Capstone Project: Biodiversity for the National Parks

By: Potoula Anagnostakos

What will be discussed

- Description of the data.
- Significance calculation for endangered status between the different categories of species.
- Sample size determination from foot and mouth disease study.
- The project was conducted in Jupyter Notebooks.

Analyzed Data

- species_info.csv: contains 5,824 rows
- observations.csv: contains 23,296 rows

Description of Data

- Below are the columns from the species_info.csv and what the data is sorted into:
  - Category
  - Scientific_name
  - Common_names
  - Conservation_status

Description of Data

- Below are the columns from the species_info.csv and what the data is sorted into:
  - Category
    - Total of 7 categories: ‘Mammal’, ‘Reptile’, ‘Bird’, ‘Fish’, ‘Vascular Plant’, ‘Nonvascular Plant’, and ‘Amphibian’
  - Scientific_name
  - Common_names
  - Conservation_status

Description of Data

- Below are the columns from the species_info.csv and what the data is sorted into:
  - Category
  - Scientific_name
    - Total of 5,541 names
  - Common_names
  - Conservation_status

Description of Data

- Below are the columns from the species_info.csv and what the data is sorted into:
  - Category
  - Scientific_name
  - Common_names
    - Total of 5,504 unique common_names
  - Conservation_status

Description of Data

- Below are the columns from the species_info.csv and what the data is sorted into:
  - Category
  - Scientific_name
  - Common_names
  - Conservation_status
    - Total of 5 statuses: ‘Endangered’, ‘Species of Concern’, ‘Threatened’, ‘Null’, and ‘In Recovery’

Description of Data

- Below are the columns from the observations.csv and what the data is sorted into:
  - Scientific_name
  - Park_name
  - Observations

Description of Data

- Below are the columns from the observations.csv and what the data is sorted into:
  - Scientific_name
    - Total of 5,541
  - Park_name
  - Observations

Description of Data

- Below are the columns from the observations.csv and what the data is sorted into:
  - Scientific_name
  - Park_name
    - Total of 4 national parks: ‘Yellowstone National Park’, ‘Yosemite National Park’, ‘Great Smoky Mountains National Park’, and ‘Bryce National Park’
  - Observations

