In Search of Endangered Sheep

Biodiversity Capstone Project
Piccinini Alessandro
10/04/2018

Objective

- The reason and the aim of this data analysis are :

    1. To perform some **investigation on the conservation statuses** of some species in the National Parks and to investigate if there are species that are likely to be endangered

    2. To help a team of ruminant-enthused scientists to **locate the sheep** in the national parks

    3. To help the Yellowstone National Park Rangers to see if the program to **reduce the rate of foot and mouth disease** is working. And help them to find a way to detect disease reductions of at least 5 percentage points.

- To complete the points above we are going to propose some bar chart and some data's observations. The two data files used to reach this propose have been:

    1. species_info.csv

    2. observations.csv

1- Investigation on the conservation statuses

Observations from data

1. species_info.csv

	category	scientific_name	common_names	conservation_status
0	Mammal	Clethrionomys gapperi gapperi	Gapper's Red-Backed Vole	nan
1	Mammal	Bos bison	American Bison, Bison	nan
2	Mammal	Bos taurus	Aurochs, Aurochs, Domestic Cattle (Feral), Domesticated Cattle	nan
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	nan
4	Mammal	Cervus elaphus	Wapiti Or Elk	nan

The information contained in the *species_info.csv* file are a list with for each animal :

- The categories
- The scientific name
- The common name
- The conservation status

1- Investigation on the conservation statuses

Results from Analysis

From the Analysis of the the *species_info.csv* file I have found that:

- The species are 5541 If use *scientific_name* and 5504 if I use *common_names*, probably there are some *common_names* species with the same *scientific_name*.

- The categories are: 'Mammal', 'Bird', 'Reptile', 'Amphibian', 'Fish', 'Vascular Plant'.

- the conservation statuses are: 'No Intervention', 'Species of Concern', 'Endangered', 'Threatened', 'In Recovery'.
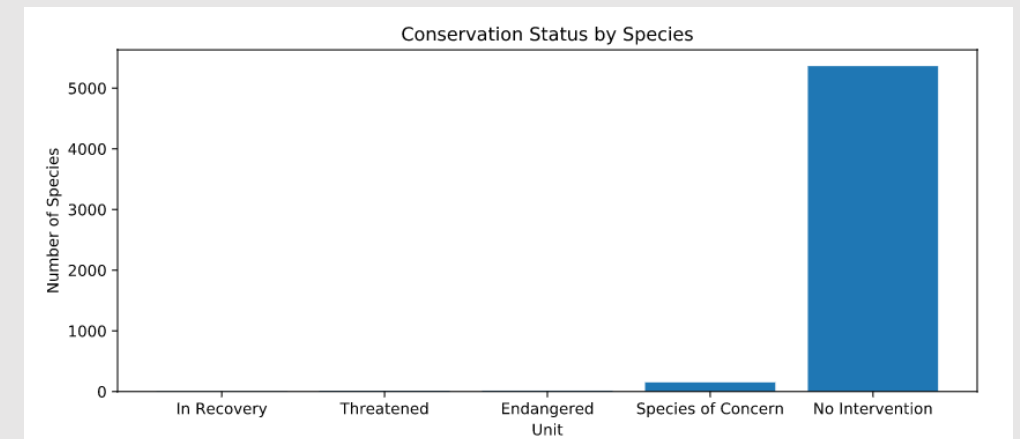
- Thanks to the table above we can say that the main part of the species 5363, are in the "No intervention" category but in any case 15 species are in the endangered category.

1- Investigation on the conservation statuses

Results from Analysis

- Starting from the table proposed in the previous slide I have figure out if there are certain types of species (Category) more likely to be endangered

category	not_protected	protected	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

Thanks to this table we can see that the more likely to be endangered is the mammals with the 17.04% instead the more protected are the Nonvascular Plant with the 1.07. But to ensure this affirmation we have to do a significance test to see if this difference is due to chance.

Based on the data, we have chosen the *chi-squared test*.

