Biodiversity Capstone project

Ruslan Jevpsihejev for Codecademy

Input data

Species_info.csv is a comma-separated value file containing the species category, scientific name, common name and conservation status on 5541 unique species

Input data

Along the available categories are:

- Mammal
- Bird
- Reptile
- Amphibian
- Fish
- Vascular Plant
- Nonvascular Plant

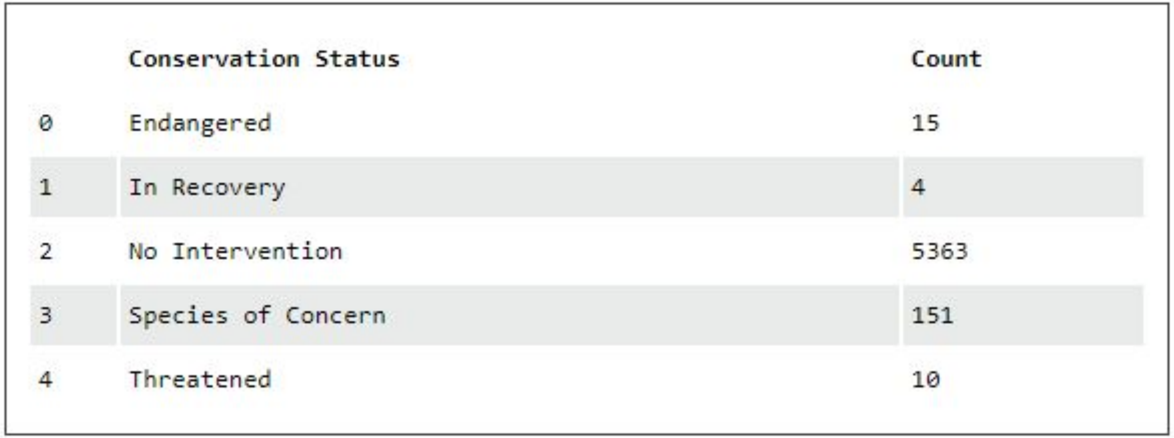
Input data

Possible conservation statuses are:

- Endangered
- Threatened
- Species of Concern
- In recovery
- Neither of those which we called ‘No intervention’

Conservation status counts

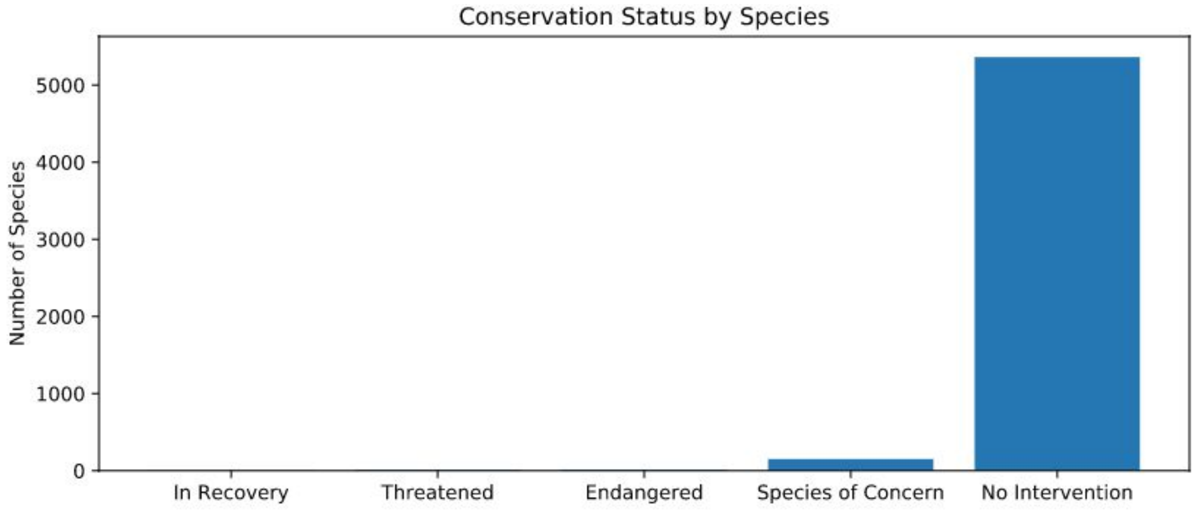
I was able to calculate how many species belong to each possible conservation status:



