Project: Capstone

Biodiversity for the National Parks

Inspecting the data (species_info.csv)

- The data from species_info.csv loaded into a DataFrame (species) contains the columns ‘category’, ‘scientific_name’, ‘common_names’, and ‘conservation_status’.
- The species DataFrame consists of 5824 rows.
- There are 5541 different (i.e. unique) species in the DF.
- These species are divided into different categories named: ‘Mammal’, 'Bird’ , 'Reptile’, 'Amphibian’, 'Fish’, 'Vascular Plant’ and 'Nonvascular Plant’.
- There are five different types of conservation statuses: ‘NaN’, ‘Species of Concern’, 'Endangered’, 'Threatened’ and 'In Recovery’.

Analyzing the data

- The different species are supposed to all fall into the different conservation statuses:

	conservation_status	scientific_name
0	Endangered	15
1	In Recovery	4
2	Species of Concern	151
3	Threatened	10

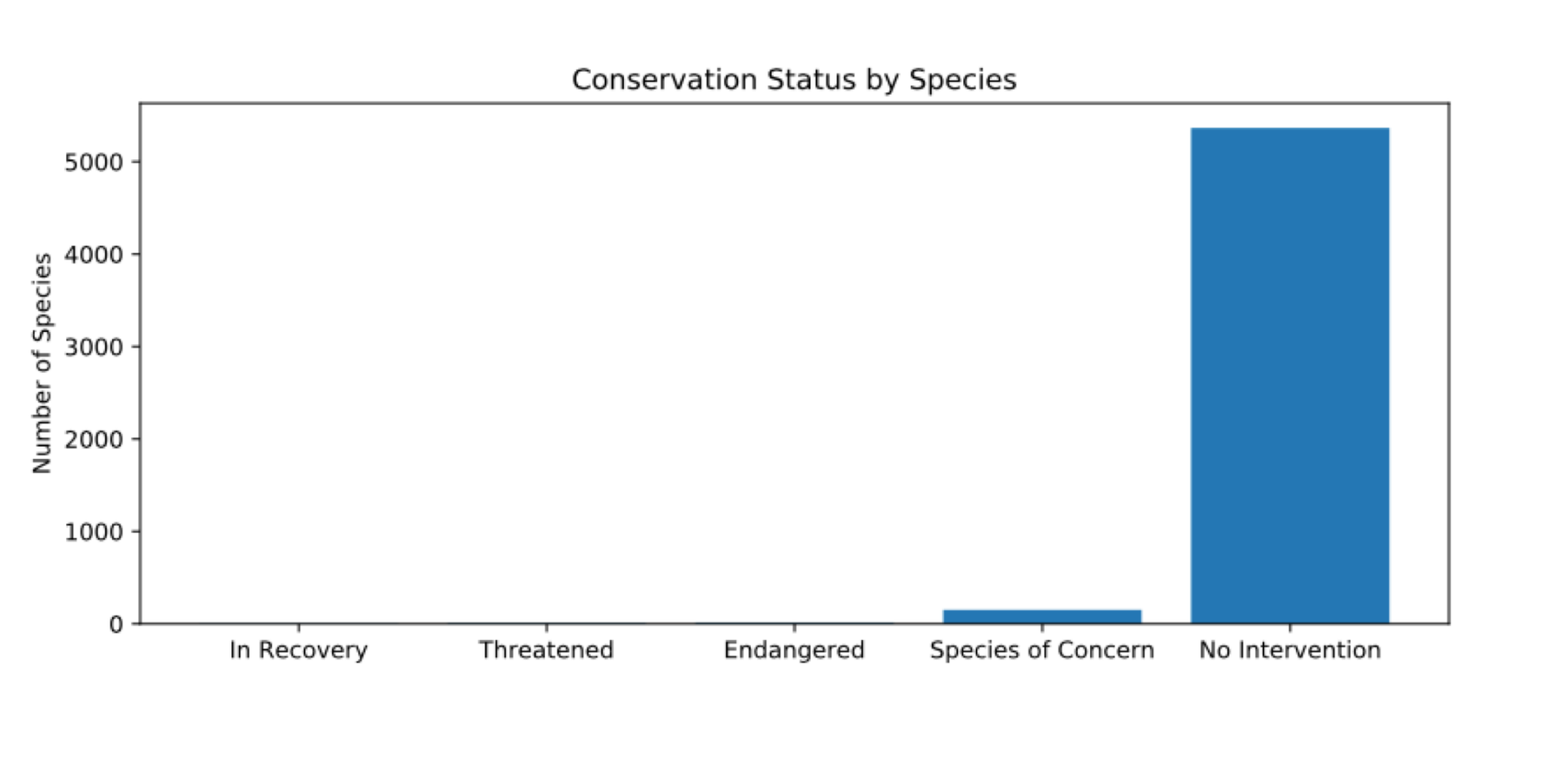
- As we can see from this table, only 180 species are categorized into the different conservation statuses. There were 5541 different species in our DF, so 5363 species don’t need some kind of protection (i.e. conservation_status is equal to NaN).

