CAPSTONE TWO

BIODIVERISTY - INVESTIGATING PROTECTED SPECIES

Student: Martin Munkholt Andersen

SECTION1: DESCRIPTION OF DATA: SPECIES_INFO.CSV

1) First step first, what sort of data is stored in species_info.csv? I printed .head() which returns these five rows

	category	scientific_name	common_names	conservation_status
0	Mammal	Clethrionomys gapperi gapperi	Gapper's Red-Backed Vole	nan
1	Mammal	Bos bison	American Bison, Bison	nan
2	Mammal	Bos taurus	Aurochs, Aurochs, Domestic Cattle (Feral), Domesticated Cattle	nan
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	nan
4	Mammal	Cervus elaphus	Wapiti Or Elk	nan

2) Findings include a species categorisation, two types of naming and a special attribute, conservation_status. Lets see what is stored there.

```
[nan 'Species of Concern' 'Endangered' 'Threatened'
'In Recovery']
  conservation_status  scientific_name
0          Endangered               15
1         In Recovery                4
2  Species of Concern              151
3          Threatened               10
```

3) We then analyse conservation_status to see what it contains. It contains what is shown in the table above, NaN is disregarded, which is not ideal.

SECTION1:
DESCRIPTION OF DATA: SPECIES_INFO.CSV

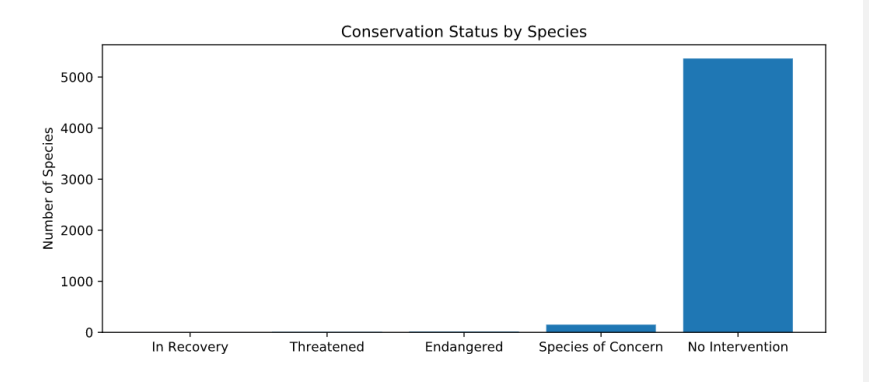
4) The operation before neglects NaN values, but we need a count of these for further analyses. So we fill NaN.

```
28  #Step 5: Fill null/NaN with string
29  species.fillna('No Intervention', inplace = True)
30
31  conservation_counts_fixed =
    species.groupby('conservation_status').scientific_na
    me.nunique().reset_index()
32  #print conservation_counts_fixed
```

5) Which returns a more useful table for doing statistics:

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

6) From experience people know better how to decode data, when it is well visualised, so here it is plottet with Matplotlib, which would work great for a presentation of key findings:



SECTION2: SIGNIFICANCE CALCULATIONS ON ENDANGERED SPECIES

1) First step was to add a new column and fill it with 'protected' or 'not protected'. This was done like this:

```
species['is_protected'] =
species.conservation_status != 'No Intervention'

category_counts = species.groupby(['category',
'is_protected']).scientific_name.nunique().reset_ind
ex()
```

2) Then in order to create a better visually looking table a np.pivot() was imployed on category_counts. Headers for the new columns was created and percent for each category was calculated:

```
13  category_pivot =
    category_counts.pivot(columns='is_protected',
14                        index='category',
15                        values='scientific_name')\
16                        .reset_index()
17
18  #This will rename columns in the order from one to
    three
19  category_pivot.columns = ['category',
    'not_protected', 'protected']
20
21  #creates new column in dataframe calculates
    percentage and passes theses to values
22  category_pivot['percent_protected'] =
    category_pivot.protected / (category_pivot.protected
    + category_pivot.not_protected)
```

3) Which results in the following nice looking table:

	category	not_protected	protected	percent_protected
0	Amphibian	72	7	0.088608
1	Bird	413	75	0.153689
2	Fish	115	11	0.087302
3	Mammal	146	30	0.170455
4	Nonvascular Plant	328	5	0.015015
5	Reptile	73	5	0.064103
6	Vascular Plant	4216	46	0.010793

