Biodiversity in National Parks

Codecademy Biodiversity Starter

Introduction

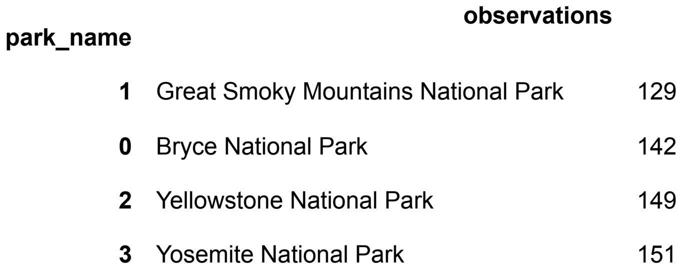
Project Scoping:

1. Data: we will start by looking at our data, what is the information that we have? .info() .head() .shape .describe(), what are the variables? which ones are we going to use? how may NaNs?
2. introductory EDA, visualization & Analysis
3. Conclusions

First look into our data

For this Project we will use 2 databases:

- observations (3 columns and 23985 entries)
- species (4 columns and 5823 entries)



	park_name	observations
1	Great Smoky Mountains National Park	129
0	Bryce National Park	142
2	Yellowstone National Park	149
3	Yosemite National Park	151

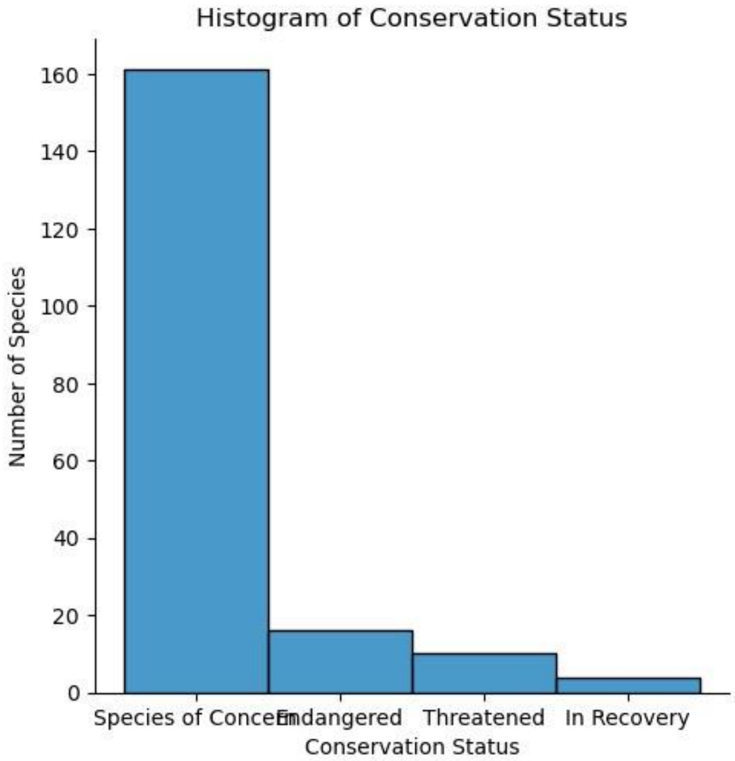
```
Observations
RangeIndex: 23296 entries, 0 to 23295
Data columns (total 3 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   scientific_name  23296 non-null  object
 1   park_name        23296 non-null  object
 2   observations     23296 non-null  int64
dtypes: int64(1), object(2)
```

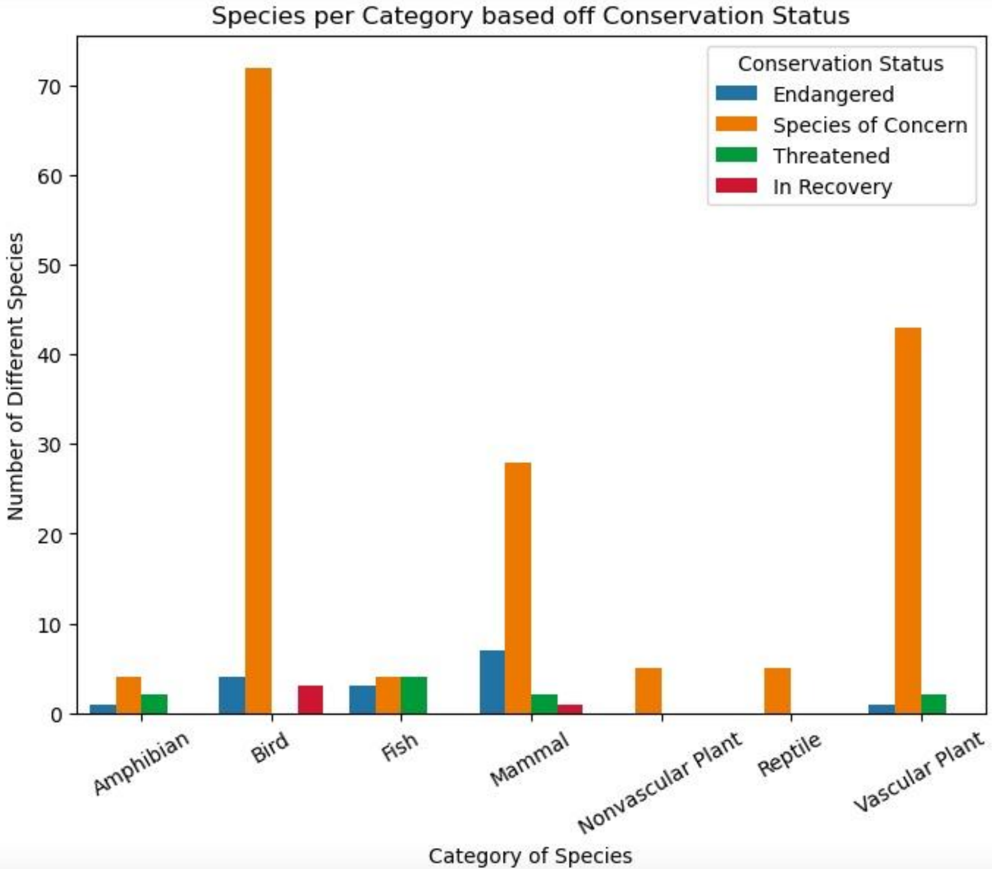
```
Species
RangeIndex: 5824 entries, 0 to 5823
Data columns (total 4 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   category             5824 non-null   object
 1   scientific_name      5824 non-null   object
 2   common_names         5824 non-null   object
 3   conservation_status  191 non-null    object
dtypes: object(4)
```

