Solar and Lunar Eclipse Milestone Report

Proposal

Problem Definition

Study the frequency of occurrence of eclipses and determine which eclipses are more common. Study the correlation between time of year and occurrence of eclipses.

Client Base and Motivation

Astronomers and astronomy-lovers will use this information to plan trips near those months that eclipses are more common. This information can also be used by astronomers to statistically confirm their theories about the occurrences of eclipses.

What data are you using? How will you acquire the data?

The data is contained in 2 csv files and was obtained from the NASA public dataset domain.

Preliminary Solution Outline

- Study distributions of months when eclipses occur.
- Analyze the data visually and statistically.
- Explain why the data shows the results found.

Deliverables

- The project will be delivered via code, this milestone report, and a slideshow.

Data Wrangling

The data was obtained from Kaggle. https://www.kaggle.com/nasa/solar-eclipses

Cleaning Steps

Steps were taken to clean two csv files: solar.csv and lunar.csv
Solar.csv - This file includes data about solar eclipses.
Lunar.csv- This file includes data about lunar eclipses.

1. Imported csv files as pandas dataframes and reindexed the rows with the column 'Catalog number' for both of the files.

2. After looking at the format of the dataframe, I created separate dataframes to include only the necessary columns for analysis.
    a. Solar now has the columns:
        i. 'Calendar Date', 'Eclipse Time', 'Delta T (s)', 'Eclipse Type','Eclipse Magnitude', 'Latitude', 'Longitude', 'Sun Altitude', 'Sun Azimuth', 'Path Width (km)', 'Central Duration'
    b. Lunar now has the columns:
        i. 'Calendar Date', 'Eclipse Time', 'Delta T (s)','Eclipse Type', 'Latitude', 'Longitude', 'Penumbral Eclipse Duration (m)', 'Partial Eclipse Duration (m)','Total Eclipse Duration (m)'

3. The data has also been reorganized by Calendar Date in ascending order.
    a. Converted the dates into a list of [year, month, day] for easier analysis of trends.
    b. Note: The years have negative values because the data is during the five millennium period -1999 to +3000 (2000 BCE to 3000 CE).

4. Missing Values
    a. Missing values appearing as '-' were replaced with NAN values when imported as dataframes.

5. No outlier data found

Data Storytelling

Questions Asked:

1. Which eclipse is more common?
    a. How many solar/lunar eclipses occur on average every year?
2. Is there a monthly trend for when solar/lunar eclipses occur?
    a. Is it different for solar vs lunar eclipses?
    b. If so why are they more/less common during certain months?
3. What months are total eclipses more common?
    a. Does it differ for solar/lunar eclipses?

1. Which eclipse is more common? How many solar/lunar eclipses per year on average?

There seem to be 166 more lunar eclipses than solar eclipses. After further investigation, it was found that on average solar and lunar eclipses had around 2 eclipses a year. There was no significant difference with solar eclipses (~2.3796/yr) and lunar eclipses (~2.4128/yr). But since this is over a 5 millennium period, the small difference can be seen affecting the total number of eclipses, that is that there have been more lunar than solar eclipses over this time frame 2000 BCE to 3000 CE.

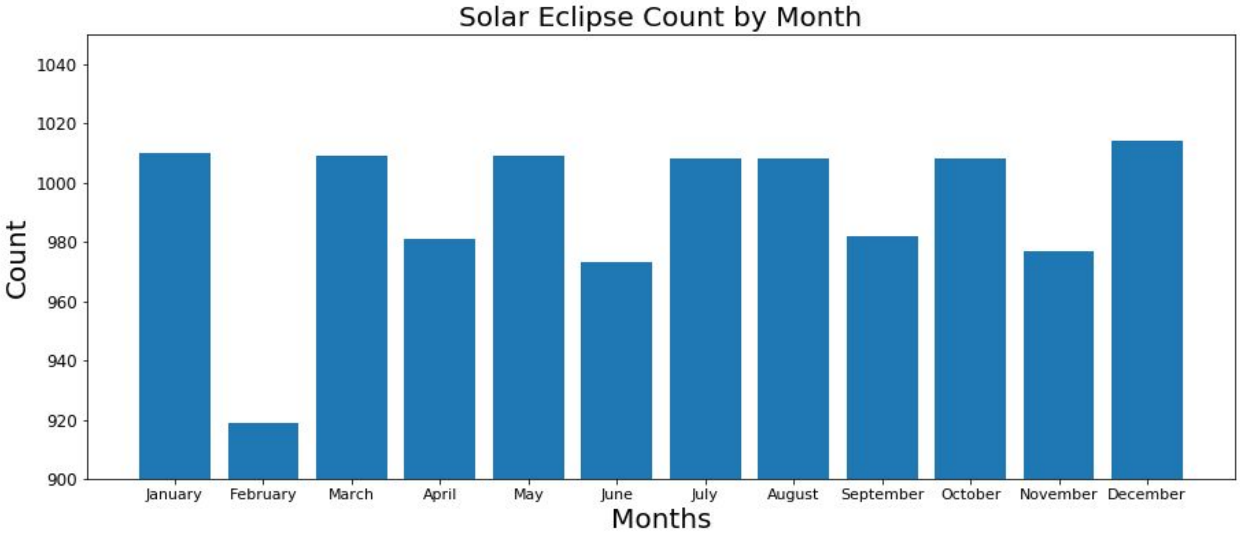
2. Is there a monthly trend for when solar/lunar eclipses occur?

