Biodiversity for National Parks

Monitoring Conservation Status and Managing Wildlife

Biodiversity for National Parks

Overview

The National Park Service observes and manages wildlife found in the national parks.

The first part of this project will consider which categories of plants and animals found in the park have species more likely to become endangered and therefore requiring some level protective intervention.

The second part of the project will focus on determining sample size to study foot and mouth disease among sheep populations at selected parks.

Biodiversity for National Parks

Tools Used for Analysis

Jupyter Notebook, Pandas, SciPy and Matplotlib all belong to the Python programming world. All of these tools are open source, easily accessible and easy to use. In addition, they are powerful tools for performing data analysis of plants and animals in the national parks.

- Jupyter Notebook - online Python programming environment.
- Python - object-oriented programming language frequently used in data analysis.
- Pandas - used to create data frames and data summarization.
- SciPy - used for chi squared test.
- Matplotlib - plotting utility to graph results produced by Pandas.

Biodiversity for National Parks

Jupyter Notebook

The Jupyter Notebook is an online Python programming environment with a graphical user interface (GUI) that combines a programming text editor window with an output window.

Biodiversity for National Parks

Python

Python is a general purpose, object-oriented programming language which is used extensively for data analysis and other scientific applications.

Biodiversity for National Parks

Pandas

Pandas is a library offering a number of functions and methods that allow you to work with data for the Python programming language.

We use Pandas to load data into the Pandas equivalent of a table called dataframe.

Pandas offers methods to sort, manipulate, and summarize data stored in dataframes.

Biodiversity for National Parks

SciPy

SciPy is a library offers a number of methods to perform scientific and statistical analysis of data for the Python programming language.

Although SciPy offers several statistical analysis tests and two specifically designed to work with categorical data, we will use the Chi2_Contingency Test in analyzing data regarding the conservation status of animals and plants in the national parks.

Reasons for selecting the Chi2_Contingency Test:

- two categories of data: Data for Group A (groups requiring no protective intervention) and data for Group B (groups requiring protective intervention).

- compare a table of results with different totals.

We will use a P-Value of 0.05 which is commonly used to determine if results are statistically significant.

Biodiversity for National Parks

Matplotlib

Matplotlib is a Python library which offers the ability to plot graphs in order to visualize data.

Matplotlib is used to make bar graphs, histograms, line graphs, pie charts.

Matplotlib was used to create the bar graphs used in our study to plot

- conservation status of species in the national parks
- sheep counts (totals) in the national parks
- sheep count by species in the national parks

Biodiversity for National Parks

Part 1: Exploring patterns in Conservation Protection

An important concern of the National Park Service regards conservation of various species in the national parks.

The conservation status of plants and animals in the national parks are recorded in *species_info.csv,* a comma separated values file.

The species_info.csv contains the following information for known animals in the parks:

- category
- scientific name
- common name
- conservation status

Biodiversity for National Parks

About species_info.csv

create and load data in csv into a Pandas data frame, species

There are a total of 5824 records

- 5541 unique scientific names; a discrepancy of 283 between number of records and number of unique scientific names (5824 - 5541).
- 5504 unique common names; a discrepancy of 320 between number of records and unique common names (5824 - 5504).
- The discrepancy between number of records and number of unique names suggests that both scientific names and common names might be repeated for animals and plants in the file.

Four columns: **category, scientific_name, common_name, conservation_status**

Category: **Mammal, Bird, Reptile, Amphibian, Fish, Vascular Plant, Nonvascular Plant**

Conservation Status: **nan, In Recovery, Species of Concern, Threatened, Endangered**

