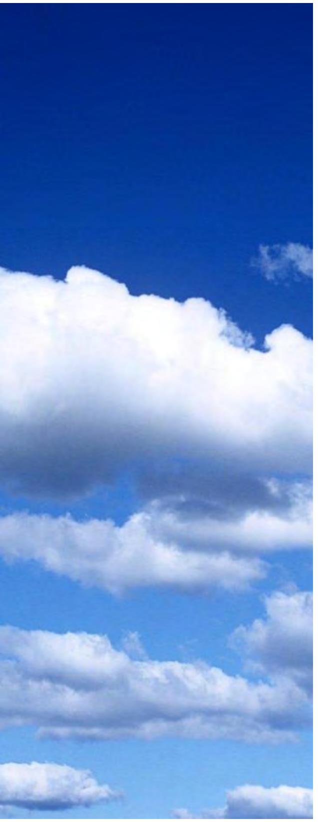


Biodiversity for the National Parks

Python Project submitted by Alan Sparks

Objectives

- Determine if there are any patterns or themes to the types of species that become endangered using field data made available in species_info.csv.

- Assist the scientific community in analyzing patterns of sheep movements in 4 national parks.

- Determine the validity of a park program to reduce in the instance of foot and mouth disease, where the objective was a reduction of 5 percentage points, from the 15% level observed last year. We will provide the field team with the number of observations required to achieve a level of significance of 90%.

Reconciling the Data in species_info

- Field data contained 5,824 entries
- Prior to analysis, data set examination revealed the following
  - 5,541 were unique, leaving 283 as unexplained

Field Data in species_info	Field Data Reconciled
5,267 truly unique entries	5,267
530 entries were found to be duplicates	530 / 2 = 265
27 entries were triplicates	27 / 3 = 9
Total raw data = 5,824	Total unique species = 5,541

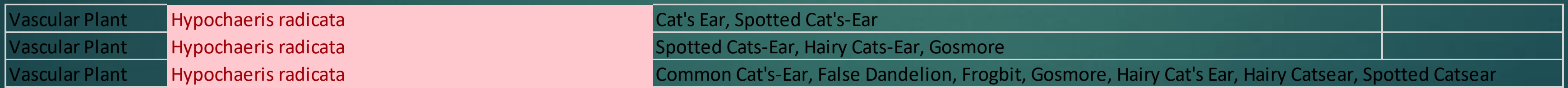
Duplicate and Triplicate Field Entries

- Of the 7 categories in the field study, 6 were affected by multiple entries
- Since our study focused only on mammals, our Python readings were potentially inflated:
  - Taxidea taxus, aka badger – duplicate entry and species of concern
  - Myotis lucifugus, aka little brown bat – triplicate entry and species of concern
  - Myotis californicus, aka presumably a California bat – duplicate entry and species of concern
  - Lasionycteris noctivagans, aka silver-haired bat – duplicate entry and species of concern
  - Eptesicus fuscus, aka big brown bat – duplicate entry and species of concern
  - Canis lupus, aka gray wolf – triplicate entry listed twice as endangered and once as in recovery.
- In total, our subsequent findings for species of concern may be inflated by 6, and endangered by 1.

Duplicate and Triplicate Field Entries

- How did this happen?

- Our conclusion is that the data entry process was poorly handled. Entries under the common_name column appear to drive the problem.
- For example, the two entries for taxidea taxus listed under common_names were:
  - American Badger, Badger
  - Badger
- A more extreme example taken directly from the field entries:

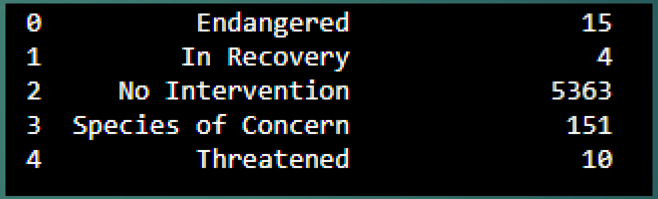


Vascular Plant	Hypochaeris radicata	Cat's Ear, Spotted Cat's-Ear	
Vascular Plant	Hypochaeris radicata	Spotted Cats-Ear, Hairy Cats-Ear, Gosmore	
Vascular Plant	Hypochaeris radicata	Common Cat's-Ear, False Dandelion, Frogbit, Gosmore, Hairy Cat's Ear, Hairy Catsear, Spotted Catsear	

(It looks like your basic garden variety dandelion, but its not. Where the field staff got Frogbit from, we aren't sure, but we digress. Back to our data analysis…)

