Basic Data Science Report: Biodiversity in US National Parks

Emircan Akyurek

June 9, 2024

Abstract

This beginner data science project explores biodiversity in four US National Parks. The analysis covers species categories and conservation status. During data analysis, a problem in data cleaning was found. The same values appeared unusually in different parks, probably due to a code or merge mistake. This report presents the project process, findings, and describes the error for future improvement. Figures are referenced by their image filenames.

1 Introduction

Biodiversity is important for ecosystems, and national parks play a role in protecting it. In this project, observation and species data from Bryce, Yosemite, Great Smoky Mountains, and Yellowstone parks were used. The aim was to count and compare species groups and their conservation status.

2 Data and Methodology

Data from two tables were joined: one with species info (including category and conservation status), and one with park observations. Data cleaning included merging on the scientific name. Counts for species by category and conservation status within each park were then calculated. Plots were made with Python matplotlib and seaborn.

3 Results

3.1 Conservation Status Counts

The conservation status plot (Figure 1) shows that most counts fall under ‘Excluded Statuses‘, with very high identical numbers (e.g. 6188) for every park. The same happens for ‘Species of Concern‘ and other categories. This exact distribution is unlikely in reality.

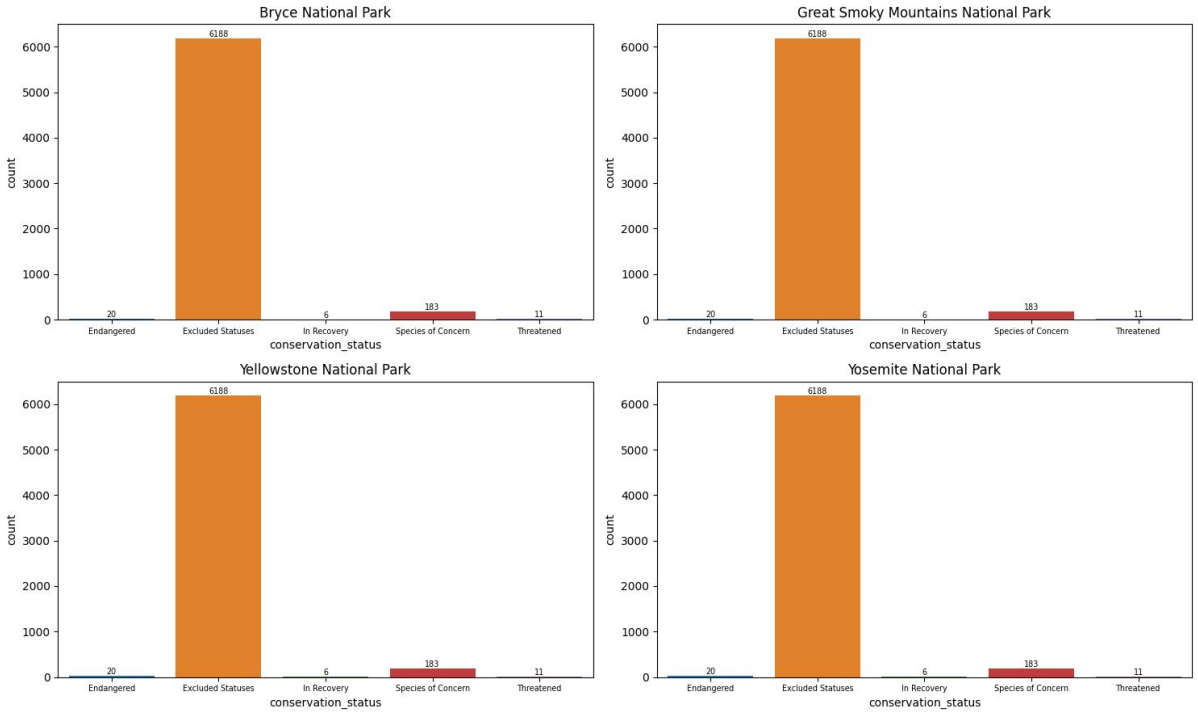

Figure 1: Conservation status by park (Source: Conservation-Status-by-Park-Name.jpg)

3.2 Species Category Counts

Plots of category per park (Figure 2) further show identical numbers for each group and park, e.g. every park has exactly 591 birds, 82 amphibians, etc. This is also quite suspicious and suggests something went wrong with the counts.

4 Error Detection and Discussion

From the results, it is clear that something is wrong with the real park counts. During data cleaning or merging, likely some operation duplicated or broadcast identical values to every park and group. Possible explanations:

- The join/merge operation matched incorrectly, causing all data to be assigned to all parks.
- A groupby/count summary accidentally used the full dataset instead of park-filtered data.
- Visualization or aggregation code may have recycled or misapplied values.

This mistake was only found after plotting and noticing the repeated patterns. It shows the importance of checking results for errors, not just running code.

