Analysis of Conservation Status of Animals observed in National Parks

Analysis Performed by: Srilakshmi Sagiraju

Inspecting Dataframe

Species_info.csv contains information of nearly 6000 different species of plants and animals. It list each animals and plants scientific name and its corresponding common name/s. The conservation status of some of the species is available while most of the conservation status of the species are missing.

Calculations Performed:

Obtained Species Count by counting the unique number of data values in scientific_name column.

species.scientific_name.nunique()

Species Count: 5541

Obtained Species type by counting the unique number of data values in category column.

species.category.nunique()

Species Type: [‘Mammal’ ‘Bird’ ‘Reptile’ ‘Amphibian’ ‘Fish’, ‘Vascular Plant’ ‘Nonvascular Plant’]

Analyze Species Conservation Status

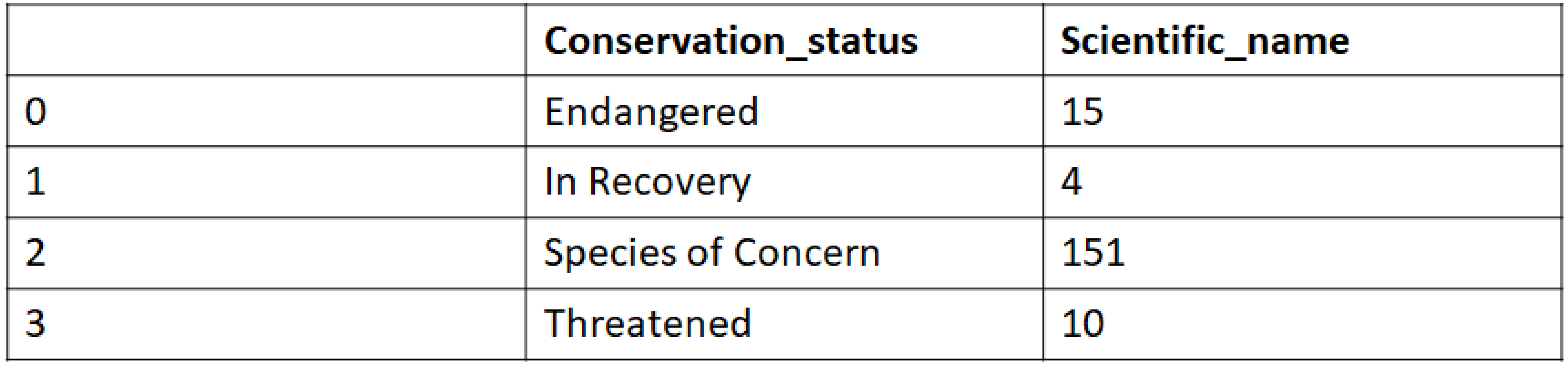
Obtained Conservation Statuses by counting the unique number of data values in conservation_status column.

```
species.conservation_status.nunique()
```

```
Conservation_statuses: [nan ‘Species of Concern’ ‘Endangered’ ‘Threatened’ ‘In Recovery’]
```

Counting species that fall into the corresponding conservation statuses

```
species.groupby(‘conservation_status’).scientific_name.nunique().reset_index()
```



	Conservation_status	Scientific_name
0	Endangered	15
1	In Recovery	4
2	Species of Concern	151
3	Threatened	10

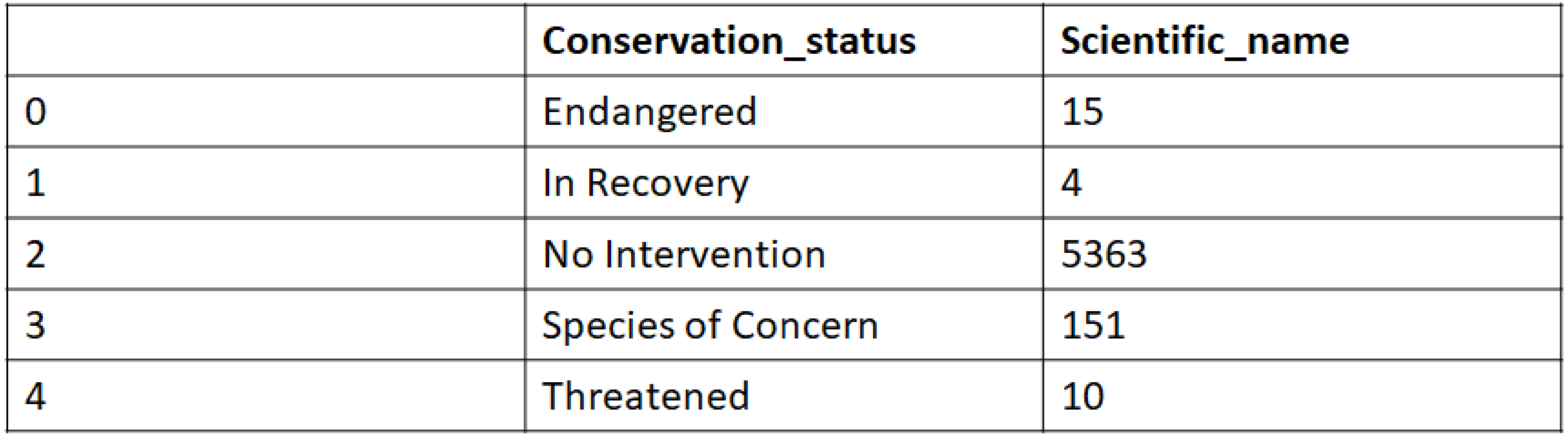
Analyze Species Conservation Status contd…

