Biodiversity Project

Jorge Rodriguez

Dataset

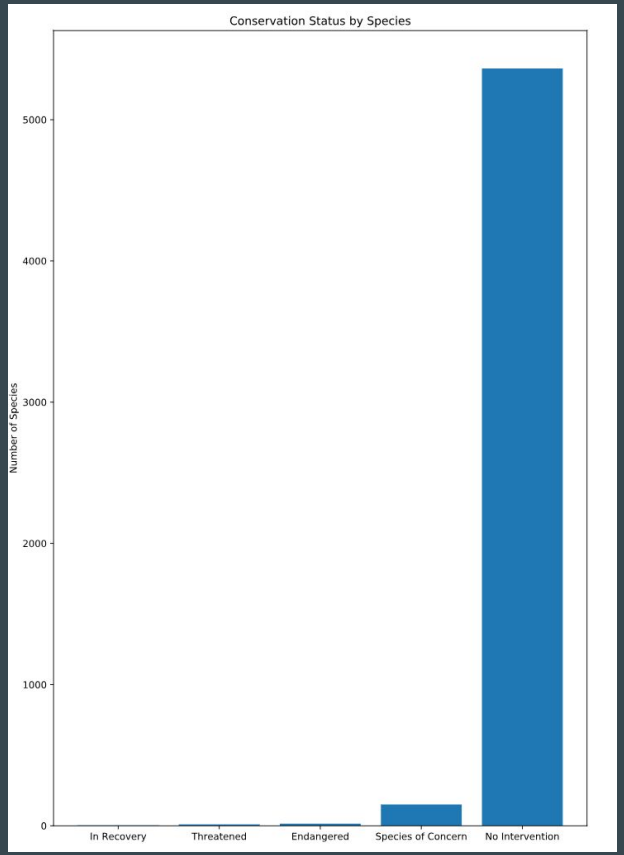
- Animal species in American National Parks, delivered in a .csv file
- Relevant data:
  - Scientific Name
  - Common Name
  - Conservation status
  - Category
- Would like to know: are there any patterns on endangered species?

Distribution of Conservation Status

- Most of species are under ‘No Intervention’, which means they’re not considered under danger
- Species of concern is also a category with a large number of species

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

Distribution of Conservation Status



Endangerement by Species

- A new field called ‘is_protected’ (boolean) was created, the logic behind it is any species not marked as ‘No Intervention’ is considered as false for this field
- Grouping by category and getting the ratio between not protected and protected species under the category, gives us this information:

	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

Endangerement by Species

- From the new table, we can gather that:
  - Mammals and birds are the most protected (endangered)
  - Plants are the least protected
  - Reptiles, fish and amphibians have similar distributions around the middle

Mammals and Birds

- We were given the null hypothesis that mammals are not significantly more endangered than birds
- Running a chi-squared test on the mammal and bird data that we gathered:

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

- Got a p-value of 0.687, which means the difference between the two is not significant. But what about the difference between mammals (17%) and reptiles (6.4%)? This one is significant with a p-value of 0.03 (<0.05)!

Observations from Analysis

- Mammals and birds are categories with a very high number of endangered species
- My suggestion would be further analyzing the species that fall under these categories to find commonalities and suggestion a plan of action

Foot & Mouth Disease

- A new dataset was sent over from the National Parks Service. It contains number of sighhtings of species by park:
  - Scientific_name: of a species
  - Park_name: where the observation was made
  - Observations: number of observations

	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

Foot & Mouth Disease

- We'd like to know observations of sheep. So we filter the data by ocurrence of 'sheep' in the species name and category of 'Mammal'

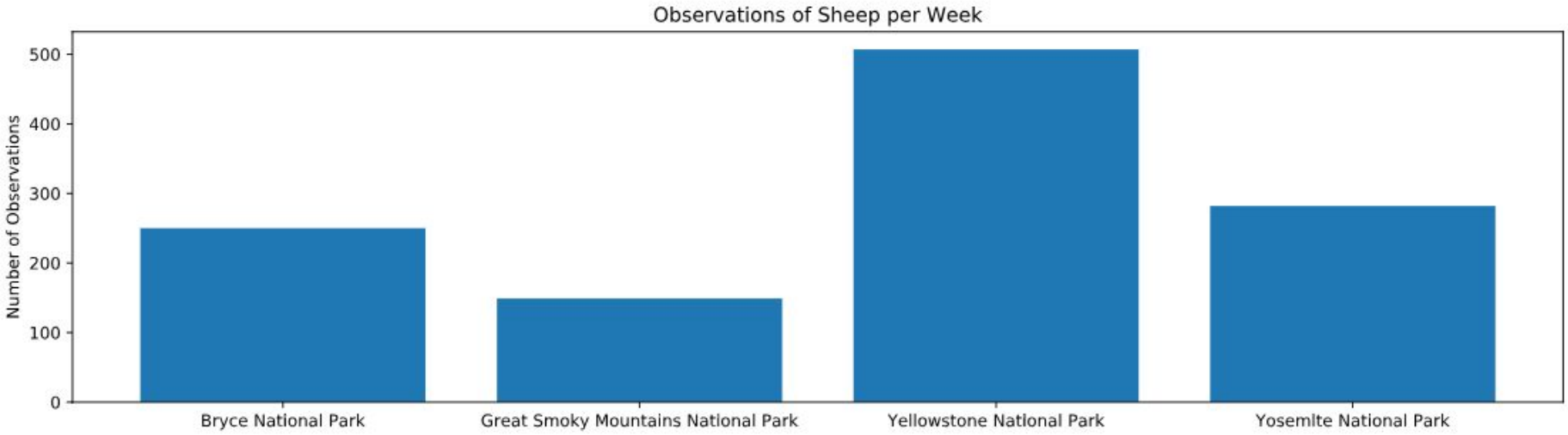
	scientific_name	park_name	observations	category	common_names	conservation_status	is_protected	is_sheep
0	Ovis canadensis	Yellowstone National Park	219	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True
1	Ovis canadensis	Bryce National Park	109	Mammal	Bighorn Sheep, Bighorn Sheep	Species 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

Foot & Mouth Disease

- Then we can group by park to get the total number of sheep observations at each park:

	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

Foot & Mouth Disease



Foot & Mouth Disease

- Now we'd like to know the number of recorded sheep sightings necessary to get a significant number of food and mouth disease ridden sheep
- We need to plug these into a sample size calculator:
  - Baseline: 15 - this is the percent of sick sheep recorded last year at Bryce National Park
  - Minimum detectable effect: 100*5./15 - we want to detect a 5% reduction
  - Level of significance: 90% - this was provided
- With this, calculator gives us a sample size of 870
- We would need 2 weeks of observing at Yellowstone and 4 at Bryce, accordingly