codecademy / PRO
INTENSIVE

Introduction to Data Analysis

Capstone Project

by Julien Best

Structure

- Presentation of the data and the analysis
- Results and recommendations for conservationists
- Sample size determination: Foot and mouth disease study

The data

The project is based on two comma seperated values (.csv): species and observations.

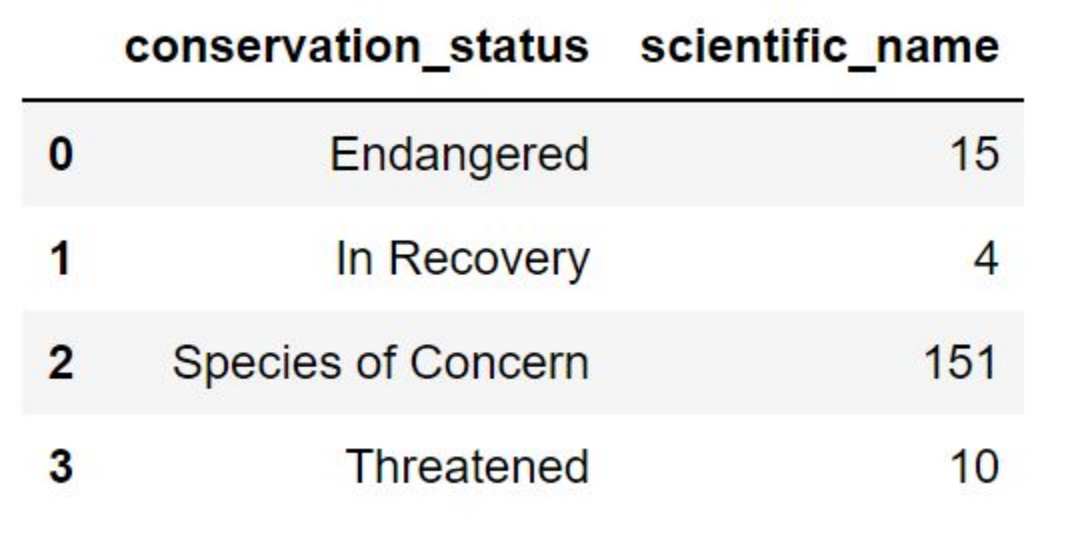
Species.csv:

Contains following categories:

- category: The kind of the species (for example: Bird)
- scientific name: The latin name of the species
- common names: The common name of the species
- conservation status: The status of the conservation: Is the species in danger?

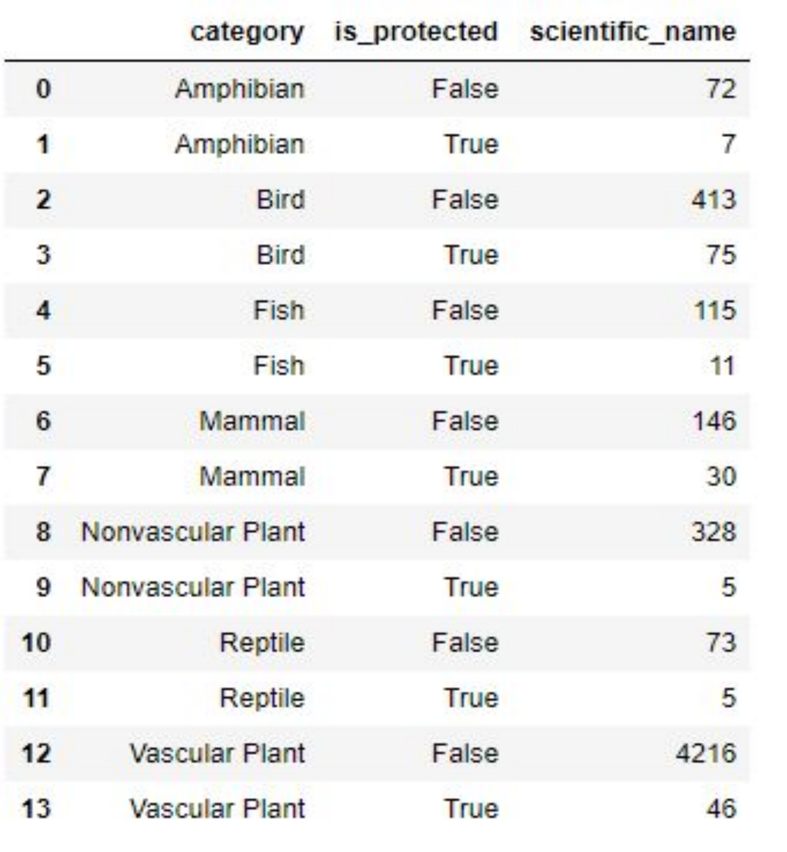
I will talk about the results of the analysis regarding this csv first, then talk about the second part of the data

