STATS 140XP Final Project EDA

Prateik Sinha

2023-03-02

Load in Data:

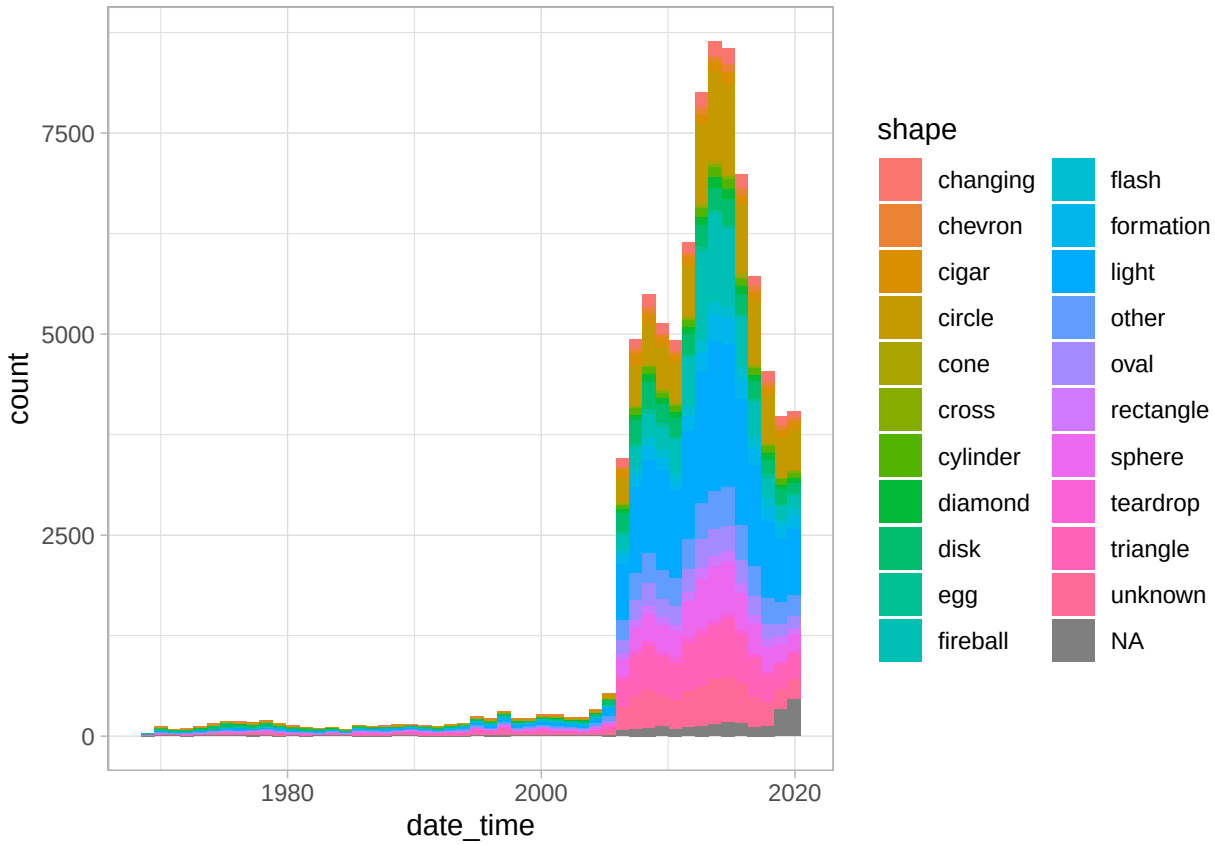
Cleaning Data

Examining the relation between the summary and text columns:

```
## Summary:
##  My wife was driving southeast on a fairly populated main side road, it was dark out side at about 6
##
##  Text:
##  My wife was driving southeast on a fairly populated main side road, it was dark out side at about 6
```