Biodiversity for National Parks

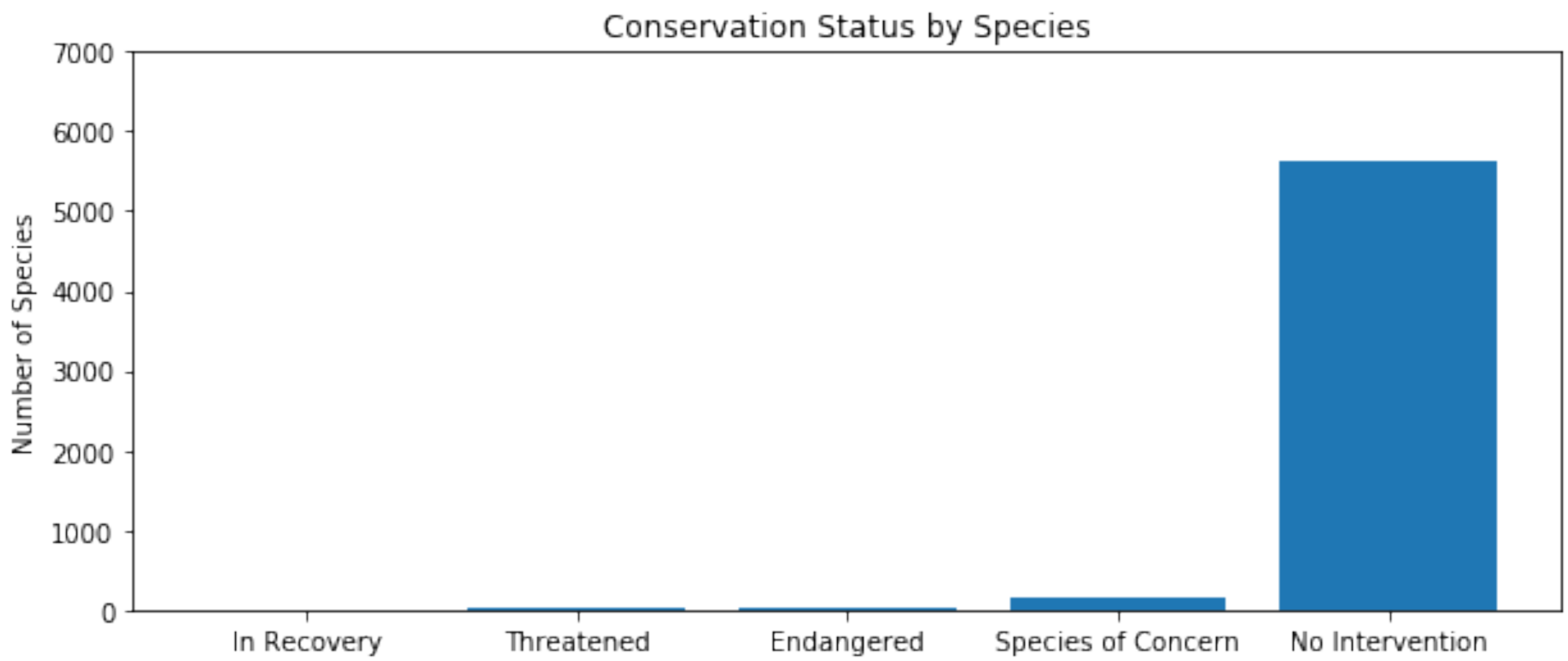
Table 1: Totals for Each Conservation Level

Conservation Status	**Totals**
In Recovery	4
Threatened	10
Endangered	16
Species of Concern	161
No Intervention	5633

Of the 5584 records in species_info.csv, only **191** species fall into a conservation status requiring protection.

The remaining 5633 species of plants and animals in the parks do not need any protection.

Biodiversity for National Parks



In this bar chart generated by Matplotlib, we can see that in comparison, most plants and animals in the parks do not require any sort of conservation measures.

Biodiversity for National Parks

Which Category has the most number of species under some level of conservation intervention?

Table 2: Number of Species in Each Category Currently Under Protection

Category	Protected	Not Protected	Totals	Percentages
Vascular Plant	46	4216	4262	0.010793
Nonvascular Plant	5	328	333	0.015015
Reptile	5	73	78	0.064103
Fish	11	115	126	0.087302
Amphibian	7	72	79	0.088608
Bird	75	413	488	0.153689
Mammal	30	146	176	0.170455

Biodiversity for National Parks

Based on the table which lists the number of species under each category currently under protection, we can say the following:

- **Vascular Plants**

  ~ 1% of total vascular plant species found in the national parks are currently protected.

