April 2018

Andrew Guindi
Introduction to Data Analysis
Code Academy

Biodiversity in our National Parks

Capstone – Option 2

Agenda

- Species
  - Data Description
  - Endangered Species
  - Chi Squared Testing
  - Recommendations for Conservation
- National Parks Observations
  - Data Description
  - Data Analysis: Sheep
  - Study: Foot & Mouth Disease
  - A/B Testing for Disease Rate Reduction

Data Description: Species

- The species_info.csv includes information on 7 categories of species which are **Mammal, Bird, Reptile, Amphibian, Fish, Vascular Plant, Nonvascular Plant**.
- Also this file delivers further information about those categories, such as scientific name, which is the rather accurate aspect to learn how many unique species we've got there, which is **5541**.
- Another characteristic is **common names** which are of course less than the scientific names **5504/5541** as we tend to ignore some differences between similar species in the common language.
- And lastly, to know the conservation status of the species, it is measured within the data file in 5 pointers: **No Intervention**, **Species of Concern**, **Endangered**, **Threatened**, and **In Recovery**.

Species: Endangered Species

Here is a representation of the status of all the species we have:

- We've counted the numbers according to the conservation statuses and sorted accordingly:

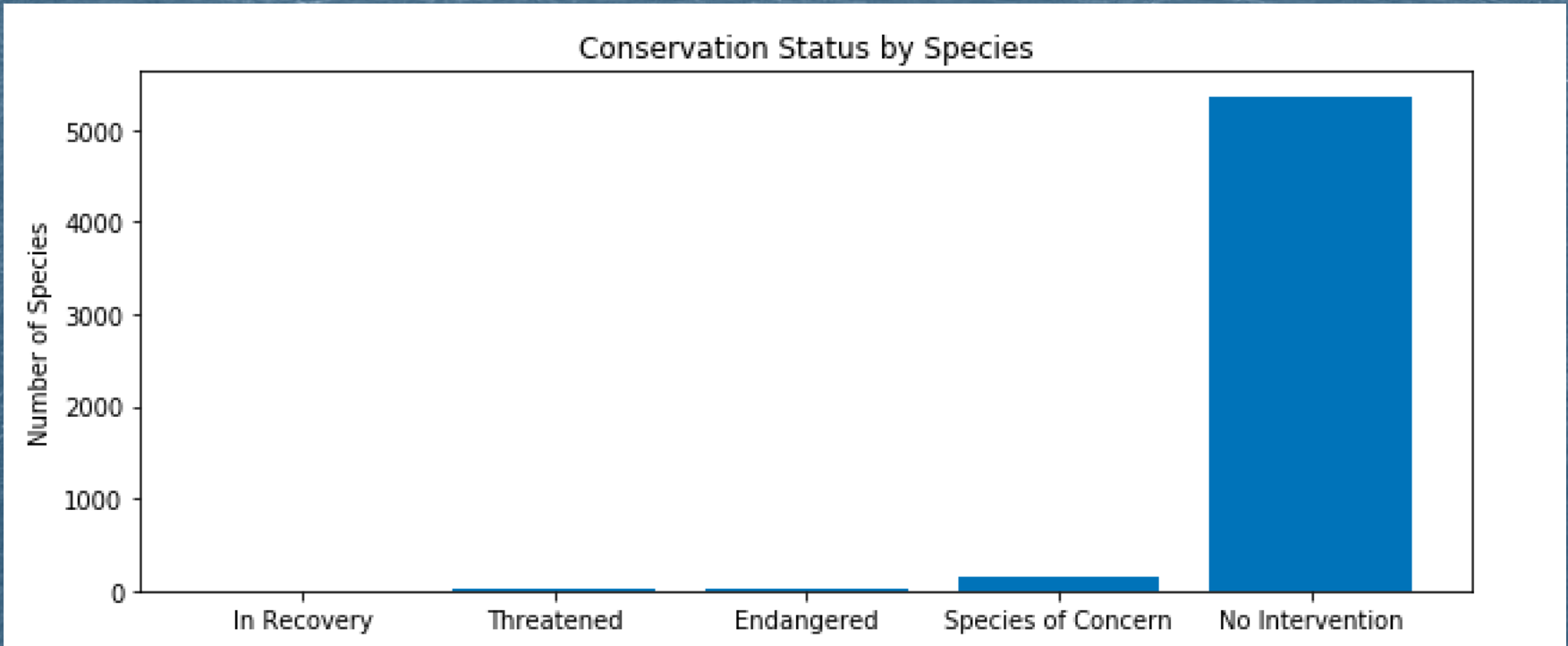


Fig 1. Conservation Status Chart

Species: Endangered Species

Taking a closer look

- According to the data represented previously here's what we have:

Status	Count
In Recovery	4
Threatened	10
Endangered	15
Species of Concern	151
No Intervention	5363

Fig 2. Conservation Status Table

- Most of the species as we can see are protected (**5363 – No Intervention**) and also **4** are in **recovery** phase, which means that soon enough they'll be at the 'No Intervention' status.
- However, **151** seem concerning and better to fix the situation at an earlier stage of concern. Those fall under '**Species of Concern'** status.
- Also we have **25** species that seem to be in real threat (**Threatened, and Endangered**) and it seems that immediate action is to be taken.

