Project: Capstone Option 2- Biodiversity for the National Park

Nicholas Foo

Content

- Section 1 – Summary of Data Provided in Species_info.csv
- Section 2 - Question: Which type of species likely to be endangered?
- Section 3 - Question: Is there a Significant Difference in Percentage of Endangered Species in Different Categories?
  - Section 3.1 - Question: Is there a Significant Difference in Percentage of Endangered Species Between Mammals (~17%) and Birds (~15%)
  - Section 3.2 - Question: Is there a Significant Difference in Percentage of Endangered Species Between Mammals (~17%) and Reptiles (~6%)
- Section 4 – Sample Size Determination for Foot and Mouth Disease Study

Section 1 - Summary of Data Provided in Species_info.csv

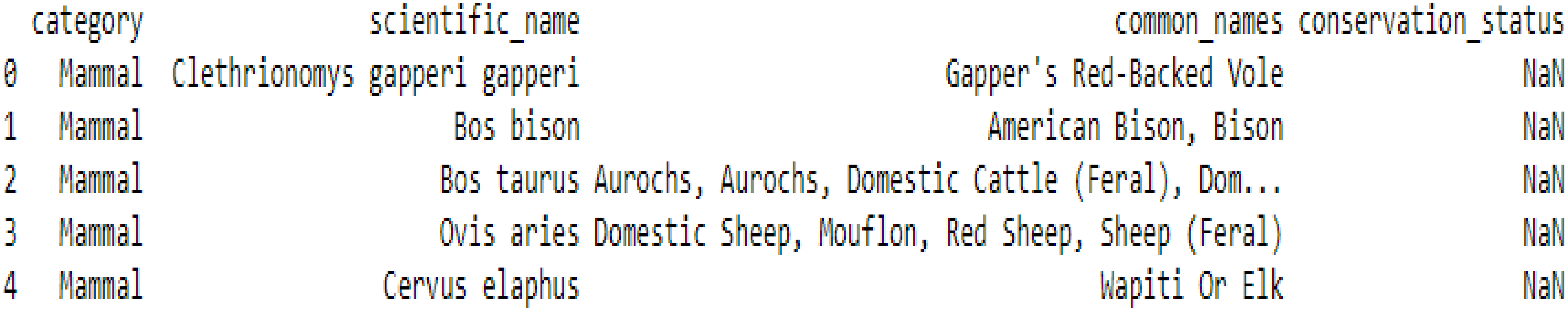
- Dataframe with 4 columns and 5824 rows
- Contains information on different species at National Park
- Column breakdown:-

Table 1: Description of Column in Dataframe Species_info.csv

Title of Column	Description of Column	Data Type of Column Entries
Category	Classification of animal/plant type	String
Scientific_name	Scientific name of animal/plant	String
Common_names	Common name of animal/plant	String
Conservation_status	Conservation status of animal/plant	String

- Sample of Dataframe entries

Table 2: First 5 Rows of Dataframe Species_info.csv



```
  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), Dom...                 NaN
3   Mammal                     Ovis aries  Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)                 NaN
4   Mammal                 Cervus elaphus                                       Wapiti Or Elk                 NaN
```

Summary(Continue - Category)

- 7 species in dataframe (Mammals, Birds, Reptiles, Amphibian, Fish, Vascular Plant, Nonvascular Plant)