Species count as seen earlier is greater than 5000, but grouping by conservation_status did not account for all species. Since groupby does not take into account the Nan status, the above data needs to be cleaned to get an accurate representation of conservation_status. Filling the data frame with the data value ‘No Intervention’ for those data values with NaN

```
species.fillna(‘No Intervention’), inplace = True)
```

Recounting species after fixing the conservation_status to ‘No Intervention’.

```
species.groupby(‘conservation_status’).scientific_name.nunique().reset_index()
```



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

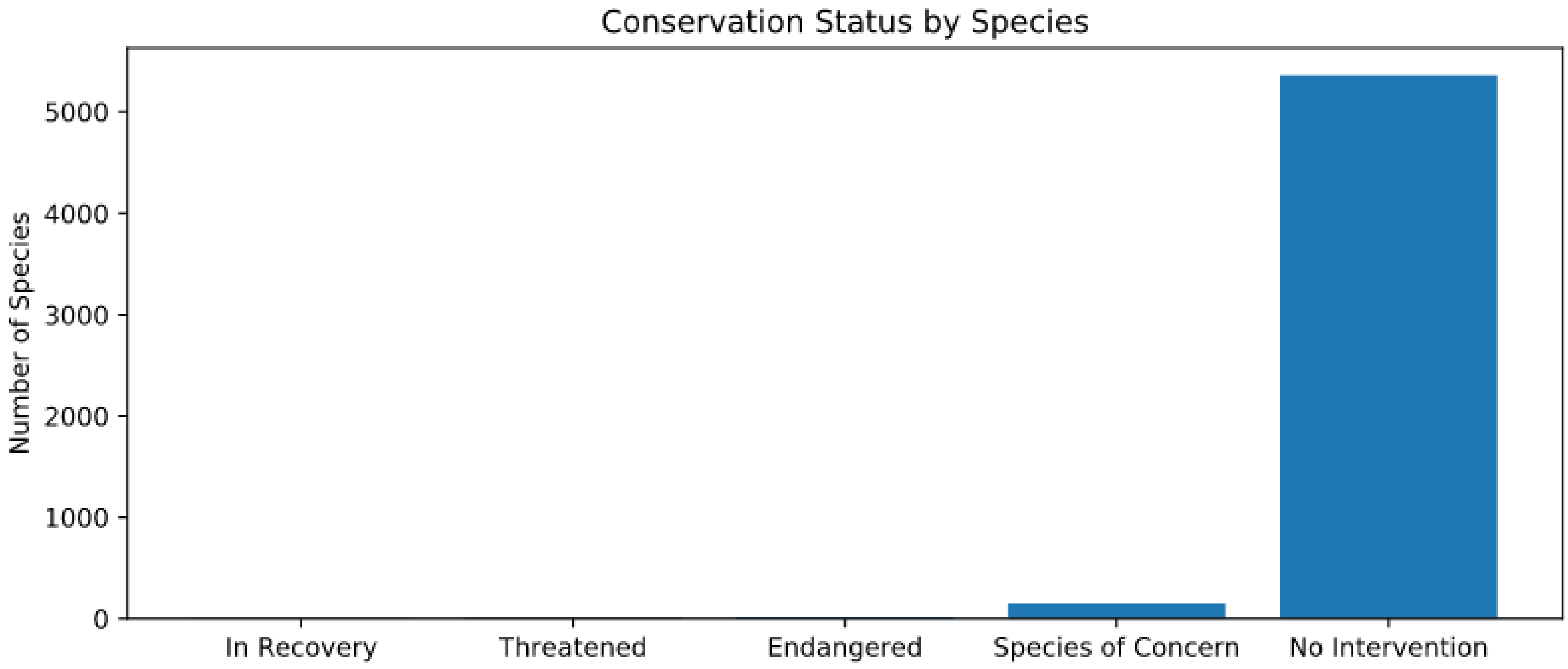
Analyze Species Conservation Status contd…

Grouping by conservation_status leads to data being sorted by conservation_status. To create a bar graph that is visually pleasing we re-sort the data frame by scientific_name to better visualize the data.

```
species.groupby(‘conservation_status’)\.scientific_name.nunique().reset_index()\.sort_values(by=‘scientific_name’)
```