	Conservation Status	Count
0	Endangered	15
1	In Recovery	4
2	No Intervention	5363
3	Species of Concern	151
4	Threatened	10

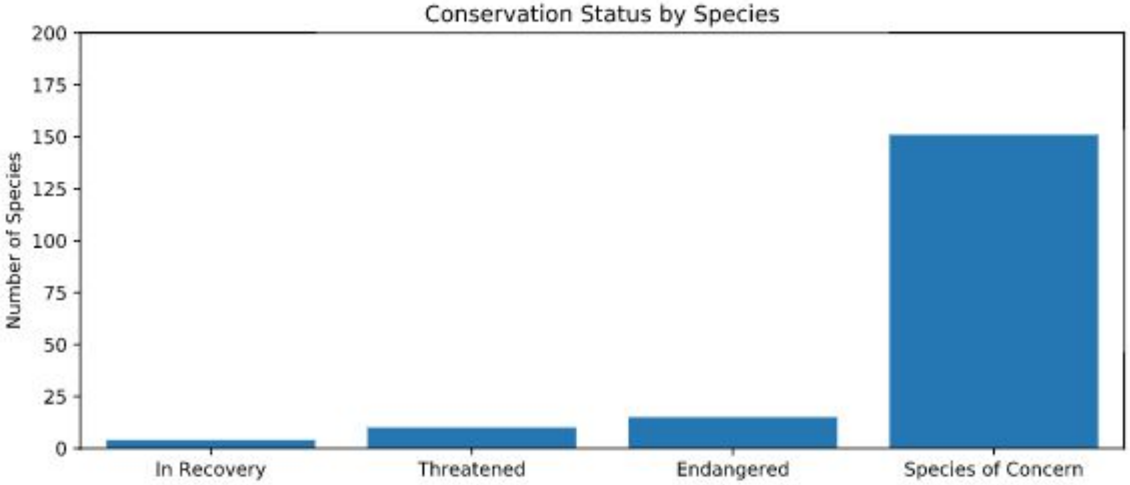
Conservation status counts

Luckily for all of us,most of the species require no human intervention in order to continue surviving. The scale of this can be seen on the graph:



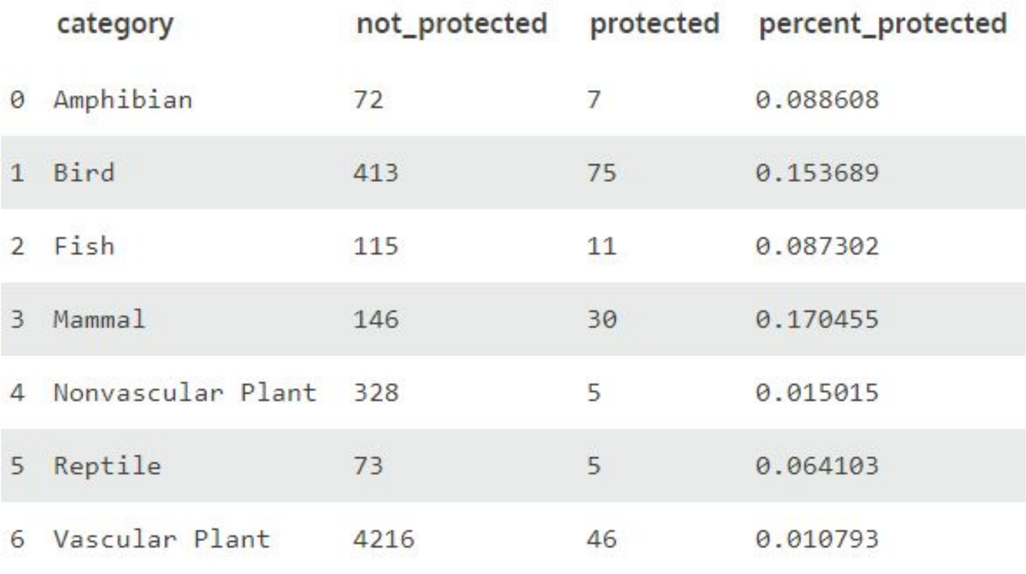
Conservation status counts

Let’s zoom in a bit on the graph to see how the species are spread between statuses that do require intervention. As we ca see, majority is again the least severe status, that’s a positive note



