IM-Analytics Qualification Test

The X-files problem

objective:

Make a recommendation on where to go to see an U.F.O or interview people that claim sightings.

Code and resource files can be found in:

https://github.com/g-vega-cl/2019_11_28-IntelimetricaRecruiting/tree/master/The_X-Files_Problem

Index

Executive Summary

The final recommendation to the owner of the company is to travel to Los angeles, California to witness an UFO or to interview people who claim to have seen UFOs, from there one can travel south to San Diego.

To come to this conclusion visual methods and frequency analysis were used to find where sightings are most common and where is most likely to be a sighting in the future.

In california there is a strong cluster of sightings around LA and San Diego, with almost 3000 sightings.

Another benefit of traveling to this cities is that since these places are well developed one can travel to and from there easily.

Introduction

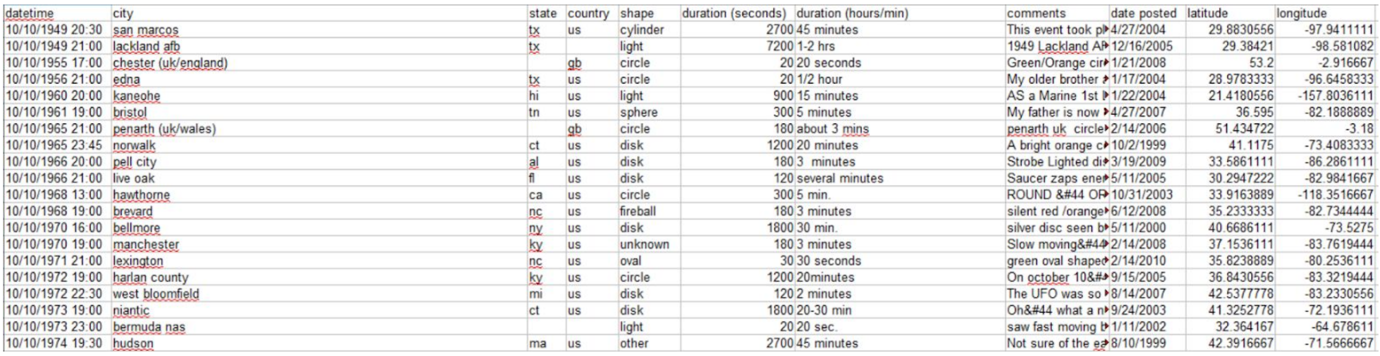
The objective of this problem is to answer the question of where should the owner of the company go to either witness an UFO sighting or to interview people that claim sightings.

This should be done as efficiently as possible since traveling takes time and resources, therefore finding a single location where there is the highest probability of finding an UFO should be priority.

Profiling of data completeness and quality

First look at the data

To profile the data firstly it was opened in a spreadsheet software to review its format and its consistency and to check if there are any clear patterns of incompleteness.



