

Biodiversity in US NP

Project overview

Interpret data from the National Parks Service about endangered species in different parks.

Data exploration

Some interesting questions would be:

- which parks maintain biodiversity?
- is there a link between conservation status and taxonomical categories?
- can we infer conservation status from observations? probably, if so can we use this to fill in missing values?
- stats on number of species (in total, per park)
- stats on number of categories (in total, per park), number of species/category.
- is there a bias (some species or park whose data are less captured)?
- which endangered species need to be protected in which park?

Raw data

Dataset 1

Field Observations:

The file **observations.csv** contains 23296 rows and 3 columns **without missing value**.

The variables are:

- 'scientific_name': categorical (nominal) variable listing species using Latin names.
- 'park_name': categorical (nominal) variable listing US National Parks.
- 'observations': numerical (integer) variable (range 9-321), reflects the number of individuals observed per species.

Dataset 2

Species Database:

The file **species_info.csv** contains 5824 rows and 4 columns with **lots of missing values in one of the variable ('conservation_status')**.

The variables are:

- 'category': categorical (nominal) variable listing species taxonomical group, which can be 'Mammal', 'Bird', 'Reptile', 'Amphibian', 'Fish', 'Vascular Plant', or 'Nonvascular Plant'
- 'scientific_name': categorical (nominal) variable listing species using Latin names. Identical to 'scientific_name' in 'observations.csv', could be used to merge both datasets
- 'common_names': categorical (nominal) variable listing common names of species observed. Single name are expected yet some entries list multiple names (e.g. 'Aurochs, Aurochs, Domestic Cattle (Feral), Dom...').
- 'conservation_status': categorical (ordinal) variable listing species conservation status, which can be 'In Recovery'>'Species of Concern'>'Endangered'>'Threatened'. Includes 5633 NaNs. Could be converted to increasing numbers for stats and visualisation purpose

Data cleaning

Combining datasets

The datasets ‘observations’ and ‘species_info’ had to be joined for statistical and data visualisation purpose.

A pre-requisite was to check for duplicates

- ‘observations’ contained 15 duplicates which were eliminated.
- ‘species_info’ contained 283 duplicated scientific names that had different ‘common_names’. I chose to only keep the 1st entry.

Both files were joined using the common variable ‘**scientific_name**’.

The merged dataset contained 23251 rows and 6 columns:

	scientific_name	park_name	observations	category	common_names	conservation_status
0	Vicia benghalensis	Great Smoky Mountains National Park	68	Vascular Plant	Purple Vetch, Reddish Tufted Vetch	NaN
1	Neovison vison	Great Smoky Mountains National Park	77	Mammal	American Mink	NaN
2	Prunus subcordata	Yosemite National Park	138	Vascular Plant	Klamath Plum	NaN
3	Abutilon theophrasti	Bryce National Park	84	Vascular Plant	Velvetleaf	NaN
4	Githopsis specuarioides	Great Smoky Mountains National Park	85	Vascular Plant	Common Bluecup	NaN
...

Replacing missing values

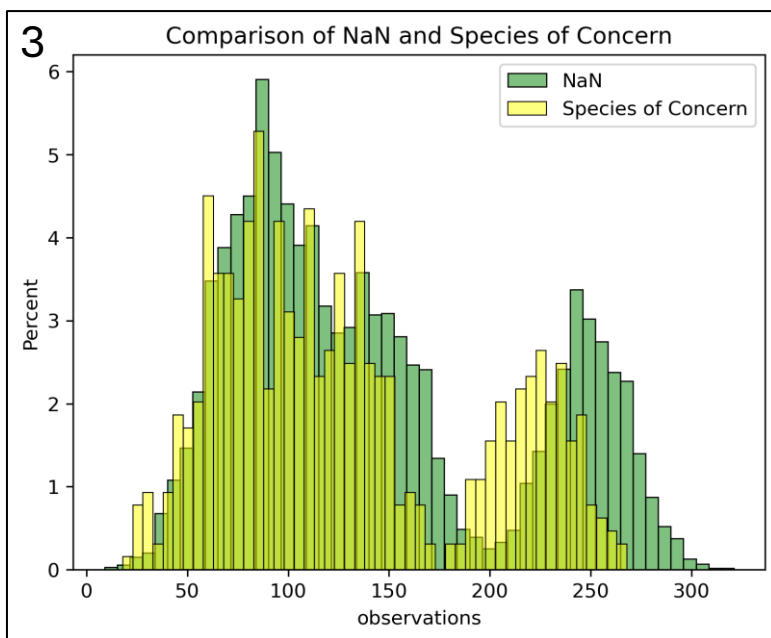
A critical step was to handle the 22,521 (97%) missing values (NaN) in 'conservation_status'.

