Introduction

The goal of this project is to show one of the ways to analyze biodiversity data from the National Parks Service. Here we will observe some part of species that are live in different national park locations.

This project was created to overview peculiarities of how much species in 4 National Parks are endangered or threatened and which of them are in need to be protected. On the ground of this work specialists, who working in these parks and other related professionals can draw some conclusions for better understanding the size of the problem and which ways could be helpful to protect endangered/threatened species in the future. Results of this analysis as well could serve an impulse for further analytical activities to continue exploring other important matters, connected to questions of preservations biodiversity of national parks in U.S.

In this study, an analysis of two data sets was carried out, which contain information about different species of animals and plants inhabiting 4 national parks of the USA.

This project will define the scope, examine, organize, visualize the data, and aim to interpret the results of the analysis.

Project goals

This project addresses issues of interest to National Park Service biodiversity analysts. The US National Park Service is committed to creating conditions for the conservation and survival of park species. Some species, wards of the service, are at risk of extinction. In order to maintain the level of biodiversity in their parks, professionals need accurate and complete visual information that reflects the current state of affairs. Therefore, the main priority results of data research for the analyst will be to understand the total number of species, how they are distributed in the parks and their current conservation status. The main questions answered during the analysis of available data are the following:

1. Which species inhabiting 4 US national parks?
2. How large is the number of endangered species? How many endangered species are there, relative to the total number of all species in general, in national parks?
3. How many species in national parks have this or that nature conservation status?
4. What categories do endangered species belong to?
5. Do all parks have endangered species?
6. Which specific species of animals and plants in national parks are endangered?
7. What was the total number of species by category that were studied?

8. Is the biodiversity of the parks under threat as such?

Data sources:

Both Observations.csv and Species_info.csv was provided by Codecademy.com.

Note: The data for this project is *inspired* by real data, but is mostly fictional.

This project contains two datasets that come with the package. The first csv file contains information about each species, and the second contains species observations with park locations. This data will be used to analyze the project objectives.

Previously, work was carried out to evaluate the data for their suitability for analysis and obtaining conclusions.

There are two datasets supplied with this project. The first CSV file contains information about different species, and the second - observations of these species, taking into account the location of the parks. These data were used to analyze the project tasks.

The total number of records in the observations file is: (23296, 3)

The total number of entries in the species shape file is: (5824, 4)

Methods of Analysis

Data analysis was performed using descriptive statistics and data visualization methods. Visualization allows you to visualize data and present digital information in a more accessible form. In turn, this facilitates the task of showing relationships between different categories and certain patterns that are difficult to calculate in a tabular presentation. Statistical inference will also be used to test whether the observed values are statistically significant. The current analysis includes the following methods:

1. Calculation

2. Distribution

3. Connection between species

4. Calculation of nature protection status of species

5. Distribution by species in parks

6. Comparison

Assessment of work done

After completing the analysis and description of the obtained conclusions, before creating a general summary, a comparison of the obtained results was also carried out and the obtained results were compared with the stated goals of the project. In this way, it was determined whether the results of the analysis correspond to the questions posed at the beginning of the analysis. In addition, at this step, possible areas of application of the obtained conclusions were considered and several additional issues were described that go beyond the scope of this analysis and require additional study, including by collecting and analyzing a larger amount of relevant data in related areas.

