# Biodiversity Project

Nathaniel Santos

Biodiversity “Analyst”

Capstone – Introduction to Data Analysis

# Objectives

- Perform data analysis on the conservation statuses of National Parks plant and animal species
- Investigate any patterns or themes that might indicate the types of species that might become endangered
- Discuss case study in sample size determination to help measure impact of initiatives in the future

# Notes on Data

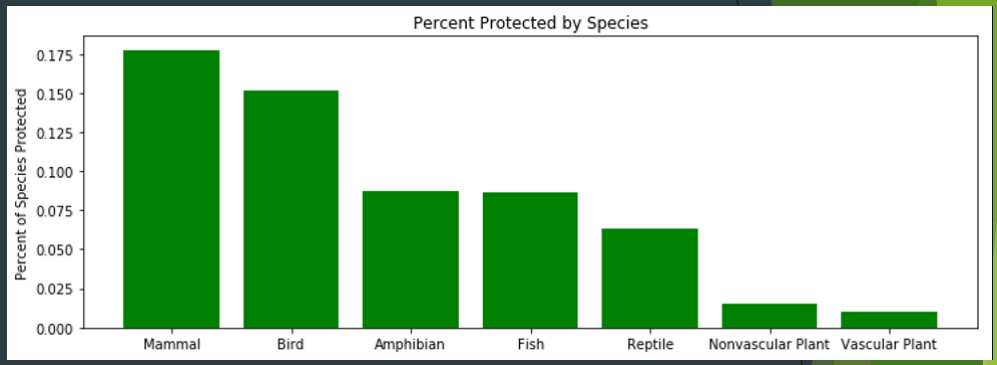
- The dataset (species_info.csv) contains a list of 5,000+ species of animals and plants found in America's National Parks, including the following details on each species:
  - Scientific and common names
  - Biological classification
  - Conservation status
- The conservation status field was populated with data for species that were under varying degrees of protection, but was left blank for species that did not require intervention
  - To make it relevant for the upcoming analysis, empty fields were filled with the string "No Intervention"; based off this information, a new field was added listing the protection status of each species

# Are certain types of species more likely to be endangered?

- Choice of significance test: Chi-squared test
  - Data: categorical
  - Pieces of data compared: 2+
- Categories compared:
  - Protected vs non-protected species per biological class
  - Pval < 0.05 indicates statistical significance, rejecting the null hypothesis
  - Goal: To determine whether certain biological classes are more likely to be endangered

# Are certain types of species more likely to be endangered?

- Result: YES, species from certain classes are more likely to be endangered than others[1,2]
- Data suggests that mammals and bird species are more at risk of being endangered relative to species of other classes




```
pval (all classes): 5.51082804731e-89
pval (animals): 0.0312971619317

pval (plants): 0.662341949138
pval (amphibians, fish, reptiles): 0.808062079153
pval (birds, mammals): 0.687594809666
```

[1]Results were significant when looking at differences between all classes and animals (pval < 0.05)

[2]Groupings determined by identifying results where the null hypothesis was accepted (pval > 0.05), i.e. classes with rates of species protection that were not significantly different from one another

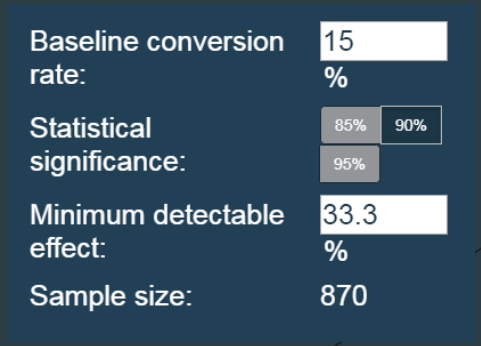
# Recommendation for conservationists

- Based on the data, conservation efforts and resources are most needed in protecting bird and mammal species
- Tracking the percentage of protected species of birds and mammals over time will show whether renewed efforts are bearing fruit

# Determining sample size: A case study

- Challenge: Identify appropriate sample size of observations to accurately determine impact of program to reduce rate of foot and mouth disease among sheep species at Yellowstone National Park

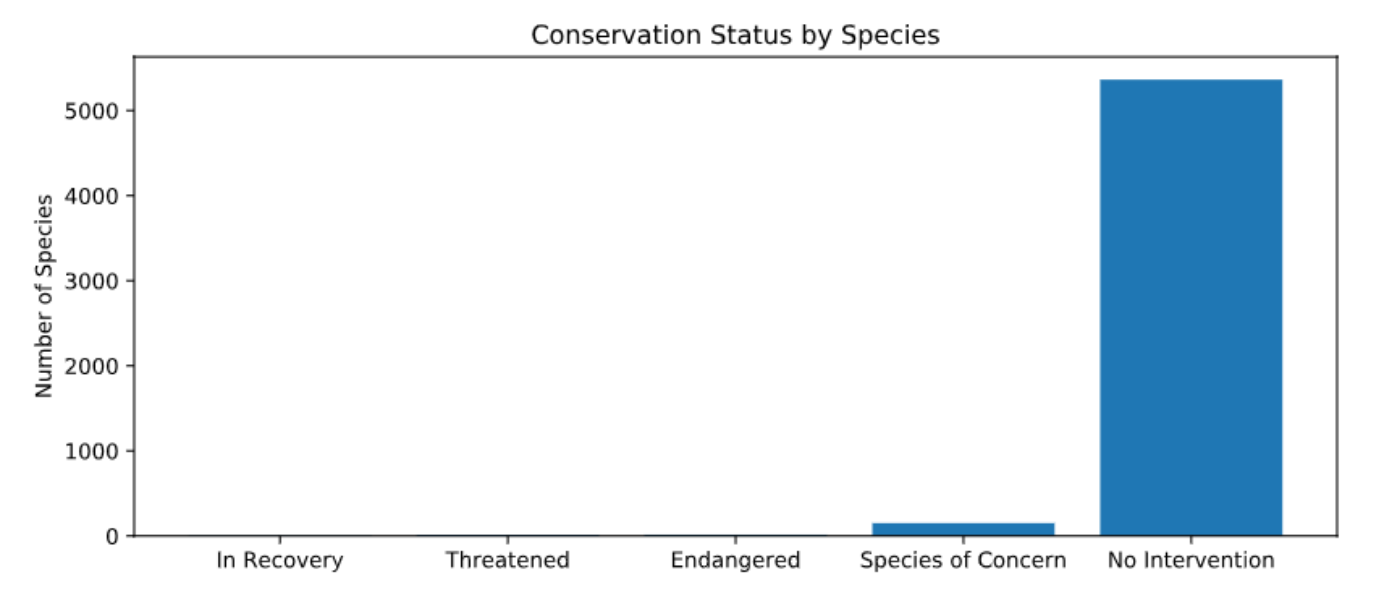
- Parameters: Utilizing historical data from Bryce National Park (15% baseline infection rate), detect at least a 5% change (33.3% change from baseline) in Yellowstone's sheep infection rate with 90% confidence



# Determining sample size: A case study

- Conclusions:
  - 870 observations needed
  - Based on historical data at Yellowstone, an observation period of ~1.7 weeks will be needed to reach the appropriate amount of sheep observations

# Appendix (1 of 2): Graphs created in Codecademy environment



# Appendix (2 of 2): Graphs created in Codecademy environment