# Close Encounters of the First Kind

## Data Science (Visualization & Storytelling) Take Home Exercise

Shahbakht Hamdani
10th August 2021

# Executive Summary

- Bright lights and flashes most easily reported as UFOs, typically lasting less than 5 minutes.
- Fourth of July and New Year's spark high number of reporting.
  - July is, by far, the month with highest reported sightings.
- Large, metropolitan cities and coastal areas drive highest number of sightings.
- Data in NUPORC is predominantly from Western countries (North America and Western Europe), but interest in UFOs is evident across the globe, especially in India, Australia & New Zealand, South Africa, and Brazil.

# Data Cleaning & Preprocessing

# Overview of Dataset

## Description of dataset and missing values pattern

- Shape of **raw** dataset is: **88,125 rows**, **12 columns**
- The number of values missing from individual columns are shown in Fig 1 with length of grey bar indicating **availability**:
  - **Latitude** and **longitude** are missing the most values (~16,000 entries)
  - **Shape**, **State** and **Duration** are also missing substantial number of entries (>2,000 each)
  - **Date_time** and **posted** columns are missing equal number of entries (1,187)
- In **Fig 2,** we look at the pattern of missing values, with dataset sorted by **date_time**:
  - **Latitude** and **Longitude** are always missing together, as are **date_time** and **posted:** this checks out as these values are reported in pairs (can't exist independent of each other).
  - The other columns are missing entries at random (as far as time-related patterns are concerned)

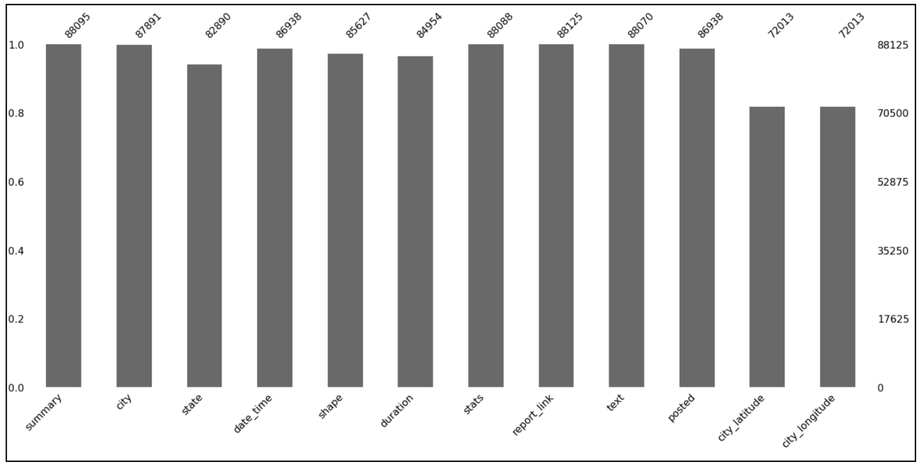



Fig 1: Missing values per column as number

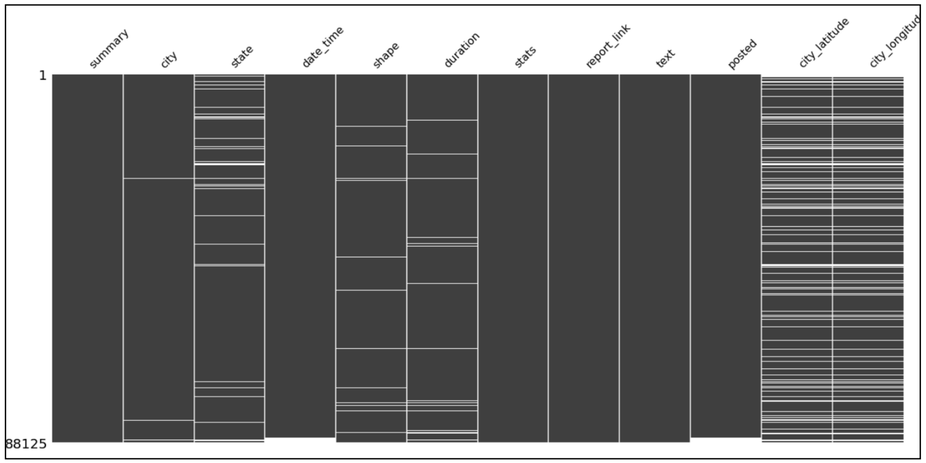



Fig 2: Missing values pattern

# Data Cleaning & Preprocessing (1/2)

**Converting date column type, removing & filling missing entries, extracting information from free-text fields**

- **Date_time** and **posted** columns were converted into **datetime64[ns]** which is **ISO 8601** representation of date time in Pandas dataframes.
- **Dropped the rows that did not have any date_time information captured, resulting in losing ~1,187 entries.**
- Filled missing entries in “**Shape”** column with “unknown” value, which is an existing category in the dataset. Shown in Fig 1.
- **Duration** column is a free-text field with user-reported entries without any format. In order to extract numerical values out of it, following transformation was done:
  - Looked for substrings “**min**”, “**sec**” and “**hour**”/“**hr**” in each entry
  - Created a separate column for each granularity of duration captured, and extracted the numerical value using **regex**
  - Converted all values to seconds (min * 60, hour * 60 * 60)
  - Clipped the upper tail of distribution to 95th percentile: [0,95] retained in order to minimize extremely large values. Any value above 95th percentile is treated the same.
  - Filled the missing duration entries with **median value.**
  - The distribution is shown in **Fig 2.**

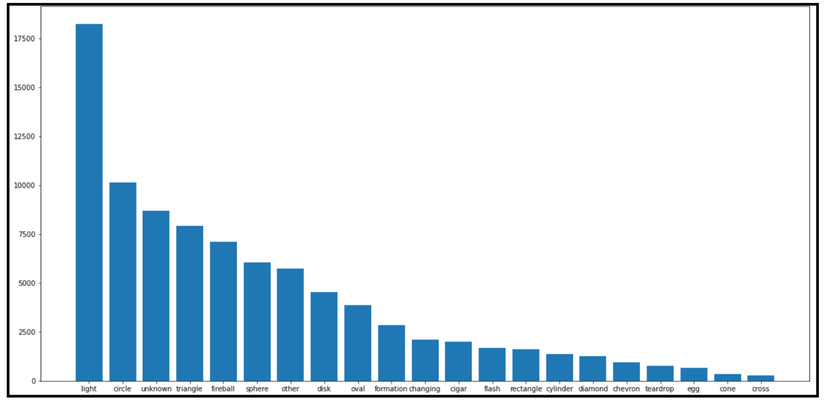



**Fig 1: Shape category count**

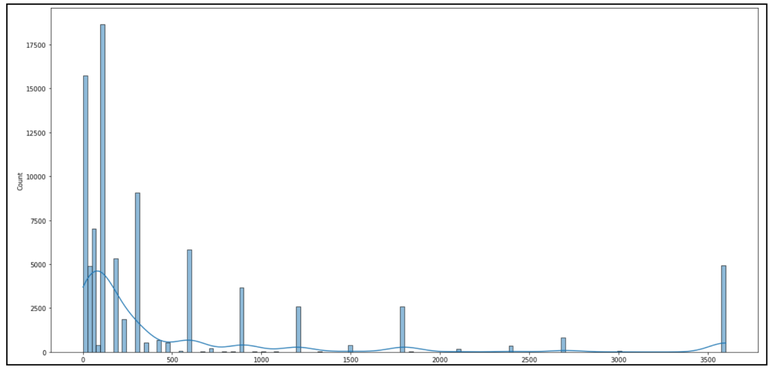



**Fig 2: Histogram of Duration (in seconds)**

1. Most values are within the 0-300 seconds range (< 5 minutes)
2. The peak at 3600 seconds is due to 95th percentile cut-off.

# Data Cleaning & Preprocessing (2/2)

## Getting “country” from location and cleaning city names

- In order to extract “**country**” from the data, following steps were taken:
  - Columns used for extraction:
    - State
    - City
  - **First pass**: using the *pycountry* library of Python, compared the **state** code with the ISO 3166-2 country subdivision code for **United States**
  - **Second pass:** same process was done for **Canada,** as Canadian states were available with their state codes
  - The first two passes captured around ~80,000 entries
  - **Third pass:** extracted the name of country that was wrongly inputted in **city** column (See Fig 1 for examples)
  - **Fourth pass:** for the country names extracted, check if all the names match with their ISO naming
- For **city column**
  - Clean the entries by removing country names from city. (see Fig 2)

| city |
| --- |
| Pleiku (Viet Nam) |
| Quang Tri (Viet Nam) |
| Catania (in countryside) (Italy) |
| Dorood (Iran) |
| Durban (South Africa) |

**Fig 1: Raw City names containing country information**

| city |
| --- |
| Pleiku |
| Quang Tri |
| Catania |
| Dorood |
| Durban |

**Fig 2: Cleaned city names**

# Cleaned Dataset

- After removing missing entries from date, and cleaning other columns, final shape of dataset is:
  - **86,938 rows** (down from 88,125 rows)
  - **13 columns** - date_time is index (“**country**” and “**duration_cleaned**” column created)
  - All other temporal features were extracted from date: hour of day, month of year, etc.

