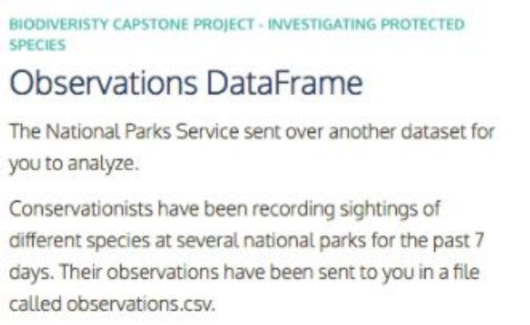
Introduction to Data Analysis
David Schrager

Capstone #2:
Biodiversity for the National Parks

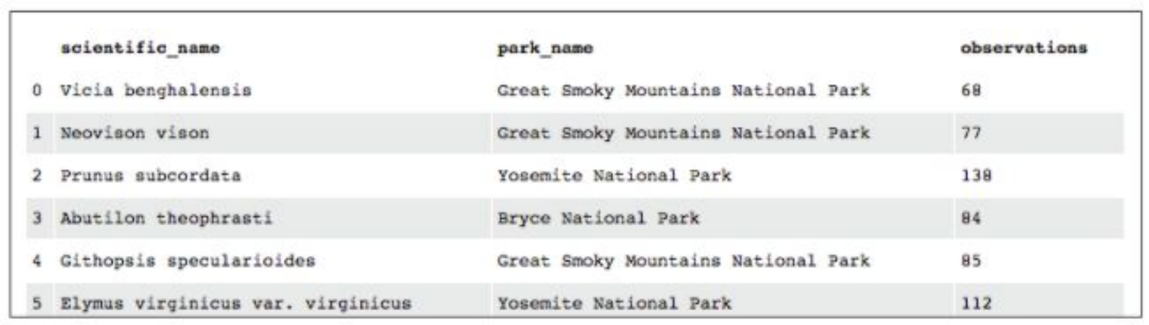


BIODIVERISTY CAPSTONE PROJECT - INVESTIGATING PROTECTED SPECIES

Observations DataFrame

The National Parks Service sent over another dataset for you to analyze.

Conservationists have been recording sightings of different species at several national parks for the past 7 days. Their observations have been sent to you in a file called observations.csv.

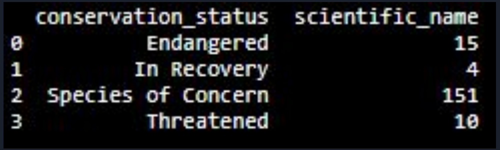


	scientific_name	park_name	observations
0	Vicia benghalensis	Great Smoky Mountains National Park	68
1	Neovison vison	Great Smoky Mountains National Park	77
2	Prunus subcordata	Yosemite National Park	138
3	Abutilon theophrasti	Bryce National Park	84
4	Githopsis specularioides	Great Smoky Mountains National Park	85
5	Elymus virginicus var. virginicus	Yosemite National Park	112

This capstone was an obvious choice for me as I have a keen interest in biodiversity and endangered species. I travelled to Africa two years ago and was fortunate enough to view the big five in their most natural habitats. It's evident that climate change is impacting our wildlife and species around the globe in ways never seen before, and unfortunately many species will continue progressing onto endangered and extinct lists. This exercise helped to highlight some of the ways in which we can pull key data points and disseminate these groupings of data

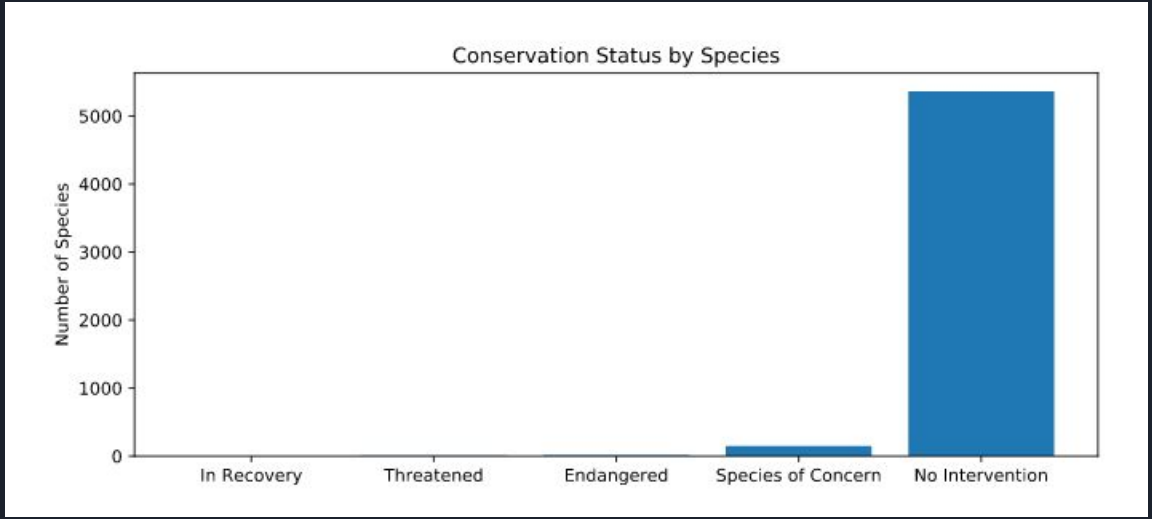
Informationally, we were able to determine the following: scientific name, conservation status, and common name of each species. On the analysis side, we were able to clearly present data in our designed pivots and bar graphs, and group names, statuses, and counts in presentable formats

