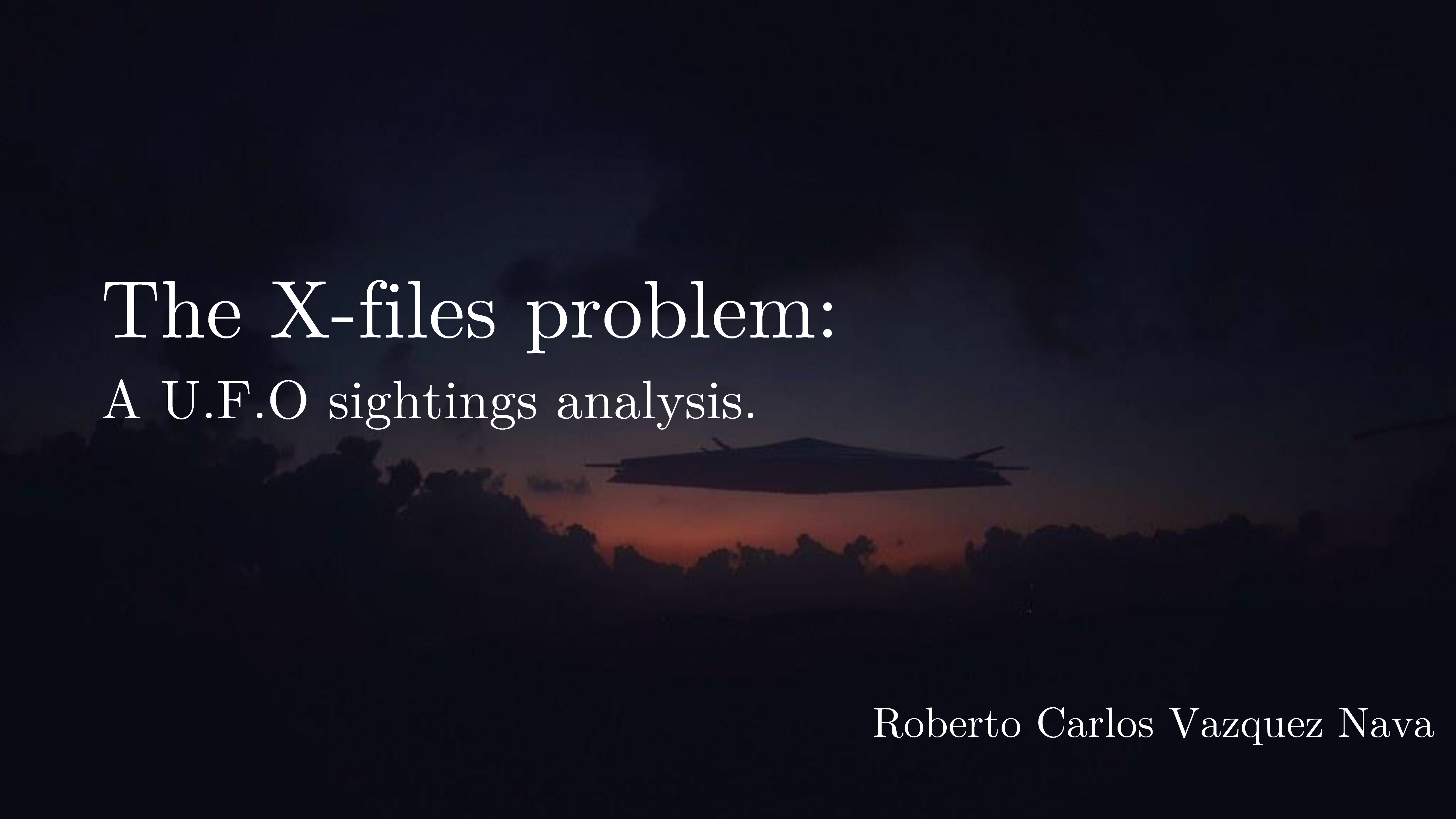


The X-files problem:

A U.F.O sightings analysis.