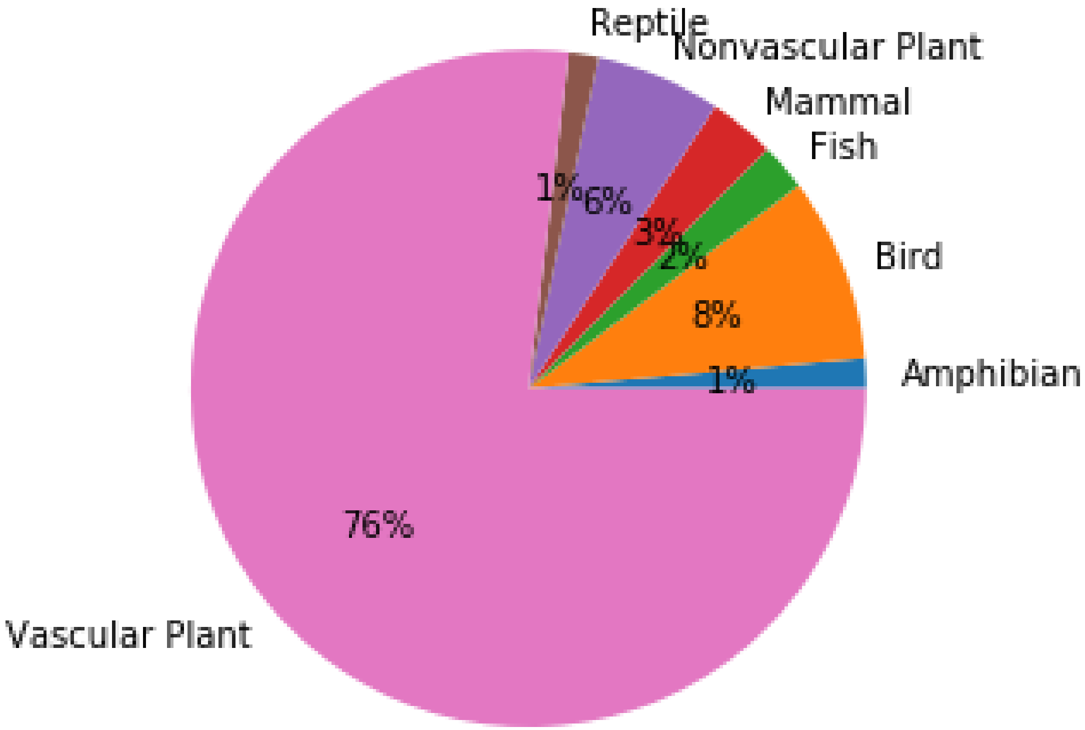


Figure 1: Pie Chart of Species Found in National Park

