Codecademy

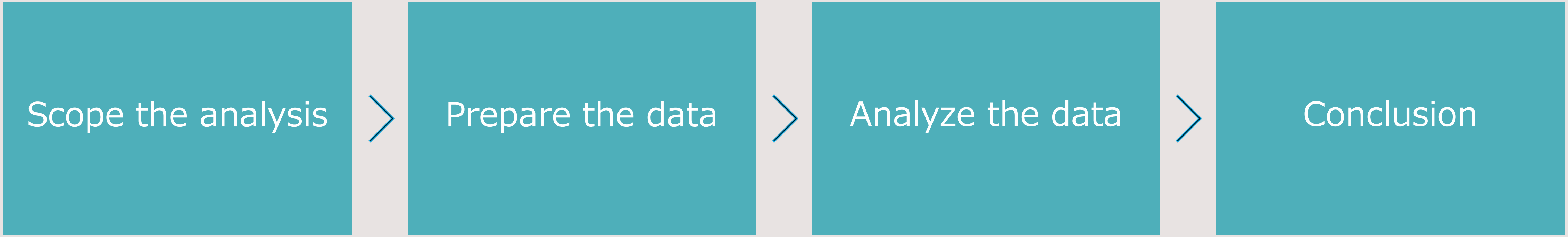
Biodiversity

Porfolio project

Duc Tran Anh

Project scoping – Overall

This project goal is to analyse the National Parks data surrounded by endangered species in different park. The project will be proceeded through those below steps:



Scope the analysis

Prepare the data

Analyze the data

Conclusion

Project scoping – Setting goals

Interpret data from the National Parks Service about endangered species in different parks, and address the below questions:

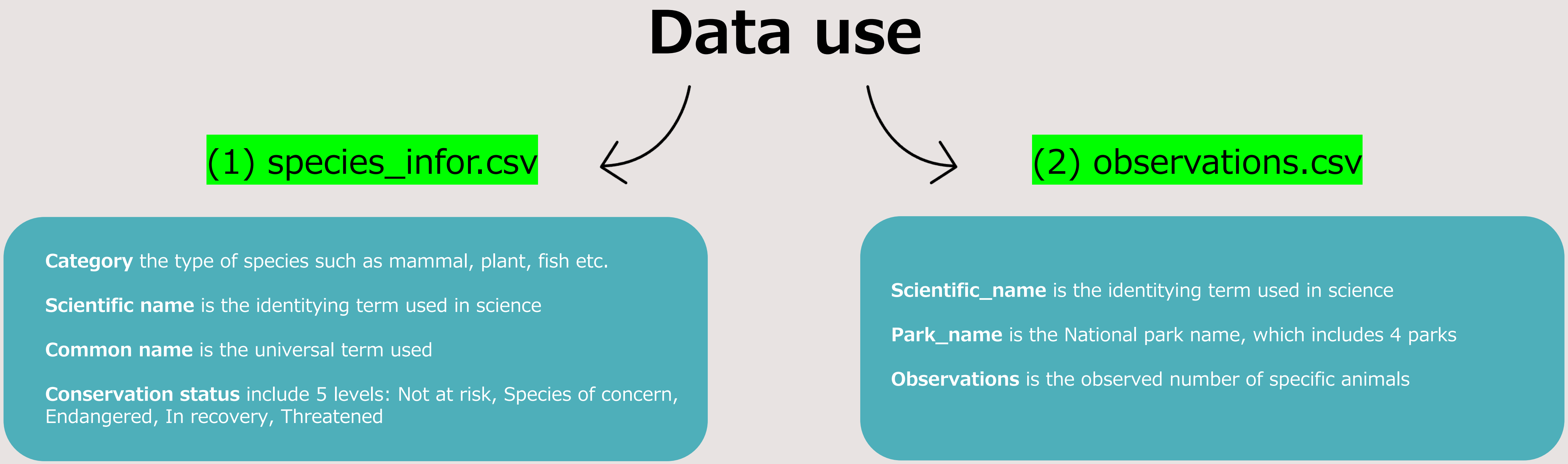
What is the distribution of species by conservation status?

What is the distribution of species among parks?

Are the differences between species and their conservation status significant?

Project scoping – Preparing the data

The data includes observations.csv and species_infor.csv, both are collected from codecademy.com The data is inspired by real data, but it is a fictional data for learning purpose.



Data use

(1) species_infor.csv

Category the type of species such as mammal, plant, fish etc.

