BIODIVERSITY FOR THE NATIONAL PARKS

species_info.csv

A CSV file species_info.csv consists of the data about different species in the National Parks, including:

- The scientific name of each species
- The common names of each species
- The species conservation status

	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

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

species_info.csv

- There are 5541 different species are in the species DataFrame.
- 'Mammal' 'Bird' 'Reptile' ‘Amphibian' 'Fish' 'Vascular Plant' 'Nonvascular Plant’ are the different values of category.
- ‘'Species of Concern' 'Endangered' ‘Threatened' 'In Recovery’ are the different values of **conservation_status**. We used the command **groupby** to see the status of each specie. As we found out the majority of species have no status assigned meaning none of the statuses available are relevant to them**.**

```
conservation_counts =
species.groupby('conservation_status').scientific
_name.nunique().reset_index()
```

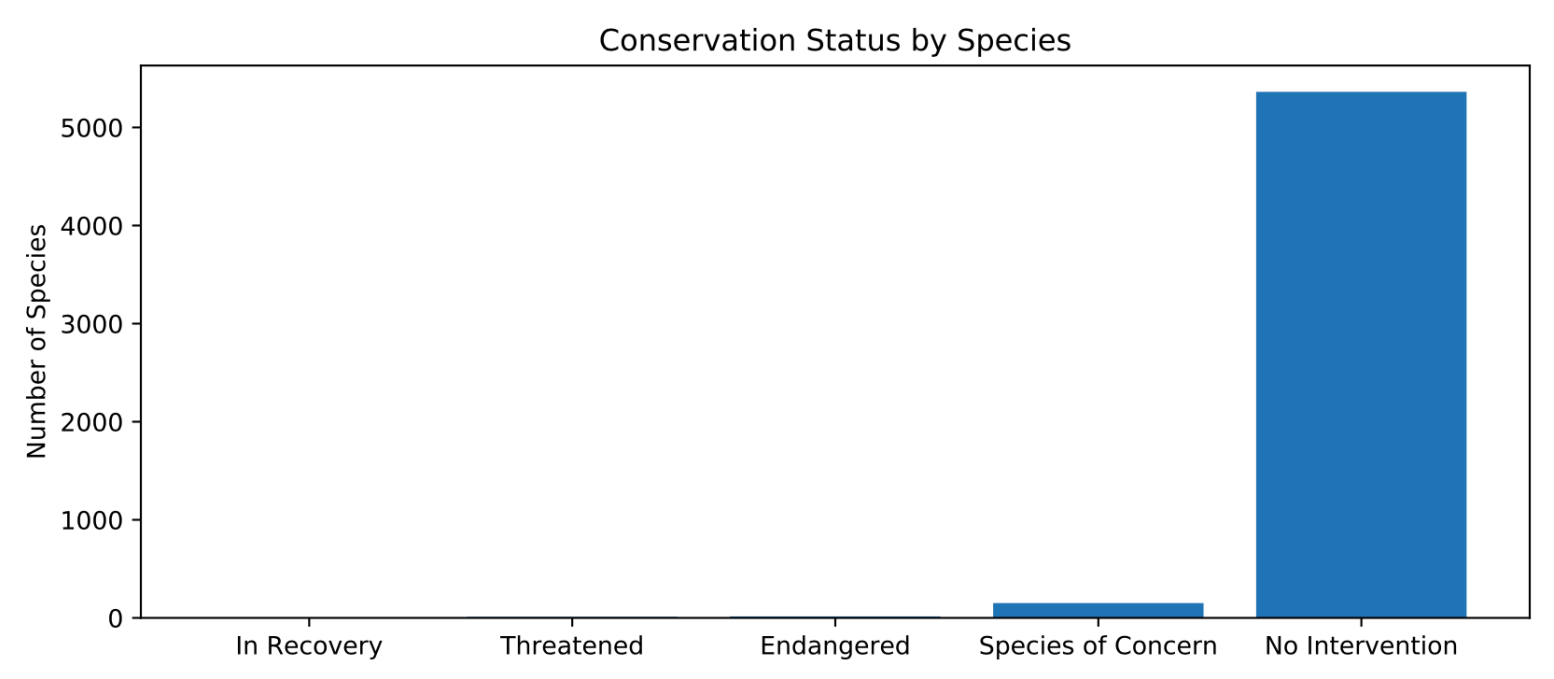
- The status ‘No intervention’ is assigned to the species that don’t fall into any of the categories above. By using the code below we can create a new status ‘No intervention’ for the species which status has been ***nan*** before.

```
species.fillna('No Intervention', inplace = True)
```

