Introduction

The purpose of this project is to analyse biodiversity data from the National Parks Service, with a direct focus on species, conservation status, and park observations. The analysis follows a structured workflow that includes project scoping, data preparation, exploratory and statistical analysis, data visualisation, and interpretation of findings. Additionally, this analysis is guided by four key questions: “How are species distributed across conservation status categories?” “Do conservation status rates differ across major taxonomic groups?” “Are observed differences in conservation status statistically significant?” “Which species group is most prevalent, and how is it distributed across national parks?”

Lastly, the analysis is performed using two datasets by Codecademy: species_info.csv, which contains species taxonomy and conservation status, and observations.csv, which records species observations across national parks.

Scoping

Before conducting the analysis, a clear scope is established to define the purpose, direction, and boundaries of the project. This scoping process outlines the analytical objectives, confirms the suitability of the available data, describes the planned analytical approach, and clarifies how findings will be evaluated. Establishing this structure ensures that the analysis remains focused, methodical, and aligned with the project aims.

Project Goals

This project is conducted from the perspective of a biodiversity analyst supporting the National Parks Service. From this standpoint, understanding species conservation status is critical for protecting at-risk species and preserving biodiversity across national parks. The primary objective of the analysis is to examine how conservation status varies across species groups and how species observations differ across park locations.

To support conservation decision-making, the analysis focuses on identifying patterns in conservation status, assessing whether certain taxonomic groups face higher levels of risk, evaluating whether observed differences are statistically meaningful, and determining how the most frequently observed species are distributed across parks.

Data

The analysis draws on two datasets provided as part of the project materials. One dataset contains species-level information, including taxonomic classification and conservation status, while the second records species observations across national park locations. Together, these datasets provide the necessary information to link species characteristics with conservation outcomes and spatial distribution.

Analysis

The analytical approach combines descriptive statistics, data visualisation, and statistical testing to uncover patterns within the data. Summary measures and visual comparisons are used to explore distributions and relationships, while inferential methods are applied to assess whether differences between species groups are statistically significant. Key comparisons include conservation status across taxonomic categories and observation patterns of prevalent species across parks.

Evaluation

The success of the analysis is assessed by how effectively it addresses the original project objectives and analytical questions. Findings are interpreted in relation to conservation relevance, with attention given to the strength and limitations of the evidence. This evaluation also considers constraints within the data and reflects on how alternative methods or additional information could improve future analyses.

Import Python Modules

Firstly, I will import the primary modules that will be used for this project:

In [1]:
```
import pandas as pd
import numpy as np
```

```
from matplotlib import pyplot as plt
import seaborn as sns
```

Loading the Data

To analyse the conservation status of species and their observations across national parks, I loaded the datasets into DataFrames for exploration and visualisation using Python. The files **Observations.csv** and **Species_info.csv** were read into DataFrames named observations and species, respectively, and their contents were initially inspected using **.head()**.

The **species_info.csv** dataset contains information about species recorded in the national parks. The variables include the taxonomic category of each species, its scientific name, common names, and conservation status.

In [2]:
```
species = pd.read_csv('species_info.csv')
species.head()
```

Out[2]:

	category	scientific_name	common_names	conservation_status
0	Mammal	Clethrionomys gapperi gapperi	Gapper's Red-Backed Vole	NaN
1	Mammal	Bos bison	American Bison, Bison	NaN
2	Mammal	Bos taurus	Aurochs, Aurochs, Domestic Cattle (Feral), Dom...	NaN
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	NaN
4	Mammal	Cervus elaphus	Wapiti Or Elk	NaN

The **observations.csv** dataset contains records of species sightings across national parks over seven days. For each record, the dataset includes the scientific name of the species, the national park in which it was observed, and the number of observations recorded during this period.

