Biodiversity_project.py

March 14, 2023

[2]:
```
#Import all the neccessary libraries for the project

import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

obs = pd.read_csv("observations.csv")
species = pd.read_csv("species_info.csv")
```

[3]:
```
obs.head()
```

[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

[4]:
```
species.head()
```

[4]:

	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

[5]:
```
print(species.groupby("category").size())
print(species["conservation_status"].isna().sum())
```

```
print(species.groupby("conservation_status").size())
```

```
category
Amphibian              80
Bird                  521
Fish                  127
Mammal                214
Nonvascular Plant     333
Reptile                79
Vascular Plant       4470
dtype: int64
5633
conservation_status
Endangered             16
In Recovery             4
Species of Concern    161
Threatened             10
dtype: int64
```

[6]:
```
#Let's fill in the NaN values and show the conservation status by groups.
species.fillna("Non observed", inplace=True)
print(species.groupby("conservation_status").size())
```

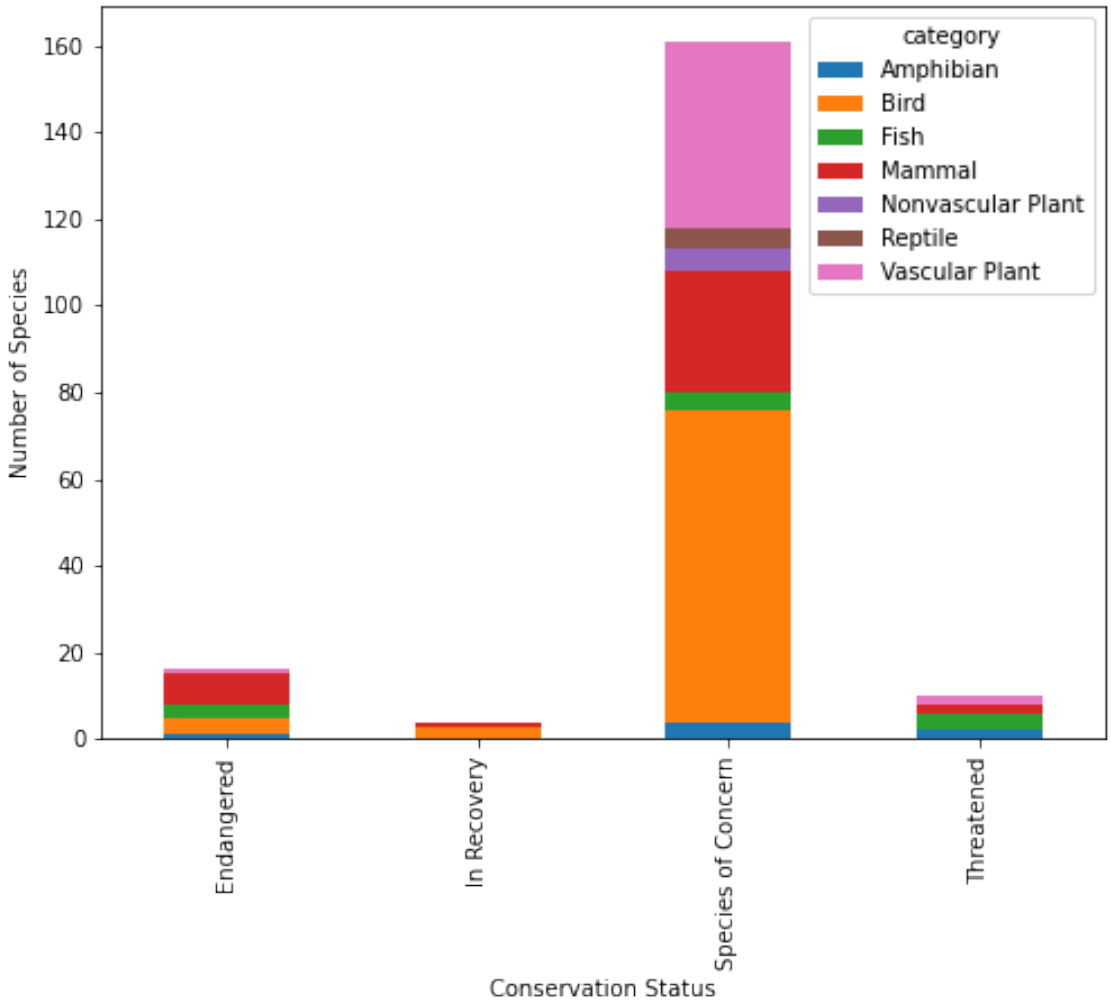
```
conservation_status
Endangered             16
In Recovery             4
Non observed         5633
Species of Concern    161
Threatened             10
dtype: int64
```

