Biodiversity Project

As a data analyst I am going to help analyze different species of National Park.

AUTHOR:
ANIRUDH KUMAR

1

As a first step,necessary python libraries were imported and analyzed input data.

Initially 2 csv documents: species and observations.

```
from matplotlib import pyplot as plt
import pandas as pd
```

```
species.head()
```

	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

1

Given dataset has 7 different species.
Species names are given below:-

```
list(species.category.unique())
```

```
['Mammal',
 'Bird',
 'Reptile',
 'Amphibian',
 'Fish',
 'Vascular Plant',
 'Nonvascular Plant']
```

Species are grouped into 5 types by conservation status: ***['NO INTERVENTION', 'SPECIES OF CONCERN', 'ENDANGERED', 'THREATENED', 'IN RECOVERY']***

2

The goal of work was to analyze dataset to get the insight about endangered animals. Table below shows sum of scientific names for each conservation status group.

	conservation_status	Sum scientific_names
1	In Recovery	4
4	Threatened	10
0	Endangered	15
3	Species of Concern	151
2	No Intervention	5363

It is seen than big percent of animals have No Intervention status.

2

But percent protected estimation per each category gives more valuable information.

	category	number_save	number_in_danger	percent_protected
3	Mammal	146	30	0.830
1	Bird	413	75	0.846
0	Amphibian	72	7	0.911
2	Fish	115	11	0.913
5	Reptile	73	5	0.936

We see that Mammals are more likely to be endangered than other categories.

Chi square test resulted that birds are almost in the same endangered level as Mammals because there is no a significant difference between datasets. Significance arises when comparing with Reptiles.

3

Sightings of different species at several national parks for the past 7 days.

```
observations = pd.read_csv("observations.csv")
observations.head(5)
```

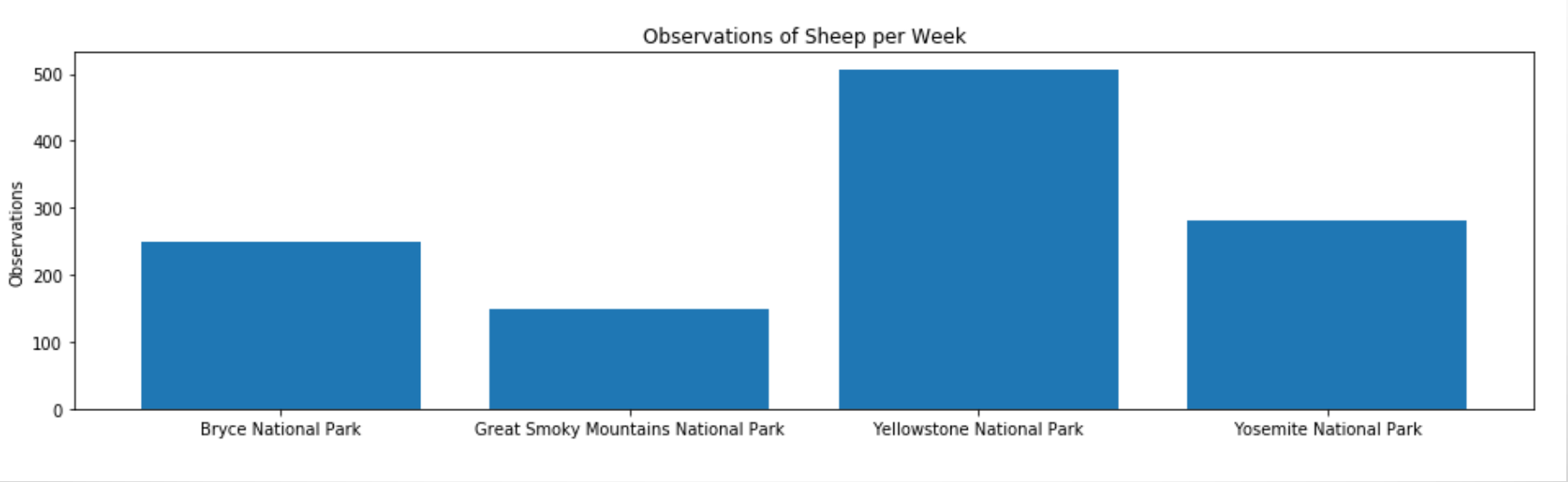
	scientific_name	park_name	observations
0	Vicia benghalensis	Great Smoky Mountains National Park	68
1	Neovison vison	Great Smoky Mountains National Park	77
2	Prunus subcordata	Yosemite National Park	138
3	Abutilon theophrasti	Bryce National Park	84
4	Githopsis specularioides	Great Smoky Mountains National Park	85

3

Scientist are studying the number of sheep sightings at different national parks.

Lets use lambda, apply and groupby functions and get table with number of sheep in each park. Yellowstone park has the biggest amount of sheep.

	park_name	observations
0	Bryce National Park	250
1	Great Smoky Mountains National Park	149
2	Yellowstone National Park	507
3	Yosemite National Park	282



3

Scientist know that 15% of sheep at Bryce National Park have foot and mouth disease. They want to be able to detect disease reductions of at least 5 percentage point. They want to know number of required observations.

Optimizely calculator helped me to estimate number of sheep that scientist need to observe - 510.

	park_name	observations
0	Bryce National Park	250
1	Great Smoky Mountains National Park	149
2	Yellowstone National Park	507
3	Yosemite National Park	282

As a result, we can conclude that scientist need 2 weeks to observe sheep in Bryce National Park and 1 week in Yellowstone National Park .

Thank You