This ETL focuses on data from a Kaggle page on species diversity found in United States National Parks. The first dataset is a CSV file that holds data for all National Parks such as park name, state, and coordinates. The second dataset used is a CSV file that holds information on different species found in the parks. This file includes information such as different species scientific name, common name, and family.

The goal of this project is to combine the two so that presence of certain species can be examined by park. It will help determine the biodiversity in each park as well as show the location of certain species. By using a PostgreSQL relational database, the two tables can be joined and queries can easily be made to explore the complete set of species information with the associated parks.

The two CSVs used contain easily readable columns with complete data. They were imported into pandas dataframes in the Jupyter notebook to clean them up and transfer them to the PostgreSQL database. The column names had to be cleaned in order to facilitate entry into the database. For example, “Species ID” was updated to “species\_id”. The index of each dataframe was also reset before being entered as a table in the database.

After both tables were added to the database, they could be easily joined to explore the data. Both datasets had a column for park name, so they were joined on that column. A view was created in SQL to combine them, taking relevant columns from each table. In the end, each record of a species in a park can be viewed revealing all species information along with the associated park. Sample queries are included in the SQL file along with the created view. For example, all plants belonging the family Fabaceae while located in Mt. Zion can be viewed at once. Querying a specific species will now show all instances of this species throughout all the parks, or a park can be queried to show all its species.