Biodiversity in US National Parks

Goal

The goal of this project to is analyze biodiversity data in US National Parks to better understand the distribution of the conservation status of species in these parks.

The project sought to answer the following questions:

- What is the distribution of conservation status for species?
- Are certain types of species more likely to be endangered?
- Are the differences between species and their conservation status significant?
- Which species are In Recovery and what is the overall distributions in the parks compared to the average observations of the species with other conservation statuses?

Data Sources

- species_info.csv
- observations.csv

Note: The data for this project is inspired by real data but is mostly fictional.
The data was provided by Codecademy.

species_info.csv

This is the file that contains information pertaining to the various species in the National Parks.

It has four columns:

- **category**: The category of taxonomy for each species
- **scientific_name**: The scientific name of the species
- **common_names**: The common name(s) of the species
- **conservation_status**: The conservation status of the species ('Species of Concern' 'Endangered' 'Threatened' 'In Recovery'). If null, then the species is not in danger

The file consists of 5824 rows and 4 columns.

species_info.csv

```
Number of Rows:5824
Number of Columns:4
```

	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

species_info.csv

	category	scientific_name	common_names	conservation_status
count	5824	5824	5824	191
unique	7	5541	5504	4
top	Vascular Plant	Castor canadensis	Brachythecium Moss	Species of Concern
freq	4470	3	7	161

species_info.csv

	category	scientific_name	common_names	conservation_status
count	5824	5824	5824	191
unique	7	5541	5504	4
top	Vascular Plant	Castor canadensis	Brachythecium Moss	Species of Concern
freq	4470	3	7	161

species_info.csv

The category column has the following 7 unique values:

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

```
category
Amphibian            80
Bird                521
Fish                127
Mammal              214
Nonvascular Plant   333
Reptile              79
Vascular Plant     4470
dtype: int64
```

species_info.csv

The conservation_status column has the following 4 unique values:

- **Species of Concern**: declining or appear to be in need of conservation
- **Endangered**: seriously at risk of extinction
- **Threatened**: vulnerable to endangerment in the near future
- **In Recovery**: formerly Endangered, but in recovery of endangerment

```
conservation_status
Endangered            16
In Recovery            4
Species of Concern   161
Threatened            10
dtype: int64
```

There are 5633 records that have a null value in conservation_status. These signify that the species is not a protected species.

species_info.csv

	category	scientific_name	common_names	conservation_status
count	5824	5824	5824	191
unique	7	5541	5504	4
top	Vascular Plant	Castor canadensis	Brachythecium Moss	Species of Concern
freq	4470	3	7	161

species_info.csv

	category	scientific_name	common_names	conservation_status
count	5824	5824	5824	191
unique	7	5541	5504	4
top	Vascular Plant	Castor canadensis	Brachythecium Moss	Species of Concern
freq	4470	3	7	161

There are 283 duplicates in the scientific name.

species_info.csv

There are 5541 records in the species_info table.

	category	scientific_name	common_names	conservation_status
count	5541	5541	5541	179
unique	7	5541	5229	4
top	Vascular Plant	Clethrionomys gapperi gapperi	Brachythecium Moss	Species of Concern
freq	4262	1	7	151

After dropping the duplicates, this is the new summary information of species_info.csv

species_info.csv

We replaced the null values in the conservation_status column (i.e. the species that are not protected) with **“Not in danger of extinction”.**

```
conservation_status
Endangered                       15
In Recovery                       3
Not in danger of extinction    5362
Species of Concern              151
Threatened                       10
dtype: int64
```

Note: Since the duplicates were removed, the numbers are a little different from the original.

This is the file that contains information from recorded sightings of different species throughout the national parks in the past 7 days.

It has three columns:

- **scientific_name**: The scientific name of the species
- **park_name**: The name of the National Park
- **observations**: The number of observations in the past 7 days

The file consists of 23296 rows and 3 columns.

