BIODIVERSITY FOR THE NATIONAL PARKS

Introduction to Data Analysis - Capstone project

By Arys Andreou

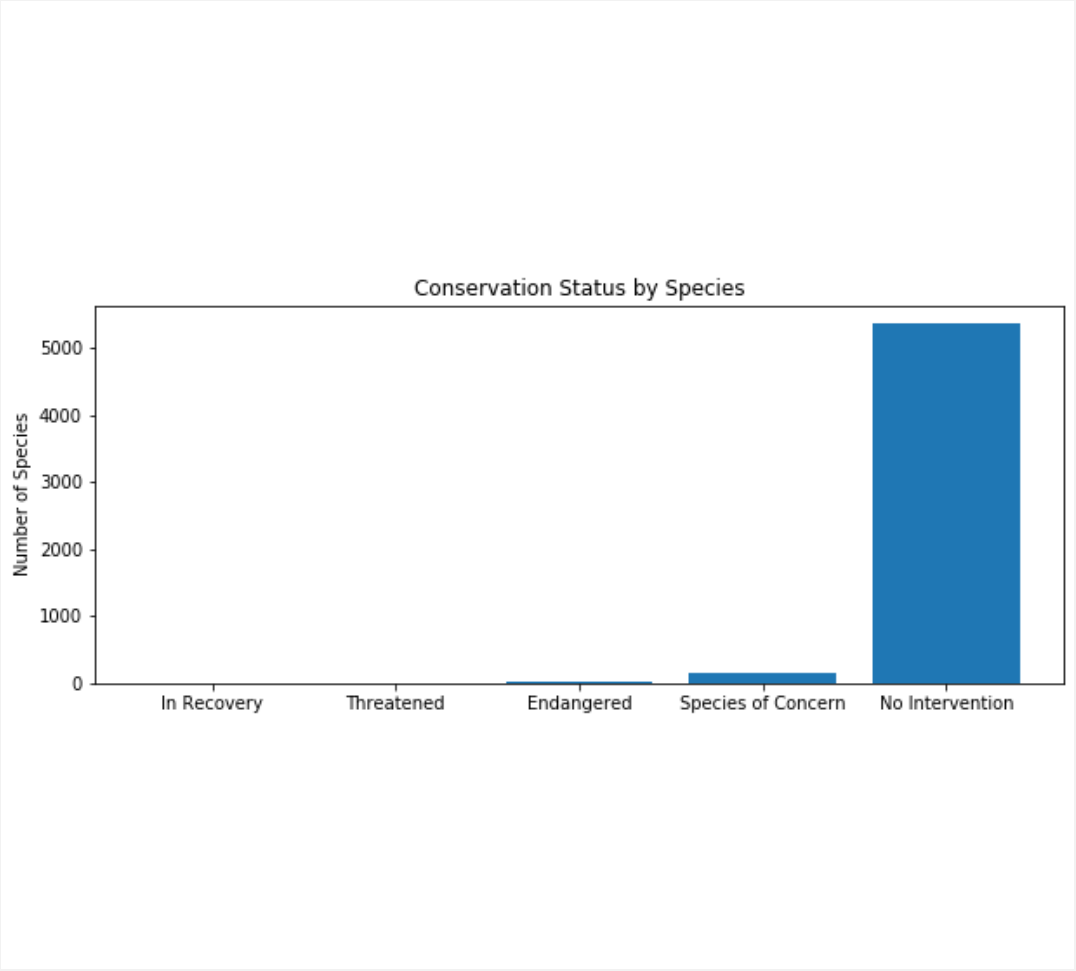
DESCRIBING THE DATA

- The species_info.csv appears to be a collection of data listing the protection status of several species in a number of national parks.
- There are 5541 unique species listed.
- These species are split into the following categories:
  - "Mammal", "Bird", "Reptile", "Amphibian", "Fish", "Vascular Plant" or "Nonvascular Plant"
- Some species are also given a conservation status of in order of importance.
  - "Species of Concern", "Threatened", "Endangered" and "In Recovery"
- The highest percentage of endangered species are the Mammals and the least are the Fish excluding the plants.

CONSERVATION STATUS

- Here we can see the populations of the species in every status category. Those without a value have been listed under "no intervention".
- Bellow is a table showing the percentage of each species which is protected. Mammals appearing to be the most endangered.

category	not_protected	protected	percet_protected
Amphibian	72	7	0,088607595
Bird	413	75	0,153688525
Fish	115	11	0,087301587
Mammal	146	30	0,170454545
Nonvascular Plant	328	5	0,015015015
Reptile	73	5	0,064102564
Vascular Plant	4216	46	0,010793055



PROTECTED SIGNIFICANCE

MAMMAL - BIRD

- Performing a Chi-Squared test on mammal and bird population to identify if the protected percentage is significantly difference returns a p-value of 0.68759.
- The difference **is not** significant. We can assume that the difference is a result of chance.

MAMMAL - REPTILE

- Performing a Chi-Squared test on mammal and reptile population to identify if the protected percentage is significantly difference returns a p-value of 0.03835.
- The difference **is** significant

RECOMMENDATION

- Based on the results of the Chi-Square test from the data given we can infer that the mammals are more likely to be endangered than reptiles.
- Further examination of the species has shown a significant difference between mammals and Fish with a p-value of ~0.056.

SAMPLE SIZE DETERMINATION

- Baseline rate as recorded from the previous year is 15%
- A significance of 90% was given
- Given a minimum percent change of 5% the detectable effect is 33.33%
- Using a sample size calculator and this data the recommended sample size is 510 sheep.
- Observations per week for Yellowstone and Bryce are 507 and 250 respectively.

This means that the recommended sample size in order to A/B test the efficiency of their disease reduction actions is **510**. In order to collect his data they would require approximately **2** weeks for Yellowstone and **4** for Bryce.

It appears that the best place for observation is Yellowstone National Park.

As demonstrated by the chart bellow and the results from the previous slide we can see that the best location for collecting samples is at the Yellowstone National Park