“Should future Generations think it’s weird we had billions to go to the moon … but not to protect endangered species on earth?”

A special data report from the National Park Service

Section 1 : Description of the dataset

(1) Description of the data

Our team provided us the data in species_info.csv. When loaded in a Data Frame, we have the following datas :

Category (`category`) - which type of species we are dealing with.

Scientific Name (`scientific_name`) - the latin name understood by every conservationist in the world.

Common Name(s) (`common_names`) - the names also commonly used to describe the species.

Conservation status (`conservation_status`) - how we are considering the species now.

(2) More description of the Data

The structure of the Dataframe is not telling us a lot. With some Python code, we were able to compute some important numbers :

```
Number of species listed in our national parks : 5541

Categories of species : ['Mammal' 'Bird' 'Reptile' 'Amphibian'
'Fish' 'Vascular Plant''Nonvascular Plant']

Situation of species : [nan 'Species of Concern' 'Endangered'
'Threatened' 'In Recovery']
```

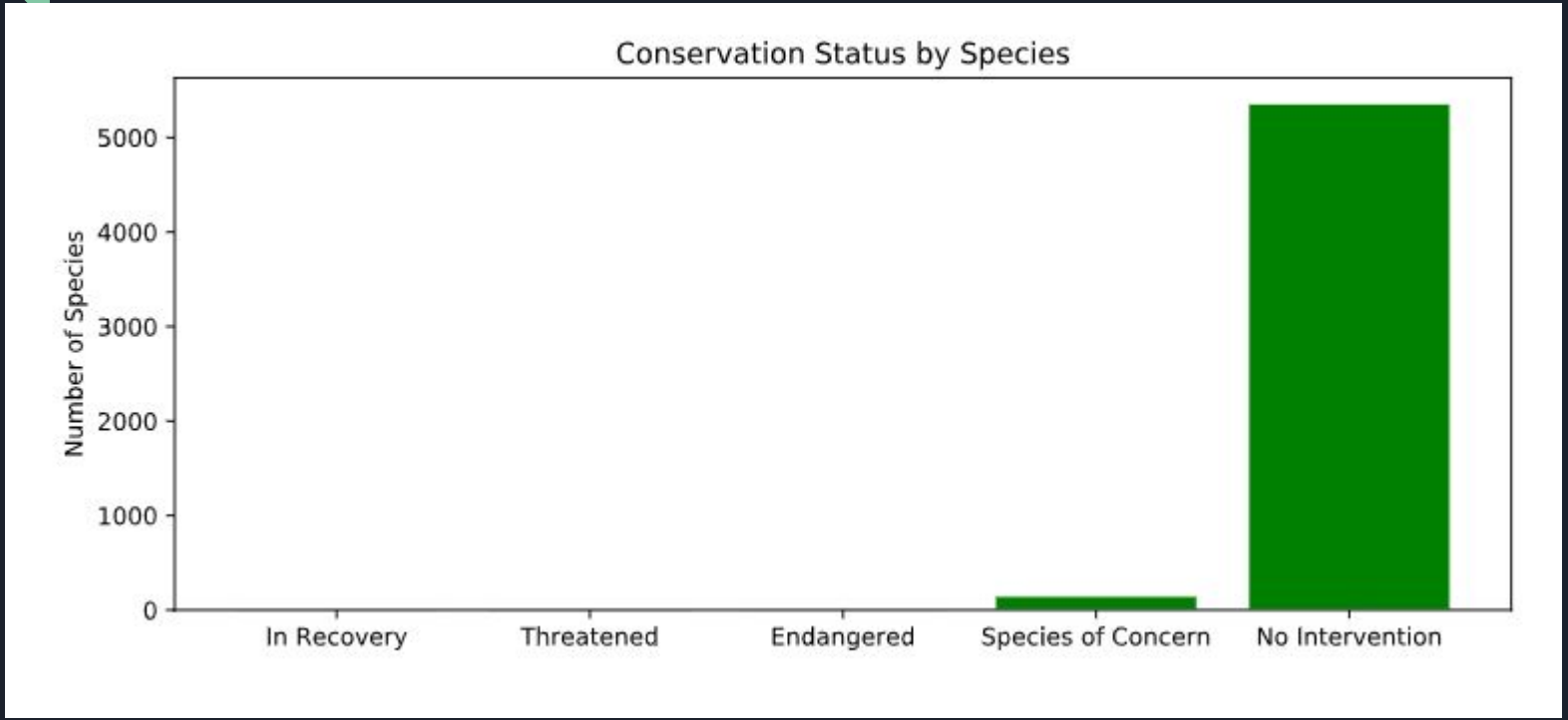
Section 2 : The real situation on endangered species

(3) The global situation

How many species are endangered in our national parks ? We needed to correct the data. Majority of species have a negative conservative status, which means no intervention.

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

(4) No intervention is the main case.



(5) Endangered species by categories

The global situation gives us already informations. We focus our efforts on near 200 species. It gives us hope, because we already success to recover some species (4).

