Maxime Laurenty - ***Introduction to Data Analysis Codecademy course***

Capstone Project: Biodiversity for the National Parks

Data description - *Species_info* (1)

The data in *species_info.csv* presents animals found in the national parks with the following characteristics : category, scientific name, common_name and conservation status.

There are 5824 animals (rows) among 5541 species.

The different categories are :

Mammal	Bird	Reptile	Amphibia	Fish	Vascular Plant	Nonvascular Plant

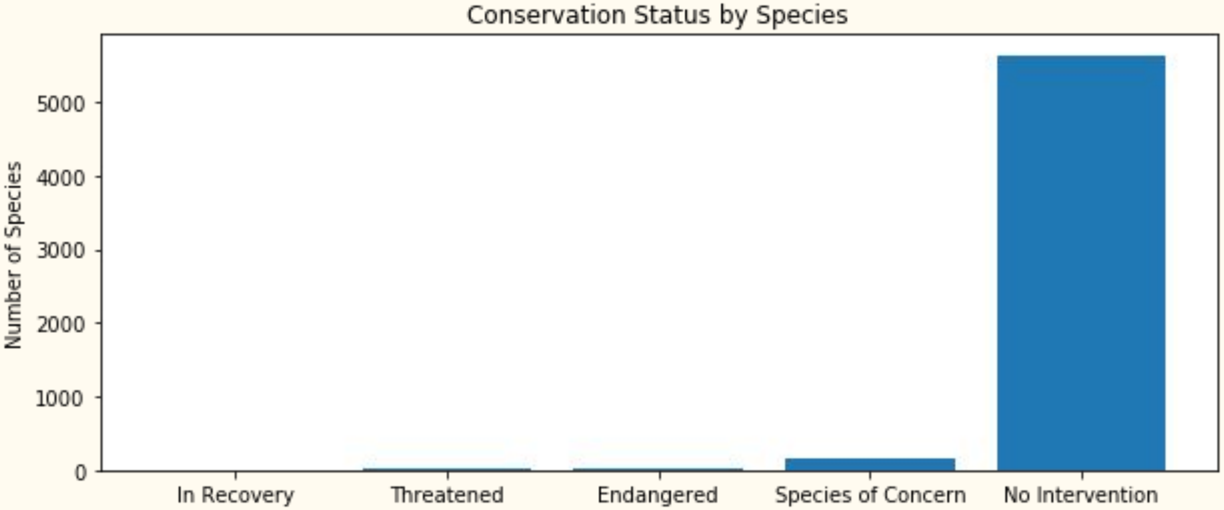
The different active conservation status are :

Species of Concern	Endangered	Threatened	In Recovery

Data description - *Species_info* (2)

The repartition among the different conservation status is the following :

Endangered	16
In Recovery	4
No Intervention	5633
Species of Concern	161
Threatened	10



Data description - *Species_info* (3)

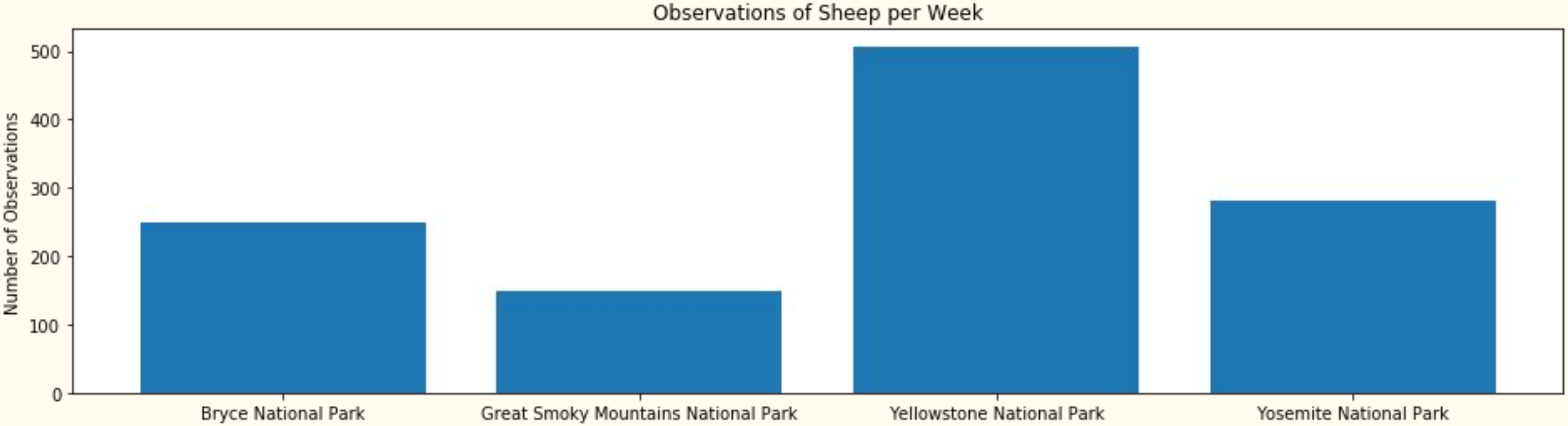
The repartition between protected & unprotected species among the different categories are:

category	not_protected	protected	percent_protected
Amphibian	72	7	0.088
Bird	413	75	0.153
Fish	115	11	0.087
Mammal	146	30	0.170
Nonvascular Plant	328	5	0.015
Reptile	73	5	0.064

Data description - *Observations*

The data in *observations.csv* presents the number of observations of each species in several national parks.

We focused on studying sheep:



(The different sheep species are: *Ovis aries*, *Ovis canadensis & Ovis canadensis sierrae*)

Endangered species - Significance calculations

To know if the repartition of endangered status (presented in slide 4) among the different categories can be compared, we performed significance tests.

This data is categorical ("yes/no" type) and we wanted to compare different samples, hence *chi squared test* was appropriate.
We choosed a pvalue threshold at 0.05. The difference observed between:
- mammal (17%) and bird (15%) **is not** significant. (pvalue = 0.69>0.05).
- mammal (17%) and reptile (6.4%) **is** significant. (pvalue = 0.038).
- mammal & amphibian **is not** significant (pvalue = 0.127).
- mammal & fish **is not** significant (pvalue = 0.056).

Endangered species - recommendation

According to the data, mammal are the ones that are most likely to be endangered. However the sizes of our categories are not big enough to be confident that bird, amphibian and fish are not as likely to be endangered as mammals.

Conservationnists should thus focus primarily on preserving mammal, bird, amphibian and fish.

Reptile and plants can be set aside if the conservationnists have limited means.

Foot & mouth disease study

Based on the provided information, we use the following parameters to determine our sample size :
Baseline conversion rate = 15%
Statistical significance = 90%
Minimum detectable effect = 5/15 = 33%

This results in a sample size of 890 sheeps to observe.

According to the number of observation in each park, this would require 2 weeks at Yellowstone National Park & 4 weeks at Bryce National Park.