Jake Thorson, Emilee Althoff

Professor Karonga

May 5, 2022

Data Analytics

Gathering Data:

Source: https://www.kaggle.com/code/dhruvkalia/ufo-sightings-dbscan/data

The data is related to UFO sightings across the world, created 2 years ago by user Dhruv Kalia. It has 11 columns relating to locations, times, and types of UFOs that were sighted. We chose this dataset because it has a large variety to work with, so creating charts will be easier. It also has an appropriate number of rows to work with, being only about 30 MB. Problems relating to this dataset that we are trying to solve are:

- Are UFO sightings correlated to geographic locations around the world?

- Are UFO sightings correlated to geographic locations within the US?

- Are UFO sightings correlated with time?

- What is the best number of clusters to use in a KNN analysis?

Preprocessing:

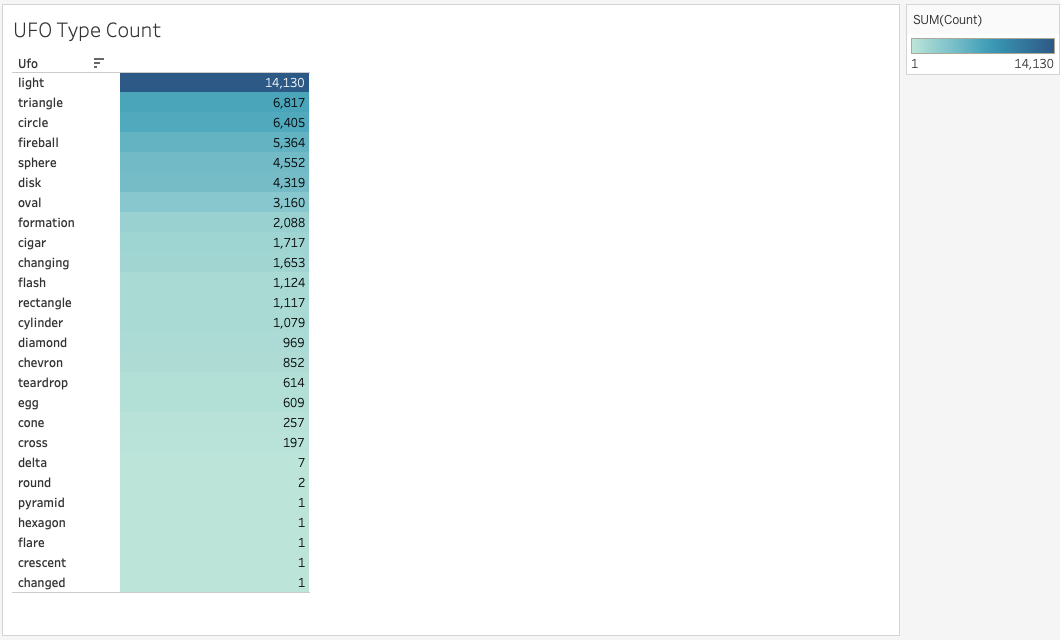
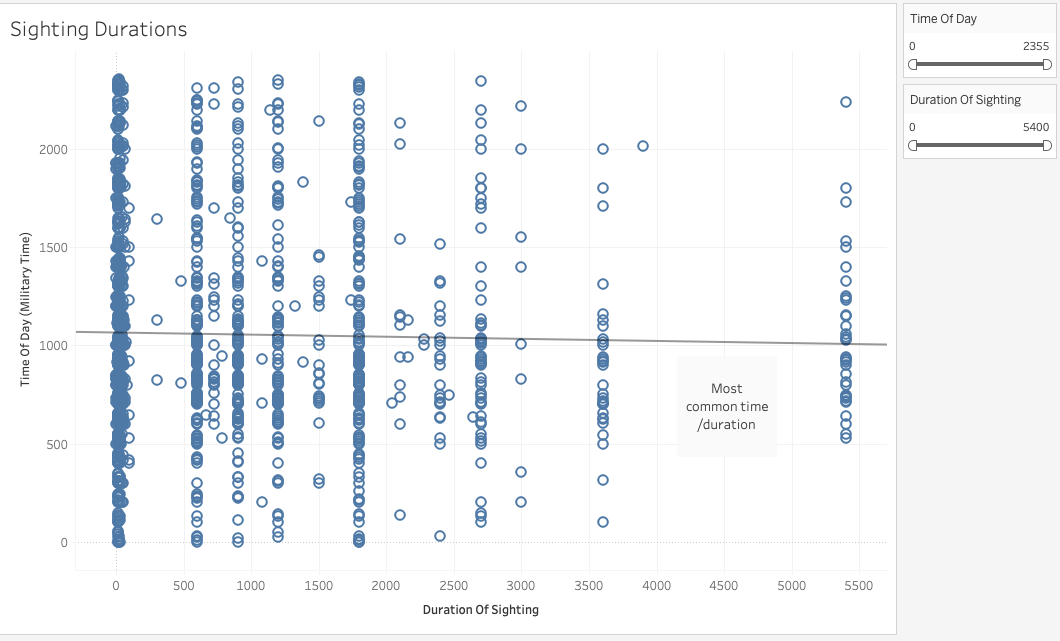
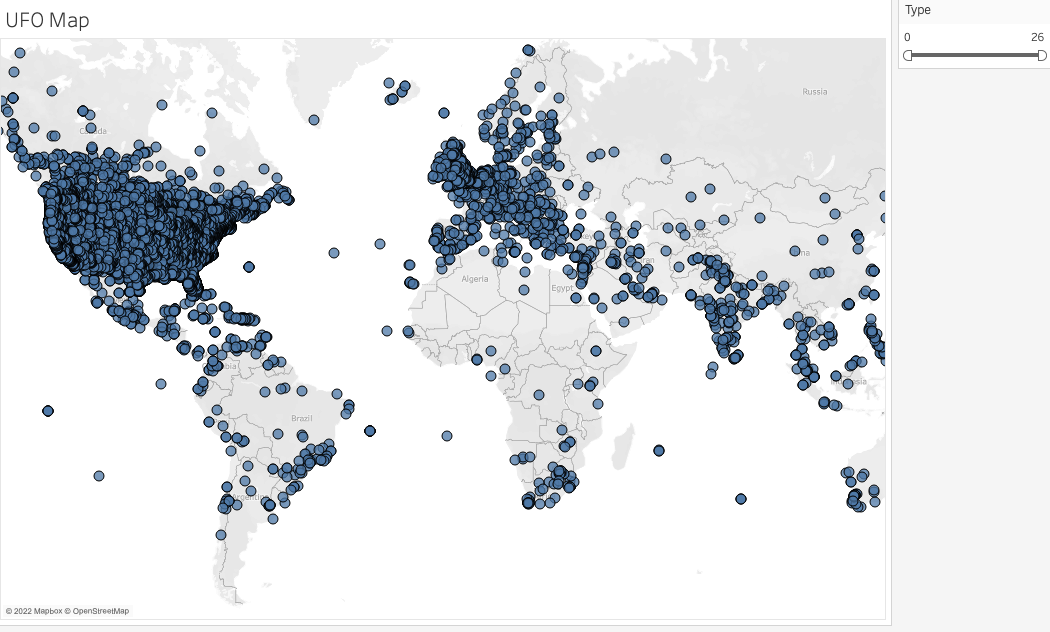
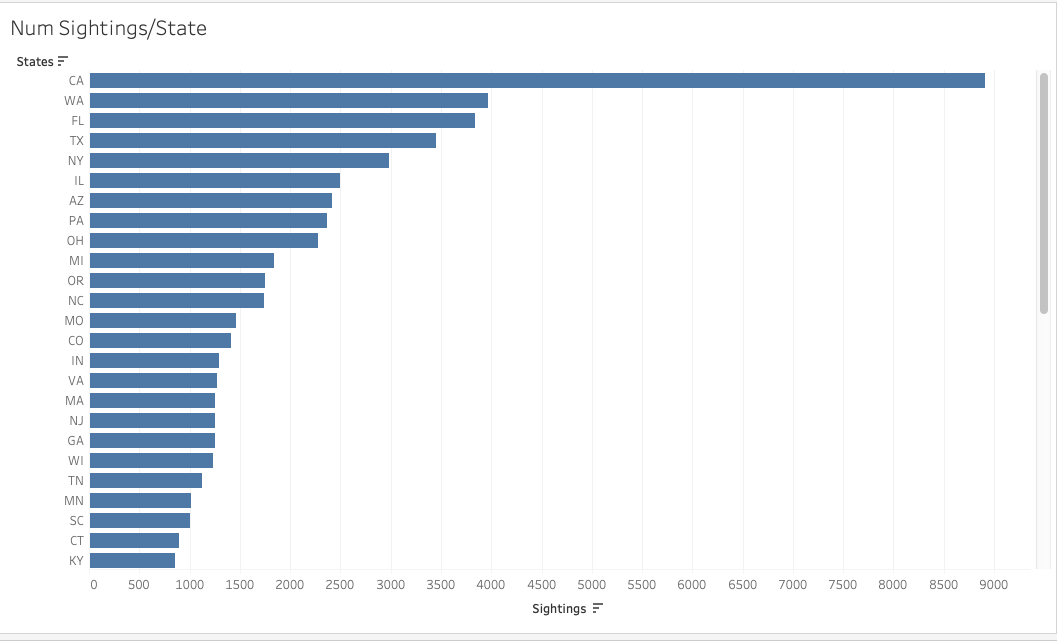
We began by changing data types, column names, and dropping unwanted values. Features such as latitude, longitude, and time should be in integer form, and the column names should be compatible with other python codes. We approached creating data specifically for charts by creating dataframes acting as subsets of the dataframe imported from Kaggle.

We used location data across the world for our first chart. We decided to replace the strings with integers and refer back to the code if we want to retranslate. This means in the future the data will be easier to work with if any machine learning techniques are to be used. We omitted unwanted values such as unknown, other, or NA.

We decided against using the csv file containing the KNN results because it did not fit with the rest of the graphs. The other graphs are relating to where and when to find a UFO, so a better fit would be a graph showing the most common type of UFO appearing. We omitted NAs and unwanted values, then ran a loop to count the number of occurrences for each different type of UFO to preprocess the data.

The last dataframe is time of day and duration of sighting. The duration of sighting column was not in a clean format, and was a combination of strings (seconds, hours, minutes) with values both as integers and strings (a few, a couple, etc). This required extensive preprocessing to convert the column in integers in a minute format. We combined it into a single dataframe with time of day in military time and UFO type. The last step was changing the times to begin at noon which makes the graph look easier, as the most common occurrences are late at night or early in the morning.

Tableau Charts



Division of Labor

We both searched through datasets from Kaggle and tried preprocessing a few but ended up using the UFO data because it had a large variety of features and was easy to work with. Jake took a bit more of the responsibility for preprocessing, and Emilee took a bit more of the responsibility for the Tableau charts.