Introduction to Data Science Final Report

UFO Sightings Dataset Analysis

Sarah Morris

IST 687

Introduction

I chose to analyze the dataset UFO sightings (https://www.kaggle.com/datasets/NUFORC/ufo-sightings/data), which is a dataset of 11 attributes and over 80,000 rows. The data stretches back over the last century and includes reports of sightings worldwide. Each row represents one instance of a sighting, which vary in location and duration.

	datetime	city	state	country	shape	duration..seconds.	duration..hours.min.	comments
1	10/10/1949 20:30	san marcos	tx	us	cylinder	2700	45 minutes	This event took place i
2	10/10/1949 21:00	lackland afb	tx		light	7200	1-2 hrs	1949 Lackland AFB&#4
3	10/10/1955 17:00	chester (uk/england)		gb	circle	20	20 seconds	Green/Orange circular
4	10/10/1956 21:00	edna	tx	us	circle	20	1/2 hour	My older brother and t
5	10/10/1960 20:00	kaneohe	hi	us	light	900	15 minutes	AS a Marine 1st Lt. flyi
6	10/10/1961 19:00	bristol	tn	us	sphere	300	5 minutes	My father is now 89 m
7	10/10/1965 21:00	penarth (uk/wales)		gb	circle	180	about 3 mins	penarth uk circle 3mi
8	10/10/1965 23:45	norwalk	ct	us	disk	1200	20 minutes	A bright orange color
9	10/10/1966 20:00	pell city	al	us	disk	180	3 minutes	Strobe Lighted disk sh

The dataset includes the following attributes:

- Datetime (char) – this shows the date and time the sighting took place.
- City (char) – the city in which the sighting occurred.
- State (char) – the state (if applicable) in which the sighting occurred.
- Country (char) – the country in which the sighting occurred.
- Shape (char) – the shape of the flying object (categorical).
- 'duration (seconds)' (double) – duration in seconds of the event.
- 'duration (hours/min) – duration in hours/minutes of the event
- Comments (char) – a longer description of the sighting.
- 'date posted' (char) – the date the sighting was added to the dataset.
- Latitude (char) – the latitude of the sighting.
- Longitude (char) – the longitude of the sighting.

Objective of the project

The purpose of this project is to see if there is any correlated data between sightings that would add validity to any of the events. There were three main questions that I set out to answer through my data analysis:

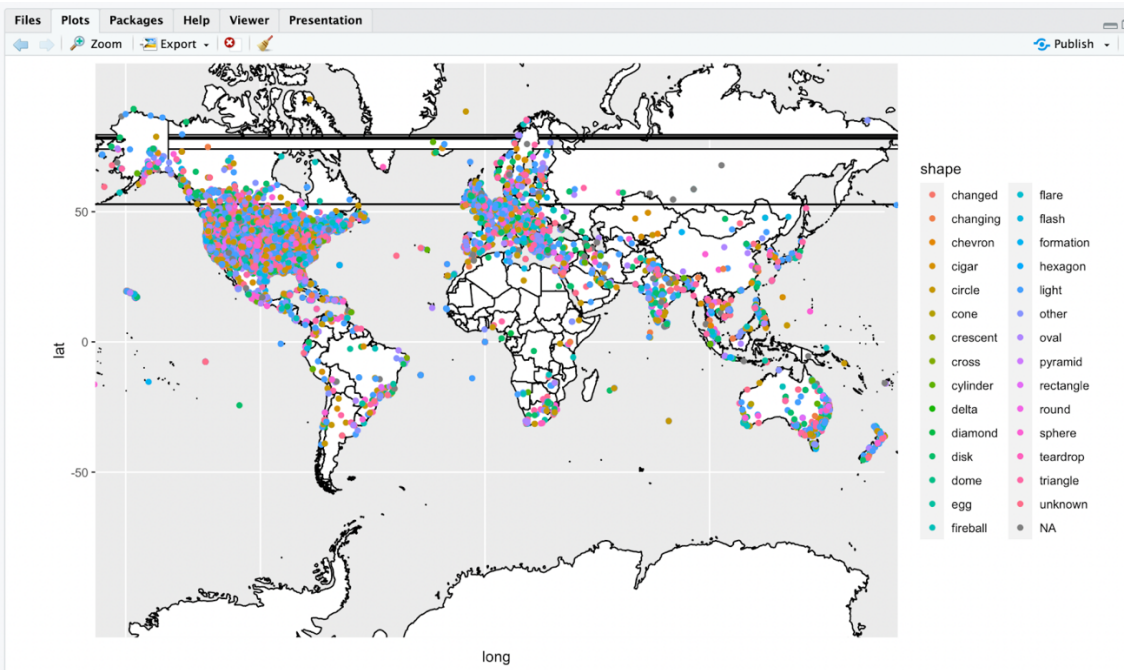
- What was the most common phenomenon reported?
- How many of these sightings occurred at the same time?
- Are there correlations between sightings?

