UFO SIGHTING REPORT INVESTIGATION

Edmund Hui, Rio Jia, Rachel Montgomery, Yuning Wu

VANDERBILT UNIVERSITY®

Data Science Institute

CONTENTS

PART 1: INTRODUCTION

Exploratory data analysis is driven by questions. So, as data scientists, we decided to ask one of the age-old questions of human civilization—are we alone in the universe?

For millennia, people have looked up to the skies and seen mysterious objects they couldn't explain. In recent years, UFOs (unidentified flying objects) have been in the news after the Pentagon declassified three videos that appear to have been UFOs recorded by the U.S. Navy. In October of this year, NASA announced a nine month study to delve into unidentified aerial phenomena, which will be entirely unclassified and within the public domain [1].

In this report, we investigated the factors that influence UFO sightings. To do so, we utilized the National UFO Reporting Center (NUFORC) database of UFO reports to summarize the associations between various reported sightings and conjecture as to whether sightings could be the result of earthly phenomena such as rainstorms, fireworks or even the release of a popular sci-fi movie.

To guide our analysis, we asked the following motivating questions:

- *What are the most common weather conditions surrounding UFO sightings?*
- *Are there any overtime trends in UFO sightings? Do they tend to be clustered/seasonal or evenly distributed?*
- *Are UFO sightings related to political affiliations? What areas of the country are most likely to report UFO sightings?*
- *Do certain cultural phenomena influence UFO sightings?*

PART 2: DATA

PART 2A: DATA COLLECTION

The National UFO Reporting Center (NUFORC) is a non-profit organization in Washington, United States that collects, records, and distributes UFO sighting reports. It is the most widely accepted national UFO reporting facility, and it partners with law enforcement agencies, military facilities, NASA, 911 dispatch centers, and national weather service offices who routinely direct all calls regarding possible UFO sightings to the NUFORC. The NUFORC releases all the data that it collects to the public [2]. In recent years, the most common way to report a UFO sighting is through its online form at https://nuforc.org/reportform/.

There are a few different input methods that the form uses to collect different types of data. For example, one type of input method is a select dropdown where users select a value from a predefined list; date, time, country, and state is collected this way. Another way that users input information is through text fields that accept any string of information; duration of the sighting, city, county, sighting summary and details, and contact information of the user is collected this way. There are also checklists where users can select all, some, or none of the options that apply; characteristics of the object such as whether the object left a trail, emitted beams, changed color, landed, made a sound, etc. and whether the sighting was a close encounter and if so whether entities were seen, missing time was experienced, marks were found on the body afterwards, etc. is collected this way. Finally, users can upload image(s) of the sighting whether that be a photograph of the object itself or an illustration or map of what happened.

On data.world, Tim Renner has a dataset that is a subset of the NUFORC data that he scraped using scrapy and merged with a city location database, MaxMind that performs geocoding to add city latitude and city longitude to each observation [3]. Rishi Damarla posted the part of this dataset that is from 1969–2019 on Kaggle, where we downloaded our original preprocessed dataset [4]. This dataset contains the following variables: summary, city, state, date_time, shape, duration, stats, report_link, text, posted, city_latitude, and city_longitude.

PART 2B: DATA CLEANING

To begin our data cleaning, we first discarded the following variables from the dataset: summary, duration, stats, posted, and report link. The reason we discarded these is mainly because they were redundant with other variables in the dataset or too messy to clean in the case of ‘duration.’ So our dataset contains the following variables: city, state, date_time, shape, text, city_latitude, and city_longitude. Because we felt that any observations that are missing values in variables related to datetime or location are unreliable, we discarded these observations. Based on our missingness analysis, these variables were all missing completely at random, so dropping them is justified (refer to figure 2). Because ‘shape’ and ‘text’ are more subjective perceptions of a UFO sighting, we decided to keep observations that were missing either or both of these fields. We furthermore created four new variables from the date_time variable: year, month, day, and hour. We did this to prepare for further analysis and to join with weather data which relies on both location data and date_time data separated into these four variables. In addition, we removed all observations prior to 2015, so our data spans the years of 2015–2019. The reason we did this is because merging with weather data is done with individual api calls for each observation and making api calls for too many rows of data is extremely time consuming. Therefore, to limit the amount of rows we were processing, we only kept the data from the last five years. Finally, during our data validation process, we discovered that some observations were located in places outside of the United States, so we dropped those observations as we were only interested in UFO sightings in the US. Our final preprocessed dataset contained 16, 811 observations.

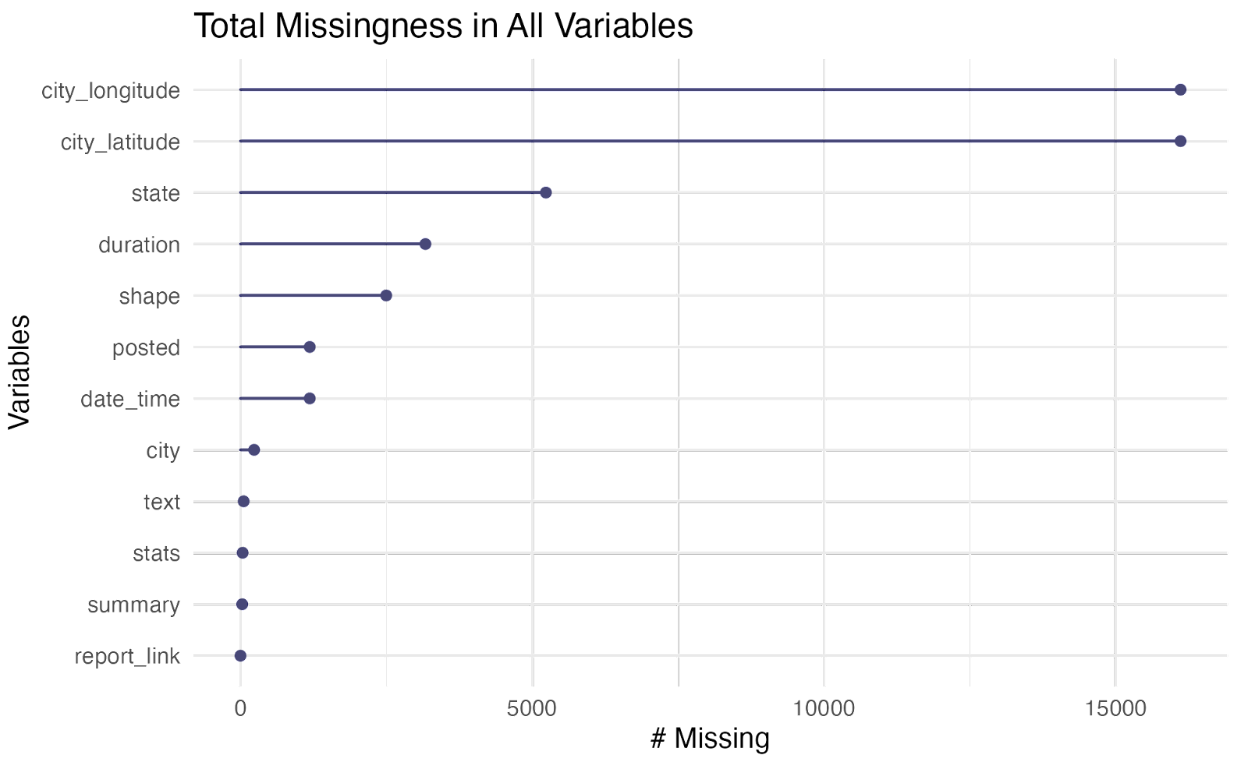


Figure 1
The y-axis shows each of the variables from the original dataset and the x-axis shows the number of observations from the original dataset that contained missing values for that particular variable. Most missingness comes from city_longitude and city_latitude columns. Approximately 20% in these 2 columns are NA values. State, duration, shape, date_time and posted have the most missing values after the first two columns, although to a much lesser extent.

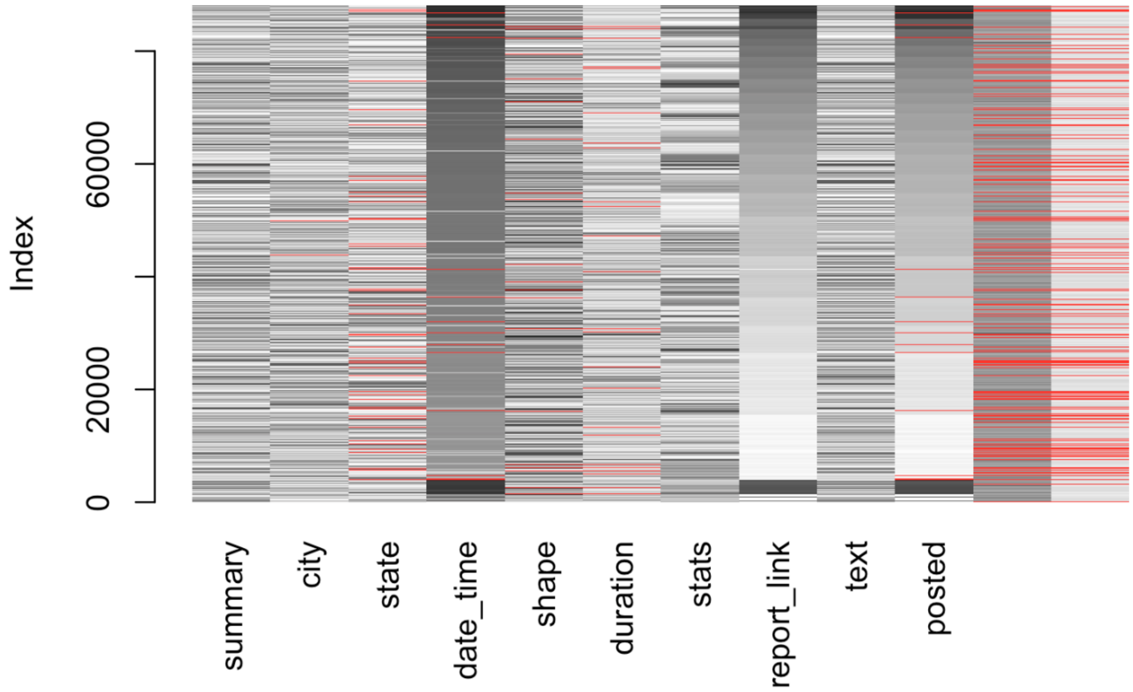


Figure 2
This is one of many matrix plots we plotted that shows that there is no pattern to missingness regardless of the variable we organize by. Therefore, the data is (Missing Completely at Random) MCAR and we are justified in dropping the NA values for each column for each respective analysis.

PART 2C: DATA PROCESSING AND JOINING

To prepare our final dataset that contains **weather data**, we used city_latitude, city_longitude and year, month, day, and hour from each row to query the meteostat api for weather data from the nearest weather station at that location and time, which includes the air temperature, relative humidity, precipitation, snow, wind direction, and wind speed. Note that weather data was not able to be retrieved for some rows (probably because it just was not recorded or available for that time / location).

To prepare our final dataset that contains societal factors, 4 datasets were imported and utilized as follows

Google trends for the words “UFO” and “Alien” were imported from trends.google.com. The monthly trends data was joined to the base dataframe grouped by month, so that we could compare trends between searches and sightings.

Sci-fi Movie Data was imported from thenumbers.com. This data included yearly gross of the sci-fi movie industry, the name of the top sci-fi movie each year and its gross and the number of sci-fi movies released each year. This data was not joined to the base dataset.

Alcohol Consumption by State data was taken from a study conducted by the National Institute on Alcohol Abuse and Alcoholism (NIT) for the year of 2018. This was joined with the base data grouped by state and count was taken per 100K population in order to see if states with a higher alcohol consumption had higher sightings.

