# Introduction to Data Analysis Capstone Project: Biodiversity for National Parks

Nicole Macon

May 20, 2018

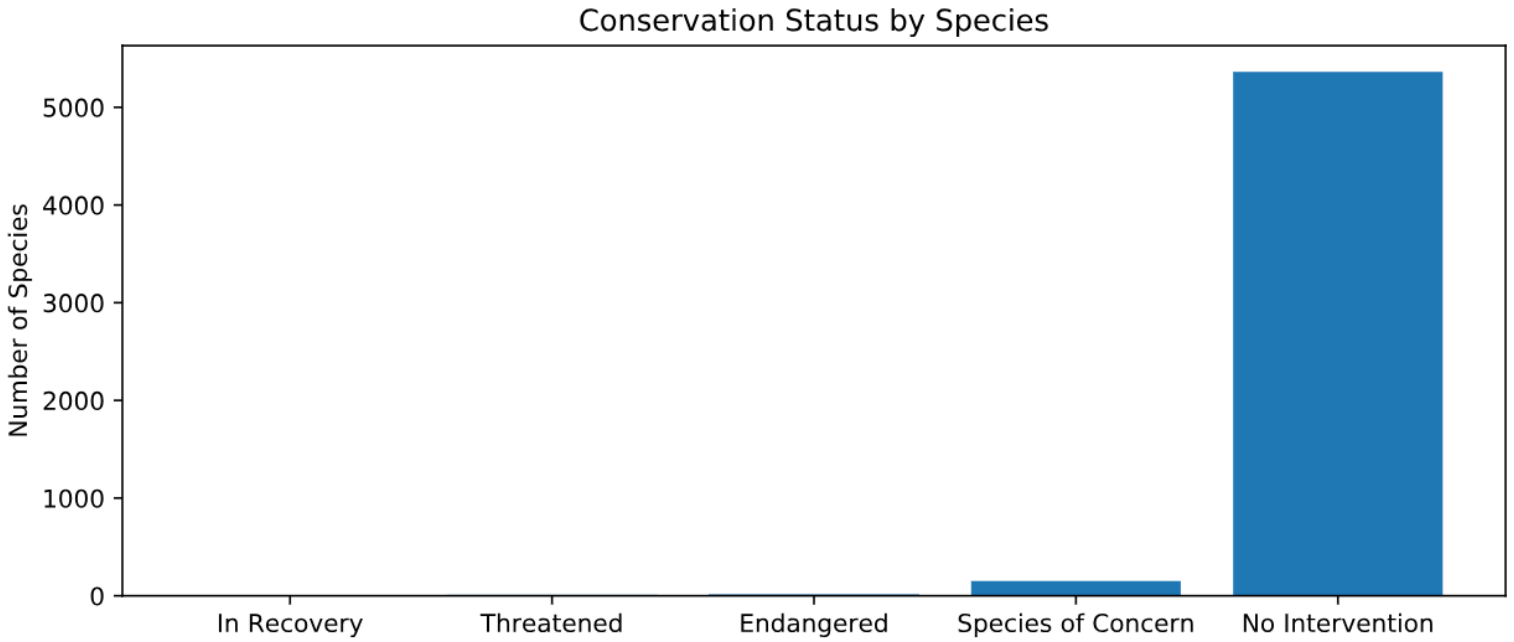
# Data in species_info.csv

- There are 5,541 unique species based on scientific name.
- The data in the file contains the following fields:
  - Category of species
  - Species' scientific name
  - Species' common name
  - Species' conservation status
- There are seven types of categories of species:
  - Amphibian
  - Bird
  - Fish
  - Mammal
  - Nonvascular Plant
  - Reptile
  - Vascular Plant
- There are five kinds of conservation status::
  - Endangered
  - In Recovery
  - No Intervention
  - Species of Concern
  - Threatened

# Significant calculations in species_info.csv

- When I grouped the scientific name of each species by conservation measure, the total number across all categories is 180, which is far below the number of unique species in this database (5,541).
- This means that most of the species in the database are not categorized as needing some form of protection.
- When I replaced the null value with "No Intervention" and re-ran my analysis, the total number of scientific species rose from 5,541 to 5,543.