First, we are interested in the deviation of the conservation status. At the bottom one can see a table describing, how many species are affected by a specific conservation status



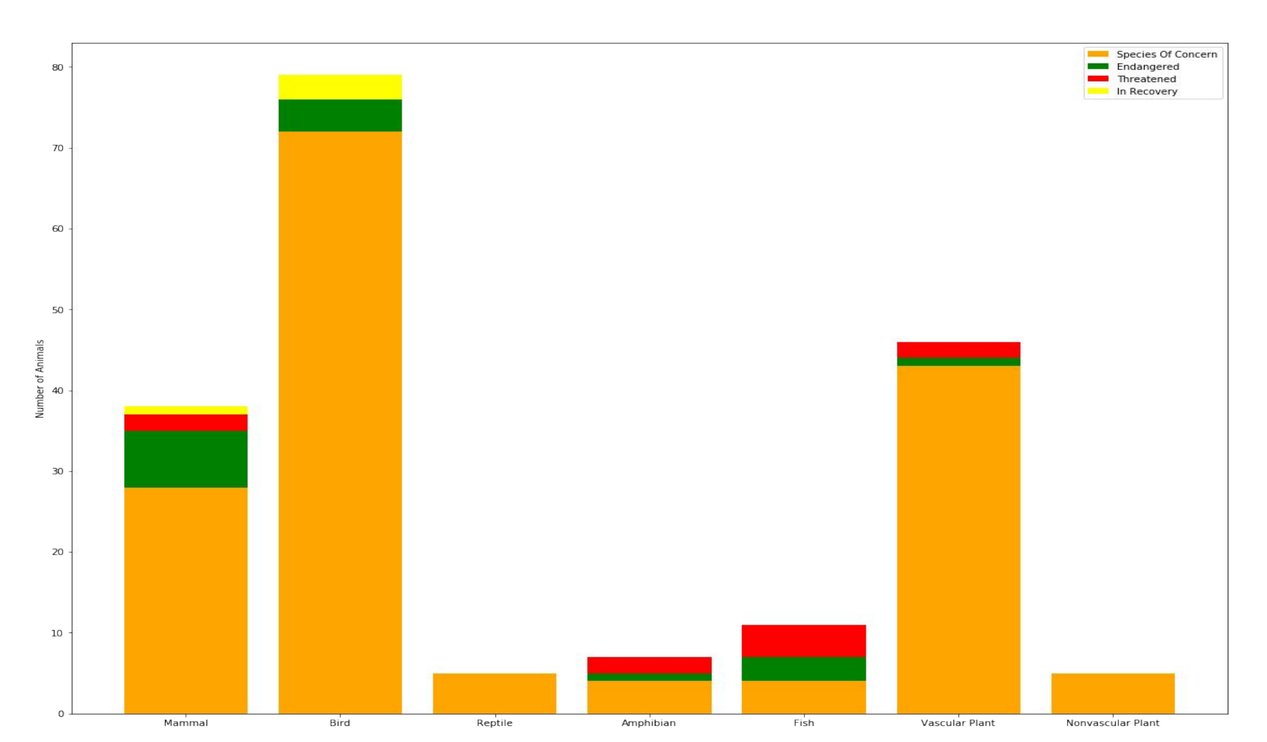
	conservation_status	scientific_name
0	Endangered	15
1	In Recovery	4
2	Species of Concern	151
3	Threatened	10

If we are not only interested in how many species have which conservation status but also in the more accurate information about the categories we can look at this table



