A Needle in a Data Haystack - Final Project - Group 44
Exploring UFO sightings data (or - ARE ALIENS REAL?)

Written by:

- Matan Cohen -- matan.cohen2@mail.huji.ac.il -- cs: matanc555
- Nir Schipper -- nir.schipper@mail.huji.ac.il -- cs: nir_s
- Ran Shaham -- ran.shaham1@mail.huji.ac.il -- cs: ransha

Online resources:

- Github repository: https://github.com/shahamran/needle
- Static rendering: https://github.com/shahamran/needle/blob/master/Aliens.ipynb

***Important**: Most of the work in this project is the notebook, found in the links above - including figures and insights regarding them.*

Description:

We set out to examine changes in UFO sightings over time and different locations, while putting emphasis on the volume of sightings and on the description of the objects.
We hypothesized that the sightings would be influenced by popular sci-fi movies and by the development of real life air and space crafts. We also assumed that the number of sightings would rise significantly in years where major cultural events (movies, tv shows, books) regarding aliens or science fiction.

Data:

UFO sighting data - We got it from Kaggle, and its source is The National UFO Reporting Center (NUFORC). We used the *scrubbed.csv* file which is a cleaner version of the dataset (without blank rows). It consists of 80,332 data rows (each row is a different UFO sighting) and the following columns:

- datetime: The date and time of the sighting
- city, state, country, longitude, latitude: The exact location in which the UFO was seen (5 columns)
- shape: The shape of the object
- duration (seconds), duration (hours/min): The duration of the sighting
- comments: A textual description of the object. This column is incomplete - probably lost in scraping.
- date posted: The date in which the sighting was *reported*.

The dataset's size is 13.9MB.
US population density and population estimates - downloaded from census.gov, in order to normalize the sightings count in each state by its population size. It comes as 2 small tables that specify the population density / estimate for each state. Its size is marginal (< 1KB each).

Solution:

We chose to write our project as a Jupyter Notebook in order to write nicely presented code and use interactive widgets to explore the data. The code is supplied via moodle, and is also hosted in a github repository. The link contains instruction on how to run the code (we supplied scripts to automate the packages installation). A static version of the notebook is available here.

All data analysis was done with python. We performed extensive statistical analysis trying to corroborate our initial hypothesis while also exploring potentially interesting findings which could serve as grounds for future research.

We extracted the following statistics:

- Number of sightings over the years
- Variations in reports of spacecraft shape through time and in different places.
- Variations in duration of sightings over the years (averaged each year) and in different places.
- Distribution of reports by time of day and time of month.
- Distribution by state, normalized by population size.

Future Work:

In this project, most of the work was to understand how the data *looks like*, and hypothesize the *reasons* for such a behavior. Therefore, future work should *test* these hypotheses. For example, we hypothesized that the increase in triangular UFO sightings in the years 1980-2010 was caused by the development of 'stealth' aircrafts, which have irregular shapes (see notebook); thus, future work can gather data about reported flights of these aircrafts and try to find correlates to them in the UFO dataset.

Another point that should be considered in future work is the correlation to popular movies. We think people's sightings are influenced by SciFi movies. E.g., the disk-shaped UFO sightings dropped in parallel to the decrease in disk-shaped UFOs in Holywood films, and so on.

And lastly, the reliance on a single source is problematic: The NUFORC's report system can be biased or noisy as it consists mainly of sightings from the US. Thus, adding another source of UFO sightings data would greatly improve the validity of our findings and conjectures.

Conclusion:

We have found no evidence to support our initial hypothesis regarding the effect of sci-fi themed cultural events, mainly because of the difficulty in getting the relevant data (types of UFOs in films, for example). The number of sightings increased dramatically since the early 90's, which we assume is related to the development of the internet and greater ease in relaying these reports. We have also found that sightings in more recent years have become shorter (in duration), possibly due to increasing difficulty in fabricating prolonged accounts of UFO sightings in an age where means for documentation are far more widespread.

Note:

No extraterrestrial beings were harmed in the making of this project (to the best of our knowledge).