Capstone Project 1: Data Wrangling

Data was obtained from Kaggle. https://www.kaggle.com/nasa/solar-eclipses

What kind of cleaning steps were performed?

Steps were taken to clean two csv files: solar.csv and lunar.csv

Solar.csv - This file includes data about solar eclipses.
Lunar.csv- This file includes data about lunar eclipses.

1. Imported csv files as panda dataframes and reindexed the rows with the column 'Catalog number' for both of the files.

2. After looking at the format of the dataframe, I created separate dataframes to include only the necessary columns for analysis.
    a. Solar now has the columns:
        i. 'Calendar Date', 'Eclipse Time', 'Delta T (s)', 'Eclipse Type','Eclipse Magnitude', 'Latitude', 'Longitude', 'Sun Altitude', 'Sun Azimuth', 'Path Width (km)', 'Central Duration'
    b. Lunar now has the columns:
        i. 'Calendar Date', 'Eclipse Time', 'Delta T (s)','Eclipse Type', 'Latitude', 'Longitude', 'Penumbral Eclipse Duration (m)', 'Partial Eclipse Duration (m)','Total Eclipse Duration (m)'

3. The data has also been reorganized by Calendar Date in ascending order.
    a. Converted the dates into a list of [year, month, day] for easier analysis of trends.
    b. Note: The years have negative values because the data is during the five millennium period -1999 to +3000 (2000 BCE to 3000 CE).

4. Missing Values
    a. Missing values appearing as '-' was replaced with NAN values when imported as dataframes.

5. No outlier data found