As part of CodeCademy’s data science program, I’ve been asked to conduct an analysis on two data sets concerning endangered species across four national parks. The hope is that, by the conclusion of this analysis, the reader might come away with a better understanding of the challenges associated with the data itself and some conclusions about the current state troubled species within the US Parks system.

**The Parks**

The parks themselves are from across the continental United States. They very in many ways, but most importantly in their unique geographic locations and park sizes. The smallest among them is Bryce Canyon National Park in Utah. It contains roughly 36,000 acres (14.500 hectars) and features a primarily rocky terrain known for its crimson-colored hoodoos (spires of rock formed from erosion). The largest, Yellowstone National Park, consists of more than 2.2 million acres (900,000 hectars) and contains geysers, hot springs, and mudpots that echo the Park’s volcanic past. Yosemite sits on 748,000 acres (303,000 hectars) on the eastern side of California bordering Nevada. Situated in the Sierra Nevada mountains, the park boasts towering Sequoia trees, monumental granite cliffs, and the beautiful Bridalveil falls. Rounding out our collection of interest is Great Smoky Mountain National Park (G.S.M.). Weighing in at 522,000 acres (211,000 hectars), this park straddles the border between North Carolina and Tennessee and hosts portions of the Appalachian Trail, year-round flower blooms, and expansive views of the mist-covered mountains. Given how diverse these parks are, one could expect quite different populations of plant and animal species.



**The Data**

Collected by the National Park Service, the data themselves are contained in two separate files called observations.csv and species\_info.csv, respectively. They are standard comma separated value files, with the first containing 23,297 rows across three columns, and the second containing 5,825 rows across four columns. (Row counts include headers.) Observations.csv lists the observations of specific species of plants and animals (identified by their scientific names) across the four parks. Species\_info.csv groups the species into categories while also providing the species common name(s) and their current conservation status.

Looking at the data, it’s clear that there are duplicates and errata. In the Observations data, for example Echinochloa Crus-Galli is listed twice at each park – often with different measurements.

Table

Description automatically generated

Similar issues exist in the Species\_info data. Canis Lupus has multiple common names and conservation statuses. What to do?

Table

Description automatically generated

While the data are not without issue, there is still a clear path ahead for the analysis. Beginning with the duplicates in the observations data, there are a few approaches one could take. One could drop or throw out all of the duplicated data, but doing so would have removed protected species from the study. In this data set, protected species make up such a small portion that I felt that throwing any out would have been excessive. Another approach could have been taking the average of the two observations – after all, the measurements are very close. But with the measurements being so close, the approach taken here was to be as conservative as possible and take the lower of the two.

In a similar manner, when cleaning up the species\_info data, the most conservative (most troubled) conservation status was kept, dropping all other rows per species. Common names were combined to make sure that the dataset was as complete and robust as possible. Finally, the data were combined into a single data frame to simplify extended analysis and charting.

After cleaning, there are 22,164 observations with 5,541 scientific names (5,236 lists of common names). Observations of these species span 7 categories, 5 conservation statuses, and 4 parks. It appears that duplication happens within the common names as well. Dicranum Moss is the common name for at least seven unique species of moss.