In [3]:
```
observations = pd.read_csv('observations.csv')
observations.head()
```

Out[3]:

	scientific_name	park_name	observations
0	Vicia benghalensis	Great Smoky Mountains National Park	68
1	Neovison vison	Great Smoky Mountains National Park	77
2	Prunus subcordata	Yosemite National Park	138
3	Abutilon theophrasti	Bryce National Park	84
4	Githopsis specularioides	Great Smoky Mountains National Park	85

Data Exploration and Characteristics

Before conducting further analysis, I explored both datasets to understand their structure, scale, and key characteristics. This step helps identify patterns, inconsistencies, and potential data quality issues that may influence the analysis.

Species

In [10]:
```
print(f'The species table has {species.shape[0]:,} rows and {species.shape[1]:,} columns')
print(f'The observations table has {observations.shape[0]:,} rows and {observations.shape[1]:,} columns')
```

```
The species table has 5,824 rows and 4 columns
The observations table has 23,296 rows and 3 columns
```

In [14]:
```
print(f'There are {species.category.nunique()} different species')
print(f'These species consist of: {species.category.unique()}')
```

```
There are 7 different species
These species consist of: ['Mammal' 'Bird' 'Reptile' 'Amphibian' 'Fish' 'Vascular Plant'
 'Nonvascular Plant']
```

In [16]:
```
for category, count in species.groupby('category').size().items():
    print(f"There are {count:,} {category}s")
```

```
There are 80 Amphibians
There are 521 Birds
There are 127 Fishs
There are 214 Mammals
There are 333 Nonvascular Plants
There are 79 Reptiles
There are 4,470 Vascular Plants
```

In [18]:
```
print(f"There are {species.conservation_status.nunique():,} conservation statuses")
print(f'These consist of: {species.conservation_status.unique()}')
```

```
There are 4 conservation statuses
These consist of: [nan 'Species of Concern' 'Endangered' 'Threatened' 'In Recovery']
```

In [24]:
```
print(f'There are {species.conservation_status.isna().sum():,} nan values in the conservation_status column.\n')
for conservation_status, count in species.groupby('conservation_status').size().items():
    print(f'There are {count} species that are considered "{conservation_status.capitalize()}"')
```

```
There are 5,633 nan values in the conservation_status column.

There are 16 species that are considered "Endangered"
There are 4 species that are considered "In recovery"
There are 161 species that are considered "Species of concern"
There are 10 species that are considered "Threatened"
```

Observations

In [25]:
```
print(f'There are {observations.park_name.nunique()} different parks.')
print(f'These parks are: {observations.park_name.unique()}')
```

```
There are 4 different parks.
These parks are: ['Great Smoky Mountains National Park' 'Yosemite National Park'
 'Bryce National Park' 'Yellowstone National Park']
```

In [28]:
```
print(f'There are {observations.observations.sum():,} observations.')
```

```
There are 3,314,739 observations.
```

Analysis

The analysis begins by examining the distribution of conservation statuses across all species. This provides a baseline understanding of how many species currently require conservation intervention before comparing differences between taxonomic groups.

In [43]:
```
# Replace missing conservation status with "No Intervention"
species['conservation_status'] = species['conservation_status'].fillna('No Intervention')

# Count each conservation status
status_counts = species['conservation_status'].value_counts()
print(status_counts)
```

```
conservation_status
No Intervention       5633
Species of Concern     161
Endangered              16
Threatened              10
In Recovery              4
Name: count, dtype: int64
```

The results show that the majority of species are classified as requiring no intervention, while a smaller subset fall under conservation categories such as Species of Concern, Threatened, Endangered, or In Recovery. This indicates that conservation efforts are concentrated on a relatively small proportion of species.

Conservation Status by Species Category

To understand whether conservation status differs across species types, species were grouped by taxonomic category and conservation status.