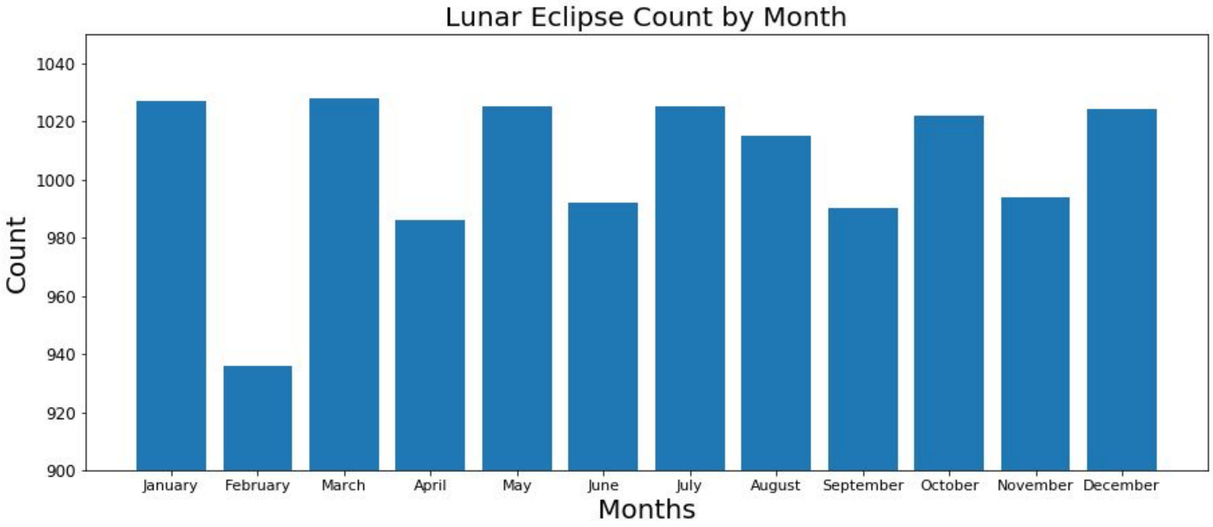
a. Is it different for solar vs lunar eclipses?

Solar and lunar eclipses seem to have relatively the same trend with every month as observed using data analysis plots.

b. If so why are they more/less common during certain months?

The trend seems directly correlated to how many days are in the month. More days in a month means more probability of an eclipse occurring during that month. That is why February seems to have the lowest number of eclipses since the month has the least number of days regardless of whether it is a leap year.





3. What months are total eclipses more common?

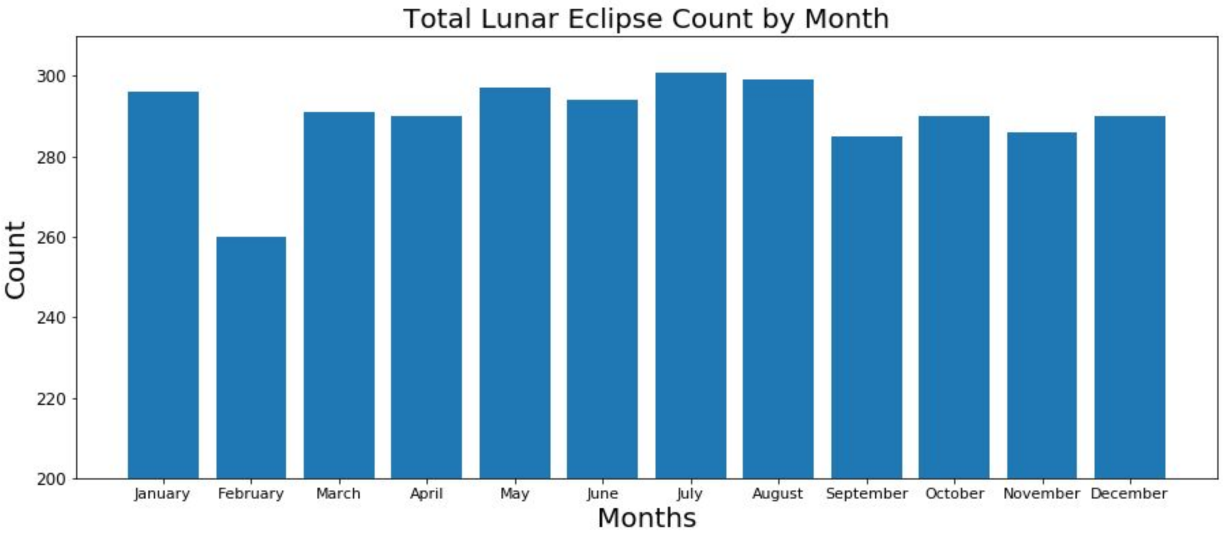
Since there are many types of eclipses, I want to focus on one type of eclipse. I will be focusing on a total eclipse for both lunar and solar eclipses. A total solar eclipse is characterized by the moon completely covering the sun in the sky. A total lunar eclipse is when the Earth's shadow completely covers the moon.