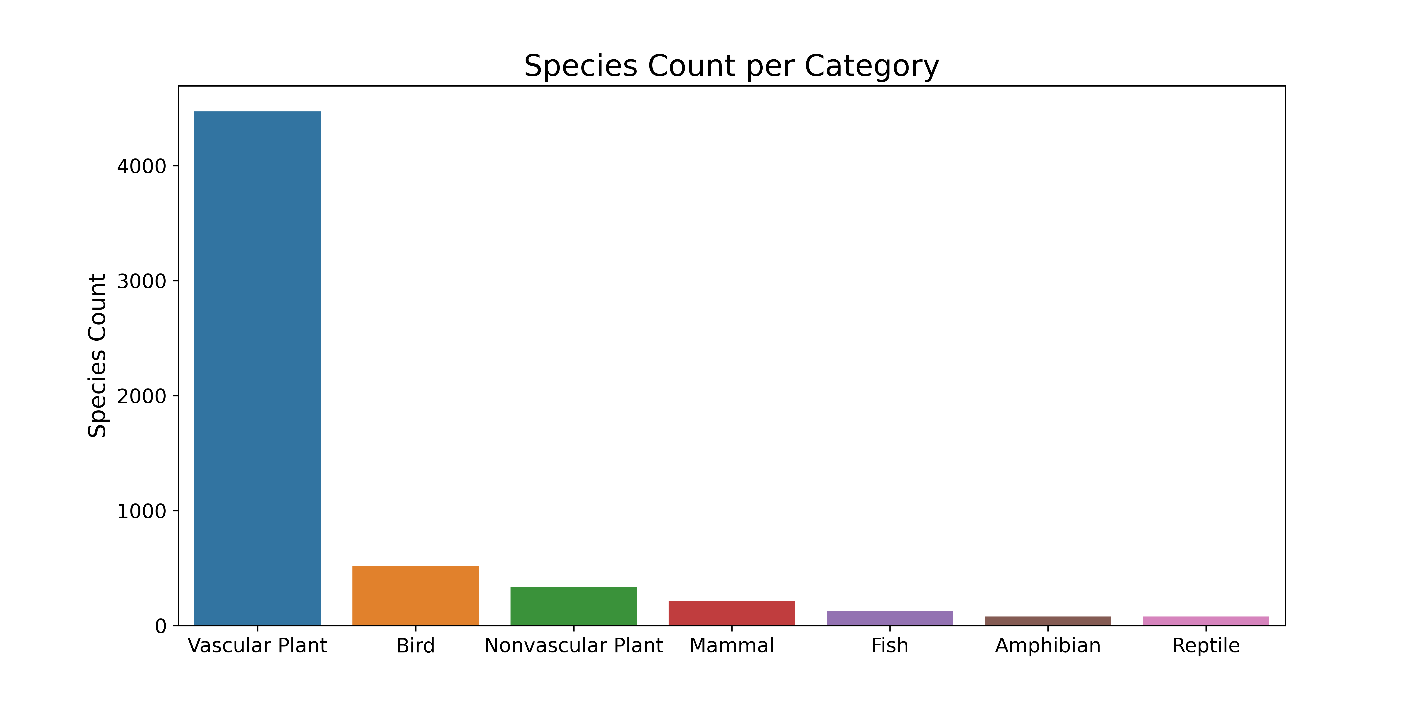
Table

Description automatically generated

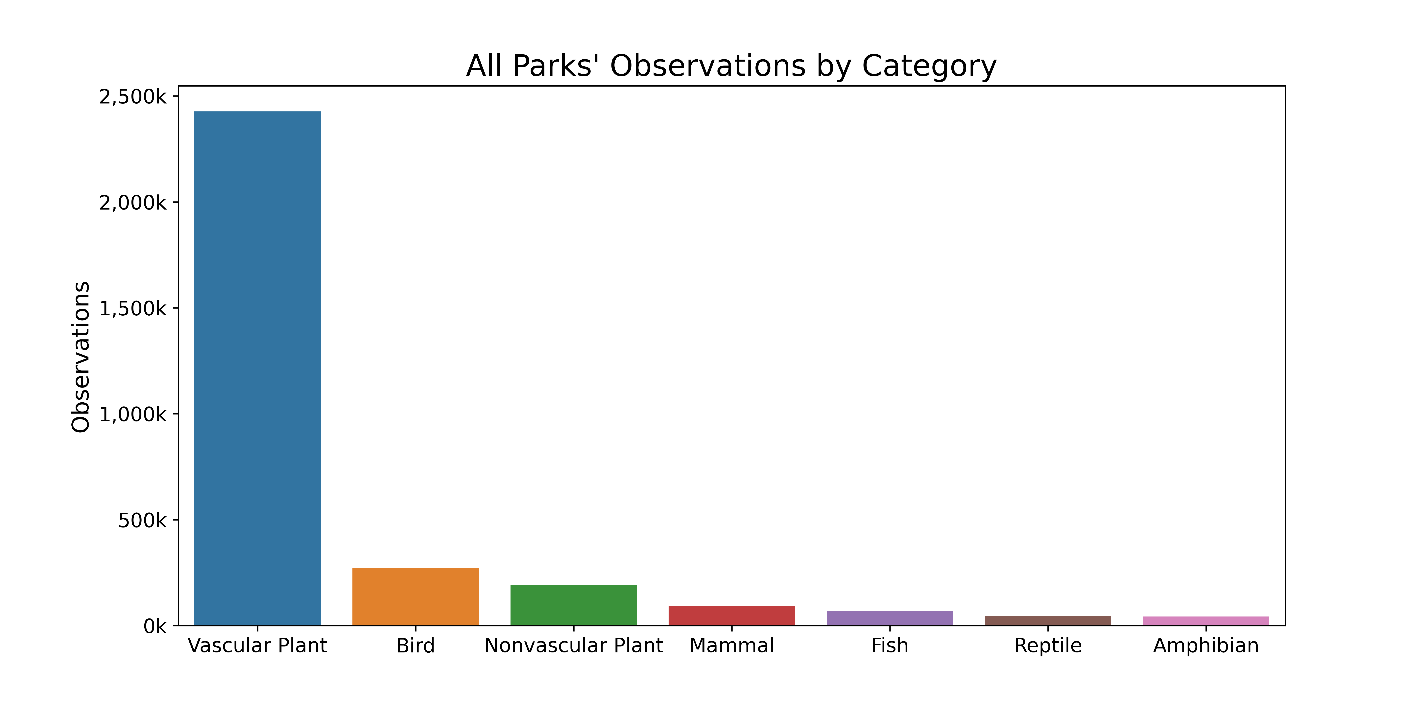
**Summary Findings**

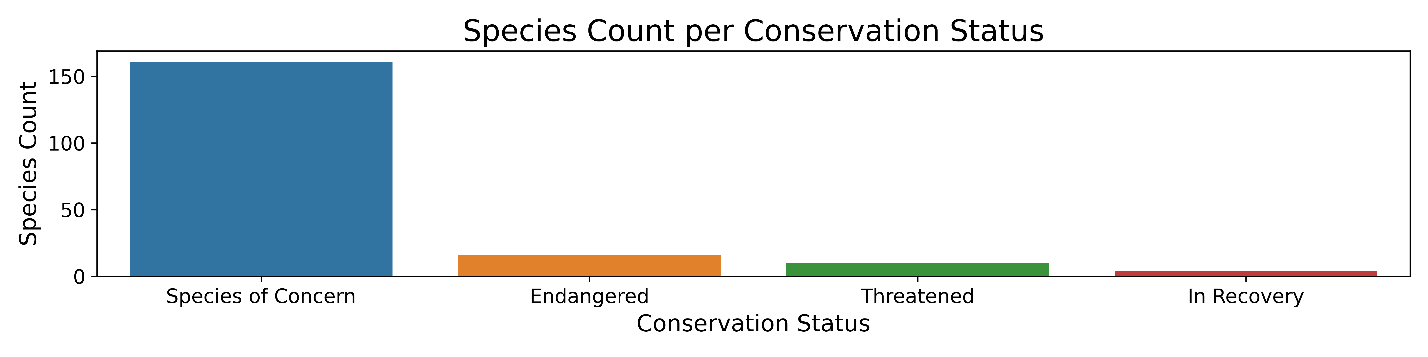
The bulk of the species (96.8%) are in no conservation group (meaning they are not in recovery, a species of concern, threatened, or endangered). Of those which are in a group; 12 (0.1%) are in recovery, 604 (2.7%) are species of concern, 40 (0.2%) are threatened, and 60 (0.3%) are endangered.

By far, the most common category was vascular plants at 76.9% of the species, followed by birds at 8.8%. Nonvascular plants made up 6.0% of the species, with the remaining four categories (mammal, fish, amphibian, and reptile) constituting the remaining 8.3%.



Not surprisingly this was almost mirrored by the total observations of each category across all parks.