Visualization and Analysis

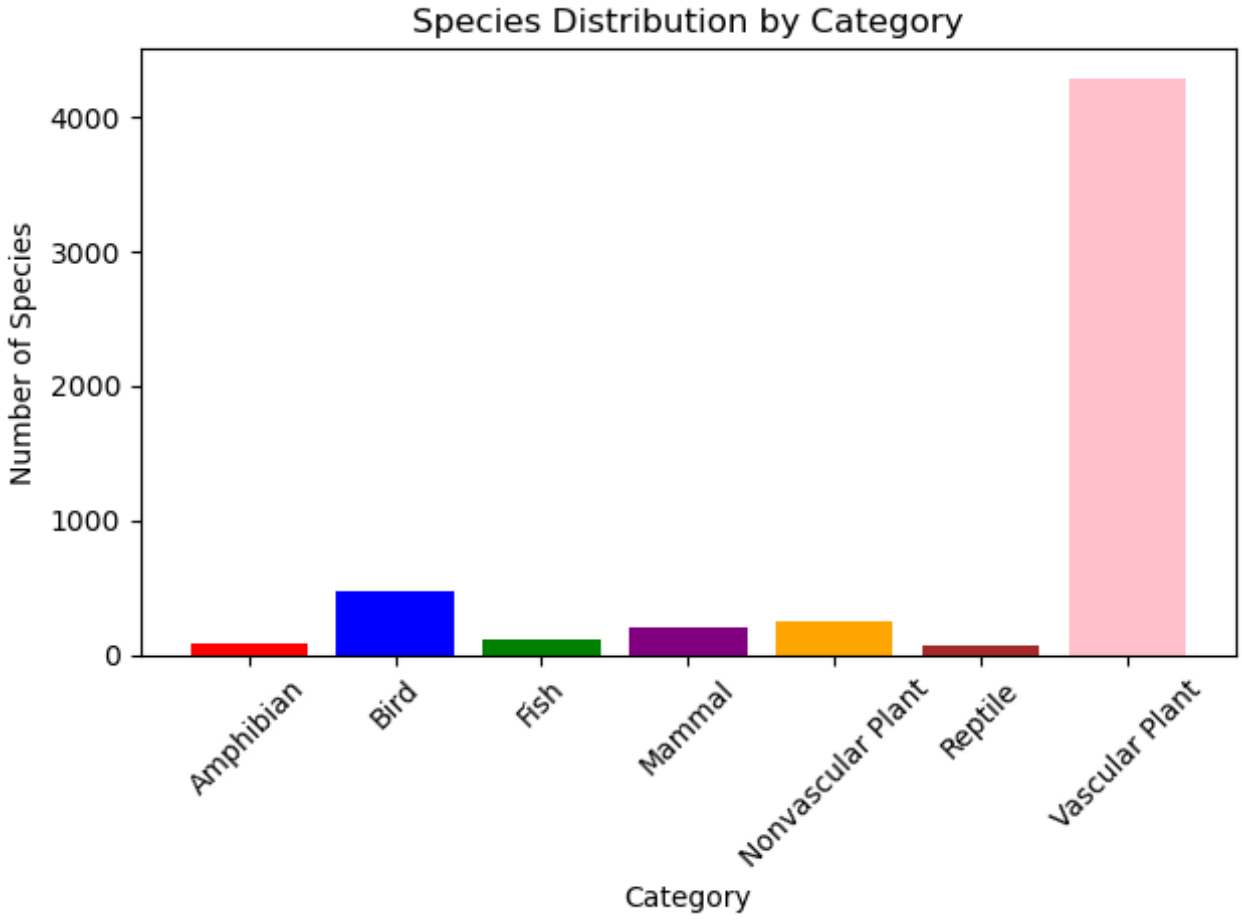
Here we can see, two documents with a total number of entries were provided for analysis in the dataset - observations: (23296, 3); and species: (5824, 4)
The species file contains 5541 unique records of species and there is also a column in which the presence of conservation statuses in some of the species that were affected is noted:
Number of species: 5541
Species, in turn, are divided into 7 categories in this file.
Names of categories:

1. Mammal
2. Bird
3. Reptile
4. Amphibian
5. Fish
6. Vascular Plant
7. Nonvascular Plant

In the process of analysis, the number of creatures in each category was also counted. This allows us to feel the order of the numbers and will help us better navigate in our further calculations.

Number of creatures in each category

Amphibian	80
Bird	521
Fish	127
Mammal	214
Nonvascular Plant	333
Reptile	79
Vascular Plant	4470
In total	**5824**



The next thing to pay attention to when analyzing the available information in the datasets is what nature protection statuses we are dealing with.

Unique conservation statuses: 'No status' (or not affected anyhow), 'Species of Concern', 'Endangered', 'Threatened', 'In Recovery'.

It is also important to note that in this analytical work we will focus the most on only one group of species, namely those that have the status of 'Endangered'. Because

