Biodiversity in the National Parks

An Analysis of the Data

Analysis of species_info.csv

The CSV file contains the following columns:

- Category (type of species)
- The scientific name of the species
- The common name of the species
- The conservation status of the species

	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), Domesticated Cattle	nan
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	nan
4	Mammal	Cervus elaphus	Wapiti Or Elk	nan

Other details in species_info.csv

Total number of species: 5541

Categories (types of species):

- Mammal
- Bird
- Reptile
- Amphibian
- Fish
- Vascular Plant
- Nonvascular Plant

Types of conservation status:

- Species of Concern
- Endangered
- Threatened
- In Recovery
- No Intervention (nan)

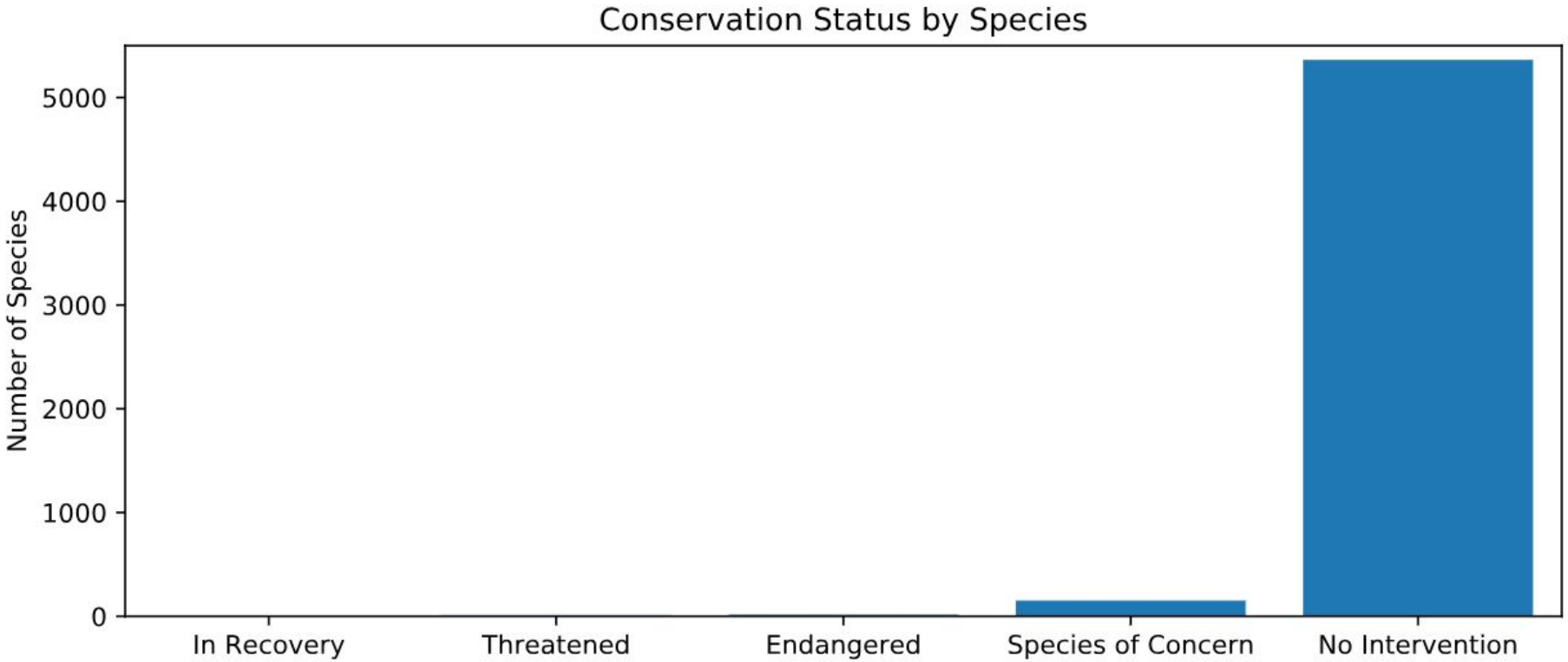
Question: How many species are in each type of conservation status?

Method: Sort the species, using scientific name to distinguish between species, by their types of conservation status and count the totals in each type of conservation status.

Result:

conservation_status	scientific_name
Endangered	15
In Recovery	4
No Intervention (nan)	5363
Species of Concern	151
Threatened	10

Result: the majority of species in the national parks require no intervention



Question: Are certain types of species more likely to be endangered?

Method: Sort the data according to species type (category) and whether the species is protected or not protected (where protected = no intervention)

Result:

category	not_protected	protected
Amphibian	72	7
Bird	413	75
Fish	115	11
Mammal	146	30
Nonvascular Plant	328	5
Reptile	73	5
Vascular Plant	4216	46

Want to know the percentage of each species that is protected

Percentage protected = # of protected species / total number of species

Result:

Mammals and Birds have the highest percentages of protected species

Potentially more likely to be endangered, but is this statistically significant?

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

Question: Are Mammals significantly more likely to be endangered than Birds?

