Biodiversity at National Parks Data Analysis

Submission for Capstone Project June 2018

Diane Weiss

xarmanla@gmail.com

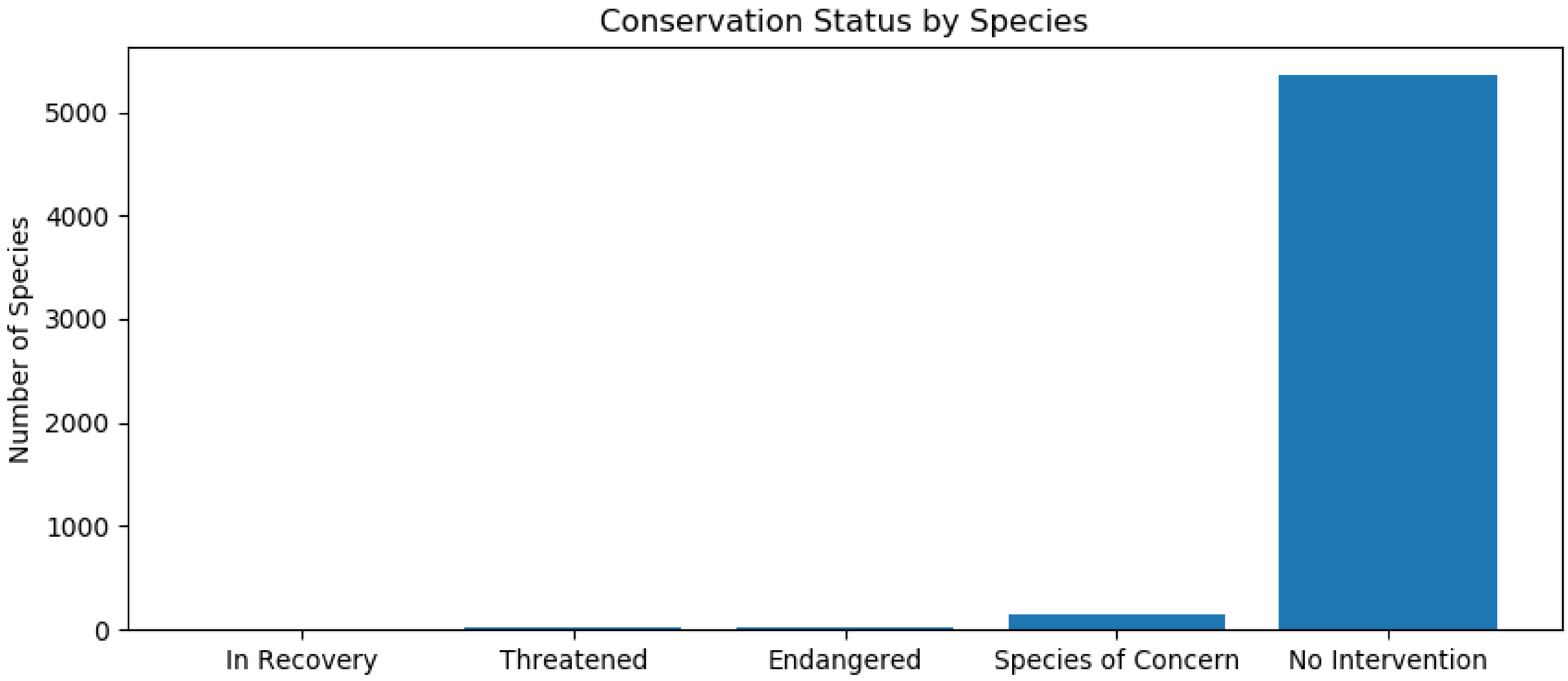
U.S. National Park
Capstone Overview

- Objective: Apply data analysis methodologies to answer real world questions regarding species found within U.S. National Parks
- Resources: Two .csv files were given, one providing information about plant and animal species found at National Parks, and one detailing sighting observations at some of the National Parks
- Methodology: python code was written using several publicly available packages to extract information, create plots, and aggregate data from both resource files. Additionally, a tool was used to determine sample size for statistic validity.
- Findings: Questions regarding determining category specifics for endangered species, ploting sighting of a specific animal, and determining duration requirements for a statistically valid disease treatment study.