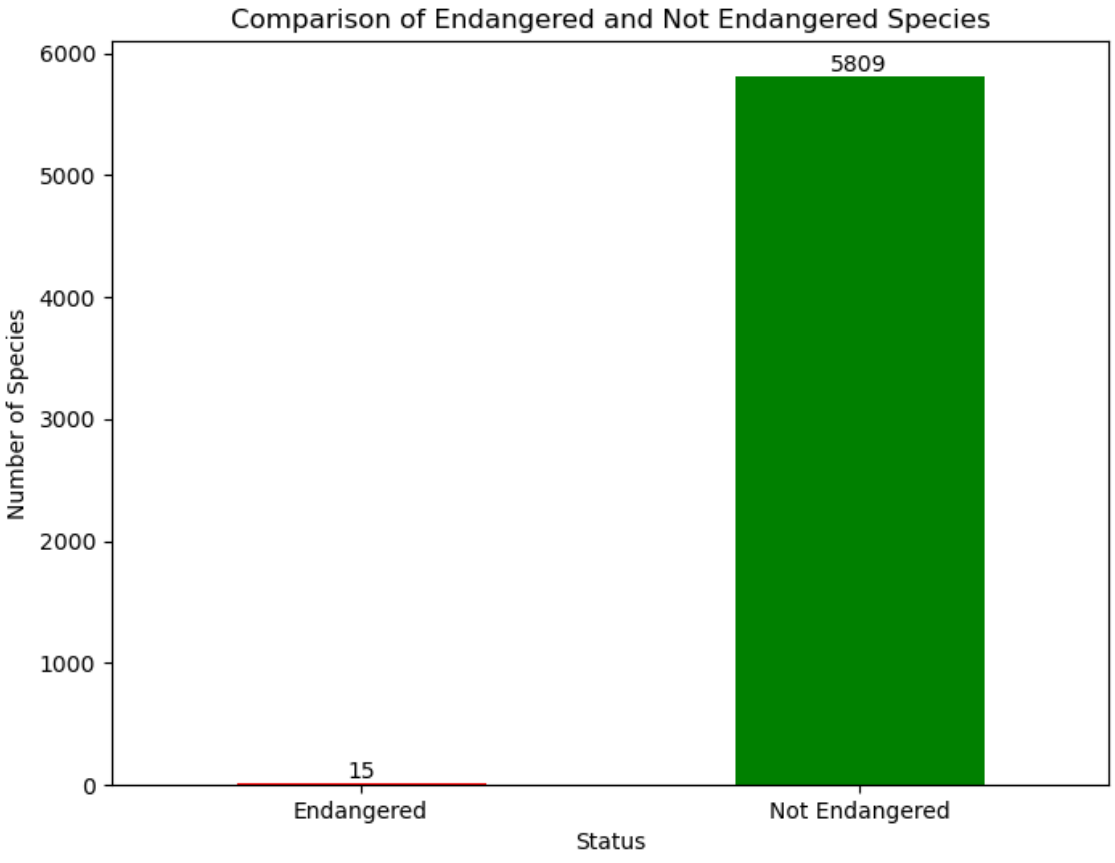
this is the most vulnerable group of species that needs urgent measures for its preservation.

In the table below you can see how many species are threatened in general. In total - 120 species - this is not such a large number in relation to 5824 species in general. Also, this table shows the number of species that have one or another nature conservation status and require additional research and the implementation of necessary measures to prevent the threat of their extinction.

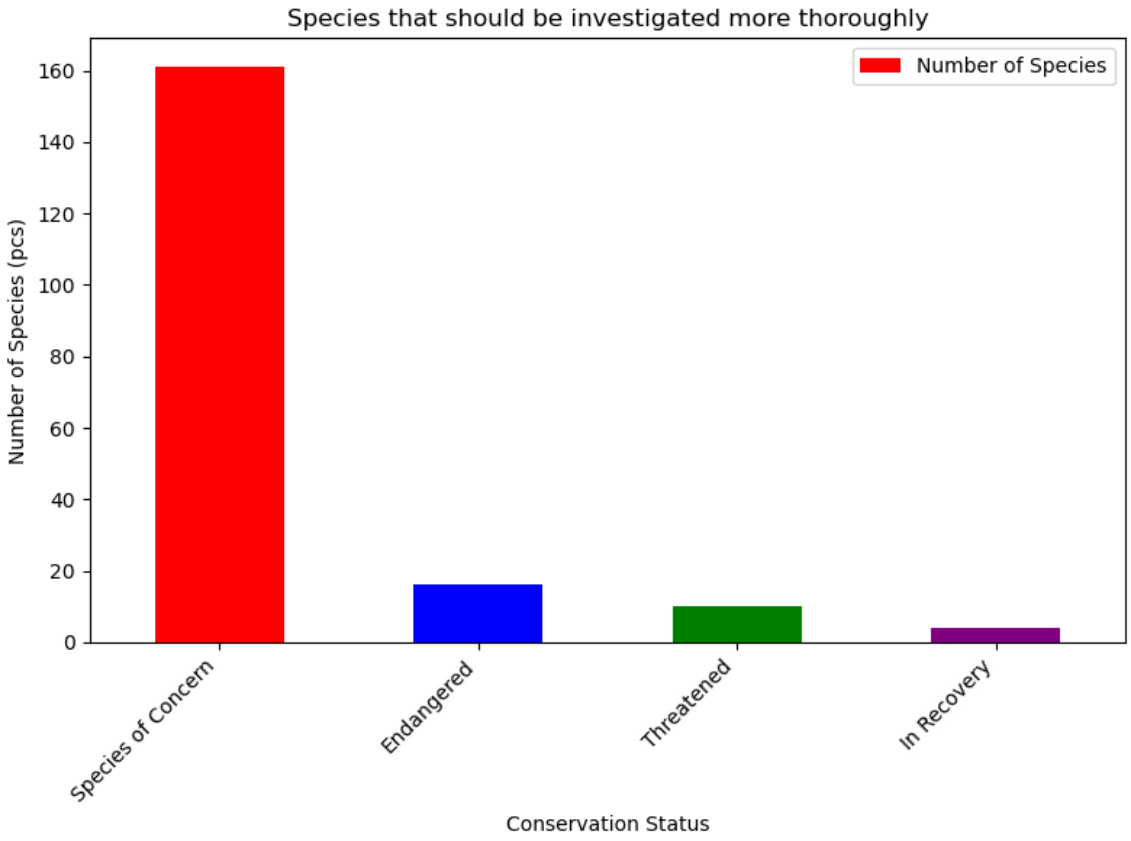
Conservation Status	Number of Species
Endangered	15
In Recovery	20
Species of Concern	40
Threatened	50

