Biodiversity in National Park Portfolio Project

Duc Minh Bui

Table of Content

1. Project Scoping

Project Scoping: Goals

This project scope is to examine the data of the interpret biodiversity data from the National Parks Service about endangered species in different parks.

The project will provide data analyses on the conservation statuses of these species and investigate if there are any patterns or themes to the types of species that become endangered.

Project Scoping: Questions to Answer

Univariate

- What is the distribution of animal observations?
- What is the difference in each Category count?
- Which park have the most observations?

Bivariate

- Does certain species are more likely to be threatened or safe?
- Count of 'Castor canadensis' in each Park, the most famous scientific name among the species

2. Data Information and Acknowledgement

Data Acknowledgement

The data is acquired through the Codecademy's Business Intelligence Data Analyst Course.

Link:
https://www.codecademy.com/paths/bi-data-analyst/tracks/dsf-portfolio-project/modules/dscp-biodiversity-in-national-parks/kanban_projects/biodiversity-in-national-parks-portfolio-project

Data Information

The Project works with two main datasets:

species_info.csv:

- category - class of animal
- scientific_name - the scientific name of each species
- common_name - the common names of each species
- conservation_status - each species' current conservation status

observations.csv:

- scientific_name - the scientific name of each species
- park_name - Park where species were found
- observations - the number of times each species was observed at park

3. Examining and Cleaning Data

Examining Data

First, we examine the datasets' first rows and columns' names

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

In [4]:
```
observations.head()
```

Out[4]:

	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

In [5]:
```
species.head()
```

Out[5]:

	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

In [6]:
```
print(observations.columns)
print(species.columns)
```

```
Index(['scientific_name', 'park_name', 'observations'], dtype='object')
Index(['category', 'scientific_name', 'common_names', 'conservation_status'], dtype='object')
```

Examining Data

Next, we examine the dataframes' summary statistics and shape

In [7]:
```
species.describe()
```

Out[7]:

	category	scientific_name	common_names	conservation_status
count	5824	5824	5824	191
unique	7	5541	5504	4
top	Vascular Plant	Castor canadensis	Brachythecium Moss	Species of Concern
freq	4470	3	7	161

In [8]:
```
observations.describe()
```

Out[8]:

	observations
count	23296.000000
mean	142.287904
std	69.890532
min	9.000000
25%	86.000000
50%	124.000000
75%	195.000000
max	321.000000

In [9]:
```
print(observations.shape)
print(species.shape)
```

```
(23296, 3)
(5824, 4)
```

Handling Missing Data

Then, we can check for missing data

```
observations.isna().sum()
```

```
scientific_name    0
park_name          0
observations       0
dtype: int64
```

```
species.isna().sum()
```

```
category                0
scientific_name         0
common_names            0
conservation_status    5633
dtype: int64
```

We can see that the conservation status data of the has many missing values. As we inspect the data, we could know that this occured because these species does not face any threats. This is Structurally Missing Data (SMD). Hence, we can change the missing data in this column to "Safe"

```
species = species.fillna(value={'conservation_status':'Safe'})
```

Inspect the species dataset again:

```
species.head(10)
```

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

Cleaning and Tidying Data

We need to drop duplicated rows in the datasets

Examine if there is any duplicate values in the datasets

```
observations.duplicated()
```

```
0        False
1        False
2        False
3        False
4        False
         ...
23291    False
23292    False
23293    False
23294    False
23295    False
Length: 23296, dtype: bool
```

```
observations = observations.drop_duplicates()
```

```
species.duplicated()
```

```
0       False
1       False
2       False
3       False
4       False
        ...
5819    False
5820    False
5821    False
5822    False
5823    False
Length: 5824, dtype: bool
```

```
species = species.drop_duplicates()
```

Cleaning and Tidying Data

Next, we inspect the data again to change the data types is necessary

```
observations.dtypes
```

```
scientific_name    object
park_name          object
observations        int64
dtype: object
```

```
species.dtypes
```

```
category               object
scientific_name        object
common_names           object
conservation_status    object
dtype: object
```

All the columns have the correct data type, so there is no need to change the data type of any column.

