Exploratory Analysis of UFO Sightings from NUFORC Dataset

Shreyansh Misra, Lawrence Meng

2024-12-2

The NUFORC Databank is the largest independently collected set of unidentified flying object (UFO) and unidentified aerial phenomenon (UAP) sighting reports available on the internet. This investigation aimed to uncover trends and patterns in UFO sightings by addressing three primary questions:

1. **Where are UFOs most likely to be sighted?** Are they concentrated in specific countries, near landmarks, or certain distances from the equator?

2. **When are UFOs most likely to be sighted?** Are sightings tied to specific seasons, holidays, or days of the week?

3. **What are the most common UFO descriptions?** What shapes, patterns, and accounts are commonly reported?

1 Data Preperation

1.1 Overview

```
ufo.data <- read.csv("scrubbed.csv")
```

The dataset contains 80,332 records of UFO sightings, with variables detailing sighting locations, times, and descriptions.

1.1.1 Variables

- **Datetime**: When the sighting occurred.
- **City/State/Country**: The geographical location of the sighting.
- **Shape**: Reported shape of the UFO.
- **Duration (seconds)**: Length of the sighting.
- **Latitude/Longitude**: Geographic coordinates of the sighting.
- **Comments**: Eyewitness accounts.

1.1.2 Variables of Interest

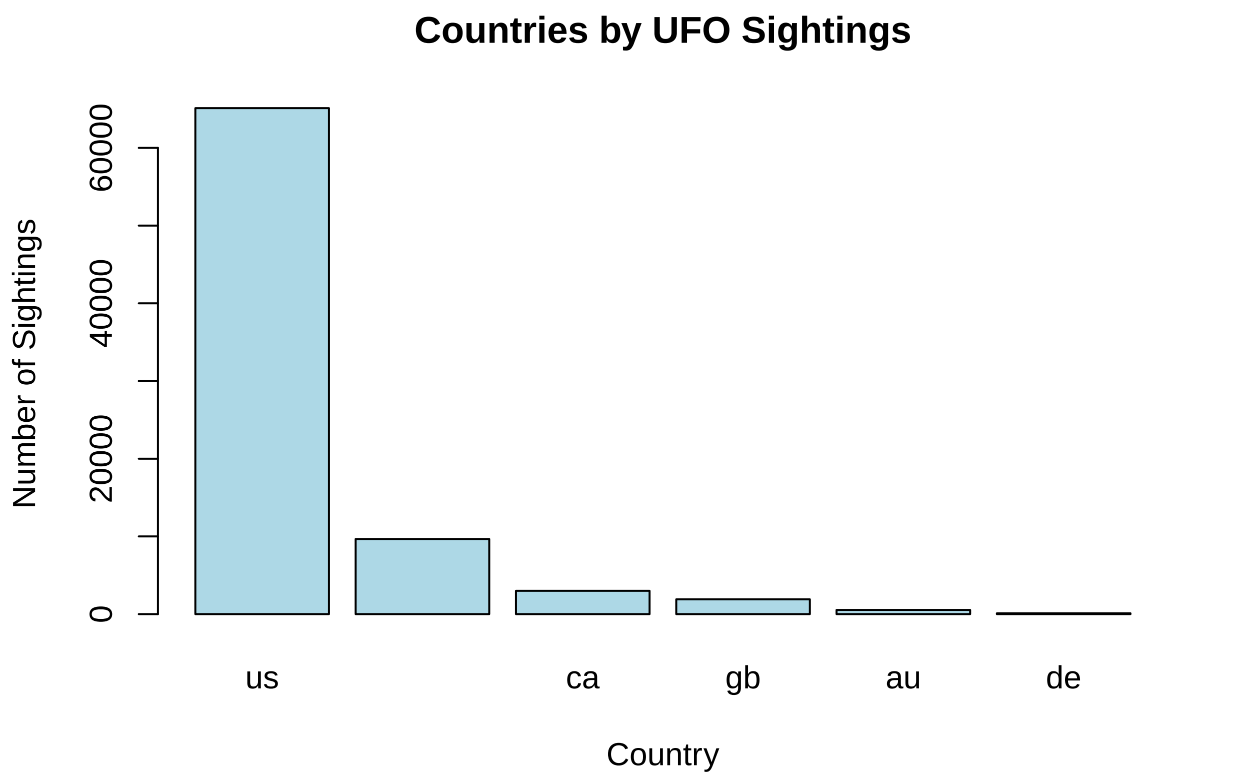
- **Datetime**
- **City/State/Country**
- **Latitude/Longitude**
- **Shape**