- **Nonvascular Plants**

  ~ 2% of total nonvascular plant species found in the national parks are currently protected.

- **Reptile**

  ~ 6% of total reptile species found in the national parks are currently protected.

Biodiversity for National Parks

- **Amphibian**

  ~ 9% of amphibian species are currently under some level of conservation protection.

- **Fish**

  ~ 9% of fish species are currently under some level of conservation protection.

- **Bird**

  ~ 15% of bird species are currently under some level of conservation protection.

- **Mammal**

  ~ 17% of mammal species are currently under some level of conservation protection.

Biodiversity for National Parks

Table 3: Summary of our findings regarding categories requiring intervention and no intervention:

Group A: No Intervention	Group B: Intervention Required
Vascular Plants (~ 1%)	Amphibians (~ 9%)
Nonvascular Plants (~ 2%)	Fish (~ 9%)
Reptile (~ 6%)	Bird (~ 15%)
	Mammal (~ 17%)

It seems like the categories of plants and animals in national parks fall into two broad conservation protocols:

Group A: categories requiring little or practically no intervention

- 6% or less of its species are currently under some level of conservation protection

Group B: categories requiring intervention

- 9% or more of its species are currently under some level of conservation protection.

Biodiversity for National Parks

Questions

1. Are the variations between members within Group A and within Group B of statistical significance?

2. Are the variations between members of Group A and Group B of statistical significance?

Statistical Significance

- Null Hypothesis I: Variations **within** a group are due to random variation and are not statistically significant.

- Null Hypothesis II: Variations **between** groups are due to random variation and are not statistically significant.

- We shall use a p-value of 0.05 to either accept or reject the Null Hypotheses.

Biodiversity for National Parks

Table 4: Number of Protected Species in Each Category for Group A and Group B.

Group A	Group B
Vascular Plant (1%) Protected: 46 Not Protected: 4216	Amphibian (9%) Protected: 7 Not Protected 72
Nonvascular Plant (2%) Protected: 5 Not Protected 328	Fish (9%) Protected: 11 Not Protected: 115
Reptile (6%) Protected: 5 Not Protected: 73	Bird (15%) Protected: 75 Not Protected: 413
	Mammal (17%) Protected: 30 Not Protected: 146

Biodiversity for National Parks

Procedure for Chi2_contingency test.

Create a contingency table for each test group

- Group A contingency table: [[46, 4216], [5, 328], [5, 73]]
- Group B contingency table: [[7, 72], [11, 115], [75, 413], [30, 146]]
- Group A and Group B: [[46, 4216], [5, 328], [5, 73], [7, 72], [11, 115], [75, 413], [30, 146]]
- Mammal and Bird: [[30, 146], [75, 413]]
- Reptile and Mammal: [[5, 73], [30, 146]]
- Bird and Fish: [75, 413], [11, 115]]
- Vascular Plant and Nonvascular Plant: [[46, 4216], [5, 328]]
- Nonvascular Plant and Reptile: [[5, 328], [5, 73]]

Use the contingency table for each group in SciPy's chi2_contingency test. The test returns four values. Use four variables to capture the four values returned by the chi2_contingency test. The p-value is the second value returned by the chi2_contingency test.

- chi_value, pvalue, def_of_freedom, expected_frequencies = chi2_contingency(insert_contingecy_table_here)

Biodiversity for National Parks

Table 5: Results of SciPy's chi2_contingency test. Are the variations between members within Group A and within Group B of statistical significance? In order to be statistically significant, the p-value must by less than 0.05.

Test Groups	P-values	Statistical Significance?
Within Group A	p-value Group A 8.8563422494e-05	Yes
Within Group B	Group B is : 0.0829759602065	No
Group A and Group B (Combine)	p-value is 5.51082804731e-89	Yes
Mammal (B) and Bird (B)	0.68759480966613362	No
Reptile (A) and Mammal (B)	0.038355590229698977	Yes
Bird (B) and Fish (B)	0.076681995690571936	No

Biodiversity for National Parks

Conclusion

Is there statistical significance within a group?