By piecing together different accounts and looking at the events on a map, my goal is to understand better the number of instances, location, and correlations between sightings.

Data Analysis: Adding Context

Before I began attempting to answer the prior questions, I spent some time familiarizing myself with the data by doing some simple visualizations and looking for any patterns.

First, I plotted all sightings on a map with colors by “shape”.



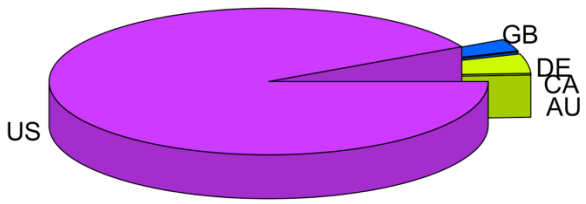

```
#plotting ufo sightings on a world map
world<-map_data("world")
ggplot() + geom_polygon(data=world, color="black", fill="White", aes(x=long, y=lat, group=group))
head(world) + coord_map()
ufo$longitude<-as.numeric(ufo$longitude)
ufo$latitude<-as.numeric(ufo$latitude)
head(ufo)
ggplot() + geom_polygon(data=world, color="black", fill="White", aes(x=long, y=lat, group=group))
+coord_map()+geom_point(data=ufo, aes(x=longitude, y=latitude, color=shape))
#based on the generated map, we can see that there are an overwhelming amount of sightings in the northern hemisphere.
```

I noticed here that most sightings are in the Northern Hemisphere. Otherwise, the map was quite busy and does not tell us much. I decided to instead create a pie chart of the locations of sightings to visualize this data in a different manner. After plotting the pie chart by country, I found that the large majority of sightings are in the US, followed by Great Britain, Denmark, Canada, and Australia.

```
perCountry
#There are 593 UFO sightings in Australia.
#There are 3266 UFO sightings in Canada.
#There are 112 UFO sightings in Denmark.
#There are 2050 UFO sightings in Great Britain.
#There are 70293 UFO sightings in the US.
# 3D Exploded Pie Chart
#install.packages("plotrix")
library(plotrix)
slices <- c(593,3266,112,2050,70293)
lbls <- c("AU", "CA", "DE", "GB", "US")
pie3D(slices,labels=lbls,explode=.1,main="UFO Sightings by Country ")

