Prepared by: Jay Haygood
April, 2022

1. Data Source:

UFO Sightings around the world | Kaggle
80,000+ documented close encounters from the past 70 years.

Data Collection
The data set is external, and should be trustworthy as it was compiled into Kaggle from NUFORC, where UFO reports have been geolocated and time standardized, for close to a century of data.

Acknowledgements
Data for above source originally curated from: https://github.com/planetsig/ufo-reports
License: CC0: Public Domain
Expected update frequency: Not specified

Context
Via the above source on Kaggle:
"Extraterrestrials, visitors, little green men, UFOs, swap gas. What do they want? Where do they come from? Do they like cheeseburgers? This dataset will likely not help you answer these questions. It does contain over 80,000 records of UFO sightings dating back as far as 1949. With the latitude and longitude data it is possible to assess the global distribution of UFO sightings (patterns could aid in planetary defense if invasion proves to be imminent). The dates and times, along with the duration of the UFO's stay and description of the craft also lend themselves to predictions. Can we find patterns in their arrival times and durations? Do aliens work on weekends? Help defend the planet and learn about your fellow earthlings (and when they are most likely to see ET)."

Overview of Data Content
- **Date_time** - standardized date and time of sighting
- **city** - location of UFO sighting
- **state/province** - the US state or Canadian province, appears blank for other locations
- **country** - Country of UFO sighting
- **UFO_shape** - a one word description of the "spacecraft"
- **length_of_encounter_seconds** - standardized to seconds, length of the observation of the UFO
- **described_duration _of_encounter** - raw description of the length of the encounter (shows uncertainty to previous column)
- **description** - text description of the UFO encounter. **Warning:* column is messy, with some curation it could lend itself to some natural language processing and sentiment analysis.
- **date_documented** - when was the UFO sighting reported
- **latitude** - latitude
- **longitude** - longitude

Explanation for data choice

I've always had a personal fascination with the history of UFO phenomena, as well as all of the theories, speculation, and possible explanations therein. While they can range from plausible to extravagant, often reading like modern science fiction, these events have been well documented and heavily investigated, over the last century in particular. Officially, the sightings are now referred to as UAP (Unidentified Arial Phenomena), due in no small part to the stigma attached to the term "UFO", after decades of conspiracy theories, big-budget movies, and seemingly countless documentary films and books on the subject. This has also gained recent traction in the press, following release of the Prelimary-Assessment-UAP-20210625.pdf (dni.gov) report, submitted to Congress June25, 2021, from the Office of the Director of National Intelligence.

2. Data Profile:

The original data set contains 80,332 rows and 11 columns.

Variable	Time Component	Data Structure	Qualitative		Quantitative	
			Nominal	Ordinal	Discrete	Continuous
Date_time	Time-invariant	Structured		X		
city	Time-invariant	Structured	X			
state/province	Time-invariant	Structured	X			
country	Time-invariant	Structured	X			
UFO_shape	Time-invariant	Structured	X			
length_of_encounter_seconds	Time-invariant	Structured				X
described_duration_of_encounter	Time-invariant	Structured				X
description	Time-invariant	*Unstructured	X			
date_documented	Time-invariant	Structured		X		
latitude	Time-invariant	Structured				X
longitude	Time-invariant	Structured				X

Consistency Checks

There are no duplicates, however there are multiple missing values as follows:

state/province	5797
country	9670
UFO_shape	1932
Description	15

The majority of blank 'state/province' entries are proper, as not being either US or Canadian territories. Missing values for 'UFO_shape' and 'Description' columns will be left blank, as this info was likely not provided for the respective sighting.

Cleaning Process

- Data type conversion error identified ‘latitude’ typo: '33q.200088', replaced with '33.200088'
- Data type conversion error identified ‘length_of_encounter_seconds’ syntax errors: 0.5`, 2`, 8`, replaced with 0.5, 2, 8, respectively.
- Converted data type for ‘latitude’ and ‘length_of_encounter_seconds’ columns to Float64.
- Replacement of blank ‘country’ data for entries with US states and Canadian territories to be performed in Excel after exporting data set, which requires multiple ‘if’ statements using Python.
- 101 entries have an undefined integer format for the ‘Date_time’. Being less than .13% of the data set, these entries will be deleted.
- Replaced blank ‘country’ entries with AU, CA, DE, GB, US, where country could be identified accordingly, using ‘city’ or ‘state_province’ data. This reduced missing ‘country’ entries from 9670 to 2680, now representing 3.33% of the data set. While a small enough percentage to remove, these entries will remain in the prepared data set for the time being.

Descriptive statistics

	length_of_encounter_seconds	latitude	longitude
count	8.033200e+04	80332.000000	80332.000000
mean	9.016889e+03	38.124416	-86.772885
std	6.202168e+05	10.469585	39.697205
min	1.000000e-03	-82.862752	-176.658056
25%	3.000000e+01	34.134722	-112.073333
50%	1.800000e+02	39.411111	-87.903611
75%	6.000000e+02	42.788333	-78.755000
max	9.783600e+07	72.700000	178.441900

Limitation and Ethics

* While the unstructured format of the ‘description’ column may present challenges, it could reveal noteworthy trends if filtered properly. The remaining missing values represent a small percentage of each variable, therefore should have minimal impact on the analysis. There are no apparent biases or ethical concerns, and no personal information within the data set.

3. Questions to explore:

- A global plot, and possible heat map showing locations of recorded UFO sightings.
- Can the duration of the UFO event be predicted from the other available data?
- Can correlations be identified, either within the existing data set, or by incorporating another external data source? (Proximity to a military base, or coinciding with a celestial event?)
- Is there an identifiable pattern to the sightings? At certain times of day, days of the week, or times of the year? (If people are on their way home from a night on the town or festive event, are they more likely to see little green men?)
- Are certain shapes, or other descriptive factors of UFO sightings, more frequently reported in various geographical regions?