Prediction Question

Can we predict UFO sightings (where they will happen, how long they will be, where they will happen) based on observed characteristics? Is there consistency amongst responses and can we find any meaningful relationships?

Results

Linear Regression

We tried multiple types of linear regression in order to look for potential patterns. As we were using linear regression, the numerical value we attempted to predict was duration of sighting. We wanted to see if there were any variables that could be used as a predictor for how long a UFO would be visible to people. After trying a few different variables, we were getting extremely low r values. The highest r-value we received would come from using the month of the year that the sighting took place. I want to note, when I say highest r-squared, it was still very low. However, this r-squared value was about 100 times the size of previous run models, which led us to want to take a second look. It was this increase in r-squared that led us to taking a closer look into our date columns.

```
y = df['duration (seconds)']
X_d = pd.get_dummies(df['month'],dtype='int')
reg = LinearRegression(fit_intercept=False).fit(X_d, y)
results = pd.DataFrame({'variable':reg.feature_names_in_, 'coefficient': reg.coef_})
print('R-squared: ', reg.score(X_d, y))
results
```

```
R-squared:  0.07122543508119494
```

```
date
2010-07-04    165
2012-07-04    158
1999-11-16    147
2013-07-04    142
2011-07-04    124
2009-09-19     84
2014-01-01     81
2013-12-31     74
2004-10-31     72
2009-07-04     69
2013-07-06     65
2011-07-03     58
```

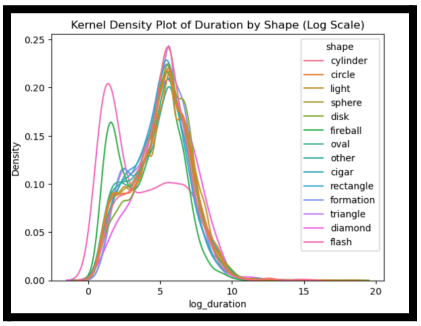
While running a linear regression using day of the year yielded another significantly low r-squared, we did notice something interesting about the dates category. By sorting the dates by

number of sightings on that date, we noticed a similarity in what days people were seeing UFOs. Of the top 15 days where the most UFOs were sighted, 11 were on holidays, 5 of which were 4th of July. This may indicate that there could be another unknown factor causing these sightings, like fireworks for example.

Decision Tree & Random Forest

The decision tree model showed that predictability is generally low. Variables such as year, duration, month, year, shape, latitude, longitude, and country often don't have the predictive power to explain each other. We used a decision tree regressor and classifier to find a more sensitive model and capture underlying relationships, since we were unable to discover anything with linear and logistic regression.

Using a decision tree classifier, we attempted to predict the observed shape of the UFO. We ended up classifying shapes on latitude, longitude, month, and year. The latitude and longitude were meant to account for location, since in our heatmap we saw that there may be a relationship between shape and country. Initially, the classifier was performing under 3% accuracy, which was worse than random.



Initial exploratory data analysis (EDA) showed that duration didn't vary across shapes, except for the fireball and flash shapes. This would make sense, for the nature of these observed forms implies the duration of no more than a moment. This meant that excluding duration (seconds), even in its log-transformed form, was more beneficial in trying to model meaningful relationships. By recognizing this, and trying to make the data less noisy, we got it up to 7.2% accuracy.

```
min_samples_leaf = 20, Accuracy = 0.0716
min_samples_leaf = 21, Accuracy = 0.0716
min_samples_leaf = 22, Accuracy = 0.0716
min_samples_leaf = 23, Accuracy = 0.0717
min_samples_leaf = 24, Accuracy = 0.0717
min_samples_leaf = 25, Accuracy = 0.0716
min_samples_leaf = 26, Accuracy = 0.0718
min_samples_leaf = 27, Accuracy = 0.0718
min_samples_leaf = 28, Accuracy = 0.0718
min_samples_leaf = 29, Accuracy = 0.0718
min_samples_leaf = 30, Accuracy = 0.0718
min_samples_leaf = 31, Accuracy = 0.0718
min_samples_leaf = 32, Accuracy = 0.0718
min_samples_leaf = 33, Accuracy = 0.0720
min_samples_leaf = 34, Accuracy = 0.0720
min_samples_leaf = 35, Accuracy = 0.0720
min_samples_leaf = 36, Accuracy = 0.0720
```

However, iterating across different numbers of min_sample leaves and adjusting the depth didn't change the performance of the model much. This means that the model is underfitting and unable to find any meaningful relationships in the data. We proceeded to create a more complicated and powerful model — a random forest — and we witnessed no improvement in the accuracy of results, which

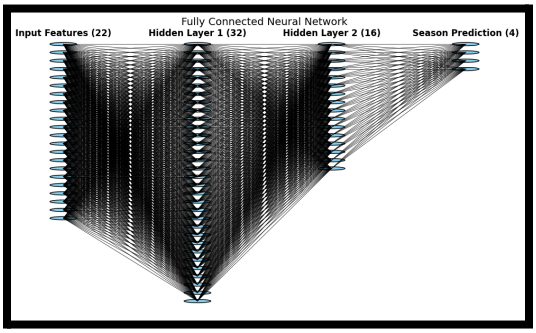
were just above random. This shows that the observations are widely inconsistent, and there exists little to no learnable pattern amongst them.

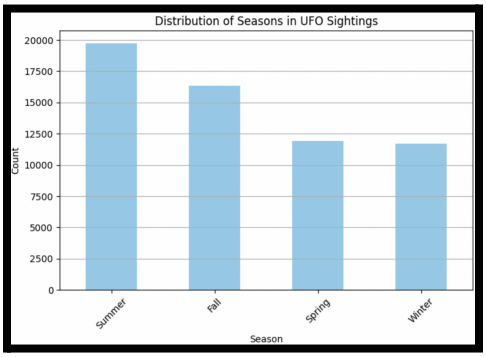
Neural Network Model

The model's performance was poor, achieving an overall validation accuracy of approximately 32%, slightly better than random guessing, and comparable to always predicting the most frequent season (~33%). Precision, recall, and F1-scores were especially low for Fall, Spring, and Winter. While the model recalled about 80% of Summer sightings, it frequently mislabeled Spring, Fall, and Winter as Summer, demonstrating strong bias toward the majority class. Overall macro-averaged precision, recall, and F1-score were around 22–27%, reflecting generally weak classification across seasons.

Season (Class)	Precision	Recall	F1-Score
Fall	25%	20%	22%
Spring	15%	10%	12%
Summer	35%	80%	49%
Winter	15%	10%	12%

Several factors likely contributed to the poor results. The model heavily favored Summer due to class imbalance (shown in the graph) and showed little ability to differentiate the other seasons. Additionally, the input features: location, duration, country, and shape, had little direct relationship with the season of a sighting, making the task inherently difficult. The lack of explicit information like the actual date or month limited predictive power.





In conclusion, the model struggled to accurately predict the season of UFO sightings because the explanatory variables used, such as location, duration, shape, and country, lacked strong predictive power. Although the dataset included the exact sighting date, we chose not to use it, as it would have directly indicated the season and introduced multicollinearity. This outcome highlights that, without direct seasonal indicators, the remaining features provide limited signal for distinguishing between seasons. As a result, the classification task remains

difficult, likely due to both data noise and weak associations between the chosen variables and the target.