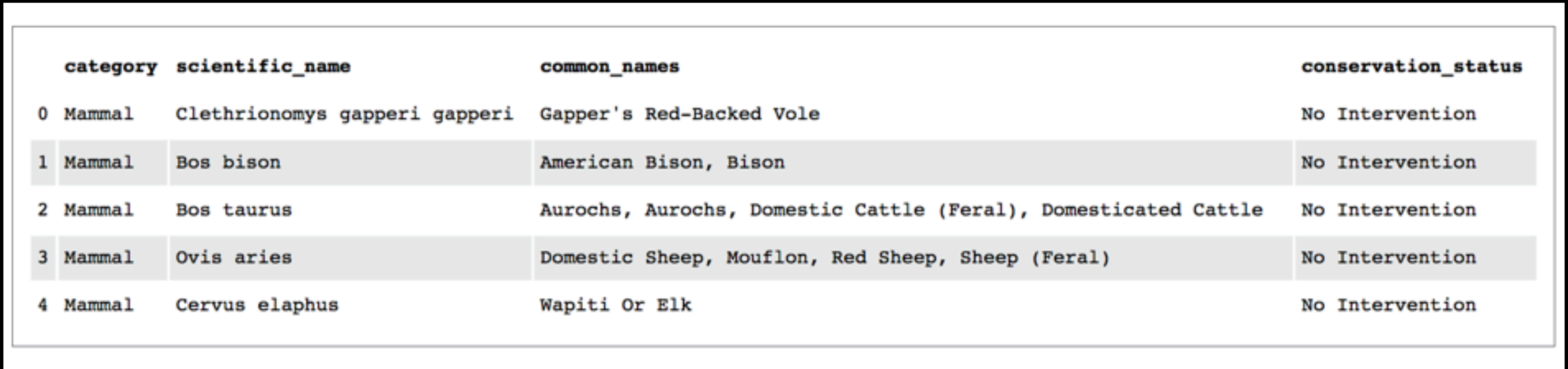
Introduction to Data Analysis: Capstone Option 2: Biodiversity for the National Parks