datetime	city	state	country	shape	duration (seconds)	duration (hours/min)	comments	date posted	latitude	longitude
10/10/1949 20:30	san marcos	tx	us	cylinder	2700	45 minutes	This event took pl	4/27/2004	29.8830556	-97.9411111
10/10/1949 21:00	lackland afb	tx		light	7200	1-2 hrs	1949 Lackland AF	12/16/2005	29.38421	-98.581082
10/10/1955 17:00	chester (uk/england)		gb	circle	20	20 seconds	Green/Orange cir	1/21/2008	53.2	-2.916667
10/10/1956 21:00	edna	tx	us	circle	20	1/2 hour	My older brother	1/17/2004	28.9783333	-96.6458333
10/10/1960 20:00	kaneohe	hi	us	light	900	15 minutes	AS a Marine 1st	1/22/2004	21.4180556	-157.8036111
10/10/1961 19:00	bristol	tn	us	sphere	300	5 minutes	My father is now	4/27/2007	36.595	-82.1888889
10/10/1965 21:00	penarth (uk/wales)		gb	circle	180	about 3 mins	penarth uk circle	2/14/2006	51.434722	-3.18
10/10/1965 23:45	norwalk	ct	us	disk	1200	20 minutes	A bright orange c	10/2/1999	41.1175	-73.4083333
10/10/1966 20:00	pell city	al	us	disk	180	3 minutes	Strobe Lighted di	3/19/2009	33.5861111	-86.2861111
10/10/1966 21:00	live oak	fl	us	disk	120	several minutes	Saucer zaps ene	5/11/2005	30.2947222	-82.9841667
10/10/1968 13:00	hawthorne	ca	us	circle	300	5 min.	ROUND &#44 OR	10/31/2003	33.9163889	-118.3516667
10/10/1968 19:00	brevard	nc	us	fireball	180	3 minutes	silent red /orange	6/12/2008	35.2333333	-82.7344444
10/10/1970 16:00	bellmore	ny	us	disk	1800	30 min.	silver disc seen b	5/11/2000	40.6686111	-73.5275
10/10/1970 19:00	manchester	ky	us	unknown	180	3 minutes	Slow moving&#44	2/14/2008	37.1536111	-83.7619444
10/10/1971 21:00	lexington	nc	us	oval	30	30 seconds	green oval shape	2/14/2010	35.8238889	-80.2536111
10/10/1972 19:00	harlan county	ky	us	circle	1200	20minutes	On october 10&#	9/15/2005	36.8430556	-83.3219444
10/10/1972 22:30	west bloomfield	mi	us	disk	120	2 minutes	The UFO was so	8/14/2007	42.5377778	-83.2330556
10/10/1973 19:00	niantic	ct	us	disk	1800	20-30 min	Oh&#44 what a n	9/24/2003	41.3252778	-72.1936111
10/10/1973 23:00	bermuda nas			light	20	20 sec.	saw fast moving b	1/11/2002	32.364167	-64.678611
10/10/1974 19:30	hudson	ma	us	other	2700	45 minutes	Not sure of the ea	8/10/1999	42.3916667	-71.5666667

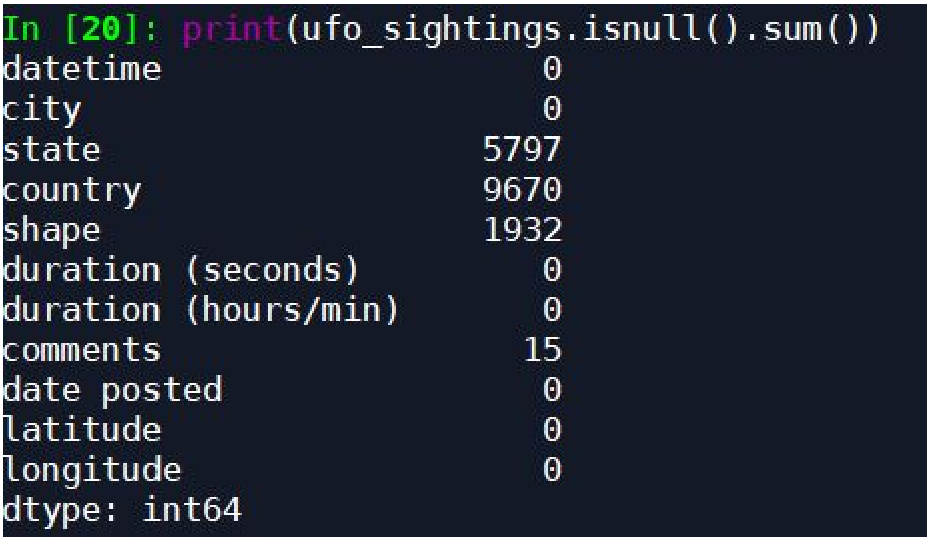
fig 1. basic format of the data.

In this simple view of the data a few incongruencies can be seen, this are:

- The city field has no clear format (some add information like sates, eg. (uk/wales)
- Only US states are filled up, and sometimes this data do not exist.
- Sometimes there is no data at all when it comes to country or state.
- The duration in hours and minutes is not congruent, luckily this data is standardized in minutes.
- The comments have no clear format.