5 Learning and Next Steps

Working through this, the main lesson was about the value of looking at results, not just trusting computations. Even simple plots can show possible errors fast. The next step is

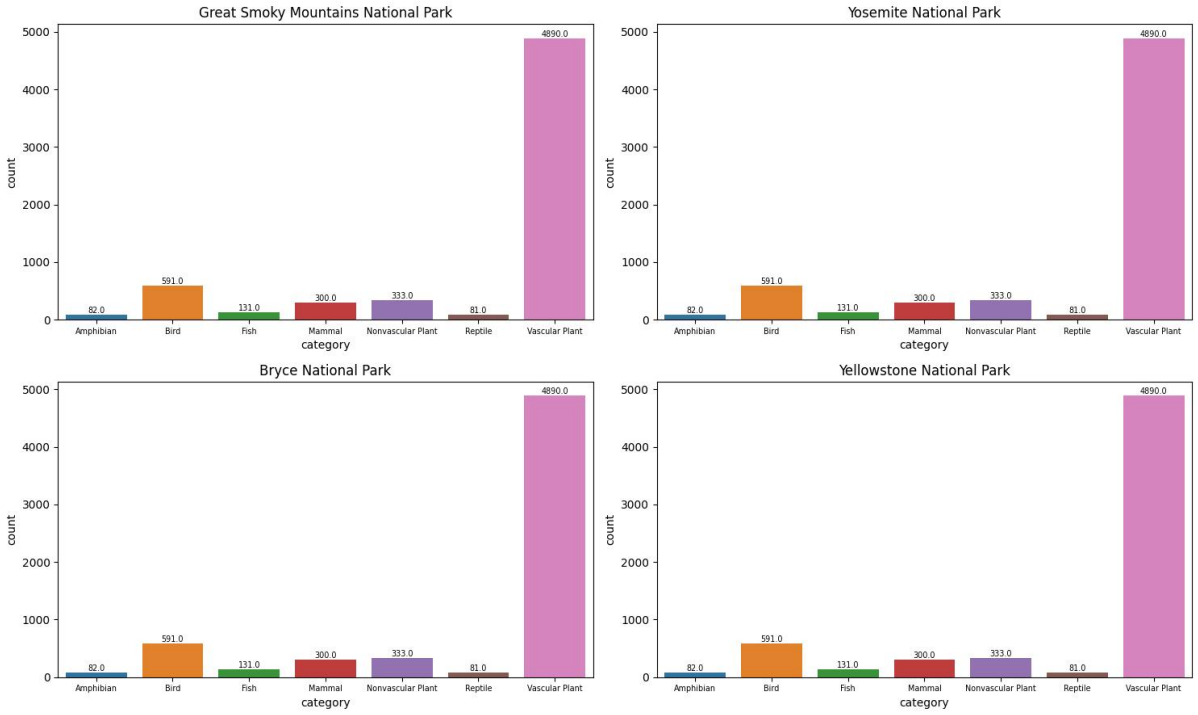

Figure 2: Category counts by park (Source: Category-by-Park-Name.jpg)

to fix the merging code, check groupby filters, and ensure that only species observed in each park are counted for that park.

6 Conclusion

This project explored biodiversity data and found a significant counting error likely caused by the data cleaning or merging process. The process showed both how basic analysis is useful, and how common small mistakes can cause big errors in results. Being able to discover and describe such problems is an important skill for future progress.

References

Figures:

- Conservation-Status-by-Park-Name.jpg
- Category-by-Park-Name.jpg
- test.jpg

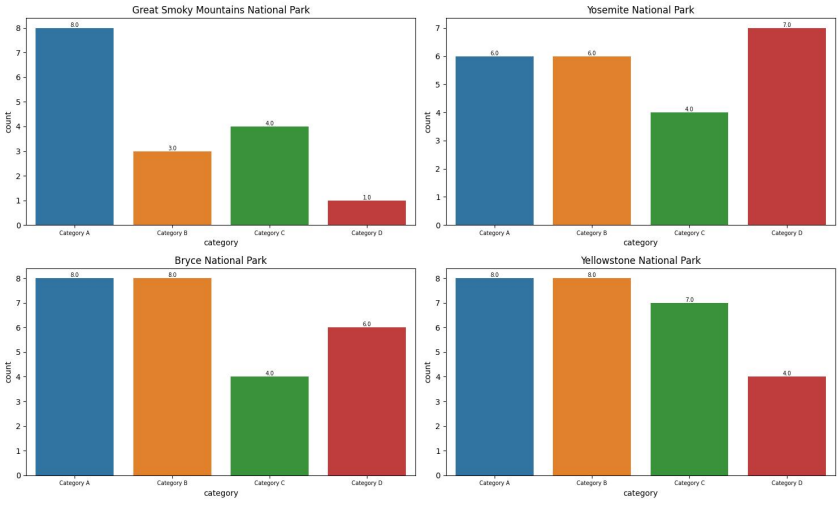

Figure 3: Test random category data by park (Source: test.jpg)