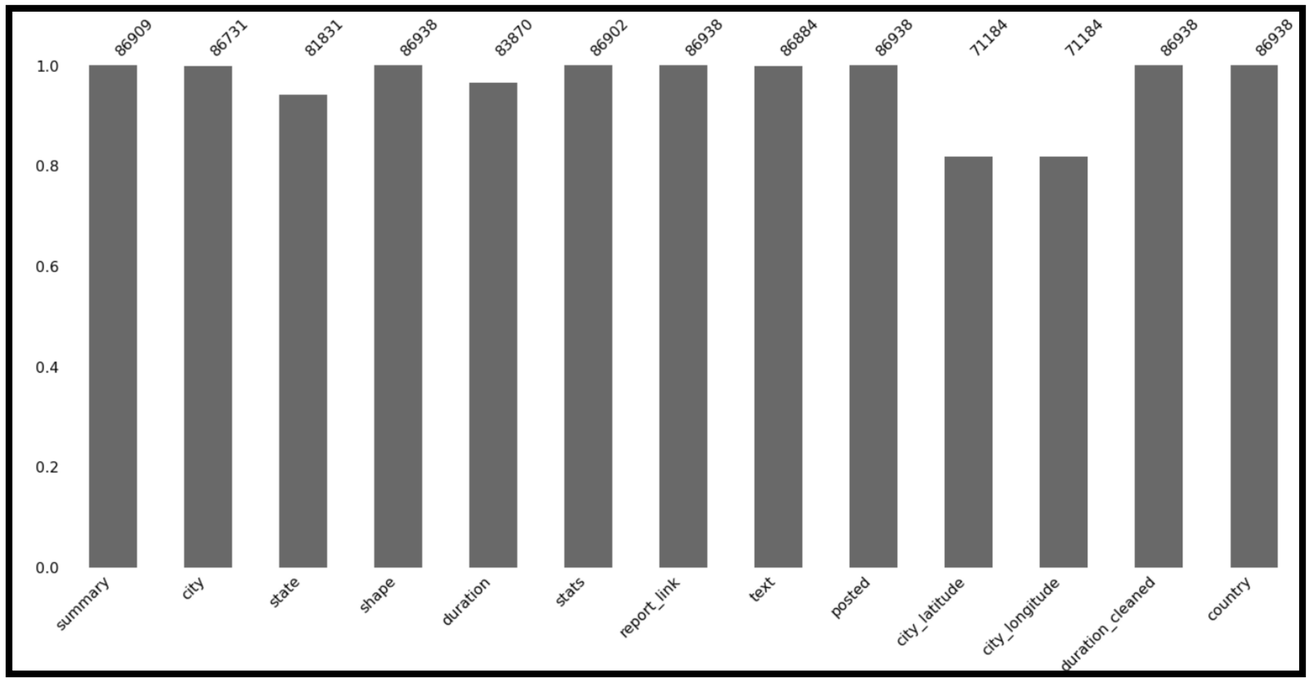
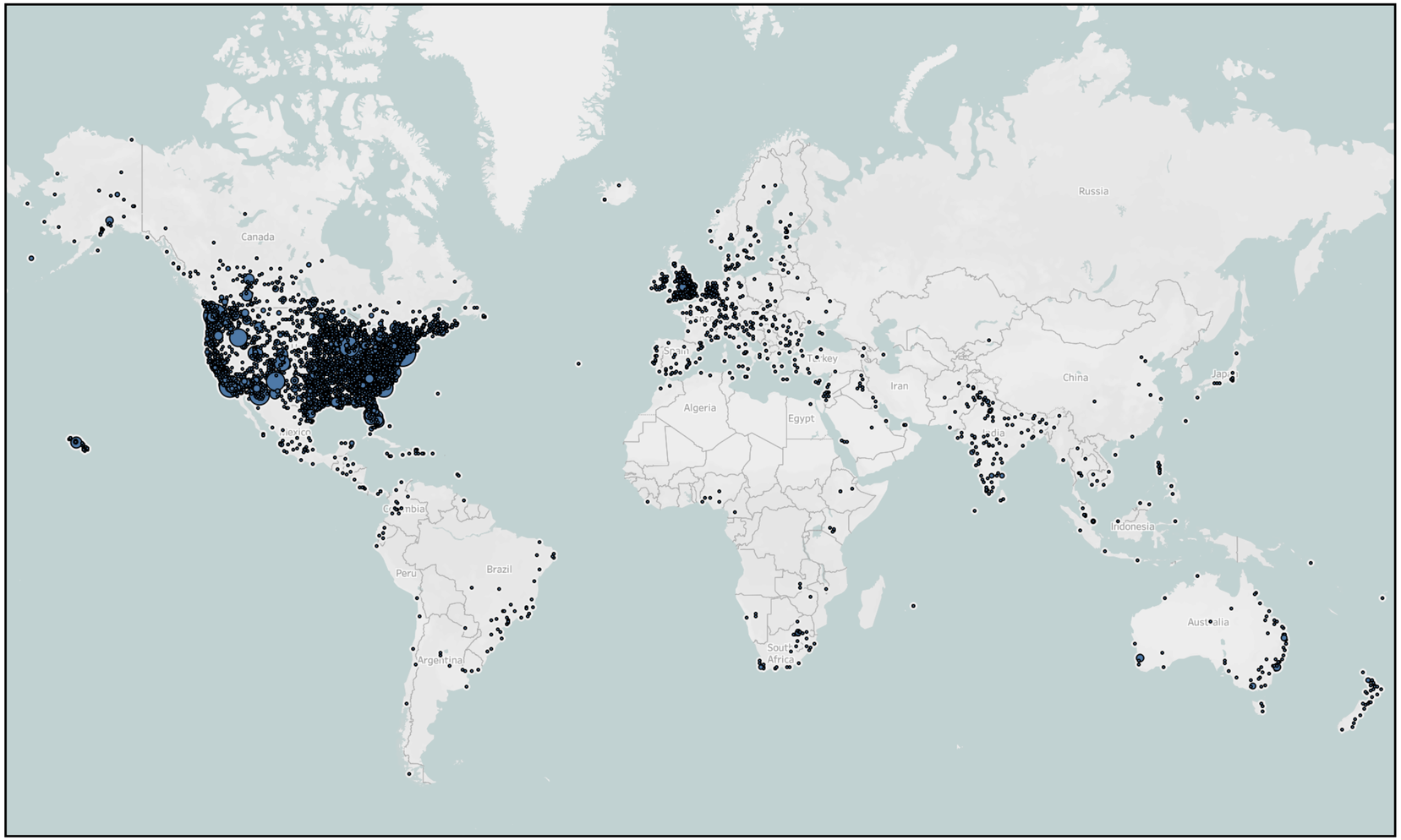



**Fig 1: Missingness of Cleaned Data Set**

# Exploratory Data Analysis: Insights

# Geographical Analysis: Worldwide

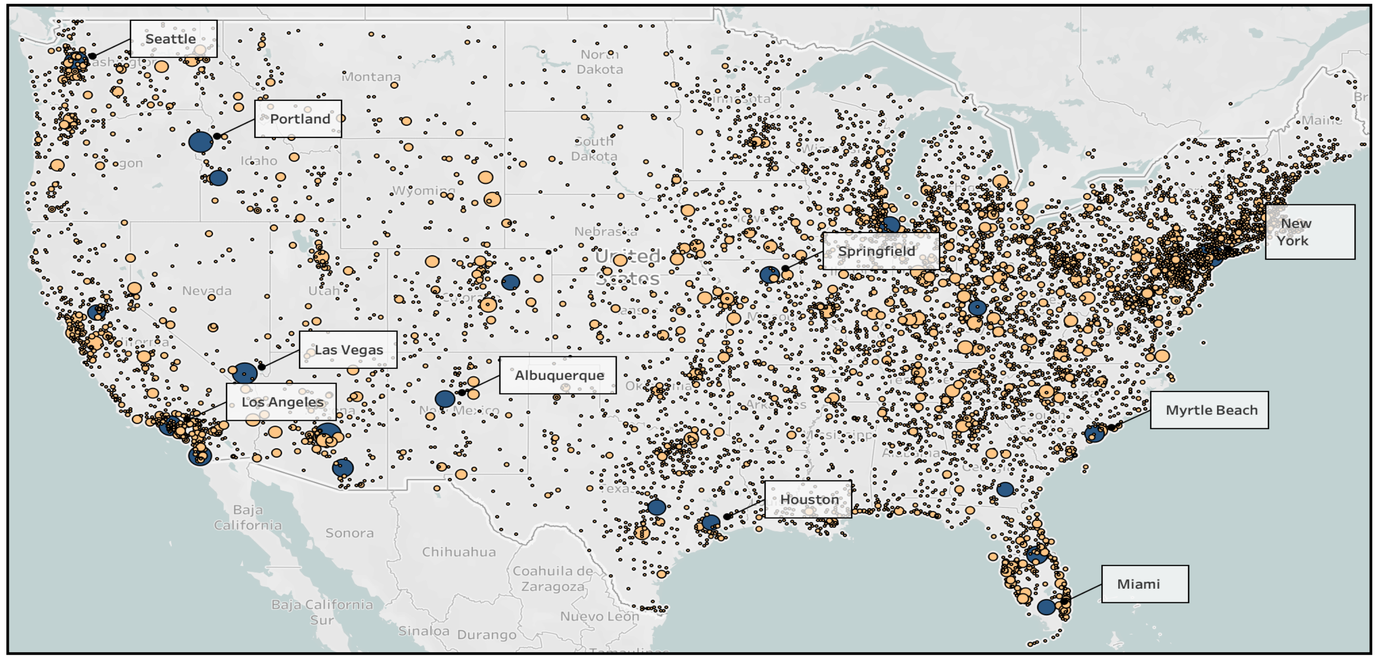
## Global concentration of sightings

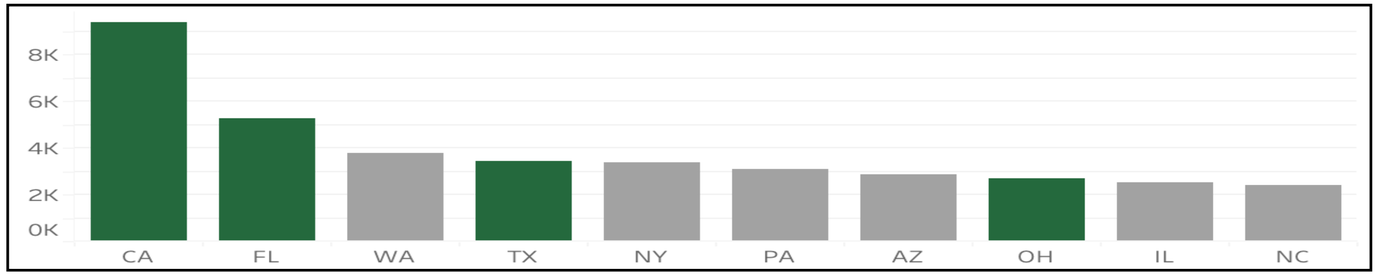



- Map on the left shows count of sightings by city and country:
  - Majority of sightings are from **United States** (~90%), with **Canada** and **South America** also having considerable sightings
  - **Europe** (with majority of **Western Europe**) is another hub
  - In Asia, **India** is leading the pack
  - **Australia** and **New Zealand** also have considerable sightings
- **Coastal areas** have much higher proportion of sightings than inland areas.
  - Noticeable both inside and outside US

