Alien Lights in California

Chloe McCaffrey, Matt Plumb, Nic McClellan, Skyler McMullen

CSCI 403: Database Management

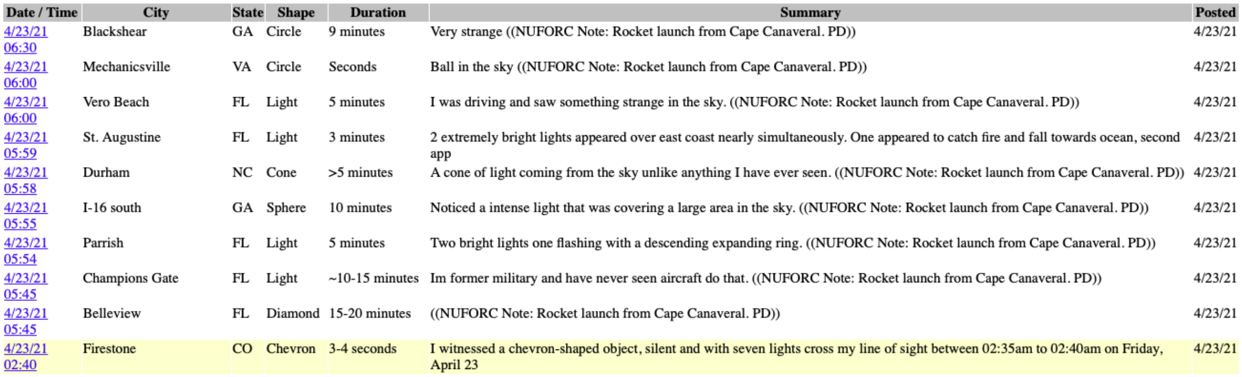
Introduction:

Our group found an entertaining dataset on reported UFO sightings in the United States. This data is extremely interesting because recently there have been many recent UFO sightings in the news, and so it was fascinating to learn more about the numbers and types of sightings that have been reported historically. There was quite a bit of information that could be extracted from this data, such as the location of the sighting, times, dates, and various information from the witnesses. Based on the data manipulation that we performed, it was discovered that UFO lights are most likely to be seen in California.

About the Dataset:

The data set was obtained from a website called National UFO Reporting Center. The data set is public and there are no licensing restrictions on this information. This data set had a collection of data entries ranging from 1561 to April 2021. There are historical accounts that have been input from prior to 1561, however these entries seem to have been input incorrectly. There are not many of these ancient historical accounts that have been added to the database, but they are interesting to look at. There have been regular reports starting in the mid 1900's. The dataset has records of sightings that occurred monthly in 1950 onward. The online database is divided into monthly reports. Each account is broken into columns such as Date/Time, City, State, Shape, Duration, Summary, and Posted date. A sample of the information that is displayed on the website can be found in Table 1.