```

View a larger version of the plot in a new window **Sightings by Country**

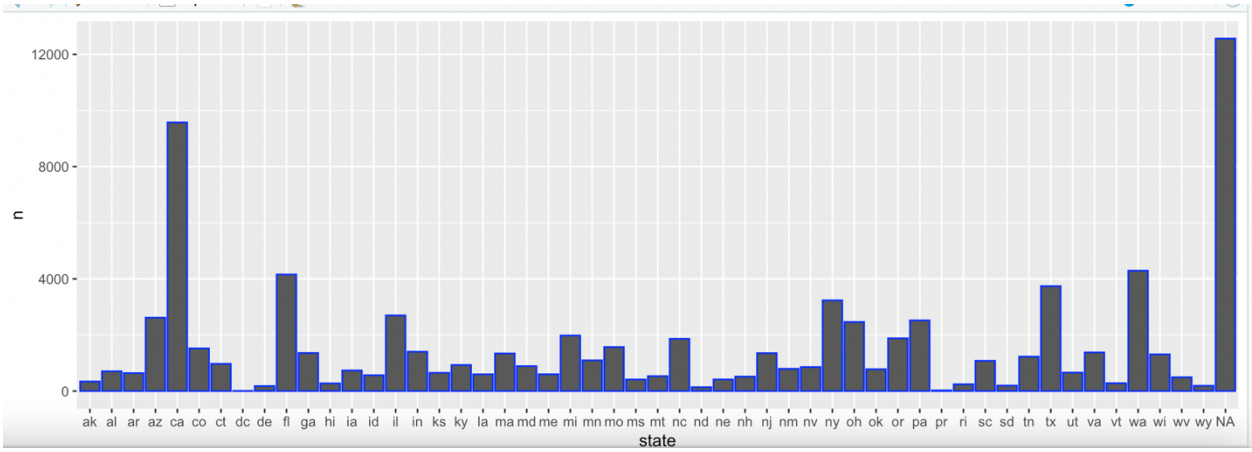


I also created one more bar chart to visualize the number of sightings per state in the US:

```
#Now counting how many observations there are per state in the us
us<-ufo[ufo$country=="us", ]

perState<-us%>%count(state)

perState
# state     n
#<chr> <int>
 # 1 ak      341
#2 al       706
#3 ar       642
#4 az      2617
#5 ca      9575
#6 co      1521
#7 ct       971
#8 dc         7
#9 de       180
#10 fl     4155
# ℹ 43 more rows
# ℹ Use `print(n = ...)` to see more rows
#creating bar chart of how many observations per state
perState%>%
  ggplot()+geom_bar(aes(x=state, y=n), stat="identity", color="blue")
```



The main takeaway from these “contextual” queries was that most sightings are clustered in the US, and that the most sightings were reported in California. It also helped me visualize different locations of various reported shapes, which is how I proceeded with my next queries.

Data Analysis: Most Common Phenomenon

To dive into our question of the most frequently seen object, I first organized the shapes column by count. The most seen shape was a “light”, followed by a triangle, and then a circle.

```
77  byShape<-ufo%>%count(shape)
78  byShape%>%arrange(desc(n))
79  #the most commonly seen shape for ufos is a light. I ar
79:1  # Project Overview
```

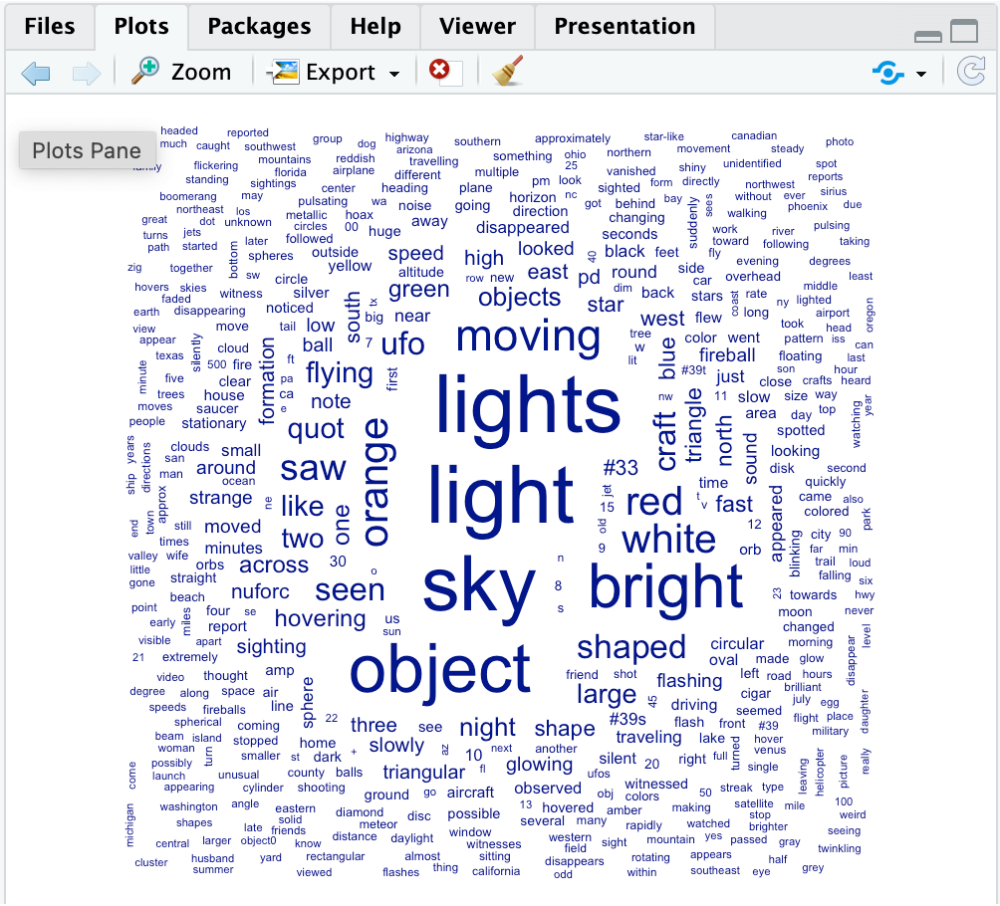
Console | Terminal | Render | Background Jobs