Analyzing the Data Set in species_info

- Of the 5,824 entries
  - 80 are amphibians
  - 521 are birds
  - 127 are fish
  - 214 are mammals
  - 79 are reptiles
  - 333 are non-vascular plants
  - 4,470 are vascular plants

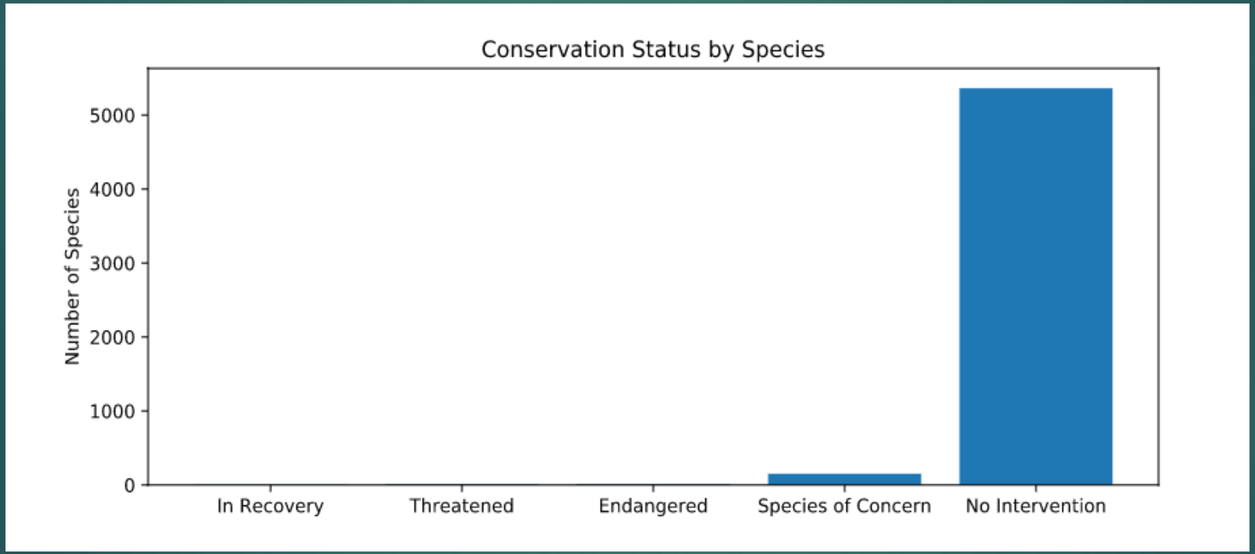


```
0          Endangered        15
1         In Recovery         4
2     No Intervention      5363
3  Species of Concern       151
4          Threatened        10
```

- Of this total, our analysis shows 5, 541 unique entries, effectively eliminating duplicates.
- This analysis focuses on mammals, and in particular mammals that are at some state of risk. The read out above shows the breakdown:

Analyzing the Data Set in species_info

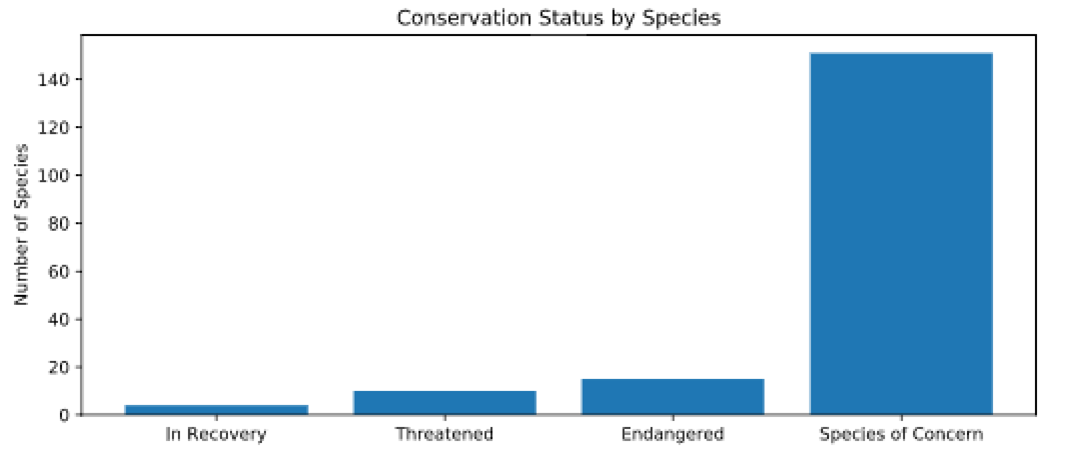
- To help visualize our data set, we use the following two charts.
- All data:



- We did not find this chart particularly useful, as the vast majority of our field data is classed as "No Intervention," meaning that it does not require any form of government protection.