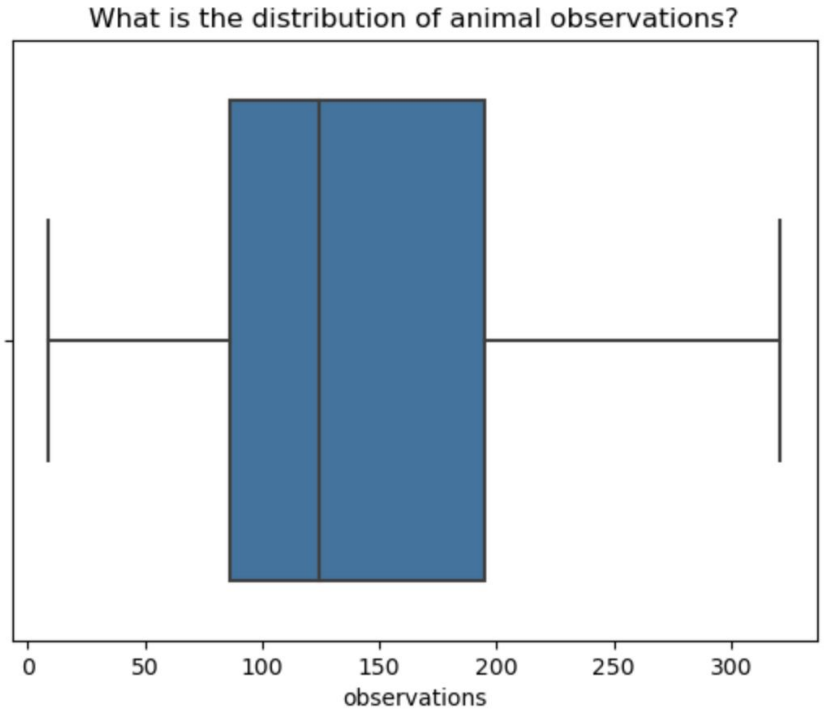
```
observations.nunique()
```

```
scientific_name    5541
park_name             4
observations        304
dtype: int64
```

4. Data Visualization and Analysis

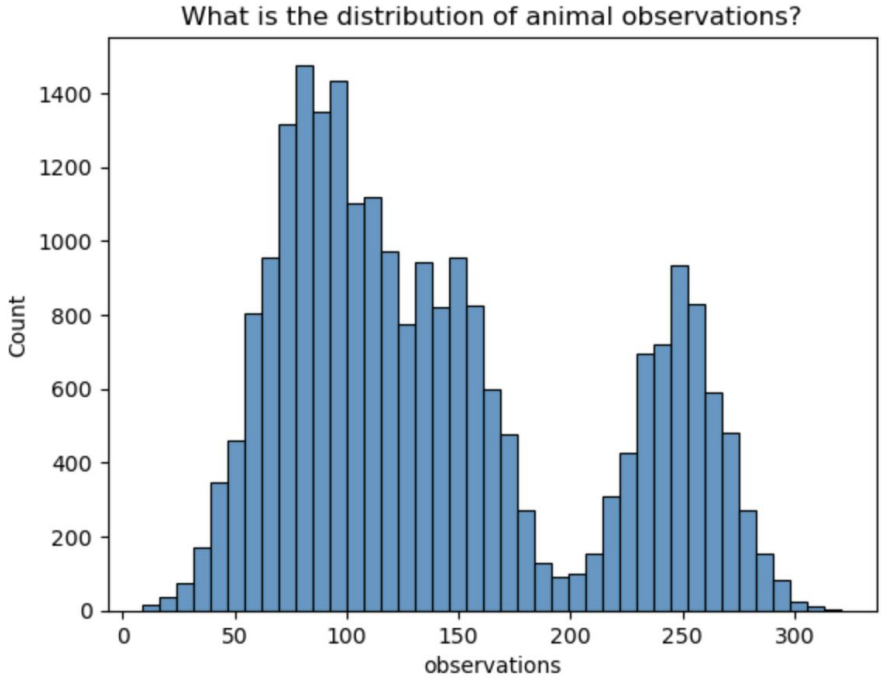
Univariate: What is the distribution of animal observations?

From the box plot, we can see that the IQR is around 90 observations, the median is 120, minimum and maximum values are 10 to 325. We can also see that there are not many outliers. For more details, we can plot a histogram.