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

species_info.csv

As we can see from the bar chart below representing the data from spicies_info.csv, the majority of species fall into the category of ‘No Intervention’. Unfortunately, this data gives us a very limited overview of the species’ status quo. One of the issues we can try and investigate is if **certain types of species more likely to be endangered**.



species_info.csv

In order to identify if certain species are more likely to be endangered we investigated the proportion of the protected species in relation to the not protected species for each type of the specie.

First, for each category we identified which species are protected (or have a status assigned that is not equal to 'No Intervention'). Then, we could calculate the percent protected using this count.

```
species['is_protected'] =
species.conservation_status != 'No Intervention'

category_counts = species.groupby(['category',
'is_protected']).scientific_name.nunique().reset_
index()
```

category	not_protected	protected	percent_protected
Amphibian	72	7	0.088608
Bird	413	75	0.153689
Fish	115	11	0.087302
Mammal	146	30	0.170455
Nonvascular Plant	328	5	0.015015
Reptile	73	5	0.064103
Vascular Plant	4216	46	0.010793

From the pivot we can clearly see that the Mammals and the Birds have the highest number of their species being protected in comparison to the other categories.

To answer the question if **certain types of species more likely to be endangered** we can start by investigating if Mammals are more likely to be endangered than Birds.

In this case, the null hypothesis is that there's no significant correlation between the likelihood of species being endangered. We reject that hypothesis, and state that there is a significant difference between two of the datasets if we get a p-value less than 0.05.

species_info.csv

We will use the **chi-square test** for independence to determine whether there is a significant relationship between two categorical variables (Mammals and Birds).

```
contingency = [[30, 146],
               [75, 413]]

pval = chi2_contingency(contingency)[1]
```

By creating the contingency table and filling it with the values of species protected and not protected, and calculating the *p value* we can determine if there is a significant correlation between the two.

The *p value* = 0.687594809666. Therefore, we can conclude that there is no significant difference.

As a comparison we can also compare the other two categorical variables (Mammals and Reptiles) - **chi-square test 2:**

```
contingency_reptile_mammal = [[30, 146],
                              [5, 73]]

pval_reptile_mammal =
chi2_contingency(contingency_reptile_mammal)[1]
```

The *p value* = 0.0383555902297. Therefore, we can conclude that there is significant difference.

Therefore, we can conclude that certain types of species *are* more likely to be endangered than others.

species_info.csv

Recommendations for conservationists concerned about endangered species:

- By looking at the values in the percent_protected column it can be concluded that mammals and birds are the most like to become distinct.
- It is important to remember that a species with 500 animals left could be considered more endangered than one with only 300 left if that species is localised to one area and has a long reproductive cycle meaning the population cannot quickly grow.
- If species are dying out, it is an indication of the long-term health of our own species, and we need to be aware of the impact we are having on our own ecosystem.

observations.csv

The data in the observations.csv consist of information that conservationists have been recording sightings of different species at several national parks for the past 7 days.

	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

From the table above we can distinguish following types of sheep: Domestic sheep, Bighorn Sheep, Sierra Nevada Bighorn Sheep in Mammal category.

The the observations.csv only contains the scientific names of species, so we used species.csv to look for any names that refer to sheep.

