National Parks Service

Conservation Study

An analysis of different species at Yosemite, Bryce, Yellowstone, and Great Smokey Mountains National Parks.

Endangered Species Analysis

Analysis started with using the python pandas package to make a dataframe from the raw data contained in 'species_info.csv'. Investigating the data we see that the table formed by 'species_info.csv' contains 4 columns and 5541 unique species.

	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
5	Mammal	Odocoileus virginianus	White-Tailed Deer	NaN
6	Mammal	Sus scrofa	Feral Hog, Wild Pig	NaN
7	Mammal	Canis latrans	Coyote	Species of Concern
8	Mammal	Canis lupus	Gray Wolf	Endangered
9	Mammal	Canis rufus	Red Wolf	Endangered

```
number_of_species = species.scientific_name.nunique()  # 5541
```

Endangered Species Analysis

Taking this analysis a step further, we can see all the categories of species contained within this data.

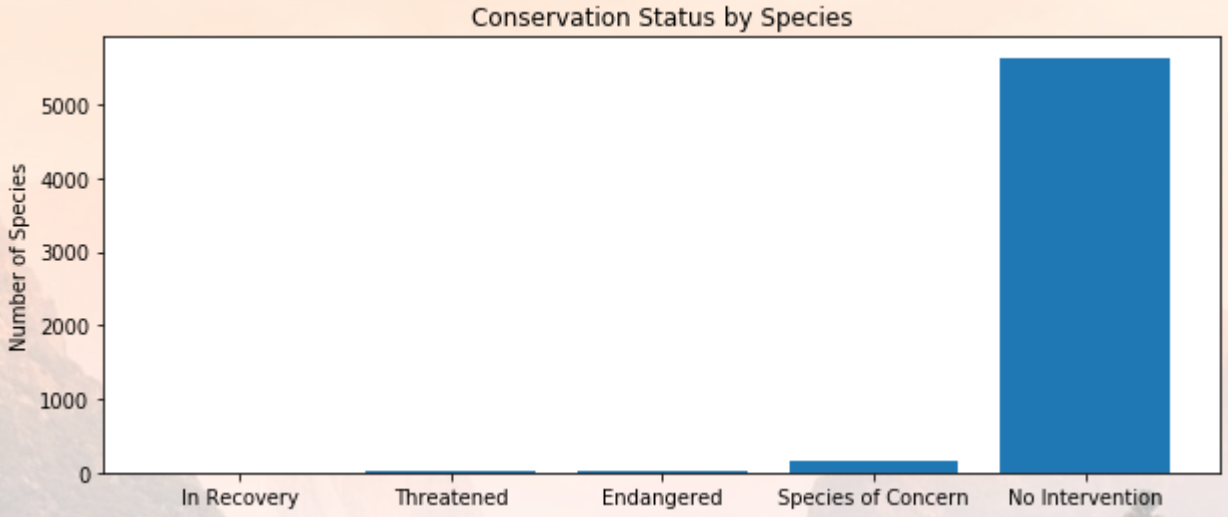
	category	species_per_category
0	Amphibian	80
1	Bird	521
2	Fish	127
3	Mammal	214
4	Nonvascular Plant	333
5	Reptile	79
6	Vascular Plant	4470

And by manipulating information about conservation status, we can summarize the conservation status per category.

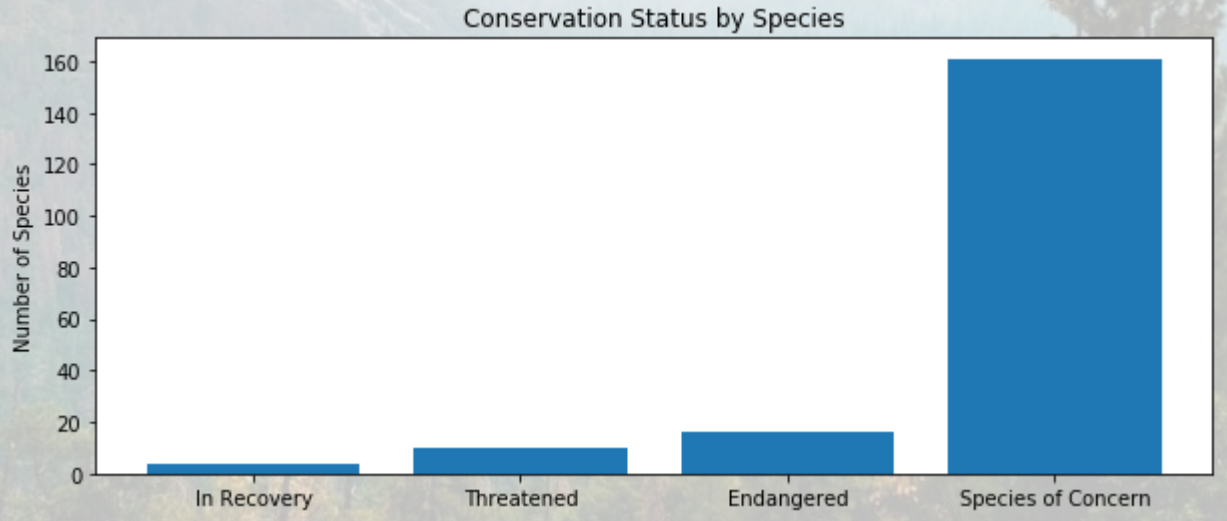
	conservation_status	species_per_conservation_status
0	In Recovery	4
1	Threatened	10
2	Endangered	16
3	Species of Concern	161
4	No Intervention	5633