observations.csv

observations.csv

```
Number of Rows:23296
Number of Columns:3
```

	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

observations.csv

	scientific_name	park_name	observations
count	23296	23296	23296.000000
unique	5541	4	NaN
top	Myotis lucifugus	Great Smoky Mountains National Park	NaN
freq	12	5824	NaN
mean	NaN	NaN	142.287904
std	NaN	NaN	69.890532
min	NaN	NaN	9.000000
25%	NaN	NaN	86.000000
50%	NaN	NaN	124.000000
75%	NaN	NaN	195.000000
max	NaN	NaN	321.000000

observations.csv

	scientific_name	park_name	observations
count	23296	23296	23296.000000
unique	5541	4	NaN
top	Myotis lucifugus	Great Smoky Mountains National Park	NaN
freq	12	5824	NaN
mean	NaN	NaN	142.287904
std	NaN	NaN	69.890532
min	NaN	NaN	9.000000
25%	NaN	NaN	86.000000
50%	NaN	NaN	124.000000
75%	NaN	NaN	195.000000
max	NaN	NaN	321.000000

observations.csv

	scientific_name	park_name	observations
1766	Canis lupus	Bryce National Park	27
7346	Canis lupus	Bryce National Park	29
9884	Canis lupus	Bryce National Park	74
10190	Canis lupus	Great Smoky Mountains National Park	15
17756	Canis lupus	Great Smoky Mountains National Park	14
20353	Canis lupus	Great Smoky Mountains National Park	30
10268	Canis lupus	Yellowstone National Park	60
10907	Canis lupus	Yellowstone National Park	67
13427	Canis lupus	Yellowstone National Park	203
1294	Canis lupus	Yosemite National Park	35
19330	Canis lupus	Yosemite National Park	117
19987	Canis lupus	Yosemite National Park	44

- There were duplicate rows with different observations.
- We summarized the observations into a single row.

observations.csv

	scientific_name	park_name	observations
3216	Canis lupus	Bryce National Park	130
3217	Canis lupus	Great Smoky Mountains National Park	59
3218	Canis lupus	Yellowstone National Park	330
3219	Canis lupus	Yosemite National Park	196

This resulted in 22164 rows and 3 columns.

The Analysis

What is the distribution of conservation status for species?

We found that 5362 species are not in the conservation program while 179 species are protected species.

Are certain types of species more likely to be endangered?

	category	not_protected	protected	percent_protected
0	Amphibian	72	7	8.860759
1	Bird	413	75	15.368852
2	Fish	114	11	8.800000
3	Mammal	146	30	17.045455
4	Nonvascular Plant	328	5	1.501502
5	Reptile	73	5	6.410256
6	Vascular Plant	4216	46	1.079305

- We found that Mammal and Bird categories had the highest percentage of protected species.

Are the differences between species and their conservation status significant?

As seen in the table, some differences in species and their conservation status are significant and some are not.

	Category1	Category2	p-value	Significant?
0	Reptile	Vascular Plant	1.450522e-04	Statistically Significant
1	Nonvascular Plant	Vascular Plant	6.623419e-01	Not Statistically Significant
2	Nonvascular Plant	Reptile	3.362698e-02	Statistically Significant
3	Mammal	Vascular Plant	1.440507e-55	Statistically Significant
4	Mammal	Reptile	3.835559e-02	Statistically Significant
5	Mammal	Nonvascular Plant	1.481869e-10	Statistically Significant
6	Fish	Vascular Plant	1.139913e-12	Statistically Significant
7	Fish	Reptile	7.286746e-01	Not Statistically Significant
8	Fish	Nonvascular Plant	4.587125e-04	Statistically Significant
9	Fish	Mammal	5.948567e-02	Not Statistically Significant
10	Bird	Vascular Plant	4.612268e-79	Statistically Significant
11	Bird	Reptile	5.313542e-02	Not Statistically Significant
12	Bird	Nonvascular Plant	1.054631e-10	Statistically Significant
13	Bird	Mammal	6.875948e-01	Not Statistically Significant
14	Bird	Fish	8.142211e-02	Not Statistically Significant

Are the differences between species and their conservation status significant?

As seen in the table, some differences in species and their conservation status are significant and some are not.

Eg: Mammals and Birds are not statistically significant while Mammals and Reptiles are statistically significant

	Category1	Category2	p-value	Significant?
0	Reptile	Vascular Plant	1.450522e-04	Statistically Significant
1	Nonvascular Plant	Vascular Plant	6.623419e-01	Not Statistically Significant
2	Nonvascular Plant	Reptile	3.362698e-02	Statistically Significant
3	Mammal	Vascular Plant	1.440507e-55	Statistically Significant
4	Mammal	Reptile	3.835559e-02	Statistically Significant
5	Mammal	Nonvascular Plant	1.481869e-10	Statistically Significant
6	Fish	Vascular Plant	1.139913e-12	Statistically Significant
7	Fish	Reptile	7.286746e-01	Not Statistically Significant
8	Fish	Nonvascular Plant	4.587125e-04	Statistically Significant
9	Fish	Mammal	5.948567e-02	Not Statistically Significant
10	Bird	Vascular Plant	4.612268e-79	Statistically Significant
11	Bird	Reptile	5.313542e-02	Not Statistically Significant
12	Bird	Nonvascular Plant	1.054631e-10	Statistically Significant
13	Bird	Mammal	6.875948e-01	Not Statistically Significant
14	Bird	Fish	8.142211e-02	Not Statistically Significant