To determine in which Parks these sheep are locating we merge the data in sheep_species and observations by using the **.merge** command:

```
sheep_observations =
observations.merge(sheep_species)
```

observations.csv

After merging our data we used .groupby comman to see the total sheep sightings (across all three species) were made at each national park:

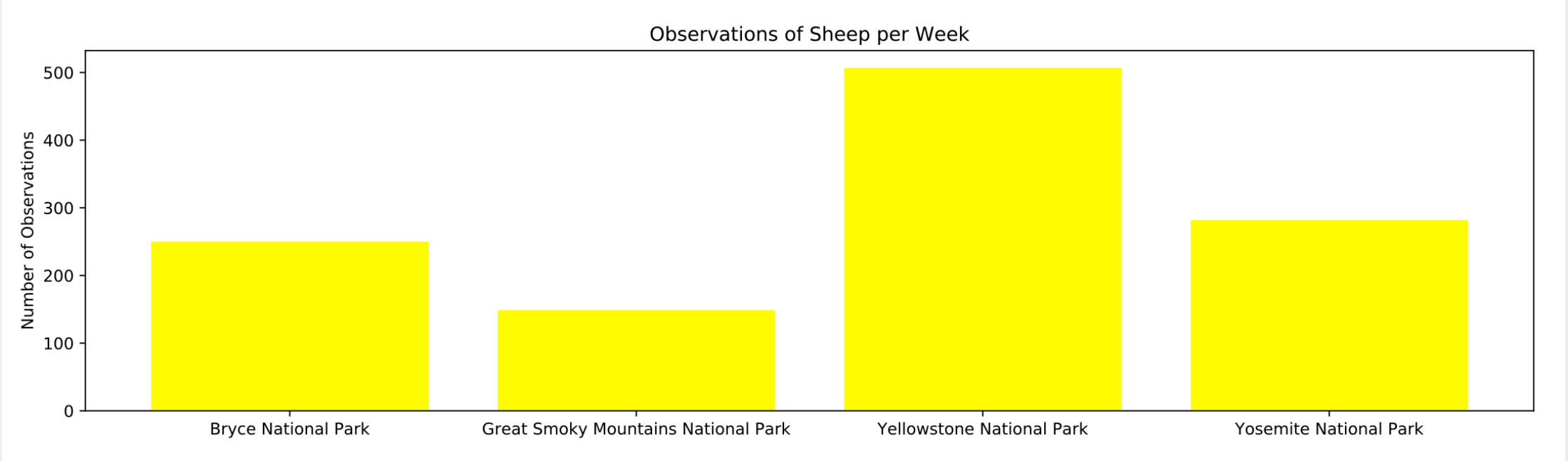
```
obs_by_park =
sheep_observations.groupby('park_name').observations.sum().reset_index()
```

This command produced the table below:

	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

observations.csv

Then we graph the sheep observation data to visualise he number of sightings at each of the four national parks under investigation:



Foot and Mouth Reduction Effort - Sample Size Determination

Park Rangers at Yellowstone National Park have been running a program to reduce the rate of foot and mouth disease at that park.

To test whether or not this program is working and to detect reductions of at least 5 – for that we used **A/B testing (**is a randomized experiment with two variants)

For the calculator we need x3 values: baseline percentage of this sample size determination, Minimum detectable effect and Statistical significance

- 15% of sheep at Bryce National Park have foot and mouth disease, therefore – this is our **baseline** variable.
- **minimum_detectable_effect** = 100*5 (o be able to detect reductions of at least 5 percentage points)/15(baseline) = 33.3

The calculator produced the value of **870** as **the sample_size_per_variant**

Foot and Mouth Reduction Effort - Sample Size Determination

To calculate how many weeks would the scientists need to spend at Yellowstone National Park to observe enough sheep we used the formulas below:

```
yellowstone_weeks_observing =
sample_size_per_variant/507.

bryce_weeks_observing =
sample_size_per_variant/250.
```

Scientist will need to spend 3 extra weeks at Bryce National Park because there is a fewer sheep available