**UFO Sightings: Predicting the Unknown**

Anastasia Volokhova

IFI8410: Programming for Business

Instructor: John Sundeep Teli

April 30, 2025

**Introduction**

UFO, an unidentified flying object, refers to a suspected alien spacecraft. UFO sightings have been reported starting in the early 1900s in various parts of the world, raising questions about life on other planets. One of the first well-known UFO sightings occurred in 1947 near Mount Rainier in Washington when businessman Kenneth Arnold claimed to see a group of objects "in a diagonally stepped-down, [echelon formation](https://en.wikipedia.org/wiki/Echelon_formation), stretched out over a distance”.

In our analysis, we will see that UFO sightings have been reported much earlier at the beginning of the 20th century, progressively increasing with time.

The objective of this project is to explore the dynamics of UFO reports through visualizations, statistical models, and predictive analysis to identify hotspots, seasonal trends, the most frequently used descriptions of UFOs, and to predict the possibility of seeing UFO at a certain time and place.

**Dataset**

The dataset named “UFO Sightings” was sourced from Kaggle. It contains 88,875 reports of UFO sightings over the last century predominantly in North America.

Reference: <https://www.kaggle.com/datasets/NUFORC/ufo-sightings/data>

**Why is this Dataset suitable?**

The dataset is well suited for our analysis due to its comprehensive and diverse structure. It includes rich historical coverage over the century of observations, particularly between 1905 and 2016, that allow us to analyze long-term trends, and timestamps to support analysis of patterns by seasons, months, day/night, and hours. The dataset contains detailed location data enabling hotspot mapping and proximity to certain landmarks, military bases, etc. Witness descriptions and diverse attributes such as UFO shapes and duration allow the use of natural language processing techniques such as shape prediction and theme clustering.

**Dataset Description**

There are 88,875 rows and eleven columns in this dataset. Each record date and time, location, shape, duration, comments, latitude, and longitude.

|  |  |  |
| --- | --- | --- |
| **ID** | **Record Information** | **Data Type** |
| *datetime* | date(dd/mm/yyyy) and time of sighting | numerical |
| *city* | city of sighting | string |
| *state* | state of sighting (if applicable) | string |
| *country* | country of sighting | string |
| *shape* | shape of UFO | string |
| *duration (seconds)* | duration of sighting in seconds | numerical |
| *duration (hours/min)* | duration of sighting in minutes/ hours | numerical, string |
| *comments* | witnesses’ comments and UFO descriptions | strings, numerical |
| *date posted* | date/place the information was posted | strings, numerical |
| *latitude* | latitude of a UFO sighting location | numerical |
| *longitude* | longitude of a UFO sighting location | numerical |

**Methods and Tools**

For this project, I have utilized Python programming language. The dataset was loaded into Google Collab. Pandas tools were used for data cleaning and grouping.

***Time Series Analysis***

I have started with a time-series plot to illustrate the annual number of reported UFO sighting cases between 1905 to 2016.

A graph with numbers and lines

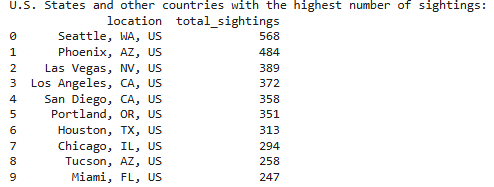
AI-generated content may be incorrect.

The analysis reveals a slow and steady increase in sightings throughout the early 20th century, with a noticeable rise beginning around the late 1940s coinciding with Kenneth Arnold's sighting report and the post-World War II era. Sighting surged in the 1960s with the highest peaks between the 2000s and 2010s.

These trends may reflect a combination of factors:

* Technological revolution: smartphones, cameras, and the Internet made it easier to record and report UFO sightings.
* Cultural factors such as Sci-fi movies and alien conspiracy theories: The X-Files series (1993-2002) turned into prime-time entertainment where agents investigated alien abductions and government cover-ups; such alien conspiracy theories as Area 51 – military base where allegedly U.S. hides aliens; The Roswell Incident (1947) when something crushed on the ranch in New Mexico and U.S. military claimed different versions only sparking the suspicion that government was hiding the truth;
* The more recent years have better record-keeping abilities.