	Conservation_status	**Scientific_name**
1	In Recovery	4
4	Threatened	10
0	Endangered	15
3	Species of Concern	151
2	No Intervention	5363

Plotting Conservation Status by Species

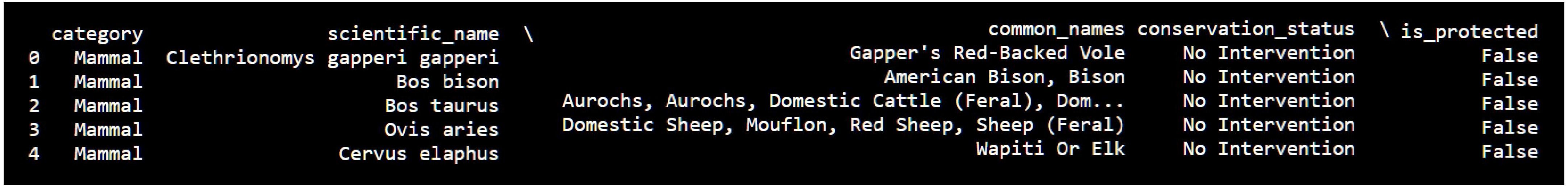


Investigating Endangered Species

Based on the analysis , it appears that around 180 species of plants and animals are likely to be endangered. To answer the question, which of these species are more likely to be endangered, we perform the following analysis.

We create a new column is_protected and set it to “False” if conservation status is equal to “No Intervetion” and “True” otherwise.

```
species['is_protected'] = species.conservation_status != 'No Intervention'
```

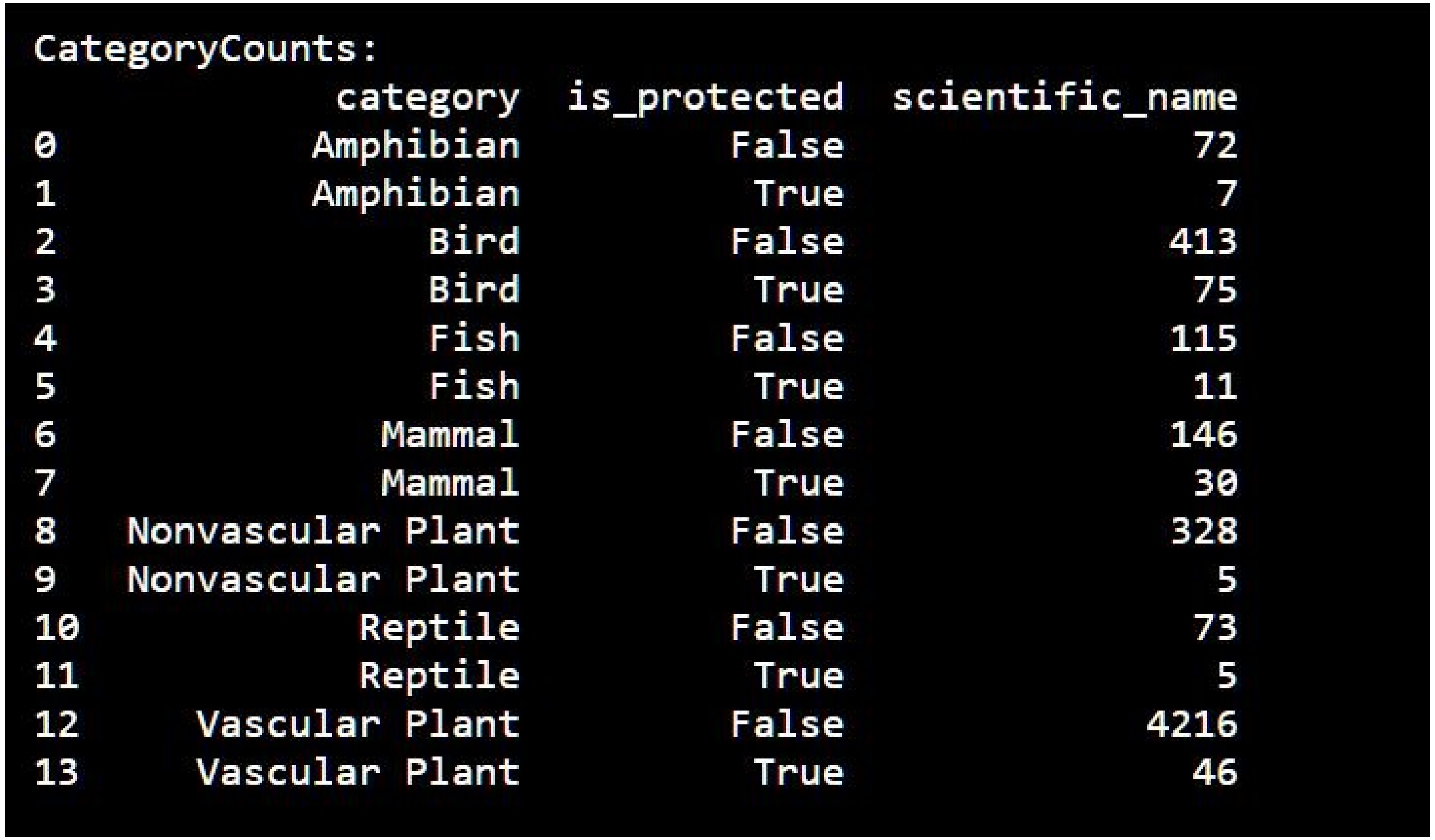


```
  category                scientific_name  \                                     common_names conservation_status  \ is_protected
0   Mammal  Clethrionomys gapperi gapperi                            Gapper's Red-Backed Vole     No Intervention          False
1   Mammal                      Bos bison                               American Bison, Bison     No Intervention          False
2   Mammal                     Bos taurus     Aurochs, Aurochs, Domestic Cattle (Feral), Dom...     No Intervention          False
3   Mammal                     Ovis aries     Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)     No Intervention          False
4   Mammal                  Cervus elaphus                                     Wapiti Or Elk     No Intervention          False
```