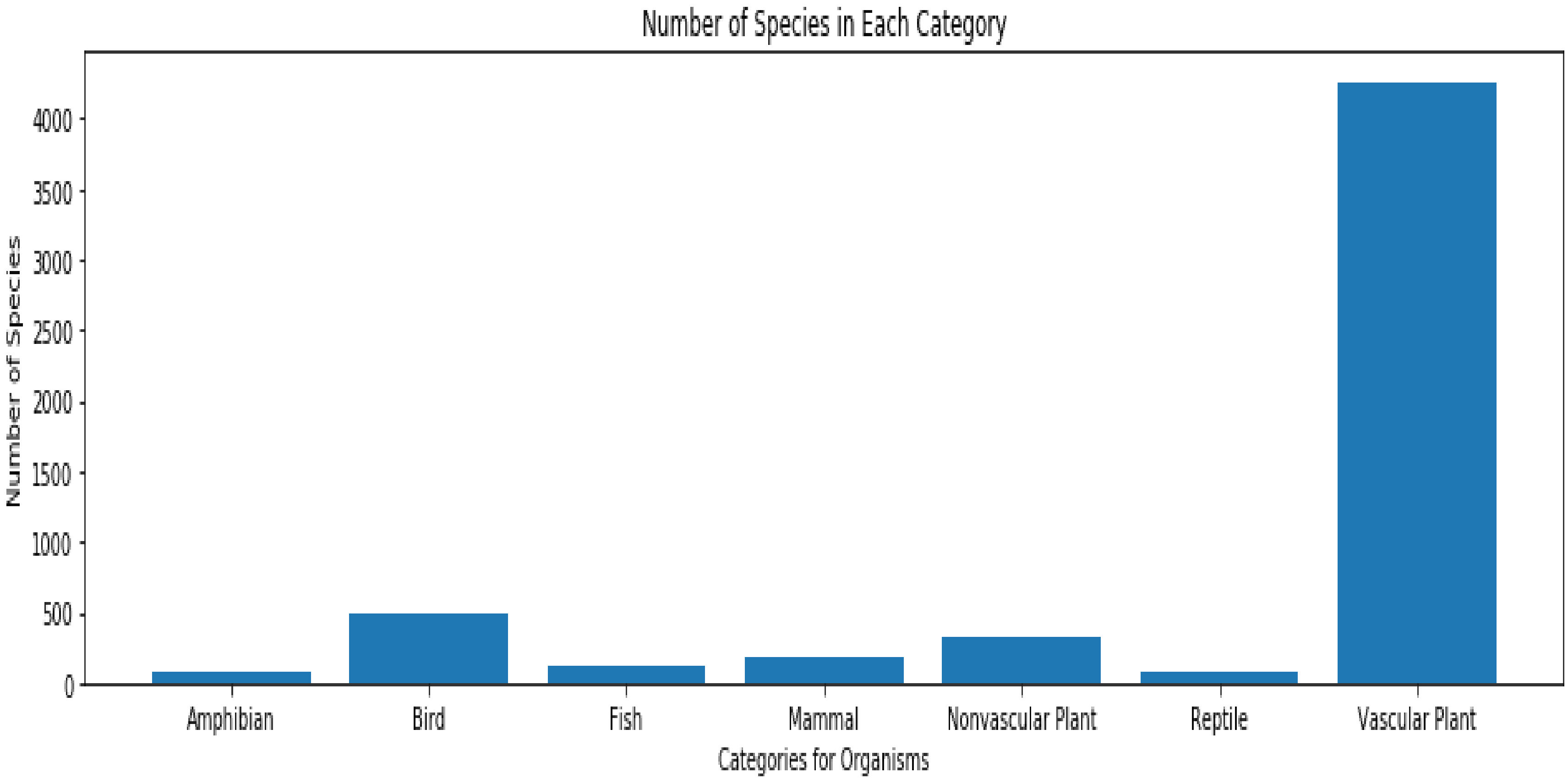
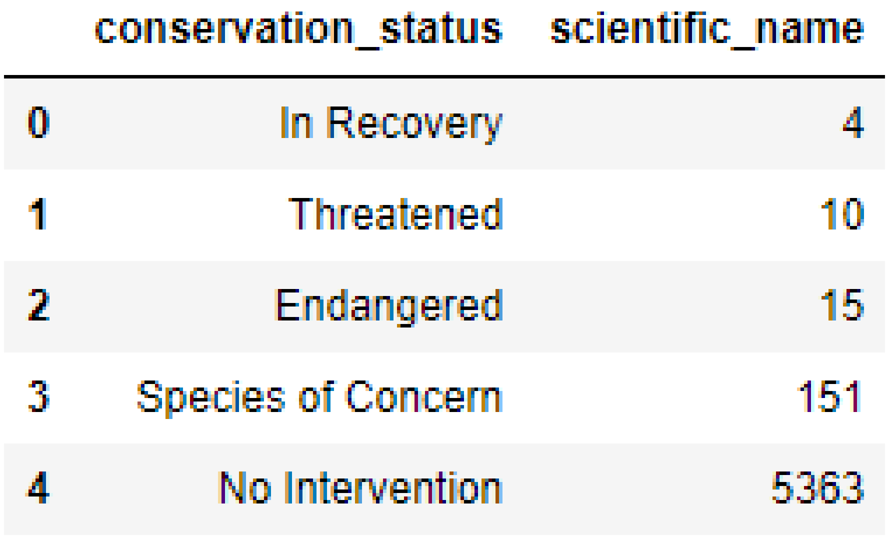