In [44]:
```
# Remove "No Intervention" species
at_risk = species[species['conservation_status'] != 'No Intervention']

# Count species by conservation status and category
status_by_category = (
    at_risk
    .groupby(['conservation_status', 'category'])
    .size()
    .unstack()
)
```

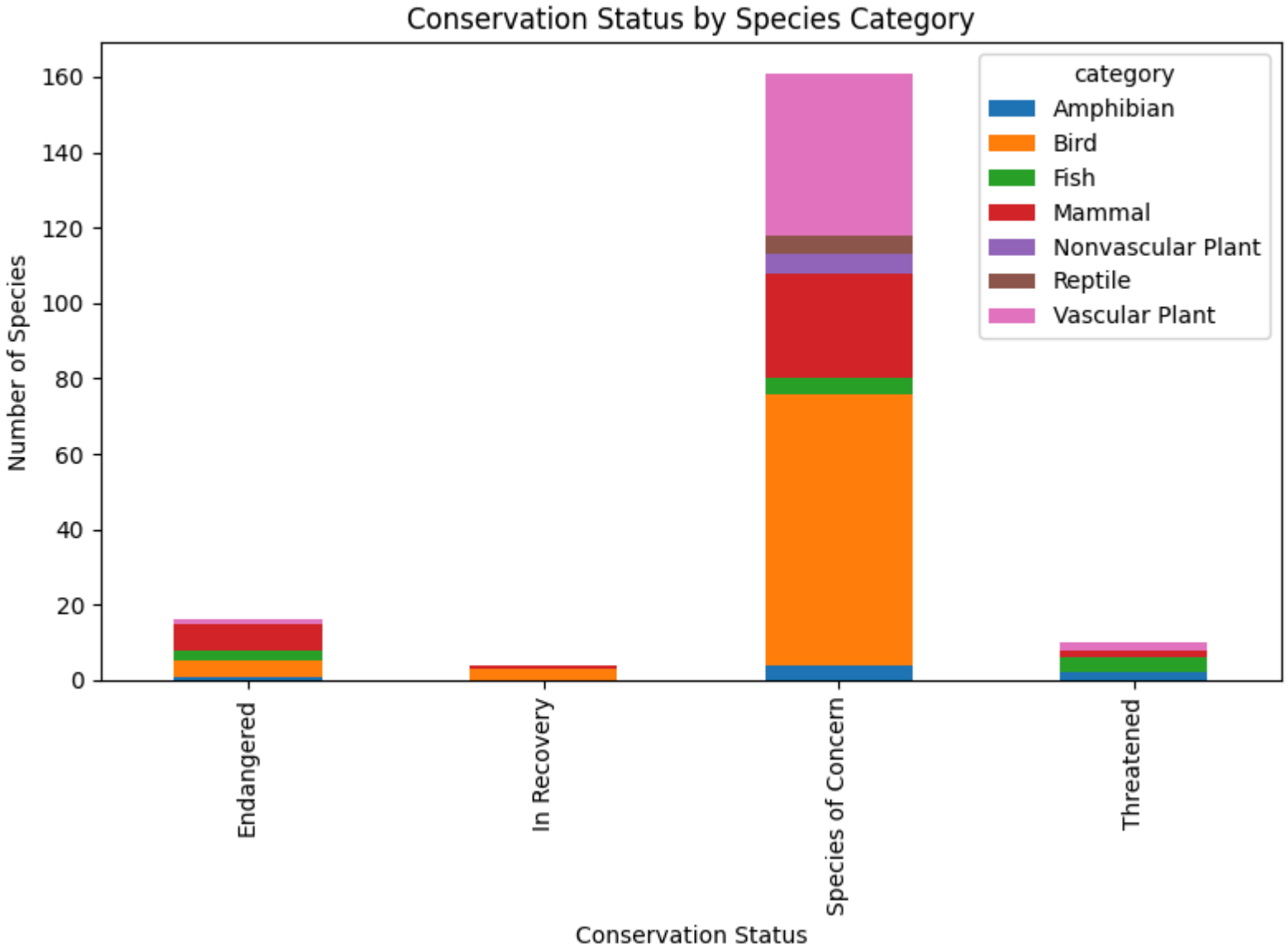
```
print(status_by_category)
```

```
category             Amphibian  Bird  Fish  Mammal  Nonvascular Plant  \
conservation_status
Endangered                 1.0   4.0   3.0     7.0                NaN
In Recovery                NaN   3.0   NaN     1.0                NaN
Species of Concern         4.0  72.0   4.0    28.0                5.0
Threatened                 2.0   NaN   4.0     2.0                NaN

category             Reptile  Vascular Plant
conservation_status
Endangered               NaN             1.0
In Recovery              NaN             NaN
Species of Concern       5.0            43.0
Threatened               NaN             2.0
```