0.687594809666

SECTION2: SIGNIFICANCE CALCULATIONS ON ENDANGERED SPECIES

	category	not_protected	protected	percent_protected
0	Amphibian	72	7	0.088608
1	Bird	413	75	0.153689
2	Fish	115	11	0.087302
3	Mammal	146	30	0.170455
4	Nonvascular Plant	328	5	0.015015
5	Reptile	73	5	0.064103
6	Vascular Plant	4216	46	0.010793

Chi-Squared test

4) Based of the pivot table described on the previous slide we calculate the statistical certainty with which we can say, one type of specie is more endagered than others. We use the values from 'not protected' and 'protected' to fill a table, contingency, with two rows for 'Mammal' and 'Bird' to calculate the p-value. It is repeated in contingenty_reptile_mammal for 'Mammal' and 'Reptile'.

- We can use a chi-squared test to check if we can be sure that some species are more likely of being endangered than others.
- We found that with a p-value of 0.69 (above 5%) mammals are

not significantly more likely to be endangered than birds, but reptiles are with a p-val of 0.03.

Conclusions on next slide

```
#Create table, fill with values
contingency = [[30, 146],
               [75, 413]]

#find p-value with chi2 test on the new table
pval = chi2_contingency(contingency)[1]
print(pval)
# No significant difference because pval > 0.05

contingency_reptile_mammal = [[30, 146],
                              [5, 73]]

pval_reptile_mammal =
chi2_contingency(contingency_reptile_mammal)[1]
print(pval_reptile_mammal)
# Significant difference! pval_reptile_mammal < 0.05
```

SECTION2: SIGNIFICANCE CALCULATIONS ON ENDANGERED SPECIES

- What statistical significance of endagered species did the analysis show?
- Data showed a slight difference between endangered status for birds and mammal. When we ran Chi2 test to find if this was a statistically significant difference it appeared to not be the case. The P-value was above 0.5.
- When we did the same test for mammals vs. Reptiles, the result showed an overwhelming statiscical significance. In conclusion, reptiles are more likely to be endangered than mammals, but mammals are not more like to be endangered than birds are.

Recommendation based on endangered tests

- It is recommended to further analyse the dataset if some types of species are more likely to be in need of protection than others. My initial research show reptiles are a candidate for a hightened conservation effort. But amphibians and others will need testing as well.
- It would be wise to identify which individual animals based on scientific_name are most likely to be endangered. It is likely that there is a spread within each species category.

SECTION 3:
SAMPLE SIZE DETERMINATION

Using the sightings table how do we use sample determination to calculate how many weeks we will need to observe a significant reduction.

We know the baseline in Bryce to be 15% and want 90% significance.

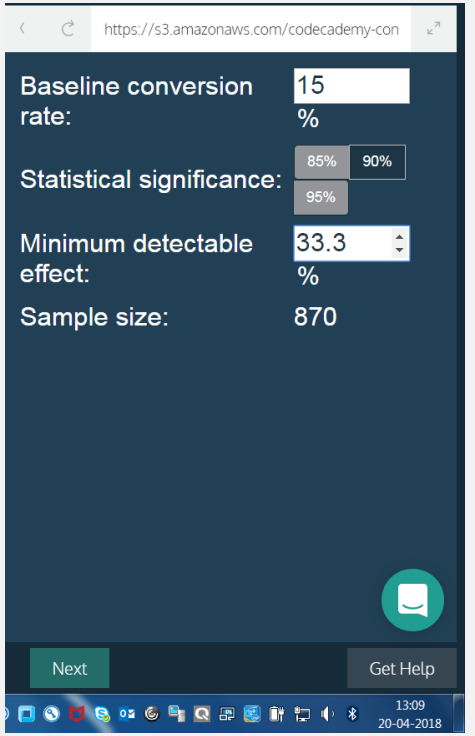
```
#The only information that the scientists currently have is that last year it was recorded that 15% of sheep at Bryce National Park have foot and mouth disease.
baseline = 15

#Minimum Detectable Effect" is a percent of the baseline, so if we wanted to observe an x% change with confidence, our minimum detectable effect would be equal to 100 * x / baseline.
minimum_detectable_effect = 100*5./15
print minimum_detectable_effect

#baseline conversion rate: 15, stat. significance: 90%, min. detectable effect: 33.3,
sample_size_per_variant = 870

#weeks needed to observe sheep at Yellowstone Park
yellowstone_weeks_observing = sample_size_per_variant/507.
print yellowstone_weeks_observing

#weeks needed to observe sheep at Bryce Park
bryce_weeks_observing = sample_size_per_variant/250.
print bryce_weeks_observing
```

