Jorge Moura Sampaio

Introduction to Data Analysis

Biodiversity Capstone project - INVESTIGATING PROTECTED SPECIES

Initial analysis of 'Species.csv'

CSV file species_info.csv contains data about different species in several National Parks, including:

- The scientific name of each species
- The common names of each species
- The species conservation status

1 – Start by importing packages

```
import codecademylib
import pandas as pd
from matplotlib import pyplot as plt
```

2 – How many different species are in the species DataFrame?

```
species = pd.read_csv('species_info.csv')
species_count = species.scientific_name.nunique()
print species_count
```

>> 5541

3 – What is the conservation status of the species?

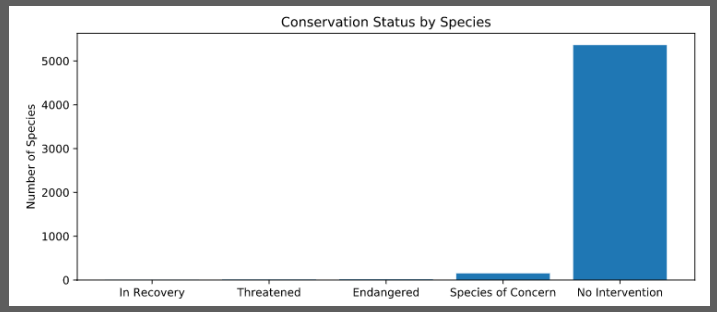
```
conservation_counts =
species.groupby('conservation_status').scientific_name.nunique().reset_index()

species.fillna('No Intervention', inplace = True)

conservation_counts_fixed =
species.groupby('conservation_status').scientific_name.nunique().reset_index()

print conservation_counts_fixed
```

>>

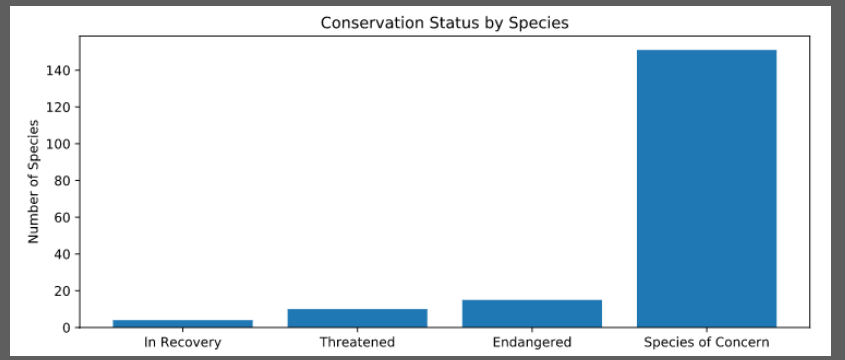
	conservation_status	scientific_name
0	Endangered	15
1	In Recovery	4
2	No Intervention	5363
3	Species of Concern	151
4	Threatened	10

Plotting Conservation Status by Species

```
species = pd.read_csv('species_info.csv')

species.fillna('No Intervention', inplace = True)

protection_counts = species.groupby('conservation_status')\
    .scientific_name.nunique().reset_index()\
    .sort_values(by='scientific_name')

plt.figure(figsize=(10, 4))
ax = plt.subplot()
plt.bar(range(len(protection_counts)),protection_counts.scientific_name.values)
ax.set_xticks(range(len(protection_counts)))
ax.set_xticklabels(protection_counts.conservation_status.values)
plt.ylabel('Number of Species')
plt.title('Conservation Status by Species')
labels = [e.get_text() for e in ax.get_xticklabels()]
plt.show()
```



'No intervention' row can be deleted for the sake of clarity when plotting the data



Are certain types of species more likely to be endangered?

Investigating Endangered Species

```
import codecademylib
import pandas as pd
from matplotlib import pyplot as plt

species = pd.read_csv('species_info.csv')

species.fillna('No Intervention', inplace = True)

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

category_counts = species.groupby(['category',
'is_protected']).scientific_name.nunique().reset_index()

category_pivot = category_counts.pivot(columns='is_protected',
                    index='category',
                    values='scientific_name')\
                    .reset_index()

category_pivot.columns = ['category', 'not_protected', 'protected']


category_pivot['percent_protected'] = category_pivot.protected /
(category_pivot.protected + category_pivot.not_protected)

print category_pivot
```

Birds and Mammals are the category of animals/plants with the highest rate of protected species

	category	not_protected	protected	percent_protected
0	Amphibian	72	7	0.0886075949367
1	Bird	413	75	0.15368852459
2	Fish	115	11	0.0873015873016
3	Mammal	146	30	0.170454545455
4	Nonvascular Plant	328	5	0.015015015015
5	Reptile	73	5	0.0641025641026
6	Vascular Plant	4216	46	0.0107930549038

It looks like Mammals are more likely to be endangered than Birds, but is it a significant difference?

