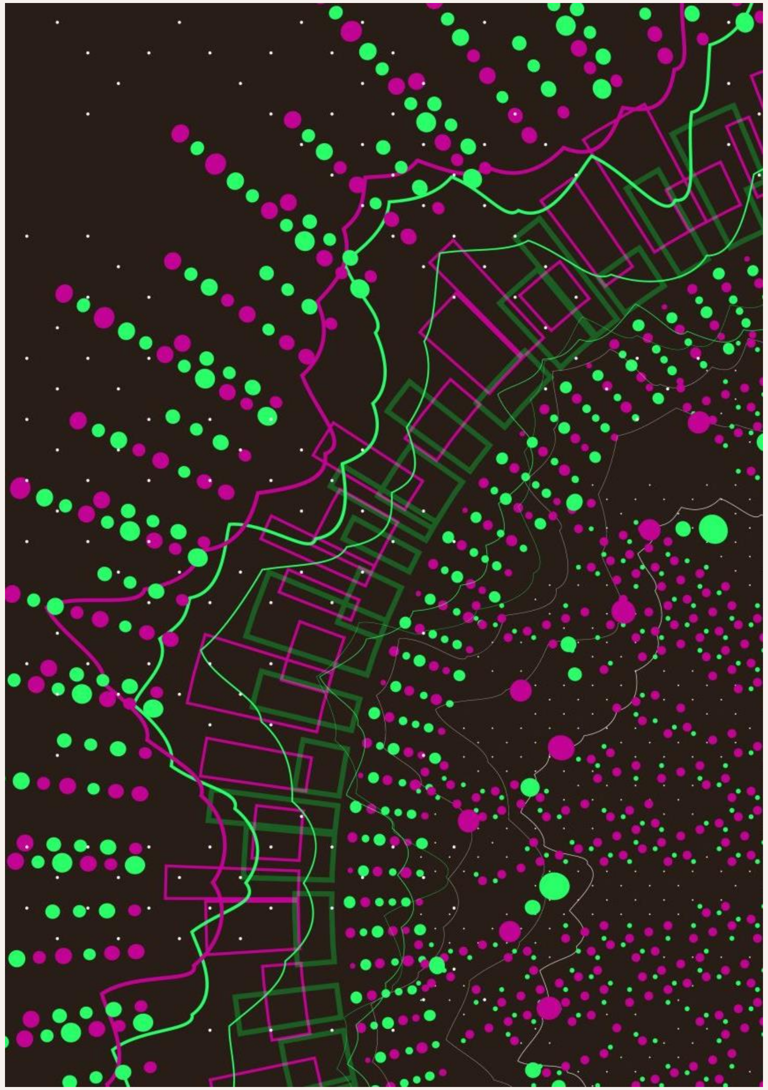
Biodiversity Analysis and Conservation Insights

ANALYZING SPECIES DATA TO DRIVE CONSERVATION DECISIONS

by

Paschalis Agapitos



Introduction

- **Objective:** Analyze biodiversity data to uncover patterns and insights

- **Dataset:** Summary of biodiversity-related data

- **Approach:**
  1. Data preparation
  2. Visualisation & Analysis
  3. Key findings

Data Overview

Datasets:

- \`observations\`: contains observations of species in various national parks, including scientific names, park names, and observation counts
- \`species_info\`: provides species information, including taxonomic categories, scientific names, common names, and conservation status

Key Observations:

- 5,824 species records.
- Conservation status available for only 191 species.
- Categories include Mammals, Birds, Reptiles, and more.