1- Investigation on the conservation statuses

Results from Analysis

- I have done the Chi Square Test comparing the Mammal and the Bird and the Reptile and mammal, the result is showed in the image below:

```
----- MAMMAL BIRD TABLE------
  category   protected   not_protected
1     Bird          75             413
3   Mammal          30             146
There isn't significant difference because the pvalue is 0.688

----- REPTILE MAMMAL TABLE------
  category   protected   not_protected
3   Mammal          30             146
5  Reptile           5              73
There is significant difference because the pvalue is 0.038
```

In the first chi square test we can see that there isn't significant difference because pval > 0.05, so our result are due to random chance and we can't say that the bird are less likely to be endangered than the mammals . Instead in the second test relative on Reptile and mammals we can see that there is a significant difference because the pval_reptile_mammal < 0.05 so our result are not due to random chance

Due this test I can conclude that certain types of species are more likely to be endangered than others but we can't assume this affermation with certainty because sometimes the pvalue is more then 0.05.

2- Locate the sheep
Observations from data

2. **observations.csv**

Conservationists have been recording sightings of different species at several national parks for the past 7 days. This observation have been recorded the csv file showed below:

	scientific_name	park_name	observations
0	Vicia benghalensis	Great Smoky Mountains National Park	68
1	Neovison vison	Great Smoky Mountains National Park	77
2	Prunus subcordata	Yosemite National Park	138
3	Abutilon theophrasti	Bryce National Park	84
4	Githopsis specularioides	Great Smoky Mountains National Park	85

. The information contained in the *osservation.csv* file are a list with for each animal :

- The scientific name
- The park name
- The observations

Thanks to this .csv file and the previous one (species_info.csv) we will be able to locate the sheep to help a ruminant-enthused scientists team.

2- Locate the sheep

Results from Analysis

To help the ruminant-enthused scientists team as first step I have observed again the *species_info.csv* in order to figure out the sheep species in the file

	category	scientific_name	common_names	conservation_status	is_protected	is_sheep
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True
3014	Mammal	Ovis canadensis	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True
4446	Mammal	Ovis canadensis sierrae	Sierra Nevada Bighorn Sheep	Endangered	True	True

How we can see above there are just 3 kinds of sheep. Merging this file and the observation file I have created a file able to represent the number of sheep observation for each National Park. Below we can see that Yellowstone National Park has been the park with more sheep sightings with 507, instead the Park with less sighting has been the Great Smoky Mountains National Park with 149 sighting

	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
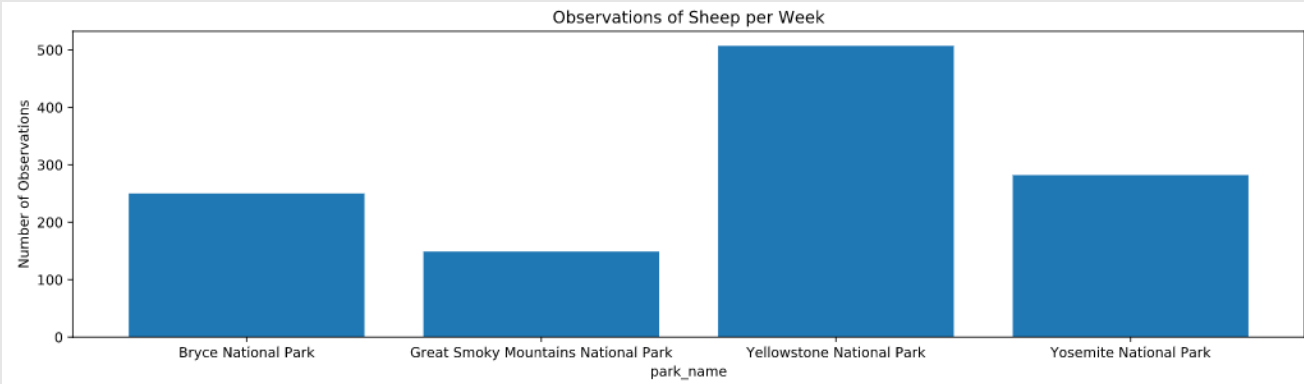
3- Reduce the rate of foot and mouth disease

Results from Analysis

The last part of the analysis has been to help the Park Rangers at Yellowstone National Park to reduce the rate of foot and mouth disease. The Rangers' want to be able to detect reductions of at least 5 percentage points. To demonstrate this reduction I have calculated all the coefficient in the image below, and thanks to an online calculator I have figure out the size of population needed to demonstrate a reduction >5% of foot and mouth disease.

Baseline conversion rate:	15	%
Statistical significance:	85% **90%** 95%	
Minimum detectable effect:	33	%
Sample size:	890	

	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

Using the observation data you analyzed earlier, I found for example that this would take 1.76 weeks at Yellowstone National Park to observe enough sheep and 3.56 weeks at Bryce National Park.

Thank You