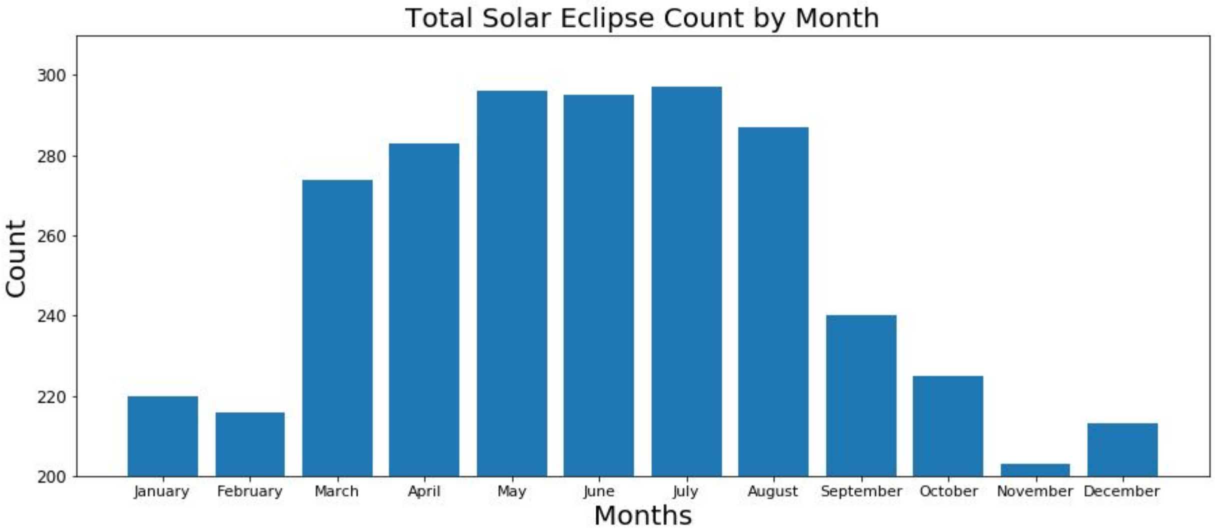
After analysis, I first found that there have been a few hundred more total lunar eclipses than solar eclipses in the past 5 millenniums. This implies that total lunar eclipses are more common.

a. Does the trend differ between lunar and solar eclipses?
   Total lunar eclipses seemed the follow the monthly trend that was seen earlier, where the number of eclipses was directly related to the number of days.

Total solar eclipses, however, seem to veer off this trend greatly. They seem to be more common during the second half of the year during the season's Fall and Winter (September - January).

This will be an area of further study using statistical testing to test whether month plays a role in the probability of a solar eclipse occurring.





Statistical Analysis

Is the probability of a total eclipse directly correlated to months in a year?

During the visual analysis above, it seemed that solar eclipses had a high count during the months of spring and summer and a low count during the months of fall and winter. But lunar eclipses seemed to be relatively the same count for all months.

For the statistical analysis, I have divided the data into two groups based on monthly analysis for both solar and lunar eclipses.

Group 1: September - February (Fall - Winter according to US seasons)

Group 2: March - August (Spring - Summer according to US season)

HYPOTHESIS TEST: To test the hypothesis that total eclipse is not directly correlated to the months in the year, we will assume the following null and alternative hypothesis. This hypothesis test will be performed for both solar and lunar eclipses.

Ho : The probability of total eclipse occurring is the same for Group 1 and Group 2. ($p1=p2$)

Ha : The probability of total eclipse occurring is not the same for Group 1 and Group 2.($p1\neq p2$)

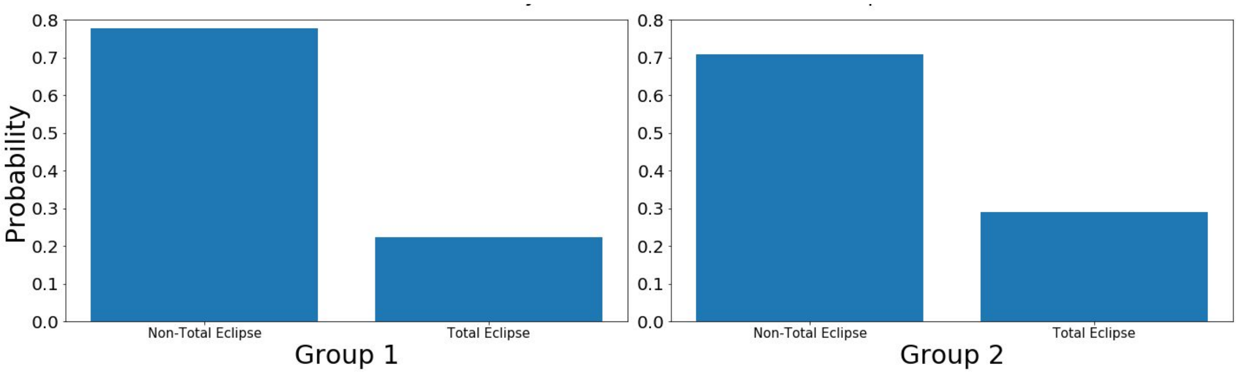
Hypothesis Test for Total Solar Eclipses