R 4.3.1 · ~/Downloads/archive (3)/

	shape	n
	<chr>	<int>
1	light	17872
2	triangle	8489
3	circle	8453
4	fireball	6562
5	unknown	6319
6	other	6247
7	disk	6005
8	sphere	5755
9	oval	4119
10	NA	3118

```
# ℹ 20 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Next, I wanted to dive into the comments column. I used text mining and a word cloud to find the most frequent words and descriptors in this column (I manually removed any numbers or odd characters that appeared in the word cloud.




```
50  #using text mining to find keywords from comments column
51  library(quanteda)
52  corpusufo<-corpus(ufo$comments, docnames=ufo$country)
53  toksufo<-tokens(corpusufo, remove_punct=TRUE)
54  toks_nostopufo<-tokens_select(toksufo, pattern = stopwords("en"), selection = "remove")
55  corpus_ufodfm<-dfm(toks_nostopufo)
56  corpus_ufodfm<-dfm_remove(corpus_ufodfm, pattern=c('#44', '1', '2', '3', '4', '5', '6'))
57  quanteda.textplots::textplot_wordcloud(corpus_ufodfm, min_count=5)
58
59
60
61
62
```

I was curious about the frequency of the words as well, so I used a textstat_frequency function to determine how often the words were repeated. I found that the most frequent words including “light”, “lights”, “bright”, and “moving” were repeated 1000s of times, showing strong correlations between different accounts.

```
63  library(quanteda.textstats)
64  freq<-textstat_frequency(corpus_ufodfm)
65  ufowordfreq<-as.matrix(freq[1:50,1:2])
66  ufowordfreq
```

71:1 # Project Overview

Console | Terminal | Render | Background Jobs

R 4.3.1 · ~/Downloads/archive (3)/

