Biodiversity in National Parks

Ming-Yuan Lu
for
Codecademy “Introduction to Data Analysis” Capstone Project
Apr 24 – Jul 17, 2018

Goal

With data from several national parks concerning the conservation status of various species, we attempt to:

- Understand the data structure and content
- Visualize the number of species in each conservation status
- Study if the protection rate demonstrates significant variation from one category to another
- Combining with observation data, we study the distribution of sheep observations across all considered nations parks
- Finally, determine the number of observations needed in two of the parks to determine, with enough confidence, if the foot and mouth disease reduction program applied at Yellowstone National Park reaches its targeted reduction rate

Species data

- Data source: species_info.csv

First 10 entries

	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
5	Mammal	Odocoileus virginianus	White-Tailed Deer	NaN
6	Mammal	Sus scrofa	Feral Hog, Wild Pig	NaN
7	Mammal	Canis latrans	Coyote	Species of Concern
8	Mammal	Canis lupus	Gray Wolf	Endangered
9	Mammal	Canis rufus	Red Wolf	Endangered

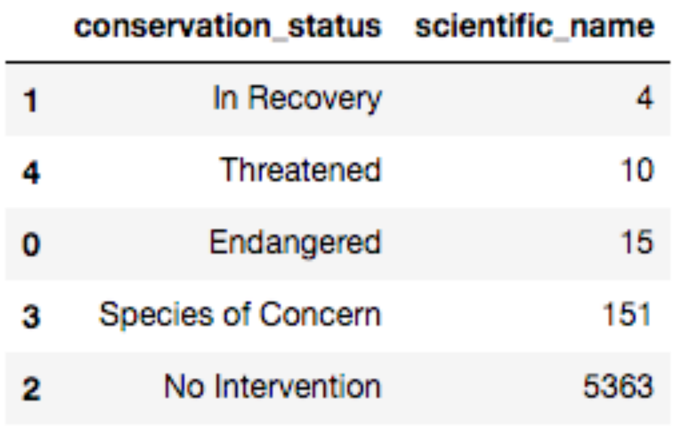
DataFrame properties:
1. `Number of entries:` 5824
2. `Number of columns:` 4
3. `Column types:`
“category” 5824 non-null object
“scientific_name” 5824 non-null object
“common_names” 5824 non-null object
“conservation_status” 191 non-null object
4. `memory usage`: 182.1+ KB

Number of unique species: 5541
This suggests some duplication of the entries (rows), since the number of unique species is less than the number of rows! The data could therefore benefit from some re-organization to make sure each species occupies only one row.

Species data

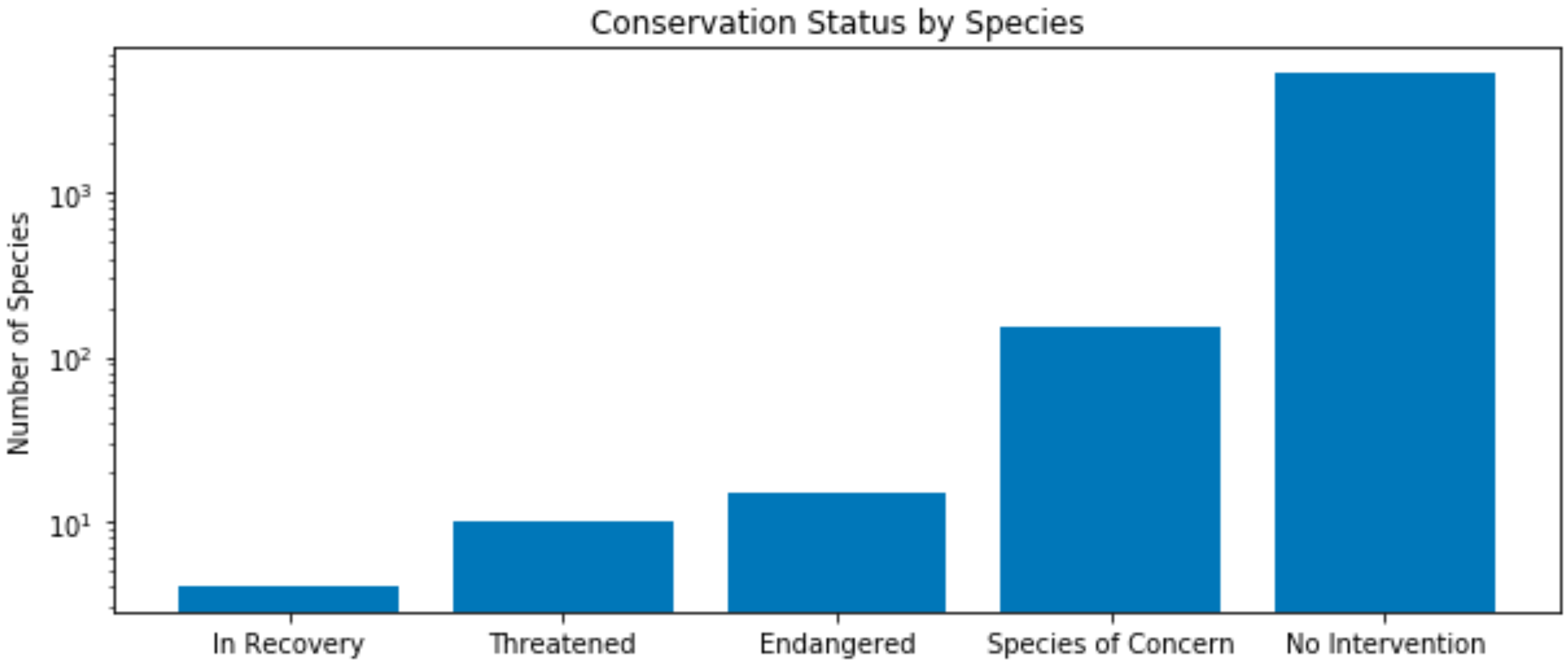
- Categories in the data:
  `Mammal`
  `Bird`
  `Reptile`
  `Amphibian`
  `Fish`
  `Vascular Plant`
  `Nonvascular plant`
- Conservation status in the data
  `nan` → indicates no intervention needed, renamed `“No Intervention”`
  `Species of Concern`
  `Endangered`
  `Threatened`
  `In Recovery`