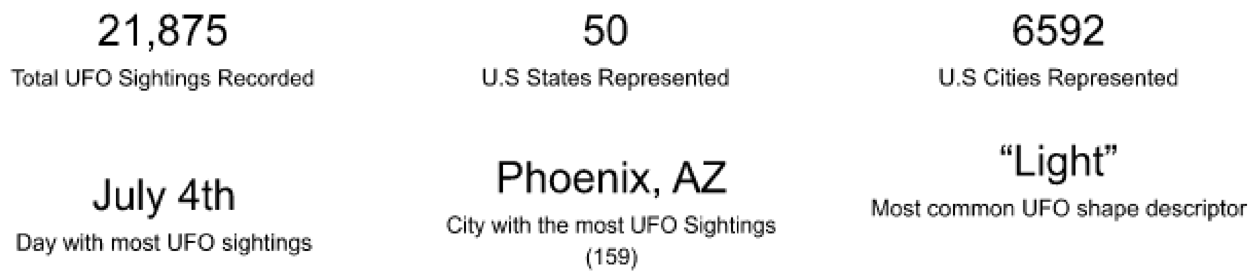
Education Data by State data was taken from a study conducted by the U.S Census Bureau for 2018 and plotted against UFO observations per 100K population. This was joined with the base data grouped by state and count was taken per 100K population to see if states with a higher education level had higher sightings.

To prepare our final dataset that contains **political data**, we imported the “2019 census US population data by state” from Kaggle [7], and the “party affiliation by state dataset” from the Pew Research Center [8]. We joined the two datasets by state. The party affiliation data came in the form of percentages, which we converted to proportions and calculated the difference in proportions between Republican and Democratic parties and stored in a new column.

To prepare our final dataset that contains **shapes data**, we summarized the number of sightings for each state by shape, dropped all observations that claimed light as a shape, because we decided that it should not be considered a shape, and kept the remaining shapes with the highest counts, ties were kept as well.

See **data dictionary** in the repository for a more detailed description of each of the variables in the final dataset.

PART 2D DATA AT A GLANCE



21,875
Total UFO Sightings Recorded

50
U.S States Represented

6592
U.S Cities Represented

July 4th
Day with most UFO sightings

Phoenix, AZ
City with the most UFO Sightings (159)

“Light”
Most common UFO shape descriptor

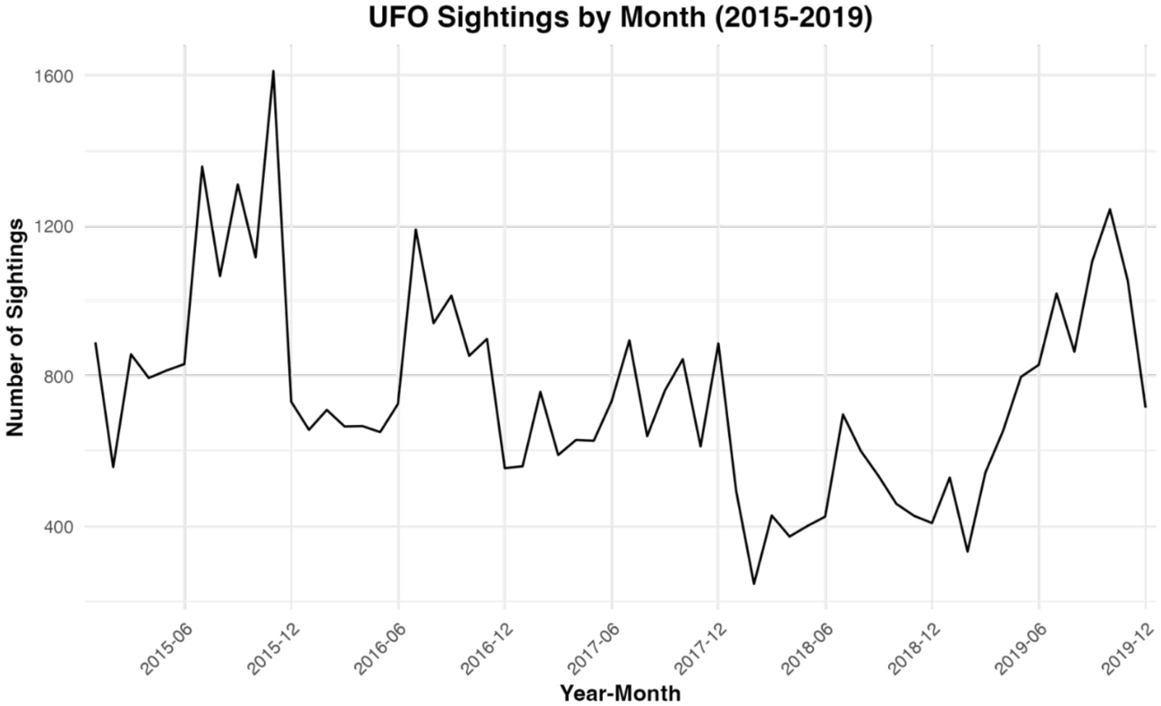


Figure 3
This time series graph depicts the number of UFO sightings reported over time from 2015–2019.

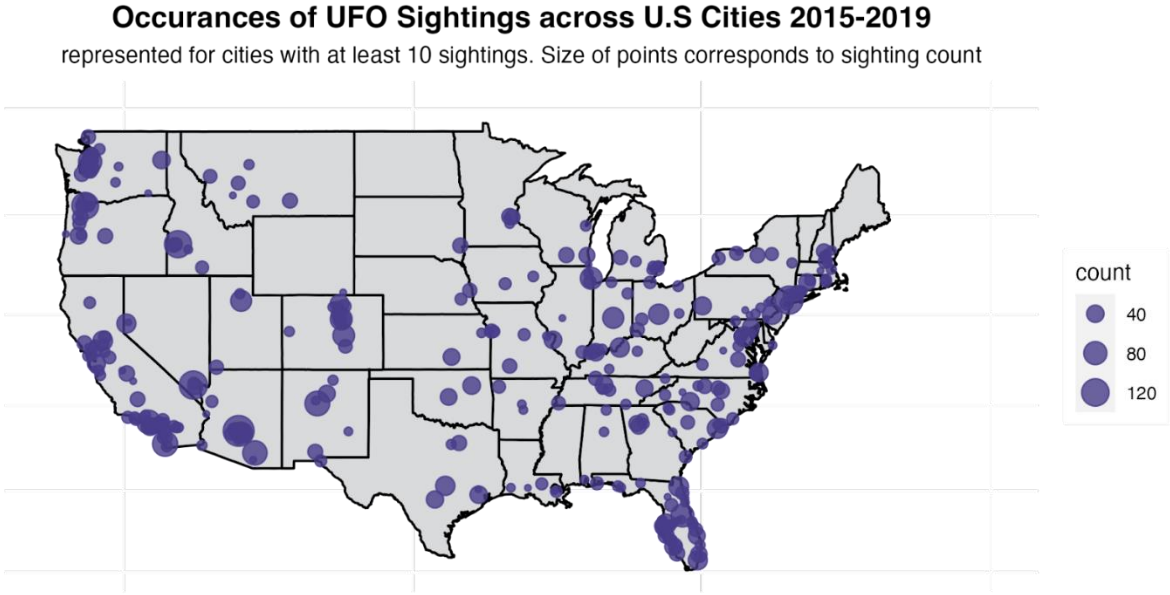

Occurances of UFO Sightings across U.S Cities 2015-2019
represented for cities with at least 10 sightings. Size of points corresponds to sighting count
count
40
80
120

Figure 4
This graphic depicts geographically where UFO sightings were reported across the United States from 2015–2019.

PART 3: ANALYSIS

PART 3A: MADAR AND WEATHER

3.A.1 MADAR: WHAT IS MADAR AND WHY ARE THERE REPORTS WHOSE DESCRIPTION IS JUST "MADAR NODE #"?

When first visually inspecting the data, we discovered that some entries only had "MADAR Node [number]" for the "text" variable which is typically a longer description of the sighting. Upon further investigation and research, we found that these observations were not UFO sightings from human reports but rather automated recordings from little MADAR device nodes that are scattered across the United States. MADAR stands for "Multiple Anomaly Detection and Automated Recording." It is a sensor that looks at the data background levels with a magnetometer and when it detects an abrupt change in the ambient magnetic field or compass heading in proximity to the device, it sends an alert to a central server and collects data much faster, recording information such as the MADAR node number, the changes in compass heading, the geomagnetic field reading, the threshold setting, and the barometric pressure all time stamped in UTC [5].

The purpose of these nodes is threefold.

1. They are configured such that when an anomaly is detected, it sends an alert to the owner of the device so they can make a human observation if they are available.
2. It allows human sighting reports (that are not initiated by nodes) to be correlated with sensor data and to see whether there are patterns across multiple nodes picking up anomalies at the same time.
3. The centralized alerting provides an automated way to report anomalies to the NUFORC database, which is why we see the MADAR node entries in our data [6].

To explore these observations from MADAR node recordings, we first looked at the percent of observations in our data that are from MADAR nodes. 630 out of 21,875 observations were taken from the nodes, comprising 2.97% of all observations. We also found that the 630 recordings from MADAR nodes were provided by 129 distinct nodes. Finally, we looked at both the raw count of observations from MADAR nodes by state and the percentage of observations from MADAR nodes to the total observations by state. Across both analyses, we found that Indiana significantly leads compared to all other states with 99 node recordings and 22.86% of the observations from Indiana being those collected from nodes. This made us wonder whether nodes are more likely to pick up anomalies in Indiana or whether Indiana just has a lot more nodes than other states. Thus, we looked at the number of distinct nodes in each state and found that indeed, Indiana possesses many more nodes than other states. Since the Command Center of the MADAR project is located at Newburgh, Indiana [4], it makes sense that the highest concentration of MADAR nodes would be located locally. And last, we found the correlation between the number of distinct nodes in a state versus the number of reports from nodes in a state and found it to be a strong .86 correlation. Therefore, states that have more MADAR nodes (like Indiana) are associated with more anomaly reports from those nodes to the NUFORC database. Although it is possible that there are states that have nodes that never picked up any anomalies in the five year period we analyzed, it is unlikely that those nodes comprise a significant amount of the total nodes. It is true that out of the nodes that did report anomalies, 62 out of those 129 nodes only made 1 report. However, out of the nodes that did make a report, the median number of reports each node made was 2 reports and the average number of reports was 4.88 reports.

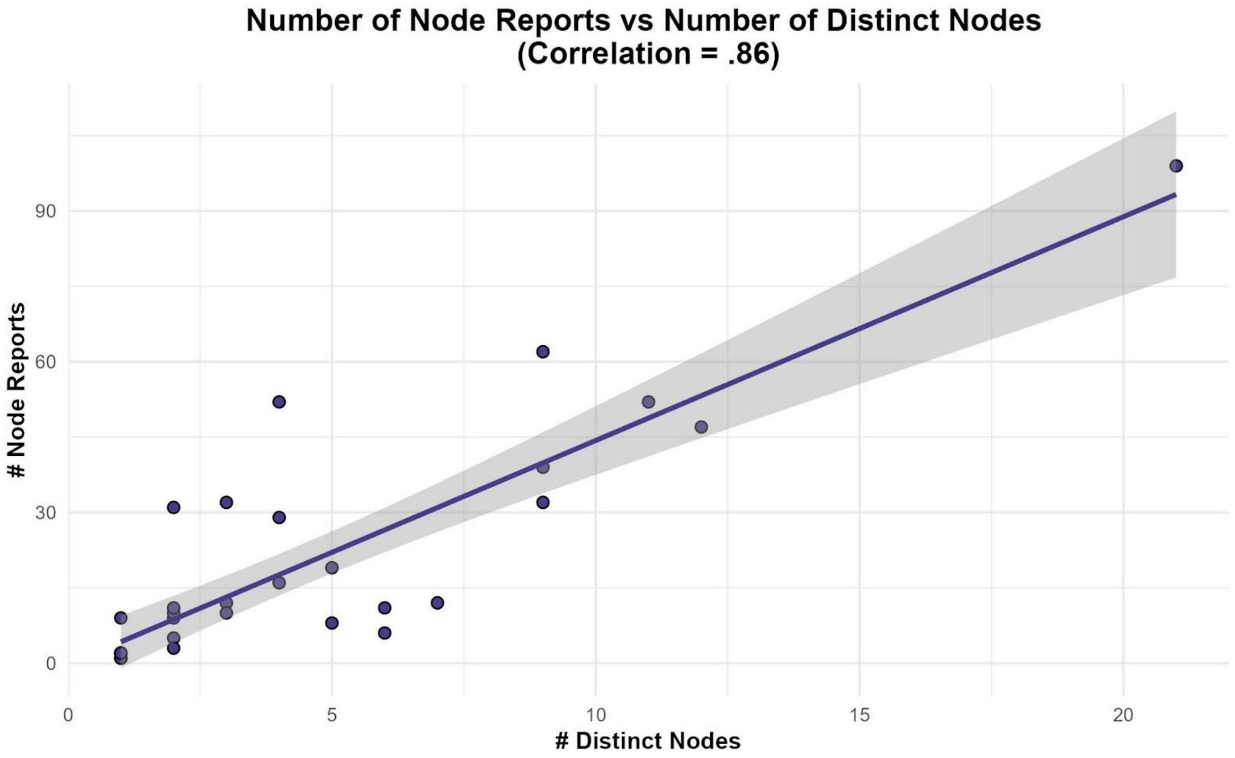


Figure 5
This plot depicts the relationship between the number of node reports and the number of distinct nodes where each point is a US state.