Since here we are mostly interested in species that are in immediate danger, we highlight: Total number of endangered species: 15

As it is presented on the diagram below, only the 15 unique species of 5824 in total are endangered in all 4 national parks in U.S.



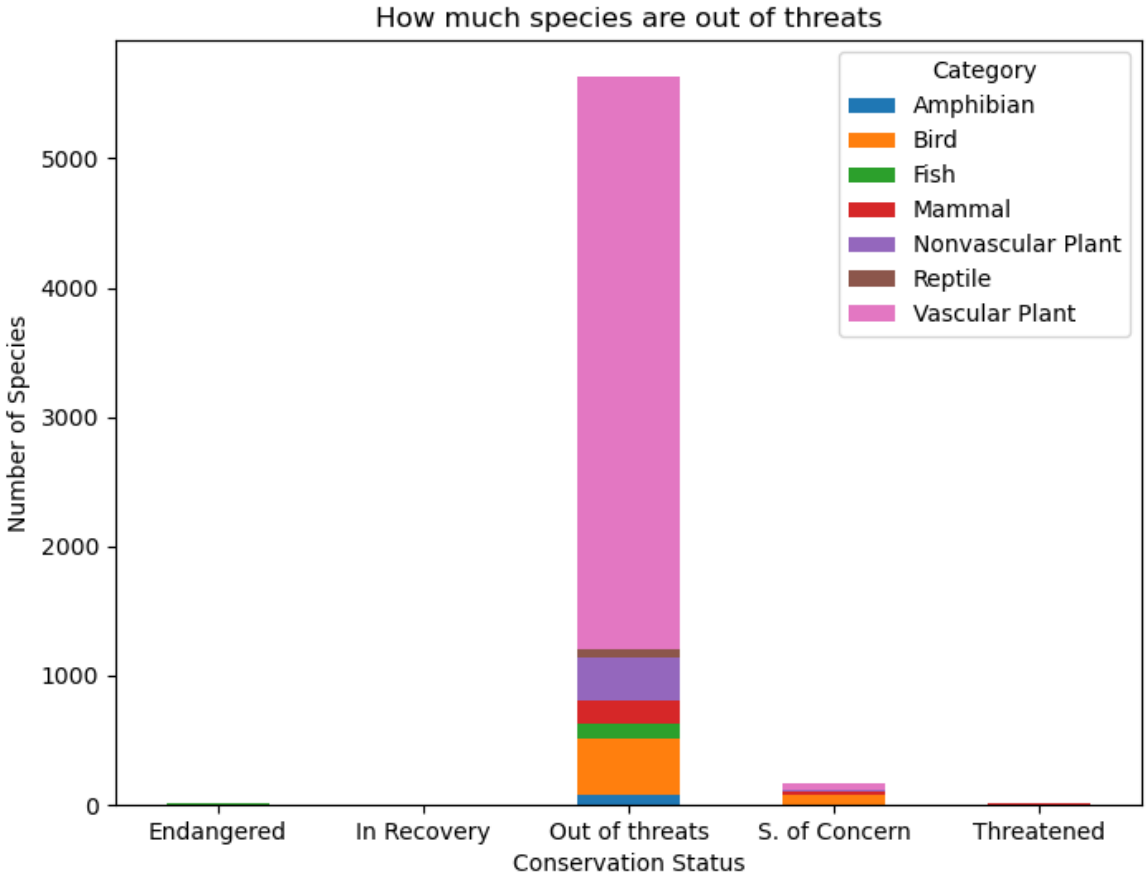
The following graph shows the distribution of species that have one or another conservation status and to which scientists and employees of national parks should pay more attention.