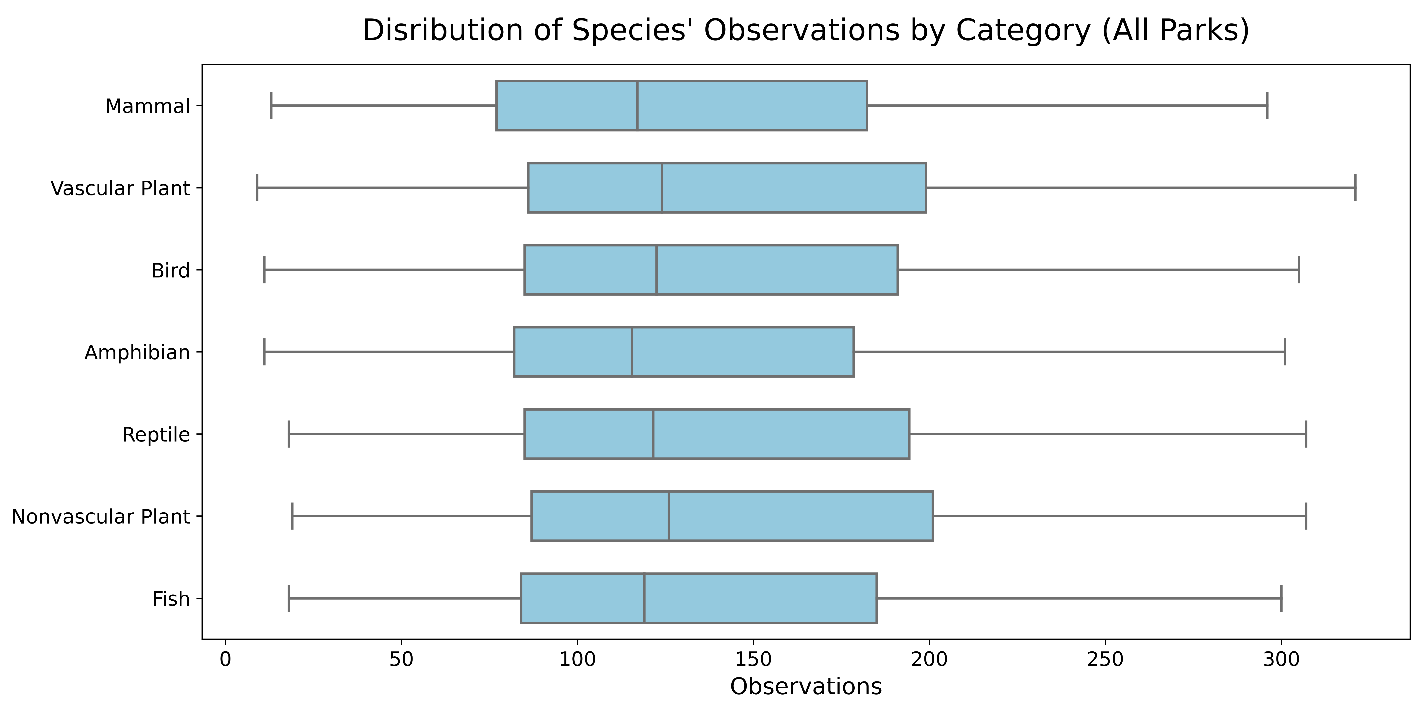


****

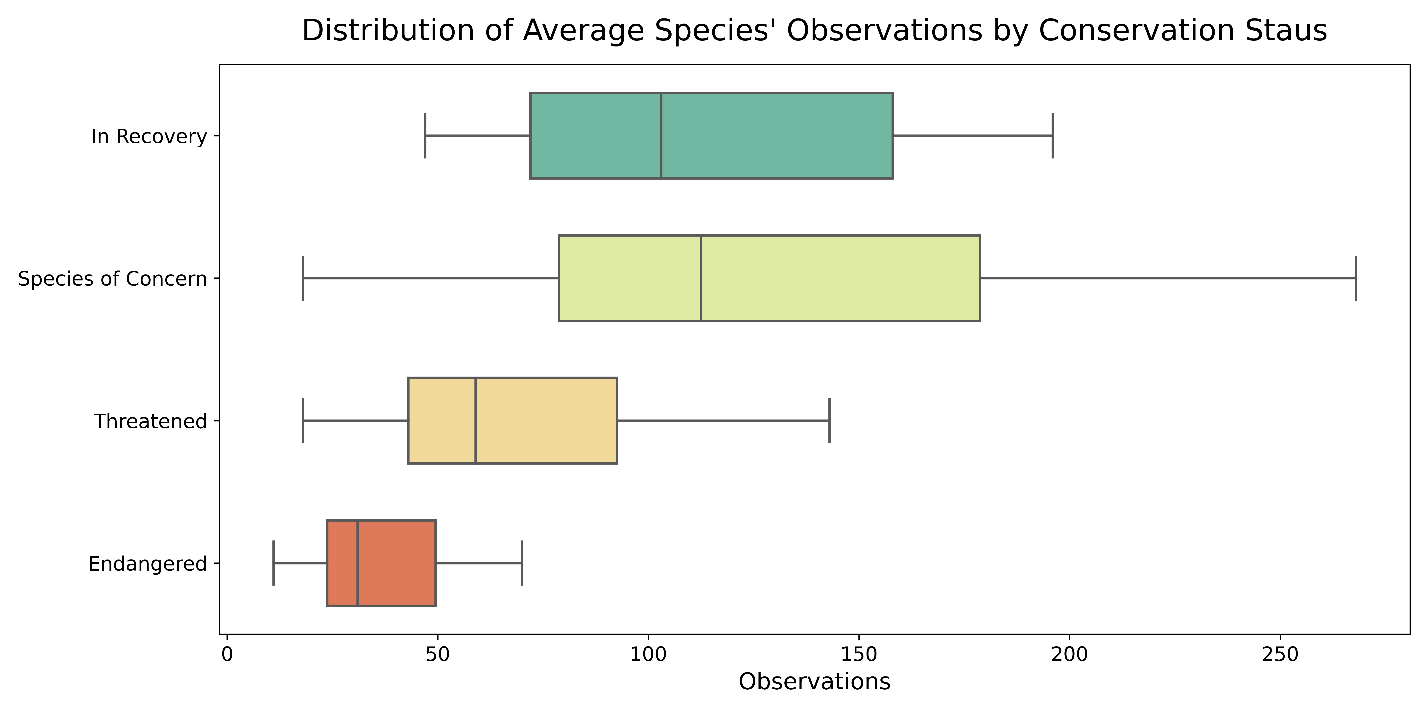
****

While one might expect that the number of species in each park might be unique, they are not. Each park records observations of all 5,541 species (leading one to conclude that this is an inventory exercise where the species are a set, predefined list).

