BIODIVERSITY PROJECT

ANALYSIS OF WILDLIFE

BIODIVERSITY PROJECT

INTRODUCTION

- A biodiversity analyst works with the National Park Service to help observe and manage wildlife.
- This analysis will use the species dataset which contains over 200 species in the wild.
- Some of the species require protection and some are endangered.
- The second part of this analysis will use the observations dataset which is observations of sheep sightings in different national parks.
- This analysis will help scientists to reduce foot and mouth disease in sheep at Yellowstone and Bryce National Parks.

BIODIVERSITY PROJECT

TOOLS USED

- Jupyter Notebook; a web application for sharing information.
- Python; a high-level computer programming language.
- Pandas; a Python package for doing data analysis.
- SciPy; an open-source Python library used for scientific computing.
- Matplotlib; a plotting library for graphs in Python.

BIODIVERSITY PROJECT

SETTING UP THE ENVIRONMENT

- Install Python on computer.
- Open Jupyter Notebook using the command line.
- Load in the modules matplotlib and pandas.
- Download the two CSV files: “species” and “observations”.
- Load “species_info.csv” into a data frame and name it “species”.
- Inspect the first 5 rows using the command “species.head()”.
- Load “species_info.csv” into a data frame and name it “observations”.
- Inspect the first 5 rows using the command “observations.head()”.

BIODIVERSITY PROJECT

FIRST ANALYSIS: THE SPECIES DATASET

- The first dataset is a CSV file, “species_info.csv”, that contains data about different species in our National Parks, including:

- The category of species

- The scientific name of each species.

- The common names of each species.

- The species conservation status.

- Using this dataset we will answer various questions related to the dataset.

BIODIVERSITY PROJECT

ABOUT THE SPECIES DATASET

- The CSV file was loaded into Python as a Pandas data frame.
- There are 5,824 rows and 4 columns.
- The column names are : category, scientific_name, common_names, conservation_status.
- There are 7 different kinds of species categories in the species dataset.
- There are five different kinds of conservation_statuses: “Species of Concern”, “Endangered”, “Threatened”, “In Recovery”, and “No Intervention”.
- 180 species were either “Species of Concern”, “Endangered”, “Threatened”, or “In Recovery”.
- 5644 species were in the “No Intervention” status.

BIODIVERSITY PROJECT

INSPECTING THE DATA

- Are certain types of species more likely to be endangered?
- Step 1. Create a new column called “is_protected” which is species that are labeled “True” in the “conservation_status” column but “False” in the “No Intervention” column.
- Step 2. Group “category” and “is_protected” using the groupby function from the Pandas module we imported into Python.
- Step 3. Save it with the variable name “category_counts”
- Step 4. Use the head function to view the first 5 observations using the command category_counts.head()

