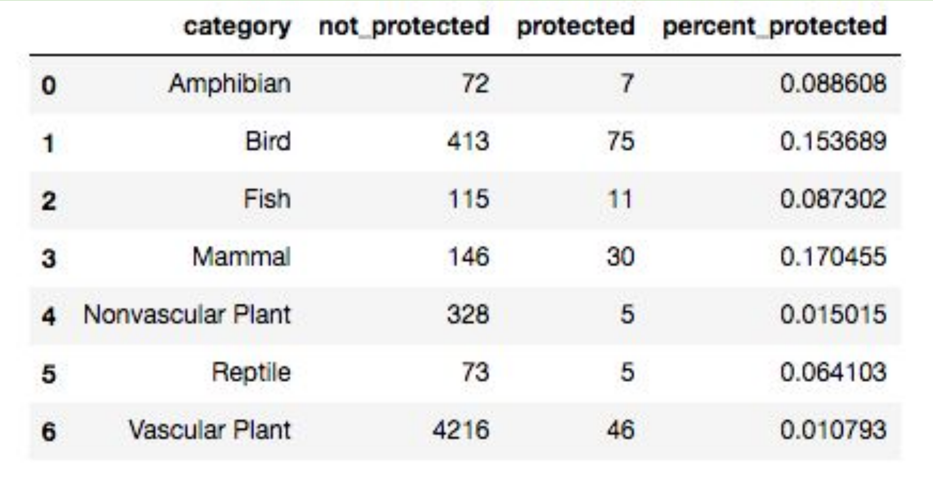
(Note: Because of the drastic value differences between species with no intervention and other categories, logarithmic scaling has been implemented on the y-axis in order to improve readability. ‘Number of Species’, refers to the values under the ‘scientific_name’ column on the left table.

Percent Protected

The next question proposed was whether or not certain species are more likely to become endangered than others.

Here we grouped each species category by ‘protected’ or ‘not_protected’.

‘protected’ was defined by any level of intervention within the Conservation Status column, while ‘not_protected’ was defined by the species having ‘No Intervention’ as its Conservation Status value.



	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

From there we were able to calculate the percentage who are protected out of each species category. As you can see, both mammals and birds (about 17% and 15% respectively) were most commonly put under some form of protection.

Significance Calculations for Endangered Status Among Species

Although it appears that mammals are more likely to be endangered than birds, we need to prove this via a significance test. With two categorical datasets, we determined that a chi-squared test was necessary to prove or disprove the null hypothesis: that there is no significant difference between data, a result of natural fluctuations or chance.

Upon running a contingency table with the mammal and bird data through a chi square test, our p-value came to about 0.69 (rounded to 2 decimal points). Because our p-value > 0.05, our null hypothesis couldn’t be disproven. Therefore there seems to be no significant difference between the data for mammals and birds.

However, after running the same test for mammals and reptile data, we discovered a p-value of about 0.04 (to 2 decimal points). Since our p-value < 0.05, we were able to reject the null hypothesis, thereby proving that mammals are more likely to become endangered (or require protection) than reptiles.

Recommendations

From the tables and significance tests describing the ratios of endangered wildlife, we can tell that while there isn’t much individual categorical difference between mammals and birds, while there is significant difference between mammals and reptiles (and likely with any other category equal-to or less-than the latter in percentage).

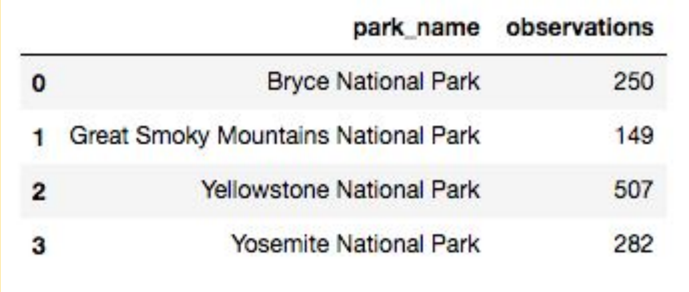
Overall mammals and birds require the most intervention out of all of the species, and therefore conservation budgets could be adjusted accordingly.

