Exploratory Data Analysis on Biodiversity in National Parks

First, let us import the datasets needed for the same.

```
import pandas as pd
```

```
species=pd.read_csv('species.csv')
species.head()
```

Now let us take a glance at our data and look at the quick summary table on it.

	Species ID	Park_Name	Category	Order	Family	Scientific_Name	Common_Names	Record Status
0	ACAD-1000	Acadia National Park	Mammal	Artiodactyla	Cervidae	Alces alces	Moose	Approved
1	ACAD-1001	Acadia National Park	Mammal	Artiodactyla	Cervidae	Odocoileus virginianus	Northern White-Tailed Deer, Virginia Deer, Whi...	Approved
2	ACAD-1002	Acadia National Park	Mammal	Carnivora	Canidae	Canis latrans	Coyote, Eastern Coyote	Approved
3	ACAD-1003	Acadia National Park	Mammal	Carnivora	Canidae	Canis lupus	Eastern Timber Wolf, Gray Wolf, Timber Wolf	Approved
4	ACAD-1004	Acadia National Park	Mammal	Carnivora	Canidae	Vulpes vulpes	Black Fox, Cross Fox, Eastern Red Fox, Fox, Re...	Approved

	Species ID	Park_Name	Category	Order	Family	Scientific_Name	Common_Names	Record Status	Occurrence	Nativeness	Abundance	Seasonality	Conservation_Status
count	119248	119248	119248	117776	117736	119248	119248	119248	99106	94203	76306	20157	4718
unique	119248	56	14	554	2332	46022	35823	54	7	5	8	24	11
top	ROMO-2575	Great Smoky Mountains National Park	Vascular Plant	Poales	Asteraceae	Falco peregrinus	None	Approved	Present	Native	Unknown	Breeder	Species of Concern
freq	1	6623	65221	11453	8843	56	27147	86254	83278	75950	28119	12214	3843

Let us import the additional dataset as well.

```
parks=pd.read_csv('parks.csv')
parks.head()
```

	Park Code	Park Name	State	Acres	Latitude	Longitude
0	ACAD	Acadia National Park	ME	47390	44.35	-68.21
1	ARCH	Arches National Park	UT	76519	38.68	-109.57
2	BADL	Badlands National Park	SD	242756	43.75	-102.50
3	BIBE	Big Bend National Park	TX	801163	29.25	-103.25
4	BISC	Biscayne National Park	FL	172924	25.65	-80.08

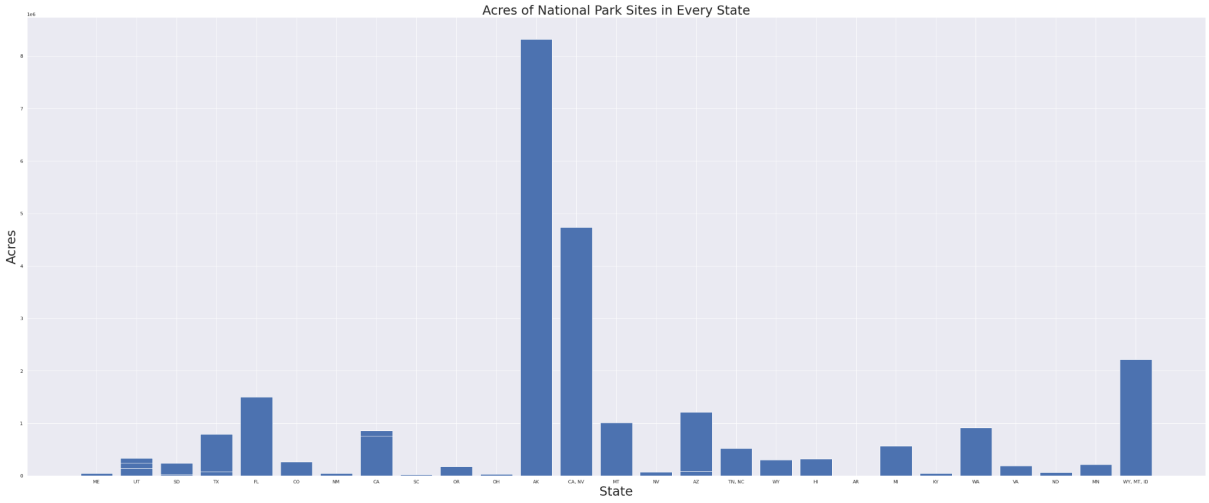
With this dataset, we can find out the top 5 largest National Parks by area.

```
Top5 = parks.sort_values(
    by=["Acres"],
    ascending = False)
Top5.head()
#Top 5 national parks with respect to Acres
```

