Capstone Option 2: Biodiversity for the National Parks

By SSamal93

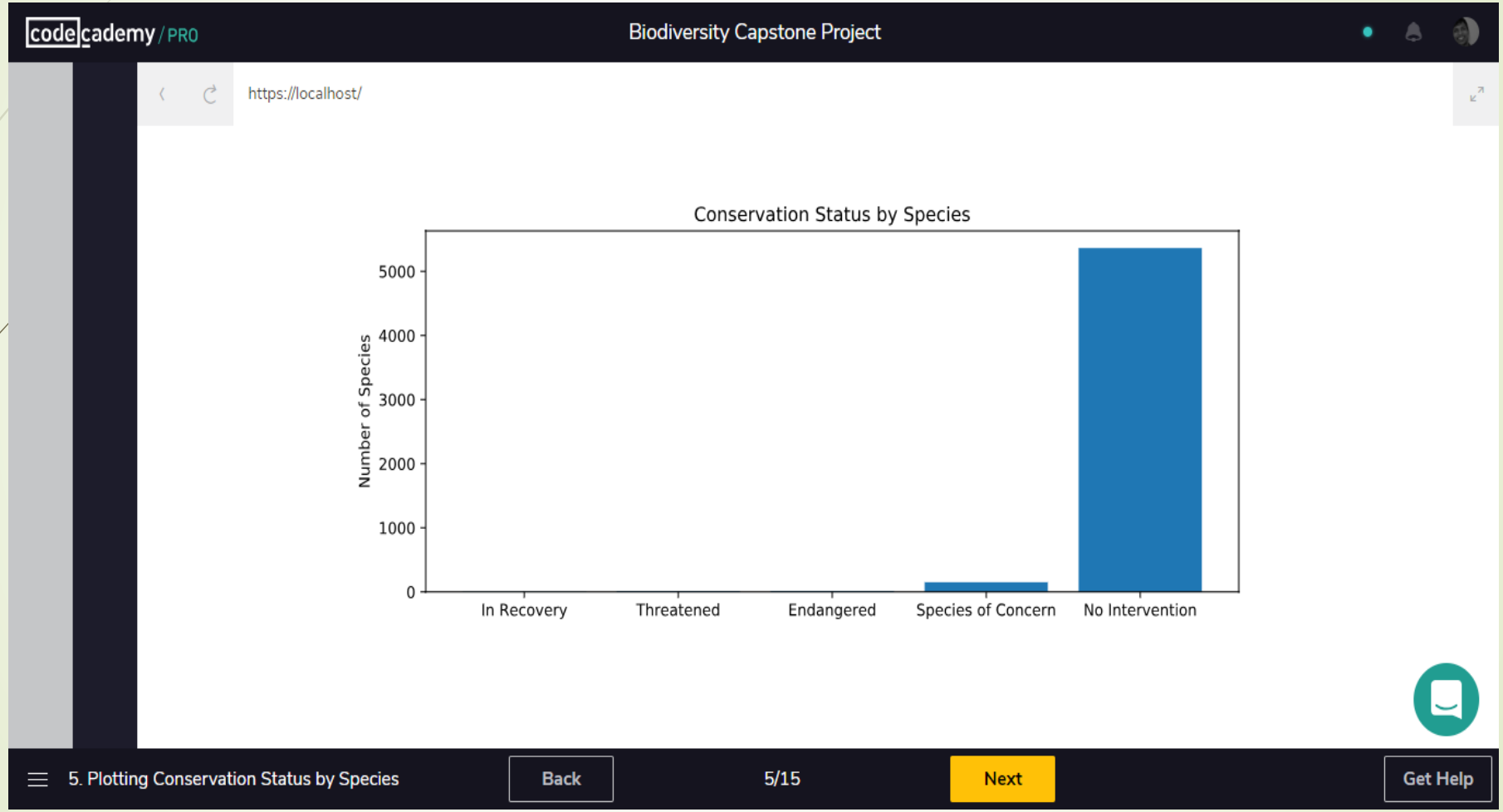
Species
species_info.csv

- Data was provided to us in the form of a CSV file **"species_info.csv"** that contained data about the various species in the National Parks. The original format of the data was:
  - The **scientific name** of each species
  - The **common names** of each species
  - The species **conservation status**
- On initial inspection, the total number of types of species, calculated based on each species' unique scientific name was found to be **5541**.