At first, I created dictionaries for each group, with {key: [month, year and values of 1 (total eclipse) and 0 (non-total eclipse)], …}. Using these dictionaries, the probability that the data was a total eclipse in each group was calculated using:

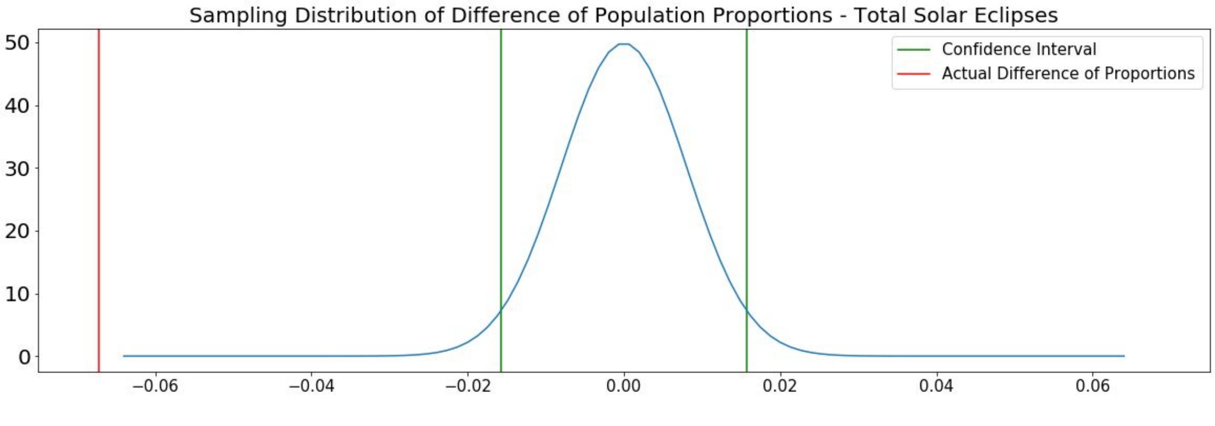
$$\frac{sum\ of\ the\ dictionary\ values\ (adding\ all\ the\ 1s\ from\ total\ eclipses)}{length\ dictionary\ (all\ eclipses)}$$

The probabilities were then used to plot a Bernoulli probability distribution for visual analysis. The x-axis consists of Total Eclipse (p1 and p2) and Non-Total Eclipses (1-p1 and 1-p2).