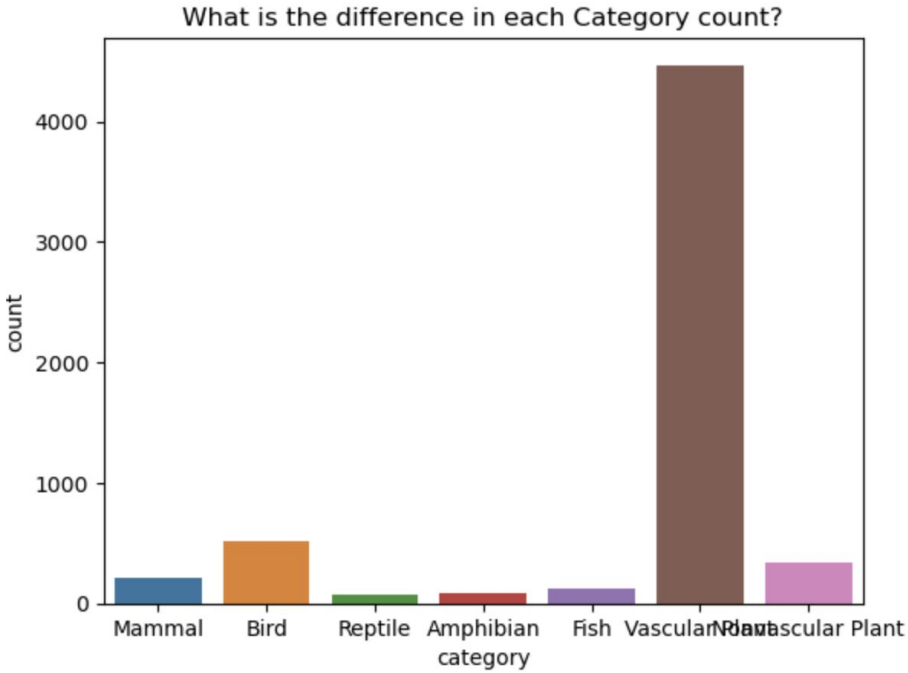
Univariate: What is the distribution of animal observations?

The histogram showed that the mode of the dataset is around 75 observations. The histogram also has two normal distributions, one around 75 and one around 250 observations.



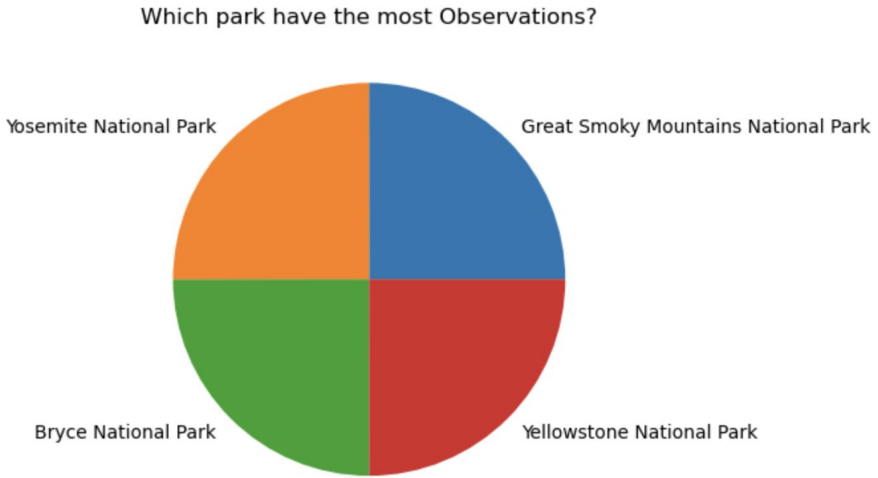
Univariate: What is the difference in each Category count?

With this bar chart, we can see that the category that appeared the most is Vascular Plan with over 4000 observations, following with bird with 600 observations.



Univariate: Which park have the most Observation?

