Introduction to Data Analysis - Capstone Project: Biodiversity for the National Parks

By Christopher Mackey

The data in species_info.csv

- The csv file contained an ID field and four fields:
  1) **category**: States one of seven different species archtypes:
     - Mammal
     - Bird
     - Reptile
     - Amphibian
     - Fish
     - Vascular Plant
     - Nonvascular Plant
  2) **scientific_name**: States the official latin name for the specific species.
  3) **common_names**: States the more common names for the specific species.
  4) **conservation_status**: States one of five statuses for the specific species:

Species of Concern	**Threatened**
Endangered	**In Recovery**
nan (No Data)	

Things of Note in species_info.csv

- The table contains 5,541 different species.
- Using a GroupBy on category, we get the following table:

ID	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

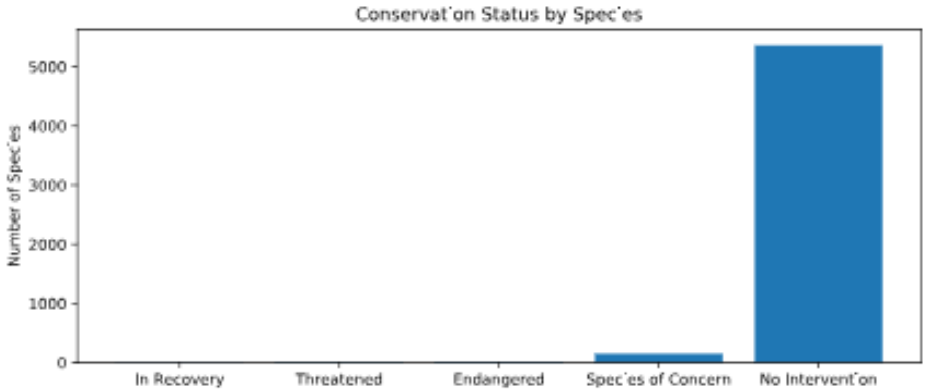
- The table shows 82.9% of the species on the table are in the plant category leaving 17.1% as a type of animal.

Conservation Status Distribution

- Next I checked the distribution of the data by conservation_status
  - Doing a Groupby on conservation_status produced the following table:

	conservation_status	scientific_name
0	Endangered	15
1	In Recovery	4
2	No Intervention	5363
3	Species of Concern	151
4	Threatened	10

  - I then created a bar graph of the above table:



Conservation Status Distribution Continued

- Both the table and the bar graph show “No Intervention” made up a majority of the data.
- I then made another groupby to better show the distribution of conservation status by category:

ID	category	conservation_status	scientific_name
0	Amphibian	Endangered	1
1	Amphibian	No Intervention	72
2	Amphibian	Species of Concern	4
3	Amphibian	Threatened	2
4	Bird	Endangered	4
5	Bird	In Recovery	3
6	Bird	No Intervention	413
7	Bird	Species of Concern	68
8	Fish	Endangered	3
9	Fish	No Intervention	115
10	Fish	Species of Concern	4
11	Fish	Threatened	4
12	Mammal	Endangered	6
13	Mammal	In Recovery	1
14	Mammal	No Intervention	146
15	Mammal	Species of Concern	22
16	Mammal	Threatened	2
17	NonVascular Plant	No Intervention	328
18	NonVascular Plant	Species of Concern	5
19	Reptile	No Intervention	73
20	Reptile	Species of Concern	5
21	Vascular Plant	Endangered	1
22	Vascular Plant	No Intervention	4216
23	Vascular Plant	Species of Concern	43
24	Vascular Plant	Threatened	2

- Discoveries from this table:
  - Mammals have the most in the ‘endangered’ category for animals.
  - Birds have the most species entries not on the ‘No Intervention’ list at 75.
  - Reptiles appear to be the least affected, only having 5 species in the ‘Species of Concern’.

