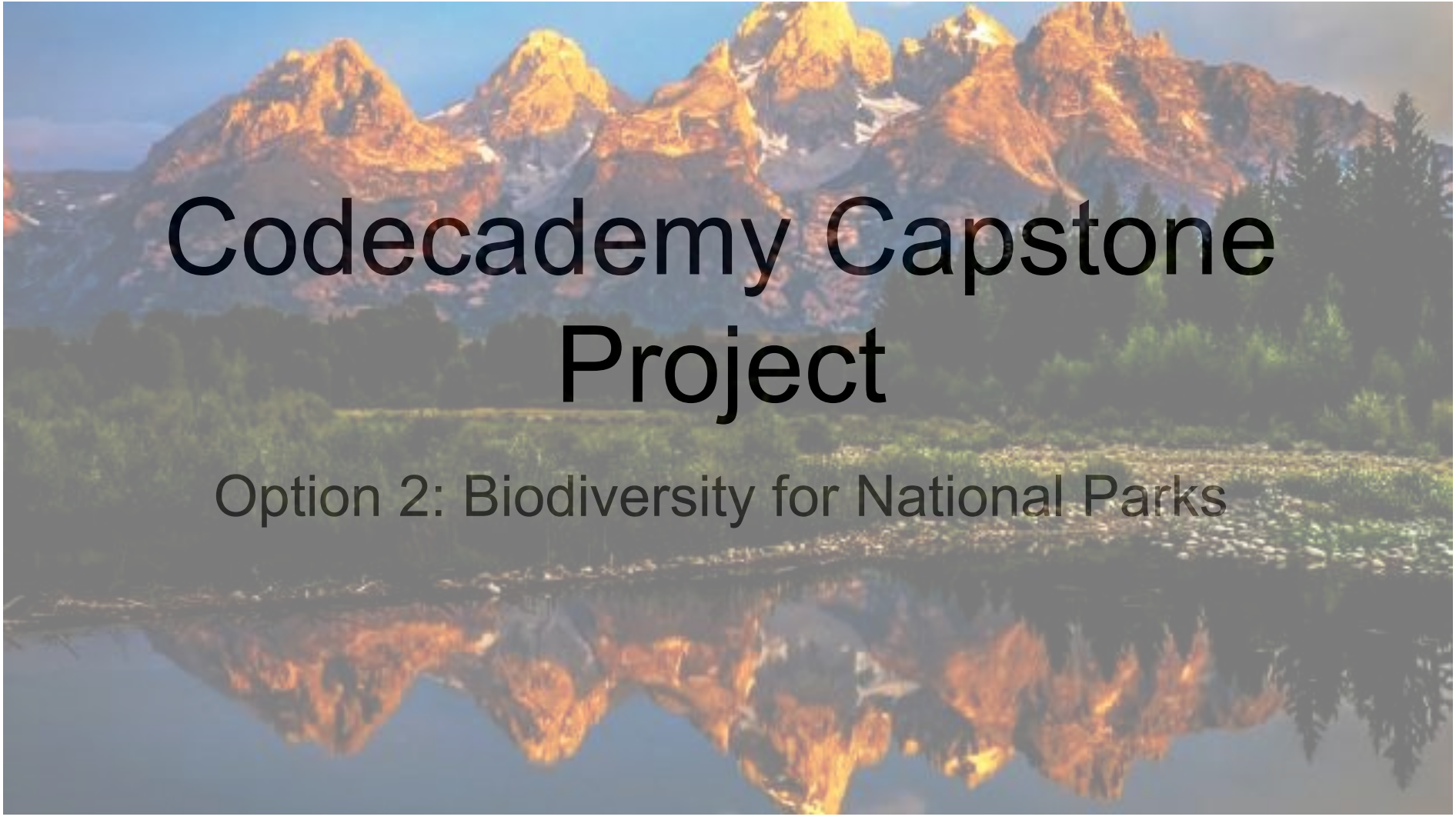


Codecademy Capstone Project

Option 2: Biodiversity for National Parks

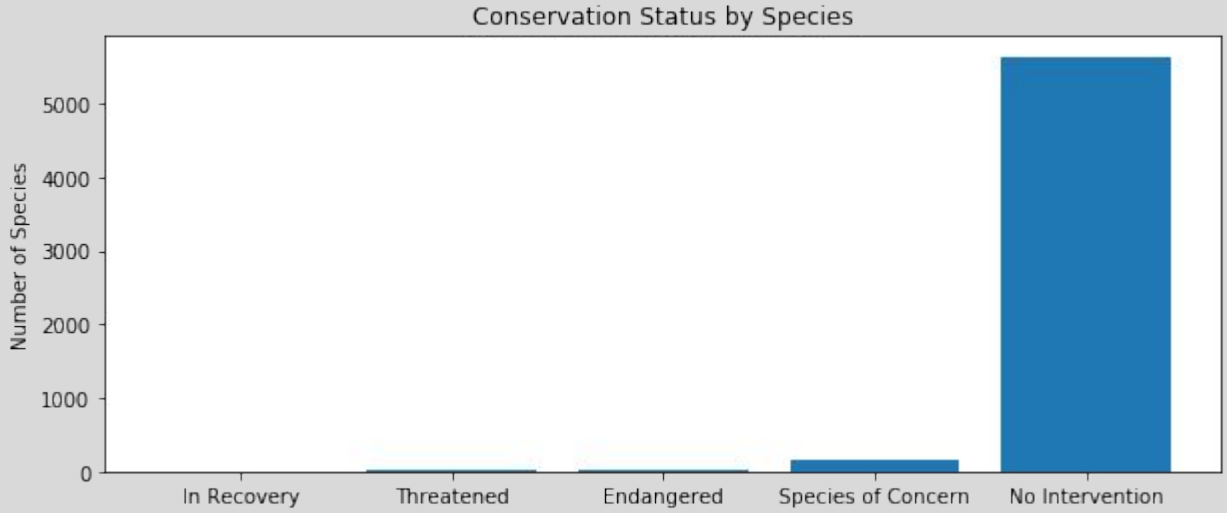
Species_info.csv

The ‘Species_info.csv’ document contains 4 columns: category, scientific name, common name, and conservation status.