- On grouping them based on their conservation status, the number of species(as per their scientific names) in each conservation status category was found to be:

	Conservation Status	Scientific Names
1	Endangered	15
2	In Recovery	4
3	No Intervention	5363
4	Species of Concern	151
5	Threatened	10

- The same table can be shown as a bar chart as below:



Probing the 'Endangered' Species

- We then grouped our species based on their category & conservation status into the following 2 broad groups:-
  - **Non Protected** : Category of Species with conservation status as
    - No Intervention
  - **Protected** : Category of Species with one of the following conservation status'
    - Endangered
    - In Recovery
    - Species of Concern
    - Threatened
- We then proceed to find the **percentage protected** for each category to get a clear indication as to which species needs most protection.

- The following table gives the **percentage protected** for each category:

is_protected	category	False	True
0	Amphibian	73	7
1	Bird	442	79
2	Fish	116	11
3	Mammal	176	38
4	Nonvascular Plant	328	5

	category	not_protected	protected
0	Amphibian	73	7
1	Bird	442	79
2	Fish	116	11
3	Mammal	176	38
4	Nonvascular Plant	328	5

	category	not_protected	protected	percent_protected
0	Amphibian	73	7	8.750000
1	Bird	442	79	15.163148
2	Fish	116	11	8.661417
3	Mammal	176	38	17.757009
4	Nonvascular Plant	328	5	1.501502

Significance calculations for 'Endangered' species

- As per the **percentage protected** in our previous table**, i**t looked like certain category of species were more likely to be endangered than others, but was this true? Or was this difference just due to chance(null hypothesis)? To answer this, we carried out significance tests between:-
  - Mammal – Birds
  - Mammal – Reptile
- We chose the **Chi-square test** that is intended to **test** how likely it is that an observed distribution is due to chance. The pvalues we got:
  - Mammal – Birds : 0.687594809666 : No significant difference (> 0.05)
  - Mammal – Reptile : 0.0383555902297 : Significant difference! (< 0.05)

