codecademy

Biodiversity

Data Analysis

Fabien Clauss

2019.11.06

Example Table of Contents

1. Data Description

1.1 Data ‘species-info.csv’

The picture beside shows the basic informations about this file.
When calling .value_counts on different columns we note that there might be duplicates (however using .duplicated show that no rows in exactly duplicated).

There are 7 categories (see picture beside).
There are 4 conservation statuses (see picture below).

```
print(species['conservation_status'].value_counts())
```

```
Species of Concern     161
Endangered              16
Threatened              10
In Recovery              4
Name: conservation_status, dtype: int64
```

```
print(species.info())
```

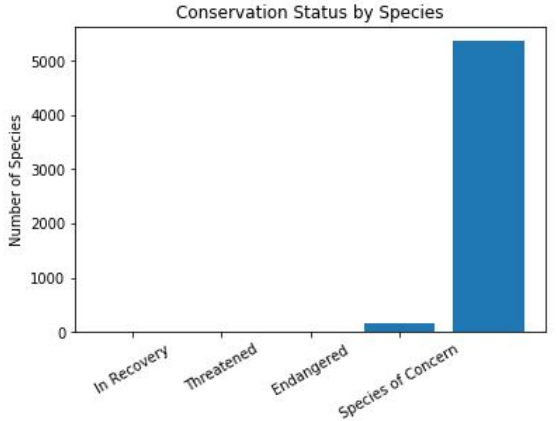
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5824 entries, 0 to 5823
Data columns (total 4 columns):
category                5824 non-null object
scientific_name         5824 non-null object
common_names            5824 non-null object
conservation_status     191 non-null object
dtypes: object(4)
memory usage: 182.1+ KB
None
```

```
print(species['category'].value_counts())
```

```
Vascular Plant        4470
Bird                   521
Nonvascular Plant      333
Mammal                 214
Fish                   127
Amphibian               80
Reptile                 79
Name: category, dtype: int64
```

1.2 Analysis of ‘species-info.csv’

In the data, about 3.4% of the species are categorized as endangered to some level. (see below)
Most of the endangered species are classified as “Species of Concern“, as shown on the bar plot.




```
print(species[['scientific_name', 'conservation_status']].groupby('conservation_status').count())
```

conservation_status	scientific_name
Endangered	16
In Recovery	4
Species of Concern	161
Threatened	10

1.3 Endangered types of species

The analysis and grouping by species type shows that:

- By number, the birds are the most endangered, followed by the vascular plants
- By percentage, the Mammal and then birds are the most endangered

The chi contingency of about 0.69 shows that the difference between the percentage of endangered mammal and birds is not significant.
When the chi contingency of 0.038 between mammals and reptiles however is significant.

is_protected	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

	category	not_protected	protected	percent_protected
0	Amphibian	72	7	8.86
1	Bird	413	75	15.37
2	Fish	115	11	8.73
3	Mammal	146	30	17.05
4	Nonvascular Plant	328	5	1.50
5	Reptile	73	5	6.41
6	Vascular Plant	4216	46	1.08

```
category_counts = species.groupby(['category', 'is_protected']).scientific_name.nunique().reset_index()
```