Investigating Endangered Species contd…

Group data by columns; category, is_protected and counting the number of scientific_name for each category and is_protected

```
species.groupby(['category', 'is_protected']).scientific_name.nunique().reset_index()
```

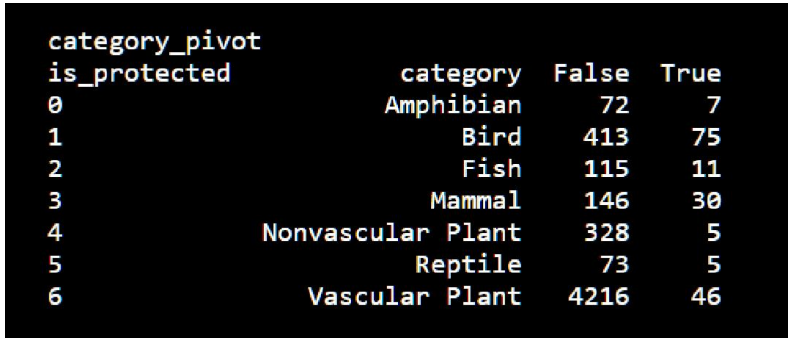


```
CategoryCounts:
             category  is_protected  scientific_name
0           Amphibian         False               72
1           Amphibian          True                7
2                Bird         False              413
3                Bird          True               75
4                Fish         False              115
5                Fish          True               11
6              Mammal         False              146
7              Mammal          True               30
8   Nonvascular Plant         False              328
9   Nonvascular Plant          True                5
10            Reptile         False               73
11            Reptile          True                5
12     Vascular Plant         False             4216
13     Vascular Plant          True               46
```

Investigating Endangered Species contd…

To get a better view of the data, pivot the dataframe so that is_protected values are columns, category is the index and values are scientific_name.

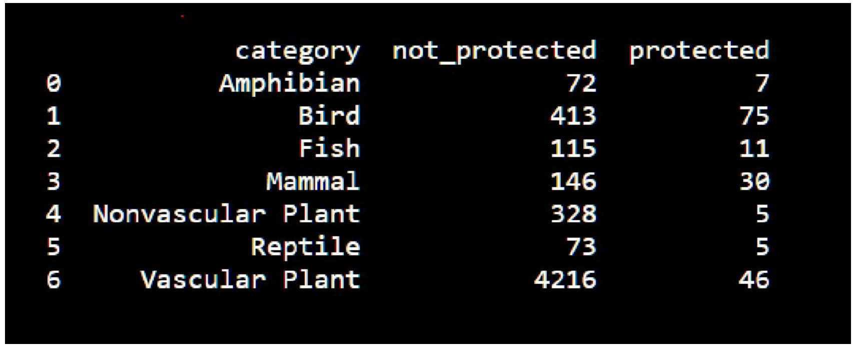
category_counts.pivot(columns='is_protected’, index='category’, values='scientific_name’).reset_index()



```
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
```

The columns False and True do not give a clear meaning, so changing the columns to not_protected and protected

category_pivot.columns = ['category', 'not_protected', 'protected'];

