Capstone Option 2:

Biodiversity for the National Parks

By Allison Dyckman

Part 1: Species Conservation Status

What kind of information is contained in this DataFrame?

1. The scientific name of each species
2. The common names of each species
3. The type of species
    - Mammal
    - Bird
    - Reptile
    - Amphibian
    - Fish
    - Vascular Plant
    - Nonvascular Plant

What kind of information is contained in this DataFrame?

4. The species conservation status
    - Endangered
    - In Recovery
    - No Intervention
    - Species of Concern
    - Threatened

5363 rows in our data frame contained a null value for `conservation_status`. We replaced all of these entries with “No Intervention,” which is more descriptive and provides a more accurate representation of the data.

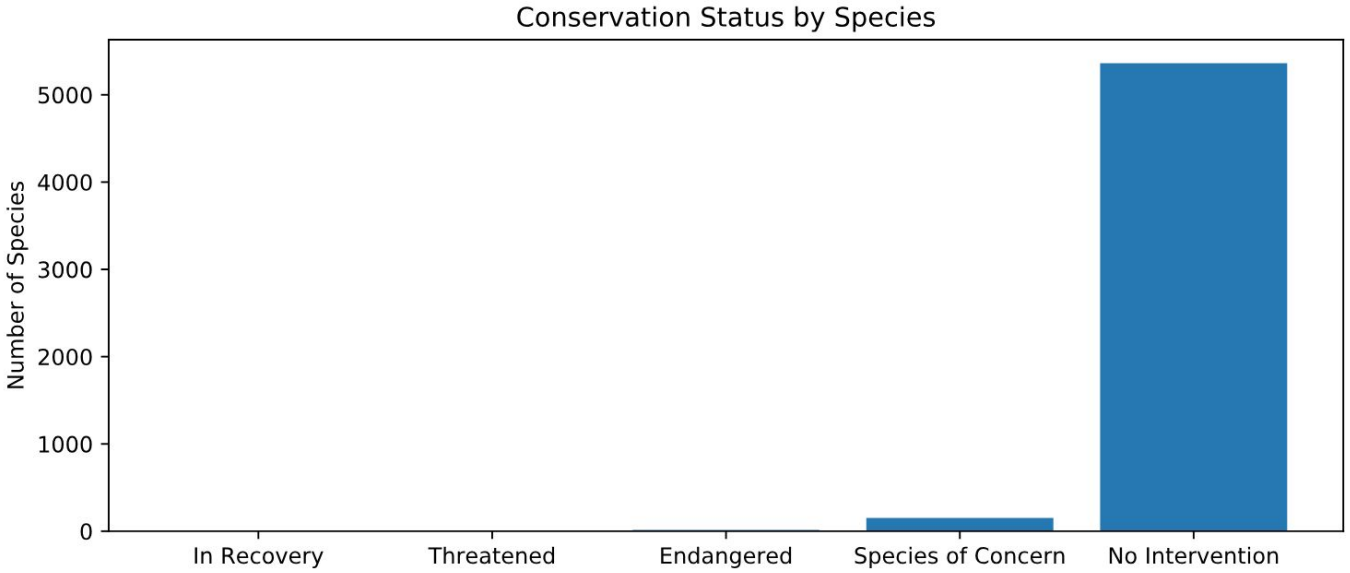
Species Conservation Status

Using this information, we calculated and sorted the total number of species per conservation status:

conservation_status	Number of Species
In Recovery	4
Threatened	10
Endangered	15
Species of Concern	151
No Intervention	5363

Graph 1 - Conservation Status by Species

We then plotted this information into a bar chart:



Species Conservation Status

Then, we created a new column in our `species` dataframe to identify whether a species' `conservation_status` indicates "No Intervention" or is protected.

From this new column, we were able to examine how many species per category were protected and how many were not protected.

Species Conservation Status

We pivoted the data to improve readability and calculated the percentage of each category that is protected.

category	not_protected	protected	percent_protected
Amphibian	72	7	8.860759
Bird	413	75	15.368852
Fish	115	11	8.730159
Mammal	146	30	17.045455
Nonvascular Plant	328	5	1.501502
Reptile	73	5	6.410256
Vascular Plant	4216	46	1.079305

Species Conservation Status

We then used a chi-squared test to determine whether or not there was a significant difference between the number of endangered mammals and the number of endangered birds.

We created the following contingency table to compare the data:

```
contingency = [[30,146], [75,413]]
```

	protected	not_protected
Mammal	30	146
Bird	75	413

Species Conservation Status

After running the chi-squared test, we return a p-value of **0.687594809666**.

As this number is greater than 0.05, we can conclude that the difference between the number of protected mammals and protected birds is **NOT** significant.

However, when we run the chi-squared test between mammals and reptiles, we see that there **IS** a significant difference as the p-value is equal to **0.0383555902297**.

Species Conservation Status

Therefore, we can confidently answer **YES** to the following question:

Are certain types of species more likely to be endangered?

Species Conservation Status

The data shows that certain types of species are more likely to be endangered than others.

Therefore, the species with the greatest likelihood of becoming endangered should be identified and prioritized by conservationists. Focus should be targeted on the types of species that have a greater probability of needing protection.