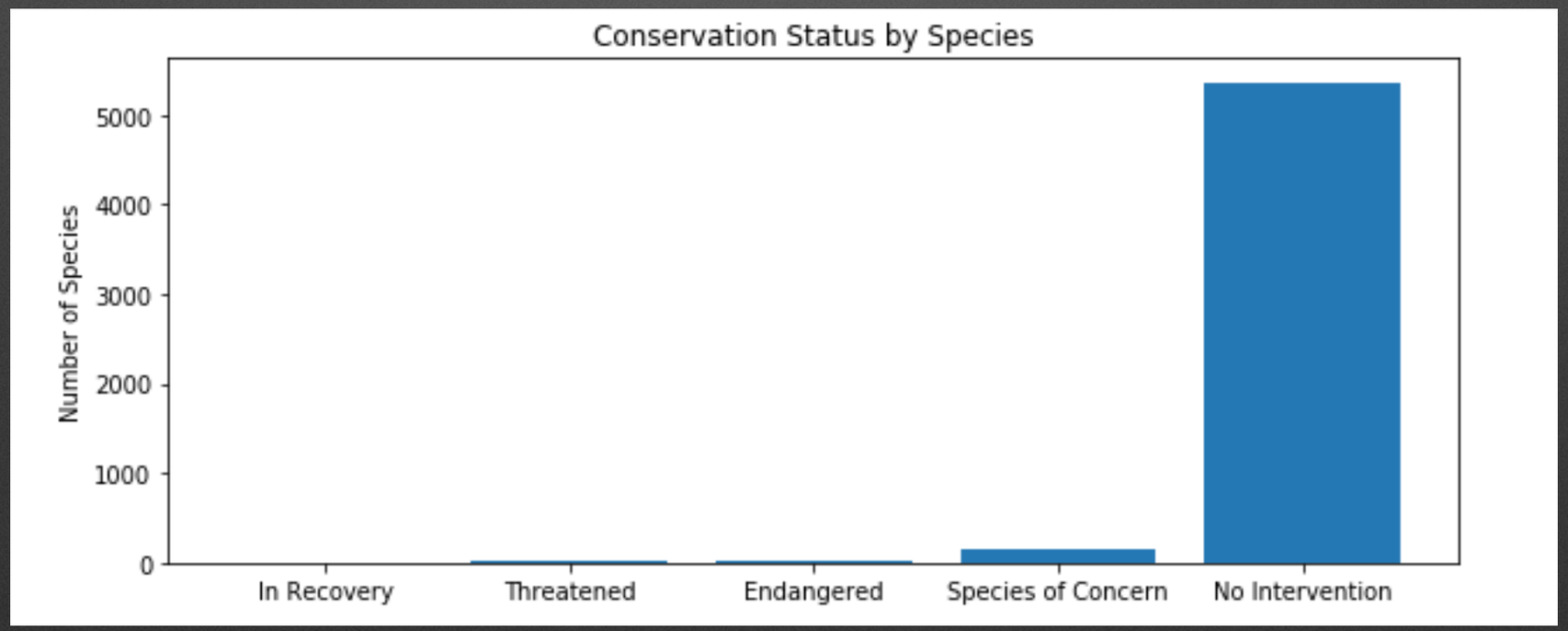
Description of Data

- Below are the columns from the observations.csv and what the data is sorted into:
    - Scientific_name
    - Park_name
    - Observations
        - Refers to total number of animals observed

Calculated Significance

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

- A [groupby] command (shown to the right) was used to calculate the number of species under each category of conservation_status.
- Using the [plt.bar] command, the results were inserted into a bar graph, shown below.



Examine Categories

Question: Are certain types more likely to be endangered?

- In [pandas], I used the [pivot] command to sort the data which species are and aren't protected.

- Columns = is_protected
- Index = category
- Values = scientific_name

Before [pivot]

	category	is_protected	scientific_name
0	Amphibian	False	72
1	Amphibian	True	7
2	Bird	False	413
3	Bird	True	75
4	Fish	False	115

After [pivot]

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

‘Mammal’ and ‘Bird’ Significance Test

	category	not_protected	protected	percent_protected
1	Bird	413	75	0.153689
3	Mammal	146	30	0.170455

- In comparison, ‘Mammal’ has a higher rate of protection than ‘Bird’. The two pieces of data are categorical and were tested through a Chi Contingency to test the significance.

Below are the results:

Contingency = [[30, 146],
[75, 413]]

→

(0.161701483165457,
0.6875948096661336, 1,
array([[27.8313253, 148.1686747],
[77.1686747, 410.8313253]]))

Conclusion:

P value is greater than 5%

Not reject hypothesis null

Difference: not significant

‘Mammal’ and ‘Reptile’ Significance Test

	category	not_protected	protected	percent_protected
3	Mammal	146	30	0.170455
5	Reptile	73	5	0.064103

- Same testing was done with ‘Mammal’ and ‘Reptile’. ‘Mammal’ has a higher protection rate than ‘Reptile’. These two pieces of data were compared and are also categorical.

Below are the results:

Contingency = [[30, 146], [5, 73]]

(4.2891830962036446,
0.03835590229698977,
1, array([[24.2519685, 151.7480315],
[10.7480315, 67.2519685]]))

Conclusion:

P value is less than 5%

Reject hypothesis null

Difference: significant

Study: Sheep Sightings

For a week, conservationists recorded sightings of different species at several national parks. In this study, sheep sightings were the main focus. The data aggregated was sorted into the chart below, which was done in [pandas]. String function was used to search for rows of ‘species’ where ‘is_sheep’ is [true] in order to only reflect the sightings of sheep in the parks.