Type of NaN:

We investigated whether they were structurally missing data and looked for a link with other variables. They were not associated to parks (χ^2 p-value = 0.999) but were associated to taxonomical categories (χ^2 p-value = 0.0). However, chart 2 doesn't show any discernable pattern so we can assume NaN are missing at random (MAR)

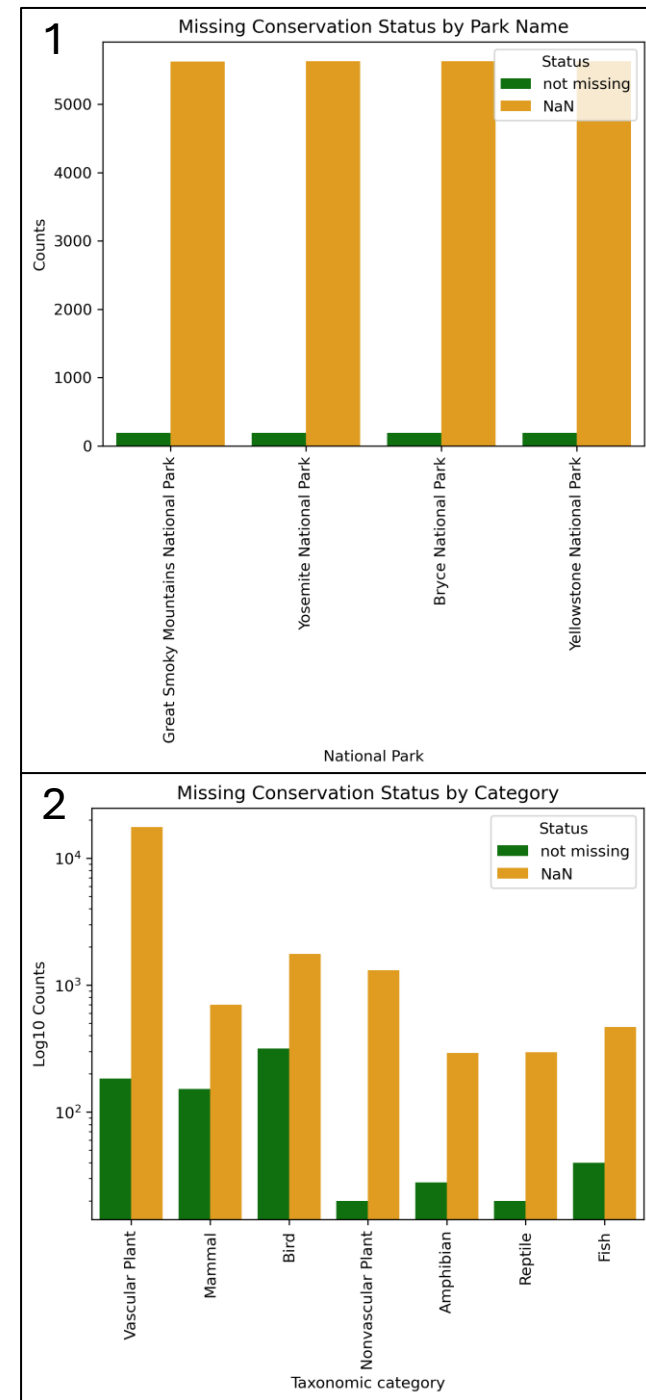
Relationship with Observations

I plotted the observation distribution for each status but we cannot infer from those an "observation" threshold to decide which status to apply to which observation. We noticed that NaN assumed a similar bimodal distribution to 'Species of Concern':



Knowing this, could I replace all NaN with 'Species of Concern'?

A more conservative and logical approach would be to consider these NaNs as structurally missing values and consider them all as "Not Threatened" which could explain why no value has been attributed.