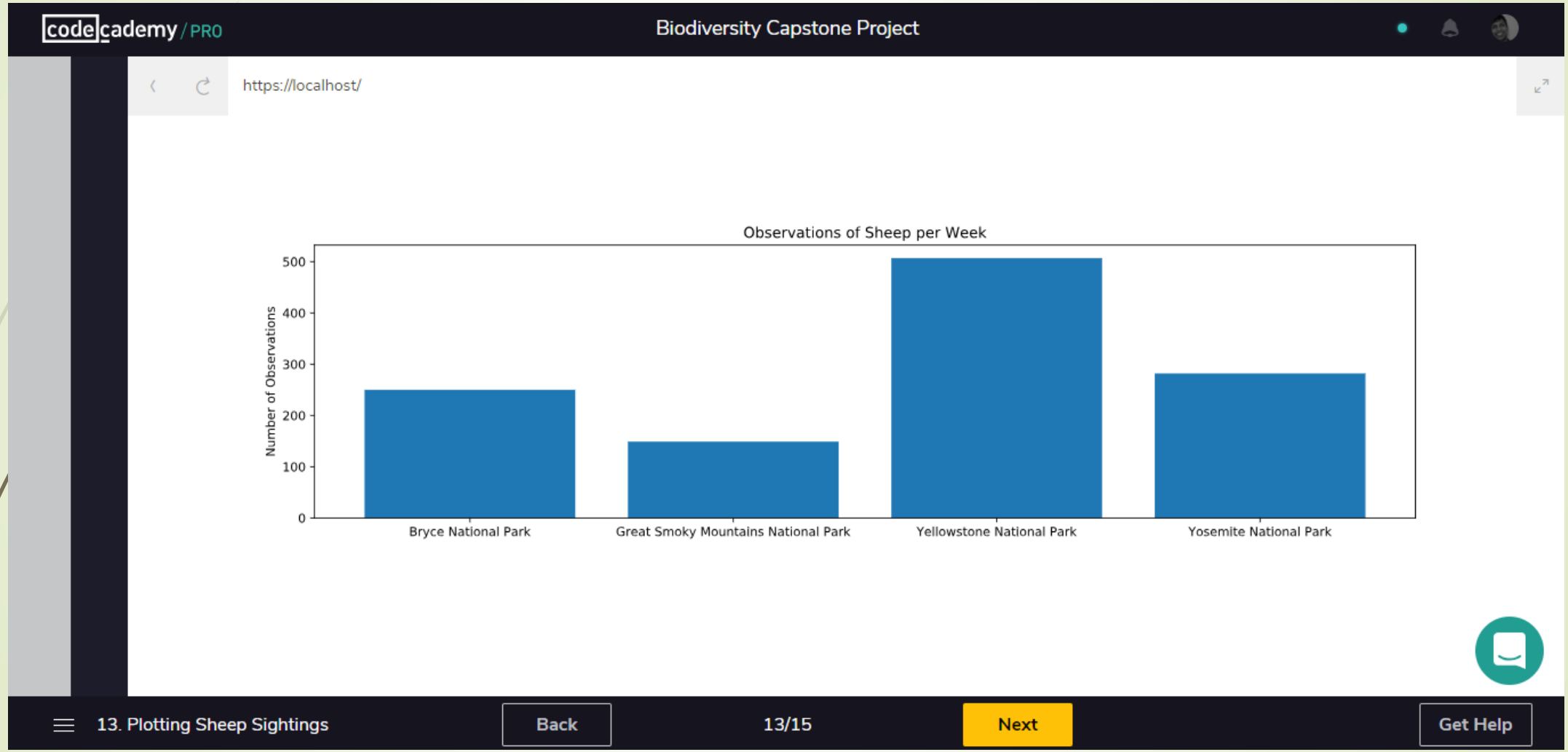
Recommendation for Conservationists

- The Chi-Square test on Mammals & Reptiles gave a p-value of ~0.038. **Therefore, we can conclude that certain types of species are more likely to be endangered than others.**
- The 38 **Mammal** species that fall under the Protected category need more protection than their Repitilian counterparts.
- Similarly, Chi Square test values can be used to find out which other species need conservation.

Observations
observations.csv

- Data was provided to us in the form of a CSV file **"observations.csv"** that contained data about sightings of different species at several national parks for a period of 7 days. The original format of the data was:
  - The **scientific name** of each species
  - The **park name** of each species
  - The number of sightings as **observations**
- We then found out the number of species that had **"Sheep"** in their **common name** and **category** mammals from Species dataframe.
- We proceeded to merge the **Species** and **Observations** dataframes to find the number of Sheep sightings/observations in the various national parks.

- The following bar chart shows the number of Sheep sightings/observations in the various national parks:



Foot and Mouth Reduction Effort Sample Size Determination

- Info provided by the Park Rangers were:
  - "The only information that the scientists currently have is that last year it was recorded that 15% of sheep at Bryce National Park have foot and mouth disease." ==> **Baseline**: 15%
  - "They want to be able to detect reductions of at least 5 percentage points." ==>**Minimum detectable effect**: 5%
  - Use the default level of significance (90%).
- Step I : If we want to observe an x% change with confidence, our **minimum detectable effect** would be equal to 100 * x / baseline, I.e, 100 * 5 / 15 = 33.33%
- Step II : We plug in **baseline, level of significance & minimum detectable effect values** into the sample size calculator & a **sample size** of **870**.

- We can further calculate the number of weeks scientists would need to spend at Yellowstone National Park to observe enough sheep as:

  Sample Size (870) / No of observations of Sheep in Yellowstone (507)

  = 1.71597633136 ~ 1.8 weeks

- The No of observations of Sheep in Yellowstone was obtained from the **Sheep Observations table** we made by merging **Sheep Species table** and **Observations table**.

The End