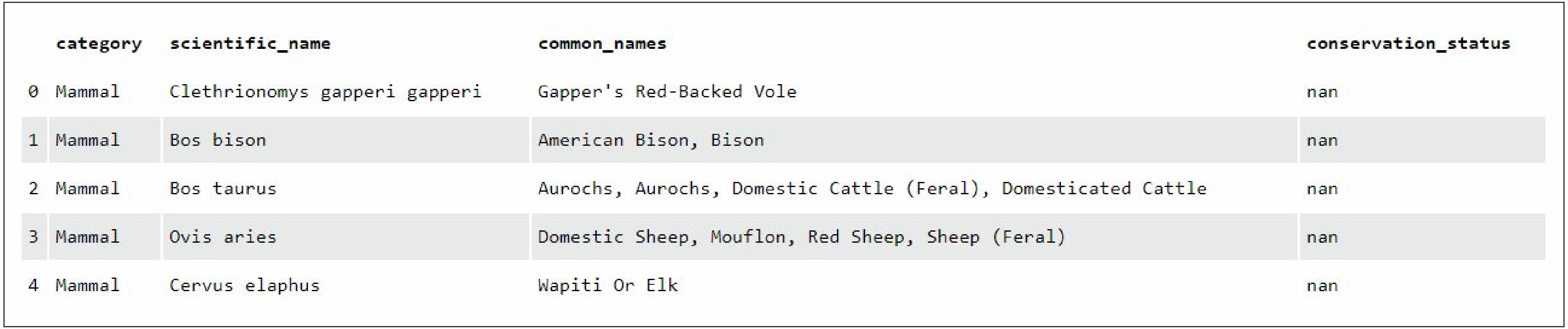
Capstone Option 2: Biodiversity for the National Parks

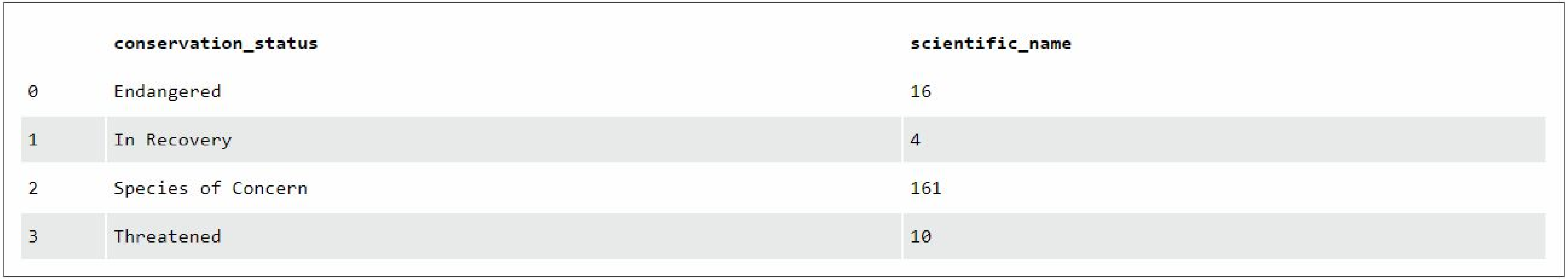
Glimpse at ‘species_info.csv’

Animals are segmented according to their category, scientific name, common name and conservation status.