The first insight that we can pull from this first look is that conservation_status has A LOT of NaNs which means that the number of species that have a conservation status is a very small compared to all of the species considered.

All national parks have more than 125 species. the park with the most species (observed) is Yosemite Park with 151 different species, followed by Yellowstone Park, then Bryce Park and last Great Smoky Mountains Park

EDA: let’s have a look at conservation status

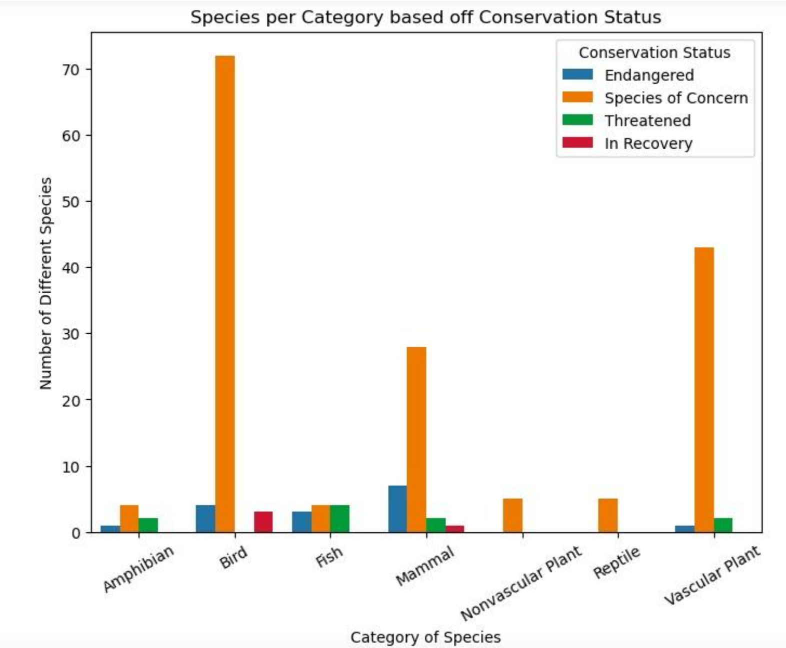


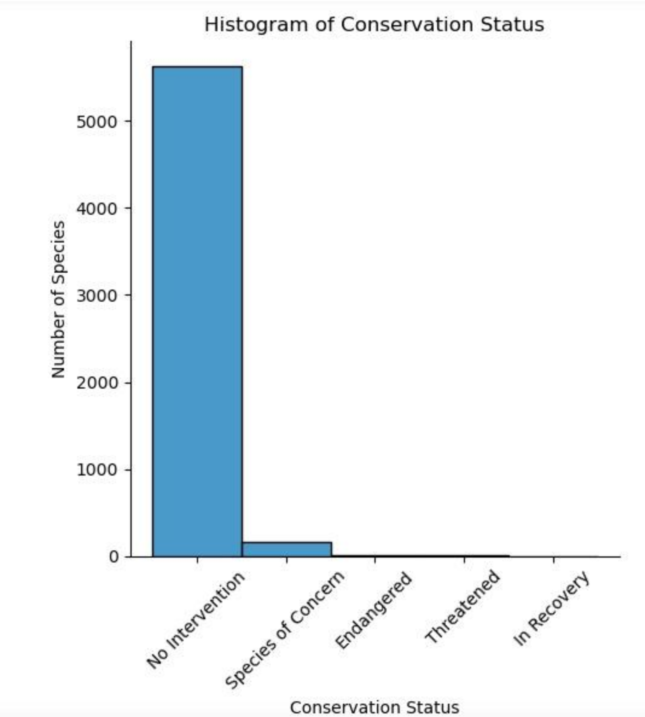


EDA: Let’s fill the NaNs of conservation status to see if there are differences in our EDA

The species that have a conservation status are 191, which is a small piece of all the species. All categories of living beings have species of concern, the category with most concern is "birds", the categories with most threatened species are amphibien, fish and vascular plant.on the other hand, mammals present the most endangered species. the differences between categories of species seems significant depending on the category, but let's remember that we dont have many data on conservation status, so its hard to make hipothesis with few data.

To all the NaNs in conservation statuswe will fill with .fillna, then we will do the distributions again
species.fillna('No Intervention', inplace = True)





Conclusions

What did you learn throughout the process?

Are the results what you expected?

What are the key findings and takeaways?
You can clearly see some major differences between the categories of species!

If you were to rank the categories from greatest to least by how much attention each category needs, then the ranking would be as follows:

1. Bird (Huge concern for nearly the entire category of species)
2. Vascular Plant (Huge concern for nearly the entire category of species)
3. Mammal (Has the most endangered species but far less species who are of concern)
4. Fish (Moderate concern)
5. Amphibian (Moderate concern)
6. Nonvascular Plant & Reptile (No Reptiles or Nonvascular Plants are endangered and very few are of concern)