Figure 2: Number of Species in Each Category

Summary (Continue – Conservation Status)

- 5 conservation status (In Recovery, Threatened, Endangered, Species of Concern, No Intervention (previously NaN in dataframe))

Table 3: Number of Scientific Names under Each Conservation Status



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

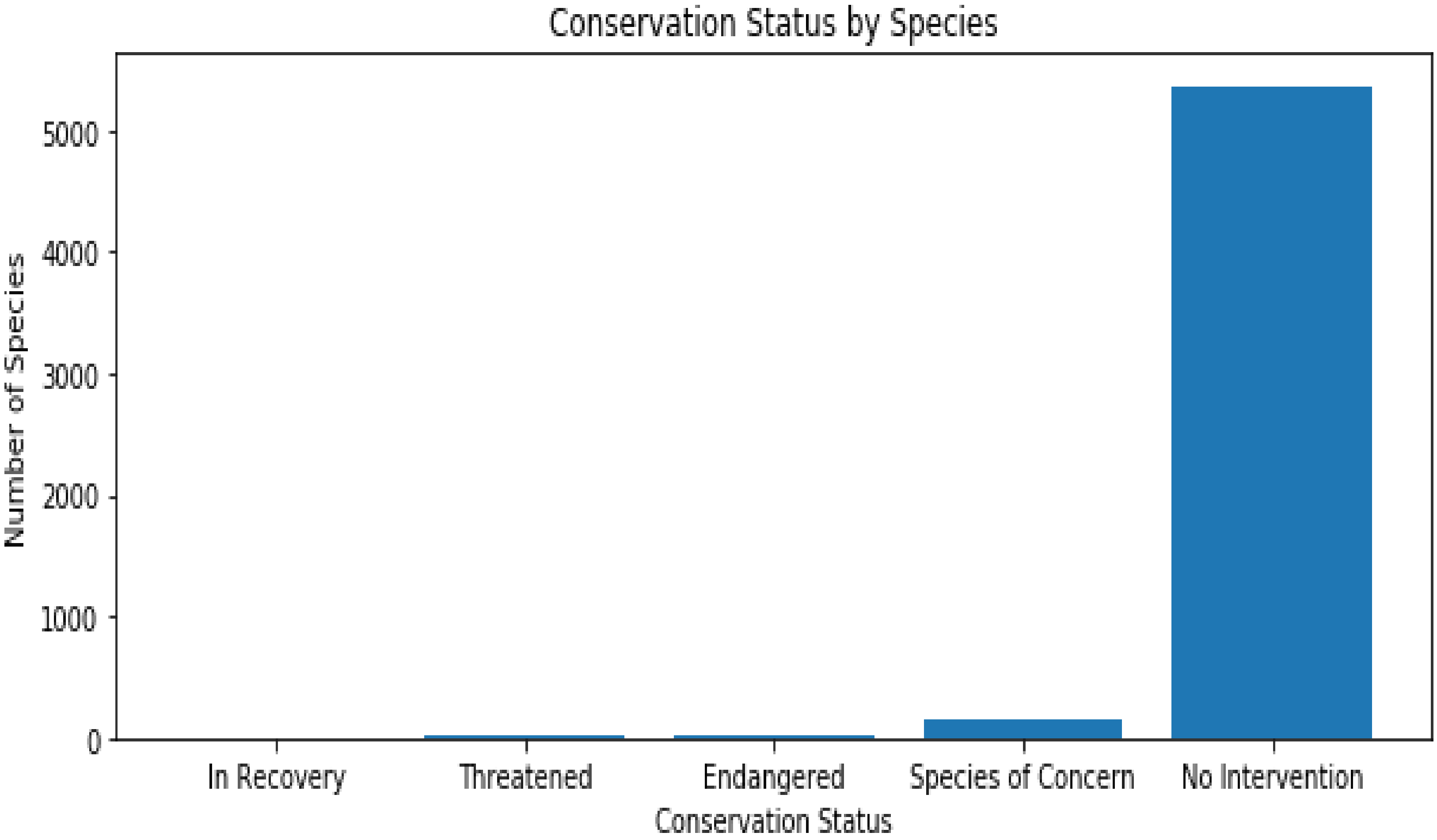
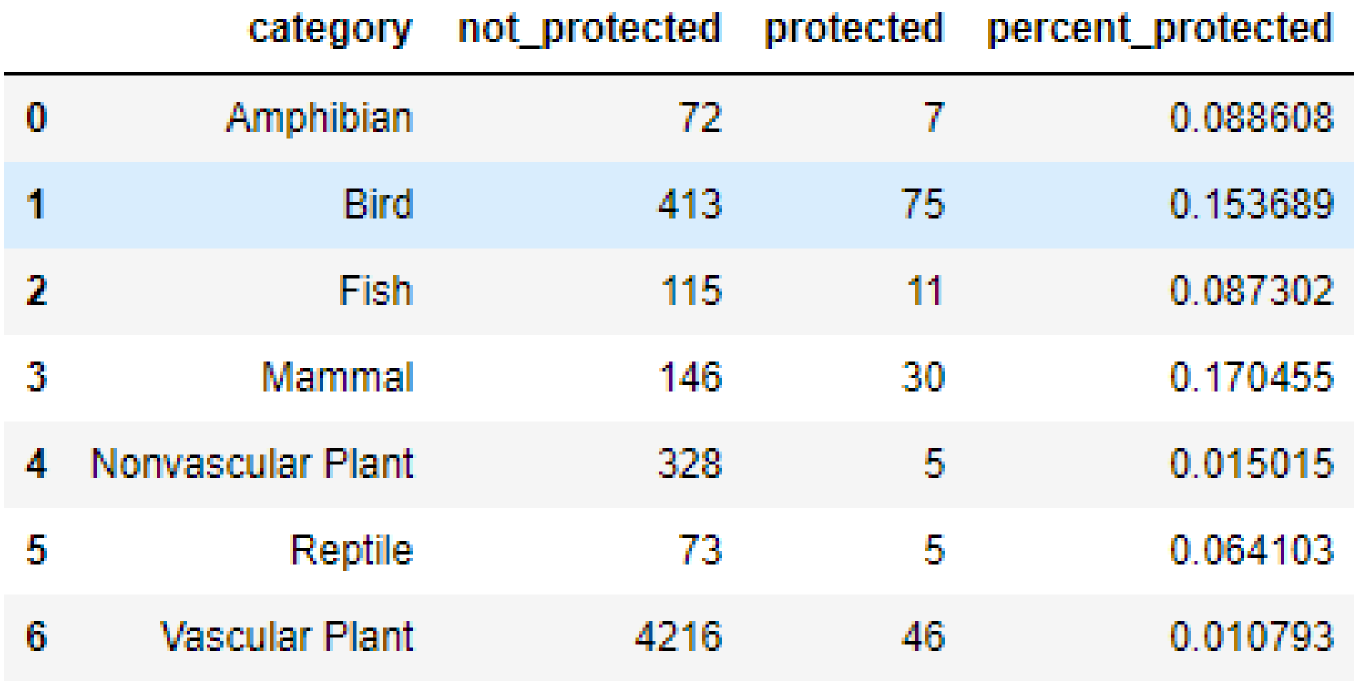


