Biodiversity for National Parks

Marie Reyes (mmreyes)
Biodiversity Analyst | 2018.05.17

Capstone Project | Codecademy Introduction to Data Analysis

Overview

- Discover any patterns or themes and which categories of plants and animals in the parks have species that are more likely to become endangered and thus, require a level of protection.

- Determine sample size to study foot and mouth disease among sheep populations at the selected parks.

Species Info for Conservation Protection

- The file, **species_info.csv**, was used as the data source to analyze the details about different species of plants and animals in the national parks.

- The data source includes details about
  - the **categories**,
  - the **scientific names**,
  - and the **common names** of each species
  - along with their **conservation status**.

Data Set

- This is a sample data set from the file, **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), Domesticated Cattle	nan
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	nan
4	Mammal	Cervus elaphus	Wapiti Or Elk	nan

- Additionally, the data set includes the following:
  - 5,824 total records
  - 5,541 unique scientific names
  - 5,504 unique common names

Data Set

- **7 Categories:**
  - Mammal, Bird, Reptile, Amphibian, Fish, Vascular Plant, Nonvascular Plant

- **5 Conservation Status Types:**
  - **Species of Concern** – declining population or appears to be in need of conservation
  - **Endangered** - seriously at risk of extinction
  - **Threatened** - vulnerable to endangerment in the near future
  - **In Recovery** - formerly Endangered, but currently not in danger of extinction throughout all or a significant portion of its inhabitable range.
  - **No Intervention –** No Intervention required

Data Clean-up: Conservation Status

- Data was enriched to include Nan (None) by using **.fillna** to fill in all of the records with Nan as "No Intervention".

```
species.fillna('No Intervention', inplace = True)
```

Before

	conservation_status	scientific_name
0	Endangered	15
1	In Recovery	4
2	Species of Concern	151
3	Threatened	10

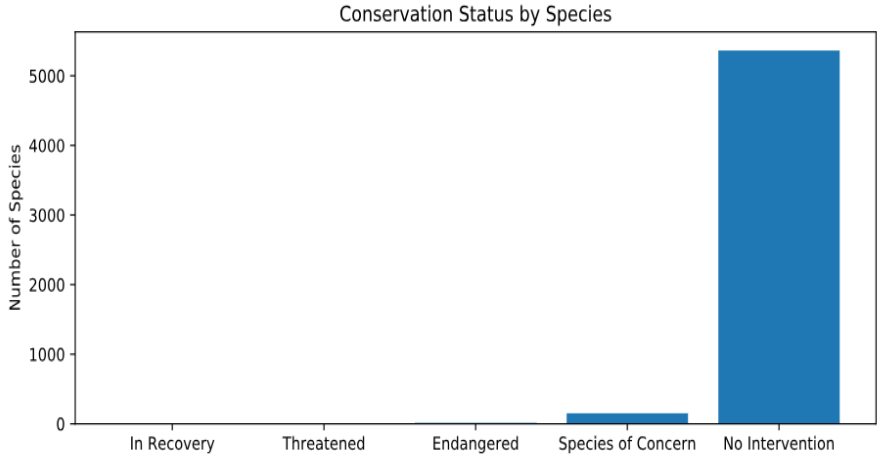
After

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

Plotting Conservation Status

- Most species (5,363) do not require intervention. 151 may be in need of conservation (species of concern) and 25 are close to extinct (endangered or threatened).

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



Investigating Endangered Species

- Are certain types of species more likely to be endangered?
- By using a **pivot**, renaming False and True **columns** to define which species are protected or not, and calculating the **percentage**, we are able to identify the percent of protection of each species and use the data for further analysis.
- Mammals (17%) are more likely to be endangered than birds (15%), but the question still remains "is there significant difference?"

Before

	category	False	True
0	Amphibian	72	7
1	Bird	413	75
2	Fish	115	11
3	Mammal	146	30
4	Nonvascular Plant	328	5
5	Reptile	73	5
6	Vascular Plant	4216	46

After

	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

Chi-Squared Test for Significance – Mammals vs Birds

- Mammals (17%) are more likely to be endangered than birds (15%), but is there significant difference?
- A **significance test** was completed. In this test, our null hypothesis is that this difference is due to chance. To reject the null hypothesis, we need a p-value < 0.05
- By creating a **contingency** table, and running a **chi2_contingency** function, we are able to calculate the **p-value**.
- With this, the null hypothesis CANNOT be rejected. There is **no significant difference** because pval > 0.05

Contingency Table

	protected	not-protected
Mammal	30	146
Bird	75	413

Chi2_ontingency

```
chi2_contingency(contingency)
pval = chi2_contingency(contingency)[1]
print(pval)
```

P-value result

Pval = 0.687594809666

Pval > 0.05

Chi-Squared Test for Significance – Mammals vs Reptiles

- Is the difference between Reptile and Mammal significant? In this test, our null hypothesis is that this difference is due to chance.
- By creating a **contingency** table, and running a **chi2_contingency** function, we are able to calculate the **p-value**.
- There is significant difference since pval < 0.05. The null hypothesis CAN be rejected.