Species_info.csv - Data



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

The bar graph below highlights the number of species in each conservation status group. ‘No Intervention’ was fortunately the largest by 35x compared to ‘Species of Concern’. Four species are currently in recovery, according to our data



Conservationists concerned about endangered species - a recommendation

Are certain types of species more likely to be endangered? Based on our significance calculations:

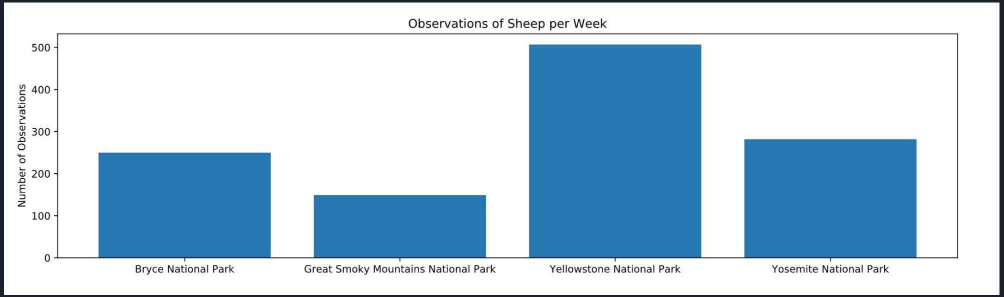
- Not significant but certainly a measurable difference between birds and mammals in the protected category
- Conclusion: Certain species are more likely, according to our data, to fall under the endangered status
- Recommendation: Prioritization should follow species type when considering protections based on conservation status

Status types, as previously highlighted:

- Species of Concern
- Threatened
- Endangered
- In Recovery
- No Intervention

Foot and mouth disease study sample size determination

- Baseline conversion rate: 15%
- Statistical significance: 90%
- Minimum detectable effect: 20%
- Sample size: 12000



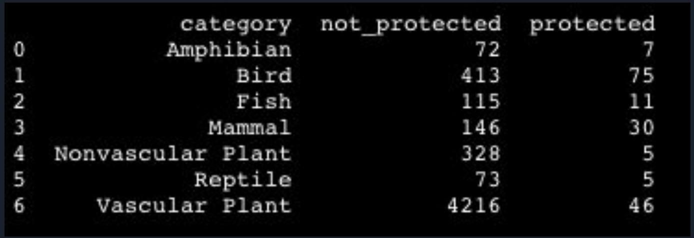
If the scientists wanted to be sure that a > 5% drop in observed cases of foot and mouth disease in the sheep at Yellowstone was significant they would have to observe at least 510 sheep

- Weeks required to see 510 sheep to test for > 5% drop in cases of foot and mouth disease:
- One week of observing in Yellowstone National Park and approximately two weeks in Bryce National Park

Significance calculations
For endangered status between different categories of species

