Biodiversity Capstone project

Describing the Data

Species dataframe contains data about the category, scientific_name, common_name and conservation_status of wildlife.

Species_count dataframe then contains a list of the number of species in the various conservation_status groups. There were 5541 species.

Species_type dataframe lists the different categories of species. Theses were mammal, bird, reptile, amphibian, fish, vascular plant and nonvascular plant.

Conservation_statuses dataframe lists the 4 different conservation status names of endangered, in recovery, species of concern and threatened.

Conservation_counts dataframe listed these conservation statuses and the number of species that fit into each group.

Looking at the data it is apparent that the vast majority of species do not require any conservation intervention, some are of concern and only a handful are endangered, threatened or in recovery.

The graph of this data is attached separately (could only save graph as a webpage!?)

Species dataframe was then added to with a new is_protected column with simple True or False entries for each species.

Category_counts dataframe showed each category of species and how many of each category was either protected or not protected.

Category_pivot reorganised this as a easier to read pivot table.

Data shows that the birds and mammals categories had the highest percentage of species that needed conservation intervention.

Running a couple of chi-squared tests showed that the difference between the percentage of mammals at risk, against the percentage of birds at risk wasn’t significant and could be put down to chance. The pvalue was ~0.688

However the difference between reptiles and mammals was very significant. Mammals were far more at risk than reptiles. The pvalue here was ~0.038

Observations dataframe showed the number of sightings of particular species (scientific_name) in a variety of National Parks.

A new column was then added to the species dataframe to highlight whether a species belonged to the ‘sheep’ family or not.

Species_is_sheep dataframe then contained ONLY the species that contained the word ‘sheep’ in their description.

Sheep_species dataframe however, filtered this further to only leave species that contained the word ‘sheep’ in their description and that were also ‘mammals’.

Sheep_observations dataframe contained data for how many sightings of each species of sheep had been recorded over a week across the 4 national parks.

Obs_by_park dataframe then displayed all the cumulative sightings across the 4 national parks.

The graph of this data is attached separately (could only save graph as a webpage!?)

Foot and Mouth reduction effort

Baseline = 15%

Minimum_detectable_effect = 100 * 5 / 15 = 33%

Statistical Significance = 90%

Sample Size = 890 (via the Codecademy site)

However the sample size changed to 520 when plugging in the same values and using the also recommended separate webpage https://www.optimizely.com/sample-size-calculator/?conversion=15&effect=33&significance=90 Quite a marked difference which would eschew my recommendations accordingly!

So….using the Codecademy sample size calculation:

Yellowstone_weeks_observing = 890/507 = ~2 wks (just less than 2 weeks but much more than 1!)

Bryce_weeks_observing = 890/250 = ~3.5 wks

However….using the optimizely website sample size calculation:

Yellowstone_weeks_observing = 520/507 = ~1 wk

Bryce_weeks_observing = 520/250 = 2 wks

So to conclude to see a >5% drop in foot and mouth disease and make sure this was significant at least 890 (or 520) sheep need to be observed. This would take approximately 2 wks (or 1 wk) in Yellowstone and 3.5 wks (2 wks) in Bryce.