Biodiversity for The National Parks

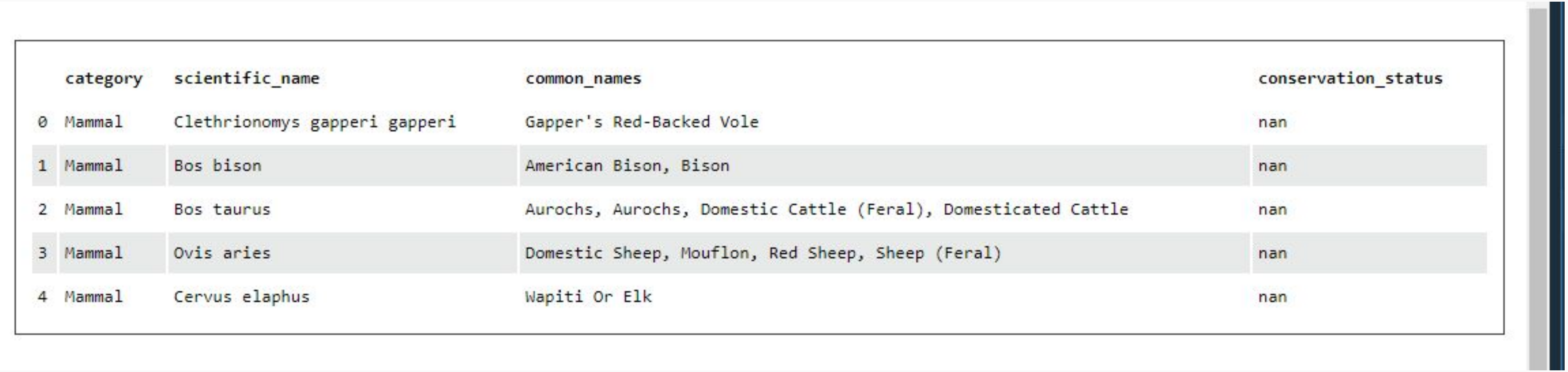
Data Analysis Capstone Project
Brendan Dangelo

Part I

Investigating Protected Species

Initial Data

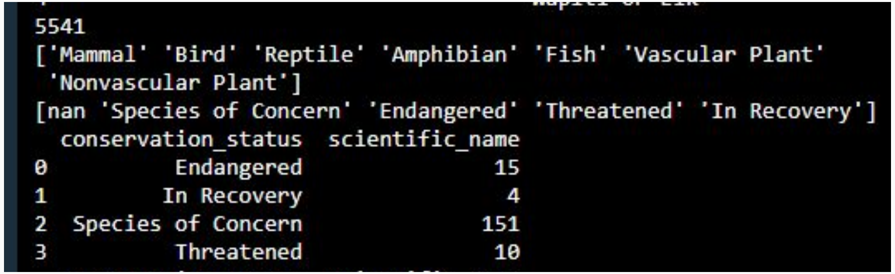
Our initial data included a list of species and their conservation statuses from a variety of US National Parks. The data was often missing in parts and needed to be organized and utilized to answer questions.



	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), Domesticated Cattle	nan
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	nan
4	Mammal	Cervus elaphus	Wapiti Or Elk	nan

Initial Findings

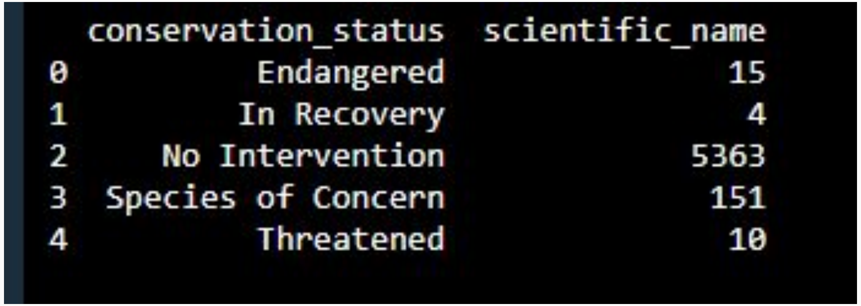
The data included 5541 species, divided into six categories, ‘Mammal’, ‘Bird’, ‘Reptile’, ‘Amphibian’, ‘Fish’, ‘Vascular Plant’ and ‘Nonvascular Plant’. Further, these were divided into Conservation Statuses, including Endangered, In Recovery, Species of Concern and Threatened



```
5541
['Mammal' 'Bird' 'Reptile' 'Amphibian' 'Fish' 'Vascular Plant'
 'Nonvascular Plant']
[nan 'Species of Concern' 'Endangered' 'Threatened' 'In Recovery']
  conservation_status  scientific_name
0          Endangered               15
1         In Recovery                4
2  Species of Concern              151
3          Threatened               10
```

