UFO sightings in USA and Canada 2016

[1]

18-12-2022

Group 13 members:

Joachim R Baumann | jobau19@student.sdu.dk
Emil Madsen | Emima20@student.sdu.dk
Phillip Nielsen | phnie19@student.sdu.dk

[1] https://www.washingtonpost.com/magazine/2021/08/11/stop-ufo-mania-no-evidence-of-aliens/

Abstract

Like many other mysteries of the ages, the sighting of UFO's has been a constant outlier throughout history, and few correlations between the sightings have been made, except vague descriptions and reports. This paper uses the methods from data visualization to identify the patterns, trends and outliers in a large dataset containing more than 5000 reports in the year of 2016 throughout the US and Canada in the hopes of discovering the mystery behind these sightings. Through several data visualization methods a R Shiny Dashboard was created, which started answering some of these mysteries, it became apparent which states the sightings are most commonly occurring, and that general words like "light" and "circle" had the most common use as shape descriptions.

Table of Contents

1. Background and Motivation

Over the years, UFO sightings have been an ever so fascinating occurrence. An unidentified flying object is the formal definition of the phenomena. Although there isn't much scientific proof of what exactly it is, it is fascinating to fantasize and wonder about the unknown in our universe, but who knows, maybe these objects are not even coming from outer space, but from our own world, created by humans. It remains unknown, and that's what makes this interesting to investigate further and for choosing this data set. To explore the unknown and find connections and the observations.

2. Project Objectives

This section contains all the questions, which the project seeks to find answers to through the use of data visualization.

The level of observations in states vary depending on a number of factors, like time of the year, what latitude and longitude they're spotted.
One of the main objectives in exploring this dataset, is to uncover if there are any links/connections between the observations, or if these are just random factors.

A list of questions that sparked curiosity in response to the dataset have been formulated, in order to explore the observations in the dataset.

1. **How often do people in the US and Canada see UFOs?**

   Is it a daily occurrence across states, does it happen more at certain latitudes and longitudes?

2. **What are the most typical seen shapes of UFOs?**

   The dataset contains the shapes observed, which can tell us if there are recurring patterns in the observations. Perhaps different observers have spotted the same object?

3. **Which time of year do UFO sightings most commonly occur?**

   Are the observations of the UFO's happening on the same day, same season, can you tell anything about when the sightings are most commonly occurring?

4. **How many UFOs are spotted in a given state/area?**

   Does the size of the state have an impact on the amount of sightings?

5. **Are there certain keywords which are more common in the sightings shape description?**

6. **How many UFOs are seen in a given state in a particular month?**

   Which months have unusual spikes in sightings?

7. **What are the outliers in the dataset?**

   Does the data show something unexpected?

3. Data

This section describes the data, and how it was processed, before being used for the visualization.

The chosen dataset was found on data.world:
https://data.world/aarranzlopez/ufo-sights-2016-us-and-canada

Throughout the dataset these are the types of variables the dataset contains (see table 1).

Table 1: Types of variables in the dataset

Variable	**Description**	**Type**
Data/Time	The date/time of the UFO sighting occurred.	Numerical - discrete
Country	Which country the UFO sighting was reported.	Categorical - nominal
City	Which City was the reported UFO sightings	Categorical - nominal
State	Which state was the UFO sighting reported.	Categorical - nominal
Shape	what shape was reported in the Sighting	Categorical - nominal
Summary	A small summary of the reported sighting	Categorical - nominal
Latitude	The Latitude of the sighting.	Numerical - discrete
Longitude	The Longitude of the sighting.	Numerical - discrete

The dataset contains 5178 different UFO sightings spread over the USA and Canada in the year 2016.

Numerical data:
There are 2 types of numerical data. Discrete and continuous. A discrete variable, is a number that takes on distinct, countable values. In theory, you should always be able to count the values of a discrete variable.
A continuous variable is a variable that can take on any value within a range. A continuous variable takes on an infinite number of possible values within a given range

Categorical data:
A nominal variable has no ordering to its categories. For example, gender is a categorical variable having two categories (male and female) with no ordering to the categories.
An ordinal variable has a clear ordering. For example, temperature as a variable with three orderly categories (low, medium and high).

3.1 Data Processing

This subsection will go through the process of extracting different data from the dataset.

There were a few problems with the data before it could be visualized. First of all, due to not having checked the rows in the set, which contained 5178 entries, the date/time was not formatted in a consistent standard way. The first 200 rows were formatted as the Cherokee Calendar, while others had formats that resembled european style date formats. Some columns had the exact timestamp the observations occurred, while others were standard American dates. This started the question, how to clean up this data?

By using a combination of excel's auto cell formatting, and using a custom python script to extract the right values. One of the python scripts created had the functionality of reading every single column in the excel data spreadsheet, extracting the day and month, the year wasn't needed, as the dataset contained sightings only from 2016, this was written to a new spreadsheet.

The new date format was added to the dataset in a new row.

Visualizing state observations

In order to display the monthly state observations for a number of states in an animated graph, it turned out to be more of a data processing task than a visualization task. The dataset did not have a total count for observations per state at each particular day of the year, nor was there a monthly count per state.