Defining and preparing the best data to work with

A simple function was used to check how many missing or corrupted values there are in the dataset:

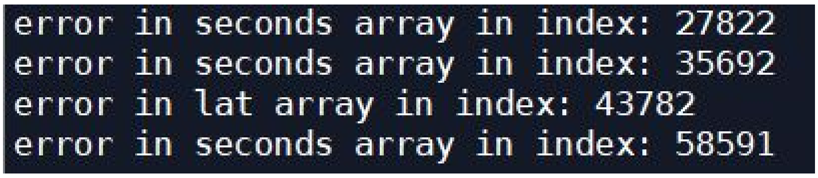


```
In [20]: print(ufo_sightings.isnull().sum())
datetime                   0
city                       0
state                   5797
country                 9670
shape                   1932
duration (seconds)         0
duration (hours/min)       0
comments                  15
date posted                0
latitude                   0
longitude                  0
dtype: int64
```

fig 2. errors in corresponding columns

Since the data that is the most complete is the date, the duration, and the location of the sightings this data will be the focus of our analysis.

Before starting to work with this data its integrity was checked, this analysis mostly consisted on testing if this columns of data had any symbols that should not be there (letters or special characters instead of numbers). In this analysis very few inconsistencies were found.



```
error in seconds array in index: 27822
error in seconds array in index: 35692
error in lat array in index: 43782
error in seconds array in index: 58591
```

fig 3. errors in corresponding indexes

To find out what caused each error the data of each index was printed.
In the first, second and fourth errors there was a “ ` ” character after the number.
The third error consisted in a letter just before the dot: “33q.200088”.

Since there were so few errors in the data considering the size of the dataset, this rows were just deleted.

Regarding the date, it was parsed into a format the compiler could understand, this mostly consisted of a snippet of code that added padding in the months or days where it was necessary and then transformed the date from a string to a datetime format that could be worked on.

To make sure the latitude and longitude data is congruent with its country labels a map visualization was made with random data points so as to see if there are any inconsistencies in this data.

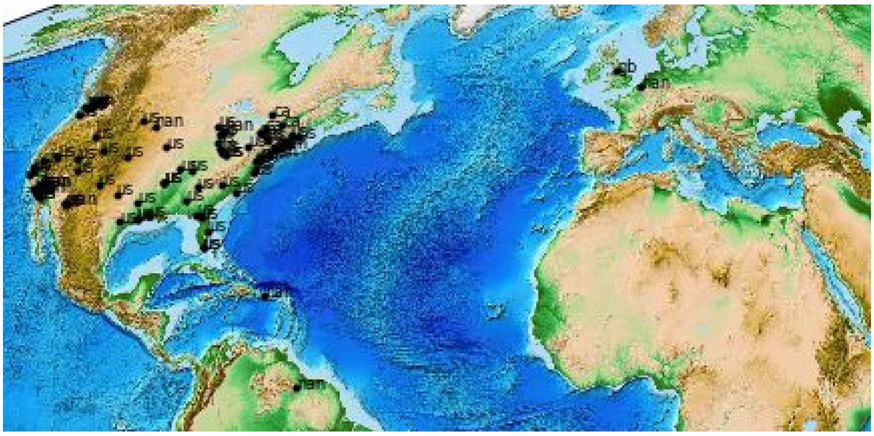


fig 4. countries and labels

With this simple visualization we can see that the labels and the coordinates of the sightings correspond with each other. (Sometimes we have no data)

Descriptive analysis and summary statistics

The analysis made in this section have the purpose of showing if there are any relations between the selected data:

- Date and Location
- Duration and Location
- Duration and Date

Date and Location

To analyze this data the sightings were divided by the following groups of decades:

- 1940 to 1960
- 1960 to 1980
- 1980 to 2000
- 2000 to most recent sighting

To visualize this data the latitude and longitude of each sighting within the time range was plotted in a world map, the results of each decade are shown next:

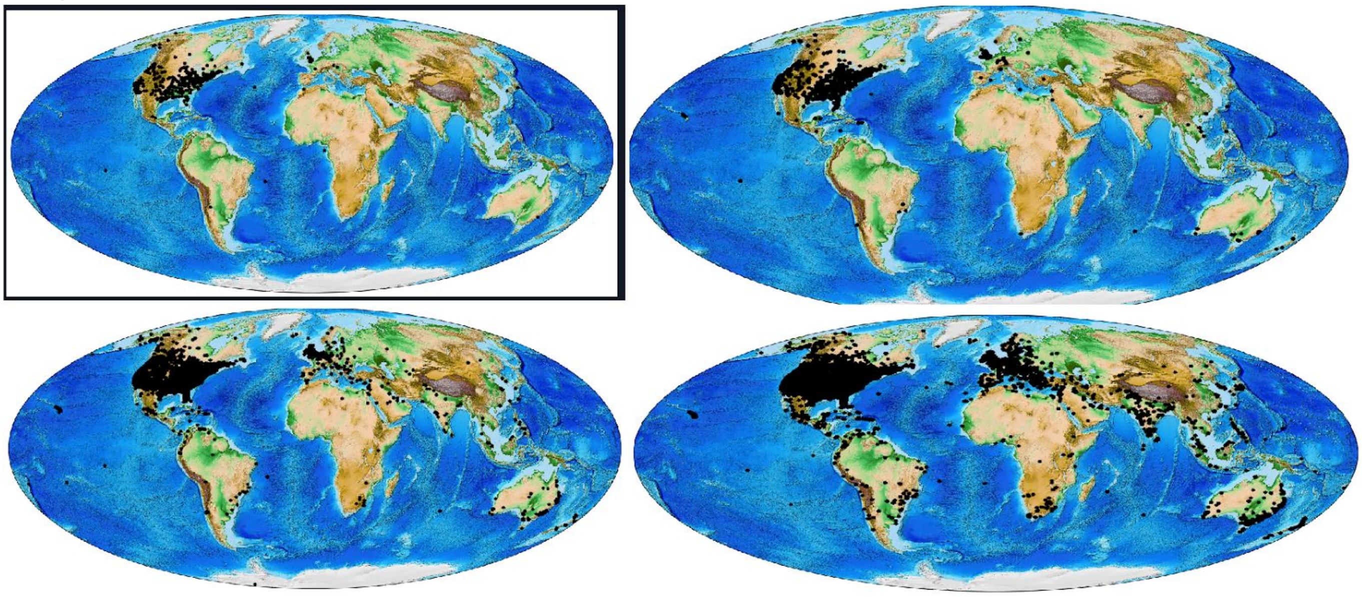


fig 5. geographical distribution of sightings by time group, 1940-1960 (top left), 1960-1980 (top right), 1980-2000 (bottom left), 2000-last sighting (bottom right).

In this maps it can be seen that more sightings have appeared over the world as time passes, (there were around 65’000 thousand from 2000 to the present, against almost 4’000 from 1960 to 1980). And although there have been more sightings around the world in recent times, the US is still the primary location for sightings.

even though there are also other factors that could explain the increase of sightings in recent times, it seems that it’s geographical distribution has not changed throughout time (even if we have sightings in India now, their number is small compared to the increase of sightings in the UK or the US for example).

Duration and Location

Here it was analyzed whether there is a relationship between where the sightings happened and the time they lasted, to do this the mean, the standard deviation and the median of the duration (in seconds) of sightings in the continental US, as well as the rest of the world were calculated:

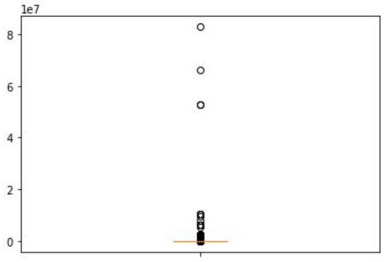
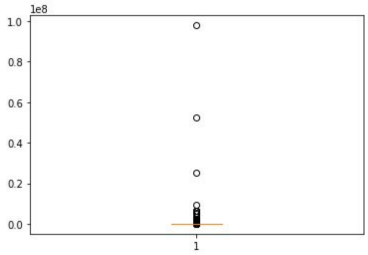
	Continental US	Rest of the world
Mean	578’400	691’800
Standard Deviation	8’320	10’230
Median	180	180
Box Plot		

Table 1. Statistical data of sighting length against sighting location

With this data one can deduce that the Duration and the location of the sightings are not really correlated with one another, since the characteristics of the data are very similar to each other.