Preliminary Investigation of the Data: Endangered Species in the National Parks

- The species file contained categorical data for a species ranging from “species of concern”, through “endangered”, to “threatened”, and finally, “in recovery”.
- Using python code with pandas, and specifically, the groupby function, along with the mathplotlib, a bar graph was produced with the height of each bar showing how many species fell into each category.
- The generated bar chart is shown in the following slide.

Bar Chart Produced with mathplotlib



Answering a Specific Question Related to Endangered Species

- After computing the percentage of protected species by category, the question arose as to whether any of these percentages were significantly different for various pairings of categories.
- At first, the percentage of protected mammals appears to be higher than the percentage of protected birds. Because the data was numeric, the chi squared test was chosen to determine if the difference in percentages was significant.
- The scipy.stats chi2_contingency function was run on a table constructed to pair mammals versus birds. The pval for that pairing was more than 0.86. This implies that the percentage difference could have been a chance occurance.

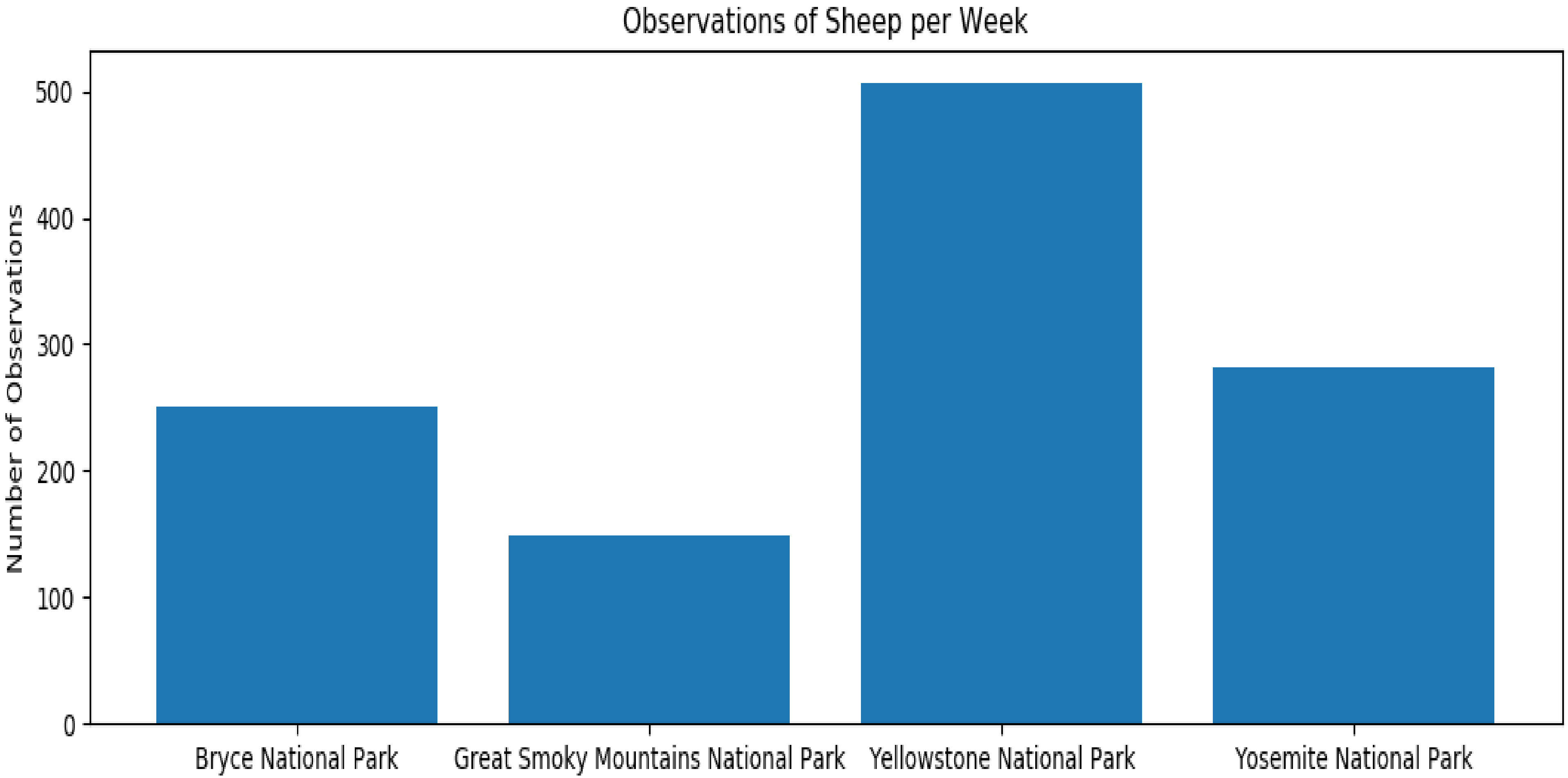
Answering a Specific Question Related to Endangered Species (continued)