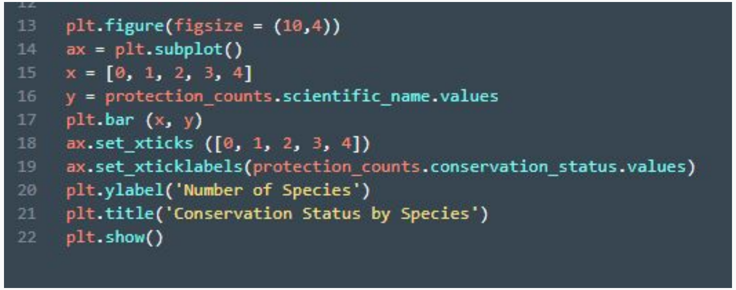
Analyzing Conservation Statuses

In looking at our data, it was shown that our species numbers did not match the amount of species categorized by Conservation Status. We re-organized our data to add the Conservation Status ‘No Intervention’ to show species that did not fall into on of the previous statuses.

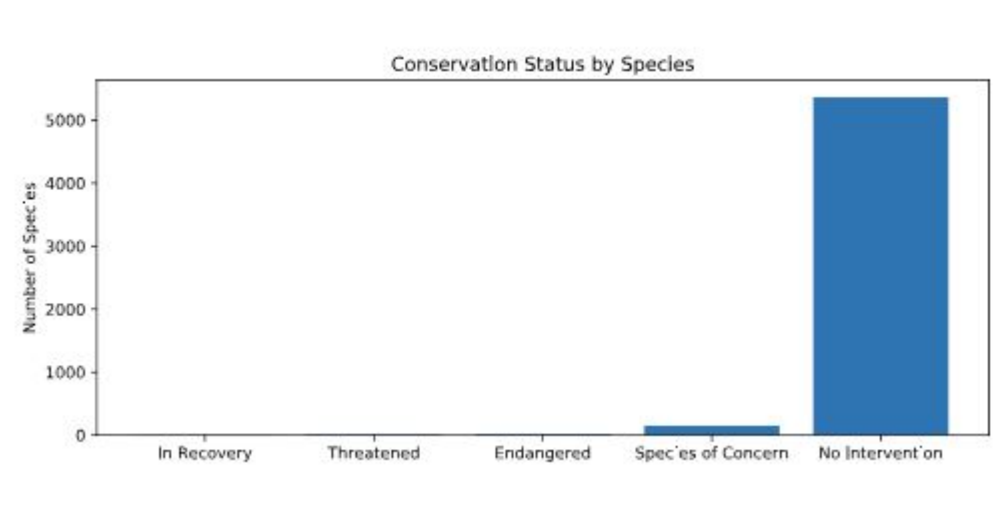


```
   conservation_status  scientific_name
0           Endangered               15
1          In Recovery                4
2      No Intervention             5363
3   Species of Concern              151
4           Threatened               10
```

Conservation Status Bar Graph



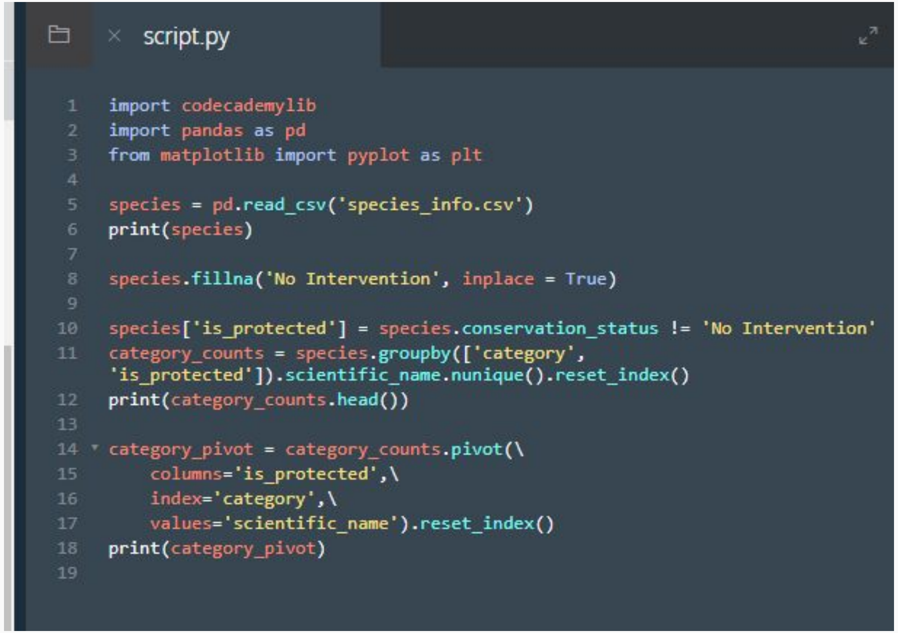
```
plt.figure(figsize = (10,4))
ax = plt.subplot()
x = [0, 1, 2, 3, 4]
y = protection_counts.scientific_name.values
plt.bar (x, y)
ax.set_xticks ([0, 1, 2, 3, 4])
ax.set_xticklabels(protection_counts.conservation_status.values)
plt.ylabel('Number of Species')
plt.title('Conservation Status by Species')
plt.show()
```