Analyzing the data

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

- We can make a visualization of the data shown above by creating a bar chart.
- The columns are sorted by the number of species in each category.
- The bar chart is shown on the next page.

Visualization of the data



Investigating Endangered Species

- Are certain types of species more likely to be endangered?
- The new column in DataFrame species called ‘is_protected’ has True values when the column ‘conservation_status’ is not equal to ‘No Intervention’ and has False values otherwise.
- The ‘True’ and ‘False’ column names are replaced with column names ‘not_protected’ and ‘protected’.
- As seen in the next table, it looks like Birds and Mammals are the most endangered species when looking at their number of protected species (75 and 30).
- With a new column called ‘percent_protected’ we can calculate the percent of protected species. A high percent means that the species concerned is more likely to be endangered as opposed to a low percent.

Investigating Endangered Species I and II

is_protected	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

is_protected	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

Chi-Squared Test for Significance

- Are certain types of species more likely to be endangered?

- We can answer this question by doing a significance test. In this test the null hypothesis is that there is no (significant) difference between our two tested species.

- The data is categorical (protected vs not_protected) and we want to compare two or more pieces of data. Therefore, we use the Chi-Squared Test. In order to run a chi-squared test, we'll need to create a contingency table. The contingency tables are shown on the next pages:

Chi-Squared Test for Significance

	protected	not-protected
Mammal	30	146
Bird	75	413

	protected	not-protected
Reptile	5	73
Mammal	30	146

- When running the chi-squared test we'll get a certain p-value. This value will tell us whether we can reject or keep the hypothesis that there is no difference between our two pieces of data.

- The p-value of 'Mammal' and 'Bird' is: 0.687594809666. So we keep the hypothesis, there is no difference between the mammal and bird species. Returning back to the original question: 'Are certain types of species more likely to be endangered?', Mammals are not more likely to be endangered than birds.

- The p-value of 'Reptile' and 'Mammal' is: 0.0383555902297. So we reject the hypothesis, there is a significant difference between the reptile and mammal species. Returning back to the original question: 'Are certain types of species more likely to be endangered?', Mammals are more likely to be endangered than reptiles

Chi-Squared Test for Significance

	protected	not-protected
Fish	11	115
Bird	75	413

	protected	not-protected
Bird	75	413
Reptile	5	73

- The p-value of ‘Fish’ and ‘Bird’ is: 0.0766819956906. So we keep the hypothesis, there is no difference between the fish and bird species. Returning back to the original question: ‘Are certain types of species more likely to be endangered?’, Birds are not more likely to be endangered than fish.

- The p-value of Bird’ and ‘Reptile’ is: 0.0531354223215. So we keep the hypothesis, there is no difference between the bird and reptile species. Returning back to the original question: ‘Are certain types of species more likely to be endangered?’, Birds are not more likely to be endangered than reptiles.

Recommendations

- Based on the significance results from the previous pages we can conclude that Mammals are the most prone species to be endangered compared to other species.
- Although almost significant ($p = 0.053$), Birds were not more endangered than Reptiles, Fish ($p = 0.077$) or Mammals ($p = 0.688$).
- Conservationists should unite and communicate the dangers for endangered species via social media with the aim of collecting sponsorship money and donations for threatened, endangered and species in recovery.
- Conservationists should organize information evenings for residents of nature reserves so that when these people witness animal cruelty they can report it via a newly introduced online forum.
- Conservationists should influence the national and local government through lobbying to protect the mammals by law.

