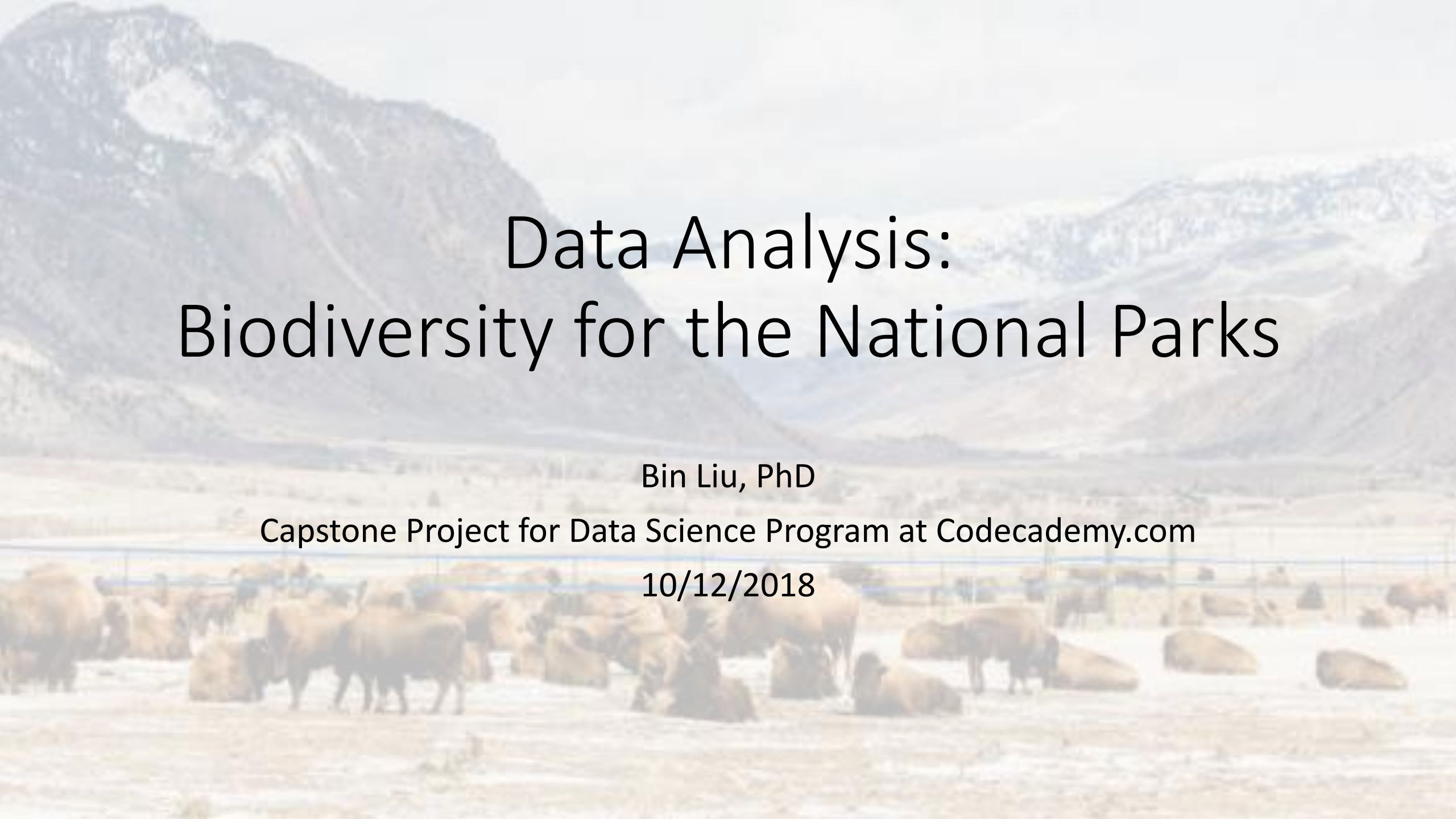


Data Analysis: Biodiversity for the National Parks

Bin Liu, PhD

Capstone Project for Data Science Program at Codecademy.com

10/12/2018

1.1 Description of the Data

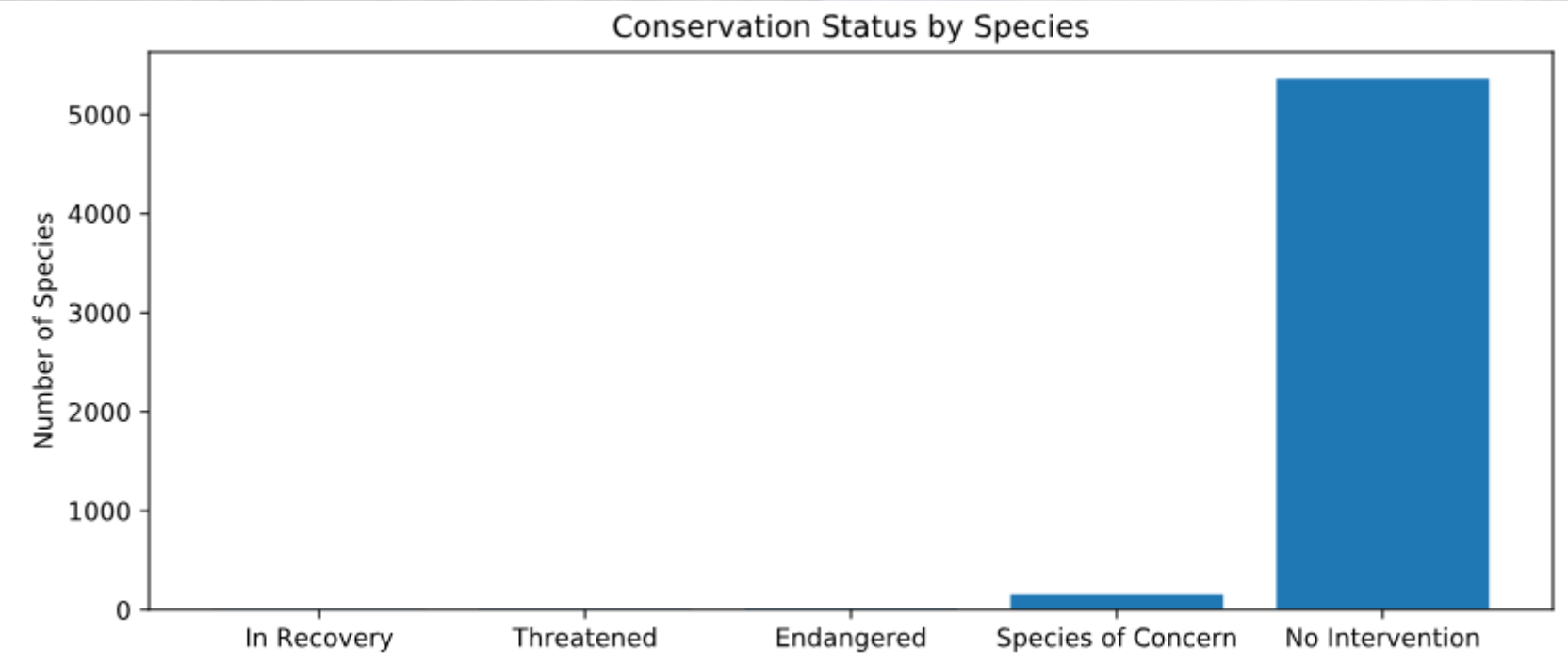
- The first file is the species_info.csv file. The data is from national park service.
- In this file, there are four important items included: category, scientific name, common name, and conservation status.
- We can do some analysis on the endangered species versus the category to investigate if there are any patterns or themes to the types of species that become endangered.
- The second set of files is the conservation.csv and species.csv, including information about species' conservation status by parks.

2.1 Endangered Species between Categories

- There are 5541 species in total. The categories are mammal, bird, reptile, amphibian, fish, vascular plant, and nonvascular plant.
- There are 4 types of conservation statuses including endangered, in recovery, species of concern, and threatened.
- Most of the species are at species of concern, and some of them are at endangered.

2.2 Conservation Status by Species

- First, I check the conservation status by species.
- Base on the pie chart below, most of the species are under no intervention.



2.3 Endangered Species by Type

- I am wondering if certain types of species are more likely to be endangered.
- I calculate the percent of endangered species for each category. The pivot table below shows Mammal and Bird have the highest percentage of endangered species.