Method: Create a contingency table to run a chi-squared test

	protected	not-protected
Mammal	30	146
Bird	75	413

Result: Pvalue for the chi-square test was 0.687594809666

Since the pvalue is not less than 0.05, the difference between the percentages of protected birds and mammals is not significant but a result of chance.

Mammals are not significantly more likely to be endangered than Birds.

Question: Is the difference between protected species of Reptile and Mammal significant?

Method: Update our contingency table with Reptiles

	protected	not-protected
Mammal	30	146
Reptile	5	73

Result: Pvalue for the chi square test was 0.0383555902297

Since the pvalue is less than 0.05, the difference between the percentages of protected reptiles and mammals is significant.

Certain types of species are more likely to be endangered than others

Recommendation

As established through this analysis, certain types of species are more likely to be endangered than others. This can be used to help guide protection efforts to help the types of species that are more likely to require it.

However, these calculations are based on the percentage of species that are already protected, so it does not account for new species that might become threatened or endangered in the future. Additionally, if we look at just absolute numbers instead of proportions, Vascular Plants have more protected species than Mammals; plants do not deserve less attention just because there are more species of them.

So my recommendation: focus on species that are more likely to be endangered, but be aware of the other species left out of this measurement that also need help.

Analysis of observations.csv

The CSV file contains the following columns:

- The scientific name of the species
- The park name where the species was seen
- The number of observations of that species in that park in the past 7 days

	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

Tracking the movements of various sheep species

Method: Filter the rows in the species DataFrame to find all the common names that contain the word ‘Sheep’

Further filter to remove the plants and focus on only those species classified as Mammals

Result: 3 species of sheep can be found in the national parks

	category	scientific_name	common_names	conservation_status	is_protected	is_sheep
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True
3014	Mammal	Ovis canadensis	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True
4446	Mammal	Ovis canadensis sierrae	Sierra Nevada Bighorn Sheep	Endangered	True	True

Question: How many total sheep sightings (across all three species) were made at each national park?

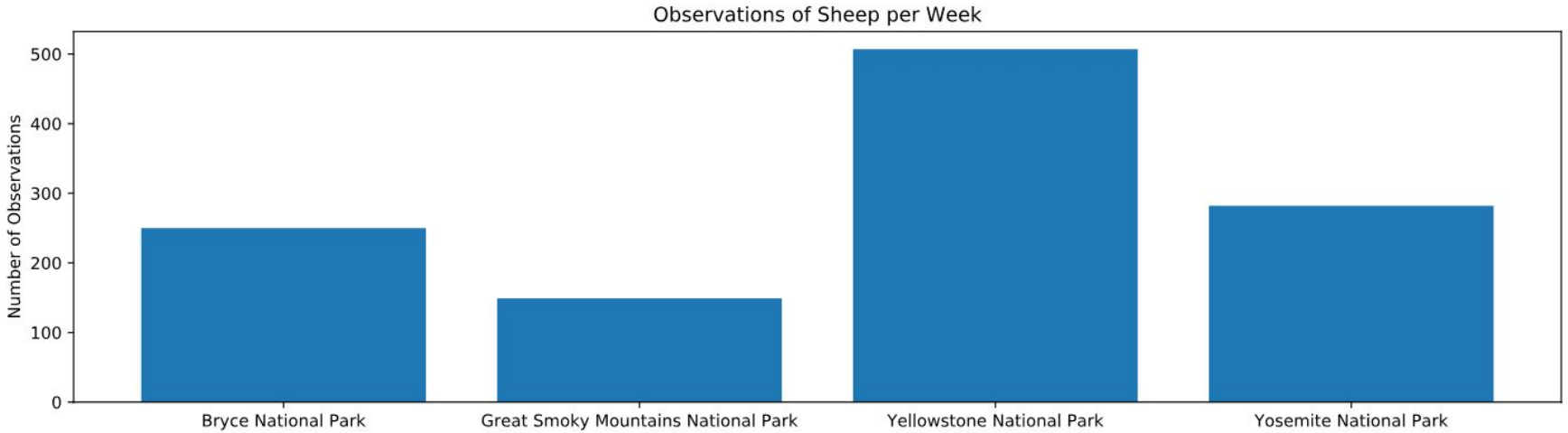
Method: Merge the information for the three sheep species with the observations DataFrame to find the number of observations for each sheep species at each national park

Group by park to get the total number of sheep observed in each national park over the past 7 days

Result:

	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

Result: Yellowstone National Park had the most sheep sightings in the last week



Question: What is the required sample size to test if the foot and mouth reduction program is working?

Goal: Want to detect reductions of at least 5% with confidence

Baseline percentage = 0.15 (or 15%)

Minimum detectable effect = 100 * 0.05 / 0.15 = 33.3333333333

Level of significance = 90%

The required sample size per variant is **870 sheep.**

Number of weeks required to observe enough sheep for the sample test:

- Yellowstone National Park = **1.716 weeks**
- Bryce National Park = **3.48 weeks**