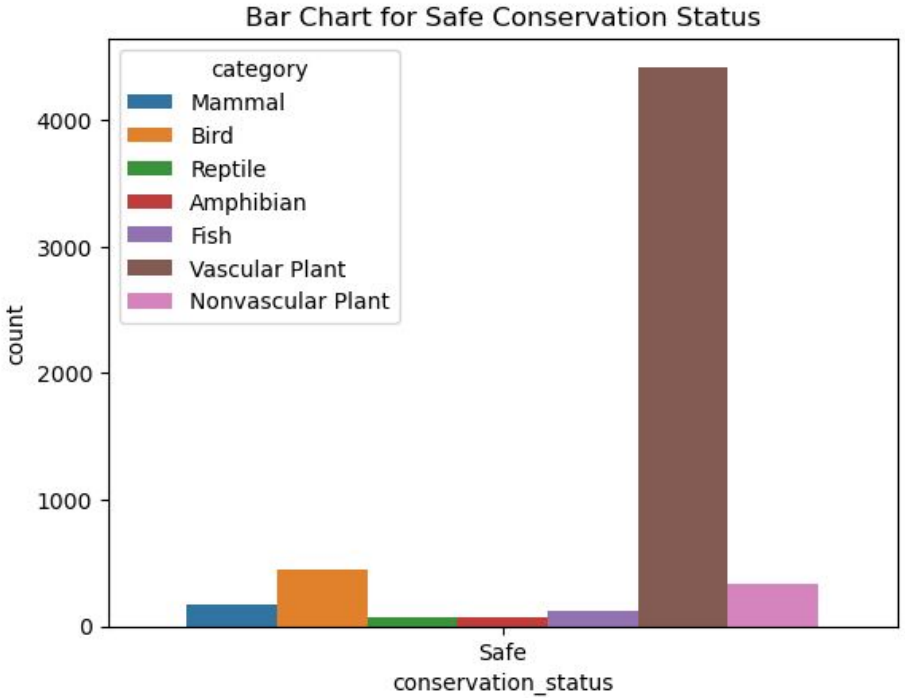
We can see that the category that appeared the most is Vascular Plan with over 4000 observations, following with bird with 600 observations.



Bivariate: Does certain species are more likely to be threatened or safe?

Safe Conservation Status

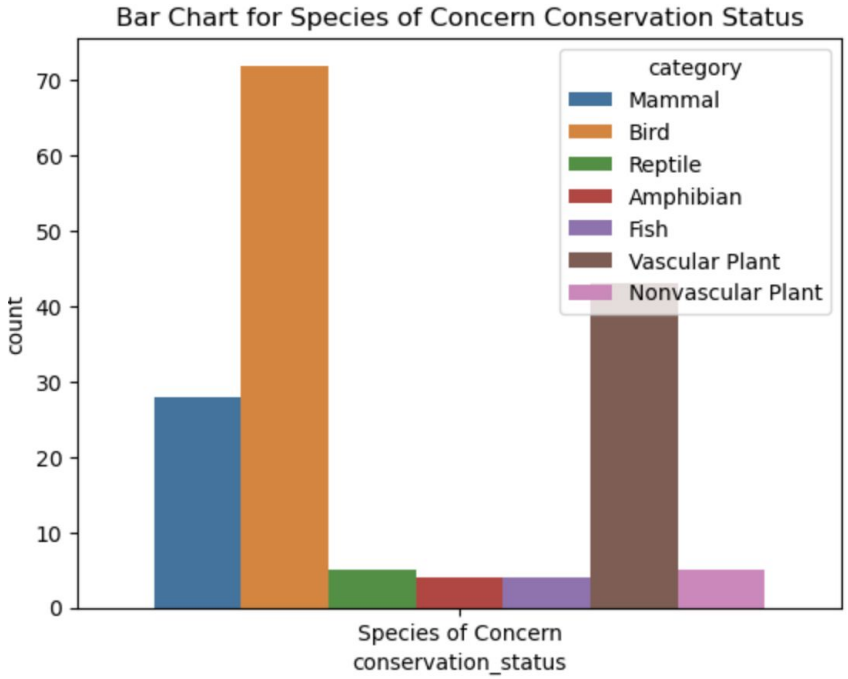
With this graph, we can see that the safest species is Vascular Plant, following with bird



Bivariate: Does certain species are more likely to be threatened or safe?