To perform our analysis of MADAR nodes, we had to do some feature engineering. First, we had to create a new variable that took on a boolean value of whether the text variable for an observation contains the word “MADAR.” This way, we could see which observations were from MADAR nodes and which observations were human reports. Next, to do our analyses that grouped by state, we had to create new variables such as the count of the reports from MADAR nodes, the count of reports (both MADAR and human), and the percentage of reports from MADAR nodes to total reports, which was the count of MADAR reports divided by the count of reports multiplied by 100.

Overall, although these reports from MADAR nodes comprise a small percentage of the total reports in our dataset and we did find anything beyond the fact that more nodes is associated with more reports from nodes for each state, it was nevertheless interesting to notice and investigate this detail.

3.A.2 WEATHER: WHAT IS THE WEATHER LIKE DURING UFO SIGHTINGS?

There were 6 variables related to weather: temperature, relative humidity, precipitation, snow, wind direction, and wind speed. To analyze weather, we first dropped all observations that were missing data for all 6 variables as that means we were not able to retrieve weather information for those observations. Then, when we analyzed each variable, we only dropped the observations that were missing data for that variable only.

The distribution of temperature across all UFO reports seems to be slightly negatively skewed towards temperatures in the warmer range. As expected, the more southern the state, the warmer the average temperature for UFO reports (figure 6). The temperature during UFO sightings seems to align with seasonal patterns across the years, as shown in figure 7.

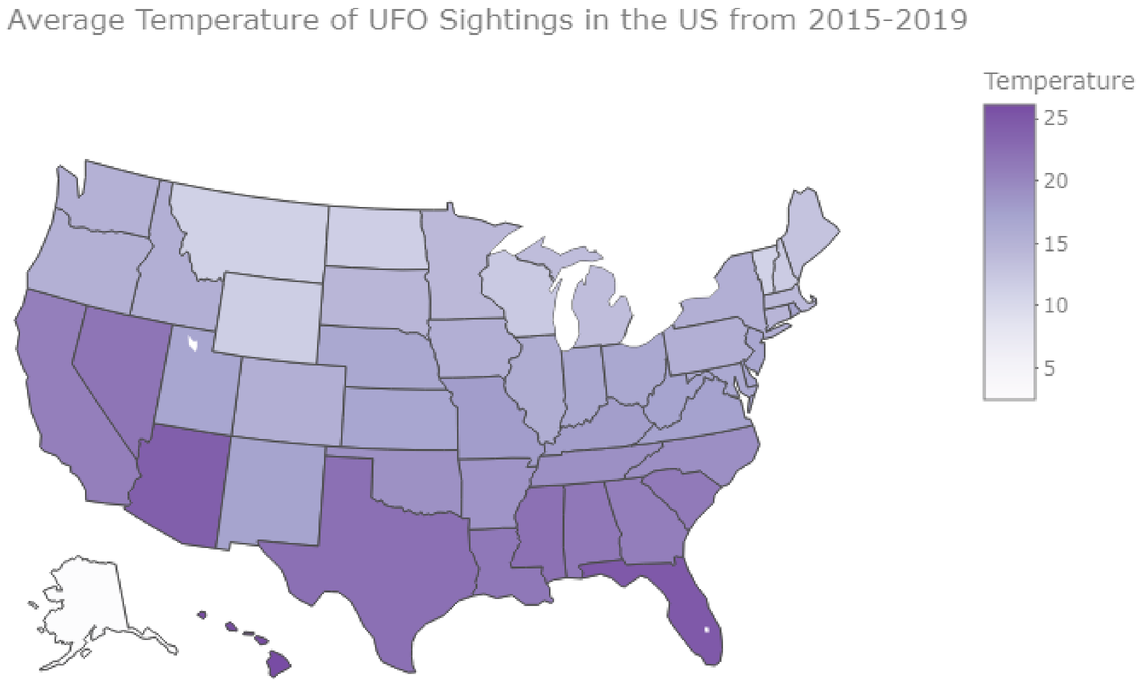

Average Temperature of UFO Sightings in the US from 2015-2019
Temperature
25
20
15
10
5

Figure 6
This plot depicts the average temperature for UFO reports for each state where the darker the color, the warmer the temperature.

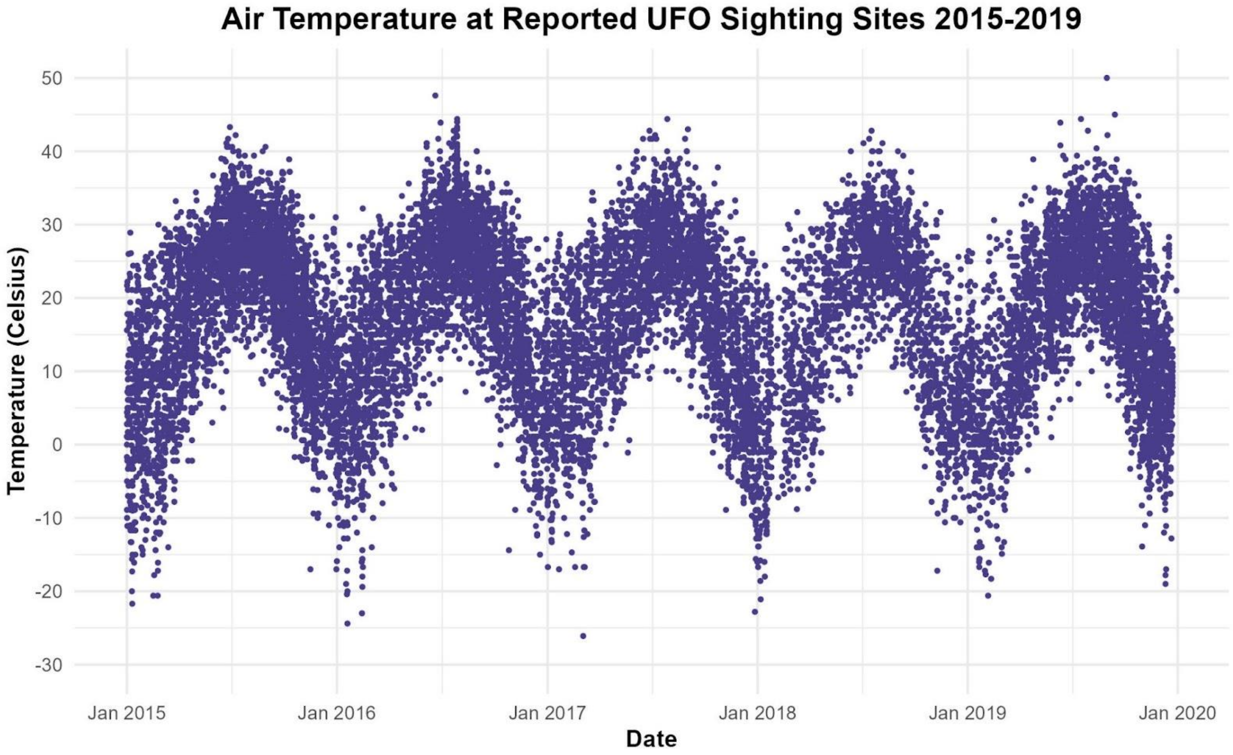


Figure 7
This plot depicts the temperature for each UFO report from January 2015–December 2019 where each point is a UFO report

Relative humidity for UFO reports seems to follow a normal distribution with the mean being 55.6% relative humidity, the median 55.0%, first quartile 39%, and third quartile 74%.

The vast majority of sightings occurred when there was no precipitation.

None of the UFO sightings occurred when it was snowing.

The wind direction for UFO sightings seems to be a fairly uniform distribution where there is no one wind direction that has more sightings than another.

The wind speed for UFO sightings seems to be a positively skewed distribution where for the majority of the time, the wind speed during sightings is low.

In conclusion, based on our analysis, there is nothing unique about the weather during UFO sightings.

PART 3B: FREQUENCY, SHAPE, AND TEXT ANALYSIS

FREQUENCY ANALYSIS

3.B.1 WHAT TIME OF THE YEAR DO MOST UFO SIGHTINGS OCCUR?

Our first question was if sightings had any seasonality. After investigation, we found that sightings were most heavily reported in the summer months of June, July and August (Figure 8). Warmer weather could be related to more people being outside participating in summer activities where there is more opportunity for them to look at the sky.

We dug a little deeper and found that July has the most sightings by quite a large margin. When we examine year to year, for four out of the five years, July has either the most or the second most sightings (Figure 8).

We drilled even deeper and discovered that the 4th of July is a huge outlier and has about 2.5x more sightings than all other days (Figure 9) in July. A possible explanation for this is that because there are many airshows and fireworks that go off on this day, releasing a lot of light and sound into the sky, some of these displays may be mistaken for UFOs. Because of these festivities, many people are also out at night looking at the sky, so a mere exposure effect could be at play. Additionally, it should be noted that since July 4th is a national holiday, one that is often celebrated with alcohol, the reliability of these reports may also be compromised in this way.

With such a high number of sightings on this holiday, this led us to wonder what other holidays may display a similar pattern of abnormal amounts of sightings. Therefore, we additionally investigated New Years and Halloween but did not find the same trend.

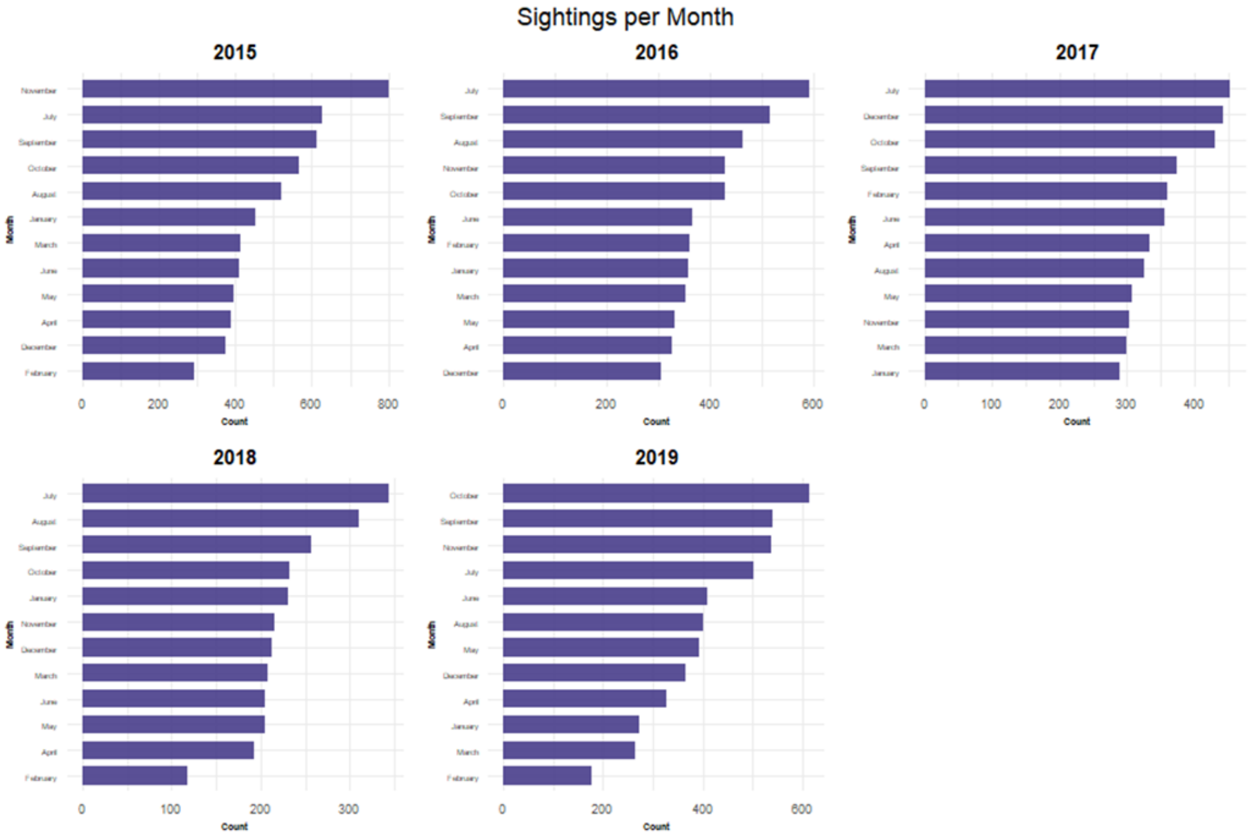


Figure 8

This plot depicts the number of sightings for each month in the years 2015-2019, organized by the months with the highest totals ascending.