# Geographical Analysis: United States (1/2)

## Cluster of Sightings



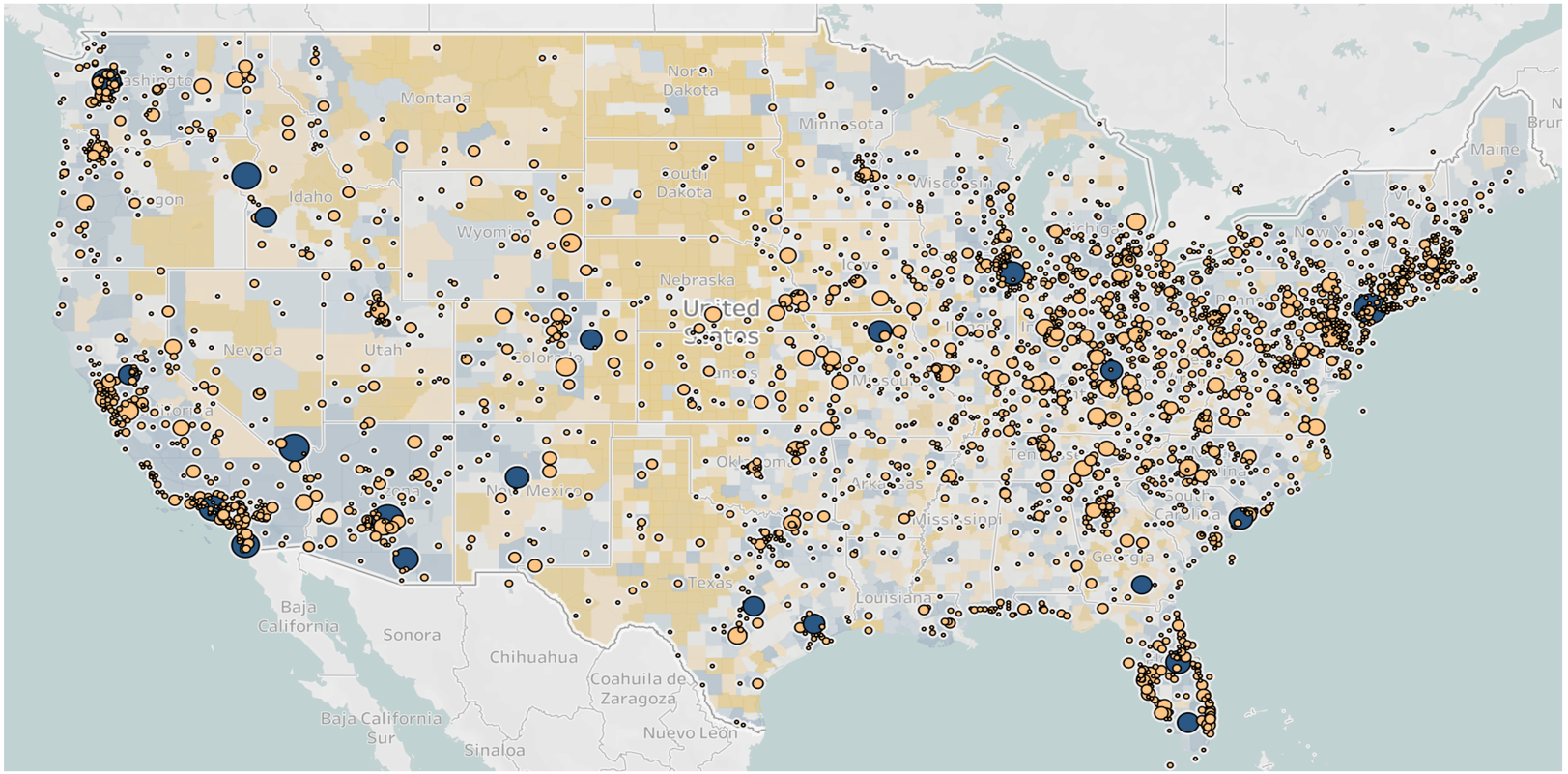




**Top 10 States with most sightings, with green indicating NASA presence.**
**4 out of 10 have either a NASA visitor center or a NASA field center or both.**

- **Blue** dots show **sightings greater than 200.**
- **Big cities/metropolitan areas** have much higher sightings, some examples are called out in the map.
- Clustering is clearly visible in **coastal cities**, and particularly in mid to eastern part of the country:
  - east coast: Miami, NYC, Myrtle Beach
  - west coast: LA, Seattle
  - south or inland: Houston, Austin, Portland,
  - **Albuquerque: interest in UFOs may be driven by the famous Roswell UFO incident which also took place in New Mexico (200 miles away)**
- States with NASA presence have high sightings as well:
  - **NASA Visitor Centers**: California, Florida, Alabama, Ohio, Texas, Virginia, Mississippi
  - **NASA Field Centers:** Virginia, Silicon Valley, Ohio, Los Angeles County, Alabama, Maryland, Mississippi, Texas, Florida

# Geographical Analysis: United States (2/2)

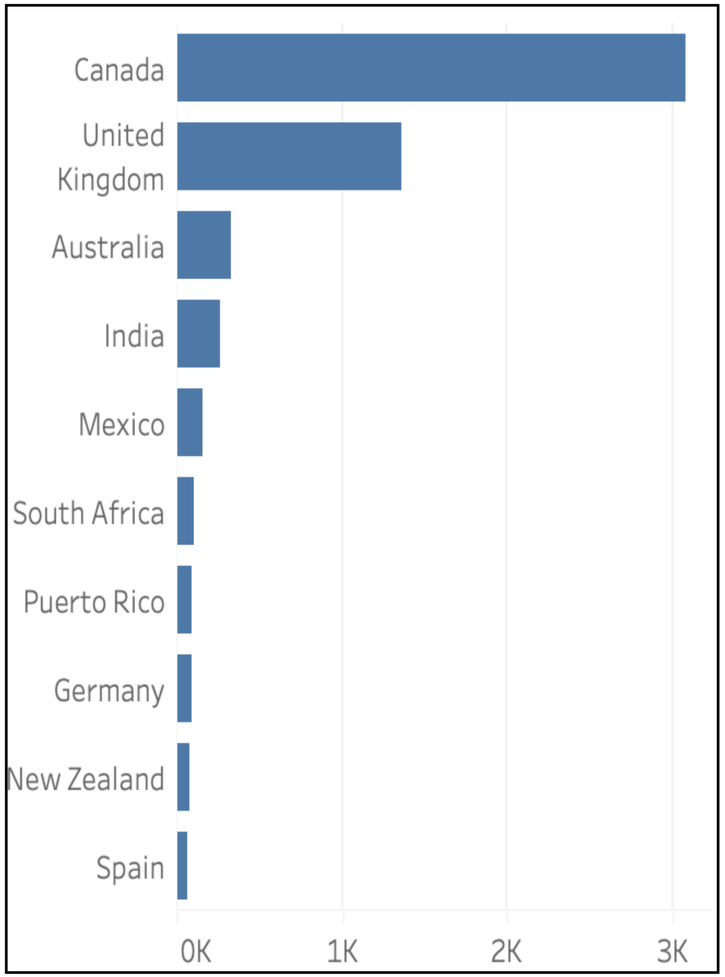
## Cluster of Sightings by Population



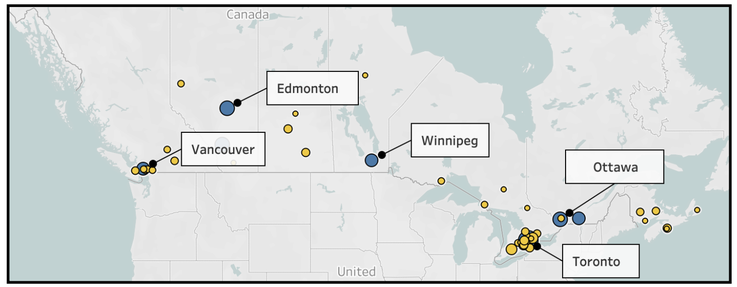

Color Grading shows population by county (yellow to blue: low to high)
This map shows that the sightings are clustered in highly populated areas.
It means that larger UFO reporting is driven by more people who are there to report, and not because aliens are getting bolder. A large amount is quite possibly just false positives.

# Geographical Analysis: Rest of the World (1/2)

### Top 10 Countries outside of USA

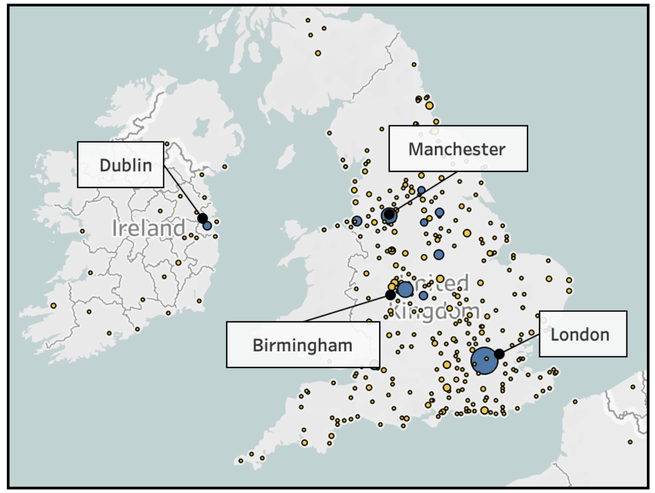



### Canada



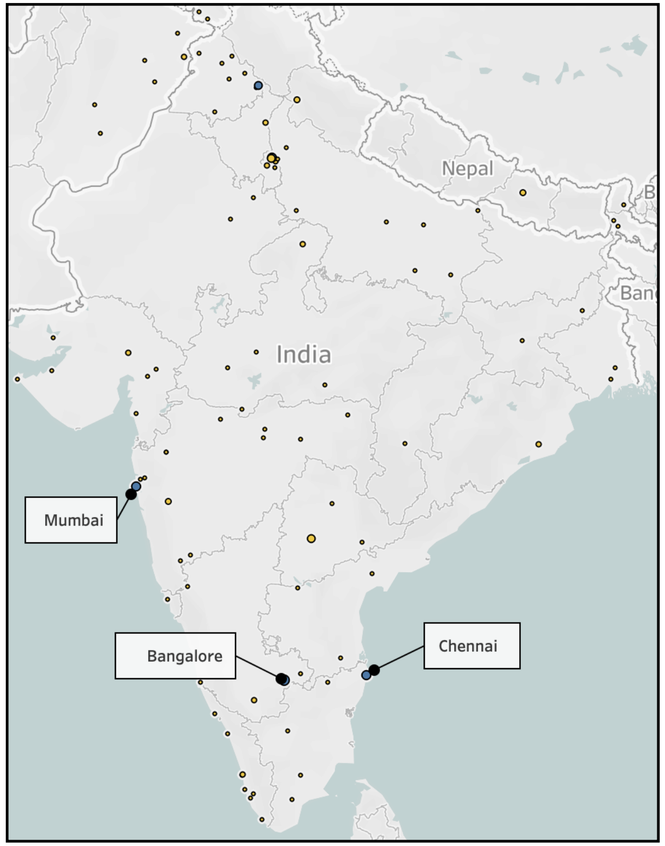

Canada's sightings are also driven by large metropolitan cities, similar to US

### United Kingdom & Ireland




UK and Ireland is also driven by large urban centers and coastal areas, but **there are a lot of inland sightings as well**, much more so than other regions.

