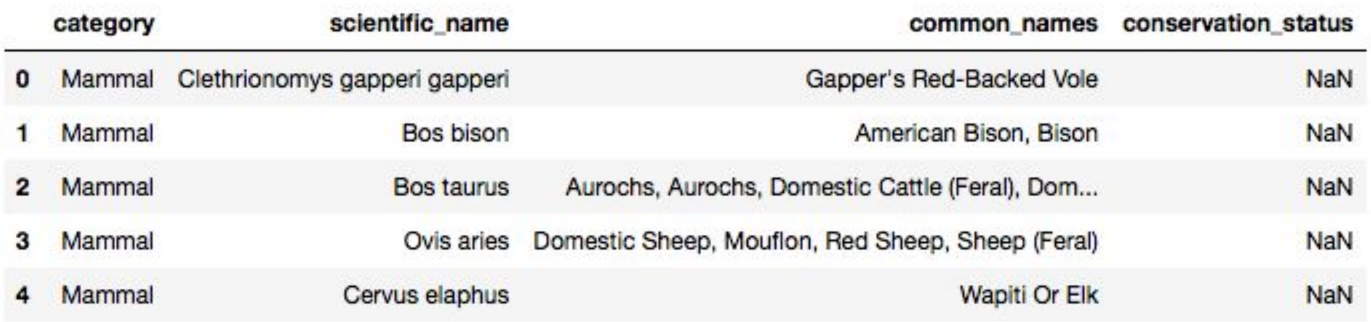
Biodiversity Project

Mia Szarvas (*miasza*)

