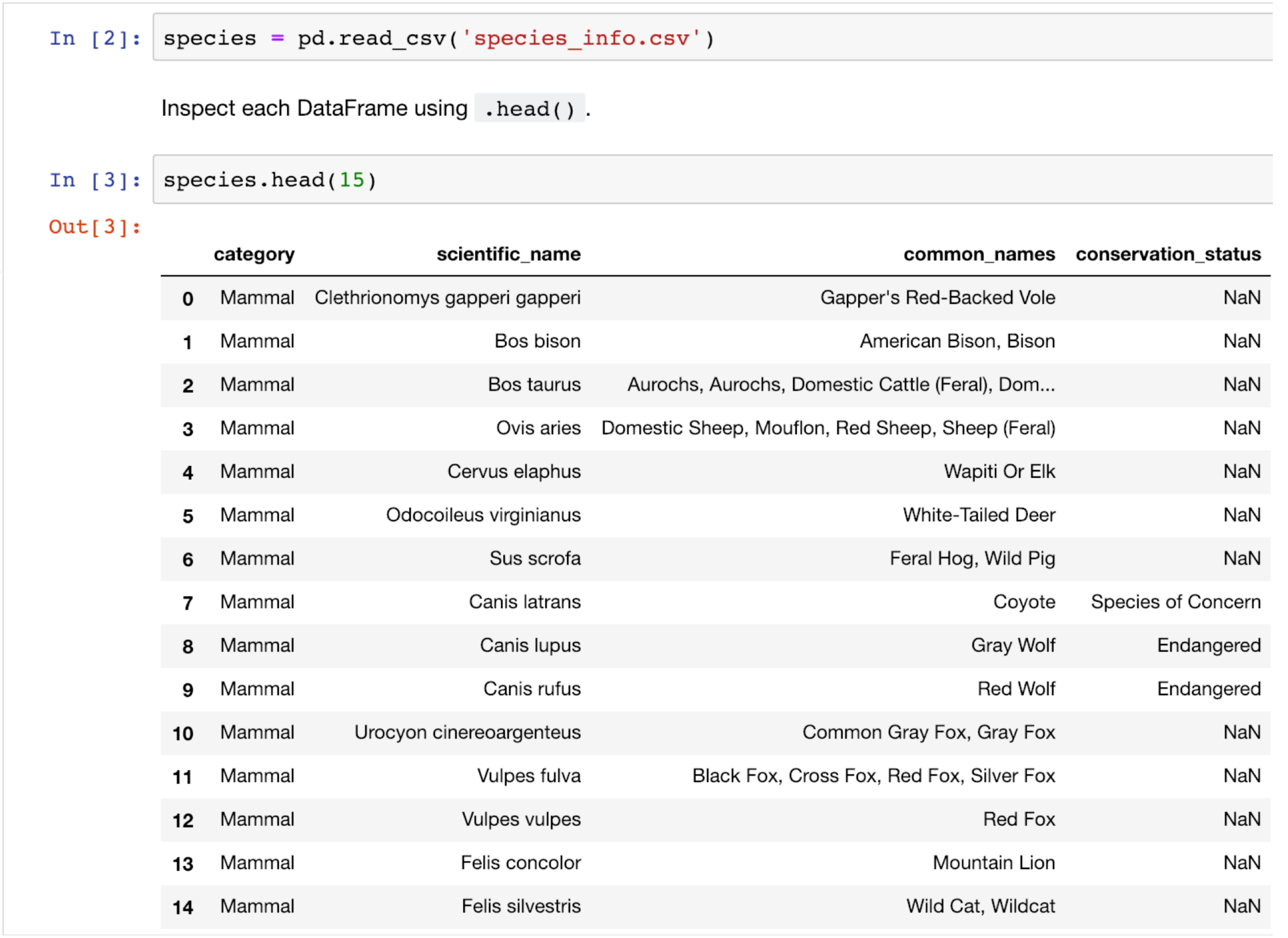
Code Academy Capstone Project

Biodiversity for the National Parks

Describing data in ‘species_info.csv’

A number of features about the data were immediately observed after inspecting the first 15 rows:

- All data is in string form (except the df index)
- Scientific name and category contained exclusively single string data
- Common_names could contain many strings, separated by commas
- Conservation_status was a mix of both strings and NaN



```
In [2]: species = pd.read_csv('species_info.csv')
```