- A second contingency table was then constructed to examine the percentage difference between mammals and reptiles.
- The percentage protected for mammals was about 17% and for reptiles, it was just over 6%.
- The chi squared test computed a pval of 0.38. Since 0.38 is less than 0.05, it is appropriate to conclude that the percentage difference is significant.
- The statistical answer to the question regarding the two pairings confirms that more mammals are protected and reptiles, but that the percentage is about the same for mammals and birds.

Investigating Sheep Sightings in U.S. National Parks

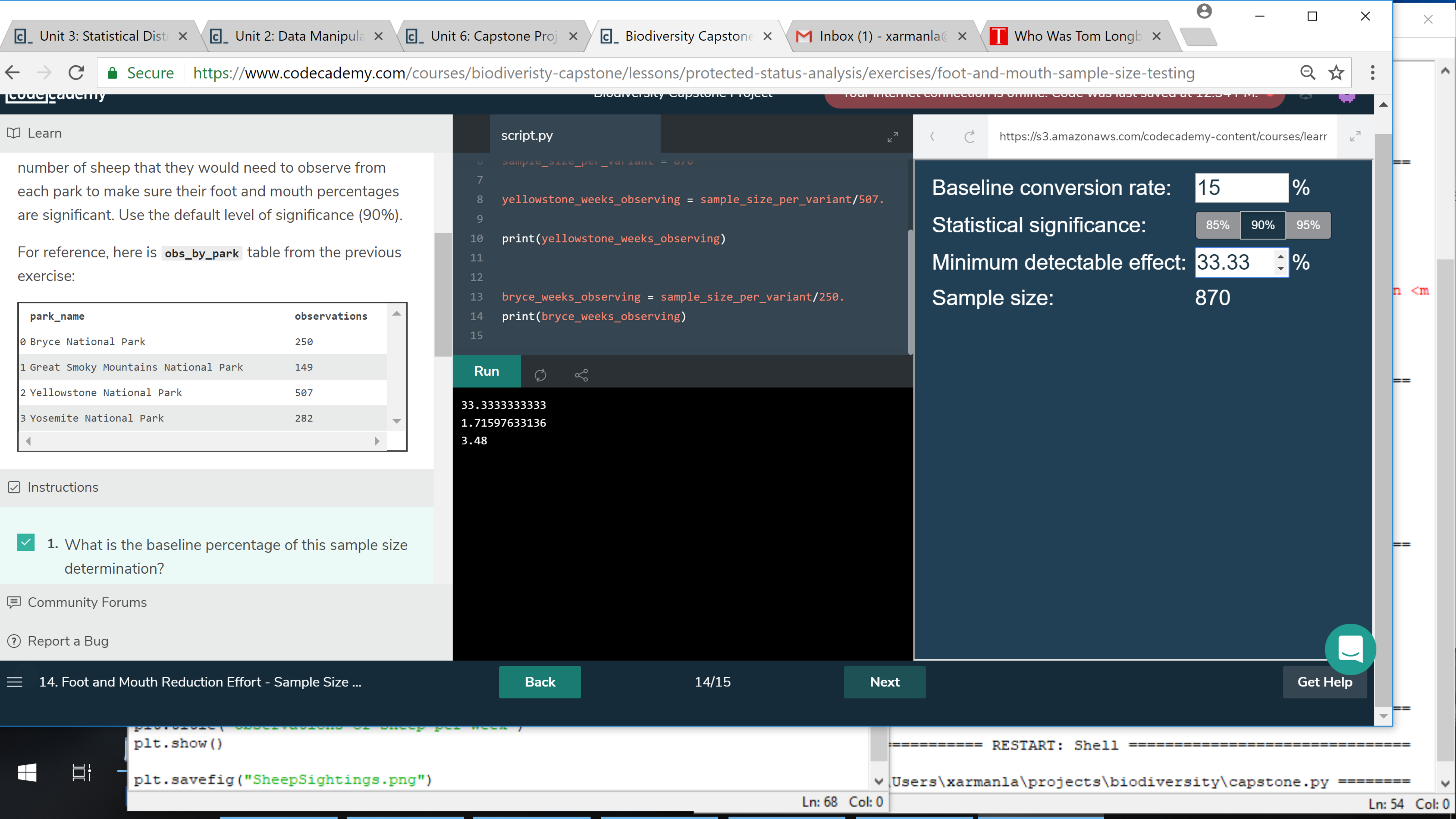
- Leaving behind protection status, we used the scientific names and common names found in the species data file to ascertain which mammals were sheep.
- We then used the merge and groupby functions of pandas to chart sheep in National Parks irrespective of species.
- We again used matplotlib to create a bar chart showing the number of sheep sightings by National Park.