### India




India has predominantly coastal areas with high sightings such as Mumbai and Chennai, and then Bangalore, which is home to Indian Space Research Organization (ISRO)

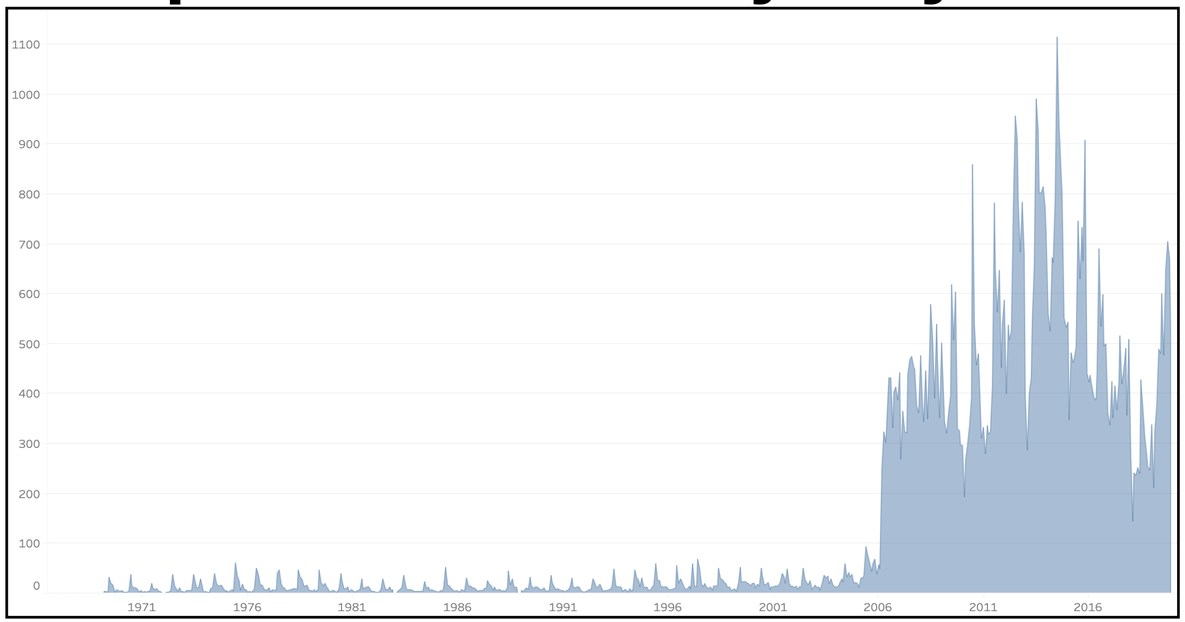
# Geographical Analysis: Rest of the World (2/2)

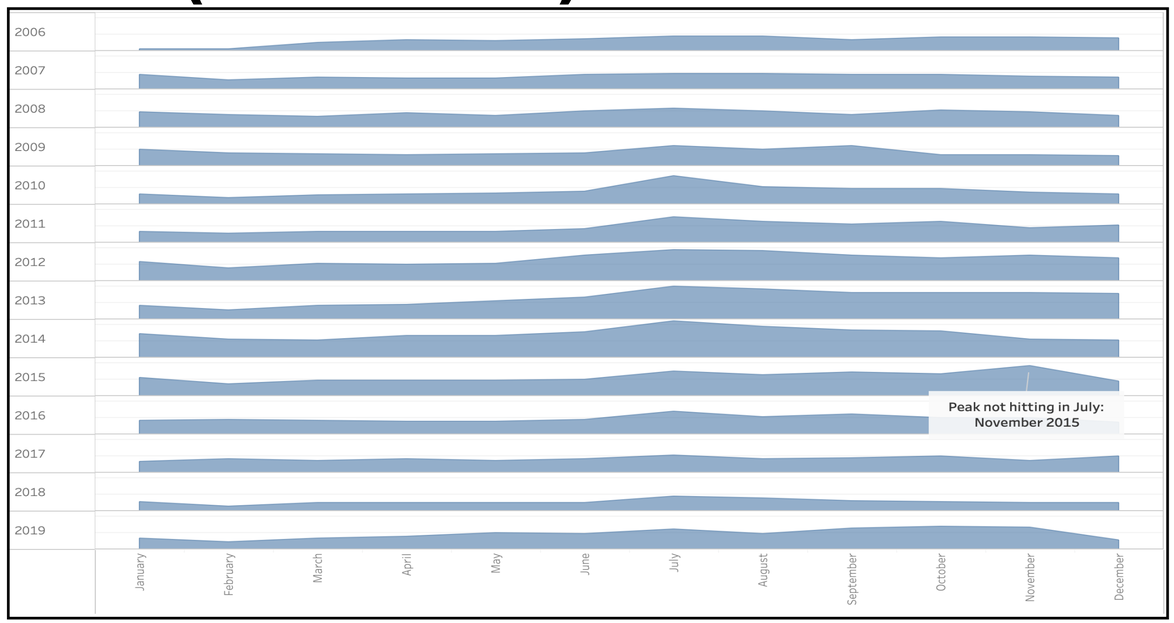
## Australia & New Zealand


**Sightings are concentrated in large metropolitan areas, plus coastal areas. This is in line with what we see across other geographies and regions.**

# Temporal Analysis: Year (1/4)

## Complete Timeline & yearly breakdown (2006-2019)



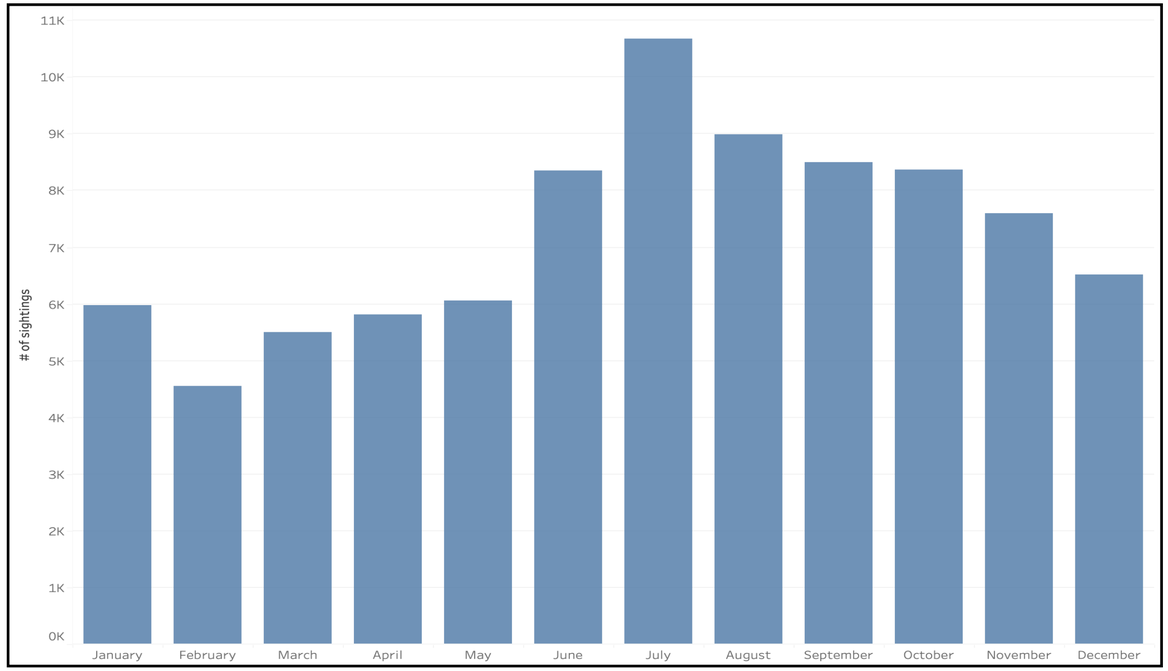




- The entire timeline of sightings for all countries is shown above:
  - Data range is from **1969** to **2019: NUFORC** was established in 1974
  - Noticeable jump in reporting since **2006**
  - Peak reporting was in **2014** with ~8700 sightings

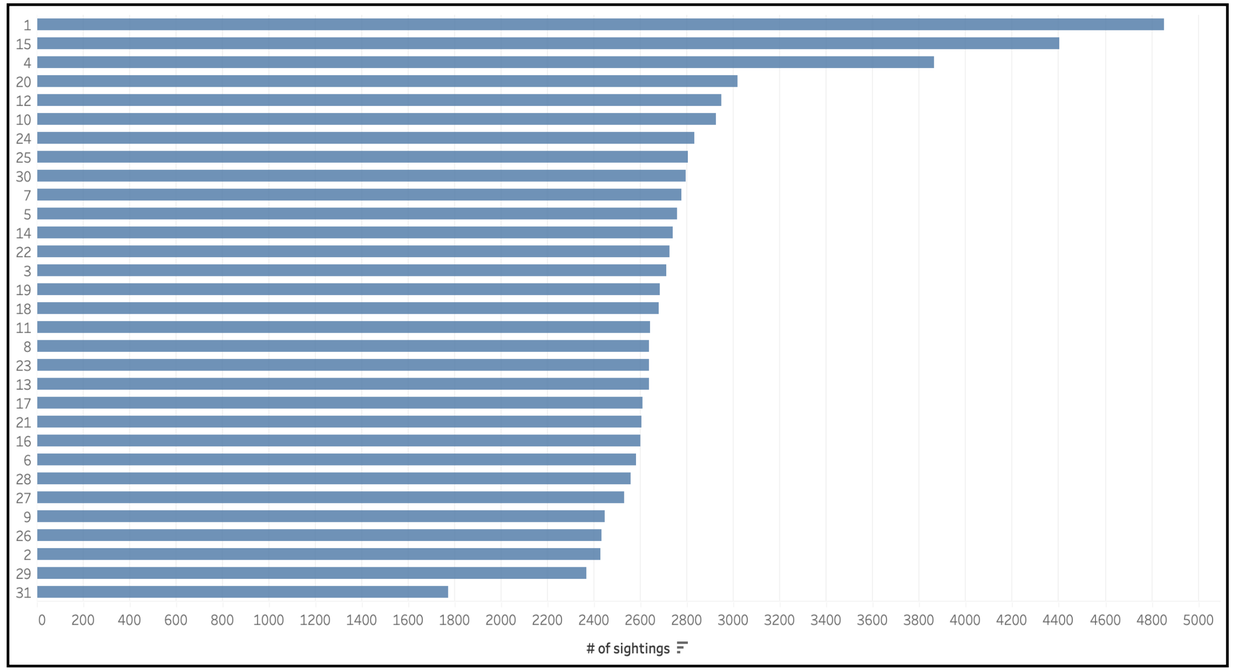
- Condensed timeline **(2006-2019)**
  - **All peaks occur in July** which corresponds to 4th of July celebrations: fireworks and drinking, prompting higher probability of reports.
  - Exception is **Nov 2015** when there was a sighting in LA prompting an unprecedented number of calls (more details to follow on that incident in next section: standout stories).