Number of species in each conservation status



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

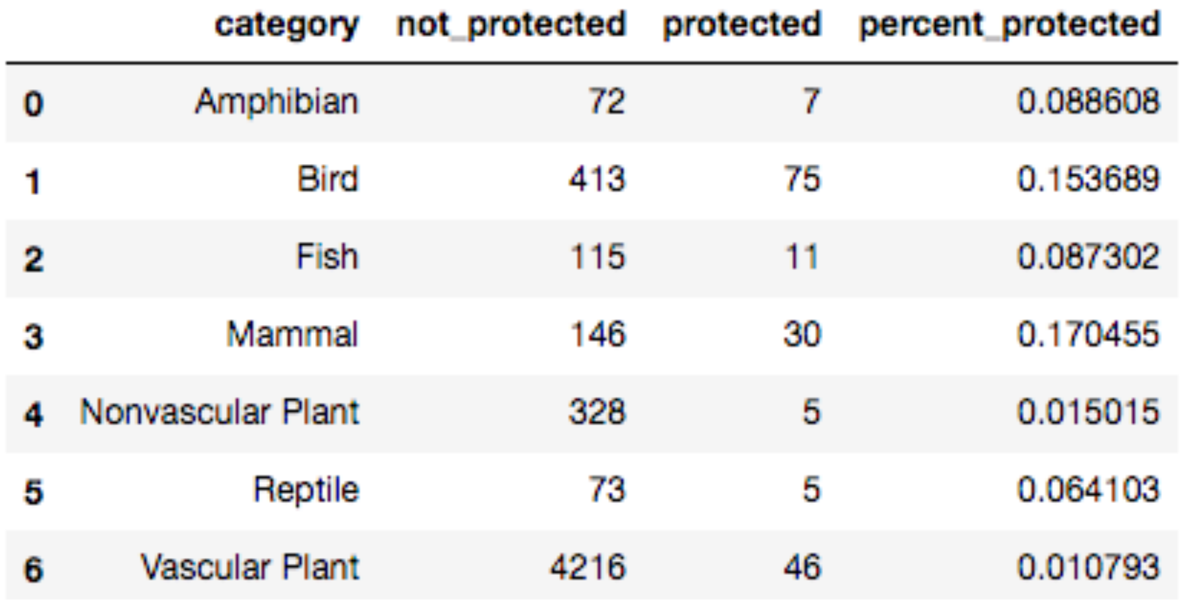
About 3.25% of all species recorded are protected



The figure is logarithmic in the y-axis for more visible low-count bins

Categorical variation in protection rate

- No intervention – not protected
  other conservation status – protected
- What’s the percentage of protected species in each 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

Birds and Mammals stand up as the most-protected categories
Vascular plants and nonvascular plants are the least-protected categories

In fact, we can almost group the data into 3 groups
1. [bird, mammal]: ~15-17% protection
2. [amphibian, fish, reptile]: ~6-9% protection
3. [vascular plant, nonvascular plant]: ~1% protection

This warrants, for future conservation studies, that distinct treatments/approaches to conserving species in these groups may be needed.