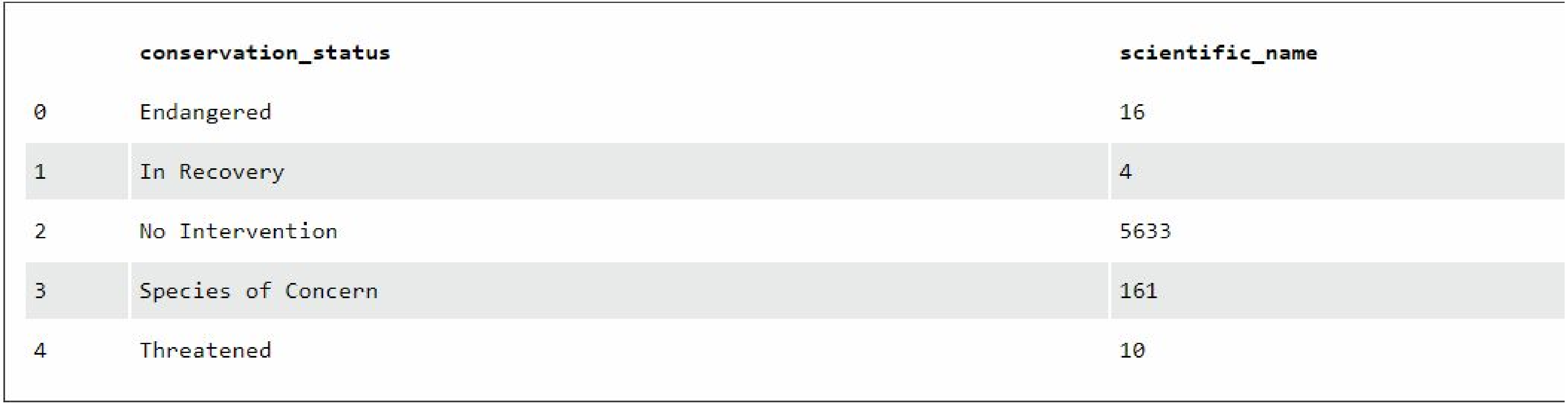
	category	scientific_name	common_names	conservation_status
0	Mammal	Clethrionomys gapperi gapperi	Gapper's Red-Backed Vole	nan
1	Mammal	Bos bison	American Bison, Bison	nan
2	Mammal	Bos taurus	Aurochs, Aurochs, Domestic Cattle (Feral), Domesticated Cattle	nan
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	nan
4	Mammal	Cervus elaphus	Wapiti Or Elk	nan

There are 4 categories in the ‘conservation_status’ column, which are ‘Endangered’, ‘In recovery’, ‘Species of Concern’, ‘Threatened’. But we have 5824 species, where did the other species go? Actually, ‘nan’, which means ‘no intervention’, is not included,



	conservation_status	scientific_name
0	Endangered	16
1	In Recovery	4
2	Species of Concern	161
3	Threatened	10

For better visualization, we reorganise the ‘conservation status’ table. Instead of ‘nan’, we use ‘no intervention’ to mark those animals which are not intervened. So the data comes out like table below.

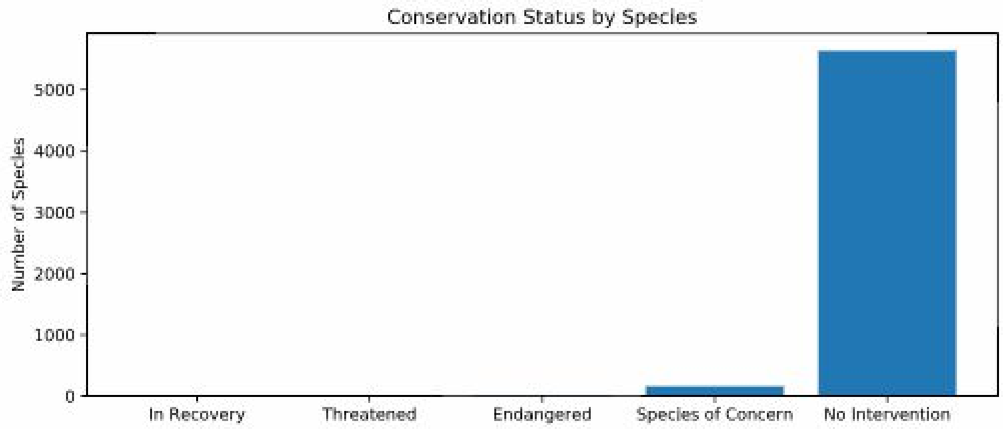


	conservation_status	scientific_name
0	Endangered	16
1	In Recovery	4
2	No Intervention	5633
3	Species of Concern	161
4	Threatened	10