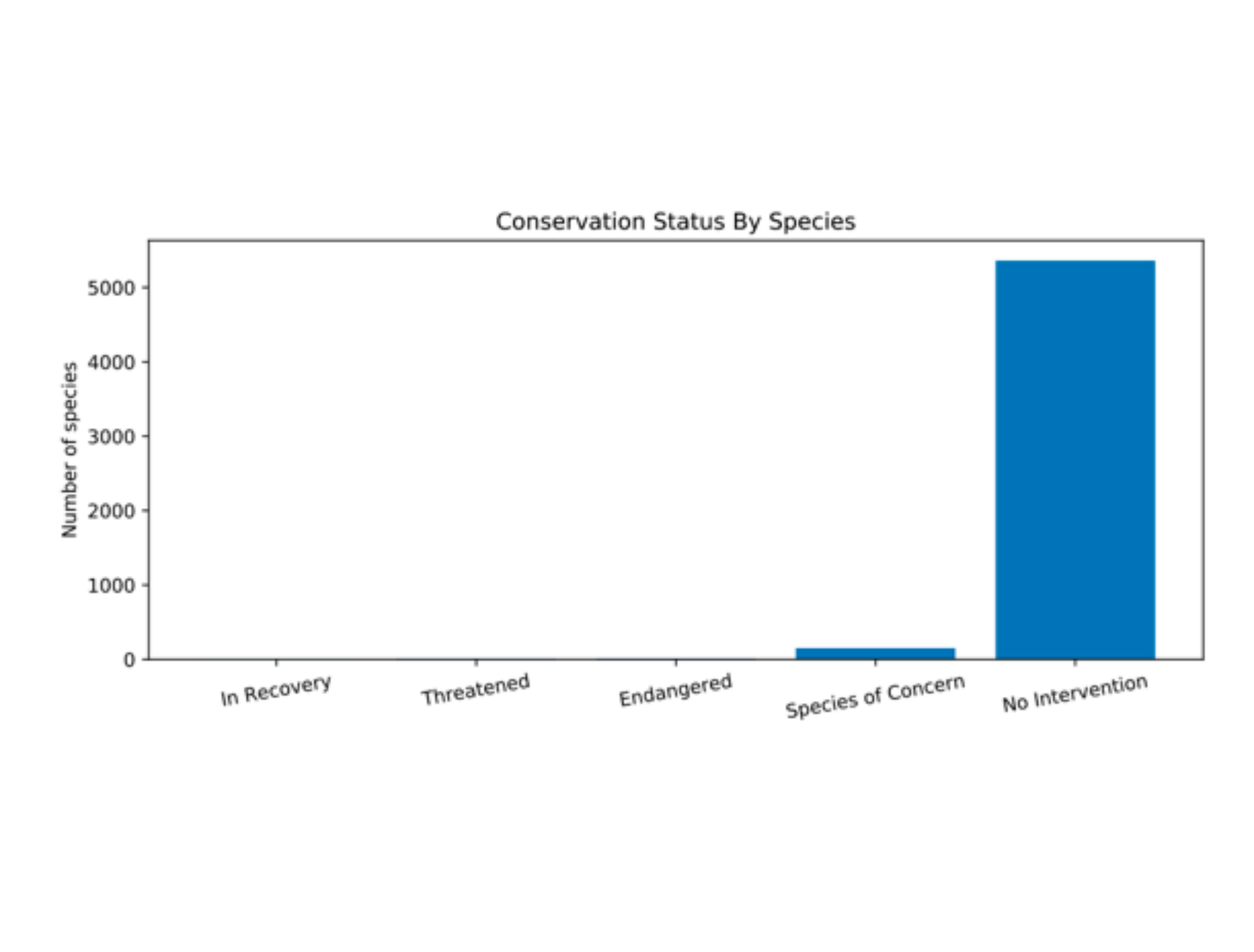
Hanbyul Jo

Species Info Data



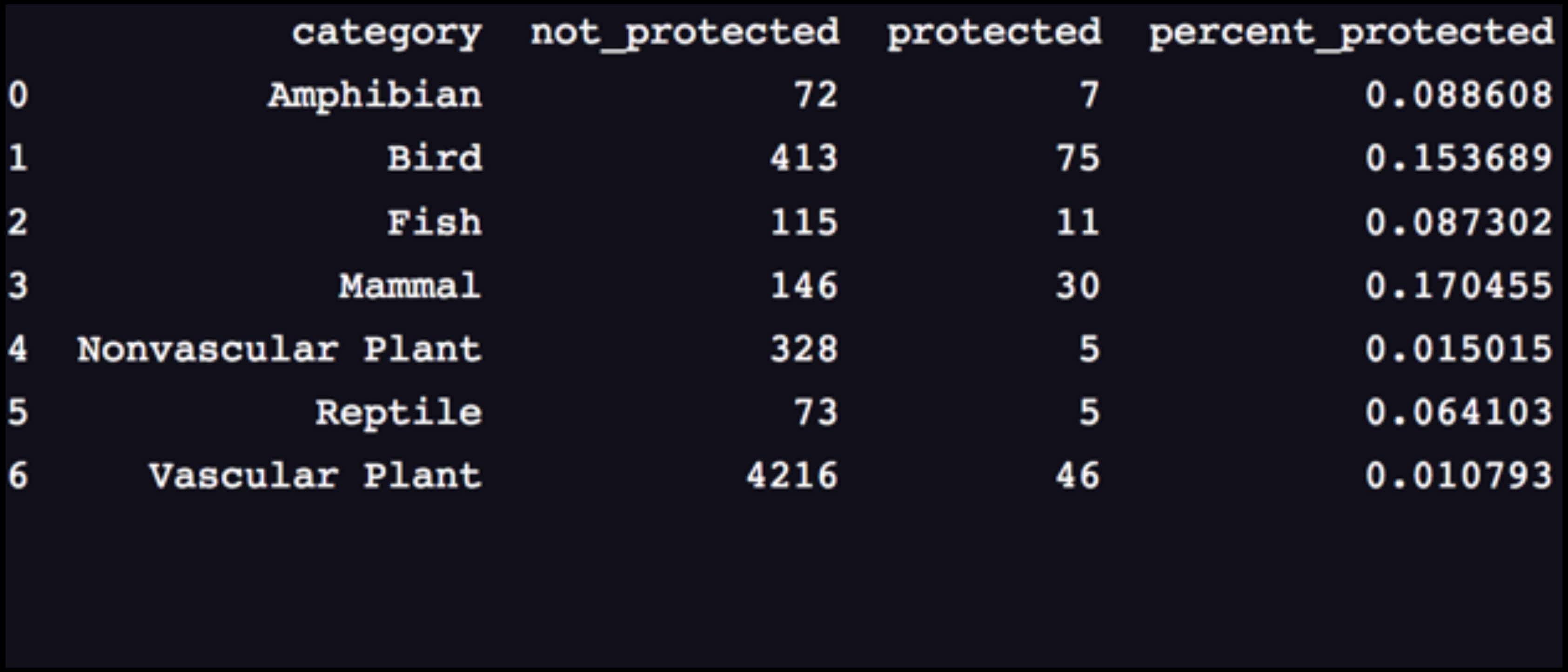
	category	scientific_name	common_names	conservation_status
0	Mammal	Clethrionomys gapperi gapperi	Gapper's Red-Backed Vole	No Intervention
1	Mammal	Bos bison	American Bison, Bison	No Intervention
2	Mammal	Bos taurus	Aurochs, Aurochs, Domestic Cattle (Feral), Domesticated Cattle	No Intervention
3	Mammal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