	Park Code	Park Name	State	Acres	Latitude	Longitude
52	WRST	Wrangell - St Elias National Park and Preserve	AK	8323148	61.00	-142.00
18	GAAR	Gates Of The Arctic National Park and Preserve	AK	7523898	67.78	-153.30
15	DEVA	Death Valley National Park	CA, NV	4740912	36.24	-116.82
32	KATM	Katmai National Park and Preserve	AK	3674530	58.50	-155.00
14	DENA	Denali National Park and Preserve	AK	3372402	63.33	-150.50

Now let us find out which state has the most acres of national park area under it.

```
import matplotlib.pyplot as plt
plt.figure(figsize=(50, 20))
plt.bar(parks.State,parks.Acres)
plt.title('Acres of National Park Sites in Every State', fontsize=30)
plt.xlabel('State', fontsize=30)
plt.ylabel('Acres', fontsize=30)
```



We can see AK which is Alaska is the state which has the most acres dedicated to national parks.

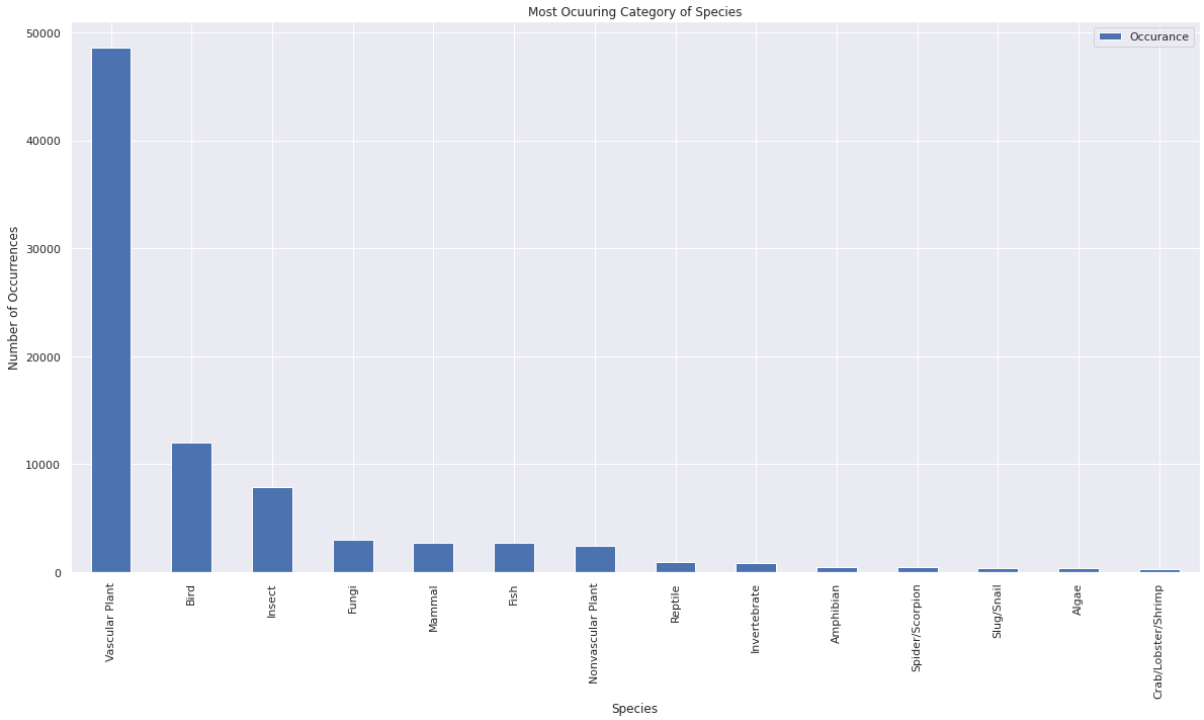
Now let us find out which is the most occurring species present in the parks.

```
occurance_present = species[(species.Occurrence=='Present')]
count_occurance_present= occurance_present['Category'].value_counts().to_frame('Occurance')
count_occurance_present
```