Scientific name is the identitying term used in science

Common name is the universal term used

Conservation status include 5 levels: Not at risk, Species of concern, Endangered, In recovery, Threatened

(2) observations.csv

Scientific_name is the identitying term used in science

Park_name is the National park name, which includes 4 parks

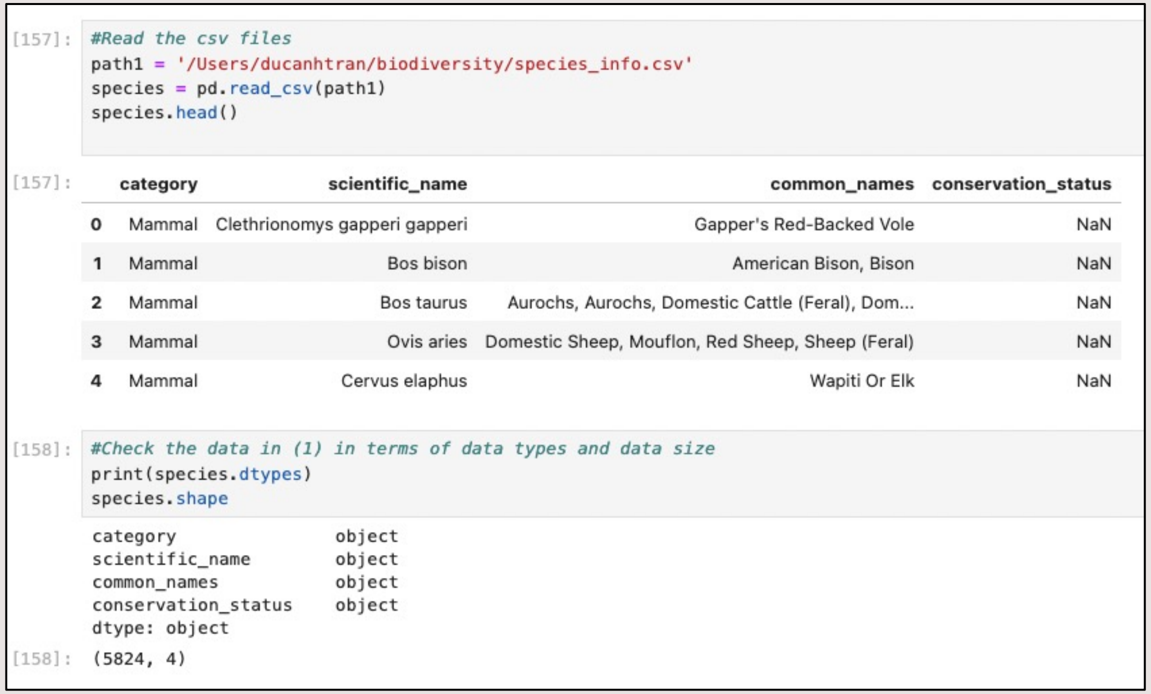
Observations is the observed number of specific animals

**Both data share the same variable named Scientific_name, which could be used to combine the two tables.*

Scanning the data frame and its shape to better understand the data. The species data has NaN value in the conservation status variable.

(1) species_infor.csv

(2) observations.csv

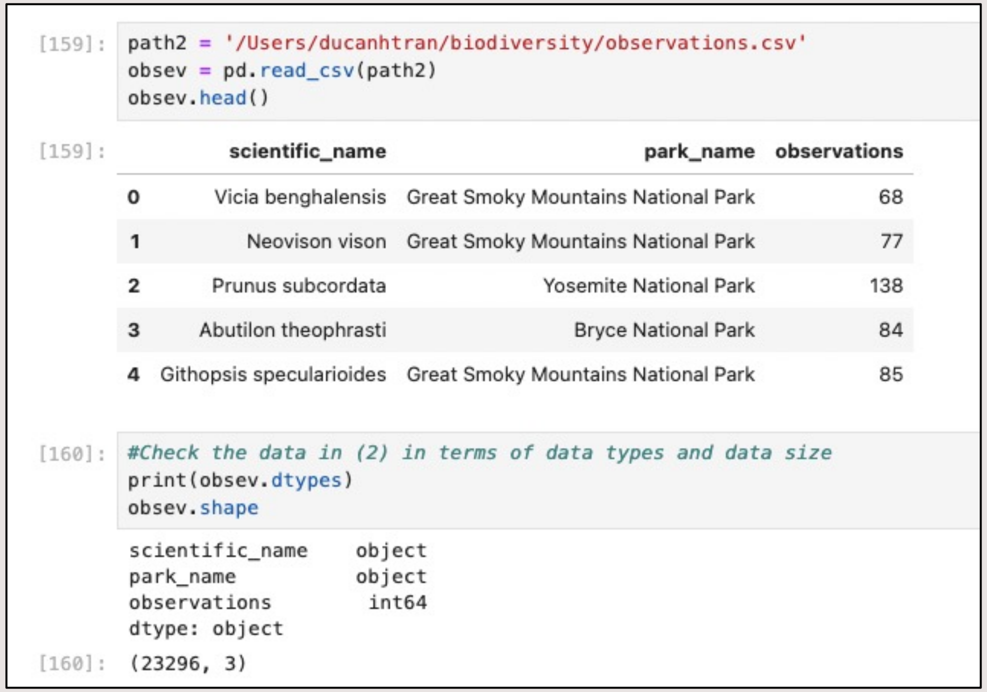


```
[157]: #Read the csv files
       path1 = '/Users/ducanhtran/biodiversity/species_info.csv'
       species = pd.read_csv(path1)
       species.head()
```