- There are than 5800 rows in species.csv. It has category, scientific_name, common_name, conservation_status columns.



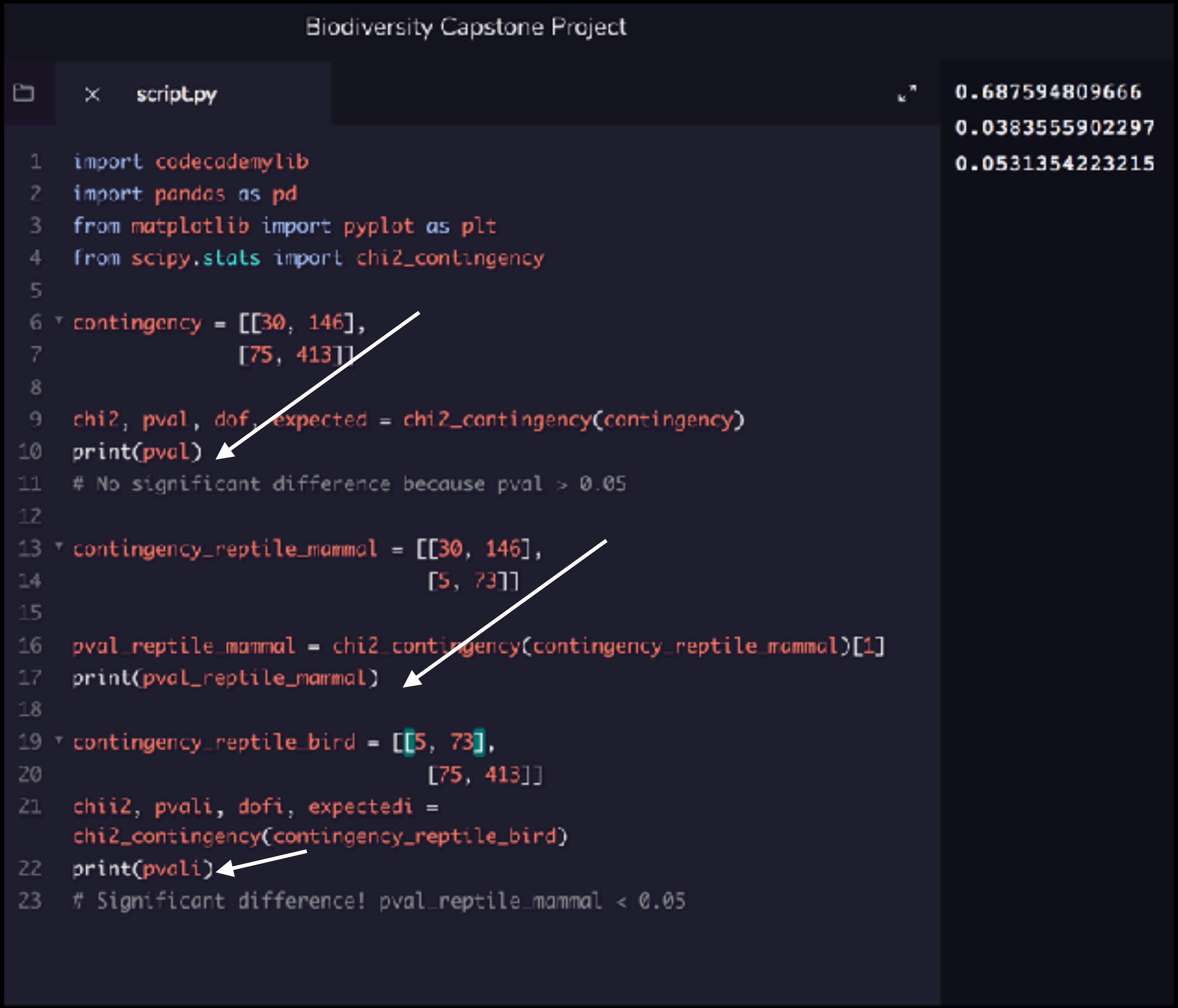
- Conservation_status column has 4 levels of status. However, less than 200 animals have valid value for conservation_status column.

