Capstone Project: Biodiversity in National Parks

Analyze data about endangered species.

By: Bibek Shah Shankhar

National Parks: A Multiuse Place

- National parks serve a two-fold purpose:

1. Outdoor recreation
2. Ecological conservation
    - Protection of animals
    - Protection of plants
    - Protection of habitat

National Parks: A Species Centric Look

- Using the species_info.csv, we can get a closer look into the species currently calling our National Parks home. Here is a sampling of just the first couple rows of data:

	category	scientific_name	common_names	conservation_status
0	Mammal	Clethrionomys gapperi gapperi	Gapper's Red-Backed Vole	NaN
1	Mammal	Bos bison	American Bison, Bison	NaN
2	Mammal	Bos taurus	Aurochs, Aurochs, Domestic Cattle (Feral), Dom...	NaN
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	NaN
4	Mammal	Cervus elaphus	Wapiti Or Elk	NaN

National Parks: A Species Centric Look cont.

- A more detailed view of the variety of the data that can be found in the species dataframe is as follows:

```
In [5]: species["category"].value_counts()

Out[5]: Vascular Plant      4470
        Bird                 521
        Nonvascular Plant    333
        Mammal               214
        Fish                 127
        Amphibian             80
        Reptile               79
        Name: category, dtype: int64
```


```
In [5]: species["category"].value_counts()

Out[5]: Vascular Plant      4470
        Bird                 521
        Nonvascular Plant    333
        Mammal               214
        Fish                 127
        Amphibian             80
        Reptile               79
        Name: category, dtype: int64
```


```
In [4]: species.nunique()

Out[4]: category           7
        scientific_name    5541
        common_names       5504
        conservation_status   4
        dtype: int64
```

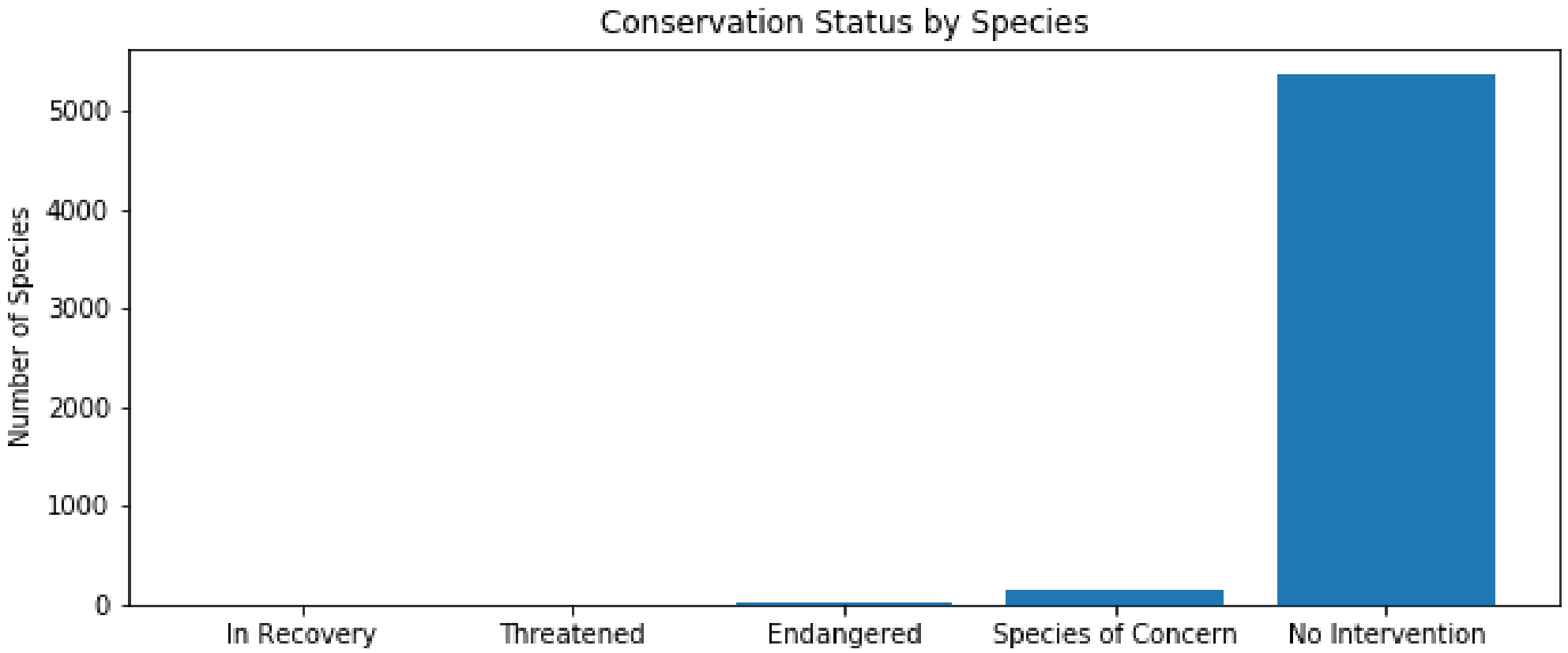
National Parks: A Species Centric Look cont.

- In regards to this analysis, one of the more important data columns is that of each species "conservation_status". With the passage of the Endangered Species Act, species were classified, based on a variety of criteria, as to whether they needed protection to survive.

Out[9]:

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

Conservation Status By Species



Conservation Status By Species cont.