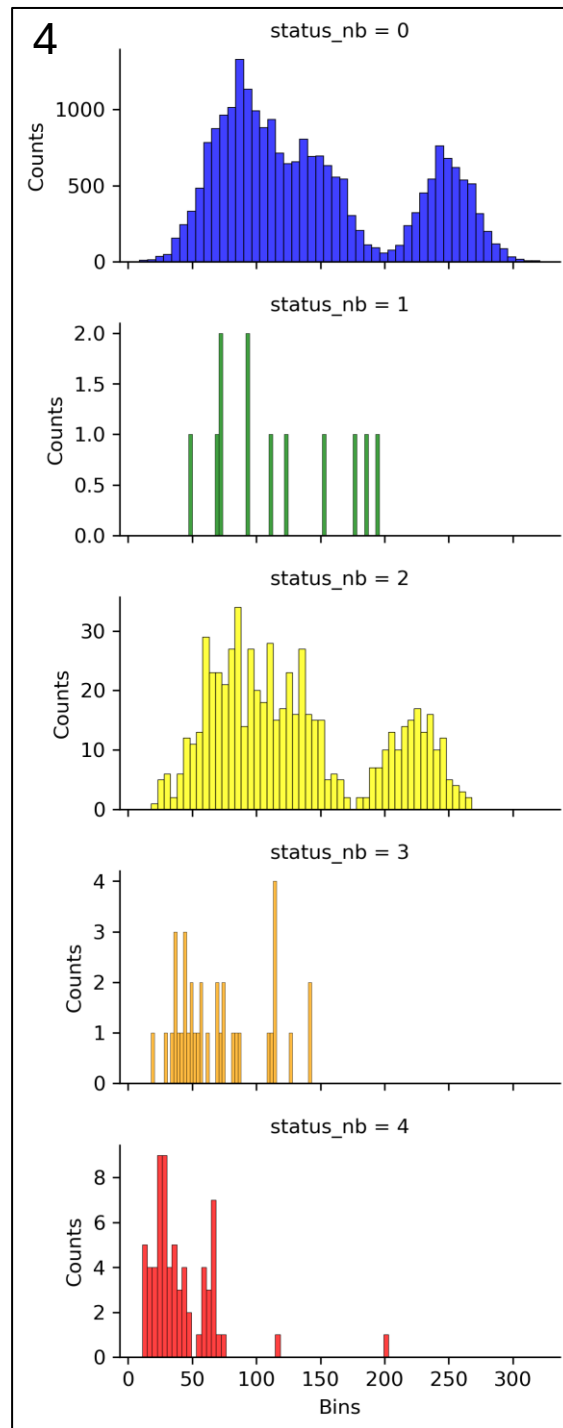
Converting status to numbers

I created a numerical variable 'status_nb' based on 'conservation_status' as follows:

- 0 : No Threatened (previously NaN)
- 1 : In Recovery
- 2 : Species of Concern
- 3 : Threatened
- 4 : Endangered

