Biodiversity in National Parks

Aspiring to become a Data Scientist in Machine Learning, I embarked on a learning journey by acquiring skills in SQL, Python, and various Python libraries through Codecademy's career path. This experience inspired me to undertake a captivating portfolio project centered around the theme of biodiversity, despite having limited knowledge in Biology.

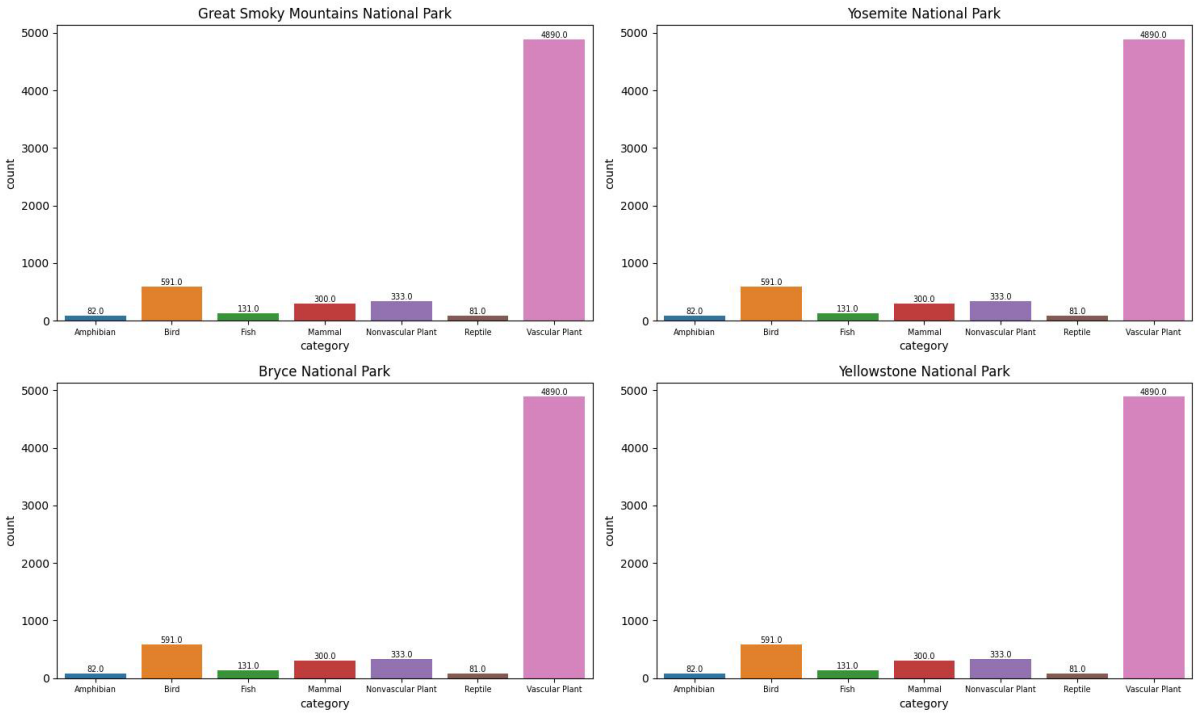
Precedence

In my project, I obtained two CSV files from Codecademy, each containing valuable information about different species. I meticulously examined the columns and skillfully merged the datasets using the "scientific name" as a common identifier.

2. observations.csv:	1. species_info:
2.1. 'scientific_name'	1.1. 'category'
2.2. 'park_name'	1.2. 'scientific_name',
2.3. 'observations'	1.3. 'common_names',
	1.4. 'conservation_status'