# Temporal Analysis: Month (2/4)

## Monthly Patterns (including United States)




**Sightings by Day of Month**

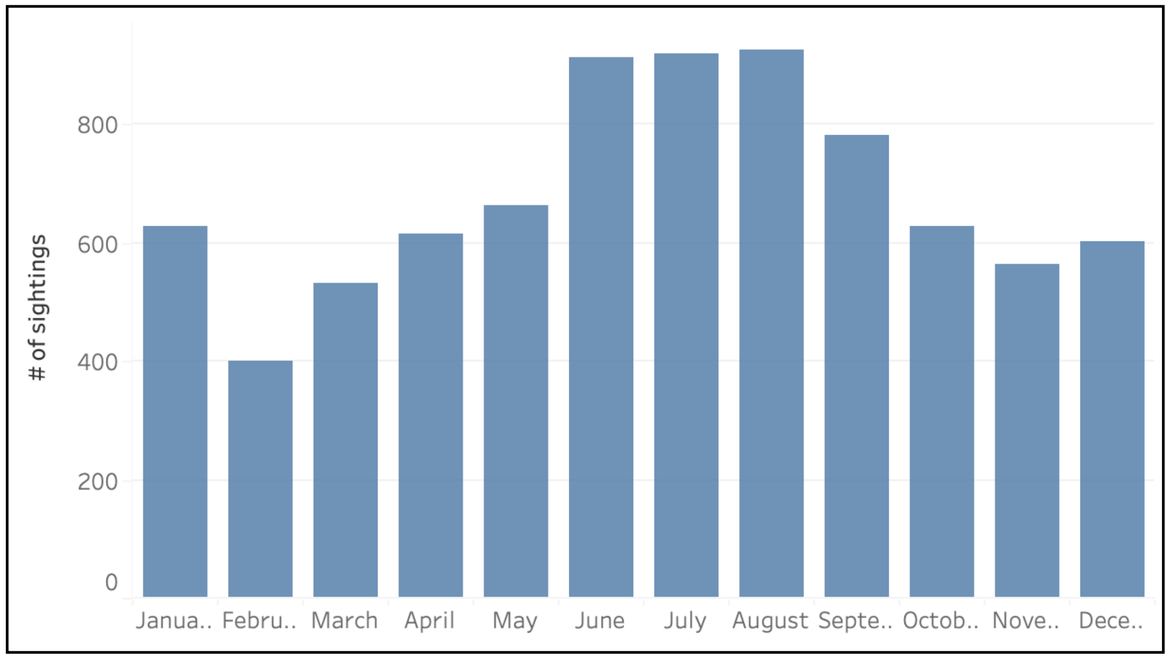



- **July** is by far the busiest month when it comes to UFO sightings:
  - Most likely due to 4th of July celebrations where fireworks are abundant
- **Summer** and **fall** months (Northern Hemisphere patterns) have a high proportion of sightings, followed by January (most likely driven by New Year celebrations similar to 4th of July)

- Most common days of month for sighting are:
  - **1st**: Possibly driven by New Year's reporting
  - **15th**: Middle of the month, possibly that is when reports are compiled in the database and that is driving up the numbers
  - **4th**: Driven by 4th of July

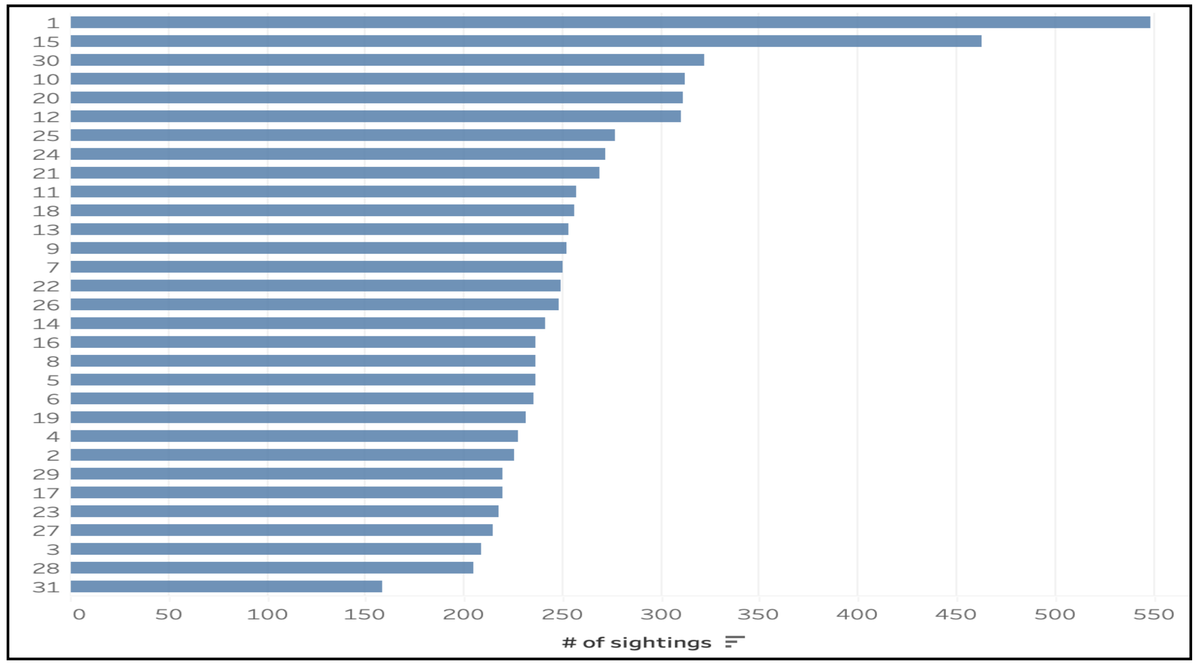
# Temporal Analysis: Month (3/4)

## Monthly Patterns (excluding United States)




- **Excluding USA,** July loses its dominance as being the month with highest reported sightings
- **Summer months** dominate, possibly because more people spend time outside, and clearer skies provide more opportunities to witness and report UFO sightings

**Sightings by Day of Month**

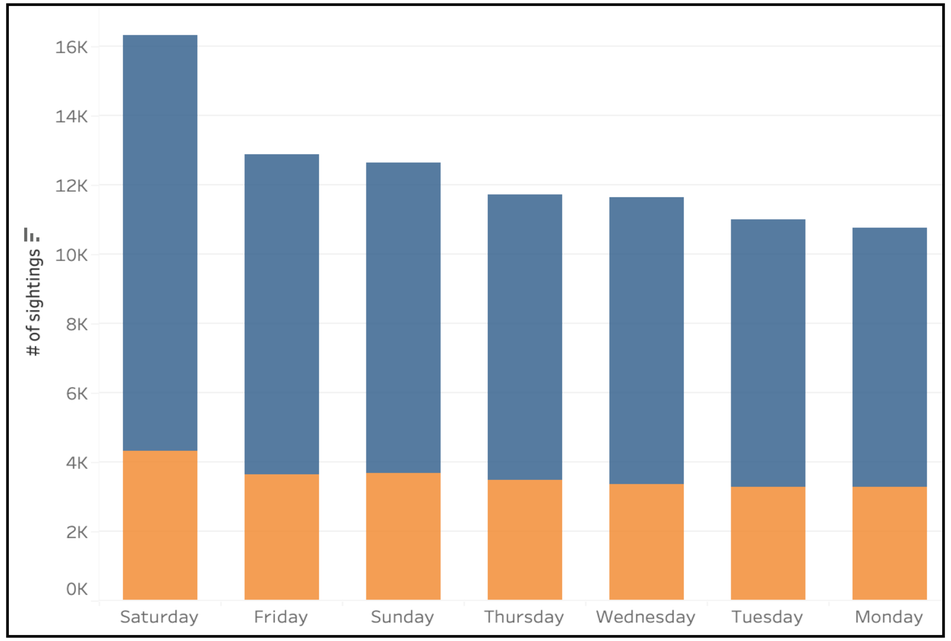



- Excluding USA, **4th** goes all the way down in terms of frequency of reporting, **as 4th of July is not a date of importance outside USA**
- Dominant date is **1st**, which indicates increased amounts of reporting due to **New Year's celebrations**

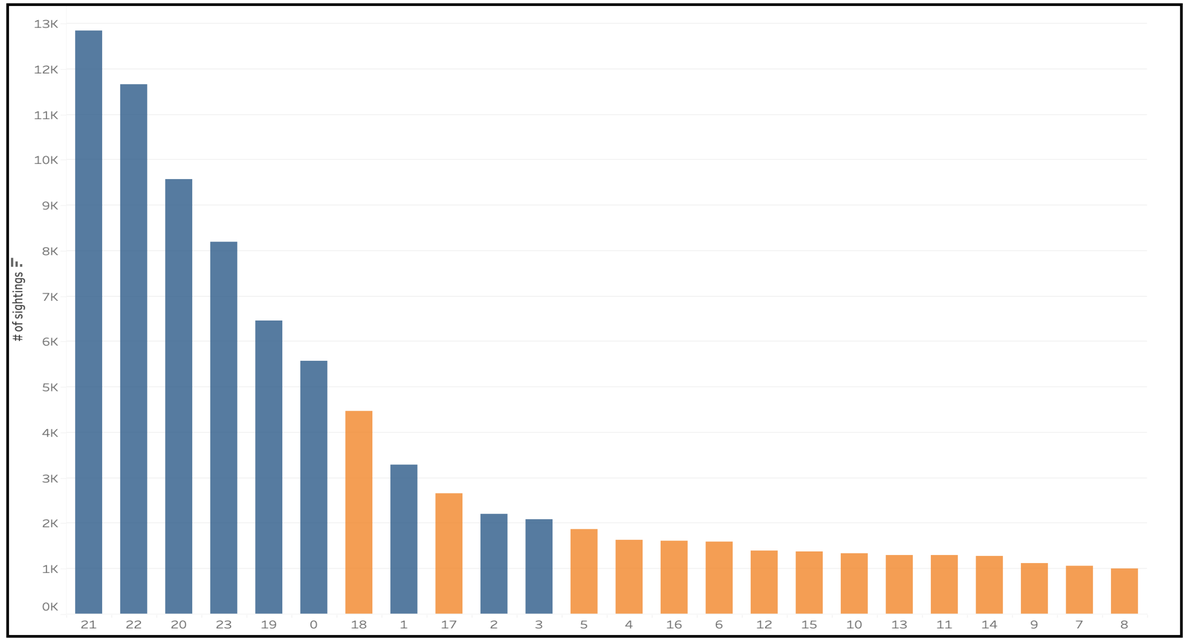
# Temporal Analysis: Day (4/4)