Analyzing these variables will help us answer the questions: When do UFO sightings take place, where are they most frequency, and what are the most common descriptions?

1.1.3 Analyzing Location

```
country.freq <- table(ufo.data$country)
country.freq <- sort(country.freq, decreasing = TRUE)

barplot(country.freq,
        main = "Countries by UFO Sightings",
        col = "lightblue",
        xlab = "Country",
        ylab = "Number of Sightings")
```



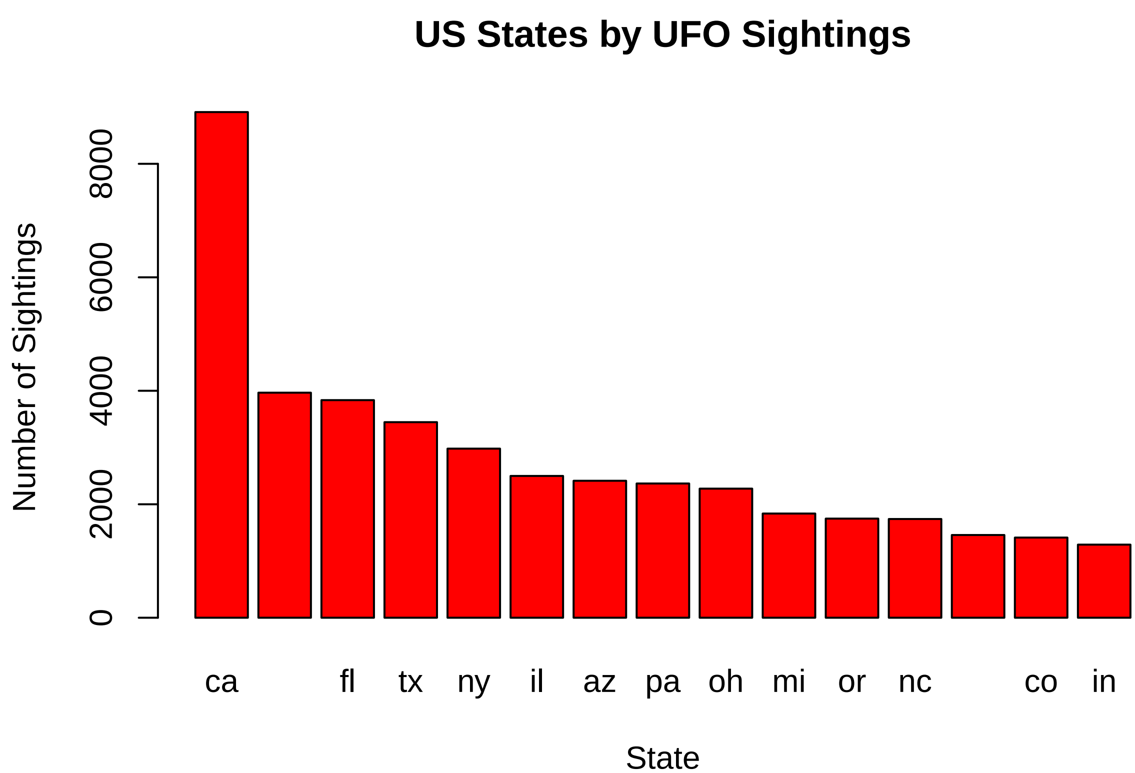
The United States overwhelmingly has the highest number of UFO Sightings reported, followed by Canada, Great Britain, Austrailia, and Germany. There are also a large number of reports where the country is left blank. The NUFORC being based in the United States is likely why most of their reports are from the United States. Since the sightings from outside the US are too few to come to reasonable conclusions with, we will only use sightings based in the United States for this investigation.

```
# Remove non-US sightings from dataset
ufo.data <- ufo.data[ufo.data$country == "us", ]

state.freq <- table(ufo.data$state)
state.freq <- state.freq[names(state.freq) != "Unknown"]
```

```r
state.freq <- sort(state.freq, decreasing = TRUE)[1:15]

barplot(state.freq,
        main = "US States by UFO Sightings",
        col = "red",
        xlab = "State",
        ylab = "Number of Sightings")
```