	Occurance
Vascular Plant	48557
Bird	12051
Insect	7920
Fungi	3001
Mammal	2763
Fish	2729
Nonvascular Plant	2480
Reptile	931
Invertebrate	892
Amphibian	461
Spider/Scorpion	434
Slug/Snail	378
Algae	369
Crab/Lobster/Shrimp	312

Now, let us plot the same.

```
count_occurance_present.plot(kind='bar')
plt.rcParams["figure.figsize"] = (20,10)
plt.title('Most Ocuuring Category of Species')
plt.ylabel('Number of Occurrences', fontsize=12)
plt.xlabel('Species', fontsize=12)
```

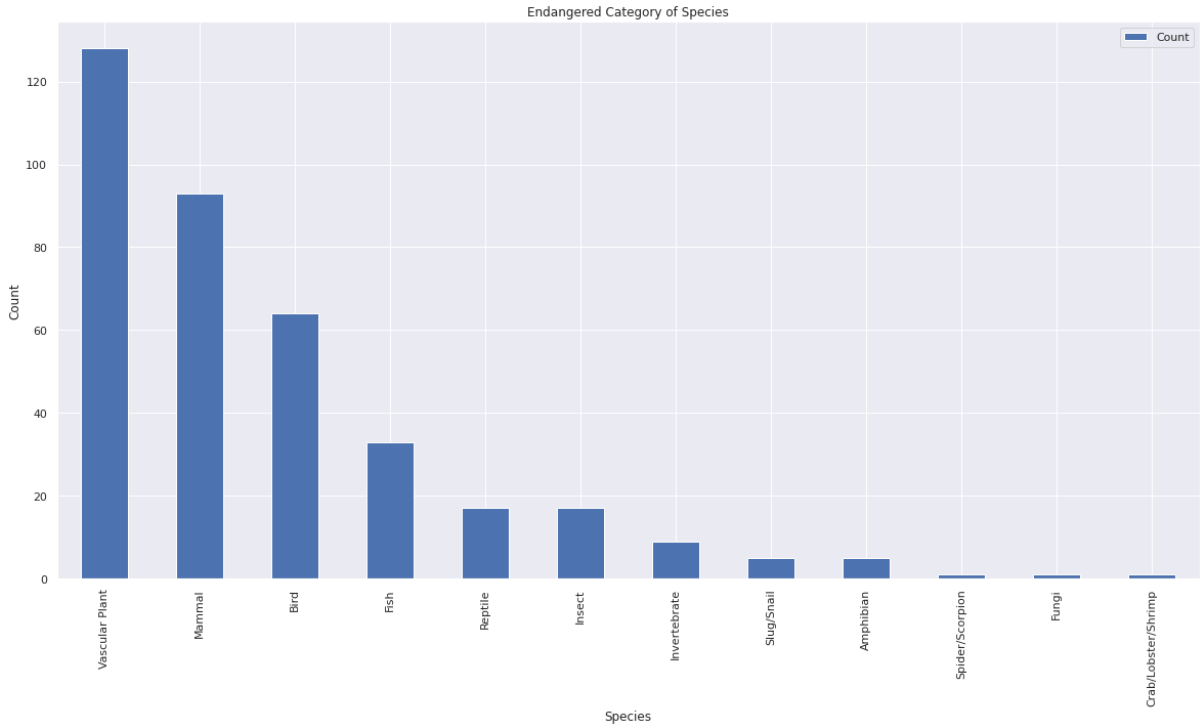


Now let us find out which species has the most endangered count

```
endangered_species = species[(species.Conservation_Status=='Endangered')]
count_endangered_species= endangered_species['Category'].value_counts().to_frame('Count')
count_endangered_species
```