Figure 3: Number of Species under Different Conservation Status

Section 2 – Question: Which type of species likely to be endangered?

Table 4: Summary of Conservation Status of Each Category



	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

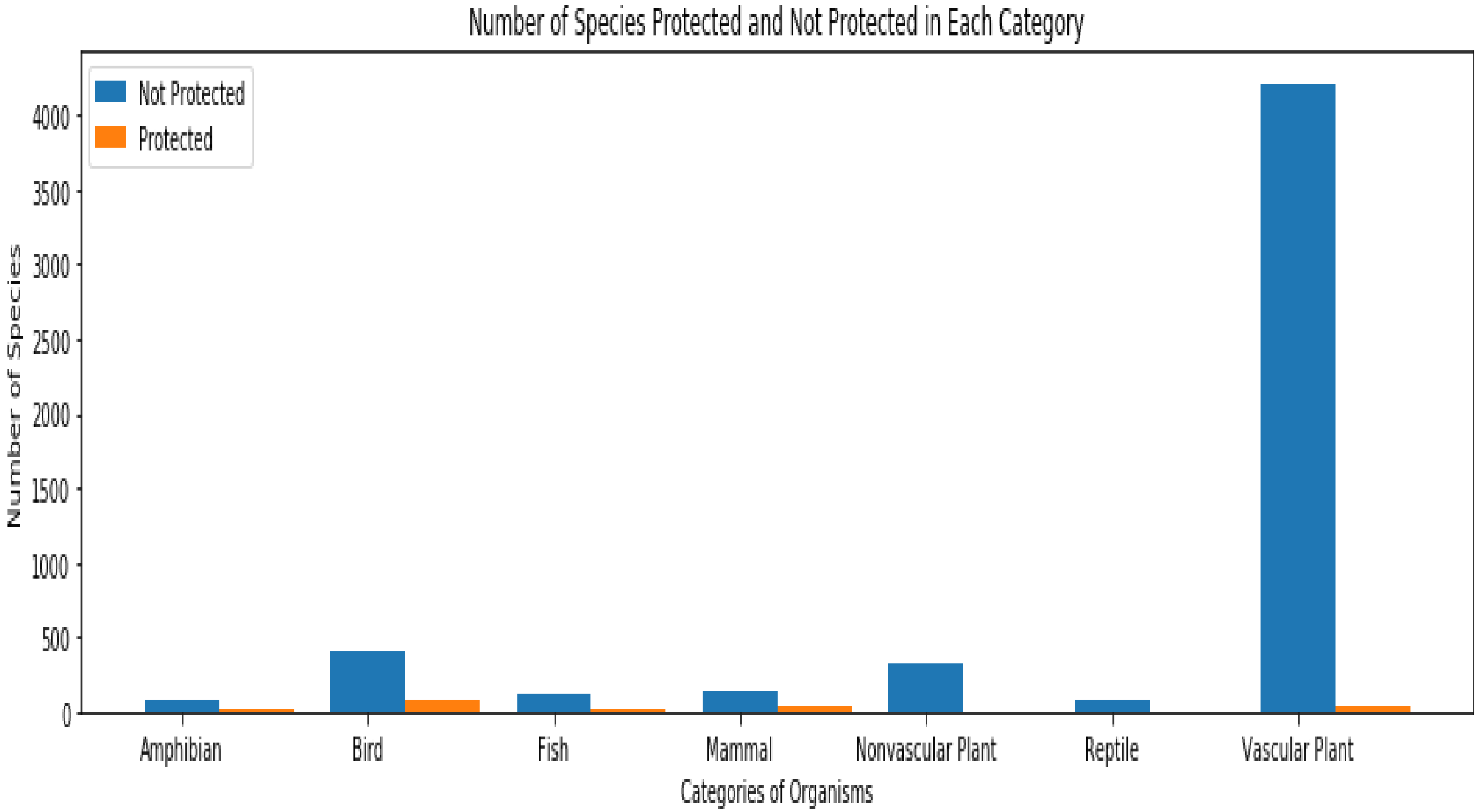


Figure 4: Number of Species Protected & Not Protected in Each Category

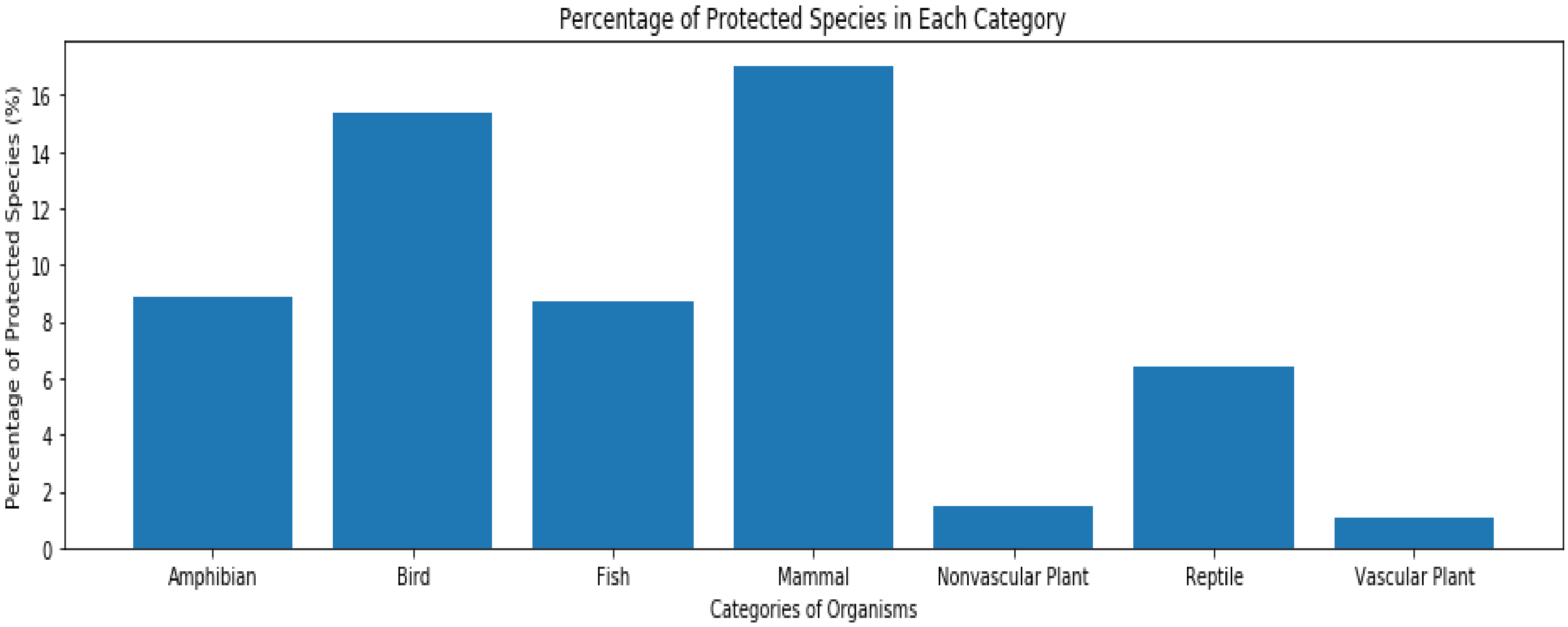


Figure 5: Percentage of Protected Species in Each Category

- Mammals (~17%) and birds (~15%), most endangered category
- Nonvascular plants (~1.5%) and vascular plants (~1.08%), least endangered category

Section 3 - Question: Is there a Significant Difference in Percentage of Endangered Species in Different Categories?

- Specifically:-
  - Between Mammals (~17%) and Birds (~15%)
  - Between Mammals (~17%) and Reptiles (~6%)
- Method = Perform a significance test
  1. Form appropriate null and alternative hypothesis
  2. Decide on appropriate hypothesis test. Significant p-value set at 0.05.