Significance Calculation Between Different species



```
            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
```

- I could see that birds and mammals have higher percentage to be endangered after aggregating the data. However, we should make sure that this observation is valid through Chi-Test.



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

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

chi2, pval, dof, expected = chi2_contingency(contingency)
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)

contingency_reptile_bird = [[5, 73],
                            [75, 413]]
chii2, pvali, dofi, expectedi = chi2_contingency(contingency_reptile_bird)
print(pvali)
# Significant difference! pval_reptile_mammal < 0.05
```

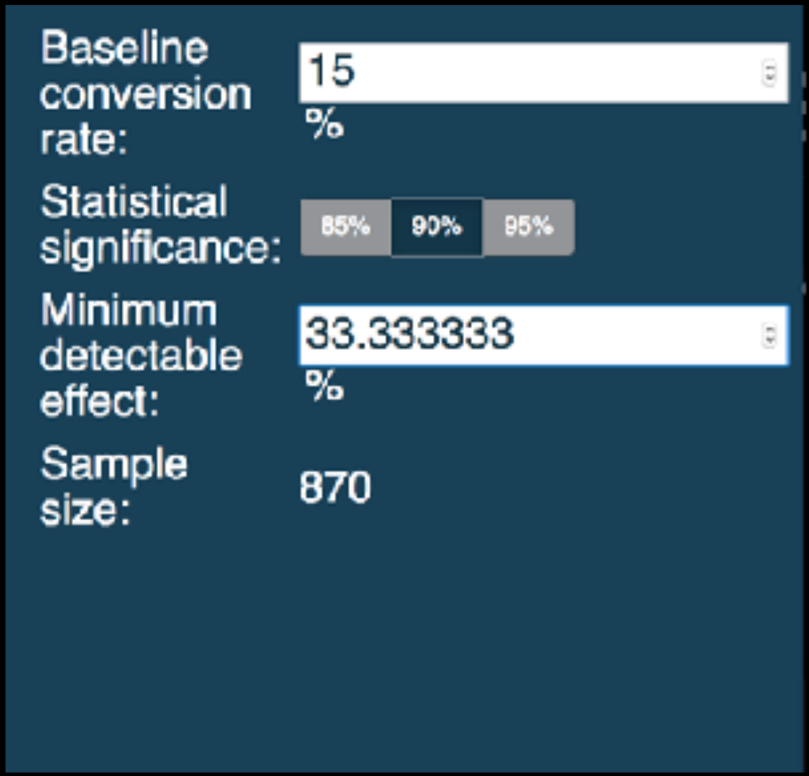
```
0.687594809666
0.0383555902297
0.0531354223215
```

- I did Chi-test between Reptile and Mammal. P-value for this test was 0.03, which is low enough to say that Mammal is more endangered than Reptile. I did additional Chi-test between Reptile and Bird. P-value was quite close to 0.05, but still higher than that, which means that we can’t confidently say that birds are more endangered than reptile.

A recommendation for conservationist

- Mammals seem to struggle the most in national parks. Pay close attention to mammals.

Sample Size Determination for the Foot and Mouth Disease



Baseline conversion rate: 15 %

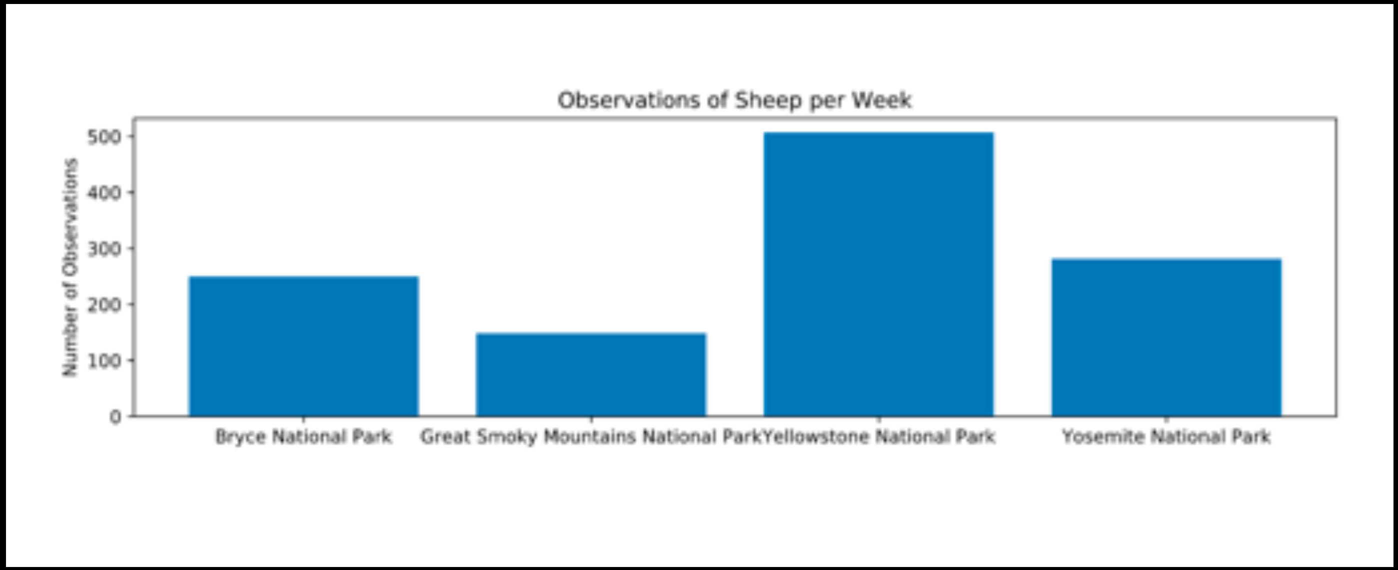
Statistical significance: 85% [✓ 90%] 95%

Minimum detectable effect: 33.333333 %

Sample size: 870

- We know that 15% of sheep had the Foot and Mouth Disease last year (Baseline conversion rate). We want to detect at least 5 percentage, so the minimum detectable effect rate is 100*5/15 = 33.3333. We want 90% confidence level. Based on these three numbers, I could find out we need at least 870 samples with sample size calculator.

+ Graph



- This is the graph showing how many sheep🐑 are observed in each national park.