Methods

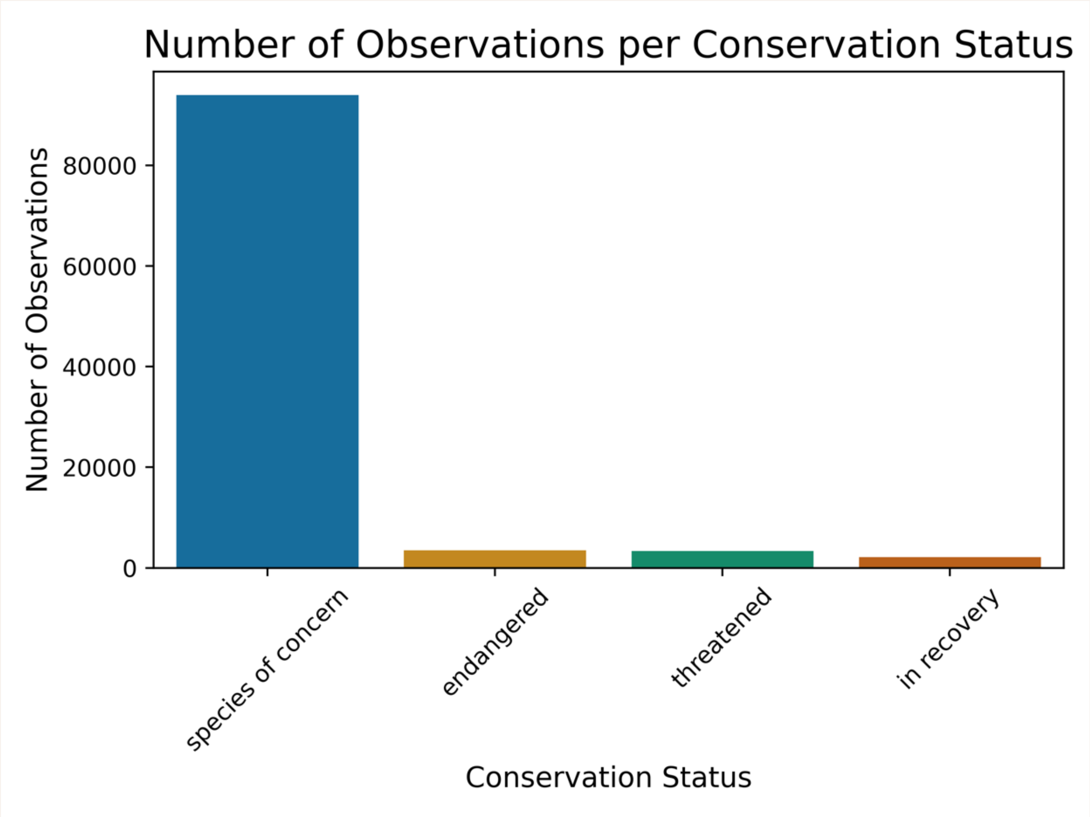
- Data Cleaning: Addressed duplicates, missing values, and ensured consistency in datasets.

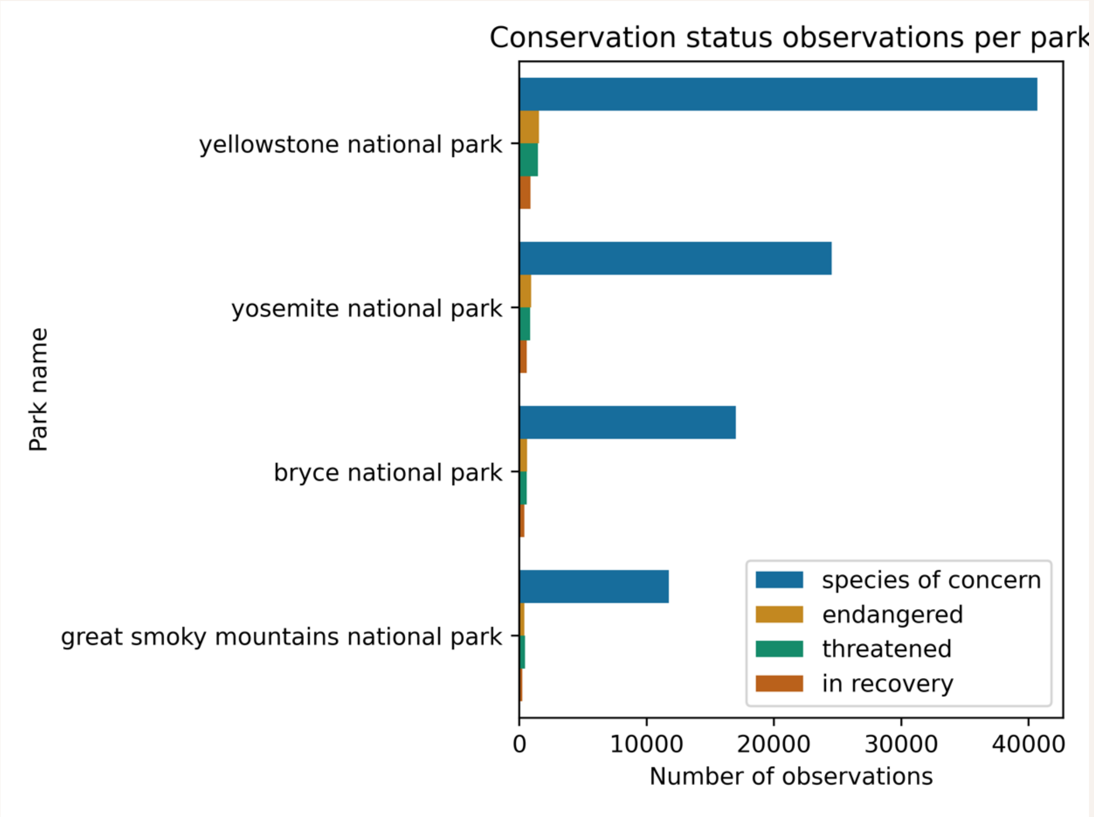
- Exploratory Data Analysis (EDA): Identified trends, patterns, and potential outliers in the biodiversity data.

- Integration of Multiple Datasets: Combined species observations with taxonomic information for comprehensive analysis.

