Biodiversity in National Parks

Erik Owens

Codecademy

Portfolio Project

Project Scope

- Goals
  a. What is the distribution of conservation_status for animals?
  b. Are certain types of species more likely to be endangered?
  c. Are the differences between species and their conservation status significant?
  d. Which species were spotted the most at each park?
- Data
  a. We will use the *observations.csv* and *species_info.csv* files supplied from Codecademy.
- Analysis
  a. We will use tools such as numpy, pandas, matplotlib, and seaborne to explore, analyse, and visualise the data.
  b. Once completed we will be able draw our conclusions based on our goals for the project.

Observations of data - Part 1, Tidiness

- General observations
  - species_info: contains 4 columns and 5,824 entries
  - observations: contains 3 columns and 23,296 entries
  - the column "scientific_name" is common to both dataframes and could be used to join them together if required
  - There are no dates so we do not know the period that these observations took place
  - More than 3 million observations were recorded of 5,541 species in 4 parks.
- Duplicates
  - species_info does not contain duplicates
  - observations contains duplicates of species, but it seems that these are observations for the sames species at different parks
- Misc
  - no null values are reported by pd.info() however *species_info.csv* does seem to have empty cells for the column *conservation_status* which show as "NaN". These have been converted to "None" as this is categorical data that we can use to classify and quantify later.

Data

- There are 5541 unique species in the *species_info.csv* file.
- There are 4 parks where observations occurred (*observations.csv*)
- A total of 3,314,739 were recorded.

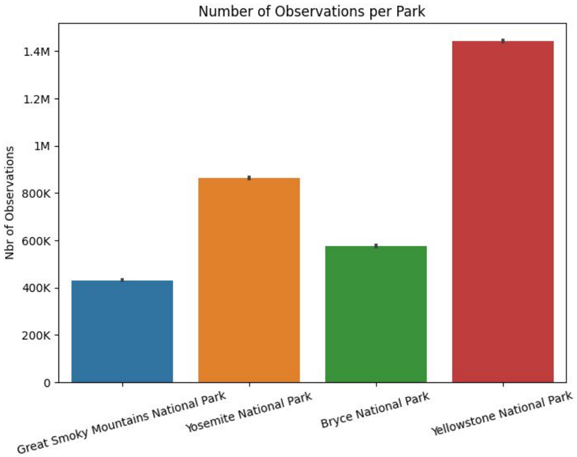
Observations of data - Part 2, Initial Analysis

- Species and Conservation Status
    - "Species of Concern" is the conservation status with by far the most number of species. This status has 7 categories of animals classified in it today. Birds, followed by Vascular Plants, and then Mammals are the top three categories of species.

Index of Visualizations

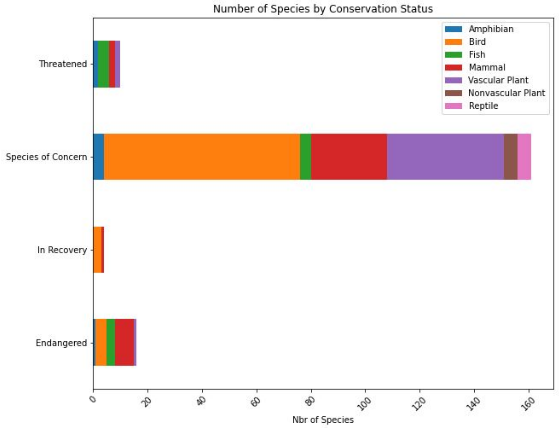
- Number of Observations per Park
- Number of Species by Conservation Status
- Distribution of Number of Species per Conservation Status
- Number of Species with status “Species of Concern”
- Number of Species ‘Endangered’
- Number of Species by Category & Conservation Status

Number of Observations per Park



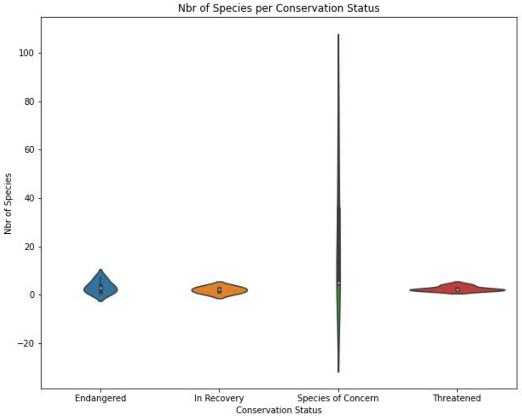
- We can see that Yellowstone National Park has a much higher number of observations than the other parks

Number of Species by Conservation Status



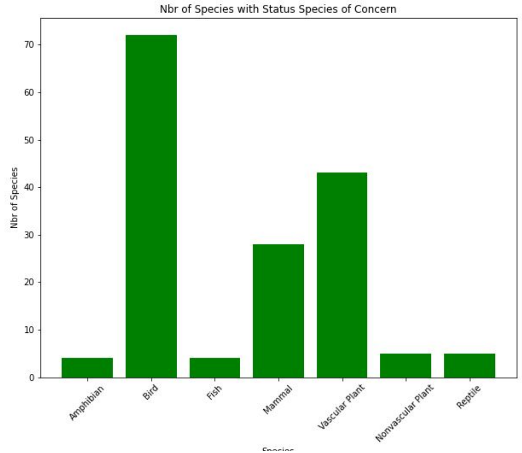
- Here I have excluded entries where the species has no conservation status.
- The data for conservation status by species is not a normal distribution since only 55.28% of the values are within 1 standard deviation of the mean.
- The conservation status “Species of Concern” merits further exploration.