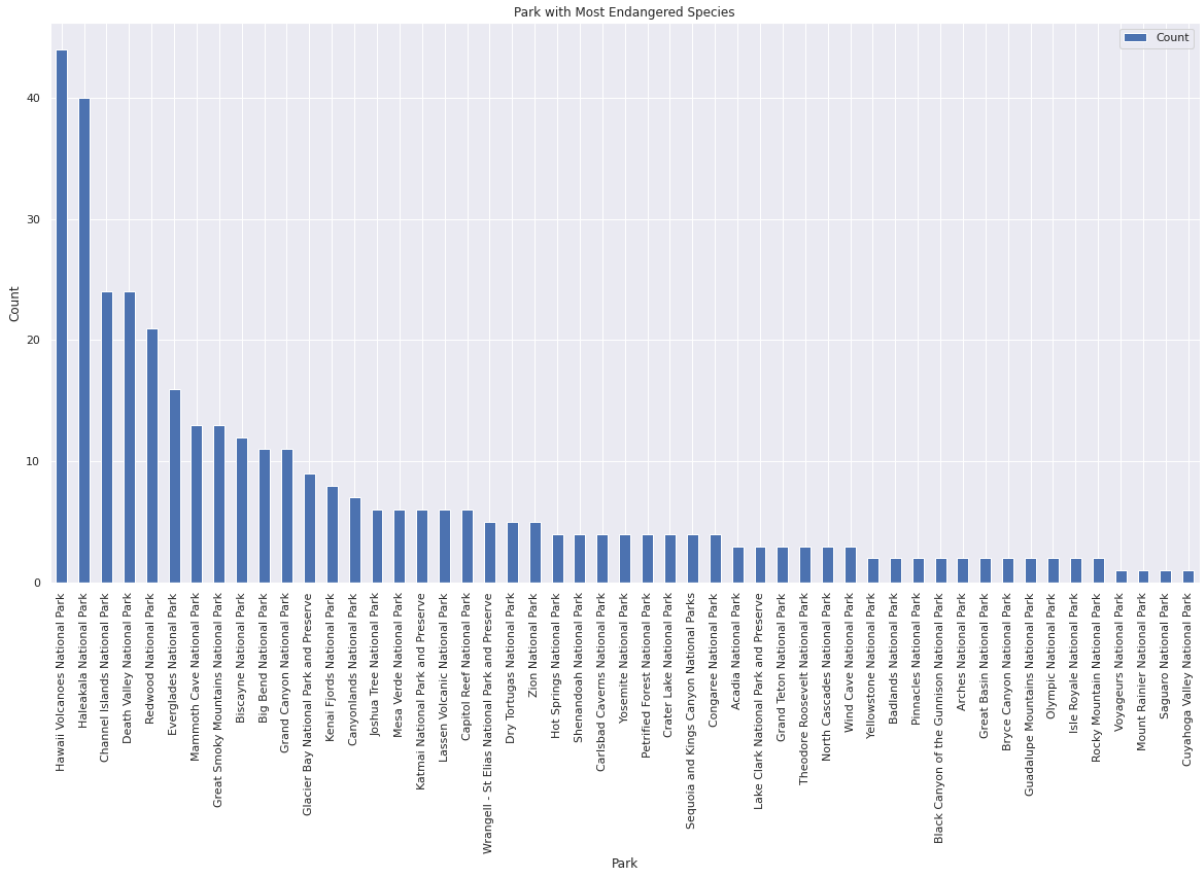
	Count
Vascular Plant	128
Mammal	93
Bird	64
Fish	33
Reptile	17
Insect	17
Invertebrate	9
Slug/Snail	5
Amphibian	5
Spider/Scorpion	1
Fungi	1
Crab/Lobster/Shrimp	1

```
count_endangered_species.plot(kind='bar')
plt.title('Endangered Category of Species')
plt.ylabel('Count', fontsize=12)
plt.xlabel('Species', fontsize=12)
```