Bernoulli Probability Distribution of Total Solar Eclipses

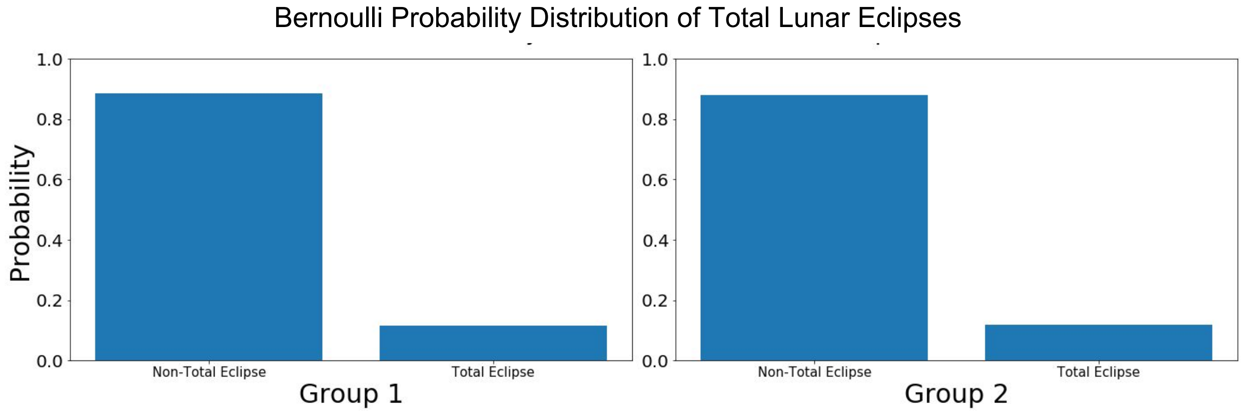


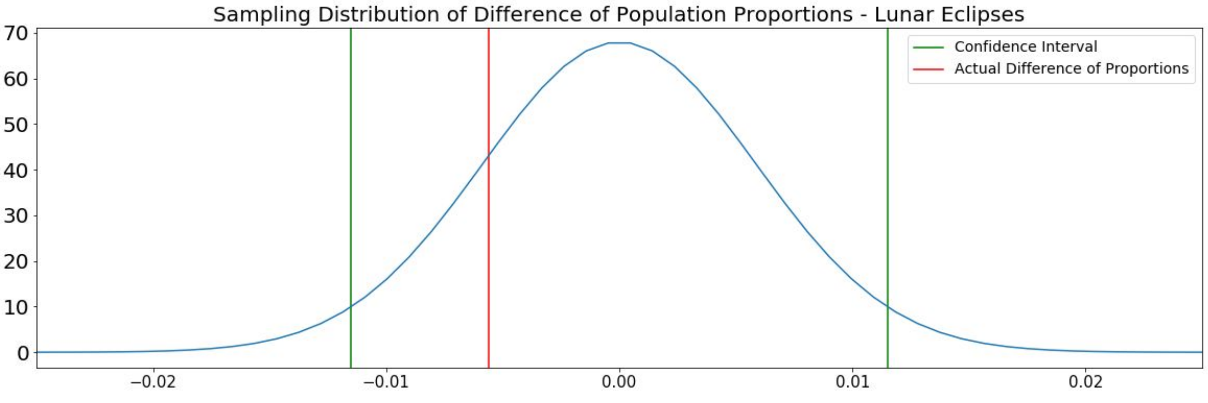
To perform the statistical frequentist test, I calculated a 95% confidence interval using the critical Z-value of 1.96 for a two-tail distribution. Using the calculated standard deviation difference between the two groups, the margin of error was calculated with z-value times the standard deviation difference: $z * \sigma_{p1-p2} \approx 0.01568$. Using the margin of error the confidence interval was calculated to be [-0.01568 0.01568] since the null hypothesis assumes that the mean of the distribution $\mu_{p_1-p_2} = p_1 - p_2 = 0$ since $p_1 = p_2$. The actual difference in proportions was calculated to be $\approx$ -0.06719, which is outside the confidence interval. Therefore, we can reject the null hypothesis which states that the total solar eclipses are not directly correlated to the months/seasons. All these calculations were plotted for better visualization.



Hypothesis Test for Total Lunar Eclipses

The same tests were performed on lunar eclipses. I calculated a 95% confidence interval using the critical Z-value of 1.96 for a two-tail distribution. Using the calculated standard deviation difference between the two groups, the margin of error was calculated with z-value times the standard deviation difference: $z * \sigma_{p1-p2} \approx 0.01149$. Using the margin of error the confidence interval was calculated to be [-0.01149 0.01149] since the null hypothesis assumes that the mean of the distribution $\mu_{p_1-p_2} = p_1 - p_2 = 0$ since $p_1 = p_2$. The actual difference in proportions was calculated to be $\approx$ -0.00560, which is inside the confidence interval. Therefore, we cannot reject the null hypothesis which states that the total solar eclipses are not directly correlated to the months/seasons. All these calculations were plotted for better visualization.





In-depth Analysis- Machine Learning

Problem

Astronomers need a metric to predict eclipses so that they can plan accordingly to gather data. The goal is to identify supervised learning techniques to predict the date of future total solar and lunar eclipses.

Hypothesis

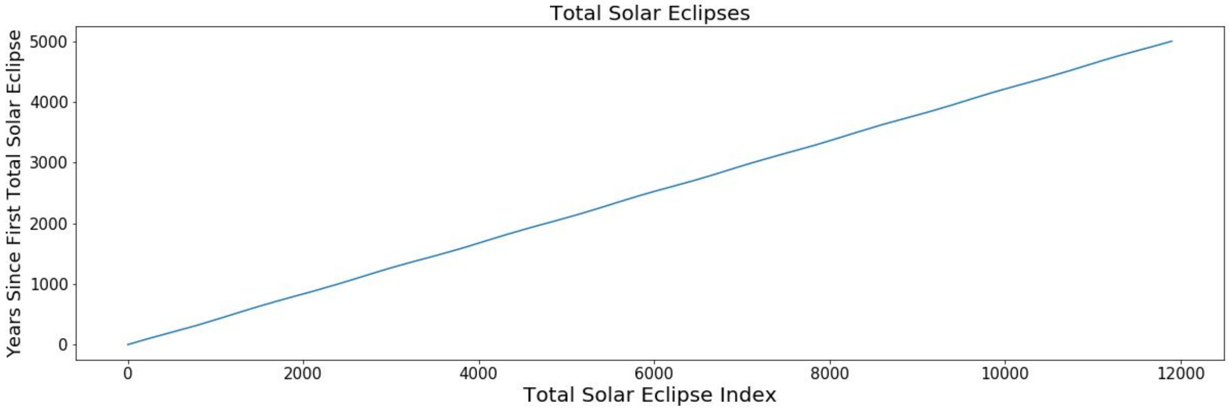
It is possible to predict the time and date of the solar and lunar eclipses from the given data leveraging machine learning regression algorithms.

Preparing Data for Analysis

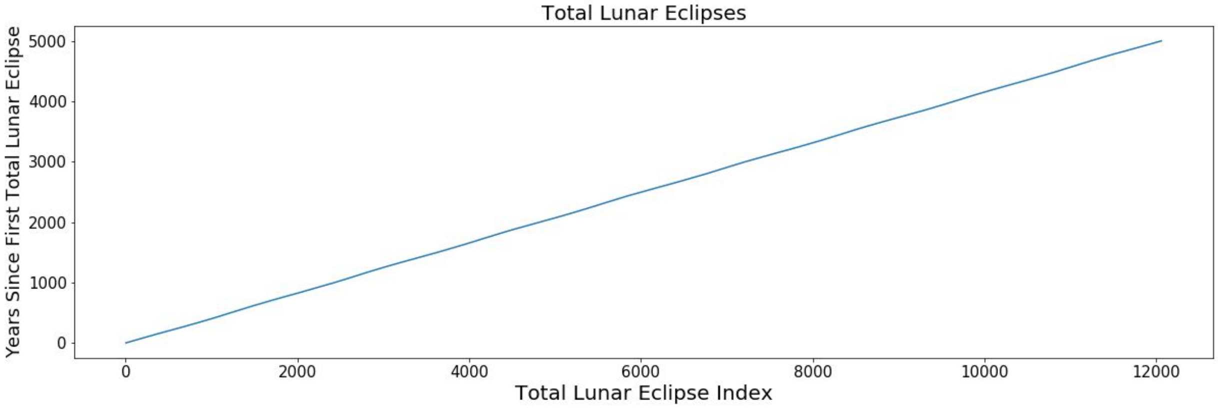
Since the statistical analysis focused on total eclipses, it was decided to do the same for the machine learning aspect of this project. The first step was to convert the total eclipse data into data that can be fit by the regression models. The feature X was defined as the eclipse index and the prediction variable y was the years from the start date of the data, which was 2000 BCE or -1999 in the csv file.