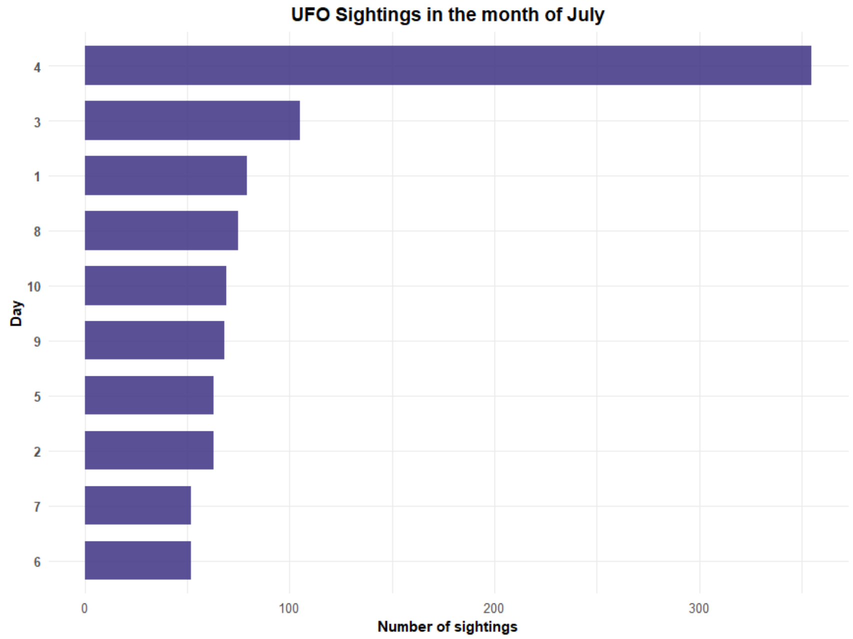


Figure 9

This plot depicts the number of sightings for each day in July in the years 2015-2019, organized by the top 10 days with the highest totals

3.B.2 WHEN ARE UFOS TYPICALLY SIGHTED?

Most sightings are at night, specifically between 6PM-11PM. The majority of the sightings occur when there is no or very little natural light. However, it is worth noting that there is a lower but steady amount of UFO sightings throughout the daytime hours as well (Figure 10).

Most sightings occur on the weekends (Figure 11). Combined with the fact that most sightings are at night, this suggests that many sightings occur during the "drinking" or social hours of the week. Additionally, the gradual increase of sightings throughout the week is similar to that of retail sales. This could suggest that the sightings are more of a function of people’s weekly schedules or habits rather than any real alien phenomenon.

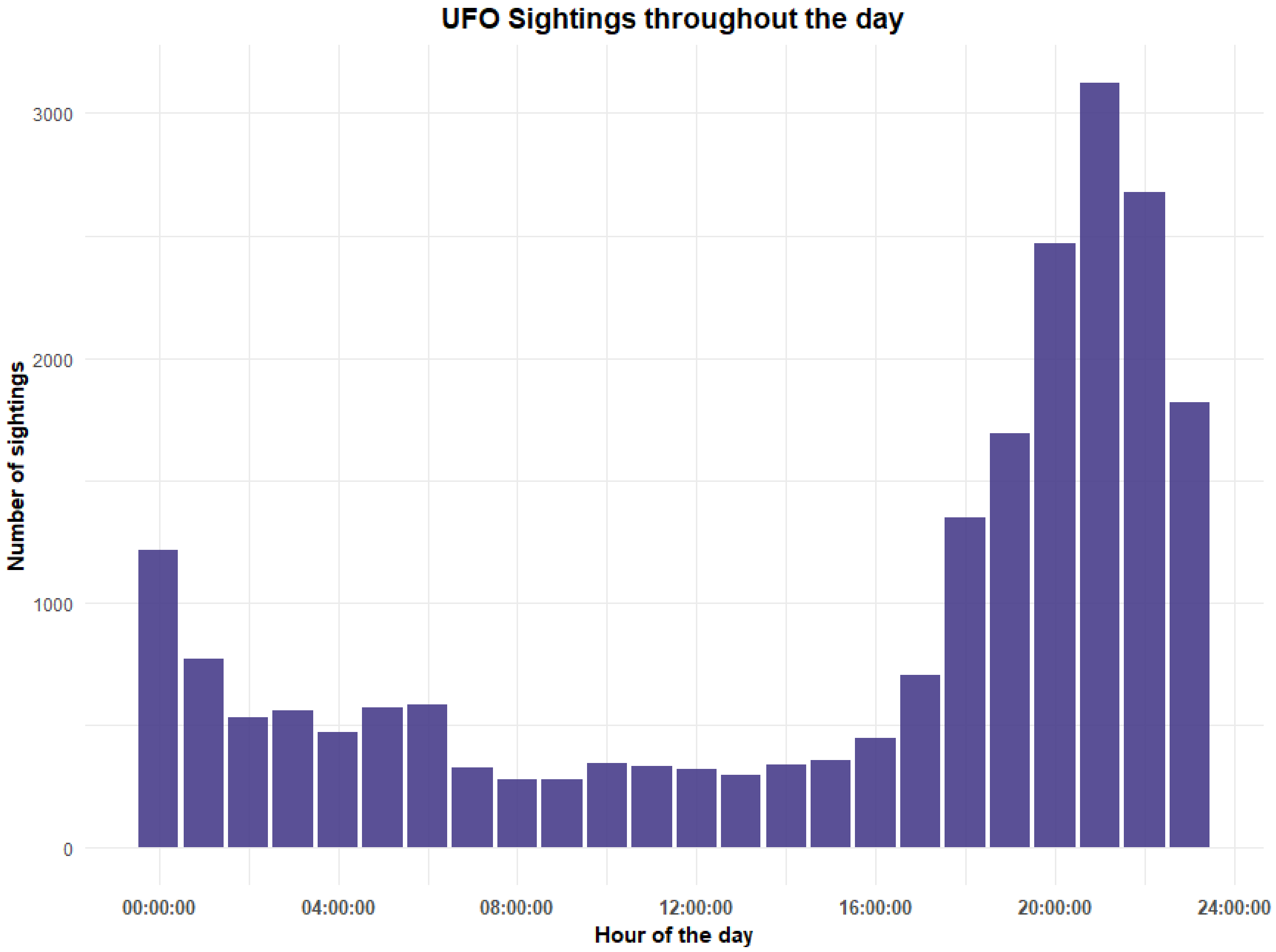


Figure 10

This plot depicts the number of sightings for each hour of the day, totalled for the years 2015-2019.

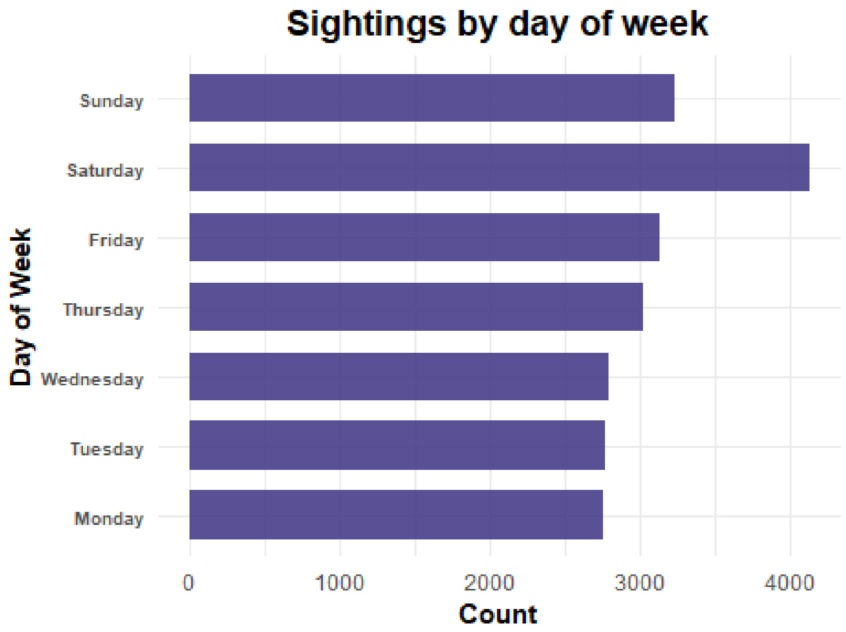


Figure 11
This plot depicts the number of sightings for each day in July in the years 2015-2019.

SHAPE ANALYSIS

3.B.3 WHAT SHAPES ARE MOST COMMON?

We were interested to know what shapes are the most prevalent in each state and if there is a pattern. From a summary of the shape variable, we found that “light,” “circle,” and “triangle” are the most popular terms regarding the shape of the UFO spotted. “Light” was removed from the list because the actual shape of “light” is ambiguous. Figure 12 shows that states had typically an even distribution of “circle” and “triangle” shapes and Figure 13 shows the most prevalent shape reported from each state. Circle appears to be the most prevalent shape in most states while triangle also appears to be the most prevalent in some southern (Tennessee, Alabama and Arkansas) and northeastern (Massachusetts and New Hampshire) states. Circle, triangle and oval are equally prevalent in sightings from Nebraska, and Sphere is most popular in Wyoming and fireball in Mississippi. Overall, there are no consistent patterns to be reported regarding the most prevalent shape from states.

