IM-Analytics Qualification Test

Gustavo Fuentes

October 2019

Contents

Chapter 1

The X-files problem

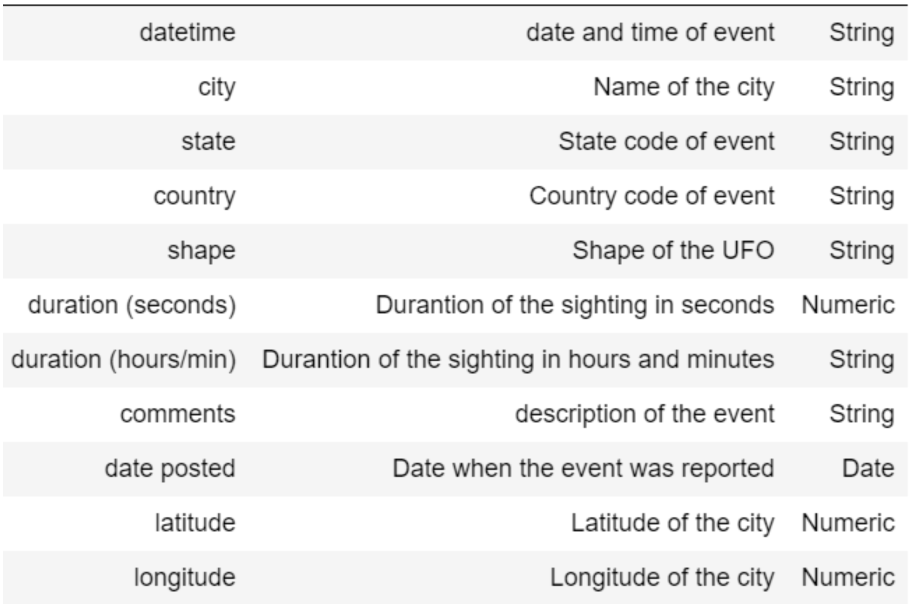
1.1 Introduction to the problem

A man who collects UFO data wondering where he could go to see one of these events or interview people that claim sightings. Our work is to make his dream come true.

1.2 Data

The data has been collected from people around the world consequently, the data is impure and it has a lack of information or misinformation.

Data features description:

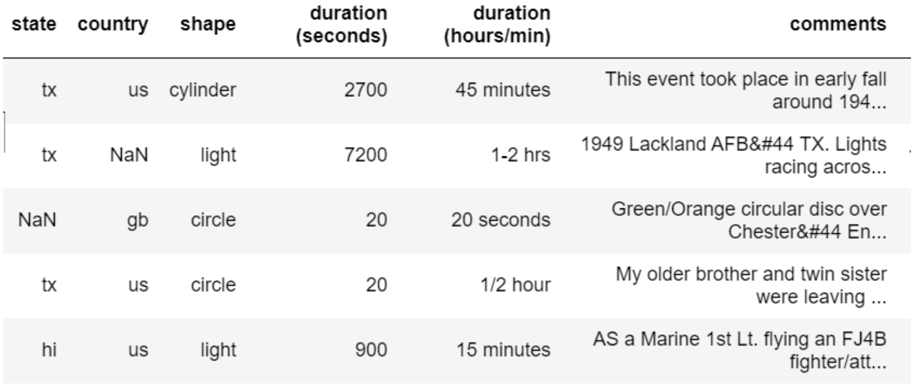


datetime	date and time of event	String
city	Name of the city	String
state	State code of event	String
country	Country code of event	String
shape	Shape of the UFO	String
duration (seconds)	Durantion of the sighting in seconds	Numeric
duration (hours/min)	Durantion of the sighting in hours and minutes	String
comments	description of the event	String
date posted	Date when the event was reported	Date
latitude	Latitude of the city	Numeric
longitude	Longitude of the city	Numeric

There are some missing values as NaN. These values could be either numeric or strings.

The data has 80332 samples by 11 features. If we find a NaN value we delete the entire row, doing this the lack of information represents around 18% the data.
We are looking for replace this values doing some statistics. As a first approx, we are going to deal with the 83% remaining and collect the data by the UFOs *shapes*.