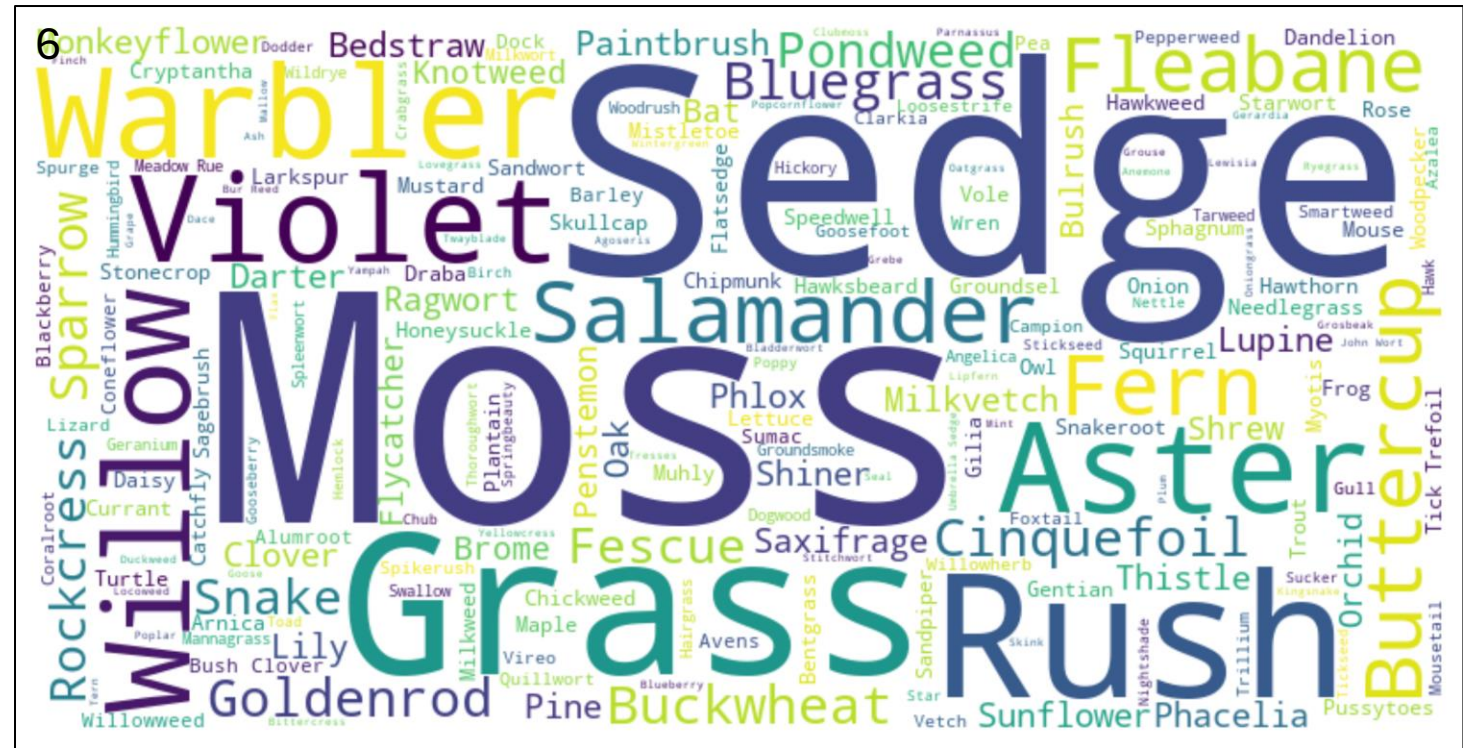
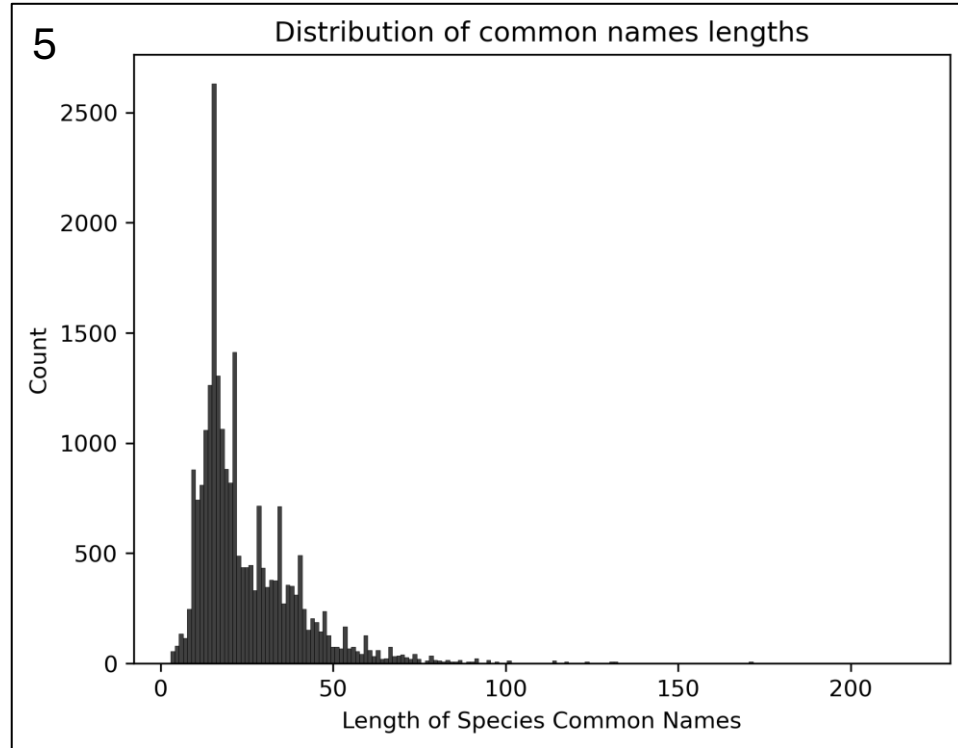
Chart 4 shows the distribution of each status based on observations. The proportions are as follows:

Status_nb	Conservation_status	counts
0	No Threatened	22521
1	In Recovery	12
2	Species of Concern	644
3	Threatened	36
4	Endangered	68



Inspecting species common names

Some common names feature very long strings. I've investigated the length distribution (chart 5). I've also isolated the last word of each common name which is the best descriptor at a high level and created a new variable 'last_common_name' (e.g. Reddish Tufted Vetch → Vetch). I've created a word cloud of them (chart 6).