While it’s clear that, overall, vascular plants rule the roost when it comes to observations, this is simply due to the overwhelming number of species contained in that category (4,262). When one takes the average of observations for species within each category, things begin to look much more normal – with all categories’ species having similar observation characteristics. The median number of all observations across all species and parks is 124, with a minimum of 9, and a maximum of 321.



As could be expected, the same cannot be said for the observations when broken down by conservation status. Here, observations of species with no conservation status (1.3m observations across 5,362 species) are omitted for emphasis, as they would otherwise squash the below results into unintelligible marks in white space.



Here we can see that species in the endangered category have fewer observations than those considered threatened, and likewise for those in species of concern. Interestingly, species categorized as in recovery tend to have fewer observations than species of concern.

Observations per Park

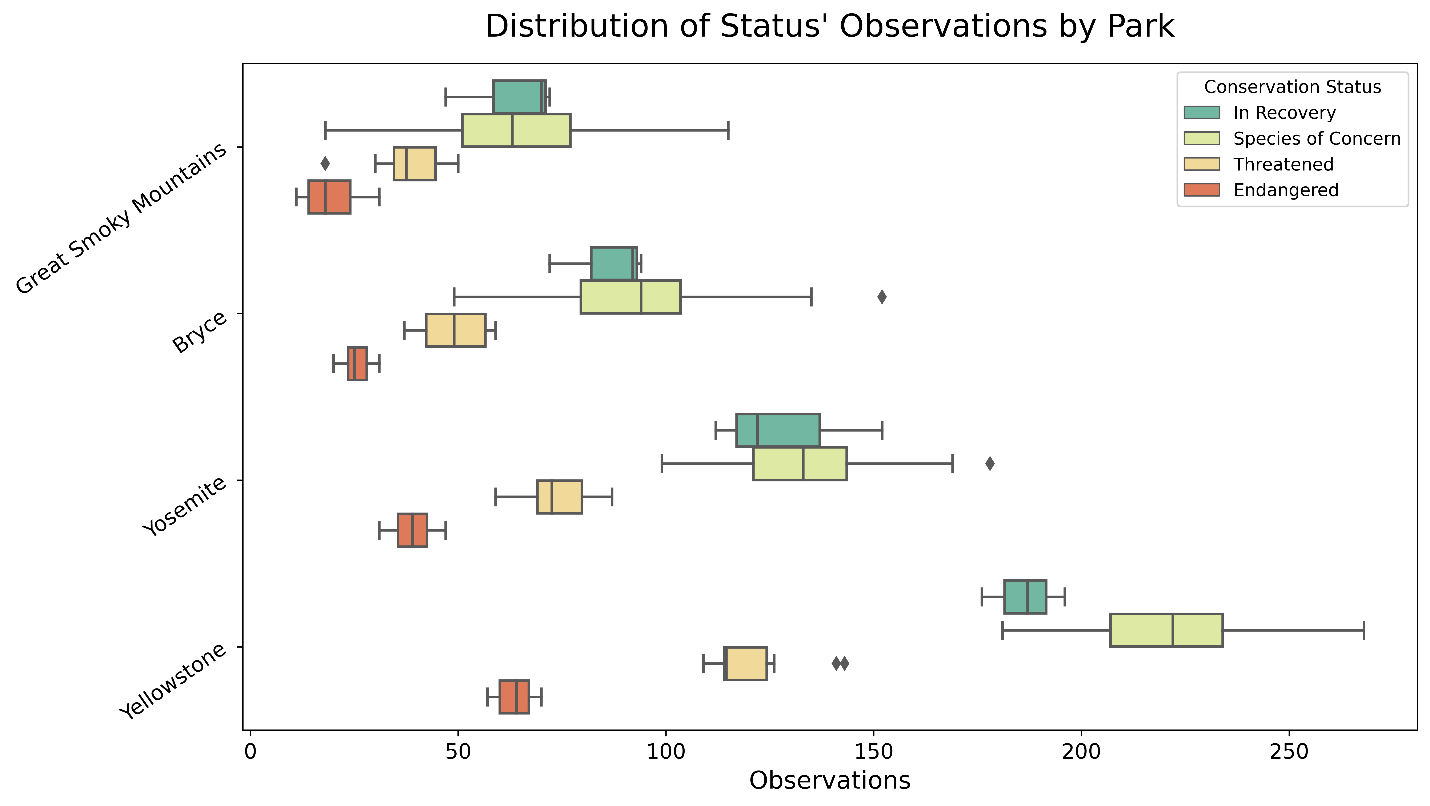
Yellowstone had the most observations across all species, followed by Yosemite, Bryce Canyon, and G.S.M. However, when looked at on a per acre basis, Bryce Canyon clearly takes the lead at 15 observations per acre. In fact, according to the data, Bryce Canyon had almost 14 times the observation density of the next nearest park, Yosemite. That’s an amazing density for such a small park, and pushes Bryce up on the list of parks I want my family to visit.



