

Biodiversity in National Parks

Биоразнообразие в Национальных парках

О проекте

В данном проекте я буду исследовать и анализировать информацию, предоставленную «National Parks Service» и связанную с различными национальными парками.

Цель проекта заключается в том, чтобы определить, существуют ли зависимость между охранным статусом обитателей национальных парков и их видом; изучить разнообразие парков и количество наблюдений в каждом из них. После анализа таблиц, необходимые данные будут для удобства вынесены в отдельные таблицы и визуализированы с помощью подходящих графиков для более лёгкого восприятия, а далее на их основе будут сделаны выводы.

Вся информация разделена на 2 таблицы. В первой таблице содержутся данные касательно наблюдений в нескольких национальных парков. Информация из второй таблицы связаны непосредственно с их обитателями.

Анализ будет проводится в Jupiter Notebook. Это веб-приложение, позволяющее очищать и преобразовывать данные, производить математические вычисления, использовать статистику для анализа данных, а также визуализировать полученные результаты.

Все операции с данными и таблицами (исследование, очистка, преобразование и визуализация) выполняется с использованием языка Python и его наиболее популярных библиотек: Pandas, NumPy, Seaborn, Matplotlib и SciPy.

Обзор исходных данных

Все данные касательно национальных парков, их обитателей и проводимых в них наблюдений распределены по 2 файлам:

1 observations.csv

2 species_info.csv

При помощи метода `read_csv()` было создано по одному DataFrame (таблице) для каждого из файлов:

1 observations

2 species

Таблица `observations` состоит из 3 колонок: данные (имя и фамилия) научного сотрудника, название национального парка и количество наблюдений, проведённых определённым сотрудником в определённом национальном парке.

Таблица `species` состоит из 4 колонок: принадлежность к группе живых существ (например, млекопитающие или птицы), данные (имя и фамилия) научного сотрудника, название живого существа и охранный статус.

Далее будут приведены примеры нескольких записей в каждой из таблиц.

Примеры записей в таблицах

Таблица observations

	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

Таблица species

	category	scientific_name	\
0	Mammal	Clethrionomys gapperi	gapperi
1	Mammal		Bos bison
2	Mammal		Bos taurus
3	Mammal		Ovis aries
4	Mammal		Cervus elaphus

	common_names	conservation_status
0	Gapper's Red-Backed Vole	NaN
1	American Bison, Bison	NaN
2	Aurochs, Aurochs, Domestic Cattle (Feral), Dom...	NaN
3	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	NaN
4	Wapiti Or Elk	NaN

Очистка данных

Перед тем, как анализировать и визуализировать данные, необходимо их очистить. Проверим каждую из таблиц на пропущенные и нулевые значения с помощью методов `describe()` и `info()`.

Ниже можно увидеть, что в таблице «`observations`» отсутствуют пропущенные значения в любой из трёх колонок.

```
Ввод [524]: print(observations.info())  
  
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 23296 entries, 0 to 23295  
Data columns (total 3 columns):  
 #   Column      Non-Null Count  Dtype     
---  --    
 0   scientific_name    23296 non-null   object    
 1   park_name        23296 non-null   object    
 2   observations     23296 non-null   int64
```

Проверим единственный числовой столбец таблицы, отражающий количество наблюдений в Национальном парке, на наличие нулевых значений. Так как минимальное значение в столбце равно 9, то можно сделать вывод об отсутствии нулевых значений в столбце.

```
          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
```

Пропущенные и нулевые значения в таблице «`observations`» отсутствуют, поэтому данные таблицы готовы к анализу и визуализации.

Очистка данных

В таблице «species» нет числовых столбцов, поэтому данную таблицу мы проверяем только на наличие пропущенных значений.

```
Ввод [523]: print(species.info())

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5824 entries, 0 to 5823
Data columns (total 4 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   category        5824 non-null    object  
 1   scientific_name 5824 non-null    object  
 2   common_names    5824 non-null    object  
 3   conservation status 191 non-null   object
```

Из рисунка видно, что в столбце `conservation_status` присутствует много значений `NaN`, т.е. пропущенных. Поэтому, для дальнейшего анализа целесообразно создать отдельную таблицу, содержащую те строки, где `conservation_status` заполнен. Для более лучшей читаемости создадим такую таблицу только с двумя столбцами: `category` и `conservation_status`.

```
Ввод [541]: category_conservation_status_not_null = \
category_conservation_status[category_conservation_status.conservations_status.isnull() != True]
print(category_conservation_status_not_null)
```

```
category conservation_status
7       Mammal Species of Concern
8       Mammal      Endangered
9       Mammal      Endangered
29      Mammal Species of Concern
30      Mammal Species of Concern
...
...
5302  Vascular Plant Species of Concern
5399  Vascular Plant Species of Concern
5426  Vascular Plant Species of Concern
5436  Vascular Plant Species of Concern
5676  Vascular Plant Species of Concern
```

[191 rows x 2 columns]

Очистка данных

В дальнейшем, для более лёгкого доступа к нужными данным, они будут перенесены в отдельные таблицы. Для этой цели нужно определить, какие из столбцов в каждой таблице интересны для рассмотрения. Один из способов, который поможет отсеять некоторые столбцы – это метод `nunique()`. С его помощью мы узнаем, количество уникальных значений в столбцах таблиц.

```
Ввод [788]: print(species.nunique())
category          7
scientific_name   5541
common_names      5504
conservation_status 4
dtype: int64

Ввод [789]: print(observations.nunique())
scientific_name   5541
park_name         4
observations      304
dtype: int64
```

На основе вывода результатов можно сказать, что столбец `scientific_name` в обеих таблицах и столбец `common_names` в таблице «`species`» не несут для нас ценности. Они содержит более 5000 уникальных значений, что бесполезно для группировки или визуализации данных. Столбцы `category`, `conservation_status` в таблице «`species`» и `park_name`, `observations` в «`observations`» мы оставляем, так как они содержат значительно меньшее число уникальных значений и могут быть визуализированы.

Анализ данных

Для начала выполним операцию `unique()` чтобы узнать, какие уникальные охранные статусы, группы живых существ и Национальные парки присутствуют в наших таблицах. Получаем:

1 Охранные статусы:

- endangered;
- in recovery;
- species of concern;
- threatened.

2 Группы живых существ :

- amphibian;
- bird;
- fish;
- mammal;
- nonvascular plant;
- reptile;
- vascular plant.

3 Национальные парки:

- Great Smoky Mountains National Park;
- Yosemite National Park;
- Bryce National Park;
- Yellowstone National Park.

Анализ данных

Одна из целей проекта заключается в том, чтобы сделать вывод о наличии (или отсутствии) зависимости между определённой группой живых существ и охранным статусом. Для этого мы будем использовать статистику и проверку статистических гипотез на Python.

В Python существует множество различных тестов для проверки статистических гипотез. Каждый из тестов используется в конкретной ситуации и при заданных условиях. Выбор того или иного теста зависит от конкретной ситуации и условий, поставленных задач, рассматриваемых переменных, области исследования и других факторов.

Для получения ответа на волнующий нас вопрос о зависимости между определённой группой живых существ и охранным статусом мы сформулируем 2 гипотезы: нулевую (H_0) и альтернативную (H_1).

H_0 : зависимость между определённой группой живых существ и охранным статусом отсутствует.

H_1 : зависимость между определённой группой живых существ и охранным статусом существует.

Указав нулевую и альтернативную гипотезы, мы установим уровень значимости (альфа или α). Это порог, за пределами которого мы понимаем, что наблюдаемые данные больше не удовлетворяют нулевой гипотезе. Часто выбирается уровень в 0.05 (5%), на нем мы и остановимся.

В результате проведения теста для проверки гипотез мы получим несколько значений, одно из которых это *r-value* - фактически это вероятность ошибки при отклонении нулевой гипотезы. Его мы и будем в дальнейшем сравнивать с уровнем значимости. Если полученное *r-value* меньше уровня значимости, то мы отвергаем нулевую гипотезу и принимаем альтернативную гипотезу (и наоборот, если *r-value* больше заданного уровня значимости).

Анализ данных

Как было сказано ранее, один из ключевых моментов при выборе теста для проверки гипотез – это вид рассматриваемых переменных (качественные или количественные).

В нашем тесте мы используем переменные category и conservation_status. Обе переменные являются качественными. Поэтому, будем использовать тест хи-квадрата (функция chi2_contingency() из библиотеки SciPy).

Для начала создадим таблицу сопряжённости (кросс-таблицу) для наших переменных с помощью функции pd.crosstab(). Выглядит она следующим образом:

category	conservation_status	Endangered	In Recovery	Species of Concern	Threatened
Amphibian		1	0	4	2
Bird		4	3	72	0
Fish		3	0	4	4
Mammal		7	1	28	2
Nonvascular Plant		0	0	5	0
Reptile		0	0	5	0
Vascular Plant		1	0	43	2

Анализ данных

Созданную сопряжённую таблицу мы задаём в качестве входного параметра функции `chi2_contingency()`.

Наша функция и условия вывода результатов будет выглядеть следующим образом:

```
Ввод [477]: chi2, pval, dof, expected = chi2_contingency(Xtab)
if pval < 0.05:
    print('P-value we get from Chi-Square Test is {pval}. \nIt is less than significance threshold of 0.05.\nSo we can say with certainty that there is strong association between two categorical variables:\ncategory and conservation status.'.format(pval = pval))
else:
    print('P-value we get from Chi-Square Test is {pval}. \nIt is more than significance threshold of 0.05.\nSo we can say with certainty that there is no association between two categorical variables: \ncategory and conservation status.'.format(pval = pval))
```

Результат

```
P-value we get from Chi-Square Test is 1.8909788349761653e-05.
It is less than significance threshold of 0.05. So we can say with certainty that there is strong association between
two categorical variables: category and conservation status.
```

Следовательно, мы можем отклонить нулевую гипотезу об отсутствие зависимости между охранным статусом и группой живых существ .

Визуализация

Чтобы приступить непосредственно к процессу визуализации данных при помощи Python, будет не лишним вынести необходимые данные в отдельные таблицы. Сперва определим, какие данные и зависимости между ними мы хотим визуализировать.

В таблице «observations» было бы интересно узнать, в каком из парков проводилось больше всего наблюдений.

Для этого создадим сводную таблицу на основе столбца park_name с помощью метода groupby(). Получаем следующую таблицу, но основе которой и будет строить графики.

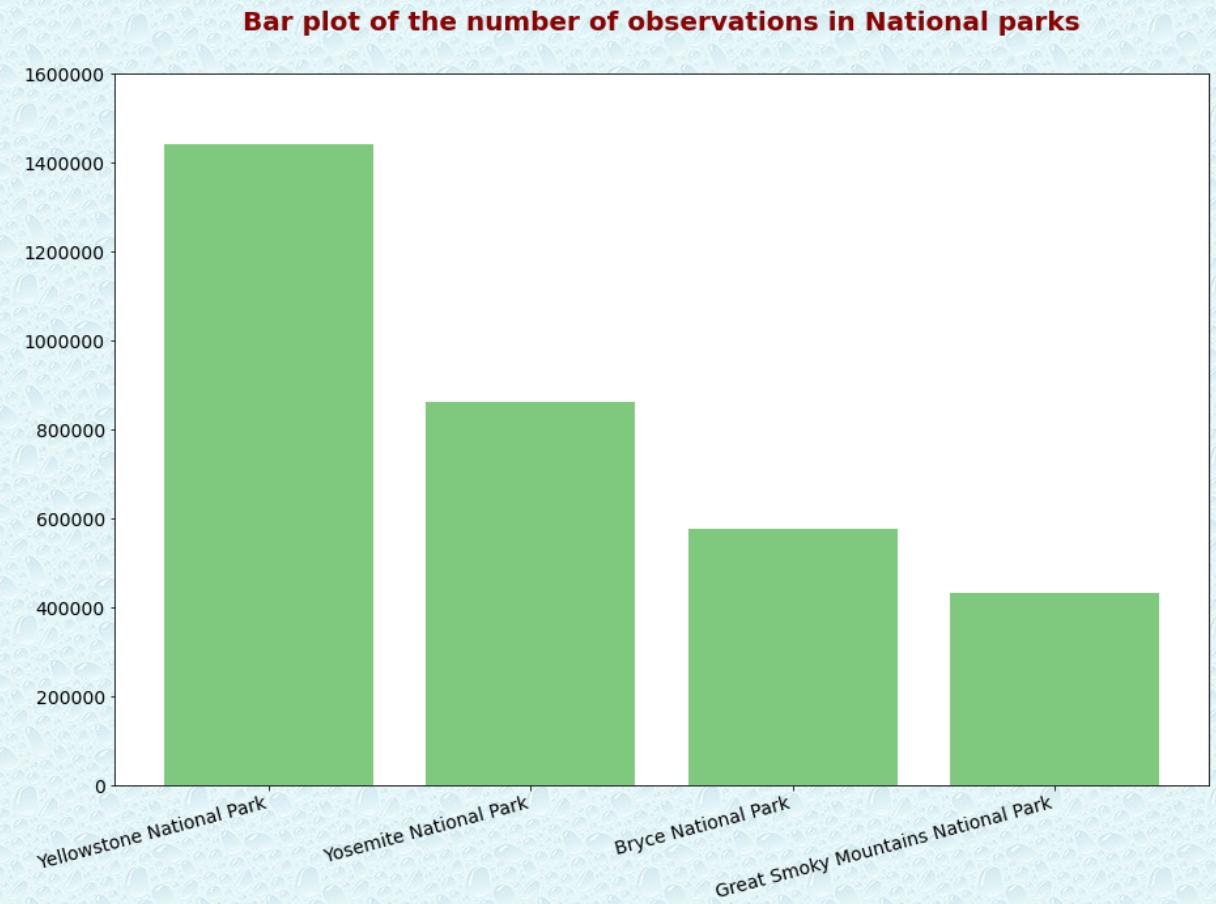
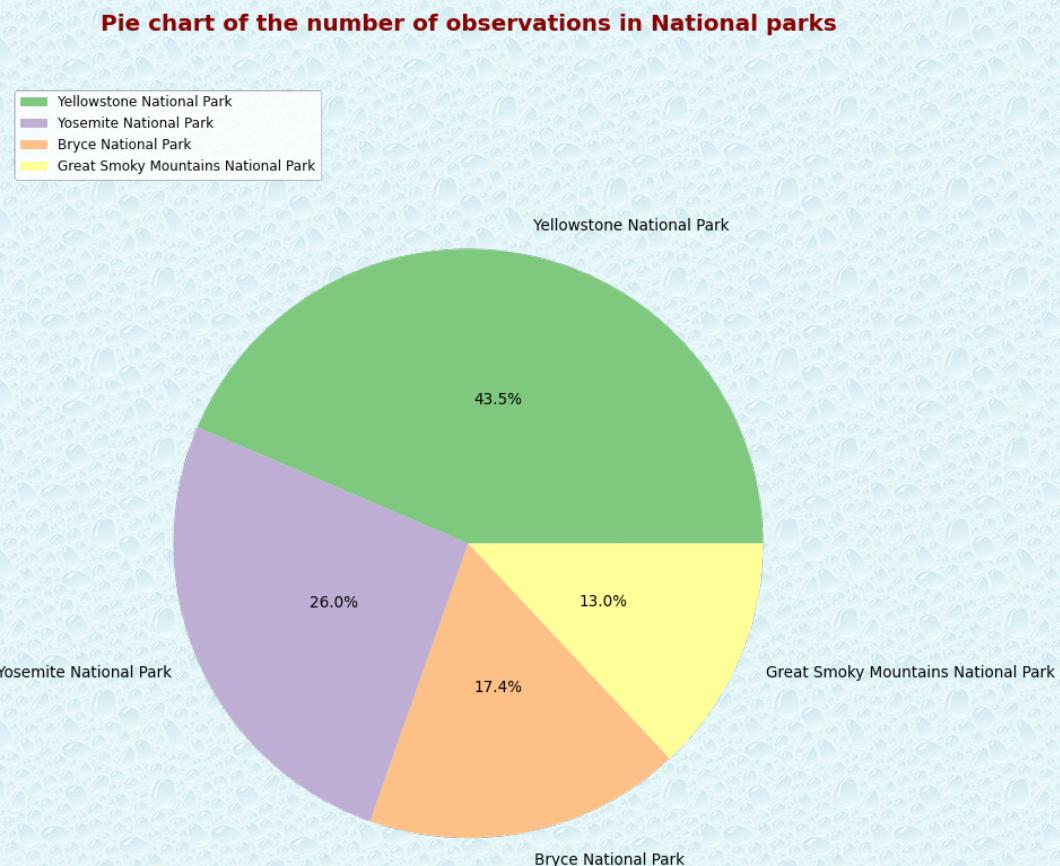
Можем заметить, что большинство наблюдений проводилось в Yellowstone National Park и Yosemite National Park. Количество наблюдений в данных парках составляет 43% и 26% от общего количества наблюдений соответственно.

	park_name	Number of observations
2	Yellowstone National Park	1443562
3	Yosemite National Park	863332
0	Bryce National Park	576025
1	Great Smoky Mountains National Park	431820

Визуализация

Изобразим количество наблюдений в каждом из Национальных парков с помощью 2 видов графиков:

- круговая диаграмма;
- столбчатая диаграмма.



Визуализация

В таблице «species» рассмотрению подлежат столбцы, отображающие группу живых существ и охранный статус. Для начала, создадим сводные таблицы. Одну на основе столбца category, вторую – на основе conservation_status.

Сводная таблица по столбцу «conservation_status»

	conservation_status	count
0	Endangered	16
1	In Recovery	4
2	Species of Concern	161
3	Threatened	10

Сводная таблица по столбцу «category»

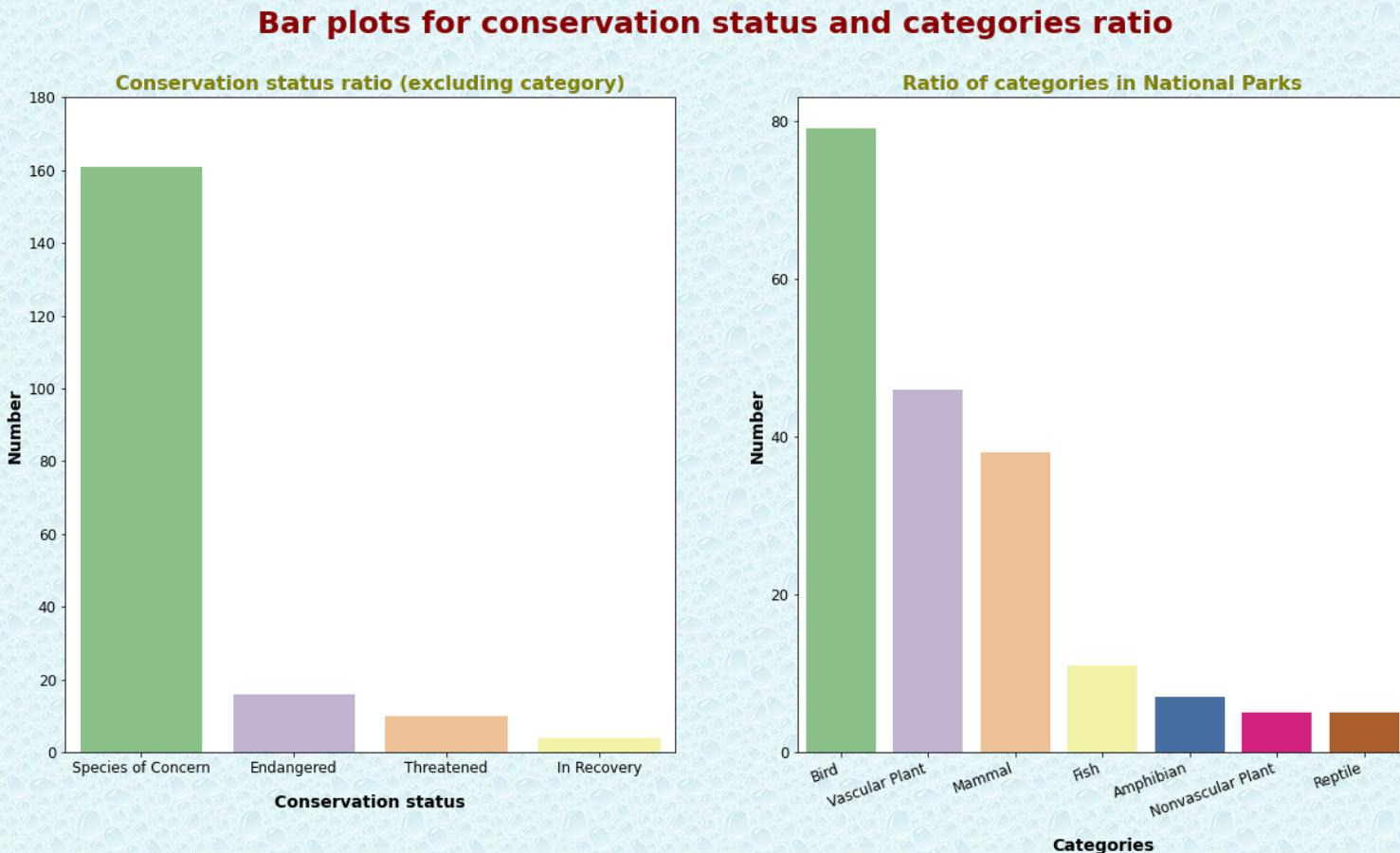
	category	count
0	Amphibian	7
1	Bird	79
2	Fish	11
3	Mammal	38
4	Nonvascular Plant	5
5	Reptile	5
6	Vascular Plant	46

Далее отсортируем эти таблицы и построим графики для каждой из них.

Визуализация

Как видно из рисунка, наиболее часто встречающийся охранный статус — виды, вызывающие озабоченность. 161 запись в таблице из 191 имеет этот статус. Наименее часто встречающийся охранный статус — в процессе восстановления (только 4 записи).

Наиболее популярная группа живых существ — это птицы. Они занимают практически половину (более 41%) всех живых существ в рассматриваемых Национальных парках. Наименее популярными являются низшие растения и рептилии.

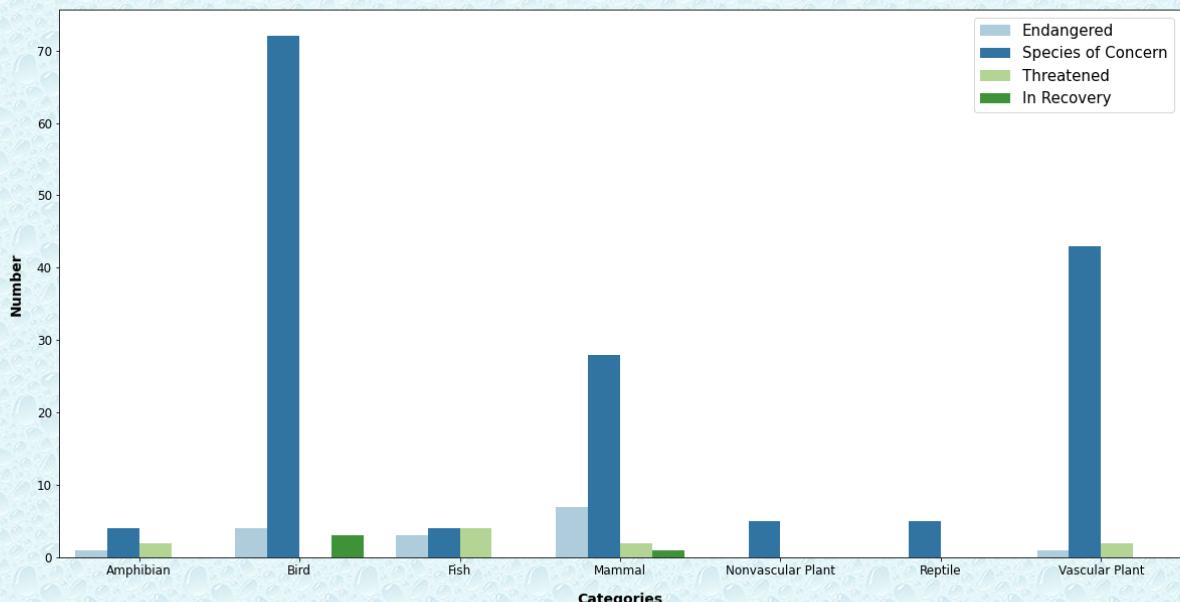


Визуализация

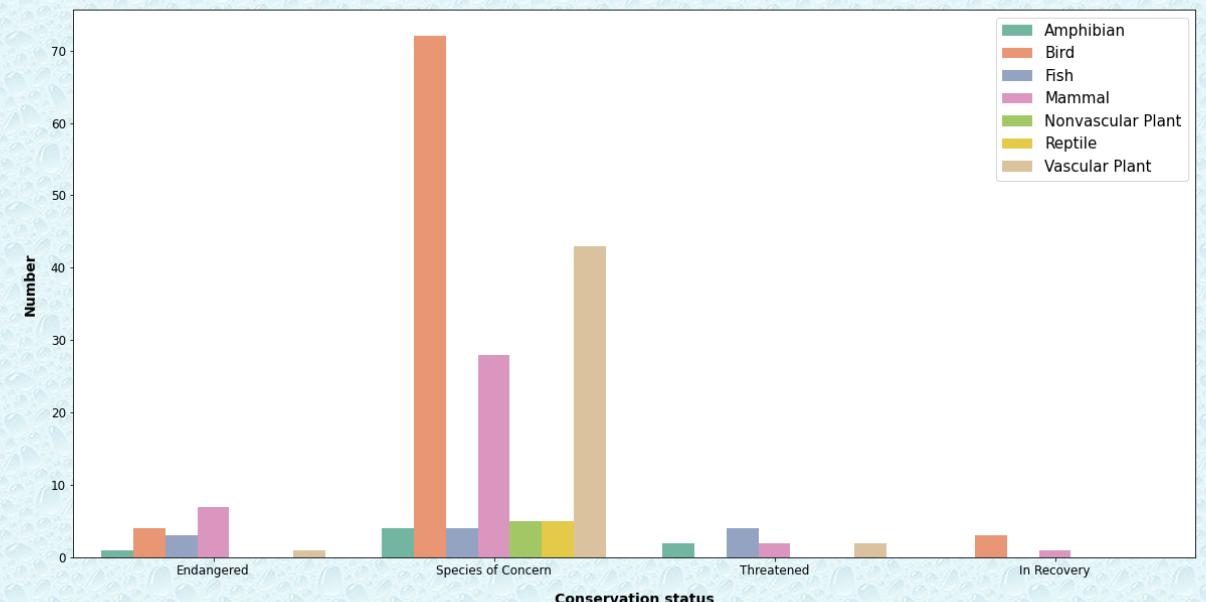
Прошлые графики содержат иллюстрацию охранного статуса, групп живых существ и позволяют наглядно показать распределение этих двух показателей по отдельности. Однако, по ним можно лишь определить, сколько записей в таблице связано с той или иной группой живых существ (без учёта делений по охранным статусам) или с тем или иным охранным статусом .

Но они будут бесполезны, если мы хотим узнать, как в рамках одного охранных статуса распределены группы живых существ или наоборот. Для этого построим новые график, добавив в функцию создания графика параметр `hue`.

Different categories by their conservation status



Different conservation statuses by their category



Выводы (1-ая часть)

В процессе выполнения проекта о биологическом разнообразии в Национальных парках мной были выполнены следующие шаги:

- 1 Исходные файлы с данными были преобразованы в более удобный формат таблицы (DataFrame). В итоге, мы имеем получили таблицы «species» и «observations», содержащие все необходимые для дальнейшего анализа данные.
- 2 Практически 95% записей в таблице «species» имели пропущенное значение в охранном статусе. Такие данные могли бы существенно исказить результаты анализа, поэтому перед непосредственно анализом был проведён этап очистки данных. Он позволил существенно снизить количество рассматриваемых записей.
- 3 В первоначальном виде таблицы содержали лишнюю информацию, которая не представляла особой ценности для визуализации и проведения тестов для проверки статистических гипотез. Поэтому, на основе таблиц «species» и «observations» были созданы новые таблицы, содержащие только необходимые данные для последующего анализа и визуализации. Новые таблицы создавались не только путём отделения нужных столбцов от ненужных. Pandas предлагает несколько вариантов группировки и обобщения данных, что в свою очередь является мощными инструментами анализа. Одними из таких инструментов, которые и были использованы в проекте, являются сводные и кросс-таблицы.

Выводы (1-ая часть)

4 Для того, чтобы определить существование зависимостей между данными, был проведён тест для проверки статистической гипотезы. В ходе проведения теста были сформулированы нулевая и альтернативная гипотезы, получено р-значение (p-value) на основе которого и принималось решение о принятии/отклонении нулевой гипотезы и наличии/отсутствии зависимости между исследуемыми данными.

5 После того, как нужные данные вынесены в отдельные таблицы и сгруппированы, было построено несколько графиков различных видов для повышения их понимания, читаемости и более обоснованных выводов.

После выполнения всех вышеперечисленных шагов, можно сделать выводы о биологическом разнообразии в Национальных парках и предоставить некоторые рекомендации касательного того, как можно улучшить сбор данных в Национальных парках.

Выводы (2-ая часть)

На основе проделанной работы, можно сделать некоторые выводы касательно биологическом разнообразии в Национальных парках и дать некоторые рекомендации по улучшению сбора данных.

С помощью теста для проверки статистических гипотез мы выяснили, что существует взаимосвязь между охранным статусом и группой живых существ. Из кросс-таблицы, представленной ниже, можно заметить значительно преобладание у всех групп живых существ охранного статуса «Species of concern». Он означает, что в отношении этих живых существ существуют некоторые опасения относительно статуса и угроз. Такое существенное преобладание является неблагоприятным и опасным, ведь существует высокий риск начала исчезновения определённых видов в природе. Особое внимание необходимо уделить таким группам, как птицы, млекопитающие и сосудистые растения. Именно они оставляют около 80% всех записей в таблице и, по сравнению с остальными группами, имеют больший шанс перейти в разряд исчезающих и находящихся под угрозой исчезновения за счёт преобладания существ с охранным статусом «Species of concern».

Также, стоит уделить внимание рептилиям и низшим растениям, ведь записей по этим двум группам недостаточно, чтобы сделать какие-либо выводы. Из таблицы можно увидеть лишь то, что у всех живых существ обеих группы существует опасения относительно статуса и угроз.

conservation_status category	Endangered	In Recovery	Species of Concern	Threatened
Amphibian	1	0	4	2
Bird	4	3	72	0
Fish	3	0	4	4
Mammal	7	1	28	2
Nonvascular Plant	0	0	5	0
Reptile	0	0	5	0
Vascular Plant	1	0	43	2

Выводы (2-ая часть)

Отдельно нужно сказать о количестве наблюдений для анализа. Хоть таблица «species» и состоит практически из 6000 записей, практически во всех из них отсутствует значение охранного статуса. Лишь в 3,3% записей не пропущен данный столбец, а следовательно, только эти данные и были использованы при анализе и визуализации. К примеру, ниже будет представлено соотношение общего количества записей для определённой группы и записей для этой же группы, в которых отсутствует охранный статус:

- сосудистые растения 4470/4424;
- птицы 521/421;
- низшие растения 333/328;
- млекопитающие 214/176;
- рыбы 127/116;
- амфибии 80/73;
- рептилии 79/74.

Ни для одной из групп количество записей с заполненным охранным статусом не превышает 20%. Хоть и с такими данными можно сделать выводы о распределении биологическом разнообразии в Национальных парках, но они будут менее точными и достоверными.

Из данного соотношения можно сказать, что сосудистые растения имеют значительное количественное преимущество по сравнению с другими группами. Однако, лишь 1% записей для этой группы подлежал рассмотрению и визуализации из-за отсутствия охранного статуса.

Выводы (2-ая часть)

Ещё один вывод будет связан с Национальными парками и количеством наблюдений в них. В таблице «*observations*» содержится более 23296 записей. При этом, между четырьмя Национальными парками они распределены одинаково, по 5824 записей для каждого.

Ранее были представлены 2 графика для демонстрации количества наблюдений в каждом из парков:

- круговая диаграмма, отражающая процент, который занимает количество наблюдений в каждом из парков в общем числе наблюдений;
- столбчатая диаграмма, созданная для этой же цели, однако, отображающая долю каждого из парков в числовой форме, а не в процентах.

Из этих двух графиков можно сказать, что больше всего наблюдений было проведено в Yellowstone National Park (43,5% или 1443562 наблюдений от общего количества). Второе место занимает Yosemite National Park (26% или 863332), третье – Bryce National Park (17,4% или 576025) и четвёртое – Great Smoky Mountains National Park (13% или 431820).

Такой отрыв по количеству наблюдений можно объяснить различиями в занимаемой парком территории. Территория Yellowstone National Park составляет $8\ 991\ km^2$, что позволяет разместить намного больше живых существ, нежели чем другие парки, территория которых не превышает отметки в $3100\ km^2$.

Рекомендации

В ходе проведения анализа были выявлены некоторые недочёты в исходных данных, которые могли повлиять на результаты анализа и визуализации. Их исправление, на мой взгляд, может улучшить качество исходных данных и позволит провести более детальный и тщательный анализ.

1 Из-за огромного количества пропущенных значений охранного статуса была потеряна значительная часть информации для анализа. Заполнив этот столбец для большинства записей, можно сделать более точные выводы касательно исчезновения и угроз для каждой из группы живых существ.

2 Таблица «species» содержит общие группы живых существ, например, рыбы, птицы или млекопитающие. Также, в таблице есть столбец `common_names`, содержащий точные имена каждого существа, но он не несёт в себе особой пользы. Он содержит слишком уточнённое название, например, «*Aurochs*, *Aurochs*, Domestic Cattle (Feral), Domesticated Cattle». Как результат, столбец содержит более 5000 уникальных значений, что достаточно проблематично визуализировать.

Для более точной и удобной классификации можно добавить столбец, уточняющий вид живого существа не так подробно, как в столбце `common_names`. К примеру, у млекопитающих он мог быть заполнен такими подгруппами, как волк, лев, верблюд, дельфин и так далее. Это не является трудным для заполнения, но позволит строить более детальные графики и сводные таблицы.

3 С помощью существующей информации мы можем оценить и сравнить Национальные парки только по количеству наблюдений в них. Таблица «species» не содержит уточнение, в каком из Национальных парков находится то или иное живое существо. Чтобы узнать более подробную характеристику и статистику каждого из парков, можно добавить в таблицу «species» столбец с названием парка. Это позволит нам делать выводы о том, какая группа существ преобладает в каждом из парков, в каком из них можно увидеть больше всего существ, находящихся под угрозой исчезновения и так далее.