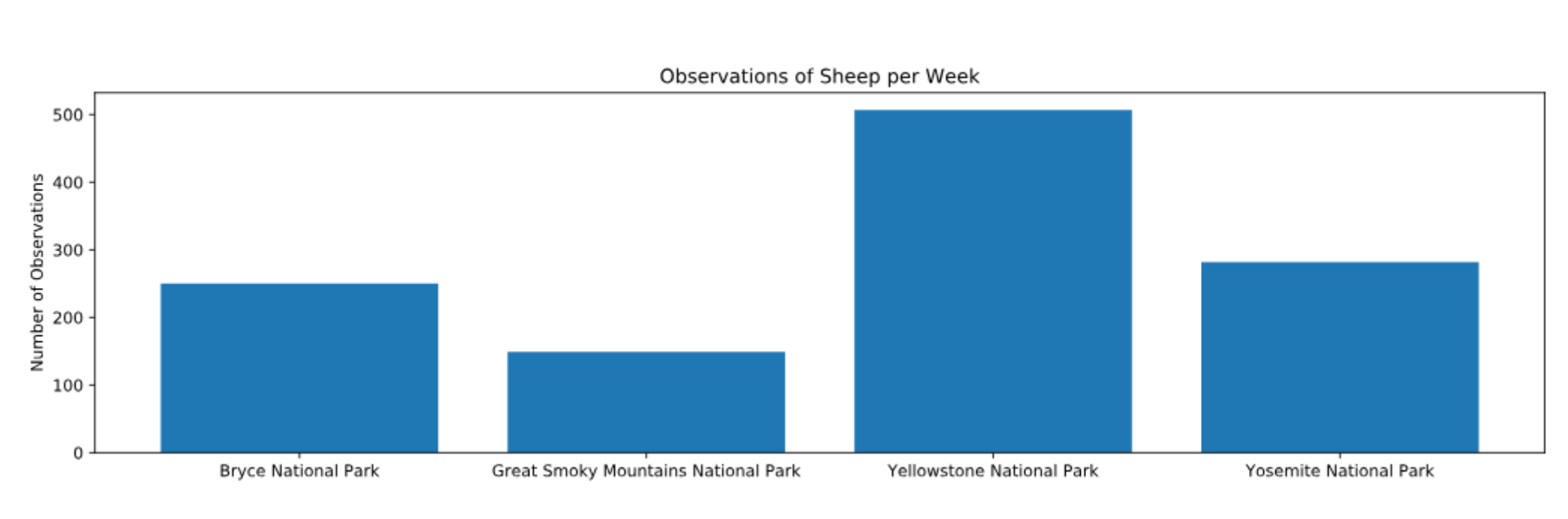
Category	Amphibian	Bird	Fish	Mammal	Nonvascular Plant	Reptile	Vascular Plant
Percent of Species Endangered	8.9%	**15.4%**	8.7%	**17.0%**	1.5%	6.4%	1.1%

2.4 Significant Difference or Not

- Although Mammal and Bird have the highest percentage of endangered species, but I still want to check if there is a significant difference between them.
- I use Chi-Squared Test for the significance. The null hypothesis is that the significance is due to chance.
- After the Chi-Squared test, I get the p value of **0.69**>0.05, which means I can not reject the null hypothesis.
- Therefore, there is no significant difference between Mammal and Bird for their percentage of endangered species.
- However, I repeat the test for Mammal and Reptile and get the p value of **0.038**<0.05, which means there is significant difference between Mammal and Reptile.

3.1 The Sheep in Parks

- I inspected the observations of sheep in each park. The bar chart shows the number of observations per park.



3.2 Sample Size Determination

- I would like to do the analysis for if the rate of foot and mouth disease has been reduced by a special program. The first thing to do is to decide the sample size.
- Last year, 15% of sheep at Bryce National Park have the disease on the record. This will be our **baseline (15).**
- I want to be able to detect reductions of at least 4 percentage points. Therefore, the minimum detectable effect will be **100*5/15 = 33.3**.
- The default level of significance is **90%**.
- Using the online calculator of the sample size determination, the sample size is **870**.

3.3 Time to Get Enough Data

- Base on the calculation, the sample size should be at least **870** to detect if the program was working to reduce the foot and mouth disease in sheep at Bryce National Park.
- The number of observation per week at Bryce National Park is 250. Therefore, it will takes 870/250 = **3.5** weeks to get the enough observations of sheep.
- The number of observation per week at Yellowstone National Park is 507. Therefore, it will takes 870/507 = **1.7** weeks to get the enough observations of sheep.