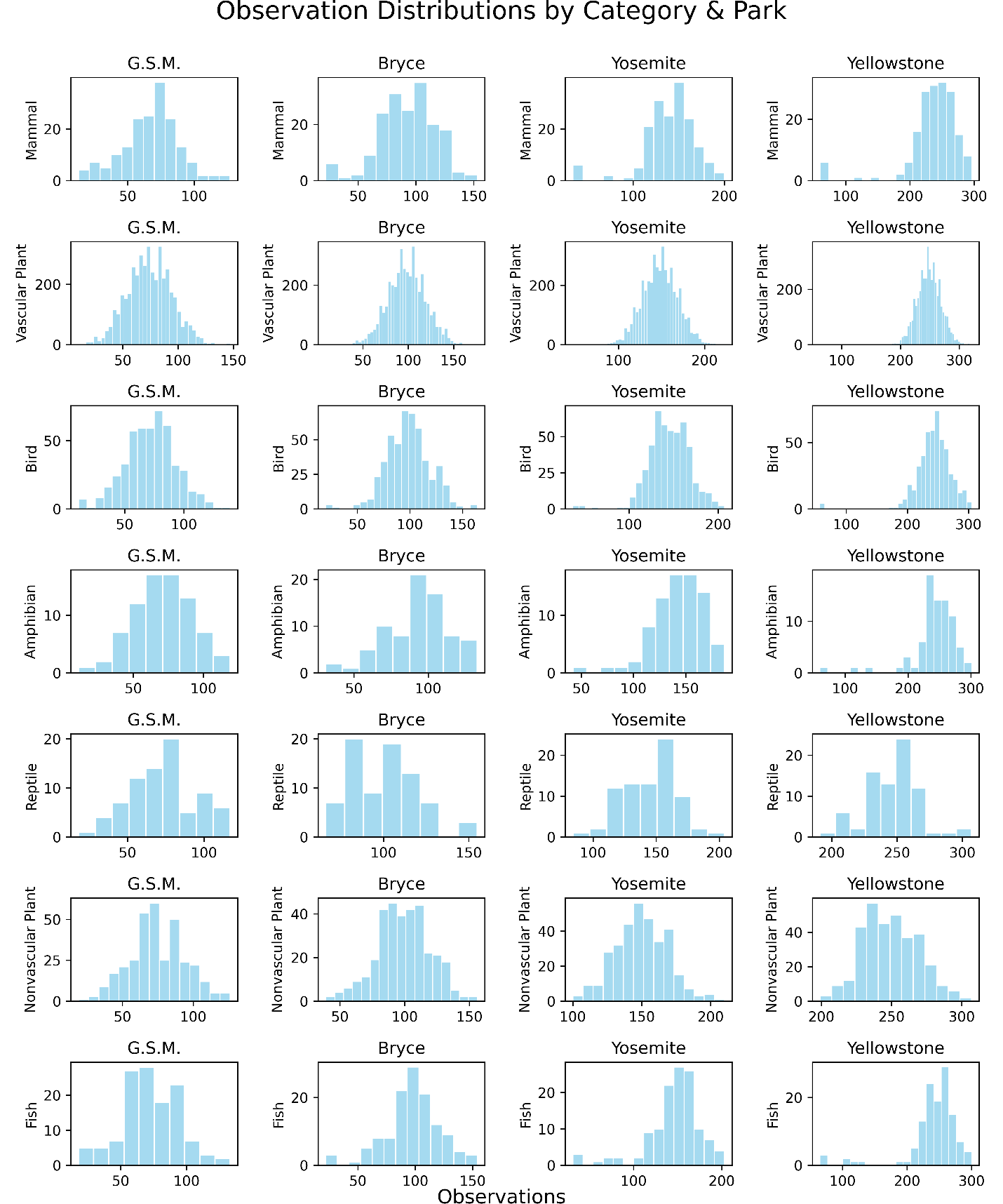


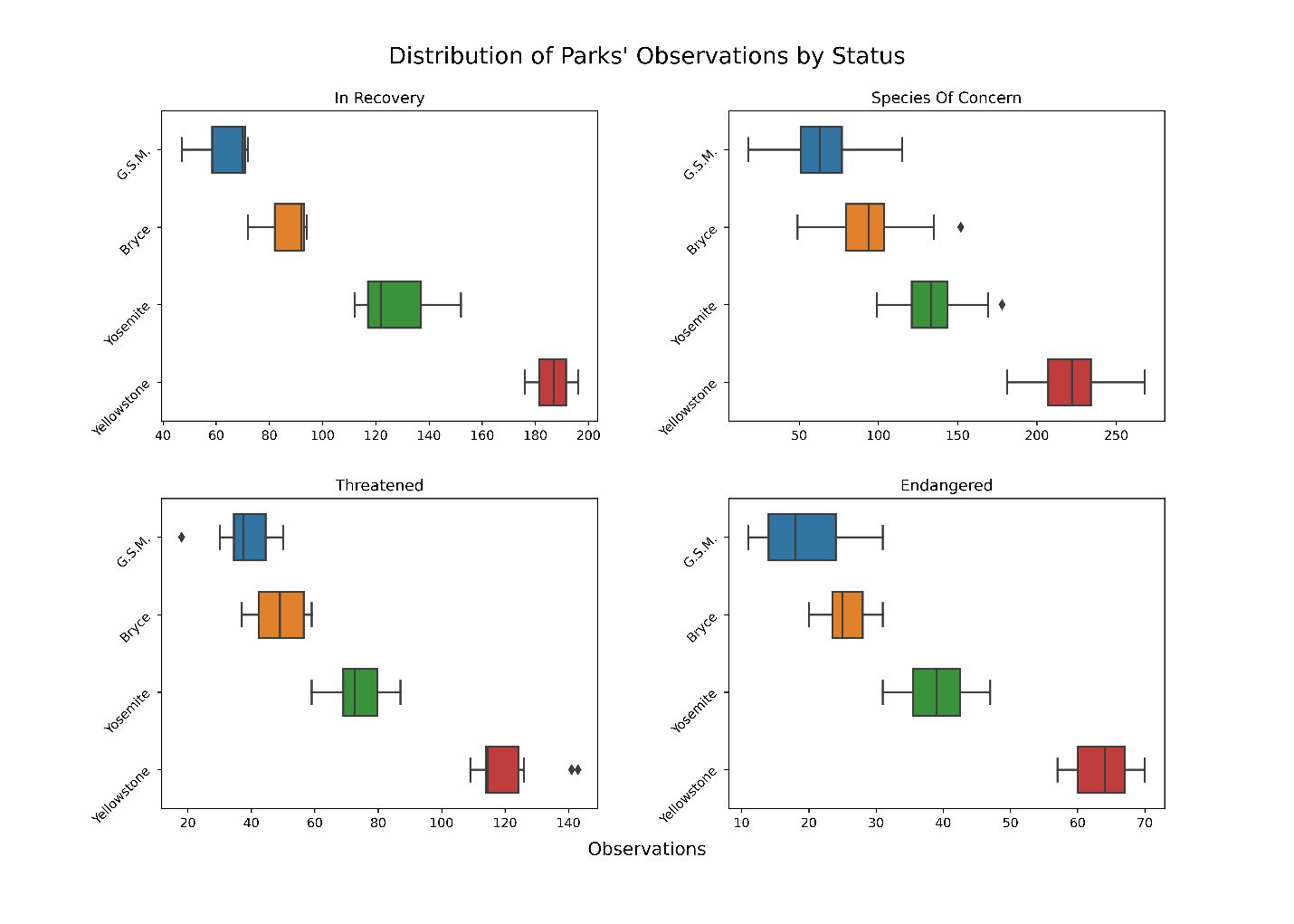


Observations for protected status by park stays on trend – mirroring the total observations by park.

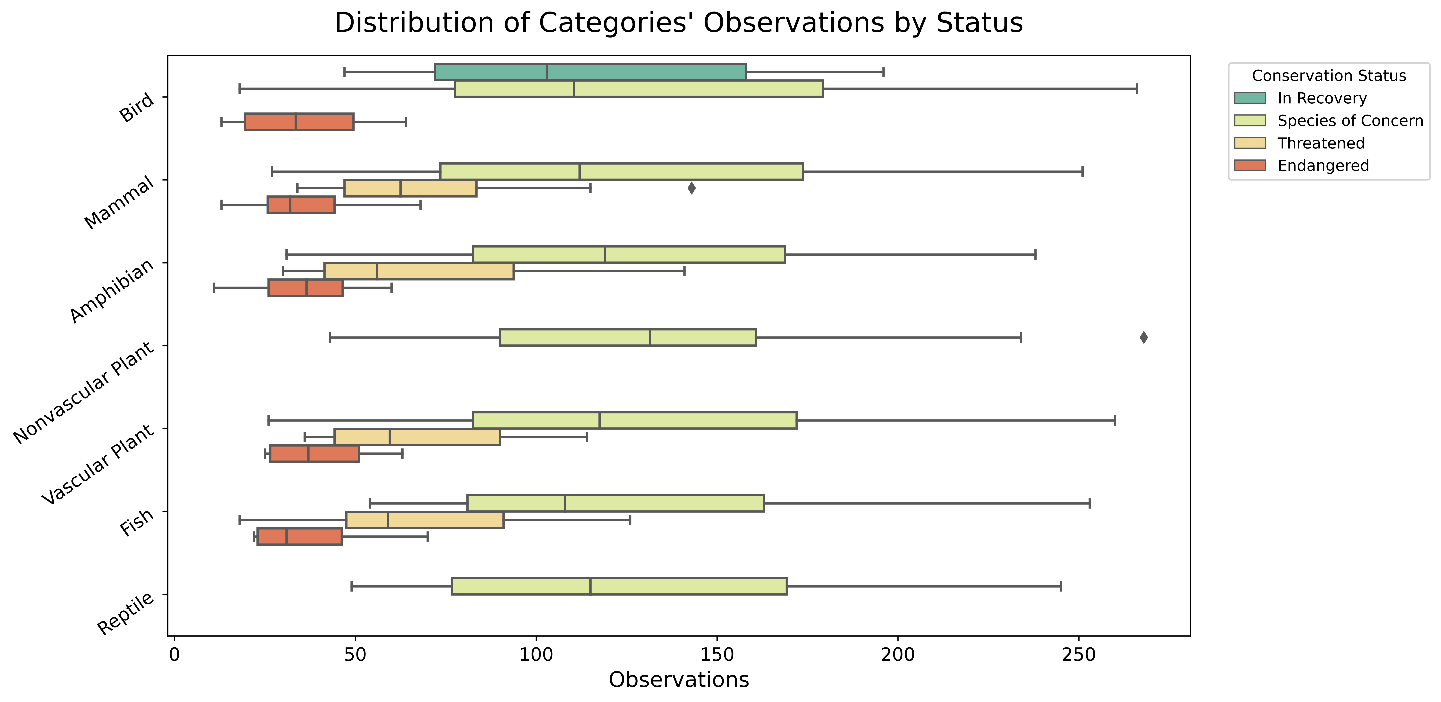


Given the geographic diversity of the parks, one might expect that they contain vastly different amounts of species. However, that is not the case. They observations by species category again mirror the Parks’ overall observation numbers. And a similar pattern emerges when we view all distributions of all animal classes across all parks.





So far, the data is looking a bit boring. I’m honestly starting to question the validity of the dataset itself (which I still do). But let’s keep pushing on and see if there’s anything interesting to uncover.



Now we’re getting somewhere. We can see very clearly that birds are the only category that have a portion of their population in recovery. We can also see that birds have no observations of species who are threatened. That could be a good thing or a bad thing. It would be good if there were no threatened birds to observe. It would be bad if no one counted any ‘threatened’ birds. Let’s see if the data contain any threatened birds with no counts.

**Graphical user interface, text, application, email

Description automatically generated**

So good news! The empty data frame tells us that there aren’t any birds in the threatened conservation status to count. The bad news is that the birds are the only population with *any* species in recovery. However, there are other elements of silver lining here. Both reptiles and nonvascular plants only have observations in the species of concern categories – and none in the threatened or endangered categories. To confirm that this is good news, let’s check the data once more.