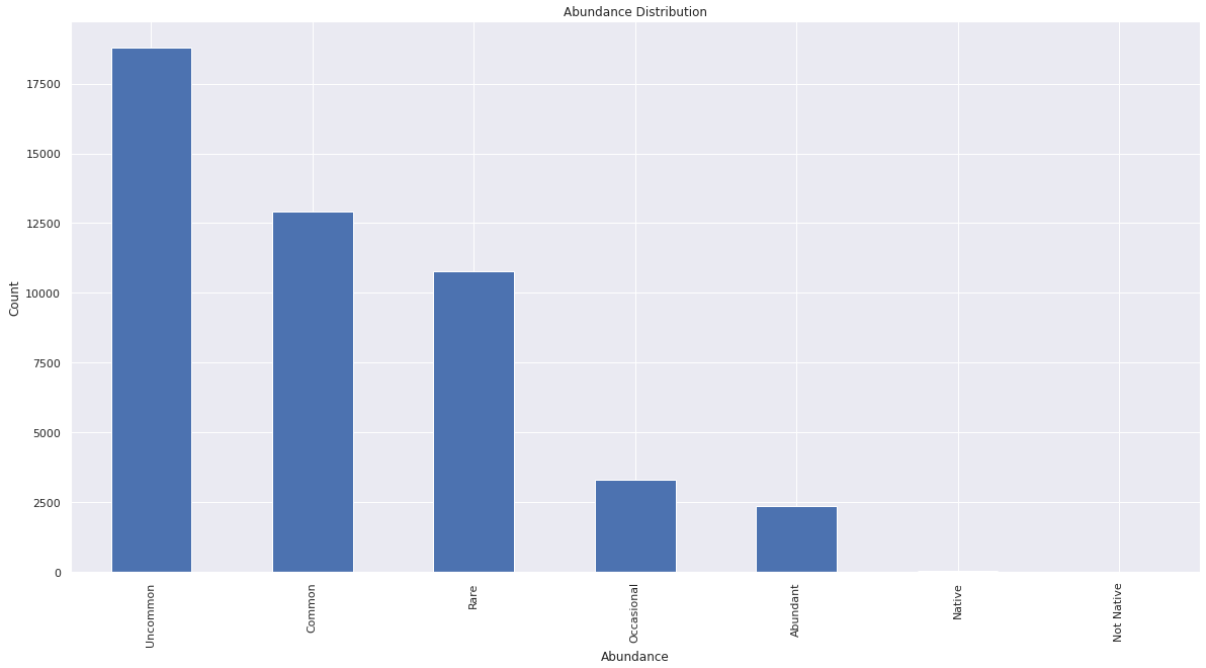
Now let us find out which national park houses the most endangered species.

```
endangered_park = species[(species.Conservation_Status=='Endangered')]
count_endangered_park= endangered_park['Park_Name'].value_counts().to_frame('Count')
plt.rcParams["figure.figsize"] = (20,10)
count_endangered_park.plot(kind='bar')
plt.title('Park with Most Endangered Species')
plt.ylabel('Count', fontsize=12)
plt.xlabel('Park', fontsize=12)
```

