As an aspiring data scientist enrolled in CodeCademy’s DS program, I’ve been asked to conduct an analysis on two data sets collected by the National Park Service 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 of the US Parks system.

Data Description

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 four United States national parks. Species\_info.csv groups the species into categories while also providing the species common name(s) and their current conservation status.

Data Challenges

Almost immediately, it became clear that there were duplicates and errata among the data. In the Observations data, for example Echinochloa Crus-Galli is listed twice at each park – often with different measurements.

Graphical user interface, text

Description automatically generated

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

Graphical user interface, application, table

Description automatically generated

While the data are not without issue, there is still a clear path ahead for the analysis. While one could have taken the average of the observations data, the approach here was to try and remain as conservative as possible. In that vein, the lowest of all observations for each species within each park was kept, and all other observations removed. 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 the data, 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

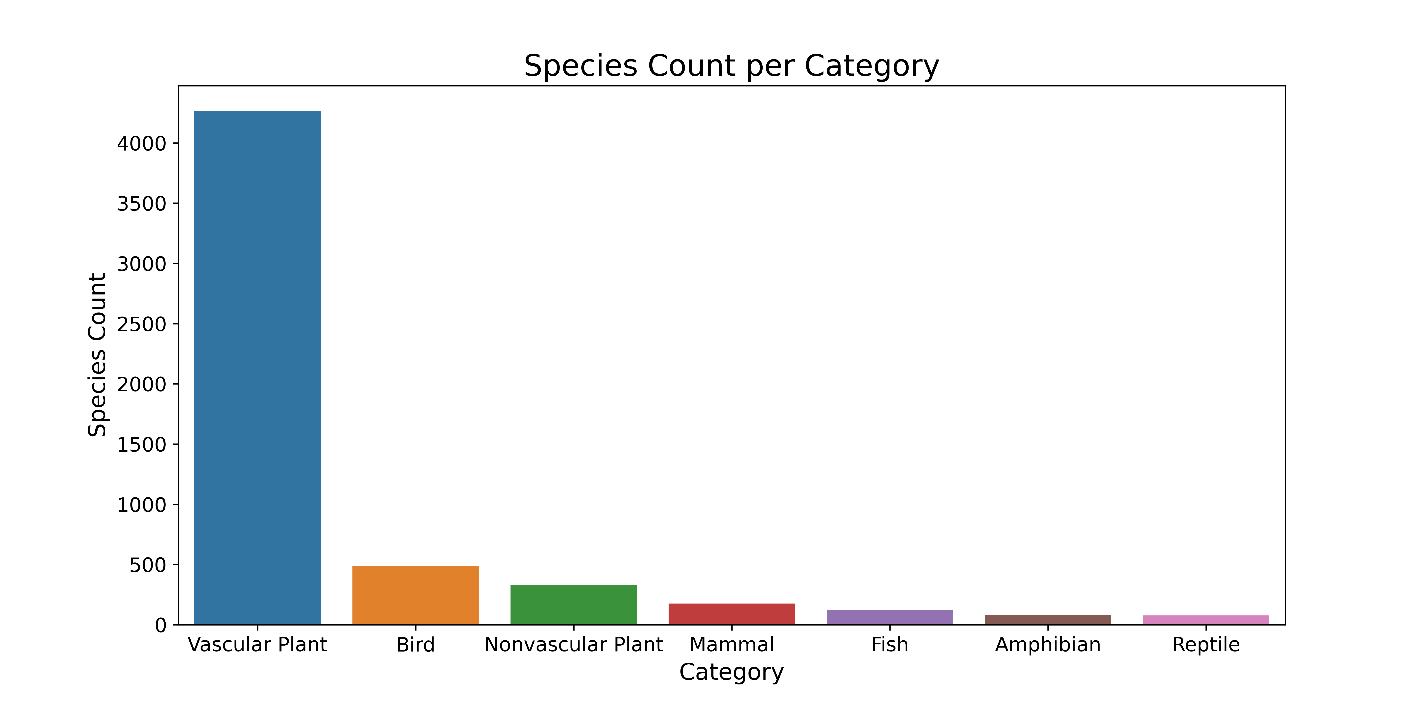
Description automatically generated

For a full description of the data cleaning process (including code) see Appendix A (LINK HERE).