Analyzing the Data Set in species_info

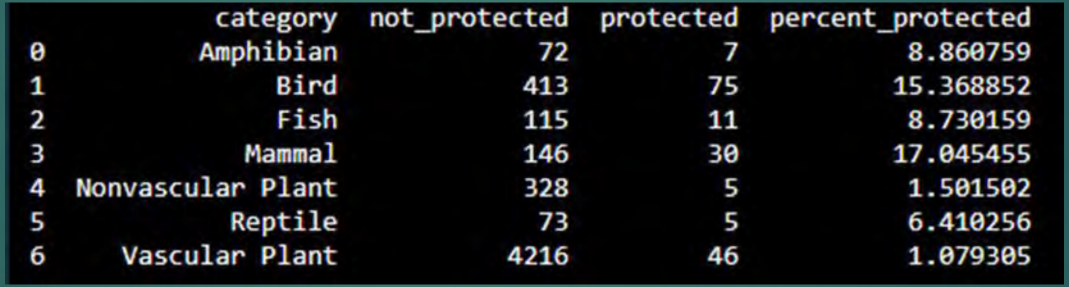
- Eliminating the No Intervention classification reveals important information.



- Only a small percentage of species of concern migrate to the more critical endangered and threatened stages, both precursors to extinction.
- At least some species recover sufficiently to be re-classed as in recovery.

Analyzing the Data Set in species_info

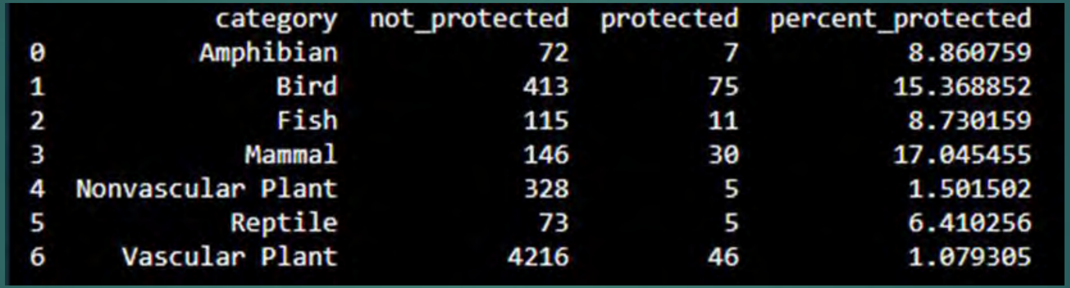
- Further insight of this select data demonstrates the level of concern broken down by category:



```
            category  not_protected  protected  percent_protected
0          Amphibian             72          7           8.860759
1               Bird            413         75          15.368852
2               Fish            115         11           8.730159
3             Mammal            146         30          17.045455
4  Nonvascular Plant            328          5           1.501502
5            Reptile             73          5           6.410256
6     Vascular Plant           4216         46           1.079305
```

- We look closely at the species having some sort of protection and see that mammals and birds are almost twice as likely as other species, and that plants are the least likely.
- Before presenting our findings to the corporate office, we were asked to analyze the differences in data, knowing that several members of corporate are skilled analysts themselves.

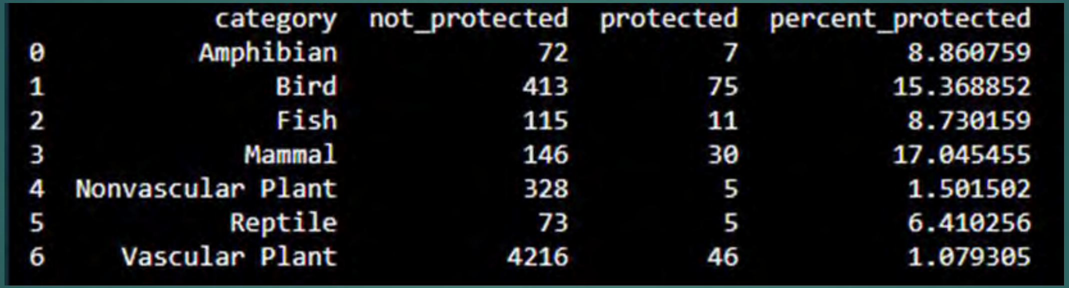
Analyzing the Data Set in species_info



```
             category  not_protected  protected  percent_protected
0           Amphibian             72          7           8.860759
1                Bird            413         75          15.368852
2                Fish            115         11           8.730159
3              Mammal            146         30          17.045455
4   Nonvascular Plant            328          5           1.501502
5             Reptile             73          5           6.410256
6      Vascular Plant           4216         46           1.079305
```

- First, we looked at the relationship between birds and mammals. Mammals have a higher propensity to become at risk than birds. We wanted to understand if this difference was significant, or due to chance.
- Using Chi-Square Testing, we set the null hypothesis: The difference is due to chance.
- The result is 0.688 (68.8%), far higher than the required 5% threshold. Therefore, the null hypothesis must be accepted. The field data differences between the risk factors of birds and mammals is statistically insignificant and is due to chance.

