**PROJECT REPORT**

**1. EXTRACT:**

**Data sources:**

* National Park Trails: Kaggle -<https://www.kaggle.com/planejane/national-park-trails>
* Biodiversity in National Parks: Kaggle (Includes parks and species datasets) -<https://www.kaggle.com/nationalparkservice/park-biodiversity>
* US National Parks Visitation 1904 - 2016 (with boundaries):  Data World -<https://data.world/inform8n/us-national-parks-visitation-1904-2016-with-boundaries/discuss/national-parks-totals-by-year/g44tenrr>
* North American Bear Attacks: Data World

https://data.world/ajsanne/north-america-bear-killings

**2. TRANSFORM**

Dependencies used:

# Import libraries

# import pandas as pd

# from sqlalchemy import create\_engine

# from sqlalchemy import inspect

# from sqlalchemy.orm import Session

# from ast import literal\_eval

**National\_Parks dataset**

·         Renamed columns

·         Removed duplicates

·         Reset index

·         Renamed index *id* in order to use it as a primary key in our database

**Trails dataset**

• Dropping rows that contain the parks listed below:

Clayton Co International Park, Jonesboro GA, Fort Hunt National Park and Wolf Trap National Park for the Performing Arts

• Created dictionary to replace some of the parks’ names

• Created three different datasets: activities, features and trails

• Renamed columns in each dataset

• Reset index in each dataset

**Species dataset**

·         Dropped unwanted columns

·         Kept species that are currently present in parks

·         Renamed columns

·         Reset index

·         Renamed index *id* in order to use it as a primary key in our database

**Parks visitations dataset**

·         Grabbed only National Parks from the dataframe

·         Filtered out rows that do not contain a numeric year

·         Created filtered dataframe with specific columns

·         Renamed the column headers for clarification

·         Reset the index

·         Renamed index *id* in order to use it as a primary key in our database

**Bear attacks dataset**

·         Filtered dataframe for Canada’s and US National Parks

·         Checked dataframe for duplicates

·         Reset index

·         Used *split* function to change the parks columns in order to make sure that the values match other dataset

**3. LOAD**

·        We each created our tables’ schemata.

·        We combined them into one schema.

·         We combined our jupyter notebooks

·         We loaded our tables into the database

**Why did we choose these datasets?**

·         Common interest

·         We thought it would be interesting and challenging to combine the different datasets.

·         We also thought that once in the database, we would be able to perform different queries by joining different tables. (As shown in the presentation).