Test Group A: p-value is less than 0.05 therefore, the variation in number of species protected between the three categories is statistically significant. This is unexpected. Could this be a result of throwing in the Reptile category along with the Vascular and Nonvascular Plants? Both plant categories require minimal intervention. Only 1% - 2% of vascular and nonvascular plants need protection. However, 6% of reptile species require protection.

To explore, do further chi2_contingency tests on members of Group A to see if the p-value changes.

Vascular Plant and Nonvascular Plant: Both are in Group A. The p-value is 0.66234194913819855. The p-value is greater than 0.05 therefore we can accept the Null Hypothesis. There is no statistical significance in number of species protected in these two categories.

Nonvascular Plant and Reptile. Both are in Group A. The p-value is 0.03362698310726171З. The p-value is less than 0.05 therefore we have to reject the Null Hypothesis between these categories. The difference in variation between these two groups is statistically significant.

Biodiversity for National Parks

Conclusion (continued)

Is there statistical significance within a group?

Test Group B: the p-value is greater than 0.05 and suggests there is no statistical significance to the variation in numbers of species protected. The variation in numbers of species protected for each category is due to random variation. This suggest we have to accept the Null Hypothesis.

Test Group Mammal and Bird: Both of these categories are in Group B. The p-value is greater than 0.05 and suggests there is no statistical significance in variation in number of species protected in these categories

Test Group Bird and Fish: Both of these categories are in Group B. The p-value is greater than 0.05 and suggests again there is no statistical significance in variation in number of species protected in these categories.

In all of these tests within group B, the variation in number of species protected for categories within a group is due to random chance. Variation in numbers of species protected for categories in each group is due to random fluctuation and is not statistically significant. In essence, we have to accept the Null Hypothesis for variation in number of species protected within a group.

Biodiversity for National Parks

Conclusion (continued)

Is there statistical significance between Group A and Group B?

Test Group A and Group B (combine): the p-value is less than 0.05 therefore, there is statistical significance in the number of species protected for each category of organism in the national parks.

Test Group Reptile and Mammal: Reptile is in Group A and Mammal is in Group B. The p-value is less than 0.05, therefore there is statistical significance in the numbers of species protected for each category of organism in the national parks. It seems Mammal species require more intervention than Mammal species.

When we perform the chi2_contingency test on categories from Group A and Group B, we see that there is statistical significance in the variation in numbers of species protected for categories in Group A and Group B. Species in Group B require intervention while those in Group A (except for Reptile) require very minimal intervention.

Biodiversity for National Parks

Conclusion (continued)

Tabular data suggests that 17% of Mammal species in the national parks are under some level of conservation protection. This suggests that mammal species frequently need protective intervention to maintain healthy populations in the park.

The next most protected category is the Bird category. It seems that 15% of bird species also are under some level of conservation protection.

Amphibians and Fish both have 9% of species in their respective categories under some level of conservation protection. 6% of Reptile species are under some form of conservation protection.

On the low end of conservation protection, Vascular Plants (1%) and Nonvascular Plants (2%) do not have many species under conservation protection.

Biodiversity for National Parks

Conclusion (continued)

If we were to create a range of conservation protection:

- Mammal and Bird categories have the most number of species under some level of conservation protection.

- Amphibians, Fish and Reptiles are in the middle with regard to numbers of species under some level of conservation protection.

- Nonvascular and Vascular plant species require the least level of conservation intervention.

Biodiversity for National Parks

Questions to explore further

1. Look at each category and determine the actual level of conservation protection of the different species in each category. For example, explore which bird species are under conservation protection and the level of that conservation protection.

2. Can we say anything about differences in level of conservation protection between the categories? For example, can we say that bird species are most commonly at endangered levels while mammals species are typically at species-of-concern level of protective intervention?