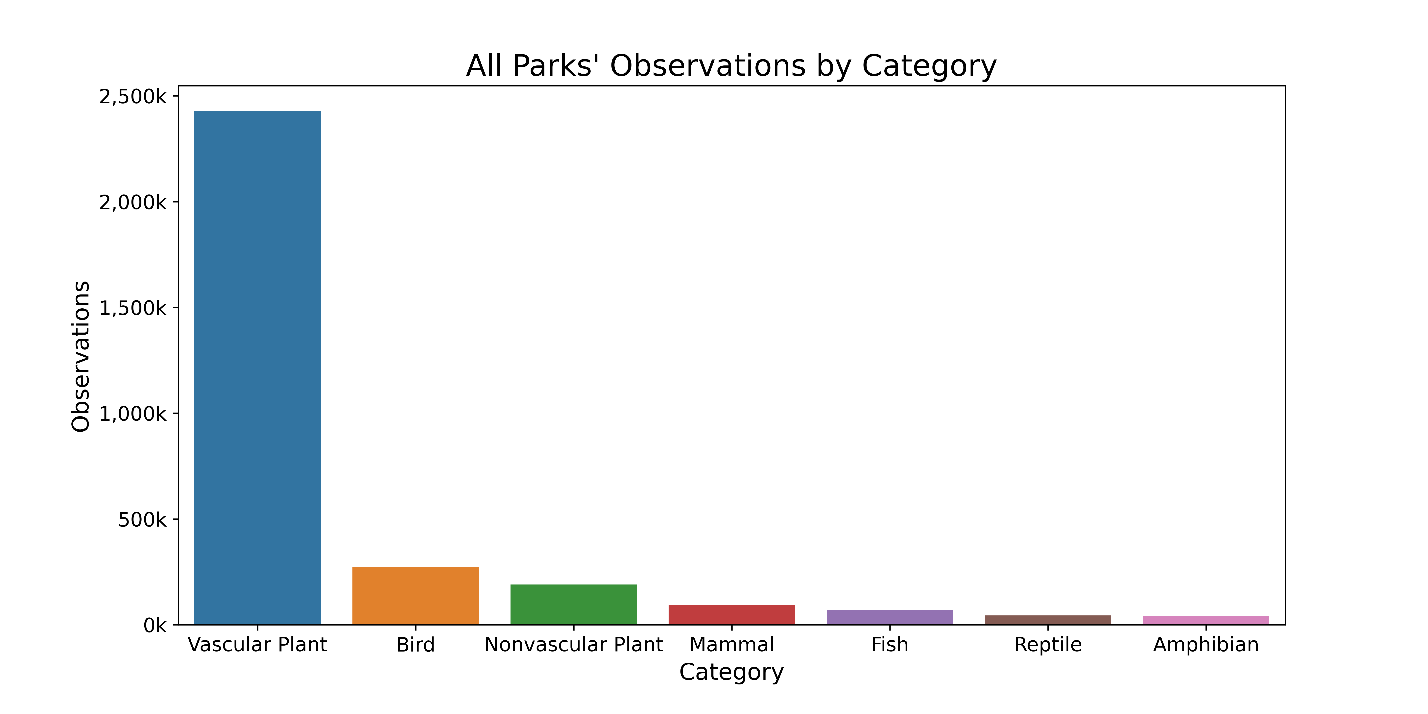
**Summary Findings**

The bulk of the species (96.8%) are in no conservation group (meaning they are not a in recovery, a species of concern, threatened, or endangered). Of those which are in a group (do have a conservation status); 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%.