Investigating Endangered Species

Here we are trying to understand whether certain categories of species are more likely to be endangered. First let’s get the data:

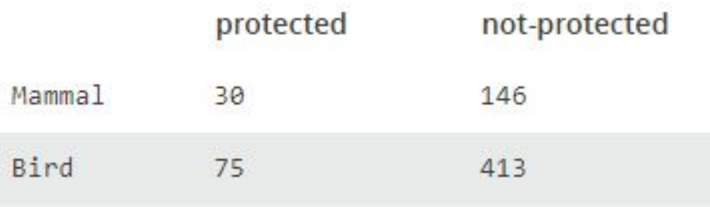


	category	not_protected	protected	percent_protected
0	Amphibian	72	7	0.088608
1	Bird	413	75	0.153689
2	Fish	115	11	0.087302
3	Mammal	146	30	0.170455
4	Nonvascular Plant	328	5	0.015015
5	Reptile	73	5	0.064103
6	Vascular Plant	4216	46	0.010793

Investigating Endangered Species

We see that some categories indeed have more species protected percent-wise, but there are also differences in absolute values, what if it’s like that only due to the diversity of the category itself?

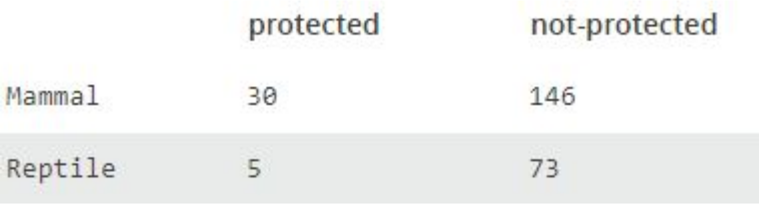
For example, let’s take two species with similar percentages, but different absolute values: Mammals and birds. Mammals are less common, yet the percentage of protected species is roughly the same - 17% as opposed to 15% for birds. Let’s use chi-square test.



	protected	not-protected
Mammal	30	146
Bird	75	413

Investigating Endangered Species

The result of chi-square test showed that there is no significant difference between birds and mammals, so one is not more likely to be endangered. Now let’s compare some other categories: same mammals against reptiles There’s a bigger difference in percentages 17% for mammals against 6% for reptiles



	protected	not-protected
Mammal	30	146
Reptile	5	73

Chi-square shows that actually mammals *are* more likely to become endangered than reptiles! If there are two, that there can be others as well.

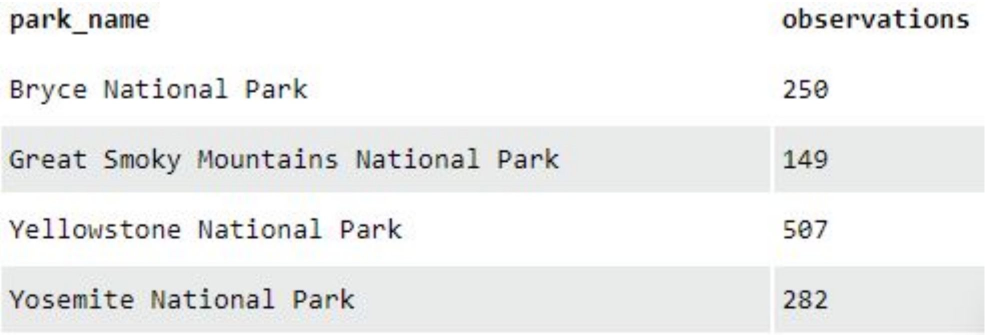
Pt. 2 Input data

Next file I was offered to analyze observations.csv This file contains data on how many times each species has been observed in some national parks. Columns are scientific name, park name and observations.