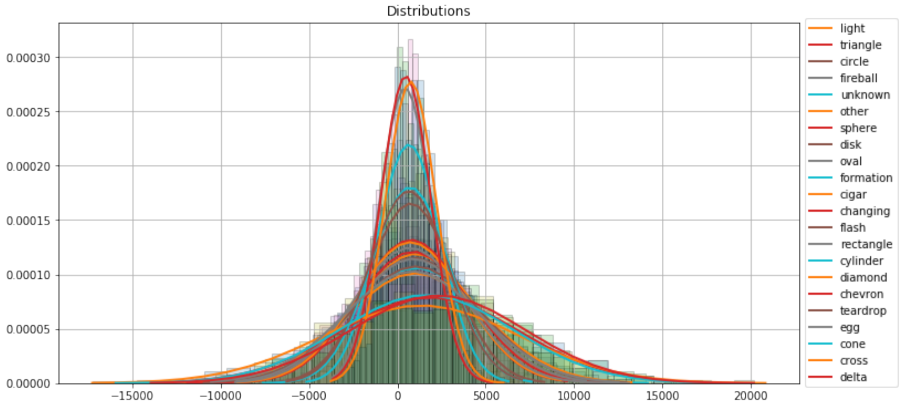
Such as: 'light', 'triangle', 'circle', 'fireball', 'unknown', 'other', 'sphere','disk', 'oval', 'formation', 'cigar', 'changing', 'flash', 'rectangle', 'cylinder', 'diamond', 'chevron', 'teardrop', 'egg', 'cone', 'cross','delta', 'round', 'crescent', 'pyramid', 'flare', 'hexagon', 'changed'

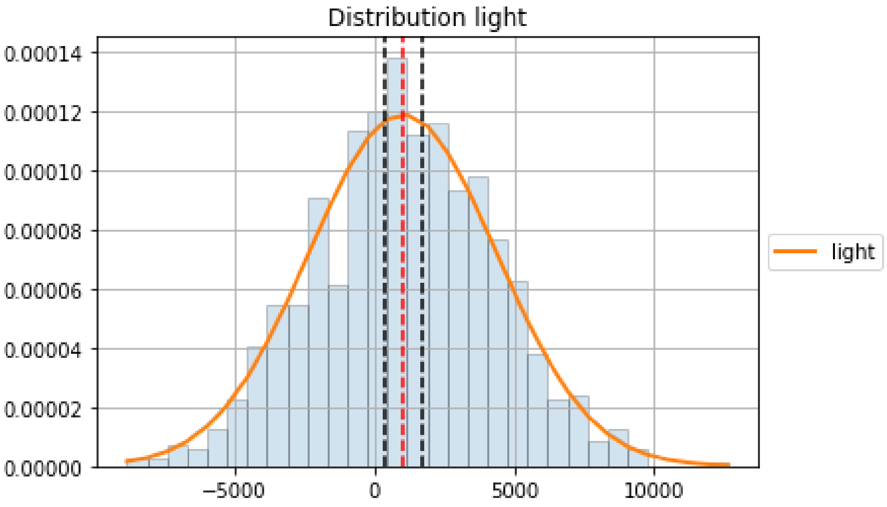


state	country	shape	duration (seconds)	duration (hours/min)	comments
tx	us	cylinder	2700	45 minutes	This event took place in early fall around 194...
tx	NaN	light	7200	1-2 hrs	1949 Lackland AFB&#44 TX. Lights racing acros...
NaN	gb	circle	20	20 seconds	Green/Orange circular disc over Chester&#44 En...
tx	us	circle	20	1/2 hour	My older brother and twin sister were leaving ...
hi	us	light	900	15 minutes	AS a Marine 1st Lt. flying an FJ4B fighter/att...

