# Biodiversity Capstone Project

**Introduction to Data Analysis | Codecademy**

Christina Morgenstern, PhD
cmorgenstern@science-impuls.at

# The Biodiversity Analyst

- Christina Morgenstern, PhD
- Austria
- Task: investigating protected species
- Employer: National Parks Service

# Import relevant Python libraries

- Pandas
- Matplotlib

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

# Import data from .csv

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

| category | scientific_name | common_names | conservation_status |
|---|---|---|---|
| Mammal | *Bos taurus* | Aurochs, deomestic cattle | nan |
| | | | |

Example of one dataset

# How many different species are there in the table?

```
species_count = species.scientific_name.nunique()
print species_count
```

In total there are **5541** different species listed in the table.

# What are the different categories?

- The following classes of species are listed:
  - Mammal
  - Bird
  - Reptile
  - Amphibian
  - Fish
  - Vascular Plant
  - Nonvascular Plant

```
species_type = species.category.unique()
print species_type
```

# What are the possible conservation statuses?

- Species are assigned one of the following conservation status:
  - nan (not a number)
  - Species of Concern
  - Endangered
  - Threatened
  - In Recovery

```
conservation_statuses = species.conservation_status.unique()
print conservation_statuses
```

# The meaning of the different conservation statuses

- *Species of Concern:* declining population or appears to be in need of conservation.
- *Threatened:* vulnerable to endangerment in the near future.
- *Endangered:* seriously at risk of extinction.
- *In Recovery:* formerly endangered, but currently not in danger of extinction throughout all or a significant portion of its inhabitable range.

# Grouping of species by their conservation status

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

| | conservation_status | scientific_name |
|---|---|---|
| 0 | Endangered | 15 |
| 1 | In Recovery | 4 |
| 2 | Species of Concern | 151 |
| 3 | Threatened | 10 |

- 15 species are seriously at risk of extinction.
- 4 species are currently not endangered but have been previously.
- 151 species show a decline in their population numbers.
- 10 species are threatened.

# A more accurate representation

- Fill places of *nan* with argument ‘No intervention‘
- **5363** species need no intervention

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

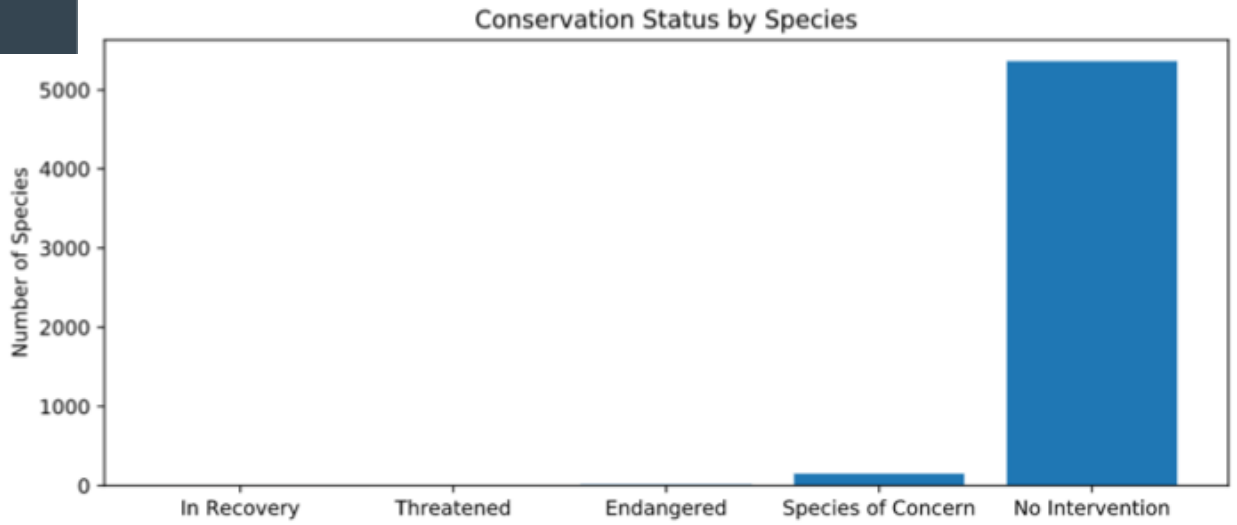
```
conservation_counts_fixed =
species.groupby('conservation_status').scientif
ic_name.nunique().reset_index()
print conservation_counts_fixed
```

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

# Graphical representation of conservation status per species

```
plt.figure(figsize=(10, 4))
ax = plt.subplot()
plt.bar(range(len(protection_counts)),protection_counts.scientific_name.v
alues)
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()
```

Number of species for each conservation status plotted as heights in a bar chart.



