Biodiversity in National Parks Data Explration and Analysis

In this file I will explore the data and try to answer some questions about the biodiversity in National Parks. I will use the following datasets:

- `observations.csv`
- `species_info.csv`

Project Goal

- Understand and analyze the biodiversity within national parks based on species observations and conservation status.
- Identify patterns and trends in species distribution and conservation.
- Potentially, provide insights for conservation efforts.

Analysis Plan

Descriptive

- What is the distribution of species categories across all parks?
- What is the distribution of observations across different parks?
- What are the most commonly observed species in each park?
- What are the different conservation statuses and their frequency?
- What are the common names of the most observed scientific names?

Exploratory

- How does the distribution of species categories vary between different parks?
- Are there any correlations between species category and conservation status?
- Which parks have the highest biodiversity (number of unique species)?
- Which species have the largest range (observed in the most parks)?
- Are there species that are observed very frequently in one park, but rarely in others?
- Explore the distribution of observations for species within each conservation status.

Inferential

- Is there a statistically significant difference in the number of observations between different parks?
- Is there a statistically significant association between species category and conservation status?
- Does the number of observations of a species correlate with its conservation status?
- Can we test if a specific park has a significantly higher observation number of a specific category of animal than the average of all parks?

Predictive

- Can we predict the number of observations of a species in a park based on its category and conservation status?
- Can we predict the conservation status of a species based on its observation patterns?

Observing the data

Observing these datasets we can assume that the structure of the data is as follows:

- observations.csv:

Name	Type	Description
scientific_name	String	The scientific name of the species (foreign key)
park_name	String	The name of the national park where the species was observed
Observations	Int	The number of times the species was observed in the park

- species_info.csv:

Name	Type	Description
category	String	The category of the species (mammal, bird, reptile, etc.)
scientific_name	String	The scientific name of the species (primary key)
common_names	String	The common names of the species
conservation_status	String	The species conservation status

Scientific name can be the unique identifier for the species in both datasets.

Cleaning the data

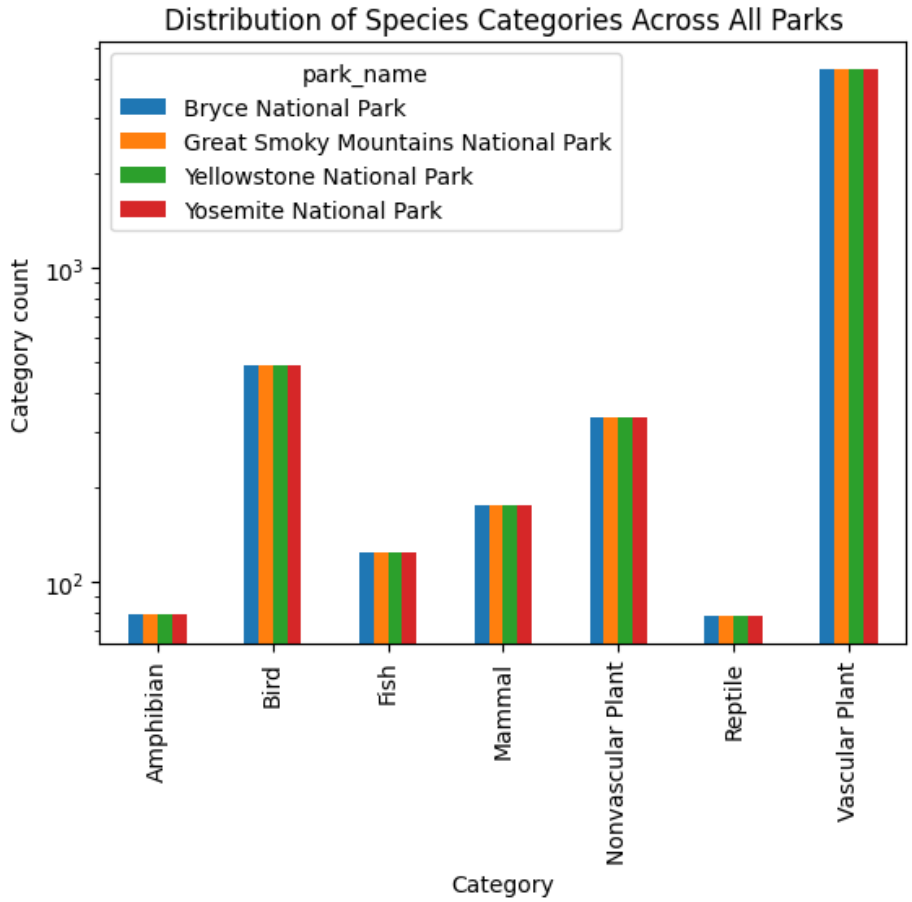
The conservation_status column contains missing values (NaNs). I will replace these with 'No Intervention' as it is likely that these species are not in danger. Also, there are duplicate scientific names in the species_info dataset, I will merge these together and sum the observations.