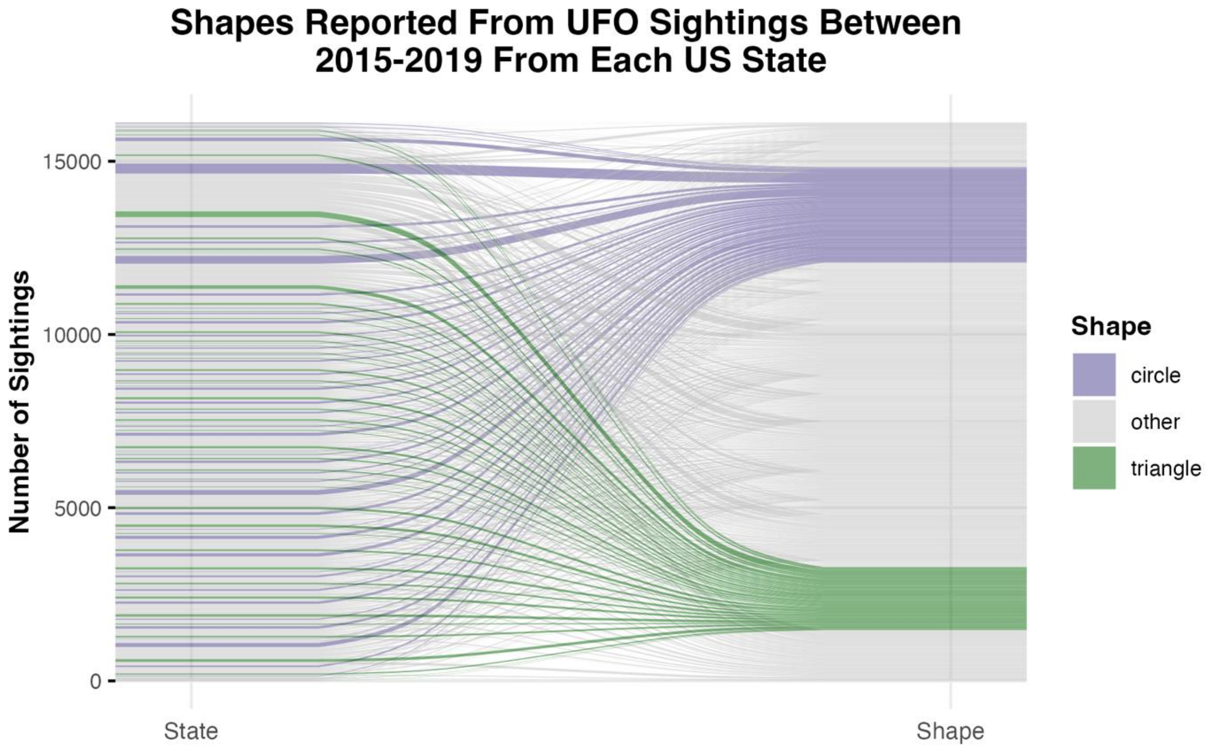


Figure 12
This plot depicts the distribution of shapes in each state. The left axis shows states, and the right axis shows shapes. “Circle” and “Triangle”, the most common shapes, are highlighted.

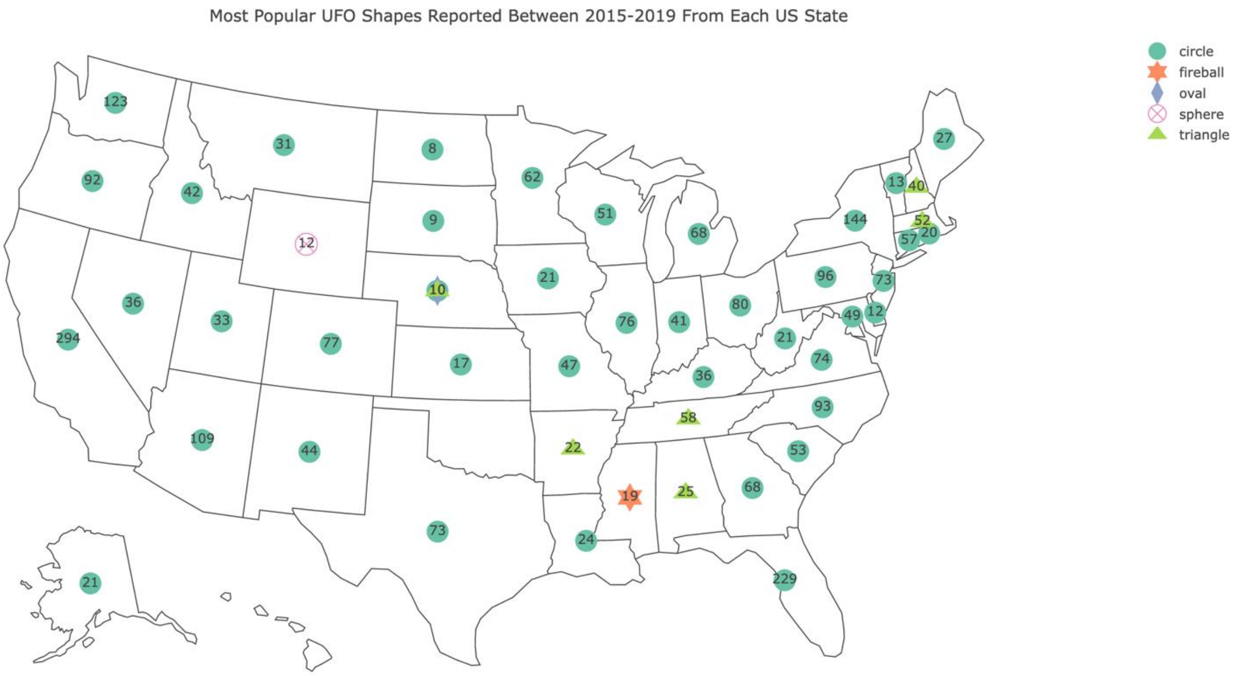


Figure 13
This plot depicts for each state, the most prevalent shape and occurrences of that shape reported from that state.

3.C.1 HOW ARE UFOS MOST COMMONLY REPORTED?

In this section, we evaluated the frequency of different words people usually use to describe UFOs. In order to evaluate the word frequency, first the reports had to be tokenized. Tokenization is the process of splitting a text document into smaller units; in our case this was splitting each report paragraph into individual words. Each of these smaller units (words) are called tokens.

On the first attempt to visualize word frequency, the most popular words were ones that are commonly used in the English language such as "a,”"the,” "is,” "are,” etc. We removed these words, called stop words, to focus on more meaningful aspects of the descriptions.

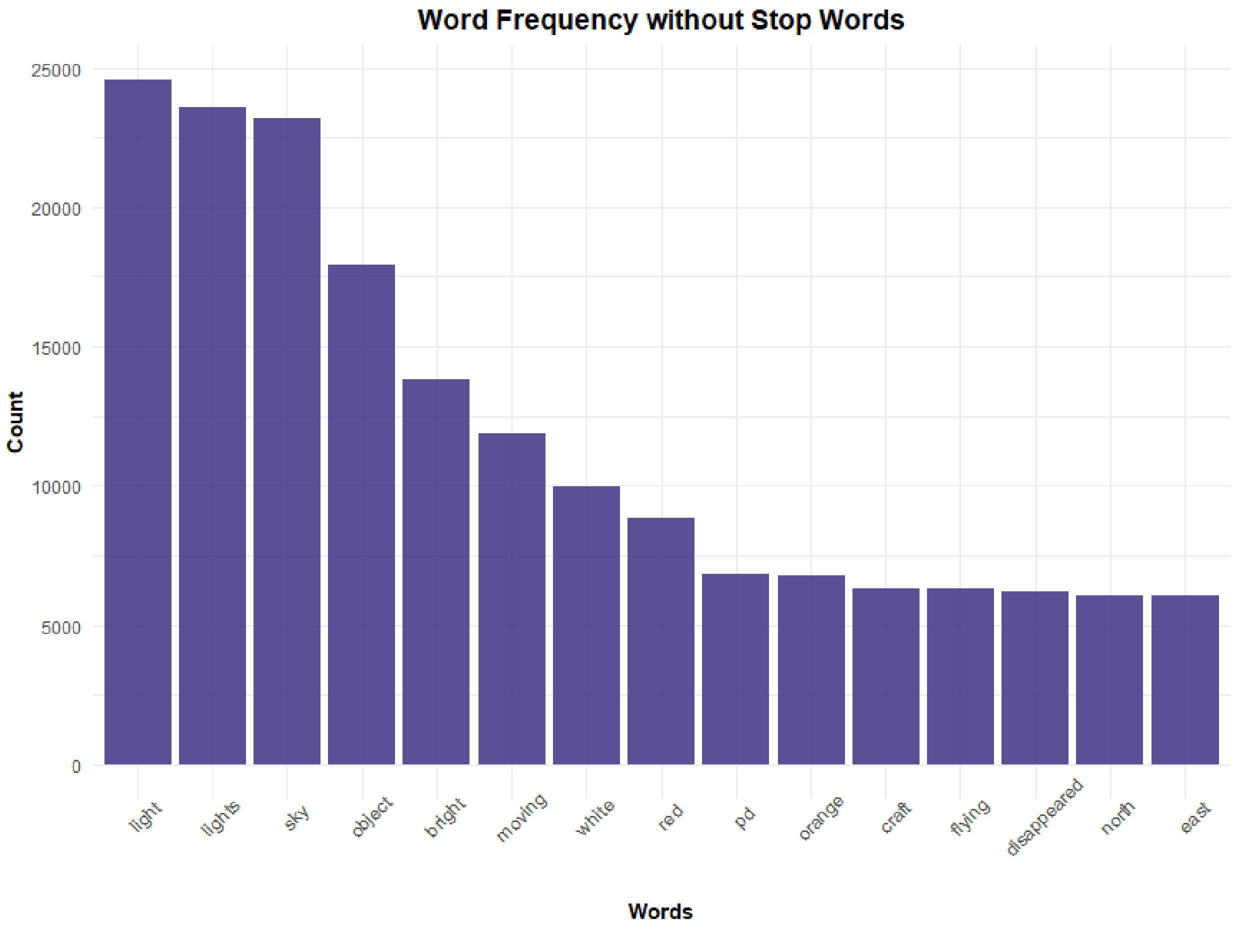


Figure 14
This plot depicts the 15 most common words that appear in the text of reports.

The most common words in reports describe the shape, movement, and color of UFOs. “Light”, “sky”, “object”, “moving”, and “looked” all make sense here. It is interesting to note that “light” and “lights” are the most common words, which make sense as we saw that most people saw UFOs at night.

Additionally, a word cloud was made to visualize the frequency of words in reports. A word cloud is a collection, or cluster, of words depicted in different sizes. The bigger and bolder the word appears, the more often it is mentioned within the entire corpus of text.

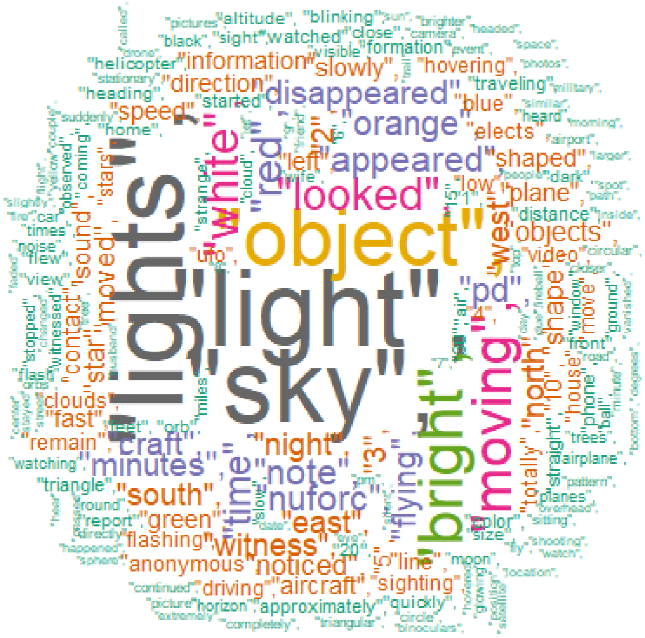


Figure 15

This plot depicts the most common words in the reports, with the relative size of each word indicating its frequency.

PART 3C: POLITICAL AFFILIATIONS

3.C.1 ARE UFO SIGHTINGS RELATED TO POLITICAL AFFILIATIONS?

As UFO sightings and discoveries are often associated with conspiracy theories, so are political ideologies. We were interested in whether people’s political party affiliations would be related to the frequency of UFO sightings. Specifically, we looked at party affiliations at the state level. We scraped data from the Pew Research Center on party affiliation by state [7], where they took a sample from each US state and asked for participants’ self-identified party affiliations, reported as percentages of Democrat, Republican, and no lean.

We joined the party affiliation data with our UFO sightings data by state and graphed the proportions of UFO sightings per 1000 residents in each state and their corresponding party affiliations. Each point on the dot plot is a state and the size is the party affiliation strength, which is defined by the absolute difference between the percentage of Democrats and the percentage of Republicans. The larger the dot, the stronger affiliation that state has to its corresponding party. One thing to note is that for the states that have a percentage difference between Democratic and Republican that is less than 3%, they were grouped in the “Swing” group.