```
   feature     frequency
1  "light"     "19519"
2  "lights"    "19512"
3  "sky"       "18487"
4  "object"    "16063"
5  "bright"    "14382"
6  "moving"    " 9712"
7  "orange"    " 8768"
8  "white"     " 8468"
9  "red"       " 8323"
10 "saw"       " 7487"
11 "shaped"    " 6518"
```

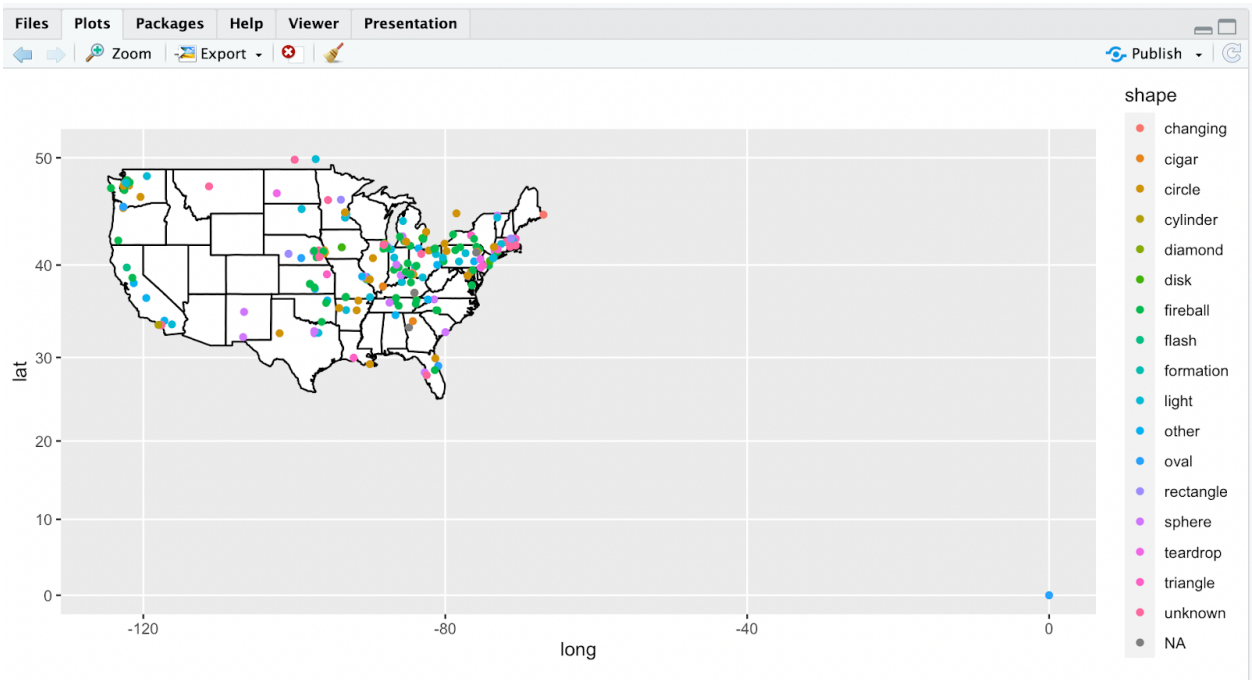
Data Analysis: Comparing different sightings

I wanted to now look at the datetime column to find certain days with many different reports in a certain location. To do this, I first needed to extract the date from the time and create new columns.