Endangered Species Analysis

No one likes tables – the previous table can easily be represented by a bar chart.



Clearly the majority of species are not endangered! Removing 'No Intervention' status, we can read the categories of interest more easily:



Endangered Species Analysis

As we saw in from the last bar chart, there are about 200 species that are classified as 'Species of Concern' or worse. In order to make conservation efforts for these species, we'll need to gather more information from this data. First, we can create a new column to isolate all of the species that we're concerned with:

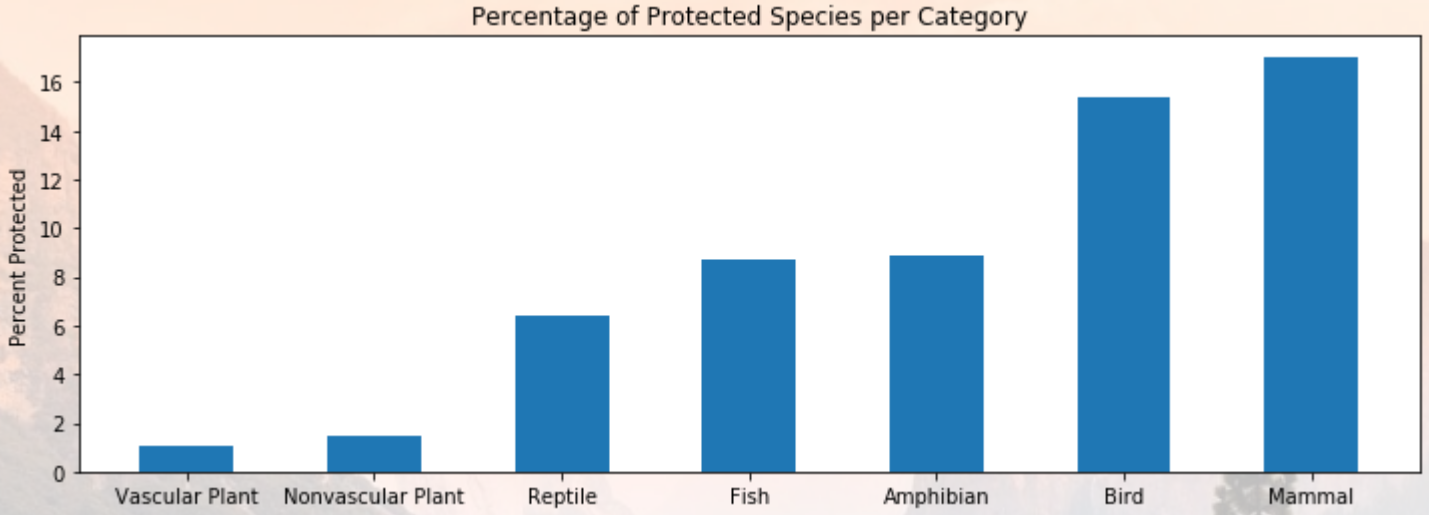
	category	scientific_name	common_names	conservation_status	is_protected
0	Mammal	Clethrionomys gapperi gapperi	Gapper's Red-Backed Vole	No Intervention	False
1	Mammal	Bos bison	American Bison, Bison	No Intervention	False
2	Mammal	Bos taurus	Aurochs, Aurochs, Domestic Cattle (Feral), Dom...	No Intervention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

Using this column and our knowledge of pivot tables, we can create a count of the species of concern in each category and compare it with the total to get a percentage :

	category	not_protected	protected	percent_protected
0	Amphibian	72	7	8.860759
1	Bird	413	75	15.368852
2	Fish	115	11	8.730159
3	Mammal	146	30	17.045455
4	Nonvascular Plant	328	5	1.501502
5	Reptile	73	5	6.410256
6	Vascular Plant	4216	46	1.079305

Endangered Species Analysis

We can also graph the percentage of protected species per category to get a visual understanding of the species affected:



Here, we see ‘Mammals’ has a higher percentage of protected species than ‘Birds’, but does that mean that mammals are more *likely* to be at risk of endangerment than birds?