Graph of Conservation Status by Species

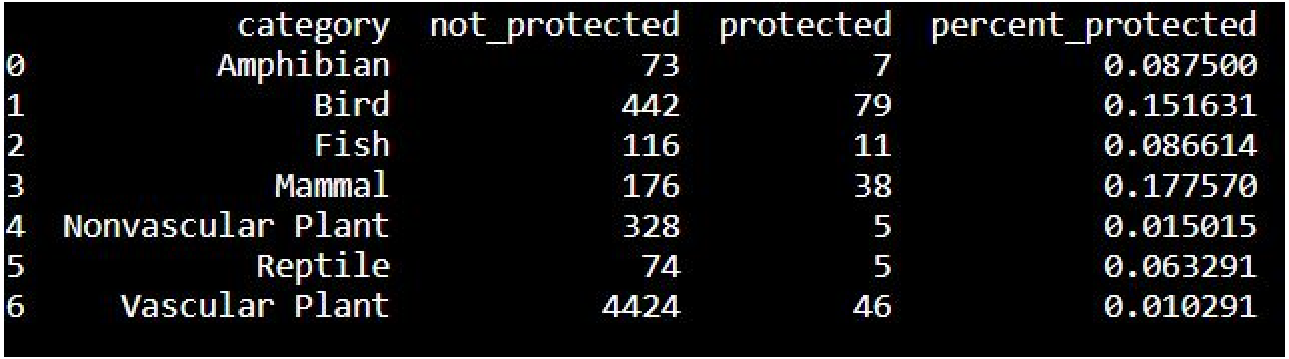
Then we make a bar chart via matplotlib.

Most animals are not intervened



Pivot the Table

But we want to know how is the conservation status of each animal category, so we add another column ‘is_protected’, and then group by ‘category’ and ‘is_protected’. For better visualization, we pivot the table and got table below.



```
             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
5             Reptile             74          5           0.063291
6      Vascular Plant           4424         46           0.010291
```

Are certain types of species more likely to be endangered?

We want to know if certain types of species are more likely to be endangered.

We apply Chi-Squared Test and see the Null Hypothesis be “the difference is due to chance”. I put all data from the previous data into Chi-Squared Test,

```
contingency = [[75, 413], [30, 146], [7,72], [11,115],[5,328],[5,73],[46,4216]]

pval = chi2_contingency(contingency)[1]

print(pval)
```

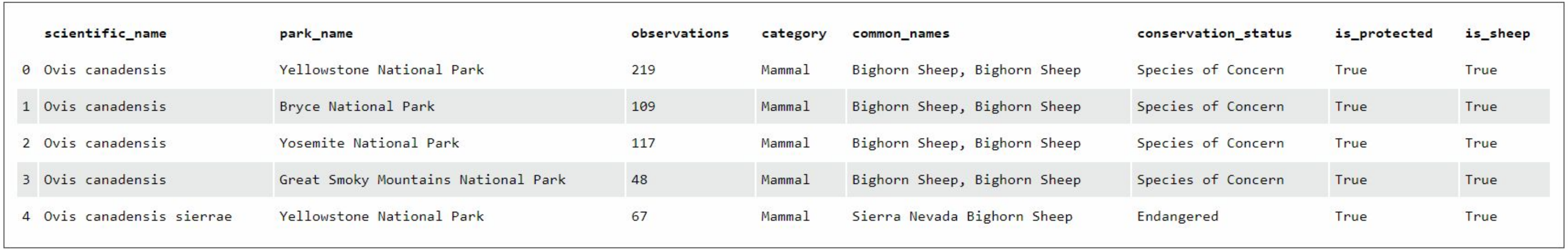
Pval = 5.51082804731e-89

This means there is significant difference between this data. So the answer is positive.

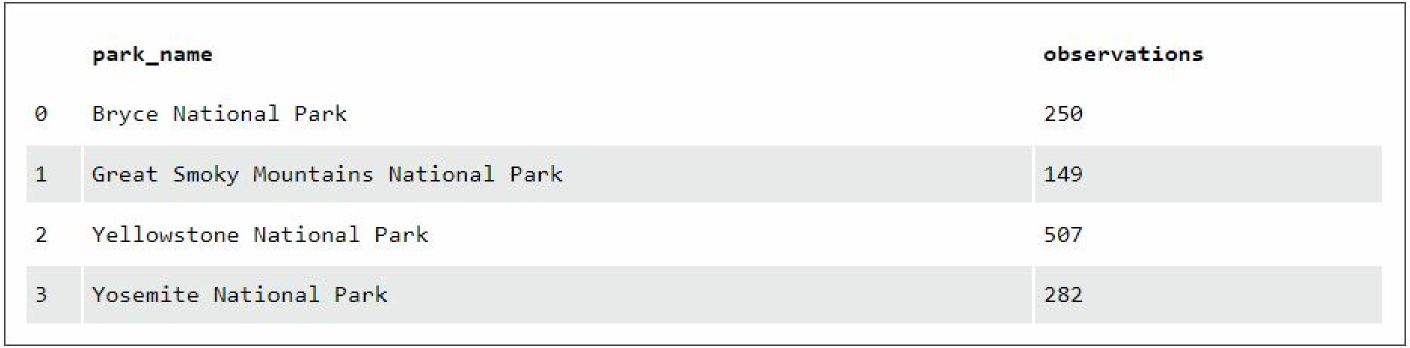
There is certain types of species are more likely to be endangered.