California had the most UFO sightings reported, followed by Washington, Flordia, Texas, and New York. It is significant to note that California, Texas, Florida, and New York are the four most populous states in that order.

1.1.4 Analyzing Descriptions

There were 29 different "shapes" that people reported that the UFO looked like. While they are labelled as "shapes" in the dataset, a more accurate way to define the data column would be "descriptions" of UFOs. This is because some data points include "fireball", "light", "flash", and "flare", which are good descriptions of what the UFO would have looked like but are not explicitly shapes.

```r
ufo.data$shape <- factor(ufo.data$shape, levels = unique(ufo.data$shape))
levels(ufo.data$shape)
```

```
##  [1] "cylinder"  "circle"    "light"     "sphere"    "disk"      "fireball"
##  [7] "unknown"   "oval"      "other"     "rectangle" "chevron"   "formation"
## [13] "triangle"  "cigar"     ""          "delta"     "changing"  "diamond"
```

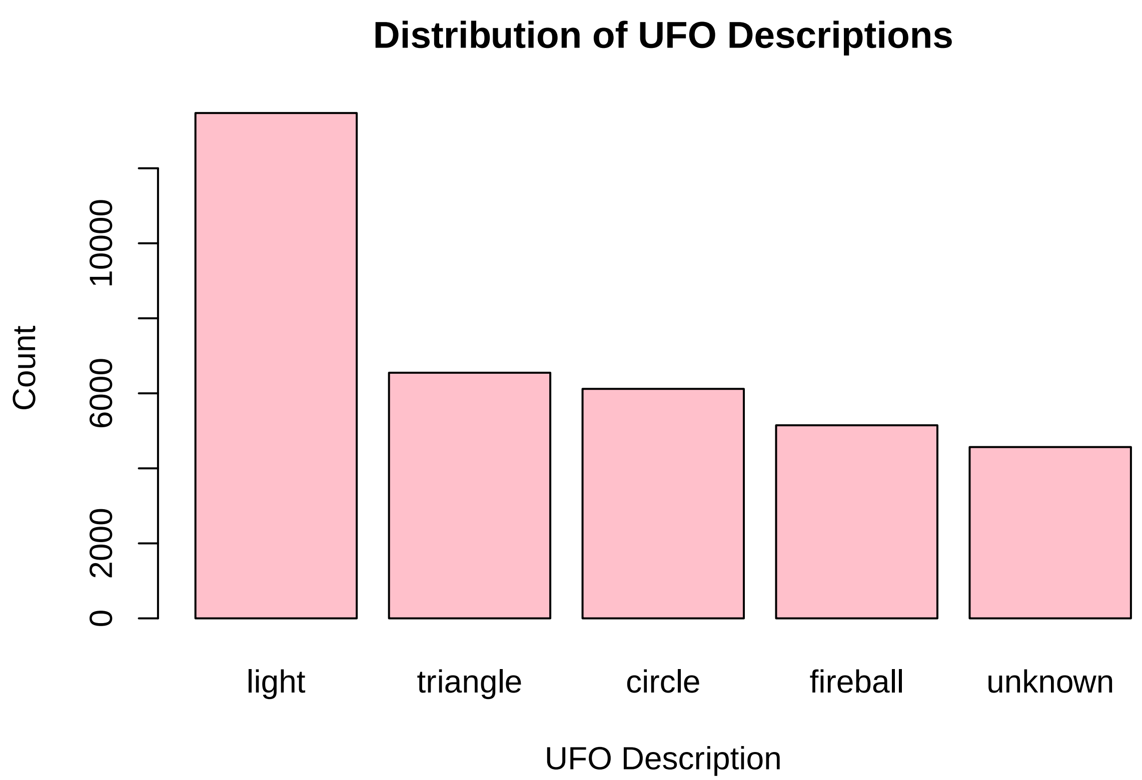
```
## [19] "flash"     "egg"       "teardrop"  "cone"      "cross"     "pyramid"
## [25] "round"     "flare"     "hexagon"   "crescent"  "changed"
```

```r
# Remove n/a values for barplot
ufo.data$shape[ufo.data$shape == ""] <- "Unknown"
```

```
## Warning in `[<-.factor`(`*tmp*`, ufo.data$shape == "", value = structure(c(1L,
## : invalid factor level, NA generated
```

```r
ufo.data$shape <- factor(ufo.data$shape)
shape.freq <- table(ufo.data$shape)
shape.freq <- sort(shape.freq, decreasing = TRUE)[1:5]

barplot(
  shape.freq,
  col = "pink",
  main = "Distribution of UFO Descriptions",
  xlab = "UFO Description",
  ylab = "Count"
)
```