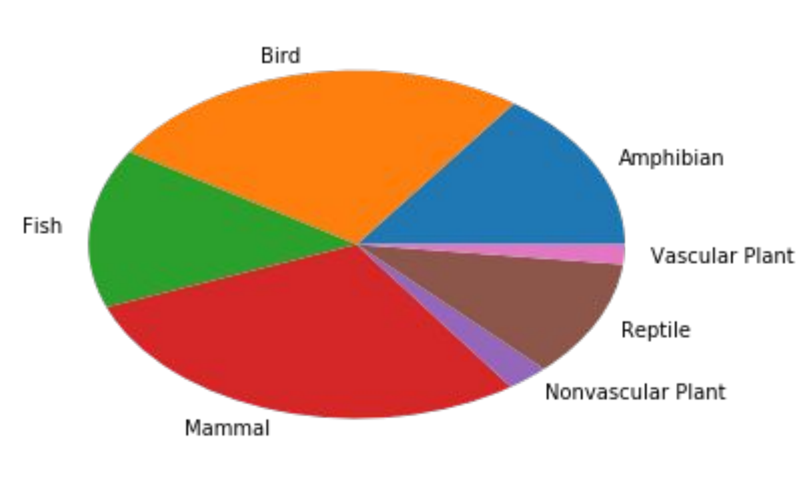
	category	is_protected	scientific_name
0	Amphibian	False	72
1	Amphibian	True	7
2	Bird	False	413
3	Bird	True	75
4	Fish	False	115
5	Fish	True	11
6	Mammal	False	146
7	Mammal	True	30
8	Nonvascular Plant	False	328
9	Nonvascular Plant	True	5
10	Reptile	False	73
11	Reptile	True	5
12	Vascular Plant	False	4216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

To get a better idea of the deviation of the conservation status deviation I decided to also do a stacked bar chart



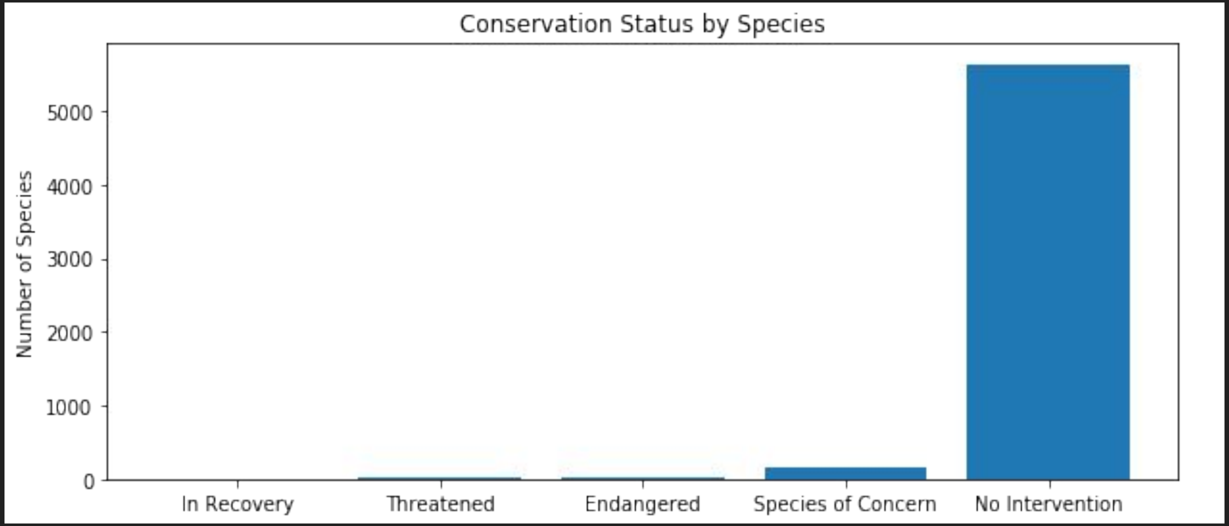
You can see that if you look at these 4 categories, most of the species are a species of concern. An important thing to notice is that the Fish is the most affected category if you look at the number of threatened and endangered fish in comparison to the number of fish in the category “Species of Concern”!

But not only Fish is not protected enough. Here is a pie chart describing the percentage of the species of each category being protected