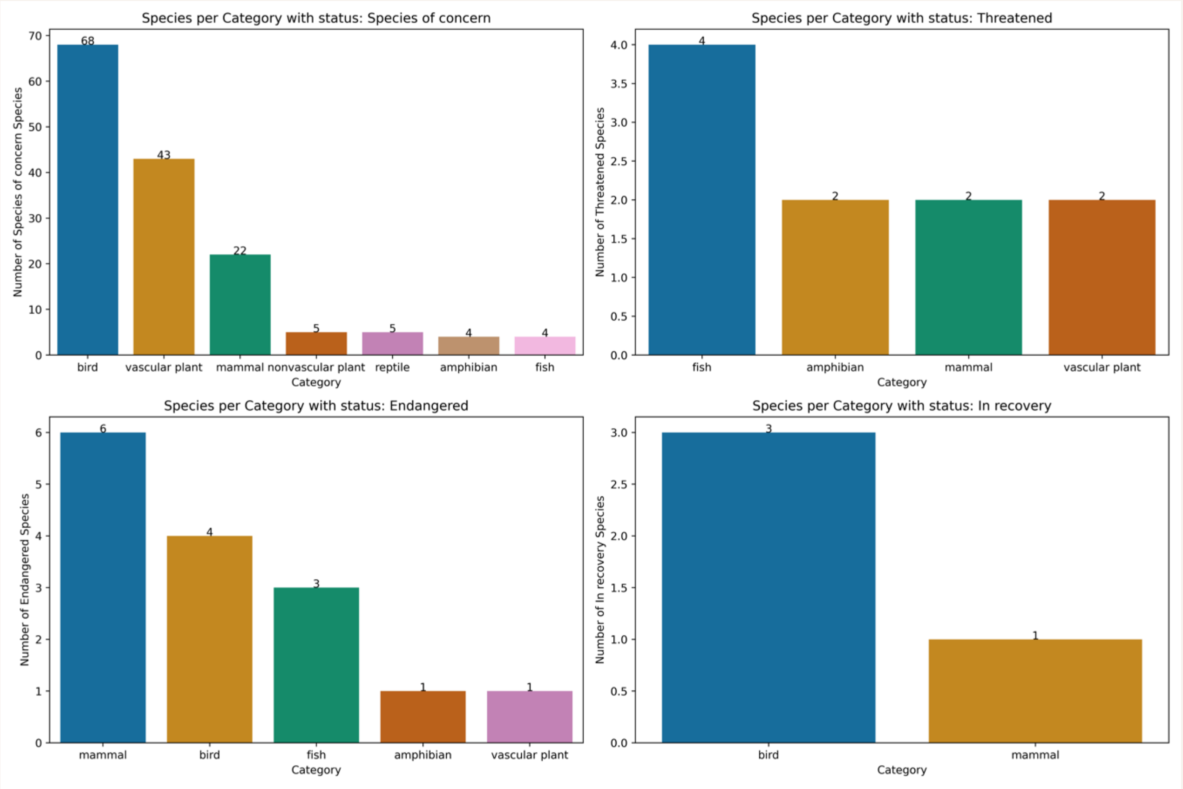
- Statistical Insights: Extracted metrics such as species frequency, regional diversity, and conservation status distribution.

Exploratory Data Analysis (1)





Exploratory Data Analysis (2)



Species per Category with status: Species of concern

Species per Category with status: Threatened

Species per Category with status: Endangered

Species per Category with status: In recovery

- species of concern dominate the conservation status across all parks (93962).
- birds are more likely to be in recover or of concern
- mammals are more likely to be endangered
- fishes most likely are threatened species' categories

Biodiversity Analysis (1)

- **Species richness**: total number of different species in a specific area (**alpha diversity**)
- **Species evenness**: how evenly distributed are species within a given ecosystem (i.e., the relative number of individuals of each species in an area).
- **Shannon-Wiener Index (H')**: $H' = -\sum(p \times \ln(p_i))$
  - combines evenness and richness

park name	species richness	species evenness	H'
Bryce	5541	1.014	8.742
Great Smoky Mountains	5541	1.012	8.725
Yellowstone	5541	1.016	8.760
Yosemite	5541	1.015	8.754

Biodiversity Analysis – Findings

- The slight differences in H' values are driven by how evenly species are distributed across the parks.

- Yellowstone has the most even distribution.

- Great Smoky Mountains is slightly less balanced.

Is there an association between the species' cateogories and the conservation status?

Data Preparation: Removed missing values, "in recovery" are labeled as "safe" (24) and the rest as "danger" (856)

Chi-Square Test Findings:

- Null hypothesis (H_0): There is no association between a species' category (e.g., bird, mammal) and its status (e.g., danger, safe).
- Alternative hypothesis (H_1): There is some association between category and status
  - P-value: ≈ **0.026** (Significant at $\alpha = 0.05$).

Interpretation:

- H_0 is rejected, indicating a significant association between species category and their conservation status.

Which categories drive the association?

Post-hoc analysis findings:

- mammals which have more safe individuals than expected (2.255)
- vascular plants which have fewer safe individuals than expected (2.240)

Interpretation:

- the species categories that drive the association observed in the significance test (chi-squared test) are mammals and vascular plants