Part 2: Foot and Mouth Reduction Effort

Foot and Mouth Reduction Effort

From the information gathered last year by scientists at Bryce National Park, we knew that the baseline percentage of this sample size determination was 15%. Additionally, we knew to test for 90% statistical significance.

In order for scientists to confidently know whether their foot and mouth reduction program was working, they wanted to be able to detect reductions of at least 5 percentage points.

We used this information to calculate the Minimum Detectable Effect as follows:

$$\frac{100 * 5}{15} = 33.333333\%$$

Foot and Mouth Reduction Effort

Plugging this information into a sample size calculator, we determined that the sample size needed per variant for our experiment was **870**.

Foot and Mouth Reduction Effort

Next, we needed to determine how many observations of a particular species (sheep) had been made per each of the following national parks:

- Bryce National Park
- Great Smoky Mountains National Park
- Yellowstone National Park
- Yosemite National Park

Foot and Mouth Reduction Effort

Examining the observations.csv dataframe, we saw that it includes the following information:

- The scientific name of each species
- The park name where that species was observed
- The number of observations of that species at that park

However, it did not include the common name of each species. We had to go back to the `species` dataframe for this information.

Foot and Mouth Reduction Effort

We applied a lambda function to the `species` dataframe to determine whether the `common_names` column contained ‘sheep’. A new column was created that indicates `True` if the `common_names` column includes ‘sheep’ and `False` if it does not.

From there, we were able to select all of the rows that return a value of `True` within the new `is_sheep` column.

Foot and Mouth Reduction Effort

However, we noticed that many of the results in this new table were actually plants.

As we were only interested in the mammals that include ‘sheep’ in their common name, we selected only the rows that return a value of `True` within the new `is_sheep` column AND have a `category` of ‘Mammal’.

```
sheep_species = species[(species.is_sheep) & (species.category == 'Mammal')]
```

Foot and Mouth Reduction Effort

Now we had only the mammals whose `common_names` included ‘sheep’.

category	scientific_name	common_names	conservation_status	is_protected	is_sheep
Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True
Mammal	Ovis canadensis	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True
Mammal	Ovis canadensis sierrae	Sierra Nevada Bighorn Sheep	Endangered	True	True

Foot and Mouth Reduction Effort

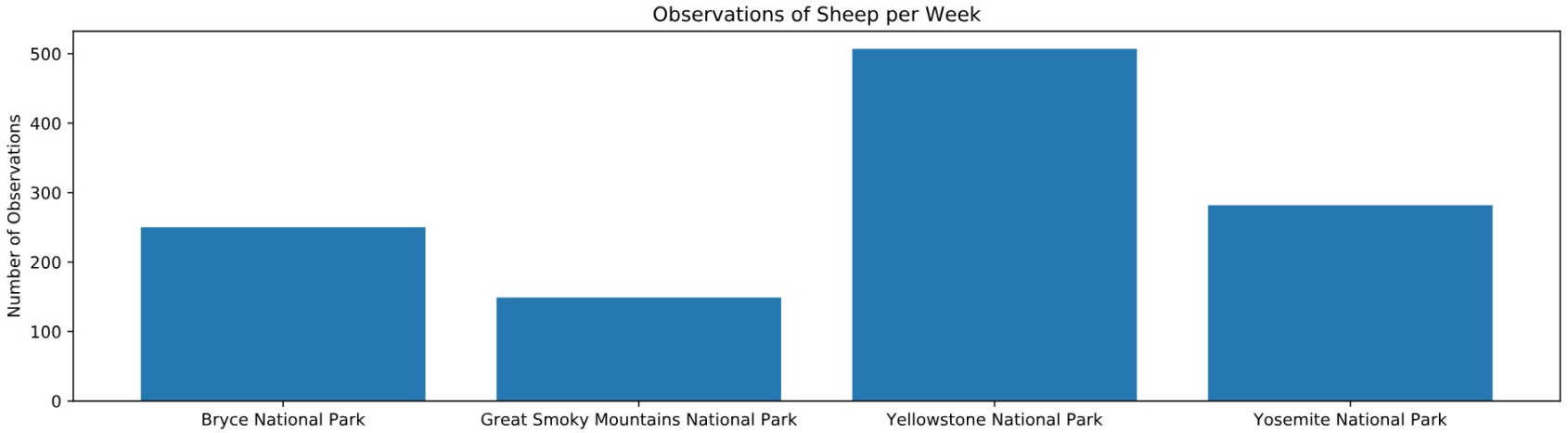
Next, we merged this new table with the `observations` dataframe. Each row then included the national park and number of observations.

We then grouped by `park_name` and the sum of `observations` for each `park_name` to view the total number of sheep observed in each park over the past 7 days.

park_name	observations
Bryce National Park	250
Great Smoky Mountains National Park	149
Yellowstone National Park	507
Yosemite National Park	282

Graph 2 - Observations of Sheep per Week

We then plotted this information into a bar chart:



Foot and Mouth Reduction Effort

Finally, we wanted to know how many weeks of observation at each park it would take to meet our sample size of **870**.

By dividing our sample size by the number of observations in a week at each park, we determined how many weeks of observation must occur for our study.

Park Name	Weeks of Observation to Achieve Sample Size
Bryce National Park	3.48
Great Smoky Mountains National Park	5.8389261745
Yellowstone National Park	1.71597633136
Yosemite National Park	3.08510638298

Thank you!