Looking at Endangered Species

- I then created a pivot table to make it easier to read how many species were protected in each category:

ID	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

Significance Testing of Endangered Species

- Using the numbers from the pivot table, I did a Chi-squared Test to determine if certain species are more likely to be endangered or if it's by chance:
  - Comparing Birds and Mammals created a score of ~0.68 which isn't significant.
    - Comparing the percentages of protected Birds and protected Mammals showed it wasn't significant and is a result of chance.
  - Comparing Mammals to Reptiles created a score of ~0.038 which is significant.
    - Comparing the percentages of protected Mammals and protected Reptiles showed it was significant and not by chance.
- In conclusion through the Chi-squared test: Conservationists should know that certain species are more likely to be endangered than others.

observations.csv

- The next section introduced the “observations.csv” file to use in conjunction with the “species_info.csv” file in order to perform more specific analysis on sheep.
- I first narrowed the species table to only show sheep species.

	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

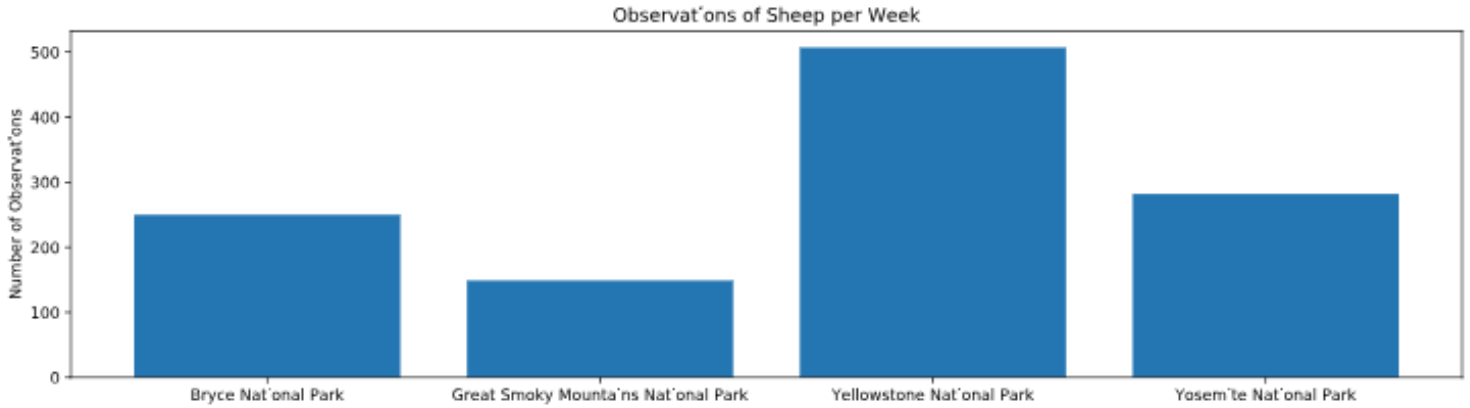
- I then combined the newly created table with the observation data by scientific name to show each sheep observation with its conservation status in each park:

	category	scientific_name	common_names	conservation_status	is_protected	is_sheep	park_name	observations
0	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True	Yosemite National Park	126
1	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True	Great Smoky Mountains National Park	76
2	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True	Bryce National Park	119
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True	Yellowstone National Park	221
4	Mammal	Ovis canadensis	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True	Yellowstone National Park	219

Number of Observations of Sheep

- I then created a table showing the total number of observations of sheep in each park as well as a bar chart to better visualize the data.

	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



Identifying the Sample Size for Foot and Mouth Reduction effort

- To test if Yellowstone's Foot and Mouth Disease program is working, I needed to decipher how large the sample size would need to be for the test. To do this I identified three things:
  - The Baseline: This is last year's observation of the total population of sheep have foot and mouth disease which is 15%.
  - Minimum Detectable Effect: This is 100*(% of effect [5])/(Baseline: 15) which equals 33.33333
  - The Statistical Significance: This is to be set at 90%, the default amount.
- The result was a sample size of 870 would be needed in order to properly determine if the program is working.
  - For Yellowstone this would take a 1.72 weeks to get the necessary number of observations.
  - For Bryce it would take 3.48 weeks to get the number of observations.