Torrey Capobianco

Asteroid Summary

Milestone 5

For the final project for the class, Data Preparation, I selected three different types of data sources and information about Asteroids. The first data source, the main collection of asteroids, was a CSV file downloaded from online. The preparation for this file was the first introduction to creating pandas dataframes. From there, I became familiar with mapping in replacing headers, converting yes and no values to Boolean, and filling missing values with a number. I found the importance of dictionaries in data preparation, and the value they have in the process of many data manipulation techniques. Learning about dictionaries in the intro to Python class cannot be stressed enough.

The second data source was from a website on the value and profit of asteroids if we were to mine them. Here, I learned how to scrape a website using BeautifulSoup. Two new manipulation methods I applied on this dataframe were regular expressions and binning. Regular expressions was used to find a series of letters in a column value and remove them from the value. There is more I need to learn and practice with regular expressions but was a great start. Binning can hold a lot of value in grouping items, in this case by the buckets the asteroids are worth: thousands, millions, billions, trillions, or over trillions.

The third data source required connecting to an API to request data on asteroids and their near Earth approach date. This was challenging in learning the JSON structure of an API, selecting which keys are necessary and unpacking the dictionaries to convert to a dataframe. Along with other processes in this project, there were times that I needed to convert columns to a pandas series to make modifications and then concatenate the series back with the main dataframe. Learning to drop columns was also in this process. The API source contained dates which allowed me to further practice converting columns to datetime and create new columns around the datetime features, such as day of the week and time the asteroid approached Earth. Being able to use the datetime function in this project helped further my understanding of this technique aside from the class exercises.

The last piece of the project was to merge the three dataframes into one based on a common key and store in a database. Due to needing a common key, each data source had a different way on how they identified the asteroid, whether it be the provisional designation with the year it was discovered, the order number it was discovered, or the alphabetical name it was given (very few are given a name). Through converting to a pandas series, splitting columns, lamanda functions, and concatenation, a common key name was created for all three sources. I then learned how to create a SQLite database, import the three dataframes as CSV files, convert each to a SQL table, and join all three on the designated primary key. Several visualizations of the data was explored through the use of Microsoft PowerBI, with the relationship between all three tables.