Section 3.1 - Question: Is there a Significant Difference in Percentage of Endangered Species Between Mammals (~17%) and Birds (~15%)

- Null hypothesis
  - Difference in percentage of endangered species for mammals & birds is due to chance
- Alternative hypothesis
  - Difference in percentage of endangered species for mammals & birds is not due to chance
- Hypothesis Test
  - Chi-Square Test = due to categorical data + >2 categorical dataset to compare
  - Form contingency table and perform Chi-Square Test

Result

- p-value = 0.688 (not significant)
- Accept null hypothesis, reject alternative hypothesis

Difference in percentage of endangered species for mammals & birds is due to chance

Section 3.2 - Question: Is there a Significant Difference in Percentage of Endangered Species Between Mammals (~17%) and Reptiles (~6%)

- Null hypothesis
  - Difference in percentage of endangered species for mammals & reptile is due to chance
- Alternative hypothesis
  - Difference in percentage of endangered species for mammals & reptile is not due to chance
- Hypothesis Test
  - Chi-Square Test = due to categorical data + >2 categorical dataset to compare
  - Form contingency table and perform Chi-Square Test

Result

- p-value = 0.038 (significant)
- Accept alternative hypothesis, reject null hypothesis

Difference in percentage of endangered species for mammals & reptile is not due to chance

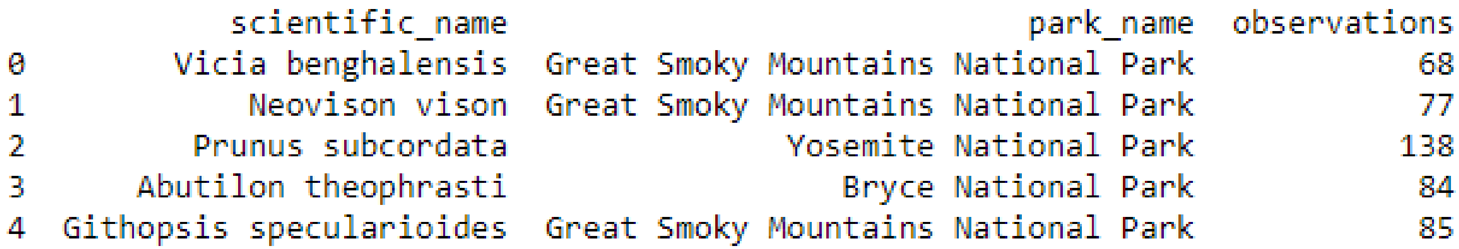
Recommendations Concerning Endangered Species

- Endangerment pressure faced by mammals and birds is similar value (based on non-significant p-value of 0.688), hence reason for similar percentage of protected species value for each category.
  - Recommendation = should look into what is the common endangerment pressure faced by both groups, to protect both groups at the same time. Maybe due to habitat needs, feeding pattern, etc.
- Mammals and birds are the most vulnerable category of organisms faced with endangerment. This is based on the significant p-value of 0.038 found when Chi-Square Test performed on percentage of protected animals in mammals and reptiles
  - Recommendation = focus conservation efforts on birds and mammals as contain the highest percentage of protected species

Section 4 – Sample Size Determination for Foot and Mouth Disease Study

- Provided additional dataframe object (observations.csv)

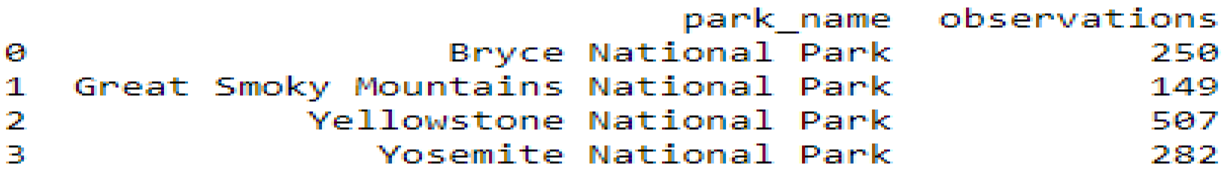
Table 5: First 5 Rows of Dataframe observations.csv



```
              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
```

- Refined species_info.csv dataframe to only include sheeps, merging with observations.csv dataframe, and calculating total number of sheep sightings at each national park per week

Table 6: Total Sheep Sightings at Each National Park per Week



```
                             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
```

