BIODIVERSITY CONSERVATION

Data Analysis Project for National Parks Service

By Yvonne So

Data in Species_info

- The Species_info csv file includes 4 fields, which are category, scientific names, common names and conservation status. There are a total of 5824 species in the file.

- Each species is categorized into mammal, bird, reptile, amphibian, fish, and vascular plant, and nonvascular plant:

	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), Dom...	NaN
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	NaN
4	Mammal	Cervus elaphus	Wapiti Or Elk	NaN

- Each species is also assigned into a conservation status of “Species in Concern”, “Threatened”, “Endangered”, “In Recovery”.

Data in Species_info

- In order to know the count of conversation_status, a group by function is performed, which returns the following:

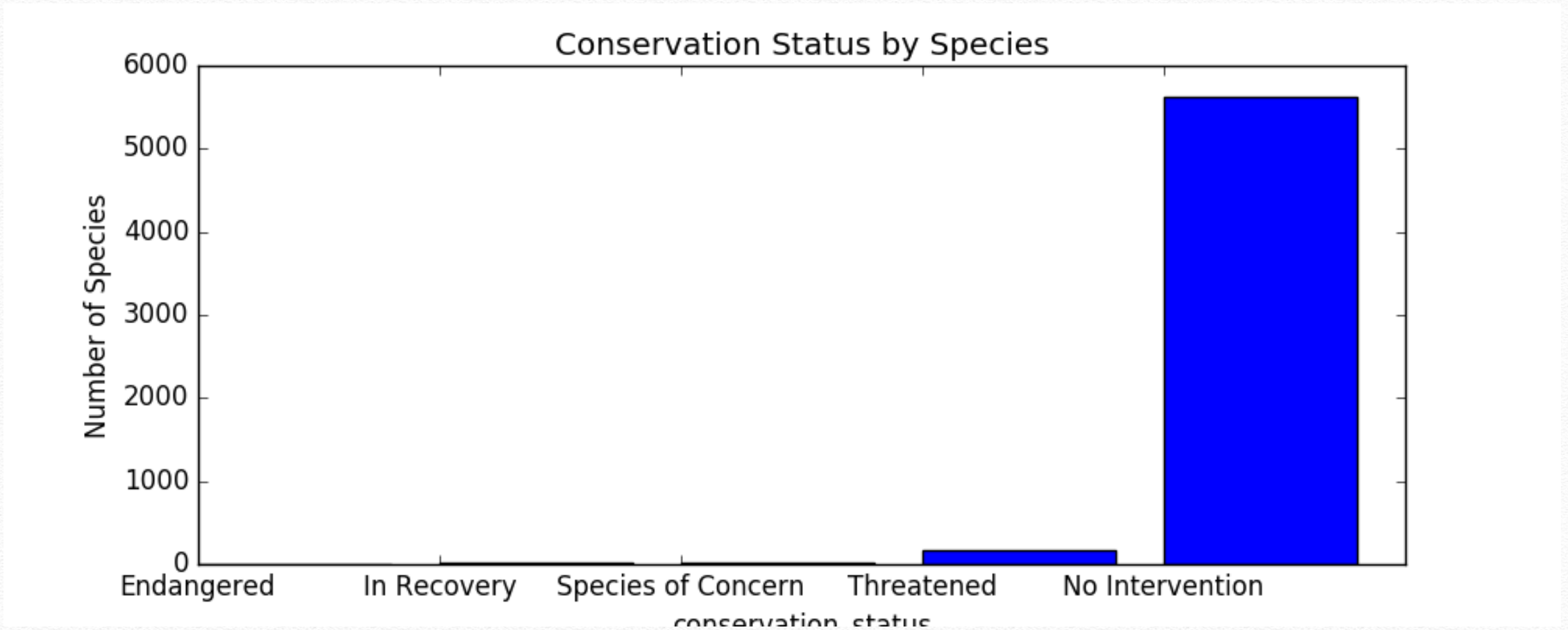
	conservation_status	scientific_name
0	Endangered	16
1	In Recovery	4
2	Species of Concern	161
3	Threatened	10