In order for us to better understand what species it was we merged observations.csv with species.csv. Now we have common name available for search in our table.

Pt. 2 Input data

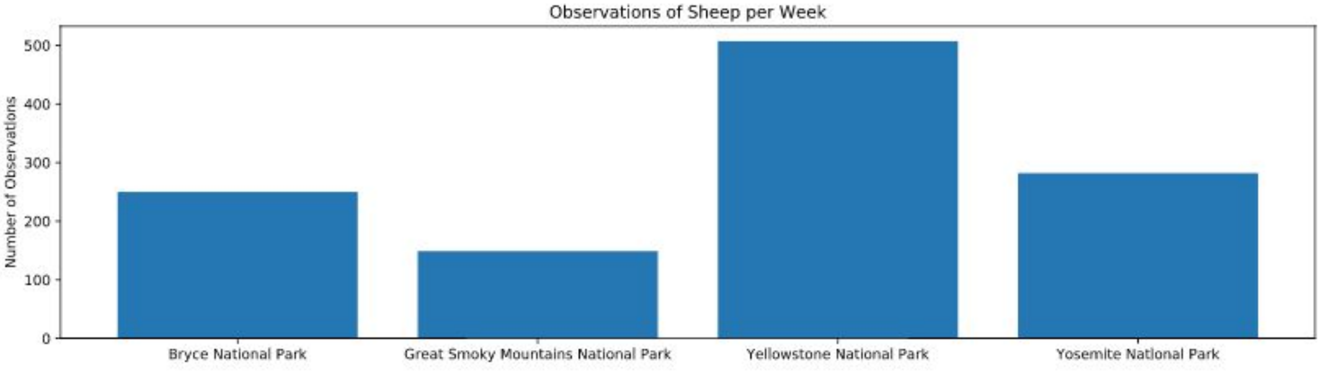
Since we were only interested in sheep we filtered out mammals that had “Sheep” in their common name. There was no need to diversify among the results, any sheep will do. Resulting table contained the number of observations of any sheep grouped by different parks:



park_name	observations
Bryce National Park	250
Great Smoky Mountains National Park	149
Yellowstone National Park	507
Yosemite National Park	282

Sheep observations by park

Let’s visualize the data in our table from the previous slide:



We can conclude that Yellowstone National Park has got the most observations of sheep. This could either mean, that there are more sheep in that par, or they are just more observant.

Foot and Mouth Reduction Effort

Park Rangers at Yellowstone National Park have been running a program to reduce the rate of foot and mouth disease at that park. But they also want to know that if reduction happens it is indeed because of the program, so they keep track of how good job they did

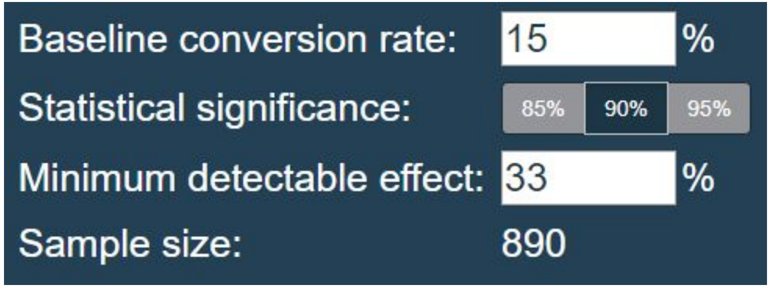
Current disease rate is 15%, so that is our baseline

They would like to drop it to at least 10%, so minimum detectable effect is 33%

Standard statistical significance is 90% there was no special need to go with another one

Foot and Mouth Reduction Effort

So, using our fancy calculator we determined that in order for the data changes to have statistical significance rangers have to observe around 890 sheep



Baseline conversion rate: 15 %

Statistical significance: 85% [✓ 90%] 95%

Minimum detectable effect: 33 %

Sample size: 890

Given the weekly rate of 507 for Yellowstone Park they would need approx. 2 weeks for the collection of the right amount of data

Thank you!