Distribution of Number of Species per Conservation Status



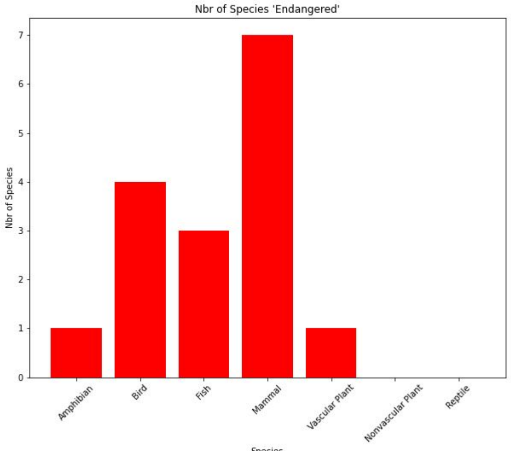
- This plot is not ideal for interpretation, but it does show that three of the four stati follow a similar distribution (tightly grouped), whereas the status “Species of Concern” contains many more species than the other stati and is highly distributed. If anything this shows that “Species of Concern” merits further exploration.

Number of Species with status “Species of Concern”



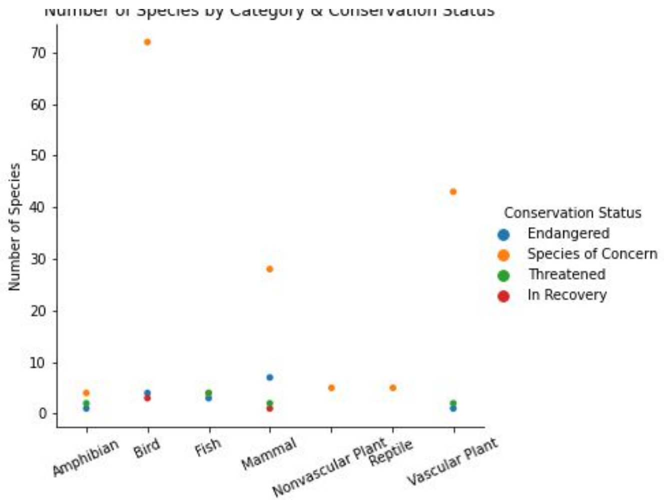
- We see here that the categories “Bird”, “Vascular Plant”, and “Mammal” contain a great deal more species than the other which contain around four or five species for each category.

Number of Species ‘Endangered’



- The pattern is similar here to what we saw for “Species of Concern”, but the category “Vascular Plant’ has moved out of the top three places to be replaced by “Fish”.

Number of Species by Category & Conservation Status



- We can clearly see that the category “Birds” has a significant number of species classified as “Species of Concern”, followed by “Vascular Plant”, and “Mammal”
  - This follows a similar pattern to what we saw for the status “Endangered”

Conclusions

What is the distribution of conservation_status for animals?

Answer: The data are not normally distributed. They are however, highly concentrated on the status "Series of Concern". However, within this conservation status we see that the data is more distributed across the seven categories of species with "Bird", "Mammal", and "Vascular Plant" making the top three, respectively.

Are certain types of species more likely to be endangered?

Answer: Yes, the category "Mammal" has the most number of species classified as "Endangered"

Are the differences between species and their conservation status significant?

Observation of question: this question lends itself to calculate the number of "scientific_name" per "category". However, unless I am mistaken, a "category" is not a species, but type of species. In other words "Cains lupus" is a species of type "Mammal". If the question were to simply be answered in terms of species ("scientific_name") per "conservation_status" we would simply see one species per "conservation_status" as the status is not differentiate per park. This would not be very interesting. We can therefore use the data from the previous question, but will have to add the conservation status "In Recovery" back in.

Answer: Yes for the categories of "Bird", "Mammal", and "Vascular Plant". The other categories do not display significant differences.

Which species were spotted the most at each park?

Answer:

Great Smoky Mountains National Park: Sonchus asper ssp. asper = 147 observations.

Yosemite National Park: Ivesia shockleyi var. shockleyi = 223 observations.

Bryce National Park: Valerianella radiata = 176 observations

Yellowstone National Park: Lycopodium tristachyum = 321 observations.