Madeleine Sharp

DSC540- Final Term Project

Write-Up - Summary

## Introduction:

National Parks are a precious source of natural land and life and are one of the United State's best efforts with respect to conservation. As someone who appreciates National Parks, I wanted to learn a bit more about them, thus selecting this for my term project topic. I obtained data from a variety of sources and source types and utilized that data to discover a few insights about National Parks in the US.

## Ethical Implications of Cleansing Data:

A large portion of this project centered on cleaning the various data sources that we obtained. The three data source types used within my project included flat-file sources, web sources, and an API source. Each of these various sources required its own cleansing steps via a variety of means and methods. For example, the ways in which I handled and cleaned the flat-file sources were not necessarily the same ways in which I handled my other two sources.

While data cleansing may be time consuming and energy intensive, to go without would be a sore misstep. If data is in incorrect format, corrupted, duplicated, irrelevant, missing, or full of outliers, this can most certainly impact the outcomes of any analysis that uses this data to obtain insights. Therefore, if data is not cleaned appropriately and handled with care, the outcomes of research, analyses, models, etc. that use that data would be inaccurate – thereby misleading anyone who consumes that information.

Given this, there is certainly a reason why those within the data field mention that data cleansing takes about 80% of the entire project process – it takes time and effort to ensure that your data is cleaned appropriately, and for good reason. Anything less than that might be a sloppy job, or worse, an inaccurate and misleading one. While data cleansing can be frustrating at times, upon completion the feeling of accomplishment for handling your data with care is unmatched.

## Summary of Steps:

Overall, to complete this project, I had to embark through the following steps:

- Identify project topic
- Locate data sources (flat-file sources, web sources, and an API source)
- Import that data
- Cleanse and transform the data for each source
- Load the data into an SQL database (as 3 tables)
- Merge those 3 loaded tables together
- Query that data, convert to a pandas dataframe, and then utilize to create 5 visualizations

**Overall Learning:**

Overall, this project taught me how important it is to ensure you have clean data, and to ensure that, if you plan to use the data together or merge it, each data source should have a variable that can be its relational variable. Having a relational variable (in the end, mine was park name) allows you to tie your data sources together by assigning/bridging that relationship.

I also learned how imperative it is to be comfortable to work with different data source types – not just individually, but in tandem. I found that was quite "real-world," and helpful to have more practice with. I know at my place of work, multiple data sources needing to be used together has been a regular occurrence.

From a project topic standpoint, I learned that a ton of data existed for National Parks, but that data was not always aligned with other sources – some data sources had all of the National Parks in addition to National Monuments, Forests, etc., and some sources did not necessarily even have all of the parks.

The visualizations from my project indicated to me the following:

- From an acreage standpoint, quite a variety exists between all National Parks.
- California has the most National Parks of any state.
- The elevation of the National Parks did not follow a normal distribution curve.
- The National Park with the largest number of visitors by far was the Great Smoky Mountains.
- The majority of species class across all National Parks is vascular plants.

I enjoyed this project and I truly feel it pushed me to grow and learn new things. The skills and experience obtained from this course are things I can with me for the road ahead.