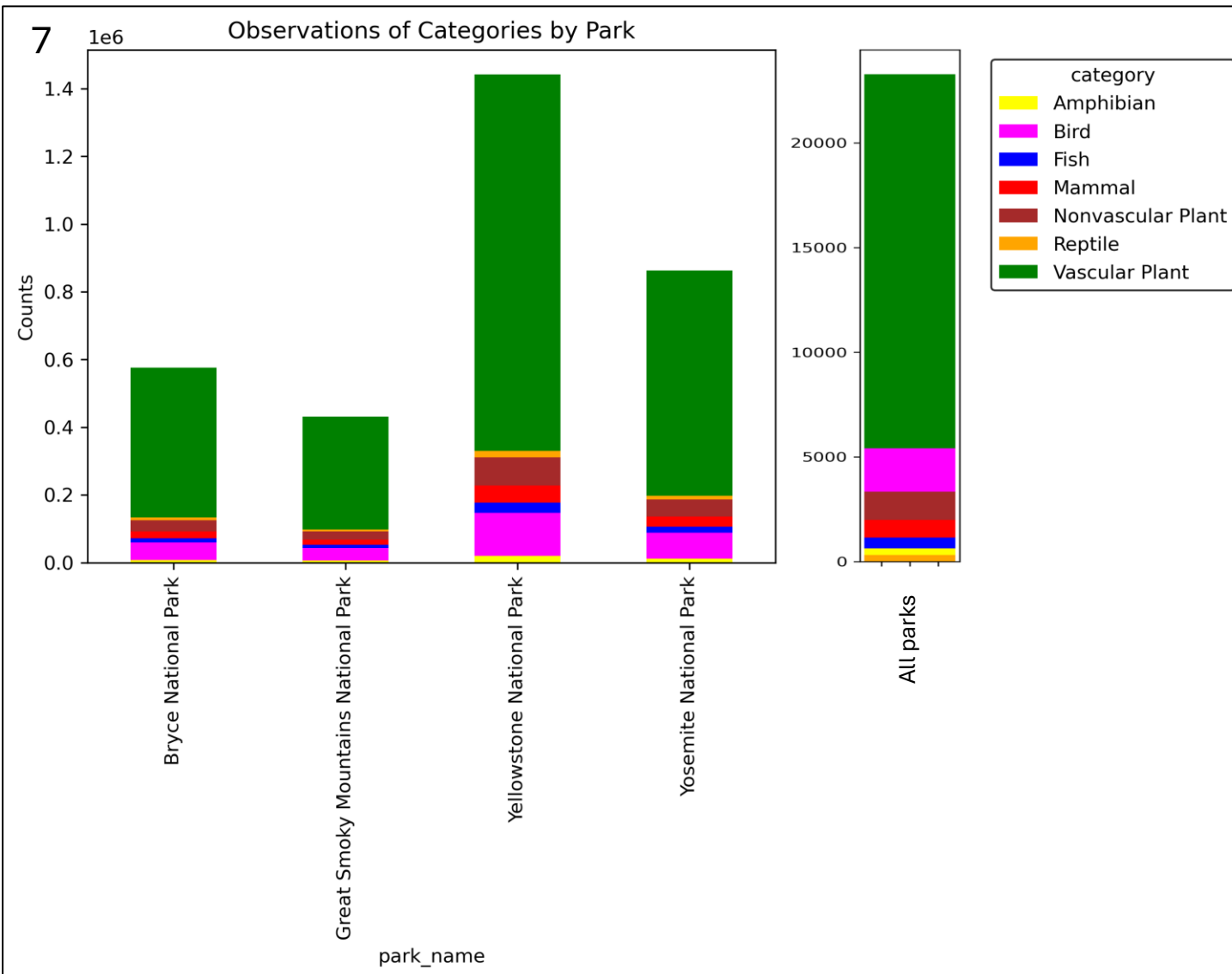


Common name lengths range 3-218 characters, with only a few very long strings. No need to take any further action. The wordcloud shows the prevalence of plants. Predominant animals are wrblers, salamanders and sparrow.

