codecademy

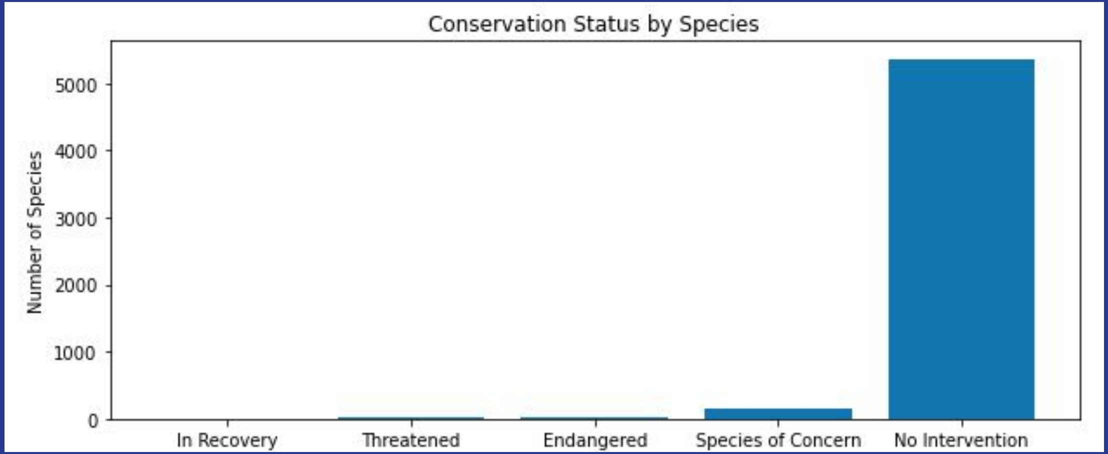
Biodiversity in National Parks

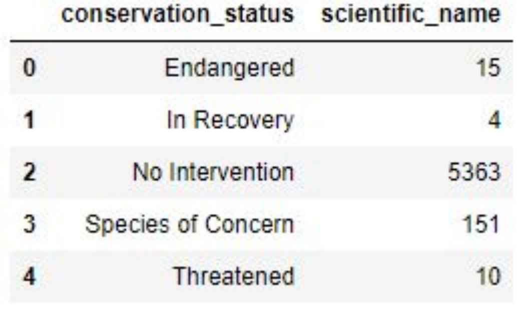
Data Analysis Capstone Project
Paul Kim
10/11/2020

Background

- For this project, we were to act as a data analyst for the National Park Service. We were to help them analyze data on endangered species from several different parks.
- One of the main asks was to analyze whether there are any patterns or themes to the types of species that become endangered.
- The source of the data to be analyzed: ‘species_info.csv’
- The file includes categories of species, their scientific and common names, and conservation status
- The different categories include: mammals, birds, reptiles, amphibians, fish, vascular and nonvascular plants
- There are a total of 5541 different species (**scientific** and **common names**)
- Majority of species fall under the conservation status of No Intervention

The conservation status of the species

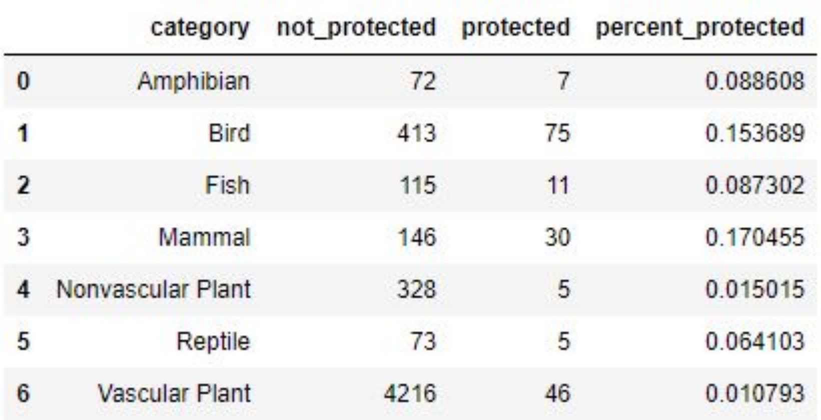




	conservation_status	scientific_name
0	Endangered	15
1	In Recovery	4
2	No Intervention	5363
3	Species of Concern	151
4	Threatened	10

- After analyzing the data using Panda Python library and visualizing the results in Matplotlib, approximately 97% of all the species in national parks do not require any protection.