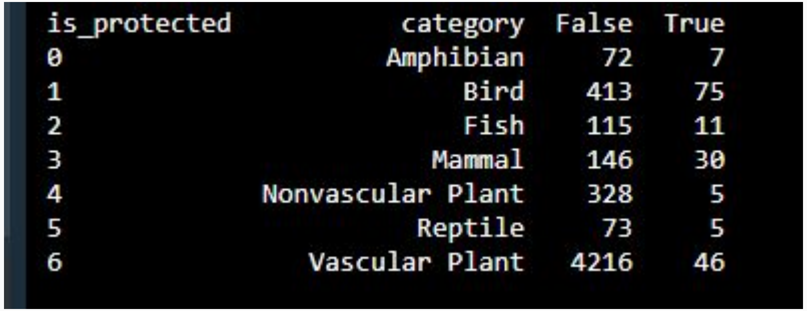
Answering the Question:

Are certain types of species more likely to be endangered?

Our data was reorganized to compare protected or non-protected status by species to help answer our question. We took our data and groupedby category and protected status.



```python
import codecademylib
import pandas as pd
from matplotlib import pyplot as plt

species = pd.read_csv('species_info.csv')
print(species)

species.fillna('No Intervention', inplace = True)

species['is_protected'] = species.conservation_status != 'No Intervention'
category_counts = species.groupby(['category',
'is_protected']).scientific_name.nunique().reset_index()
print(category_counts.head())

category_pivot = category_counts.pivot(\
    columns='is_protected',\
    index='category',\
    values='scientific_name').reset_index()
print(category_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
```

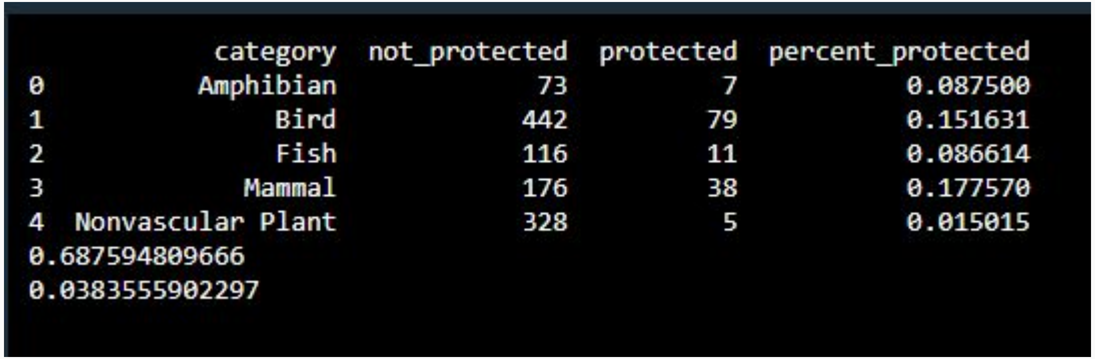
Answering the Question:

Are certain types of species more likely to be endangered?

Next we performed a Chi-Squared Test to Test for significance within our data. Our null hypothesis was to test to see if the difference between our percents protected data was due to chance.

We ran a Chi-Squared Test comparing our data on Birds vs. Mammals, then on Reptiles vs. Mammals.