Observations DataFrame / In search of Sheep

- The data from observations.csv loaded into a DataFrame (observations) contains the columns ‘scientific_name’, ‘park_name’, and ‘observations’.

- The merged DataFrame of Sheep and Observation is shown on the next page.

Merging Sheep and Observation DataFrames

	category	Scientific_name	Common_names	Conservation_status	Is_protected	Is_sheep	Park_name	observations
0	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True	Yellowstone National Park	126
1	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True	Great Smoky Mountains National Park	76
2	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True	Bryce National Park	119
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True	Yellowstone National Park	221

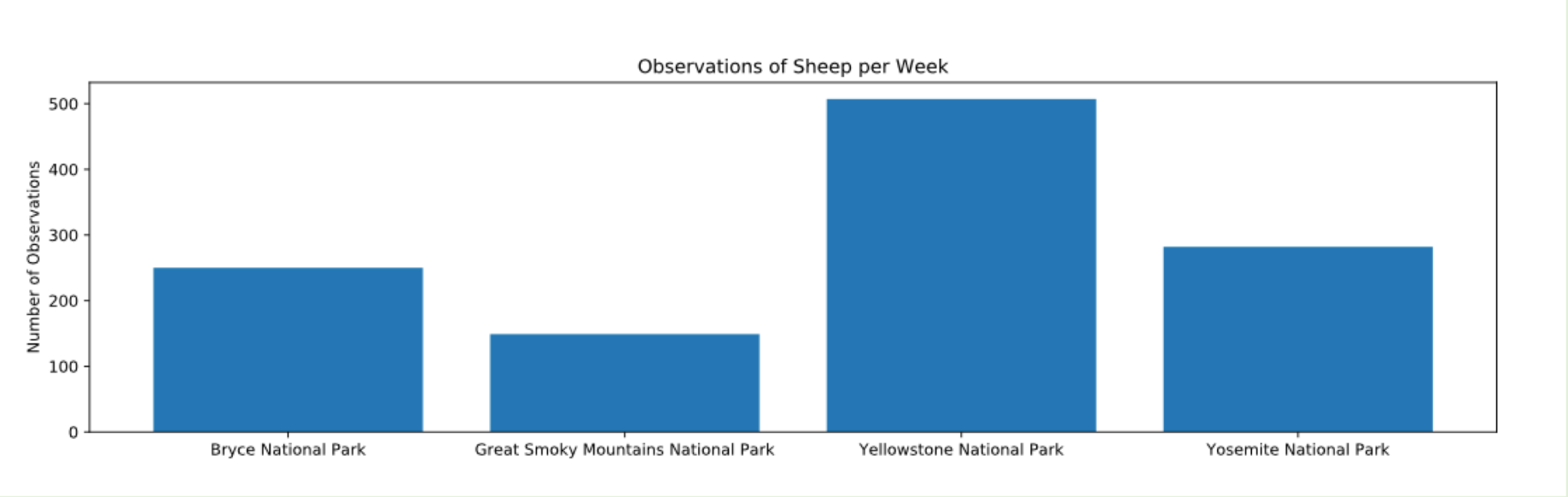
Merging Sheep and Observation DataFrames

- In the table below we see the total number of sheep observed in each park over the past 7 days.

	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

Plotting Sheep Sightings

- A bar chart can be used when we want to easily show the number of observations at each of the four national parks.



Sample Size Determination

- The baseline percentage of the sample size determination is 15%. This is based on by looking at historical data. It was said that last year 15% of sheep at Bryce National Park were recorded having foot and mouth disease.

- The definition of "Minimum Detectable Effect" is a percent of the baseline. If we wanted to observe an x% change with confidence, our minimum detectable effect would be equal to 100 * x / baseline.

- In this project our minimum detectable effect is 33% (i.e. 100*5 / 15).

- Using the sample size calculator with baseline conversion rate: 15%, statistical significance: 90% and minimum detectable effect: 33%; we calculate a sample size of 520.

- At Yellowstone National Park the scientists need to spend about 1 week to observe enough sheep
- At Bryce National Park the scientists need to spend about 2 weeks to observe enough sheep
- *At Great Smoky Mountains National Park the scientists need to spend about 3 and a half weeks to observe enough sheep
- *At Yosemite National Park the scientists need to spend about 2 weeks to observe enough sheep