Data analysis

Parks biodiversity

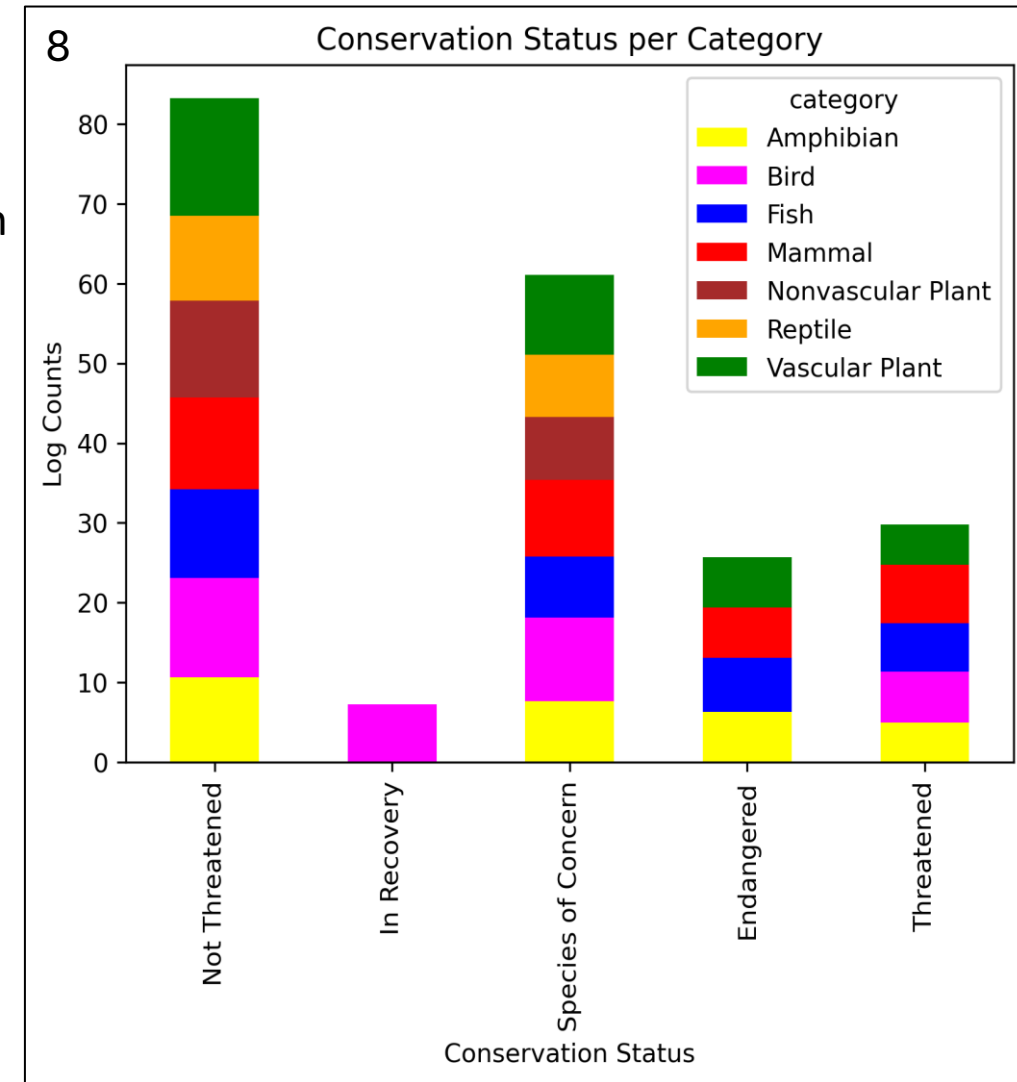
Parks feature different category distributions. As far as biodiversity is concerned, they rank as follows:
Yellowstone NP > Yosemite NP > Bryce NP > Great Smoky Mountains NP.



Are some taxonomical categories faring better than others?