Chi-Squared Test for Significance

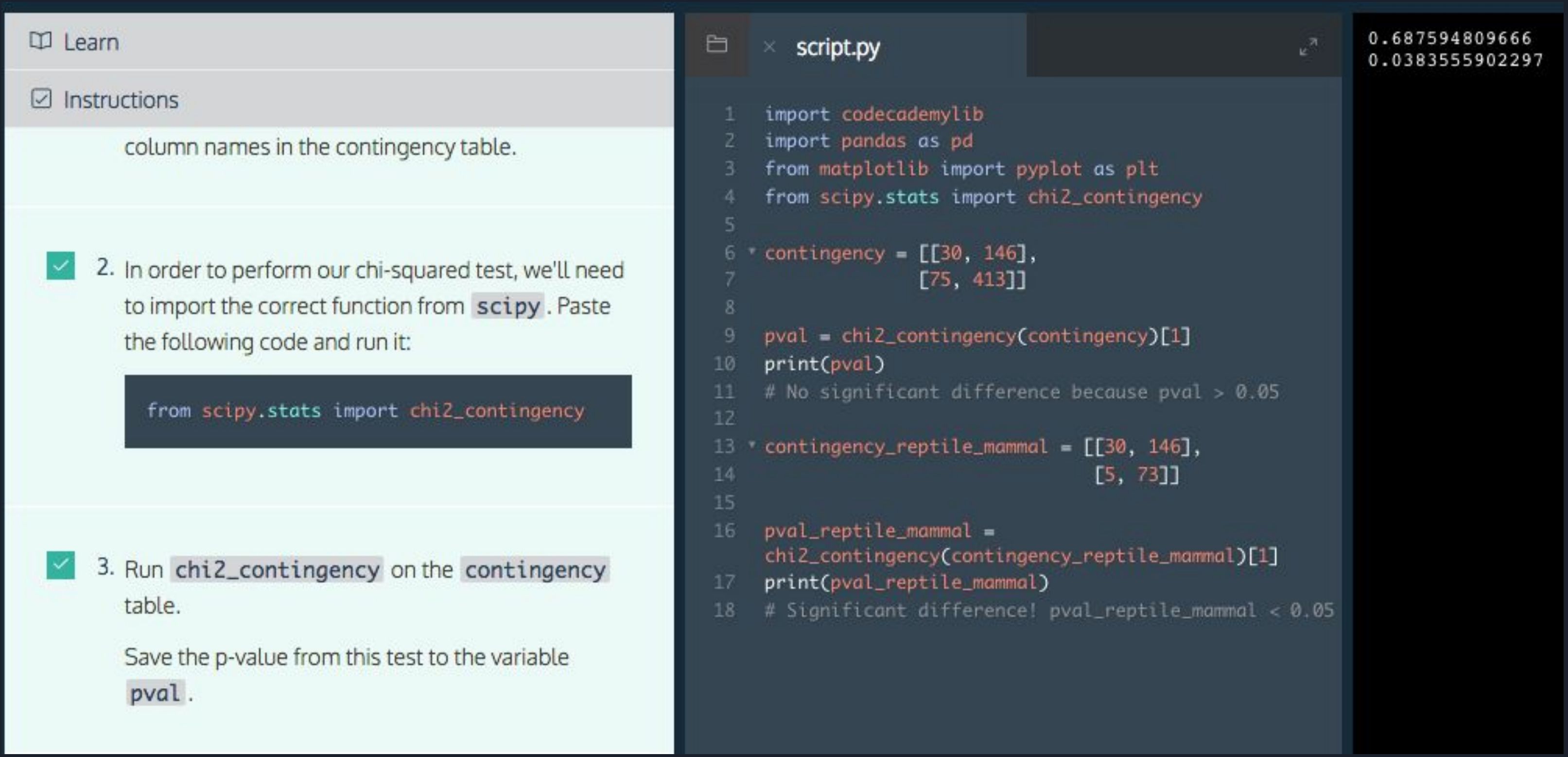
- ○ Contingency (protected birds and mammals)
  - ■ pval = 0.6875948 *(~0.688)*
  - ■ The difference between the percentages is not significant because the pval > 0.05
- ○ Protected Reptiles and Mammals
  - ■ pval_reptile_mamma = 0.03835559 *(~0.038)*
  - ■ The difference between the percentages is significant because pval_reptile_mammal < 0.05



```
            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
```

Park Name	Observations
Bryce National Park	250
Great Smoky Mountains National Park	149
Yellowstone National Park	507
Yosemite National Park	282

Scientific Name	Park Name	Observations	Category	Common Names	Conservation Status	Is Protected?	Is Sheep?
Ovis canadensis	Yellowstone National Park	219	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	TRUE	TRUE
Ovis canadensis	Bryce National Park	109	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	TRUE	TRUE
Ovis canadensis	Yosemite National Park	117	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	TRUE	TRUE
Ovis canadensis	Great Smoky Mountains National Park	48	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	TRUE	TRUE
Ovis canadensis sierrae	Yellowstone National Park	67	Mammal	Sierra Nevada Bighorn Sheep	Endangered	TRUE	TRUE



column names in the contingency table.