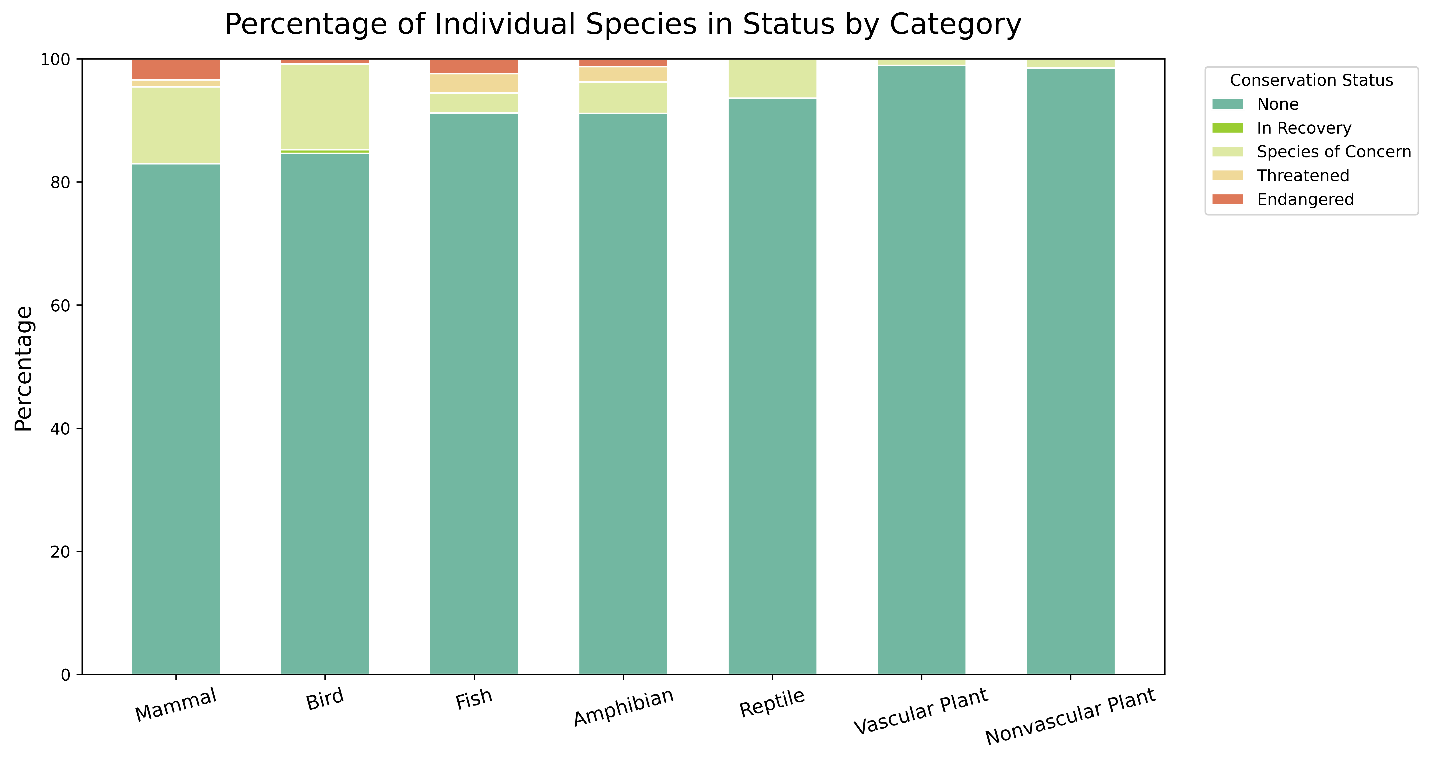
Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Great! With no species to count in these categories, we can truly feel good about this news. Let’s now look at the breakdowns of protected status across species categories.



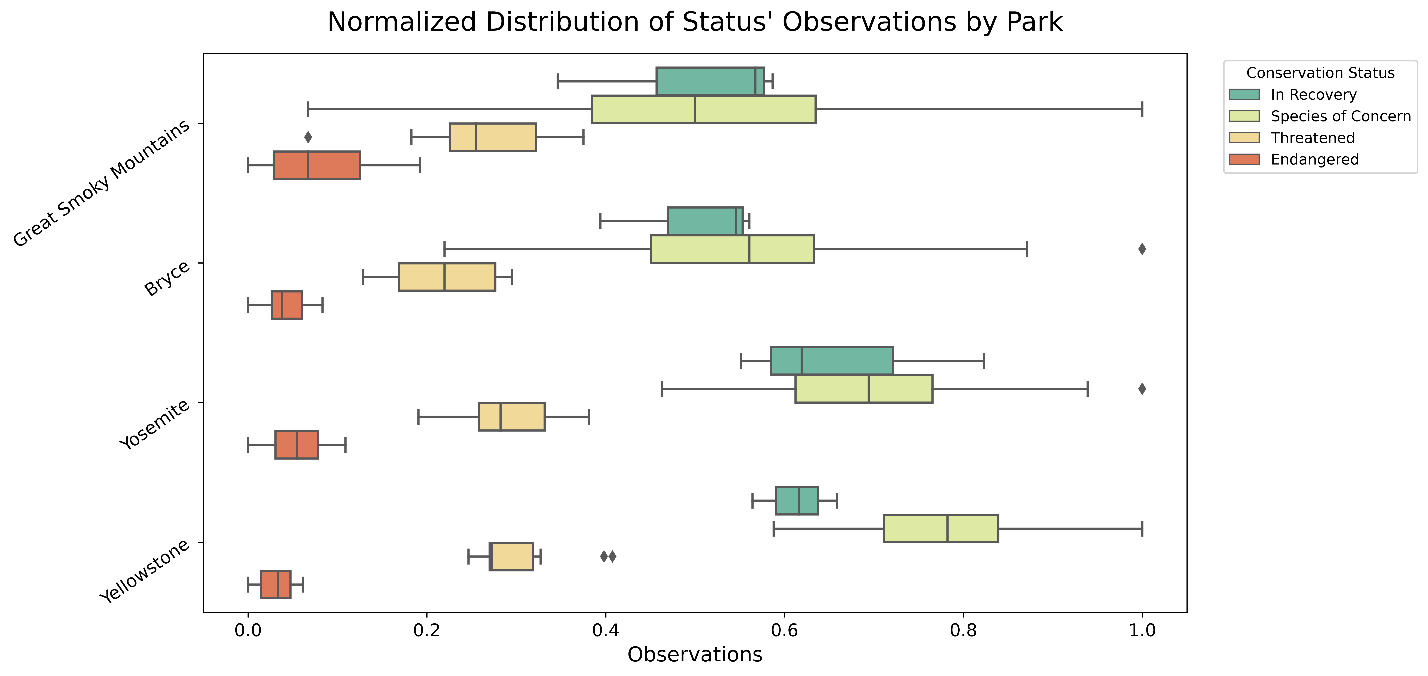
This view illustrates that mammals (the group that we humans happen to be a part of) seem to be the group currently most at risk. There are 176 individual species of mammal in the data. Of the total, 83% have no protection status (None) leaving 27% as protected. Of the protected, 13% are species of concern, 1% are threatened, and 3% are endangered. While the prognosis for birds above was leaning positive, it appears that (as a percentage of the individual species) they don’t fare as well. Of the 488 species of birds in the data, 14% are species of concern, and 1% are endangered. As we can see below, the trend is similar if we choose to look at observations of species within each category.