Inspect each DataFrame using `.head()`.

```
In [3]: species.head(15)
```

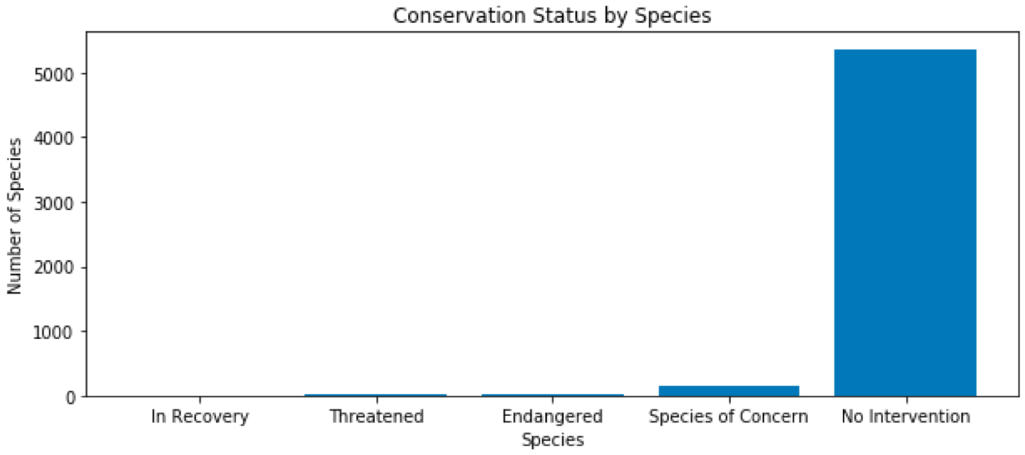
Out[3]:

	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), Dom...	NaN
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	NaN
4	Mammal	Cervus elaphus	Wapiti Or Elk	NaN
5	Mammal	Odocoileus virginianus	White-Tailed Deer	NaN
6	Mammal	Sus scrofa	Feral Hog, Wild Pig	NaN
7	Mammal	Canis latrans	Coyote	Species of Concern
8	Mammal	Canis lupus	Gray Wolf	Endangered
9	Mammal	Canis rufus	Red Wolf	Endangered
10	Mammal	Urocyon cinereoargenteus	Common Gray Fox, Gray Fox	NaN
11	Mammal	Vulpes fulva	Black Fox, Cross Fox, Red Fox, Silver Fox	NaN
12	Mammal	Vulpes vulpes	Red Fox	NaN
13	Mammal	Felis concolor	Mountain Lion	NaN
14	Mammal	Felis silvestris	Wild Cat, Wildcat	NaN

Describing data in ‘species_info.csv’

While working through the project the following was observed:

- 7 unique entries exist in ’category’
  - ['Amphibian' 'Bird' 'Fish' 'Mammal' 'Nonvascular Plant' 'Reptile’, 'Vascular Plant']
- 4 types valid entries exist in ‘conservation_status’
  - [nan nan nan ... 'In Recovery' 'Species of Concern' 'Threatened’]
  - NaN is not unique as <type 'float’>
- When using .group_by() NaNs are totally ignored
- The vast majority of the data is NaN, i.e. it does not have a conservation status, as illustrated by the plot below to the right



Significance Calculations

Breaking down the problem:

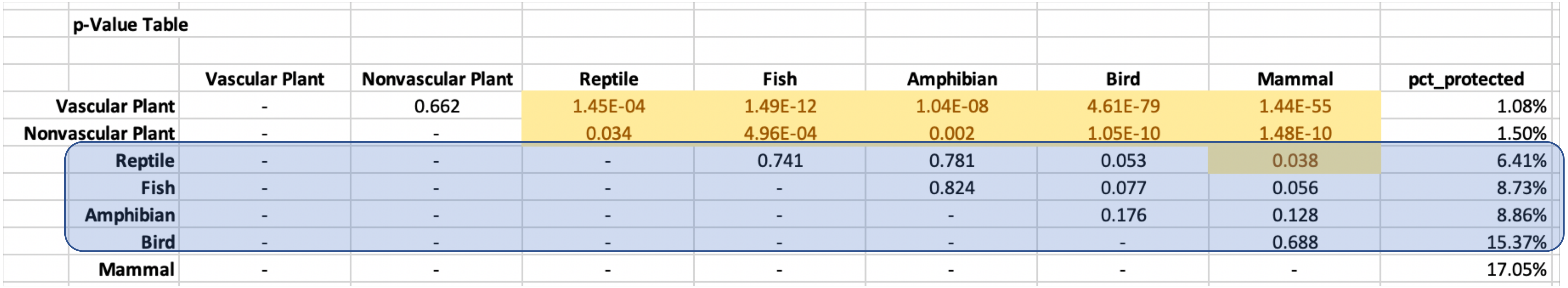
- **Fundamental Assessment :** Test if the percentage protection of two species pairs are significantly different.
- **Contingency Table :** Defined only the count of species which *are* and *are not* under protection status.
- **Null Hypothesis :** There is no significant difference between the percentage protection of each pair. Data is required for (Mammals & Birds) and (Mammals & Reptiles). Percentage protection is taken as a valid metric of probability of endangerment.
- **Results and insights :**
    - P Value for Birds vs Mammals = 0.688. Using 0.05 as a statistic significance threshold, we ***cant*** reject the null hypothesis and must accept ***there is no statistical difference in probability of protection here.***
    - P Value for Mammals vs Reptiles = 0.038. Using 0.05 as the statistical significance threshold ***we can reject the null hypothesis*** and say ***there is a statistical difference in probability of protection here.***