```
90  #how many sightings occurred at the same time in the same city
91  head(ufo)
92  datetime<-as.data.frame(str_split_fixed(ufo$datetime, ' ', 2))
93  datetime<-datetime[, 1]
94  datetime
95  ufo$date<-datetime
96  ufo$date
97  ufo%>%count(date)%>%arrange(desc(n))
98  #most commonly seen phenomenons occurred on 7/4/2010, 11/16/1999, and 7/4/2012.
```

I found that the most frequent sightings were on 7/4/2010, 11/16/1999, and 7/4/2012 (208, 195, and 192; respectively). I subsetted and then dove into sighting 1 (2010) and sighting 2 (1999) to compare the events.

First, I plotted sighting 1 using ggplot2.



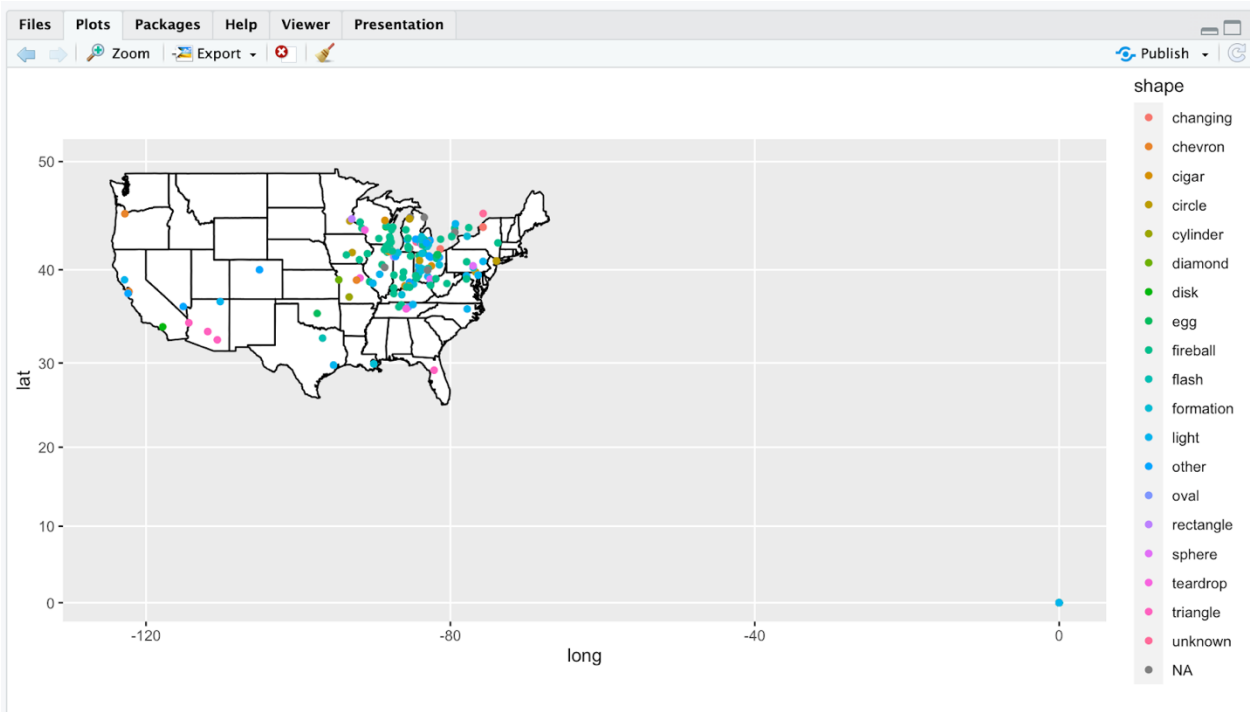
I was surprised how spread out these instances were, though they did seem to be clustered around the Midwest and East Coast. However, the most sightings were in Washington (as seen below) while the next states were Pennsylvania and Ohio for largest numbers of sightings.

```
> sighting1%>%count(state)%>%arrange(desc(n))
# A tibble: 41 × 2
   state     n
   <chr> <int>
 1 wa       23
 2 pa       15
 3 oh       13
 4 il       12
 5 ne       11
 6 ca       10
 7 mi       10
 8 mo       10
```

Finally, I was curious about the type of shape reported for this sighting. When sorting by shape, I found that most people reported a fireball, followed by a circle and light.

```
> sighting1%>%count(shape)%>%arrange(desc(n))
# A tibble: 18 × 2
   shape         n
   <chr>     <int>
 1 fireball     55
 2 circle       32
 3 light        31
 4 sphere       20
 5 other        13
 6 unknown      13
 7 oval          8
 8 triangle      8
```

I then proceeded to run the same queries for Sighting 2 (1999).



I found that this sighting was also focused on the Midwest, but much more clustered there. The majority of the sightings were in Ohio, Minnesota, and Illinois.

```
> sighting2%>%count(state)%>%arrange(desc(n))
# A tibble: 28 × 2
   state     n
   <chr> <int>
 1 oh       40
 2 mi       33
 3 il       19
 4 in       15
 5 wi       13
 6 ky       11
 7 mo        9
 8 on        8
```

There was also much more consistency among the descriptions reported. 81 people reported a fireball shape, while 23 reported a light, and 19 a formation.

```
> sighting2%>%count(shape)%>%arrange(desc(n))
# A tibble: 20 × 2
   shape         n
   <chr>     <int>
 1 fireball     81
 2 light        23
 3 formation    19
 4 other        12
 5 circle       10
 6 NA            7
 7 cigar         6
 8 triangle      6
```

Conclusion

Ultimately, the correlations between the two sightings that I investigated (centralized around a certain location, ~200 reports on each day, and common shapes/descriptions from multiple people) provides evidence for the possibility of a true UFO event. Sighting #2 (1999) was even more clustered around a centralized location that Sighting #1 (2010) and had more common descriptions that adds validity to this event. One thing that should also be noted is that many people reported a “changing” shape as well. We will never know for sure if these sightings were true, but for now we have enough data to support the possibility of UFOs existing.