Biodiversity in National Parks Analysis

Shawn Craig

Species_info.csv

- Base data for analysis
- Columns are category, scientific name, common names, and conservation status
  - Each row is a single record of an animal or plant in a national park
  - Category = organism type (mammal, bird, reptile, etc)
- Scientific name is just that, the scientific name for the organism
- Common name = common name
- Conservation status = whether the organism is threated, endangered, a species of concern, or in recovery
  - These are fairly straight forward, in ascending order of concern, species of concern, threated, and endangered

Species_info.csv Analysis

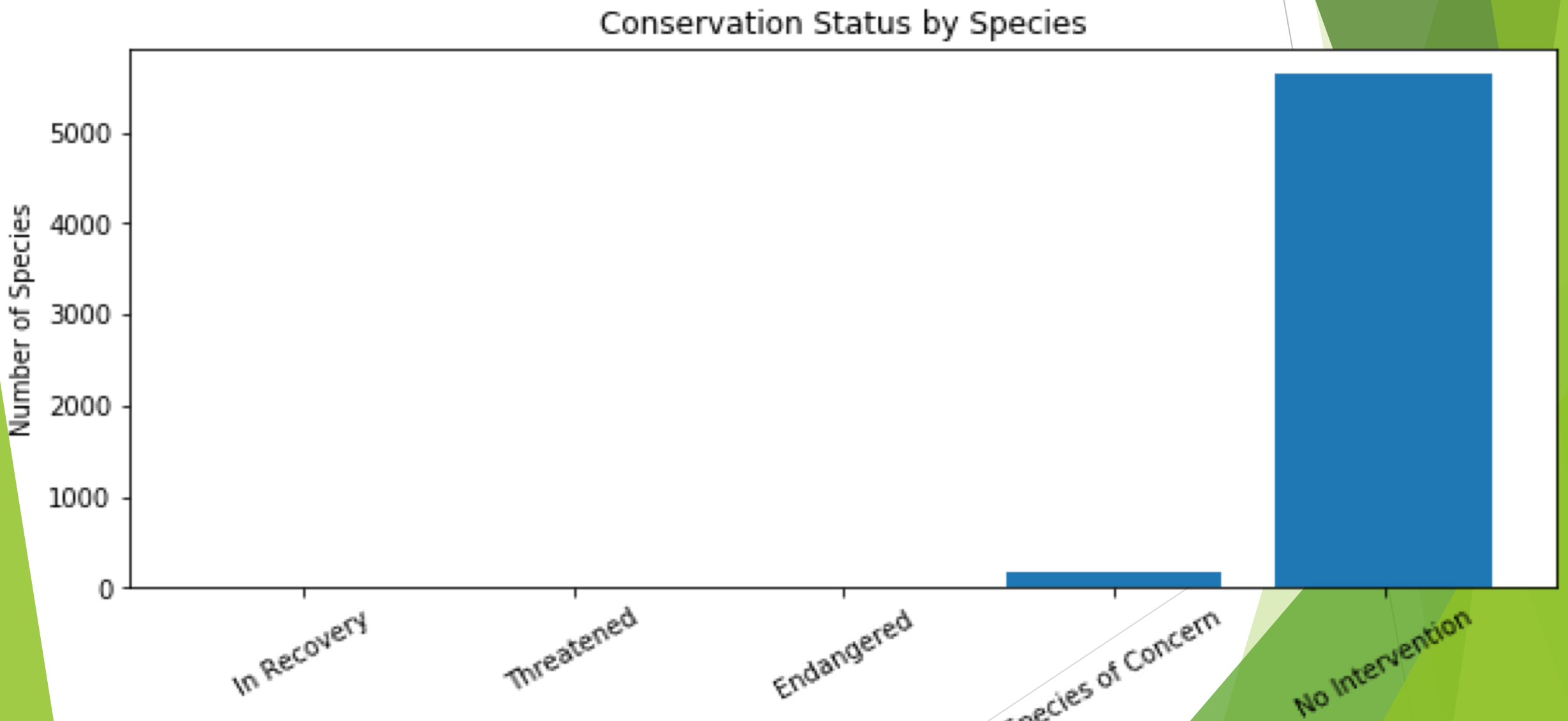
- Data was fairly complete but required minor reorganization prior to analysis
- Pandas was used with python to count unique records, determine category values, and groupby to reorder data yielded the following:
  - Data set contained 5541 unique records (species)
  - 7 categories (Mammal, Bird, Reptile, Amphibian, Fish, Vascular Plant, Nonvascular Plant)
  - Conservation statuses were: nan, Species of Concern, Endangered, Threatened, In Recovery
- Null values in conservation were taken to mean no intervention was required, so nan was replaced with ”No Intervention”