Descriptive Analysis

1. What is the distribution of species categories across all parks?

The table reveals that the distribution of species categories is consistent across all four parks, with each park containing the full range of species categories observed in the dataset.

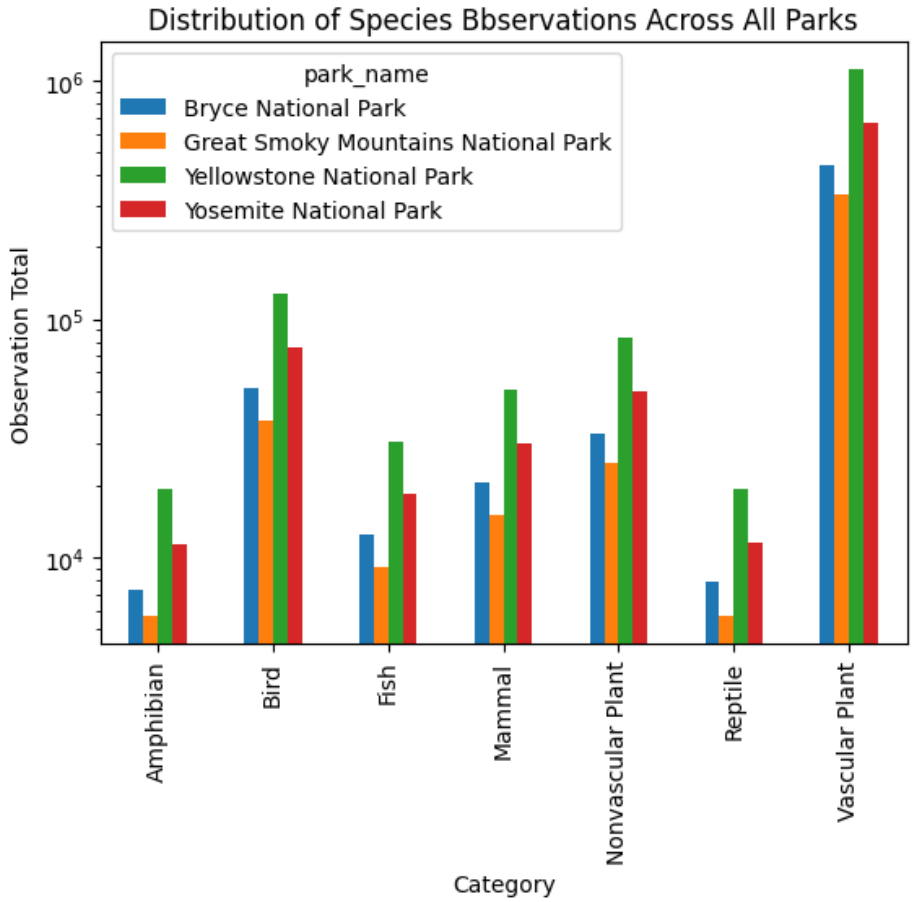
category	Bryce National Park	Great Smoky Mountains National Park	Yellowstone National Park	Yosemite National Park
Amphibian	79	79	79	79
Bird	488	488	488	488
Fish	125	125	125	125
Mammal	176	176	176	176
Nonvascular Plant	333	333	333	333
Reptile	78	78	78	78
Vascular Plant	4262	4262	4262	4262



