Codecademy Capstone Project: Biodiversity for the National Parks

João Patacas

Overview

- Description of data about different species
- Significance calculations for endangered species
- Recommendations regarding endangered species
- Foot and mouth disease study

Description of data about different species

- The *species_info.csv* file provides various data about different species, including the category, scientific name, common names, and conservation status.

category	scientific_name	common_names	conservation_status
Mammal	Clethrionomys gapperi gapperi	Gapper's Red-Backed Vole	No Intervention
Mammal	Bos bison	American Bison, Bison	No Intervention
Mammal	Bos taurus	Aurochs, Aurochs, Domestic Cattle (Feral), Dom...	No Intervention
Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention
Mammal	Cervus elaphus	Wapiti Or Elk	No Intervention

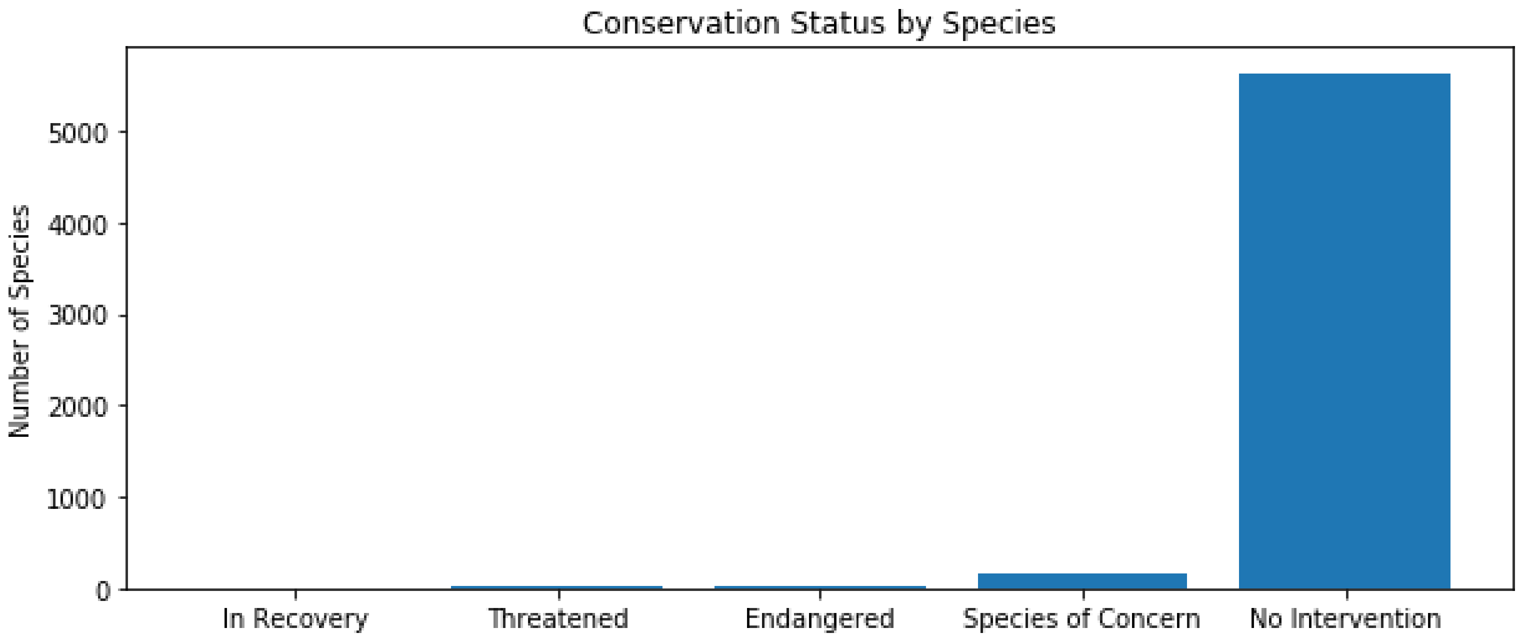
Description of data about different species

- Closer inspection of the species data reveals:
  - 5541 unique species
  - 7 unique values for different categories: *'Mammal'*, *'Bird'*, *'Reptile'*, *'Amphibian'*, *'Fish'*, *'Vascular Plant'*, and *'Nonvascular Plant'*
  - 4 different values for Conservation Status: *'Species of Concern'*, *'Endangered'*, *'Threatened'*, and *'In Recovery‘*
  - An additional value for Conservation Status was defined for species that are not in risk: *'No Intervention'*

Description of data about different species

- We can group species by *conservation_status*, and analyze the species data to find out how many species meet each of the *Conservation Status* criteria:

Conservation status	**Number of species**
Endangered	15
In Recovery	4
No Intervention	5363
Species of Concern	151
Threatened	10



Description of data about different species

- The initial data analysis revealed that the majority of species do not require intervention (5356). There are 151 species of concern, 15 endangered species, 10 threatened species, and 5 species in recovery.

Significance calculations for endangered species

- Using the data provided in the *species_info.csv* dataset, we can determine if certain types of species are more likely to be endangered.
- In order to achieve this, first we need to reshape our data. We can do this by:
    - Defining a *'is_protected'* column, which is *True* for species that need intervention (i.e. *'Species of Concern', 'Endangered', 'Threatened', and 'In Recovery'*), and *False* for species that don't need intervention (i.e. *'No Intervention'*)
    - Grouping the species data by *'category'* and *'is_protected'*, and rearranging the data using a pivot table.