Again, here we can observe that not so big number of species on this graph are shown as directly endangered. But it should be noticed the number of species of concern. This tendency, over time, when take no measurements to prevent, could lead to increasing among threatened and endangered species.

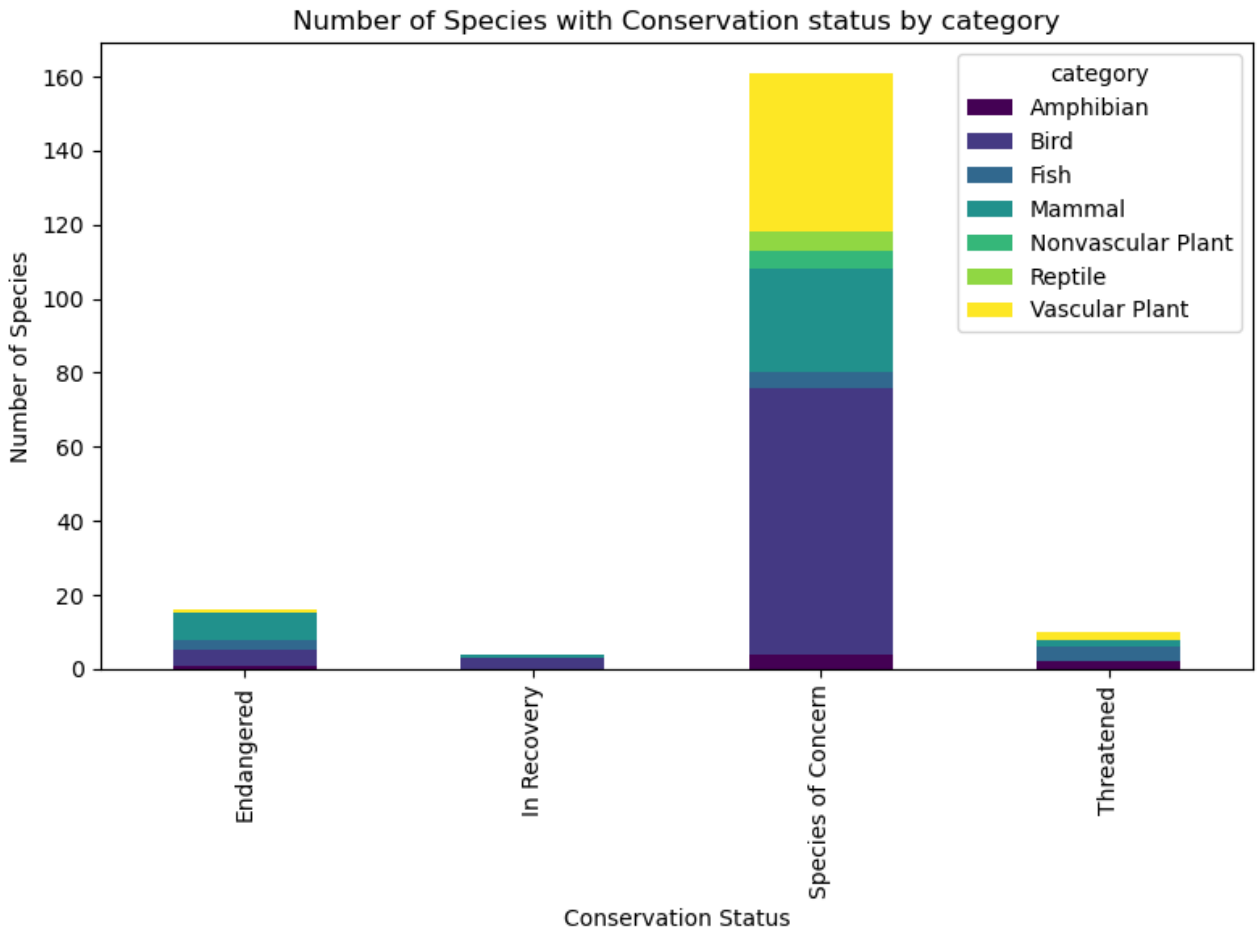
The next diagram below shows a pronounced tendency to the fact that species that are safe and not extinct or on the verge of extinction are significantly higher.

Also, this graph shows a certain increase in species in groups of birds, mammals and vascular plants that cause concern, they are more rapidly declining in national parks that species in other categories.