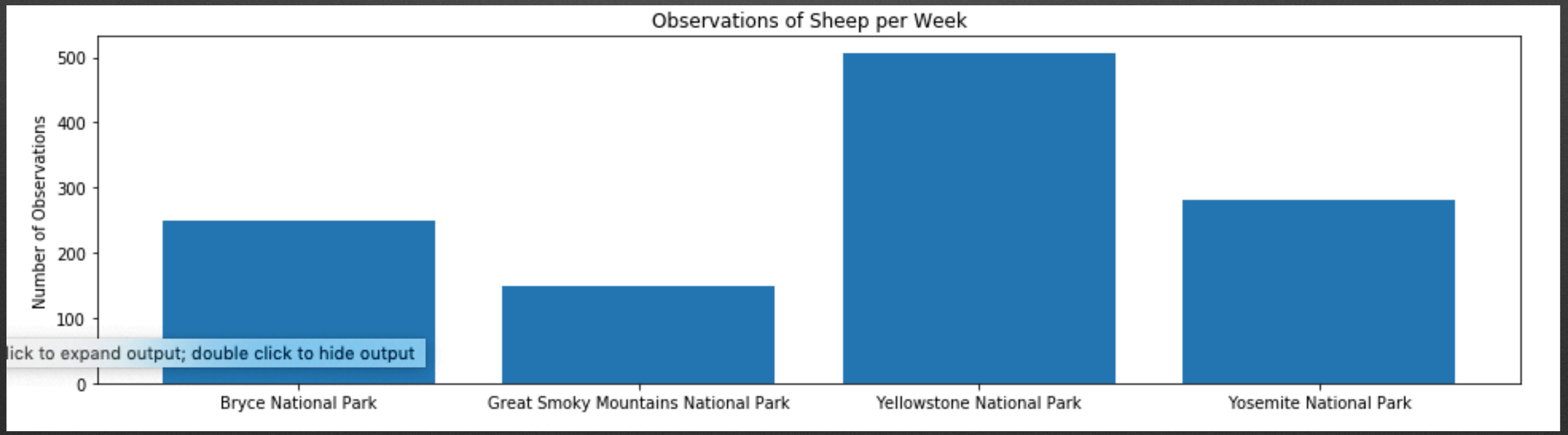
	scientific_name	park_name	observations	category	common_names	conservation_status	is_protected	is_sheep
0	Ovis canadensis	Yellowstone National Park	219	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True
1	Ovis canadensis	Bryce National Park	109	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True
2	Ovis canadensis	Yosemite National Park	117	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True
3	Ovis canadensis	Great Smoky Mountains National Park	48	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True
4	Ovis canadensis sierrae	Yellowstone National Park	67	Mammal	Sierra Nevada Bighorn Sheep	Endangered	True	True
5	Ovis canadensis sierrae	Yosemite National Park	39	Mammal	Sierra Nevada Bighorn Sheep	Endangered	True	True
6	Ovis canadensis sierrae	Bryce National Park	22	Mammal	Sierra Nevada Bighorn Sheep	Endangered	True	True
7	Ovis canadensis sierrae	Great Smoky Mountains National Park	25	Mammal	Sierra Nevada Bighorn Sheep	Endangered	True	True
8	Ovis aries	Yosemite National Park	126	Mammal	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True
9	Ovis aries	Great Smoky Mountains National Park	76	Mammal	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True
10	Ovis aries	Bryce National Park	119	Mammal	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True
11	Ovis aries	Yellowstone National Park	221	Mammal	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True

Observations of Sheep Per Week Across National Parks

- The chart to the right reflects the number of observations per national park. The [groupby] function was used in order to get the ‘sum’ of all ‘observations’ for each park_name.

	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

- The data was then put into a bar graph, shown below.



Foot and Mouth Disease

- Scientists found that 15% of sheep at Bryce National Park have foot and mouth disease. There is a program at Yellowstone National Park that reduces the rate of foot and mouth disease.

- The scientists wanted find out if this program is working or not. Ideally, they wanted to be able to detect reductions of ***at least*** 5%.

- The level of significance is at 90%.

- In order to calculate the number of sheep that scientists would need to observe from each park, the “Minimum Detectable Effect” (percent of the baseline) was found.

    - Minimum_detectable_effect: 33.33%
    - Baseline: 15%
    - Statistical_significance: 90%
        - Sample size per variation: 870

- In conclusion, approximately 3.5 weeks would need to be spent at Bryce National Park and 1.5 weeks would need to be spent at Yellowstone National Park to observe enough samples.

What should the conservationists do?

1. Observe the species that have the closest protection rates, such as ‘Mammals’ and ‘Birds’. (17% and 15%, respectively).
2. Make sure to calculate their significance. If there isn’t much of a significant difference, that it is safe to say that those species are in the same status of conservation.
3. Make sure to calculate the right sample size when looking to start the program for reducing the rate of foot and mouth disease.
4. The species that are in the ‘threatened’ and ‘in recovery’ categories should be looked at first before the other categories as they are the highest priority.


“Cherish the natural world because you’re part of it
and you depend on it.”
–Sir Richard Attenborough