## Time of day and day of week pattern

**Sightings by Day of Week**




**Sightings by Hour of Day**

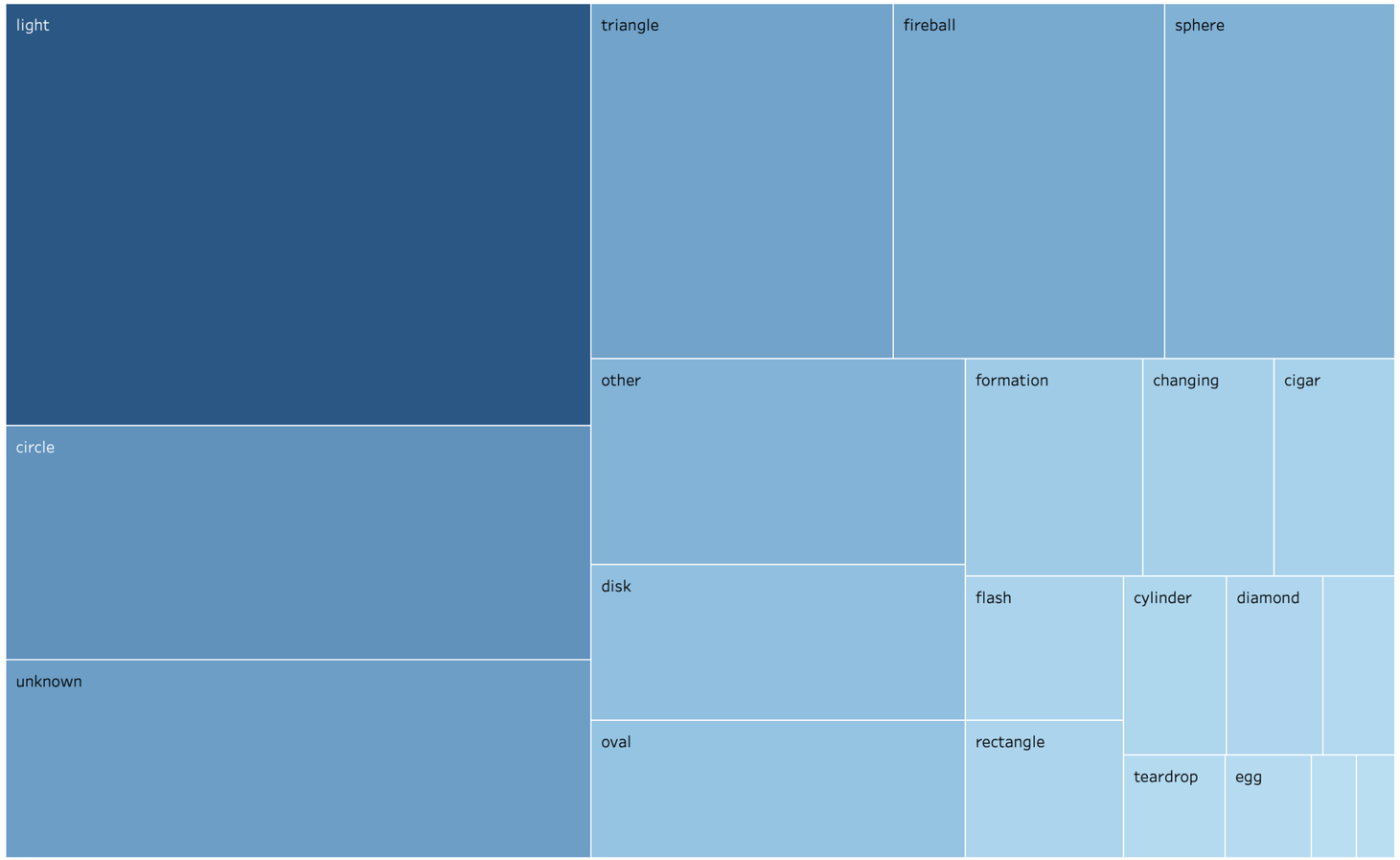



- The **blue** color indicates dark hours (**from 19:00 until 03:00**), and **orange** indicates typical daylight hours.
- The three highest days for sighting are on the weekend, with **Saturday** being the highest by far and **Monday** being the lowest
- The hours of the day with most sightings are mostly night time and dark hours:
  - Driven by increased visibility of bright objects during such hours, and confusing cosmic nighttime items with UFOs

# Shape Analysis (1/2)

## Most Common Shape Reported Overall

**Distribution of Shapes across all geographies (darker color means higher count)**

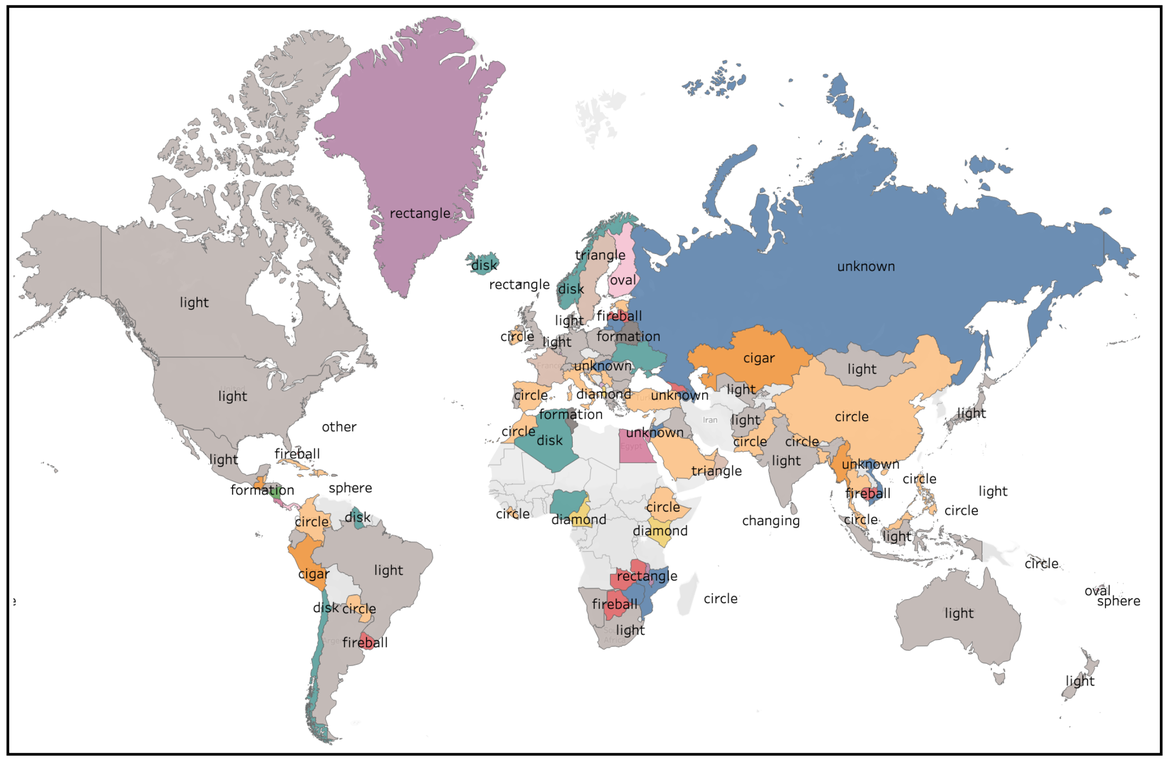



- **Light** is the most commonly reported UFO sighting
  - Possibly driven by flashes in the sky, space debris, shooting star, etc.
  - **Given the prevalence of 4th of July UFO sightings, fireworks might be getting misconstrued as UFOs.**
- The traditional **disk** shape that is associated with UFOs, popularized in pop culture in the 1950s, does not dominate the top 5 rankings, possibly because it is now considered quaint as opposed to actual representation of a UFO.
  - Also possible that **circle, oval, sphere** and **cylinder** are being used as alternates for **disk** shape and hence the sightings are distributed across these categories.

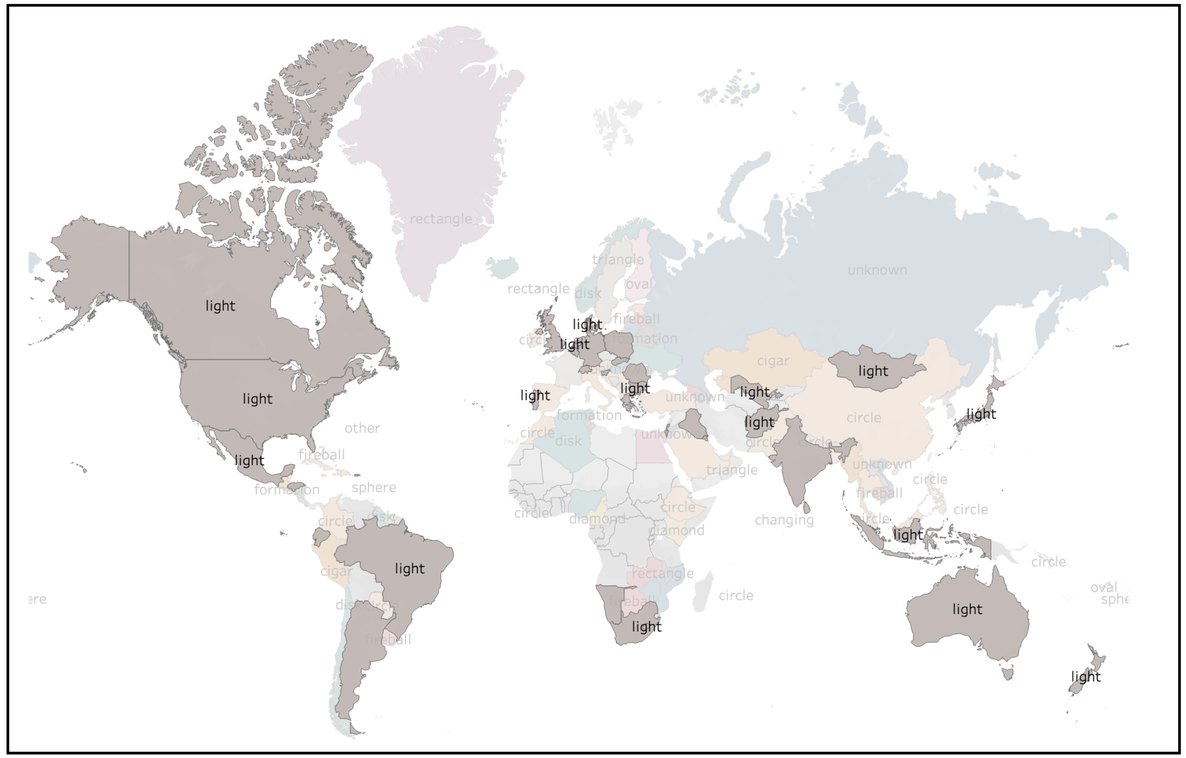
# Shape Analysis (2/2)