In Search of Sheep

A quick glimpse at the original table shows us the table below. We do not need as much data as this, so manipulate it and get the second table.

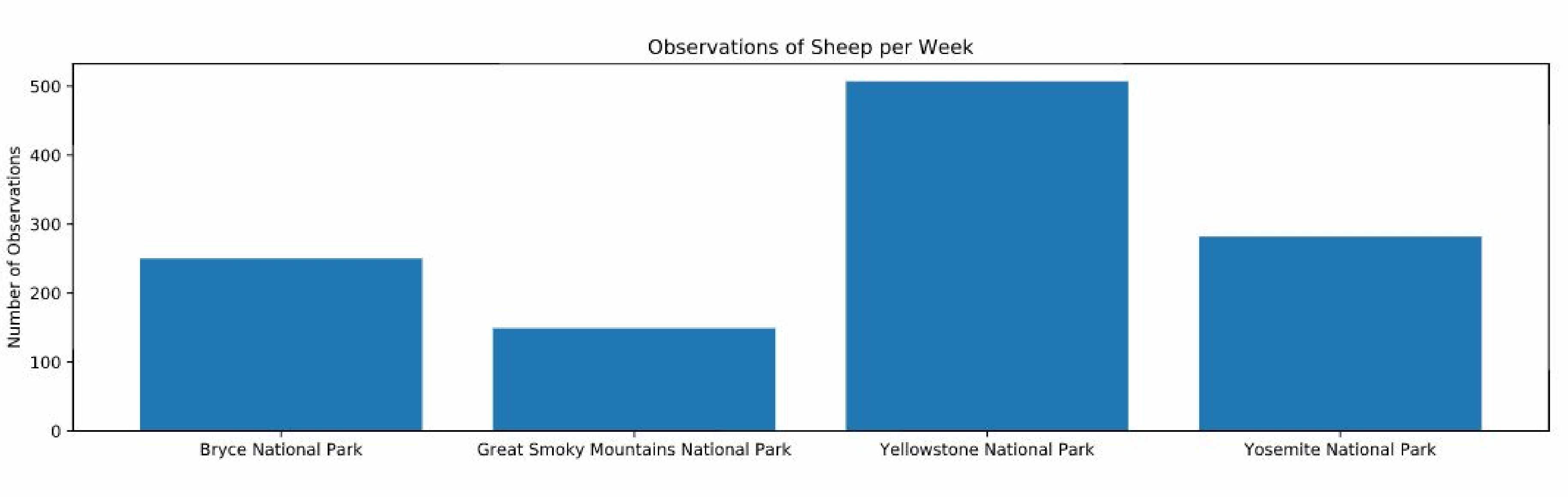


	scientific_name	park_name	observations	category	common_names	conservation_status	is_protected	is_sheep
0	Ovis canadensis	Yellowstone National Park	219	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True
1	Ovis canadensis	Bryce National Park	109	Mammal	Bighorn Sheep, Bighorn Sheep	Species 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



	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

Then we plot the previous table and get the graph below.



Foot and Mouth Reduction Effort

Based on the data below(in percentage):

Baseline = 15

Confidence line = 90%

minimum_detectble _effect = 33

We get the sample size per variant is 510

The table below shows how many week(s) every park need for the observation:

Park Name	week(s) need for observation
Bryce National Park	2
Great Smoky Mountains National Park	3.4
Yellowstone National Park	1
Yosemite National Park	1.8