A picture containing background pattern

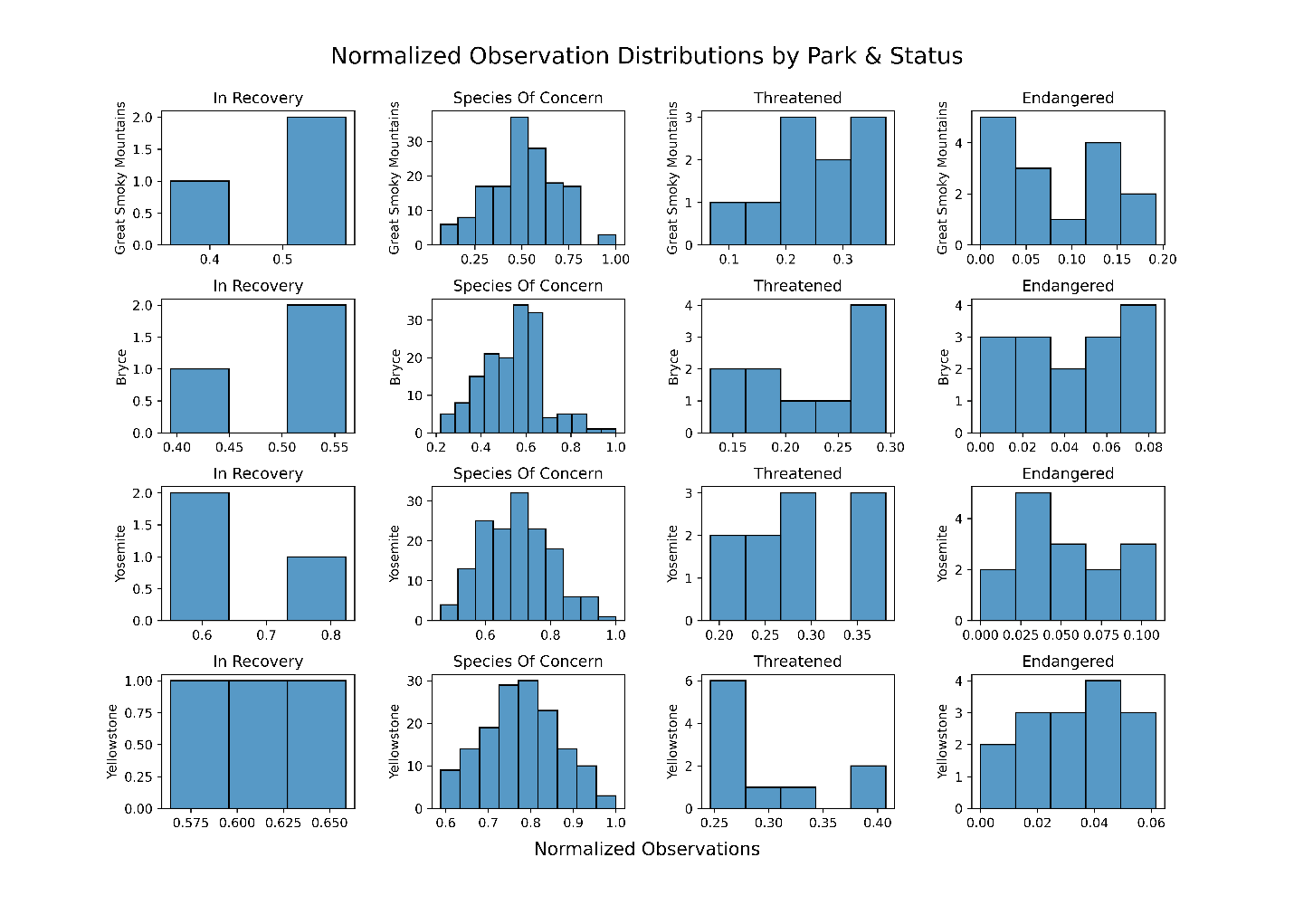
Description automatically generated

If one were advising policy makers based on this limited view of the data, one might recommend that special attention be paid to the mammals, birds, fish, and amphibians, as the reptiles, vascular plants and nonvascular plants all seem to be performing fairly well. An implementation of this could look like more frequent screening across the higher risk categories, and less frequent screening across the lower risk categories. In addition if the observation of protected species are similar across parks, one could recommend to rotate testing through individual parks each testing period in order to minimize costs. Let’s look at the normalized distributions of each protected category in each park, to see if t-tests are warranted.

If we normalize the number of observations within each park by the park’s min and max observations, we can see that they line up well. But perhaps not close enough to worry about testing means.



While most of these box plots look as if they might be normally distributed, let’s look more closely at the individual histograms.



Here we can see that many of these features are not normally distributed and actually look nothing like each other across parks. This makes sense given the relatively low number of species in most of these categories. In addition, only species of concern seems to have a large enough collection of species’ observations to warrant testing. For now it is safe to conclude that the species’ observations do not come from the same distribution and we shouldn’t recommend that the parks testing schedule be rotated.

**Potential Extensions**

This was an interesting exercise, but I would enjoy extending it by incorporating some time series data on individual species observations that might allow us to forecast what a ‘good’ number of observations would be and recommend values (per species) for the protected statuses. It would also be interesting to expand the inventory of the species to include those who were unique to each park.

Jupyter Notebook (with WAY more charts) is available here:

https://github.com/lightlySkewed/Biodiversity/blob/main/biodiversity.ipynb

Thanks for taking the time. 😉