2. What is the distribution of observations across different parks?

Vascular plants are the most observed species in all parks, followed by birds and mammals. The least observed species are amphibians and reptiles. Yellowstone national park has the most amount of observations across the board, suggesting a higher biodiversity, larger size of the park or more extensive surveying.

category	Bryce National Park	Great Smoky Mountains National Park	Yellowstone National Park	Yosemite National Park
Amphibian	79	79	79	79
Bird	488	488	488	488
Fish	125	125	125	125
Mammal	176	176	176	176
Nonvascular Plant	333	333	333	333
Reptile	78	78	78	78
Vascular Plant	4262	4262	4262	4262



3. What are the most commonly observed species in each park?

The most commonly observed species in each park are as follows:

park_name	scientific_name	observations	common_names	category	conservation_status
Bryce National Park	Columba livia	339	Rock Dove, Common Pigeon, Rock Dove, Rock Pigeon, Rock Pigeon	Bird	No Intervention
Great Smoky Mountains National Park	Streptopelia decaocto	256	Eurasian Collared Dove, Eurasian Collared Dove, Eurasian Collared-Dove, Eurasian Collared-Dove	Bird	No Intervention
Yellowstone National Park	Holcus lanatus	805	Common Velvetgrass, Yorkshire-Fog, Common Velvetgrass, Velvetgrass, Yorkshire Fog, Common Velvet Grass, Velvetgrass	Vascular Plant	No Intervention
Yosemite National Park	Hypochaeris radicata	505	Common Cat's-Ear, False Dandelion, Frogbit, Gosmore, Hairy Cat's Ear, Hairy Catsear, Spotted Catsear, Cat's Ear, Spotted Cat's-Ear, Spotted Cats-Ear, Hairy Cats-Ear, Gosmore	Vascular Plant	No Intervention

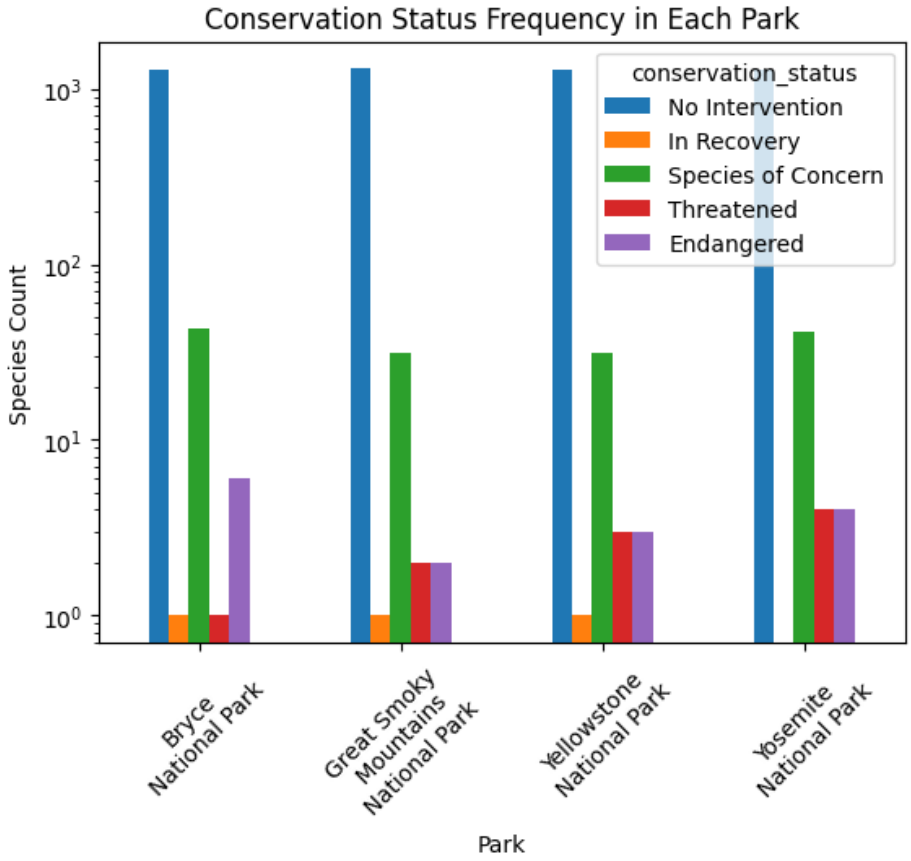
No intervention means that the species is not in danger and does not require any conservation efforts.