Roberto Carlos Vazquez Nava

U.F.O sightings

Infinito

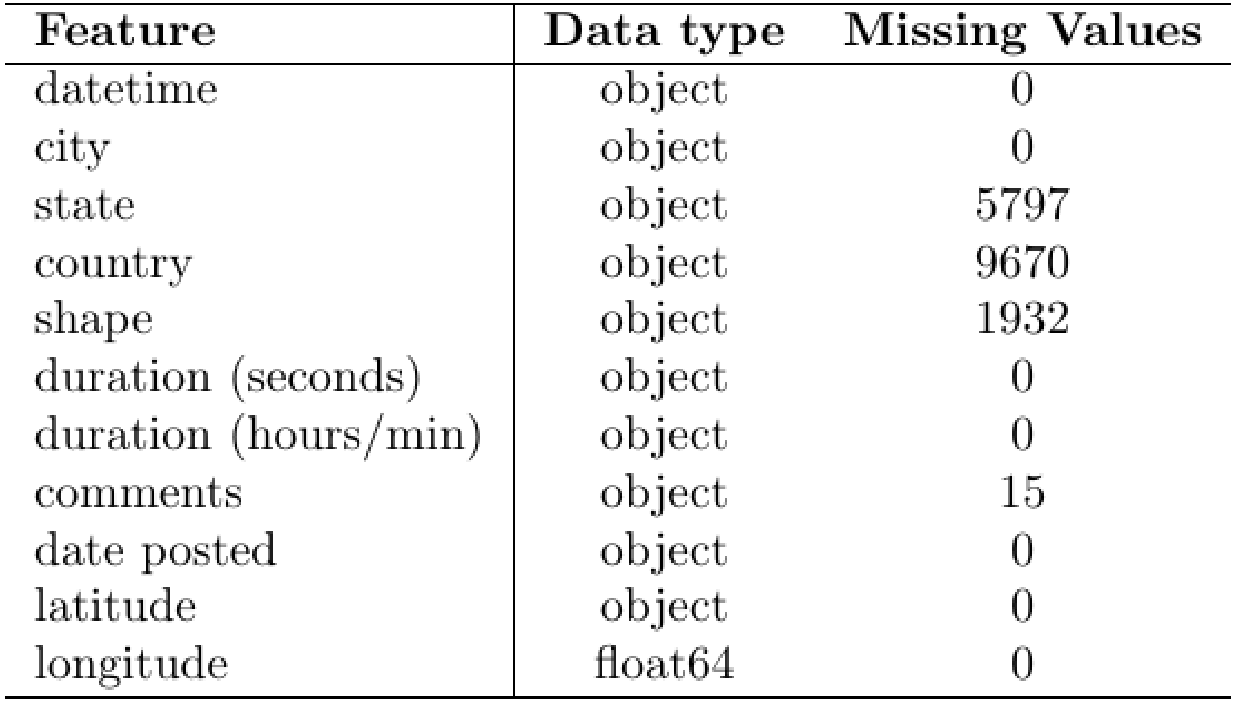
Infinito is a company which has U.F.O. sightings information collected by people from around the world. They go to the company's website and fill out a form about their U.F.O. experience.

CHALLENGE.

To know where it is the best place to see a U.F.O. or interview people that claim sightings.

Data Profiling

Database description



Feature	Data type	Missing Values
datetime	object	0
city	object	0
state	object	5797
country	object	9670
shape	object	1932
duration (seconds)	object	0
duration (hours/min)	object	0
comments	object	15
date posted	object	0
latitude	object	0
longitude	float64	0

The U.F.O database has 80,332 instances and 11 features. The table shows a summary of data types and missing values amount.

Cleaning database I

Wrong Format

The datetime feature data type changed to “*datetime*”. Some instances in this feature had this format "24:00" in time.

Unwanted Characters

The duration (seconds) feature data type changed to “*numeric*”. Three instances have this character “`”.

Wrong Data

The latitude feature data type change to “*numeric*”. There was an instance with this value: 33q.200088 which as imputed by “33.157440” given the city.

Cleaning database II

Cleaning text

The country feature has missing values. Some values are found in city feature. Therefore, a cleaning text process was performed to extract the information.

Remove NaN

The instances with NaN values in country feature (not imputed) were removed. They represented less than 1% of the whole database.

Homogenization

The values from state feature were homogenized (values from USA and Canada mainly).

