UFO sightings analysis

Andrew Jalnine

Abstract

Used UFO sightings dataset, geographical datasets and USA weather archive

Research questions:

1. Find patterns in distribution of UFO sightings
2. Find relations between UFO sightings and weather conditions

Methods:

Visualizations, cluster analysis, classification, distribution analysis, etc

Analysis covers 2010-2014 years

Motivation

Any unexpected object in atmosphere can be dangerous to aircrafts.

At least because of this UFO phenomena exploration is important.

Datasets

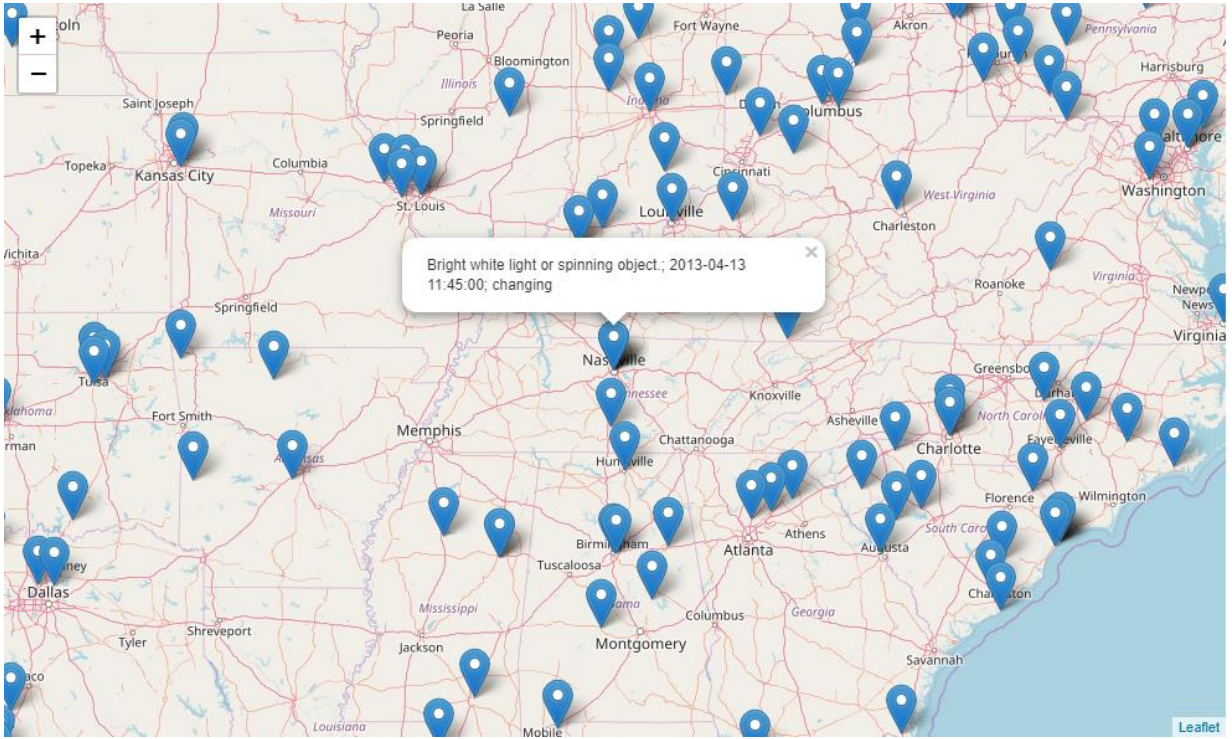
1. US cities coordinates. Found at simplemaps
2. Weather archive. Contains worldwide daily measures with related weather station codes. Used data for 2010-2014 years. Found at National centers for environmental information. Contains measures: precipitations, snowfalls, temperatures.
3. Weather stations metadata. Contains station codes and coordinates. Found also at National centers for environmental information

Datasets (continue)

4. UFO sightings:

Contains coordinates, timestamp and brief description

Found at [Kaggle]



Data Preparation and Cleaning

Many data transformations implemented, because analysis based on several independent heterogeneous datasets. Used dataset merging, pivot, concatenation, filtering, grouping operations etc.

UFO sightings dataset contains mistakes of hand input and required checking, parsing, error filtering and type convertions.

Main problem is large size of weather data (before any transformations and filtering is about 6 Gb). Due to RAM limitations used chunked operations.

Binding each sighting to closest weather station or city requires join tables by minimal distance on earth (calculated via haversine formula), that is very slow operation and required some optimisations.