[ ]:
```

```

[7]:
```
#Lets look at the conservation status table without looking at the "Nan" values.
ConservationCategory = species[species["conservation_status"] != "Non␣
↪observed"].groupby(["conservation_status", "category"])['scientific_name'].
↪count().unstack(level=1)
```

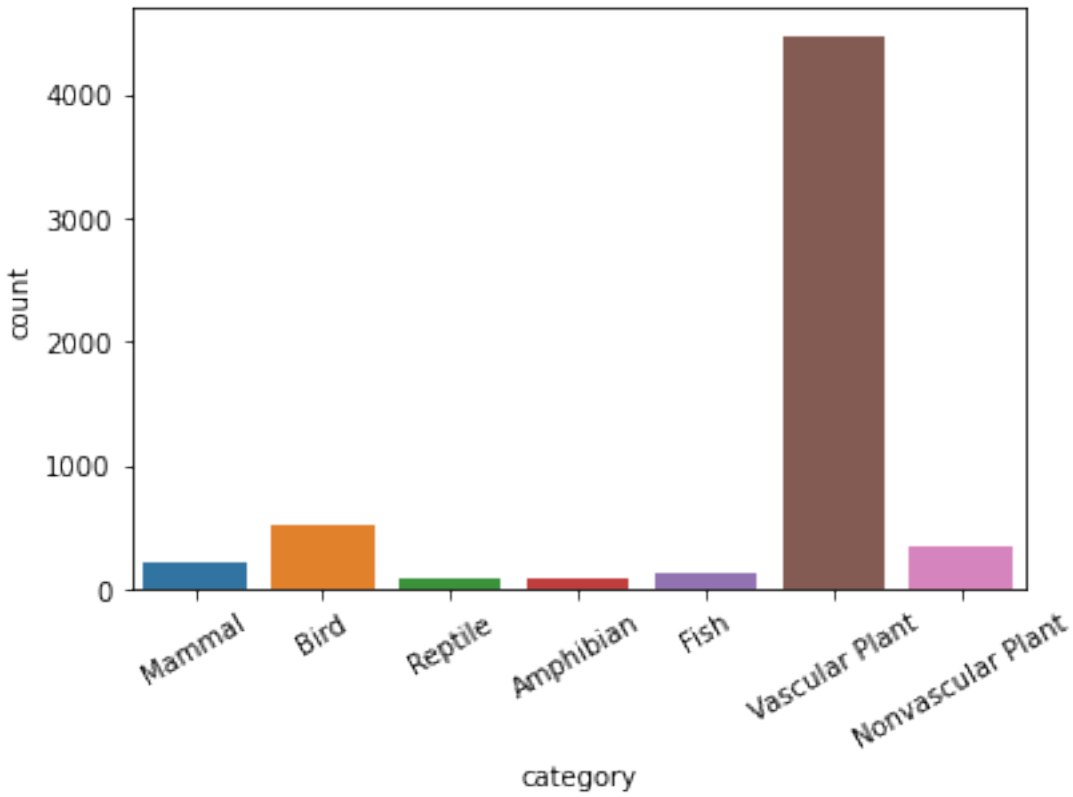
[8]:
```
ax = ConservationCategory.plot(kind='bar', figsize=(8,6), stacked=True)

ax.set_xlabel("Conservation Status")
ax.set_ylabel("Number of Species")
```

