Author: Era Ebhodaghe
DSC 540

Final Project Summary

For this project, I worked on the National Parks data, implementing various data wrangling techniques to clean my dataset and combining all 3 data sets to create meaningful visualizations.

I performed multiple steps to clean the datasets to ensure that only the relevant data was kept for the data to be joined seamlessly. Some of the changes made on all data datasets include: dropping unwanted columns, changing header names, modifying datatypes. The most important transformation done, was to ensure that the unique id (park name) was the same across datasets by removing "National Park" at the end of the Park Name and also stripping white spaces before and after the string.

Building on my experience with CSV files, I expanded my skillset by learning web scraping with BeautifulSoup to process HTML data. Scraping posed challenges due to the web's organic growth, which integrates various technologies and formats. I also gained expertise in working with web APIs using the `requests` library. This allowed me to send HTTP requests, analyze response objects, and extract JSON data for analysis. By inspecting response headers and status codes, I could determine the success of the request.

I developed proficiency in data manipulation with Pandas, including filtering, sorting, merging, and joining data frames. These techniques were essential for cleaning and preparing data for analysis. I also learned to use popular visualization libraries like Matplotlib and Seaborn to create clear, informative charts that effectively communicated trends and insights from the data.

In this project, I gained hands-on experience with SQLite3 to create databases and tables. I also learned to use SQLAlchemy to interact with the database programmatically, enabling me to import data into Pandas data frames for analysis.

Based on this topic, there are no legal or regulatory guidelines. All information obtained is public and can be used to provide informative insights on US National Parks. One of my data sources was verified for credibility by pulling data directly from the National Parks API, a US government website. The other 2 sources were from credible organizations as well. For the Wikipedia data, I performed spot checks on the data using google searches to verify that the elevations shown for the Parks were indeed accurate.

The data was acquired in an ethical way from public datasets. In addition, I took precaution to ensure that the data obtained was not manipulated with any sort of bias in favour of a State, Park or any other specific variable.