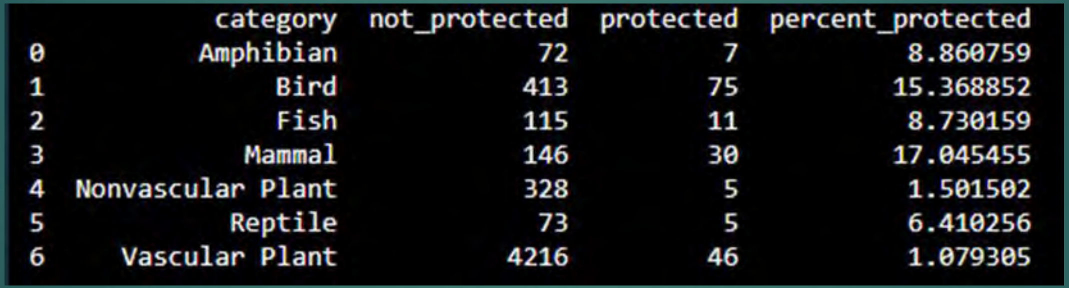
Analyzing the Data Set in species_info



```
            category  not_protected  protected  percent_protected
0          Amphibian             72          7           8.860759
1               Bird            413         75          15.368852
2               Fish            115         11           8.730159
3             Mammal            146         30          17.045455
4  Nonvascular Plant            328          5           1.501502
5            Reptile             73          5           6.410256
6     Vascular Plant           4216         46           1.079305
```

- We next tested the relationship between reptiles and mammals. Again, mammals have a higher propensity to become at risk than reptiles, this time by a factor of more than 2 ½ x. It would seem that in this case, the difference is NOT due to chance.
- Using Chi-Square Testing, we again set the null hypothesis: The difference is due to chance.
- The result is 0.038 (3.8%), well below the required 5% threshold. Therefore, the null hypothesis must be rejected. The field data differences between the risk factors of birds and mammals is statistically significant and is not due to chance. This also aligns with the observed 2 ½ x difference.

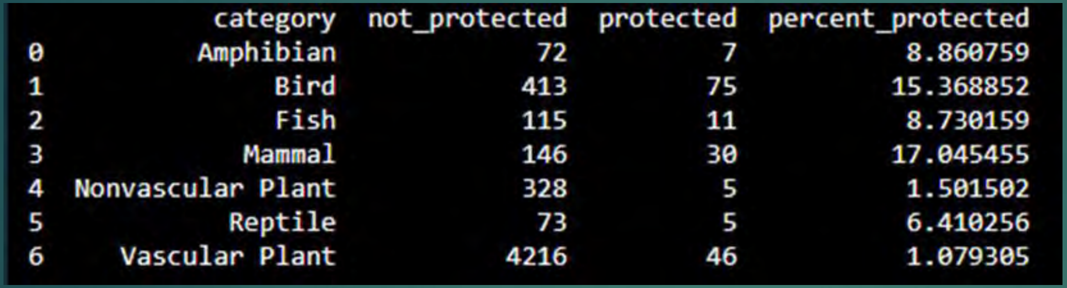
Analyzing the Data Set in species_info



```
             category  not_protected  protected  percent_protected
0           Amphibian             72          7           8.860759
1                Bird            413         75          15.368852
2                Fish            115         11           8.730159
3              Mammal            146         30          17.045455
4  Nonvascular Plant            328          5           1.501502
5             Reptile             73          5           6.410256
6      Vascular Plant           4216         46           1.079305
```

- Further testing between species gave similar results
  - Reptiles and fish showed no significant differences. The null hypothesis was not rejected. (0.741)
  - Vascular and non-vascular plants also showed no significant differences (0.662)
  - The null hypothesis between plants and reptiles was rejected (0.034). The differences in this case are not due to chance.

