



CodeAcademy BioDiversity

Are some species more likely to be endangered than others?"

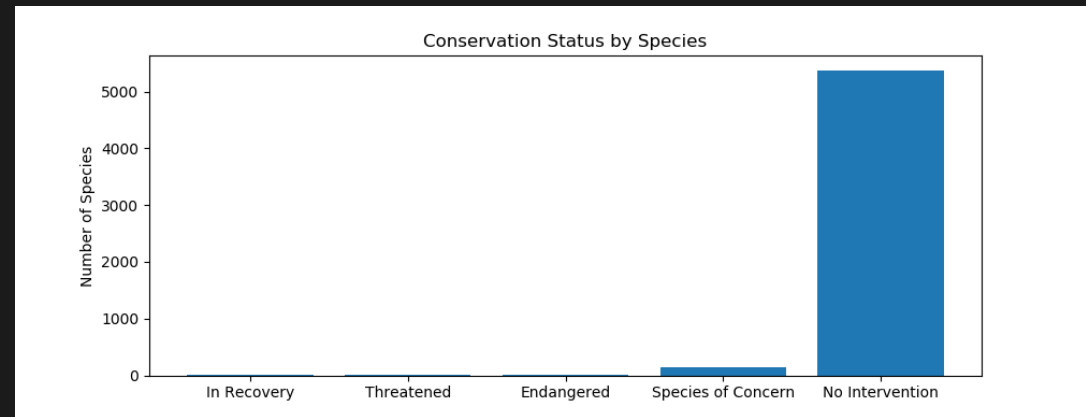
In the national parks data

Project Overview

- The Project had 2 sets of data as 2 csv files.
- 1) In the form of species info which contained information regarding species which contained columns like name, category, common name and conservation status.
- 2) The other file was observations.csv which contained data regarding type of species and their observation location and frequency.

Part 1

- The first part of the research involved counting the number of unique species and unique categories of species along with different status of conservation . Which Were Endangered ,in recovery, species of concern, Threatened and nan.
- We used the fillna function to replance nan with a new category called no intervention .
- We found the following data
- Conservation Status : Number of species
- Endangered 15
- In Recovery 4
- No Intervention 5363
- Species of Concern 151
- Threatened 10



Part 1

- We then wanted to ascertain how much percentage of the species in each category was protected, we did this by grouping category and scientific name to get the total number of species where the conservation status was no intervention and the total number where the conservation status was not no intervention which was placed in the protected column, we then divided protected/protected+nonprotected to get the percentage of protected species in each category. We got the following result

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

Part 1

We then used the chi2 test the null hypothesis that the difference was due to chance

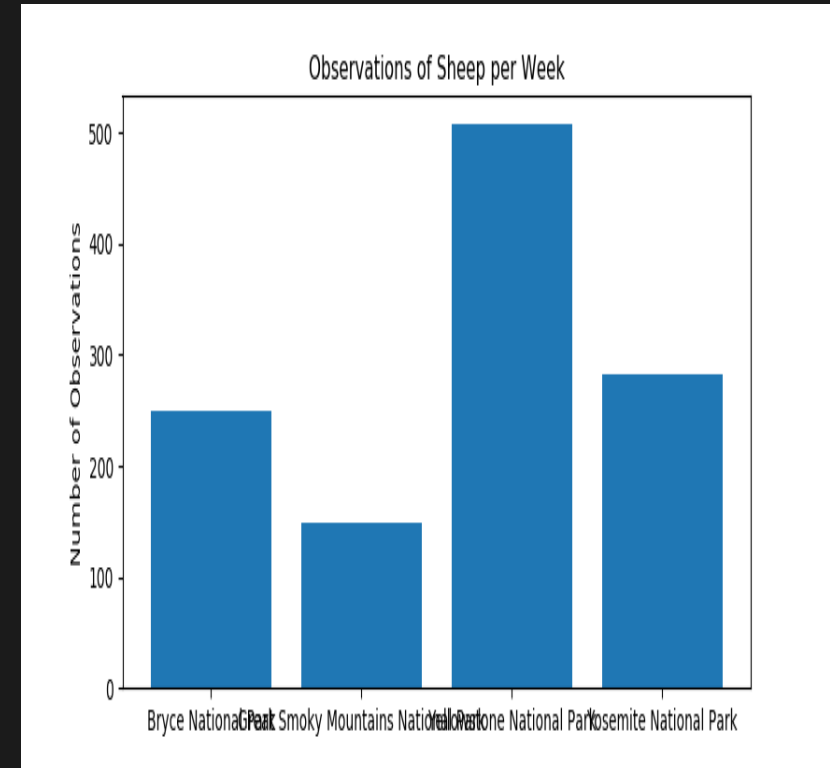
We used sample data from to categories Mammals and reptiles,

We found that the pvalue results from the chi2 test were significant

0.0383555902297 was the pvalue which was below 0.05 and therefore our null hypothesis was rejected and we could conclude certain species were more likely to be in endangered

Part 2

- In the second part the national park send us another data set observations. The Data Set used scientific names of the species so we had to we had to create new column which would check if the animal was type of sheep from the original species_info.csv , Once we tested weather a particular species was sheep using the lambda function , We then merged it with the observation dataframe which contained location and number observations
- We merged the two dataframes and grouped it by parkname to determine the sum of observations in each park



The foot and mouth Disease

Baseline

- The baseline was set to 15 % because as per the exercise last year 15 % of the sheep had foot and mouth disease at the Bryce national park . They want to check for reduction

Minimum Detectable effect

- This year it was 10 % at Yellowstone that is a drop of 5/15 , 33 % which is significant , In order to confirm their findings they wanted determine the sample size to see if their program was working

Statistical significance

- We are choosing 90 % as the statistical significance as this is how sure of the result that we derive from the sample size.

The sample size was 510 as per the calculator and the 3 parameters mentioned.

Conclusion

- Based on the result we have concluded the following
- There is a significant difference between the levels of endangerment between different types of species . In our case it was reptiles vs mammals where the Pvalues were below 0.05 ,Certain species of Mamals and birds seemed to be most in danger.
- I would recommend the national parks to be careful about poaching and provide more protection for birds and mamals .
- It could also be weather conditions that might not suit species and therefore shifting might be in order to protect them.