Which species are In Recovery and what is the overall distributions in the parks compared to the average observations of the species with other conservation statuses?

category	Amphibian	Bird	Fish	Mammal	Nonvascular Plant	Reptile	Vascular Plant
conservation_status							
Endangered	1.0	4.0	3.0	6.0	NaN	NaN	1.0
In Recovery	NaN	3.0	NaN	NaN	NaN	NaN	NaN
Species of Concern	4.0	68.0	4.0	22.0	5.0	5.0	43.0
Threatened	2.0	NaN	4.0	2.0	NaN	NaN	2.0

Which species are In Recovery and what is the overall distributions in the parks compared to the average observations of the species with other conservation statuses?

category	Amphibian	Bird	Fish	Mammal	Nonvascular Plant	Reptile	Vascular Plant
conservation_status							
Endangered	1.0	4.0	3.0	6.0	NaN	NaN	1.0
In Recovery	NaN	3.0	NaN	NaN	NaN	NaN	NaN
Species of Concern	4.0	68.0	4.0	22.0	5.0	5.0	43.0
Threatened	2.0	NaN	4.0	2.0	NaN	NaN	2.0

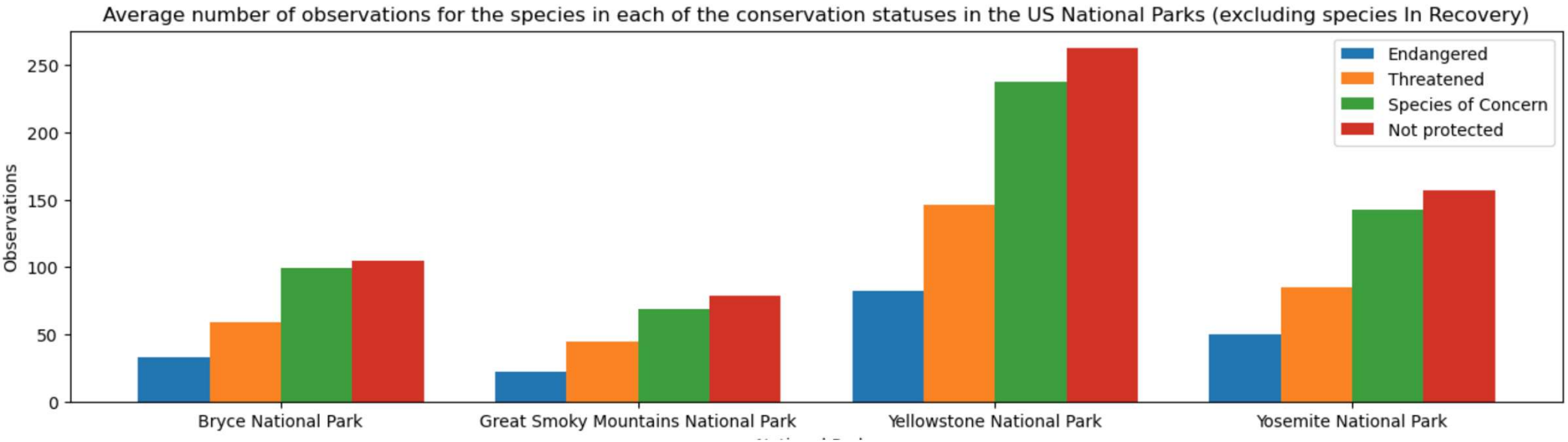
Which species are In Recovery and what is the overall distributions in the parks compared to the average observations of the species with other conservation statuses?

	scientific_name	common_names	park_name	observations
0	Falco peregrinus anatum	American Peregrine Falcon	Bryce National Park	72
1	Falco peregrinus anatum	American Peregrine Falcon	Great Smoky Mountains National Park	70
2	Falco peregrinus anatum	American Peregrine Falcon	Yellowstone National Park	176
3	Falco peregrinus anatum	American Peregrine Falcon	Yosemite National Park	152
4	Haliaeetus leucocephalus	Bald Eagle	Bryce National Park	94
5	Haliaeetus leucocephalus	Bald Eagle	Great Smoky Mountains National Park	72
6	Haliaeetus leucocephalus	Bald Eagle	Yellowstone National Park	187
7	Haliaeetus leucocephalus	Bald Eagle	Yosemite National Park	112
8	Pelecanus occidentalis	Brown Pelican	Bryce National Park	92
9	Pelecanus occidentalis	Brown Pelican	Great Smoky Mountains National Park	47
10	Pelecanus occidentalis	Brown Pelican	Yellowstone National Park	196
11	Pelecanus occidentalis	Brown Pelican	Yosemite National Park	122