Summarizing species_info and Recommendation



```
             category  not_protected  protected  percent_protected
0           Amphibian             72          7           8.860759
1                Bird            413         75          15.368852
2                Fish            115         11           8.730159
3              Mammal            146         30          17.045455
4  Nonvascular Plant            328          5           1.501502
5             Reptile             73          5           6.410256
6      Vascular Plant           4216         46           1.079305
```

- Restate our objective: Determine if there are any patterns or themes to the types of species that become endangered
- Answer and Recommnedation:
  - Yes. Mammals and birds have the highest likelihood of risk, followed generally by amphibians, fish, and reptiles as a group. Plants are the least likely to be susceptible to risk.
  - Recommendation: Concentrate available resources in the first group (mammals and birds. Use remaining resources on the middle group. Do not spend any significant time or money on plants.

Intermission

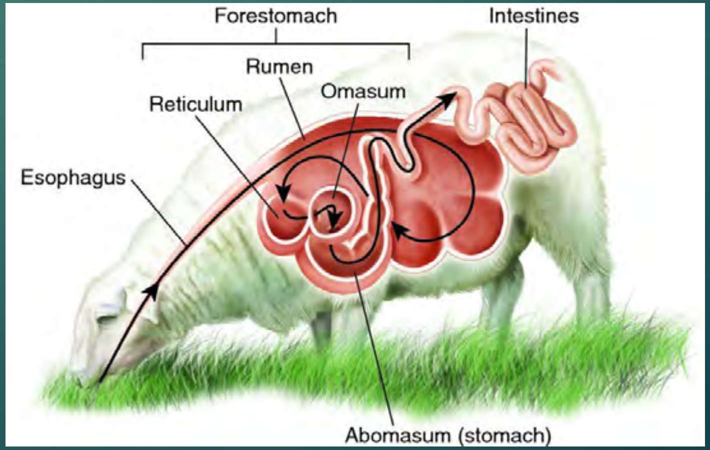
As background, please consider the noun *ruminant*. An unusual word, not often used in the company of civilized society.

- An even-toed ungulate mammal that chews the cud regurgitated from its rumen. The ruminants comprise the cattle, sheep, antelopes, deer, giraffes, and their relatives.

  - Charming. Who knew? Sign up for Python and learn about ruminants.

  - Apparently, there are a breed of scientists that are enthused about ruminants. They don't often get invited to parties.

  - Here's a working diagram.