But as we continued the analysis, we discovered big differences between the categories of species.

By using groupby and pivot on Python, we were able to generate the following table :

(6) Endangered species by categories

	category	not_protected	protected
0	Amphibian	72	7
1	Bird	413	75
2	Fish	115	11
3	Mammal	146	30
4	Nonvascular Plant	328	5
5	Reptile	73	5
6	Vascular Plant	4216	46

Section 3 : Is there any species more in danger than others ?
(Hint : sadly, yes.)

(7) Chi-squared test

To verify the hypotheses that some species are more in danger than others, we decided to apply a Chi-squared test to identify the significance.

When p-value > 0.05, the hypothesis is result of a chance.

When p-value < 0.05, the hypothesis is significant.

We decided to conduct several chi-squared to draw conclusions.

(8) Results of our chi-squared tests

Mammal / Reptile : p = 0,038 (significant)

Amphibian / Non-vascular Plant : p = 0,01 (significant)

Bird / Reptile : p = 0,053 (significant)

Mammal / Fish : p = 0,0056 (significant)

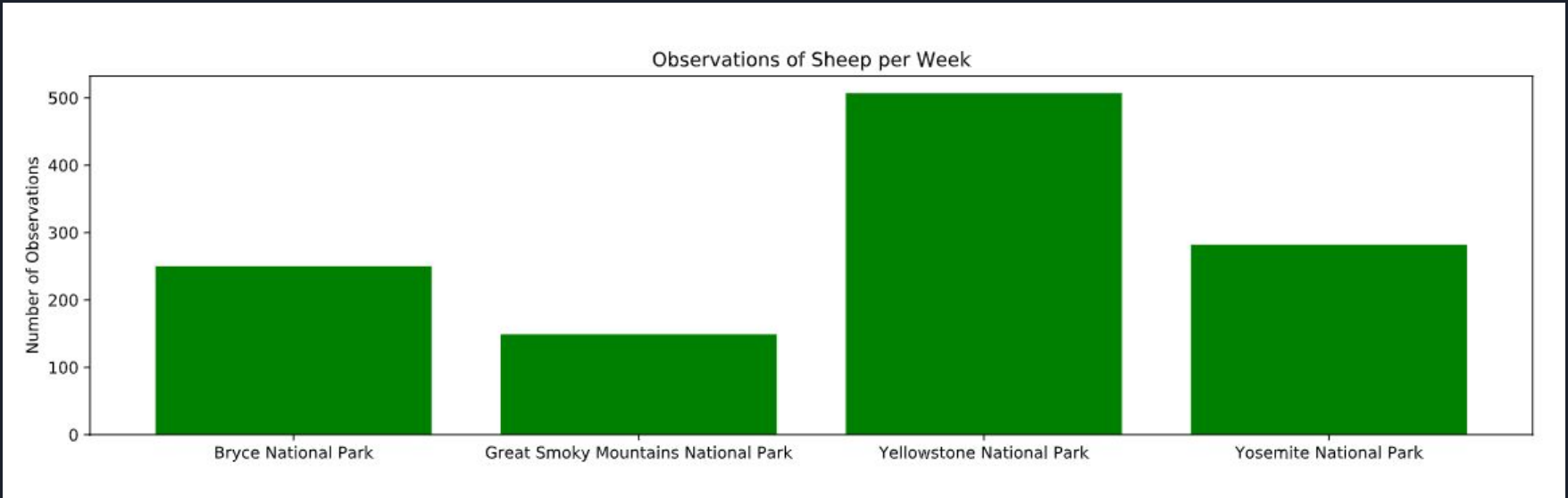
(9) Recommendations for conservationists

- Some species are indeed more at risk than others. The difference between the percentages of protected reptiles and mammals is significant.
- We analyzed especially the difference between the animal and vegetable worlds. It seems that in our national parks, the animal biodiversity is more threatened than the vegetable ones.
- We need to stay humble about our datas. How many species are listed in “no intervention” when an intervention should be needed ?

Section 4 : results of our observations in our parks

(10) Description of the dataset

Conservationists had observed different species at several national parks for the past 7 days.



(11) Foot and Mouth injuries reduction effort

We know two things :

- the goal : to be able to detect reductions of diseases of at least 5%
- the actual situation : last year, 15% of sheep at Bryce National Park have foot and mouth disease.

Baseline percentage of this sample size determination : it's the actual situation, so 15%.

```
Baseline = 15
```

Minimum detectable effect is a percent of the Baseline : we want 5% change with confidence. So it is

```
minimum_detectable_effect = 100 * 5. / 15
```

(12) Foot and Mouth injuries reduction effort

Sample Size by variant with a level of significance to 90% :
870 (if you take Minimum Detectable Effect = 33,3) or 890 (if you take MDE = 33)

From these datas, we need to reuse obs_by_park :

	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

(13) Foot and Mouth injuries reduction effort

So, the week needed to observe enough subjects per park are :

Bryce = 3,48

Great Smoky Mountains = 5,83

Yellowstone = 1,71

Yosemite = 3,08