In case of the boxplots, their purpose was to show the outliers of the data, in this case, it is clear that the outliers have no correlation with the location of the sighting

Duration and date

This same analysis can was done regarding the duration of the sightings and the time of the sightings, the results are shown in the following table:

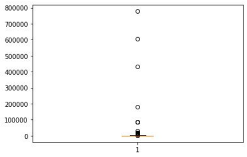
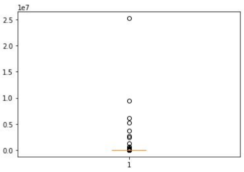
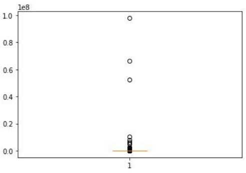
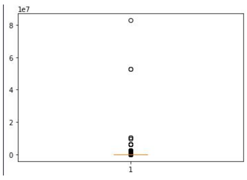
	1940-1960	1960-1980	1980-2000	2000-2020
Mean	5’082	17’550	24’660	5’800
Standard Deviation	46’900	471’750	1’228’600	448’800
Median	300	300	180	180
Box Plot				

Table 2.Statistical data of sighting length against sighting date

There is a variation in the duration with the pass of time, but this data shows no trend or structure, therefore in this report it will be considered that the duration has no correlation with the date.

In case of the boxplots, their purpose was to show the outliers of the data, in this case, it is clear that the outliers have no correlation with the dates of the sightings.

Hypotheses and Modeling

Hypothesis and results

To answer the main question of the owner of the company, the hypotheses is that certain geographical locations have a higher likelihood of witnessing a sighting, to do this the data was clustered by their longitude and latitude.

After doing this a simple histogram showed that there were locations where the frequency of sightings is higher, using this histogram a frequency value was selected to filter the sightings and a second histogram was built:

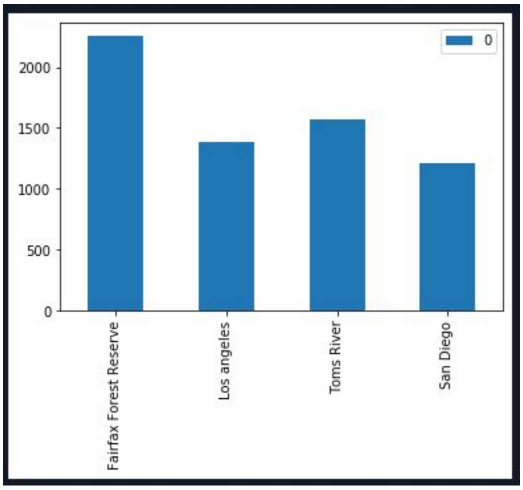


fig 6. Histogram of locations with most sightings

This histogram shows that the Fairfax Forest reserve has the highest frequency of sightings, but by checking the coordinates of this locations it showed that Los Angeles (Latitude: 34.5, Longitude: -118.5) and San Diego(Latitude: 33.5, Longitude: -117.5) are very close together, where between them is Hillgrove (Latitude: 34, Longitude: -118).

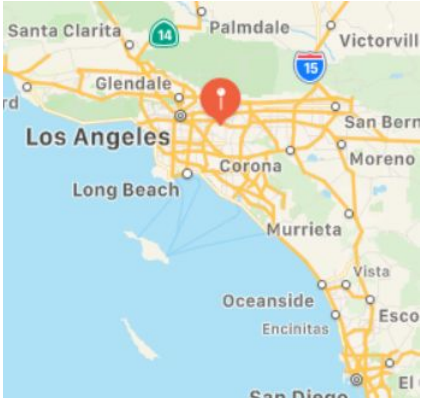


fig 7. Location of Los Angeles and San Diego

To make sure no cluster of locations was overlooked a new map built only with the top 9 highest frequency locations.