Cleaning database III

Tokenize and lemmatization

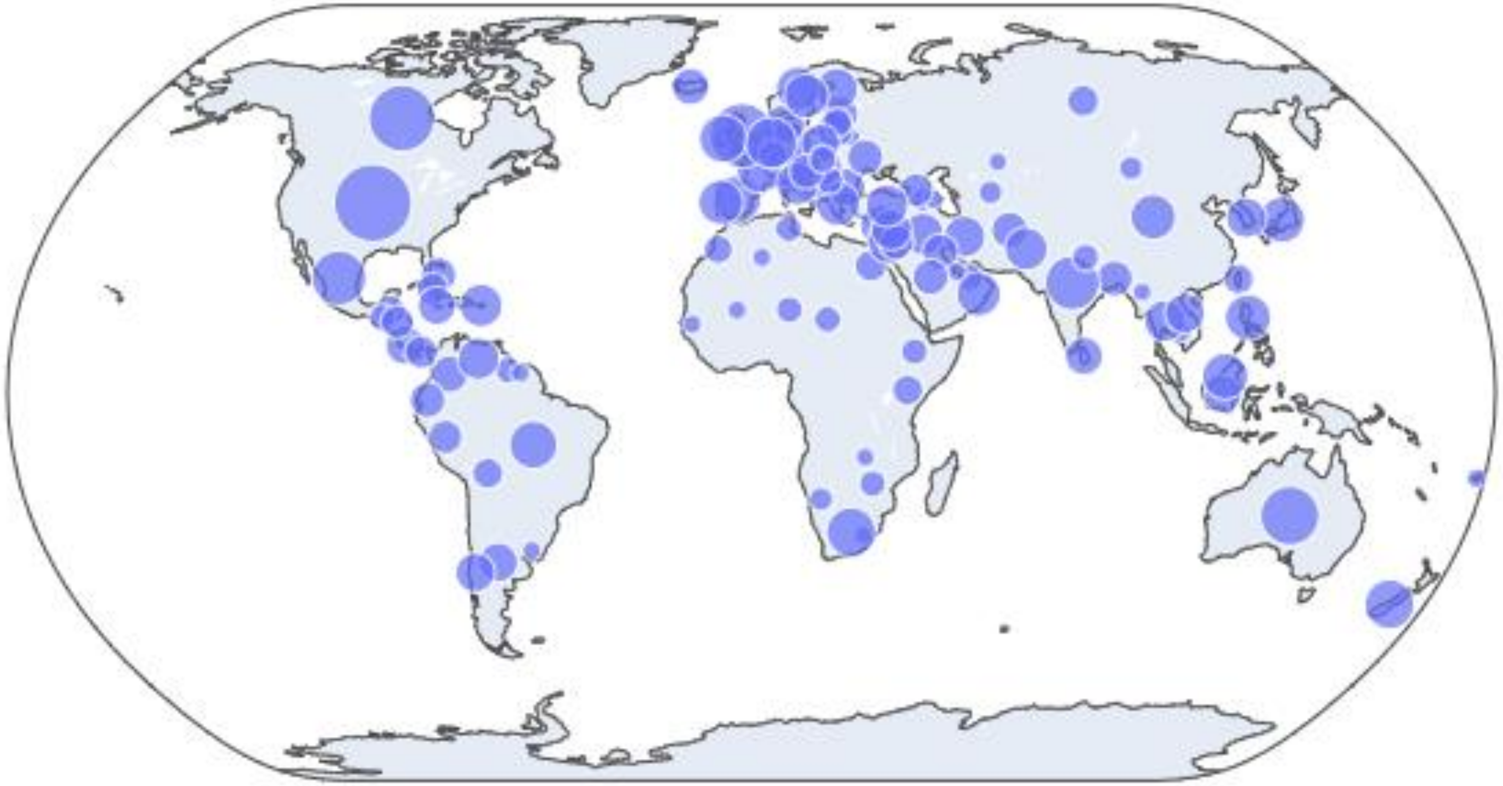
The values from comments feature have symbols that are not useful. Therefore, the cleaning process was performed. The instances with missing values were removed and numbers and non-alphanumeric symbols too. In the end, tokenize and lemmatization techniques were implemented.

Remove NaN

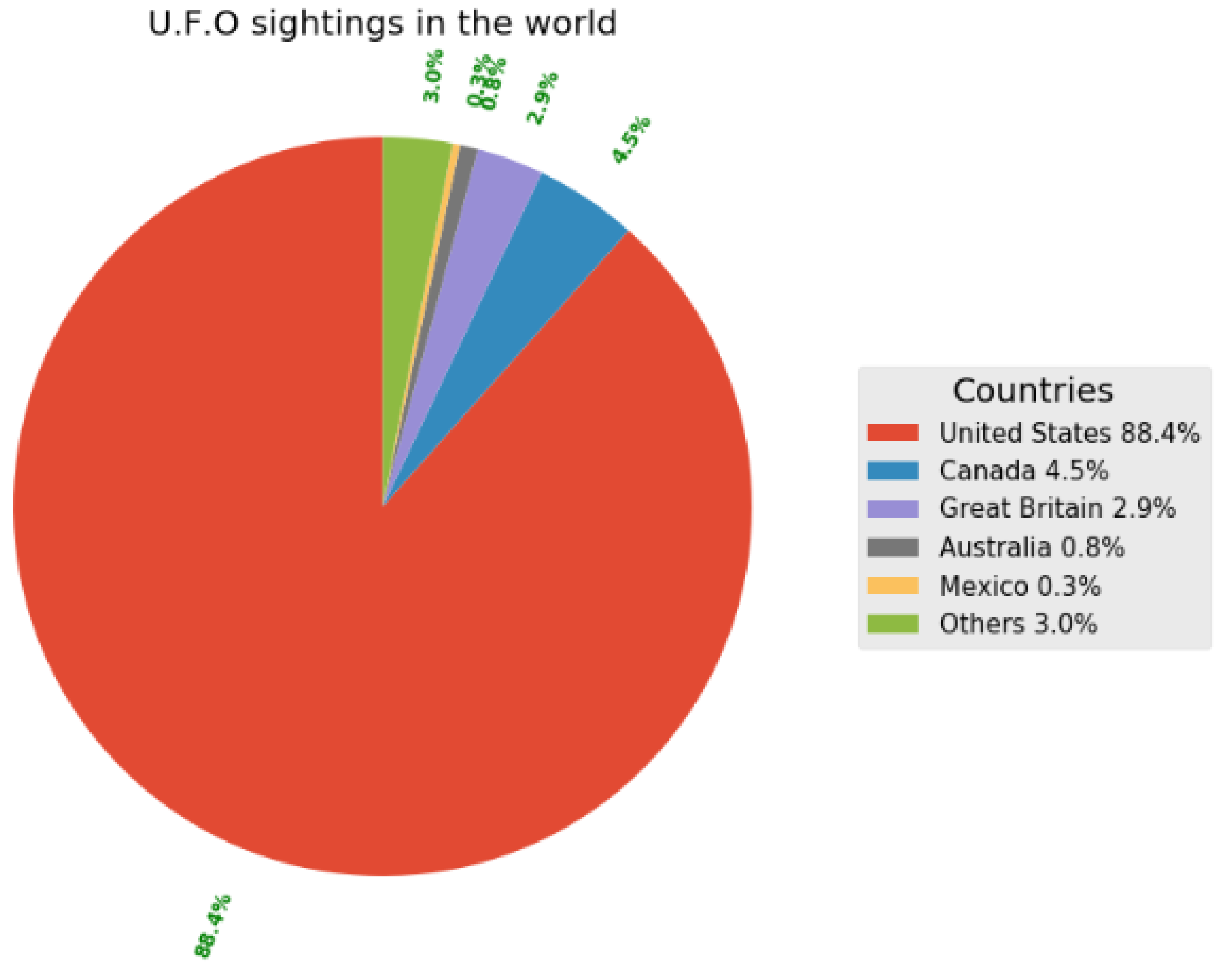
To impute the missing values from shape feature, key elements were searched within comments feature for each instance, so that if there was a match, the imputation was made. The default value for those instances not imputed was "unknown".

