Assignment: Exploratory Analysis of NUFORC UFO Sightings

Overview

You've been asked to brief a curious journalist: Where do different kinds of UFO sightings concentrate, and what words do people use when they describe them? The National UFO Reporting Center (NUFORC) has collected and published first-hand UFO/UAP sighting reports since 1974. Each record typically contains the event date and time, geographic location, a witness-reported "shape," and a free-text narrative. For this assignment, you will use the CSV file `nuforc_sightings.csv`. Your task is to transform these reports into forms suitable for quantitative analysis. You will apply **principal component analysis (PCA)** to uncover dominant patterns and interpret what the components mean. In the first part of the assignment, you will focus on reported shapes across U.S. states. In the second part, you will turn the free-text summaries into keyword features and repeat the PCA.

At each step, you should not only produce the requested plots and tables but also provide short written interpretations of what you observe.

What to Turn In

1. Reproducible source code: a well-commented **R script plus** a short `README.md` explaining exactly how to run it to reproduce all outputs.
2. A rendered document in **one** of: **HTML** (`.html`), **PDF** (`.pdf`), or **Word** (`.docx` , "doc").
3. Include a **Dockerfile** that installs R and all packages you used so the grader can build and run your analysis in a clean environment.

Task 1 — Build the shape table

Begin by loading the `nuforc_sightings.csv` data. Your analysis will focus on sightings within the United States. Filter the dataset to include only reports from the USA, using the country and state columns to identify these records. From this U.S-only subset, construct a data frame that quantifies the number of sightings for each reported UFO Shape. Clean and standardize the shape column, addressing any missing, blank, or inconsistently formatted entries in a reasoned manner. The goal is to create a matrix where each row corresponds to a U.S. state, and each column represents a distinct UFO shape category. The cell values should indicate the frequency of each shape's sightings for each state.

How many different known shapes are in the dataset (excluding "Other" or "Unknown")? Which state has the most sightings of the "Circle" shape?

Task 2 — PCA on the shape table

With your state-by-shape count matrix prepared, perform a Principal Component Analysis (PCA) to explore the main patterns in the data.

Before running PCA, convert your counts to proportions per state (row-normalize) so that the analysis reflects differences in the distribution of shapes rather than just overall sighting volume.

- Can the variety of UFO sightings be boiled down to a few key patterns? Make a scree plot showing the proportion of variance explained by each principal component. What do you find?
- Plot the first two principal components as a scatterplot, with each point representing a state. States that are close together have more similar UFO shape distributions. Interpret whether you see regional clusters or outliers.
- Examine the first two columns of the PCA rotation (the loadings matrix) and explain which shapes contribute the most to PC1 and PC2.

Task 3 — Clean and tokenize the summaries

For Summary:

- Convert to lowercase.
- Remove non-ASCII characters: Use a regular expression to find and replace any characters that are not standard ASCII text with an empty string (“”).
- Trim whitespace from the edges and use a regular expression to replace any repeated internal whitespace with a single space.
- Break each summary into an array of words.
- Generate a histogram or a table showing the most frequent words across the entire dataset.
- (Optional) You may also create a word cloud as an alternative or supplementary visualization.

What do you observe in this initial output? Do the most frequent words give you any immediate insight, or are they mostly common, generic English words?

- Install the stopwords package and remove stopwords from your tokens. Re-make the histogram and comment on changes. (Include stopwords in your Dockerfile.)

After stopword removal, which words feel most characteristic of these reports?

Task 4 — Build the keyword table and repeat PCA

- Define a vocabulary of the top 100 most frequent words from your cleaned summaries (from Task 3), ensuring each word is at least 3 characters long.
- Create a wide table where each row is a state and each column is a word from your vocabulary. The values should be the count of each word for that state.
- Repeat the core analysis from Task 2 on this new state-by-keyword table:
- Compare this analysis with the shape-based PCA from Task 2. Do the top words that drive the clusters align with the top shapes that drove the clusters in your first PCA? Discuss the similarities and differences.