Actual Plots Before Fitting - For reference

Total Solar Eclipses



Total Lunar Eclipses



Models For Analysis

Since the data seemed to be very linear, I used a linear regression model to fit the data.

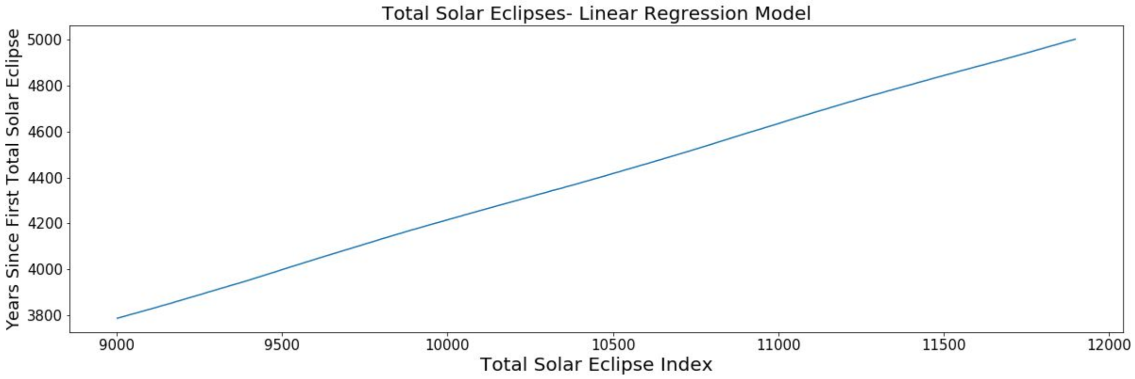
Linear Regression

Steps

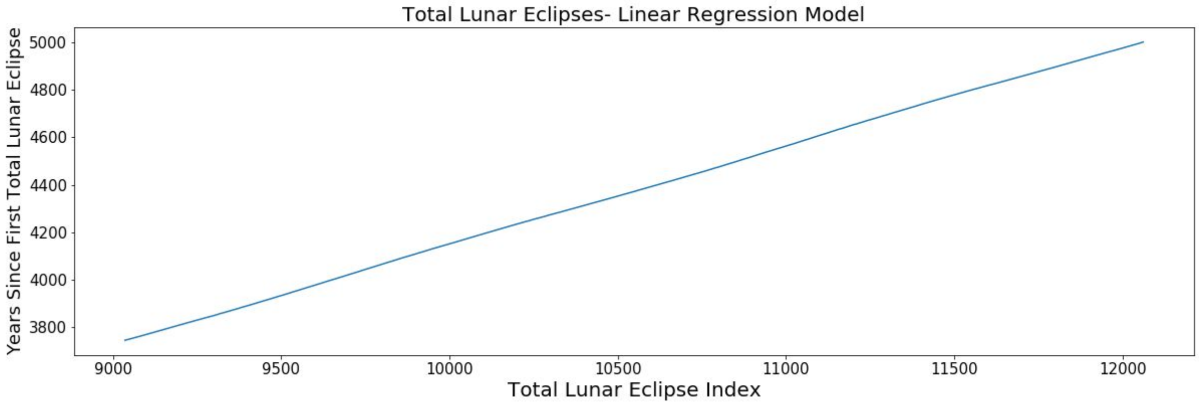
1. Created a function that converts the data into the appropriate linear regression format. X = total eclipse index , y = years from start date of dataset
2. Split the data in order into train and test sets. Train = 0.75 Test = 0.25
    a. It was done in order to keep the data in order of date.
3. Fitted the training data to linear regression model.
4. Predicted the model on testing data and computed root mean square error