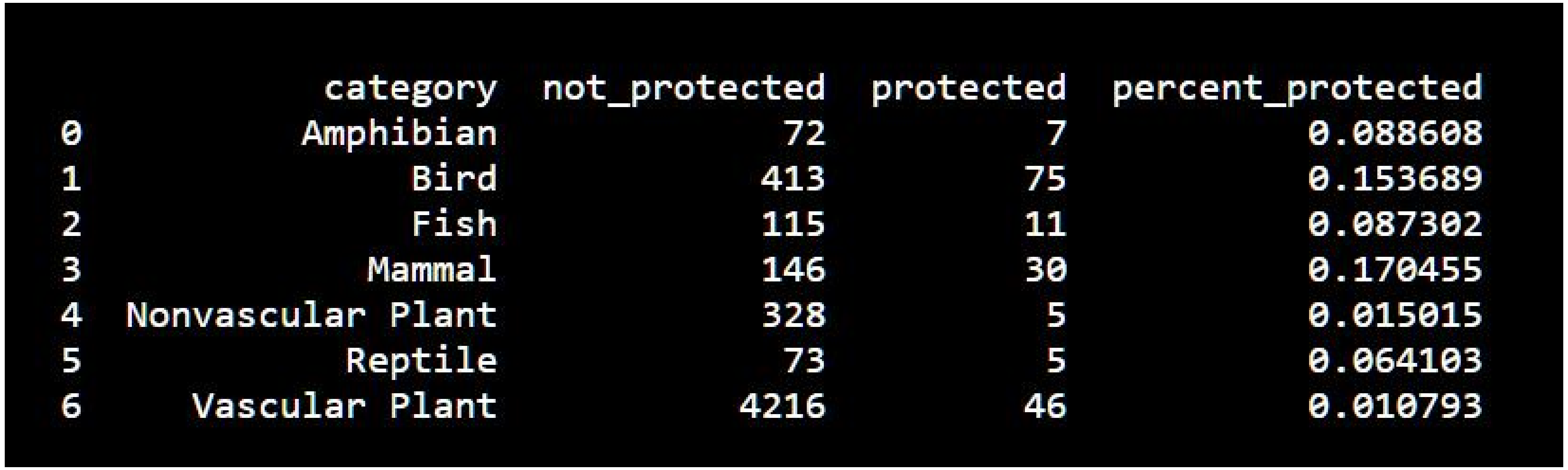


```
            category  not_protected  protected
0          Amphibian             72          7
1               Bird            413         75
2               Fish            115         11
3             Mammal            146         30
4  Nonvascular Plant            328          5
5            Reptile             73          5
6     Vascular Plant           4216         46
```