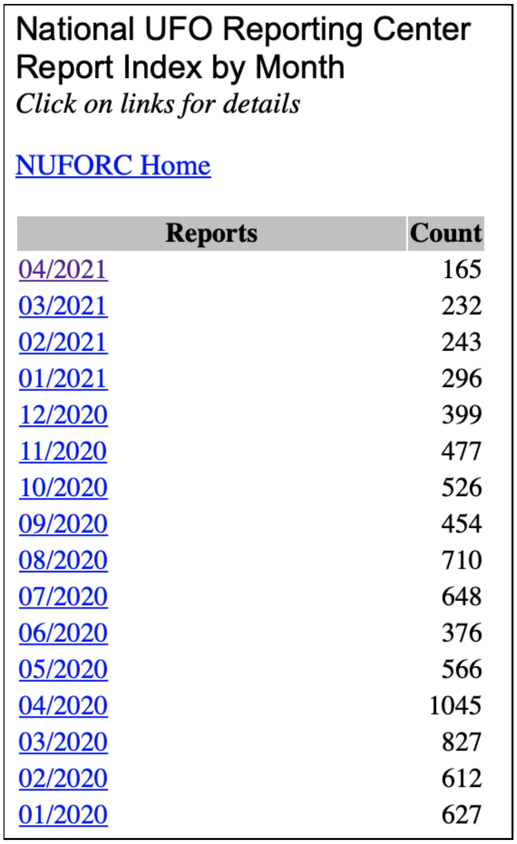
Table 1: Sample of the online database



Date / Time	City	State	Shape	Duration	Summary	Posted
4/23/21 06:30	Blackshear	GA	Circle	9 minutes	Very strange ((NUFORC Note: Rocket launch from Cape Canaveral. PD))	4/23/21
4/23/21 06:00	Mechanicsville	VA	Circle	Seconds	Ball in the sky ((NUFORC Note: Rocket launch from Cape Canaveral. PD))	4/23/21
4/23/21 06:00	Vero Beach	FL	Light	5 minutes	I was driving and saw something strange in the sky. ((NUFORC Note: Rocket launch from Cape Canaveral. PD))	4/23/21
4/23/21 05:59	St. Augustine	FL	Light	3 minutes	2 extremely bright lights appeared over east coast nearly simultaneously. One appeared to catch fire and fall towards ocean, second app	4/23/21
4/23/21 05:58	Durham	NC	Cone	>5 minutes	A cone of light coming from the sky unlike anything I have ever seen. ((NUFORC Note: Rocket launch from Cape Canaveral. PD))	4/23/21
4/23/21 05:55	I-16 south	GA	Sphere	10 minutes	Noticed a intense light that was covering a large area in the sky. ((NUFORC Note: Rocket launch from Cape Canaveral. PD))	4/23/21
4/23/21 05:54	Parrish	FL	Light	5 minutes	Two bright lights one flashing with a descending expanding ring. ((NUFORC Note: Rocket launch from Cape Canaveral. PD))	4/23/21
4/23/21 05:45	Champions Gate	FL	Light	~10-15 minutes	Im former military and have never seen aircraft do that. ((NUFORC Note: Rocket launch from Cape Canaveral. PD))	4/23/21
4/23/21 05:45	Belleview	FL	Diamond	15-20 minutes	((NUFORC Note: Rocket launch from Cape Canaveral. PD))	4/23/21
4/23/21 02:40	Firestone	CO	Chevron	3-4 seconds	I witnessed a chevron-shaped object, silent and with seven lights cross my line of sight between 02:35am to 02:40am on Friday, April 23	4/23/21

The creators of the website have split up the data according to month, which is demonstrated by Figure 1. This made it difficult to manually access the information because each month would have had to be individually accessed. With the large and divided dataset, it was not feasible to manually access all of the reports. In order to attain the information, the group developed a web scraping algorithm to grab the data and then cleaned the data by reformatting cells. A sample of this algorithm can be seen in Figure 2. This was particularly used for populating NULL in empty cells, and making the time duration a uniform format. After the data was scraped, there were

99704 rows of data that we were able to work with. This dataset was uploaded to the chloemccaffrey schema and the table is called final_project. It was also uploaded to the skylermcmullen schema and the table in this schema is called final_project.

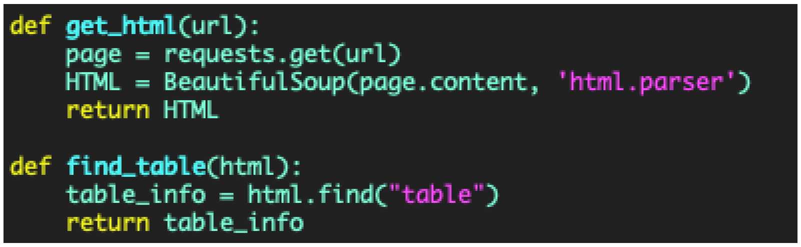


National UFO Reporting Center
Report Index by Month
Click on links for details

NUFORC Home

Reports	Count
04/2021	165
03/2021	232
02/2021	243
01/2021	296
12/2020	399
11/2020	477
10/2020	526
09/2020	454
08/2020	710
07/2020	648
06/2020	376
05/2020	566
04/2020	1045
03/2020	827
02/2020	612
01/2020	627