In [45]:
```
status_by_category.plot(
    kind='bar',
    stacked=True,
    figsize=(8, 6)
)

plt.xlabel('Conservation Status')
plt.ylabel('Number of Species')
plt.title('Conservation Status by Species Category')
plt.tight_layout()
plt.show()
```



Mammals and birds account for a large share of species requiring conservation intervention. In particular, mammals appear most frequently in the Endangered category, while birds are more prominent in the In Recovery category. This suggests variation in conservation outcomes across taxonomic groups.

Protected Species by Category

Absolute counts alone do not account for differences in the total number of species within each category. To address this, species were classified as either protected or not protected, and protection rates were calculated for each taxonomic group.

In [58]:
```
# Exclude species with no conservation concern
protected_species = species[species['conservation_status'] != 'No Intervention']

# Count protected vs not protected
protection_counts = (
    species
    .groupby(['category', 'is_protected'])
    .size()
    .unstack(fill_value=0)
    .reset_index()
)

protection_counts = protection_counts.rename(
    columns={False: 'not_protected', True: 'protected'}
)

print(protection_counts)
```

```
is_protected           category  not_protected  protected
0                     Amphibian             73          7
1                          Bird            442         79
2                          Fish            116         11
3                        Mammal            176         38
4             Nonvascular Plant            328          5
5                       Reptile             74          5
6                Vascular Plant           4424         46
```

In [59]:
```
# Calculate percentage protected
protection_counts['percent_protected'] = (
    protection_counts['protected'] /
    (protection_counts['protected'] + protection_counts['not_protected'])
) * 100

print(protection_counts)
```

is_protected	category	not_protected	protected	percent_protected
0	Amphibian	73	7	8.750000
1	Bird	442	79	15.163148
2	Fish	116	11	8.661417
3	Mammal	176	38	17.757009
4	Nonvascular Plant	328	5	1.501502
5	Reptile	74	5	6.329114
6	Vascular Plant	4424	46	1.029083

When expressed as percentages, mammals and birds show the highest proportion of protected species. This indicates that these groups face relatively higher conservation risk compared to other categories.

Statistical Significance of Conservation Differences

To determine whether observed differences in conservation status between species groups are statistically significant, chi-squared tests were conducted.

In [60]:
```
from scipy.stats import chi2_contingency

# Mammals vs Birds
contingency_mb = [
    [30, 146],  # Mammals: protected, not protected
    [75, 413]   # Birds: protected, not protected
]

chi2, p, dof, expected = chi2_contingency(contingency_mb)
print("Mammals vs Birds p-value:", p)
```

```
Mammals vs Birds p-value: 0.6875948096661336
```

In [61]:
```
# Mammals vs Reptiles
contingency_mr = [
    [30, 146],  # Mammals
    [5, 73]     # Reptiles
]

chi2, p, dof, expected = chi2_contingency(contingency_mr)
print("Mammals vs Reptiles p-value:", p)
```

```
Mammals vs Reptiles p-value: 0.038355590229699
```

The comparison between mammals and birds did not yield a statistically significant result, suggesting similar protection rates between these groups. However, the comparison between mammals and reptiles produced a statistically significant result, indicating that mammals are more likely to require conservation protection than reptiles.

Conclusion

- This project used data visualisation and statistical analysis to examine species and conservation status across four national parks included in the dataset.
- The distribution of conservation status showed that most species were not protected:
  - 5,633 species were not part of conservation
  - 191 species were classified as protected
- Analysis by species category showed that:
  - Mammals and birds had the highest proportions of protected species
  - The difference in conservation status between mammals and birds was not statistically significant
  - A statistically significant difference was found between mammals and reptiles, with mammals more likely to be protected
- Overall, the analysis identified clear patterns in conservation status and species distribution within the limits of the provided data.