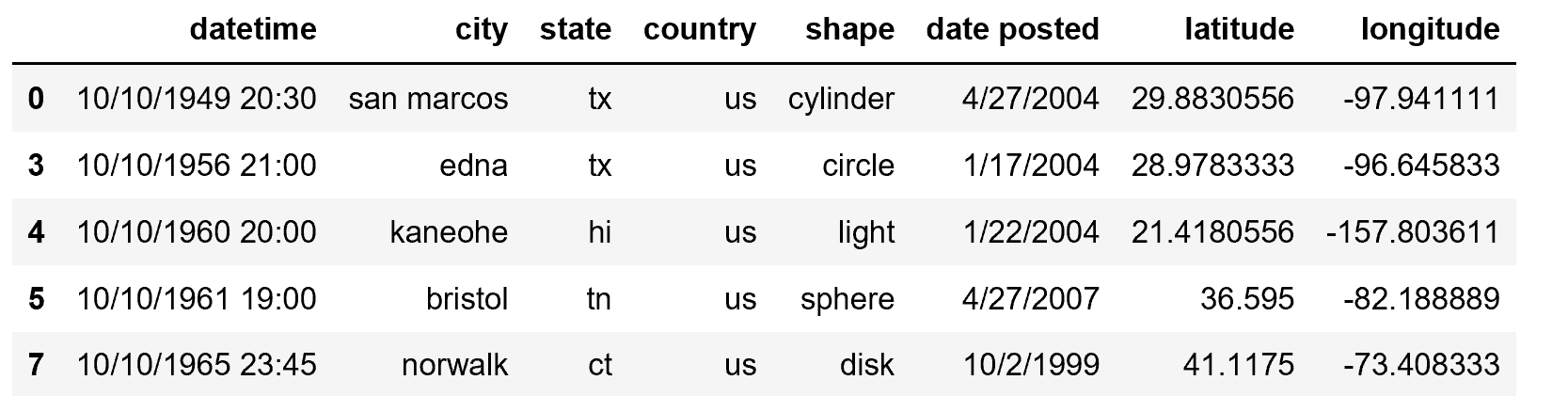
Extract:

For this project we decided to do an analysis of UFO Sightings and compare this with the location of military databases in the United States. For the analysis we used two csv files imported from Kaggle and Open Data Soft. Both files were imported into Jupyter notebook for analysis and transformation.

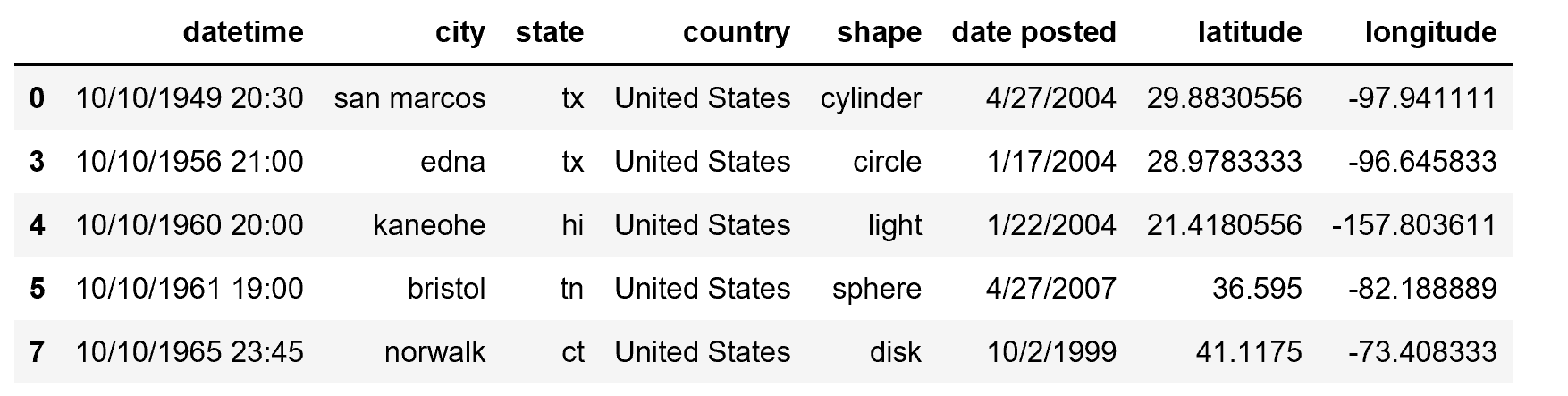
* Data Source: Military Bases (<https://public.opendatasoft.com/explore/dataset/military-bases/export/>)
  + Format: CSV
* Data Source: UFO Sightings (<https://www.kaggle.com/NUFORC/ufo-sightings>)
  + Format: CSV

Transform:

Our first steps in cleaning up the datasets involved figuring out which variables were not relevant. For the UFO sightings data set we dropped the duration (seconds), duration (hours/min) and comments columns. After dropping the columns we filtered by country selecting the US only.



After this we dropped duplicate rows with and rows with NA values and renamed the country from “us” to “United States”.



Moving on to the military bases file 8 columns were dropped and renamed the remaining columns to Base Status, state, Base Name, Base Type, latitude, country and longitude. We filtered the remaining information by country keeping only data for the United States. After this we dropped duplicates keeping only the first row of each “Base name”. This was done to ensure we didn’t have duplicate coordinates for the location of each base. Finally NA values were dropped to ensure only rows with information were kept.



For the analysis two different methods were considered to join the tables, one calculating the distance between the UFO sighting and the military base and using the result as the joining element and the other using the state and coordinates to add an additional column for the city for both tables using this new column as the joining element.