Species of Concern Conservation Status

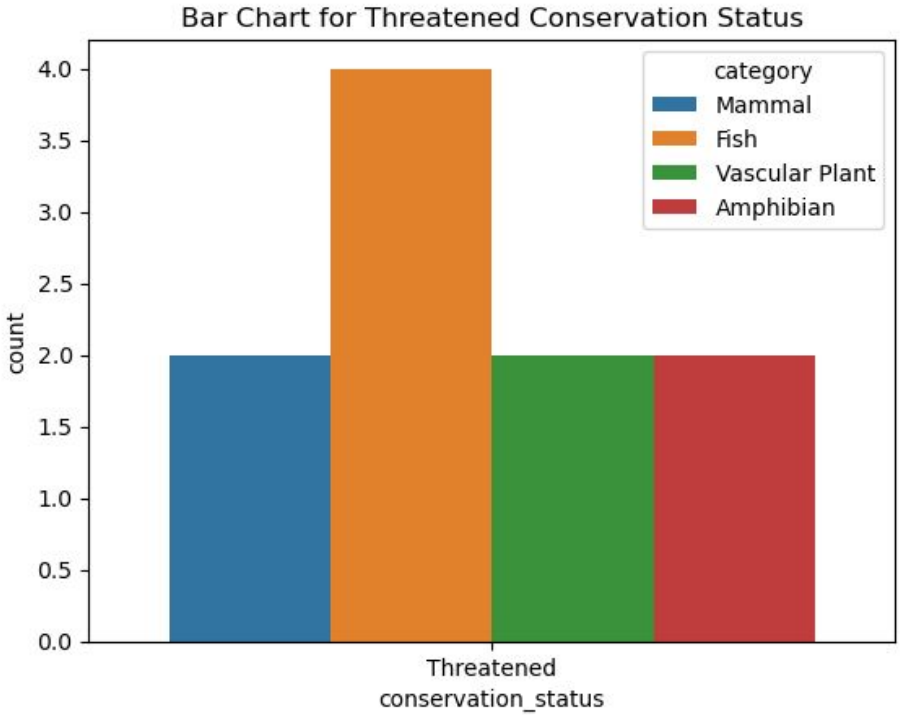
The species that dominated the conservation status Species of Concern is Bird, following with Vascular Plant. This alligns with the previous graph of the Safe status, as these 2 species are not threatened in a high level.



Bivariate: Does certain species are more likely to be threatened or safe?

Threatened Conservation Status

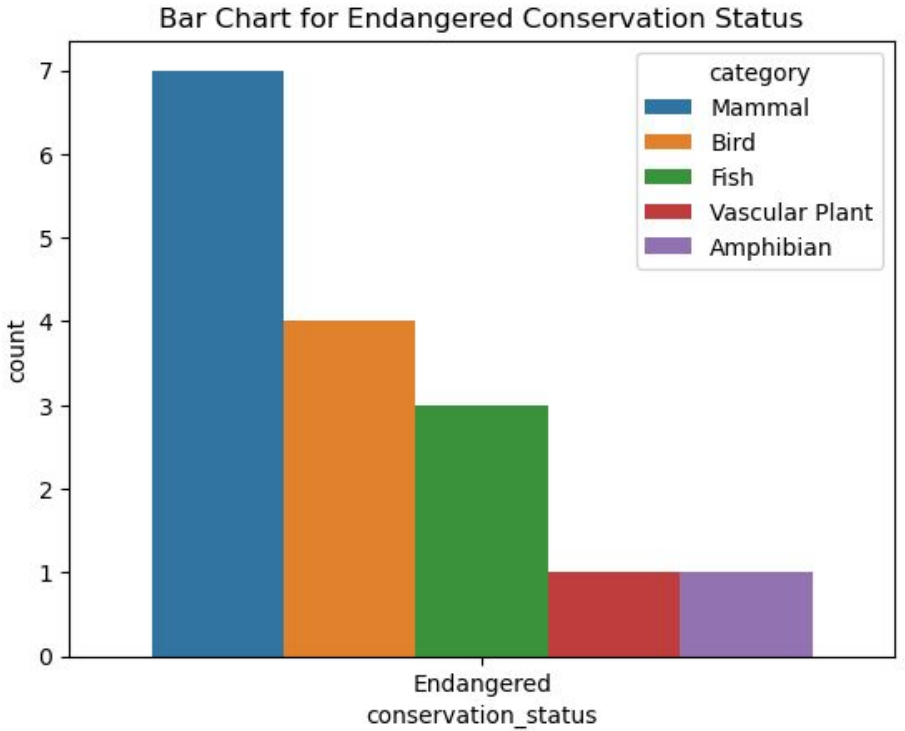
As we can see, some species does not appear in the Threatened status, which is the most negative status of the Conservation statuses. The most threatened specie is Fish, which may suggest that the quality of water in these area are threatening this specie.