Mammals and birds might be on the receiving end of the most detrimental impact. It would be helpful to know whether or not the causation is man-made or otherwise. Collecting information that elucidates this causation of endangerment would benefit the dataset immensely.

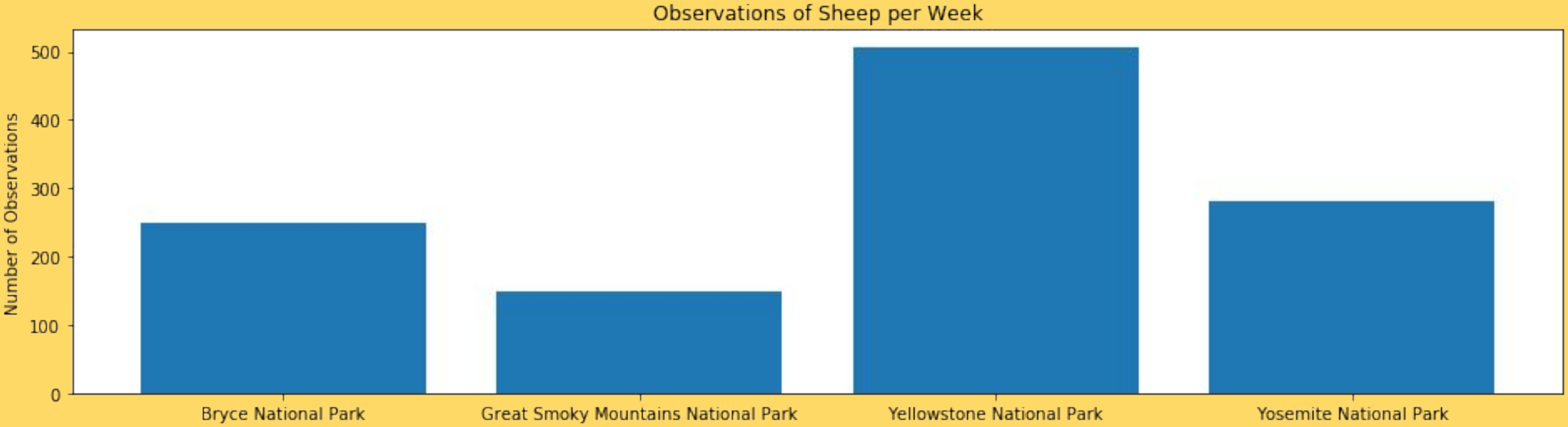
Sheep Observations by National Park

The dataset ‘observations.csv’, contains the number of observations of species made by conservationists in different national parks over a seven day period.

To find the total count of sheep species per national park over this week-long period, we selected the different species of sheep in ‘species.csv’ and merged them with results with ‘observations’.csv in order to create the table ‘sheep observations’. We then grouped the observation numbers by national park, shown here:



	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



Sample size determination for foot and mouth disease study

The data outlines a foundation to make a sample size determination for the foot and mouth disease study by the park rangers at Yellowstone National Park.

- 15% of sheep at Bryce National Park have foot and mouth disease, so this figure will act as our baseline conversion rate.
- Scientists need to detect a difference within 5 percentage points with confidence.
- Minimum Detectable Effect = 100 x (0.05 / 0.15) = 33% (rounded to nearest integer)
- With statistical significance set at 90%, we used the sample size calculator at Optimizely.com to determine that our required sample size would be 520 sheep.

Therefore to calculate the amount of weeks it would take to observe our sample size at each park, we divide the sample size by the number of sheep observed per week at each park:

- Bryce National Park would be (520 / 250) = ~ 2 weeks
- Yellowstone would be (520 / 507) = ~ 1 week

Conclusion

This concludes our presentation on data analysis for the National Park Service.

A copy of the analysis code is included within the submitted folder.

Thank You!