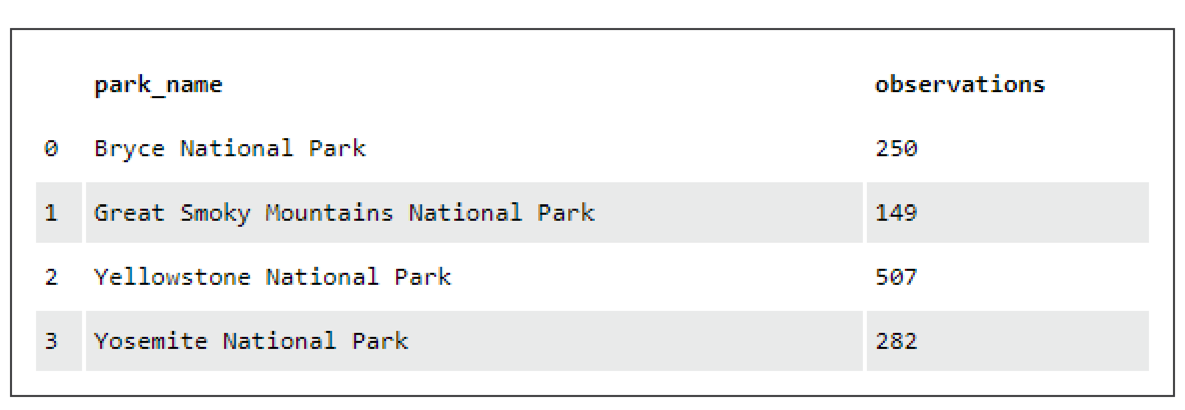


Augmenting species_info with Sheep Field Information File; observations

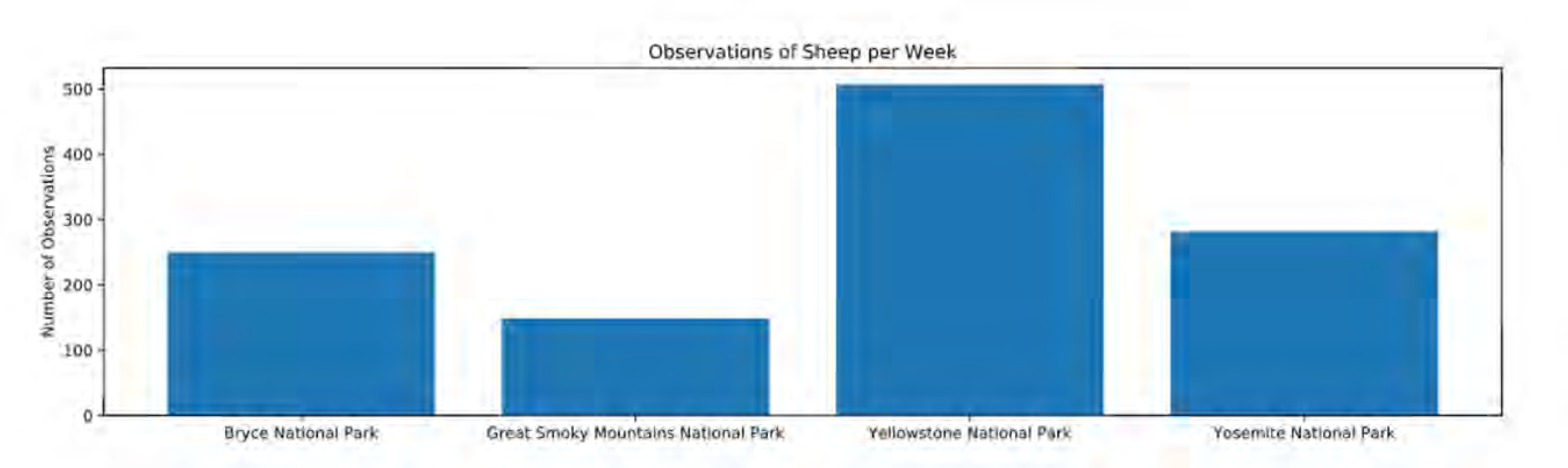
- The field data on sheep contains limited data
  - Scientific name only of thousands of various species
  - Park where observations took place
  - Number of sightings for each species, by park

- Analysis required the following steps to be performed on the field data in the original species_info file in order to:
  - Isolate common sheep names from non-sheep names and saved to a new file entitled ‘species_is_sheep
  - Further, that plant names that contain sheep, ie, Fineleaf Sheep Fescue, be isolated from mammal names containing the word sheep, and then saved to a file entitled ‘sheep_species’
  - That the modified sheep_species file then be merged with the observations file, then sorted by park name in order to count the number of sheep spotted in each park

Results of the Analytics on Sheep Sightings, by Park



	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

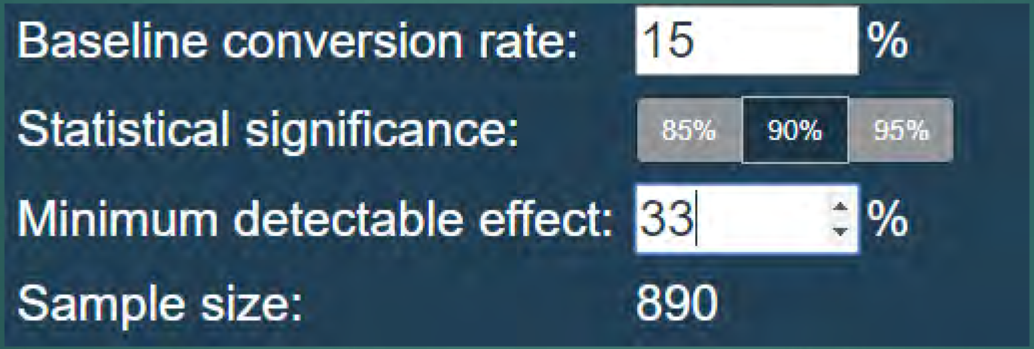


Foot and Mouth Disease Testing

- To assist the field team, a Python online sample size calculator was utilized.
- The calculator is a relatively simple tool that outputs a sample size result given a user adjustable confidence level. The tool takes 3 inputs:
  - A baseline conversion rate: This is the reference point at which a starting baseline is set. In this case, the 15% is the field measurement that was observed last year in one of the parks. The field team is hoping its program will improve this years results by 5 percentage points.
  - Statistical significance: This is our confidence level that the sample size is enough to give us a level of certainty that we are achieving the desired results. Usually set at 85%, 90%, or 95%. The higher the statistical significance, the more samples are needed.
  - The minimum detectable effect: The desired percent of improvement, also known as "lift." In our case, it is 5% divided by the baseline conversion rate, or 33%.

Foot and Mouth Disease Testing

- The tool yields a sample size of 890, meaning that the field team will need to observe and test 890 sheep to be 90% certain that their program is working.