## Most Common Shape Reported By Country

**Most common shape in every country**



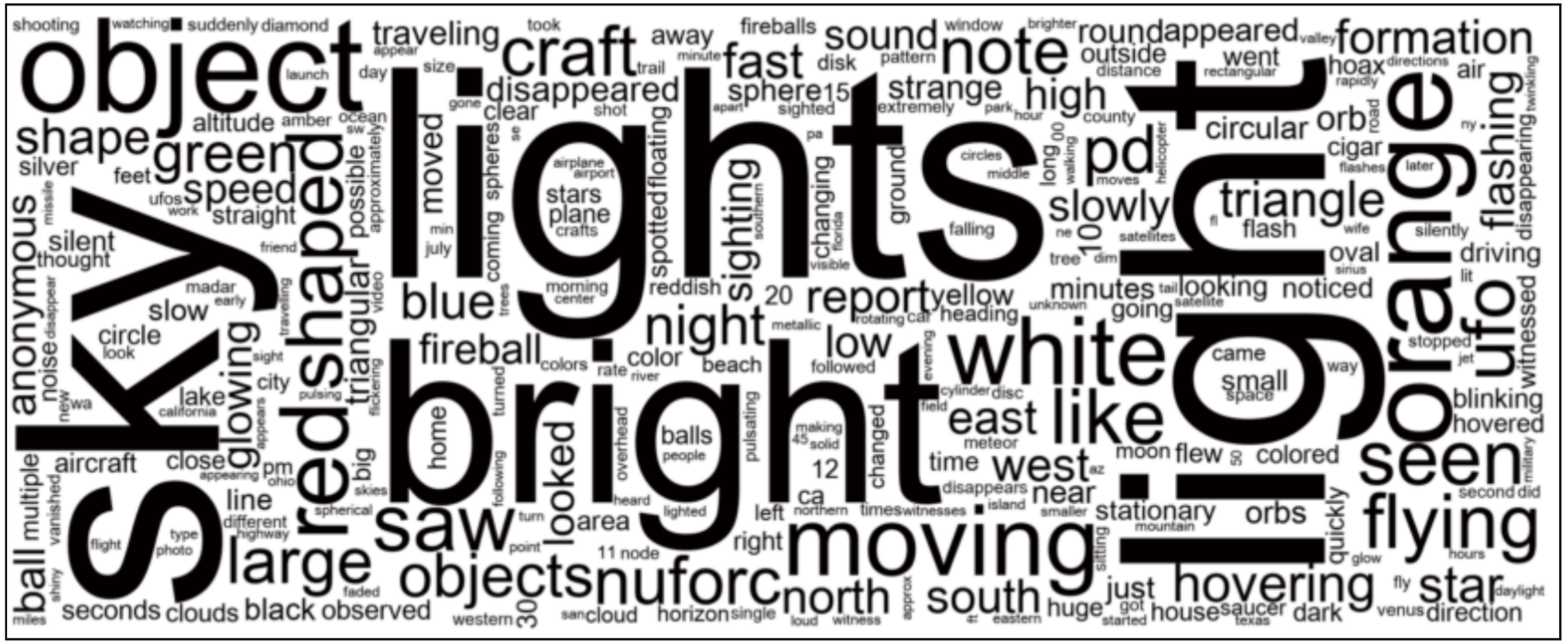

**“Light” category highlighted**




- **Light** is the most common sighting reported even when broken down by geography
- Backs the hypothesis: it is easy to conflate or confuse flashes in the sky to something unknown and by the looks of data, that instinct is universal

# Word Cloud

## Top 300 words extracted from “summary” field




**From the word cloud, it is clear that “bright lights” is a very popular term associated with UFO sightings.**
**There are a lot of mentions of “objects”, “shape”, “sky”, “glowing” and “flashing”.**
**There are some words related to movement as well: “hovering”, “moving”, “flying”, “fast” and “slowly”.**

# Synthesizing Insights

## INSIGHTS

- Sightings concentrated in **large, urban metropolitan cities** with high population.
- Sightings demonstrably go upward during events such as **New Year's** and **4th of July**.
- **Summer months** see high sightings on average as more people are expected to be outside during the time.
- Most sightings occur at **nighttime hours, peaking around 9 pm**.
- **Flashing, bright lights** are most commonly reported as UFOs followed by a **circular** shape.
- The sightings typically last **less than 5 minutes**.
- Most of the reports in NUFORC are from **United States**, as NUFORC is a US organization.

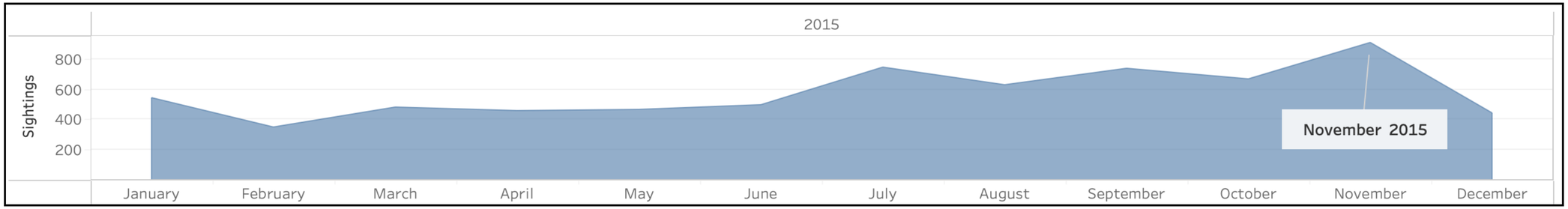
## ACTIONS

- **Emergency service** hotline officials (like 911) in large metropolitan areas should be on the lookout for **high proportion of calls on Saturday nights, especially around 4th of July and New Year's.**
    - The increased volume of UFO related calls could lead to **legitimate calls not getting requisite attention or having increased wait times** and a possible uptick in crime as a consequence.
    - To counter that, plan for increased capacity during those times, especially **after dark hours**.
- National security advisors should be on the lookout for **increased activity across coastal regions** for any nefarious party trying to capitulate on the confusion and chaos caused by increased UFO reporting.
- **Local officials** should clarify to the public that **increased firework activities during New Year's and 4th of July** causing high reporting of UFOs is **expected and nothing to panic about** regarding alien life visiting earth during its festivities.
- **Students** and **interested individuals** in UFOs should be informed that bright, flashing lights are most easily confused as UFOs, and to **not fall into that trap of false positives**.

# Standout Stories

# November 7, 2015

## The year that broke the “July” streak




- **November 2015** was the month with highest reporting of sightings for 2015, beating July’s reign.
- Since 2006, **July** had been the most populous month for reported sightings.
- **Explanation:** November 2015 was an anomaly, driven by a sighting in West coast of US that quickly started trending on social media.
- As reported by **earthsky.org**:

  ***Witnesses from Southern California to Reno, Nevada reportedly “went into a tizzy” when they spotted a bright, blue-green, mysterious, long-lasting light in the sky on Saturday evening (November 7, 2015).***

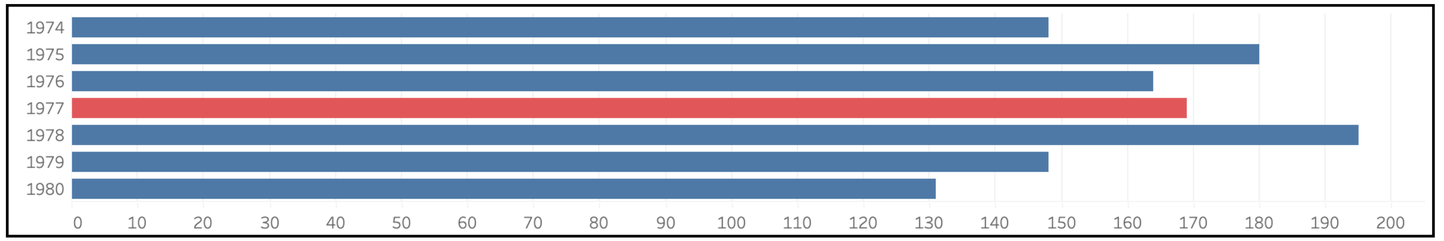

**Fig 1: An image capturing the strange phenomenon that was witnessed on West Coast USA on Nov 7, 2015**

# Aliens & the Movies

## Correlating famous UFO/alien centric movies against UFO reporting before and after the release year

**Close Encounters of the Third Kind**

**Released: November 16, 1977**

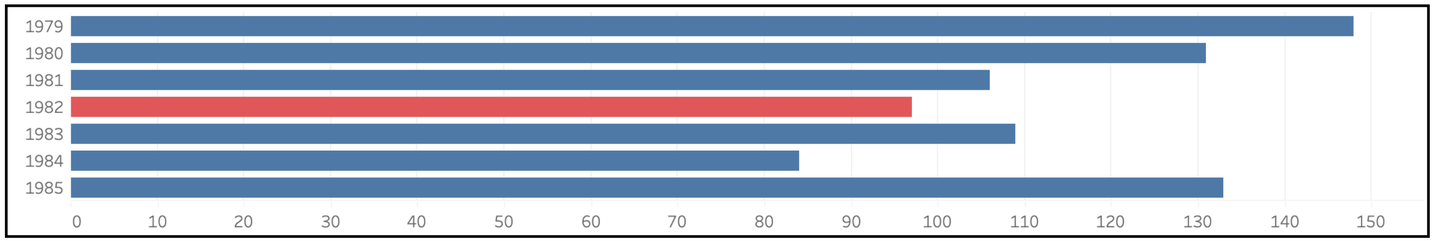




There is an immediate increase following the year of release (1978 versus 1977), but the effect wore off couple of years after that.