Investigating Endangered Species contd…

Creating a new column percent_protected and calculating the percent of endangered protected species

```
category_pivot['percent_protected'] = category_pivot.protected / (category_pivot.protected + category_pivot.not_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-Square Test for Significance

Are Mammals more likely to be endangered than Birds?

Pearson’s chi-squared test is a statistical test applied to sets of categorical data to evaluate how likely it is that any observed difference between the sets arose by chance. Here we are testing Mammals and Birds, protected and non_protected status establishing a null hypothesis that this difference is due to chance.

small chi square value - no definite correlation between the two variables

large chi square value – definite correlation between the two variables.

Creating a contingency table and including the values for mammals and birds (protected and not_protected)

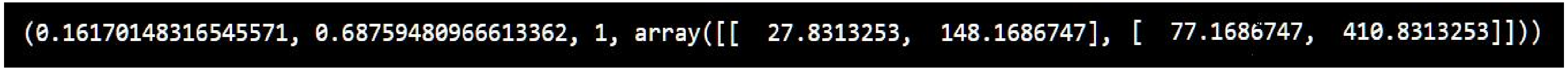
contingency = [[30,146], [75,413]]

chi2_contingency function of scipy.stats computes the chi-square statistic and p-value for the hypothesis test of independence of the observed frequencies in the contingency table

Chi-Square Test for Significance contd..

scipy.stats.chi2_contingency returns a 4 element tuple, where the second element is the p-value.

pval = chi2_contingency(contingency)



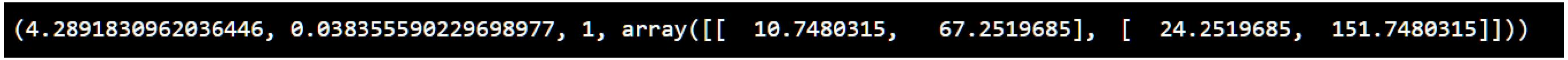
```
(0.16170148316545571, 0.68759480966613362, 1, array([[  27.8313253,  148.1686747], [  77.1686747,  410.8313253]]))
```

There is no significant difference since the p-value 0.69 > 0.05

Testing to see if the observed difference between reptiles and mammals is by chance.

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

pval_reptile_mammal = chi2_contingency(reptile_mammal_contingency)



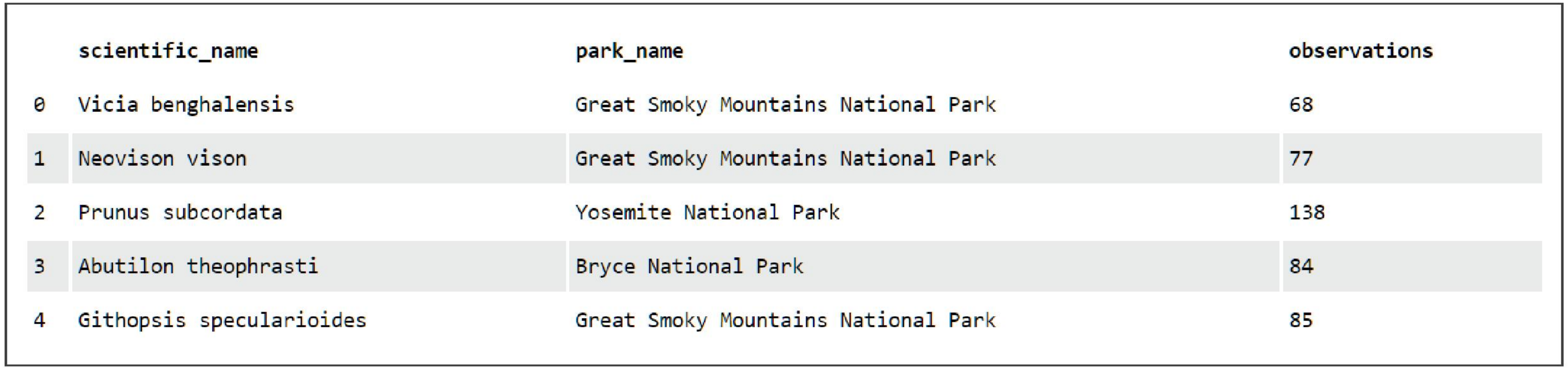
```
(4.2891830962036446, 0.038355590229698977, 1, array([[  10.7480315,   67.2519685],  [  24.2519685,  151.7480315]]))
```

There is significant difference since the p-value 0.04 < 0.05

Therefore we can conclude that certain types of species are more likely to be endangered than others.

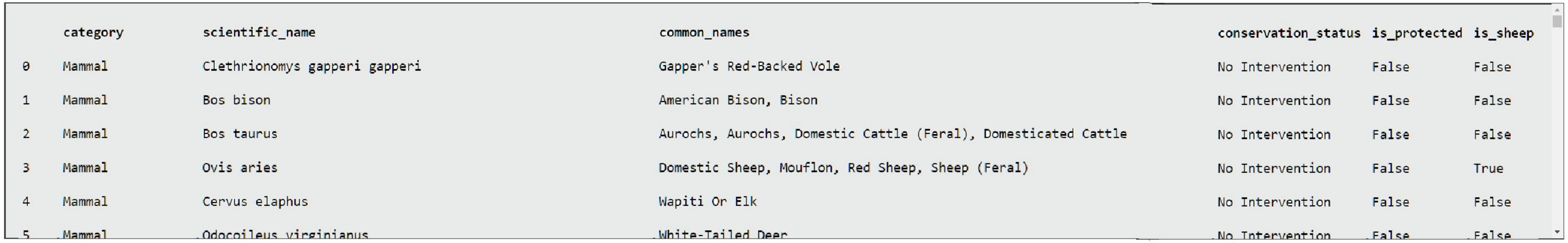
Observations Data frame

The observations data frame contains information about the national park and the number of observed animals with their scientific name.



	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

Manipulating the species data frame to add a column is_sheep and populating with ‘True″ where the common_name column contains sheep as a substring.



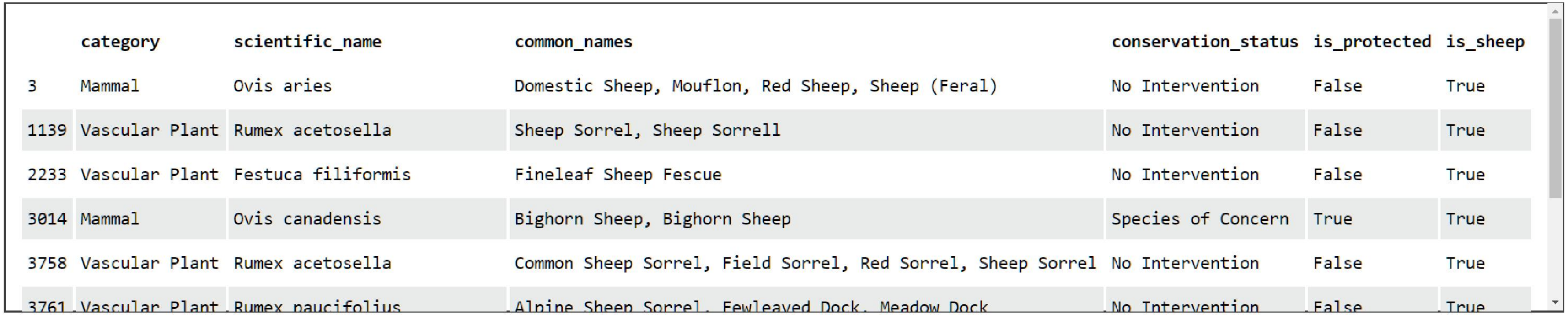
	category	scientific_name	common_names	conservation_status	is_protected	is_sheep
0	Mammal	Clethrionomys gapperi gapperi	Gapper's Red-Backed Vole	No Intervention	False	False
1	Mammal	Bos bison	American Bison, Bison	No Intervention	False	False
2	Mammal	Bos taurus	Aurochs, Aurochs, Domestic Cattle (Feral), Domesticated Cattle	No Intervention	False	False
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True
4	Mammal	Cervus elaphus	Wapiti Or Elk	No Intervention	False	False
5	Mammal	Odocoileus virginianus	White-Tailed Deer	No Intervention	False	False