[157]:

	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

```
[158]: #Check the data in (1) in terms of data types and data size
       print(species.dtypes)
       species.shape

       category               object
       scientific_name        object
       common_names           object
       conservation_status    object
       dtype: object
[158]: (5824, 4)
```



```
[159]: path2 = '/Users/ducanhtran/biodiversity/observations.csv'
       obsev = pd.read_csv(path2)
       obsev.head()
```

[159]:

	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

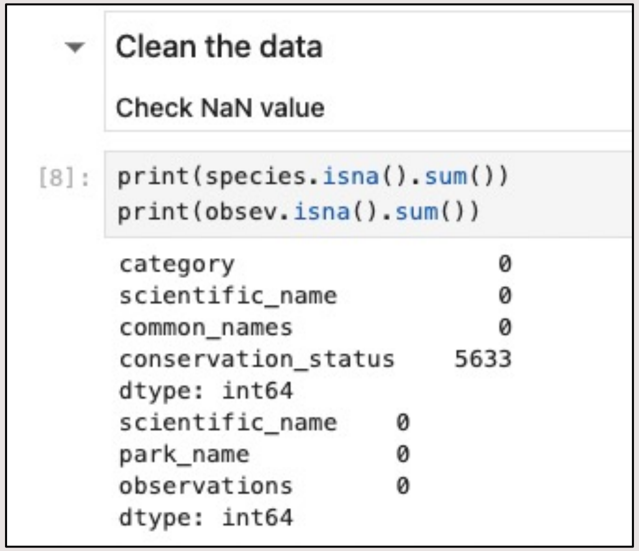
```
[160]: #Check the data in (2) in terms of data types and data size
       print(obsev.dtypes)
       obsev.shape

       scientific_name    object
       park_name          object
       observations        int64
       dtype: object
[160]: (23296, 3)
```

Project scoping – Preparing the data

96.7% of the conservation_status is **NaN** in species table.

1 *My assumption is that it might mean that these species are **not considered to be at risk** (i.e., not 'Species of Concern', 'Endangered', 'Threatened', or 'In Recovery').*

2 *Replacing those data with 'Not at risk' value and continue the analysis*

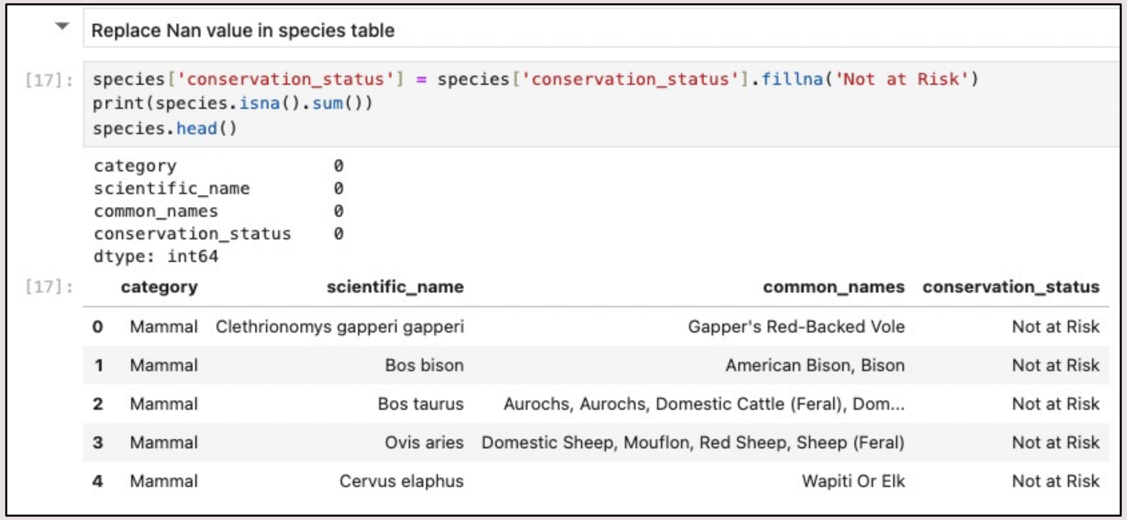


Clean the data

Check NaN value

```
[8]: print(species.isna().sum())
     print(obsev.isna().sum())
```

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



Replace Nan value in species table

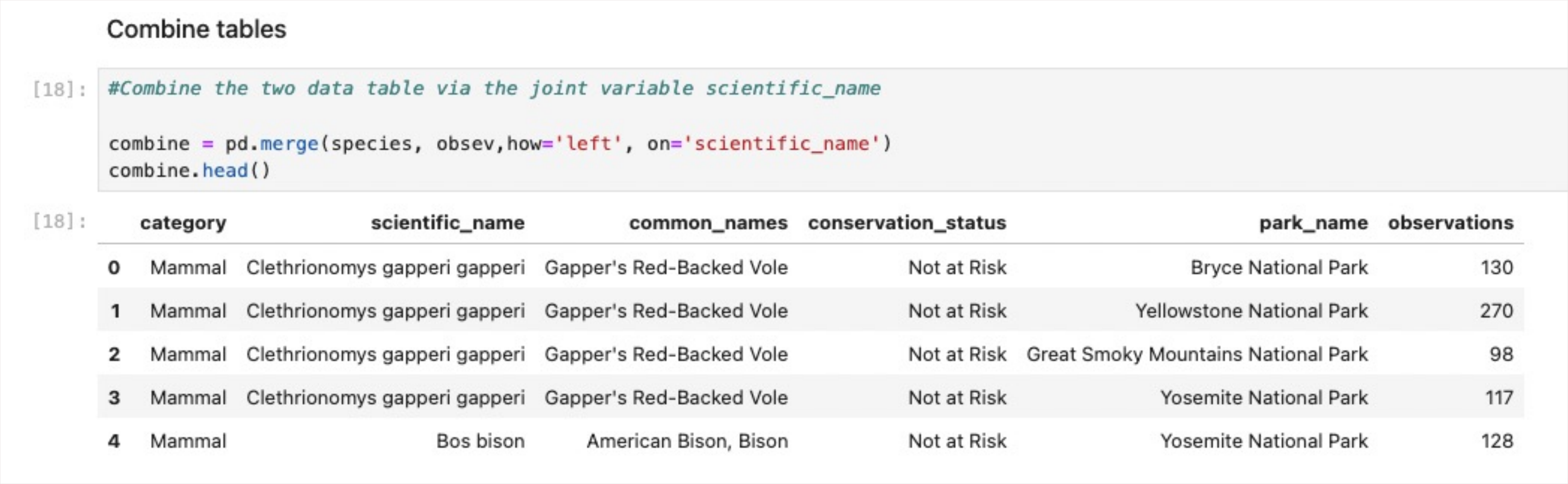
```
[17]: species['conservation_status'] = species['conservation_status'].fillna('Not at Risk')
      print(species.isna().sum())
      species.head()
```

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

[17]:

	category	scientific_name	common_names	conservation_status
0	Mammal	Clethrionomys gapperi gapperi	Gapper's Red-Backed Vole	Not at Risk
1	Mammal	Bos bison	American Bison, Bison	Not at Risk
2	Mammal	Bos taurus	Aurochs, Aurochs, Domestic Cattle (Feral), Dom...	Not at Risk
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	Not at Risk
4	Mammal	Cervus elaphus	Wapiti Or Elk	Not at Risk

Combine table and proceed analysis.



Combine tables

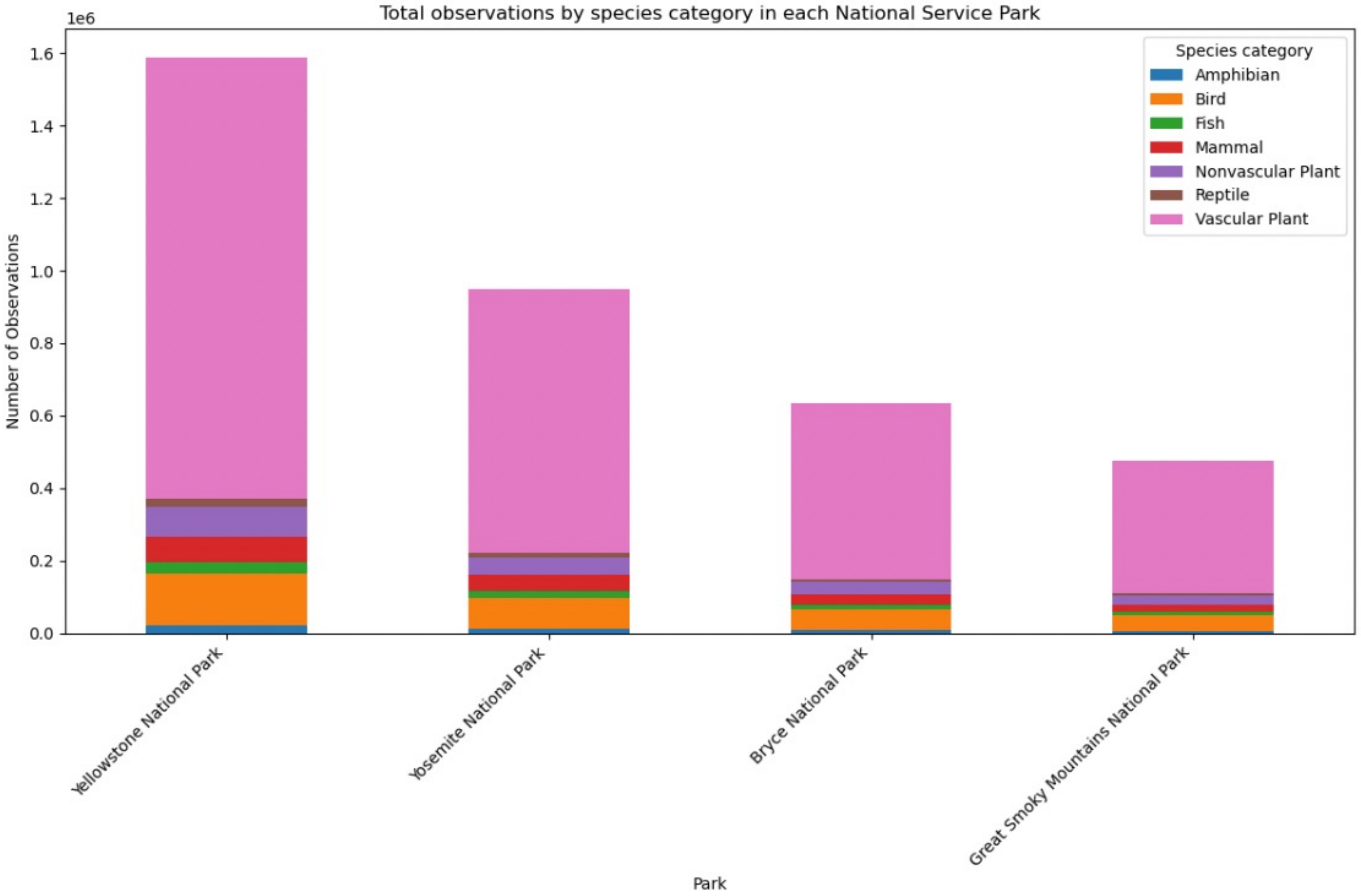
[18]:

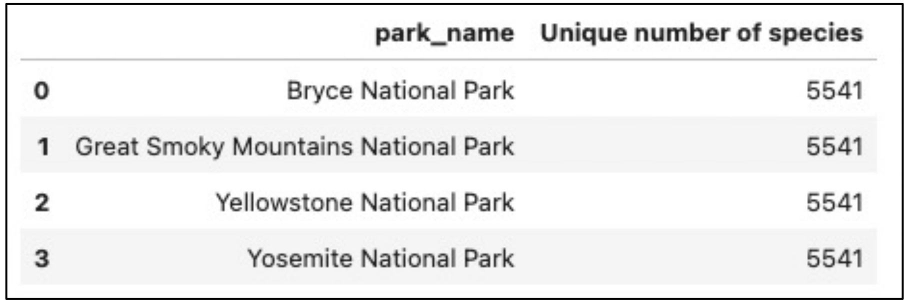
```python
#Combine the two data table via the joint variable scientific_name

combine = pd.merge(species, obsev,how='left', on='scientific_name')
combine.head()
```

[18]:

	category	scientific_name	common_names	conservation_status	park_name	observations
0	Mammal	Clethrionomys gapperi gapperi	Gapper's Red-Backed Vole	Not at Risk	Bryce National Park	130
1	Mammal	Clethrionomys gapperi gapperi	Gapper's Red-Backed Vole	Not at Risk	Yellowstone National Park	270
2	Mammal	Clethrionomys gapperi gapperi	Gapper's Red-Backed Vole	Not at Risk	Great Smoky Mountains National Park	98
3	Mammal	Clethrionomys gapperi gapperi	Gapper's Red-Backed Vole	Not at Risk	Yosemite National Park	117
4	Mammal	Bos bison	American Bison, Bison	Not at Risk	Yosemite National Park	128

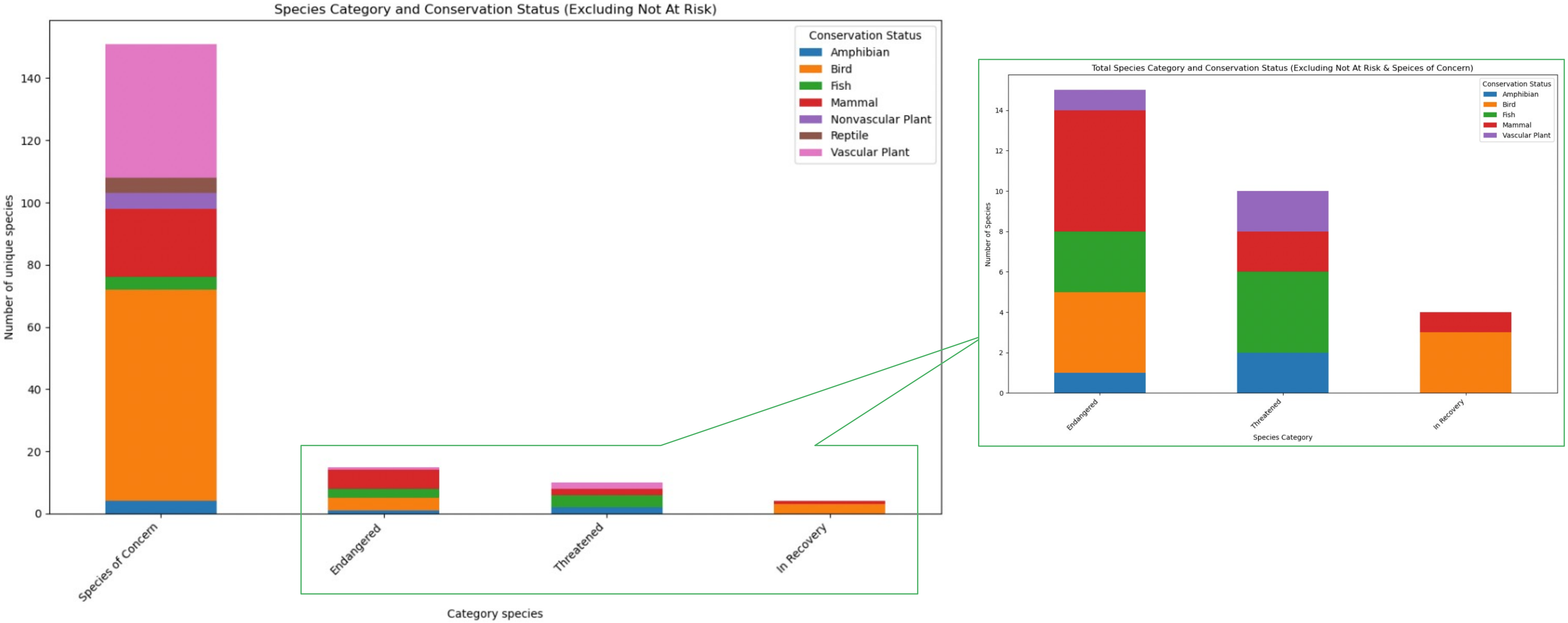
The diversity among parks are similar in terms of species . And **Yellowstone National Park** has the greatest number of species across categories and conservation status.



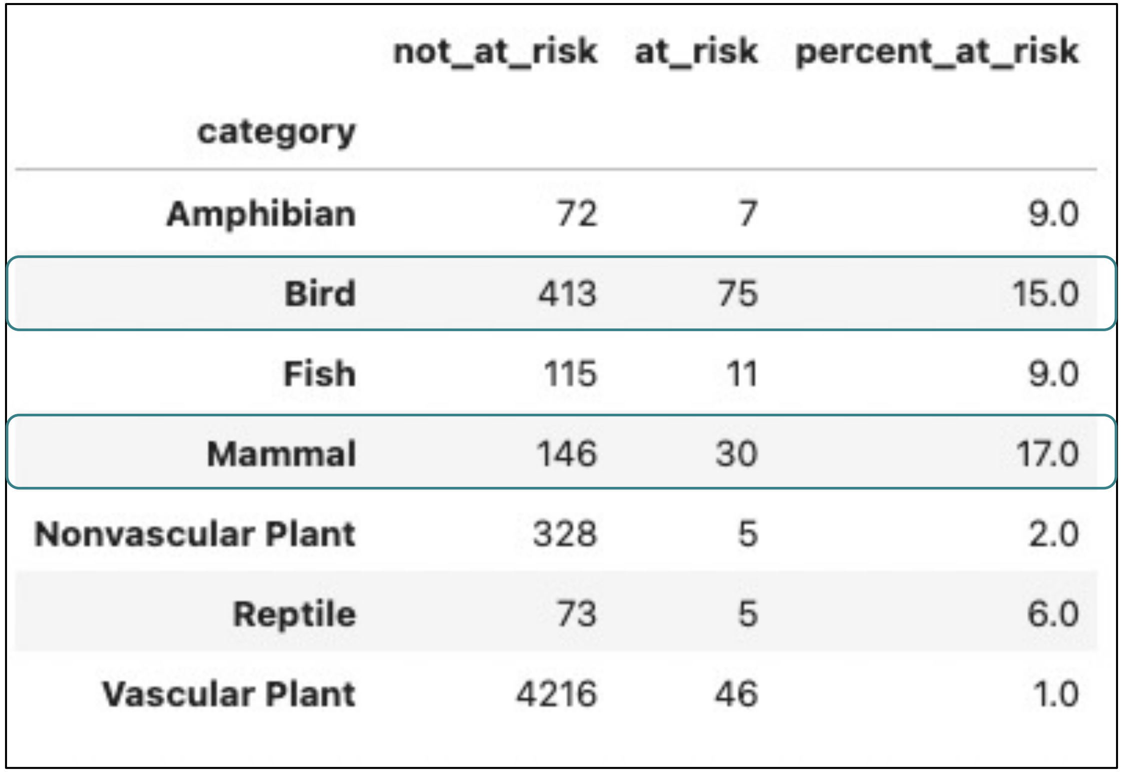


	park_name	Unique number of species
0	Bryce National Park	5541
1	Great Smoky Mountains National Park	5541
2	Yellowstone National Park	5541
3	Yosemite National Park	5541

Bird, Vascular plant, and Mammal are top 3 having the most species of concern, in which, mammal is both threatened and endangered.



There are 5,363 species that are not at risk, while other 179 species need strict conservation to protect. In which, **birds and mammals** are the 2 groups with the highest at-risk proportions.



	not_at_risk	at_risk	percent_at_risk
category			
Amphibian	72	7	9.0
Bird	413	75	15.0
Fish	115	11	9.0
Mammal	146	30	17.0
Nonvascular Plant	328	5	2.0
Reptile	73	5	6.0
Vascular Plant	4216	46	1.0

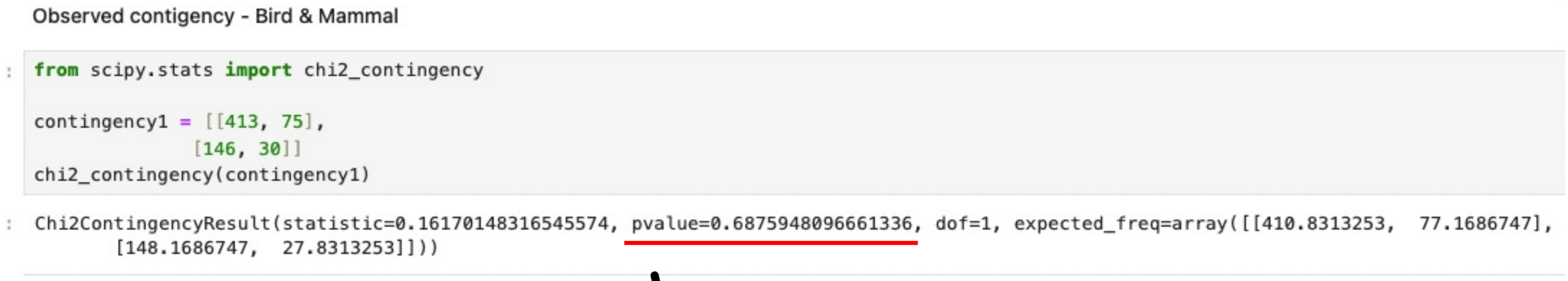
Whether or not there is association between species and their risk status?

Examinating the association between the risk status and the different species. Since both Bird and Mammal have the greatest risk percentage in their population, let start **the chi square test** with the abovementioned species.

There are 2 hypothesises:

H0: There is no association between risk status and the species, also understand that the distribution of the risk status by species happen by chance.

H1: There is associated relationship between the risk status and the species, indicating the underlying relationship between the 2 variables.



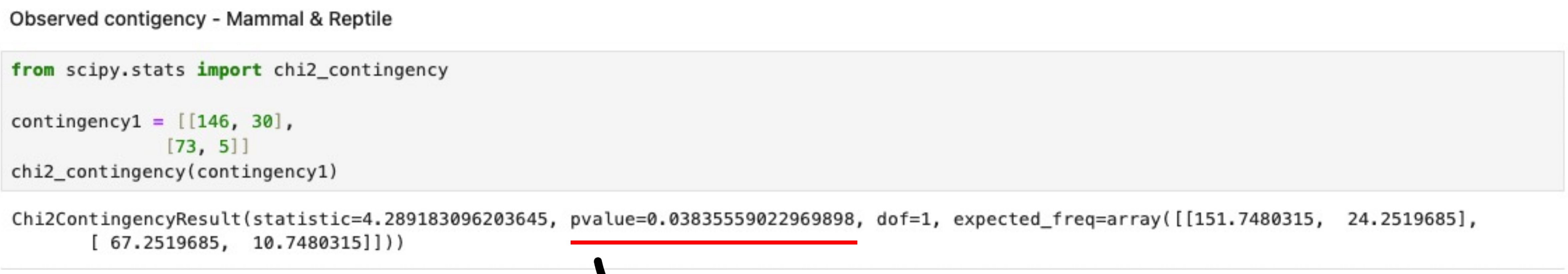
Observed contigency - Bird & Mammal

```
from scipy.stats import chi2_contingency

contingency1 = [[413, 75],
                [146, 30]]
chi2_contingency(contingency1)
```

```
Chi2ContingencyResult(statistic=0.16170148316545574, pvalue=0.6875948096661336, dof=1, expected_freq=array([[410.8313253,  77.1686747],
       [148.1686747,  27.8313253]]))
```

Since the p-value (0.687) is more than the significance level (0.05), we agree to the null hypothesis.

Examining the association between the risk status and the different species of the pair of mammal and reptile.



Observed contigency - Mammal & Reptile

```
from scipy.stats import chi2_contingency

contingency1 = [[146, 30],
               [73, 5]]
chi2_contingency(contingency1)
```

```
Chi2ContingencyResult(statistic=4.289183096203645, pvalue=0.03835559022969898, dof=1, expected_freq=array([[151.7480315,  24.2519685],
       [ 67.2519685,  10.7480315]]))
```

Since the p-value (0.0383) is less than the significance level (0.05), we reject the null hypothesis.

This means that there is a significant association between being a mammal or a reptile and their risk status (at risk or not at risk).

Addressing the project questions:

What is the distribution of species by conservation status?

There are 5363 species that are not at risk, while other 179 species need strict conservation to protect. Birds and mammals being the 2 groups with the highest at-risk proportions.

What is the distribution of species among parks?

The diversity among parks are similar in terms of species, and, Yellow National Park has the most number of observations.

Are the differences between species and their conservation status significant?

There is a significant association between being a mammal or a reptile and their risk status (at risk or not at risk).

Further research & verifications:

- The assumption regarding the 5,363 species considered not-at-risk requires verification through further research.

- Investigate why mammals are more likely to be at risk compared to reptiles. This could involve looking into factors such as habitat loss, hunting pressures, environmental changes, etc.

- The data is the observation in selected parks in the last 7 days, it would be curious to see how the obsservtions change overtime.

Learning from Biodiversity project

Project Architecture

> I have learned to plan a data analysis project in a structured and well-organized manner, clearly defining the scope of work and focusing on key aspects. For example, setting the project goal with investigative questions is crucial to ensure that my analysis remains relevant and central.

Understanding the Role of Each Scope in the Holistic Data Project:

> Understand the importance of clearly defining the project scope and goals to maintain focus and relevance throughout the analysis.

Loading and Preparing Data:

> Recalled how to statistically summarize data, handle missing values, and make assumptions based on the data. This step is vital for ensuring data quality and integrity.

Data Analysis:

> Practiced descriptive analysis to understand the overall information.

> Performed exploratory analysis to gain deeper insights, using techniques such as the chi-square test to examine associations between two categorical variables.

Conclusion & Further Research:

> Formulated conclusions to address the questions set at the beginning of the project.

> Acknowledging limitations in the information and understanding, I identified areas for further research to broaden my understanding.

This structured approach ensures a comprehensive and effective data analysis project, from initial scoping to drawing conclusions and identifying future research directions.

Technical Skills

> I have applied various techniques to modify tables, perform aggregations, and carry out calculations to summarize statistics effectively.

Furrther action

> Prepare for final porfolio project