Are the different rates significant?

- Following the observation from the last slide, we want to know if the rate differences among categories are statistically significant
- We will test these combinations:
  [mammal, bird]
  [reptile, mammal]
  [reptile, nonvascular plant]
- Chi-square test will be applied with a significance level α=0.05
- Null hypothesis – the difference in rates is not significant
  Alternative hypothesis – the difference in rates is significant

[mammal, bird]

- We build contingency table

	Protected	Not protected
Mammal	30	146
Bird	75	413

- Chi-square test results:
  chi-square = 0.162
  p-value = 0.688 > α=0.05
- We reject the alternative hypothesis that the protection rate difference between mammals and birds is significant
- This suggests that even though the protection rate is higher for mammals than for birds, the difference is likely statistical

[reptile, mammal]

- We build contingency table

	Protected	Not protected
Mammal	30	146
Reptile	5	73

- Chi-square test results:
  chi-square = 4.289
  p-value = 0.038 < α=0.05
- We accept the alternative hypothesis that the protection rate difference between mammals and reptiles is significant
- This suggests that more resources and efforts should be allocated to the conservation of mammals (& birds) than reptiles

[reptile, nonvascular plant]

- We build contingency table

	Protected	Not protected
Nonvascular plant	5	328
Reptile	5	73

- Chi-square test results:
  chi-square = 4.514
  p-value = 0.034 < α=0.05
- We accept the alternative hypothesis that the protection rate difference between reptiles and nonvascular plants is significant
- This suggests that more resources and efforts should be allocated to the conservation of reptiles than nonvascular plants

Observation data

- Data source: observations.csv
- Data contains observation data in several national parks in the past week from conservationists
-

First 10 entries

	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
5	Elymus virginicus var. virginicus	Yosemite National Park	112
6	Spizella pusilla	Yellowstone National Park	228
7	Elymus multisetus	Great Smoky Mountains National Park	39
8	Lysimachia quadrifolia	Yosemite National Park	168
9	Diphyscium cumberlandianum	Yellowstone National Park	250

DataFrame properties:
1. `Number of entries:` 23296
2. `Number of unique species:` 5541
3. `Number of columns:` 3
4. `Column types:`
"scientific_name" 23296 non-null object
"park_name" 23296 non-null object
"observations" 23296 non-null object
5. `Memory usage:` 546.1+ KB

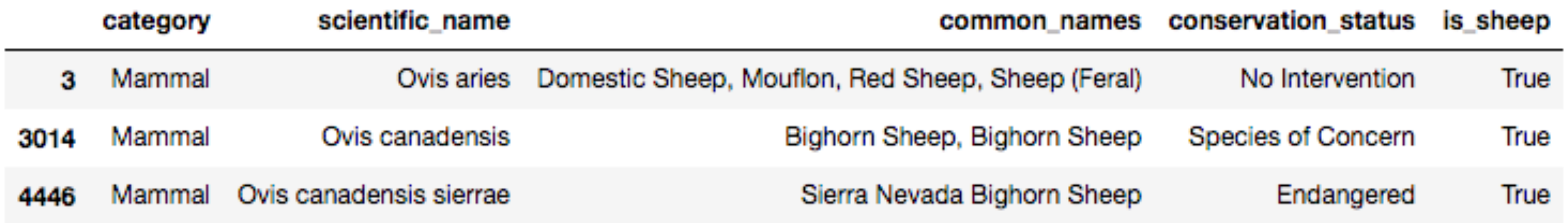
Observations in parks

- 4 nations parks:
  Great Smoky Mountains
  Yosemite
  Bryce
  Yellowstone
- Number of unique species logged in each of the parks = 5541
- Number of observations in each park:

Park name	N_{obs}
Bryce	576025
Great Smoky Mountains	431820
Yellowstone	1443562
Yosemite	863332

