CODECADEMY – INTRODUCTION TO DATA ANALYSIS BIODIVERSITY CAPSTONE PROJECT

BY: CHELSEY POPEJOY

AUGUST 14TH COHORT

WHAT DATA DOES OUR SPECIES_INFO FILE CONTAIN?

- This .csv file, based on data from the National Parks Service, provides content for the analyst to explore and find patterns or themes about a multitude of animal species
  - Rows within this data file include:
    - The set of row indices, category classifications (mammals, reptiles, amphibians, bird, fish, vascular plants, and nonvascular plants), scientific names, common names, and conservation status, where one applies (species of concern, threatened, endangered, or in recovery)
    - All this for 5541 different species!
- While working with this file you may notice that the vast majority of animals do not have a 'conservation_status' distinction and are listed instead as 'nan', meaning most species are not actively the focus of conservation efforts – that's good news!

SIGNIFICANCE OF ENDANGERED STATUS ACROSS SPECIES CATEGORIES

	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

- At first glance, the proportion of protected (True) or unprotected (False) species appear higher for certain categories like birds, mammals, and reptiles; but are those differences across categories statistically significant? Before changing tactics for conservation or rerouting efforts and funds to protect species groups, the National Park Service needs to determine whether these differences are due to a real variance in the species or whether the difference we're seeing could be a result of random error or sampling bias.
- We can test this hypothesis by using a Chi Square Test, appropriate for data sets with two or more categorical variables, like true/false
- Results of the Chi Square test reveal that the difference between mammals (17% protected) and birds (15% protected) are not significant while the difference between mammals and reptiles (6% protected) is!

Recommendation:

- Based on the findings, conservationists should focus their efforts first on mammals as their propensity to need protection is higher than all other categories and is significantly higher than some other categories. Ways to mitigate factors leading to species decline need to be identified and enforced for this category in particular
  - A secondary target would be the bird species category

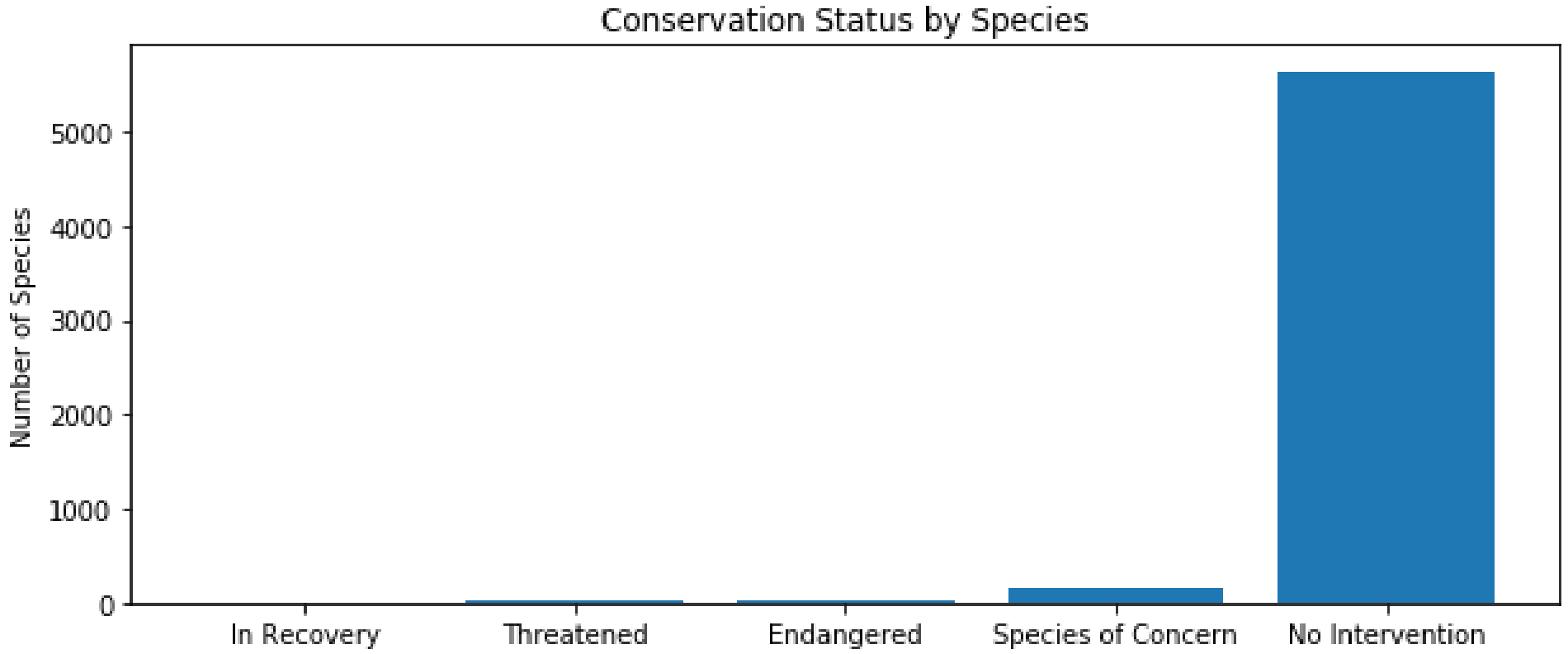
HOW MANY SHEEP OBSERVATIONS ARE NEEDED TO DETECT DISEASE REDUCTION?

- Park Rangers across Bryce and Yellowstone National Parks have concentrated their efforts to reduce Foot and Mouth Disease among the local sheep populations
- To determine the sample size needed to test the effectiveness of their efforts the analyst first needs to define a few data points
  - **Baseline percentage:** current level of the focus of the test and what we test against for significance in the delta
    - BP = 15% because that proportion of sheep were recorded as having Foot and Mouth Disease in Bryce National Park last year
  - **Minimal detectable effect:** the relative minimum improvement over the baseline that you're willing to detect in an experiment
    - MDE = 5% as stated, the Rangers would like to detect changes of 5 percentage points in the proportion of sick sheep
  - **Statistical significance/confidence interval:** the percentage at which we are 90% confident that any significant changes are due to actual differences (or changes) in the sample and not due to error or sampling bias
    - CI = 90% as stated, the Rangers would like to be 90% confident in their conclusions. This is also often the standard level for CI in research.
- **Findings:** Based on these figures, the Rangers would need a random sampling of 870 sheep per park, in order to determine if any recorded changes in the baseline percentage of disease are significant
  - Keeping in mind the average number of sheep sightings per week in each park, to reach this sample the Rangers would need to monitor Bryce and Yellowstone National Parks for about 3.5 and 1.5 weeks, respectively

APPENDIX

CHARTS CREATED THROUGHOUT THE BIODIVERSITY CAPSTONE PROJECT

CONSERVATION STATUS BY SPECIES



OBSERVATIONS OF SHEEP PER WEEK