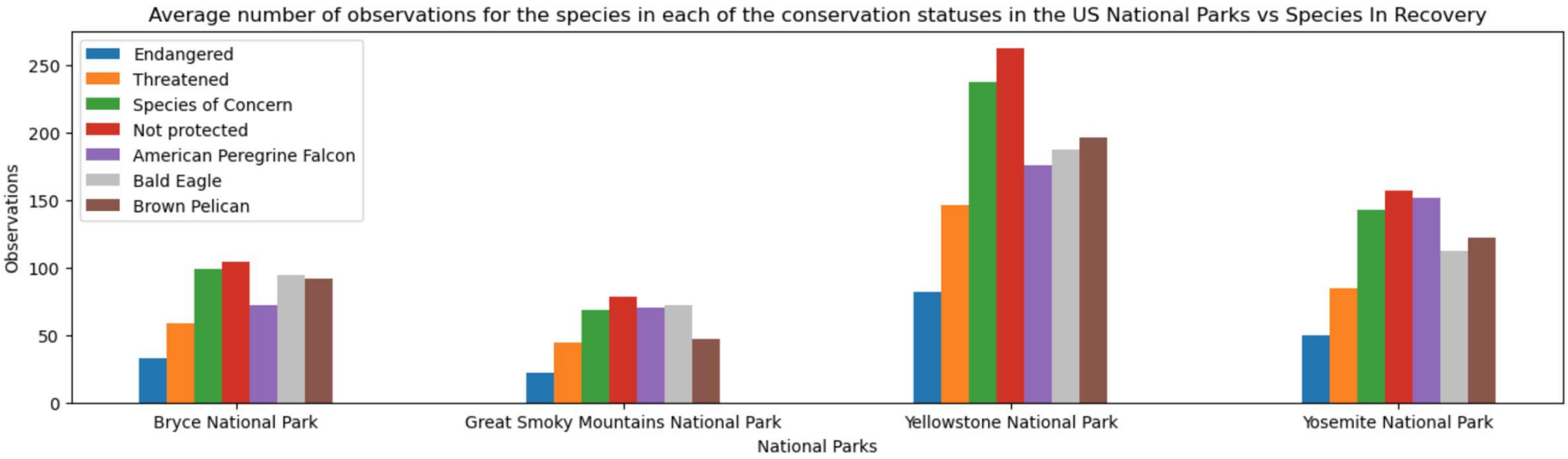
The three species **In Recovery**:

- Falco peregrinus anatum (American Peregrine Falcon)
- Haliaeetus leucocephalus (Bald Eagle)
- Pelecanus occidentalis (Brown Pelican)

Which species are In Recovery and what is the overall distributions in the parks compared to the average observations of the species with other conservation statuses?