Figure 1: Quantities of monthly UFO Reports dating back to 2020



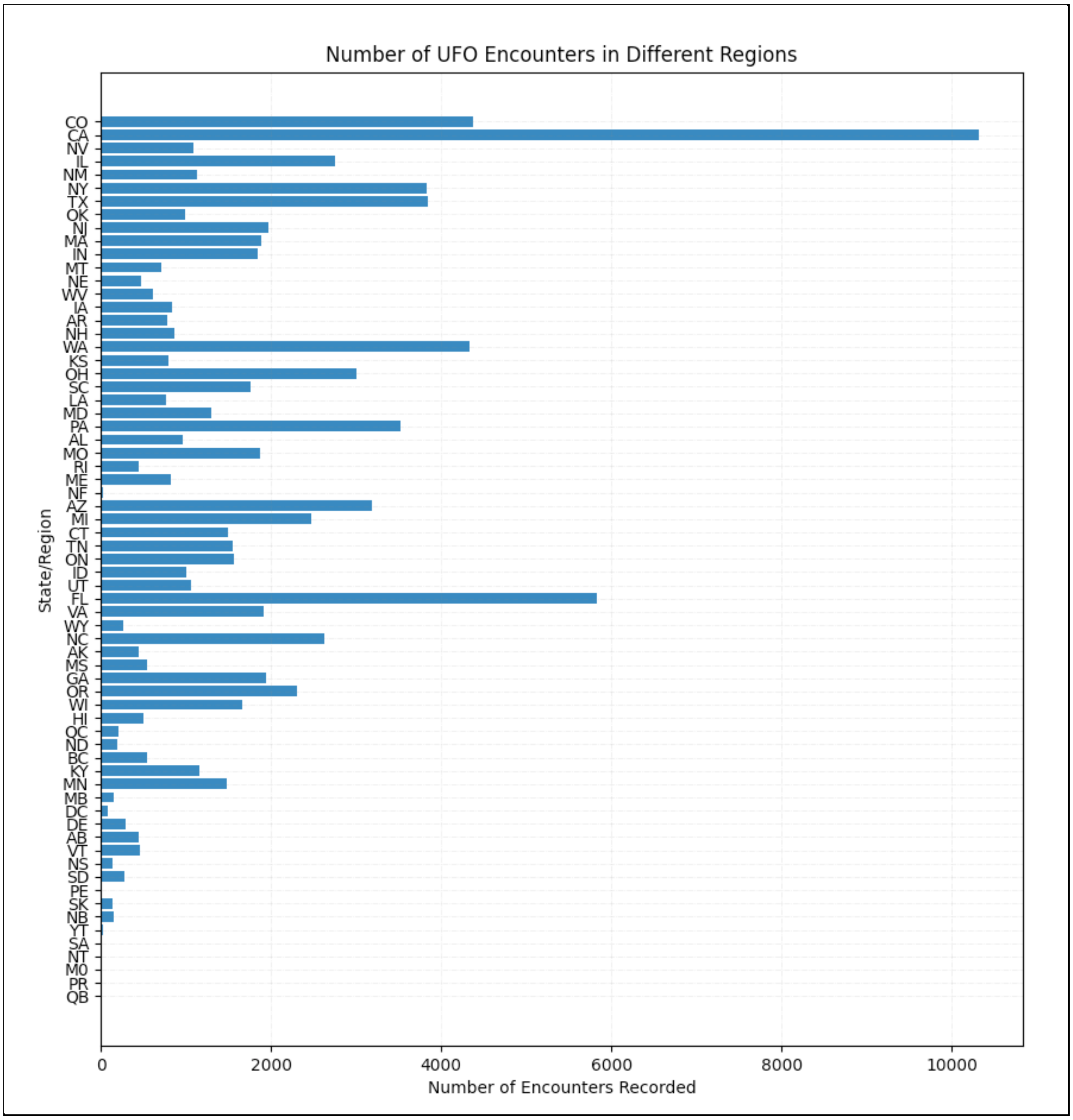
```
def get_html(url):
    page = requests.get(url)
    HTML = BeautifulSoup(page.content, 'html.parser')
    return HTML

def find_table(html):
    table_info = html.find("table")
    return table_info
```

