Biodiversity in National Parks

By: Chenyuan Li

The Codecademy Data Analysis Capstone Project

Background Information

- Acting as a job of a data analyst for the National Park Service and helping them analyze data on endangered species from several different parks.
- Performing some data analysis on the conservation statuses of these species and investigating if there are any patterns or themes to the types of species that become endangered. For this project, I will analyze, clean up, and plot data, pose questions and seek to answer them in a meaningful way.

Files that are provided for this project:

1. species_info.csv
2. observation.csv

species_info.csv

- The *species_info.csv* file is one of the main files that we use in order to analyze for our topic in the biodiversity in National Parks
- Using the species.scientific_name.nunique() function in Python, we acknowledge that there is a total of 5541 unique species in these national parks.
- The title rows for the data are category, scientific_name, common_names, and conservation_status
- In total, seven categories of species are collected. These are “Mammal”, “Bird”, “Reptile”, “Amphibian”, “Fish”, “Vascular”, and “Nonvascular Plant”
- There are also five different conservation status for the species in the parks. These conservation statuses are “Species of Concern”, “Endangered”, “Threatened”, “In Recovery”, and “No Intervention”

observation.csv

- This file's row consists of scientific_name,park_name, and observations
- The information regarding the animal's name, the name of the park, and the number of species' appearance were stored under each row correspondingly

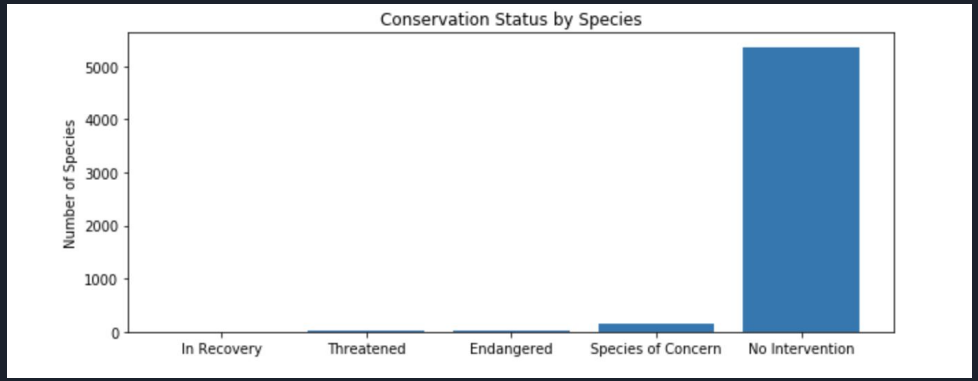
Categories of Species(Conservation Status)

Combining the two charts together, we can easily see that there are only four species in recovery, and we have to mainly focus on the endangered species. I also want to point out that the graph in this case is not as effective as the table due to the “No intervention” category being an outlier which then cause the minimization of other categories on the graph.

The table below shows the number of species by their conservation status.

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

The graph below is a visual representation of the data from the left



Regarding to endangered species:

In regards to the endangered species, the ones that need to be protected the most, the chart below shows a the percentage of protected species in each category:

	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

Looking at the percent_protected column, the one with the highest percent_protected rate, the mammal category, can be viewed as the category of species that are most likely to be endangered

Significance Tests

- Base on the previous slide, the percent_protected for mammal(~0.17) and bird (~0.15) are relatively close. But do we know if mammals are more likely to be endangered than the birds?
- To validate the statement that mammals are more likely to be endangered than the birds based on their protected percentage, we need to perform the Two Chi-Square Significance Tests

Test #1:

- ***Chi-Square Test:***
  - Null Hypothesis: There are no significant differences between the Mammals and Birds
- In order to reject the null hypothesis, we need a *p-value less than 0.05
- In python, we use the function chi2_contingency(contingency) to determine the p-value
- After calculation, we get a p-value of 0.69
- Since 0.69 > 0.05
- We cannot reject our null hypothesis and therefore there are no significant differences between the mammals and the birds

Test #2:

- ***Chi-Square Test:*** Comparing Mammals and Reptiles
  - Null Hypothesis: There are no significant differences between the Mammals and Reptiles
- Like test #1, we use the function chi2_contingency(contingency) to determine the p-value
- After calculation, we get a p-value of 0.04
- Since 0.0.4 < 0.05
- We reject our null hypothesis and therefore there are significant differences between the mammals and the reptiles

Based on the Significance Tests

- From looking at the percent_protected column, the higher the the percentage, the more likely the species is becoming endangered species
- It appears that vascular and nonvascular plants are the least likely to become endangered given by their low percent_protected values
- Even though mammals and birds do not have significant difference between each other based on the previous chi-square significance tests, they are still two of the species that are endangered

Sheep Observation

We are given the information that some scientists are studying the number of sheep sightings at different national parks which are stored in the observation.csv file

Table and graph below show the number of sheep appearance

	scientific_name	park_name	observations
0	Vicia benghalensis	Great Smoky Mountains National Park	68
1	Neovison vison	Great Smoky Mountains National Park	77
2	Prunus subcordata	Yosemite National Park	138
3	Abutilon theophrasti	Bryce National Park	84
4	Githopsis specularioides	Great Smoky Mountains National Park	85

Base on the figures above, Yellowstone National Park has the most sheep appearance

Foot and Mouth Disease

- We are given the information that 15% of the sheep at Bryce National Park have the Foot and Mouth Disease
- We are also given the information that Park Rangers at Yellowstone National Park are running a program to reduce the rate of Foot and Mouth Disease. They aim to reduce the rate for about 5%
- We want to verify that this program that they are running is actually effective and working(A/B Test)
- Using the given resource: Codecademy's sample size calculator, we need to input the following data:
    - Baseline Conversion Rate: 15%
    - Minimal Detectable Effect: 33.33%
    - Statistical Difference: 90%
- The given calculator calculated a sample of 510 per variation, which means that they need 510 sheep observation for both Bryce and Yellowstone National Park
- From the previous table, Bryce National Park has 250 sheep, which means we need, 510/250, roughly two weeks to verify the effectiveness of the program.
- We also need about, 510/507, roughly one week at Yellowstone National park to verify the effectiveness of the program