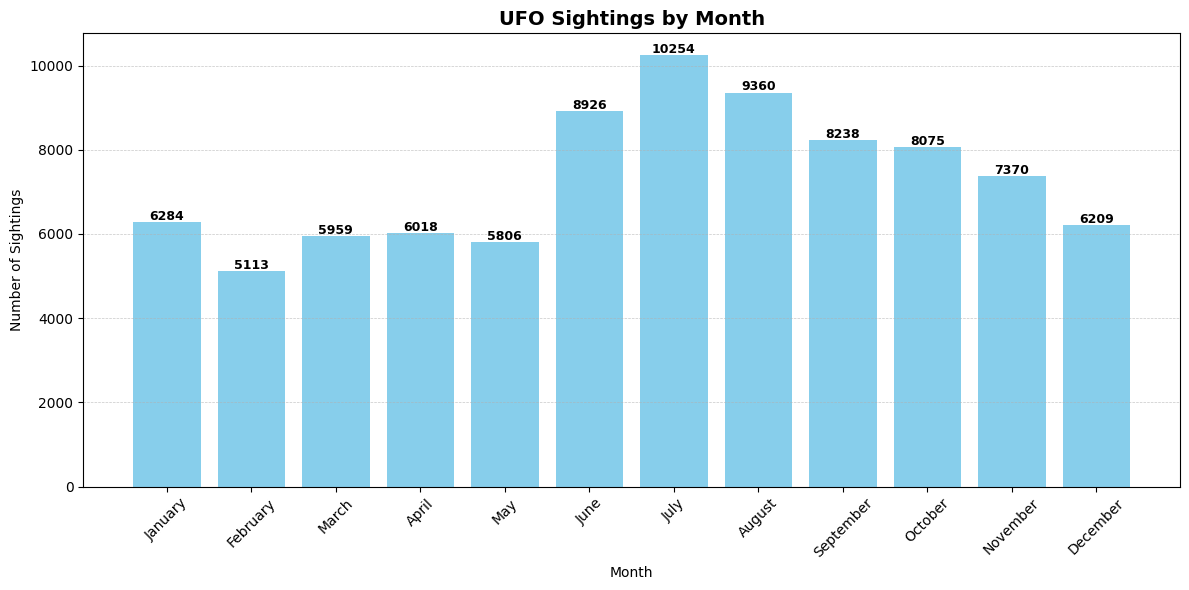
The next step was to identify the locations with the highest number of sightings across the Northern American Region.

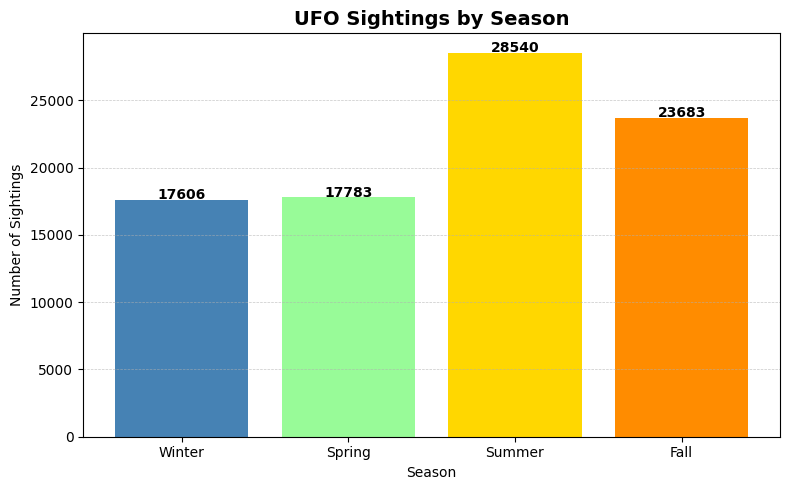


Most reports originate from the United States, particularly in Washington, California, Nevada, Arizona, Oregon, Texas, Illinois, and Florida. These results can be explained that these areas are highly populated and have communities dedicated to the aerial phenomenon. In particular, the Mutual UFO Network was founded in Illinois in 1969 to investigate UFO sightings. Additionally, such states as California, Arizona, and Nevada are known for their dry climates and high sunshine hours. Florida and Texas also have good visibility especially outside of storm seasons.

***Exploratory Analysis***

Having the basic distribution patterns, I have moved into the exploration phase which is focused on identifying trends, patterns, and relations between variables.