Figure 2: Sample of the web scraping algorithm used to obtain the data

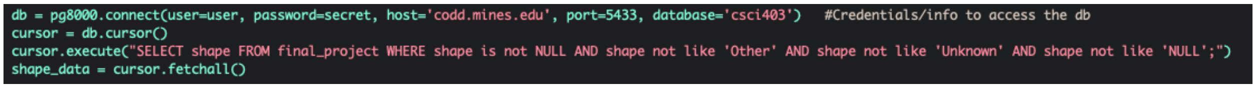
Cool Things:

This dataset provided quite a bit of information to work from. It was discovered that based on the data in the dataset, you are most likely to see a UFO if you are in California. The eyewitness accounts of UFO sightings in California are over two times larger than the sightings in other states. Florida was the second most popular UFO sighting state, and Colorado came in third place. These distributions can be seen in Table 2.

Table 2: Number of UFO sightings in various regions



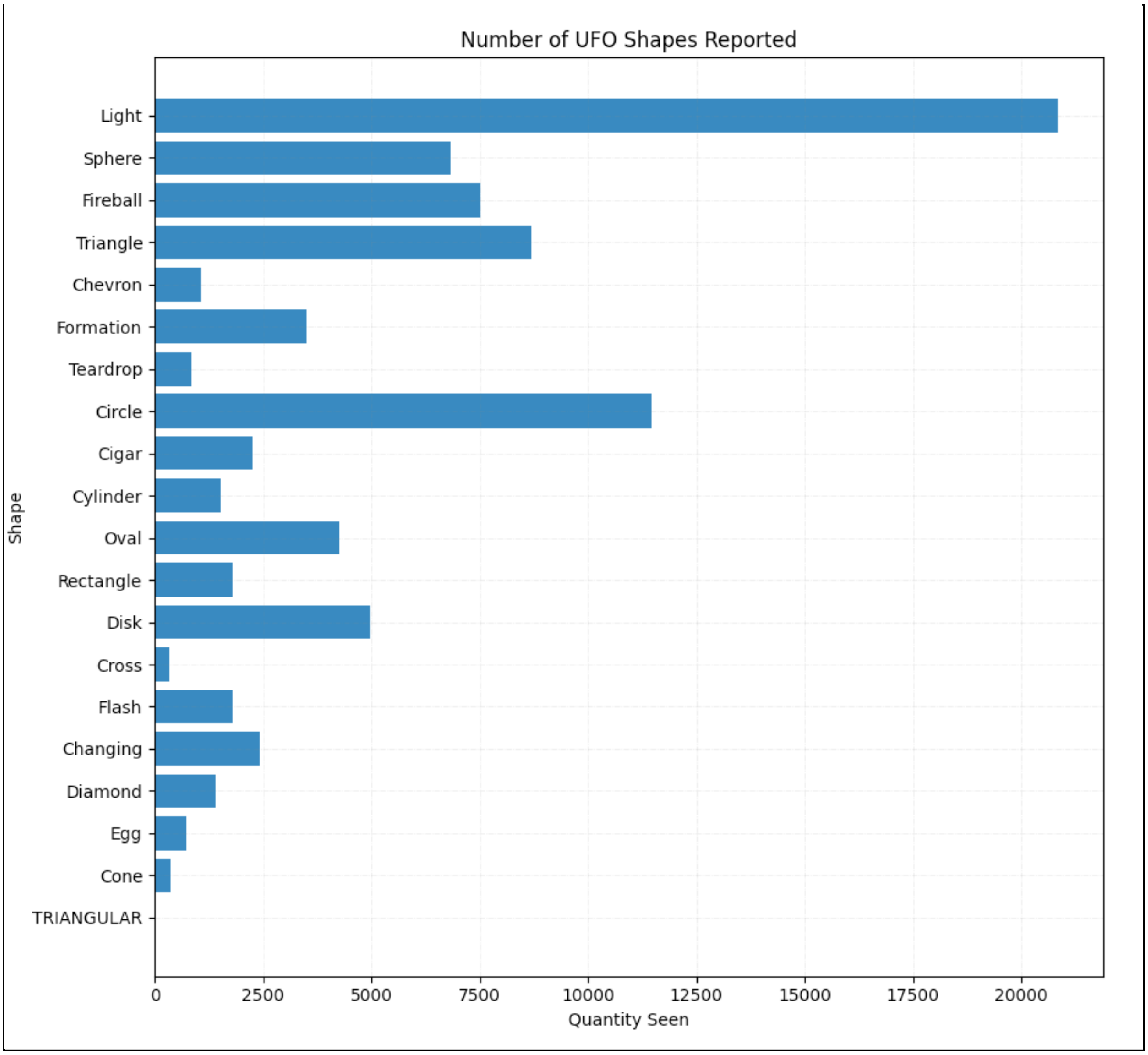
The data was also manipulated to determine the most common shape that was reported. The query that was used to determine this can be seen in Figure 4. According to the reports in the dataset, most people described the shapes that they saw as a light. Circles, triangles, fireballs, and spheres were also common shapes that people used to describe their encounters. This data can be seen in Table 3.



```
db = pg8000.connect(user=user, password=secret, host='codd.mines.edu', port=5433, database='csci403')   #Credentials/info to access the db
cursor = db.cursor()
cursor.execute("SELECT shape FROM final_project WHERE shape is not NULL AND shape not like 'Other' AND shape not like 'Unknown' AND shape not like 'NULL';")
shape_data = cursor.fetchall()
```