Test 1 showed a pval of ~.68, showing the difference was not significant and was a result of chance.



```
             category  not_protected  protected  percent_protected
0           Amphibian             73          7           0.087500
1                Bird            442         79           0.151631
2                Fish            116         11           0.086614
3              Mammal            176         38           0.177570
4   Nonvascular Plant            328          5           0.015015
0.687594809666
0.0383555902297
```

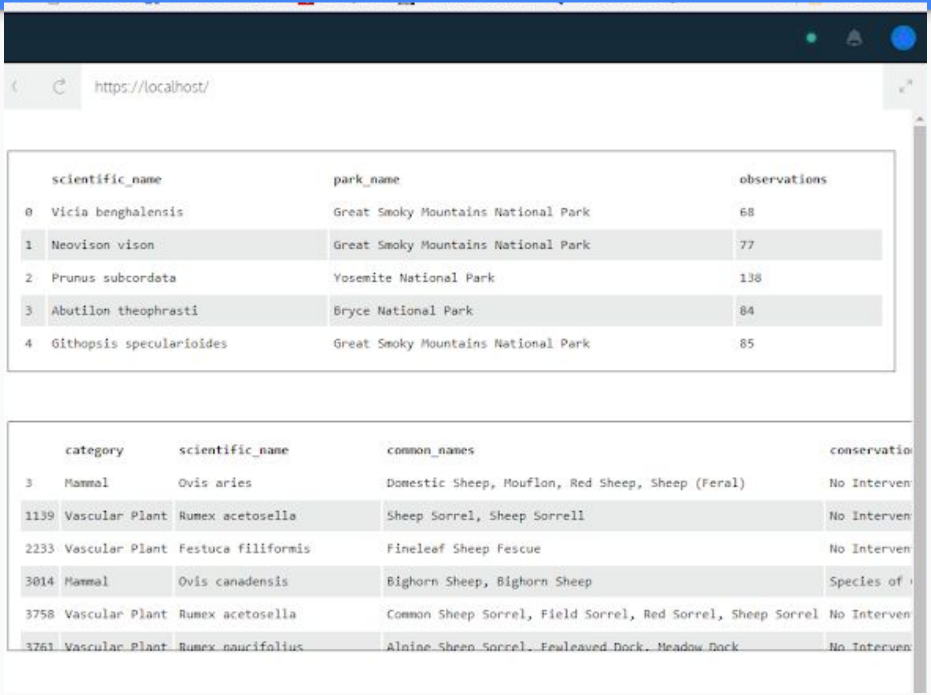
~.038

Our second test showed a pval of ~.038, which is significant. Our conclusion is thus: Certain types of species are more susceptible to be endangered than others.

Part 2:
Observing Sheep in the National Parks

Our second data set included information regarding number of sheep observed in National Parks and taking and analyzing that data.

Our first step was to take two Data Frames and combine them, while using a lambda function to find the sheep species.

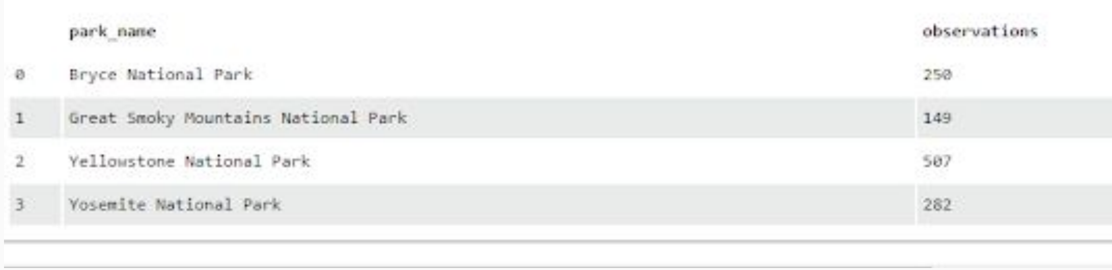


	scientific_name	park_name	observations
0	Vicia benghalensis	Great Smoky Mountains National Park	68
1	Neovison vison	Great Smoky Mountains National Park	77
2	Prunus subcordata	Yosemite National Park	138
3	Abutilon theophrasti	Bryce National Park	84
4	Githopsis specularioides	Great Smoky Mountains National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

	category	scientific_name	common_names	conservatio
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Interven
1139	Vascular Plant	Rumex acetosella	Sheep Sorrel, Sheep Sorrell	No Interven
2233	Vascular Plant	Festuca filiformis	Fineleaf Sheep Fescue	No Interven
3014	Mammal	Ovis canadensis	Bighorn Sheep, Bighorn Sheep	Species of
3758	Vascular Plant	Rumex acetosella	Common Sheep Sorrel, Field Sorrel, Red Sorrel, Sheep Sorrel	No Interven
3761	Vascular Plant	Rumex paucifolius	Alpine Sheep Sorrel, Fewleaved Dock, Meadow Dock	No Interven

Groupby National Park

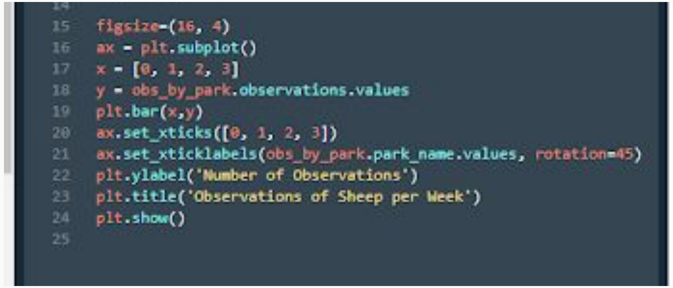
Next, we took our data and combined it to Groupby the National Park to find the number of observed sheep.