# Significant calculations in species_info.csv




- The bar graph above demonstrates that the species listed in the species_info file are overwhelmingly not protected species.

# Significant calculations in species_info.csv

## Investigating Protected Species

The chart below organizes categories of species by whether that species is protected. It also includes the percentage of the species that are protected.

| Category | Not Protected | Protected | Percentage Protected |
|---|---|---|---|
| Amphibian | 72 | 7 | 8.86% |
| Bird | 413 | 75 | 15.37% |
| Fish | 115 | 11 | 8.73% |
| Mammal | 146 | 30 | 17.05% |
| Nonvascular Plant | 328 | 5 | 1.50% |
| Reptile | 73 | 5 | 6.41% |
| Vascular Plant | 4216 | 46 | 1.10% |

Mammals and birds seem the most likely to be protected.

# Significant calculations in species_info.csv

**Determining Statistical Significance**

- Although the chart in the previous slide indicated that mammals and birds were the most likely to be prot4ected, we need to run a Chi Squared Test to determine whether the difference between the two is significant.
- The p-value of the Chi Squared Test between Mammals and Birds was 0.687549, which is not statistically significant.
- However the difference between Mammals and Reptiles is statistically significant, as the p-value for that Chi Squared Test is 0.386556

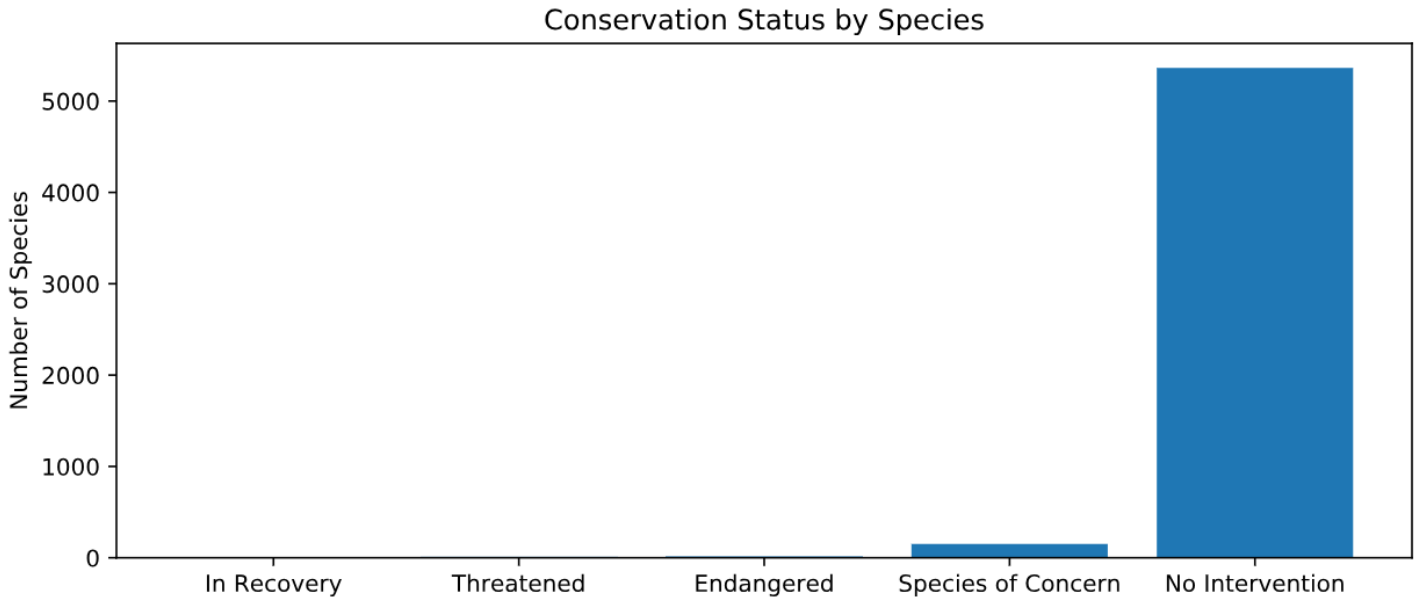
# Recommendations for conservationists based on species_info.csv data

