Pre-Analysis Plan for Global UFO Sightings

Avery, Katarina, Zoe

To what extent can UFO sighting and its characteristics be predicted/modeled?
What is an observation in your study?

The dataset we are looking at is a collection of self-reported UFO sightings. Each observation is a UFO sighting record. It lists the date and time of the sighting, the location, the shape and duration of the sighting, and comments as well.

How are we cleaning the data?

The data was relatively clean when we received it. Initially, we removed NaN values and ensured consistent capitalization. Next, we structured the data for analysis by splitting the date into separate day and year columns. We also had to convert some integer values, such as duration, to numeric format to enable numerical analysis. We also plan to include a month column to facilitate modeling, as we cannot one-hot encode datetime values for use as explanatory variables. Additionally, some countries in our dataset do not have states, as they are not part of the United States. To maintain consistency, we will impute these missing values by replacing NaN with "non-US" in the state column. This variable was overlooked during EDA, but we will clean it before building our models.

Are you doing supervised or unsupervised learning? Classification or regression?

We are conducting supervised learning since our data is clearly labeled, and we aim to analyze which variables can predict sightings and their characteristics. For this reason, our project will primarily focus on regression. By building regression models using different variables, we can identify significant predictors, quantify their impact, and improve our understanding of the factors influencing sightings.

What models or algorithms do you plan to use in your analysis? How?

We plan to use linear regression, logistic regression, and decision trees in our analysis to examine the ability of various factors to predict UFO sightings. Our predictor variables include location (country, state, city), time of year, shape, and duration.

For logistic regression, we will assess the credibility of a UFO sighting based on duration, country, month, and shape. Credibility will be determined using a newly created binary variable, multiple sightings (yes/no). If a sighting has been reported multiple times, it is considered more credible. Thus, we aim to predict multiple sightings using these explanatory variables. Additionally, we will use logistic regression to predict the likelihood of a UFO sighting occurring based on categorical inputs such as location, month, and shape.

For **linear regression**, we will predict the duration of UFO sightings using country, month, and UFO shape. Our exploratory data analysis (EDA) revealed significant variation in the kernel density plots of log-transformed duration across different countries, suggesting potential patterns. If we obtain a high R^2, it will indicate a strong correlation between these factors and sighting duration.

However, linear regression assumes a linear relationship between predictors and the outcome. Given the variability, subjectivity, and potential fabrication of UFO sightings, we will also apply a **decision tree model**. Decision trees are useful for capturing complex, non-linear relationships such as differences in sighting duration across countries that other models might miss.

How will you know if your approach "works"? What does success mean?

If our regression models yield accurate predictors (high R^2 scores and low RMSEs), we will have found an approach that “works.” The goal of our project is to determine whether UFO sightings can be predicted and understood, which could suggest either a pattern in UFO appearances or, more likely, the influence of a third variable; such as sky events (e.g., fireworks) or a rise in UFO conspiracy theories. Additionally, we hope our results provide insight into the credibility of these reports. If no model proves effective, it may suggest that UFO sightings are entirely random or simply made up.

What are weaknesses that you anticipate being an issue? How will you deal with them if they come up? If your approach fails, what might you learn from this unfortunate outcome?

The nature of our question highlights a key weakness in our dataset: we don’t know if UFO sightings follow any detectable patterns in the first place. We expect our approach to fail, but this is an important insight. Our search for patterns will involve trial and error as we test different combinations of variables and models, and in the process, we may uncover something unexpected.

For example, we might find that sightings are more frequent in certain geographic regions, suggesting environmental or cultural influences. Alternatively, we could discover that specific time periods, such as summer months or holidays, correlate with increased reports, potentially linking UFO sightings to human activity rather than extraterrestrial presence. We may also identify trends that align with historical events, such as spikes in sightings following major UFO related media coverage. Even if no strong predictive model emerges, understanding the randomness (or lack thereof) in UFO reports can still provide valuable insights into the social and psychological factors driving these claims.