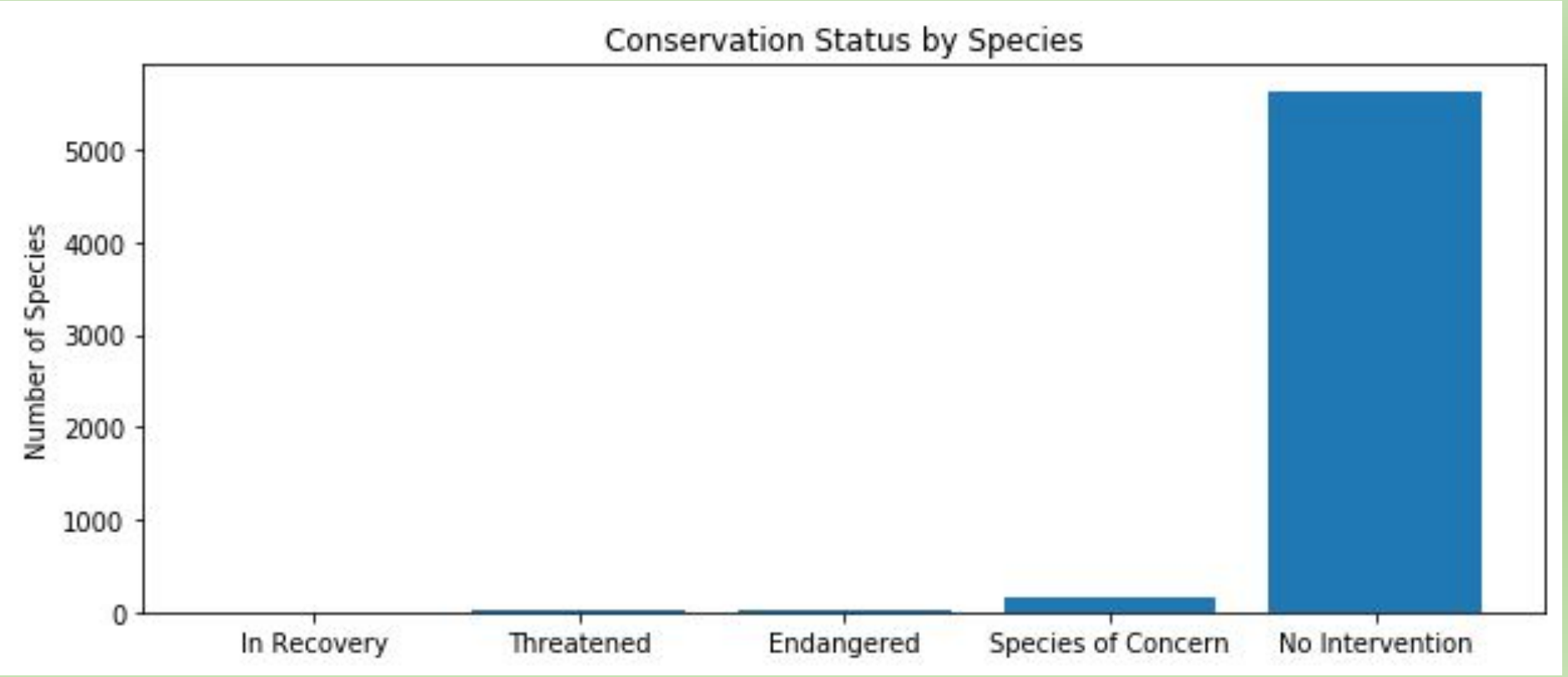
Biodiversity Project

- The species_info.csv file simply contains Category, Scientific Name, Common Name, and Conservation status for various organisms.
- There are 5,541 unique organisms listed.
- Categorized as Mammals, Birds, Reptiles, Amphibians, Fish, Vascular Plants, and Non vascular plants.
- The conservation statuses are listed as nan, Species of concern, Endangered, Threatened, and In Recovery.

To see a breakdown of the data, it needed to be grouped together. This was accomplished by grouping the data by conservation status. In order to show all the data the null or ‘nan’ values needed to be populated. This was done using the “.fillna” process in python. Then we plotted the new table in a bar graph shown below.



- We then wanted to see if certain types of organisms were more likely to be endangered than others. This was accomplished by creating a new column in the table called ‘is protected’ by which was set to ‘False’ if conservation status was ‘No Intervention’ and ‘True’ for any other status.
- This created a table with two values for each category with totals for True and False, so to better represent the data we needed to create a pivot table.