☑ 2. In order to perform our chi-squared test, we'll need to import the correct function from `scipy`. Paste the following code and run it:

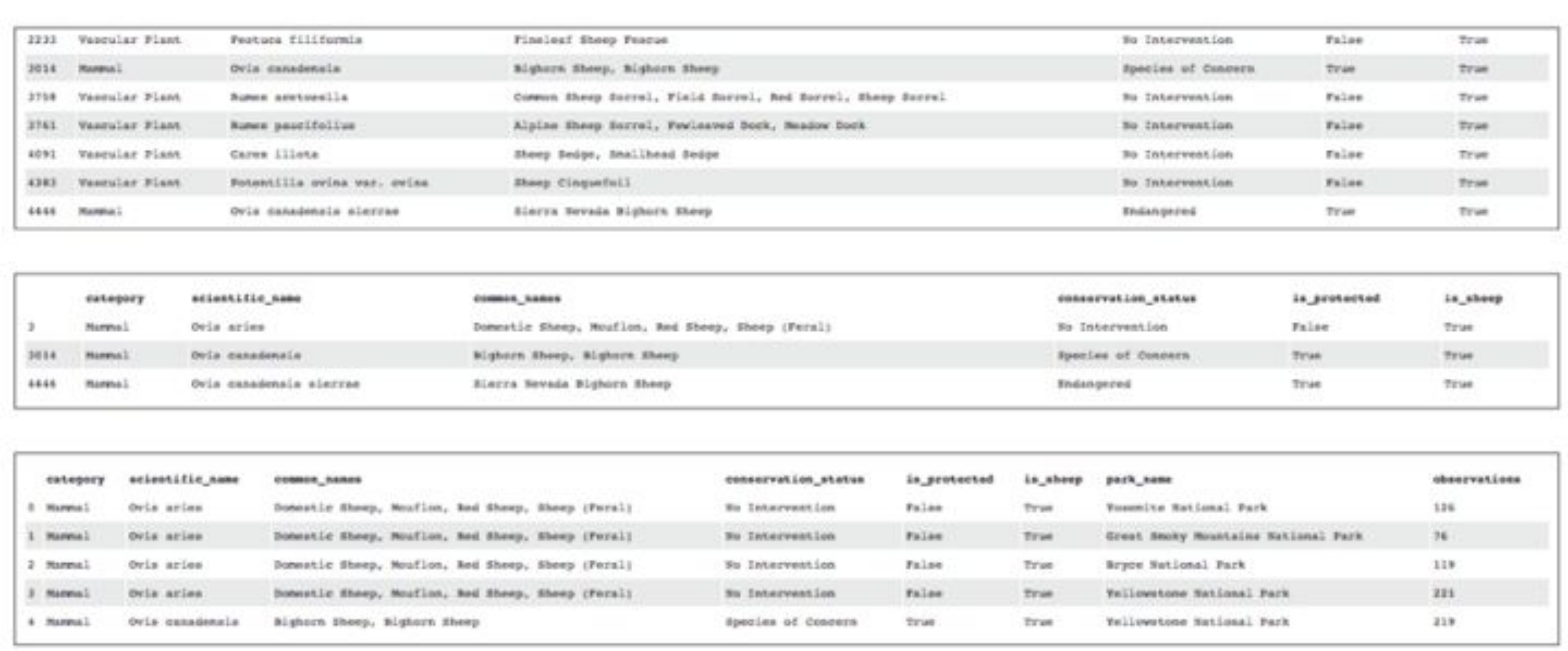
```
from scipy.stats import chi2_contingency
```

☑ 3. Run `chi2_contingency` on the `contingency` table.

Save the p-value from this test to the variable `pval`.

```
import codecademylib
import pandas as pd
from matplotlib import pyplot as plt
from scipy.stats import chi2_contingency

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

pval = chi2_contingency(contingency)[1]
print(pval)
# No significant difference because pval > 0.05

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

pval_reptile_mammal = chi2_contingency(contingency_reptile_mammal)[1]
print(pval_reptile_mammal)
# Significant difference! pval_reptile_mammal < 0.05
```

```
0.687594809666
0.0383555902297
```



2233	Vascular Plant	Festuca filiformis	Fineleaf Sheep Fescue	No Intervention	False	True
3014	Mammal	Ovis canadensis	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True
3758	Vascular Plant	Rumex acetosella	Common Sheep Sorrel, Field Sorrel, Red Sorrel, Sheep Sorrel	No Intervention	False	True
3761	Vascular Plant	Rumex paucifolius	Alpine Sheep Sorrel, Fewleaved Dock, Meadow Dock	No Intervention	False	True
4091	Vascular Plant	Carex illota	Sheep Sedge, Smallhead Sedge	No Intervention	False	True
4383	Vascular Plant	Potentilla ovina var. ovina	Sheep Cinquefoil	No Intervention	False	True
4446	Mammal	Ovis canadensis sierrae	Sierra Nevada Bighorn Sheep	Endangered	True	True

	category	scientific_name	common_names	conservation_status	is_protected	is_sheep
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True
3014	Mammal	Ovis canadensis	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True
4446	Mammal	Ovis canadensis sierrae	Sierra Nevada Bighorn Sheep	Endangered	True	True

	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	Yosemite 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
4	Mammal	Ovis canadensis	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True	Yellowstone National Park	219