Codecademy Intro to Data Analysis

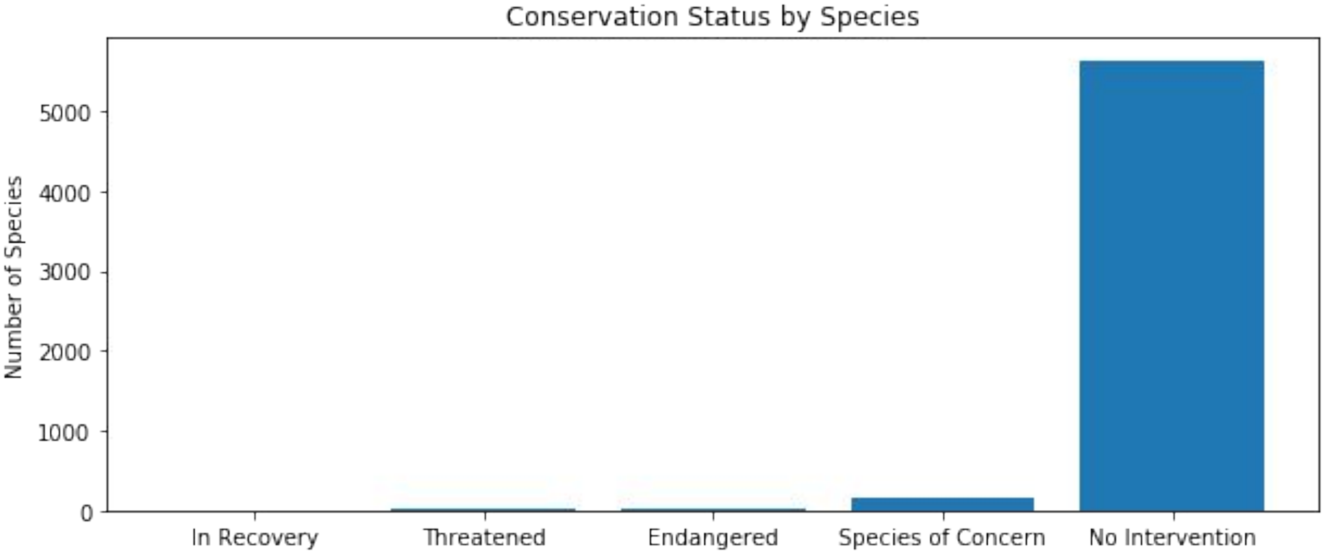
About species.info

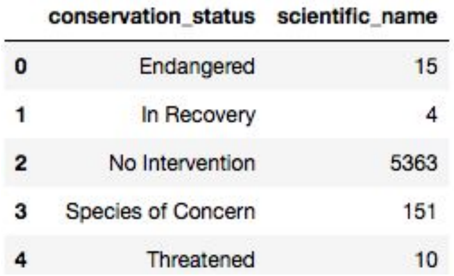
- Species.info is a fictional dataset about different species in United States National Parks
- It includes columns that list the species’ ID, category (common name for species class e.g. ‘Mammal’ for Mammalia class), scientific name of the individual species, common names, and conservation status (Fig. 1).
- There are 5541 unique species in the dataset.
- There are 7 unique category values which are: 'Mammal', 'Bird', 'Reptile', 'Amphibian', 'Fish', 'Vascular Plant', 'Nonvascular Plant'.
- There are 5 unique conservation status values which are ‘NaN’, 'Species of Concern', 'Endangered', 'Threatened', 'In Recovery'. ‘NaN’ stands for species that do not require intervention (Fig. 3). In Fig. 2 and Fig. 3 ‘NaN’ has been replaced with ‘No Intervention’.
- Most species in the dataset do not require invention, as illustrated in Fig. 2.



	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

Fig.1 (left)
species.head( )



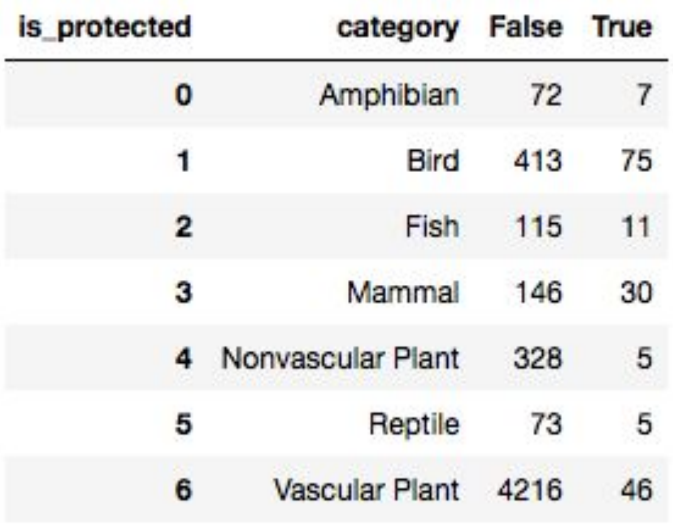


	conservation_status	scientific_name
0	Endangered	15
1	In Recovery	4
2	No Intervention	5363
3	Species of Concern	151
4	Threatened	10

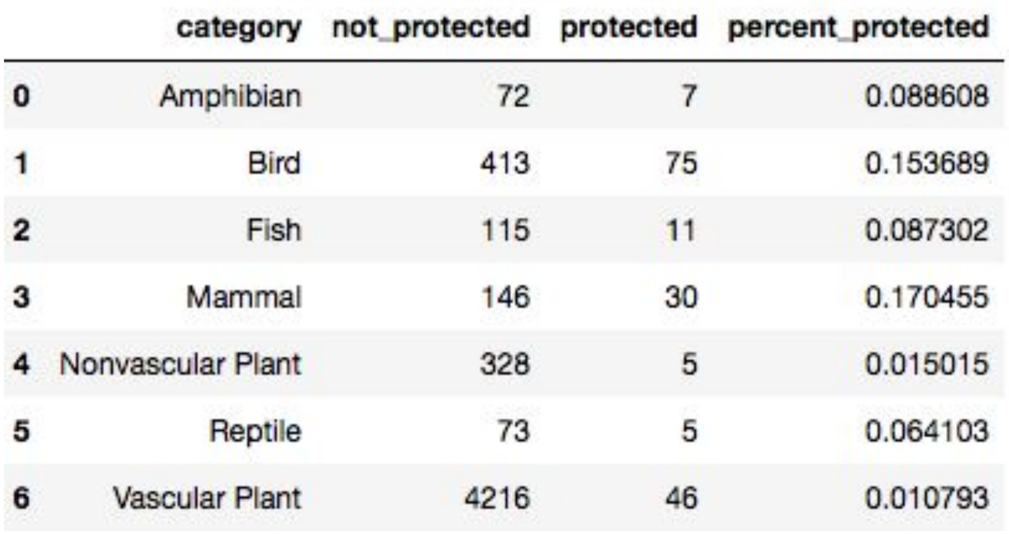
Fig. 2 (left) and Fig. 3 (above)