Results

Total Solar Eclipses RMSE = 4.8 yrs



Total Lunar Eclipses RMSE = 5.4 yrs



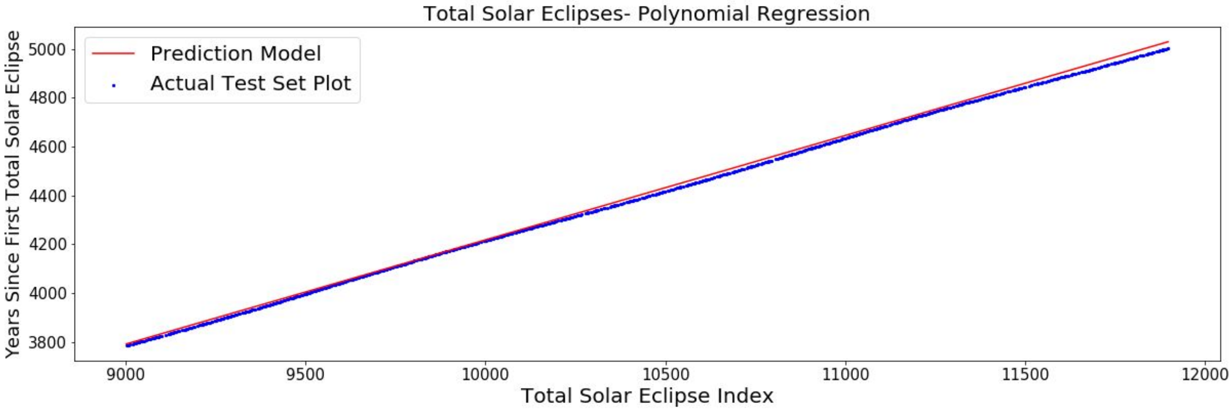
Polynomial Regression

Steps

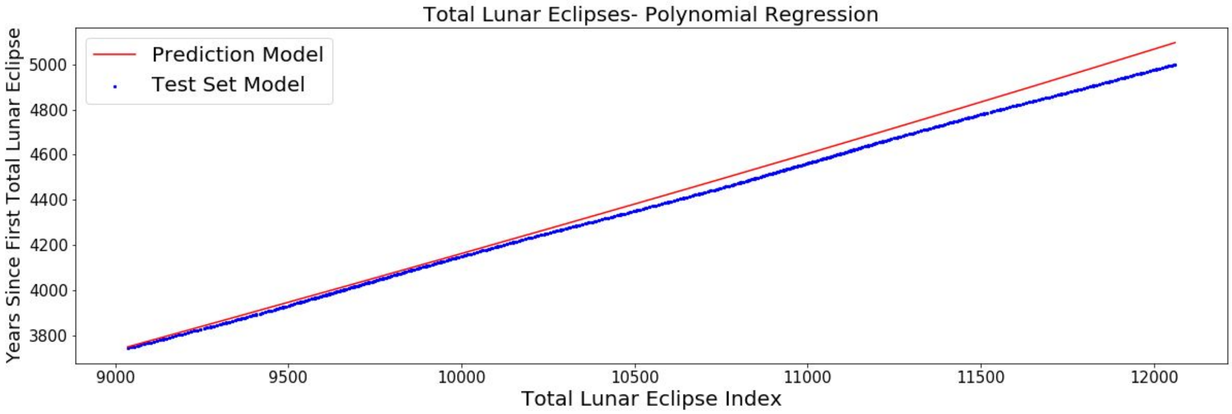
1. The data was already converted into the correct format and split into train and test sets in previous steps.
2. Performed a manual gridsearch to test for the best degree parameter for the Polynomial Features. Best degree = 4
3. Fitted the training data to polynomial regression model.
4. Predicted the model on testing data and computed root mean square error.

Results

Total Solar Eclipses RMSE = 4.6 yrs



Total Lunar Eclipses RMSE = 4.4 yrs



Ridge Regression

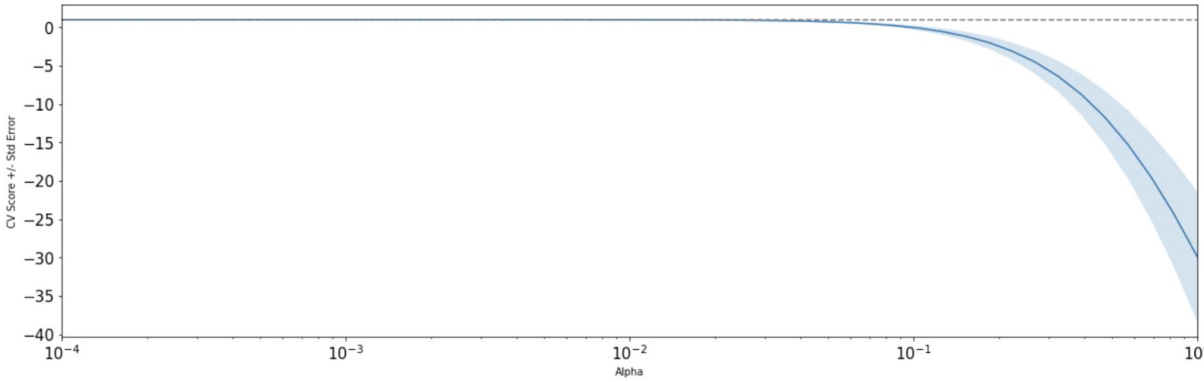
Steps

1. The data was already converted into the correct format and split into train and test sets in previous steps.
2. Performed GridSearchCV to find the best alpha parameter for the model and cross validation of 5.
3. Plotted the alphas vs the cross validation of 10 scores for each alpha for visual analysis.
4. Fitted the training data to ridge regression model.
5. Predicted the model on testing data and computed root mean square error.

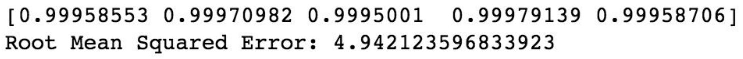
Results

Ridge regression is just linear regression with a normalization term to help reduce overfitting. For both solar and lunar eclipses models, it was observed that the best alpha parameter seemed to be the lowest alpha. It implies that the best model is one where you use as little regularization as possible.