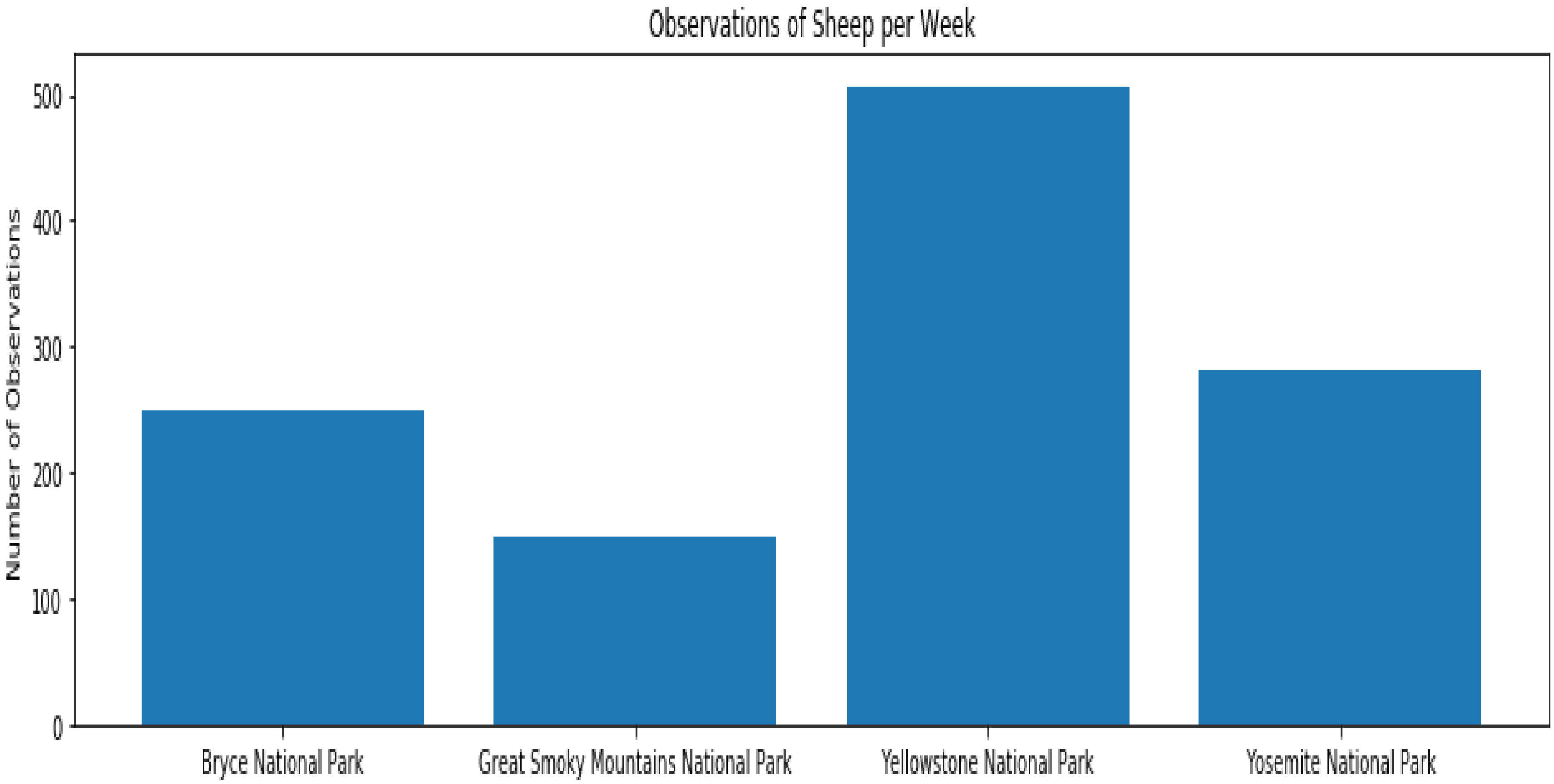


Figure 6: Number of Sheep Observed per Week in Each National Park

Sample Size Determination and Number of Study Weeks

- Baseline conversion rate = 15%
- Minimum detectable effect = $100 \times \frac{5\%}{baseline\ converstion\ rate} = 33.33\%$
  - Use 5% as scientist want to be able to detect reductions of Foot & Mouth of at least 5 %
- Statistical Significance = 90%
- Sample Size Per Variant (calculated using sample size calculator at Optimizely) = 510 sheep
- Number of weeks required by scientist to observe enough sheep at:-
  - Yellowstone National Park = $\frac{510}{507} = 1.01\ weeks$
  - Bryce National Park = $\frac{510}{250} = 2.04\ weeks$

End of Presentation

- Thank you and look forward to comments & criticism.