- The above table shows that a large number of rows is not assigned into any conversation status, we've decided the fill those rows with "No Intervention":

	conservation_status	scientific_name
1	In Recovery	4
4	Threatened	10
0	Endangered	16
3	Species of Concern	161
2	No Intervention	5633

Data in Species_info

A bar chart is created to display the conservation status by each category:



Calculations for endangered species

- To calculate the percentage of species that are engendered by category, a column called “is_protected” was created, which is False if the species needs no intervention, and otherwise True:

- **is_protected = lambda x: True if x <> 'No Intervention' else False: species[‘is_protected'] = species.conservation_status.apply(is_protected)**

- A table is created by grouping columns “is_protected” and “category”, while using scientific_names as count

	is_protected	category	scientific_name
0	False	Amphibian	73
1	False	Bird	442
2	False	Fish	116
3	False	Mammal	176
4	False	Nonvascular Plant	328

- The above table was modified into a pivot table. The “False” from “is_protected” column is changed into “not_protected”, and “True” into “protected”:

is_protected	category	False	True
0	Amphibian	73	7
1	Bird	442	79
2	Fish	116	11
3	Mammal	176	38
4	Nonvascular Plant	328	5
5	Reptile	74	5
6	Vascular Plant	4424	46

Calculations for endangered species

- The percentage of protected and not_protected is then calculated for each category in the pivot table by using: **category_pivot['percent_protected'] = category_pivot.protected/category_pivot.not_protected * 100:**

is_protected	category	not_protected	protected	percent_protected
0	Amphibian	73	7	9.589041
1	Bird	442	79	17.873303
2	Fish	116	11	9.482759
3	Mammal	176	38	21.590909
4	Nonvascular Plant	328	5	1.524390
5	Reptile	74	5	6.756757
6	Vascular Plant	4424	46	1.039783

Bases on the above table, it looks like species in category Mammal are more likely to be endangered than species in Bird. Since we are comparing categorial data (endangered or not endangered) between two species, we will perform a chi-square test in the following slide.

Recommendations for conservationists concerned about endangered species

- A contingency table is created based on the above table:

	protected	not protected
Mammal	?	?
Bird	?	?

- In our case, our contingency table would be the percentages of protected and not protected by mammal and bird:

- [[79, 442],
- [38, 176]]

- The following code is used to run the chi-square test:
- **contingency = [[79, 442],**
- **[38, 176]]**
- **from scipy.stats import chi2_contingency**
- **chi2, pval, dof, expected = chi2_contingency(contingency)**
- This returns a p value of: 0.445901703047.

- Since the p value is larger than 0.05, the test result indicates that there is not a significant difference in terms of being endangered between mammal and bird, therefore we reject our initial hypothesis that Mammal are more likely to be endangered than species in Bird.

Recommendations for conservationists concerned about endangered species

- Let's now look at mammal and reptile: our contingency table would be:

- X = [[38, 176],
- [5, 74]]

- The following code is used to run the chi-square test:
- **X = [[38, 176],**
- **[5, 74]]**
- **chi2, pval, dof, expected = chi2_contingency(X)**
- which returns a p value of: 0.0233846521487

- Since the p value is less than 0.05, the test result indicates that is a significant difference in terms of being endangered between Mammal and Reptile, and therefore Reptile is more likely to be danger than Mammal.

Recommendations for conservationists concerned about endangered species

- After running more chi-square tests, we discovered that the following categories all have a p-value smaller than 0.05 when being measured against Mammal:

Category	P-value
Fish	0.031145
Nonvascular Plant	1.68189E-11
Vascular Plant	1.734911E-70

- Based on the data, I would suggest conservationists to take active measure to increase the number of protected species that belongs to the Fish, Nonvascular Plant, Vascular Plant, and the Reptile category, since they are more likely to be endangered than Mammal.