Baseline conversion rate: 15 %

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

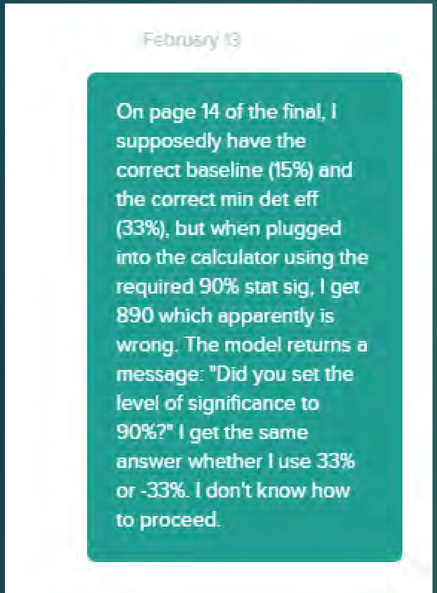
Minimum detectable effect: 33 %

Sample size: 890

Foot and Mouth Disease Testing: Results

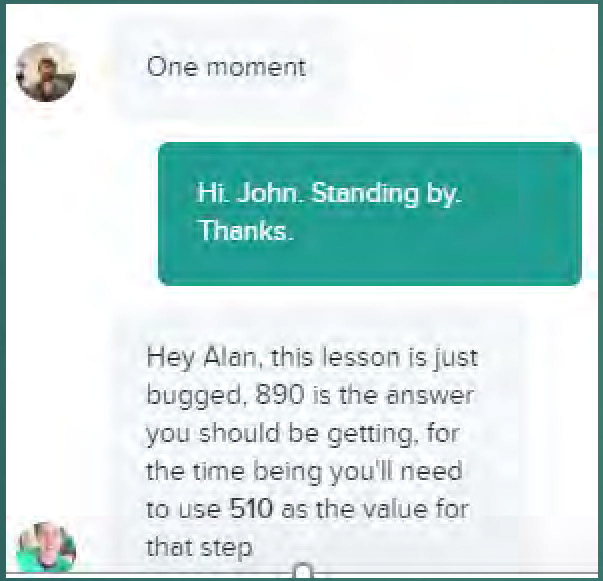
- The results yielded a sample size of 890.
- Sadly, the field team found this output unacceptable. The Python sample size calculator was scorned, mocked, and shouted at. One scientist even yelled ni and he threw a herring at me.
- Data Analyst Sparks was forced to call the NPS corporate offices to report this unhappy situation.
- Following is a rough transcript of the difficult conversation.

Foot and Mouth Disease Testing: Lesson Bugged



February 13

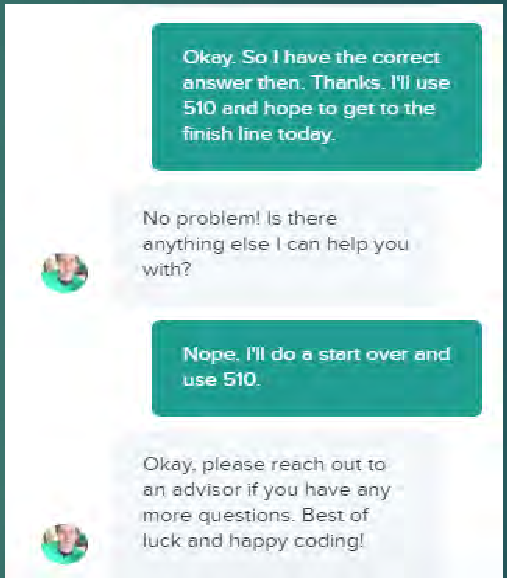
On page 14 of the final, I supposedly have the correct baseline (15%) and the correct min det eff (33%), but when plugged into the calculator using the required 90% stat sig, I get 890 which apparently is wrong. The model returns a message: "Did you set the level of significance to 90%?" I get the same answer whether I use 33% or -33%. I don't know how to proceed.



One moment

Hi. John. Standing by. Thanks.

Hey Alan, this lesson is just bugged, 890 is the answer you should be getting, for the time being you'll need to use 510 as the value for that step



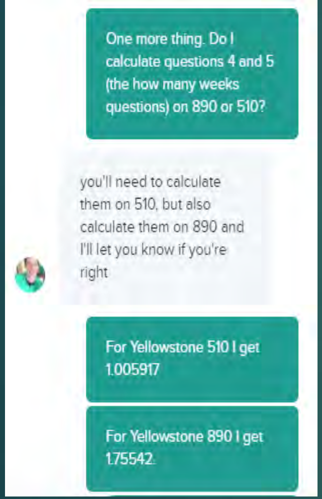
Okay. So I have the correct answer then. Thanks. I'll use 510 and hope to get to the finish line today.

No problem! Is there anything else I can help you with?

Nope. I'll do a start over and use 510.