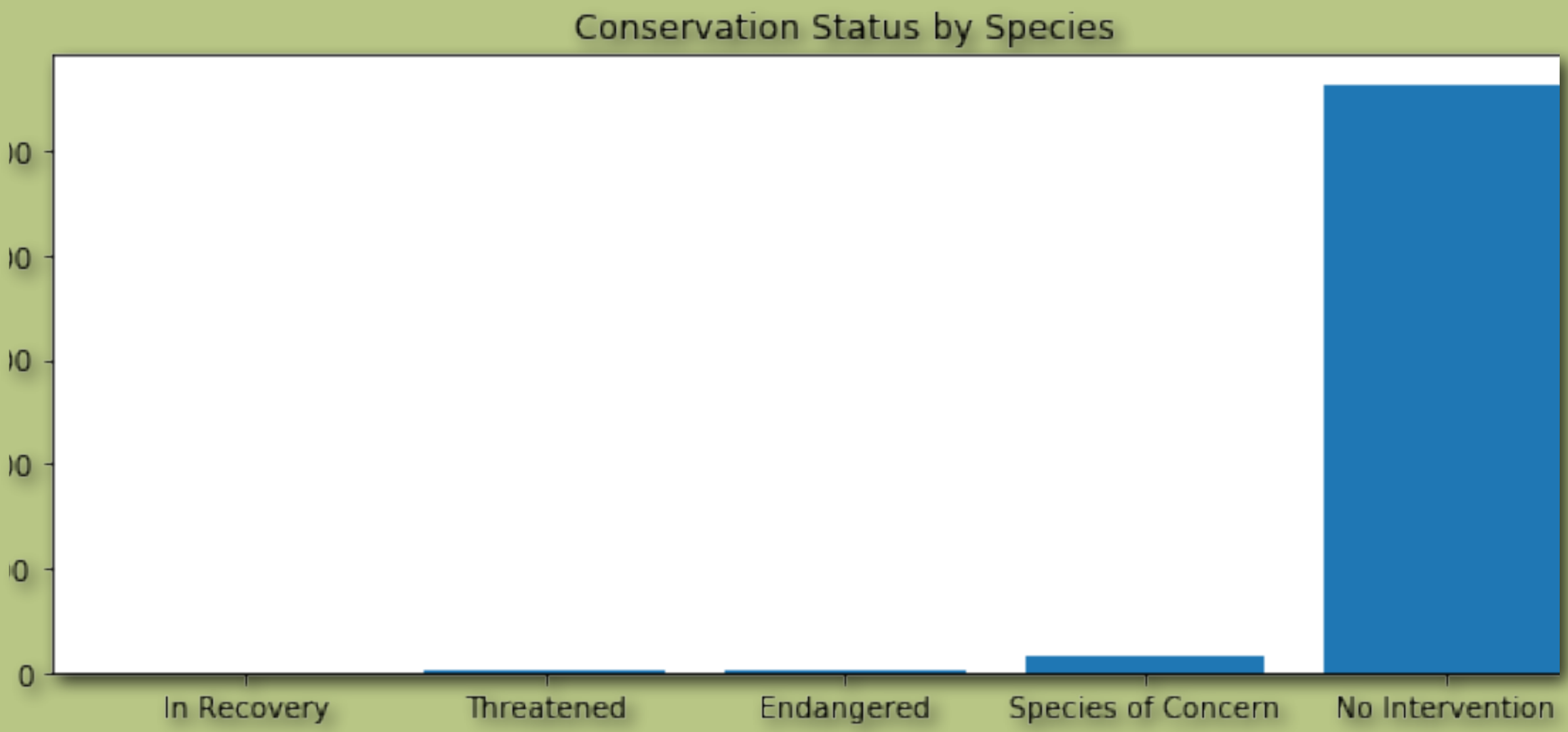
BIODIVERSITY PROJECT

OUTPUT OF CATEGORY_COUNTS.HEAD()

	category	is_protected	scientific_name
0	Amphibian	False	72
1	Amphibian	True	7
2	Bird	False	413
3	Bird	True	75
4	Fish	False	115

BIODIVERSITY PROJECT

CONSERVATION STATUS BY SPECIES



BIODIVERSITY PROJECT

SECOND ANALYSIS: THE SPECIES DATASET

- The second dataset is a CSV file, “observations.csv” that contains data about recorded sightings of different species at several national parks for the past 7 days.

- There are 3 columns: “scientific_name”, “park_name”, and “observations”.

- There are several different scientific names for different types of sheep.

- We will use both the “species” dataset from the first analysis and the “observation” dataset.

BIODIVERSITY PROJECT

ABOUT THE OBSERVATIONS DATASET

- The CSV file was loaded into Python as a Pandas data frame.
- There are 452 rows and 3 columns.
- The column names are : scientific_name, park_names, observations.
- Conservationists recorded sightings of different species at several national parks for the past 7 days.
- There are different scientific names for different types of sheeps.
- We need to know which rows of species are referring to sheep.