# Are certain types of species more likely to be endangered?

- A new column with the title ‘is_protected’ was added. This is TRUE if the conservation status is not equal to no intervention and FALSE otherwise.
- The data has been grouped by category and is_protected
- From these data **Birds** seem more likely to be endangered

| is_protected | category | False | True |
|---|---|---|---|
| 0 | Amphibian | 72 | 7 |
| 1 | Bird | 413 | 75 |
| 2 | Fish | 115 | 11 |
| 3 | Mammal | 146 | 30 |
| 4 | Nonvascular Plant | 328 | 5 |
| 5 | Reptile | 73 | 5 |
| 6 | Vascular Plant | 4216 | 46 |

# Performing a calculation on tabular data

- To make the data more readable the table was transformed into a pivot table
- Calculations of the percent of protected species were performed
- Results suggest that Mammals are more likely to be endangered than Birds

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

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

print category_pivot
```

| | 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 I: Mammals and Birds

- Test for significance between Mammals and Birds using Chi-Squared Test
- Comparison of two categorical datasets
- Null Hypothesis = the difference between the two species is due to chance

```
contingency = [[30, 146],
               [75, 413]]

from scipy.stats import chi2_contingency

chi2, pval, dof, expected = chi2_contingency(contingency)
print pval
```

- p-value = 0.687594809666
- Not significant
- Accept Null Hypothesis
- Result is due to chance

# Chi-Squared Test II: Mammals and Reptiles

```
contingency_2 = [[30, 146],
                 [5, 73]]

chi2, pval_reptile_mammal, dof, expected =
chi2_contingency(contingency_2)
print pval_reptile_mammal
```

- p-value: 0.0383555902297
- Significant value (because p-value < 0.05)
- Reject Null Hypothesis
- **Conclusion:** Different species are more likely to be endangered than others.

# Inspection of observations DataFrame

The data in the DataFrame *observations.csv* record sightings of different species at several national parks for the past 7 days.

| | scientific_name | park_name | observations |
|---|---|---|---|
| 0 | Vicia benghalensis | Great Smoky Mountains National Park | 68 |
| 1 | Neovison vison | Great Smoky Mountains National Park | 77 |
| 2 | Prunus subcordata | Yosemite National Park | 138 |
| 3 | Abutilon theophrasti | Bryce National Park | 84 |
| 4 | Githopsis specularioides | Great Smoky Mountains National Park | 85 |

# Apply Lambda function to select for species containing ‘Sheep‘

```
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]

print species_is_sheep

sheep_species = species[(species.is_sheep) & (species.category ==
'Mammal')]

print sheep_species
```

Focus of further analysis is on “sheep”.
Perform a Lambda function to select for data entries containing the word “sheep”

| | category | scientific_name | common_names | conservation_status | is_protected | is_sheep |
|---|---|---|---|---|---|---|
| 3 | Mammal | Ovis aries | Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral) | No Intervention | False | True |
| 1139 | Vascular Plant | Rumex acetosella | Sheep Sorrel, Sheep Sorrell | No Intervention | False | True |
| 2233 | Vascular Plant | Festuca filiformis | Fineleaf Sheep Fescue | No Intervention | False | True |
| 3014 | Mammal | Ovis canadensis | Bighorn Sheep, Bighorn Sheep | Species of Concern | True | True |
| 3758 | Vascular Plant | Rumex acetosella | Common Sheep Sorrel, Field Sorrel, Red Sorrel, Sheep Sorrel | No Intervention | False | True |
| 3761 | Vascular Plant | Rumex paucifolius | Alpine Sheep Sorrel, Fewleaved Dock, Meadow Dock | No Intervention | False | True |

# Selection of mammalian sheep species

Filter for sheep species that are mammals.

| | category | scientific_name | common_names | conservation_status | is_protected | is_sheep |
|---|---|---|---|---|---|---|
| 3 | Mammal | Ovis aries | Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral) | No Intervention | False | True |
| 3014 | Mammal | Ovis canadensis | Bighorn Sheep, Bighorn Sheep | Species of Concern | True | True |
| 4446 | Mammal | Ovis canadensis sierrae | Sierra Nevada Bighorn Sheep | Endangered | True | True |

# Merge of 2 DataFrames