Okay, please reach out to an advisor if you have any more questions. Best of luck and happy coding!

Foot and Mouth Disease Testing: Lesson Bugged

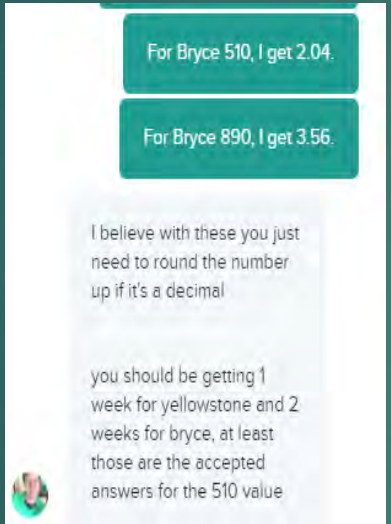


One more thing. Do I calculate questions 4 and 5 (the how many weeks questions) on 890 or 510?

you'll need to calculate them on 510, but also calculate them on 890 and I'll let you know if you're right

For Yellowstone 510 I get 1.005917

For Yellowstone 890 I get 1.75542.

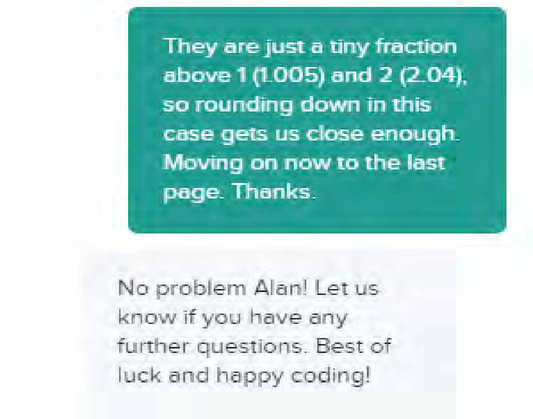


For Bryce 510, I get 2.04.

For Bryce 890, I get 3.56.

I believe with these you just need to round the number up if it's a decimal

you should be getting 1 week for yellowstone and 2 weeks for bryce, at least those are the accepted answers for the 510 value



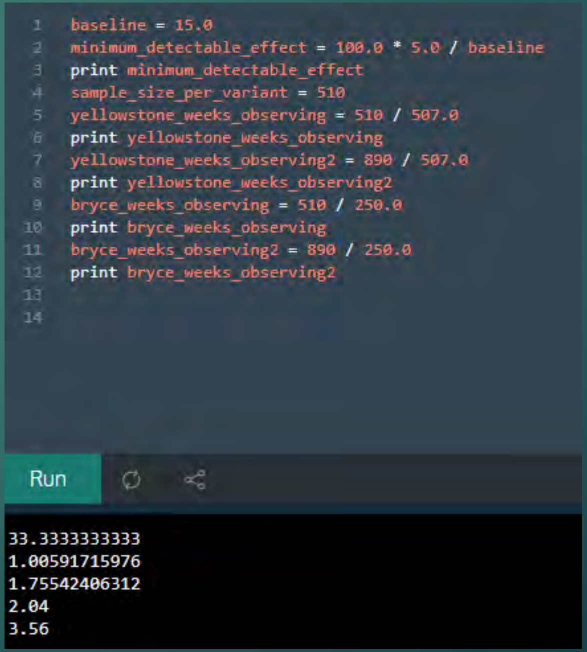
They are just a tiny fraction above 1 (1.005) and 2 (2.04), so rounding down in this case gets us close enough. Moving on now to the last page. Thanks.

No problem Alan! Let us know if you have any further questions. Best of luck and happy coding!

- The field team were pleased with the corporate response. They only have to find 510 sheep in each park now.
- This concludes the great bugged ruminant project of 2018. The multiple outputs follow on the next slide.

Foot and Mouth Disease Testing: Lesson Bugged

- This project required multiple outputs in order to reach conclusion. The output demonstrates the simple calculation to determine the minimum detectable effect, followed by the length, in weeks, of the corporate forced sample size for each park, and then the larger calculated sample size for each park.



```
baseline = 15.0
minimum_detectable_effect = 100.0 * 5.0 / baseline
print minimum_detectable_effect
sample_size_per_variant = 510
yellowstone_weeks_observing = 510 / 507.0
print yellowstone_weeks_observing
yellowstone_weeks_observing2 = 890 / 507.0
print yellowstone_weeks_observing2
bryce_weeks_observing = 510 / 250.0
print bryce_weeks_observing
bryce_weeks_observing2 = 890 / 250.0
print bryce_weeks_observing2
```

```
33.3333333333
1.00591715976
1.75542406312
2.04
3.56
```