Capstone project 2: Biodiversity

Cedric De Leersnijder

October 2018

Description of species_info.csv

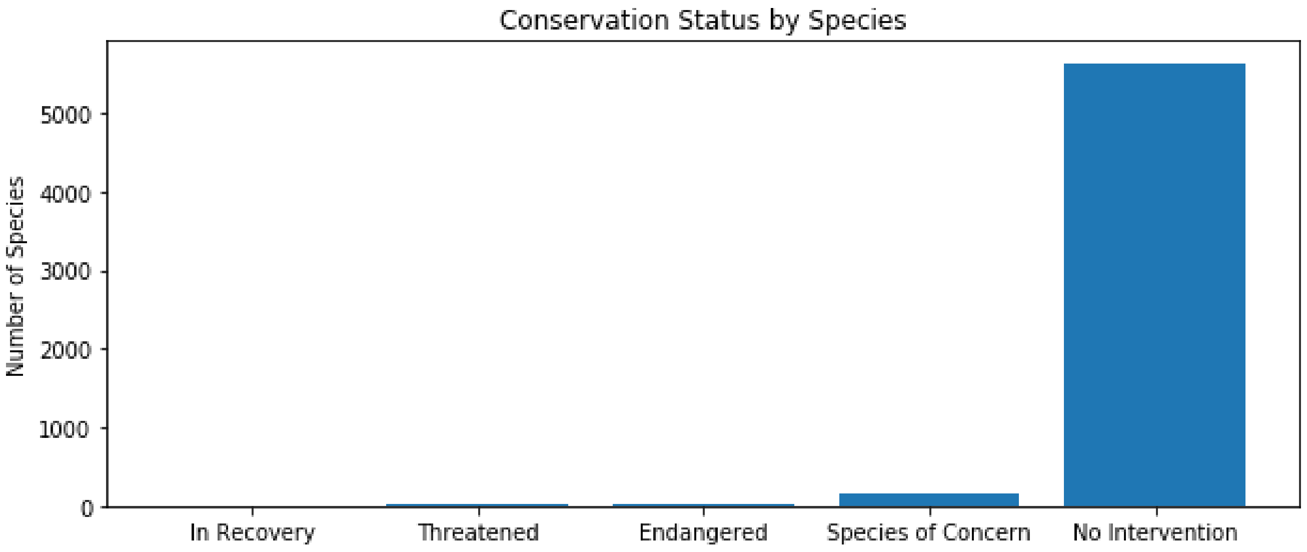
- The csv file represents data about species in National Parks, especially the conservation status (5824 rows and 4 columns)
- Column names are ‘category’, ‘scientific_name’, ‘common_names’ and ‘conservation_status’
- 5824 rows but 5541 unique values in column ‘scientific_name’
- 283 rows from double or multiple entries in column ‘scientific_name’
  - 274 species have 2 entries (Agrostis capillaris, Agrostis gigantea, …)
  - 9 species have 3 entries (Canis lupus, Castor canadensis, …)

Description of species_info.csv

- The species are grouped by 7 categories: Mammal, Bird, Reptile, Amphibian, Fish, Vascular Plant and Nonvascular Plant
- The column conservation_status has several possible values:
  - **Species of Concern**: declining or appear to be in need of conservation.
  - **Threatened**: vulnerable to endangerment in the near future.
  - **Endangered**: seriously at risk of extinction.
  - **In Recovery**: formerly Endangered, but currently neither in danger of extinction throughout all or a significant portion of its range.

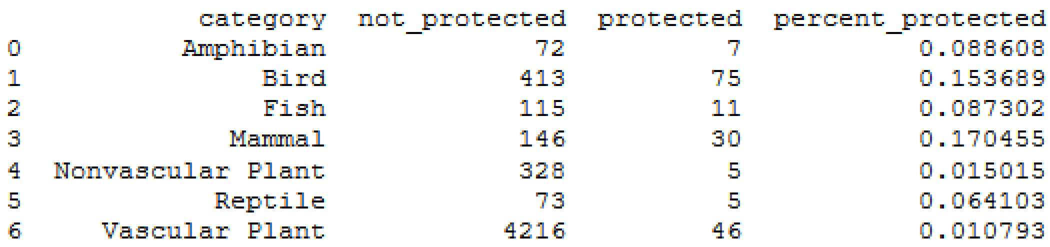
Conservation status of species set

- A lot of species are missing a value for conservation_status (nan). Later on those missing values were filled in with the value ‘**No Intervention**’.
- 180 species need special attention/protection