Analysis

U.F.O sightings in the world until 2014

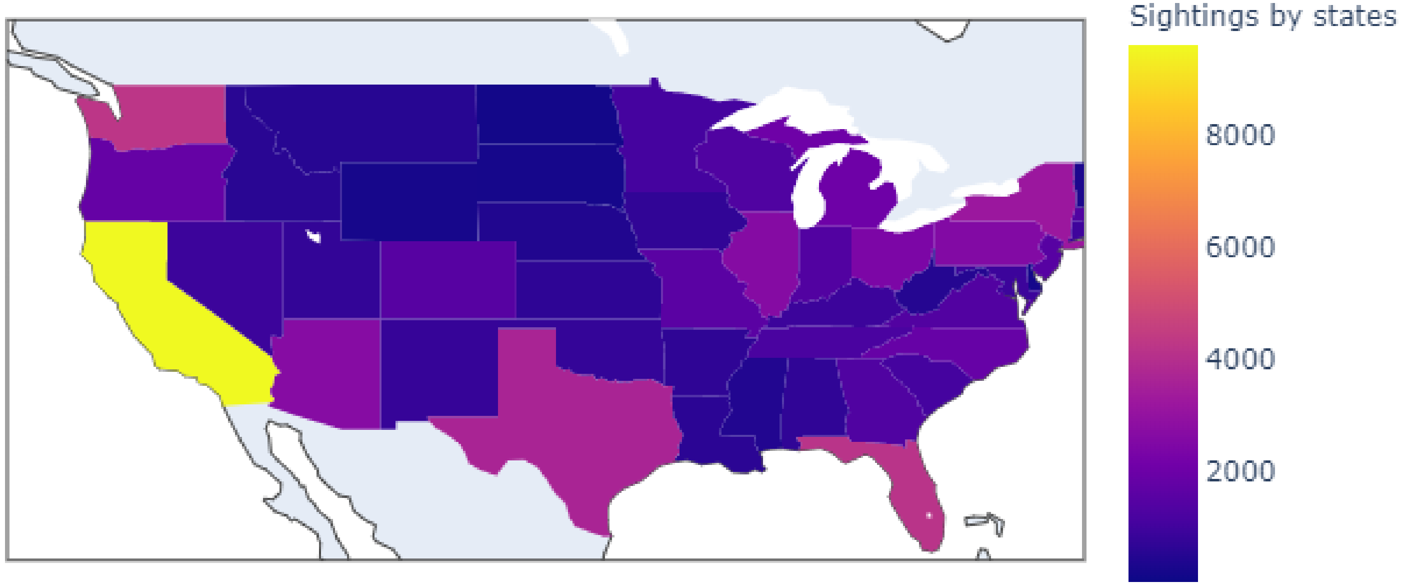


U.F.O sightings in the world until 2014



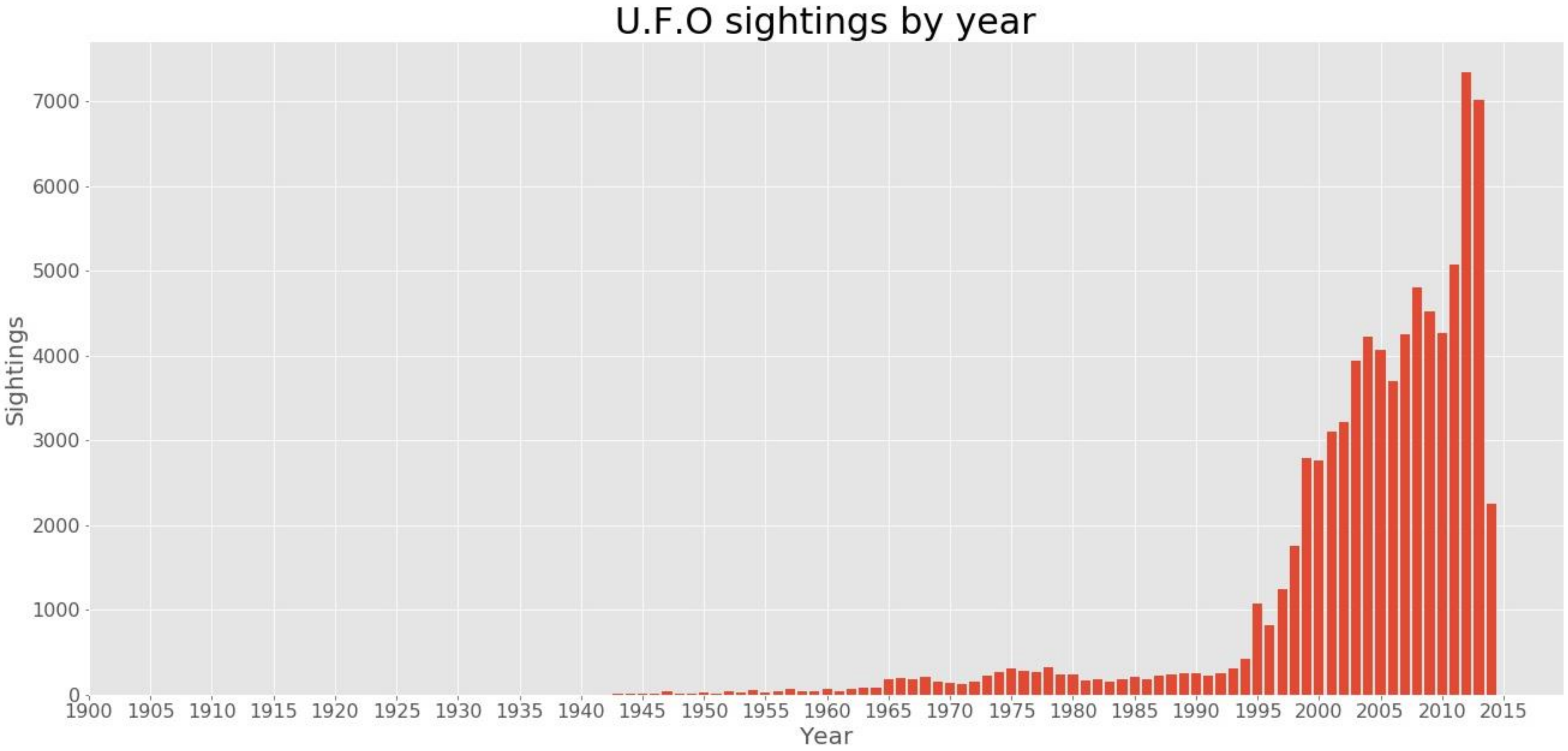
The USA is the country where there