

Nick Swyter

Introduction to Data Analysis - Codecademy
Biodiversity Capstone Project
June 18, 2018

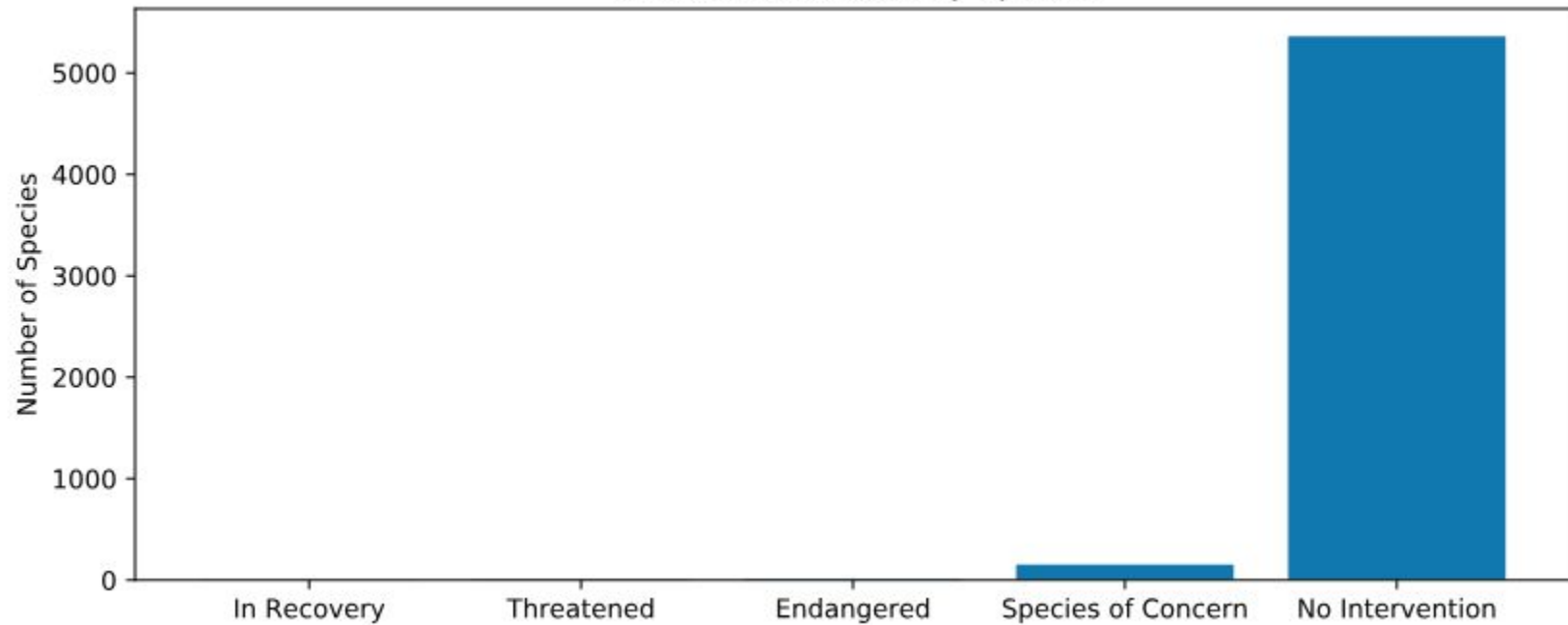
Part I: Observations of species_info.csv

- Dataframe contains 5541 species
 - 79 amphibians
 - 488 birds
 - 125 fish
 - 176 mammals
 - 78 reptiles
 - 333 nonvascular plants
 - 4262 vascular plants
- Dataframe contains 6 conservation statuses
 - 15 species are **endangered**
 - 4 species are **in recovery**
 - 151 are **species of concern**
 - 10 species are **threatened**
 - 5363 species have **no intervention**

```
17 species_type_counts =  
   species.groupby('category').scientific_name.nunique().reset_index()  
18 print(species_type_counts)
```

```
25 species.fillna('No Intervention', inplace  
   = True)  
26  
27 conservation_counts_fixed =  
   species.groupby('conservation_status').scientific_name.nunique().reset_index()  
28 print(conservation_counts_fixed)
```

Conservation Status by Species



More observations of species_info.csv

- **8.86%** of amphibians have protected status
- **15.37%** of birds have protected status
- **8.73%** of fish have protected status
- **17.05%** of mammals have protected status
- **1.50%** of nonvascular plants have protected status
- **6.41%** of reptiles have protected status
- **1.08%** of vascular plants have protected status

```
13 category_pivot =  
    category_counts.pivot(columns='is_protected',  
14                          index='category',  
15                          values='scientific_name')\  
16                          .reset_index()  
17  
18 category_pivot.columns = ['category',  
    'not_protected', 'protected']  
19  
20 category_pivot['percent_protected'] =  
    (category_pivot.protected /  
    (category_pivot.protected +  
    category_pivot.not_protected)) * 100  
21 print(category_pivot)
```

Are certain types of species more likely to have protected status?