Contingency Table

	protected	not-protected
Mammal	30	146
Reptile	5	73

Chi2_ontingency

```
pval_reptile_mammal =
chi2_contingency(contingency_reptile_mammal)
[1]
print(pval_reptile_mammal)
```

P-value result

Pval = 0.0383555902297

Pval < 0.05

Conclusion: Species Conservation

- There is no significant difference between birds and mammals. Thus, birds and mammals should be protected equally in terms of conservation efforts especially since both are still at the top two that require extra attention.

- There is statistical significant difference between the level of protection that mammals and reptiles need (less than 5% chance that they are the same). Therefore, certain types of species are more likely to be endangered and so, conservation efforts should be treated accordingly.

Observations Data Frame: Sheep Observations

- The file, **observations.csv**, contains recorded sightings of different species at several national parks for the past 7 days.

	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

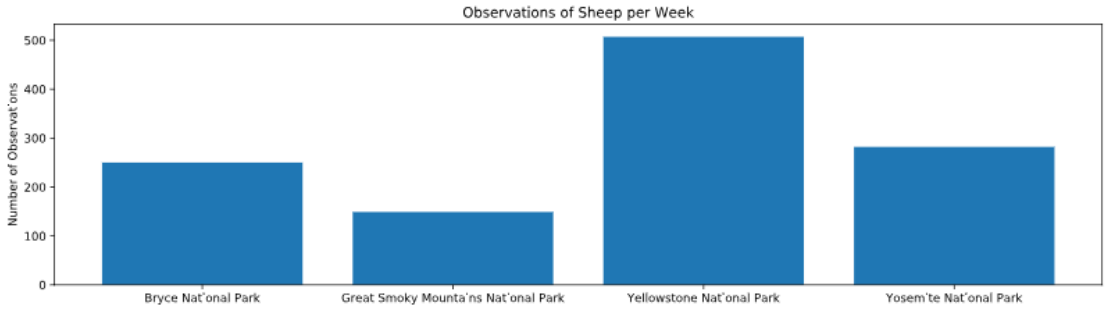
In Search of Sheep

- Because the observation DataFrame only contains the scientific names of species, we also used the species DataFrame to look for any names that refer to sheep.
- **Apply** and **Lambda** functions were used to identify species that were truly sheep in a new column 'is_sheep'

	category	scientific_name	common_names	conservation_status	is_protected	is_sheep
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True
3014	Mammal	Ovis canadensis	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True
4446	Mammal	Ovis canadensis sierrae	Sierra Nevada Bighorn Sheep	Endangered	True	True

Sheep sightings in National Parks

- Observations and Sheep species were **merged** and by using **groupby**, we were able to know the total sheep sightings (across all three species) at each national park.



	park_name	observations	is_protected	is_sheep
0	Bryce National Park	250	2.0	3.0
1	Great Smoky Mountains National Park	149	2.0	3.0
2	Yellowstone National Park	507	2.0	3.0
3	Yosemite National Park	282	2.0	3.0

Foot & Mouth Reduction Effort – Sample Size Determination

- 15% of sheep at Bryce National Park have Foot & Mouth disease.
- Park rangers at Yellowstone National Park have been running a program to reduce the rate of foot and mouth disease at that park.
- To determine if the program is working, an A/B test is required to detect reductions of at least 5 percent.

Sample Size for A/B Test

- Using the calculator, the sample size was determined to identify the number of sheep observations that are required to be made at Bryce and Yellowstone.

- For the calculator:
  - Baseline conversation rate: 15%
  - Minimum Detectable Effect: 33.33%
  - Statistical Difference: 90%

Conclusion

- It was calculated that a sample size of 510 sheep variation for both Bryce and Yellowstone.

- To observe enough sheep
  - ~2 weeks is required at Bryce National Park (510/250 = 2.04 weeks)
  - ~1 week at Yellowstone National Park (1.0059 weeks)