Significance calculations for endangered species

- Finally we can obtain the following table, which shows how many species are protected or not protected for each category
- We can use this table as a basis for performing significance tests

category	not_protected	protected	percent_protected
Amphibian	72	7	8.860759
Bird	413	75	15.368852
Fish	115	11	8.730159
Mammal	146	30	17.045455
Nonvascular Plant	328	5	1.501502
Reptile	73	5	6.410256
Vascular Plant	4216	46	1.079305

Significance calculations for endangered species

- We can perform a significance test to determine if certain types of species are more likely to be endangered than others.
- For example, we can determine if species in the category *‘Mammal’* are more likely to be endangered that species in the category *‘Bird’*. Since the data is split in two categories (i.e. *protected* and *not protected*), we can use the Chi square test to determine if the statement is true.

Significance calculations for endangered species

- We can obtain the following Chi square contingency table from the previous table and use it to perform the Chi square test:

	Protected	**Not Protected**
Mammal	30	146
Bird	75	413

- The result of the Chi square test is **0.688 ≥ 0.05** which indicates that the difference between *Mammal* and *Bird* is not significant, i.e. *Mammals* are not more likely to be endangered than *Birds*.

Significance calculations for endangered species

- We can perform a similar test to determine if the difference between *Reptile* and *Mammal* is significant
- We can obtain the following Chi square contingency table and use it to perform the Chi square test:

	Protected	**Not Protected**
Reptile	5	73
Mammal	30	146

- The result of the Chi square test is **0.038 < 0.05** which indicates that the difference between *Reptile* and *Mammal* is significant, i.e. *Reptiles* are more likely to be endangered than *Mammals*

Recommendations regarding endangered species

- Results from the significance calculations performed in this study show that Reptiles are more likely to be endangered than Mammals. Conservationists should take this into account in their efforts to preserve endangered species.

Foot and mouth disease study

- In order to know with confidence how many sheep have foot and mouth disease in each National Park, we can use the dataset from *observations.csv* to determine the number of sheep to be observed and for how long they should be observed.
- Since the *observations.csv* file provides observation data for all species, first we need to filter *Sheep* observations. This can be achieved using *apply* and a *lambda function* to create a new column in *species* called *is_sheep* which is *True* if the *common_names* contains *'Sheep'*, and *False* otherwise.

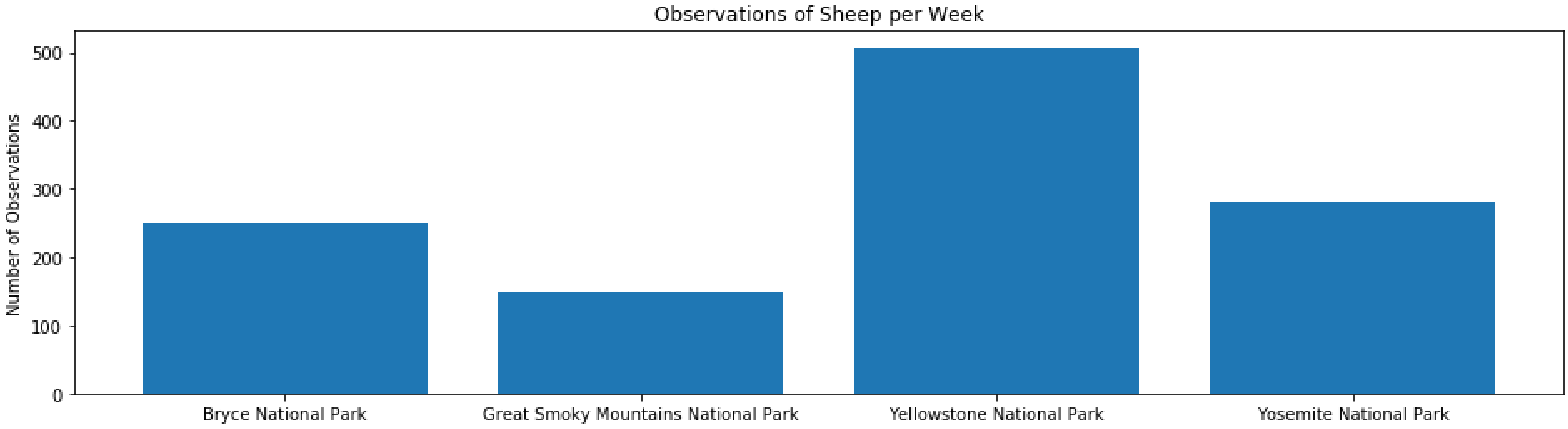
Foot and mouth disease study

- Since we are looking at *Sheep* which are *Mammals*, we can define a *sheep_species* table by selecting the rows of species where *is_sheep* is *True* and category is *Mammal*:

category	scientific_name	common_names	conservation_status	is_protected	is_sheep
Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True
Mammal	Ovis canadensis	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True
Mammal	Ovis canadensis sierrae	Sierra Nevada Bighorn Sheep	Endangered	True	True

Foot and mouth disease study

- Since we are interested in observations of sheep by National Park, we merge the *sheep_species* table with the *observations* dataset and group observations by National Park:



Foot and mouth disease study

- Considering a baseline of **15%** of sheep that have foot and mouth disease;
- We can calculate the *minimum detectable effect* by:

$$mde = \frac{0.05}{baseline} \times 100 = \mathbf{33.3\%}$$

- For a **90%** *statistical significance;*

- Using the online sample calculator we obtain a *sample size* of: **870**

Foot and mouth disease study

- In order to determine how many weeks we need to observe sheep at each park in order to observe enough sheep we can calculate:

- $weeks = \frac{sample\ size}{observations\ in\ National\ Park}$

- For Bryce National Park: $weeks = \frac{870}{250}$ ≅ **3.5 weeks**
- For Yellowstone National Park: $weeks = \frac{507}{250}$ ≅ **1.7 weeks**

Thank You

João Patacas