About species.info con’t

- The category with the most unique species is ‘Vascular Plant’, followed by ‘Bird’, ‘Non-Vascular Plant’, ‘Mammal’, ‘Fish’, ‘Amphibian’, and ‘Reptile’, in descending order (Fig. 4)
- The first category, ‘Vascular Plant’, contains more than 4000 species, the next largest category, ‘Bird’, has just under 500 (Fig. 4)
- There are almost equal amounts of species in the ‘Amphibians’ category and the ‘Reptiles’ category in the data set (Fig. 4)
- The category ‘Mammal’ has the largest percent of species protected (meaning either ‘endangered’, ‘threatened’, ‘species of concern’, or ‘in recovery’) out of all the categories. The ‘Vascular Plant’ category has the lowest percentage of species protected (Fig. 5)



is_protected	category	False	True
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



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

Fig. 4 (above)
Species totals per category were calculated by adding the values from the False and True columns for each row.

Fig. 5 (above)

Significance Calculations I

Part 1: Are species in 'Mammals' category more likely to be endangered than those 'Birds'?

- At first glance, it looks like species in the category 'Mammal' are more likely to be endangered than species in the category 'Bird' (Fig. 5). I ran a significance test to see if this statement was true.
- Because the data was categorical, and I was comparing multiple pieces of data, I chose to run a chi squared test. First I created a contingency table whose inputs were as follows:
  - [[mammal protected, mammal not_protected], [bird protected, bird not_protected]]
- I then ran the chi square test to find the pval. Since the pval was greater than 0.05 (in fact, much greater : ~0.69) I concluded that the difference is not significant, and the initial statement was false.

Species in the category 'Mammal' are not more likely to be endangered than species in the category 'Bird'.

Significance Calculations II

Part 1: Are species in ‘Mammals’ category more likely to be endangered than those ‘Reptiles’?

- It also appeared that species in the category ‘Mammal’ are more likely to be endangered than species in the category ‘Reptile’ (Fig. 5). I ran a significance test to see if this statement was true. For the same reasons as before, I chose a chi squared test following a similar input format, running the test to find the pval.
- Since the pval was less than 0.05 (~0.04) I found that this difference is significant, and the statement above is true.

Species in the category ‘Mammal’ are more likely to be endangered than species in the category ‘Reptile’.

Recommendation for Conservationists

Based on preceding significance calculations

According to our data, mammals are no more likely to be endangered than birds. Thus, conservationists should focus their efforts equally on protecting mammals and birds.

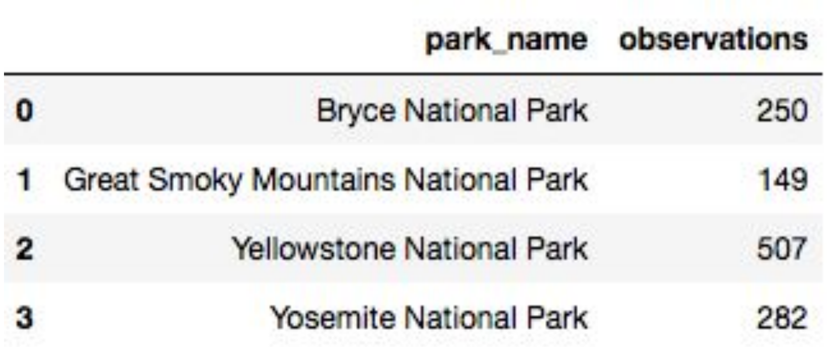
However, mammals are more likely to be endangered than reptiles. Conservationists at our parks should focus their efforts on protecting mammals over reptiles.

Sample Size Determination *foot and mouth disease in sheep*

- It is known that 15% of sheep at Bryce National Park have foot and mouth disease. At Yellowstone, park rangers want to reduce the rate of foot and mouth disease. They want to detect reductions of at least 5% points.
- First, I determined the baseline conversion rate as 15%, from the data above.
- The statistical significance was given (90%)
- The minimum detectable effect was calculated as follows:
- (100 * percentage points) / baseline = (100 * 5) / 15 = 33
- I plugged these numbers into Codecademy’s sample size calculator and found the sample size to be 890.