3. Are categories which have little or no protective intervention genuinely faring well in the park or is there a bias for certain categories by park rangers?

Biodiversity for National Parks

Part 2:
Determining Sample Size

Biodiversity for National Parks

Part 2: Determining Sample Size

Scientists are studying the incidence of Foot and Mouth disease among sheep populations in the national parks.

In order to do conduct this study, we will need to know the different sheep species that are currently present in the national parks.

Next, we will need to have population counts for the different species of sheep found in the national parks.

Using this information, we can determine a sample size to see if the incidence of foot and mouth disease among sheep populations is decreasing to desired levels.

Biodiversity for National Parks

We can formalize the goals for this section by posing and answering the following questions:

1. How many sheep species are present in the national parks and what are their scientific names?

2. What is the combined total of sheep populations at each park?

3. What is the population of each species of sheep at each park?

4. How do we determine sample size to test sheep populations for hoof and mouth disease?

Biodiversity for National Parks

To do this study, we will use data stored in two .csv files

- species_info.csv
    - 5824 records
    - create and load data from species_info.csv into a pandas data frame species
    - species data frame will store data in the following columns: category, scientific_name, common_name, conservation_status.
    - category column has seven possible values: Mammal, Bird, Reptile, Amphibian, Fish, Vascular Plant, Nonvascular Plant
- observations.csv
    - 23, 296 records
    - create and load data from observations.csv file into a pandas data frame observations
    - data is stored in the following columns: **scientific_name, park_name, observations**
    - park names (4) : **Bryce National Park, Yellowstone National Park, Yosemite National Park, Great Smoky Mountains National Park**
    - observations column contains the number of times a species has been observed in a park.

Biodiversity for National Parks

Question: How many sheep species are present in the national parks and what are their scientific names?

Table 6: Sheep Species Present in National Parks. Excerpted from the sheep_species dataframe generated by querying the species dataframe.

Scientific Name	Common Name
Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep(feral)
Ovis canadensis	Bighorn Sheep
Ovis canadensis sierare	Sierra Nevada Bighorn Sheep

We have three different sheep species that are present in the national parks: Ovis aries, Ovis canadensis and Ovis canadensis sierrae.

Biodiversity for National Parks

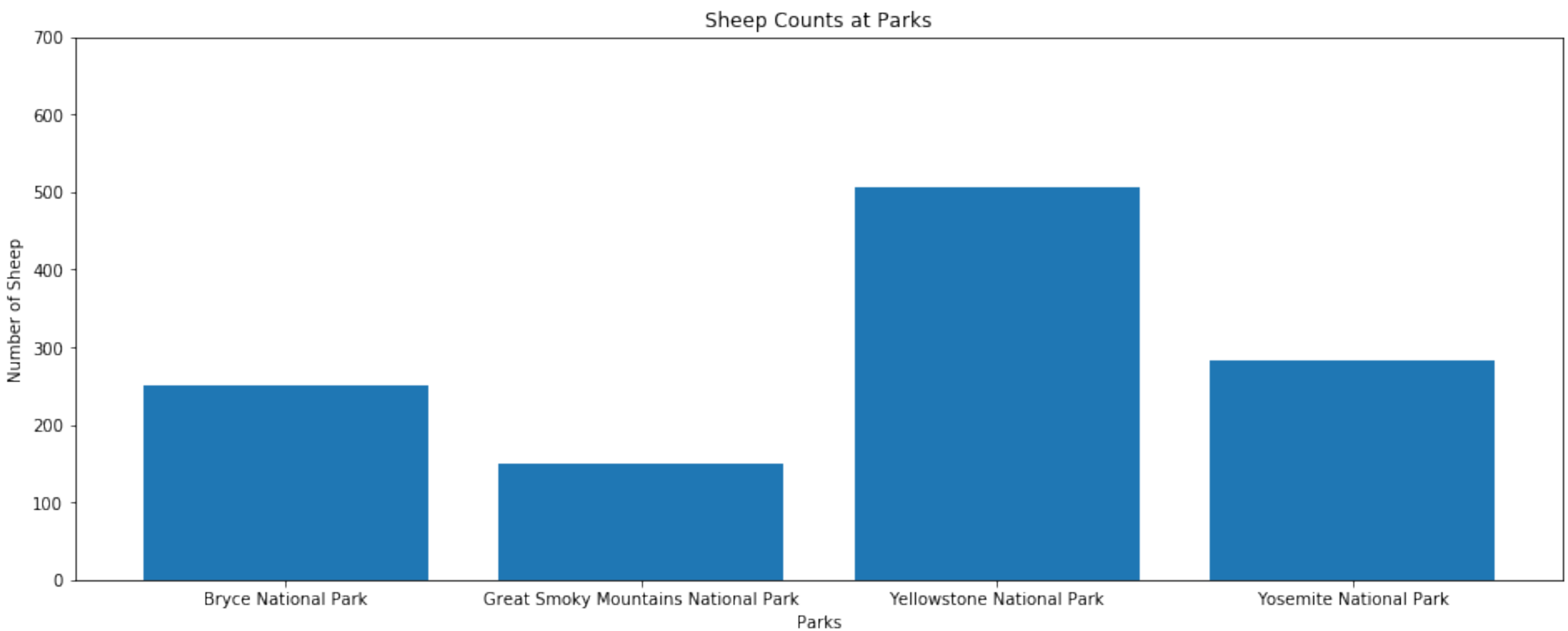
Question: What is the combined total of sheep populations at each park?