Recommendation : There are statically less reptile species under protection status than mammals. If choosing at random to protect one species or mammal or reptile, it is more likely to helpful for conservation to protect a mammal. When choosing between Birds and Mammals there is no statistical difference. Further, they should do a full analysis of every protection ratio. So I did that.

Significance Calculations

Deeper analysis : Comparing the statistical significance between all pairs reveals some interesting insights :

- *Mammals, Birds, Amphibians and Fish* all have **statistically similar rates of protection**. This implies that when choosing one thing at random from any of these groups to all conservation efforts to, you are equally likely to pick a protected thing.
- *Mammals* are **statistically significantly more protected** than *reptiles***.** This implies that should you pick one thing at random from a category to protect, you should chose from the Mammals.
- *Mammals, Birds, Amphibians, Fish* and *Reptiles* are **statistically significantly more protected** than both *Vascular and Non-Vascular plants.*
- There is **no statistical significance between** the rates of protection of *Vascular* and *Non-Vascular* plants.



p-Value Table

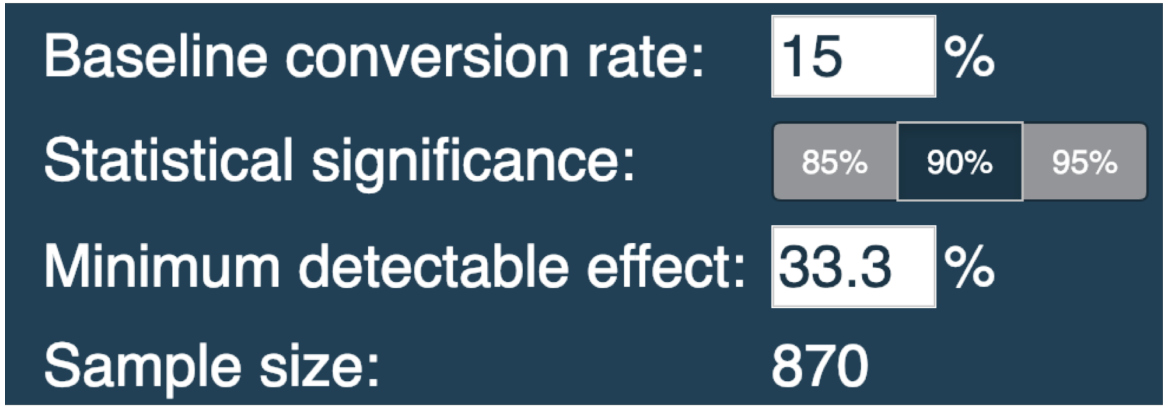
	Vascular Plant	Nonvascular Plant	Reptile	Fish	Amphibian	Bird	Mammal	pct_protected
Vascular Plant	-	0.662	1.45E-04	1.49E-12	1.04E-08	4.61E-79	1.44E-55	1.08%
Nonvascular Plant	-	-	0.034	4.96E-04	0.002	1.05E-10	1.48E-10	1.50%
Reptile	-	-	-	0.741	0.781	0.053	0.038	6.41%
Fish	-	-	-	-	0.824	0.077	0.056	8.73%
Amphibian	-	-	-	-	-	0.176	0.128	8.86%
Bird	-	-	-	-	-	-	0.688	15.37%
Mammal	-	-	-	-	-	-	-	17.05%

Sample Size Determination

Breaking down the problem :

- **Fundamental Assessment :** Determine how many sheep must be observed to determine if a 5% change has been made to the baseline rate of foot and mouth
- **Variables :**
  - **Baseline conversion rate : 15%** is the rate on which we need to define a difference.
  - **Minimum Detectable Effect :** 5% out of 15% which gives 5/15 = ***33%***
  - **Statistical Significance :** Default significance of ***90%*** is used