Species_info.csv Analysis II

- Reorganizing yielded the following data:

Conservation Status	Scientific Name
In Recovery	4
Threatened	10
Endangered	16
Species of Concern	161
No Intervention	5633

Conservation Status Bar Graph



Species_info.csv Analysis III

- To further determine what was happening in the data I
  - Added an ‘is protected’ column to help sort the data
  - Grouped, pivoted, renamed, added % protected to the dataframe to produce:

Category	Not protected	Protected	Percent protected
Amphibian	72	7	0.088
Bird	413	75	0.153
Fish	115	11	0.087
Mammal	146	30	0.1704
Nonvascular Plant	328	5	0.015
Reptile	73	5	0.064
Vascular Plant	4216	46	0.0107

Testing, Are Mammals more likely to be endangered than Birds?

- Because we have two or more categorical datasets the best test to apply is the Chi Squared test
  - To run a chi squared test, we first need to construct a contingency table with our data
  - contingency = [[146, 30], [413, 75]]
- Running the test we get a p value of **0.687**, which is far to large to indicate a significant difference

Testing 2, What about Reptiles compared to Mammals?

- Again, this is two sets of categorical data so I used the Chi Squared Test
  - Our contingency table was : contingency2 = [[146, 30], [73, 5]]
  - p value was 0.038
- This p value, generally anything 0.05 or lower, indicates there is a significant difference between Reptiles and Mammals
- Mammals are more likely to be endangered than Reptiles

Recommendation

- Based on the available data and my analysis:
- Conservationists should focus their efforts on Mammals
- However, birds are also threaten at nearly the same level as Mammals so initiatives that benefit both should be prioritized

Observations.csv

- Observations.csv was a dataset containing conservationists sightings of different species at several national parks over the course of a week
- We specifically focused our efforts on examining data related to sheep
- We modified our species dataframe with a new column called is_sheep
  - This column returned a True or False indicating if the word sheep was present in a each record (row)
- Using this new column we filtered our dataframe to show only records that were both mammals and sheep

Category	Scientific Name	Common Name	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

Merging Dataframes

- After reconfiguring our observations data frame we merged it with our species dataframe we filtered to display just sheep

Scientific Name	Park Name	Observations	Category	Common Names	Conservation Status	Is Protected	Is Sheep
Ovis canadensis	Yellowstone National Park	219	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True
Ovis canadensis	Bryce National Park	109	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True
Ovis canadensis	Yosemite National Park	117	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True
Ovis canadensis	Great Smoky Mountains National Park	48	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True
Ovis canadensis sierrae	Yellowstone National Park	67	Mammal	Bighorn Sheep, Bighorn Sheep	Endangered	True	True