```
category_pivot = category_counts.pivot(index = 'category', columns = 'is_protected', values = 'scientific_name').reset_
```

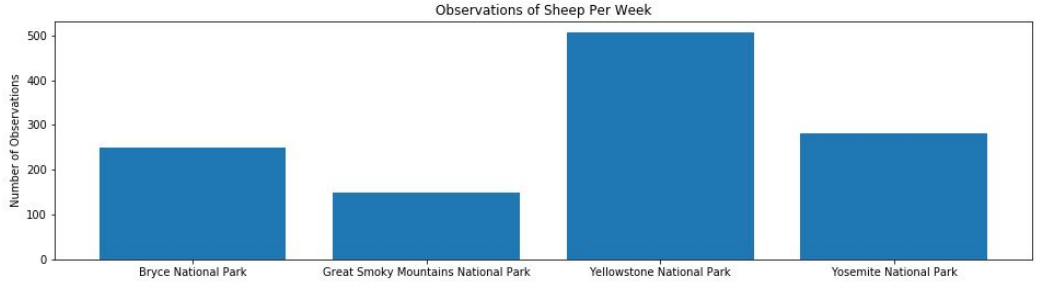
```
category_pivot['percent_protected'] =
        round(category_pivot['protected'] / (category_pivot['protected'] + category_pivot['not_protected']), 4) * 100
```

1.4 Analysis of ‘observations.csv’

After filtering out all the sheep observations, we get a fairly small table with all the observations of sheeps.
By grouping by park_name and plotting the data we see that most observations by far happened in the Yellowstone Park, the least in the Great Smoky Mountains Park.

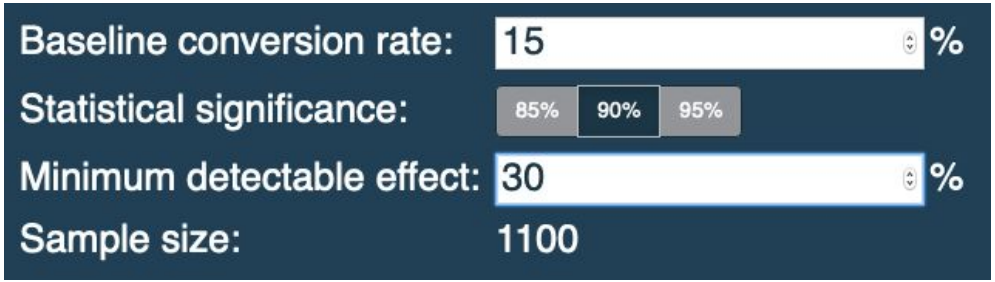
```
from IPython.display import display, HTML
sheep_observations = pd.merge(sheep_species, observations)
display(HTML(sheep_observations.to_html()))
```

	category	scientific_name	common_names	conservation_status	is_protected	is_sheep	park_name	observations
0	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True	Yosemite National Park	126
1	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True	Great Smoky Mountains National Park	76
2	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True	Bryce National Park	119
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True	Yellowstone National Park	221
4	Mammal	Ovis canadensis	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True	Yellowstone National Park	219
5	Mammal	Ovis canadensis	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True	Bryce National Park	109
6	Mammal	Ovis canadensis	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True	Yosemite National Park	117
7	Mammal	Ovis canadensis	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True	Great Smoky Mountains National Park	48
8	Mammal	Ovis canadensis sierrae	Sierra Nevada Bighorn Sheep	Endangered	True	True	Yellowstone National Park	67
9	Mammal	Ovis canadensis sierrae	Sierra Nevada Bighorn Sheep	Endangered	True	True	Yosemite National Park	39
10	Mammal	Ovis canadensis sierrae	Sierra Nevada Bighorn Sheep	Endangered	True	True	Bryce National Park	22
11	Mammal	Ovis canadensis sierrae	Sierra Nevada Bighorn Sheep	Endangered	True	True	Great Smoky Mountains National Park	25



1.5 Foot and Mouth disease

Using the sample size calculator, we see that a minimum of 1100 sheep observations are necessary to detect a 5% reduction in the disease.
Knowing that at Bryce National Park about 250 sheep observations occurs weekly, we conclude that 5 weeks (4.4 precisely) of observations will be necessary.



2. Conclusions and recommendations

- Over 3% of the species listed in the data are in danger, the vast majority of them are classified as “Species of Concern”
- Mammal and fish are the most often classified as endangered to some level with over 15% each. The question here would be if these categories are so much more endangered of just more often observed and therefore their decline is easier to notice. Another factor to take into account is the small amount of mammal and birds listed (176 and 488 respectively) compare for example to the vascular plants with 4262 species listed.
- Concerning the foot and mouth disease, since we do not have further information about the observations, it could be useful to analyse the number of observations for each park adjusted for the size of the park. This way the program can be run where the minimal number of observations can be reach in a minimal amount of time and with the least resources.