Figure 4: Query for the most common shapes seen

Table 3: Most common shapes reported



Queries were used to determine the most popular words that were used to describe the UFO sightings. This analysis was performed on the entire dataset, which can be seen in Table 3, as well as Colorado data, which can be seen in Table 4. Both analyses indicated that “lights” was the most common word used in the summary section of the dataset.

Table 3: Top 50 words in summary data (with 1000 most common English words subtracted) for all observations

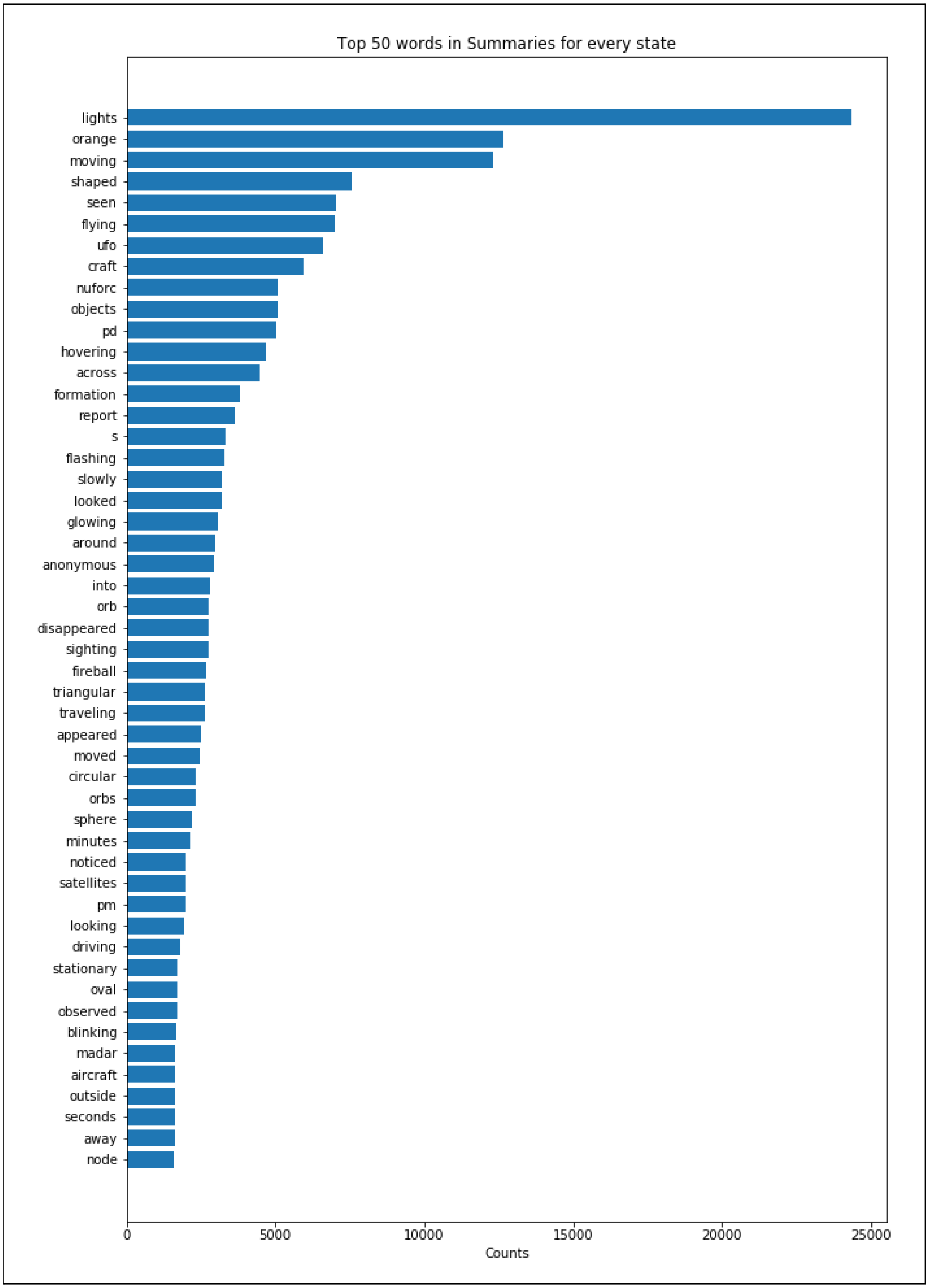
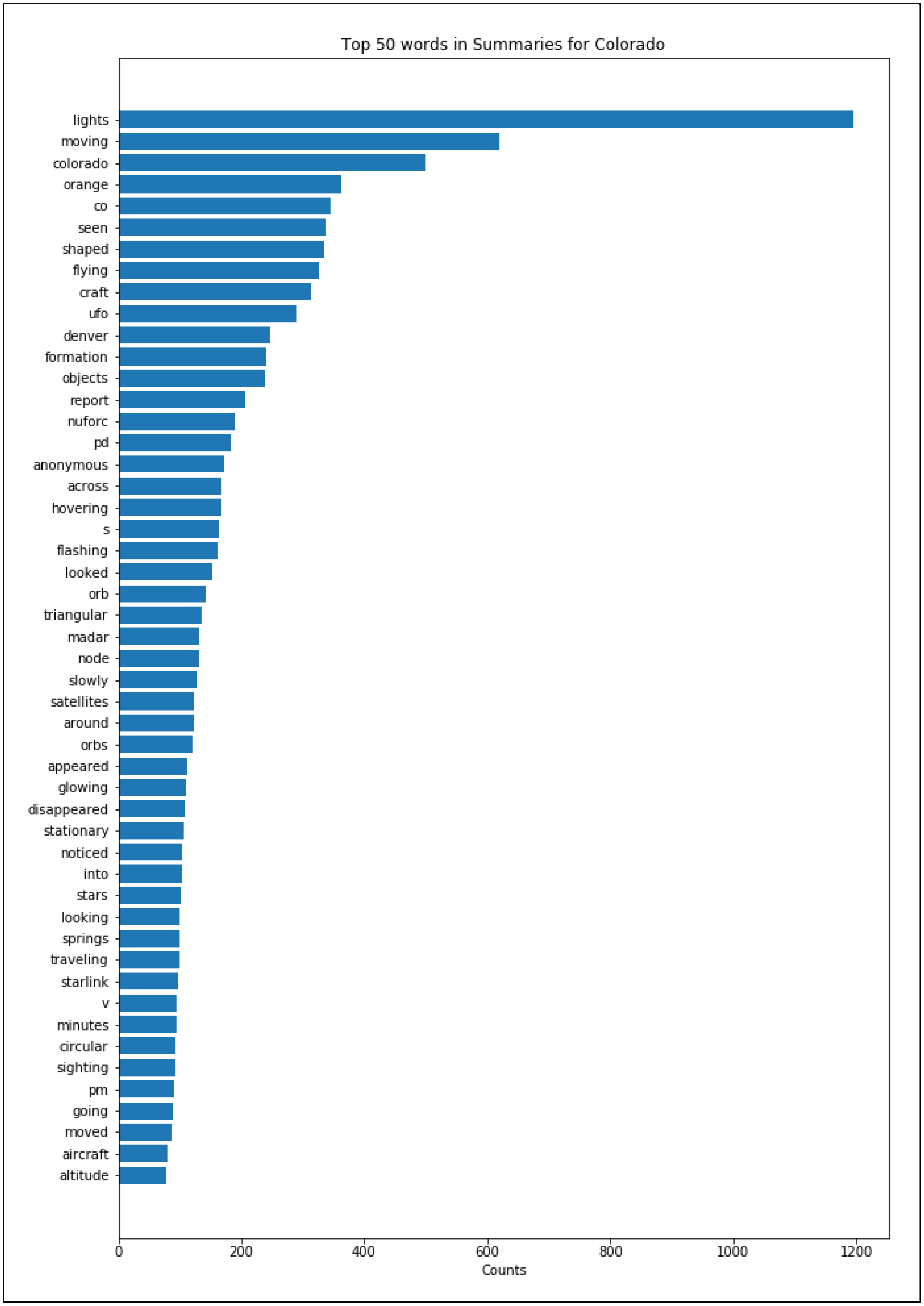


