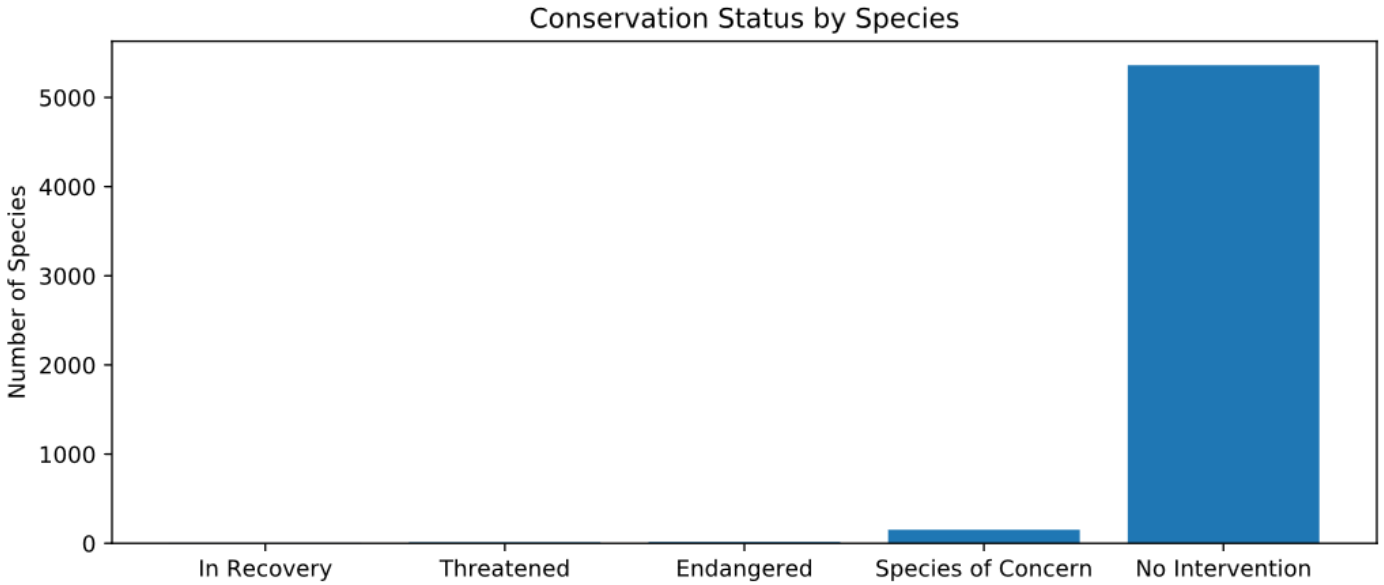
Biodiversity for the National Parks

Capstone Option 2 for Introduction to Data Analysis

By Ronald Reyes

At first inspection, species_info.csv only seems to list entries with their category, scientific and common names, and what kind of protection status they have (if any).

However, with a few lines of code, we can see how many species are in each category. In addition, by checking which rows had a null value in their protection status, we could see what percentage of each species had a form of protection status.



Attempting to plot how many species are in each conservation status doesn't easily show us how many are in recovery, threatened, or endangered due to how skewed the graph is. Perhaps the sheer numbers of species that don't require intervention is a minor comfort in its own right.

Significance testing

Performing significance tests would allow us to see if one category of species is more likely to have a protection status than others. Because we would be comparing multiple categories, a Chi Square test should be most appropriate.

Although the percentage of protected mammals was slightly higher than the percentage of protected birds, the probability value of .68 that the Chi Square test returned was too high to say that this wasn't due to luck.

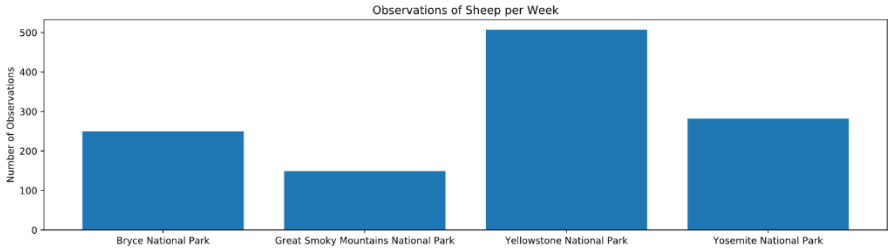
However, the probability value of .03 returned for Mammals and Reptiles allows us to comfortably say that one category of species is more likely to require protection than at least one other category.

In that regard, if a concerned conservationist were to closely examine the conditions of one or two categories of species, it would probably be most prudent to observe mammals and birds.

Sample Size Determination for Disease Study

Given a baseline of 15% for sheep that had foot and mouth disease, we could determine a sample size large enough to confidently say that that percentage had changed by at least 5%. Rather, our minimum detectable effect was 33.3%.

Using a sample size calculator, it would seem that you would need to observe 870 sheep to ensure the percentages have a 90% level of significance.



As each national park has a differing amount of sheep observations, the amount of time it would take for each park to complete the measurements would also differ. Yellowstone could finish its observations in a mere 1.7 weeks, while the Great Smoky Mountains National Park would need up to 5.8 weeks to complete its measurements.