have been reported more U.F.O sightings.

U.F.O sightings in the world until 2014



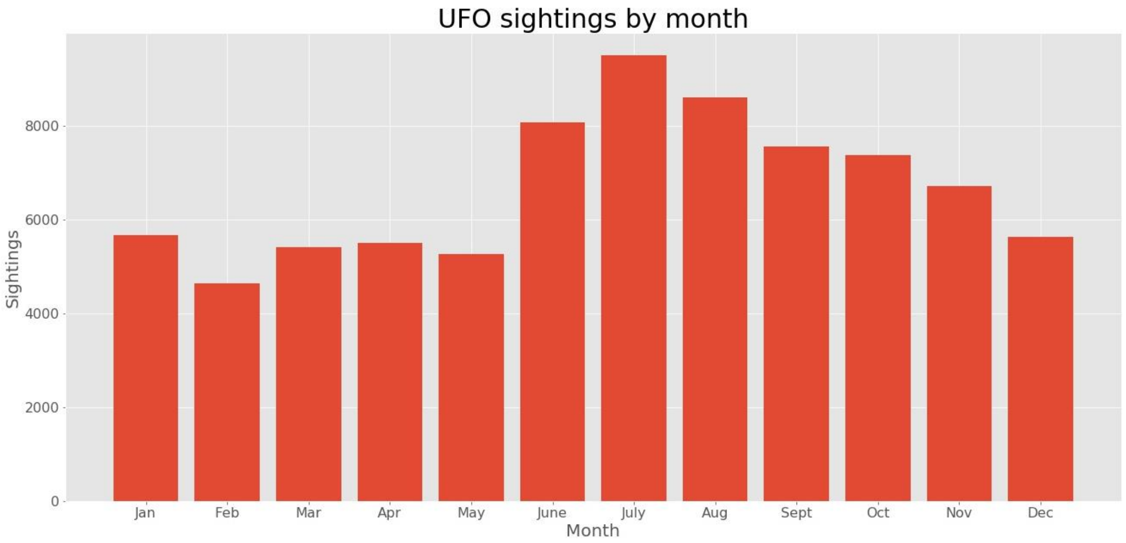
California is the USA's state where there have been reported more U.F.O sightings.