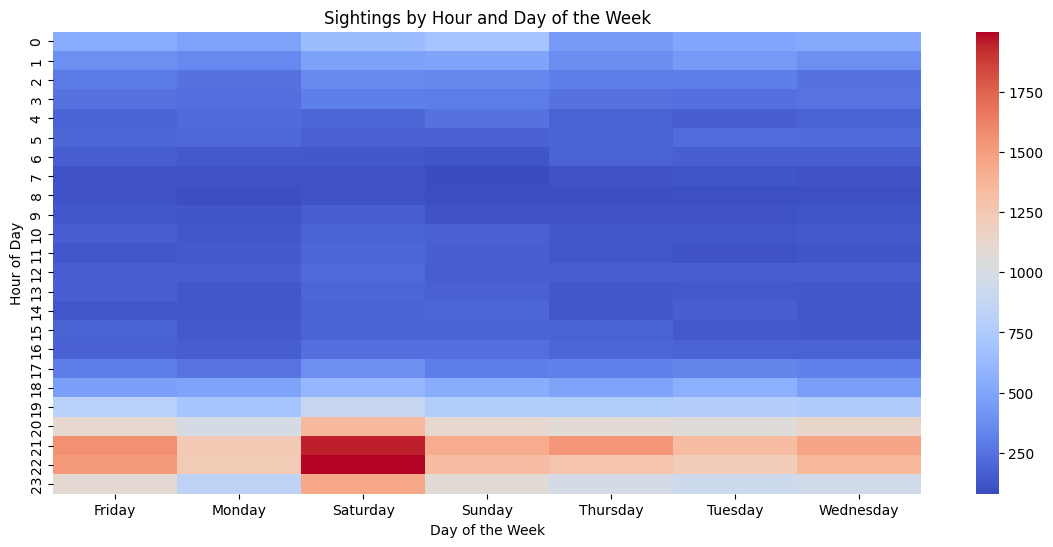
A comparison of a graph

AI-generated content may be incorrect.

The analysis of UFO sightings by seasons, months, day/night, and weekdays/weekends reveals that the patterns found may heavily rely on environmental and behavioral factors. With a sighting peak in the summer season, particularly in July and August, weekends and night hours, it could be explained by the favorable weather when more people spend time outside in the evenings doing barbecues, fireworks, camping, etc. Fireworks around the holidays like the 4th of July could be mistakenly taken as a sighting. In general, summer and weekends are associated with vacations and outdoor activities: people spend time in unfamiliar places focusing on observing things around them, which also may raise the number of UFO reports. Winter months are usually associated with colder weather and unfavorable conditions that keep people indoors, limiting visibility and reducing the chances of seeing anything.

***Correlation Analysis***

In the correlation heatmap, Friday, and Saturday demonstrate the highest correlation with evening hours, between 8:00 pm to 11:00 pm, meaning that the likelihood of spotting UFOs in the sky is the highest during these days and times. Other weekdays demonstrate a weaker correlation with evening hours; however, the trend is still observed. This means, with all other things being equal, the best time to observe UFOs is during the evening hours.

******

***Text Mining***

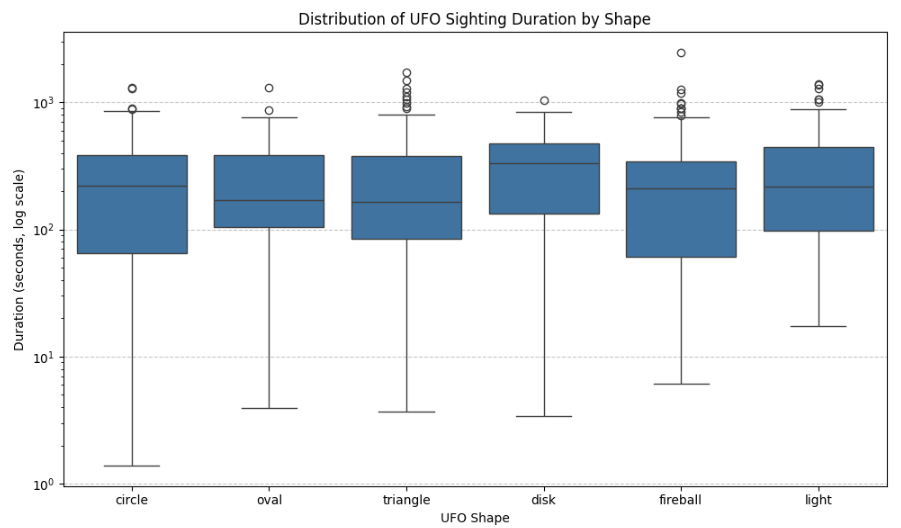
Going further in the exploratory analysis, Natural Language Processing tools were used to identify recurring themes and shapes that appeared throughout the reports.

Among the various shapes reported, light, triangles, and circles appear significantly more often than others. These findings may reflect the association with cliché UFO images promoted in the media and culture. Such categories as “unknown” and “other,” if put together, take the 2d place in the rating. However, “Unknown” typically means that observers could not identify the object, while “other” means that appearances do not fit previous descriptions. It could be due to unusual forms or vague nature.

A graph with numbers and a green background

AI-generated content may be incorrect.

Additionally, the possible relationship between UFO sighting duration and their reported shapes was checked. The y-axis was log-scaled due to a wide range of duration values and extreme outliers.



All shape categories are strongly right-skew with short-duration values, but a few extend into longer periods (minutes and more). While most shapes have similar median durations, triangles show more variability and longer upper whiskers assuming they were often reported with longer durations compared to others. Outliers appear across all UFO shapes, indicating substantial variability in sighting durations. This suggests that, regardless of shape, some sightings can last far longer than average.