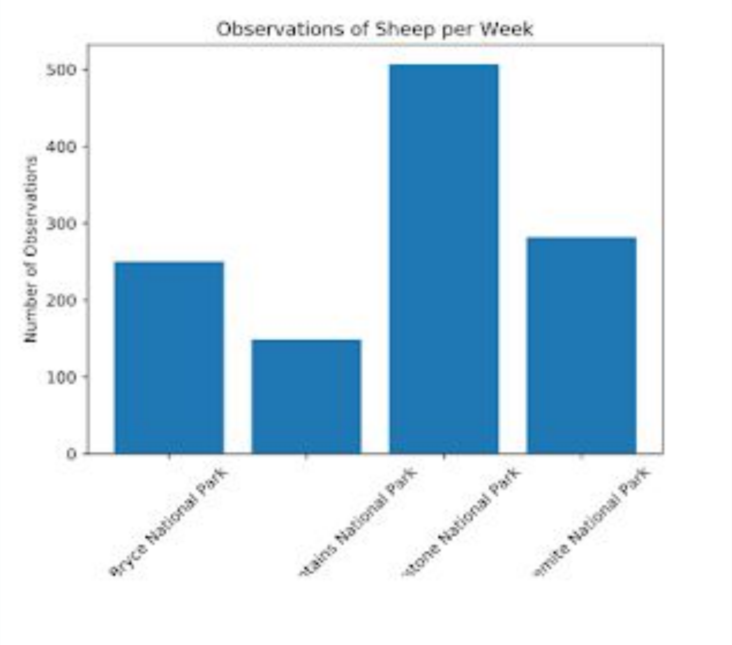
	park_name	observations
0	Bryce National Park	250
1	Great Smoky Mountains National Park	149
2	Yellowstone National Park	507
3	Yosemite National Park	282

Sheep Per Week

We used the following code to plot the bar graph on the left, showing the number of sheep observed per park for a week.



```
figsize=(16, 4)
ax = plt.subplot()
x = [0, 1, 2, 3]
y = obs_by_park.observations.values
plt.bar(x,y)
ax.set_xticks([0, 1, 2, 3])
ax.set_xticklabels(obs_by_park.park_name.values, rotation=45)
plt.ylabel('Number of Observations')
plt.title('Observations of Sheep per Week')
plt.show()
```

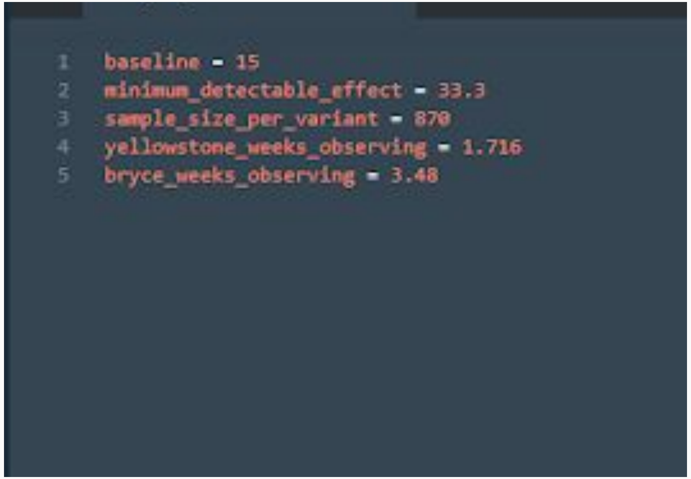


Foot and Mouth Reduction

Our final task was to create a sample size to help understand Foot and Mouth disease reductions in the parks.

Our calculations are found on the left.

We found that scientists would need to observe at least 510 sheep for a sample size. This would take approximately one week of observing in Yellowstone and two weeks in Bryce



```
baseline = 15
minimum_detectable_effect = 33.3
sample_size_per_variant = 870
yellowstone_weeks_observing = 1.716
bryce_weeks_observing = 3.48
```

Recommendations

For scientists working with endangered species, we suggest offering a heightened awareness for certain species over others.

This will lead to careful consideration of conservation efforts.