Research Questions

1. Find patterns in distribution of UFO sightings. Includes subquestions:

    Check if distribution can be divided on clusters

    Determine places with highest density of sightings

2. Try to find relations between UFO sightings and weather conditions.

If model can predict UFO sighting by generic weather parameters with low errors, then sighting can be classified as weather phenomena

Methods

1. Cluster analysis. Used K-means with determination of number of clusters by elbow method. Clusters visualized on map with confidence circles which radius calculated by cluster distances standard deviation.

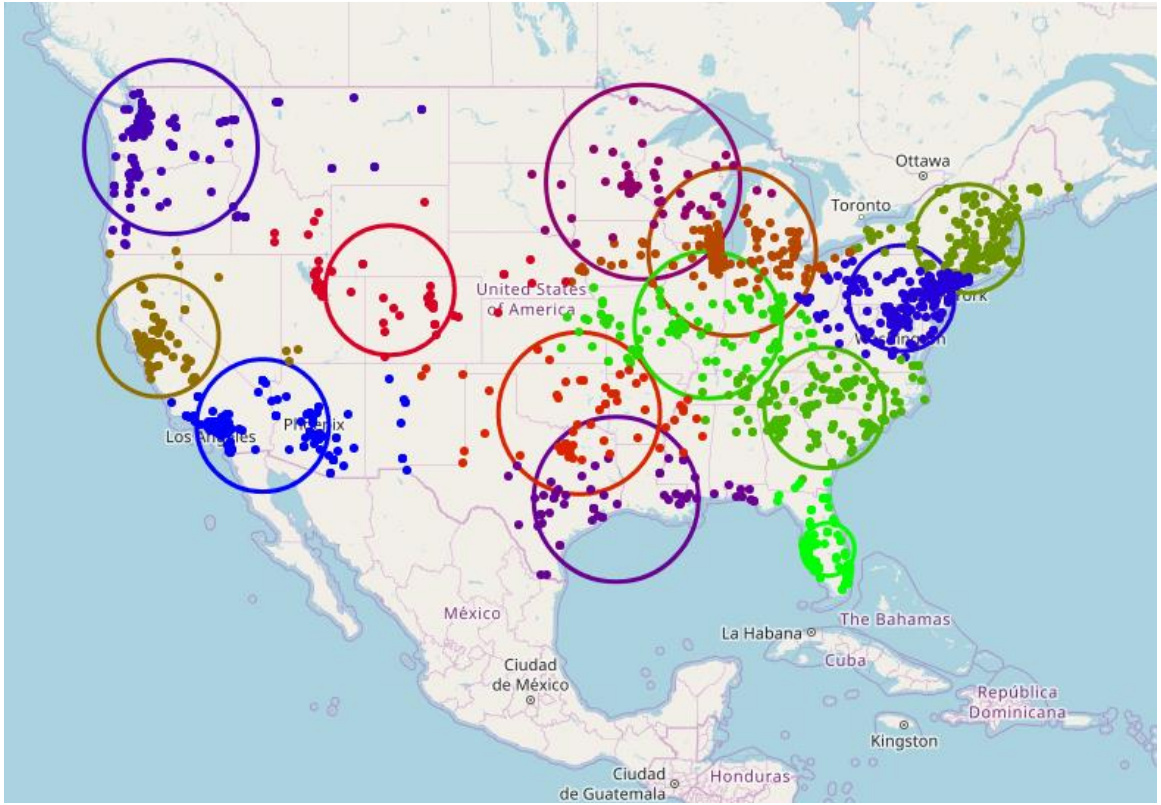
2. Basic statistics, analysis of distributions. Too wide latitude range can reduce weather model quality because of different climatic conditions, this requires latitude distribution analysis and data filtering).

3. Decision Tree Classification. Used to find relations between weather parameters and UFO observations. By model error we can determine how much part of sightings is weather phenomena.

4. Visualizations – maps, violin plots, etc

Findings

Found that UFO sightings can be clustered on map with optimal number of clusters **k=15**



Findings

Clusters formed

around this cities*:

Anchorage
Beaumont
Chicopee
Daly City
Decatur
Gary
Grand Junction
Lancaster
Lodi
McKinney
Poinciana
Rock Hill
Woodbury
Yakima

*with population at least 50 000 peoples

Findings

Cities* with highest sighting rate:

Around Wilmington and Portland

average sighting rate about

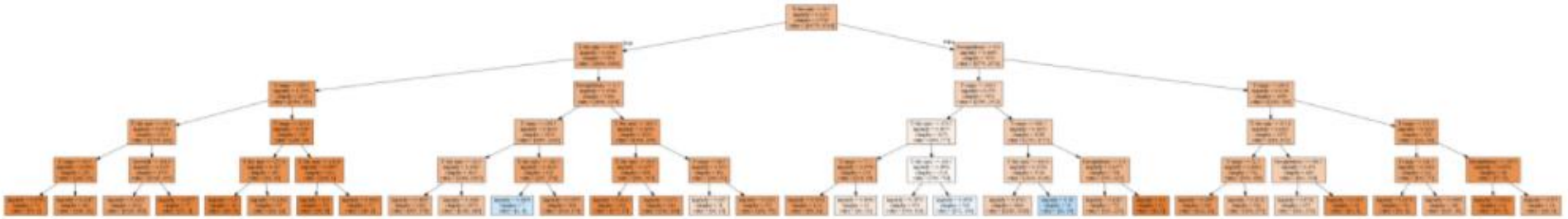
1 sighting per week

*with population at least 50 000 peoples

Findings

Decision tree classification **can’t** predict UFO sighting by weather conditions

Only 1.75% of predictions is correct for test data (depth of tree = 5).



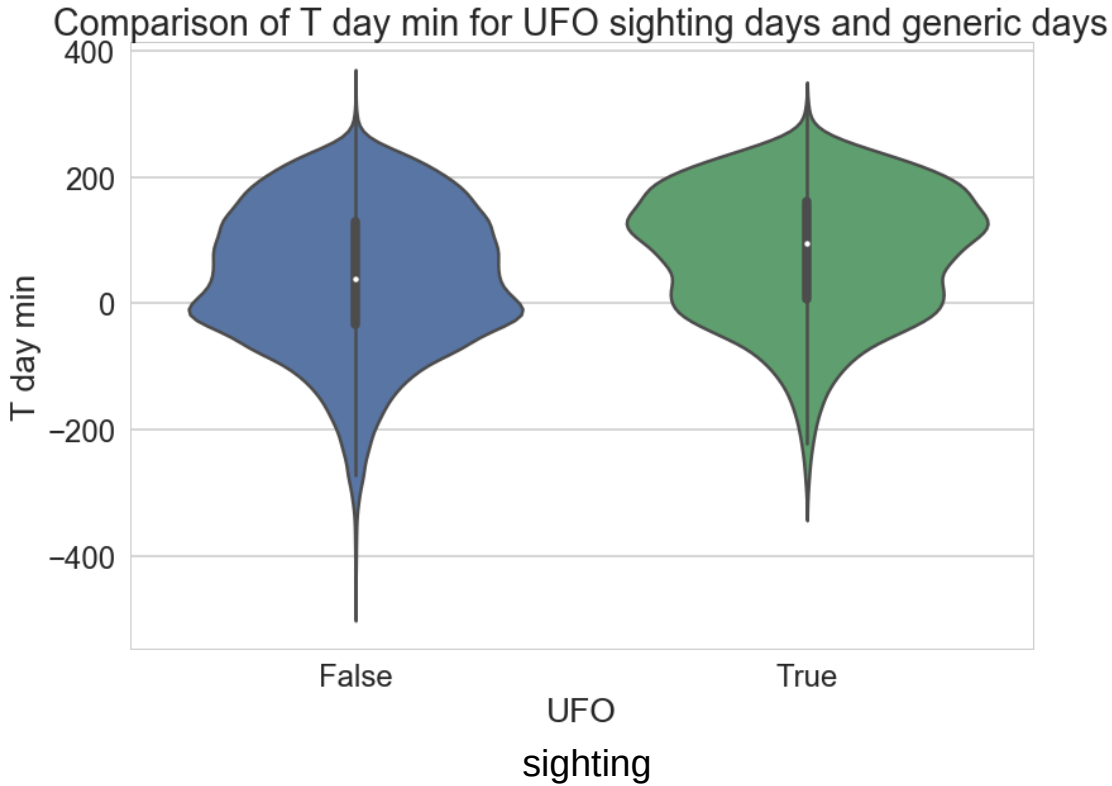
Findings

Sightings often occurs

in **warm** days

It can be explained

by farther visibility



Limitations

Many stations does not records other parameters, and filtering dataset by non-empty values for all parameters (in case of wider parameters set) reduces dataset to unappropriate size. Because of this, only few weather parameters used for analysis.

With per-hour weather data results would be more correct, but this requires more resources than available.

Conclusions

1. UFO sightings can be clustered around 15 cities
2. Found cities with very high sighting rate (about 1 per week)
3. UFO sighting can not be predicted by generic weather parameters
4. UFO sighting often occurs in warm days
5. Deeper analysis requires more resources than available

Acknowledgements

Thanks to:

- Kaggle,
- National centers for environmental information,
- Simplemaps

for free access to datasets

References

NumPy is the fundamental package for scientific computing with Python.

pandas: powerful Python data analysis toolkit

Scikit-learn Machine Learning in Python

Matplotlib is a Python 2D plotting library

Folium

Seaborn: statistical data visualization