[](https://bootcampspot.com/)

# **Student**: Reynalyn Gaoaen

# **Project**: ETL Project

# **Timeframe**: May 4, 2019 to May 9, 2019

## PROJECT DESCRIPTION:

Create a database that includes a list of 2019 first day hike recommendation from the California Department of Parks and Recreation website and a log of hikes that you have personally completed.

## EXTRACT:

Source #1: [www.parks.ca.gov/?page\_id=28055](http://www.parks.ca.gov/?page_id=28055)

Use BeautifulSoup to scrape the list of parks and website links from the California Department of Parks and Recreation website

* + Start by querying for id = “center\_content” in a “div”
  + Then further query for data with ‘a’ tag
  + There are multiple links in the website that are not related to parks, further clean the data by querying only for links where the first 25 characters starts with “https://www.parks.ca.gov/” and the text of the link starts with “https”
  + Append the data to a dictionary

Source #2: Resources/parkshiked.csv

Use python to upload a data from a csv file. Note that this csv file is your own personal log of hikes that you have done. (To keep the scope of this project small, it is assumed that the list of parks in your personal log is the same as the list of parks in the California Department of Parks and Recreation website)

## TRANSFORM:

Source #1

* Create a pandas DataFrame to house the data.
* Clean the park column by removing the semi column at the end of each value
* Drop the record does not have website link. Note that one of the record is a link to PDF file.

Source #2

* Create a pandas DataFrame to house the data.
* If the value under the comment column is blank, change it to “future hike”

## LOAD:

Upload the data into tables in sqlite. The expected data are all string and in tabular formats which is perfect for a sql database.

Type of database: sqlite

Database name: CalHikes

Tables:

* Source #1: name: park. Columns: “index” integer primary key, “park” string, “link” string
* Source #2: name: hike. Columns: “index” integer primary key, “park” string, hikedt string, comments string