```
sheep_observations = pd.merge(sheep_species, observations)
print sheep_observations.head()
```

| | category | scientific_name | common_names | conservation_status | is_protected | is_sheep | park_name | observations |
|---|---|---|---|---|---|---|---|---|
| 0 | Mammal | Ovis aries | Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral) | No Intervention | False | True | Yosemite National Park | 126 |
| 1 | Mammal | Ovis aries | Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral) | No Intervention | False | True | Great Smoky Mountains National Park | 76 |
| 2 | Mammal | Ovis aries | Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral) | No Intervention | False | True | Bryce National Park | 119 |
| 3 | Mammal | Ovis aries | Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral) | No Intervention | False | True | Yellowstone National Park | 221 |
| 4 | Mammal | Ovis canadensis | Bighorn Sheep, Bighorn Sheep | Species of Concern | True | True | Yellowstone National Park | 219 |

The previous DataFrame sheep_species is merged with the observations DataFrame in order to have all relevant information of sheep in one table.

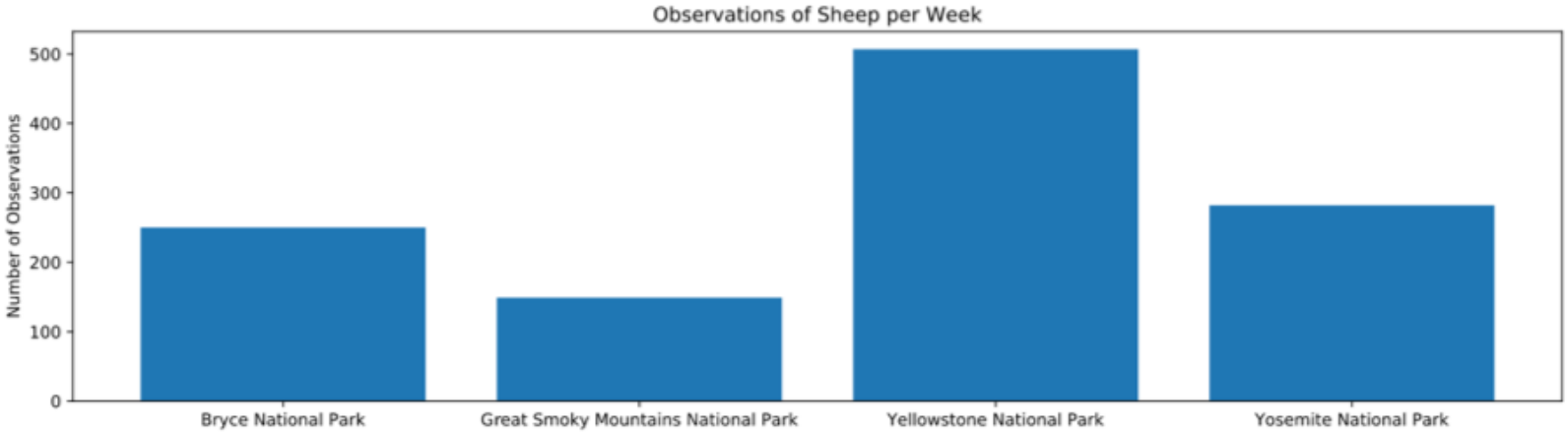
# Total Sheep observations grouped by park_name

```
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 |

The new DataFrame obs_by_park contains information about the different National Parks and the recorded sheep observations.

# Plot sheep observations per week using a bar chart




The bar char shows the number of sheep observations per week as the height of the bars at the respective National Parks.

# Foot-and-mouth disease

- Infectious viral disease that affects cloven-hoofed animals
- Symptoms: high fever for 2-6 days followed by blisters inside the mouth and on the feet
- Highly infectious
- Implementation of monitoring to avoid spreading of the disease


Image Courtes:  Texas A&M University College of Veterinary Medicine

# Foot-and-mouth disease reduction effort

- baseline = 15%
- minimum_detectable_effect = 33
- sample_size_per_variant = 520
- yellowstone_weeks_observing = 1
- bryce_weeks_observing = 2

At Bryce National Park the occurrence of foot-and-mouth disease in sheep is **15%** (baseline).
In order to observe a >5% drop in cases of foot-and-mouth-disease at Yellowstone National Park we would have to observe **520** sheep.
Using this data we would have to observe sheep for 1 week in Yellowstone National Park and 2 weeks in Bryce National Park.