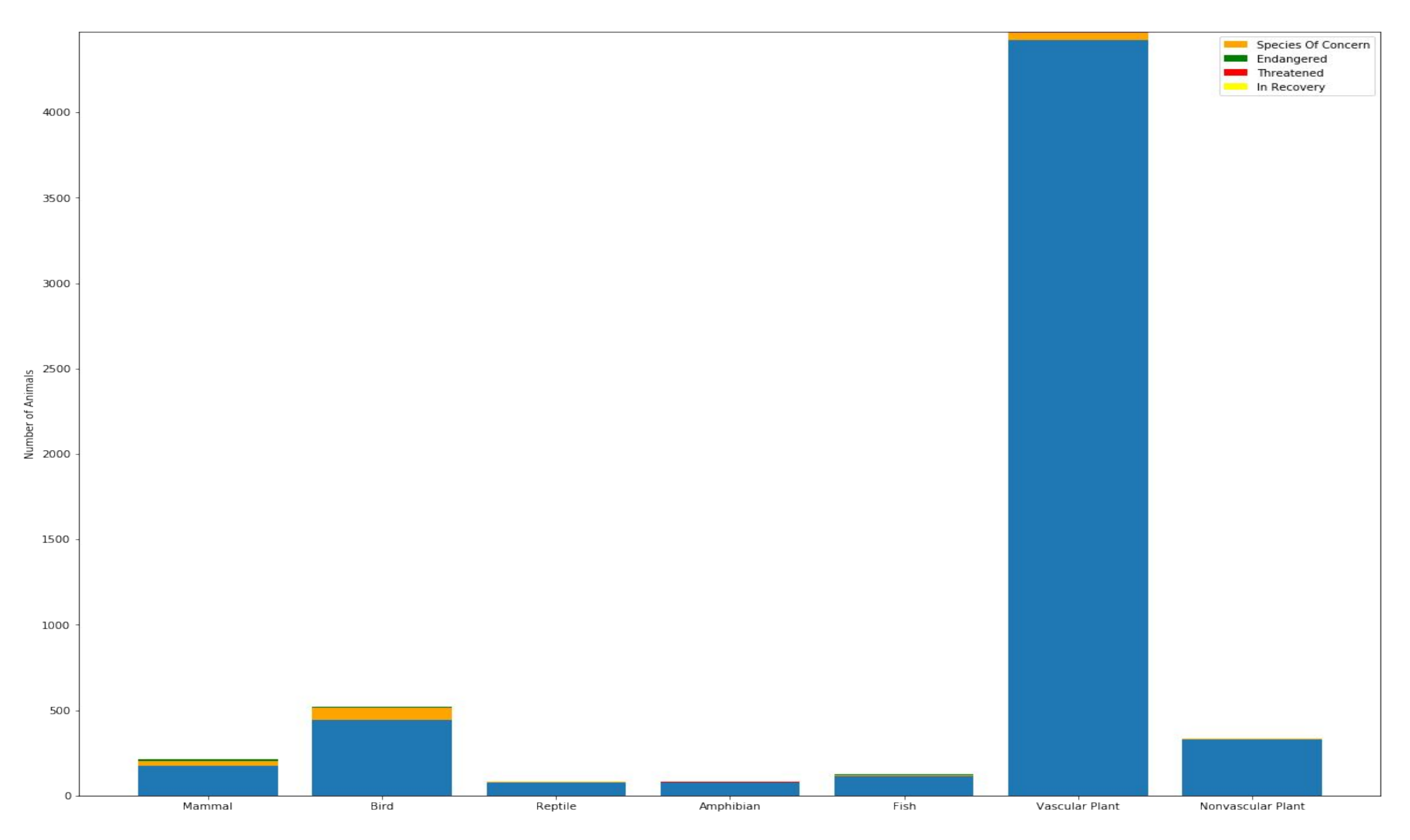


For example: In the category Vascular Plant are the fewest protected and in the category Mammal are the most species protected

If we also look at the species with no intervention, the deviation looks like this:



If we look again at a stacked bar chart it gets obvious how many species are in the category “No Intervention”. Some other categories are not even visible anymore



Recommendations for conservationists

- They should look out more for species that are not protected and in danger. Especially fish are obviously not cared enough for.
- Regarding the conservation status “In Recovery” which takes a *very low* percentage one can assume that the success rate of interventions can be optimized

observations.csv

contains following categories:

- scientific name
- park name
- observations

Sample size determination

Sample size determination is a calculation on how many samples do you need in order to have a confident result

To calculate that you need a few information

- The baseline conversion rate describes for example the ratio between the amount of people visiting a website and the amount of people buying something. In the foot and mouth disease study this ratio is 15% so that means that 15 out of 100 sheep have a foot or mouth disease
- The minimum detectable effect is dependent on the baseline and is calculated by multiplying the quotient of the difference goal and the baseline conversion rate with 100. In that case it’s 33.3
- The statistical significance is most of the time between 85% and 95% to get a trustworthy result