Bar Chart produced by matplotlib



Determining Minimum Sample Size for Treatment of Foot and Mouth Disease

- The final question addressed using the sightings data was determining the minimum sample size in a treatment study.
- Yellowstone Park personnel knew the baseline data from Bryce Canyon and also knew that they wished to achieve a 5% reduction in foot and mouth disease in Yellowstone Park.
- We used a standard tool for determining minimum sample size based on “baseline conversion rate”, “statistical significance”, “minimum detectable effect”, and “sample size”.
- A screen shot of the tool and the code is in the next slide.



number of sheep that they would need to observe from each park to make sure their foot and mouth percentages are significant. Use the default level of significance (90%).

For reference, here is `obs_by_park` table from the previous exercise:

	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

```
sample_size_per_variant = 870

yellowstone_weeks_observing = sample_size_per_variant/507.

print(yellowstone_weeks_observing)


bryce_weeks_observing = sample_size_per_variant/250.
print(bryce_weeks_observing)

```

```
33.3333333333
1.71597633136
3.48
```

Baseline conversion rate: 15 %

Statistical significance: 85% [✓ 90%] 95%

Minimum detectable effect: 33.33 %

Sample size: 870

☑ 1. What is the baseline percentage of this sample size determination?

14. Foot and Mouth Reduction Effort - Sample Size ...

14/15

```
plt.show()

plt.savefig("SheepSightings.png")
```

Ln: 68 Col: 0

```
=========== RESTART: Shell ==================================
Users\xarmanla\projects\biodiversity\capstone.py ========
```

Ln: 54 Col: 0

Foot and Mouth Disease Treatment Study Duration

- Using a 90% statistical significance and a baseline of 15% in Bryce Canyon, the tool determined that 870 sheep needed to be observed.
- Combining the 870 sheep needed with the number of sheep observed in each park, we computed the length of time need to reach 870. Using the data for one week, we determined that the study had to last almost two weeks in Yellowstone Park and three and a half weeks in Bryce Canyon.