Table 4: Top 50 words in summary data (with 1000 most common English words subtracted) for all observations in Colorado



Challenges:

As was discussed earlier, in order to collect the data from the National UFO Reporting Center Online Database, we needed to scrape the contents of every table on the NUFORC site. The site did not provide a way of displaying a single table containing the contents of the database. However, users can access the data through links that index the data in several ways: EVENT_DATE, STATE, SHAPE_OF_UFO, DATE_POSTED. After selecting a link for a particular means of indexing, there were links to a table for each index. For example, if one chose to index the data by EVENT_DATE, they would be brought to a webpage containing links to pages containing tables for every month and year in the database (e.g. '04/2021'). To see all this data, a user would need to click through every link of the page. We automated this process by grabbing every url in the webpage indexed by EVENT_DATE and scraping the tables at these links. We stored each table in a .csv file, then combined every file into a single .csv file for upload into the database.

Once we retrieved the data, we needed to clean it before loading into the database. Each column contained string data. Fortunately, the columns involving dates/times were already formatted appropriately, so we did not have so any conversion on these. We did notice that there were duplicate rows in the database, and all of these were removed. Furthermore, we replaced every empty string with a 'NULL' string. The only column that needed significant cleaning was the duration column. The duration column represented the duration of each sighting in the database. These were all entered as text, and most took the following forms: 5s, 5sec, 5 seconds, 55 minutes, 5.5 hours, 55.55 etc. We used regular expressions to recognize these types of patterns and convert them into their corresponding time in seconds. Occasionally, the duration was entered as an interval like '5-10' minutes, which we changed accordingly to '7.5 minutes' before converting to seconds. Any entry that did not match these patterns was changed to 'NULL'.

In some instances, we found that the regions of an observation were unclear. For example, 'colorado (above; in flight)' was recorded as a city, which makes the region data slightly inconsistent. This did not impact our ability to upload to the database, and was not changed. However, this issue could be resolved outside, in Python for instance, by enforcing all the 'City' data to correspond to an existing city. There were also some data inputs that were referencing regions outside of the United States, even though this is supposed to be a United States dataset. When used in the geographic analysis, this was fixed by comparing the state names to state abbreviations and casting out data not within the United States.

Conclusion:

The data that was collected was not easy to get into a usable format. There was a significant amount of cleaning and modifying in order to get it in a format that could be worked with. Once this data was converted, it was discovered that most of the UFO sightings reported came from California and the words "light" and "lights" were most commonly used to describe these encounters. This means that if you decide to move to California, be on the lookout for some weird light patterns in the sky!