The top five "descriptions" of UFOs were light, triangle, circle, fireball, and unkown respectively.

1.1.5 Analyzing Time

The dataset includes historical reports dating back to 1910. For the context of this investigation, we will be eliminating reports collected before 1990.

```r
# Initial Range
ufo.data$datetime <- as.Date(ufo.data$datetime, format = "%m/%d/%Y")
range(ufo.data$datetime)
```

```
## [1] "1910-01-01" "2014-05-08"
```

```r
# Filtered Range
ufo.data <- ufo.data %>%
  filter(datetime >= as.Date("1990-01-01") & datetime <= as.Date("2014-05-01"))
range(ufo.data$datetime)
```

```
## [1] "1990-01-03" "2014-05-01"
```

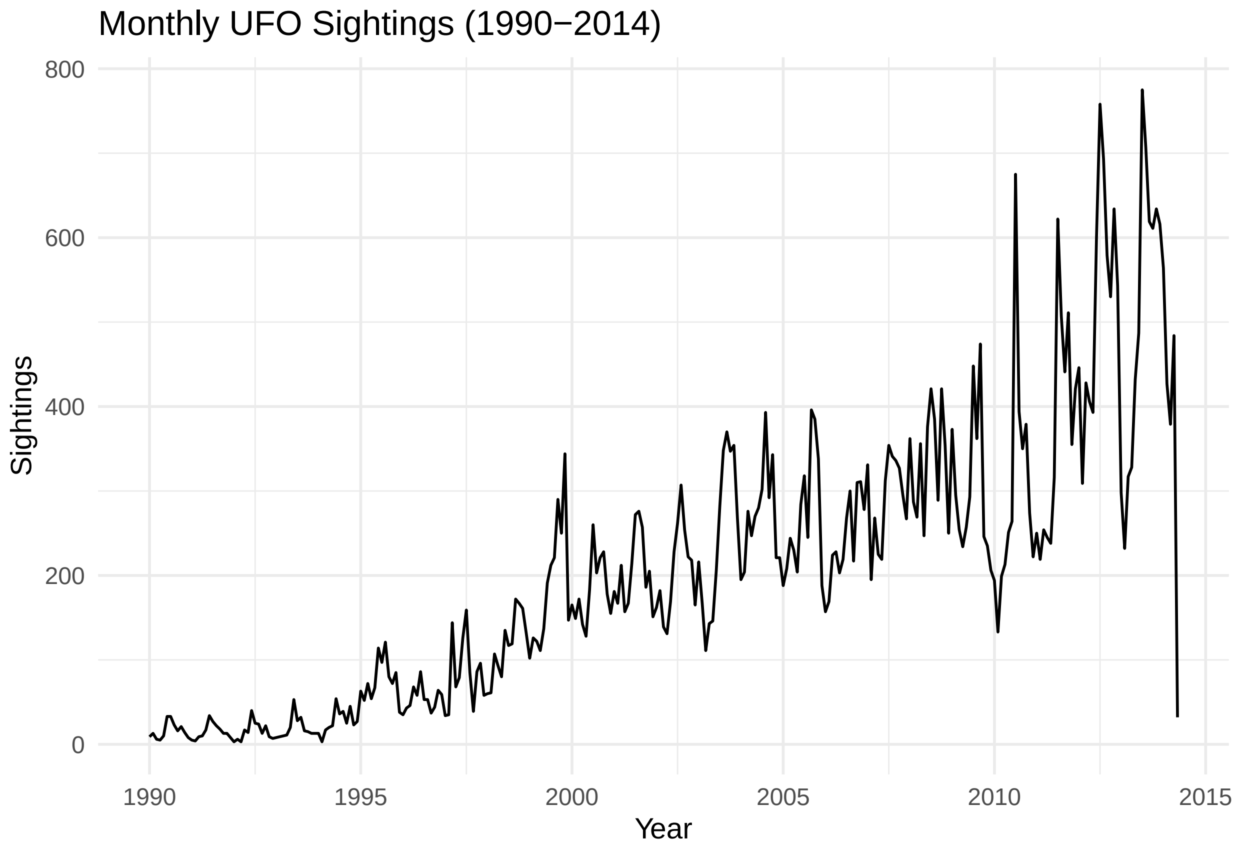
```r
# Creating a year-month column
ufo.data$year_month <- format(ufo.data$datetime, "%Y-%m")

# Aggregating sightings by year-month
ufo.monthly <- ufo.data %>%
  group_by(year_month) %>%
  summarize(sightings = n())

# Converting year-month to Date type for plotting
ufo.monthly$date <- as.Date(paste0(ufo.monthly$year_month, "-01"))
```

