Biodiversity Case Study

Konrad Hughes

>>> The Task

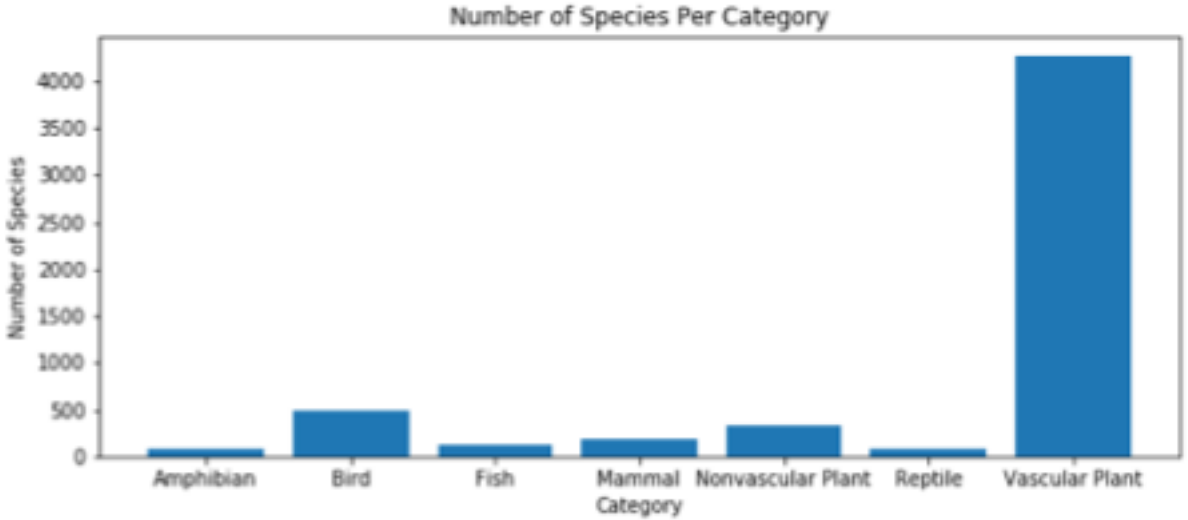
- Perform data analysis for the National Park Service using the data provided in species.csv and observations.csv
- Investigate the conservation status of endangered species from several different parks
- Answer the question: **Are certain types of species more likely to be endangered?**
- Help advise scientists **how big a sample needs to be** to gather data to test whether a foot and mouth reduction program in sheep is working

>>> The Data & Initial Analysis

- To perform the analysis for the National Park Service the data in **species_info.csv** was provided.
- This provided a list of of different species categorized into:
  - The **scientific name** of each species
  - The **common name** of each species
  - The **species** conservation status

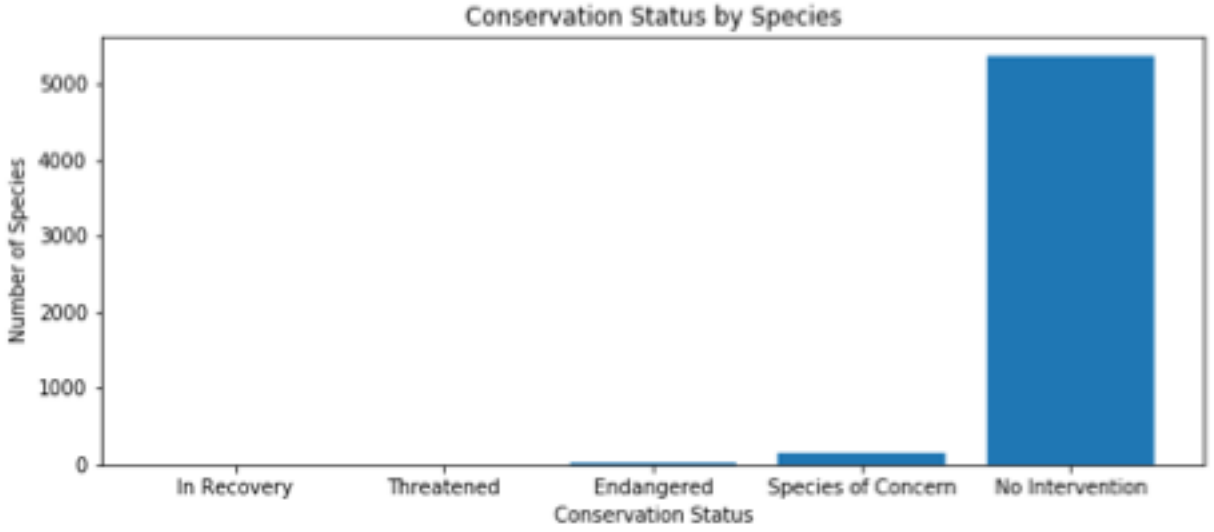
- In the data set there are 5541 different unique species which are categorized into the different category below and then plotted on a graph.
- In terms of numbers biodiversity is highest in the vascular plant category, followed by a wide range of birds in the different parks, then by nonvascular plants.

	category	scientific_name
0	Amphibian	79
1	Bird	488
2	Fish	125
3	Mammal	176
4	Nonvascular Plant	333
5	Reptile	78
6	Vascular Plant	4262



>>> The Data & Initial Analysis

- Each species is then then classified into different conservation statuses which are:
  - **Species of concern** - declining population or appears to be in need of conservation– 151 species
  - **Endangered** – seriously at risk of extinction – 15 species
  - **Threatened** – vulnerable to endangerment in the near future – 10 species
  - **In recovery** – formerly Endangered, but currently not in danger of extinction throughout all or a significant portion of its inhabitable range – 4 species
- However this only showed around 200 species, so the remaining ones were relabeled to ‘No Intervention:
  - **No Intervention** – 5363 species
- This data is shown visually in the graph shown below
- An initial observation suggests that overall the biodiversity of the parks is very healthy with many species needing no intervention, but it is crucial to help those that do.



>>> The Calculations

- Having had an initial look at the data and added it a dataframe to summarize it, it raises an interesting question: **Are certain types of species more likely to be endangered?**
- Several steps are required before we can carry out a significance test to see if we can reject the Null Hypothesis that different species are not more likely to be endangered,
- First is to add a new column to the data in Species to state whether a species needs any kind of intervention or not.
- This data is then pivoted by category to allow an easy comparison and an additional column added to the pivot to show the percent protected as below.
- An initial view of the pivot suggests that some species are far more likely to be in the protected category and so endangered, but until we do a significance test we cannot answer the question definitively

	category	not_protected	protected	percent_protected
0	Amphibian	73	7	87.5%
1	Bird	442	79	15.2%
2	Fish	116	11	86.6%
3	Mammal	176	38	17.8%
4	Nonvascular Plant	328	5	15.0%
5	Reptile	74	5	63.2%
6	Vascular Plant	4424	46	10.3%

>>> The Calculations

- Following from this, given that the data is categorical and we are comparing the protected vs not protected a chi square test is the best form of significance test to perform here
- The tests performed compare Mammals vs Birds and then Reptiles vs Mammals to see if there is a significant difference between their likelihood of being endangered.
  - Chi Squared between Mammals and Birds, **pval = 0.446**
    - No significant difference so cannot reject Null Hypothesis
  - Chi Squared between Mammals and Reptiles, **pval = 0.0235**
    - There is a significant difference between Mammals and Reptiles

- Whilst the first test may fail to be significant, with the result of the second test **we can reject our Null Hypothesis** and state that there are indeed significant differences between different species.
- **As a result we can conclude that certain types of species are more likely to be endangered than others.**

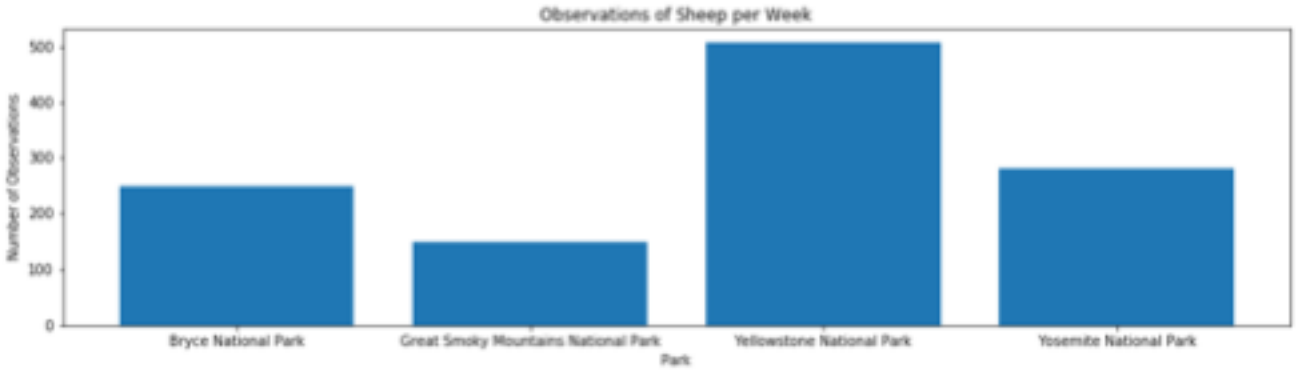
>>> Recommendations

- As a result of the analysis in the previous slides several recommendations can be made:
  - As certain types of species are more likely to be endangered than others it is key to help those that are more likely to be endangered.
- Looking back at the pivot the species with the most protected species (and so needing most help) are amphibians, fish followed by reptiles
- A start would be to **create conversation programs** across different National Parks focused on protecting these species and what steps can be taken to move them to No Intervention
- Having initially focused on these species, **this analysis should be repeated at regular intervals** to ensure progress is happening and that are species are not becoming endangered as well
- By taking steps to improve species more likely to be endangered and tracking progress it should **ensure that the biodiversity in the parks continues to thrive**.

>>> Sample Size Determination

- For the second part of the case study the task is to investigate the different sightings of sheep in various national parks and investigate if a program to reduce foot and mouth is working at the different parks.
- In order to do this data was provided in observations.csv which contains observations of different species at the different parks.
    - The data here used scientific names so it had to be merged with species.csv so that the observations of sheep could be analyzed, leading to the below output, which was then put into a graph:

	park_name	observations
0	Bryce National Park	250
1	Great Smoky Mountains National Park	149
2	Yellowstone National Park	507
3	Yosemite National Park	282



>>> Sample Size Determination

- With the data now gathered categorized it is possible to do the calculations about the foot and mouth reduction program.
- The parks that will be initially observed are Yellowstone and Bryce
- We know 507 sheep are observed per week in Yellowstone and 250 per week in Bryce
- The key aspect is how big the sample size should be to test significance for the reduction program.
- Given the information provided:
  - **Baseline of 15%** of sheep have foot and mouth disease
  - We are looking to see if there has been a **5% reduction** so **minimum detectable effect is 33%**
  - **Statistical significance of 90%**
  - Using the sample size calculator at at Optimizely, putting in the above information suggests a **sample size of 510 sheep to test significance**
- As a result of this to get the correct sample sizes:
  - Sheep in Yellowstone need to be observed for 1.01 weeks
  - Sheep in Bryce need to be observed for 2.04 weeks
- With this the scientists can observe the sheep and once this data is gathered the significance tests can be carried out to see if the program is working as planned.