Significance calculations on endangered status between different categories of species

- Data is **categorical**: protected or 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
```

- Column ‘Protected’ is sum of unique species with conservation status ‘Species of concern’, ‘Threatened’, ‘Endangered’ and ‘In recovery’.
- The higher percent_protected, the more likely to become endangered
- Comparison of 2 or more datasets: **Chi-Square test**

Significance calculations on endangered status between different categories of species

- Testing significance difference between Mammal and Bird
  - My contingency table: contingency = [[146,413],[30,75]]

	Mammal	Bird
Not protected	146	413
Protected	30	75

  - Pval = 0.6875948096661336
  - P-value > 0.5 : difference isn't significant.

Significance calculations on endangered status between different categories of species

- Testing significance difference between Reptile and Mammal
  - My contingency table: contingency = [[73,146],[5,30]]

	Reptile	Mammal
Not protected	73	146
Protected	5	30

  - Pval = 0.03835559022969898 ≈ 0.04
  - P-value < 0.5 : Null-hypothesis rejected, there is a significant difference.
  - Mammals are more likely to be endangered than Reptiles

Recommendation for conservationists, based on previous significance calculations

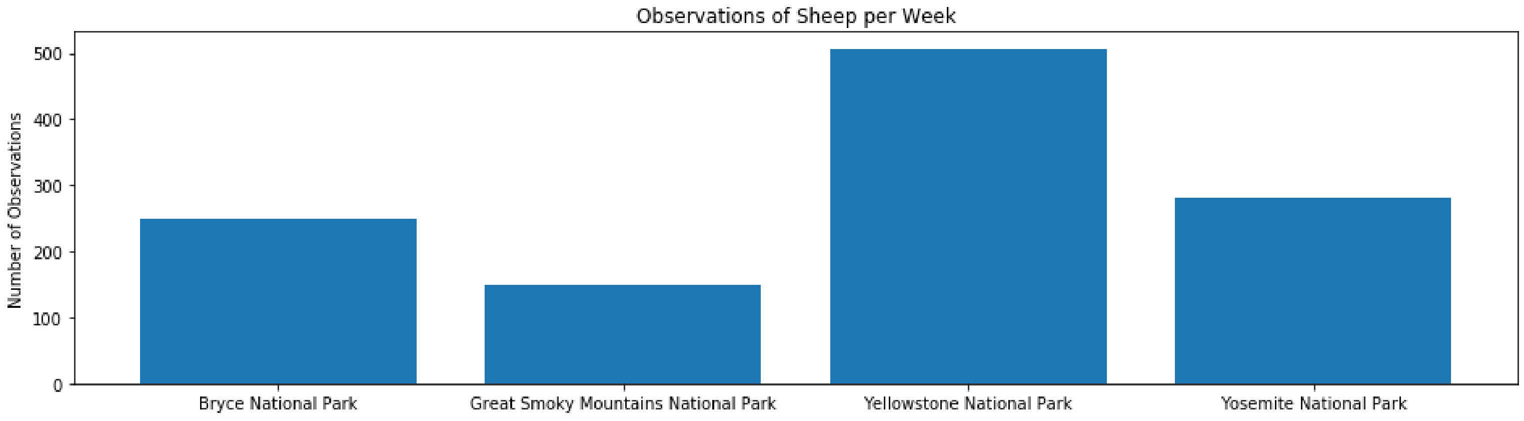
- Mammals and birds need most attention because they are most likely to become endangered (although there is no significance difference between the two categories).
- Vascular and non vascular species are least likely to become endangered.

Sample Size Determination
Food and mouth disease study

- Historical data: 15% of sheep at Bryce National Park have foot and mouth disease
- Yellowstone National Park program to reduce the rate of the disease.
  - They want confidence to detect reductions of at least 5 percentage points (example 15% to 10%).
- Amount of sheep to test?
  - Baseline = 15% (based on historical data)
  - Statistical difference = 90%
  - Minimum Detectable Effect = 33.33% -> (0.05 / 0.15)*100
  - Sample size = 510 -> Calculated with the sample size calculator from Optimizely (as referred in the Slack forum)

Sample Size Determination
Food and mouth disease study

- Note: In one week they observed 250 and 507 sheep respectively in Bryce National Park and Yellowstone National Park (3 same species in each park).



- How many weeks to observe enough sheep?
  - At Bryce National Park: 510.0 / 250 = 2.04 ≈ 2 weeks
  - At Yellowstone National Park: 510.0 / 507 = 1.01 ≈ 1 week