Exploratory data analysis of UFO data set

Anita Li, Jacob McFarlane, Steffen Pentelow, Chirag Rank

20/11/2020

Summary of the data set

The data set used in this project are records of UFO sightings in British Columbia, Canada and Washington States, USA, which is provided by America's foremost UFO Reporting Agency since 1974 and can be found here. Each row in the data set represents an observation of UFO sighting, and features recorded include place and time, shape of UFO, duration of sightings, and a short descriptive summary. There are 4710 observations and 7 features in the data set. However, there are many records with invalid shape and duration. After removing invalid records, there are 2682 observations left. Below is a summary of duration for each UFO shape that has more than 30 observations.

Table 1: Table 1. Statistic summary on the duration of sightings for each shape

Shape	median	count	min	max
Chevron	15.0	34	2	1800
Flash	27.5	68	1	90000
Fireball	120.0	250	1	14400
Rectangle	120.0	43	1	7200
Cigar	150.0	56	1	7200
Formation	180.0	113	3	10800
Light	180.0	798	1	86400
Oval	180.0	155	1	7200
Sphere	180.0	233	1	25200
Triangle	180.0	252	2	10800
Circle	300.0	370	1	54000
Cylinder	300.0	48	3	10800
Disk	300.0	138	1	331200
Diamond	420.0	51	5	14400

Exploratory analysis on the data set

We choose median to represent the average duration for each sighting because the distribution of duration is skewed. To look at weather the median of duration are different for different shape, we plotted the distribution of duration (on log scale) for each shape.

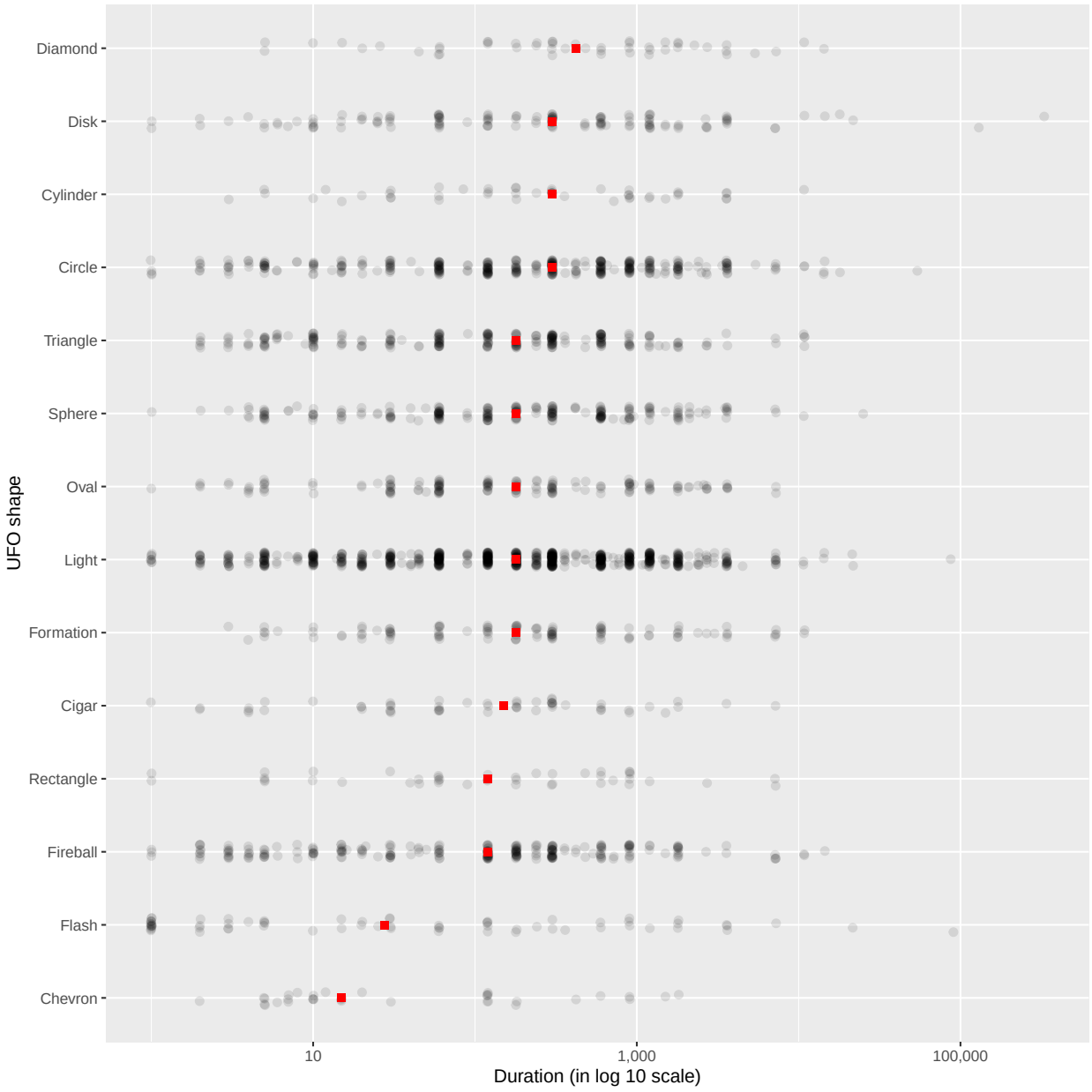

Figure 1: Distribution of time duration (in log scale) for each UFO shape