Percentage of Endangered Species

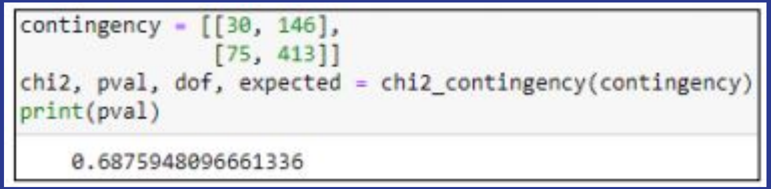


	category	not_protected	protected	percent_protected
0	Amphibian	72	7	0.088608
1	Bird	413	75	0.153689
2	Fish	115	11	0.087302
3	Mammal	146	30	0.170455
4	Nonvascular Plant	328	5	0.015015
5	Reptile	73	5	0.064103
6	Vascular Plant	4216	46	0.010793

- The number and percentage of species that are protected were found by finding species that were NOT labeled with ‘No Intervention’

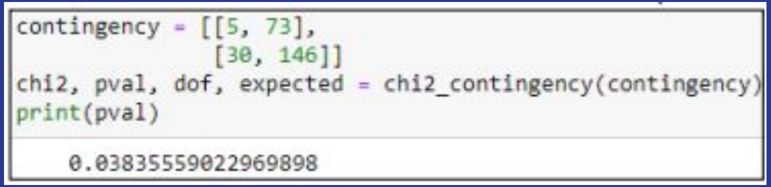
- Based on the output, it seems, for example, that Mammal is more likely to be endangered than Amphibians (17% protected vs. 8.8% respectively).

Chi Square Significance Test



```
contingency = [[30, 146],
               [75, 413]]
chi2, pval, dof, expected = chi2_contingency(contingency)
print(pval)

    0.6875948096661336
```



```
contingency = [[5, 73],
               [30, 146]]
chi2, pval, dof, expected = chi2_contingency(contingency)
print(pval)

    0.03835559022969898
```

- We used a Chi Square significance test to verify the hypothesis that mammals are more likely to be endangered than birds

- When comparing two pieces of categorical data, a Chi Square test is a good method to utilize