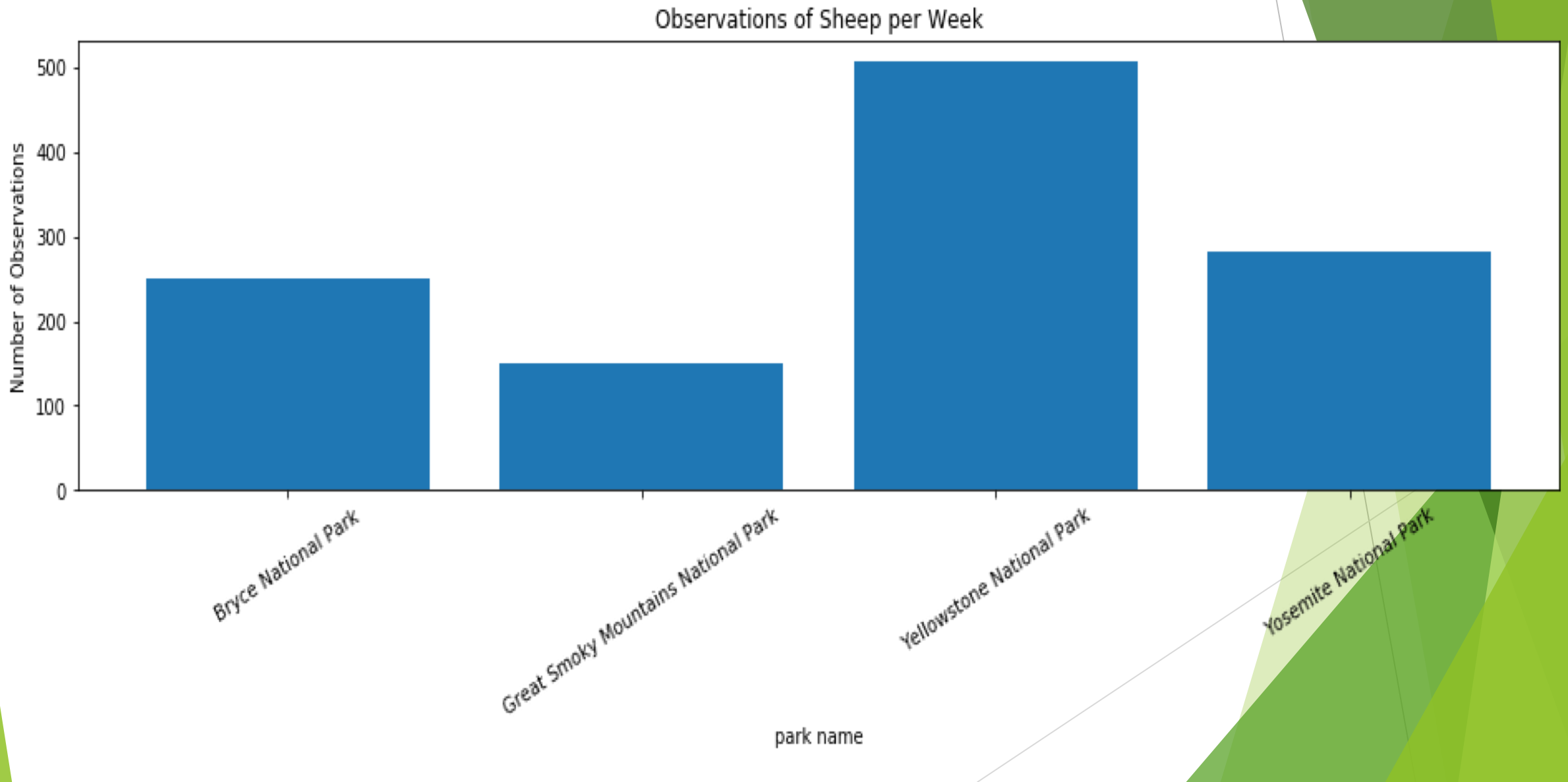
Merged Dataframe Analysis

- After merging we were able to use the groupby command to create the following dataframe, Observations by Park

Park Name	Observations
Bryce National Park	250
Great Smoky Mountains National Park	159
Yellowstone National Park	507
Yosemite National Park	282

- This dataframe was plotted to create the chart found on the next slide

Observations by Park



Foot and Mouth Disease Rates

- Scientists have reported that roughly 15% of the observed sheep in Bryce National Park have foot and mouth disease. Further it's estimated 10% of the population at Yellowstone have the disease as well.
- Scientists at Bryce have instituted a program to reduce the rate of foot and mouth disease
- Based on our observation numbers and the rates of disease supplied to us by park scientists I was able to calculate the required sample size to test whether their disease eradication program was having any effect

Foot and Mouth Disease II

- Using a sample size calculator, and a significance of 90%, I determined a sample size of **520** is required for Bryce National Park
- Using the same methods, a sample size of **370** would be required for Yellowstone National Park
- Using the previously supplied data regarding observations per week:
- It will take roughly **two weeks** to determine whether the program is having any effect in Bryce National Park
- It will take roughly **1 week** (about 5 days actually) to determine whether the program is having any effect in Yellowstone National Park