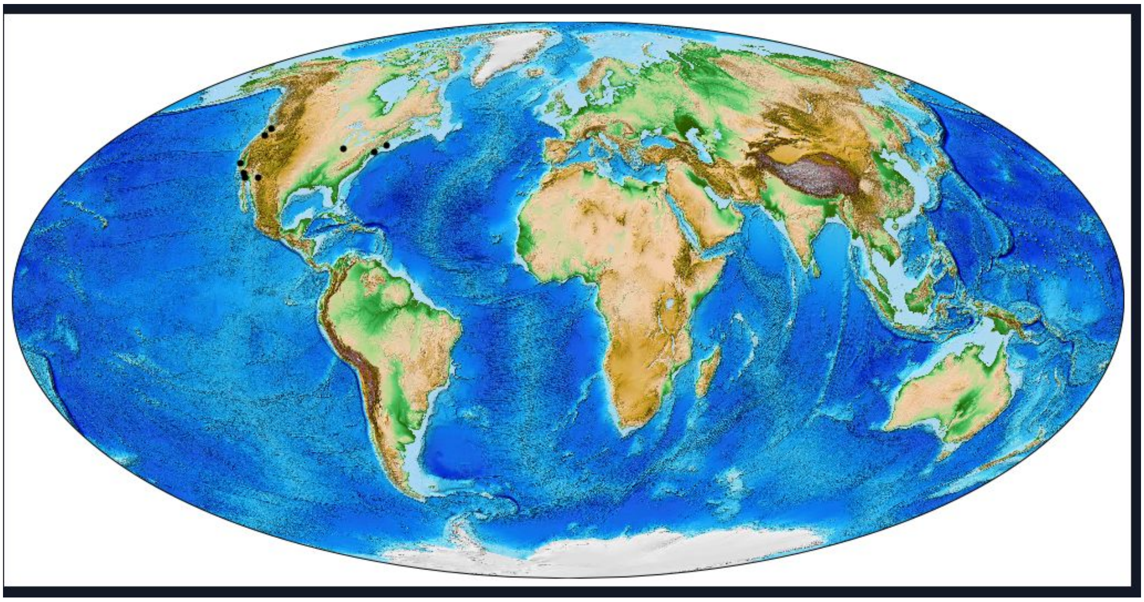


fig 8. Location of top 9 highest frequency locations

The map shows how even though there are a couple more clusters of high-frequency locations, the sightings are more concentrated in the California area.

With this information the recommendation of where to travel next to either witness an UFO sighting or to interview people that claim sightings is Los angeles, California, and travel south to San Diego from there.

Model discussion and advantages/disadvantages

The model used was a simple clustering and frequency analysis. This model was chosen mostly because of it simplicity to follow and build in a short amount of time, which has advantages as well as disadvantages, for example:

advantages of the model

- Simple to build and visually understand: This model was built with a visual and qualitative approach, which makes it easy to explain and present to an audience.
- The model does not require knowledge of complex statistical formulas or computational algorithms, which makes it accessible to a broader audience.

disadvantages of the model

- Little hard quantitative justification in certain aspects: some values were taken from qualitative criteria that might be difficult to reproduce or justify if needed.
- Outliers were not cleaned: for the sake of properly visualizing every sighting, some outliers were not removed from the data, which causes distortion in other statistical values (eg. averages).
- The comments importance of each sighting is overlooked.

Conclusion and next steps

Conclusion

Even with its simplicity, the recommendation given in this analysis (Go to Los angeles, California, US and travel south to San Diego from there) has solid bases and should be taken.
Data was properly analyzed for relationships and a sound approach based on location was followed.

Next steps

Regarding the next steps to analyze this data there are a couple recommendations to be made:

- Improve the date analysis with more clusters, instead of just using 4 time periods, sightings could be divided per year, or even per month to see if there is any seasonal pattern in the data.
- Clean outliers from the data.
- More robust statistical models could be applied (eg. ANOVA).
- Relatively new and experimental Unsupervised and Reinforcement learning models could be used to find underlying clusters (eg. K-means) or to build a model to predict the most likely locations of the next sightings (eg. deep-q-learning)