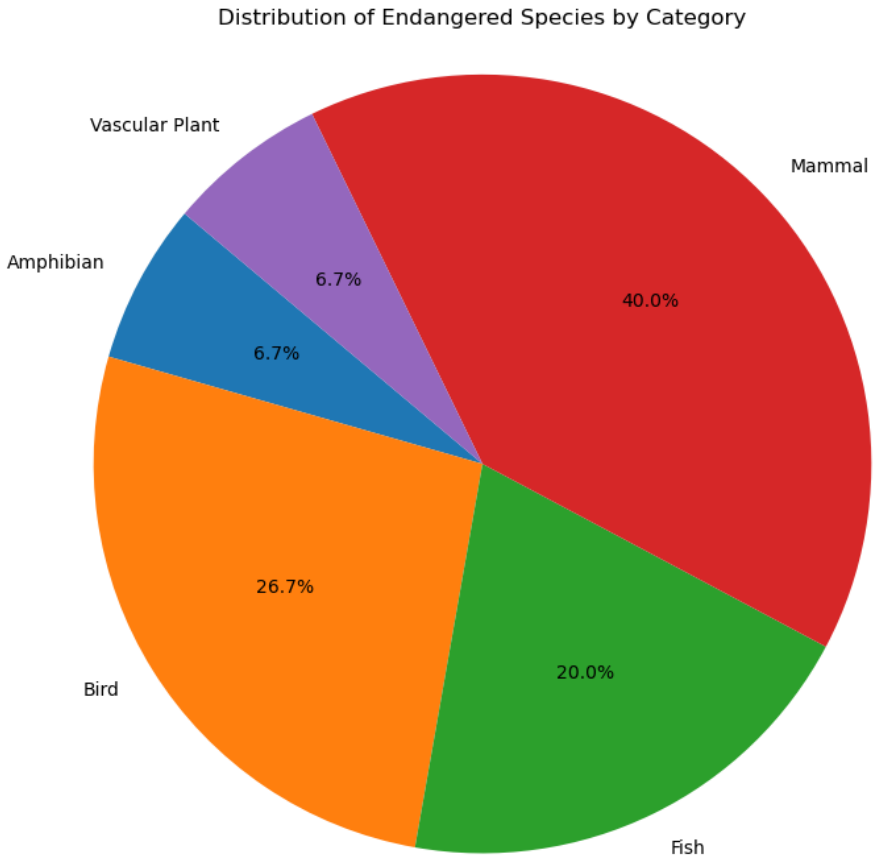


This situation shows that although there is a phenomenon of species extinction in national parks, it is not large-scale at this stage and there is still an opportunity to apply the necessary measures to stop the extinction of species and preserve them.

The diagram below shows the number of species in every category that have a conservation status. Here it is also possible to track the proportion between certain categories of species are endangered. It could be observed the content of each conservation status.