Which print these three values to the console:

33 procent of the baseline of 15:
33.33

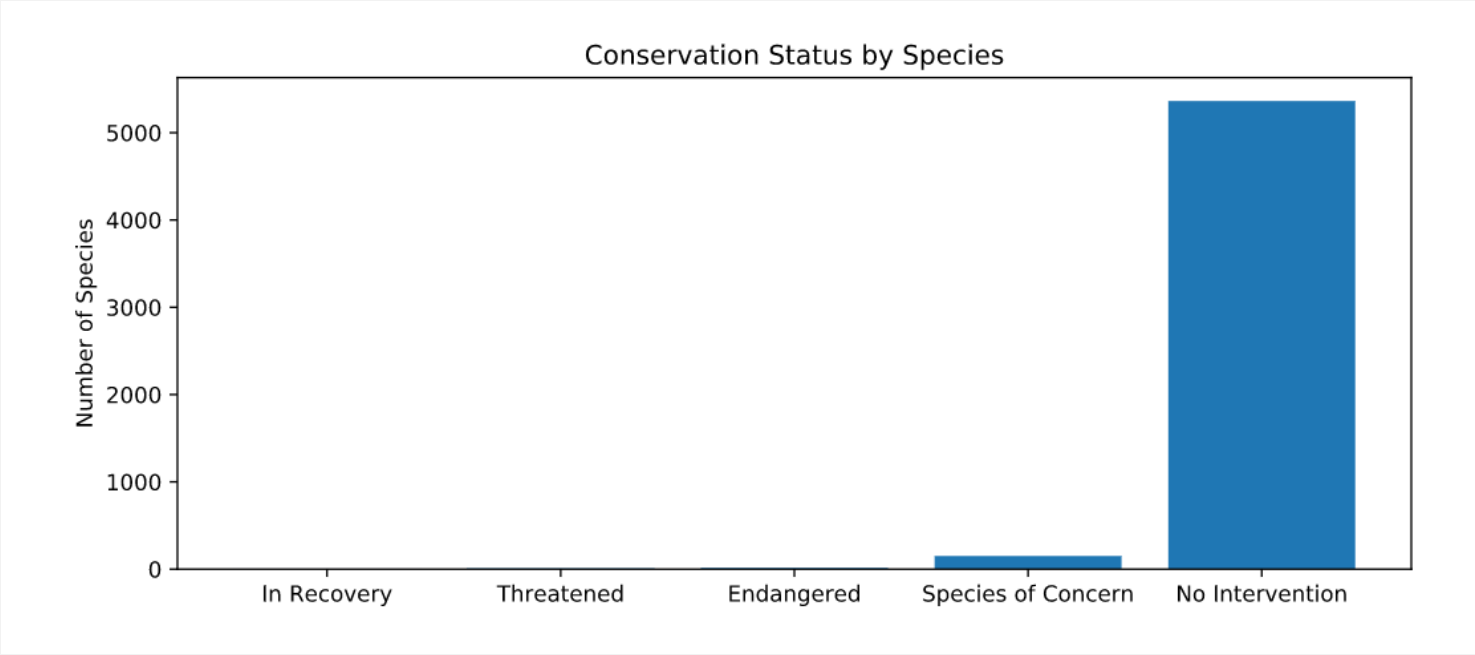
Weeks needed to observe at Yellowstone:
1.72

Weeks needed to observe at Bryce National:
3.48



ALL GRAPHS – IN ORDER OF OCCURENCE (2 GRAPHS)

LEASON 5: PLOTTING CONSERVATION STATUS BY SPECIES



LEASON 10: PLOTTING SHEEP SIGHTINGS