BIODIVERSITY PROJECT

INSPECTING THE DATA

- How many total sheep observations were made at each national park?
- Step 1. Save the observations file into a dataset named “observations” using the command: observations = pd.read_csv(‘observations.csv').
- Step 2. Use apply and lambda function to create a new column in the previous dataset “species” called “is_sheep”, using the command: species['is_sheep'] = species.common_names.apply(lambda x: 'Sheep' in x).
- Step 3. Many of the results are plants, so to isolate only the sheep use this command: sheep_species = species[(species.is_sheep) & (species.category == ‘Mammal')].
- Step 4. Now merge “sheep_species” and “observations” to get just observations of sheep using this command: sheep_observations = observations.merge(sheep_species).

BIODIVERSITY PROJECT

INSPECTING THE DATA

- Step 5. Use the “groupby” function to get the sum of “observations” for each park name. Save the answer to the variable name “obs_by_park”.

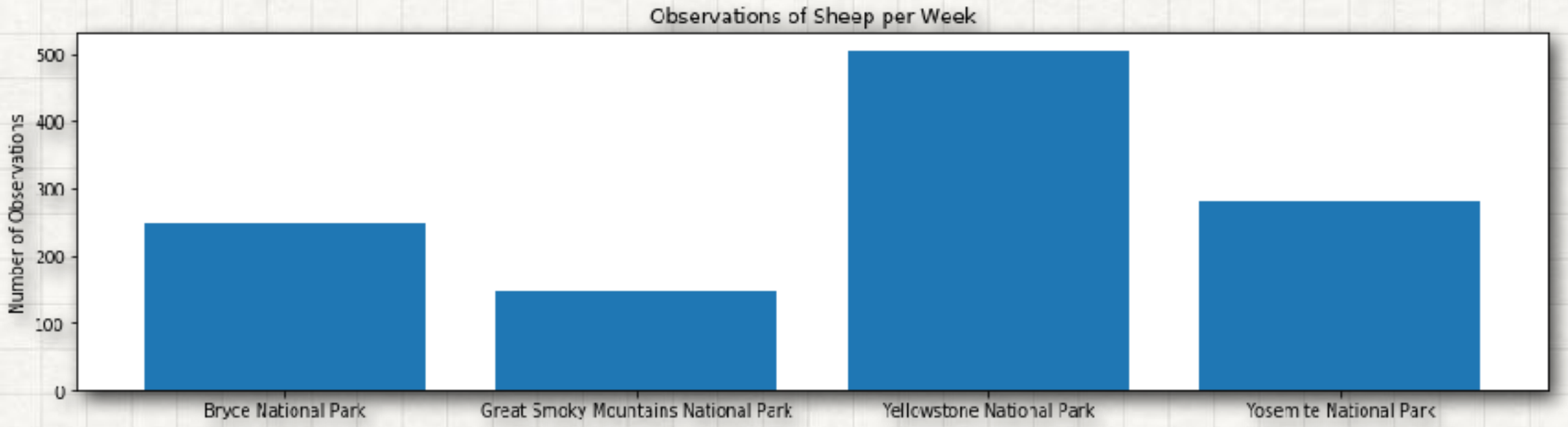
- Step 6. Use this command: obs_by_park = sheep_observations.groupby('park_name').observations.sum().reset_index()

- Here is the top 5 result:

	category	is_protected	scientific_name
0	Amphibian	False	72
1	Amphibian	True	7
2	Bird	False	413
3	Bird	True	75
4	Fish	False	115

BIODIVERSITY PROJECT

HISTOGRAM OF OBSERVATIONS



BIODIVERSITY PROJECT

INSPECTING THE DATA

- How many weeks would scientists need to observe enough sheep in order to observe with confidence of a big enough sample size of sheep?

- This was done to see if their program to reduce foot and mouth disease was working or not.

- A significance level of 90% was used.

- The baseline of the “Minimum Detectable Effect” was 15.

- The sample size was determined by the online calculator at optimizely.com.

BIODIVERSITY PROJECT

RECOMMENDATION

- Our recommendation is to observe sheep for 2 weeks in the Bryce National Park to observe enough sheep to test the program.
- We also recommend to observe sheep for 1 week in the Yellowstone National Park to observe enough sheep to test the program.
- The scientists will then be able to reach their goal of reductions of at least 5 percentage points.
- We can make this recommendation with 90% confidence.