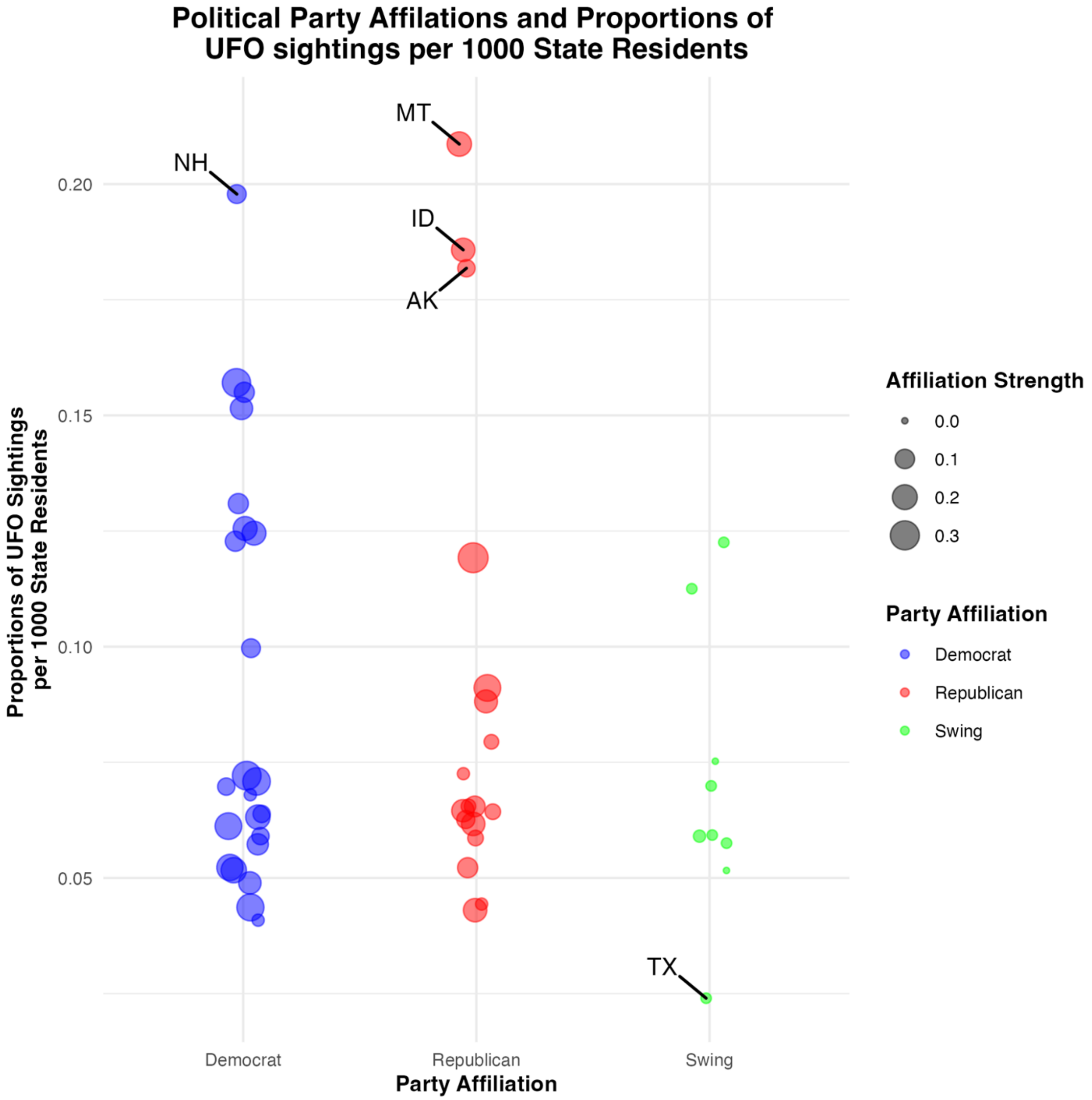


Figure 16
Political affiliations and UFO sightings per population in all US states. Each dot represents a state; outliers are denoted. The size of the dot corresponds to the affiliation strength and the color of the dot corresponds to the party affiliation.

There are some interesting takeaways from Figure 16. First, there is no strong correlation between party affiliations and UFO sighting frequency: both Democratic and Republican parties have states that have high and low frequencies of sightings. It is worth noting that most states, regardless of their party affiliations, are clustered in the low-frequency end of UFO sightings, including the swing states. Democratic states furthermore seem to have a more even spread in the frequency spectrum compared to Republican states which have two clusters on both ends.

The outliers on the high-frequency end are labeled on this plot. Notice that all of these outliers are states that are located in the northern part of the country (New Hampshire, Montana, Idaho, and Alaska). While on the lower-sighting-frequency end, the outlier is Texas, which is located on the southern border.

Although there may be some relationship between UFO sighting frequencies and party affiliation or geography, it is not entirely clear from this analysis what it is, if any.

PART 3D: SOCIETAL FACTORS

3.D.1 INTRODUCTION & MOTIVATION

An interesting aspect to consider is the relationship between societal factors and the prevalence of UFO sightings. This part is concerned with how external factors influence individuals within a society (in this case the U.S populace) that may affect the frequency of UFO sightings. Under this umbrella we identified and explored three potentially relevant societal factors to see if they informed UFO sighting frequencies.

1. **General Interest towards UFO's**. Because "interest" itself is an unquantifiable metric, we have developed individual proxy indicators that inform to some degree the societal interest in UFOs over time. Our two proxy indicators included google search trends for two adjacent search terms, and data from the Sci-fi movie industry.

2. **Alcohol Consumption between states** and its correlation between the number of sightings for a given state.

3. **Level of Education between states** and the correlation between the number of sightings for a given state.

3.D.2 CAN THE VOLUME OF GOOGLE SEARCHES INFORM SIGHTINGS?

Upwards of 75% of the total U.S population use Google [8]. Assuming that individuals search for topics of which they are at least partially interested in, the google trends data for "UFO" and its related topics becomes a reasonable proxy to gauge interest in UFOs overall. By definition, a trend lasts for only a short period of time, so we looked at whether searches and sightings coincide in specific instances in time.

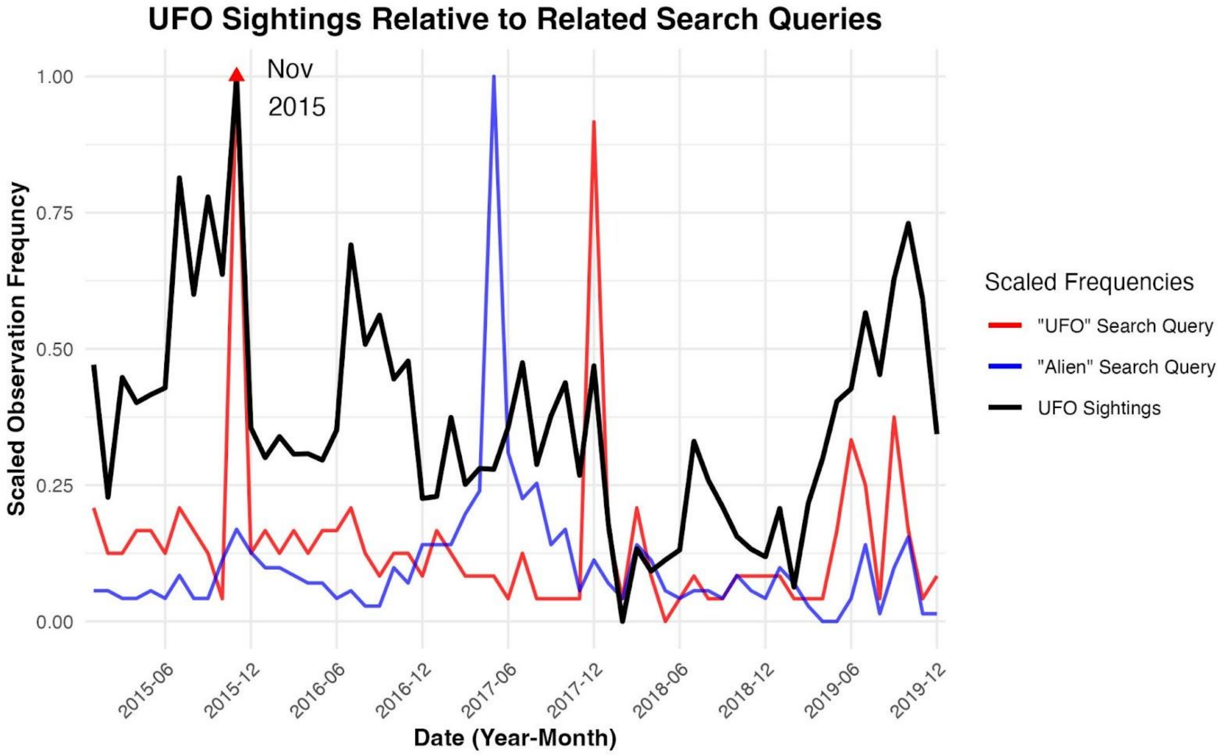


Figure 17
Time Series of Scaled Search Frequencies and Google Queries

We extracted google trends data openly available from trends.google.com. This uses a 0 to 100 relative scale to indicate how much a keyword is being searched for. We elected on the keywords “Alien” and of course, “UFO” to plot against UFO sighting count. Data for all three values were scaled between 0 and 1 to ensure comparability

We found that although peaks in the search trends did not match up with peaks in sightings for May 2017 and December 2017, the peak search month for the word “UFO” did coincide with the peak sighting frequency in November 2015 (Figure 17).

3.D.3 CAN WE INFER SIGHTINGS FROM THE SIZE OF THE SCI-FI MOVIE INDUSTRY?

The scale and selection of modern Sci-Fi movies is a reflection of our obsession with supernatural events. Due to the deep impressions that these movies have on viewers, we speculate that the monetary size of the Sci-Fi movie industry and the number of Sci-Fi films released may have a positive correlation with UFO sighting frequency.

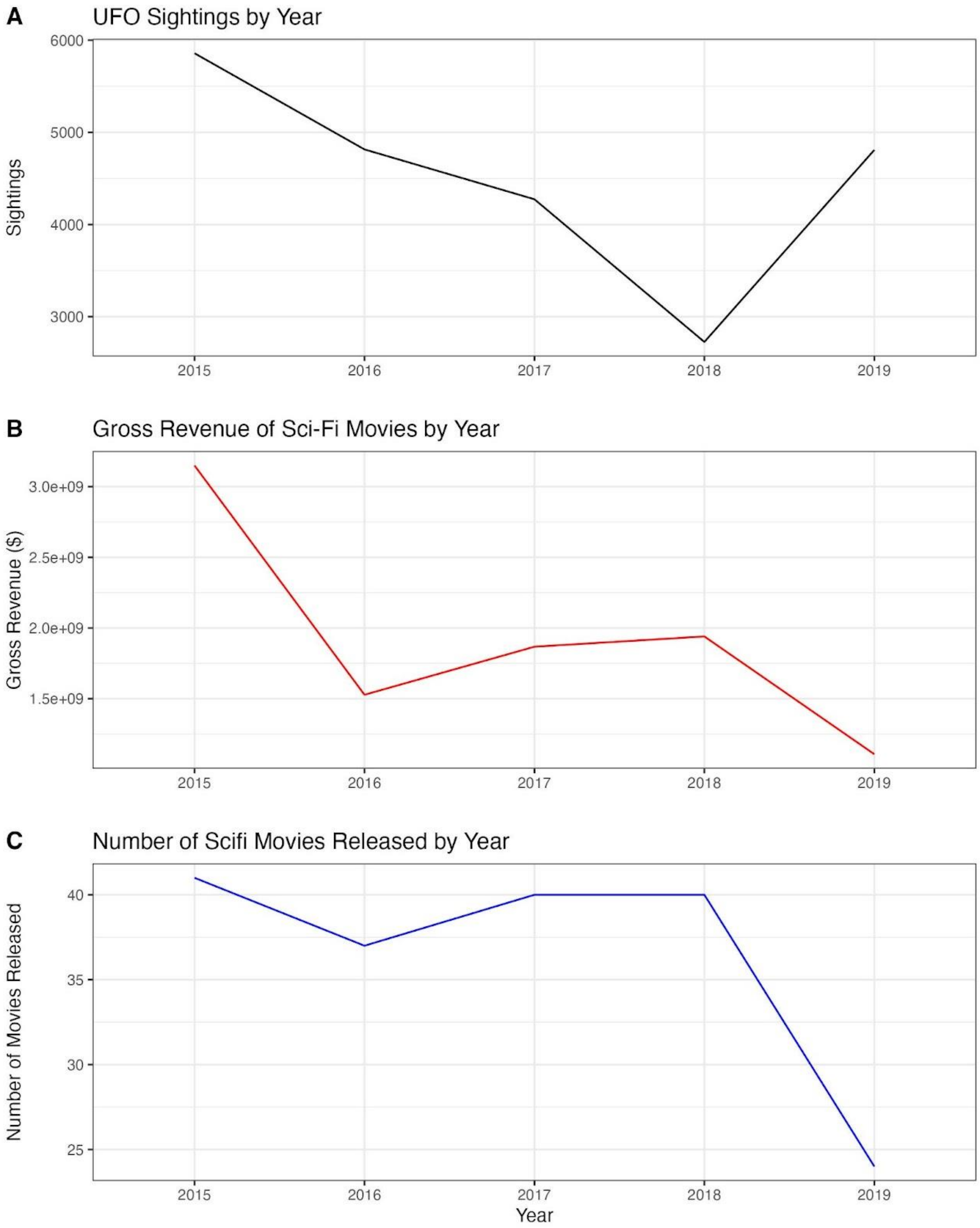


Figure 18

Comparisons of Sightings to Sci-Fi Movie Data

The yearly gross revenue and yearly releases of sci-fi movies per year were recorded by *The Numbers* [9], a freely available resource for movie data, and compared with yearly UFO sightings. We plotted these figures side-by-side with yearly sighting count, which we obtained by grouping sightings by year and then applying a COUNT aggregation function.

