BioDiversity Analysis

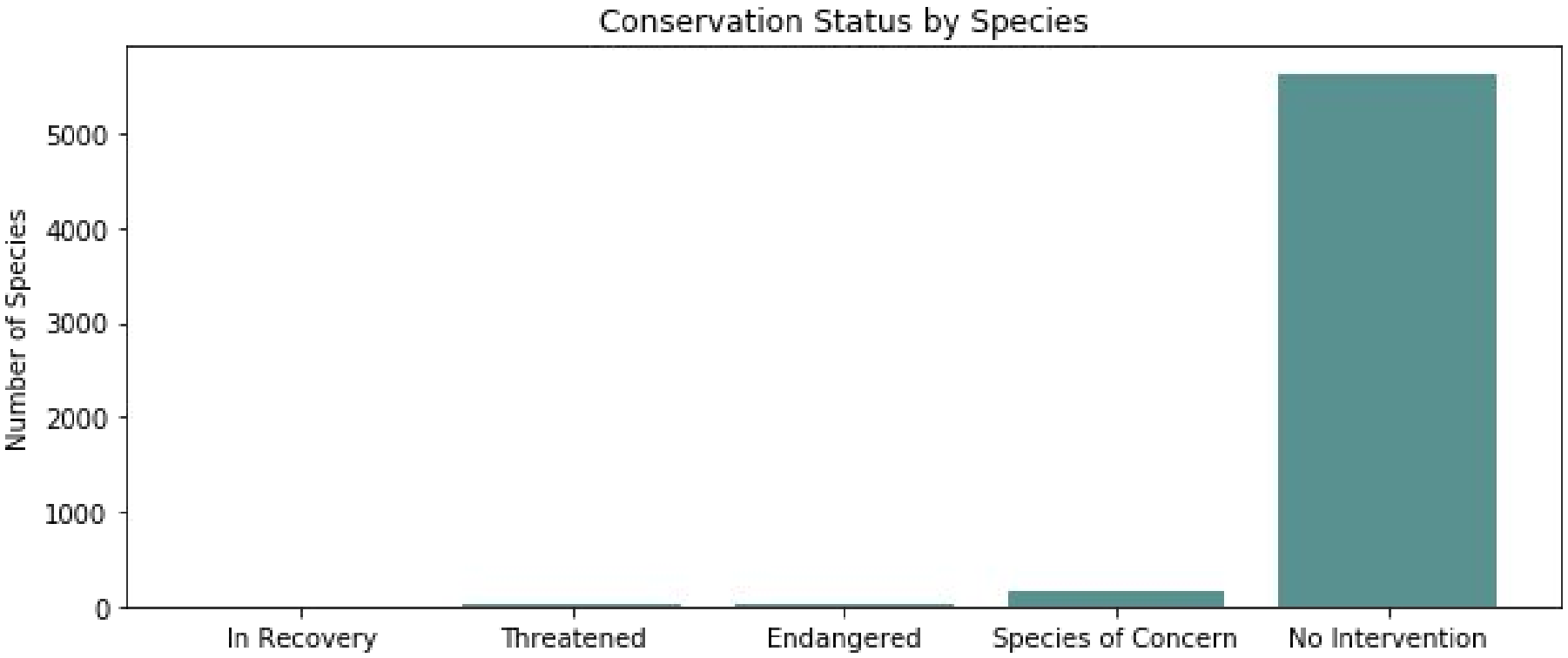
Capstone by Sharmeen Islam

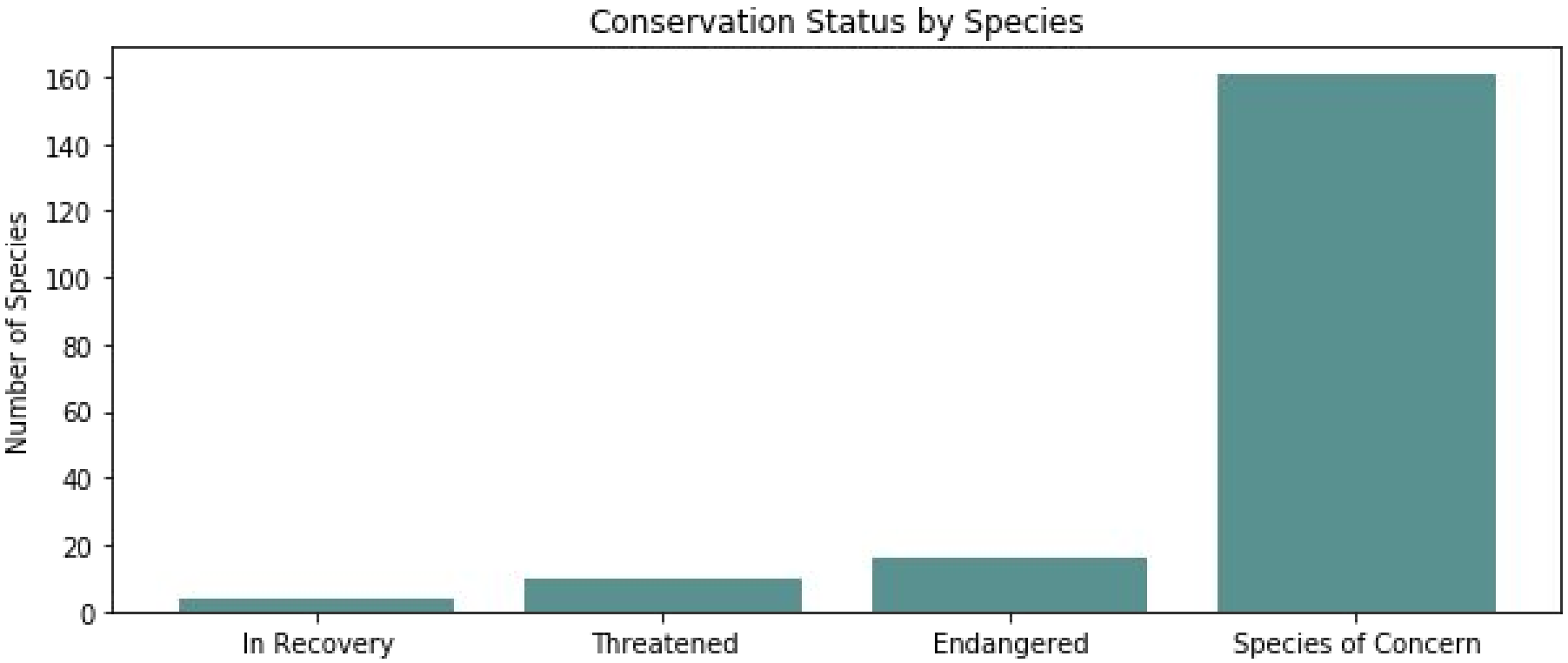
The DataSet

After converting the species_info.csv file into a Data Frame labeled “species”, I was able to look at the various table headings (category, scientific name, common names, and conservation status) as well as adding new column headings (is protected to separate at risk species from those that need no intervention, as well as if a species was a sheep for the use in the later Observations portion).

In this analysis, we looked at data collected on 5,541 species split into 7 categories (Mammal, Bird, Reptile, Amphibian, Fish, Vascular Plant, and Nonvascular Plant) with 5 conservation statuses (Species of Concern, Threatened, Endangered, In Recovery, and No Intervention).

The data showed that while there are about 180 species in need of some form of intervention, there are 5,363 species in no need of intervention at all.


Conservation Status by Species
Number of Species
5000
4000
3000
2000
1000
0
In Recovery
Threatened
Endangered
Species of Concern
No Intervention



To get a better view of the number of species that qualify for intervention, here are the bars not clearly visible in the previous slide.

Endangered Species

To highlight the relationship between categories of species to determine if there is significance in relation, we compared the number of protected Mammals to Birds, and a comparison between Reptiles and Mammals.

We used Chi Square Test to determine significance.

Observation: the numbers generated in Jupyter Notebook were different from Codecademy. I was informed that Codecademy had an error in script when calculating the number of protected species, so the advisors suggested clicking through until offered the “Get Code” option.

Pval = 0.688

Mammals versus Birds

pval> 0.05

There is no significance in the difference between the two categories.

Pval = 0.038

Reptiles versus Mammals

pval<0.05

The difference between Reptiles and Mammals is significant.

Pval = 0.053

Birds versus Reptiles

pval>0.05

Recommendation

While the percentage of Mammals population in protection (17.05%) is greater than Birds (15.37%) and Reptiles (6.41%), the Chi Square Tests showed that Mammals are no more likely to become endangered than Birds. However, the comparison between Reptiles and Mammals show that Reptiles are much more likely to become endangered than Mammals. Further analysis showed that the difference between Reptiles and Birds is also not significant, which leads me to believe that Birds are also at little risk of becoming endangered compared to Reptiles.

This analysis leads me to recommend that conservationists focus on programs that protection Reptiles over ones that focus on Birds and Mammals. This could include more funding, placing more Reptiles in conservations/zoos, or increasing education on the treatment of Reptiles to preserve the numbers that are left.