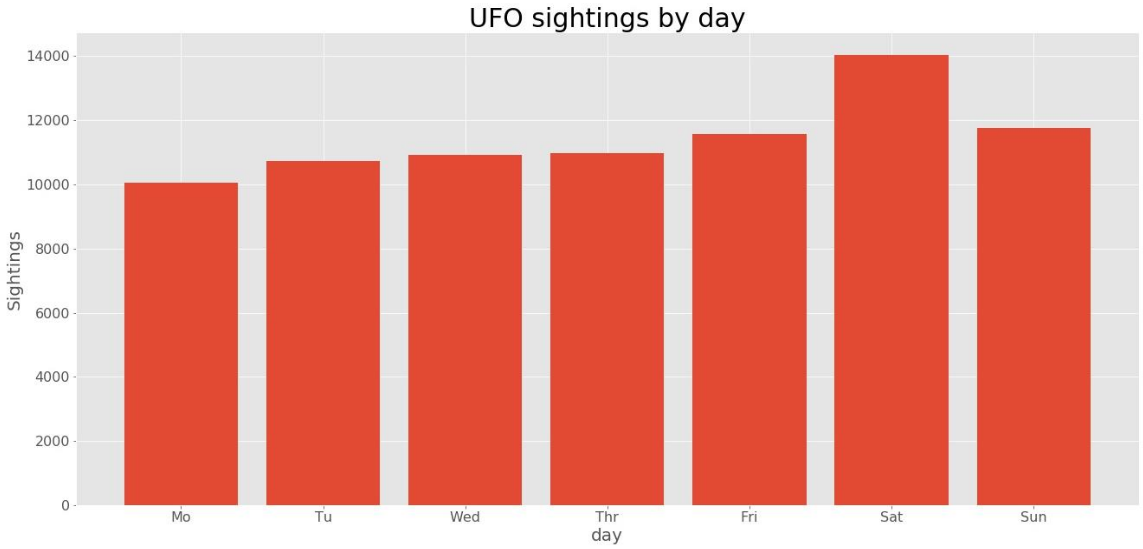
U.F.O sightings by year



The U.F.O sightings start to increase rapidly after 1995.