For two or more categorical datasets, we use the **Chi-Square** test to answer this question.

Endangered Species Analysis

Here is the data that we compare using the Chi-Square test:

	protected	not protected
Mammal	30	146
Bird	75	413

Which returns a p-value (probability that the sample data doesn't represent reality) of 68.75%. Here we conclude that there is no significant difference between the data sets, and mammals are not more at risk than birds.

We can run the same test against any other category, like 'Reptiles'.

	protected	not protected
Mammal	30	146
Reptile	5	73

This time the test returns a p-value = 3.84%, meaning there is a significant difference between these datasets. Mammals are more likely to be a species of concern compared to reptiles.

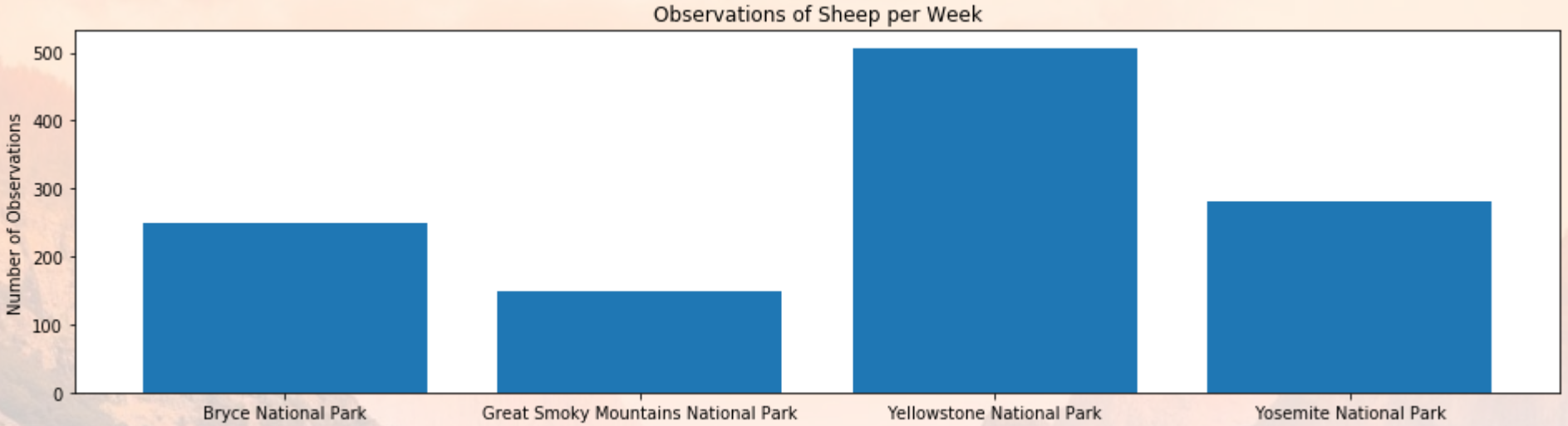
Typically p-values less than 5% is deemed as significant.

Endangered Species Recommendation

Based on the overall breakdown of this information, I would recommend that conservation efforts be focused on mammals and birds. These categories have a high percentage of their species at risk, and are significantly more likely to be at risk when compared to other categories.

It should also be noted that while vascular plants have a relatively high number of species at risk, this comprises a very small percentage of their overall population.

Foot and Mouth Disease Analysis



Data shows that 15% of the sheep at Bryce and Yellowstone national park have foot and mouth disease. A program has been running to reduce the number of sheep infected by the disease.

Our scientists want to be able to detect reductions of 5% with 90% confidence. By setting our values as shown below, we can determine we need a sample size of 510 observations to achieve this.

- Baseline Conversion Rate = 15%
- Minimum Detectable Effect = (Old - New)/Old = 33%
- Statistical Significance = 90%

This study will take approximately 1 week at Yellowstone, and approximately 2 weeks at Bryce National Park.


codecademy