Total Solar Eclipses RMSE = 4.9

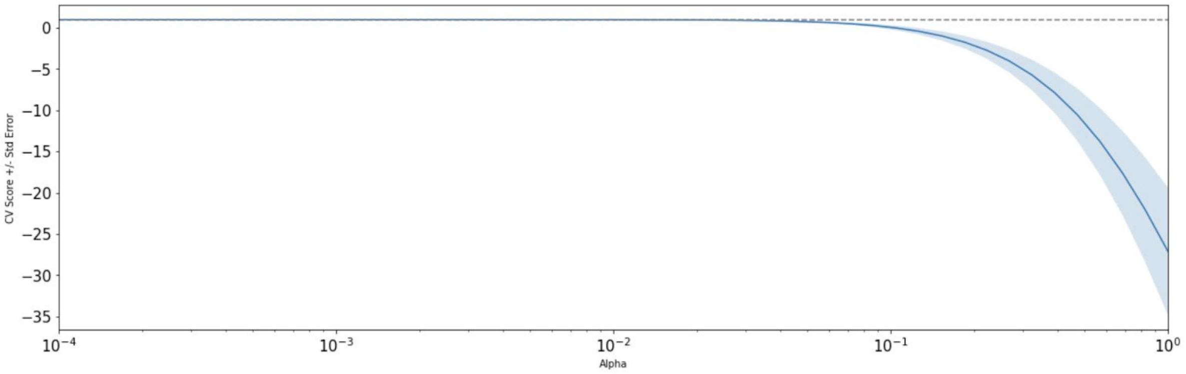


Grid search and RMSE results

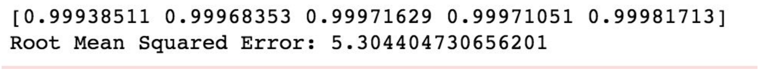


```
[0.99958553 0.99970982 0.9995001  0.99979139 0.99958706]
Root Mean Squared Error: 4.942123596833923
```

Total Lunar Eclipses RMSE = 5.3



Grid search and RMSE results



```
[0.99938511 0.99968353 0.99971629 0.99971051 0.99981713]
Root Mean Squared Error: 5.304404730656201
```

Lasso Regression

Steps

1. The data was already converted into the correct format and split into train and test sets in previous steps.
2. Performed GridSearchCV to find the best alpha parameter for the model and cross validation of 5.
3. Plotted the alphas vs the cross validation of 10 scores for each alpha for visual analysis.
4. Fitted the training data to lasso regression model.
5. Predicted the model on testing data and computed root mean square error.

Results

Lasso regression had a similar result as ridge regression with the best alpha parameter being the smallest alpha, or no normalization.

Total Solar Eclipses RMSE = 4.8 yrs

Total Lunar Eclipses RMSE = 5.3 yrs

Bayesian Regression

Steps

1. The data was already converted into the correct format and split into train and test sets in previous steps.
2. Performed GridSearchCV to find the best alpha and lambda parameters for the model with cross validation of 5.
3. Fitted the training data to bayesian regression model.
4. Predicted the model on testing data and computed root mean square error.

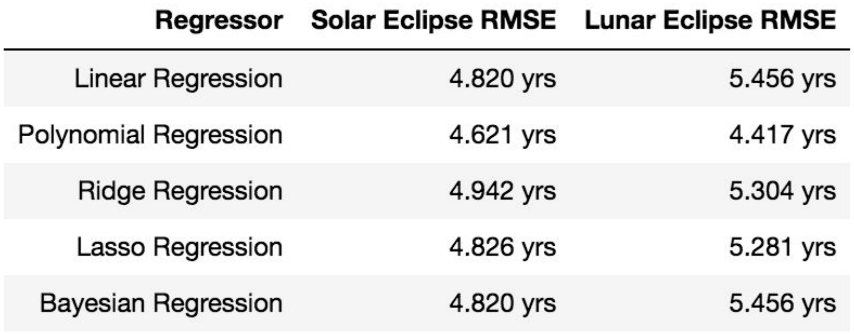
Results

Total Solar Eclipses RMSE = 4.8

Total Lunar Eclipses RMSE = 5.4

Conclusion

The results of our analysis using different Machine Learning models are tabulated below. This shows that our data is highly linear and hence polynomial regression performs poorly. The other regressors are all linear predictors - Ridge and Lasso regressions are just Linear Regression with L1 and L2 norm penalty terms and hence the hyperparameter for these learned from grid search result in convergence of these two models to linear regression. We see that Bayesian Regression also does not show an increase in performance, despite our 4-hyperparameter grid search. Hence, the best modelling technique given this data is probably linear regression - we conclude that further investigation into the data must be performed as the current data is not suitable for modelling.



Regressor	Solar Eclipse RMSE	Lunar Eclipse RMSE
Linear Regression	4.820 yrs	5.456 yrs
Polynomial Regression	4.621 yrs	4.417 yrs
Ridge Regression	4.942 yrs	5.304 yrs
Lasso Regression	4.826 yrs	5.281 yrs
Bayesian Regression	4.820 yrs	5.456 yrs