**Milestone 5 learnings Journal**

Pradeep Kumar Vallepalli

Bellevue University

DSC 540 – Data Preparation

Professor. Catherine Williams

August 13, 2022

This project has significantly increased my knowledge of python and how to handle data from various data sources by presenting multiple challenges throughout the project. Instead of using the provided flat files for this project, I was able to get the data from various sources. Beginning with cleaning a flat file in Milestone 2, through each milestone, I improved my data cleaning skills by fixing null values, formatting problems, and handling outliers in the data and other issues that needed to be fixed to get the data ready for merging. Web scraping, which is covered in Milestone 3, was a new skill I learned through this project and this course. I extracted the population data from a table on the world population review page using requests and beautifulsoup modules in python. Then I converted the results into a Pandas data frame that I could use to work with the data. One challenge I came across with web scraping was identifying a website that doesn't generate the table data dynamically using JavaScript.

Gaining more comfortability with APIs was also very helpful as most of the applications and sites nowadays were internally driven by APIs to get data. In Project Milestone 4, I had to work through a challenge of API rate limits when I started a lookup of reverse geolocation through API. The site blocked me due to making several API requests too frequently. Eventually, I found another site that has a better API performance and doesn't block US based on the number of API requests. Although I found a working API, due to the large size of my flat file dataset, it took me around 6 hours to fetch the ~60,000 records and pull it into a data frame for data cleaning. I believe that my knowledge of SQL has significantly increased throughout the course of this term.

I learned more about the ethical considerations that go into every data science project. In any data science project, we must consider multiple things, such as whether the data is ethically sourced, whether the data cleaning will end up with misleading information or if the results end up providing a negative impact on specific demographics or shouldn't contain any data that can be used of identification. For example, when collecting the geolocation information, I excluded the exact address of the rental listing in the dataset to avoid any information that could be used for identification.