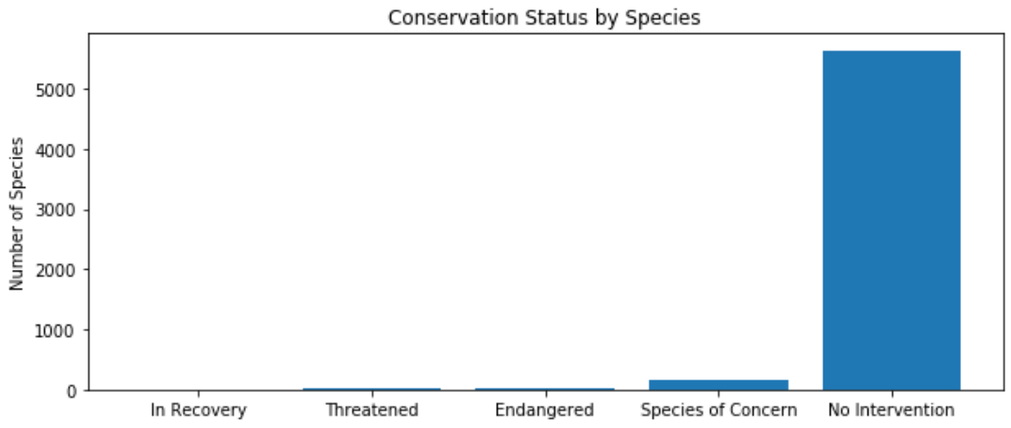
Capstone Project No.2

Biodiversity

We are going to analyse some data about species at different national parks location.
We are given a dataset in form of a .csv file which contains information about several species, with their scientific names, common names, conservation status and category.
More precisely there are a number of different species listed in the dataframe, which is calculate with the code different_species = species.scientific_name.unique()
The result is 5541 species. While categories are 7: ['Mammal' 'Bird' 'Reptile' 'Amphibian' 'Fish' 'Vascular Plant' 'Nonvascular Plant']
Then we were asked to groupby conservation status and here is the table and the bar plot for it.

1)

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



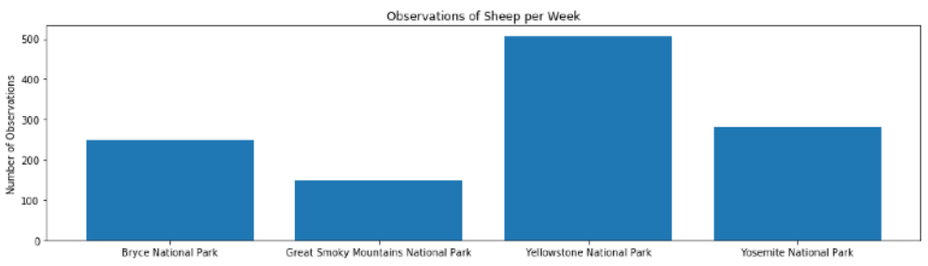
Then we were asked the relevant question: are certain species more likely to be endangered? Thanks to some manipulation of the existing dataset including a pivoting we ended up with a table summarising the global picture: for each category how many species are protected and how many are not? The answer is shown in table 2), with also percentage of protected categories.

2)

	category	not_protected	protected	percent_protected
0	Amphibian	73	7	8.750000
1	Bird	442	79	15.163148
2	Fish	116	11	8.661417
3	Mammal	176	38	17.757009
4	Nonvascular Plant	328	5	1.501502
5	Reptile	74	5	6.329114
6	Vascular Plant	4424	46	1.029083

The following analysis is focusing then on whether `Mammal` are more likely to be endangered than species in `Bird`. We are requested to perform some analysis on categorical data, more specifically belonging to this 2 categories, mammal and bird, hence a Chi_square method was adopted to evaluate whether the observed fluctuations are statistical or random.
The pval resulted from this test (0.445901703047) is well above the threshold 0.05 hence, in the case of mammal and bird there is no significative difference statistically between the distribution of those 2 categories.
Applying the same chi square method between reptile and mammal instead resulted in a pval = 0.0233846521487 so the null hypothesis can be rejected and we can say there is a significative difference in the stats.

In the second half of the project we focused on sheeps. By manipulating and merging our data we extracted information on the weekly observations done for each species of sheep, the location and their conservation status



From data bighorn sheep is a species of concern, while the Sierra Nevada bighorn sheep is an endangered specie. For ovis aries (domestic sheep) there is no concern.
To carry out studies for the mouth and foot disease I considered the following numbers:
Baseline = 15 % (the percentage of sick sheeps)
Since they want to be able to detect a 5% variation (the lift), I calculated the Minimum detectable effect (in per cent) as: 0.15/0.05 = 33.3 %. The significance is given as a 90% confidence level.
Then, based on a sample calculator, we estimated the number of weeks for observing a possible effect of the cure adopted to reduce a mouth and foot diseases affecting sheeps.
Given the different population size in the different locations, the number of weeks would be different, that is (rounded number):

- **2 weeks of observation for sheeps at Bryce National Park**
- **1 week for sheeps at Yellowstone National Park**

From this analysis looks like the situation for the Sierra Nevada Bighorn sheep at Bryce National Park is serious because it is an endangered species and also there are only 22 observations. Hence it is recommendable to focus on this item and maybe investing more time and resources.