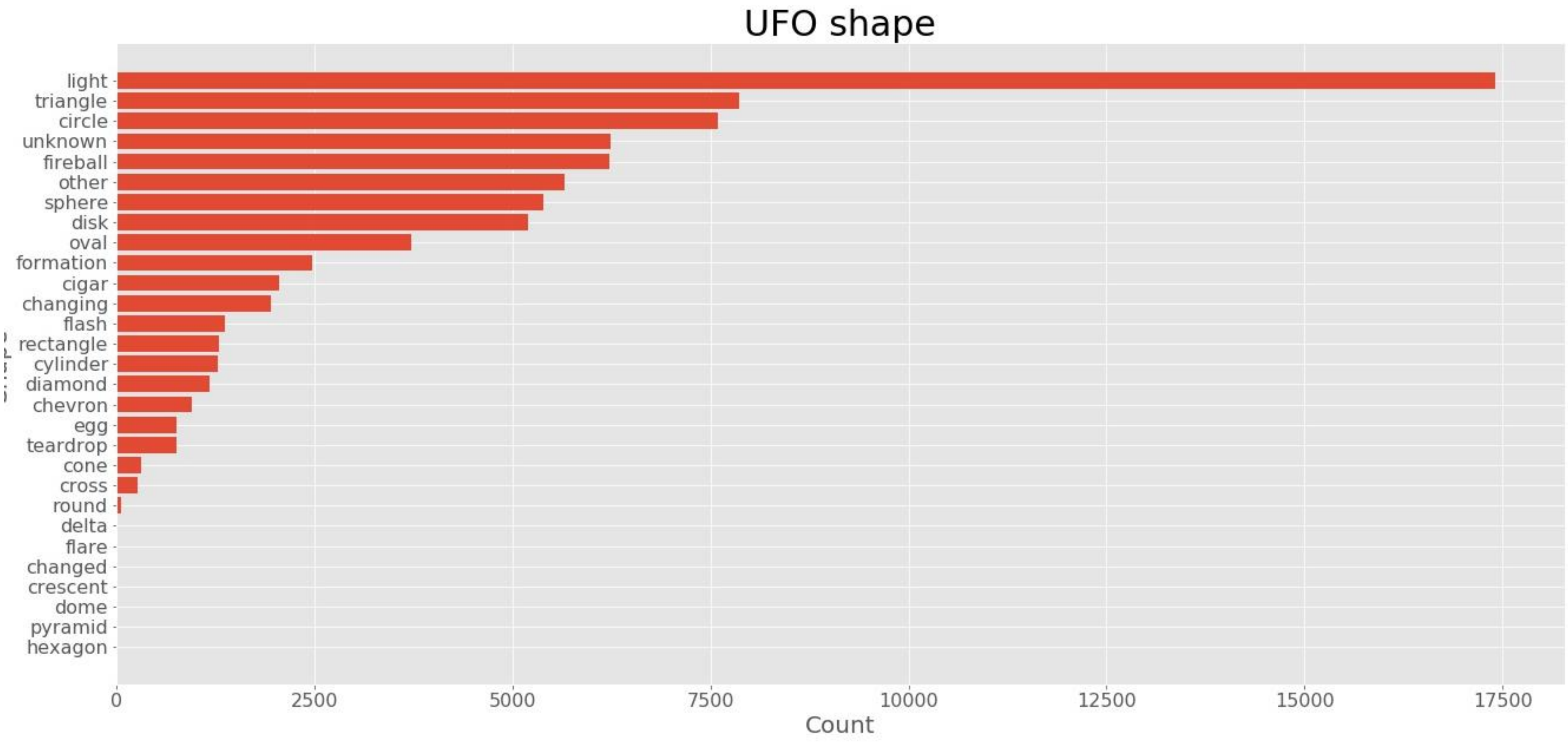
U.F.O sightings by month and day





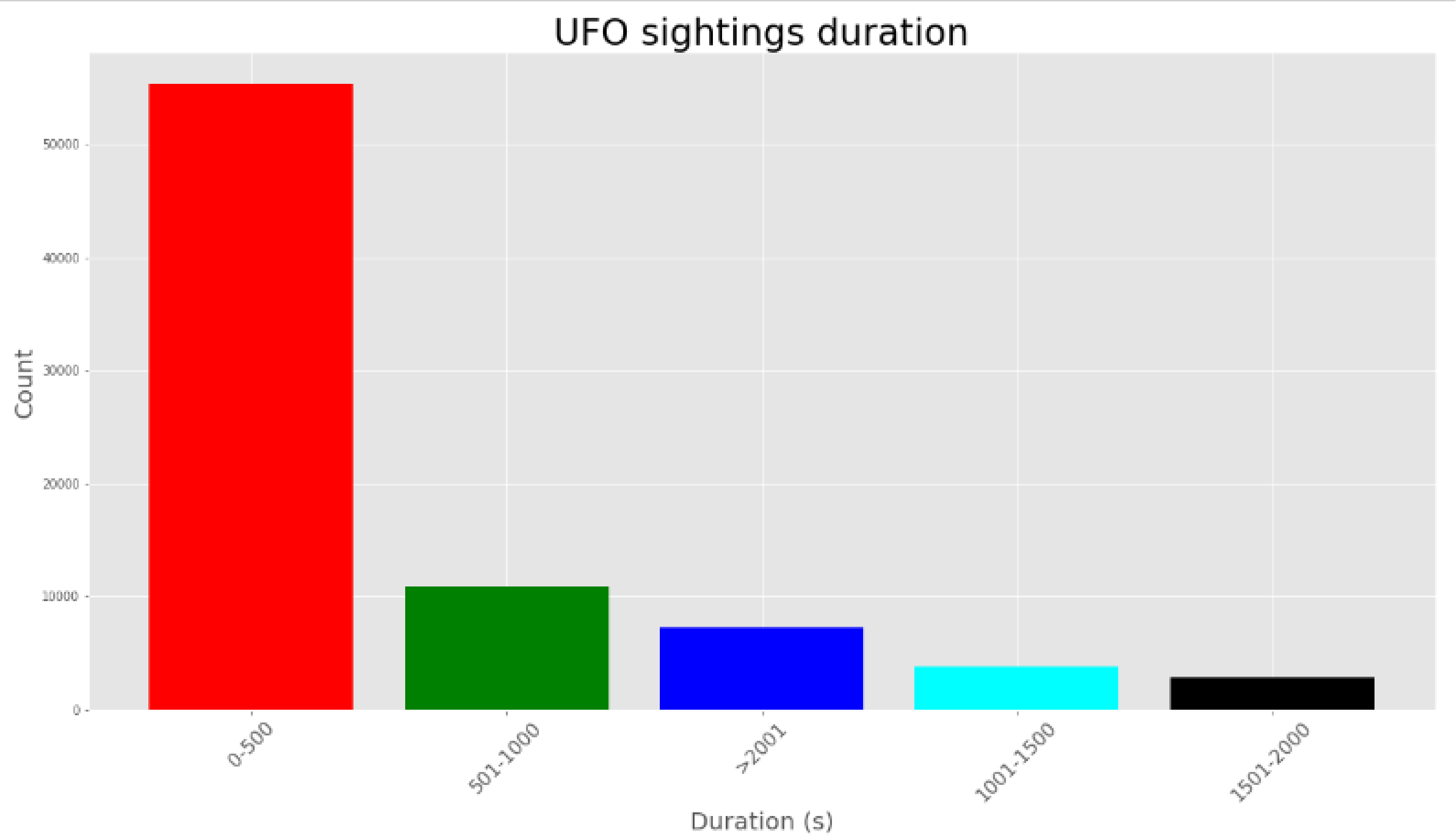
July and Saturday are the month and the day respectively with the most U.F.O sightings.