To assess whether sighting durations differ by UFO shape, a **Kruskal-Wallis H-test** was conducted. The test yielded a p-value of 0.3191. This p-value exceeds the conventional alpha level of 0.05, indicating that the observed differences in sighting durations across shapes are not statistically significant. The shapes do not appear to have an impact on the duration of sighting. Duration variability could be influenced by other factors such as geographical location, environmental conditions, or subjective perception by witnesses.

***Words Frequency Analysis and Bigram Analysis***

A close up of words

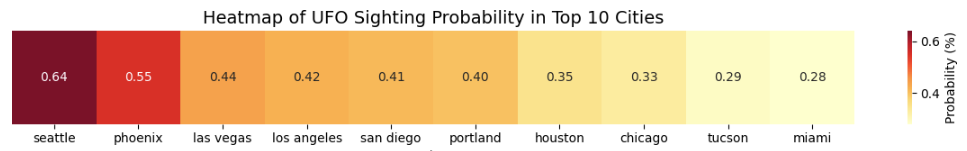
AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

The word cloud visually highlighted dominant terms. Larger and bolder words indicate more frequent usage, affirming the patterns found in our quantitative analysis. Many sightings involve light-based phenomena observed in the night sky that align with classic UFO tropes.

The bigrams reinforced the visual and dynamic nature of sightings, often associated with unusual movements, brightness, and mysterious appearances.

***Probability-Based Modeling***

Building on all the findings, the likelihood of UFO sightings by location was computed.

***Naïve Bayes Classifier***

To move beyond historical patterns and descriptive statistics, a Naïve Bayes probabilistic approach was implemented to estimate the likelihood of UFO sightings occurring under specific conditions: the probability of a sighting occurring in Georgia, on a Wednesday, and at night. The likelihood of a UFO sighting occurring in Georgia is approximately 1.94%, while the probability of such sightings occurring on a Wednesday is around 13.78%. The chances of a UFO sighting taking place at night are quite high, at about 67.08%. When combining these factors, the relative probability of a UFO sighting occurring in Georgia on a Wednesday night is approximately 0.18%. This suggests that, while UFO sightings in Georgia on Wednesday nights are not extremely common, they are still a measurable event within the dataset.

***Geospatial analysis***

There have been long-standing associations between UFO sightings and Area 51 that are often linked to conspiracy theories and extraterrestrial activity. I wanted to explore whether there was any notable geographic clustering of UFO reports near this infamous location. The distance between each sighting and Area 51’s coordinates were calculated. The closest sightings were visualized on an interactive map using the Folium library. Each sighting is represented by a circle color-coded marker. The analysis revealed that there is a noticeable cluster of UFO sightings reported within a few hundred kilometers of Area 51

The QR-code below provides direct access to the interactive map:

A qr code with black squares

AI-generated content may be incorrect.

**Limitations**

Some limitations may affect the reliability and interpretability of the analysis results. Subjectivity is one of the major factors as witness reports are often influenced by personal beliefs and emotions leading to misidentifications of ordinary objects or phenomena. Shapes classification as “unknown” and “other” highlight the limitations when dealing with real-world observations. It emphasizes the necessity of flexible data collection methods to account for a wide range of human perceptions. The lack of visual confirmation makes it harder to confirm the claims outside of the witness’s testimony. Technological limitations such as low-quality recordings hinder the collection of evidence.

Conditions, locations, and timing of sightings persist to be inconsistent making it hard to establish patterns. The absence of standardized reportingcriteria makes it even more complicated, as different organizations may collect disparate data types, resulting in unclear or incomplete information. Media plays a significant role in shaping our memories and co-beliefs creating cultural narratives that influence how people perceive and interpret UFO sightings.

**Conclusion**

In this project, a comprehensive dataset of UFO sightings spanning over a century was explored. Analysis was focused on determining temporal, spatial, and descriptive patterns. Sightings steadily increased after the 1940s, with geographic distribution hotspots in such states as Washington, Arizona, Illinois, California, Nevada, and Florida.

Patterns of higher sighting rates were detected during the summer months, weekends, and evening hours, potentially influenced by weather, social activities, and environmental visibility. The NLP techniques highlighted recurring themes and shapes, reinforcing the influence of media and popular culture on people’s perceptions. The probability of sightings based on specific conditions was calculated.

The set of limitations contributes to the UFO perception and reporting. Among those are subjectivity, lack of physical evidence, inconsistent reporting, collection methods, and cultural influences.

Ultimately, as the dataset evaluation revealed meaningful discoveries associated with UFO sightings, it also highlighted the need for standardized reporting and continued research. Future studies can refine predictive models and offer new perspectives on the nature of UFO sightings, helping to differentiate genuine phenomena from misperceptions, and suggesting a scientifically grounded understanding of the topic.