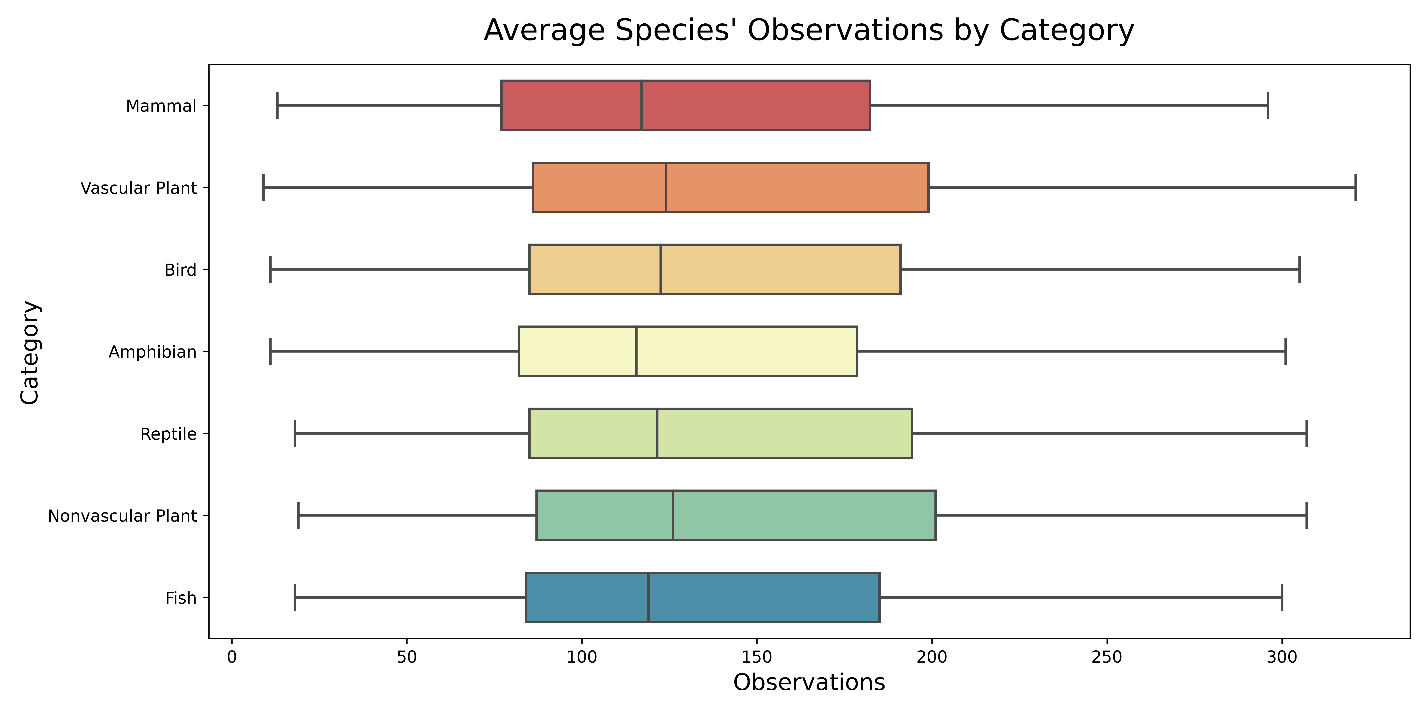


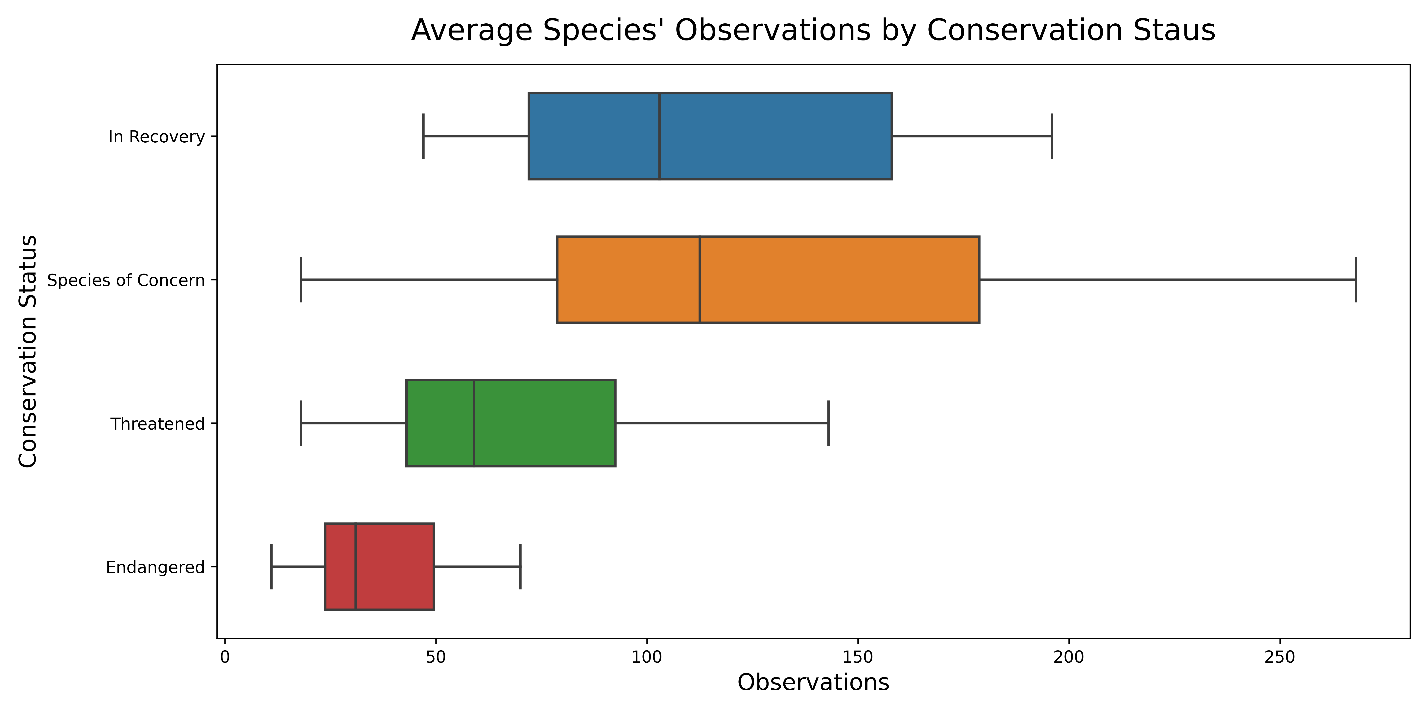
Not surprisingly this was almost mirrored by the total observations of each category across all parks.



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.



Observations per Park

Park Size

Observations per Acre

Having a feel for the data at a high level…

**Conclusions**

**Appendix A – Cleaning the Data**

* Link to notebook on git