4. What are the different conservation statuses and their frequency?

This shows that the most common conservation status is 'No Intervention' across all parks, with bryce national park having the most species of concern and endangered species. With Bryce National Park having the highest count of species that are non 'No Intervention'.

park_name	No Intervention	In Recovery	Species of Concern	Threatened	Endangered
Bryce National Park	1277	1	43	1	6

park_name	No Intervention	In Recovery	Species of Concern	Threatened	Endangered
Great Smoky Mountains National Park	1296	1	31	2	2
Yellowstone National Park	1284	1	31	3	3
Yosemite National Park	1281	0	41	4	4



5. What are the common names of the most observed scientific names?

scientific_name	common_names	category	observations
Holcus lanatus	Common Velvetgrass, Yorkshire-Fog, Common Velvetgrass, Velvetgrass, Yorkshire Fog, Common Velvet Grass, Velvetgrass	Vascular Plant	805
Streptopelia decaocto	Eurasian Collared Dove, Eurasian Collared Dove, Eurasian Collared-Dove, Eurasian Collared-Dove	Bird	771
Puma concolor	Cougar, Mountain Lion, Puma, Mountain Lion, Panther (Mountain Lion)	Mammal	753
Procyon lotor	Raccoon, Common Raccoon, Raccoon, Common Raccoon, Northern Raccoon, Raccoon	Mammal	745
Hypochaeris radicata	Common Cat's-Ear, False Dandelion, Frogbit, Gosmore, Hairy Cat's Ear, Hairy Catsear, Spotted Catsear, Cat's Ear, Spotted Cat's-Ear, Spotted Cats-Ear, Hairy Cats-Ear, Gosmore	Vascular Plant	726