U.F.O shape



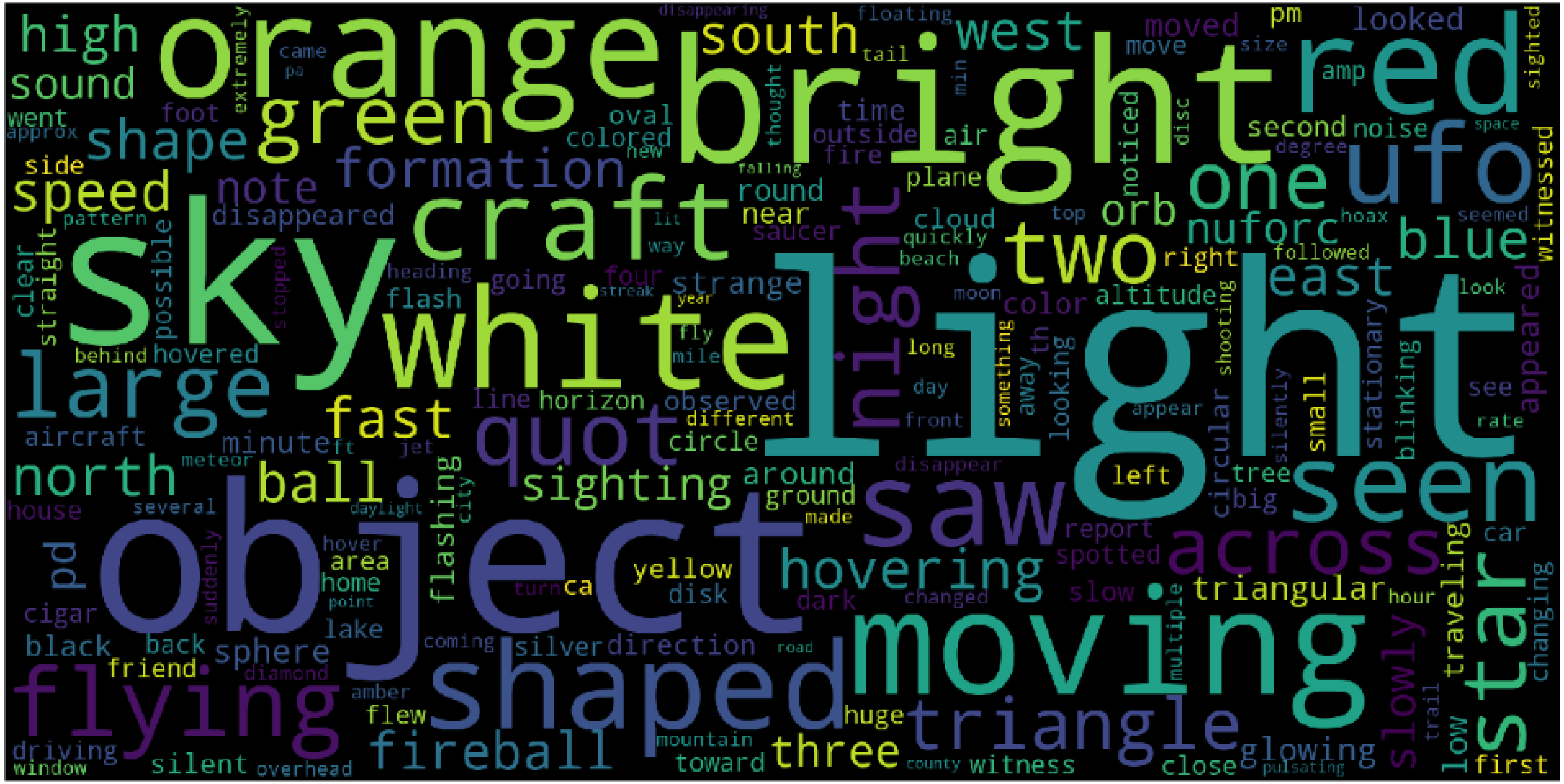
Most people describe a UFO as a light.

U.F.O duration



The duration of a U.F.O sighting is less than 500 seconds.

Words more often used when a sighting is described.



1. light
2. object
3. sky
4. bright
5. moving
6. orange
7. white
8. red
9. shaped
10. saw
11. craft
12. ufo
13. seen
14. like
15. flying

Conclusions

Conclusion

- The most likely place to see a U.F.O is California, USA.
- July is the best month for a U.F.O sighting, on Saturday night.
- The most likely shape of a U.F.O is a light or a triangle, which color could be orange, white or red.
- If it is no possible to go to the USA, then you should go to Ontario, Canada.

Future work

- Carry out a deeper analysis of other countries. It is possible you do not live in America.