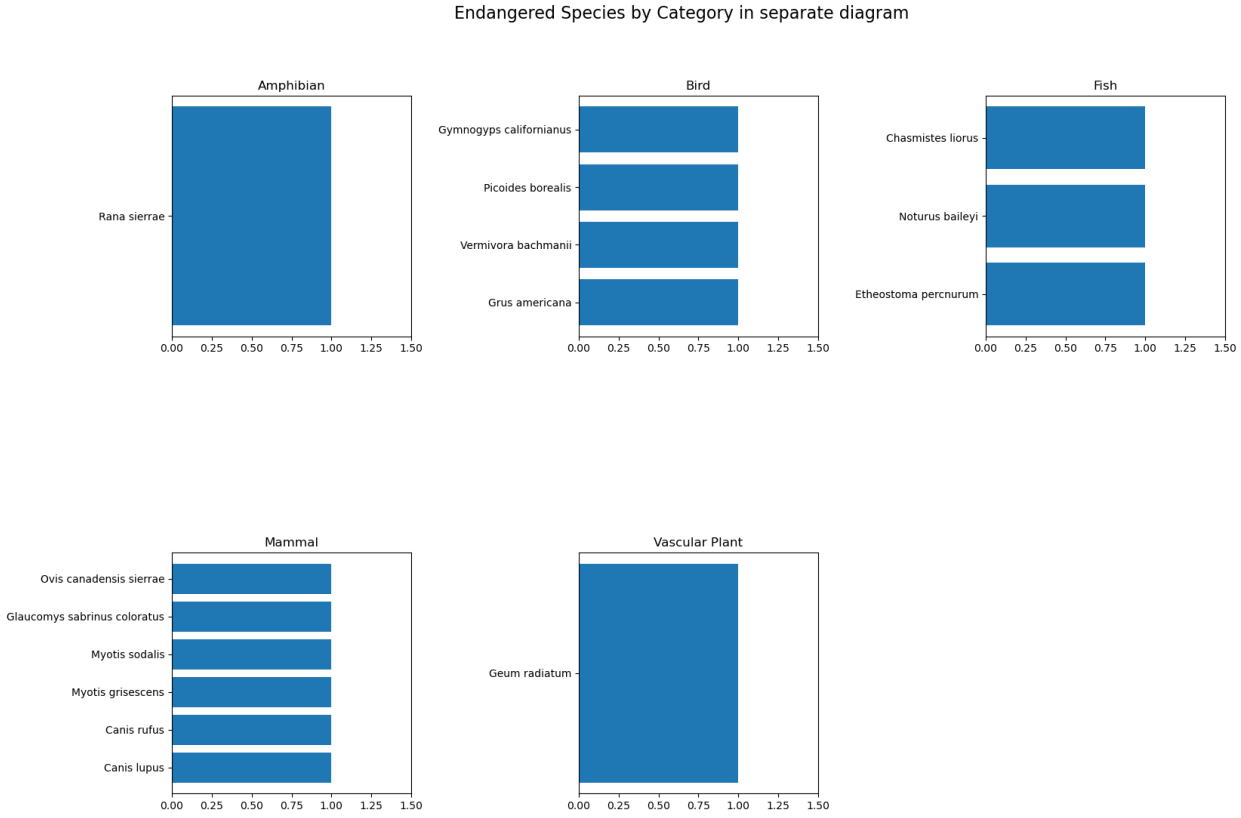


The next step will be to look at the percentage distribution of endangered species, depending on the category they belong to.

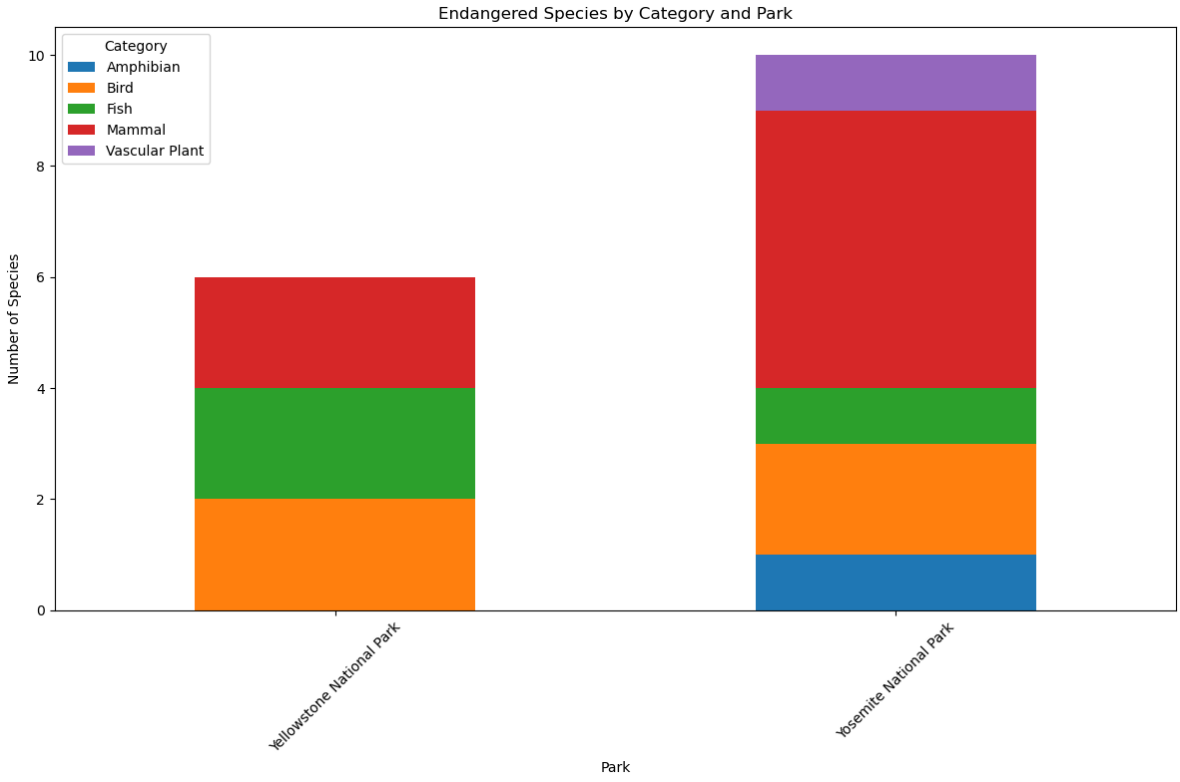


On the diagram of endangered species, which is broken down into categories and displayed by a separate figure, we can easily see which specific species require special attention of national park specialists. Six species from category of mammals and four species of birds are clearly displayed here, which quantitatively stand out among other endangered species.

Further in this analysis, we will see that Canis lupus is a specie that is endangered in two of four national parks and this can be a clue to understand the reason of the notion.