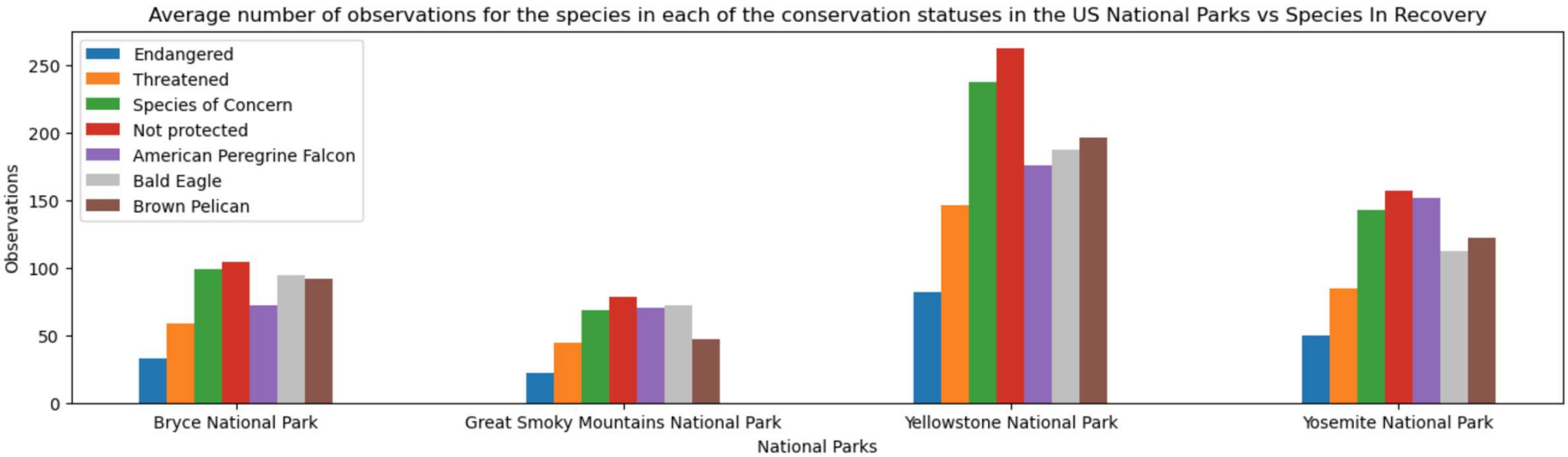


Which species are In Recovery and what is the overall distributions in the parks compared to the average observations of the species with other conservation statuses?



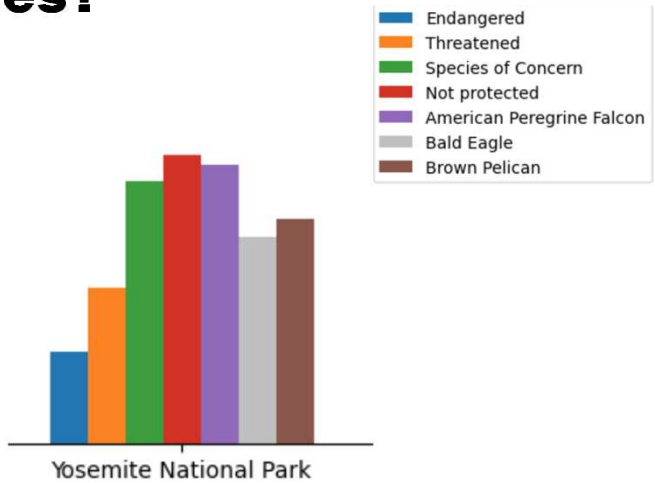
Which species are In Recovery and what is the overall distributions in the parks compared to the average observations of the species with other conservation statuses?

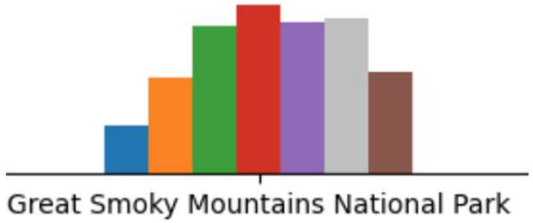
The observations for the **In Recovery** species fell mostly between the average value for *Species of Concern* and *Threatened* or close to the average value for *Species of Concern* which is what is expected. However, there were a few instances where that was not the case.



Which species are In Recovery and what is the overall distributions in the parks compared to the average observations of the species with other conservation statuses?

- In *Yosemite*,
  - The **American Peregrine Falcon** had observations that were higher than the average observations for *Species of Concern* species but lower than the average observations for *Not Protected* species in the National Park.
  - The **Bald Eagle** and the **Brown Pelican** fall in between *Threatened* and *Species of Concern*.
- In *Great Smoky Mountains*,
  - The observations for the **American Peregrine Falcon** were around the average observations of the *Species of Concern.*
  - The observations for the **Bald Eagle** were higher than the average observations for *Species of Concern* species but lower than the average observations for *Not Protected* species in the National Park.
  - The **Brown Pelican** observations fall in between *Threatened* and *Species of Concern*.





Further Analysis

- This dataset only included observations from the last 7 days. It would be curious to see how the conservation status for various species changes over time.
- Another piece that is missing is the Area of each park, it can be assumed that Yellowstone National Park might be much larger than the other parks which would mean that it would exhibit more observations and greater biodiversity.
- Lastly, if precise locations were recorded, the spatial distribution of the species could also be observed and test if these observations are spatially clustered.

Thank you

Appendix

- Duplicates in common names were observed in species_info.csv due to spellings and other common user errors. The common names were not a field that made a huge impact in our analysis, so the duplicates were ignored.

	category	scientific_name	common_names	conservation_status
count	5824	5824	5824	191
unique	7	5541	5504	4
top	Vascular Plant	Castor canadensis	Brachythecium Moss	Species of Concern
freq	4470	3	7	161