Capstone Project Submission:
Biodiversity for the National Parks

Nelson Fleig

Created for Codecademy's Intensive Course "Introduction to Data Analysis"

Project Objectives

The aim of this project is to gain insights about the biodiversity of parks in the United States.

Methods employed were significance testing, sample size determination, and data visualization.

The work on the data was done on a Jupyter Notebook (Python).

Dataset Description

Dataset Description:

▶ The dataset in the CSV file **species_info.csv** contains data about different species in different national parks in the United States

Variables	Datatypes
category	String
scientific_name	String
common_names	String
conservation_status	String

Total # of Records = 5824

▶ The dataset in the CSV file observations.csv contains data about recorded sightings of different species at several national parks for the past 7 days.

Variables	Datatypes
scientific_name	String
park_name	String
observations	Integer

Total # of Records = 23296

Significance Testing

Two Chi Square tests were done to compare if there is a significant difference between the endangerment risk of species categories. This particular test was chosen because of the categorical nature of the data.

Species categories compared:

1) Mammal and Bird

2) Mammal and Reptile

- The **null hypothesis** is that there is no significance difference between the species categories

Significance Testing

Test results

The difference between categories is significant if the P value is less than 5%

1) Mammal and Bird: 44.59% => Not significant

2) Mammal and Reptile: 2.34 % => Significant

Recommendations

The significance calculations showed that there is a significance difference between the likelihood of mammals of becoming endangered compared to reptiles.

Some suggestions for conservationists would be:

1) Focus more research on the health of mammals in the park

2) Research if mammals are suffering from some kind of disease

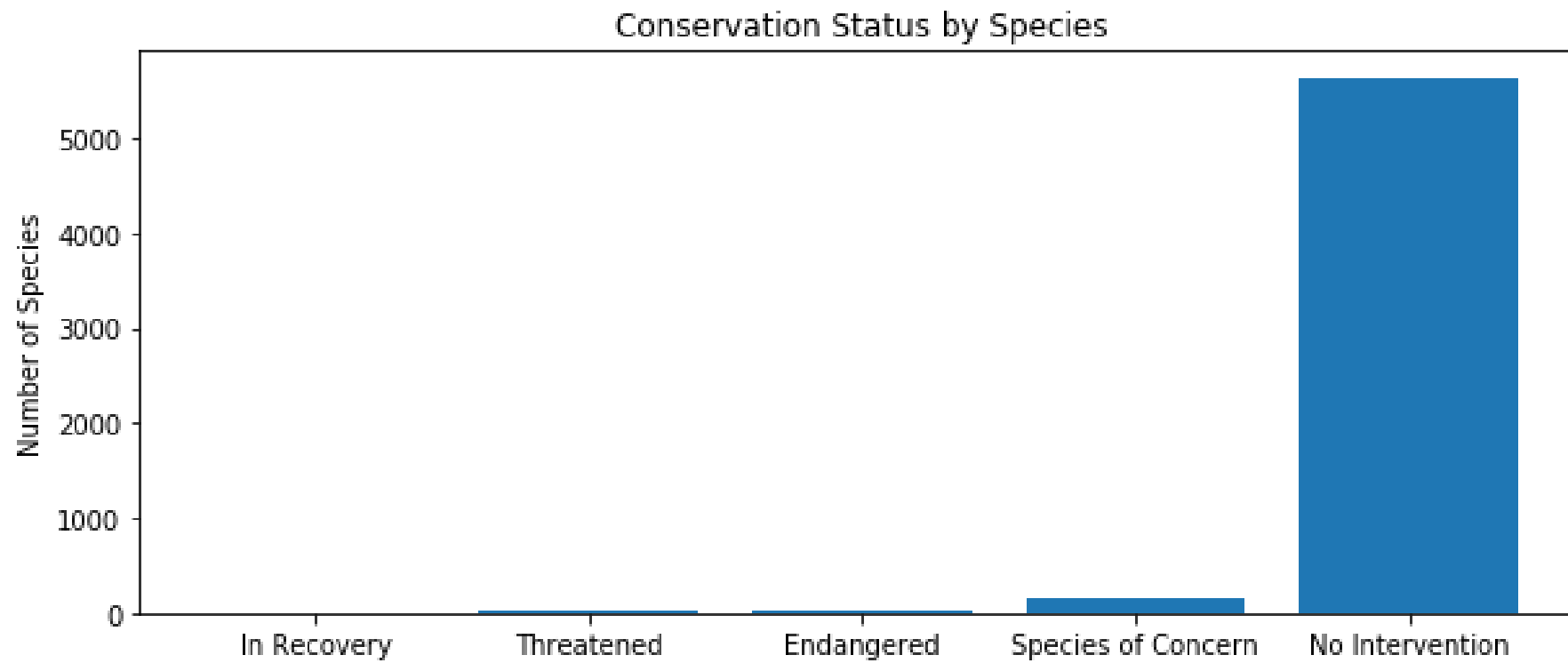3) Find if there is a possible correlation between the conservations status of mammals and climate change

Sample Size Determination

Baseline conversion rate: 15 %

Statistical significance: 85% **90%** 95%

Minimum detectable effect: 33 %

Sample size: 890

The sample size determination calculator shows that a sample size of 840 is necessary to be able to determine with a confidence of 90% that the number of sheep with disease has gone down by 5% (a detectable effect of 33.3% from the baseline)

In order to reach the required sample size, it would take 4 weeks of sampling in Bryce National Park or 2 weeks in Yellowstone National Park.

Graphs

Conservation Status by Species

Graphs

Observations of Sheep per Week