Species: Endangered Species

Taking a closer look on the protection percentage per category of species:

Category	Percent Protected
Mammal	17.045455
Bird	15.368852
Amphibian	8.860759
Fish	8.730159
Reptile	6.410256
Nonvascular Plant	1.501502
Vascular Plant	1.079305

Fig 3. Protection % Table

And now let's find out if the differences between those numbers are significant!

Chi-Squared Significance Testing: Protection % for different Categories

- Test Results:
- 1. Mammal, Bird:
  - Pval= 0.68 > 0.05
  - H_0 is accepted
  - No significant difference!
- 2. Reptile, Mammal:
  - Pval= 0.038 < 0.05
  - H_0 is rejected
  - There *is* a significant difference!

- Conditions and Procedure:
  - Null hypothesis (H_0) states that there is no significant difference
  - To be able to reject H_0 , we need the test's pval < 0.05
  - Contingency table is created and the test is run using scipy.stats.chi2_contingency()

Recommendations for Conservation:

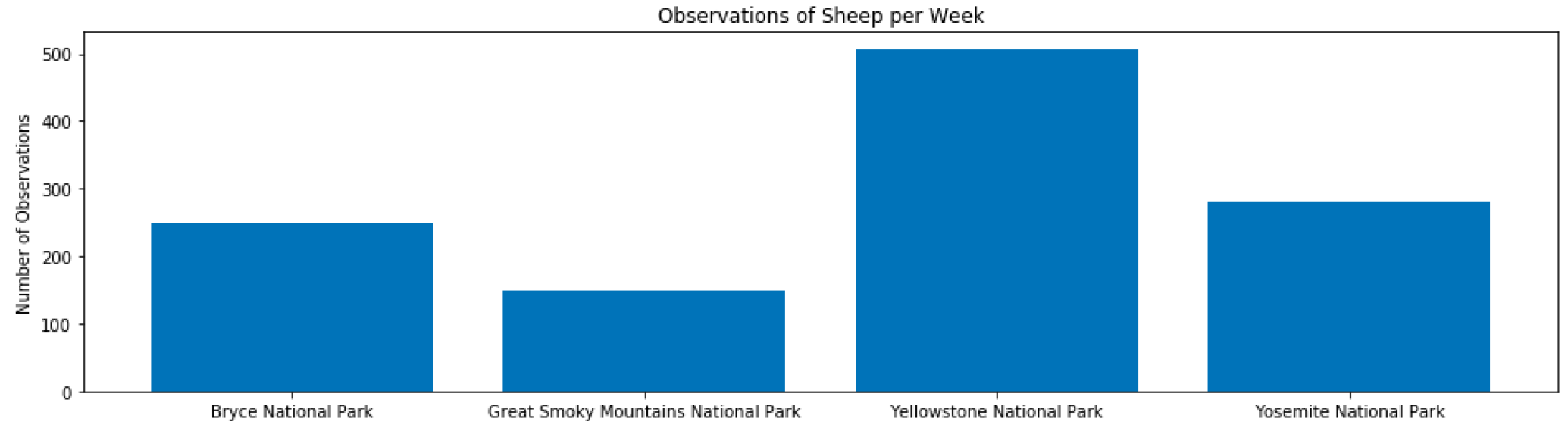
- Species that are more likely to become endangered would be the top ones on the % protected table (fig.3). Yet, there is no significant difference in their protection percentages according to the chi square test performed.
- From the same table, we can also deduce that the least to become endangered are the bottom of the table, which are Vascular and Non-Vascular plants.
- Also comes the Reptile family as a less-likely to become endangered, since we've seen before that there is a significant different between its % of protection and Mammals' which also indicates that the threat status should be quite different between them as well.

Data Description: Sheep Observations

- The observations.csv file includes information on sightings of different species within multiple national parks during the period of a week.
- To filter the information to sheep (which is a mammal), we've created a column called is_sheep which contains a Boolean value to indicate whether the species is a sheep or not.
- We've come to find that there are 3 different species of sheep observed:
  - Domestic Sheep - AKA Mouflon, Red Sheep, Sheep (Feral)
  - Bighorn Sheep
  - Sierra Nevada Sheep

Data Analysis: Sheep Observations

- The observations of sheep across different parks were found to be as follows: (Data is sorted in alphabetical order of parks)



Park Name	Observations
Bryce National Park	250
Great Smoky Mountains National Park	149
Yellowstone National Park	507
Yosemite National Park	282

Fig 4.Sheep Observations Across Parks Chart and Table

Case Study: Foot & Mouth Disease

- Park Rangers at Yellowstone National Park have been running a program to reduce the rate of foot and mouth disease at that park. The scientists want to test whether or not this program is working.
- They want to be able to detect reductions of at least 5 percentage points. For instance, if 10% of sheep in Yellowstone have foot and mouth disease, they'd like to be able to know this, with confidence.
- According to the data received, we know that 15% of the sheep observed in Bryce National Park, are sick with the Foot & Mouth Disease (We shall use that as a our baseline conversion rate).

A/B Testing
Reduction of Foot & Mouth Disease

- Time needed to observe the sample:
  - Bryce Park:
    - Observations/week=250
    - Time to observe the sample: 510/250= 2.04 ~ 2 weeks
  - Yellowstone Park:
    - Observations/week=507
    - Time to observe the sample: 510/507 = 1.005 ~1 week

- Sample Size:
  - Baseline conversion rate: 15%
  - Statistical Significance: 90%
  - Min. Detectable Effect = $\frac{5}{15} \times 100 = 33.3\%$
  - Sample size (using calculator) = 510

Thank you.

Andrew Guindi