Finally, we came to the analysis of the distribution of endangered species in the four national parks of the United States. On the presented diagram it is easy to see that the problem of endangered species does not exist in each of the four investigated national parks. The Great Smoky Mountains National Park and the Bryce National Park that are not presented in the table, but they are displayed in data sets. At the same time as the Yellowstone National Park and the Yosemite National Park are contain data about endangered species.



This table contains 16 endangered species because Canis lupus is listed as endangered in both parks at the same time.

Endangered species by category and park

#	Park	Category	Species
1	Yellowstone National Park	Bird	Grus americana
2	Yellowstone National Park	Bird	Vermivora bachmanii
3	Yellowstone National Park	Fish	Etheostoma percnurum
4	Yellowstone National Park	Fish	Noturus baileyi
5	Yellowstone National Park	Mammal	Canis lupus
6	Yellowstone National Park	Mammal	Canis rufus

7	Yosemite National Park	Amphibian	Rana sierrae
8	Yosemite National Park	Bird	Picoides borealis
9	Yosemite National Park	Bird	Gymnogyps californianus
10	Yosemite National Park	Fish	Chasmistes liorus
11	Yosemite National Park	Mammal	Myotis grisescens
12	Yosemite National Park	Mammal	Myotis sodalis
13	Yosemite National Park	Mammal	Glaucomys sabrinus coloratus
14	Yosemite National Park	Mammal	Ovis canadensis sierrae
15	Yosemite National Park	Mammal	Canis lupus
16	Yosemite National Park	Vascular Plant	Geum radiatum

Analysis of statistical significance

Chi2 Statistic: 0.1617014831654557

P-value: 0.6875948096661336

Degrees of Freedom: 1

Expected Frequencies:

[[27.8313253 148.1686747]

[77.1686747 410.8313253]]

The null hypothesis (H0) in the chi-square test states that there is no association between the two variables that were analyzed - between the conservation status of the species and the park. A high p-value (0.6876) means that there is insufficient evidence to reject the null hypothesis, that is, there is no statistically significant association between these variables.

1. How does it affect the conclusions of the analysis?

- Since the p-value (0.6876) significantly exceeds the threshold value of 0.05, the null hypothesis cannot be rejected. This means that there is not enough evidence to say that there is an association between the two variables. In other words, the variables that were analyzed are independent of each other.

2. Is this high or low statistical significance?

- P-value 0.6876 indicates low statistical significance. In this case, since the p-value is significantly greater than 0.05, there is no reason to believe that the observed frequencies differ significantly from the expected frequencies.

Conclusions

Several data visualizations were performed during data analysis. This work made it possible to assess the general state of biodiversity in 4 US parks and draw certain conclusions regarding the composition of endangered species and their habitats.

In addition to the fact that this project provides clear answers to the questions that were raised at the beginning, several interesting details were additionally found that may be useful in further work on the preservation of species of flora and fauna of national parks.

It turned out that the percentage of species with nature protection status in relation to the total biodiversity of the parks is not very high, the majority of the flora and fauna of the national parks is not affected by extinction. However, certain species of animals and plants were identified in the data set, which require special attention and urgent application of appropriate conservation measures.

Also, it became obvious what distribution of nature protection status each of the species has, which categories of animals are the most vulnerable.

Although this analysis focused on researched questions related to the most vulnerable group of species and the definition of their habitats, however, additionally, when familiarizing with its results, one can see a tendency towards the growth of groups of species that cause concern that are approaching threatened marks in their populations. This may indicate that groups of extinct species may soon be replenished with new categories of animals and plants.

This data analysis also highlighted that the two most threatened categories in national parks are the mammal and bird categories. However, another interesting observation was that in the Great Smoky Mountains National Park and the Bryce National Park, the situation with the threat of species extinction is not so dire, and there are no groups that are closest to extinction from the 'Endangered' group. It is possible that this situation may be due to insufficient data. However, if this is not the case, then there is an opportunity to conduct additional research to identify the real reasons. It is hoped that certain conditions created in these national parks have led to better conservation of species and can be applied in places where such a problem exists.

Further research

The data set received for analysis contains observations for the last 7 days. In this way, it is impossible to follow any changes over time.

With a larger sample, where data are presented over a period of a year or even longer periods of time, it would be possible to follow whether the trends discussed in this paper are temporary and how animal and plant populations change their quantitative limits and see other trends related to with nature conservation status.

Due to the limitations of the obtained data sets, it is also difficult to analyze the factors that could affect the populations of the species. Spatial factors, the impact of climate change, cross-species competition, animal migration and many others remain unclear.

Continued accumulation of data and further deepening of their analysis could provide more insight into what specific factors affect biodiversity and what measures could be taken to help preserve it.