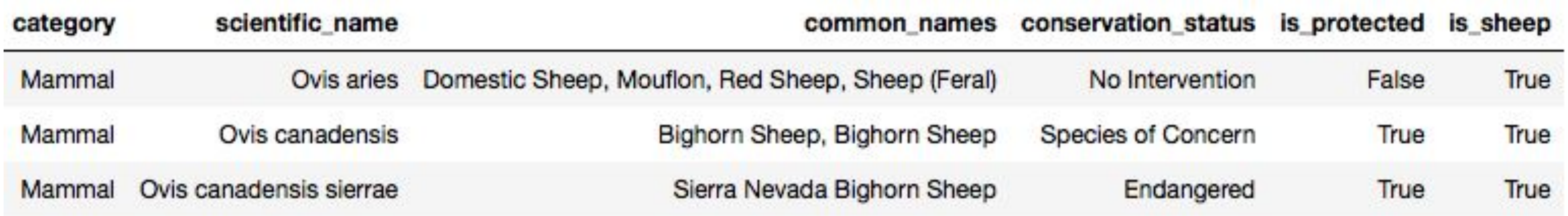
Sample Size Determination, con't

- Scientists wanted to know how many weeks they would need to observe sheep at Bryce and Yellowstone National parks.
- This could be calculated as follows:
  - Weeks observing = sample size / observations per week
- In order to make this calculation I needed to isolate the observations of sheep species, per week, at each National Park.
- I imported data from the csv observations.csv (Fig. 6), a dataset which contains data about animal sightings at different national parks. I then isolated sightings of sheep species by using a lambda function to create a new column called 'is_sheep' and selecting rows of 'species' where 'is_sheep' is True and 'category' is 'Mammal', saving it in the variable 'sheep_species'. Then I merged 'sheep_species' with 'observations' to create the DataFrame 'sheep_observations' (Fig. 7). From there, I was able to use groupby and sum to see the number of sheep observations (per week) by park (Fig. 8)



	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

Fig. 6 (left)
observations.head( )



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

Fig. 7 *sheep_species*

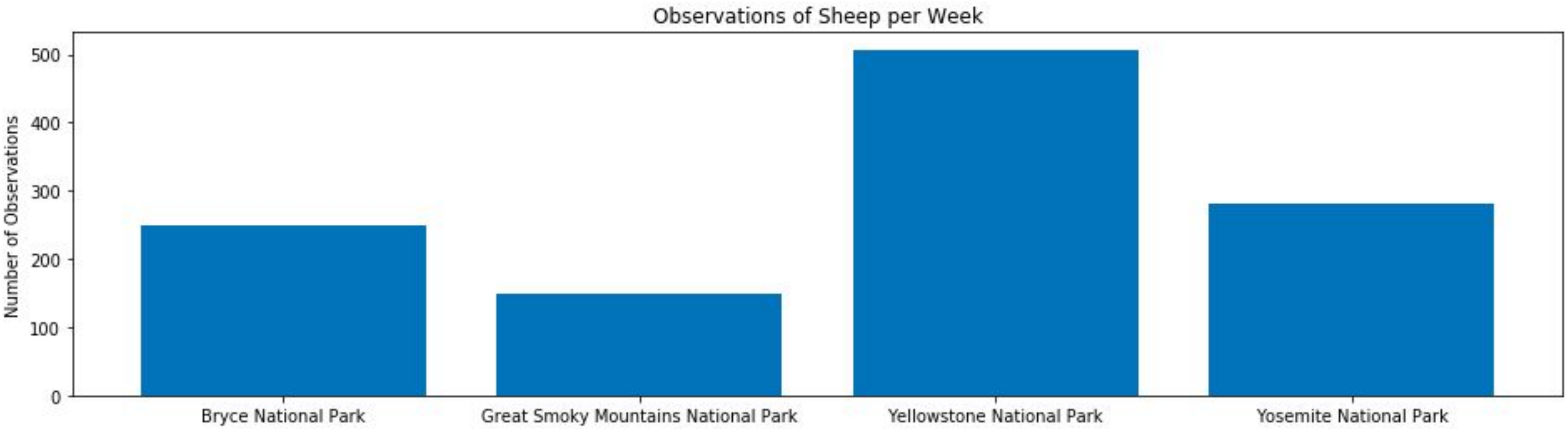

Observations of Sheep per Week
Number of Observations
0
100
200
300
400
500
Bryce National Park
Great Smoky Mountains National Park
Yellowstone National Park
Yosemite National Park

Fig. 8

Sample Size Determination, con’t

- Finally I was able to compute:
  - yellowstone_weeks_observing = sample_size_per_variant/507. = ~1.76
  - bryce_weeks_observing = sample_size_per_variant/250 = 3.56

The scientists at Yellowstone National Park would need to observe sheep for about 2 weeks to reach the required sample size.

The scientists at Bryce National Park would need to observe sheep for about 4 weeks to reach the required sample size.

Fin