Chart 8 shows that:

- Unthreatened species (status_nb = 0) are equally represented across all taxonomical categories.
 - Only birds are in recovery (status_nb = 1), across all parks (12 entries) with 47-196 sightings.
 - All taxons feature Species of Concern (status_nb = 2).
 - Threatened species (status_nb = 3) belong to Amphibians, Fish, Mammal and Vascular Plant.
 - Endangered species (status_nb = 4) belong to Amphibians, Birds, Fish, Mammal and Vascular Plant.
- Mammals are the most numerous (13-203 sightings), featuring wolves, squirrels, bats and mountain sheep (chart 10).



Recovering species

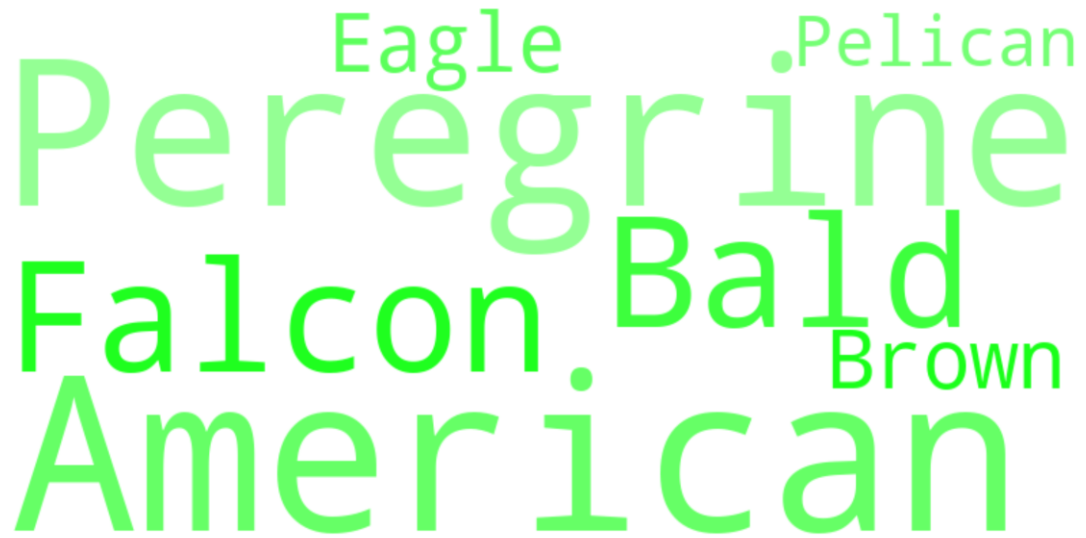
Only birds are in recovery (chart 9).

Those birds are:

- American Peregrine Falcon
- Bald Eagle
- Brown Pelican

Why has this category been more successful than others?

9: In recovery



A word cloud containing the names of the three bird species in recovery: 'Peregrine', 'Falcon', 'American', 'Bald', 'Eagle', and 'Brown Pelican'. The words are in various shades of green and are arranged in a way that they are partially overlapping and difficult to read as a single phrase.

Critically endangered species

The most critically endangered species are represented in chart 10.

The 5 most critically endangered species were common across all parks, they should be part of national program to protect them.

They are:

- Canis rufus
- Rana sierrae
- Myotis grisescens
- Picoides borealis
- Etheostoma percnurum

10: Critically endangered



Conclusions

General conclusions

The datasets provided were merged, duplicates were removed, and missing values were replaced by a Not Threatened label.

Data analysis and visualisation showed that:

- Parks feature different category distributions. As far as biodiversity is concerned, they rank as follows: Yellowstone NP > Yosemite NP > Bryce NP > Great Smoky Mountains NP.
- Observation values are associated with the different parks and status but not taxonomic categories
- Only 'American Peregrine Falcon', the 'Bald Eagle', and the 'Brown Pelican' birds are in recovery
- Amphibians, Birds, Fish, Vascular Plant and Mammal in particular feature Endangered species.
- Mammals are the most numerous, featuring wolves, squirrels, bats and mountain sheep.
- The 5 most critically endangered species are:
 - *Canis rufus*
 - *Rana sierrae*
 - *Myotis grisescens*
 - *Picoides borealis*
 - *Etheostoma percnurum*

Future endeavours

More data should be acquired across more US National Parks and the analysis reproduced.

Federal and state governments should be given a list of the most critically endangered species so that they can put in place protective measures.