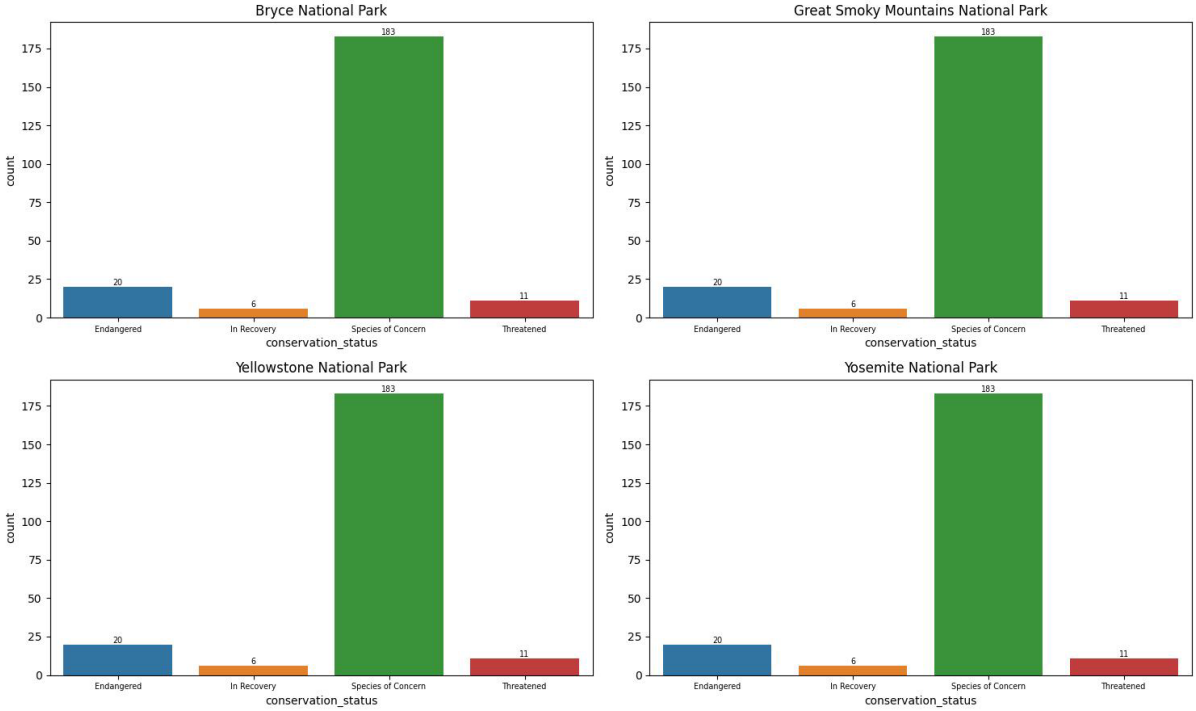
What I found

At first, I wasn't sure if I could write the code properly. But I managed to create some bar plots using a for loop. The data showed that there are more vascular plants than other types of species. This data might not be 100% accurate, but it tells us that each park has similar kinds of species. So, if we want to see a specific type of animal, it doesn't matter which park we visit.



It has MNAR Nulls because and I replaced them with “Excluded Status”. I don’t show it in the graphics because it was 6188 with this bar we would not see the other ones. Then I also looked for the conservation’s status in each park. There are also the same counts for the different status. Most of the species are concerned to be extinguished in the future and there are 11 species that are threatened today and

soon. The national parks try to recover this species. Like we see on the graphics. There are six species in recovery. This must be a concern for us.