Gross Revenue: In the period 2015-2018, gross revenue declined 34% whilst UFO sightings declined 53%. However, between 2018-2019 the trends diverged, with UFO sightings increasing 35% whilst gross revenue declined 42%.

3.D.4 DO UFO SIGHTING PEAKS COINCIDE WITH MAJOR SCI-FI MOVIE RELEASES?

To examine the relationship more closely between movies and sightings, we plot sightings per month to examine whether its movement coincided with the release of the highest grossing Sci-Fi movie in each year.

We obtain monthly sightings by grouping by year and month, followed by applying a COUNT aggregation function. We mark the release dates of major movies manually to ensure the figure maintains readability.

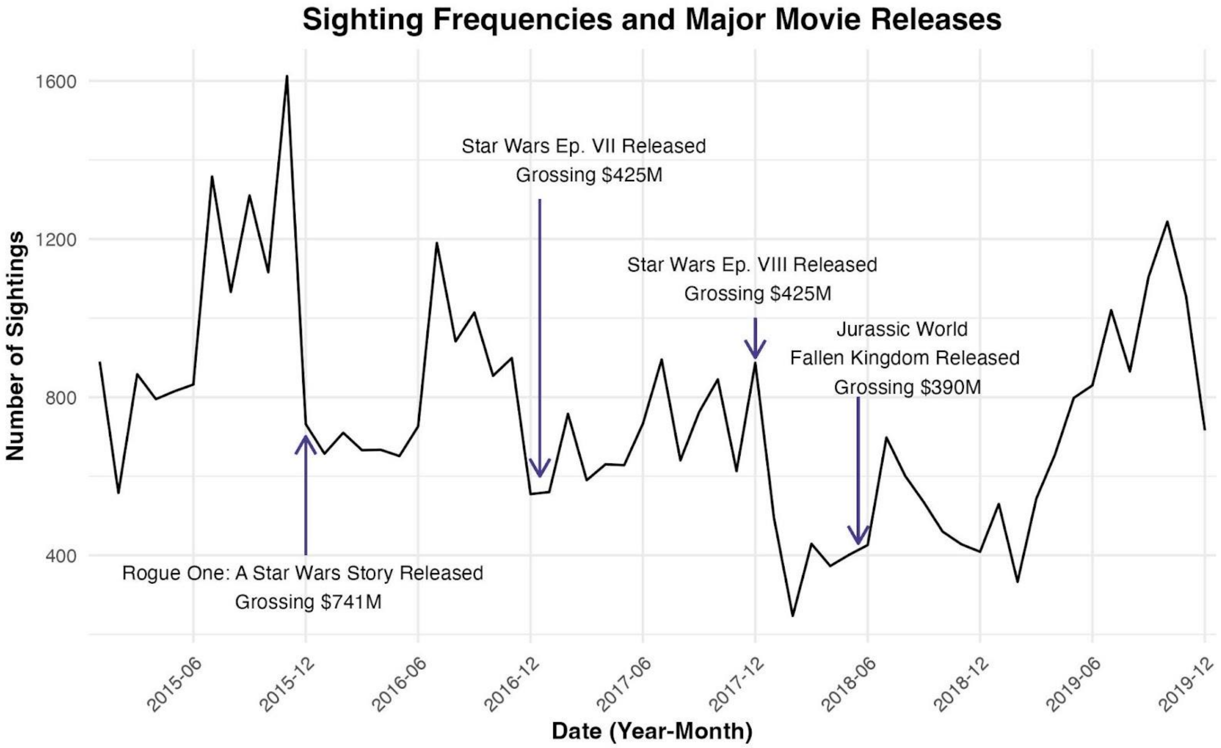


Figure 19
Sighting Frequencies After Big Movie Releases

We observe that the release dates of these major Sci-Fi films do not necessarily coincide with upticks in UFO sightings. Particularly interesting is that the release of Star Wars episode VII occurred in the month prior to that of the lowest recorded sightings (February 2018).

One major limitation to this analysis is that the genre of Sci-Fi films covers movies which contain topics reaching far beyond that which is relevant to UFOs. For instance, the Jurassic World movie which grossed highest in 2018 holds little to no indication of the American populace's interest in UFOs. A way to remedy this is to utilize a more focused dataset towards UFOs and extraterrestrial themes. However, those on the web fitting these criteria are often ill-defined or incomplete.

3.D.5 DOES A STATE'S ALCOHOL CONSUMPTION INFORM SIGHTING FREQUENCY?

It is widely known that alcohol affects an individual's perception [10]. In light of this, we deemed the relationship between alcohol consumption and UFO sightings worth investigating. We have already pushed the idea in section 3B (frequency analysis) that alcohol may have been a contributing factor to the relatively high number of sightings on weekend evenings, as well as the abnormally high sighting rate on July 4th. Here, we aim to challenge this speculation with more rigorous data.

Data for alcoholic beverage consumption by state for 2020 was taken from a study conducted by the National Institute on *Alcohol Abuse and Alcoholism (NIT)* for the year of 2018, and recorded in gallons of ethanol per capita [12]. We plot this data against the UFO observations per 100K population in each state to observe their relationship.

To plot the data, we first group sightings by state, followed by a COUNT aggregation function. We then divide these by the population of each state, to get the per capita. Note that this is not dissimilar to the wrangling done in 3.D.1. A left join is then done with the alcohol data on the state variable from which we can plot the following data. The mean consumption across states is also found and marked on our scatter plot.

In line with our previous hypothesis, we observe a moderate positive correlation between alcohol consumption and UFO observations per capita, with a correlation coefficient of 0.539. We note that New Hampshire outdrink every other state by 4 standard deviations and at the same time, rank second in UFO sightings. On the other hand, Utah, which is the state that consumes the least alcohol, actually has a lot more UFO sightings than one would expect from the trend but is still slightly below the mean.

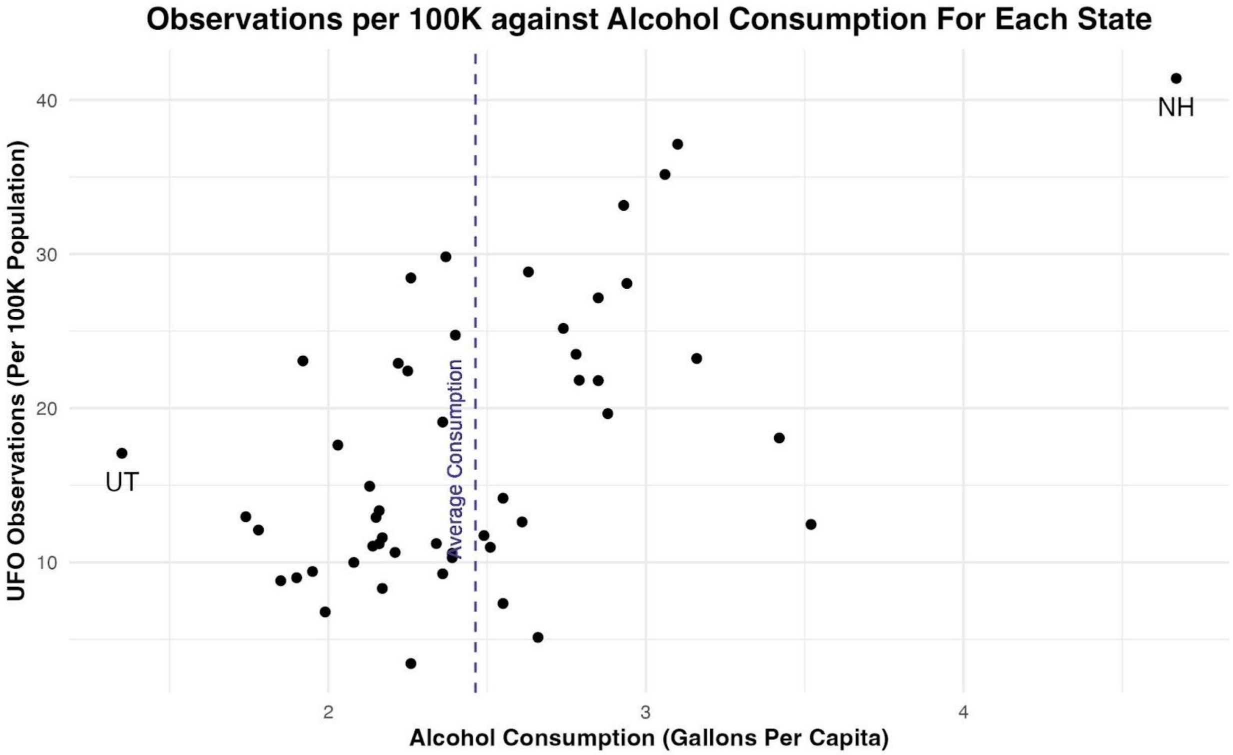


Figure 20
Alcohol Consumption per State against their Sightings Per Capita

3.D.6 DOES A STATE'S EDUCATION LEVEL INFORM SIGHTING FREQUENCY?

We also investigate the education level of states and plot against UFO sightings, similar to the analysis with alcohol above. We speculated that a higher education level within a state may be related to less UFO reports from that state. This is because more educated individuals may first eliminate more rational deductions about abnormal objects they see before settling on the conclusion that they have seen a UFO.

The metric that we used to measure education level is the percentage of the populace that graduated from high school. We use data collected by the *U.S Census Bureau* for 2018 and plot this against UFO observations per 100K population for each state.

Transformations for the following figure are similar to the one for alcohol above with the exception that the per capita sightings by state are joined with the percentage of population graduating high school to create the below plot.

Contrary to the initial hypothesis, UFO observation count actually has a moderate-low positive correlation with the percentage of the population graduating high school, with a correlation coefficient of 0.44. In other words, higher education levels in a state actually correspond to higher per capita sightings. States particularly noteworthy are Montana, who have the highest percentage of population graduating high school, whilst also having the highest number of UFO sightings per capita. Texas on the other hand are the second lowest in percentage graduating high school and have the lowest number of UFO sightings per capita.

