**ETL Report**

By: Serena Baker, Chelsea Chaussee, Cindy Jones, and Lauren Toothaker

**Extract:**

With spring around the corner, we wanted to look at parks and wildlife. We found a biodiversity of national parks csv from Kaggle (<https://www.kaggle.com/nationalparkservice/park-biodiversity>) and historical reservation data for fiscal year 2020 from Recreation.gov (<https://ridb.recreation.gov/download>). \*\*\* The reservation csv is exceptionally large and cannot be pushed to github. We each downloaded the file in order to look through the data\*\*\*

**Transform:**

Starting with the parks.csv we read in the csv file and created an empty list to store our dictionary values. We kept all columns from this file.

Text

Description automatically generated

The 2020 recreation data csv is too large to be pushed to github. The columns we kept were park, facility latitude and longitude, nights stayed, and which state the facility is in.

Text, letter

Description automatically generated

For the species.csv we kept park name, category, scientific name, common names, nativeness, abundance, and conservation status columns. As the previous csv’s we uploaded the data in similar fashion into Mongodb.

Text, letter

Description automatically generated

**Load:**

Each data source file has been created to be housed in it's own unique Mongo database.

Latitude and longitude coordinates could not be used as primary keys due to some referencing the welcome center of the park, while others pointed to a campsite within the park.

**Summary:**

If we were to rework our files, a more usable form of this data would have been creating one Mongo database with multiple collections, one for each of the source datasets. Setting up the data this way would provide a more functional way to answer our original question: What parks could you visit to see certain animals?

Ideally, we would be setting up a collection that holds the unique Park Names and an array of all the species that can be found there. Additionally, we would build a collection of unique species and an array of all of the parks that they can be found in. Our third collection would be a unique list of park names and an array of all the geocoordinates of reservable facilities to stay at.

With this set up, you could identify all of the places to go to see a specific species, or you could pick a park to visit and know ahead of time what species you might see. Using the geocoordinates, you can identify parks that might have a large number of visitors at one time, or maybe are less busy due to less reservable locations.