From the last two slides, it is clear to see that:

The species listed under “No Intervention” comprised most of the dataset

- At 96.8%

“Species of Concern” had the second highest representation

- At 2.75

Conservation Status By Category

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

- Let's pivot and look if there is any significant difference between conservation status amongst the various groups/categories. Here is a pivot table:

Conservation Status By Category cont.

A couple takeaways from that pivot table on the previous slide:

- The number of protected species in each category is not uniform
- The two groups, at least by visual inspection, that have a much higher likelihood of being protected are:
  - Mammals
  - Birds

    We need perform a significance test

Conservation Status By Category cont.

Significance Test:

Utilized a chi squared test

- When comparing mammals and birds, the chi squared test reveals:
  - (0.1617014831654557, 0.6875948096661336) => This difference is not significant
- When comparing mammals and reptiles, the chi squared test reveals:
  - (4.289183096203645, 0.038355590229698) => This difference is significant
    - We can conclude that mammals are much more likely to be protected than reptiles

National Parks: Sheep Observations

- The “observations.csv” consists of sighting data for various species throughout the National Parks system.
- Here is a cursory view of the dataframe:

Out[24]:

	scientific_name	park_name	observations
0	Vicia benghalensis	Great Smoky Mountains National Park	68
1	Neovison vison	Great Smoky Mountains National Park	77
2	Prunus subcordata	Yosemite National Park	138
3	Abutilon theophrasti	Bryce National Park	84
4	Githopsis specularioides	Great Smoky Mountains National Park	85

National Parks: Sheep Observations cont.

Recently, there has been an outbreak of foot and mouth disease among various sheep species. So the getting an idea of the number of sheep in each park is paramount.

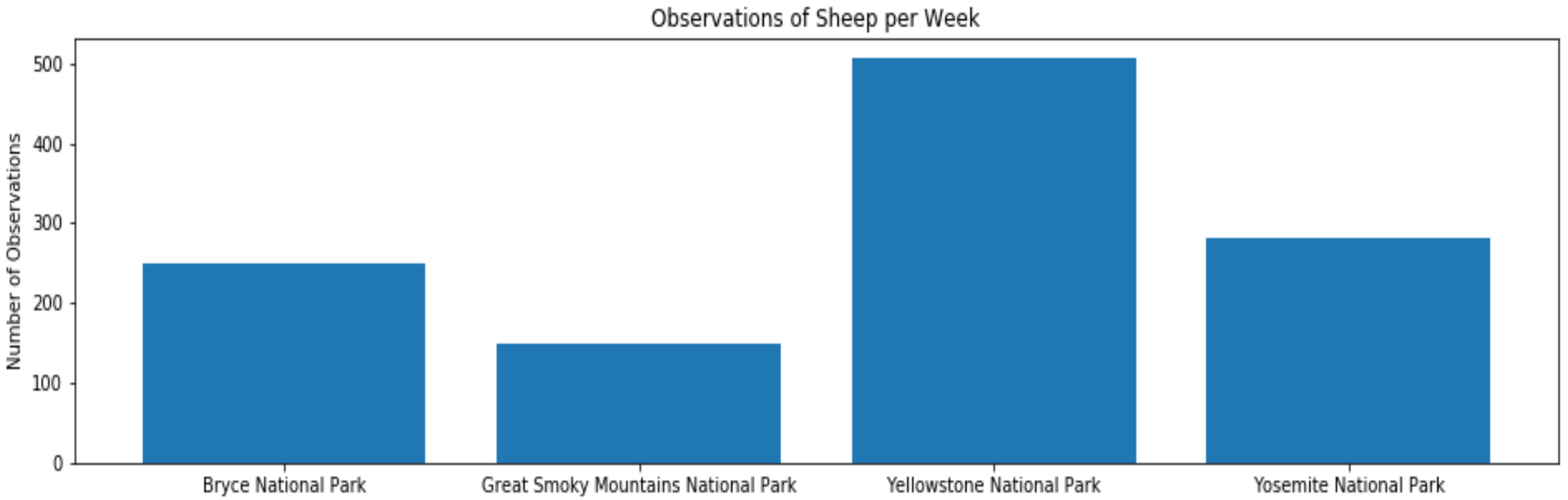
In order to get an idea of the number of sighting of sheep per park, we need to perform a couple data manipulations:

- Filter the species data for just species of sheep
- Merge the observations data with the filtered sheep dataframe
- Finally group by “park_name” to reveal the total number of sheep sighting for each park

Out[31]:

	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

National Parks: Sheep Observations cont.



National Parks: Sheep Observations cont.

- A graphical representation of the same data as the previous slide. Yellowstone NP has by far the most sighting of sheep at 507.

National Parks: Sheep Observations cont.

- Unfortunately, there have been reports that 15% of sheep at Bryce National Park have foot and mouth disease. Actions are being to mitigate the outbreak. Understandably, the scientists want test whether or not these actions working.
- They want to be able to detect reductions of at least 5 percentage points
  - Baseline percentage = 15%
  - Minimum detectable effect = 100 * 0.05 / 0.15 = 33.33
  - Calculated sample size = 870
- Using these values we can calculate the how many week s it would take to observe the needed amount of sheep
  - Bryce NP: 870 / 250 = 3.48 weeks
  - Yellowstone NP: 810 / 507 = 1.6 weeks