Conclusion

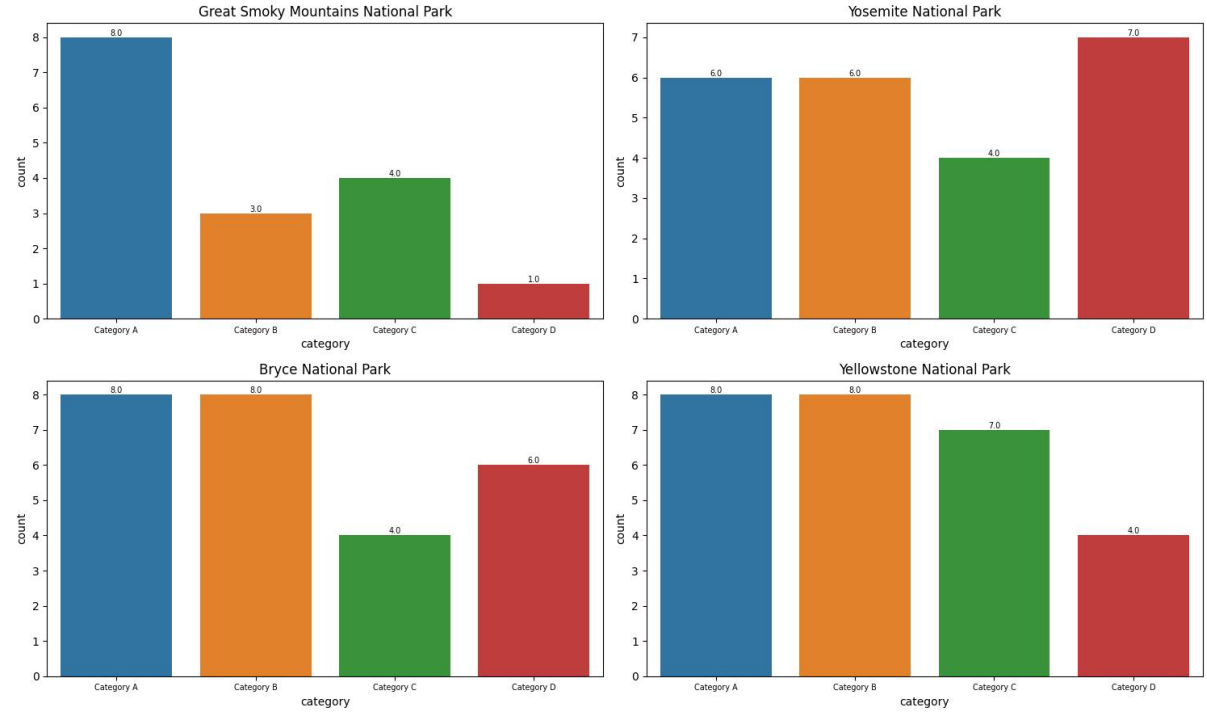
National parks have a lot of different animal species, but some of them are in danger. We should be careful not to disturb plants or animals that are at risk when we visit these parks. The parks are doing their best to help these species, and it's important for us to support their efforts. By being mindful of the wildlife and our actions, we can play a role in protecting these valuable species for the future.

I think the same values comes that this is a inference data and the national parks have the similar animals

Appendix

```
park_names = ['Great Smoky Mountains National Park', 'Yosemite National Park',
'Bryce National Park', 'Yellowstone National Park']

categories = ['Category A', 'Category B', 'Category C', 'Category D']

data = []
for park_name in park_names:
    for category in categories:
        count = np.random.randint(1, 10)
        data.append({'park_name': park_name, 'category': category, 'count':
count})

test_df = pd.DataFrame(data)

plt.figure(figsize=(15, 9))

for i, park_name in enumerate(park_names, 1):
```

```
    plt.subplot(2, 2, i)

    sns.barplot(x="category", y="count", data=test_df[test_df["park_name"] ==
park_name])
    plt.title(park_name)
    plt.xticks(fontsize="x-small")

    ax = plt.gca()
    for p in ax.patches:
        height = p.get_height()
        ax.annotate(f'{height}', (p.get_x() + p.get_width() / 2., height),
                    ha='center', va='bottom', fontsize='x-small')

plt.tight_layout()
plt.savefig("test.jpg")
plt.show()
plt.clf()
```



```
groups = merged_df.groupby('park_name')['category'].count()

print(groups)

sorted_df = merged_df.sort_values(by='category')
groups = sorted_df.groupby('park_name')

for key, group in groups:
    count = len(group)
    print(f"Group '{key}': Count = {count}")
```

```
park_name
Bryce National Park                    6408
Great Smoky Mountains National Park    6408
Yellowstone National Park              6408
Yosemite National Park                 6408
Name: category, dtype: int64
Group 'Bryce National Park': Count = 6408
Group 'Great Smoky Mountains National Park': Count = 6408
Group 'Yellowstone National Park': Count = 6408
Group 'Yosemite National Park': Count = 6408
```