Table 7: Sheep populations at National Parks

Park Name	Sheep Totals
Great Smoky Mountains National Park	149
Bryce National Park	250
Yosemite National Park	282
Yellowstone National Park	507

Great Smoky Mountains National Park has the smallest sheep population and Yellowstone National Park has the largest sheep population.

Biodiversity for National Parks



We can see that Yellowstone National Park has the highest population of sheep of the four national parks in our study and that Great Smoky Mountains National Park has the lowest sheep population.

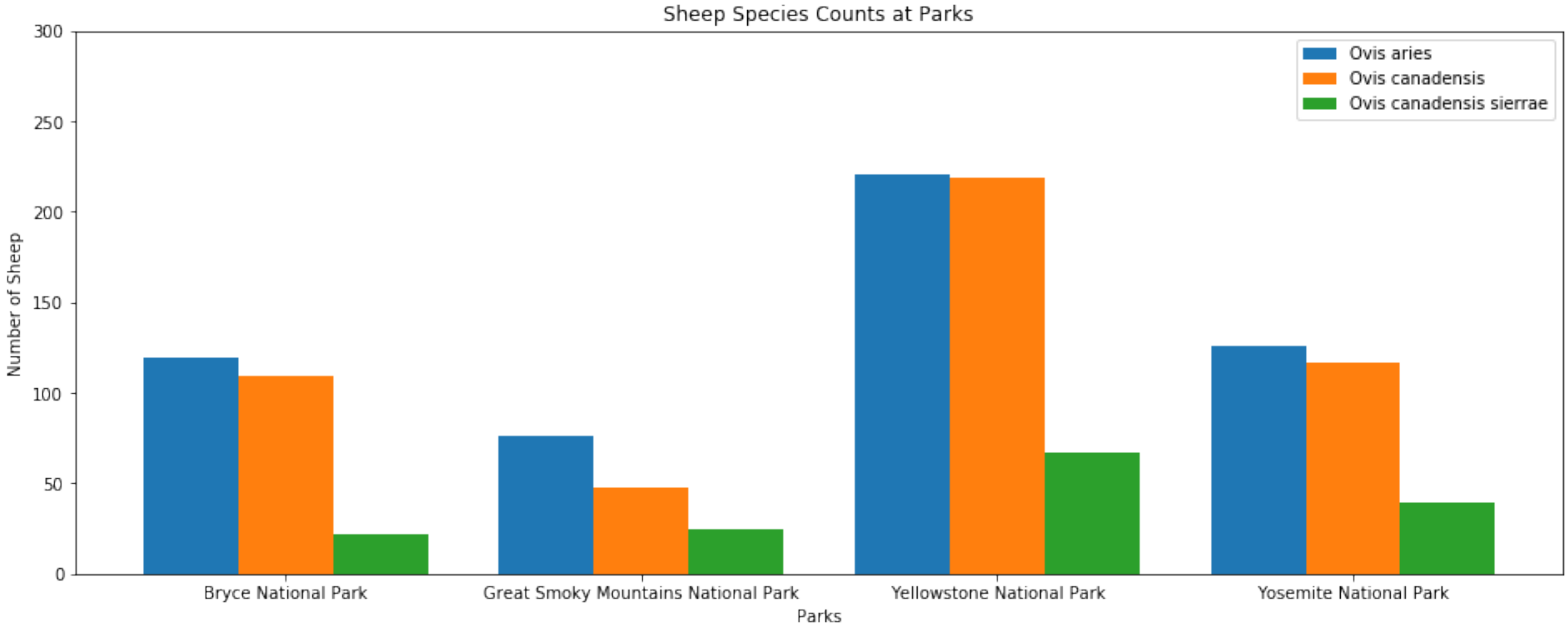
Biodiversity for National Parks

Question: What is the breakdown by species of sheep populations at the four national parks in our study?

Table 8: Breakdown of Sheep Populations at National Parks by Sheep Species.

	Ovis aries	Ovis canadensis	Ovis canadensis sierrae
Great Smoky Mountains National Park	76	48	25
Bryce National Park	119	109	22
Yosemite National Park	126	117	39
Yellowstone National Park	221	219	67

Biodiversity for National Parks



Yellowstone National Park has the highest population of sheep while Great Smoky Mountains National Park has the lowest population of sheep.

Ovis aries and Ovis canadensis are more common than Ovis canadensis sierrae at all four parks.

Biodiversity for National Parks

Question: Determine sample size to test sheep populations for hoof and mouth disease.

Background

Scientists at Bryce National Park are studying sheep populations for foot and mouth disease. 15% of the sheep population at Bryce National Park have foot and mouth disease.

At a different national park, Yellowstone National Park, park rangers have implemented a treatment protocol to help reduce the incidence of foot and mouth disease in sheep populations.