[8]: Text(0, 0.5, 'Number of Species')



[9]:
```
ax = sns.countplot(species["category"])
ax.set_xticklabels(species["category"].unique(),rotation=30, fontsize=10)
plt.show()
```

```
/Users/oskarwigen/opt/anaconda3/lib/python3.9/site-
packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variable
as a keyword arg: x. From version 0.12, the only valid positional argument will
be `data`, and passing other arguments without an explicit keyword will result
in an error or misinterpretation.
  warnings.warn(
```



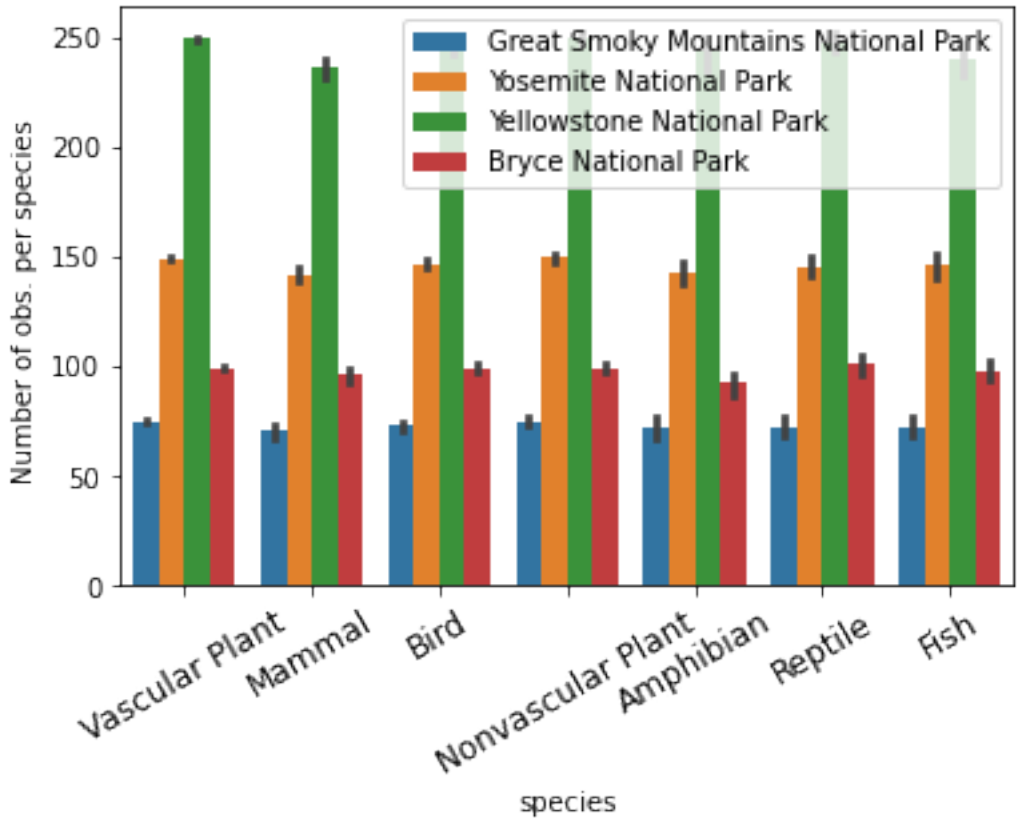
[10]:
```
merged_data = pd.merge(obs, species)

merged_data.head()
```

[10]:

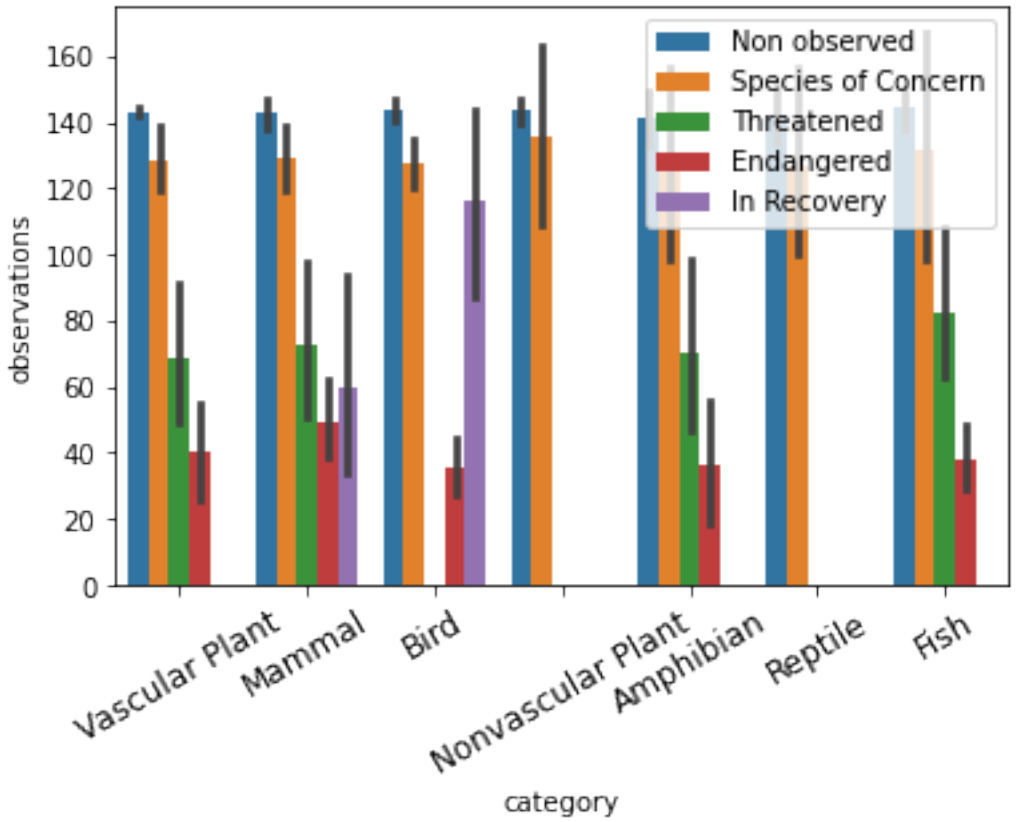
	scientific_name	park_name	observations	category	common_names	conservation_status
0	Vicia benghalensis	Great Smoky Mountains National Park	68	Vascular Plant	Purple Vetch, Reddish Tufted Vetch	Non observed
1	Vicia benghalensis	Yosemite National Park	148	Vascular Plant	Purple Vetch, Reddish Tufted Vetch	Non observed
2	Vicia benghalensis	Yellowstone National Park	247	Vascular Plant	Purple Vetch, Reddish Tufted Vetch	Non observed
3	Vicia benghalensis	Bryce National Park	104	Vascular Plant	Purple Vetch, Reddish Tufted Vetch	Non observed
4	Neovison vison	Great Smoky Mountains National Park	77	Mammal	American Mink	Non observed

[11]:
```
#Which species are seen the most at each park
ax = sns.barplot(x = "category", y = "observations", hue = "park_name", data =␣
↪merged_data)
```

```
ax.set_xticklabels(merged_data.category.unique(), rotation = 30, fontsize = 12)
ax.legend()
ax.set_xlabel("species")
ax.set_ylabel("Number of obs. per species ")
plt.show()

#It looks like YellowStone is the biggest national park since it has the most␣
 ↪observations in every species category.
```



[12]:
```
#Lets see if theres any correlation between the number of a species and the␣
 ↪risk of concern.
# We
ax = sns.barplot(x = "category", y= "observations", hue = "conservation_status"␣
 ↪,data=merged_data)
ax.set_xticklabels(merged_data.category.unique(), rotation = 30, fontsize = 12)
ax.legend(loc= "upper right")
plt.show()
```


Non observed
Species of Concern
Threatened
Endangered
In Recovery
observations
0
20
40
60
80
100
120
140
160
Vascular Plant
Mammal
Bird
Nonvascular Plant
Amphibian
Reptile
Fish
category

[ ]:

[ ]: