Capstone Option 2: Biodiversity for the National Parks

Mayank Naithani

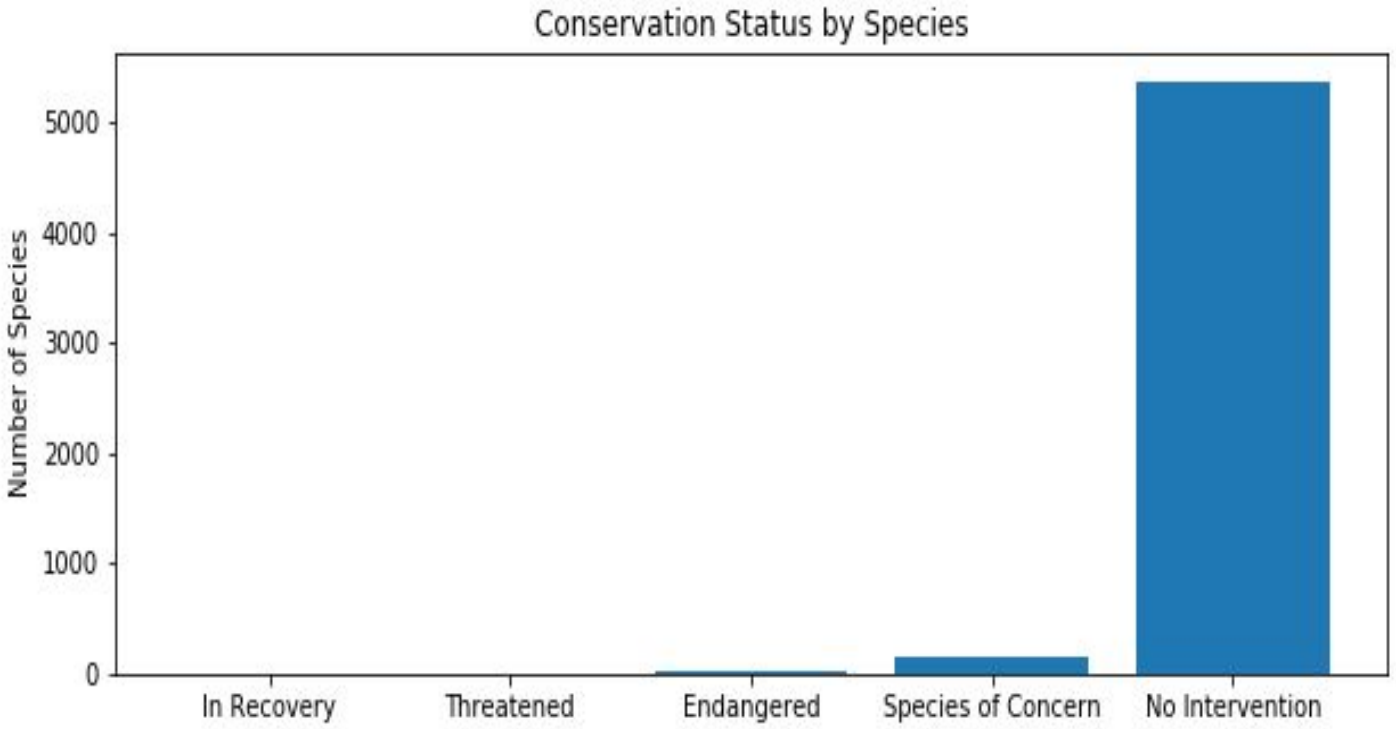
First data set given is sepcies_info.csv, upon importing it to species dataset via pandas and inspecting it can be found that

- It contains scientific names of each species
- Common name of each species
- Conservation status

species.info() return the below insights:

- That data species has 5824 rows (indexing from 0 to 5823)
- 4 columns (Category, scientific_name, common_names and conservation status)
- Conservation status has large number of null values (only 191 non-null values are there in 5824 rows)

- There are 5541 different species in species data frame
- There are 7 types of different categories these species fall into (‘Mammal’, ‘Bird’, ‘Reptile’, ‘Amphibian’, ‘Fish’, ‘Vascular Plant’ and ‘Nonvascular Plant’
- There is 4 types of conservation_status for these species (other than missing values) (‘Species of concern’, ‘Endangered’, ‘Threatened’ and ‘In Recovery’)
- Missing values for conservation_status is replace by ‘No Intervention’ (pd.fillna)
- Using matplotlib a histogram is created (refer next slide please) which depicts number of species against Conservation Status
- It can be seen in chart that ‘No Intervention’ has the way more number of species then rest of the categories, which kind of make sense as it's not common to find species which are either’ Species of concern’, ‘Endangered’, ‘Threatened’ or ‘In Recovery’


Conservation Status by Species
Number of Species
5000
4000
3000
2000
1000
0
In Recovery
Threatened
Endangered
Species of Concern
No Intervention

Investigating endangered species

Now we will investigate whether certain type of species is more like to be endangered than others, as these are categorical variables we will use Chi-Square test for our hypothesis

For performing Chi Square test another column ‘is_protected’ is added which is True if conservation_status is not equal to ‘No Intervention’ else False

Using groupby on ‘Category’ and ‘is_protected’ and applying pivot on that we can create a data frame category_count which will help me creating contingency table for Chi Square test

Investigating endangered species : Cont.

Based on pivot table contingency table is created for Bird and Mammal and upon performing Chi square test p value is found to be 0.68 which is considerable higher than 0.05 so its not significant and thus we can’t say there is no significant difference between Bird and Mammal in terms

Chi Square test between Reptile and Mammal gave p-value of 0.03 (which is smaller than 0.05) hence there is a significant difference between Reptile and Mammal

Conclusion on Protected Species

When we ran our chi-squared test, we found a p-value of around 0.688, so we can conclude that the difference between the percentages of protected birds and mammals is not significant and is a result of chance.

But, when we compared the percentages of protected reptiles and mammals and ran the same chi-squared test, we calculated a p-value of ~ 0.038, which is significant.

Therefore, we can conclude that certain types of species are more likely to be endangered than others.

Observations Dataframe

‘Observations.csv contains’ data of sighting of different species at several national parks for the past 7 days

It contains 3 columns ‘scientifc_name’, ‘park_name’ and ‘observations’

We are interested in animal sheep hence applying lambda function new column ‘is_sheep’ is created which contains all the common names with word ‘sheep’

Many of the results are actually plants hence we further filtered species dataframe for ‘Mammal’ category and saved it in sheep_species

```
sheep_species = species[(species.is_sheep == True) & (species.category == 'Mammal')]
```

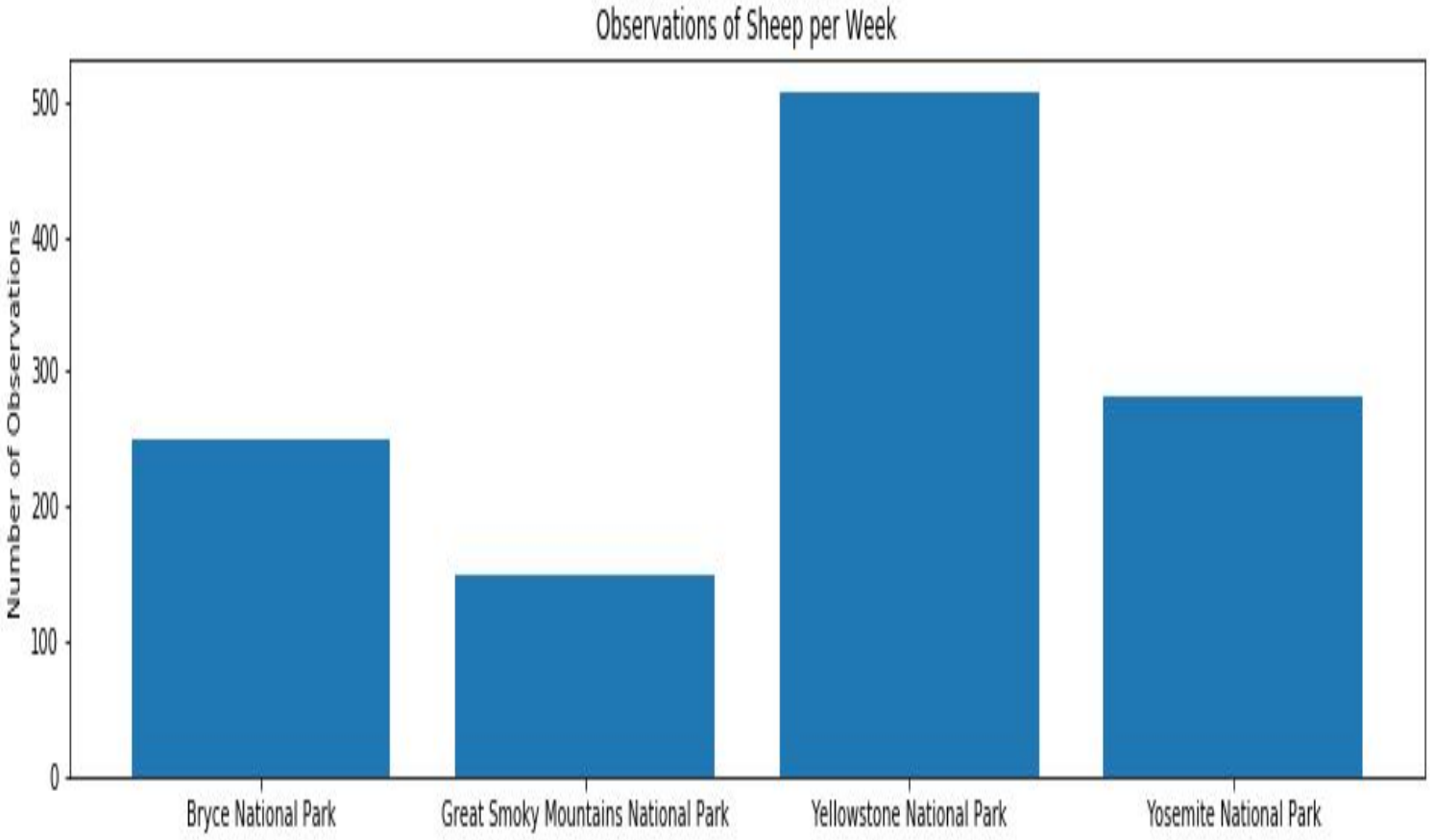
To know where these sheeps are located (park) we will combine sheep_species data frame with observations data frame using merge and save the result into sheep_observations

```
sheep_observations = observations.merge(sheep_species)
```

Using group by on ‘park_name’ on sheep_observations dataframe we get the total number of sheeps observed in various parks

obs_by_park = sheep_observations.groupby('park_name').observations.sum().reset_index()

‘Yellowstone National Park’ has highest number of sheep sighting with 507 observations


Observations of Sheep per Week
Number of Observations
0
100
200
300
400
500
Bryce National Park
Great Smoky Mountains National Park
Yellowstone National Park
Yosemite National Park

Foot and Mouth Reduction Effort

Given a baseline of 15% occurrence of foot and mouth disease in sheep at Bryce National Park, we found that if the scientists wanted to be sure that a >5% drop in observed cases of foot and mouth disease in the sheep at Yellowstone was significant they would have to observe at least 510 sheep.

This would take approximately one week of observing in Yellowstone to see that many sheep, or approximately two weeks in Bryce to see that many sheep