```
species['is_sheep'] = species.common_names.apply(lambda x: 'Sheep' in x)
```

Observations Data frame contd…

```
species_is_sheep = species[species.is_sheep]
```

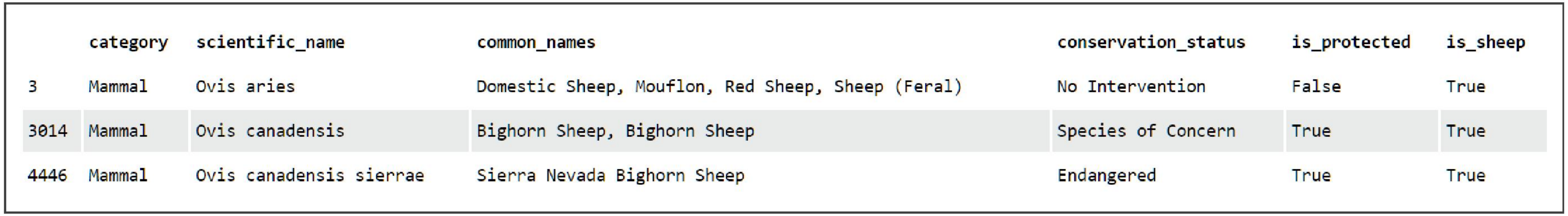
Selecting data where is_sheep is true, it appears that there are some categories of plants included.



	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
1139	Vascular Plant	Rumex acetosella	Sheep Sorrel, Sheep Sorrell	No Intervention	False	True
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

Selecting data where is_sheep is “True” and category “Mammal”

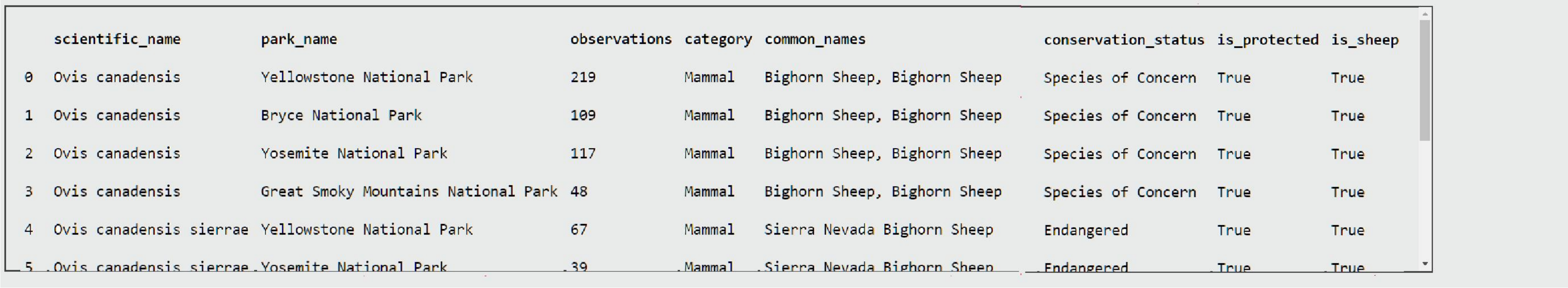
```
sheep_species = species[(species.is_sheep) & (species.category == 'Mammal’)]
```



	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

Merging Sheep and Observation Data frames

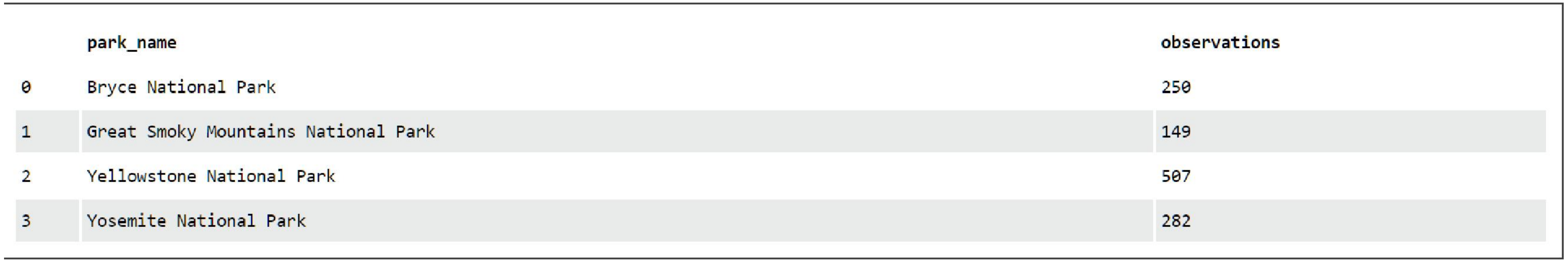
sheep_observations = observations.merge(sheep_species)



	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

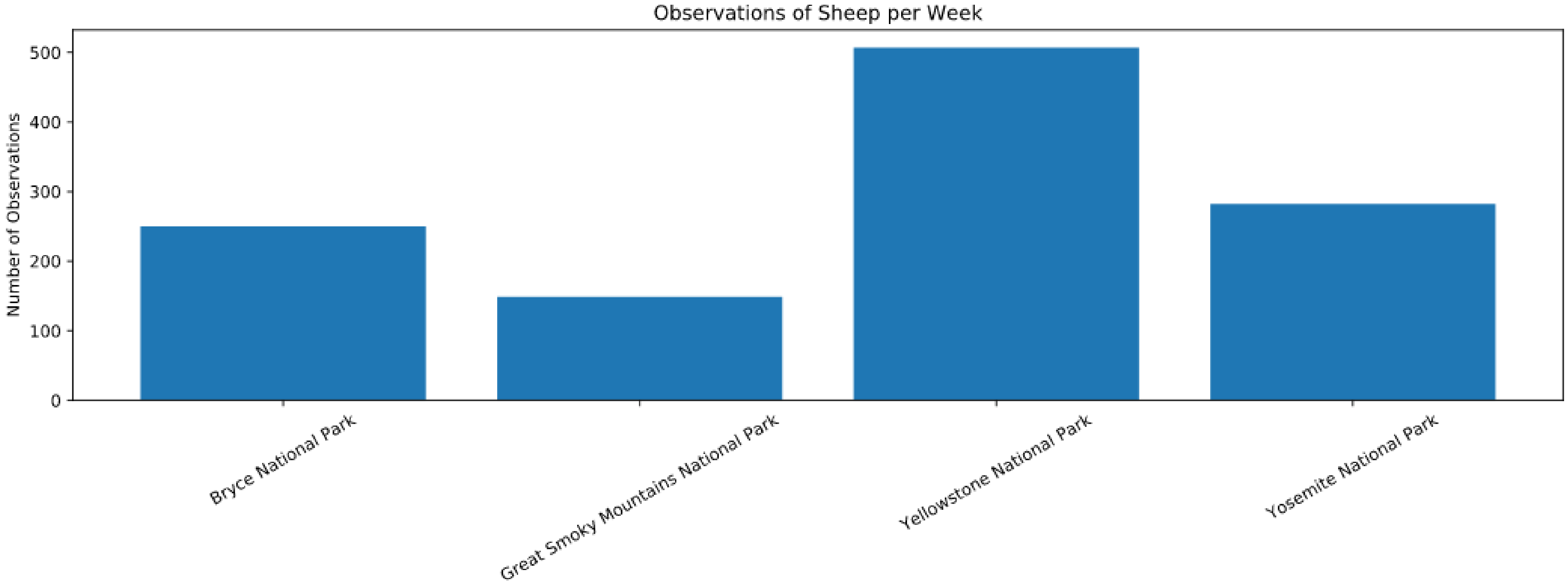
Three species of sheep are observed at four different national parks. Grouping by park the number of sheep observed is shown below

sheep_observations.groupby('park_name').observations.sum().reset_index();



	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

Bar chart showing the number of observations per week at each park.



Foot and Mouth Reduction Effort - Sample Size Determination.

baseline = 15

minimum_detectable_effect = 100 * 5/15 = 33.33

plugging in the baseline and minimum detectable effect into the sample size calculator

sample_size_per_variant = 510

Total number of sheep observed at Yellow Stone National park over a period of 7 days is 507. Therefore the number of weeks observing 510 sheep would be 1 week.

yellowstone_weeks_observing = 1

Total number of sheep observed at Bryce National park over a period of 7 days is 250. Therefore the number of weeks observing 510 sheep would be 2 weeks.

yellowstone_weeks_observing = 2

Conclusion: Foot and Mouth Reduction Effort - Sample Size Determination

Given a baseline of 15% occurrence of foot and mouth disease in sheep at Bryce National Park, if the scientists wanted to be sure that a >5% drop is needed to be considered significant at Yellow Stone National park they would need to observe 510 sheep which would take approximately 1 week or approximately 2 weeks at Bryce National Park.