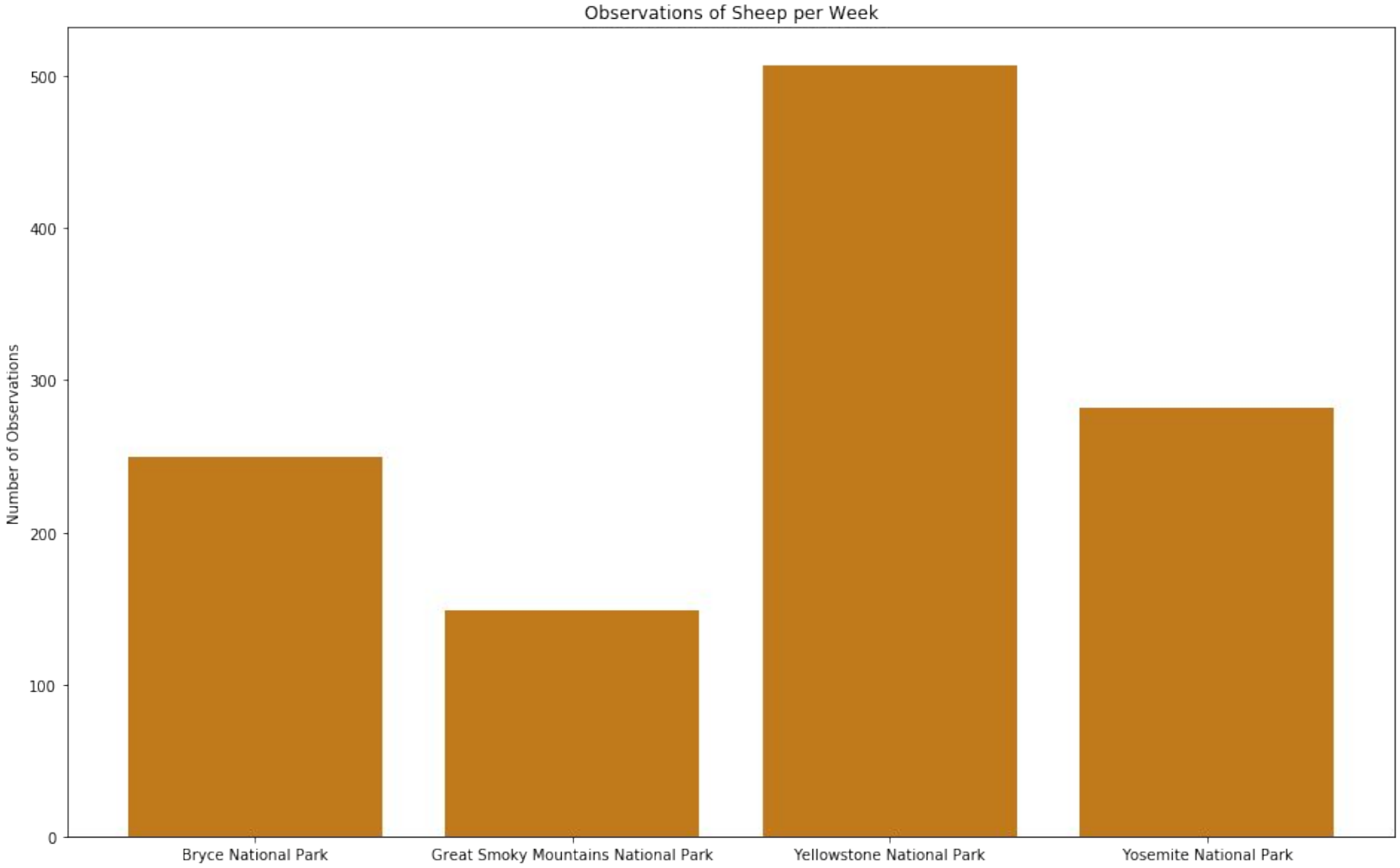
Observations of Sheep

Sample of Sheep

The next portion of analysis was focused on observations over a week made of sheep in 4 national parks (Bryce, Great Smoky Mountain, Yellowstone, and Yosemite) to extrapolate occurrences of Foot and Mouth disease in order to implement preventive measures. This is of importance since there are sheep that are protected and at risk of being endangered.

This involved merging the species data we had with park observation data to determine which sheep have been observed, where and whether they were a protected species.

Based on observations of sheep made at Bryce National Park, we determined sample population through calculations made on Optimizely. The baseline was set at 15, the minimum detectable effect was set to 33, and the sample size was determined to be 520 sheep.


Observations of Sheep per Week
Number of Observations
0
100
200
300
400
500
Bryce National Park
Great Smoky Mountains National Park
Yellowstone National Park
Yosemite National Park

In order to determine if efforts made at Yellowstone National Park to decrease the occurrence of Foot and Mouth Disease were bringing about significant change, the observations would have to run for nearly a week in Yellowstone and nearly 2 weeks in Bryce to achieve the desired sample size of 520 sheep in each park.

Reflection

I originally did this assignment on Codecademy, and then tried to replicate it in Jupyter Notebook. I found it easier to manipulate the data in Jupyter, especially when it came to changing the graphs for clarification of data and aesthetics.

At the same time, I appreciated the instant feedback on Codecademy if the code I had written was correct. The advisors were very helpful at all hours of the day and night to troubleshoot issues. Since this was my very first data analytics project, the instant feedback and help made it easier to understand what I needed to do for this project.

The main issue I came across during this project was the discrepancies that occured between the two platforms due to programing issues and data being incorrect on Codecademy. I had an issue first with determining pvals for the first part, and in the second the native calculator for determining sample size gave a completely different answer than Optimizely. Being able to work with all the correct information in Jupyter was very helpful in understanding the data and what recommendations to make.