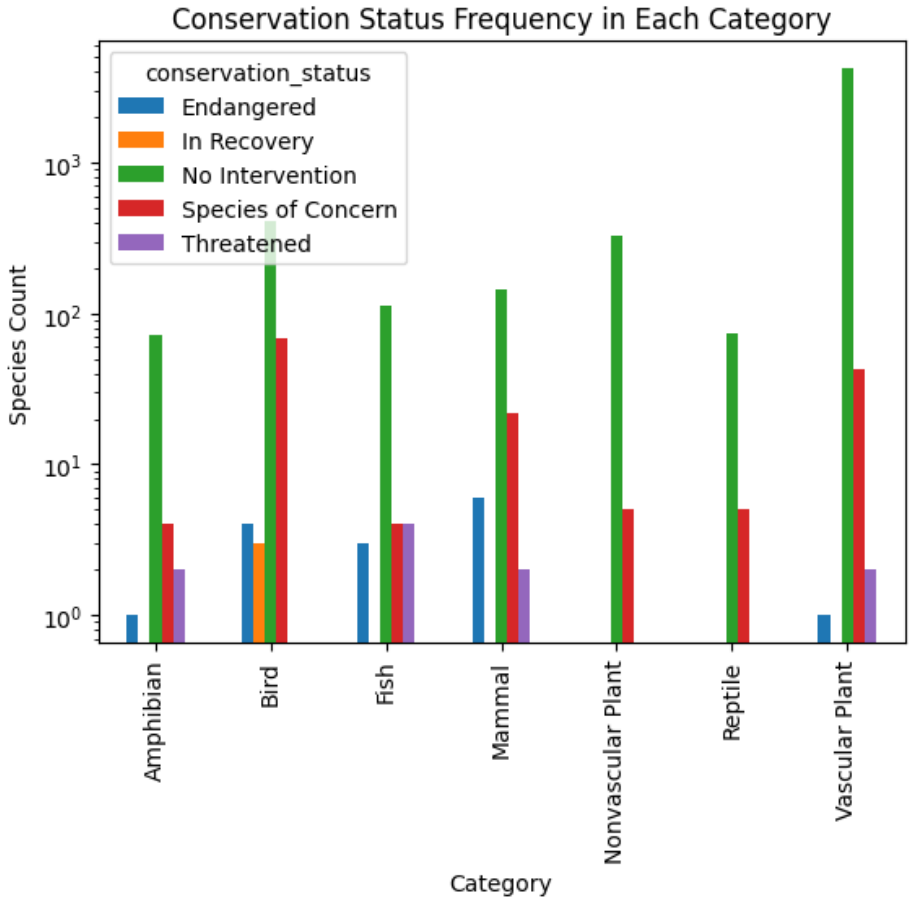
Exploratory

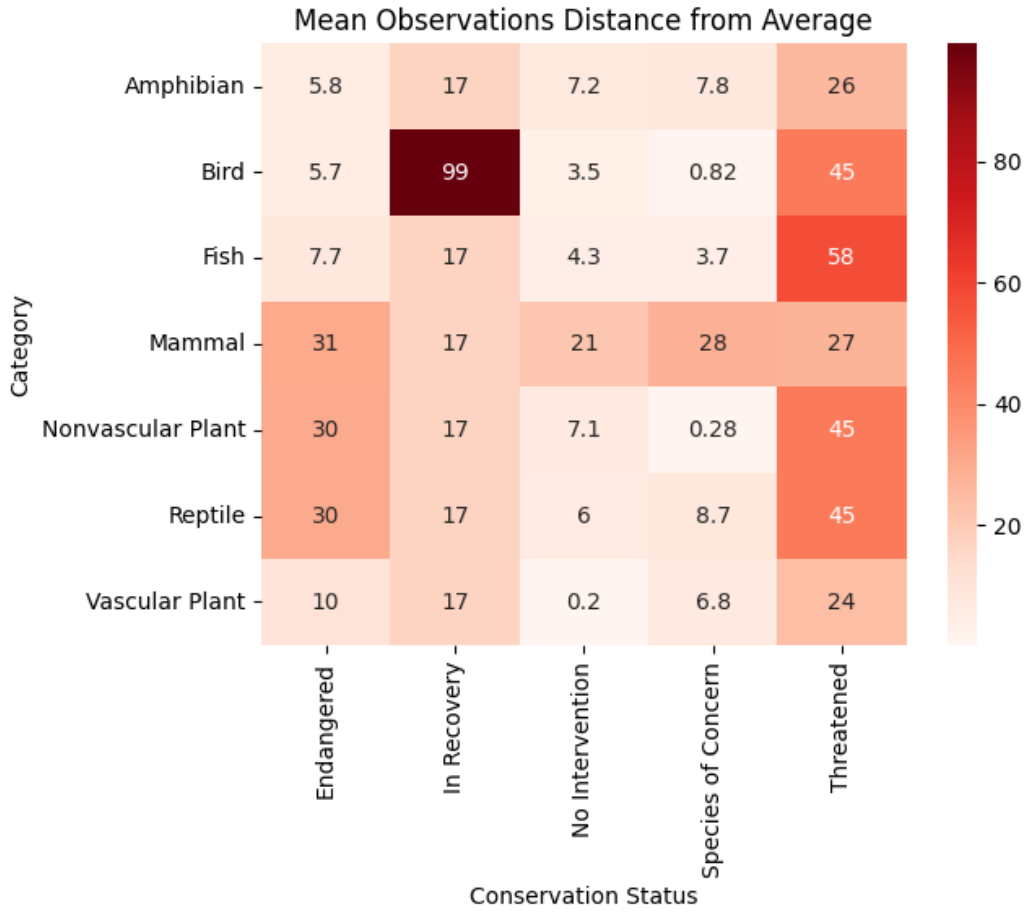
2. Are there any correlations between species category and conservation status?