Sample size required : 870 sheep



Baseline conversion rate: 15 %

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

Minimum detectable effect: 33.3 %

Sample size: 870

Plots

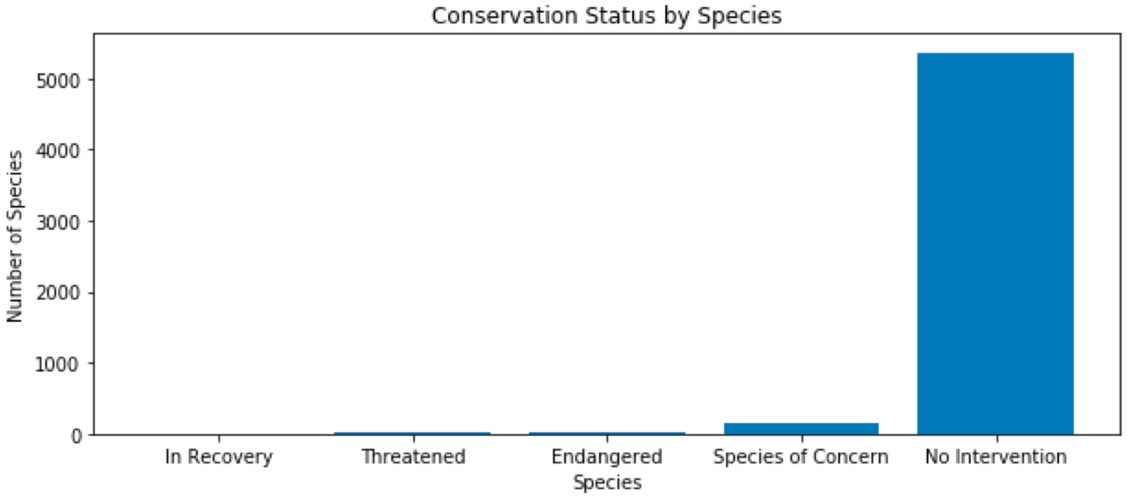


Fig 1 : Conservation status by species

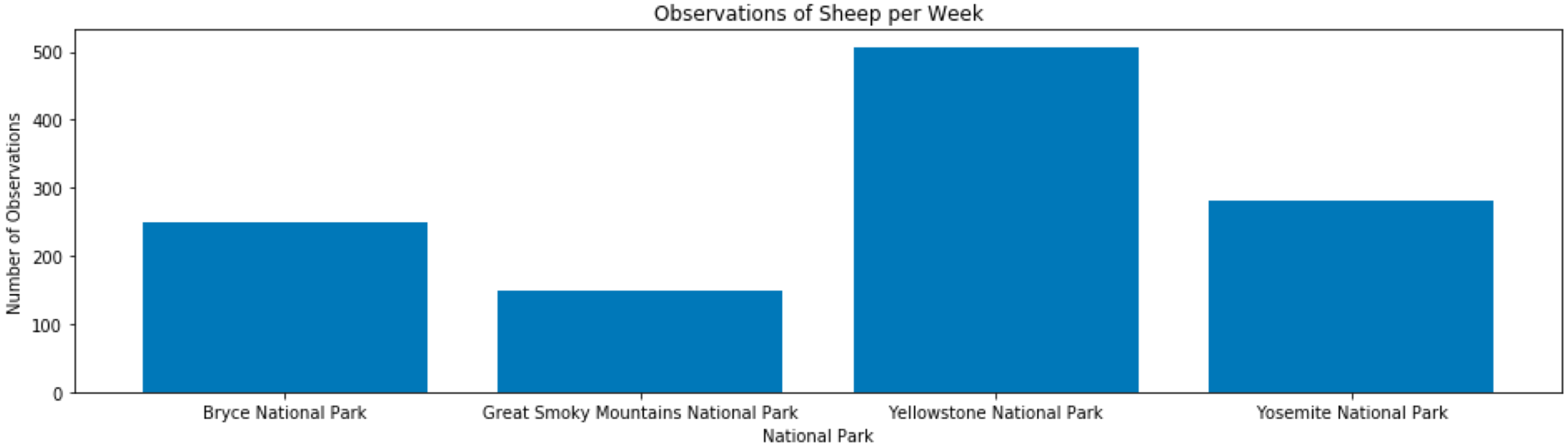


Fig 2 : Sheep observations by national park