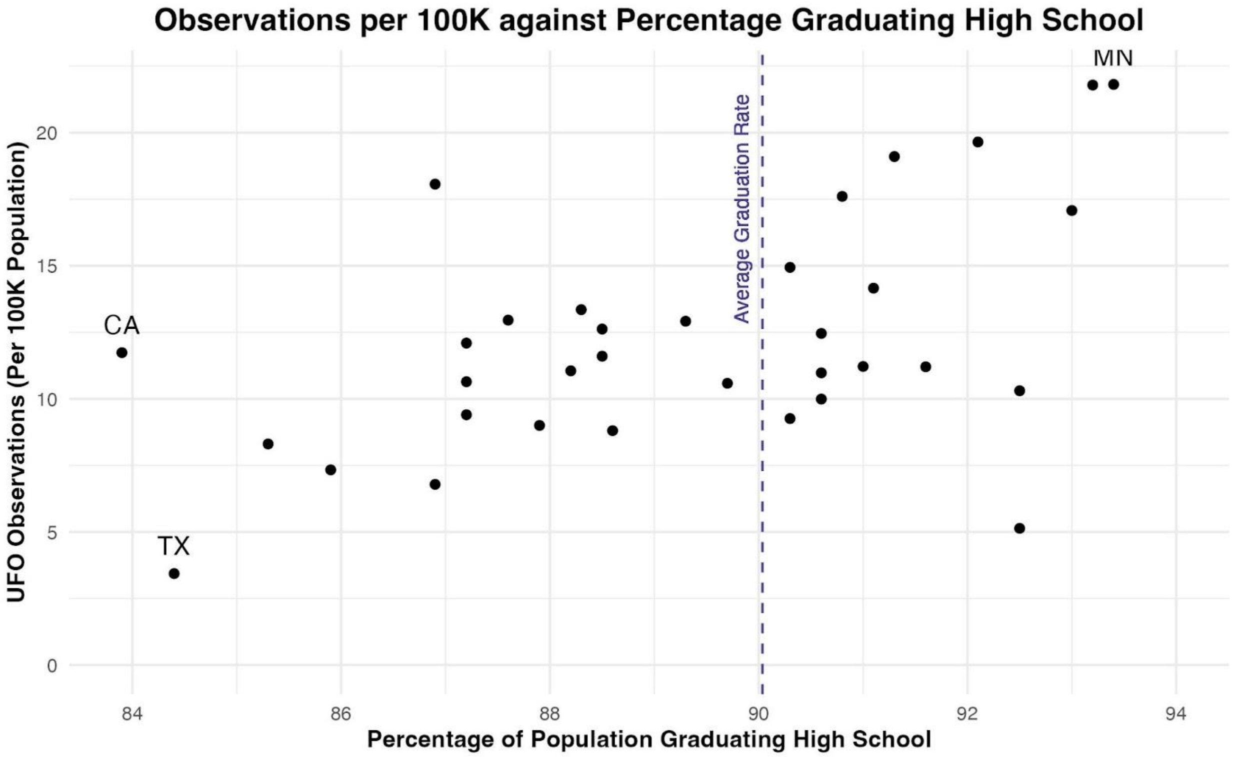


Figure 21
% Graduating High School per State against their Sightings Per Capita

PART 4 MAIN FINDINGS TL;DR

4.A MADAR AND WEATHER ANALYSIS

4.A.1 MADAR

States that have more MADAR nodes are strongly associated ($R^2 = 0.86$) with more anomaly reports from those nodes to the NUFORC database.

4.A.2 WEATHER

No significant patterns emerged from our investigation in weather patterns across UFO sightings.

4.B FREQUENCY SHAPE AND TEXT ANALYSIS

4.B.1 FREQUENCY ANALYSIS

Most sightings are reported in the summer months (June, July and August), among which July has the most sightings by a large margin. July 4th, in particular, has the most sightings out of any other day in the year. Finally, the majority of sightings occur at night and on weekends.

4.B.2 SHAPE ANALYSIS

The most popular shapes reported are “light,” “circle,” and “triangle”. There is an even spread of shape reporting in most states. “Circle” is the most prevalent shape in the majority of the states.

4.B.3 TEXT ANALYSIS

The most common words in reports describe the shape, movement, and color of UFOs. “Light,” “sky,” “object”, “bright,” and “moving” are the words with the highest frequencies.

4.C POLITICAL ANALYSIS

There is no strong correlation between party affiliations and UFO sighting frequency.

4.D SOCIETAL ANALYSIS

4.D.1 GOOGLE SEARCHES

Out of the three peaks in Google search for key terms “UFO” and “Alien,” peaks for May 2017 and December 2017 do not match frequency peaks of UFO sightings. However, the November 2015 peak aligns for the two.

4.D.2 SCI-FI MOVIE INDUSTRY

We found mixed results looking at the sci-fi movie industry’s gross revenue through 2015-2019 and the frequency of UFO sighting reports. In the period of 2015-2018, UFO sightings changed in the same direction as sci-fi movie revenues (both declining). However, between 2018-2019 the trends diverged, with UFO sightings increasing whilst gross revenue continued to decline. The release dates of major Sci-Fi films do not coincide with upticks in UFO sightings.

4.D.3 ALCOHOL CONSUMPTION

There is a moderate positive correlation between alcohol consumption and UFO observations per capita, with a correlation coefficient of 0.539.

4.D.4 EDUCATION

UFO sighting frequency has a moderate-low positive correlation with the percentage of the population graduating high school per state, with a correlation coefficient of 0.44.

PART 5: LIMITATIONS

PART 5A: LIMITATIONS OF DATA

SAMPLING BIAS TOWARDS THE SAMPLE OF UFO REPORTERS

It is of paramount importance to note that the sample of UFO sightings represented by our data is a biased sample of the entire population of sightings. Specifically, since our data predominantly relies on individual reports to the NUFORC, our data will be biased towards individuals who are more likely to make such reports.

The most glaring example of this bias can be observed in the distribution of sightings across countries. It should come as no surprise that sightings located in the U.S make up over 92% of total sightings. After all, as an American organization NUFORC's UFO reporting system will be much better known in the U.S as compared to the other countries recorded (who may have their own mainstream UFO reporting system).

The above reasoning could be extrapolated to other factors. It is plausible that citizens of a certain state may be more likely to report a sighting, or individuals that leave home at a certain time of day etc. These are all underlying factors causing bias in the data generation that we cannot account for in subsequent analysis.

It is thus important to reiterate that the data and any analysis thereof **represents only the sightings recorded by NUFORC's reporting system and may not be representative of any generalized conclusions about UFO sightings.**

FALSIFIED REPORTS

As of the writing of this report, the assertion of legitimate physical existence of UFOs is still an outlandish and niche claim to make. As such, any report of UFO sightings must also be treated with the same scrutiny.

There are many reasons that a report may be intentionally falsified including: elevating the personal fame of the observer, manufacturing the next big news story or other incentives. Nevertheless it is important to consider the conspiratorial nature of sightings and reports.

RELIABILITY OF PERCEPTUAL DATA

Even if we were to consider the case where no reports have been intentionally falsified, we are still left with the **inaccuracy produced by the perceptual nature of the data**. Each record of the data is susceptible to natural human error, and likely culprits include shape, time of day and description, which may be highly subjective to the individual reporting.

Following this logic, the existence of each row of data is also indeed dependent on the accuracy of human perception. This phenomena rears its head in different aspects of our dataset. For example, the majority of sightings occur in the evening, when a potentially intoxicated individual may mistake something else in the night sky for a UFO.

Sarah Scoles, author of *They're Already Here: UFO Culture and Why We See Saucers* offers insight into why many inaccurate sightings exist [14]:

> *"For most believers, it's mainly just fun. So, if they see something others might assume was a military test, they're probably not going to go read up on it and accept an official explanation. They want to believe."*

Ultimately, what constitutes an inaccurate report, and whether these reports are useful may be completely subjective to the purposes of the reader. But **it is important to note the inherent probability of "unreliability" in each observation of our data.**

PART 5B: LIMITATIONS OF ANALYSIS

In addition to issues with the data, we must also evaluate the robustness of our analyses. In this section, we list out some of what we deem to be the major weaknesses with both the analysis and conclusions drawn thereof of each of our main analysis categories. Note that this is not an exhaustive list and we recognize that the limitations may stretch beyond what we describe.

MADAR AND WEATHER: UNCERTAINTY BEHIND MADAR NODES

Whereas human observations are quite well defined and relatable, the observations made by MADAR nodes remains a mystery. For more rigorous analysis it would be best to know the reliability of their observations, especially relative to the more prevalent human observations. Furthermore, it is unknown how many of these nodes exist in total and whether there are active nodes that did not pick up a single sighting and thus were not present in the dataset. The presence of these may partially compromise the analysis done in figure 5.

MADAR AND WEATHER: SEPARATING SIGHTING WEATHER FROM EXPECTED WEATHER

A common challenge when breaking down weather data for sightings was identifying the uniqueness of the weather for a given sighting. Because weather changes throughout the year (at least for the U.S), plotting the weather over time or weather of sightings for states (figures 6 and 7) reflects the weather of a given time or state more than the uniqueness of the weather in that sighting.

FREQUENCY ANALYSIS: AMBIGUOUS UNDERLYING GENERATION

Although we observed certain patterns when exploring frequency of sightings by temporal factors (day of week, day of month) etc., it is *hard to derive conclusions on whether a given behavior was caused by the presence of more observants or more observations*. To explain with a concrete example: we found that July 4th was the day of the year with the most UFO observations, however we do not know whether this tells us that there were more UFO like objects to observe or whether there were simply more people outdoors which raised the chance of any individual observing a UFO. It is also notable that most sightings are at night and on weekends, which are typical "drinking" or social hours.

SOCIETAL FACTORS: RELIABILITY OF PROXY VARIABLES TO MEASURE INTEREST

In section 3.D.1 - 3.D.4, proxy variables were introduced as a measure of the level of interest of the American populace towards UFOs at a given instance in time. However, precisely how well each of these measures such an interest is ill-defined. For example, a google search may not necessarily be a result of a strong interest towards a topic but rather a fleeting thought or another related topic and the exact interest remains hard to quantify. Furthermore, as discussed, the genre of sci-fi encompasses broader subject matter than extraterrestrials and UFOs so using this as a proxy may cause inaccuracies.

SOCIETAL FACTORS: CAUSE OR COINCIDENCE?

The line between an actual relationship and pure coincidence were particularly blurred in the societal factors section, fundamentally due to a lack of data. Using the relationships in section 3.D.3 as the most glaring example, there are in total only five data points which is absolutely not enough to make definitive conclusions about the variables relationships. The same could be said for the state-wise alcohol consumption and education plots. Although there is more data, it is still insufficient to rule out purely random behavior causing the relationships observed.

PART 6 CONCLUSION

The goal of this report was to answer our driving question of what factors are related to UFO sightings. Although we generated some interesting insights through our investigation, it is still largely unclear whether there really is a relationship between UFO sightings and any of the factors we looked at even if there was some correlation (may be spurious). It is difficult to discern the degree to which each limitation we outlined played a role and how different factors may interact. We truly wonder what percentage of the reports in our UFO sighting dataset were true positives. Regardless, from our analysis, we now know some human factors and behaviors that are related, or not, to UFO sighting reports.

WORKS CITED

[1] https://www.nasa.gov/feature/nasa-announces-unidentified-aerial-phenomena-study-team-members/
[2] https://nuforc.org/about-us/
[3] https://data.world/timothyrenner/ufo-sightings
[4] https://www.kaggle.com/datasets/rishidamarla/ufo-sightings-approx-100000
[5] https://madar.site/madar/more.html
[6] https://www.enigmaticdevices.com/hunting-ufos-with-the-madar-iii/
[7] https://www.pewresearch.org/religion/religious-landscape-study/compare/party-affiliation/by/state/

[8] https://www.kaggle.com/datasets/peretzcohen/2019-census-us-population-data-by-state

[8] https://review42.com/resources/google-statistics-and-facts/
[9] https://www.the-numbers.com/market/creative-type/Science-Fiction
[10] https://www.narconon.org/blog/the-effects-of-alcohol-use-on-the-senses.html
[11] https://worldpopulationreview.com/state-rankings/alcohol-consumption-by-state
[12] https://pubs.niaaa.nih.gov/publications/surveillance115/CONS18.htm#fig7
[13] https://www.census.gov/data/tables/2018/demo/education-attainment/cps-detailed-tables.html
[14] https://www.discovermagazine.com/the-sciences/reports-of-rising-ufo-sightings-are-greatly-exaggerated