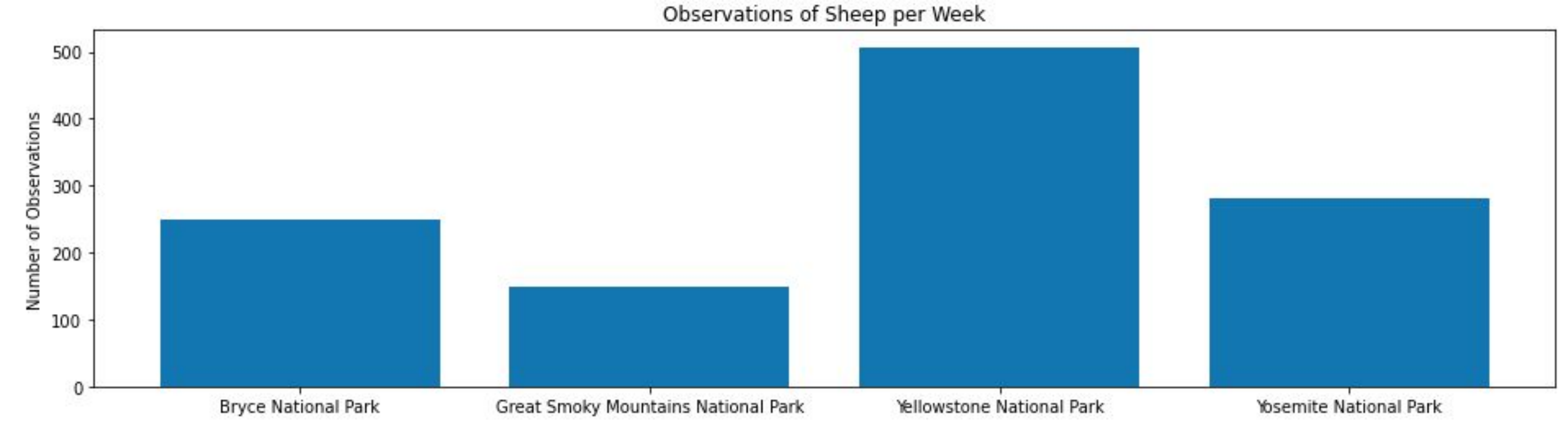
- The first test (screenshot 1) shows a pvalue > 0.05, thus the difference between Mammal and Bird categories is NOT signifncant
- The second test (screenshot 2) shows a pvalue < 0.05, thus the difference between Mammal and Reptile IS significant

Recommendation

- Based on the results, I would advise the conservations to focus their efforts on Mammals instead of Reptiles
- Similarly, Birds show to be the second most endangered categories shadowing Mammals
- Even though 97% of the species do not require protection, special attention should be placed to the vulnerable species (that 3%) to preserve their existence

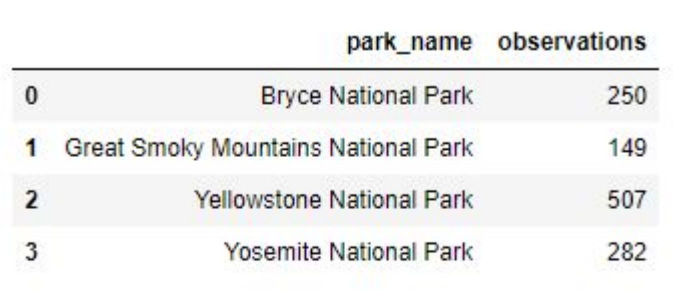
Sheep in National Parks

- Another dataset that was analyzed is from: ‘observations.csv’
- This dataset contained the number of observations of the species in each National Park that was used in the previous graphs
- The below graph illustrates the amount of observations of sheep in each National Park per week



Sheep in National Parks

- Three type of sheep species were identified in the observed ata:
  - **Ovis aries** = Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)
  - **Ovis canadensis** = Bighorn Sheep
  - **Ovis canadensis sierrae** = Sierra Nevada Bighorn Sheep
- Based on the data set, we found that 15% of the sheep at Bryce National Park have foot and mouth disease
- Park Rangers at Yellowstone National Park have been trying to reduce the rate of disease in their park
- The scientists want to test whether the program is working
- They want to be able to detect reductions >= 5% points



	park_name	observations
0	Bryce National Park	250
1	Great Smoky Mountains National Park	149
2	Yellowstone National Park	507
3	Yosemite National Park	282

Sheep in National Parks - Sample size

- To determine the sample size, we were instructed to use the Codcademy sample size calculator found here: https://s3.amazonaws.com/codecademy-content/courses/learn-hypothesis-testing/a_b_sample_size/index.html

- Inputs for the calculation, as follows:
  - Baseline conversion rate = 15%
  - Statistical significance = 90%
  - Minimum detectable effect = 5 * 100/15 = 33.3%
- The sample size produced = 870
- Taking this sample size and the table from the previous slide, the number of weeks to observe enough sheep was calculated for Bryce National Park and Yellowstone National Park:
  - Bryce National Park = 870 / 250 = 3.48
  - Yellowstone National Park = 870 / 507 = 1.71

THANK YOU!