**ETL Project - Data Analytics Bootcamp - 14er Hiking and Camping**

**Motivation for gathering the data:**

* + Summertime is approaching, which is hiking season here in Colorado. We wanted to have a dataset ready for users to be able to plan their trip into the Rocky Mountains to hike one of the 58 mountain peaks that are 14,000 feet high or higher (called a 14er). We wanted inexperienced or visiting hikers to have a dataset where one stop planning could occur.

**Specific use cases for the data:**

* + The use case of the data is to combine datasets, so that any person can go camping and hike a 14er in a Colorado mountain range. There are 58 14ers in Colorado and we wanted to provide the user with a dataset that will give distance from a campground to a 14er, basic information about the 14er and the most frequently used trails, and trailheads by which it can be accessed. In addition, we have provided road type to ensure the user has the correct vehicle to access the trail head.
  + We utilized a hard coded 20 mile distance. We felt this was an ideal distance for driving from the campground to the trailhead. If a user interface were developed, the user could adjust the distance input to determine campsites within various distances, but at this point it is hard coded.
  + A user will be able to access:
    - Campground Name
    - Distance from Campsite to Mountain Peak (miles)
    - Trailhead
    - Road Type
    - Mountain Range
    - Elevation\_ft
    - Peak Latitude
    - Peak Longitude
    - Standard Route
    - Standard Route Distance (miles)
    - Elevation Gain\_ft
    - Difficulty

**The sources of data we used to extract was:**

* + Southwest\_camp\_data.csv from <http://www.uscampgrounds.info/takeit.html> to which we added headers.
    - Columns used: Long, Lat, Name
  + 14ers\_data.csv from <https://www.kaggle.com/mikeshout/14erpeaks>
    - Columns used: Mountain Peak, Mountain Range, Lat, Long
  + 14ers trailhead by road difficulty scraped from <https://www.14ers.com/php14ers/trailheads_bydifficulty.php>

**How we manipulated/transformed the data:**

* + To scrape the trailheads from 14ers.com:
    - We scraped 14ers trailhead by road difficulty tables from the website noted above, renamed columns headers, and added road types. Since the trailheads listed had multiple peaks that could be summited from, we then split mountain peaks into separate columns to have a dataset of unique trailheads and dropped duplicates. Next we combined the split mountains into one column by running a for loop which took the trailhead dataset and each mountain peak column, appending them to a list, concatenating the results, dropping duplicates and null values, and then resetting the index.
  + To determine distance from campground to mountain peak and create distance\_df:
    - We used southwest\_camp\_data.csv, and 14ers\_data.csv into a jupyter notebook using pandas. Then we constructed a for loop to calculate distances from the latitude and longitude of the campsites to the latitude and longitude of the mountain peaks and transformed the list results into a dataframe. Next we dropped duplicates.
  + We made all mountain peak names consistent in the various csv files by removing leading spaces and replacing “Mount” with “Mt.”, as necessary.
  + To arrive at the final database:
    - We used distance\_df, mountains, and camps and merged them into a final dataframe and renamed the columns as appropriate to make them consistent and more user friendly. The three dataframes were merged on Mountain Peak (the name of the 14er).
    - We decided to use pymongo to turn the data frame into a Mongo database because it was the fastest method and there was just one table.