Un-pivoted Table

	category	is_protected	scientific_name
0	Amphibian	False	72
1	Amphibian	True	7
2	Bird	False	413
3	Bird	True	75
4	Fish	False	115

Pivoted Table

is_protected	category	False	True
0	Amphibian	72	7
1	Bird	413	75
2	Fish	115	11
3	Mammal	146	30
4	Nonvascular Plant	328	5
5	Reptile	73	5
6	Vascular Plant	4216	46

- To test to see if these numbers were significant we could run the Chi Squared test.
- We first check between mammals and birds. Which returned these results:

```
(0.1617014831654557, 0.6875948096661336, 1L, array([[ 27.8313253, 148.1686747],
       [ 77.1686747, 410.8313253]]))
```

- The p-values are both above .05 making them not significant.
- So we check again for Reptile and mammal. The test results this time showed as:

```
(4.289183096203645, 0.03835559022969898, 1L, array([[ 24.2519685, 151.7480315],
       [ 10.7480315,  67.2519685]]))
```

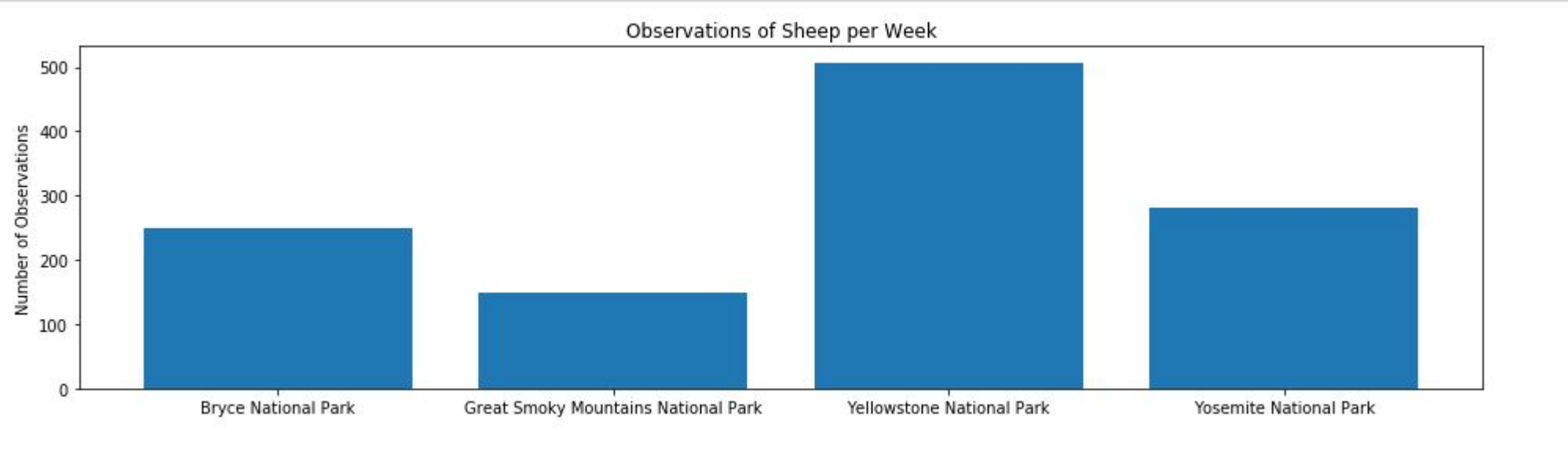
- This time we have a p-value that is less than .05 showing significance.

Observations

- We look at the observation portion of the biodiversity project, and we see that oberservation.csv has columns for scientific name, park name, and number of observations.
- We're interested in only 'sheep' so we look for the word sheep in the speicies_info.csv using the 'IN' operator. Returning only 'sheep'.
- We then merge the new 'sheep' table with the 'observation' table, and group them by park name.

	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

Observations of Sheep per Week



We then create a bar chart showing the different number of observations per week at each park.

- To determine the how long it’ll take to complete the study we need to see roughly how many sheep we’d be able to check at each park.
- Since we know that 15% of sheep at Bryce National Park have foot and mouth disease. We also know we’re looking for at least a 5% reduction we can calculate a minimum detectable effect.
- (MDE) minimum detectable effect = 100 * .05 / .15 = 33.3333
- So using the A/B Test Sample calculator. We enter 15% Baseline Conversion Rate, our calculated 33% (MDE), and use the 90% Statistical Significance to get our Sample Size per Variation of 520!
- This leaves us with simple math: Bryce 520/ 250 = 2.08 and Yellowstone 520/507 = 1.026
- So we’ll need about 2 weeks at Bryce and 1 week at Yellowstone.