The new data now looked like (table 2), but it was necessary to extract some values to get the needed table.

Table 2: Modified table with separation of Date and Time

Date	Time	State	Shape	Summary	lat	lng
12-21-16	19:15	VA	Sphere	Bright round object hovering in sky. ETC	38,0652286	-78,9058876

There is no data for daily observations in a particular state, which is needed. Loop through every row in the dataset, extract the date and state, save the date and state in a list together, if the next column contains the same data and stat, a counter will increase for that day/state.
Another script to extract, sort and count the daily, monthly, state observations was required. (Appendix: Script 1)

Table 3 shows what our data looks like when counted and sorted.

Table 3: Table of States and number of observations pr. day.

Date	State	Observations
12-21-16	VA	1
12-20-16	CA	1
12-19-16	AZ	2

Data processing for word cloud visualizations

In order to gain insight into what people described when they observed a UFO, a way to analyze the summaries was needed. The summaries of all the rows were gathered in a single text file, filler words and conjunctions were removed, all individual words, plus the amount of times they occurred had been counted.

Shapes observed

By using the earlier python script (see Appendix: Script 1) and modifying it slightly, it was easy to count the amount of observations of a particular shape. This gave as a new table to work with (table 4):

Table 4: A table showing how many times each shape has been mentioned

ID	individual	value
1	Circle	320
2	Cone	240

4. Visualization/Dashboard:

This section contains the official requirements of the project, and how these requirements will be met, through the dashboard.

Before starting to make the actual visualization of the data, a plan for the visualization was made, in order to keep track of the end goal.
As this paper is being written as a mandatory project a few formal requirements had to be met, which are the following:

- You must have at least three types of graphs (i.e barchart, time series plot or boxplots)
- at least one animated graph (using for example gganimate).
- In total at least 8 graphs. Provide clear and well-referenced images showing the key design and interaction elements.
- A link to the dashboard/Visualization must also be included in the report.
- An option to download the report as a manual from the dashboard

In order to effectively visualize the data in our set, and answering the questions with said visualizations, some ideas on how to visualize the different things were brainstormed and prioritized after "must haves" and "could haves".

It was quickly decided that a map with each sighting was a "must have", and it would be a useful functionality to have some mouseover effect to show each sighting, together with the rows of data.

The visualization should be able to show how many sightings were in each state for the year, this could be done with a bar plot, and a choropleth map, both gives a good view of the number of sightings. it should also be possible to see the frequencies of sightings over the different months, to see if some months were more UFO active than others. This should be animated to further visualize the different patterns there are in the observations.

For showing the shapes and amounts of sightings for the UFOs, both a bar-plot and word cloud would be nice to have, as these both give a good visualization of the different shapes, the bar plot would show how many sightings of the different shapes, where a "word cloud/tag cloud/weighted list" would show the difference, by making the more common shapes bigger than the less seen shapes, word cloud is often used with visualizations of text data, which in this case is a perfect fit.

5. Results

The R Shiny Dashboard

The R Shiny application primarily consists of 2 components, the Server and the UI. To make the appearance of a Dashboard with different sections to select from, the “shinydashboard” package was used for the main layout of the app. The dashboard has three main parts: a header, a sidebar, and a body. The sidebar is where categories are shown, while the body has the functionality to display visualization content. This is done by adding boxes which function as a container, inside this container it is possible to display the desired content.

How often do people in the US and Canada see UFOs?

This depends entirely on the particular state in the country, as the sightings are varying in frequency. Inspecting the monthly sightings frequency graph for 4 chosen states, it is clear that the range of monthly sightings lies between 1-25 occurrences.
However, this is only for a select few states, and looking at the mapview of the yearly observations it is clear that states such as north dakota and montana are experiencing less sightings.

What are the most typical seen shapes of UFOs?

Shape distribution bar plots and circle plots aim to visualize the different shapes and their frequencies. The most observed shape was clear to be of “light”, which was found odd, as “light” is not really a shape of any form, but more a natural phenomena. The second most spotted shape was of the circle shape, which was to be expected, as most the stereotypical description of a UFO is the “flying plate”.

When are UFO sightings most commonly occurring?

There are no clear visualizations to tell, if there are any recurring patterns or any particular day of the year/season which the sightings are most commonly occurring.

When inspecting monthly state observations, the frequencies seem to be of random order, with no dependency/connection to certain time periods having more observations.

How many UFOs are spotted in a given state/area?

The barplot with the yearly state observations gives a clear picture of the yearly UFOs spotted.

Looking at the mapview of our total yearly observation per state and comparing it to a population density map, it is possible that there is a resemblance between the two.
There seems to be a connection between the amount of citizens in a state and the amount of observations. Could it be that the number of citizens increases the chances of a UFO being spotted, as there are more people observing?