A successful treatment protocol will see a reduction in incidence of foot and mouth disease to about 10% of the sheep population in the parks.

Biodiversity for National Parks

Calculating Sample Size

We will use an online A/B Sample Size calculator found at

- https://www.optimizely.com/sample-size-calculator/

We are using an A/B Sample size calculator since the sheep population falls into two categories: those with foot and mouth disease and those without the disease.

Biodiversity for National Parks

Calculating Sample Size

We will provide the online A/B Sample Size Calculator with the following information:

- baseline conversion rate: 15%
- minimum detectable effect: 33.3%
- statistical significance: 90%

Biodiversity for National Parks

Calculating Sample Size: Baseline Conversion Rate

The baseline conversion rate of 15% represents the current rate of incidence of foot and mouth disease among sheep populations at Bryce National Park.

Biodiversity for National Parks

Calculating Sample Size: Minimum Detectable Effect

Minimum detectable effect represents a percent change that will show that the treatment protocol to reduce incidence of foot and mouth disease is effective.

To calculate minimum detectable effect

- 15%, the current rate of incidence of foot and mouth disease in sheep populations

- 10%, a desirable rate of incidence of foot and mouth disease:

- 100 * (10 - 15) / 15 = 33.3333

Biodiversity for National Parks

Sample Size

510

(if we use a minimum detectable effect of 33.3%)

520

(if we round minimum detectable effect to 33%)

Biodiversity for National Parks

Question: How much time will we need to test sheep for incidence of foot and mouth disease?

Bryce National Park

- Bryce has a sheep population of 250
- Approximately 2 weeks (510 / 250 = 2.04)

Yellowstone National Park

- Yellowstone has a sheep population of 507
- Approximately 1 week (510 / 507 = 1.00)