**E.T.: The Extra Terrestrial**

**Released: June 11, 1982**





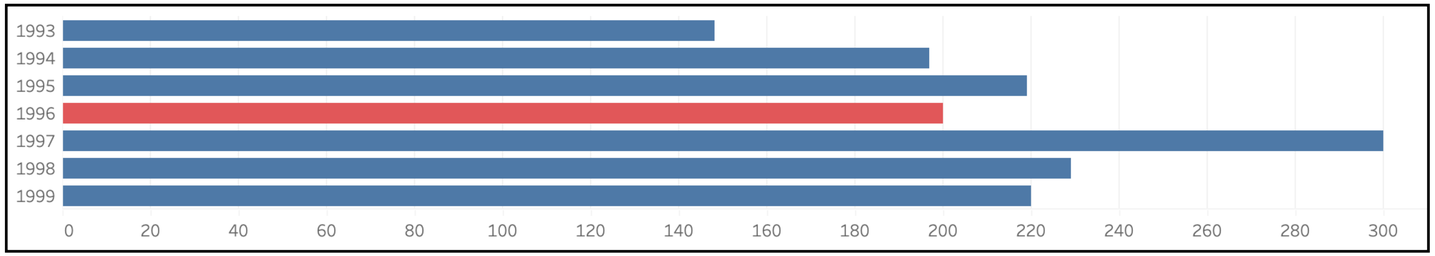
Increase in reporting (1983 vs. 1982) observed after the release, but the effect tapered off in 1984, but revived in 1985 possibly by the release of “Starman” and “2010: The Year We Make Contact” in 1984.

**Independence Day**

**Released: July 3, 1996**

**Contact**

**Released: July 11, 1997**



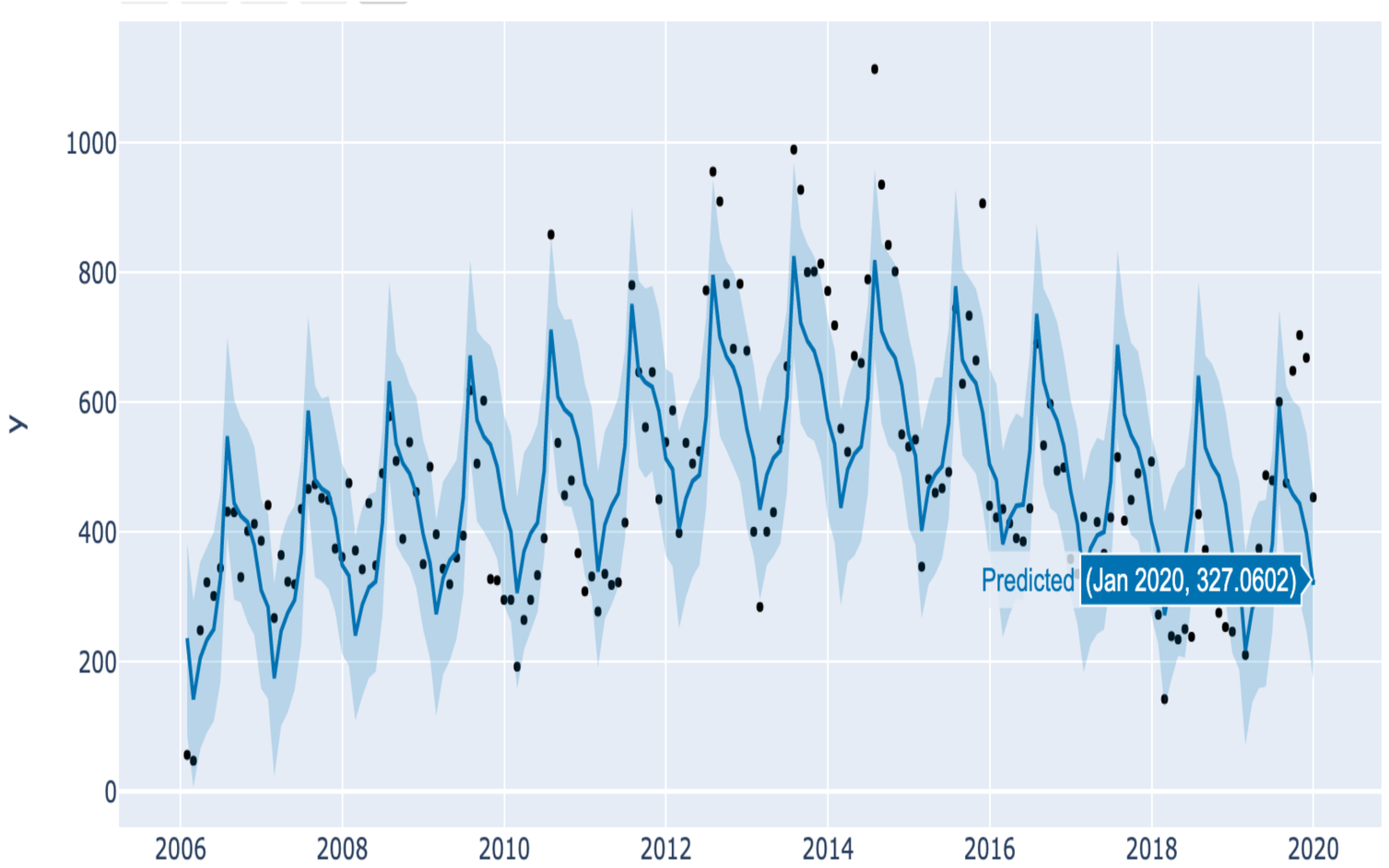



Independence Day (1997 vs. 1996) had a huge effect on the reported sightings, but Contact (1998 vs. 1997) failed to sustain that excitement. It was still higher as compared to early 90s, but lower than the bump from Independence Day.

# January 2020 Forecast

# January 2020 Forecast

## 327 sightings forecasted for January 2020

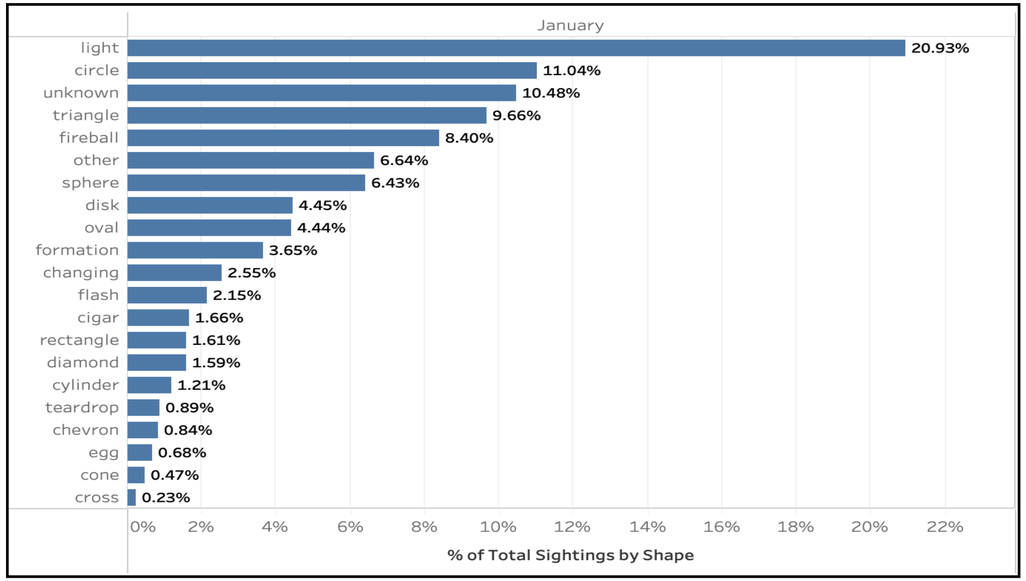



- **The prediction for January 2020 is: 327 sightings**
- Time period used is from **2006** to **2019** in order to reduce the noise from earlier years when reporting was very low
- Time series forecasting done using **Facebook's Prophet** based on historic monthly volume of sightings that uses **decomposable time series model**.
- Methodology:
  - Aggregated sightings into months (Jan-2006, Feb-2006, etc).
  - Volume was based on number of rows.
  - Built a **time series model** with **monthly seasonality.**
  - Turns prediction for volume of sightings into a time series forecasting problem where historic volumetric trends are used to predict future value.

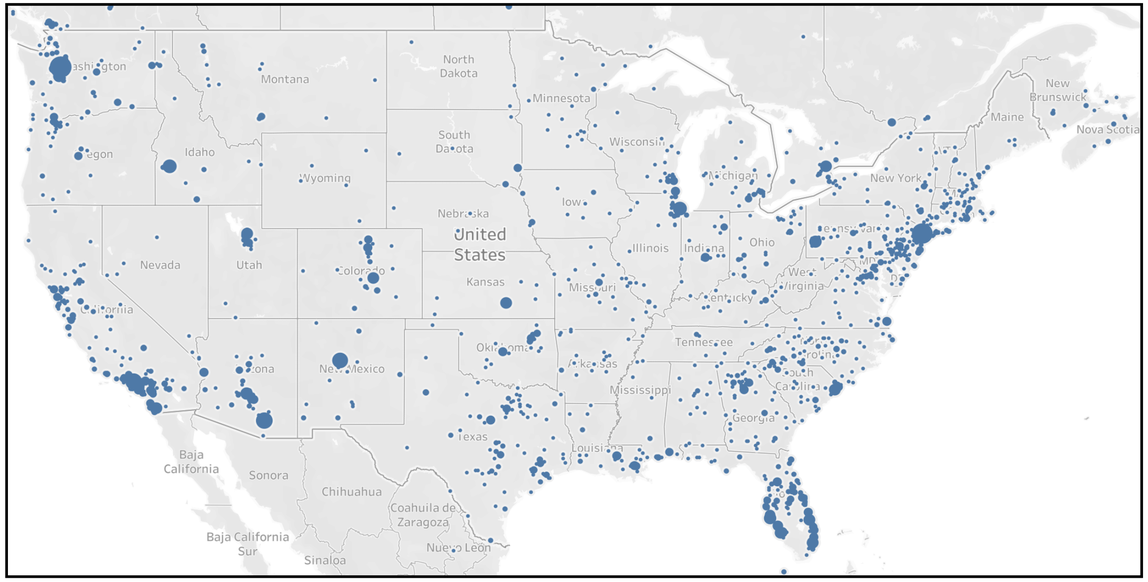
# January 2020 Forecast

## Shape and Geographic Distribution

### Proportion of Reported Shapes across all January months




### Geographic distribution of sightings in all January months




**<u>"Light" shapes in Jan 2020: "68" sightings (20.93% of 327)</u>**

The proportion of "light" shape in all January months since 2006 overwhelmingly outnumbers other shapes (21%).
The second highest is circle, followed by unknown/not-reported shape.

**The distribution of sightings across all Januarys suggest that we except coastal areas, and large urban centers to carry the bulk of the reporting.** This follows from the macro trend we established throughout the timeline.

# Tech Stack Used

# Tools Used for Analysis

**Data Cleaning & Preprocessing**

- Python
  - Pandas
  - Numpy
  - Matplotlib
  - Seaborn
  - missingno
  - pycountry
- Jupyter Lab

**Visualization**

- Tableau
- Matplotlib
- Seaborn
- Word cloud
  - sklearn (tf-idf)
  - wordcloud

**Forecast Model**

- Facebook's Prophet Library
  - Monthly/ seasonal forecast model