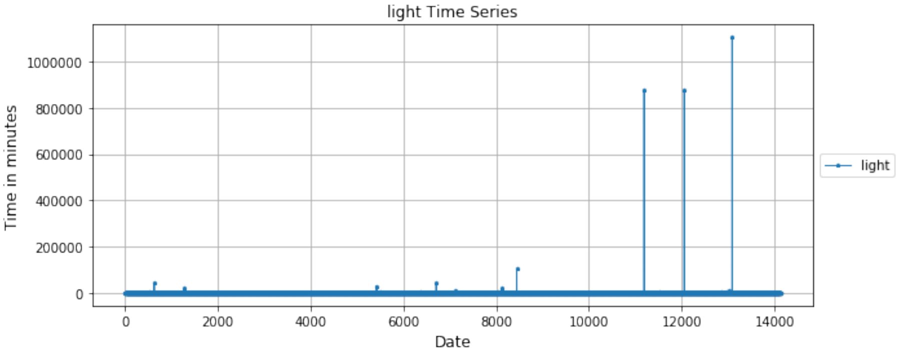
1.3 Some statistics

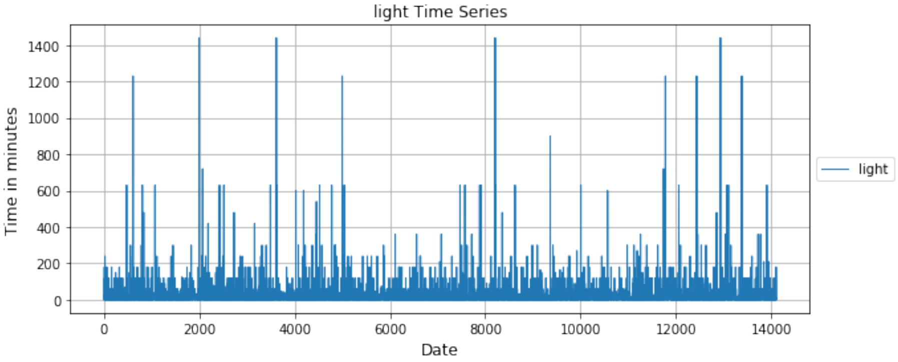
We will use mean μ and standard deviation σ to estimate a Gaussian distribution





Where the mean is the center, red line, and the standard deviation the width, in these case the black lines represents $\mu \pm \frac{1}{5}\sigma$

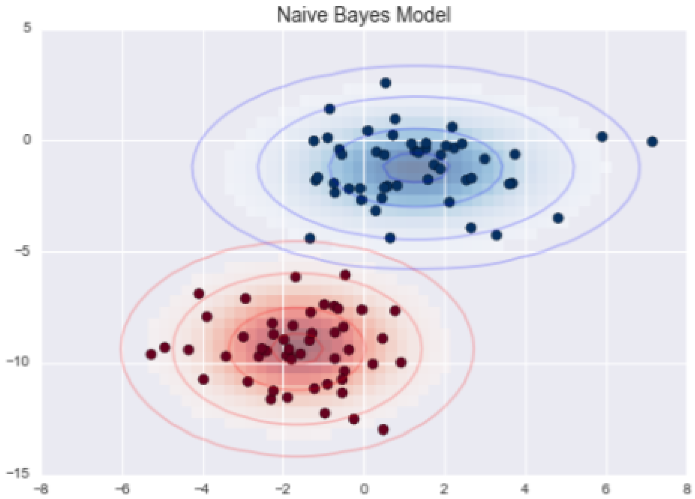




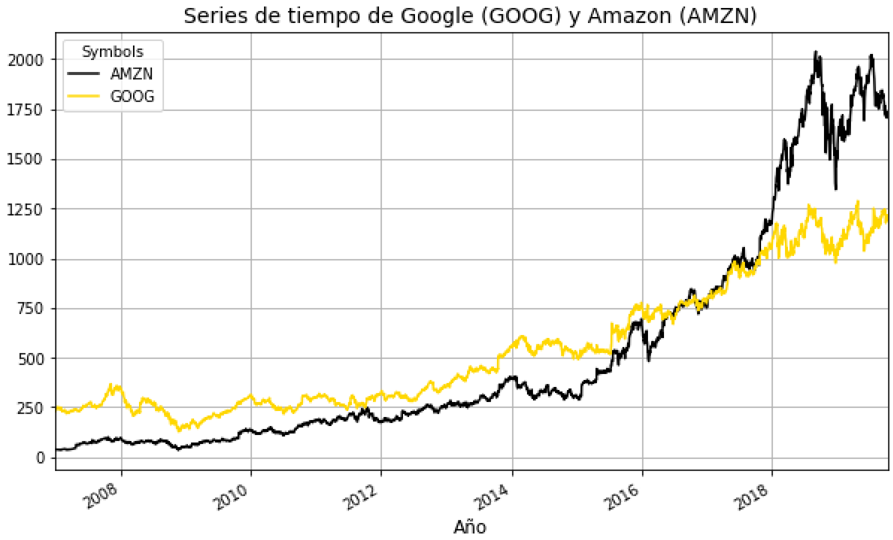
1.4 Hypotheses and Modeling

1.5 Conclusion and next steps

Conclusions so far, if you want to see some UFOs, the best a place is California with around 13% of chances, and you will see lights as the most probably shape. Why is this model bad ???. I would not say bad at all, apparently there is not relation between shape and time, I would see a ship or a similar object I would expect this object remains more time visible than a flashing light. Lights is the most probably shape due to in the night any brilliant object is very visible. My impressions so far, this data is impure. In my experience, It is frequently find patterns working with data and humans. I can do something else What is next ?? We can classify our data by state, country or shape given a state, what is the probability to see certain shape? or given a shape, what is the probability to see it in some state? This is called a Naive Bayes Method. I did clusters to collect duration given a shape. Time Series Analysis. This data in particular is sorted by dates, We can



obtain time series by states and look for correlation, trend, pattern and more.


Series de tiempo de Google (GOOG) y Amazon (AMZN)
Symbols
AMZN
GOOG
Año