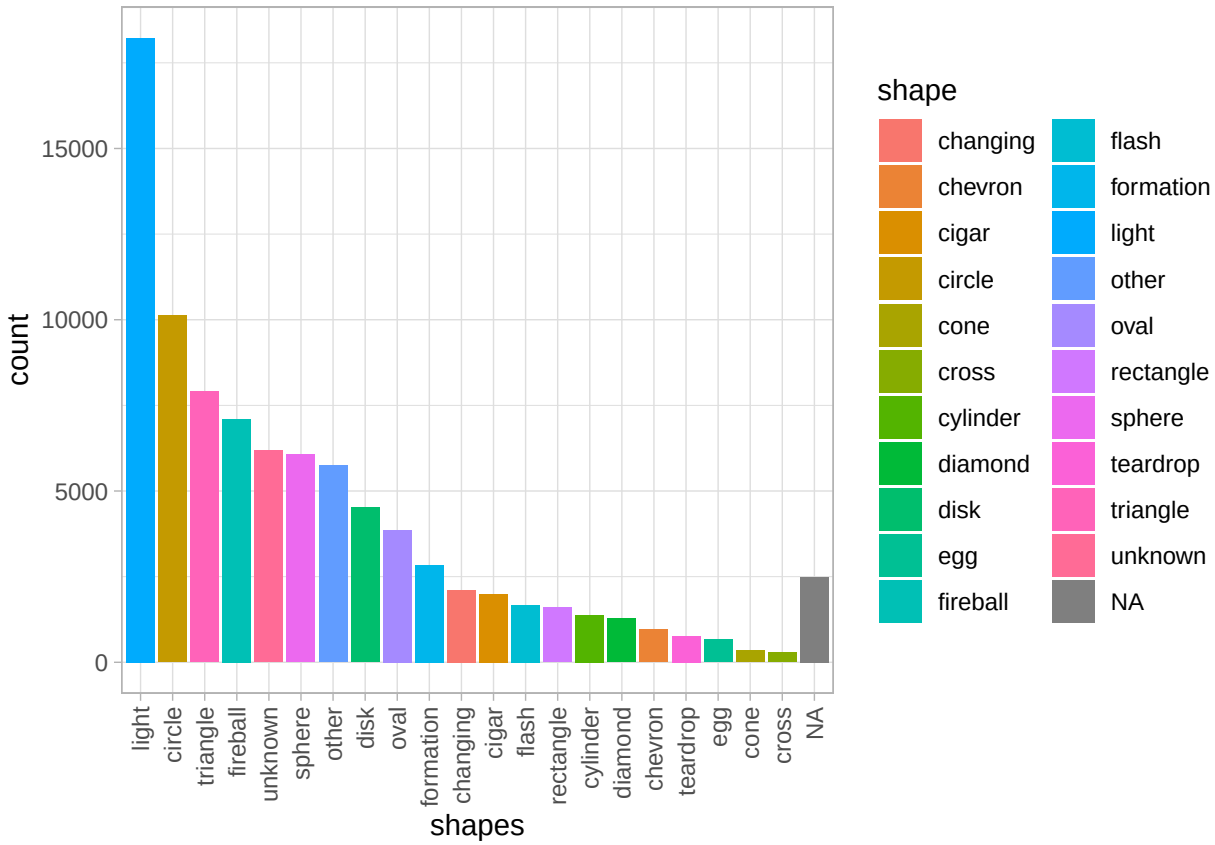
On inspecting the data, we can see that the summary is always a substring of the text column. Since it provides no additional information we can remove the summary column.

Let us look at the distribution of these reports with respect to time:



We can see that there was a peak in the mid-2010s, with the distribution appearing fairly normal from ~2006 onwards. There were very few reports before that, but these reports were submitted online and that may be due to the fact that the internet was not as popular was it in recent years.

Although the relative proportions of the shapes of UFOs being reported seems to stay fairly constant throughout the years, we can look further into which shapes are more frequent occurrences and how their popularity changes over time:



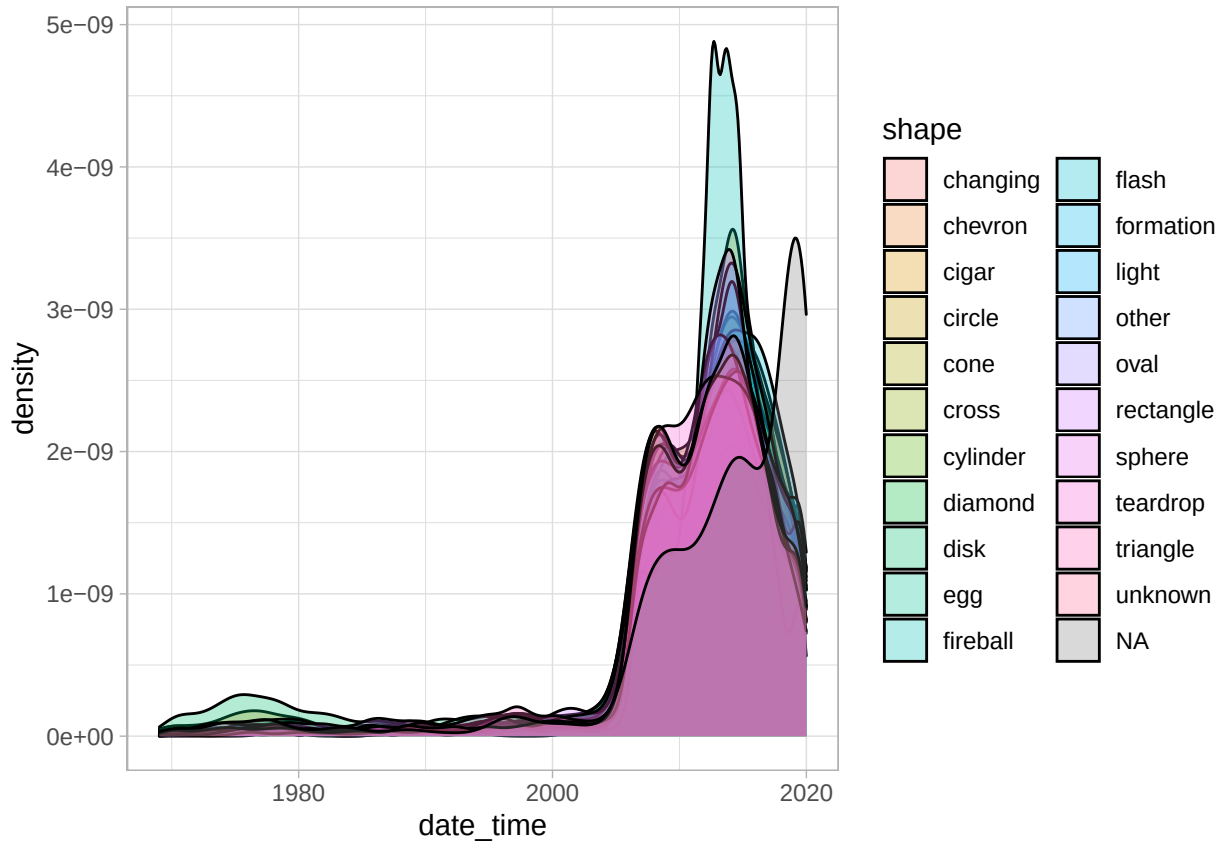
Looking at their frequency, we can see that light is by far the most popular, which may be due to the fact that this it is a more vague term that could be used to describe many possible shapes rather than something like chevron or crone, which is very specific.

The distant second place goes to circle, which may possess the same advantage due to its abstractness. Many may describe oval, disk and egg shapes as circles rather than place them in their own specific categories.

The rest of the labels smoothly taper off in terms of popularity, ending with cone and cross. Roughly 2.8346099 of the reports did not mention a shape, meaning the vast majority of UFO encounters do have people making clear visual contact of the vessel.

An interesting thesis to consider would be to find the relation between the shapes and occurrences of UFOs reported by people and coverage of them in news, media and other popular culture to see how related they are, and if it data indicates that one may be based off another.

```
## Warning: Removed 1187 rows containing non-finite values (stat_density).
```



From the above graph we can see:

- The distribution appears to be tri-modal
- the second mode is the largest, and contains a narrow and steep peak of reports of a light-shaped UFO. This warrants further investigation into real life events at the time to see if this can be explained by events happening at the time.
- The third peak of NA shapes seems to be the most recent. This could be a problem with data collection and also warrants investigation