4. Conclusions

- Certain types of species are more likely to be endangered. The top tow categories are **mammal** and **bird**.
- To do further analysis of the foot and mouth disease reduction program on sheep, a minimum sample size of **870** is needed to detect the significant difference. It may take weeks to get enough observations data for parks. Yellowstone National Park, which has the largest number of observations per week, is the one that scientist can get enough data with least time.

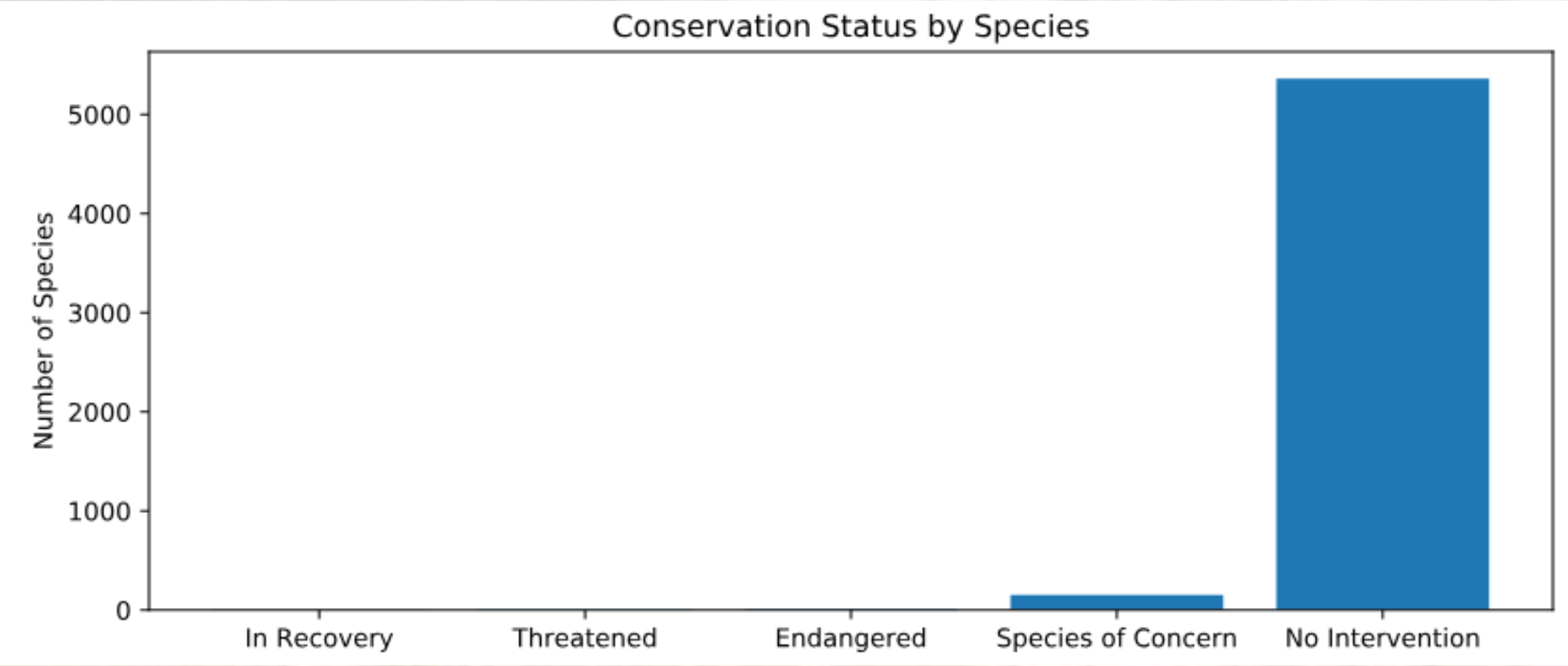
5. Recommendations

- Since certain types of species are more likely to be endangered, scientist need to put more attention and more funding on these categories, such as **mammal** and **birds**.
- **Yellowstone National Park** is the best place to do the observation of foot and mouth disease of sheep, since it has the largest observation per week and allow the scientists to get enough data soon.

6. Sample Size Determination

- Baseline is **15**. (Based on last year's record, 15% of the sheep got disease.)
- Minimum detective effect is **33.3**.(It requires at least 5 percent point of reduction based on a 15% of rate, so 100 * 5 / 15 = 33.3)
- Level of significance is **90%**.
- Using sample size determination calculator, the sample size is **870**. (Baseline 15, minimum detective effect 33.3, level of significance 90%)

7.1. Graphs – conservation status by species



7.2. Graphs- observations of sheep per week