- **Yes**
- For example, there is a significant difference between the number of protected mammals and:
 - Reptiles (0.038 chi-squared test)
 - Nonvascular plants (1.48×10^{-10} chi-squared test)
 - Vascular plants (1.44×10^{-10} chi-squared test)
- However, there is not a significant difference between the number of protected mammals and:
 - Birds (0.688 chi-squared test)
 - Amphibians (0.128 chi-squared test)
 - Fish (0.056 chi-squared test)

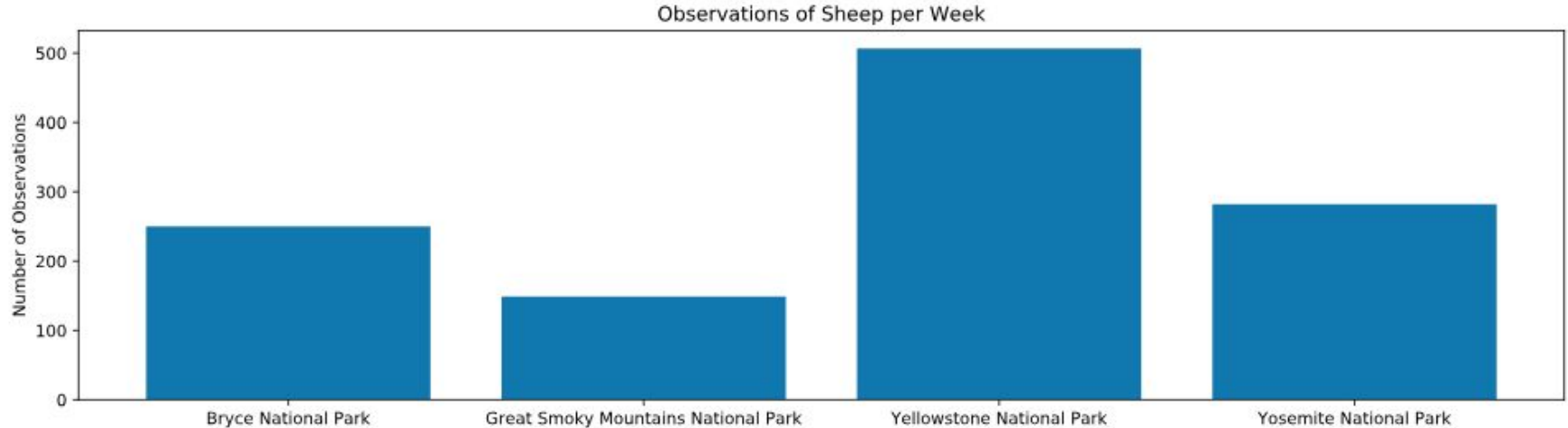
Recommendations for conservationists

- Conservation efforts should deliberately target mammals because:
 - The category has the highest percentage of species with protected status **(17.05%)**
 - There is a statistically significant difference between the number of mammals with protected status compared to reptiles **(0.038 chi-squared test)**, nonvascular plants **(1.48e-10 chi-squared test)** and vascular plants **(1.44e-10 chi-squared test)**

Part II: How scientists determine whether a foot and mouth disease reduction program for sheep is working

- The only information they have is that 15 percent of sheep at Bryce National Park have foot and mouth disease
 - Scientists will use **15 percent** as the baseline
- They want to detect reductions of 5 percentage points
 - Scientists will use **33.33 percent** as the minimum detectable effect ($5/15 = 33.33\%$)
- There is default level of significance
 - Scientists will use **90 percent** as the default level of significance
- Based on these sample size calculations, scientists must observe **870 sheep** to determine whether the foot and mouth reduction program is working

How many sheep are observed at national parks each week?



How long will it take for scientists to determine whether the foot and mouth disease reduction program for sheep is working?

Understanding that the sample size for the experiment is **870 sheep**, it will take scientists:

- **1.72 weeks** to observe sheep at Yellowstone National Park
- **3.48 weeks** to examine sheep at Bryce National Park
- **5.84 weeks** to examine sheep at Great Smoky Mountains National Park
- **3.09 weeks** to examine sheep at Yosemite National Park