- Because there was a significant difference between the percentage protected between mammals and reptiles, conservationists may want to study environmental changes to determine whether these changes could have an impact on species development.
- Conservationists may also want to examine the criteria for protecting a species to determine whether this favors some species over others.

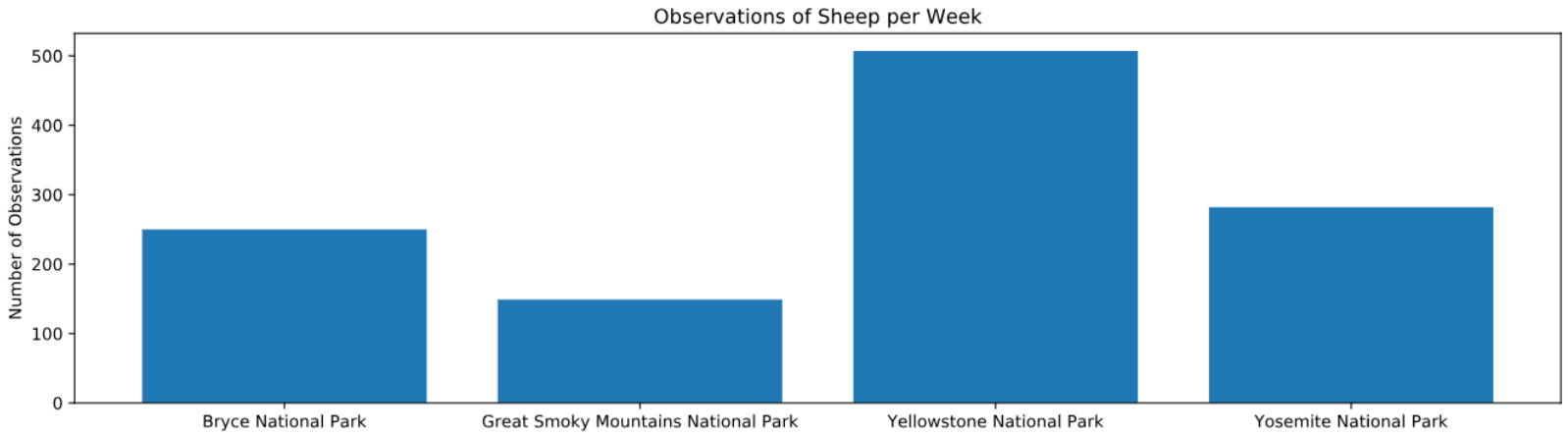
# Sample size determination for foot and mouth disease

- We can determine the **baseline percentage** because we have last year's percentage of foot and mouth disease at Bryce National Park—15%.
- We can determine the **minimum detectable effect** by dividing the margin of error (5%) by the baseline percentage (15) and multiplying that value by 100. The resulting number is 33.33%.
- Using the baseline percentage and minimum detectable effect, I used the Sample Size Calculator to determine the correct sample size for Bryant National Park (870).
- To determine the sample size for Yellowstone National Park, I would need the baseline percentage for that park.

# Graphs


Conservation Status by Species
Number of Species
5000
4000
3000
2000
1000
0
In Recovery
Threatened
Endangered
Species of Concern
No Intervention

# Graphs


Observations of Sheep per Week
Number of Observations
0
100
200
300
400
500
Bryce National Park
Great Smoky Mountains National Park
Yellowstone National Park
Yosemite National Park