Chi-squared test examples

```
import codecademylib
import pandas as pd
from matplotlib import pyplot as plt
from scipy.stats import chi2_contingency

contingency_bird_mammal = [[30, 146],
        [75, 413]]

pval_bird_mammal =
chi2_contingency(contingency_bird_mammal)[1]
print(pval_bird_mammal)
```

>> 0.68

Birds vs. Mammals

No significant difference because pval > 0.05

Reptiles vs. Mammals

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

pval_reptile_mammal =
chi2_contingency(contingency_reptile_mammal)[1]
print(pval_reptile_mammal)
```

>> 0.0383

Significant difference! pval_reptile_mammal < 0.05

Chi-squared test examples

Vascular vs. Non-Vascular Plants

```
contingency_vascular_nonvascular = [[5, 328],
                                    [46, 4216]]

pval_vascular_nonvascular =
chi2_contingency(contingency_vascular_nonvascular)[1]
print(pval_vascular_nonvascular)
```

>> 0.66

No significant difference because pval > 0.05

Reptiles vs. Anphibians

```
contingency_reptile_anphibian = [[5, 73],
                                 [7, 72]]

pval_reptile_anphibian =
chi2_contingency(contingency_reptile_anphibian)[1]
print(pval_reptile_anphibian)
```

>> 0.78

No significant difference because pval > 0.05

Advice for Conservationists

Both birds and mammals seem to have more probability of being endangered when compared to reptiles. The reason for decline in the population of these two categories may be connected and should be investigated

```
contingency_reptile_birds = [[75, 413],
                             [5, 73]]

pval_reptile_birds =
chi2_contingency(contingency_reptile_birds)[1]
print(pval_reptile_birds)

>> 0.05
```

Significant difference! pval_reptile_birds = 0.05

Analyzing the ‘observation’ and ‘species’ DataFrames to help track sheep locations.

The National Parks Service – Analyzing ‘observations.csv’

```
species = pd.read_csv('species_info.csv')
species.fillna('No Intervention', inplace = True)
species['is_protected'] = species.conservation_status != 'No Intervention'
observations = pd.read_csv('observations.csv')

species['is_sheep'] = species.common_names.apply(lambda x: 'Sheep' in x)
species_is_sheep = species[species.is_sheep]
sheep_species = species[(species.is_sheep) & (species.category == 'Mammal')]
sheep_observations = observations.merge(sheep_species)
```

print sheep_observations.head()

	scientific_name	park_name	observations	category	common_names	conservation_status	is_protected	is_sheep
0	Ovis canadensis	Yellowstone National Park	219	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True
1	Ovis canadensis	Bryce National Park	109	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True
2	Ovis canadensis	Yosemite National Park	117	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True
3	Ovis canadensis	Great Smoky Mountains National Park	48	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True
4	Ovis canadensis sierrae	Yellowstone National Park	67	Mammal	Sierra Nevada Bighorn Sheep	Endangered	True	True

Analyzing the ‘observation’ and ‘species’ DataFrames to help track sheep locations.

The National Parks Service – Analyzing ‘observations.csv’

```
obs_by_park = sheep_observations.groupby('park_name').observations.sum().reset_index()
```

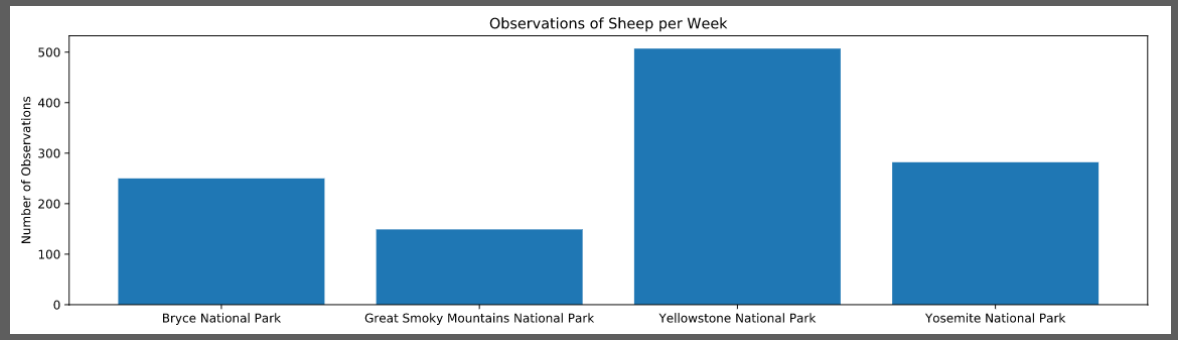
print obs_by_park

	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

Plotting Sheep Sightings

```
import codecademylib
import pandas as pd
from matplotlib import pyplot as plt

species = pd.read_csv('species_info.csv')
species['is_sheep'] = species.common_names.apply(lambda
x: 'Sheep' in x)
sheep_species = species[(species.is_sheep) &
(species.category == 'Mammal')]

observations = pd.read_csv('observations.csv')

sheep_observations = observations.merge(sheep_species)

obs_by_park =
sheep_observations.groupby('park_name').observations.sum()
.reset_index()
```

```
plt.figure(figsize=(16, 4))
ax = plt.subplot()
plt.bar(range(len(obs_by_park)),
        obs_by_park.observations.values)
ax.set_xticks(range(len(obs_by_park)))
ax.set_xticklabels(obs_by_park.park_name.values)
plt.ylabel('Number of Observations')
plt.title('Observations of Sheep per Week')
plt.show()
```



Sample Size Determination

baseline = 15

minimum_detectable_effect = 100*5./15

sample_size_per_variant = 870

yellowstone_weeks_observing = sample_size_per_variant/507.

bryce_weeks_observing = sample_size_per_variant/250.

great_smokey_observing = sample_size_per_variant/149.

yosemite_observing = sample_size_per_variant/282.

Given a baseline of 15% occurrence of foot and mouth disease in sheep at Bryce National Park, if a >5% drop in observed cases of foot and mouth disease in the sheep at Yellowstone was to be considered significant, then scientists should at least observe 510 sheep.

The observation data states that this would take approximately:

- one week of observing in Yellowstone
- two weeks in Bryce to see that many sheep
- Three and a half weeks in the Great Smoky Mountains National Park
- Almost two weeks in Yosemite