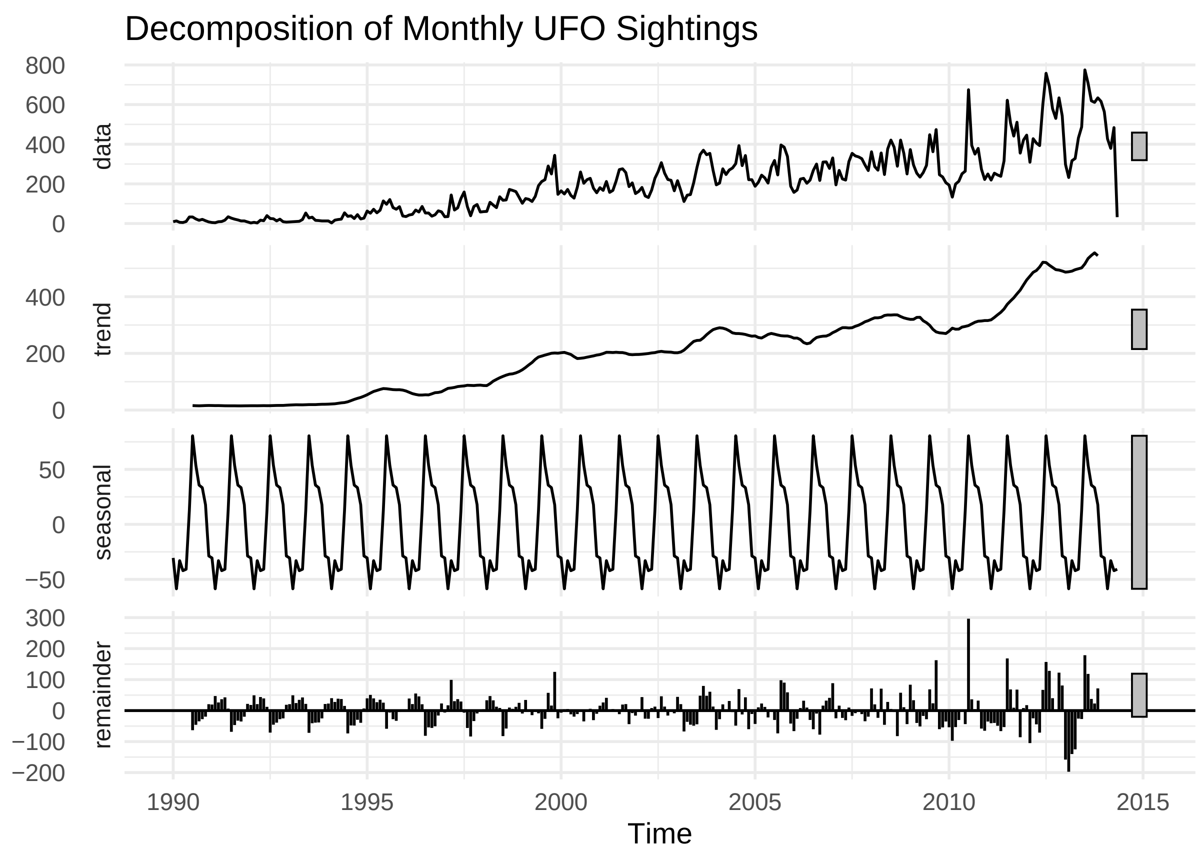
```r
# Create a time series object
ufo.ts <- ts(
  ufo.monthly$sightings,
  frequency = 12, # Monthly frequency
  start = c(as.numeric(format(min(ufo.monthly$date), "%Y")),
            as.numeric(format(min(ufo.monthly$date), "%m")))
)

# Plotting the time series
autoplot(ufo.ts) +
  labs(title = "Monthly UFO Sightings (1990-2014)", x = "Year", y = "Sightings") +
  theme_minimal()
```




```r
# Decomposing the time series
ufo.decomp <- decompose(ufo.ts)

# Plotting the decomposed time series
autoplot(ufo.decomp) +
  labs(title = "Decomposition of Monthly UFO Sightings") +
  theme_minimal()
```




```r
# Augmented Dickey-Fuller Test
adf.test(ufo.ts)
```

```
## Warning in adf.test(ufo.ts): p-value smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  ufo.ts
## Dickey-Fuller = -7.0054, Lag order = 6, p-value = 0.01
## alternative hypothesis: stationary
```

```r
# KPSS Test
kpss.test(ufo.ts)
```

```
## Warning in kpss.test(ufo.ts): p-value smaller than printed p-value
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  ufo.ts
## KPSS Level = 4.1877, Truncation lag parameter = 5, p-value = 0.01
```

For the ADF test, the p-value is very small, so the null hypothesis is rejected, and the time series is stationary.

For the KPSS test, the p-value is very small, so the null hypothesis is rejected, and the time series is not stationary.

These two are contradicting, so we apply differencing.

```r
# Differencing the time series
ndiffs(ufo.ts)
```

```
## [1] 1
```

```r
ufo.ts.diff <- diff(ufo.ts, differences = 1)

# ADF test
adf.test(ufo.ts.diff)
```

```
## Warning in adf.test(ufo.ts.diff): p-value smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  ufo.ts.diff
## Dickey-Fuller = -10.44, Lag order = 6, p-value = 0.01
## alternative hypothesis: stationary
```

```r
# KPSS test
kpss.test(ufo.ts.diff)
```

```
## Warning in kpss.test(ufo.ts.diff): p-value greater than printed p-value
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  ufo.ts.diff
## KPSS Level = 0.059493, Truncation lag parameter = 5, p-value = 0.1
```

Now, the time series is stationary.

1.2 References

- https://nuforc.org/databank/
- https://www.geeksforgeeks.org/how-to-sort-a-dataframe-by-date-in-r/