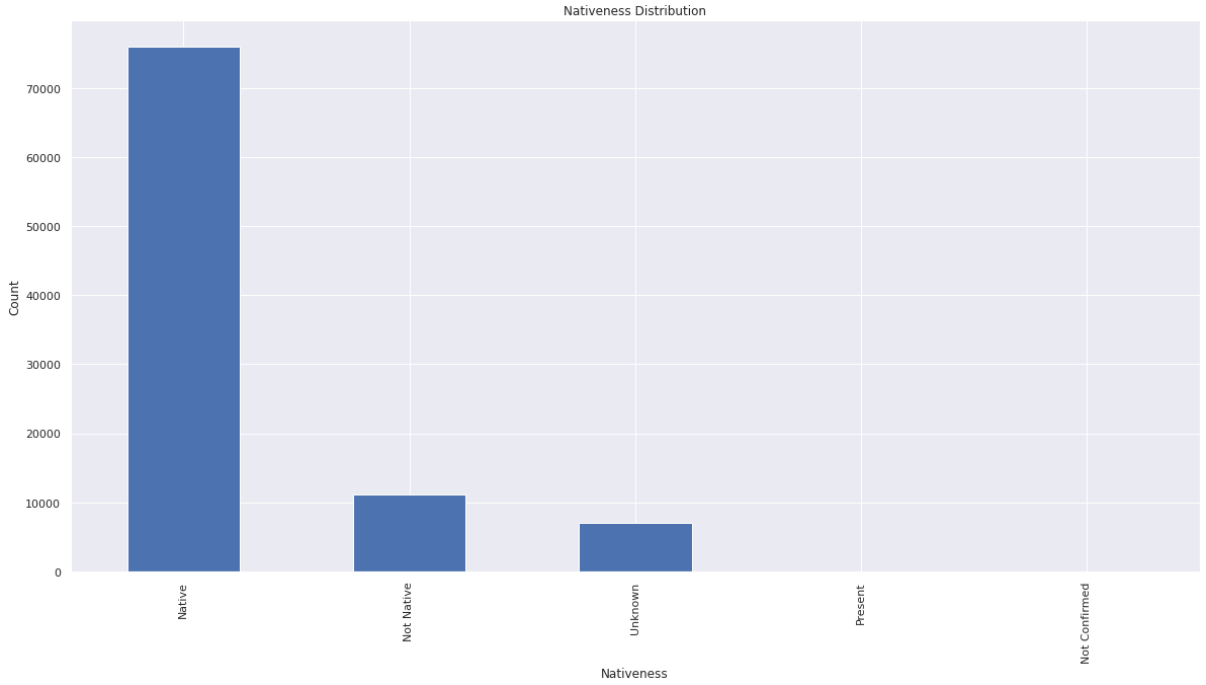


And finally, let us find the abundance and nativeness distribution of the different species in the parks.

```
abundance_check = species[(species.Abundance!='Unknown')]
abundance_check['Abundance'].value_counts().plot.bar()
plt.title('Abundance Distribution')
plt.ylabel('Count', fontsize=12)
plt.xlabel('Abundance', fontsize=12)
```




```
species['Nativeness'].value_counts().plot.bar()
plt.title('Nativeness Distribution')
plt.ylabel('Count', fontsize=12)
plt.xlabel('Nativeness', fontsize=12)
```



Conclusion: There are a number of takeaways from this EDA, they are

- Wrangell - St Elias National Park and Preserve in Alaska is the largest national park with respect to area and 4 of the top 5 national parks are in Alaska.
- Due to the above reason Alaska is also the state with the most area dedicated to national parks.
- Vascular plants are the species with the highest occurrence in the parks followed by birds, insects and fungi respectively.
- With respect to endangered species, the order is vascular plants, mammals, birds and fishes. Here we have to take into consideration the occurrence number as well. Mammals were not in the top 4 of the occurrence number but are in the endangered top 4. This shows an increase in the number of endangered species in mammals.
- We also found out Hawaii Volcanos National Park houses the most endangered species with Haleakala National Park coming in close at second.
- We performed an abundance check which showed a good number of uncommon and rare species found in the parks
- We also performed a nativeness check which showed most species were native to the region

Dataset:https://www.kaggle.com/nationalparkservice/park-biodiversity

- Nikhil Thomas A
  2139446