Exploratory data analysis of UFO data set

Anita Li, Jacob McFarlane, Steffen Pentelow, Chirag Rank

21/11/2020

Summary of the data set

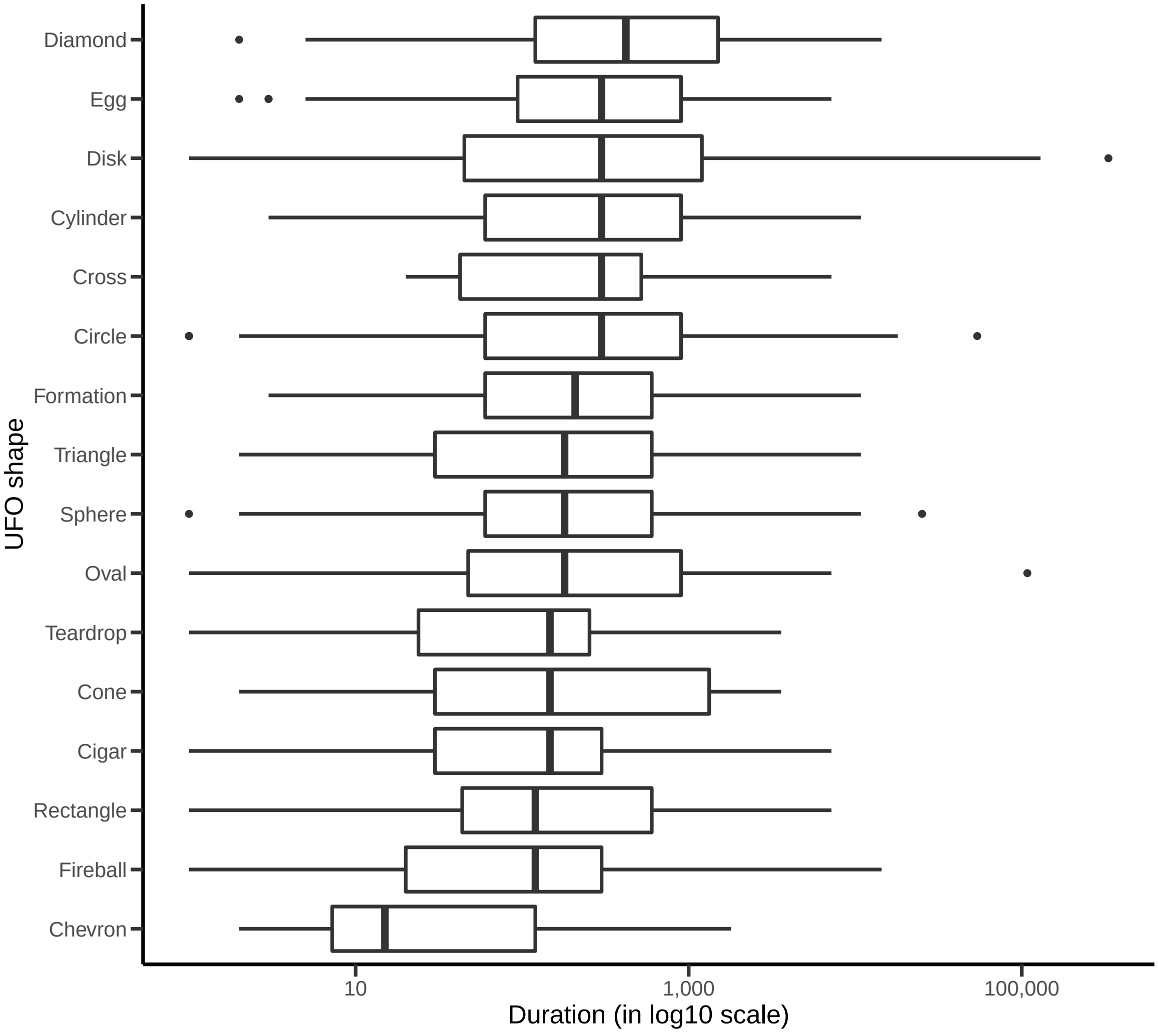
The data set used in this project are records of UFO sightings in British Columbia, Canada and Washington State, USA, which is provided by America's foremost UFO Reporting Agency since 1974. Each row in the data set represents an observation of UFO sighting, and features recorded include place and time, 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 or durations. After removing invalid records, there are 1846 observations left. This project will only consider UFO shapes that have more than 30 observations. Table 1 summarizes the duration for each UFO shape.

Table 1: Table 1. Summary on the duration (seconds) of sightings for each shape

Shape	Numer of observations	Median	Minimum	Maximum
Chevron	34	15	2	1800
Fireball	253	120	1	14400
Rectangle	44	120	1	7200
Cigar	56	150	1	7200
Cone	16	150	2	3600
Teardrop	20	150	1	3600
Oval	159	180	1	108000
Sphere	239	180	1	25200
Triangle	257	180	2	10800
Formation	114	210	3	10800
Circle	374	300	1	54000
Cross	11	300	20	7200
Cylinder	48	300	3	10800
Disk	141	300	1	331200
Egg	28	300	2	7200
Diamond	52	420	2	14400

Exploratory analysis on the data set

The median duration of sighting has been selected as the preferred measure of central tendency for this project because the distributions of durations are skewed. Figure 1 illustrates the distribution of durations for each shape through boxplot of each different. A log_{10} scale was used for duration axis so that the distribution of observations could be seen more clearly.



From Table 1 and Figure 1, it is noted that several shapes share similar median durations. For example, both ‘Fireball’ and ‘Rectangle’ have median durations of 120 seconds or 2 minutes. Multiple shapes also shared median durations of 3 minutes and 5 minutes. Based on a review of the raw data, it appears that many observers reported durations to the nearest minute which explains these duplications and the ‘binning’ of points around particular values observed in Figure 1.