Bivariate: Does certain species are more likely to be threatened or safe?

Endangered Conservation Status

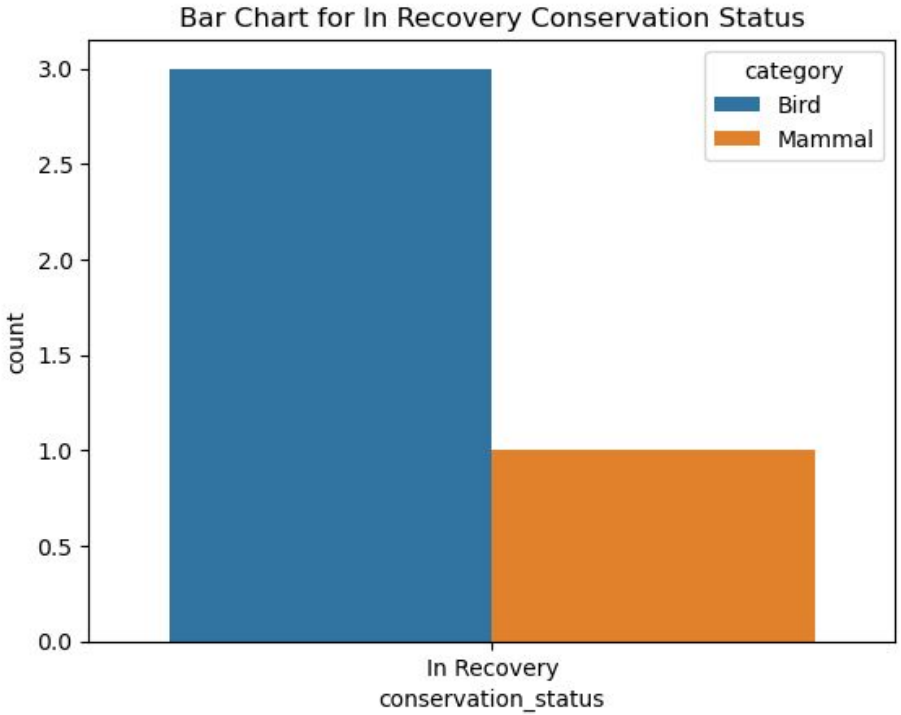
In this graph, we can only see 2 species, which are Bird and Mammal. This means that the other species are either Safe or is experiencing some degree of Danger and are not able to recover yet. Bird is the specie that is recovering the most.



Bivariate: Does certain species are more likely to be threatened or safe?

In Recovery Conservation Status

In this graph, we can only see 2 species, which are Bird and Mammal. This means that the other species are either Safe or is experiencing some degree of Danger and are not able to recover yet. Bird is the specie that is recover the most.

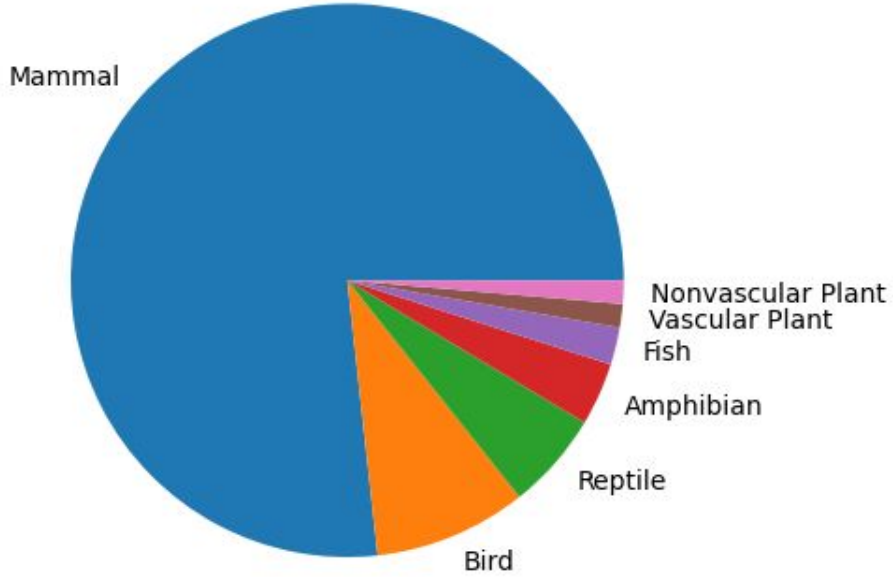


Bivariate: Does certain species are more likely to be threatened or safe?