This table shows the number of species in each category and conservation status. The most common conservation status is 'No Intervention' across all species categories. The least common conservation status is 'Endangered' across all species categories.

category	No Intervention	In Recovery	Species of Concern	Threatened	Endangered

category	No Intervention	In Recovery	Species of Concern	Threatened	Endangered
Amphibian	72	0	4	2	1
Bird	413	3	68	0	4
Fish	114	0	4	4	3
Mammal	146	0	22	2	6
Nonvascular Plant	328	0	5	0	0
Reptile	73	0	5	0	0
Vascular Plant	4216	0	43	2	1





This graphs shows this calculation:

As there are multiple entries of observations for each species, i have taken the mean observation count for each species category and conservation status and used that as for the next part of the analysis. Using these numbers i calcuated the distance from the average for each conservation status. Then i calculated the distance from the average for each species category and conservation status.

$$ \text{Mean Observation Distance} = {\text{Mean Observations for status}} - {\text{Mean Observations for status and category}}$$

This shows us when a value is a considerable distance from the average and alerts us to a potential outlier.

For example, birds in recovery have an average ovservation count of 116, with the rest of the species in recovery have a mean observation count of 0. this makes the average observation count for in recovery 16.6. making the distance from the average 99.4. this suggests that birds in recovery are a potential outlier and should be investigated further.

4. Which species have the largest range (observed in the most parks)?

This table shows the species that have the biggest diffence in observation count acorss parks

scientific_name	Bryce National Park	Great Smoky Mountains National Park	Yellowstone National Park	Yosemite National Park	range	category	conservation_status
Holcus lanatus	296	216	805	463	589	Vascular Plant	No Intervention
Columba livia	339	169	722	423	553	Bird	No Intervention
Hypochaeris radicata	294	195	726	505	531	Vascular Plant	No Intervention
Streptopelia decaocto	301	256	771	457	515	Bird	No Intervention
Puma concolor	311	239	753	408	514	Mammal	No Intervention

5. Are there species that are observed very frequently in one park, but rarely in others?

This table shows the species that have a large difference in observation count across parks. This suggests that these species are more common in one park than the others.

scientific_name	Bryce National Park	Great Smoky Mountains National Park	Yellowstone National Park	Yosemite National Park	min	range	range-min	category	conservation_status
Columba livia	339	169	722	423	169	553	384	Bird	No Intervention
Holcus lanatus	296	216	805	463	216	589	373	Vascular Plant	No Intervention
Leucosticte tephrocotis	170	81	530	341	81	449	368	Bird	No Intervention
Equisetum hyemale var. affine	139	65	496	312	65	431	366	Vascular Plant	No Intervention
Cerastium fontanum ssp. vulgare	225	98	544	343	98	446	348	Vascular Plant	No Intervention

Inferential

1. Is there a statistically significant difference in the number of observations between different parks?

Null hypothesis: There is no statistically significant difference in the number of observations between different parks. p-value < 0.05 indicates a statistically significant difference between the means of the two groups being compared.

Using the t-test to compare the means of the number of observations between different parks, we can see that there is a statistically significant difference in the number of observations between some parks.

Statistically significant difference in the number of observations between different parks

park1	park2	t_stat	p_val	significant?
Bryce National Park	Great Smoky Mountains National Park	0.847093	0.39698	False
Bryce National Park	Yellowstone National Park	-5.09611	3.58206e-07	True
Bryce National Park	Yosemite National Park	-1.68771	0.0915237	False
Great Smoky Mountains National Park	Yellowstone National Park	-6.83701	8.95635e-12	True
Great Smoky Mountains National Park	Yosemite National Park	-2.91601	0.00355958	True
Yellowstone National Park	Yosemite National Park	1.72107	0.0852935	False

this test shows that there is a statistically significant difference between the number of observations in the great smokey mountains and all parks. A bref look at the size of the parks, this does not follow the statistical significance of the observation counts suggeesting that there is different reason for this significance.

Statistically significant difference in the number of observations between different parks and all parks

park	t_stat	p_val	significant?	mean observation count
Bryce National Park	-1.48418	0.137818	False	103.957
Great Smoky Mountains National Park	-2.68188	0.0073427	True	77.9318
Yellowstone National Park	1.82384	0.0682297	False	260.524

park	t_stat	p_val	significant?	mean observation count
Yosemite National Park	0.156432	0.875698	False	155.808

This statistcal significance could be due the the low mean observation count of the species in the park.