Sample size determination or foot and mouth disease study

Upon investigating the Observations.csv, we discover there are 3 fields in the file:

	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

A lambda function to create a new column in the species file, which is called “is_sheep’. It is true if the common_names contains 'Sheep', and False otherwise.

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

Sample size determination or foot and mouth disease study

- The rows where column “is_sheep” is True and category belongs to Mammal are then extracted using the following code: **sheep_species = species[(species.is_sheep) & (species.category == 'Mammal')]**

- This returns the following table called “sheep_species”:

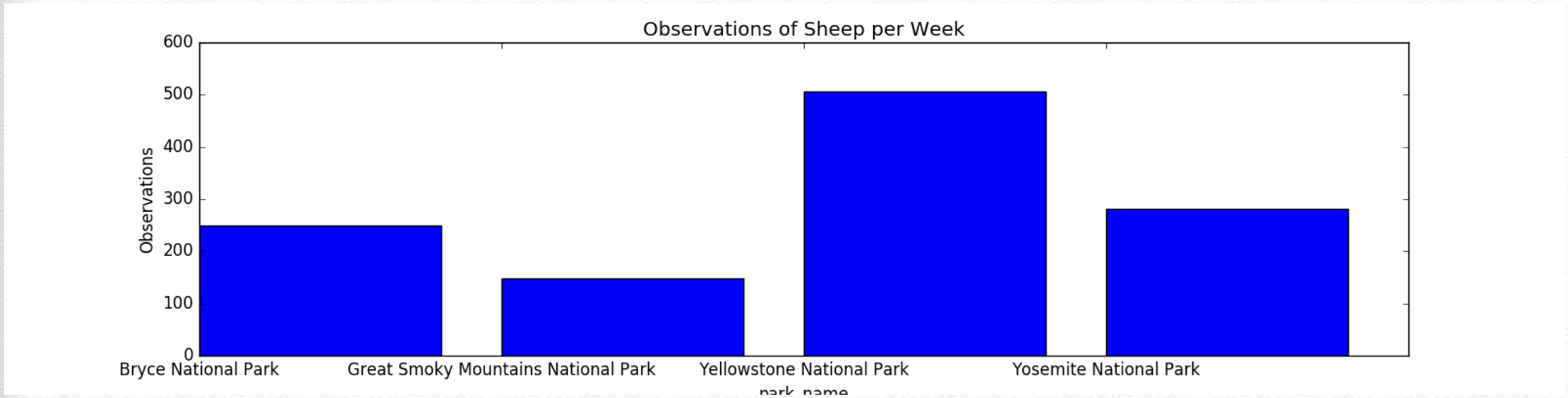
	category	scientific_name	common_names	conservation_status	is_protected	is_sheep
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True
3014	Mammal	Ovis canadensis	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True
4446	Mammal	Ovis canadensis sierrae	Sierra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

- The data frame of sheep_species and observations are then merged into a table called “sheep_observations”. A group function is performed to group the column park_name and use the sum of observation of sheep in each park name as count:

	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 or foot and mouth disease study

- A bar is then created to show the number of observations of sheep per week in each park:



Sample size determination or foot and mouth disease study

Scientists know that 15% of sheep at Bryce National Park have foot and mouth disease. Park rangers at Yellowstone National Park want to reduce the rate of foot and mouth disease at that park. Scientists want to be able to detect reductions of at least 5 percentage point.

In order to calculate the number of sheep that they would need to observe from each park, using a 90% statical significance, the minimum detectable effect would be = 5.0/15*100 = 33.33.

Using this number and the A/B testing sample size calculator at Optimizely, it generates a sample size of 510.

Therefore, at the Bryce National Park, we would need 510 / 250 (number of sheep observed) ~= 2 weeks to observe enough sheep.

At Yellowstone National Park, we would need 510 / 507 ~= 1 week to observe enough sheep.