Bias Discussion

However, after we graping these data, we can notice that there may be bias in the data, as the count of each species are not the same. With the pie chart, we can see that the Mammal category dominates the other category.

Hence, this could introduce bias to our data, as there are too many mammals compare to other categories



Bivariate: Count of 'Castor canadensis' in each Park

First, we create a new dataframe that contains only the scientific name of 'Castor canadensis':

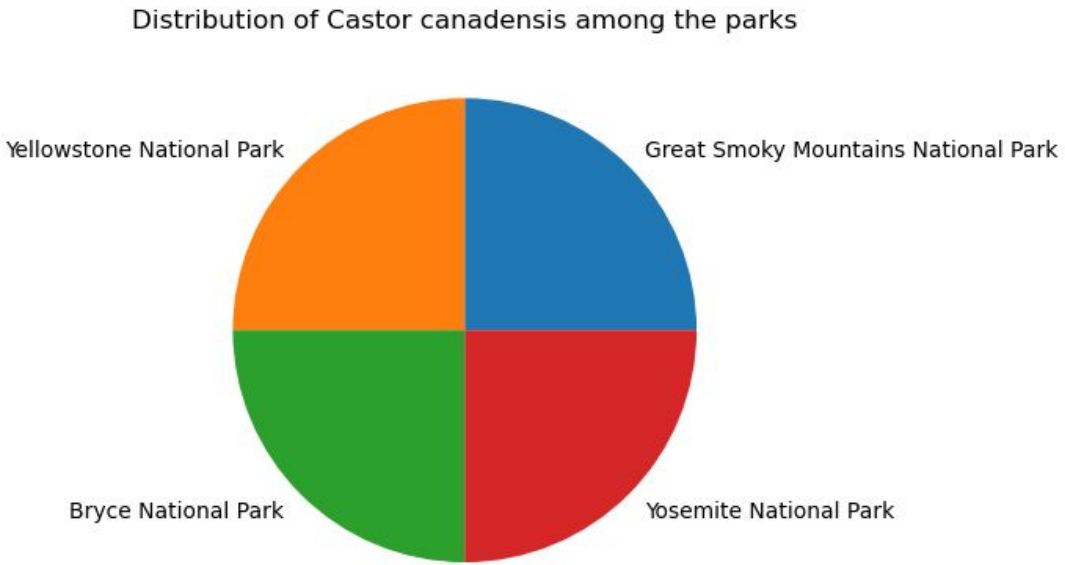
```
castor_df = observations[observations['scientific_name'] == 'Castor canadensis']

castor_df.head()
```

	scientific_name	park_name	observations
951	Castor canadensis	Great Smoky Mountains National Park	95
3792	Castor canadensis	Great Smoky Mountains National Park	62
6186	Castor canadensis	Yellowstone National Park	183
6303	Castor canadensis	Bryce National Park	70
9811	Castor canadensis	Yellowstone National Park	256

Bivariate: Count of 'Castor canadensis' in each Park

This is an interesting fact, as we can see that both the original observations data and the castor dataframe have the same equal park distribution. This means that the data of each scientific name have the equal park distribution.



5. Conclusions and Recommendations for Conservationists

Conclusions and Recommendations for Conservationists

Based on the analysis of the conservation status of each species, we have these important findings and actions for conservationist with each finding:

1. **The safest species are Vascular Plant and Bird**

   ⇒ Conservationists should allocate less resources in protecting these species, as they need less care than other species

2. **The most threatened species are Fish and Mammal**

   ⇒ Conservationists should give extra care to these species

3. **Mammal counts dominates the other category**

   ⇒ Conservationists should regulate Mammals strictly, dividing the category into sub-categories for better classifycation

4. **Bird is in recovery the most, and only 2 species are recovering**

   ⇒ As bird is one of the safest species, conservationists should also find more animals which are hurting to assists these species in recovering