Initial data quality review of the CSD streetlight location data set

I chose to explore the https://data.sandiego.gov/datasets/streetlight_inventory/ to look for insights on where future LED upgrades could be most beneficial. After reviewing the data set, I shifted my efforts to focus on key locations with missing or incorrect data. It is important to know the quality of a data set before using those data for planning, policy, etc.

Findings/recommendations:

- If this is the most current data set the City has, I'd recommend thorough examination and validation. I might caution using streetlight_locations_datasd_v1 for planning purposes until the dataset is updated.
- I identified at least 117 duplicated data rows
- All rows contain latitude and longitude values, but most rows are missing some metadata
  - About 20% of the rows are missing the SAP ID
  - About 20% of the rows are missing streetlight types making it difficult to decide which lights should be retrofitted
- There is a group of streetlights that map to Carlsbad, outside the City limits
- There may be one streetlight that maps south of the US border, and there might be one streetlight that is just east of the City border

Assumptions:

- The data set is old (2017) so it might not be the most current version. The same QA could be applied to the most recent data if available.
- The likelihood of large datasets containing errors is >0
- There are often common patterns of errors to look for
- Spatial data also can be QA'd visually using interactive mapping

Methodology/approach:

- I analyzed the data set and plotted some results using R
  - For a detailed, step-by-step record of my methodology, please refer to the code
- To get an understanding of the quality of these data I looked for common data errors
  - Missing values
  - Duplicate or repeated data
  - Values that are out of range (e.g. lamp voltage <120V or >240V)
  - Spatial datasets like this one often have incorrect lat/lons associated with the asset
  - Misplaced or mis-assigned data (e.g. longitude values in the latitude column)
- The data set is large (>60k rows) and most rows were missing some data, so I decided to focus on rows most in need of updating – those missing SAP IDs and location descriptions
  - This subset of the lowest-quality data was about 16% of the entire dataset >9500 rows
- With these data, I could plot out the streetlight locations on an interactive map to allow internal City stakeholders to identify where resources should be focused to update the main dataset. Please see: https://nickharing.shinyapps.io/streetlights/.