Image 1:UFO sightings pr. State

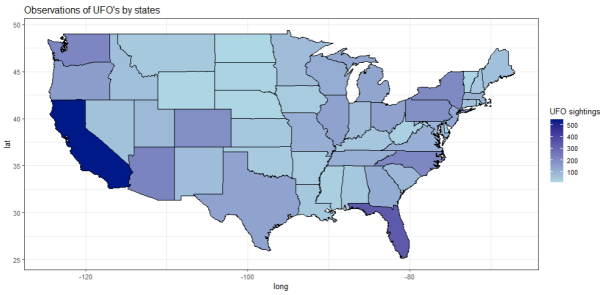
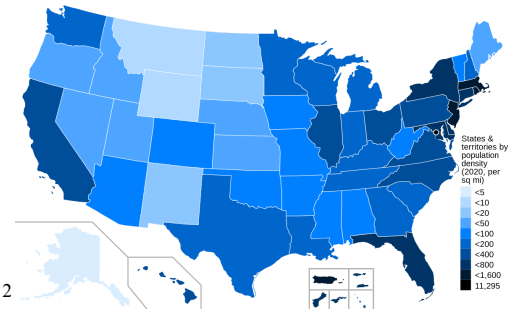


Image 2: U.S. population density map



Are there certain keywords which are more common in the sightings summary?

In the chosen dataset, there is a row which is a summary of the sighting. By gathering all the summaries into one file and analyzing these shapes, it is possible to find out which words are used to describe the most common sightings.

The word cloud adds interactivity to the visualizations, it is possible to sort and analyze the words used more in depth by using the sliders. The frequency slider gives the viewer the opportunity to sort the words all the way from 1 to 1000 frequencies, as well as the amount of words displayed from 1-80.

How many UFOs are seen in a given state in a particular month?

It was a goal to animate this for a set of states to show the observations in each state monthly. Showing all 50 states in a single graph would be too overwhelming, the option of choosing states which resembled each other the most in population size was used.

Are there outliers or anomalies in the dataset?

At first glance, it might appear like the state of california is an outlier in the dataset, as the amount of observations are much higher compared to the rest of the states. Inspecting the monthly observations animated graph, California is always the state with the most observations per month. However, due to the population density of the different states, it is not necessarily an outlier, as the population density in California is much greater compared to other states such as Montana.

[2] Image 2:
https://en.wikipedia.org/wiki/List_of_states_and_territories_of_the_United_States_by_population_density#/media/File:U.S._states_and_territories_by_population_density.svg

6. Conclusion

In hindsight, the chosen dataset had very few dependent variables, and there was a clear consensus that in future data visualization projects there would be less focus on the quantity of data, and instead the group would prioritize datasets with more dependant, numerical values, allowing for more in-depth analysis.

After grieving our choice of dataset, it was decided to focus on the values we had, to see if it was possible to do some data visualization and answer some of the questions the group had originally hoped to answer.

Starting with the shapes of the UFOs, both a barplot and a word cloud were created, both of these clearly indicated that the words “light” and “circle” were the most common descriptors.

The most observations made was in California, but this was partly due to the overwhelming population size compared to other states, making observations more frequent as there are people living in this region to observe it.

7. Discussion

We choose to make our dashboard by using R Shiny, which none of the group members had any experience in. R Shiny, when getting started, is a very powerful tool for quick visualization of data. The way rendering plots/graphs in the UI worked, felt very natural and making changes to existing code as well as adding features was made easy in an intuitive way. The ability to host the R Shiny application in your web browser was another very clear advantage of the Package. This also gives the opportunity to (in the future) host our own dashboard on the web (shinyapps.io).

In future courses, the lessons should keep using the powerful R Shiny package, as it doesn't require too much coding skills to actually make some nice visualizations, although it is nice to have the opportunity to choose whichever language and framework to use, Shiny is best for beginners.

Appendix

(Script 1)

```
class Entry:

 def __init__(self, date, state):
   self.date = date
   self.state = state
   self.observations = 0

 def __str__(self):
    return self.date + ", " + self.state + ", " + str(self.observations) + ", "

class EntryHolder:

   entryList = []

   def exists(self, entry):
      for x in range(len(self.entryList)):
         if self.entryList[x].date == entry.date and self.entryList[x].state == entry.state:
            self.entryList[x].observations = self.entryList[x].observations + 1
            return
      self.__addEntry(entry)

   def __addEntry(self, entry):
      self.entryList.append(entry)

   def getEntries(self):
      return self.entryList


#main script here

if __name__ == '__main__':

   # load excel with its path
   wrkbk = openpyxl.load_workbook("Date_State.xlsx")

   sh = wrkbk.active

   entryHolder = EntryHolder()

   # iterate through excel and display data
   #max_row = 5178
   for row in sh.iter_rows(min_row=2, min_col=1, max_row=5178, max_col=2):
      date = row[0].value
      state = row[1].value
      e = Entry(date, state)
      #print("Entry: " + date + ", " + state)
      entryHolder.exists(e)
```

```
print("finished counting/sorting, writing to text file")
with open('date_state_observations.txt', 'w') as f:
    for x in entryHolder.getEntries():
        #print(x.__str__())
        f.write(x.__str__())
        f.write('\n')
```