After inspection, we’ve found that there are six categories of species: mammals, birds, reptiles, amphibian, fish, and vascular plants.

Both scientific names and common names are self-explanatory.

For conservation status, we’ve edited the results so that the previous ‘NaN’ returns ‘no intervention’ instead. The other conservation statuses are: threatened, species of concern, in recovery, and endangered. For the purposes of this analysis, we will be addressing these four statuses uniformly.



We decided to ask the question: **“Are certain types of of species more likely to be endangered?”**

Continuing forward, we created a new column in our ‘species_info’ document called ‘is protected’. We deemed species with no intervention as protected and anything else as not protected.

Using the same dataset, we calculated percentages of those protected and grouped them by their category.

	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

Both birds and mammals returned with the highest results with 15% and 17% respectively. Was this due to chance (null hypothesis) or are birds and mammals less likely to be endangered?

We decided to run a chi-squared test because we the datasets were categorical and there were more than 2 datasets to compare.

Comparing *mammals* to *birds*, we found a p-value of 0.688 which was higher than 0.5. Not what we were hoping for.

Yet, when we ran the datasets of *mammals* to *reptiles* we achieved a p-value of 0.038. From this finding we can conclude that certain types of species are more likely to be endangered than others.

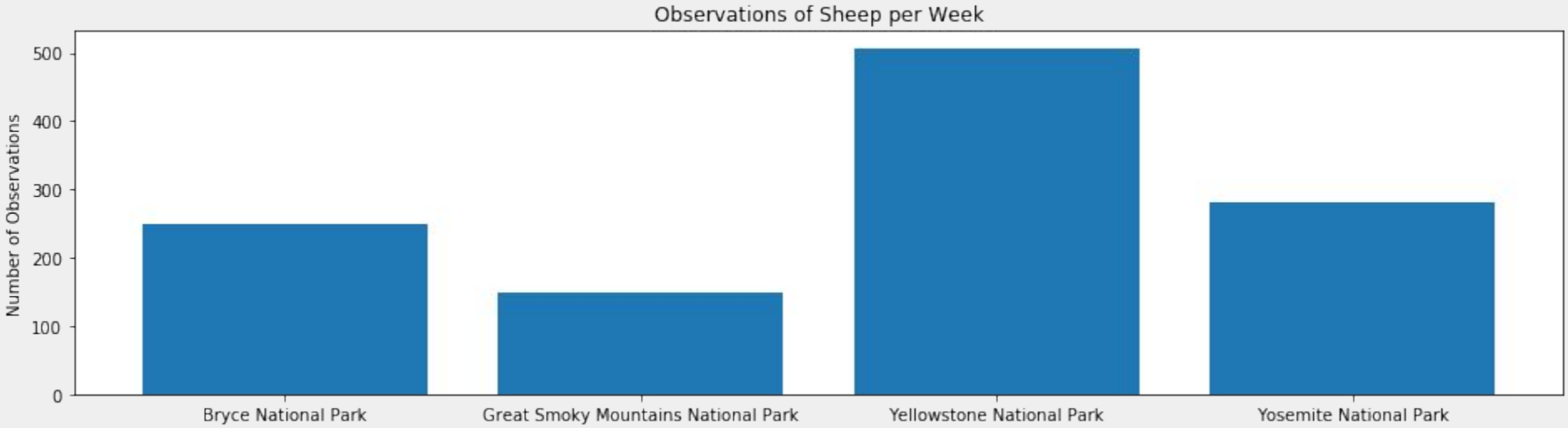
In a third test, we took a look at comparing *vascular* and *nonvascular plants* which had the lowest protection rates in the group. Surprisingly, our p-value came out to be 0.662. *Amphibians* and *fish* which shared similar protection rates also resulted in a p-value of 0.824.

With more tests having a p-value of over 0.05, does that make our conclusion more or less inconclusive? Or does simply finding a result below 0.05 an adequate answer?

Observations.csv

With this document we are analyzing the types of sheep species that were observed over a 1 week span. We are combining the numbers from four different national parks to determine whether any sheep were more prone to ‘foot-and-mouth’ disease.

We filtered our previous ‘species_info.csv’ document to select only the rows and which were both mammals and contained the word ‘sheep’. We then grouped by national parks and their respective number of observations.



Last year, scientists recorded that 15% of the observed sheep population to have ;foot-and-mouth' disease. Using that as our baseline, we concluded that our minimum detectable effect is 33.3% and our sample size per variant would be 870.

We were able to determine how long it would take to properly observe sheep for signs of the disease. For example, at Yellowstone it would take 1.7 weeks of observation. At Bryce, nearly double that rate at 3.48 weeks.