2. Is there a statistically significant association between species category and conservation status?

null hypothesis: There is no statistically significant association between species category and conservation status. p-value < 0.05 indicates a statistically significant association between the two categorical variables.

Is there a statistically significant association between species category and conservation status?

```
Chi2 value:  591.1513416161314
P value:  1.6816257270187694e-109
Degrees of Freedom:  24

Significant?:  True
```

Expected:

	0	1	2	3	4
0	0.21386	0.0427721	76.4479	2.15286	0.142574
1	1.32106	0.264212	472.235	13.2987	0.880707
2	0.338387	0.0676773	120.962	3.40642	0.225591
3	0.476448	0.0952897	170.314	4.79625	0.317632
4	0.901462	0.180292	322.243	9.07472	0.600975
5	0.211153	0.0422306	75.4802	2.12561	0.140769
6	11.5376	2.30753	4124.32	116.145	7.69175

There is sufficeint evicence to reject the null hypothessis therefore there is a connection between the species category and conservation status.

3. Does the number of observations of a species correlate with its conservation status?

```
Correlation:  -0.06641341665896287
Significant?:  False
```

This shows that there is no correlation btween the number of observations of a species and its consrvation status.

4. Can we test if a specific park has a significantly higher observation number of a specific category of animal than the average of all parks?

park	category	t_stat	p_val	significant?
Bryce National Park	Nonvascular Plant	-2.15734	0.0316966	True
Bryce National Park	Amphibian	-2.02011	0.0468053	True
Great Smoky Mountains National Park	Vascular Plant	-2.74459	0.00608402	True
Great Smoky Mountains National Park	Nonvascular Plant	-3.49193	0.000544421	True
Great Smoky Mountains National Park	Bird	-2.43217	0.0153678	True
Great Smoky Mountains National Park	Amphibian	-2.77809	0.00684664	True

park	category	t_stat	p_val	significant?
Great Smoky Mountains National Park	Reptile	-3.04293	0.0032038	True
Great Smoky Mountains National Park	Fish	-2.85293	0.00507741	True
Yellowstone National Park	Nonvascular Plant	5.33773	1.74779e-07	True
Yellowstone National Park	Amphibian	2.31978	0.0229697	True
Yellowstone National Park	Reptile	2.98964	0.00374825	True
Yellowstone National Park	Fish	2.13785	0.0344909	True

This shows table shows the categories that have a statistically significant difference in the number of observations between the park and the average of all parks.

Predictive

1. Can we predict conservation status based on observations, species category, and park name?

Within the jupyter notebook I experiemente with linear regression between two variables. The models produced a high accuracy however due to the skew in the dataset it only outputed 'No Intervention' as the conservation status. The accuracy for this model was 96%, however the percentge of 'No Intervention' species in the dataset is 96%. This shows that the model is not a good predictor of conservation status.

Conclusion

In this report, I explored the biodiversity of national parks using the observations.csv and species_info.csv datasets. My analysis revealed that while the distribution of species categories is consistent across all parks, the number of observations varies significantly, with Yellowstone National Park showing the highest overall counts. Vascular plants are the most observed species. I identified the most commonly observed species in each park, with Holcus lanatus, Streptopelia decaocto, Puma concolor, Procyon lotor, and Hypochaeris radicata being the top five.

The majority of species fall under the 'No Intervention' conservation status, but there are differences in the frequency of other statuses across parks. I found a statistically significant association between species category and conservation status. There was no correlation between the number of observations of a species and its conservation status. Further, I identified specific parks and species categories that have significantly higher or lower observation numbers than the average across all parks.

Attempts to predict conservation status based on observation patterns were unsuccessful due to the overwhelming prevalence of the 'No Intervention' category in the dataset, which skewed the model's predictions.

Further research could explore the factors contributing to the variation in species observations between parks, such as park size, habitat diversity, and survey effort. Additionally, a more balanced dataset with a wider representation of conservation statuses would be necessary to develop a reliable predictive model for conservation status.