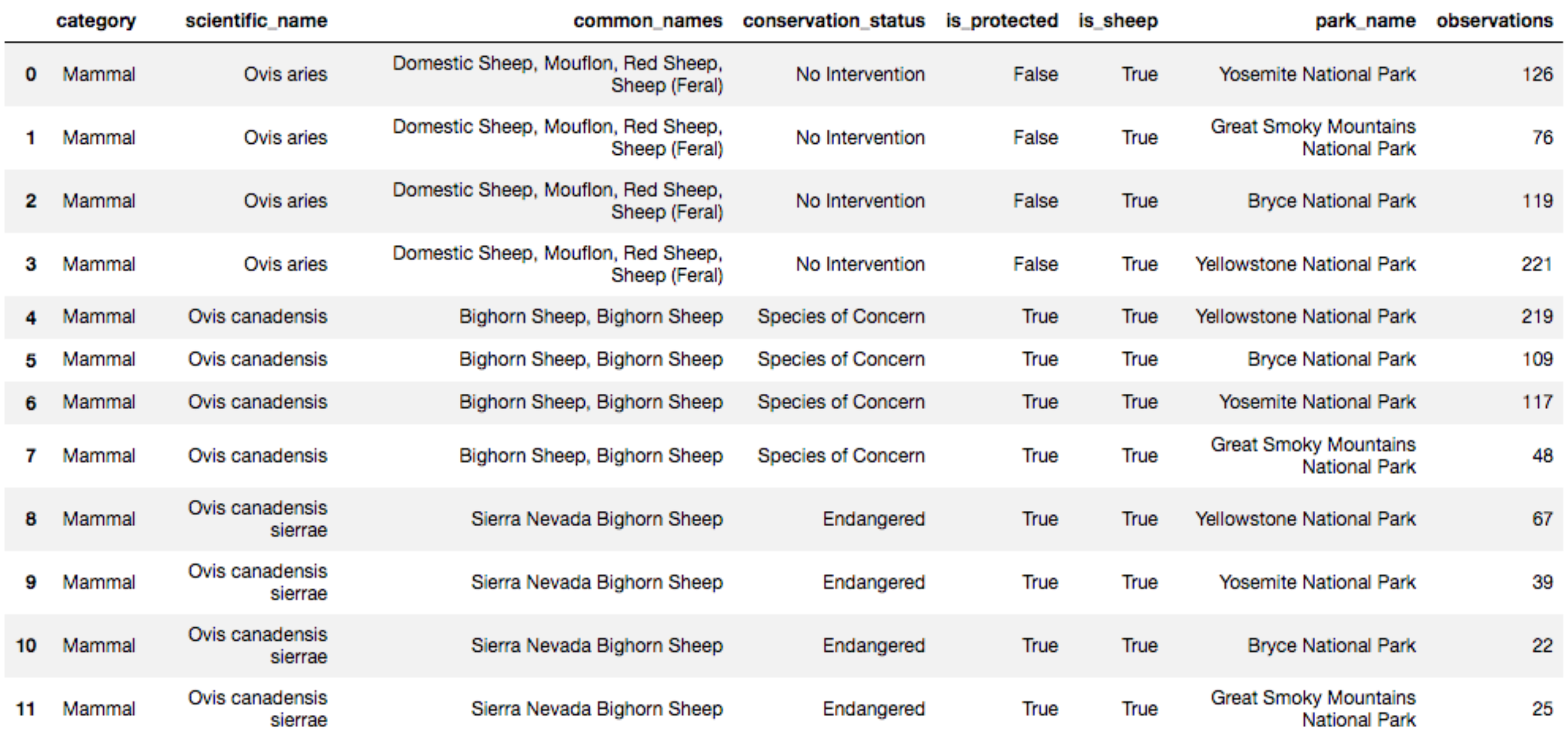
Sheep in national parks

- From species data we have 3 sheep species:



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

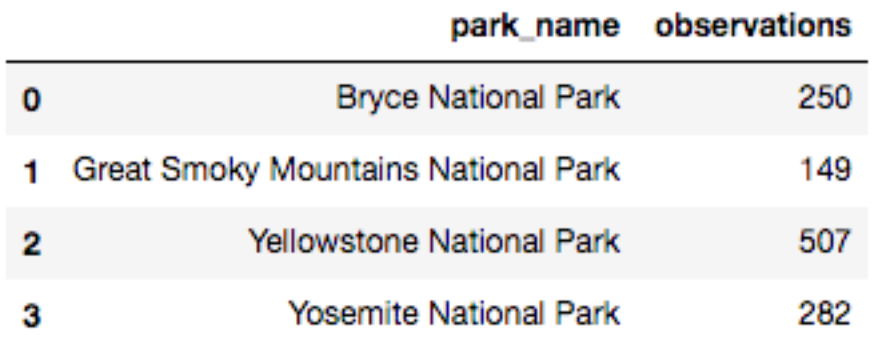
- Merging this with the observation data, we have:



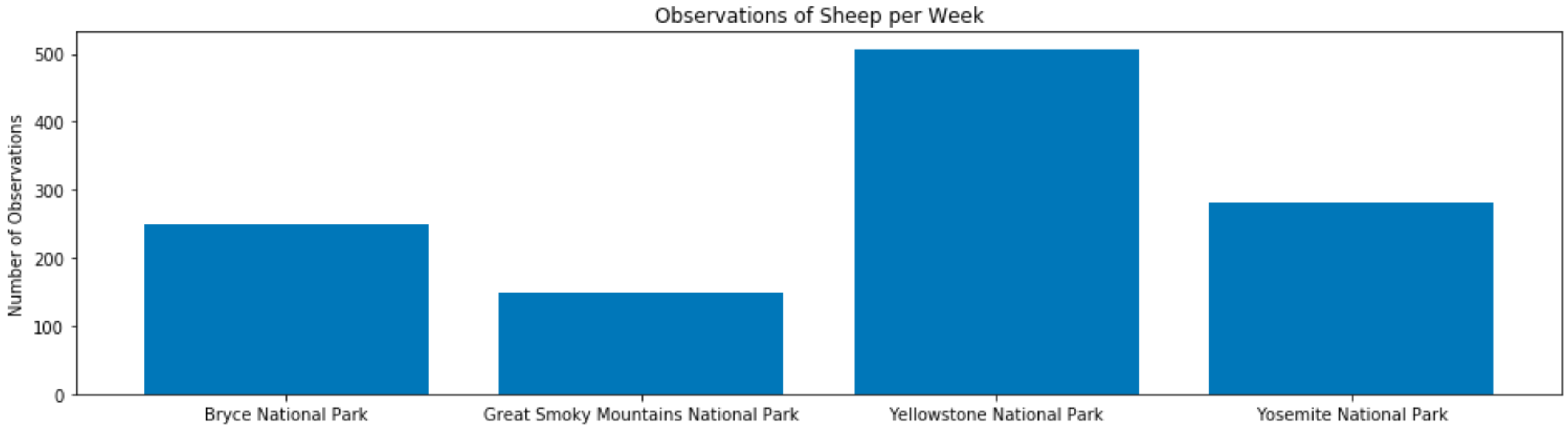
	category	scientific_name	common_names	conservation_status	is_protected	is_sheep	park_name	observations
0	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True	Yosemite National Park	126
1	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True	Great Smoky Mountains National Park	76
2	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True	Bryce National Park	119
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True	Yellowstone National Park	221
4	Mammal	Ovis canadensis	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True	Yellowstone National Park	219
5	Mammal	Ovis canadensis	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True	Bryce National Park	109
6	Mammal	Ovis canadensis	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True	Yosemite National Park	117
7	Mammal	Ovis canadensis	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True	Great Smoky Mountains National Park	48
8	Mammal	Ovis canadensis sierrae	Sierra Nevada Bighorn Sheep	Endangered	True	True	Yellowstone National Park	67
9	Mammal	Ovis canadensis sierrae	Sierra Nevada Bighorn Sheep	Endangered	True	True	Yosemite National Park	39
10	Mammal	Ovis canadensis sierrae	Sierra Nevada Bighorn Sheep	Endangered	True	True	Bryce National Park	22
11	Mammal	Ovis canadensis sierrae	Sierra Nevada Bighorn Sheep	Endangered	True	True	Great Smoky Mountains National Park	25

Sheep observation in the past week

- We can group by parks in the previous dataframe to get the number of sheep observations in each park in the last 7 days



	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



Foot and mouth disease reduction program

- 15% of sheep at Bryce National Park have foot and mouth disease
- A program to decrease the rate of the disease at Yellowstone National Park by 5% is being run
- How many sheep observation is needed to detect this effect?

	%
Baseline conversion rate	15
Minimum detectable effect	33.33
Statistical significance	90

- This gives a sample size of 510
- Given that there are 250 sheep observations in Bryce NP, and 507 observations in Yellowstone NP in the past week, we can estimate the time it takes to attain 510 observations in each park:
  Bryce: 510 / 250 = 2.04 weeks
  Yellowstone: 510 / 507 = 1.01 weeks

Conclusion

- We studied species and there conservation status in several national parks
- About 3.25% of all species recorded are protected
- Protection rate shows variation across categories. Using statistical significance test, we made suggestions for how future conservation efforts should prioritize
- Combining with observation data, we determined the sample size (number of observations) needed to test if a new disease treatment program for sheep is effective